Design Engineering

ISSN: 0011-9342
Issue: 4 | Pages: 1738 - 1746

Big Data Preprocessing Frameworks: Tools and Techniques

[1]Vinaya Keskar, [2]Shokhjakhon Abdufattokhov, [3]Khongdet Phasinam, [4]Alex Wenda,
[5]Dr. Santosh T. Jagtap, [6]Randy Joy Magno Ventayen
[1]Research Scholar, Savitribai Phule Pune University, Pune & Assistant Professor, ATSS's College of Business Studies and Computer Applications, Pune.
vasanti.keskar@gmail.com
[2]Phd, Department of Computer Science, Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Uzbekistan
[3]Assistant Professor, Faculty of Food and Agricultural Technology, Pibulsongkram Rajabhat University, Phitsanulok, Thailand
phasinam@psru.ac.th
[4]Department of Electrical Engineering, Faculty of Science and Technology,State Islamic University of Sultan Syarif Kasim Riau, Indonesia
alexwenda@uin-suska.ac.id
[5]Assistant Professor, Department of Computer Science, Prof. Ramkrishna More College, Pradhikaran, Pune, Maharashtra, India
St.jagtap@gmail.com
[6]University Director, Department of Public Relations, Publication and Information Office, Pangasinan State University, Pangasinan, Philippines
https://orcid.org/0000-0002-0952-7795

Abstract

Big Data Analysis blended with computational algorithms is a novel tendency in feature abstraction. This involves acquiring knowledge from reliable data sources, rapidity in processing information, and future prediction. Big Data analytics is dynamically evolving with variant features of velocity (analysis time has drastically decreased subsequently), volume (corpus size raise from Big Data to Bigger Data) and Vectors (consonance to dissonance). Organizations now focus on analyzing data that are getting accumulated and are interested in deploying analytics to withstand forthcoming challenges.

Big data involves a massive volume of data that are so large, and it is difficult to process using traditional database and software techniques. In the use of big data applications, a technical barrier is encountered when moving the data across various locations, which is very expensive, and it requires large main memory for holding data for computing. Big data includes transaction and interaction of datasets based on the size and complexity that exceed the regular technical capability in capturing, organizing and processing data in cloud environment. To encounter this problem data preprocessing is necessary. Various data mining and machine learning techniques are used for preprocessing. This paper provides an in-depth study of data preprocessing for big data analytics. It also contains generalized framework and tools used.

Keywords: Big data, data processing, data preprocessing, big data frameworks

ISSN: 0011-9342
Issue: 4 | Pages: 1738 - 1746

1. Introduction

Big data are processed with high-performance clusters in real time, data intensive. Big data networking is used to spread data around different places. This method is very costly and requires a huge storage room to store the records. Big data includes scale and complexity-based transaction and interaction databases that transcend regularly technological capacity to collect, organize and process data at a fair cost in a cloud environment. Big data calculation and data exchange are efficiently carried out using cloud data preprocessing. Data collection has expanded exponentially in applications involving large data [1] [2].

The method for collecting and exchanging information with greater memory usage involves broad data applications. An analysis of huge quantities of information and the extraction of information or expertise for potential operations is the biggest challenge of major database applications. With the aid of preprocessing, the unacceptable noises collected from different data sources are eliminated, which minimizes the time required for calculation and enhances the information exchange. The distributed data extraction on a vast volume of cloud data requires minimal overhead computing and connectivity costs. Higher than the normal range of databases [3], large amounts are often represented as large data.

The Big Data principle provides information about volume, speed and variety. Volume is associated with the enormous amount of the data required for valuable information to be processed.

