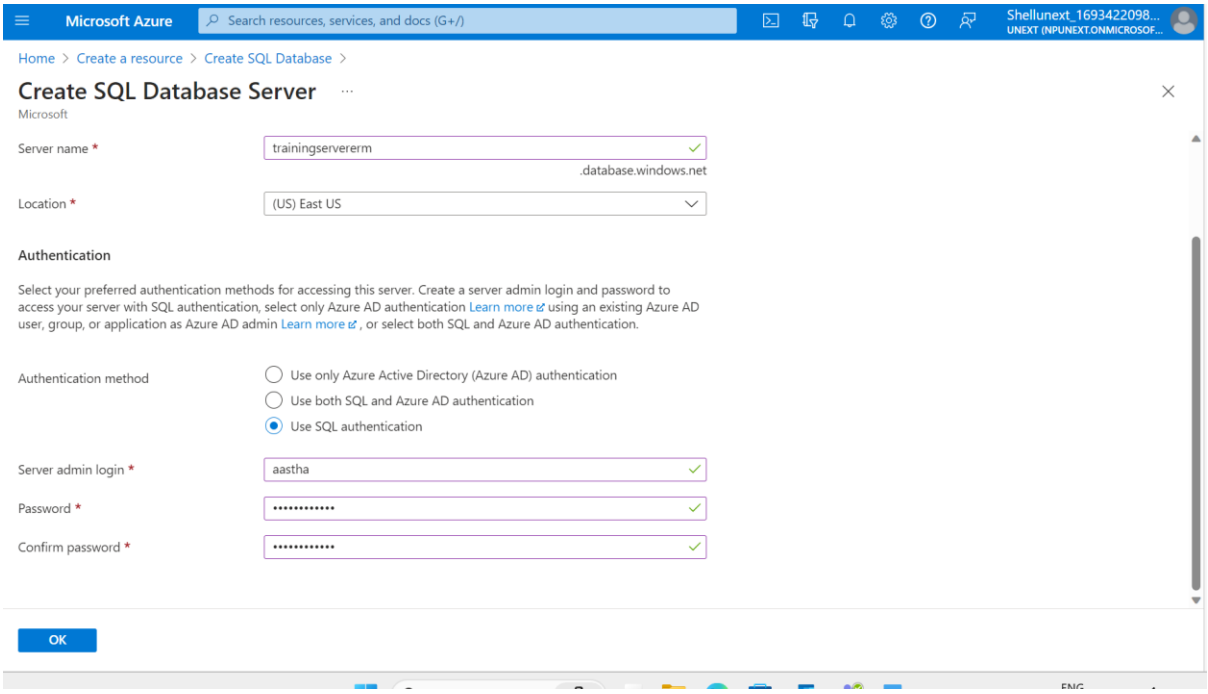
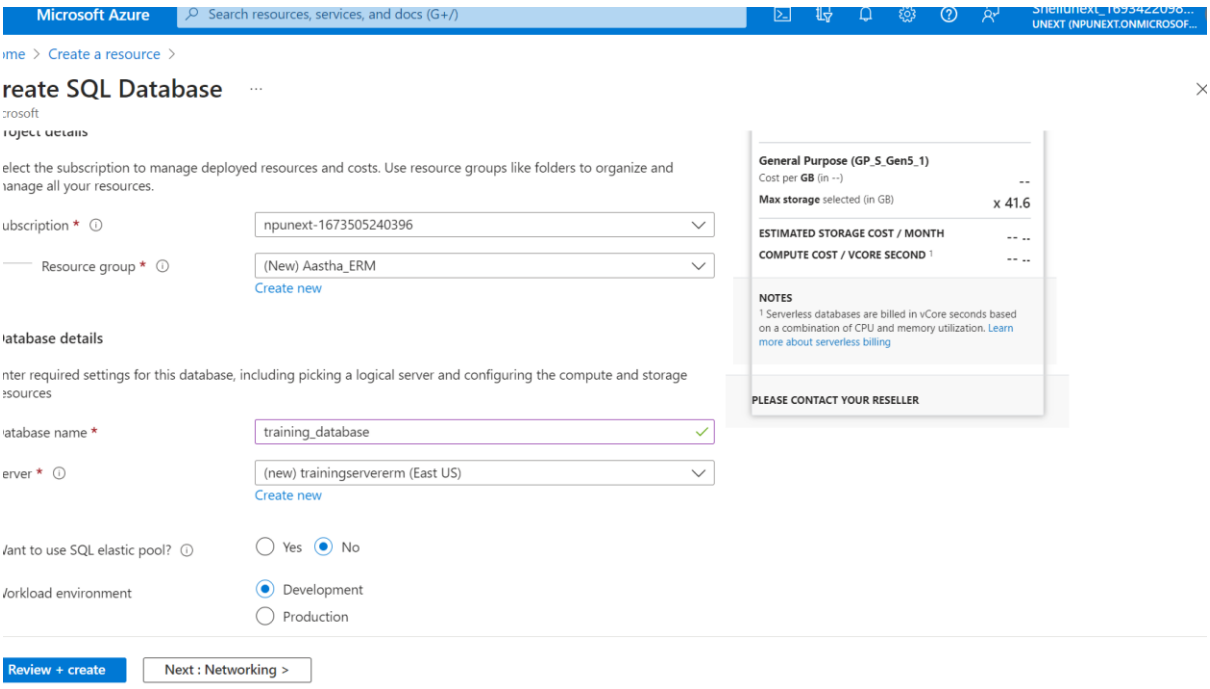
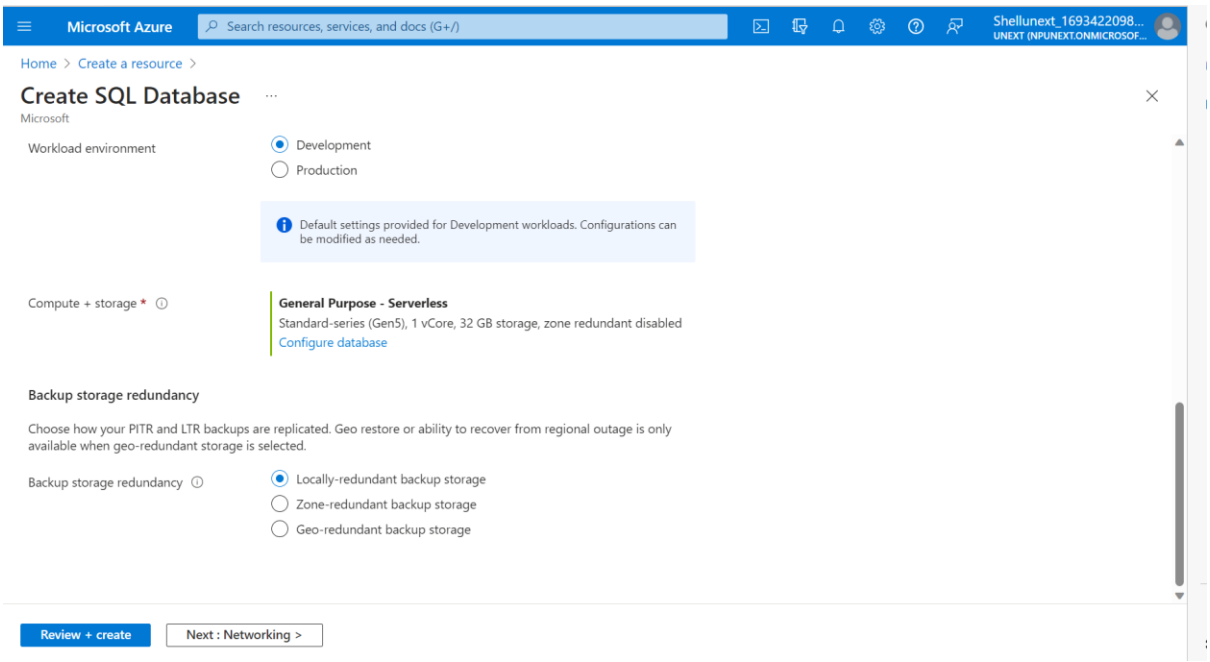
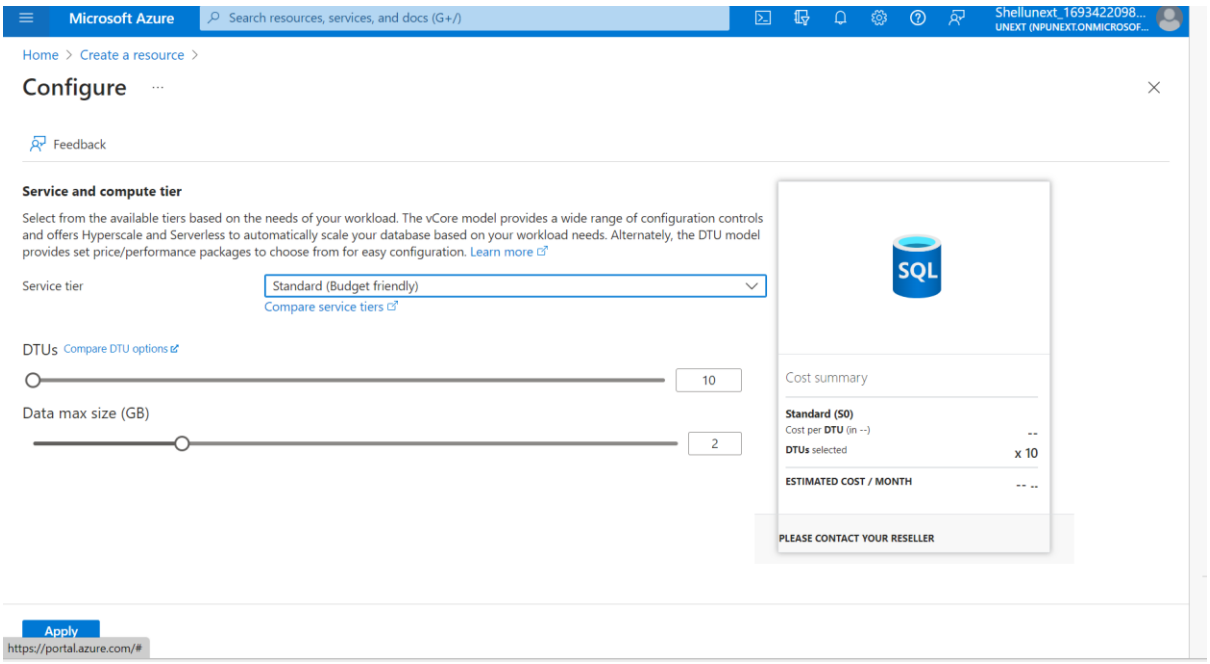
31/08/2023 DAY 3 CUSTOM TRAINING

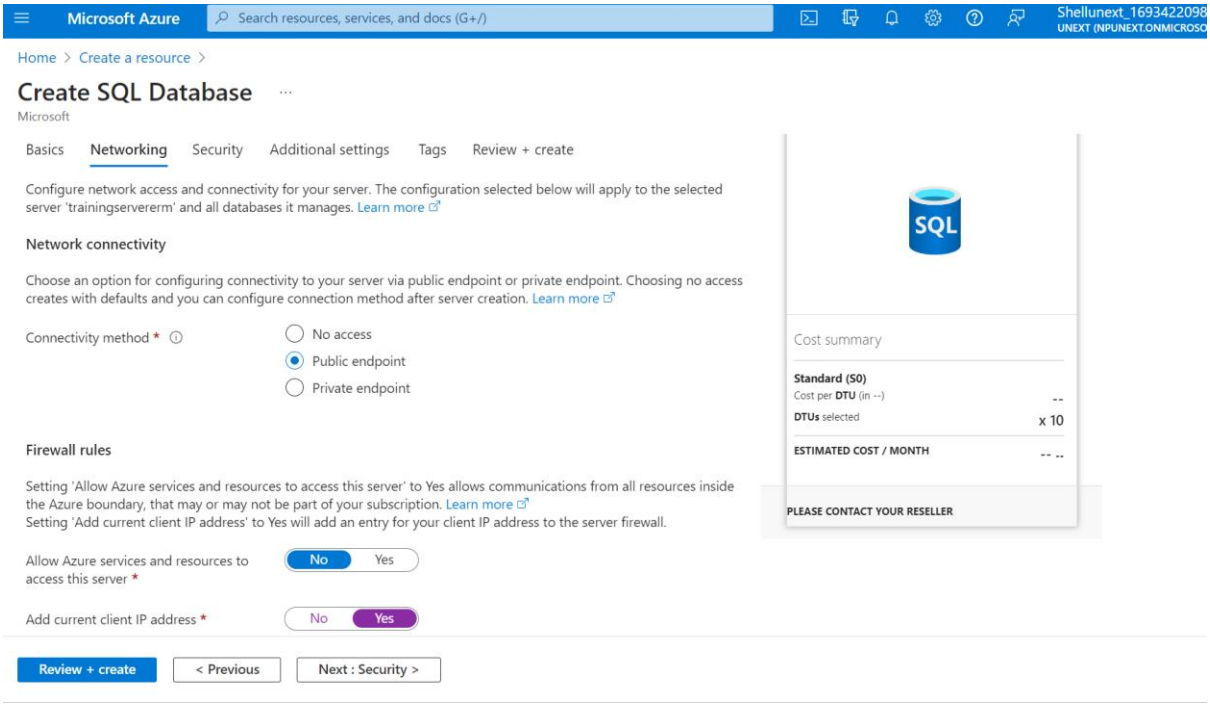
Server



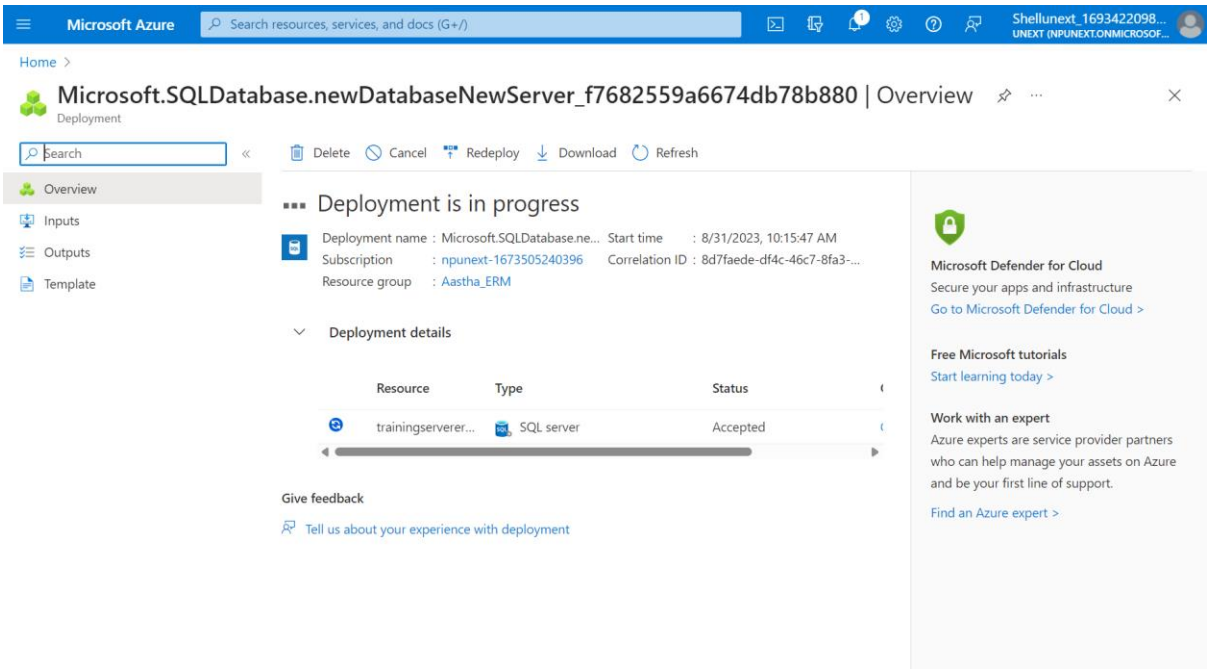




Microsoft Azure
Search resources, services, and docs (G+/)
Shellunext_1693422098...
UNEXT (NPUNEXT.ONMICROSOF...
Home > Create a resource >
Create SQL Database
Microsoft
Workload environment
Development
Production
Default settings provided for Development workloads. Configurations can be modified as 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+/)
Shellunext_1693422098...
UNEXT (NPUNEXT.ONMICROSOF...
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 (S0)
Cost per DTU (in --)
--
DTUs selected
x 10
ESTIMATED COST / MONTH
-- --
PLEASE CONTACT YOUR RESELLER
Apply
https://portal.azure.com/#



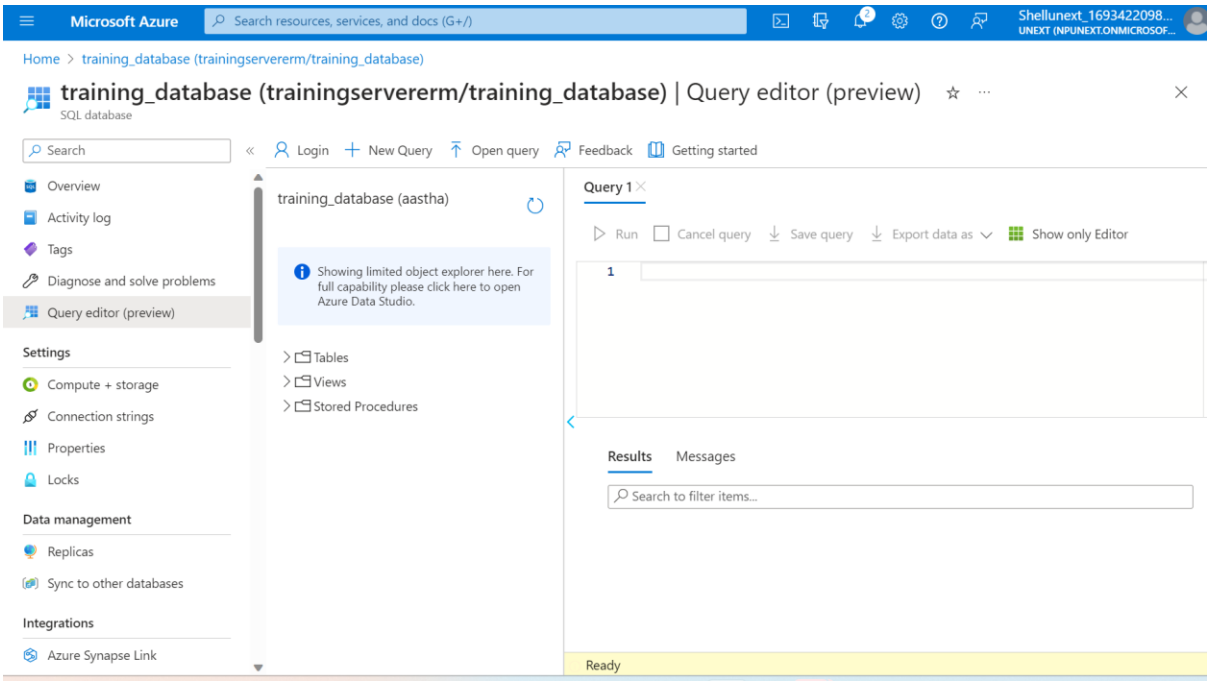

Click on Review and create then create




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]




Deployment Models: single, elastic pools, managed instance

Sql commands category- 4/5 categories:

1.DDL- create, alter, drop, truncate

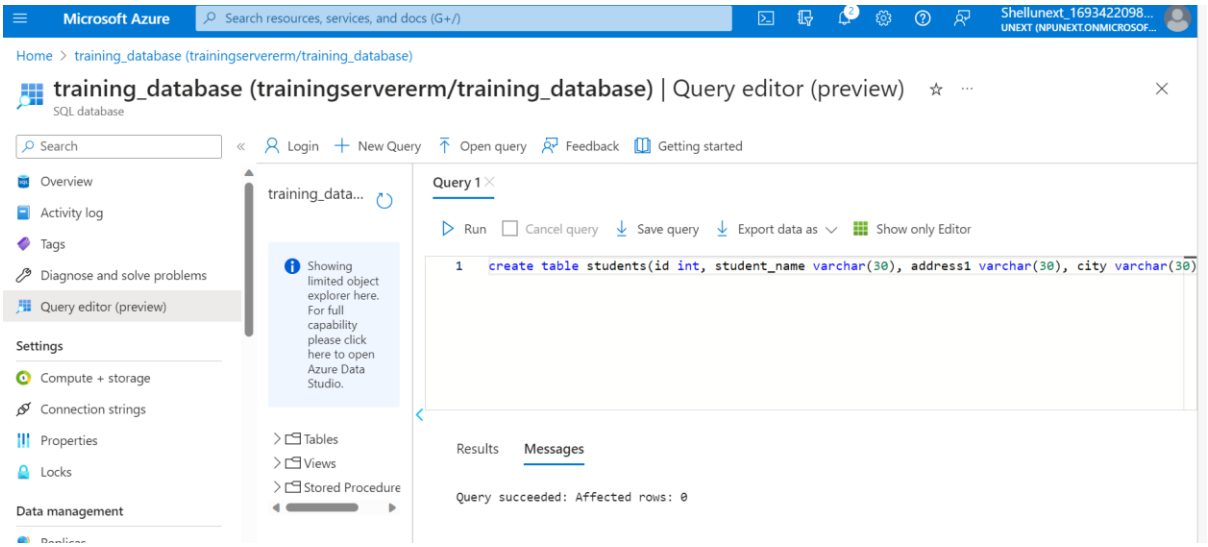
Diff between truncate(will delete all the rows) and delete(used with conditions and can be undone)

2.DML- insert, update, delete, select

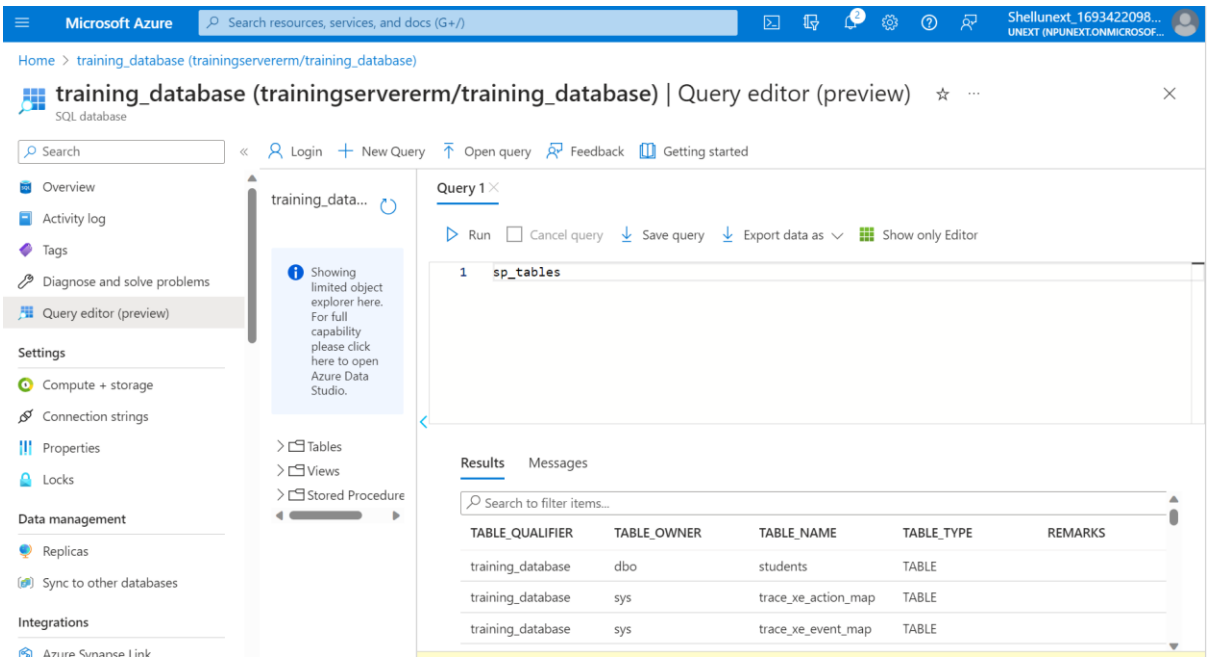
3.DCL- grant, revoke

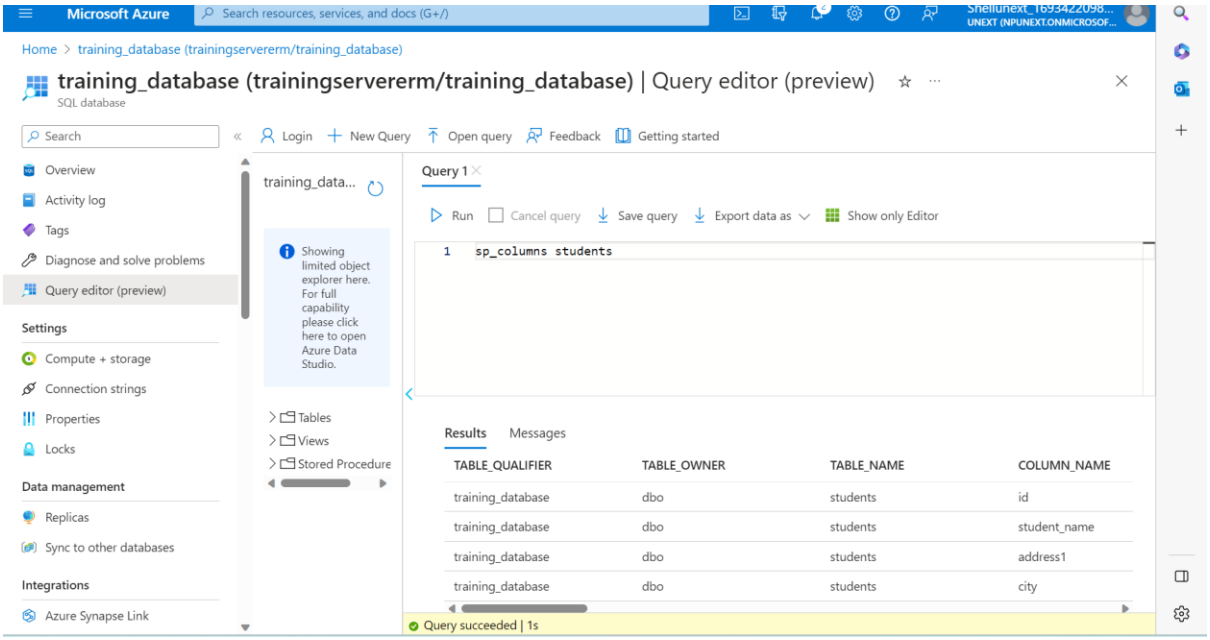
4.ECL-commit, rollback, save point

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

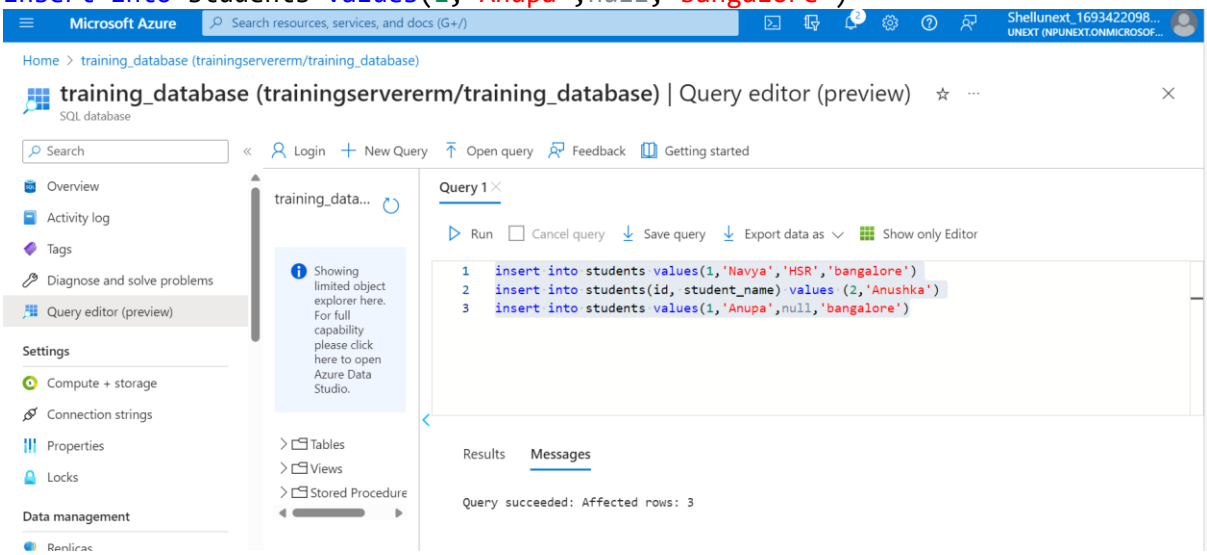



sp_tables- show tables

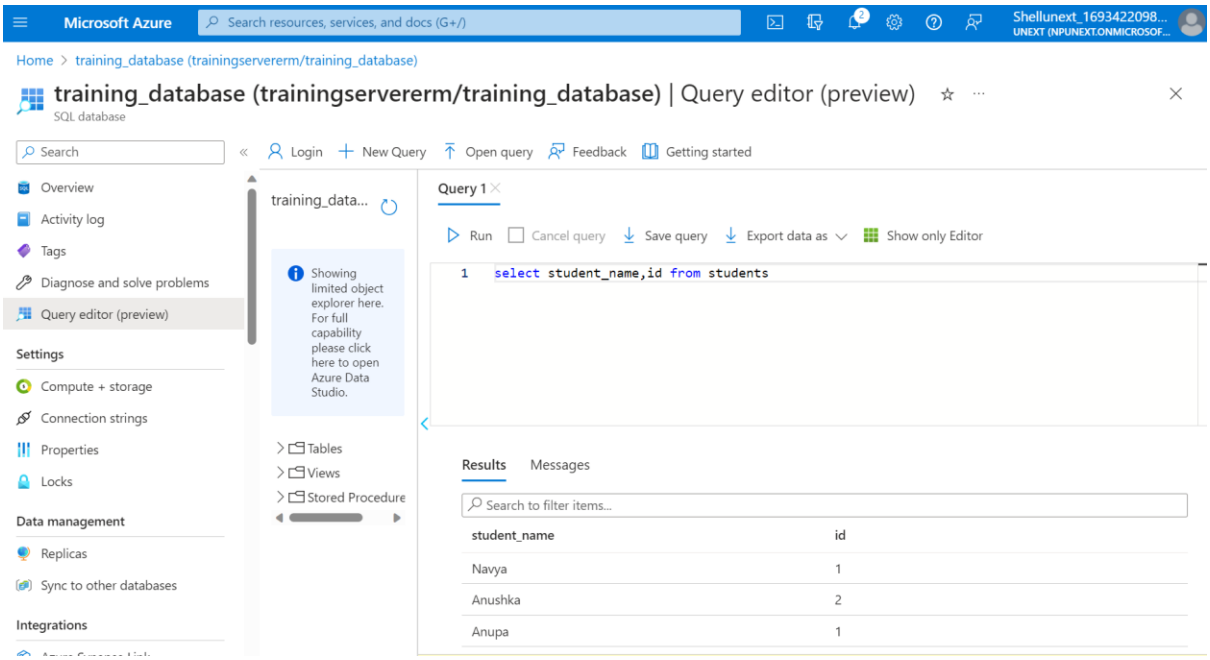





```
insert into students values(1,'Navya','HSR','bangalore')
insert into students(id, student_name) values (2,'Anushka')
insert into students values(1,'Anupa',null,'bangalore')
```




