International Journal of Advanced Computer Technology (IJACT)

ISSN:2319-7900

BIG DATA APPLICATIONS AND CHALLENGES

BAKRI HUSSAIN AL-AWAJI, Najran University, Najran, Saudi Arabia

Abstract

Big Data has revolutionized scientific research in many ways. It is capable of not only revolutionizing research but every field of education also. It is believed that usage of IT can minimize the healthcare costs and also improve the quality by making the care process more personalized and preventive. Big Data has also found significance in the fields of urban planning; intelligent transportation; environmental modeling; energy saving; smart materials; financial systemic risk analysis; homeland security; computer security etc. This paper examines the application and challenges of Big Data in current market.

Introduction

Every facet of human life is being driven by Big Data today ranging from consumers to enterprises and from government to science. It is a multiple-step process to create value from Big Data. It comprises of steps like acquisition, data integration, information cleaning and extraction, analysis and modeling and deployment and interpretation. Some of the challenges faced by researchers with respect to Big Data are data heterogeneity, privacy, timeliness, collaboration and visualization [1]. Case studies are showing that numerous rewards would be ushered on the ones capable of using Big Data in the right manner. Data has become indispensible in today's world where it is being gathered at an unprecedented scale in multiple applications. Decisions that had been taken previously based on reality models or were just guessed can be made today based on the accumulated data. Big Data analysis is presently driving each and every facet of the modern day society comprising of physical sciences, life sciences, financial services, manufacturing, retail and mobile services.

The prospective advantages of Big Data are significant and real but there are various technical challenges that have to be paid attention to. The key challenge is the huge size of data. Some organizations believe that apart from Volume challenges, there are other issues with Velocity and Variety. Velocity refers to the time within which the data must be dealt with and the arrival rate of the data [9]. Variety refers to heterogeneity, semantic interpretation and representation of the different types of data. These challenges are evident but there are other requirements like usability and privacy. The main objective of this paper is to critically examine the application and challenges of Big Data.

Applications of Big Data

Big Data analysis is all about varying phases introducing different challenges. Most of the people focus unfortunately only on the modeling/analysis phase, which even comprises of complexities with respect to multi-tenanted clusters running multiple user programs concurrently. Some of the challenges are found in other phases as well. For instance, management of Big Data might be noisy, might not comprise of an upfront model and might be heterogeneous in nature [9]. This mandates tracking of provenance and efforts to deal with error and uncertainty. This would also require better support and smarter systems for analysis pipeline and user interaction. Presently there are a number of bottlenecks to the total number of individuals that are empowered to enquire about the data and its analysis [2]. This number can be increased drastically by supporting different engagement levels with data. These problems can be resolved by devising suitable ways to deal with data analysis.

A number of opportunities are generated by Big Data for improving the understanding for human behavior in order to support global development in 3 different ways. First is early warning about the anomalies existing in the ways digital services and populations to enable speedy response during crisis use devices. Second is real-time awareness as Big Data is capable of painting a fine-grained and latest representation of the facts informing the targeting and design of the policies and programs [2]. Last is real-time feedback which is acquired by monitoring a population for understanding the failure of programs and policies and making the required adjustments.

All these applications are quite promising but it should be noted that nothing is actually automatic about converting the Big Data sources to actionable information. Big Data just like any other information technology is capable of bringing about significant cost reductions, improvements in the time needed for completing any computing task or service offerings and new product. It also supports the business decision taken internally. The concepts and technologies linked with Big Data permit organizations to fulfill multiple objectives [3].

Organizations that pursue Big Data are strong believers that terabyte and MIPS storage with respect to structured data can be delivered cheaply via Big Data technologies such as Hadoop Clusters. However, data security methods associated with the Hadoop cluster has not yet been developed completely. Organizations focusing on cost reduction took the decision of adopting the Big Data tools based on economic and technical criteria. IT groups might also involve few of the sponsors and users for debating the disadvantages and advantages of data management. However, cost reduction is the secondary objective after the fulfillment of other objectives. For example, the primary goal of an organization was innovating new services products via big data.

The company on achieving this objective might be willing to study ways in which this can be done in a cost-effective manner. This was the case with GroupM, which is the media-buying subsidiary for WPP, an advertising conglomerate. This organization purchases much more media compared to any other global company using a number of big data tools also [3]. The key problem is that there are about 120 offices of GroupM all over the globe with each of its office having its own approach and technological viewpoint towards big data analytics.

