

DATA + ANALYTICS

Analytics Lab and Starflow

Jinyoung Kim

Head of Development, Analytics Platform

Morningstar, Inc.

Vision Inspiration

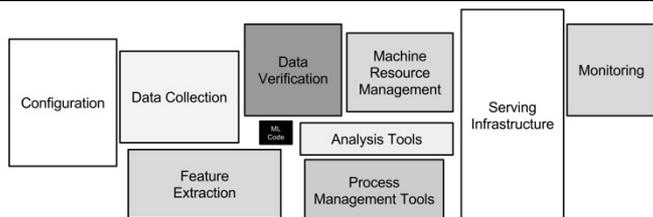
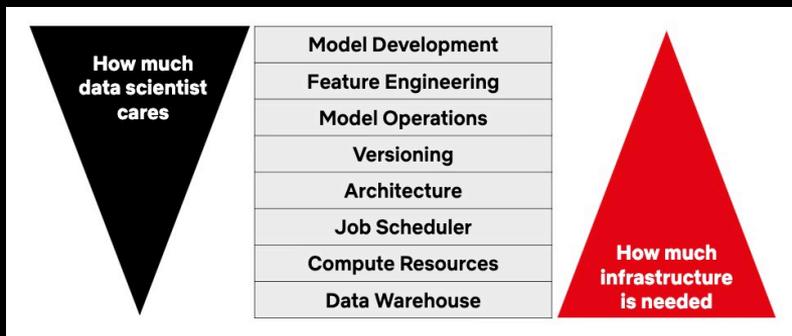


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

<https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf>



<https://docs.metaflow.org/introduction/what-is-metaflow>

Frameworks

“How do you build your webapp?”

2000



2004



2008



symfony



2015



Data/ML Frameworks

“How do you build your models?”

2013

SPSS
STATA

sas
MATLAB

R
python



Java



2015

TensorFlow

scikit
learn

PyTorch

2019

Kubeflow

mlflow

METAFLOW

Frameworks + Platform as a Service



+



STARFLOW

+

Analytics Platform

Starflow App Example

Flask

Flask is a lightweight [WSGI](#) web application framework. It is designed to make getting started quick and easy, with the ability to scale up to complex applications. It began as a simple wrapper around [Werkzeug](#) and [Jinja](#) and has become one of the most popular Python web application frameworks.

Flask offers suggestions, but doesn't enforce any dependencies or project layout. It is up to the developer to choose the tools and libraries they want to use. There are many extensions provided by the community that make adding new functionality easy.

```
from flask import Flask, escape, request

app = Flask(__name__)

@app.route('/')
def hello():
    name = request.args.get("name", "World")
    return f'Hello, {escape(name)}!'
```

```
1 from typing import Dict, Any
2 import starflow
3
4 app = starflow.App()
5
6 @app.run(run_type='add')
7 def add(x: int, y: int) -> Dict[str, Any]:
8     return {"sum": x+y}
9
```

Starflow API (Use Model)

GET	/status	The status, latency and details of various services on which CAP depend to keep the starflow model app up and running. A 200 is returned even if some dependent services are down.	
GET	/metadata	Metadata information like git commit, branch, version and environment of the starflow model app.	🔒
POST	/files	Generates a signed URL to upload the file.	🔒
GET	/files	Get a paginated list of files.	🔒
GET	/files/{id}	A secure signed url to download the file matching the unique file identifier.	🔒
GET	/validations/{id}	Get validation details for the given validation id.	🔒
GET	/validations	Get a paginated list of validations.	🔒
POST	/validations	Submit data for validation of inputs for different run types.	🔒
GET	/runs/{id}	Get details for the given run id.	🔒
GET	/runs	Get a paginated list of runs.	🔒
POST	/runs	Submit data to queue a run.	🔒
GET	/statistics	App Statistics like average and standard deviation of run-time for all run types of the starflow model app. Statistics are calculated from executions in the last 6 months.	🔒
POST	/schedules	Schedule a Starflow run.	🔒
GET	/schedules	Get a paginated list of schedules.	🔒
GET	/schedules/{id}	Get details for the given schedule id.	🔒
DELETE	/schedules/{id}	Archive a schedule for a Starflow run. Archived schedules do not trigger runs. They can still be accessed with GET /schedules/{id} but cannot be updated by PATCH /schedules/{id}.	🔒
PATCH	/schedules/{id}	Update a schedule for a Starflow run.	🔒

