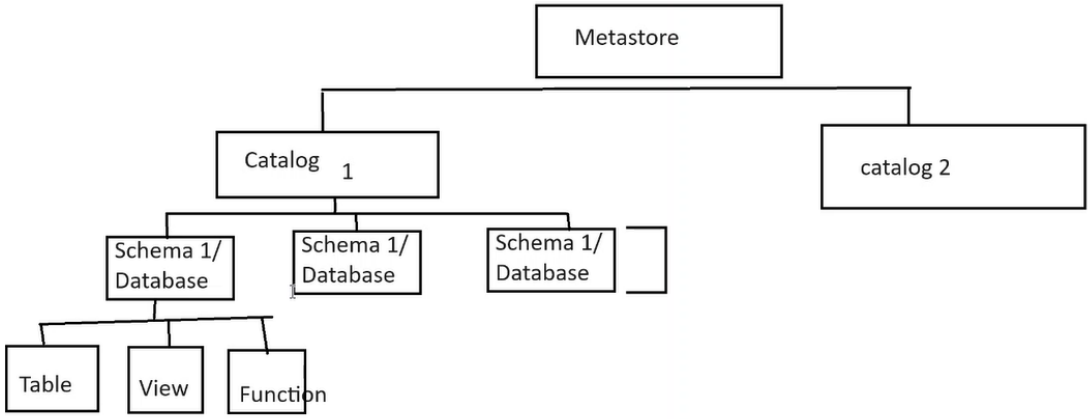
Resources - It consists of memory, storage, and compute power
Workload - Data Engineering (job, DLT), Data Science and ML, Data Analyst SQL (Modern DW)
Databricks does not have its own Storage(it has to rely on data lake) and Compute power(it has to rely on cloud providers). Hence it works with Azure, AWS, or GCP.
We can create dashboards using SQL queries in databricks.
Sparks + Databricks + Lakehouse (Data Lake + Data Warehouse)
Mounting technique (notebook authentication) -

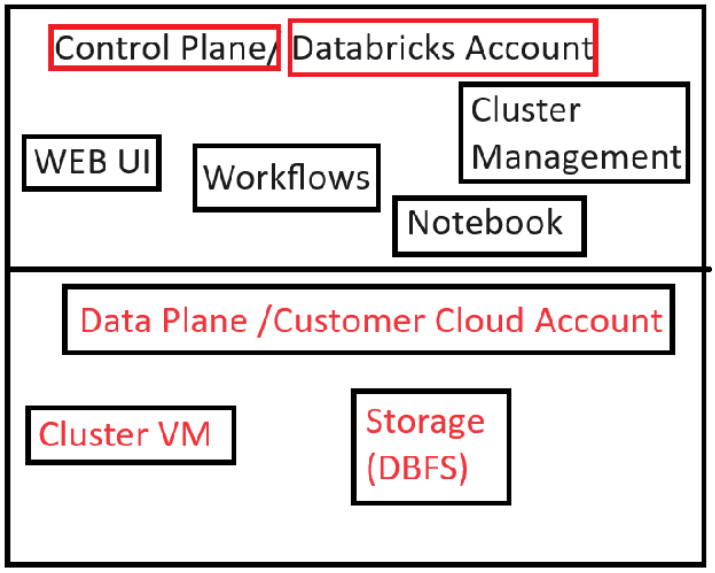
1. Access Key (Deprecated)
2. SAS (Shared Access Key)
3. Service Principal(Recommend)

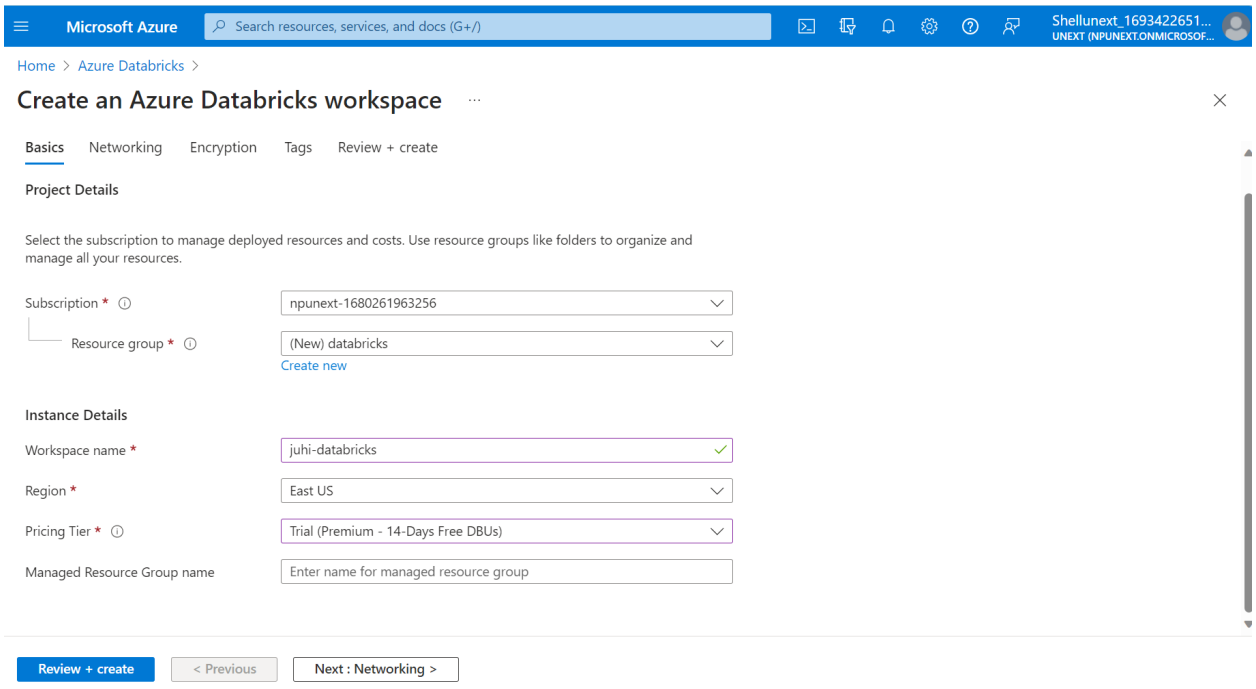
Databricks uses three level namespace : catalog.database.objectname

