ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

IMPLEMENTATION CHALLENGES INVOLVED IN BIG DATA ANALYTICS

Meghna Sangtani, Shannon Michelle D'souza, Mriganka Harsh, Satish Chander, *P. Vijaya

* Waljat College of Applied Sciences, Muscat, Sultanate of Oman.

DOI: 10.5281/zenodo.163317

ABSTRACT

Big Data is probably the most talked-about topic in the IT world today. Currently, information is produced and stored at a rapidly exceeding rate. According to the current analysis, there are over 2 billion internet users and almost the double who own mobile phones. It is predicted that by 2020 the data produced will be nearly 44 times greater than that today. As the data is generated every second by everything around us, the volume of data increases simultaneously, and with this come numerous challenges in handling such large data sets, such as the increasing volume of data, transfer speed, heterogeneous data, and security.

By overcoming such challenges and adopting the right implementation ways, Big Data is said to revolutionize the IT world. This paper is intended to provide a detailed view about Big Data and the phases of big data analytics, as well as the challenges encompassed in its implementation.

KEYWORDS: Big Data, Diverse Data, Big Data Analytics.

INTRODUCTION

Every field of the contemporary world including the production industry, hotel industry etc. is using big data analysis. Handling such exponential and complex data being generated at an unmatched scale is now the biggest challenge in the industry. As a result, it is giving data scientists a great opportunity to discover new techniques to analyze Big Data.

Now, the ultimate challenge is to merge data into something that can enhance the economic growth, on the basis of structured and unstructured data. Big Data analytics is used by leading companies like Google, Yahoo and IBM to accomplish success by escaping all possible business risks.

One can't deny the momentous benefits of Big Data, but still there persists countless technical challenges that demand to be looked upon to make it easier to implement the data collected. The biggest challenge is the "every second" increasing size of data. Also, privacy, heterogeneity, scalability etc. are some important factors that need to be addressed.

Several distinct stages are involved in the analysis of Big Data, each stage introducing a new challenge. This leads to poorly understood difficulties in the analysis phase due to various user programs running in parallel.

Big Data must be managed in context as it is heterogeneous and lacks suitable models. This frequently leads to countless uncertainties and errors.

Luckily, to solve some aspects of this Big Data problem we can apply the continuing computational methods either directly or by adding a few extensions to it. As new solutions are being developed to many new challenges of Big Data, it becomes tedious to combine these former system technologies.

In this paper, we first briefly examine the five stages in the analysis pipeline, then focus on the challenges faced by the data analysts and conclude it with the case study of a prime Middle East Company using Big Data.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

THE BIG DATA ANALYSIS PIPELINE

First, identify the right software which will be used (SQL or NoSQL) and the required analytics tools (Apache Hadoop, Spunk, etc.). This will set the base of the technology infrastructure. However, this completes only half the start-up battle on "Big Data" analytics.

The next step must focus on using a strategy that will be able to exploit Big Data efficiently.

The transition from Simple Data analysis to Big Data analysis is a challenging task as:

1. It must include the "fit" features from existing systems, and
2. It must create new solutions to combat the upcoming challenges of Big Data

Thus, we "dumb down" the whole Big Data Analysis as five stages in a pipeline to ensure a smooth deployment:

Acquiring and Recording of Data

Big data is being generated by various data sources in large volume every day. However, most of this data is of no interest. Thus, we must find a way to extract "meaning" from meaningless data. The ability to recognize strategic data will lead to precise analytical insights.

In this stage, data reduction techniques are applied which can intelligently process raw data and scale it down for the user to grasp it alongside not missing out on probable vital information. Streaming data, on the other side is 'Data with velocity'. For such data, "on-line" analyzing techniques are additionally required as this analysis type is one-pass (Data can be analyzed only once).

