

📖 06a_timeseries_forecasting_vertex_ai_notebook_execution.md

🔗 Timeseries Forecasting using sessions in Serverless Spark through Vertex AI

Following are the lab modules:

- [1. Understanding Data](#)
- [2. Solution Architecture](#)
- [3. Execution](#)
- [4. Logging](#)

🔗 1. Understanding Data

🔗 Data Files

The datasets used for this project are:

- train.csv: This file contains the date, store, item, sales data.

date - Date of the sale data. There are no holiday effects or store closures.

store - Store ID

item - Item ID

sales - Number of items sold at a particular store on a particular date.

- test.csv:

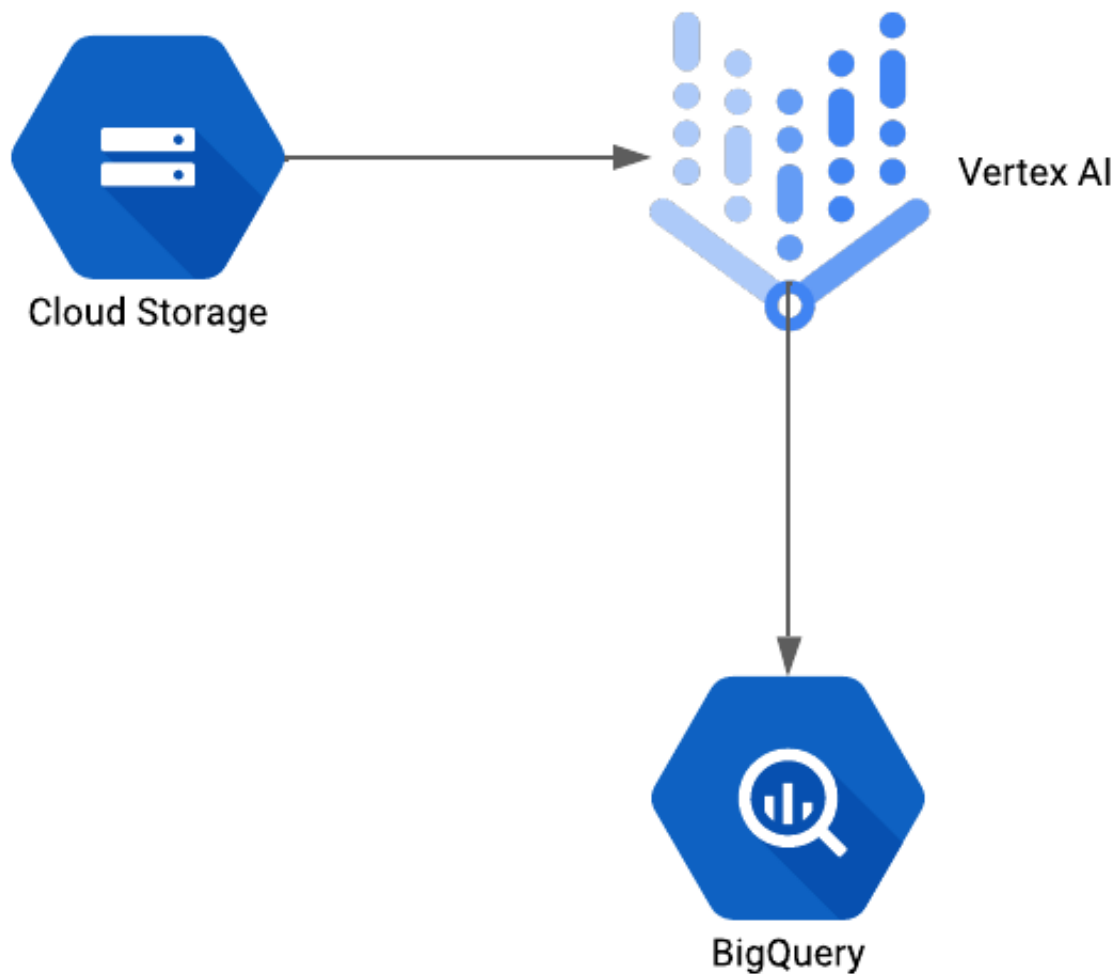
id- Unique identifier

date - Date of the sale data. There are no holiday effects or store closures.

store - Store ID

item - Item ID

🔗 2. Solution Architecture



3. Execution

3.1. Run the Batch by creating session.

Creating Notebook in Vertex AI

Select Workbench from the left scroll bar of the Vertex AI main page. Select the Managed Notebooks tab. In the Managed Notebooks tab, click the New Notebook icon.