The cost for allowing every office to deploy their personally selected big data tool would amount to a minimum of $1 million for every site. Thus, GroupM instead of using this decentralized approach aims at offering centralized big data services from its office in New York. The key focus would be on the twenty-five markets all over the world expecting to spend 1/3rd of the total amount required for every site based on the decentralized approach.

Time reduction is the second most important objective associated with Big Data solutions and technologies. The merchandise pricing optimization application used by departmental store, Macy, is the best example for minimizing the cycle time for large-scale and complicated analytical calculations from some hours or days to minutes to even seconds. The store has been successful in reducing the time required for optimizing the prices of about 73 million products from 27 hours to 1 hour [3].

This capability has made it possible for the store to reprice its products frequently for adapting to the changing conditions within the retail markets. The Hadoop cluster is taken out by big data analytics application and put into the parallel computing and in-built memory software architectures. Macy had achieved about 70% of cost reduction with respect to its hardware. Another important objective with respect to time reduction is interacting with the customers using data and analytics acquired from consumer experiences. Targeted services and offers fail to be effective when the customers leave the building. This indicates towards rapid analytics, processing, aggregation and capture of data.

Implementation of Big Data in the development of new goods and service offerings is the most aspiring step an organization can take. This approach is usually employed by online firms that need to deploy data-based goods and services. Google is a strong contender for the development of goods and services using Big Data. It makes use of big data for refining its core ad-serving and search algorithms [4].

Google is constantly developing new goods and services as it has huge data algorithms for ad and search placements at its core like Google Apps, Google Plus, and Gmail etc. Few of the product developments usually pay off whereas some others are simply discontinued. Some of the organizations beyond the online industry are few other examples that use this phenomenon. The firms that had been interviewed under this research considered GE as the most important trigger for developing new services depending on big data. The key focus of GE is to focus on optimization of service maintenance and contracts intervals for the industrial products.

System Architecture

Business intelligence is largely being appreciated, valued and used by most of the companies today. There are a number of purposes for which business data has been analyzed like performing social media analytics and system log analytics for brand management, customer retention, risk assessment etc. [4]. These tasks have typically been managed by distinct systems in spite of every system comprising of a number of common steps for predictive and statistical modeling, relational-like processing, data cleaning, information extraction and suitable visualization and exploration tools.

Using Big Data helps in using distinct systems making the entire process quite expensive based on the huge size of data sets. This expenditure is not due only to the system costs but also to the time for loading the data into various systems. Thus, Big Data has made it essential to have heterogeneous workloads run on a solo infrastructure, which is flexible enough to deal with such workloads. The key challenge is to develop a system, which is ideally suitable for every processing task. The most urgent need is flexibility of the system architecture such that its components express the different processing tasks tuning them such that they can run efficiently on varying workloads. This particular section would emphasize on the different programmability requirements. It is important for users to have suitable high-level of primitives for specifying their needs in flexible systems if complicated analytical pipelines are to be built and composed using Big Data. In this sense, the framework of MapReduce has turned out to be highly valuable but is the primary step [5]. Declarative languages like Pig Latin exploit it and are found at quite a low level with respect to complicated analysis tasks. Identical declarative specifications are needed at a higher level to fulfill the composition and programmability needs of the analysis pipeline. Declarative specification is required basically for pipeline composition and also for individual operations. Every operation runs potentially on an enormous data set. Also, every operation is quite complicated as multiple optimizations and choices are possible for its implementation [6].

Significant work has been done in the databases on optimization of individual operations like joins. It is quite obvious that multiple levels of magnitude differences are possible in the cost of the varying ways of executing this query. The user fortunately does not need to make this choice as this task is accomplished by the database system. Such optimizations with respect to Big Data might be more complicated as all the operations would not be I/O intensive like the databases. Some of the operations might be CPU intensive or may be a mix [6]. This does not make sense for the direct usage of standard database optimization methods. Nevertheless, the development of new methods can be possible with respect to the Big Data operations driven by database methods. Big Data analysis comprises of numerous phase highlights, which is the key challenge arising in, practice regularly. Complicated workflows or analytic pipelines must be run by the production systems at regular intervals. It is important to take into consideration new data taking into account the outcomes of pre-existing data and prior analysis.

It is important to preserve provenance including the different phases within the analytic pipeline. No or little support is offered by the current systems for the Big Data pipelines, which is quite a challenging objective.

Challenges of Big Data

Application of Big Data analytics to the process of development faces a number of challenges. Few of them are related with the data comprising of its sharing and acquisition and concern about privacy. This section would focus on the salient challenges.

