Big Data Challenges

Article · September 2015

DOI: 10.4172/2324-9307.1000133

2 authors:

Nasser Thabet
Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Dubai
2 PUBLICATIONS **51** CITATIONS

SEE PROFILE

Tariq Rahim Soomro
Institute of Business Management
120 PUBLICATIONS **2,136** CITATIONS

SEE PROFILE

Nasser and Tariq, J Comput Eng Inf Technol 2015, 4:3
http://dx.doi.org/10.4172/2324-9307.1000135

Journal of Computer Engineering & Information Technology

Research Article

A SCITECHNOL JOURNAL

Big Data Challenges

Nasser T* and Tariq RS

Abstract

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. However this data with very special attributes can't be managed and processed by the current traditional software systems, which became a real problem. This study will discuss all different challenges of Big Data categorized into three main groups: Data, process and management challenges. Data challenges are the group of the challenges relates to the characteristics of the data itself. Process group includes all the challenges encountered while processing the Big Data; started with capture step and ended with presenting the output to clients.

The management group comprises the legal and ethical issues related to accessing data. Layered architecture reference called "big data technology stack" will be presented as theoretical solution framework for the challenges of the Big Data. Each layer will provide the technologies required to overcome different challenge but collectively all these layers provide the complete solution. Continues evolution of technology necessitate innovating new Big Data analytics to dig more deeper into the data looking for more valuable insights and releasing new Big Data version 2.0.

Keywords

Big Data; Challenges; Layered

Introduction

The subject of this study is the challenges of Big Data and to realize those challenges two important questions need to be answered what is Big Data? And what are the characteristics of Big Data? The term "Big Data" is a little bit an inaccurate designation because it means that the preexisting data is small while it is really not and also it indicates that the data size is the one challenge we have [1]. Simply Big Data refers to the data and information which can't be handled or processed through the current traditional software systems. Big Data is large sets of structured and unstructured data which needs be processed by advanced analytics and visualization techniques to uncover hidden patterns and find unknown correlations to improve the decision making process. Many organization have huge volume of data, but they can't utilize it because it still in raw, semi-structured or unstructured format which is difficult to realize. Business faces a real challenge as data piles up, the percentage of data which can be processed is decreasing [2]. People now live in the Big Data era due to the rapid evolution of technology. Instrumentation helped us to sense everything around us and with the capability of saving the measured data fully or partially. Also advances in telecommunication and networking technology ease the interconnectivity of people and things, for example people can monitor remotely from office the project progress taking place thousands of miles away. Also the continuous drop down of electronic circuits helped to add intelligence everywhere [3]. For example railway companies start to install sensors in their cars to track the status of all components and all events faced and to reduce the possibility of train accidents. The readings are stored and analyzed to be used later to identify for example the parts which should be repaired or replaced before causing more damage [4]. Also intelligence added to rails themselves, sensors installed every few feet to provide online monitoring for all external events.

The roads are equipped with sensors to read all weather related events to avoid any expected disasters. The automated system covering the transportation network will generate huge, diverse and raw volumes of data which can't be fully processed by traditional systems [5]; here a Big Data problem is created. Initially Big Data was characterized by 3Vs: volume, variety and velocity, as shown in Figure 1 [6].

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 [7]. 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

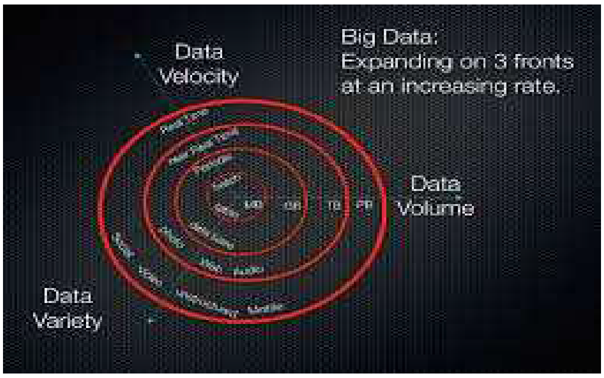

Figure 1: Big Data characterized by its volume, variety and velocity.

*Corresponding author: Nasser T, Department of Computer Science, SZABIST Dubai, United Arab Emirates, Tel: +971-43507376, +971-551231314; E-mail: Nasser.Thabet@omv.com, tariq@szabist.ac.ae

Received: July 16, 2015 Accepted: August 24, 2015 Published: August 31, 2015

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 [8].

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 [9]. 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 tool 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 [10]. 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 the 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 [11]. Also the creation of attractive products and services from Big Data is another major opportunities and there are many examples of successful products and services deriving from Big Data. For example, at LinkedIn there is a particular feature which has certainly added high value to the company is the people you may know (PYMK). Most of LinkedIn users have already make use of the PYMK feature which suggests to LinkedIn members some other members whom they many want to connect. The PYMK features collected these connections details by running multifactor approach to find out suggested members based on criteria like shared schools, connections, universities and geographies. The number of customers for LinkedIn has increased a lot by the PYMK feature. LinkedIn found that PYMK messages have achieved 30% click-through rate higher than other prompts sent to encourage people to revisit the site again [12].