Extraction and Cleaning of Data

The data amassed in the previous stage is mostly unstructured and cannot be used instantly in analysis. The term 'unstructured' refers to any format of stored information which is nearly impossible to process. Thus, to counter this complexity, a Data Extraction Process, that extracts the required data from a variety of sources, should be defined. Presently, this data also contains images and soon, in the near future, may even include videos. It must convert the raw data in a structured format for a smooth analysis. But, the implementation is still a constant technical challenge.

To obtain the right analytical outputs, data scientists define a necessary set of constraints in advance. Once defined, they can assess the amount of complexity created and the work required to convert the raw data into applicable and productive findings. Existing work on data cleaning will be performed using these well-recognized constraints on relevant data or well-understood error models; but, for most impending Big Data domains these are not applicable.

The most common delusion of Big Data is that it "always tells the truth", but this is quite distant from reality. Analysis of 'clean' data doesn't necessarily guarantee correct answers.

Data Integration, Aggregation and Representation

Given that Big Data is heterogeneous, it is not enough to simply record and store it.

Data Integration combines data from various dissimilar sources into valuable information. Thus, it provides a consolidated view of data. **Data Aggregation**, on the other hand, searches and gathers data findings and then presents it in a summarized format, making it easier for the end user. Big Data, in its raw form, is hardly of any value and thus, this process is implemented on it. With more data being processed, it should be represented in a format which allows a user to interpret it easily. This is known as **Data Representation**. Big Data can be available in the following formats:

- Numbers
- Text
- Graphics of many varieties (stills, video, animation)
- Sound

To apply an effective large-scale scrutiny, 'Integration, Aggregation and Representation' of data must occur in a completely automated manner. However, bringing this entire process to effect is a major challenge. To analyze even a single data set, a suitable data model must be designed. Reason: Every single design will have a little edge over the others for precise goals, and perhaps drawbacks for the other purposes. Thus, database design is nowadays a fine art, and is carefully executed in the enterprise context by the above highly-paid experts.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

2.4 Query Processing and Data Mining

Big Data: Heterogeneous, Dynamic, Inter-related, Noisy and Untrustworthy. Ironically, this data becomes valuable when we discover patterns and thus, establish relations between various types of data. This process is called **Data Mining**. It utilizes advanced mathematical algorithms to:

1. Segment and sort through the data
2. Identify patterns and establish relationships, and
3. Evaluate the probability of future events
4. Analyze relationships that have not yet been discovered

The advantage of data mining is that it enhances the quality to provide trusted data, comprehend its semantics, and provide intelligent query functions. Here, analysts create the queries and algorithms in order to generate the desired outputs. The more accurate the queries are, the less redevelopment is required. However, this monotonous procedure of "exporting data from the database, performing a non-SQL process and bringing the data back" is a very slow process. Also, mining requires simultaneously analyzing unstructured data – transactions, social media interactions, machine data, log data, and so on.

To summarize, Data Mining provides answers to problems that cannot be addressed through simple querying and reporting techniques.

Interpretation

This stage will require all the users to be considered and taken care of. Having the ability to analyze Big Data is hardly of any value if users cannot understand the analysis. Ultimately, a decision-maker, provided with the result of analysis, has to interpret these results.

Usually, interpretation involves examining all the assumptions made and retracing the analysis. Sometimes providing the results is hardly sufficient. One must provide additional details to explain how every single result was derived, on the basis of the inputs. Such information is called the Provenance of the data.

Furthermore, it should be made easier for the user to get the desired information at a fast rate. It means that, users must also know the reason behind the displayed results and not simply "see" those results. But raw provenance, considering the stage in the analysis pipeline, would probably be too technical for most users to grasp completely.

There's no one way to ensure big data analytics success. But by following a set of frameworks and best practices, can help organizations keep their big data initiatives on the right track.

CHALLENGES INVOLVED IN BIG DATA ANALYTICS

We now shift to some common challenges that prevail in the stages the Analysis Pipeline.

Heterogeneity and Lack of Availability

The problems with heterogeneity and incomplete data can be handled manually to some extent. But, the same does not work for computer systems. When dealing with several items, they require the data that is identical in structure and size to work efficiently.

