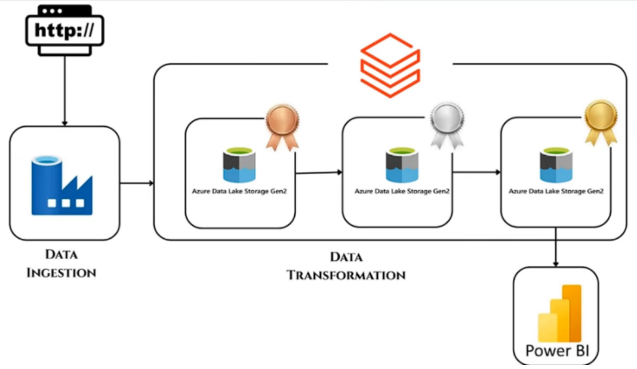
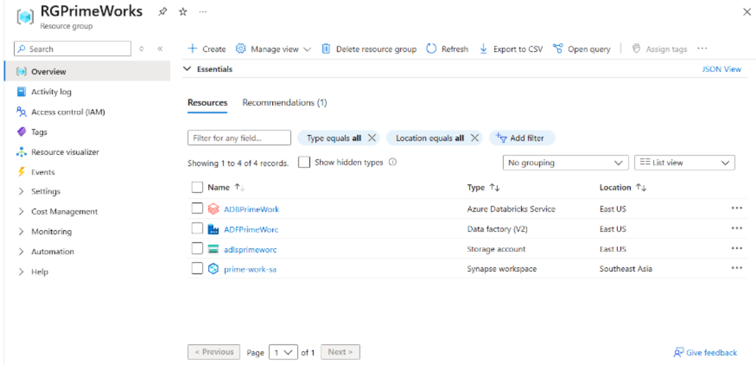
End-to-End Azure ETL Pipeline Project

Thilina Sasanka
UOP

1) ETL Pipeline Overview


http://
Azure Data Lake Storage Gen2
Azure Data Lake Storage Gen2
Azure Data Lake Storage Gen2
DATA INGESTION
DATA TRANSFORMATION
Power BI

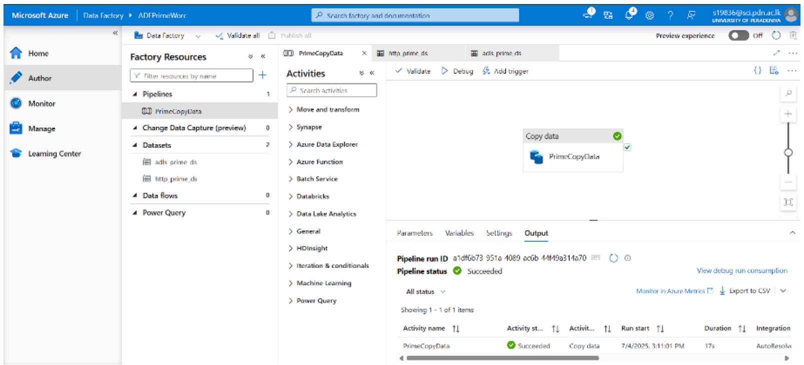
2) Resource Group Creation



Showing 1 to 4 of 4 records.

Name	Type	Location
ADBPrimeWork	Azure Databricks Service	East US
ADFPrimeWorc	Data factory (V2)	East US
adlsprimeworc	Storage account	East US
prime-work-sa	Synapse workspace	Southeast Asia

