

31/08/2023 DAY 3 CUSTOM TRAINING

Server

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

Home > Create a resource > Create SQL Database >

Create SQL Database Server

Microsoft

Server name * .database.windows.net

Location *

Authentication

Select your preferred authentication methods for accessing this server. Create a server admin login and password to access your server with SQL authentication, select only Azure AD authentication [Learn more](#) using an existing Azure AD user, group, or application as Azure AD admin [Learn more](#), or select both SQL and Azure AD authentication.

Authentication method Use only Azure Active Directory (Azure AD) authentication
 Use both SQL and Azure AD authentication
 Use SQL authentication

Server admin login *

Password *

Confirm password *

OK

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

Home > Create a resource >

Create SQL Database

Microsoft

Project details

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

Subscription *

Resource group * [Create new](#)

Database details

Enter required settings for this database, including picking a logical server and configuring the compute and storage resources.

Database name *

Server * [Create new](#)

Want to use SQL elastic pool? Yes No

Workload environment Development
 Production

General Purpose (GP_S_Gen5_1)
Cost per GB (in --) --
Max storage selected (in GB) x 41.6
ESTIMATED STORAGE COST / MONTH -- --
COMPUTE COST / VCORE SECOND ! -- --

NOTES
↑ Serverless databases are billed in vCore seconds based on a combination of CPU and memory utilization. [Learn more about serverless billing](#)

PLEASE CONTACT YOUR RESELLER

Review + create **Next : Networking >**

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

Home > Create a resource >

Create SQL Database

Workload environment Development Production

Default settings provided for Development workloads. Configurations can be modified as needed.

Compute + storage * General Purpose - Serverless
Standard-series (Gen5), 1 vCore, 32 GB storage, zone redundant disabled
[Configure database](#)

Backup storage redundancy
Choose how your PITR and LTR backups are replicated. Geo restore or ability to recover from regional outage is only available when geo-redundant storage is selected.

Backup storage redundancy Locally-redundant backup storage
 Zone-redundant backup storage
 Geo-redundant backup storage

[Review + create](#) [Next : Networking >](#)

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

Home > Create a resource >

Configure

Feedback

Service and compute tier
Select from the available tiers based on the needs of your workload. The vCore model provides a wide range of configuration controls and offers Hyperscale and Serverless to automatically scale your database based on your workload needs. Alternately, the DTU model provides set price/performance packages to choose from for easy configuration. [Learn more](#)

Service tier Standard (Budget friendly) [Compare service tiers](#)

DTUs [Compare DTU options](#)
10

Data max size (GB)
2

SQL

Cost summary

Standard (\$0)	Cost per DTU (in --)	--
DTUs selected	x 10	--
ESTIMATED COST / MONTH	--	--

PLEASE CONTACT YOUR RESELLER

[Apply](#)

Create SQL Database

Networking

Configure network access and connectivity for your server. The configuration selected below will apply to the selected server 'trainingserver' and all databases it manages. [Learn more](#)

Network connectivity

Choose an option for configuring connectivity to your server via public endpoint or private endpoint. Choosing no access creates with defaults and you can configure connection method after server creation. [Learn more](#)

Connectivity method * No access Public endpoint Private endpoint

Firewall rules

Setting 'Allow Azure services and resources to access this server' to Yes allows communications from all resources inside the Azure boundary, that may or may not be part of your subscription. [Learn more](#)

Setting 'Add current client IP address' to Yes will add an entry for your client IP address to the server firewall.

Allow Azure services and resources to access this server * No Yes

Add current client IP address * No Yes

Cost summary

Standard (\$0)

Cost per DTU (in --) DTUs selected x 10 ESTIMATED COST / MONTH --

PLEASE CONTACT YOUR RESELLER

[Review + create](#) < Previous Next : Security >

Click on Review and create then create

Microsoft SQL Database newDatabaseNewServer_f7682559a6674db78b880 | Overview

Deployment

Deployment is in progress

Resource	Type	Status
trainingserver...	SQL server	Accepted

Give feedback [Tell us about your experience with deployment](#)

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 >](#)

Then go to home and you can view your resource group and sqldatabase.

Click on view and click on query editor preview and give user id and pass

[Query SQL Database with Query editor in the Azure portal - Azure SQL Database | Microsoft Learn](#)

The screenshot shows the Microsoft Azure portal interface. At the top, there's a navigation bar with 'Microsoft Azure' and a search bar. Below it, the URL 'Home > training_database (trainingservererm/training_database)' is visible. The main content area is titled 'training_database (trainingservererm/training_database) | Query editor (preview)'. On the left, there's a sidebar with various database management options like Overview, Activity log, Tags, Diagnose and solve problems, and Query editor (preview). The 'Query editor (preview)' option is currently selected. The main pane shows the database name 'training_database (aastha)' and a message: 'Showing limited object explorer here. For full capability please click here to open Azure Data Studio.' Below this, there are three expandable sections: Tables, Views, and Stored Procedures. To the right of the sidebar is a large query editor window titled 'Query 1'. It has tabs for 'Run', 'Cancel query', 'Save query', 'Export data as', and 'Show only Editor'. The editor area contains a single digit '1'. At the bottom of the editor is a 'Results' tab and a search bar. A status bar at the bottom right says 'Ready'.

Deployment Models: single, elastic pools, managed instance

Sql commands category- 4/5 categories:

1.DDL- create, alter, drop, truncate

Diff between truncate(will delete all the rows) and delete(used with conditions and can be undone)

2.DML- insert, update, delete, select

3.DCL- grant, revoke

4.ECL-commit, rollback, save point

```
create table students(id int, student_name varchar(30), address1 varchar(30), city varchar(30))
```

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

Home > training_database (trainingservererm/training_database)

training_database (trainingservererm/training_database) | Query editor (preview)

SQL database

Search Login New Query Open query Feedback Getting started

Overview Activity log Tags Diagnose and solve problems Query editor (preview) Settings Compute + storage Connection strings Properties Locks Data management Backstage

training_data... (Showing limited object explorer here. For full capability please click here to open Azure Data Studio.)

Tables Views Stored Procedure

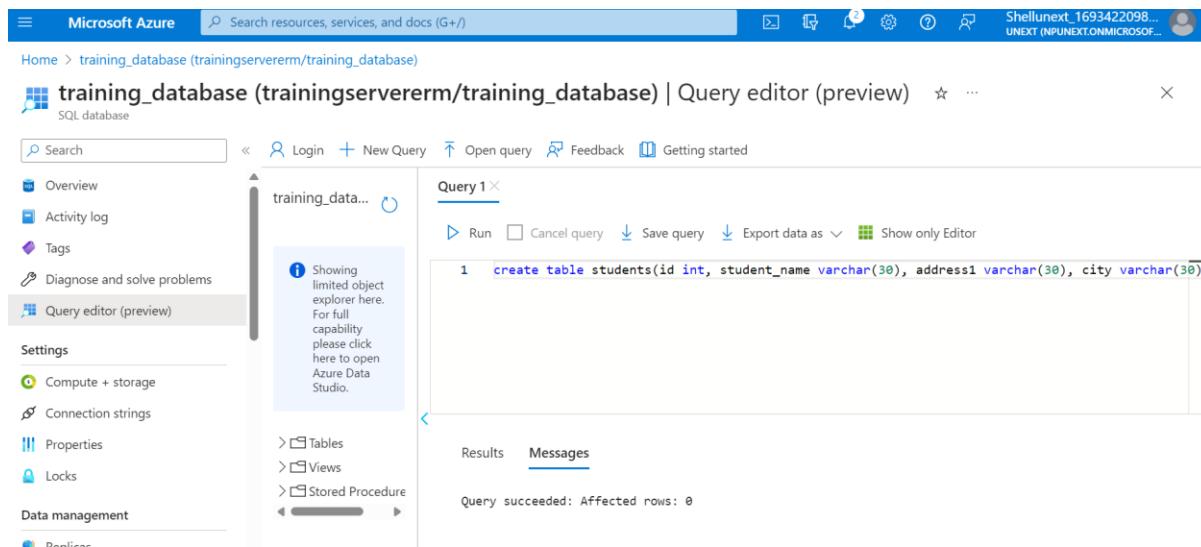
Query 1

Run Cancel query Save query Export data as Show only Editor

```
1 create table students(id int, student_name varchar(30), address1 varchar(30), city varchar(30))
```

Results Messages

Query succeeded: Affected rows: 0



sp_tables- show tables

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

Home > training_database (trainingservererm/training_database)

training_database (trainingservererm/training_database) | Query editor (preview)

SQL database

Search Login New Query Open query Feedback Getting started

Overview Activity log Tags Diagnose and solve problems Query editor (preview) Settings Compute + storage Connection strings Properties Locks Data management Replicas Sync to other databases Integrations Azure Synapse Link

training_data... (Showing limited object explorer here. For full capability please click here to open Azure Data Studio.)

Tables Views Stored Procedure

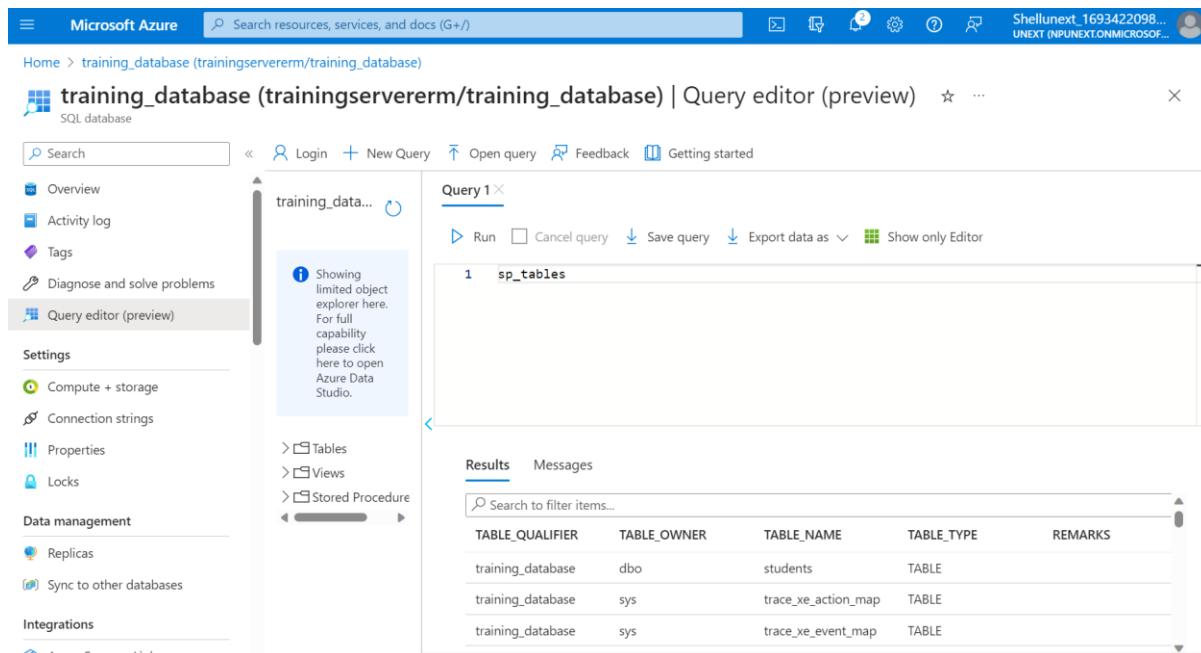
Query 1

Run Cancel query Save query Export data as Show only Editor

```
1 sp_tables
```

Results Messages

TABLE_QUALIFIER	TABLE_OWNER	TABLE_NAME	TABLE_TYPE	REMARKS
training_database	dbo	students	TABLE	
training_database	sys	trace_xe_action_map	TABLE	
training_database	sys	trace_xe_event_map	TABLE	



Microsoft Azure Search resources, services, and docs (G+/-) Home > training_database (trainingservererm/training_database)

training_database (trainingservererm/training_database) | Query editor (preview)

Search Login New Query Open query Feedback Getting started

Overview Activity log Tags Diagnose and solve problems Query editor (preview) Settings Compute + storage Connection strings Properties Locks Replicas Sync to other databases Integrations Azure Synapse Link

Showing limited object explorer here. For full capability please click here to open Azure Data Studio.

Tables Views Stored Procedure

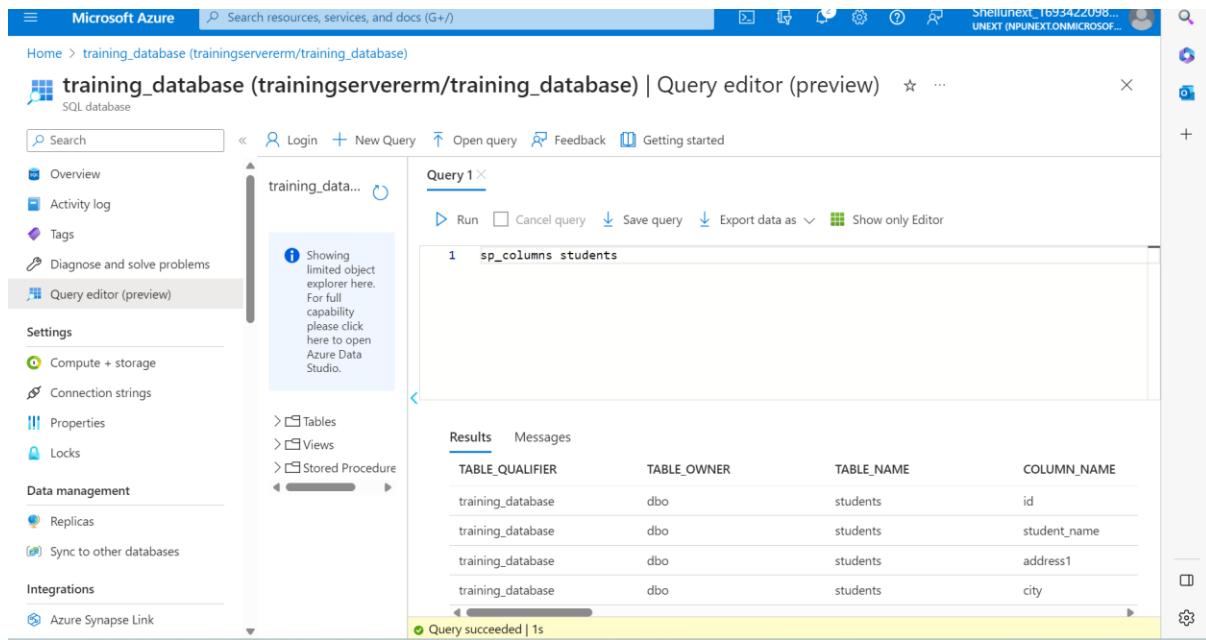
Query 1

Run Cancel query Save query Export data as Show only Editor

1 sp_columns students

TABLE_QUALIFIER	TABLE_OWNER	TABLE_NAME	COLUMN_NAME
training_database	dbo	students	id
training_database	dbo	students	student_name
training_database	dbo	students	address1
training_database	dbo	students	city

Query succeeded | 1s



```
insert into students values(1,'Navya','HSR','bangalore')
insert into students(id, student_name) values (2,'Anushka')
insert into students values(1,'Anupa',null,'bangalore')
```

Microsoft Azure Search resources, services, and docs (G+/-) Home > training_database (trainingservererm/training_database)

training_database (trainingservererm/training_database) | Query editor (preview)

Search Login New Query Open query Feedback Getting started

Overview Activity log Tags Diagnose and solve problems Query editor (preview) Settings Compute + storage Connection strings Properties Locks Replicas Sync to other databases Integrations Azure Synapse Link

Showing limited object explorer here. For full capability please click here to open Azure Data Studio.

Tables Views Stored Procedure

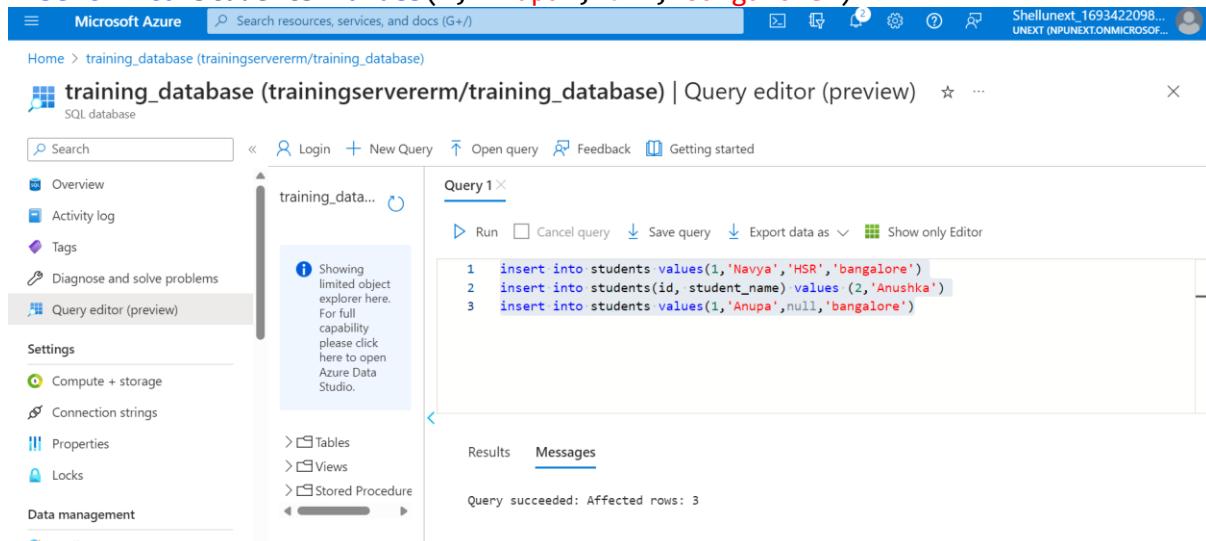
Query 1

Run Cancel query Save query Export data as Show only Editor

```
1 insert into students values(1,'Navya','HSR','bangalore')
2 insert into students(id, student_name) values (2,'Anushka')
3 insert into students values(1,'Anupa',null,'bangalore')
```

Results Messages

Query succeeded: Affected rows: 3



```
select student_name,id from students
```

The screenshot shows the Microsoft Azure Query editor (preview) interface. The left sidebar lists database management options like Overview, Activity log, Tags, Diagnose and solve problems, and the active Query editor (preview). The main area displays a query window titled 'Query 1' containing the SQL command: 'select student_name,id from students'. Below the query, the results pane shows a table with two columns: 'student_name' and 'id', listing three rows: Navya (id 1), Anushka (id 2), and Anupa (id 1). Navigation arrows on the right side of the results pane allow for scrolling through the data.

```

select * into student1 from students
select * from students
NO SS AS NET NOT WORKING
select * into student2 from students where 10=20
insert into students2 select * from students
select * from students
NO SS AS NET NOT WORKING

```

Basically we need to use existing table to create new table

```

select id as student_id,student_name into students4 from students
NO SS AS NET NOT WORKING
select id as student_id,student_name into students5 from students
insert into students5 select id, student_name from students.
NO SS AS NET NOT WORKING

```

```

Alter table student7 add city varchar(30)
Alter table student7 add email...
Alter table student7 drop column email
Alter table student7
Alter column name varchar(100)
Alter column name varchar(100) not null

```

To rename an existing col- Sp_rename 'student7.name', student_name, 'column'
 To rename whole table name with a new table name- sp_rename student7,
 student_back

Constraints- not null, unique, primary key, check constraint.

Missed 1-hour stuff

```
Create table suppliers(supplier_id int primary key, supplier_name varchar(30));
Create table prod(product_id int primary key, product_name varchar(30), supplier_id int constraint
prod_fk foreign key, references suppliers(supplier_id))

Insert into suppliers values (100,'P&G')
Insert into suppliers values (100,'HUL')
Insert into suppliers values (100,'J&J')
Insert into prod values (1,'paste',100)
Insert into prod values (1,'perfume',101)
Insert into prod values (2,'deodorant',102)
Insert into prod values (3,'baby powder',103)
Delete from suppliers where supplier_id =100
Drop table suppliers
```

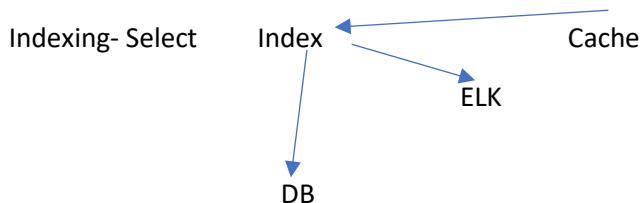
[EMP and DEPT table script – SQL Server – Data Analytics and BI WORLD \(wordpress.com\)](#)

Day- 05/09/2023

Azure Storage services, pic on mobile

SMB server- client is ASA

Find slide on lumen about this



1mb-cache

1gb- RAM

1TB- HDD

Create resource

Microsoft Azure Search resources, services, and docs (G+) Shellunext_1693422098...
UNEXT (NPUNEXT.ONMICROSOFT.COM)

Home > Resource groups >

Create a resource group

Validation passed.

Basics Tags Review + create

Basics

Subscription npunext-1673505240396
Resource group AS009
Region East US

Tags

None

Create

< Previous

Next >

Download a template for automation

Microsoft Azure Search resources, services, and docs (G+) Shellunext_1693422098...
UNEXT (NPUNEXT.ONMICROSOFT.COM)

Home > Storage accounts >

Create a storage account

Basics Advanced Networking Data protection Encryption Tags Review

Enable storage account key access

Default to Azure Active Directory authorization in the Azure portal

Minimum TLS version

Permitted scope for copy operations (preview)

Hierarchical Namespace

Hierarchical namespace, complemented by Data Lake Storage Gen2 endpoint, enables file and directory semantics, accelerates big data analytics workloads, and enables access control lists (ACLs) [Learn more](#)

Enable hierarchical namespace

Review

< Previous

Next : Networking >

Give feedback

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

Home > Storage accounts >

Create a storage account

Basics Advanced Networking Data protection Encryption Tags Review

Enable soft delete for blobs
Soft delete enables you to recover blobs and directories that were previously marked for deletion. [Learn more](#)
Days to retain deleted blobs

Enable soft delete for containers
Soft delete enables you to recover containers that were previously marked for deletion. [Learn more](#)
Days to retain deleted containers

Enable soft delete for file shares
Soft delete enables you to recover file shares that were previously marked for deletion. [Learn more](#)
Days to retain deleted file shares

Tracking
Manage versions and keep track of changes made to your blob data.

[Review](#) [< Previous](#) [Next : Encryption >](#) [Give feedback](#)

2 encryption keys- MMK and CMK

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

Home > Storage accounts >

Create a storage account

Basics Advanced Networking Data protection Encryption Tags Review

Encryption type * Microsoft-managed keys (MMK) Customer-managed keys (CMK)

Enable support for customer-managed keys Blobs and files only All service types (blobs, files, tables, and queues)
⚠ This option cannot be changed after this storage account is created.

Enable infrastructure encryption

Microsoft Azure Search resources, services, and docs (G+) Shellunext_1693422098... UNEXT (NPUNEXTONMICROSOFT)

Home > storageas1099_1693889463083 | Overview

storageas1099_1693889463083 Deployment

Search Delete Cancel Redeploy Download Refresh

Overview Deployment Inputs Outputs Template

Your deployment is complete

Deployment name: storageas1099_169... Start time: 9/5/2023, 10:21:11 AM
Subscription: npunext-1673505240396 Correlation ID: f0a035f4-1110-4318-adc4-5477855
Resource group: AS009

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.

Microsoft Azure Search resources, services, and docs (G+) Shellunext_1693422098... UNEXT (NPUNEXTONMICROSOFT)

Home > storageas1099

storageas1099 | Containers

Storage account

Search Container Change access level Restore containers Refresh Delete Give feedback

Search containers by prefix Show deleted containers

Name	Last modified	Anonymous access level	Lease state
\$logs	9/5/2023, 10:21:43 AM	Private	Available

Containers File shares Queues Tables

Data storage

Containers (selected) File shares Queues Tables

Security + networking

After enabled static website, give index.html and error.html name and save

Home > storageas1099

storageas1099 | Static website

Storage account

Search Save Discard Give feedback

Overview Activity log Tags Diagnose and solve problems Access Control (IAM) Data migration Events Storage browser

Data storage Containers File shares Queues Tables

Security + networking Networking

Enabling static websites on the data lake service allows you to host static content. Webpages may include static content and client-side scripts. Server-side scripting is not supported. If a static website is enabled on a data lake service, ACLs will not be honored. As data is replicated asynchronously from primary to secondary regions, files at the secondary endpoint may not be immediately available or in sync with files at the primary endpoint. [Learn more](#)

Static website

Disabled Enabled

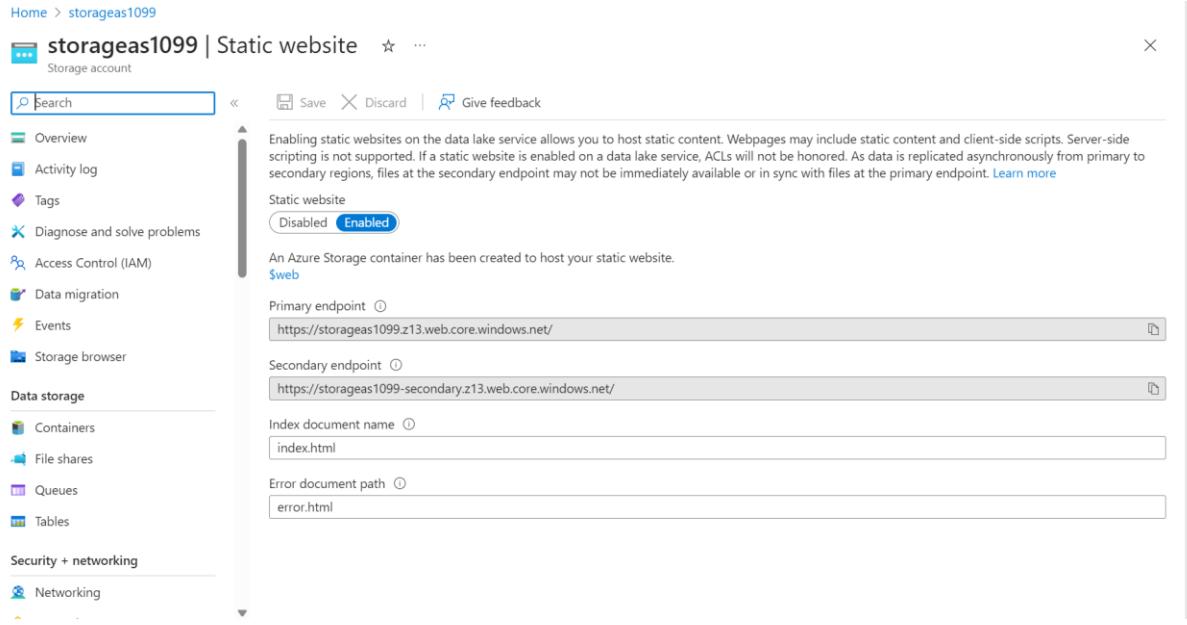
An Azure Storage container has been created to host your static website. \$web

Primary endpoint https://storageas1099.z13.web.core.windows.net/

Secondary endpoint https://storageas1099-secondary.z13.web.core.windows.net/

Index document name index.html

Error document path error.html



Then click on \$web and upload the files

Creation of alert

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

Home > st1099 | Alerts > Create an alert rule

Scope Condition Actions Details Tags Review + create

Configure when the alert rule should trigger by selecting a signal and defining its logic.

Signal name * Egress See all signals

Alert logic

Threshold Static Dynamic

Aggregation type Total

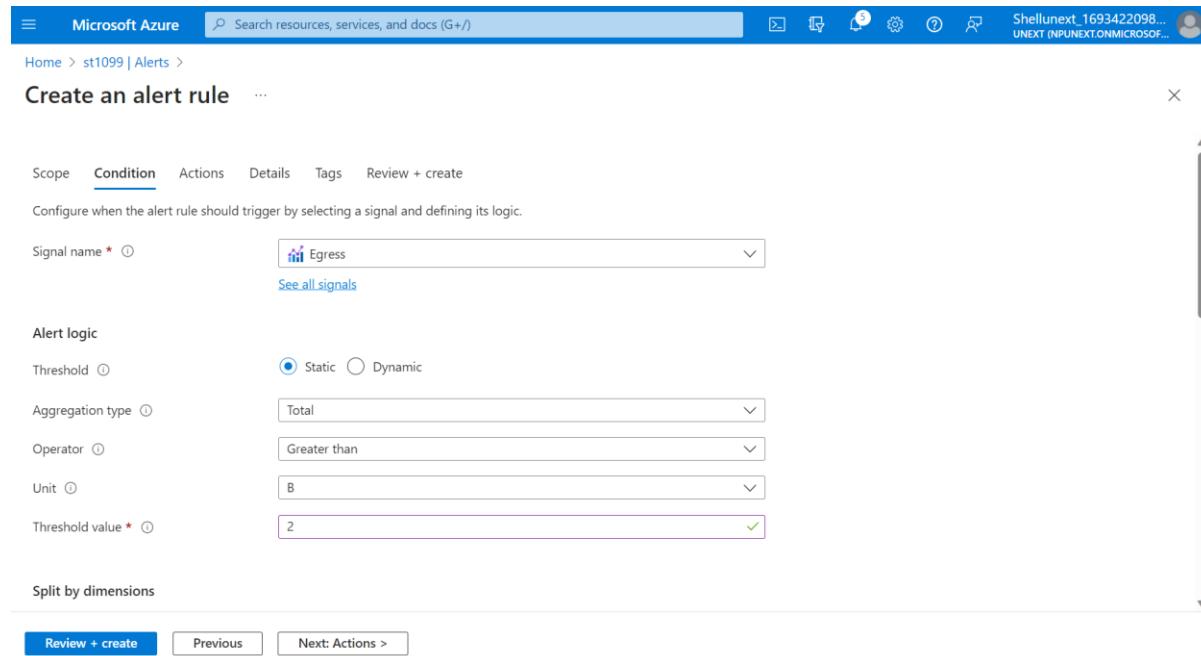
Operator Greater than

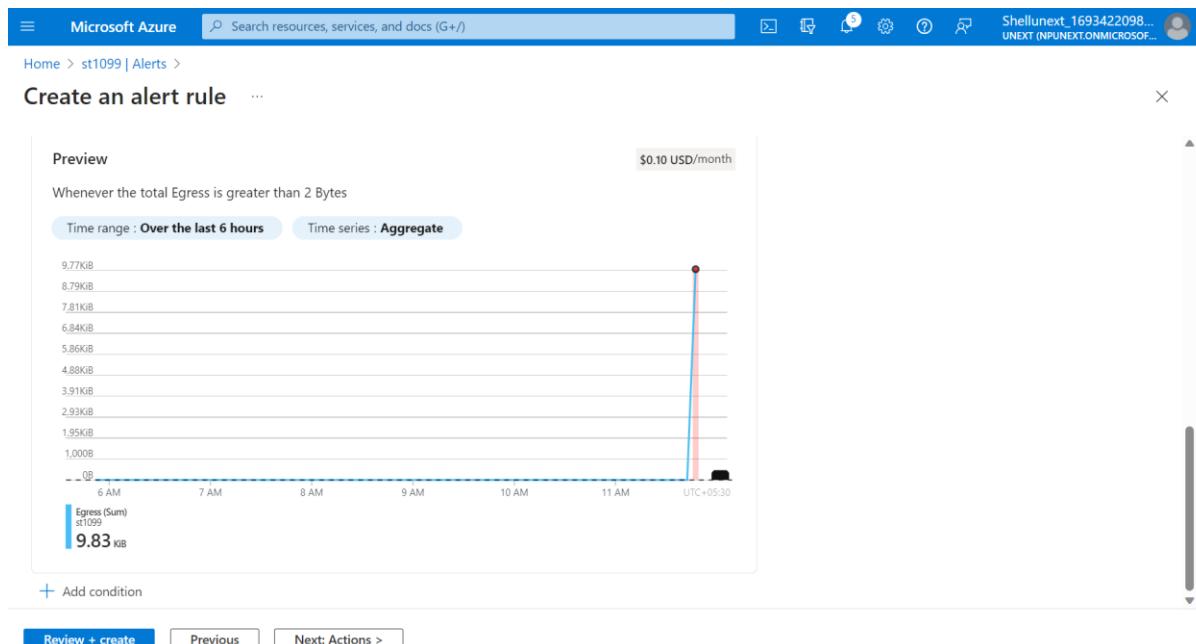
Unit B

Threshold value * 2

Split by dimensions

Review + create Previous Next: Actions >





In containers make a new container file and in \$web and that container upload the index.html and error.html files and these files can be viewed through static website links provided.

The screenshot shows the 'Containers' page for the storage account 'st1099'. The left sidebar includes options like Overview, Activity log, Tags, Diagnose and solve problems, Access Control (IAM), Data migration, Events, and Storage browser. The main area lists containers with columns for Name, Last modified, Anonymous access level, and Lease state. The container '\$web' is selected, indicated by a blue border. The 'Show deleted containers' toggle is off.

The screenshot shows the 'Containers' page for the container 'as1099'. The left sidebar includes Overview, Diagnose and solve problems, Access Control (IAM), Settings (Shared access tokens, Manage ACL, Access policy, Properties, Metadata), and a blob search bar. The main area shows a table of blobs with columns for Name, Modified, Access tier, Archive status, and Blob type. Two blobs are listed: 'error.html' and 'index.html', both of which are 'Block blob' type.

After that you can see the fired alert in the alert section.

The screenshot shows the Microsoft Azure Alerts page for a storage account named 'st1099'. A single alert titled 'alert1099' is listed under the 'Total alerts' section. The alert details show it was fired at 9/5/2023, 11:54 AM, with a severity of 3 - Informational. The affected resource is 'st1099'. A chart below shows a sharp increase in Egress (Sum) from 0B to 64.2 kB between 11:45 AM and 11:54 AM. The alert condition is marked as 'Fired'.

Deleting a resource group

The screenshot shows the Microsoft Azure Resource groups page for a resource group named 'as1099'. A modal window titled 'Deleting resource group as1099' is open, indicating the action is in progress. The main table lists two resources: 'alert1099' (Metric alert rule) and 'st1099' (Storage account).

Name	Type	Location	Actions
alert1099	Metric alert rule	Global	...
st1099	Storage account	East US	...

Mstsc- server connector

Practical 2-Now create a new resource group in it click on create select virtual machine and then create machine with the required specification.

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

Home > vm1009 > Marketplace > Virtual machine > Create a virtual machine

Project details

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

Subscription * Resource group * [Create new](#)

Instance details

Virtual machine name * Region * Availability options Security type [Configure security features](#)

Image * [See all images](#) | [Configure VM generation](#)

[Review + create](#) [< Previous](#) [Next : Disks >](#) [Give feedback](#)

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

Home > vm1009 > Marketplace > Virtual machine > Create a virtual machine

Create a virtual machine

Username * Password * Confirm password *

Inbound port rules

Select which virtual machine network ports are accessible from the public internet. You can specify more limited or granular network access on the Networking tab.

Public inbound ports * None Allow selected ports

Select inbound ports *

All traffic from the internet will be blocked by default. You will be able to change inbound port rules in the VM > Networking page.

Licensing

[Review + create](#) [< Previous](#) [Next : Disks >](#) [Give feedback](#)

Microsoft Azure Search resources, services, and docs (G+) Shellunext_1693422098... UNEXT (NPUNEXT.ONMICROSOFT.COM)

Home > CreateVm-MicrosoftWindowsDesktop.Windows-10-win10-20230905121313 | Overview

Deployment

Search Delete Cancel Redeploy Download Refresh

Overview Deployment

Your deployment is complete

Deployment name: CreateVm-MicrosoftWindowsDes... Start time: 9/5/2023, 12:15:59 PM
Subscription: npunext-1673505240396 Correlation ID: 57e518ac-8853-4b4e
Resource group: vm1009

Deployment details

Next steps

Setup auto-shutdown Recommended
Monitor VM health, performance and network dependencies Recommended
Run a script inside the virtual machine Recommended

Go to resource Create another VM

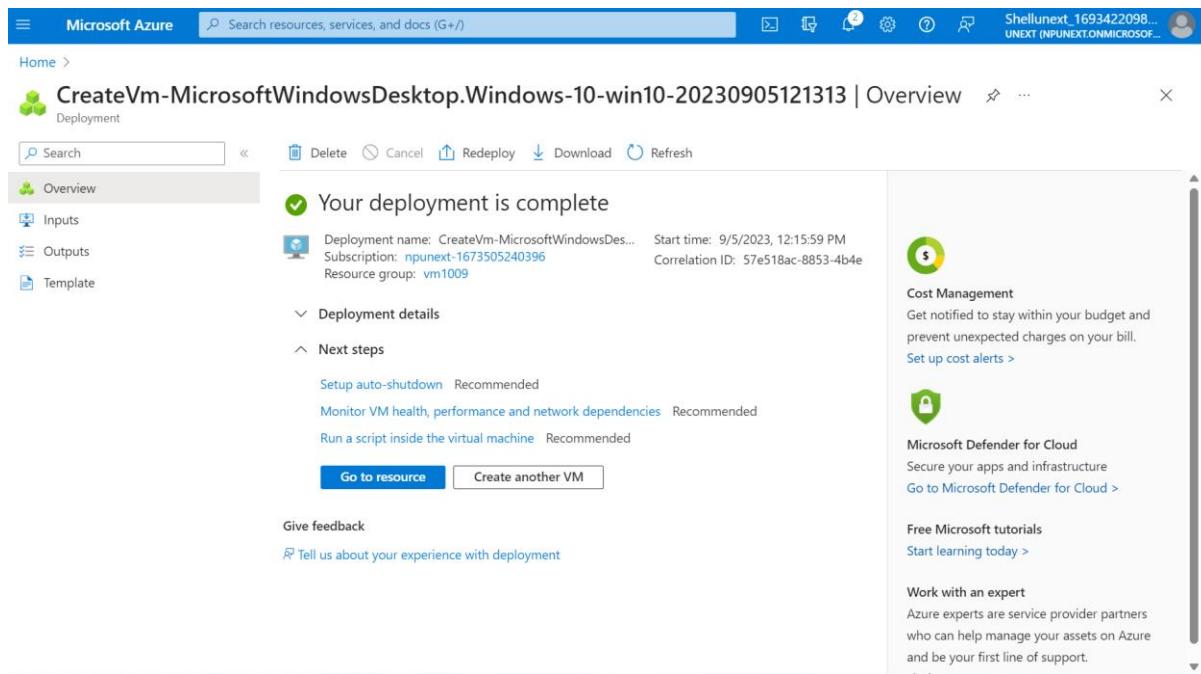
Give feedback Tell us about your experience with deployment

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 >

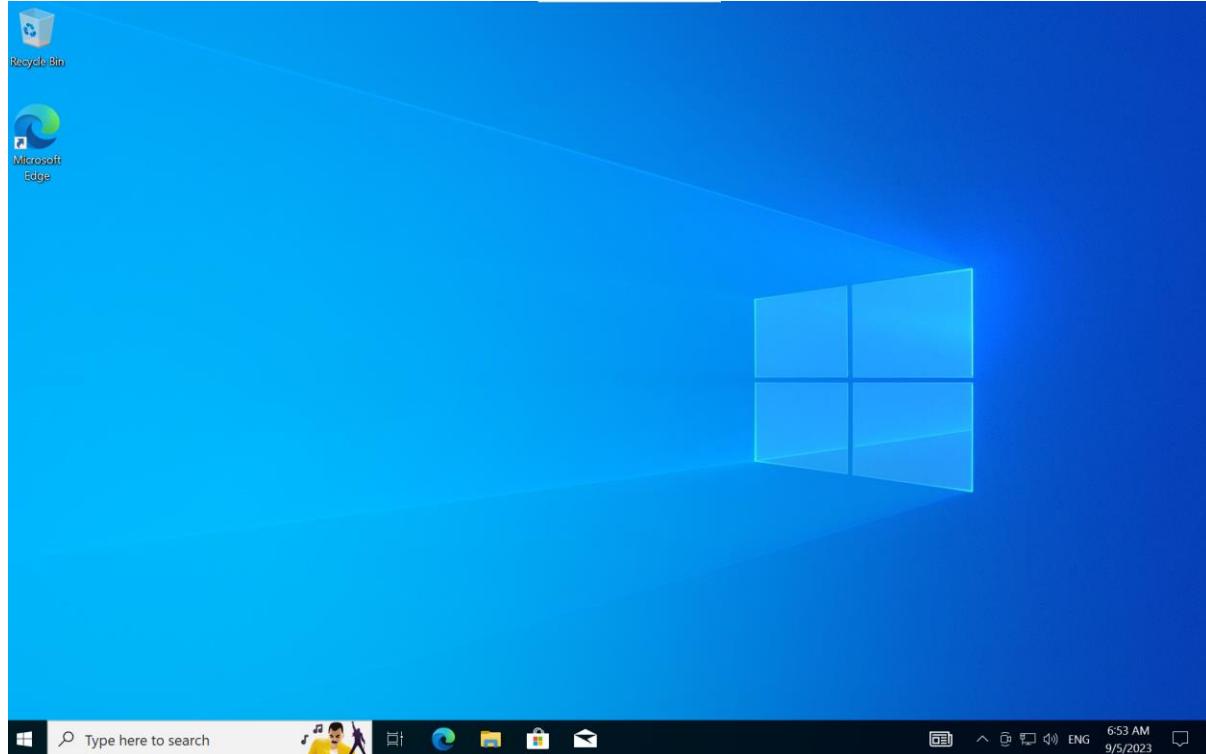
Free Microsoft tutorials Start learning today >

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.



After this in run type mstsc and the copy the virtual machine ip address and user password and log in

You would see this next



Open cmd and download

ADF will take data from storage and pass to SQL. ADF link service bring others to ADF.

