

Mallu Sravana Sandhya

04/01/24

Assessment-26

ETL workload on Azure Databricks

Creating Azure Databricks workspace

Microsoft Azure

Search resources, services, and docs (G+J)

azuser921_mml.local@il...
IHHT (IHHT@ONMICROSOFT.COM)

Home > Azure Databricks >

Create an Azure Databricks workspace

manage all your resources.

Subscription *

Azure subscription 1

Resource group *

rg-azuser921_mml.local-yvpvZ

Create new

Instance Details

Workspace name *

AzureDB-921

Region *

Central India

Pricing Tier *

Premium (+ Role-based access controls)

We selected the recommended pricing tier for your workspace. You can change the tier based on your needs.

Managed Resource Group name

AzureDB-921

Review + create

< Previous

Next : Networking >

Microsoft Azure

Search resources, services, and docs (G+J)

azuser921_mml.local@il...
IHHT (IHHT@ONMICROSOFT.COM)

Home > Azure Databricks >

Create an Azure Databricks workspace

Validation Succeeded

Summary

Basics

Workspace name

AzureDB-921

Subscription

Azure subscription 1

Resource group

rg-azuser921_mml.local-yvpvZ

Region

Central India

Pricing Tier

premium

Managed Resource Group name

AzureDB-921

Networking

Deploy Azure Databricks workspace with Secure Cluster Connectivity (No Public IP)

No

Deploy Azure Databricks workspace in your own Virtual Network (VNet)

No

Create

< Previous

Download a template for automation

Waiting for management.azure.com...

Submitting deployment...

Submitting the deployment template for resource group 'rg-azuser921_mml.local-yvpvZ'.

30°C

Partly sunny

Search

ENG IN

13:57

04-01-2024

Microsoft Azure portal showing the overview of the AzureDB-921 Azure Databricks Service. The page displays the status (Active), resource group (rg-azuser921_mml.local-yVpvZ), location (Central India), subscription (Azure subscription 1), and subscription ID (984f097c-963c-4eb6-a20d-839457ae9f08). A large red Databricks logo is centered, with a "Launch Workspace" button below it. The left sidebar shows navigation options like Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Settings, Virtual Network Peering, Encryption, Networking, Properties, Locks, Monitoring, and Diagnostic settings. The top navigation bar includes the Microsoft Azure logo and a search bar.

Creating cluster in Azure Databricks

Azure Databricks workspace interface showing the "Get started with Databricks" onboarding screen. The screen includes a "Start the SQL warehouse" section with a "Start warehouse" button, and an "Explore sample projects" section with links to "SQL and visualizations", "KPI dashboard", and "Data analysis in Python". The left sidebar shows navigation options like New, Workspace, Recents, Catalog, Workflows, Compute, SQL, SQL Editor, Queries, Dashboards, Alerts, Query History, SQL Warehouses, Data Engineering, Job Runs, Data Ingestion, and Delta Live Tables. The top navigation bar includes the Microsoft Azure logo and a search bar.

Microsoft Azure

databricks

Search data, notebooks, recents, and more...

CTRL + P

AzureDB-921

azuser921_mml.local@iihtl.onmicro...

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

Compute > UI preview > Send feedback

azuser921_mml.local@iihtl.onmicrosoft.com's Cluster

More

Terminate

Edit

Configuration

Notebooks (0)

Libraries

Event log

Spark UI

Driver logs

Metrics

Apps

Spark compute UI - Master

Policy

Personal Compute

Access mode

Single user access

Single user

azuser921_mml.local@iihtl.onmicrosoft.c...

Performance

Databricks Runtime Version

13.3 LTS (includes Apache Spark 3.4.1, Scala 2.12)

13.3 LTS (includes Apache Spark 3.4.1, Scala 2.12)

Use Photon Acceleration

Node type

Standard_DS3_v2

14 GB Memory, 4 Cores

Terminate after

4320

minutes of inactivity

Tags

Summary

1 Driver

14 GB Memory, 4 Cores

Runtime

13.3.x-scala2.12

Standard_DS3_v2

0.75 DBU/h

UI | JSON

30°C

Partly sunny

Search

ENG IN

14:02

04-01-2024

Microsoft Azure

databricks

Search data, notebooks, recents, and more...

CTRL + P

AzureDB-921

azuser921_mml.local@iihtl.onmicro...

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

Compute

All-purpose compute

Job compute

SQL warehouses

Pools

Policies

Filter compute you have...

Created by

Create with Personal Compute

Create compute

State	Name	Policy	Runtime	Active m...	Active co...	Active D...	Source	Creator	Notebooks
	azuser921_mml.local@iihtl.onmi...	Personal Comp	13.3	14 GB	4 cores	0.75	UI	azuser921_mml...	

