BVG Trust

J. Sci. Trans. Environ. Technov. 2016, 9(3) : 115-118

Scientific Transactions in Environment and Technovation

Big Data Analytics-A revolution in the digital world

P. Shanmugavadivu*

*Department of Computer Science and Applications, Gandhigram Rural Institute - Deemed University, Gandhigram – 624 302, Dindigul Dt., Tamil Nadu, India.

Abstract

'Big data ' is a popular, but poorly defined marketing buzzword, that describes the exponential growth, availability and use of information, both structured and unstructured. The big data trend and how it can serve as the basis for innovation, differentiation and growth. Looking at big data is that it represents the large and rapidly growing volume of information that is mostly untapped by existing analytical applications and data warehousing systems. Examples of this data include high-volume sensor data and social networking information from web sites such as Google, Face Book, LinkedIn, Yahoo, Amazon and Twitter. The exponential growth in the amount of biological data needed for data management, analysis and accessibility.

Keywords: social media, structured data, unstructured data, Hadoop, NoSQL, Data, warehousing

INTRODUCTION

This is the first time when the world is witnessing volumes of data, from so disparate sources, in very different frequency, bombarding everyone in such a way that, it is virtually impossible to understand or analyse every possible piece of data. Big Data Analytics is an effort in the direction of getting the data to make sense, to the right person, at the right time, to the right level, for the right purpose, at the right place. It is a means to help one make meaningful, informed decision.

A. Definition of Big Data Analytics

A collection of data that is huge (volume), of heterogeneous type (variety), and getting collected at a fast pace (velocity) is in a broad term referred to as Big Data. The process of probing Big Data to get some relationship, significance, and coherence that may be useful for decision making, is referred to as Big Data Analytics. Though volume alone is not the criteria for measurement, data of the size of minimum of 1 Petabytes (1,000 Terabytes) is usually considered as Big Data.

B. Explosion of Big Data

Data collection and analysis has always been happening though at an extremely small proportions as compared to the Big Data times. Whether it was a storage of the citizens' and the financial data of a country during the early 1900s, or storage of the various books, magazines, research papers, etc. during the mid-1900s, the data has been growing steadily. Naturally, as the need arises the means to process such data also is invented. However, it was only in mid-1900s an alarm was raised stating that as the size of libraries were doubling every 16 years, the storage and retrieval was getting increasingly difficult. With the expansion of scientific knowledge and technology advancements, the size of the storage devices shrunk and hence the capacity increased, and was cost-effective.

With increase in availability of cost effective technology, more businesses started using centralised processing data centres. As more businesses started generating, collecting, storing and analysing data, and with the booming broadcasting industry, more data started getting stored. It was during mid-1970s that the earlier forms of RDBMS was commercially available, which made the storage and retrieval of data extremely easier as compared to that during the earlier decades. This encouraged more usage, which was supported by the cost effective storage facilities that were becoming more as a norm. With increase in the number of applications, there was a need to consolidate the data for effective and meaningful analysis. The mid-1980s saw the start of data warehousing, and the coining of the term "Business Intelligence" in 1990.storage, retrieval, analysis,interpretation, reporting, usefulness, etc. started surfacing. It was in the year 1997 that the term "Big Data" was used to refer to such large volumes of data. In the year 1999, it was estimated that the world produced about 1.5 Exabytes (1 Billion GB) of data, and was growing rapidly. Extraordinary situations need extraordinary solutions and thus the Hadoop platform was created in 2006 to handle such huge volumes of disparate data that are stream in fast. The objective was to be able to store, analyse, and scale as per the need, without having a limit

During the early years, data was more confined to space and defence. At the next stage, data comprised of centralised accounting in organisations, government records, and the related areas. When the reach of

*Corresponding Author :
email: *psvadivu67@gmail.com*

computers improved, data increased to organisational level from multiple departments, government data resulting from automation of several functions, bank transactions, medical records, insurance records, organisation level e-mails, phone calls and records, television streams, etc. The following years saw open source data, chats, public e-mail facilities, instant messages, songs, etc. However, with the advent of social media, and with cost effective availability of high speed internet, the scene is left open to a deluge of data which can only be spoken in terms of Petabytes (1 Million Terabytes). This data includes comments, emotions, review, feedback, discussions, debates, votes, sentiments, photos, videos, images, documents, news, songs, speeches, e-mails, video calls, chats, instant messages, traffic, websites, urban services such as smart cities, IoT, geographical information, etc.

