**BIG DATA**

**Introduction**

Today, large of volume of data is being generated at rate that is unprecedented from several sources and this trend will only increase with time. This is a result of continuous technology trends like Cloud Computing and internet of things. Big data is a term that is used when traditional data mining and data handling methods are not able to process and analyze huge volume of data available. The data is too large and transmitted at a high rate, which cannot be managed by relational database engines. To solve this problem, a different and innovative approach is used which uses massive parallelism. (Techopedia.com, 2018)

The need for big data is down to many reasons, one of these is that most data that is collected and needs to be processed is mostly unstructured and hence needs different storage and processing solutions. Additionally, the computational power that is available is extremely huge and can be harnessed to tackle the problem. The data that is produced each day, is raw and has no value. It needs to be processed in order to be useful and of value. (Techopedia.com, 2018).

Big Data comprises of the three Vs namely, Volume, Velocity and Variety. Volume refers to the amount of data that is collected. This data can be unstructured and might be as big as tens of terabytes. The number varies according to the area. The second V is the velocity. Velocity is nothing but the rate at which data needs to be transmitted or travel from one place to another. Since data is in the form of complex unstructured format, it is important for it to be sorted properly and sent with optimal reliability i.e. without any errors. The third V is variety, which refers to the different types of data available. Earlier, data was structured enough and hence it was not really a problem to manage large amount of data. But since the introduction of unstructured and semi-structured data, requirement of stronger tools is essential. (Oracle.com, 2018)

Big data has its own benefits, one of them is that it provides more clear and complete answers to problems since there is more information available. Secondly, since the data available is more complete, it gives more confidence in tackling problems. (Oracle.com, 2018)

Big data can be a useful resource in managing data and business activities. Some of the use cases are:-

1. Product development

Industries like Netflix make use of big data to predict customer demand based on the type of media the user chooses to watch. Based on this pattern, Netflix makes use of the data collected and puts up a list of movies that the user will be most likely to watch. Predictive models are built along with several services by classifying key attributes of previous products and services and a relationship is laid out. Similarly, many industries use the data collected to predict the type of advertisements that needs to be viewed for the user. (Oracle.com, 2018)

1. Predictive maintenance(Oracle.com, 2018)

In case of avoiding mechanical failures, structured data can provide information about certain factors like equipment year, make, model etc. But unstructured data can give an even better in-depth analysis and information like log entries, sensor data etc. Analyzing these issues can be extremely time saving and cost effective. (Oracle.com, 2018)

1. Customer Experience

Big data allows user experience to be even more improved than it used to be by allowing the users to gather data from various sources on the internet and consequently improve interaction. Personalization of products and services allow better user experience and quality of service.(Oracle.com, 2018)

1. Machine Learning

Machine learning is nothing but a technology available right now that allows machines to learn and train themselves instead of being programmed to perform tasks. With abundant data available, machines learning is on a rise leading to new and innovative technologies to rise. (Oracle.com, 2018)

**Solutions**

1. Apache Cassandra

Cassandra provides an effective management of huge amounts of data. The database provides continuous availability and scalability without tampering the performance of the hardware and cloud infrastructure. Cassandra is known for its fault tolerance, decentralized nodes, durability and elasticity. (Big Data Made Simple - One source. Many perspectives, 2018)

1. Hadoop

Hadoop is also developed by Apache and it is widely used by industries just like its counterparts. It is known for its high processing power of large volume of data sets in clusters. Alongside, it carries an advanced software library coupled with highly effective programming models. Regular updates are an added advantage for the developers. (Big Data Made Simple - One source. Many perspectives, 2018)

1. Cloudera

Cloudera is a data repository that is used by corporate users. It can be accessed and be used to reduce business risks and ultimately giving the companies a competitive advantage. (Big Data Made Simple - One source. Many perspectives, 2018)

1. Microsoft HDinsight

A solution from Microsoft and powered by Apache Hadoop, HDinsight uses Windows Azure Blob to store data by default and also provide high availability with low cost. (Data, 2018)

1. NoSQL

The traditional SQL can be used to handle a large amount of data, however SQL is useful for structured data. In case of unstructured data, NoSQL (Not only SQL) is used. NoSQL stores data with no schema and it is possible to have set of column values for each row. (Data, 2018)