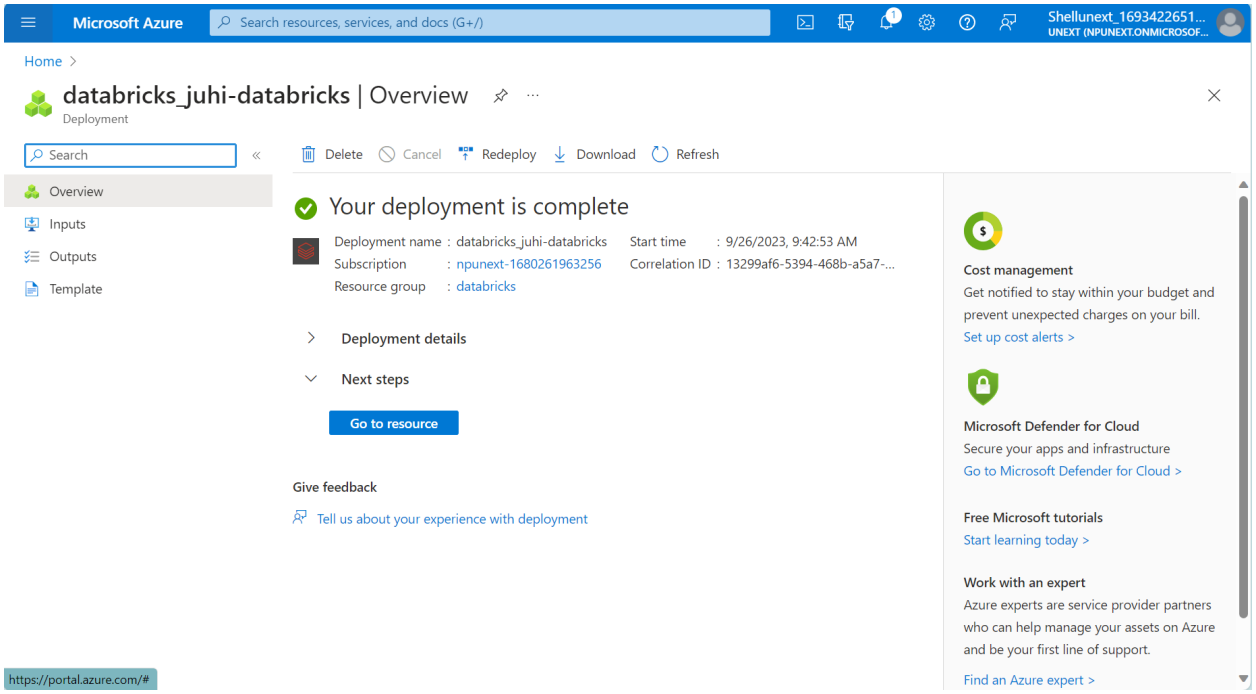


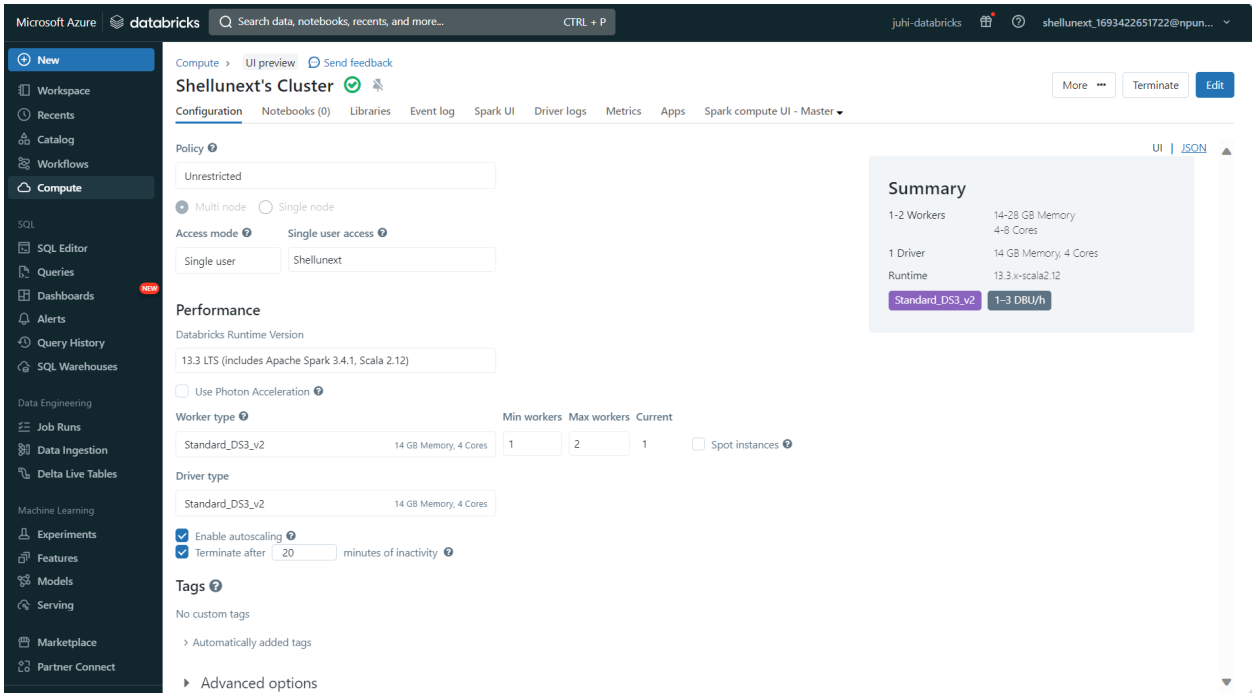
Databricks

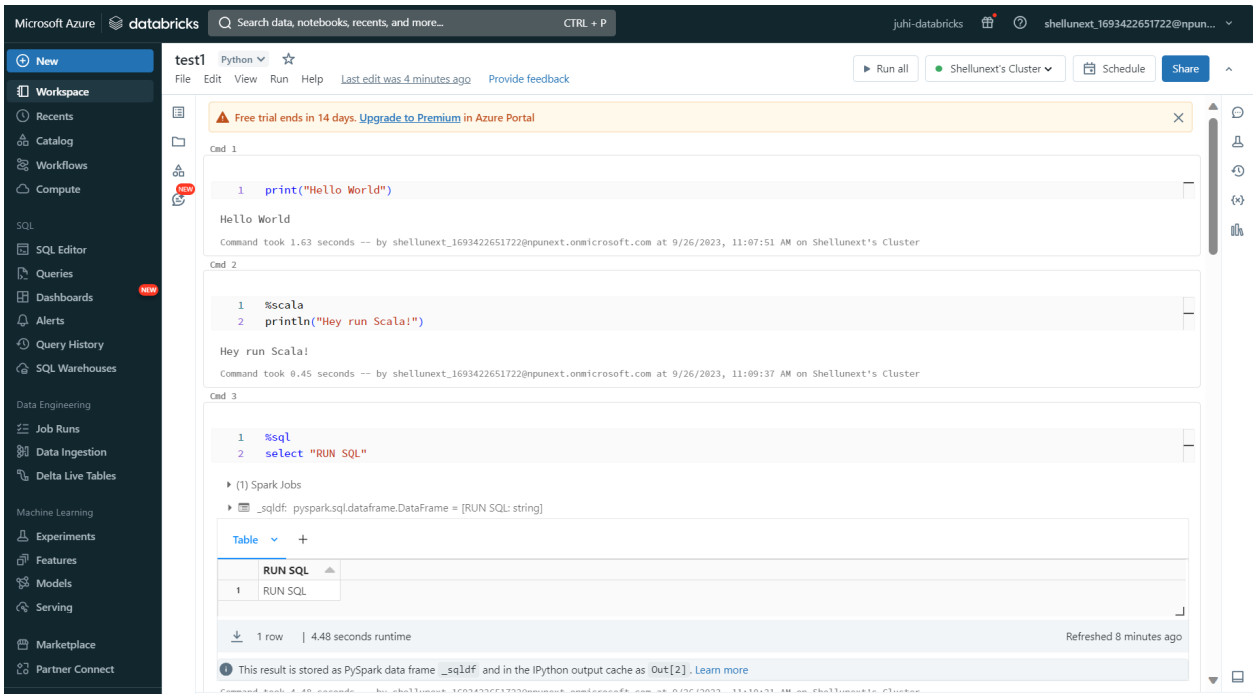
Architecture

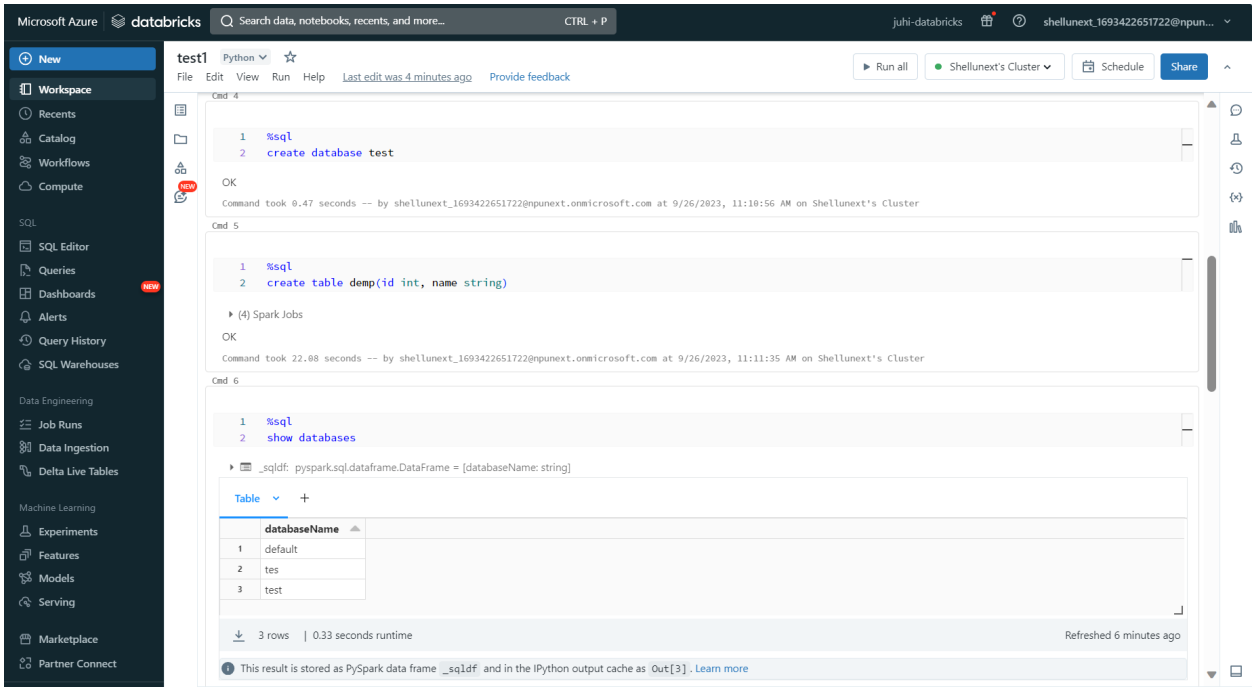


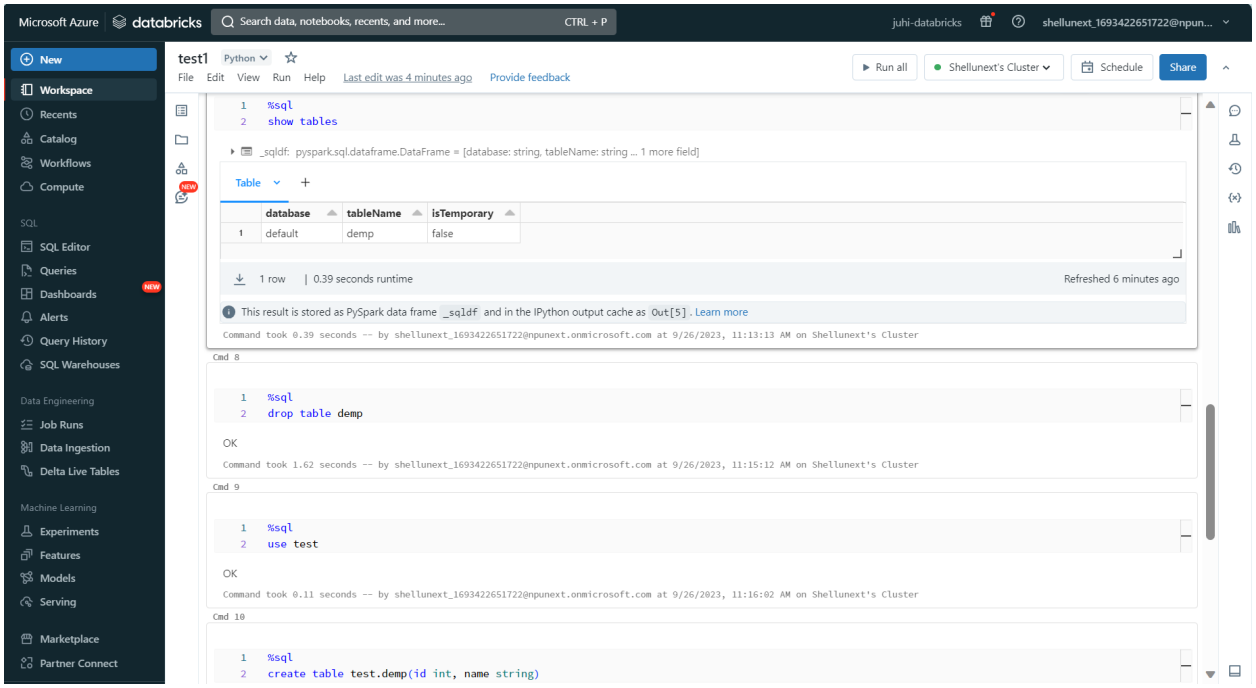


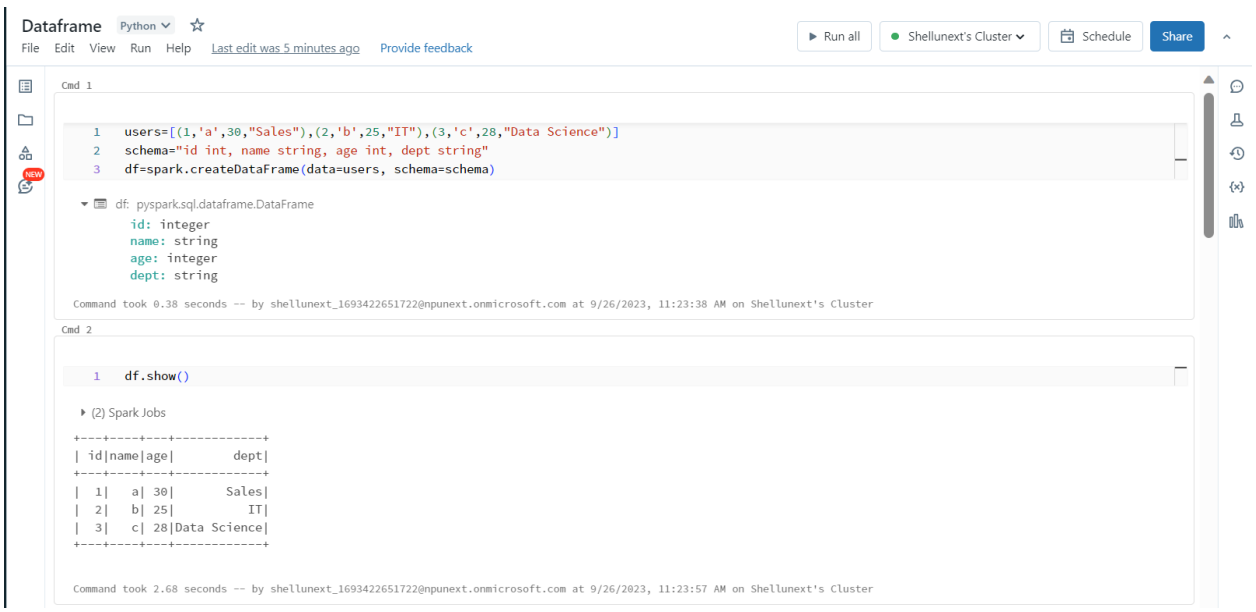

Microsoft Azure
Search resources, services, and docs (G+/)
Shellunext_1693422651...
UNEXT (NPUNEXT.ONMICROSOF...
Home
databricks_juhi-databricks | Overview
Deployment
Search
Overview
Inputs
Outputs
Template
Delete
Cancel
Redeploy
Download
Refresh
Your deployment is complete
Deployment name : databricks_juhi-databricks
Subscription : npunext-1680261963256
Resource group : databricks
Start time : 9/26/2023, 9:42:53 AM
Correlation ID : 13299af6-5394-468b-a5a7-...
Deployment details
Next steps
Go to resource
Give feedback
Tell us about your experience with deployment
Cost management
Get notified to stay within your budget and prevent unexpected charges on your bill.
Set up cost alerts >
Microsoft Defender for Cloud
Secure your apps and infrastructure
Go to Microsoft Defender for Cloud >
Free Microsoft tutorials
Start learning today >
Work with an expert
Azure experts are service provider partners who can help manage your assets on Azure and be your first line of support.
Find an Azure expert >
https://portal.azure.com/#


Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
juhi-databricks
shellunext_1693422651722@npun...
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Experiments
Features
Models
Serving
Marketplace
Partner Connect
Compute
UI preview
Send feedback
Shellunext's Cluster
More
Terminate
Edit
Configuration
Notebooks (0)
Libraries
Event log
Spark UI
Driver logs
Metrics
Apps
Spark compute UI - Master
Policy
Unrestricted
Multi node
Single node
Access mode
Single user access
Single user
Shellunext
Performance
Databricks Runtime Version
13.3 LTS (includes Apache Spark 3.4.1, Scala 2.12)
Use Photon Acceleration
Worker type
Min workers
Max workers
Current
Standard_DS3_v2
14 GB Memory, 4 Cores
1
2
1
Spot instances
Driver type
Standard_DS3_v2
14 GB Memory, 4 Cores
Enable autoscaling
Terminate after
20
minutes of inactivity
Tags
No custom tags
Automatically added tags
Advanced options
UI | JSON
Summary
1-2 Workers
14-28 GB Memory
4-8 Cores
1 Driver
14 GB Memory, 4 Cores
Runtime
13.3.x-scala2.12
Standard_DS3_v2
1–3 DBU/h


Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
juhi-databricks
shellunext_1693422651722@npun...
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Experiments
Features
Models
Serving
Marketplace
Partner Connect
test1
Python
File Edit View Run Help
Last edit was 4 minutes ago
Provide feedback
Run all
Shellunext's Cluster
Schedule
Share
Free trial ends in 14 days. Upgrade to Premium in Azure Portal
Cmd 1
print("Hello World")
Hello World
Command took 1.63 seconds -- by shellunext_1693422651722@npunext.onmicrosoft.com at 9/26/2023, 11:07:51 AM on Shellunext's Cluster
Cmd 2
%scala
println("Hey run Scala!")
Hey run Scala!
Command took 0.45 seconds -- by shellunext_1693422651722@npunext.onmicrosoft.com at 9/26/2023, 11:09:37 AM on Shellunext's Cluster
Cmd 3
%sql
select "RUN SQL"
(1) Spark Jobs
_sqldf: pyspark.sql.dataframe.DataFrame = [RUN SQL: string]
Table
RUN SQL
1 RUN SQL
1 row | 4.48 seconds runtime
Refreshed 8 minutes ago
This result is stored as PySpark data frame _sqldf and in the IPython output cache as Out[2]. Learn more


Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
juhi-databricks
shellunext_1693422651722@npun...
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Experiments
Features
Models
Serving
Marketplace
Partner Connect
test1
Python
File Edit View Run Help
Last edit was 4 minutes ago
Provide feedback
Run all
Shellunext's Cluster
Schedule
Share
Cmd 4
%sql
create database test
OK
Command took 0.47 seconds -- by shellunext_1693422651722@npunext.onmicrosoft.com at 9/26/2023, 11:10:56 AM on Shellunext's Cluster
Cmd 5
%sql
create table demp(id int, name string)
(4) Spark Jobs
OK
Command took 22.08 seconds -- by shellunext_1693422651722@npunext.onmicrosoft.com at 9/26/2023, 11:11:35 AM on Shellunext's Cluster
Cmd 6
%sql
show databases
_sqldf: pyspark.sql.dataframe.DataFrame = [databaseName: string]
Table
databaseName
1 default
2 tes
3 test
3 rows | 0.33 seconds runtime
Refreshed 6 minutes ago
This result is stored as PySpark data frame _sqldf and in the IPython output cache as Out[3]. Learn more


Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
juhi-databricks
shellunext_1693422651722@npun...
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Experiments
Features
Models
Serving
Marketplace
Partner Connect
test1
Python
File Edit View Run Help
Last edit was 4 minutes ago
Provide feedback
Run all
Shellunext's Cluster
Schedule
Share
1 %sql
2 show tables
_sqldf: pyspark.sql.dataframe.DataFrame = [database: string, tableName: string ... 1 more field]
Table
database
tableName
isTemporary
1 default demp false
1 row | 0.39 seconds runtime
Refreshed 6 minutes ago
This result is stored as PySpark data frame _sqldf and in the IPython output cache as Out[5]. Learn more
Command took 0.39 seconds -- by shellunext_1693422651722@npunext.onmicrosoft.com at 9/26/2023, 11:13:13 AM on Shellunext's Cluster
Cmd 8
1 %sql
2 drop table demp
OK
Command took 1.62 seconds -- by shellunext_1693422651722@npunext.onmicrosoft.com at 9/26/2023, 11:15:12 AM on Shellunext's Cluster
Cmd 9
1 %sql
2 use test
OK
Command took 0.11 seconds -- by shellunext_1693422651722@npunext.onmicrosoft.com at 9/26/2023, 11:16:02 AM on Shellunext's Cluster
Cmd 10
1 %sql
2 create table test.demp(id int, name string)


Dataframe
Python
File Edit View Run Help
Last edit was 5 minutes ago
Provide feedback
Run all
Shellunext's Cluster
Schedule
Share
Cmd 1
1 users=[(1,'a',30,"Sales"),(2,'b',25,"IT"),(3,'c',28,"Data Science")]
2 schema="id int, name string, age int, dept string"
3 df=spark.createDataFrame(data=users, schema=schema)
df: pyspark.sql.dataframe.DataFrame
id: integer
name: string
age: integer
dept: string
Command took 0.38 seconds -- by shellunext_1693422651722@npunext.onmicrosoft.com at 9/26/2023, 11:23:38 AM on Shellunext's Cluster
Cmd 2
1 df.show()
(2) Spark Jobs
+---+----+---+------------+
| id|name|age| dept|
+---+----+---+------------+
1	a	30	Sales
2	b	25	IT
3	c	28	Data Science
+---+----+---+------------+
Command took 2.68 seconds -- by shellunext_1693422651722@npunext.onmicrosoft.com at 9/26/2023, 11:23:57 AM on Shellunext's Cluster

```
1  df.display()
```