Starflow App Management API (Deploy Model)

POST	<code>/apps</code>	Create a starflow application in the Core Analytics Platform.	🔒
GET	<code>/apps/{appName}</code>	Get app information	🔒
PATCH	<code>/apps/{appName}</code>	Update partial starflow application info	🔒
POST	<code>/apps/{appName}/environments</code>	Create an environment to run the starflow application.	🔒
PATCH	<code>/apps/{appName}/environments/{envName}</code>	Update the starflow application runtime environment	🔒
GET	<code>/apps/{appName}/environments/{envName}</code>	Get app environment level information	🔒
POST	<code>/gitevent</code>	Send a git event notification of the starflow application repository.	
GET	<code>/apps/environments</code>	Get app name and environment name for the matching S3 bucket path	🔒
POST	<code>/apps/{appName}/environments/{envName}/execute</code>	Trigger the app build in the desired environment.	
POST	<code>/apps/{appName}/environments/{envName}/promote</code>	Promote a non-prod app environment to production.	

Analytics Lab

The screenshot displays the Analytics Lab interface. On the left is a sidebar with a 'Data Explorer' section containing 'Select Dataset' (Equity) and 'Select Database' (All databases). Below this is a 'Tables' section with a search bar and a list of tables including 'Aggregation Daily Dividend Metrics', 'Aggregation Efficiency Ratios', 'Aggregation Enterprise Value Calculations', 'Aggregation Financial Health Ratios', 'Aggregation Financial Statements Simple Summation', 'Aggregation Financial Statements Weighted Average', 'Aggregation Leverage', 'Aggregation Margins', 'Aggregation Market Capital', 'Aggregation Monthly Dividend Metrics', 'Aggregation Price Multiples', 'Aggregation Price Yields', 'Aggregation Profitability', 'Aggregation Residual Risk And Return Sensitivity', 'Aggregation Trailing Return', 'Average Rates', and 'Average Rates'. The main content area features the Morningstar logo and the title 'Analytics Lab'. It includes a 'Launcher' section with three notebook options: 'Basic Notebook' (Starting out fresh?), 'Direct Notebook' (Building for Direct?), and 'Presentation Studio Notebook' (Building a report?). Below this is a 'Morningstar Notebooks' section with five tiles: 'Firm Diversity', 'Investment Product Launches', 'Manager History', 'Security Ownership Analysis', and 'Time Series Regression Tools'. A 'Utilities' section follows with five tiles: 'Show Contextual Help', 'Terminal', 'Python 3 (ipykernel)', 'Octave', 'R', 'Data Quality', 'Text File', 'Markdown File', and 'VS Code IDE [?]'.

Analytics Lab (beta) File Edit View Run Kernel Git Tabs Settings Help Analytics Lab

Data Explorer Direct User Objects

Select Dataset
Equity

Select Database
All databases

Tables

- Aggregation Daily Dividend Metrics
- Aggregation Efficiency Ratios
- Aggregation Enterprise Value Calculations
- Aggregation Financial Health Ratios
- Aggregation Financial Statements Simple Summation
- Aggregation Financial Statements Weighted Average
- Aggregation Leverage
- Aggregation Margins
- Aggregation Market Capital
- Aggregation Monthly Dividend Metrics
- Aggregation Price Multiples
- Aggregation Price Yields
- Aggregation Profitability
- Aggregation Residual Risk And Return Sensitivity
- Aggregation Trailing Return
- Average Rates
- Average Rates

Launcher

MORNINGSTAR

Analytics Lab

Drive efficiencies, mitigate risk, & uncover new insights by building scalable, custom analytics with Morningstar data.

Open a notebook to get started.

Basic Notebook

Starting out fresh? Use this blank notebook with the Morningstar data integration established for you.

Direct Notebook

Building for Direct? Start with this notebook. It has guidance on best practices for publishing your notebook to Direct.

Presentation Studio Notebook

Building a report? This notebook highlights best practices for building notebooks to work with Presentation Studio.

