

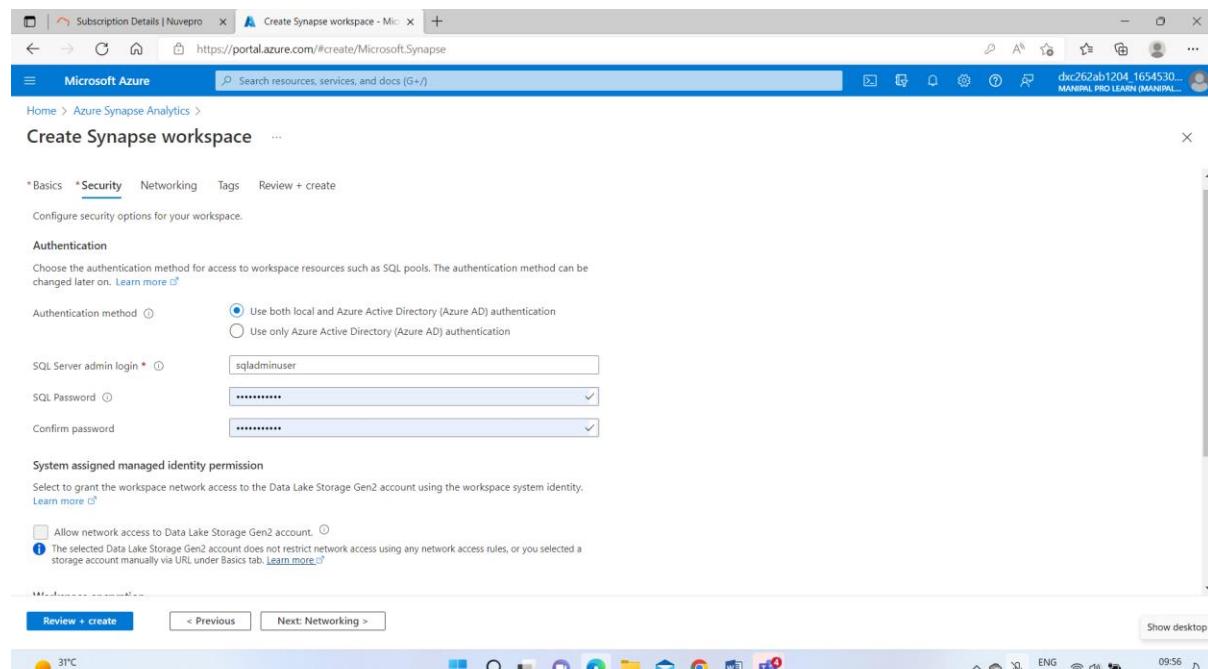
ASSIGNMENT 9

NAME: Aishee Bhattacharya

BATCH: DXC-262-Analytics-B12-Azure

DATE: 09/06/2022

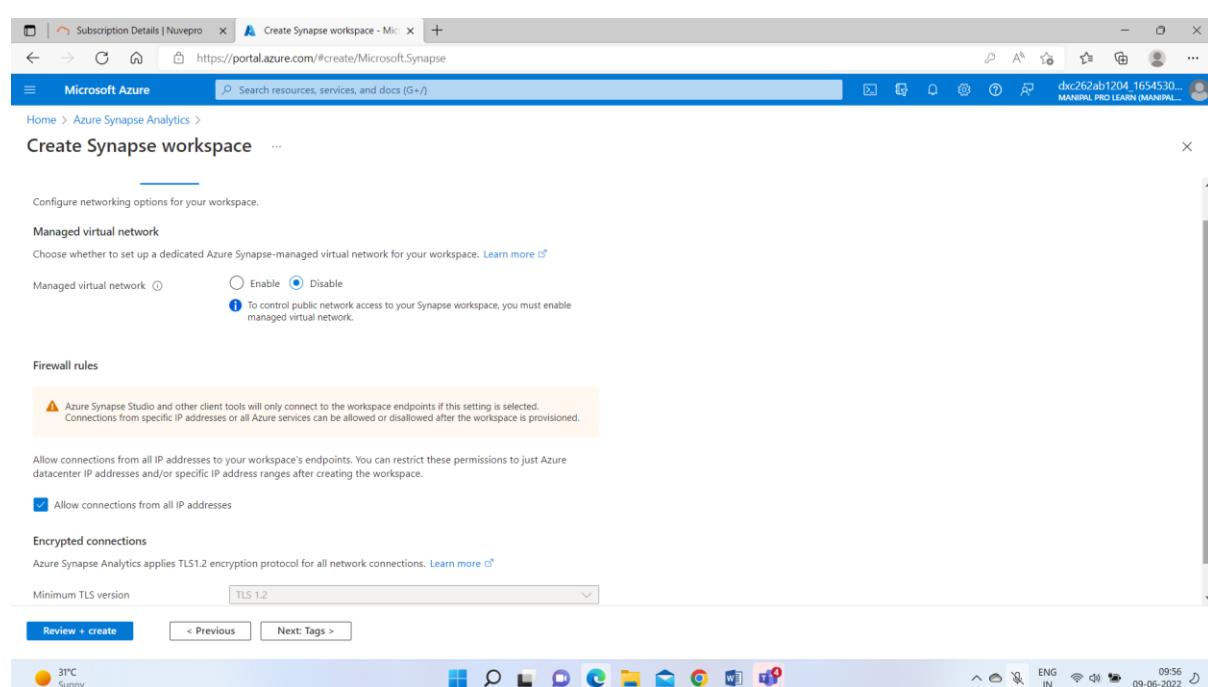
1. Explain the steps with screenshots how to AzureSynapse analytics?



