Journey Presentation

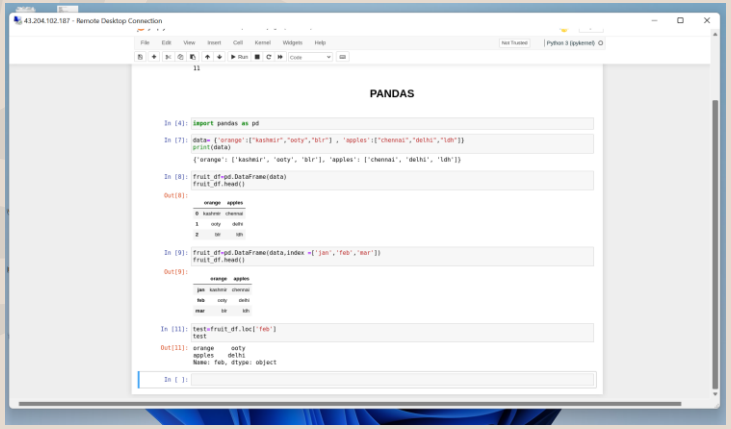
Harsh Thakur

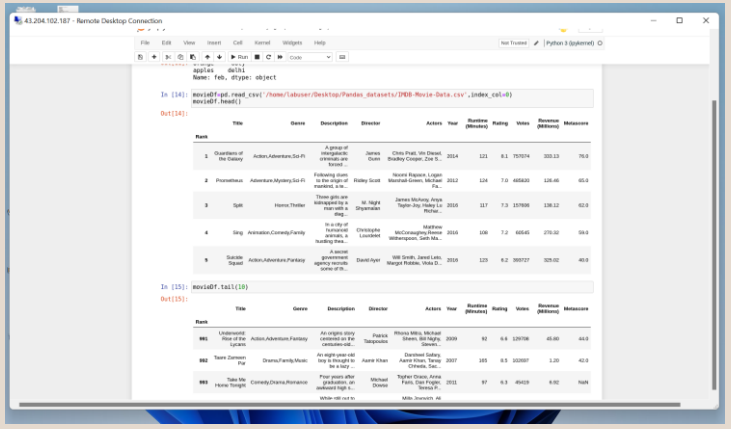
Harsh.Thakur@shell.com

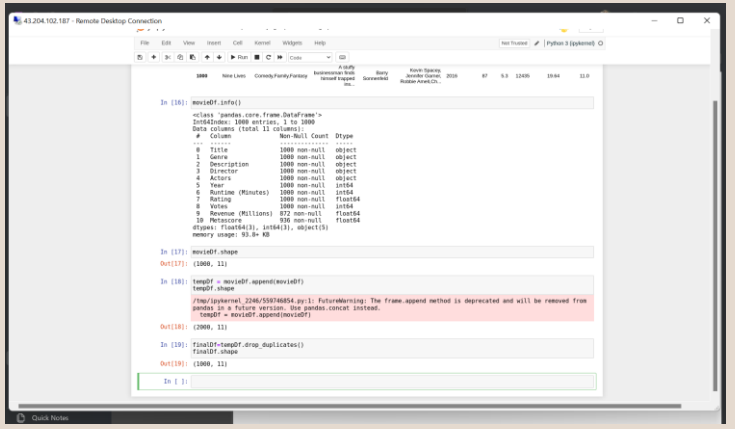
DAY-15

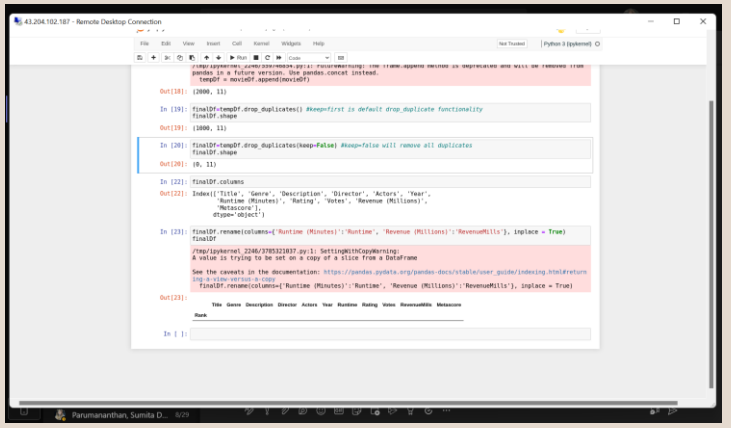
Python – Day 3

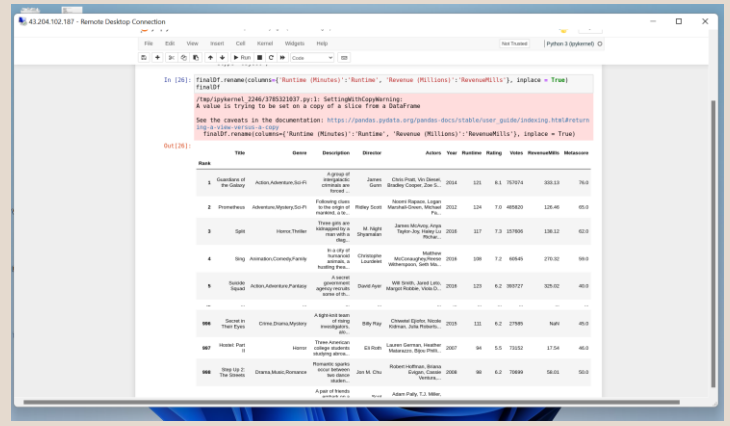
1. Pandas : Panel Datasets
    a. Dataframe
    b. Functions on dataframe
    c. Changing column to lowercase
    d. Dropping Null Values
    e. Dropping columns with Null Values
    f. Filling null values
    g. Describe columns
    h. Value counts()
    i. Selected rows or columns of dataframe
    j. Conditional selection
2. Matplotlib
3. Arrays In Python

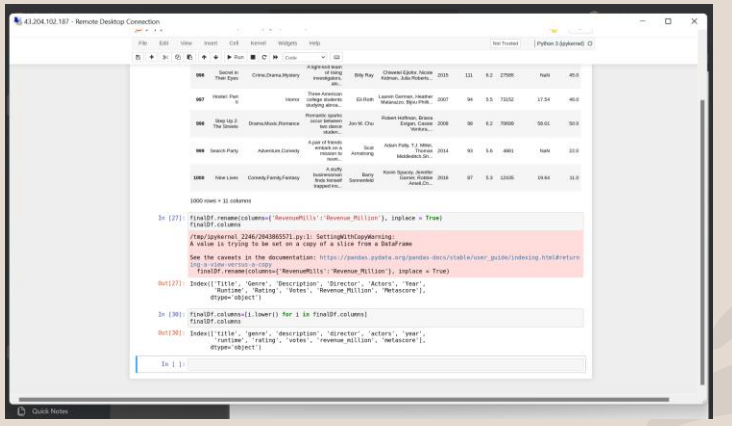

43.204.102.187 - Remote Desktop Connection
PANDAS
In [4]: import pandas as pd
In [7]: data= {'orange':['kashmir','ooty','blr'] , 'apples':['chennai','delhi','ldh']}
print(data)
{'orange': ['kashmir', 'ooty', 'blr'], 'apples': ['chennai', 'delhi', 'ldh']}
In [8]: fruit_df=pd.DataFrame(data)
fruit_df.head()
Out[8]:
orange apples
0 kashmir chennai
1 ooty delhi
2 blr ldh
In [9]: fruit_df=pd.DataFrame(data,index ='jan','feb','mar'])
fruit_df.head()
Out[9]:
orange apples
jan kashmir chennai
feb ooty delhi
mar blr ldh
In [11]: test=fruit_df.loc['feb']
test
Out[11]: orange ooty
apples delhi
Name: feb, dtype: object

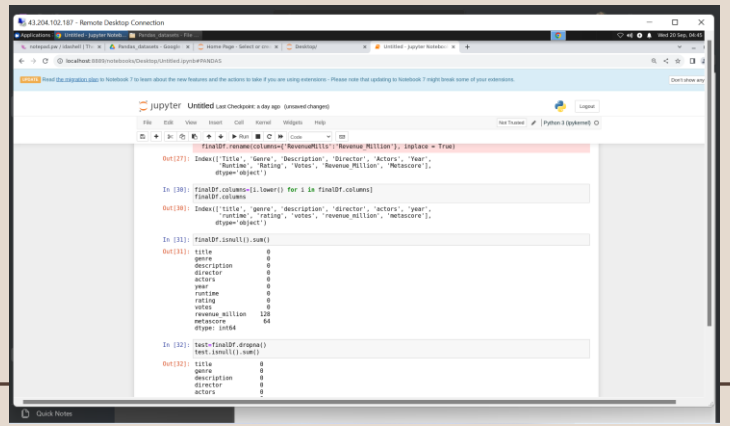

43.204.102.187 - Remote Desktop Connection
In [14]: movieDf=pd.read_csv('/home/labuser/Desktop/Pandas_datasets/IMDB-Movie-Data.csv',index_col=0)
movieDf.head()
In [15]: movieDf.tail(10)

