
Modin

Release 0.14.1+0.gd7eb019b.dirty

Modin contributors

May 04, 2022

CONTENTS

1	Installation	3
1.1	Installing with pip	3
1.2	Installing with conda	4
1.3	Installing from the GitHub master branch	5
1.4	Windows	5
1.5	Building Modin from Source	5
2	Using Modin	7
2.1	Using Modin Locally	7
2.2	Using Modin in a Cluster	9
3	Why Modin?	13
3.1	How does Modin differ from pandas?	13
3.2	Out-of-memory data with Modin	16
3.3	Modin vs. Dask DataFrame vs. Koalas	17
4	Examples and Resources	21
4.1	Usage Examples	21
4.2	Tutorials	21
4.3	Talks & Podcasts	22
4.4	Community contributions	22
5	Frequently Asked Questions (FAQs)	23
5.1	FAQs: Why choose Modin?	23
5.2	FAQs: How to use Modin?	24
6	Troubleshooting	27
6.1	Frequently encountered issues	27
6.2	Common errors	30
7	Quick Start Guide	33
8	Example: Instant Scalability with No Extra Effort	35
8.1	Faster Data Loading with read_csv	35
8.2	Faster concat across multiple dataframes	36
8.3	Faster apply over a single column	36
9	Summary	39
10	Usage Guide	41
10.1	Modin Configuration Settings	41

10.2	Modin Usage Examples	45
10.3	Advanced Usage	46
10.4	Optimization Notes	70
11	Supported APIs	75
11.1	Questions on implementation details	75
12	Development	91
12.1	Contributing	91
12.2	System Architecture	94
12.3	Partition API in Modin	131
12.4	pandas on Ray	132
12.5	pandas on Dask	133
12.6	pandas on Python	133
12.7	OmniSci	134
12.8	PyArrow on Ray	134
12.9	Modin SQL API	134
13	Contact	137
13.1	Slack	137
13.2	Discussion forum	137
13.3	Mailing List	137
13.4	Issues	137
14	Scale your pandas workflow by changing a single line of code	139
15	Installation and choosing your compute engine	141
16	Faster pandas, even on your laptop	143
17	Modin is a DataFrame for datasets from 1MB to 1TB+	145
	Python Module Index	147
	Index	149



Note:

Estimated Reading Time: 10 minutes

You can follow along this tutorial in a Jupyter notebook [here](#).

INSTALLATION

Note:

Estimated Reading Time: 15 minutes

If you already installed Modin on your machine, you can skip this section.

There are several ways to install Modin. Most users will want to install with `pip` or using `conda` tool, but some users may want to build from the master branch on the [GitHub repo](#). The master branch has the most recent patches, but may be less stable than a release installed from `pip` or `conda`.

1.1 Installing with pip

1.1.1 Stable version

Modin can be installed with `pip` on Linux, Windows and MacOS. To install the most recent stable release run the following:

```
pip install -U modin # -U for upgrade in case you have an older version
```

Modin can be used with *Ray*, *Dask*, or *OmniSci* engines. If you don't have *Ray* or *Dask* installed, you will need to install Modin with one of the targets:

```
pip install modin[ray] # Install Modin dependencies and Ray to run on Ray
pip install modin[dask] # Install Modin dependencies and Dask to run on Dask
pip install modin[all] # Install all of the above
```

Modin will automatically detect which engine you have installed and use that for scheduling computation! See below for OmniSci engine installation.

1.1.2 Release candidates

Before most major releases, we will upload a release candidate to test and check if there are any problems. If you would like to install a pre-release of Modin, run the following:

```
pip install --pre modin
```

These pre-releases are uploaded for dependencies and users to test their existing code to ensure that it still works. If you find something wrong, please raise an [issue](#) or email the bug reporter: bug_reports@modin.org.

1.1.3 Installing specific dependency sets

Modin has a number of specific dependency sets for running Modin on different execution engines and storage formats or for different functionalities of Modin. Here is a list of dependency sets for Modin:

```
pip install "modin[ray]" # If you want to use the Ray execution engine
```

```
pip install "modin[dask]" # If you want to use the Dask execution engine
```

1.1.4 Installing on Google Colab

Modin can be used with Google [Colab](#) via the `pip` command, by running the following code in a new cell:

```
!pip install modin[all]
```

Since Colab preloads several of Modin's dependencies by default, we need to restart the Colab environment once Modin is installed by either clicking on the "RESTART RUNTIME" button in the installation output or by run the following code:

```
# Post-install automatically kill and restart Colab environment
import os
os.kill(os.getpid(), 9)
```

Once you have restarted the Colab environment, you can use Modin in Colab in subsequent sessions.

Note that on the free version of Colab, there is a [limit on the compute resource](#). To leverage the full power of Modin, you may have to upgrade to Colab Pro to get access to more compute resources.

1.2 Installing with conda

1.2.1 Using conda-forge channel

Modin releases can be installed using conda from conda-forge channel. Starting from 0.10.1 it is possible to install modin with chosen engine(s) alongside. Current options are:

Package name in conda-forge	Engine(s)	Supported OSs
modin	Dask	Linux, Windows, MacOS
modin-dask	Dask	Linux, Windows, MacOS
modin-ray	Ray	Linux, Windows
modin-omnisci	OmniSci	Linux
modin-all	Dask, Ray, OmniSci	Linux

For installing Dask and Ray engines into conda environment following command should be used:

```
conda install -c conda-forge modin-ray modin-dask
```

All set of engines could be available in conda environment by specifying:

```
conda install -c conda-forge modin-all
```

or explicitly:

```
conda install -c conda-forge modin-ray modin-dask modin-omnisci
```

conda may be slow installing modin-omnisci and hence modin-all packages so it's worth trying to set `channel_priority` to `strict` prior the installation process:

```
conda config --set channel_priority strict
```

1.2.2 Using Intel® Distribution of Modin

With conda it is also possible to install [Intel Distribution of Modin](#), a special version of Modin that is part of Intel® oneAPI AI Analytics Toolkit. This version of Modin is powered by *OmniSci* engine that contains a bunch of optimizations for Intel hardware. More details to get started can be found in the [Intel Distribution of Modin Getting Started](#) guide.

1.3 Installing from the GitHub master branch

If you'd like to try Modin using the most recent updates from the master branch, you can also use `pip`.

```
pip install git+https://github.com/modin-project/modin
```

This will install directly from the repo without you having to manually clone it! Please be aware that these changes have not made it into a release and may not be completely stable.

1.4 Windows

All Modin engines except *OmniSci* are available both on Windows and Linux as mentioned above. Default engine on Windows is *Ray*. It is also possible to use Windows Subsystem For Linux ([WSL](#)), but this is generally not recommended due to the limitations and poor performance of Ray on WSL, a roughly 2-3x worse than native Windows.

1.5 Building Modin from Source

If you're planning on [contributing](#) to Modin, you will need to ensure that you are building Modin from the local repository that you are working off of. Occasionally, there are issues in overlapping Modin installs from pypi and from source. To avoid these issues, we recommend uninstalling Modin before you install from source:

```
pip uninstall modin
```

To build from source, you first must clone the repo. We recommend forking the repository first through the GitHub interface, then cloning as follows:

```
git clone https://github.com/<your-github-username>/modin.git
```

Once cloned, cd into the modin directory and use pip to install:

```
cd modin  
pip install -e .
```

USING MODIN

In this section, we show how Modin can be used to accelerate your pandas workflows on a single machine up to multiple machines in a cluster setting.

2.1 Using Modin Locally

Note:

Estimated Reading Time: 5 minutes

You can follow along this tutorial in a Jupyter notebook *here*

<<https://github.com/modin-project/modin/tree/master/examples/quickstart.ipynb>>.

In our quickstart example, we have already seen how you can achieve considerable speedup from Modin, even on a single machine. Users do not need to know how many cores their system has, nor do they need to specify how to distribute the data. In fact, users can **continue using their existing pandas code** while experiencing a considerable speedup from Modin, even on a single machine.

To use Modin on a single machine, only a modification of the import statement is needed. Once you've changed your import statement, you're ready to use Modin just like you would pandas, since the API is identical to pandas.

```
# import pandas as pd
import modin.pandas as pd
```

That's it. You're ready to use Modin on your previous pandas workflows!

2.1.1 Optional Configurations

When using Modin locally on a single machine or laptop (without a cluster), Modin will automatically create and manage a local Dask or Ray cluster for the executing your code. So when you run an operation for the first time with Modin, you will see a message like this, indicating that a Modin has automatically initialized a local cluster for you:

```
df = pd.DataFrame({'col1': [1, 2], 'col2': [3, 4]})
```

```
UserWarning: Ray execution environment not yet initialized. Initializing...
To remove this warning, run the following python code before doing dataframe
operations:
```

```
import ray
```

(continues on next page)

(continued from previous page)

```
ray.init()
```

If you prefer to use Dask over Ray as your execution backend, you can use the following code to modify the default configuration:

```
import modin
modin.config.Engine.put("Dask")
```

```
df = pd.DataFrame({'col1': [1, 2], 'col2': [3, 4]})
```

UserWarning: Dask execution environment not yet initialized. Initializing...
To remove this warning, run the following python code before doing dataframe operations:

```
from distributed import Client

client = Client()
```

Finally, if you already have an Ray or Dask engine initialized, Modin will automatically attach to whichever engine is available. If you are interested in using Modin with OmniSci engine, please refer to [these instructions](#). For additional information on other settings you can configure, see [this page](#) for more details.

2.1.2 Advanced: Configuring the resources Modin uses

Modin automatically check the number of CPUs available on your machine and sets the number of partitions to be equal to the number of CPUs. You can verify this by running the following code:

```
import modin
print(modin.config.NPartitions.get()) #prints 16 on a laptop with 16 physical cores
```

Modin fully utilizes the resources on your machine. To read more about how this works, see [this page](#) for more details.

Since Modin will use all of the resources available on your machine by default, at times, it is possible that you may like to limit the amount of resources Modin uses to free resources for another task or user. Here is how you would limit the number of CPUs Modin used in your bash environment variables:

```
export MODIN_CPUS=4
```

You can also specify this in your python script with `os.environ`:

```
import os
os.environ["MODIN_CPUS"] = "4"
import modin.pandas as pd
```

If you're using a specific engine and want more control over the environment Modin uses, you can start Ray or Dask in your environment and Modin will connect to it.

```
import ray
ray.init(num_cpus=4)
import modin.pandas as pd
```

Specifying `num_cpus` limits the number of processors that Modin uses. You may also specify more processors than you have available on your machine; however this will not improve the performance (and might end up hurting the performance of the system).

Note: Make sure to update the MODIN_CPUS configuration and initialize your preferred engine before you start working with the first operation using Modin! Otherwise, Modin will opt for the default setting.

2.2 Using Modin in a Cluster

Note:

Estimated Reading Time: 15 minutes

You can follow along in a Jupyter notebook in this two-part tutorial: [\[Part 1\]](#), [\[Part 2\]](#).

Often in practice we have a need to exceed the capabilities of a single machine. Modin works and performs well in both local mode and in a cluster environment. The key advantage of Modin is that your notebook does not change between local development and cluster execution. Users are not required to think about how many workers exist or how to distribute and partition their data; Modin handles all of this seamlessly and transparently.

2.2.1 Starting up a Ray Cluster

Modin is able to utilize Ray's built-in autoscaled cluster. To launch a Ray cluster using Amazon Web Service (AWS), you can use [this file](#) as the config file.

```
pip install boto3
aws configure
```

To start up the Ray cluster, run the following command in your terminal:

```
ray up modin-cluster.yaml
```

This configuration script starts 1 head node (m5.24xlarge) and 7 workers (m5.24xlarge), 768 total CPUs. For more information on how to launch a Ray cluster across different cloud providers or on-premise, you can also refer to the Ray documentation [here](#).

2.2.2 Connecting to a Ray Cluster

To connect to the Ray cluster, run the following command in your terminal:

```
ray attach modin-cluster.yaml
```

The following code checks that the Ray cluster is properly configured and attached to Modin:

```
import ray
ray.init(address="auto")
from modin.config import NPartitions
assert NPartitions.get() == 768, "Not all Ray nodes are started up yet"
ray.shutdown()
```

Congratulations! You have successfully connected to the Ray cluster. See more on the Modin in the Cloud documentation page.

2.2.3 Using Modin on a Ray Cluster

Now that we have a Ray cluster up and running, we can use Modin to perform pandas operation as if we were working with pandas on a single machine. We test Modin's performance on the 120GB [NYC Taxi dataset](#) that was provided as part of our [cluster setup script](#). We can time the following operation in a Jupyter notebook:

```
%%time
df = pd.read_csv("big_yellow.csv", quoting=3)

%%time
count_result = df.count()

%%time
groupby_result = df.groupby("passenger_count").count()

%%time
apply_result = df.applymap(str)
```

Modin performance scales as the number of nodes and cores increases. The following chart shows the performance of the above operations with 2, 4, and 8 nodes, with improvements in performance as we increase the number of resources Modin can use.



2.2.4 Advanced: Configuring your Ray Environment

In some cases, it may be useful to customize your Ray environment. Below, we have listed a few ways you can solve common problems in data management with Modin by customizing your Ray environment. It is possible to use any of Ray's initialization parameters, which are all found in [Ray's documentation](#).

```
import ray
ray.init()
import modin.pandas as pd
```

Modin will automatically connect to the Ray instance that is already running. This way, you can customize your Ray environment for use in Modin!

WHY MODIN?

In this section, we explain the design and motivation behind Modin and why you should use Modin to scale up your pandas workflows. We first describe the architectural differences between pandas and Modin. Then we describe how Modin can also help resolve out-of-memory issues common to pandas. Finally, we look at the key differences between Modin and other distributed dataframe libraries.

3.1 How does Modin differ from pandas?

Note:

Estimated Reading Time: 10 minutes

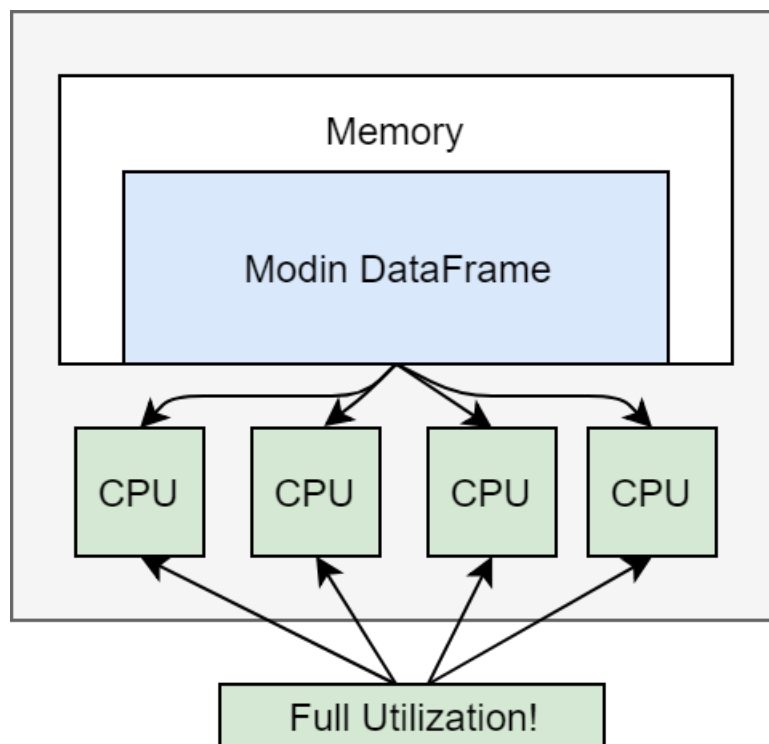
In the earlier tutorials, we have seen how Modin can be used to speed up pandas workflows. Here, we discuss at a high level how Modin works, in particular, how Modin's dataframe implementation differs from pandas.

3.1.1 Scalability of implementation

Modin exposes the pandas API through `modin.pandas`, but it does not inherit the same pitfalls and design decisions that make it difficult to scale. The pandas implementation is inherently single-threaded. This means that only one of your CPU cores can be utilized at any given time. In a laptop, it would look something like this with pandas:

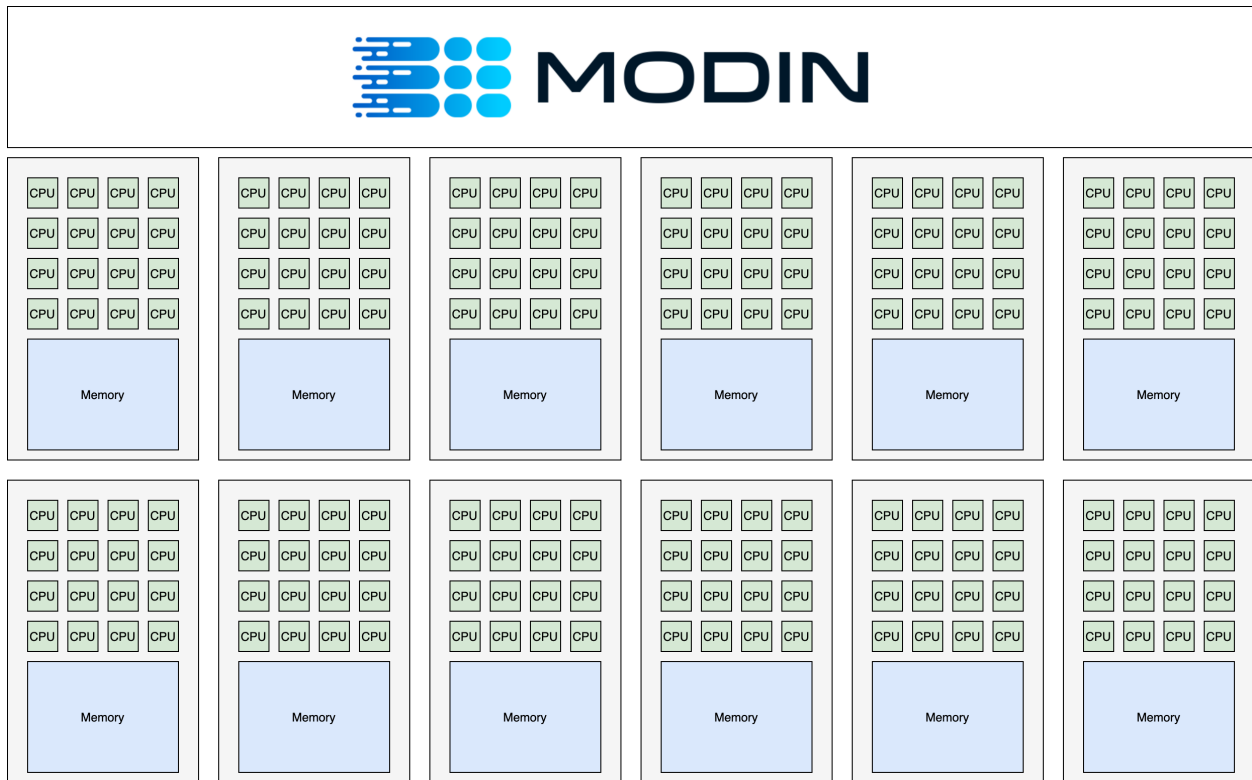


However, Modin's implementation enables you to use all of the cores on your machine, or all of the cores in an entire cluster. On a laptop, it will look something like this:



The additional utilization leads to improved performance, however if you want to scale to an entire cluster, Modin

suddenly looks something like this:



Modin is able to efficiently make use of all of the hardware available to it!

3.1.2 Memory usage and immutability

The pandas API contains many cases of “inplace” updates, which are known to be controversial. This is due in part to the way pandas manages memory: the user may think they are saving memory, but pandas is usually copying the data whether an operation was inplace or not.

Modin allows for inplace semantics, but the underlying data structures within Modin’s implementation are immutable, unlike pandas. This immutability gives Modin the ability to internally chain operators and better manage memory layouts, because they will not be changed. This leads to improvements over pandas in memory usage in many common cases, due to the ability to share common memory blocks among all dataframes.

Modin provides the inplace semantics by having a mutable pointer to the immutable internal Modin dataframe. This pointer can change, but the underlying data cannot, so when an inplace update is triggered, Modin will treat it as if it were not inplace and just update the pointer to the resulting Modin dataframe.

3.1.3 API vs implementation

It is well known that the pandas API contains many duplicate ways of performing the same operation. Modin instead enforces that any one behavior have one and only one implementation internally. This guarantee enables Modin to focus on and optimize a smaller code footprint while still guaranteeing that it covers the entire pandas API. Modin has an internal algebra, which is roughly 15 operators, narrowed down from the original >200 that exist in pandas. The algebra is grounded in both practical and theoretical work. Learn more in our [VLDB 2020 paper](#). More information about this algebra can be found in the [architecture](#) documentation.

3.2 Out-of-memory data with Modin

Note:

Estimated Reading Time: 10 minutes

When using pandas, you might run into a memory error if you are working with large datasets that cannot fit in memory or perform certain memory-intensive operations (e.g., joins).

Modin solves this problem by spilling over to disk, in other words, it uses your disk as an overflow for memory so that you can work with datasets that are too large to fit in memory. By default, Modin leverages out-of-core methods to handle datasets that don't fit in memory for both Ray and Dask engines.

3.2.1 Motivating Example: Memory error with pandas

pandas makes use of in-memory data structures to store and operate on data, which means that if you have a dataset that is too large to fit in memory, it will cause an error on pandas. As an example, let's create a 80GB DataFrame by appending together 40 different 2GB DataFrames.

```
import pandas
import numpy as np
df = pandas.concat([pandas.DataFrame(np.random.randint(0, 100, size=(2**20, 2**8))) for _
↪ in range(40)]) # Memory Error!
```

When we run this on a laptop with 32GB of RAM, pandas will run out of memory and throw an error (e.g., `MemoryError, Killed: 9`).

The [pandas documentation](#) has a great section on recommendations for scaling your analysis to these larger datasets. However, this generally involves loading in less data or rewriting your pandas code to process the data in smaller chunks.

3.2.2 Operating on out-of-memory data with Modin

In order to work with data that exceeds memory constraints, you can use Modin to handle these large datasets.

```
import modin.pandas as pd
import numpy as np
df = pd.concat([pd.DataFrame(np.random.randint(0, 100, size=(2**20, 2**8))) for _ in
↪ range(40)]) # 40x2GB frames -- Working!
df.info()
```

Not only does Modin let you work with datasets that are too large to fit in memory, we can perform various operations on them without worrying about memory constraints.

3.2.3 Advanced: Configuring out-of-core settings

By default, out-of-core functionality is enabled by the compute engine selected. To disable it, start your preferred compute engine with the appropriate arguments. For example:

```
import modin.pandas as pd
import ray

ray.init(_plasma_directory="/tmp") # setting to disable out of core in Ray
df = pd.read_csv("some.csv")
```

If you are using Dask, you have to modify local configuration files. Visit the [Dask documentation](#) on object spilling for more details.

3.3 Modin vs. Dask DataFrame vs. Koalas

Libraries such as [Dask DataFrame](#) (DaskDF for short) and [Koalas](#) aim to support the pandas API on top of distributed computing frameworks, Dask and Spark respectively. Instead, Modin aims to preserve the pandas API and behavior as is, while abstracting away the details of the distributed computing framework underneath. Thus, the aims of these libraries are fundamentally different.

Specifically, Modin enables pandas-like

- row and column-parallel operations, unlike DaskDF and Koalas that only support row-parallel operations
- indexing & ordering semantics, unlike DaskDF and Koalas that deviate from these semantics
- eager execution, unlike DaskDF and Koalas that provide lazy execution

As a result, Modin's coverage is [more than 90%](#) of the pandas API, while DaskDF and Koalas' coverage is about 55%.

For more technical details please see our VLDB 2022 research paper, referenced [here](#).

3.3.1 Brief Overview of DaskDF and Koalas

Dask's [DataFrame](#) (DaskDF) is effectively a meta-DataFrame, partitioning and scheduling many smaller `pandas.DataFrame` objects. Users construct a task graph of dataframe computation step by step and then trigger computation using the `compute` function.

Spark's [Koalas](#) provides the pandas API on Spark, leveraging the preexisting Spark SQL optimizer to execute select pandas commands. Like DaskDF, Koalas also employs lazy computation, only triggering computation when the user requests to see the results.

3.3.2 Partitioning and Parallelization

Modin, DaskDF, Koalas are all examples of parallel dataframe systems. Parallelism is achieved by partitioning a large dataframe into smaller ones that can be operated on in parallel. As a result, the partitioning scheme chosen by the system dictates the pandas functions that can or can not be supported.

DaskDF and Koalas only support row-oriented partitioning and parallelism. This approach is analogous to relational databases. The dataframe is conceptually broken down into horizontal partitions along rows, where each partition is independently processed if possible. When DaskDF or Koalas are required to perform column-parallel operations

that to be done on columns independently (e.g., dropping columns with null values via `dropna` on the column axis), they either perform very poorly with no parallelism or do not support that operation.

Modin supports both row, column, and cell-oriented partitioning and parallelism. That is, the dataframe can be conceptually broken down as groups of rows, groups of columns, or both groups of rows and groups of columns (effectively a block or sub-matrix). Modin will transparently reshape the partitioning as necessary for the corresponding operation, based on whether the operation is row-parallel, column-parallel, or cell-parallel (independently applied to each unit cell). This allows Modin to support more of the pandas API and do so efficiently. Due to the finer-grained control over the partitioning, Modin can support a number of operations that are very challenging to parallelize in row-oriented systems (e.g., transpose, median, quantile). This flexibility in partitioning also gives Modin tremendous power to implement efficient straggler mitigation and improve utilization over the entire cluster.

3.3.3 API Coverage

One of the key benefits of pandas is its versatility, due to the wide array of operations, with more than 600+ API operations for data cleaning, feature engineering, data transformation, data summarization, data exploration, and machine learning. However, it is not trivial to develop scalable implementations of each of these operations in a dataframe system. **DaskDF and Koalas only implements about 55% of the pandas API;** they do not implement certain APIs that would deviate from the row-wise partitioning approach, or would be inefficient with the row-wise parallelization. For example, Dask does not implement `iloc`, `MultiIndex`, `apply(axis=0)`, `quantile` (only approximate quantile is available), `median`, and more. Given DaskDF's row-oriented architecture, `iloc`, for example, can technically be implemented, but it would be inefficient, and column-wise operations such as `apply(axis=0)` would be impossible to implement. Similarly, Koalas does not implement `apply(axis=0)` (it only applies the function per row partition, giving a different result), `quantile`, `median` (only approximate quantile/median is available), `MultiIndex`, `combine`, `compare` and more.

Modin supports all of the above pandas API functions, as well as others, with more than 90% coverage of the pandas API. Modin additionally acts as a drop-in replacement for pandas, such that even if the API is not yet supported, it still works by falling back to running vanilla pandas. One of the key features of being a drop-in replacement is that not only will it work for existing code, if a user wishes to go back to running pandas directly, they are not locked in to using Modin and can switch between Modin and pandas at no cost. In other words, scripts and notebooks written in Modin can be converted to and from pandas as the user desires by simply replacing the import statement.

3.3.4 Execution Semantics

DaskDF and Koalas make use of lazy evaluation, which means that the computation is delayed until users explicitly evaluate the results. This mode of evaluation places a lot of optimization responsibility on the user, forcing them to think about when it would be useful to inspect the intermediate results or delay doing so. Specifically, DaskDF's API differs from pandas in that it requires users to explicitly call `.compute()` to materialize the result of the computation. Often if that computation corresponds to a long chain of operators, this call can take a very long time to execute. Overall, the need to explicitly trigger computation makes the API less convenient to work with, but gives DaskDF and Koalas the opportunity to perform holistic optimizations over the entire dataflow graph. However, to the best of our knowledge, neither DaskDF nor Koalas actually leverage holistic optimizations.

Modin employs eager evaluation, like pandas. Eager evaluation is the default mode of operation for data scientists when working with pandas in an interactive environment, such as Jupyter Notebooks. Modin reproduces this familiar behavior by performing all computations eagerly as soon as it is issued, so that users can inspect intermediate results and quickly see the results of their computations without having to wait or explicitly trigger computation. This is especially useful during interactive data analysis, where users often iterate on their dataframe workflows or build up their dataframe queries in an incremental fashion. Modin also supports lazy evaluation via the OmniSci engine, you can learn more about it here. We also have developed techniques for [opportunistic evaluation](#) that bridges the gap between lazy and eager evaluation that will be incorporated in Modin in the future.

3.3.5 Ordering Semantics

By default, pandas preserves the order of the dataframe, so that users can expect a consistent, ordered view as they are operating on their dataframe.

Both DaskDF and Koalas make no guarantees about the order of rows in the DataFrame. This is because DaskDF sorts the index for optimization purposes to speed up computations that involve the row index; and as a result, it does not support user-specified order. Likewise, Koalas [does not support ordering](#) by default because it will lead to a performance overhead when operating on distributed datasets.