The screenshot shows the 'Create Synapse workspace' wizard on the Microsoft Azure portal. The current step is 'Security'. The configuration includes:

- Authentication:** Use both local and Azure Active Directory (Azure AD) authentication.
- SQL Server admin login:** sqldadminuser
- SQL Password:** [REDACTED]
- Confirm password:** [REDACTED]

A note states: "Select to grant the workspace network access to the Data Lake Storage Gen2 account using the workspace system identity." There is a warning: "Allow network access to Data Lake Storage Gen2 account. The selected Data Lake Storage Gen2 account does not restrict network access using any network access rules, or you selected a storage account manually via URL under Basics tab." Buttons at the bottom include 'Review + create', '< Previous', and 'Next: Networking >'.



The screenshot shows the 'Create Synapse workspace' wizard on the Microsoft Azure portal. The current step is 'Networking'. The configuration includes:

- Managed virtual network:** Enable (selected)
- Firewall rules:** A warning message: "Azure Synapse Studio and other client tools will only connect to the workspace endpoints if this setting is selected. Connections from specific IP addresses or all Azure services can be allowed or disallowed after the workspace is provisioned."
- Allow connections from all IP addresses:** Selected

Other settings include 'Encrypted connections' (TLS 1.2 selected) and 'Minimum TLS version' (TLS 1.2). Buttons at the bottom include 'Review + create', '< Previous', and 'Next: Tags >'.

Subscription Details | Nuvepro | Create Synapse workspace - Microsoft Azure | https://portal.azure.com/#create/Microsoft.Synapse

Microsoft Azure | Search resources, services, and docs (G+) | dxc262ab1204_1654530... | MANIPAL PRO LEARN (MANIPAL)

Home > Azure Synapse Analytics > Create Synapse workspace ...

Validation succeeded

* Basics * Security Networking Tags Review + create

Product Details

Azure Synapse Analytics workspace by Microsoft Serverless SQL est. cost/TB ⓘ

Terms of use | Privacy policy

By clicking Create, I (a) agree to the legal terms and privacy statement(s) associated with the Marketplace offering(s) listed above; (b) authorize Microsoft to bill my current payment method for the fees associated with the offering(s), with the same billing frequency as my Azure subscription; and (c) agree that Microsoft may share my contact, usage and transactional information with the provider(s) of the offering(s) for support, billing and other transactional activities. Microsoft does not provide rights for third-party offerings. For additional details see [Azure Marketplace Terms](#). ⓘ

Basics

Subscription	Azure-DXC262AB12Lab
Resource group	(new) dxcrg1224
Region	East US
Workspace name	(new) dxcsynapse1220

Create < Previous Next > Download a template for automation

31°C Sunny ENG IN 09:57 09-06-2022

Subscription Details | Nuvepro | Microsoft.Azure.SynapseAnalytics | Overview - Microsoft Azure | https://portal.azure.com/#view/HubsExtension/DeploymentDetailsBlade/~/overview/id%2Fsubscriptions%2F4780d990-8511-4a8c-a667-d8eb89f11900%2Fres... | dxc262ab1204_1654530... | MANIPAL PRO LEARN (MANIPAL)

Home > Microsoft.Azure.SynapseAnalytics-20220609094953 | Overview ...

Deployment

Search (Ctrl+ /) Delete Cancel Redeploy Refresh

We'd love your feedback! →

Your deployment is complete

Deployment name: Microsoft.Azure.SynapseAnalytics-20220609094953 Start time: 6/9/2022, 9:58:16 AM

Subscription: Azure-DXC262AB12Lab Correlation ID: a4be6158-ccba-41a0-9c4c-0048fd618785

Resource group: dxcrg1224

Deployment details (Download) Next steps Go to resource group

Deployment succeeded Deployment 'Microsoft.Azure.SynapseAnalytics-20220609094953' to resource group 'dxcrg1224' was successful.