3) Azure Data Factory Creation



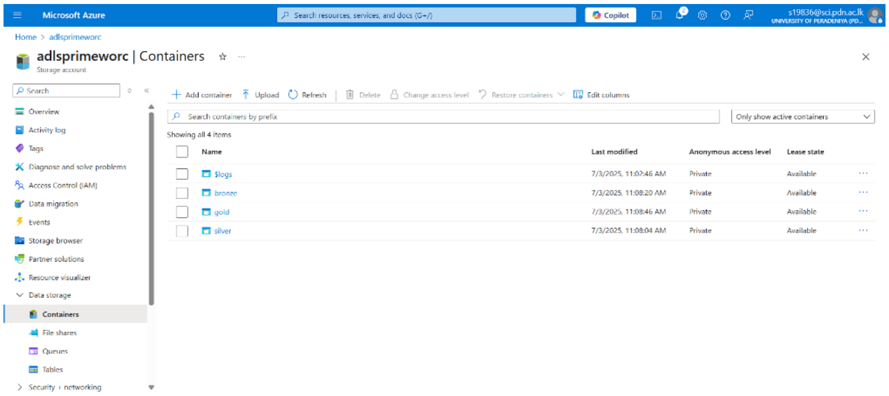
Pipeline run ID a1df6b73-951a-4089-ac6b-44f49a314a70
Pipeline status Succeeded

Showing 1 - 1 of 1 items

Activity name	Activity st...	Activit...	Run start	Duration	Integration
PrimeCopyData	Succeeded	Copy data	7/4/2025, 3:11:01 PM	17s	AutoResolv

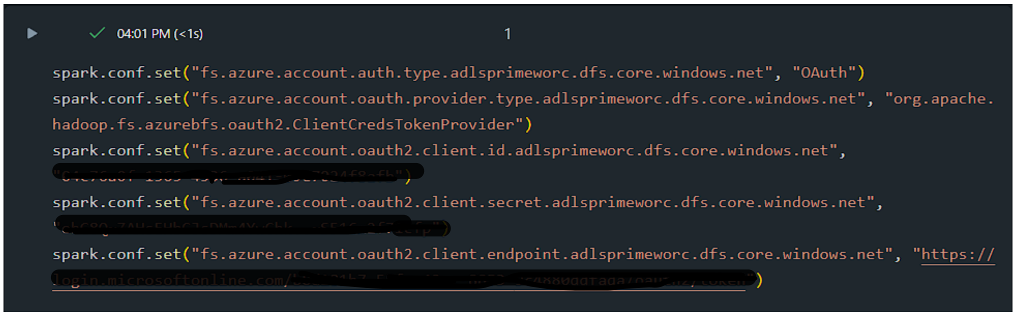
4) Azure Databricks Creation

5) Azure Databricks- Silver Layer - Cleaning & Transformations

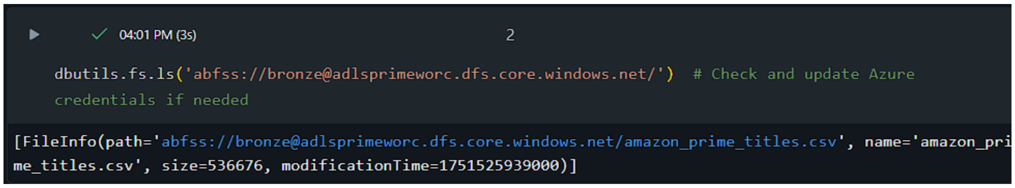


Showing all 4 items

Name	Last modified	Anonymous access level	Lease state
$logs	7/3/2025, 11:02:46 AM	Private	Available
bronze	7/3/2025, 11:08:20 AM	Private	Available
gold	7/3/2025, 11:08:46 AM	Private	Available
silver	7/3/2025, 11:08:04 AM	Private	Available

