APPLICATIONS FOR BIG DATA BY USING OPTIMIZATION TECHNIQUES

KAPPALA SHANTHI LATHA
Research Scholar
Dept of Computer Science & Engg
Arni University-Himachal Pradesh

DR. PRASADU PEDDI
Research Supervisor
Dept of Computer Science & Engg
Arni University-Himachal Pradesh

DR. MARAM ASHOK
Co- Supervisor
Dept of Computer Science & Engg
Principal & Professor
Malla Reddy Institute of Engineering & Technology,
Hyderabad, Telangana

Abstract

As the world is getting digitized the speed in which the amount of data is over owing from different sources in different format, it is not possible for the traditional system to compute and analysis this kind of big data for which big data tool like Hadoop is used which is an open-source software. It stores and computes data in a distributed environment. In the last few years developing Big Data Applications has become increasingly important. In fact, many organizations are depending upon knowledge extracted from huge amount of data. However traditional data technique shows a reduced performance, accuracy, slow responsiveness and lack of scalability. It stores and computes data in a divided environment. Since a decade Big Data Application development has become increasingly paramount. Many organizations are relied on getting knowledge essence from a huge amount of data. However classic data technique demonstration includes reduced performance, accuracy, slow responsiveness and lack of scalability. To resolve the complicated Big Data problem, many of the work has been carried out. For that various types of technologies have been developed. This research paper focuses on the survey of recent optimization technologies and their Applications developed for Big Data.

Keywords: *Big Data, Hadoop, Optimization*

Introduction

Data is one of the most important and vital aspect of different activities in today's world. Therefore, vast amount of data is generated in each and every second. A quick development of data in current time in different domains requisite an intellectual data analysis tool that would be useful to fulfill the requirement to analysis a vast quantity of data. Big data is a set of data groups that are so huge and complex so it is risky to manage the database using just one tool or conventional data processing APPs. Challenges comprise capture, duration, storage, research, participation, transport, analysis, and visualization. The trend to huge data-sets is due to the additional information derived from the analysis of one large set of relevant data, compared to smaller separate groups with the same total data volume from 2012 the size specified on the data sets suitable for processing in a plausible amount of time was subject to the exabyte measurement unit. Scientists often face many constraints due to big data sets in many areas, including meteorology, genetics, complex physical simulation, and biological and environmental research. Restrictions also affect Internet search, business technology, and finance. Data sets are increasing in volume on one side due to their influence in sensor wireless, sensor networks, transmitter frequency sensors, mobile information sensors, microphones, program registers, and cameras.

Literature review

Kenglung Hsu [2022] Communication operators are paying more and more attention to the value of data and are demanding more and bigger data technologies. Many companies have started to take advantage of their resources to tap the value of data and develop their own core business. Then the corresponding scheduling module architecture process is designed and built, the corresponding scheduling rules and related scheduling information field tables are designed, and the data aggregation storage is improved. The program scheduling module was designed to be more lightweight and easier to use, and the data migration module increased the timeliness of data migration.

Mohammad Alhowaidi [2021] Named Data Networking (NDN) is a promising approach to provide fast in-network access to compact muon solenoid (CMS) datasets. It proposes a content-centric rather than a host-centric approach to data retrieval. Data packets with unique and immutable names are retrieved from a content store (CS) using Interest packets. The current NDN architecture relies on forwarding strategies that are only dependent upon on-path caching. Such a design does not take advantage of the cached content available on the adjacent off-path routers in the network, thus reducing data transfer efficiency. In this work, we propose a software-defined, storage-aware routing mechanism that leverages NDN router cache-states, software defined networking (SDN) and multipath forwarding strategies to improve the efficiency of very large data transfers.

Big data

Big data refers to data that contains complex, extensive data sets within it, typically from newer data sources on the market. It encompasses a variety of data from various sources, including artificial intelligence, social media, the internet, smartphones and apps. The term big data is a reference to the massive amounts of data these streams process, so much so that most processing software programs can't manage all of this information.

Big Data Optimization

For large business enterprises, it is indispensable to process large-scale data (Big Data) to get better insight into the business. There is no doubt that, processing of Big Data has become a challenging task, although many tools and techniques are available to process this flood of data. However, to process Big Data in an optimized way such that its overall performance doesn't degrade is a challenging task. Increasing rate of data will become critical to handle in future, hence, proper optimization techniques need to be applied for Big Data processing.

A. Volume

The main feature that makes data "big" is the phenomenal volume and in these days, we are in the situation where every second we are generating a huge amount of data from Twitter messages, WhatsApp, Facebook, photos, sensor data and video clips that we produce and post it on social media every second. We are talking about zettabytes or brontobytes of data. The huge volume of the data needs space and different processing technologies than classic storage and processing space.

B. Variety

One of the reasons why we call data the huge data is the variety. We get the Big Data from a big variety of provenance and in general that data are three species:

1. Structured Data The structured data is data that can be stored, accessed it and processed it in the form of a stable format. Structured data Indicates to high-level data

types of organization, like information in a relational database.