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




```
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
**Constrains**- 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

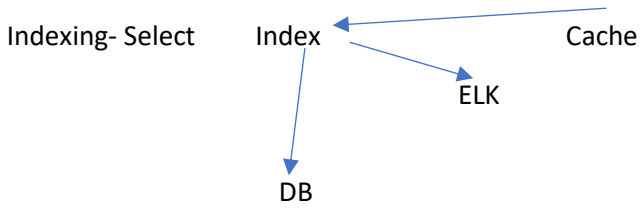
[EMP and DEPT table script – SQL Server – Data Analytics and BI WORLD (wordpress.com)]

Day- 05/09/2023

Azure Storage services, pic on mobile

SMB server- client is ASA

Find slide on lumen about this

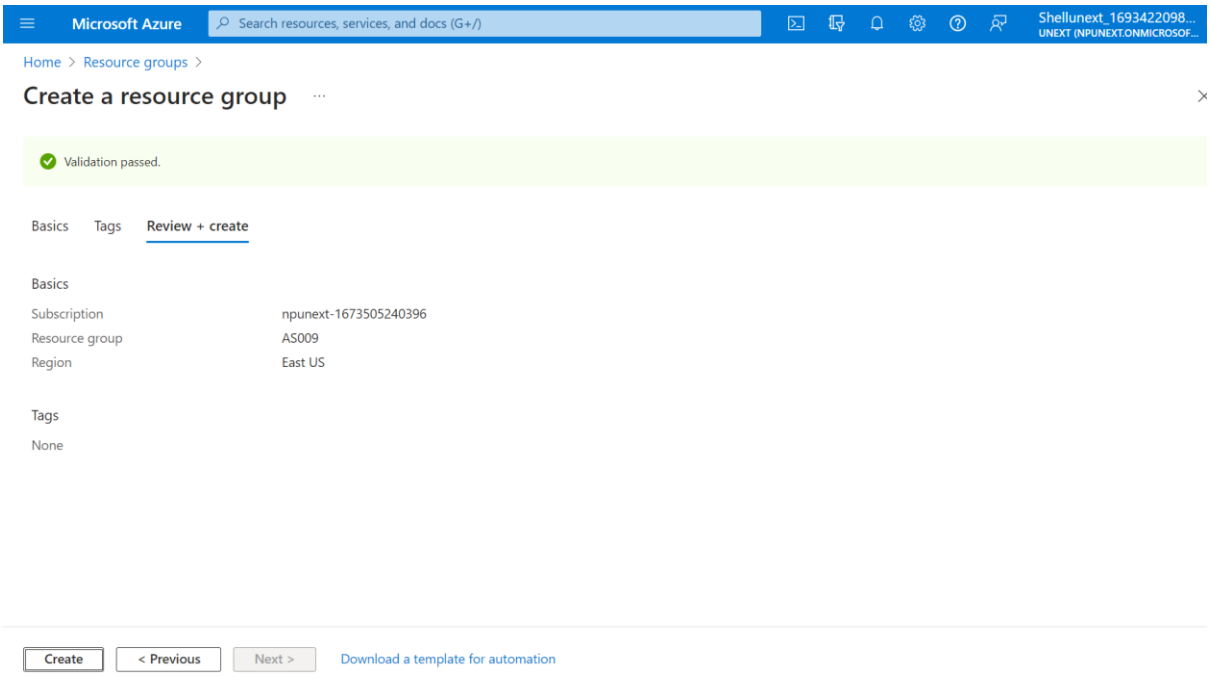


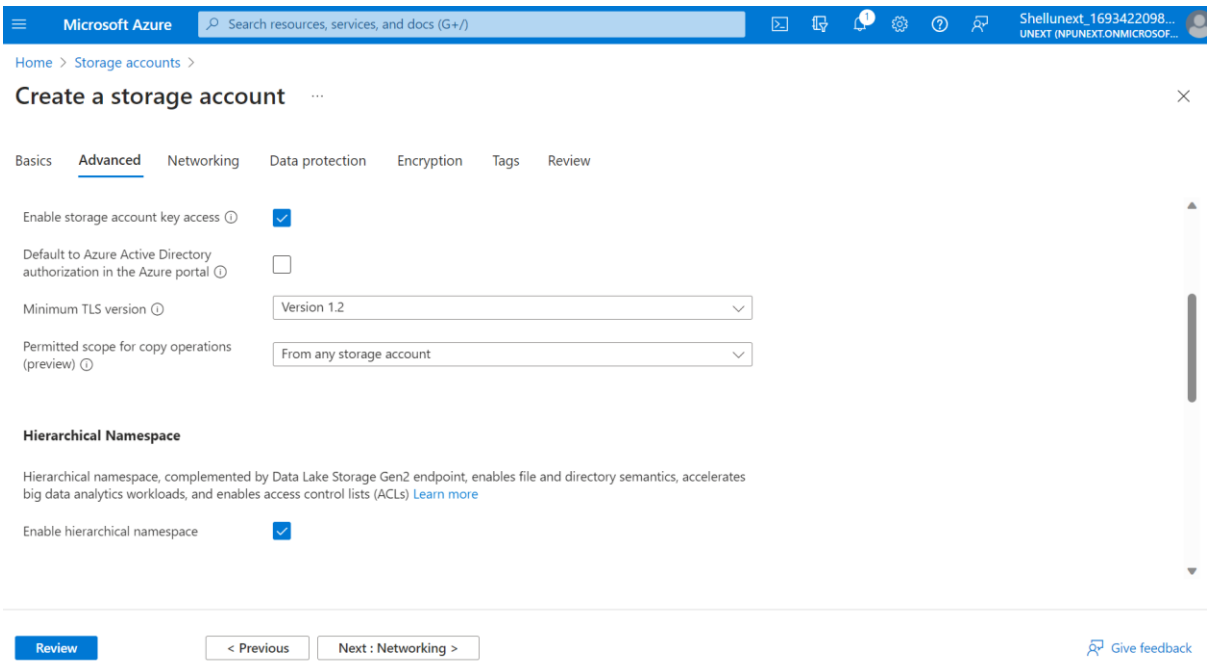

1mb-cache

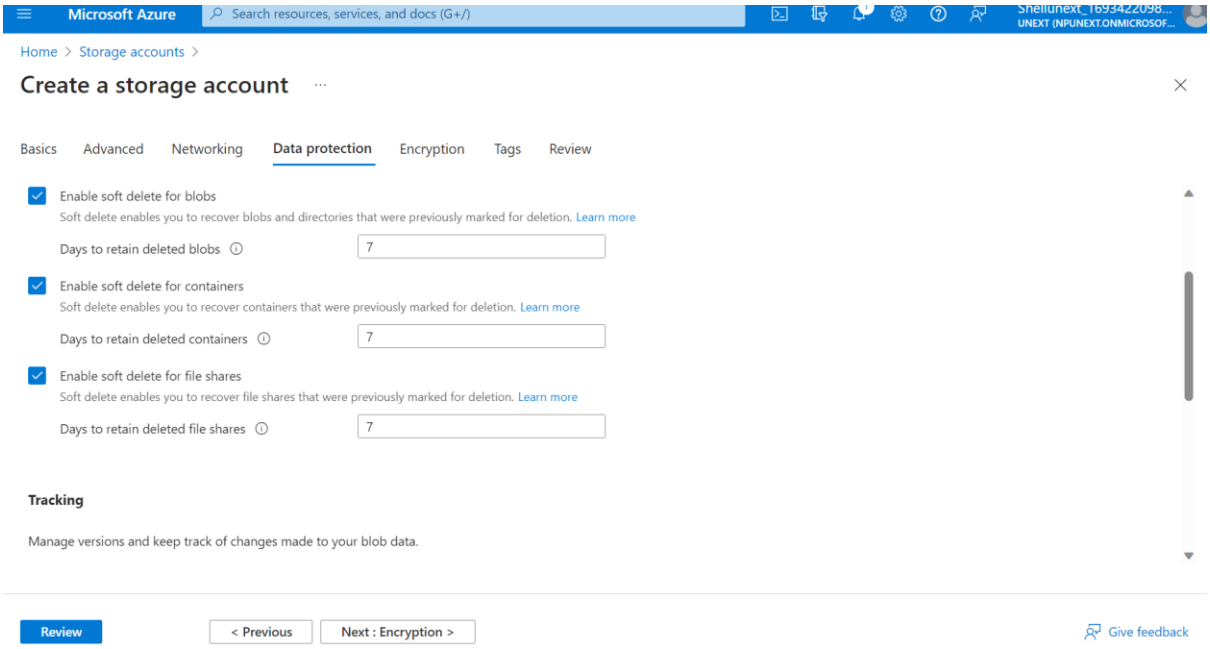
1gb- RAM

1TB- HDD

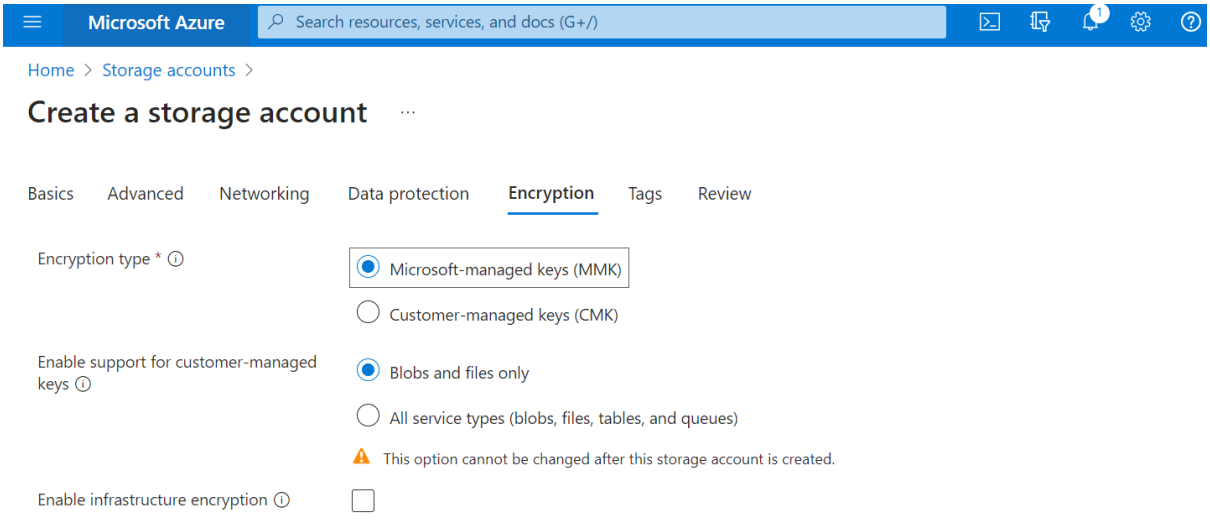
Create resource

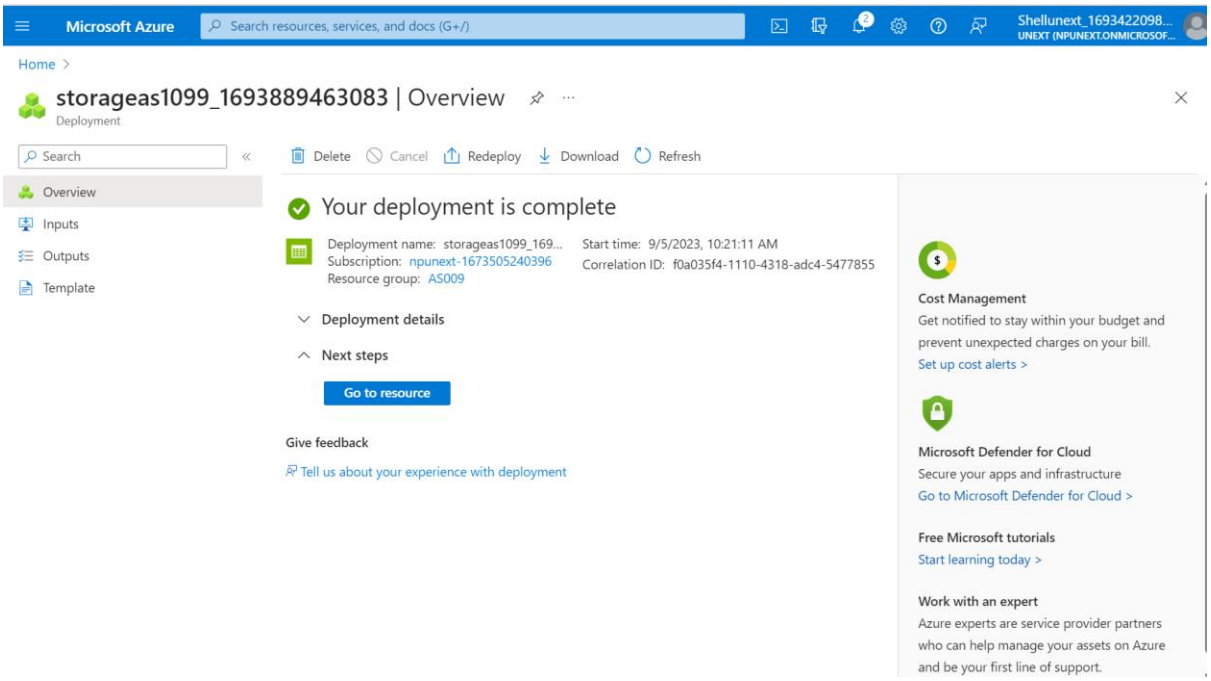

Microsoft Azure
Search resources, services, and docs (G+/)
Shellunext_1693422098...
UNEXT (NPUNEXT.ONMICROSOF...
Home > Resource groups >
Create a resource group
Validation passed.
Basics Tags Review + create
Basics
Subscription npunext-1673505240396
Resource group AS009
Region East US
Tags
None
Create
< Previous
Next >
Download a template for automation

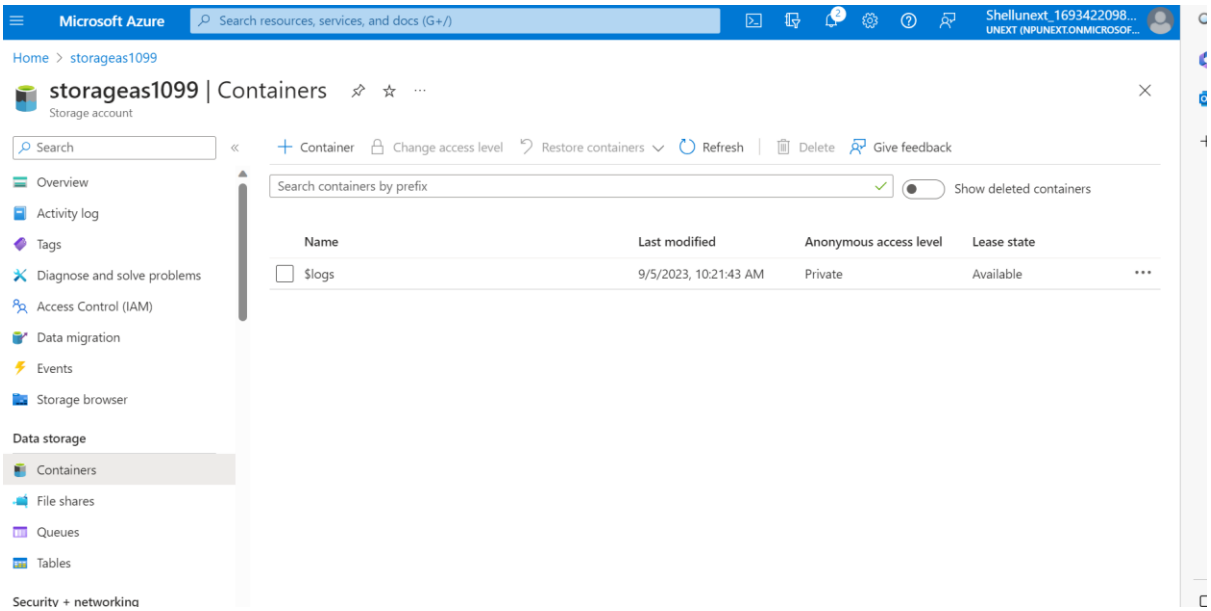

Microsoft Azure
Search resources, services, and docs (G+/)
Shellunext_1693422098...
UNEXT (NPUNEXT.ONMICROSOF...
Home > Storage accounts >
Create a storage account
Basics Advanced Networking Data protection Encryption Tags Review
Enable storage account key access
Default to Azure Active Directory authorization in the Azure portal
Minimum TLS version
Version 1.2
Permitted scope for copy operations (preview)
From any storage account
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



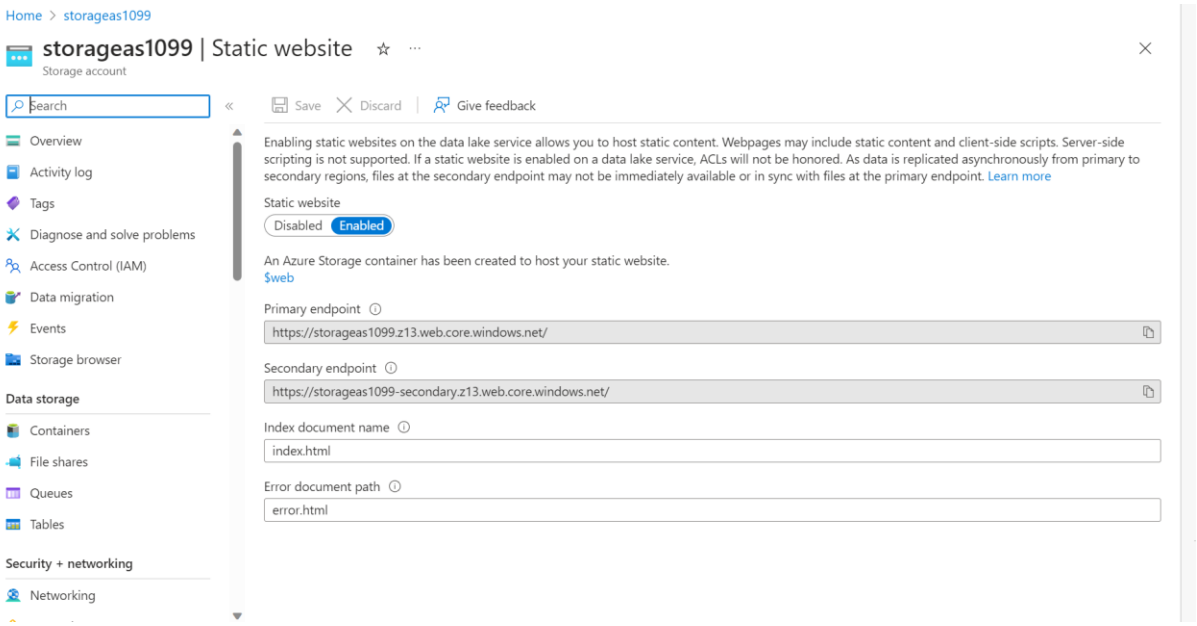

2 encryption keys- MMK and CMK





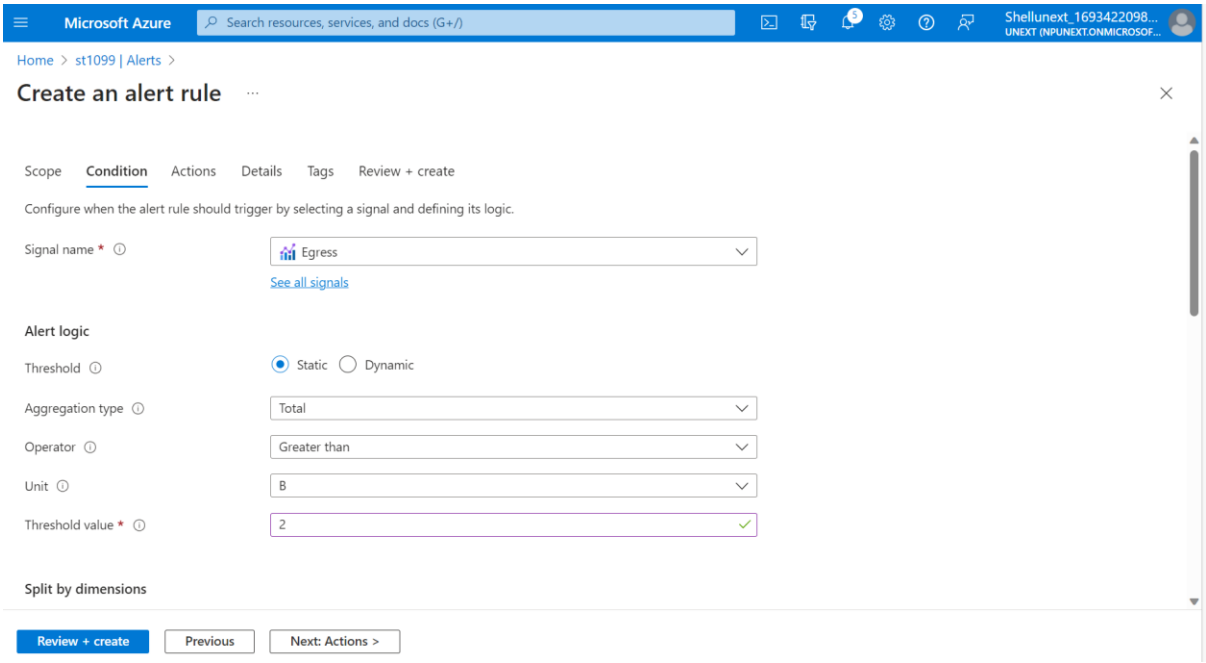



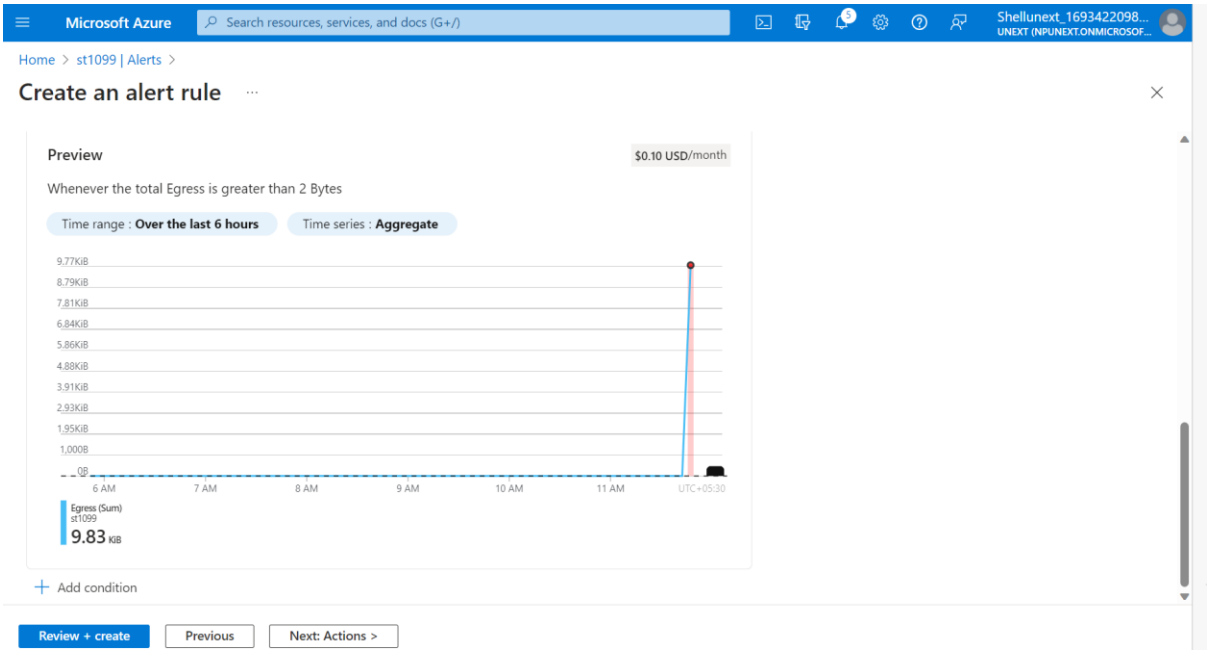

After enabled static website, give index.html and error.html name and save



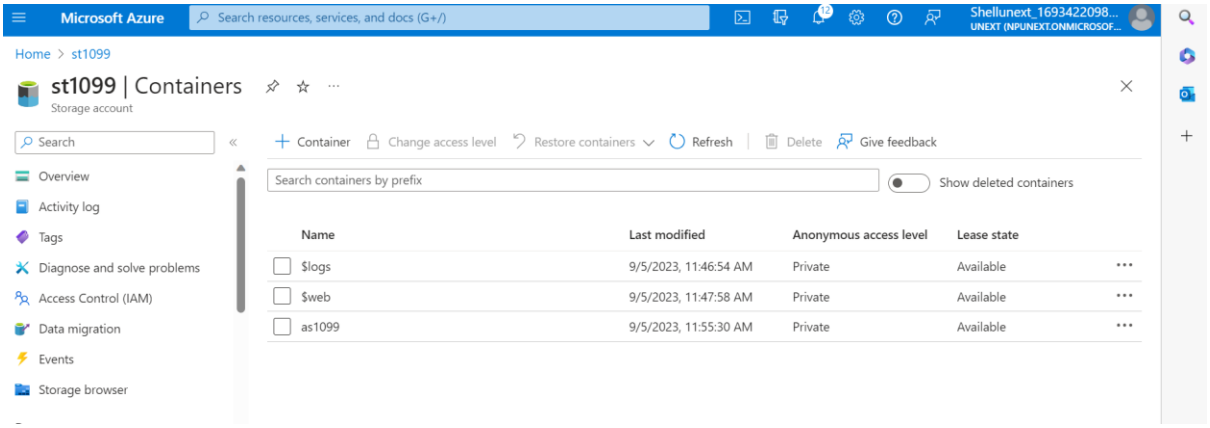

Then click on $web and upload the files

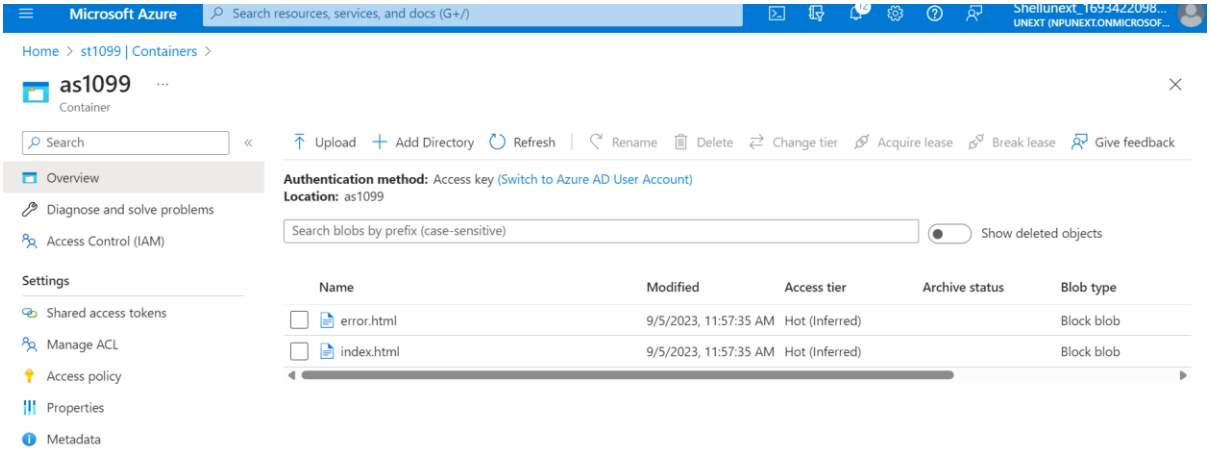
Creation of alert



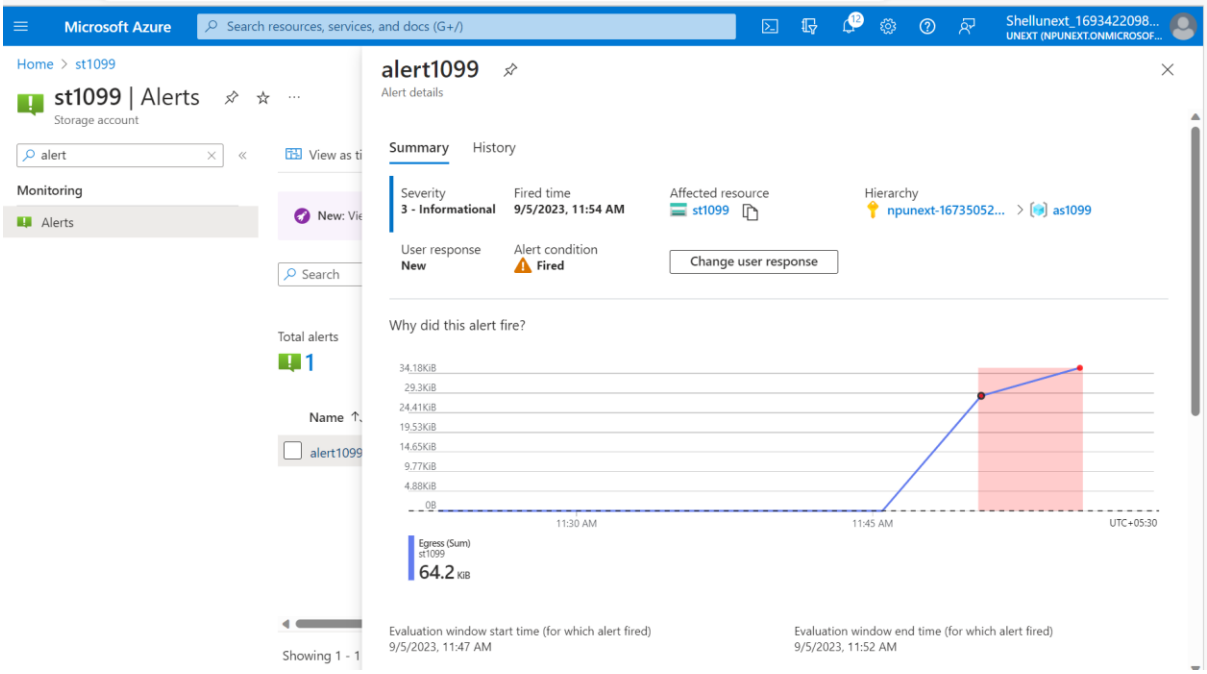



In containers make a new container file and in $web and that container upload the index.html and error.html files and these files can be viewed through static website links provided.

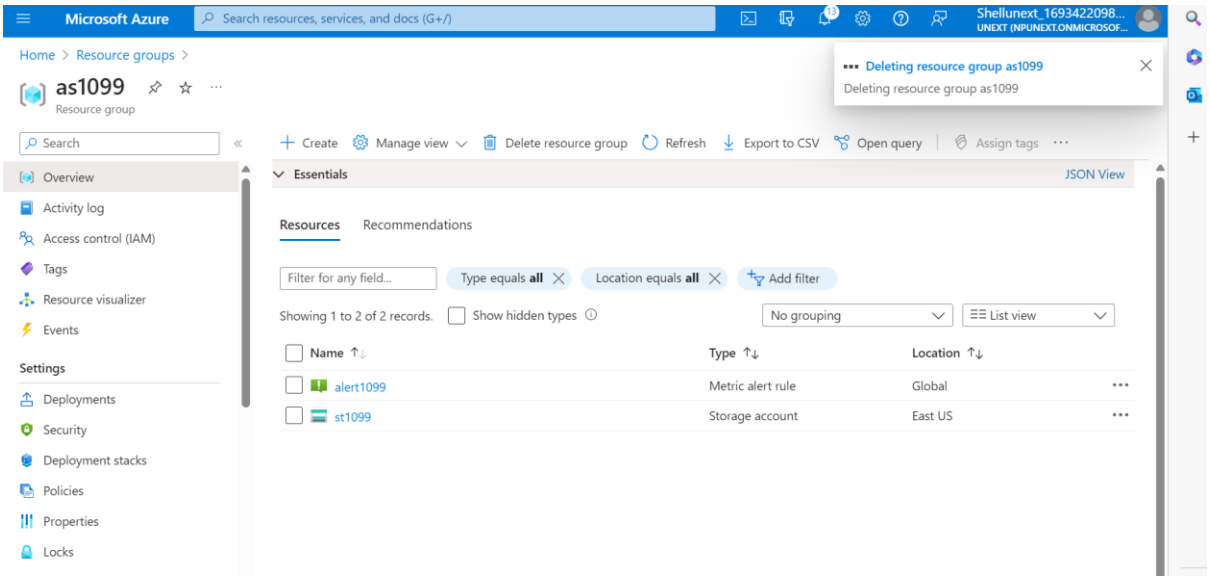


After that you can see the fired alert in the alert section.

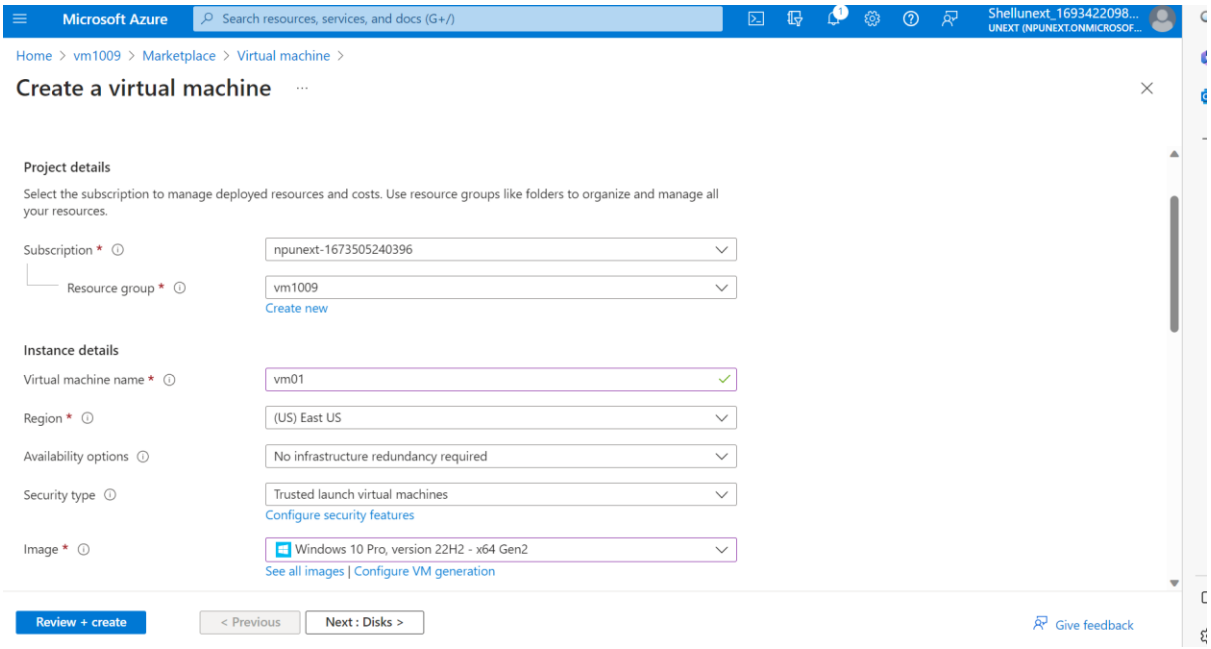


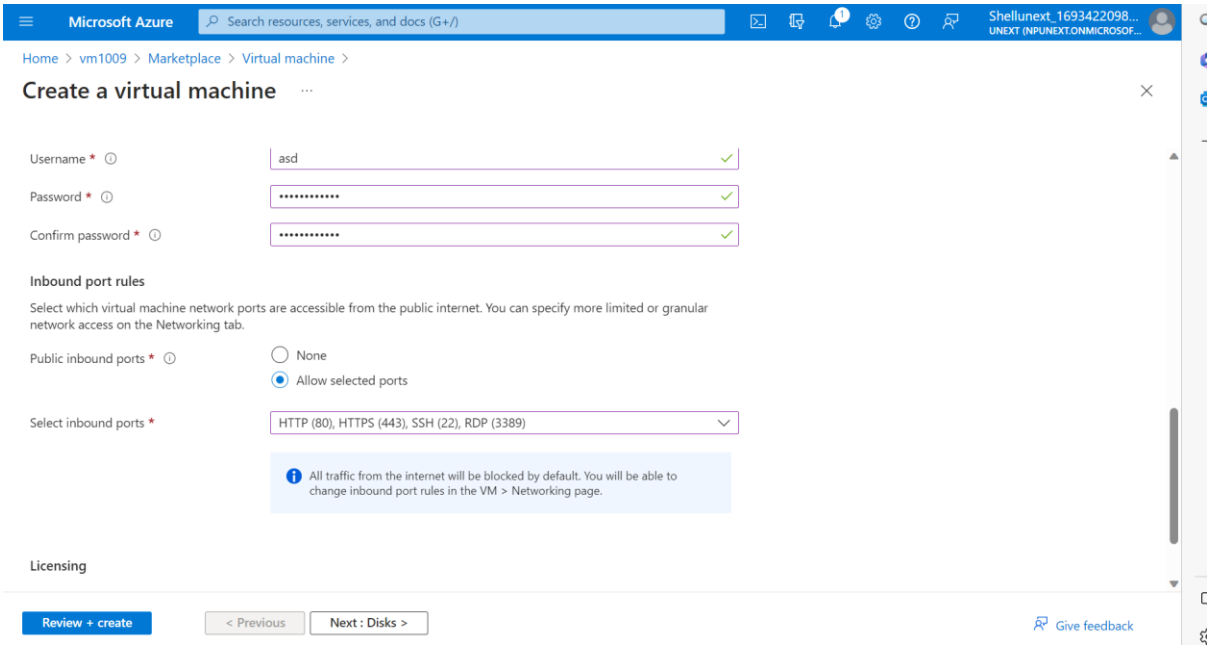

Deleting a resource group

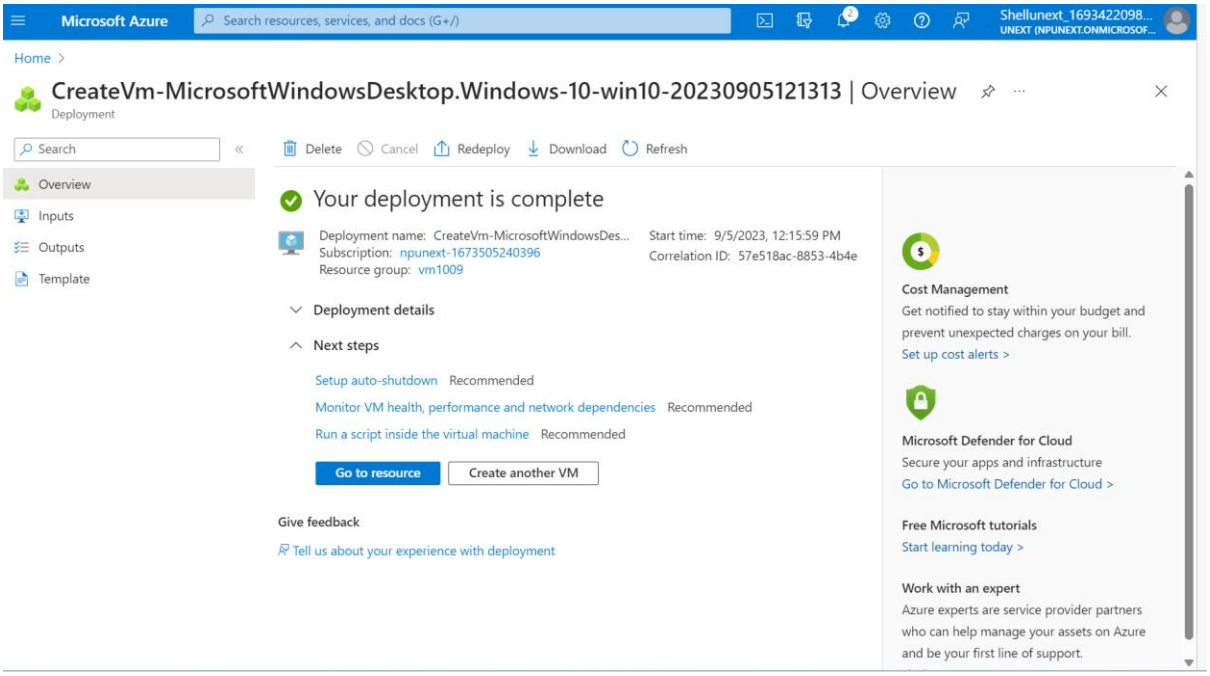



Mstsc- server connector

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

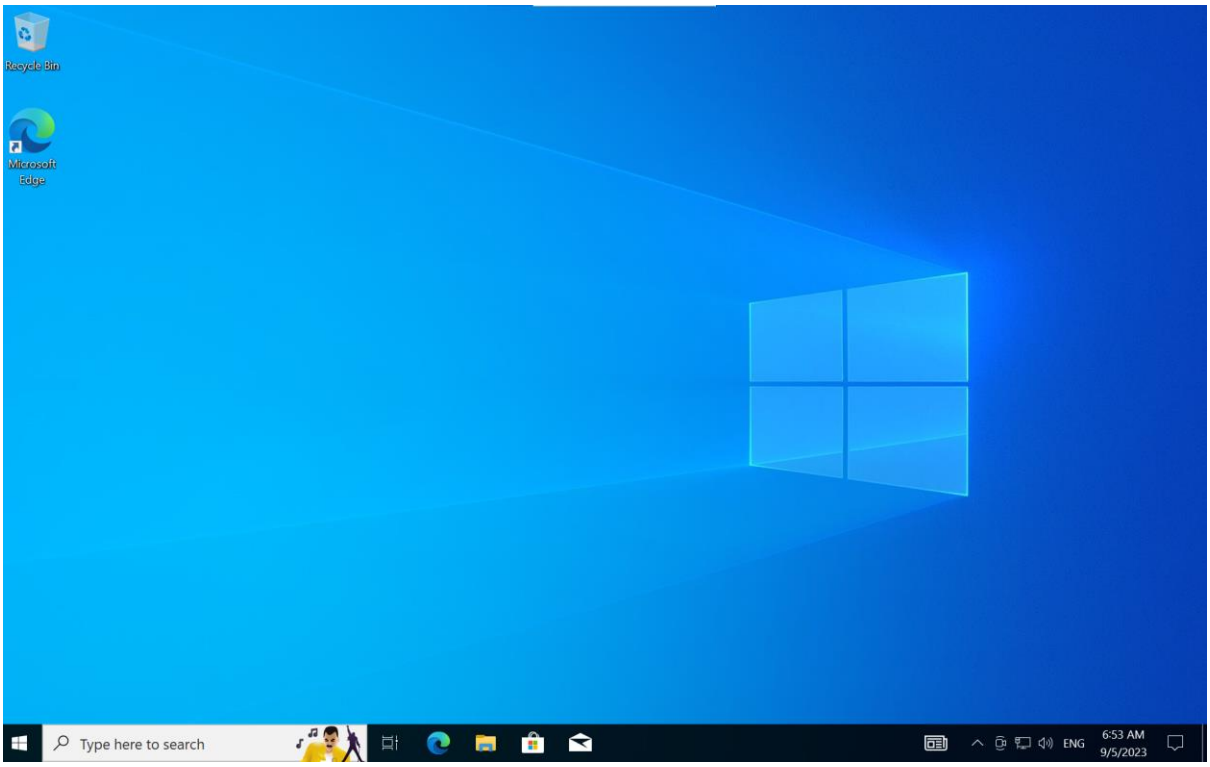

Microsoft Azure
Search resources, services, and docs (G+/)
Shellunext_1693422098...
UNEXT (NPUNEXT.ONMICROSOF...
Home > vm1009 > Marketplace > Virtual machine >
Create a virtual machine
Project details
Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.
Subscription *
npunext-1673505240396
Resource group *
vm1009
Create new
Instance details
Virtual machine name *
vm01
Region *
(US) East US
Availability options
No infrastructure redundancy required
Security type
Trusted launch virtual machines
Configure security features
Image *
Windows 10 Pro, version 22H2 - x64 Gen2
See all images | Configure VM generation
Review + create
< Previous
Next : Disks >
Give feedback


Microsoft Azure
Search resources, services, and docs (G+/)
Shellunext_1693422098...
UNEXT (NPUNEXT.ONMICROSOF...
Home > vm1009 > Marketplace > Virtual machine >
Create a virtual machine
Username *
asd
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 *
HTTP (80), HTTPS (443), SSH (22), RDP (3389)
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




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

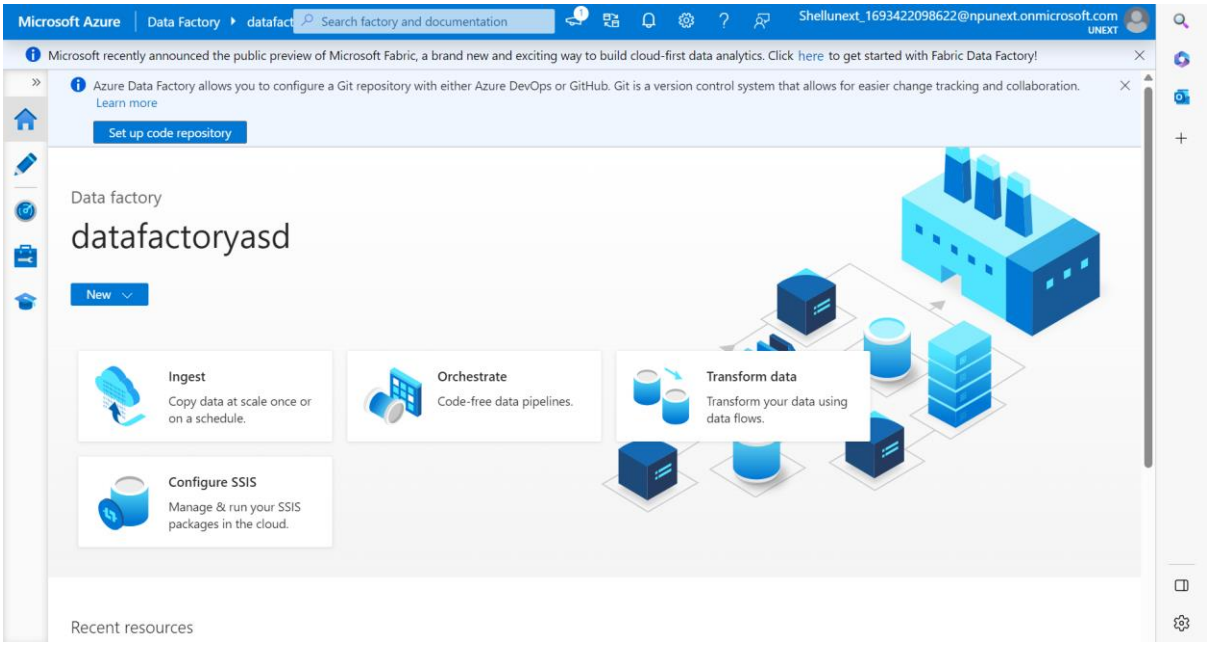
You would see this next




Open cmd and download

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

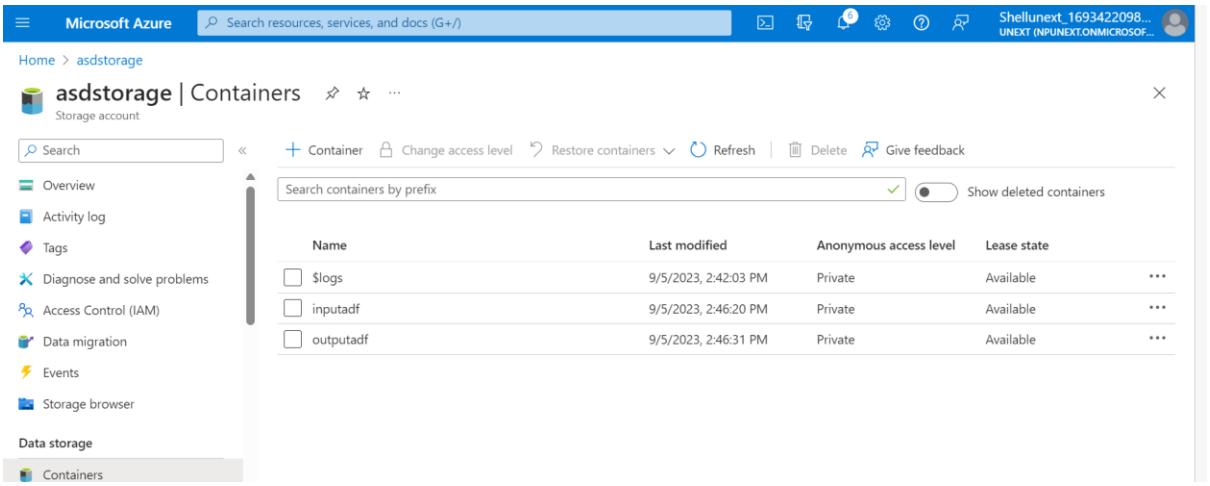
Server list/link and adf

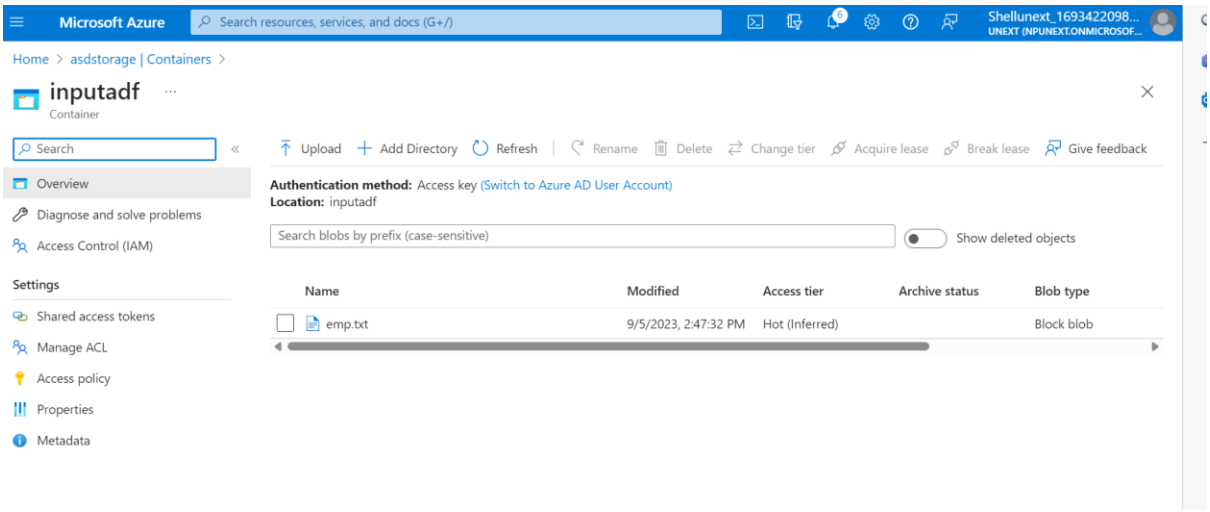



Created a adf under resource group

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

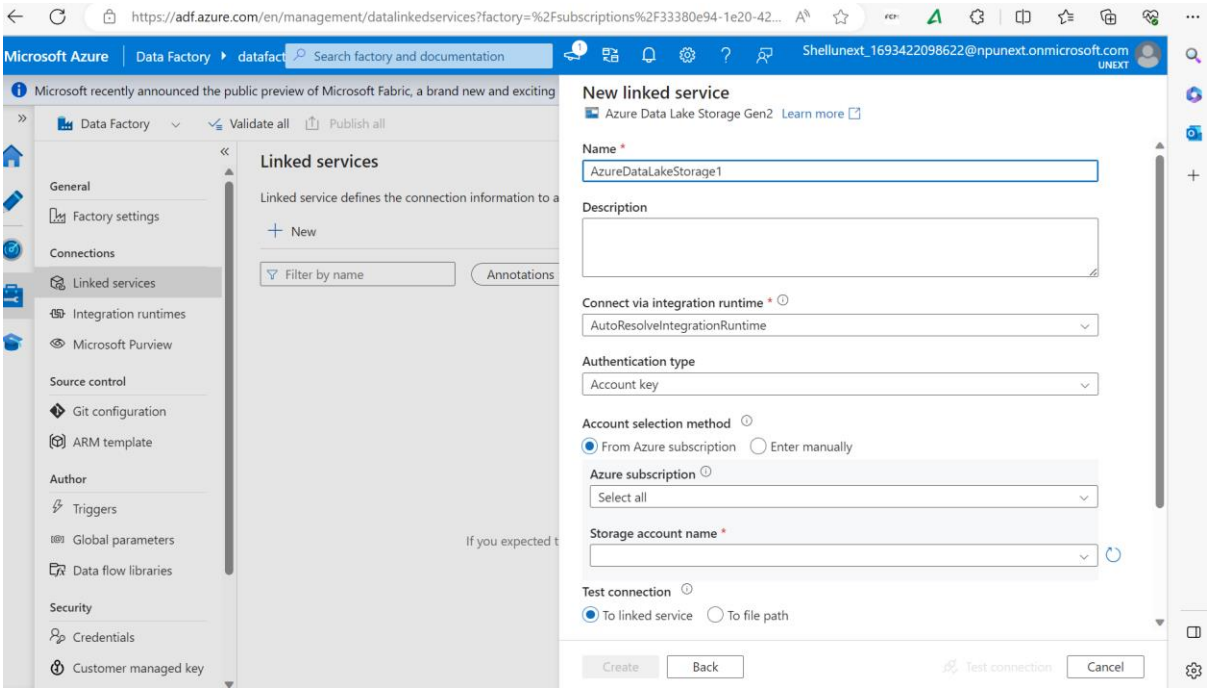
Basically create- RG, ADF, SA, 2 CONTAINERS(input and output), Emp.txt, link services for storage account in adf under manage tab, adf-author-select pipeline-copy task-from where you want to copy and paste, then just execute and check

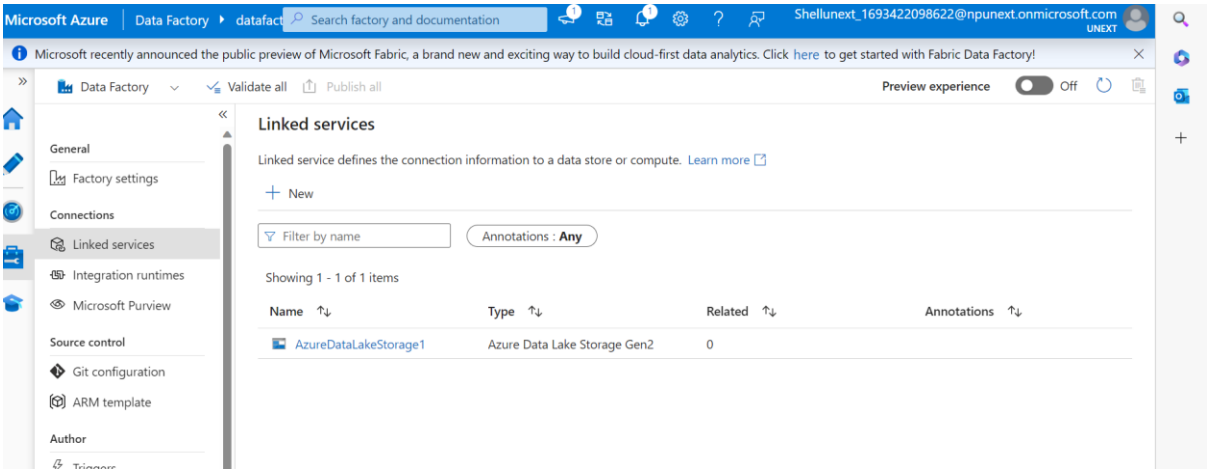





Basic sense is through adf we need to transfer input txt file to output container.

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



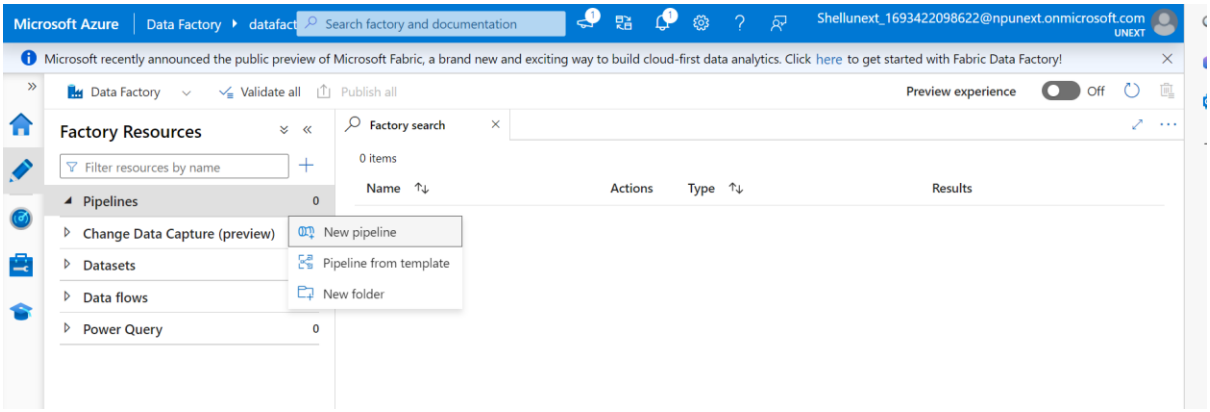



8.1-pipeline

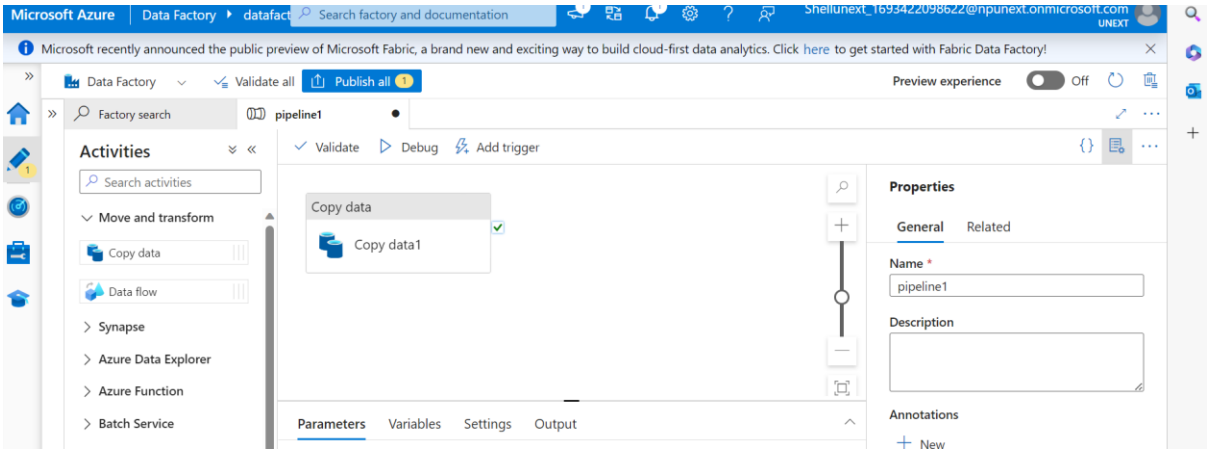
8.2-activity you want to execute

8.3-source

8.4-sink

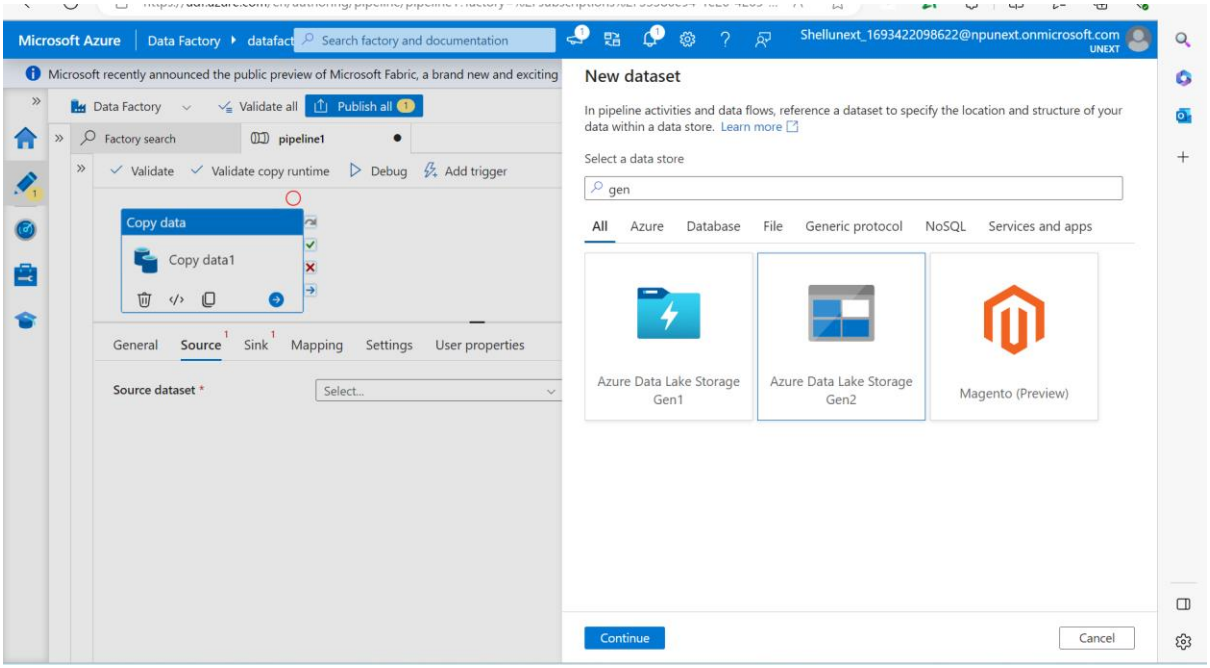


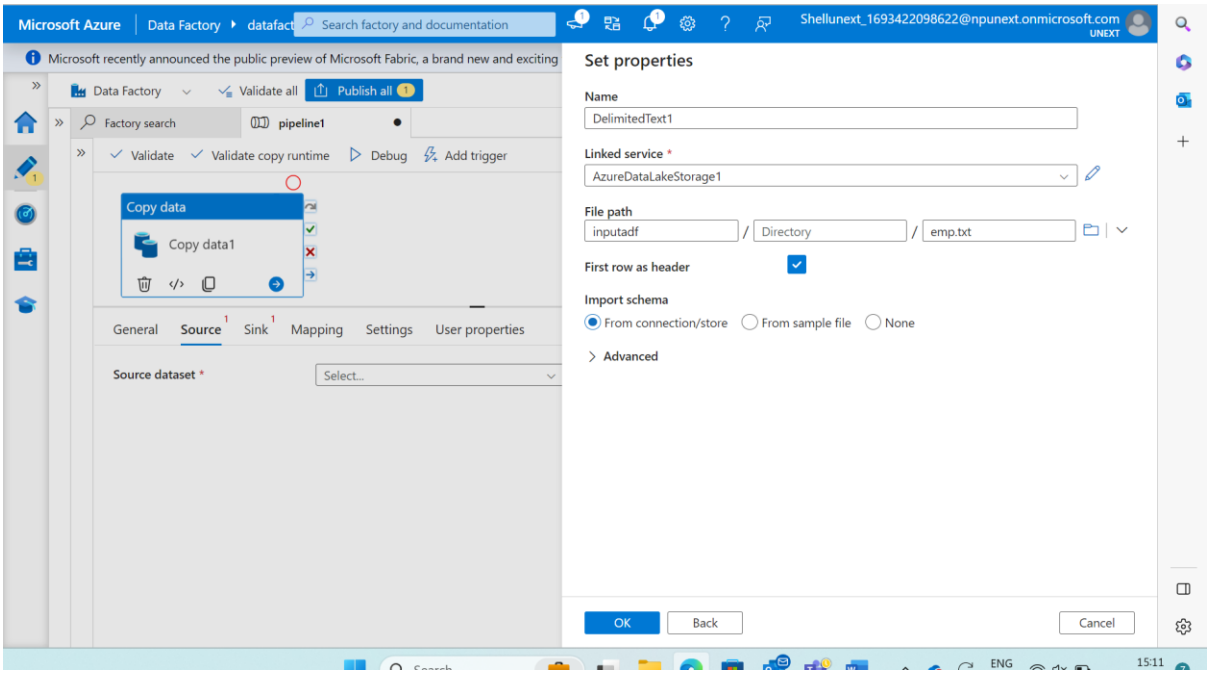

Under move and transform drag and drop copy data. Now 8.1 and 8.2 is done.

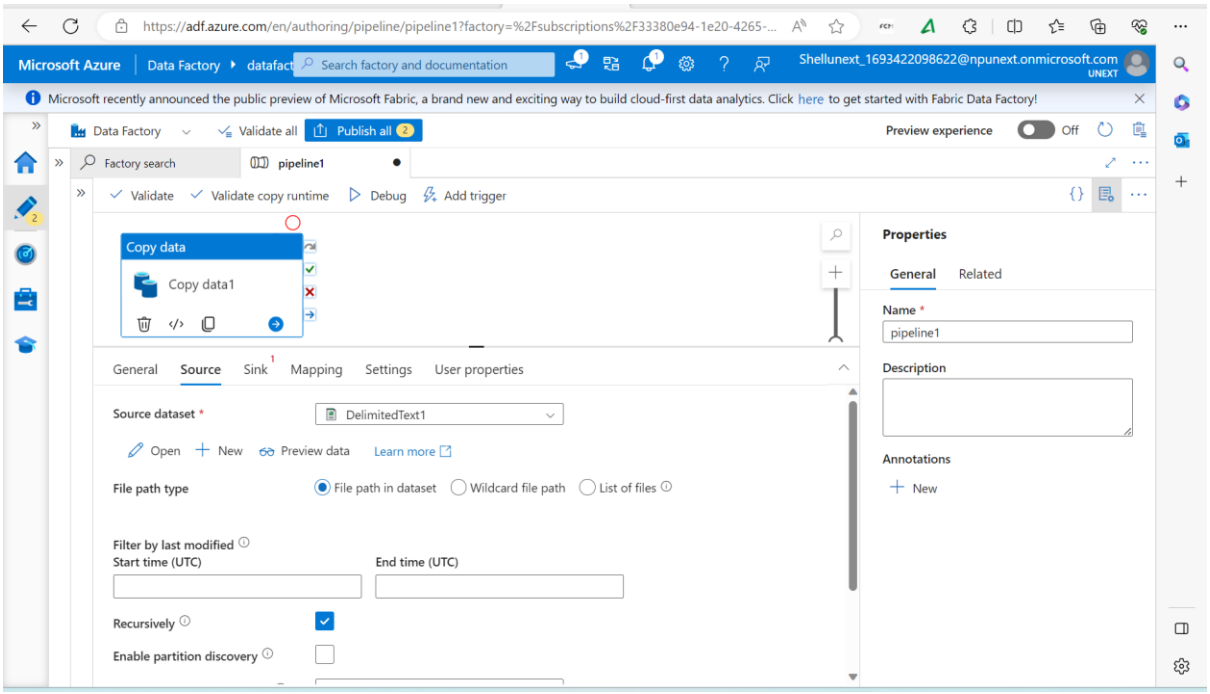



For 8.3 and 8.4 click on copy data and then under source click to make new dataset under gen2.

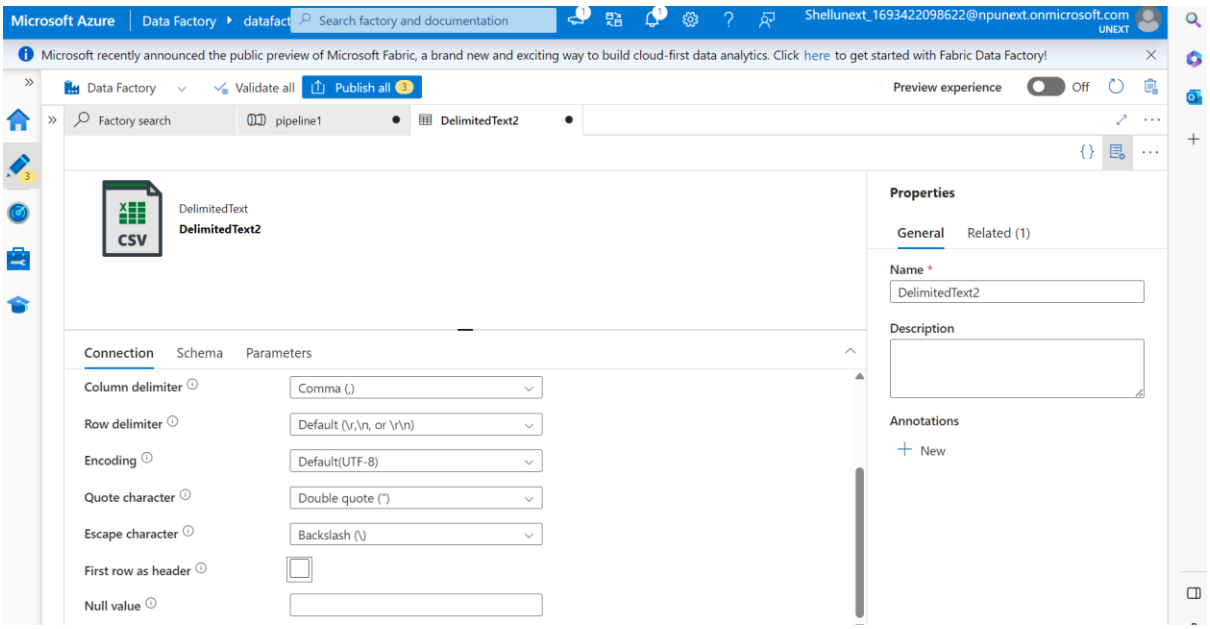
Under it select csv file type and from it select linked service azure file and under file path browsing select input file emp text and click okay.






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

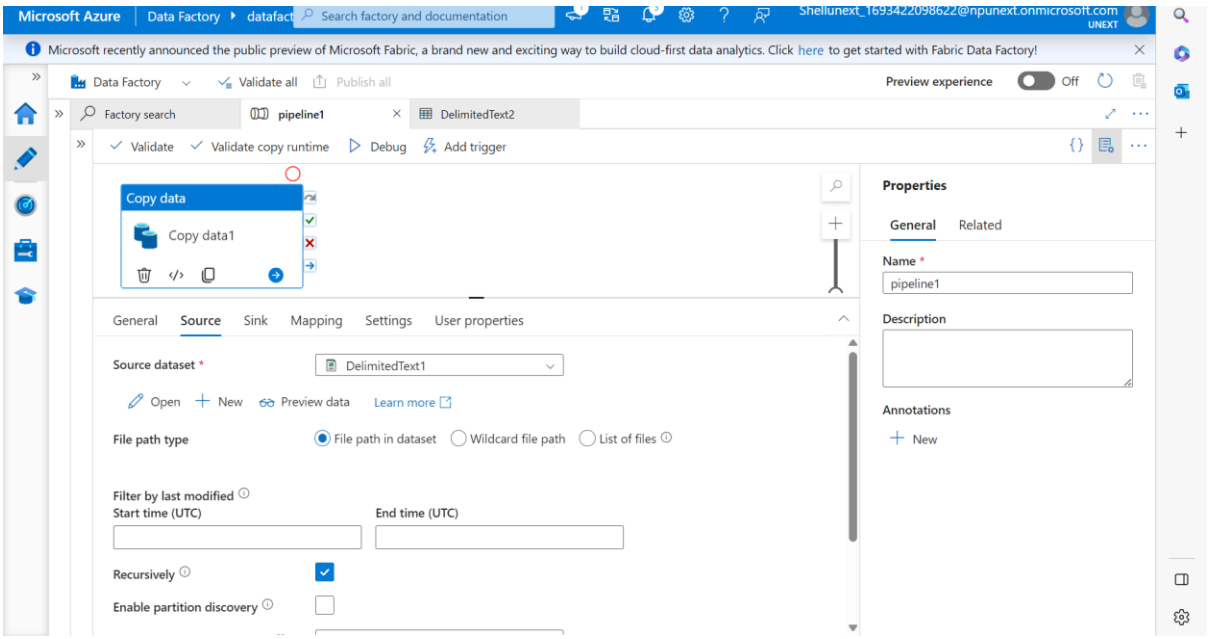


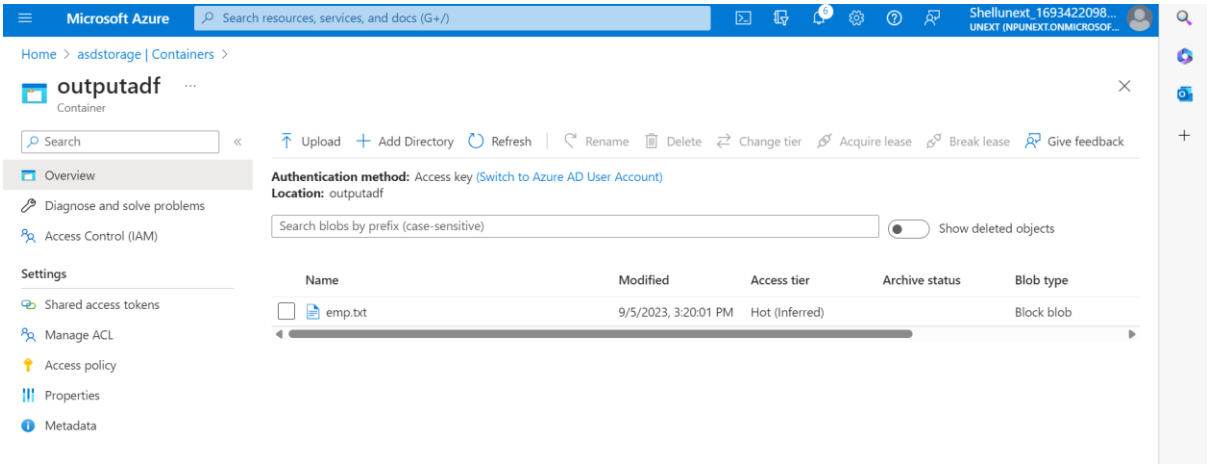

Now in open file deselect header checkbox.

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.




DAY-6TH September 2023

ADF-

Topic- DataOps

Azure has multiple formats-UI(need portal.azure.com), CLI(need Azure 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

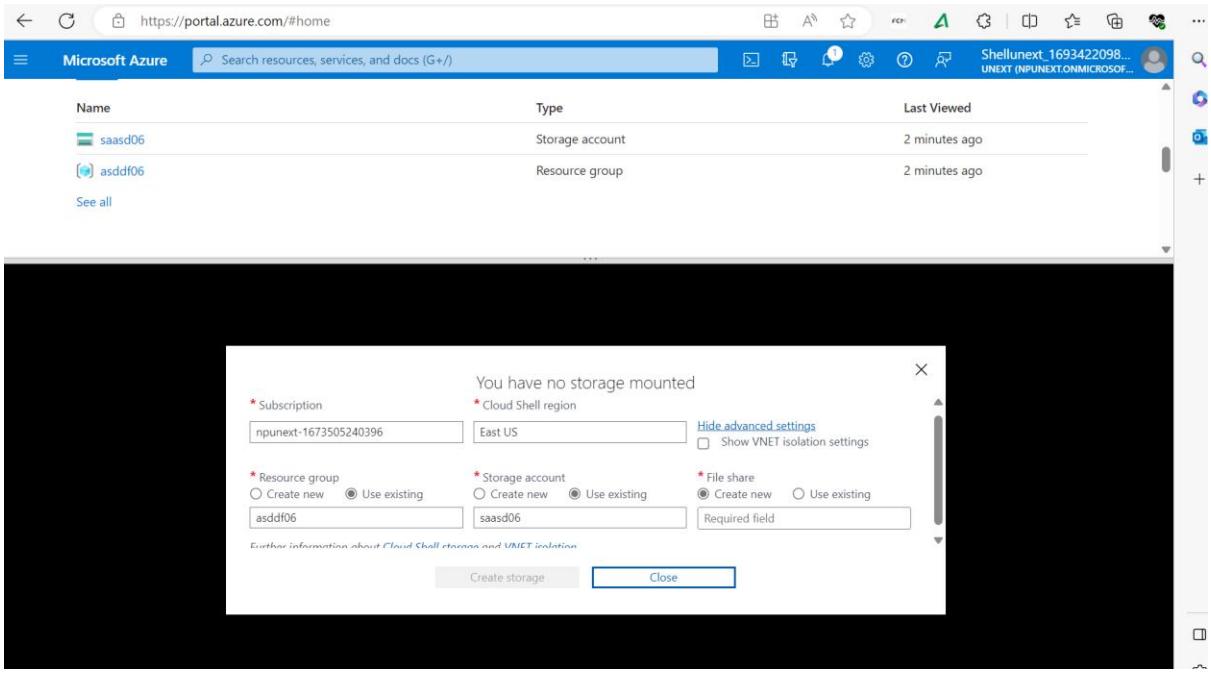
Commands-

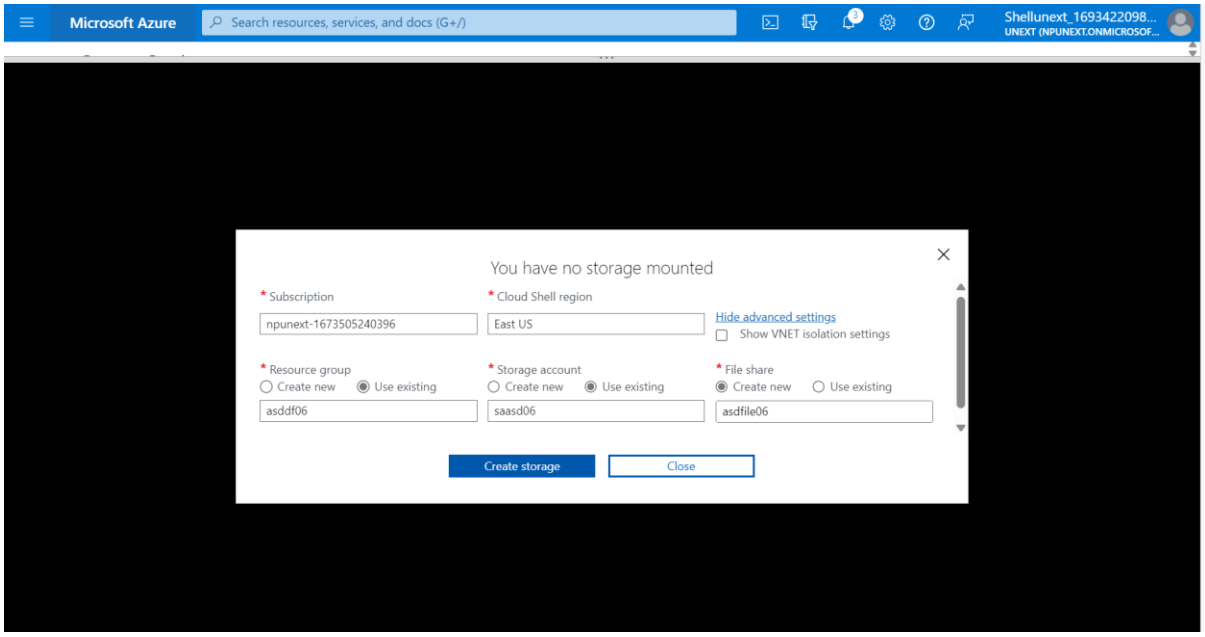
- az login
- CloudVendorCode(az) \
  ServiceName(ResGrp) \
  FeatureName[] \
  OperationName(Create/Delete/Upload) \
  Properties(--Key Value) \
  Properties(--Key Value) \
  Properties(--Key Value) \
  Properties(--Key Value)
- az group create \
  --name asdrg \
  --location eastus
- az \ ->used to 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 containerasd06 \
  --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 containerasd06 \
--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






Git hub

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

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

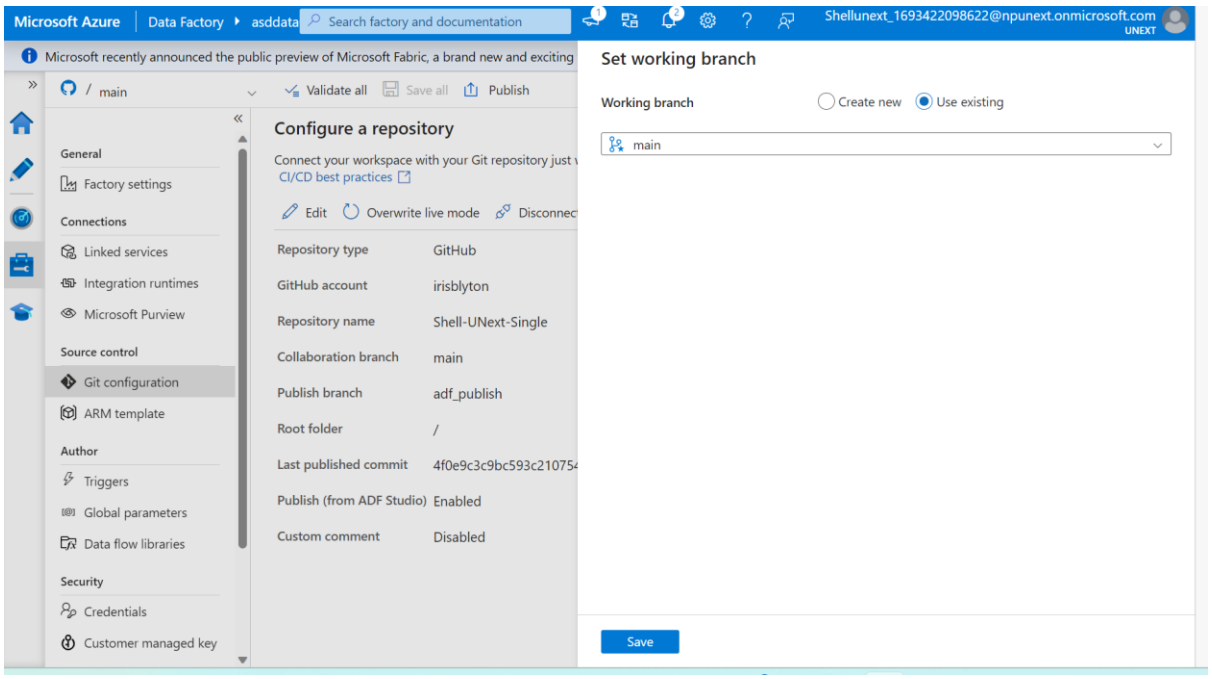
Now go to code and copy url

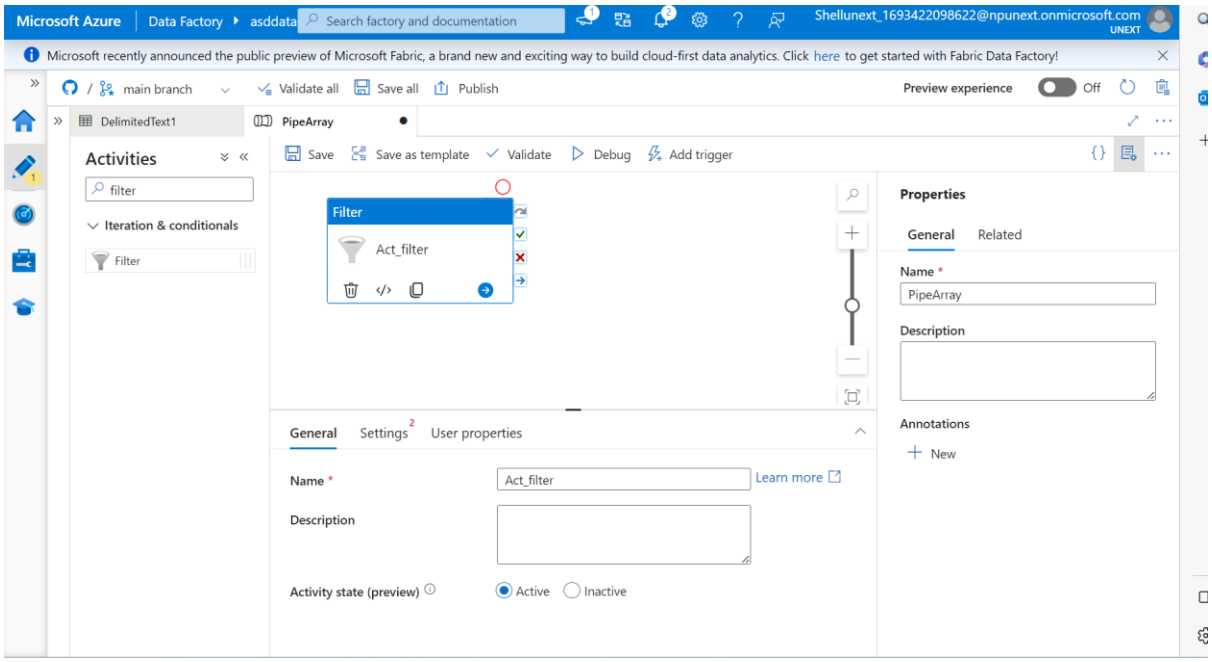
Go to adf under git configuration->repository owner mein write github name

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.

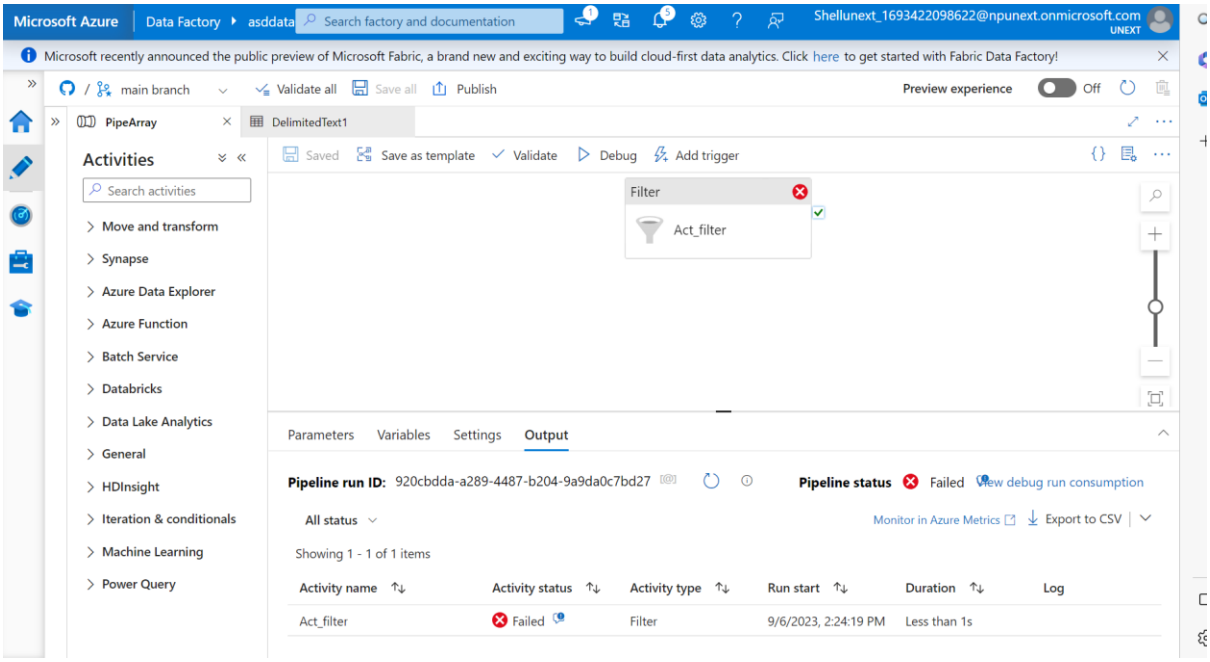
Then click on save and the next screen shot happened.

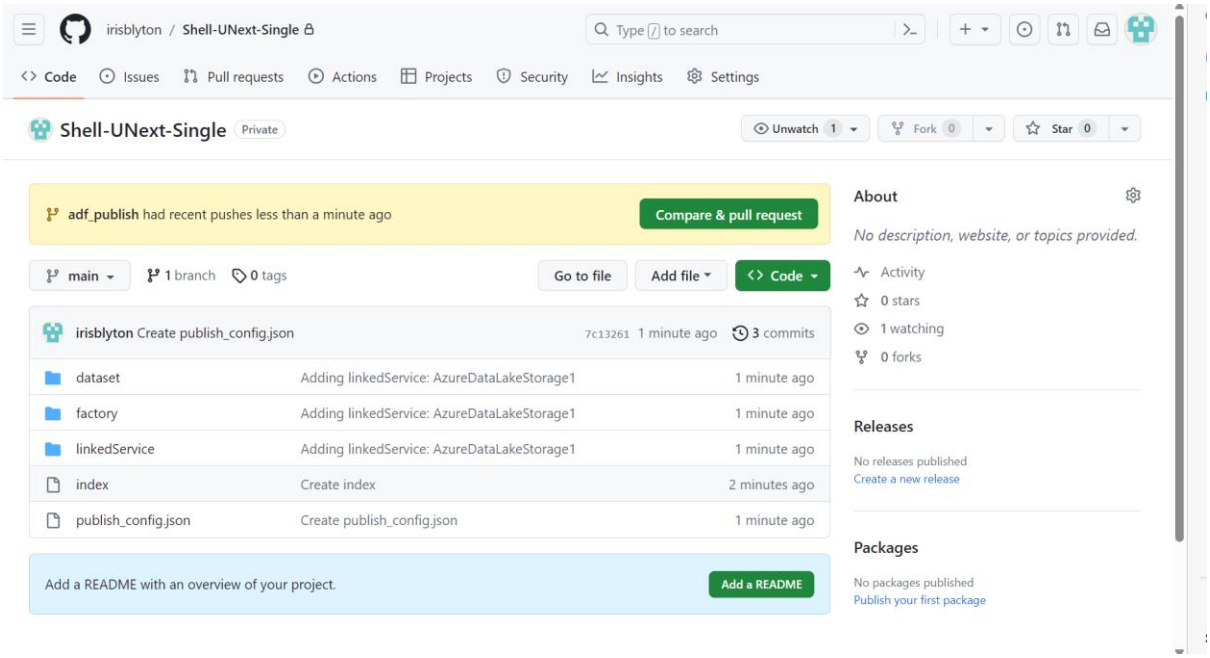


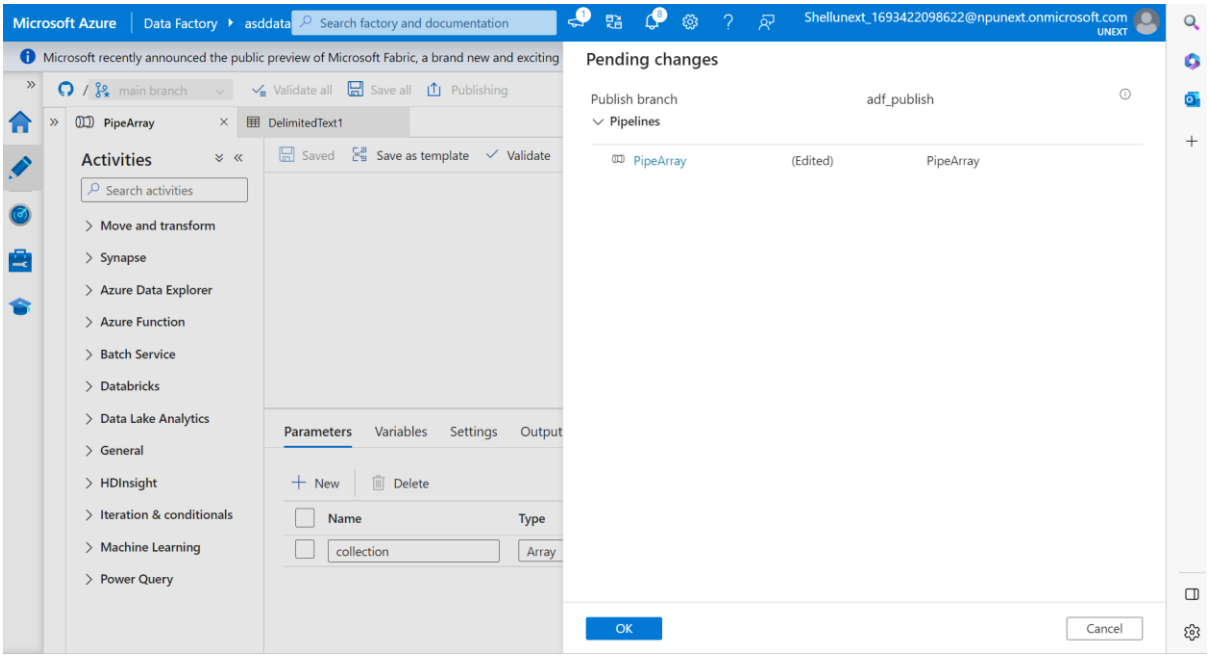



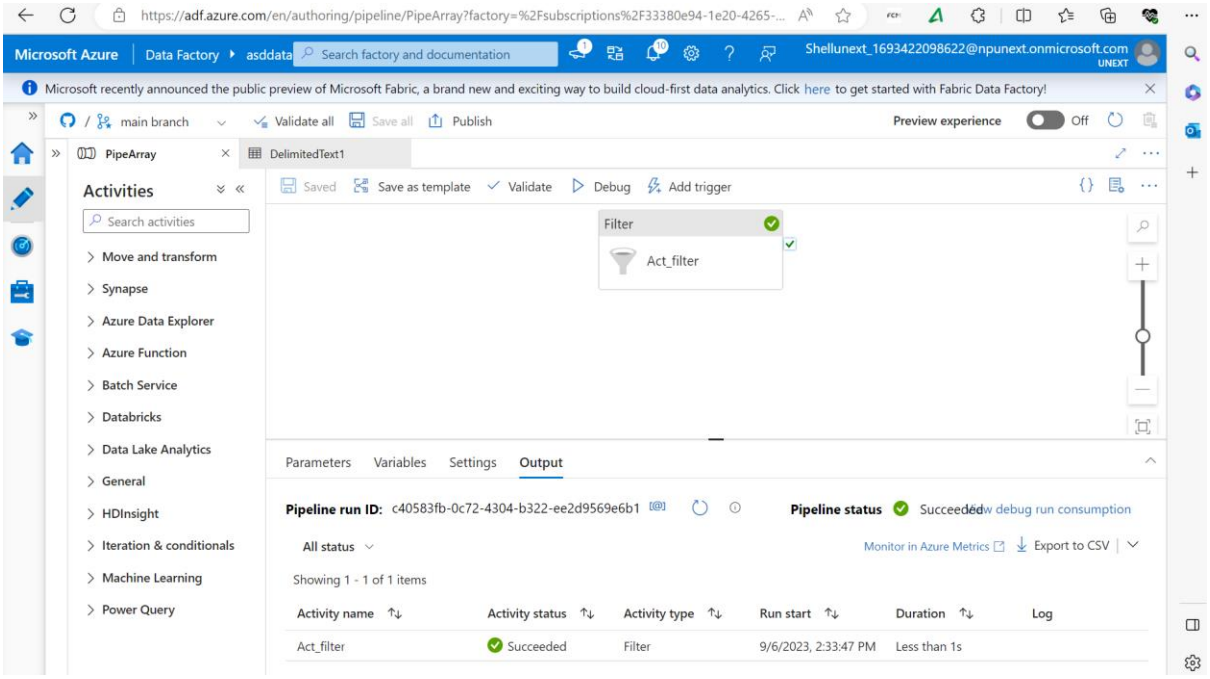

Now after making a new pipeline-

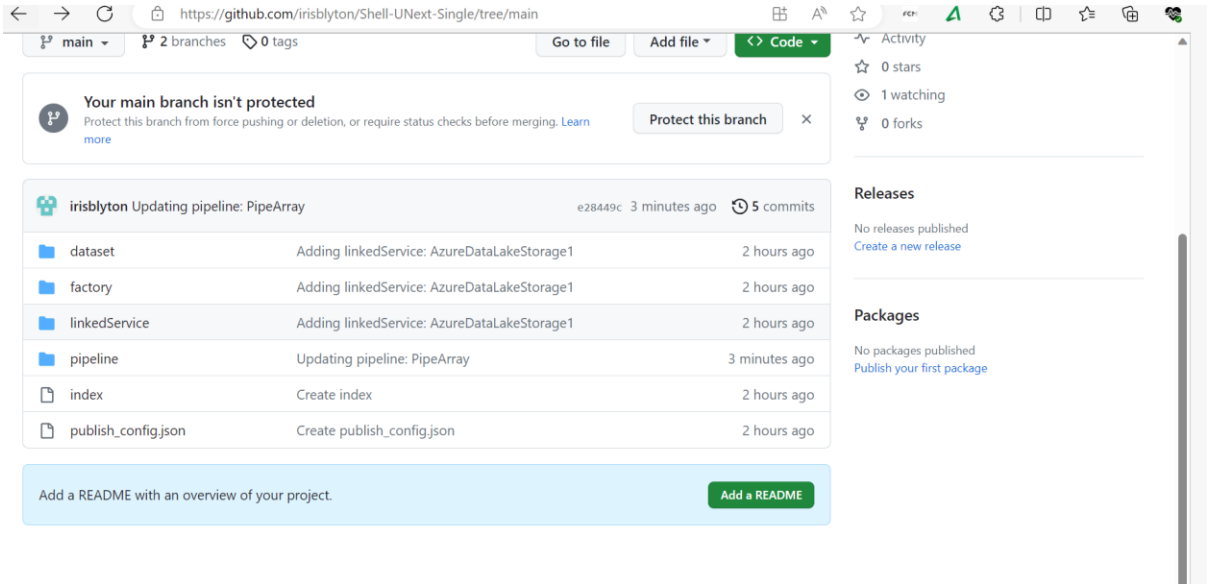
- Pipeline->variable section->name-data array, type-array, default-[“aaa”,”bbb”,”ccc”, “ddd”]
- Picture clicked on mobile whatsapp

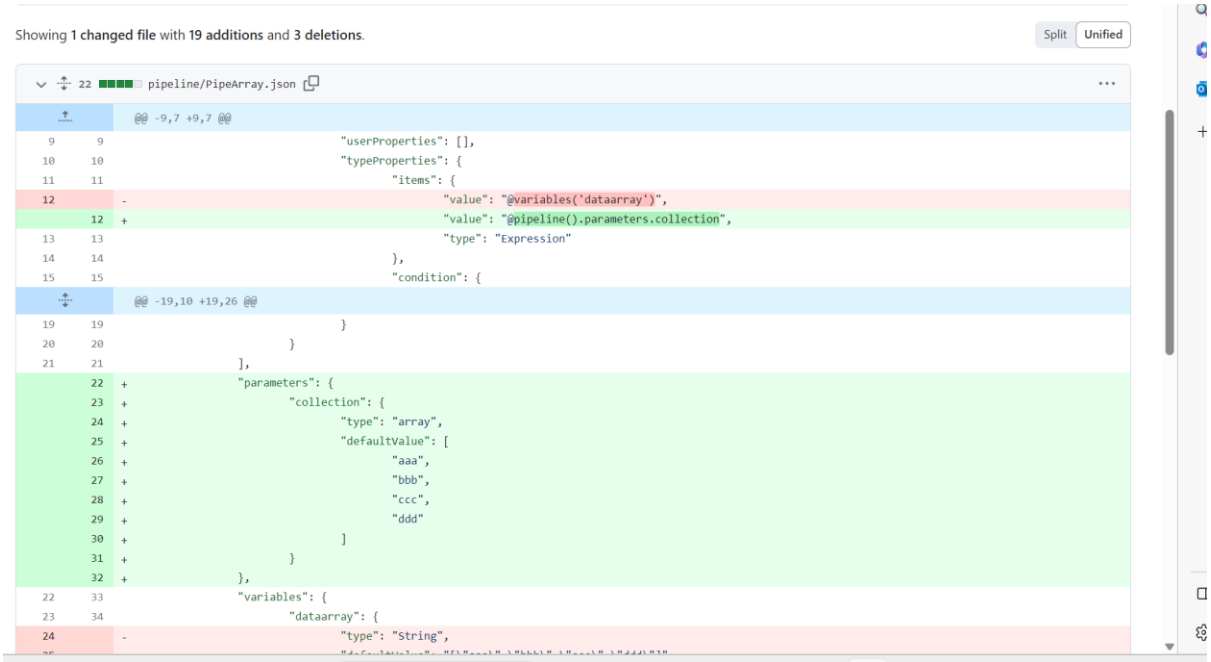

Microsoft Azure
Data Factory
asddata
Search factory and documentation
Shellunext_1693422098622@npunext.onmicrosoft.com