Pin to dashboard Go to resource group

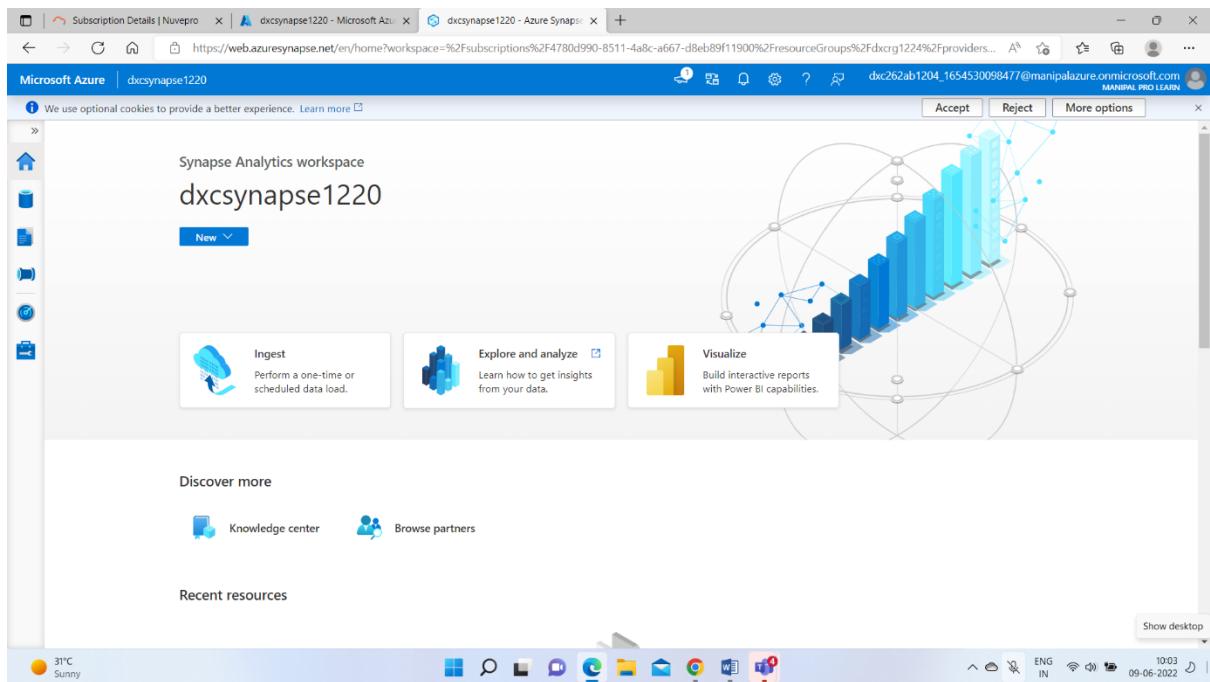
Cost Management Get notified to stay within your budget and prevent unexpected charges on your bill. Set up cost alerts >

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 >

31°C Sunny ENG IN 10:02 09-06-2022



2. Explain the steps with screenshots how to SQL Pool in AzureSynapse analytics?

id	updated	confirmed	confirmed_change	deaths	deaths_change	recovered
338995	2020-01-21T00:00:00Z	262	(NULL)	0	(NULL)	(NULL)
338996	2020-01-22T00:00:00Z	313	51	0	0	(NULL)
338997	2020-01-23T00:00:00Z	578	265	0	0	(NULL)
338998	2020-01-24T00:00:00Z	841	263	0	0	(NULL)
338999	2020-01-25T00:00:00Z	1320	479	0	0	(NULL)

The screenshot shows the Microsoft Azure portal interface for an Azure Synapse workspace. The left sidebar navigation includes options like Analytics pools, External connections, Integration, Security, Configurations + libraries, and Data flow libraries. The main content area is titled 'SQL pools' and shows a single entry: 'Built-in' (Type: Serverless, Status: Online, Size: Auto). A 'Filter by name' input field is present at the top of the list.

3. Explain the steps with screenshots how to import COVID19 dataset in AzureSynapse analytics. and run sample 500 rows & dispaly the output?

The screenshot shows the Microsoft Azure portal interface for an Azure Synapse workspace. The left sidebar navigation includes options like Database templates, Datasets, Notebooks, SQL scripts, and Pipelines. The main content area is titled 'Datasets' and shows a grid of available datasets. The datasets listed are:

- Bing COVID-19 Data
- Boston Safety Data
- COVID Tracking Project
- Chicago Safety Data
- European Centre for Disease Prevention and Control (ECDC) COVID-19 Cases
- NOAA Integrated Surface Data (ISD)
- NYC Taxi & Limousine Commission - For-Hire Vehicle (FHV) trip records
- NYC Taxi & Limousine Commission - green taxi trip records
- NYC Taxi & Limousine Commission - yellow taxi trip records
- New York City Safety Data

Each dataset entry includes a brief description, an icon, and a unique ID.

The screenshot shows the Microsoft Azure Synapse Studio interface. On the left, there's a navigation pane with icons for Home, Data, Pipelines, Databricks, and Notebooks. The 'Data' section is selected, showing a tree view of resources under 'Linked'. A context menu is open over a dataset named 'bing-covid-19-data'. The menu options include 'New SQL script', 'Select TOP 100 rows' (which is highlighted), 'Create external table', 'Edit', 'Delete', and 'Properties'. At the top of the screen, there are tabs for 'Subscription Details | Nuvepro', 'dxcynapse1220 - Microsoft Az...', and 'dxcynapse1220 - Azure Synapse'. The address bar shows the URL: https://web.azuresynapse.net/en/authoring/explore/linked/blobstorage/opendatasets%252Fbing-covid-19-data?workspace=%2Fsubscriptions%2F4780d990-8511-4... . The status bar at the bottom right shows the date as 09-06-2022 and the time as 10:13.

This screenshot shows the Microsoft Azure Synapse Studio interface with a pipeline run. The pipeline is named 'Pipeline 1'. In the center, there's a 'SQL script 4' tab with the following code:

```
1 -- This is auto-generated code
2
3 SELECT
4     TOP_500 *
5 FROM
6     OPENROWSET(
7         BULK 'https://pandemictodata.blob.core.windows.net/public/curated/covid-19/bing_covid-19_data.parquet'
8     ) AS [result];
```

To the right of the code editor, there's a 'Properties' panel with tabs for 'General' and 'Related (0)'. Under 'General', the 'Name' is set to 'SQL script 4' and the 'Type' is listed as 'sql script'. Below the properties, there are sections for 'Size' (262 bytes), 'Results settings per query (0)', and two radio button options: 'First 5000 rows (default)' (selected) and 'All rows'. At the bottom of the screen, there's a status bar showing the date as 09-06-2022 and the time as 16:16.

