THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

A Journey through Big Data Analytics

M. Kavitha
Assistant Professor, Department of Information Technology, SRM University, Chennai, India

Abstract:
The eon of Big Data has been arrived. Big data was the noted slogan of the year 2012. Before everyone realize and appreciate what is big data it went out of hill. In the current era most of the enterprises are flooding in flooding data.The increase in usage of mobile devices, medical data analysis, Web analytics, social websites and other types of emerging technologies is creating a new path for the research in big data analytics. With data in hand, one can start analysis, but where do you begin? And which type of analytics to be followed is most appropriate for your big data environment? This survey paper gives overall view of what is big data, different big data analytics used in real scenario and some of the tools used.

Keywords: *Lorem ipsum, dolor sit amet, consectetuer (key words) (Separate by semicolon) (9 pt)*

1. Introduction

The Data is also growing exponentially due to the outburst of machine-generated data like data records, web-log files, etc., and from growing human engagement within the social network like Facebook, twitter, LinkedIn, etc., [1]. In 2004, Wal-Mart claimed to have the largest data warehouse with 500 terabytes storage. In 2009, eBay storage amounted to eight petabytes. Before proceeding for the further discussion one should know how much is equal to how much bytes.

1KB (Kilobytes)	1024 Bytes
1MB (Megabytes)	1024 KB
1GB (Gigabytes)	1024 MB
1TB (Terabytes)	1024 GB
1 PB (Petabyte)	1024 TB
1EB (Exabytes)	1024 PB
1ZB (Zettabyte)	1024 EB

Table 1: conversions

A research study was made by IBM which says that IBM [3] estimates 2.5 quintillion bytes (2.5 exabytes) of data are created every day from a variety of sources. Futuristic analysis by CISCO says that the projected increase of global internet traffic in the middle of 2015 and 2016 alone is more than 330 exabytes, which is almost equal to the total amount of global IP traffic generated in 2011 (369 exabytes)" which is a unexpectable increase.

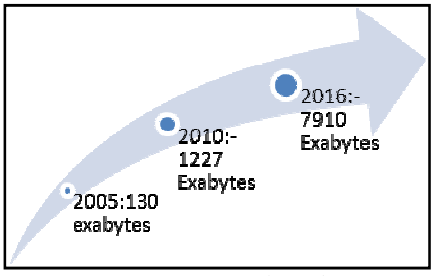

Figure 1: Growth of data

Big data is typically highlighted by 6 V''s

1. Volume: Storage of excessive data. Where massive volume of data is stored.
2. Velocity: Rate at which data is flowing. Data is flowing in at extraordinary speed.
3. Variety: combination of data formats(text, images, audio, video, etc.)

4. Veracity: uncertainty of data
5. Validity: correct data and accurate for the intended use.
6. Volatility: how long is data valid and how long should it be stored.

Before getting into big data analytics we will clearly understand what is data science, big data and big data Analytics.

	Data Science	Big Data	Data Analytics
What is what	Processing to extract knowledge or insights from data. Simply related to data cleansing and analysis	extremely large data sets that may be analyzed computationally to reveal patterns	Reveal patterns and decision making in business moves at right timeline.
Usage	Web search, advertisements	Mobile data, medical data	Health care, Government,
Skills required	Working with unstructured data, indepth knowledge in SAS and R, python and hadoop	Creativity and business skills	Mathematics, machine learning skills and data visualization skills

Table 2: Study on Data Science, Big Data and Data Analytics

2. Types of Analytics

Big data analytics is of three types

A. Descriptive analytics: the simplest class of analytics, that allows you to crush big data into smaller

B. Predictive analytics is projection of what might happen in future, because they are probabilistic in nature.

C. *Prescriptive* analytics: is done by comparing or making analysis with various information. It predicts what will happen, when it will happen and why it will happen, and then how to take value of predictive future

3. Challenges of Big Data

The challenges [3] include capture, store, search, sharing, transfer, analysis and visualization. challenges can be viewed in 3 dimensions 1.data 2. process and 3. management.

Data Challenges

1. Volume: How to deal with the size of big data?
2. Variety: How to handle multiple data types, formats etc.?
3. Velocity: How to react with information overflow with in a time period.
4. Veracity: data reliability, data quality, data availability: How can we handle uncertainty, missing values, etc.? How good is the data? How broad is the analysis? How sufficient is the sampling resolution? How timely are the readings? How well understood are the sampling biases?
5. Data discovery: How to find good quality or relevant data in the data flood?
6. Data Assumptions: Is any assumptions made? How the assumption is relevant to data processing?
7. Data comprehensiveness: What are the implications of uncovered data?
8. data Scalability: How far data scaling is done?

Process challenges

1. Data capture: how data is captured from different sources?
2. Data Arrangement: how data is brought into line from different sources?
3. Data Transformation: when and how data should be transformed form original data to data suitable for analysis?
4. Data modelling: what technique or Modelling or simulation is used?
5. Data visualization: how data is visualized? Is that a normal person can understand the output?