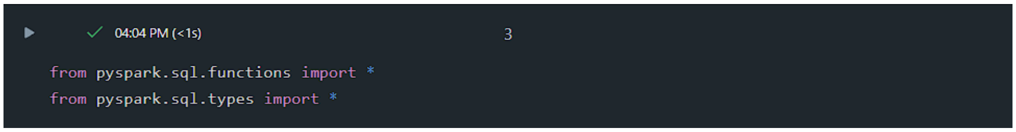

04:01 PM (<1s) 1

```
spark.conf.set("fs.azure.account.auth.type.adlsprimeworc.dfs.core.windows.net", "OAuth")
spark.conf.set("fs.azure.account.oauth.provider.type.adlsprimeworc.dfs.core.windows.net", "org.apache.
hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider")
spark.conf.set("fs.azure.account.oauth2.client.id.adlsprimeworc.dfs.core.windows.net",
[redacted]
spark.conf.set("fs.azure.account.oauth2.client.secret.adlsprimeworc.dfs.core.windows.net",
[redacted]
spark.conf.set("fs.azure.account.oauth2.client.endpoint.adlsprimeworc.dfs.core.windows.net", "https://
[redacted])
```

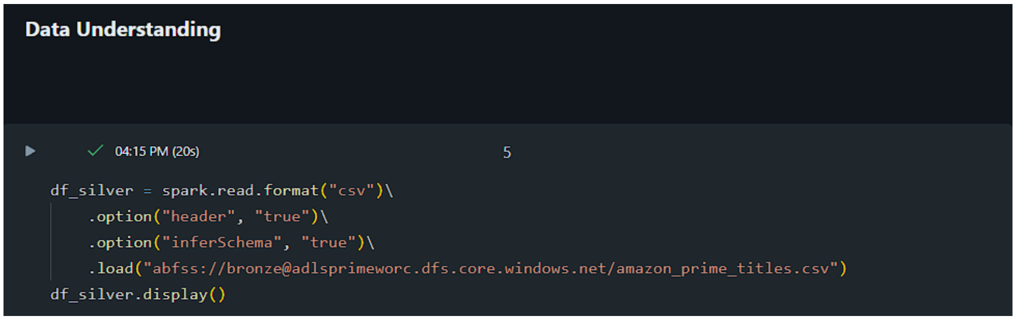

04:01 PM (3s) 2

```
dbutils.fs.ls('abfss://bronze@adlsprimeworc.dfs.core.windows.net/')  # Check and update Azure
credentials if needed
```

```
[FileInfo(path='abfss://bronze@adlsprimeworc.dfs.core.windows.net/amazon_prime_titles.csv', name='amazon_pri
me_titles.csv', size=536676, modificationTime=1751525939000)]
```

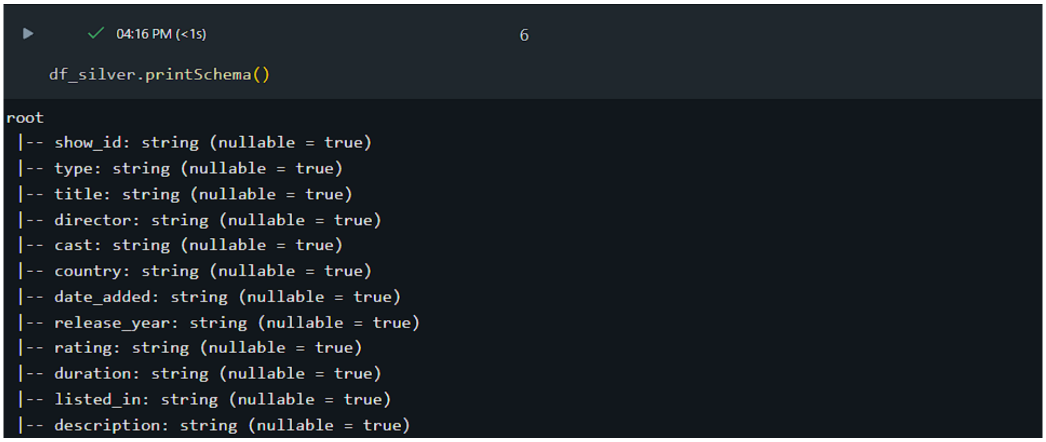

04:04 PM (<1s) 3

```
from pyspark.sql.functions import *
from pyspark.sql.types import *
```


