

**ASSIGNMENT-----SOLUTION SUBMISSION**  
**ON**  
**AZURE ANALYTICS**  
**BY**

**NAME : SAI KIRAN ANCHE**

**BATCH:DXC-262-ANALYTICS-B12-**  
**AZURE**

**TRAINING UNDER : MANIPAL PRO**  
**LEARN**

**DATE OF SUBMISSION : 09-06-2022**

**EMPLOYEE DOMAIN - AZURE**  
**ANALYTICS**

**ROLL NO: DXC262AB12021**

**COMPANY – DXC TECHNOLOGY**

**TRAINER NAME – MR. AJAY KUMAR**

**NO OF QUESTIONS :11**

Assignment - 9th June 2022:

- 
1. Explain the steps with screenshots how to AzureSynapse analytics?
  2. Explain the steps with screenshots how to SQL Pool in AzureSynapse analytics?
  3. Explain the steps with screenshots how to import COVID19 dataset in AzureSynapse analytics and run sample 500 rows & display the output?
  4. Explain the steps with screenshots how to input Boston Safety datasets into AzureSynapse analytics?  
using Notebooks ?
  5. Explain the steps with screenshots how to create Spark pool in AzureSynapse analytics? ?
  6. Explain the steps with screenshots how to create pipeline in AzureSynapse analytics? ?
  7. Explain the steps with screenshots how to automate the pipelines in AzureSynapse analytics??
  8. Explain the steps with screenshots how to Databricks ?
  9. Explain the steps with screenshots how to create notebooks in Databricks ?
  10. Explain the steps with screenshots how to insert data into databricks notebook & display the result?
  11. Explain the steps with screenshots how to create cluster in databricks ?

Please create a word / pdf document, and send it to : avyuktitraining1@gmail.com

## **INTRODUCTION**

This Assignment is given by manipal pro learn team on the basis of the training done in the forenoon session of this morning. The main objective behind this assignment is to master the theory and enhance knowledge over creating the azure synapse analytics, datafactory , databricks etc..

There are 11 questions and they are of easy to moderately difficult level. All the questions have been focused on what the trainer taught in the earlier sessions. All the demonstrations have been done successfully and documented except question 6 and 7 due to the unavailability of lab access after 6:00 pm I have tried a lot to do the task even till 6:33 PM but im unable to access the azure platform.

This assignment gave me immense confidence in mastering the domain that has been assigned to me. Special thanks to Unext team for providing the lab access.

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

A:

Step1: go to <https://portal.azure.com/#home> and select azure synapse analytics

The screenshot shows the Microsoft Azure portal homepage. At the top, there are two tabs: "Subscription Details | Nuvepro" and "Home - Microsoft Azure". The "Home - Microsoft Azure" tab is active. Below the tabs, the URL "portal.azure.com/#home" is displayed. The page has a blue header bar with the Microsoft Azure logo and a search bar that says "Search resources, services, and docs (G+/)".

The main content area is divided into sections:

- Azure services:** This section contains icons for various Azure services: "Create a resource" (blue plus icon), "Azure Databricks" (red cubes icon), "Azure Synapse Analytics" (yellow hexagon icon with a green "S" inside), "Virtual networks" (green arrows icon), "Storage accounts" (teal bar icon), and "Data factories" (blue factory icon). The "Azure Synapse Analytics" icon is highlighted with a yellow oval.
- Resources:** This section has tabs for "Recent" (underlined) and "Favorite". It lists a single resource: "dxcdatabrick12" (Resource group).
- Navigate:** This section contains a "Get started" button and links to "Azure services", "Marketplace", "My account", and "Logout".

Step2: Click on create button

The screenshot shows the Microsoft Azure portal interface for managing Azure Synapse Analytics resources. The top navigation bar includes tabs for 'Subscription Details | Nuvepro' and 'Azure Synapse Analytics - Microsoft'. The URL in the address bar is [portal.azure.com/#view/HubsExtension/BrowseResource/resourceType/Microsoft.Synapse%2Fworkspaces](https://portal.azure.com/#view/HubsExtension/BrowseResource/resourceType/Microsoft.Synapse%2Fworkspaces). The main content area is titled 'Azure Synapse Analytics' and shows a message 'No Azure Synapse Analytics to display'. Below this, there is a brief description of what Synapse Analytics is and a 'Create Synapse workspace' button.

Subscription Details | Nuvepro    Azure Synapse Analytics - Microsoft

portal.azure.com/#view/HubsExtension/BrowseResource/resourceType/Microsoft.Synapse%2Fworkspaces

Microsoft Azure

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

Home >

Azure Synapse Analytics

Manipal Pro Learn (manipalazure.onmicrosoft.com)

Create Manage view Refresh Export to CSV Open query Assign tags

Filter for any field... Subscription == all Resource group == all Location == all Add filter

Name ↑↓	Type ↑↓	Resource group
---------	---------	----------------

No Azure Synapse Analytics to display

Synapse Analytics is a fully-managed service to build modern data warehouses. It brings together SQL, Apache Spark, Orchestration, and Ingestion into a single time to build an analytics solution.

Create Synapse workspace

Step3: select the resource group , workspace, workspacename,account name etc and click on review and create.

The screenshot shows the Microsoft Azure portal interface for creating a Synapse workspace. The top navigation bar includes the Microsoft Azure logo and a search bar. Below the navigation, the breadcrumb trail shows 'Home > Azure Synapse Analytics > Create Synapse workspace'. The main title is 'Create Synapse workspace'. A horizontal menu bar below the title includes tabs for 'Basics' (which is selected), 'Security', 'Networking', 'Tags', and 'Review + create'. A descriptive text below the tabs reads: 'Create a Synapse workspace to develop an enterprise analytics solution in just a few clicks.'

**Project details**

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage your resources.

Subscription \*  (highlighted in yellow)

Resource group \*  (highlighted in yellow)  
Create new

Managed resource group

**Workspace details**

Name your workspace, select a location, and choose a primary Data Lake Storage Gen2 file system to serve as the default location for logs and job output.

Workspace name \*  (highlighted in yellow)

Region \*

Select Data Lake Storage Gen2 \*  From subscription  Manually via URL

Account name \*  (highlighted in yellow)  
Create new

**Actions**

**Review + create** (highlighted in green)  
< Previous  
Next: Security >

Step4: once the validation is completed click on create button.

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](#)

### Terms

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 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	dxcdatabrick12
Region	West US
Workspace name	(new) assignmentsynapse
Data Lake Storage Gen2 account	(new) <a href="https://synapse1234.dfs.core.windows.net">https://synapse1234.dfs.core.windows.net</a>
Data Lake Storage Gen2 file system	(new) forsynapse12

**Create**

< Previous

Next >

Download a template for automation

Step5: your deployment is completed and you can enjoy the services now.

---

soft.Azure.SynapseAnalytics-20220609170131 | Overview ⚡ ...

Delete Cancel Redeploy Refresh

We'd love your feedback! →

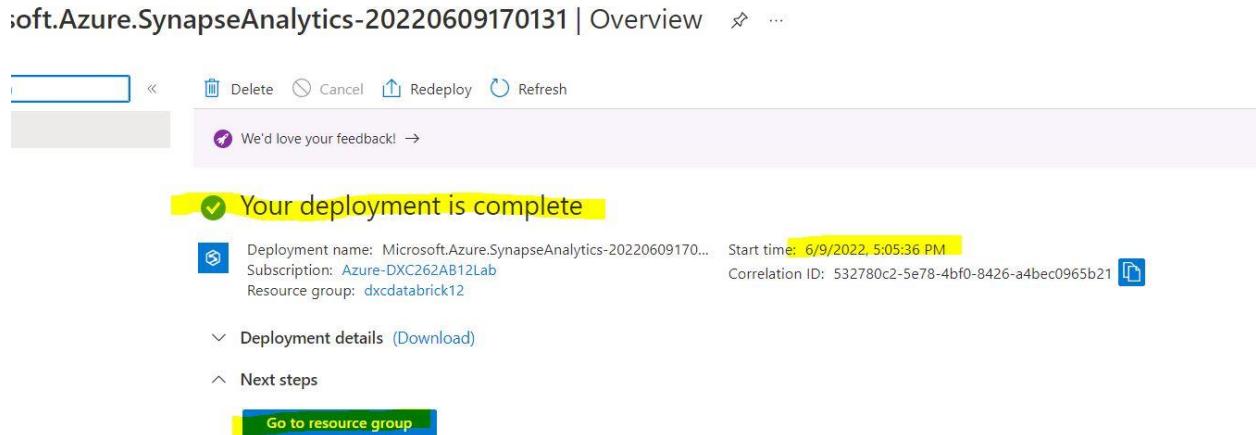
>Your deployment is complete

Deployment name: Microsoft.Azure.SynapseAnalytics-20220609170... Start time: 6/9/2022, 5:05:36 PM  
Subscription: Azure-DXC262AB12Lab Correlation ID: 532780c2-5e78-4bf0-8426-a4bec0965b21

Deployment details (Download)

Next steps

Go to resource group

A screenshot of the Azure portal showing a deployment status page. The title bar says "soft.Azure.SynapseAnalytics-20220609170131 | Overview". Below it are standard actions: Delete, Cancel, Redeploy, and Refresh. A feedback link "We'd love your feedback! →" is present. A prominent yellow banner at the top says "Your deployment is complete" with a checkmark icon. Below the banner, deployment details are listed: Deployment name (redacted), Start time (6/9/2022, 5:05:36 PM), Subscription (Azure-DXC262AB12Lab), Correlation ID (532780c2-5e78-4bf0-8426-a4bec0965b21), and a copy icon. There are two expandable sections: "Deployment details (Download)" and "Next steps". A blue button labeled "Go to resource group" is at the bottom.

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

A:

Step1: open homepage of azure synapse and click on the synapse work space.

The screenshot shows the Microsoft Azure portal interface. At the top, there's a search bar and a navigation bar with various icons. Below the search bar, the URL 'microsoft.Azure.SynapseAnalytics-20220609170131' is visible. The main area is titled 'cdatabrick12' and shows a 'Resource group' view. On the left, there's a sidebar with 'Essentials' expanded, showing a subscription (move) to 'Azure-DXC262AB12Lab', a Subscription ID, and a Tags section. The main content area displays a 'Resources' table with two records:

Name	Type	Location
assignmentsynapse	Synapse workspace	West US
synapse1234	Storage account	West US

## Step2: open the synapse studio.

crosoft.Azure.SynapseAnalytics-20220609170131 > dxcdatabrick12 >

### gnmentsynapse

workspace

Ctrl + /

New dedicated SQL pool    New Apache Spark pool    New Data Explorer pool (preview)    Refresh    Rese

Essentials

Resource group ( <a href="#">move</a> )	: dxcdatabrick12	Networking
Status	: Succeeded	Primary ADLS Gen
Location	: West US	Primary ADLS Gen
Subscription ( <a href="#">move</a> )	: Azure-DXC262AB12Lab	SQL admin user
Subscription ID	: 3a28cdce-3bd7-4219-858e-23ff20f8b998	SQL Active Direct
Managed virtual network	: No	Dedicated SQL er
Managed Identity object ...	: 7e5c1971-1780-4ace-a9ac-3cb5abb588c6	Serverless SQL en
Workspace web URL	: <a href="https://web.azuresynthesize.net?workspace=%2fsubscriptions%2f3a...">https://web.azuresynthesize.net?workspace=%2fsubscriptions%2f3a...</a>	Development enc
Tags ( <a href="#">edit</a> )	: <a href="#">Click here to add tags</a>	

Getting started

 Open Synapse Studio  
Start building your fully-integrated analytics solution and unlock new insights.  
[Open ↗](#)

 Read documentation  
Learn how to be productive quickly. Explore concepts, tutorials, and samples.  
[Learn more ↗](#)

Analytics pools

Search to filter items...