1. Big data in EXCEL

Excel is widely used to do analysis and hence data analysis is done with the help of Hadoop connected with Excel. (Data, 2018)

**Comparison Of NoSQL Databases HBase, Cassandra & MongoDB**

**Hbase:**

Apache Hbase runs on top of HDFS and is a NoSQL key/value database. It does not use MapReduce and instead runs in real-time on its own database. In Hbase, the columns are split into families after being partitioned into tables. The column families are declared in the schema and the column families combine with other set of columns. Hbase stores data as key and value supporting four primary operations that are put to add or update rows, scan to retrieve a range of cells, get to return cells for a particular row and delete to remove the rows. (Anon, 2018)

Advantages:

1. Excellent Optimzation for read.
2. Range based scan is good.
3. Consistent
4. Faster read and write operations

(Anon, 2018)

Disadvantages:

1. Transactional operations and relational analytics
2. Full table scan is needed for applications

(Anon, 2018)

**Cassandra:**

Cassandra is a non-relational database developed by Apache. It is a distributed database that is used to manage a large amount of structured data across multiple servers. It provides high availability, no single point of failure, continuous availability, linear scale performance, operational simplicity and easy distribution of data across multiple zones. There is no master node and all nodes play an identical role where all the nodes work with each other and communicate with each other equally. Due to no single point of failure, it offers continuous availability and zero downtime by simply adding new nodes to an existing cluster without taking it down. (DataStax Academy: Free Cassandra Tutorials and Training, 2018)

Advantages:

1. Easy setup
2. Maintenance is efficient.
3. Parsing is flexible.
4. No need for more than one secondary index.

(Anon, 2018)

Disadvantages:

1. Secondary index.
2. Relational data
3. Rollback and commit operations.
4. Authorization needed on data
5. Low latency.

(Anon, 2018)

**MongoDB:**

MongoDB is a document database with the scalability and flexibility that is needed while querying and indexing a dataset. It stores its data in the form of JSON like documents. Since data can vary from document to document, MongoDB allows changing in structure of the data over time. It has the ability to map the objects in the application code making data management more efficient to work with. It also consists of Ad hoc queries, indexing and real time aggregation, providing a powerful method to analyze data. MongoDB is a distributed database and hence it has a built in characteristics like high availability, horizontal scaling and geographic distribution. (MongoDB, 2018)

Advantages:

1. Efficient option to replace RDBMS for web applications.
2. Content management is semi-structured.
3. Real-time analytics
4. Caching and high scalability.
5. High-speed logging.

(Anon, 2018)

Disadvantages:

1. Transactional operations and systems.
2. Foreign key limitations for applications with traditional database management systems.

(Anon, 2018)

Table 1 shows a comparison analysis of Cassandra, Hbase and MongoDB based on operational and analytical workload. (DataStax: always-on data platform | NoSQL | Apache Cassandra, 2018)