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 [13]. 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 as seen in Figure 2 [14].

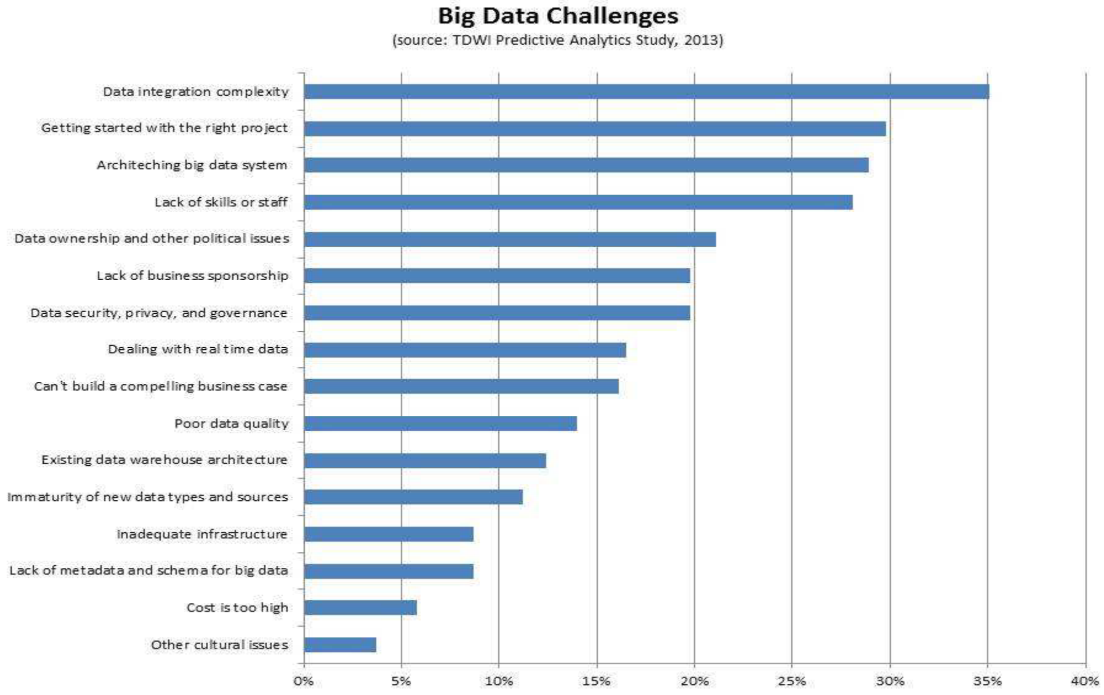

Figure 2:Big Data challenges.

doi:http://dx.doi.org/10.4172/2324-9307.1000135

The purpose of this study is to discuss all various types of challenges of Big Data and out of that to extrapolate and build layered reference architecture to be used as conceptual solution to efficiently manage Big Data regardless of all encountered difficulties.

After understanding what "Big Data" means and why it is important, it is time to know the challenges of Big Data. The study section 2 will discuss in details each challenge. Then in section 3 it will present all possible means to overcome each different challenge. After that in section 4 it will discuss the future of Big Data before it comes finally to the conclusion along with the recommendations and suggestions.

Challenges

Big data offer organization with massive insight; however terabytes or petabytes of data flowing every day to an organization have revealed that current infrastructures and architectures are not sufficient to meet the challenge. IT scientists are responsible to provide the technology capable of managing all technical requirements of tremendous streams of data. IT specialists are getting more calls as data grows; the requests are for more Ad-Hoc analysis and summarized reports. Decision makers can't wait for hours or days to find replies to queries if possible. Also end users will need means to access, understand and analyze this data by themselves without the need to return back to IT for every request [15]. These are examples of the challenges of Big Data which can be grouped into three main categories based on the data life cycle: data, process and management challenges. Data Challenges are the ones pertain to the characteristics of the data itself, for example data volume, variety, velocity, veracity, volatility, quality, discovery and dogmatism. The second group is the process challenges that are related to series of how techniques: how to capture data, how to integrate data, how to transform data, how to select the right model for analysis and how to provide the results. The third category is the management challenges which cover all privacy, security, governance and ethical aspects [16].

Data challenges

Data challenges are the group of the challenges pertains to the characteristics of the data itself and these characteristics are as follows:

Volume: The volume of data being stored is dramatically increasing every single minute, about 800,000 PB of data stored all over the world in 2000, expected to jump to 35 ZB in 2020. Facebook generates about 10 TB every day, Twitter generates about 7 TB and some enterprises generate terabytes every single hour [17].