DaskDF additionally does not support multi-indexing or sorting. DaskDF sorts the data based on a single set of row labels for fast row lookups, and builds an indexing structure based on these row labels. Data is both logically and physically stored in the same order. As a result, DaskDF does not support a *sort* function.

Modin reproduces the intuitive behavior in pandas where the order of the DataFrame is preserved, and supports multi-indexing. Enforcing ordering on a parallel dataframe system like Modin requires non-trivial effort that involves decoupling of the logical and physical representation of the data, enabling the order to be lazily kept up-to-date, but eagerly computed based on user needs (See Section 4.2 in [our recent paper](#)). Modin abstracts away the physical representation of the data and provides an ordered view that is consistent with user's expectations.

3.3.6 Compatibility with Computational Frameworks

DaskDF and Koalas are meant to be run on Dask and Spark respectively. They are highly tuned to the corresponding frameworks, and cannot be ported to other computational frameworks.

Modin's highly modular design is architected to run on a variety of systems, and support a variety of APIs. The goal for the extensible design is that users can take the same notebook or script and seamlessly move between different clusters and environments, with Modin being able to support the pandas API on your preexisting infrastructure. Currently, Modin support running on Dask's compute engine in addition to Ray. The modular design makes it easier for developers to different execution engines or compile to different memory formats. Modin can run on a Dask cluster in the same way that DaskDF can, but they differ in the ways described above. In addition, Modin is continually expanding to support popular data processing APIs (SQL in addition to pandas, among other DSLs for data processing) while leveraging the same underlying execution framework. Modin's flexible architecture also means that as the [pandas API continues to evolve](#), Modin can quickly move towards supporting new versions of the pandas API.

3.3.7 Performance Comparison

On operations supported by all systems, Modin provides substantial speedups. Thanks to its optimized design, Modin is able to take advantage of multiple cores relative to both Koalas and DaskDF to efficiently execute pandas operations. It is notable that Koalas is often slower than pandas, due to the overhead of Spark.

Modin provides substantial speedups even on operators not supported by other systems. Thanks to its flexible partitioning schemes that enable it to support the vast majority of pandas operations — be it row, column, or cell-oriented - Modin provides benefits on operations such as `join`, `median`, and `infer_types`. While Koalas performs `join` slower than Pandas, Dask failed to support `join` on more than 20M rows, likely due poor support for [shuffles](#). Details of the benchmark and additional join experiments can be found in [our paper](#).

Modin is built on many years of research and development at UC Berkeley. For more information on how this works underneath the hoods, check out our publications in this space:

- [Flexible Rule-Based Decomposition and Metadata Independence in Modin \(VLDB 2021\)](#)
- [Enhancing the Interactivity of Dataframe Queries by Leveraging Think Time \(IEEE Data Eng 2021\)](#)

- [Dataframe Systems: Theory, Architecture, and Implementation](#) (PhD Dissertation 2021)
- [Scaling Data Science does not mean Scaling Machines](#) (CIDR 2021)
- [Towards Scalable Dataframe Systems](#) (VLDB 2020)

EXAMPLES AND RESOURCES

Here you can find additional resources to learn about Modin. To learn more about advanced usage for Modin, please refer to [this section](#).

4.1 Usage Examples

The following notebooks demonstrate how Modin can be used for scalable data science:

- Quickstart Guide to Modin [[Source](#)]
- Using Modin with the NYC Taxi Dataset [[Source](#)]
- Modin for Machine Learning with scikit-learn [[Source](#)]

4.2 Tutorials

The following tutorials cover the basic usage of Modin. [Here](#) is a one hour video tutorial that walks through these basic exercises.

- Exercise 1: Introduction to Modin [[Source PandasOnRay](#), [Source PandasOnDask](#), [Source Omnis-ciOnNative](#)]
- Exercise 2: Speed Improvements with Modin [[Source PandasOnRay](#), [Source PandasOnDask](#), [Source Omnis-ciOnNative](#)]
- Exercise 3: Defaulting to pandas with Modin [[Source PandasOnRay](#), [Source PandasOnDask](#), [Source Omnis-ciOnNative](#)]

The following tutorials covers more advanced features in Modin:

- Exercise 4: Experimental Features in Modin (Spreadsheet, Progress Bar) [[Source PandasOnRay](#), [Source PandasOnDask](#)]
- Exercise 5: Setting up Modin in a Cluster Environment [[Source PandasOnRay](#)]
- Exercise 6: Running Modin in a Cluster Environment [[Source PandasOnRay](#)]

How to get required dependencies for the tutorial notebooks and to run them please refer to the respective [README.md](#) file.

4.3 Talks & Podcasts

- [Scaling Interactive Data Science with Modin and Ray](#) (20 minute, Ray Summit 2021)
- [Unleash The Power Of Dataframes At Any Scale With Modin](#) (40 minute, Python Podcast 2021)
- [\[Russian\] Distributed Data Processing and XGBoost Training and Prediction with Modin](#) (30 minute, PyCon Russia 2021)
- [\[Russian\] Efficient Data Science with Modin](#) (30 minute, ISP RAS Open 2021)
- [Modin: Scaling the Capabilities of the Data Scientist, not the Machine](#) (1 hour, RISE Camp 2020)
- [Modin: Pandas Scalability with Devin Petersohn](#) (1 hour, Software Engineering Daily Podcast 2020)
- [Introduction to the DataFrame and Modin](#) (20 minute, RISECamp 2019)
- [Scaling Interactive Pandas Workflows with Modin](#) (40 minute, PyData NYC 2018)

4.4 Community contributions

Here are some blogposts and articles about Modin:

- [Anaconda Blog: Scale your pandas workflow with Modin](#) by Vasilij Litvinov
- [The Modin view of Scaling Pandas](#) by Devin Petersohn
- [Data Science at Scale with Modin](#) by Areg Melik-Adamyan
- [Speed up Pandas using Modin](#) by Eric D. Brown, D.Sc.
- [Explore Python Libraries: Make Your DataFrames Parallel With Modin](#) by Zachary Bennett
- [Get faster pandas with Modin, even on your laptops](#) by Parul Pandey
- [How to speedup pandas by changing one line of code](#) by Shrivarsheni
- [How To Accelerate Pandas With Just One Line Of Code](#) by Analytics India

Here are some articles contributed by the international community:

- [\[Chinese\] Modin pandas](#) by Python Chinese Community
- [\[German\] Was ist Modin?](#) by Dipl.-Ing. (FH) Stefan Luber
- [\[Russian\] Pandas modin](#) by
- [\[Korean\] modin pandas](#) by

If you would like your articles to be featured here, please [submit a pull request](#) to let us know!

FREQUENTLY ASKED QUESTIONS (FAQS)

Below, you will find answers to the most commonly asked questions about Modin. If you still cannot find the answer you are looking for, please post your question on the #support channel on our [Slack](#) community or open a Github [issue](#).

5.1 FAQs: Why choose Modin?

5.1.1 What's wrong with pandas and why should I use Modin?

While pandas works extremely well on small datasets, as soon as you start working with medium to large datasets that are more than a few GBs, pandas can become painfully slow or run out of memory. This is because pandas is single-threaded. In other words, you can only process your data with one core at a time. This approach does not scale to larger data sets and adding more hardware does not lead to more performance gain.

The [DataFrame](#) is a highly scalable, parallel DataFrame. Modin transparently distributes the data and computation so that you can continue using the same pandas API while being able to work with more data faster. Modin lets you use all the CPU cores on your machine, and because it is lightweight, it often has less memory overhead than pandas. See this [page](#) to learn more about how Modin is different from pandas.

5.1.2 Why not just improve pandas?

pandas is a massive community and well established codebase. Many of the issues we have identified and resolved with pandas are fundamental to its current implementation. While we would be happy to donate parts of Modin that make sense in pandas, many of these components would require significant (or total) redesign of the pandas architecture. Modin's architecture goes beyond pandas, which is why the pandas API is just a thin layer at the user level. To learn more about Modin's architecture, see the [architecture](#) documentation.

5.1.3 How much faster can I go with Modin compared to pandas?

Modin is designed to scale with the amount of hardware available. Even in a traditionally serial task like `read_csv`, we see large gains by efficiently distributing the work across your entire machine. Because it is so light-weight, Modin provides speed-ups of up to 4x on a laptop with 4 physical cores. This speedup scales efficiently to larger machines with more cores. We have several published [papers](#) that include performance results and comparisons against pandas.

5.1.4 How much more data would I be able to process with Modin?

Often data scientists have to use different tools for operating on datasets of different sizes. This is not only because processing large dataframes is slow, but also pandas does not support working with dataframes that don't fit into the available memory. As a result, pandas workflows that work well for prototyping on a few MBs of data do not scale to tens or hundreds of GBs (depending on the size of your machine). Modin supports operating on data that does not fit in memory, so that you can comfortably work with hundreds of GBs without worrying about substantial slowdown or memory errors. For more information, see [out-of-memory support](#) for Modin.

5.1.5 How does Modin compare to Dask DataFrame and Koalas?

TLDR: Modin has better coverage of the pandas API, has a flexible backend, better ordering semantics, and supports both row and column-parallel operations. Check out this [page](#) detailing the differences!

5.1.6 How does Modin work under the hood?

Modin is logically separated into different layers that represent the hierarchy of a typical Database Management System. User queries which perform data transformation, data ingress or data egress pass through the Modin Query Compiler which translates queries from the top-level pandas API Layer that users interact with to the Modin Core Dataframe layer. The Modin Core DataFrame is our efficient DataFrame implementation that utilizes a partitioning schema which allows for distributing tasks and queries. From here, the Modin DataFrame works with engines like Ray or Dask to execute computation, and then return the results to the user.

For more details, take a look at our system [architecture](#).

5.2 FAQs: How to use Modin?

5.2.1 If I'm only using my laptop, can I still get the benefits of Modin?

Absolutely! Unlike other parallel DataFrame systems, Modin is an extremely light-weight, robust DataFrame. Because it is so light-weight, Modin provides speed-ups of up to 4x on a laptop with 4 physical cores and allows you to work on data that doesn't fit in your laptop's RAM.

5.2.2 How do I use Jupyter or Colab notebooks with Modin?

You can take a look at this Google Colab installation [guide](#) and this notebook [tutorial](#). Once Modin is installed, simply replace your pandas import with Modin import:

```
# import pandas as pd
import modin.pandas as pd
```

5.2.3 Which execution engine (Ray or Dask) should I use for Modin?

Modin lets you effortlessly speed up your pandas workflows with either [Ray](#)'s or [Dask](#)'s execution engine. You don't need to know anything about either engine in order to use it with Modin. If you only have one engine installed, Modin will automatically detect which engine you have installed and use that for scheduling computation. If you don't have a preference, we recommend starting with Modin's default Ray engine. If you want to use a specific compute engine, you can set the environment variable `MODIN_ENGINE` and Modin will do computation with that engine:

```
pip install "modin[ray]" # Install Modin dependencies and Ray to run on Ray
export MODIN_ENGINE=ray # Modin will use Ray

pip install "modin[dask]" # Install Modin dependencies and Dask to run on Dask
export MODIN_ENGINE=dask # Modin will use Dask
```

This can also be done with:

```
from modin.config import Engine

Engine.put("ray") # Modin will use Ray
Engine.put("dask") # Modin will use Dask
```

We also have an experimental OmniSciDB-based engine of Modin you can read about [here](#). We plan to support more execution engines in future. If you have a specific request, please post on the [#feature-requests](#) channel on our [Slack](#) community.

5.2.4 How do I connect Modin to a database via `read_sql`?

To read from a SQL database, you have two options:

- 1) Pass a connection string, e.g. `postgresql://reader:NWDMCE5xdipIjRrp@hh-pgsql-public.ebi.ac.uk:5432/pfmegrnargs`
- 2) Pass an open database connection, e.g. `for psycopg2, psycopg2.connect("dbname=pfmegrnargs user=reader password=NWDMCE5xdipIjRrp host=hh-pgsql-public.ebi.ac.uk")`

The first option works with both Modin and pandas. If you try the second option in Modin, Modin will default to pandas because open database connections cannot be pickled. Pickling is required to send connection details to remote workers. To handle the unique requirements of distributed database access, Modin has a distributed database connection called `ModinDatabaseConnection`:

```
import modin.pandas as pd
from modin.db_conn import ModinDatabaseConnection

con = ModinDatabaseConnection(
    'psycopg2',
    host='hh-pgsql-public.ebi.ac.uk',
    dbname='pfmegrnargs',
    user='reader',
    password='NWDMCE5xdipIjRrp')
df = pd.read_sql("SELECT * FROM rnc_database",
    con,
    index_col=None,
    coerce_float=True,
    params=None,
    parse_dates=None,
    chunksize=None)
```

The `ModinDatabaseConnection` will save any arguments you supply it and forward them to the workers to make their own connections.

5.2.5 How can I contribute to Modin?

Modin is currently under active development. Requests and contributions are welcome!

If you are interested in contributing please check out the [Contributing Guide](#) and then refer to the [Development Documentation](#), where you can find system architecture, internal implementation details, and other useful information. Also check out the [Github](#) to view open issues and make contributions.

TROUBLESHOOTING

We hope your experience with Modin is bug-free, but there are some quirks about Modin that may require troubleshooting. If you are still having issues, please post on the [#support](#) channel on our [Slack](#) community or open a [Github issue](#).

6.1 Frequently encountered issues

This is a list of the most frequently encountered issues when using Modin. Some of these are working as intended, while others are known bugs that are being actively worked on.

6.1.1 Warning during execution: defaulting to pandas

Please note, that while Modin covers a large portion of the pandas API, not all functionality is implemented. For methods that are not yet implemented, such as `asfreq`, you may see the following:

`UserWarning: `DataFrame.asfreq` defaulting to pandas implementation.`

To understand which functions will lead to this warning, we have compiled a list of [currently supported methods](#). When you see this warning, Modin defaults to pandas by converting the Modin dataframe to pandas to perform the operation. Once the operation is complete in pandas, it is converted back to a Modin dataframe. These operations will have a high overhead due to the communication involved and will take longer than pandas. When this is happening, a warning will be given to the user to inform them that this operation will take longer than usual. You can learn more about this [here](#).

If you would like to request a particular method be implemented, feel free to [`open an issue`](#). Before you open an issue please make sure that someone else has not already requested that functionality.

6.1.2 Hanging on `import modin.pandas as pd`

This can happen when Ray fails to start. It will keep retrying, but often it is faster to just restart the notebook or interpreter. Generally, this should not happen. Most commonly this is encountered when starting multiple notebooks or interpreters in quick succession.

Solution

Restart your interpreter or notebook kernel.

Avoiding this Error

Avoid starting many Modin notebooks or interpreters in quick succession. Wait 2-3 seconds before starting the next one.

6.1.3 Importing heterogeneous data by read_csv

Since Modin `read_csv` imports data in parallel, it can occur that data read by different partitions can have different type (this happens when columns contains heterogeneous data, i.e. column values are of different types), which are handled differently. Example of such behaviour is shown below.

```
import os
import pandas
import modin.pandas as pd
from modin.config import NPartitions

NPartitions.put(2)

test_filename = "test.csv"
# data with heterogeneous values in the first column
data = """one,2
3,4
5,6
7,8
9.0,10
"""

kwargs = {
    # names of the columns to set, if `names` parameter is set,
    # header inferring from the first data row/rows will be disabled
    "names": ["col1", "col2"],

    # explicit setting of data type of column/columns with heterogeneous
    # data will force partitions to read data with correct dtype
    # "dtype": {"col1": str},
}

try :
    with open(test_filename, "w") as f:
        f.write(data)

    pandas_df = pandas.read_csv(test_filename, **kwargs)
    pd_df = pd.read_csv(test_filename, **kwargs)

    print(pandas_df)
    print(pd_df)
finally:
    os.remove(test_filename)
```

Output:

```
pandas_df:
  col1  col2
0  one     2
1    3     4
2    5     6
3    7     8
4  9.0    10
```

(continues on next page)

(continued from previous page)

```

pd_df:
  col1  col2
0  one     2
1    3     4
2    5     6
3  7.0     8
4  9.0    10

```

In this case *DataFrame* read by pandas in the column `col1` contain only `str` data because of the first string value (“one”), that forced pandas to handle full column data as strings. Modin the first partition (the first three rows) read data similar to pandas, but the second partition (the last two rows) doesn’t contain any strings in the first column and its data is read as floats because of the last column value and as a result 7 value was read as 7.0, that differs from pandas output.

The above example showed the mechanism of occurrence of pandas and Modin `read_csv` outputs discrepancy during heterogeneous data import. Please note, that similar situations can occur during different data/parameters combinations.

Solution

In the case if heterogeneous data is detected, corresponding warning will be showed in the user’s console. Currently, the discrepancies of such type doesn’t properly handled by Modin, and to avoid this issue, it is needed to set `dtype` parameter of `read_csv` function manually to force correct data type definition during data import by partitions. Note, that to avoid excessive performance degradation, `dtype` value should be set fine-grained as it possible (specify `dtype` parameter only for columns with heterogeneous data).

Setting of `dtype` parameter works well for most of the cases, but, unfortunately, it is ineffective if data file contain column which should be interpreted as index (`index_col` parameter is used) since `dtype` parameter is responsible only for data fields. For example, if in the above example, `kwargs` will be set in the next way:

```

kwargs = {
    "names": ["col1", "col2"],
    "dtype": {"col1": str},
    "index_col": "col1",
}

```

Resulting Modin *DataFrame* will contain incorrect value as in the case if `dtype` is not set:

```

col1
one     2
3       4
5       6
7.0     8
9.0    10

```

In this case data should be imported without setting of `index_col` parameter and only then index column should be set as index (by using `DataFrame.set_index` function for example) as it is shown in the example below:

```

pd_df = pd.read_csv(filename, dtype=data_dtype, index_col=None)
pd_df = pd_df.set_index(index_col_name)
pd_df.index.name = None

```

6.1.4 Using Modin with python multiprocessing

We strongly recommend not to mix the use of Modin with Ray or Dask engine selected in conjunction with python multiprocessing because that can lead to undefined behavior. One of such examples is shown below:

```
import modin.pandas as pd

# Ray engine is used by default
df = pandas.DataFrame([1, 2, 3])

def f(arg):
    return df + arg

if __name__ == '__main__':
    from multiprocessing import Pool

    with Pool(5) as p:
        print(p.map(f, [1]))
```

Even if this example may work on your machine, we do not recommend similar scenarios. The python multiprocessing will cause conflicts with excessive resource use by launching duplicated Ray clusters on the same machine.

6.2 Common errors

6.2.1 Error when using OmniSci engine along with pyarrow.gandiva: LLVM ERROR: inconsistency in registered CommandLine options

This can happen when you use OmniSci engine along with pyarrow.gandiva:

```
import modin.config as cfg
cfg.Engine.put("Native") # 'omniscidbe'/'dbe' would be imported with dlopen flags
cfg.StorageFormat.put("OmniSci")
cfg.IsExperimental.put(True)
import modin.pandas as pd
import pyarrow.gandiva as gandiva # Error
# CommandLine Error: Option 'enable-vfe' registered more than once!
# LLVM ERROR: inconsistency in registered CommandLine options
# Aborted (core dumped)
```

Solution

Do not use OmniSci engine along with pyarrow.gandiva.

6.2.2 Error when using Dask engine: `RuntimeError: if __name__ == '__main__':`

The following *script.py* uses Modin with Dask as an execution engine and produces errors:

```
# script.py
import modin.pandas as pd
import modin.config as cfg

cfg.Engine.put("dask")

df = pd.DataFrame([0,1,2,3])
print(df)
```

A part of the produced errors by the script above would be the following:

```
File "/path/python3.9/multiprocessing/spawn.py", line 134, in _check_not_importing_main
    raise RuntimeError('')
RuntimeError:
    An attempt has been made to start a new process before the
    current process has finished its bootstrapping phase.

    This probably means that you are not using fork to start your
    child processes and you have forgotten to use the proper idiom
    in the main module:

        if __name__ == '__main__':
            freeze_support()
            ...

    The "freeze_support()" line can be omitted if the program
    is not going to be frozen to produce an executable.
```

This happens because Dask Client uses `fork` to start processes.

Solution

To avoid the problem Dask Client creation needs to be moved into `__main__` scope of the module.

The corrected *script.py* would look like:

```
# script.py
import modin.pandas as pd
import modin.config as cfg

cfg.Engine.put("dask")

if __name__ == "__main__":
    df = pd.DataFrame([0, 1, 2, 3]) # Dask Client creation is hidden in the first call of
    ↪Modin functionality.
    print(df)
```

or

```
# script.py
from distributed import Client
```

(continues on next page)

(continued from previous page)

```
import modin.pandas as pd
import modin.config as cfg

cfg.Engine.put("dask")

if __name__ == "__main__":
    client = Client() # Explicit Dask Client creation.
    df = pd.DataFrame([0, 1, 2, 3])
    print(df)
```

QUICK START GUIDE

To install the most recent stable release for Modin run the following:

```
pip install modin[all]
```

For further instructions on how to install Modin with conda or for specific platforms or engines, see our detailed [installation guide](#).

Modin acts as a drop-in replacement for pandas so you simply have to replace the import of pandas with the import of Modin as follows to speed up your pandas workflows:

```
# import pandas as pd
import modin.pandas as pd
```


EXAMPLE: INSTANT SCALABILITY WITH NO EXTRA EFFORT

When working on large datasets, pandas becomes painfully slow or *runs out of memory*. Modin automatically scales up your pandas workflows by parallelizing the dataframe operations, so that you can more effectively leverage the compute resources available.

For the purpose of demonstration, we will load in modin as `pd` and pandas as `pandas`.

```
import modin.pandas as pd
import pandas

#####
### For the purpose of timing comparisons ###
#####
import time
import ray
ray.init()
#####
```

In this toy example, we look at the NYC taxi dataset, which is around 120MB in size. You can download [this dataset](#) to run the example locally.

```
# This may take a few minutes to download
import urllib.request
s3_path = "https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2021-01.csv"
urllib.request.urlretrieve(s3_path, "taxi.csv")
```

8.1 Faster Data Loading with `read_csv`

```
start = time.time()

pandas_df = pandas.read_csv(s3_path, parse_dates=["tpep_pickup_datetime", "tpep_dropoff_
↳datetime"], quoting=3)

end = time.time()
pandas_duration = end - start
print("Time to read with pandas: {} seconds".format(round(pandas_duration, 3)))
```

By running the same command `read_csv` with Modin, we generally get around 4X speedup for loading in the data in parallel.

```
start = time.time()

modin_df = pd.read_csv(s3_path, parse_dates=["tpep_pickup_datetime", "tpep_dropoff_
↳datetime"], quoting=3)

end = time.time()
modin_duration = end - start
print("Time to read with Modin: {} seconds".format(round(modin_duration, 3)))

print("Modin is {}x faster than pandas at `read_csv`!".format(round(pandas_duration /
↳modin_duration, 2)))
```

8.2 Faster concat across multiple dataframes

Our previous `read_csv` example operated on a relatively small dataframe. In the following example, we duplicate the same taxi dataset 100 times and then concatenate them together, resulting in a dataset around 19GB in size.

```
start = time.time()

big_pandas_df = pandas.concat([pandas_df for _ in range(25)])

end = time.time()
pandas_duration = end - start
print("Time to concat with pandas: {} seconds".format(round(pandas_duration, 3)))
```

```
start = time.time()

big_modin_df = pd.concat([modin_df for _ in range(25)])

end = time.time()
modin_duration = end - start
print("Time to concat with Modin: {} seconds".format(round(modin_duration, 3)))

print("Modin is {}x faster than pandas at `concat`!".format(round(pandas_duration /
↳modin_duration, 2)))
```

Modin speeds up the concat operation by more than 60X, taking less than a second to create the large dataframe, while pandas took close to a minute.

8.3 Faster apply over a single column

The performance benefits of Modin becomes apparent when we operate on large gigabyte-scale datasets. For example, let's say that we want to round the number across a single column via the `apply` operation.

```
start = time.time()
rounded_trip_distance_pandas = big_pandas_df["trip_distance"].apply(round)

end = time.time()
```

(continues on next page)

(continued from previous page)

```
pandas_duration = end - start
print("Time to apply with pandas: {} seconds".format(round(pandas_duration, 3)))

start = time.time()

rounded_trip_distance_modin = big_modin_df["trip_distance"].apply(round)

end = time.time()
modin_duration = end - start
print("Time to apply with Modin: {} seconds".format(round(modin_duration, 3)))

print("Modin is {}x faster than pandas at `apply` on one column!".format(round(pandas_
↳ duration / modin_duration, 2)))
```

Modin is more than 30X faster at applying a single column of data, operating on 130+ million rows in a second.

In short, Modin provides orders of magnitude speed up over pandas for a variety of operations out of the box.

SUMMARY

Hopefully, this tutorial demonstrated how Modin delivers significant speedup on pandas operations without the need for any extra effort. Throughout example, we moved from working with 100MBs of data to 20GBs of data all without having to change anything or manually optimize our code to achieve the level of scalable performance that Modin provides.

Note that in this quickstart example, we've only shown `read_csv`, `concat`, `apply`, but these are not the only pandas operations that Modin optimizes for. In fact, Modin covers [more than 90% of the pandas API](#), yielding considerable speedups for many common operations.

USAGE GUIDE

This guide describes both basic and advanced Modin usage, including usage examples, details regarding Modin configuration settings, as well as tips and tricks on how you can further optimize the performance of your workload with Modin.

10.1 Modin Configuration Settings

To adjust Modin's default behavior, you can set the value of Modin configs by setting an environment variable or by using the `modin.config` API. To list all available configs in Modin, please run `python -m modin.config` to print all Modin configs with descriptions.

10.1.1 Public API

Potentially, the source of configs can be any, but for now only environment variables are implemented. Any environment variable originates from [EnvironmentVariable](#), which contains most of the config API implementation.

class `modin.config.envvars.EnvironmentVariable`

Base class for environment variables-based configuration.

classmethod `get()`

Get config value.

Returns Decoded and verified config value.

Return type Any

classmethod `get_help()` → str

Generate user-presentable help for the config.

Return type str

classmethod `get_value_source()`

Get value source of the config.

Return type int

classmethod `once(onvalue, callback)`

Execute *callback* if config value matches *onvalue* value.

Otherwise accumulate callbacks associated with the given *onvalue* in the `_once` container.

Parameters

- **onvalue** (Any) – Config value to set.

- **callback** (*callable*) – Callable that should be executed if config value matches *onvalue*.

classmethod put (*value*)

Set config value.

Parameters **value** (*Any*) – Config value to set.

classmethod subscribe (*callback*)

Add *callback* to the *_subs* list and then execute it.

Parameters **callback** (*callable*) – Callable to execute.

10.1.2 Modin Configs List

Config Name	Env. Variable Name	Default Value	Description	Options
AsvDataSizeConfig	MODIN_ASV_DATA_SIZE_CONFIG	SIZE_CONFIG	Allows to override default size of data (shapes).	
AsvImplementation	MODIN_ASV_USE_IMPL	Modin	Allows to select a library that we will use for testing performance.	('modin', 'pandas')
BenchmarkMode	MODIN_BENCHMARK_MODE	False	Whether or not to perform computations synchronously.	
CpuCount	MODIN_CPUS	2	How many CPU cores to use during initialization of the Modin engine.	
DoLogRpyc	MODIN_LOG_RPYC		Whether to gather RPyC logs (applicable for remote context).	
DoTraceRpyc	MODIN_TRACE_RPYC		Whether to trace RPyC calls (applicable for remote context).	
DoUseCalcite	MODIN_USE_CALCITE	True	Whether to use Calcite for OmniSci queries execution.	
Engine	MODIN_ENGINE	Ray	Distribution engine to run queries by.	('Ray', 'Dask', 'Python', 'Native')
GpuCount	MODIN_GPUS		How many GPU devices to utilize across the whole distribution.	
IsDebug	MODIN_DEBUG		Force Modin engine to be "Python" unless specified by \$MODIN_ENGINE.	
IsExperimental	MODIN_EXPERIMENTAL		Whether to Turn on experimental features.	
IsRayCluster	MODIN_RAY_CLUSTER		Whether Modin is running on pre-initialized Ray cluster.	
Memory	MODIN_MEMORY		How much memory (in bytes) give to an execution engine.	

10.1.3 Usage Guide

See example of interaction with Modin configs below, as it can be seen config value can be set either by setting the environment variable or by using config API.

```
import os

# Setting `MODIN_STORAGE_FORMAT` environment variable.
# Also can be set outside the script.
os.environ["MODIN_STORAGE_FORMAT"] = "OmniSci"

import modin.config
import modin.pandas as pd

# Checking initially set `StorageFormat` config,
# which corresponds to `MODIN_STORAGE_FORMAT` environment
# variable
print(modin.config.StorageFormat.get()) # prints 'OmniSci'

# Checking default value of `NPartitions`
print(modin.config.NPartitions.get()) # prints '8'

# Changing value of `NPartitions`
modin.config.NPartitions.put(16)
print(modin.config.NPartitions.get()) # prints '16'
```

10.2 Modin Usage Examples

This section shows Modin usage examples in different scenarios like Modin on a local/remote cluster, Modin in the cloud, the use of Modin spreadsheet.

