A Review on Data Generation using Big Data

Om Kumar[1], Dr.Ravi Kumar Singh Pippal, Abhinav Shukla[3]
M.Tech (CSE) Scholar, Department of CSE, RKDF University Bhopal, India[1]
Head of Department, Department of CSE, RKDF University Bhopal, India[2]
Assistant Professor, Department of ECE, RKDF University Bhopal, India[3]
omkumar2201@gmail.com[1], ravesingh@gmail.com[2], abhinav.shukla@hotmail.com[3]

Abstract - The term "big data" is increasingly prevalent in today's discussions. Big data and social networks are closely linked, as much of the data generated nowadays comes from social networking sites. However, the challenge lies not in collecting this data but in effectively managing it. Many fields and subjects, spanning from everyday life to traditional research areas, present various issues related to big data. The proliferation of different types of networks has diversified the types of data, issues, and solutions associated with big data.In this review, recent research on data types, storage models, privacy, data security, analysis methods, and applications related to network big data is explored. Finally, we summarize the challenges and developments in the field of big data, aiming to predict current and future trends..

I. INTRODUCTION

Big data is closely associated with the speed at which we can interact with and incorporate vast amounts of information. It is a growing trend in many industries, offering a means to improve and streamline business processes. It spans various fields and sectors, including economic and business activities, public administration, national security, and scientific research. Explore our data analytics platform to understand why it is considered one of the most advanced technologies for handling big data in the world..

II.PROBLEM STATEMENT

The large number of fields and subjects, ranging from traditional research fields involve different problems. Big data and social networks are interdependent, because most of today's data are generated from social networking sites.The popularizing of various types of network has diversified types, issues, and solutions for big data more than ever before. Review recent research in data types, Storage models, privacy, data security, analysis methods, and applications related to network big data.

III. OBJECTIVES

The main focuses of network data is online social network data, such as Facebook. This focus has expanded with developments in the data analysis.Many studies have been performed with OSN using knowledge about representative characteristics at the macro level, for instance, small world features for the potential micro processes are not well represented in these studies.

IV. PURPOSE OF SYSTEM

Organizations are living in an era of big data and it incorporates endless amount of information. Many industries, it is growing, providing a means to improve and streamline business. The many fields and sectors, ranging from economic and business activities to public administration, from national security to scientific research in many areas, are involved in the problem.

V.SYSTEM ARCHITECTURE

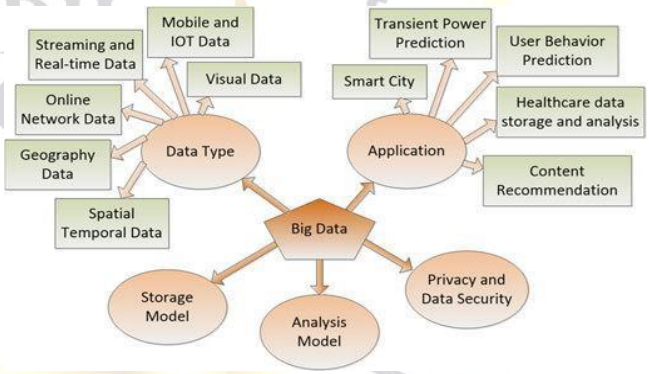

Fig 5.0 Architecture of Big Dataanalytics

The era of big data has produced a variety of datasets from different sources in different domains. Big data analytics is classified into five types.

- Data Types
- Storage Model
- Analysis Model
- Privacy and Data Security
- Application

A. Data Types

Datasets consist of multiple modalities, each of which has a different representation, distribution, scale, and density. It provide to unlock the power of knowledge from disparate datasets is of paramount importance in big data research, essentially distinguishing big data from traditional data mining tasks.

a.

The main focuses of network big data is online social network data, such as face book and focus has expanded with developments in the data analysis. Many studies have been performed with OSNs using knowledge about representative characteristics at the macro level, for instance, small world features. However, factors for the features of potential micro processes are not well represented in these studies.

b. Mobile and IoT Data

Atrend in network big data is the analysis of mobile and IoT data with the development of 5G technology, converged mobile networks have resulted in significant improvements in machine to- machine communications performance. Integrated mobile webs share unlicensed spectrum bands in cellulite networks,such as long-term evolution-advanced, by using cognitive radio technology. This network generates large volumes of data, compared to former mobile networks.

c.Geography Data

OSN data will soon include geographic data along with OSN interaction, for example, geo-tag real-time geographic data .Location-based data will soon expand beyond terrain. One study addresses the gauntlets of major forms of technologyfor three-dimensional (3-D) interaction and volume rendering technology on the basis of graphics processing unit (GPU) technology.This work explores visual software for the hydrological environment based on data orientation and it produces ocean plans, contour mapping of surfaces, element field mapping, and dynamic simulation of the existing field .To better present features in space and achieve real-time upgrading of a large amount of hydrological environment data, the study constructs nodes on the spot for the control of geometry to achieve dynamic mapping of high properties.

d. Spatial Temporal Data

It can be accompanied by online streaming services, network big data changes from simple OSN data to spatial temporal data. In the recent years, the volume of available data from space has increased substantially. Data are classified in many categories on the basis of features and differences.