Vertex AI Workbench

1 NEW NOTEBOOK **2** MANAGED NOTEBOOKS USER-MANAGED NOTEBOOKS EXECUTIONS SCHEDULES

Managed notebooks provide JupyterLab services and flexible computing resources integrated with Google Cloud services. [Learn more](#)

Region: us-central1 (Iowa)

Filter: Enter property name or value

Notebook name	Location	Owner	Last modified
...

Next, fill in the following values in the Notebook creation window as shown in the images below:

- **Notebook Name** - A unique identifier for your Notebook
- **Region** - The region name provided by the Admin team
- **Permission Type** - Single User Only (Single user only mode restricts access to the specified user)
- Provide a name and region to the notebook and select 'Single User Only' and click 'Create'. We will let the 'Advanced Settings' remain as the default values.

←

Create a managed notebook

Notebook name *

63-char limit with lowercase letters, digits, or '-' only. Must start with a letter. Cannot end with a '-'.

Region

us-central1 (Iowa)

▼

?

Permission

JupyterLab access modes determine who can use a notebook instance and which credentials are used to call Google APIs. This cannot be changed once the notebook is created.

☐ Service account

Service account mode allows anyone who is granted the iam.serviceAccounts.actAs permission on the specified service account to access the JupyterLab UI. [Learn more](#)

☒ Single user only

Single user only mode restricts access to the user specified below. [Learn more](#)

User email *

Advanced settings

▼

CREATE

CANCEL

- Once the notebook is running, click the 'OPEN JUPYTERLAB' option next to the Notebook name as shown below

<input type="checkbox"/>	<input checked="" type="checkbox"/>	Notebook name ↑		Location	Owner	Last modified
<input type="checkbox"/>	<input checked="" type="checkbox"/>		OPEN JUPYTERLAB	us-central1-a	Service account	28 Apr 2022, 13:11:01