UNEXT
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!
main branch
Validate all
Save all
Publish
Preview experience
Off
PipeArray
DelimitedText1
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Saved
Save as template
Validate
Debug
Add trigger
Filter
Act_filter
Parameters
Variables
Settings
Output
Pipeline run ID: 920cbdda-a289-4487-b204-9a9da0c7bd27
Pipeline status
Failed
View debug run consumption
All status
Monitor in Azure Metrics
Export to CSV
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

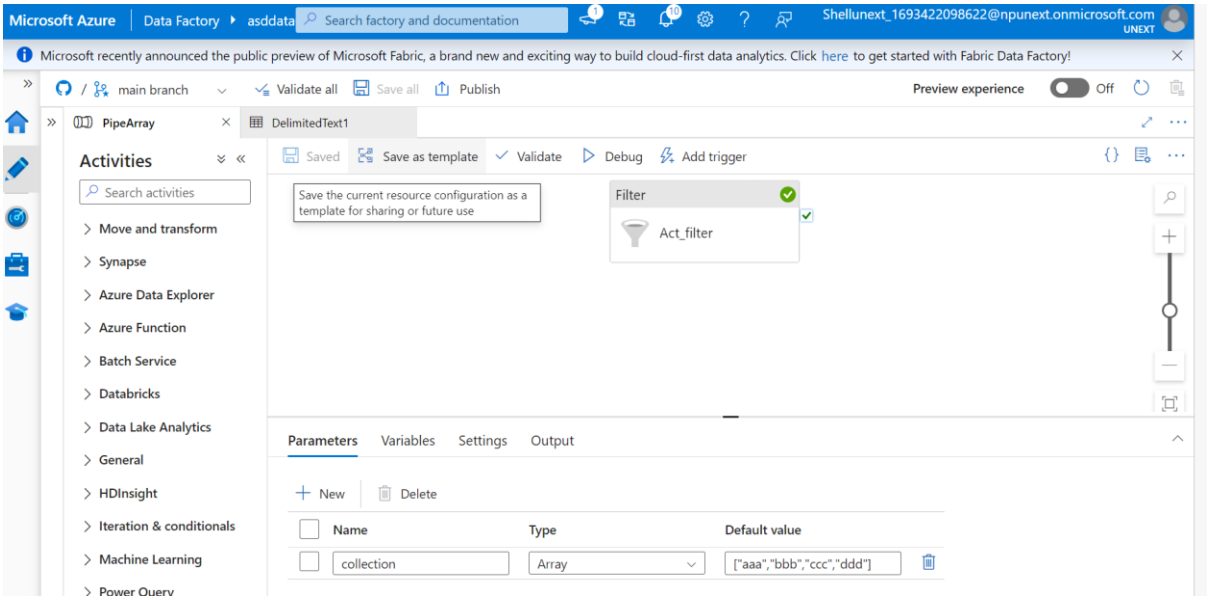

irisblyton / Shell-UNext-Single
Type / to search
Code
Issues
Pull requests
Actions
Projects
Security
Insights
Settings
Shell-UNext-Single
Private
Unwatch 1
Fork 0
Star 0
adf_publish had recent pushes less than a minute ago
Compare & pull request
main
1 branch
0 tags
Go to file
Add file
Code
irisblyton 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
Add a README with an overview of your project.
Add a README
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

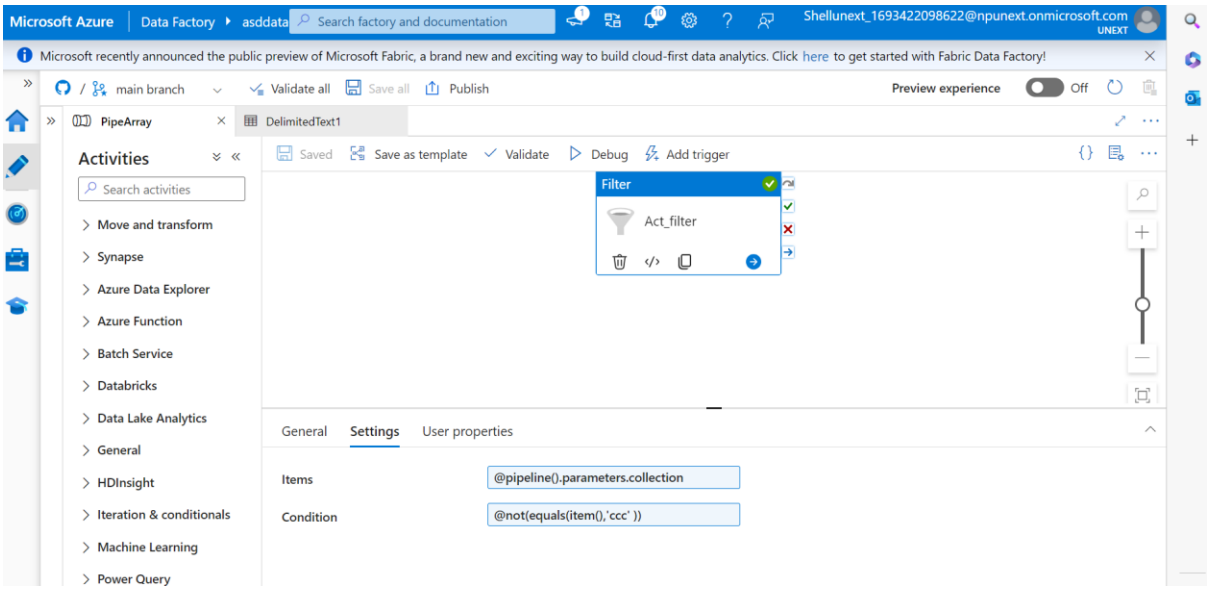

Microsoft Azure
Data Factory
asddata
Search factory and documentation
Shellunext_1693422098622@npunext.onmicrosoft.com
UNEXT
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting
main branch
Validate all
Save all
Publishing
PipeArray
DelimitedText1
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Saved
Save as template
Validate
Parameters
Variables
Settings
Output
New
Delete
Name
Type
collection
Array
Pending changes
Publish branch
adf_publish
Pipelines
PipeArray
(Edited)
PipeArray
OK
Cancel

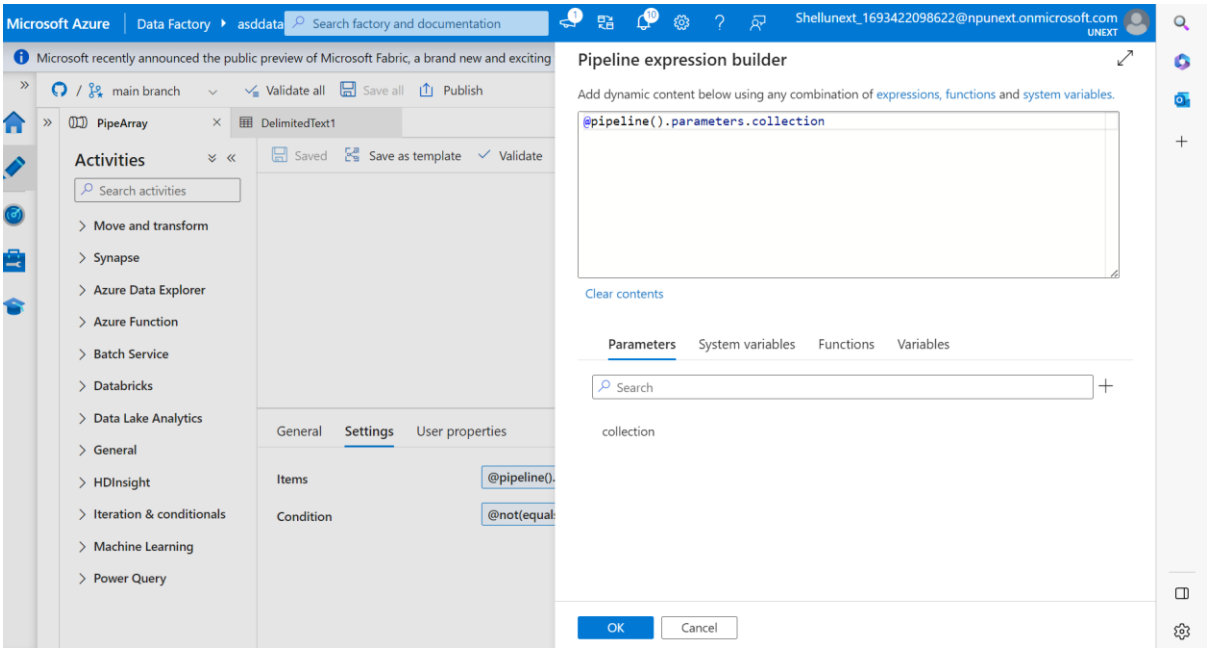

https://adf.azure.com/en/authoring/pipeline/PipeArray?factory=%2Fsubscriptions%2F33380e94-1e20-4265-...
Microsoft Azure
Data Factory
asddata
Search factory and documentation
Shellunext_1693422098622@npunext.onmicrosoft.com
UNEXT
Microsoft recently announced the public preview of Microsoft Fabric, a brand 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
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Saved
Save as template
Validate
Debug
Add trigger
Filter
Act_filter
Parameters
Variables
Settings
Output
Pipeline run ID: c40583fb-0c72-4304-b322-ee2d9569e6b1
Pipeline status
Succeeded
w debug run consumption
All status
Monitor in Azure Metrics
Export to CSV
Showing 1 - 1 of 1 items
Activity name
Activity status
Activity type
Run start
Duration
Log
Act_filter
Succeeded
Filter
9/6/2023, 2:33:47 PM
Less than 1s

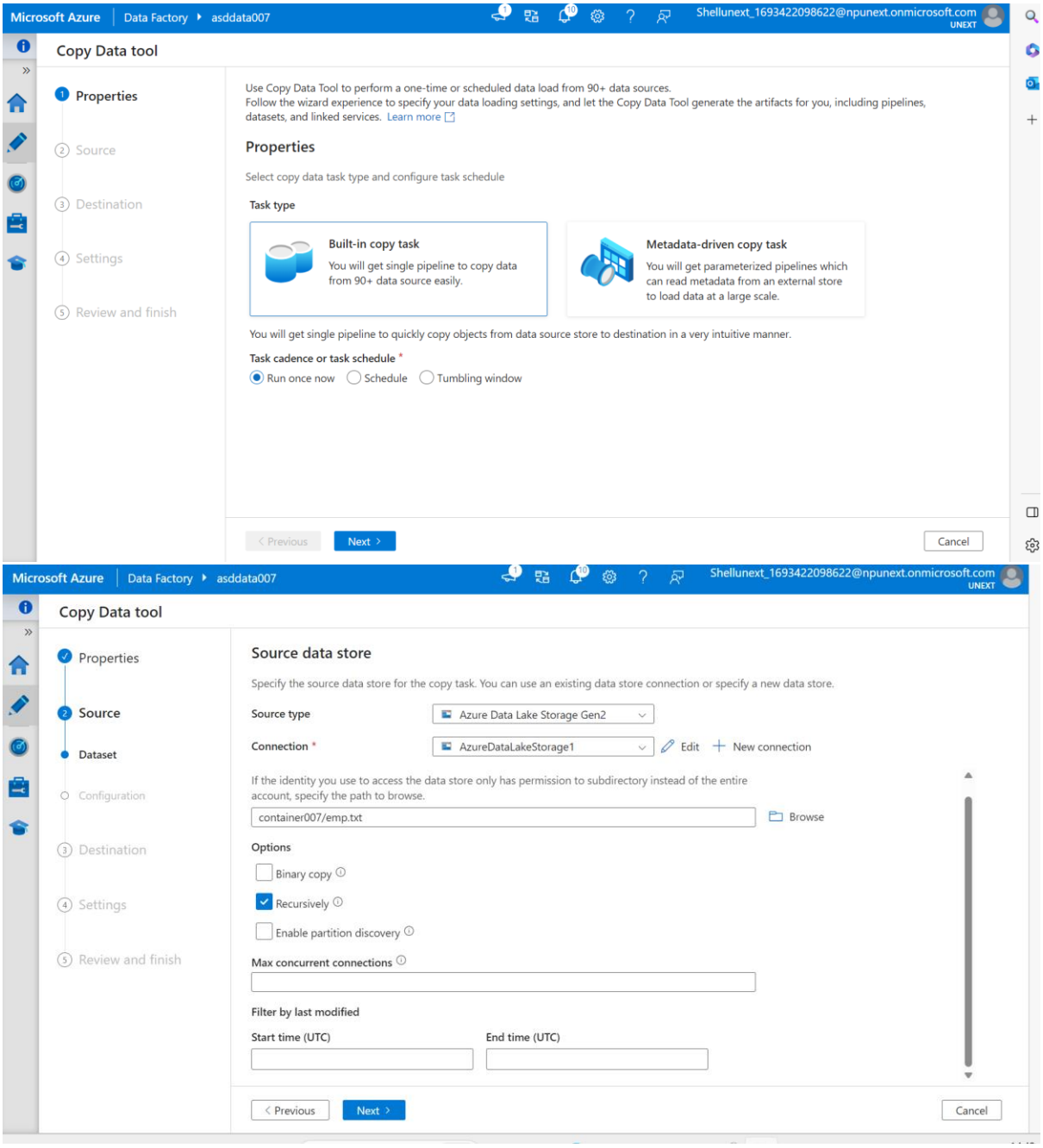

https://github.com/irisblyton/Shell-UNext-Single/tree/main
main
2 branches
0 tags
Go to file
Add file
Code
Your main branch isn't protected
Protect this branch from force pushing or deletion, or require status checks before merging. Learn more
Protect this branch
irisblyton Updating pipeline: PipeArray
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
Activity
0 stars
1 watching
0 forks
Releases
No releases published
Create a new release
Packages
No packages published
Publish your first package


Showing 1 changed file with 19 additions and 3 deletions.
Split
Unified
22
pipeline/PipeArray.json
@@ -9,7 +9,7 @@
"userProperties": [],
"typeProperties": {
"items": {
- "value": "@variables('dataarray')",
+ "value": "@pipeline().parameters.collection",
"type": "Expression"
},
"condition": {
@@ -19,10 +19,26 @@
}
}
],
+ "parameters": {
+ "collection": {
+ "type": "array",
+ "defaultValue": [
+ "aaa",
+ "bbb",
+ "ccc",
+ "ddd"
+ ]
+ }
+ },
"variables": {
"dataarray": {
- "type": "String",

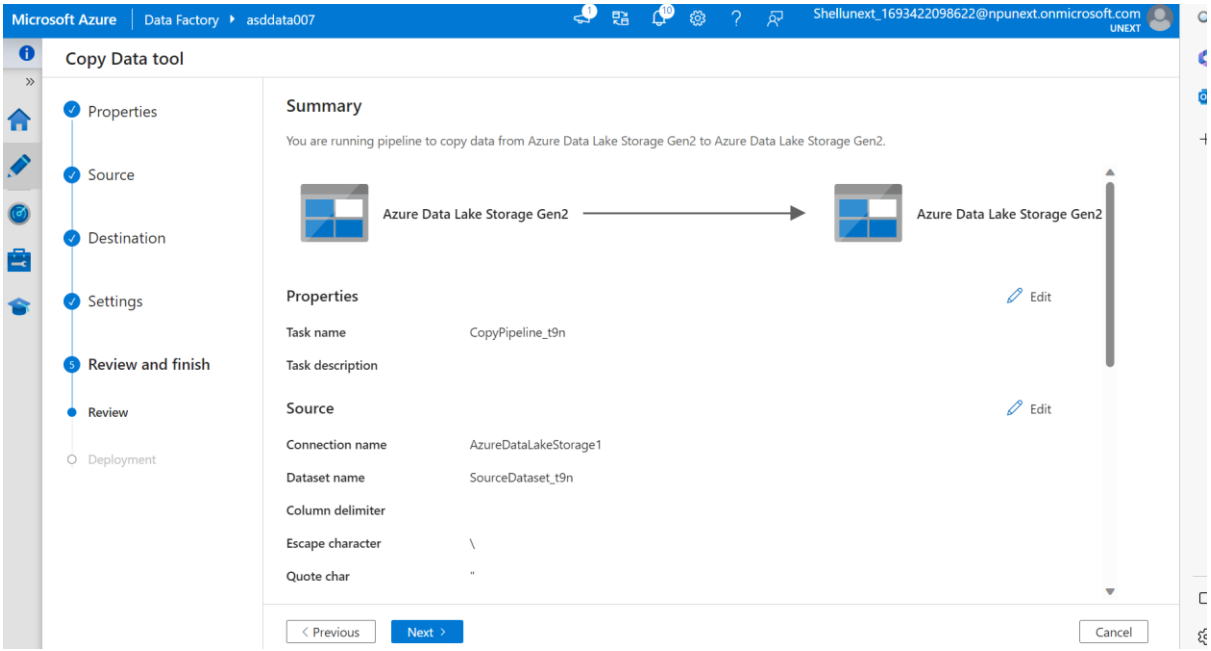

Microsoft Azure
Data Factory
asddata
Search factory and documentation
Shellunext_1693422098622@npunext.onmicrosoft.com
UNEXT
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!
main branch
Validate all
Save all
Publish
Preview experience
Off
PipeArray
DelimitedText1
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Saved
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"]

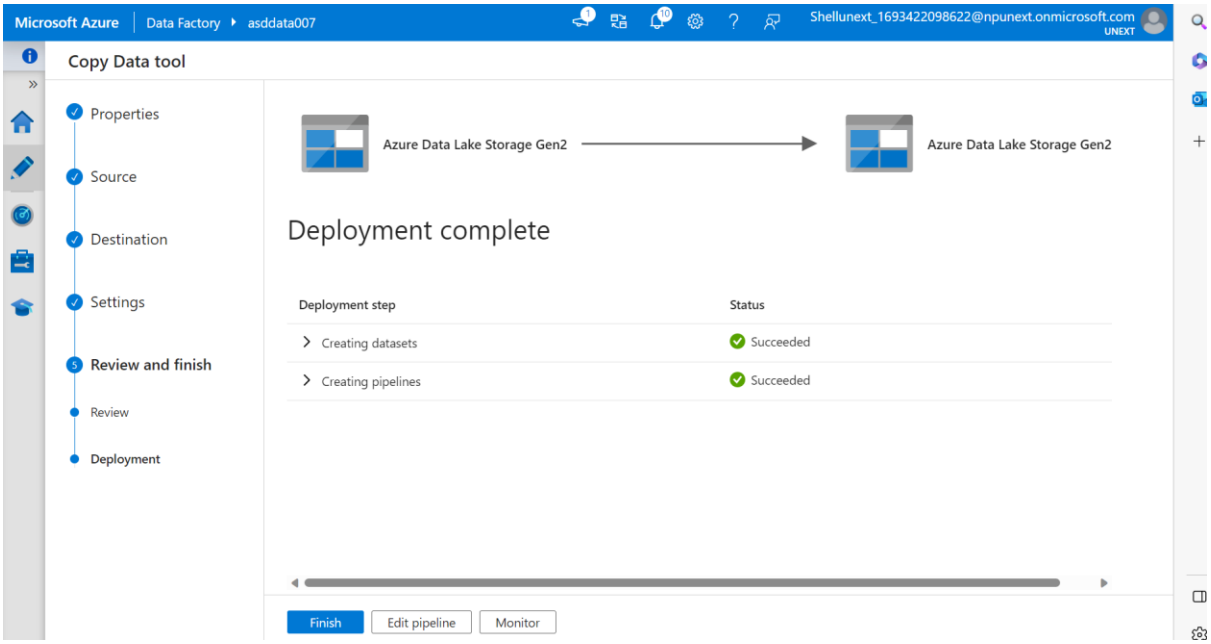

Microsoft Azure
Data Factory
asddata
Search factory and documentation
Shellunext_1693422098622@npunext.onmicrosoft.com
UNEXT
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!
main branch
Validate all
Save all
Publish
Preview experience
Off
PipeArray
DelimitedText1
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Saved
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
asddata
Search factory and documentation
Shellunext_1693422098622@npunext.onmicrosoft.com
UNEXT
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting
main branch
Validate all
Save all
Publish
PipeArray
DelimitedText1
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Saved
Save as template
Validate
General
Settings
User properties
Items
@pipeline().
Condition
@not(equal
Pipeline expression builder
Add dynamic content below using any combination of expressions, functions and system variables.
@pipeline().parameters.collection
Clear contents
Parameters
System variables
Functions
Variables
Search
collection
OK
Cancel

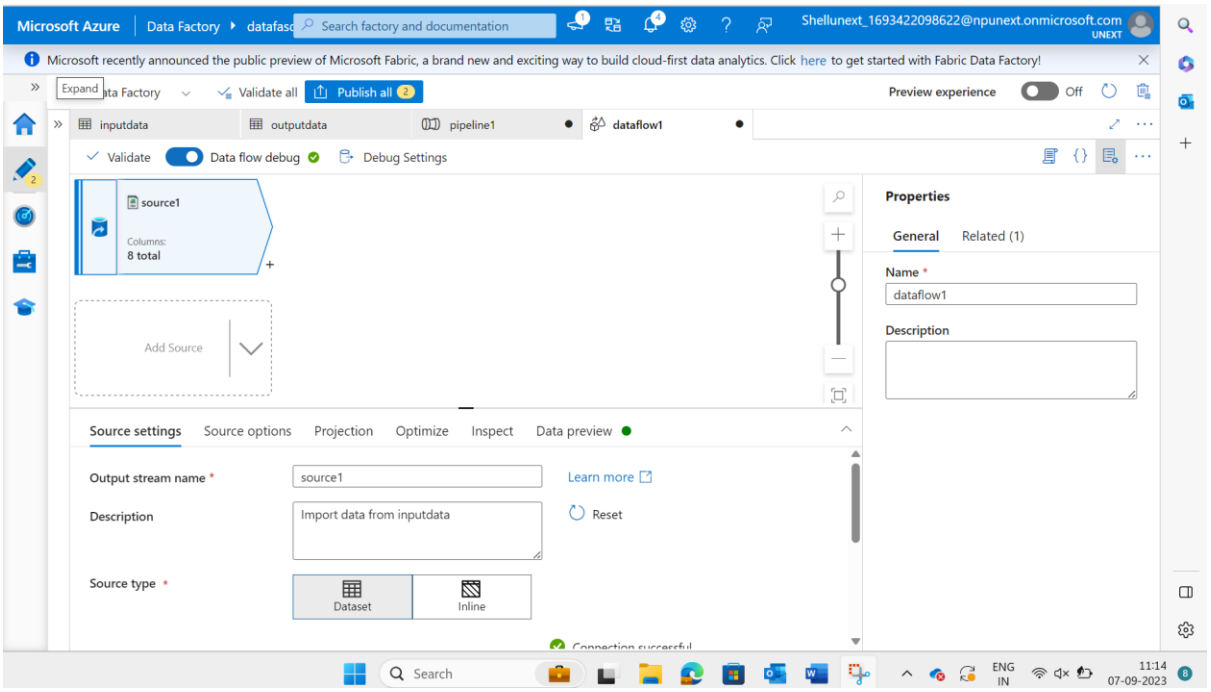


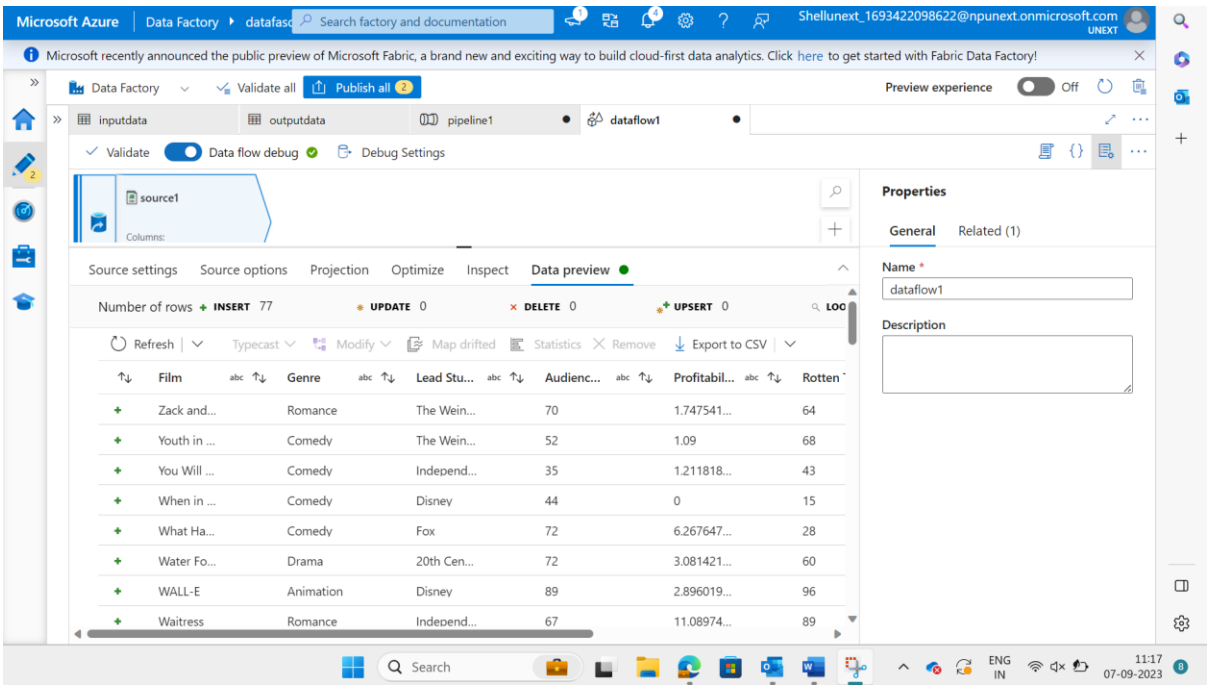

Click on next then again for file format setting click on next or click on

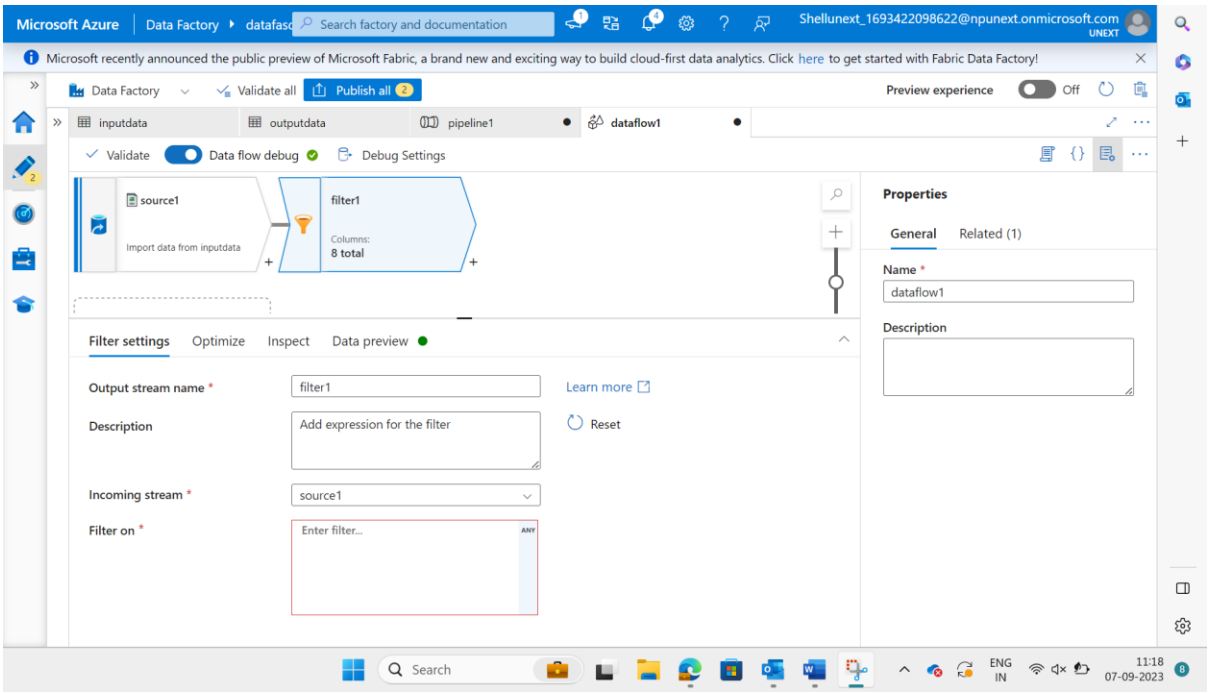


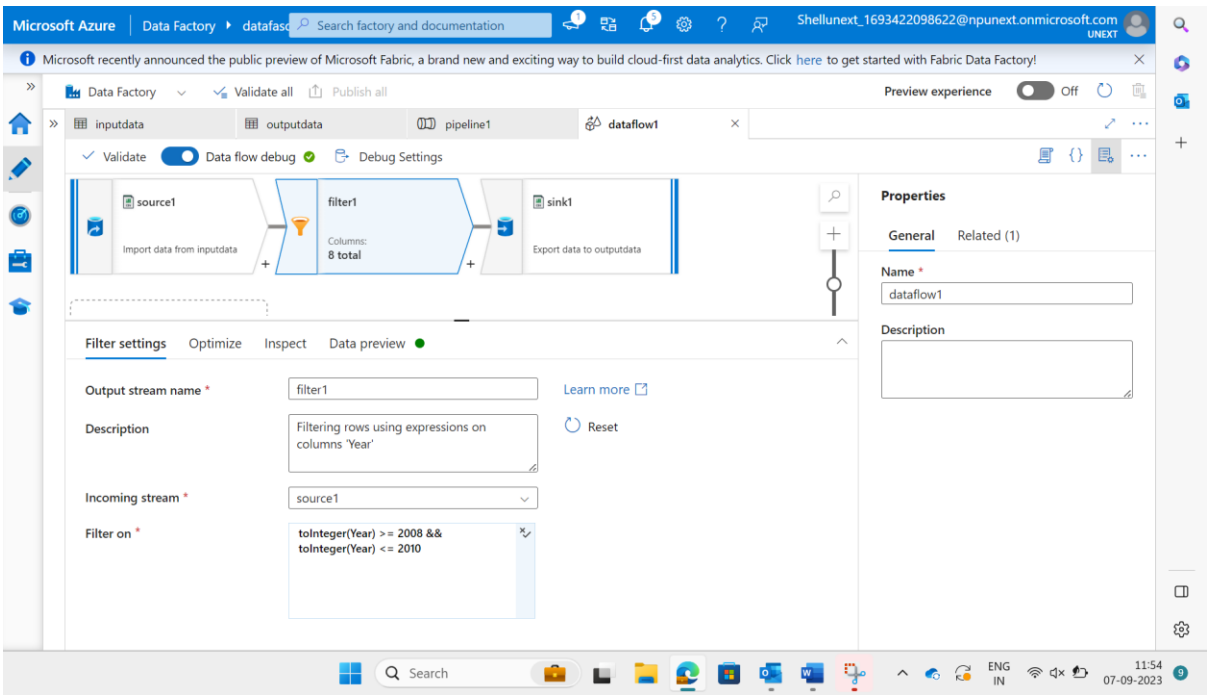



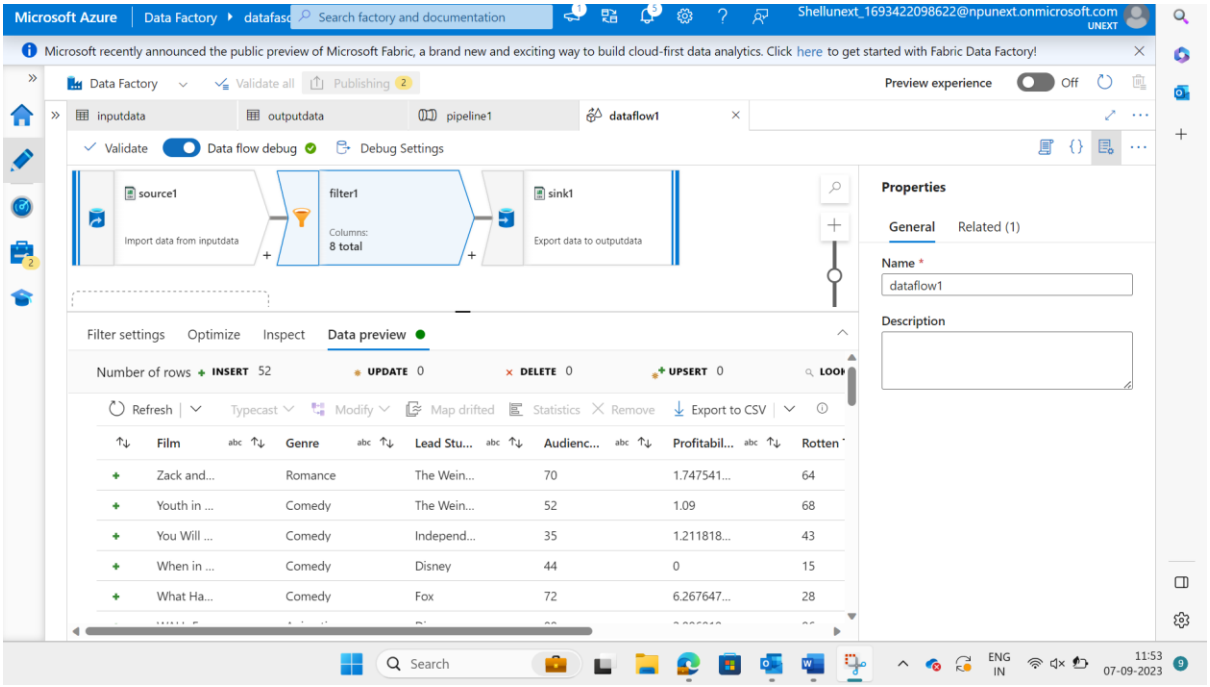

DAY- 7th September 2023

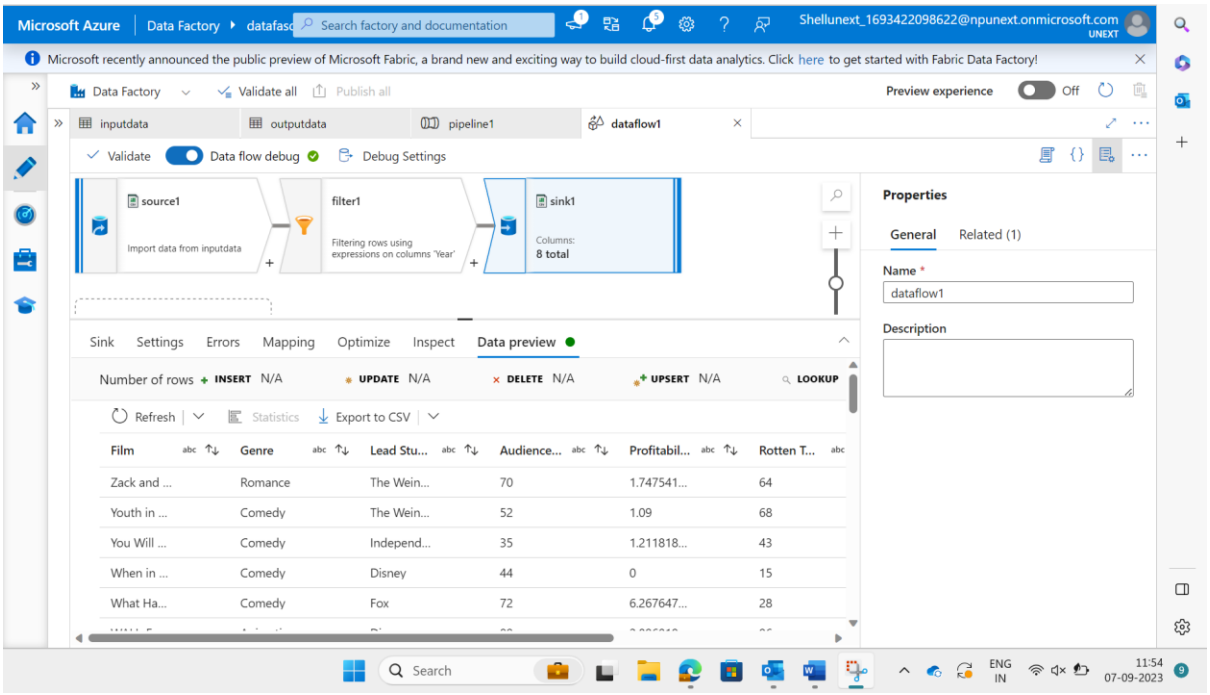

Microsoft Azure
Data Factory
Search factory and documentation
Shellunext_1693422098622@npunext.onmicrosoft.com
UNEXT
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!
Expand
Validate all
Publish all 2
Preview experience
Off
inputdata
outputdata
pipeline1
dataflow1
Validate
Data flow debug
Debug Settings
source1
Columns:
8 total
Add Source
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
Search
ENG
IN
11:14
07-09-2023

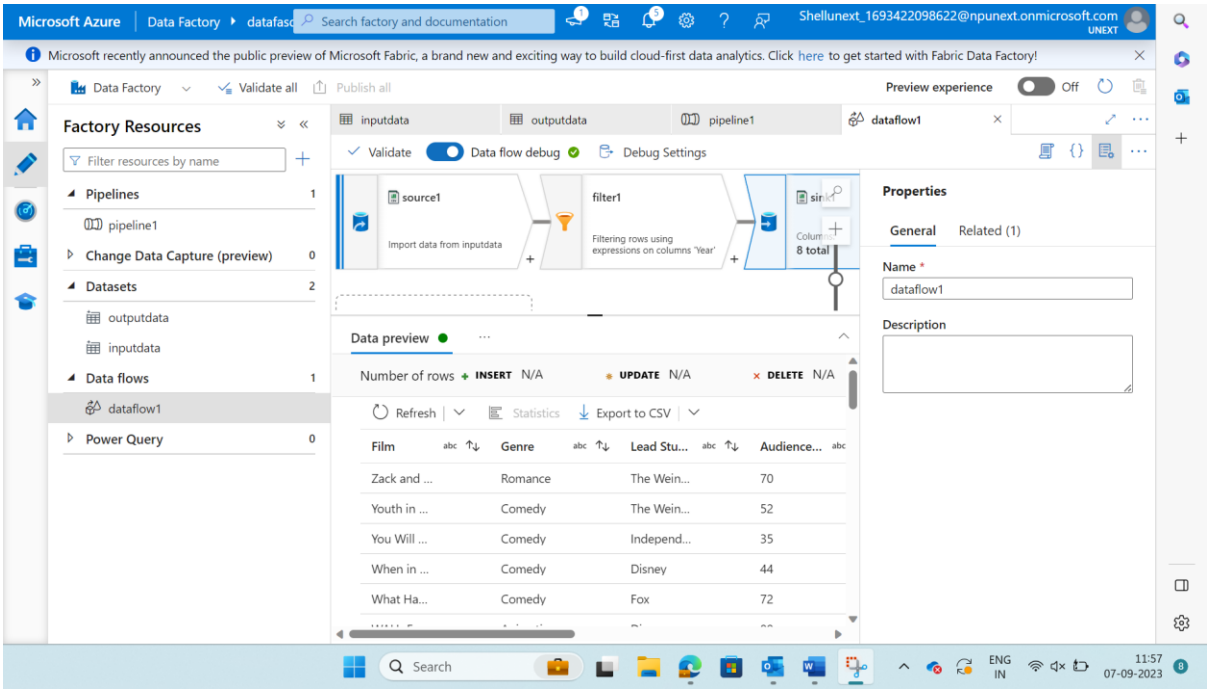

Microsoft Azure
Data Factory
Search factory and documentation
Shellunext_1693422098622@npunext.onmicrosoft.com
UNEXT
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!
Validate all
Publish all 2
Preview experience
Off
inputdata
outputdata
pipeline1
dataflow1
Validate
Data flow debug
Debug Settings
source1
Columns:
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
Search
ENG
IN
11:17
07-09-2023

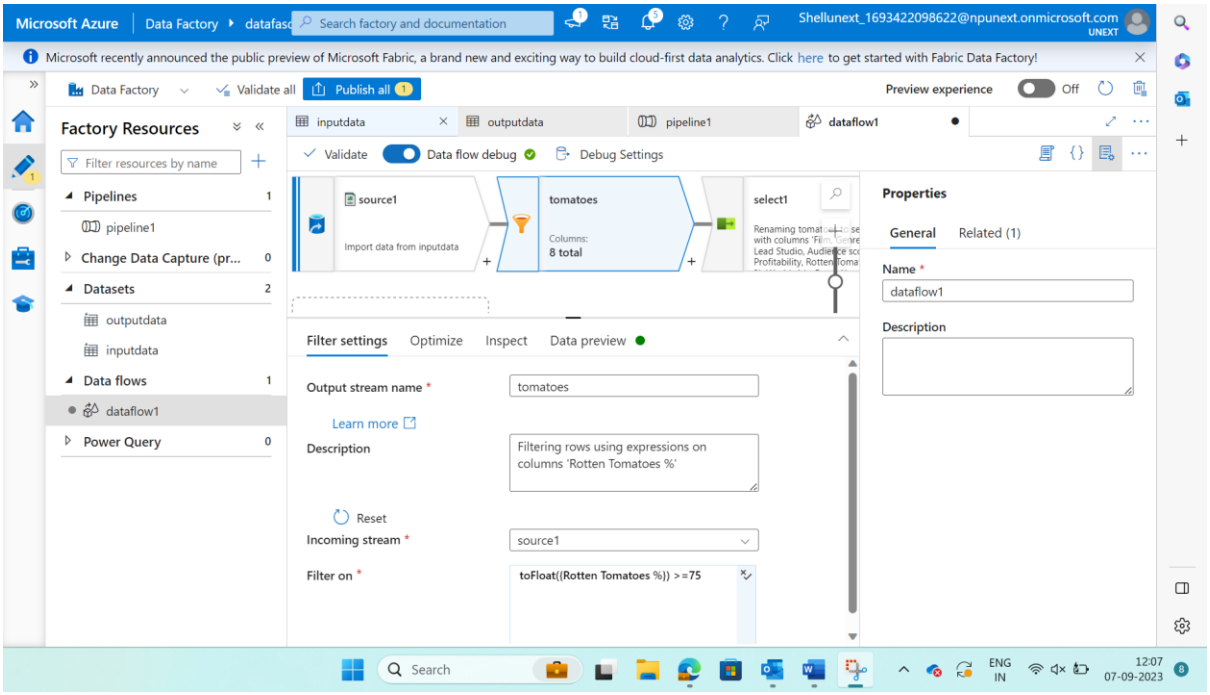

Microsoft Azure
Data Factory
Search factory and documentation
Shellunext_1693422098622@npunext.onmicrosoft.com
UNEXT
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!
Data Factory
Validate all
Publish all 2
Preview experience
Off
inputdata
outputdata
pipeline1
dataflow1
Validate
Data flow debug
Debug Settings
source1
Import data from inputdata
filter1
Columns:
8 total
Properties
General
Related (1)
Name *
dataflow1
Description
Filter settings
Optimize
Inspect
Data preview
Output stream name *
filter1
Learn more
Description
Add expression for the filter
Reset
Incoming stream *
source1
Filter on *
Enter filter...
ANY
Search
ENG
IN
11:18
07-09-2023

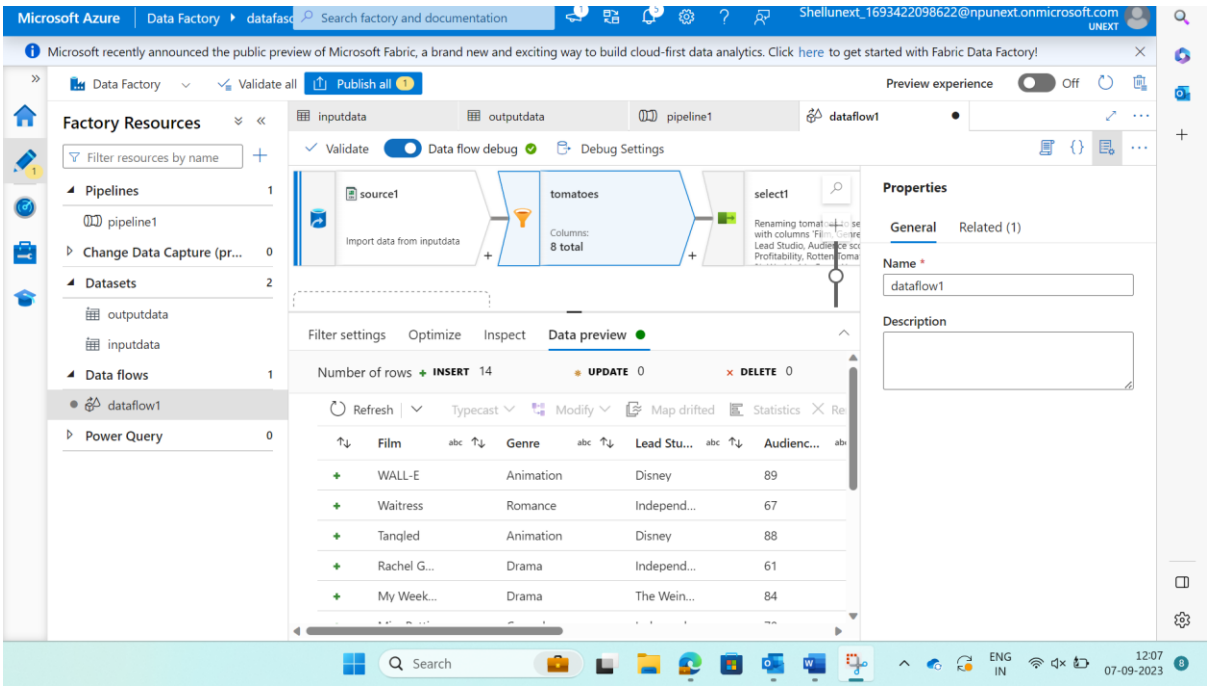

Microsoft Azure
Data Factory
Search factory and documentation
Shellunext_1693422098622@npunext.onmicrosoft.com
UNEXT
