(https://databricks.com)

Video Game Sales Data Pipeline

In this notebook, we will create a data pipeline to store, transfer, query, and visualize video game sales data. The pipeline will consist of the following stages:

1. Source data - storage and ingest
2. ETL transformations
3. Storage of data for analytics
4. Query and visualization

Source Data: Storage and Ingest

In this step, we will read the CSV file containing video game sales data and load it into a DataFrame using the PySpark library.

```
from pyspark.sql import SparkSession

spark = SparkSession.builder.master("local").appName("Video Game Sales").getOrCreate()
data = spark.read.csv("/FileStore/tables/Video_Games_Sales_as_at_22_Dec_2016-1.csv", header=True,
inferSchema=True)
data.show()
```

```
+--------------------+--------+--------------+-----------+--------------------+--------+--------
+--------+-----------+-----------+-----------+-----------+----------+----------+--------------
----+------+
|                Name|Platform|Year_of_Release|       Genre|           Publisher|NA_Sales|EU_Sales
|JP_Sales|Other_Sales|Global_Sales|Critic_Score|Critic_Count|User_Score|User_Count|          Devel
oper|Rating|
+--------------------+--------+--------------+-----------+--------------------+--------+--------
+--------+-----------+-----------+-----------+-----------+----------+----------+--------------
----+------+
|          Wii Sports|     Wii|          2006|      Sports|            Nintendo|   41.36|   28.96
|    3.77|       8.45|       82.53|          76|          51|         8|       322|           Nint
endo|     E|
|   Super Mario Bros.|     NES|          1985|    Platform|            Nintendo|   29.08|    3.58
|    6.81|       0.77|       40.24|        null|        null|      null|      null|
null|  null|
|      Mario Kart Wii|     Wii|          2008|      Racing|            Nintendo|   15.68|   12.76
```

```
|    3.79|       3.29|       35.52|          82|          73|       8.3|       709|           Nint
endo|     E|
|   Wii Sports Resort|     Wii|          2009|      Sports|           Nintendo|   15.61|   10.93
|    3.28|       2.95|       32.77|          80|          73|         8|       192|           Nint
endo|     E|
```

ETL Transformations

In this step, we will perform two ETL transformations on the dataset:

1. Filter out rows with missing values.
2. Add a new column indicating the average sales across all regions.

```
data_cleaned = data.dropna()
data_cleaned.show()
```

```
+--------------------+--------+--------------+--------+--------------------+--------+--------+---
-----+-----------+------------+------------+------------+----------+----------+------------------+
-+------+
|                Name|Platform|Year_of_Release|   Genre|           Publisher|NA_Sales|EU_Sales|JP_
Sales|Other_Sales|Global_Sales|Critic_Score|Critic_Count|User_Score|User_Count|           Develope
r|Rating|
+--------------------+--------+--------------+--------+--------------------+--------+--------+---
-----+-----------+------------+------------+------------+----------+----------+------------------
-+------+
|          Wii Sports|     Wii|          2006|  Sports|            Nintendo|   41.36|   28.96|
3.77|       8.45|       82.53|          76|          51|         8|       322|            Nintendo
|     E|
|      Mario Kart Wii|     Wii|          2008|  Racing|            Nintendo|   15.68|   12.76|
3.79|       3.29|       35.52|          82|          73|       8.3|       709|            Nintendo
|     E|
|   Wii Sports Resort|     Wii|          2009|  Sports|            Nintendo|   15.61|   10.93|
3.28|       2.95|       32.77|          80|          73|         8|       192|            Nintendo
|     E|
|New Super Mario B...|      DS|          2006|Platform|            Nintendo|   11.28|    9.14|
6.5|       2.88|        29.8|          89|          65|       8.5|       431|            Nintendo|
E|
```

```
from pyspark.sql.functions import col

data_cleaned = data_cleaned.withColumn("Average_Sales", (col("NA_Sales") + col("EU_Sales") +
col("JP_Sales") + col("Other_Sales")) / 4)
data_cleaned.show()
```

```
+--------------------+--------+--------------+--------+--------------------+--------+--------+---
-----+-----------+------------+------------+------------+----------+----------+------------------
-+------+------------------+
|                Name|Platform|Year_of_Release|   Genre|           Publisher|NA_Sales|EU_Sales|JP_
Sales|Other_Sales|Global_Sales|Critic_Score|Critic_Count|User_Score|User_Count|           Develope
r|Rating|     Average_Sales|
+--------------------+--------+--------------+--------+--------------------+--------+--------+---
```

```
-----+-----------+------------+------------+------------+----------+----------+------------------
-+------+-----------------+
|          Wii Sports|     Wii|           2006|  Sports|            Nintendo|   41.36|   28.96|
3.77|       8.45|       82.53|          76|           51|         8|       322|            Nintendo
|     E|20.634999999999998|
|      Mario Kart Wii|     Wii|           2008|  Racing|            Nintendo|   15.68|   12.76|
3.79|       3.29|       35.52|          82|           73|       8.3|       709|            Nintendo
|     E| 8.879999999999999|
|   Wii Sports Resort|     Wii|           2009|  Sports|            Nintendo|   15.61|   10.93|
3.28|       2.95|       32.77|          80|           73|         8|       192|            Nintendo
|     E|            8.1925|
|New Super Mario B...|      DS|           2006|Platform|            Nintendo|   11.28|    9.14|
6.5|       2.88|        29.8|          89|           65|       8.5|       431|            Nintendo|
```

