7/6/2017

Problems of storage, transfer and processing of Big Data

Course: Computing clusters, grids, and clouds

Lecturer: Andrey Shevel

Kristian Kouros
ITMO UNIVERSITY

1. Introduction

Massive, fast and diverse data moving quickly everywhere creating what is known as "Big Data" era. This data becomes very important source for valuable insights and ultimately helping to make more informed decision. Big data has existed as a term for a long time and it refers to data sets that are so large and complex that traditional data processing application software is inadequate to deal with them. This term has become very popular the recent years with the advancement of technology and introduction of new services.

The sources that produce this kind and amount of data are very diverse. Cheap and numerous information-sensing Internet of things devices such as mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers, Social Media and wireless sensor networks are some of the main contributors in Big Data generation. Usually this type of information is a complex combination of structured, semi-structured, unstructured, homogeneous and heterogeneous data. This variation of data needs be processed by advanced analytics and visualization techniques to uncover hidden patterns and find unknown correlations to improve the decision-making process.

2. Characteristics

The "3V's", Volume, Velocity and Variety, represent key elements that are considered vital regarding the characteristics of Big Data systems.

The first characteristic of Big Data, which is **Volume**, refers to the quantity of data that is being manipulated and analyzed in order to obtain the desired results. It represents a challenge because in order to manipulate and analyze a big volume of data requires a lot of resources that will eventually materialize in displaying the requested results. For example, a computer system is limited by current technology regarding the speed of processing operations. The size of the data that is being processed can be unlimited, but the speed of processing operations is constant. To achieve higher processing speeds more computer power is needed and so, the infrastructure must be developed, but at higher costs. By trying to compress huge volumes of data and then analyze it, is a tedious process which will ultimately prove more ineffective. To compress data it takes time, almost the same amount of time to decompress it in order to analyze it so it can be displayed, by doing this, displaying the results will be highly delayed.

Velocity is all about the speed that data travels from point A, which can be an end user interface or a server, to point B, which can have the same characteristics as point A is described. This is a key issue as well due to high requests that end users have for streamed data over numerous devices (laptops, mobile phones, tablets etc.). For companies this is a challenge that most of them can't keep up to. Usually data transfer is done at less than the capacity of the systems. Transfer rates are limited but requests are unlimited, so streaming data in real-time or close to real-time is a big challenge. The only solution at this point is to shrink the data that is being sent. A good example is Twitter. Interaction on Twitter consists of text, which can be easily compressed at high rates. But, as in the case of "Volume" challenge, this operation is still time-consuming and there will still be delay in sending-receiving data. The only solution to this right now is to invest in infrastructure.

Variety is the third characteristic of Big Data. It represents the type of data that is stored, analyzed and used. The type of data stored and analyzed varies and it can consist of location coordinates, video files, data sent from browsers, simulations etc. The challenge is how to sort all this data so it can be "readable" by all users that access it and does not create ambiguous results. The mechanics of sorting has two key variables at the beginning: the system that transmits data and the system that receives it and interpret it so that can be later displayed. The issue of these two key aspects is that they might not be compatible regarding the content of the data transferred between them. For example, a browser can send data that consists of user's location, favorite search terms and so on. Meanwhile, the Big Data system receives all this information unsorted, so it's difficult to for it to understand whether this user is from "London" or from "orange". To avoid this "mess" created in Big Data solutions, all systems that send data should be standardized so that can send data in a logical array that, afterwards, it can be easily analyzed and displayed in a proper manner. After the "3V's" another two "V's" where added as key aspects of Big Data systems.

The fourth "V" is **Value** and is all about the quality of data that is stored and the further use of it. Large quantity of data is being stored from mobile phones call records to TCP/IP logs. The question is if all together can have any commercial value. There is no point in storing large amount of that if it can't be properly managed and the outcome can't offer insights for a good development.