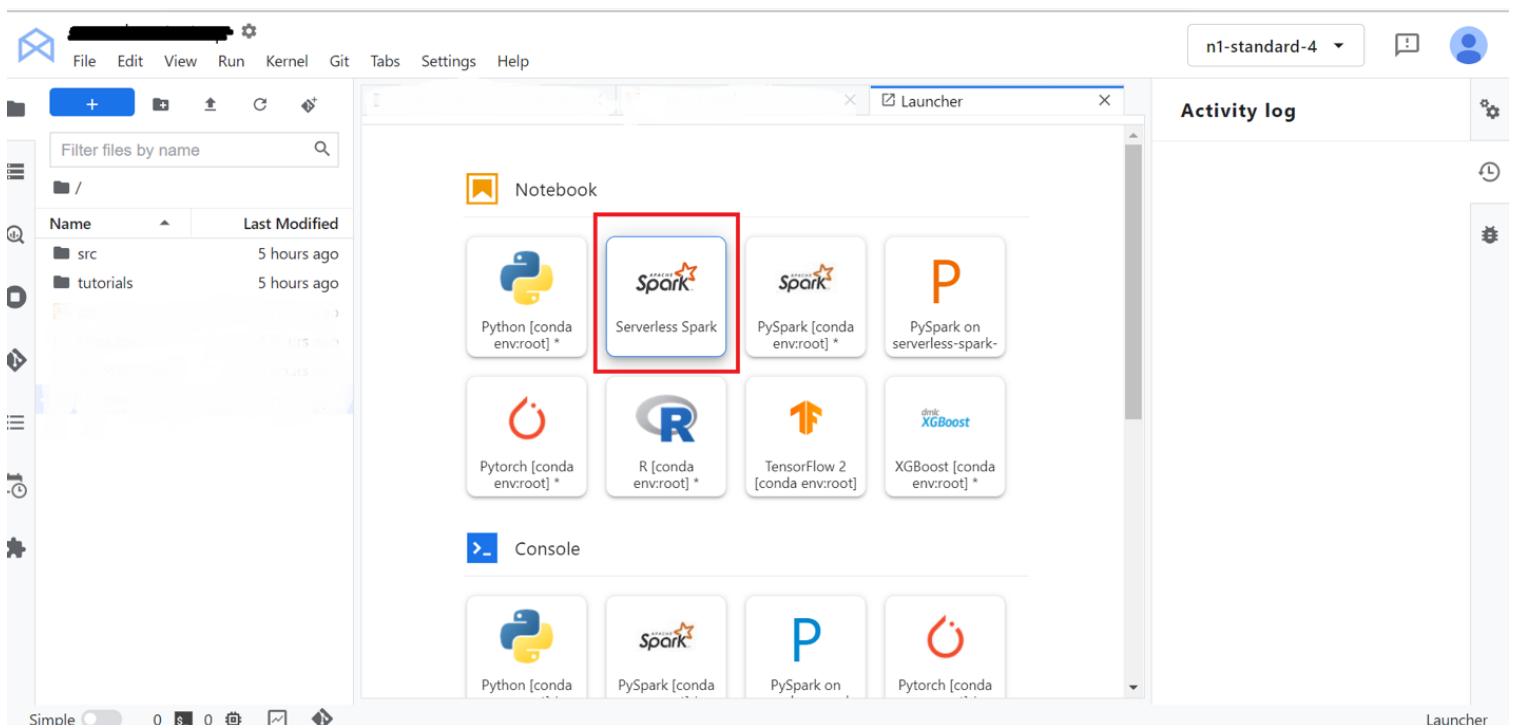
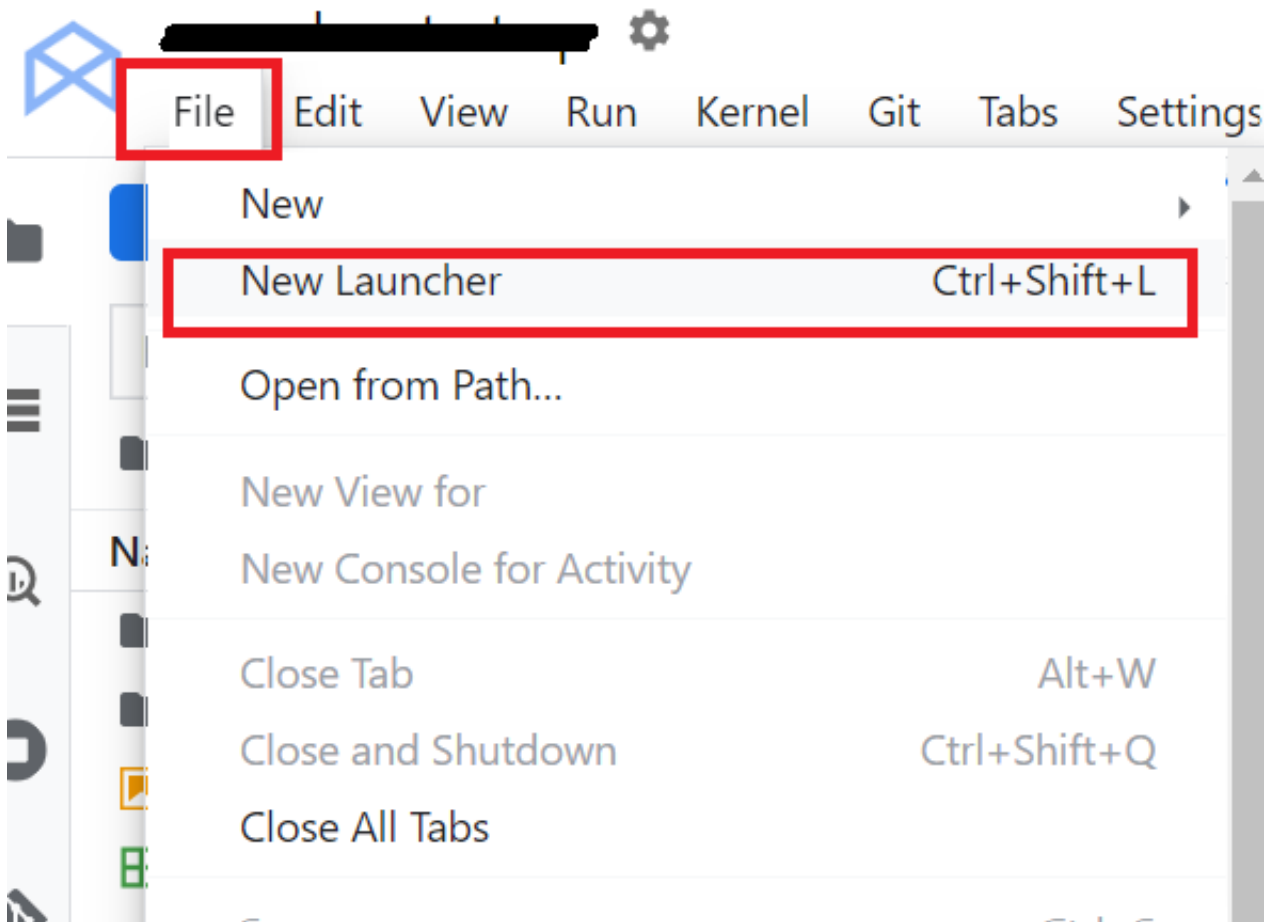
- Follow the on screen instructions to launch the JupyterLab session