So it becomes normal to have data capacity in terms of petabytes (PB).

Nowadays we are tracking and recording everything environmental data, business data, medical data, surveillance data, etc [18]. Due to instrumentation, automated machines are recording every event for example, ATM machine store all backing transactions, access doors store every access event, and airline systems store all check-in requests, Monitoring Cameras store over-speed violations. Therefore we have massive amounts of data which can't be managed by current traditional system. The challenge becomes clear here as the data flows in the enterprise is increasing and the percentage of data which can be processed is decreasing, thereby creating what is known as blind zone [19]. This zone indicate the data "you don't know" which can be of great importance or can be nothing at all, as shown in Figure 3 [1].

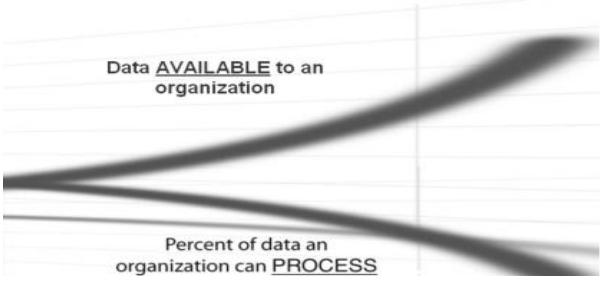

Figure 3: The volume of data today is on rise, while the percent of data they can analyze is on decline.

Variety: The massive volume of data caused by Big Data phenomenon presented new challenge which is the variety of data types and format. The tremendous spread out of sensors, smart devices and social collaboration technologies has made the data in the enterprise complex to deal with because it includes not only traditional data type but also raw, semi structured and unstructured data collected from WebPages, emails, social media forums, search indexes, audio and video …etc [20]. Only about 20% of data can be processed by current traditional systems and the remaining 80% are not analyzed and thereby not utilized for decision making and insight processes. Another challenge of Big Data appears here due to the variety characteristic [1].

Velocity: The velocity characteristic can be defined by the capacity of the current software application to handle and process data stream generated continuously and constantly at a pace which becomes critical due to the short shelf-life of the data which need to be analyzed in near real time if we plan to find insight in that data. This adds new challenge to the Big Data which needs to be analyzed while it is in motion [21].

Veracity: This refers to the biases, uncertainties, impression, untruths and missing values in the data. This feature measures the precision of the data and the possibility to use it for analysis. The correctness level of the data sets accumulating to our systems will determine how important this data for the problem being studied and some researchers believes this is the biggest challenge of Big Data [22].

Volatility: Data volatility denotes how long the data is valid, for how long we should keep it in our databases. Our world now is more relying on real-time data and it becomes important to know at which point the data is no longer applicable for analysis [23].

Quality: Quality characteristic measures how the data is reliable to be used for making decision. Saying that the quality of data is high or low is basically dependent on four parameters:

a) Complete: all relevant data are available, for example all details of vendors like name, address, bank account etc. exist

b) Accurate: data is free of misspelling, typos, wrong terms and abbreviations

c) Available: data is available when requested and easy to find

d) Timely: data is up to date and ready to support decision [24].

Discovery: This refers to how to filter out high-quality data pertained to the concerned problem from the massive non-stop stream of data [5].

Dogmatism: Valuable insights with be extracted out of the analysis of big data, but we must not be always obligated to the numbers. We should consult domain experts, apply common sense and react to events around us. If would not right for example to take preventative actions to flu outbreak only when Google Flu Trends told us [25].

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 Figure 4 [26].

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 terabyte per day. Likewise nowadays the scientific experiments can generate petabytes of data. Are we interested for 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