Server list/link and adf

The screenshot shows the Microsoft Azure Data Factory portal. At the top, there's a navigation bar with 'Microsoft Azure', 'Data Factory', and a search bar. A notification bar at the top right says 'Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!' Below the navigation, there's a sidebar with icons for Home, New, and Recent resources. The main area is titled 'datafactoryasd' and features a large diagram illustrating the Data Factory process: Ingest (Copy data at scale once or on a schedule), Orchestrate (Code-free data pipelines), and Transform data (Transform your data using data flows). There are also links for 'Set up code repository' and 'Configure SSIS'.

Created a adf under resource group

What we will do now is use adf to connect to 2 different Storage account.

Basically create- RG, ADF, SA, 2 CONTAINERS(input and output), Emp.txt, link services for storage account in adf under manage tab, adf-author-select pipeline-copy task-from where you want to copy and paste, then just execute and check

The screenshot shows the Microsoft Azure Storage account 'asdstorage'. The left sidebar has options like Overview, Activity log, Tags, Diagnose and solve problems, Access Control (IAM), Data migration, Events, and Storage browser. The 'Containers' option is selected. The main area shows a list of containers with columns for Name, Last modified, Anonymous access level, and Lease state. Three containers are listed: '\$logs', 'inputadf', and 'outputadf'. Each container has a checkbox next to its name.

Name	Last modified	Anonymous access level	Lease state
\$logs	9/5/2023, 2:42:03 PM	Private	Available
inputadf	9/5/2023, 2:46:20 PM	Private	Available
outputadf	9/5/2023, 2:46:31 PM	Private	Available

The screenshot shows the Microsoft Azure Storage Explorer interface. The top navigation bar includes the Microsoft Azure logo, a search bar, and user information. Below the navigation bar, the path 'Home > asdstorage | Containers > inputadf' is displayed. The main area shows the 'inputadf' container details, including an 'Overview' section, 'Diagnose and solve problems', and 'Access Control (IAM)'. On the left, a 'Settings' sidebar lists options like Shared access tokens, Manage ACL, Access policy, Properties, and Metadata. The central table lists blobs with columns for Name, Modified, Access tier, Archive status, and Blob type. One blob, 'emp.txt', is listed with its details.

Basic sense is through adf we need to transfer input txt file to output container.

Now go to adf and under manage find linked services and create a new service under gen 2 data lake.

The screenshot shows the Microsoft Azure Data Factory 'Linked services' creation page. The left sidebar lists various factory settings like General, Connections, and Source control. The main area is titled 'New linked service' and is configured for 'Azure Data Lake Storage Gen2'. The 'Name' field is set to 'AzureDataLakeStorage1'. The 'Description' field is empty. Under 'Connect via integration runtime', 'AutoResolveIntegrationRuntime' is selected. Under 'Authentication type', 'Account key' is selected. Under 'Account selection method', 'From Azure subscription' is selected. Under 'Azure subscription', 'Select all' is chosen. Under 'Storage account name', the dropdown is empty. At the bottom, there are 'Create', 'Back', 'Test connection', and 'Cancel' buttons.

8.1-pipeline

8.2-activity you want to execute

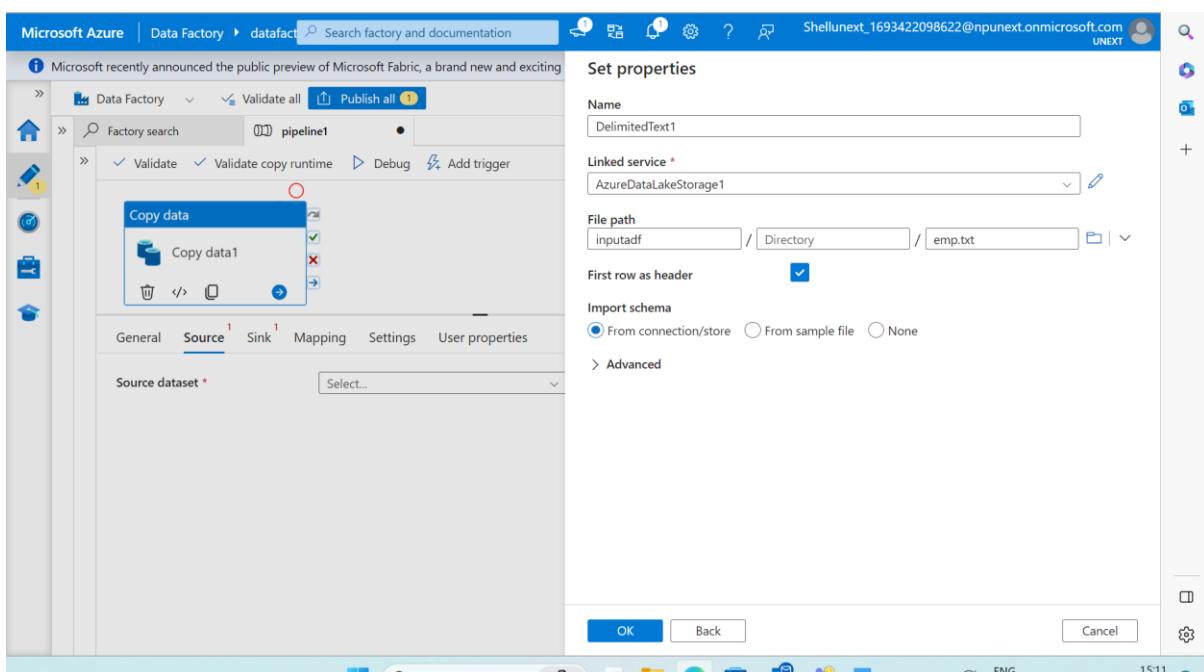
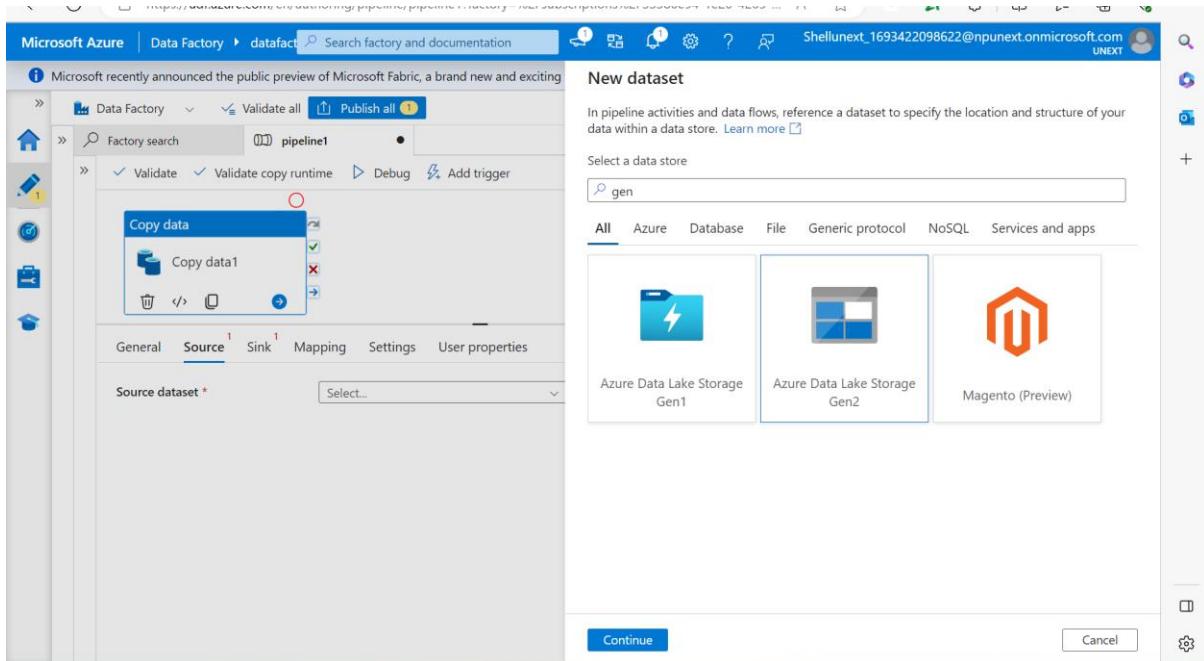
8.3-source

8.4-sink

Under move and transform drag and drop copy data. Now 8.1 and 8.2 is done.

For 8.3 and 8.4 click on copy data and then under source click to make new dataset under gen2.

Under it select csv file type and from it select linked service azure file and under file path browsing select input file emp.txt and click okay.



Now for 8.4 under sink do the same for output file as an option.

Now in open file deselect header checkbox.

Then validate all and then publish all

After that in pipeline add trigger button click on trigger now

Then after okay can check the input file in output storage container.

Microsoft Azure | Data Factory > datafactory Search factory and documentation

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

Preview experience: Off

Properties

- General Related
- Name * pipeline1
- Description
- Annotations + New

Copy data

Copy data1

Validate all Publish all

Validate Validate copy runtime Debug Add trigger

General Source Sink Mapping Settings User properties

Source dataset * DelimitedText1

File path type File path in dataset Wildcard file path List of files

Filter by last modified Start time (UTC) End time (UTC)

Recursively

Enable partition discovery

outputadf Container

Search

Upload Add Directory Refresh Rename Delete Change tier Acquire lease Break lease Give feedback

Overview Diagnose and solve problems Access Control (IAM)

Authentication method: Access key (Switch to Azure AD User Account)
Location: outputadf

Search blobs by prefix (case-sensitive)

Show deleted objects

Name	Modified	Access tier	Archive status	Blob type
emp.txt	9/5/2023, 3:20:01 PM	Hot (Inferred)		Block blob

DAY-6TH September 2023

ADF-

Topic- DataOps

Azure has multiple formats-UI([need portal.azure.com](https://portal.azure.com)), CLI([need Azure CLI](https://learn.microsoft.com/en-us/cli/) can be used in -Cloud Shell, VM, Laptop)

Plan to use today cloud shell- power shell, bash. Cloud shell is serverless.

To make ADF we know manual process now we will learn automated process. For this we can use GIT version.

Automated Enterprise Level Cloud Data Integration Solution- CLI+ GIT

To upload –

- Make a file
- Login in azure account
- Create RG- SA- Container-folder-file upload

Cloud Shell needs –

RG-SA-Cloud shell.

Command imp- Az login-> URL->Code

Commands-

- az login
- CloudVendorCode(az) \

ServiceName(ResGrp) \

FeatureName[] \

OperationName(Create/Delete/Upload) \

Properties(--Key Value) \

Properties(--Key Value) \

Properties(--Key Value) \

Properties(--Key Value)
- az group create \

--name asdrg \

--location eastus
- az \
->used to write in multiple lines for eg->

az\

group\

create
- copy shrtct-> Ctrl+C -> Ctrl+Ins
- paste shrtct-> Ctrl+V -> Shift+Ins
- create storage account->

az storage account create \

--resource group asddf06 \

--name saasdf06 \

--location eastus
- create SA container->

az storage container create \

--resource-group asddf06 \

--account-name saasdf06 \

--name containeras06 \

--auth-mode key
- create file->

cat > emp.txt

welcome to employee file

then press ctrl+d
- Upload container in GITHub->

Az storage blob upload \

```
--account-name saasdf06 \
--container-name containeras06 \
--name input/emp.txt \ (source ya destination address mein se ek hai yeh)
--file emp.txt \
--auth-mode key
```

Now in CLI only we will create ADF

RG->ADF-> Linked service-> datasets->folders->file(upload)

- #Create ADF Workspace->
`az datafactory create \
--resource-group asdf06 \
--name adf06`
- Create connection string in photo whatsapp
- Create linked service from JSON file->
`Az datafactory linked-service create \
--resource-group asddf06 \
--factory-name asdf06 \
--linked-service-name LSStorageAccountGen2 \
--properties @AzureStorageLinkedService.json`

The screenshot shows the Microsoft Azure portal interface. At the top, there's a navigation bar with back, forward, and search icons. Below it is the Azure search bar. The main content area displays a table of resources:

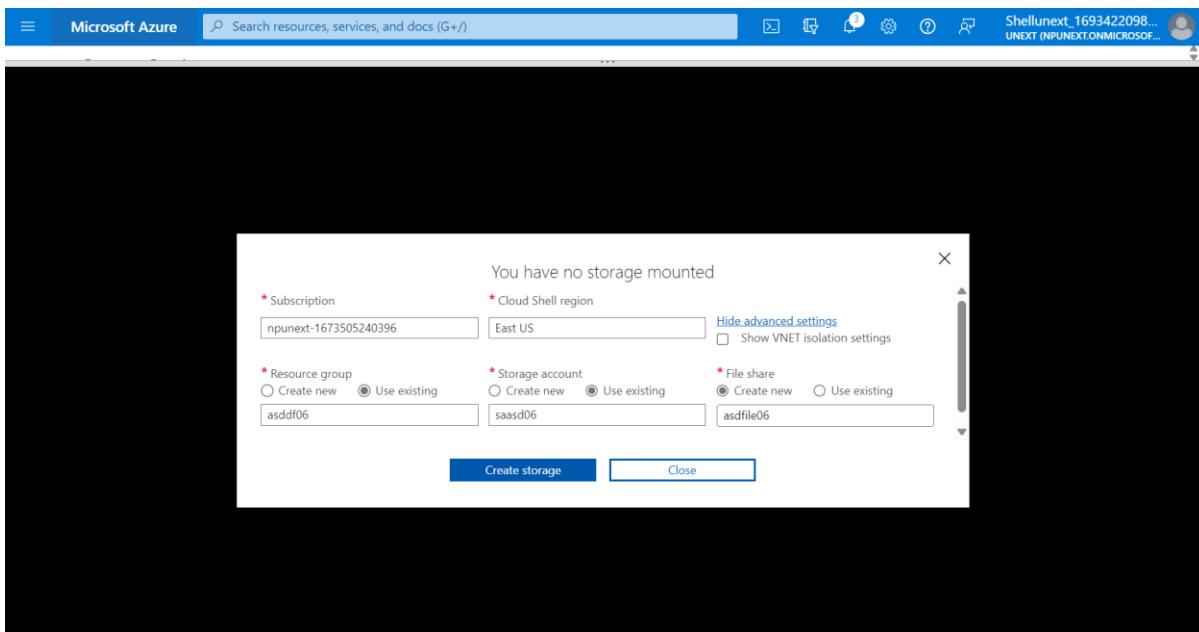
Name	Type	Last Viewed
saasdf06	Storage account	2 minutes ago
asddf06	Resource group	2 minutes ago

Below the table, there's a link to "See all".

Overlaid on the page is a modal dialog box for creating storage. The dialog has the following fields:

- Subscription:** npunext-1673505240396
- Cloud Shell region:** East US
- Storage account:** saasdf06
- File share:** saasdf06
- Resource group:** asddf06 (radio button selected)
- Advanced settings:** Hide advanced settings (checkbox)
- VNET isolation settings:** Show VNET isolation settings (checkbox)
- Buttons:** Create storage (disabled), Close

The dialog also contains a note: "You have no storage mounted".



Git hub

New repository- give name-> Shell-UNext-Single-> create new click

After that settings-> general-> dange->disable branch protection rules

Now go to code and copy url

Go to adf under git configuration->repository owner mein write github name

then create a new branch main under it create a new file

now in setting up git congifuration under man branch select the branches / file...not sure what happened as samne wale ne kiya.

Then click on save and the next screen shot happened.

Configure a repository

Connect your workspace with your Git repository just via [CI/CD best practices](#)

Repository type GitHub
GitHub account irisblyton
Repository name Shell-UNext-Single
Collaboration branch main
Publish branch adf_publish
Root folder /
Last published commit 4f0e9c3c9bc593c210754
Publish (from ADF Studio) Enabled
Custom comment Disabled

Save

Activities

- filter
- Iteration & conditionals
- Filter

PipeArray

Properties

General

Name * PipeArray
Description
Activity state (preview) Active (radio button selected) Inactive

Now after making a new pipeline-

- Pipeline->variable section->name-data array, type-array, default-[“aaa”, “bbb”, “ccc”, “ddd”]
- Picture clicked on mobile whatsapp

Microsoft Azure | Data Factory > asddata Search factory and documentation

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

main branch Validate all Save all Publish Preview experience Off

PipeArray DelimitedText1 Activities

Move and transform Synapse Azure Data Explorer Azure Function Batch Service Databricks Data Lake Analytics General HDInsight Iteration & conditionals Machine Learning Power Query

Filter Act_filter

Saved Save as template Validate Debug Add trigger

Parameters Variables Settings Output

Pipeline run ID: 920cbdda-a289-4487-b204-9a9da0c7bd27 Pipeline status Failed View debug run consumption Monitor in Azure Metrics Export to CSV

All status Showing 1 - 1 of 1 items

Activity name	Activity status	Activity type	Run start	Duration	Log
Act_filter	Failed	Filter	9/6/2023, 2:24:19 PM	Less than 1s	

irisblyton / Shell-UNext-Single

Type / to search

Code Issues Pull requests Actions Projects Security Insights Settings

Shell-UNext-Single Private

adf_publish had recent pushes less than a minute ago

Compare & pull request

main 1 branch 0 tags Go to file Add file Code

irisblyton Create publish_config.json 7c13261 1 minute ago 3 commits

dataset	Adding linkedService: AzureDataLakeStorage1	1 minute ago
factory	Adding linkedService: AzureDataLakeStorage1	1 minute ago
linkedService	Adding linkedService: AzureDataLakeStorage1	1 minute ago
index	Create index	2 minutes ago
publish_config.json	Create publish_config.json	1 minute ago

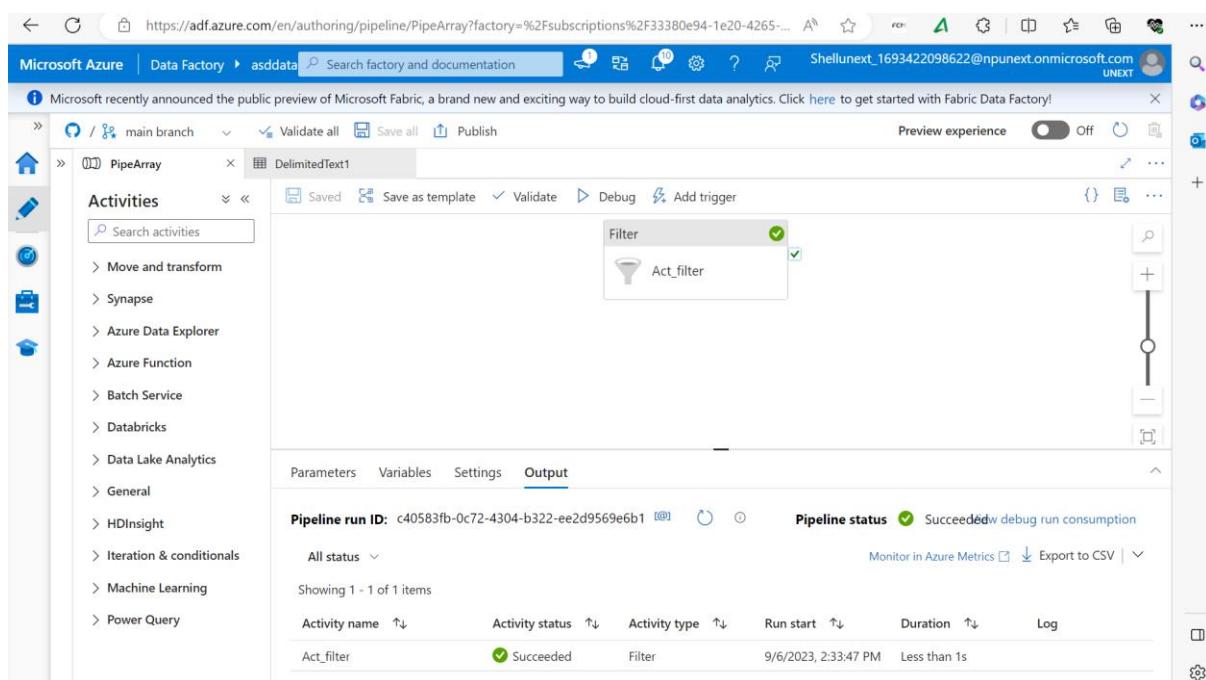
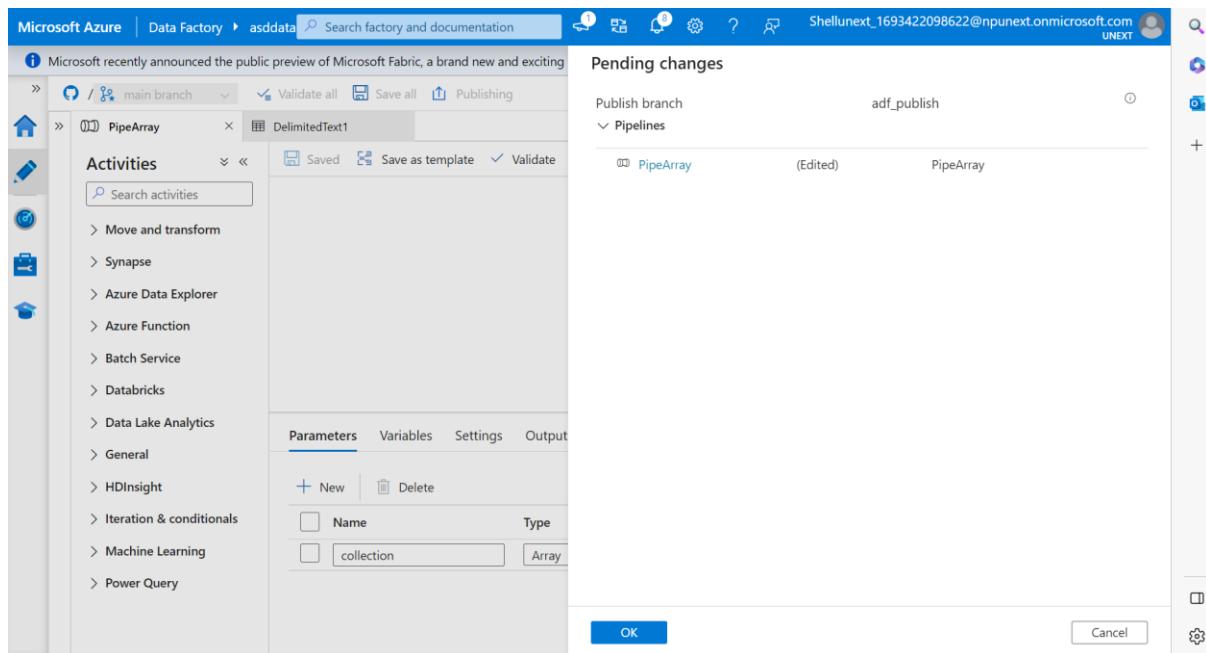
About No description, website, or topics provided.

Activity 0 stars 1 watching 0 forks

Releases No releases published Create a new release

Packages No packages published Publish your first package

Add a README with an overview of your project. Add a README



https://github.com/irisblyton/Shell-UNext-Single/tree/main

main 2 branches 0 tags Go to file Add file Code

Your main branch isn't protected
Protect this branch

Activity 0 stars 1 watching 0 forks

Releases No releases published Create a new release

Packages No packages published Publish your first package

irisblyton Updating pipeline: PipeArray e28449c 3 minutes ago 5 commits

dataset	Adding linkedService: AzureDataLakeStorage1	2 hours ago
factory	Adding linkedService: AzureDataLakeStorage1	2 hours ago
linkedService	Adding linkedService: AzureDataLakeStorage1	2 hours ago
pipeline	Updating pipeline: PipeArray	3 minutes ago
index	Create index	2 hours ago
publish_config.json	Create publish_config.json	2 hours ago

Add a README with an overview of your project. Add a README



Showing 1 changed file with 19 additions and 3 deletions.

Unified

pipeline/PipeArray.json

```
@@ -9,7 +9,7 @@
 9   9           "userProperties": [],
10  10          "typeProperties": {
11  11            "items": [
12  12              {
13  13                "value": "@variables('dataArray')",
14  14                "value": "@pipeline().parameters.collection",
15  15                "type": "Expression"
16  16              },
17  17              "condition": [
18  18                {
19  19                  }
20  20                }
21  21              ],
22  22              "parameters": {
23  23                "collection": {
24  24                  "type": "array",
25  25                  "defaultValue": [
26  26                    "aaa",
27  27                    "bbb",
28  28                    "ccc",
29  29                    "ddd"
30  30                  ]
31  31                }
32  32              },
33  33              "variables": {
34  34                "dataArray": {
35  35                  "type": "String",
36  36                  "value": "test"
37  37                }
38  38              }
39  39            ]
40  40          }
41  41        ]
42  42      ]
43  43    ],
44  44    "type": "Object"
45  45  }
46  46  }
47  47  }
48  48  }
49  49  }
50  50  }
51  51  }
52  52  }
53  53  }
54  54  }
55  55  }
56  56  }
57  57  }
58  58  }
59  59  }
60  60  }
61  61  }
62  62  }
63  63  }
64  64  }
65  65  }
66  66  }
67  67  }
68  68  }
69  69  }
70  70  }
71  71  }
72  72  }
73  73  }
74  74  }
75  75  }
76  76  }
77  77  }
78  78  }
79  79  }
80  80  }
81  81  }
82  82  }
83  83  }
84  84  }
85  85  }
86  86  }
87  87  }
88  88  }
89  89  }
90  90  }
91  91  }
92  92  }
93  93  }
94  94  }
95  95  }
96  96  }
97  97  }
98  98  }
99  99  }
100 100  }
101 101  }
102 102  }
103 103  }
104 104  }
105 105  }
106 106  }
107 107  }
108 108  }
109 109  }
110 110  }
111 111  }
112 112  }
113 113  }
114 114  }
115 115  }
116 116  }
117 117  }
118 118  }
119 119  }
120 120  }
121 121  }
122 122  }
123 123  }
124 124  }
125 125  }
126 126  }
127 127  }
128 128  }
129 129  }
130 130  }
131 131  }
132 132  }
133 133  }
134 134  }
135 135  }
136 136  }
137 137  }
138 138  }
139 139  }
140 140  }
141 141  }
142 142  }
143 143  }
144 144  }
145 145  }
146 146  }
147 147  }
148 148  }
149 149  }
150 150  }
151 151  }
152 152  }
153 153  }
154 154  }
155 155  }
156 156  }
157 157  }
158 158  }
159 159  }
160 160  }
161 161  }
162 162  }
163 163  }
164 164  }
165 165  }
166 166  }
167 167  }
168 168  }
169 169  }
170 170  }
171 171  }
172 172  }
173 173  }
174 174  }
175 175  }
176 176  }
177 177  }
178 178  }
179 179  }
180 180  }
181 181  }
182 182  }
183 183  }
184 184  }
185 185  }
186 186  }
187 187  }
188 188  }
189 189  }
190 190  }
191 191  }
192 192  }
193 193  }
194 194  }
195 195  }
196 196  }
197 197  }
198 198  }
199 199  }
200 200  }
201 201  }
202 202  }
203 203  }
204 204  }
205 205  }
206 206  }
207 207  }
208 208  }
209 209  }
210 210  }
211 211  }
212 212  }
213 213  }
214 214  }
215 215  }
216 216  }
217 217  }
218 218  }
219 219  }
220 220  }
221 221  }
222 222  }
223 223  }
224 224  }
225 225  }
226 226  }
227 227  }
228 228  }
229 229  }
230 230  }
231 231  }
232 232  }
233 233  }
234 234  }
235 235  }
236 236  }
237 237  }
238 238  }
239 239  }
240 240  }
241 241  }
242 242  }
243 243  }
244 244  }
245 245  }
246 246  }
247 247  }
248 248  }
249 249  }
250 250  }
251 251  }
252 252  }
253 253  }
254 254  }
255 255  }
256 256  }
257 257  }
258 258  }
259 259  }
260 260  }
261 261  }
262 262  }
263 263  }
264 264  }
265 265  }
266 266  }
267 267  }
268 268  }
269 269  }
270 270  }
271 271  }
272 272  }
273 273  }
274 274  }
275 275  }
276 276  }
277 277  }
278 278  }
279 279  }
280 280  }
281 281  }
282 282  }
283 283  }
284 284  }
285 285  }
286 286  }
287 287  }
288 288  }
289 289  }
290 290  }
291 291  }
292 292  }
293 293  }
294 294  }
295 295  }
296 296  }
297 297  }
298 298  }
299 299  }
300 300  }
301 301  }
302 302  }
303 303  }
304 304  }
305 305  }
306 306  }
307 307  }
308 308  }
309 309  }
310 310  }
311 311  }
312 312  }
313 313  }
314 314  }
315 315  }
316 316  }
317 317  }
318 318  }
319 319  }
320 320  }
321 321  }
322 322  }
323 323  }
324 324  }
325 325  }
326 326  }
327 327  }
328 328  }
329 329  }
330 330  }
331 331  }
332 332  }
333 333  }
334 334  }
335 335  }
336 336  }
337 337  }
338 338  }
339 339  }
340 340  }
341 341  }
342 342  }
343 343  }
344 344  }
345 345  }
346 346  }
347 347  }
348 348  }
349 349  }
350 350  }
351 351  }
352 352  }
353 353  }
354 354  }
355 355  }
356 356  }
357 357  }
358 358  }
359 359  }
360 360  }
361 361  }
362 362  }
363 363  }
364 364  }
365 365  }
366 366  }
367 367  }
368 368  }
369 369  }
370 370  }
371 371  }
372 372  }
373 373  }
374 374  }
375 375  }
376 376  }
377 377  }
378 378  }
379 379  }
380 380  }
381 381  }
382 382  }
383 383  }
384 384  }
385 385  }
386 386  }
387 387  }
388 388  }
389 389  }
390 390  }
391 391  }
392 392  }
393 393  }
394 394  }
395 395  }
396 396  }
397 397  }
398 398  }
399 399  }
400 400  }
401 401  }
402 402  }
403 403  }
404 404  }
405 405  }
406 406  }
407 407  }
408 408  }
409 409  }
410 410  }
411 411  }
412 412  }
413 413  }
414 414  }
415 415  }
416 416  }
417 417  }
418 418  }
419 419  }
420 420  }
421 421  }
422 422  }
423 423  }
424 424  }
425 425  }
426 426  }
427 427  }
428 428  }
429 429  }
430 430  }
431 431  }
432 432  }
433 433  }
434 434  }
435 435  }
436 436  }
437 437  }
438 438  }
439 439  }
440 440  }
441 441  }
442 442  }
443 443  }
444 444  }
445 445  }
446 446  }
447 447  }
448 448  }
449 449  }
450 450  }
451 451  }
452 452  }
453 453  }
454 454  }
455 455  }
456 456  }
457 457  }
458 458  }
459 459  }
460 460  }
461 461  }
462 462  }
463 463  }
464 464  }
465 465  }
466 466  }
467 467  }
468 468  }
469 469  }
470 470  }
471 471  }
472 472  }
473 473  }
474 474  }
475 475  }
476 476  }
477 477  }
478 478  }
479 479  }
480 480  }
481 481  }
482 482  }
483 483  }
484 484  }
485 485  }
486 486  }
487 487  }
488 488  }
489 489  }
490 490  }
491 491  }
492 492  }
493 493  }
494 494  }
495 495  }
496 496  }
497 497  }
498 498  }
499 499  }
500 500  }
501 501  }
502 502  }
503 503  }
504 504  }
505 505  }
506 506  }
507 507  }
508 508  }
509 509  }
510 510  }
511 511  }
512 512  }
513 513  }
514 514  }
515 515  }
516 516  }
517 517  }
518 518  }
519 519  }
520 520  }
521 521  }
522 522  }
523 523  }
524 524  }
525 525  }
526 526  }
527 527  }
528 528  }
529 529  }
530 530  }
531 531  }
532 532  }
533 533  }
534 534  }
535 535  }
536 536  }
537 537  }
538 538  }
539 539  }
540 540  }
541 541  }
542 542  }
543 543  }
544 544  }
545 545  }
546 546  }
547 547  }
548 548  }
549 549  }
550 550  }
551 551  }
552 552  }
553 553  }
554 554  }
555 555  }
556 556  }
557 557  }
558 558  }
559 559  }
560 560  }
561 561  }
562 562  }
563 563  }
564 564  }
565 565  }
566 566  }
567 567  }
568 568  }
569 569  }
570 570  }
571 571  }
572 572  }
573 573  }
574 574  }
575 575  }
576 576  }
577 577  }
578 578  }
579 579  }
580 580  }
581 581  }
582 582  }
583 583  }
584 584  }
585 585  }
586 586  }
587 587  }
588 588  }
589 589  }
590 590  }
591 591  }
592 592  }
593 593  }
594 594  }
595 595  }
596 596  }
597 597  }
598 598  }
599 599  }
600 600  }
601 601  }
602 602  }
603 603  }
604 604  }
605 605  }
606 606  }
607 607  }
608 608  }
609 609  }
610 610  }
611 611  }
612 612  }
613 613  }
614 614  }
615 615  }
616 616  }
617 617  }
618 618  }
619 619  }
620 620  }
621 621  }
622 622  }
623 623  }
624 624  }
625 625  }
626 626  }
627 627  }
628 628  }
629 629  }
630 630  }
631 631  }
632 632  }
633 633  }
634 634  }
635 635  }
636 636  }
637 637  }
638 638  }
639 639  }
640 640  }
641 641  }
642 642  }
643 643  }
644 644  }
645 645  }
646 646  }
647 647  }
648 648  }
649 649  }
650 650  }
651 651  }
652 652  }
653 653  }
654 654  }
655 655  }
656 656  }
657 657  }
658 658  }
659 659  }
660 660  }
661 661  }
662 662  }
663 663  }
664 664  }
665 665  }
666 666  }
667 667  }
668 668  }
669 669  }
670 670  }
671 671  }
672 672  }
673 673  }
674 674  }
675 675  }
676 676  }
677 677  }
678 678  }
679 679  }
680 680  }
681 681  }
682 682  }
683 683  }
684 684  }
685 685  }
686 686  }
687 687  }
688 688  }
689 689  }
690 690  }
691 691  }
692 692  }
693 693  }
694 694  }
695 695  }
696 696  }
697 697  }
698 698  }
699 699  }
700 700  }
701 701  }
702 702  }
703 703  }
704 704  }
705 705  }
706 706  }
707 707  }
708 708  }
709 709  }
710 710  }
711 711  }
712 712  }
713 713  }
714 714  }
715 715  }
716 716  }
717 717  }
718 718  }
719 719  }
720 720  }
721 721  }
722 722  }
723 723  }
724 724  }
725 725  }
726 726  }
727 727  }
728 728  }
729 729  }
730 730  }
731 731  }
732 732  }
733 733  }
734 734  }
735 735  }
736 736  }
737 737  }
738 738  }
739 739  }
740 740  }
741 741  }
742 742  }
743 743  }
744 744  }
745 745  }
746 746  }
747 747  }
748 748  }
749 749  }
750 750  }
751 751  }
752 752  }
753 753  }
754 754  }
755 755  }
756 756  }
757 757  }
758 758  }
759 759  }
760 760  }
761 761  }
762 762  }
763 763  }
764 764  }
765 765  }
766 766  }
767 767  }
768 768  }
769 769  }
770 770  }
771 771  }
772 772  }
773 773  }
774 774  }
775 775  }
776 776  }
777 777  }
778 778  }
779 779  }
780 780  }
781 781  }
782 782  }
783 783  }
784 784  }
785 785  }
786 786  }
787 787  }
788 788  }
789 789  }
790 790  }
791 791  }
792 792  }
793 793  }
794 794  }
795 795  }
796 796  }
797 797  }
798 798  }
799 799  }
800 800  }
801 801  }
802 802  }
803 803  }
804 804  }
805 805  }
806 806  }
807 807  }
808 808  }
809 809  }
810 810  }
811 811  }
812 812  }
813 813  }
814 814  }
815 815  }
816 816  }
817 817  }
818 818  }
819 819  }
820 820  }
821 821  }
822 822  }
823 823  }
824 824  }
825 825  }
826 826  }
827 827  }
828 828  }
829 829  }
830 830  }
831 831  }
832 832  }
833 833  }
834 834  }
835 835  }
836 836  }
837 837  }
838 838  }
839 839  }
840 840  }
841 841  }
842 842  }
843 843  }
844 844  }
845 845  }
846 846  }
847 847  }
848 848  }
849 849  }
850 850  }
851 851  }
852 852  }
853 853  }
854 854  }
855 855  }
856 856  }
857 857  }
858 858  }
859 859  }
860 860  }
861 861  }
862 862  }
863 863  }
864 864  }
865 865  }
866 866  }
867 867  }
868 868  }
869 869  }
870 870  }
871 871  }
872 872  }
873 873  }
874 874  }
875 875  }
876 876  }
877 877  }
878 878  }
879 879  }
880 880  }
881 881  }
882 882  }
883 883  }
884 884  }
885 885  }
886 886  }
887 887  }
888 888  }
889 889  }
890 890  }
891 891  }
892 892  }
893 893  }
894 894  }
895 895  }
896 896  }
897 897  }
898 898  }
899 899  }
900 900  }
901 901  }
902 902  }
903 903  }
904 904  }
905 905  }
906 906  }
907 907  }
908 908  }
909 909  }
910 910  }
911 911  }
912 912  }
913 913  }
914 914  }
915 915  }
916 916  }
917 917  }
918 918  }
919 919  }
920 920  }
921 921  }
922 922  }
923 923  }
924 924  }
925 925  }
926 926  }
927 927  }
928 928  }
929 929  }
930 930  }
931 931  }
932 932  }
933 933  }
934 934  }
935 935  }
936 936  }
937 937  }
938 938  }
939 939  }
940 940  }
941 941  }
942 942  }
943 943  }
944 944  }
945 945  }
946 946  }
947 947  }
948 948  }
949 949  }
950 950  }
951 951  }
952 952  }
953 953  }
954 954  }
955 955  }
956 956  }
957 957  }
958 958  }
959 959  }
960 960  }
961 961  }
962 962  }
963 963  }
964 964  }
965 965  }
966 966  }
967 967  }
968 968  }
969 969  }
970 970  }
971 971  }
972 972  }
973 973  }
974 974  }
975 975  }
976 976  }
977 977  }
978 978  }
979 979  }
980 980  }
981 981  }
982 982  }
983 983  }
984 984  }
985 985  }
986 986  }
987 987  }
988 988  }
989 989  }
990 990  }
991 991  }
992 992  }
993 993  }
994 994  }
995 995  }
996 996  }
997 997  }
998 998  }
999 999  }
1000 1000  }
```

Microsoft Azure | Data Factory > asddata Search factory and documentation

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

main branch Validate all Save all Publish Preview experience Off

PipeArray DelimitedText1

Activities

Save as template Validate Debug Add trigger

Save the current resource configuration as a template for sharing or future use

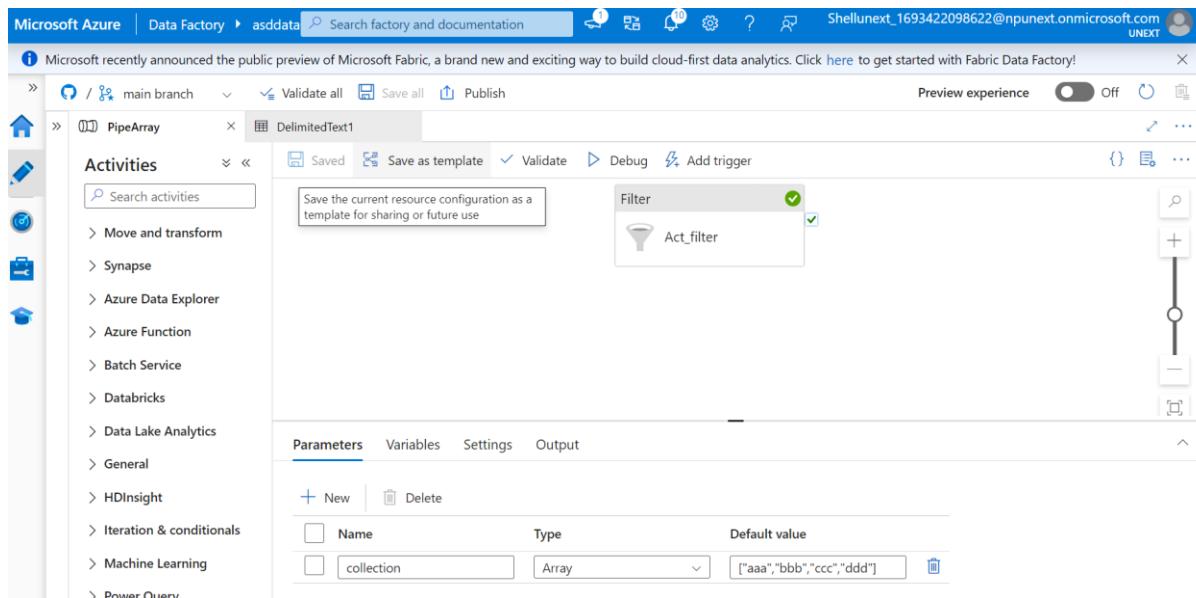
Filter Act_filter

Parameters Variables Settings Output

New Delete

Name Type Default value

collection Array ["aaa","bbb","ccc","ddd"]



Microsoft Azure | Data Factory > asddata Search factory and documentation

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

main branch Validate all Save all Publish Preview experience Off

PipeArray DelimitedText1

Activities

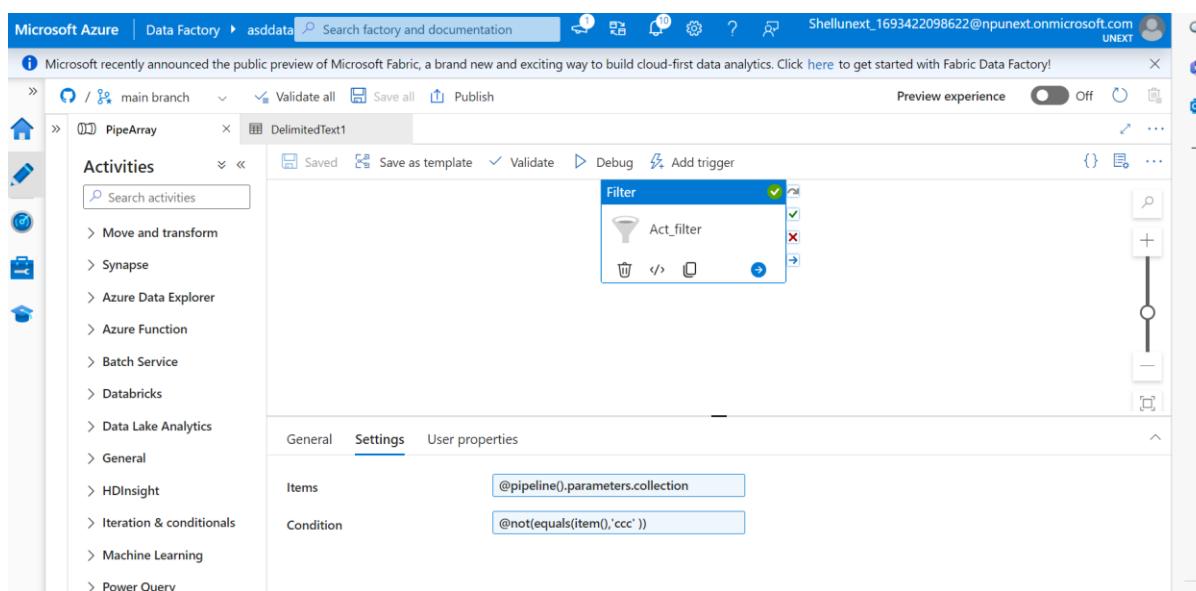
Save as template Validate Debug Add trigger

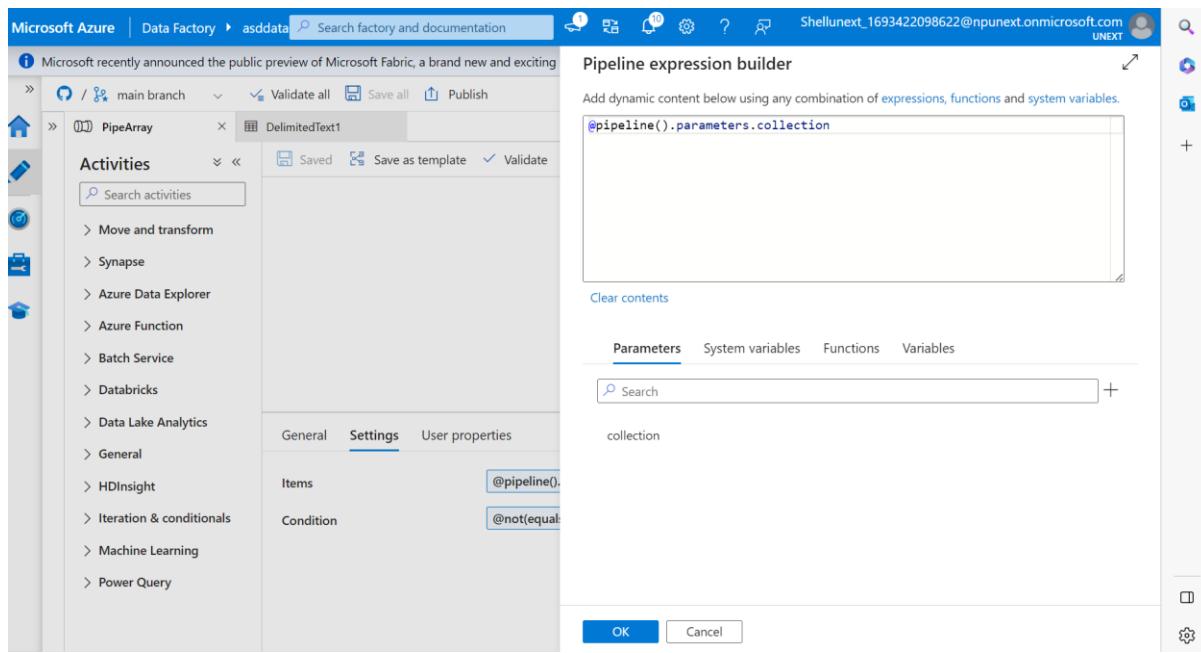
Filter Act_filter

General Settings User properties

Items @pipeline().parameters.collection

Condition @not>equals(item(),'ccc'))





Microsoft Azure | Data Factory > asodata007

Copy Data tool

Properties

Source

Destination

Settings

Review and finish

Use Copy Data Tool to perform a one-time or scheduled data load from 90+ data sources. Follow the wizard experience to specify your data loading settings, and let the Copy Data Tool generate the artifacts for you, including pipelines, datasets, and linked services. [Learn more](#)

Properties

Select copy data task type and configure task schedule

Task type

Built-in copy task
You will get single pipeline to copy data from 90+ data source easily.