Unfortunately, the potential users still do not have enough knowledge about the availability and usage of tools and data. However, it has become increasingly easy to get the desired data to say nearly anything.

As we move towards automated decision-making, bigger and even complex problems arise. In the process of extracting suitable data from raw data, in order to understand the methods and suitable models to process the data, detecting and correcting errors in analysis becomes more difficult.

Even after data cleaning and correction, some incompleteness and errors are likely to remain that must be handled during the analysis process. But achieving this is a big challenge.

The changing Internet scale

The first and one of the major challenges is the constantly evolving internet. For example, recent stats show that Google processes over 40,000 search queries every second on average, meaning over approximately 3 billion

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

searches per day. Dealing with such large data requires new tools and techniques such as parallel computing, cloud computing, etc. Managing the increasing volumes of data has thus become a challenge faced by the data scientists today.

Moreover, underway in the internet is adapting cloud computing that groups multiple contrasting workloads with changing performance goals into large clusters. Sharing of resources, in this case, requires new and cost-effective ways to analyze data and to handle frequent system failures.

Data Timeliness

When the data to be processed is in bulk, then it takes a longer time to analyze. This leads to both, acquisition rate challenge and timeliness challenge.

There are plenty of situations in which the analysis result is required immediately. For example, if a fake credit card transaction is suspected or reported, then it should be blocked before the transaction takes place. In this situation, the full purchase history of the user is not practical.

In order to analyze a large data set, it is essential to discover new tools and techniques to support the queries and design suitable data models. Designing such models becomes more challenging with the continuously growing data and strict time limits.

Privacy and Security

In the Internet environment, public accessibility and openness helps facilitate crowd-sourcing data improvements. A more flexible view toward ensuring data quality through continuous improvement, transparency, and open accessibility is more consistent with Internet culture. Much of the data that is collected under their authorization or voluntarily provided is made available only under the condition that confidentiality and privacy be protected.

Big Data privacy is a major concern today. Many internet users are worried about the inappropriate use of their personal data. Managing these privacy concerns is a big challenge. One solution to this is to not make the entire user data publically available. Instead limit the access to user information by third parties. Such strategies have been used quite effectively in the past to enable academics to access confidentiality-protected survey data. Researchers who work with the raw data are obligated to not disclose the information at a level of disaggregation that would allow the raw data to be uncovered. Additionally, there may be technical ways to hide the data by recoding or selective sampling to preserve its originality for certain types of analysis, while still protecting data privacy.

A common example is the information of the user collected from location-based services that ask the users to share their location, and so the services get full access to their personal information, which is a major privacy concern. With the current technology, it is easier to track the user's location information depending on the use of cell phones. Similarly, other details can also be revealed by observing the user's movement pattern.

To reduce the risk factor associated with the user's privacy, a method called differential privacy can be implemented, but the drawback is that it reduces the content of the information that may be useful in the future. In addition, the data keeps growing and changing over time. This leads to the unavailability of suitable paradigms to solve this problem.

Human Collaboration

As computers become more powerful and new technologies are more able to harness the complexity of human life, data-intensive research is becoming more prominent. As a result, the way in which data scientists deal with data must also change.

As data complexity increases, the number of experts required to manage this data also increase. These experts may be from different domains. Thus, the data system must be designed in such a way that it accepts the input and support their association.

Another method for human collaboration to resolve problems is crowd-sourcing. For example, most often, the information provided by Wikipedia is correct. But, it also allows the users to update the information accordingly.

There is a possibility that some individuals provide fake information. In such cases, a data structure is required that can be used to analyze such crowd-sourced data with contradictory statements.

CASE STUDY: SHARE DIMENSION & VOX CINEMAS

Share Dimension, a leading software development company established in the Netherlands, specializes in Business Intelligence and cinematic predictive analysis applications. It has released Cinema Intelligence, a software suite for movie exhibitors, which helps to enhance forecasting, planning and scheduling of movies and events.