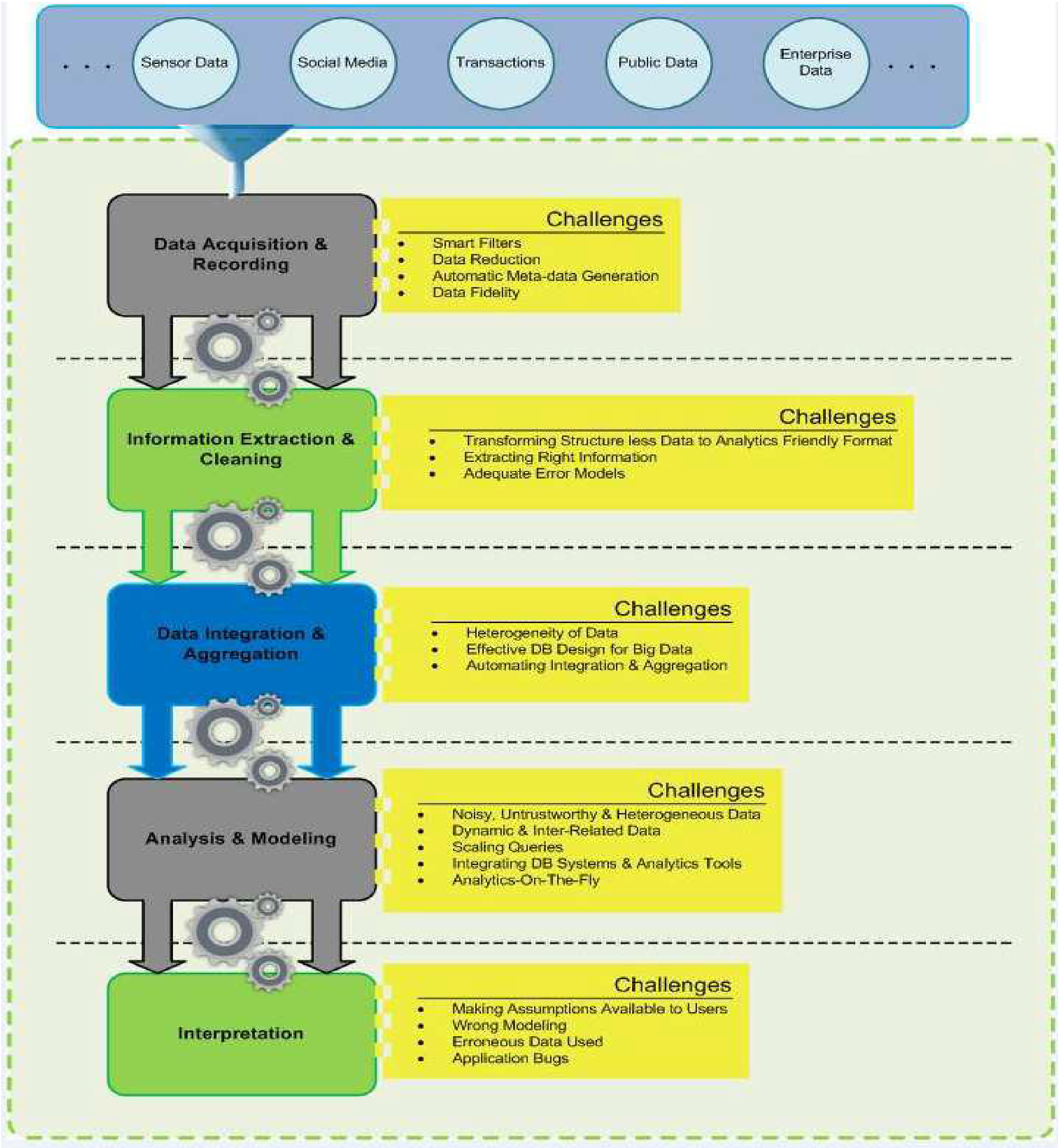
. . .
Sensor Data
Social Media
Transactions
Public Data
Enterprise Data
. . .
Data Acquisition & Recording
Challenges
Smart Filters
Data Reduction
Automatic Meta-data Generation
Data Fidelity
Information Extraction & Cleaning
Challenges
Transforming Structure less Data to Analytics Friendly Format
Extracting 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

Figure 4: Big Data processing pipeline and challenges.

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 [27].

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 mis-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 [28].

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 advantages 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 [29].

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. For instance a physician may diagnose a patient case as “DVT” which used to refer to both “Diverticulitis” and “deep vein thrombosis” which are two dissimilar medical conditions. Digging into related data such symptoms or medications will help to determine what the physician meant.

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.

doi:http://dx.doi.org/10.4172/2324-9307.1000135

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 [30].

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. This supplementary information is called the provenance of the data. However more research should be performed to find out the best methods to capture, store and query provenance in conjunction with the techniques to create adequate metadata. By this users will have the infrastructure required both to interpret the results of analysis and to redo the analysis with different sets of assumptions, parameters and data.

Software systems with rich visualizations choices are crucial to convey the results of queries in most understandable way for each different domain. Early business intelligence systems mostly presented the results in tabular format, nowadays analysts need to express their results in more suitable visualized format that help users during the interpretation.. In addition the system should a powerful interface enabling the user with a few clicks to drill down into each single bit of the data and understand its provenance which is vital to understand the data itself. It is not for users to display the results, but he need also to understand why they are getting those results. The data provenance for different phases of the analytical pipeline is too technical for many users to understand. One solution for this is to grant users the option to make small changes to the steps, for example to modify values for some parameter so the users can see the results of those changes. Achieving this will need special systems that provide flexible facilities to specify analyses [31].