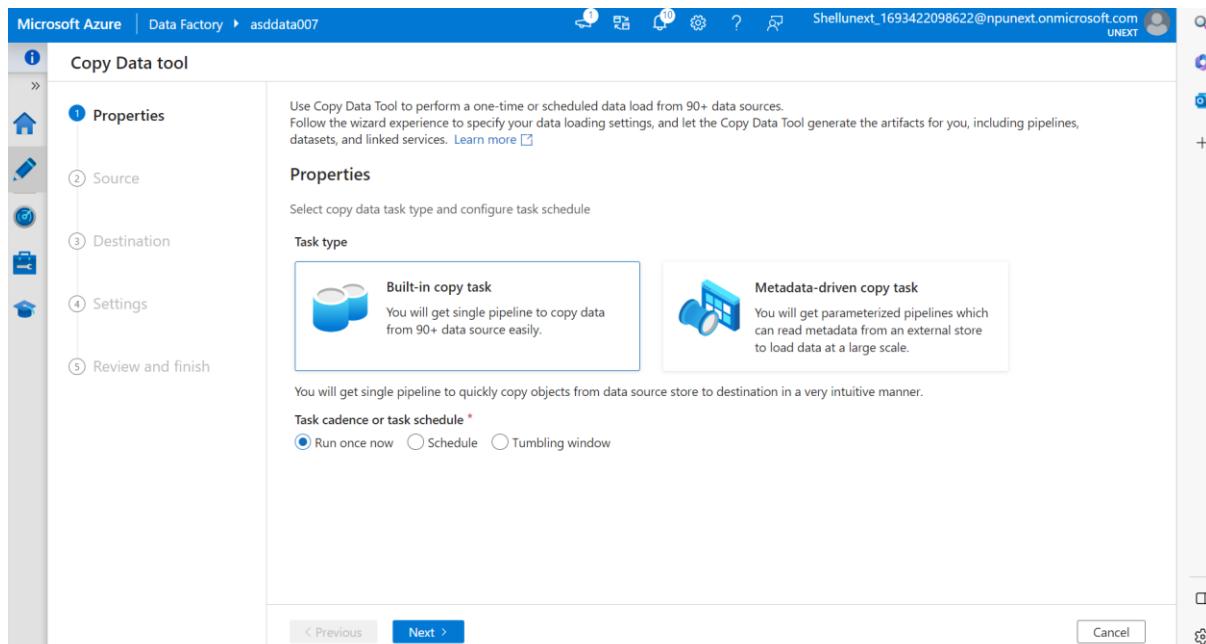
Metadata-driven copy task
You will get parameterized pipelines which can read metadata from an external store to load data at a large scale.

You will get single pipeline to quickly copy objects from data source store to destination in a very intuitive manner.

Task cadence or task schedule *

Run once now Schedule Tumbling window

< Previous Next > Cancel



Microsoft Azure | Data Factory > asodata007

Copy Data tool

Properties

Source

Dataset

Configuration

Destination

Settings

Review and finish

Source data store

Specify the source data store for the copy task. You can use an existing data store connection or specify a new data store.

Source type: Azure Data Lake Storage Gen2

Connection *: AzureDataLakeStorage1

If the identity you use to access the data store only has permission to subdirectory instead of the entire account, specify the path to browse.

container007/emp.txt

Options

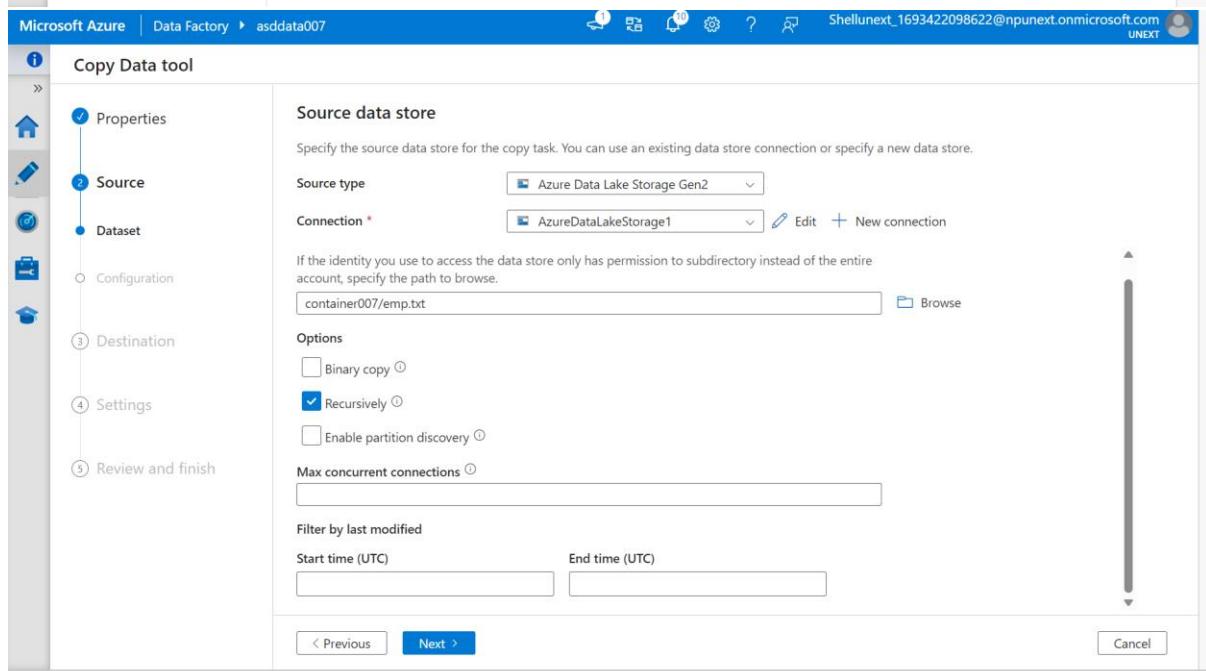
Binary copy (1)
 Recursively (1)
 Enable partition discovery (1)

Max concurrent connections:

Filter by last modified

Start time (UTC) End time (UTC)

< Previous Next > Cancel



Click on next then again for file format setting click on next or click on

Microsoft Azure | Data Factory > asdata007

Copy Data tool

Properties

Source

Destination

Settings

Review and finish

Review

Deployment

Summary

You are running pipeline to copy data from Azure Data Lake Storage Gen2 to Azure Data Lake Storage Gen2.

Azure Data Lake Storage Gen2 → Azure Data Lake Storage Gen2

Properties

Task name: CopyPipeline_t9n

Task description:

Source

Connection name: AzureDataLakeStorage1

Dataset name: SourceDataset_t9n

Column delimiter:

Escape character: \

Quote char: "

Edit

Next >

Cancel

Copy

Microsoft Azure | Data Factory > asdata007

Copy Data tool

Properties

Source

Destination

Settings

Review and finish

Review

Deployment

Azure Data Lake Storage Gen2 → Azure Data Lake Storage Gen2

Deployment complete

Deployment step Status

> Creating datasets Succeeded

> Creating pipelines Succeeded

Finish Edit pipeline Monitor

DAY- 7th September 2023

Microsoft Azure | Data Factory > datafactory Search factory and documentation

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

Expand Data Factory Validate all Publish all

Preview experience Off

Properties

General Related (1)

Name * dataflow1

Description

Source settings Source options Projection Optimize Inspect Data preview

Output stream name * source1 Learn more

Description Import data from inputdata Reset

Source type * Dataset Inline

Connection successful

11:14 07-09-2023

Source settings Source options Projection Optimize Inspect Data preview

Number of rows + INSERT 77 * UPDATE 0 × DELETE 0 + UPSERT 0

Film	Genre	Lead Stu...	Audienc...	Profitabil...	Rotten...
Zack and...	Romance	The Wein...	70	1.747541...	64
Youth in ...	Comedy	The Wein...	52	1.09	68
You Will ...	Comedy	Independ...	35	1.211818...	43
When in ...	Comedy	Disney	44	0	15
What Ha...	Comedy	Fox	72	6.267647...	28
Water Fo...	Drama	20th Cen...	72	3.081421...	60
WALL-E	Animation	Disney	89	2.896019...	96
Waitress	Romance	Independ...	67	11.08974...	89

Microsoft Azure | Data Factory > datafactory Search factory and documentation

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

Expand Data Factory Validate all Publish all

Preview experience Off

Properties

General Related (1)

Name * dataflow1

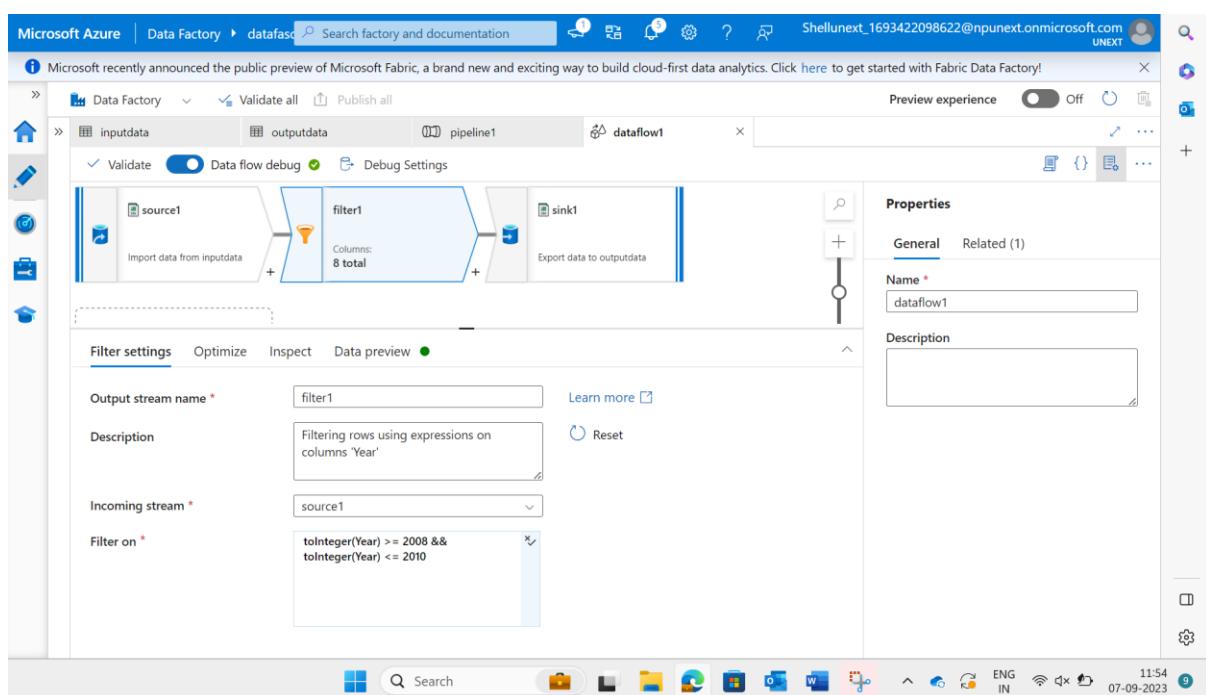
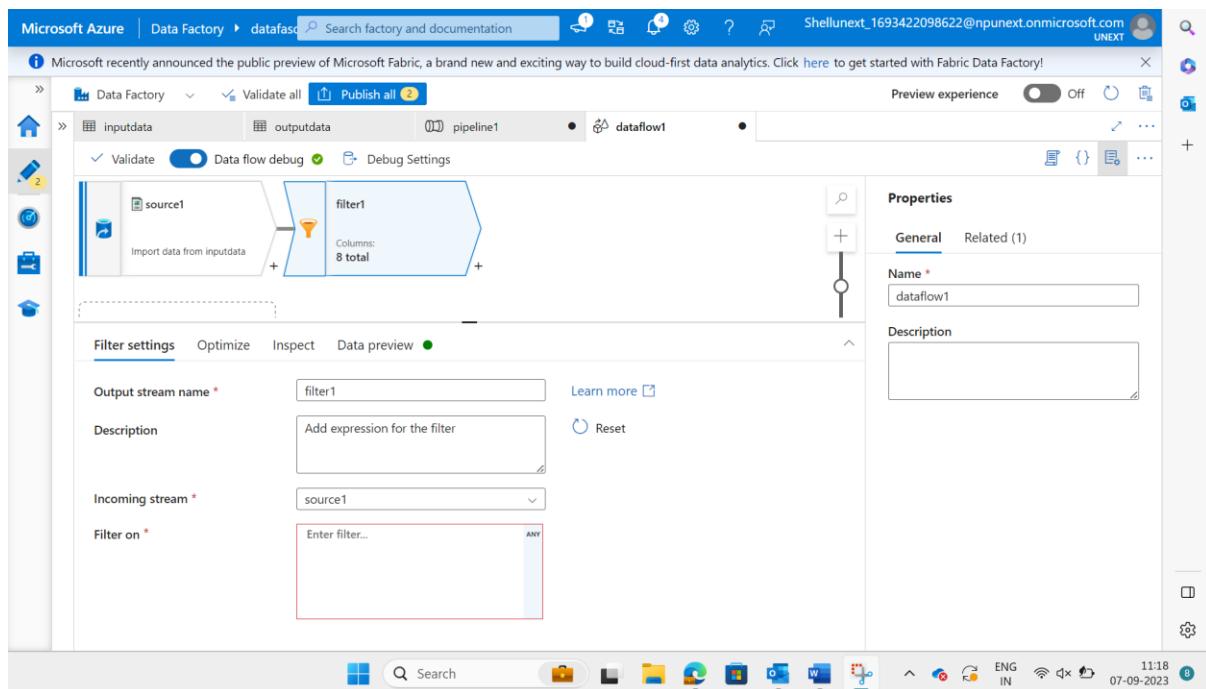
Description

Source settings Source options Projection Optimize Inspect Data preview

Number of rows + INSERT 77 * UPDATE 0 × DELETE 0 + UPSERT 0

Refresh | Typecast | Modify | Map drifted | Statistics | Remove | Export to CSV |

Film	Genre	Lead Stu...	Audienc...	Profitabil...	Rotten...
Zack and...	Romance	The Wein...	70	1.747541...	64
Youth in ...	Comedy	The Wein...	52	1.09	68
You Will ...	Comedy	Independ...	35	1.211818...	43
When in ...	Comedy	Disney	44	0	15
What Ha...	Comedy	Fox	72	6.267647...	28
Water Fo...	Drama	20th Cen...	72	3.081421...	60
WALL-E	Animation	Disney	89	2.896019...	96
Waitress	Romance	Independ...	67	11.08974...	89



Microsoft Azure | Data Factory > dataflow1 Search factory and documentation

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

Data Factory Validate all Publishing

inputdata outputdata pipeline1 dataflow1

Validate Data flow debug Debug Settings

source1 Import data from inputdata filter1 Columns: 8 total sink1 Export data to outputdata

Properties General Related (1)

Name * dataflow1

Description

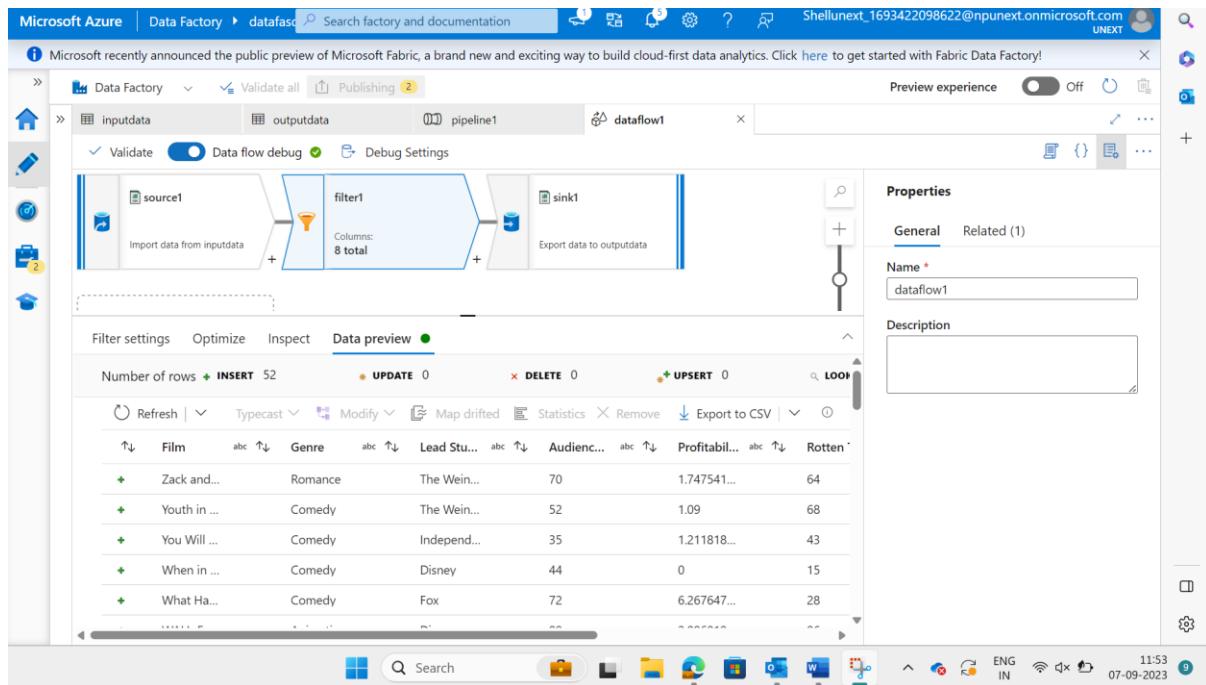
Filter settings Optimize Inspect Data preview

Number of rows + INSERT 52 UPDATE 0 DELETE 0 UPSERT 0 LOOP 0

Refresh Typecast Modify Map drifted Statistics Remove Export to CSV

Film	Genre	Lead Stu...	Audience...	Profitabil...	Rotten T...
Zack and...	Romance	The Wein...	70	1.747541...	64
Youth in ...	Comedy	The Wein...	52	1.09	68
You Will ...	Comedy	Independ...	35	1.211818...	43
When in ...	Comedy	Disney	44	0	15
What Ha...	Comedy	Fox	72	6.267647...	28

11:53 07-09-2023



Microsoft Azure | Data Factory > dataflow1 Search factory and documentation

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

Data Factory Validate all Publish all

inputdata outputdata pipeline1 dataflow1

Validate Data flow debug Debug Settings

source1 Import data from inputdata filter1 Filtering rows using expressions on columns 'Year' sink1 Columns: 8 total

Properties General Related (1)

Name * dataflow1

Description

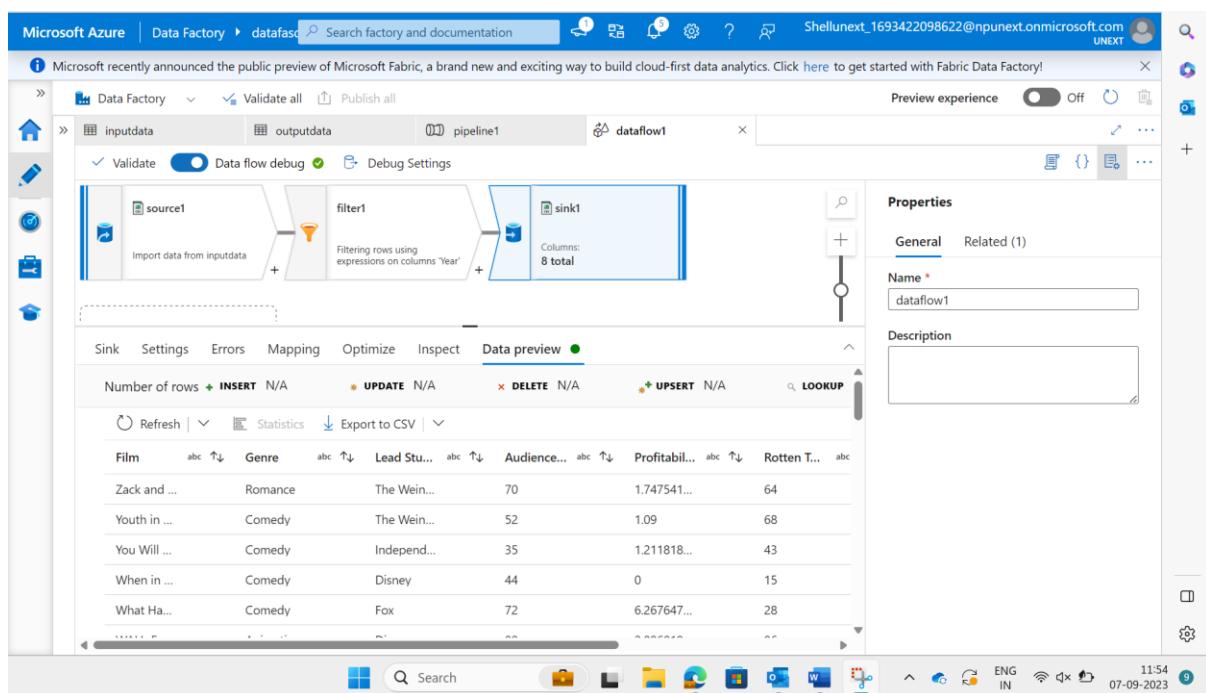
Sink Settings Errors Mapping Optimize Inspect Data preview

Number of rows + INSERT N/A UPDATE N/A DELETE N/A UPSERT N/A LOOKUP

Refresh Statistics Export to CSV

Film	Genre	Lead Stu...	Audience...	Profitabil...	Rotten T...
Zack and ...	Romance	The Wein...	70	1.747541...	64
Youth in ...	Comedy	The Wein...	52	1.09	68
You Will ...	Comedy	Independ...	35	1.211818...	43
When in ...	Comedy	Disney	44	0	15
What Ha...	Comedy	Fox	72	6.267647...	28

11:54 07-09-2023



Microsoft Azure | Data Factory > datafactory Search factory and documentation

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

Data Factory > Validate all Publish all

Factory Resources > Filter resources by name

- Pipelines: pipeline1
- Change Data Capture (preview): 0
- Datasets: outputdata, inputdata
- Data flows: dataflow1
- Power Query: 0

inputdata > pipeline1 > dataflow1

Properties: General Related (1)

Name: dataflow1

Description:

Preview experience: Off

Data flow debug: On

Debug Settings

Number of rows: N/A

INSERT: N/A

UPDATE: N/A

DELETE: N/A

Data preview:

Film	abc	Genre	abc	Lead Studio	abc	Audience...	abc
Zack and ...		Romance		The Wein...		70	
Youth in ...		Comedy		The Wein...		52	
You Will ...		Comedy		Independ...		35	
When in ...		Comedy		Disney		44	
What Ha...		Comedy		Fox		72	

Refresh | Statistics | Export to CSV | ...

11:57 07-09-2023

```
graph LR; source1[source1] --> filter1(filter1); filter1 --> sink1[sink1];
```

The screenshot shows a data flow named 'dataflow1' within a pipeline. It consists of three main components: 'source1', 'filter1', and 'sink1'. The 'source1' component is connected to the 'filter1' component, which is then connected to the 'sink1' component. The 'filter1' component has a note indicating it is filtering rows using expressions on columns 'Year'. The 'sink1' component has a note indicating it is summing up values. A 'Data preview' section shows a table with five rows of movie data, including columns for Film, Genre, Lead Studio, Audience, and abc.

Microsoft Azure | Data Factory > datafactory Search factory and documentation

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

Data Factory > Validate all Publish all

Factory Resources > Filter resources by name

- Pipelines: pipeline1
- Change Data Capture (preview): 0
- Datasets: outputdata, inputdata
- Data flows: dataflow1
- Power Query: 0

inputdata > pipeline1 > dataflow1

Properties: General Related (1)

Name: dataflow1

Description:

Preview experience: Off

Data flow debug: On

Debug Settings

Output stream name: tomatoes

Description: Filtering rows using expressions on columns 'Rotten Tomatoes %'

Learn more

Reset Incoming stream: source1

Filter on: toFloat(Rotten Tomatoes %) >= 75

Filter settings | Optimize | Inspect | Data preview

12:07 07-09-2023

```
graph LR; source1[source1] --> tomatoes[tomatoes]; tomatoes --> select1(select1);
```

The screenshot shows a data flow named 'dataflow1' within a pipeline. It consists of two main components: 'source1' and 'select1'. The 'source1' component is connected to the 'tomatoes' component, which is then connected to the 'select1' component. The 'tomatoes' component has a note indicating it is renaming columns. A 'Filter settings' section shows a 'Filter on' condition: 'toFloat(Rotten Tomatoes %) >= 75'. The 'Description' field for the filter is 'Filtering rows using expressions on columns "Rotten Tomatoes %"'.

Microsoft Azure | Data Factory | datafactory | Search factory and documentation

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

Data Factory Validate all Publish all

Factory Resources Filter resources by name

Pipelines 1 pipeline1

Change Data Capture (pr... 0

Datasets 2 outputdata inputdata

Data flows 1 dataflow1

Power Query 0

inputdata outputdata pipeline1 dataflow1

Validate Data flow debug Debug Settings

source1 Import data from inputdata tomatoes select1 Renaming tomatoes to select1 with columns 'Film', 'Genre', 'Lead Studio', 'Audience score %', 'Profitability', 'Rotten Tomatoes %'

Columns: 8 total

Filter settings Optimize Inspect Data preview

Number of rows: 14 INSERT 14 UPDATE 0 DELETE 0

Refresh Typecast Modify Map drifted Statistics Refresh

	Film	Genre	Lead Studio	Audience...	Profitability	Rotten Tomatoes %
+	WALL-E	Animation	Disney	89		
+	Waitress	Romance	Independ...	67		
+	Tangled	Animation	Disney	88		
+	Rachel G...	Drama	Independ...	61		
+	My Week...	Drama	The Wein...	84		

Properties General Related (1)

Name * dataflow1

Description

12:07 07-09-2023

Microsoft Azure | Data Factory | datafactory | Search factory and documentation

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

Data Factory Validate all Publish all

Factory Resources Filter resources by name

Pipelines 1 pipeline1

Change Data Capture (pr... 0

Datasets 2 outputdata inputdata

Data flows 1 dataflow1

Power Query 0

inputdata outputdata pipeline1 dataflow1

Validate Data flow debug Debug Settings

source1 Import data from inputdata tomatoes select1 Renaming tomatoes to select1 with columns 'Film', 'Genre', 'Lead Studio', 'Audience score %', 'Profitability', 'Rotten Tomatoes %'

Filtering rows using expressions on columns 'Rotten Tomatoes %'

Columns: 8 total

filter1 Filtering rows using expressions on columns 'Year'

Add Source

Select settings Output stream name * select1

Learn more Description Renaming tomatoes to select1 with columns 'Film', 'Genre', 'Lead Studio', 'Audience score %', 'Profitability', 'Rotten Tomatoes %'

Reset Incoming stream * tomatoes

Properties General Related (1)

Name * dataflow1

Description

12:07 07-09-2023

Microsoft Azure | Data Factory > datafactory1 Search factory and documentation

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

Data Factory Validate all Publish all

Factory Resources

- Pipelines 1
 - pipeline1
- Change Data Capture (pr... 0
- Datasets 2
 - outputdata
 - inputdata
- Data flows 1
 - dataflow1
- Power Query 0

inputdata outputdata pipeline1 dataflow1

Validate Data flow debug Debug Settings

Select settings Options

- Optimize
- Inspect
- Data preview

Preview experience Off

Properties

General Related (1)

Name * dataflow1

Description

Input columns *

Auto mapping Reset Add mapping Delete 8 mappings: All inputs mapped

tomatoes's column	Name as
abc Film	Film
abc Genre	Genre
abc Lead Studio	Lead Studio
abc Audience score %	Audience score %
abc Profitability	Profitability
abc Rotten Tomatoes %	Rotten Tomatoes %
abc Worldwide Gross	Worldwide Gross
abc Year	Year

Search ENG IN 12:08 07-09-2023

Microsoft Azure | Data Factory > datafactory1 Search factory and documentation

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

Data Factory Validate all Publish all

Activities

- General
 - Web
 - WebHook

inputdata outputdata pipeline1 dataflow1 WebAPI

Validate Debug Add trigger

Properties

General Related

Name * WebAPI

Description

Annotations + New

Web

Web1

General Settings User properties

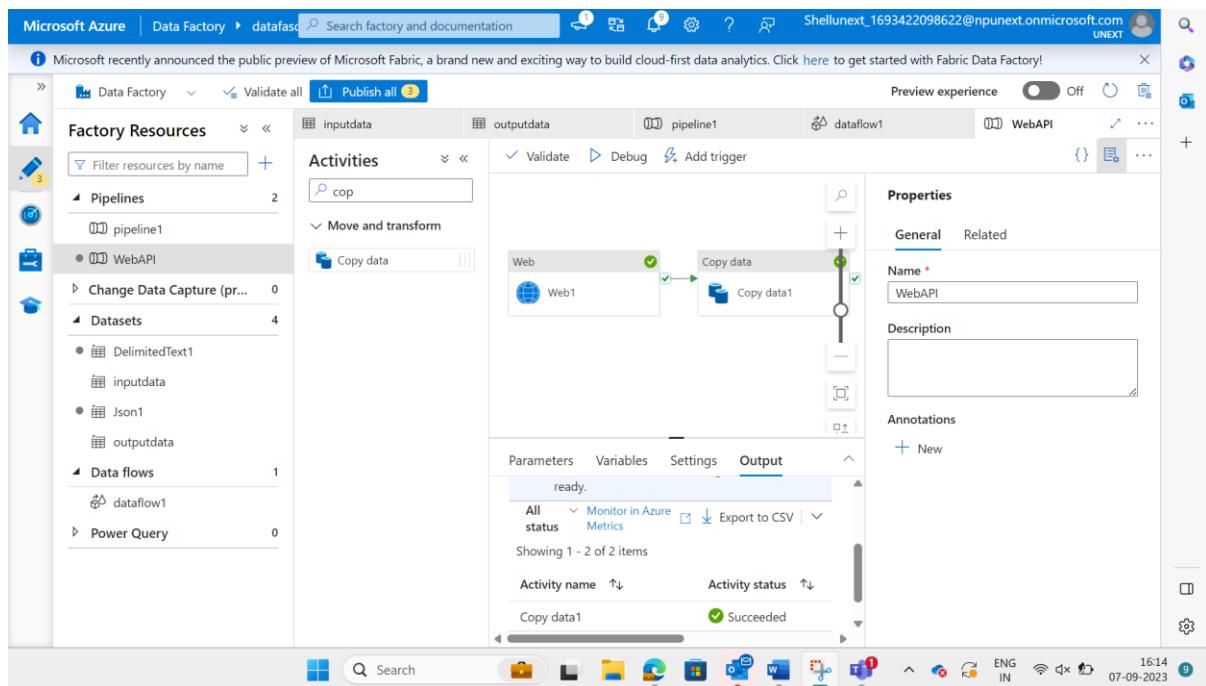
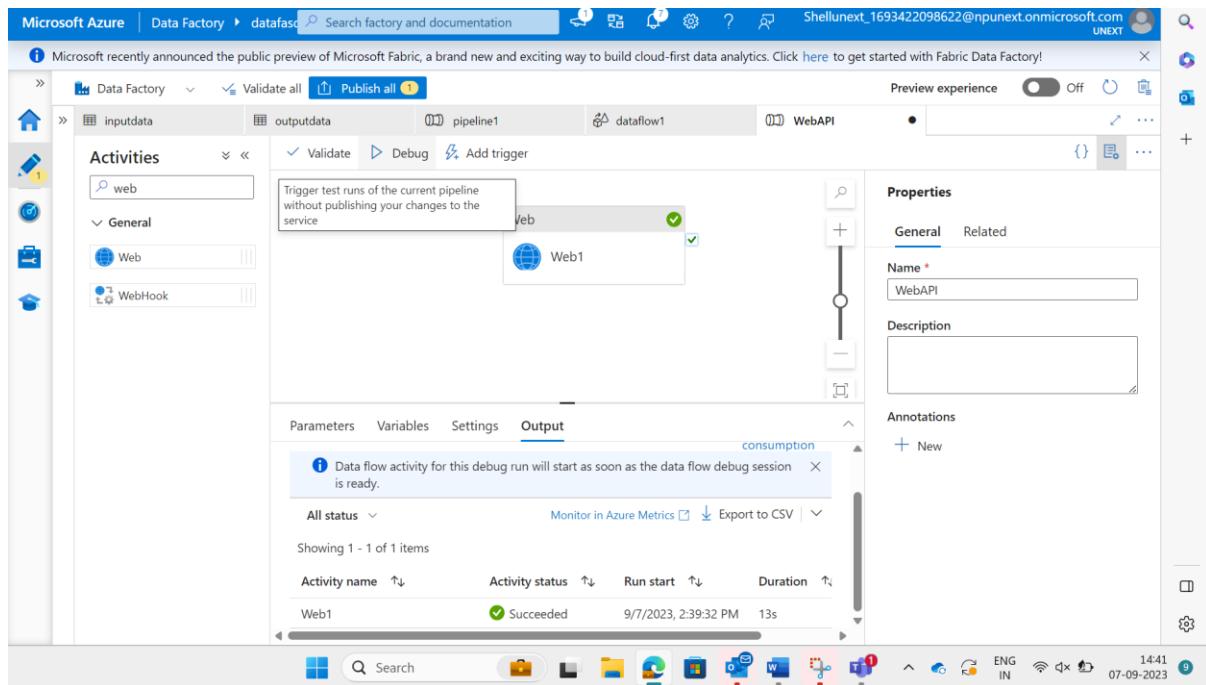
URL * https://atlas.microsoft.com/weather/curr...

Method * GET

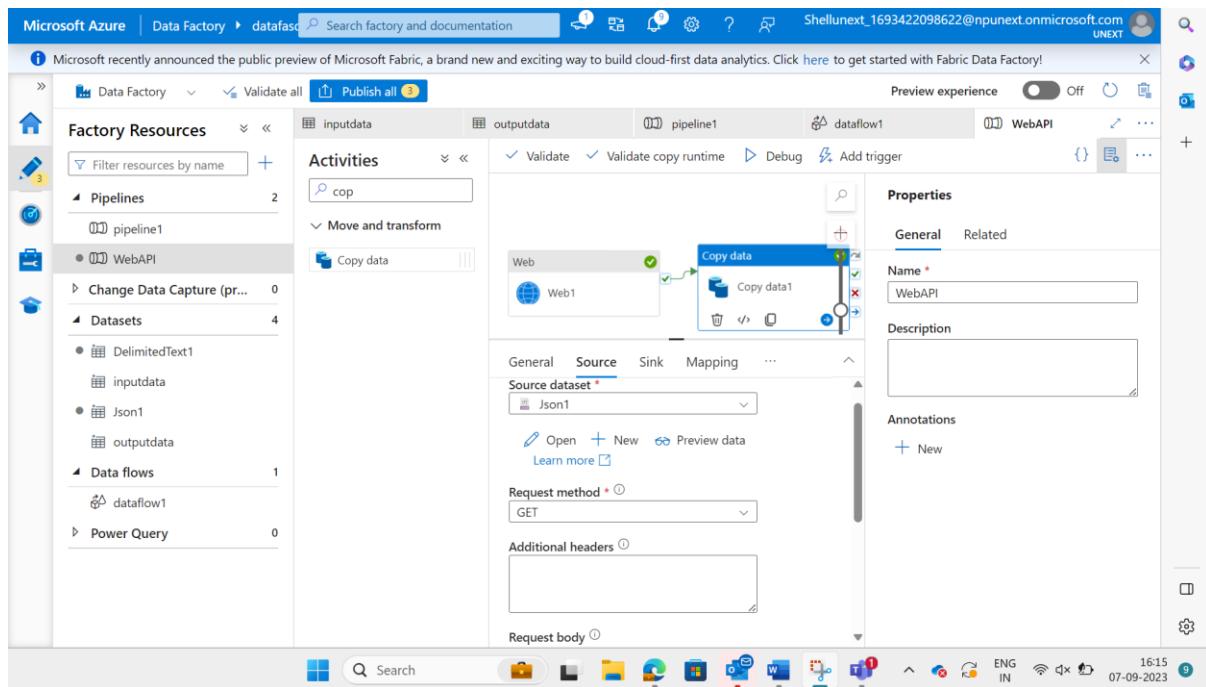
Authentication None

Headers + New

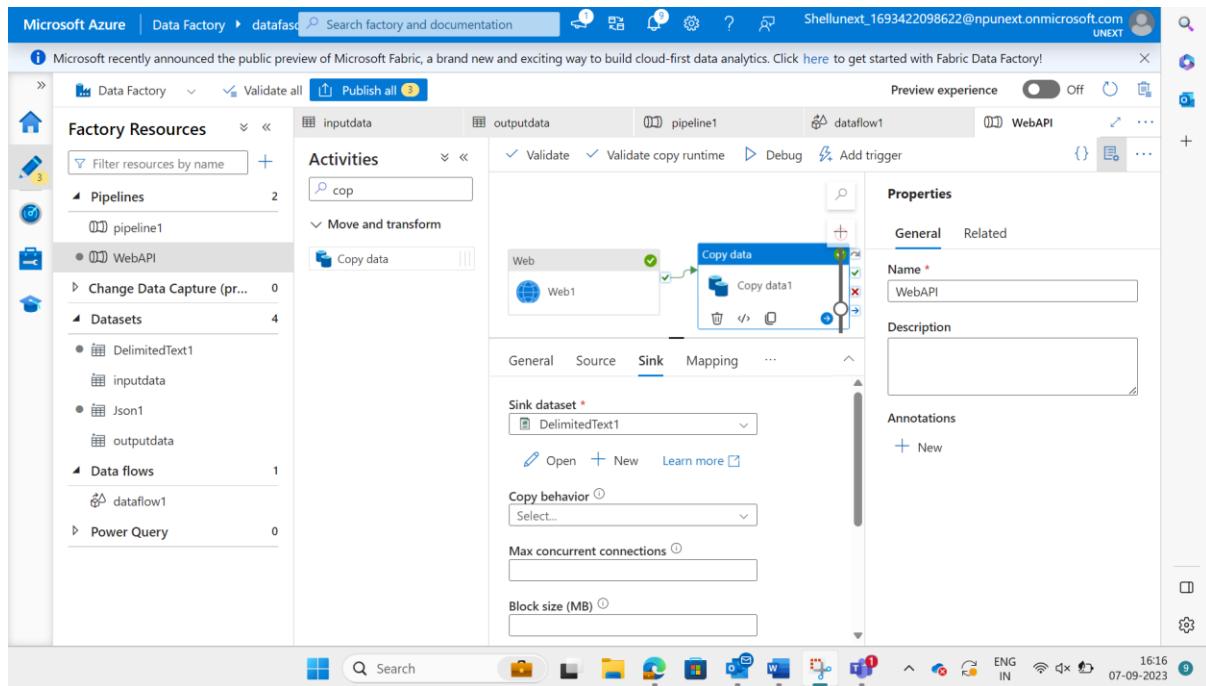
Search ENG IN 14:41 07-09-2023



Okay for this first we need to drag copy data then in source choose HTTP.