🔗 Create Serverless Spark Session

http://localhost:6419/

Page 4 of 11

- Click on the File and the New launcher and select Serverless Spark



🔗 Follow the on screen instructions to create Session

🔗 3.2. Provide the details for the Session

Next, fill in the following values in the session creation window as shown in the images below:

- **Session Name** - A unique identifier for your session
- **Region** - The region name provided by the Admin team
- **Language** - Pyspark
- **Autoshutdown** - 24 hours
- **Service Account** - <UMSA_NAME>@<PROJECT_ID>.iam.gserviceaccount.com
- **Network Configuration** - Select the network and subnetwork provided by the Admin team
- **History Server Cluster** -
projects/<PROJECT_ID>/regions/<REGION_NAME>/clusters/<HISTORY_SERVER_NAME>
- **Container** - gcr.io/<PROJECT_ID>/<CONTAINER_IMAGE>:1.0.1
- Click the **SUBMIT** button to create the session.

Create Serverless Spark Session

[PREVIEW](#)

Basic info

Session name *

[REDACTED]

Up to 128 lowercase letters, numbers, or underscores.

Language

PySpark



Region

us-west1

Autoshutdown

24h

The session will automatically shutdown after 24 hours.

Execution configuration

Service Account

Enter your service account

If not provided, the default GCE service account will be used. [Learn More](#)

Network configuration

Private IP Google Access must be enabled on the network.

- ☒ Networks in this project
- ☐ Networks shared from host project: "undefined"

Network

Subnetwork

^ ADVANCED OPTIONS

Container

Custom container image

gcr.io/

Peripheral configuration

Metastore Service

None

We recommend this option to persist table metadata when a cluster is shut down, for a metastore shared by different clusters, or for metadata operability across GCP products.