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!
Data Factory
Validate all
Publish all
Preview experience
Off
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
Output stream name *
filter1
Learn more
Description
Filtering rows using expressions on columns 'Year'
Reset
Incoming stream *
source1
Filter on *
toInteger(Year) >= 2008 &&
toInteger(Year) <= 2010
Search
ENG
IN
11:54
07-09-2023

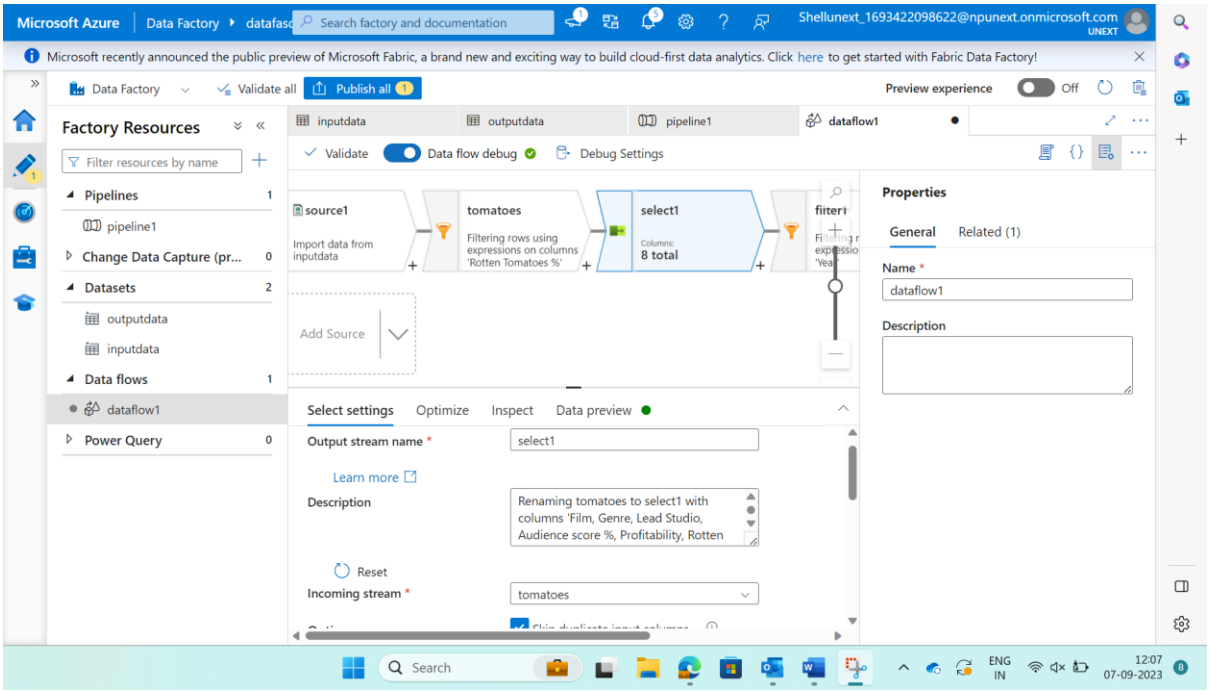

Microsoft Azure
Data Factory
Search factory and documentation
Shellunext_1693422098622@npunext.onmicrosoft.com
UNEXT
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!
Data Factory
Validate all
Publishing 2
Preview experience
Off
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
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
Search
ENG
IN
11:53
07-09-2023

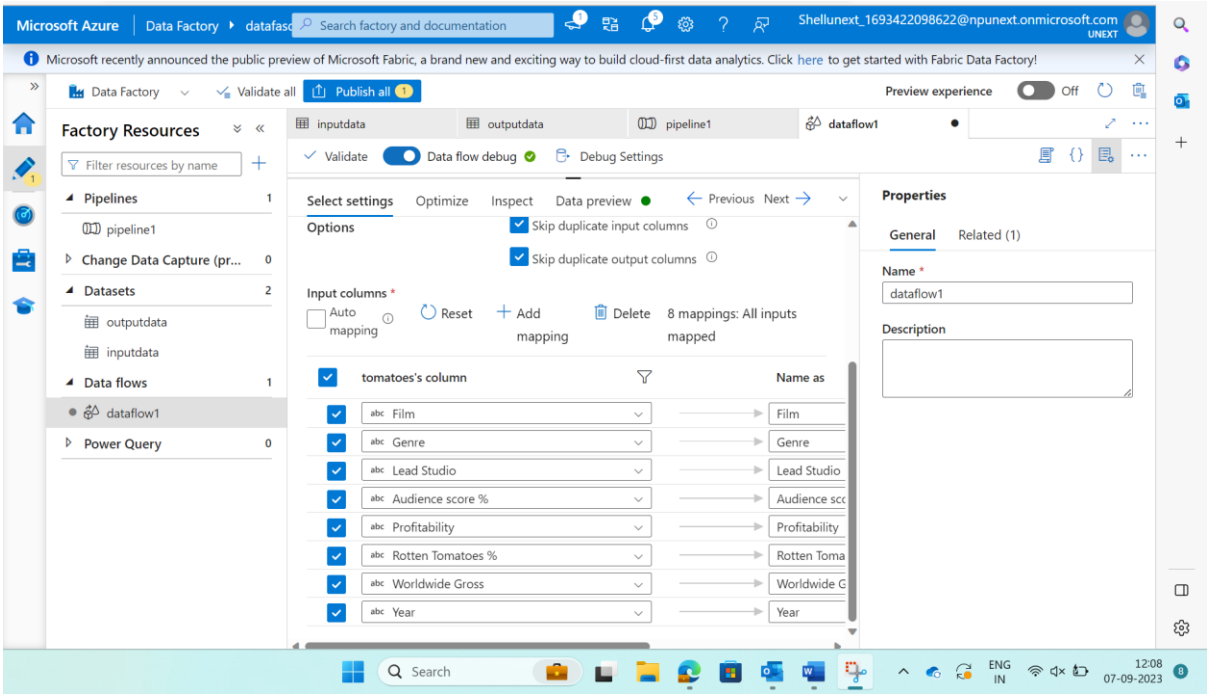

Microsoft Azure
Data Factory
Search factory and documentation
Shellunext_1693422098622@npunext.onmicrosoft.com
UNEXT
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!
Data Factory
Validate all
Publish all
Preview experience
Off
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
Search
ENG
IN
11:54
07-09-2023

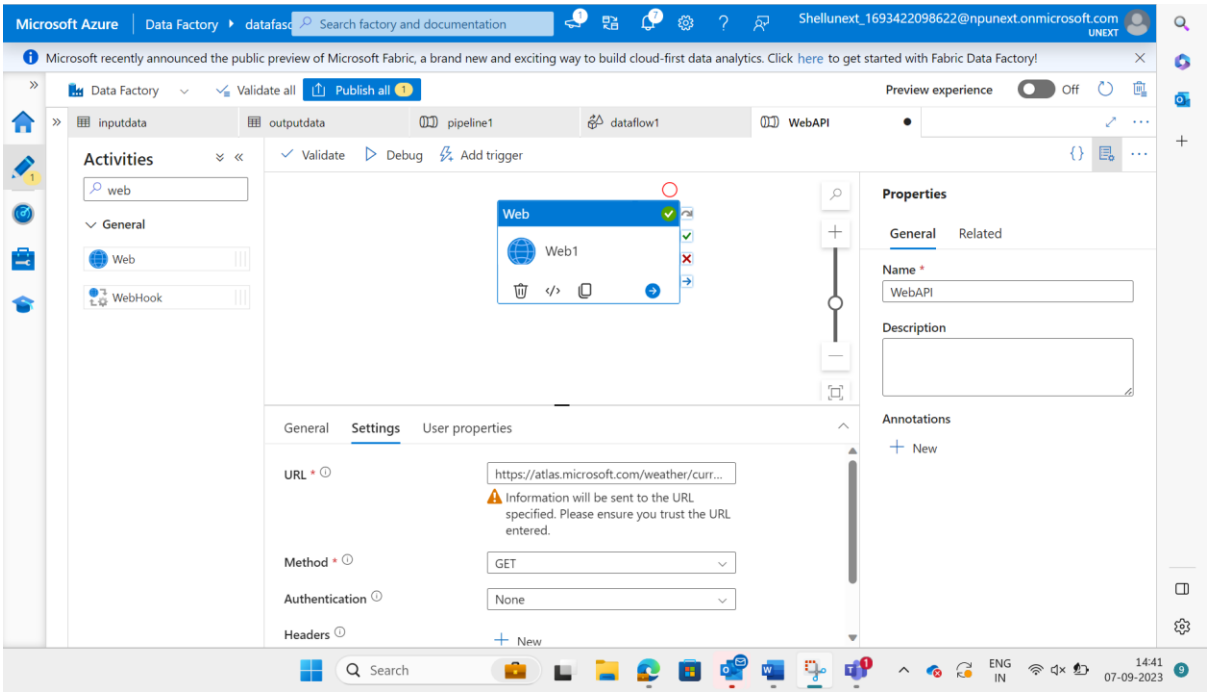

Microsoft Azure
Data Factory
Search factory and documentation
Shellunext_1693422098622@npunext.onmicrosoft.com
UNEXT
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!
Data Factory
Validate all
Publish all
Preview experience
Off
Factory Resources
Filter resources by name
Pipelines 1
pipeline1
Change Data Capture (preview) 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
filter1
Filtering rows using expressions on columns 'Year'
8 total
Properties
General
Related (1)
Name *
dataflow1
Description
Data preview
Number of rows + INSERT N/A
UPDATE N/A
DELETE N/A
Refresh
Statistics
Export to CSV
Film
Genre
Lead Stu...
Audience...
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
Search
ENG
IN
11:57
07-09-2023

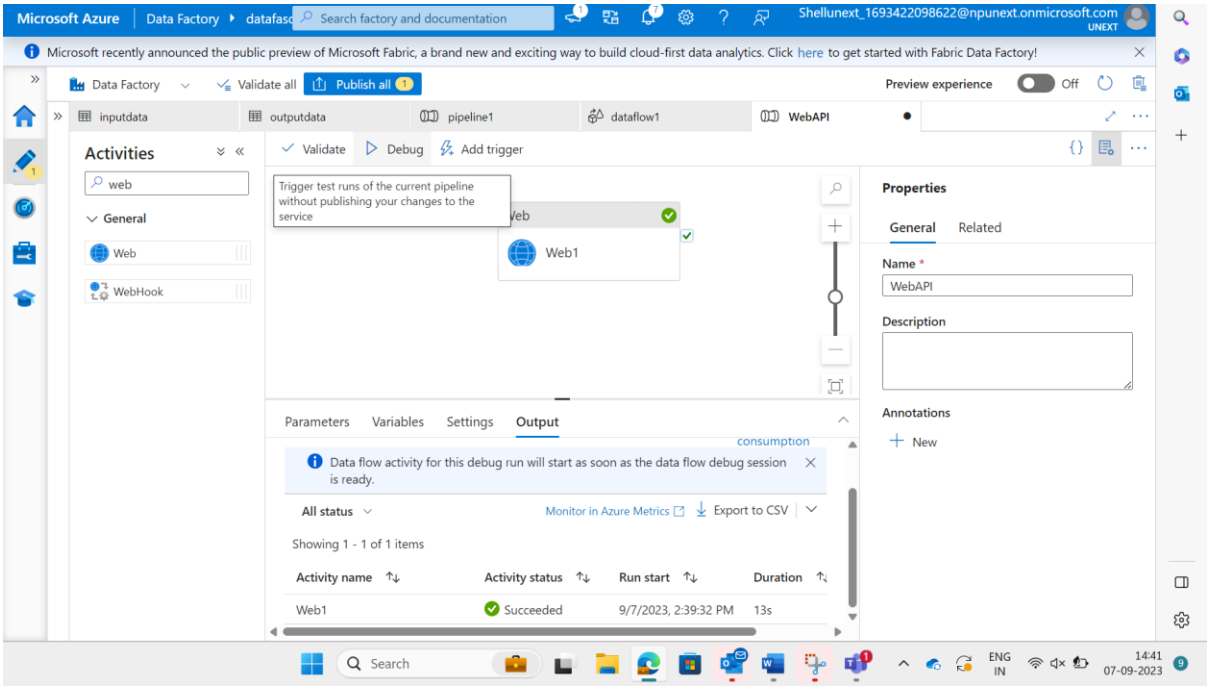

Microsoft Azure
Data Factory
Search factory and documentation
Shellunext_1693422098622@npunext.onmicrosoft.com
UNEXT
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!
Data Factory
Validate all
Publish all 1
Preview experience
Off
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
Columns:
8 total
select1
Properties
General
Related (1)
Name *
dataflow1
Description
Filter settings
Optimize
Inspect
Data preview
Output stream name *
tomatoes
Learn more
Description
Filtering rows using expressions on columns 'Rotten Tomatoes %'
Reset
Incoming stream *
source1
Filter on *
toFloat((Rotten Tomatoes %)) >= 75
Search
ENG
IN
12:07
07-09-2023

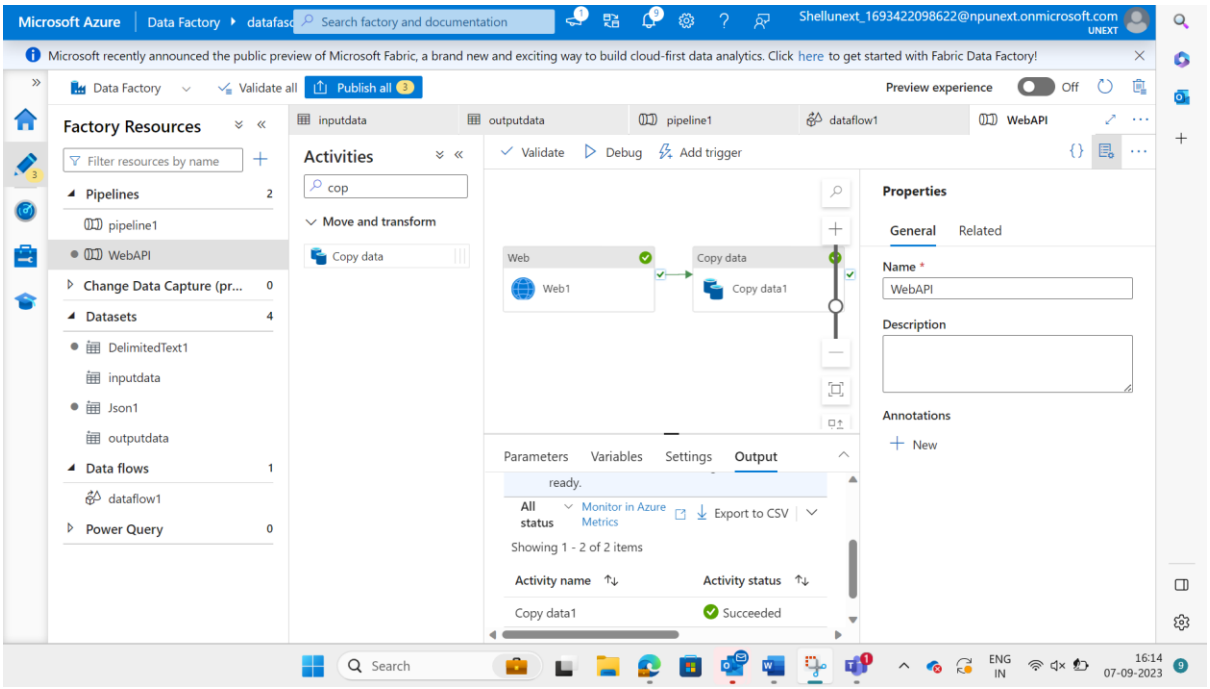

Microsoft Azure
Data Factory
Search factory and documentation
Shellunext_1693422098622@npunext.onmicrosoft.com
UNEXT
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!
Data Factory
Validate all
Publish all 1
Preview experience
Off
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
Columns:
8 total
select1
Properties
General
Related (1)
Name *
dataflow1
Description
Filter settings
Optimize
Inspect
Data preview
Number of rows + INSERT 14
UPDATE 0
DELETE 0
Refresh
Typecast
Modify
Map drifted
Statistics
Film
Genre
Lead Stu...
Audienc...
WALL-E
Animation
Disney
89
Waitress
Romance
Independ...
67
Tangled
Animation
Disney
88
Rachel G...
Drama
Independ...
61
My Week...
Drama
The Wein...
84
Search
ENG
IN
12:07
07-09-2023


Microsoft Azure
Data Factory
Search factory and documentation
Shellunext_1693422098622@npunext.onmicrosoft.com
UNEXT
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 1
Preview experience
Off
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
Filtering rows using expressions on columns 'Rotten Tomatoes %'
select1
Columns:
8 total
Add Source
Properties
General
Related (1)
Name *
dataflow1
Description
Select settings
Optimize
Inspect
Data preview
Output stream name *
select1
Learn more
Description
Renaming tomatoes to select1 with columns 'Film, Genre, Lead Studio, Audience score %, Profitability, Rotten
Reset
Incoming stream *
tomatoes
Search
ENG
IN
12:07
07-09-2023

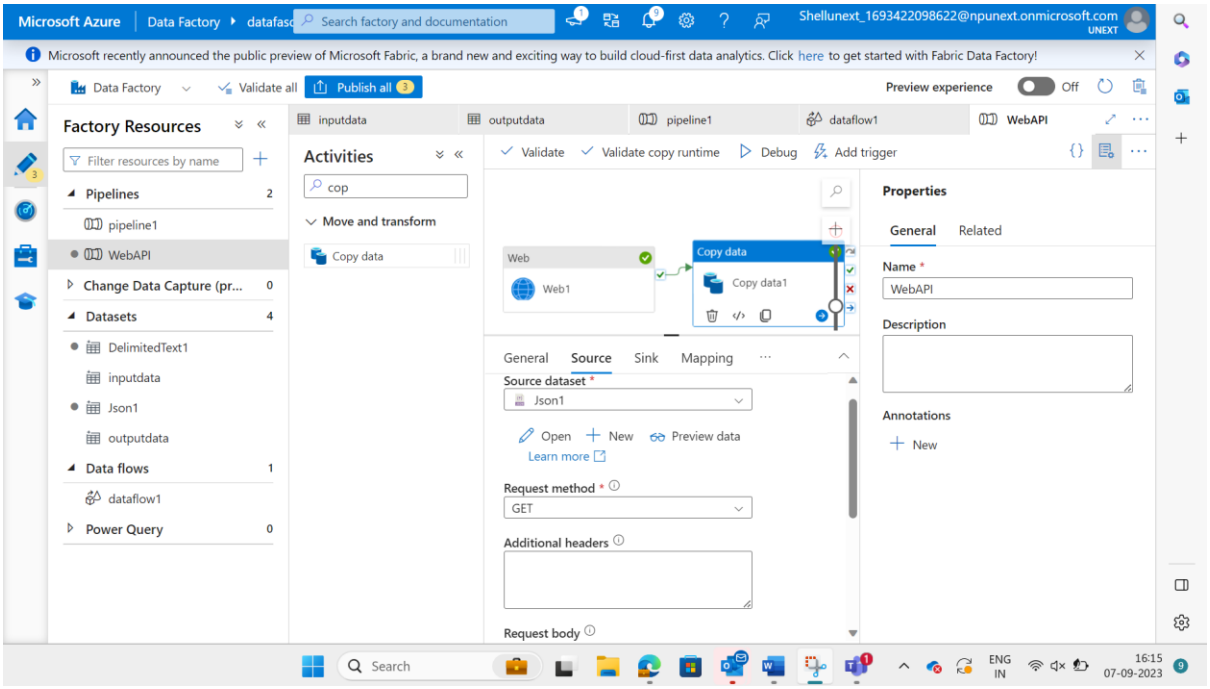

Microsoft Azure
Data Factory
Search factory and documentation
Shellunext_1693422098622@npunext.onmicrosoft.com
UNEXT
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!
Data Factory
Validate all
Publish all 1
Preview experience
Off
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
Select settings
Optimize
Inspect
Data preview
Previous
Next
Options
Skip duplicate input columns
Skip duplicate output columns
Input columns *
Auto mapping
Reset
Add mapping
Delete
8 mappings: All inputs mapped
tomatoes's column
Name as
Film
Genre
Lead Studio
Audience score %
Profitability
Rotten Tomatoes %
Worldwide Gross
Year
Properties
General
Related (1)
Name *
dataflow1
Description
Search
ENG
IN
12:08
07-09-2023


Microsoft Azure
Data Factory
Search factory and documentation
Shellunext_1693422098622@npunext.onmicrosoft.com
UNEXT
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 1
Preview experience
Off
inputdata
outputdata
pipeline1
dataflow1
WebAPI
Activities
web
General
Web
WebHook
Validate
Debug
Add trigger
Web
Web1
General
Settings
User properties
URL *
https://atlas.microsoft.com/weather/curr...
Information will be sent to the URL specified. Please ensure you trust the URL entered.
Method *
GET
Authentication
None
Headers
New
Properties
General
Related
Name *
WebAPI
Description
Annotations
New
Search
ENG
IN
14:41
07-09-2023

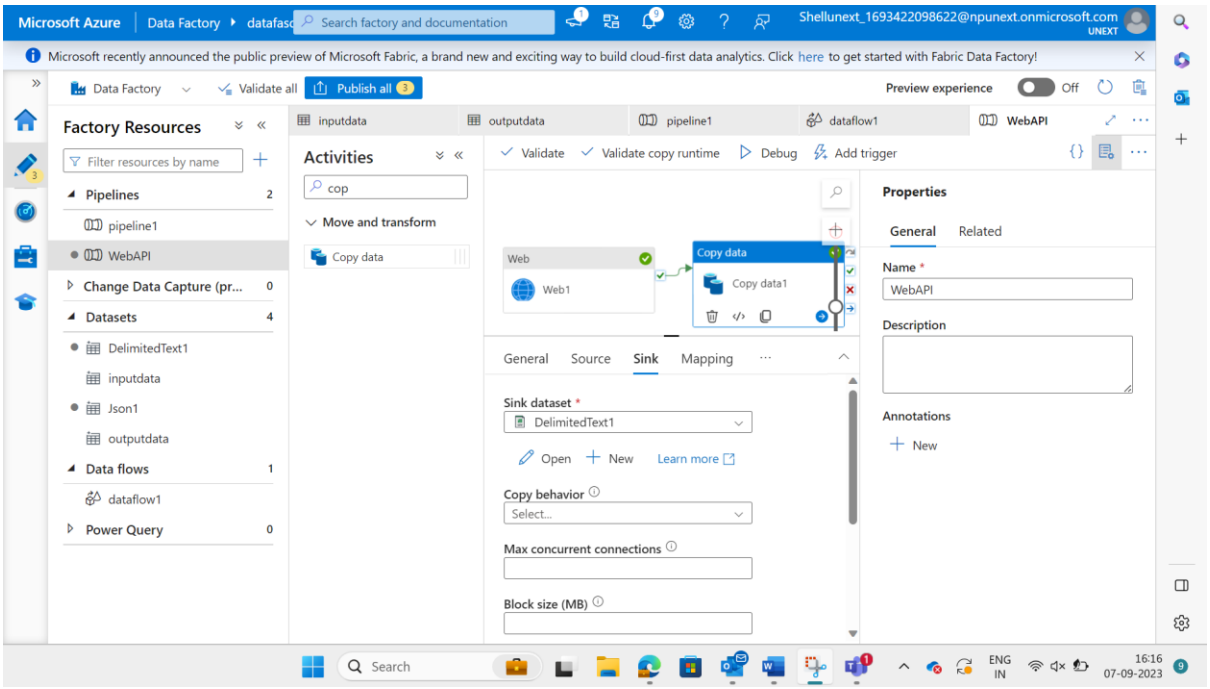


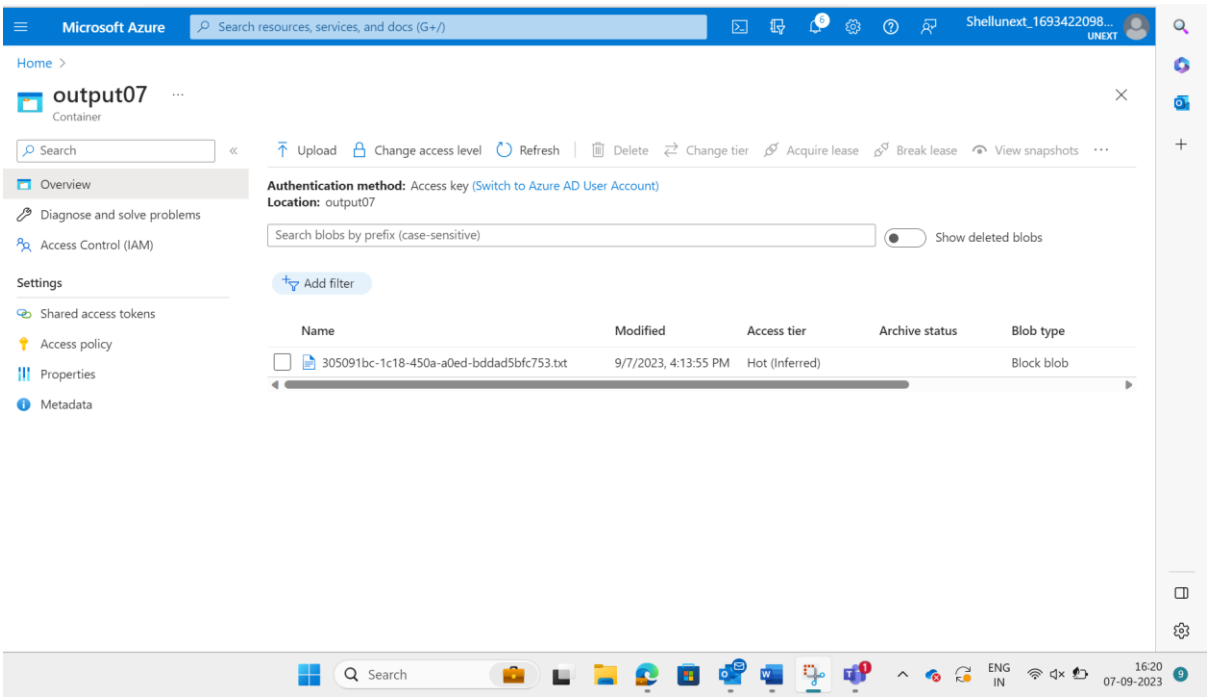

Okay for this first we need to drag copy data then in source choose HTTP.




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.




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)

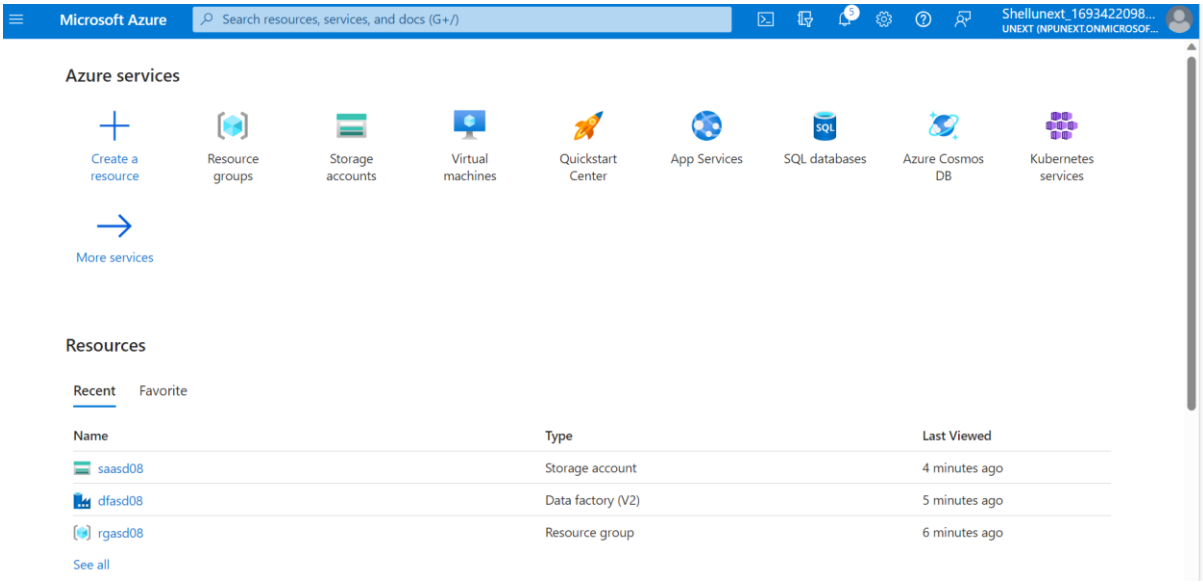
GUI->

Central authentication->AAD

User, password authentication->see in pic

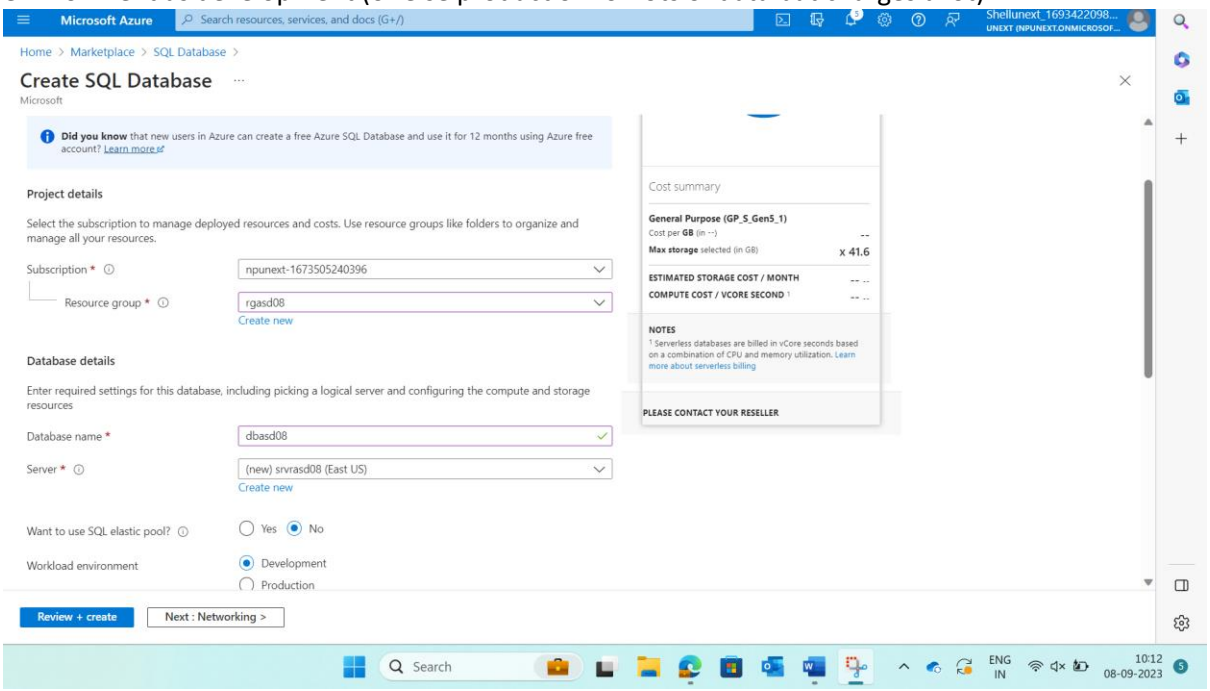
<u>Steps for today:</u>

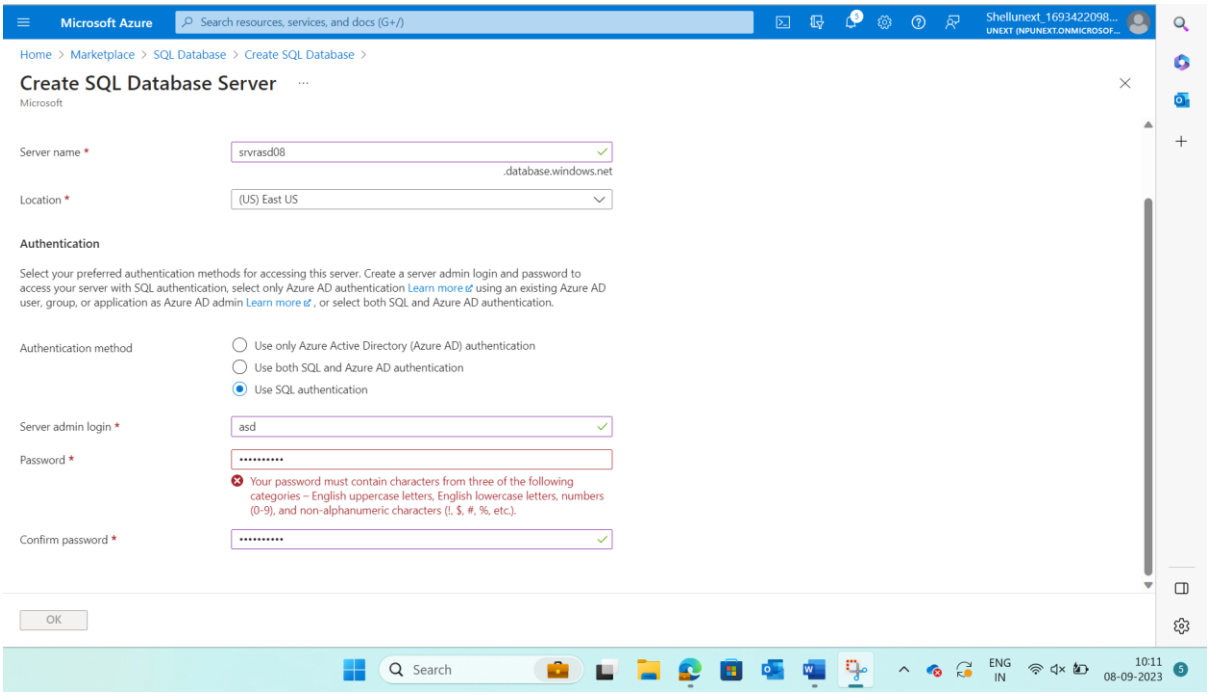
1. Create RA->SA(Gen2)->container(i/p, o/p)->ADF



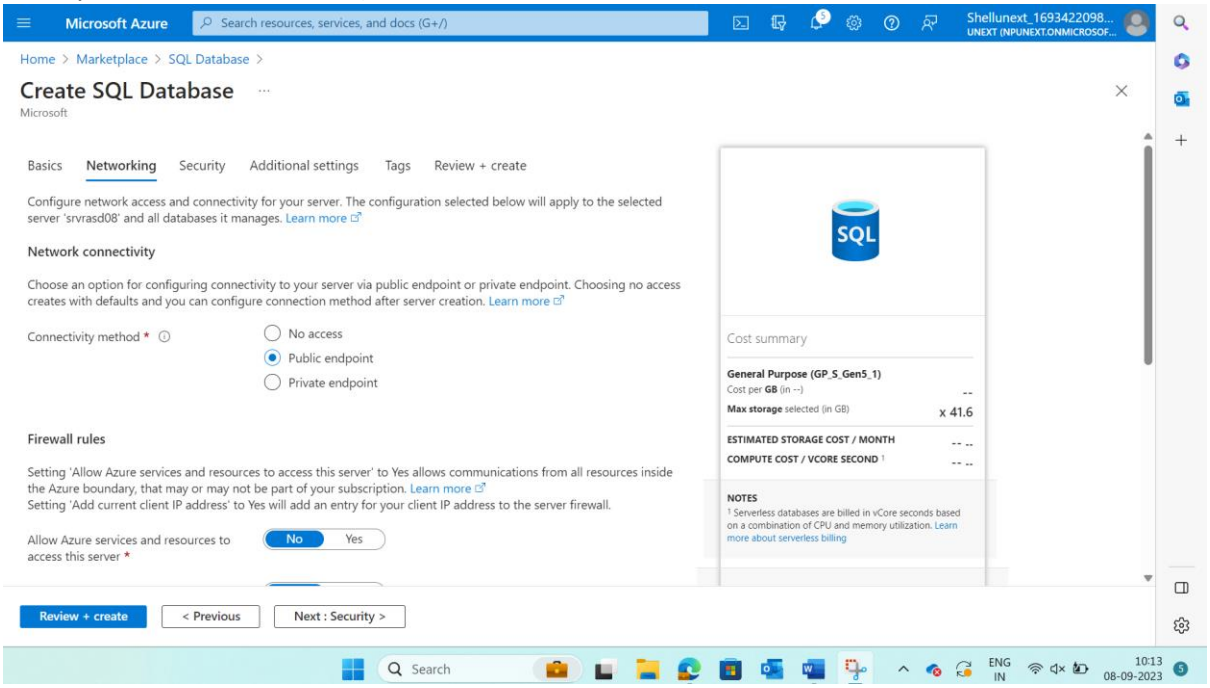

2. Create server host-> srvrasd08.database.windows.net

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)

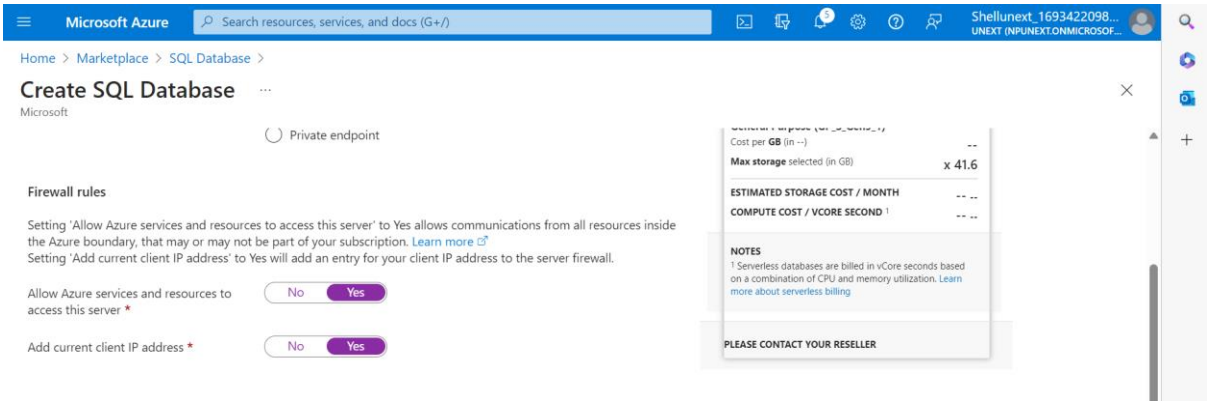





-> in networking select connectivity method-> select Public endpoint(don't use this in company as they have confidential data)

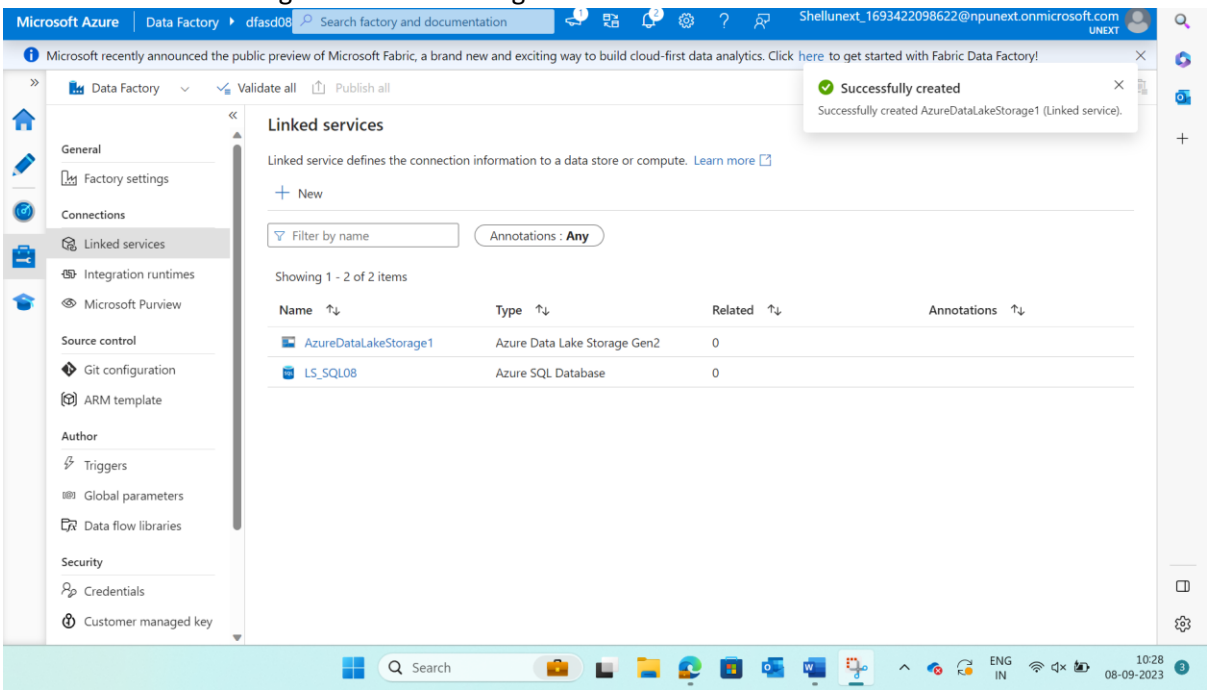



-> next in firewall rules there are 2 rules-> allow azure services and resources to access this server and add current client IP address.




- > 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 suser name and password.-> test connection-> create.
5. Create new LS-> select gen 2-> select storage account-> create it.



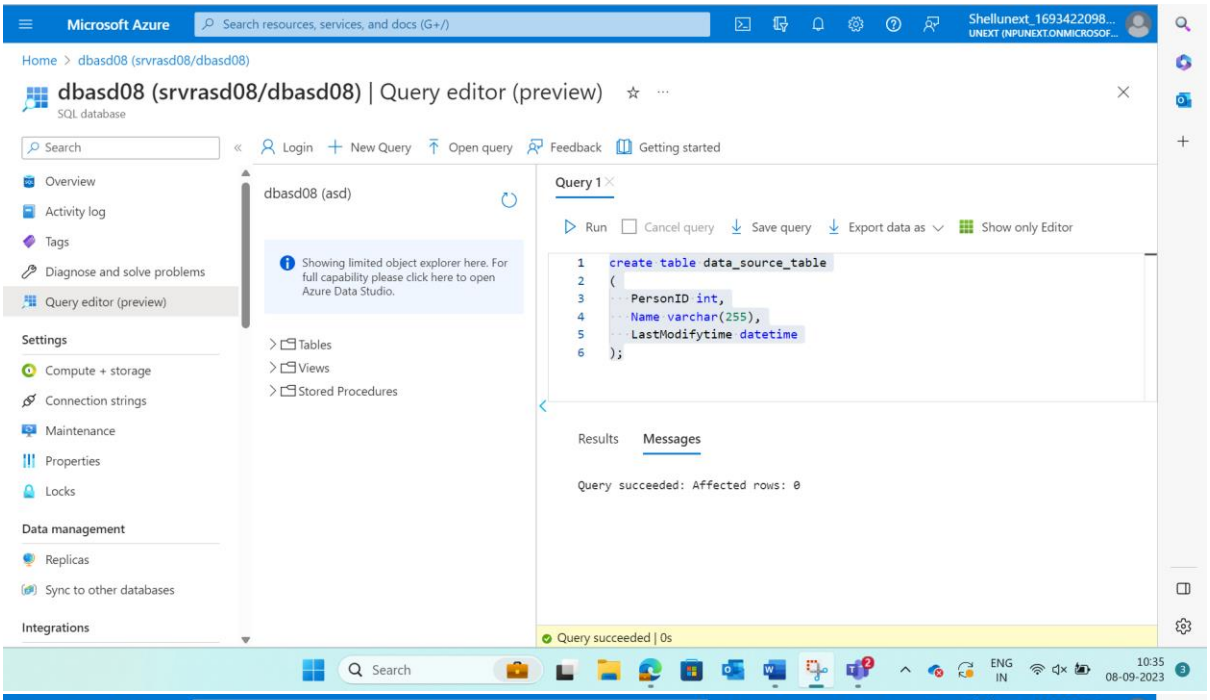

6. Now under datasets->Azure SQL database->name DS_SQL_input->select LS ->now for table name-> go back to azure portal and click on sql database-> under query editor we need to give user id and password->type sql query and insert and run

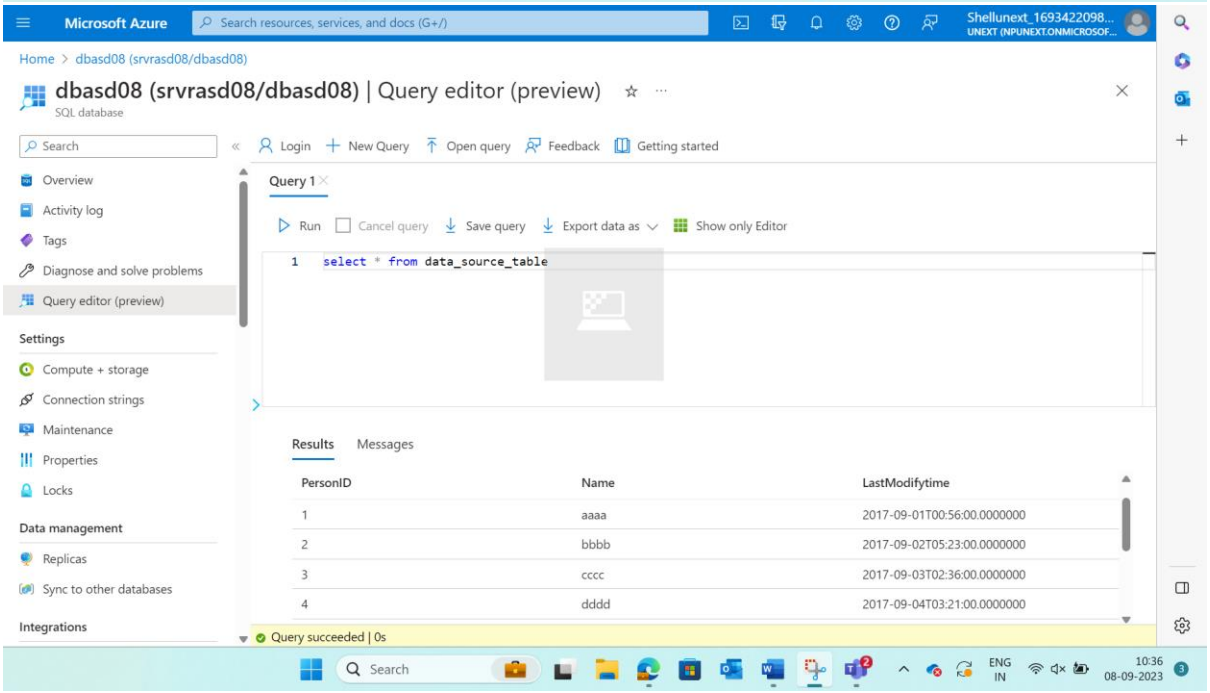
The sql query is->[10:29] Singh, Aastha SBOBNG-PTIY/RE