Management challenges

This group tackles the legal and ethical issues related to accessing data:

Privacy: Privacy is a major concern especially when in the context of Big Data. For instance, in the health sector there are already laws that govern the privacy of the patients. There is an increasing fear of inappropriate use of personal data especially when combining this data from multiple sources. For example the data extracted from location-based services that asks subscribers to share their location resulting in clear privacy concerns. Some people think that hiding their identity alone without hiding their location would not properly address privacy concerns which are not true. The location-based service provider can conclude the identity of the subscriber by tracing subsequent location information. It resembles the user leaving a trail of packer crumbs behind him which could be linked to a certain office location or residence and therefore used to detect the user's identity. Also other types of private information such health details (e.g. frequent visit to cancer treatment center) or religious preferences (e.g. presence in a church) can be disclosed by monitoring movement of anonymous users and analyzing movement patterns over time. It is more difficult to hide user location than his identity because to utilize location-based services he/she needs to expose your location. Today many private information are shared through online services like Facebook, Twitter, etc., and until now many people don't understand what it means to share data, how this data can be associated together to come up with more personal details not intended to share [32].

Security: More companies are building big computing environment to store, aggregate and analyze the growing amount of Big Data. It becomes known that Big Data helps businesses to tailor their products and services according to customer needs and enhance enterprise efficiencies. Consequently the number of large repositories of Big Data has been increased with comparative increase of related security concerns. The impact of such security breach is big as suggested "Big Data" itself and criminal groups are targeting Big Data repository to gain big payoffs, imagine terabytes of data in those repositories which may include the company original jewels: customer data, employee data and business secrets. The recent security breach of Big Data cost the company about $1.1 billion, the loss will be much higher if financial institutions or healthcare providers are targeted.

- Despite the high value of Big Data target, securing Big Data has its own unique challenges which are not fundamentally different from those associated with traditional data. In fact there are incremental not fundamental differences between the security challenges of Big Data and traditional data which include:

- The data: special variety, velocity and volume attributes of big data amplifies the security management challenges than those address in traditional environment.

- The Infrastructure: The distributed nature of big data environments is another challenge since they more complicated and more vulnerable to attack.

- The Technology: security was not considered in mind when Big Data technologies like Hadoop and NoSQL were originally designed which creates vulnerabilities for authentication and network security [33].

Governance: The demand for big data continues to grow and the big data governance becomes critical issue. Big data is rich with personal information and confidential enterprise data, and data governance is required to ensure that information is highly secured. Big data governance is essential to ensure the quality of Big Data and consequently the quality of the processing and analysis built upon. Applying superior Big Data governance will help to make decision with confidence, to plan accurately for future , to avoid costs resulted from low quality data and need to re-do the work again, and provide big data reporting compatible with government standards like Sarbanes. The management of Big Data governance presents some challenges to the organizations. Big Data governance will require native compliance of the security protocol for big data computing-environment, however most organizations operate their own legacy engines which have their own proprietary security techniques and basically don't integrate and can't scale to meet the needs of big data [34].

Solutions

The task of getting useful insights out of Big Data is not easy. It is a matter of developing comprehensive environment that includes hardware, infrastructure software, operational software, management software and application programming interface (API) to provide fully functional model managing the Big Data. The conceptual representation of this environment represented as layered reference architecture is called big data technology stack, as shown in Figure 5 [35]. This technology stack provides layered functional model for Big Data in which each layer will present the proper technology to be used to tackle the special need and challenges of Big Data, now we will start descripting what is special for each of these layers for Big Data implementation and how it participated to provide an overall optimal solution [34].

Layer 1 - Redundant physical infrastructure (Data challenges)

It is mainly about the new technology infrastructure to overcome the challenges arising from data characteristics like high-volume, high data-variety, high-velocity. IT Physical infrastructure will provide necessary hardware systems with adequate storage, processing power and communication speed matching the requirements of Big Data [36]. The optimal IT infrastructure which will suffice your Big Data implementation will be clearly determined after you set your requirements against each of the following criteria:

Performance: This measure the responsive degree of system, as the system performance increases the cost of infrastructure increases

Availability: Do you need your system to be up and running for 24/7 with no interruption? If yes this means you need high availability infrastructure which is also expensive.

Scalability: You need to set the size of your infrastructure, the storage capacity and the computing power. Also you need to consider additional scale for future challenges.

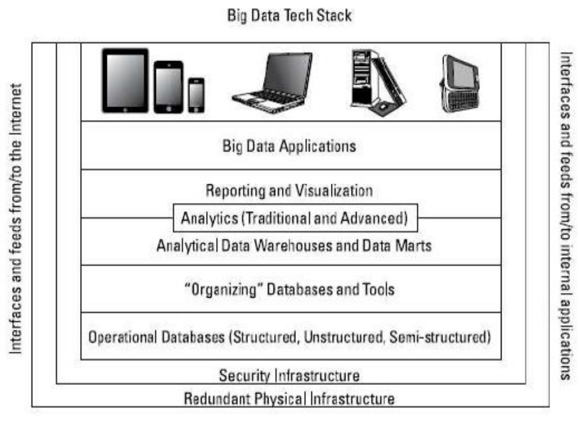
Big Data Tech Stack
Interfaces and feeds from/to the Internet
Interfaces and feeds from/to internal applications
Big Data Applications
Reporting and Visualization
Analytics (Traditional and Advanced)
Analytical Data Warehouses and Data Marts
"Organizing" Databases and Tools
Operational Databases (Structured, Unstructured, Semi-structured)
Security Infrastructure
Redundant Physical Infrastructure

Figure 5: The Big Data technology stack.