Data Understanding

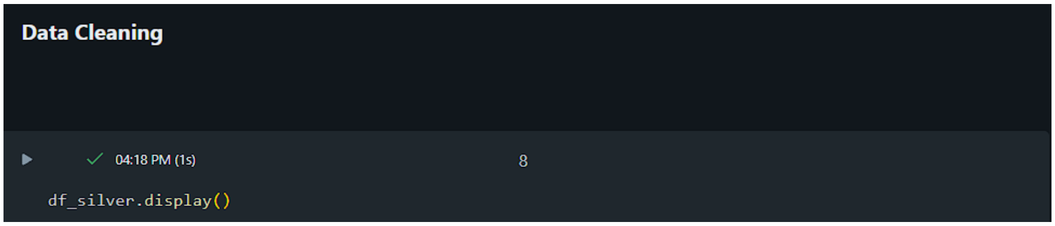
04:15 PM (20s) 5

```
df_silver = spark.read.format("csv")\
    .option("header", "true")\
    .option("inferSchema", "true")\
    .load("abfss://bronze@adlsprimeworc.dfs.core.windows.net/amazon_prime_titles.csv")
df_silver.display()
```



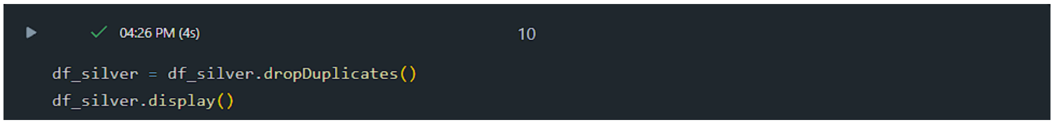
```
df_silver.printSchema()
```

```
root
 |-- show_id: string (nullable = true)
 |-- type: string (nullable = true)
 |-- title: string (nullable = true)
 |-- director: string (nullable = true)
 |-- cast: string (nullable = true)
 |-- country: string (nullable = true)
 |-- date_added: string (nullable = true)
 |-- release_year: string (nullable = true)
 |-- rating: string (nullable = true)
 |-- duration: string (nullable = true)
 |-- listed_in: string (nullable = true)
 |-- description: string (nullable = true)
```

