DATA WHICH IS HUGE

BIG DATA

These days, one thing that is being produced in tremendous amount is data. Every time we turn to our search engines for answers, we produce data. Talking about the facts we produce 2.5 quintillion bytes of data each day and the most interesting and mind boggling thing is that, 90% of the data in the world today has been created in the last two years alone. Last year, more than 3.5 million text messages were sent every minute. This year however, it's around 15.2 million texts, a 334 percent increase. Other facts are like: 103,447,520 spam emails are sent every minute, Google 3.6 million searches, Spotify adding 13 new songs, we tweet 456,000 times, Uber taking 45,787 trips each minute, post 46,740 Instagram photos, and publish 600 new page edits on Wikipedia each minute. All this complex, structured and unstructured data is collectively called as big data. How to handle this large amount of data that was a big question. Thanks to Doug Cutting, Mike Cafarella and Google, they formed the foundation to answer this question. Google published technical papers detailing its Google File System (GFS) and MapReduceprogramming framework in 2003 and 2004, respectively, Cutting and Cafarella modified earlier technology plans and developed a Java-based MapReduce implementation. In early 2006, those elements were split off from Nutch and became a separate Apache subproject, which Cutting named Hadoop after his son's stuffed elephant. At the same time, Cutting was hired by internet services company Yahoo, which became the first production user of Hadoop later in 2006. Use of the framework grew over the next few years, and three independent Hadoop vendors were founded: Cloudera in 2008, MapR a year later and Hortonworks as a Yahoo spinoff in 2011. In addition, AWS launched a Hadoop cloud service called Elastic MapReduce in 2009. That was all before Apache released Hadoop 1.0.0, which became available in December 2011 after a succession of 0.x releases.

Hadoop is an open source distributed processing framework that manages data processing and storage for big data applications running in clustered systems. It is at the centre of a growing ecosystem of big data technologies that are

primarily used to support advanced analytics initiatives, including predictive analytics, data mining and machine learning applications. Hadoop can handle various forms of structured and unstructured data, giving users more flexibility for collecting, processing and analysing data. Hadoop runs on clusters of commodity servers and can scale up to support thousands of hardware nodes and massive amounts of data. It uses a namesake distributed file system that's designed to provide rapid data access across the nodes in a cluster, plus fault-tolerant capabilities so applications can continue to run if individual nodes fail. The Basic Components of Hadoop Architecture are Hadoop Distributed File System (HDFS), Map Reduce, Yarn and different Apache Hadoop Frameworks are Hive, Ambari, HBase, Pig, ZooKeeper. Hadoop has solved one of the biggest problems of the generation but looking at the data generation rate the question is for how long it will work?