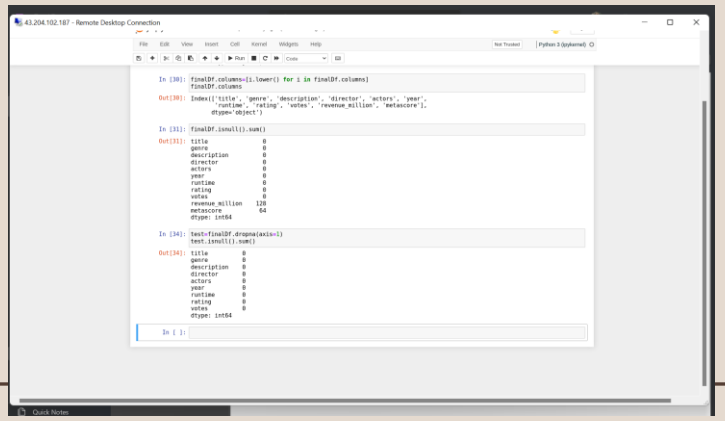

43.204.102.187 - Remote Desktop Connection
In [16]: movieDf.info()
In [17]: movieDf.shape
Out[17]: (1000, 11)
In [18]: tempDf = movieDf.append(movieDf)
tempDf.shape
In [19]: finalDf=tempDf.drop_duplicates()
finalDf.shape
Out[19]: (1000, 11)

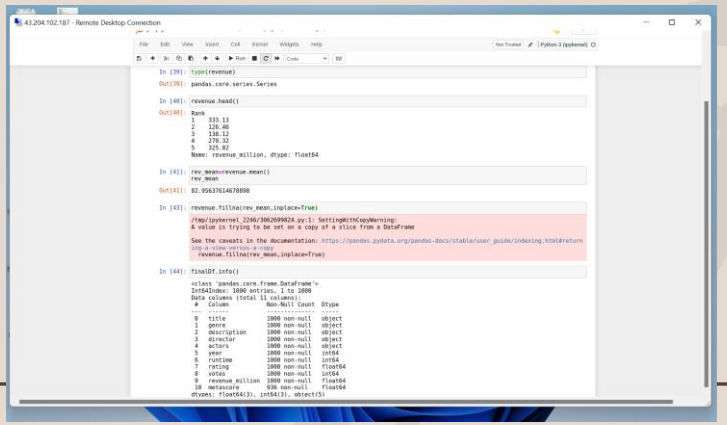

43.204.102.187 - Remote Desktop Connection
Out[18]: (2000, 11)
In [19]: finalDf=tempDf.drop_duplicates() #keep=first is default drop_duplicate functionality
finalDf.shape
Out[19]: (1000, 11)
In [20]: finalDf=tempDf.drop_duplicates(keep=False) #keep=false will remove all duplicates
finalDf.shape
Out[20]: (0, 11)
In [22]: finalDf.columns
In [23]: finalDf.rename(columns={'Runtime (Minutes)':'Runtime', 'Revenue (Millions)':'RevenueMills'}, inplace = True)
finalDf

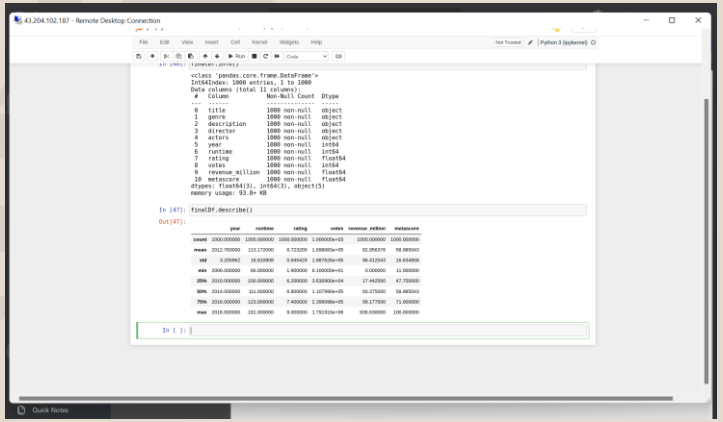

43.204.102.187 - Remote Desktop Connection
In [26]: finalDf.rename(columns={'Runtime (Minutes)':'Runtime', 'Revenue (Millions)':'RevenueMills'}, inplace = True)
finalDf
Out[26]:

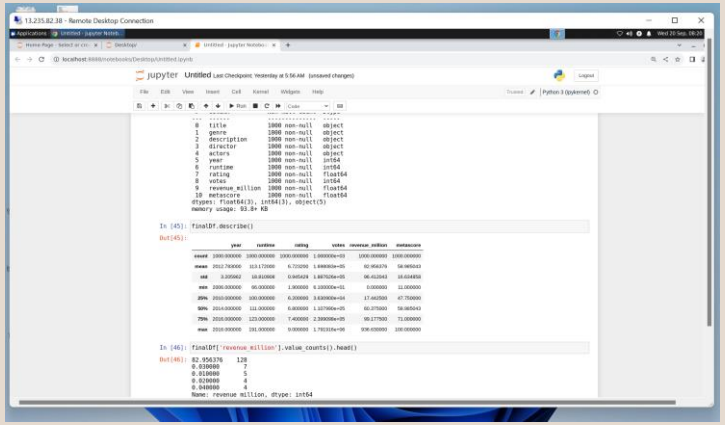

43.204.102.187 - Remote Desktop Connection
In [27]: finalDf.rename(columns={'RevenueMills':'Revenue_Million'}, inplace = True)
finalDf.columns
In [30]: finalDf.columns=[i.lower() for i in finalDf.columns]
finalDf.columns

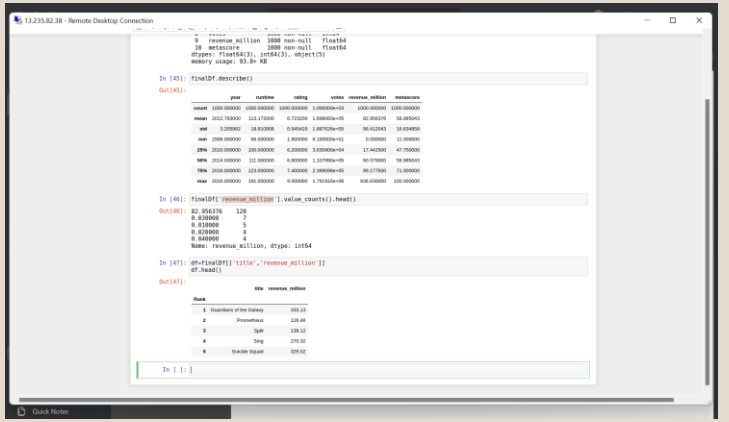

43.204.102.187 - Remote Desktop Connection
jupyter Untitled
In [31]: finalDf.isnull().sum()
In [32]: test=finalDf.dropna()
test.isnull().sum()

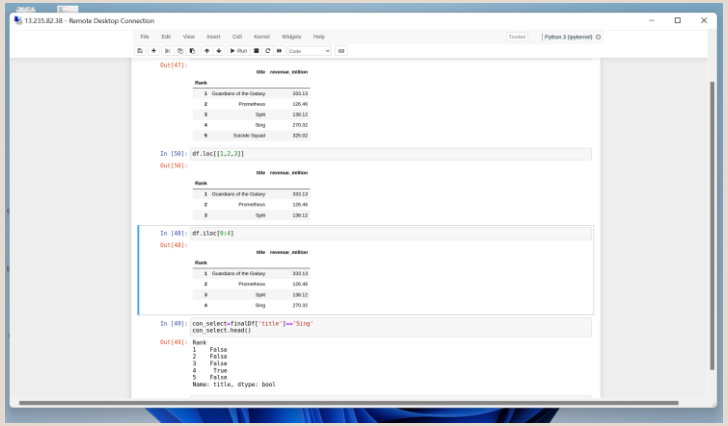

43.204.102.187 - Remote Desktop Connection
In [34]: test=finalDf.dropna(axis=1)
test.isnull().sum()

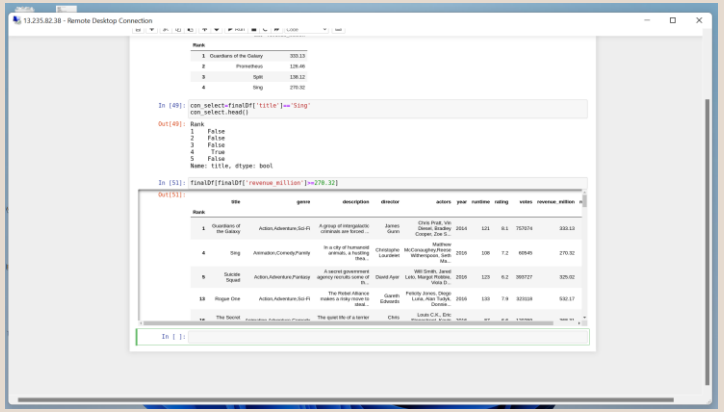

43.204.102.187 - Remote Desktop Connection
In [39]: type(revenue)
Out[39]: pandas.core.series.Series
In [40]: revenue.head()
In [41]: rev_mean=revenue.mean()
rev_mean
In [43]: revenue.fillna(rev_mean,inplace=True)
In [44]: finalDf.info()