Morningstar Notebooks

- Firm Diversity
- Investment Product Launches
- Manager History
- Security Ownership Analysis
- Time Series Regression Tools

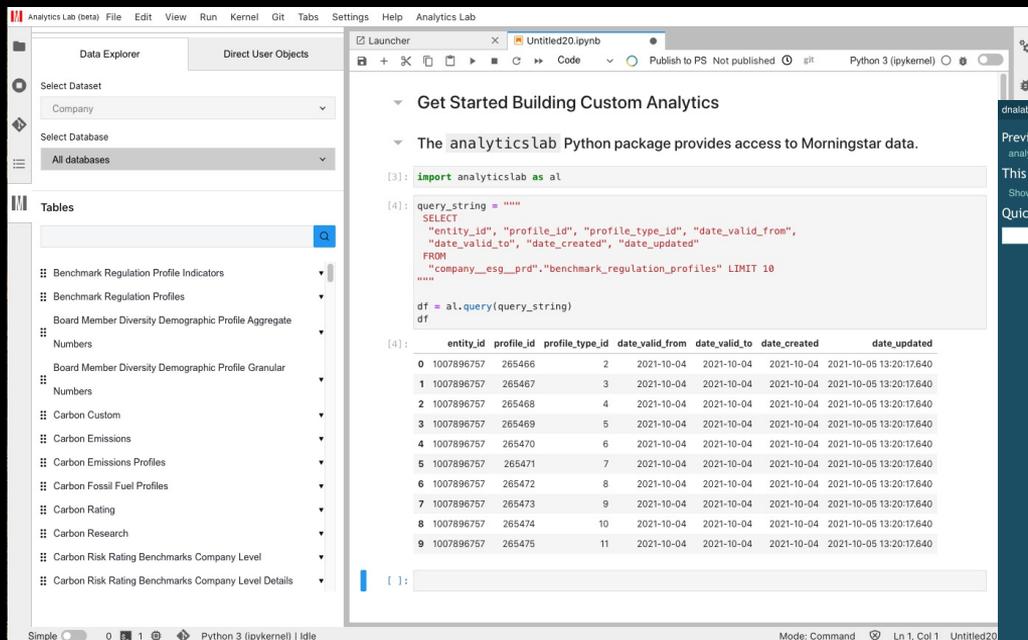
Utilities

- Show Contextual Help
- Terminal
- Python 3 (ipykernel)
- Octave
- R
- Data Quality
- Text File
- Markdown File
- VS Code IDE [?]

Simple 0 0 0

Feedback

Easy Access to Data and Compute



The screenshot shows the Analytics Lab interface. On the left, there is a sidebar with a 'Data Explorer' section containing 'Direct User Objects', 'Select Dataset', 'Select Database', and 'Tables'. The main area displays a Jupyter notebook titled 'Untitled20.ipynb' with the following content:

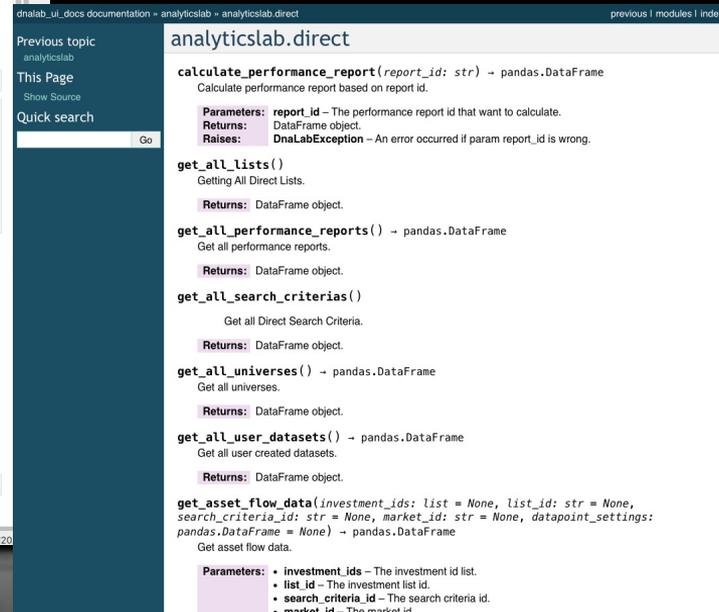
```
Get Started Building Custom Analytics
```

