Big Challenges? Big Data …

Sahil R. Kalra
Student,
Pune Institute of Computer Technology,
Pune, India.

Aarati Mahajan
Lecturer,
Vivekanand Education Society's Polytechnic,
Mumbai, India.

ABSTRACT

In today‘s world, every tiny gadget is a potential data source, adding to the huge data bank. Every day, we create 2.5 quintillion bytes of data – structured and unstructured, so much that 90% of the data in the world today has been created in the last two years alone. This data generated through large customer transactions, social networking sites is varied, voluminous and rapidly generating. All this data prove a storage and processing crisis for the enterprises. While more data enables realistic analysis and thus help in making accurate business decisions / goals, it is equally difficult to manage and analyze such a huge amount of data.

This document provides insights on the challenges of managing such a huge Data – popularly known as Big Data, the solutions offered by Big Data management tools/ techniques and the opportunities it has created.

General Terms

Big Data, Big Data Opportunities, Big Data Challenges

Keywords

Hadoop, ParAccel, Storm, GridGain, MapReduce, HPCC Systems, Hortonworks, Dremel, Apache Drill, Calpont, Oracle TimesTen, GreenplumHD, Zettaset, Stratosphere, Big Data, Big Data Challenges, Big Data Analytical Tools, Big Data and Traditional BI Analytics, Big Data Opportunities, Three V's of Big Data

1. INTRODUCTION

Most people today have accounts in different social networking sites such as Facebook, Twitter, Instagram, LinkedIn, etc. In Facebook alone there are 1.393 billion monthly active users among which 890 million are daily active users. On an average, 350 million photos are uploaded and 4.75 billion items are shared in Facebook on a daily basis. As per the statistics of November 2014, Facebook stores 300 petabytes of data and takes in an average amount of 600 terabytes of data each day. This data never gets deleted. Instead, it increases in such a way that the rate of increase in the data itself gets increased. Such large amounts of data are termed as Big Data.

Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. And big data may be as important to business – and society – as the Internet has become. Why? More data may lead to more accurate analysis. The Business intelligence community is facing a greatest analytics challenge in addressing the data management issues of larger volume of data. The volume, velocity, and variety of data are growing exponentially for every microsecond. All the organizations need to manage the large volume of dataset, extract the value and knowledge from the data if they want to compete. Organizations are now starting to understand the potential that Big Data has to drive unprecedented improvements in their business. But they are struggling with how to turn raw Big Data into real value in their organization.

“Big data” word itself describes that it is a huge amount of data, but this is not the complete explanation of big data, if you want to understand it properly. For a complete understanding of big data you have to study all the basic properties of it.

2. CHARACTERISTICS OF BIG DATA

The Three V's of Big Data are as following:

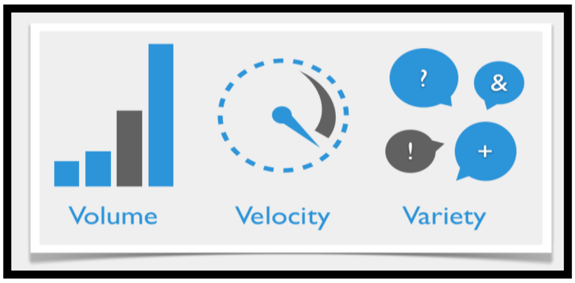


Fig 1: Three V's of Big Data

2.1 Volume (Scale)

Big data implies enormous volume of data. Volume defines the quantity of Big Data. The size of this data ranges from terabytes and petabytes, to even Exabyte. Few companies were generating data, all others were consuming data.

Now all of us are generating data, and all of us are consuming via machines, networks and human interaction on systems like social media the volume of data to be analyzed is massive. As the size of data increases, so does the difficulty in analyzing those data using traditional methods.

This brings about the need for introducing better and efficient methods to analyze Big Data.

