ASSIGNMENT-5

NAME-KONAKALLA AKHILA ROLL NUMBER-DXC262AB1225

BATCH-DXC-262-ANALYTICS-B12-AZURE

COMPANY-DXC TECHNOLOGY

EMPLOYEE DOMAIN-AZURE ANALYTICS

TRAINING UNDER-MANIPAL PRO LEARN

TRAINER NAME-MR.AJAY KUMAR

DATE OF SUBMISSION -6th JUNE 2022

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

In-memory computation (or in-memory computing) is the technique of running computer calculations entirely in computer memory.

Processing in memory is one approach to overcoming the von Neumann bottleneck,which is limitation on throughput caused bythe latency inherent in the standard computer architecture.

In-memory computing primarily relies on keeping data in servers RAM as means of processing at faster speeds.

2.Explain adavantages od spark framework ?

1. Speed
2. Ease of Use
3. Advanced Analytics
4. Dynamic in Nature
5. Apache Spark is powerful
6. Increased access to Big data
7. Demand for Spark Developers
8. Open-source community

3. Explain components of spark with block diagram?

1. Spark Core
2. Spark SQL
3. Spark Streaming
4. MLlib (Machine learning library)
5. GraphX

|  |
| --- |
| Spark SQL  Structure data |

|  |
| --- |
| Spark Streamning  Real-time |

|  |
| --- |
| Mlib Machine learning |
| Graph X  Graph processing | |

|  |
| --- |
| Spark core |

4. Explain benfitis of in-memory computation ?

Ability to reduce cost

Grow revenue

Reduce risk

Best suited for performing real-time analytics.

Faster in decision making

Avoid movement of detailed data.

Calculate first and ,then move the results.

5.Explain major difference between hadoop and spark ?

|  |  |
| --- | --- |
| Hadoop | Spark |
| Hadoop is an open source framework which uses a MapReduce algorithm | Spark is lightning fast cluster computing technology, which extends the MapReduce model to efficiently use with more type of computations. |
| Hadoop is designed to handle batch processing efficiently | Spark is designed to handle real-time data efficiently. |
| With Hadoop MapReduce, a developer can only process data in batch mode only | Spark can process real-time data, from real time events like twitter, facebook |

6.Explain features of spark ?

**1. Lighting-fast processing speed**

Big Data processing is all about processing large volumes of complex data. Hence, when it comes to Big Data processing, organizations and enterprises want such frameworks that can process massive amounts of data at high speed. As we mentioned earlier, Spark apps can run up to 100x faster in memory and 10x faster on disk in Hadoop clusters.

### 2. Ease of use

### Spark allows you to write scalable applications in Java, Scala, Python, and R. So, developers get the scope to create and run Spark applications in their preferred programming languages. Moreover, Spark is equipped with a built-in set of over 80 high-level operators.

### 3. It is flexible

### Spark can run independently in cluster mode, and it can also run on Hadoop YARN, Apache Mesos, Kubernetes, and even in the cloud. Furthermore, it can access diverse data sources.

### 7.Write a py-spark program to create Dataframe from RDD &explain with screenshot steps?

### Creation of dataframe using pandas

### First import pandas as

### Import pandas as pd

### Pandas\_df=pd.DataFrame({

### ‘a’:[2,3,4],

### ‘b’:[‘string1’,’string2’,’string3,],

### ‘c’:[date(2022,6,6), date(2022,6,7),

### date(2022,6,8),],

### ‘d’:[datetime(2022,6,6,4,40), datetime(2022,6,7,4,41), datetime(2022,6,8,4,42)]

### })

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

### Resilient Distributed Datasets (RDD) is a fundamental data structure of Spark. It is an immutable distributed collection of objects. Each dataset in RDD is divided into logical partitions, which may be computed on different nodes of the cluster. RDDs can contain any type of Python, Java, or Scala objects, including user-defined classes.

### Formally, an RDD is a read-only, partitioned collection of records. There are two ways to create RDDs − parallelizing an existing collection in your driver program, or referencing a dataset in an external storage system, such as a shared file system, HDFS, HBase, or any data source offering a Hadoop Input Format.

### RDD is a basic data structure used in Spark to execute the MapReduce operations faster and efficiently. Data sharing in MapReduce take a lot of time because of replication, serialization, and disk IO. Hadoop applications take over 90 percent of the time in read-write operations. So, researchers came up with this RDD concept that uses in-memory processing computation. Using RDDs increased the data sharing in memory by 10 to 100 times faster than network and disk.

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

Convert column to upper case in pyspark – upper() function

### from pyspark.sql import column

### from pyspark.sql.functions import upper

### type(df.a)==type(upper(df.c))==type(df.a.isnull())

### df.withcolumn(‘upper\_a’,upper(df.a)).show()