BIG DATA, ITS CHALLENGES AND LARGE DATA STORAGE

PERCCOM 2017-19

Course: Cluster, Grid & Clouds

Course by: Andrey Shevel
ITMO University, St. Petersburg

Ijlal Ahmed Niazi
Ijlal.Niazi@student.lut.fi

Big Data:

Big data refers to data of high magnitude and complexity which cannot be treated using conventional data analytics. It is a term which explains the high volume of data that are both structured and unstructured.

With the turn of the century we have moved to a digital world. With the evolution of technology and the advent of mobile phone communication, the number of connected people increased dramatically. Then came the social media age which changed the world we currently live in by making it a global village. Almost everyone uses one or another type of social media application and thus remain connected with friends and colleagues all over the world. Now technology has gone a step further. Not only people are connected but devices are also connected and communicating with each other. This is the age of IoT and smart cities where every entity will be connected and sharing information between itself and other entities to make facilities better for the user and to reduce the impact on the environment.

All of these connections and communications between people and devices mean that data is transferred at a rapid rate. It is generated from the connected devices and sensors and sent over the network to somewhere, where it can be stored and processed. Considering the number of nodes connected across the globe, this generated data is huge in terms of volume. And it is important to derive some kind of information from this data through analysis. Both normal statistical models and analysis techniques cannot be applied to analyze such huge amounts of data being generated at such a high rate.

Following are some characteristics which indicate the growth of data in recent years:

- The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s [2]
- Data generated by devices would reach 507.5 zettabytes (ZB) per year (42.3 ZB per month) by 2019, 269 times higher than the amount of data being transmitted to datacenters from end-user devices and 49 times higher than total datacenter traffic. [3]
- There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days." –Eric Schmidt Google [4]

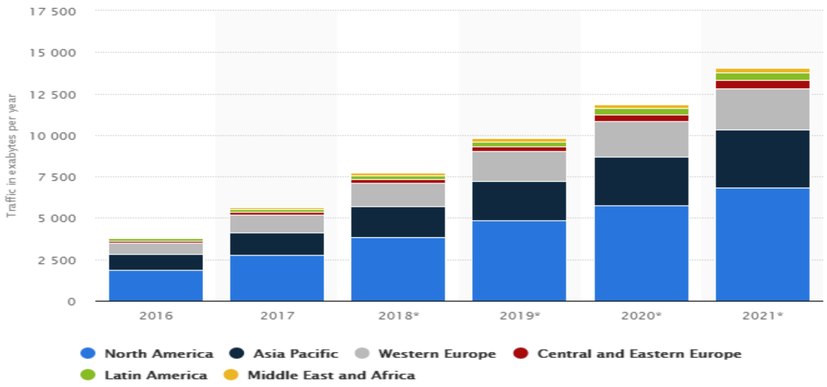
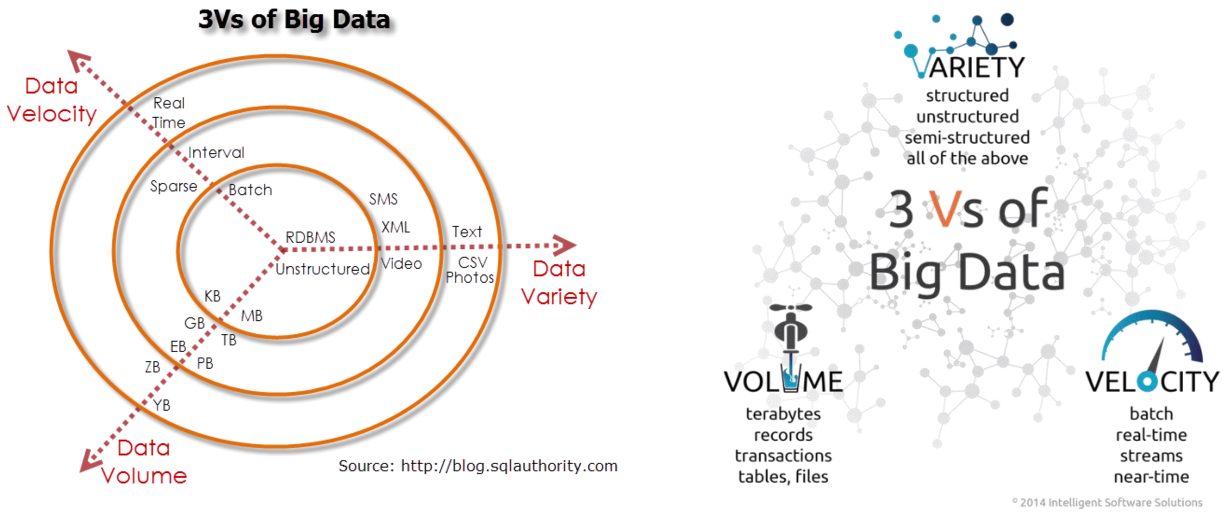


