International Journal of Advanced Science and Research
ISSN: 2455-4227
Impact Factor: RJIF 5.12
www.allsciencejournal.com
Volume 3; Issue 1; January 2018; Page No. 90-92

A study on various challenges in big data

Gagan Madaan
Assistant Professor, Department of Computer Science & Applications, S.U.S. Panjab University Constituent College, Guru Harsahai, Punjab, India

Abstract
Big Data is the field of database management system that has been used for data collection at large levels. Millions of records have been stored on the big data servers that are indexed properly. The relation between various data instances has been built so that access to the particular data must be easy. Due to vast large value of data big data has various challenges that have to be overcome. These challenges are scalability, availability, security and confidentiality. In this paper various challenges has been discussed that cause various issues in utilization of big data concept.

Keywords: big data, hadoop, scalability, velocity and volume

1. Introduction

1.1 Big Data

Big Data is a popular phrase used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process with traditional database and software techniques. The characteristics which broadly distinguish Big Data are the "3 V's": more volume, more variety and higher rates of velocity. In earlier a relatively small volume of analog data was produced and made available through a limited number of channels, today a massive amount of data is regularly being generated and flowing from various sources, through different channels, every minute in today's Digital Age. In fact, today, we are afloat in Data Ocean. In a broad range of application areas, data is being collected at unprecedented scale.

1.2 Data Management in Big Data Science

Emergence of computer aided research methods is transforming the way how research is done and scientific data are used. The following types of scientific data are defined:

- Raw data collected from observation and from experiment (according to an initial research model)
- Structured data and datasets that went through data filtering and processing (supporting some particular formal model)
- Published data that supports one or another scientific hypothesis, research result or statement
- Data linked to publications to support the wide research consolidation, integration, and openness

1.3 Challenges

Big data offer organization with massive insight; however terabytes or pet bytes 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. 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

2. Review of Literature

Shashi Shekhar *et al* [1] "Spatial big-data challenges intersecting mobility and cloud computing" Increasingly, location-aware datasets are of a size, variety, and update rate that exceeds the capability of spatial computing technologies. This paper addresses the emerging challenges posed by such datasets, which we call Spatial Big Data (SBD). SBD examples include trajectories of cell- phones and GPS devices, vehicle engine measurements, temporally detailed road maps, etc. SBD has the potential to transform society via next-generation routing services such as eco-routing. However, the envisaged SBD-based next-generation routing services pose several significant challenges for current routing techniques. SBD magnifies the impact of partial information and ambiguity of traditional routing queries specified by a start location and an end location. In addition, SBD challenges the assumption that a single algorithm utilizing a specific dataset is appropriate for all situations. Alexandru adrian *et al* [2] "Big

Data Challenges", The amount of data that is traveling across the internet today, not only that is large, but is complex as well. Companies, institutions, he althcare system etc., all of them use piles of data which are further used for creating reports in order to ensure continuity regarding the services that they have to offer. The process behind the results that these entities requests represents a challengefor software developers and companies that provide IT infrastructure. The challenge is how to manipulate an impressive volume of data that has to be securely delivered through the internet and reach its destination intact. This paper treats the challenges that Big Data creates. Nrusimham Ammu Mohd rfanuddin *et al* [3] "Big Data Challenges" Big Data, an umbrella term for the explosion in the quantity and diversity of high frequency digital data. The Big Data may be logs, mobile-banking transactions; online user-generated content such as blog posts and Tweets, online searches, satellite images, etc.-into actionable information requires using computational techniques to unveil trends and patterns within and between these extremely large social economic datasets. These data hold the potential to allow decision makers to track development progress, improve social protection, and understand where existing policies and programmers require adjustment. This paper presents the novel challenges and opportunities associated with Big Data necessitate rethinking many aspects of these data management platforms, while retaining other desirable aspects. Nasser T and Tariq RS *et al* [4] "Big Data Challenges", 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. Yuri Demchenko, Zhiming Zhao *et al* [5] "Addressing Big Data Challenges for Scientific Data Infrastructure" This paper discusses the challenges that are imposed by Big Data Science on the modern and future Scientific Data Infrastructure (SDI). The paper refers to different scientific communities to define requirements on data management, access control and security. The paper introduces the Scientific Data Lifecycle Management (SDLM) model that includes all the major stages and reflects specifics in data management in modern e Science. The paper proposes the SDI generic architecture model that provides a basis for building interoperable data or project centric SDI using modern technologies and best practices. The paper explains how the proposed models SDLM and SDI can be naturally implemented using modern cloud based infrastructure services provisioning model.

3. Challenges

Volume: The volume of data, especially machine-generated data, is exploding, how fast that data is growing every year, with new sources of data that are emerging. For example, in the year 2000, 800,000 petbytes (PB) of data were stored in the world, and it is expected to reach 35 zettabytes (ZB) by 2020.

