ISSN:0975-9646

Data Mining Application for Big Data Analysis

[1]Bharati Punjabi [2]Prof. Sonal Honale

[1]*Scholar - M.Tech. CSE,Abha Giakwad Patil College of Engineering,Nagpur.*
[2]*Professor – Deptt. Of Mtech CSE Abha Giakwad Patil College of Engineering,Nagpur.*

Abstract : Data mining is the application of specific algorithms for extracting patterns from data. Big Data is a new term used to identify the datasets that due to their large size and complexity, we cannot manage them with our current methodologies or data mining software tools. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. Processing or analyzing the huge amount of data or extracting meaningful information is a challenging task. The term "Big data" is used for large data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Difficulties include capture, storage, search, sharing, analytics and visualizing. Typical examples of big data includes web logs, RFID generated data, sensor networks, , social data from social networks, Internet text and documents, Internet search indexing, call detail records, astronomy, atmospheric science, genomics, biogeochemical, biological, and other complex and/or interdisciplinary scientific research, military. Surveillance, medical records, photography archives, video archives, and large-scale ecommerce.

Keywords— Big Data Problem, Hadoop Distributed File System, Parallel Processing, Hadoop cluster, MapReduce.

I. INTRODUCTION

Applications where data collection has grown tremendously and is beyond the capability of commonly used software tools to capture, manage, and process within a "tolerable elapsed time." The most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions . In many situations, the knowledge extraction process has to be very efficient and close to real time because storing all observed data is nearly infeasible.

Data is being produced at an ever increasing rate. This growth in data production is being driven by individuals and their increased use of media; organisations; the switch from analogue to digital technologies; and the proliferation of internet connected devices and systems.

There has also been an acceleration in the proportion of machine-generated and unstructured data (photos , videos, social media feeds and so on) compared to structured data such that 80% or more of all data holdings are now unstructured and new approaches and technologies are required to access, link, manage and gain insight from these data sets.

The commonly accepted definition of big data comes from Gartner who define it as high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight, decision making, and process optimization. These are known as the "three Vs". Some analysts also discuss big data in terms of value (the economic or political worth of data) and veracity (uncertainty introduced through data quality issues). Government agencies hold or have access to an ever increasing wealth of data including spatial and location data, as well as data collected from and by citizens. Experience suggests that such data can be utilised in ways that have the potential to transform service design and delivery so that personalised and streamlined services, that accurately and specifically meet individual's needs, can be delivered to them in a timely manner. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it.

Apache Hadoop – It is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are commonplace and thus should be automatically handled in software by the framework.

The core of Apache Hadoop consists of a storage part (Hadoop Distributed File System (HDFS)) and a processing part (MapReduce). Hadoop splits files into large blocks and distributes them amongst the nodes in the cluster. To process the data, Hadoop MapReduce transfers packaged code for nodes to process in parallel, based on the data each node needs to process. This approach takes advantage of data locality—nodes manipulating the data that they have on hand—to allow the data to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are connected via high-speed networking.

The base Apache Hadoop framework is composed of the following modules:

Hadoop Common – contains libraries and utilities needed by other Hadoop modules;

Hadoop Distributed File System (HDFS) – is a distributed file system providing fault tolerance and designed to run on commodity hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. Hadoop provides a distributed file system (HDFS) that can store data across thousands of servers, and a means of running work (Map/Reduce jobs) across those machines, running the work near the data. HDFS has master/slave architecture. Large data is automatically split into chunks which are managed by different nodes in the hadoop cluster.

HBASE-HBase is a column-oriented database management system that runs on top of HDFS. It is well suited for sparse data sets, which are common in many big data use cases. Unlike relational database systems, HBase does not support SQL.In fact, HBase isn't a relational database at all. HBase applications are written In Java much like a typical MapReduce application.

Hadoop MapReduce – a programming model for large scale data processing. Map reduce is a software framework introduced by Google in 2004 to support distributed computing on large data sets on clusters of computers. Map Reduce is a programming model for processing and generating large data sets. Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs and a reduce function that merges all intermediate values associated with the same intermediate key.