▸ (2) Spark Jobs

Table +

	id	name	age	dept
1	1	a	30	Sales
2	2	b	25	IT
3	3	c	28	Data Science

3 rows | 0.53 seconds runtime — Refreshed 20 minutes ago

Command took 0.53 seconds -- by shellunext_1693422651722@npunext.onmicrosoft.com at 9/26/2023, 11:24:04 AM on Shellunext's Cluster

Cmd 4

```
1  display(df)
```

▸ (2) Spark Jobs

Table +

	id	name	age	dept
1	1	a	30	Sales
2	2	b	25	IT
3	3	c	28	Data Science

3 rows | 0.39 seconds runtime — Refreshed 20 minutes ago

Command took 0.39 seconds -- by shellunext_1693422651722@npunext.onmicrosoft.com at 9/26/2023, 11:24:18 AM on Shellunext's Cluster

```
1  from pyspark.sql.functions import *
```

Command took 0.10 seconds -- by shellunext_1693422651722@npunext.onmicrosoft.com at 9/26/2023, 11:26:07 AM on Shellunext's Cluster

Cmd 6

```
1  df1=df.withColumnRenamed("id","emp_id").withColumn("current_date",current_date())
```

▾ df1: pyspark.sql.dataframe.DataFrame

```
emp_id: integer
name: string
age: integer
dept: string
current_date: date
```

Command took 0.10 seconds -- by shellunext_1693422651722@npunext.onmicrosoft.com at 9/26/2023, 11:36:53 AM on Shellunext's Cluster

Cmd 7

```
1  display(df1)
```

▸ (2) Spark Jobs

Table +

	emp_id	name	age	dept	current_date
1	1	a	30	Sales	2023-09-26
2	2	b	25	IT	2023-09-26
3	3	c	28	Data Science	2023-09-26

3 rows | 0.66 seconds runtime — Refreshed 7 minutes ago

Command took 0.66 seconds -- by shellunext_1693422651722@npunext.onmicrosoft.com at 9/26/2023, 11:36:55 AM on Shellunext's Cluster

```
1  df1.write.mode("overwrite").saveAsTable("test.emp_table")
```

▸ (6) Spark Jobs

Command took 4.31 seconds -- by shellunext_1693422651722@npunext.onmicrosoft.com at 9/26/2023, 11:38:46 AM on Shellunext's Cluster

Cmd 9

SQL

```
1  %sql
2  use test;
3  select * from emp_table
```

▸ (3) Spark Jobs

▸ _sqldf: pyspark.sql.dataframe.DataFrame = [emp_id: integer, name: string ... 3 more fields]

Table +

	emp_id	name	age	dept	current_date
1	3	c	28	Data Science	2023-09-26
2	1	a	30	Sales	2023-09-26
3	2	b	25	IT	2023-09-26

3 rows | 1.66 seconds runtime — Refreshed 4 minutes ago

This result is stored as PySpark data frame `_sqldf` and in the IPython output cache as `Out[16]`. Learn more

Command took 1.66 seconds -- by shellunext_1693422651722@npunext.onmicrosoft.com at 9/26/2023, 11:40:55 AM on Shellunext's Cluster

Service Principal

Python

File Edit View Run Help Last edit was 3 minutes ago Provide feedback

Run all | Shellunext's Cluster | Schedule | Share

Cmd 2

```
1  dbutils.secrets.listScopes()
```

```
[SecretScope(name='key')]
```

Command took 0.36 seconds -- by shellunext_1693422035770@npunext.onmicrosoft.com at 9/26/2023, 3:15:59 PM on Shellunext's Cluster

Cmd 3

```
1  dbutils.secrets.get(scope="key",key="clientid")
```

```
'[REDACTED]'
```

Command took 0.43 seconds -- by shellunext_1693422035770@npunext.onmicrosoft.com at 9/26/2023, 3:16:27 PM on Shellunext's Cluster