Variety, Combining Multiple Data Sets

More than 80% of today's information is unstructured and it is typically too big to manage effectively. What does it mean? David Gorbet explains: It used to be the case that all the data an organization needed to run its operations effectively was structured data that was generated within the organization. Things like customer transaction data, ERP data, etc. Today, companies are looking to leverage a lot more data from a wider variety of sources both inside and outside the organization. Things like documents, contracts, machine data, sensor data, social media, health records, emails, etc. The list is endless really.

Velocity

Shilpa Lawande of Vertica defines this challenge nicely [4]: "as businesses get more value out of analytics, it creates a success problem they want the data available faster, or in other words, want real-time analytics and they want more people to have access to it, or in other words, high user volumes."

Scalability

Shilpa Lawande explains: "techniques like social graph analysis, for instance leveraging the influencers in a social network to create better user experience are hard problems to solve at scale. All of these problems combined create a perfect storm of challenges and opportunities to create faster, cheaper and better solutions for Big Data analytics than traditional approaches can solve."

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.

Data integrity

Data Analytics, and Data Mining technologies are only as good as the data. Properly managing, and gathering the data in the first place will save money and time, reduce errors, and simplify the process of gleaning knowledge from a data set. Properly managing data involves a variety of approaches that maintain Data Integrity throughout the lifecycle. When presenting data, it is important to maintain the integrity of that data in order to reduce discrimination and promote inclusion. This, of course, connotes a dangerous realm within this idea of big data, one that enables presentation manipulation and disparate representation, which is particularly relevant to the Asian-American demographic in the United States. The term "Asian-American" represents a group for which there is very

little data integrity maintenance because this data is often presented in a largely aggregated and, therefore, misleading way.

Confidentiality

Confidentiality is an issue of trust that makes interactions possible. For instance, I have given my credit card information to several online companies because one, it makes interactions with them much easier if that information is already stored, and two, I feel confident that they'll keep that information private. There are several exceptions and limits to confidentiality, however, that need to be discussed. The first is that in conducting transactions, companies do share limited amounts of information with third parties. For example, if a person goes to make a large purchase, it's not uncommon for the vendor to call the bank and ask whether the person has enough money in the bank to make that purchase.

4. Conclusion

Big data is the field of data warehousing used to store multiple files and records that occupies thousands of Gigabytes on memory. This data has been stored at different servers so that data can be fetching easily at every location. Amazon, flipkart, Google and Facebook are the major resources that use the concept of big data. Various platforms that are Hadoop, H-Base and many others have been used for database management system. In this paper various challenges have been described that must be resolved for better performance of big data concepts.

5. References

1. Shashi Shekhar "Spatial big-data challenges intersecting mobility and cloud computing", Proceedings of the 11th ACM International Workshop on Data Engineering for Wireless and Mobile Access - In Conjunction with ACM SIGMOD / PODS 2012 (pp. 1-6).
2. Alexandru Adrian "Big Data Challenges", Database Systems Journal vol. I, no 3/2013.
3. Nrusimham Ammu Mohd rfanuddin **"**Big Data Challenges", international Journal of Advanced Trends in Computer Science and Engineering. 2013; 2(1):613-615.
4. Nasser T, Tariq RS. "Big Data Challenges", Nasser and Tariq, J Comput Eng Inf Technol. 2015, 4:3.
5. Yuri Demchenko, Zhiming Zhao Addressing Big Data Challenges for Scientific Data Infrastructure. 2016, 120-130.
6. Jeremy J. Harwood Spectral ageing in the era of big data: integrated versus resolved models, Monthly Notices of the Royal Astronomical Society. 2888-2894.
7. Patrick Wils, Boris T, Gänsicke, Andrew J, Drake John, Southworth "Data mining for dwarf novae in SDSS, GALEX and astrometric catalogues", Monthly Notices of the Royal Astronomical Society. 2010, 436-446.
8. Dodd RJ. Monthly Notices of the Royal Astronomical Society. 2016, 959-972.
9. Vid Podpečan, Monika Zemenova, Nada Lavrač "Orange4WS Environment for Service-Oriented Data Mining" The Computer Journal. 2012, 82-98.
10. Zhenyu Zhou, Houjian Yu, Chen Xu, Yan Zhang, Shahid Mumtaz, Jonathan Rodriguez "Dependable Content Distribution in D2D-Based Cooperative Vehicular Networks: A Big Data-Integrated Coalition Game Approach" IEEE Transactions on Intelligent Transportation Systems. 2018, 1-12.
11. Ivano Notarnicola, Ying Sun, Gesualdo Scutari, Giuseppe Notarstefano "Distributed big-data optimization via block-iterative convexification and averaging", Decision and Control (CDC), 2017.
12. Mostafa Rahmani, George Atia. "Robust and Scalable Column/Row Sampling from Corrupted Big Data" Computer Vision Workshop (ICCVW). 2017, 112-125.