Q1. Explain 5Vs ?

Ans1. The concept of big data deals with the data set whose size is beyond the ability of commonly used software tools to capture, manage and process the data within a tolerable elapsed time. It has three fundamental V's popularly known as 3V's which defines the whole analysis process.

* **Volume:** The whole concept of calling the big data as huge revolves around volume. Volume is nothing but the measurement of data generated every second. It can be very well assumed that the total data generated till 2003 is now merely generated in two days. Isn't it big enough? Hence volume is the biggest constraint in the analysis of big data.
* **Velocity:** Velocity is one of the greatest challenge. The data coming up every second has broken all records of past. Thousands of app are installed every single second. Thus the velocity of incoming data is rising exponentially.
* **Variety:** Undoubtedly in this 21st century, we are exposed to million sets of data & so there lies a huge scope to define variety. Here variety is not restricted to text, image & audio files rather it has some precise definitions.
* **Structured:** The data sets which have some schema are structured data. They are easy to analyze but unfortunately only 10-20% data is structured. Ex- Father table
* **Unstructured:** Data which doesn't have any schema. Obviously, it will be tough job to analyze it & its a sheer counterpart that around 80% data is unstructured. Ex- Web blogs
* **Semi Structured Data:** Data which have some schema but doesn't possess full schema. Ex- Xml tags

Apart from these 3V's, 2 more V's are added to complete the definition of big data.

* **Veracity:** Veracity refers to the quality of data. If we are required to fetch the data, then there is always a question mark in the certainty of data. The whole pool of data is quite corrupted & hence the quality and accuracy of the data is not assured.
* **Value:** In this competitive world, data adds value. More is the data, more is the value. This is directly proportional to marketing. If you have analyzed n number of users then you can specifically advertise their choice of interest instead of random advertisement. Hence more is the value of analyzed data, more structured marketing system you will have.

Q2. Difference between Commodity Hardware & Enterprise Edition ?

Ans 2. Commodity Hardware: . Computer **hardware** that is affordable and easy to obtain. Typically it is a low-performance system that is IBM PC-compatible and is capable of running Microsoft Windows, Linux, or MS-DOS without requiring any special devices or equipment. Commodity hardware are usually meant for low cost missions & hence it is carried out.

Enterprise Edition: A flexible and heterogeneous data integration platform for big data processing that enables you to transform, enrich, and govern raw data for faster, more informed decision-making. It works as a server & hence it is expensive to use. Enterprise edition deals with name node which is a master for assigning all other jobs.

Q3. Define Data Locality ?

Ans3. Data Locality deals with the idleness of data. Whenever any mission executes, the bulky data sits idle & the light weight code gets copied & moves to data for analyzing. This reduces network traffic.

The beauty of Hadoop is that it works on local data as far as possible which means that when we invoke any map reduce job, the logic( map and reduce code) is sent to each data node in the cluster. So suppose if you have a text file of size 1 GB and you have written a map reduce code to convert all text in that file to upper case then first the file will be broken into chunks and the logic to cover the text to Upper case would be available to each data node. Now the tasktracker on each node would only run the map reduce code the data block/s present on that local node. This is known as data locality.

Q4. Define Scalability ?

Ans 4. Scalability can be defined as the flexibility to increase or decrease the size of nodes according to the burden.

To further understand scalability, here are two examples. First, a basic anti-virus program can become premium and be used by enterprises through downloading certain add-ons or paying for subscription. Because more resources may be added to it, it is considered scalable. On the other hand, more computers and servers can be added to a network in order to increase throughput or intensify security. This makes the network scalable.