1

20 / page

30°C

Partly sunny

Search

ENG IN

14:05

04-01-2024

Creating notebook to write the code

The screenshot shows the Databricks web interface. The 'New' menu is open, displaying options like Workspace, Recents, Catalog, Workflows, Compute, SQL, SQL Editor, Queries, Dashboards, Alerts, Query History, SQL Warehouses, Data Engineering, Job Runs, Data Ingestion, and Delta Live Tables. The 'Compute' option is selected, leading to a page with a 'Create with Personal Compute' button and a table of existing clusters.

Cluster	Policy	Runtime	Active m...	Active co...	Active D...	Source	Creator	Notebooks
SQL Warehouse	Personal Comp	13.3	14 GB	4 cores	0.75	UI	azuser921_mml...	

Configuring autoloader to ingest data to delta lake

The screenshot shows a Databricks notebook titled 'Untitled Notebook 2024-01-04 14:06:00'. The code in the notebook is as follows:

```
6 username = spark.sql("SELECT regexp_replace(current_user(), '[_a-zA-Z0-9]', '_').first()[0]
7 table_name = f'{username}_etl_quickstart'
8 checkpoint_path = f'/tmp/{username}/_checkpoint/etl_quickstart'
9
10 # Clear out data from previous demo execution
11 spark.sql(f'DROP TABLE IF EXISTS {table_name}')
12 dbutils.fs.rm(checkpoint_path, True)
13
14 # Configure Auto Loader to ingest JSON data to a Delta table
15 (spark.readStream
16   .format("cloudFiles")
17   .option("cloudFiles.format", "json")
18   .option("cloudFiles.schemaLocation", checkpoint_path)
19   .load(file_path)
20   .select("?", col("_metadata.file_path").alias("source_file"), current_timestamp().alias("processing_time"))
21   .writeStream
22   .option("checkpointLocation", checkpoint_path)
23   .trigger(availableNow=True)
24   .toTable(table_name))
```

The notebook also shows a Spark job running, with the output indicating the job is a streaming query.

To query the table that was just created

The screenshot shows a Databricks notebook interface. The left sidebar contains navigation options like Workspace, Recents, Catalog, Workflows, Compute, SQL, SQL Editor, Queries, Dashboards, Alerts, Query History, SQL Warehouses, Data Engineering, Job Runs, Data Ingestion, Delta Live Tables, and Machine Learning. The main area displays a notebook titled "Untitled Notebook 2024-01-04 14:06:00" with a Python kernel. A code cell contains the following code:

```
<pyspark.sql.streaming.query.StreamingQuery at 0x7fc937f4fe80>
Command complete

Cmd 2

1 df = spark.read.table(table_name)
```

The output shows the schema of the DataFrame:

```
df: pyspark.sql.dataframe.DataFrame
Schema Details History
action: string
time: string
_rescued_data: string
source_file: string
processing_time: timestamp
```

Below the schema, it indicates the command took 0.29 seconds and was executed by azuser921_mml.local@ih1l.onmicrosoft.com on 1/4/2024 at 2:08:09 PM on the azuser921_mml.local@ih1l.onmicrosoft.com's Cluster. The bottom status bar shows the system temperature as 30°C and the time as 14:08 on 04-01-2024.

To preview the table

The screenshot shows the same Databricks notebook interface. The code cell now contains:

```
_rescued_data: string
source_file: string
processing_time: timestamp

Cmd 3

1 display(df)
```

The output displays a table with 7 rows and 5 columns. The columns are action, time, _rescued_data, source_file, and processing_time. The data is as follows:

	action	time	_rescued_data	source_file	processing_time
1	Close	1469679568	null	/databricks-datasets/structured-streaming/events/file-49.json	2024-01-04T08:37:12.356Z
2	Close	1469679568	null	/databricks-datasets/structured-streaming/events/file-49.json	2024-01-04T08:37:12.356Z
3	Open	1469679569	null	/databricks-datasets/structured-streaming/events/file-49.json	2024-01-04T08:37:12.356Z
4	Close	1469679571	null	/databricks-datasets/structured-streaming/events/file-49.json	2024-01-04T08:37:12.356Z
5	Open	1469679571	null	/databricks-datasets/structured-streaming/events/file-49.json	2024-01-04T08:37:12.356Z
6	Open	1469679571	null	/databricks-datasets/structured-streaming/events/file-49.json	2024-01-04T08:37:12.356Z
7	Close	1469679572	null	/databricks-datasets/structured-streaming/events/file-49.json	2024-01-04T08:37:12.356Z

Below the table, it indicates 10,000 rows, truncated data, and a 2.08 seconds runtime. The bottom status bar shows the system temperature as 30°C and the time as 14:09 on 04-01-2024.

Creating a Schedule to perform the job

The screenshot displays the Databricks workspace interface. The top navigation bar includes the Microsoft Azure logo, the Databricks logo, a search bar, and the user profile 'azuser921_mml.local@iitl.onmicro...'. The left sidebar shows the workspace structure: New, Workspace, Recents, Catalog, Workflows, Compute, SQL, SQL Editor, Queries, Dashboards, Alerts, Query History, SQL Warehouses, Data Engineering, Job Runs, Data Ingestion, and Delta Live Tables.