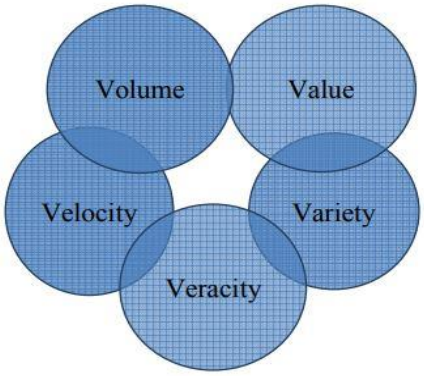

Figure 1: Characteristics of Big data

The velocity function analyzes the big data which is necessary to respond in a logical period of time. Similarly, variation refers to the various data type that composes the data quantity. Data classification is commonly used for the provision of expertise and information to consumers in an extensive amount of accurate and productive way. While large datasets exist, traditional methods of classification do not produce satisfactory results. The task is to assess and to consider the special features of big data sets by searching for useful geometrical and mathematical models in the classification of big data.

Big data has acquired considerable science importance due to the availability of comprehensive information and the advantages of data analysis. Large data applications are analyzed using MapReduce [4][5][6] with the flexible essence of data.

ISSN: 0011-9342
Issue: 4 | Pages: 1738 - 1746

2. Big Data Processing

For examining large amounts of data, the cluster architecture is used. Cluster nodes are divided into three groups of front end nodes, data processing and data loading nodes. In front end nodes, the parameterized queries are pre-processed and passed on to the data processing with analyzing of further processing nodes. The Scatter-Gather data processing style is then implemented. If an appraisal request is sent from one of the front end nodes, the request is broken down into sections and parts are assigned to information retrieval with computer analysis nodes.

In front-end nodes used by the database sender to select correct data processing with parallel query processing nodes, metadata about data delivery are stored. A popular database engine that stores the metadata processes all the front-end nodes. Hot standby technology is used to protect the database with parallel data analysis to maintain high reliability. Each data processing using the analyzing node manages a sub-set of data and is responsible for local indicator estimation, local time series testing and partial results for fusion in the frontends. A customer's keeping benefit is determined by data nodes and the partial results are fusioned with the front end nodes to achieve the final result.

Parallel processing of results, as shown in Figure, has been planned by authors in [7]. Data analyze processes manage a vast amount of historical information in this architecture and the data are rarely modified to use non-transactional database retrieval engines to monitor them in the analysis of data.

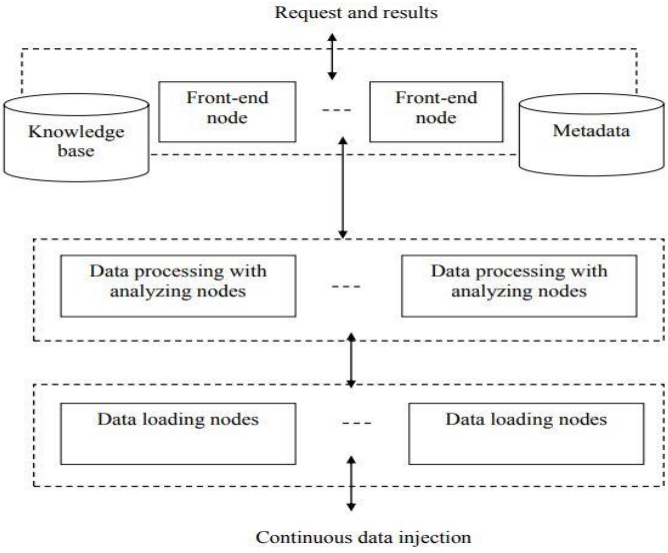

Figure 2: Data Processing Architecture

MySQL is the basic storage engine. Through following the basis of the code, transaction processing problems are overcome. In the whole transaction processing block, the database structures are analyzed and hence the storage engine has less weight and the data access rate has been improved. The nodes for data

ISSN: 0011-9342
Issue: 4 | Pages: 1738 - 1746

loading are responsible to divide, convert and load fresh information. Each data partition is simulated in the data propagation scheme to achieve strong fault tolerances at least three nodes. MapReduce, Google's massive data processing platform, is the inspiration of the parallel architecture of data processing.