Privacy is quite a sensitive issue comprising of technological, legal and conceptual implications. It is defined as the individual right to influence or control the information linked with them and its disclosure. It also comprises of the organizational desire to safeguard their consumers, states and competitiveness to preserve their citizens and sovereignty. Thus, it has become a serious concern having multiple implications for the ones that desire to utilize Big Data for the purpose of Development [7]. It is also the fundamental right of every human being having both instrumental and intrinsic values. Authors emphasize the need for ensuring suitable degree of privacy for companies, individuals and the society at large. Privacy has become quite important for the modern society to flourish. Basic freedom, innovation, pluralism, diversity and safety would be at risk without privacy. Such risks are of concern for every individual that does not have anything to hide. The need for expanding at length on the sensitivity and significance of information for states and corporations is not needed [8].

With respect to individual privacy, it can be said that the primary producers like the devices that generate data and service users are quite unaware of their activities and usage. For instance, individuals routinely consent towards the use and collection of web-generated data on ticking a box not realizing the misuse or usage of their data. For example, it is not yet clear if Twitter users and bloggers consent to analysis of their data. Also, recent research work shows that it is possible to deanonymize the datasets that had been anonymized previously [7].

Another major concern is the enormous individual-level information being held by Facebook, Google, credit card companies and some mobile phone firms. It should be noted that privacy is nothing but the pillar of democracy and individuals should stay alert to likeliness of it being compromised by new technologies putting into place the essential safeguards.

Majority of the online data available publicly has significant value for further development but there is much more valuable data, which is held closely by the corporations and cannot be accessed for commitments made in this paper. The key challenge here is the reluctance exhibited by private firms and related institutions for sharing data regarding their users and clients and their personal operations. The key obstacles might comprise of reputational or legal considerations and the need for safeguarding their competitiveness, lack of information structures and right incentives and culture to maintain secrets.

Some technical and institutional challenges also exist like difficult access to and transfer of the stored data. For instance, Nathan Eagle, the MIT professor usually describes weeks spent in the basements of cell-phone firms in Africa seeking through thousands of boxes filled of magnetic backup-tapes for accumulation of data. It has been estimated by some Indonesian mobile carrier that almost half work day would be taken up for extracting backup data worth of one day stored within magnetic tapes. It might prove to be difficult in UN system to have the agencies share the program data stored with them based on the reasons given above.

Data streams can be accessed reliably and back-up data can be accessed for data training and retrospective analysis by partnering with suitable partners in private and public sectors for accessing non-public data. Some of the other technical problems are of inter-operability of the systems

ISSN:2319-7900

and inter-comparability of the data but these problems might not be that complex to resolve. Problems like acquiring formal agreement or access on licensing issues related with the data are critical issues [8].

There are a number of critical challenges for the development of Big Data in order to acquire traction. Initiatives that are taken for identifying the salient privacy issues and significance of managing the data should make sure that privacy would not be compromised. Such concerns might shape and nurture the on-going debates on data privacy in today's digital age constructively for devising strong rules and principles baked by suitable systems and tools for ensuring privacy-preserving analysis.

Nevertheless, this promise would be unfulfilled if the institutions basically private corporations are not willing to share the data. For example, Global Pulse in this light is promoting the notion of 'data philanthropy' stating that corporations will be taking the initiative of anonymizing the data sets and providing the data to the social innovators for data mining to acquire the trends, patterns and insights in real time. The success of the data philanthropy concept is not sure but it surely points out towards the avenues and challenges to be considered in the times to come. It can also be expected to consider further alternative models and refinements for determining the ways to deal with data share and privacy.

Heterogeneity is acceptable to a great extent when human beings consume information. The richness and nuance associated with natural language provides great depth. Nevertheless, machine analysis algorithms not tolerating nuance expect homogenous data. The primary step for data analysis is careful structuring of data. For instance, assume a patient undergoing numerous medical procedures in the hospital [8].

One record can be created for every laboratory test or medical procedure; one record for the entire stay at the hospital or one record for every interaction that a patient has with the hospital for their lifetime. However, the number of lab tests and medical procedures would vary for every patient. The design choices that have been stated indicate greater variety and less structure. Most of the traditional systems for data analysis required greater structure. Nevertheless, less-structured designs are more likely to be effective in multiple ways. Computer systems are much more efficient as multiple items can be stored by them in identical structure and size. Further work has to be done on analysis, access and efficient representation of semi-structured data.

Assume a database design for electronic health record having specific fields for blood type, occupation and birth date of every patient. The question is about the absence of any of this information with respect to the concerned patient. It is obvious that the database has all the health records but the equivalent attribute values are NULL [8]. Data analysis aimed at classifying the patients must consider the patients lacking the information. In spite of error correction and data cleaning, some errors and incompleteness in the data will be found. These errors and incompleteness should be dealt with at the time of data analysis. It is quite a challenge to do this correctly.