After that in sink make a csv data set in output folder container in SA.



Then validate and debug as can be seen first ss.

Microsoft Azure

Search resources, services, and docs (G+)

Shellunext_1693422098... UNEXT

Home > output07 Container

Overview Diagnose and solve problems Access Control (IAM)

Authentication method: Access key (Switch to Azure AD User Account)
Location: output07

Search blobs by prefix (case-sensitive) Show deleted blobs

Name	Modified	Access tier	Archive status	Blob type
305091bc-1c18-450a-a0ed-bddad5bfc753.txt	9/7/2023, 4:13:55 PM	Hot (Inferred)		Block blob

Add filter

Shared access tokens Access policy Properties Metadata

Search

16:20 07-09-2023

Output container now has a new http file created.

Mostly the task will be to copy http content into a csv file.

DAY-08th September 2023

Lab-> JS-> Remote Desktop(mstpc)->IP,

JS-> Jump server required to access the private resources.

ADF<->SQL/SA

ANSI SQL->MS and Oracle

In MS-> server(DB) connects to client through it

Client-> GUI, CLI, Web

Web->Query Editor

CLI-> SQL(cmd)

GUI->

Central authentication->AAD

User, password authentication->see in pic

Steps for today:

1. Create RA->SA(Gen2)->container(i/p, o/p)->ADF

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

Shellunext_1693422098...
UNEXT (NPUNEXT.ONMICROSOFT.COM)

Azure services

- [Create a resource](#)
- [Resource groups](#)
- [Storage accounts](#)
- [Virtual machines](#)
- [Quickstart Center](#)
- [App Services](#)
- [SQL databases](#)
- [Azure Cosmos DB](#)
- [Kubernetes services](#)

[More services](#)

Resources

Recent Favorite

Name	Type	Last Viewed
saasd08	Storage account	4 minutes ago
dfasd08	Data factory (V2)	5 minutes ago
rgasd08	Resource group	6 minutes ago

[See all](#)

2. Create server host-> svrzasd08.database.windows.net
3. Create SQL database-> select create new server->enter server name-> select use SQL authentication-> set user id and password->now after server create database name->workload environment as development (or else production for lots of data but charges a lot)

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

Home > Marketplace > SQL Database >

Create SQL Database

Did you know that new users in Azure can create a free Azure SQL Database and use it for 12 months using Azure free account? [Learn more](#)

Project details

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

Subscription *

Resource group * [Create new](#)

Database details

Enter required settings for this database, including picking a logical server and configuring the compute and storage resources.

Database name *

Server * [Create new](#)

Want to use SQL elastic pool? Yes No

Workload environment Development Production

[Review + create](#) [Next : Networking >](#)

Cost summary

General Purpose (GP_S_Gen5_1)
Cost per GB (in -)
Max storage selected (in GB)

ESTIMATED STORAGE COST / MONTH
COMPUTE COST / VCORE SECOND

NOTES
† Serverless databases are billed in vCore seconds based on a combination of CPU and memory utilization. Learn more about serverless billing

PLEASE CONTACT YOUR RESELLER

10:12 08-09-2023 ENG IN

Microsoft Azure

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

Home > Marketplace > SQL Database > Create SQL Database >

Create SQL Database Server

Microsoft

Server name * svrasd08 .database.windows.net

Location * (US) East US

Authentication

Select your preferred authentication methods for accessing this server. Create a server admin login and password to access your server with SQL authentication, select only Azure AD authentication [Learn more](#) using an existing Azure AD user, group, or application as Azure AD admin [Learn more](#), or select both SQL and Azure AD authentication.

Authentication method Use only Azure Active Directory (Azure AD) authentication Use both SQL and Azure AD authentication Use SQL authentication

Server admin login * asd

Password * *****

Your password must contain characters from three of the following categories – English uppercase letters, English lowercase letters, numbers (0-9), and non-alphanumeric characters (!, \$, #, %, etc.).

Confirm password * *****

OK

-> in networking select connectivity method-> select Public endpoint(don't use this in company as they have confidential data)

Microsoft Azure

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

Home > Marketplace > SQL Database >

Create SQL Database

Microsoft

Basics Networking Security Additional settings Tags Review + create

Configure network access and connectivity for your server. The configuration selected below will apply to the selected server 'svrasd08' and all databases it manages. [Learn more](#)

Network connectivity

Choose an option for configuring connectivity to your server via public endpoint or private endpoint. Choosing no access creates with defaults and you can configure connection method after server creation. [Learn more](#)

Connectivity method * No access Public endpoint Private endpoint

Firewall rules

Setting 'Allow Azure services and resources to access this server' to Yes allows communications from all resources inside the Azure boundary, that may or may not be part of your subscription. [Learn more](#)

Setting 'Add current client IP address' to Yes will add an entry for your client IP address to the server firewall.

Allow Azure services and resources to access this server * No Yes

Cost summary

General Purpose (GP_S_Gen5_1)

Cost per GB (in -) --

Max storage selected (in GB) x 41.6

ESTIMATED STORAGE COST / MONTH -- --

COMPUTE COST / VCORE SECOND -- --

NOTES

! Serverless databases are billed in vCore seconds based on a combination of CPU and memory utilization. Learn more about serverless billing

Review + create < Previous Next : Security >

-> next in firewall rules there are 2 rules-> allow azure services and resources to access this server and add current client IP address.

Private endpoint

Firewall rules

Setting 'Allow Azure services and resources to access this server' to Yes allows communications from all resources inside the Azure boundary, that may or may not be part of your subscription. [Learn more](#)

Setting 'Add current client IP address' to Yes will add an entry for your client IP address to the server firewall.

Allow Azure services and resources to access this server No Yes

Add current client IP address * No Yes

Cost per GB (in -) Max storage selected (in GB) x 41.6

ESTIMATED STORAGE COST / MONTH COMPUTE COST / VCORE SECOND

NOTES Serverless databases are billed in vCore seconds based on a combination of CPU and memory utilization. [Learn more about serverless billing](#)

PLEASE CONTACT YOUR RESELLER

- > then review and create

4. Coming to ADF back-> create linked service and under linked service select Azure Sql db-> name the service LS_Sql-> now select subscription and server and data base.-> now select SQL authentication and give the same user name and password.-> test connection-> create.
5. Create new LS-> select gen 2-> select storage account-> create it.

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

Successfully created Successfully created AzureDataLakeStorage1 (Linked service).

Name	Type	Related	Annotations
AzureDataLakeStorage1	Azure Data Lake Storage Gen2	0	
LS_SQL08	Azure SQL Database	0	

6. Now under datasets->Azure SQL database->name DS_SQL_input->select LS
->now for table name-> go back to azure portal and click on sql database-> under query editor we need to give user id and password->type sql query and insert and run

The sql query is->[10:29] Singh, Aastha SBOBNG-PTIY/RE

```
create table data_source_table
(
    PersonID int,
    Name varchar(255),
    LastModifytime datetime
);
```

```
INSERT INTO data_source_table(PersonID, Name, LastModifytime) VALUES  
(1, 'aaaa','9/1/2017 12:56:00 AM'),  
(2, 'bbbb','9/2/2017 5:23:00 AM'),  
(3, 'cccc','9/3/2017 2:36:00 AM'),  
(4, 'dddd','9/4/2017 3:21:00 AM'),  
(5, 'eeee','9/5/2017 8:06:00 AM');
```

```
create table watermarktable  
(  
    TableName varchar(255),  
    WatermarkValue datetime,  
);
```

```
INSERT INTO watermarktable VALUES ('data_source_table','1/1/2010 12:00:00 AM')
```

```
Select * from watermarktable
```

```
CREATE PROCEDURE usp_write_watermark @LastModifiedtime datetime, @TableName  
varchar(50)  
AS  
BEGIN  
    UPDATE watermarktable  
    SET [WatermarkValue] = @LastModifiedtime  
    WHERE [TableName] = @TableName
```

END

The screenshot shows two separate sessions of the Microsoft Azure SQL Database Query editor (preview). Both sessions are for the database 'dbasd08'.

Session 1:

```
1 create table data_source_table
2 (
3     PersonID int,
4     Name varchar(255),
5     LastModifytime datetime
6 );
```

Session 2:

```
1 select * from data_source_table
```

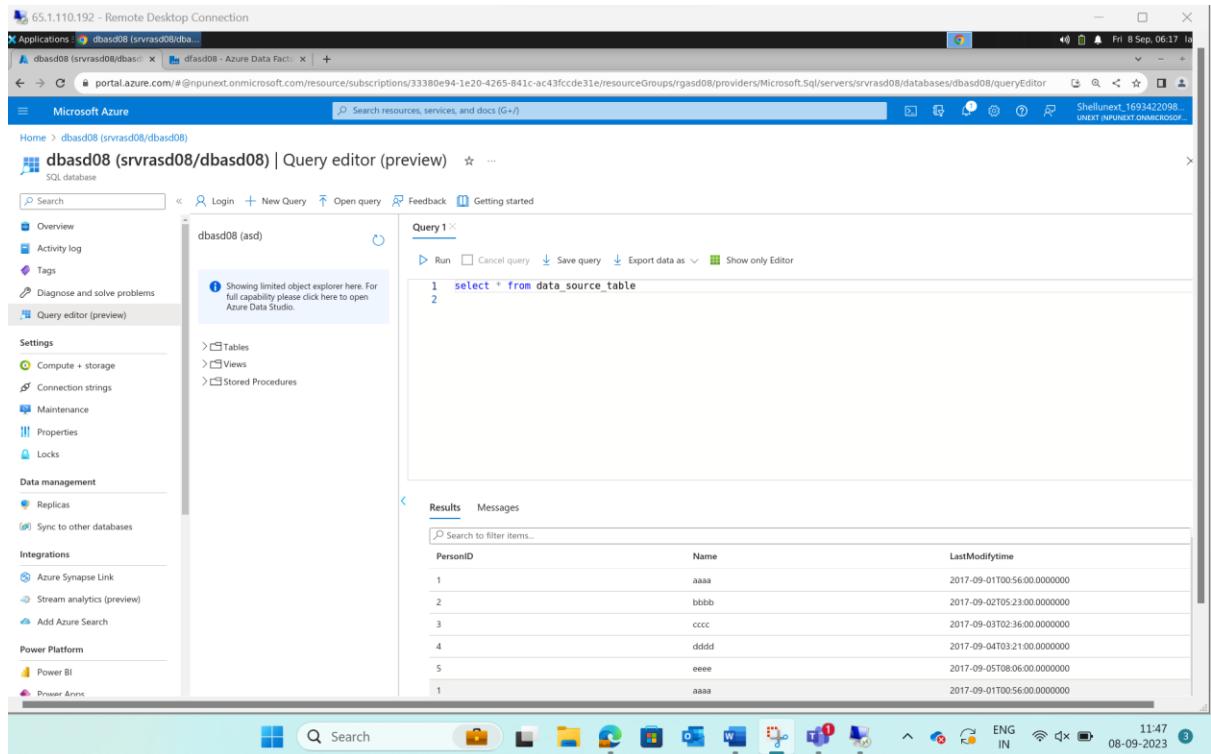
The results of the second query are displayed in a table:

PersonID	Name	LastModifytime
1	aaaa	2017-09-01T00:56:00.0000000
2	bbbb	2017-09-02T05:23:00.0000000
3	cccc	2017-09-03T02:36:00.0000000
4	dddd	2017-09-04T03:21:00.0000000

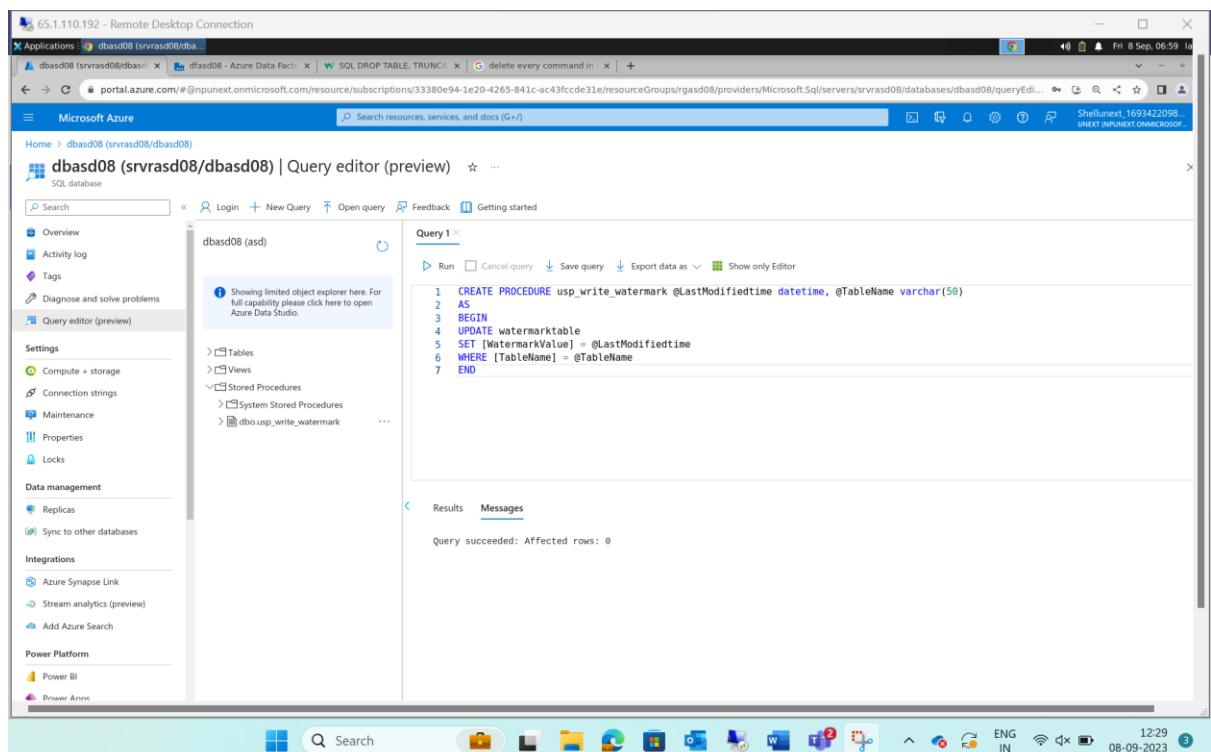
->Then refresh table name in ADF and select and create.

7. Then create another dataset for output-> gen2->from container select output folder and csv file option before this.

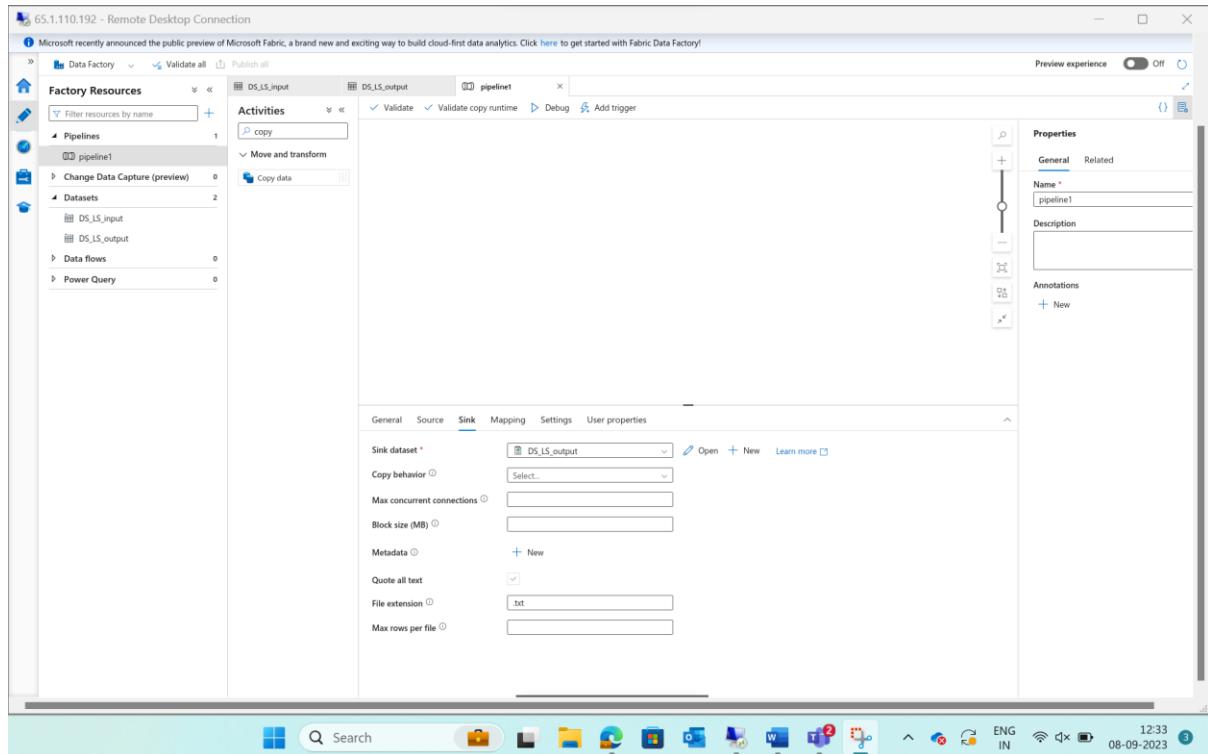
We needed to connect this through the VM



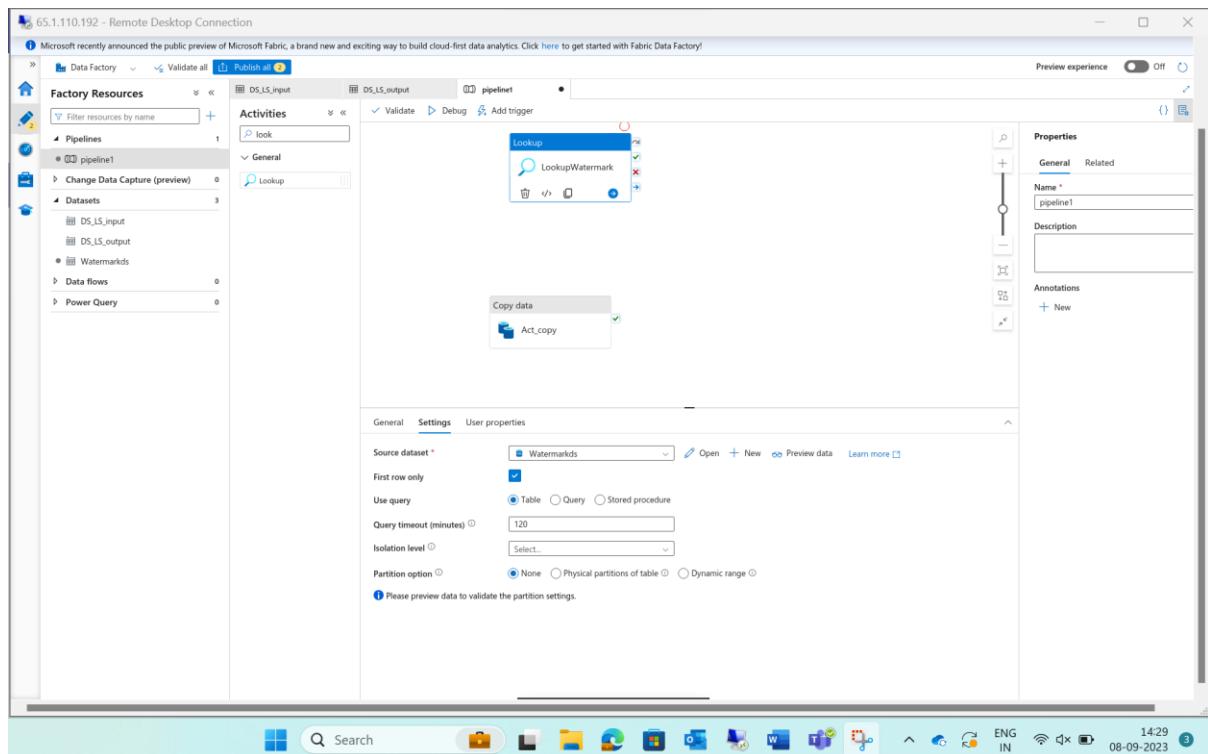
Run the whole table and then create procedure table



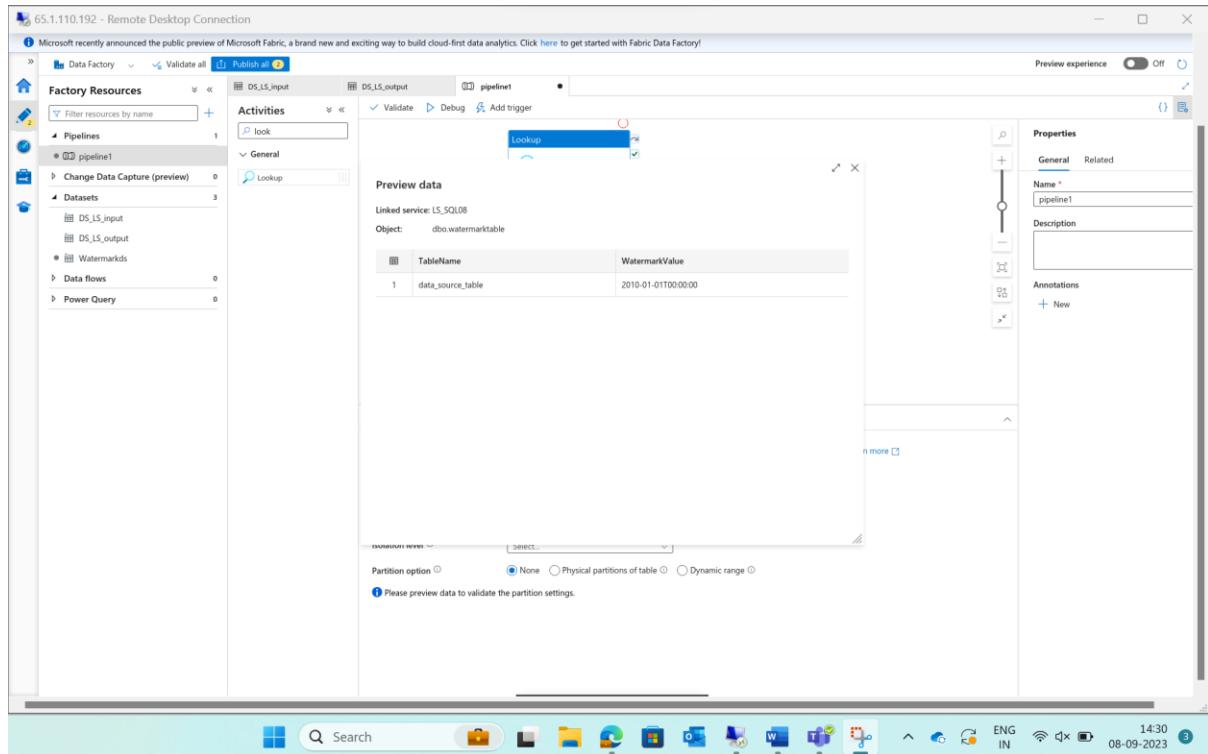
- Now in pipeline add source and sink and then validate and publish



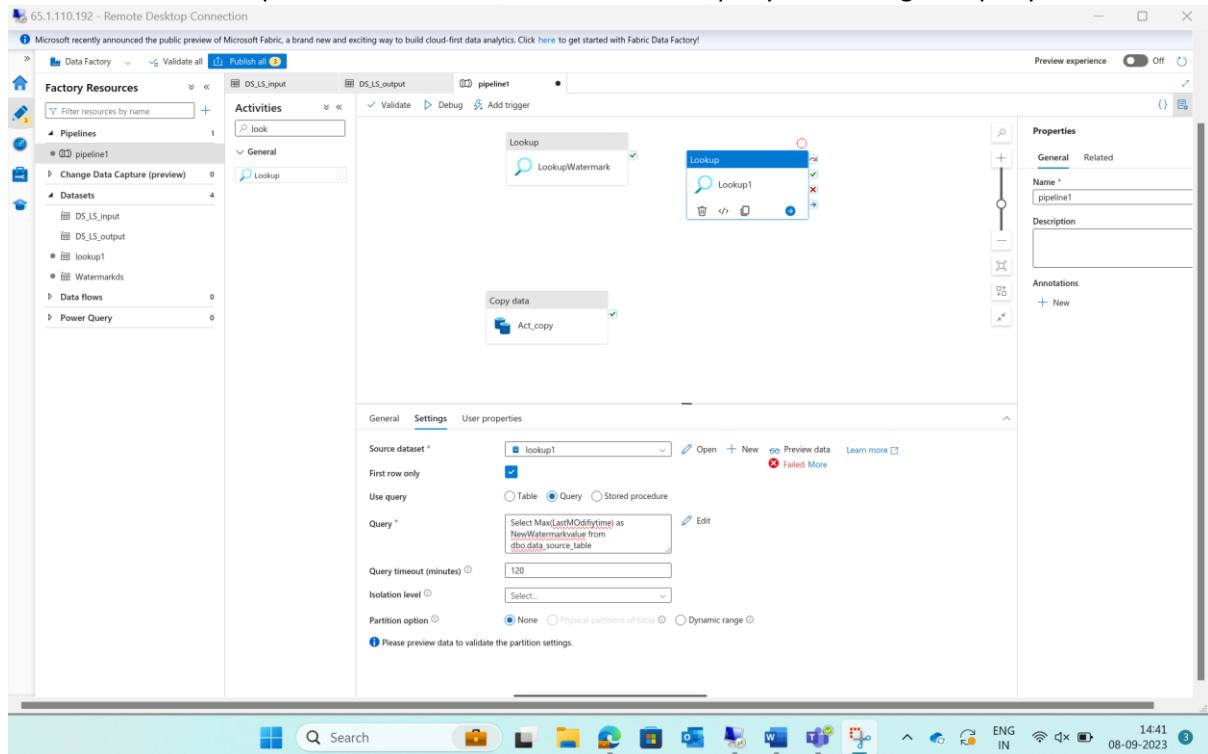
9. Next in pipeline search and drag lookup and in settings create a new dataset with azure sql database then under it name it watermarkds and then under it select dbo.watermark one.

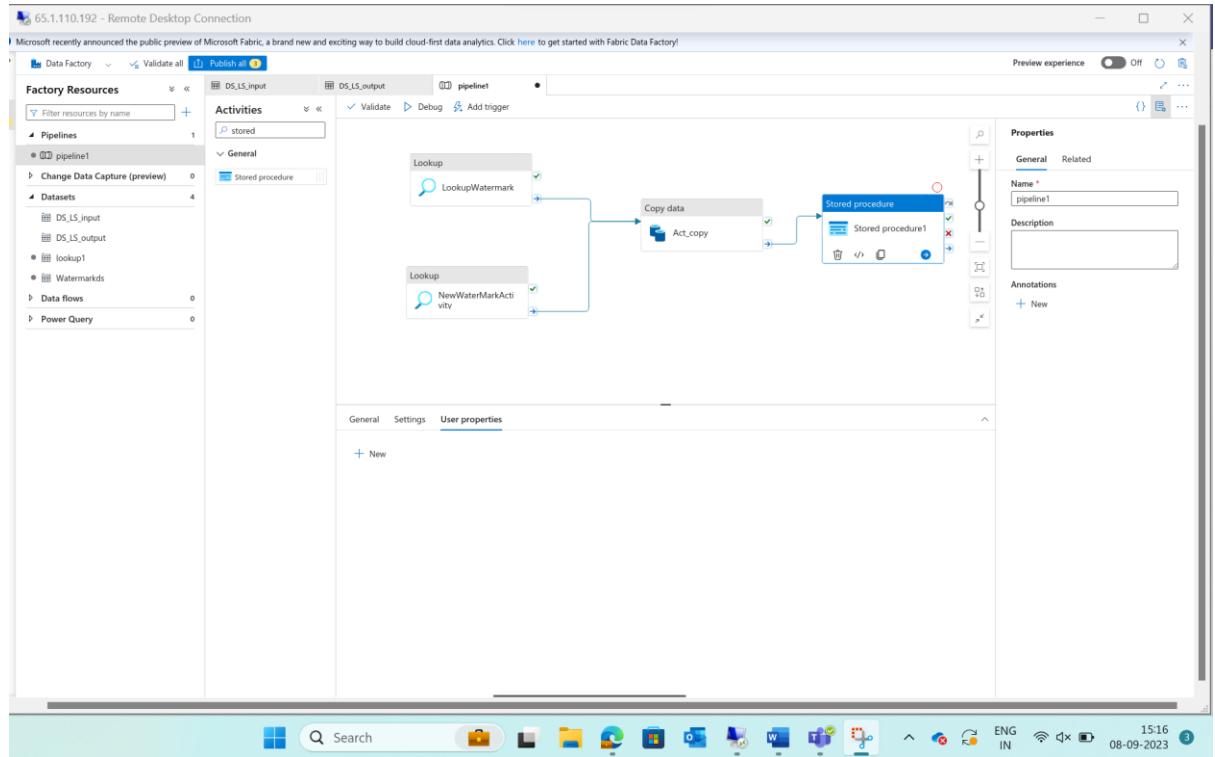


We can also preview data and see that data set has been copied.



10. Create another lookup for data source table. In that under query write the given query.





anyname.database.windows.net-> is used to access any server

WEEK 3

DAY-1-> 12th September 2023

Azure Synapse

Data-> live and dead

In dead-> Traditional, Big and New data.

It brings together the best of azure technologies used in data warehouse, apache spark and azure data explorer for log and times series analytics.

Synapse-dwh, big data and data explorer

Azure-> subscription-> resource provider-> search(synapse)-> registration

3.108.237.220 - Remote Desktop Connection

Applications npunext-1673505240396

portal.azure.com/#/npunext.onmicrosoft.com/resource/subscriptions/33380e94-1e20-4265-841c-ac43fccde31e/resourceproviders

Microsoft Azure

npunext-1673505240396 | Resource providers

Subscription

Cost Management

Cost analysis

Settings

Provider

Resource groups

Resources

Policies

Resource providers

Resource locks

Status

Microsoft.Synapse

registered

Search resources, services, and docs (G+)

Search resources, services, and docs (G+)

Register Unregister Refresh Feedback

Subscriptions global filter

My role == all

Status == all

Add filter

Subscription name ↴

npunext-1673505240396

Shellunext_16934212098_UNEXT_NPUNEXT_ONMICROSOFT

Tue 12 Sep, 07:02:14

12:32 ENG IN 12-09-2023

Register-> create resource group->create storage account(gen 2)

Qwerty1234

3.108.237.220 - Remote Desktop Connection

Applications asdsynapse - Azure Synapse

yellow_tripdata_2023-01 TLC Trip Record Data - TLC

https://d37c16vzurychx.c...

Microsoft Azure Synapse Analytics asdsynapse

Accept Reject More options

Synapse live Validate all Public

Develop Filter resources by name

SQL scripts SQL script 1

Run Undo Publish Query plan Connect to Built-in Use database master

```

1 SELECT TOP 100 *
2 FROM OPENROWSET(
3   [REBROADCAST://asdstoragesynapse.blob.core.windows.net/asfdata/yellow_tripdata_2023-01.parquet],
4   FORMAT = 'PARQUET'
5 ) AS[Result];
6
7 CREATE DATABASE DataexplorationDB
8   COLLATE Latin1_General_100_BIN2_UTF8
9
10 CREATE EXTERNAL DATA SOURCE t1l2septfs
11   WITH (LOCATION = 'https://t1l2septfs.dfs.core.windows.net')
12
13 CREATE LOGIN user_data_explorer WITH PASSWORD='localhost@12345'
14
15 CREATE user usrdb_data_explorer FOR LOGIN user_data_explorer;
16 GO
17
18 GRANT ADMINISTER DATABASE BULK OPERATIONS TO usrdb_data_explorer;
19 GO
20
21
  
```

Properties

General Related (0)

Name * SQL script 1

Description

Type sql script

Size 0 bytes

Results settings per query (1)

First 5000 rows (default)

All rows

Search

15:15 ENG IN 12-09-2023

Manage-> SQL Pool, Spark Pool

Develop-> SQL, Notebook-> spark->SQL, Python.

SQL-> SQL Pool, Notebook-> Spark Pool(language- sql, python)

Synapse see later

Day 02(week 3) – 13th September 2023

Starts with Power BI lab today

In VM we need to open power bi lab and check whether we can copy and paste text and files from original desktops to remote desktop and the open sql and power bi softwares. In sql softwares we need to make sure that DB creation works. For it right click on db and create new db.

Data centre-> storage,compute,n/w(LAN, PAN), security(Authentication, authorisation, transformation), etc. and it can have one or multiple storage.

OLTP->DB.SQL(Row)

OLAP->DB.NoSql(Column)

PowerBi-> Microsoft provide these 3 services

1. Desktop(easy and compact)
2. Service(Integration)->API, CLI,
3. Embedded mobile(light weight)

Target for today-> create a DB and connect it to PowerBI

Now first in SQL management

The screenshot shows the Microsoft SQL Server Management Studio (SSMS) interface. The title bar reads "13.233.153.95 - Remote Desktop Connection" and "EC2AMAZ-OIIIOIMA\SQLEXPRESS.sampledb - Diagram_0* - Microsoft SQL Server Management Studio (Administrator)".

The Object Explorer on the left shows the database structure:

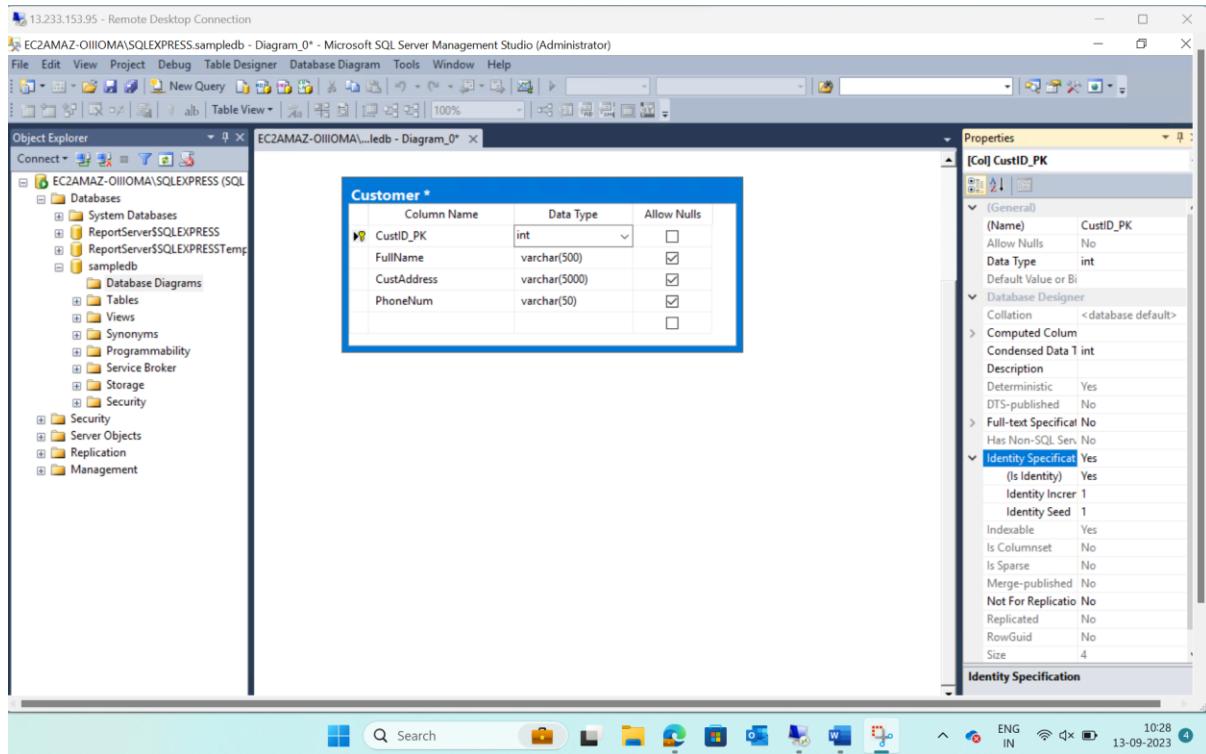
- EC2AMAZ-OIIIOIMA\SQLEXPRESS (SQL Server Database)
 - Databases
 - System Databases
 - ReportServer\$SQLEXPRESS
 - ReportServer\$SQLEXPRESSTemp
 - sampledb
 - Database Diagrams
 - Tables
 - Customer
 - Views
 - Synonyms
 - Programmability
 - Service Broker
 - Storage
 - Security
 - Security
 - Server Objects
 - Replication
 - Management

The central pane displays the table definition for "Customer".

Column Name	Data Type	Allow Nulls
CustID_PK	int	<input type="checkbox"/>
FullName	varchar(500)	<input checked="" type="checkbox"/>
CustAddress	varchar(5000)	<input checked="" type="checkbox"/>
PhoneNum	varchar(50)	<input checked="" type="checkbox"/>

The Properties pane on the right shows the table properties:

- (Identity)**
 - Name: Customer
 - Database Name: sampledb
 - Description:
 - Schema: dbo
 - Server Name: ec2amaz-oiiioima\sql
- Database Designer**
 - Identity Column: CustID_PK
 - Indexable: Yes
 - Lock Escalation: Table
 - Regular Data Space: PRIMARY
 - Replicated: No
 - Row GUID Column:
 - Text/Image Filegroup: PRIMARY



Same hi ss hai

For above in a new db select database diagrams and under it new table. Add columns and on side in properties for identify column click on it and under it select yes.

