Trino and Starburst Training Series:

Transformation processing with PyStarburst

v1.0.0

Session 5 of 5: [Full Series Information]

Prior sessions

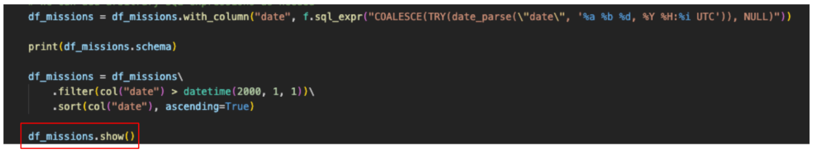
1. https://www.starburst.io/resources/creating-querying-data-lake-tables-on-demand/
2. https://www.starburst.io/resources/modern-table-formats-apache-iceberg-on-demand/
3. https://www.starburst.io/resources/data-pipelines-views-data-products/
4. https://www.starburst.io/resources/experience-warp-speed-in-action/

This workshop is focused on introducing & exercising the PyStarburst DataFrame API. These are the goals for this session.

- Differential SQL-based data engineering from programming-based.
- Understand how PyStarburst implements lazy execution with Starburst Galaxy.
- Explore the DataFrame API.
- Write Python code to perform analytical questions and transformation processing.

Table of Contents

PyStarburst Overview

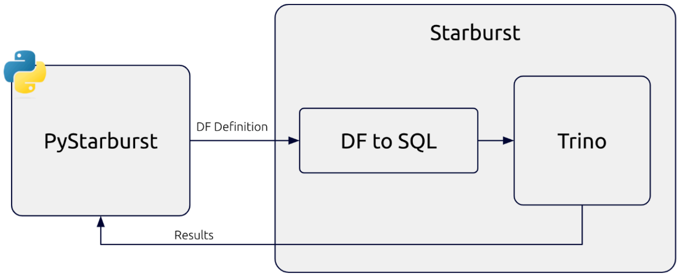


```
df_missions = df_missions.with_column("date", f.sql_expr("COALESCE(TRY(date_parse(\"date\", '%a %b %d, %Y %H:%i UTC')), NULL)"))

print(df_missions.schema)

df_missions = df_missions\
    .filter(col("date") > datetime(2000, 1, 1))\
    .sort(col("date"), ascending=True)

df_missions.show()
```



- PySpark-like Syntax
- Lazy execution
- Python gets converted to SQL
- Heavy lifting done by Trino

Links: API Docs & Example Code

Currently supported on Starburst Galaxy and Starburst Enterprise.

Create Starburst Galaxy account

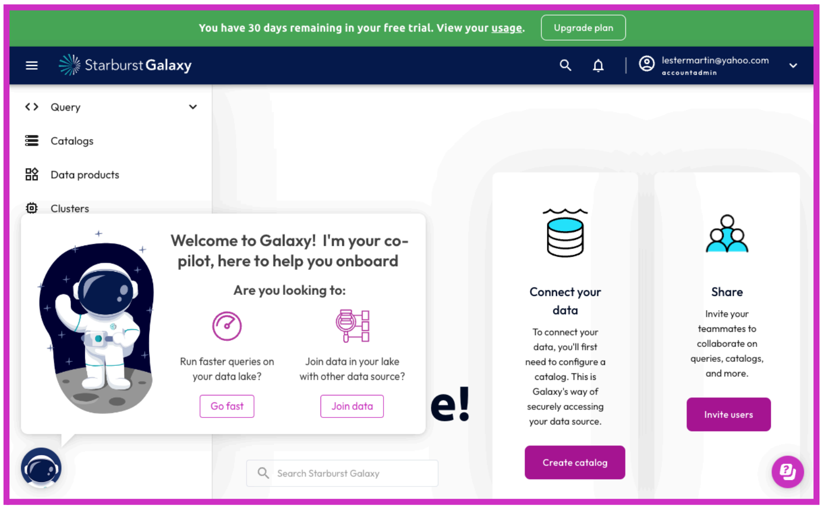
For each webinar in the 5-part series, you will be using your own Starburst Galaxy environment. ***If you have already registered for Starburst Galaxy, you may skip this FIRST lab.***

To sign up for Starburst Galaxy, follow the instructions on the free registration page at https://www.starburst.io/platform/starburst-galaxy/start/.

Note: You will receive an "invitation" email. Please check your spam or junk folder if it does not immediately arrive in your inbox.

After you have entered your confirmation code, set a password, and selected your new domain name, you will be presented with a series of questions about your desired usage for Starburst Galaxy. Complete these with whatever you choose to share.

Eventually, you will likely be presented with a page similar to the following screenshot.



You have 30 days remaining in your free trial. View your usage.