```
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
```

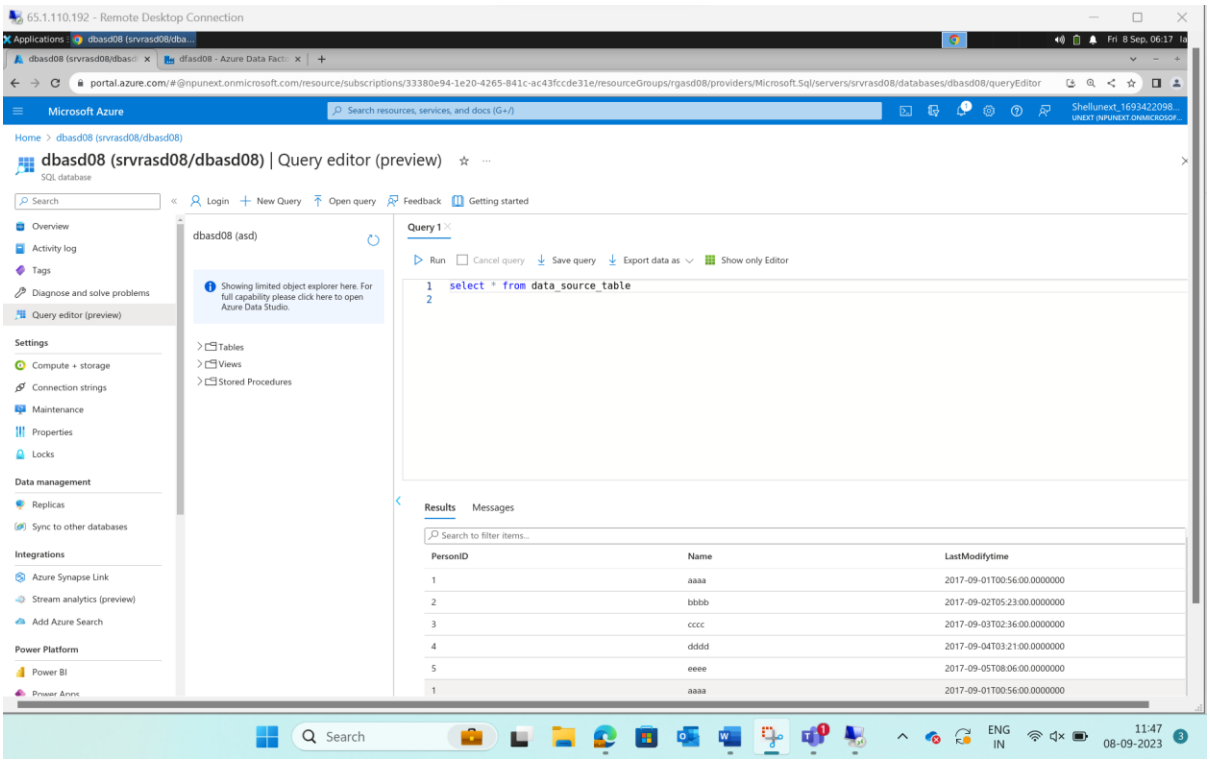
END



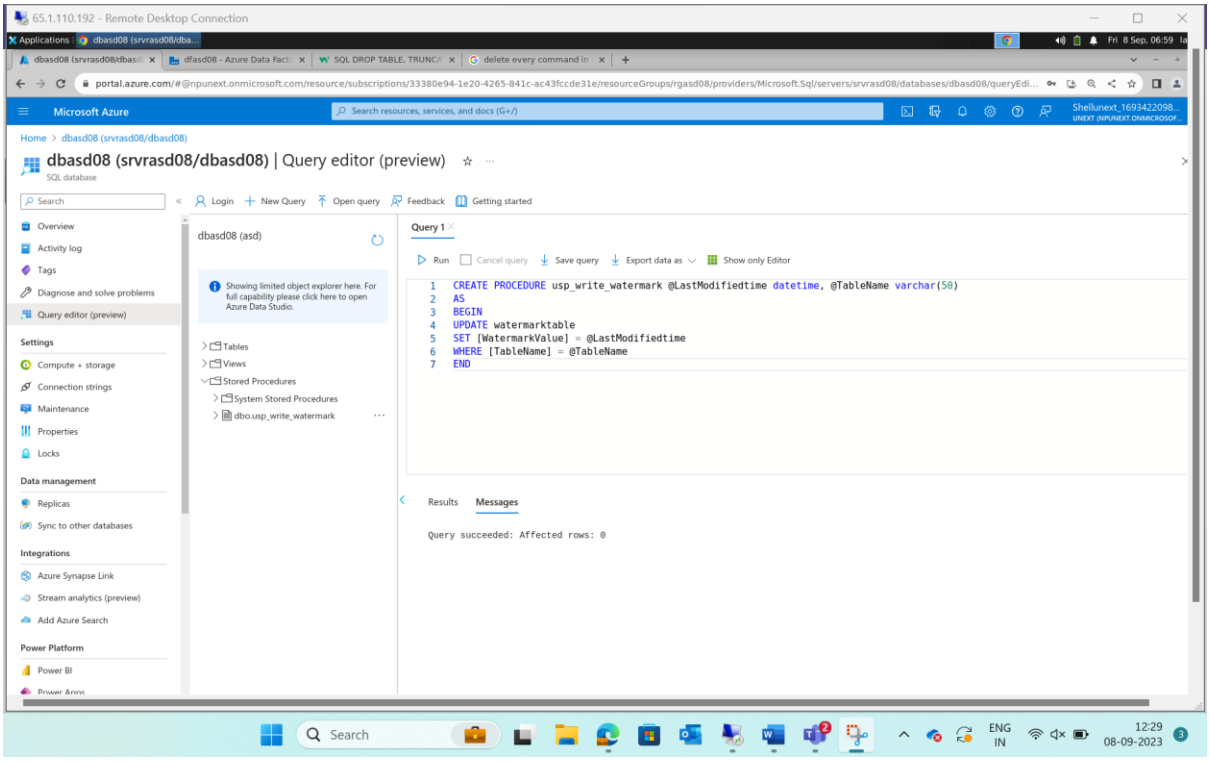




- >Then refresh table name in ADF and select and create.

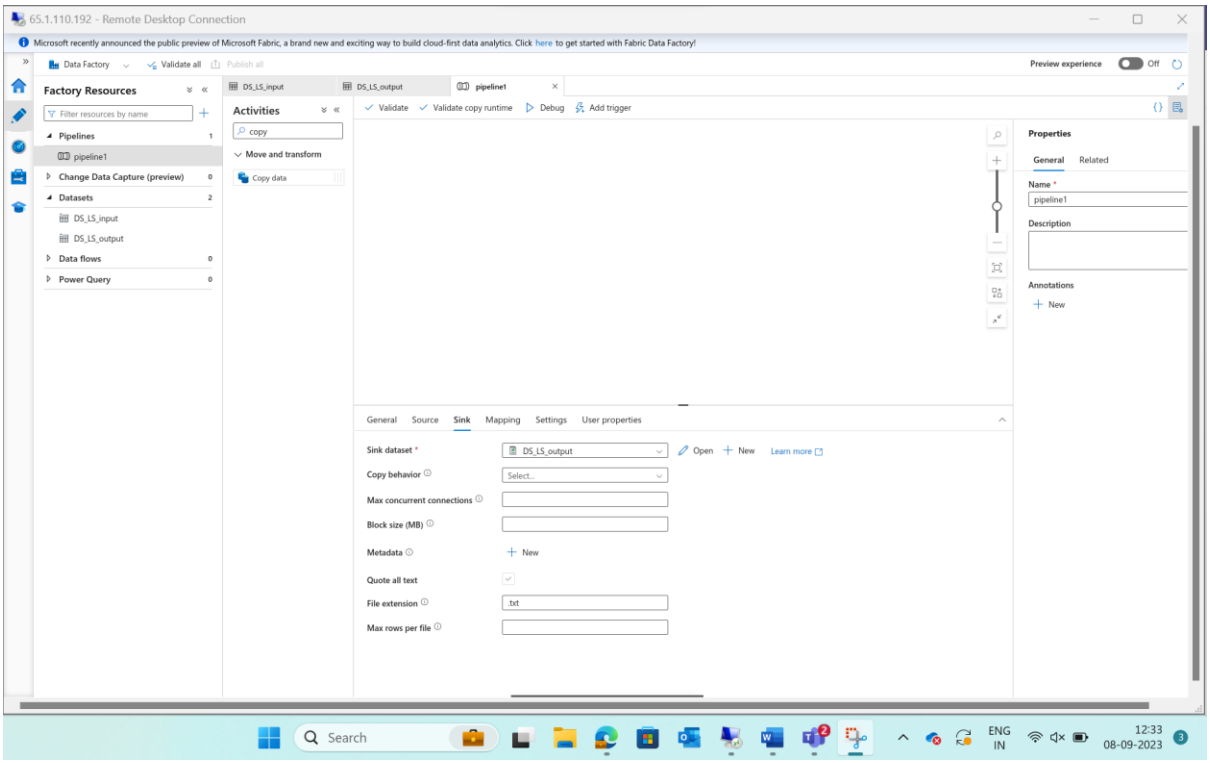
7. Then create another dataset for output-> gen2->from container select output folder and csv file option before this.
   We needed to connect this through the VM



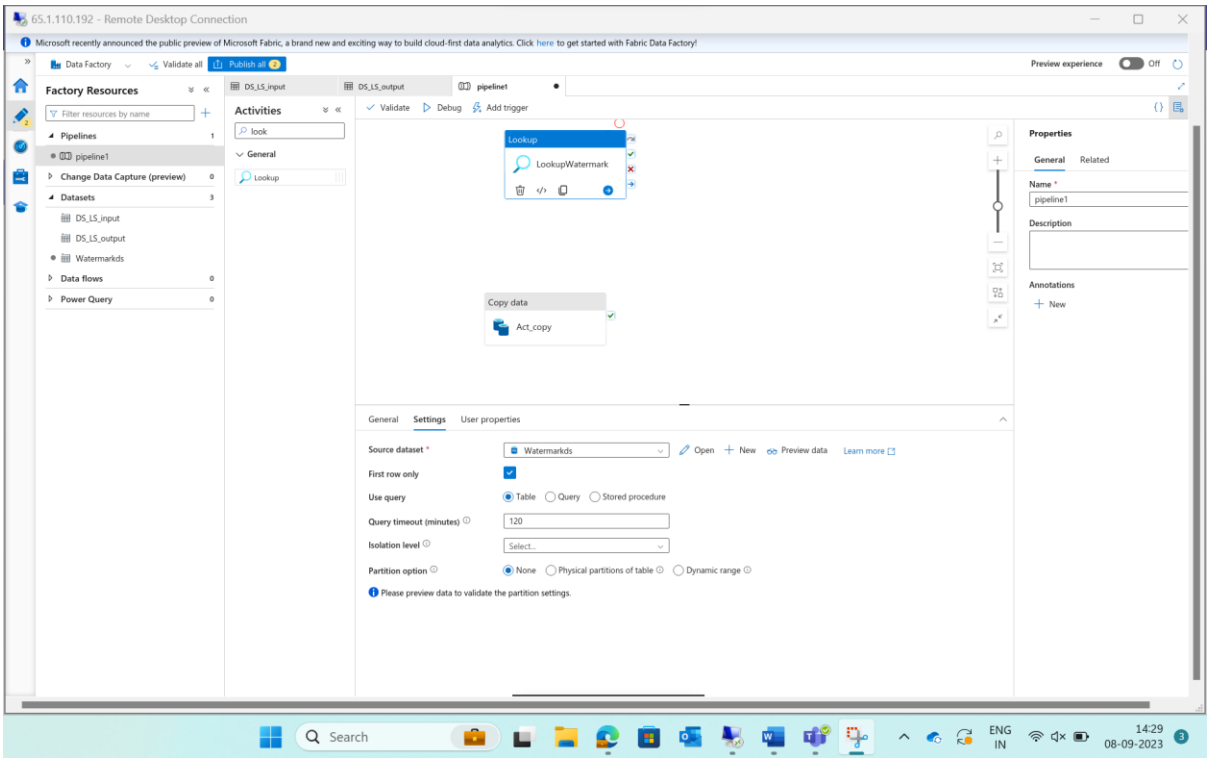

Run the whole table and then create procedure table



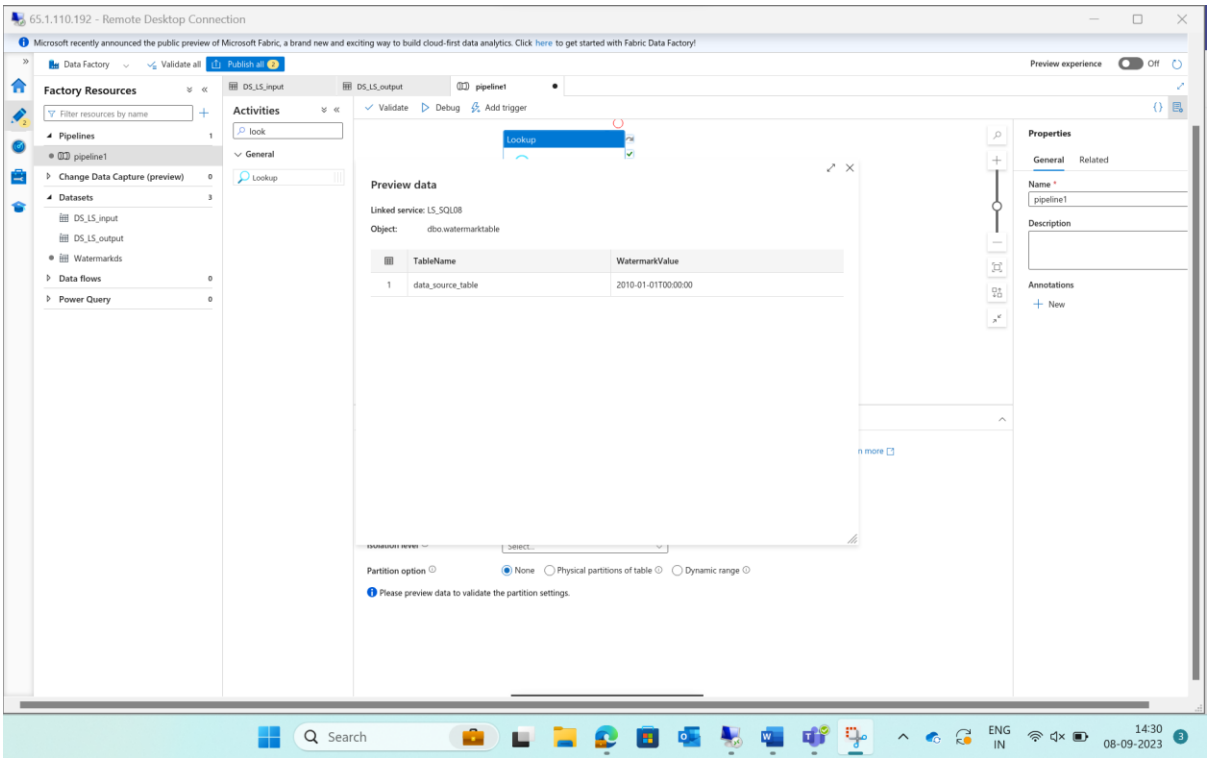

8. Now in pipline add source and sink and then validate and publish



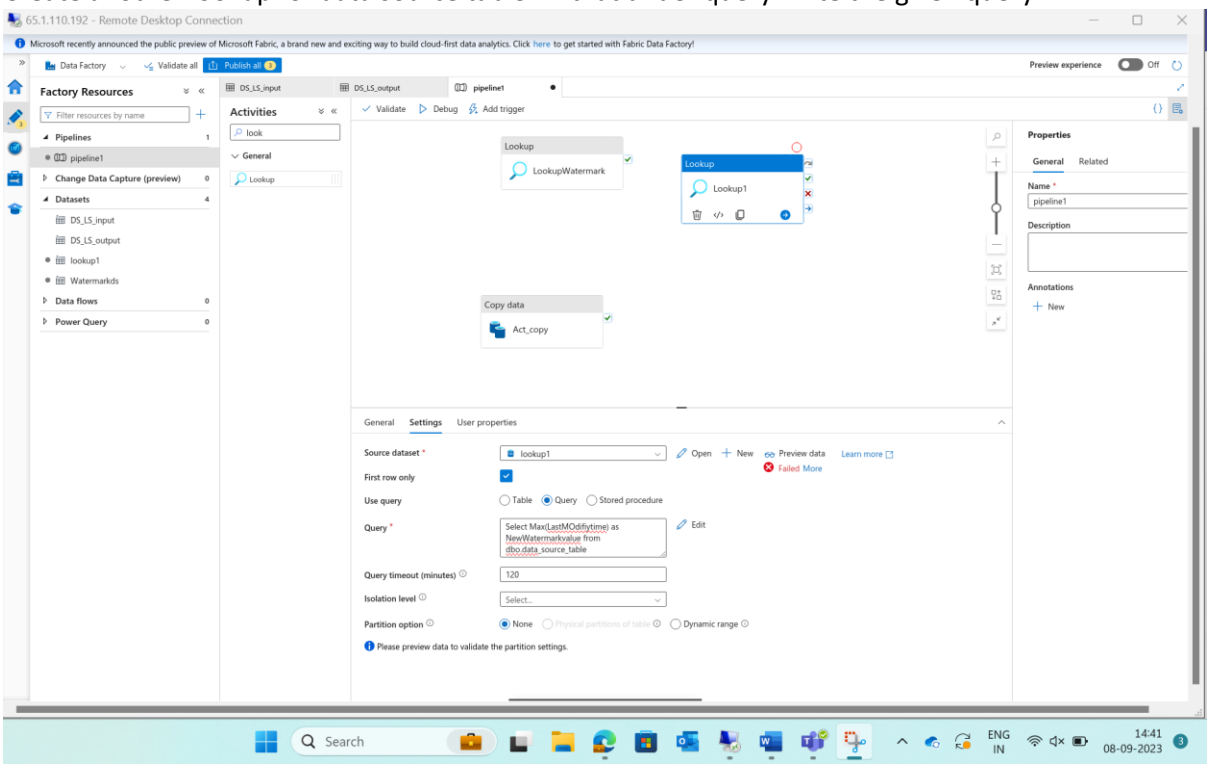

9. Next in pipline search and drag lookup and in settings create a new dataset with azure sql database then under it name it watermarkds and then under it select dbo.watermark one.

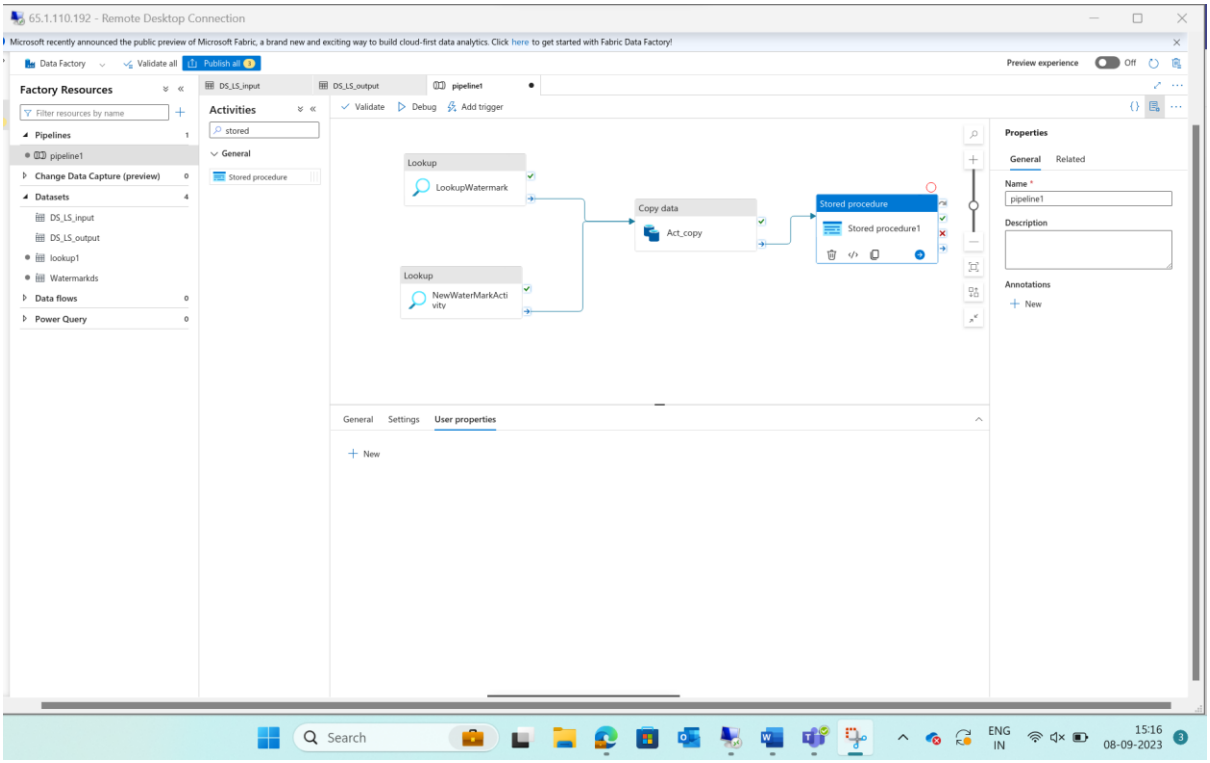



We can also preview data and see that data set has been copied.




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

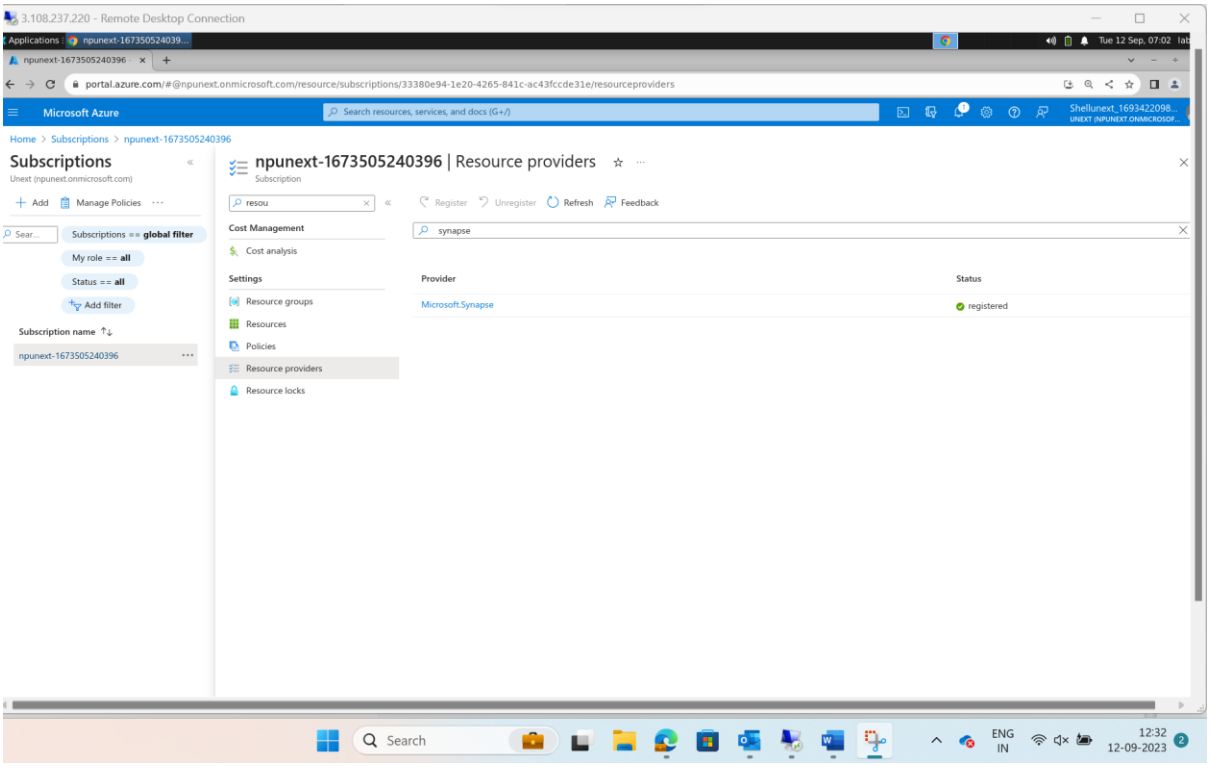
Data-> live and dead

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.

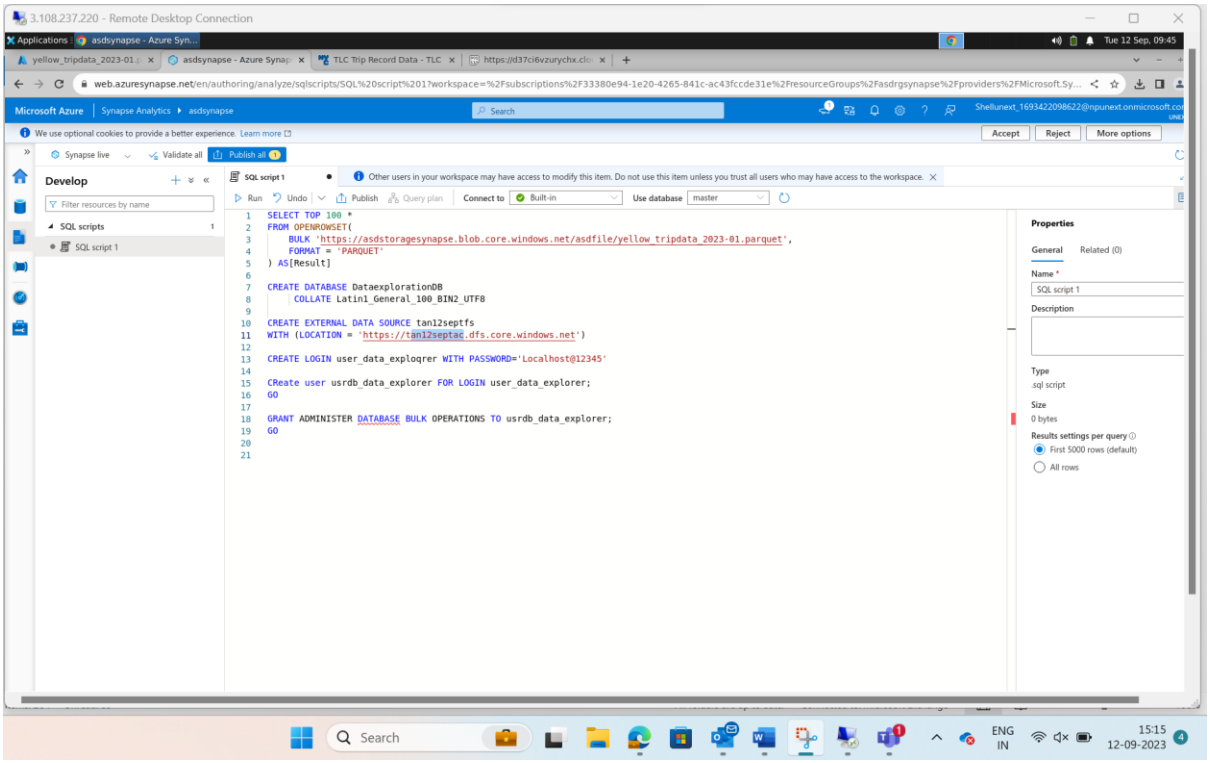
Synapse-dwh, big data and data explorer

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




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

Qwerty1234




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, authoriasation, transformation), etc. and it can have one or multiple storage.

OLTP->DB.SQL(Row)

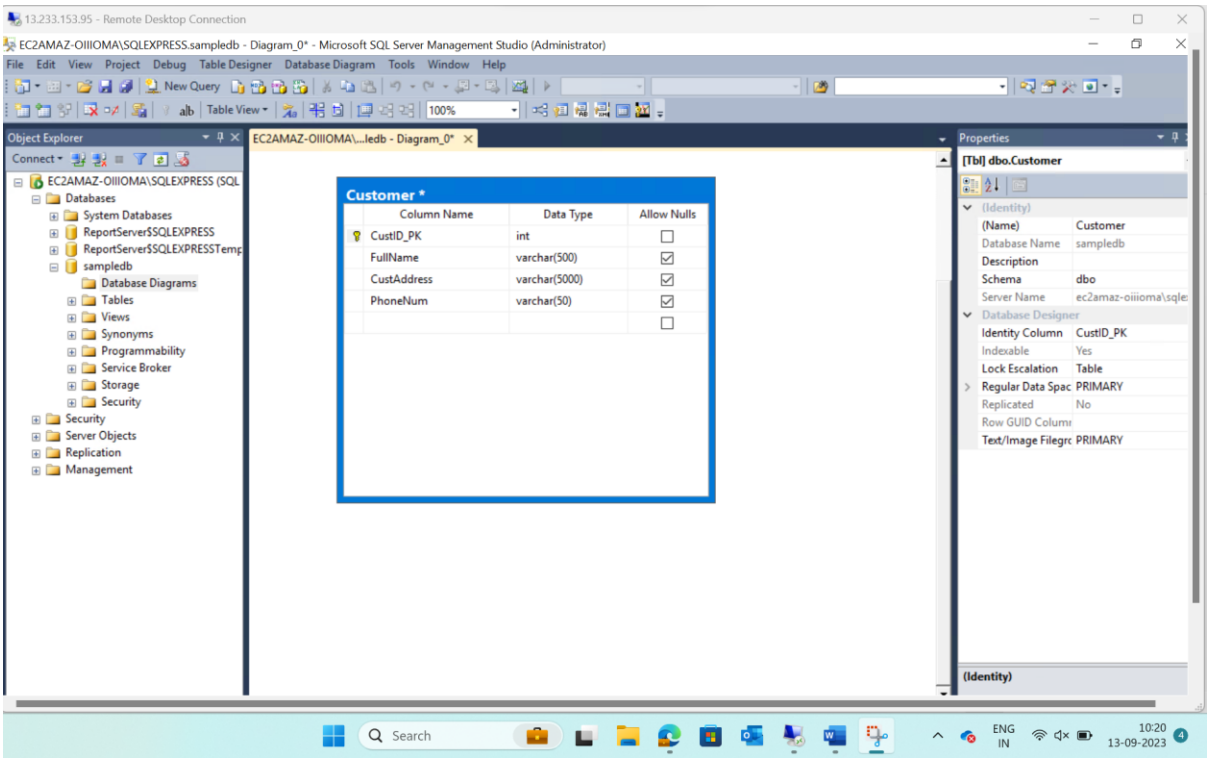
OLAP->DB.NoSql(Column)

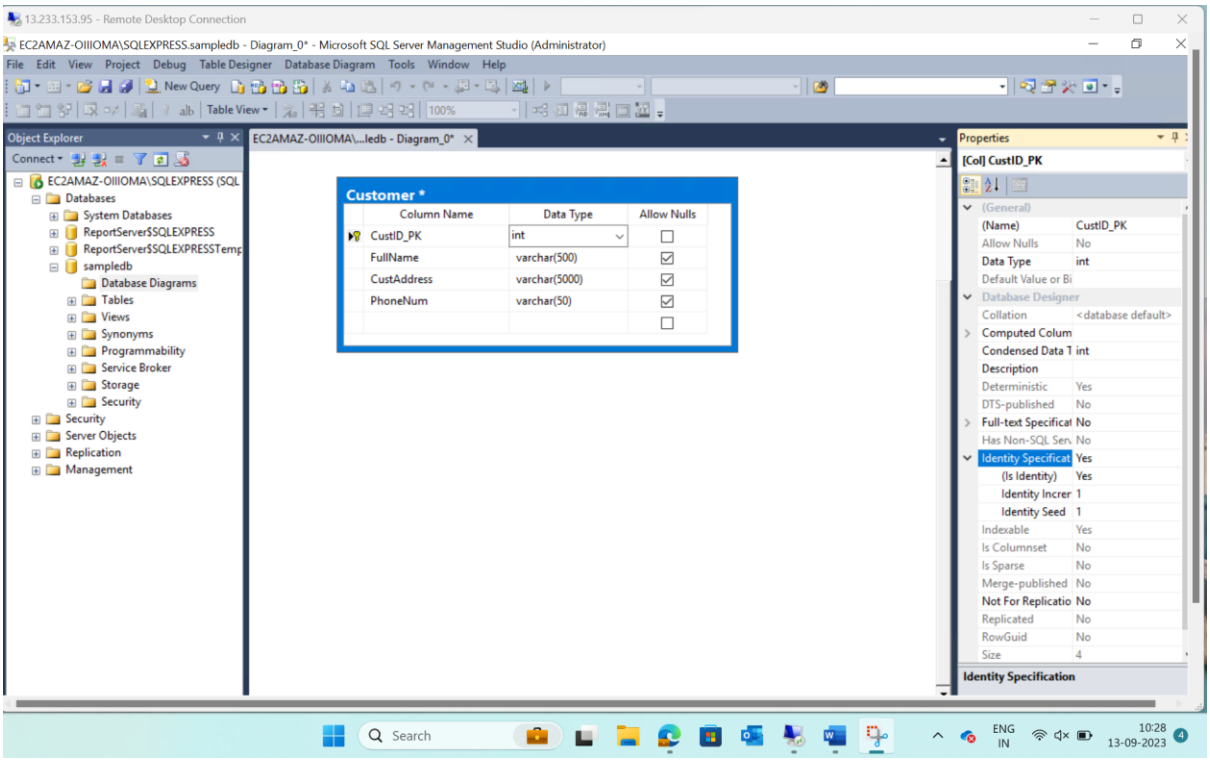
PowerBi-> Microsoft provide these 3 services

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






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

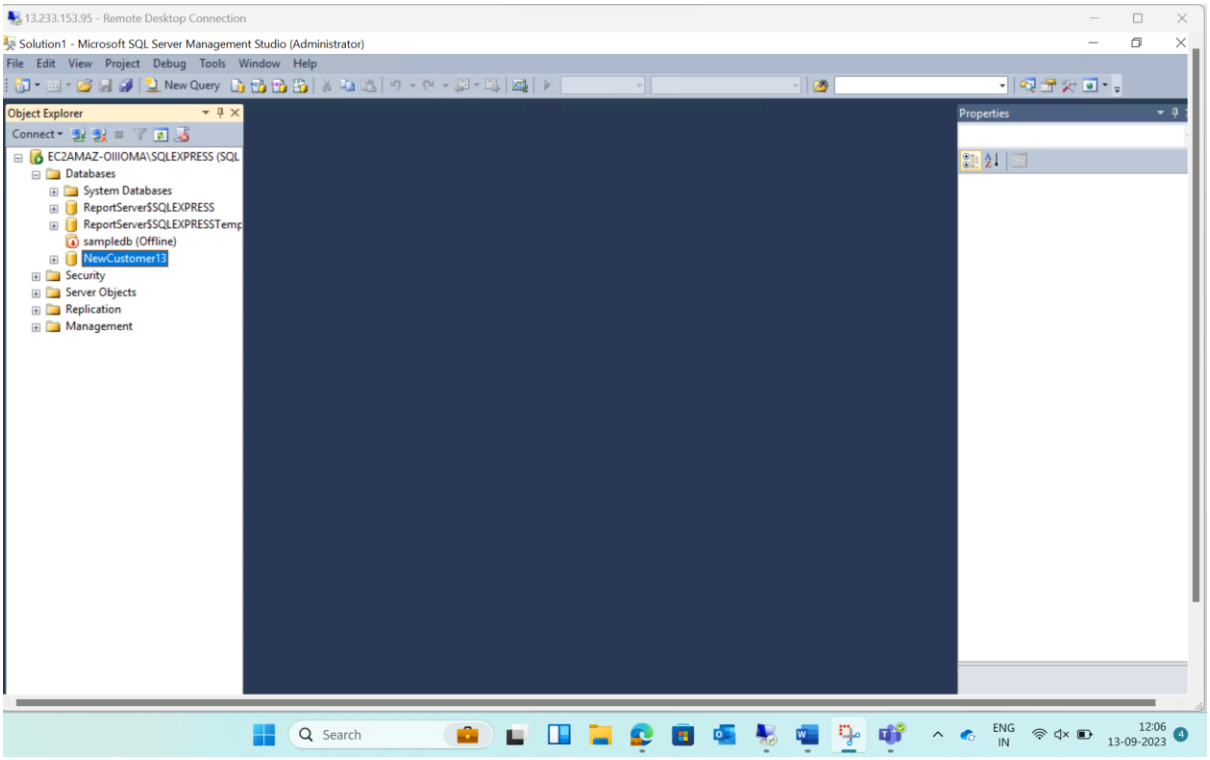
1. https://www.bankrate.com/retirement/best-and-worst-states-for-retirement/#best-and-worst
2. https://github.com/Microsoft/sql-server-samples/releases/tag/adventureworks

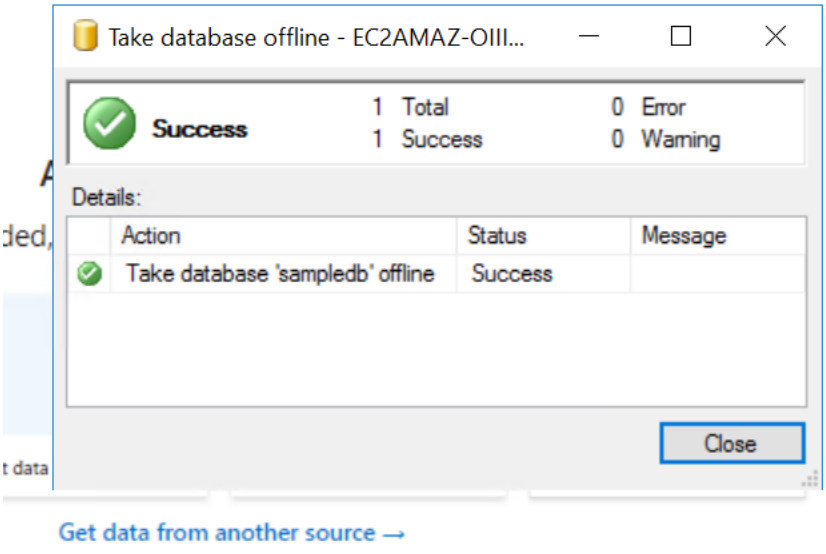
Now open Power BI app

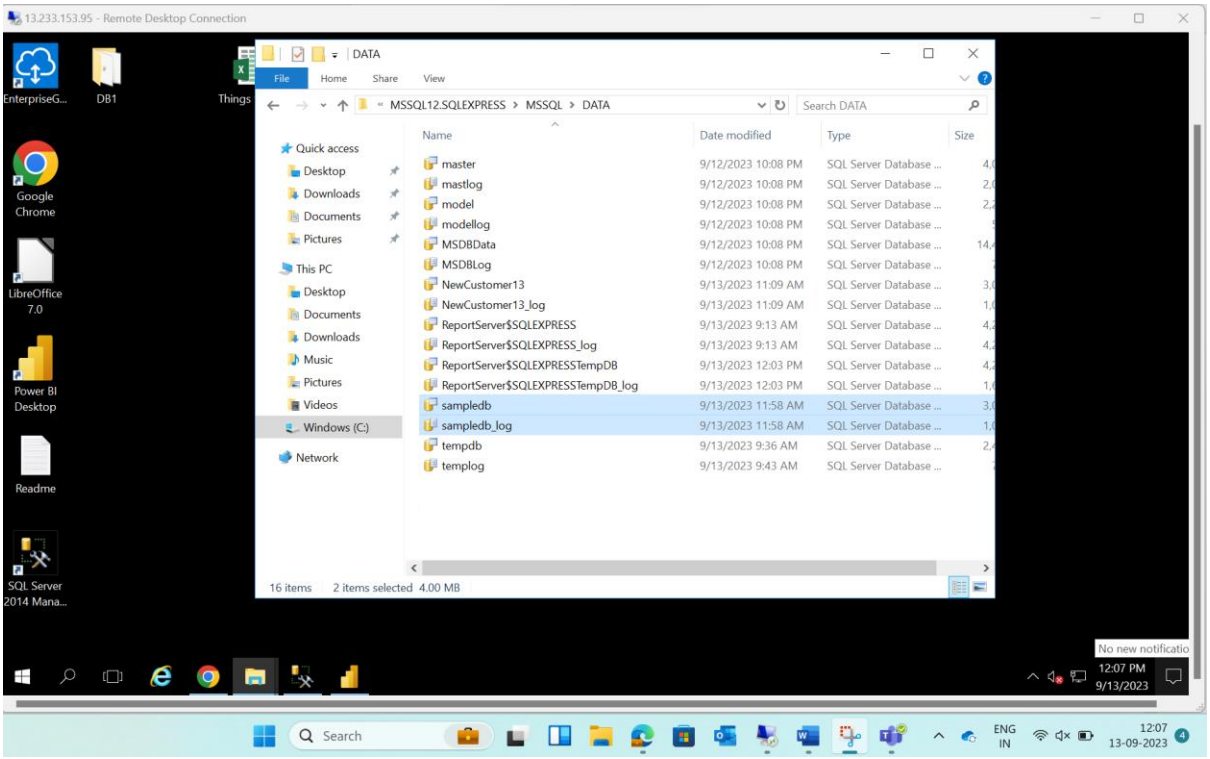
Before that make backup

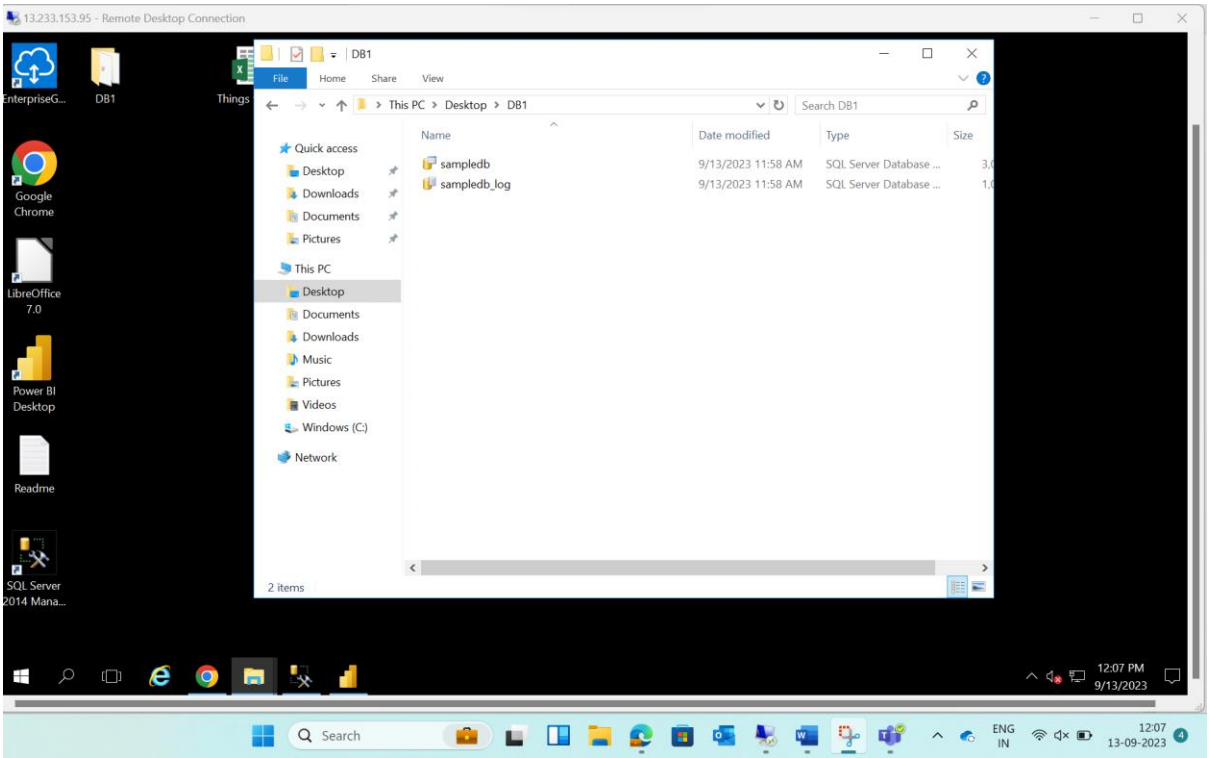
Select database jaha table hai and right click and under tasks

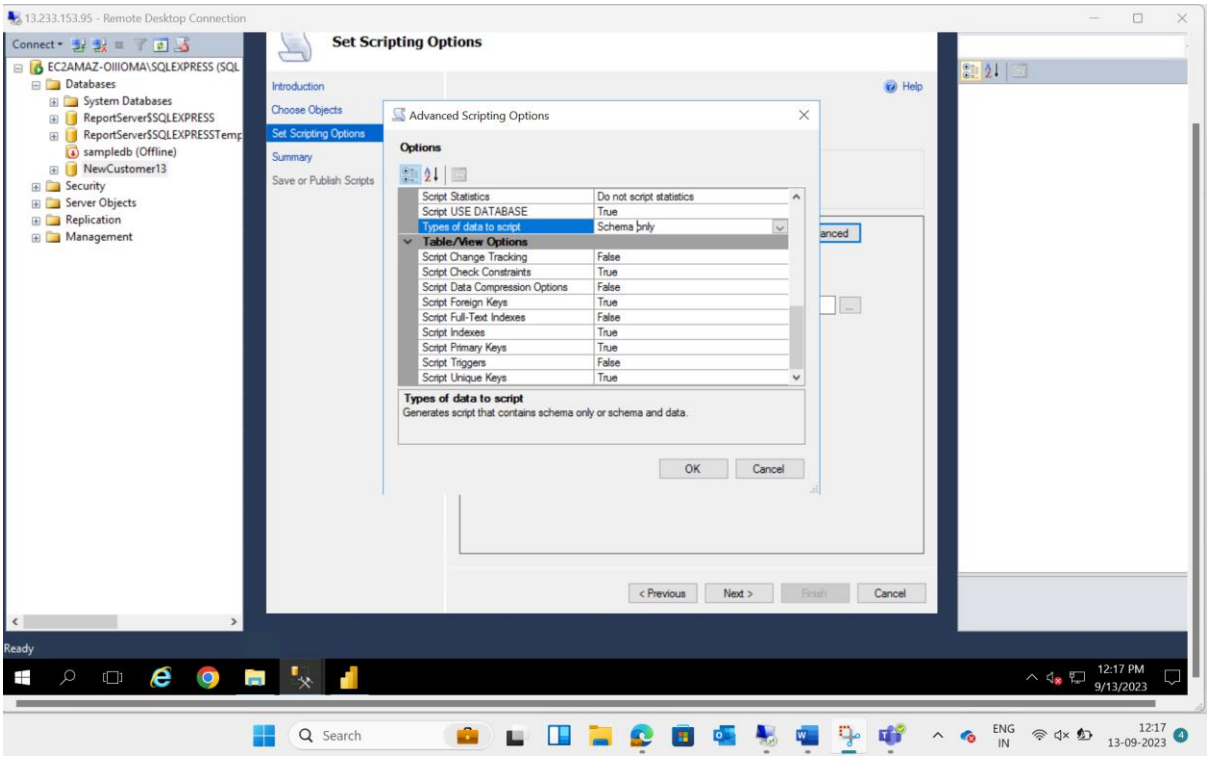
Tasks-> take offline to copy->then after that go to vm ka files explorer in it windows c drive -> programs files->Microsoft sql server-> data-> the table files both mdf and log files of data base and paste it on new folder created in desktop of vm

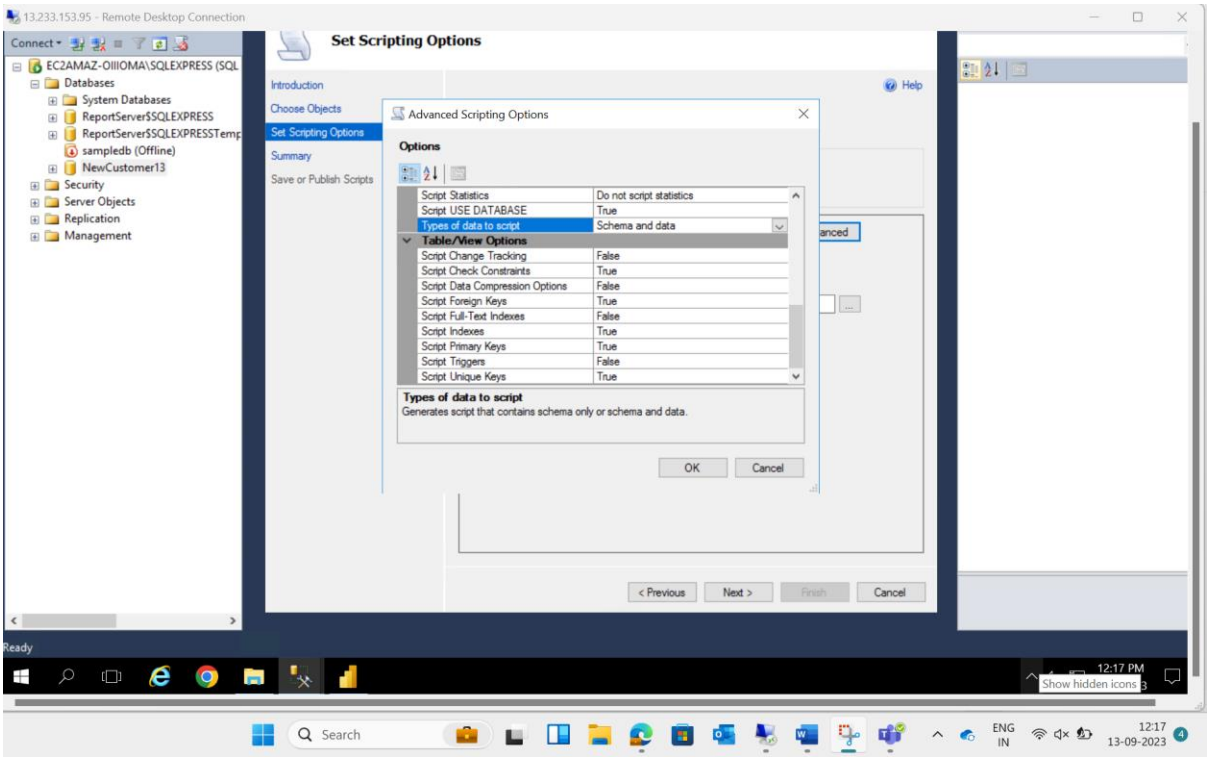



13.233.153.95 - Remote Desktop Connection
DATA
File Home Share View
MSSQL12.SQLEXPRESS › MSSQL › DATA
Search DATA
Quick access
Desktop
Downloads
Documents
Pictures
This PC
Desktop
Documents
Downloads
Music
Pictures
Videos
Windows (C:)
Network
Name
Date modified
Type
Size
master 9/12/2023 10:08 PM SQL Server Database ...
mastlog 9/12/2023 10:08 PM SQL Server Database ...
model 9/12/2023 10:08 PM SQL Server Database ...
modellog 9/12/2023 10:08 PM SQL Server Database ...
MSDBData 9/12/2023 10:08 PM SQL Server Database ...
MSDBLog 9/12/2023 10:08 PM SQL Server Database ...
NewCustomer13 9/13/2023 11:09 AM SQL Server Database ...
NewCustomer13_log 9/13/2023 11:09 AM SQL Server Database ...
ReportServer$SQLEXPRESS 9/13/2023 9:13 AM SQL Server Database ...
ReportServer$SQLEXPRESS_log 9/13/2023 9:13 AM SQL Server Database ...
ReportServer$SQLEXPRESSTempDB 9/13/2023 12:03 PM SQL Server Database ...
ReportServer$SQLEXPRESSTempDB_log 9/13/2023 12:03 PM SQL Server Database ...
sampledb 9/13/2023 11:58 AM SQL Server Database ...
sampledb_log 9/13/2023 11:58 AM SQL Server Database ...
tempdb 9/13/2023 9:36 AM SQL Server Database ...
templog 9/13/2023 9:43 AM SQL Server Database ...
16 items 2 items selected 4.00 MB
No new notificatio
12:07 PM
9/13/2023
Search
ENG
IN
12:07
13-09-2023


13.233.153.95 - Remote Desktop Connection
DB1
File Home Share View
This PC › Desktop › DB1
Search DB1
Quick access
Desktop
Downloads
Documents
Pictures
This PC
Desktop
Documents
Downloads
Music
Pictures
Videos
Windows (C:)
Network
Name
Date modified
Type
Size
sampledb 9/13/2023 11:58 AM SQL Server Database ...
sampledb_log 9/13/2023 11:58 AM SQL Server Database ...
2 items
12:07 PM
9/13/2023
Search
ENG
IN
12:07
13-09-2023



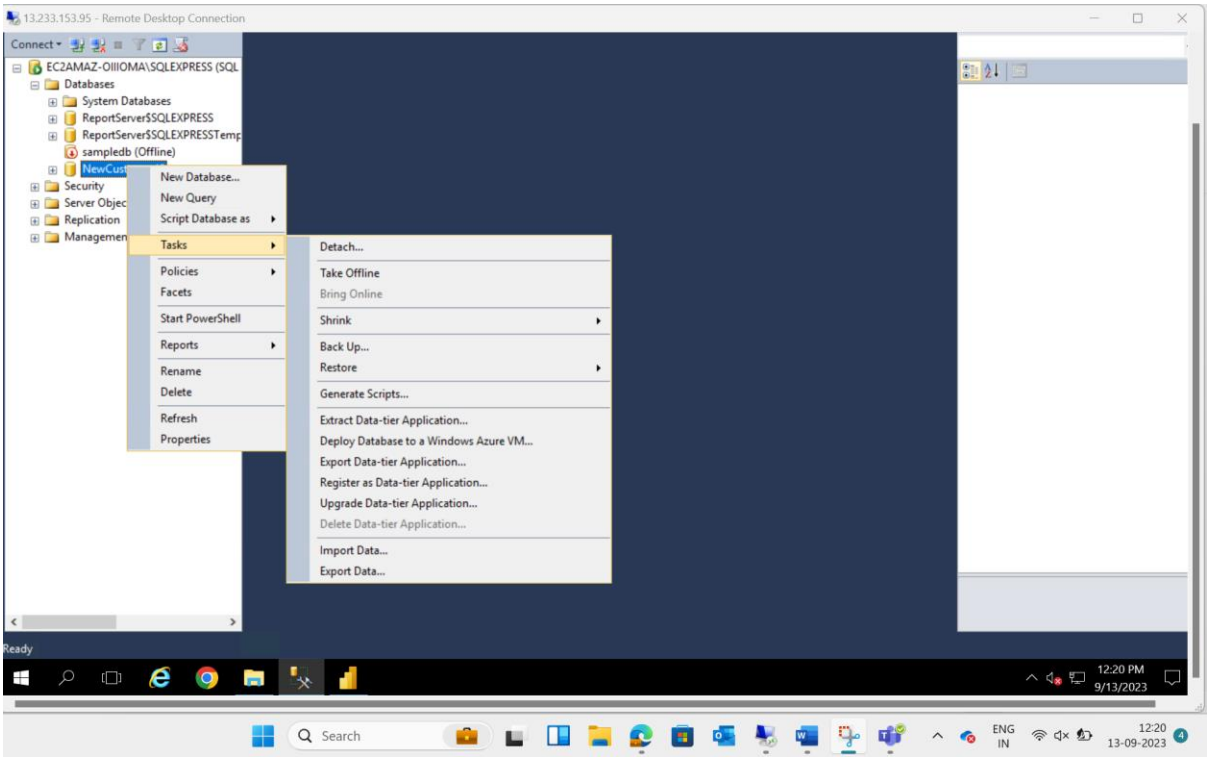



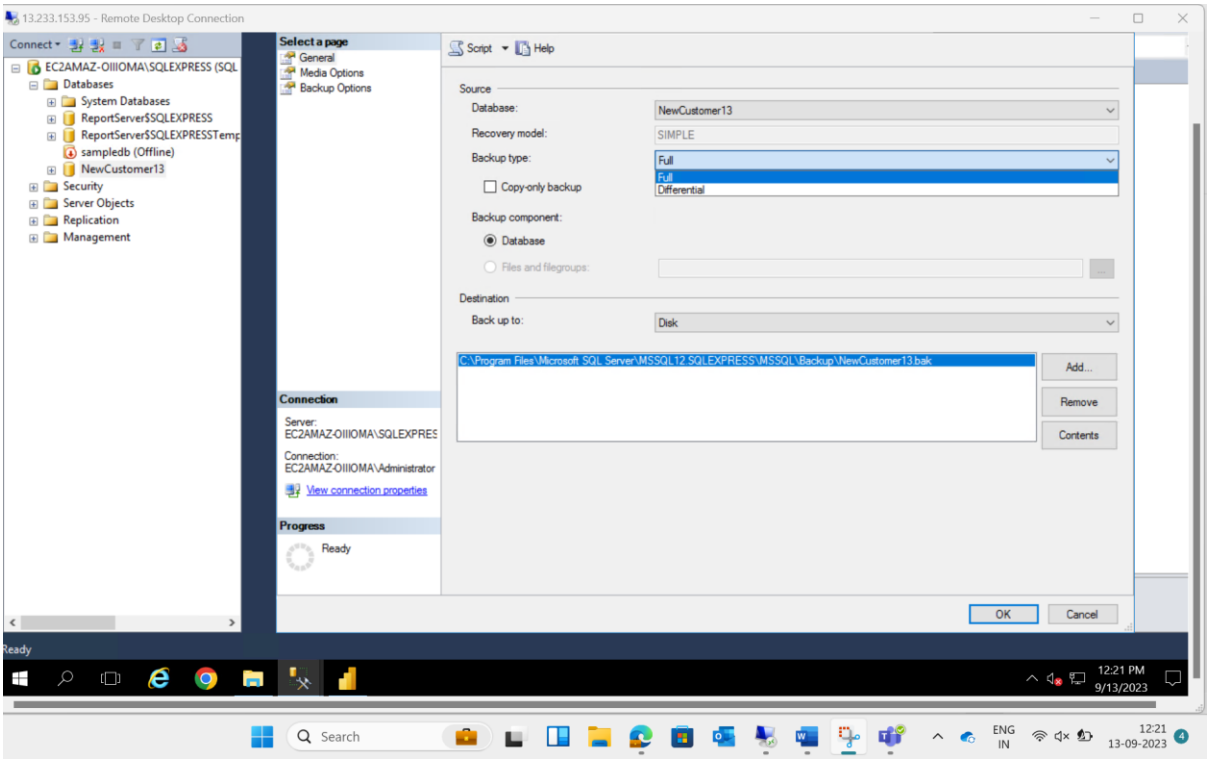

Select both option in it

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

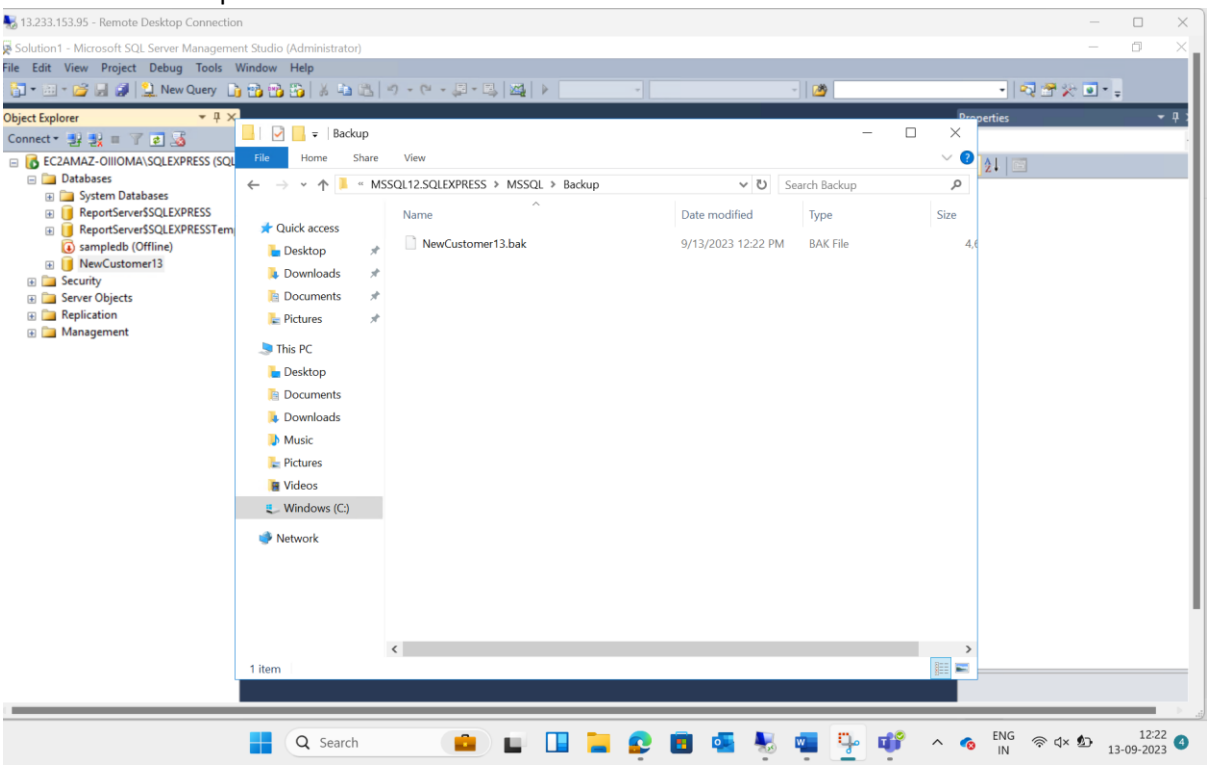
C:\Users\Administrator\Documents\script.sql

Then under same db-> tasks-> backup

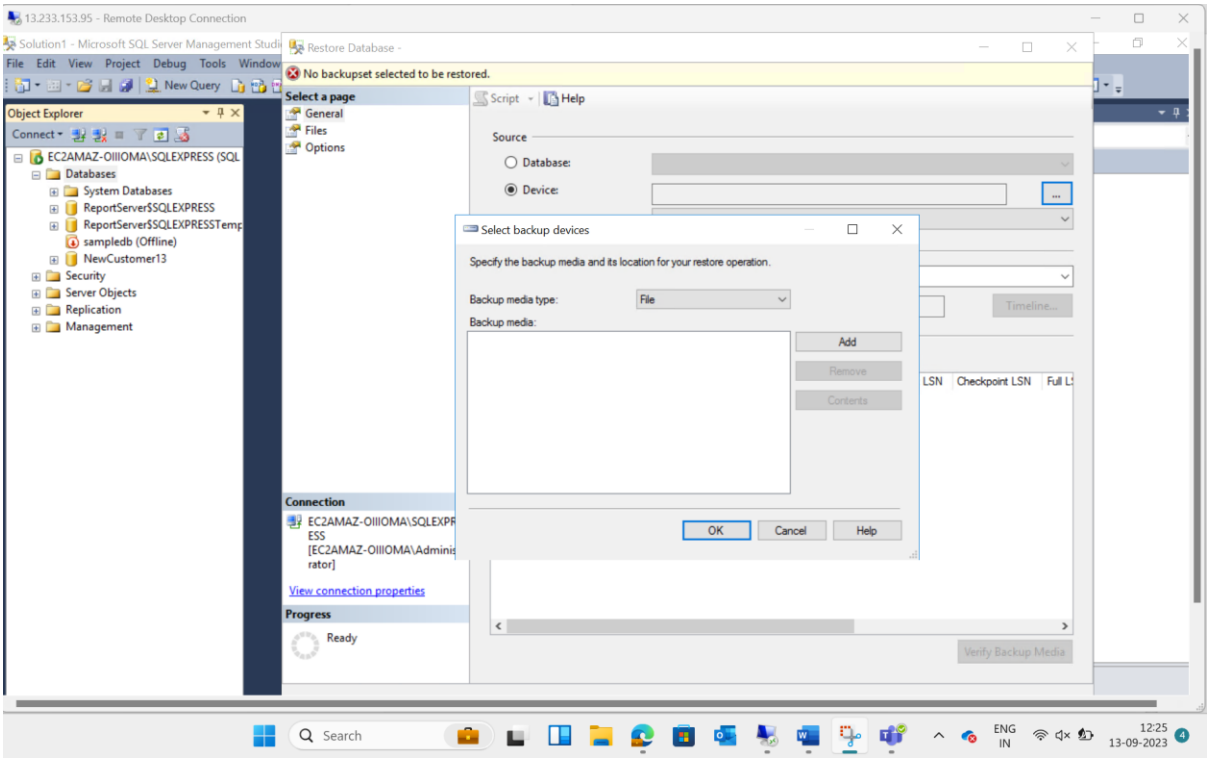


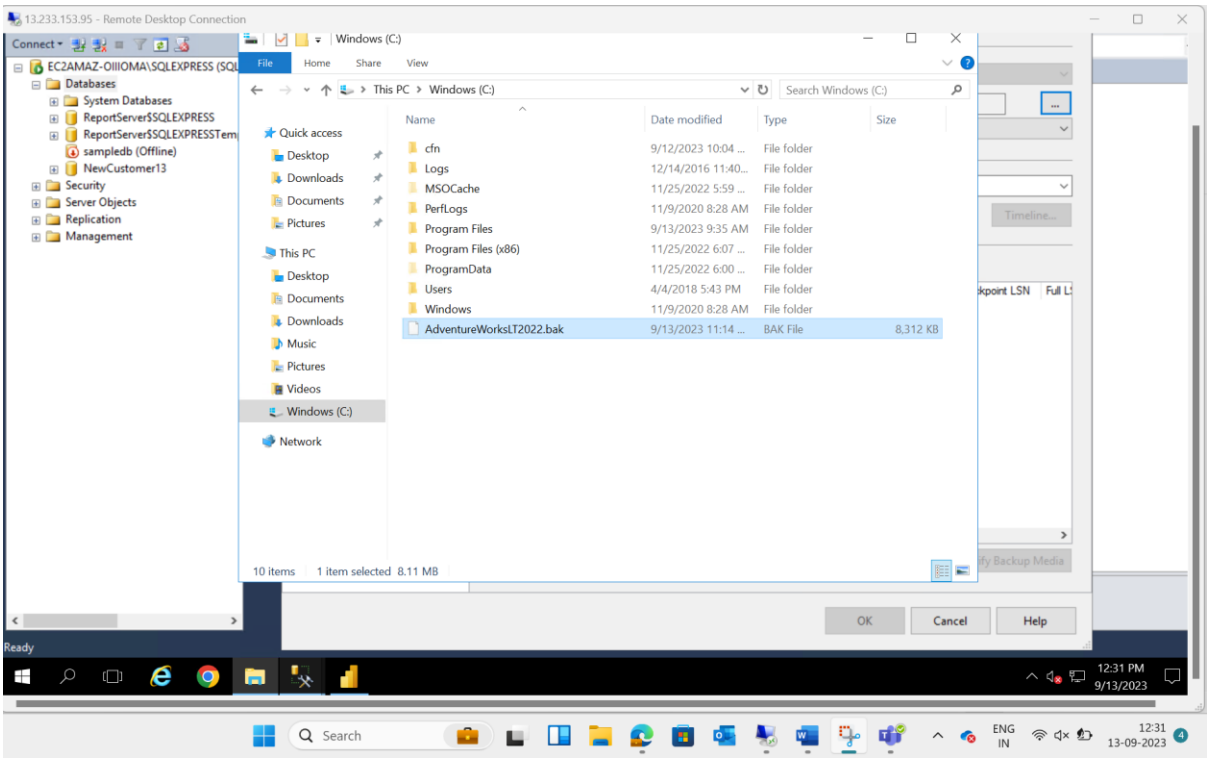
Can see the backup file in folder



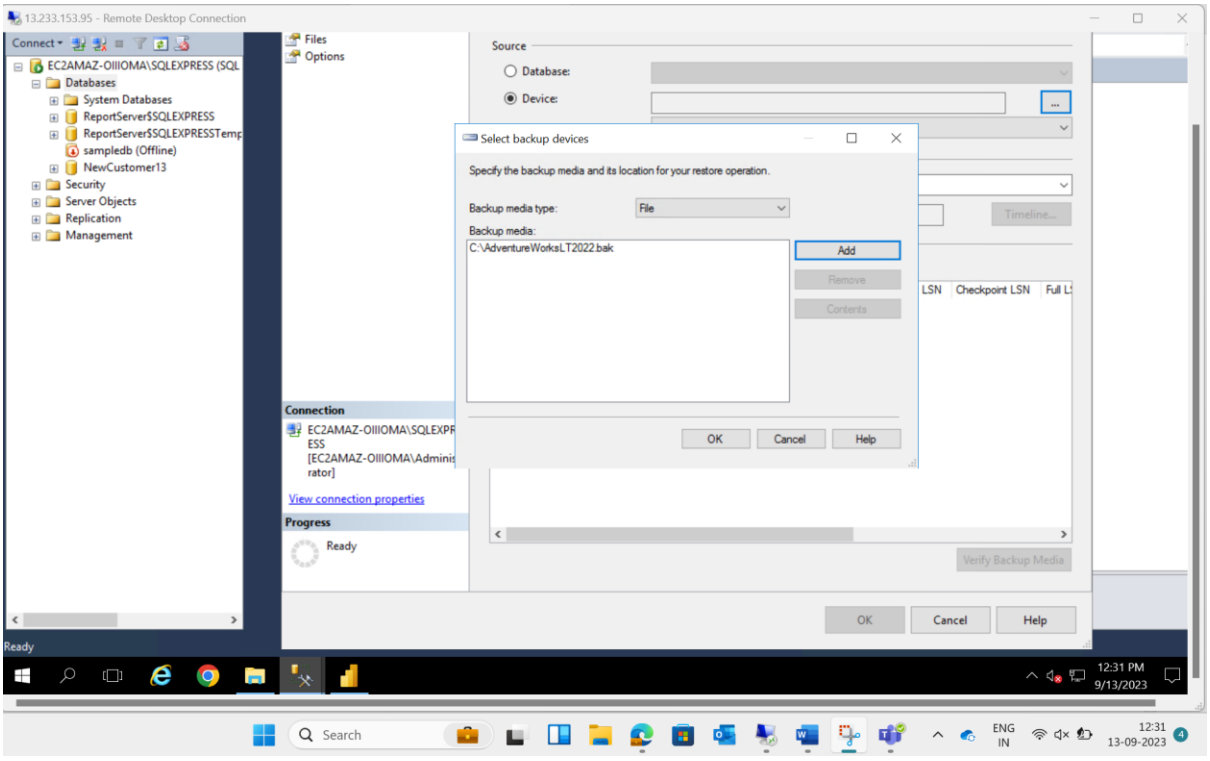

Under main db to restore the entire db we will do the following steps

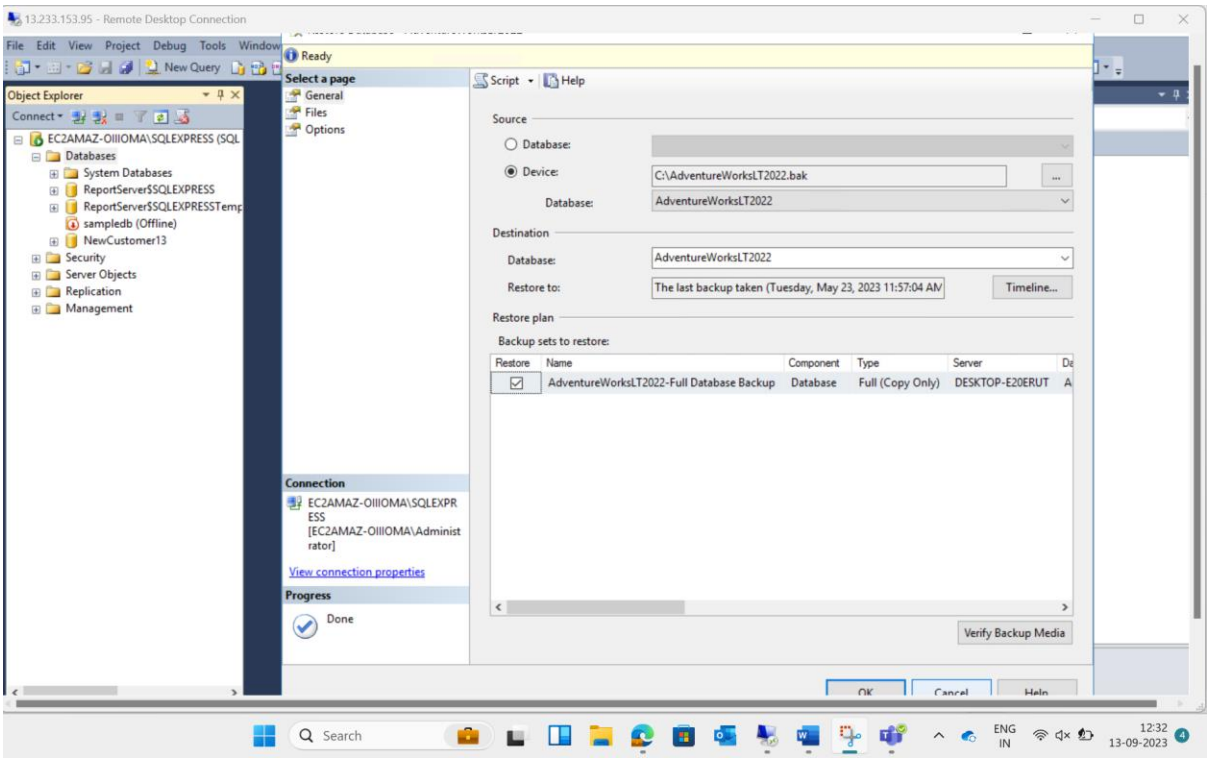


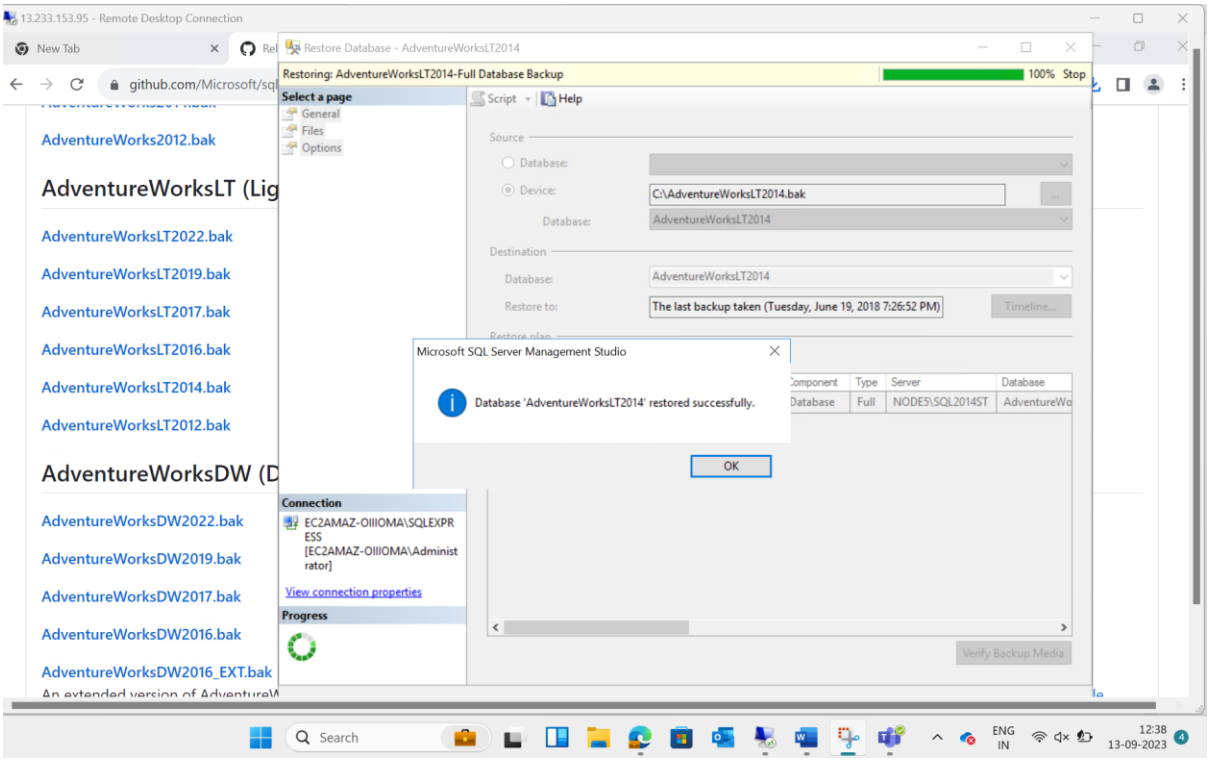

Copy the adventure works db downloaded file to c drive



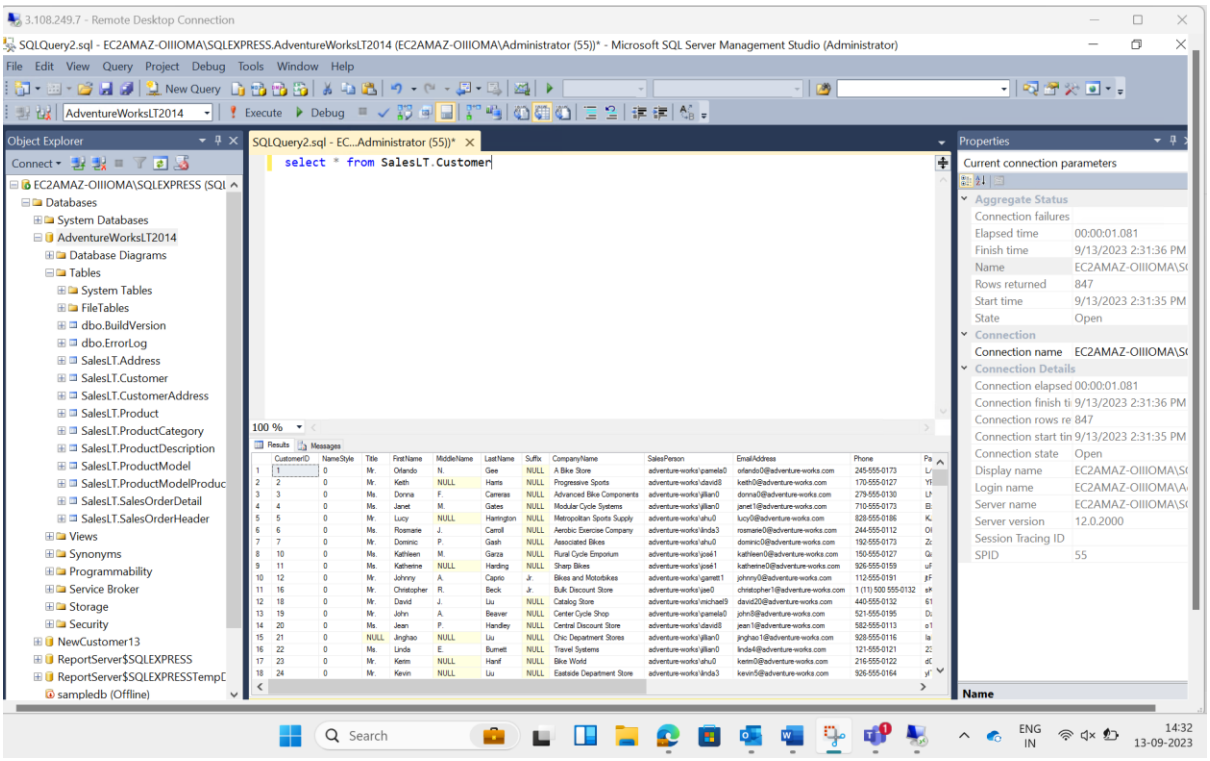

Then go to add and select the file-

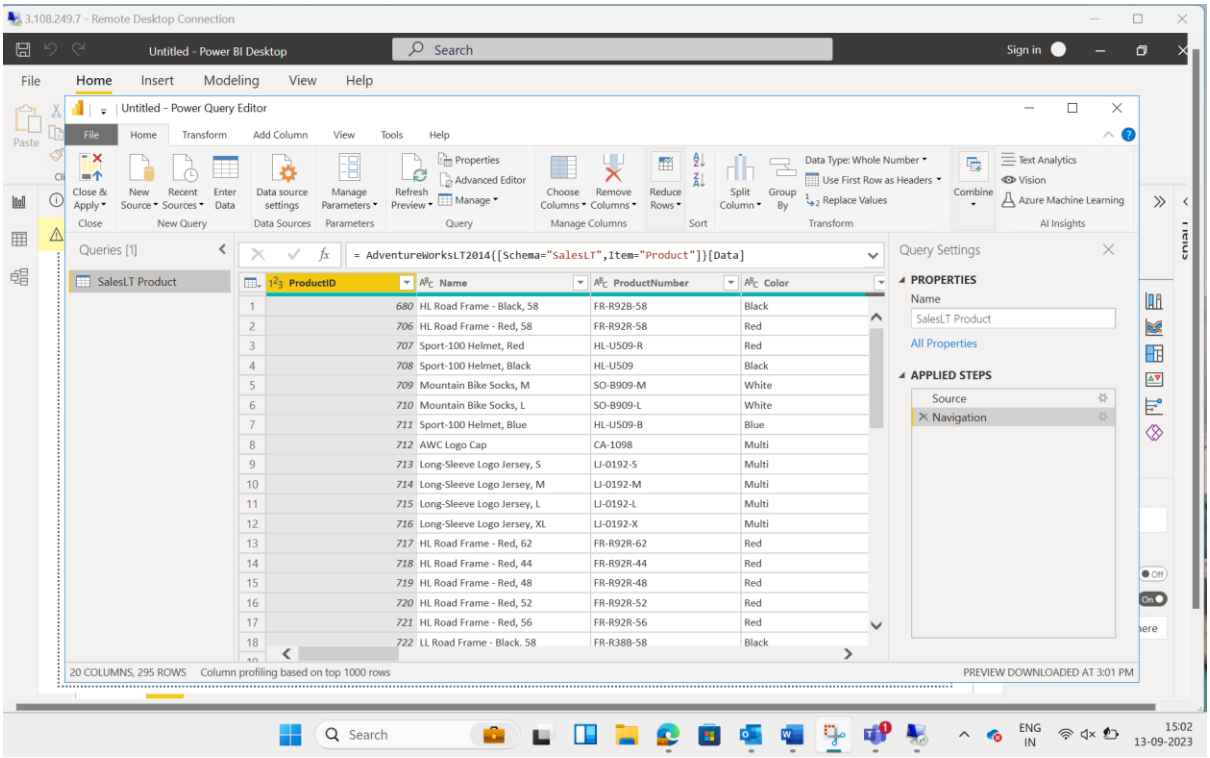





Now we will run queries under adventure db



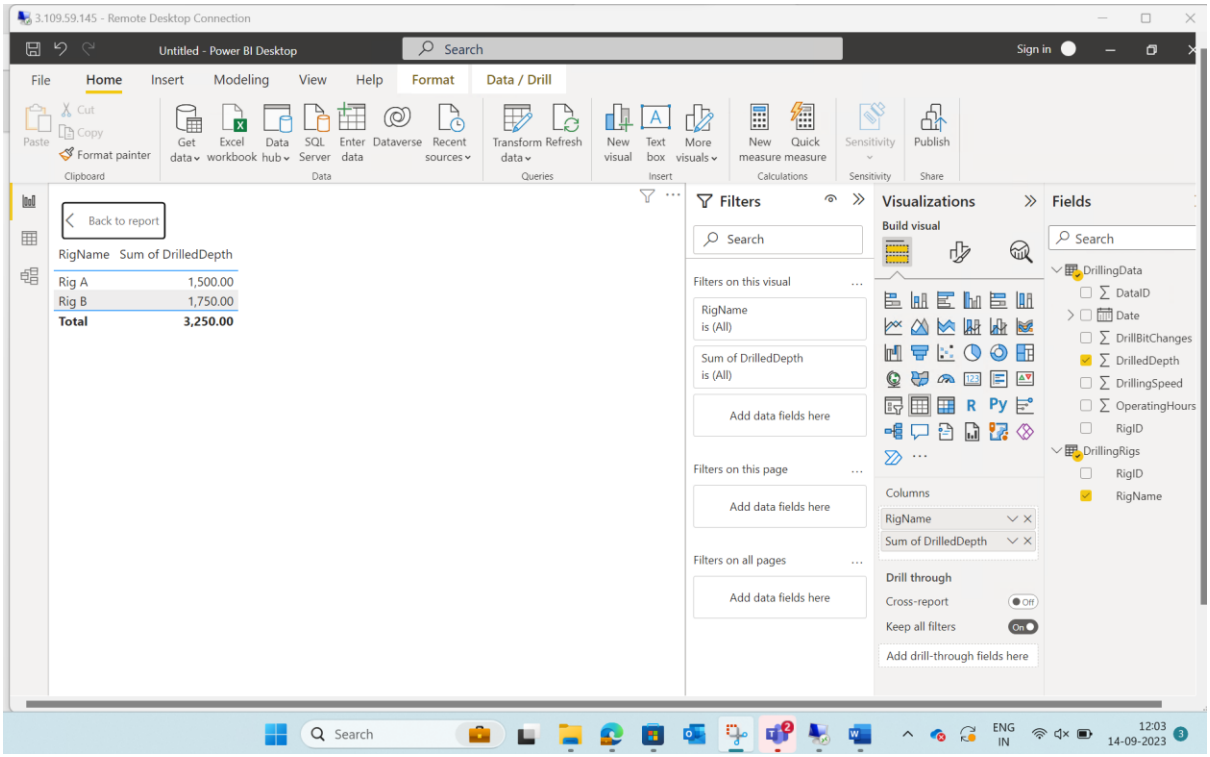




Now in PowerBI we need to go on sql data base-> type name of the main data base click next and create.

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




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 Jupter 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

Microsoft-> HD Insight

AWS-> EMR

GCP-> BigQuery

Data bricks (own cloud)

Storage<-Hadoop->Java

Memory<-Spark->Python/SQL

100x,R, JAVA, Distributed->Compute,storage

Compute->clustering->master slave architecture

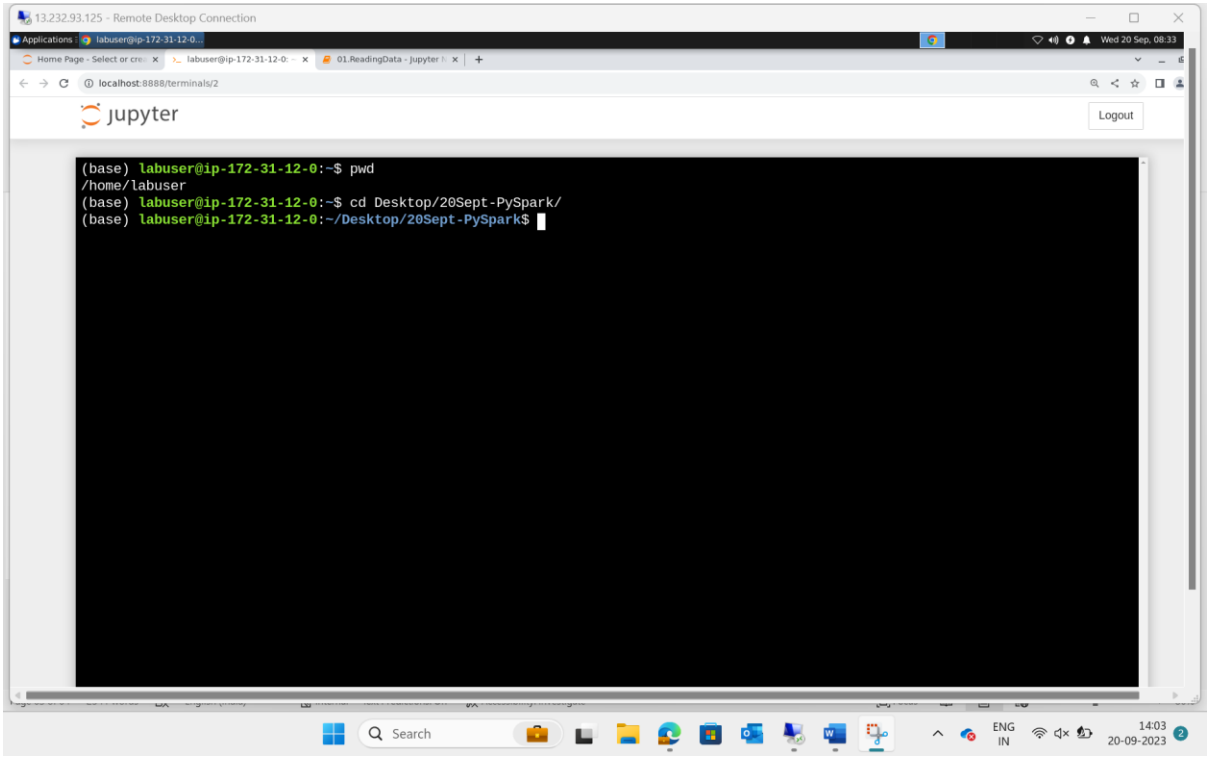
Difference between frameworks and library

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-

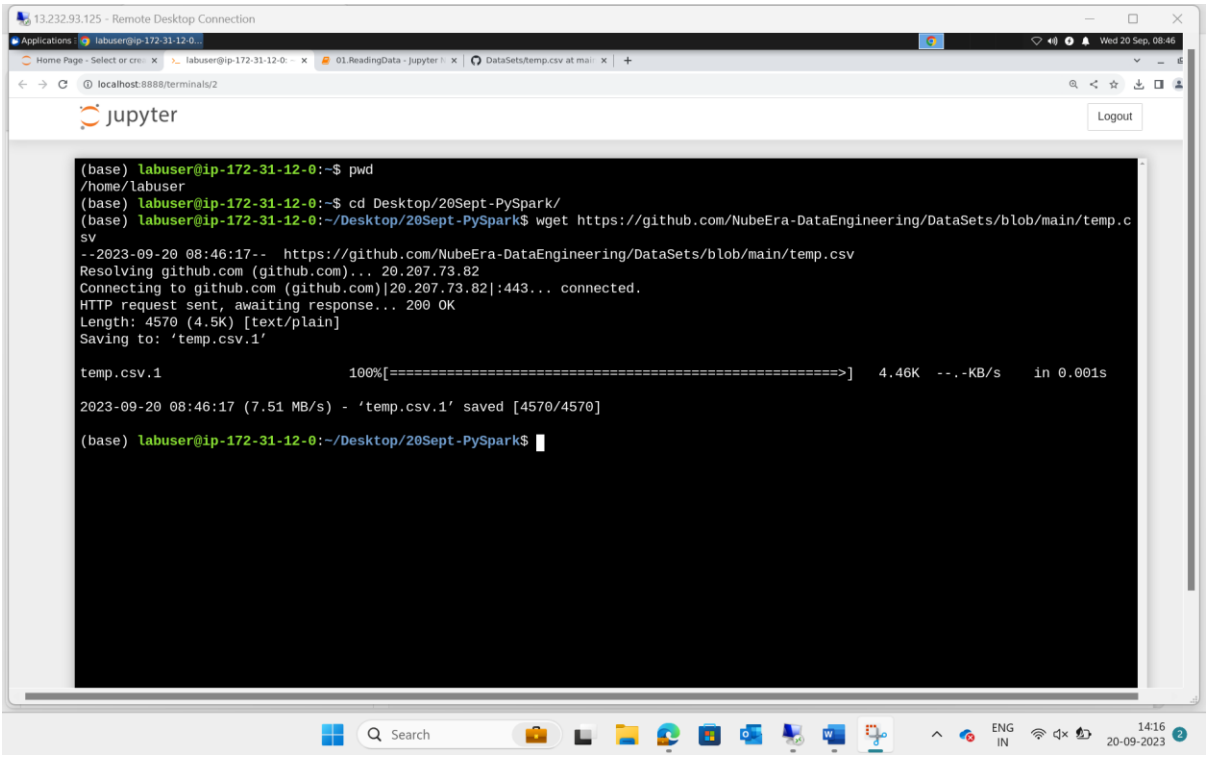
- Task is a single operation applied to a single partition.
- It is executed as a single thread/process in an executer.
- Why spark-
  -> You can perform batch processing in spark.
  -> it will support interactive sql.
  -> real time analysis
  -> structured, unstructured and semi structured data can 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

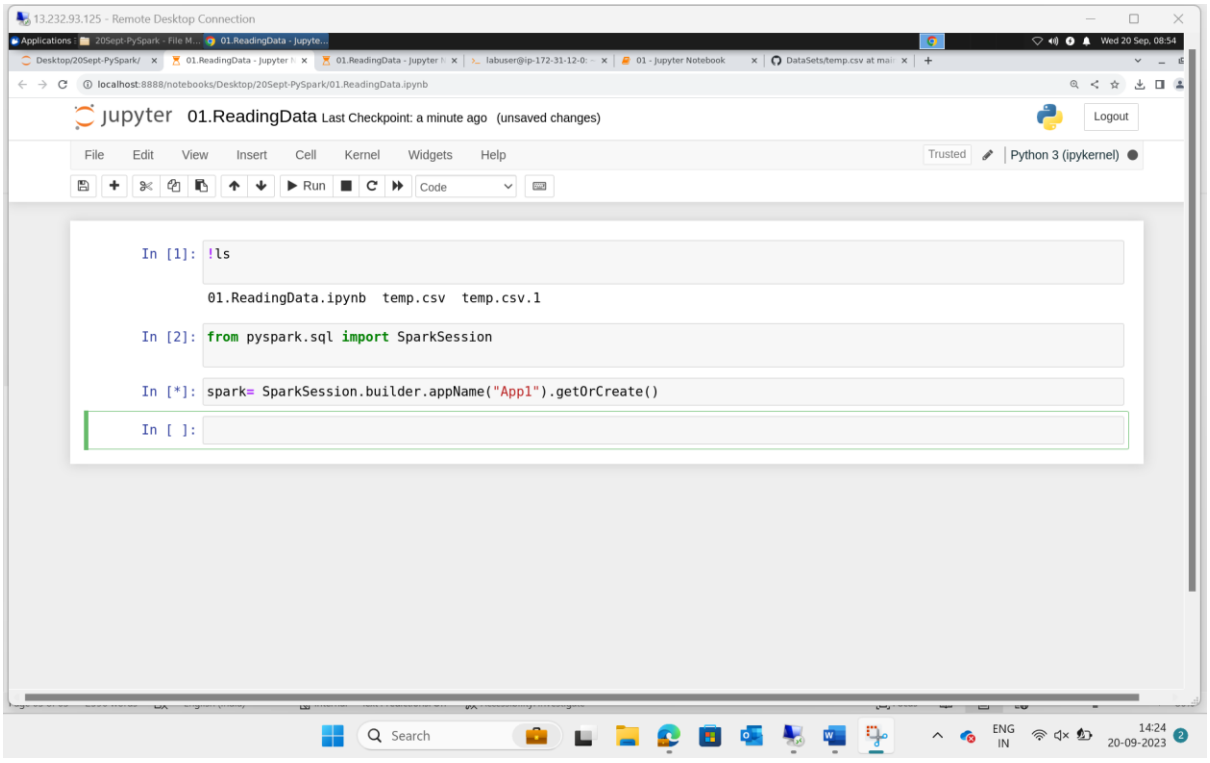
https://github.com/NubeEra-DataEngineering/DataSets

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

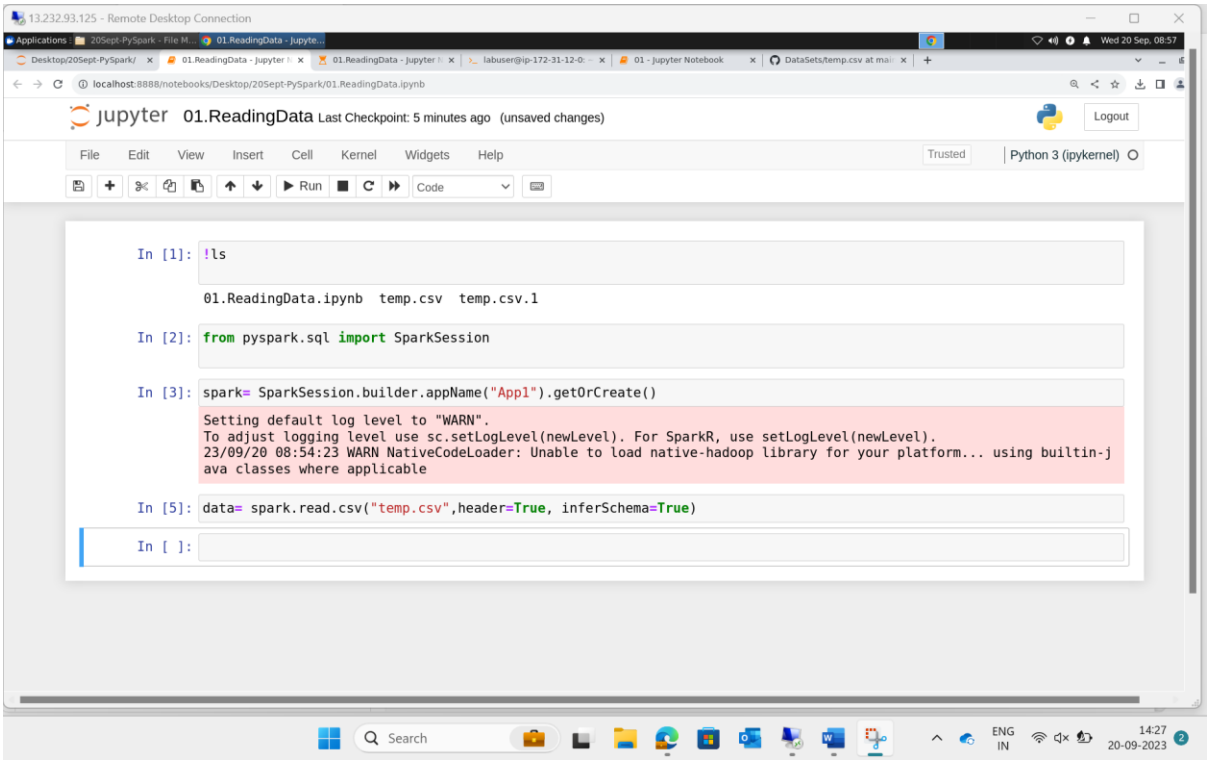


Using wget paste the url of the file

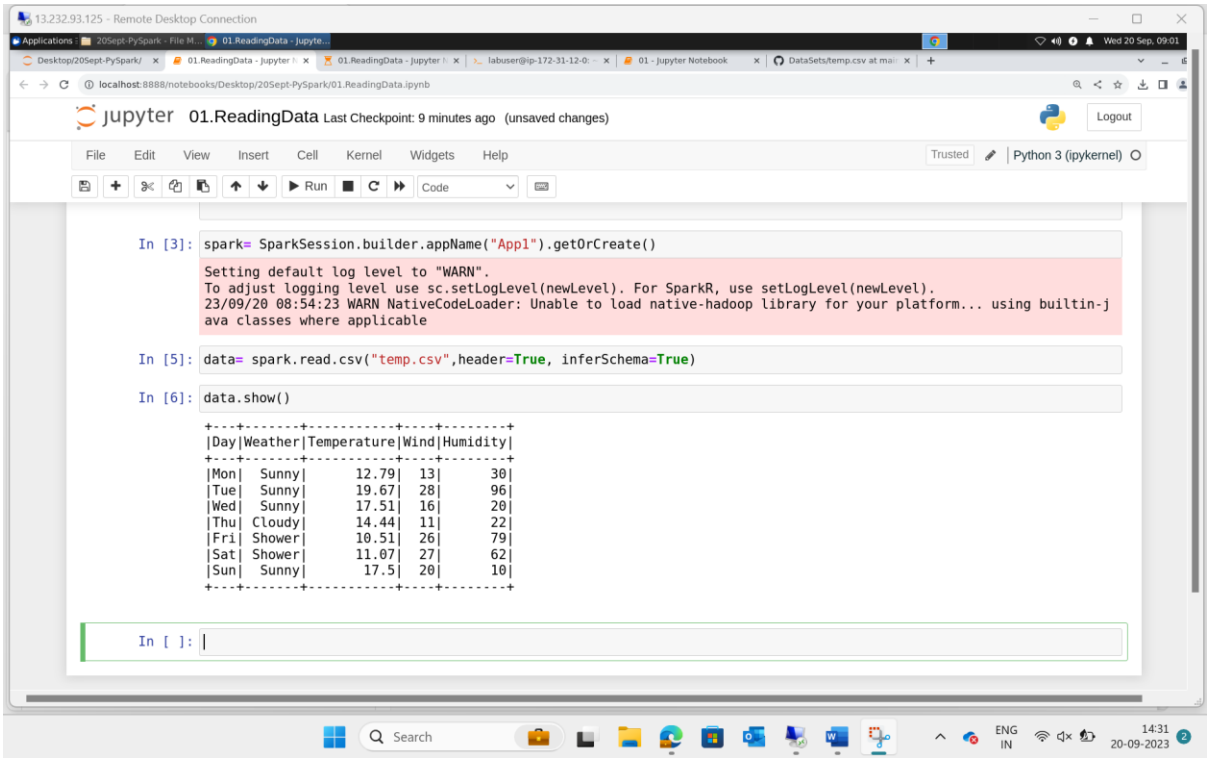
Now in a new python window type !ls to view all files. And create the file in the same desktop file



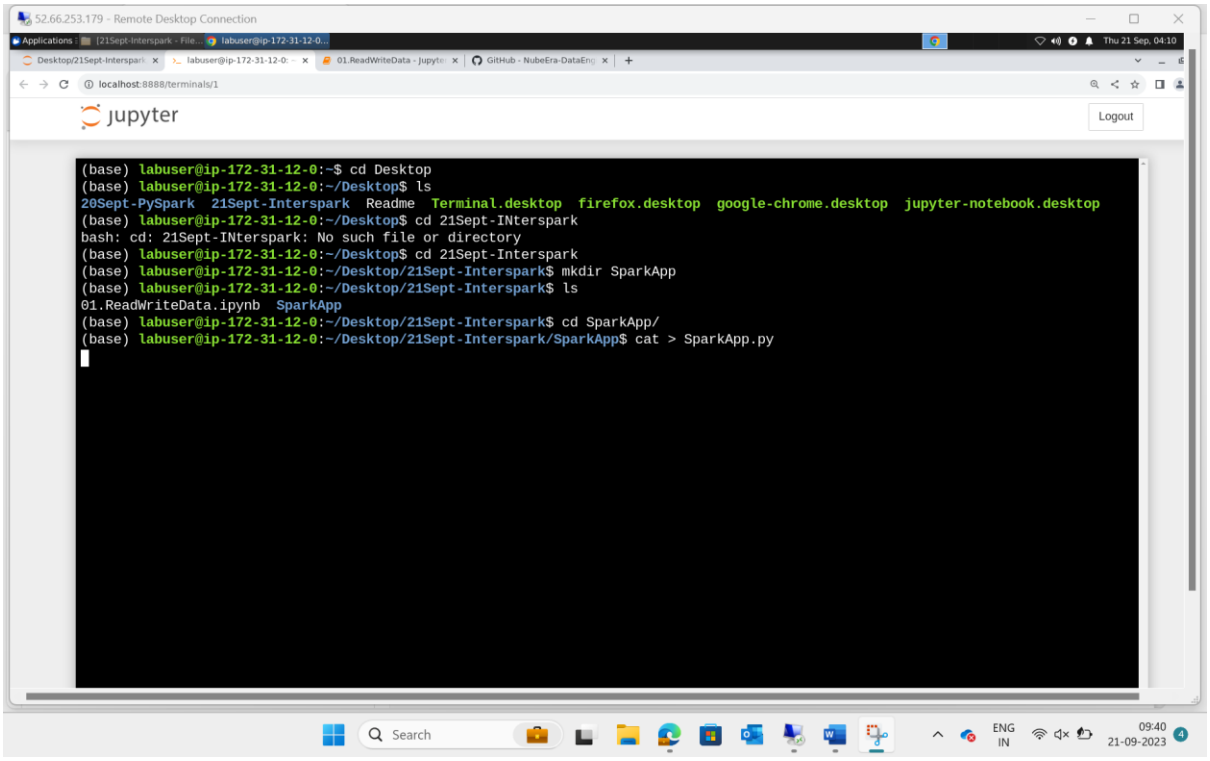

Ignore the warning

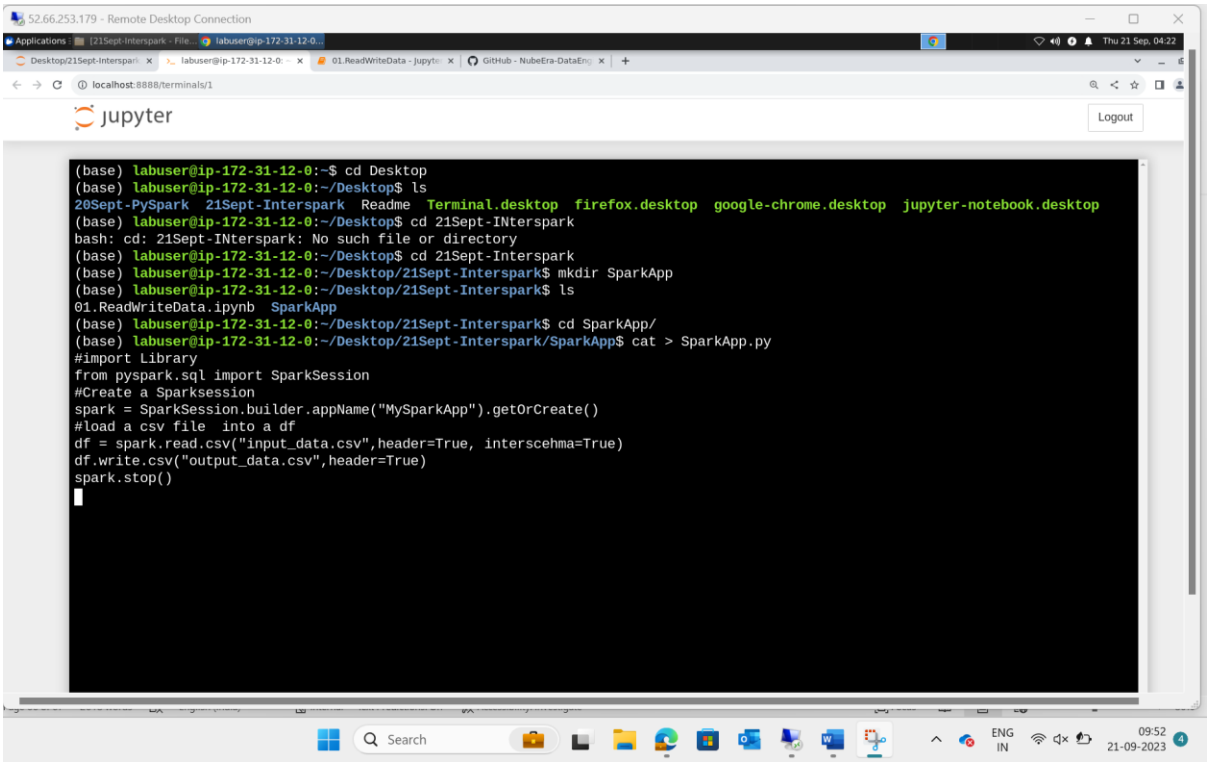


There are 2 types of path relative(/desktop/foldername/filename) and absolute path(.temp.csv).

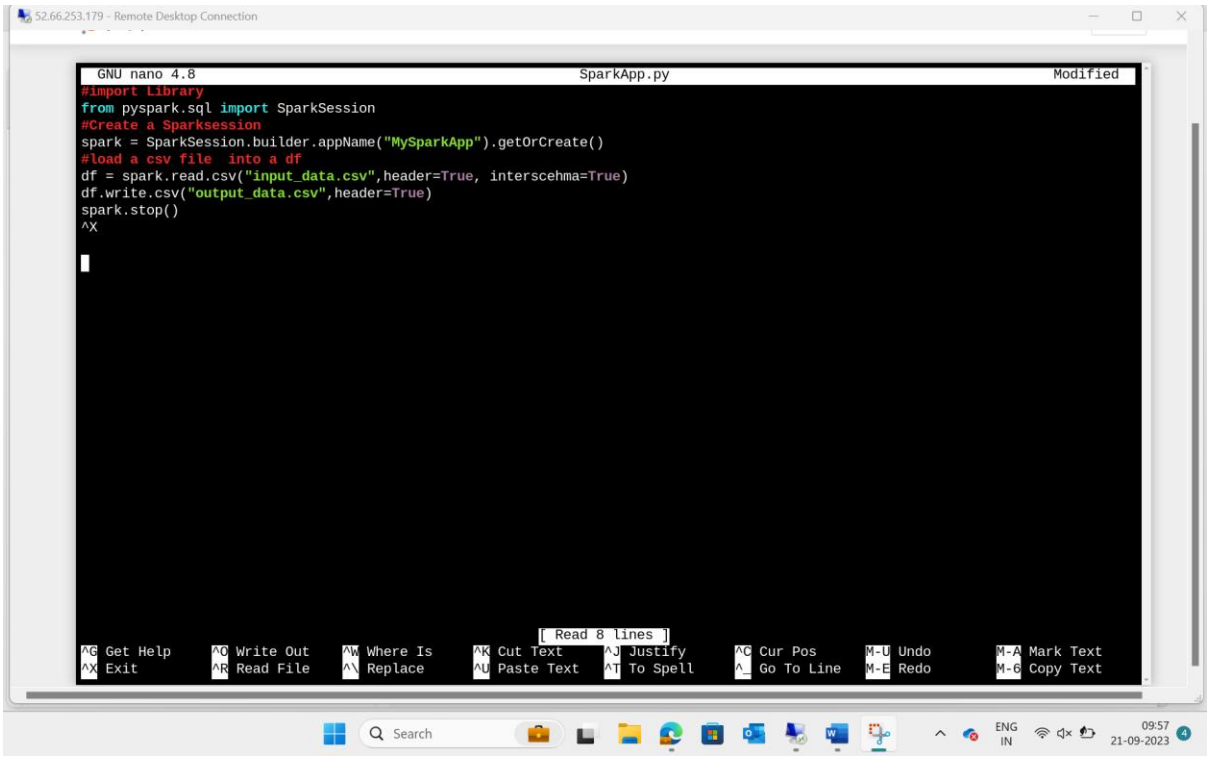



**Day 03(week 4) – 21th September 2023**





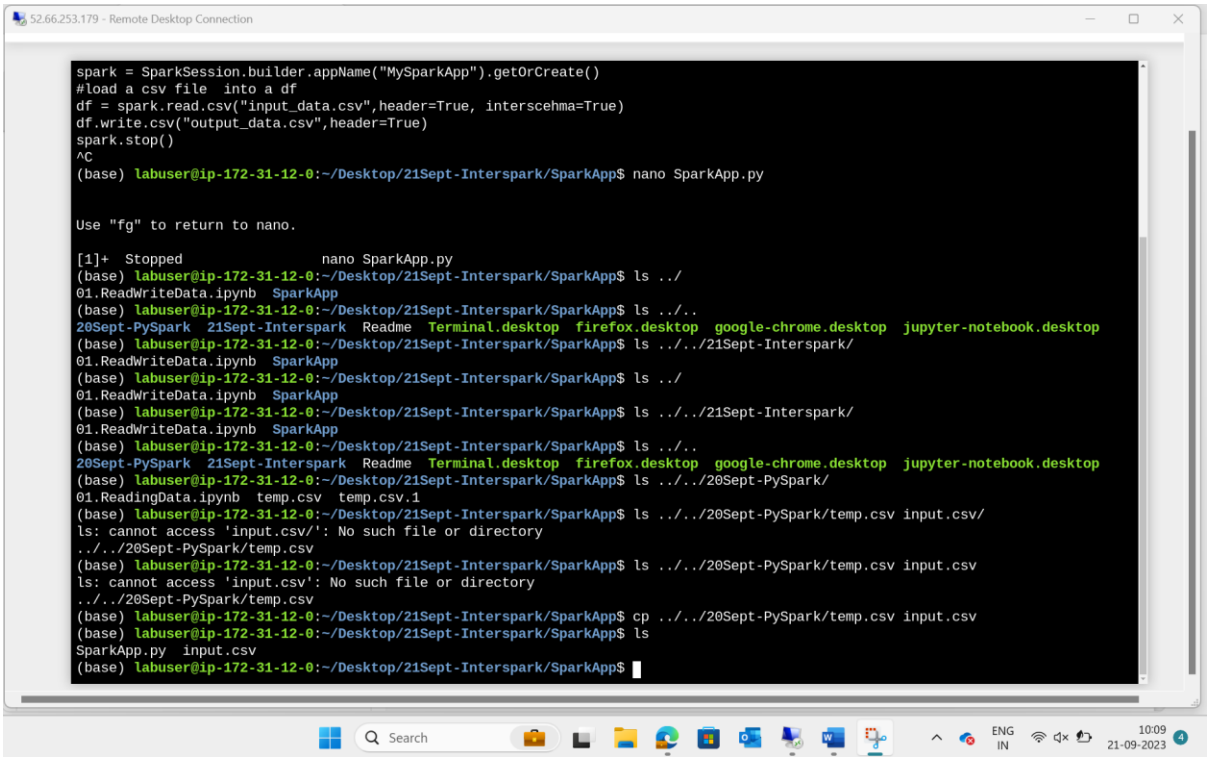

Can use nano SparkApp.py to write under this window as well it's the same baat




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, interscehma=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()
```