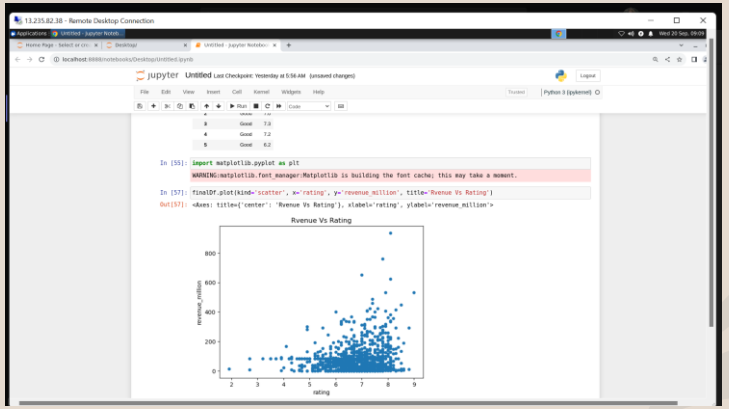

43.204.102.187 - Remote Desktop Connection


13.235.82.38 - Remote Desktop Connection


13.235.82.38 - Remote Desktop Connection


13.235.82.38 - Remote Desktop Connection

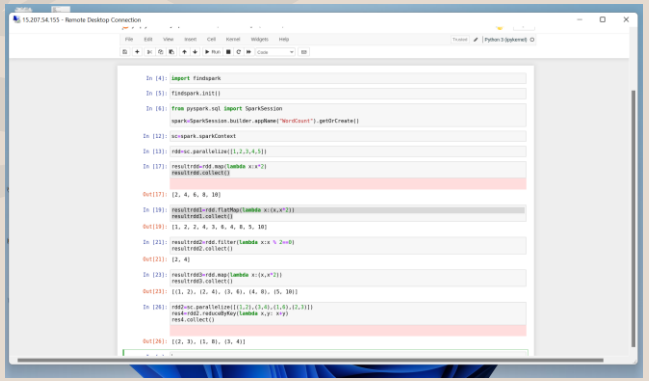

13.235.82.38 - Remote Desktop Connection

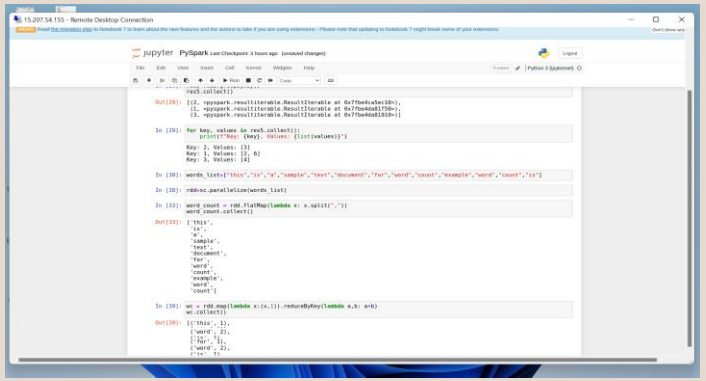

13.235.82.38 - Remote Desktop Connection
Revenue Vs Rating

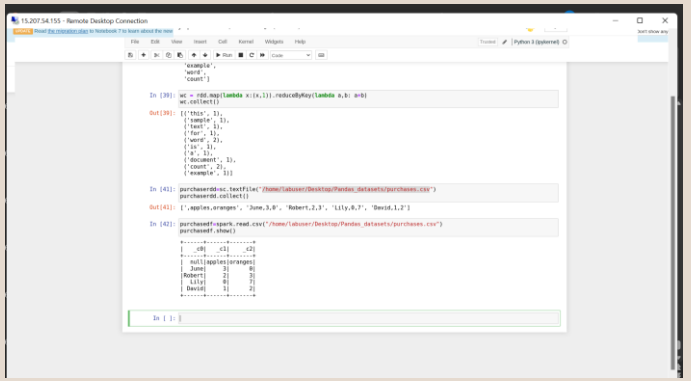
DAY-16

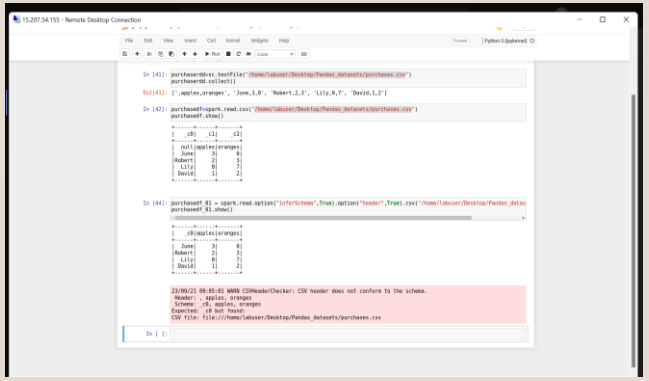
Apache PySpark – Day 1

1. Driver process
2. Executors
3. Cluster Manager
4. Partition
5. Job, Stage and Task
6. Lazy Evaluation
7. RDD (Resilient Distributed Dataset)
8. Caching
9. Catalyst and Tungsten
10. RDD vs Dataframe
11. Deployment Types
    1. Client Mode
    2. Cluster Mode


15.207.54.155 - Remote Desktop Connection
In [4]: import findspark
In [5]: findspark.init()
In [6]: from pyspark.sql import SparkSession
spark=SparkSession.builder.appName("WordCount").getOrCreate()
In [12]: sc=spark.sparkContext
In [13]: rdd=sc.parallelize([1,2,3,4,5])
In [17]: resultrdd=rdd.map(lambda x:x*2)
resultrdd.collect()
Out[17]: [2, 4, 6, 8, 10]
In [19]: resultrdd1=rdd.flatMap(lambda x:(x,x*2))
resultrdd1.collect()
Out[19]: [1, 2, 2, 4, 3, 6, 4, 8, 5, 10]
In [21]: resultrdd2=rdd.filter(lambda x:x % 2==0)
resultrdd2.collect()
Out[21]: [2, 4]
In [23]: resultrdd3=rdd.map(lambda x:(x,x*2))
resultrdd3.collect()
Out[23]: [(1, 2), (2, 4), (3, 6), (4, 8), (5, 10)]
In [26]: rdd2=sc.parallelize([(1,2),(3,6),(1,4),(2,3)])
res4=rdd2.reduceByKey(lambda x,y: x+y)
res4.collect()
Out[26]: [(2, 3), (1, 8), (3, 4)]


15.207.54.155 - Remote Desktop Connection
jupyter PySpark Last Checkpoint: 3 hours ago (unsaved changes)
Logout
Out[28]: [(2, <pyspark.resultiterable.ResultIterable at 0x7fbe4ca3ac10>),
(1, <pyspark.resultiterable.ResultIterable at 0x7fbe4da08170>),
(3, <pyspark.resultiterable.ResultIterable at 0x7fbe4da01010>)]
In [29]: for key, values in res5.collect():
print(f"Key: {key}, Values: {list(values)}")
Key: 2, Values: [3]
Key: 1, Values: [2, 6]
Key: 3, Values: [4]
In [30]: words_lists=["this","is","a","sample","text","document","for","word","count","example","word","count","is"]
In [38]: rdd=sc.parallelize(words_list)
In [33]: word_count = rdd.flatMap(lambda x: x.split(","))
word_count.collect()
Out[33]: ['this',
'is',
'a',
'sample',
'text',
'document',
'for',
'word',
'count',
'example',
'word',
'count']
In [39]: wc = rdd.map(lambda x:(x,1)).reduceByKey(lambda a,b: a+b)
wc.collect()
Out[39]: [('this', 1),
('is', 1),
('text', 1),
('for', 1),
('word', 2),
('is', 1),
('a', 1),
('document', 1),
('count', 2),
('example', 1)]

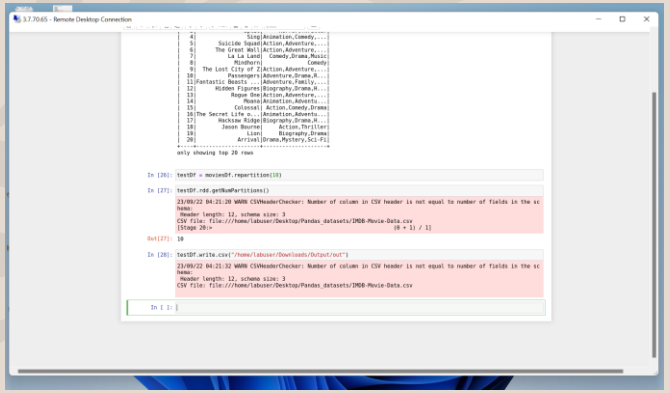

15.207.54.155 - Remote Desktop Connection
'example',
'word',
'count']
In [39]: wc = rdd.map(lambda x:(x,1)).reduceByKey(lambda a,b: a+b)
wc.collect()
Out[39]: [('this', 1),
('is', 1),
('text', 1),
('for', 1),
('word', 2),
('is', 1),
('a', 1),
('document', 1),
('count', 2),
('example', 1)]
In [41]: purchasesrdd=sc.textFile("/home/labuser/Desktop/Pandas_datasets/purchases.csv")
purchasesrdd.collect()
Out[41]: [',apples,oranges', 'June,3,0', 'Robert,2,3', 'Lily,0,7', 'David,1,2']
In [42]: purchasesdf=spark.read.csv("/home/labuser/Desktop/Pandas_datasets/purchases.csv")
purchasesdf.show()
+-----+----+-------+
| _c0| _c1| _c2|
+-----+----+-------+
| null|apples|oranges|
| June| 3| 0|
|Robert| 2| 3|
| Lily| 0| 7|
|David| 1| 2|
+-----+----+-------+
In [ ]:

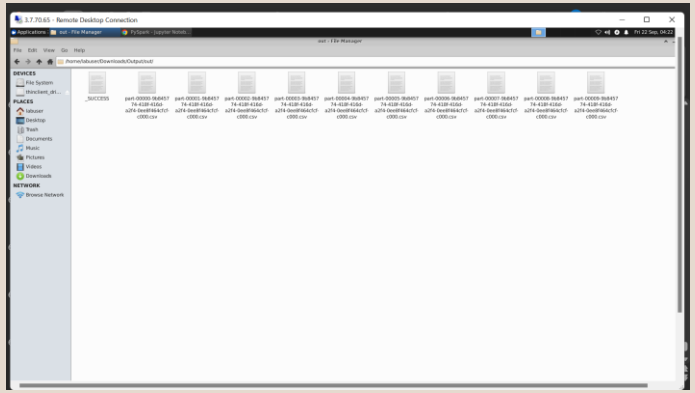

15.207.54.155 - Remote Desktop Connection
In [41]: purchasesrdd=sc.textFile("/home/labuser/Desktop/Pandas_datasets/purchases.csv")
purchasesrdd.collect()
Out[41]: [',apples,oranges', 'June,3,0', 'Robert,2,3', 'Lily,0,7', 'David,1,2']
In [42]: purchasesdf=spark.read.csv("/home/labuser/Desktop/Pandas_datasets/purchases.csv")
purchasesdf.show()
+-----+----+-------+
| _c0| _c1| _c2|
+-----+----+-------+
| null|apples|oranges|
| June| 3| 0|
|Robert| 2| 3|
| Lily| 0| 7|
|David| 1| 2|
+-----+----+-------+
In [44]: purchasesdf_01 = spark.read.option("inferSchema",True).option("header",True).csv("/home/labuser/Desktop/Pandas_datas
purchasesdf_01.show()
+-----+----+-------+
| _c0|apples|oranges|
+-----+----+-------+
| June| 3| 0|
|Robert| 2| 3|
| Lily| 0| 7|
|David| 1| 2|
+-----+----+-------+
23/09/21 09:05:01 WARN CSVHeaderChecker: CSV header does not conform to the schema.
Header: , apples, oranges
Schema: _c0, apples, oranges
Expected: _c0 but found:
CSV file: file:///home/labuser/Desktop/Pandas_datasets/purchases.csv
In [ ]:

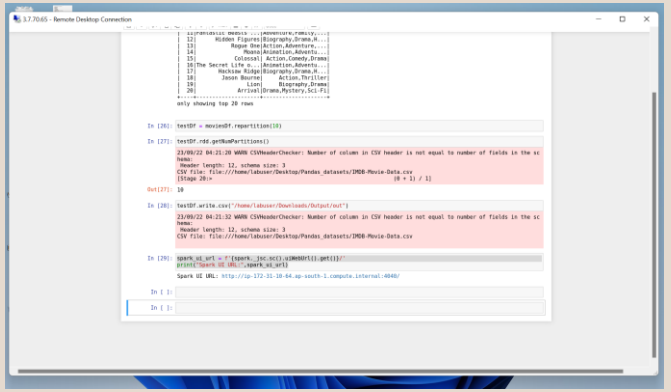
DAY-17

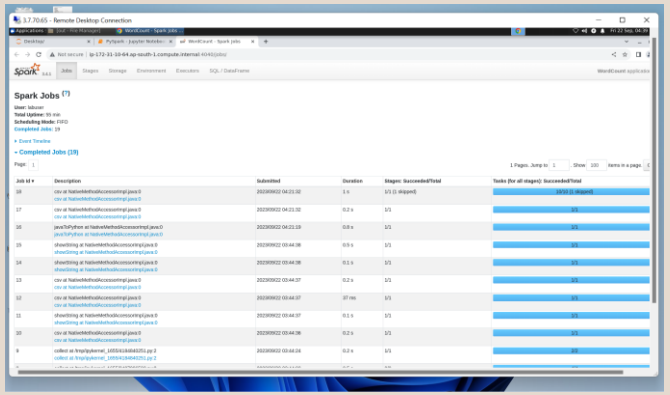
Apache PySpark – Day 2

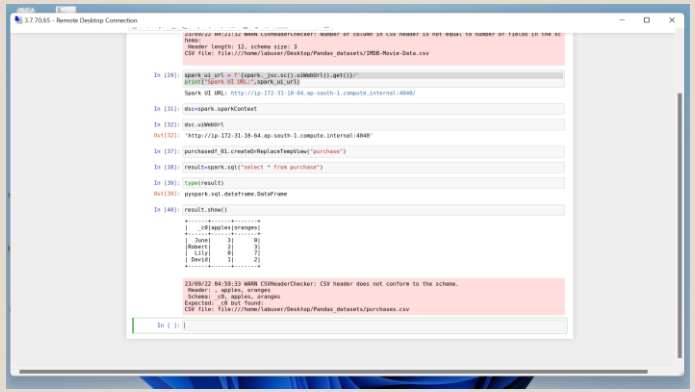
1. Partitioning
2. Spark UI URL
3. SQL in PySpark
4. Drop duplicates
5. When – Otherwise condition
6. Concat and Functions
7. Caching
8. Joins
9. Group By

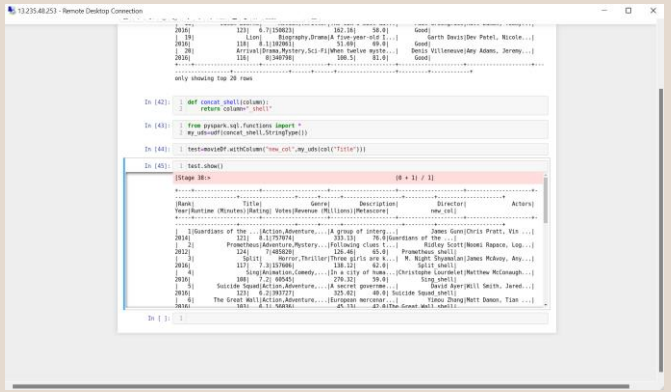


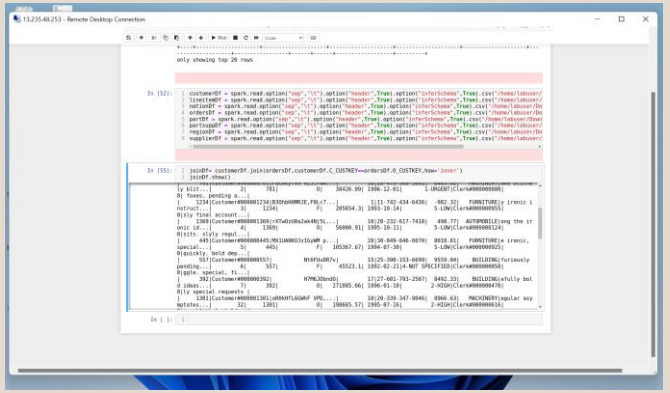


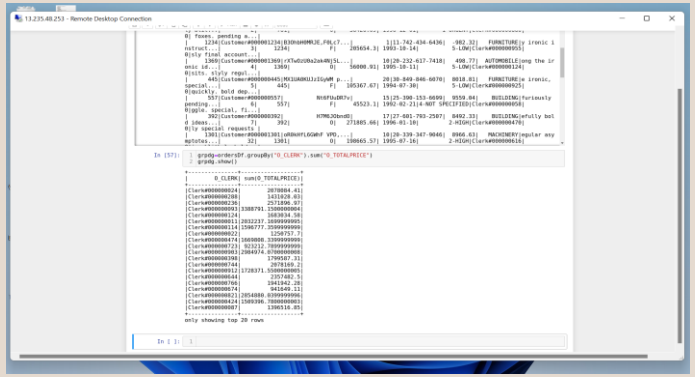












DAY-18

Apache PySpark – Day 3

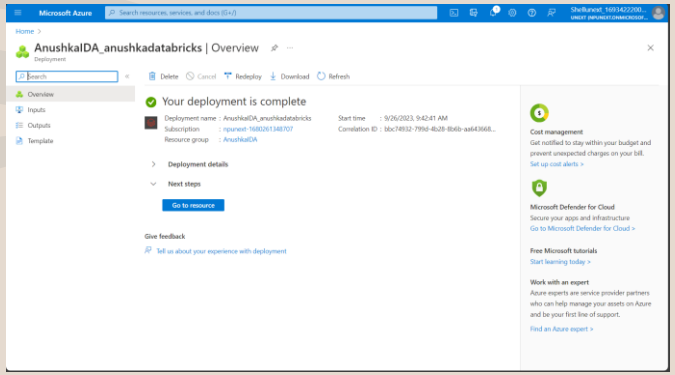
1. Caching
2. Persist and different storage levels
3. Aggregation
4. Managed and External Tables
5. Partitioning
6. For-Each