PROJECTION FOR DATA SIZE OF BIG DATA

It was estimated in the year 2006 that the world created about 161 Exabytes of data. It was projected that the data would double every 18 months and by the year 2010 the data size would be 6 times that of the size as in the year 2006 (about 988 Exabytes). However, the actual data was much larger, to about 1,227 Exabytes in 2010, and about 2,837 Exabytes in 2012. In the year 2008 it was projected that the data in 2015 would be about 50 times that of the data in 2006 (about 1 Zettabyte or 1 Billion Terabytes). It is now estimated that the world creates about 1,200 Exabytes of unstructured data every year and is growing, and most of that is being wasted without treating that to any meaningful analysis. It is estimated that by the year 2020, about 33% of data will contain information that might be valuable if analysed. At the moment, only about 3% of all data has been tagged and ready for manipulation, and only about 0.5% are treated through a meaningful analysis .

As the spending in infrastructure such as hardware, software and networking increases, the cost of investment per GB of data reduces. However, the cost of storage, security, etc. grows substantially. It is estimated that the by the year 2020, every person (man, woman, child) in the world would contribute on average of about 5.2 Terabytes of data, which

will add up to a total of about 40,000 Exabytes of data. The share of the emerging markets (Asia) will continue to increase at a more rapid pace as compared to that of the mature markets (USA, Western Europe, Japan, Australia, New Zealand). The challenge is not just in storage or such disparate and voluminous data, but even more in feeding the voracious appetite of consumers of data in forms of meaningful information, by processing them appropriately. It is estimated that about 4 Billion Terabytes of data would be stored and processed by the end of year 2016 . It is expected that the share of data from the emerging markets will surpass that from the mature markets very soon, by some estimates as soon as year 2017.

TRADITIONAL RDBMS ANALYSIS Vs BIG DATA ANALYTICS

Traditional RDBMS are not suitable for Big Data Analytics. Some of the reasons are listed below:

1.RDBMS can store only structured data whereas Big Data by definition contains variety of data that are clearly unstructured.

2.RDBMS is in general used for transactional processing whereas Big Data is for batch processing because of the extent of time it takes.

3.Performance is far better in Big Data database as it has distributed computing capabilities, as compared to that in RDBMS, because of the sheer size of data.

4.Scalability is expensive in RDBMS in terms of infrastructure and also involves redesign if the data types don't match, which is not a guarantee in Big Data.

RDBMS and Big Data databases can complement but not replace each other.

APPLICATIONS OF BIG DATA ANALYTICS

Big Data Analytics can be used anywhere the characteristics of Big Data exist. Some of those include: Government, Business, Media, Healthcare, Education, Research, Sports, Banking, Airlines, Manufacturing, Internet of Things, and Science and Technology such as early warning systems, astronomy, geographical information, intelligence, surveillance and search engines.

RISKS AND CHALLENGES

Any technological advancement has its risks and challenges associated with it and Big Data implementations are no different. It is important that these are minimised and addressed at the first detection. The following can be listed as some of the risks and challenges:

1. Information is power – to the owners, and unfortunately even more to the hackers. Providing a fool-proof security mechanism and constantly monitoring and upgrading is a challenge.
2. In case of data leak, there is a high risk of monitory loss, especially if the data is from a bank, financial institution, and insurance companies.

3. Since most of the systems are closely integrated, providing a fault tolerant system is very important but is not easy, as providing redundancy increases the cost. Hence alternate mechanisms have to be worked out which are complex to manage and protect, and is a challenge.
4. The use of open source in several Big Data implementations is definitely a security risk.
5. A major challenge in Big Data implementations is the high cost involved in terms of storage and security.
6. Security vulnerabilities are not fully understood as this is a new and emerging technology, with numerous software being made available as open source.
7. Risk of being legally implicated due to a security loophole and data loss or compromise

SECURITY

The explosive growth in the volume of data is unfortunately not aligning with the security requirements. It was estimated that at least about 80% of all data is unprotected, which means only less than about 20% of all data is protected whereas at least about 35% needed protective measures. The data in the mature markets are better protected than that in the emerging markets, by only marginally.