The distinction is that it uses code base engines instead of a file structure for data storage. Since transaction abilities are disabled, the storage engine achieves higher hierarchical data processing speeds. When the number of nodes increases, the mapping metadata is manually updated, which means that the data distribution is executed appropriately. Allow a time signature-based dividing technique to divide data between data nodes. After partitioning, the data is assigned by applying the round tap technique to neighboring nodes. The technique of data partitioning ignores the case where transaction data from a client are collected on one node only.

The time-marked partitioning strategy then mobilizes the entire data processing in parallel to analyze nodes, which increases the efficiency of the parallel data processing architecture.

Big data comprises a vast range of self-contained data sets. Big data is being used more rapidly in all research and engineering fields of physical, biological and biomedical sciences as a result of rapid networking, data storage and data capability. HACE theorem by [7] classifies the characteristics of the big data revolution and designs a large-scale data processing model. It includes the collection of sources of content, mining and interpretation based on demand.

Big data and cloud computing are the main issues which enable better efficiency of computing resources using IT services. The needs for computer hardware, space and applications are reduced by cloud computing.

Big data is an integral [9] model which can be extended to databases, and the scale or sophistication of these datasets is outside the computer software and hardware resources usually used. Big data analysis problems with the cloud computing infrastructure, cloud computing functions, service models, implementation models and software architecture are discussed by enterprise data management and data processing procedures.

Big data is called by [10] to process and store very large quantities of data. The success of large data systems is affected by anomalies and failures in the cloud platform. Big data mostly focuses on the output of Big Data systems to assess the consistency variables. The output analyzes findings are crucial in determining the cause of the Big Date apps and cloud deterioration. Service level arrangements enhance large data applications across resource planning phases. For big-data applications, a performance measurement model is developed that blends the output of the software.

The [11] along with the I/O path are a plurality of potential data analytics placement plans. Flexible data Analytics (FlexAnalytics) is designed to determine the best plan to minimize data movement in a specific

ISSN: 0011-9342
Issue: 4 | Pages: 1738 - 1746

situation. The prototype framework FlexAnalytics is designed for positioning of analytics. It enhances the I/O stack scalability and usability of HEC data preprocessing, data interpretation and visualization tools.

Researchers at [12] also clarified that entropy is not exempted and that the distribution of information in big data is to be defined. For data translation within a single phase calculation, the multi-interval discretization process is complex. The most critical role in the data mining process is discretization. The key goal is to minimize the importance of the continuous data in massive datasets. This decreases the noise and incoherence of data in various data sources. This increases the complexity of the processing time and space during cloud data sharing

The distributed and concurrent data management and analysis of petabyte-scale datasets by [13] with Hadoop in the bio-informatics community introduces large data technologies such as Apache Hadoop project.

Authors in [14] provided a cloud storage solution that can consistently store a wide range of heterogeneous data. The hybrid architecture is designed to improve data collection, query, and retrieval for document and object-orientated plans.

A scientific workflow framework for cloud computing, Map-Reduce and service-orientated architecture has also been developed by authors in 15. HBase is deployed in distributed computers for the storage and control of large geoscience data. Parallel processing of geoscience data is maintained in the MapReduce based algorithm.

3. Framework and Techniques Used for Big Data Processing and analytics

Machine learning is used for data creation. The key goal is to build algorithms that enable a design based on observed data to be built on computers. Big data is used to generally managed and unmonitored learning algorithms. The most basic and is known as the supporting vector machine is used for clarification of the data (SVM).

But the SVM disadvantage including scalability, time and memory issues is now being solved parallel vector support machines. Artificial Neural Networks Networks ANN are mature methods used in adaptive regulation, image processing and the identification of patterns. Today, ANNs are used to manage control theory and statistical inference of artificial intelligence. Big data for studying communication, social psychology and economy are used in Social Network analysis