Welcome to Galaxy! I'm your co-pilot, here to help you onboard

Are you looking to:

Run faster queries on your data lake?

Join data in your lake with other data source?

Connect your data

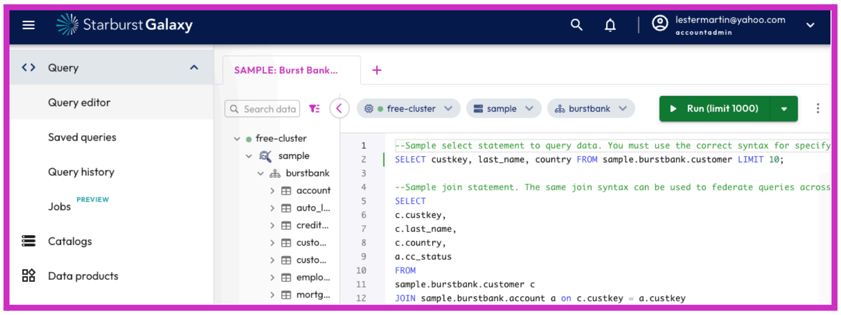
To connect your data, you'll first need to configure a catalog. This is Galaxy's way of securely accessing your data source.

Share

Invite your teammates to collaborate on queries, catalogs, and more.

Click on the astronaut helmet icon in the lower-left to silence the pop-up coming from it.

At this point, you should see something similar to this to indicate you are fully configured.



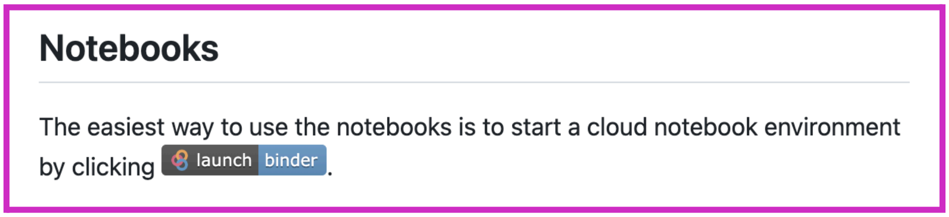
```
--Sample select statement to query data. You must use the correct syntax for specify
SELECT custkey, last_name, country FROM sample.burstbank.customer LIMIT 10;

--Sample join statement. The same join syntax can be used to federate queries across
SELECT
c.custkey,
c.last_name,
c.country,
a.cc_status
FROM
sample.burstbank.customer c
JOIN sample.burstbank.account a on c.custkey = a.custkey
```

PyStarburst lab exercises

The code used in this webinar is in a Jupyter notebook file that is available at https://github.com/starburstdata/pystarburst-examples/blob/main/notebooks/tpch.ipynb.

If you already have access to a jupyter instance, you can simply upload this notebook into it. If you do not, visit https://github.com/starburstdata/pystarburst-examples and click on the button labeled "**launch binder**".

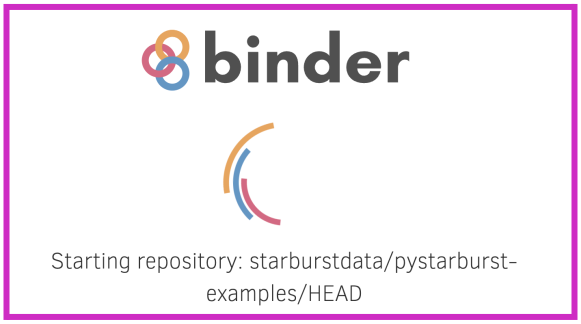


Notebooks

The easiest way to use the notebooks is to start a cloud notebook environment by clicking launch binder.

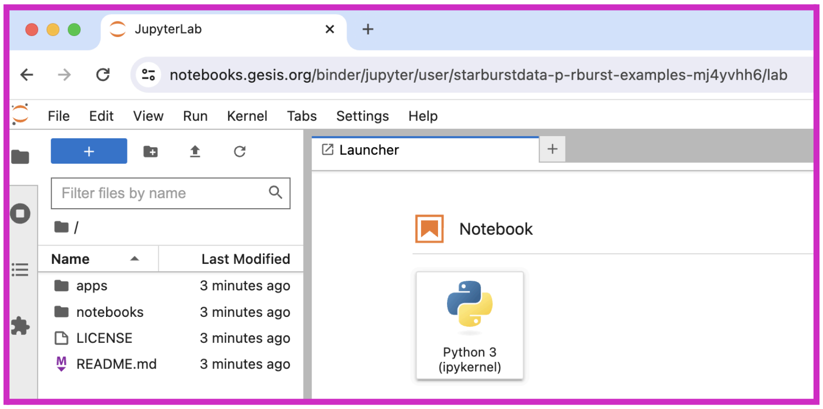
Alternatively, click on https://mybinder.org/v2/gh/starburstdata/pystarburst-examples/HEAD if you cannot access GitHub directly.