Management challenges: data privacy, security, governance and ethical issues.

a) Data privacy: how data privacy is preserved? What Level of privacy?
b) Data security: How secure is my enterprise data? How my data is used?
c) Data ethics: what legal issues is in my data? What about my ethical concern?
d) Data governance: how quality is my result? Any risk in applying any model?

Table 3

4. Levels of Big Data

big data analysis can be done in 4 levels

1. Educate: it deals with data assembly and business interpretations
2. Explore or discover: identify the relevant data from data source and spot the business needs and challenges.
3. Engage: use statistical modeling or any analytical technique to make some insights out of big data
4. Implementation or Execute: Deploy the analytics for more big data initiatives

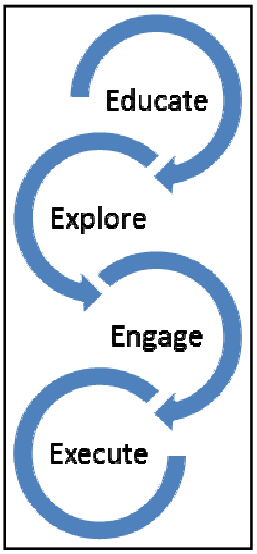

Figure 1: Levels of Analytics

Real time implementation of big data analytics and its outcomes

1. analyzing big data allow researchers to decode human DNA in minutes, predict where terrorists plan to attack, determine which gene is mostly likely to be responsible for certain diseases and, of course, which ads you are most likely to respond to on Facebook. This is called datamation
2. applying big data analytics to improve customer retention, help with product development and gain a competitive advantage.
3. In United States of America big data analysis played a large role in Barack Obamas successful 2012 re-election campaign. [3]
4. Decoding the human genome originally took 10 years to process, now it can be achieved in less than a day. The DNA sequencers have divided the sequencing cost by 10,000 in the last ten years, which is 100 times cheaper than the reduction in cost predicted by Moore's Law.
5. The NASA Center for Climate Simulation (NCCS) stores 32 petabytes of climate observations and simulations on the Discover supercomputing cluster[5].
6. The movie MoneyBall demonstrates how big data could be used to scout players and also identify undervalued players [4]

5. Tools Typically Used in Big Data Scenario

The majority of raw data, particularly big data, doesn't offer a lot of value in its unprocessed state. Of course, by applying the right set of tools, we can pull powerful insights from them. some tools are paid version and some are open sources. Sample tools are NoSQL: DatabasesMongoDB, CouchDB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riak, ZooKeeper. MapReduce: Hadoop, Hive, Pig, Cascading, Cascalog, mrjob, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum. Storagetools: S3, Hadoop Distributed File System. Servers tools: EC2, Google App Engine, Elastic, Beanstalk, Heroku. Processing tools: R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, BigSheets, Tinkerpop.

6. Tools Used in Big Data Analytics

Before you can really mine your data for insights you need to clean it up. Even though it's always good practice to create a clean, well-structured data set, sometimes it's not always possible. Data sets can come in all shapes and sizes (some good, some not so good!), especially when you're getting it from the web. The tools below will help you refine and reshape the data into a useable data set.

Tool name	Purpose
Open Refine	an open source tool that is dedicated to cleaning messy data.
Data Cleaner	transforms messy semi-structured data sets into clean readable data sets
RapidMiner	for predictive analysis.
IBM SPSS Modeler	text analysis, entity analytics, decision management and optimization
Qubole	Qubole simplifies, speeds and scales big data analytics workloads against data stored on AWS, Google, or Azure clouds.
BigML	BigML is attempting to simplify machine learning.
Statwing	takes data analysis to a new level providing everything from beautiful visuals to complex analysis.
Tableau	create maps, bar charts, scatter plots and more without the need for programming. They recently released a web connector that allows you to connect to a database or API thus giving you the ability to get live data in a visualisatio
Silk	It allows you to bring your data to life by building interactive maps and charts
CartoDB	is a data visualization tool that specializes in making maps
Chartio	allows you to combine data sources and execute queries in-browser
Datawrapper	It's an open source tool that creates embeddable charts in minutes
Import.io	Import.io is the number one tool for data extraction.

Table 4: Usage of Different Tools

7. Conclusion

Deem Big data is simply a matter of size, and analytics provides the opportunity to find insights in new and emerging types of data and content. This paper describes a systematic flow of survey on the big data Analytics. Finally, the most enormous question "why big data Analytics?" is answered. Massive data scalability and low latency data access and decision making is the final solution of big data Analytics.

8. References

i. IDC iView "Extracting Value from Chaos," June 2011, sponsored by EMC.
ii. IDC iView "Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East," December 2012, sponsored by EMC.
iii. IBM Global Business Services an Executive Report 'Analytics: The real-world use of big data"
iv. Rich Miller. "The Lessons of Moneyball for Big Data Analysis".*www.datecenterknowledge.com*. Retrieved 12 December 2015.
v. "NASA - NASA Goddard Introduces the NASA Center for Climate Simulation". *Retrieved 13 April 2016.*
vi. Delort P., OECD ICCP Technology Foresight Forum, 2012
vii. Marianela, Ulrik, Magnus, Vladimir Vlassov "Towards Automatic Veracity Assessment of Open Source Information" 2015 IEEE international congress on big data.