ISSN: 0011-9342
Issue: 4 | Pages: 1738 - 1746

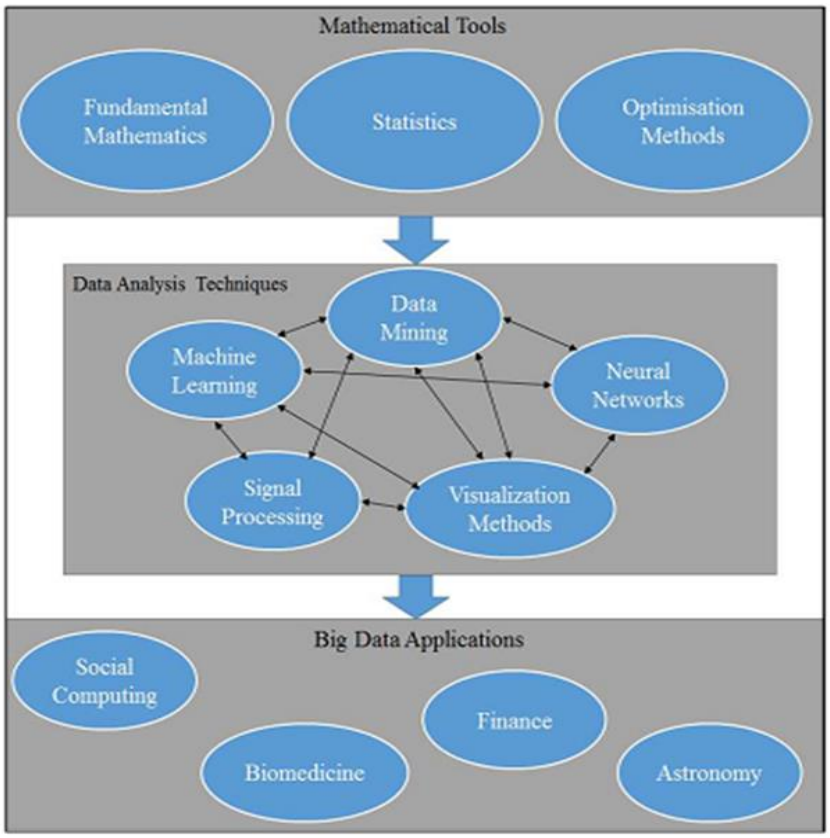

Figure 3: Big Data Techniques

We always have to be successful in managing and processing the giant to tackle big data in order to get better ideas and judgements. "Hadoop" comes first in mind as we talk about Big Data technologies.

For distributed storage and retrieval applications, Hadoop offers a free-source computing architecture in java on very big data sets. For questions writing and data processing pivers, Hadoop framework consists of higher level declarative languages. Hadoop manages and analyzes large numbers of unstructured logs and activities by about 63 per cent of organizations [16]. Hadoop consists of several modules, though two mainly HadoOp Distributed File System and MapReduce() components are used in Big Data. Hadoop has many components. The other elements include additional resources and higher abstraction levels.

The Hadoop system's principal component, MapReduce () is the one that is used on the distributed or parallel algorithm cluster to process and produce large datasets. This is a paradigm for programming used to handle huge quantities of data by splitting the job into separate nodes. A program MapReduce () is two jobs, A processMap(), that includes data collection, filtering and sorting, A procedure Reduce (), which involve summary finding and the final result. A program MapReduce () The machine MapReduce () manages all communications, parallel data sharing, replication and fault tolerance as well. [17]

HDFS is used to store massive data files, which are mostly to be stored in gigabyte to terabyte on a single computer. The HDFS file systems for Hadoop framework [18] is distributed, modular and compact, written

ISSN: 0011-9342
Issue: 4 | Pages: 1738 - 1746

on Java. It keeps the data durability of multiple hosts, making concurrent processing easier, dividing a file into blocks that are stored across multiple devices. The HDFS cluster has a master-slave link to a single name-node and many data-nodes.

Apache Spark has opened the source tool originally created at UC Berkley's AMP Lab. It offers analytics in memory that is better than Hadoop (up to 100 times). It has been developed for the use of iterative algorithms and immersive analysis. The Hadoop's Storage APIs are extremely compatible. The Hadoop Cluster Setup can be installed. Developers can use several Spark programming languages to write driver programs. The use of Spark is important because iterative machine learning algorithms will boost results per time [19].