Flexibility: This relates to how fast you can add more resources to infrastructure or you can recover from failures. The flexibility degree is pragmatically proportional with the cost. Due to the non-stop flow of data for Big Data projects the physical infrastructure must be both redundant and resilient [35].

Layer 2 - Security infrastructure (Management challenges)

In additional to the traditional security measures applied by operation systems and applications to control access to data and protect data during communication, special measures should be taken to consider the special attributes of Big Data. The main solution to ensure the data remains protected is to apply encryption. The remove of any personal identifiable details is crucially required to protect user's privacy which can be achieved by anonym zing the data records and removing all personal sensitive data before saving those records. In addition it worth saying that if the Big Data is not controlled by clear governance framework, we will end up with misguiding data leading to unexpected loss, however the Big Data still new concept and no standards or policies are currently in place [36,37].

Layer 3 - Operational databases (Process challenges)

Big Data architecture will still need to have operational database to fulfill the business needs. For example looks to

Facebook environment if you changed your relationship status, this change will start to propagate to timeline of your family members, partners, immediate friends and so on until reach remote friends. The Facebook uses MySQL database to track such changes and also to track all changes performed on its home page. Database technology provide different engines and query languages and according to the requirements of Big Data implementation to implement one or more engines and to select SQL language or any other languages [38], More details about characteristics of SQL an NoSQL databases are seen in Table 1 [34]. It is crucial to understand the types of the data which will be manipulated by the database and if it supports the transactional behavior termed by "ACID" that stands for:

- Atomicity: Atomic means all or nothing. If any part of transaction fails, the whole transaction also fails.

- Consistency: Transaction with valid data will be implemented in database in consistent with defined rules and constraints.

- Isolation: Each transaction will run until end according to its order and concurrent and multiple transactions will not interfere.

- Durability: Once the transaction data written to the database, it remains forever regardless power loss, crashes or any other type of errors [39].

Layer 4 - Organizing data services and tools (Data & process challenges)

The real driver to transform this data from 0's and 1's to informed insights is still missing; this is simply the software technology. Organizing data services and tools capture, validate and organize the elements of Big Data. One software technology very well known in the Big Data field is the Apache Hadoop [40]. Apache Hadoop is an open source software framework targets to handle massive amount of data in real time by distributing the processing of data-sets among cluster of machines. It scale up from one server to thousands of computers each has its own storage and computing power, see

Table 1: Important characteristics of SQL and NoSQL Databases.

Engine	Query Language	Map Reduce	Data Types	Transactions	Examples
Relational	SQL, Python, C	No	Typed	ACID	Postgre SQL, Oracle, DB/2
Columnar	Ruby	Hadoop	Predefined and Typed	Yes, if enabled	HBase
Graph	Walking, Search, Cypher	No	Untyped	ACID	Neo4J
Document	Commands	JavaScript	Typed	No	MongoDB, CouchDB
Key-Value	Lucene, Commands	JavaScript	BLOB, Semityped	No	Riak, Redis

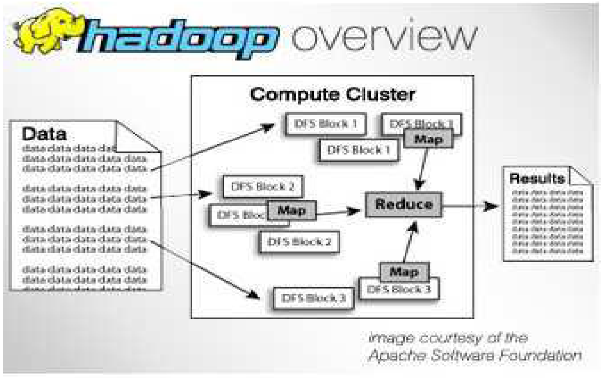

Figure 6: Hadoop overview.

Figure 6 [41]. Also it did not rely on the hardware redundancy to provide system availability as it detected failure at the application layer and takes the necessary healing actions [42].

The Apache Hadoop consists of four main modules:

Hadoop Distributed File System (HDFS): This is distributed file system responsible of storing data on multiple servers which sum up to provide very high bandwidth across the cluster.

Hadoop yarn: This module manages the computing resources in the cluster and to control assigning resources to clients' applications.