4. Explain the steps with screenshots how to input Boston Safety datasets into AzureSynapse analytics using Notebooks ?

The screenshot shows the Microsoft Azure Datasets page. The 'Datasets' tab is selected. A search bar at the top right contains the text 'Search'. Below the search bar, there are tabs for 'Database templates', 'Datasets', 'Notebooks', 'SQL scripts', and 'Pipelines'. A 'Filter by keyword' input field and a 'Tags : All' dropdown are also present. The main area displays several dataset cards:

- Bing COVID-19 Data**: Bing COVID-19 data includes confirmed, fatal, and recovered cases from all regions, updated daily. ID: bing-covid-19-data
- Boston Safety Data**: Read data about 311 calls reported to the city of Boston. This dataset is stored in Parquet format and is updated daily. ID: city_safety_boston
- COVID Tracking Project**: The COVID Tracking Project dataset provides the latest numbers on tests, confirmed cases, hospitalizations, and patient outcomes from every US state and...
- Chicago Safety Data**: Read data about 311 calls reported to the city of Chicago. This dataset is stored in Parquet format and is updated daily. ID: covid-tracking
- European Centre for Disease Prevention and Control (ECDC) Covid-19 Cases**: The latest available public data on geographic distribution of COVID-19 cases worldwide from the Euro...
- NOAA Integrated Surface Data (ISD)**: NOAA Integrated Surface Data (ISD) provides Worldwide hourly weather history data sourced from the National Oceanic and Atmospheric...
- NYC Taxi & Limousine Commission - For-Hire Vehicle (FHV) trip records**: The For-Hire Vehicle trip records include fields capturing the dispatching base license number a...
- NYC Taxi & Limousine Commission - green taxi trip records**: The green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop...
- NYC Taxi & Limousine Commission - yellow taxi trip records**: The yellow taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop...
- New York City Safety Data**: This dataset contains all New York City 311 service requests from 2010 to the present. It's stored in Parquet format and updated daily. ID: city_safety_newyork

At the bottom left is a 'Continue' button, and at the bottom right is a 'Close' button. The status bar at the bottom shows the date '09-06-2022' and time '18:08'.

The screenshot shows the Microsoft Azure Notebook page. The 'Notebook' tab is selected. The interface includes a toolbar with 'Accept', 'Reject', and 'More options' buttons, and a status bar at the bottom showing '36°C Cloudy' and the date '09-06-2022'.

The left sidebar shows the 'Data' section with 'Workspace' and 'Linked' tabs. Under 'Linked', there are sections for 'Sample Datasets' (bing-covid-19-data, city_safety_boston, nyc_tlc_yellow) and 'Azure Data Lake Storage Gen2' (dxcynapse1220). The status bar at the bottom shows the date '09-06-2022' and time '16:55'.

The main workspace contains a code editor with the following PySpark code:

```

1 from azureml.opendatasets import BostonSafety
2
3 data = BostonSafety()
4 df = data.to_spark_dataframe()
5 # Display 10 rows
6 df.write.mode("overwrite").saveAsTable("default.YourTableName")

```

The code editor has tabs for 'Code' and 'Markdown'.

5. Explain the steps with screenshots how to create Spark pool in AzureSynapse analytics?

The screenshot shows the Azure portal interface for creating an Apache Spark pool. The left sidebar navigation includes Synapse live, Validate all, Publish all, Analytics pools, SQL pools, Apache Spark pools (selected), Data Explorer pools (pre...), External connections, Linked services, Microsoft Purview, Integration, Triggers, Integration runtimes, Security, Access control, Credentials, Managed private endpoints, Configurations + libraries, Workspace packages, Data flow libraries (previ...), and Apache Spark configurat... The main content area is titled 'New Apache Spark pool' under the 'Basics' tab. It prompts the user to create an Apache Spark pool with configurations. The 'Apache Spark pool name' field is set to 'dxcsparkpool1'. The 'Isolated compute' setting is 'Disabled'. The 'Node size family' is 'Memory Optimized'. The 'Node size' is 'Small (4 vCores / 32 GB)'. The 'Autoscale' setting is 'Enabled'. The 'Number of nodes' is set to 3. The 'Estimated price per hour' section shows an error message: 'Failed to fetch billing info'. The 'Dynamically allocate executors' setting is 'Disabled'. At the bottom, there are 'Review + create' and 'Next: Additional settings >' buttons, along with a 'Cancel' button.

The screenshot shows the Azure portal interface after the Apache Spark pool has been created. The left sidebar navigation is identical to the previous screenshot. The main content area displays the 'Apache Spark pool' list. A single entry is shown: 'Name' is 'dxcsparkpool1', 'Node size family' is 'Memory Optimized', and 'Size' is 'Small (4 vCores / 32 GB) - 3 to 3 nodes'. The status bar at the bottom indicates it's 11:38 on 09-06-2022.