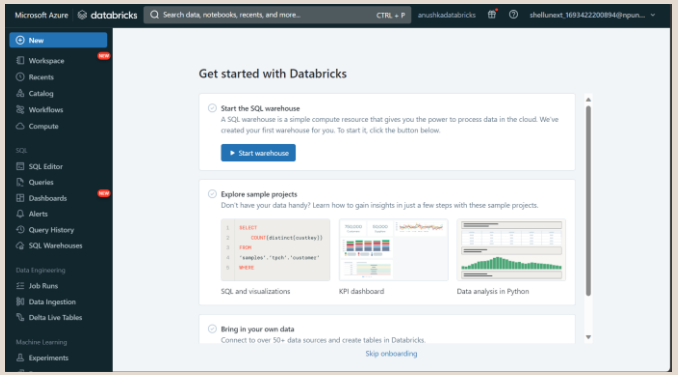
** Note : Only 4 hours training due to OneIDA event. **

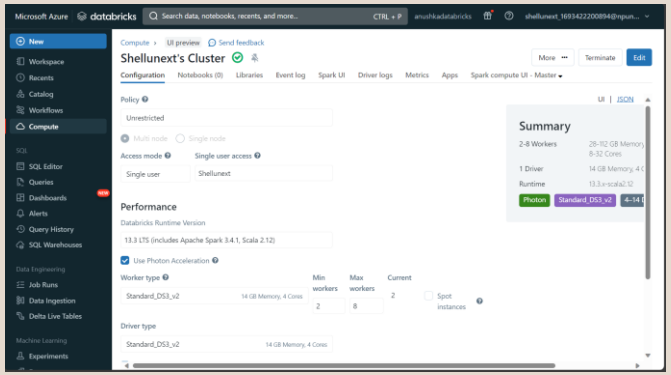
DAY-19

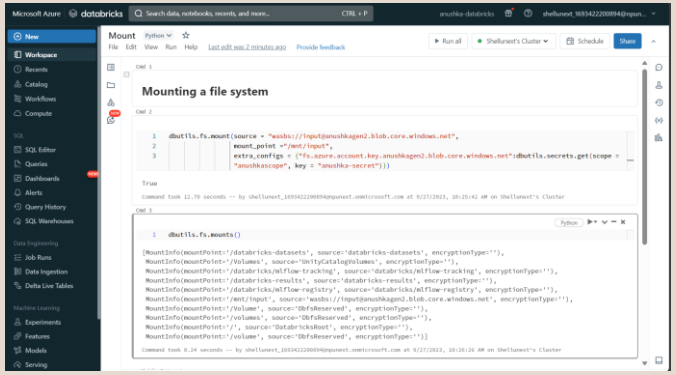
Azure Databricks – Day 1

1. What is databricks?
2. Setting Up azure databricks
3. Overview of Databricks Workspace
4. Cluster configuration
5. Mounting
6. Widgets
7. Parameters
8. Functions
9. %run, %md, %language
10. Structured streaming


AnushkaIDA_anushkadatabricks | Overview
Your deployment is complete
Deployment details
Next steps
Go to resource
Give feedback


Get started with Databricks
Start the SQL warehouse
Start warehouse
Explore sample projects
Bring in your own data
Skip onboarding

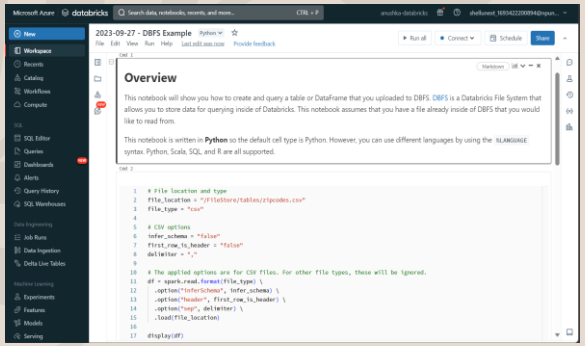

Shellunext's Cluster
Summary
Performance

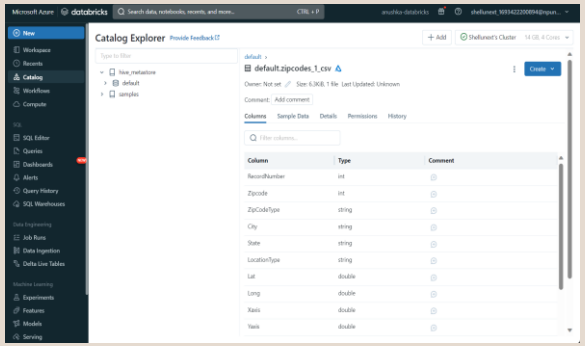

Mount
Mounting a file system
dbutils.fs.mount(source = "wasbs://input@anushkagen2.blob.core.windows.net",
dbutils.fs.mounts()

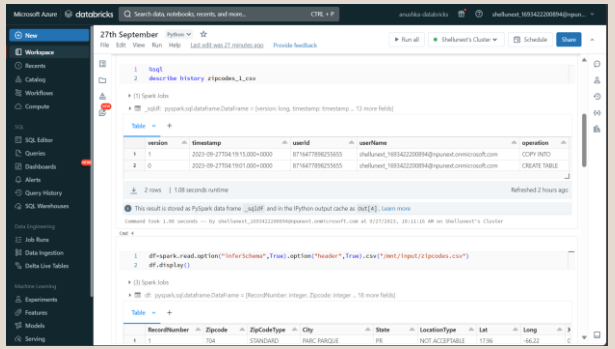
DAY-20

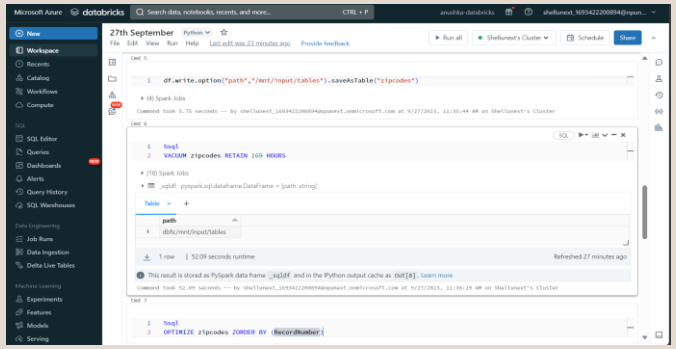
Azure Databricks – Day 2

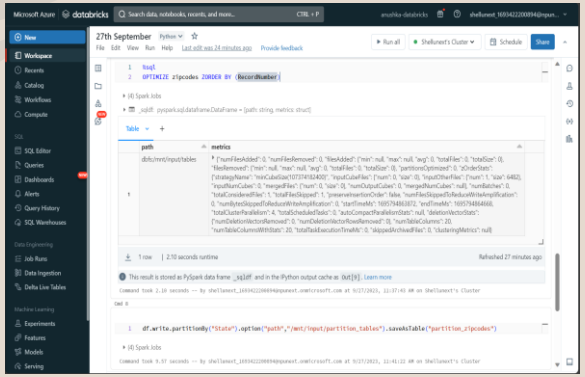
1. Structured streaming
2. DBFS from sample data
3. Delta table & Parquet table
4. Partition Tables
5. Medallion / Multi hop Architecture
6. Workflows → Jobs
7. Accessing data from Azure SQL database
8. Unity catalog in databricks
9. Metastore

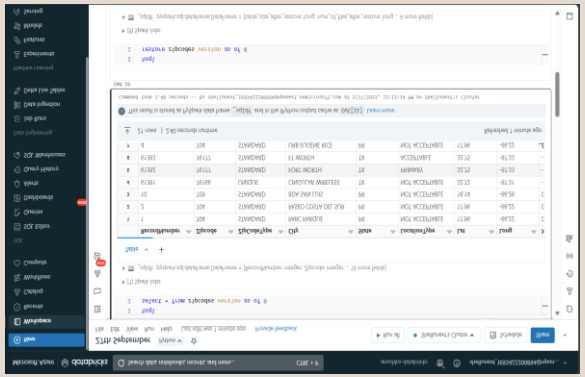

2023-09-27 - DBFS Example
Overview
This notebook will show you how to create and query a table or DataFrame that you uploaded to DBFS. DBFS is a Databricks File System that allows you to store data for querying inside of Databricks. This notebook assumes that you have a file already inside of DBFS that you would like to read from.
This notebook is written in Python so the default cell type is Python. However, you can use different languages by using the %LANGUAGE syntax. Python, Scala, SQL, and R are all supported.
# File location and type
file_location = "/FileStore/tables/zipcodes.csv"
file_type = "csv"
# CSV options
infer_schema = "false"
first_row_is_header = "false"
delimiter = ","
# The applied options are for CSV files. For other file types, these will be ignored.
df = spark.read.format(file_type) \
.option("inferSchema", infer_schema) \
.option("header", first_row_is_header) \
.option("sep", delimiter) \
.load(file_location)
display(df)

