TRAINING ASSIGNMENT

(6th JUNE 2022)

NAME – ROHIT ARORA ROLLNO – DXC-262AB-1209

BATCH – DXC-262-ANALYTICS-B12-AZURE COMPANY – DXC TECHNOLOGY

EMPLOYEE DOMAIN – AZURE ANALYTICS TRAINER NAME – MR. AJAY KUMAR

DATE OF SUBMISSION – 6RD JUNE 2022 NO.OF QUESTIONS: 7

1. Explain what is in-Memory computation in detail?

In-Memory computation is a concept of increasing the processing of data. This works by keeping the data which is extensively used by applications like reporting , preprocessing , big data processing in the RAM itself so through the processing speed of data is switched to maximum.

2. Explain advantages of Spark framework ?

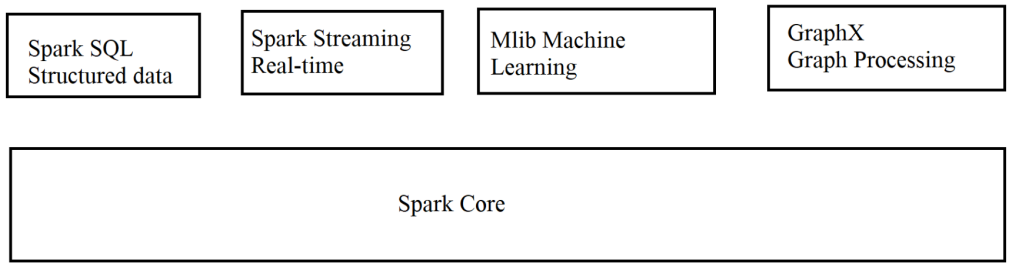
Spark framework is an open source distributed and computing data processing process for big data analysis.

Advantages

- Spark is specially designed to process big data therefore it's fast in processing huge chunks of data.
- Supported and available in multiple languages like Python , Java , Scala , R etc
- Apache Spark carries easy-to-use APIs for operating on large datasets. It offers over 80 high-level operators that make it easy to build parallel apps.
- Apache Spark can handle many analytics challenges because of its low-latency in-memory data processing capability. It has well-built libraries for graph analytics algorithms and machine learning.

3. Explain components of Spark with block diagrams?

Components of Apache Spark:



4. Explain benefits of in-Memory computation ?

The benefits of in-Memory computation are it increases the processing speed of data as the data is directly being stored in the RAM itself instead of slow hard drives therefore the speed of data processing and analysing increases.

5. Explain major differences between Hadoop & Spark ?

Hadoop	Spark
1. Hadoop is an open source framework which uses Map Reduce Algorithm	1. Spark is a lightning fast clustering algorithm , it is an extension of Hadoop's Map Reduce algorithm and gives more types of computation.
2. Hadoop is slow as it reads data from hard disks.	2. Spark is fast as it uses in-memory data storing technique.
3. Hadoop is designed to handle batch processing efficiently	3. Spark is designed to handle real-time data efficiently.

6. Explain features of Spark?

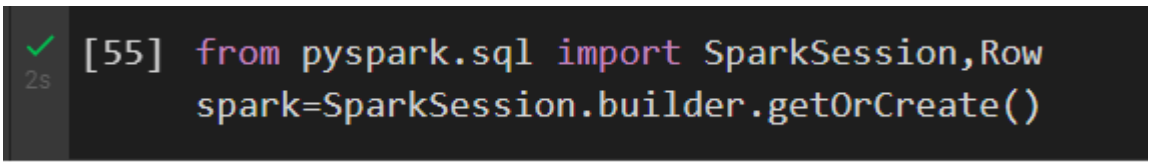
- Speed — Spark runs up to 100 times faster than Hadoop MapReduce for large-scale data processing. It is also able to achieve this speed through controlled partitioning.
- Powerful Caching — Simple programming layer provides powerful caching and disk persistence capabilities.
- Deployment — It can be deployed through Mesos, Hadoop via YARN, or Spark's own cluster manager.

- Real-Time — It offers Real-time computation & low latency because of in-memory computation.
- Polyglot — Spark provides high-level APIs in Java, Scala, Python, and R. Spark code can be written in any of these four languages. It also provides a shell in Scala and Python.

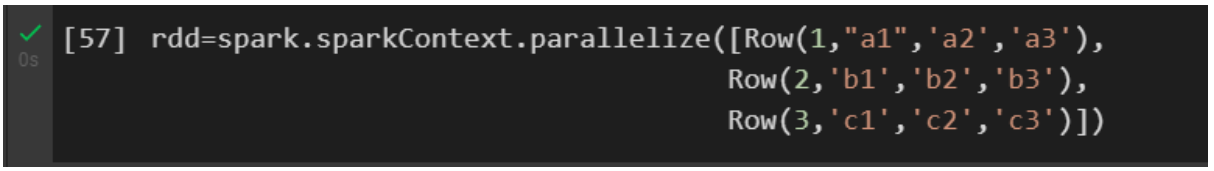
7. Write a Py-Spark program to create Dataframe from RDD & explain with screenshots & steps ?

For creating a RDD data frame here we are going to use Python language using Jupyter notebook to create a sample RDD dataframe. Firstly we will install the pyspark library into the environment or in your system.

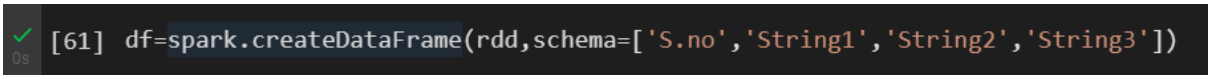
After pyspark is installed we will import SparkSession and Row from spark module which is used to create the spark session and Row is used to define a row of dataframe



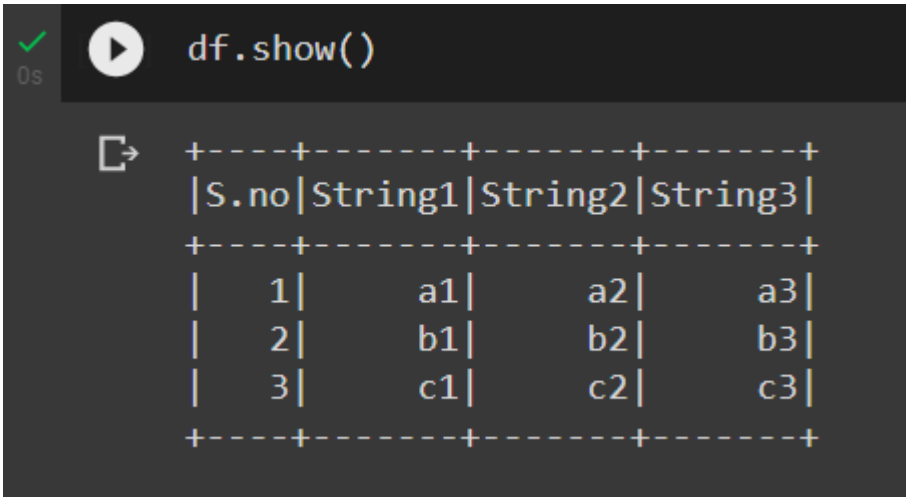
After the session/instance is created now we need to create a RDD instance of the data



After the process is done now we will finally create the dataframe



Printing the Dataframe in a tabular form


0s
df.show()
+----+-------+-------+-------+
|S.no|String1|String2|String3|
+----+-------+-------+-------+
1	a1	a2	a3
2	b1	b2	b3
3	c1	c2	c3
+----+-------+-------+-------+