To sum it all up, Cinema Intelligence is the driving force behind the success of Share Dimension. Their Technology Partners include: Zendesk; Rotten Tomatoes; Microsoft Windows Azure; Weather Underground; Google Trends.

In the Middle East, VOX Cinemas (UAE's numero uno exhibitor) is an "intriguing" client primarily due to their running VIP and GOLD concepts which requires special scheduling.

Share Dimension uses software tools proposed to grasp the entire lifetime of a movie in a theatre environment, starting from the announcement by the studios till the moment the movie stops playing in the theatre. This entire process consists of main tools: Predict, Book and Schedule.

Predict:
By using several software tools, VOX is able to guesstimate the box office performance of a new release, a few weeks or months before the movie is released. For example, they can predict the total number of people watching the movie, per theatre and per week, and analyze locations based on the data acquired from the past movie performances and trends. The software is able to identify the patterns and analyze the movie's performance for each VOX theatre using predictions and signals from social media platforms.

Booking:
This helps VOX comprehend which theatre should play a particular movie, along with the best show timings for the movie and the auditoriums to schedule in. After attaining the analysis of past performance of a theatre, they can accordingly plan new releases and even create automated show timings and schedules easily.

Schedule:
This tool implements the smart algorithms that observes every show and performance per week and plans the next week's schedule (with optimized occupancy and maximized box office performance of each theatre). The booking team then modifies the schedule and releases it to POS.

VOX, using Cinema Scheduler, can generate schedules spontaneously to enhance every single show, every single day. Thus, every show gets the best auditorium for its forecasted attendance. So Friday's schedule looks different than Sunday's schedule for example.

The Cinema Scheduler implements 2 algorithms:
Forecasting algorithm:
It will look at every show, every movie from last week's box office data and predict what the box office is going to be next week. It does so using a combination of machine learning from the theatre historical box office performance combined with analysis on weather, Rotten Tomatoes rating pattern, public holidays, school holidays, special events such as a football in the stadium next to the location and social media.

Scheduling algorithm:-
It decides what show time and auditorium is best for every movie, every show. It does so by making sure it fulfills all constraints (like, making sure that no two movies start at the same or that it allows for 30min between the shows of the same movie or for enough pre-show time or cleaning time). But it also looks at scheduling requirements such as minimum shows committed to the studio, etc.

Modern organizations, including cinema exhibitors are inundated with flood of data. According to the records, 1 terabyte contains 2,000 hours of CD-quality music. Data is gathered from all possible directions: from box office operational systems, schedule management systems, customer contact points, social networking sites and the Web.

"Imagine that using all that data, enhanced with historical data performance and you can now picture a future where for example VOX is able to change the advertising that runs as pre-show just 5 minutes before the show starts and fine tune it to the exact audience in the room. As an advertiser you now have the perfect environment to project your brand in an auditorium in front of people you can be certain like your brand and have undivided attention. That's extremely powerful and exhibitors are pushing for this change. It's not yet here, we need much more powerful infrastructure for that to happen, but there is no doubt in my mind that this is where we are going." - Gabriel Tanasescu – CEO, Share Dimension.

CONCLUSION

This world of ours is drastically transforming into a world of Big Data. While Big Data will lead to more efficient, connected and widely acceptable product and service industries, there are many technical challenges to be dealt with in its way. The possibility for making quicker advancements can be increased in various scientific fields and profitability can be enhanced in several enterprises, through better analysis of Big Data. But, with smart solutions come complex systems because the Data Scientists need ideal processing power and analytical skills to churn value from Big Data. The challenges include not only the evident issues of scale, but also privacy, heterogeneity, timeliness, non-availability, visualization and cost, at all the phases of analysis pipeline. Moreover, these challenges will require brand new solutions and suitable design models. In order to accomplish plentiful benefits of Big Data, we should encourage further research towards finding more tools and techniques to overcome such challenges.