Ls ../

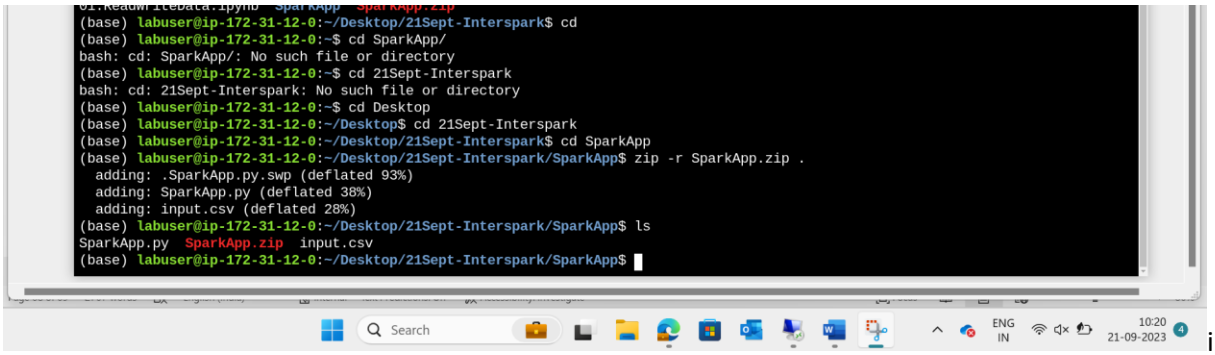
Ls ../../

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

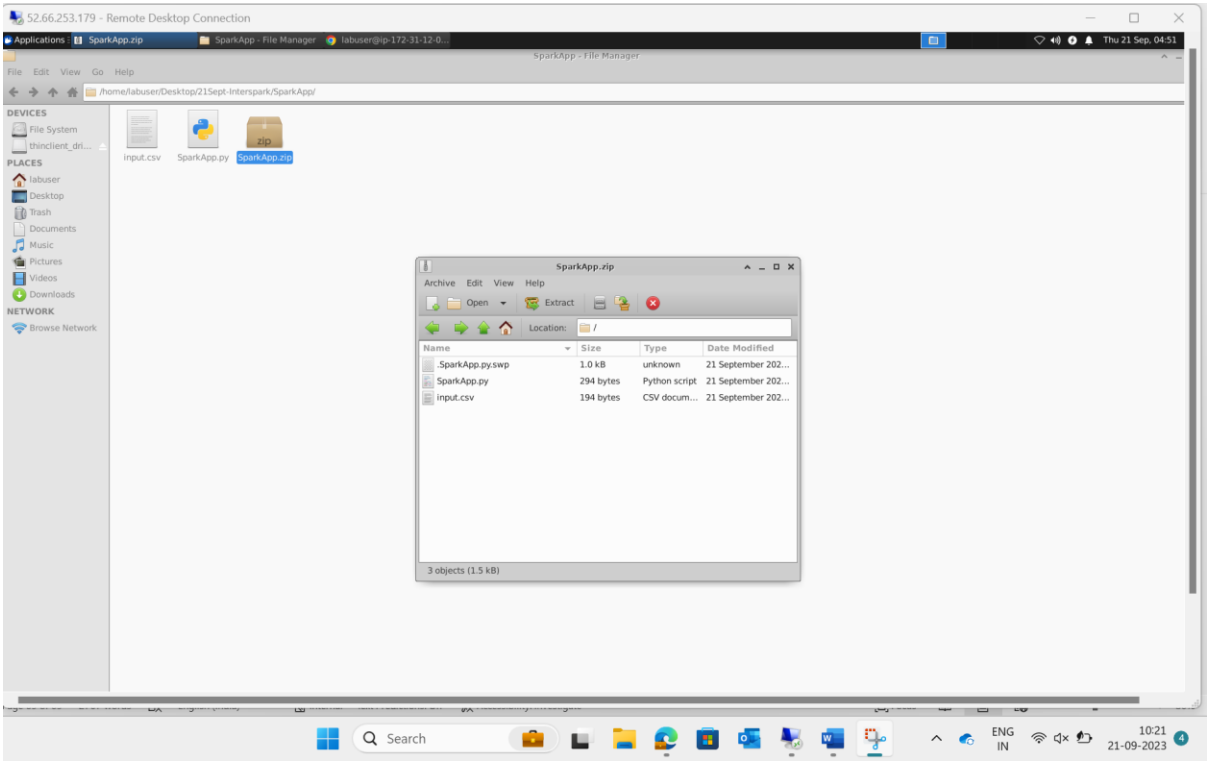
Now we need to go back one step after ls

Using cd ..

Then type zip -r SparkApp.zip SparkApp/




in desktop you can see this




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

SparkApp.py

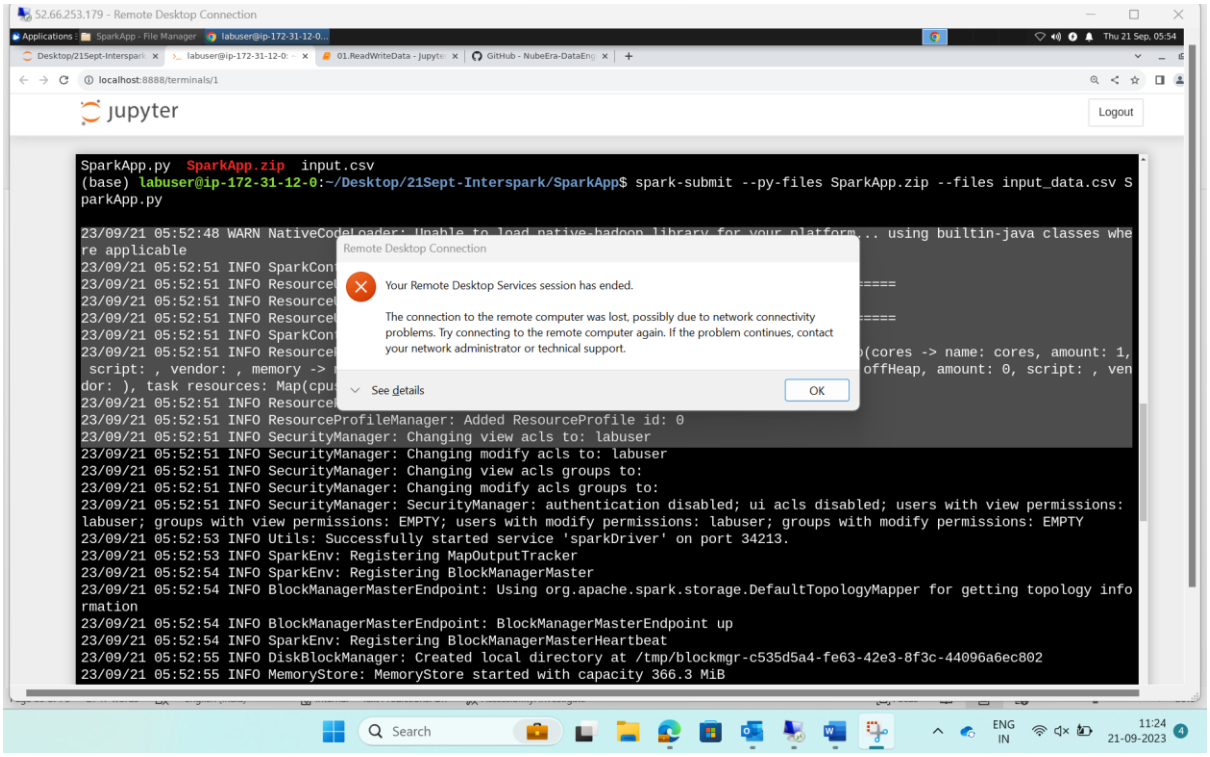
Cp \

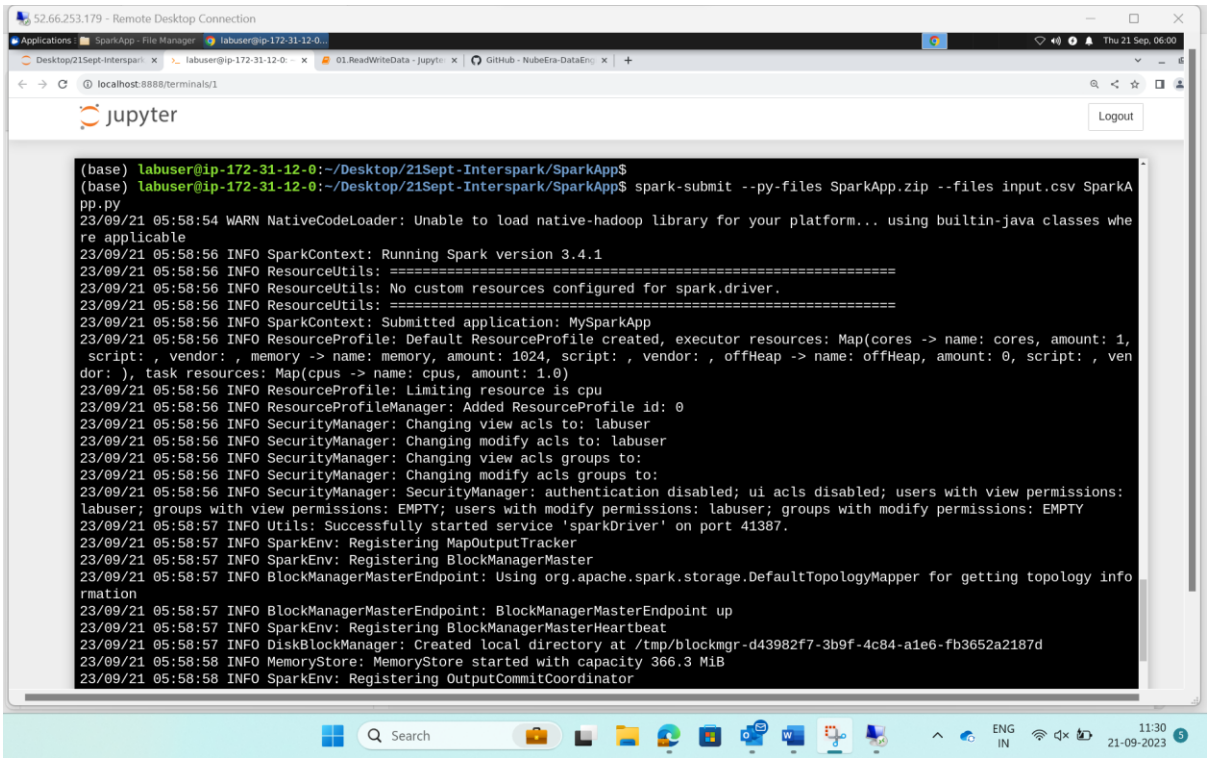
Source_path \

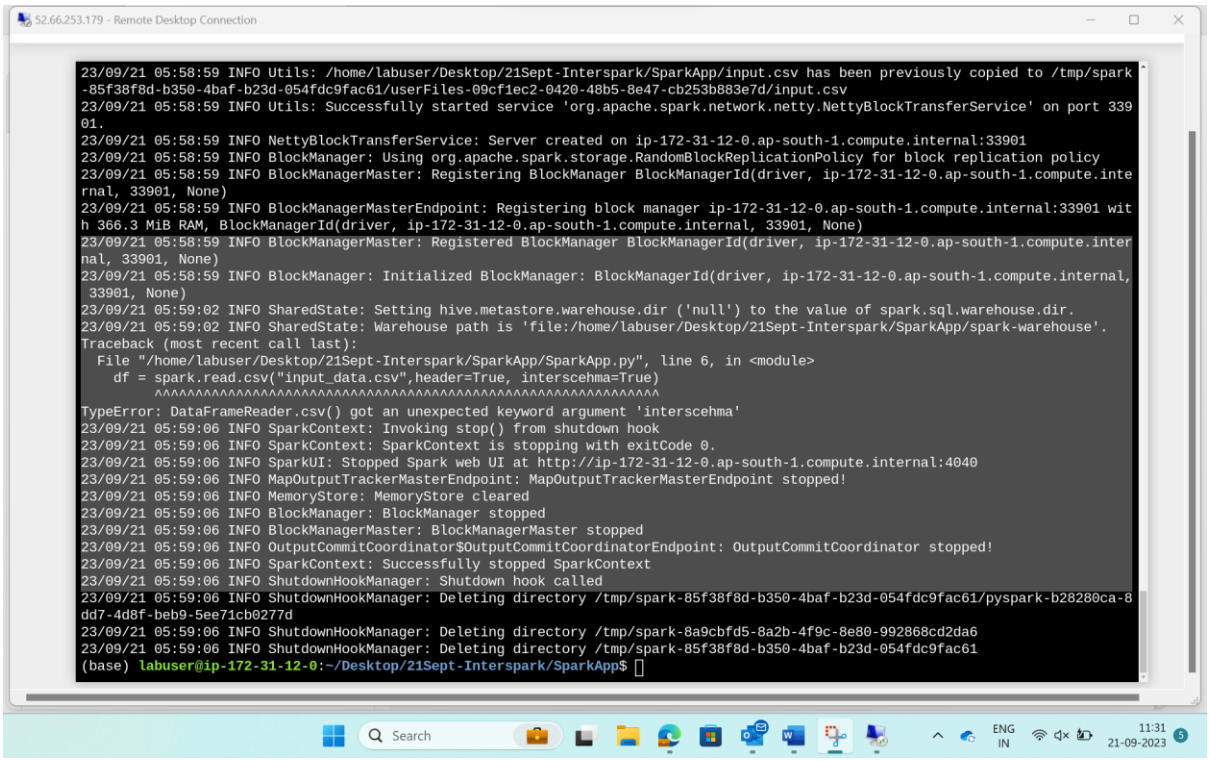
Target_path

➔ 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
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 SparkApp.py
23/09/21 05:52:48 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Remote Desktop Connection
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.
See details
OK
23/09/21 05:52:51 INFO ResourceProfileManager: 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 information
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
Search
ENG IN
11:24 21-09-2023


52.66.253.179 - Remote Desktop Connection
jupyter
Logout
(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: Default ResourceProfile 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(cpus -> 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 MiB
23/09/21 05:58:58 INFO SparkEnv: Registering OutputCommitCoordinator
Search
ENG IN
11:30 21-09-2023




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())

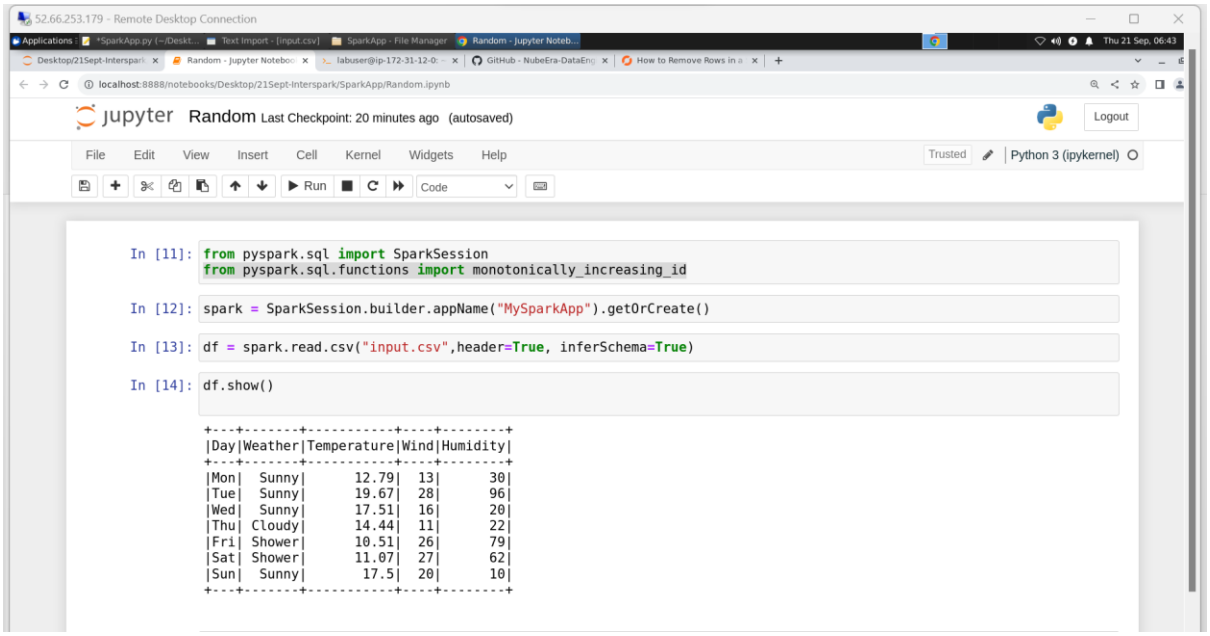
rows_to_remove = [0,1,2] } *for all rows type out*

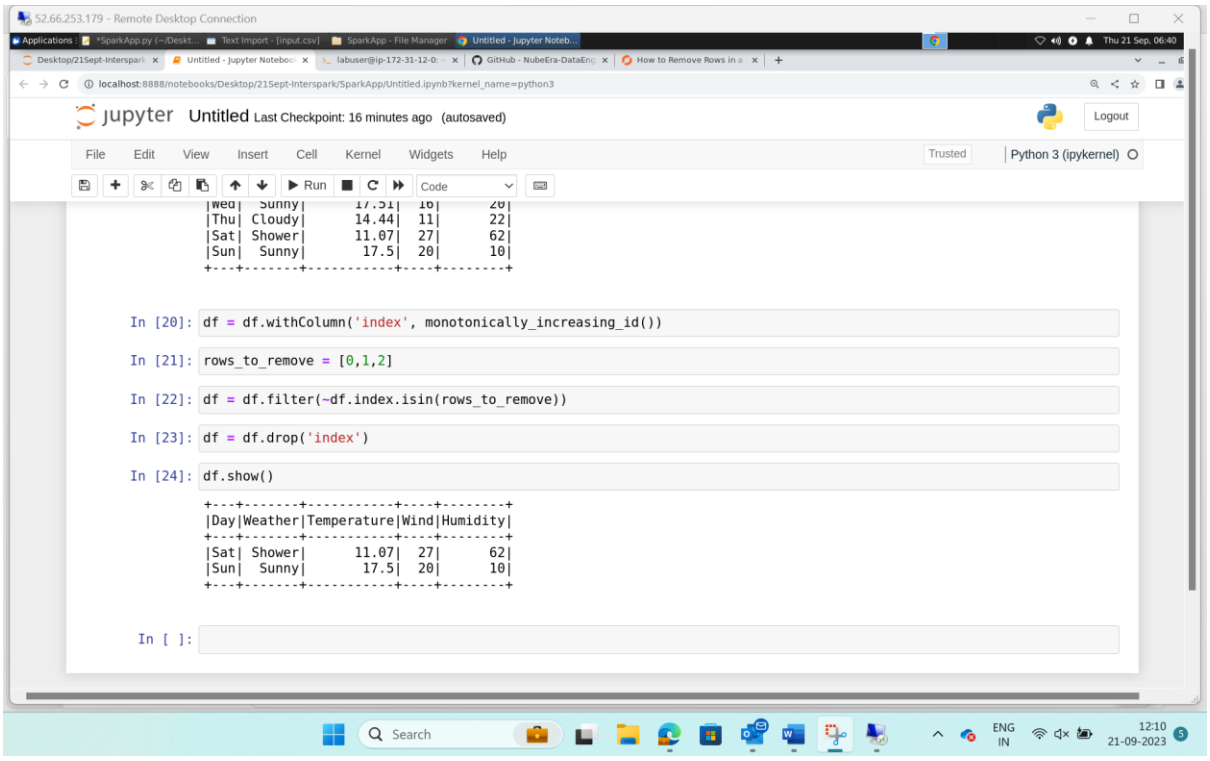
df = df.filter(~df.index.isin(rows_to_remove))

df = df.drop('index')

df.show()

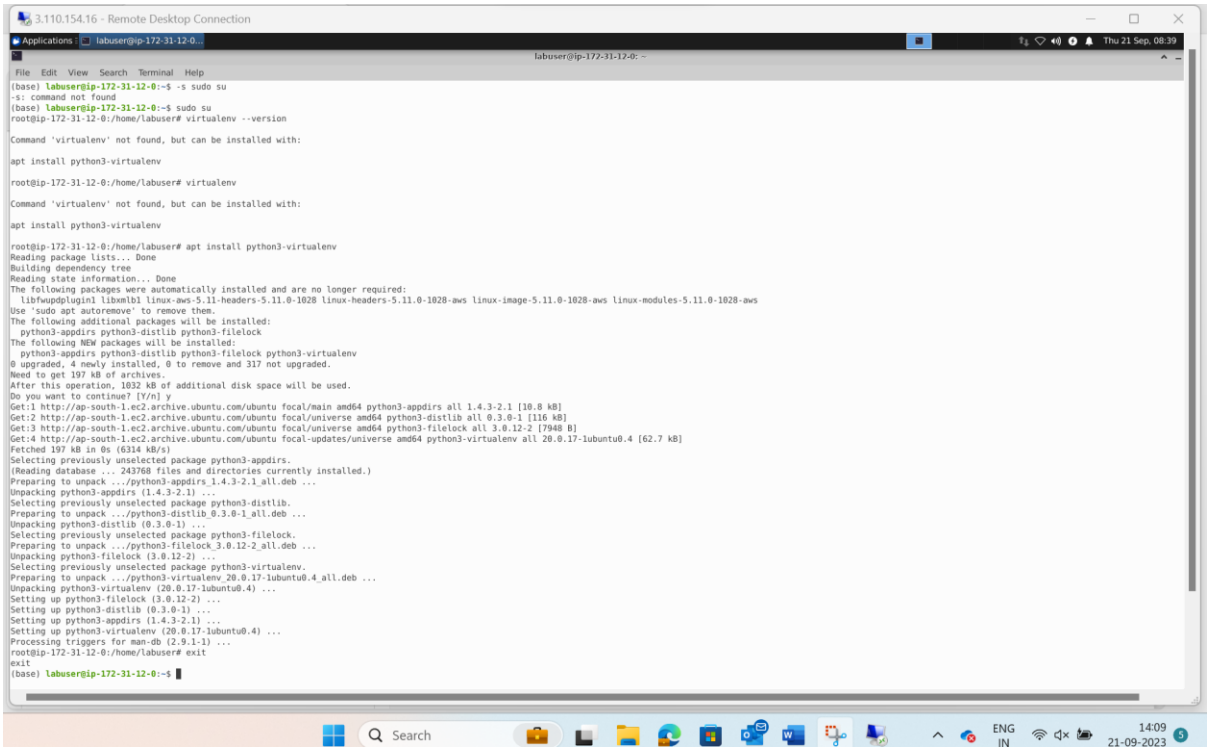
temp way to remove a row in pyspark

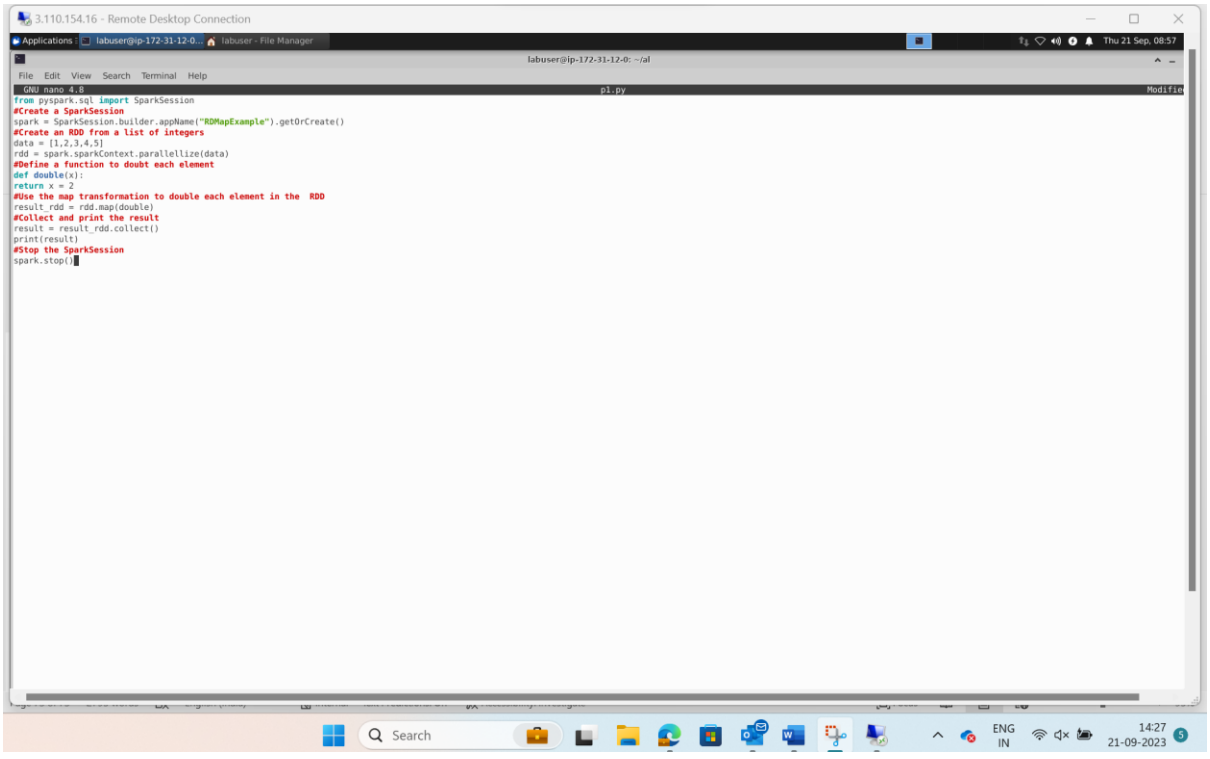


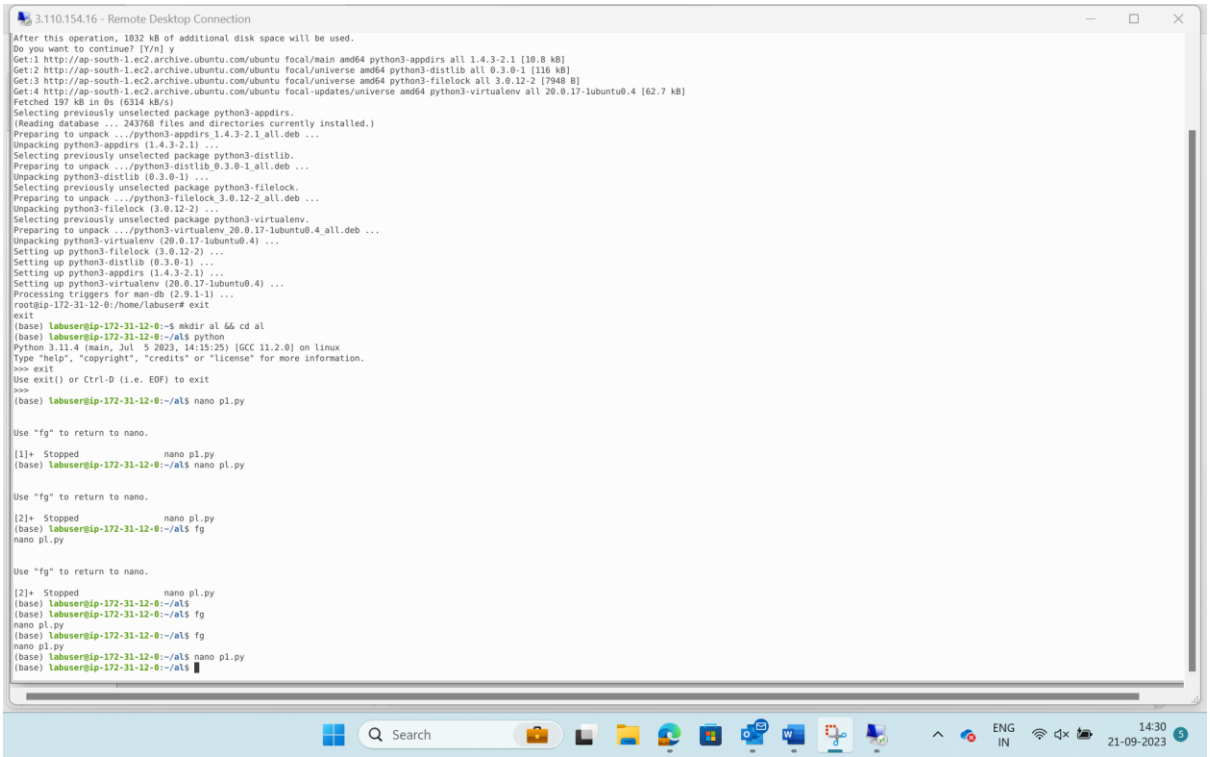




Task 2- convert csv to parquet file

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


3.110.154.16 - Remote Desktop Connection
Applications
labuser@ip-172-31-12-0...
Thu 21 Sep, 08:39
labuser@ip-172-31-12-0: ~
File Edit View Search Terminal Help
(base) labuser@ip-172-31-12-0:~$ -s sudo su
-s: command not found
(base) 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 libxmlb1 linux-aws-5.11-headers-5.11.0-1028 linux-headers-5.11.0-1028-aws 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-appdirs python3-distlib python3-filelock
The following NEW packages will be installed:
python3-appdirs python3-distlib python3-filelock python3-virtualenv
0 upgraded, 4 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] 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 B]
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.
(Reading database ... 243768 files and directories currently installed.)
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-filelock (3.0.12-2) ...
Setting up python3-distlib (0.3.0-1) ...
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:~$
Search
ENG
IN
14:09
21-09-2023


3.110.154.16 - Remote Desktop Connection
Applications
labuser@ip-172-31-12-0...
labuser - File Manager
Thu 21 Sep, 08:57
labuser@ip-172-31-12-0: ~/al
File Edit View Search Terminal Help
GNU nano 4.8
p1.py
Modifie
from pyspark.sql import SparkSession
#Create a SparkSession
spark = SparkSession.builder.appName("RDMapExample").getOrCreate()
#Create an RDD from a list of integers
data = [1,2,3,4,5]
rdd = spark.sparkContext.parallelize(data)
#Define a function to doubt each element
def double(x):
return x = 2
#Use the map transformation to double each element in the RDD
result_rdd = rdd.map(double)
#Collect and print the result
result = result_rdd.collect()
print(result)
#Stop the SparkSession
spark.stop()
Search
ENG
IN
14:27
21-09-2023




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

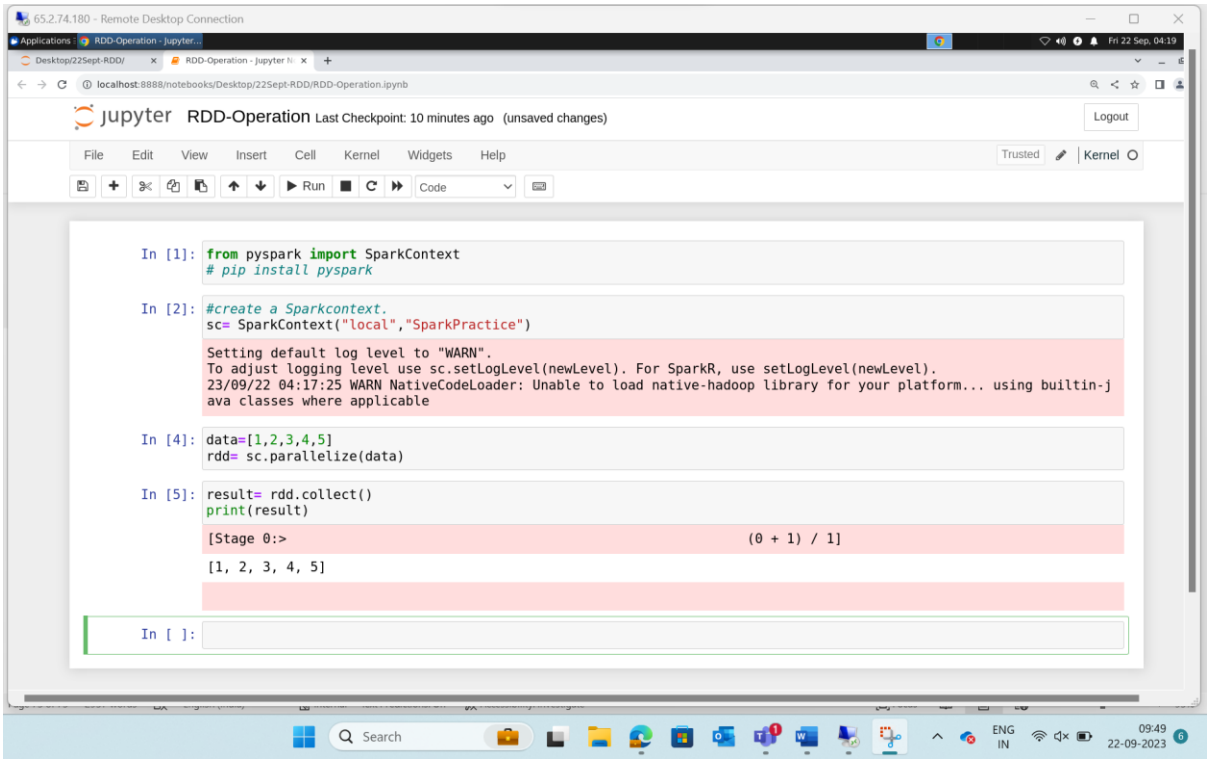
**Day 04(week 4) – 22nd September 2023**

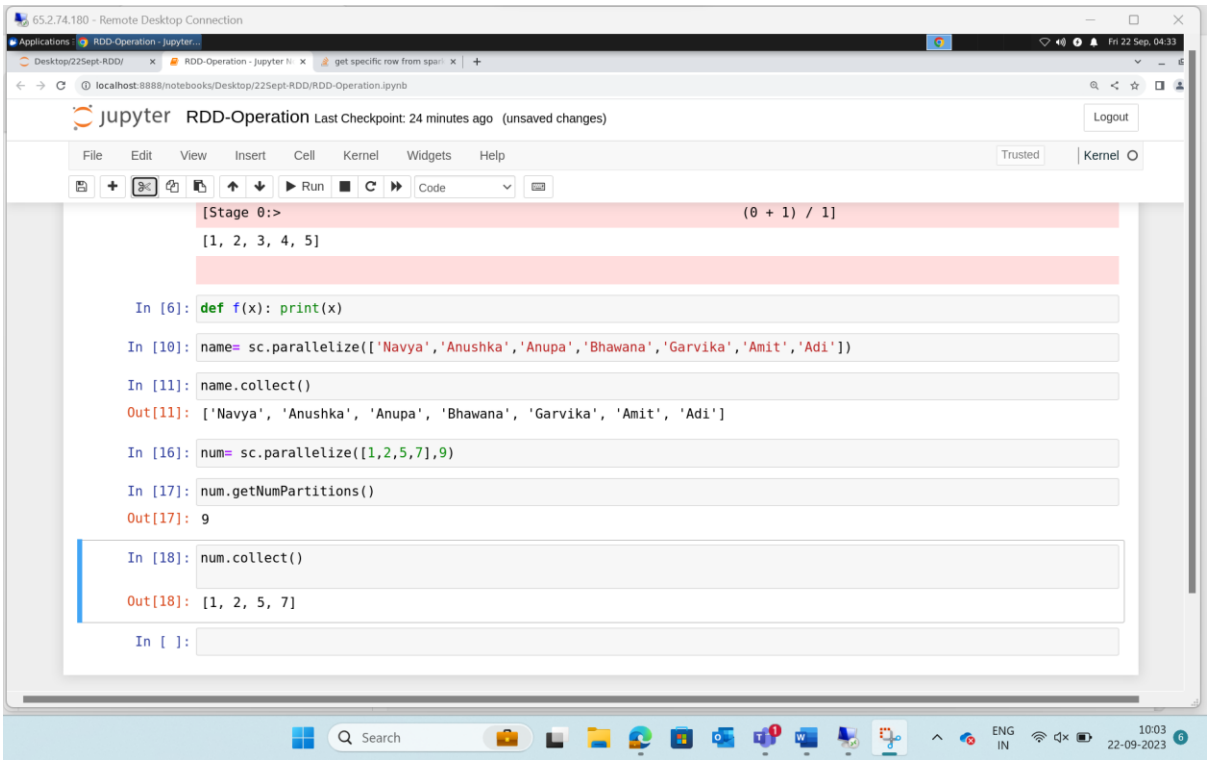
Spark Context is the main entry point and consists of all basic functionality of spark.

- ➔ 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.

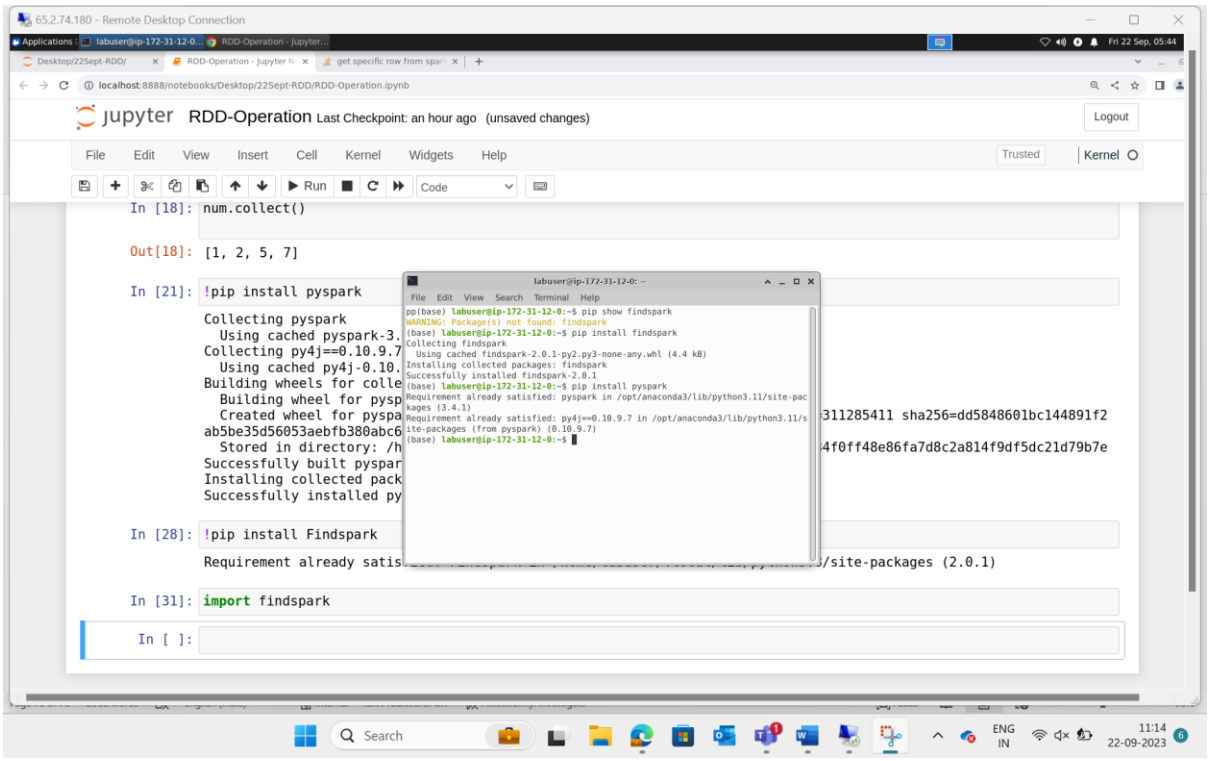
Create a new file in a new folder named rdd and create a spark context



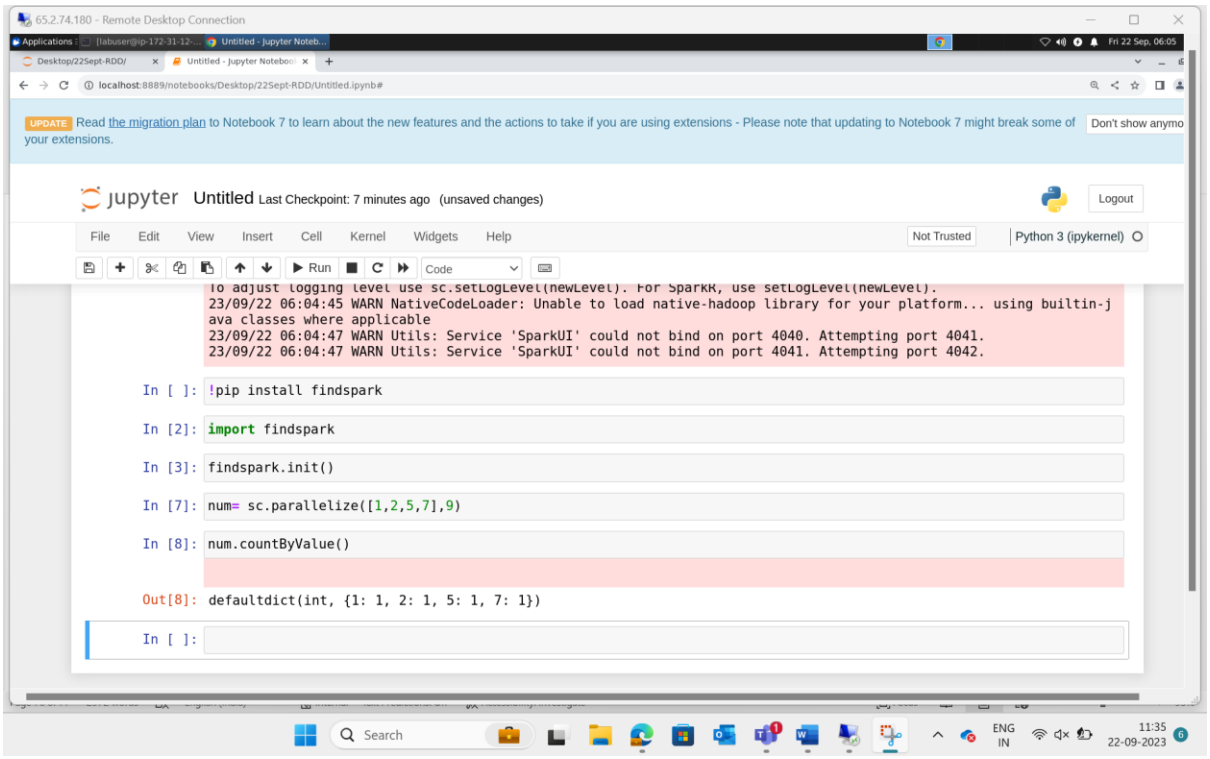




To get partition use getNumPartition

In case findspark doesn't work in terminal of vm type



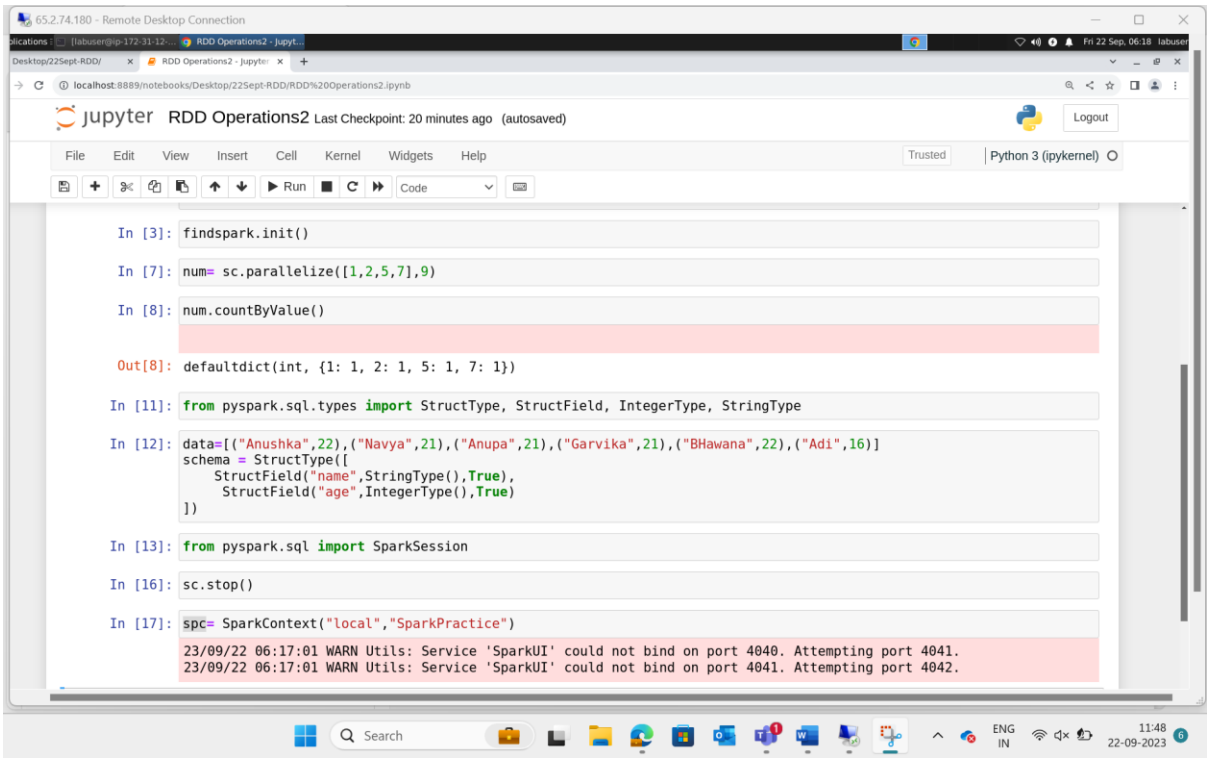

```
Out[18]: [1, 2, 5, 7]

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 [ ]:
```

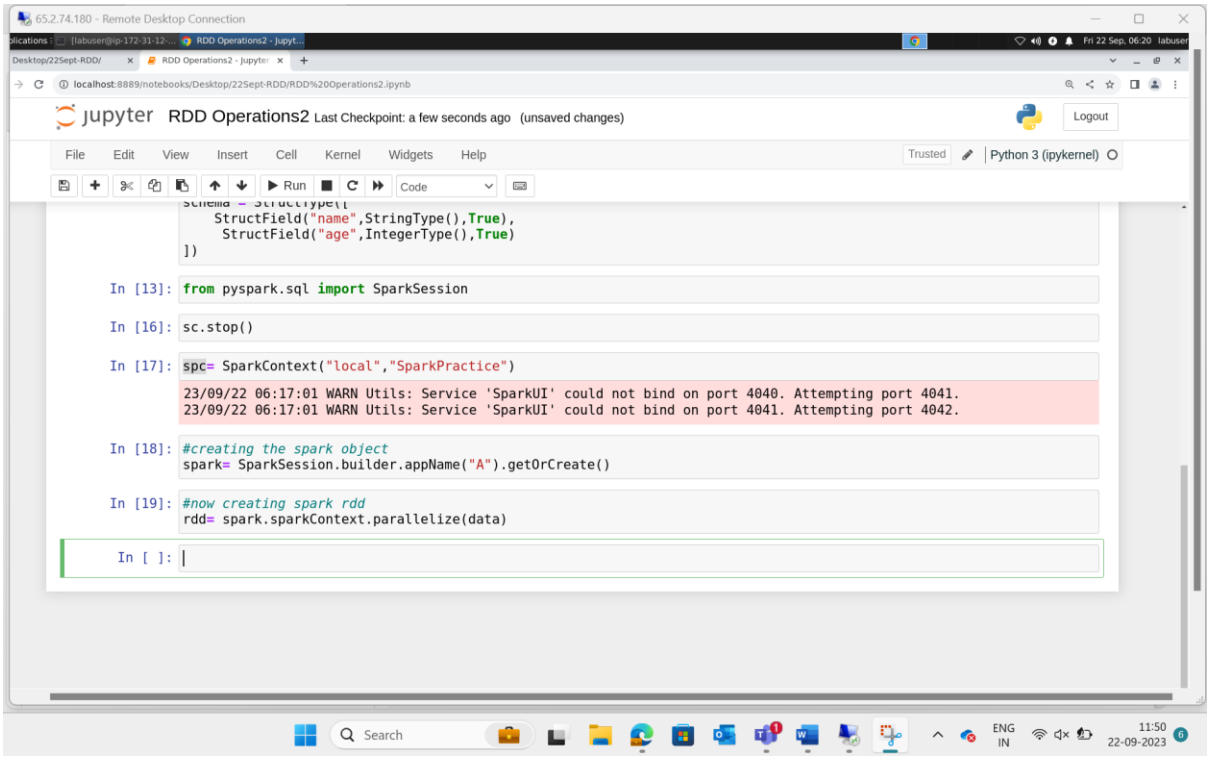



After find spark in installed in main terminal make sure to clear all previous outputs and after importing findspark in program, type findspark.init()

Then the code will work.

