Q1. What are the various characteristics of BIg Data ?

Back in 2001, Gartner analyst Doug Laney listed the 3 ‘V’s of Big Data – Variety, Velocity, and Volume. Let’s discuss the characteristics of big data.

These characteristics, isolatedly, are enough to know what is big data. Let’s look at them in depth:

1) Variety

Variety of Big Data refers to structured, unstructured, and semistructured data that is gathered from multiple sources. While in the past, data could only be collected from spreadsheets and databases, today data comes in an array of forms such as emails, PDFs, photos, videos, audios, SM posts, and so much more. Variety is one of the important characteristics of big data.

Variety

As Discussed before, Big Data is generated in multiple varieties. Compared to the traditional data like phone numbers and addresses, the latest trend of data is in the form of photos, videos, and audios and many more, making about 80% of the data to be completely unstructured

Structured data is just the tip of the iceberg.

2) Velocity

Velocity essentially refers to the speed at which data is being created in real-time. In a broader prospect, it comprises the rate of change, linking of incoming data sets at varying speeds, and activity bursts.

Velocity

Last but never least, Velocity plays a major role compared to the others, there is no point in investing so much to end up waiting for the data. So, the major aspect of Big Dat is to provide data on demand and at a faster pace.

3) Volume

Volume is one of the characteristics of big data. We already know that Big Data indicates huge ‘volumes’ of data that is being generated on a daily basis from various sources like social media platforms, business processes, machines, networks, human interactions, etc. Such a large amount of data are stored in data warehouses. Thus comes to the end of characteristics of big data.

Volume

Volume refers to the unimaginable amounts of information generated every second from social media, cell phones, cars, credit cards, M2M sensors, images, video, and whatnot. We are currently using distributed systems, to store data in several locations and brought together by a software Framework like Hadoop.

Facebook alone can generate about billion messages, 4.5 billion times that the “like” button is recorded, and over 350 million new posts are uploaded each day. Such a huge amount of data can only be handled by Big Data Technologies

4) Veracity

Veracity basically means the degree of reliability that the data has to offer. Since a major part of the data is unstructured and irrelevant, Big Data needs to find an alternate way to filter them or to translate them out as the data is crucial in business developments

5) Value

Value is the major issue that we need to concentrate on. It is not just the amount of data that we store or process. It is actually the amount of valuable, reliable and trustworthy data that needs to be stored, processed, analyzed to find insights.

Q2. How big data differs from traditional data ?

|  | Traditional Data | Big Data |
| --- | --- | --- |
| Types of data | Traditional database systems are based on structured data i.e traditional data is stored in fixed formats or fields in a file. | Big data uses the semi-structured and unstructured data and improves the variety of the data gathered from different sources |
| Volume of Data | The traditional system database can store only small amount of data ranging from gigabytes to terabytes. | big data helps to store and process large amount of data which consists of hundreds of terabytes of data or petabytes of data and beyond. |
| Data Schema | The traditional database is based on the fixed schema which is static in nature. In traditional database data cannot be changed once it is saved and this is only done during write operations | Big data uses the dynamic schema for data storage. Both the un-structured and structured information can be stored and any schema can be used since the schema is applied only after a query is generated. |
| Data Relationship | In the traditional database system relationship between the data items can be explored easily as the number of informations stored is small. | big data contains massive or voluminous data which increase the level of difficulty in figuring out the relationship between the data items |
| Scaling | achieving the scalability in the traditional database is very difficult because the traditional database runs on the single server and requires expensive servers to scale up | Big data is based on the scale out architecture under which the distributed approaches for computing are employed with more than one server. So, the load of the computation is shared with single application based system. |
| Cost | Traditional database system requires complex and expensive hardware and software in order to manage large amount of data. Also moving the data from one system to another requires more number of hardware and software resources which increases the cost significantly. | While in case of big data as the massive amount of data is segregated between various systems, the amount of data decreases. So use of big data is quite simple, makes use of commodity hardware and open source software to process the data |
| Accuracy and Confidentiality | Under the traditional database system it is very expensive to store massive amount of data, so all the data cannot be stored. This would decrease the amount of data to be analyzed which will decrease the result’s accuracy and confidence. | While in big data as the amount required to store voluminous data is lower. Therefore the data is stored in big data systems and the points of correlation are identified which would provide high accurate results. |
| Flexibility | Less flexible | More flexible |
| Data Source | Centralized | Fully distributed |
| Data Integration | Easy | Difficult |
| Generated Rate | Per hour, per day | Per second |
| System Configuration | Normal system configuration is sufficient to process | High system configuration is required to process |
| Tools required | A traditional database tool is enough | Special kinds of tools are required |
| Functions | Normal functions are enough to manipulate the data | Requires special kind of functions to manipulate the data |

Q3. Explain Map Reduce framework with the help of example ?

MapReduce is a framework using which we can write applications to process huge amounts of data, in parallel, on large clusters of commodity hardware in a reliable manner.

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model.

The Algorithm

Generally MapReduce paradigm is based on sending the computer to where the data resides!

MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

Map stage − The map or mapper’s job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.

Reduce stage − This stage is the combination of the Shuffle stage and the Reduce stage. The Reducer’s job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

During a MapReduce job, Hadoop sends the Map and Reduce tasks to the appropriate servers in the cluster.

The framework manages all the details of data-passing such as issuing tasks, verifying task completion, and copying data around the cluster between the nodes.

Most of the computing takes place on nodes with data on local disks that reduces the network traffic.

After completion of the given tasks, the cluster collects and reduces the data to form an appropriate result, and sends it back to the Hadoop server.

MapReduce is a programming framework that allows us to perform distributed and parallel processing on large data sets in a distributed environment.

MapReduce consists of two distinct tasks — Map and Reduce.

As the name MapReduce suggests, reducer phase takes place after the mapper phase has been completed.

So, the first is the map job, where a block of data is read and processed to produce key-value pairs as intermediate outputs.

The output of a Mapper or map job (key-value pairs) is input to the Reducer.

The reducer receives the key-value pair from multiple map jobs.

Then, the reducer aggregates those intermediate data tuples (intermediate key-value pair) into a smaller set of tuples or key-value pairs which is the final output.

Advantages of Mpreduce:

1. Parallel processing
2. Data locality