Veracity is the fifth characteristic of Big Data and came from the idea that the possible consistency of data is good enough for Big Data. For example, if A is sending an email to B, B will have the exact content that A sent it, if else, the email service will not be reliable and people will not use it. In Big Data, if there is a loss regarding the data stored from one geo-location, is not an issue, because there a hundred more that can cover that information. Current technologies software technologies try to overcome the challenges that "V's" raises. One of these is Apache Hadoop, which is open source software that its main goal is to handle large amounts of data in a reasonable time. What Hadoop does is dividing data across a multiple systems infrastructure in order to be processed. Also, Hadoop creates a map of the content that is scattered so it can be easily found and accessed.

3. Importance and Examples

The Big Data technology is highly advantageous as it provides business with three main values: *cost reduction, decision-making improvement and improvements in products and services*.

The *cost reduction* obtained from Big Data can be directly or indirectly. Directly some companies follow the idea that processing power (MIPs: million instructions per second) and storage capacities (terabytes) are cheaper if delivered by technologies of Big Data like Hadoop system. For example, one company made comparison between the annual costs of one terabyte on different systems and it found that it costs $37k for traditional database system, $5K for a database appliance and only $2K for Hadoop cluster. Indirectly the analysis of Big Data helped to make decision cutting the cost down. For example, the United Parcel Service (UPS), which is the global largest package delivery company, has been utilizing Big Data captured to track the movement of packages since 1980s and it maintain over 16 petabytes collected by sensor installed on about 46,000 vehicles. This data not used only to monitor UPS drivers' performance but mainly to redesign the route path of the drivers and reconfigure the pick-ups and drop-off operation in real time. This project is called On Road Integrated Optimization and Navigation (ORION) which saved UPS in 2011 more than 8.4 million gallons of fuel, about $30 millions.

A recent study found that 90% of executives believe that data becomes the fourth factor of production for business essential like land, labor and capital. Consequently, the term "*data-driven decision making*" comes to describe the process of collecting and analyzing data to guide or improve decisions. This involves the analysis of non-transactional and unstructured data like products ideas or reviews generated by consumers. In fact, data specialists explore the big data collected for example from social media to do field research to test a particular hypothesis and as results of that they can determine the value, validity and feasibility of these ideas and prepare the plans for executing them. Decision scientists are using several listening tools to conduct text and sentiment analysis and through these tool companies can measure certain aspects of interest about their products and taking necessary rectifying or improving actions. For instance, before a product put in market the marketing people would like to know how the consumer will feel about the price, how this sentiment will change from on area to another and how this feeling will change over time.

Based on the analysis of these tests' results the marketers can adjust prices to ensure high rate of marking of the product. Another example, Caesars a leading gaming company has embraced Big Data Technology to improve its decision-making process. The company is collecting data about its customers through sources like Total Rewards loyalty program, web click streams and real-time play in slot machines. It is using this data to understand customers, but the problem was how to interpret this data and act accordingly in real time while the customer is still standing at the slot machine. Caesars has realized that if a new customer was unlucky at the slots; it is likely he will never come back again. However, if the company present him, for example a free meal coupon before he left the slot machine, he is much more likely to revisit the casino again. The concept here is the requirement of real time analysis of the situation and to offer the coupon before the customer unhappily turns away. Caesars has implemented Hadoop clusters and necessary software analytics, of course in addition to hiring few data specialist to operate its analytics system.

The analysis process of Big Data comprised multiple phases including data acquisition and capture, extraction of information and cleaning, data integration, aggregation and representation, query processing and data modeling and analysis, and interpretation and presentation. Each phase has its own obstacles. Data is increasing and flowing very quickly generated by mobile devices, sensors, social media, emails, web site…etc. which are contributing to the Big Data explosion. However, organizations need to gather, store and drive value out of this stream of data which present group of challenges.