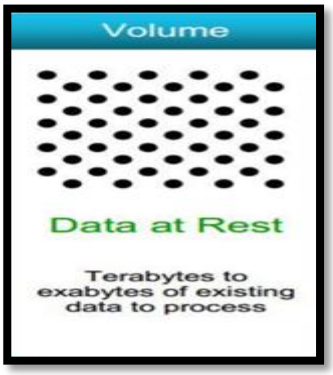


Fig 2: Exponential Increase in Generated Data

2.2 Variety (Complexity)

Variety is one of the most important properties of big data. The basic identification of big, data is based on a variety of data. Big Data may not contain only structured data as in traditional database systems. Variety refers to many sources and types of data both structured and unstructured. We used to store data from sources like spreadsheets and databases. Now data comes in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc.

The challenges involved in using such varieties of data are in storing and retrieving the data quickly and efficiently.

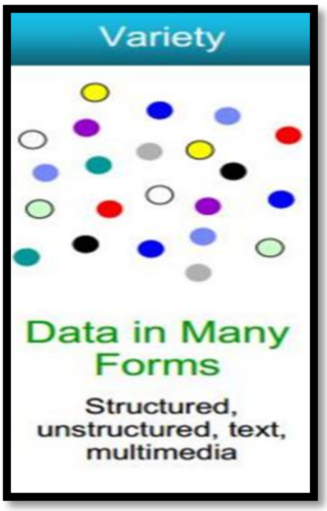


Fig 3: Various Formats, Types and Structures of Data

2.3 Velocity (Speed)

As the generation of data is rapid, the process of acquiring, processing and analyzing it requires fast mechanisms. Big Data Velocity deals with the pace at which data flows in from sources like business processes, machines, networks and human interaction with things like social media sites, mobile devices, etc. The flow of data is massive and continuous.

This real-time data can help researchers and businesses make valuable decisions that provide strategic competitive advantages. Velocity specifies the data that is in motion. In addition to this, velocity has a third dimension which specifies the speed at which the data should be stored and retrieved.

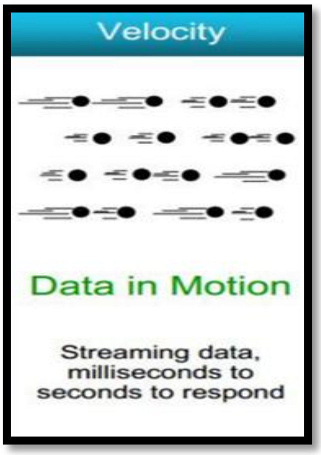


Fig 4: Data Being Generated and Processed Fast

3. KNOW & UNDERSTAND YOUR DATA

More accurate analysis may lead to more confident decision making. And better decisions can mean greater operational efficiencies, cost reductions and reduced risk.

Acquiring the huge amount of data is no more an issue but it is certainly what you do with the data that counts. Some of the major goals behind acquiring this voluminous data are 1) cost reductions, 2) time reductions, 3) new product development and optimized offerings, and 4) smarter business decision making.

For instance, by combining big data and high-powered analytics, it is possible to:

- Determine root causes of failures, issues and defects in near-real time, potentially saving billions of dollars annually.
- Optimize routes for many thousands of package delivery vehicles while they are on the road.
- Analyze millions of SKUs to determine prices that maximize profit and clear inventory.
- Generate retail coupons at the point of sale based on the customer's current and past purchases.
- Send tailored recommendations to mobile devices while customers are in the right area to take advantage of offers.
- Recalculate entire risk portfolios in minutes.
- Quickly identify customers who matter the most.
- Use clickstream analysis and data mining to detect fraudulent behavior

The past decade's successful web start-ups are prime examples of big data used as an enabler of new products and services.

For example, by combining a large number of signals from a user's actions and those of their friends, Facebook has been able to craft a highly personalized user experience and create a new kind of advertising business. It's no coincidence that the lion's share of ideas and tools underpinning big data has emerged from Google, Yahoo, Amazon and Facebook.