2. Semi-structured Data

The semi-structured data is data or information that has not resided in a relational database but it has some Organizational characteristics that make it easy for us to analyze and store it in a relational database. Examples of this type: JSON and XML documents are semi-structured, NoSQL databases also as semi-structured.

3. Unstructured Data If the data do not have structure or any Organizational characteristics to classify it as semi-structured referred to as unstructured data. In other words, any data has unknown form or unknown structure is categorize as unstructured data. It often includes multimedia and text content.

C. Veracity

Veracity refers that the data being analyzed is of high quality and accurate, which are reflected in sound engineering, decision making based on analysis of these data. In contrast, data of low resolution and low quality contain a lot of data to be discarded or what we called noise In order to get rid of this data, which is not valuable and because it is of large size and has high speed, we need to use advanced tools.

D. Velocity

Though, we call data large data, it must be generated very fast at the same time we should analyze it and process it fast also to get the information the faster we process your information we can make the right decision at the right time.

Scalability: Big data scalability issue leads to cloud computing. There is a high-end resource sharing involved in it which is expensive and also challenges to efficiently run various jobs so that the goal of each work load cost effectively. System failure is also dealt efficiently while working on large data clusters. These combined factors make it difficult to write program, even machine learning complex tasks. Changes are being made in technologies used, solid state drive is used in place of hard drive and phase change technologies are not performing as good as sequential and random transfer. Thus it's a big question that what kind of storage devices should be used to store data.

Quality of Data: Collecting and storing huge amount of data are costly; if more data is stored for predictive analysis and decision making in business then it will yield better results. Big data has interest in having quality rather than unused data so that better conclusion and result can be drawn. This further arise the question on whether the data is relevant or not, can the stored data be accurate enough to produce right result, how much data would be enough in making decision.

Big Data Analytics

Nowadays, the data that need to be analyzed are big, contained heterogeneous data types, and even including streaming data which may change the statistical and data analysis approaches. Therefore, Traditional tools cannot be able to analyze this category of data. So, New approach of big data called 'Big Data Analytics' was born. "Big data analytics" refers to advanced technologies designed to work with large volumes of heterogeneous data in order to improve the traditional Data Analytics Process mentioned in below figure.

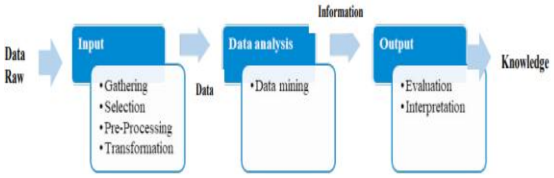

Figure: Data analytics process

A lot of researchers talk about sophisticated types of analytics techniques as shown in figure:

Descriptive Analytics, Predictive Analytics and Prescriptive Analytics:

Descriptive Analytics: Gives information about What happened. In this technique, based on historical data, new insights are developed using statistical descriptions (such as Statistic Summary, Correlations and Sampling…) and Clustering (such as K-means…)

Predictive Analytics: Predicts the future outcomes using new statistical methods and predictive algorithms such as 'Decision Tree'. It provides information on what is likely happen in the future and what actions can be taken.

Prescriptive Analytics: It is a type of predictive analytics. It helps to derive a best possible outcome by analyzing the possible outcomes by responding to the question So what? Now What?

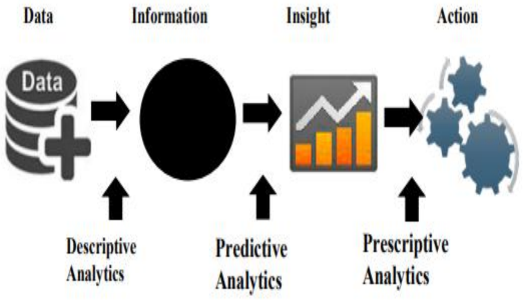

Figure: Types of Analytics Techniques

Big Data Applications

big data and business process management (bpm) synergy solutions

Business Process (BP) is a succession of activities designed by human and systems that intends to achieve business goals4]. Business Process Management (BPM) is a way to collect and treat data outcoming from processes in real time to support decision making. The lifecycle of BPM Project contains successive steps as it's shows in

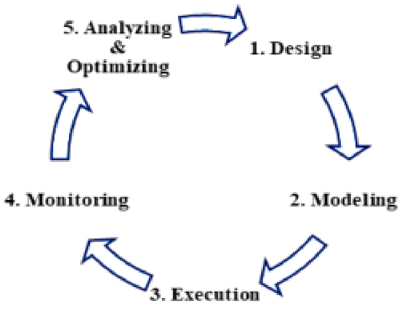

Figure: BPM lifecycle

Design: Modeling the design of what is currently being used and what will be used.

Modeling: Analyzing process and performing "what if" analysis and then comparing the various process options to determine optimal improvements.