10.2.1 Tutorials

The following tutorials cover the basic usage of Modin. [Here](#) is a one hour video tutorial that walks through these basic exercises.

- Exercise 1: Introduction to Modin [[Source PandasOnRay](#), [Source PandasOnDask](#), [Source OmniSciOnNative](#)]
- Exercise 2: Speed Improvements with Modin [[Source PandasOnRay](#), [Source PandasOnDask](#), [Source OmniSciOnNative](#)]
- Exercise 3: Defaulting to pandas with Modin [[Source PandasOnRay](#), [Source PandasOnDask](#), [Source OmniSciOnNative](#)]

The following tutorials covers more advanced features in Modin:

- Exercise 4: Experimental Features in Modin (Spreadsheet, Progress Bar) [[Source PandasOnRay](#), [Source PandasOnDask](#)]
- Exercise 5: Setting up Modin in a Cluster Environment [[Source PandasOnRay](#)]
- Exercise 6: Running Modin in a Cluster Environment [[Source PandasOnRay](#)]

How to get required dependencies for the tutorial notebooks and to run them please refer to the respective [README.md](#) file.

10.2.2 Data Science Benchmarks

- Using Modin with the NYC Taxi Dataset [[Source](#)]
- Using Modin with the Census Dataset (coming soon...)
- Using Modin with the Plasticc Dataset (coming soon...)

10.2.3 Modin in the Cloud

- Using Experimental Modin in the cloud with the NYC Taxi Dataset on an AWS cluster [[Source](#)]

10.2.4 Modin Spreadsheets

- Using Modin along with the Spreadsheets API [[Source](#)]

10.2.5 Modin with scikit-learn

- Modin for Machine Learning with scikit-learn [[Source](#)]

10.3 Advanced Usage

10.3.1 Pandas partitioning API

This page contains a description of the API to extract partitions from and build Modin Dataframes.

`unwrap_partitions`

`modin.distributed.dataframe.pandas.unwrap_partitions(api_layer_object, axis=None, get_ip=False)`

Unwrap partitions of the `api_layer_object`.

Parameters

- **`api_layer_object`** (`DataFrame` or `Series`) – The API layer object.
- **`axis`** (`{None, 0, 1}`, `default: None`) – The axis to unwrap partitions for (0 - row partitions, 1 - column partitions). If `axis` is `None`, the partitions are unwrapped as they are currently stored.
- **`get_ip`** (`bool`, `default: False`) – Whether to get node ip address to each partition or not.

Returns A list of `Ray.ObjectRef/Dask.Future` to partitions of the `api_layer_object` if Ray/Dask is used as an engine.

Return type `list`

Notes

If `get_ip=True`, a list of tuples of `Ray.ObjectRef/Dask.Future` to node ip addresses and partitions of the `api_layer_object`, respectively, is returned if Ray/Dask is used as an engine (i.e. `[(Ray.ObjectRef/Dask.Future, Ray.ObjectRef/Dask.Future), ...]`).

from_partitions

`modin.distributed.dataframe.pandas.from_partitions(partitions, axis, index=None, columns=None, row_lengths=None, column_widths=None)`

Create DataFrame from remote partitions.

Parameters

- **partitions** (*list*) – A list of `Ray.ObjectRef/Dask.Future` to partitions depending on the engine used. Or a list of tuples of `Ray.ObjectRef/Dask.Future` to node ip addresses and partitions depending on the engine used (i.e. `[(Ray.ObjectRef/Dask.Future, Ray.ObjectRef/Dask.Future), ...]`).
- **axis** (*{None, 0 or 1}*) – The axis parameter is used to identify what are the partitions passed. You have to set:
 - `axis=0` if you want to create DataFrame from row partitions
 - `axis=1` if you want to create DataFrame from column partitions
 - `axis=None` if you want to create DataFrame from 2D list of partitions
- **index** (*sequence, optional*) – The index for the DataFrame. Is computed if not provided.
- **columns** (*sequence, optional*) – The columns for the DataFrame. Is computed if not provided.
- **row_lengths** (*list, optional*) – The length of each partition in the rows. The “height” of each of the block partitions. Is computed if not provided.
- **column_widths** (*list, optional*) – The width of each partition in the columns. The “width” of each of the block partitions. Is computed if not provided.

Returns DataFrame instance created from remote partitions.

Return type `modin.pandas.DataFrame`

Notes

Pass `index`, `columns`, `row_lengths` and `column_widths` to avoid triggering extra computations of the metadata when creating a DataFrame.

Example

```
import modin.pandas as pd
from modin.distributed.dataframe.pandas import unwrap_partitions, from_partitions
import numpy as np
data = np.random.randint(0, 100, size=(2 ** 10, 2 ** 8))
df = pd.DataFrame(data)
partitions = unwrap_partitions(df, axis=0, get_ip=True)
print(partitions)
new_df = from_partitions(partitions, axis=0)
print(new_df)
```

10.3.2 Modin Spreadsheets API

Getting started

Install Modin-spreadsheet using pip:

```
pip install modin[spreadsheet]
```

The following code snippet creates a spreadsheet using the FiveThirtyEight dataset on labor force information by college majors (licensed under CC BY 4.0):

```
import modin.pandas as pd
import modin.spreadsheet as mss
df = pd.read_csv('https://raw.githubusercontent.com/fivethirtyeight/data/master/college-
➔majors/all-ages.csv')
spreadsheet = mss.from_dataframe(df)
spreadsheet
```

Basic Manipulations through User Interface

The Spreadsheet API allows users to manipulate the DataFrame with simple graphical controls for sorting, filtering, and editing.

Here are the instructions for each operation:

- **Sort:** Click on the column header of the column to sort on.
- **Filter:** Click on the filter button on the column header and apply the desired filter to the column. The filter dropdown changes depending on the type of the column. Multiple filters are automatically combined.
- **Edit Cell:** Double click on a cell and enter the new value.
- **Add Rows:** Click on the “Add Row” button in the toolbar to duplicate the last row in the DataFrame. The duplicated values provide a convenient default and can be edited as necessary.
- **Remove Rows:** Select row(s) and click the “Remove Row” button. Select a single row by clicking on it. Multiple rows can be selected with Cmd+Click (Windows: Ctrl+Click) on the desired rows or with Shift+Click to specify a range of rows.

Some of these operations can also be done through the spreadsheet’s programmatic interface. Sorts and filters can be reset using the toolbar buttons. Edits and adding/removing rows can only be undone manually.

```
import modin.pandas as pd
import modin.spreadsheet as mss
df = pd.read_csv('https://raw.githubusercontent.com/fivethirtyeight/data/master/college-majors/all-ages.csv')
spreadsheet = mss.from_dataframe(df)
spreadsheet
```

Add Row	Remove Row	Clear History	Filter History	Reset Filters	Reset Sort	
	Major_code	Major	Major_category	Total	Employed	Employed_fu
0	1100	GENERAL AGRICULT...	Agriculture & Natural ...	128148	90245	74078
1	1101	AGRICULTURE PROD...	Agriculture & Natural ...	95326	76865	64240
2	1102	AGRICULTURAL ECO...	Agriculture & Natural ...	33955	26321	22810
3	1103	ANIMAL SCIENCES	Agriculture & Natural ...	103549	81177	64937
4	1104	FOOD SCIENCE	Agriculture & Natural ...	24280	17281	12722
5	1105	PLANT SCIENCE AND...	Agriculture & Natural ...	79409	63043	51077
6	1106	SOIL SCIENCE	Agriculture & Natural ...	6586	4926	4042
7	1199	MISCELLANEOUS AG...	Agriculture & Natural ...	8549	6392	5074
8	1301	ENVIRONMENTAL SC...	Biology & Life Science	106106	87602	65238
9	1302	FORESTRY	Agriculture & Natural ...	69447	48228	39613
10	1303	NATURAL RESOURC...	Agriculture & Natural ...	83188	65937	50595
11	1401	ARCHITECTURE	Engineering	294692	216770	163020
12	1501	AREA ETHNIC AND C...	Humanities & Liberal ...	103740	75798	50530
13	1901	COMMUNICATIONS	Communications & Jo...	987676	790696	595739
14	1902	JOURNALISM	Communications & Jo...	418104	314438	235407
15	1903	MASS MEDIA	Communications & Jo...	244010	170474	105400

```
# ---- spreadsheet transformation history ----
unfiltered_df = df.copy()
```

Virtual Rendering

The spreadsheet will only render data based on the user's viewport. This allows for quick rendering even on very large DataFrames because only a handful of rows are loaded at any given time. As a result, scrolling and viewing your data is smooth and responsive.

Transformation History and Exporting Code

All operations on the spreadsheet are recorded and are easily exported as code for sharing or reproducibility. This history is automatically displayed in the history cell, which is generated below the spreadsheet whenever the spreadsheet widget is displayed. The history cell is displayed on default, but this can be turned off. Modin Spreadsheet API provides a few methods for interacting with the history:

- `SpreadsheetWidget.get_history()`: Retrieves the transformation history in the form of reproducible code.
- `SpreadsheetWidget.filter_relevant_history(persist=True)`: Returns the transformation history that contains only code relevant to the final state of the spreadsheet. The `persist` parameter determines whether the internal state and the displayed history is also filtered.
- `SpreadsheetWidget.reset_history()`: Clears the history of transformation.

Customizable Interface

The spreadsheet widget provides a number of options that allows the user to change the appearance and the interactivity of the spreadsheet. Options include:

- Row height/Column width
- Preventing edits, sorts, or filters on the whole spreadsheet or on a per-column basis
- Hiding the toolbar and history cell
- Float precision
- Highlighting of cells and rows
- Viewport size

Converting Spreadsheets To and From Dataframes

```
modin.experimental.spreadsheet.general.from_dataframe(dataframe, show_toolbar=None,
                                                         show_history=None, precision=None,
                                                         grid_options=None, column_options=None,
                                                         column_definitions=None,
                                                         row_edit_callback=None)
```

Renders a DataFrame or Series as an interactive spreadsheet, represented by an instance of the SpreadsheetWidget class. The SpreadsheetWidget instance is constructed using the options passed in to this function. The dataframe argument to this function is used as the df kwarg in call to the SpreadsheetWidget constructor, and the rest of the parameters are passed through as is.

If the dataframe argument is a Series, it will be converted to a DataFrame before being passed in to the SpreadsheetWidget constructor as the df kwarg.

Return type SpreadsheetWidget

Parameters

- **dataframe** (DataFrame) – The DataFrame that will be displayed by this instance of SpreadsheetWidget.
- **grid_options** (dict) – Options to use when creating the SlickGrid control (i.e. the interactive grid). See the Notes section below for more information on the available options, as well as the default options that this widget uses.
- **precision** (integer) – The number of digits of precision to display for floating-point values. If unset, we use the value of `pandas.get_option('display.precision')`.
- **show_toolbar** (bool) – Whether to show a toolbar with options for adding/removing rows. Adding/removing rows is an experimental feature which only works with DataFrames that have an integer index.
- **show_history** (bool) – Whether to show the cell containing the spreadsheet transformation history.
- **column_options** (dict) – Column options that are to be applied to every column. See the Notes section below for more information on the available options, as well as the default options that this widget uses.
- **column_definitions** (dict) – Column options that are to be applied to individual columns. The keys of the dict should be the column names, and each value should be the column options for a particular column, represented as a dict. The available options for each col-

umn are the same options that are available to be set for all columns via the `column_options` parameter. See the Notes section below for more information on those options.

- **row_edit_callback** (*callable*) – A callable that is called to determine whether a particular row should be editable or not. Its signature should be `callable(row)`, where `row` is a dictionary which contains a particular row’s values, keyed by column name. The callback should return `True` if the provided row should be editable, and `False` otherwise.

Notes

The following dictionary is used for `grid_options` if none are provided explicitly:

```
{
  # SlickGrid options
  'fullWidthRows': True,
  'syncColumnCellResize': True,
  'forceFitColumns': False,
  'defaultColumnWidth': 150,
  'rowHeight': 28,
  'enableColumnReorder': False,
  'enableTextSelectionOnCells': True,
  'editable': True,
  'autoEdit': False,
  'explicitInitialization': True,

  # Modin-spreadsheet options
  'maxVisibleRows': 15,
  'minVisibleRows': 8,
  'sortable': True,
  'filterable': True,
  'highlightSelectedCell': False,
  'highlightSelectedRow': True
}
```

The first group of options are SlickGrid “grid options” which are described in the [SlickGrid documentation](#).

The second group of option are options that were added specifically for modin-spreadsheet and therefore are not documented in the SlickGrid documentation. The following bullet points describe these options.

- **maxVisibleRows** The maximum number of rows that modin-spreadsheet will show.
- **minVisibleRows** The minimum number of rows that modin-spreadsheet will show
- **sortable** Whether the modin-spreadsheet instance will allow the user to sort columns by clicking the column headers. When this is set to `False`, nothing will happen when users click the column headers.
- **filterable** Whether the modin-spreadsheet instance will allow the user to filter the grid. When this is set to `False` the filter icons won’t be shown for any columns.
- **highlightSelectedCell** If you set this to `True`, the selected cell will be given a light blue border.
- **highlightSelectedRow** If you set this to `False`, the light blue background that’s shown by default for selected rows will be hidden.

The following dictionary is used for `column_options` if none are provided explicitly:

```
{
    # SlickGrid column options
    'defaultSortAsc': True,
    'maxWidth': None,
    'minWidth': 30,
    'resizable': True,
    'sortable': True,
    'toolTip': "",
    'width': None

    # Modin-spreadsheet column options
    'editable': True,
}
```

The first group of options are SlickGrid “column options” which are described in the [SlickGrid documentation](#).

The `editable` option was added specifically for `modin-spreadsheet` and therefore is not documented in the SlickGrid documentation. This option specifies whether a column should be editable or not.

See also:

set_defaults Permanently set global defaults for the parameters of `show_grid`, with the exception of the `dataframe` and `column_definitions` parameters, since those depend on the particular set of data being shown by an instance, and therefore aren’t parameters we would want to set for all `SpreadsheetWidget` instances.

set_grid_option Permanently set global defaults for individual grid options. Does so by changing the defaults that the `show_grid` method uses for the `grid_options` parameter.

SpreadsheetWidget The widget class that is instantiated and returned by this method.

`modin.experimental.spreadsheet.general.to_dataframe(spreadsheet)`

Get a copy of the `DataFrame` that reflects the current state of the `spreadsheet` `SpreadsheetWidget` instance UI. This includes any sorting or filtering changes, as well as edits that have been made by double clicking cells.

Return type *DataFrame*

Parameters `spreadsheet` (*SpreadsheetWidget*) – The `SpreadsheetWidget` instance that `DataFrame` that will be displayed by this instance of `SpreadsheetWidget`.

Further API Documentation

class `modin_spreadsheet.grid.SpreadsheetWidget(**kwargs)`

The widget class which is instantiated by the `show_grid` method. This class can be constructed directly but that’s not recommended because then default options have to be specified explicitly (since default options are normally provided by the `show_grid` method).

The constructor for this class takes all the same parameters as `show_grid`, with one exception, which is that the required `data_frame` parameter is replaced by an optional keyword argument called `df`.

See also:

show_grid The method that should be used to construct `SpreadsheetWidget` instances, because it provides reasonable defaults for all of the `modin-spreadsheet` options.

df

Get/set the DataFrame that's being displayed by the current instance. This DataFrame will NOT reflect any sorting/filtering/editing changes that are made via the UI. To get a copy of the DataFrame that does reflect sorting/filtering/editing changes, use the `get_changed_df()` method.

Type *DataFrame*

grid_options

Get/set the grid options being used by the current instance.

Type dict

precision

Get/set the precision options being used by the current instance.

Type integer

show_toolbar

Get/set the show_toolbar option being used by the current instance.

Type bool

show_history

Get/set the show_history option being used by the current instance.

Type bool

column_options

Get/set the column options being used by the current instance.

Type bool

column_definitions

Get/set the column definitions (column-specific options) being used by the current instance.

Type bool

add_row(row=None)

Append a row at the end of the DataFrame. Values for the new row can be provided via the `row` argument, which is optional for DataFrames that have an integer index, and required otherwise. If the `row` argument is not provided, the last row will be duplicated and the index of the new row will be the index of the last row plus one.

Parameters `row` (*list* (default: *None*)) – A list of 2-tuples of (column name, column value) that specifies the values for the new row.

See also:

SpreadsheetWidget.remove_rows The method for removing a row (or rows).

change_grid_option(option_name, option_value)

Change a SlickGrid grid option without rebuilding the entire grid widget. Not all options are supported at this point so this method should be considered experimental.

Parameters

- **option_name** (*str*) – The name of the grid option to be changed.
- **option_value** (*str*) – The new value for the grid option.

change_selection(rows=[])

Select a row (or rows) in the UI. The indices of the rows to select are provided via the optional `rows` argument.

Parameters `rows` (*list* (default: [])) – A list of indices of the rows to select. For a multi-indexed DataFrame, each index in the list should be a tuple, with each value in each tuple corresponding to a level of the MultiIndex. The default value of [] results in the no rows being selected (i.e. it clears the selection).

edit_cell(index, column, value)

Edit a cell of the grid, given the index and column of the cell to edit, as well as the new value of the cell. Results in a `cell_edited` event being fired.

Parameters

- **index** (*object*) – The index of the row containing the cell that is to be edited.
- **column** (*str*) – The name of the column containing the cell that is to be edited.
- **value** (*object*) – The new value for the cell.

get_changed_df()

Get a copy of the DataFrame that was used to create the current instance of SpreadsheetWidget which reflects the current state of the UI. This includes any sorting or filtering changes, as well as edits that have been made by double clicking cells.

Return type *DataFrame*

get_selected_df()

Get a DataFrame which reflects the current state of the UI and only includes the currently selected row(s). Internally it calls `get_changed_df()` and then filters down to the selected rows using `iloc`.

Return type *DataFrame*

get_selected_rows()

Get the currently selected rows.

Return type List of integers

off(names, handler)

Remove a modin-spreadsheet event handler that was registered with the current instance's `on` method.

Parameters

- **names** (*list, str, All* (default: All)) – The names of the events for which the specified handler should be uninstalled. If names is All, the specified handler is uninstalled from the list of notifiers corresponding to all events.
- **handler** (*callable*) – A callable that was previously registered with the current instance's `on` method.

See also:

SpreadsheetWidget.on The method for hooking up instance-level handlers that this `off` method can remove.

on(names, handler)

Setup a handler to be called when a user interacts with the current instance.

Parameters

- **names** (*list*, *str*, *All*) – If names is All, the handler will apply to all events. If a list of str, handler will apply to all events named in the list. If a str, the handler will apply just the event with that name.
- **handler** (*callable*) – A callable that is called when the event occurs. Its signature should be handler(event, spreadsheet_widget), where event is a dictionary and spreadsheet_widget is the SpreadsheetWidget instance that fired the event. The event dictionary at least holds a name key which specifies the name of the event that occurred.

Notes

Here’s the list of events that you can listen to on SpreadsheetWidget instances via the on method:

```
[
    'cell_edited',
    'selection_changed',
    'viewport_changed',
    'row_added',
    'row_removed',
    'filter_dropdown_shown',
    'filter_changed',
    'sort_changed',
    'text_filter_viewport_changed',
    'json_updated'
]
```

The following bullet points describe the events listed above in more detail. Each event bullet point is followed by sub-bullets which describe the keys that will be included in the event dictionary for each event.

- **cell_edited** The user changed the value of a cell in the grid.
 - **index** The index of the row that contains the edited cell.
 - **column** The name of the column that contains the edited cell.
 - **old** The previous value of the cell.
 - **new** The new value of the cell.
- **filter_changed** The user changed the filter setting for a column.
 - **column** The name of the column for which the filter setting was changed.
- **filter_dropdown_shown** The user showed the filter control for a column by clicking the filter icon in the column’s header.
 - **column** The name of the column for which the filter control was shown.
- **json_updated** A user action causes SpreadsheetWidget to send rows of data (in json format) down to the browser. This happens as a side effect of certain actions such as scrolling, sorting, and filtering.
 - **triggered_by** The name of the event that resulted in rows of data being sent down to the browser. Possible values are change_viewport, change_filter, change_sort, add_row, remove_row, and edit_cell.
 - **range** A tuple specifying the range of rows that have been sent down to the browser.
- **row_added** The user added a new row using the “Add Row” button in the grid toolbar.
 - **index** The index of the newly added row.

- **source** The source of this event. Possible values are **api** (an api method call) and **gui** (the grid interface).
- **row_removed** The user added removed one or more rows using the “Remove Row” button in the grid toolbar.
 - **indices** The indices of the removed rows, specified as an array of integers.
 - **source** The source of this event. Possible values are **api** (an api method call) and **gui** (the grid interface).
- **selection_changed** The user changed which rows were highlighted in the grid.
 - **old** An array specifying the indices of the previously selected rows.
 - **new** The indices of the rows that are now selected, again specified as an array.
 - **source** The source of this event. Possible values are **api** (an api method call) and **gui** (the grid interface).
- **sort_changed** The user changed the sort setting for the grid.
 - **old** The previous sort setting for the grid, specified as a dict with the following keys:
 - * **column** The name of the column that the grid was sorted by
 - * **ascending** Boolean indicating ascending/descending order
 - **new** The new sort setting for the grid, specified as a dict with the following keys:
 - * **column** The name of the column that the grid is currently sorted by
 - * **ascending** Boolean indicating ascending/descending order
- **text_filter_viewport_changed** The user scrolled the new rows into view in the filter dropdown for a text field.
 - **column** The name of the column whose filter dropdown is visible
 - **old** A tuple specifying the previous range of visible rows in the filter dropdown.
 - **new** A tuple specifying the range of rows that are now visible in the filter dropdown.
- **viewport_changed** The user scrolled the new rows into view in the grid.
 - **old** A tuple specifying the previous range of visible rows.
 - **new** A tuple specifying the range of rows that are now visible.

The event dictionary for every type of event will contain a **name** key specifying the name of the event that occurred. That key is excluded from the lists of keys above to avoid redundancy.

See also:

on Same as the instance-level **on** method except it listens for events on all instances rather than on an individual **SpreadsheetWidget** instance.

SpreadsheetWidget.off Unhook a handler that was hooked up using the instance-level **on** method.

remove_row(rows=*None*)

Alias for **remove_rows**, which is provided for convenience because this was the previous name of that method.

remove_rows(rows=None)

Remove a row (or rows) from the DataFrame. The indices of the rows to remove can be provided via the optional rows argument. If the rows argument is not provided, the row (or rows) that are currently selected in the UI will be removed.

Parameters rows (*list (default: None)*) – A list of indices of the rows to remove from the DataFrame. For a multi-indexed DataFrame, each index in the list should be a tuple, with each value in each tuple corresponding to a level of the MultiIndex.

See also:

SpreadsheetWidget.add_row The method for adding a row.

SpreadsheetWidget.remove_row Alias for this method.

toggle_editable()

Change whether the grid is editable or not, without rebuilding the entire grid widget.

10.3.3 Progress Bar

The progress bar allows users to see the estimated progress and completion time of each line they run, in environments such as a shell or Jupyter notebook.

Quickstart

The progress bar uses the *tqdm* library to visualize displays:

```
pip install tqdm
```

Import the progress bar into your notebook by running the following:

```
import modin.pandas as pd
from tqdm import tqdm
from modin.config import ProgressBar
ProgressBar.enable()
```

10.3.4 SQL on Modin Dataframes

MindsDB has teamed up with Modin to bring in-memory SQL to distributed Modin Dataframes. Now you can run SQL alongside the pandas API without copying or going through your disk. What this means is that you can now have a SQL solution that you can seamlessly scale horizontally and vertically, by leveraging the incredible power of Ray.

A Short Example Using the Google Play Store

```
import modin.pandas as pd
import modin.experimental.sql as sql

# read google play app store list from csv
gstore_apps_df = pd.read_csv("https://tinyurl.com/googleplaystorecsv")
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up

Imagine that you want to quickly select from ‘gstore_apps_df’ the columns App, Category, and Rating, where Price is ‘0’.

```
# You can then define the query that you want to perform
sql_str = "SELECT App, Category, Rating FROM gstore_apps WHERE Price = '0'"

# And simply apply that query to a dataframe
result_df = sql.query(sql_str, gstore_apps=gstore_apps_df)

# Or, in this case, where the query only requires one table,
# you can also ignore the FROM part in the query string:
query_str = "SELECT App, Category, Rating WHERE Price = '0' "

# sql.query can take query strings without FROM statement
# you can specify from as the function argument
result_df = sql.query(query_str, from=gstore_apps_df)
```

Writing Complex Queries

Let’s explore a more complicated example.

```
gstore_reviews_df = pd.read_csv("https://tinyurl.com/gstorereviewscsv")
```

	App	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
0	10 Best Foods for You	I like eat delicious food. That's I'm cooking ...	Positive	1.00	0.533333
1	10 Best Foods for You	This help eating healthy exercise regular basis	Positive	0.25	0.288462
2	10 Best Foods for You	NaN	NaN	NaN	NaN
3	10 Best Foods for You	Works great especially going grocery store	Positive	0.40	0.875000

Say we want to retrieve the top 10 app categories ranked by best average ‘sentiment_polarity’ where the average ‘sentiment_subjectivity’ is less than 0.5.

Since 'Category' is on the **gstore_apps_df** and sentiment_polarity is on **gstore_reviews_df**, we need to join the two tables, and operate averages on that join.

```
# Single query with join and group by
sql_str = """
SELECT
category,
avg(sentiment_polarity) as avg_sentiment_polarity,
avg(sentiment_subjectivity) as avg_sentiment_subjectivity
FROM (
SELECT
    category,
    CAST(sentiment as float) as sentiment,
    CAST(sentiment_polarity as float) as sentiment_polarity
FROM gstore_apps_df
    INNER JOIN gstore_reviews_df
    ON gstore_apps_df.app = gstore_reviews_df.app
) sub
GROUP BY category
HAVING avg_sentiment_subjectivity < 0.5
ORDER BY avg_sentiment_polarity DESC
LIMIT 10
"""

# Run query using apps and reviews dataframes,
# NOTE: that you simply pass the names of the tables in the query as arguments

result_df = sql.query( sql_str,
                        gstore_apps_df = gstore_apps_df,
                        gstore_reviews_df = gstore_reviews_df)
```

Or, you can bring the best of doing this in python and run the query in multiple parts (it's up to you).

```
# join the items and reviews

result_df = sql.query( """
SELECT
    category,
    sentiment,
    sentiment_polarity
FROM gstore_apps_df INNER JOIN gstore_reviews_df
ON gstore_apps_df.app = gstore_reviews_df.app """ ,
gstore_apps_df = gstore_apps_df,
gstore_reviews_df = gstore_reviews_df )

# group by category and calculate averages

result_df = sql.query( """
SELECT
    category,
    avg(sentiment_polarity) as avg_sentiment_polarity,
    avg(sentiment_subjectivity) as avg_sentiment_subjectivity
GROUP BY category
HAVING CAST(avg_sentiment_subjectivity as float) < 0.5
```

(continues on next page)

(continued from previous page)

```
ORDER BY avg_sentiment_polarity DESC
LIMIT 10""",
from = result_df)
```

If you have a cluster or even a computer with more than one CPU core, you can write SQL and Modin will run those queries in a distributed and optimized way.

Further Examples and Full Documentation

In the meantime, you can check out our [Example Notebook](#) that contains more examples and ideas, as well as this [blog](#) explaining Modin SQL usage.

10.3.5 Distributed XGBoost on Modin

Modin provides an implementation of [distributed XGBoost](#) machine learning algorithm on Modin DataFrames. Please note that this feature is experimental and behavior or interfaces could be changed.

Install XGBoost on Modin

Modin comes with all the dependencies except `xgboost` package by default. Currently, distributed XGBoost on Modin is only supported on the Ray execution engine, therefore, see the installation page for more information on installing Modin with the Ray engine. To install `xgboost` package you can use `pip`:

```
pip install xgboost
```

XGBoost Train and Predict

Distributed XGBoost functionality is placed in `modin.experimental.xgboost` module. `modin.experimental.xgboost` provides a drop-in replacement API for `train` and `Booster.predict` `xgboost` functions.

Module holds public interfaces for Modin XGBoost.