Tools-> Options-> Designer-> uncheck prevent saving changes option to prevent saving error

DB->

1. Media descriptor file
2. LDS(logs file)
3. BAK-> used to take archival backups and restore

Now download datasets from 2 links sir ne share kiya

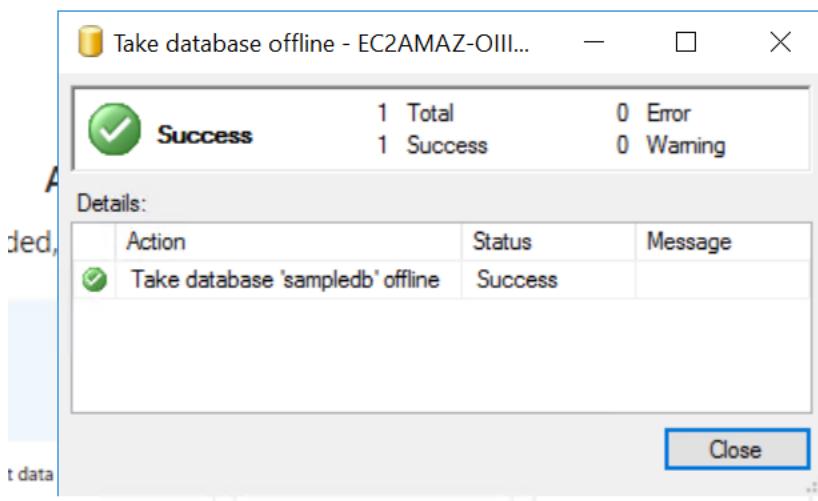
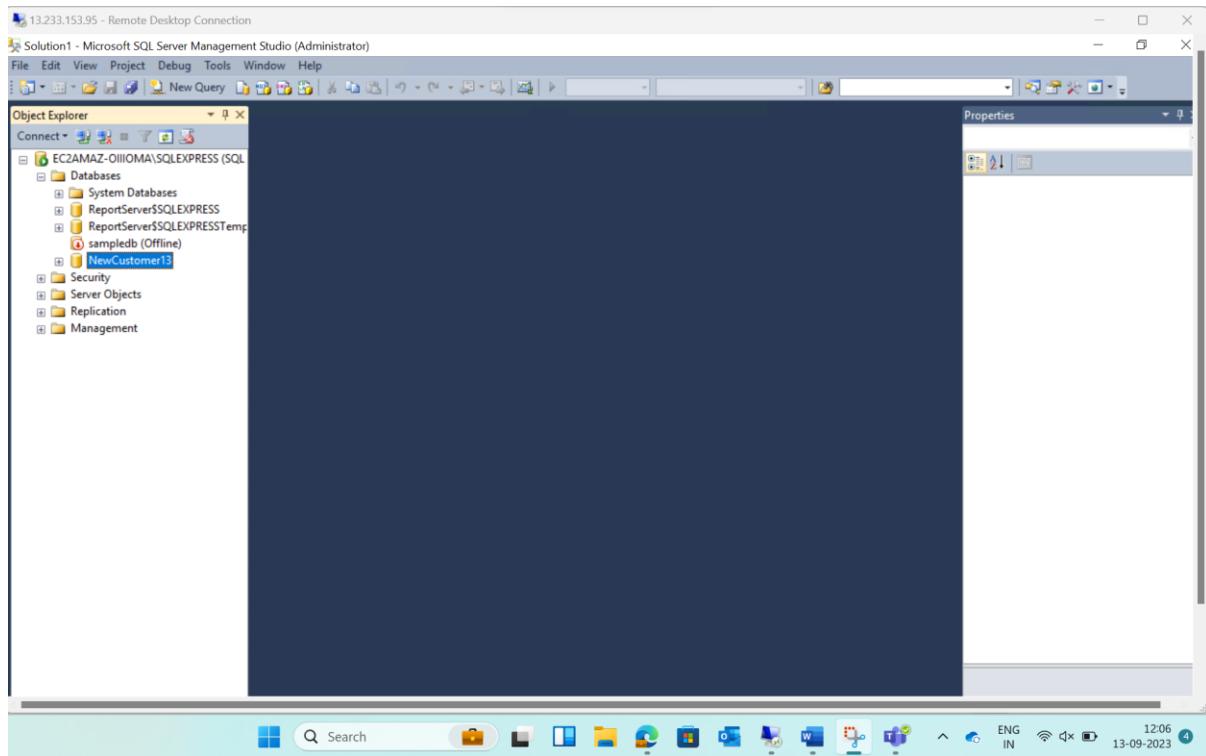
1. <https://www.bankrate.com/retirement/best-and-worst-states-for-retirement/#best-and-worst>
2. <https://github.com/Microsoft/sql-server-samples/releases/tag/adventureworks>

Now open Power BI app

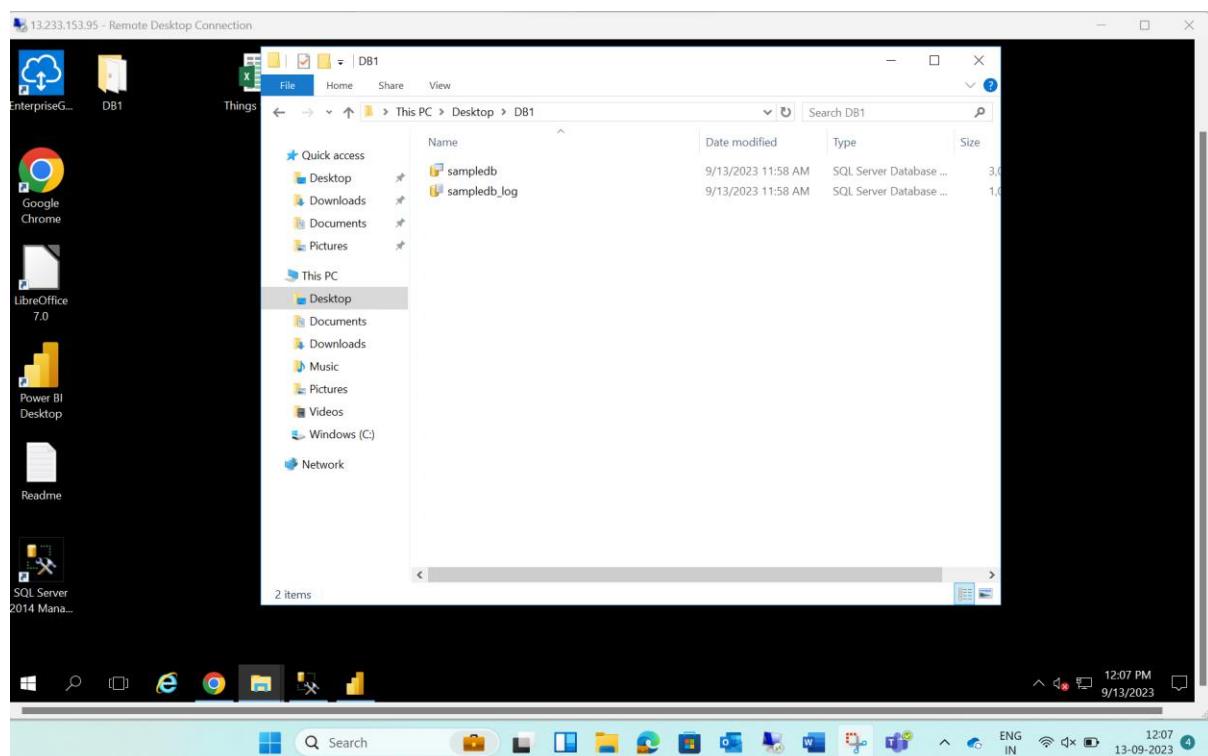
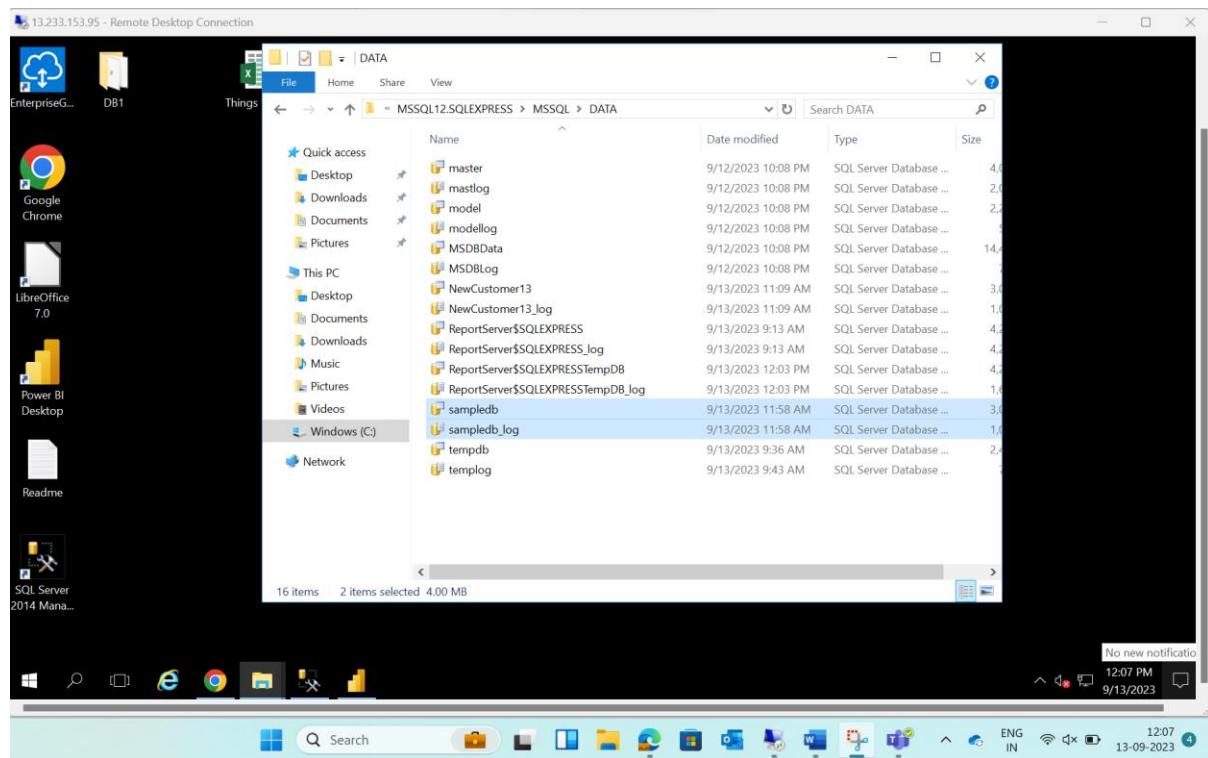
Before that make backup

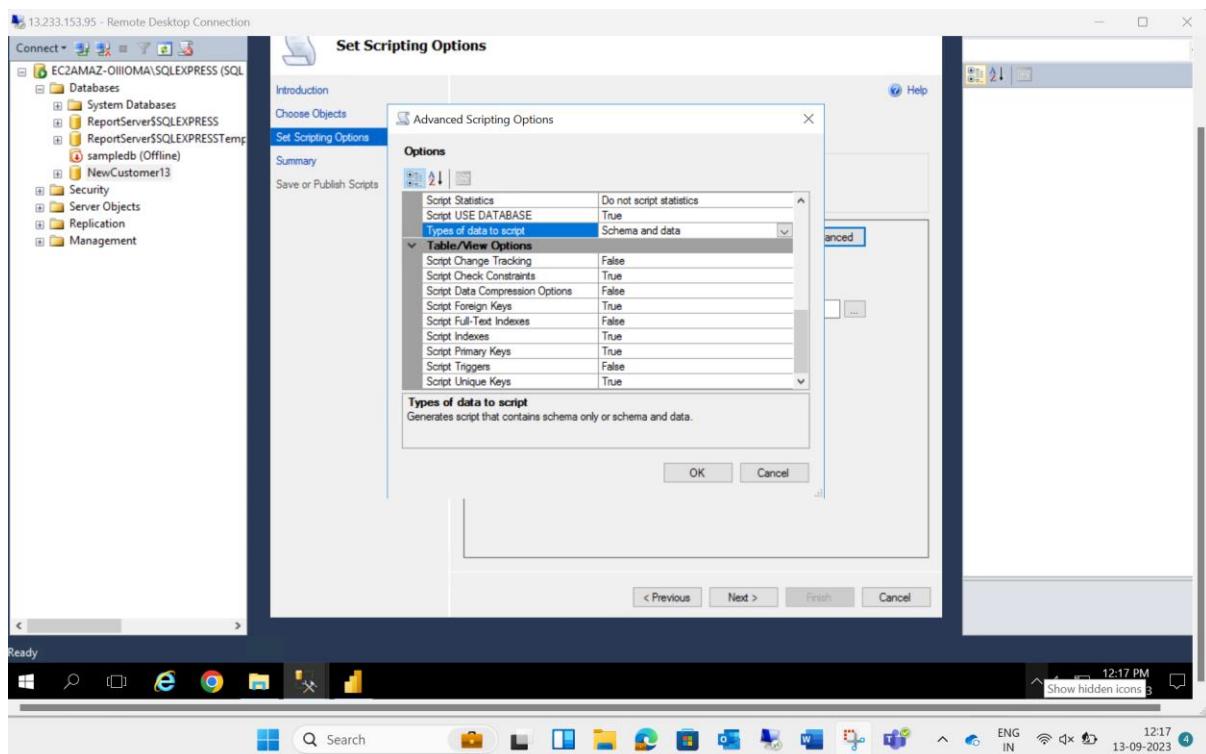
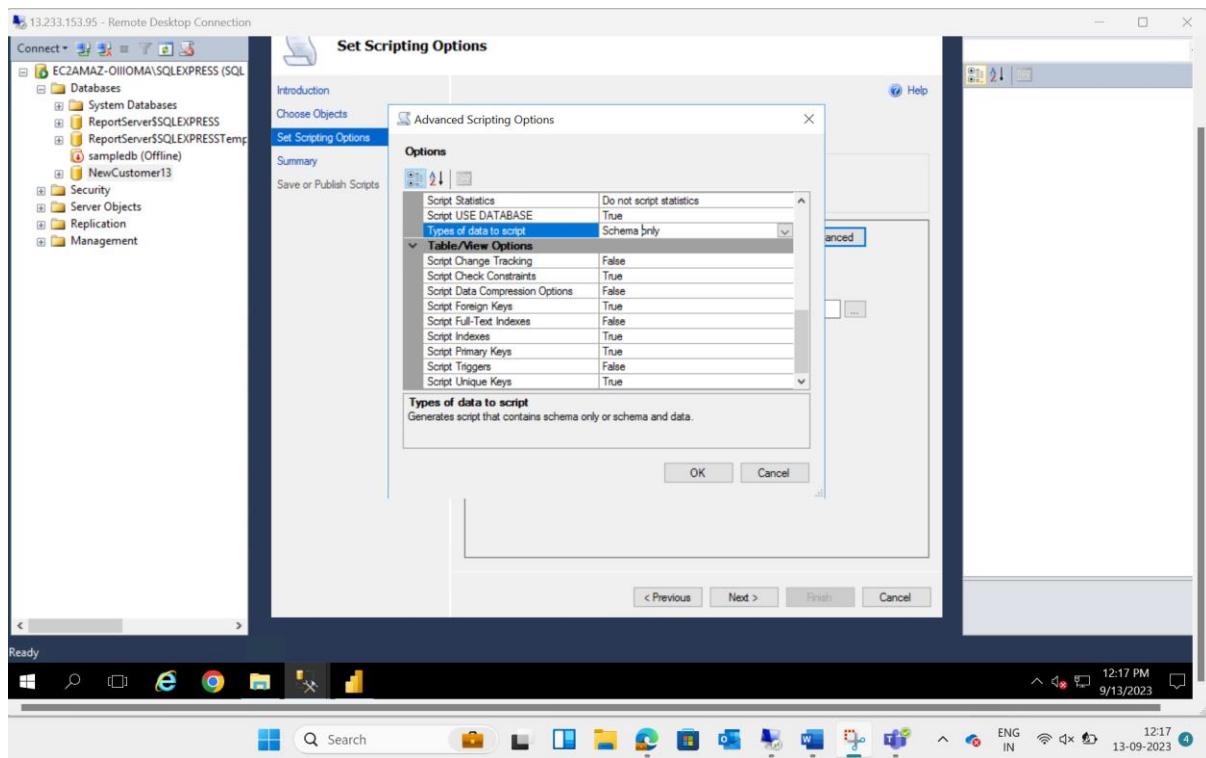
Select database jaha table hai and right click and under tasks

Tasks-> take offline to copy->then after that go to vm ka files explorer in it windows c drive -> programs files->Microsoft sql server-> data-> the table files both mdf and log files of data base and paste it on new folder created in desktop of vm



[Get data from another source →](#)



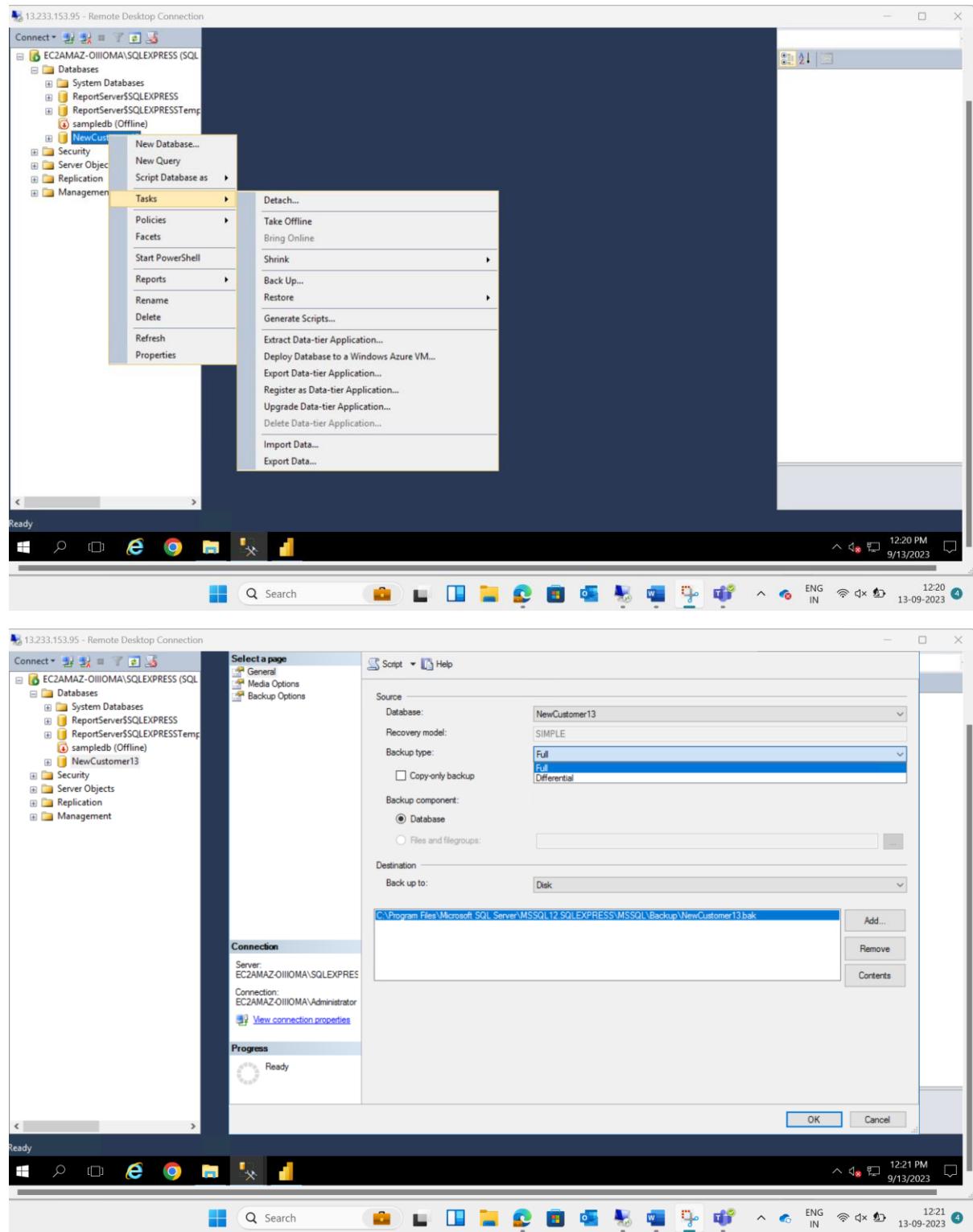


Select both option in it

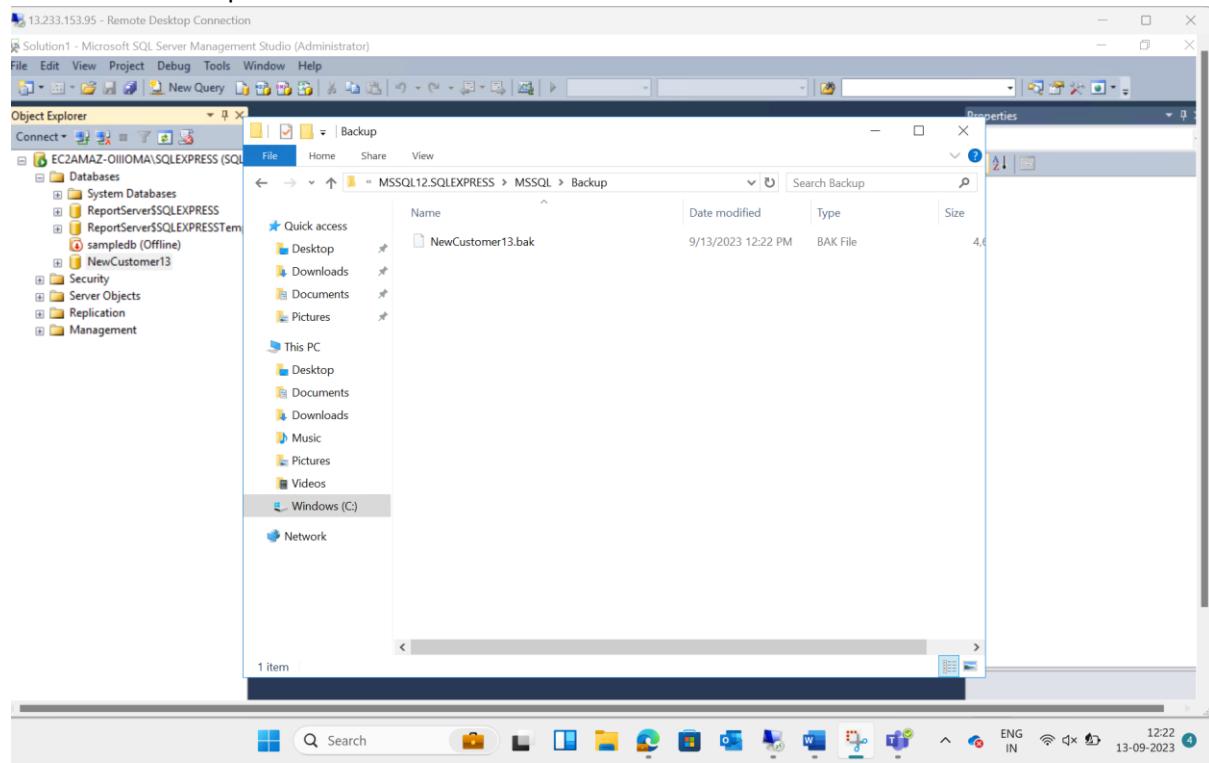
Now we created another db new customer 13 in which under tasks-> generate scripts-> under options select ss option and save

C:\Users\Administrator\Documents\script.sql

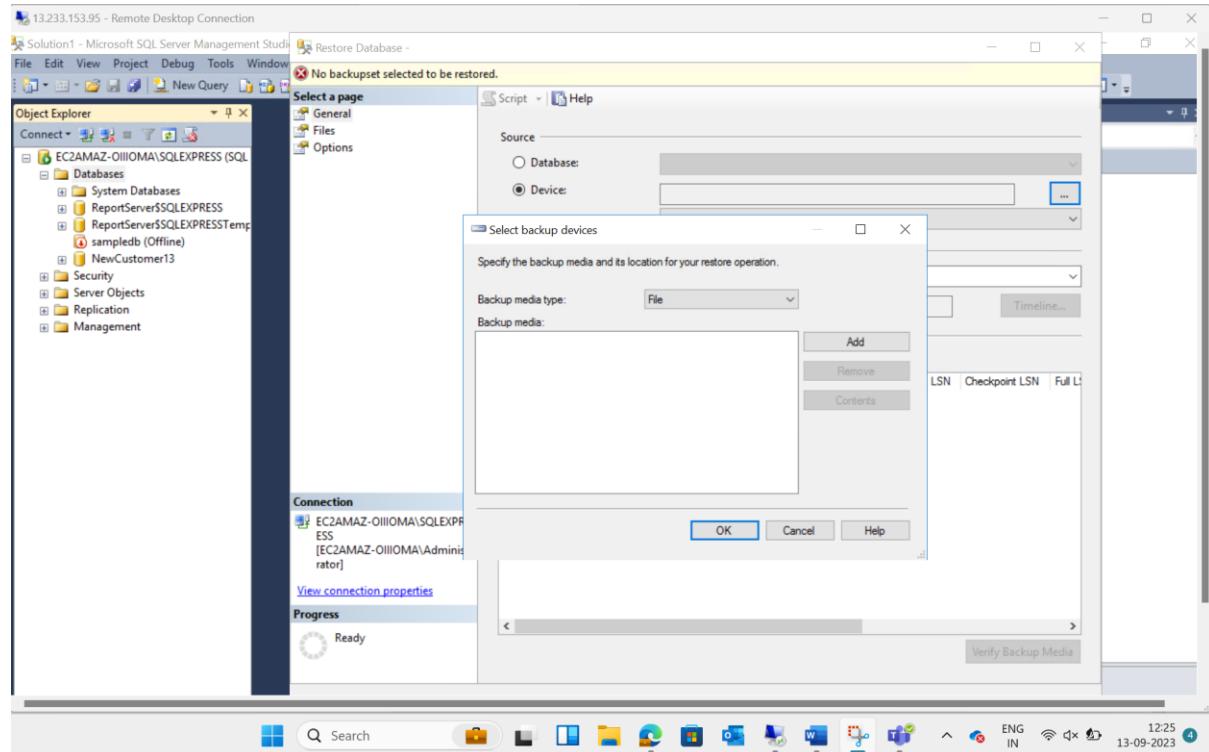
Then under same db-> tasks-> backup



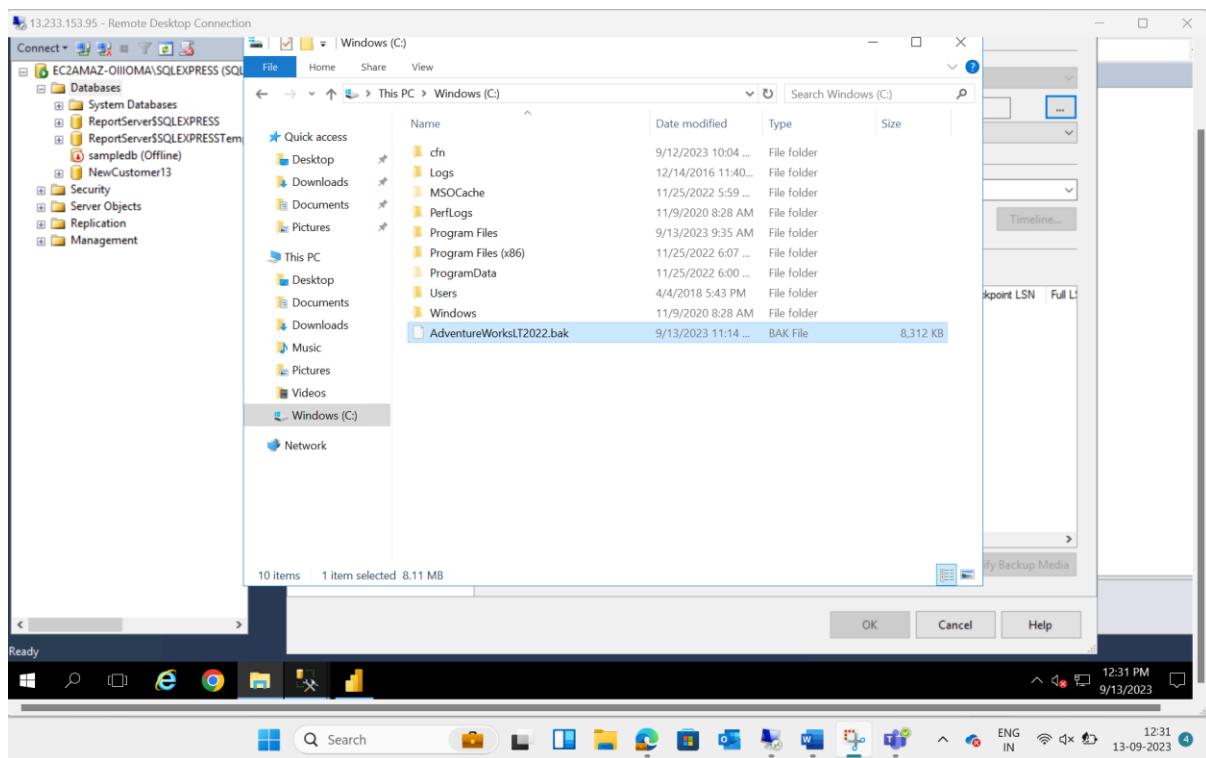
Can see the backup file in folder



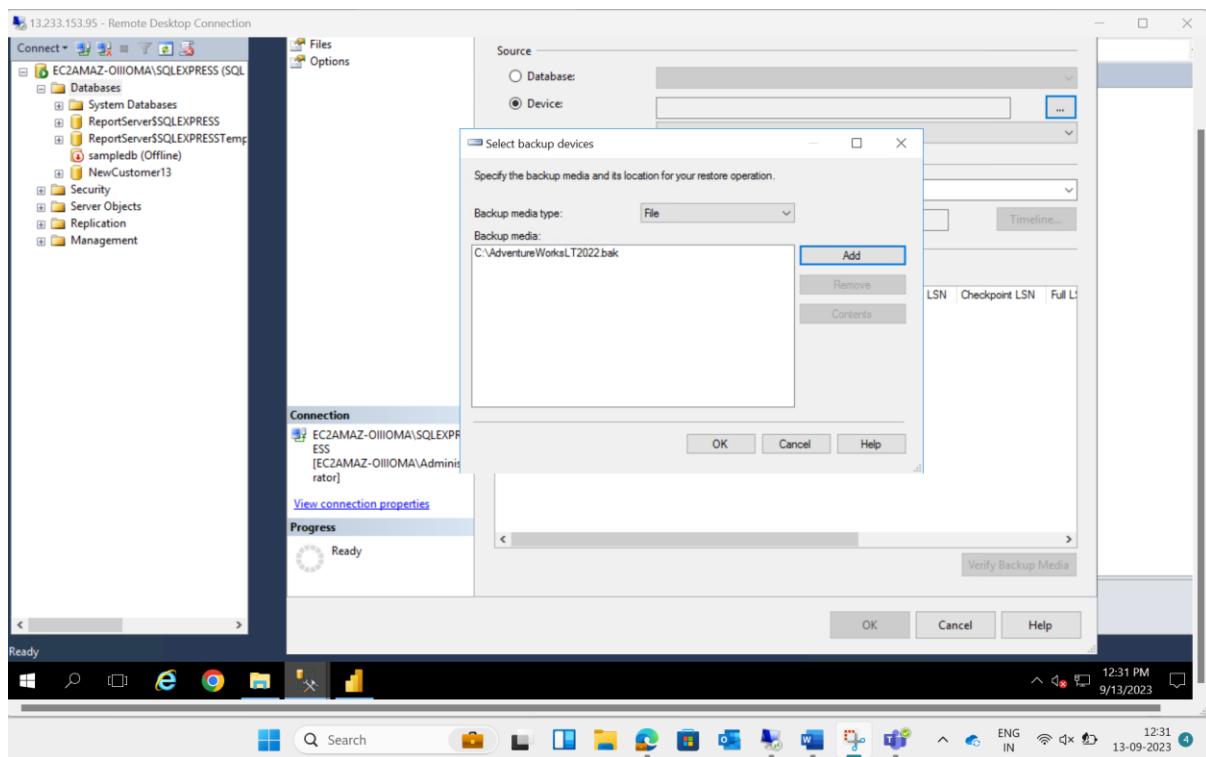
Under main db to restore the entire db we will do the following steps

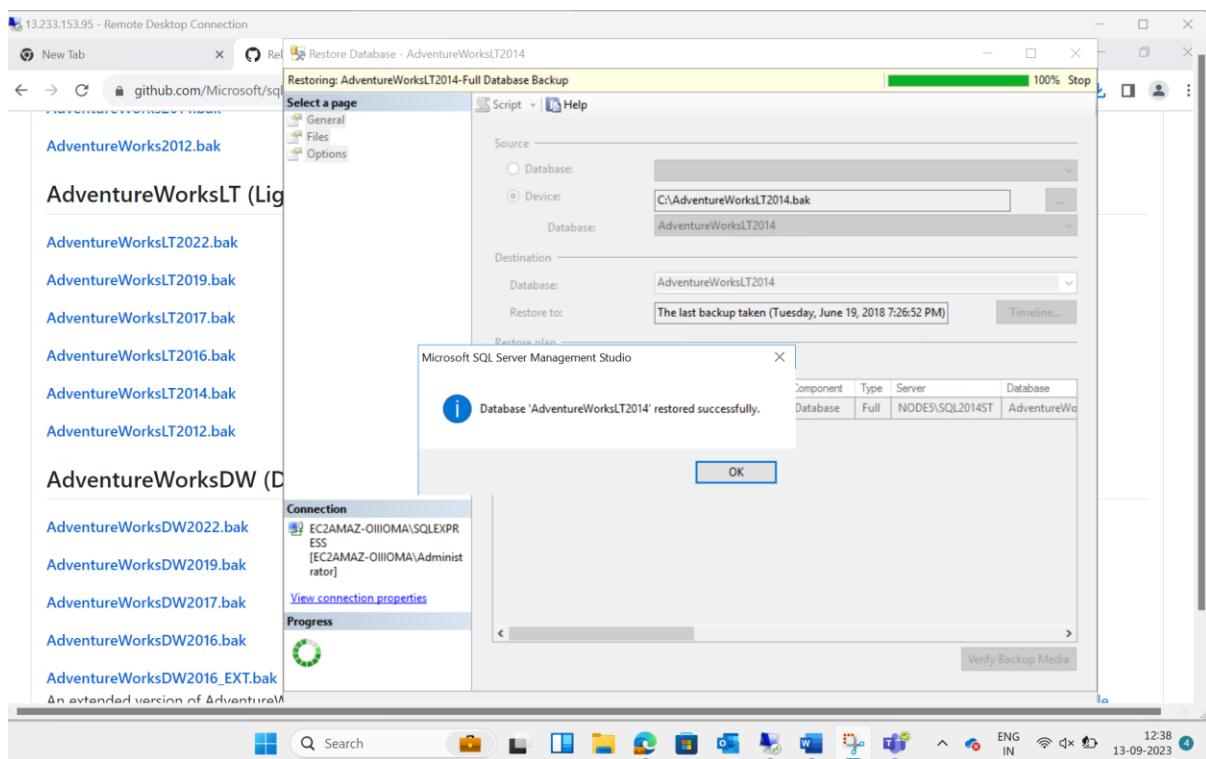
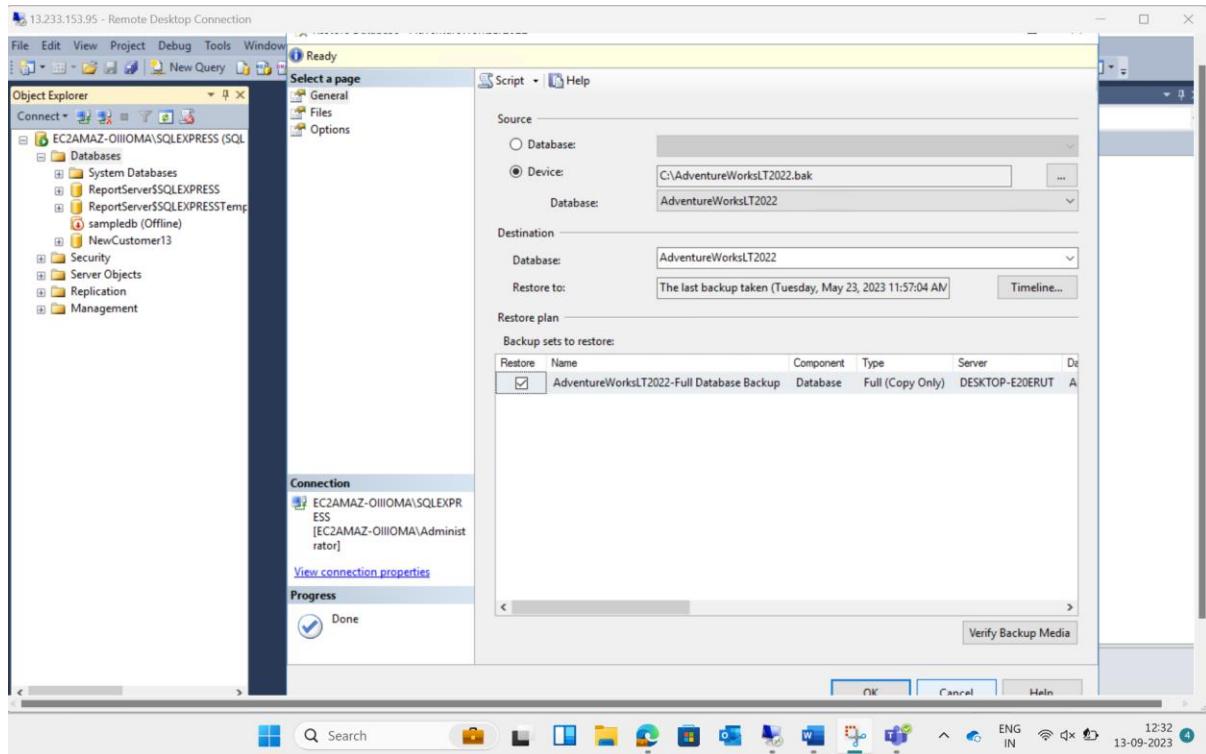


Copy the adventure works db downloaded file to c drive



Then go to add and select the file-





Now we will run queries under adventure db

3.108.249.7 - Remote Desktop Connection

SQLQuery2.sql - EC2AMAZ-OIIIOAMA\SQLEXPRESS.AdventureWorksLT2014 (EC2AMAZ-OIIIOAMA\Administrator (55)) - Microsoft SQL Server Management Studio (Administrator)

File Edit View Query Project Debug Tools Window Help

AdventureWorksLT2014 Execute Debug

Object Explorer

EC2AMAZ-OIIIOAMA\SQLEXPRESS (SQL Server)

- Databases
 - System Databases
 - AdventureWorksLT2014
 - Database Diagrams
- Tables
 - System Tables
 - FileTables
 - dbo.BuildVersion
 - dbo.ErrorLog
 - SalesLT.Address
 - SalesLT.Customer
 - SalesLT.CustomerAddress
 - SalesLT.Product
 - SalesLT.ProductCategory
 - SalesLT.ProductDescription
 - SalesLT.ProductModel
 - SalesLT.ProductModelProduct
 - SalesLT.SalesOrderDetail
 - SalesLT.SalesOrderHeader
- Views
- Synonyms
- Programmability
- Service Broker
- Storage
- Security
- NewCustomer13
- ReportServer\$SQLEXPRESS
- ReportServer\$SQLEXPRESSTempC
- sampledb (Offline)

SQLQuery2.sql - EC...Administrator (55)*

```
select * from SalesLT.Customer
```

Results Messages

CustomerID	NameStyle	Title	FirstName	MiddleName	LastName	Suffix	CompanyName	SalesPerson	EmailAddress	Phone	Pa
1	0	Mr.	Orlando	N.	Gee	NULL	A Bike Store	adventure-works\panelli	orlando@adventure-works.com	245-555-0173	Y
2	0	Mr.	Keth	NULL	Hans	NULL	Progressive Spots	adventure-works\david8	keth@adventure-works.com	170-555-0127	L
3	3	Ms.	Donna	F.	Camras	NULL	Advanced Bike Components	adventure-works\jill0	donna@adventure-works.com	279-555-0170	U
4	4	Ms.	Jean	M.	Gates	NULL	Modular Cycle Systems	adventure-works\jean0	jean@adventure-works.com	710-555-0171	B
5	5	Ms.	Julie	NULL	Hartman	NULL	Mountain Bike Supply	adventure-works\juli0	juli@adventure-works.com	180-555-0166	A
6	6	Ms.	Rosemarie	J.	Carroll	NULL	Aero Exercise Company	adventure-works\linda3	rosemarie@adventure-works.com	244-555-0116	O
7	7	Ms.	Dominic	P.	Gashi	NULL	Pural Cycle Emporium	adventure-works\luh0	dominic@adventure-works.com	192-555-0173	Z
8	10	Ms.	Kathleen	M.	Garza	NULL	Pural Cycle Emporium	adventure-works\josh0	kathleen@adventure-works.com	192-555-0173	Z
9	11	Ms.	Katherine	NULL	Herring	NULL	Sharp Bikes	adventure-works\josh1	katherine@adventure-works.com	926-555-0159	U
10	12	Mr.	Johnny	A.	Capito	Jr.	Bikes and Motobikes	adventure-works\gant0	johnny@adventure-works.com	120-555-0127	JF
11	16	Mr.	Christopher	R.	Beck	NULL	Bulk Discount Store	adventure-works\christopher0	christopher@adventure-works.com	110-555-0132	E
12	18	Mr.	Dawn	J.	Lu	NULL	Central Cycle Shop	adventure-works\lu0	dawn@adventure-works.com	445-555-0133	E1
13	19	Mr.	John	A.	Beaver	NULL	Center Cycle Shop	adventure-works\john0	john@adventure-works.com	521-555-0195	D
14	20	Ms.	Jean	P.	Handley	NULL	Central Discount Store	adventure-works\david8	jean@adventure-works.com	582-555-0113	a1
15	21	Mr.	NULL	Jinghao	Lu	NULL	Chic Department Stores	adventure-works\jill0	jinghao@adventure-works.com	928-555-0116	Ia
16	22	Ms.	Unde	E.	Burnett	NULL	Travel Systems	adventure-works\luh0	unde@adventure-works.com	121-555-0121	Z1
17	23	Mr.	Kem	NULL	Harff	NULL	Bike World	adventure-works\lu0	kem@adventure-works.com	216-555-0122	oC
18	24	Mr.	Kevin	NULL	Lu	NULL	Eastside Department Store	adventure-works\linda3	kevin@adventure-works.com	926-555-0164	y'

Properties

Current connection parameters

Aggregate Status

Connection failures

Elapsed time 00:00:01.081

Finish time 9/13/2023 2:31:36 PM

Name EC2AMAZ-OIIIOAMA\SC

Rows returned 847

Start time 9/13/2023 2:31:35 PM

State Open

Connection

Connection name EC2AMAZ-OIIIOAMA\SC

Connection elapsed 00:00:01.081

Connection finish ti 9/13/2023 2:31:36 PM

Connection rows re 847

Connection start fin 9/13/2023 2:31:35 PM

Connection state Open

Display name EC2AMAZ-OIIIOAMA\SC

Login name EC2AMAZ-OIIIOAMA\SC

Server name EC2AMAZ-OIIIOAMA\SC

Server version 12.0.2000

Session Tracing ID

SPID 55

14:32 13-09-2023

3.108.249.7 - Remote Desktop Connection

Untitled - Power BI Desktop

File Home Insert Modeling View Help

File Home Transform Add Column View Tools Help

Queries [1]

= AdventureWorksLT2014([Schema="SalesLT",Item="Product"])[Data]

ProductID	Name	ProductNumber	Color
680	HL Road Frame - Black, 58	FR-R928-58	Black
706	HL Road Frame - Red, 58	FR-R92R-58	Red
707	Sport-100 Helmet, Red	HL-US09-R	Red
708	Sport-100 Helmet, Black	HL-US09	Black
709	Mountain Bike Socks, M	SO-8909-M	White
710	Mountain Bike Socks, L	SO-8909-L	White
711	Sport-100 Helmet, Blue	HL-US09-B	Blue
712	AWC Logo Cap	CA-1098	Multi
713	Long-Sleeve Logo Jersey, S	Li-0192-5	Multi
714	Long-Sleeve Logo Jersey, M	Li-0192-M	Multi
715	Long-Sleeve Logo Jersey, L	Li-0192-L	Multi
716	Long-Sleeve Logo Jersey, XL	Li-0192-X	Multi
717	HL Road Frame - Red, 62	FR-R92R-62	Red
718	HL Road Frame - Red, 44	FR-R92R-44	Red
719	HL Road Frame - Red, 48	FR-R92R-48	Red
720	HL Road Frame - Red, 52	FR-R92R-52	Red
721	HL Road Frame - Red, 56	FR-R92R-56	Red
722	LL Road Frame - Black, 58	FR-R38B-58	Black

20 COLUMNS, 295 ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED AT 3:01 PM

Properties

APPLIED STEPS

Source Navigation

15:02 13-09-2023

Now in PowerBI we need to go on sql data base-> type name of the main data base click next and create.