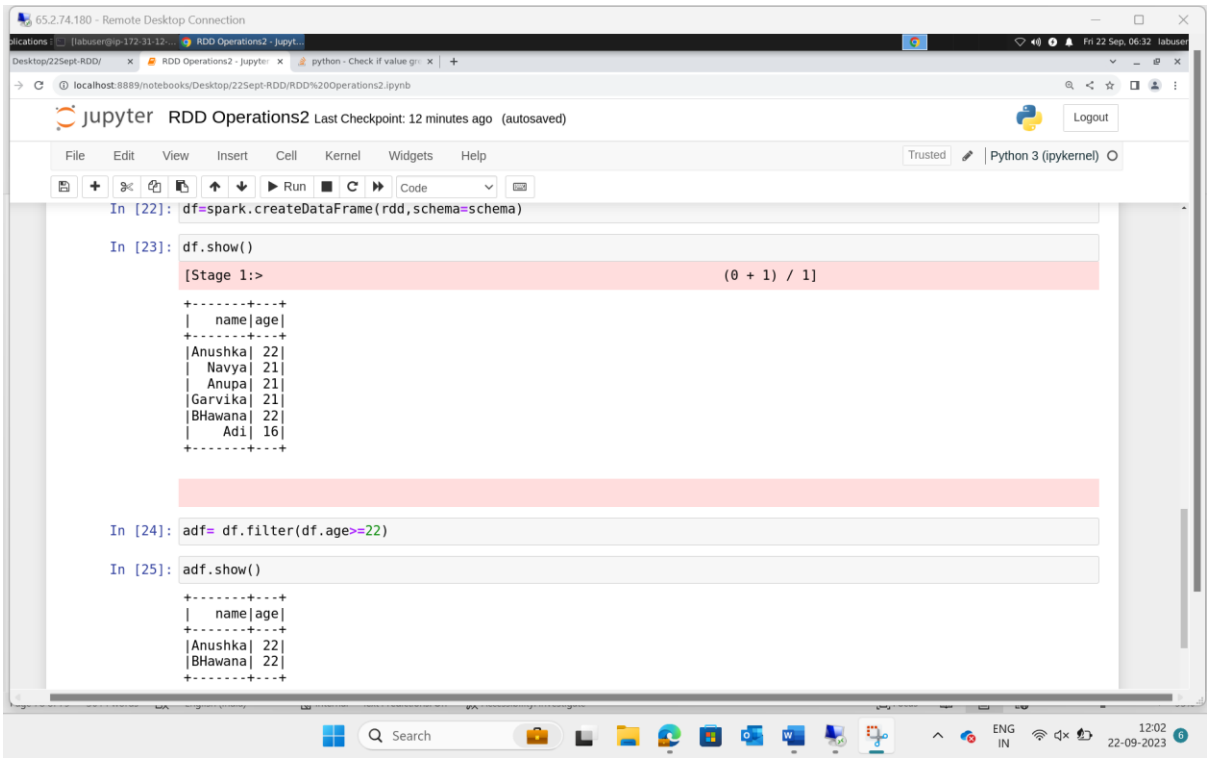
After creating schema we will now create spark RDD




Difference between dataframe and RDD

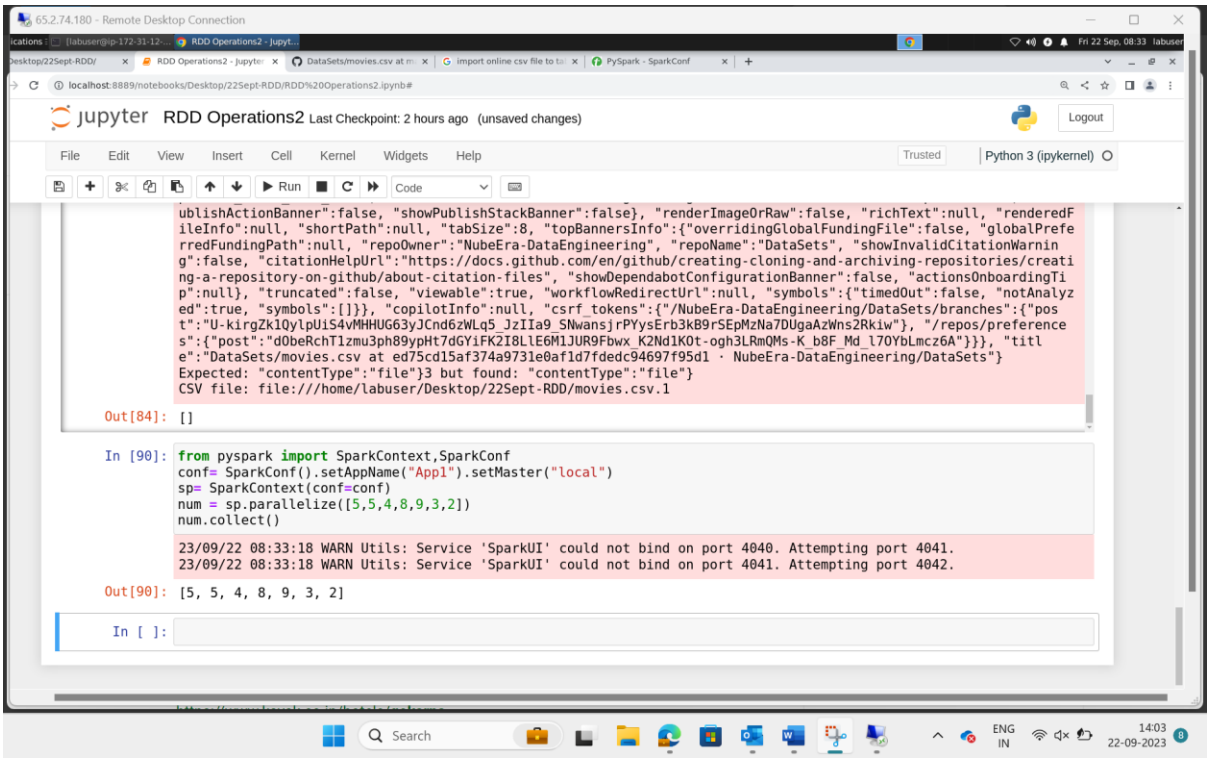
```
rdd= spark.sparkContext.parallelize(data)

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 [ ]:
```



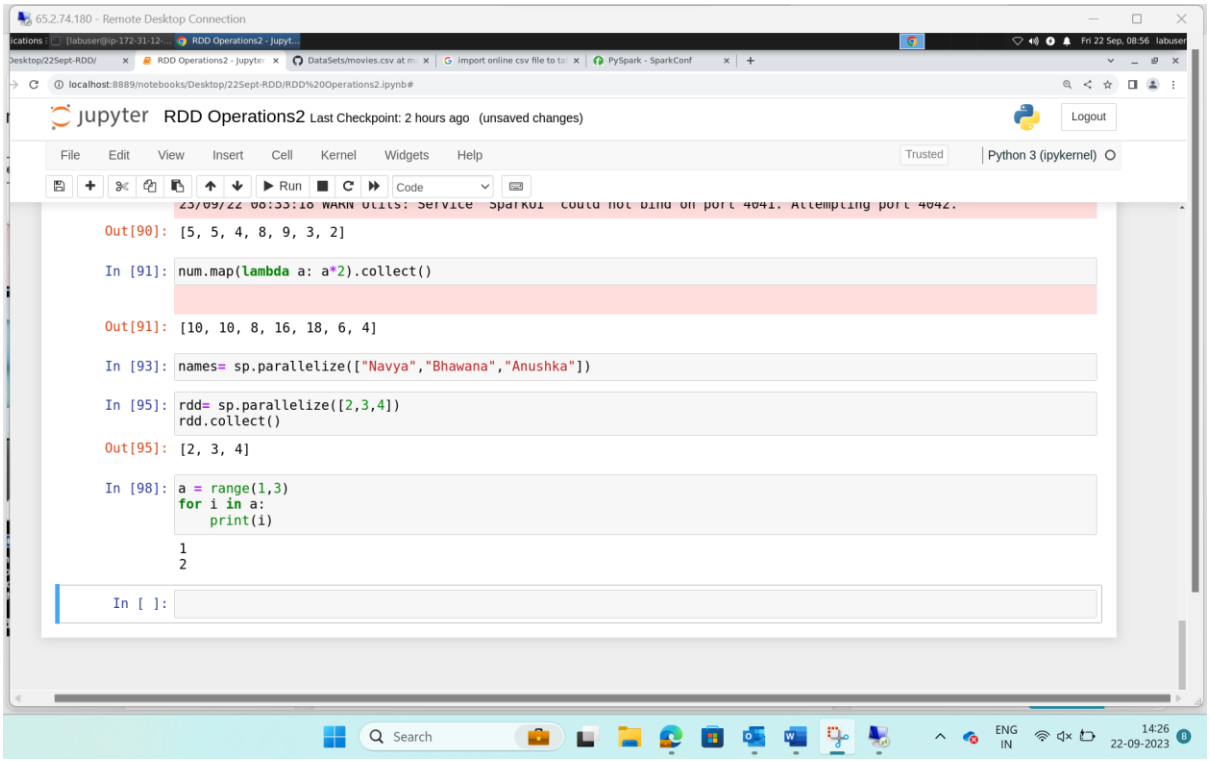

After that can write spark.stop()

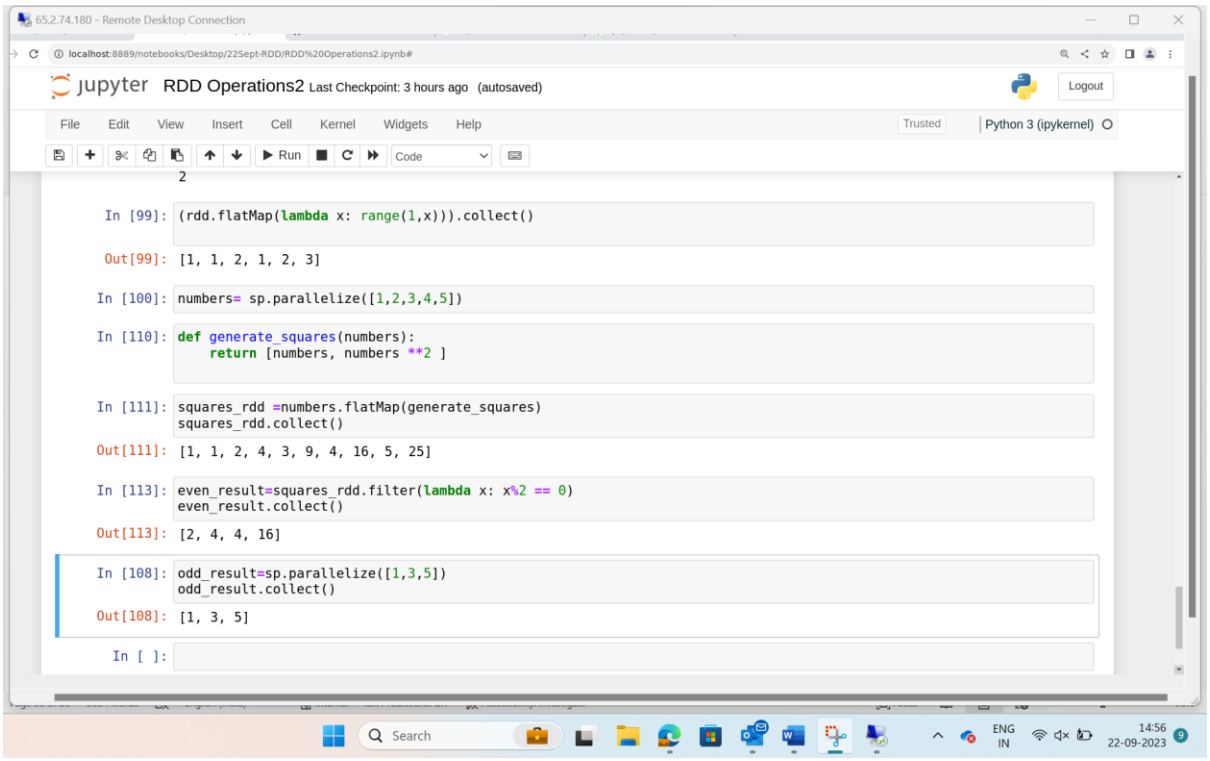
Task:

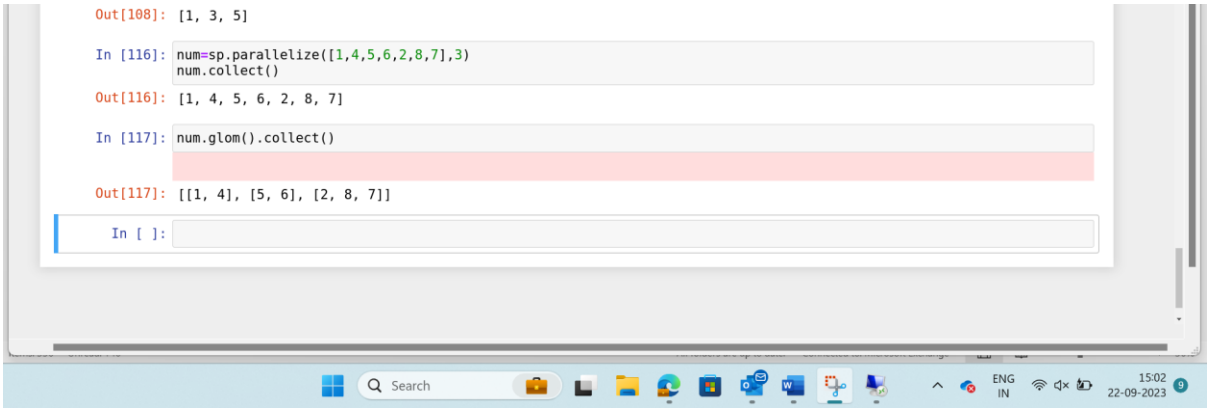
SparkSession-> RDD-> DF->TempView(SQL)->Select



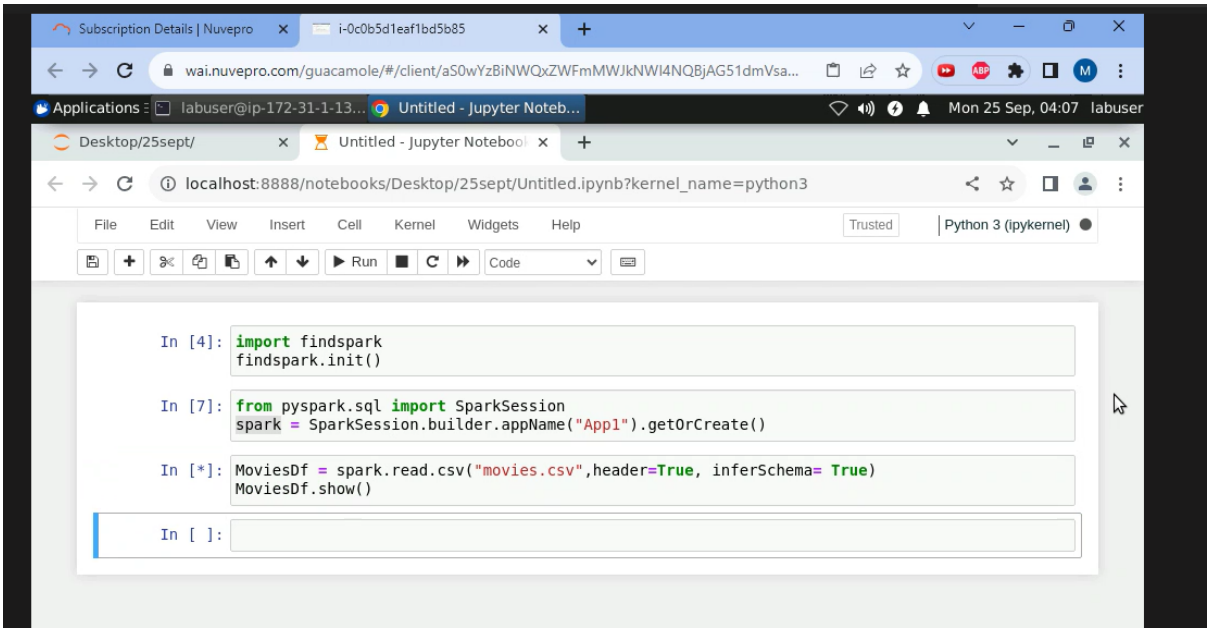

We need to stop spc spark before executingthis

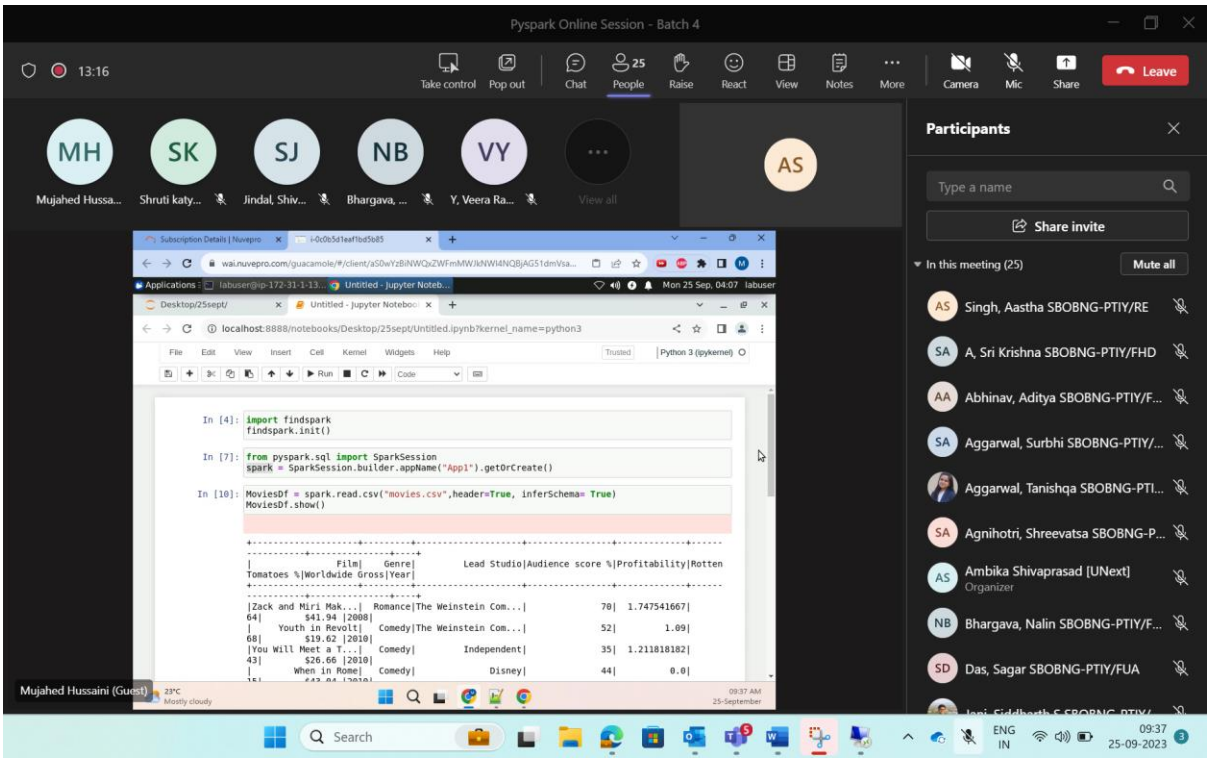


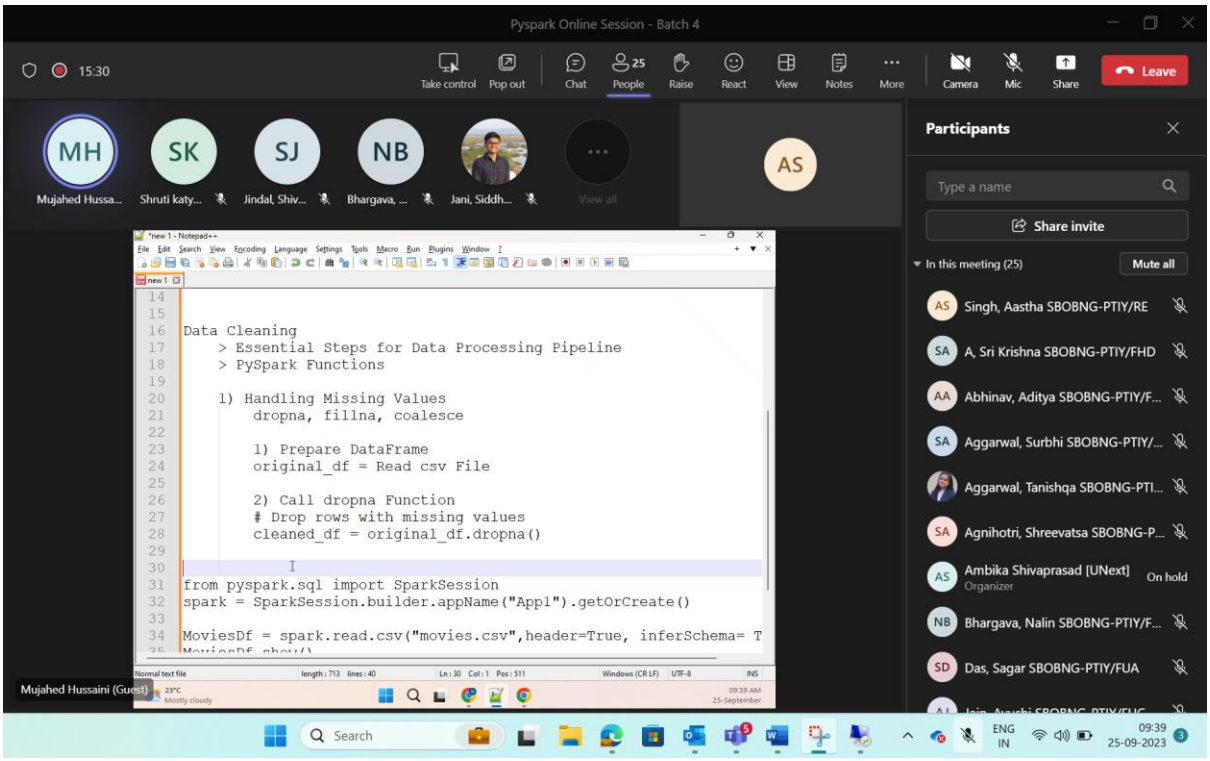


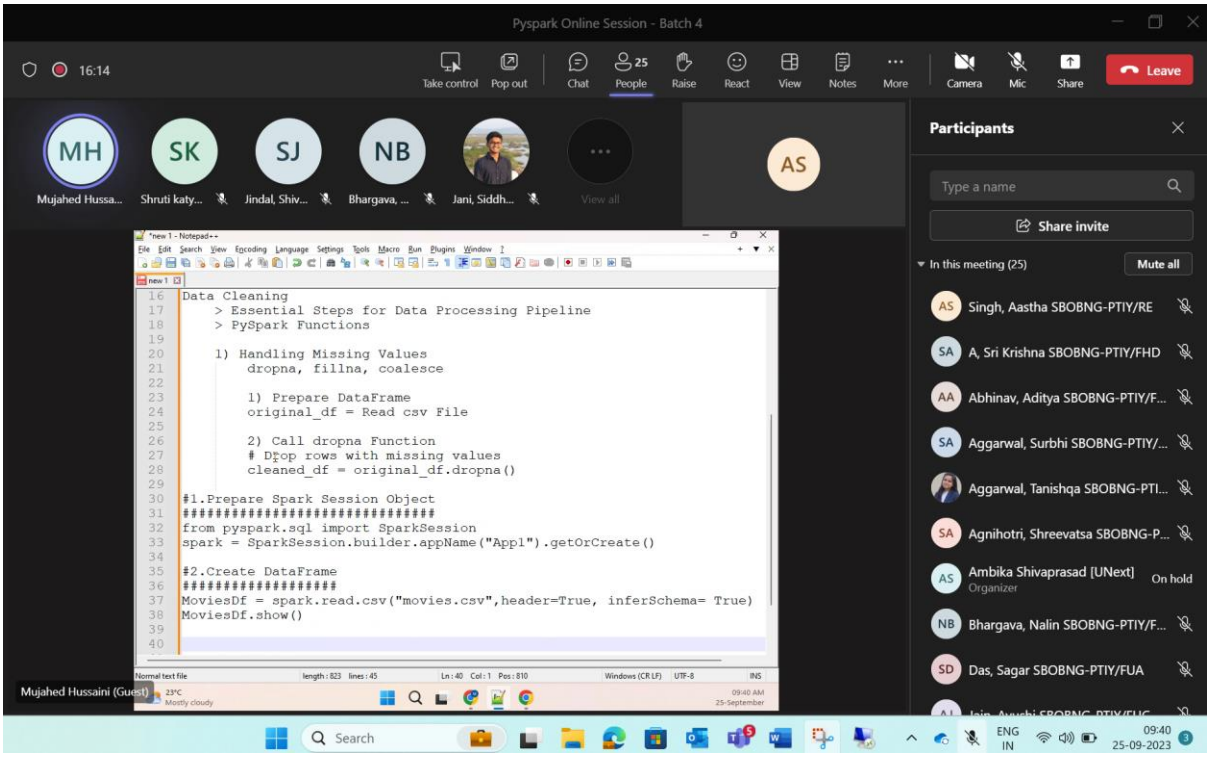



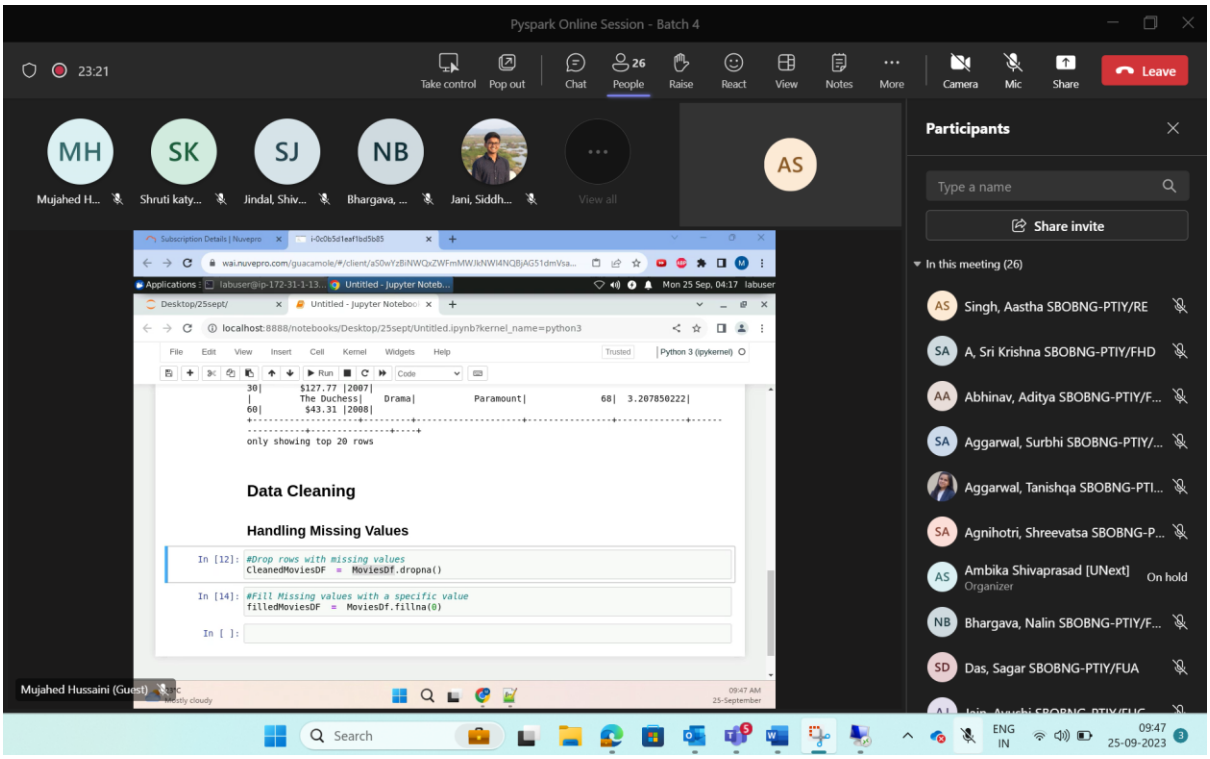

**Day 01(week 5) – 25th September 2023**

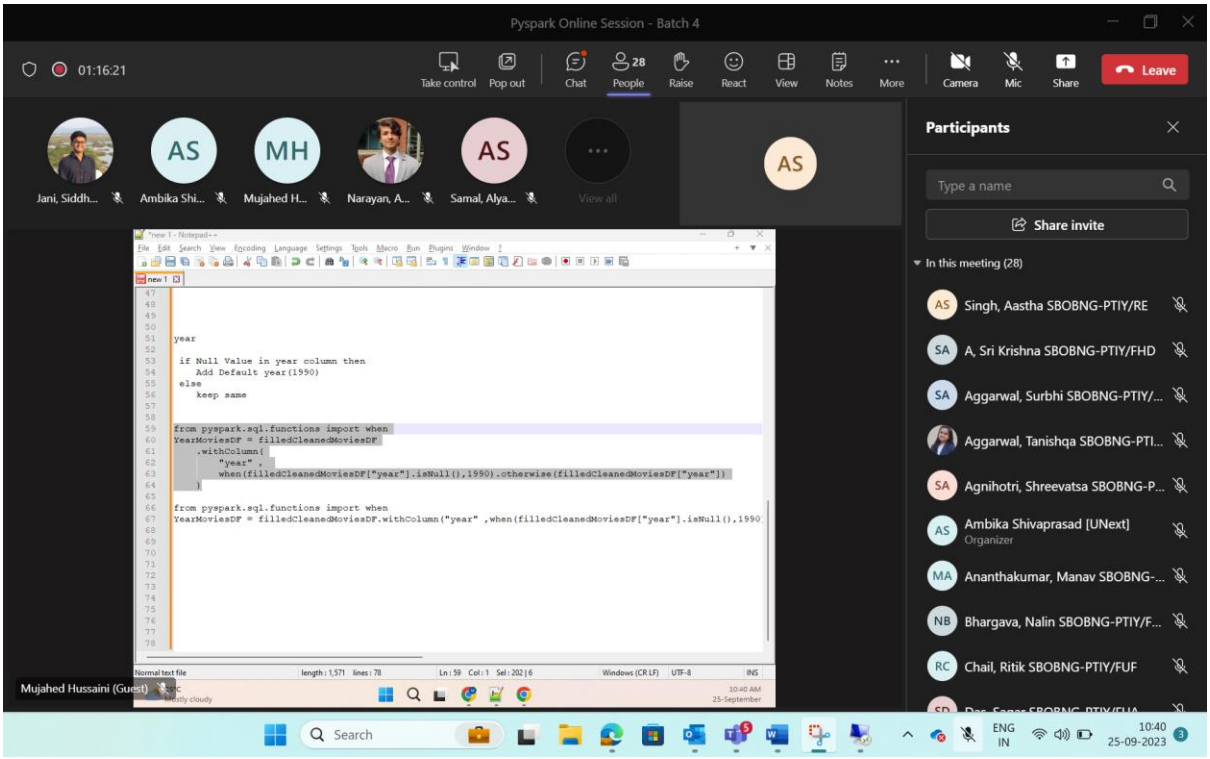

Subscription Details | Nuvepro
i-0c0b5d1eaf1bd5b85
wai.nuvepro.com/guacamole/#/client/aS0wYzBiNWQxZWFmMWJkNWI4NQBjAG51dmVsa...
Applications
labuser@ip-172-31-1-13...
Untitled - Jupyter Noteb...
Mon 25 Sep, 04:07 labuser
Desktop/25sept/
Untitled - Jupyter Noteboo
localhost:8888/notebooks/Desktop/25sept/Untitled.ipynb?kernel_name=python3
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
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 [ ]:

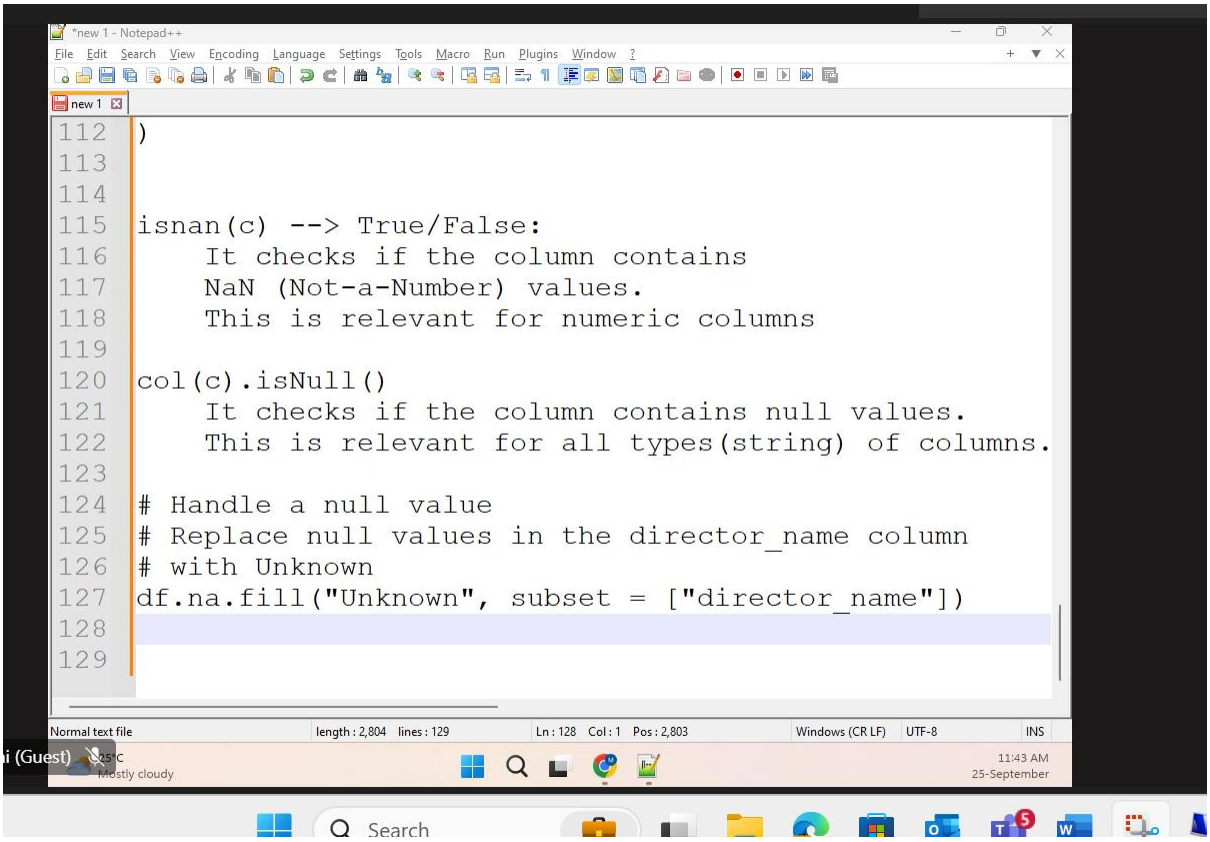

Pyspark Online Session - Batch 4
13:16
Take control
Pop out
Chat
People
Raise
React
View
Notes
More
Camera
Mic
Share
Leave
MH
SK
SJ
NB
VY
AS
Mujahed Hussa...
Shruti katy...
Jindal, Shiv...
Bhargava, ...
Y, Veera Ra...
View all
Participants
Type a name
Share invite
In this meeting (25)
Mute all
Singh, Aastha SBOBNG-PTIY/RE
A, Sri Krishna SBOBNG-PTIY/FHD
Abhinav, Aditya SBOBNG-PTIY/F...
Aggarwal, Surbhi SBOBNG-PTIY/...
Aggarwal, Tanishqa SBOBNG-PTI...
Agnihotri, Shreevatsa SBOBNG-P...
Ambika Shivaprasad [UNext]
Organizer
Bhargava, Nalin SBOBNG-PTIY/F...
Das, Sagar SBOBNG-PTIY/FUA
Mujahed Hussaini (Guest)
Search
ENG
IN
09:37
25-09-2023

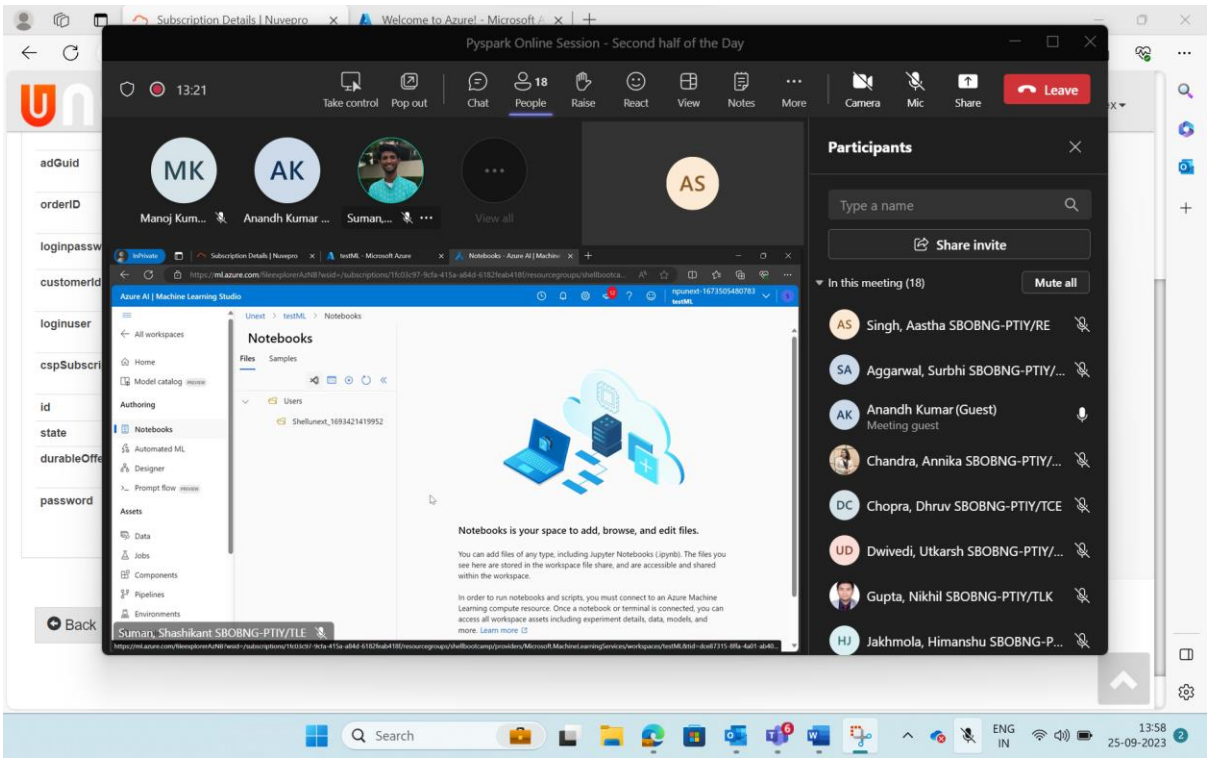

Pyspark Online Session - Batch 4
15:30
Take control
Pop out
Chat
People
Raise
React
View
Notes
More
Camera
Mic
Share
Leave
Participants
Type a name
Share invite
In this meeting (25)
Mute all
Singh, Aastha SBOBNG-PTIY/RE
A, Sri Krishna SBOBNG-PTIY/FHD
Abhinav, Aditya SBOBNG-PTIY/F...
Aggarwal, Surbhi SBOBNG-PTIY/...
Aggarwal, Tanishqa SBOBNG-PTI...
Agnihotri, Shreevatsa SBOBNG-P...
Ambika Shivaprasad [UNext]
Organizer
On hold
Bhargava, Nalin SBOBNG-PTIY/F...
Das, Sagar SBOBNG-PTIY/FUA
Mujahed Hussaini (Guest)
Search
ENG
IN
09:39
25-09-2023
Data Cleaning
> Essential Steps for Data Processing Pipeline
> PySpark Functions
1) Handling Missing Values
dropna, fillna, coalesce
1) Prepare DataFrame
original_df = Read csv File
2) Call dropna Function
# Drop rows with missing values
cleaned_df = original_df.dropna()
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("App1").getOrCreate()
MoviesDf = spark.read.csv("movies.csv",header=True, inferSchema= T

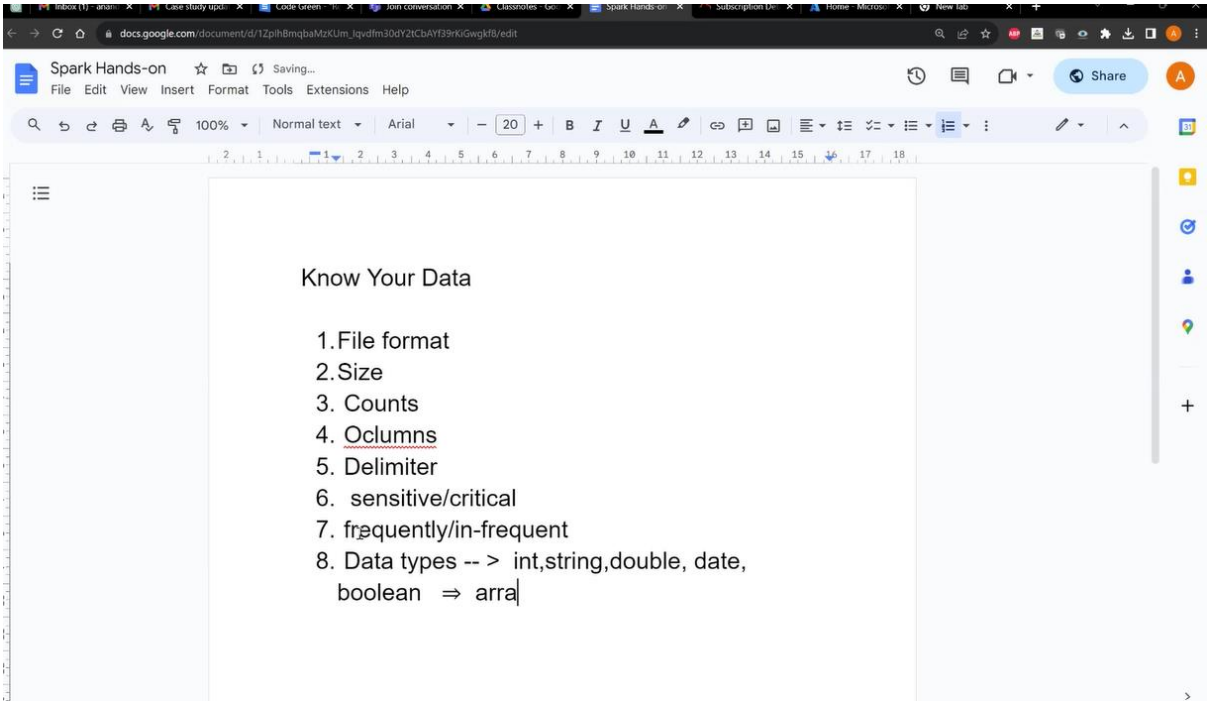

Pyspark Online Session - Batch 4
16:14
Take control
Pop out
Chat
People
Raise
React
View
Notes
More
Camera
Mic
Share
Leave
Participants
Type a name
Share invite
In this meeting (25)
Mute all
Singh, Aastha SBOBNG-PTIY/RE
A, Sri Krishna SBOBNG-PTIY/FHD
Abhinav, Aditya SBOBNG-PTIY/F...
Aggarwal, Surbhi SBOBNG-PTIY/...
Aggarwal, Tanishqa SBOBNG-PTI...
Agnihotri, Shreevatsa SBOBNG-P...
Ambika Shivaprasad [UNext]
Organizer
On hold
Bhargava, Nalin SBOBNG-PTIY/F...
Das, Sagar SBOBNG-PTIY/FUA
Mujahed Hussaini (Guest)
Search
ENG
IN
09:40
25-09-2023
Data Cleaning
> Essential Steps for Data Processing Pipeline
> PySpark Functions
1) Handling Missing Values
dropna, fillna, coalesce
1) Prepare DataFrame
original_df = Read csv File
2) Call dropna Function
# Drop rows with missing values
cleaned_df = original_df.dropna()
#1.Prepare Spark Session Object
################################
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("App1").getOrCreate()
#2.Create DataFrame
####################
MoviesDf = spark.read.csv("movies.csv",header=True, inferSchema= True)
MoviesDf.show()

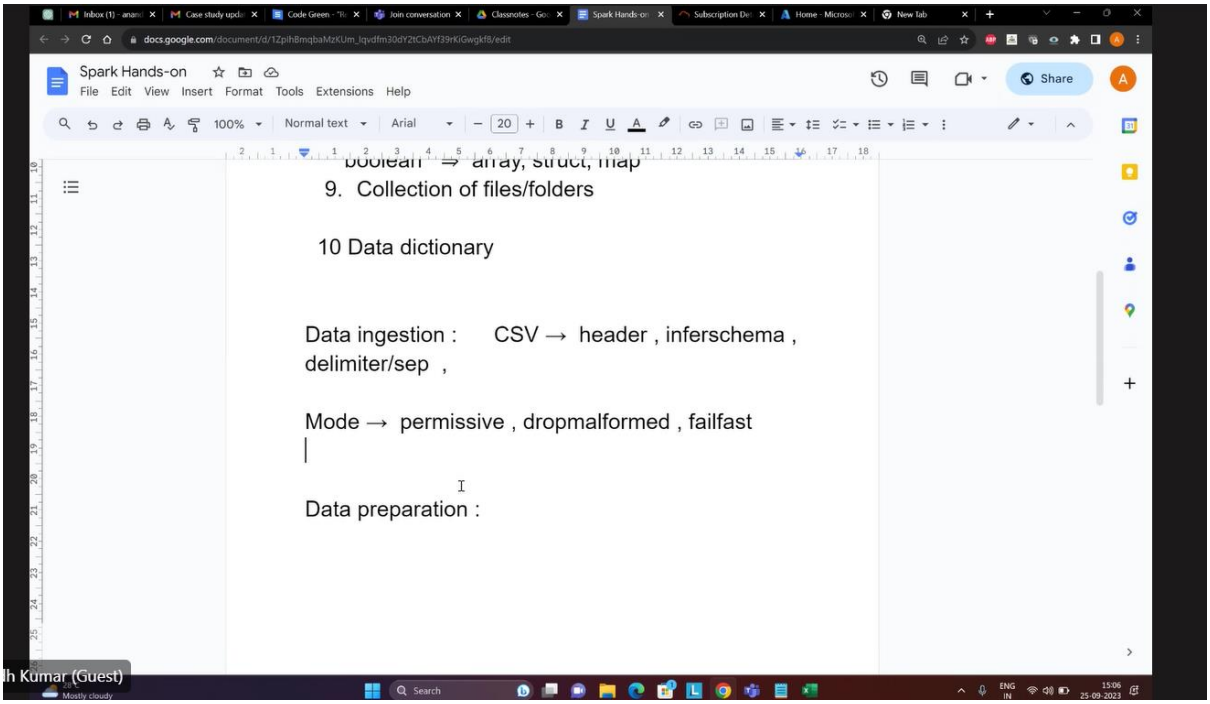

Pyspark Online Session - Batch 4
23:21
Take control
Pop out
Chat
People
Raise
React
View
Notes
More
Camera
Mic
Share
Leave
MH
SK
SJ
NB
AS
Mujahed H...
Shruti katy...
Jindal, Shiv...
Bhargava, ...
Jani, Siddh...
View all
Participants
Type a name
Share invite
In this meeting (26)
Singh, Aastha SBOBNG-PTIY/RE
A, Sri Krishna SBOBNG-PTIY/FHD
Abhinav, Aditya SBOBNG-PTIY/F...
Aggarwal, Surbhi SBOBNG-PTIY/...
Aggarwal, Tanishqa SBOBNG-PTI...
Agnihotri, Shreevatsa SBOBNG-P...
Ambika Shivaprasad [UNext]
Organizer
On hold
Bhargava, Nalin SBOBNG-PTIY/F...
Das, Sagar SBOBNG-PTIY/FUA
Mujahed Hussaini (Guest)
Data Cleaning
Handling Missing Values
#Drop rows with missing values
CleanedMoviesDF = MoviesDf.dropna()
#Fill Missing values with a specific value
filledMoviesDF = MoviesDf.fillna(0)
only showing top 20 rows
Search
ENG
IN
09:47
25-09-2023

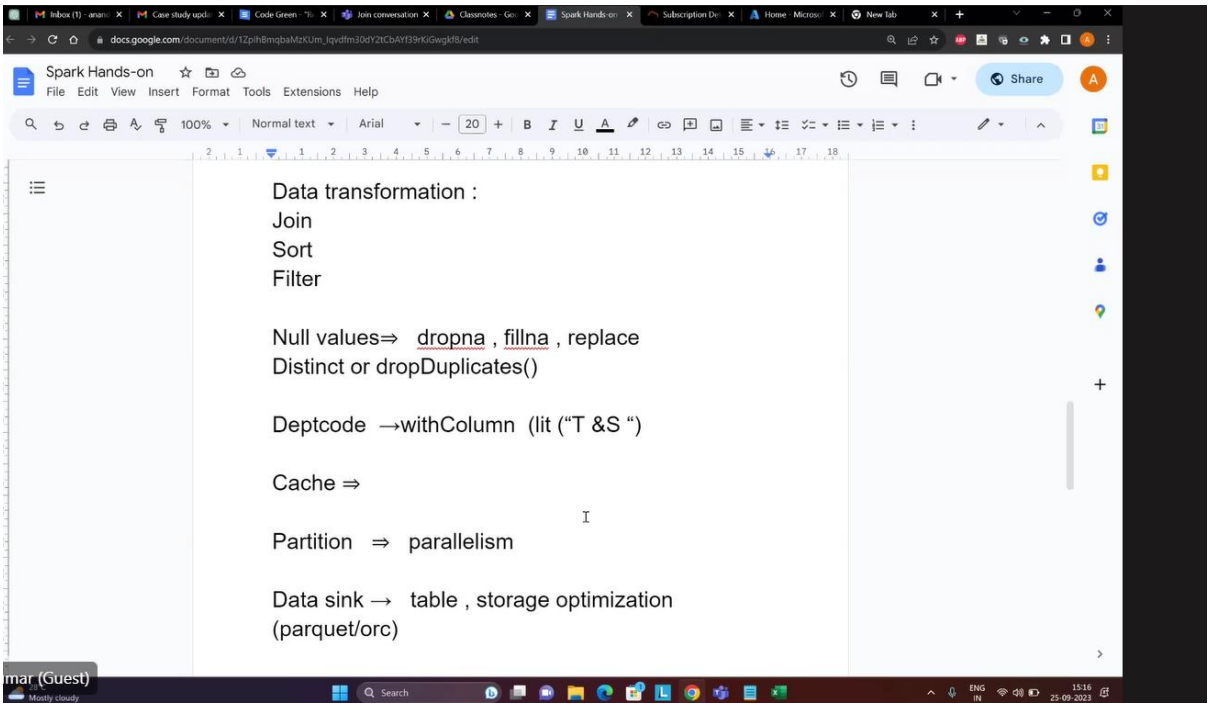

Pyspark Online Session - Batch 4
01:16:21
Take control
Pop out
Chat
People
Raise
React
View
Notes
More
Camera
Mic
Share
Leave
Jani, Siddh...
AS
MH
Narayan, A...
Samal, Alya...
View all
Participants
Type a name
Share invite
In this meeting (28)
Singh, Aastha SBOBNG-PTIY/RE
A, Sri Krishna SBOBNG-PTIY/FHD
Aggarwal, Surbhi SBOBNG-PTIY/...
Aggarwal, Tanishqa SBOBNG-PTI...
Agnihotri, Shreevatsa SBOBNG-P...
Ambika Shivaprasad [UNext]
Organizer
Ananthakumar, Manav SBOBNG-...
Bhargava, Nalin SBOBNG-PTIY/F...
Chail, Ritik SBOBNG-PTIY/FUF
Mujahed Hussaini (Guest)
year
if Null Value in year column then
Add Default year(1990)
else
keep same
from pyspark.sql.functions import when
YearMoviesDF = filledCleanedMoviesDF
.withColumn(
"year" ,
when(filledCleanedMoviesDF["year"].isNull(),1990).otherwise(filledCleanedMoviesDF["year"])
)
from pyspark.sql.functions import when
YearMoviesDF = filledCleanedMoviesDF.withColumn("year" ,when(filledCleanedMoviesDF["year"].isNull(),1990
Search
ENG
IN
10:40
25-09-2023


*new 1 - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
new 1
)
isnan(c) --> True/False:
It checks if the column contains
NaN (Not-a-Number) values.
This is relevant for numeric columns
col(c).isNull()
It checks if the column contains null values.
This is relevant for all types(string) of columns.
# Handle a null value
# Replace null values in the director_name column
# with Unknown
df.na.fill("Unknown", subset = ["director_name"])
Normal text file
length : 2,804 lines : 129
Ln : 128 Col : 1 Pos : 2,803
Windows (CR LF)
UTF-8
INS
Mostly cloudy
11:43 AM
25-September
Search


Pyspark Online Session - Second half of the Day
Participants
Type a name
Share invite
In this meeting (18)
Mute all
Notebooks
Notebooks is your space to add, browse, and edit files.
Suman, Shashikant SBOBNG-PTIY/TLE


Spark Hands-on
Know Your Data
1.File format
2.Size
3. Counts
4. Oclumns
5. Delimiter
6. sensitive/critical
7. frequently/in-frequent
8. Data types -- > int,string,double, date,
boolean ⇒ arra




## 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**

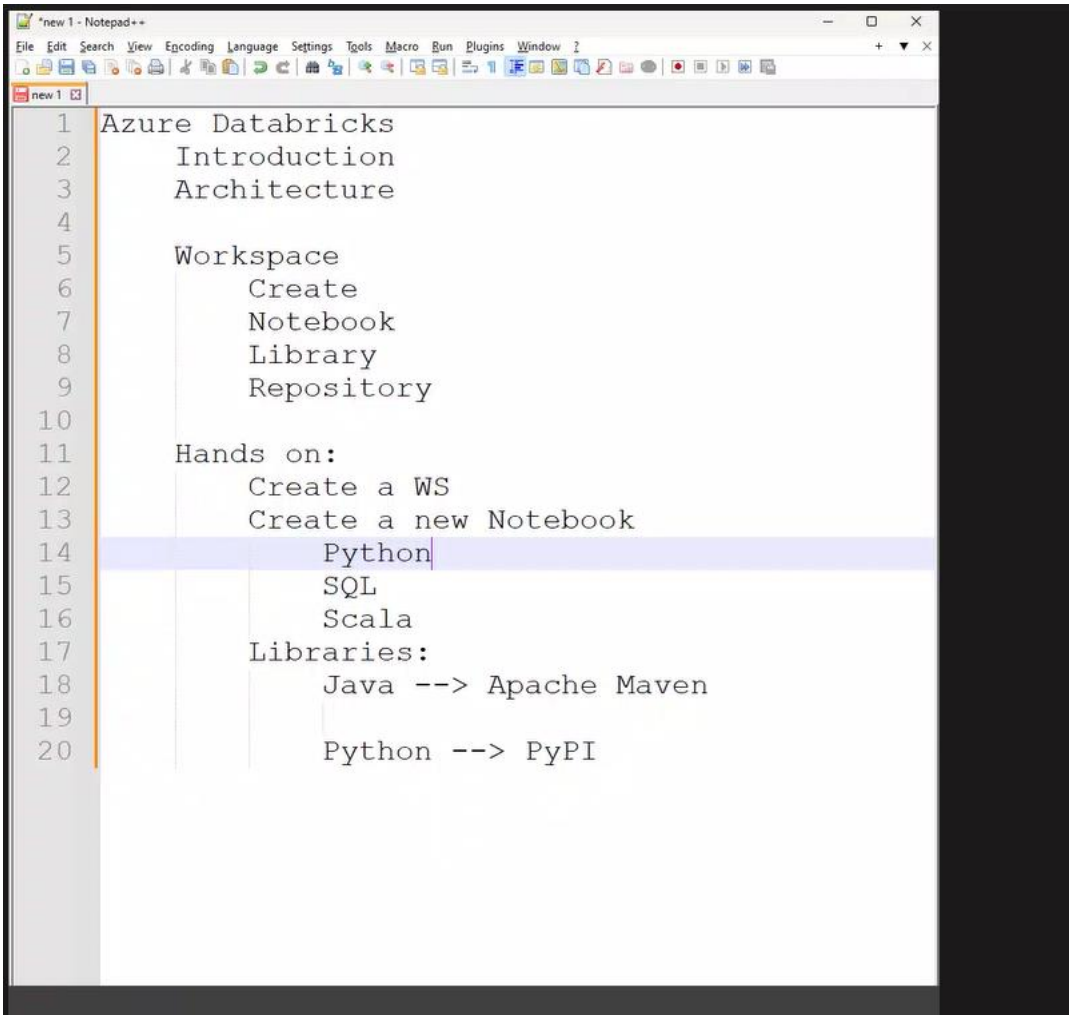
**Writer ⇒ ignore , append and overwrite**

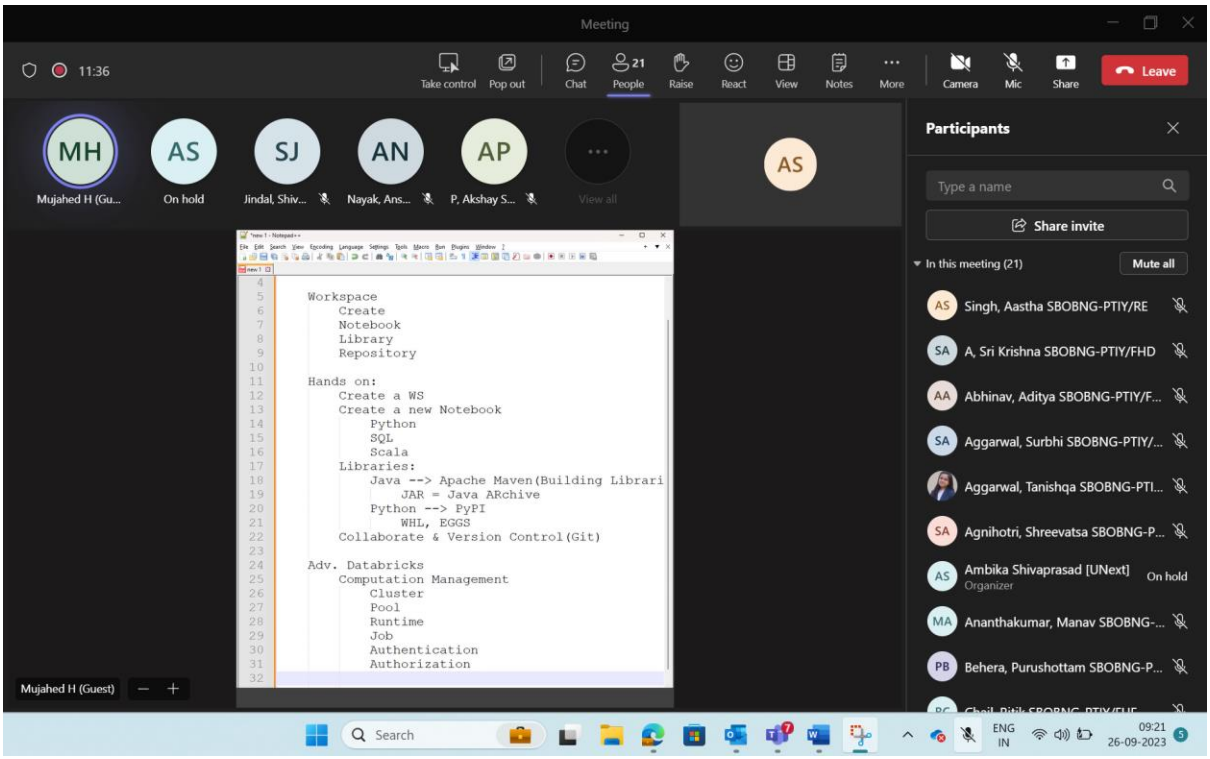
notepad.pw / shellidasep23 | The napkin of the internet.

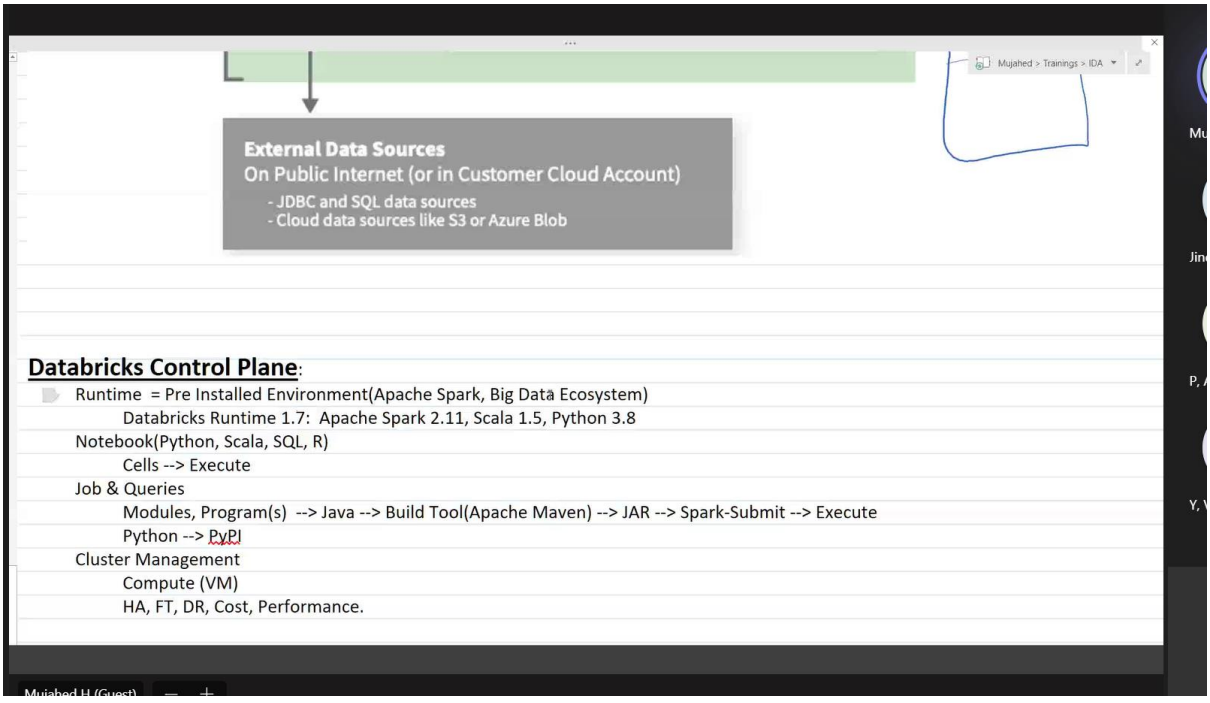
notepad.pw / shellidasep23 | The napkin of the internet.

