**HADOOP**

**Big data**

As the years have gone on, the necessity for using technology has increased rapidly to the point where human society would be doomed without it. With each technology comes its own concerns such as how will we hold, make sense and use all that data that is being produced? There is a lot of data being produced constantly such as news, stocks, the forex market

**Apache Hadoop**

Hadoop is an open-source software Java-based framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs. Hadoop is comprised of 2 systems HDFS and MapReduce.

**HDFS**

HDFS is the file system of the Hadoop framework, and stands for Hadoop Distributed File System. HDFS’s intended purpose was to manage and store huge volumes of data in an efficient way.

**User-space File System**

Traditionally the file system is inbuilt within the operating systems kernel and this runs as an OS process however, HDFS it is not inbuilt within the operating system kernel and does not run as an OS process. HDFS runs as a user process within the process space allocated for user processes which is on the OS process table.

**Block size**

On a traditional file system, the block sizes are between 4KB to 8KB however with HDFS due to the environment it is used in the default block size is 64MB.

**Hadoop Cluster**

The Hadoop cluster is formed of 2 entities: Master Node and Slave Node. The set-up is one master node with one or many slave nodes. The master and slaves each have a different job to do.

The master is responsible of the Resource manager and Name node as well as the secondary name node which is a backup of the name node.

The resource manager tracks the resources in a cluster and schedules applications. An example of scheduling an application would be deciding when and which map reduce job will take place. The name node manages the HDFS storage. The name node has information on every bit of data in the Hadoop cluster. Its main duty is getting the data stored with the help of the data nodes. When a new piece of data arrives to the cluster, the name node splits it up and identifies which data nodes will be able to hold that specific partition of the overall data. Having split up and shared the data, the name node then uses an allocation table to record where each piece of data is within the cluster i.e. what bit of data is on what data node. Whenever a call is made on HDFS the name node checks the allocation table and uses that information to show the file.

The slave node is responsible for the data node and node manager. Data nodes are actual storage locations in a cluster. The Hadoop cluster can have as many data nodes as it needs. Data nodes store data as and when commanded by the name node. The node manager is responsible for the execution of the tasks on all the data nodes.

**MapReduce**

Google came up with the algorithm for being able to deal with large chunks of data which led to Hadoop creating the MapReduce. Map reduce works in such as way whereby the huge chunks of data is split into smaller pieces and then the same computation is worked on all pieces of data which is then put back together at the end.