Figure 1: Global cloud traffic worldwide from 2016 to 2021

3 V Model for Big Data:

Big data is specified by mainly 3 characteristics. These characteristics are explained as the three V’s of big data:



Source: http://blog.sqlauthority.com

© 2014 Intelligent Software Solutions

Figures 2&3: 3 V Model of Big Data

Velocity:

Data comes from everywhere i.e. from multiple sensors, mobile devices, telecommunication networks etc. and it arrives at a very high rate. The data from such different sources means it can be of any type such as real time, in batches or in data streams. Because of such different data types, we need lots of handling mechanisms to store and analyze such high-speed data.

Volume:

The amount of data that is generated by various sources has reached the threshold of petabytes. Different sources generate big data every day ranging from commerce and education to social media and telecommunication devices. For example, AT&T transfers about 30 petabytes of data through its network each day

Variety:

Another main feature of big data is its variety. As it is generated from a number of different sources, it can be of any type. It ranges from personal files such as photos and videos to data coming from sensors of various types to signals received from satellites. It can be structured numerical data gathered in various databases to unstructured text documents email and transaction data.

In addition to these three characteristics, various others have been used to describe big data such as veracity and data value among others.

Sources of Big Data:

The sources of big data are as different and versatile as big data itself. They span different fields of everyday life. In education and research, data is generated each day by various experiments being conducted all over the world in different fields of science. One such example is the data generated at CERN by the experiments conducted at the Large Hadron Collider. Business and economics is another field where data is constantly generated at a rapid pace with the ever-changing stock indexes and millions of transactions being carried out each second. Other sources include communication infrastructures and social media platforms where people communicate and share data everyday such as Facebook, Google, Twitter etc. Such companies and organizations deal with gigabytes and petabytes of data on a daily basis.

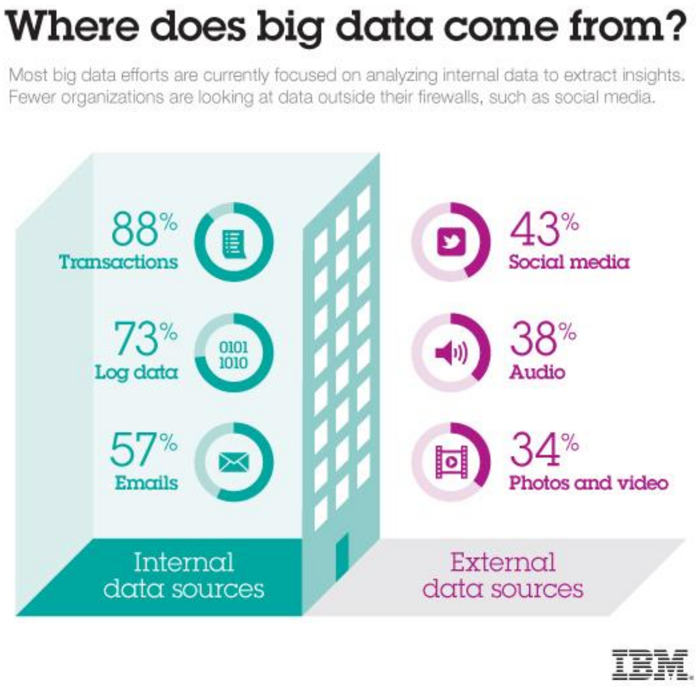


Figure 4: Generation of big data

Challenges in Big data:

The main characteristics of big data are also its challenges. Due to the huge size of data it is hard to gather, transmit and store. As it comes from various sources and different types, being structured, unstructured and real time, it is hard to devise models to process it and store it. The speed of generation also makes it a challenge for the data to be stored and processed in real time as we need more resources to store and derive value from it while at the same time it is being constantly generated. Next, we discover the challenges involved in various stages of big data:

Challenges in Storage:

There are some fundamental features that big data storage systems must have. Some of these features are as follows:

- Scalability: The storage systems must be scalable to account for the new data being produced. As mentioned earlier, the speed of data being generated is very high and this new data is constantly sent to the storage systems. These systems must be dynamic enough so that they can account for new data and provide the necessary resources. And in case of loss of data or data deletion, the storage systems should be able to be efficient and provide resources for data with higher priority.
- Input/Output rates: The storage systems must be efficient enough to cope with the high demands coming from high volumes and high velocity. They should be fast enough to save and move data around to analytics system so that new data can be preserved. Depending on the application, the storage systems must provide specific read/write speeds. If these requirements are not fulfilled, then the constant incoming data might overwhelm the systems which can result in data loss. This data loss can be disastrous for organizations.
- Backup and archiving: In order to prevent loss of data, backups and archiving is essential. But it becomes a problem with big data because of the sheer size and rates of acquisition. On one hand backups may be essential as the loss of data might be drastic and cause financial losses or loss of valuable information, on the other hand this data increases the resources required to store it and hence the expenses.

Current Storage Architectures:

- Data Centers are the main architectures in use for handling of big data. Cloud providers such as Drop-box, Google Drive, Google Docs, and Microsoft cloud provide storage for their clients in the form of Storage-as-a-Service (StaaS). To meet the challenges imposed by big data storage, cloud providers employ a huge number of commodity disk, across several data centers. For example, Amazon S3 aims to provide scalability, high availability, and low latency at commodity costs. The system provides efficient data loading from different sources, flexible data partitioning scheme, index and parallel sequential scan.
- Network Attached Storage (NAS) is designed for file sharing. It enhances storing, retrieving, and accessing files for applications and clients. It uses a high level abstraction that enables cross-platform data sharing. NAS comes with a processor and software for management and backup of data. It is efficient, reliable and comes with huge storage capacity.

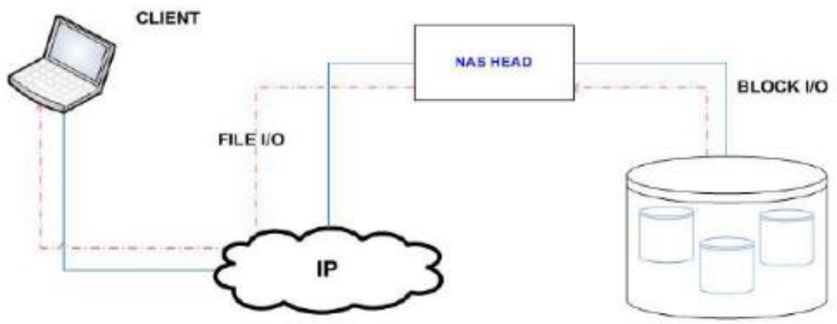


Figure 5: Architectural Overview of NAS

- Storage Area Network over IP (IP-SAN) provides the platform where thousands of computers connect to share a large amount of storage devices that range from simple disks to large, high-performance and high-functioning storage systems. The widespread of IP networks make it attractive to many organizations. It is less expensive and can span over a wide geographical area.