Cmd 4

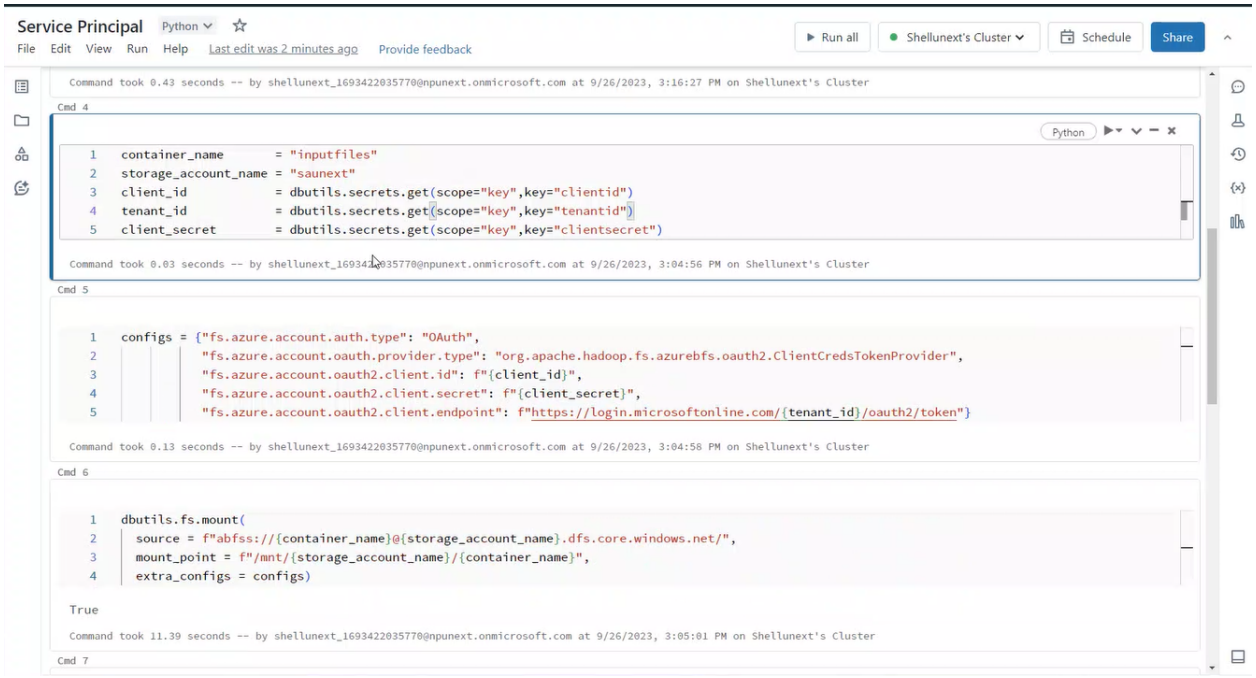
Python

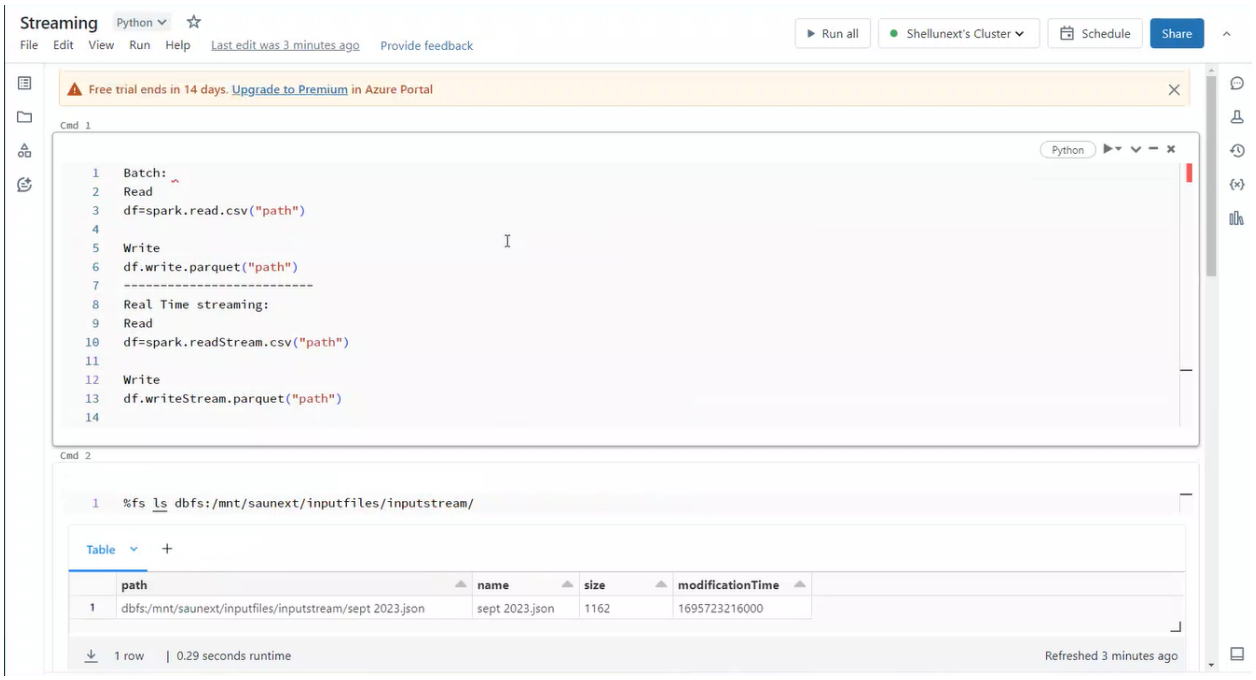
```
1  container_name       = "inputfiles"
2  storage_account_name = "saunext"
3  client_id            = dbutils.secrets.get(scope="key",key="clientid")
4  tenant_id            = dbutils.secrets.get(scope="key",key="tenantid")
5  client_secret        = dbutils.secrets.get(scope="key",key="clientsecret")
```

Command took 1.23 seconds -- by shellunext_1693422035770@npunext.onmicrosoft.com at 9/26/2023, 3:17:33 PM on Shellunext's Cluster

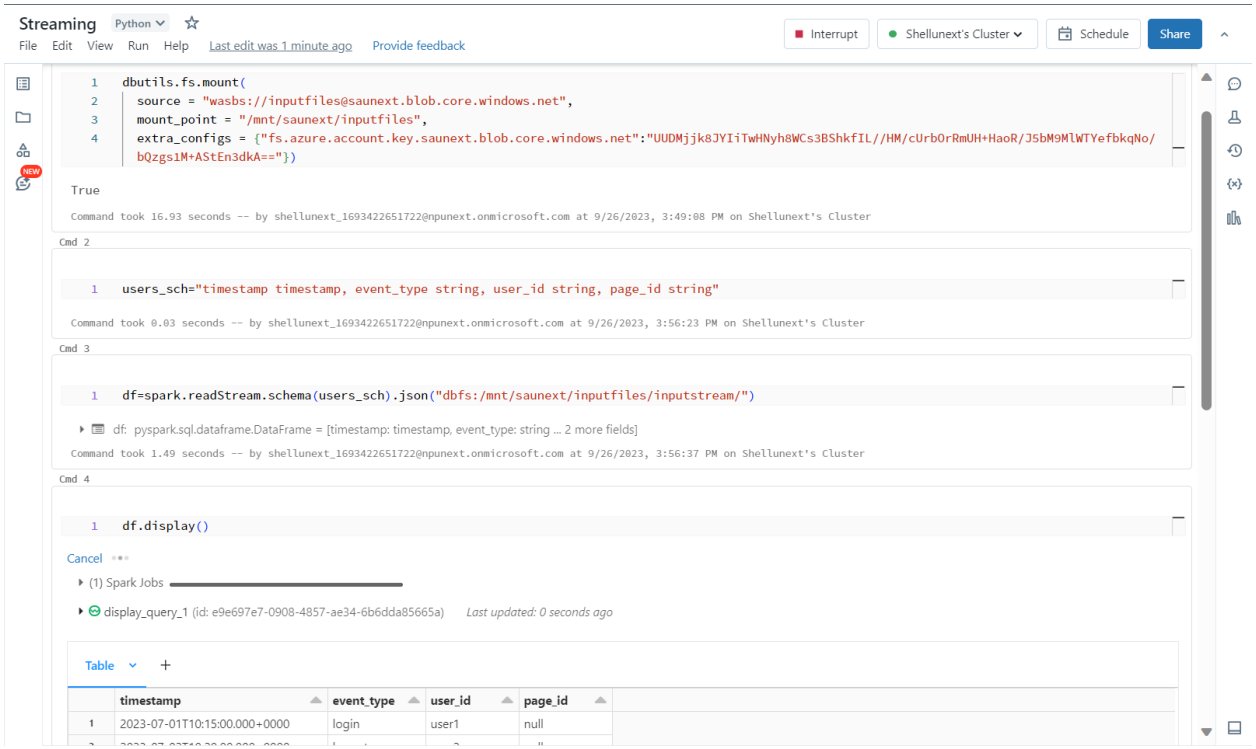
Cmd 5

```
1  configs = {"fs.azure.account.auth.type": "OAuth",
2             "fs.azure.account.oauth.provider.type": "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider",
3             "fs.azure.account.oauth2.client.id": f"{client_id}",
4             "fs.azure.account.oauth2.client.secret": f"{client_secret}",
5             "fs.azure.account.oauth2.client.endpoint": f"https://login.microsoftonline.com/{tenant_id}/oauth2/token"}
```