```
modin.experimental.xgboost.train(params: Dict, dtrain: modin.experimental.xgboost.xgboost.DMatrix,
                                  *args, evals=(), num_actors: Optional[int] = None, evals_result:
                                  Optional[Dict] = None, **kwargs)
```

Run distributed training of XGBoost model.

During work it evenly distributes *dtrain* between workers according to IP addresses partitions (in case of not even distribution of *dtrain* over nodes, some partitions will be re-distributed between nodes), runs `xgb.train` on each worker for subset of *dtrain* and reduces training results of each worker using Rabbit Context.

Parameters

- **params** (*dict*) – Booster params.
- **dtrain** (*modin.experimental.xgboost.DMatrix*) – Data to be trained against.
- ***args** (*iterable*) – Other parameters for *xgboost.train*.
- **evals** (*list of pairs (modin.experimental.xgboost.DMatrix, str), default: empty*) – List of validation sets for which metrics will be evaluated during training. Validation metrics will help us track the performance of the model.

- **num_actors** (*int*, *optional*) – Number of actors for training. If unspecified, this value will be computed automatically.
- **evals_result** (*dict*, *optional*) – Dict to store evaluation results in.
- ****kwargs** (*dict*) – Other parameters are the same as *xgboost.train*.

Returns A trained booster.

Return type `modin.experimental.xgboost.Booster`

class `modin.experimental.xgboost.Booster`(*params=None*, *cache=()*, *model_file=None*)

A Modin Booster of XGBoost.

Booster is the model of XGBoost, that contains low level routines for training, prediction and evaluation.

Parameters

- **params** (*dict*, *optional*) – Parameters for boosters.
- **cache** (*list*, *default: empty*) – List of cache items.
- **model_file** (*string/os.PathLike/xgb.Booster/bytearray*, *optional*) – Path to the model file if it's string or PathLike or xgb.Booster.

predict(*data: modin.experimental.xgboost.xgboost.DMatrix*, ***kwargs*)

Run distributed prediction with a trained booster.

During execution it runs `xgb.predict` on each worker for subset of *data* and creates Modin DataFrame with prediction results.

Parameters

- **data** (*modin.experimental.xgboost.DMatrix*) – Input data used for prediction.
- ****kwargs** (*dict*) – Other parameters are the same as for `xgboost.Booster.predict`.

Returns Modin DataFrame with prediction results.

Return type `modin.pandas.DataFrame`

ModinDMatrix

Data is passed to `modin.experimental.xgboost` functions via a Modin `DMatrix` object.

Module holds public interfaces for Modin XGBoost.

class `modin.experimental.xgboost.DMatrix`(*data*, *label=None*, *missing=None*, *silent=False*,
feature_names=None, *feature_types=None*,
feature_weights=None, *enable_categorical=None*)

DMatrix holds references to partitions of Modin DataFrame.

On init stage unwrapping partitions of Modin DataFrame is started.

Parameters

- **data** (*modin.pandas.DataFrame*) – Data source of DMatrix.
- **label** (*modin.pandas.DataFrame or modin.pandas.Series*, *optional*) – Labels used for training.
- **missing** (*float*, *optional*) – Value in the input data which needs to be present as a missing value. If None, defaults to `np.nan`.
- **silent** (*boolean*, *optional*) – Whether to print messages during construction or not.

- **feature_names** (*list, optional*) – Set names for features.
- **feature_types** (*list, optional*) – Set types for features.
- **feature_weights** (*array_like, optional*) – Set feature weights for column sampling.
- **enable_categorical** (*boolean, optional*) – Experimental support of specializing for categorical features.

Notes

Currently DMatrix doesn't support *weight*, *base_margin*, *nthread*, *group*, *qid*, *label_lower_bound*, *label_upper_bound* parameters.

property **feature_names**

Get column labels.

Return type Column labels.

property **feature_types**

Get column types.

Return type Column types.

get_dmatrix_params()

Get dict of DMatrix parameters excluding *self.data/self.label*.

Return type dict

get_float_info(name)

Get float property from the DMatrix.

Parameters **name** (*str*) – The field name of the information.

Return type A NumPy array of float information of the data.

num_col()

Get number of columns.

Return type int

num_row()

Get number of rows.

Return type int

set_info(*, *label=None, feature_names=None, feature_types=None, feature_weights=None*) → None

Set meta info for DMatrix.

Parameters

- **label** (*modin.pandas.DataFrame or modin.pandas.Series, optional*) – Labels used for training.
- **feature_names** (*list, optional*) – Set names for features.
- **feature_types** (*list, optional*) – Set types for features.
- **feature_weights** (*array_like, optional*) – Set feature weights for column sampling.

Currently, the Modin DMatrix supports `modin.pandas.DataFrame` only as an input.

A Single Node / Cluster setup

The XGBoost part of Modin uses a Ray resources by similar way as all Modin functions.

To start the Ray runtime on a single node:

```
import ray
ray.init()
```

If you already had the Ray cluster you can connect to it by next way:

```
import ray
ray.init(address='auto')
```

A detailed information about initializing the Ray runtime you can find in [starting ray](#) page.

Usage example

In example below we train XGBoost model using [the Iris Dataset](#) and get prediction on the same data. All processing will be in a *single node* mode.

```
from sklearn import datasets

import ray
ray.init() # Start the Ray runtime for single-node

import modin.pandas as pd
import modin.experimental.xgboost as xgb

# Load iris dataset from sklearn
iris = datasets.load_iris()

# Create Modin DataFrames
X = pd.DataFrame(iris.data)
y = pd.DataFrame(iris.target)

# Create DMatrix
dtrain = xgb.DMatrix(X, y)
dtest = xgb.DMatrix(X, y)

# Set training parameters
xgb_params = {
    "eta": 0.3,
    "max_depth": 3,
    "objective": "multi:softprob",
    "num_class": 3,
    "eval_metric": "mlogloss",
}
steps = 20

# Create dict for evaluation results
evals_result = dict()

# Run training
```

(continues on next page)

(continued from previous page)

```
model = xgb.train(  
    xgb_params,  
    dtrain,  
    steps,  
    evals=[(dtrain, "train")],  
    evals_result=evals_result  
)  
  
# Print evaluation results  
print(f'Evals results:\n{evals_result}')  
  
# Predict results  
prediction = model.predict(dtest)  
  
# Print prediction results  
print(f'Prediction results:\n{prediction}')
```

10.3.6 Modin in the Cloud

Modin implements functionality that allows to transfer computing to the cloud with minimal effort. Please note that this feature is experimental and behavior or interfaces could be changed.

Prerequisites

Sign up with a cloud provider and get credentials file. Note that we supported only AWS currently, more are planned. ([AWS credentials file format](#))

Setup environment

```
pip install modin[remote]
```

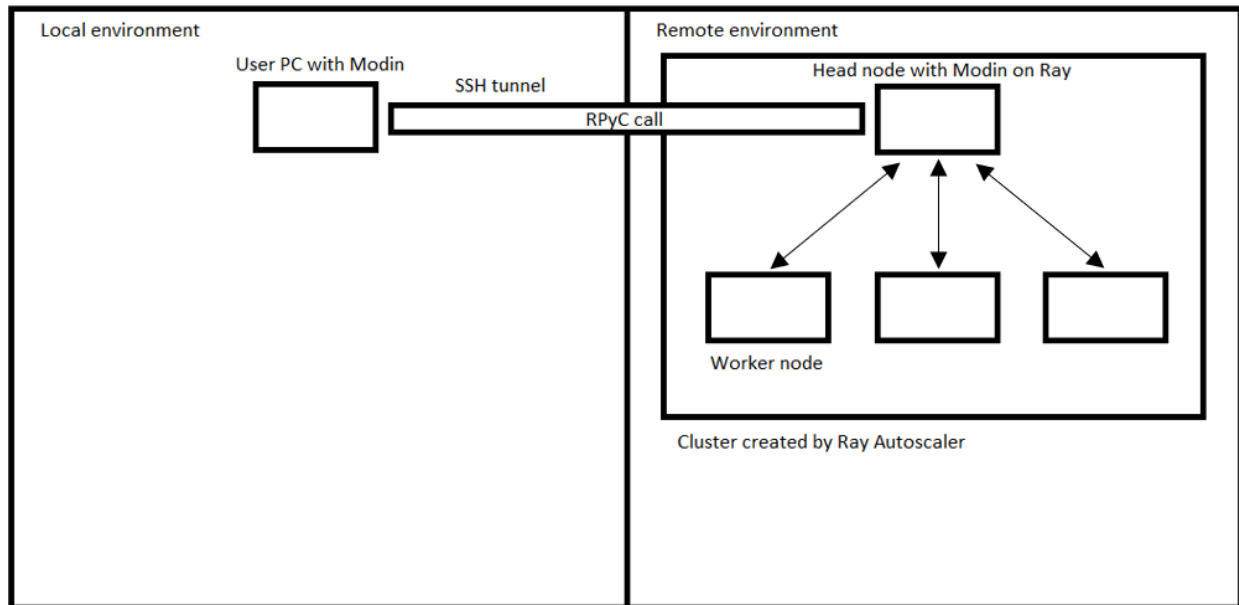
This command install the following dependencies:

- [RPyC](#) - allows to perform remote procedure calls.
- [Cloudpickle](#) - allows pickling of functions and classes, which is used in our distributed runtime.
- [Boto3](#) - allows to create and setup AWS cloud machines. Optional library for Ray Autoscaler.

Notes:

- It also needs Ray Autoscaler component, which is implicitly installed with Ray (note that Ray from conda is now missing that component!). More information in [Ray docs](#).

Architecture



Notes:

- To get maximum performance, you need to try to reduce the amount of data transferred between local and remote environments as much as possible.
- To ensure correct operation, it is necessary to ensure the equivalence of versions of all Python libraries (including the interpreter) in the local and remote environments.

Public interface

exception `modin.experimental.cloud.CannotDestroyCluster`(*args, cause: *Optional[BaseException]* = None, traceback: *Optional[str]* = None, **kw)

Raised when cluster cannot be destroyed in the cloud

exception `modin.experimental.cloud.CannotSpawnCluster`(*args, cause: *Optional[BaseException]* = None, traceback: *Optional[str]* = None, **kw)

Raised when cluster cannot be spawned in the cloud

exception `modin.experimental.cloud.ClusterError`(*args, cause: *Optional[BaseException]* = None, traceback: *Optional[str]* = None, **kw)

Generic cluster operating exception

`modin.experimental.cloud.create_cluster`(provider: *Union[modin.experimental.cloud.cluster.Provider, str]*, credentials: *Optional[str]* = None, region: *Optional[str]* = None, zone: *Optional[str]* = None, image: *Optional[str]* = None, project_name: *Optional[str]* = None, cluster_name: *str* = 'modin-cluster', workers: *int* = 4, head_node: *Optional[str]* = None, worker_node: *Optional[str]* = None, add_conda_packages: *Optional[list]* = None, cluster_type: *str* = 'rayscale') → `modin.experimental.cloud.cluster.BaseCluster`

Creates an instance of a cluster with desired characteristics in a cloud. Upon entering a context via with statement Modin will redirect its work to the remote cluster. Spawned cluster can be destroyed manually, or it will be destroyed when the program exits.

Parameters

- **provider** (*str* or *instance of Provider class*) – Specify the name of the provider to use or a Provider object. If Provider object is given, then credentials, region and zone are ignored.
- **credentials** (*str*, *optional*) – Path to the file which holds credentials used by given cloud provider. If not specified, cloud provider will use its default means of finding credentials on the system.
- **region** (*str*, *optional*) – Region in the cloud where to spawn the cluster. If omitted a default for given provider will be taken.
- **zone** (*str*, *optional*) – Availability zone (part of region) where to spawn the cluster. If omitted a default for given provider and region will be taken.
- **image** (*str*, *optional*) – Image to use for spawning head and worker nodes. If omitted a default for given provider will be taken.
- **project_name** (*str*, *optional*) – Project name to assign to the cluster in cloud, for easier manual tracking.
- **cluster_name** (*str*, *optional*) – Name to be given to the cluster. To spawn multiple clusters in single region and zone use different names.
- **workers** (*int*, *optional*) – How many worker nodes to spawn in the cluster. Head node is not counted for here.
- **head_node** (*str*, *optional*) – What machine type to use for head node in the cluster.
- **worker_node** (*str*, *optional*) – What machine type to use for worker nodes in the cluster.
- **add_conda_packages** (*list*, *optional*) – Custom conda packages for remote environments. By default remote modin version is the same as local version.
- **cluster_type** (*str*, *optional*) – How to spawn the cluster. Currently spawning by Ray autoscaler (“rayscale” for general and “omnisci” for Omnisci-based) is supported

Returns The object that knows how to destroy the cluster and how to activate it as remote context. Note that by default spawning and destroying of the cluster happens in the background, as it’s usually a rather lengthy process.

Return type BaseCluster descendant

Notes

Cluster computation actually can work when proxies are required to access the cloud. You should set normal “http_proxy”/“https_proxy” variables for HTTP/HTTPS proxies and set “MODIN SOCKS_PROXY” variable for SOCKS proxy before calling the function.

Using SOCKS proxy requires Ray newer than 0.8.6, which might need to be installed manually.

`modin.experimental.cloud.get_connection()`

Returns an RPyC connection object to execute Python code remotely on the active cluster.

Usage examples

```

"""
This is a very basic sample script for running things remotely.
It requires `aws_credentials` file to be present in current working directory.
On credentials file format see https://docs.aws.amazon.com/cli/latest/userguide/cli-configure-files.html#cli-configure-files-where
"""

import logging
import modin.pandas as pd
from modin.experimental.cloud import cluster
# set up verbose logging so Ray autoscaler would print a lot of things
# and we'll see that stuff is alive and kicking
logging.basicConfig(format="%(asctime)s %(message)s")
logger = logging.getLogger()
logger.setLevel(logging.DEBUG)
example_cluster = cluster.create("aws", "aws_credentials")
with example_cluster:
    remote_df = pd.DataFrame([1, 2, 3, 4])
    print(len(remote_df)) # len() is executed remotely

```

Some more examples can be found in `examples/cloud` folder.

Modin aims to not only optimize pandas, but also provide a comprehensive, integrated toolkit for data scientists. We are actively developing data science tools such as DataFrame spreadsheet integration, DataFrame algebra, progress bars, SQL queries on DataFrames, and more. Join us on [Slack](#) and [Discourse](#) for the latest updates!

10.3.7 Experimental APIs

Modin also supports these experimental APIs on top of pandas that are under active development.

- `read_csv_glob()` – read multiple files in a directory
- `read_sql()` – add optional parameters for the database connection
- `read_pickle_distributed()` – read multiple files in a directory
- `to_pickle_distributed()` – write to multiple files in a directory

10.3.8 DataFrame partitioning API

Modin DataFrame provides an API to directly access partitions: you can extract physical partitions from a [DataFrame](#), modify their structure by reshuffling or applying some functions, and create a DataFrame from those modified partitions. Visit [pandas partitioning API](#) documentation to learn more.

10.3.9 Modin Spreadsheet API

The Spreadsheet API for Modin allows you to render the dataframe as a spreadsheet to easily explore your data and perform operations on a graphical user interface. The API also includes features for recording the changes made to the dataframe and exporting them as reproducible code. Built on top of Modin and SlickGrid, the spreadsheet interface is able to provide interactive response times even at a scale of billions of rows. See our [Modin Spreadsheet API documentation](#) for more details.

10.3.10 Progress Bar

Visual progress bar for Dataframe operations such as groupby and fillna, as well as for file reading operations such as read_csv. Built using the [tqdm](#) library and Ray execution engine. See [Progress Bar documentation](#) for more details.


```
In [6]: df
```

```
Out[6]:
```

	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	pickup_longitude	pickup_latitude	RateCodeID	store_and_fwd
0	2	2015-01-15 19:05:39	2015-01-15 19:23:42	1	1.59	-73.993896	40.750111	1	
1	1	2015-01-10 20:33:38	2015-01-10 20:53:28	1	3.30	-74.001648	40.724243	1	
2	1	2015-01-10 20:33:38	2015-01-10 20:43:41	1	1.80	-73.963341	40.802788	1	
3	1	2015-01-10 20:33:39	2015-01-10 20:35:31	1	0.50	-74.009087	40.713818	1	
4	1	2015-01-10 20:33:39	2015-01-10 20:52:58	1	3.00	-73.971176	40.762428	1	
...
12748981	1	2015-01-10 19:01:44	2015-01-10 19:05:40	2	1.00	-73.951988	40.786217	1	
12748982	1	2015-01-10 19:01:44	2015-01-10 19:07:26	2	0.80	-73.982742	40.728184	1	
12748983	1	2015-01-10 19:01:44	2015-01-10 19:15:01	1	3.40	-73.979324	40.749550	1	
12748984	1	2015-01-10 19:01:44	2015-01-10 19:17:03	1	1.30	-73.999565	40.738483	1	
12748985	1	2015-01-10 19:01:45	2015-01-10 19:07:33	1	0.70	-73.960350	40.766399	1	

12748986 rows x 19 columns

```
In [7]: df.groupby("passenger_count").count()
```

Estimated completion of line 1: 100%  12/12 [00:06<00:00, 6.64s/it]

```
Out[7]:
```

	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	trip_distance	pickup_longitude	pickup_latitude	RateCodeID	store_and_fwd_flag	drop
passenger_count	0	6565	6565	6565	6565	6565	6565	6565	6565
1	8993870	8993870	8993870	8993870	8993870	8993870	8993870	8993870	8993870
2	1814594	1814594	1814594	1814594	1814594	1814594	1814594	1814594	1814594
3	528486	528486	528486	528486	528486	528486	528486	528486	528486
4	253228	253228	253228	253228	253228	253228	253228	253228	253228
5	697645	697645	697645	697645	697645	697645	697645	697645	697645

10.3.11 Dataframe Algebra

A minimal set of operators that can be composed to express any dataframe query for use in query planning and optimization. See our [paper](#) for more information, and full documentation is coming soon!

10.3.12 SQL on Modin Dataframes

Read about Modin Dataframe support for SQL queries in this recent [blog post](#). Check out the [Modin SQL documentation](#) as well!

```

1  # Importing in memory SQL engine as mdsql
2  # in the next couple of days you will import as follows
3  # - import modin.experimental.sql as mdsql
4
5  # For now you can pip install the sql engine 'pip install pdsql' and include it as follow
6  from dfsql import sql_query
7  import modin.sql as mdsql
8  mdsql.query = sql_query
9
10 # You can then define the query that you want to perform
11 sql_str = "SELECT App,Category,Rating FROM gstore_apps WHERE Price = '0'"
12
13 # And simply apply that query to a dataframe
14 result_df = mdsql.query(sql_str, gstore_apps=gstore_apps_df)
15
16 # Or, in this case, where the query only requires one table,
17 # you can also ignore the FROM part in the query string:
18 query_str = "SELECT App, Category, Rating WHERE Price = '0' "
19
20 # mdsql.query can take query strings without FROM statement
21 # you can specify from as the function argument
22 result_df = mdsql.query(query_str, from=gstore_apps_df)

```

10.3.13 Distributed XGBoost on Modin

Modin provides an implementation of [distributed XGBoost](#) machine learning algorithm on Modin DataFrames. See our [Distributed XGBoost on Modin documentation](#) for details about installation and usage, as well as Modin XGBoost architecture documentation for information about implementation and internal execution flow.

10.4 Optimization Notes

Modin has chosen default values for a lot of the configurations here that provide excellent performance in most cases. This page is for those who love to optimize their code and those who are curious about existing optimizations within Modin. Here you can find more information about Modin's optimizations both for a pipeline of operations as well as for specific operations.

10.4.1 Understanding Modin's partitioning mechanism

Modin's partitioning is crucial for performance; so we recommend expert users to understand Modin's partitioning mechanism and how to tune it in order to achieve better performance.

How Modin partitions a dataframe

Modin uses a partitioning scheme that partitions a dataframe along both axes, resulting in a matrix of partitions. The row and column chunk sizes are computed independently based on the length of the appropriate axis and Modin's special *configuration variables* (`NPartitions` and `MinPartitionSize`):

- `NPartitions` is the maximum number of splits along an axis; by default, it equals to the number of cores on your local machine or cluster of nodes.
- `MinPartitionSize` is the minimum number of rows/columns to do a split. For instance, if `MinPartitionSize` is 32, the column axis will not be split unless the amount of columns is greater than 32. If it is greater, for example, 34, then the column axis is sliced into two partitions: containing 32 and 2 columns accordingly.

Beware that `NPartitions` specifies a limit for the number of partitions *along a single axis*, which means, that the actual limit for the entire dataframe itself is the square of `NPartitions`.

Full-axis functions

Some of the aggregation functions require knowledge about the entire axis, for example at `.apply(foo, axis=0)` the passed function `foo` expects to receive data for the whole column at once.

When a full-axis function is applied, the partitions along this axis are collected at a single worker that processes the function. After the function is done, the partitioning of the data is back to normal.

Note that the amount of remote calls is equal to the number of partitions, which means that since the number of partitions is decreased for full-axis functions it also decreases the potential for parallelism.

Also note, that reduce functions such as `.sum()`, `.mean()`, `.max()`, etc, are not considered to be full-axis, so they do not suffer from the decreasing level of parallelism.

How to tune partitioning

As you can see from the examples above, the more the dataframe's shape is closer to a square, the closer the number of partitions to the square of NPartitions. In the case of NPartitions equals to the number of workers, that means that a single worker is going to process multiple partitions at once, which slows down overall performance.

If your workflow mainly operates with wide dataframes and non-full-axis functions, it makes sense to reduce the NPartitions value so a single worker would process a single partition.

Copy-pastable example, showing how tuning NPartitions value for wide frames may improve performance on your machine:

```
from multiprocessing import cpu_count
from modin.distributed.dataframe.pandas import unwrap_partitions
import modin.config as cfg
import modin.pandas as pd
import numpy as np
import timeit

# Generating data for a square-like dataframe
data = np.random.randint(0, 100, size=(5000, 5000))

# Explicitly setting `NPartitions` to its default value
cfg.NPartitions.put(cpu_count())

# Each worker processes `cpu_count()` amount of partitions
df = pd.DataFrame(data)
print(f"NPartitions: {cfg.NPartitions.get()}")
# Getting raw partitions to count them
partitions_shape = np.array(unwrap_partitions(df)).shape
print(
    f"The frame has {partitions_shape[0]}x{partitions_shape[1]}={np.prod(partitions_
↪shape)} partitions "
    f"when the CPU has only {cpu_count()} cores."
)
print(f"10 times of .abs(): {timeit.timeit(lambda: df.abs(), number=10)}s.")
# Possible output:
#   NPartitions: 112
#   The frame has 112x112=12544 partitions when the CPU has only 112 cores.
#   10 times of .abs(): 23.64s.

# Taking a square root of the the current `cpu_count` to make more even partitioning
cfg.NPartitions.put(int(cpu_count() ** 0.5))

# Each worker processes a single partition
df = pd.DataFrame(data)
print(f"NPartitions: {cfg.NPartitions.get()}")
# Getting raw partitions to count them
partitions_shape = np.array(unwrap_partitions(df)).shape
print(
    f"The frame has {partitions_shape[0]}x{partitions_shape[1]}={np.prod(partitions_
↪shape)} "
    f"when the CPU has {cpu_count()} cores."
```

(continues on next page)

(continued from previous page)

```

)
print(f"10 times of .abs(): {timeit.timeit(lambda: df.abs(), number=10)}s.")
# Possible output:
#   NPartitions: 10
#   The frame has 10x10=100 partitions when the CPU has 112 cores.
#   10 times of .abs(): 0.25s.

```

10.4.2 Avoid iterating over Modin DataFrame

Use `df.apply()` or other aggregation methods when possible instead of iterating over a dataframe. For-loops don't scale and forces the distributed data to be collected back at the driver.

Copy-pastable example, showing how replacing a for-loop to the equivalent `.apply()` may improve performance:

```

import modin.pandas as pd
import numpy as np
from timeit import default_timer as timer

data = np.random.randint(1, 100, (2 ** 10, 2 ** 2))

md_df = pd.DataFrame(data)

result = []
t1 = timer()
# Iterating over a dataframe forces to collect distributed data to the driver and doesn't
↪ scale
for idx, row in md_df.iterrows():
    result.append((row[1] + row[2]) / row[3])
print(f"Filling a list by iterating a Modin frame: {timer() - t1:.2f}s.")
# Possible output: 36.15s.

t1 = timer()
# Using `.apply()` perfectly scales to all axis-partitions
result = md_df.apply(lambda row: (row[1] + row[2]) / row[3], axis=1).to_numpy().tolist()
print(f"Filling a list by using '.apply()' and converting the result to a list: {timer() -
↪ t1:.2f}s.")
# Possible output: 0.22s.

```

10.4.3 Use Modin's Dataframe Algebra API to implement custom parallel functions

Modin provides a set of low-level parallel-implemented operators which can be used to build most of the aggregation functions. These operators are present in the algebra module. Modin DataFrame allows users to use their own aggregations built with this module. Visit the appropriate section of the documentation for the steps to do it.

10.4.4 Avoid mixing pandas and Modin DataFrames

Although Modin is considered to be a drop-in replacement for pandas, Modin and pandas are not intended to be used together in a single flow. Passing a pandas DataFrame as an argument for a Modin's DataFrame method may either slowdown the function (because it has to process non-distributed object) or raise an error. You would also get an undefined behavior if you pass a Modin DataFrame as an input to pandas methods, since pandas identifies Modin's objects as a simple iterable, and so can't leverage its benefits as a distributed dataframe.

Copy-pastable example, showing how mixing pandas and Modin DataFrames in a single flow may bottleneck performance:

```
import modin.pandas as pd
import numpy as np
import timeit
import pandas

data = np.random.randint(0, 100, (2 ** 20, 2 ** 2))

md_df, md_df_copy = pd.DataFrame(data), pd.DataFrame(data)
pd_df, pd_df_copy = pandas.DataFrame(data), pandas.DataFrame(data)

print("concat modin frame + pandas frame:")
# Concatenating modin frame + pandas frame using modin '.concat()'
# This case is bad because Modin have to process non-distributed pandas object
time = timeit.timeit(lambda: pd.concat([md_df, pd_df]), number=10)
print(f"\t{time}s.\n")
# Possible output: 0.44s.

print("concat modin frame + modin frame:")
# Concatenating modin frame + modin frame using modin '.concat()'
# This is an ideal case, Modin is being used as intended
time = timeit.timeit(lambda: pd.concat([md_df, md_df_copy]), number=10)
print(f"\t{time}s.\n")
# Possible output: 0.05s.

print("concat pandas frame + pandas frame:")
# Concatenating pandas frame + pandas frame using pandas '.concat()'
time = timeit.timeit(lambda: pandas.concat([pd_df, pd_df_copy]), number=10)
print(f"\t{time}s.\n")
# Possible output: 0.31s.

print("concat pandas frame + modin frame:")
# Concatenating pandas frame + modin frame using pandas '.concat()'
time = timeit.timeit(lambda: pandas.concat([pd_df, md_df]), number=10)
print(f"\t{time}s.\n")
# Possible output: TypeError
```

10.4.5 Operation-specific optimizations

merge

merge operation in Modin uses the broadcast join algorithm: combining a right Modin DataFrame into a pandas DataFrame and broadcasting it to the row partitions of the left Modin DataFrame. In order to minimize interprocess communication cost when doing an inner join you may want to swap left and right DataFrames.