Storage of Data for Analytics

After cleaning and transforming the data, we will save the resulting DataFrame as a Parquet file for efficient storage and analytics.

```
data_cleaned.write.parquet("/FileStore/tables/parquetfile.parquet")
```

Query and Visualization

In this step, we will perform two queries on the dataset and visualize the results:

1. Find the top 10 best-selling games globally.
2. Find the total global sales by genre.

```
top10_games = data_cleaned.select("Name",
"Global_Sales").orderBy(col("Global_Sales").desc()).limit(10)
top10_games.show()
```

```
+--------------------+------------+
|                Name|Global_Sales|
+--------------------+------------+
|          Wii Sports|       82.53|
|      Mario Kart Wii|       35.52|
|   Wii Sports Resort|       32.77|
|New Super Mario B...|        29.8|
|            Wii Play|       28.92|
|New Super Mario B...|       28.32|
|       Mario Kart DS|       23.21|
|             Wii Fit|        22.7|
|  Kinect Adventures!|       21.81|
|        Wii Fit Plus|       21.79|
+--------------------+------------+
```

```
from pyspark.sql.functions import sum as _sum

genre_sales =
data_cleaned.groupBy("Genre").agg(_sum("Global_Sales").alias("Total_Global_Sales")).orderBy(col("T
otal_Global_Sales").desc())
genre_sales.show()
```