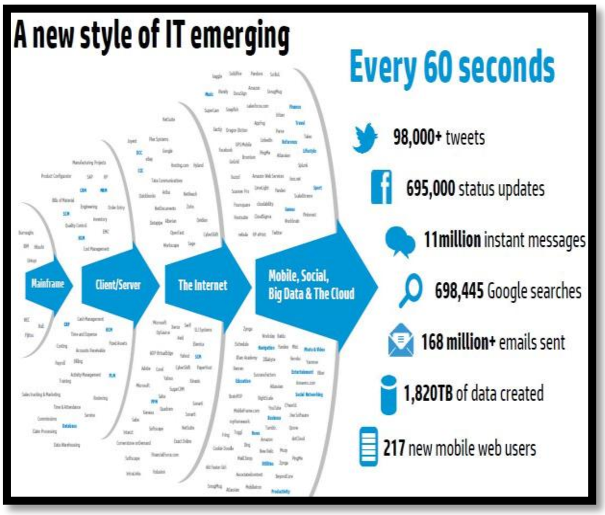


Fig 5: A new style of IT emerging

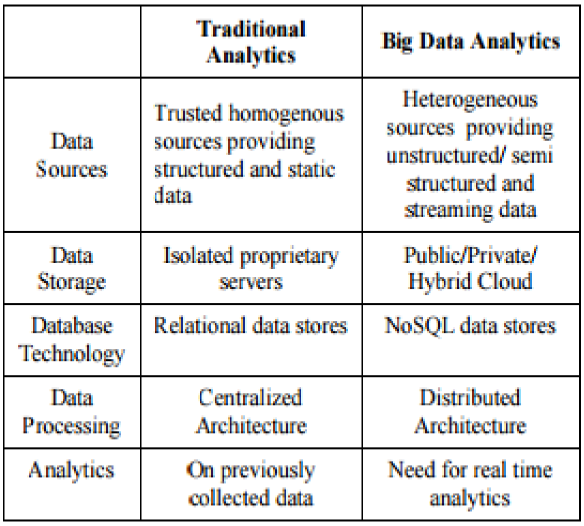
4. TRADITIONAL DATA ANALTICS V/S BIG DATA ANALYTICS

In Traditional Analytics, the analysis used to be done on the known data topography which was well understood. Most of the data warehouses have an elaborate ETL processes and database constraints, which means the data that is loaded inside a data warehouse is well understood, cleansed and in line with the business metadata. It is built on top of the relational data model, relationships between the subjects of interests have been created inside the system and the analysis is done based on them. But the traditional analytic systems are challenged as data becomes bigger, faster and increasingly unstructured. The massive volume of both structured and unstructured data is difficult to process using traditional database and software techniques. In most enterprise scenarios the data is too big or it moves too fast or it exceeds the current processing capacity.

The biggest advantage of the Big Data is that it is targeted at unstructured data outside of traditional means of capturing the data. Which means there is no guarantee that the incoming data is well formed, clean and without of any errors. In typical world, it is very difficult to establish relationship between all the information in a formal way, and hence unstructured data in the form of images, videos and mobile generated information etc. have to be considered in big data analytics which is aiming at near real time analysis of the data.

The technologies associated with big data analytics include NoSQL databases, Hadoop, ParAccel, Storm, GridGain, HPCC Systems, Hortonworks, Dremel, Calpont, Zettaset, MapReduce, GreenplumHD, Stratosphere and Oracle TimesTen etc. Which makes it more challenging but at the same time it gives a scope for much more insight into the data.

Table 1: Comparison between Traditional Analytics & Big Data Analytics



	Traditional Analytics	Big Data Analytics
Data Sources	Trusted homogenous sources providing structured and static data	Heterogeneous sources providing unstructured/ semi structured and streaming data
Data Storage	Isolated proprietary servers	Public/Private/ Hybrid Cloud
Database Technology	Relational data stores	NoSQL data stores
Data Processing	Centralized Architecture	Distributed Architecture
Analytics	On previously collected data	Need for real time analytics

5. CHALLENGES IN HANDLING BIG DATA

Big data analytics is a solution, not a product. To really get big data security analytics one needs to have a pretty deep understanding of technical elements like switching/routing, operating systems, logs, flows, IP packet metadata, DNS applications, DHCP, network/endpoint forensics, malware properties, malware behavior and known threat vectors.