As dealing with novel data sources results in a few analytical challenges we will have some cases to discuss. The severity and relevance of the challenges would vary based on the analysis being carried out and on the decisions taken based on the data. The main question here is about the message being conveyed by the accumulated data. This question is the core of evidence-based policymaking and social science research. However, it is generally perceived that novel digital data sources results in acute challenges. It is important to spell out such concerns in a completely transparent style. All the challenges are difficult to consider and interview in isolation. However, they can be classified into three categories for ensuring clarity. They are acquiring the correct picture by summarizing the entire data, making sense or interpreting the data via inferences and detecting and defining anomalies [9].

Acquiring the correct picture reminds one of Plato's allegories regarding the cave which refers to data that is seen by the analyst as the shadow of the objects that trespass the fire. The chief question here is the accuracy of the message reflected by the data. The data may sometimes be fabricated or false. For instance, unverified bloggers or reporters might publish false information. Facts may also be falsified or fabricated by journalists, bloggers and citizen reporters as they are individuals who speak based on their real identity. External factors or actors may interfere making the data paint seem like a misleading image of reality.

For instance, the perpetrators would make active efforts to suppress reporting when SMS streams measure public violence. The SMS streams would not only measure the location of cell phones but would also measure the location of cell phones, which cannot be suppressed by the perpetrators. There would be many other false negative zones exhibiting no violence and no SMS traffic at all [9]. This would result in highly duplicated and dense reports of the visible events having multiple observers and almost no efforts to suppress texting. In such cases, a willingness to change the perception about reality occurs. This particular challenge is highly salient with text-based data that is user-generated and unstructured like social media messages, news and blogs due to its loose verification steps and spontaneous nature.

ISSN:2319-7900

Also, most of the share of novel digital data sources making up Big Data seems to be derived from individual perceptions of individuals. For instance, it indicates towards information, which is extracted from health hotlines, calls and online searches for disease symptoms. Individual perceptions vary from feelings as they seem to express objective facts like health symptoms. On the other hand, perceptions can be misleading and inaccurate [9].

The best example here is that of Google Flu Trends and its ability of detecting influenza epidemics in fields having huge population of the web search users has been discussed previously. Google Flu Trends have been compared with data by a team comprising of medical experts from year 2003 to 2008 comprising of data acquired from two different surveillance networks. It was seen that Google Flu Trends have been highly successful in prediction of non-specific respiratory illnesses. Thus, problems in varying out sentiment analysis would be organized in varying ways.

One particular perspective differentiates challenges associated with conceptualization like definition of clusters, categories; measurement like assignment of clusters and categories to the unstructured data and verification like assessing the success of steps 1 and 2 in extraction of significant information. The key focus is also laid on the chief challenges faced during the selection of target documents; identification of the expressed sentiment within the target documents and presentation of such sentiments in a summarized manner.

On the whole, the basic challenge is to reach the actual intention of the statement with respect to intensity, polarity etc. This might be impeded by most of the obstacles ranging from usage of slang; sarcasm; irony; hyperboles and local dialect to lack of key words. These are referred to as measurement or technical challenges becoming easier to deal with as the extent to complication associated with sentiment analysis algorithms advances. However, the classification and conceptualization of the analysis being conducted is not that trivial. For example, this indicates on deciding if presence or frequency of the key words is important.

Therefore, the most important thing is the input of the human analyst. It should be noted that classification is the most generic and central of the conceptual exercises. Big Data analysis is presently driving each and every facet of the modern day society comprising of physical sciences, life sciences, financial services, manufacturing, retail and mobile services. There cannot be any advanced conceptualization, language, reasoning, data analysis or social science research or any other kind of research work in absence of classification [7].

The chief aspect of Big Data is its enormous size. It is quite a challenging task to manage rapidly increasing and large data volumes. This particular challenge had been mitigated in the past by faster processors based on Moore's law for providing the required resources to deal with the huge data. However, there has been a rapid shift with data volume scaling speedily compared with computed resources and static CPU speeds. There has been a dramatic shift in the processor technology over the previous 5 years. The traditional data processing systems were concerned about the parallelism associated with the nodes within a cluster. Today, a single node is used to manage parallelism. The data processing methods that had been applied previously for data processing cannot be applied anymore for intra-node parallelism. This is because the architecture seems quite different with multiple hardware resources like processor memory channels and processor caches shared across a single node [7]. The shift towards packing of numerous sockets has made it much more complicated for intra-node parallelism. Such changes have made it essential to rethink ways of designing, building and operating the data processing elements.