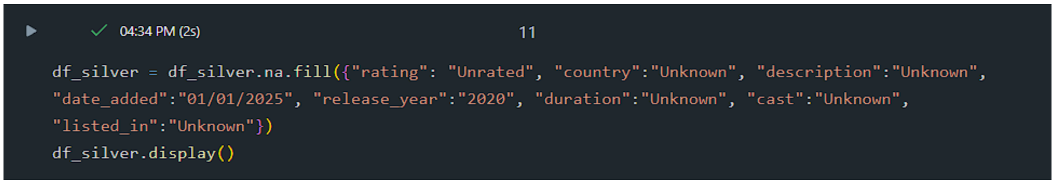


Data Cleaning

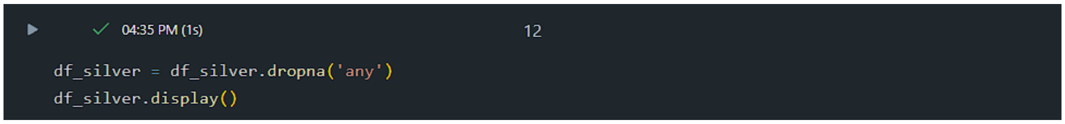
```
df_silver.display()
```



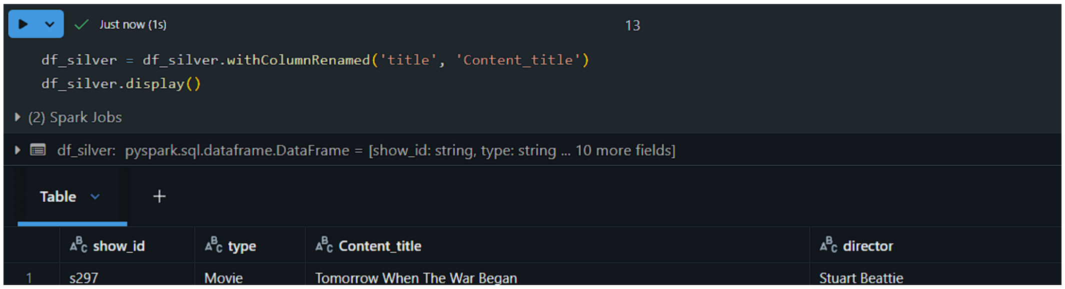
```
df_silver = df_silver.dropDuplicates()
df_silver.display()
```



```
df_silver = df_silver.na.fill({"rating": "Unrated", "country":"Unknown", "description":"Unknown",
"date_added":"01/01/2025", "release_year":"2020", "duration":"Unknown", "cast":"Unknown",
"listed_in":"Unknown"})
df_silver.display()
```



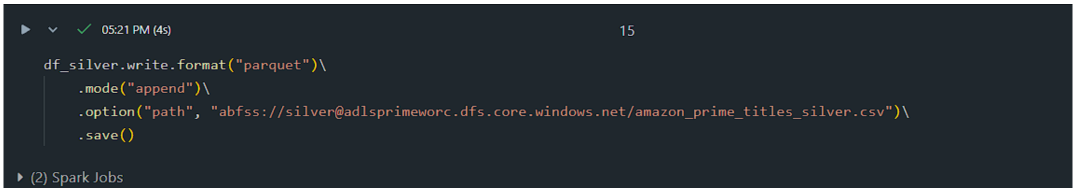
```
df_silver = df_silver.dropna('any')
df_silver.display()
```



```
df_silver = df_silver.withColumnRenamed('title', 'Content_title')
df_silver.display()
```

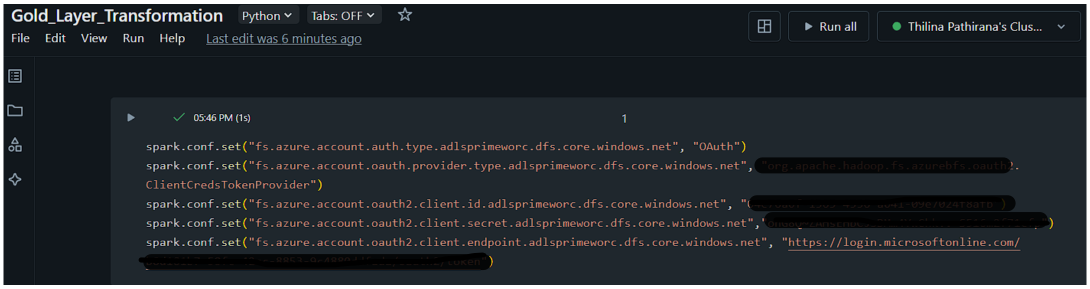
df_silver: pyspark.sql.dataframe.DataFrame = [show_id: string, type: string ... 10 more fields]

	show_id	type	Content_title	director
1	s297	Movie	Tomorrow When The War Began	Stuart Beattie



