ASSIGNMENT 6

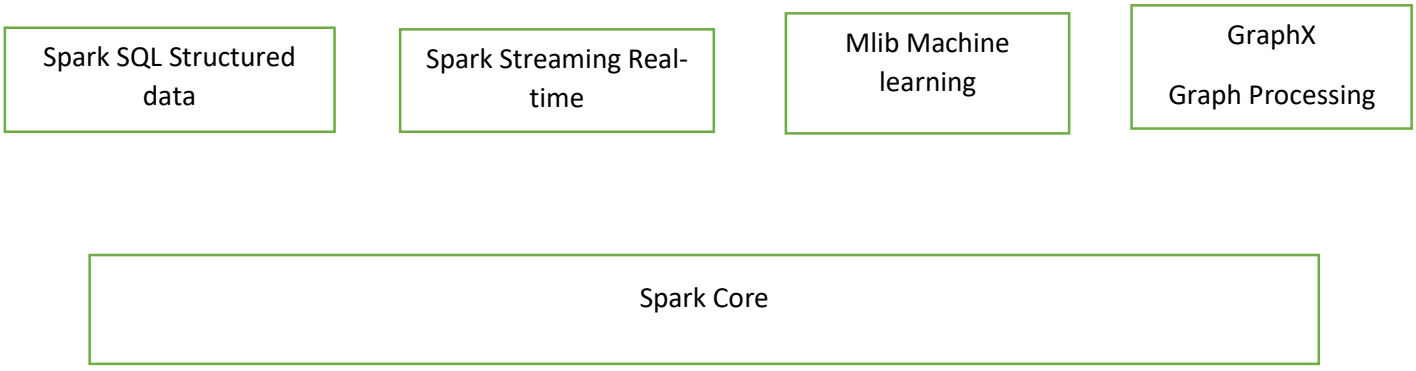
NAME: Aishee Bhattacharya

BATCH: DXC-262-Analytics-B12-Azure

DATE: 06/06/2022

1. Explain what is in-Memory computation in details?
    - Processing in memory is one approach to overcoming the von Neumann bottle neck which is a limitation throughout caused by latency inherent in the standard computer architechture.
    - In-memory computing primary relies on keeping data in a sever's RAM as means of processing at faster speeds.
    - In-memory computation especially applies to problems that require extensive access to data analytics, reporting or data warehousing and big data applications.
2. Explain advantages of Spark framework ?
    - Apache Spark is an open-source cluster computing(In-memory) framework.
    - Spark was built on the top of Hadoop MapReduce, Spark process the data much quicker than Hadoop framework.
    - Hadoop 1 x times faster- Spark will be 100xtimes faster.