Challenges in Transmission:

Due to the high volumes, data transfer has become more of a challenge. Conventional networks which provide bandwidths of the order of hundreds of megabytes are not feasible for data of the order of petabytes. The data transfer itself would take too much time. In addition, such transfer networks need to be reliable and robust in order to prevent loss of data while transfer.
One solution to such data transfer issues is "edge computing". The theory of edge computing states that instead of transferring all the data to storage devices for analysis, some processing can be done at the data acquisition nodes such as sensors. After getting some information from the raw data, this can be transferred over the network for further analysis instead of transferring the whole volume of data. This will not only reduce the network load but also reduce the quantity to be stored.

Challenges in Processing:

After acquiring the data and storing it, the next step is to process it. Analysis of the data is fundamental as without it the data cannot be converted into information and no value can be extracted from it. But again, because the data is so huge and complex there are a lot of problems faced when efforts are made to analyze it. Normal statistical analysis methods are not developed to handle such large volumes and such high complexity of data. Hence, we cannot use them to generate information. So, we have to look beyond conventional methods and come up with new ways. Also, when such data must be stored for a long time for analysis at a later stage, the analysis tools developed might have already been outdated. Hence there is a need to revolutionize big data analysis along with other categories.

Processing of big data is a multi stage process and each stage has its own challenges to overcome. The following figure gives an insight into issues faced at each stage:

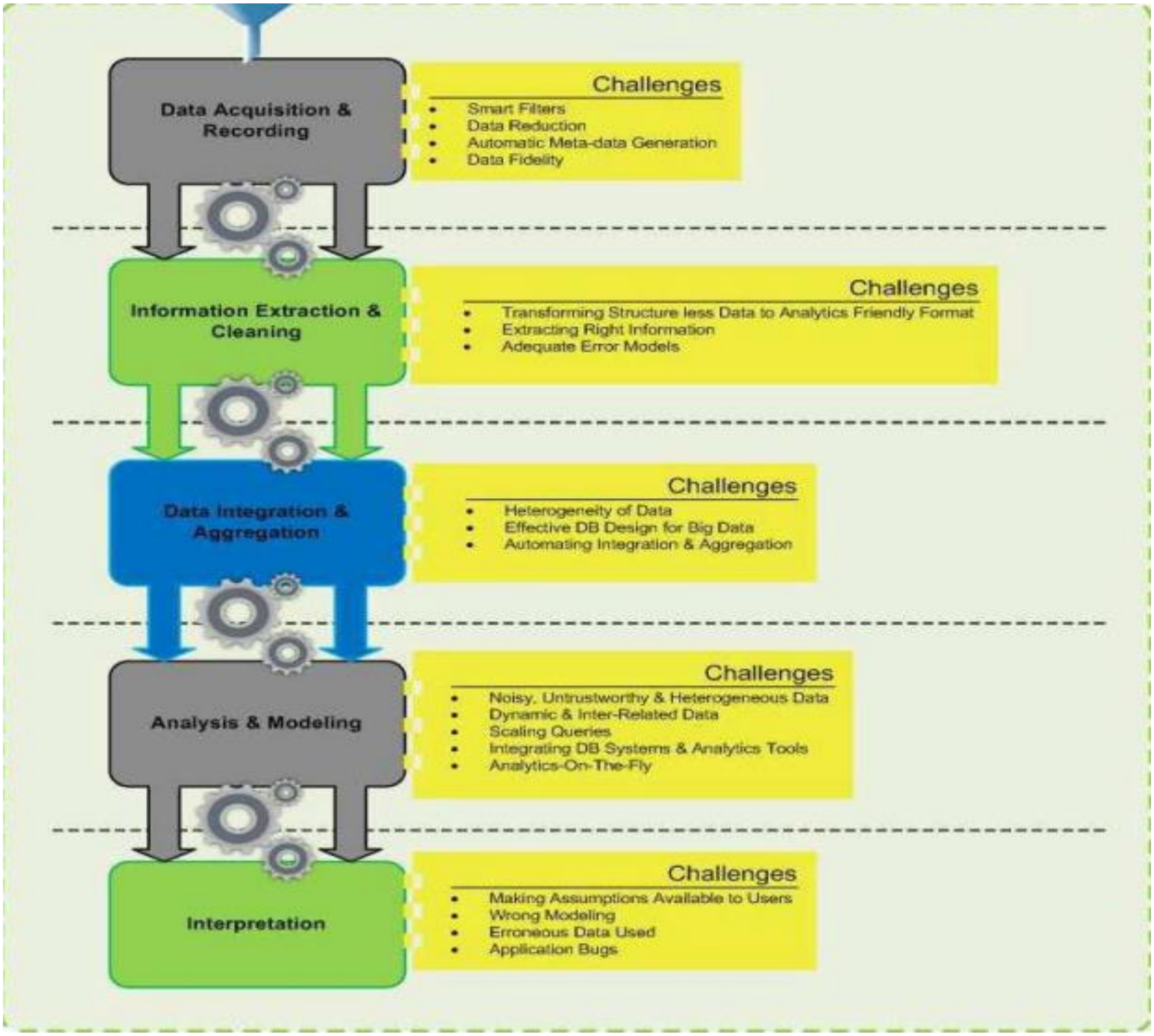

Data Acquisition & Recording
Challenges
Smart Filters
Data Reduction
Automatic Meta-data Generation
Data Fidelity
Information Extraction & Cleaning
Challenges
Transforming Structure less Data to Analytics Friendly Format
Extracting Right Information
Adequate Error Models
Data Integration & Aggregation
Challenges
Heterogeneity of Data
Effective DB Design for Big Data
Automating Integration & Aggregation
Analysis & Modeling
Challenges
Noisy, Untrustworthy & Heterogeneous Data
Dynamic & Inter-Related Data
Scaling Queries
Integrating DB Systems & Analytics Tools
Analytics-On-The-Fly
Interpretation
Challenges
Making Assumptions Available to Users
Wrong Modeling
Erroneous Data Used
Application Bugs

Figure 6: Big data processing and challenges

- The current solutions available for solving issues in big data processing involve machine learning methods to extract valuable information from data. It is important to introduce artificial intelligence in processing of huge amounts of data because it surpasses the abilities of human efforts.
- High performance computing clusters (HPCC's) with large amount of resources are also used to host such machine learning algorithms.
- There are also some online analytical processing (OLAP) tools available which make use of cloud infrastructures to provide processing capabilities for big data solutions.

Conclusion:

This report provides a general overview of big data and its characteristics. Various challenges involved in big data stages are highlighted with possible solutions. Large data storage and its challenges are the main focus of the report.

Big data is the data of the future. With IoT on the horizon there is no looking back from the technological revolutions happening right now. These revolutions allow the generation of huge volumes of data. And while this data is valuable in terms of its importance and ability to impact the world, there are a lot of challenges that must be overcome before it can be utilized to propel ourselves in the future.

References:

- J. Wu, S. Guo, J. Li and D. Zeng, "Big Data Meet Green Challenges: Big Data Toward Green Applications",*IEEE Systems Journal*, vol. 10, no. 3, pp. 888-900, 2016.
- J. Wu, S. Guo, J. Li and D. Zeng, "Big Data Meet Green Challenges: Greening Big Data", IEEE Systems Journal, vol. 10, no. 3, pp. 873-887, 2016.
- R. Agrawal and C. Nyamful, "Challenges of big data storage and management", Global Journal of Information Technology, vol. 6, no. 1, 2016.
- *Challenges of big data storage and management (PDF Download Available)*. Available from: https://www.researchgate.net/publication/298433319_Challenges_of_big_data_storage_and_management [accessed Jun 07 2018].