```
import modin.pandas as pd
import numpy as np

left_data = np.random.randint(0, 100, size=(2**8, 2**8))
right_data = np.random.randint(0, 100, size=(2**12, 2**12))

left_df = pd.DataFrame(left_data)
right_df = pd.DataFrame(right_data)
%timeit left_df.merge(right_df, how="inner", on=10)
3.59 s  107 ms per loop (mean std. dev. of 7 runs, 1 loop each)

%timeit right_df.merge(left_df, how="inner", on=10)
1.22 s  40.1 ms per loop (mean std. dev. of 7 runs, 1 loop each)
```

Note that result columns order may differ for first and second merge.

SUPPORTED APIS

For your convenience, we have compiled a list of currently implemented APIs and methods available in Modin. This documentation is updated as new methods and APIs are merged into the master branch, and not necessarily correct as of the most recent release.

To view the docs for the most recent release, check that you're viewing the [stable version](#) of the docs.

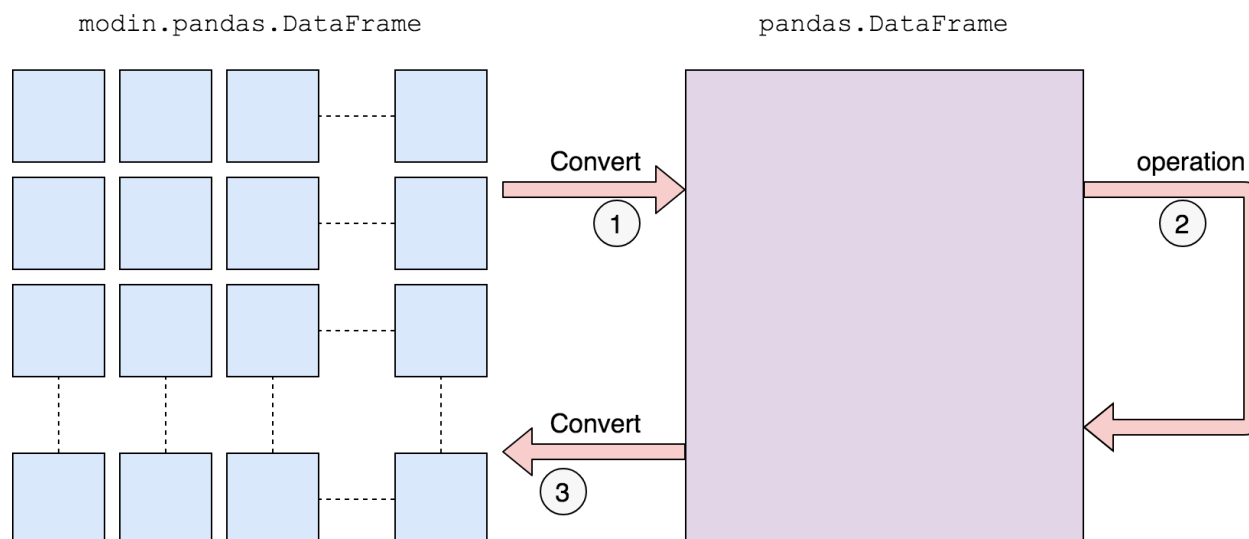
In order to install the latest version of Modin, follow the directions found on the [installation page](#).

11.1 Questions on implementation details

If you have a question about the implementation details or would like more information about an API or method in Modin, please contact the Modin [developer mailing list](#).

11.1.1 Defaulting to pandas

Currently Modin does not support distributed execution for all methods from pandas API. The remaining unimplemented methods are being executed in a mode called “default to pandas”. This allows users to continue using Modin even though their workloads contain functions not yet implemented in Modin. Here is a diagram of how we convert to pandas and perform the operation:



We first convert to a pandas DataFrame, then perform the operation. **There is a performance penalty for going from a partitioned Modin DataFrame to pandas because of the communication cost and single-threaded nature of pan-**

das. Once the pandas operation has completed, we convert the DataFrame back into a partitioned Modin DataFrame. This way, operations performed after something defaults to pandas will be optimized with Modin.

The exact methods we have implemented are listed in the respective subsections:

- *DataFrame*
- *Series*
- *utilities*
- *I/O*

We have taken a community-driven approach to implementing new methods. We did a [study on pandas usage](#) to learn what the most-used APIs are. Modin currently supports **93%** of the pandas API based on our study of pandas usage, and we are actively expanding the API. **To request implementation, send an email to feature_requests@modin.org or file an issue at <https://github.com/modin-project/modin/issues>.**

11.1.2 pd.DataFrame supported APIs

The following table lists both implemented and not implemented methods. If you have need of an operation that is listed as not implemented, feel free to open an issue on the [GitHub repository](#), or give a thumbs up to already created issues. Contributions are also welcome!

The following table is structured as follows: The first column contains the method name. The second column contains link to a description of corresponding pandas method. The third column is a flag for whether or not there is an implementation in Modin for the method in the left column. Y stands for yes, N stands for no, P stands for partial (meaning some parameters may not be supported yet), and D stands for default to pandas.

Note: Currently third column reflects implementation status for Ray and Dask engines. By default, support for a method in the Omnisai engine could be treated as D unless Notes column contains additional information. Similarly, by default Notes contains information about Ray and Dask engines unless Omnisai is explicitly mentioned.

DataFrame method	pandas Doc link	Implemented? (Y/N/P/D)	Notes for Current implementation
T	T	Y	
abs	abs	Y	
add	add	Y	Ray and Dask: Shuffles data in operations between DataFrames. Omnisci: P, support binary operations on scalars and projections of the same frame, otherwise D
add_prefix	add_prefix	Y	
add_suffix	add_suffix	Y	
agg / aggregate	agg / aggregate	P	<ul style="list-style-type: none"> • Dictionary func parameter defaults to pandas • Numpy operations default to pandas

continues on next page

Table 1 – continued from previous page

align	align	D	
all	all	Y	
any	any	Y	
append	append	Y	Omnisci: Y but sort and ignore_index parameters ignored
apply	apply	Y	See agg
applymap	applymap	Y	
asfreq	asfreq	D	
asof	asof	Y	
assign	assign	Y	
astype	astype	Y	Omnisci: P, int``<-> ``float supported
at	at	Y	
at_time	at_time	Y	
axes	axes	Y	
between_time	between_time	Y	
bfill	bfill	Y	
bool	bool	Y	
boxplot	boxplot	D	
clip	clip	Y	
combine	combine	Y	
combine_first	combine_first	Y	
compare	compare	Y	
copy	copy	Y	
corr	corr	Y	Correlation floating point precision may slightly differ from pandas. For now pearson method is available only. For other methods defaults to pandas
corrwith	corrwith	D	
count	count	Y	Omnisci: P, only default params supported, otherwise D
cov	cov	Y	Covariance floating point precision may slightly differ from pandas
cummax	cummax	Y	
cummin	cummin	Y	
cumprod	cumprod	Y	
cumsum	cumsum	Y	
describe	describe	Y	
diff	diff	Y	
div	div	Y	See add
divide	divide	Y	See add
dot	dot	Y	
drop	drop	Y	Omnisci: P since row drop unsupported
droplevel	droplevel	Y	
drop_duplicates	drop_duplicates	D	

continues on next page

Table 1 – continued from previous page

dropna	dropna	Y	Omnisci: P since thresh and axis params unsupported
dtypes	dtypes	Y	Omnisci: Y
duplicated	duplicated	Y	
empty	empty	Y	
eq	eq	Y	See add
equals	equals	Y	Requires shuffle, can be further optimized
eval	eval	Y	
ewm	ewm	D	
expanding	expanding	D	
explode	explode	Y	
ffill	ffill	Y	
fillna	fillna	P	value parameter of type DataFrame defaults to pandas. Omnisci: P, params limit, downcast and method unsupported. Also only axis = 0 supported for now
filter	filter	Y	
first	first	Y	
first_valid_index	first_valid_index	Y	
floordiv	floordiv	Y	See add
from_dict	from_dict	D	
from_records	from_records	D	
ge	ge	Y	See add
get	get	Y	
groupby	groupby	Y	Not yet optimized for all operations. Omnisci: P. count, sum, size, mean, nunique, std, skew supported, otherwise D
gt	gt	Y	See add
head	head	Y	
hist	hist	D	
iat	iat	Y	
idxmax	idxmax	Y	
idxmin	idxmin	Y	
iloc	iloc	Y	Omnisci: P, read access fully supported, write access: no row and 2D assignments support
infer_objects	infer_objects	D	
info	info	Y	
insert	insert	Y	
interpolate	interpolate	D	
isin	isin	Y	
isna	isna	Y	
isnull	isnull	Y	

continues on next page

Table 1 – continued from previous page

items	items	Y	
iteritems	iteritems	P	Modin does not parallelize iteration in Python
iterrows	iterrows	P	Modin does not parallelize iteration in Python
itertuples	itertuples	P	Modin does not parallelize iteration in Python
join	join	P	When on is set to right or outer it defaults to pandas
keys	keys	Y	
kurt	kurt	Y	
kurtosis	kurtosis	Y	
last	last	Y	
last_valid_index	last_valid_index	Y	
le	le	Y	See add
loc	loc	P	We do not support: boolean array, callable. Omnisci: P, read access fully supported, write access: no row and 2D assignments support
lookup	lookup	D	
lt	lt	Y	See add
mad	mad	Y	
mask	mask	D	
max	max	Y	Omnisci: P, only default params supported, otherwise D
mean	mean	P	Modin defaults to pandas if given the level param. Omnisci: P. D for level, axis, skipna and numeric_only params
median	median	P	Modin defaults to pandas if given the level param.
melt	melt	Y	
memory_usage	memory_usage	Y	

continues on next page

Table 1 – continued from previous page

merge	merge	P	Implemented the following cases: <code>left_index=True</code> and <code>right_index=True</code> , <code>how=left</code> and <code>how=inner</code> for all values of parameters except <code>left_index=True</code> and <code>right_index=False</code> or <code>left_index=False</code> and <code>right_index=True</code> . Defaults to pandas otherwise. Omnisci: P, only non-index joins for <code>how=left</code> and <code>how=inner</code> with explicit <code>on</code> are supported
min	min	Y	Omnisci: P, only default params supported, otherwise D
mod	mod	Y	See add
mode	mode	Y	
mul	mul	Y	See add
multiply	multiply	Y	See add
ndim	ndim	Y	
ne	ne	Y	See add
nlargest	nlargest	Y	
notna	notna	Y	
notnull	notnull	Y	
nsmallest	nsmallest	Y	
nunique	nunique	Y	Omnisci: P, no support for <code>axis!=0</code> and <code>dropna=False</code>
pct_change	pct_change	D	
pipe	pipe	Y	
pivot	pivot	Y	
pivot_table	pivot_table	Y	
plot	plot	D	
pop	pop	Y	
pow	pow	Y	See add; Omnisci: D
prod	prod	Y	
product	product	Y	
quantile	quantile	Y	
query	query	P	Local variables not yet supported
radd	radd	Y	See add
rank	rank	Y	
rdiv	rdiv	Y	See add; Omnisci: D
reindex	reindex	Y	Shuffles data
reindex_like	reindex_like	D	
rename	rename	Y	
rename_axis	rename_axis	Y	

continues on next page

Table 1 – continued from previous page

reorder_levels	reorder_levels	Y	
replace	replace	Y	
resample	resample	Y	
reset_index	reset_index	Y	Omnisci: P. D for level parameter
rfloordiv	rfloordiv	Y	See add; Omnisci: D
rmod	rmod	Y	See add; Omnisci: D
rmul	rmul	Y	See add
rolling	rolling	Y	
round	round	Y	
rpow	rpow	Y	See add; Omnisci: D
rsub	rsub	Y	See add; Omnisci: D
rtruediv	rtruediv	Y	See add; Omnisci: D
sample	sample	Y	
select_dtypes	select_dtypes	Y	
sem	sem	P	Modin defaults to pandas if given the level param.
set_axis	set_axis	Y	
set_index	set_index	Y	
shape	shape	Y	Omnisci: Y
shift	shift	Y	
size	size	Y	
skew	skew	P	Modin defaults to pandas if given the level param
slice_shift	slice_shift	Y	
sort_index	sort_index	Y	
sort_values	sort_values	Y	Shuffles data; Omnisci: Y
sparse	sparse	N	
squeeze	squeeze	Y	
stack	stack	Y	
std	std	P	Modin defaults to pandas if given the level param.
style	style	D	
sub	sub	Y	See add
subtract	subtract	Y	See add; Omnisci: D
sum	sum	Y	Omnisci: P, only default params supported, otherwise D
swapaxes	swapaxes	Y	
swaplevel	swaplevel	Y	
tail	tail	Y	
take	take	Y	
to_clipboard	to_clipboard	D	
to_csv	to_csv	Y	
to_dict	to_dict	D	
to_excel	to_excel	D	
to_feather	to_feather	D	
to_gbq	to_gbq	D	
to_hdf	to_hdf	D	
to_html	to_html	D	

continues on next page

Table 1 – continued from previous page

to_json	to_json	D	
to_latex	to_latex	D	
to_parquet	to_parquet	D	
to_period	to_period	D	
to_pickle	to_pickle	D	Experimental implementation: to_pickle_distributed
to_records	to_records	D	
to_sql	to_sql	Y	
to_stata	to_stata	D	
to_string	to_string	D	
to_timestamp	to_timestamp	D	
to_xarray	to_xarray	D	
transform	transform	Y	
transpose	transpose	Y	
truediv	truediv	Y	See add
truncate	truncate	Y	
tshift	tshift	Y	
tz_convert	tz_convert	Y	
tz_localize	tz_localize	Y	
unstack	unstack	Y	
update	update	Y	
values	values	Y	
value_counts	value_counts	D	
var	var	P	Modin defaults to pandas if given the level param.
where	where	Y	

11.1.3 pd.Series supported APIs

The following table lists both implemented and not implemented methods. If you have need of an operation that is listed as not implemented, feel free to open an issue on the [GitHub repository](#), or give a thumbs up to already created issues. Contributions are also welcome!

The following table is structured as follows: The first column contains the method name. The second column is a flag for whether or not there is an implementation in Modin for the method in the left column. Y stands for yes, N stands for no, P stands for partial (meaning some parameters may not be supported yet), and D stands for default to pandas. To learn more about the implementations that default to pandas, see the related section on [Defaulting to pandas](#).

Note: Currently, the second column reflects implementation status for Ray and Dask engines. By default, support for a method in the `Omnisci` engine could be treated as D unless `Notes` column contains additional information. Similarly, by default `Notes` contains information about Ray and Dask engines unless `Omnisci` is explicitly mentioned.

Series method	Modin Implementation? (Y/N/P/D)	Notes for Current implementation
abs	Y	
add	Y	Omnisci: P, support binary operations on scalars and projection
add_prefix	Y	
add_suffix	Y	
agg	Y	

Table 2 – continued from previous page

aggregate	Y	
align	D	
all	Y	
any	Y	
append	Y	Omnisci: Y but sort and ignore_index parameters ignored
apply	Y	
argmax	Y	
argmin	Y	
argsort	D	
array	D	
asfreq	D	
asobject	D	
asof	Y	
astype	Y	Omnisci: P, int<->float supported
at	Y	
at_time	Y	
autocorr	Y	
axes	Y	
base	D	
between	D	
between_time	Y	
bfill	Y	
bool	Y	
cat	D	
clip	Y	
combine	Y	
combine_first	Y	
compare	Y	
compress	D	
copy	Y	
corr	Y	Correlation floating point precision may slightly differ from pandas
count	Y	Omnisci: P, only default params supported, otherwise D
cov	Y	Covariance floating point precision may slightly differ from pandas
cummax	Y	
cummin	Y	
cumprod	Y	
cumsum	Y	
data	D	
describe	Y	
diff	Y	
div	Y	See add
divide	Y	See add
divmod	Y	
dot	Y	
drop	Y	Omnisci: D
drop_duplicates	Y	
droplevel	Y	
dropna	Y	Omnisci: P since thresh and axis params unsupported
dt	Y	Omnisci: P, only year, month, day and hour supported, otherwise D
dtype	Y	

Table 2 – continued from previous page

dtypes	Y	Omnisci: Y
duplicated	Y	
empty	Y	
eq	Y	See add
equals	Y	
ewm	D	
expanding	D	
explode	Y	
factorize	D	
ffill	Y	
fillna	Y	Omnisci: P, params limit, downcast and method unsupported
filter	Y	
first	Y	
first_valid_index	Y	
flags	D	
floordiv	Y	See add
from_array	D	
ftype	Y	
ge	Y	See add
get	Y	
get_dtype_counts	Y	
get_ftype_counts	Y	
get_value	D	
get_values	D	
groupby	D	Omnisci: P. count, sum, size supported, otherwise D
gt	Y	See add
hasnans	Y	
head	Y	
hist	D	
iat	Y	
idxmax	Y	
idxmin	Y	
iloc	Y	Omnisci: P, read access fully supported, write access: no row
imag	D	
index	Y	
infer_objects	D	
interpolate	D	
is_monotonic	Y	
is_monotonic_decreasing	Y	
is_monotonic_increasing	Y	
is_unique	Y	
isin	Y	
isna	Y	
isnull	Y	
item	Y	
items	Y	
itemsize	D	
iteritems	Y	
keys	Y	
kurt	Y	

Table 2 – continued from previous page

kurtosis	Y	
last	Y	
last_valid_index	Y	
le	Y	See add
loc	Y	Omnisci: P, read access fully supported, write access: no row
lt	Y	See add
mad	Y	
map	Y	
mask	D	
max	Y	Omnisci: P, only default params supported, otherwise D
mean	P	Modin defaults to pandas if given the level param. Omnisci:
median	P	Modin defaults to pandas if given the level param.
memory_usage	Y	
min	Y	Omnisci: P, only default params supported, otherwise D
mod	Y	See add
mode	Y	
mul	Y	See add
multiply	Y	See add
name	Y	
nbytes	D	
ndim	Y	
ne	Y	See add
nlargest	Y	
nonzero	Y	
notna	Y	
notnull	Y	
nsmallest	Y	
nunique	Y	Omnisci: P, no support for axis!=0 and dropna=False
pct_change	D	
pipe	Y	
plot	D	
pop	Y	
pow	Y	See add; Omnisci: D
prod	Y	
product	Y	
ptp	D	
put	D	
quantile	Y	
radd	Y	See add
rank	Y	
ravel	Y	
rdiv	Y	See add; Omnisci: D
rdivmod	Y	
real	D	
reindex	Y	
reindex_like	Y	
rename	Y	
rename_axis	Y	
reorder_levels	D	
repeat	D	

Table 2 – continued from previous page

replace	Y	
resample	Y	
reset_index	Y	Omnisci: P, D for level parameter
rfloordiv	Y	See add; Omnisci: D
rmod	Y	See add; Omnisci: D
rmul	Y	See add
rolling	Y	
round	Y	
rpow	Y	See add; Omnisci: D
rsub	Y	See add; Omnisci: D
rtruediv	Y	See add; Omnisci: D
sample	Y	
searchsorted	Y	
sem	P	Modin defaults to pandas if given the level param.
set_axis	Y	
set_value	D	
shape	Y	Omnisci: Y
shift	Y	
size	Y	
skew	P	Modin defaults to pandas if given the level param.
slice_shift	Y	
sort_index	Y	
sort_values	D	Omnisci: Y
sparse	Y	
squeeze	Y	
std	P	Modin defaults to pandas if given the level param.
str	Y	
strides	D	
sub	Y	See add
subtract	Y	See add; Omnisci: D
sum	Y	Omnisci: P, only default params supported, otherwise D
swapaxes	Y	
swaplevel	Y	
tail	Y	
take	Y	
to_clipboard	D	
to_csv	D	
to_dict	D	
to_excel	D	
to_frame	Y	
to_hdf	D	
to_json	D	
to_latex	D	
to_list	D	
to_numpy	D	
to_period	D	
to_pickle	D	
to_sql	Y	
to_string	D	
to_timestamp	D	

Table 2 – continued from previous page

to_xarray	D	
tolist	D	
transform	Y	
transpose	Y	
truediv	Y	See add
truncate	Y	
tshift	Y	
tz_convert	Y	
tz_localize	Y	
unique	Y	
unstack	Y	
update	Y	
valid	D	
value_counts	Y	The indices order of resulting object may differ from pandas.
values	Y	
var	P	Modin defaults to pandas if given the level param.
view	D	
where	Y	

11.1.4 pandas Utilities Supported

If you run `import modin.pandas as pd`, the following operations are available from `pd.<op>`, e.g. `pd.concat`. If you do not see an operation that pandas enables and would like to request it, feel free to [open an issue](#). Make sure you tell us your primary use-case so we can make it happen faster!

The following table is structured as follows: The first column contains the method name. The second column is a flag for whether or not there is an implementation in Modin for the method in the left column. Y stands for yes, N stands for no, P stands for partial (meaning some parameters may not be supported yet), and D stands for default to pandas.

Note: Currently, the second column reflects implementation status for Ray and Dask engines. By default, support for a method in the Omnisci engine could be treated as D unless Notes column contains additional information. Similarly, by default Notes contains information about Ray and Dask engines unless Omnisci is explicitly mentioned.

Utility method	Modin (Y/N/P/D)	Implementation?	Notes for Current implementation
<code>pd.concat</code>	Y		Omnisci: Y but <code>sort</code> and <code>ignore_index</code> parameters ignored
<code>pd.eval</code>	Y		
<code>pd.unique</code>	Y		
<code>pd.value_counts</code>	Y		The indices order of resulting object may differ from pandas.
<code>pd.cut</code>	D		
<code>pd.to_numeric</code>	D		
<code>pd.factorize</code>	D		
<code>pd.qcut</code>	D		
<code>pd.match</code>	D		
<code>pd.to_datetime</code>	D		
<code>pd.get_dummies</code>	Y		
<code>pd.date_range</code>	D		
<code>pd.bdate_range</code>	D		
<code>pd.to_timedelta</code>	D		
<code>pd.options</code>	Y		
<code>pd.datetime</code>	D		

Other objects & structures

This list is a list of objects not currently distributed by Modin. All of these objects are compatible with the distributed components of Modin. If you are interested in contributing a distributed version of any of these objects, feel free to open a [pull request](#).

- Panel
- Index
- MultiIndex
- CategoricalIndex
- DatetimeIndex
- Timedelta
- Timestamp
- NaT
- PeriodIndex
- Categorical
- Interval
- UInt8Dtype
- UInt16Dtype
- UInt32Dtype
- UInt64Dtype
- SparseDtype
- Int8Dtype

- Int16Dtype
- Int32Dtype
- Int64Dtype
- CategoricalDtype
- DatetimeTZDtype
- IntervalDtype
- PeriodDtype
- RangeIndex
- Int64Index
- UInt64Index
- Float64Index
- TimedeltaIndex
- IntervalIndex
- IndexSlice
- TimeGrouper
- Grouper
- array
- Period
- DateOffset
- ExcelWriter
- SparseArray
- SparseSeries
- SparseDataFrame

11.1.5 `pd.read_<file>` and I/O APIs

A number of IO methods default to pandas. We have parallelized `read_csv` and `read_parquet`, though many of the remaining methods can be relatively easily parallelized. Some of the operations default to the pandas implementation, meaning it will read in serially as a single, non-distributed DataFrame and distribute it. Performance will be affected by this.

The following table is structured as follows: The first column contains the method name. The second column is a flag for whether or not there is an implementation in Modin for the method in the left column. Y stands for yes, N stands for no, P stands for partial (meaning some parameters may not be supported yet), and D stands for default to pandas.

Note: Currently, the second column reflects implementation status for Ray and Dask engines. By default, support for a method in the `OmniSci` engine could be treated as D unless `Notes` column contains additional information.

IO method	Modin implementation? (Y/N/P/D)	Notes for Current implementation
read_csv	Y	Omnisci: P, only basic cases and parameters supported: filepath_or_buffer can be local file only, sep, delimiter, header (partly) names, usecols, dtype, true/false_values, skiprows (partly) skip_blank_lines (partly), parse_dates (partly), compression (inferred automatically, should not be specified), quotechar, escapechar, doublequote, delim_whitespace
read_table	Y	
read_parquet	Y	
read_json	P	Implemented for lines=True
read_hdf	D	
read_clipboard	D	
read_excel	D	
read_hdf	D	
read_feather	Y	
read_msgpack	D	
read_stata	D	
read_sas	D	
read_pickle	D	Experimental implementation: read_pickle_distributed
read_sql	Y	

DEVELOPMENT

12.1 Contributing

12.1.1 Getting Started

If you're interested in getting involved in the development of Modin, but aren't sure where start, take a look at the issues tagged [Good first issue](#) or [Documentation](#). These are issues that would be good for getting familiar with the codebase and better understanding some of the more complex components of the architecture. There is documentation here about the [architecture](#) that you will want to review in order to get started.

Also, feel free to join the discussions on the [developer mailing list](#).

If you want a quick guide to getting your development environment setup, please use [the contributing instructions on GitHub](#).

12.1.2 Certificate of Origin

To keep a clear track of who did what, we use a *sign-off* procedure (same requirements for using the signed-off-by process as the Linux kernel has <https://www.kernel.org/doc/html/v4.17/process/submitting-patches.html>) on patches or pull requests that are being sent. The sign-off is a simple line at the end of the explanation for the patch, which certifies that you wrote it or otherwise have the right to pass it on as an open-source patch. The rules are pretty simple: if you can certify the below:

CERTIFICATE OF ORIGIN V 1.1

“By making a contribution to this project, I certify that:

1.) The contribution was created in whole or in part by me and I have the right to submit it under the open source license indicated in the file; or 2.) The contribution is based upon previous work that, to the best of my knowledge, is covered under an appropriate open source license and I have the right under that license to submit that work with modifications, whether created in whole or in part by me, under the same open source license (unless I am permitted to submit under a different license), as indicated in the file; or 3.) The contribution was provided directly to me by some other person who certified (a), (b) or (c) and I have not modified it. 4.) I understand and agree that this project and the contribution are public and that a record of the contribution (including all personal information I submit with it, including my sign-off) is maintained indefinitely and may be redistributed consistent with this project or the open source license(s) involved.”

This is my commit message

Signed-off-by: Awesome Developer <developer@example.org>

Code without a proper signoff cannot be merged into the master branch. Note: You must use your real name (sorry, no pseudonyms or anonymous contributions.)

The text can either be manually added to your commit body, or you can add either `-s` or `--signoff` to your usual `git commit` commands:

```
git commit --signoff
git commit -s
```

This will use your default git configuration which is found in `.git/config`. To change this, you can use the following commands:

```
git config --global user.name "Awesome Developer"
git config --global user.email "awesome.developer@example.org"
```

If you have authored a commit that is missing the signed-off-by line, you can amend your commits and push them to GitHub.