Hadoop map reduce: This is the major engine for processing large scale data.

Hadoop common: A tool box of libraries and utilities required by other Hadoop modules [43].

Layer 5 – Analytical data warehouses (Process challenges)

Data warehouse is a relational database built by integrating data from operational databases, historical databases and other sources for the purpose of running query and analysis without having any impact on the performance of online transactional databases. The data warehouse is key source of information for decision makers especially for strategic planning. Actually it is not only consolidated database that makes the data warehouse environment but, this include extraction, transportation, transformation and loading processes to compile the data in the form ready for analysis [44]. The data warehouse is repository of highly structured data while big data consists of different data types: structured, semi-structured and unstructured and cannot be merged together due to the structural difference of data and also to that the traditional data warehouse can't process the flow stream of big data. Organizations will continue using the traditional data warehouse to analyze the key values, trends and patterns. The relationship between the data warehouses and Big Data is complementary to become like hybrid structure. This hybrid structure will include highly structured data managed by traditional data warehouse and highly distributed and prone to change data in real time is managed by Hadoop-controlled solution [45].

Layer 6 - Big Data analytics (Process challenges)

Big Data analytics is the process of examining massive amount of potentially real-time and disparate data in order to for example uncover hidden trends, market trends and customer preferences. The ultimate target of Big Data analytics is to assist organizations to make more informed business decisions [46]. Generally there are three main classes of analytics tools which can be used individually or together by organization to gain real business value and these classes are:

Reporting and dashboards: User-friendly tool to represent information collected from various sources. This area is still evolving to support Big Data needs and currently they access new database technology known as "NoSQL" [34].

Visualization: These can be seen as advanced reporting tools which provide pictorial or graphical representation for data and help users to easily understand the data and relationship of several variables at the same time. The output of these tools is greatly interactive and dynamic. These tools have employed new techniques to enable users to watch the data as being changing in real time [47].

Analytics and advanced analytics: This is group of analytical techniques based on mathematical principles used to sum and count historical events in organizational data warehouse in order to identify trends and detect patterns. These techniques help to predict the future outcome and to make the necessary decisions either to support it or take corrective measure to avoid the anticipated impact. Predictive analytics, simulation analytics and optimization analytics are common examples of those techniques [48].

Layer 7 - Big Data applications (Process challenges)

Users are most interested with the technology products relevant to this layer as they are the end products which they interact with. These products can be third-party applications or in-house applications which can fulfill common requirements of multiple industries or the requirement of particular industry. Some examples of well-known groups are log data application (Splunk, Loggly), marketing applications (Bloomreach, Myrrix) and advanced Media applications (Bluefin, DataXu). Building custom products for Big Data should follow up proper software development standards, for example to include well-defined API interface which will help the developers to access the functionalities exposed by each layer through those interfaces [49].

Discussion and Future Work

Now it is evident that Big Data is too big, too fast and too disparate data which could not be processed by the traditional

doi:http://dx.doi.org/10.4172/2324-9307.1000135

database architectures and need special sophisticated systems. The "Big Data" still considered as concept and the market of that technology still growing but eventually will evolve during the next five years. Researches

show that under normal state conditions the market of Big Data will nearly double over the period 2015-2020, increase from about $38 billion in 2015 to more than $76 billion in 2020 [50].

It becomes obvious that developing the Big Data technologies to process bigger and real-time data and from more devices than ever would be the standard plan for the future, while researchers look to that as only extension for the current approach. Actually researchers are looking forward to leverage data on deeper level to get more informed insights than the normal data points which are being extracted now. Data are now being analyzed to find out answers to traditional questions: "when? What? How many? How long?" which answered by numeric data points and called "Big Data 1.0". In the future researchers plan to analyze the Big Data deeper to find answer to the question "why?" which studies the emotion laying behind the calculated numbers; researchers called this "Big Data 2.0". The new Big Data would be achievable after developing the text analytics which will enable processing the data looking for phrases and patterns explaining the motivations behind concluded results [51].

It is clear that we are still in the beginning of Big Data era and there are still many things to discover. Big Data is nice keyword but many people don't know a lot about. Until now Big software corporations do not own or do not market solutions for Big Data, for instance companies like Google does not utilize its Big Data solution in commercial way. If companies want to utilize Big Data, they should build well defined strategy for implementation. There are different approaches to implement Big Data solution. The first is the revolutionary approach which simply means that company establishes a new Big Data computing environment and moves all the data to the new platform, so all processing, analysis, reporting and modeling are performed through the new business intelligence and analytics. Several analytics driven companies already implement this approach and move their data to Hadoop environment and on the top of that they build business intelligence solutions. Also there is the evolutionary approach in which the Big Data is processed using the current traditional BI platform. The data is gathered and analyzed through structured and unstructured tools and then output forwarded to the data warehouse.