The screenshot shows the Microsoft Azure Synapse Analytics Notebook interface. On the left, the 'Data' sidebar is open, showing 'Linked' resources: 'Azure Blob Storage' (containing 'bing-covid-19-data', 'city_safety_boston', and 'nyc_tlc_yellow') and 'Azure Data Lake Storage Gen2' (containing 'dxcsynapse1220'). The main area displays a PySpark notebook titled 'Notebook 1' attached to 'dxsparkpool1'. The code cell contains the following PySpark code:

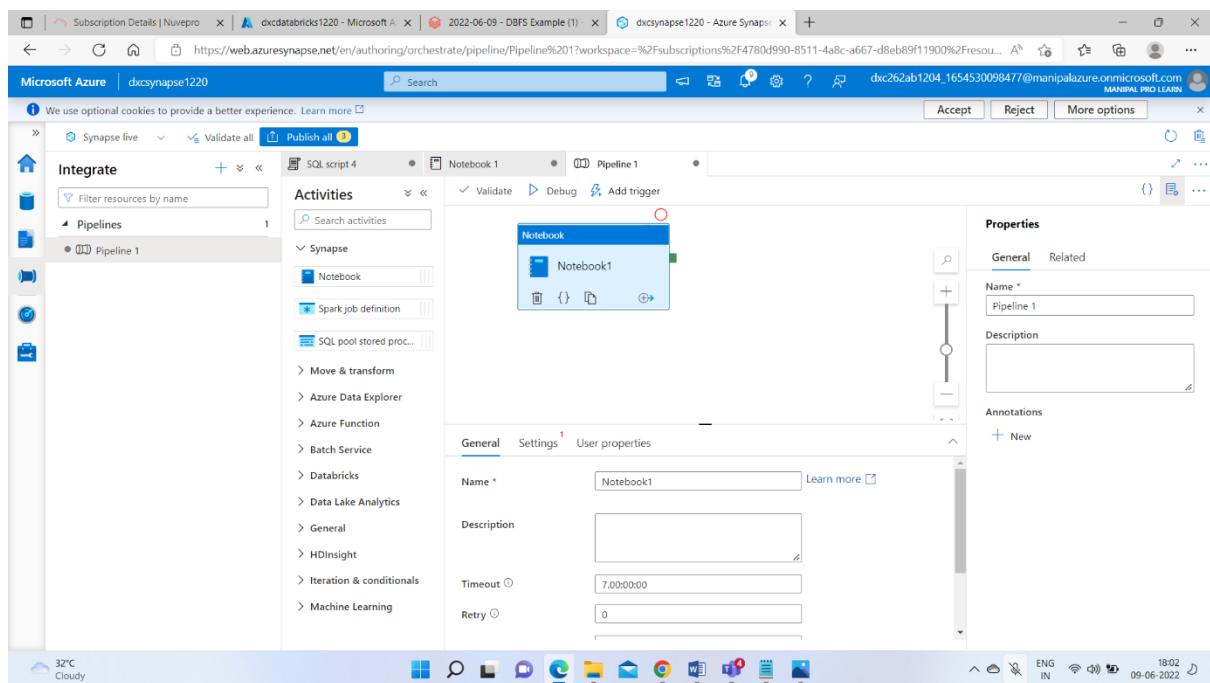
```

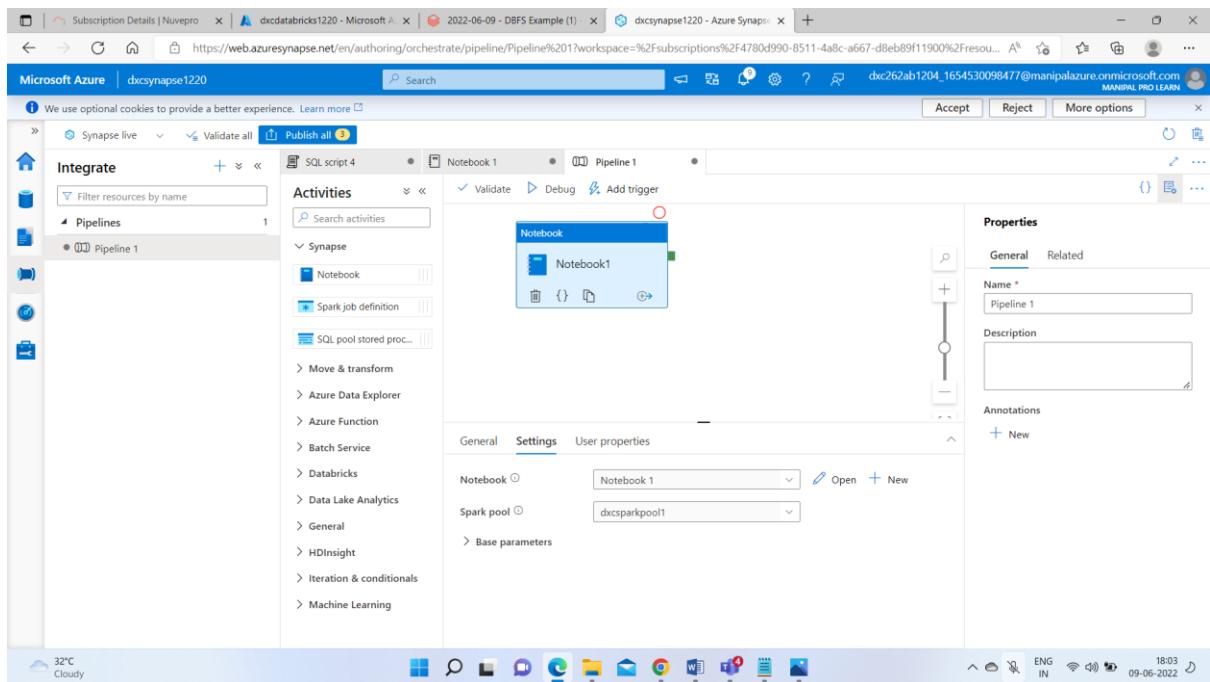
1 from azureml.opendatasets import NycTlcYellow
2
3 data = NycTlcYellow()
4 df = data.to_spark_dataframe()
5 # Display 10 rows
6 df.write.mode("overwrite").saveAsTable("default.YourTableName")

```

The status bar at the bottom indicates it's 35°C and sunny.