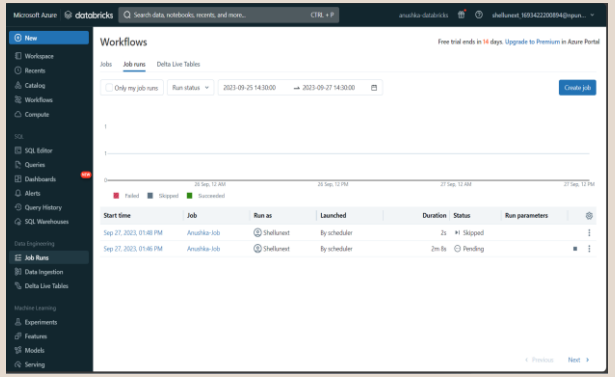

Catalog Explorer
default.zipcodes_1_csv
Column
Type
Comment
RecordNumber
Zipcode
ZipCodeType
City
State
LocationType
Lat
Long
Xaxis
Yaxis

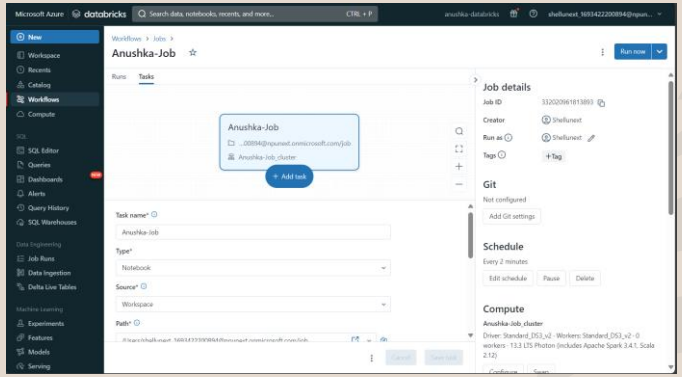

27th September
%sql
describe history zipcodes_1_csv
df=spark.read.option("inferSchema",True).option("header",True).csv("/mnt/input/zipcodes.csv")
df.display()


27th September
df.write.option("path","/mnt/input/tables").saveAsTable("zipcodes")
%sql
VACUUM zipcodes RETAIN 169 HOURS
%sql
OPTIMIZE zipcodes ZORDER BY (RecordNumber)


27th September
%sql
OPTIMIZE zipcodes ZORDER BY (RecordNumber)
df.write.partitionBy("State").option("path","/mnt/input/partition_tables").saveAsTable("partition_zipcodes")



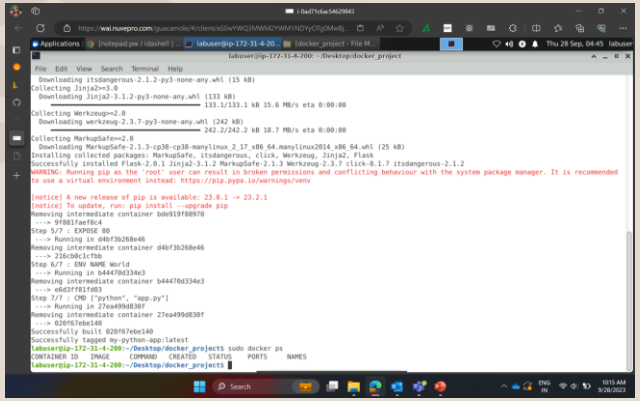

Workflows
Job runs
Delta Live Tables
Create job
Start time
Job
Run as
Launched
Duration
Status
Run parameters

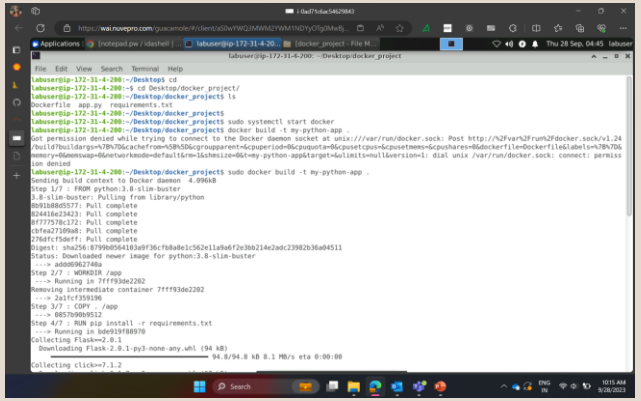

Anushka-Job
Runs
Tasks
Job details
Job ID
Creator
Run as
Tags
Git
Not configured
Schedule
Every 2 minutes
Compute

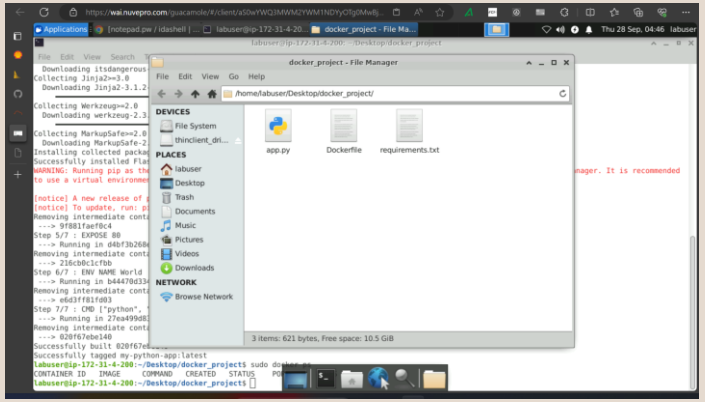
DAY-21

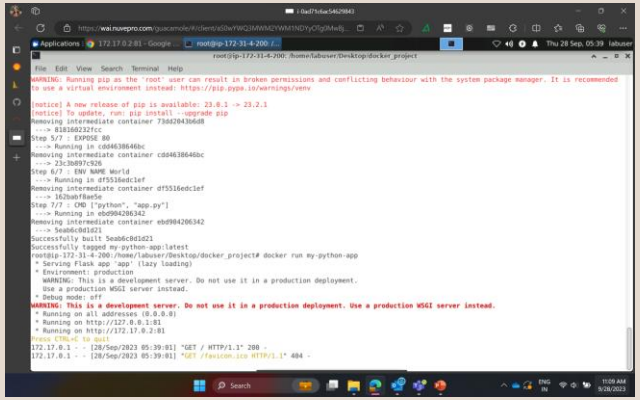
Docker Kubernetes

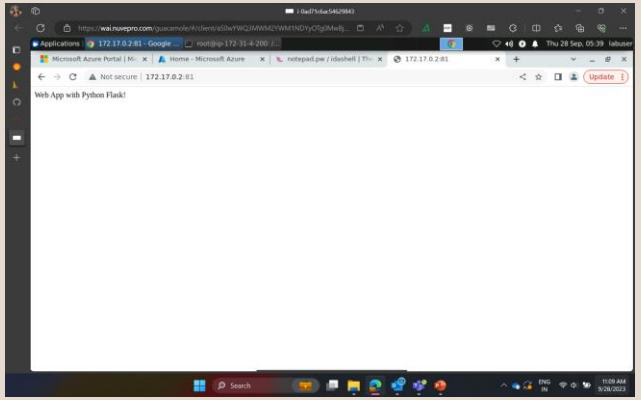
1. What is Docker ?
2. Docker initialization in VM
3. Build Docker image
   - Application
   - Requirement.txt
   - Dockerfile
4. Image created and Docker push
5. Azure container registry (ACR)
6. Kubernetes
7. AKS
8. Deploy a single-image application using code and UI