Execution: Once the processes have been designed and simulated, they will be integrated into the information system for execution.

Monitoring: Managing and supervising the processes.

Analysis and optimization: Iterate for continuous improvement.

By the appearance of Big Data, organisations will need to integrate mobile data, social networks, digital video, and sensor data into their Business Process. So, they must be aware of Big Data challenges in order to make an efficient and intelligent BP that is aims to bring huge value for process decision makers and process actors. Some decisions are based on subjective judgment however, increasingly important decisions need to be based on hard data based on big data analytics.

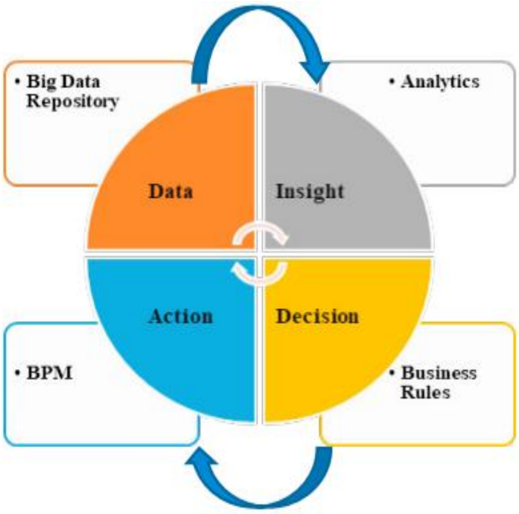

Figure: Big data-BPM lifecycle

The big data opens up BPM to new data sources and opportunities to analyze processes more deeply and simulate potential improvements:

Network Optimization: Today, with big data analytics techniques, operators seek to collect, store and analyze data generated by user devices and network devices like: routers, switches, base stations, and so one. All of these data and others Key performance Indicators (KPIs) can help operators to well monitor network performance to ensure proper operation without problem.

Data Monetization: Some of telecommunication companies see these massive volumes of data as a marketable resource to generate new revenue by aggregating and selling these data, they search to turn that data into money. Many telecom operators also seek to commercially exploit this customer information, i.e. to generate new revenue by aggregating and selling this data. Some of them consider this an excellent financial opportunity. To sum up, promoting loyalty, anticipating and reducing churn, offering upselling and cross-selling, optimizing network and personalization services are key areas where telecom operators can take advantage of Big Data Analytics.

Conclusion

The Big data concept has progressively become the next evolutionary phase in batch processing, storing, manipulation and relations visualization in vast number of records. The era of big data is upon us, bringing with it an urgent need for advanced data acquisition, management, and analysis mechanisms. In this study, we have presented the concept of big data, challenges, related technologies, applications and highlighted the big data value chain, which covers the entire big data lifecycle. Although major innovations in analytical techniques for big data have not yet taken place. Our future work is planned in the sense of the contribution to the good governance of "Big Data Analytics Systems" since Big Data is still in the development and its related techniques and tools are far from mature. For instance, real-time analytics will likely become a profitable field of research because of the remarkable growth in location-aware social media and mobile apps.

References

1. *Kenglung Hsu (2022) "Big data analysis and optimization and platform components, Journal of King Saud University - Science, ISSN 1018-3647, Volume 34, Issue 4, 2022, 101945, https://doi.org/10.1016/j.jksus.2022.101945.*
2. *Mohammad Alhowaidi (2021) "Cache management for large data transfers and multipath forwarding strategies in Named Data Networking", Computer Networks, ISSN:1389-1286, Volume 199, 108437, https://doi.org/10.1016/j.comnet.2021.108437.*
3. *M. A. Memon, S. Soomro, A. K. Jumani, and M. A. Kartio, (2017) "Big Data Analytics and Its Applications," vol. 1, no. 1.*
4. *M. Dave and H. Gianey, (2017) "Different clustering algorithms for Big Data analytics: A review," Proc. 5th Int. Conf.*

Syst. Model. Adv. Res. Trends, SMART 2016, pp. 328–333.

5. *C. M. Chen, (2016) "Use cases and challenges in telecom big data analytics," APSIPA Trans. Signal Inf. Process., vol. 5, no. 2016, pp. 1–7.*
6. *Eyman Yosef (2018) "Big Data Flow Adjustment Using Knapsack Problem", Journal of Computer and Communications, ISSN: 2327-5227, Vol.6, No.10, Pp.30-39.*
7. *Hao Peng (2018) "Optimizing the induction chemotherapy regimen for patients with locoregionally advanced nasopharyngeal Carcinoma: A big-data intelligence platform-based analysis", Oral oncology, ISSN: 1879-0593, Volume.79, Pp.40-46. doi: 10.1016/j.oraloncology.2018.02.011.*
8. *Muhammad Habib ur Rehman (2016) "Big Data Reduction Methods: A Survey", Data Science and Engineering, Volume.1, pages265–284, https://doi.org/10.1007/s41019-016-0022-0.*