Step3:

Click on data and then click on + and later sql data base.

The screenshot shows the Microsoft Azure portal interface for creating a new workspace. The top navigation bar displays "Microsoft Azure" and the workspace name "assignmentsynapse". The search bar on the right contains the text "Search". Below the navigation bar, there are several buttons: a double arrow icon, "Synapse live" with a dropdown arrow, "Validate all" with a checkmark icon, "Publish all" with an upward arrow icon, and a magnifying glass icon labeled "Search".

The main area is titled "Data" and features a "Workspace" tab selected. A yellow box highlights the "Data" icon in the sidebar and the "+" button above the workspace list. The workspace list includes "Workspace" (selected), "Link", and "Linked". Under "Linked", there are four options: "SQL database" (highlighted with a yellow box), "Lake database", "Data Explorer database (preview)", and "Browse gallery". On the left side, there is a vertical sidebar with icons for Home, Storage, Functions, Logic Apps, and App Services. The "Storage" icon is highlighted with a yellow box.

Step4: name your database and click okay

0ZI0XCDATAHCKTZ%0ZI providers%0ZI ...

dxc262ab1227\_1654530006055@manipalazure.onmicrosoft.com  
MANIPAL PRO LEARN

## Create SQL database

Create database to organize your workload into databases and database objects.

Select SQL pool type \*

Serverless (i)

Dedicated (i)

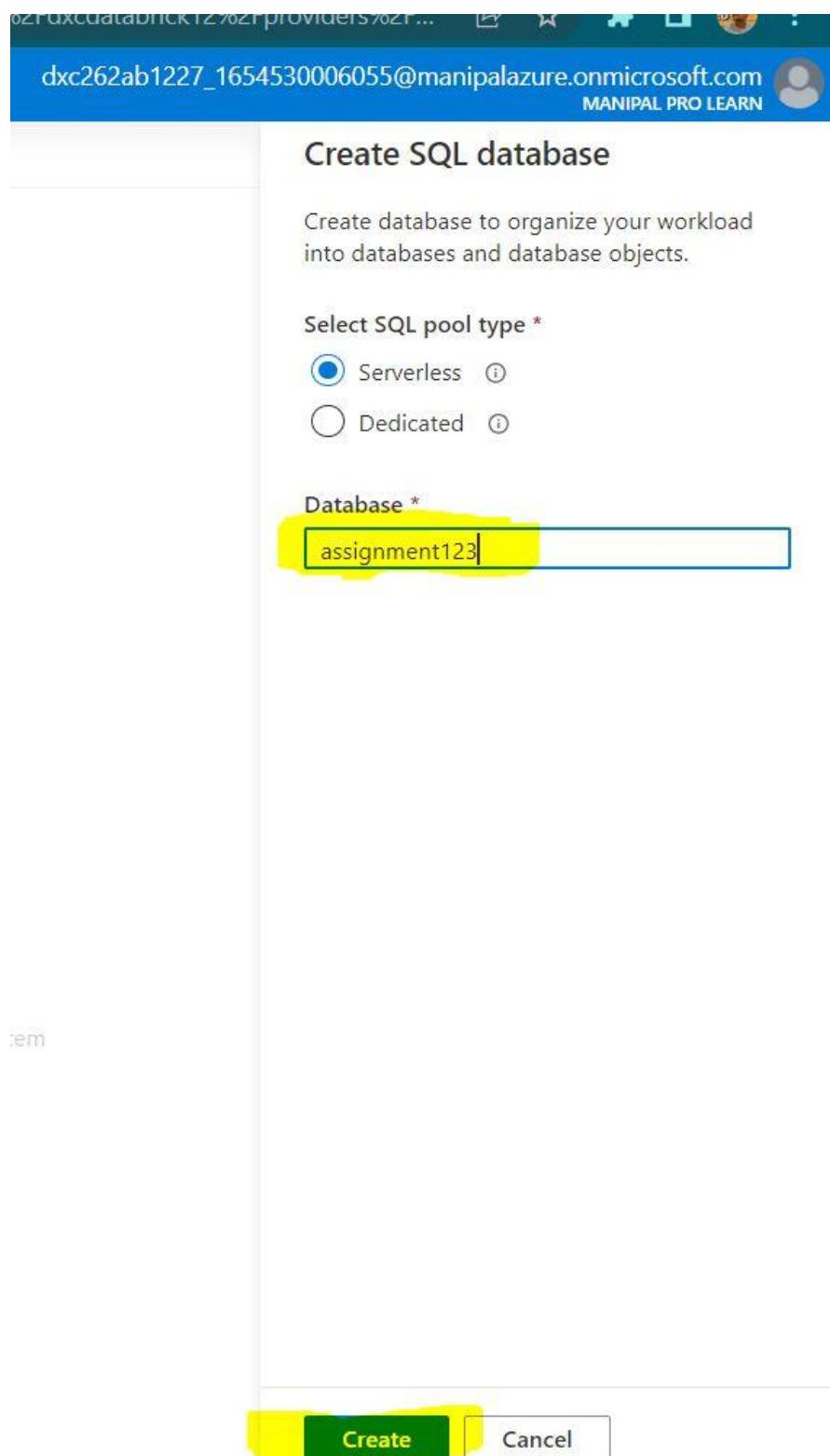
Database \*

assignment123

tem

---

**Create** **Cancel**



Step5: your sql database is ready

Microsoft Azure | assignmentsynapse

» Synapse live ▾ Validate all Publish all

**Data** + ⌂ <<

**Workspace** Linked

Filter resources by name

SQL database ...

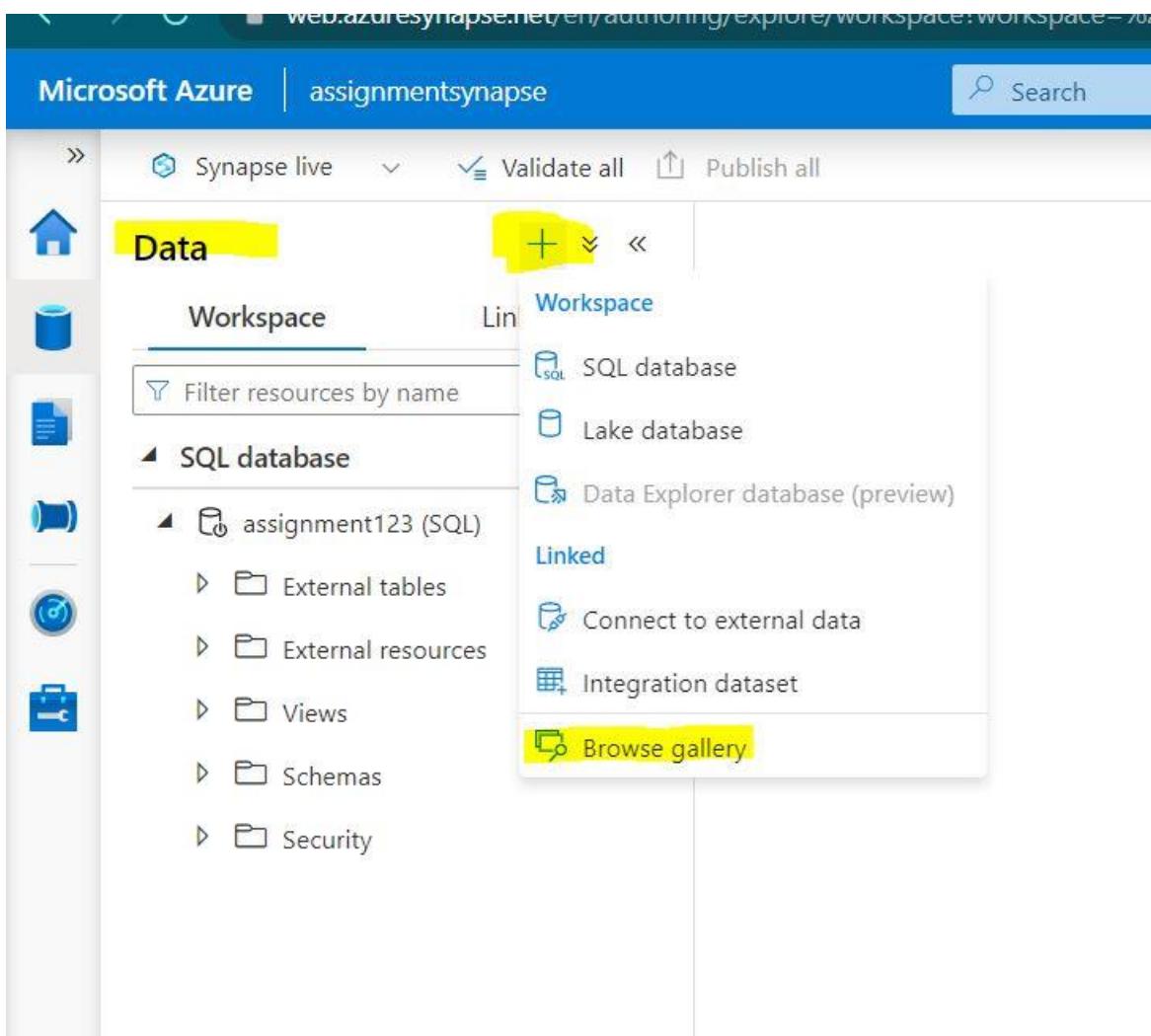
assignment123 (SQL)

The screenshot shows the Microsoft Azure portal interface for a workspace named 'assignmentsynapse'. On the left, there's a vertical navigation bar with icons for Home, Storage, Files, Functions, and App Services. The main area is titled 'Data' and has tabs for 'Workspace' (which is selected) and 'Linked'. A search bar at the top says 'Filter resources by name'. Below it, a section for 'SQL database' lists one item: 'assignment123 (SQL)', which is highlighted with a yellow box. At the top of the main content area, there are buttons for 'Synapse live', 'Validate all', and 'Publish all'. The overall theme is light blue and white.

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

A:

Step1: open azure synapse studio, click data and then on + later click on browse gallery.



