International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

Big Data Analytics for Deriving Business Intelligence Rules

Chandrashekar D K, Srikantaiah K C, Venugopal K R,

Abstract*: Big data is a large volume of data pool and processing and analyzing these data is tedious jobs. The aim of fulfilling huge information storage needs is that the structural transformation of repository system using traditional approaches to NoSQL technology. However, the existing technologies for storage are inefficient since, they do not generated data that are scalable, consistent and solutions for rapidly evolving diversified data. The primary method for storing huge amounts of data is used for analytics in real time applications like healthcare, scientific experiments, e-business and networks. In this paper, it is in sighted the characteristics, application, tools of big data, Technologies, Big data analytics, challenges and issues in Big data.

Keywords: Big Data Analytics, Hadoop, Map Reduce Structured Data, Semi Structured and Unstructured Data

I. INTRODUCTION

Big data is a huge volume of data of different types of data from different sources for high speed processing for producing good quality and accuracy of data. It is characterized based on 4V's model refers to the Variety, Volume, Velocity and Veracity. Volume is referred with the amount of big data, variety with the size and velocity with the speed, Veracity refers the quality of the data. Nowadays, data are increasing rapidly from different sources like apps from mobile devices, satellite pictures, Health care, education etc. The collection of data from all these fields becomes a large volume of data, but processing these data is tedious job and obtaining accurate inform for processing data is a meticulous job.

An unstructured and semi-structured data types will not support the traditional data storage which has been designed based on the relational databases to data set which is in structured form. When data warehouse is not able to handle the demands assigned by sets of big data which comes continuously and updated frequently. In real time application the activities of online users and trading stock exchange or apps of mobile users.

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author
Chandrashekar D K*, Assistant Professor in the Department of Computer Science and Engineering at S J B Institute of Technology, Bangalore, India.
Srikantaiah K C, Professor in the Department of Computer Science and Engineering at S J B Institute of Technology, Bangalore, India.
Venugopal K R, Vice-Chancellor of Bangalore, University.

Here, big data analytics concept introduces, analytics is used for complex problems for processing and examining large and various data sets and to explore uncover data such as patterns hidden in mining, correlation of unknown, trends in market and preferences of customer *etc.*, this information helps the organization to establish their decision for business. When high computing power systems are used to designed for specialized system of analytics. Big data analytics offers many benefits for business are new revenue opportunities, more effective marketing, better customer service, improved operational efficiency, Competitive advantages over rivals by increasing accuracy it is easy to make decision confidently and that will result in efficiency and reduces the cost and risk. The analytics methods are used to generate actionable business intelligence rules.

II. CHARACTERISTICS OF BIG DATA

Data means which is always used and stored forever, consider any application that stored the data in a decade's means which can be stored forever and when the data required it can be retrieved from the application. There are so many problems in data due to the increasing size of data exponentially and data organization. Big data can be characterized by the following terms:

Volume- The amount of data generated and stored. The value and potential understanding of the data decides whether it can be treated as big data or not.

Variety- The nature and category of the data. It assists people who evaluate it completely resulting in insight of data. The important sources include images, text, audio, video and also the missing data can be found by binding of data.

Velocity- It refers to the speed of generating and processing of data in order to attain the path of development and also the growth of data. For real time applications they are used.

Veracity- It refers to the quality of the data. It denotes the accuracy and also truthfulness of the data.

Data should be handled with modern tools (algorithms an analytics) to bring out the weighted information. The factory people must check the visible and invisible matters with different elements to operate it. Algorithms that generate the information should recognize and mark the invisible elements like degradation of machine, component wear in the factory floor.

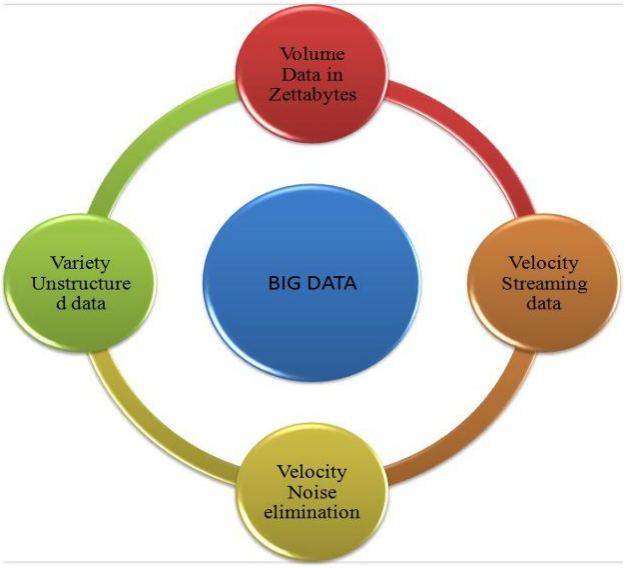

Fig.1 Characteristics of Big Data.