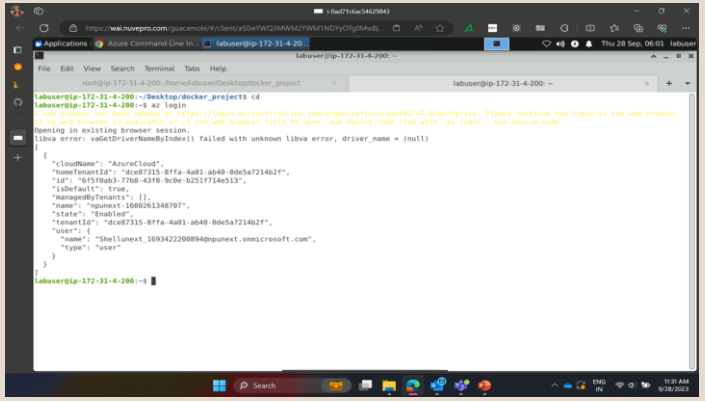


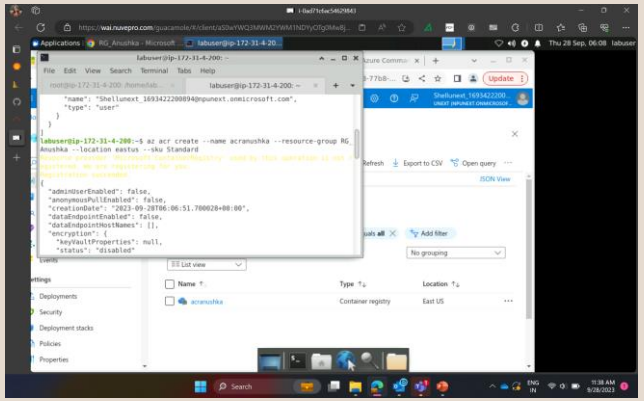


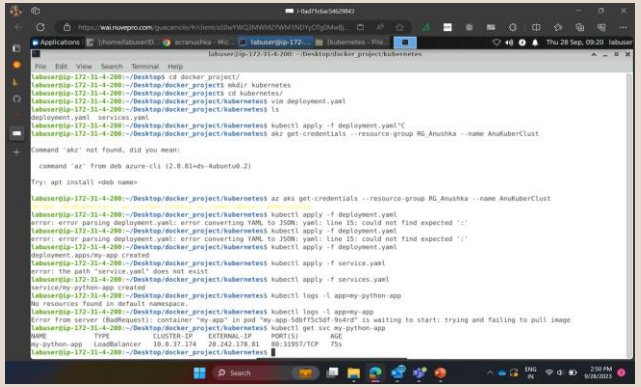


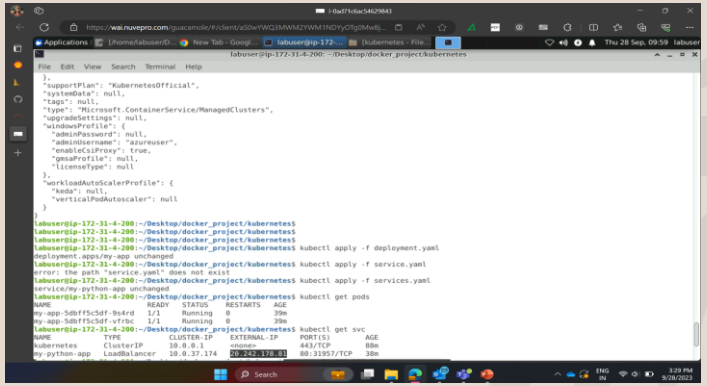


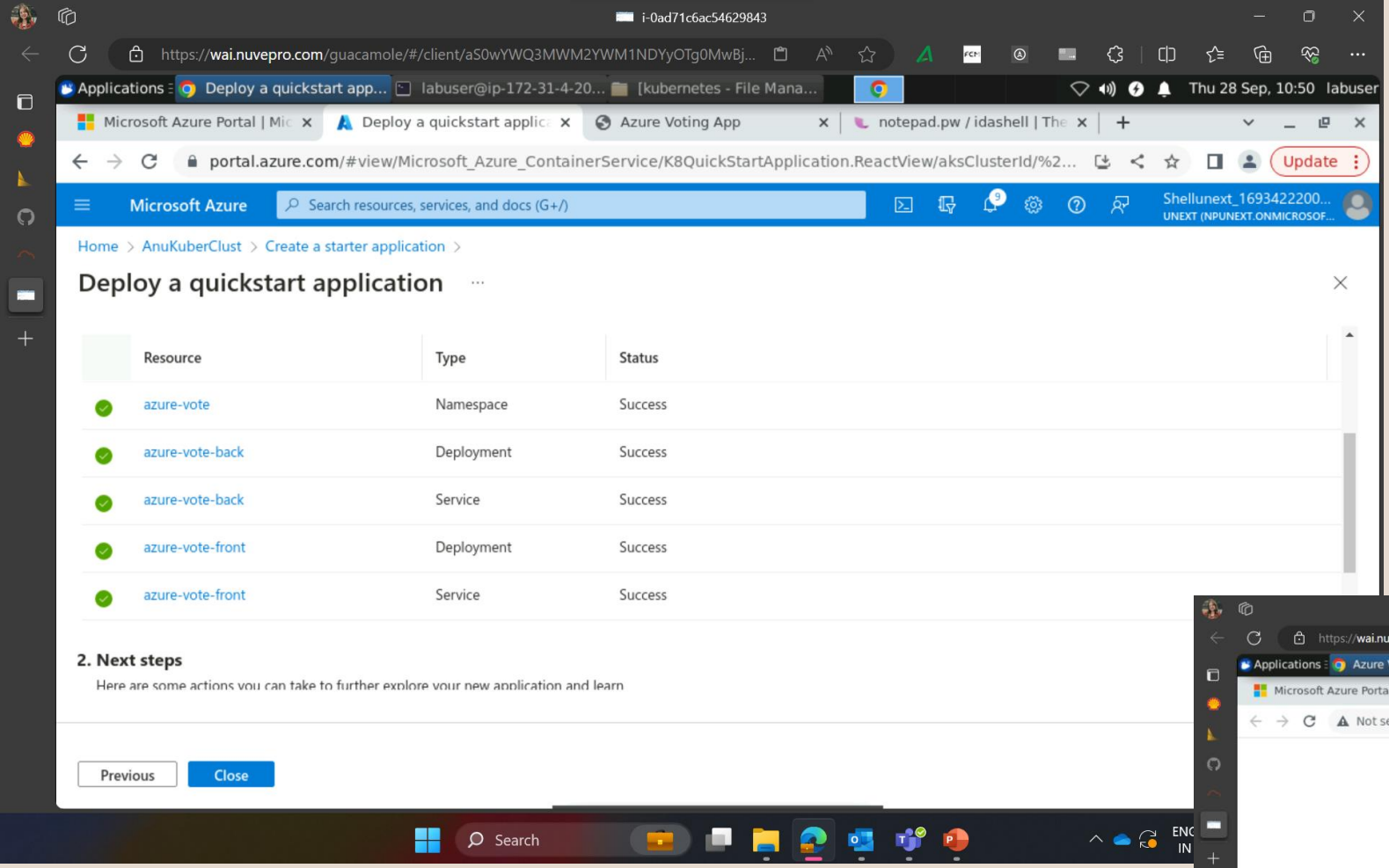


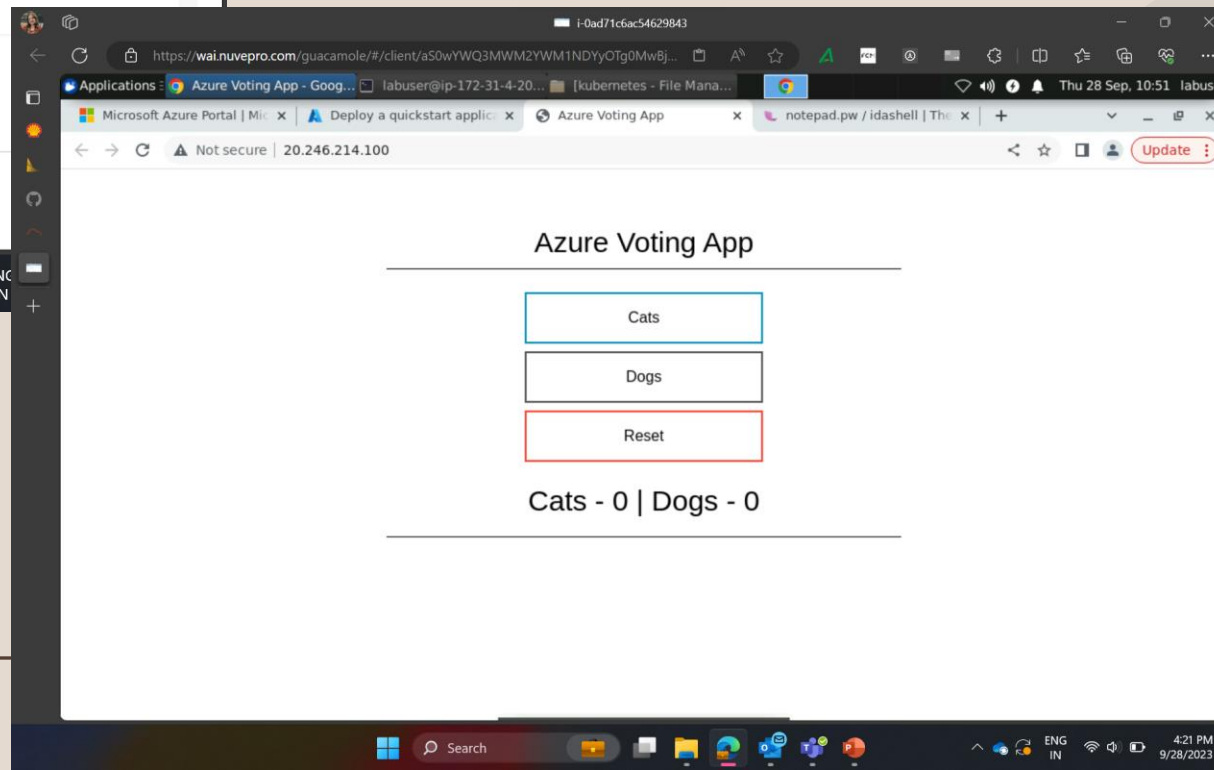








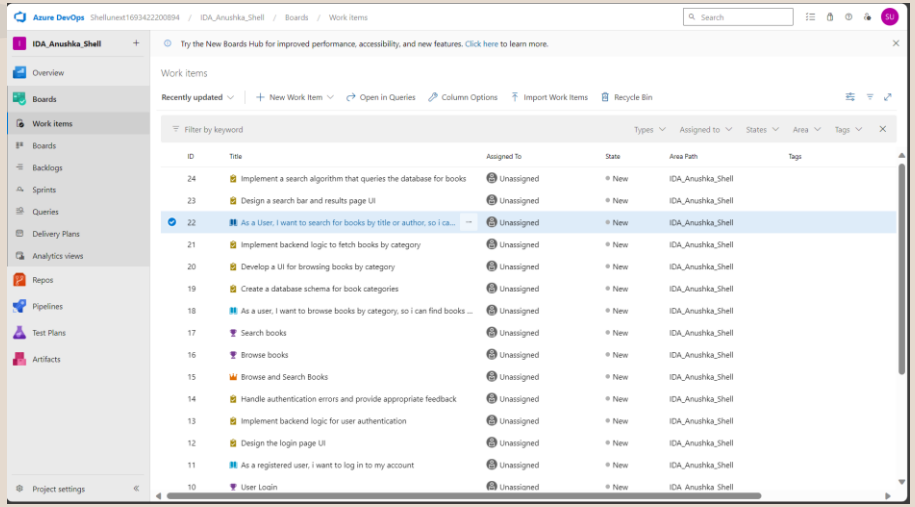

i-0ad71c6ac54629843
https://wai.nuvepro.com/guacamole/#/client/aS0wYWQ3MWM2YWM1NDYyOTg0MwBj...
Applications
Deploy a quickstart app...
labuser@ip-172-31-4-20...
[kubernetes - File Mana...
Thu 28 Sep, 10:50
labuser
Microsoft Azure Portal | Mic
Deploy a quickstart applic
Azure Voting App
notepad.pw / idashell | The
portal.azure.com/#view/Microsoft_Azure_ContainerService/K8QuickStartApplication.ReactView/aksClusterId/%2...
Update
Microsoft Azure
Search resources, services, and docs (G+/)
Shellunext_1693422200...
UNEXT (NPUNEXT.ONMICROSOF...
Home > AnuKuberClust > Create a starter application >
Deploy a quickstart application
Resource
Type
Status
azure-vote
Namespace
Success
azure-vote-back
Deployment
Success
azure-vote-back
Service
Success
azure-vote-front
Deployment
Success
azure-vote-front
Service
Success
2. Next steps
Here are some actions you can take to further explore your new application and learn
Previous
Close
Search
ENG
IN

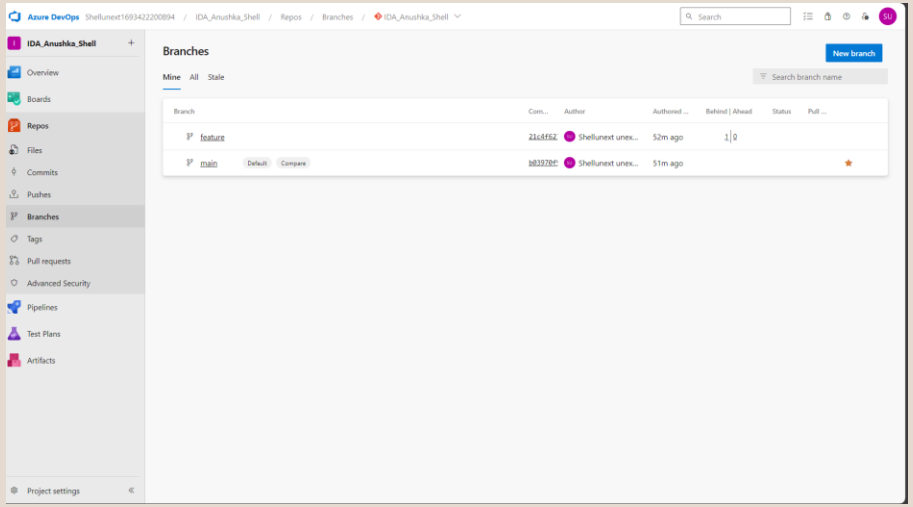

i-0ad71c6ac54629843
https://wai.nuvepro.com/guacamole/#/client/aS0wYWQ3MWM2YWM1NDYyOTg0MwBj...
Applications
Azure Voting App - Goog...
labuser@ip-172-31-4-20...
[kubernetes - File Mana...
Thu 28 Sep, 10:51
labus
Microsoft Azure Portal | Mic
Deploy a quickstart applic
Azure Voting App
notepad.pw / idashell | The
Not secure | 20.246.214.100
Update
Azure Voting App
Cats
Dogs
Reset
Cats - 0 | Dogs - 0
Search
ENG
IN
4:21 PM
9/28/2023

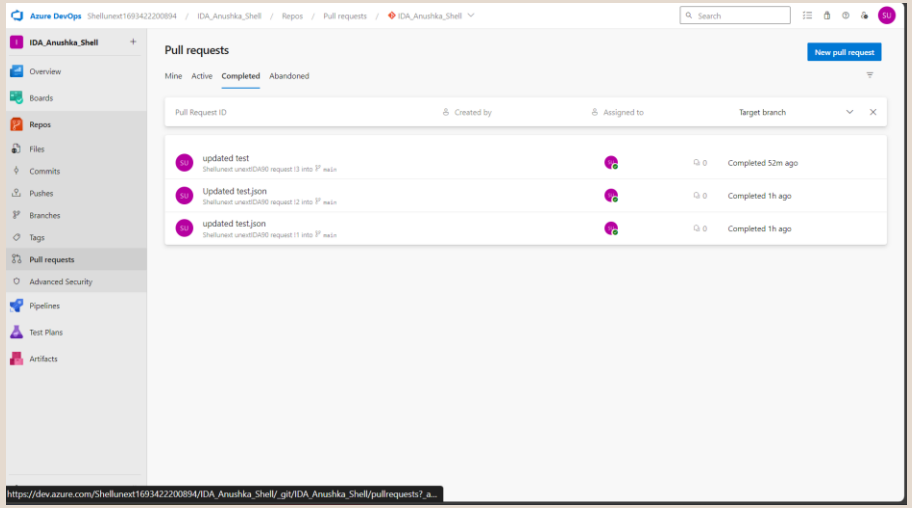