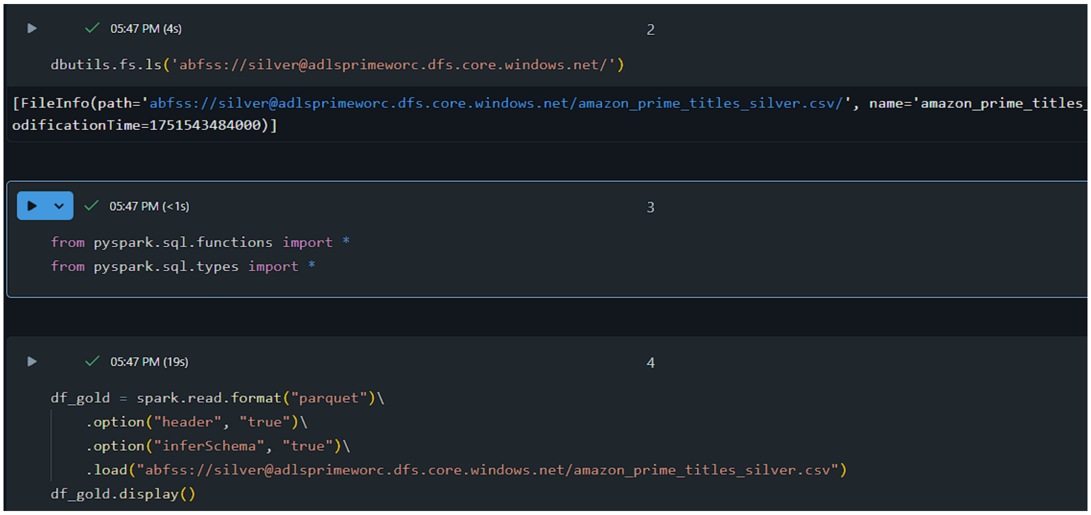
```
df_silver.write.format("parquet")\
    .mode("append")\
    .option("path", "abfss://silver@adlsprimeworc.dfs.core.windows.net/amazon_prime_titles_silver.csv")\
    .save()
```

6) Azure Databricks - Gold Layer – Transformations



Gold_Layer_Transformation
Python
Tabs: OFF
File
Edit
View
Run
Help
Last edit was 6 minutes ago
Run all
Thilina Pathirana's Clus...
05:46 PM (1s)
1

```
spark.conf.set("fs.azure.account.auth.type.adlsprimeworc.dfs.core.windows.net", "OAuth")
spark.conf.set("fs.azure.account.oauth.provider.type.adlsprimeworc.dfs.core.windows.net", [redacted].
ClientCredsTokenProvider")
spark.conf.set("fs.azure.account.oauth2.client.id.adlsprimeworc.dfs.core.windows.net", "[redacted]
spark.conf.set("fs.azure.account.oauth2.client.secret.adlsprimeworc.dfs.core.windows.net",[redacted])
spark.conf.set("fs.azure.account.oauth2.client.endpoint.adlsprimeworc.dfs.core.windows.net", "https://login.microsoftonline.com/
[redacted])
```

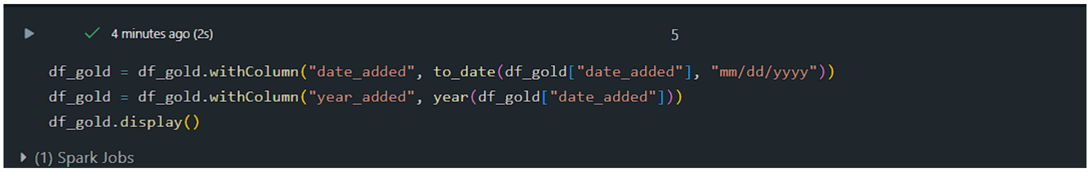


```
dbutils.fs.ls('abfss://silver@adlsprimeworc.dfs.core.windows.net/')
```

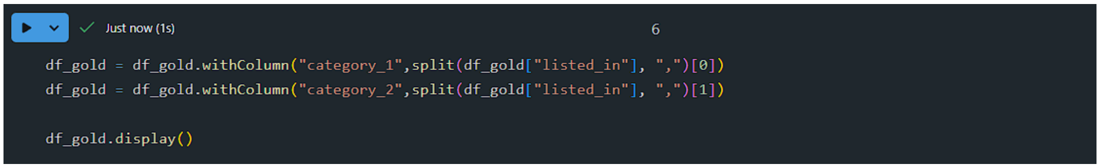
```
[FileInfo(path='abfss://silver@adlsprimeworc.dfs.core.windows.net/amazon_prime_titles_silver.csv/', name='amazon_prime_titles_
odificationTime=1751543484000)]
```

```
from pyspark.sql.functions import *
from pyspark.sql.types import *
```

```
df_gold = spark.read.format("parquet")\
    .option("header", "true")\
    .option("inferSchema", "true")\
    .load("abfss://silver@adlsprimeworc.dfs.core.windows.net/amazon_prime_titles_silver.csv")
df_gold.display()
```