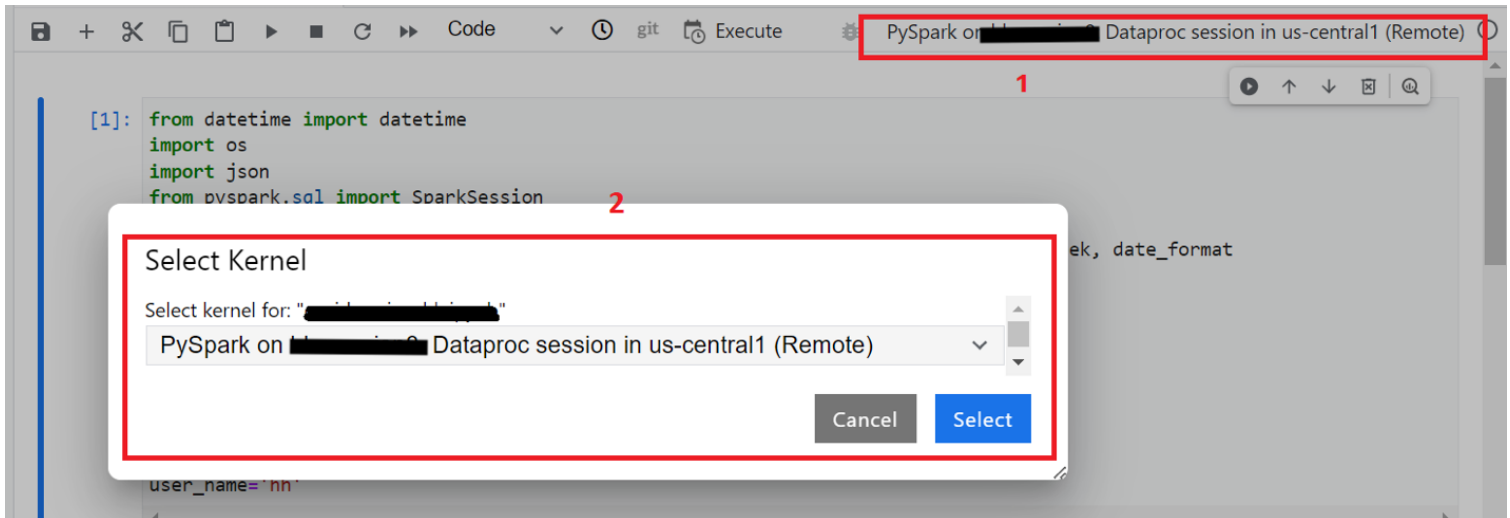
History server cluster

Choose a history server cluster to store logs in.

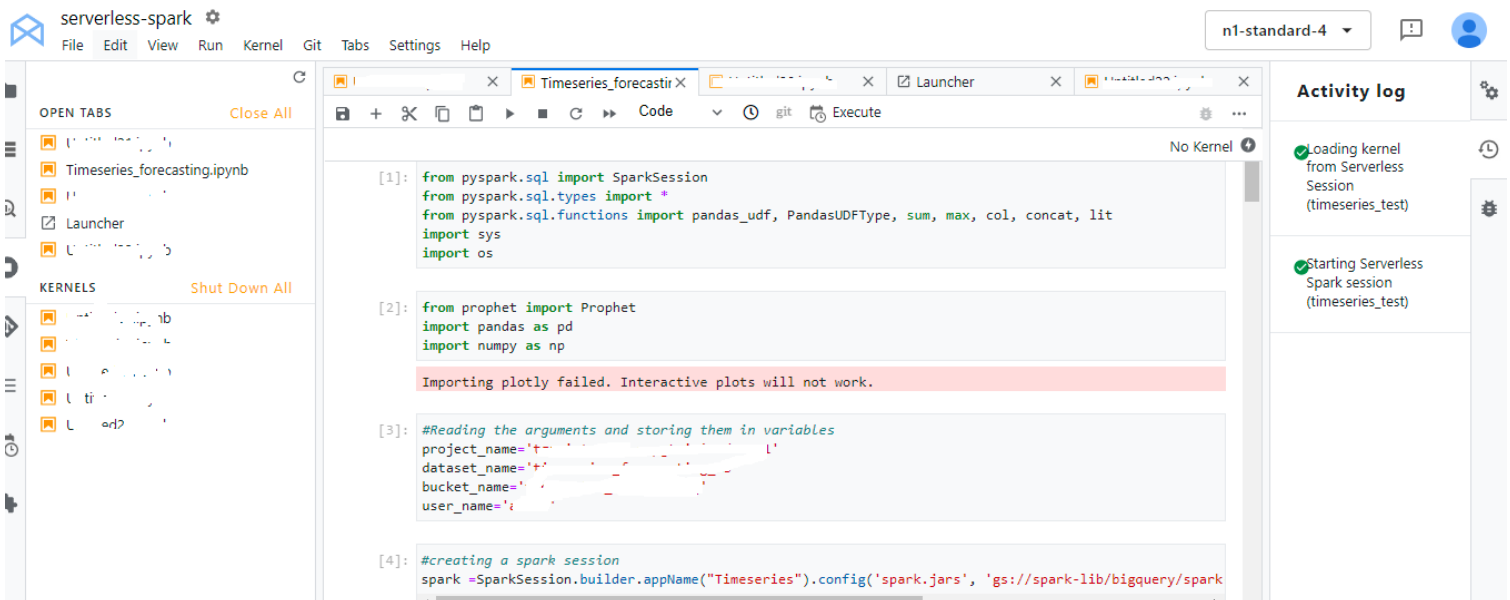
History server cluster

projects/ /regions/us-central1/clust...

