EGT 305: BIG DATA PROCESSING & APPLICATION

BOSS

WHAT IS APACHE SPARK

- Apache Spark is an open-source distributed computing system designed for big data processing and analytics.
- Unified analytics engine for large-scale data processing with speed, ease of use, and versatility.
- Spark supports various programming languages including Scala, Java, Python, and R, making it accessible to a wide range of developers.
- Run queries and machine learning workflows on petabytes of data, which is impossible to do on your local device.

Source: https://www.oracle.com/sg/big-data/what-is-big-data/#:~:text=What%20exactly%20is%20big%20data,especially%20from%20new%20data%20sources

PRESENTATION TITLE

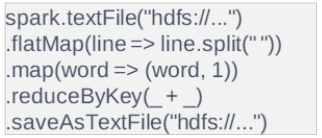
HOW APACHE SPARK WORKS? – FROM RDD TO DATAFRAME

Resilient Distributed Datasets (RDDs) - low level data representation in Spark

DataFrame API was introduced in 2015 – RDD + schema + advance interfaces

✓ Incredibly powerful API:

- **Hide low level distributed operations.** Map-reduce tasks that used to take thousands of lines of code to express could be reduced to dozens.



```
spark.textFile("hdfs://...")
.flatMap(line => line.split(" "))
.map(word => (word, 1))
.reduceByKey(_ + _)
.saveAsTextFile("hdfs://...")
```

VS

Written in Java for MapReduce it has around 50 lines of code

- Conceptually equivalent to a table in a relational database or a data frame in R/Python, but with richer **optimizations** under the hood.
- Supporting multiple general-purpose programming languages (Java, Python, Scala).

✓ Ability to scale

No need to change code: from kilobytes of data on a single laptop to petabytes on a large cluster**.**

✓ Support not only batch data, but also **streaming data**

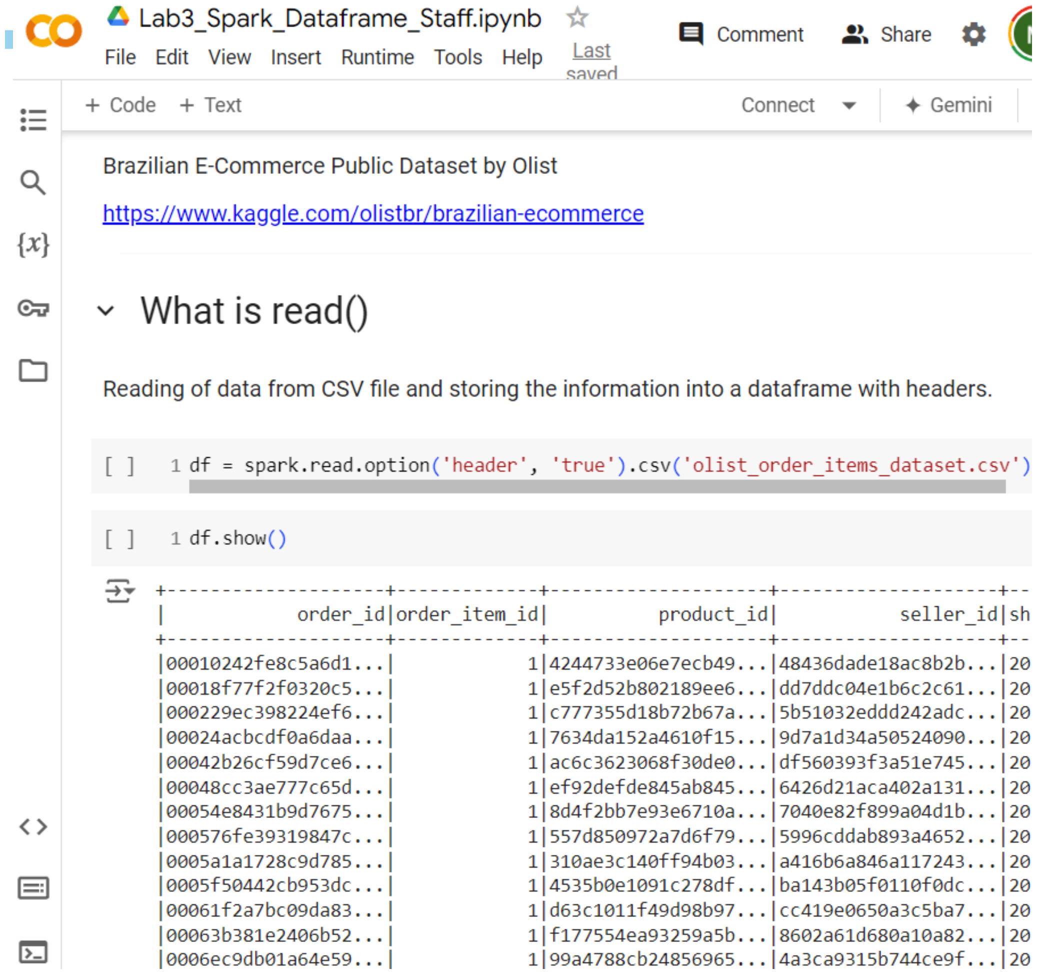
- To write streaming jobs the same way you write batch jobs.
- Join streams against historical data, or run ad-hoc queries on stream state.
- Build powerful interactive applications, not just analytics.

