Big Data is the nomer used when the data set needs more space for storage than 1 machine can provide. Performing actions on this data set would also take up more computing power than one computer provides. Hadoop is a framework that allows the storage and manipulation of this data that uses a multiple linear machine strategy, that is it uses many computers to store information and also uses the computation power of these machines as well. Hadoop has many programs that allow for the successful manipulation of this data.

Some of the many programs Hadoop consists of is as follows:

Data Storage - HDFS, HBASE

Data Processing - Mapreduce, YARN

Data Access - HIVE, PIG, SQOOP

Data Management - Flume, Zookeeper

Data is gathered in two methods, transactional and analytical. THis data is then stored in one of the multiple machines for recall at a later point. The machines are nodes. It is fail safe uch that if one machine goes down, all the data can be recalled on another machine.

Databases can be stored in two different ways, SQL and NOSQL. The difference is that SQL follows a schema, a categorical layout of tables which dara can be merged and recalled through primary keys and SQL queries. NOSQL databases such as MongoDB are formatted through a key:value pair.

The Hadoop Daemons are:

Name Node - this stores metadata information

Data Node - this tores the actual physical information

Secondary Node - Used for Checkpointing

Resource Manager - Resource Allocation

Node Manager - Task Execution