Even if one knows all of the topics listed above, he/she still have to figure out how to glue it all together in the organization. What data should be collected? How one will collect it? Does one have the right processes and procedures to design, deploy, and operate big data security analytics? Where to start and how to proceed?

Though big data can yield extremely useful information, it also presents new challenges with respect to how much data to store, how much it will cost, whether the data will be secure and how long it must be maintained. For example, both companies and law enforcement agencies increasingly rely on video data for surveillance and criminal investigation. Closed-circuit television (CCTV) is ubiquitous in many commercial buildings and public spaces. Police cars have cameras to record pursuits and traffic stops, as well as dash-cams for complaint handling. Because all of these devices can quickly generate a large amount of data, which can be expensive to store and time-consuming to process, operators must decide whether it is more cost-effective to let them run continuously or only capture selective images or scenes.

Another major challenge of big data is preserving individual privacy. Since big multimedia datasets become commonplace, the boundaries between public and private space will blur. During day to day activities, the digital footprints are left behind most of the time, which when combined, could denote unique aspects about individuals. Big data analytics will draw on aspects of home, work and social lives to make assumptions beyond typical market segmentations and delve deep into ontological questions such as, "Who are you?". Life can still continue with many of these uncertainties for now with the hope that the benefits of big data will outweigh the harms, however an individual shouldn't blind oneself to the possible irreversibility of changes whether good or bad to society.

6. BIG DATA – TOOLS & TECHNOLOGIES

The major challenge is that as things stand today, there is no single technology that can cope with all the characteristics of big data – volume, velocity and variety all at once. Big data analytics is the application of advanced analytic techniques to very big data sets. Advanced analytics is a collection of techniques and tool types, including predictive analytics, data mining, statistical analysis, complex SQL, data visualization, artificial intelligence, natural language processing, and database methods that support analytics (such as MapReduce, in-database analytics, in-memory database, columnar data stores).

Big Data analytics requires massive performance and scalability- common problems that old platforms can't scale to big data volumes, load data too slowly, respond to queries too slowly, lack processing capacity for analytics and can't handle concurrent mixed workloads.

There are already so many open source tools are available for Big Data processing and analytics. Some of the major players are –

6.1 Hadoop

Hadoop is the most well-known big data open source tool around at the moment. It supports data-intensive distributed applications that can run simultaneously on large clusters of normal, commodity, hardware. It is licensed under the Apache v2 license.

A Hadoop network is reliable and extremely scalable and it works according to the computational model MapReduce. Hadoop is written in the Java programming language and is used by a global community of distributors.

6.2 ParAccel

The ParAccel data analytics platform helps organizations enhance their performance with interactive big data analytics. This platform offers columnar storage, adaptive compression, in-memory processing, and on the fly compilation. Thus making it easy to work with and adaptable.

6.3 Storm

Storm, which is now owned by Twitter, is a real-time distributed computation system. It works the same way as Hadoop's batch processing as it uses a set of general primitives for performing real-time analysis. Storm is easy to use and it works with any programming language. It is very scalable and fault-tolerant.

6.4 GridGain

GridGain is an enterprise open source grid computing made for Java. It is compatible with Hadoop DFS and it offers a substitute to Hadoop's MapReduce. GridGain offers a distributed, in-memory and scalable data grid, which is the link between data sources and different applications. An open source version is available on Github or a commercial version can be downloaded from their homepage.

6.5 MapReduce

MapReduce was originally developed by Google but has now been adapted by many big data tools, among others Hadoop. It is a software framework and model that can process vast amounts of data parallel on a large system of different computer nodes. The MapReduce libraries have been written in many programming languages and it therefore can work with all of them. MapReduce can work with structured and unstructured data.

6.6 HPCC Systems

HPCC means 'high performance computing cluster' and was developed by LexisNexis Risk Solutions. It is a similar version of Hadoop, but it claims to offer 'superior performance'. There is a free and paid version available. It works with structured and unstructured data and it is scalable from 1-1000s of nodes. It therefore also offers high-performance, parallel big data processing.