REFERENCES

[1] W. Lehr, 'Measuring the Internet', *OECD Digital Economy Papers*, 2012.
[2] J. Gould, 'Big data: Automation and collaboration', *NATUREJOBS BLOG*. 2015 [Online]. Available: http://blogs.nature.com/naturejobs/2015/10/19/big-data-automation-and-collaboration. [Accessed: 27- Nov- 2015]
[3] A. Labrinidis and H. Jagadish, 'Challenges and opportunities with big data', *Proc. VLDB Endow.*, vol. 5, no. 12, pp. 2032-2033, 2012.
[4] SAP News Center, 'Millions of Middle East Mobile Users to Benefit from Big Data and Analytics Solutions with Zain Group and SAP', 2015. [Online]. Available: http://news.sap.com/millions-middle-east-mobile-users-benefit-big-data-analytics-solutions-zain-group-sap/. [Accessed: 27- Nov- 2015]
[5] Security Middle East, 'Public Sector is largest Middle East producer of Big Data', 2015. [Online]. Available: http://securitymiddleeast.com/2015/03/13/public-sector-is-largest-middle-east-producer-of-big-data/. [Accessed: 27- Nov- 2015]
[6] GE Healthcare The Pulse, 'For the Middle East's Evolving Healthcare Landscape, Big Data is a Big Deal', 2015. [Online]. Available: http://newsroom.gehealthcare.com/middle-easts-evolving-healthcare-landscape-big-data-big-deal/. [Accessed: 27- Nov- 2015]
[7] Search Business Analytics, 'Five first steps to creating an effective 'big data' analytics program', 2015. [Online]. Available: http://searchbusinessanalytics.techtarget.com/feature/Five-first-steps-to-creating-an-effective-big-data-analytics-program. [Accessed: 27- Nov- 2015]
[8] MongoDB, 'Big Data Explained', 2015. [Online]. Available: https://www.mongodb.com/big-data-explained. [Accessed: 27- Nov- 2015]
[9] Zdnet3.cbsistatic.com, 2015. [Online]. Available: http://zdnet3.cbsistatic.com/hub/i/2014/10/05/0257bfa7-4c25-11e4-b6a0-d4ae52e95e57/8e9c411d640af9dff745a52355f63518/angusbigdata2.png. [Accessed: 27- Nov- 2015]
[10] Image.slidesharecdn.com, 2015. [Online]. Available: http://image.slidesharecdn.com/introductiontobigdatav1-150223035902-conversion-gate01/95/introduction-to-big-data-29-638.jpg?cb=1424770750. [Accessed: 27- Nov- 2015]
[11] R. Kolar, 'Big Data - A Study on Big Data Implementation Challenges', *ResearchGate*, 2015. [Online]. Available: http://www.researchgate.net/publication/276234024_Big_Data_-_A_Study_on_Big_Data_Implementation_Challenges. [Accessed: 27- Nov- 2015]
[12] Sciencedirect.com, 'Data-intensive applications, challenges, techniques and technologies: A survey on Big Data', 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025514000346. [Accessed: 28- Nov- 2015]
[13] N. Khan, I. Yaqoob, I. Hashem, Z. Inayat, W. Mahmoud Ali, M. Alam, M. Shiraz and A. Gani, 'Big Data: Survey, Technologies, Opportunities, and Challenges', *The Scientific World Journal*, vol. 2014, pp. 1-18, 2014.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

[14] Cioinsight.com, "Why Extracting Value from Big Data Is Difficult", 2015. [Online]. Available: http://www.cioinsight.com/it-management/expert-voices/why-extracting-value-from-big-data-is-difficult.html. [Accessed: 15- Dec- 2015]

[15] Cioupdate.com, "How to Extract Information from the Sea of Big Data – Part I -- CIO Update", 2015. [Online]. Available: http://www.cioupdate.com/technology-trends/how-to-extract-information-from-the-sea-of-big-data-part-i.html. [Accessed: 15- Dec- 2015]