LAB 3:

DATAFRAME

- SPARK DATAFRAME FUNDAMENTALS:
  - read()
  - explain()
  - printSchema()
  - inferSchema()
  - DoubleData Type
  - Other Datattypes
  - Data Transformation
  - Data ETL using PySpark

LAB 3


CO
Lab3_Spark_Dataframe_Staff.ipynb
File Edit View Insert Runtime Tools Help
Last saved
Comment
Share
+ Code
+ Text
Connect
Gemini

Brazilian E-Commerce Public Dataset by Olist

https://www.kaggle.com/olistbr/brazilian-ecommerce

What is read()

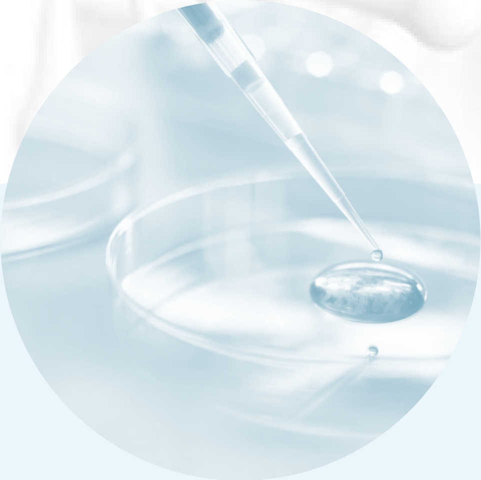
Reading of data from CSV file and storing the information into a dataframe with headers.

```
df = spark.read.option('header', 'true').csv('olist_order_items_dataset.csv')
```

```
df.show()
```

```
+--------------------+-------------+--------------------+--------------------+--
|            order_id|order_item_id|          product_id|           seller_id|sh
+--------------------+-------------+--------------------+--------------------+--
|00010242fe8c5a6d1...|            1|4244733e06e7ecb49...|48436dade18ac8b2b...|20
|00018f77f2f0320c5...|            1|e5f2d52b802189ee6...|dd7ddc04e1b6c2c61...|20
|000229ec398224ef6...|            1|c777355d18b72b67a...|5b51032eddd242adc...|20
|00024acbcdf0a6daa...|            1|7634da152a4610f15...|9d7a1d34a50524090...|20
|00042b26cf59d7ce6...|            1|ac6c3623068f30de0...|df560393f3a51e745...|20
|00048cc3ae777c65d...|            1|ef92defde845ab845...|6426d21aca402a131...|20
|00054e8431b9d7675...|            1|8d4f2bb7e93e6710a...|7040e82f899a04d1b...|20
|000576fe39319847c...|            1|557d850972a7d6f79...|5996cddab893a4652...|20
|0005a1a1728c9d785...|            1|310ae3c140ff94b03...|a416b6a846a117243...|20
|0005f50442cb953dc...|            1|4535b0e1091c278df...|ba143b05f0110f0dc...|20
|00061f2a7bc09da83...|            1|d63c1011f49d98b97...|cc419e0650a3c5ba7...|20
|00063b381e2406b52...|            1|f177554ea93259a5b...|8602a61d680a10a82...|20
|0006ec9db01a64e59...|            1|99a4788cb24856965...|4a3ca9315b744ce9f...|20
```

- Download file from Brightspace
- Transfer file into Google Colab
- Start running the code



THANK YOU

SLEEPYHEAD
SLEPPYMEL@GMAIL.COM