Since the differences in data determine the success of the analysis, they play an essential role. Different features are also applied to search for the same features and some studies attach importance to time changing data and data with time sequences. With respect to network big data,

some same-feature searching methods for time-changing data were discussed and explored.

e. Streaming and Real-Time Data

The rise in online streaming services, network big data has evolved from spatial temporal data to real time spatial temporal data. Network surveys in general require ongoing statistics over large capacity data streams.

f. Visual Data

In the era of big data, the most difficult problems that remain to be solved are how to efficiently deal with large quantities and varieties of data. There are many analytical theories and models. In this section, recent discoveries in big data storage and analysis models are surveyed.

B. Storage Model.

The increasing amount of image data has posed significant challenges to modern image analysis and retrieval. Proposed weakly semi supervised deep learning for the multi label image annotation approach which was inspired by recent advances in deep learning research. In weighted pair wise ranking loss is effectively utilized to handle weakly labelled images while a triplet similarity loss is employed to harness unlabelled images.

C. Privacy and Data Security

The scope of big data, safety and privacy protection is a crucial problem .There may be risks of privacy violations at each step and many methods for privacy protection, e.g., encryption. The popularity of big data depends on a complete understanding of the safety problems inherent within the system. Safety is a new concern, and this paper mainly introduces the concept of privacy using new problems, and focuses on efficiency and privacy protection.

The study specializes in the structure of data analytics, demonstrating the requirements for privacy protection, it explains the safety protection cosine similarity agreement in data mining and requirements. In the area wearable sensors collect data that are often sensitive and must be protected. Compared to former methods, this method provides more reliable and available privacy protection. The experiments prove that this method has sufficient privacy protection, even when hackers have adequate knowledge of the system.

D. Analysis Method

The methods used for big data analysis are Map Reduce-related. For data control in the past, instruments for analysing data were insufficient in depot and exploring systems. The models used by big data researchers are usually inspired by mathematical ease of exposition. By virtue of the essence of big data, it is memorized in a dispersed document system framework. Hadoop and HDFS by Apache are extensively applied in memorizing and controlling big data. These technologies are applied in MapReduce.

E. Applications

A data and interactions are generated in every form of human behaviour, big data is used in almost all aspects of life. Big data increasingly benefits both research and industrial fields, such as Healthcare, financial series and

a. Transient Power Prediction

The prediction of transient power is valid in both distributed and streaming data. ML was used in the study. In the classifier cultivation stage, researchers regard the tremendous amount of data from the past as a dispersed study target, and establish evaluation principles regularly.Designed a naive Bayes-category approach based on MapReduce handling, creating a map-and-decrease procedures method for calculating the chance rate of being tested in advance and the chance rate for conditions in dispersed.

b. User Behaviour Prediction

Many of the network big data predictions are based on data from OSNs. Big data is used for predictions based on ranked data, such as elections, car performance, and other areas in business and politics. One study discussed modelling and analysis approaches to democracy, as well as various cases of big data from elections;

c. Healthcare Data Storage and Analysis

Big data in health and biology to tackle the challenges in new models is becoming significant. One study introduced two uses of health, which gathers electronic medical records that are used for health services terminals .One is a blended system that enhances the user experience in high- pressured oxygen halls using virtual reality (VR) glasses, which creates the feeling of being inside it. The other is a sound interaction game that is used by patients as a possible measurement for supplementary recovery tools. It is possible to analyse recordings of the sounds made by patients to assess long-term recovery results and further forecast the recovery process.

d. Content Recommendation

This study presents a movie recommendation system based on scores provided by users and view of the movie evaluation system, the impacts of access control and multimedia security are analyzed, and a secure hybrid cloud storage architecture is presented. Mobile-edge computing technology is used in the public cloud, which guarantees high-efficiency requirements for the transmission of multimedia content.

The processes of the system, including registration, user login, role assignment, data encryption, and data decryption, are also described. Personalized travel sequence recommendation was proposed in another study, which uses travelogues, community contributed photos, and heterogeneous Meta data associated with the photos.

e. Smart City

A 3-D Shenzhen city web platform based on a network virtual reality geographic information system (GIS) was put forward. A 3-D worldwide browser is applied to load different kinds of required data from a city, such as 3-D construction model data, inhabitants' messages, and traffic data from the past and present. Traffic visual analysis systems based on a virtual reality GIS represent the standard by which traffic data are controlled and developed. Aside from the fundamental GIS mutual functions, the system put forward also contains smart functions for visual analysis and forecast accuracy.