When it comes to real time results, Hadoop's efficiency decreases marginally. Data stream processing is required during real time processing. For processing data in real-time, big data platform tools such as Strom [20], SQL stream [20] and Stream Cloud [21] are used.

The popular Cloud Computing paradigm is service-oriented computing which enables its shared remote users to access data storage and processing on demand, pay-as-you-use, safe storage management, simple and flexible growth, and computer resources [22].

Cloud Computing is fully virtual for consumers who need minimal effort to operate with functionalities such as on-demand, scalability, stability, maintenance, economic efficiency and accessibility. Service delivery to customers by leveraging the Internet and resource exchange can be carried out using a remote server network for data storage, administration and processing using a distributed data management system. His service-driven architecture embraces 'anything like a service,' provides his 'technology' with networks, network and applications as a service according to various models.

Big data and cloud computing systems are evolutionary and additive. The advantages of integrated use are: scalability, mobility and on-demand elastic data availability [23]. The Big Data environment requires a cloud cluster to manage the resources for processing high-speed and varied data volumes. Cloud infrastructure offers this form of service in a cost-effective manner, with server cluster, storage and networking services that can scale or down as required. A single server can support many users by using cloud computing to download and update the data without paying for various apps.

4. Conclusion

We are now in the Big Data age, and modern technologies and strategies are needed to process data efficiently. In modern developments a leapfrogging will occur and a new wave can be witnessed in the computer system. In this study some big data issues, the methods and techniques that are used for the processing of big data were addressed. While little technology is still in the making, today's innovations

ISSN: 0011-9342
Issue: 4 | Pages: 1738 - 1746

produce respectable performance. Big data often means large processes, big confrontations and large income, so further analysis in the sub-components is needed to overcome this.

We take care of the work and progress of Big Data and nobody can help without it. The development of Big Data is based on human resources, financial investments and new concepts. This paper includes a thorough analysis in Big Data Analytics data preprocessing. The system and methods used are also widespread.

5. References

[1] Research Note 6, pp. 70, February 2001.

[2] Hilbert, Martin, and Priscila López. "The world's technological capacity to store, communicate, and compute information.", Science 332.6025, pp. 60-65, April 2011.

[3] "Big data as the bedrock of the future economy", January 11 2017, European Political Strategy Centre, [Online] https://ec.europa.eu/ epsc/publications/strategic-notes/enter-data-economy_en, last accessed 05-04-2017.

[4] "What is Big Data", 2016 [Online] http://www.ibm.com/bigdata/us/en/,last accessed 05-04-17.

[5] "Big Data 3 V's: Volume, Variety, Velocity (Infographic)", 2013, [Online] http://whatsthebigdata.com/2013/07/25/big-data-3-vs-volumevariety-velocity-infographic/, last accessed 05-04-2017.

[6] "Big Data", 2016, [Online] https://www.sas.com/en_us/insights/bigdata/what-is-big-data.html, last accessed 05-04-2017.

[7] Xiongpai, Q, Huiju, W, Xiaoyong, D & Shan, W 2010, 'Parallel techniques for large data analysis in a futures trading evaluation service system', in 9th International Conference on Grid and Cooperative Computing (GCC), pp. 179-184.

[8] Wu, X, Zhu, X, Wu, G-Q & Ding, W 2014, 'Data mining with big data', IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107.

[9] Reddy, VA & Reddy, GR 2015, 'Study and Analysis of Big Data in Cloud Computing', International Journal of Advance Research in Computer Science and Management Studies, vol. 3, no. 6, pp. 416-422.

[10] Bautista Villalpando, L, April, A & Abran, A 2014, 'Performance analysis model for big data applications in cloud computing', Journal of Cloud Computing: Advances, Systems and Applications, vol. 3, no. 1, pp. 19-38.