Map step: The master node takes the input, partitions it up into smaller sub-problems, and distributes them to the worker nodes. A worker node may repeat in turn, which leads to a multi-level tree structure. The worker node processes the smaller problem, and then passes the answer back to the node acting as a master node for it . Map takes one pair of data with a type in one data domain, and returns a list of pairs in a different domain: Map (k1, v1) → list (K2, v2)

Reduce step: The master node in turn collects the answers to all the smaller sub-problems and then combines them in such a way to form the output – the answer to the problem it was originally trying to solve. The Reduce function is then applied in parallel to each group, which in turn produces a collection of values in the same domain: Reduce (K2, list (v2)) → list (v3)

II. RELATED WORK

Various research work in this area are made by the authors in the following national projects that all involve Big Data components:

Integrating and mining biodata from multiple sources in biological networks, sponsored by th US National Science Foundation, Medium Grant No. CCF-0905337, 1 October 2009 - 30 September 2013.

Issues and significance. We have integrated and mined biodata from multiple sources to decipher and utilize the structure of biological networks to shed new insights on the functions of biological systems. We address the theoretical underpinnings and current and future enabling technologies for integrating and mining biological networks. We have expanded and integrated the techniques and methods in information acquisition, transmission, and processing for information networks. We have developed methods for semantic-based data integration, automated hypothesis generation from mined data, and automated scalable analytical tools to evaluate simulation results and refine models.

Big Data Fast Response. Real-time classification of Big Data Stream, sponsored by the Australian Research Council (ARC), Grant No. DP130102748, 1 January 2013 - 31 Dec. 2015.

Issues and significance. We propose to build a stream-based Big Data analytic framework for fast response and real-time decision making. The key challenges and research issues include: - designing Big Data sampling mechanisms to reduce Big Data volumes to a manageable size for processing; - building prediction models from Big Data streams. Such models can adaptively adjust to the dynamic changing of the data, as well as accurately predict the trend of the data in the future; and - a knowledge indexing framework to ensure real-time data monitoring and classification for Big Data applications. .

Pattern matching and mining with wildcards and length constraints, sponsored by the National Natural Science Foundation of China, Grant Nos. 60828005 (Phase 1, 1 January 2009 - 31 December 2010) and 61229301 (Phase 2, 1 January 2013 - 31 December 2016).

Issues and significance. We perform a systematic investigation on pattern matching, pattern mining with wildcards, and application problems as follows: - exploration of the NP-hard complexity of the matching and mining problems, - multiple pattern matching with wildcards, - approximate pattern matching and mining, and - application of our research onto ubiquitous personalized information processing and bioinformatics.

Key technologies for integration and mining of multiple, heterogeneous data sources, sponsored by the National High Technology Research and Development Program (863 Program) of China, Grant No. 2012AA011005, 1 January 2012 - 31 December 2014.

Issues and significance. We have performed an investigation on the availability and statistical regularities of multisource, massive and dynamic information, including cross-media search based on information extraction, sampling, uncertain information querying, and cross-domain and cross-platform information polymerization. To break through the limitations of traditional data mining methods, we have studied heterogeneous information discovery and mining in complex inline data, mining in data streams, multigranularity knowledge discovery from massive multisource data, distribution regularities of massive knowledge, quality fusion of massive knowledge. Group influence and interactions in social networks, sponsored by the National Basic Research 973 Program of China, Grant No. 2013CB329604, 1 January 2013 - 31 December 2017.

Issues and significance. We have studied group influence and interactions in social networks, including - employing group influence and information diffusion models, and deliberating group interaction rules in social networks using dynamic game theory, studying interactive individual selection and effect evaluations under social networks affected by group emotion, and analyzing emotional interactions and influence among individuals and groups, and - establishing an interactive influence model and its computing methods for social network groups, to reveal the interactive influence effects and evolution of social networks.

III. PROPOSED SYSTEM

The Proposed system uses Hbase. It’s a column oriented database management system that runs on top of HDFS. It is well suited for sparse data sets, which are common in many big data use cases. Unlike relational database systems, HBase does not support SQL.In fact, HBase isn’t a relational database at all. HBase applications are written In Java much like a typical MapReduce application.