Another dramatic shift that is taking place is adoption of cloud computing that aggregates a number of disparate workloads with different performance goals into huge clusters. Sharing the resources on large and expensive clusters mandates novel ways to determine ways to execute and run data processing jobs for fulfilling the goal of minimizing workload cost and dealing with system failures. Dependence on optimization of user-driven programs result in poor utilization of clusters as the users is not aware of the other programs. It is important to have transparent programs for system-driven holistic optimization.

Another dramatic shift is transformative changes in the traditional I/O subsystem. Hard Disk Drives (HDDs) have been used over the last few years for storing persistent data. They exhibit sluggish random IO performance compared with sequential IO performance. Today, solid state drives are replacing the HDDs with technologies like Phase Change Memory also playing a role. Such novel storage technologies fail to have a huge performance spread between random and sequential I/O performance. This would require rethinking about the designing of the storage subsystems [7]. Changes in the storage system affect every facet of data processing like query processing algorithms, database design, query scheduling, recovery methods and concurrency control methods.

Conclusion

Today, it is the phase of Big Data. Even though there are a number of new ways available for better analysis of the

ISSN:2319-7900

huge data volume, there are chances of making rapid advances in numerous scientific disciplines and enhancing the success and profitability of most of the enterprises. Nevertheless, most of the technical challenges that have been stated in this paper should be addressed prior to the realization of this potential [5]. The key challenges are the scale issue, heterogeneity, and error-handling, absence of structure, privacy, provenance, visualization and timelines. These challenges stem in all the phases of analysis pipeline ranging from acquisition of data to interpretation of result. Such technical challenges are quite common in multiple application domains and are not cost-efficient for addressing with respect to just one domain.

These challenges also require a number of transformative solutions and cannot be addressed naturally by the upcoming industrial products. Fundamental research should be encouraged and supported for addressing such technical challenges for achieving the promised advantages of Big Data. Apart from the fundamental technical need, there are robust business imperatives also. Big Data processing would be typically outsourced. Declarative specifications are needed for enabling service level agreements that are technically meaningful. Further work has to be done on analysis, access and efficient representation of semi-structured data along with usage of Big Data in the right manner.

References

[1] A. Labrinidis, "Challenges and Opportunities with Big Data 2011-1," Cyber Cente Technical Reports, 2011.

[2] A. Pentland, S. Berinato, "With Big Data Comes Big Responsibility," Harvard Business Review, 2014, 92(11),100-104.

[3] G. Fulgoni, "Big Data: Friend or Foe of Digital Advertising? Five Ways Marketers Should Use Digital Big Data to Their Advantage," Journal Of Advertising Research, 53(4), 372-376. 2013.doi: 10.2501/JAR-53-4-372-376.

[4] H. JAGADISH, G.GEHRKE, A. LABRINIDIS, Y PAPAKONSTANTINOU, J. PATEL, R. RA MAKRISHNAN & C. SHAHABI, "Big Data and Its Technical Challenges," Communications Of The ACM, 57(7), 86-94. doi:10.1145/2611567, 2014.

[5] P. Goes, "Big Data and IS Research," MIS Quarterly, 38(3), iii-viii, 2014.

[6] D. Nunan, D. Domenico, "Market research and the ethics of big data," International Journal Of Market Research, 55(4), 2-13, 2013.

[7] S. Rodríguez-Vaamonde, A. Torre-Bastida, E. Gar rote, "TECNOLOGÍAS BIG DATA PARA ANÁLISIS Y RECUPERACIÓN DE IMÁGENES WEB," (Spanish). El Profesional De La Infor mación, 23(6), 2014, 567-574. doi:10.3145/epi.2014.nov.02.

[8] W S.TIRUNILLAI, G. TELLIS, "Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation," Journal Of Marketing Research (JMR), 2014, 51(4), 463-479.

Biographies

BAKRI HUSSAIN AL-AWAJI received the B.S. degree in Computer Science from King Abdul Aziz University, Jeddah, Saudi Arabia, in 2003, the M.S. degree in Computer Science from the University of Colorado at Denver, Denver, Colorado, U.S.A, in 2015. Currently, He is a lecturer of Computer Science at Najran University and he is seeking the PhD degree in the same field. His teaching and research areas include Cloud computing, Big Data, Data Mining, Artificial Intelligence, Human and Computer Interaction, Software and system modeling, and embedded system design. Mr. Bakri HUSSAIN AL-Awaji may be reached at Balawaji@nu.edu.sa.