4. Big Data Problems

a. Storage

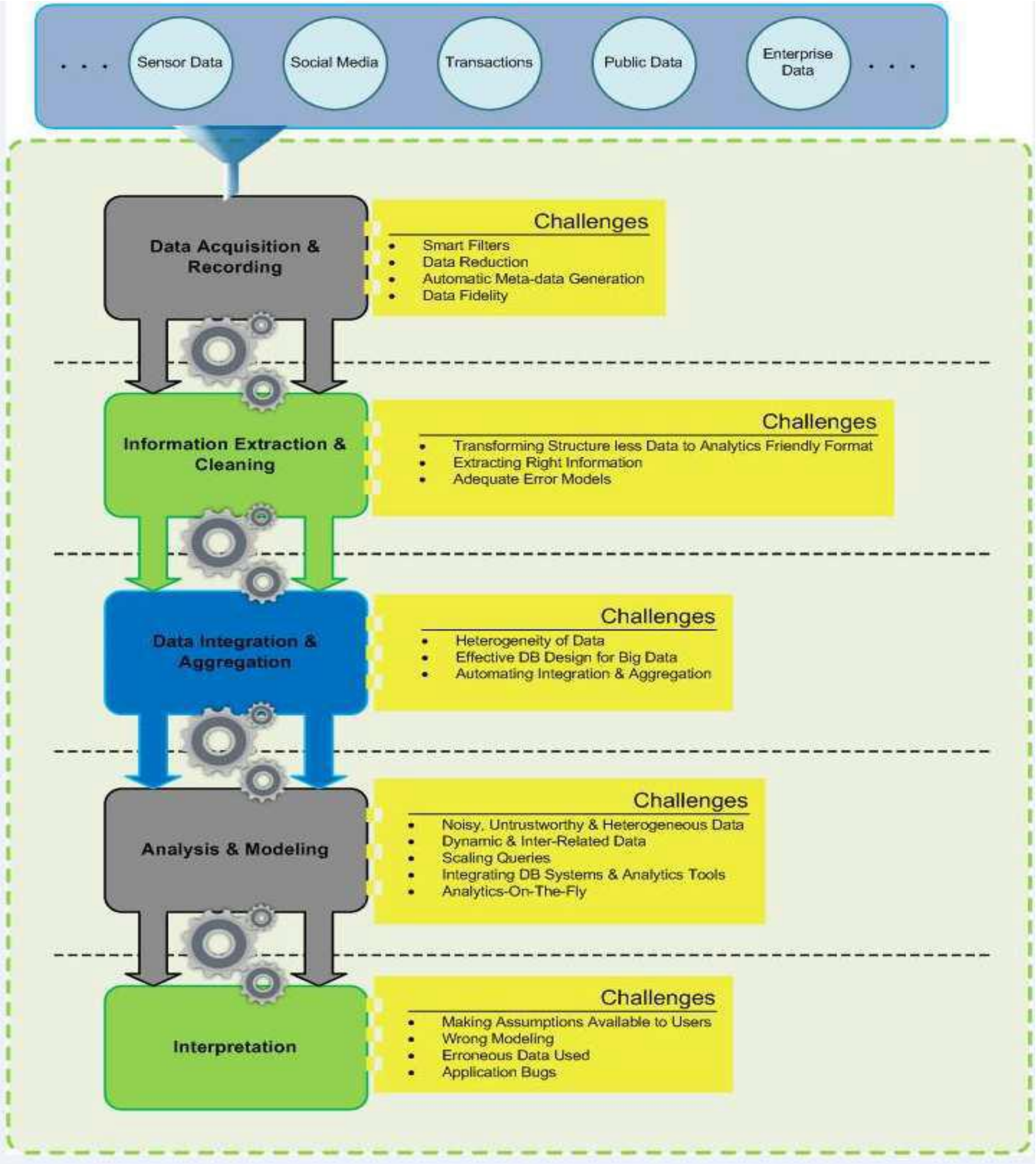
The quantity of data has exploded each time we have invented a new storage medium. What is different about the most recent explosion, due largely to social media, is that there has been no new storage medium. Moreover, data is being created by everyone and everything, not just, as before, by professionals such as scientist, journalists, writers, etc. Current disk technology limits are about 4 terabytes per disk. So, 1 exabyte would require 25,000 disks. Even if an exabyte of data could be processed on a single computer system, it would be unable to directly attach the requisite number of disks. The most suitable solution to Big Data storage is the usage of hyperscale computing environments that can be extended and be flexible based on our needs.

b. Transfer

Access to that data would overwhelm current communication networks. Assuming that a 1 gigabyte per second network has an effective sustainable transfer rate of 80%, the sustainable bandwidth is about 100 megabytes. Thus, transferring an exabyte would take about 2800 hours, if we assume that a sustained transfer could be maintained. It would take longer to transmit the data from a collection or storage point to a processing point than it would to actually process it! Two solutions manifest themselves. First, process the data “in place” and transmit only the resulting information. In other words, “bring the code to the data”, vs. the traditional method of “bring the data to the code.” Second, perform triage on the data and transmit only that data which is critical to downstream analysis. In either case, integrity and provenance metadata should be transmitted along with the actual data.

c. Process challenges

This group includes all the challenges encountered while processing the Big Data; starts with capture step and ends with presenting the output to clients, to understand the overall picture, as shown in the figure.