Traditional modeling and reporting utilities now can access thoughts and records streaming from social media sources. However even if the evolutionary approach fulfils many of the requirements of Big Data environment, it still has as well most of the problems of classic BI which might become a bottleneck between the information streaming from Big Data sources and the analyzing power of transitional BI or data warehouse. The third approach is hybrid one in which both traditional and new Big Data technologies are used and data are distributed between the two, one example of such approach is Hana solution from SAP [52].

References

1. Dirk R, Chris E, George L, Paul Z, Tom D (2011) Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, (1st edtn), McGraw-Hill Osborne Media, New York.
2. Mongo DB (2014) Big Data Explained.
3. Teich P (2012) Big Data is Extra Sensory Correlation.
4. Rijmenam M (2013) Union Pacific Railroad Turned To The Industrial Internet To Stay On Track.
5. SQL stream. Intelligent Transportation.
6. Soubra D (2012) The 3Vs that define Big Data.
7. Chacón D (2013) Why Big Data Is Important to You and Your Organization.
8. Davenport TH, Jill D (2013) Big Data in Big Companies.
9. Power DJ (2014) What is data-driven decision making?.
10. Parise S (2012) Four strategies to capture and create value from big data.
11. Davenport T (2012) Three big benefits of big data analytics.
12. Harvard Business review (2013) Big Data: The Future of Information and Business.
13. Jaseena KU and Julie MD (2014) Issues, Challenges, and Soltuions: Big Data Mining.
14. Shields A (2014) Overview: What is "big data"?.
15. SAS Institute Inc. (2013) Five big data challenges.
16. Zicari RV (2013) The challenges and opportunities of big data.
17. Orlova A (2014) Telcos Gain Valuable Insight with "Big Data".
18. Perficient. IBM Big Data Solution.
19. Hulme T (2012) Understanding Big Data.
20. Datamation Magazine (2013) Exploring the Big Data Stack.
21. Weiss A (2012) Data at the Speed of Life.
22. Hamoudy E (2014) Analyzing 6Vs of Big Data using System Dynamics. The 2nd Scientific Conference of the College of Science, 75-83.
23. Normandeau K (2013) Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity.
24. IBM (2013) Data Quality.
25. Zicari RV (2012) Big Data for Good.
26. Kahn R (2013) Big Data Processing Pipeline & Challenges Infographic.
27. Agrawal D, Philip B, Elisa B, Susan D, Umeshwas D, et al. (2011) Challenges and Opportunities with Big Data Purdue e-Pubs 1-16.
28. Labrinidis A, Jagadish HV (2011) Challenges and Opportunities with Big Data.
29. Khan E (2013) Addressing Big Data Problems using Semantics and Natural Language Understanding.
30. Pradeepa A, Thanamani AS (2013) Significant Trends of Big Data Analytics in Social Network. Int J Adv Res Com Sci Soft Eng 3: 510-514.
31. Srivatsa P (2014) Big Data and Data Science: Case Studies. Int J Res Sci Inn 1: 22-26.
32. Ohlhorst FJ (2012) Big Data Analytics: Turning Big Data into Big Money, John Wiley & Sons, 176.
33. Markey J (2014) How to Manage Big Data's Big Security Challenges.
34. Talend. Big Data Governance.
35. Judith H, Alan N, Fern H, Marcia K (2013) Big Data for Dummies, John Wiley &Sons, Inc., New Jersey, USA.
36. Terma Software Labs (2013) Challenges and solutions in big data analytics.
37. Lafuente G (2014) Big Data Security - Challenges & Solutions.
38. Dave P (2013) Big Data – Operational Databases Supporting Big Data – RDBMS and NoSQL – Day 12 of 21.
39. Kothari V (2014) Operational Database in Big Data.
40. Adrian AT (2013) Big Data Challenges. Dat Sys J 4: 31-40.
41. Dorf M (2011) Intro to Hadoop.
42. Apache Hadoop (2014) Hadoop.

43. Bappalige SP (2014) An introduction to Apache Hadoop for big data.
44. Oracle (2002) Data Warehousing Concepts.
45. Hurwitz J, et al. (2014) Integrate Big Data with the Traditional Data Warehouse.
46. Rouse M (2014) Big Data Analytics.
47. SAS. Data Visualization: Making Big Data Approachable and Valuable.
48. IBM. What is Advanced Analytics ?.
49. Hurwitz J. Big Data Applications.
50. Newswire PR (2014) The Future of Big Data Analytics - Global Market and Technologies Forecast - 2015-2020.
51. Spiegel B (2014) The Future of Big Data – Big Data 2.0.
52. Trifu MR, Ivan ML (2013) Big Data: present and future Database Sys J 5: 32-41.

Author Affiliation

Top

Department of Computer Science, SZABIST Dubai, United Arab Emirates