Regardless of which route you choose, your browser will look like this once launched.



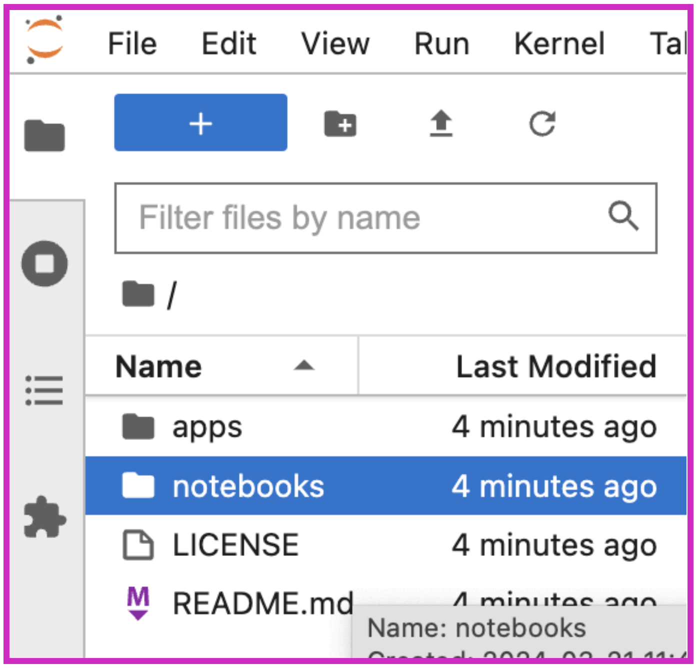
Starting repository: starburstdata/pystarburst-examples/HEAD

After a short time, a jupyter web-based notebook system will load.



Name	Last Modified
apps	3 minutes ago
notebooks	3 minutes ago
LICENSE	3 minutes ago
README.md	3 minutes ago

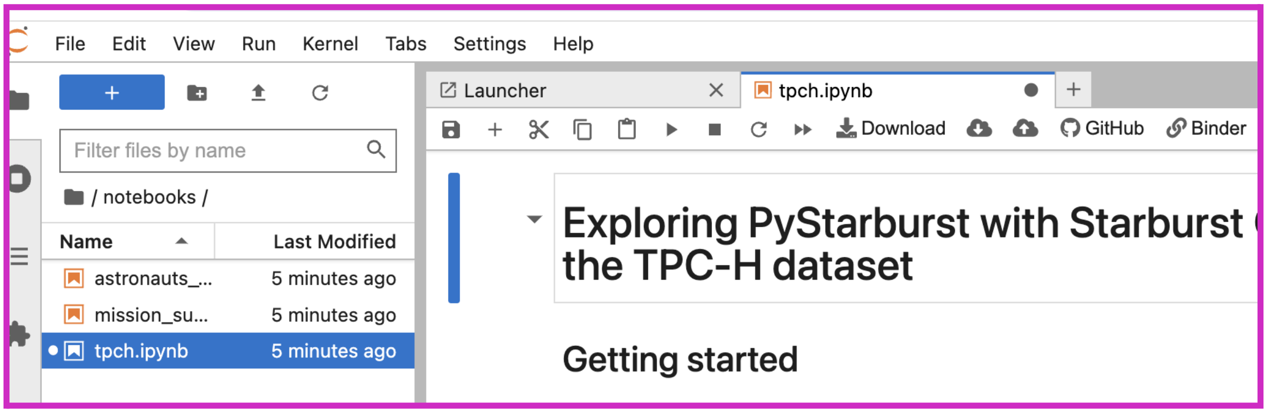
Now, double-click on **notebooks** in the explorer pane on the left.



Name	Last Modified
apps	4 minutes ago
notebooks	4 minutes ago
LICENSE	4 minutes ago
README.md	4 minutes ago

Name: notebooks
Created: 2024-02-21 11:

Then double-click on **tpch.ipynb** in the same explore pane on the left which will render the notebook that contains the code for this lab.



Name	Last Modified
astronauts_...	5 minutes ago
mission_su...	5 minutes ago
tpch.ipynb	5 minutes ago

Exploring PyStarburst with Starburst (the TPC-H dataset

Getting started

The instructor will walk through executing this notebook, cell by cell, and will provide additional information and discussions on each one. If reviewing this document after the live webinar, please feel free to view the following YouTube video with the same discussions.

▶ Using PyStarburst on Starburst Galaxy with the TPC-H dataset