. . .
Sensor Data
Social Media
Transactions
Public Data
Enterprise Data
. . .
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

Generally, the process challenges are:

- *Data acquisition and recording*: Big Data does not come from space, there is should a source producing this data. We can sense anything around us, starting form measuring the heart rate of elderly citizen, to checking the existence of toxins in the air and to the global next-generation raid telescope known as kilometer square array telescope expected to generate one million terabytes per day. Likewise, nowadays the scientific experiments can generate petabytes of data. Are we interested in all this data? The answer is no, we can filter out and compress it by order of magnitude. How to define these filters is real challenge as they should be smart to distinguish between what is useful to capture and what is useless to discard. For instance, assume one sensor is giving readings different of the rest, this can be possibly caused by that the sensor is faulty, however how we can ensure that is not an artifact which needs further intention. Also, this data gathered from these sensors are most often spatially and temporally related, for example traffic sensors installed on the same street. Most research should conduct in the science of data reduction which can process the data into manageable volume and at the same time to preserve the user from being lost. In addition, on-line analytical algorithms are required to process the streaming data and to reduce data before storing it. The second major challenge in this domain is the automatic generation of the metadata that describes the data recorded and how it is recorded and measured. For Instance, in scientific experiments many details about the conditions and procedures are required to properly interrupt the results and it is essential to save this metadata with the observational data, special metadata acquisition systems are required to minimize human load. The data provenance becomes major issue since recording the origin of the data and it movement in the processing pipeline will help to determine the next processing steps which are clearly dependent on the current step. For example, if processing error happened at one stage, all the subsequent analysis will be useless. Therefore, research should be also conducted here to develop systems to generate suitable metadata and to carry out the provenance of data thought the various stages of data analysis pipeline.

- *Information extraction and cleaning:* The collected data is mostly not in the format required for processing. For example, the health records of hospital comprising of medical reports, prescriptions, readings captured from sensors and monitoring machines and image data like x-rays. Can we utilize this data effectively while they are in different forms? The answer is no. We need to build an extraction process that pulls out the required information from the Big Data source and formulate it in a standard and structured form ready for analysis. Creating and maintaining this process correctly is continuous challenge. The design of extraction process is highly dependable on the application area, for example the data you pull out from MRI is different of that pulled out of picture of the stars. Moreover, due to the ubiquity of the surveillance cameras and popularity of GPS-enabled devices like cameras, mobiles, navigators and other portal devices, rich and location dependent data can be extracted. The Big Data is not always telling the truth, it may carry some fake information. For example, patients may intentionally hide some risky behaviors or symptoms which might lead the physician to miss-diagnose the condition; or patients may sometime give wrong names of the drugs they were taking before which leads to inaccurate medical records. This will necessitate using data cleaning techniques which comprise of well-controlled constraints to valid data and well verified error models to ensure the quality of the data. However, quality-control models for most Big Data Models are still unavailable which represent another major challenge.

- *Data integration and aggregation:* The stream of Big Data is heterogeneous, so it is not enough to capture it and save in our repository. For example, if we take the data of several scientific experiments, it would be useless to save them as bunch of data sets. It is not likely that someone will find this data or include it in any analysis. However, if the data has adequate metadata, it might be used but the challenge still arise from the differences on the experimental details and the hosting data record structure. Data analysis is a sophisticated process and more than simply finding, identifying, understanding and citing data. Perform data analysis in large scale requires automating all these steps. This needs to express different data structures and semantics in form that computer can understand and then resolve automatically. A lot of work has been conducted in the field of data integration, however still more additional efforts required achieving automatic error-free different solution. The efficiency of the data analysis is mainly dependent on the database design. The same data set can be stored in different means, some designs will advantage over others for certain domains and possibly disadvantages for other domains, look for example the difference in the structure of bioinformatics databases hosting information about similar entities like genes. Database design becomes an art and people who are responsible for that role in big organizations should be highly paid experts. On the other hand, the domain experts can create effective database designs by themselves either to provide them with intelligent tools to help them in the design process or to entirely skip the design process and develop techniques to use database effectively.