Step2: select bing covid data. And click continue

Microsoft Azure | assignmentsynapse

Search

Gallery

Database templates Datasets Notebooks SQL scripts Pipelines

Filter by keyword Tags : All

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

Continue

Step3: click continue

[Database templates](#)[Datasets](#)[Notebooks](#)[SQL scripts](#)[Pipeline](#)[Filter by keyword](#)

Tags : All



### 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 report to the city of Boston. This database is stored in Parquet format and is updated daily.

[ID: city\\_safety\\_boston](#)

### NOAA Integrated Surface Data (ISD)

NOAA Integrated Surface Data (ISD) provides Worldwide hourly weather history data sourced from the National Oceanic and Atmospheric...

[ID: isd](#)

### NYC Taxi & Limousine Commission - For-Hire Vehicle (FHV) trip records

The For-Hire Vehicle trip records include fields capturing the dispatching base license number

[ID: nyc\\_tlc\\_fhv](#)[Continue](#)

Step4: click on add dataset



## Bing COVID-19 Data

### Description

Bing COVID-19 data includes confirmed, fatal, and recovered cases from all regions, updated daily. This data is reflected in the [Bing COVID-19 Tracker](#).

Bing collects data from multiple trusted, reliable sources, including the World Health Organization (WHO), Centers for Disease Control and Prevention (CDC), national and state public health departments, BNO News, 24/7 Wall St., and Wikipedia.

For more information and original source data see this [link](#). For license terms see this [link](#).

### Datasets:

Modified datasets are available in CSV, JSON, JSON-Lines, and Parquet.  
[https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/bing\\_covid-19\\_data/latest/bing\\_covid-19\\_data.csv](https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/bing_covid-19_data/latest/bing_covid-19_data.csv)  
[https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/bing\\_covid-19\\_data/latest/bing\\_covid-19\\_data.json](https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/bing_covid-19_data/latest/bing_covid-19_data.json)  
[https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/bing\\_covid-19\\_data/latest/bing\\_covid-19\\_data.jsonl](https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/bing_covid-19_data/latest/bing_covid-19_data.jsonl)  
[https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/bing\\_covid-19\\_data/latest/bing\\_covid-19\\_data.parquet](https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/bing_covid-19_data/latest/bing_covid-19_data.parquet)

All modified datasets have ISO 3166 subdivision codes and load times added, and use lower case column names with underscore separators.

### Raw data:

[https://pandemicdatalake.blob.core.windows.net/public/raw/covid-19/bing\\_covid-19\\_data/latest/Bing-COVID19-Data.csv](https://pandemicdatalake.blob.core.windows.net/public/raw/covid-19/bing_covid-19_data/latest/Bing-COVID19-Data.csv)

### Previous versions of modified and raw data:

[https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/bing\\_covid-19\\_data/](https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/bing_covid-19_data/)

### Data Volume

All datasets are updated daily. As of May 11, 2020 they contained

### Preview

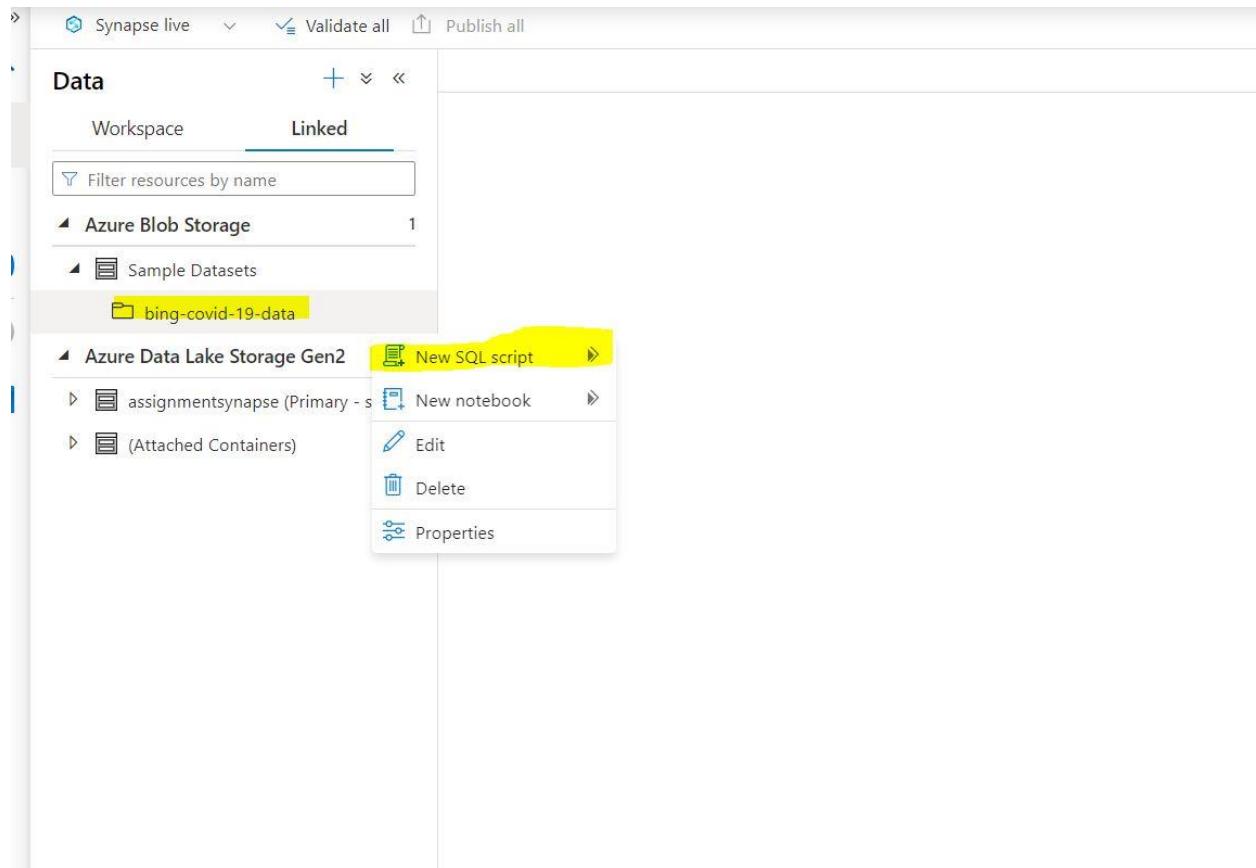
id	updated	confirmed	deaths	iso2	iso3
338995	2020-01-21	262	0	null	null
338996	2020-01-22	313	0	null	null
338997	2020-01-23	578	0	null	null
338998	2020-01-24	841	0	null	null
338999	2020-01-25	1320	0	null	null
339000	2020-01-26	2014	0	null	null
339001	2020-01-27	2798	0	null	null
339002	2020-01-28	4593	0	null	null
339003	2020-01-29	6065	0	null	null
339004	2020-01-30	7818	0	null	null

▼

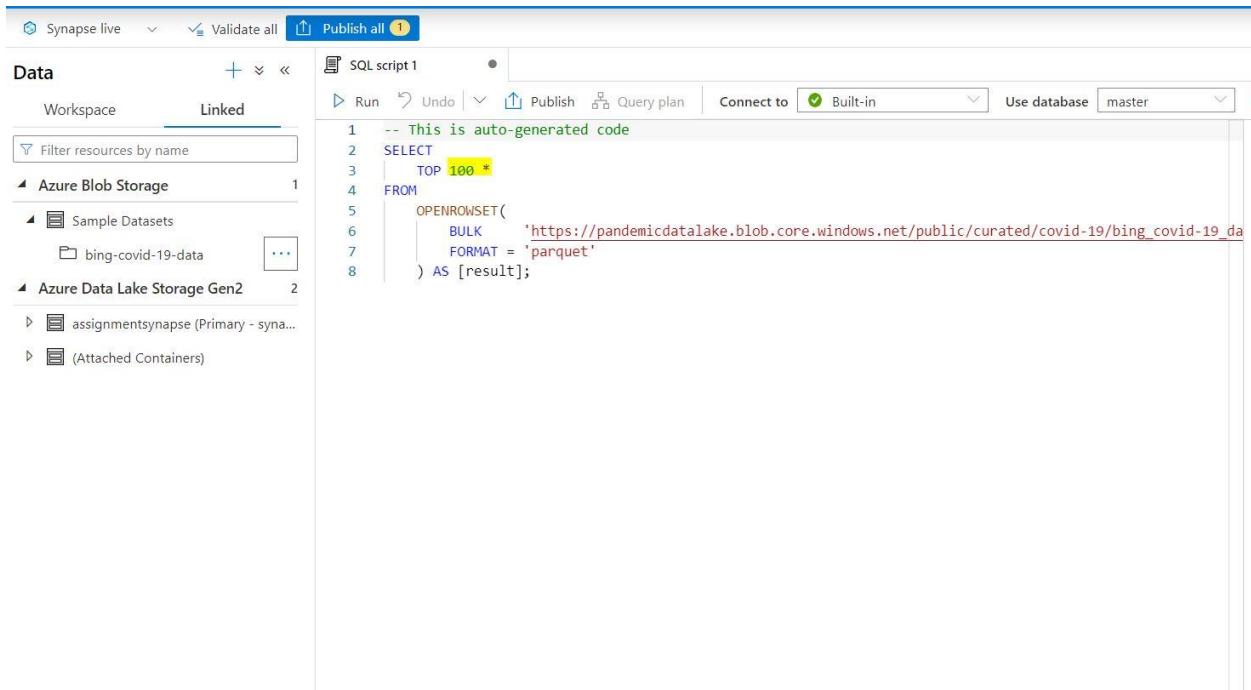
Add dataset

Back

Step5: select data and then select our database and later click on new sql script. And elect top 100.