The `analyticslab` Python package provides access to Morningstar data.

```
[3]: import analyticslab as al
```

```
[4]: query_string = """
      SELECT
        "entity_id", "profile_id", "profile_type_id", "date_valid_from",
        "date_valid_to", "date_created", "date_updated"
      FROM
        "company_esg_prd"."benchmark_regulation_profiles" LIMIT 10
      """
      df = al.query(query_string)
      df
```

	entity_id	profile_id	profile_type_id	date_valid_from	date_valid_to	date_created	date_updated
[4]:	0	1007896757	265466	2	2021-10-04	2021-10-04	2021-10-04 2021-10-05 13:20:17.640
	1	1007896757	265467	3	2021-10-04	2021-10-04	2021-10-04 2021-10-05 13:20:17.640
	2	1007896757	265468	4	2021-10-04	2021-10-04	2021-10-04 2021-10-05 13:20:17.640
	3	1007896757	265469	5	2021-10-04	2021-10-04	2021-10-04 2021-10-05 13:20:17.640
	4	1007896757	265470	6	2021-10-04	2021-10-04	2021-10-04 2021-10-05 13:20:17.640
	5	1007896757	265471	7	2021-10-04	2021-10-04	2021-10-04 2021-10-05 13:20:17.640
	6	1007896757	265472	8	2021-10-04	2021-10-04	2021-10-04 2021-10-05 13:20:17.640
	7	1007896757	265473	9	2021-10-04	2021-10-04	2021-10-04 2021-10-05 13:20:17.640
	8	1007896757	265474	10	2021-10-04	2021-10-04	2021-10-04 2021-10-05 13:20:17.640
	9	1007896757	265475	11	2021-10-04	2021-10-04	2021-10-04 2021-10-05 13:20:17.640



The screenshot shows the documentation for `analyticslab.direct`. The page title is `analyticslab.direct`. The documentation includes the following sections:

- calculate_performance_report(report_id: str) - pandas.DataFrame**
Calculate performance report based on report id.
Parameters: report_id - The performance report id that want to calculate.
Returns: DataFrame object.
Raises: DnaLabException - An error occurred if param report_id is wrong.
- get_all_lists()**
Getting All Direct Lists.
Returns: DataFrame object.
- get_all_performance_reports() - pandas.DataFrame**
Get all performance reports.
Returns: DataFrame object.
- get_all_search_criterias()**
Get all Direct Search Criteria.
Returns: DataFrame object.
- get_all_universes() - pandas.DataFrame**
Get all universes.
Returns: DataFrame object.
- get_all_user_datasets() - pandas.DataFrame**
Get all user created datasets.
Returns: DataFrame object.
- get_asset_flow_data(investment_ids: list = None, list_id: str = None, search_criteria_id: str = None, market_id: str = None, datapoint_settings: pandas.DataFrame = None) - pandas.DataFrame**
Get asset flow data.
Parameters:
 - investment_ids - The investment list id.
 - list_id - The investment list id.
 - search_criteria_id - The search criteria id.
 - market_id - The market id.

Speed of Delivery of Insights

Forbes

Subscribe Sign In

Sep 28, 2021, 11:43am EDT | 25,905 views

Speed Matters: How Morningstar Accelerates Innovation With Data

aws Amazon Web Services BRANDVOICE | Paid Program
Leadership

f When the New York Stock Exchange (NYSE) announced plans to delist several large Chinese telecom firms in January 2021, investors grew concerned. The move was a response to an executive order targeting companies suspected of helping the Chinese military, and it required investors to divest of the affected stocks by November.

t

in



ADVERTISEMENT

aws

Reinvent your business with data

Leverage data insights to drive better, faster outcomes

Learn how it works

Empowering Investors

Pioneering a new way to empower investors

The financial services company Morningstar has always been a leader in investment research and insight. But the way it was storing and using its data was starting to hold it back.

The screenshot shows the Morningstar website's analytics lab banner. At the top, there is a search bar with the text 'Search Quotes and Site' and a 'Contact Us' button. Below the search bar are navigation links: 'Software & Data Services', 'Money Management', 'Solutions for Business', and 'All Products & Services'. The main banner has a purple and orange background with the text 'Morningstar Notebooks in Analytics Lab' and 'Automate your innovation with advanced analytics at the click of a button'. Below this, it says 'Discover Notebooks with your free Morningstar Direct trial today!' and a button that says 'Get Morningstar Notebook Email Updates'. At the bottom of the banner, there is a line of text: 'To succeed in the financial services industry, you need to transform vast amounts of data into actionable insights every day, which can be challenging and time-consuming. Our newest Direct workspaces centralizes data, analytics, and...'.