DAY-22

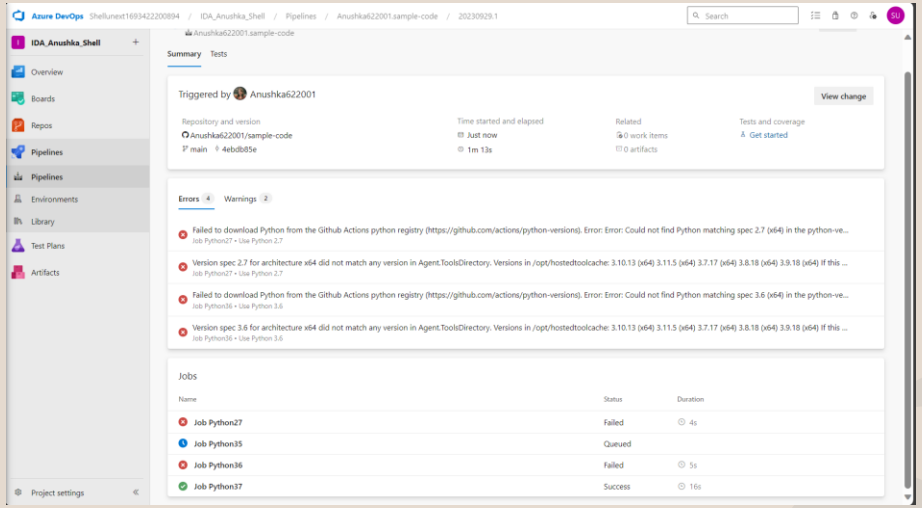
Azure Deployment / DevOps

1. What is DevOps ?
    1. Collaborative Branch (Main branch)
    2. Feature branch
2. Azure DevOps
    1. Boards
    2. Repos
    3. Pipelines
    4. Test plans
3. Creating a project Board in Azure DevOps
    - Work items
        - Epic
        - Feature
        - User story
        - Tasks
4. Azure Repos
5. Azure Pipelines


Azure DevOps
IDA_Anushka_Shell
Boards
Work items
Try the New Boards Hub for improved performance, accessibility, and new features. Click here to learn more.
Work items
New Work Item
Open in Queries
Column Options
Import Work Items
Recycle Bin
Implement a search algorithm that queries the database for books
Design a search bar and results page UI
As a User, I want to search for books by title or author, so i ca...
Implement backend logic to fetch books by category
Develop a UI for browsing books by category
Create a database schema for book categories
As a user, I want to browse books by category, so i can find books ...
Search books
Browse books
Browse and Search Books
Handle authentication errors and provide appropriate feedback
Implement backend logic for user authentication
Design the login page UI
As a registered user, I want to log in to my account
User Login


Azure DevOps
IDA_Anushka_Shell
Branches
New branch
Mine
All
Stale
feature
main


Azure DevOps
IDA_Anushka_Shell
Pull requests
New pull request
Mine
Active
Completed
Abandoned
updated test
Updated test.json
updated test.json
Completed 52m ago
Completed 1h ago
Completed 1h ago


Azure DevOps
IDA_Anushka_Shell
Summary
Tests
Triggered by Anushka622001
View change
Errors
Warnings
Jobs
Job Python27
Job Python35
Job Python36
Job Python37
Failed
Queued
Failed
Success

Thank You