Service Principal
Python
File Edit View Run Help Last edit was 2 minutes ago Provide feedback
Run all
Shellunext's Cluster
Schedule
Share
Command took 0.43 seconds -- by shellunext_1693422035770@npunext.onmicrosoft.com at 9/26/2023, 3:16:27 PM on Shellunext's Cluster
Cmd 4
container_name = "inputfiles"
storage_account_name = "saunext"
client_id = dbutils.secrets.get(scope="key",key="clientid")
tenant_id = dbutils.secrets.get(scope="key",key="tenantid")
client_secret = dbutils.secrets.get(scope="key",key="clientsecret")
Command took 0.03 seconds -- by shellunext_1693422035770@npunext.onmicrosoft.com at 9/26/2023, 3:04:56 PM on Shellunext's Cluster
Cmd 5
configs = {"fs.azure.account.auth.type": "OAuth",
"fs.azure.account.oauth.provider.type": "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider",
"fs.azure.account.oauth2.client.id": f"{client_id}",
"fs.azure.account.oauth2.client.secret": f"{client_secret}",
"fs.azure.account.oauth2.client.endpoint": f"https://login.microsoftonline.com/{tenant_id}/oauth2/token"}
Command took 0.13 seconds -- by shellunext_1693422035770@npunext.onmicrosoft.com at 9/26/2023, 3:04:58 PM on Shellunext's Cluster
Cmd 6
dbutils.fs.mount(
source = f"abfss://{container_name}@{storage_account_name}.dfs.core.windows.net/",
mount_point = f"/mnt/{storage_account_name}/{container_name}",
extra_configs = configs)
True
Command took 11.39 seconds -- by shellunext_1693422035770@npunext.onmicrosoft.com at 9/26/2023, 3:05:01 PM on Shellunext's Cluster
Cmd 7


Streaming
Python
File Edit View Run Help Last edit was 3 minutes ago Provide feedback
Run all
Shellunext's Cluster
Schedule
Share
Free trial ends in 14 days. Upgrade to Premium in Azure Portal
Cmd 1
Batch:
Read
df=spark.read.csv("path")
Write
df.write.parquet("path")

Real Time streaming:
Read
df=spark.readStream.csv("path")
Write
df.writeStream.parquet("path")
Cmd 2
%fs ls dbfs:/mnt/saunext/inputfiles/inputstream/
Table
path name size modificationTime
1 dbfs:/mnt/saunext/inputfiles/inputstream/sept 2023.json sept 2023.json 1162 1695723216000
1 row | 0.29 seconds runtime
Refreshed 3 minutes ago


Streaming
Python
File Edit View Run Help
Last edit was 1 minute ago
Provide feedback
Interrupt
Shellunext's Cluster
Schedule
Share
1 dbutils.fs.mount(
2 source = "wasbs://inputfiles@saunext.blob.core.windows.net",
3 mount_point = "/mnt/saunext/inputfiles",
4 extra_configs = {"fs.azure.account.key.saunext.blob.core.windows.net":"UUDMjjk8JYIiTwHNyh8WCs3BShkfIL//HM/cUrbOrRmUH+HaoR/J5bM9MlWTYefbkqNo/
bQzgs1M+AStEn3dkA=="})
True
Command took 16.93 seconds -- by shellunext_1693422651722@npunext.onmicrosoft.com at 9/26/2023, 3:49:08 PM on Shellunext's Cluster
Cmd 2
1 users_sch="timestamp timestamp, event_type string, user_id string, page_id string"
Command took 0.03 seconds -- by shellunext_1693422651722@npunext.onmicrosoft.com at 9/26/2023, 3:56:23 PM on Shellunext's Cluster
Cmd 3
1 df=spark.readStream.schema(users_sch).json("dbfs:/mnt/saunext/inputfiles/inputstream/")
df: pyspark.sql.dataframe.DataFrame = [timestamp: timestamp, event_type: string ... 2 more fields]
Command took 1.49 seconds -- by shellunext_1693422651722@npunext.onmicrosoft.com at 9/26/2023, 3:56:37 PM on Shellunext's Cluster
Cmd 4
1 df.display()
Cancel
(1) Spark Jobs
display_query_1 (id: e9e697e7-0908-4857-ae34-6b6dda85665a) Last updated: 0 seconds ago
Table
timestamp
event_type
user_id
page_id
1
2023-07-01T10:15:00.000+0000
login
user1
null