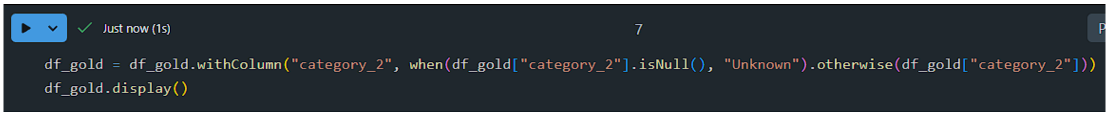


```
df_gold = df_gold.withColumn("date_added", to_date(df_gold["date_added"], "mm/dd/yyyy"))
df_gold = df_gold.withColumn("year_added", year(df_gold["date_added"]))
df_gold.display()
```

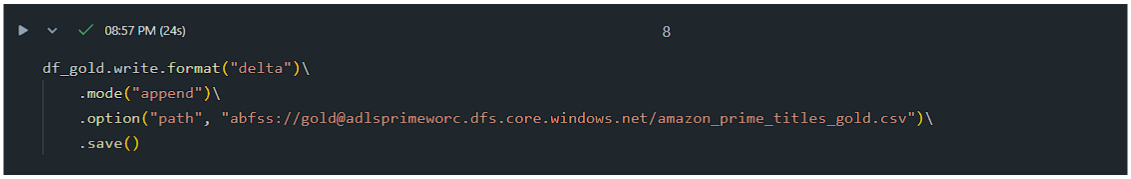


```
df_gold = df_gold.withColumn("category_1",split(df_gold["listed_in"], ",")[0])
df_gold = df_gold.withColumn("category_2",split(df_gold["listed_in"], ",")[1])

df_gold.display()
```

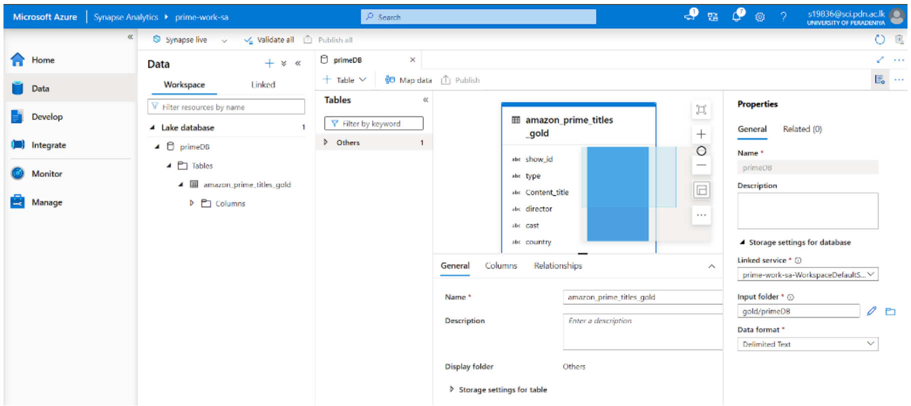


```
df_gold = df_gold.withColumn("category_2", when(df_gold["category_2"].isNull(), "Unknown").otherwise(df_gold["category_2"]))
df_gold.display()
```



```
df_gold.write.format("delta")\
    .mode("append")\
    .option("path", "abfss://gold@adlsprimeworc.dfs.core.windows.net/amazon_prime_titles_gold.csv")\
    .save()
```