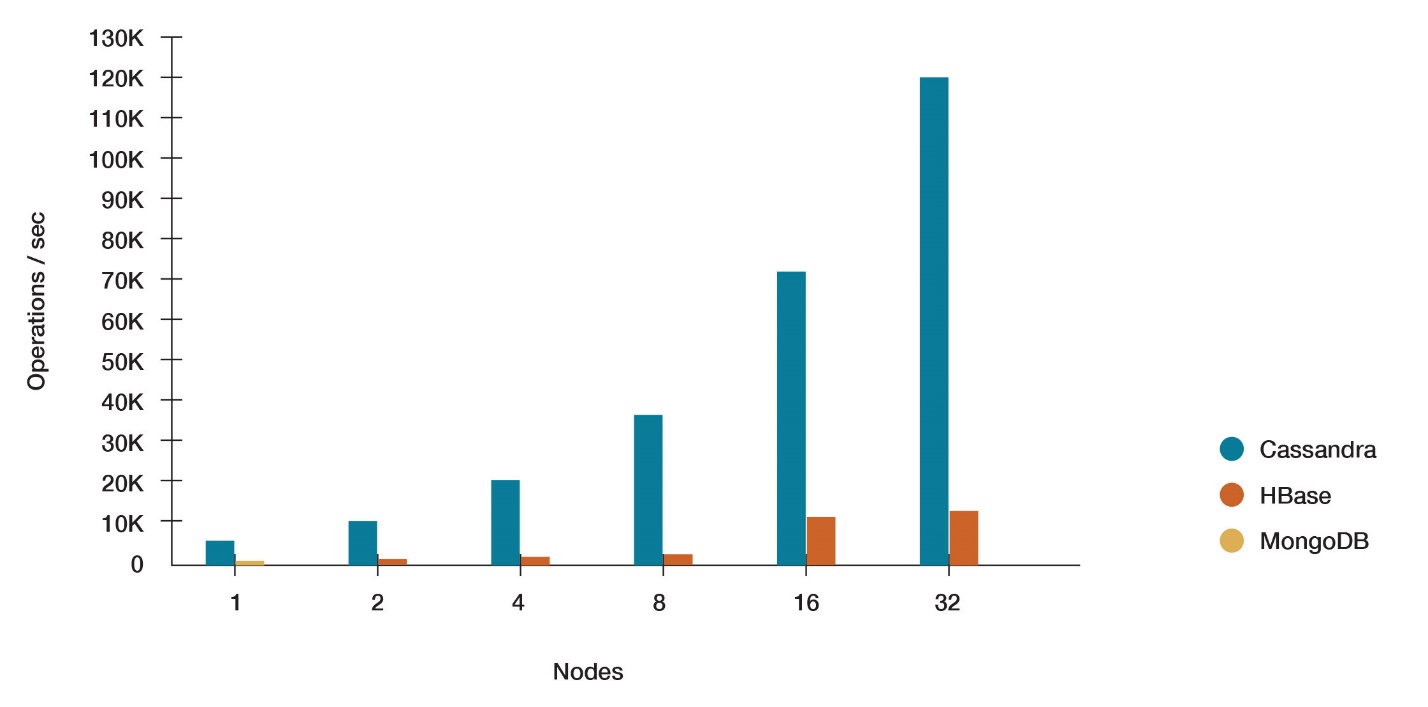


Table 1 (DataStax: always-on data platform | NoSQL | Apache Cassandra, 2018)

**Conclusion:**

Concluding, it is no doubt that big data is rising force in the industries today and the need for technologies like Cassandra, Hbase and MongoDB is essential. Data is being produced at a growing rate and with technologies like Artificial intelligence and Machine learning it is no doubt that the rate will grow even more. Since the inception of the internet, newer technologies are creating innovative mechanisms to bring useful products and services that the industries can use. Cloud Computing is another factor where virtualization has allowed excellent data management.

**References:**

DataStax Academy: *Free Cassandra Tutorials and Training*. (2018). What is Apache Cassandra. [online] Available at: https://academy.datastax.com/planet-cassandra/what-is-apache-cassandra [Accessed 2 May 2018].

MongoDB. (2018). *What Is MongoDB?.* [online] Available at: https://www.mongodb.com/what-is-mongodb [Accessed 2 May 2018].

Techopedia.com. (2018). *What is Big Data? - Definition from Techopedia*. [online] Available at: https://www.techopedia.com/definition/27745/big-data [Accessed 2 May 2018].

Oracle.com. (2018). *What is Big Data? | Oracle*. [online] Available at: https://www.oracle.com/big-data/guide/what-is-big-data.html [Accessed 2 May 2018].

Big Data Made Simple - One source. Many perspectives. (2018). *Top 10 tools for working with big data for successful analytics developers*. [online] Available at: http://bigdata-madesimple.com/top-10-tools-for-working-with-big-data-for-successful-analytics-developers-2/ [Accessed 2 May 2018].

data, T. (2018). *Top big data technologies used to store and analyse data - Crayon Blog*. [online] Crayon Blog. Available at: http://www.crayondata.com/blog/top-big-data-technologies-used-store-analyse-data/ [Accessed 2 May 2018].

Anon, (2018). [online] Available at: https://www.linkedin.com/pulse/real-comparison-nosql-databases-hbase-cassandra-mongodb-sahu/ [Accessed 2 May 2018].

DataStax: always-on data platform | NoSQL | Apache Cassandra. (2018*). NoSQL Comparison Benchmarks*. [online] Available at: https://www.datastax.com/nosql-databases/benchmarks-cassandra-vs-mongodb-vs-hbase [Accessed 2 May 2018].