The main area shows the 'ETL workload' job configuration. The job name is 'ETL workload'. The schedule is set to 'Manual' with a frequency of 'Every Day' at '14:48' in the 'Asia/Calcutta' time zone. The cluster is 'azuser921_mml.local@iitl.onmicrosoft.com's Cluster 14 GB · 4 Cores · DBR 13.3 LTS · S...'. The parameters are empty. The alerts are 'azuser921_mml.local@iitl...' with checkboxes for Start, Success, and Failure. The job status is 'Job - Paused - ETL workload' with 'Last run: No runs'. The 'Run now' button is visible.

The job execution results are shown below the configuration. The command took 2.08 seconds. The results table has 10,000 rows and 2.08 seconds runtime. The table columns are 'action', 'time', '_rescued_data', 'source_file', and 'processing_time'. The data is as follows:

action	time	_rescued_data	source_file	processing_time
1 Close	1469679568	null	/databricks-datasets/structured-streaming/events/file-49.json	2024-01-04T08:37:12.356Z
2 Close	1469679568	null	/databricks-datasets/structured-streaming/events/file-49.json	2024-01-04T08:37:12.356Z
3 Open	1469679569	null	/databricks-datasets/structured-streaming/events/file-49.json	2024-01-04T08:37:12.356Z
4 Close	1469679571	null	/databricks-datasets/structured-streaming/events/file-49.json	2024-01-04T08:37:12.356Z
5 Open	1469679571	null	/databricks-datasets/structured-streaming/events/file-49.json	2024-01-04T08:37:12.356Z
6 Open	1469679571	null	/databricks-datasets/structured-streaming/events/file-49.json	2024-01-04T08:37:12.356Z
7 Close	1469679572	null	/databricks-datasets/structured-streaming/events/file-49.json	2024-01-04T08:37:12.356Z

The bottom status bar shows the system clock as 14:49 on 04-01-2024.

Microsoft Azure databricks

Search data, notebooks, recents, and more... CTRL + P

AzureDB-921 azuser921_mml.local@ihitl.onmicro...

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

Workflows > Jobs > ETL workload ☆

Runs Tasks

Runs

Start date < Previous Next >

There are no runs for this job yet

Run now

Job details

Job ID 951357206085456

Creator @ azuser921_mml.local@ihitl.onmicro...

Run as @ azuser921_mml.local@ihitl.on...

Tags +Tag

Description Add description

Git Not configured

Add Git settings

Schedule Paused - At 02:48 PM (UTC+05:30 — undefined)

Edit schedule Resume Delete

While running Scheduled job

Microsoft Azure databricks

Search data, notebooks, recents, and more... CTRL + P

AzureDB-921 azuser921_mml.local@ihitl.onmicro...

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

Workflows > Jobs > ETL workload ☆

Runs Tasks

Runs

Run total duration

19s

9s

ETL_workload

Tasks

Cancel runs

Start time	Run ID	Launched	Duration	Spark	Status	Run paramet...
Jan 04, 2024, 0...	1759646...	Manually	21s	Spark UI / Logs / Metrics	Runni...	

Job details

Job ID 951357206085456

Creator @ azuser921_mml.local@ihitl.onmicro...

Run as @ azuser921_mml.local@ihitl.on...

Tags +Tag

Description Add description

Git Not configured

Add Git settings

Schedule Paused - At 02:48 PM (UTC+05:30 — undefined)

Edit schedule Resume Delete

After successful run of scheduled job

The screenshot displays the Databricks web interface. The top navigation bar includes the Microsoft Azure logo, the Databricks logo, a search bar, and the user profile 'azuser921_mml.local@iihtl.onmicro...'. The left sidebar contains navigation options: New, Workspace, Recents, Catalog, Workflows (selected), Compute, SQL, SQL Editor, Queries, Dashboards, Alerts, Query History, SQL Warehouses, Data Engineering, Job Runs, Data Ingestion, Delta Live Tables, and Machine Learning.

The main content area is titled 'ETL workload' and shows a 'Runs' tab. A chart displays the 'Run total duration' for 'ETL_workload' on 'Jan 04', with a green bar indicating a duration of 58s. Below the chart, a table lists the job runs:

Start time	Run ID	Launched	Duration	Spark	Status	Run paramet...
Jan 04, 2024, 0...	1759646...	Manually	59s	Spark UI / Logs / Metrics	Success	

The right sidebar shows 'Job details' for the selected run, including Job ID, Creator, Run as, Tags, Description, Git settings, and Schedule. The schedule is currently 'Paused - At 02:48 PM (UTC+05:30 — undefined)'.

The bottom of the image shows a Windows taskbar with the date '04-01-2024' and time '14:51'.