Protection mechanism for data depends on the type of data. Some data (such as contact details and e-mail addresses) need minimal security whereas some others (sensitive data such as those related to identity, bank records, medical records, etc.) require the maximum possible security. However, this is easier said than done. In the earlier years, the systems were confined within an organisation and hence with a well-configured firewall, most of the security issues could be handled. However, with the advent of open source code in Big Data implementations, the security

With the rapid and huge exchange of data in social media, there is a real challenge in protecting the first two types of data above. The following are some of the security controls that may be needed

1. Maintaining, monitoring, and analysis of audit logs consistently across the enterprise in real time.
2. Using secure configurations of hardware and software, including secure versions of open-source software;using traditional security mechanisms such as firewall, strong user credentials and password policy, intrusion detection, antivirus, etc.
3. Providing access only to select group of users for data stores. For example, data ownership can be with the IT department, but the information ownership perhaps should be with the business owner.
4. Providing pre-defined automated reports instead of giving access to the data to the end users.

TRENDS IN BIG DATA ANALYTICS

The tools for Big Data Analytics are still emerging, but are evolving very quickly for the comfort of most business houses. The following are the trends :

1. Hybrid of on-premises and Cloud based Analytics, from the present in-house only database.
2. Hadoop becoming the new enterprise data operating system, which is an enterprise data hub wherein many different data manipulations and analytics operations can be performed by plugging them into Hadoop as the distributed file storage system.
3. Incremental build of large-scale database by loading all the data into Data Lakes to start with. This is opposite of the traditional database practice of designing before loading data. This will provide flexibility to the designers to build the structure based on the data analysis.
4. Improvement in Predictive Analytics where predictions can be made about unknown future events.
5. Availability of faster and better SQL on Hadoop for easier and cost-effective solutions.
6. Availability of more and better NoSQL (Not only SQL) databases to give more and better capabilities in an open-source database.
7. Deep Learning combined with Neural Networks is a set of machine learning techniques and is used in image recognition, speech recognition and natural language processing .
8. Hybrid Transaction Analytical Processing, which is processing the data and storing it in memory itself for quicker response. Traditional database storage may only be used in future to store data that are not used frequently.

TOOLS FOR BIG DATA ANALYTICS

The following open source tools are popular at present, but is evolving fast with more options:

1. Platforms and Tools: Hadoop, MapReduce, GridGain, HPCC Systems, Storm
2. Databases: Cassandra, HBase, MongoDB, Neo4j, CouchDB, OrientDB, Terrastore, FlockDB, Hibari, Riak, Hypertable, Blazegraph, Hive, InfoBright Community Edition, InfiniSpan, Redis
3. Business Intelligence Tools: Talend, Jaspersoft, Jedox, Pentaho, SpagoBI, KNIME, BIRT
4. Mining Tools: RapidMiner, Mahout, Orange, Weka, DataMelt, KEEL, SPMF, Rattle
5. File Systems and Programming Languages: Gluster, Hadoop Distributed File System, Pig, R, ECL
6. Tools: Transfer and Aggregate: Lucene, Solr, Sqoop, Flume, Chukwa
7. Miscellaneous Big Data Tools: Terracotta, Avro, Oozie, Zookeeper

CONCLUSION

With technological advancements, low cost availability of hardware, open source software, and almost freely available huge data cache, the time is right to reap the benefits in terms of patterns, forecasts, productivity, and efficiency, which lead to better profitability and value addition to the stakeholder. The emerging trends are even more promising, with convergence of different disparate technology systems. No doubt there are challenges, but it is worth to take up the challenges in order to move ahead leaps and bounds.

REFERENCES

http://www.winshuttle.com/big-data-timeline/

http://www.cloudera.com/content/dam/cloudera/Resources/PDF/obson_IQT_Quarterly_ Spring_ 2010.pdf

https://www.domo.com/learn/infographic-the-physical-size-of-big-data

https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

http://www.theguardian.com/news/datablog/2012/dec/19/big-data-study-digital-universe-global-volume

http://www.computerworld.com/article/2473980/data-storage-solutions/data-storage-solutions-143723-storage-now-and-then.html

http://www.computerweekly.com/feature/How-to-tackle-big-data-from-a-security-point-of-view

http://neuralnetworksanddeeplearning.com/

http://www.computerworld.com/article/2690856/big-data/8-big-trends-in-big-data-analytics.html

http://www.datamation.com/data-center/50-top-open-source-tools-for-big-data-1.htm