Reading Assignment Paper 2

February 7, 2020

1 Big Data : Then vs Now

This paper gives insights on the evolution of technologies in big data technology, from disk management systems to distributed computing systems.Adam clearly states that the definition of big data keeps changing with the evolution of technology and increase in data points, given the developers has the edge cutting techniques to solve such big data systems with hardware and algorithms. Earlier, Big data was stored in disks and the amount of was not really large to manipulate data. In case the data is more than expected, then the storage was an issue. But later, when researched in depth, the problems behind running query faster was independent of the storage database systems, and it was behind manipulation of data. Manipulation of transaction data like web search, retail store data, user behaviour have to deal with large amount of data handling. With these roadblocks, huge storage space and efficient hardware is required to carry the process and analysis for business needs. Redundancy issues are mostly faced with huge amount of data, to solve this data was normalized to increase performance of the database systems. The challenges is not only with storage systems and hardware based, but also client side applications used for analytics and manipulations. Applications such as Excel Sheets were used to manipulate such data to make visualizations on data results. Then growth of data scaling issues led to distributed system strategies to tackle pathologies of big data. Nodes are not reliable though since when there is multiple copies of data, it is not easy to manage.

Now, technological development have started to support big data, storage and memory management issues. From physical to virtual cloud systems, all are built to overcome such storage issues. Retrieval of data has also been made lightning fast. Even cost of hardwares are coming down leading to more big cluster computers and distributed systems. Fault tolerance, scalability are the major factors to these clustered systems that data is stored in more than one nodes to handle any failures. With Google's Cloud and Amazon Web Services in the present market, developers have no struggles working on the algorithms and strategies to handle large amounts of data. The analysis of big data has been easier with the invent of hadoop and spark with NoSQL column wise data stores like Cassandra. The queries are designed to read 1 million key values from a partition for Cassandra using key-value mappings distributed over a cluster in a partition of the database. This is a significant performance boost compared to MySQL query retrievals. Traditional databases are mainly being retrieved by HDFS to process large amount of structured and unstructured data. With these advancements and featured services like Amazon Web Services, Big data management has evolved and not only used in processing and visualizing data but also in machine learning algorithms and artificial intelligence. Only with the confidence of faster processing rate of data, such advancements in technologies can be assured. So, the definition of big data is never constant with the ever increasing horizons of knowledge.