3. Explain components of Spark with block diagram ?

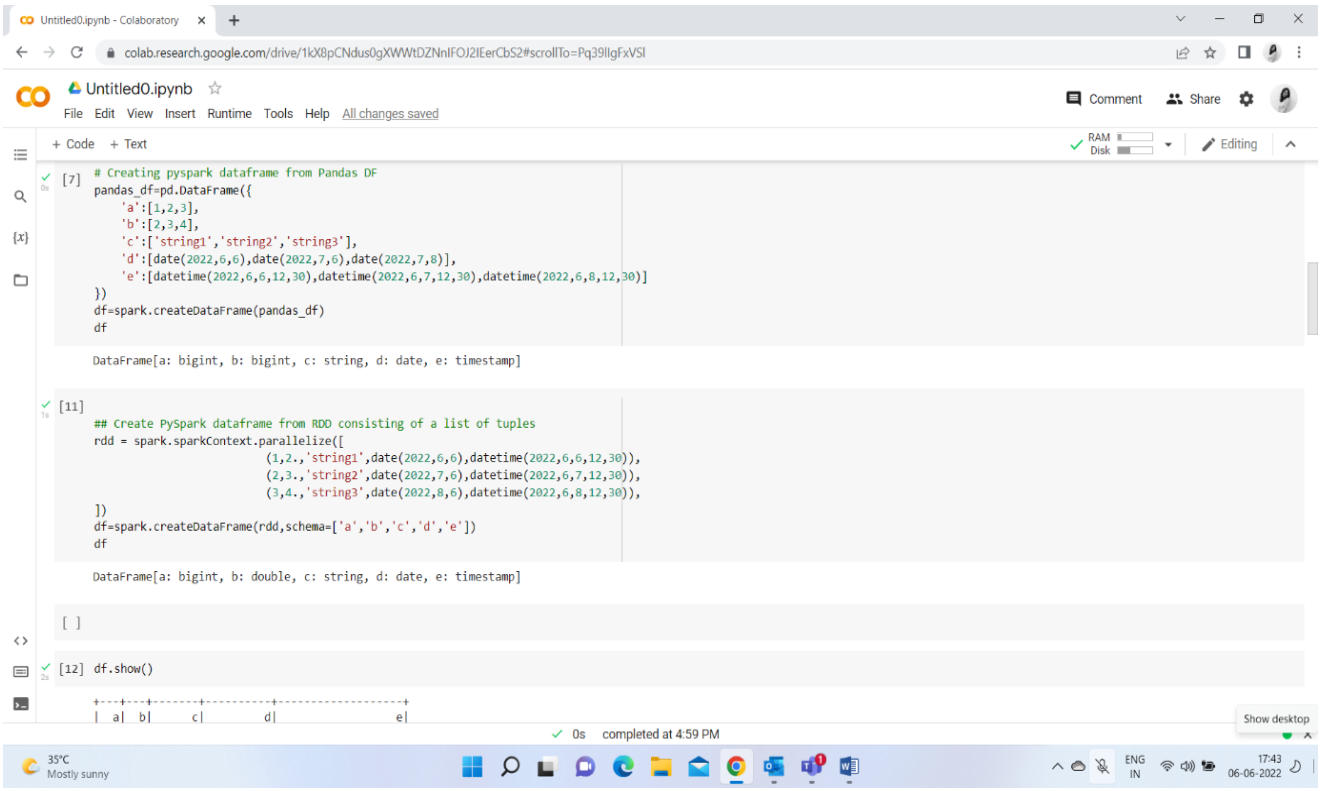


4. Explain benifits of in-Memory computation ?
    - Better faster decision making
    - Ability to reduce cost.
    - Identify competitive opportunities
    - Grow revenue
    - More efficient application
    - Reduce risk
    - It's best suited for performing real-time analytics and developing and deploying real-time applications.
    - In-memory computing imperative:
        1. Avoid movement of detailed data
        2. Calculate first ,then move the results.

5. Explain major difference between Hadoop & Spark ?
   Spark is faster compared to Hadoop because it uses RAM instead of reading and writing intermediate data to disks. Whereas, Hadoop stores data on multiple sources and processes it in batches via MapReduce.

6. Explain features of Spark?
    - Fast: It provides high performance for both batch and streaming data,using a state of the art DAG scheduler, a query optimizer, and a physical execution engine.
    - Easy to use: It supports various languages like Java, Python ,Scala ,Sql ,R. It facilitates to write the application in Java ,Scala,Python,R, and SQL.It provides more than 80 high-level operations.
    - Supports various libraries: It provides a collection of libraries including SQL and DataFrames, MLib for machine learning, GraphX and Spark streaming.
    - Supports Realtime streaming
    - Lightweight: It is a light unified analytics engine which is used for large scale data processing.
    - Runs everywhere- It can easily run on Hadoop ,Apache Mesos, Kubernetes, standalone or in the cloud.

7. Write a Py-Spark program to create Dataframe from RDD & explain with screenshots & steps ?



8. Explain what is RDD & why it is needed ?

- It is basic building block of Spark
- The RDD(Resilient Distributed Dataset) is the Spark's core abstraction.
- It is a collection of elements, partitioned across nodes of the cluster so that we can execute various parallel operations on it.
- There are two ways to create RDDs:
  1. Parallelizing an existing data in the driver program
  2. Referencing a dataset in an external storage system, such as filesystem, HDFS, HBase or any data sourceoffering a Hadoop Input Format.

9. Write a Py-Spark program to make the column in Upper case & explain with screenshots & steps ?