[11] Zou, H, Yu, Y, Tang, W & Chen, H-wM 2014, 'FlexAnalytics : A Flexible Data Analytics Framework for Big Data Applications with I / O Performance Improvement', Big Data Research, vol. 1, pp. 4-13.

[12] Ramirez-Gallego, S, Garcia, S, Mourino-Talin, H, Martinez-Rego, D, Bolon-Canedo, V, Alonso-Betanzos, A, Benitez, JM & Herrera, F 2015, 'Distributed Entropy Minimization Discretizer for Big Data Analysis under Apache Spark', Proceedings - 14th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2015, vol. 2, pp. 33-40.

ISSN: 0011-9342
Issue: 4 | Pages: 1738 - 1746

[13] O'Driscoll, A, Daugelaite, J & Sleator, RD 2013, ''Big data', Hadoop and cloud computing in genomics', Journal of biomedical informatics, vol. 46, no. 5, pp. 774-781.

[14] O'Driscoll, A, Daugelaite, J & Sleator, RD 2013, ''Big data', Hadoop and cloud computing in genomics', Journal of biomedical informatics, vol. 46, no. 5, pp. 774-781.

[15] Li, Z, Yang, C, Jin, B, Yu, M, Liu, K, Sun, M & Zhan, M 2015, 'Enabling big geoscience data analytics with a cloud-based, mapreduce-enabled and service-oriented workflow framework', PLoS ONE, vol. 10, no. 3, pp. e0116781-e0116781.

[16] "Hadoop Eco-system: Hadoop tools for crunching Big Data", October 2016, [Online] https://www.edureka.co/blog/hadoop-ecosystem, last accessed 05-04-2017.

[17] "Perform the computation using MapReduce jobs", April 2016, [Online] https://examples.javacodegeeks.com/enterprise-java/apachehadoop/how-does-hadoop-work/, last accessed 05-04-2017.

[18] "Big Data in the cloud: Converging Technologies", April 2015, Intel IT Center, [Online] https://www.intel.com/content/dam/www/public/ us/en/documents/product-briefs/big-data-cloud-technologies-brief.pdf, last accessed 05-04-2017.

[19] V Srinivas Jonnalagadda, P Srikanth, Krishnamachari Thumati, Sri Hari Nallamala, "A Review Study of Apache Spark in Big Data Processing", International Journal of Computer Science Trends and Technology (IJCST) – vol. 4, issue 3, May - Jun 2016.

[20] "Pentaho Business Analytics", Pentaho, 2017, [Online] http://www. pentaho.com/product/business-visualization-analytics, last accessed 05-04-2017.

[21] Diana Samuels, "Skytree: Machine Learning Meets Big Data", February 23, 2012, [Online] https://www.bizjournals.com/sanjose/blog /2012/02/skytree-machine-learning-meets-big-data.html, last accessed 05-04-2017.

[22] Bharathi KK, "Converging Technologies of Cloud and Big Data", IJCSET (www.ijcset.net), vol 6, issue 1, pp. 75-77, January 2016.

[23] Mathur R, "Integrating Big Data in Cloud Environment: A Review", International Journal of Innovations in Engineering and Technology (IJIET), volume 7, issue 1, pp. 513-517, June 2016.

[24] S. Abdufattokhov, K. Ibragimova, D. Gulyamova, K. Tulaganov, "Gaussian Processes Regression based Energy System Identification of Manufacturing Process for Model Predictive Control," *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 9, pp. 4927-4932, 2020, doi=10.1109/ICISCT47635.2019.9012025

[25] S. Abdufattokhov and B. Muhiddinov, "Probabilistic Approach for System Identification using Machine Learning," *2019 International Conference on Information Science and Communications Technologies (ICISCT)*, 2019, pp. 1-4, doi: 10.1109/ICISCT47635.2019.9012025

[26] S. Abdufattokhov and B. Muhiddinov, "Stochastic Approach for System Identification using Machine Learning," *2019 Dynamics of Systems, Mechanisms and Machines (Dynamics)*, 2019, pp. 1-4, doi: 10.1109/Dynamics47113.2019.8944452.