- Once the Session is created select 'No Kernel' from the kernel dropdown list and then delete the notebook



- Next, using the browser option from JupyterLab, navigate to the Notebook file located at: <bucket_name> > 'timeseries_forecasting' > 00-scripts > timeseries_forecasting.ipynb
- From the kernel dropdown list, select the kernel for the session created in section 3.2
- Pass the values to the variables project_name, dataset_name, bucket_name as provided by the Admin and replace user_name by your username
- Next, hit the **Execute** button as shown below to run the code in the notebook.



3.3. Check the output table in BigQuery

Navigate to BigQuery Console, and check the **timeseries_forecasting** dataset.

Once the code has successfully executed, new table '<your_name_here>_global_predictions' will be created :

To query the data to find the list of stocks with highest stringency Index, run the following query -

```
select * from `<GCP-PROJECT-NAME>.<BQ-DATASET-NAME>.<user_name>_global_predictions`
```

Note: Edit all occurrences of and to match the values of the variables PROJECT_ID,user_name and BQ_DATASET_NAME respectively

Explorer

+ ADD DATA

K

Type to search

?

Viewing projects.

timeseries_forecasting_ds

_global_predictions

wordcount_dataset

Editor

*Unsaved ...y 2

+ +

RUN

SAVE

SHARE

SCHEDULE

MORE

This query will process 1,007.23

1

SELECT * FROM timeseries_forecasting_ds._global_predictions LIMIT 1000

Query results

SAVE RESULTS

EXPLORE DATA

JOB INFORMATION

RESULTS

JSON

EXECUTION DETAILS

Row	store	item	ds	yhat
1	1	1	2018-01-01	12.983731047884683
2	1	1	2018-01-02	15.64013857392832
3	1	1	2018-01-03	16.283219994725421
4	1	1	2018-01-04	16.944170700952707
5	1	1	2018-01-05	18.493730273560921

Results per page: 50

1 - 50 of 1000

REFRESH

PERSONAL HISTORY

PROJECT HISTORY

SAVED QUERIES

4. Logging

4.1 Persistent History Server logs

To view the Persistent History server logs, click the 'View History Server' button on the Sessions monitoring page and the logs will be shown as below:

As the session is still in active state , we will be able to find the logs in show incomplete applications.

Dataproc

Jobs on clusters

Clusters

Jobs

Workflows

Auto-scaling policies

Serverless

Batches

Sessions

Session details

PREVIEW

TERMINATE

VIEW LOGS

SPARK HISTORY SERVER

JUPYTER SESSION

cel-session-5

ca05ab02-3dbc-49f5-a07c-d6a00001a078

Active

28 Apr 2022

Properties

dataproc:jupyter.notebook.gcs.dir

gs://[REDACTED]

spark:spark.jars

gs://spark-lib/bigquery/spark-bigquery-with-dependencies_2.12-0.22.2.jar

spark:spark.executor.instances

2

spark:spark.driver.cores

4

spark:spark.executor.cores

4

spark:spark.eventLog.dir

gs://[REDACTED]phs/ca05ab02-3dbc-49f5-a07c-d6a00001a078/spark-job-history

History Server

Event log directory:

gs://[REDACTED]phs/*spark-job-history

Last updated:

2022-04-04 16:52:29

Client local time zone:

Asia/Calcutta

Search:

Version	App ID	App Name	Driver Host	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.2.1	[REDACTED]	[REDACTED]	10.122.15.217	2022-04-04 16:35:43	2022-04-04 16:36:44	1.0 min	spark	2022-04-04 16:36:45	<div>Download</div>

Showing 1 to 1 of 1 entries

Show incomplete applications

History Server

Event log directory:

gs://[REDACTED]phs/*spark-job-history

Last updated:

2022-04-04 16:52:29

Client local time zone:

Asia/Calcutta

Search:

Version	App ID	App Name	Driver Host	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.2.1	app-20220404110546-0000	[REDACTED]	10.122.15.217	2022-04-04 16:35:43	2022-04-04 16:36:44	1.0 min	spark	2022-04-04 16:36:45	<div>Download</div>

Showing 1 to 1 of 1 entries

Show incomplete applications