Then in query editor write the query after selecting adventure table under navigator window and select sales product table and click on apply changes

Day 03(week 3) – 14th September 2023

Click on Get data-> Web-> type url ‘<https://www.bankrate.com/retirement/best-and-worst-states-for-retirement/#best-and-worst>’-> click okay-> navigator opens

The screenshot shows the Power BI Desktop interface. On the left, there's a report pane with a table visual titled 'RigName Sum of DrilledDepth'. The table contains three rows: Rig A (1,500.00), Rig B (1,750.00), and a total row (3,250.00). To the right of the report pane are several toolbars and panes. The 'Filters' pane shows filters for 'RigName' (is (All)) and 'Sum of DrilledDepth' (is (All)). The 'Visualizations' pane displays a grid of visualization icons. The 'Fields' pane lists fields from two datasets: 'DrillingData' and 'DrillingRigs'. Under 'DrillingData', fields include DataID, Date, DrillBitChanges, DrilledDepth, DrillingSpeed, OperatingHours, and RigID. Under 'DrillingRigs', fields include RigID and RigName. The bottom of the screen shows the Windows taskbar with various pinned icons.

Insert-> buttons-> under action give an event

Select any dataset->visualisation->py->write all commands in the block->

Tomorrow case study 😊

Day 04(week 3) – 15th September 2023

Case study done

Today Python starts and we will be doing it in anaconda Jupyter terminal.

Day 01(week 4) – 19th September 2023

Absent but did some python handson

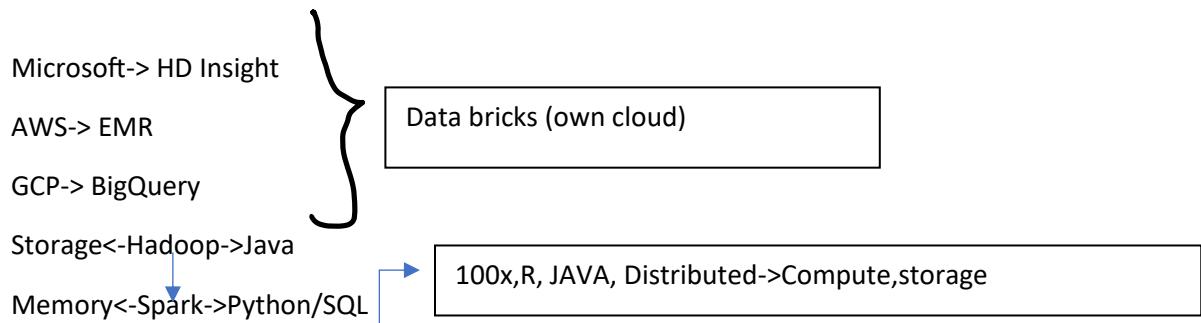
Day 02(week 4) – 20th September 2023

Big data

Big data Problem-

- Processing->framework, cluster
- Storage->distributed, i/o operations

Big data platform



Compute->clustering->master slave architecture

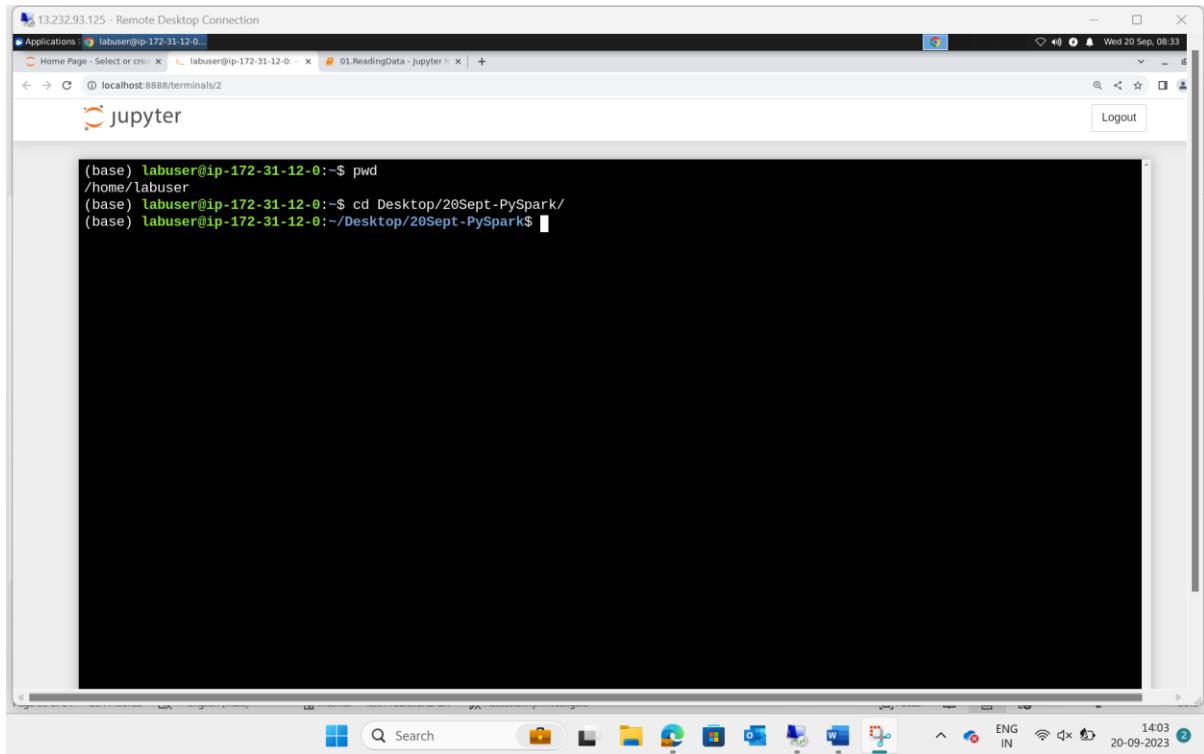
Difference between frameworks and library

Lifecycle of spark

1. Submit an application either interactive or program
2. Initialization of program-> when application is submitted, spark program is launched on a cluster node. The driver program initialized the spark context which is responsible for coordinating job execution.
3. Job and stage creation-> spark app is divided into one or more jobs. Each job consists of stages
4. DAG Gen-> spark constructs dag that are logical execution plans... idk kya bola. DAG contains information about dependencies between stages and tasks.
5. Task Schedule-> The scheduler takes the DAG and schedules tasks for execution.
6. Task execution-> tasks are executed on worker node in parallel or distributed mode.
7. Shuffle and Data exchange->
8. Task completion-> as tasks completes, they produce immediate results. These results are cached or persisted in manually for subsequent task or stages to use which reduces the need for de-compilation.
9. Stage completion-> stages are marked as complete when all of their tasks are finished successfully.
10. Job completion->

SPARK-

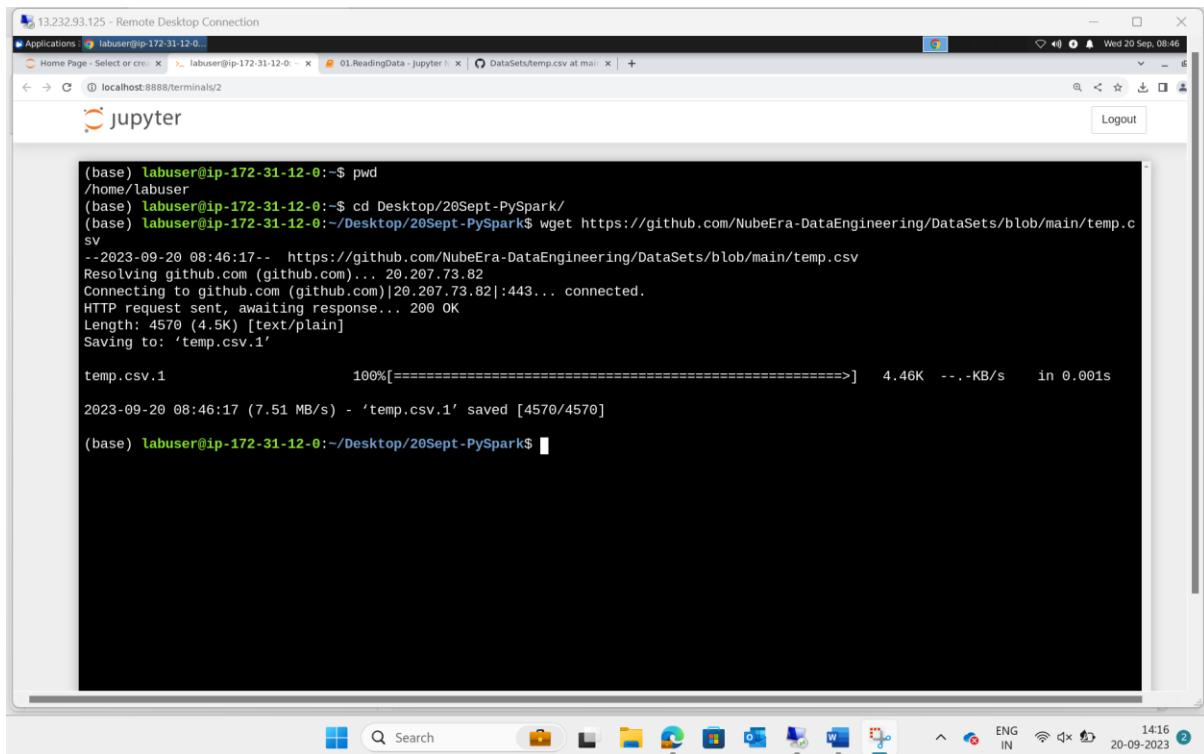
- Task is a single operation applied to a single partition.
- It is executed as a single thread/process in an executer.
- Why spark-
 - > You can perform batch processing in spark.
 - > it will support interactive sql.
 - > real time analysis
 - > structured, unstructured and semi structured data can be processed.
 - > streaming
 - > can perform graph oriented data base
 - > can perform machine learning and python operations
 - > poly lang support.
 - > graph oriented operations
-



Open terminal in python and then open new folder which you created in desktop

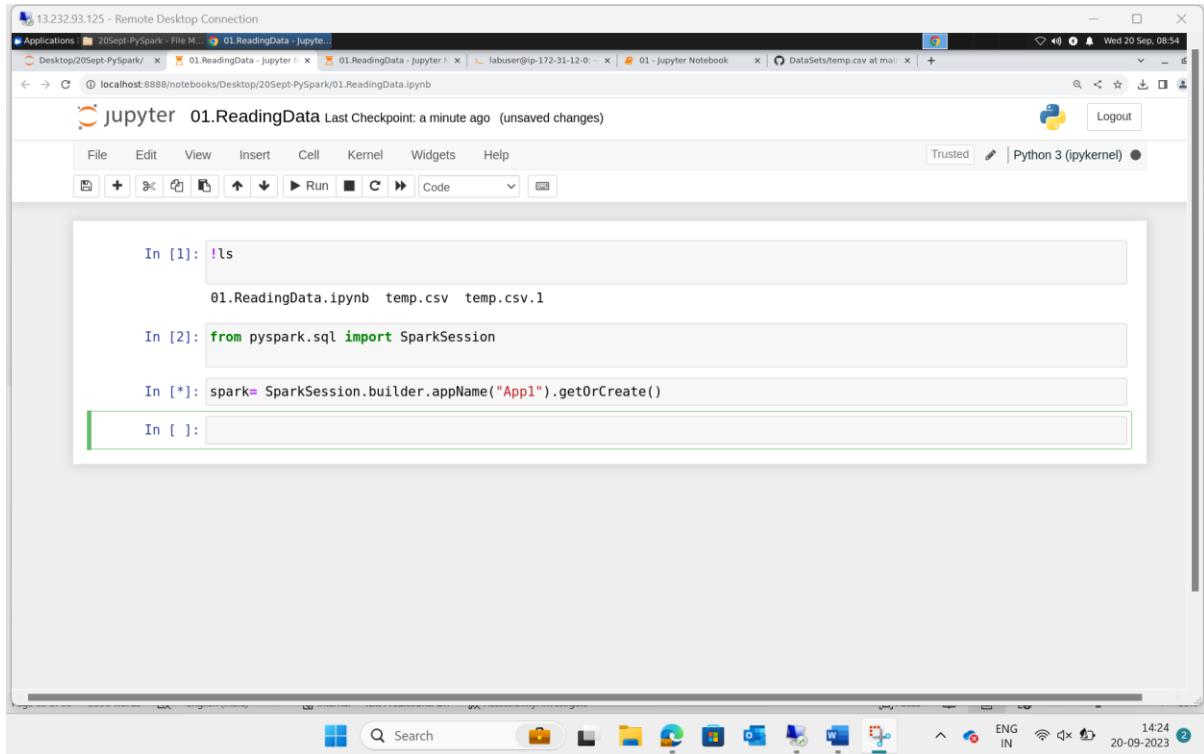
<https://github.com/NubeEra-DataEngineering/DataSets>

download file from above url into the new folder created in the vm desktop



Using wget paste the url of the file

Now in a new python window type !ls to view all files. And create the file in the same desktop file

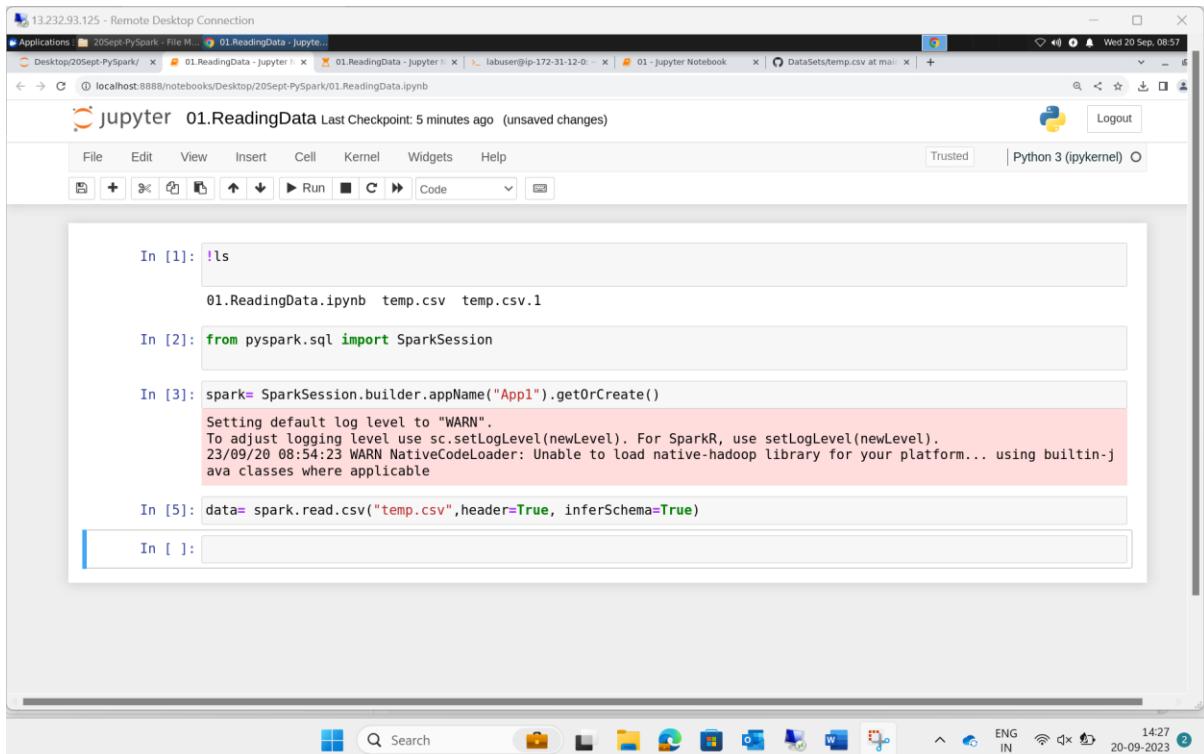


The screenshot shows a Jupyter Notebook interface running on a Windows desktop. The title bar reads "13.232.93.125 - Remote Desktop Connection". The notebook has three cells:

- In [1]: `!ls`
Output: `01.ReadingData.ipynb temp.csv temp.csv.1`
- In [2]: `from pyspark.sql import SparkSession`
- In [*]: `spark= SparkSession.builder.appName("App1").getOrCreate()`

The bottom cell is currently active, indicated by a green border.

Ignore the warning



The screenshot shows a Jupyter Notebook interface running on a Windows desktop. The title bar reads "13.232.93.125 - Remote Desktop Connection". The notebook has five cells:

- In [1]: `!ls`
Output: `01.ReadingData.ipynb temp.csv temp.csv.1`
- In [2]: `from pyspark.sql import SparkSession`
- In [3]: `spark= SparkSession.builder.appName("App1").getOrCreate()`
Output (highlighted in pink):

```
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/09/20 08:54:23 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
```
- In [4]: `df= spark.read.csv("temp.csv",header=True, inferSchema=True)`
- In [5]: `In []:`

There are 2 types of path relative(/desktop/foldername/filename) and absolute path(.temp.csv).

The screenshot shows a Jupyter Notebook interface running on a remote desktop connection. The notebook has three cells:

- In [3]:

```
spark= SparkSession.builder.appName("App1").getOrCreate()
```

Output:

```
Setting default log level to "WARN".  
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
```
- In [5]:

```
data= spark.read.csv("temp.csv",header=True, inferSchema=True)
```
- In [6]:

```
data.show()
```

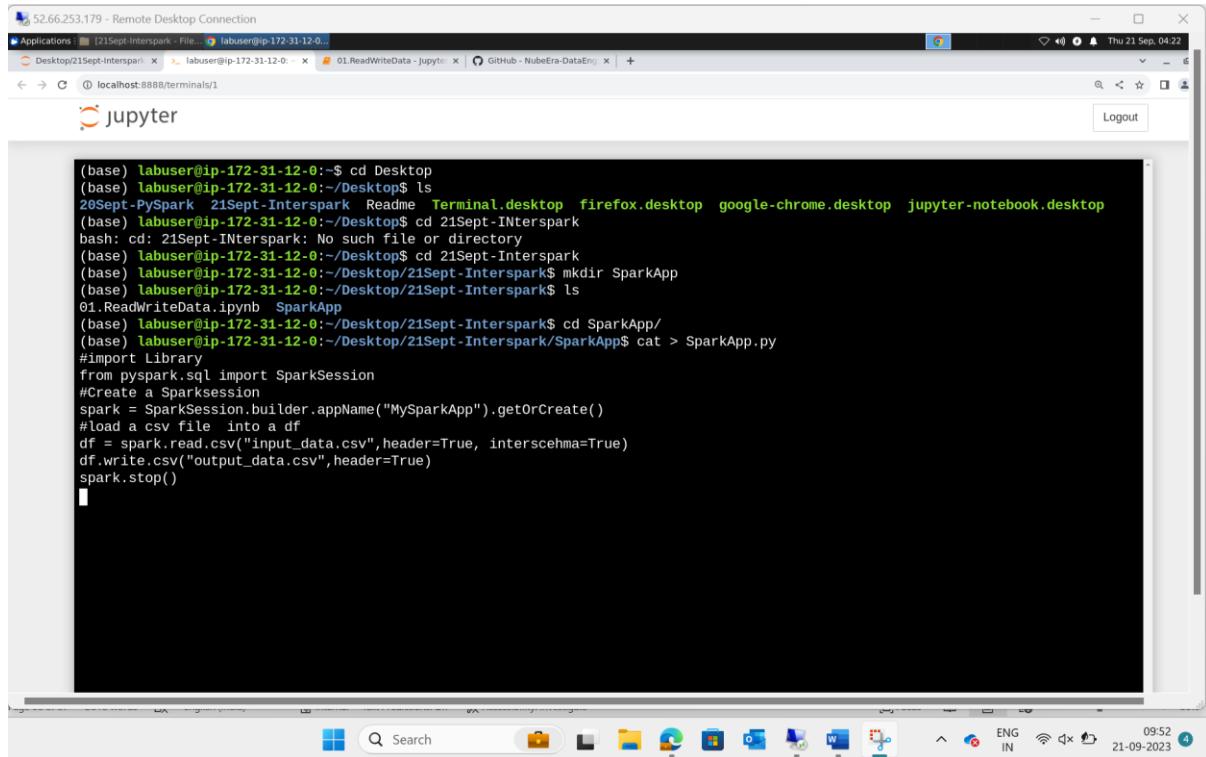
Output:

Day	Weather	Temperature	Wind	Humidity
Mon	Sunny	12.79	13	30
Tue	Sunny	19.67	28	96
Wed	Sunny	17.51	16	20
Thu	Cloudy	14.44	11	22
Fri	Shower	10.51	26	79
Sat	Shower	11.07	27	62
Sun	Sunny	17.5	20	10

Day 03(week 4) – 21th September 2023

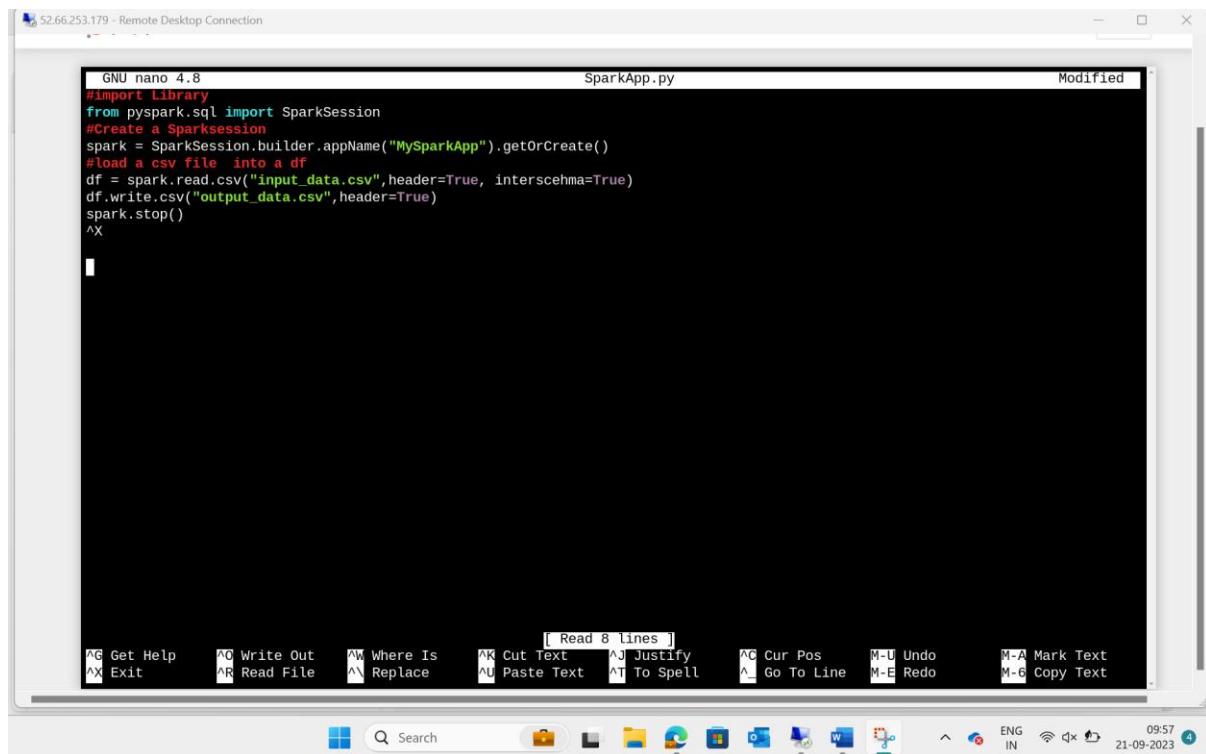
The screenshot shows a terminal window on a remote desktop connection. The user is navigating through a directory structure and creating files:

```
(base) labuser@ip-172-31-12-0:~$ cd Desktop
(base) labuser@ip-172-31-12-0:~/Desktop$ ls
20Sept-PySpark  21Sept-Interspark  Readme  Terminal.desktop  firefox.desktop  google-chrome.desktop  jupyter-notebook.desktop
(base) labuser@ip-172-31-12-0:~/Desktop$ cd 21Sept-Interspark
bash: cd: 21Sept-Interspark: No such file or directory
(base) labuser@ip-172-31-12-0:~/Desktop$ cd 21Sept-Interspark
(base) labuser@ip-172-31-12-0:~/Desktop$ mkdir SparkApp
(base) labuser@ip-172-31-12-0:~/Desktop$ cd SparkApp/
01.ReadWriteData.ipynb  SparkApp
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark$ cd SparkApp/
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark$ cat > SparkApp.py
```



```
(base) labuser@ip-172-31-12-0:~$ cd Desktop
(base) labuser@ip-172-31-12-0:~/Desktop$ ls
20Sept-PySpark 21Sept-Interspark  Readme  Terminal.desktop  firefox.desktop  google-chrome.desktop  jupyter-notebook.desktop
(base) labuser@ip-172-31-12-0:~/Desktop$ cd 21Sept-INTerspark
bash: cd: 21Sept-INTerspark: No such file or directory
(base) labuser@ip-172-31-12-0:~/Desktop$ cd 21Sept-Interspark
(base) labuser@ip-172-31-12-0:~/Desktop$ mkdir SparkApp
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark$ ls
01.ReadWriteData.ipynb  SparkApp
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark$ cd SparkApp/
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark/SparkApp$ cat > SparkApp.py
#import Library
from pyspark.sql import SparkSession
#Create a SparkSession
spark = SparkSession.builder.appName("MySparkApp").getOrCreate()
#load a csv file into a df
df = spark.read.csv("input_data.csv",header=True, inferSchema=True)
df.write.csv("output_data.csv",header=True)
spark.stop()
```

Can use nano SparkApp.py to write under this window as well it's the same baat



```
GNU nano 4.8
import Library
from pyspark.sql import SparkSession
#Create a SparkSession
spark = SparkSession.builder.appName("MySparkApp").getOrCreate()
#load a csv file into a df
df = spark.read.csv("input_data.csv",header=True, inferSchema=True)
df.write.csv("output_data.csv",header=True)
spark.stop()
^X
```

Ctrl Z to exit this window

```
#import Library
```

```

from pyspark.sql import SparkSession

#Create a Sparksession

spark = SparkSession.builder.appName("MySparkApp").getOrCreate()

#load a csv file into a df

df = spark.read.csv("input_data.csv",header=True, inferSchema=True)

#perform transformation(eg., filter, aggregation,etc.) on the df

#save the result to a csv file
df.write.csv("output_data.csv",header=True)

#stop the SparkSession
spark.stop()

```

```

52.66.253.179 - Remote Desktop Connection

spark = SparkSession.builder.appName("MySparkApp").getOrCreate()
#load a csv file into a df
df = spark.read.csv("input_data.csv",header=True, inferSchema=True)
df.write.csv("output_data.csv",header=True)
spark.stop()
^C
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark/SparkApp$ nano SparkApp.py

Use "fg" to return to nano.

[1]+ Stopped nano SparkApp.py
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark/SparkApp$ ls ../
01.ReadWriteData.ipynb SparkApp
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark/SparkApp$ ls ../..
20Sept-PySpark 21Sept-Interspark Readme Terminal.desktop firefox.desktop google-chrome.desktop jupyter-notebook.desktop
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark/SparkApp$ ls ../../21Sept-Interspark/
01.ReadWriteData.ipynb SparkApp
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark/SparkApp$ ls ../..
01.ReadWriteData.ipynb SparkApp
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark/SparkApp$ ls ../../21Sept-Interspark/
01.ReadWriteData.ipynb SparkApp
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark/SparkApp$ ls ../..
20Sept-PySpark 21Sept-Interspark Readme Terminal.desktop firefox.desktop google-chrome.desktop jupyter-notebook.desktop
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark/SparkApp$ ls ../../20Sept-PySpark/
01.ReadingData.ipynb temp.csv temp.csv.v1
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark/SparkApp$ ls ../../20Sept-PySpark/temp.csv input.csv/
ls: cannot access 'input.csv': No such file or directory
../20Sept-PySpark/temp.csv
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark/SparkApp$ ls ../../20Sept-PySpark/temp.csv input.csv
ls: cannot access 'input.csv': No such file or directory
../20Sept-PySpark/temp.csv
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark/SparkApp$ cp ../../20Sept-PySpark/temp.csv input.csv
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark/SparkApp$ ls
SparkApp.py input.csv
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark/SparkApp$ 

```

Ls .. /

Ls ../../

cp ../../20Sept-PySpark/temp.csv input.csv

Now we need to go back one step after ls

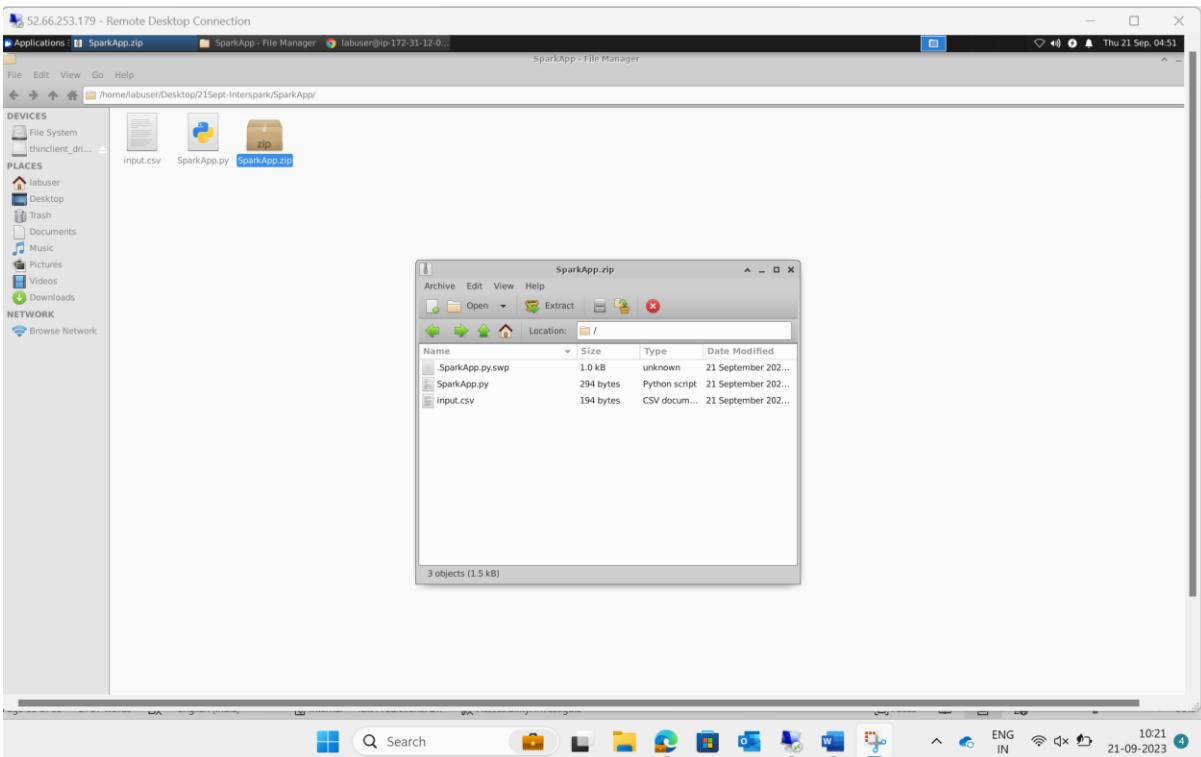
Using cd ..

Then type zip -r SparkApp.zip SparkApp/

01.ReadwriteData.ipynb SparkApp.zip

```
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark$ cd
(base) labuser@ip-172-31-12-0:~$ cd SparkApp/
bash: cd: SparkApp/: No such file or directory
(base) labuser@ip-172-31-12-0:~$ cd 21Sept-Interspark
bash: cd: 21Sept-Interspark: No such file or directory
(base) labuser@ip-172-31-12-0:~$ cd Desktop
(base) labuser@ip-172-31-12-0:~/Desktop$ cd 21Sept-Interspark
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark$ cd SparkApp
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark$ zip -r SparkApp.zip .
  adding: .SparkApp.py.swp (deflated 93%)
  adding: SparkApp.py (deflated 38%)
  adding: input.csv (deflated 28%)
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark$ ls
SparkApp.py  SparkApp.zip  input.csv
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark$
```

desktop you can see this



```
Spark-submit \
-- master spark://MasterIP:7077 \
--py-files SparkApp.zip
--files input_data.csv
```

SparkApp.py

Cp \

Source_path \

Target_path

➔ Resource allocation contains RAM,CORE and Temp storage(\temp)

spark-submit --py-files SparkApp.zip --files input_data.csv SparkApp.py

52.66.253.179 - Remote Desktop Connection

Applications : SparkApp - File Manager labuser@ip-172-31-12-0

Desktop/21Sept-Interspark x labuser@ip-172-31-12-0: ~ 01.ReadWriteData - Jupyter GitHub - NubeEra-DataEng +

localhost:8888/terminals/1

jupyter

Logout

```
SparkApp.py  SparkApp.zip  input.csv
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark/SparkApp$ spark-submit --py-files SparkApp.zip --files input_data.csv $S
parkApp.py

23/09/21 05:52:48 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes whe
re applicable
23/09/21 05:52:51 INFO SparkCon
23/09/21 05:52:51 INFO Resource
23/09/21 05:52:51 INFO Resource
23/09/21 05:52:51 INFO Resource
23/09/21 05:52:51 INFO SparkCon
23/09/21 05:52:51 INFO Resource
script: , vendor: , memory ->
dor: ), task resources: Map(cpu
23/09/21 05:52:51 INFO Resource
23/09/21 05:52:51 INFO ResourceManager: Added ResourceProfile id: 0
23/09/21 05:52:51 INFO SecurityManager: Changing view acls to: labuser
23/09/21 05:52:51 INFO SecurityManager: Changing modify acls to: labuser
23/09/21 05:52:51 INFO SecurityManager: Changing view acls groups to:
23/09/21 05:52:51 INFO SecurityManager: Changing modify acls groups to:
23/09/21 05:52:51 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: labuser; groups with view permissions: EMPTY; users with modify permissions: labuser; groups with modify permissions: EMPTY
23/09/21 05:52:53 INFO Utils: Successfully started service 'sparkDriver' on port 34213.
23/09/21 05:52:53 INFO SparkEnv: Registering MapOutputTracker
23/09/21 05:52:54 INFO SparkEnv: Registering BlockManagerMaster
23/09/21 05:52:54 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology info
rmation
23/09/21 05:52:54 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
23/09/21 05:52:54 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
23/09/21 05:52:55 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-c535d5a4-fe63-42e3-8f3c-44096a6ec802
23/09/21 05:52:55 INFO MemoryStore: MemoryStore started with capacity 366.3 MiB
```

Your Remote Desktop Services session has ended.

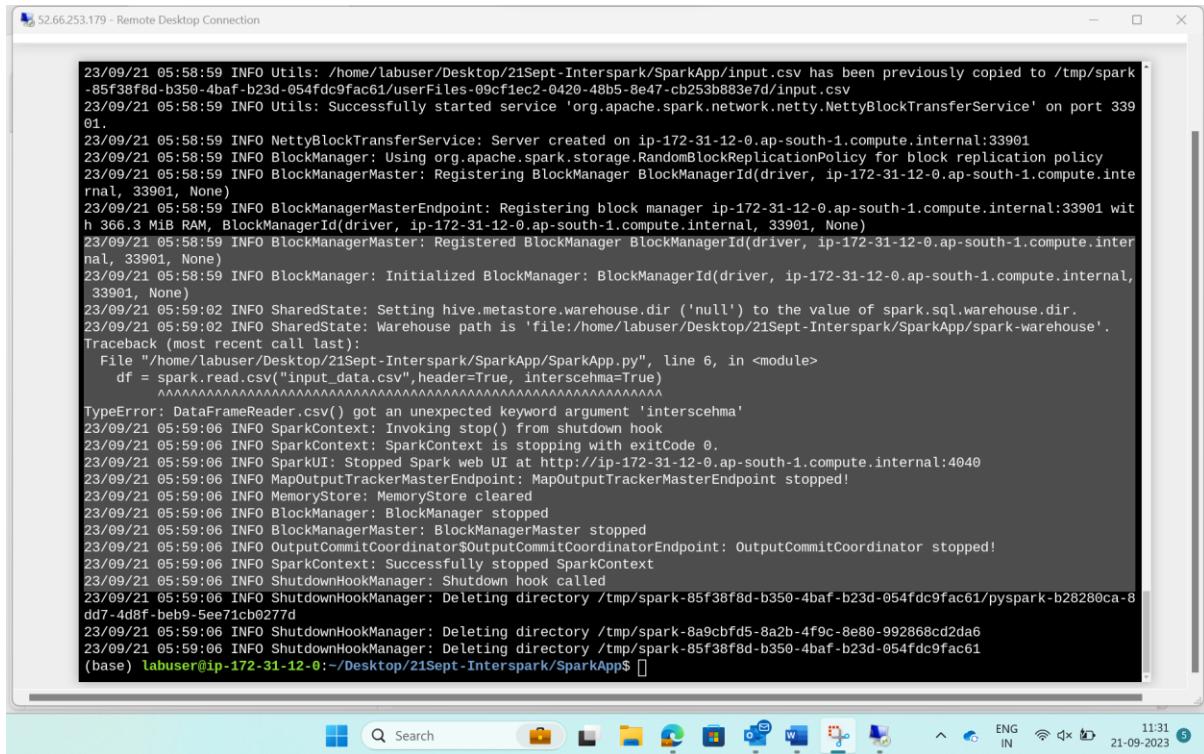
The connection to the remote computer was lost, possibly due to network connectivity problems. Try connecting to the remote computer again. If the problem continues, contact your network administrator or technical support.