VI. New Technologies

The Big Data produced by social networks can be analyzed by current computer technologies .MapReduce, Hadoop and NoSQL techniques have supported distributed data storage, parallel data retrieval and processing. Many analytical methods and algorithms are designed for business analytics, such as: k-means clustering, Association rules, Linear/logistic regression, and Time series.

a. Hadoop

Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

It is suitable for the distributed storage and processing. Hadoop provides a command interface to interact with HDFS. The built-in servers of name node and data node help users to easily check the status of cluster. Streaming access to file system data and HDFS provides file permissions and authentication.

b. NoSQL

NoSQL is a non-relational DMS that does not require a fixed schema, avoids joins, and is easy to scale. NoSQL database is used for distributed data stores with humongous data storage needs. It is used for Big data and real-time web apps.Handle structured, semi-structured, and unstructured data with equal effect and a non-relational DMS, that does not require a fixed schema, avoids joins, and is easy to scale.

The concept of NoSQL databases became popular with Internet giants like Google, Facebook, Amazon, etc. who deal with huge volumes of data. Databases never follow the relational model it is either schema-free or has relaxed schemas.

It can serve as the primary data source for online applications. big data which manages data velocity, variety, volume, and complexity, Excels at distributed database and multi-data center operations. Eliminates the need for a specific caching layer to store data.

c. MapReduce

MapReduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault- tolerant manner.A MapReduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then

input to the reduce tasks. Typically both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks.

The compute nodes and the storage nodes are the same, that is, the MapReduce framework and the Hadoop Distributed File System running on the same set of nodes. This configuration allows the framework to effectively schedule tasks on the nodes where data is already present, resulting in very high aggregate bandwidth across the cluster.

MapReduce framework consists of a single masterJobTracker and one slave Task Tracker per cluster-node. The master is responsible for scheduling the jobs' component tasks on the slaves, monitoring them and re-executing the failed tasks. The slaves execute the tasks as directed by the master and the value of big data continues to remain a challenge, other practical challenges including funding and return on investment and skills continue to remain at the forefront for a number of different industries.

1. Trying to decide whether there is true value in big data or not

2. Evaluating the size of the market opportunity

3. Developing new services and products that will utilize big data

4. Already utilizing big data solutions Re-positioning existing services and products to utilize big data

5. Already utilizing big data solutions

Consumers expect rich media on-demand in different formats and in a variety of devices, some big data challenges in the communications, media and entertainment industry include:

- Collecting, analyzing, and utilizing consumer insights
- Leveraging mobile and social media content

VII. Result and Analysis

The ability to analyze big data provides unique opportunities for the organization as well and able to expand the kind of analysis can do. Instead of being limited to sampling large data sets, you can now use much more detailed and complete data to do your analysis. However, analyzing big data can also be challenging. Changing algorithms and technology, even for basic data analysis, often has to be addressed with big data.

VIII. CONCLUSION

Big data is a field focused on techniques for systematically analyzing, extracting information from, or handling data sets that are too large or complex for traditional data-processing software. Data with many cases can provide greater statistical power, but data with higher complexity may lead to a higher rate of false discoveries. Challenges in this field include capturing data, data storage, data analysis, search, sharing, and more.

Big data encompasses various aspects including data transfer, visualization, querying, updating, information privacy, and data sources. Building on existing research studies, this research aims to analyze, synthesize, and provide a comprehensive structured analysis of big data to guide future research directions.

REFERENCES

[1] 1] C. Min, M. Shiwen, and L. Yunhao, "Big data: A survey," Mobile Netw. Appl., no. 2, pp. 171–209, Apr. 2014.

[2] 2] S. Chris, U. Matzat, and U.-D. Reips, "Big data: Big gaps of knowledge in the field of internet science," Int. J. Internet Sci., vol. 7, no. 1, pp. 1–5,2012.

[3] 3] J.Minho, T. Maksymyuk, R. L. Batista, T. F. Maciel, A. L. F. de Almeida, and M. Klymash, "A survey of converging solutions for heterogeneous mobile networks," IEEE Wirel. Commun., vol. 21, no. 6, pp. 54–62, Dec. 2014.

[4] 4] J. Minho, L. Han, N. D. Tan, and H. P. In, "A survey: Energy exhausting attacks in MAC protocols in WBANs," Telecommun. Syst., vol. 58, no. 2, pp. 153–164, 2015

[5] 5] B. Mitra, N. Meratnia, and P. J. M. Havinga, "On the use of mobility data for discovery and description of social ties," in Proc. 2013 IEEE/ACM Int Conf. Adv. Social Netw. Anal. Mining, 2013, pp. 1229–1236

[6] 6] M. Marco and S. Valtolina,"Towards a user-friendly loading system for the analysis of big data in the internet of things," in Proc. 2014 IEEE 38th Int. Comput. Softw. Appl. Conf. Workshops, 2014, pp. 312–317.

[7] 7] S. H. Thiago, P. O. S. Vaz DeMelo, J.M. Almeida, and A. A. F. Loureiro,"Large-scale study of city dynamics and urban social behavior using participatory sensing," IEEE Wirel. Commun, vol. 21, no. 1, pp. 42–51, Feb.2014.

[8] 8] H. Guo, X. Li, W. Wang, Z. Lv, Wu, and W. Xu, "An event-driven dynamic updating method for 3D geo-databases.