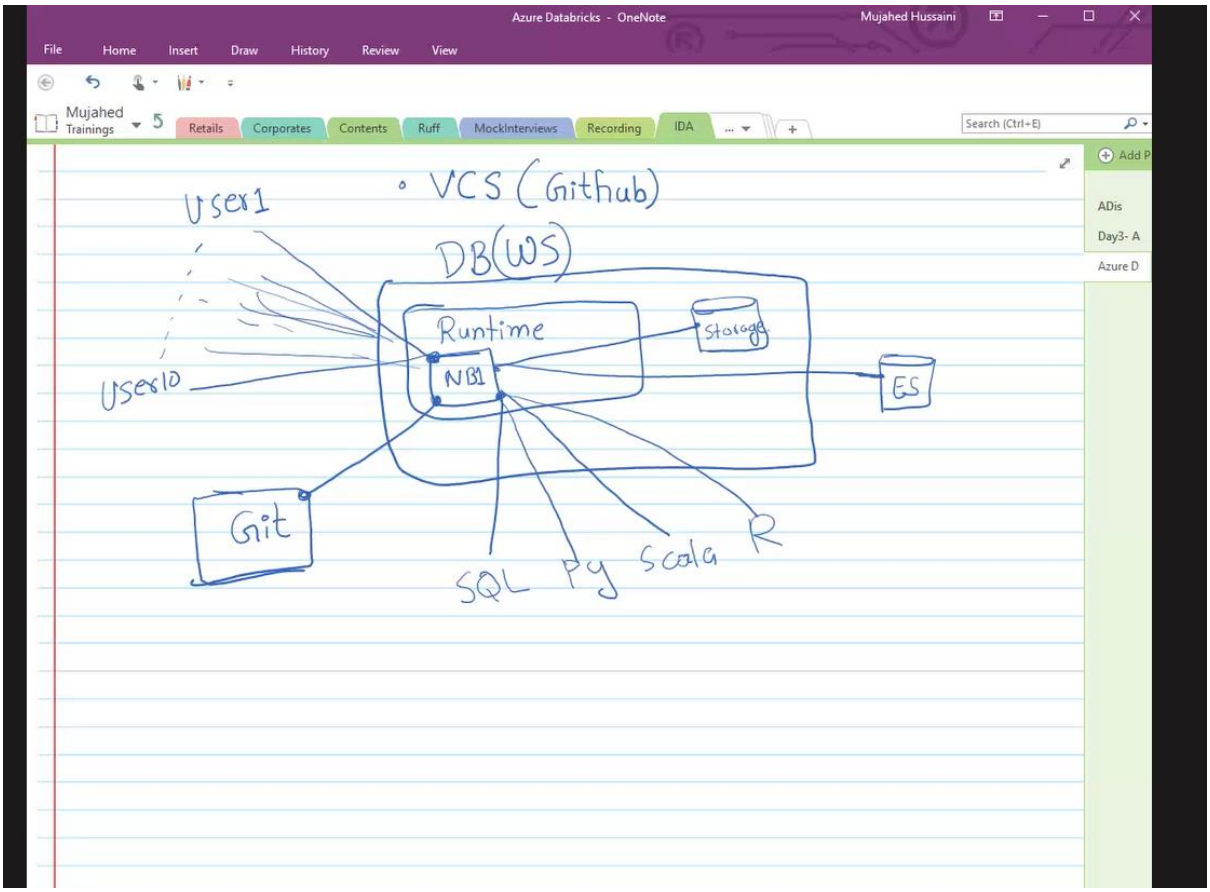
GitHub - akgeoinsys/retail

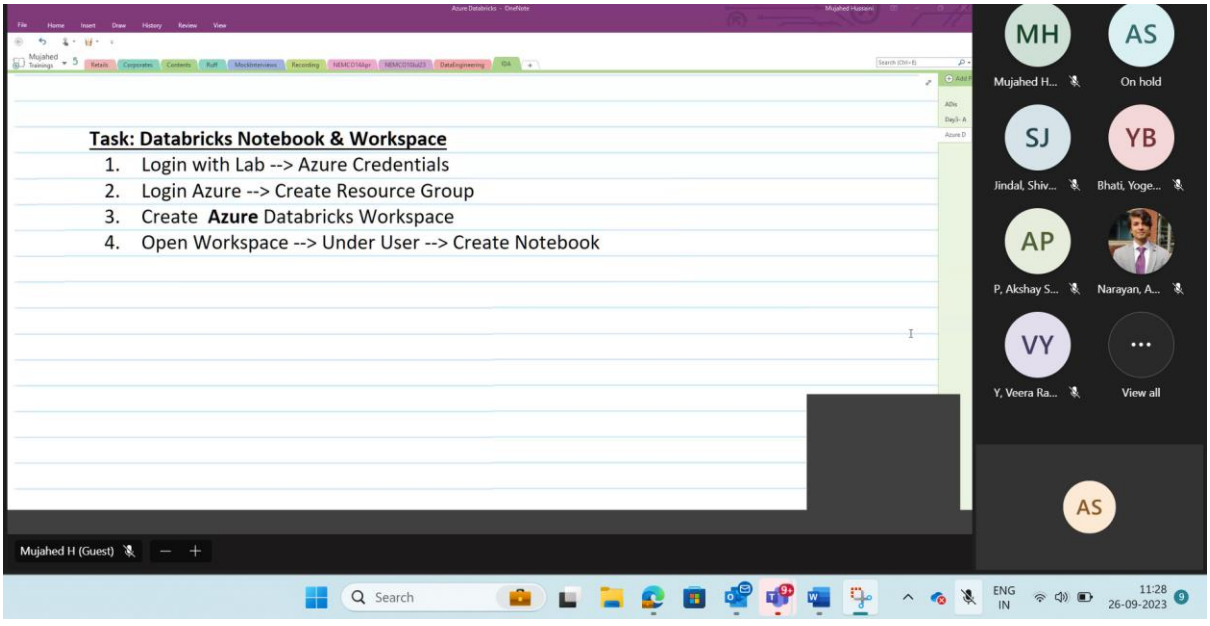
**Day 02(week 5) – 26th September 2023**

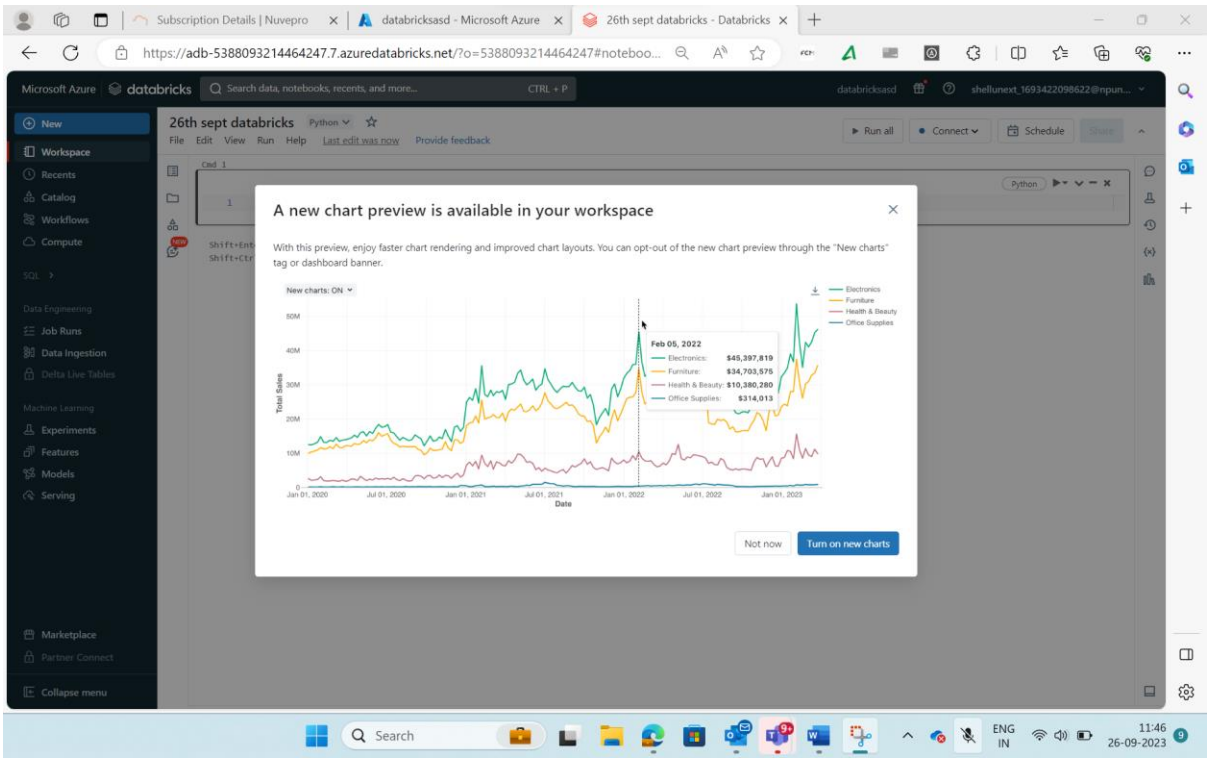

*new 1 - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
new 1
Azure Databricks
Introduction
Architecture
Workspace
Create
Notebook
Library
Repository
Hands on:
Create a WS
Create a new Notebook
Python
SQL
Scala
Libraries:
Java --> Apache Maven
Python --> PyPI

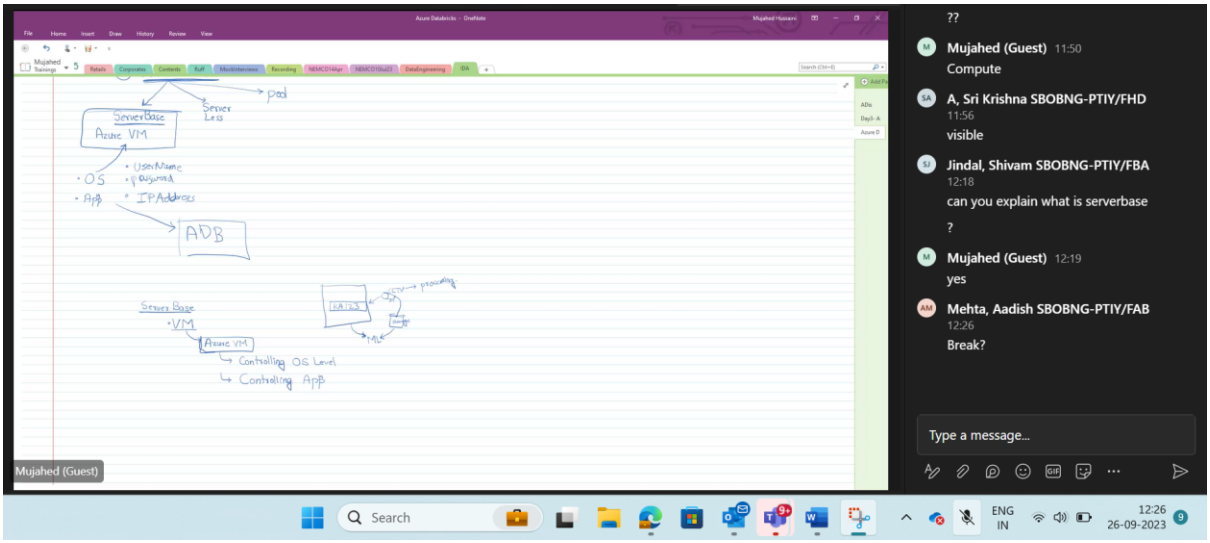

Meeting
11:36
Take control
Pop out
Chat
People
Raise
React
View
Notes
More
Camera
Mic
Share
Leave
MH
Mujahed H (Gu...
AS
On hold
SJ
Jindal, Shiv...
AN
Nayak, Ans...
AP
P. Akshay S...
View all
AS
Participants
Type a name
Share invite
In this meeting (21)
Mute all
Singh, Aastha SBOBNG-PTIY/RE
A, Sri Krishna SBOBNG-PTIY/FHD
Abhinav, Aditya SBOBNG-PTIY/F...
Aggarwal, Surbhi SBOBNG-PTIY/...
Aggarwal, Tanishqa SBOBNG-PTI...
Agnihotri, Shreevatsa SBOBNG-P...
Ambika Shivaprasad [UNext]
Organizer
On hold
Ananthakumar, Manav SBOBNG-...
Behera, Purushottam SBOBNG-P...
Workspace
Create
Notebook
Library
Repository
Hands on:
Create a WS
Create a new Notebook
Python
SQL
Scala
Libraries:
Java --> Apache Maven(Building Librari
JAR = Java ARchive
Python --> PyPI
WHL, EGGS
Collaborate & Version Control(Git)
Adv. Databricks
Computation Management
Cluster
Pool
Runtime
Job
Authentication
Authorization
Mujahed H (Guest)
Search
ENG
IN
09:21
26-09-2023

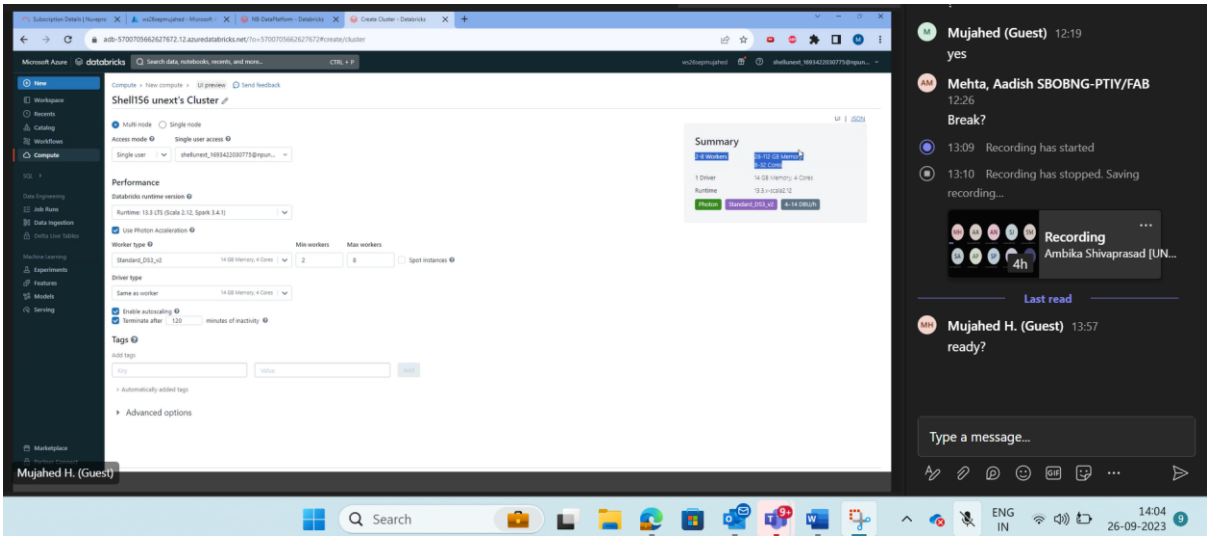

Mujahed > Trainings > IDA
External Data Sources
On Public Internet (or in Customer Cloud Account)
- JDBC and SQL data sources
- Cloud data sources like S3 or Azure Blob
Databricks Control Plane:
Runtime = Pre Installed Environment(Apache Spark, Big Data Ecosystem)
Databricks Runtime 1.7: Apache Spark 2.11, Scala 1.5, Python 3.8
Notebook(Python, Scala, SQL, R)
Cells --> Execute
Job & Queries
Modules, Program(s) --> Java --> Build Tool(Apache Maven) --> JAR --> Spark-Submit --> Execute
Python --> PyPI
Cluster Management
Compute (VM)
HA, FT, DR, Cost, Performance.

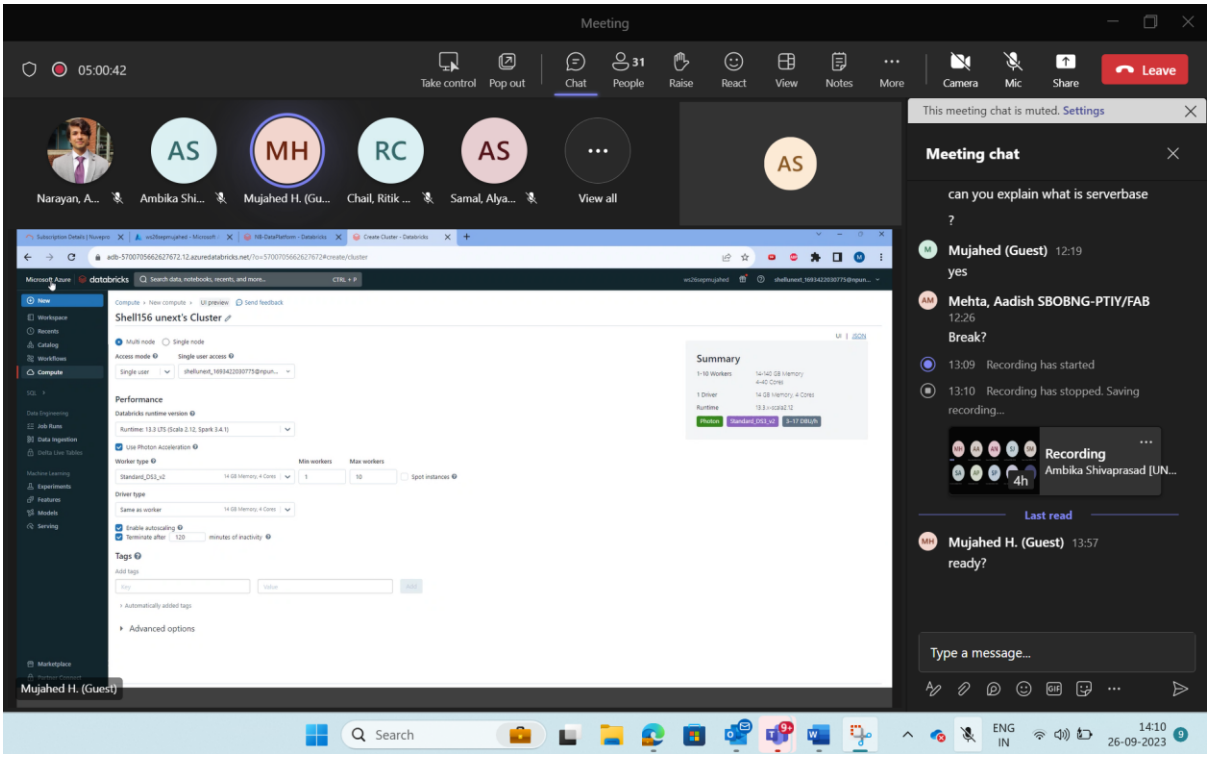

Azure Databricks - OneNote
Mujahed Hussaini
File
Home
Insert
Draw
History
Review
View
Mujahed Trainings
Retails
Corporates
Contents
Ruff
MockInterviews
Recording
IDA
Search (Ctrl+E)
User1
VCS (Github)
DB(WS)
Runtime
Storage
NB1
User10
ES
Git
SQL
Py
Scala
R
ADis
Day3- A
Azure D

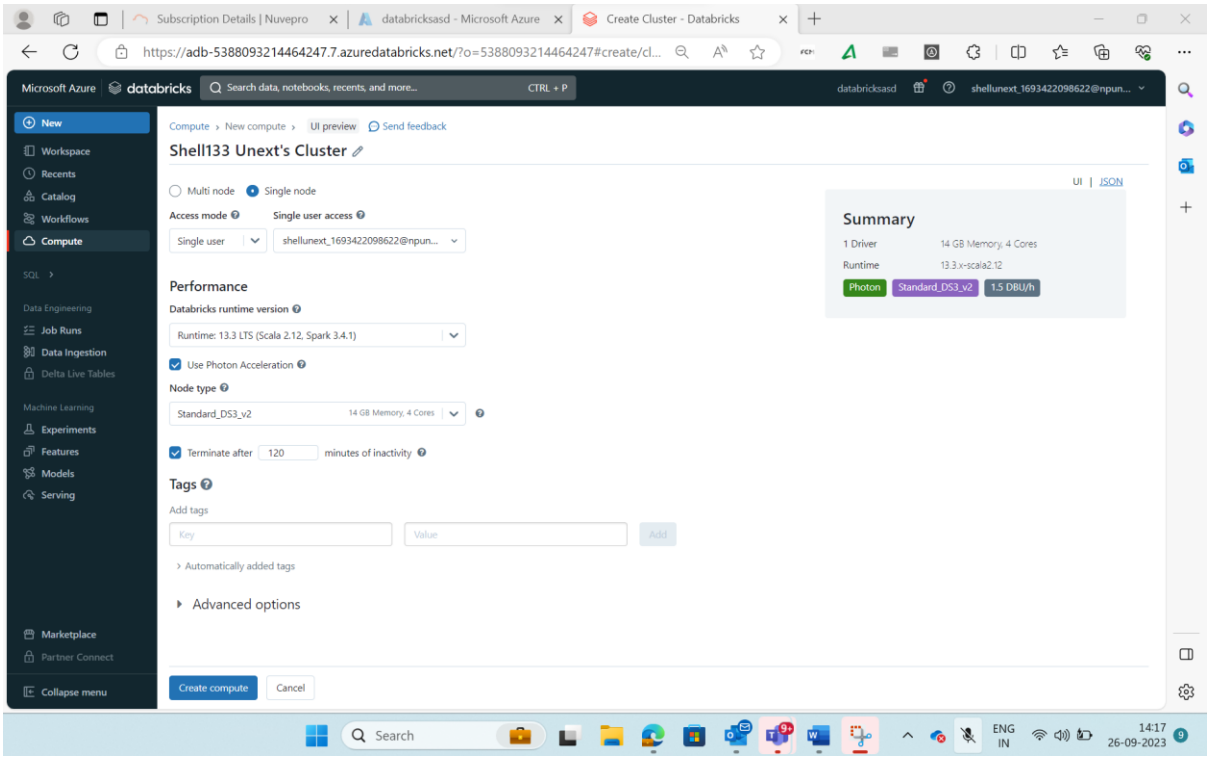

Task: Databricks Notebook & Workspace
1. Login with Lab --> Azure Credentials
2. Login Azure --> Create Resource Group
3. Create **Azure** Databricks Workspace
4. Open Workspace --> Under User --> Create Notebook
Mujahed H (Guest)
MH
Mujahed H...
AS
On hold
SJ
Jindal, Shiv...
YB
Bhati, Yoge...
AP
P, Akshay S...
Narayan, A...
VY
Y, Veera Ra...
View all
AS
Search
ENG
IN
11:28
26-09-2023

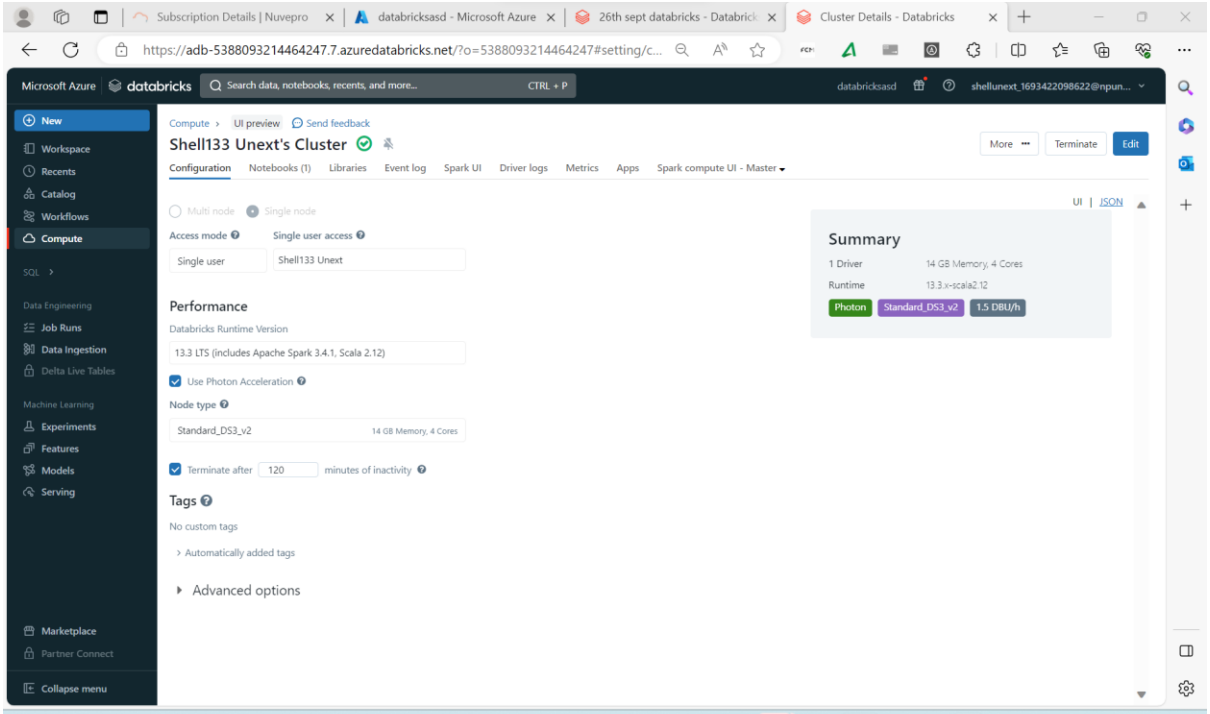

Subscription Details | Nuvepro
databricksasd - Microsoft Azure
26th sept databricks - Databricks
https://adb-5388093214464247.7.azuredatabricks.net/?o=5388093214464247#noteboo...
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
26th sept databricks
Python
File Edit View Run Help Last edit was now Provide feedback
Run all
Connect
Schedule
Share
New
Workspace
Recents
Catalog
Workflows
Compute
Job Runs
Data Ingestion
Delta Live Tables
Experiments
Features
Models
Serving
Marketplace
Partner Connect
Collapse menu
A new chart preview is available in your workspace
With this preview, enjoy faster chart rendering and improved chart layouts. You can opt-out of the new chart preview through the "New charts" tag or dashboard banner.
New charts: ON
Electronics
Furniture
Health & Beauty
Office Supplies
Feb 05, 2022
Electronics: $45,397,819
Furniture: $34,703,575
Health & Beauty: $10,380,280
Office Supplies: $314,013
Total Sales
Date
Not now
Turn on new charts
Search
ENG
IN
11:46
26-09-2023

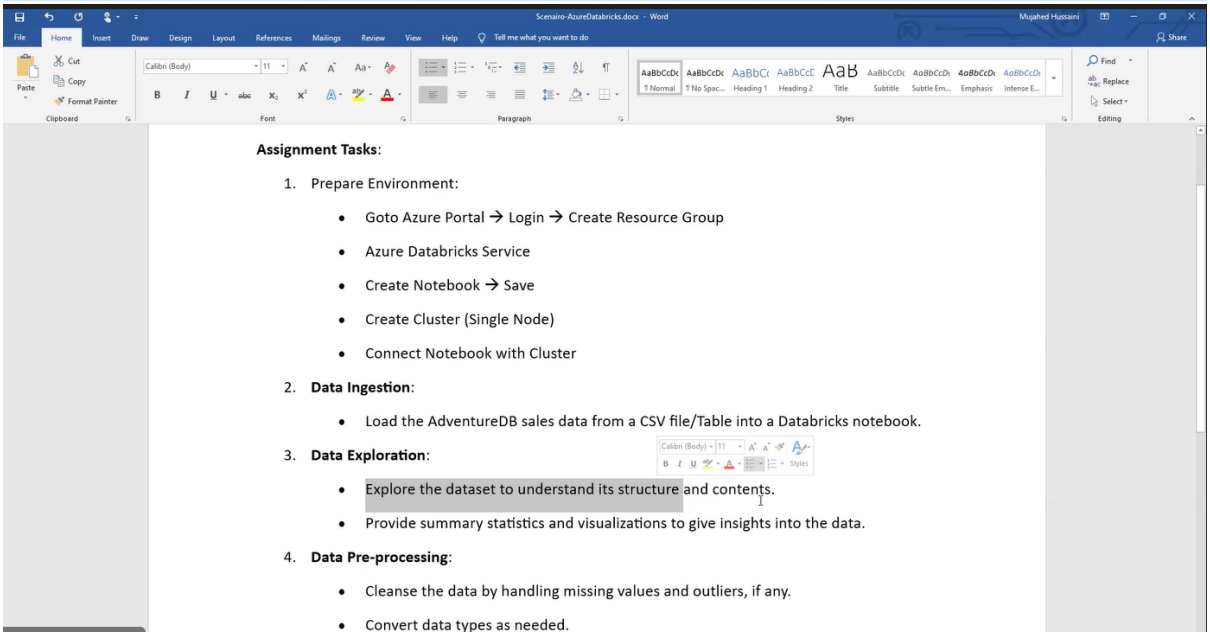

ServerBase
Azure VM
Server Less
pool
OS
App
UserName
Password
IPAddress
ADB
Server Base
VM
Azure VM
Controlling OS Level
Controlling App
??
Mujahed (Guest) 11:50
Compute
A, Sri Krishna SBOBNG-PTIY/FHD
11:56
visible
Jindal, Shivam SBOBNG-PTIY/FBA
12:18
can you explain what is serverbase
?
Mujahed (Guest) 12:19
yes
Mehta, Aadish SBOBNG-PTIY/FAB
12:26
Break?
Type a message...
Mujahed (Guest)
Search
ENG
IN
12:26
26-09-2023


Shell156 unext's Cluster
Summary
Performance
Tags
Advanced options
Mujahed H. (Guest)
Mujahed (Guest) 12:19
yes
Mehta, Aadish SBOBNG-PTIY/FAB
12:26
Break?
13:09 Recording has started
13:10 Recording has stopped. Saving recording...
Recording
Ambika Shivaprasad [UN...
Last read
Mujahed H. (Guest) 13:57
ready?
Type a message...
Search
ENG
IN
14:04
26-09-2023


Meeting
05:00:42
Take control
Pop out
Chat
People
Raise
React
View
Notes
More
Camera
Mic
Share
Leave
Narayan, A...
Ambika Shi...
Mujahed H. (Gu...
Chail, Ritik ...
Samal, Alya...
View all
This meeting chat is muted. Settings
Meeting chat
can you explain what is serverbase ?
Mujahed (Guest) 12:19
yes
Mehta, Aadish SBOBNG-PTIY/FAB
12:26
Break?
13:09 Recording has started
13:10 Recording has stopped. Saving recording...
Recording
Ambika Shivaprasad [UN...
Last read
Mujahed H. (Guest) 13:57
ready?
Type a message...
Shell156 unext's Cluster
Summary
Performance
Tags
Advanced options
Mujahed H. (Guest)
Search
ENG
IN
14:10
26-09-2023

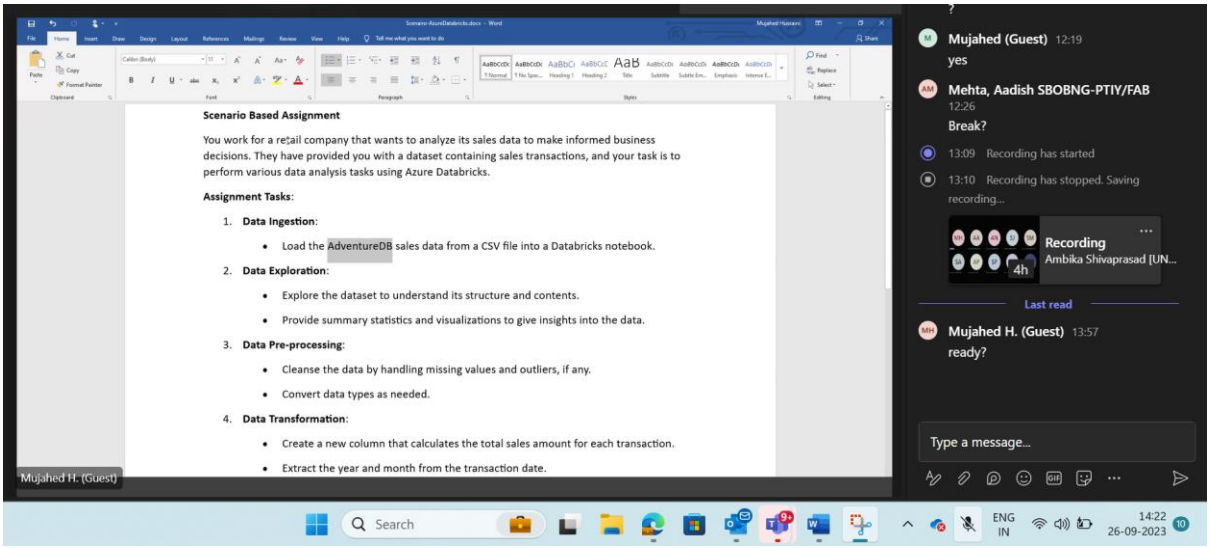

Subscription Details | Nuvepro
databricksasd - Microsoft Azure
Create Cluster - Databricks
https://adb-5388093214464247.7.azuredatabricks.net/?o=5388093214464247#create/cl...
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
databricksasd
shellunext_1693422098622@npun...
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
Collapse menu
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...
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
UI | JSON
Summary
1 Driver
14 GB Memory, 4 Cores
Runtime
13.3.x-scala2.12
Photon
Standard_DS3_v2
1.5 DBU/h
Search
ENG
IN
14:17
26-09-2023

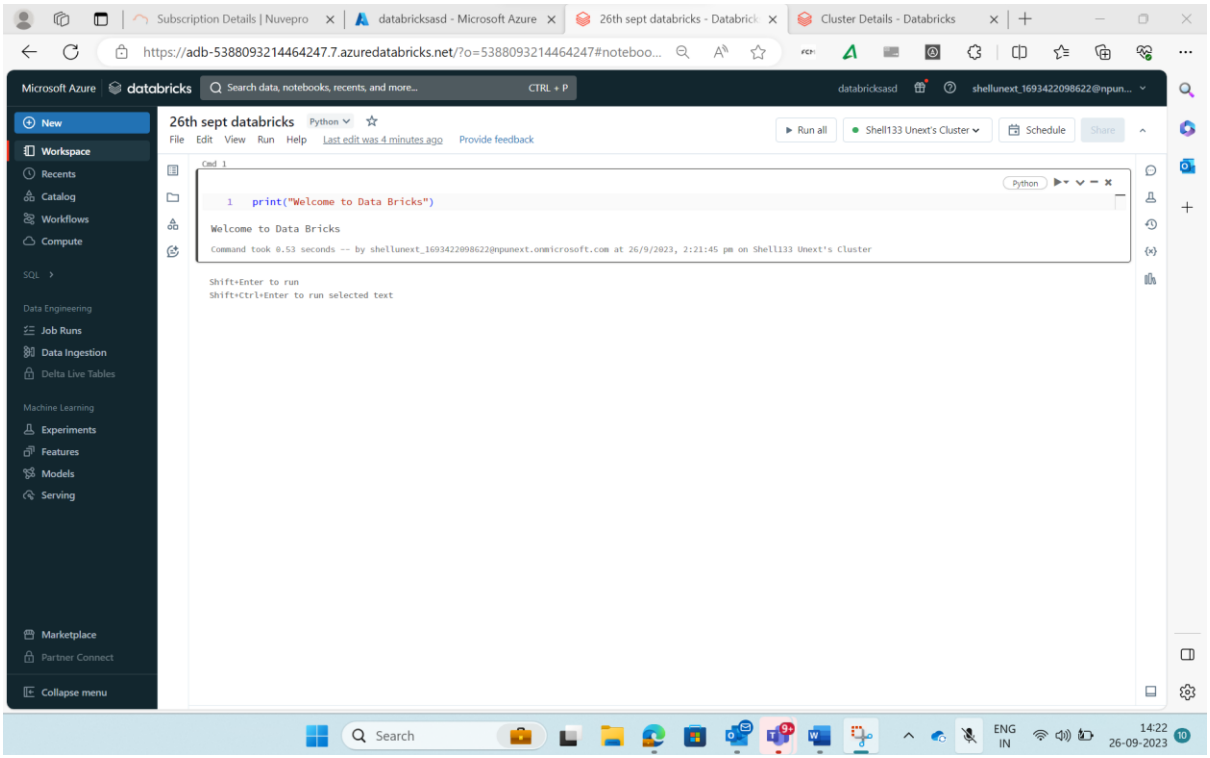


**Assignment Tasks**:

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.

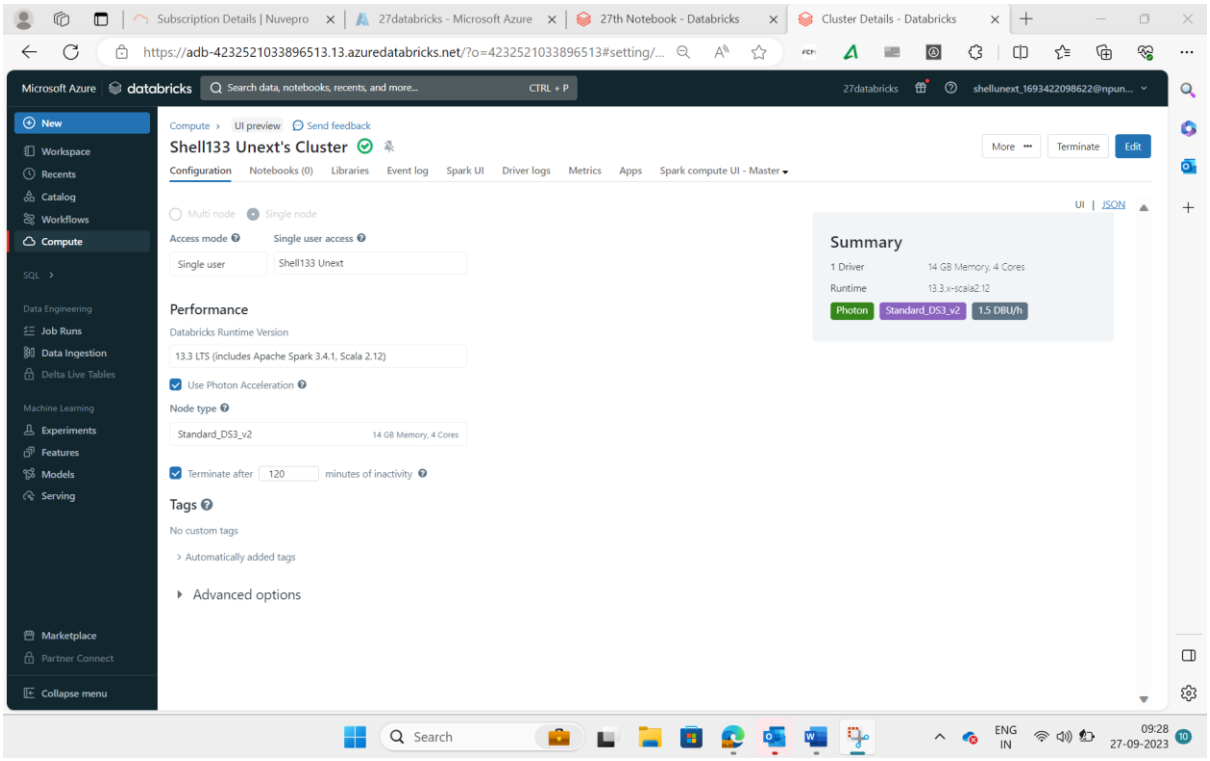


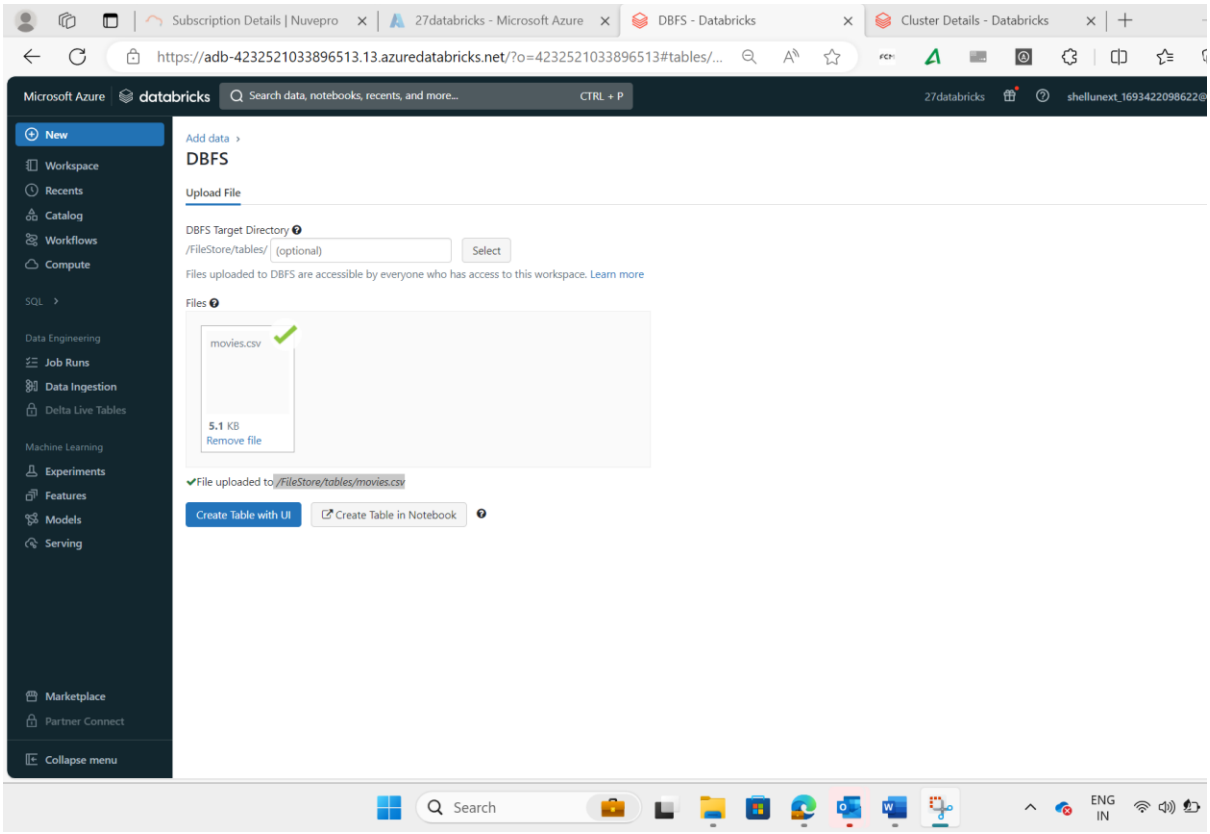

[Scenairo-AzureDatabricks.docx - Google Docs]

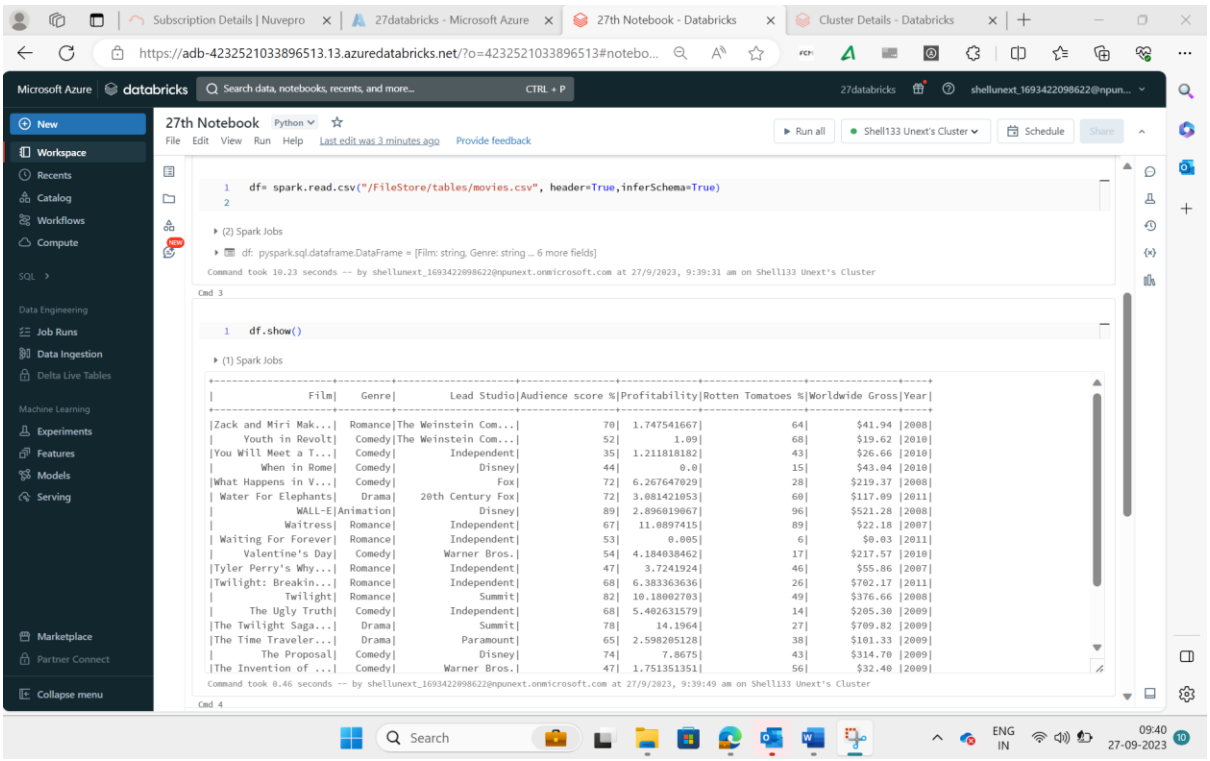
[Load data using the add data UI | Databricks on AWS]

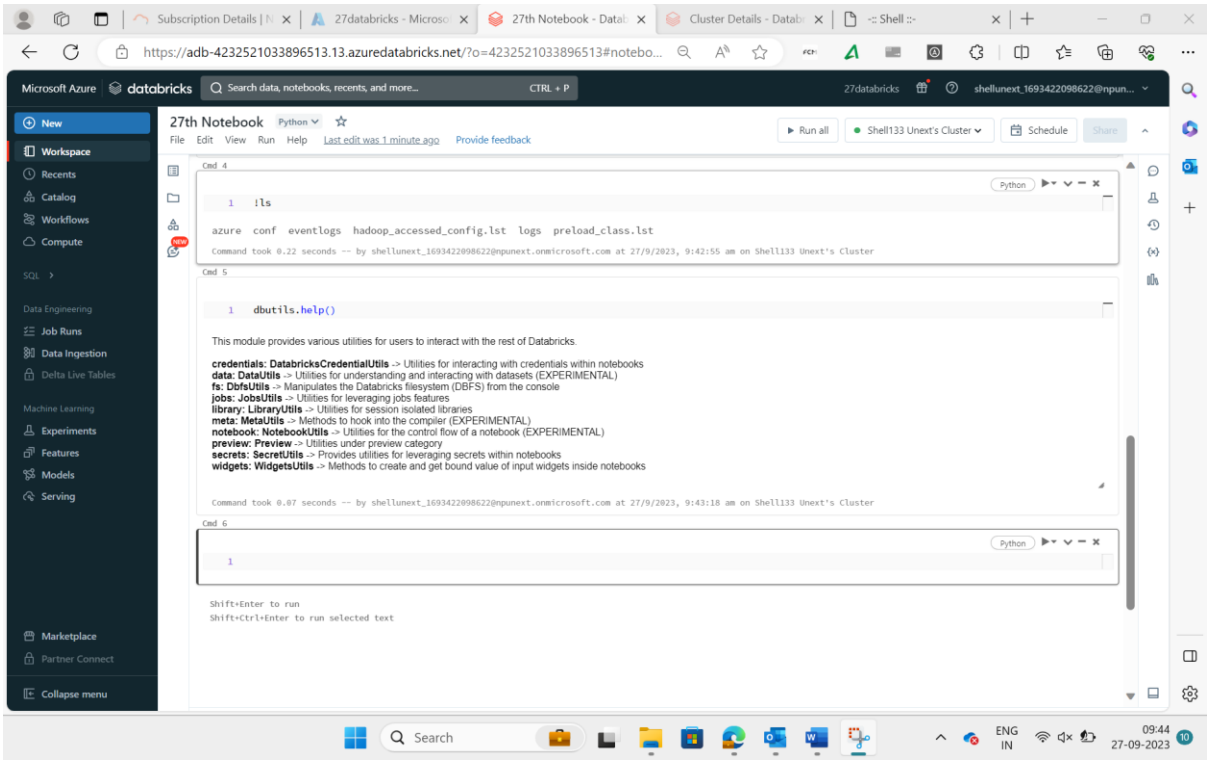
**Day 03(week 5) – 27th September 2023**

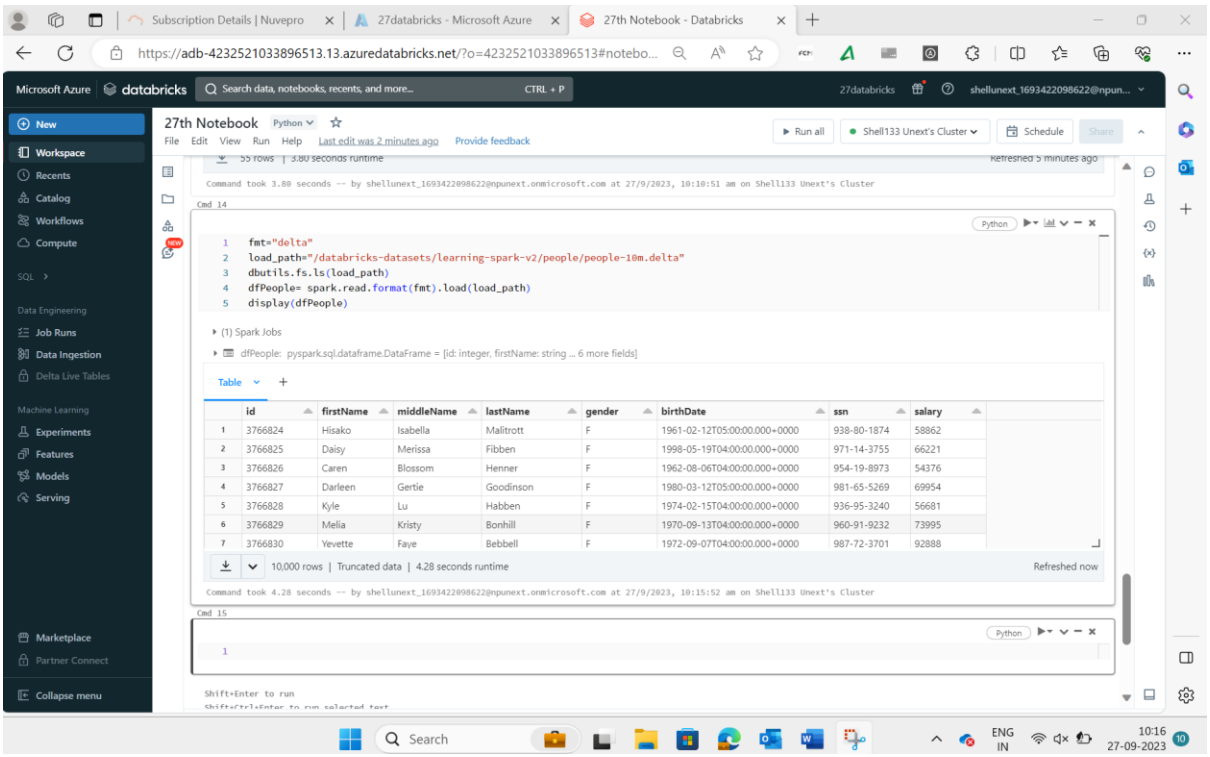



*/FileStore/tables/movies.csv*




Subscription Details | Nuvepro
27databricks - Microsoft Azure
27th Notebook - Databricks
Cluster Details - Databricks
https://adb-4232521033896513.13.azuredatabricks.net/?o=4232521033896513#notebo...
Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
27databricks
shellunext_1693422098622@npun...
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
Collapse menu
27th Notebook
Python
File Edit View Run Help
Last edit was 3 minutes ago
Provide feedback
Run all
Shell133 Unext's Cluster
Schedule
Share
1 df= spark.read.csv("/FileStore/tables/movies.csv", header=True,inferSchema=True)
2
(2) Spark Jobs
df: pyspark.sql.dataframe.DataFrame = [Film: string, Genre: string ... 6 more fields]
Command took 10.23 seconds -- by shellunext_1693422098622@npunext.onmicrosoft.com at 27/9/2023, 9:39:31 am on Shell133 Unext's Cluster
Cmd 3
1 df.show()
(1) Spark Jobs
+--------------------+---------+--------------------+----------------+-------------+-----------------+----------------+----+
| Film| Genre| Lead Studio|Audience score %|Profitability|Rotten Tomatoes %|Worldwide Gross|Year|
+--------------------+---------+--------------------+----------------+-------------+-----------------+----------------+----+
|Zack and Miri Mak...| Romance|The Weinstein Com...| 70| 1.747541667| 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| 1.211818182| 43| $26.66 |2010|
| When in Rome| Comedy| Disney| 44| 0.0| 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| 60| $117.09 |2011|
| WALL-E|Animation| Disney| 89| 2.896019067| 96| $521.28 |2008|
| Waitress| Romance| Independent| 67| 11.0897415| 89| $22.18 |2007|
| Waiting For Forever| Romance| Independent| 53| 0.005| 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| $702.17 |2011|
| Twilight| Romance| Summit| 82| 10.18002703| 49| $376.66 |2008|
| The Ugly Truth| Comedy| Independent| 68| 5.402631579| 14| $205.30 |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 shellunext_1693422098622@npunext.onmicrosoft.com at 27/9/2023, 9:39:49 am on Shell133 Unext's Cluster
Cmd 4
Search
ENG IN
09:40 27-09-2023

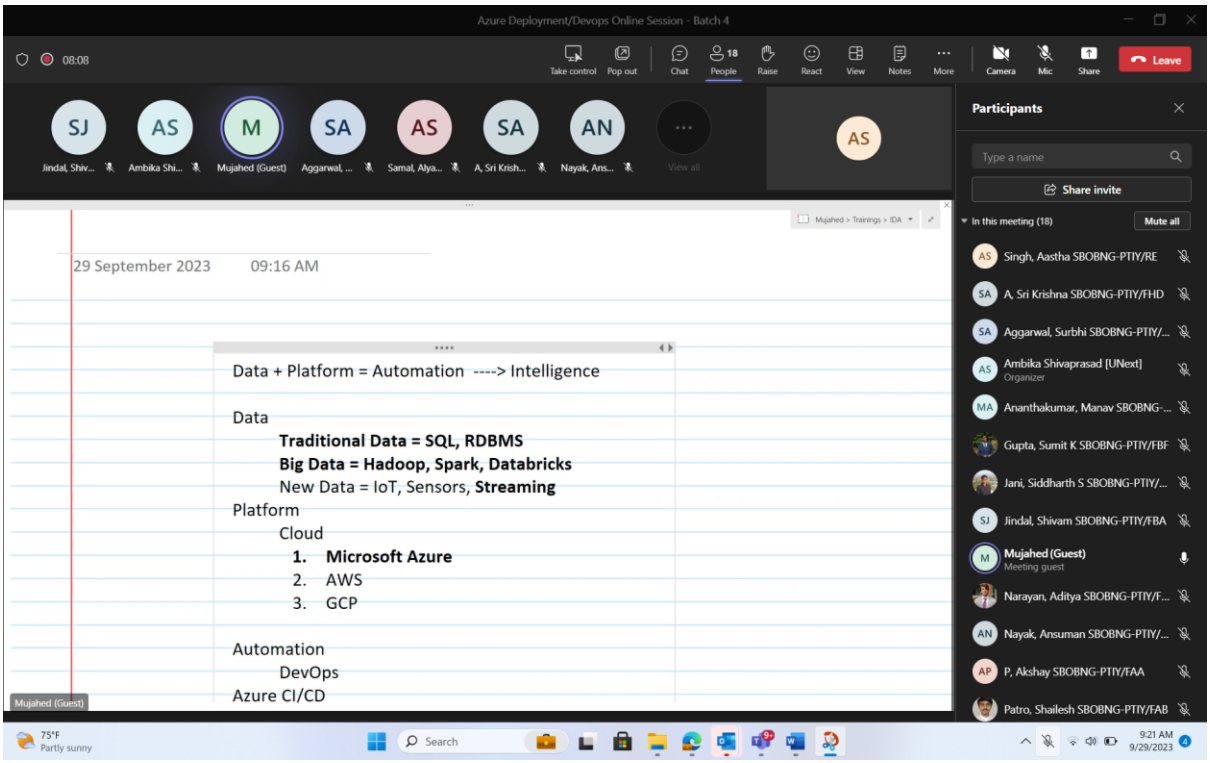

Subscription Details | N
27databricks - Microso
27th Notebook - Datab
Cluster Details - Databr
Shell ::-
https://adb-4232521033896513.13.azuredatabricks.net/?o=4232521033896513#notebo...
Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
27databricks
shellunext_1693422098622@npun...
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
Collapse menu
27th Notebook
Python
File Edit View Run Help
Last edit was 1 minute ago
Provide feedback
Run all
Shell133 Unext's Cluster
Schedule
Share
Cmd 4
1 !ls
azure conf eventlogs hadoop_accessed_config.lst logs preload_class.lst
Command took 0.22 seconds -- by shellunext_1693422098622@npunext.onmicrosoft.com at 27/9/2023, 9:42:55 am on Shell133 Unext's Cluster
Cmd 5
1 dbutils.help()
This module provides various utilities for users to interact with the rest of Databricks.
credentials: DatabricksCredentialUtils -> Utilities for interacting with credentials within notebooks
data: DataUtils -> Utilities for understanding and interacting with datasets (EXPERIMENTAL)
fs: DbfsUtils -> Manipulates the Databricks filesystem (DBFS) from the console
jobs: JobsUtils -> Utilities for leveraging jobs features
library: LibraryUtils -> Utilities for session isolated libraries
meta: MetaUtils -> Methods to hook into the compiler (EXPERIMENTAL)
notebook: NotebookUtils -> Utilities for the control flow 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 notebooks
Command took 0.07 seconds -- by shellunext_1693422098622@npunext.onmicrosoft.com at 27/9/2023, 9:43:18 am on Shell133 Unext's Cluster
Cmd 6
1
Shift+Enter to run
Shift+Ctrl+Enter to run selected text
Search
ENG IN
09:44 27-09-2023

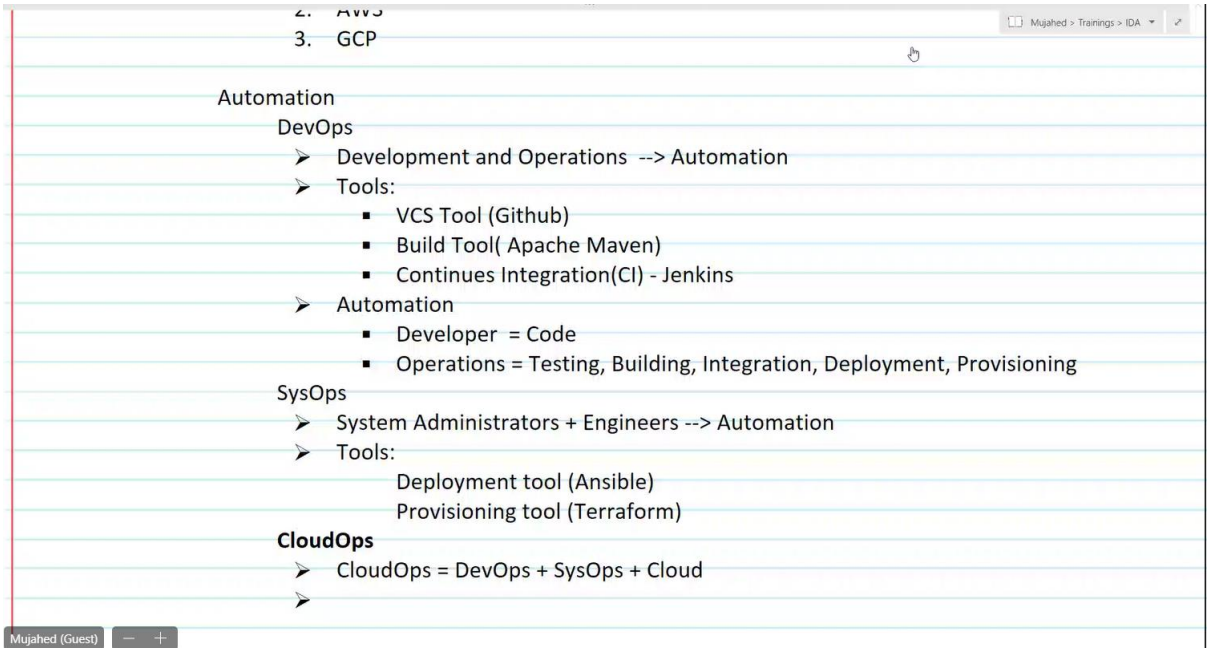


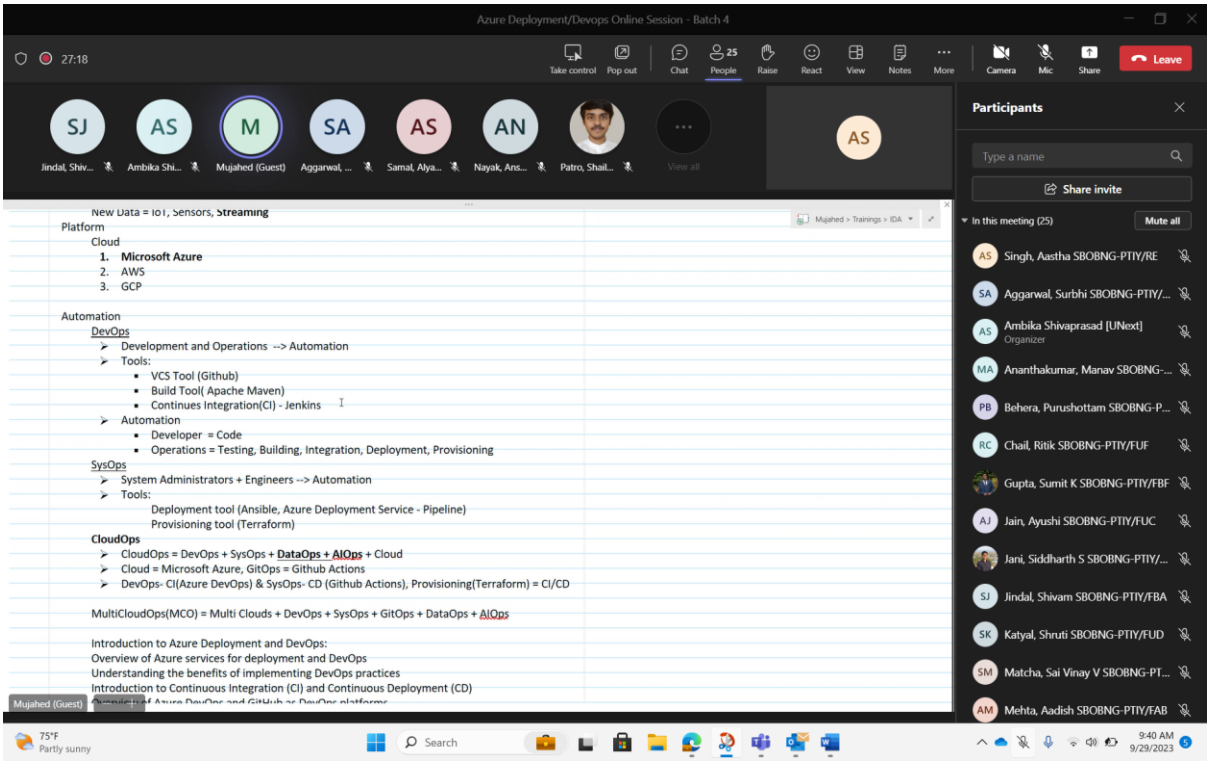

**Day 04(week 5) – 28th September 2023**

**Went to EcoWorld to collect laptop**

**Day 05(week 5) – 29th September 2023**


Azure Deployment/Devops Online Session - Batch 4
Participants
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


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
➢


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, Shail...
View all
Participants
Type a name
Share invite
In this meeting (25)
Mute all
Singh, Aastha SBOBNG-PTIY/RE
Aggarwal, Surbhi SBOBNG-PTIY/...
Ambika Shivaprasad [UNext]
Organizer
Ananthakumar, Manav SBOBNG-...
Behera, Purushottam SBOBNG-P...
Chail, Ritik SBOBNG-PTIY/FUF
Gupta, Sumit K SBOBNG-PTIY/FBF
Jain, Ayushi SBOBNG-PTIY/FUC
Jani, Siddharth S SBOBNG-PTIY/...
Jindal, Shivam SBOBNG-PTIY/FBA
Katyal, Shruti SBOBNG-PTIY/FUD
Matcha, Sai Vinay V SBOBNG-PT...
Mehta, Aadish SBOBNG-PTIY/FAB
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)
75°F
Partly sunny
Search
9:40 AM
9/29/2023