OK

```
52.66.253.179 - Remote Desktop Connection
Applications : SparkApp - File Manager labuser@ip-172-31-12-0...
Desktop/21Sept-Interspark x labuser@ip-172-31-12-0- x 01.ReadWriteData - Jupyter x GitHub - NubeEra-DataEng x + 
localhost:8888/terminals/1
Logout

jupyter

(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark/SparkApp$ 
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark/SparkApp$ spark-submit --py-files SparkApp.zip --files input.csv SparkApp.py
23/09/21 05:58:54 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23/09/21 05:58:56 INFO SparkContext: Running Spark version 3.4.1
23/09/21 05:58:56 INFO ResourceUtils: =====
23/09/21 05:58:56 INFO ResourceUtils: No custom resources configured for spark.driver.
23/09/21 05:58:56 INFO ResourceUtils: =====
23/09/21 05:58:56 INFO SparkContext: Submitted application: MySparkApp
23/09/21 05:58:56 INFO ResourceProfile: DefaultResourceProfile created, executor resources: Map(cores => name: cores, amount: 1, script: , vendor: , memory -> name: memory, amount: 1024, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpu -> name: cpus, amount: 1.0)
23/09/21 05:58:56 INFO ResourceProfile: Limiting resource is cpu
23/09/21 05:58:56 INFO ResourceProfileManager: Added ResourceProfile id: 0
23/09/21 05:58:56 INFO SecurityManager: Changing view acls to: labuser
23/09/21 05:58:56 INFO SecurityManager: Changing modify acls to: labuser
23/09/21 05:58:56 INFO SecurityManager: Changing view acls groups to:
23/09/21 05:58:56 INFO SecurityManager: Changing modify acls groups to:
23/09/21 05:58:56 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: labuser; groups with view permissions: EMPTY; users with modify permissions: labuser; groups with modify permissions: EMPTY
23/09/21 05:58:57 INFO Utils: Successfully started service 'sparkDriver' on port 41387.
23/09/21 05:58:57 INFO SparkEnv: Registering MapOutputTracker
23/09/21 05:58:57 INFO SparkEnv: Registering BlockManagerMaster
23/09/21 05:58:57 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
23/09/21 05:58:57 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
23/09/21 05:58:57 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
23/09/21 05:58:57 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-d43982f7-3b9f-4c84-a1e6-fb3652a2187d
23/09/21 05:58:58 INFO MemoryStore: MemoryStore started with capacity 366.3 MB
23/09/21 05:58:58 INFO SparkEnv: Registering OutputCommitCoordinator
```



```

23/09/21 05:58:59 INFO Utils: /home/labuser/Desktop/21Sept-Interspark/SparkApp/input.csv has been previously copied to /tmp/spark-85f38f8d-b350-4ba5-b23d-054fdc9fac61/userFiles-09cf1ec2-0420-48b5-8e47-cb253b883e7d/input.csv
23/09/21 05:58:59 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 33901.
23/09/21 05:58:59 INFO NettyBlockTransferService: Server created on ip-172-31-12-0.ap-south-1.compute.internal:33901
23/09/21 05:58:59 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
23/09/21 05:58:59 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, ip-172-31-12-0.ap-south-1.compute.internal, 33901, None)
23/09/21 05:58:59 INFO BlockManagerMasterEndpoint: Registering block manager ip-172-31-12-0.ap-south-1.compute.internal:33901 with 366.3 MiB RAM, BlockManagerId(driver, ip-172-31-12-0.ap-south-1.compute.internal, 33901, None)
23/09/21 05:58:59 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, ip-172-31-12-0.ap-south-1.compute.internal, 33901, None)
23/09/21 05:58:59 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, ip-172-31-12-0.ap-south-1.compute.internal, 33901, None)
23/09/21 05:59:02 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.warehouse.dir.
23/09/21 05:59:02 INFO SharedState: Warehouse path is 'file:/home/labuser/Desktop/21Sept-Interspark/SparkApp/spark-warehouse'.
Traceback (most recent call last):
  File "/home/labuser/Desktop/21Sept-Interspark/SparkApp/SparkApp.py", line 6, in <module>
    df = spark.read.csv("input_data.csv", header=True, interschema=True)
      ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
TypeError: DataFrameReader.csv() got an unexpected keyword argument 'interschema'
23/09/21 05:59:06 INFO SparkContext: Invoking stop() from shutdown hook
23/09/21 05:59:06 INFO SparkContext: SparkContext is stopping with exitCode 0.
23/09/21 05:59:06 INFO SparkUI: Stopped Spark web UI at http://ip-172-31-12-0.ap-south-1.compute.internal:4040
23/09/21 05:59:06 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
23/09/21 05:59:06 INFO MemoryStore: MemoryStore cleared
23/09/21 05:59:06 INFO BlockManager: BlockManager stopped
23/09/21 05:59:06 INFO BlockManagerMaster: BlockManagerMaster stopped
23/09/21 05:59:06 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
23/09/21 05:59:06 INFO SparkContext: Successfully stopped SparkContext
23/09/21 05:59:06 INFO ShutdownHookManager: Shutdown hook called
23/09/21 05:59:06 INFO ShutdownHookManager: Deleting directory /tmp/spark-85f38f8d-b350-4ba5-b23d-054fdc9fac61/pyspark-b28280ca-8dd7-4d0f-beb9-5ee71cb0277d
23/09/21 05:59:06 INFO ShutdownHookManager: Deleting directory /tmp/spark-8a9cbfd5-8a2b-4f9c-8e80-992868cd2da6
23/09/21 05:59:06 INFO ShutdownHookManager: Deleting directory /tmp/spark-85f38f8d-b350-4ba5-b23d-054fdc9fac61
(base) labuser@ip-172-31-12-0:~/Desktop/21Sept-Interspark/SparkApp$ 
```

Task 1- filter csv file

<https://saturncloud.io/blog/how-to-remove-rows-in-a-spark-dataframe-based-on-position-a-comprehensive-guide/>

```

from pyspark.sql.functions import monotonically_increasing_id

df = df.withColumn('index', monotonically_increasing_id())

rows_to_remove = [0,1,2] } for all rows type out

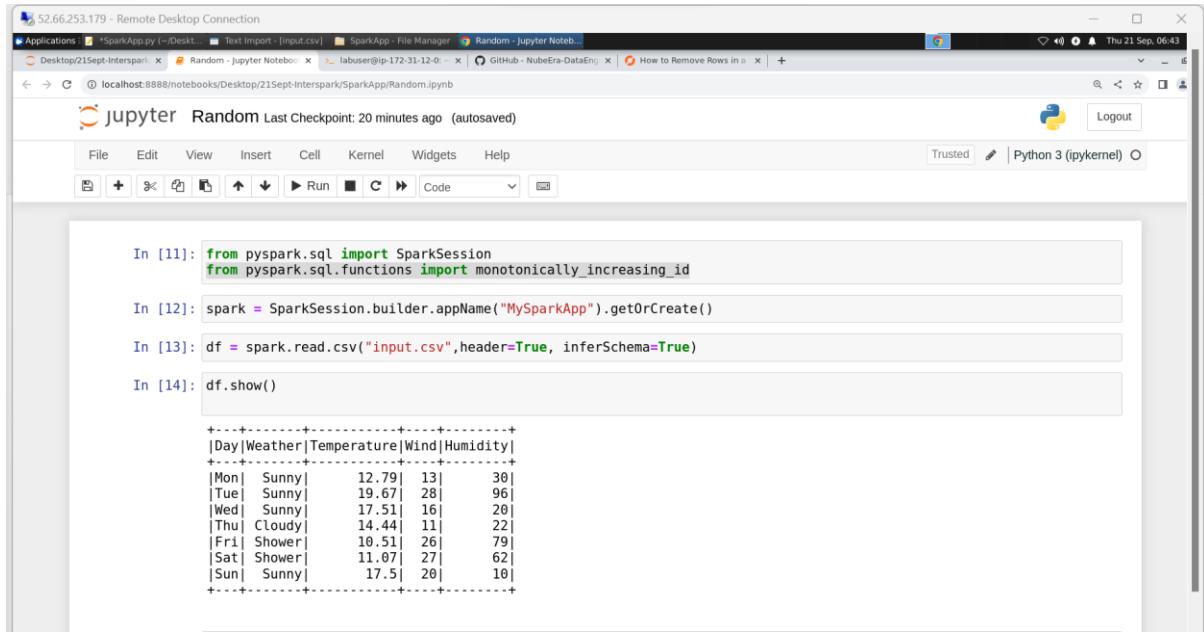
df = df.filter(~df.index.isin(rows_to_remove))

df = df.drop('index')

df.show()

```

temp way to remove a row in pyspark



In [11]:

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import monotonically_increasing_id
```

In [12]:

```
spark = SparkSession.builder.appName("MySparkApp").getOrCreate()
```

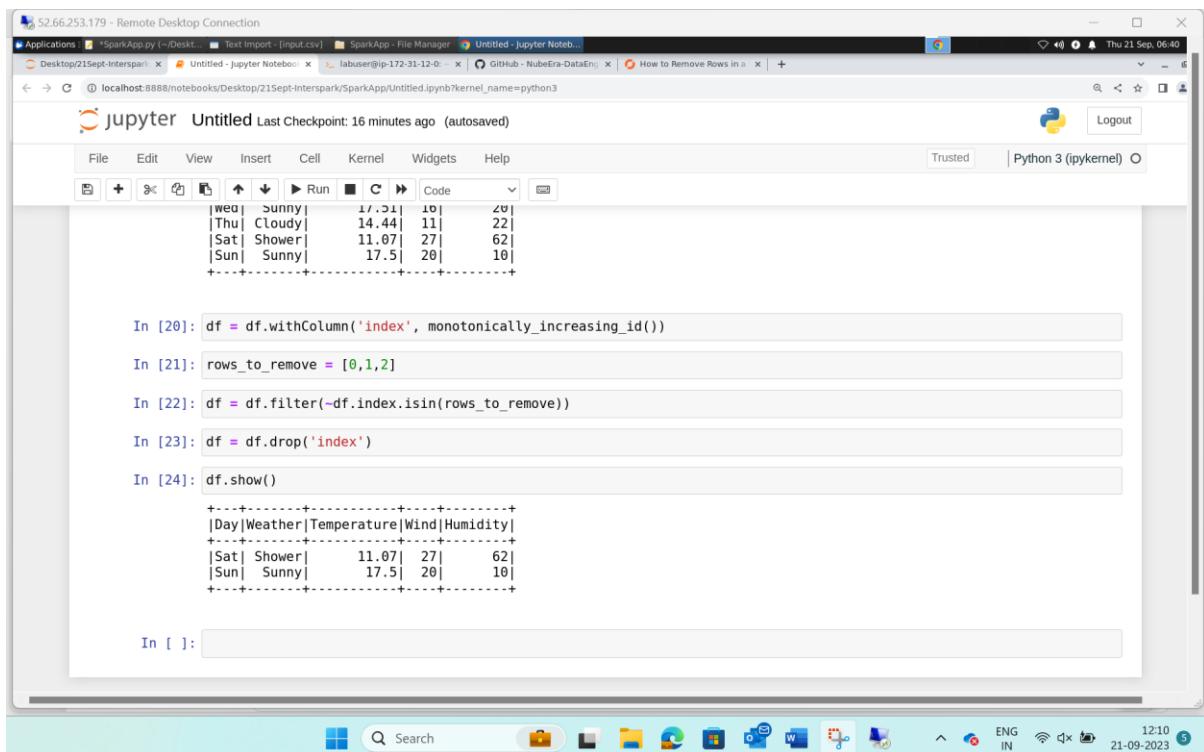
In [13]:

```
df = spark.read.csv("input.csv", header=True, inferSchema=True)
```

In [14]:

```
df.show()
```

	Day	Weather	Temperature	Wind	Humidity
[Mon]	Sunny	12.79	13	30	
[Tue]	Sunny	19.67	28	96	
[Wed]	Sunny	17.51	16	20	
[Thu]	Cloudy	14.44	11	22	
[Fri]	Shower	10.51	26	79	
[Sat]	Shower	11.07	27	62	
[Sun]	Sunny	17.5	20	10	



In [20]:

```
df = df.withColumn('index', monotonically_increasing_id())
```

In [21]:

```
rows_to_remove = [0,1,2]
```

In [22]:

```
df = df.filter(~df.index.isin(rows_to_remove))
```

In [23]:

```
df = df.drop('index')
```

In [24]:

```
df.show()
```

	Day	Weather	Temperature	Wind	Humidity
[Sat]	Shower	11.07	27	62	
[Sun]	Sunny	17.5	20	10	

In []:

Task 2- convert csv to parquet file

<https://mungingdata.com/python/writing-parquet-pandas-pyspark-koalas/>

3.110.154.16 - Remote Desktop Connection

File Edit View Search Terminal Help labuser@ip-172-31-12-0: ~

```
(base) labuser@ip-172-31-12-0:~$ -s sudo su
(sudo) labuser@ip-172-31-12-0:~$ sudo su
root@ip-172-31-12-0:/home/labuser# virtualenv --version
Command 'virtualenv' not found, but can be installed with:
apt install python3-virtualenv
root@ip-172-31-12-0:/home/labuser# virtualenv
Command 'virtualenv' not found, but can be installed with:
apt install python3-virtualenv
root@ip-172-31-12-0:/home/labuser# apt install python3-virtualenv
Reading package lists... Done
Building dependency tree...
Reading state information... Done
The following packages were automatically installed and are no longer required:
  libfwupdplugin1 libxenial linux-aws-5.11-headers-5.11.0-1028 linux-image-5.11.0-1028-aws linux-modules-5.11.0-1028-aws
Use 'sudo apt autoremove' to remove them.
The following additional packages will be installed:
  python3-apipkg python3-distlib python3-filelock
The following NEW packages will be installed:
  python3-apipkg python3-distlib python3-virtualenv
0 upgraded, 0 newly installed, 0 to remove and 317 not upgraded.
Need to get 197 kB of archives.
After this operation, 1032 kB of additional disk space will be used.
Do you want to continue? [Y/n]
Get:1 http://ap-south-1.ec2.archive.ubuntu.com/ubuntu focal/main amd64 python3-apipkg all 1.4.3-2.1 [10.8 kB]
Get:2 http://ap-south-1.ec2.archive.ubuntu.com/ubuntu focal/universe amd64 python3-distlib all 0.3.0-1 [116 kB]
Get:3 http://ap-south-1.ec2.archive.ubuntu.com/ubuntu focal/universe amd64 python3-filelock all 3.0.12-2 [7948 B]
Get:4 http://ap-south-1.ec2.archive.ubuntu.com/ubuntu focal/universe amd64 python3-virtualenv all 20.0.17-1ubuntu0.4 [62.7 kB]
Fetched 197 kB in 0s (631 kB/s)
Selecting previously unselected package python3-apipkg.
(Reading database ... 243768 files and directories currently installed.)
Preparing to unpack .../python3-apipkg_1.4.3-2.1_all.deb ...
Unpacking python3-apipkg (1.4.3-2.1) ...
Selecting previously unselected package python3-distlib.
Preparing to unpack .../python3-distlib_0.3.0-1_all.deb ...
Unpacking python3-distlib (0.3.0-1) ...
Selecting previously unselected package python3-filelock.
Preparing to unpack .../python3-filelock_3.0.12-2_all.deb ...
Unpacking python3-filelock (3.0.12-2) ...
Selecting previously unselected package python3-virtualenv.
Preparing to unpack .../python3-virtualenv_20.0.17-1ubuntu0.4_all.deb ...
Unpacking python3-virtualenv (20.0.17-1ubuntu0.4) ...
Setting up python3-filelock (3.0.12-2) ...
Setting up python3-distlib (0.3.0-1) ...
Setting up python3-apipkg (1.4.3-2.1) ...
Setting up python3-virtualenv (20.0.17-1ubuntu0.4) ...
Processing triggers for man-db (2.9.1-1) ...
root@ip-172-31-12-0:/home/labuser# exit
exit
(base) labuser@ip-172-31-12-0:~
```

3.110.154.16 - Remote Desktop Connection

File Edit View Search Terminal Help labuser@ip-172-31-12-0: ~/pl.py

```
#!/usr/bin/env python3
from pyspark.sql import SparkSession
#Create a SparkSession
spark = SparkSession.builder.appName("RDMapExample").getOrCreate()
#Create a RDD with a list of integers
data = [1,2,3,4,5]
rdd = spark.sparkContext.parallelize(data)
#Define a function to double each element
def doubler(x):
    return x * 2
#Use the map transformation to double each element in the RDD
result_rdd = rdd.map(doubler)
#Collect and print the result
result = result_rdd.collect()
print(result)
#Stop the SparkSession
spark.stop()
```

```

3.110.154.16 - Remote Desktop Connection
After this operation, 1022 kB of additional disk space will be used.
Do you want to continue? [Y/n] y
Get:1 http://ap-south-1.ec2.archive.ubuntu.com/ubuntu focal/main amd64 python3-appdirs all 1.4.3-2.1 [10.8 kB]
Get:2 http://ap-south-1.ec2.archive.ubuntu.com/ubuntu focal/universe amd64 python3-distlib all 0.3.0-1 [116 kB]
Get:3 http://ap-south-1.ec2.archive.ubuntu.com/ubuntu focal/universe amd64 python3-filelock all 3.0.12-2 [7948 kB]
Get:4 http://ap-south-1.ec2.archive.ubuntu.com/ubuntu focal-updates/universe amd64 python3-virtualenv all 20.0.17-1ubuntu0.4 [62.7 kB]
Fetched 197 kB in 0s (6314 kB/s)
Selecting previously unselected package python3-appdirs.
Preparing to unpack .../python3-appdirs_1.4.3-2.1_all.deb ...
Unpacking python3-appdirs (1.4.3-2.1) ...
Selecting previously unselected package python3-distlib.
Preparing to unpack .../python3-distlib_0.3.0-1_all.deb ...
Unpacking python3-distlib (0.3.0-1) ...
Selecting previously unselected package python3-filelock.
Preparing to unpack .../python3-filelock_3.0.12-2_all.deb ...
Unpacking python3-filelock (3.0.12-2) ...
Selecting previously unselected package python3-virtualenv.
Preparing to unpack .../python3-virtualenv_20.0.17-1ubuntu0.4_all.deb ...
Unpacking python3-virtualenv (20.0.17-1ubuntu0.4) ...
Setting up python3-distlib (3.0.12-2) ...
Setting up python3-appdirs (1.4.3-2.1) ...
Setting up python3-virtualenv (20.0.17-1ubuntu0.4) ...
Processing triggers for man-db (2.9.1-1) ...
root@ip-172-31-12-0:/home/labuser# exit
exit
(base) labuser@ip-172-31-12-0:~$ nano pl.py
Use 'fg' to return to nano.

[1]+  Stopped                  nano pl.py
(base) labuser@ip-172-31-12-0:~$ nano pl.py
Use 'fg' to return to nano.

[2]+  Stopped                  nano pl.py
(base) labuser@ip-172-31-12-0:~$ fg
nano pl.py
(base) labuser@ip-172-31-12-0:~$ fg
nano pl.py
(base) labuser@ip-172-31-12-0:~$ fg
nano pl.py
(base) labuser@ip-172-31-12-0:~$ nano pl.py
(base) labuser@ip-172-31-12-0:~$ █

```

RDD Operators:

- Actions: Collects, first, reduce, count, SaveAs Text
Actions are operations on RDD that triggers the execution of spark computation.
They force evaluation of the transformations and produce a result(perform an action)
They are used to retrieve results, save data and trigger side effects in sparks.
- Transformation: map, filter, reduceByKey, Join
they are operations on RDD that create new RDD result.
They are lazy evaluated.
they are used to build computational pipelines and define data pipeline sequences.

Why RDD:

Best performance Optimization

Day 04(week 4) – 22nd September 2023

Spark Context is the main entry point and consists of all basic functionality of spark.

- Spark driver contains DAG scheduler and task Scheduler and block manager.
- The main purpose of which are responsible for ... user written codes on jobs that are executed upon cluster.

Use analytics Vidhya to learn about DAG scheduler, driver program etc.

Create a new file in a new folder named rdd and create a spark context

```
In [1]: from pyspark import SparkContext  
# pip install pyspark  
  
In [2]: #create a Sparkcontext.  
sc= SparkContext("local","SparkPractice")  
  
Setting default log level to "WARN".  
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).  
23/09/22 04:17:25 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j  
ava classes where applicable  
  
In [4]: data=[1,2,3,4,5]  
rdd= sc.parallelize(data)  
  
In [5]: result= rdd.collect()  
print(result)  
  
[Stage 0:> (0 + 1) / 1]  
[1, 2, 3, 4, 5]
```

```
[Stage 0:> (0 + 1) / 1]  
[1, 2, 3, 4, 5]  
  
In [6]: def f(x): print(x)  
  
In [10]: name= sc.parallelize(['Navya','Anushka','Anupa','Bhawana','Garvika','Amit','Adi'])  
  
In [11]: name.collect()  
Out[11]: ['Navya', 'Anushka', 'Anupa', 'Bhawana', 'Garvika', 'Amit', 'Adi']  
  
In [16]: num= sc.parallelize([1,2,5,7],9)  
  
In [17]: num.getNumPartitions()  
Out[17]: 9  
  
In [18]: num.collect()  
Out[18]: [1, 2, 5, 7]
```

To get partition use `getNumPartition`

In case `findspark` doesn't work in terminal or vm type

65.2.74.180 - Remote Desktop Connection

Applications labuser@ip-172-31-12-0... RDD-Operation - Jupyter...
Desktop/22Sept-RDD/ RDD-Operation - Jupyter N... get specific row from spa... +

localhost:8888/notebooks/Desktop/22Sept-RDD/RDD-Operation.ipynb

jupyter RDD-Operation Last Checkpoint: an hour ago (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help Logout Trusted Kernel O

In [18]: num.collect()

Out[18]: [1, 2, 5, 7]

In [21]: !pip install pyspark

```
Collecting pyspark
  Using cached pyspark-3.4.1.tar.gz (310.8 MB)
  Collecting py4j==0.10.9.7
    Using cached py4j-0.10.9.7-py2.py3-none-any.whl (200 kB)
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
    Created wheel for pyspark: filename=pyspark-3.4.1-py2.py3-none-any.whl size=311285411 sha256=dd5848601bc144891f2
ab5be35d56053aebfb380abc6
  Stored in directory: /tmp/pip-ephem-wheel-cache-1qkzv1tj/pip-wheel-2.8.1
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9.7 pyspark-3.4.1
```

In [28]: !pip install Findspark

Requirement already satisfied: Findspark in /opt/anaconda3/lib/python3.11/site-packages (2.0.1)

In [31]: import findspark

In []:

11:14 Fri 22 Sep, 05:44

311285411 sha256=dd5848601bc144891f2
4f0ff48e86fa7d8c2a814f9df5dc21d79b7e

J/site-packages (2.0.1)

Out[18]: [1, 2, 5, 7]

In [21]: !pip install pyspark

```
Collecting pyspark
  Using cached pyspark-3.4.1.tar.gz (310.8 MB)
  Collecting py4j==0.10.9.7
    Using cached py4j-0.10.9.7-py2.py3-none-any.whl (200 kB)
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
    Created wheel for pyspark: filename=pyspark-3.4.1-py2.py3-none-any.whl size=311285411 sha256=dd5848601bc144891f2
ab5be35d56053aebfb380abc60b15ff0200389dbed83f
  Stored in directory: /home/labuser/.cache/pip/wheels/07/9f/04/fc2c478c8c87334f0ff48e86fa7d8c2a814f9df5dc21d79b7e
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9.7 pyspark-3.4.1
```

In [28]: !pip install Findspark

Requirement already satisfied: Findspark in /home/labuser/.local/lib/python3.8/site-packages (2.0.1)

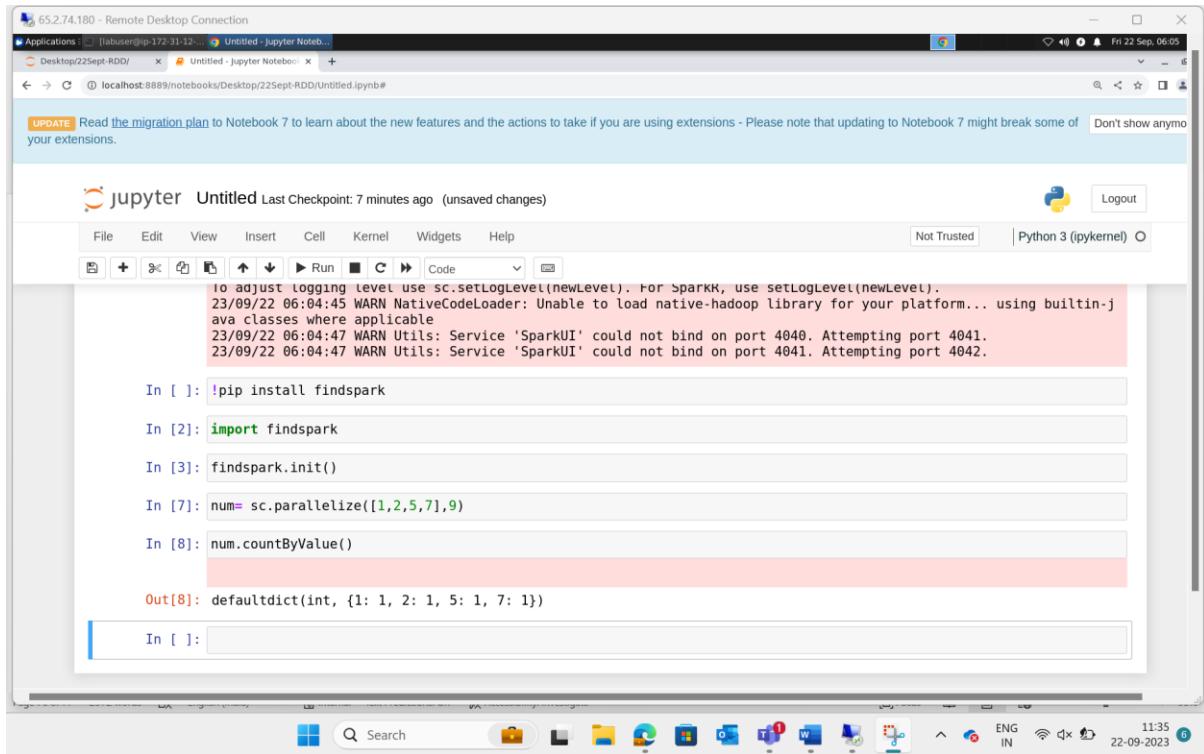
In [31]: import findspark

In []:

11:14 Fri 22 Sep, 05:44

311285411 sha256=dd5848601bc144891f2
4f0ff48e86fa7d8c2a814f9df5dc21d79b7e

J/site-packages (2.0.1)



After find spark is installed in main terminal make sure to clear all previous outputs and after importing findspark in program, type findspark.init()

Then the code will work.

After creating schema we will now create spark RDD

```
In [3]: findspark.init()

In [7]: num=sc.parallelize([1,2,5,7],9)

In [8]: num.countByValue()

Out[8]: defaultdict(int, {1: 1, 2: 1, 5: 1, 7: 1})

In [11]: from pyspark.sql.types import StructType, StructField, IntegerType, StringType

In [12]: data=[("Anushka",22),("Navya",21),("Anupa",21),("Garvika",21),("Bhawana",22),("Adi",16)]
    schema = StructType([
        StructField("name",StringType(),True),
        StructField("age",IntegerType(),True)
    ])

In [13]: from pyspark.sql import SparkSession

In [16]: sc.stop()

In [17]: spc=SparkContext("local","SparkPractice")
23/09/22 06:17:01 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
23/09/22 06:17:01 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
```

```
Schema: StructType(
    StructField("name",StringType(),True),
    StructField("age",IntegerType(),True)
)

In [13]: from pyspark.sql import SparkSession

In [16]: sc.stop()

In [17]: spc=SparkContext("local","SparkPractice")
23/09/22 06:17:01 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
23/09/22 06:17:01 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.

In [18]: #creating the spark object
spark=SparkSession.builder.appName("A").getOrCreate()

In [19]: #now creating spark rdd
rdd=spark.sparkContext.parallelize(data)

In [ ]:
```

Difference between dataframe and RDD

```
rdd= spark.sparkContext.parallelize(data)
In [20]: #to read value from rdd use collect
rdd.collect()
Out[20]: [('Anushka', 22),
 ('Navya', 21),
 ('Anupa', 21),
 ('Garvika', 21),
 ('BHawana', 22),
 ('Adi', 16)]
```

In []:

```
In [22]: df=spark.createDataFrame(rdd,schema=schema)
In [23]: df.show()
In [24]: adf=df.filter(df.age>=22)
In [25]: adf.show()
```

name	age
Anushka	22
Navya	21
Anupa	21
Garvika	21
BHawana	22
Adi	16

name	age
Anushka	22
BHawana	22

After that can write spark.stop()

Task:

SparkSession-> RDD-> DF->TempView(SQL)->Select

The screenshot shows a Jupyter Notebook interface running on a remote desktop connection. The notebook has a single open cell with the following Python code:

```
    "useActionBanner":false,"showPublishStackBanner":false}, "renderImageOrRaw":false, "richText":null, "renderedFileInfo":null, "shortPath":null, "tabSize":8, "topBannersInfo":{ "overridingGlobalFundingFile":false, "globalPrefe rredFundingPath":null, "repoOwner":"NubeEra-DataEngineering", "repoName":"DataSets", "showInvalidCitationWarning":false, "citationHelpUrl":"https://docs.github.com/en/github/creating-cloning-and-archiving-repositories/creati ng-a-repository-on-github/about-citation-files", "showDependabotConfigurationBanner":false, "actionsOnboardingTi p":null}, "truncated":false, "viewable":true, "workflowRedirectUrl":null, "symbols":{ "timedOut":false, "notAnalyzed":true, "symbols":[]}}, "copilotInfo":null, "csrf_tokens":{"/NubeEra-DataEngineering/DataSets/branches":{ "pos t": "U-kirgZkI0ylpuis54VMHUUG63yJChd6zWLq5_J2Il9_SNwansiPYYsErbsk89rEpMzNa7DUgaA2ws2Rklw"}, "/repos/preference s":{ "post": "d0beRchTlzmup89yph7dGY1FK2I8LE6M1JUR9Fbxw_K2Nd1K0t_ogh3LrmQMs-K_bBF_Mw_l70YbLmc26A"}}, "titl e": "DataSets/movies.csv at ed75cd15af374a9731e0af1d7fdec94697f95d1 · NubeEra-DataEngineering/DataSets"}  
Expected: "contentType":"file"}3 but found: "contentType":"file"}  
CSV file: file:///home/labuser/Desktop/22Sept-RDD/movies.csv
```

The output of the cell is empty: [].

The next cell (In [90]) contains the following code:

```
from pyspark import SparkContext,SparkConf  
conf= SparkConf().setAppName("App1").setMaster("local")  
sp= SparkContext(conf=conf)  
num = sp.parallelize([5,5,4,8,9,3,2])  
num.collect()
```

The output of this cell shows two warning messages:

```
23/09/22 08:33:18 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.  
23/09/22 08:33:18 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
```

The final output of the cell is [5, 5, 4, 8, 9, 3, 2].

We need to stop spark before executing this

The screenshot shows a Jupyter Notebook interface running on a remote desktop connection. The notebook has several cells with the following Python code:

```
23/09/22 08:33:18 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.  
Out[90]: [5, 5, 4, 8, 9, 3, 2]  
  
In [91]: num.map(lambda a: a*2).collect()  
  
Out[91]: [10, 10, 8, 16, 18, 6, 4]  
  
In [93]: names= sp.parallelize(["Navya", "Bhawana", "Anushka"])  
  
In [95]: rdd= sp.parallelize([2,3,4])  
rdd.collect()  
  
Out[95]: [2, 3, 4]  
  
In [98]: a = range(1,3)  
for i in a:  
    print(i)
```

The output of the last cell shows the numbers 1 and 2.

The screenshot shows a Jupyter Notebook interface running on a remote desktop connection. The notebook is titled "RDD Operations2" and has a last checkpoint from 3 hours ago. The Python kernel version is Python 3 (ipykernel). The session contains the following code and output:

```
In [99]: (rdd.flatMap(lambda x: range(1,x))).collect()
Out[99]: [1, 1, 2, 1, 2, 3]

In [100]: numbers= sp.parallelize([1,2,3,4,5])
In [110]: def generate_squares(numbers):
    return [numbers, numbers **2 ]

In [111]: squares_rdd =numbers.flatMap(generate_squares)
squares_rdd.collect()
Out[111]: [1, 1, 2, 4, 3, 9, 4, 16, 5, 25]

In [113]: even_result=squares_rdd.filter(lambda x: x%2 == 0)
even_result.collect()
Out[113]: [2, 4, 4, 16]

In [108]: odd_result=sp.parallelize([1,3,5])
odd_result.collect()
Out[108]: [1, 3, 5]

In [ ]:
```



```
Out[108]: [1, 3, 5]

In [116]: num=sp.parallelize([1,4,5,6,2,8,7],3)
num.collect()
Out[116]: [1, 4, 5, 6, 2, 8, 7]

In [117]: num.glom().collect()
Out[117]: [[1, 4], [5, 6], [2, 8, 7]]

In [ ]:
```

Day 01(week 5) – 25th September 2023

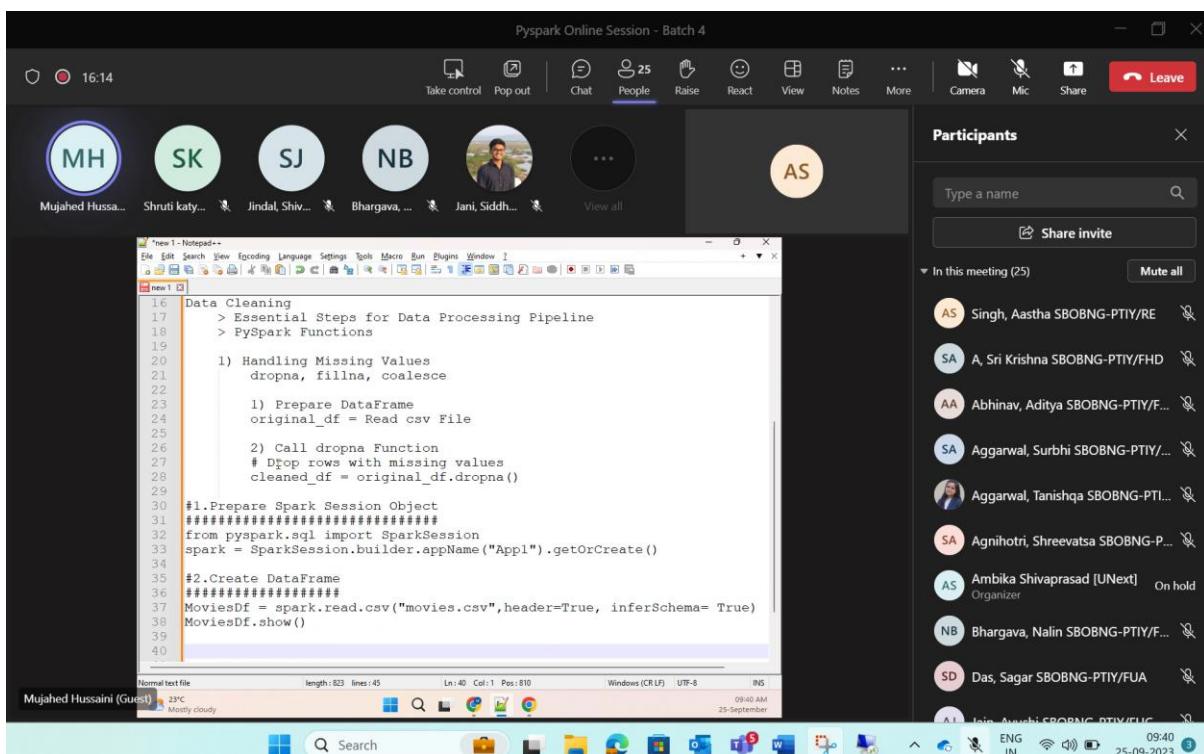
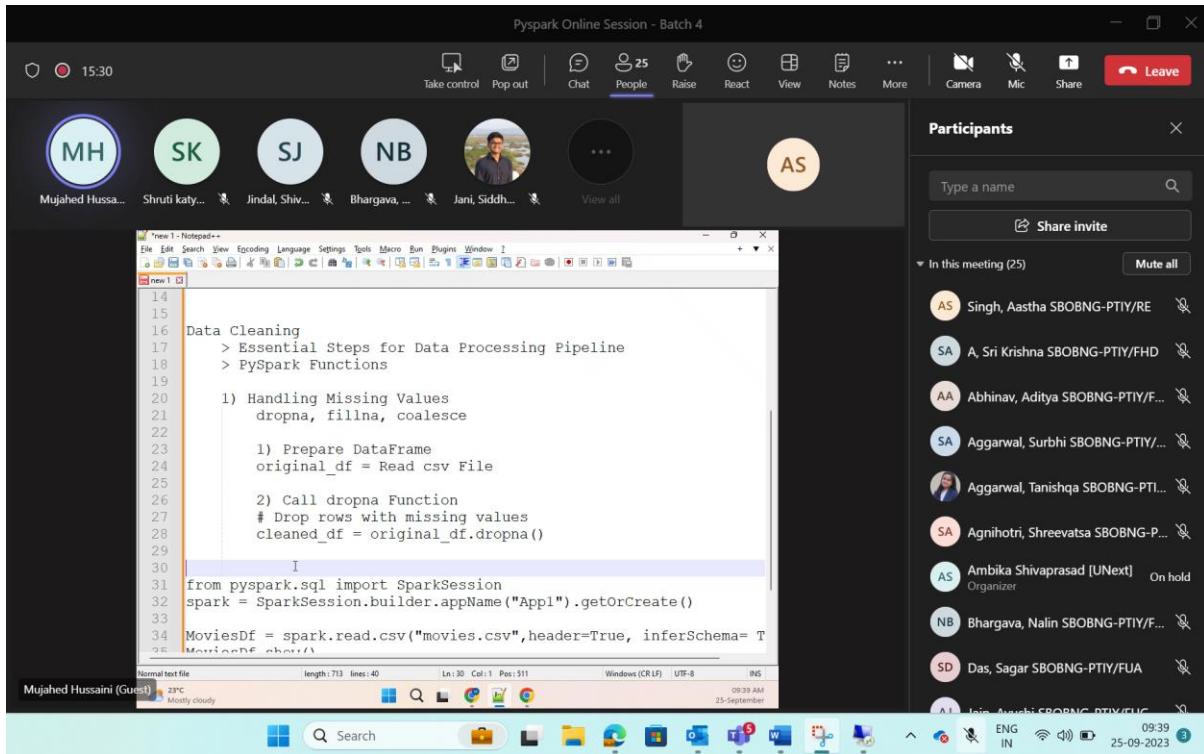
A screenshot of a Jupyter Notebook interface. The top bar shows the URL `wai.nuvepro.com/guacamole/#/client/aS0wYzBiNWQxZWVmMWJkNWl4NQbjAG51dmVs...` and the date/time `Mon 25 Sep, 04:07 labuser`. The notebook tab is titled `Untitled - Jupyter Notebo...`. Below the tabs, there's a toolbar with File, Edit, View, Insert, Cell, Kernel, Widgets, Help, Run, and Code buttons. A status bar indicates `Trusted | Python 3 (ipykernel) |`. The main area contains four code cells:

```
In [4]: import findspark  
findspark.init()  
  
In [7]: from pyspark.sql import SparkSession  
spark = SparkSession.builder.appName("App1").getOrCreate()  
  
In [*]: MoviesDF = spark.read.csv("movies.csv",header=True, inferSchema= True)  
MoviesDF.show()  
  
In [ ]:
```

A screenshot of a video conference interface titled "Pyspark Online Session - Batch 4". The top bar shows the date/time `Mon 25 Sep, 04:07 labuser` and a list of participants: MH, SK, SJ, NB, VY, AS, and others. The main window shows a Jupyter Notebook with the same code as the previous screenshot. The output of the `show()` command is displayed as a table:

	Film	Genre	Lead Studio	Audience score	% Profitability	Rotten Tomatoes	% Worldwide Gross Year
64	[Zack and Miri Meet ...]	[Romance]	[The Weinstein Com...]	70	1.747541667	74	[2008]
68	[Youth in Revolt]	[Comedy]	[The Weinstein Com...]	52	1.09	\$19.62	[2010]
43	[You Will Meet a T...	[Comedy]	[Independent]	35	1.211818182	\$26	[2010]
1	[What's in a Name]	[Comedy]	[Disney]	44	0.0		

The bottom of the screen shows a taskbar with various icons and system status indicators.



Pyspark Online Session - Batch 4

01:16:21 Take control Pop out Chat People Raise React View Notes More Camera Mic Share Leave

Jani, Siddh... Ambika Shi... Mujahed H... Narayan, A... Samal, Alya... View all AS MH AS *** AS

Participants

Type a name Share invite

In this meeting (28)

- AS Singh, Aastha SBOBNG-PTIY/RE
- SA A, Sri Krishna SBOBNG-PTIY/FHD
- SA Aggarwal, Surbhi SBOBNG-PTIY/...
- Aggarwal, Tanishqa SBOBNG-PTI...
- SA Agnihotri, Shreevatsa SBOBNG-P...
- AS Ambika Shivaprasad [UNext] Organizer
- MA Ananthakumar, Manav SBOBNG-...
- NB Bhargava, Nalin SBOBNG-PTIY/F...
- RC Chail, Ritik SBOBNG-PTIY/FUF
- SD Dose, Saara SBOBNG-PTIY/ELA

Notebook

New 1 - Notepad++

```
new 1 | Encoding: Language: Settings: Tools: Macro: Run: Plugins: Window: Help
```

```
47
48
49
50
51 year
52
53     if Null Value in year column then
54         Add Default year(1990)
55     else
56         keep same
57
58
59 from pyspark.sql.functions import when
60 YearMoviesDF = filledCleanedMoviesDF
61     .withColumn(
62         "year",
63         when(filledCleanedMoviesDF["year"].isNull(),1990).otherwise(filledCleanedMoviesDF["year"]))
64
65
66 from pyspark.sql.functions import when
67 YearMoviesDF = filledCleanedMoviesDF.withColumn("year" ,when(filledCleanedMoviesDF["year"].isNull()),1990)
68
69
70
71
72
73
74
75
76
77
78
```

Normal text file length : 1,571 lines : 78 Ln : 59 Col : 1 Sel : 202/16 Windows (CR/LF) UTF-8 INS

10:40 AM 25-September-2023

Mujahed Hussaini (Guest) Sunc mostly cloudy

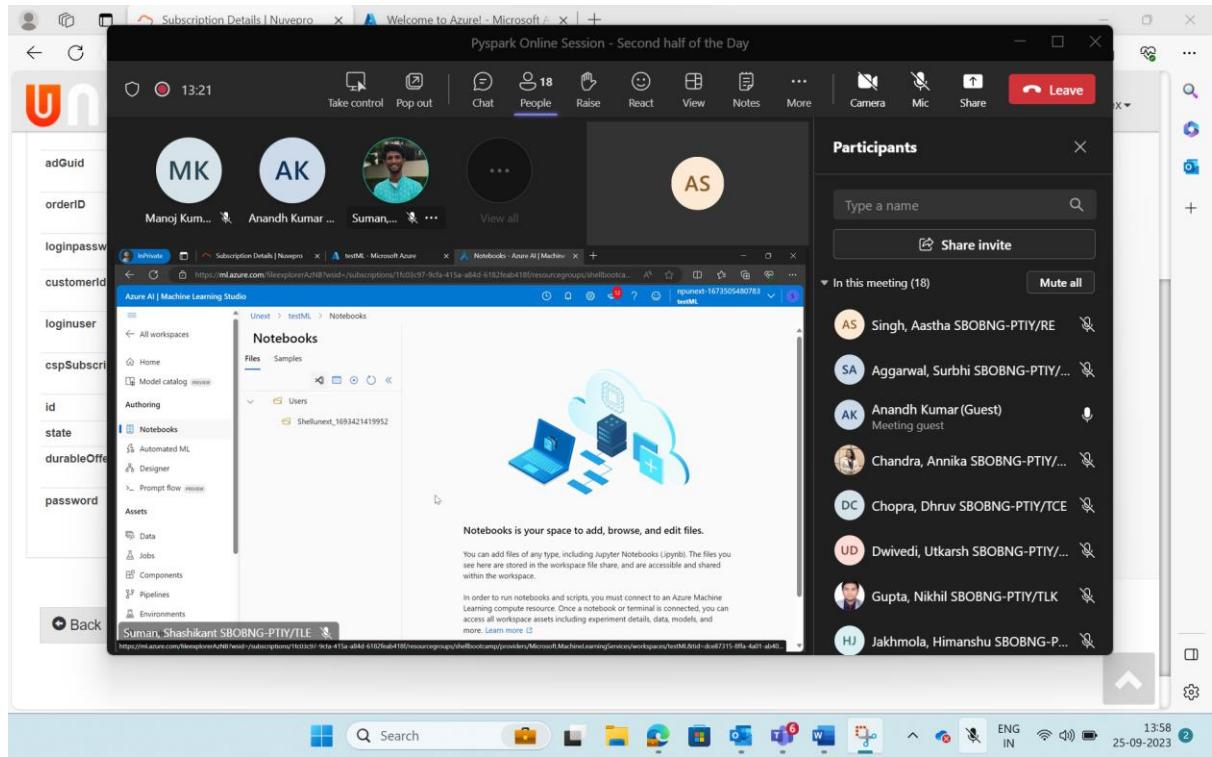
Search

10:40 25-09-2023 3

The screenshot shows a Notepad++ window titled "new 1 - Notepad++". The code in the editor is as follows:

```
112 )
113
114
115 isnan(c) --> True/False:
116     It checks if the column contains
117     NaN (Not-a-Number) values.
118     This is relevant for numeric columns
119
120 col(c).isNull()
121     It checks if the column contains null values.
122     This is relevant for all types(string) of columns.
123
124 # Handle a null value
125 # Replace null values in the director_name column
126 # with Unknown
127 df.na.fill("Unknown", subset = ["director_name"])
128
129
```

The status bar at the bottom displays: Normal text file, length : 2,804 lines : 129, Ln : 128 Col : 1 Pos : 2,803, Windows (CR LF), UTF-8, INS.



Inbox (1) anandh@nuvero.com Case study update... Code Green... Join conversation Classnotes Ge... Spark Hands-on Subscription Details Home Microsoft New tab

docs.google.com/document/d/1ZphBmqbaMzkUm_lqvdIm30dy2CbAYF39Kigwgf8/edit

Spark Hands-on Saving...

File Edit View Insert Format Tools Extensions Help

Share

Normal text Arial 100% 20 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Know Your Data

1. File format
2. Size
3. Counts
4. Ocolumns
5. Delimiter
6. sensitive/critical
7. fr̄equently/in-frequent
8. Data types -- > int, string, double, date, boolean ⇒ arra|

Spark Hands-on

File Edit View Insert Format Tools Extensions Help

Normal text Arial 20 B I U A G S E

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

9. Collection of files/folders

10 Data dictionary

Data ingestion : CSV → header , inferSchema , delimiter/sep ,

Mode → permissive , dropMalformed , failFast

|

I

Data preparation :

In

Spark Hands-on

File Edit View Insert Format Tools Extensions Help

Normal text Arial 20 B I U A G S E

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Data transformation :

Join

Sort

Filter

Null values→ dropna , fillna , replace
Distinct or dropDuplicates()

Deptcode → withColumn (lit ("T & S "))

Cache ⇒

I

Partition ⇒ parallelism

Data sink → table , storage optimization
(parquet/orc)

In

Know Your Data

1. File format
2. Size
3. Counts

4. **Columns**
5. **Delimiter**
6. **sensitive/critical**
7. **frequently/in-frequent**
8. **Data types -- > int,string,double, date, boolean ⇒ array, struct, map**
9. **Collection of files/folders**

10 Data dictionary

Data ingestion : CSV → header , inferschema , delimiter/sep ,

Recommend : user defined schema

Mode → permissive , dropmalformed , failfast

Null →

Timestamp →

Data preparation :

Wide tables --1600 →

Sal 10m →

int,

abc

Data cleansing /scrubbing :

Data transformation :

Join

Sort

Filter

Null values⇒ dropna , fillna , replace

Distinct or dropDuplicates()

Deptcode →withColumn (lit (“T &S “)

Cache ⇒

Partition ⇒ parallelism

Data sink → table , storage optimization (parquet/orc)

Partition ⇒ parallelism

Low latency ⇒ 10 mins → 100ns

=====

Spark jobs → job , spark context/spark session

Stages , task , executor

Partition , driver , master , worker

Transformation and action ⇒

Narrow and wide transformation ⇒

=====

Cost estimation :

Customer , line items , nation , region , orders

Handle the bad records

Drop the duplicates

Drop null values

Type cast ⇒

Sort /filter / select()

nation , region ⇒join

Create a temp view → line items

Run a → aggregation in sql query

Set the number of partition to 5

Write the df in append mode in a parquet

Reader ⇒ permissive , dropmalformed and failfast

Writer ⇒ ignore , append and overwrite

[notepad.pw / shellidasep23 | The napkin of the internet.](#)

[notepad.pw / shellidasep23 | The napkin of the internet.](#)

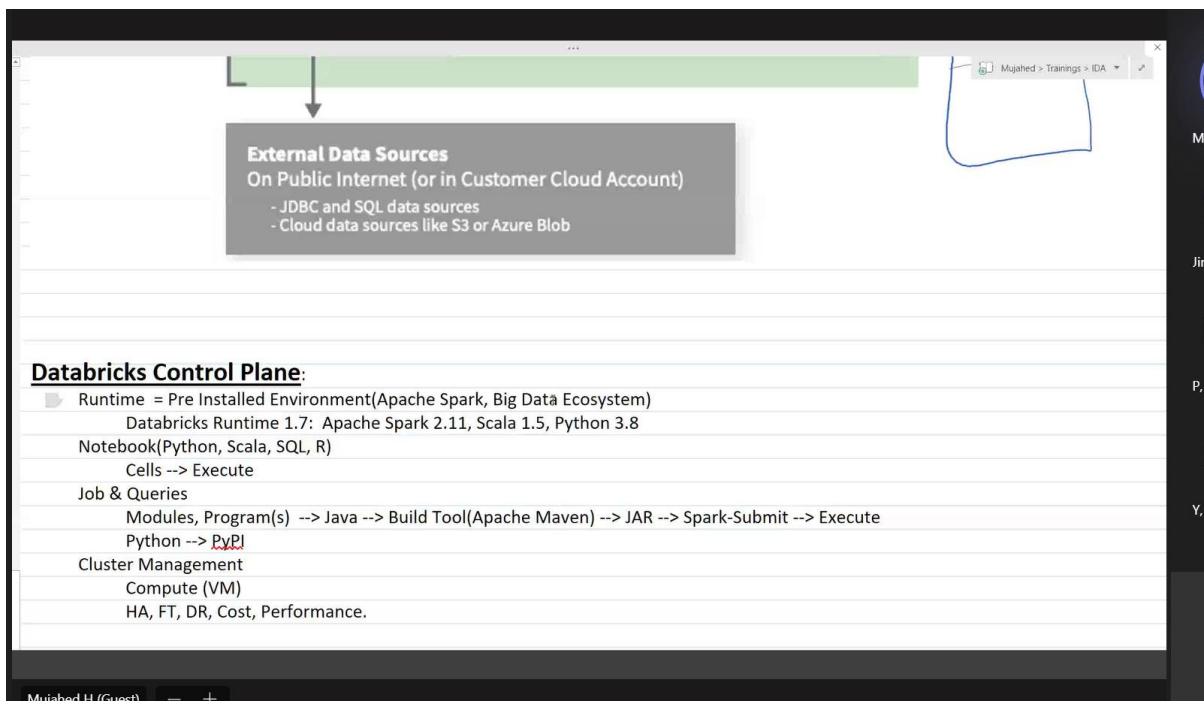
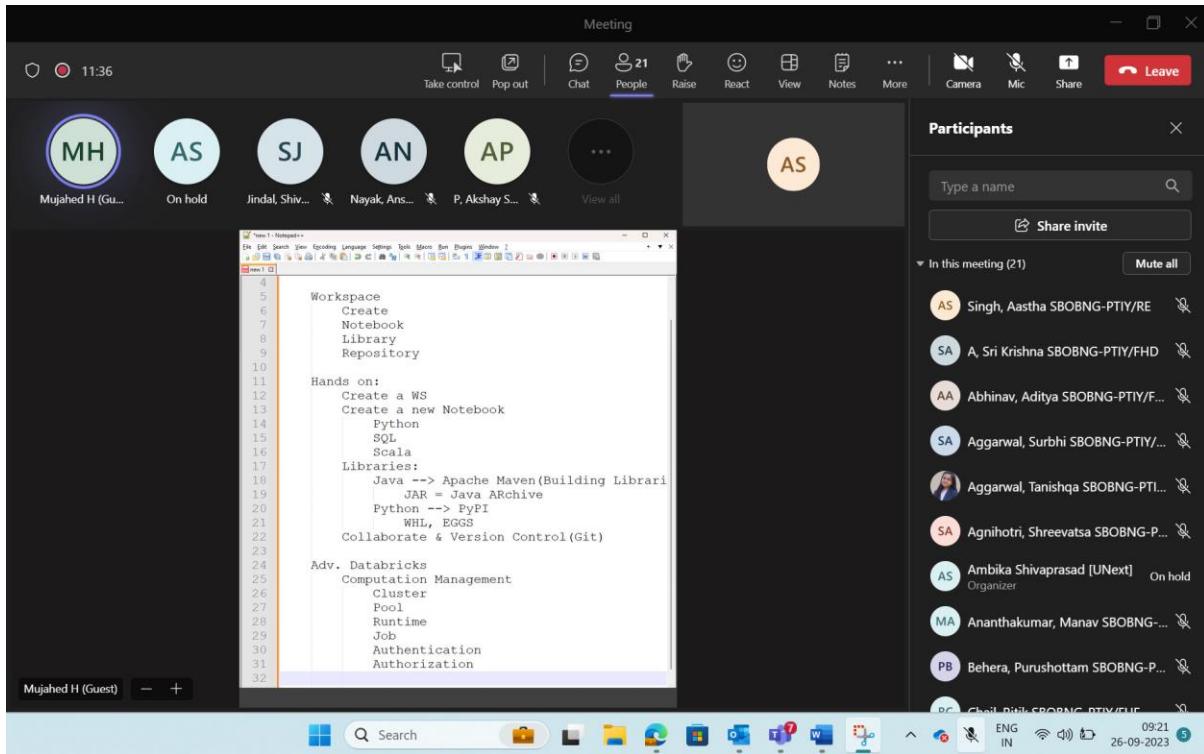
[GitHub - akgeoinsys/retail](#)

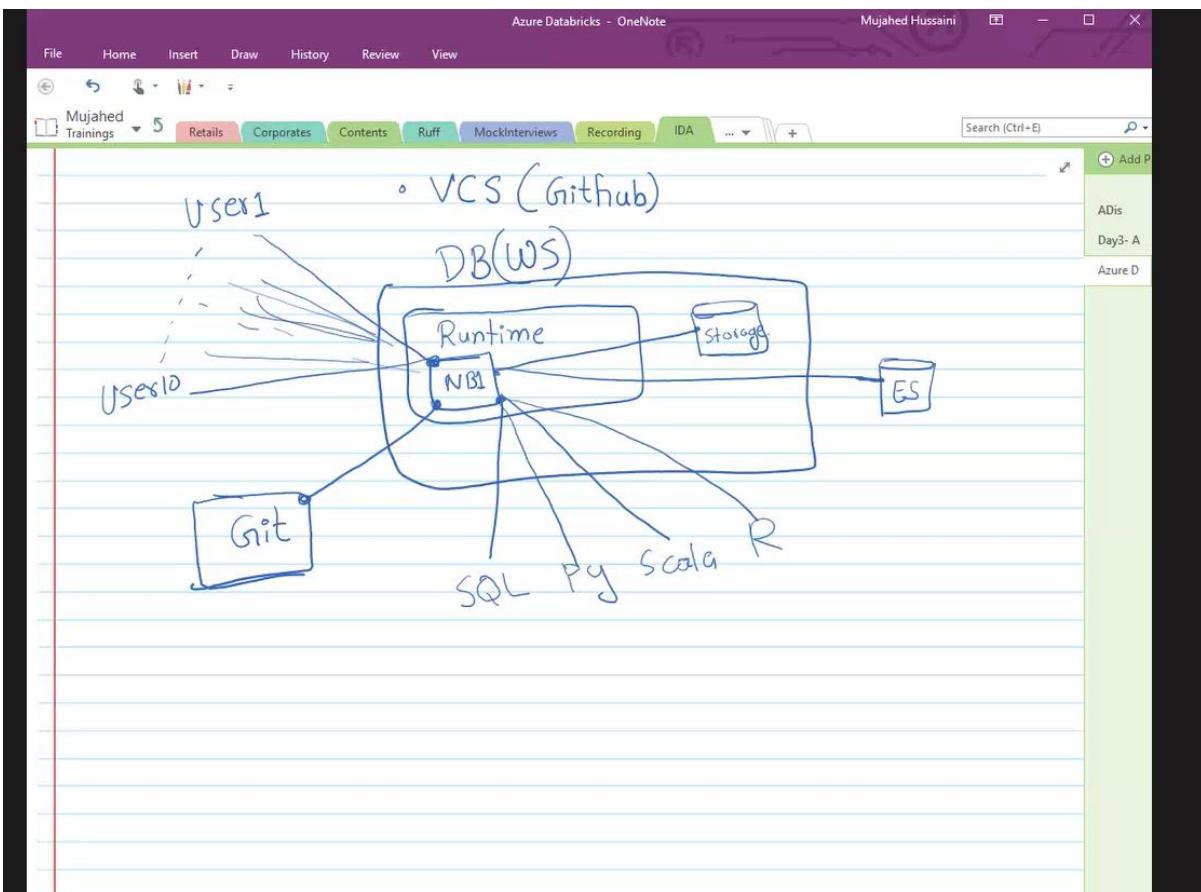
Day 02(week 5) – 26th September 2023

The screenshot shows a Notepad++ window with the title bar "new 1 - Notepad++". The menu bar includes File, Edit, Search, View, Encoding, Language, Settings, Tools, Macro, Run, Plugins, Window, and Help. The toolbar below the menu has various icons for file operations like Open, Save, Find, and Print.

The main text area contains the following code:

```
1 Azure Databricks
2     Introduction
3     Architecture
4
5     Workspace
6         Create
7         Notebook
8         Library
9         Repository
10
11    Hands on:
12        Create a WS
13        Create a new Notebook
14        Python|          ← The cursor is here
15        SQL
16        Scala
17    Libraries:
18        Java --> Apache Maven
19
20        Python --> PyPI
```





Azure Databricks - OneNote

Mujahed Hussaini

File Home Insert Draw History Review View

Mujahed Trainings Retails Corporates Contents Ruff MockInterviews Recording MEMO/Slides REMOTE/SLIDES DataEngineering IDA ... + Search (Ctrl+E)

Task: Databricks Notebook & Workspace

1. Login with Lab --> Azure Credentials
2. Login Azure --> Create Resource Group
3. Create **Azure Databricks Workspace**
4. Open Workspace -- Under User --> Create Notebook

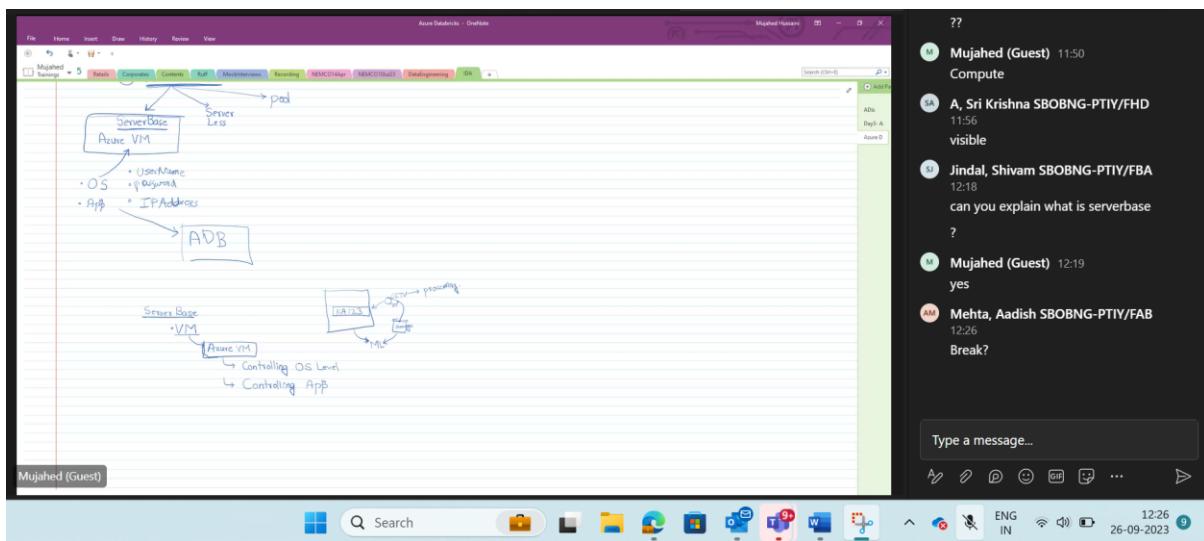
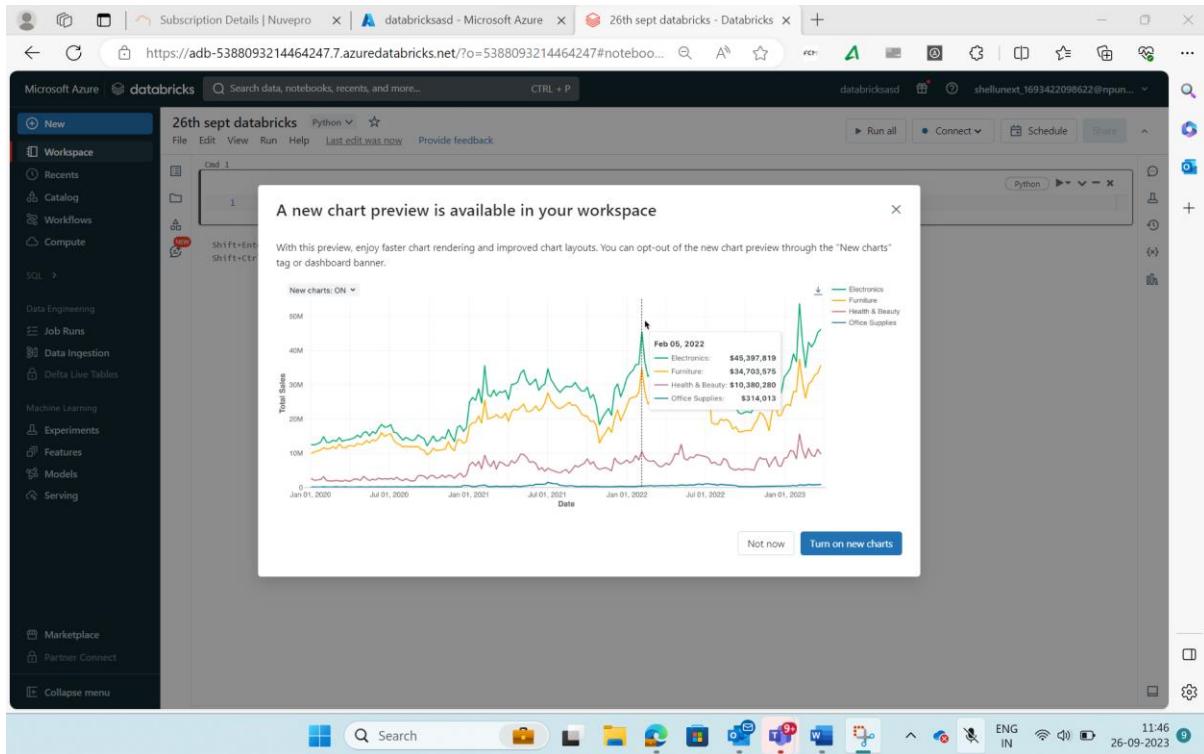
Mujahed H (Guest) - +

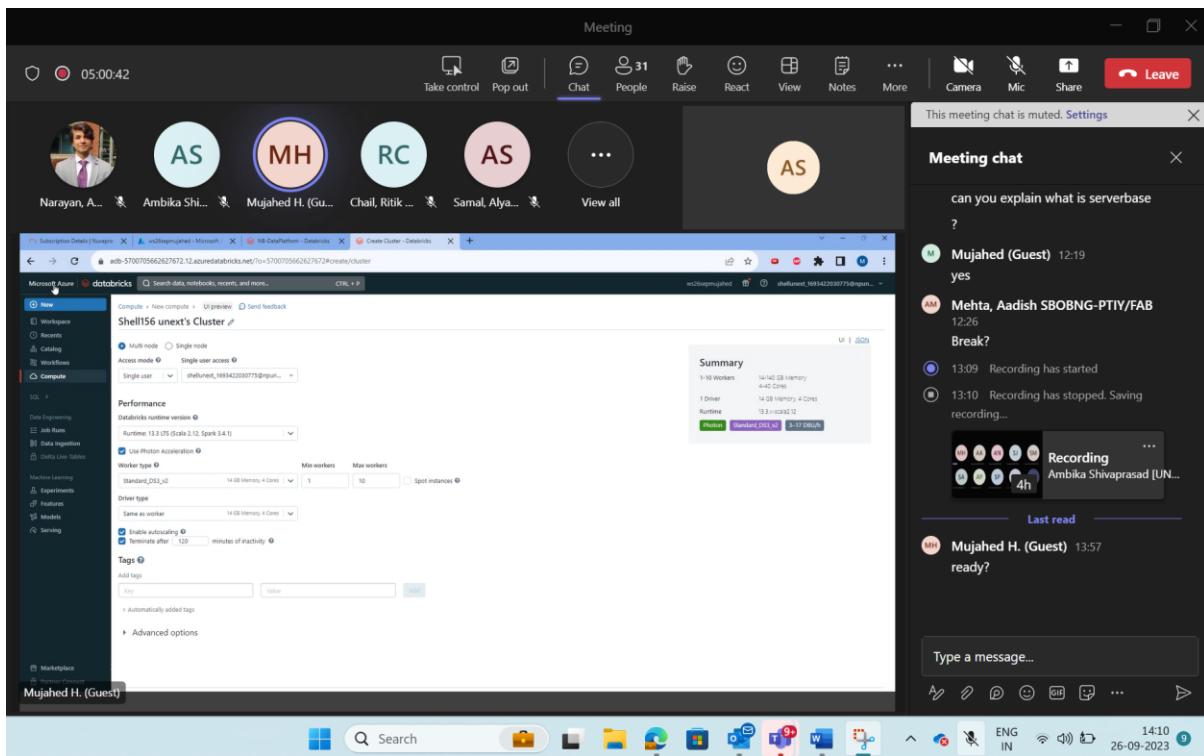
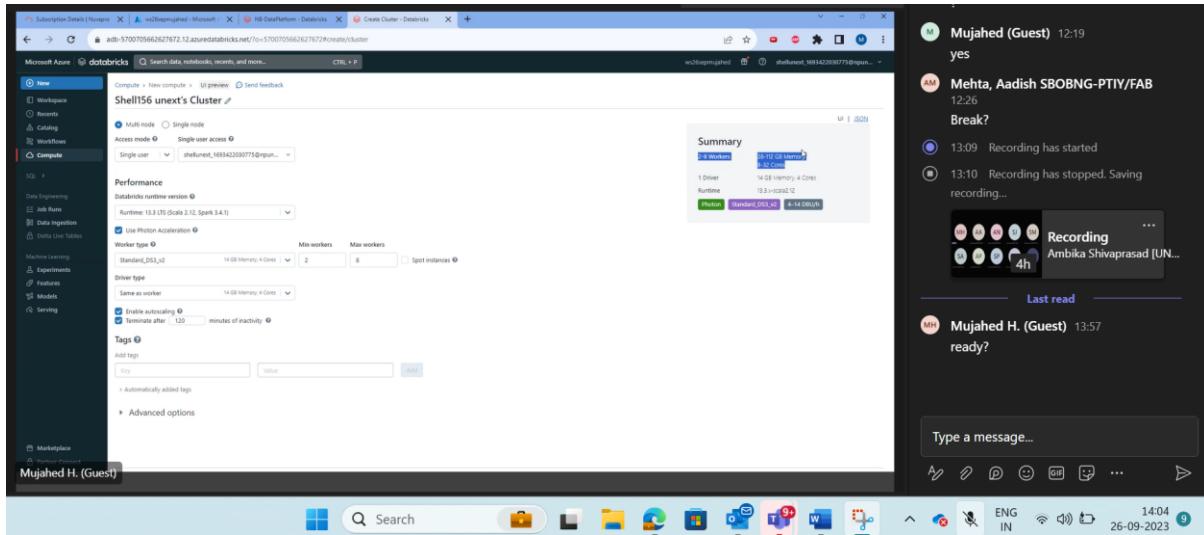
Search

11:28 26-09-2023

AS

Profile icons for users: MH (On hold), AS, SJ, YB, AP, Narayan, A., VY, ..., AS.





The screenshot shows the Microsoft Azure Databricks 'Create Cluster' interface. The left sidebar has a 'Compute' section selected. The main area shows the configuration for a cluster named 'Shell133 Unext's Cluster'. The 'Summary' panel indicates 1 Driver (14 GB Memory, 4 Cores) and Runtime (13.3.x-scala2.12). The 'Photon' node type is selected. The 'Performance' section includes settings for Databricks runtime version (13.3 LTS (Scala 2.12, Spark 3.4.1)), Use Photon Acceleration (checked), Node type (Standard_DS3_v2), and Terminate after (120 minutes of inactivity). The 'Tags' section allows adding key-value pairs. A 'Create compute' button is at the bottom.

Subscription Details | Nuvelpro | databricksasd - Microsoft Azure | Create Cluster - Databricks

Microsoft Azure | databricks | Search data, notebooks, recents, and more... | CTRL + P | databricksasd | shellunext.1693422098622@npun... | UI | JSON

Compute > New compute > UI preview | Send feedback

Shell133 Unext's Cluster

Multi node Single node

Access mode Single user access

Single user shellunext_1693422098622@npun...

Summary

1 Driver 14 GB Memory, 4 Cores
Runtime 13.3.x-scala2.12
Photon Standard_DS3_v2 1.5 DBU/h

Performance

Databricks runtime version Runtime: 13.3 LTS (Scala 2.12, Spark 3.4.1)

Use Photon Acceleration

Node type Standard_DS3_v2 14 GB Memory, 4 Cores

Terminate after 120 minutes of inactivity

Tags

Add tags

Key Value Add

Automatically added tags

Advanced options

Create compute Cancel

Screenshot of Microsoft Azure Databricks Cluster Configuration:

Compute > Shell133 Unext's Cluster

Configuration tab selected.

Access mode: Single user access (Single user: Shell133 Unext).

Performance: Databricks Runtime Version: 13.3 LTS (includes Apache Spark 3.4.1, Scala 2.12). Use Photon Acceleration: checked. Node type: Standard_DS3_v2 (14 GB Memory, 4 Cores). Terminate after: 120 minutes of inactivity.

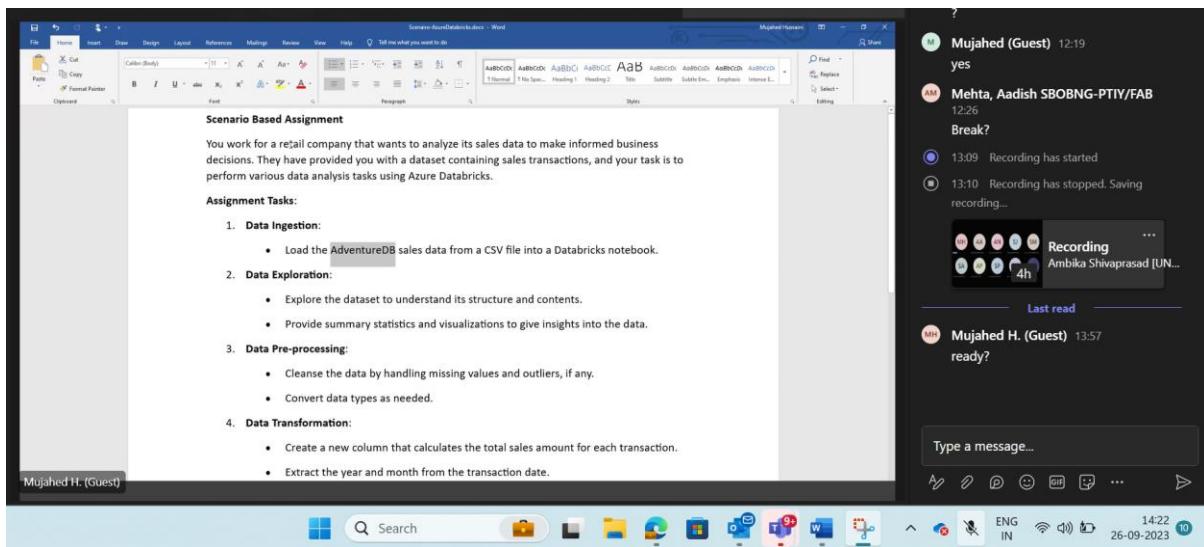
Tags: No custom tags. Advanced options.

Summary: 1 Driver, 14 GB Memory, 4 Cores, Runtime: 13.3.x-scala2.12. Photon, Standard_DS3_v2, 1.5 DBU/h.

Word Document: Scenario-AzureDatabricks.docx - Word

Assignment Tasks:

- 1. Prepare Environment:**
 - Goto Azure Portal → Login → Create Resource Group
 - Azure Databricks Service
 - Create Notebook → Save
 - Create Cluster (Single Node)
 - Connect Notebook with Cluster
- 2. Data Ingestion:**
 - Load the AdventureDB sales data from a CSV file/Table into a Databricks notebook.
- 3. Data Exploration:**
 - Explore the dataset to understand its structure and contents.
 - Provide summary statistics and visualizations to give insights into the data.
- 4. Data Pre-processing:**
 - Cleanse the data by handling missing values and outliers, if any.
 - Convert data types as needed.



```
1 print("Welcome to Data Bricks")
Welcome to Data Bricks
Command took 0.53 seconds -- by shellunext_1693422098622@npunext.onmicrosoft.com at 26/9/2023, 2:21:45 pm on Shell133 Unext's Cluster
```

Scenairo-AzureDatabricks.docx - Google Docs

[Load data using the add data UI | Databricks on AWS](#)

Day 03(week 5) – 27th September 2023

The screenshot shows the Databricks Cluster Details page. On the left, a sidebar menu is open under the 'Compute' section, showing options like Workspace, Recents, Catalog, Workflows, and SQL. The main area displays the configuration for a cluster named 'Shell133 Unixt's Cluster'. The 'Configuration' tab is selected. Key settings include:

- Access mode:** Single user (selected), Single user access (Shell133 Unixt).
- Performance:** Databricks Runtime Version: 13.3 LTS (includes Apache Spark 3.4.1, Scala 2.12). Node type: Standard_DS3_v2 (14 GB Memory, 4 Cores). Photon Acceleration is checked.
- Tags:** No custom tags. Advanced options are visible.
- Summary:** 1 Driver, 14 GB Memory, 4 Cores, Runtime: 13.3.x-scala2.12, Photon, Standard_DS3_v2, 1.5 DBU/h.

The browser address bar shows the URL: <https://adb-4232521033896513.13.azuredatabricks.net/?o=4232521033896513#setting/>.

/FileStore/tables/movies.csv

The screenshot shows the Databricks DBFS upload interface. The sidebar menu is open under the 'Compute' section. The main area shows the 'DBFS' section with the 'Upload File' tab selected. A file named 'movies.csv' is shown in the 'DBFS Target Directory' field, which is set to '/FileStore/tables/'. The file is described as 5.1 KB and has a 'Remove file' option. A message at the bottom indicates the file was uploaded successfully to '/FileStore/tables/movies.csv'. There are buttons for 'Create Table with UI' and 'Create Table in Notebook'.

The browser address bar shows the URL: <https://adb-4232521033896513.13.azuredatabricks.net/?o=4232521033896513#tables/>.

The screenshot shows a Microsoft Azure Databricks notebook titled "27th Notebook" running Python code. The code reads a CSV file from the FileStore and displays its contents. The notebook interface includes a sidebar with various workspace options like Recents, Catalog, Workflows, Compute, SQL, Data Engineering, Job Runs, Data Ingestion, Delta Live Tables, Machine Learning, Experiments, Features, Models, Serving, Marketplace, and Partner Connect. The main area shows the command history and the resulting DataFrame.

```
df = spark.read.csv("/FileStore/tables/movies.csv", header=True, inferSchema=True)
```

```
df.show()
```

	Film	Genre	Lead Studio	Audience score %	Profitability	Rotten Tomatoes %	Worldwide Gross	Year
Zack and Miri Mak...	Romance	The Weinstein Com...		70 1.747514667	64	\$41.94	2008	
Youth in Revolt	Comedy	The Weinstein Com...		52 1.09	68	\$19.62	2010	
You Will Meet a T...	Comedy	Independent		35 1.21181812	43	\$26.66	2010	
When in Rome	Comedy	Disney		44 0.8	15	\$43.04	2010	
What Happens in V...	Comedy	Fox		72 6.267647029	28	\$219.37	2008	
Water for Elephants	Drama	20th Century Fox		72 3.081421053	68	\$117.09	2011	
WALL-E	Animation	Disney		89 2.89601967	96	\$521.28	2008	
Waitress	Romance	Independent		67 11.8997415	89	\$22.18	2007	
Waiting For Forever	Romance	Independent		53 0.095	6	\$0.03	2011	
Valentine's Day	Comedy	Warner Bros.		54 4.184038462	17	\$217.57	2010	
Tyler Perry's Why...	Romance	Independent		47 3.7241924	46	\$55.86	2007	
Twilight: Breakin...	Romance	Independent		68 6.383363636	26	\$762.17	2011	
Twilight	Romance	Summit		82 19.18902703	49	\$376.66	2008	
The Ugly Truth	Comedy	Independent		68 5.482631579	14	\$265.30	2009	
The Twilight Saga...	Drama	Summit		78 14.1964	27	\$709.82	2009	
The Time Traveler...	Drama	Paramount		65 2.598205128	38	\$101.33	2009	
The Proposal	Comedy	Disney		74 7.8675	43	\$314.70	2009	
The Invention of ...	Comedy	Warner Bros.		47 1.751351351	56	\$32.40	2009	

Command took 0.46 seconds -- by shellnext_169342298622@npunext.onmicrosoft.com at 27/9/2023, 9:39:49 am on Shell133 Unext's Cluster

The screenshot shows a Microsoft Azure Databricks notebook titled "27th Notebook" running in Python. The notebook interface includes a left sidebar with navigation links like "New", "Workspace", "Recents", "Catalog", "Workflows", "Compute", "SQL", "Data Engineering", "Job Runs", "Data Ingestion", "Delta Live Tables", "Machine Learning", "Experiments", "Features", "Models", "Serving", "Marketplace", "Partner Connect", and "Collapse menu". The main workspace displays three code cells:

- Cell 4:** Contains the command `!ls` followed by the output of the Azure command `azure conf eventlogs hadoop_accessed_config.lst logs preload_class.lst`. A note indicates the command took 0.22 seconds.
- Cell 5:** Contains the command `dbutils.help()`. The output provides a detailed list of available utilities:
 - credential:** `DatabricksCredentialUtils` - Utilities for interacting with credentials within notebooks
 - data:** `DataUtils` - Utilities for understanding and interacting with datasets (EXPERIMENTAL)
 - fs:** `DatabricksFilesystemClient` - Manipulates the Databricks filesystem (DBFS) from the console
 - jobs:** `JobsUtils` - Utilities for leveraging job features
 - library:** `LibraryUtils` - Utilities for session isolated libraries
 - meta:** `MetadataUtils` - Methods to hook into the compiler (EXPERIMENTAL)
 - notebook:** `NotebookClient` - Utilities for the content view of a notebook (EXPERIMENTAL)
 - preview:** `Preview` - Utilities under preview category
 - secrets:** `SecretUtils` - Provides utilities for leveraging secrets within notebooks
 - widgets:** `WidgetsUtils` - Methods to create and get bound value of input widgets inside notebooksA note indicates the command took 0.07 seconds.
- Cell 6:** Contains the command `1` and a note: "Shift+Enter to run" and "Shift+Ctrl+Enter to run selected text".

The screenshot shows a Microsoft Edge browser window with multiple tabs open. The active tab is a Databricks notebook titled "27th Notebook" in Python. The notebook contains the following code:

```
1 fmt="delta"
2 load_path="/databricks-datasets/learning-spark-v2/people/people-10m.delta"
3 dbutils.fs.ls(load_path)
4 dfPeople=spark.read.format(fmt).load(load_path)
5 display(dfPeople)
```

The output of the code shows a DataFrame named "dfPeople" with the following schema:

```
dfPeople: pyspark.sql.DataFrame[{"id": integer, "firstName": string, "middleName": string, "lastName": string, "gender": string, "birthDate": timestamp, "ssn": string, "salary": double}]
```

The DataFrame has 10 columns: id, firstName, middleName, lastName, gender, birthDate, ssn, and salary. The data is a truncated version of the "people-10m" dataset, containing 10,000 rows. The runtime for this command was 3.80 seconds.

Day 04(week 5) – 28th September 2023

Went to EcoWorld to collect laptop

Day 05(week 5) – 29th September 2023

Azure Deployment/Devops Online Session – Batch 4

08:08

Take control Pop out Chat People Raise React View Notes More Camera Mic Share Leave

SJ AS M SA AS SA AN ... AS

Jindal, Shiv... Ambika Shi... Mujahed (Guest) Aggarwal, ... Samal, Alya... A, Sri Krish... Nayak, Ans... View all

Participants

Type a name Share invite

In this meeting (18) Mute all

AS Singh, Aastha SBOBNG-PTIV/RE
SA A, Sri Krishna SBOBNG-PTIV/FHD
SA Aggarwal, Surbhi SBOBNG-PTIV/...
AS Ambika Shivaprasad [UNext] Organizer
MA Ananthakumar, Manav SBOBNG-...
Gupta, Sumit K SBOBNG-PTIV/FBD
Jani, Siddharth S SBOBNG-PTIV/...
SJ Jindal, Shivam SBOBNG-PTIV/FBA
M Mujahed (Guest)
Narayan, Aditya SBOBNG-PTIV/F...
AN Nayak, Ansuman SBOBNG-PTIV/...
AP P, Akshay SBOBNG-PTIV/FAA
Patro, Shailesh SBOBNG-PTIV/FAB

29 September 2023 09:16 AM

Data + Platform = Automation ----> Intelligence

Data

Traditional Data = SQL, RDBMS
Big Data = Hadoop, Spark, Databricks
New Data = IoT, Sensors, Streaming

Platform

Cloud

1. Microsoft Azure
2. AWS
3. GCP

Automation

DevOps

Azure CI/CD

Mujahed (Guest)

75°F Partly sunny

Search

9:21 AM 9/29/2023

2. AWS

3. GCP

Automation

DevOps

- Development and Operations --> Automation
- Tools:
 - VCS Tool (Github)
 - Build Tool(Apache Maven)
 - Continues Integration(CI) - Jenkins
- Automation
 - Developer = Code
 - Operations = Testing, Building, Integration, Deployment, Provisioning

SysOps

- System Administrators + Engineers --> Automation
- Tools:
 - Deployment tool (Ansible)
 - Provisioning tool (Terraform)

CloudOps

- CloudOps = DevOps + SysOps + Cloud
-

Mujahed (Guest) — +

Azure Deployment/Devops Online Session – Batch 4

27:18

Take control Pop out Chat People Raise React View Notes More Camera Mic Share Leave

Jindal, Shiv... Ambika Shi... Mujahed (Guest) Aggarwal, ... Samal, Alya... Nayak, Ans... Patro, Shall... View all

New Data = IoT, Sensors, Streaming

Platform

Cloud

- 1. Microsoft Azure
- 2. AWS
- 3. GCP

Automation

DevOps

- Development and Operations --> Automation
- Tools:
 - VCS Tool (Github)
 - Build Tool Apache Maven
 - Continues Integration(CI) - Jenkins
- Automation
 - Developer = Code
 - Operations = Testing, Building, Integration, Deployment, Provisioning

SysOps

- System Administrators + Engineers --> Automation
- Tools:
 - Deployment tool (Ansible, Azure Deployment Service - Pipeline)
 - Provisioning tool (Terraform)

CloudOps

- CloudOps = DevOps + SysOps + **DataOps + AIOps** + Cloud
- Cloud = Microsoft Azure, GitOps = Github Actions
- DevOps- CI(Azure DevOps) & SysOps- CD (Github Actions), Provisioning(Terraform) = CI/CD

MultiCloudOps(MCO) = Multi Clouds + DevOps + SysOps + GitOps + DataOps + **AIOps**

Introduction to Azure Deployment and DevOps:
Overview of Azure services for deployment and DevOps
Understanding the benefits of implementing DevOps practices
Introduction to Continuous Integration (CI) and Continuous Deployment (CD)

Mujahed (Guest)

Participants

Type a name

Share invite

In this meeting (25) Mute all

- AS Singh, Aastha SBOBNG-PTIY/RE
- SA Aggarwal, Surbhi SBOBNG-PTIY/...
- MA Ambika Shivaprasad [UNext] Organizer
- MA Ananthakumar, Manav SBOBNG-...
- PB Behera, Purushottam SBOBNG-P...
- RC Chail, Ritik SBOBNG-PTIY/FUF
- Gupta, Sumit K SBOBNG-PTIY/FBF
- AJ Jain, Ayushi SBOBNG-PTIY/FUC
- Jani, Siddharth S SBOBNG-PTIY/...
- SJ Jindal, Shivam SBOBNG-PTIY/FBA
- SK Katyal, Shruti SBOBNG-PTIY/FUD
- SM Matcha, Sai Vinay V SBOBNG-PT...
- AM Mehta, Aadish SBOBNG-PTIY/FAB

75°F Partly sunny Search

9:40 AM 9/29/2023