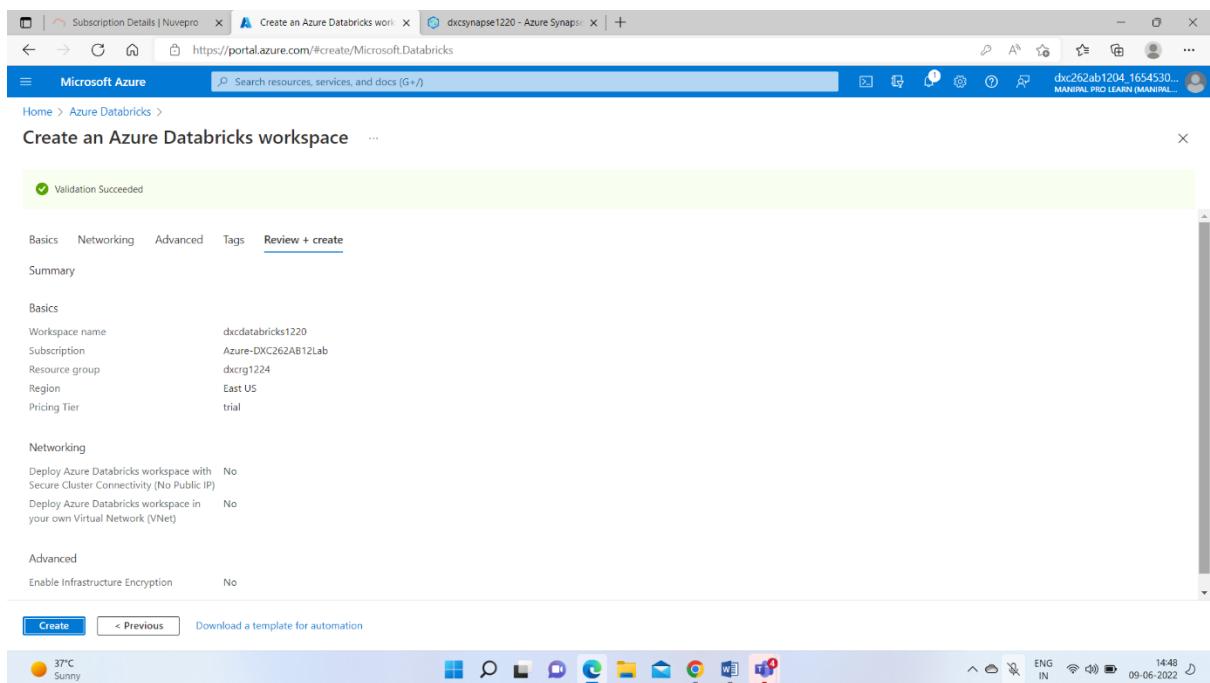
6. Explain the steps with screenshots how to create pipeline in AzureSynapse analytics?





8. Explain the steps with screenshots how to Databricks ?

- Scalable analytics performed in azure
- Based on Apache spark
- Workflows and workspaces for data users
- Native integration with other azure services



Subscription Details | Nuverro x dxcdatabricks1220 - Microsoft Azure x dcsynapse1220 - Azure Synapse +

https://portal.azure.com/#@manipalazure.onmicrosoft.com/resource/subscriptions/4780d990-8511-4a8c-a667-d8eb89f11900/resourceGroups/dxcrg1224/prov... ↗ ⓘ

Microsoft Azure Search resources, services, and docs (G+)

Home > dxcrg1224_dxcdatabricks1220 >

dxcdatabricks1220 Azure Databricks Service

Search (Ctrl+F) Delete

Overview Activity log Access control (IAM) Tags

Essentials

Status : Active
Resource group : [dxcrg1224](#)
Location : East US
Subscription : [Azure-DXC262A812Lab](#)
Subscription ID : 4780d990-8511-4a8c-a667-d8eb89f11900

Managed Resource Group : [databricks-rg-dxcdatabricks1220-7l5hr62lyftqw](#)
URL : <https://adb-25947125729108.azuredatabricks.net>
Pricing Tier : Trial (Premium - 14-Days Free DBUs)