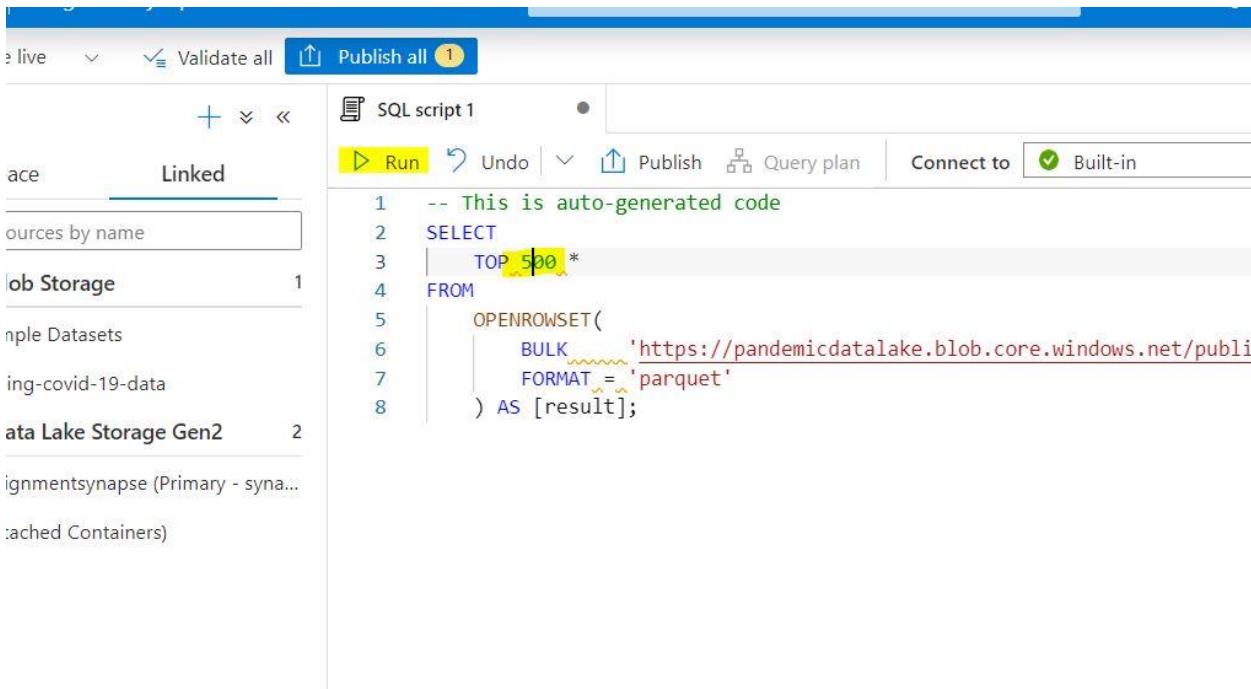


## Step6: change the 100 in the code to 500



```
-- This is auto-generated code
SELECT
| TOP 100 *
FROM
OPENROWSET(
    BULK 'https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/bing_covid-19_da
FORMAT = 'parquet'
) AS [result];
```

## Step7: click on run button



```
-- This is auto-generated code
SELECT
| TOP 500 *
FROM
OPENROWSET(
    BULK 'https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/bing_covid-19_da
FORMAT = 'parquet'
) AS [result];
```

Step8: here are the results.

The screenshot shows the SSMS interface with a query window open. The code in the script pane is:

```
-- This is auto-generated code
SELECT
    TOP 500 *
FROM
    OPENROWSET(
        BULK 'https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/bing_covid-19',
        FORMAT = 'parquet'
    ) AS [result];
```

The results pane displays a table with the following data:

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)
339000	2020-01-26T00:00:00Z	2014	694	0	0	(NULL)
339001	2020-01-27T00:00:00Z	2798	784	0	0	(NULL)
339002	2020-01-28T00:00:00Z	4593	1795	0	0	(NULL)

A status message at the bottom of the results pane says "00:00:15 Query executed successfully."

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

A:

Step1: click on data , +and browse gallery

The screenshot shows the Azure Synapse Analytics Data workspace interface. At the top, there's a navigation bar with 'On Azure' and 'assignmentsynapse'. Below it, there are buttons for 'Synapse live', 'Validate all', and 'Publish all'. The main area is titled 'Data' and has tabs for 'Workspace' and 'Link'. A '+' button is visible above a list of resources. To the right, there's a search bar and some status indicators. The resource list includes 'SQL database', 'Lake database', 'Data Explorer database (preview)', 'Linked', 'Connect to external data', 'Integration dataset', and 'Browse gallery'. The 'Browse gallery' item is highlighted with a yellow box.

Step2: select boston secyrity data and click continue.

Gallery

Database templates   **Datasets**   Notebooks   SQL scripts   Pipelines

Filter by keyword   Tags : All



**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



**NOAA Integrated Surface Data (ISD)**  
NOAA Integrated Surface Data (ISD) provides Worldwide hourly weather history data sourced from the National Oceanic and Atmospheric...

ID: isd



**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...

ID: nyc\_tlc\_fhv



**Continue**

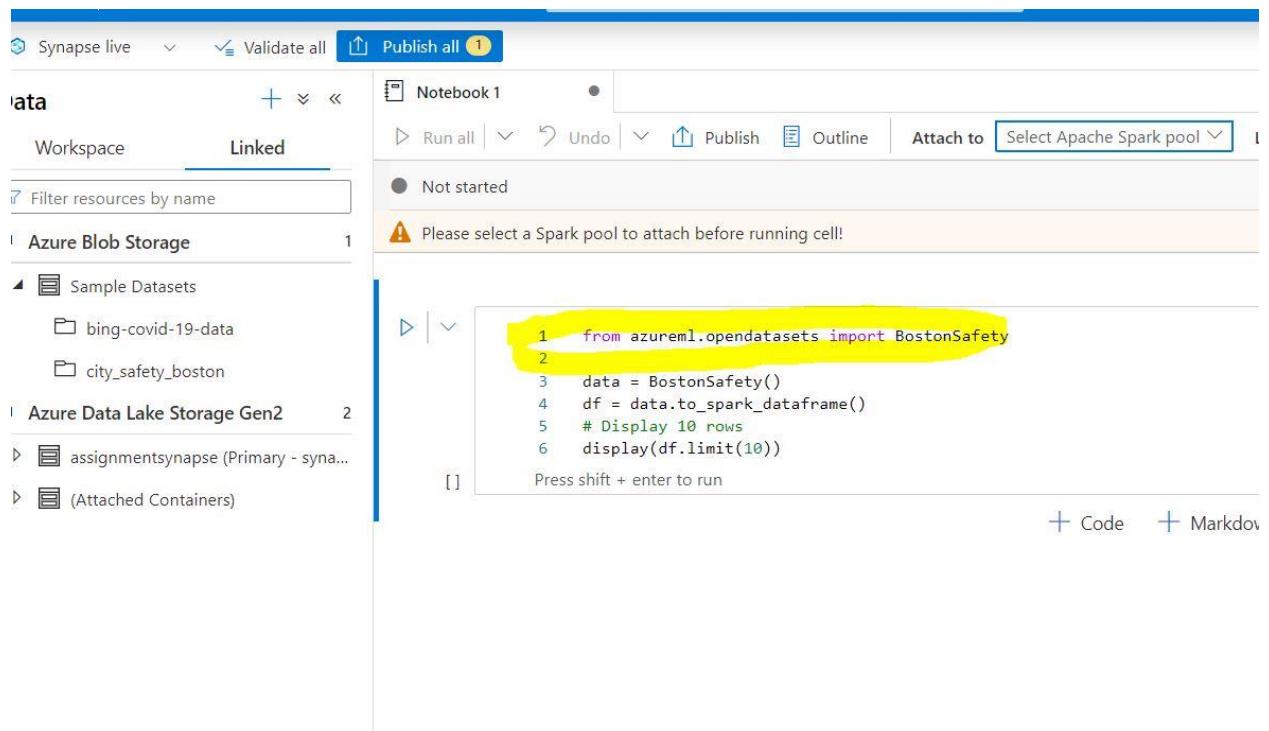
Step3: click on add dataset

Microsoft Azure   assignmentsynapse		Search
<h2>Boston Safety Data</h2>		
<h3>Description</h3> <p>311 calls reported to the city of Boston.</p> <p>Refer to this link to learn more about <a href="#">BOS:311</a>.</p>		Preview
<h3>Volume and Retention</h3> <p>This dataset is stored in Parquet format. It is updated daily, and contains about 100K rows (10MB) in total as of 2019.</p> <p>This dataset contains historical records accumulated from 2011 to the present. You can use parameter settings in our SDK to fetch data within a specific time range.</p>		dataType
<h3>Storage Location</h3> <p>This dataset is stored in the East US Azure region. Allocating compute resources in East US is recommended for affinity.</p>		Safety
<h3>Additional Information</h3> <p>This dataset is sourced from city of Boston government. More details can be found from <a href="#">here</a>. Reference <a href="#">Open Data Commons Public Domain Dedication and License (ODC PDDL)</a> for the license of using this dataset.</p>		Safety
<h3>Notices</h3> <p>MICROSOFT PROVIDES AZURE OPEN DATASETS ON AN "AS IS" BASIS. MICROSOFT MAKES NO WARRANTIES, EXPRESS OR IMPLIED, GUARANTEES OR CONDITIONS WITH RESPECT TO YOUR USE OF THE DATASETS. TO THE EXTENT PERMITTED UNDER YOUR LOCAL LAW, MICROSOFT DISCLAIMS ALL LIABILITY FOR ANY DAMAGES OR LOSSES, INCLUDING DIRECT, CONSEQUENTIAL, SPECIAL, INDIRECT, INCIDENTAL OR PUNITIVE, RESULTING FROM YOUR USE OF THE DATASETS.</p>		Safety
<p>This dataset is provided under the original terms that Microsoft</p>		Safety
<a href="#">Add dataset</a>		<a href="#">Back</a>

Step4: click on azure blobstorage and then city\_safety\_boston and click on new note book and later load to Dataframe.

The screenshot shows the Azure Synapse Studio interface. At the top, there are buttons for 'Synapse live', 'Validate all', and 'Publish all'. Below this is the 'Data' section with tabs for 'Workspace' and 'Linked'. The 'Linked' tab is selected. A search bar says 'Filter resources by name'. Under 'Linked', there is a list of storage accounts: 'Azure Blob Storage' (selected), 'Azure Data Lake Storage Gen2', 'assignmentsynapse (Primary - s)', and '(Attached Containers)'. 'Azure Blob Storage' has a sub-list: 'Sample Datasets' (with 'bing-covid-19-data' and 'city\_safety\_boston' selected). A context menu is open over 'city\_safety\_boston', with options: 'New SQL script', 'New notebook' (highlighted), 'Edit', 'Delete', and 'Properties'. The 'Load to DataFrame' option is also highlighted in the menu.

Step5: we can see the input dataframe in the notebook



The screenshot shows the Azure Synapse Notebooks interface. On the left, there's a sidebar titled 'Data' with sections for 'Workspace' and 'Linked'. Under 'Linked', there are entries for 'Azure Blob Storage' (with 'bing-covid-19-data' and 'city\_safety\_boston' sub-folders) and 'Azure Data Lake Storage Gen2' (with 'assignmentsynapse' and '(Attached Containers)'). The main area is titled 'Notebook 1' and shows a single cell with the following Python code:

```
1 from azureml.opendatasets import BostonSafety
2
3 data = BostonSafety()
4 df = data.to_spark_dataframe()
5 # Display 10 rows
6 display(df.limit(10))
```

A yellow box highlights the first two lines of code. Below the code cell, it says 'Press shift + enter to run'. At the bottom right, there are buttons for '+ Code' and '+ Markdown'.

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

A:

Step1: open synapse studio and click on synapse live and click on apache spark pools and on new button.

The screenshot shows the Microsoft Azure Synapse Studio interface. The top navigation bar displays "Microsoft Azure" and the workspace name "assignmentsynapse". Below the navigation bar, there are several icons representing different service categories: Home, Analytics pools, SQL pools, Apache Spark pools (which is highlighted with a yellow box), Data Explorer pools (preview), External connections, Linked services, Microsoft Purview, Integration, Triggers, Integration runtimes, Security, Access control, Credentials, and Managed private endpoints. On the right side of the screen, under the "Apache Spark pools" section, there is descriptive text about Apache Spark pools and a "New" button. A "Refresh" button is also present. A "Filter by name" input field is available. The message "Showing 0-0 of 0 item" indicates no pools have been created yet. The overall layout is clean and organized, typical of a cloud-based management tool.