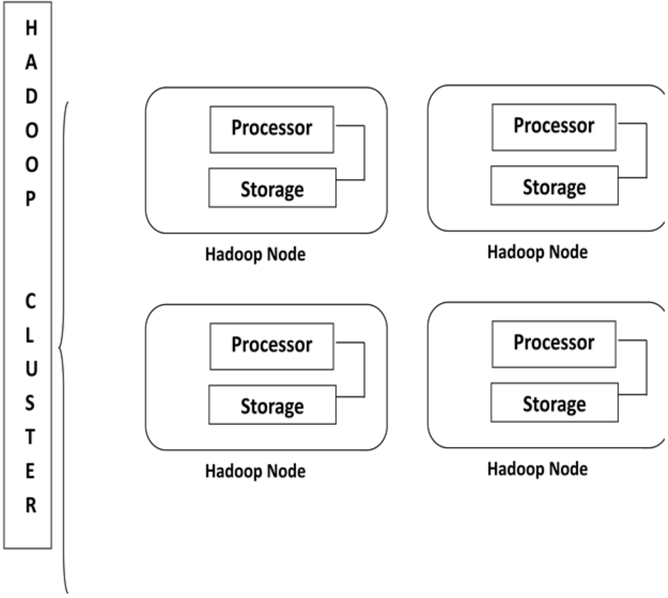


Fig. 1 Proposed system

IV. REQUIREMENTS

Input: - What acts as input to the system?

For instance: - Unformatted data.

Output: - What exactly is the expected output from the analyzed input?

Functional Requirements:-It defines the function of the System (BigData). In this case, we are analyzing large scale of Data Sets and their retrieval.

Non-Functional Requirements:-It specifies the criteria that can be used. For instance: - Hadoop and MapReduce, in this case. Simple, it defines how a system is supposed to perform.

V. DESIGNING

Components of the system:-

- Client System:-System used by user for accessing, through a browser.
- Interface Layer:-Provides the user interface for the client system. From here the user can access all the functional operation of this system.
- Processing Component:-It will provide different functional capabilities to perform different operations on data such as reading, display analyzed input, performing queries etc.
- Hadoop Cluster:-Stores and manages the huge amount of data.

VI. RESULT OF PROJECT

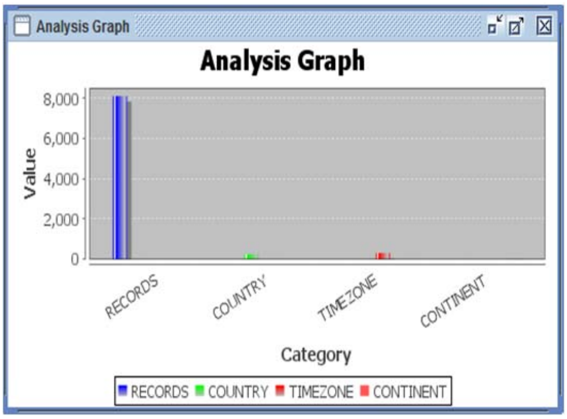
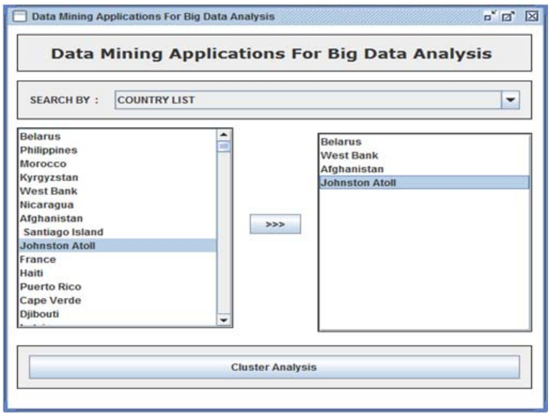


Fig. 2 Overall data analysis graph



Data Mining Applications For Big Data Analysis

SEARCH BY : COUNTRY LIST

Belarus
Philippines
Morocco
Kyrgyzstan
West Bank
Nicaragua
Afghanistan
Santiago Island
Johnston Atoll
France
Haiti
Puerto Rico
Cape Verde
Djibouti