7) Load data to the Synapse Analytics



Name * amazon_prime_titles_gold

Description Enter a description

Display folder Others

Storage settings for table

Name * primeDB

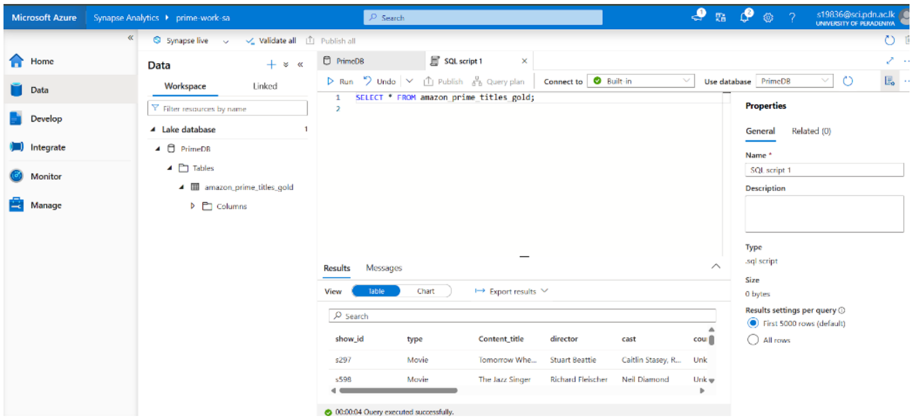
Description

Storage settings for database

Linked service * prime-work-sa-WorkspaceDefaultS...

Input folder * gold/primeDB

Data format * Delimited Text



Connect to Built-in

Use database PrimeDB

```
SELECT * FROM amazon_prime_titles_gold;
```

show_id	type	Content_title	director	cast	cou
s297	Movie	Tomorrow Whe...	Stuart Beattie	Caitlin Stasey, R...	Unk
s598	Movie	The Jazz Singer	Richard Fleischer	Neil Diamond	Unk

00:00:04 Query executed successfully.

Name * SQL script 1

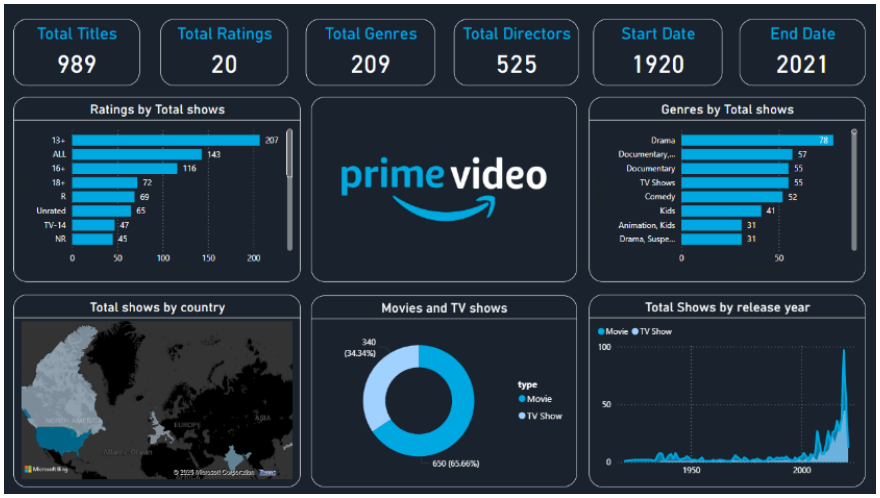
Description

Type
.sql script

Size
0 bytes

Results settings per query
● First 5000 rows (default)
○ All rows

8) Connecting to Power BI


Total Titles
Total Ratings
Total Genres
Total Directors
Start Date
End Date
Ratings by Total shows
ALL 143
R 69
Unrated 65
TV-14 47
NR 45
prime video
Genres by Total shows
Drama 78
Documentary,... 57
Documentary 55
TV Shows 55
Comedy 52
Kids 41
Animation, Kids 31
Drama, Suspe... 31
Total shows by country
NORTH AMERICA
EUROPE
ASIA
Atlantic Ocean
Microsoft Bing
© 2025 Microsoft Corporation Terms
Movies and TV shows
340 (34.34%)
650 (65.66%)
type
Movie
TV Show
Total Shows by release year
Movie
TV Show