Step2: name the spark pool and select node size , pool name and number of nodes.

## New Apache Spark pool

Basics • Additional settings \* Tags Review + create

Create an Synapse Analytics Apache Spark pool with your preferred configurations. Complete the Basics tab then go to Review + Create to provision with smart defaults, or visit each tab to customize.

### Apache Spark pool details

Name your Apache Spark pool and choose its initial settings.

Apache Spark pool name \* assignments1

Node size family \* Memory Optimized

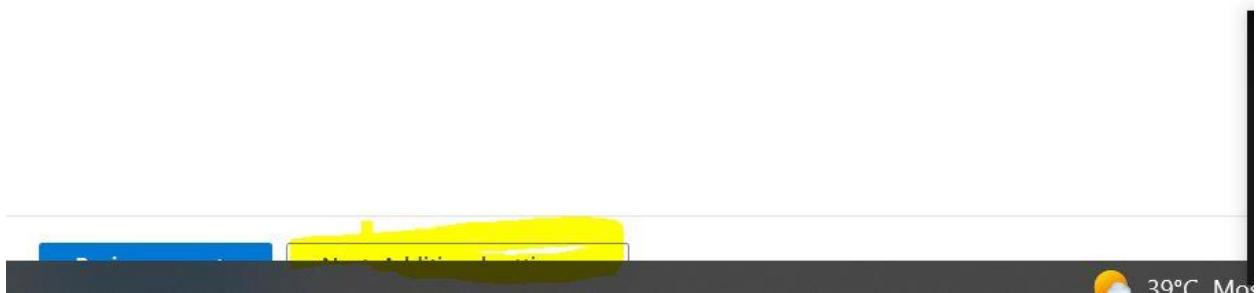
Node size \* Small (4 vCores / 32 GB)

Autoscale \*  Enabled  Disabled

Number of nodes \* 3

Estimated price ⓘ Est. cost per hour  
✖ Failed to fetch billing info

Dynamically allocate executors \*  Enabled  Disabled



Step3:

After successful validation click on create .

New Apache Spark pool

Validation succeeded.

Basics \* Additional settings \* Tags Review + create

Product details

Azure Synapse Analytics Apache Spark pool by Microsoft

Est. cost per hour Failed to fetch billing info

Terms of use | Privacy policy

Terms

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	dxcdatabrick12
Apache Spark pool name	assignments1
Node size family	Memory Optimized
Node size	Small (4 vCores / 32 GB)
Autoscale	Enabled

Create < Previous Download template for automation

## Step4: your spark pool is created.

The screenshot shows a user interface for managing Apache Spark pools. At the top, there are navigation links: 'Validate all' and 'Publish all'. Below this, a header bar displays the title 'Apache Spark pool'. A sub-header provides a brief description: 'Apache Spark pools can be tuned to run different kinds of Apache Spark workloads using specific configuration libraries, permissions, etc. Learn more'. There are two buttons: '+ New' and 'Refresh'. A search bar labeled 'Filter by name' is present. The main content area shows a table with one item. The table has columns: 'Name', 'Node size family', and 'Size'. The single row contains the value 'assignments1' under 'Name', 'Memory Optimized' under 'Node size family', and 'Small (4 vCores / 32 GB) - 3 to 3 nodes' under 'Size'. The entire row is highlighted with a yellow background.

Name	Node size family	Size
assignments1	Memory Optimized	Small (4 vCores / 32 GB) - 3 to 3 nodes

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

A:

Step1:

7. Explain the steps with screenshots how to automate the pipelines in AzureSynapse analytics??

A:

Step1:

Step1:

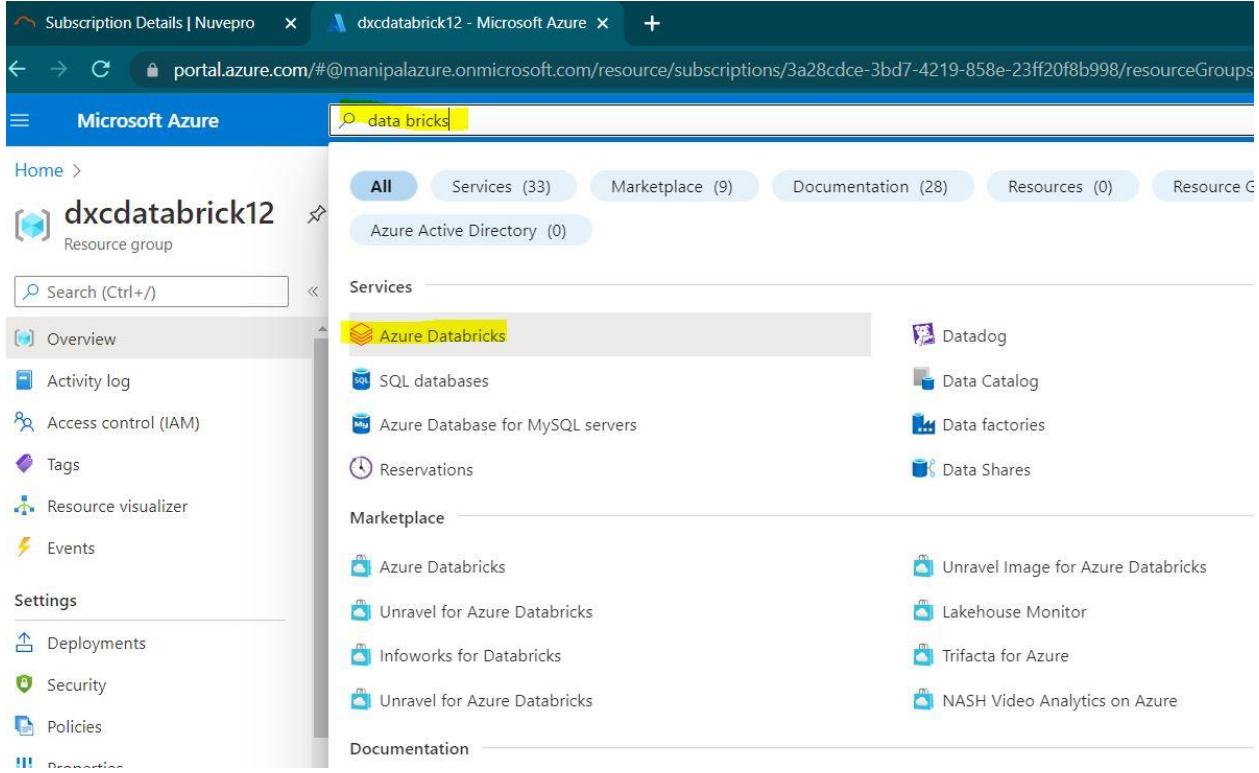
Step1:

Step1:

## 8. Explain the steps with screenshots how to Create Databricks ?

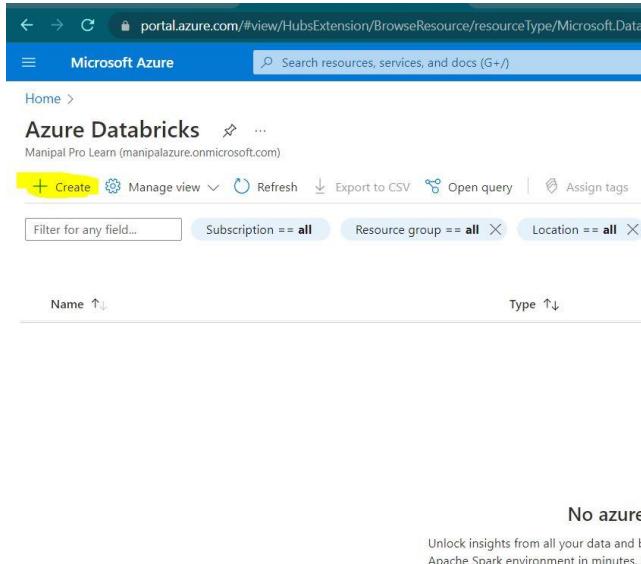
A:

Step1: Go to <https://portal.azure.com/> and search for Databricks.



The screenshot shows the Microsoft Azure portal interface. At the top, there is a search bar with the text "data bricks". Below the search bar, there are several tabs: "All", "Services (33)", "Marketplace (9)", "Documentation (28)", "Resources (0)", and "Resource G". Under the "Services" tab, the "Azure Databricks" service is highlighted with a yellow box. To the right of the main search results, there is a sidebar with various Azure services listed, such as Datadog, Data Catalog, Data factories, Data Shares, Unravel Image for Azure Databricks, Lakehouse Monitor, Trifacta for Azure, and NASH Video Analytics on Azure. On the left side of the screen, there is a navigation menu for the "dxcdatabrick12" resource group, which includes sections for Overview, Activity log, Access control (IAM), Tags, Resource visualizer, Events, Deployments, Security, Policies, and Documentation.

Step2: click on create button



The screenshot shows the Microsoft Azure portal interface, specifically the "Azure Databricks" resource list. At the top, there is a search bar with the text "Search resources, services, and docs (G+J)". Below the search bar, there are several buttons: "+ Create", "Manage view", "Refresh", "Export to CSV", "Open query", and "Assign tags". There are also filters for "Subscription == all", "Resource group == all", and "Location == all". The main area displays a table with columns for "Name" and "Type". A message at the bottom of the table says "No azure". Below the table, there is a brief description: "Unlock insights from all your data and build Apache Spark environment in minutes;".

Step3: select resource group, name the workspace , select the pricing tier and click on review and create

Home > Azure Databricks >

## Create an Azure Databricks workspace ...

Basics Networking Advanced Tags Review + create

### Project Details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription \* ⓘ

Azure-DXC262AB12Lab

Resource group \* ⓘ

dxcdatabrick12



Create new

### Instance Details

Workspace name \*

assignmentbrick1



Region \*

East US



Pricing Tier \* ⓘ

Trial (Premium - 14-Days Free DBUs)



Review + create

< Previous

Next : Networking >

Step4: after validation succeeded click on create button.

Home > Azure Databricks >

## Create an Azure Databricks workspace



Validation Succeeded

Basics Networking Advanced Tags Review + create

### Summary

#### Basics

Workspace name	assignmentbrick1
Subscription	Azure-DXC262AB12Lab
Resource group	dxcdatabrick12
Region	East US
Pricing Tier	trial

#### Networking

Deploy Azure Databricks workspace with Secure Cluster Connectivity (No Public IP)	No
Deploy Azure Databricks workspace in your own Virtual Network (VNet)	No

#### Advanced

Enable Infrastructure Encryption	No
----------------------------------	----

**Create**

< Previous

Download a template for automation

Step5: your deployment is completed.click goto resource.

Home > **dxcdatabrick12\_assignmentbrick1 | Overview** ⚙ ...