Belarus
West Bank
Afghanistan
Johnston Atoll

Fig. 3 Countrywise data display

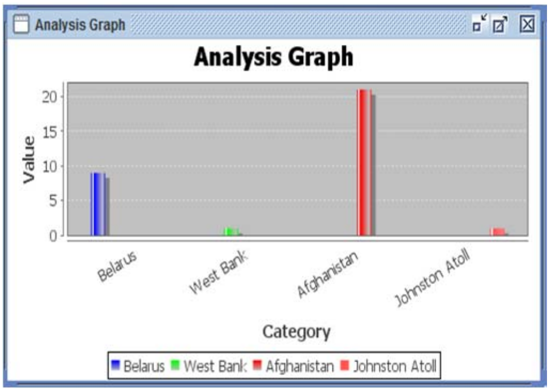


Fig. 5 graphical display

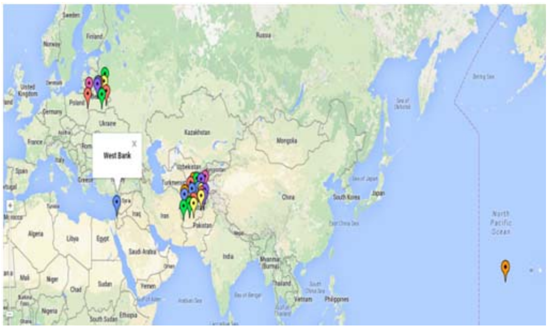


Fig.6 Map representation

VII. EXPERIMENTS. RESULTS

Considering with the airlines data as a input, experiments were performed on both java and hadoop mapreduce environment for analysis of the dataset of the airports from various countries. The size of data is not huge as the experiments were performed in low configuration computers. The results obtained are as below:

File Size (apprx.)	Time Taken	
	In Java(apprx)	In Hadoop Environment(apprx)
2 Mb	300 sec	100 sec
20 Mb	2700 sec	800 sec
360 Mb	38 minutes	10 minutes
800 Mb	1 hr. 25minutes	25 minutes

VIII. CONCLUSION

We have examined the design and architecture of Hadoop's MapReduce framework in great detail. Particularly, our analysis has focused on data processing. We would conclude by saying that bigdata is the new buzz word and Hadoop Mapreduce is the best tool available for processing data and its distributed, column-oriented database, HBase which uses HDFS for its underlying storage, and support provides more efficiency to the system.

REFERENCES

[1] Bharati Punjabi*, Prof. Sonal S. Honale "Survey On Big Data Analysis & Processing With Data Mining Methodology" IJESRT Mar 2015

[2] "Data Mining with Big Data" - Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE..IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, January 2014 97

[3] R. Ahmed and G. Karypis, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012.

[4] M.H. Alam, J.W. Ha, and S.K. Lee, "Novel Approaches to Crawling Important Pages Early," Knowledge and Information Systems, vol. 33, no. 3, pp 707-734, Dec. 2012.

[5] S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," Science, vol. 337, pp. 337-341, 2012.

[6] A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," ACM Crossroads, vol. 19, no. 1, pp. 20-23, 2012.

[7] S. Banerjee and N. Agarwal, "Analyzing Collective Behavior from Blogs Using Swarm Intelligence," Knowledge and Information Systems, vol. 33, no. 3, pp. 523-547, Dec. 2012.

[8] E. Birney, "The Making of ENCODE: Lessons for Big-Data Projects," Nature, vol. 489, pp. 49-51, 2012.

[10] J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," J. Computational Science, vol. 2, no. 1, pp. 1-8, 2011.

[11] S. Borgatti, A. Mehra, D. Brass, and G. Labianca, "Network Analysis in the Social Sciences," Science, vol. 323, pp. 892-895, 2009.

[12] J. Bughin, M. Chui, and J. Manyika, Clouds, Big Data, and Smart Assets: Ten Tech-Enabled Business Trends to Watch. McKinSey Quarterly, 2010.

[13] E.Y. Chang, H. Bai, and K. Zhu, "Parallel Algorithms for Mining Large-Scale Rich-Media Data," Proc. 17th ACM Int'l Conf. Multimedia, (MM '09,) pp. 917-918, 2009.

[14] R. Chen, K. Sivakumar, and H. Kargupta, "Collective Mining of Bayesian Networks from Distributed Heterogeneous Data," Knowledge and Information Systems, vol. 6, no. 2, pp. 164-187, 2004.