Virtual Network Peerings Tags (edit) Click here to add tags

JSON View

Automation

Tasks (preview)
Export template

Support + troubleshooting

New Support Request

Launch Workspace

Upgrade to Premium

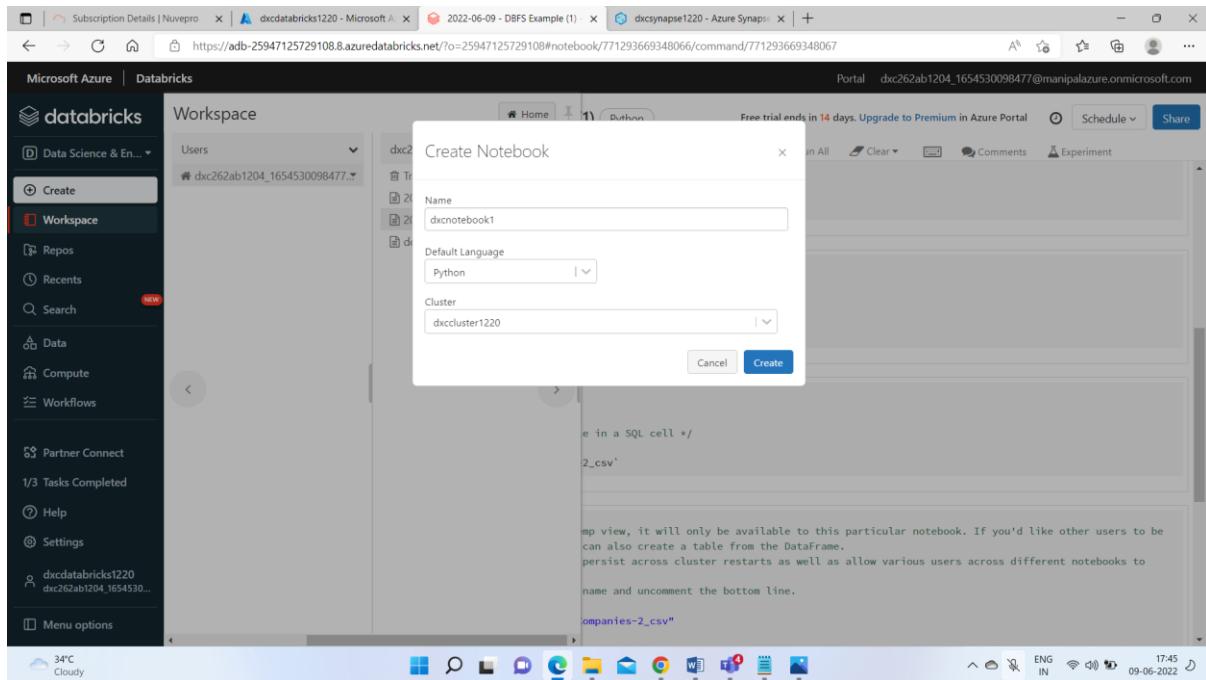
Documentation Getting Started Import Data from File Import

37°C Sunny

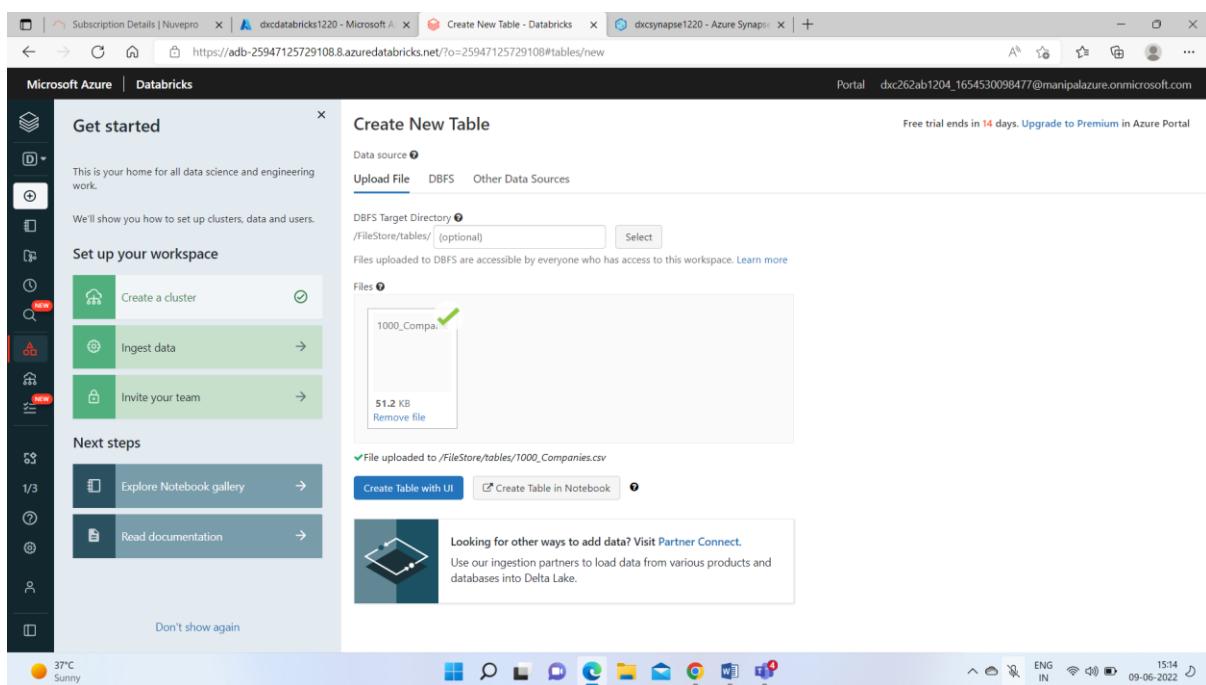
General 05:58:59 RJ

9. Explain the steps with screenshots how to create notebooks in Databricks ?

The screenshot shows the Microsoft Azure Databricks workspace interface. On the left, there's a sidebar with various navigation options like Data Science & Engineering, Create, Workspace, Repos, Recents, Search, Data, Compute, Workflows, Partner Connect, Tasks Completed, Help, Settings, and Menu options. The main area is titled 'Workspace' and shows a list of notebooks: '2022-06-09 - DBFS Example', '2022-06-09 - DBFS Example (2)', and 'demonotebook1'. A context menu is open over the first notebook, with 'Python' selected as the language. The menu includes options: Create (Notebook, Library, Folder, MLflow Experiment), Clone, Import, Export, Permissions, and Copy Link Address. The URL in the browser bar is <https://adb-25947125729108.azuredatabricks.net/?o=25947125729108#notebook/771293669348066/command/771293669348067>.