Deployment

Search (Ctrl+ /) << Delete Cancel Redeploy Refresh

Overview We'd love your feedback! →

Inputs

Outputs

Template

✓ Your deployment is complete

Deployment name: dxcdatabrick12\_assignmentbrick1  
Subscription: Azure-DXC262AB12Lab  
Resource group: dxcdatabrick12

Start time: 6/9/2022, 4:05:39 PM  
Correlation ID: d12790a4-69a9-424d-bec5-c0594bf305e0

Deployment details (Download)

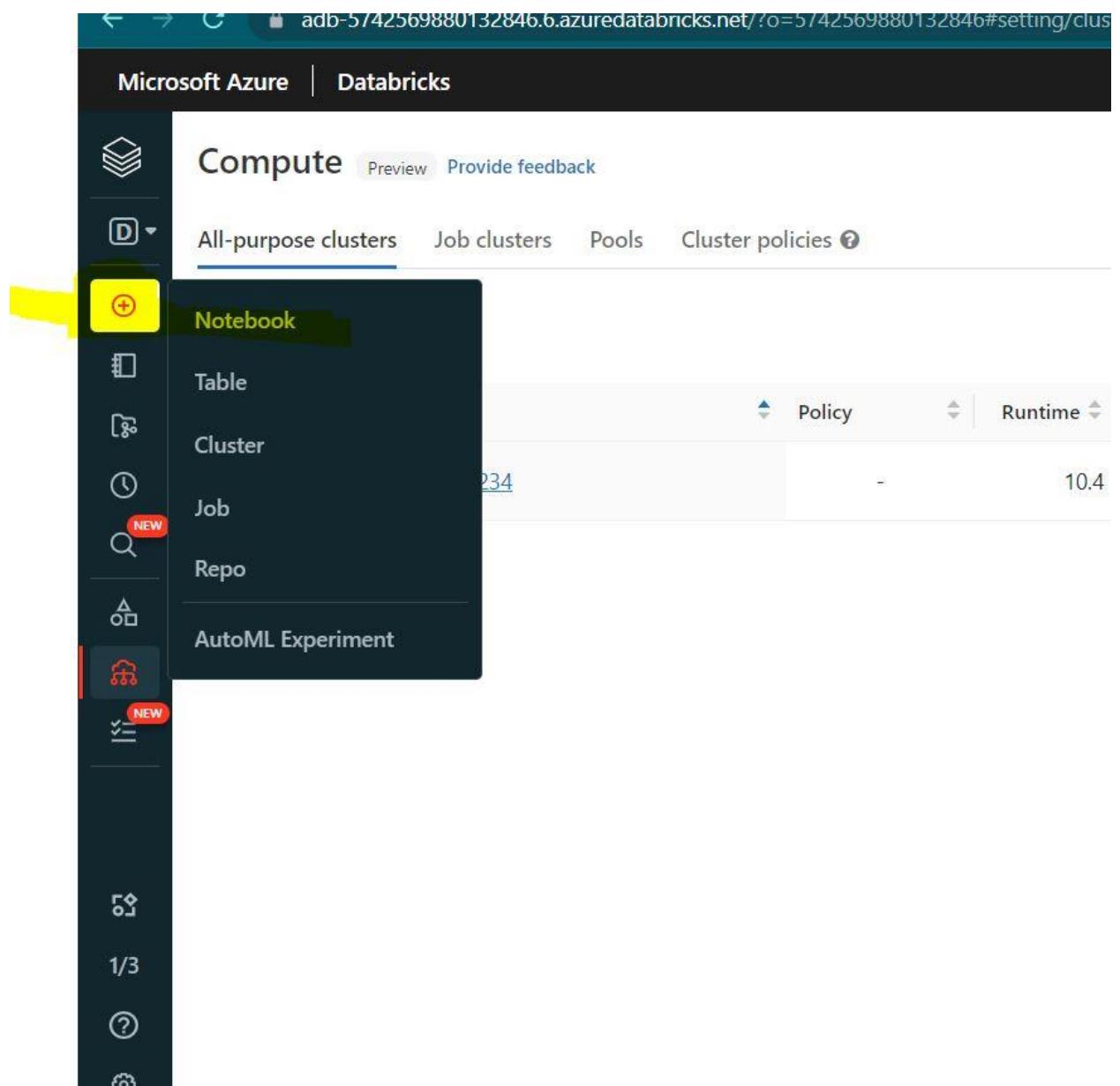
Next steps

Go to resource

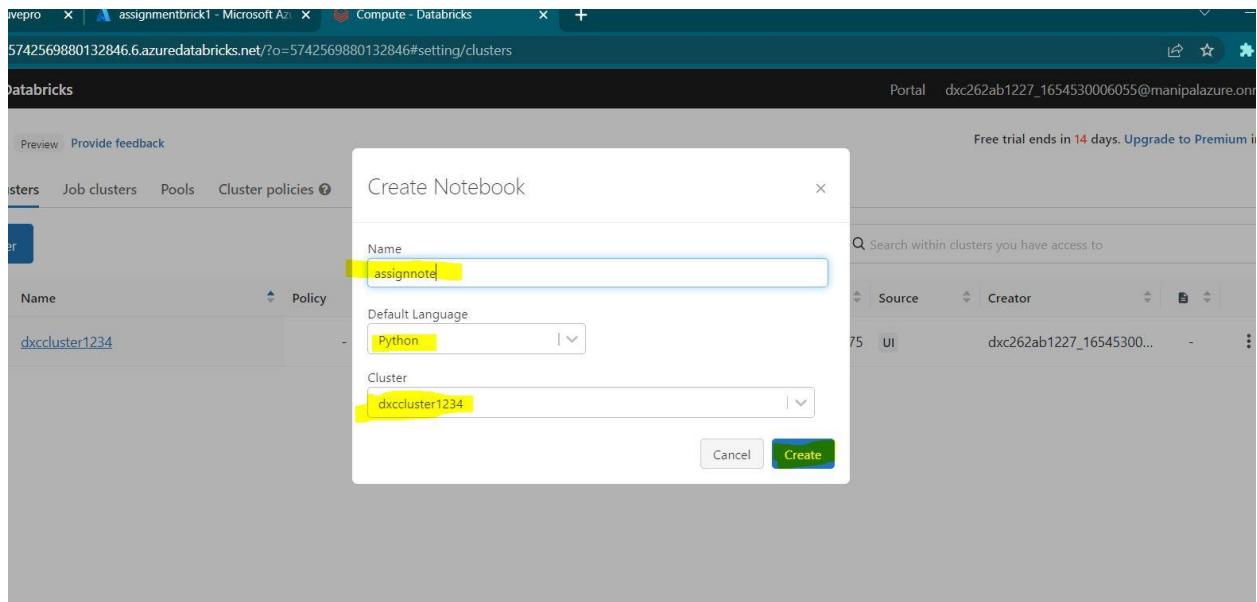
9. Explain the steps with screenshots how to create notebooks in Databricks ?

A:

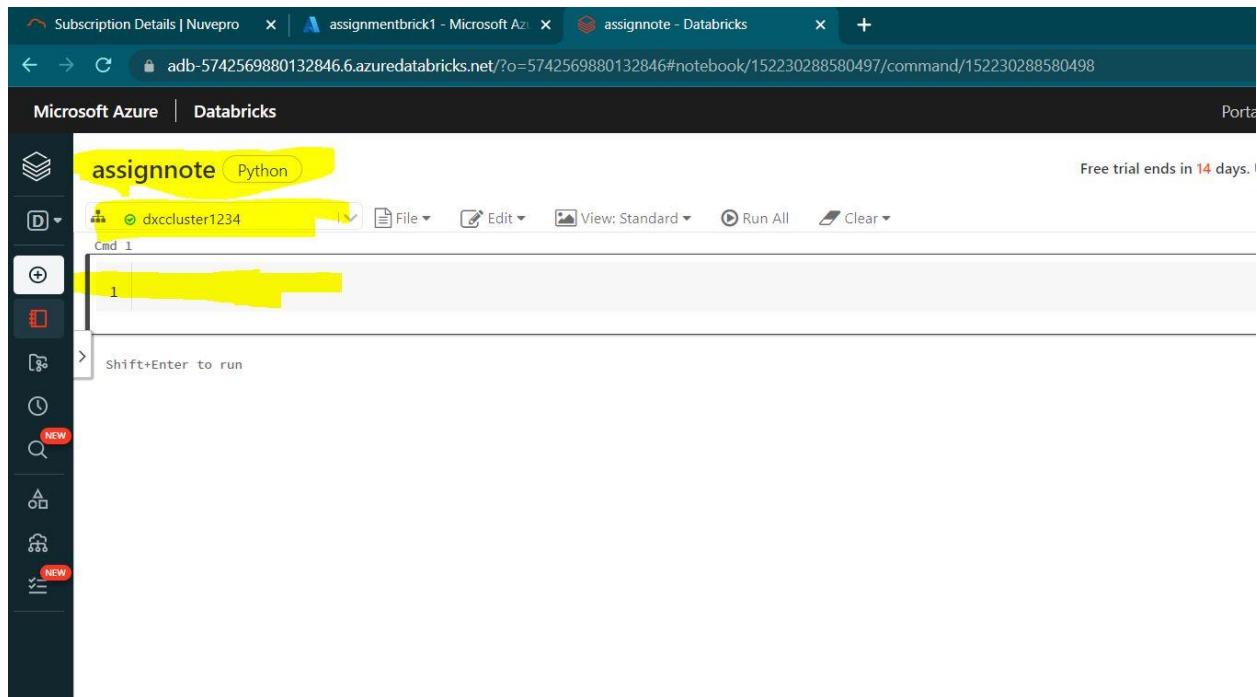
Step1: Click on create button in the workspace and then on notebook



Step2: give the name , slect the cluster and also language and finally click on create button.