```
git commit --amend --signoff
```

If you've pushed your changes to GitHub already you'll need to force push your branch after this with `git push -f`.

12.1.3 Commit Message formatting

To ensure that all commit messages in the master branch follow a specific format, we enforce that all commit messages must follow the following format:

```
FEAT-#9999: Add `DataFrame.rolling` functionality, to enable rolling window operations
```

The FEAT component represents the type of commit. This component of the commit message can be one of the following:

- FEAT: A new feature that is added
- DOCS: Documentation improvements or updates
- FIX: A bugfix contribution
- REFACTOR: Moving or removing code without change in functionality
- TEST: Test updates or improvements

The #9999 component of the commit message should be the issue number in the Modin GitHub issue tracker: <https://github.com/modin-project/modin/issues>. This is important because it links commits to their issues.

The commit message should follow a colon (:) and be descriptive and succinct.

12.1.4 General Rules for committers

- Try to write a PR name as descriptive as possible.
- Try to keep PRs as small as possible. One PR should be making one semantically atomic change.
- Don't merge your own PRs even if you are technically able to do it.

12.1.5 Development Dependencies

We recommend doing development in a virtualenv or conda environment, though this decision is ultimately yours. You will want to run the following in order to install all of the required dependencies for running the tests and formatting the code:

```
conda env create --file environment-dev.yml
# or
pip install -r requirements-dev.txt
```

12.1.6 Code Formatting and Lint

We use [black](#) for code formatting. Before you submit a pull request, please make sure that you run the following from the project root:

```
black modin/ asv_bench/benchmarks scripts/doc_checker.py
```

We also use [flake8](#) to check linting errors. Running the following from the project root will ensure that it passes the lint checks on Github Actions:

```
flake8 modin/ asv_bench/benchmarks scripts/doc_checker.py
```

We test that this has been run on our [Github Actions](#) test suite. If you do this and find that the tests are still failing, try updating your version of black and flake8.

12.1.7 Adding a test

If you find yourself fixing a bug or adding a new feature, don't forget to add a test to the test suite to verify its correctness! More on testing and the layout of the tests can be found in our testing documentation. We ask that you follow the existing structure of the tests for ease of maintenance.

12.1.8 Running the tests

To run the entire test suite, run the following from the project root:

```
pytest modin/pandas/test
```

The test suite is very large, and may take a long time if you run every test. If you've only modified a small amount of code, it may be sufficient to run a single test or some subset of the test suite. In order to run a specific test run:

```
pytest modin/pandas/test::test_new_functionality
```

The entire test suite is automatically run for each pull request.

12.1.9 Performance measurement

We use [Asv](#) tool for performance tracking of various Modin functionality. The results can be viewed here: [Asv dashboard](#).

More information can be found in the [Asv readme](#).

12.1.10 Building documentation

To build the documentation, please follow the steps below from the project root:

```
cd docs
pip install -r requirements-doc.txt
sphinx-build -b html . build
```

To visualize the documentation locally, run the following from *build* folder:

```
python -m http.server <port>
# python -m http.server 1234
```

then open the browser at *0.0.0.0:<port>* (e.g. *0.0.0.0:1234*).

12.1.11 Contributing a new execution framework or in-memory format

If you are interested in contributing support for a new execution framework or in-memory format, please make sure you understand the [architecture](#) of Modin.

The best place to start the discussion for adding a new execution framework or in-memory format is the [developer mailing list](#).

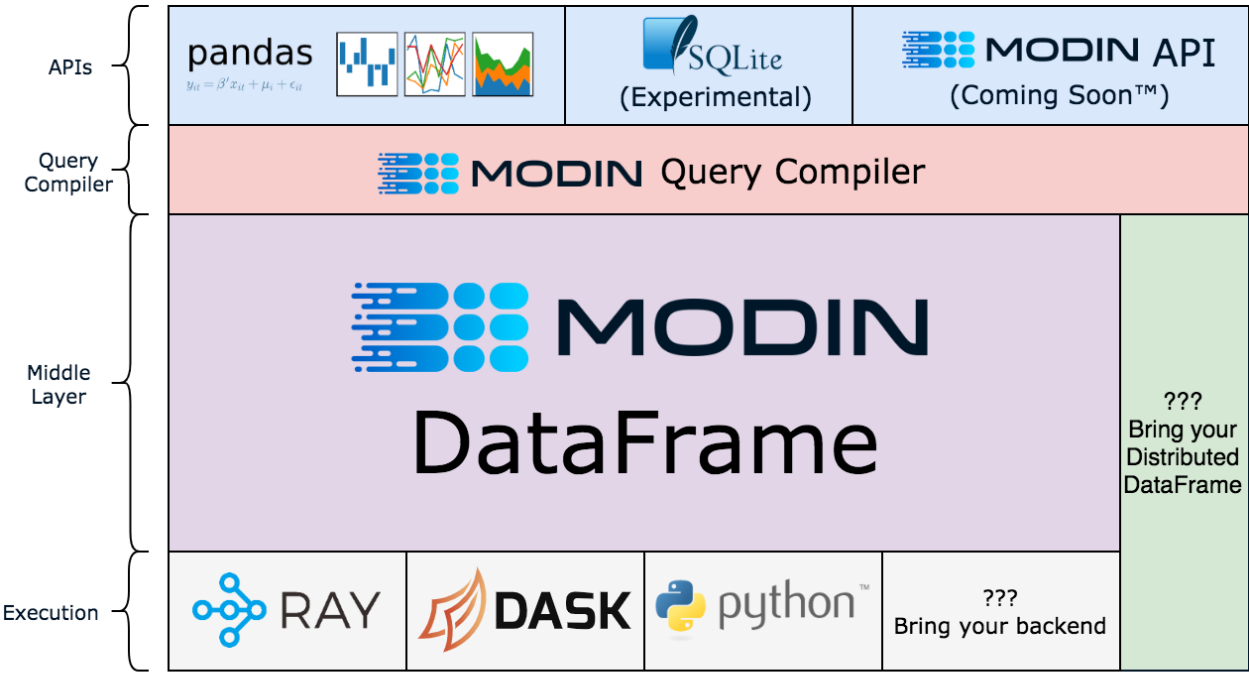
More docs on this coming soon...

12.2 System Architecture

In this section, we will lay out the overall system architecture for Modin, as well as go into detail about the component design, implementation and other important details. This document also contains important reference information for those interested in contributing new functionality, bugfixes and enhancements.

12.2.1 High-Level Architectural View

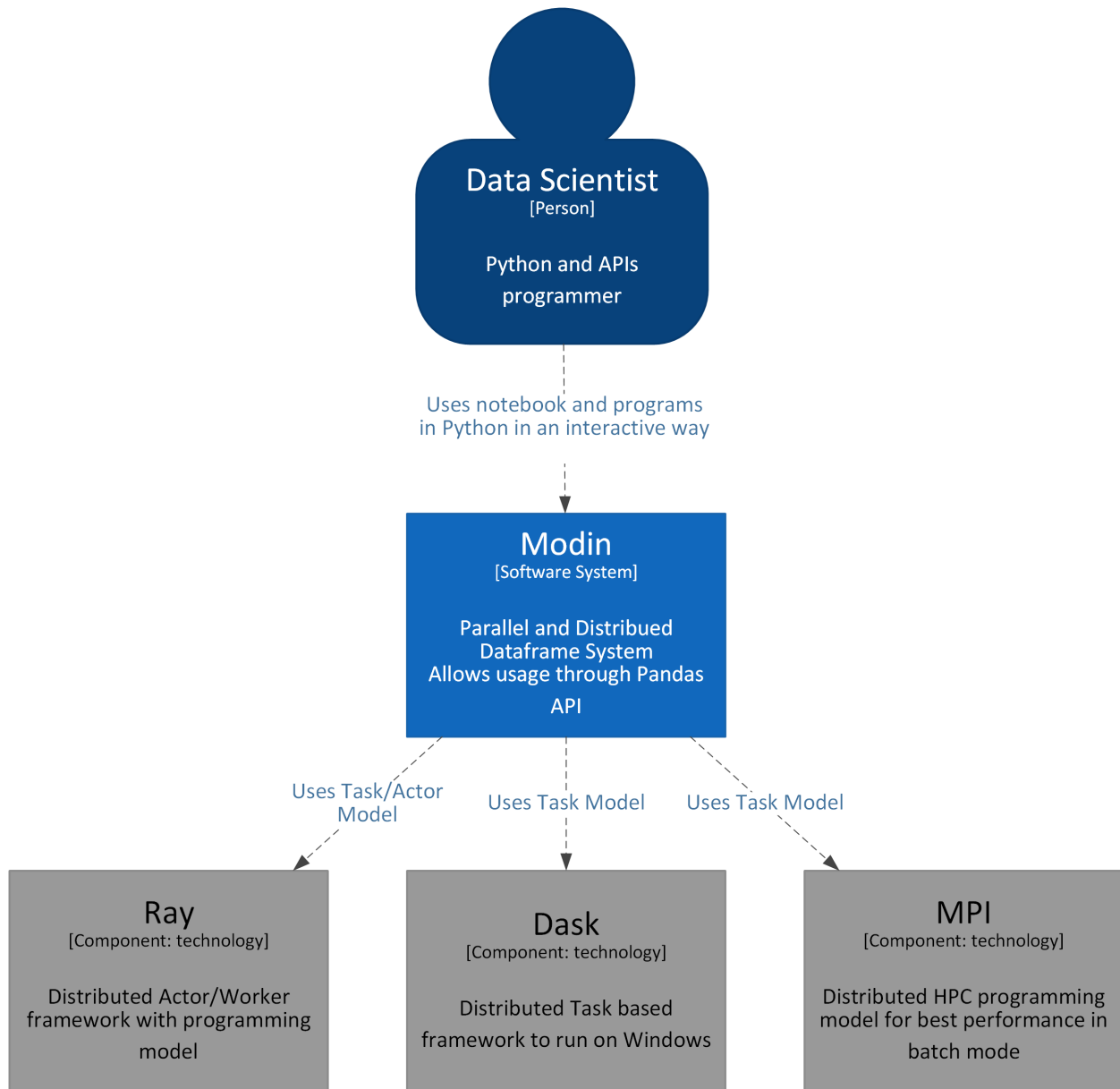
The diagram below outlines the general layered view to the components of Modin with a short description of each major section of the documentation following.



Modin is logically separated into different layers that represent the hierarchy of a typical Database Management System. Abstracting out each component allows us to individually optimize and swap out components without affecting the rest of the system. We can implement, for example, new compute kernels that are optimized for a certain type of data and can simply plug it in to the existing infrastructure by implementing a small interface. It can still be distributed by our choice of compute engine with the logic internally.

12.2.2 System View

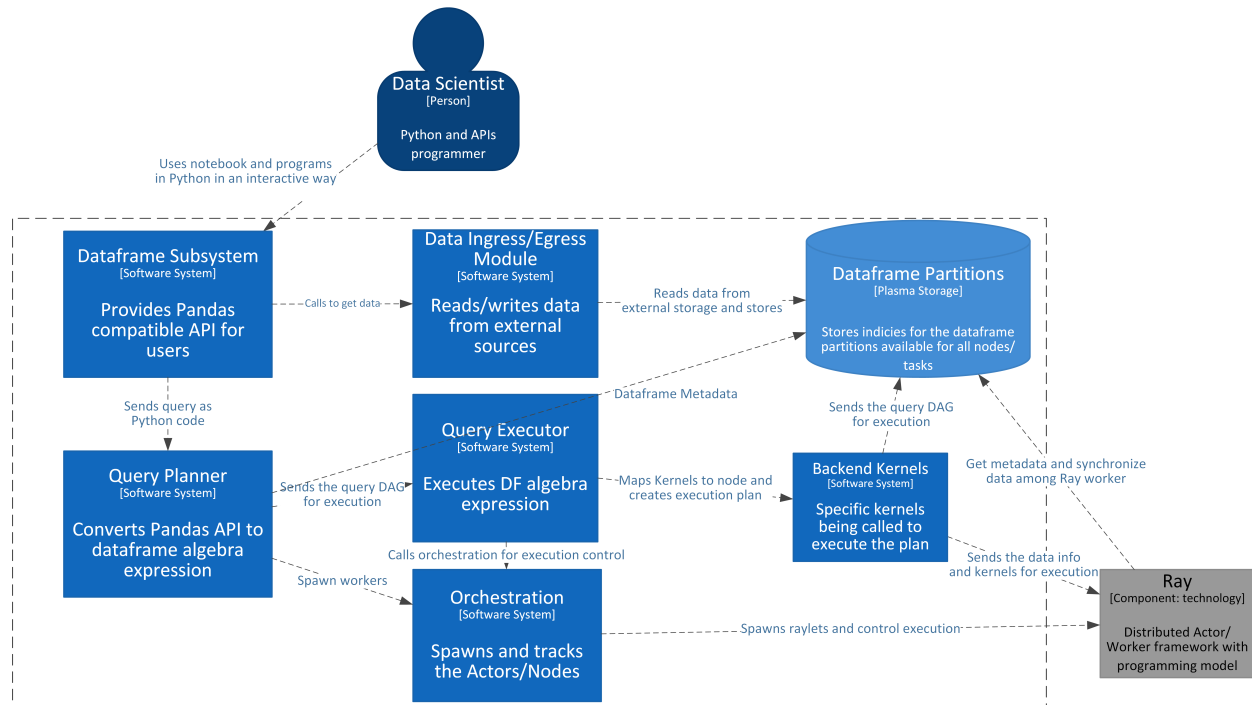
A top-down view of Modin’s architecture is detailed below:



The user - Data Scientist interacts with the Modin system by sending interactive or batch commands through API and Modin executes them using various execution engines: Ray, Dask and MPI are currently supported.

12.2.3 Subsystem/Container View

If we click down to the next level of details we will see that inside Modin the layered architecture is implemented using several interacting components:



For the simplicity the other execution systems - Dask and MPI are omitted and only Ray execution is shown.

- Dataframe subsystem is the backbone of the dataframe holding and query compilation. It is responsible for dispatching the ingress/egress to the appropriate module, getting the pandas API and calling the query compiler to convert calls to the internal intermediate Dataframe Algebra.
- Data Ingress/Egress Module is working in conjunction with Dataframe and Partitions subsystem to read data split into partitions and send data into the appropriate node for storing.
- Query Planner is subsystem that translates the pandas API to intermediate Dataframe Algebra representation DAG and performs an initial set of optimizations.
- Query Executor is responsible for getting the Dataframe Algebra DAG, performing further optimizations based on a selected storage format and mapping or compiling the Dataframe Algebra DAG to and actual execution sequence.
- Storage formats module is responsible for mapping the abstract operation to an actual executor call, e.g. pandas, PyArrow, custom format.
- Orchestration subsystem is responsible for spawning and controlling the actual execution environment for the selected execution. It spawns the actual nodes, fires up the execution environment, e.g. Ray, monitors the state of executors and provides telemetry

12.2.4 Component View

User queries which perform data transformation, data ingress or data egress pass through the Modin components detailed below. The path the query takes is mostly similar across execution systems, with some minor exceptions like OmnisciOnNative.

Data Transformation

Query Compiler

The Query Compiler receives queries from the pandas API layer. The API layer is responsible for ensuring a clean input to the Query Compiler. The Query Compiler must have knowledge of the compute kernels and in-memory format of the data in order to efficiently compile the query.

The Query Compiler is responsible for sending the compiled query to the Core Modin Dataframe. In this design, the Query Compiler does not have information about where or when the query will be executed, and gives the control of the partition layout to the Modin Dataframe.

In the interest of reducing the pandas API, the Query Compiler layer closely follows the pandas API, but cuts out a large majority of the repetition.

Core Modin Dataframe

At this layer, operations can be performed lazily. Currently, Modin executes most operations eagerly in an attempt to behave as pandas does. Some operations, e.g. `transpose` are expensive and create full copies of the data in-memory. In these cases, we can wait until another operation triggers computation. In the future, we plan to add additional query planning and laziness to Modin to ensure that queries are performed efficiently.

The structure of the Core Modin Dataframe is extensible, such that any operation that could be better optimized for a given execution can be overridden and optimized in that way.

This layer has a significantly reduced API from the QueryCompiler and the user-facing API. Each of these APIs represents a single way of performing a given operation or behavior.

Core Modin Dataframe API

More documentation can be found internally in the [code](#). This API is not complete, but represents an overwhelming majority of operations and behaviors.

This API can be implemented by other distributed/parallel DataFrame libraries and plugged in to Modin as well. Create an [issue](#) or discuss on our [Discourse](#) or [Slack](#) for more information!

The Core Modin Dataframe is responsible for the data layout and shuffling, partitioning, and serializing the tasks that get sent to each partition. Other implementations of the Modin Dataframe interface will have to handle these as well.

Partition Manager

The Partition Manager can change the size and shape of the partitions based on the type of operation. For example, certain operations are complex and require access to an entire column or row. The Partition Manager can convert the block partitions to row partitions or column partitions. This gives Modin the flexibility to perform operations that are difficult in row-only or column-only partitioning schemas.

Another important component of the Partition Manager is the serialization and shipment of compiled queries to the Partitions. It maintains metadata for the length and width of each partition, so when operations only need to operate on or extract a subset of the data, it can ship those queries directly to the correct partition. This is particularly important for some operations in pandas which can accept different arguments and operations for different columns, e.g. `fillna` with a dictionary.

This abstraction separates the actual data movement and function application from the Dataframe layer to keep the Core Dataframe API small and separately optimize the data movement and metadata management.

Partitions

Partitions are responsible for managing a subset of the Dataframe. As mentioned below, the Dataframe is partitioned both row and column-wise. This gives Modin scalability in both directions and flexibility in data layout. There are a number of optimizations in Modin that are implemented in the partitions. Partitions are specific to the execution framework and in-memory format of the data, allowing Modin to exploit potential optimizations across both. These optimizations are explained further on the pages specific to the execution framework.

Execution Engine

This layer performs computation on partitions of the data. The Modin Dataframe is designed to work with [task parallel](#) frameworks, but integration with data parallel frameworks should be possible with some effort.

Storage Format

The storage format describes the in-memory partition type. The base storage format in Modin is pandas. In the default case, the Modin Dataframe operates on partitions that contain `pandas.DataFrame` objects.

Data Ingress

Note: Data ingress operations (e.g. `read_csv`) in Modin load data from the source into partitions and vice versa for data egress (e.g. `to_csv`) operation. Improved performance is achieved by reading/writing in partitions in parallel.

Data ingress starts with a function in the pandas API layer (e.g. `read_csv`). Then the user's query is passed to the Factory Dispatcher, which defines a factory specific for the execution. The factory for execution contains an IO class (e.g. `PandasOnRayIO`) whose responsibility is to perform a parallel read/write from/to a file. This IO class contains class methods with interfaces and names that are similar to pandas IO functions (e.g. `PandasOnRayIO.read_csv`). The IO class declares the Modin Dataframe and Query Compiler classes specific for the execution engine and storage format to ensure the correct object is constructed. It also declares IO methods that are mix-ins containing a combination of the engine-specific class for deploying remote tasks, the class for parsing the given file format and the class handling the chunking of the format-specific file on the head node (see dispatcher classes implementation details). The output from the IO class data ingress function is a Modin Dataframe.

Data Egress

Data egress operations (e.g. `to_csv`) are similar to data ingress operations up to execution-specific IO class functions construction. Data egress functions of the IO class are defined slightly different from data ingress functions and created only specifically for the engine since partitions already have information about its storage format. Using the IO class, data is exported from partitions to the target file.

Supported Execution Engines and Storage Formats

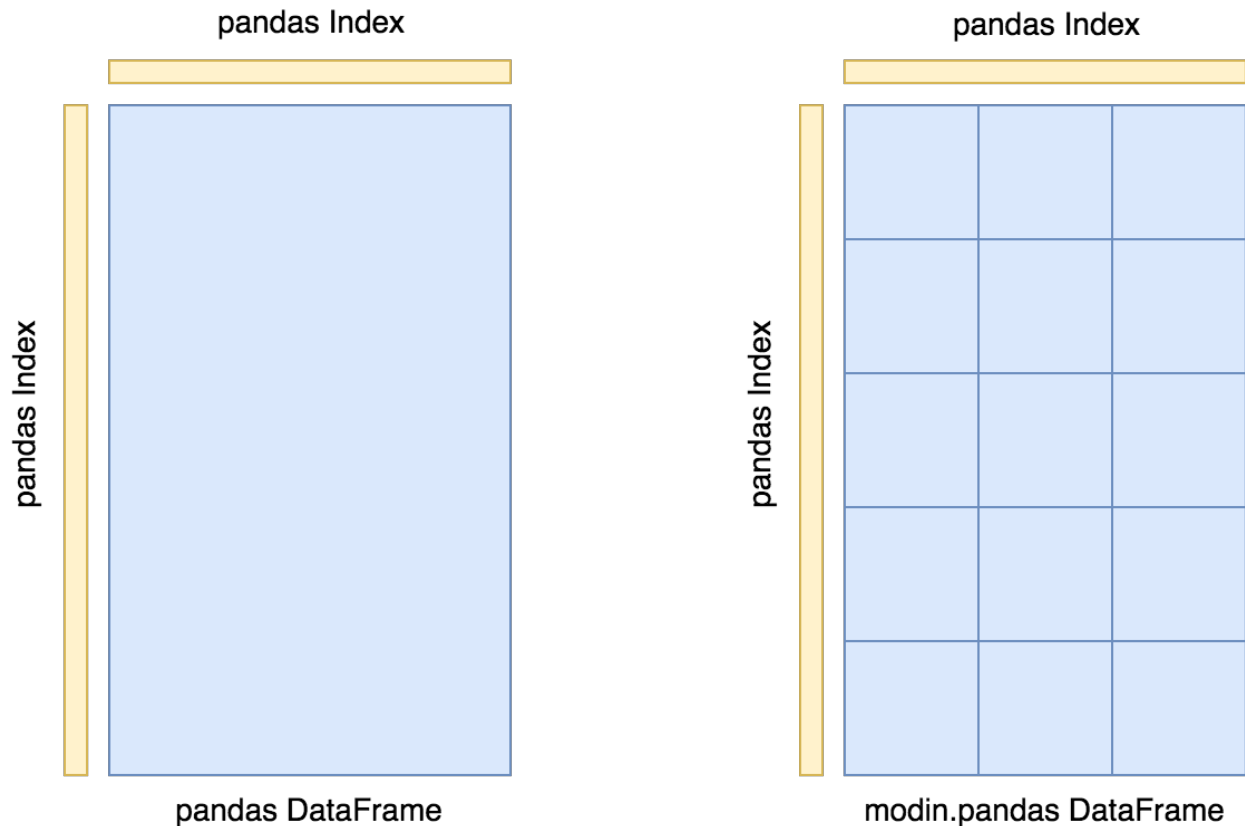
This is a list of execution engines and in-memory formats supported in Modin. If you would like to contribute a new execution engine or in-memory format, please see the documentation page on [contributing](#).

- ***pandas on Ray***
 - Uses the [Ray](#) execution framework.
 - The storage format is *pandas* and the in-memory partition type is a pandas DataFrame.
 - For more information on the execution path, see the [pandas on Ray](#) page.
- ***pandas on Dask***
 - Uses the [Dask Futures](#) execution framework.
 - The storage format is *pandas* and the in-memory partition type is a pandas DataFrame.
 - For more information on the execution path, see the [pandas on Dask](#) page.
- ***pandas on Python***
 - Uses native python execution - mainly used for debugging.
 - The storage format is *pandas* and the in-memory partition type is a pandas DataFrame.
 - For more information on the execution path, see the [pandas on Python](#) page.
- **pandas on Ray (experimental)**
 - Uses the [Ray](#) execution framework.
 - The storage format is *pandas* and the in-memory partition type is a pandas DataFrame.
 - For more information on the execution path, see the [experimental pandas on Ray](#) page.
- ***OmniSci on Native* (experimental)**
 - Uses OmniSciDB as an engine.
 - The storage format is *omnisci* and the in-memory partition type is a pyarrow Table. When defaulting to pandas, the pandas DataFrame is used.
 - For more information on the execution path, see the [OmniSci on Native](#) page.
- ***Pyarrow on Ray* (experimental)**
 - Uses the [Ray](#) execution framework.
 - The storage format is *pyarrow* and the in-memory partition type is a pyarrow Table.
 - For more information on the execution path, see the [Pyarrow on Ray](#) page.
- **cuDF on Ray (experimental)**
 - Uses the [Ray](#) execution framework.

- The storage format is *cudf* and the in-memory partition type is a cuDF DataFrame.
- For more information on the execution path, see the cuDF on Ray page.

12.2.5 DataFrame Partitioning

The Modin DataFrame architecture follows in the footsteps of modern architectures for database and high performance matrix systems. We chose a partitioning schema that partitions along both columns and rows because it gives Modin flexibility and scalability in both the number of columns and the number of rows. The following figure illustrates this concept.



Currently, the main in-memory format of each partition is a [pandas DataFrame](#) (pandas storage format). Omnisci, PyArrow and cuDF are also supported as experimental in-memory formats in Modin.

12.2.6 Index

We currently use the `pandas.Index` object for indexing both columns and rows. In the future, we will implement a distributed, pandas-compatible Index object in order to remove this scaling limitation from the system. Most workloads will not be affected by this scalability limit since it only appears when operating on more than 10's of billions of columns or rows. **Important note:** If you are using the default index (`pandas.RangeIndex`) there is a fixed memory overhead (~200 bytes) and there will be no scalability issues with the index.

12.2.7 API

The API is the outer-most layer that faces users. The following classes contain Modin's implementation of the pandas API:

Base pandas Dataset API

The class implements functionality that is common to Modin's pandas API for both `DataFrame` and `Series` classes.

Public API

`class modin.pandas.base.BasePandasDataset`

Implement most of the common code that exists in `DataFrame`/`Series`.

Since both objects share the same underlying representation, and the algorithms are the same, we use this object to define the general behavior of those objects and then use those objects to define the output type.

Notes

See pandas API documentation for [pandas.DataFrame](#), [pandas.Series](#) for more.

`abs()`

Return a *BasePandasDataset* with absolute numeric value of each element.

Notes

See pandas API documentation for [pandas.DataFrame.abs](#), [pandas.Series.abs](#) for more.

`add(other, axis='columns', level=None, fill_value=None)`

Return addition of *BasePandasDataset* and *other*, element-wise (binary operator *add*).

Notes

See pandas API documentation for [pandas.DataFrame.add](#), [pandas.Series.add](#) for more.

`agg(func=None, axis=0, *args, **kwargs)`

Aggregate using one or more operations over the specified axis.

Notes

See pandas API documentation for [pandas.DataFrame.aggregate](#), [pandas.Series.aggregate](#) for more.

`aggregate(func=None, axis=0, *args, **kwargs)`

Aggregate using one or more operations over the specified axis.

Notes

See pandas API documentation for [pandas.DataFrame.aggregate](#), [pandas.Series.aggregate](#) for more.

align(*other*, *join*='outer', *axis*=None, *level*=None, *copy*=True, *fill_value*=None, *method*=None, *limit*=None, *fill_axis*=0, *broadcast_axis*=None)

Align two objects on their axes with the specified join method.

Notes

See pandas API documentation for [pandas.DataFrame.align](#), [pandas.Series.align](#) for more.

all(*axis*=0, *bool_only*=None, *skipna*=True, *level*=None, ***kwargs*)

Return whether all elements are True, potentially over an axis.

Notes

See pandas API documentation for [pandas.DataFrame.all](#), [pandas.Series.all](#) for more.

any(*axis*=0, *bool_only*=None, *skipna*=True, *level*=None, ***kwargs*)

Return whether any element is True, potentially over an axis.

Notes

See pandas API documentation for [pandas.DataFrame.any](#), [pandas.Series.any](#) for more.

apply(*func*, *axis*=0, *broadcast*=None, *raw*=False, *reduce*=None, *result_type*=None, *convert_dtype*=True, *args*=(), ***kws*)

Apply a function along an axis of the *BasePandasDataset*.

Notes

See pandas API documentation for [pandas.DataFrame.apply](#), [pandas.Series.apply](#) for more.

asfreq(*freq*, *method*=None, *how*=None, *normalize*=False, *fill_value*=None)

Convert time series to specified frequency.

Notes

See pandas API documentation for [pandas.DataFrame.asfreq](#), [pandas.Series.asfreq](#) for more.

asof(*where*, *subset*=None)

Return the last row(s) without any NaNs before *where*.

Notes

See pandas API documentation for [pandas.DataFrame.asof](#), [pandas.Series.asof](#) for more.

astype(*dtype*, *copy=True*, *errors='raise'*)

Cast a Modin object to a specified dtype *dtype*.

Notes

See pandas API documentation for [pandas.DataFrame.astype](#), [pandas.Series.astype](#) for more.

property at

Get a single value for a row/column label pair.

Notes

See pandas API documentation for [pandas.DataFrame.at](#), [pandas.Series.at](#) for more.

at_time(*time*, *asof=False*, *axis=None*)

Select values at particular time of day (e.g., 9:30AM).

Notes

See pandas API documentation for [pandas.DataFrame.at_time](#), [pandas.Series.at_time](#) for more.

backfill(*axis=None*, *inplace=False*, *limit=None*, *downcast=None*)

Synonym for *DataFrame.fillna* with *method='bfill'*.

Notes

See pandas API documentation for [pandas.DataFrame.backfill](#), [pandas.Series.backfill](#) for more.

between_time(*start_time*, *end_time*, *include_start: bool_t | NoDefault = NoDefault.no_default*, *include_end: bool_t | NoDefault = NoDefault.no_default*, *inclusive: str | None = None*, *axis=None*)

Select values between particular times of the day (e.g., 9:00-9:30 AM).

Notes

See pandas API documentation for [pandas.DataFrame.between_time](#), [pandas.Series.between_time](#) for more.

bfill(*axis=None*, *inplace=False*, *limit=None*, *downcast=None*)

Synonym for *DataFrame.fillna* with *method='bfill'*.

Notes

See pandas API documentation for [pandas.DataFrame.backfill](#), [pandas.Series.backfill](#) for more.

bool()

Return the bool of a single element *BasePandasDataset*.

Notes

See pandas API documentation for [pandas.DataFrame.bool](#), [pandas.Series.bool](#) for more.

clip(*lower=None, upper=None, axis=None, inplace=False, *args, **kwargs*)

Trim values at input threshold(s).

combine(*other, func, fill_value=None, **kwargs*)

Perform combination of *BasePandasDataset*-s according to *func*.

Notes

See pandas API documentation for [pandas.DataFrame.combine](#), [pandas.Series.combine](#) for more.

combine_first(*other*)

Update null elements with value in the same location in *other*.

Notes

See pandas API documentation for [pandas.DataFrame.combine_first](#), [pandas.Series.combine_first](#) for more.

convert_dtypes(*infer_objects: modin.pandas.base.BasePandasDataset.bool = True, convert_string: modin.pandas.base.BasePandasDataset.bool = True, convert_integer: modin.pandas.base.BasePandasDataset.bool = True, convert_boolean: modin.pandas.base.BasePandasDataset.bool = True, convert_floating: modin.pandas.base.BasePandasDataset.bool = True*)

Convert columns to best possible dtypes using dtypes supporting `pd.NA`.

Notes

See pandas API documentation for [pandas.DataFrame.convert_dtypes](#), [pandas.Series.convert_dtypes](#) for more.

copy(*deep=True*)

Make a copy of the object's metadata.

Notes

See pandas API documentation for [pandas.DataFrame.copy](#), [pandas.Series.copy](#) for more.

count(*axis=0, level=None, numeric_only=False*)

Count non-NA cells for *BasePandasDataset*.

Notes

See pandas API documentation for [pandas.DataFrame.count](#), [pandas.Series.count](#) for more.

cummax(*axis=None, skipna=True, *args, **kwargs*)

Return cumulative maximum over a *BasePandasDataset* axis.

Notes

See pandas API documentation for [pandas.DataFrame.cummax](#), [pandas.Series.cummax](#) for more.

cummin(*axis=None, skipna=True, *args, **kwargs*)

Return cumulative minimum over a *BasePandasDataset* axis.

Notes

See pandas API documentation for [pandas.DataFrame.cummin](#), [pandas.Series.cummin](#) for more.

cumprod(*axis=None, skipna=True, *args, **kwargs*)

Return cumulative product over a *BasePandasDataset* axis.

Notes

See pandas API documentation for [pandas.DataFrame.cumprod](#), [pandas.Series.cumprod](#) for more.

cumsum(*axis=None, skipna=True, *args, **kwargs*)

Return cumulative sum over a *BasePandasDataset* axis.

Notes

See pandas API documentation for [pandas.DataFrame.cumsum](#), [pandas.Series.cumsum](#) for more.

describe(*percentiles=None, include=None, exclude=None, datetime_is_numeric=False*)

Generate descriptive statistics.

Notes

See pandas API documentation for [pandas.DataFrame.describe](#), [pandas.Series.describe](#) for more.

diff(*periods=1, axis=0*)

First discrete difference of element.

Notes

See pandas API documentation for [pandas.DataFrame.diff](#), [pandas.Series.diff](#) for more.

div(*other*, *axis*='columns', *level*=None, *fill_value*=None)

Get floating division of *BasePandasDataset* and *other*, element-wise (binary operator *truediv*).

Notes

See pandas API documentation for [pandas.DataFrame.truediv](#), [pandas.Series.truediv](#) for more.

divide(*other*, *axis*='columns', *level*=None, *fill_value*=None)

Get floating division of *BasePandasDataset* and *other*, element-wise (binary operator *truediv*).

Notes

See pandas API documentation for [pandas.DataFrame.truediv](#), [pandas.Series.truediv](#) for more.

drop(*labels*=None, *axis*=0, *index*=None, *columns*=None, *level*=None, *inplace*=False, *errors*='raise')

Drop specified labels from *BasePandasDataset*.

Notes

See pandas API documentation for [pandas.DataFrame.drop](#), [pandas.Series.drop](#) for more.

drop_duplicates(*keep*='first', *inplace*=False, ***kwargs*)

Return *BasePandasDataset* with duplicate rows removed.

Notes

See pandas API documentation for [pandas.DataFrame.drop_duplicates](#), [pandas.Series.drop_duplicates](#) for more.

droplevel(*level*, *axis*=0)

Return *BasePandasDataset* with requested index / column level(s) removed.

Notes

See pandas API documentation for [pandas.DataFrame.droplevel](#), [pandas.Series.droplevel](#) for more.

dropna(*axis*=0, *how*='any', *thresh*=None, *subset*=None, *inplace*=False)

Remove missing values.

Notes

See pandas API documentation for [pandas.DataFrame.dropna](#), [pandas.Series.dropna](#) for more.

eq(*other*, *axis*='columns', *level*=None)

Get equality of *BasePandasDataset* and *other*, element-wise (binary operator *eq*).

Notes

See pandas API documentation for [pandas.DataFrame.eq](#), [pandas.Series.eq](#) for more.

ewm(*com*: float | None = None, *span*: float | None = None, *halflife*: float | TimedeltaConvertibleTypes | None = None, *alpha*: float | None = None, *min_periods*: int | None = 0, *adjust*: bool_t = True, *ignore_na*: bool_t = False, *axis*: Axis = 0, *times*: str | np.ndarray | BasePandasDataset | None = None, *method*: str = 'single') → ExponentialMovingWindow

Provide exponentially weighted (EW) calculations.

Notes

See pandas API documentation for [pandas.DataFrame.ewm](#), [pandas.Series.ewm](#) for more.

expanding(*min_periods*=1, *center*=None, *axis*=0, *method*='single')

Provide expanding window calculations.

Notes

See pandas API documentation for [pandas.DataFrame.expanding](#), [pandas.Series.expanding](#) for more.

explode(*column*, *ignore_index*: modin.pandas.base.BasePandasDataset.bool = False)

Transform each element of a list-like to a row.

Notes

See pandas API documentation for [pandas.DataFrame.explode](#), [pandas.Series.explode](#) for more.

ffill(*axis*=None, *inplace*=False, *limit*=None, *downcast*=None)

Synonym for *DataFrame.fillna* with *method*='ffill'.

Notes

See pandas API documentation for [pandas.DataFrame.pad](#), [pandas.Series.pad](#) for more.

filter(*items*=None, *like*=None, *regex*=None, *axis*=None)

Subset the *BasePandasDataset* rows or columns according to the specified index labels.

Notes

See pandas API documentation for [pandas.DataFrame.filter](#), [pandas.Series.filter](#) for more.

first(*offset*)

Select initial periods of time series data based on a date offset.

Notes

See pandas API documentation for [pandas.DataFrame.first](#), [pandas.Series.first](#) for more.

first_valid_index()

Return index for first non-NA value or None, if no non-NA value is found.

Notes

See pandas API documentation for [pandas.DataFrame.first_valid_index](#), [pandas.Series.first_valid_index](#) for more.

property flags

Get the properties associated with this *BasePandasDataset*.

Notes

See pandas API documentation for [pandas.DataFrame.flags](#), [pandas.Series.flags](#) for more.

floordiv(*other*, *axis*='columns', *level*=None, *fill_value*=None)

Get integer division of *BasePandasDataset* and *other*, element-wise (binary operator *floordiv*).

Notes

See pandas API documentation for [pandas.DataFrame.floordiv](#), [pandas.Series.floordiv](#) for more.

ge(*other*, *axis*='columns', *level*=None)

Get greater than or equal comparison of *BasePandasDataset* and *other*, element-wise (binary operator *ge*).

Notes

See pandas API documentation for [pandas.DataFrame.ge](#), [pandas.Series.ge](#) for more.

get(*key*, *default*=None)

Get item from object for given key.

Notes

See pandas API documentation for [pandas.DataFrame.get](#), [pandas.Series.get](#) for more.

gt(*other*, *axis*='columns', *level*=None)

Get greater than comparison of *BasePandasDataset* and *other*, element-wise (binary operator *gt*).

Notes

See pandas API documentation for [pandas.DataFrame.gt](#), [pandas.Series.gt](#) for more.

head(*n*=5)

Return the first *n* rows.

Notes

See pandas API documentation for [pandas.DataFrame.head](#), [pandas.Series.head](#) for more.

property iat

Get a single value for a row/column pair by integer position.

Notes

See pandas API documentation for [pandas.DataFrame.iat](#), [pandas.Series.iat](#) for more.

idxmax(*axis*=0, *skipna*=True)

Return index of first occurrence of maximum over requested axis.

Notes

See pandas API documentation for [pandas.DataFrame.idxmax](#), [pandas.Series.idxmax](#) for more.

idxmin(*axis*=0, *skipna*=True)

Return index of first occurrence of minimum over requested axis.

Notes

See pandas API documentation for [pandas.DataFrame.idxmin](#), [pandas.Series.idxmin](#) for more.

property iloc

Purely integer-location based indexing for selection by position.

Notes

See pandas API documentation for [pandas.DataFrame.iloc](#), [pandas.Series.iloc](#) for more.

property index

Get the index for this DataFrame.

Returns The union of all indexes across the partitions.

Return type pandas.Index

infer_objects()

Attempt to infer better dtypes for object columns.

Notes

See pandas API documentation for [pandas.DataFrame.infer_objects](#), [pandas.Series.infer_objects](#) for more.

isin(values)

Whether elements in *BasePandasDataset* are contained in *values*.

Notes

See pandas API documentation for [pandas.DataFrame.isin](#), [pandas.Series.isin](#) for more.

isna()

Detect missing values.

Notes

See pandas API documentation for [pandas.DataFrame.isna](#), [pandas.Series.isna](#) for more.

isnull()

Detect missing values.

Notes

See pandas API documentation for [pandas.DataFrame.isna](#), [pandas.Series.isna](#) for more.

kurt(*axis*: *Axis* | *None* | *NoDefault* = *NoDefault.no_default*, *skipna*=*True*, *level*=*None*, *numeric_only*=*None*, ***kwargs*)

Return unbiased kurtosis over requested axis.

Notes

See pandas API documentation for [pandas.DataFrame.kurt](#), [pandas.Series.kurt](#) for more.

kurtosis(*axis*: *Axis* | *None* | *NoDefault* = *NoDefault.no_default*, *skipna*=*True*, *level*=*None*, *numeric_only*=*None*, ***kwargs*)

Return unbiased kurtosis over requested axis.

Notes

See pandas API documentation for [pandas.DataFrame.kurt](#), [pandas.Series.kurt](#) for more.

last(offset)

Select final periods of time series data based on a date offset.

Notes

See pandas API documentation for [pandas.DataFrame.last](#), [pandas.Series.last](#) for more.

last_valid_index()

Return index for last non-NA value or None, if no non-NA value is found.

Notes

See pandas API documentation for [pandas.DataFrame.last_valid_index](#), [pandas.Series.last_valid_index](#) for more.

le(*other*, axis='columns', level=None)

Get less than or equal comparison of *BasePandasDataset* and *other*, element-wise (binary operator *le*).

Notes

See pandas API documentation for [pandas.DataFrame.le](#), [pandas.Series.le](#) for more.

property loc

Get a group of rows and columns by label(s) or a boolean array.

Notes

See pandas API documentation for [pandas.DataFrame.loc](#), [pandas.Series.loc](#) for more.

lt(*other*, axis='columns', level=None)

Get less than comparison of *BasePandasDataset* and *other*, element-wise (binary operator *lt*).

Notes

See pandas API documentation for [pandas.DataFrame.lt](#), [pandas.Series.lt](#) for more.

mad(*axis=None*, skipna=True, level=None)

Return the mean absolute deviation of the values over the requested axis.

Notes

See pandas API documentation for [pandas.DataFrame.mad](#), [pandas.Series.mad](#) for more.

mask(*cond*, *other=nan*, *inplace=False*, *axis=None*, *level=None*, *errors='raise'*, *try_cast=NoDefault.no_default*)

Replace values where the condition is True.

max(*axis: int | None | NoDefault = NoDefault.no_default*, *skipna=True*, *level=None*, *numeric_only=None*, *kwargs*)**

Return the maximum of the values over the requested axis.

Notes

See pandas API documentation for [pandas.DataFrame.max](#), [pandas.Series.max](#) for more.

mean(*axis*: int | None | NoDefault = NoDefault.no_default, *skipna*=True, *level*=None, *numeric_only*=None, ***kwargs*)

Return the mean of the values over the requested axis.

Notes

See pandas API documentation for [pandas.DataFrame.mean](#), [pandas.Series.mean](#) for more.

median(*axis*: int | None | NoDefault = NoDefault.no_default, *skipna*=True, *level*=None, *numeric_only*=None, ***kwargs*)

Return the median of the values over the requested axis.

Notes

See pandas API documentation for [pandas.DataFrame.median](#), [pandas.Series.median](#) for more.

memory_usage(*index*=True, *deep*=False)

Return the memory usage of the *BasePandasDataset*.

Notes

See pandas API documentation for [pandas.DataFrame.memory_usage](#), [pandas.Series.memory_usage](#) for more.

min(*axis*: int | None | NoDefault = NoDefault.no_default, *skipna*=True, *level*=None, *numeric_only*=None, ***kwargs*)

Return the minimum of the values over the requested axis.

Notes

See pandas API documentation for [pandas.DataFrame.min](#), [pandas.Series.min](#) for more.

mod(*other*, *axis*='columns', *level*=None, *fill_value*=None)

Get modulo of *BasePandasDataset* and *other*, element-wise (binary operator *mod*).

Notes

See pandas API documentation for [pandas.DataFrame.mod](#), [pandas.Series.mod](#) for more.

mode(*axis*=0, *numeric_only*=False, *dropna*=True)

Get the mode(s) of each element along the selected axis.

Notes

See pandas API documentation for [pandas.DataFrame.mode](#), [pandas.Series.mode](#) for more.

mul(*other*, *axis*='columns', *level*=None, *fill_value*=None)

Get multiplication of *BasePandasDataset* and *other*, element-wise (binary operator *mul*).

Notes

See pandas API documentation for [pandas.DataFrame.mul](#), [pandas.Series.mul](#) for more.

multiply(*other*, *axis*='columns', *level*=None, *fill_value*=None)

Get multiplication of *BasePandasDataset* and *other*, element-wise (binary operator *mul*).

Notes

See pandas API documentation for [pandas.DataFrame.mul](#), [pandas.Series.mul](#) for more.

ne(*other*, *axis*='columns', *level*=None)

Get Not equal comparison of *BasePandasDataset* and *other*, element-wise (binary operator *ne*).

Notes

See pandas API documentation for [pandas.DataFrame.ne](#), [pandas.Series.ne](#) for more.

notna()

Detect existing (non-missing) values.

Notes

See pandas API documentation for [pandas.DataFrame.notna](#), [pandas.Series.notna](#) for more.

notnull()

Detect existing (non-missing) values.

Notes

See pandas API documentation for [pandas.DataFrame.notna](#), [pandas.Series.notna](#) for more.

nunique(*axis*=0, *dropna*=True)

Return number of unique elements in the *BasePandasDataset*.

Notes

See pandas API documentation for [pandas.DataFrame.nunique](#), [pandas.Series.nunique](#) for more.

pad(*axis*=None, *inplace*=False, *limit*=None, *downcast*=None)

Synonym for *DataFrame.fillna* with *method*='ffill'.

Notes

See pandas API documentation for [pandas.DataFrame.pad](#), [pandas.Series.pad](#) for more.

pct_change(*periods=1*, *fill_method='pad'*, *limit=None*, *freq=None*, ***kwargs*)

Percentage change between the current and a prior element.

Notes

See pandas API documentation for [pandas.DataFrame.pct_change](#), [pandas.Series.pct_change](#) for more.

pipe(*func*, **args*, ***kwargs*)

Apply chainable functions that expect *BasePandasDataset*.

Notes

See pandas API documentation for [pandas.DataFrame.pipe](#), [pandas.Series.pipe](#) for more.

pop(*item*)

Return item and drop from frame. Raise *KeyError* if not found.

Notes

See pandas API documentation for [pandas.DataFrame.pop](#), [pandas.Series.pop](#) for more.

pow(*other*, *axis='columns'*, *level=None*, *fill_value=None*)

Get exponential power of *BasePandasDataset* and *other*, element-wise (binary operator *pow*).

Notes

See pandas API documentation for [pandas.DataFrame.pow](#), [pandas.Series.pow](#) for more.

quantile(*q=0.5*, *axis=0*, *numeric_only=True*, *interpolation='linear'*)

Return values at the given quantile over requested axis.

Notes

See pandas API documentation for [pandas.DataFrame.quantile](#), [pandas.Series.quantile](#) for more.

radd(*other*, *axis='columns'*, *level=None*, *fill_value=None*)

Return addition of *BasePandasDataset* and *other*, element-wise (binary operator *add*).

Notes

See pandas API documentation for [pandas.DataFrame.add](#), [pandas.Series.add](#) for more.

rank(*axis=0*, *method: str = 'average'*, *numeric_only: bool_t | None | NoDefault = NoDefault.no_default*, *na_option: str = 'keep'*, *ascending: bool_t = True*, *pct: bool_t = False*)

Compute numerical data ranks (1 through n) along axis.

Notes

See pandas API documentation for [pandas.DataFrame.rank](#), [pandas.Series.rank](#) for more.

rdiv(*other*, *axis*='columns', *level*=None, *fill_value*=None)

Get floating division of *BasePandasDataset* and *other*, element-wise (binary operator *rtruediv*).

Notes

See pandas API documentation for [pandas.DataFrame.rtruediv](#), [pandas.Series.rtruediv](#) for more.

reindex(*index*=None, *columns*=None, *copy*=True, ***kwargs*)

Conform *BasePandasDataset* to new index with optional filling logic.

Notes

See pandas API documentation for [pandas.DataFrame.reindex](#), [pandas.Series.reindex](#) for more.

reindex_like(*other*, *method*=None, *copy*=True, *limit*=None, *tolerance*=None)

Return an object with matching indices as *other* object.

Notes

See pandas API documentation for [pandas.DataFrame.reindex_like](#), [pandas.Series.reindex_like](#) for more.

rename_axis(*mapper*=None, *index*=None, *columns*=None, *axis*=None, *copy*=True, *inplace*=False)

Set the name of the axis for the index or columns.

Notes

See pandas API documentation for [pandas.DataFrame.rename_axis](#), [pandas.Series.rename_axis](#) for more.

reorder_levels(*order*, *axis*=0)

Rearrange index levels using input order.

Notes

See pandas API documentation for [pandas.DataFrame.reorder_levels](#), [pandas.Series.reorder_levels](#) for more.

resample(*rule*, *axis*=0, *closed*=None, *label*=None, *convention*='start', *kind*=None, *loffset*=None, *base*:
Optional[int] = None, *on*=None, *level*=None, *origin*: Union[str, Timestamp, datetime.datetime,
numpy.datetime64, int, numpy.int64, float] = 'start_day', *offset*: Optional[Union[Timedelta,
datetime.timedelta, numpy.timedelta64, int, numpy.int64, float, str]] = None)

Resample time-series data.

Notes

See pandas API documentation for [pandas.DataFrame.resample](#), [pandas.Series.resample](#) for more.

reset_index(*level=None, drop=False, inplace=False, col_level=0, col_fill=""*)

Reset the index, or a level of it.

Notes

See pandas API documentation for [pandas.DataFrame.reset_index](#), [pandas.Series.reset_index](#) for more.

rfloordiv(*other, axis='columns', level=None, fill_value=None*)

Get integer division of *BasePandasDataset* and *other*, element-wise (binary operator *rfloordiv*).

Notes

See pandas API documentation for [pandas.DataFrame.rfloordiv](#), [pandas.Series.rfloordiv](#) for more.

rmod(*other, axis='columns', level=None, fill_value=None*)

Get modulo of *BasePandasDataset* and *other*, element-wise (binary operator *rmod*).

Notes

See pandas API documentation for [pandas.DataFrame.rmod](#), [pandas.Series.rmod](#) for more.

rmul(*other, axis='columns', level=None, fill_value=None*)

Get multiplication of *BasePandasDataset* and *other*, element-wise (binary operator *mul*).

Notes

See pandas API documentation for [pandas.DataFrame.mul](#), [pandas.Series.mul](#) for more.

rolling(*window, min_periods=None, center=False, win_type=None, on=None, axis=0, closed=None, method='single'*)

Provide rolling window calculations.

Notes

See pandas API documentation for [pandas.DataFrame.rolling](#), [pandas.Series.rolling](#) for more.

round(*decimals=0, *args, **kwargs*)

Round a *BasePandasDataset* to a variable number of decimal places.

Notes

See pandas API documentation for [pandas.DataFrame.round](#), [pandas.Series.round](#) for more.

rpow(*other*, *axis*='columns', *level*=None, *fill_value*=None)

Get exponential power of *BasePandasDataset* and *other*, element-wise (binary operator *rpow*).

Notes

See pandas API documentation for [pandas.DataFrame.rpow](#), [pandas.Series.rpow](#) for more.

rsub(*other*, *axis*='columns', *level*=None, *fill_value*=None)

Get subtraction of *BasePandasDataset* and *other*, element-wise (binary operator *rsub*).

Notes

See pandas API documentation for [pandas.DataFrame.rsub](#), [pandas.Series.rsub](#) for more.

rtruediv(*other*, *axis*='columns', *level*=None, *fill_value*=None)

Get floating division of *BasePandasDataset* and *other*, element-wise (binary operator *rtruediv*).

Notes

See pandas API documentation for [pandas.DataFrame.rtruediv](#), [pandas.Series.rtruediv](#) for more.

sample(*n*=None, *frac*=None, *replace*=False, *weights*=None, *random_state*=None, *axis*=None, *ignore_index*=False)

Return a random sample of items from an axis of object.

Notes

See pandas API documentation for [pandas.DataFrame.sample](#), [pandas.Series.sample](#) for more.

sem(*axis*=None, *skipna*=True, *level*=None, *ddof*=1, *numeric_only*=None, ***kwargs*)

Return unbiased standard error of the mean over requested axis.

Notes

See pandas API documentation for [pandas.DataFrame.sem](#), [pandas.Series.sem](#) for more.

set_axis(*labels*, *axis*=0, *inplace*=False)

Assign desired index to given axis.

Notes

See pandas API documentation for [pandas.DataFrame.set_axis](#), [pandas.Series.set_axis](#) for more.

set_flags(**copy*: *modin.pandas.base.BasePandasDataset.bool = False*, *allows_duplicate_labels*:
Optional[modin.pandas.base.BasePandasDataset.bool] = None)

Return a new *BasePandasDataset* with updated flags.

Notes

See pandas API documentation for [pandas.DataFrame.set_flags](#), [pandas.Series.set_flags](#) for more.

shift(*periods=1*, *freq=None*, *axis=0*, *fill_value=NoDefault.no_default*)

Shift index by desired number of periods with an optional time *freq*.

Notes

See pandas API documentation for [pandas.DataFrame.shift](#), [pandas.Series.shift](#) for more.

property size

Return an int representing the number of elements in this *BasePandasDataset* object.

Notes

See pandas API documentation for [pandas.DataFrame.size](#), [pandas.Series.size](#) for more.

skew(*axis: int | None | NoDefault = NoDefault.no_default*, *skipna=True*, *level=None*, *numeric_only=None*,
***kwargs*)

Return unbiased skew over requested axis.

Notes

See pandas API documentation for [pandas.DataFrame.skew](#), [pandas.Series.skew](#) for more.

sort_index(*axis=0*, *level=None*, *ascending=True*, *inplace=False*, *kind='quicksort'*, *na_position='last'*,
sort_remaining=True, *ignore_index: modin.pandas.base.BasePandasDataset.bool = False*, *key*:
Optional[Callable[[Index], Union[Index, ExtensionArray, numpy.ndarray, Series]]] = None)

Sort object by labels (along an axis).

Notes

See pandas API documentation for [pandas.DataFrame.sort_index](#), [pandas.Series.sort_index](#) for more.

sort_values(*by*, *axis=0*, *ascending=True*, *inplace: modin.pandas.base.BasePandasDataset.bool = False*,
kind='quicksort', *na_position='last'*, *ignore_index*:
modin.pandas.base.BasePandasDataset.bool = False, *key: Optional[Callable[[Index],*
Union[Index, ExtensionArray, numpy.ndarray, Series]]] = None)

Sort by the values along either axis.

Notes

See pandas API documentation for [pandas.DataFrame.sort_values](#), [pandas.Series.sort_values](#) for more.

std(*axis=None*, *skipna=True*, *level=None*, *ddof=1*, *numeric_only=None*, ***kwargs*)

Return sample standard deviation over requested axis.

Notes

See pandas API documentation for [pandas.DataFrame.std](#), [pandas.Series.std](#) for more.

sub(*other*, *axis='columns'*, *level=None*, *fill_value=None*)

Get subtraction of *BasePandasDataset* and *other*, element-wise (binary operator *sub*).

Notes

See pandas API documentation for [pandas.DataFrame.sub](#), [pandas.Series.sub](#) for more.

subtract(*other*, *axis='columns'*, *level=None*, *fill_value=None*)

Get subtraction of *BasePandasDataset* and *other*, element-wise (binary operator *sub*).

Notes

See pandas API documentation for [pandas.DataFrame.sub](#), [pandas.Series.sub](#) for more.

swapaxes(*axis1*, *axis2*, *copy=True*)

Interchange axes and swap values axes appropriately.

Notes

See pandas API documentation for [pandas.DataFrame.swapaxes](#), [pandas.Series.swapaxes](#) for more.

swaplevel(*i=-2*, *j=-1*, *axis=0*)

Swap levels *i* and *j* in a *MultiIndex*.

Notes

See pandas API documentation for [pandas.DataFrame.swaplevel](#), [pandas.Series.swaplevel](#) for more.

tail(*n=5*)

Return the last *n* rows.

Notes

See pandas API documentation for [pandas.DataFrame.tail](#), [pandas.Series.tail](#) for more.

take(*indices*, *axis=0*, *is_copy=None*, ***kwargs*)

Return the elements in the given *positional* indices along an axis.

Notes

See pandas API documentation for [pandas.DataFrame.take](#), [pandas.Series.take](#) for more.

to_clipboard(*excel=True, sep=None, **kwargs*)

Copy object to the system clipboard.

Notes

See pandas API documentation for [pandas.DataFrame.to_clipboard](#), [pandas.Series.to_clipboard](#) for more.

to_csv(*path_or_buf=None, sep=',', na_rep="", float_format=None, columns=None, header=True, index=True, index_label=None, mode='w', encoding=None, compression='infer', quoting=None, quotechar='"', line_terminator=None, chunksize=None, date_format=None, doublequote=True, escapechar=None, decimal='.', errors: str = 'strict', storage_options: Optional[Dict[str, Any]] = None*)

Write object to a comma-separated values (csv) file.

Notes

See pandas API documentation for [pandas.DataFrame.to_csv](#), [pandas.Series.to_csv](#) for more.

to_dict(*orient='dict', into=<class 'dict'>*)

Convert the *BasePandasDataset* to a dictionary.

Notes

See pandas API documentation for [pandas.DataFrame.to_dict](#), [pandas.Series.to_dict](#) for more.

to_excel(*excel_writer, sheet_name='Sheet1', na_rep="", float_format=None, columns=None, header=True, index=True, index_label=None, startrow=0, startcol=0, engine=None, merge_cells=True, encoding=None, inf_rep='inf', verbose=True, freeze_panes=None, storage_options: Optional[Dict[str, Any]] = None*)

Write object to an Excel sheet.

Notes

See pandas API documentation for [pandas.DataFrame.to_excel](#), [pandas.Series.to_excel](#) for more.

to_hdf(*path_or_buf, key, format='table', **kwargs*)

Write the contained data to an HDF5 file using HDFStore.

Notes

See pandas API documentation for [pandas.DataFrame.to_hdf](#), [pandas.Series.to_hdf](#) for more.

to_json(*path_or_buf=None, orient=None, date_format=None, double_precision=10, force_ascii=True, date_unit='ms', default_handler=None, lines=False, compression='infer', index=True, indent=None, storage_options: Optional[Dict[str, Any]] = None*)

Convert the object to a JSON string.

Notes

See pandas API documentation for [pandas.DataFrame.to_json](#), [pandas.Series.to_json](#) for more.

to_latex(*buf=None, columns=None, col_space=None, header=True, index=True, na_rep='NaN', formatters=None, float_format=None, sparsify=None, index_names=True, bold_rows=False, column_format=None, longtable=None, escape=None, encoding=None, decimal='.', multicolumn=None, multicolumn_format=None, multirow=None, caption=None, label=None, position=None*)

Render object to a LaTeX tabular, longtable, or nested table.

Notes

See pandas API documentation for [pandas.DataFrame.to_latex](#), [pandas.Series.to_latex](#) for more.

to_markdown(*buf=None, mode: str = 'wt', index: modin.pandas.base.BasePandasDataset.bool = True, storage_options: Optional[Dict[str, Any]] = None, **kwargs*)

Print *BasePandasDataset* in Markdown-friendly format.

Notes

See pandas API documentation for [pandas.DataFrame.to_markdown](#), [pandas.Series.to_markdown](#) for more.

to_numpy(*dtype=None, copy=False, na_value=NoDefault.no_default*)

Convert the *BasePandasDataset* to a NumPy array.

Notes

See pandas API documentation for [pandas.DataFrame.to_numpy](#), [pandas.Series.to_numpy](#) for more.

to_period(*freq=None, axis=0, copy=True*)

Convert *BasePandasDataset* from DatetimeIndex to PeriodIndex.

Notes

See pandas API documentation for [pandas.DataFrame.to_period](#), [pandas.Series.to_period](#) for more.

to_pickle(*path, compression: Optional[Union[Literal['infer', 'gzip', 'bz2', 'zip', 'xz', 'zstd'], Dict[str, Any]]] = 'infer', protocol: int = 5, storage_options: Optional[Dict[str, Any]] = None*)

Pickle (serialize) object to file.

Notes

See pandas API documentation for [pandas.DataFrame.to_pickle](#), [pandas.Series.to_pickle](#) for more.

to_sql(*name, con, schema=None, if_exists='fail', index=True, index_label=None, chunksize=None, dtype=None, method=None*)

Write records stored in a *BasePandasDataset* to a SQL database.

Notes

See pandas API documentation for [pandas.DataFrame.to_sql](#), [pandas.Series.to_sql](#) for more.

to_string(buf=None, columns=None, col_space=None, header=True, index=True, na_rep='NaN',
formatters=None, float_format=None, sparsify=None, index_names=True, justify=None,
max_rows=None, min_rows=None, max_cols=None, show_dimensions=False, decimal='.',
line_width=None, max_colwidth=None, encoding=None)

Render a *BasePandasDataset* to a console-friendly tabular output.

Notes

See pandas API documentation for [pandas.DataFrame.to_string](#), [pandas.Series.to_string](#) for more.

to_timestamp(freq=None, how='start', axis=0, copy=True)

Cast to DatetimeIndex of timestamps, at *beginning* of period.

Notes

See pandas API documentation for [pandas.DataFrame.to_timestamp](#), [pandas.Series.to_timestamp](#) for more.

to_xarray()

Return an xarray object from the *BasePandasDataset*.

Notes

See pandas API documentation for [pandas.DataFrame.to_xarray](#), [pandas.Series.to_xarray](#) for more.

transform(func, axis=0, *args, **kwargs)

Call func on self producing a *BasePandasDataset* with the same axis shape as self.

Notes

See pandas API documentation for [pandas.DataFrame.transform](#), [pandas.Series.transform](#) for more.

truediv(other, axis='columns', level=None, fill_value=None)

Get floating division of *BasePandasDataset* and *other*, element-wise (binary operator *truediv*).

Notes

See pandas API documentation for [pandas.DataFrame.truediv](#), [pandas.Series.truediv](#) for more.

truncate(before=None, after=None, axis=None, copy=True)

Truncate a *BasePandasDataset* before and after some index value.

Notes

See pandas API documentation for [pandas.DataFrame.truncate](#), [pandas.Series.truncate](#) for more.

tshift(*periods=1, freq=None, axis=0*)

Shift the time index, using the index's frequency if available.

Notes

See pandas API documentation for [pandas.DataFrame.tshift](#), [pandas.Series.tshift](#) for more.

tz_convert(*tz, axis=0, level=None, copy=True*)

Convert tz-aware axis to target time zone.

Notes

See pandas API documentation for [pandas.DataFrame.tz_convert](#), [pandas.Series.tz_convert](#) for more.

tz_localize(*tz, axis=0, level=None, copy=True, ambiguous='raise', nonexistent='raise'*)

Localize tz-naive index of a *BasePandasDataset* to target time zone.

Notes

See pandas API documentation for [pandas.DataFrame.tz_localize](#), [pandas.Series.tz_localize](#) for more.

value_counts(*subset: Optional[Sequence[Hashable]] = None, normalize: modin.pandas.base.BasePandasDataset.bool = False, sort: modin.pandas.base.BasePandasDataset.bool = True, ascending: modin.pandas.base.BasePandasDataset.bool = False, dropna: modin.pandas.base.BasePandasDataset.bool = True*)

Count unique values in the *BasePandasDataset*.

Notes

See pandas API documentation for [pandas.DataFrame.value_counts](#), [pandas.Series.value_counts](#) for more.

property values

Return a NumPy representation of the *BasePandasDataset*.

Notes

See pandas API documentation for [pandas.DataFrame.values](#), [pandas.Series.values](#) for more.

var(*axis=None, skipna=True, level=None, ddof=1, numeric_only=None, **kwargs*)

Return unbiased variance over requested axis.

Notes

See pandas API documentation for [pandas.DataFrame.var](#), [pandas.Series.var](#) for more.

DataFrame Module Overview

Modin's pandas.DataFrame API

Modin's pandas.DataFrame API is backed by a distributed object providing an identical API to pandas. After the user calls some DataFrame function, this call is internally rewritten into a representation that can be processed in parallel by the partitions. These results can be e.g., reduced to single output, identical to the single threaded pandas DataFrame method output.

Public API