- *Query processing, data modeling, and analysis:* Techniques to query and mine Big Data are significantly different of those used for the analysis traditional data sets. Big data is often noisy, unreliable, heterogeneous, dynamic and inter-connected data. However, the noisy Big Data might more useful than small samples of data since general statistics cab be extracted from repeated patterns and interrelation analysis usually overwhelm the individual variations and reveal more hidden knowledge. In addition, Big Data forms a large interconnect network of heterogeneous information, redundant information can be analyzed to compensate the missing data, to check unreliable relationships, to verify contradicting conditions and to disclose hidden models. There several requirements for data mining like cleaned, integrated, reliable and easily accessed data, declarative query interface, scalable mining algorithm and powerful computing environment. Simultaneously the data mining itself can assist to improve quality and reliability of the data, explain its semantics, and suggest intelligent query functions. As seen already medical records are of heterogeneous nature distributed across multiple systems and have errors. Here the importance of Big Data analysis is realized when applied in health care for example robustly considering all previous hard conditions. On the other hand, knowledge extracted during analysis and mining can help to correct errors and remove ambiguity Big Data is the main enabler for the next generation of interactive data analysis which provides answers in real-time. This new generation will enable querying the streams of Big Data for example the content of websites to populate hot lists, provide instant recommendations and to provide ad hoc analysis to decide if it worth to store or discard a dataset. The query process techniques should be developed to meet the scaling complexity of terabytes of Big Data and to enable interactive response time, more research need to be conducted. Analysts of Big Data complain the lack of coordination between the database system hosting the data and having SQL querying with analytics packages which conduct different type of non-SQL processing such data mining and statistical analysis. Nowadays analysts are delayed by the slow process of exporting data first from database, doing Non-SQL processing and finally importing data back to database. This represents a real obstacle for running the interactive option of the first generation of SQL-driven OLAP systems. Future analytics packages will have declarative query languages to enhance the performance of analysis and in turn to make the right decision on time.

- *Interpretation:* The analysis will be of limited value if it can't be understood by users. At the end of the day the result of analysis will be presented to the decision makers to interpret. It includes often testing all the assumptions made and reviewing the analysis. As shown already errors can emerge from different sources: faults in computer systems, assumptions made for the models and erroneous data on which the results based. End-user needs to understand and verify the results the computer systems generate and computer system must ease the job for the user. However due to the complexity of Big Data this becomes a challenge. Assumptions are there since the beginning of journey; important assumptions are made initially behind the captured data and also again through all various steps of the analytical pipelines are based on built-in assumptions. As example the recent mortgage-related crisis to the financial system clearly justifies the need for decision-maker diligence who should examine closely all possible assumptions at the various stages of analysis. In conclusion, it is not enough to give just the results; instead one must provide additional information describing how each step is derived and what inputs are used.

4. Conclusion

To conclude everything we have described earlier, we can say that Big Data is here to stay. In the current amounts, managing Big Data is feasible, but the predictions about the future situation suggest that new algorithms, techniques and technologies are required in order to cope with the vast generated data amounts, in a distributed, real time processing approach.

References

[1] “Big Data Challenges” Nasser and Tariq, J Computer Engineering Information Technology 2015, 4:3

[2] “Big Data in Cloud Computing: features and issues “, Pedro Caldeira Neves, Bradley Schmerl , Jorge Bernardino and Javier Cámara

[3] “Big Data: Issues and Challenges Moving Forward,” S. Kaisler, F. Armour, J. a Espinosa, and W. Money, 46th Hawaii Int. Conf. Syst. Sci., pp. 995–1004, 2013.

[4] "Big data", Wikipedia, 2017. [Online]. Available: https://en.wikipedia.org/wiki/Big_data