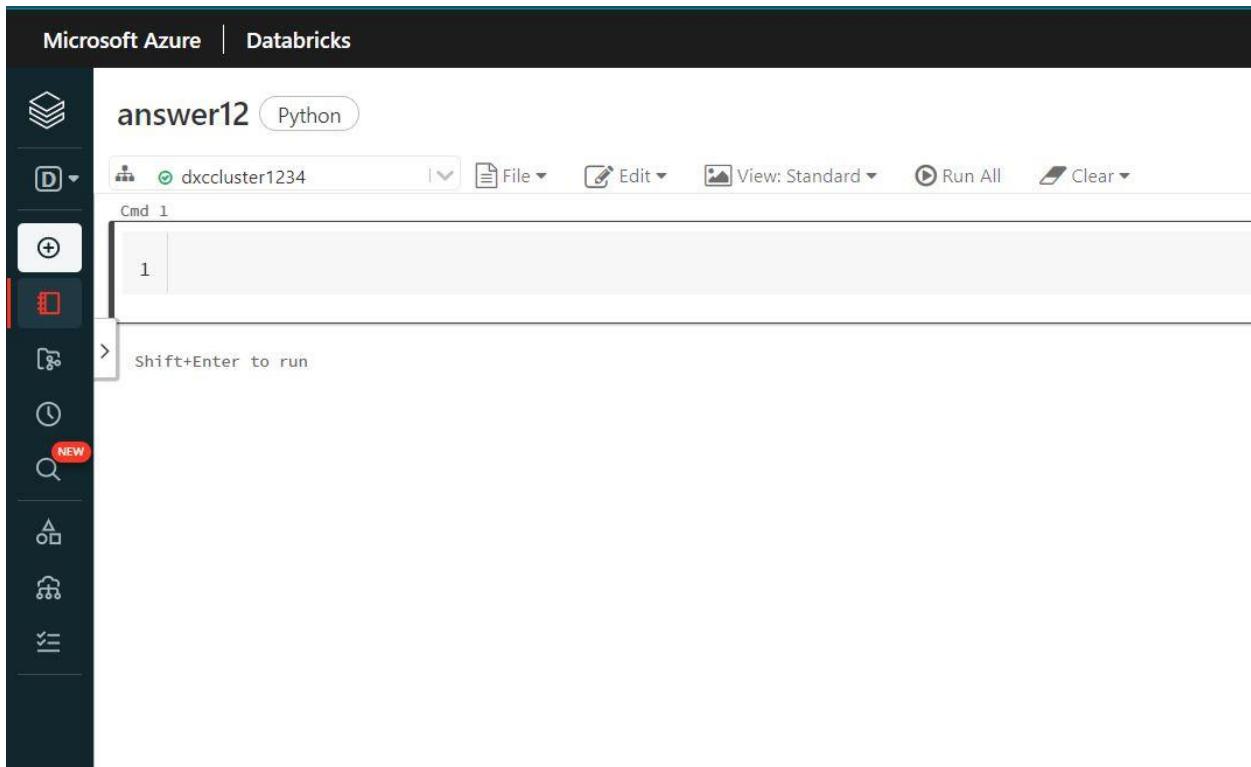
Step3: your notebook is created.



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

A:

Step1: open the note book



The screenshot shows the Microsoft Azure Databricks interface. At the top, it says "Microsoft Azure | Databricks". Below that is a notebook titled "answer12" in Python. The notebook has one cell, labeled "Cmd 1", which contains the number "1". Below the cell, there is a prompt: "Shift+Enter to run". On the left side, there is a sidebar with various icons: a cluster icon, a plus sign, a minus sign, a refresh, a clock, a search bar with a red "NEW" badge, a triangle, and a three-line menu icon.

Step2: click on create and then table , create table

The screenshot shows the Microsoft Azure Databricks interface. On the left, there's a sidebar with various options: Data Science & Eng., Create (highlighted with a yellow box), Workspace, Repos, Recents, Search (with a NEW badge), Data (highlighted with a green box), Compute, Workflows, Partner Connect, 1/3 Tasks Completed, Help, Settings, and assignmentbrick1. The main area is titled 'Data' and has tabs for 'Databases' (selected) and 'Tables' (highlighted with a yellow box). A 'Create Table' button is located at the top right of the 'Tables' section. Below it, there's a search bar labeled 'Filter Databases' and a dropdown menu showing 'default'. The message 'No Tables' is displayed.

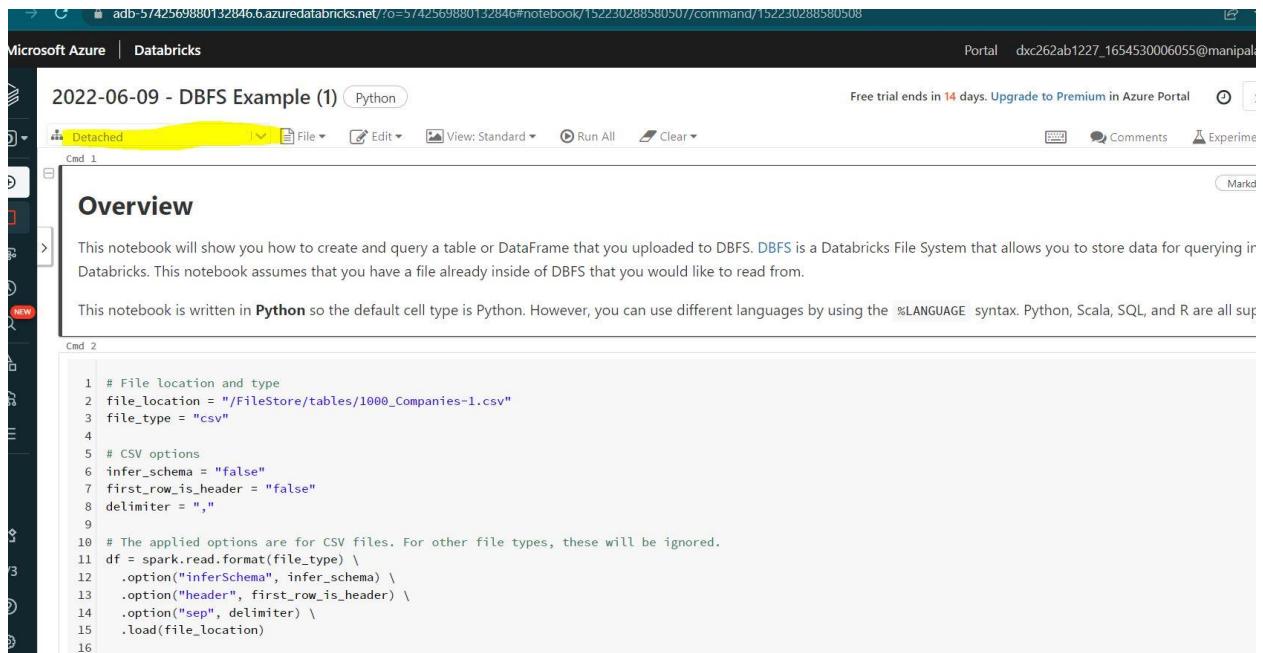
### Step3: upload the data which you want.

The screenshot shows the Databricks interface for creating a new table. In the top navigation bar, it says "Azure | Databricks" and "Portal dxc26ab1227\_16545300". A message at the top right says "Free trial ends in 14 days". The main section is titled "Create New Table" and has tabs for "Upload File", "DBFS", and "Other Data Sources". Under "Upload File", there's a dropdown for "DBFS Target Directory" set to "FileStore/tables/" with a "(optional)" note. Below that is a file upload area with a placeholder "Drop files to upload, or click to browse". To the right of this area is a Windows File Explorer window showing the "Downloads" folder. The "1000\_Companies" file is selected. A blue arrow points from the "Drop files to upload, or click to browse" area to the "1000\_Companies" file in the File Explorer.

Step4: after uploading successfully click on create table in notebook

The screenshot shows the 'Create New Table' page in the Microsoft Azure Databricks interface. On the left, there is a sidebar with various icons for navigation and workspace management. The main area has a title 'Create New Table' and a 'Data source' section with tabs for 'Upload File' (which is selected), 'DBFS', and 'Other Data Sources'. Under 'Upload File', there is a 'DBFS Target Directory' input field containing '/FileStore/tables/' with '(optional)' text next to it, and a 'Select' button. Below this, a note states 'Files uploaded to DBFS are accessible by everyone who has access to this workspace.' A 'Learn more' link is provided. The 'Files' section shows a single file named '1000\_Compan' with a yellow checkmark icon overlaid on it, indicating success. The file size is listed as '51.2 KB' and there is a 'Remove file' link. At the bottom of the 'Files' section, a green success message says '✓ File uploaded to /FileStore/tables/1000\_Companies-1.csv'. There are two buttons at the bottom: 'Create Table with UI' (blue) and 'Create Table in Notebook' (yellow, with a yellow highlight box around it). A small icon of a diamond shape with arrows is to the left of the buttons. To the right of the buttons, there is a callout box with the text 'Looking for other ways to add data? Visit Partner Connect.' and 'Use our ingestion partners to load data from various products and databases into Delta Lake.'

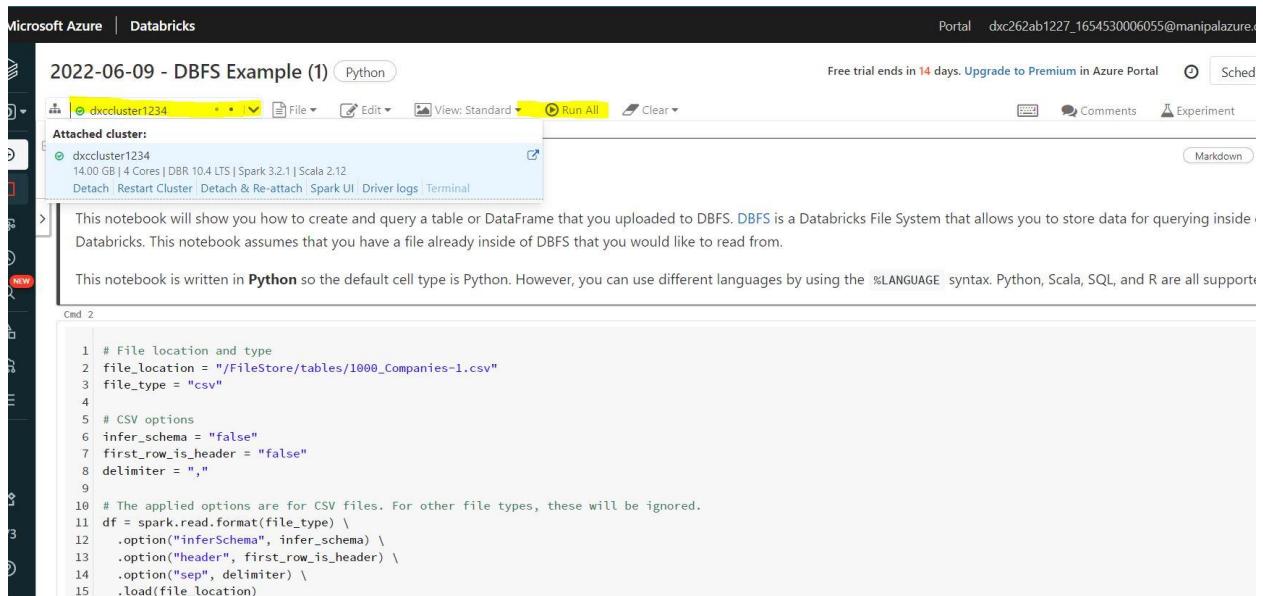
## Step5: attach the cluster in the menu bar.



The screenshot shows a Microsoft Azure Databricks notebook titled "2022-06-09 - DBFS Example (1)". The notebook is set to Python. The "Overview" section contains introductory text about DBFS and its capabilities. Below it, a code cell (Cmd 1) contains the following Python code:

```
1 # File location and type
2 file_location = "/FileStore/tables/1000_Companies-1.csv"
3 file_type = "csv"
4
5 # CSV options
6 infer_schema = "false"
7 first_row_is_header = "false"
8 delimiter = ","
9
10 # The applied options are for CSV files. For other file types, these will be ignored.
11 df = spark.read.format(file_type) \
12     .option("inferSchema", infer_schema) \
13     .option("header", first_row_is_header) \
14     .option("sep", delimiter) \
15     .load(file_location)
```

## Step6: after attaching the cluster click on run all



The screenshot shows the same Microsoft Azure Databricks notebook interface as before, but now with an attached cluster named "dxccluster1234" selected in the menu bar. The notebook title remains "2022-06-09 - DBFS Example (1)" and the code cell content is identical to Step 5.

