1. **Various sources of Big data**

various sources of Big Data that organizations can use to forward their business, divided into 9 sources, indicating whether these sources are internal or external, structured or unstructured (on a continuum), their velocity, variety and volume, and whether they leverage APIs or not.

1. **Archives**

Archives of scanned documents, statements, insurance forms, medical records and customer correspondence, paper archives, and print stream files that contain original systems of record between organizations and their customers.

1. **Docs**

XLS,PDF,CSV,PPT,HTML,HTML5,XML,JSON,email,Word,plain text etc.

1. **Media**

Images, videos, Audio, Flash, live streams, podcasts, etc.

1. **Data Storage**

SQL, NoSQL, Hadoop, doc repository, file systems, etc.

1. **Business Apps**

Project Management, marketing automation, productivity, CRM, ERP content

Management systems, HR, storage, talent management, Google Docs, intranets, portals, etc.

1. **Public Web**

Government, weather, competitive, traffic, regulatory, compliance, health care services, economic, census, public finance, stock, OSINT, the World Bank, SEC/Edgar, Wikipedia, IMDb, and other web services.

1. **Social Media**

Twitter, LinkedIn, Facebook, Tumblr, Blog, Slide share, YouTube, Google+, Instagram, Flickr, Pinterest, Vimeo, Wordpress, IM, RSS, review, Chatter, Jive, Yammer, etc.

1. **Machine Log Data**

Event logs, server data, application log, business process logs, audit logs, CDRs, mobile location, mobile app usage, click stream data etc.

1. **Sensor Data**

Medical devices, smart electric meters, car sensors, road cameras, satellites, traffic recording devices, processors found within vehicles, video games, cable boxes and household appliances, assembly lines, office buildings, cell towers and jet engines, air conditioning units, refrigerators, trucks, farm, machinery, etc.

1. **3 V's of Big Data**

Big Data is defined by three V’s:

1. **Volume**

Large amount of data.

1. **Variety**

The data comes in different forms, including traditional databases, images, documents, and complex records.

1. **Velocity**

The content of the data is constantly changing, through the absorption of complementary data collections, through the introduction of previously archived data or legacy collections, and from streamed data arriving from multiple sources.

1. **Horizontal Scaling and Vertical Scaling**

**Horizontal Scaling:**

Horizontal scaling means that you scale by adding more machines into your pool of references.

In a database world horizontal-scaling is often based on partitioning of the data i.e. each node contains only part of the data, With horizontal-scaling it is often easier to scale dynamically by adding more machines into the existing pool.

**Vertical Scaling:**

**Vertical scaling means that you scale by adding more power (CPU, RAM) to an existing machine**.

Vertical-scaling the data resides on a single node and scaling is done through multi-core i.e. spreading the load between the CPU and RAM resources of that machine. Vertical-scaling is often limited to the capacity of a single machine, scaling beyond that capacity often involves downtime and comes with an upper limit.

1. **Need and Working of Hadoop**

**Need:**

Hadoop is an open source platform that provides excellent data management provision. It is a framework that supports the processing of large data sets in a distributed computing environment. It is designed to expand from single servers to thousands of machines, each providing computation and storage.  Its distributed file system facilitates rapid data transfer rates among nodes and allows the system to continue operating uninterrupted in case of a node failure, which minimizes the risk of catastrophic system failure, even if a significant number of nodes become out of action.

**Working:**

Hadoop is an ecosystem of libraries, and each library has its own dedicated tasks to perform. HDFS writes data once to the server and then reads and reuses it many times. When comparing it with continuous multiple read and write actions of other file systems, HDFS exhibits speed with which Hadoop works and hence is considered as a perfect solution to deal with voluminous variety of data.

Job Tracker is the master node which manages all the Task Tracker slave nodes and executes the jobs. Whenever some data is required, request is sent to NameNode which is the master node (smart node of the cluster) of HDFS and manages all the DataNode slave nodes. The request is passed on all the DataNode which serves the required data. There is concept of Heartbeat in Hadoop, which is sent by all the slave nodes to their master nodes, which is an indication that the slave node is alive.

MapReduce or YARN, are used for scheduling and processing. Hadoop MapReduce executes a sequence of jobs, where each job is a Java application that runs on the data. Instead of MapReduce, using querying tools like Pig Hadoop and Hive Hadoopgives the data hunters strong power and flexibility.