6.7 Hortonworks

Hortonworks is a pure open source Hadoop Distribution system. It is built on top of Hadoop and it allows users to capture, process and share data at any scale and in any format in a simple and cost-effective manner. Apache Hadoop is a core component of the Hortonworks architecture.

6.8 Dremel

Dremel is an interactive ad-hoc query system, which is developed by Google. IT offers analysis of read-only nested data. The system is extremely scalable; to 1000s of PCs and petabytes of data. It can process a collection of queries over massive,trillion-row, tables in just a matter of seconds by combining multi-level execution trees and columnar data layout.

6.9 Apache Drill

Apache Drill is part of the Apache Incubator and it offers a distributed system to perform interactive analysis of large-scale datasets that are based on Dremel. At the moment it is still incubating but the goals is to eventually become a massive scalable platform that can process petabytes of data in seconds over up to 10.000 servers.

6.10 Calpont

Calpont is a database management platform that runs a scalable software only columnar database for analytical applications. InfiniDB, the name of the platform that is enabled via MySQL, enables fast loading and query times. It scales with any type of hardware as the analytical needs and data grows over time, due to its Massive Parallel Processing. It is made for business intelligence, data warehousing and read-intensive needs.

6.11 Oracle TimesTen

Oracle TimesTen In-Memory Database is a memory-optimized, relational database with persistence and recoverability. TimesTen was acquired by Oracle in 2005. It allows applications a high throughput, which is often necessary for database-intensive applications. All the data will be located in the RAM section and applications get access to TimesTen via standard SQL interfaces.

6.12 GreenplumHD

Greenplum HD allows users to start with big data analytics without the need to build an entire new project. Greenplum HD is offered as software or can be used in a pre-configured Data Computing Appliance Module. It exists as a complete data analysis platform and it combines Hadoop and Greenplum database into a single Data Computing Appliance.

6.13 Zettaset

Zettaset offers a Hadoop management platform that is built on enterprise software. It works with any open source Apache Hadoop-base distribution. The focus of Zettaset lays on high availability and security of the data, easy Hadoop deployment and lower operational IT expenses. Zettaset basically offers a simple tool for managing Hadoop clusters.

6.14 Stratosphere

Stratosphere is a distributed data processing runtime, similar to Apache Hadoop. It extends the known MapReduce programming model by more operators for improved performance. Most applications naturally fit the versatile data flow graph that connects these operators.

Developers and Data Scientists can write jobs using a Java or a Scala programming interface. Stratosphere runs in existing infrastructures, it can operate from HDFS, runs in YARN (Hadoop 2.2) and is very easy to install in existing clusters.

The system is Apache-licensed and actively developed and maintained.

7. BIG DATA OPPORTUNITIES

Though there are great challenges there are even great opportunities that big data presents. McKinsey calls big data "the next frontier for innovation, competition and productivity." The questions which were beyond reach in the past can be answered with big data. Big data offers multiple opportunities like insight and knowledge identify trends and use the data to improve productivity, gain competitive advantage and create substantial value for the world economy. The challenges with big data are limited compared to the potential benefits, which are limited only by our creativity and ability to make connections among the trillions of bytes of data we have access to.

8. CONCLUSIONS

The era of Big Data has just begun. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products.

9. REFERENCES

[1] Douglas, Laney. "The Importance of 'Big Data': A Definition". Gartner. Retrieved 21 June 2012.

[2] Jean Yan, April 9, 2013 ―Big Data, Bigger Opportunities Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems.

[3] Oracle: http://www.oracle.com/technetwork/database

[4] Digital Magazine Information Week, available at: http://www.informationweek.com/big-data-analytics.asp

[5] SAS: http://www.sas.com/en_us/insights/big-data/what-is-big-data.html

[6] SAS: http://www.sas.com/resources/asset/Big-Data-in-Big-Companies

[7] Oreilly: http://strata.oreilly.com/2012/01/what-is-big-data.html

[8] Tom White, "Meet Hadoop, " in Hadoop: The definitive guide, 3rd Edition, California: O"Reilly Media, 2012, ch.1, pp. 9-15

[9] What Is Apache Hadoop? Available: https://hadoop.apache.org/