BIG DATA ANALYTICS ASSIGNMENT I

Name: Aditya Mahajan
Roll No: A - 10

Reg No - 2017BCS001

1. What is Big Data?
Answer: It describes the large volume of Data both Structured and Unstructured. The term Big Data refers to the simple use of predictive analytics, user behavior analytics and other advanced data analytics methods.

It is extract value from data and seldom to a particular size to the data set. The challenge includes capture, storage, search, sharing, transfer, analysis, creation.

2. What do you know about the term "Big Data"?
Answer: Big Data is a term associated with complex and large datasets. A relational database cannot handle big data, and that's why special tools and methods are used to perform operations on a vast collection of data. Big data enables companies to understand their business better and helps them derive meaningful information from the unstructured and raw data collected on a regular basis. Big data also allows the companies to make better business decisions backed by data.

3. Explain the NameNode recovery process?
Answer: The NameNode recovery process involves the below mentioned steps to make the Hadoop cluster running: In the first step in the recovery process, the file system metadata replica (FsImage) starts a new NameNode. The next step is to configure the DataNodes and Clients. These DataNodes and Clients will then acknowledge the new NameNode.

During the final step, the new NameNode starts serving the client on the completion of last checkpoint FsImage loading and receiving block reports from the DataNodes. Note : Don't forget to mention, this NameNode recovery process consumes a lot of time on large Hadoop clusters. Thus, it makes routine maintenance difficult. For this reason, HDFS high availability architecture is recommended to use.

4. What is the purpose of the JPS command in Hadoop?

Answer: The JBS command is used to test whether all Hadoop daemons are running correctly or not. It specifically checks daemons in Hadoop like the NameNode, DataNode, ResourceManager, NodeManager, and others.

5. Explain the core methods of a Reducer?

Answer: There are three core methods of a reducer. They are-
setup() – Configures different parameters like distributed cache, heap size, and input data. reduce() – A parameter that is called once per key with the concerned reduce task cleanup() – Clears all temporary files and called only at the end of a reducer task.

6. Where does Big Data come from?

Answer: There are three sources of Big Data

Social Data : It comes from the social media channel's insights on consumer behavior. Machine Data : It consists of real-time data generated from sensors and weblogs. It tracks user behavior online.

Transaction Data : It generated by large retailers and B2B Companies frequent basis.

7. How are file systems checked in HDFS?

Answer: File system is used to control how data are stored and retrieved. Each file system has a different structure and logic properties of speed, security, flexibility, size.

Such kind of file system designed in hardware. This file includes NTFS, UFS, XFS, HDFS.

8. What are the four features of Big Data?

Answer: The four V's renders the perceived value of data. It is as valuable as the business results bringing improvements in operational efficiency.

- Volume
- Velocity
- Variety
- Veracity

9. What are some of the interesting facts about Big Data?

Answer: According to the experts of the industry, digital information will grow to 40 zettabytes by 2020. Surprisingly, every single minute of a day, more than 500 sites come into existence. This data is certainly vital and also awesome

With the increase in the number of smartphones, companies are funneling their money into it by carrying mobility to the business with apps

It is said that Walmart collects 2.5 petabytes of data every hour from its consumer transactions

10. How will you define checkpoint?

Answer: It is the main part of maintaining filesystem metadata in HDFS. It creates checkpoints of file system metadata by joining fsimage with edit log. The new version of the image is named Checkpoint.

11. Pig Latin contains different relational operations; name them?
Answer: The important relational operations in Pig Latin are:

- group
- distinct
- join • for each
- order by
- filters
- limit.

12. What is the meaning of big data and how is t different?
Answer: Big data is the term to represent all kind of data generated on the internet. On the internet over hundreds of GB of data is generated only by online activity. Here, online activity implies web activity, blogs, text, video/audio files, images, email, social network activity, and so on. Big data can be referred to as data created from all these activities. Data generated online is mostly in unstructured form. Big data will also include transaction data in the database, system log files, along with data generated from smart devices such as sensors, IoT, RFID tags, and so on in addition to online activities.

Big data needs specialized systems and software tools to process all unstructured data. In fact, according to some industry estimates almost 85% data generated on the internet is unstructured. Hence, RDBMS processing can be quickly done using a query language such as SQL. On the other hand, big data is very large and is distributed across the internet and hence processing big data will need distributed systems and tools to extract information from them. Big data needs specialized tools such as Hadoop, Hive, or others along with high-performance hardware and networks to process them.

13. Why is big data important for organizations?

Answer: Big data is important because by processing big data, organizations can obtain insight information related to:

- Cost reduction
- Improvements in products or servic es
- To understand customer behavior and markets
- Effective decision making
- To become more competitive

14. What is big data solution implementation?

Answer: Big data solutions are implemented at a small scale first, based on a concept as appropriate for the business. From the result, which is a prototype solution, the business solution is scaled further. Some of the best practices followed in the industry include,

- To have clear project objectives and to collaborate wherever necessary
- Gathering data from the right sources
- Ensure the results are not skewed because this can lead to wrong conclusions
- Be prepared to innovate by considering hybrid approaches in processing by including data from structured and unstructured types, include both internal and external data sources

- Understand the impact of big data on existing information flows in the organization. (company)

15. Which hardware configuration is most beneficial for Hadoop jobs?

Answer: It is best to use dual processors or core machines with 4 / 8 GB RAM and ECC memory for conducting Hadoop operations. Though ECC memory cannot be considered low-end, it is helpful for Hadoop users as it does not deliver any checksum errors.

The hardware configuration for different Hadoop jobs would also depend on the process and workflow needs of specific projects and may have to be customized accordingly.

16. What is Hive Metastore?

Answer: Hive megastore is a database that stores metadata about your Hive tables (eg. Table name, column names and types, table location, storage handler being used, number of buckets in the table, sorting columns if any, partion columns if any, etc.).

When you create a table, this megastore gets updated with the information related to the new table which gets queried when you issue queries on that table.

Hive is a central repository of hive metadata. it has 2 parts of services and data. by default, it uses derby DB in local disk. it is referred to as embedded megastore configuration. It tends to the limitation that only one session can be served at any given point of time.