PySpark Learning Summary: Healthcare Dataset

1. Installing PySpark

!pip install pyspark
- Installs Apache Spark Python API (PySpark) in Colab.

2. Creating a Spark Session

from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("Healthcare Data Ingestion").getOrCreate()
- Initializes a Spark session for processing data.

3. Uploading Files in Google Colab

from Google.Colab import files
uploaded = files.upload()
- Enables browser-based file upload in Colab.

4. Checking Files in the Directory

import os
os.listdir()
- Lists all files in the current working directory.

5. Reading the CSV File Using PySpark

df = spark.read.csv("healthcare_dataset.csv", header=True, inferSchema=True)
df.show(5)
- Reads a CSV into a PySpark DataFrame and shows the top 5 rows.

6. Transforming Columns

df_transformed = df.selectExpr("Age as Patient_Age", "Gender", "`Medical Condition` as Disease")
df_transformed.show()

- Renames and selects relevant columns using SQL-style expressions.

7. Filtering Data

```
df_filtered = df_transformed.filter("Patient_Age > 40")
df_filtered.show(10)
```

- Filters patients aged over 40.

8. Final Output Example

```
+-----------+--------+---------+
|Patient_Age|Gender  |Disease  |
+-----------+--------+---------+
|62         |Male    |Obesity  |
|76         |Female  |Obesity  |
|43         |Female  |Cancer   |
|82         |Male    |Asthma   |
+-----------+--------+---------+
```

Summary Table

Concept	Description
SparkSession	Used to initialize and configure the PySpark session
File Upload	Used google. Colab.files to upload CSV files
CSV Reading	Used .read.csv() with schema inference
Column Transformation	Renamed columns using selectExpr()
Filtering	Filtered data using filter()
DataFrame Ops	Explored with show(), selectExpr(), and filter()