```
class modin.pandas.dataframe.DataFrame(data=None, index=None, columns=None, dtype=None,  
copy=None, query_compiler=None)
```

Modin distributed representation of pandas.DataFrame.

Internally, the data can be divided into partitions along both columns and rows in order to parallelize computations and utilize the user's hardware as much as possible.

Inherit common for DataFrame-s and Series functionality from the *BasePandasDataset* class.

Parameters

- **data** (*DataFrame, Series, pandas.DataFrame, ndarray, Iterable or dict, optional*) – Dict can contain Series, arrays, constants, dataclass or list-like objects. If data is a dict, column order follows insertion-order.
- **index** (*Index or array-like, optional*) – Index to use for resulting frame. Will default to RangeIndex if no indexing information part of input data and no index provided.
- **columns** (*Index or array-like, optional*) – Column labels to use for resulting frame. Will default to RangeIndex if no column labels are provided.
- **dtype** (*str, np.dtype, or pandas.ExtensionDtype, optional*) – Data type to force. Only a single dtype is allowed. If None, infer.
- **copy** (*bool, default: False*) – Copy data from inputs. Only affects pandas.DataFrame / 2d ndarray input.
- **query_compiler** (*BaseQueryCompiler, optional*) – A query compiler object to create the DataFrame from.

Notes

See pandas API documentation for [pandas.DataFrame](#) for more. DataFrame can be created either from passed *data* or *query_compiler*. If both parameters are provided, data source will be prioritized in the next order:

- 1) Modin DataFrame or Series passed with *data* parameter.
- 2) Query compiler from the *query_compiler* parameter.
- 3) Various pandas/NumPy/Python data structures passed with *data* parameter.

The last option is less desirable since import of such data structures is very inefficient, please use previously created Modin structures from the first two options or import data using highly efficient Modin IO tools (for example `pd.read_csv`).

Usage Guide

The most efficient way to create Modin DataFrame is to import data from external storage using the highly efficient Modin IO methods (for example using `pd.read_csv`, see details for Modin IO methods in the separate section), but even if the data does not originate from a file, any pandas supported data type or `pandas.DataFrame` can be used. Internally, the DataFrame data is divided into partitions, which number along an axis usually corresponds to the number of the user's hardware CPUs. If needed, the number of partitions can be changed by setting `modin.config.NPartitions`.

Let's consider simple example of creation and interacting with Modin DataFrame:

```
import modin.config

# This explicitly sets the number of partitions
modin.config.NPartitions.put(4)

import modin.pandas as pd
import pandas

# Create Modin DataFrame from the external file
pd_dataframe = pd.read_csv("test_data.csv")
# Create Modin DataFrame from the python object
# data = {'fcol{x}': [fcol{x}_{y}' for y in range(100, 356)] for x in range(4)}
# pd_dataframe = pd.DataFrame(data)
# Create Modin DataFrame from the pandas object
# pd_dataframe = pd.DataFrame(pandas.DataFrame(data))

# Show created DataFrame
print(pd_dataframe)

# List DataFrame partitions. Note, that internal API is intended for
# developers needs and was used here for presentation purposes
# only.
partitions = pd_dataframe._query_compiler._modin_frame._partitions
print(partitions)

# Show the first DataFrame partition
print(partitions[0][0].get())
```

Output:

```
# created DataFrame
```

	col0	col1	col2	col3
0	col0_100	col1_100	col2_100	col3_100
1	col0_101	col1_101	col2_101	col3_101
2	col0_102	col1_102	col2_102	col3_102
3	col0_103	col1_103	col2_103	col3_103
4	col0_104	col1_104	col2_104	col3_104

(continues on next page)

(continued from previous page)

```

..      ...      ...      ...      ...
251  col0_351  col1_351  col2_351  col3_351
252  col0_352  col1_352  col2_352  col3_352
253  col0_353  col1_353  col2_353  col3_353
254  col0_354  col1_354  col2_354  col3_354
255  col0_355  col1_355  col2_355  col3_355

[256 rows x 4 columns]

# List of DataFrame partitions

[<modin.core.execution.ray.implementations.pandas_on_ray.partitioning.partition.
↪PandasOnRayDataframePartition object at 0x7fc554e607f0>]
[<modin.core.execution.ray.implementations.pandas_on_ray.partitioning.partition.
↪PandasOnRayDataframePartition object at 0x7fc554e9a4f0>]
[<modin.core.execution.ray.implementations.pandas_on_ray.partitioning.partition.
↪PandasOnRayDataframePartition object at 0x7fc554e60820>]
[<modin.core.execution.ray.implementations.pandas_on_ray.partitioning.partition.
↪PandasOnRayDataframePartition object at 0x7fc554e609d0>]]

# The first DataFrame partition

      col0      col1      col2      col3
0  col0_100  col1_100  col2_100  col3_100
1  col0_101  col1_101  col2_101  col3_101
2  col0_102  col1_102  col2_102  col3_102
3  col0_103  col1_103  col2_103  col3_103
4  col0_104  col1_104  col2_104  col3_104
..      ...      ...      ...      ...
60  col0_160  col1_160  col2_160  col3_160
61  col0_161  col1_161  col2_161  col3_161
62  col0_162  col1_162  col2_162  col3_162
63  col0_163  col1_163  col2_163  col3_163
64  col0_164  col1_164  col2_164  col3_164

[65 rows x 4 columns]

```

As we show in the example above, Modin DataFrame can be easily created, and supports any input that pandas DataFrame supports. Also note that tuning of the DataFrame partitioning can be done by just setting a single config.

Series Module Overview

Modin's pandas.Series API

Modin's pandas.Series API is backed by a distributed object providing an identical API to pandas. After the user calls some Series function, this call is internally rewritten into a representation that can be processed in parallel by the partitions. These results can be e.g., reduced to single output, identical to the single threaded pandas Series method output.

Public API

```
class modin.pandas.series.Series(data=None, index=None, dtype=None, name=None, copy=False,
                                  fastpath=False, query_compiler=None)
```

Modin distributed representation of *pandas.Series*.

Internally, the data can be divided into partitions in order to parallelize computations and utilize the user's hardware as much as possible.

Inherit common for DataFrames and Series functionality from the *BasePandasDataset* class.

Parameters

- **data** (*modin.pandas.Series*, array-like, *Iterable*, *dict*, or scalar value, optional) – Contains data stored in Series. If data is a dict, argument order is maintained.
- **index** (array-like or *Index (1d)*, optional) – Values must be hashable and have the same length as *data*.
- **dtype** (*str*, *np.dtype*, or *pandas.ExtensionDtype*, optional) – Data type for the output Series. If not specified, this will be inferred from *data*.
- **name** (*str*, optional) – The name to give to the Series.
- **copy** (*bool*, default: *False*) – Copy input data.
- **fastpath** (*bool*, default: *False*) – *pandas* internal parameter.
- **query_compiler** (*BaseQueryCompiler*, optional) – A query compiler object to create the Series from.

Notes

See pandas API documentation for [pandas.Series](#) for more.

Usage Guide

The most efficient way to create Modin Series is to import data from external storage using the highly efficient Modin IO methods (for example using `pd.read_csv`, see details for Modin IO methods in the separate section), but even if the data does not originate from a file, any pandas supported data type or `pandas.Series` can be used. Internally, the Series data is divided into partitions, which number along an axis usually corresponds to the number of the user's hardware CPUs. If needed, the number of partitions can be changed by setting `modin.config.NPartitions`.

Let's consider simple example of creation and interacting with Modin Series:

```
import modin.config

# This explicitly sets the number of partitions
modin.config.NPartitions.put(4)

import modin.pandas as pd
import pandas

# Create Modin Series from the external file
pd_series = pd.read_csv("test_data.csv", header=None).squeeze()
```

(continues on next page)

(continued from previous page)

```

# Create Modin Series from the python object
# pd_series = pd.Series([x for x in range(256)])
# Create Modin Series from the pandas object
# pd_series = pd.Series(pandas.Series([x for x in range(256)]))

# Show created `Series`
print(pd_series)

# List `Series` partitions. Note, that internal API is intended for
# developers needs and was used here for presentation purposes
# only.
partitions = pd_series._query_compiler._modin_frame._partitions
print(partitions)

# Show the first `Series` partition
print(partitions[0][0].get())

```

Output:

```

# created `Series`

0      100
1      101
2      102
3      103
4      104
...
251    351
252    352
253    353
254    354
255    355
Name: 0, Length: 256, dtype: int64

# List of `Series` partitions

[<modin.core.execution.ray.implementations.pandas_on_ray.partitioning.partition.
↳PandasOnRayDataframePartition object at 0x7fc554e607f0>]
[<modin.core.execution.ray.implementations.pandas_on_ray.partitioning.partition.
↳PandasOnRayDataframePartition object at 0x7fc554e9a4f0>]
[<modin.core.execution.ray.implementations.pandas_on_ray.partitioning.partition.
↳PandasOnRayDataframePartition object at 0x7fc554e60820>]
[<modin.core.execution.ray.implementations.pandas_on_ray.partitioning.partition.
↳PandasOnRayDataframePartition object at 0x7fc554e609d0>]]

# The first `Series` partition

0
0    100
1    101
2    102
3    103

```

(continues on next page)

(continued from previous page)

```

4    104
..    ..
60   160
61   161
62   162
63   163
64   164

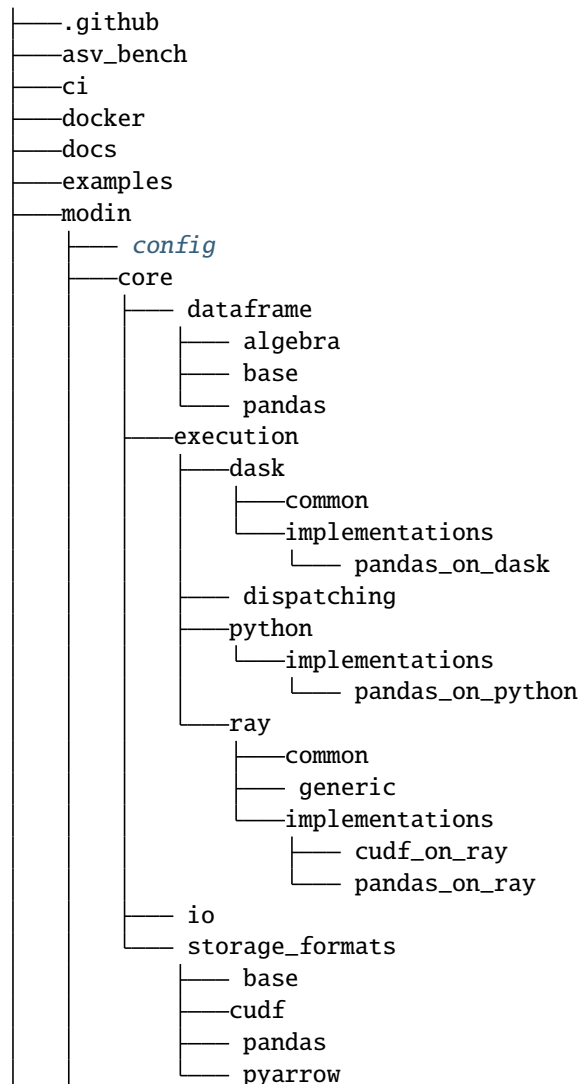
```

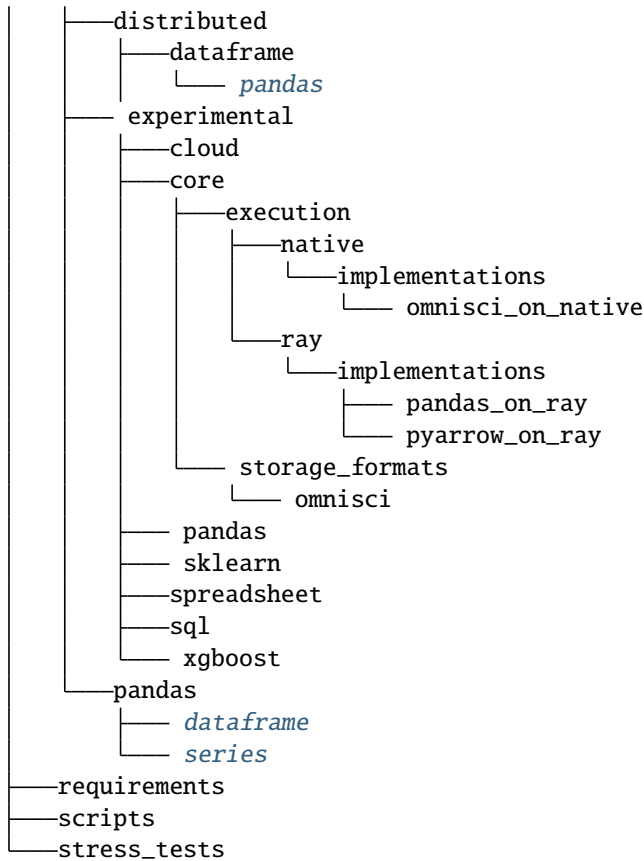
```
[65 rows x 1 columns]
```

As we show in the example above, Modin Series can be easily created, and supports any input that pandas Series supports. Also note that tuning of the Series partitioning can be done by just setting a single config.

12.2.8 Module/Class View

Modin's modules layout is shown below. Click on the links to deep dive into Modin's internal implementation details. The documentation covers most modules, with more docs being added everyday!





12.3 Partition API in Modin

When you are working with a [DataFrame](#), you can unwrap its remote partitions to get the raw futures objects compatible with the execution engine (e.g. `ray.ObjectRef` for Ray). In addition to unwrapping of the remote partitions we also provide an API to construct a `modin.pandas.DataFrame` from raw futures objects.

12.3.1 Partition IPs

For finer grained placement control, Modin also provides an API to get the IP addresses of the nodes that hold each partition. You can pass the partitions having needed IPs to your function. It can help with minimizing of data movement between nodes.

12.3.2 Partition API implementations

By default, a [DataFrame](#) stores underlying partitions as `pandas.DataFrame` objects. You can find the specific implementation of Modin's Partition Interface in [pandas Partition API](#).

12.3.3 Ray engine

However, it is worth noting that for Modin on Ray engine with pandas in-memory format IPs of the remote partitions may not match actual locations if the partitions are lower than 100 kB. Ray saves such objects (≤ 100 kB, by default) in in-process store of the calling process (please, refer to [Ray documentation](#) for more information). We can't get IPs for such objects while maintaining good performance. So, you should keep in mind this for unwrapping of the remote partitions with their IPs. Several options are provided to handle the case in [How to handle Ray objects that are lower 100 kB](#) section.

12.3.4 Dask engine

There is no mentioned above issue for Modin on Dask engine with pandas in-memory format because Dask saves any objects in the worker process that processes a function (please, refer to [Dask documentation](#) for more information).

12.3.5 How to handle Ray objects that are lower than 100 kB

- If you are sure that each of the remote partitions being unwrapped is higher than 100 kB, you can just import Modin or perform `ray.init()` manually.
- If you don't know partition sizes you can pass the option `_system_config={"max_direct_call_object_size": <nbytes>, }`, where `nbytes` is threshold for objects that will be stored in in-process store, to `ray.init()`.
- You can also start Ray as follows: `ray start --head --system-config='{"max_direct_call_object_size":<nbytes>}'`

Note that when specifying the threshold the performance of some Modin operations may change.

12.4 pandas on Ray

This section describes usage related documents for the pandas on Ray component of Modin.

Modin uses pandas as a primary memory format of the underlying partitions and optimizes queries ingested from the API layer in a specific way to this format. Thus, there is no need to care of choosing it but you can explicitly specify it anyway as shown below.

One of the execution engines that Modin uses is Ray. If you have Ray installed in your system, Modin also uses it by default to distribute computations.

If you want to be explicit, you could set the following environment variables:

```
export MODIN_ENGINE=ray
export MODIN_STORAGE_FORMAT=pandas
```

or turn it on in source code:

```
import modin.config as cfg
cfg.Engine.put('ray')
cfg.StorageFormat.put('pandas')
```


12.5 pandas on Dask

This section describes usage related documents for the pandas on Dask component of Modin.

Modin uses pandas as a primary memory format of the underlying partitions and optimizes queries ingested from the API layer in a specific way to this format. Thus, there is no need to care of choosing it but you can explicitly specify it anyway as shown below.

One of the execution engines that Modin uses is Dask. To enable the pandas on Dask execution you should set the following environment variables:

```
export MODIN_ENGINE=dask
export MODIN_STORAGE_FORMAT=pandas
```

or turn them on in source code:

```
import modin.config as cfg
cfg.Engine.put('dask')
cfg.StorageFormat.put('pandas')
```

12.6 pandas on Python

This section describes usage related documents for the pandas on Python component of Modin.

Modin uses pandas as the primary memory format of the underlying partitions and optimizes queries from the API layer in a specific way to this format. Since it is a default, you do not need to specify the pandas memory format, but we show how to explicitly set it below.

One of the execution engines that Modin uses is Python. This engine is sequential and used for debugging. To enable the pandas on Python execution you should set the following environment variables:

```
export MODIN_ENGINE=python
export MODIN_STORAGE_FORMAT=pandas
```

or turn a debug mode on:

```
export MODIN_DEBUG=True
export MODIN_STORAGE_FORMAT=pandas
```

or do the same in source code:

```
import modin.config as cfg
cfg.Engine.put('python')
cfg.StorageFormat.put('pandas')
```

```
import modin.config as cfg
cfg.IsDebug.put(True)
cfg.StorageFormat.put('pandas')
```

12.7 OmniSci

This section describes usage related documents for the OmniSciDB-based engine of Modin.

This engine uses analytical database [OmniSciDB](#) to obtain high single-node scalability for specific set of dataframe operations. To enable this engine you can set the following environment variable:

```
export MODIN_STORAGE_FORMAT=omnisci
```

or use it in your code:

```
import modin.config as cfg
cfg.StorageFormat.put('omnisci')
```

Since OmniSci is run through its native engine, Modin automatically sets `MODIN_ENGINE=Native` and you might not specify it explicitly. If for some reasons Native engine is explicitly set using `modin.config` or `MODIN_ENGINE` environment variable, make sure you also tell Modin that Experimental mode is turned on (`export MODIN_EXPERIMENTAL=true` or `cfg.IsExperimental.put(True)`) otherwise the following error occurs:

```
FactoryNotFoundError: Omnisci on Native is only accessible through the experimental API.
Run `import modin.experimental.pandas as pd` to use Omnisci on Native.
```

Note: If you encounter `LLVM ERROR: inconsistency in registered CommandLine options` error when using OmniSci, please refer to the respective section in [Troubleshooting](#) page to avoid the issue.

12.8 PyArrow on Ray

Coming Soon!

12.9 Modin SQL API

Modin's SQL API is currently a conceptual plan, Coming Soon!

12.9.1 Plans for future development

Our plans with the SQL API for Modin are to create an interface that allows you to intermix SQL and pandas operations without copying the entire dataset into a new structure between the two. This is possible due to the architecture of Modin. Currently, Modin has a query compiler that acts as an intermediate layer between the query language (e.g. SQL, pandas) and the execution (See [architecture](#) documentation for details).

We have implemented a simple example that can be found below. Feedback welcome!

```
>>> import modin.sql as sql
>>>
>>> conn = sql.connect("db_name")
>>> c = conn.cursor()
>>> c.execute("CREATE TABLE example (col1, col2, column 3, col4)")
>>> c.execute("INSERT INTO example VALUES ('1', 2.0, 'A String of information', True)")
```

(continues on next page)

(continued from previous page)

```
col1  col2                column 3  col4
0    1    2.0  A String of information  True

>>> c.execute("INSERT INTO example VALUES ('6', 17.0, 'A String of different information
↳', False)")
col1  col2                column 3  col4
0    1    2.0                A String of information  True
1    6   17.0  A String of different information  False
```


13.1 Slack

Join our [Slack](#) community to connect with Modin users and contributors, discuss, and ask questions about all things Modin-related.

13.2 Discussion forum

Check out our [discussion forum](#) to discuss release announcements, general questions, issues, use-cases, and tutorials.

13.3 Mailing List

General questions, potential contributors, and ideas can be directed to the [developer mailing list](#). It is an open Google Group, so feel free to join anytime! If you are unsure about where to ask or post something, the mailing list is a good place to ask as well.

13.4 Issues

Bug reports and feature requests can be directed to the [issues](#) page of the Modin GitHub repo.

SCALE YOUR PANDAS WORKFLOW BY CHANGING A SINGLE LINE OF CODE

Modin uses [Ray](#) or [Dask](#) to provide an effortless way to speed up your pandas notebooks, scripts, and libraries. Unlike other distributed DataFrame libraries, Modin provides seamless integration and compatibility with existing pandas code. Even using the DataFrame constructor is identical.

```
import modin.pandas as pd
import numpy as np

frame_data = np.random.randint(0, 100, size=(2**10, 2**8))
df = pd.DataFrame(frame_data)
```

To use Modin, you do not need to know how many cores your system has and you do not need to specify how to distribute the data. In fact, you can continue using your previous pandas notebooks while experiencing a considerable speedup from Modin, even on a single machine. Once you've changed your import statement, you're ready to use Modin just like you would pandas.

INSTALLATION AND CHOOSING YOUR COMPUTE ENGINE

Modin can be installed from PyPI:

```
pip install modin
```

If you don't have [Ray](#) or [Dask](#) installed, you will need to install Modin with one of the targets:

```
pip install "modin[ray]" # Install Modin dependencies and Ray to run on Ray
pip install "modin[dask]" # Install Modin dependencies and Dask to run on Dask
pip install "modin[all]" # Install all of the above
```

Modin will automatically detect which engine you have installed and use that for scheduling computation!

If you want to choose a specific compute engine to run on, you can set the environment variable `MODIN_ENGINE` and Modin will do computation with that engine:

```
export MODIN_ENGINE=ray # Modin will use Ray
export MODIN_ENGINE=dask # Modin will use Dask
```

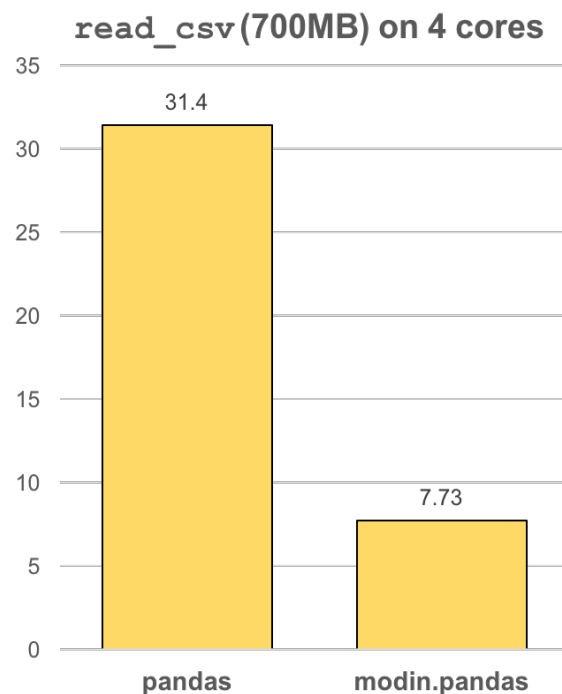
This can also be done within a notebook/interpreter before you import Modin:

```
import os

os.environ["MODIN_ENGINE"] = "ray" # Modin will use Ray
os.environ["MODIN_ENGINE"] = "dask" # Modin will use Dask

import modin.pandas as pd
```


FASTER PANDAS, EVEN ON YOUR LAPTOP



The `modin.pandas DataFrame` is an extremely light-weight parallel DataFrame. Modin transparently distributes the data and computation so that all you need to do is continue using the pandas API as you were before installing Modin. Unlike other parallel DataFrame systems, Modin is an extremely light-weight, robust DataFrame. Because it is so light-weight, Modin provides speed-ups of up to 4x on a laptop with 4 physical cores.

In pandas, you are only able to use one core at a time when you are doing computation of any kind. With Modin, you are able to use all of the CPU cores on your machine. Even in `read_csv`, we see large gains by efficiently distributing the work across your entire machine.

```
import modin.pandas as pd

df = pd.read_csv("my_dataset.csv")
```


MODIN IS A DATAFRAME FOR DATASETS FROM 1MB TO 1TB+

We have focused heavily on bridging the solutions between DataFrames for small data (e.g. pandas) and large data. Often data scientists require different tools for doing the same thing on different sizes of data. The DataFrame solutions that exist for 1MB do not scale to 1TB+, and the overheads of the solutions for 1TB+ are too costly for datasets in the 1KB range. With Modin, because of its light-weight, robust, and scalable nature, you get a fast DataFrame at 1MB and 1TB+.

Modin is currently under active development. Requests and contributions are welcome!

If you are interested in learning more about Modin, please check out the *Getting Started* guide then refer to the [Developer Documentation](#) section, where you can find system architecture, internal implementation details, and other useful information. Also check out the [Github](#) to view open issues and make contributions.

PYTHON MODULE INDEX

m

`modin.experimental.cloud`, [65](#)

INDEX

C

CannotDestroyCluster, 65

CannotSpawnCluster, 65

ClusterError, 65

create_cluster() (in module
modin.experimental.cloud), 65

D

DataFrame (class in modin.pandas.dataframe), 125

E

EnvironmentVariable (class in modin.config.envvars),
41

F

from_partitions() (in module
modin.distributed.dataframe.pandas), 47

G

get() (modin.config.envvars.EnvironmentVariable class
method), 41

get_connection() (in module
modin.experimental.cloud), 66

get_help() (modin.config.envvars.EnvironmentVariable
class method), 41

get_value_source() (modin.config.envvars.EnvironmentVariable
class method), 41

M

modin.experimental.cloud
module, 65

module
modin.experimental.cloud, 65

O

once() (modin.config.envvars.EnvironmentVariable
class method), 41

P

put() (modin.config.envvars.EnvironmentVariable class
method), 42

S

Series (class in modin.pandas.series), 128

subscribe() (modin.config.envvars.EnvironmentVariable
class method), 42

U

unwrap_partitions() (in module
modin.distributed.dataframe.pandas), 46