Step7: hence the results are executed.

```
5     .load(file_location)
6
7 display(df)
```

▶ (2) Spark Jobs  
▶ df: pyspark.sql.dataframe.DataFrame = [c0: string, c1: string ... 3 more fields]

table Data Profile

	c0	c1	c2	c3	c4
1	R&D Spend	Administration	Marketing Spend	State	Profit
2	165349.2	136897.8	471784.1	New York	192261.83
3	162597.7	151377.59	443898.53	California	191792.06
4	153441.51	101145.55	407934.54	Florida	191050.39
5	144372.41	118671.85	383199.62	New York	182901.99
6	142107.34	91391.77	366168.42	Florida	166187.94
7	131876.9	99814.71	362861.36	New York	156991.12

Truncated results, showing first 1000 rows.

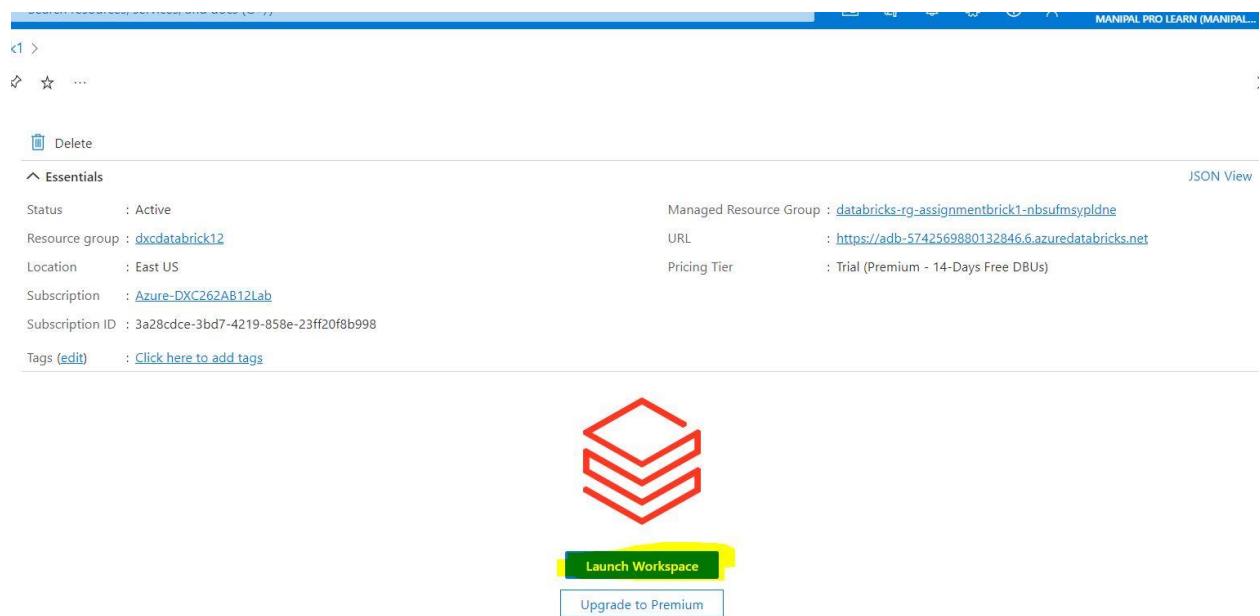
Click to re-execute with maximum result limits.



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

A:

Step1: open overview page of data bricks and click on launch work space

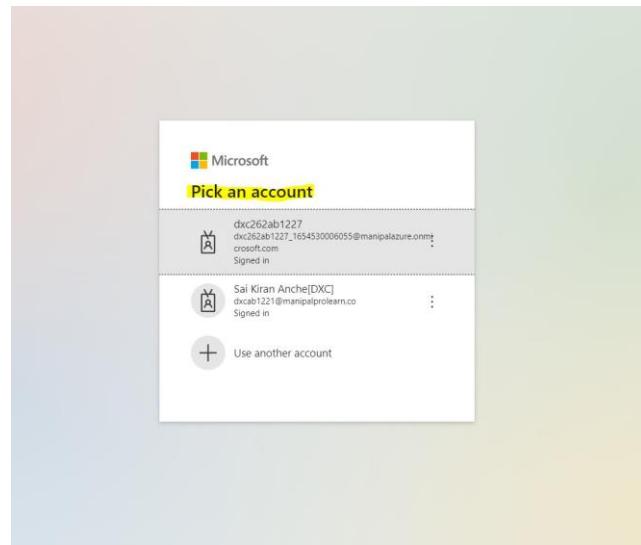


The screenshot shows the Databricks workspace overview page. At the top, there's a navigation bar with icons for back, forward, search, and other options. The title 'MANIPAL PRO LEARN (MANIPAL...)' is visible. Below the title, there's a breadcrumb trail with 'k1 >'. Underneath the title, there are several status indicators: 'Delete', 'Essentials', and 'JSON View'. The 'Essentials' section contains the following details:

Status	: Active
Resource group	: <a href="#">dxcdatabrick12</a>
Location	: East US
Subscription	: <a href="#">Azure-DXC262AB12Lab</a>
Subscription ID	: 3a28cdce-3bd7-4219-858e-23ff20f8b998
Tags (edit)	: <a href="#">Click here to add tags</a>

On the right side of the essentials section, there are two status indicators: 'Managed Resource Group' (databricks-rg-assignmentbrick1-nbsufmsypldne) and 'Pricing Tier' (Trial (Premium - 14-Days Free DBUs)). Below the essentials section, there's a large red 'Databricks' logo icon. Underneath the logo, there are two buttons: a blue 'Launch Workspace' button and a white 'Upgrade to Premium' button.

Step2: pick an account for the work space



Step3: click on compute and then click on create cluster.

The screenshot shows the Microsoft Azure Databricks Compute interface. On the left is a dark sidebar with various icons: a stack of cubes (Compute), a square with a 'D' (Databricks), a plus sign (New Cluster), a search icon (Search), and a gear (Settings). The main area has a header with the Microsoft Azure and Databricks logos. Below the header, the word "Compute" is displayed, followed by "Preview" and "Provide feedback". There are four tabs: "All-purpose clusters" (which is underlined, indicating it is selected), "Job clusters", "Pools", and "Cluster policies". A large green button labeled "Create Cluster" is prominently displayed. Below this button is a search bar with the placeholder "Name". To the right of the search bar are dropdown menus for "Policy" and "Runtime". A tooltip "Depending on you" is visible near the "Runtime" dropdown. The bottom of the sidebar shows the status "0/3" and three small icons: a question mark, a gear, and a downward arrow.

Step4: give the name and select the cluster mode , termination time etc and click on create cluster

The screenshot shows the 'New Cluster' configuration page in the Microsoft Azure Databricks interface. On the left, there's a sidebar with various icons for cluster management. The main area has the following fields:

- Cluster name:** training1234
- Cluster mode:** Single Node
- Databricks runtime version:** Runtime: 10.4 LTS (Scala 2.12, Spark 3.2.1)
- Autopilot options:** A checked checkbox for "Terminate after 30 minutes of inactivity".
- Node type:** Standard\_DS3\_v2 (14 GB Memory, 4 Cores)
- DBU / hour:** 0.75

A prominent green button at the top right says "Create Cluster". Above the "Create Cluster" button, it says "DBU / hour: 0.75" and "0 Workers: 0 GB Memory, 0 Cores 1 Driver: 14 GB Memory, 4 Cores". A message box at the bottom left indicates a "50% promotional discount applied to Photon during preview".

Step5: your cluster will be validating.

The screenshot shows the Microsoft Azure Databricks Cluster configuration interface. The cluster name is 'training1234'. The 'Configuration' tab is selected. Key settings include:

- Policy:** Unrestricted
- Cluster mode:** Single Node
- Databricks Runtime Version:** 10.4 LTS (includes Apache Spark 3.2.1, Scala 2.12)
- Autopilot options:**  Terminate after 30 minutes of inactivity
- Node type:** Standard\_DS3\_v2 (14 GB Memory, 4 Cores)
- DBU / hour:** 0.75
- Standard\_DS3\_v2** is highlighted in purple, indicating it is the selected node type.

A yellow oval highlights the cluster name 'training1234' and the node type 'Standard\_DS3\_v2'.

Step6: after successful creation of a cluster a green tick will appear

The screenshot shows the Microsoft Azure Databricks interface for managing clusters. The top navigation bar includes 'Microsoft Azure' and 'Databricks'. On the left, there's a sidebar with various icons for navigation and management, including a plus sign for creating new resources, a search icon, and a 'NEW' badge.

The main content area is titled 'Clusters / dxcluster1234'. Below the title, the cluster name 'dxcluster1234' is displayed with a green checkmark indicating it has been successfully created. A yellow highlight is placed over the checkmark.

The configuration tab is selected, showing the following settings:

- Policy:** Unrestricted
- Cluster mode:** Single Node
- Databricks Runtime Version:** 10.4 LTS (includes Apache Spark 3.2.1, Scala 2.12)
- Autopilot options:**  Terminate after 30 minutes of inactivity
- Node type:** Standard\_DS3\_v2 (14 GB Memory, 4 Cores)
- DBU / hour:** 0.75
- Standard\_DS3\_v2** (highlighted in purple, indicating the selected node type)

At the bottom, there's a link to 'Advanced options'.

## Step7: your cluster is successfully created.

The screenshot shows the Microsoft Azure Databricks Compute interface. At the top, there are three tabs: "Subscription Details | Nuvepro", "assignmentbrick1 - Microsoft A...", and "Compute - Databricks". The "Compute - Databricks" tab is active. Below the tabs, the URL is "adb-5742569880132846.6.azuredatabricks.net/?o=5742569880132846#setting/clusters". On the left, there's a sidebar with various icons. The main area is titled "Compute" and has a "Create Cluster" button. Below it, there are tabs for "All-purpose clusters", "Job clusters", "Pools", and "Cluster policies". A search bar at the top right says "Free trial ends in 14 days. Upgrade to Premium in Azure Portal". The "All-purpose clusters" table has columns: Name, Policy, Runtime, Active memory, Active cores, Active DBU / h, Source, Creator, and a delete icon. One row is highlighted with a yellow background and shows the cluster name "decluster1234". The table shows a total count of 1 cluster.

Name	Policy	Runtime	Active memory	Active cores	Active DBU / h	Source	Creator	Actions
decluster1234	-	10.4	14 GB	4 cores	0.75	UI	dxc262ab1227_1654530006055@manipalazure.onmicrosoft.com	[Delete]

## **RESULT**

Almost all the test questions have been solved and presented successfully in the present document.

## **CONCLUSIONS**

All the questions have been solved successfully with all the concepts that have been covered in the training session. It's really a great experience of learning while solving the cases. This assignment gave me immense confidence regarding my ability to upskill in new technologies.