10. Explain the steps with screenshots how to insert data into databricks notebook & display the result?



The screenshot shows a Microsoft Azure Databricks notebook interface. On the left, a sidebar titled "Get started" provides links to "Create a cluster", "Ingest data", and "Invite your team". Below this are "Next steps" links for "Explore Notebook gallery" and "Read documentation". The main area displays a Python notebook titled "2022-06-09 - DBFS Example". The notebook contains five code cells:

```
1 # Create a view or table
2
3 temp_table_name = "Companies_csv"
4
5 df.createOrReplaceTempView(temp_table_name)

Cmd 4
1 %sql
2
3 /* Query the created temp table in a SQL cell */
4
5 select * from `Companies_csv`

Cmd 5
1 # With this registered as a temp view, it will only be available to this particular notebook. If you'd like other users to be
2 # able to query this table, you can also create a table from the DataFrame.
3 # Once saved, this table will persist across cluster restarts as well as allow various users across different notebooks to
4 # query this data.
5
6 permanent_table_name = "Companies_csv"
7
8 # df.write.format("parquet").saveAsTable(permanent_table_name)
```

The notebook interface includes standard Python code completion and execution tools. The status bar at the bottom shows weather information (37°C, Sunny), system icons, and the date/time (09-06-2022, 15:59).

<https://adb-25947125729108.8.azuredatabricks.net/?o=25947125729108#tables/new>

Microsoft Azure | Databricks

Get started

This is your home for all data science and engineering work.

We'll show you how to set up clusters, data and users.

Set up your workspace

- Create a cluster
- Ingest data
- Invite your team

Next steps

- Explore Notebook gallery
- Read documentation

Don't show again

37°C Sunny

Create New Table

Specify Table Attributes

Table Name: companies_1_csv

Create in Database: default

File Type: CSV

Column Delimiter: ,

First row is header:

Infer schema:

Multi-line:

Preview Table

R_D_Spend	Administration	Marketing_Spend	State	Profit
165349.2	136897.8	471784.1	New York	192261.83
162597.7	151377.59	443898.53	California	191792.06
153441.51	101145.55	407934.54	Florida	191050.39
144372.41	118671.85	383199.62	New York	182901.99
142107.34	91391.77	366168.42	Florida	166187.94
131876.9	99814.71	362861.36	New York	156991.12
134615.46	147198.87	127716.82	California	156122.51

Create Table

Create Table in Notebook

Free trial ends in 14 days. Upgrade to Premium in Azure Portal

15:30 09-06-2022

https://adb-25947125729108.8.azuredatabricks.net/?o=25947125729108#table/hive_metastore/default/companies_1_csv

Microsoft Azure | Databricks

Get started

This is your home for all data science and engineering work.

We'll show you how to set up clusters, data and users.

Set up your workspace

- Create a cluster
- Ingest data
- Invite your team

Next steps

- Explore Notebook gallery
- Read documentation

Don't show again

37°C Sunny

Description:
Created at: 2022-06-09 10:00:17
Last modified: 2022-06-09 10:00:30
Partition columns:
Number of files: 1
Size: 40 kB

Schema:

col_name	data_type	comment
1 R_D_Spend	string	
2 Administration	string	
3 Marketing_Spend	string	
4 State	string	
5 Profit	string	
6		
7 # Partitioning		

Showing all 8 rows.

Sample Data:

R_D_Spend	Administration	Marketing_Spend	State	Profit
165349.2	136897.8	471784.1	New York	192261.83
162597.7	151377.59	443898.53	California	191792.06
153441.51	101145.55	407934.54	Florida	191050.39
144372.41	118671.85	383199.62	New York	182901.99
142107.34	91391.77	366168.42	Florida	166187.94
131876.9	99814.71	362861.36	New York	156991.12

Free trial ends in 14 days. Upgrade to Premium in Azure Portal

15:30 09-06-2022

11. Explain the steps with screenshots how to create cluster in databricks ?

The screenshot shows the 'Create a cluster' wizard in the Microsoft Azure Databricks interface. The left sidebar shows the 'Compute' section with 'Create' selected. The main panel displays the 'New Cluster' configuration screen. The 'Cluster name' field contains 'dxccluster1220'. The 'Cluster mode' dropdown is set to 'Single Node'. The 'Databricks runtime version' dropdown is set to 'Runtime: 10.4 LTS (Scala 2.12, Spark 3.2.1)'. A promotional message indicates a '50% promotional discount applied to Photon during preview'. The 'Node type' dropdown shows 'Standard_DS3_v2' selected, with '14 GB Memory, 4 Cores' and 'DBU / hour: 0.75' listed. The 'Autopilot options' section includes a checkbox for 'Terminate after 120 minutes of inactivity'. A link to 'Advanced options' is visible. The top right corner shows a free trial ends in 14 days notice and an upgrade link.

The screenshot shows the 'Compute' settings page in the Microsoft Azure Databricks interface. The left sidebar shows the 'Compute' section with 'All-purpose clusters' selected. The main panel displays a table of existing clusters. The first cluster listed is 'dxccluster1220', which was created by 'dxc262ab1204_1654530098477@manipalazure.onmicrosoft.com' using the 'UI' source. The table includes columns for Name, Policy, Runtime, Active memory, Active cores, Active DBU / h, Source, and Creator. The top right corner shows a free trial ends in 14 days notice and an upgrade link.