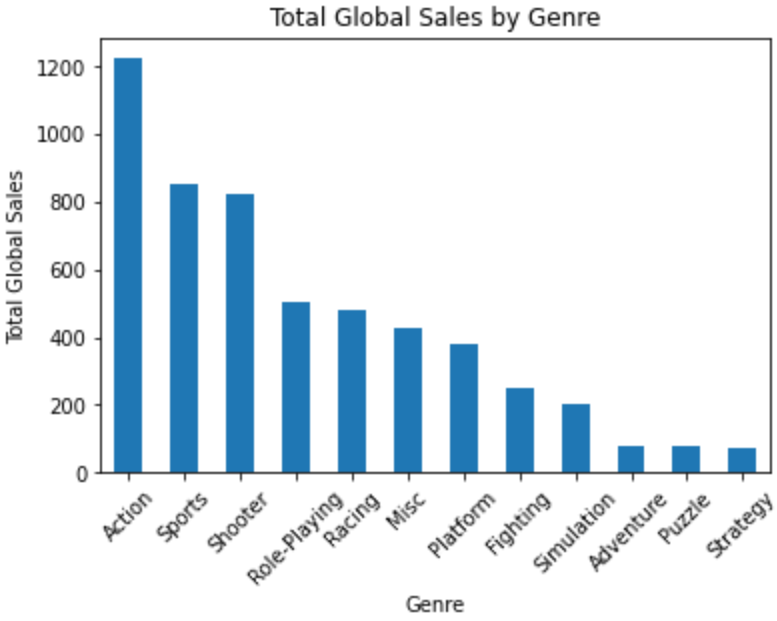
```
+------------+------------------+
|       Genre|Total_Global_Sales|
+------------+------------------+
|      Action|1224.1199999999942|
|      Sports| 850.6599999999987|
|     Shooter| 823.8099999999982|
|Role-Playing| 503.3899999999988|
|      Racing|479.79999999999944|
|        Misc| 424.6299999999999|
|    Platform|378.63999999999936|
|    Fighting|250.95000000000022|
|  Simulation|            203.52|
|   Adventure| 80.90999999999994|
|      Puzzle| 79.27000000000002|
|    Strategy| 71.00999999999993|
+------------+------------------+
```

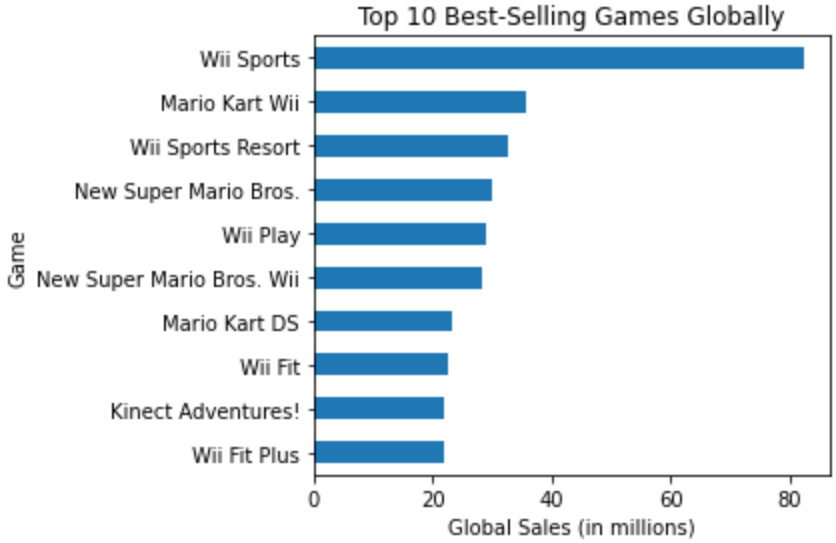
Visualization

In this step, we will use the Matplotlib library to create a bar chart that visualizes the total global sales by genre and top 10 best-selling games globally .

```
import pandas as pd
import matplotlib.pyplot as plt

genre_sales_pd = genre_sales.toPandas()
genre_sales_pd.plot.bar(x="Genre", y="Total_Global_Sales", legend=False)
plt.title("Total Global Sales by Genre")
plt.xlabel("Genre")
plt.ylabel("Total Global Sales")
plt.xticks(rotation=45)
plt.show()
```


Total Global Sales by Genre
Total Global Sales
0
200
400
600
800
1000
1200
Action
Sports
Shooter
Role-Playing
Racing
Misc
Platform
Fighting
Simulation
Adventure
Puzzle
Strategy
Genre


Top 10 Best-Selling Games Globally
Game
Wii Sports
Mario Kart Wii
Wii Sports Resort
New Super Mario Bros.
Wii Play
New Super Mario Bros. Wii
Mario Kart DS
Wii Fit
Kinect Adventures!
Wii Fit Plus
0
20
40
60
80
Global Sales (in millions)