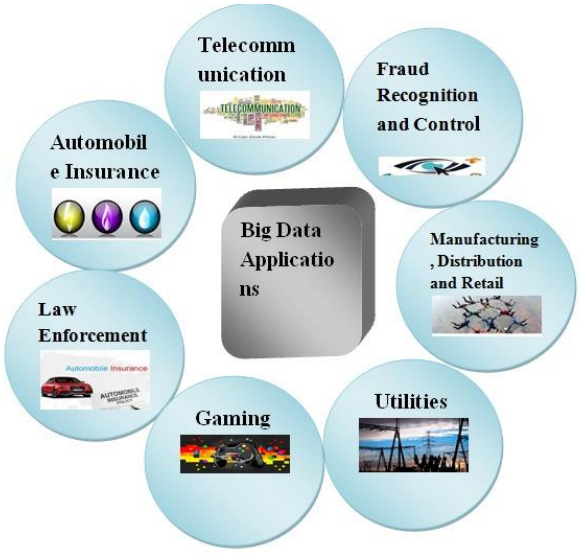

Fig.2 Applications of Big Data

III. APPLICYATIONS OF BIG DATA

Automobile insurance

Automobile industry is growing rapidly in this era and to cover insurance to where each vehicle we require a lot of data like Registration number, engine number, chassis number and details of vehicle owner. The data is so huge so that big data is used to characterize the data and store it. It also helps in keeping a record of which vehicle is owned by whom. Comparing different models with each other company can provide different kind of insurance to customer. Big data is used for known the user requirements

Telecommunications

In telecom industry the network provider always want to provide a new tariff plans or product to its consumer. So, it becomes a difficult task to provide new traffic plan along with existing plan without overlapping in the network. In such case big data can be used to overcome this issue. Implementing the latest network and increasing the bandwidth capacity.

Fraud Recognition and Control

It is a big challenge to detect the frauds in various sectors, more chances of fraud happen in banking sectors and financial sector. Big data can be used to learn the pattern of fraud transaction and control the fraud. which stops the data leak in the form of credit and debit card.

Manufacturing, distribution, and retail:

An online shopping is a trend now a day, where the information of each site is stored in big data. The help of big data we can organize the huge volume of data in structured way so that it is easy to access the data by these sites. Retail shops or market also uses huge amount which can be handled with the help of big data in systematic way.

Utilities

Big data is a supreme component to solve main business issues of utility companies. Utilities should leverage big data analytics for assisting, convert information to actionable insights authorizing high operational choices.

Transportation and logistics

In transport we require large amount of data from source to destination, to organize and classify the data in proper way we can use big data which will categorize and classify the data, thus will help in analysing the data and store the data. In logistics also we use huge amount of data like address.

.Gaming

In Today's world, Gaming is good entertainment and relaxation to mind. Which produces huge volume of data and velocity, the big data is designed to handle this velocity and helps to store the data.

Law Enforcement

Law can be maintained by using the help of big data. The law enforcement institute can have a list of criminal activity and criminal record so that whenever and criminal activity happens they can investigate their data base and check with the past records. it can also be used in watching over past criminal record and their present activity. We can use the big data in recognition of images of people violating their privacy.

Education

Big data in the field of education helps in improving the results rate and dropout rates and course offered and evaluation system. The organization using bigdata helps in Prediction and analyzing the students results and outcomes by using all the data given by analytics

Health services

Health service stores the information about each patient of their records and the videos of surgery and changes the complete health sector history. Big data plays a important role in the field of health services.

IV. TECHNOLOGIES FOR BIG DATA

There are so many emerging technologies which are used to accumulate access and interpret for large volumes of data in real time. Here we discuss a Hadoop open source technology the remaining technologies are summarized in Table I.

A. HDFS (Hadoop File System)

HDFS is developed using distributed file system and runs on an object hardware which gives high accuracy and it is cost-effective. It holds a very large amount of data which is stored in multiple machines by which it is easy to have a backup in any kind of system failure. The built-in servers of name node and data node help to easily check the status of collection, streaming access and provide file permission and authentication.

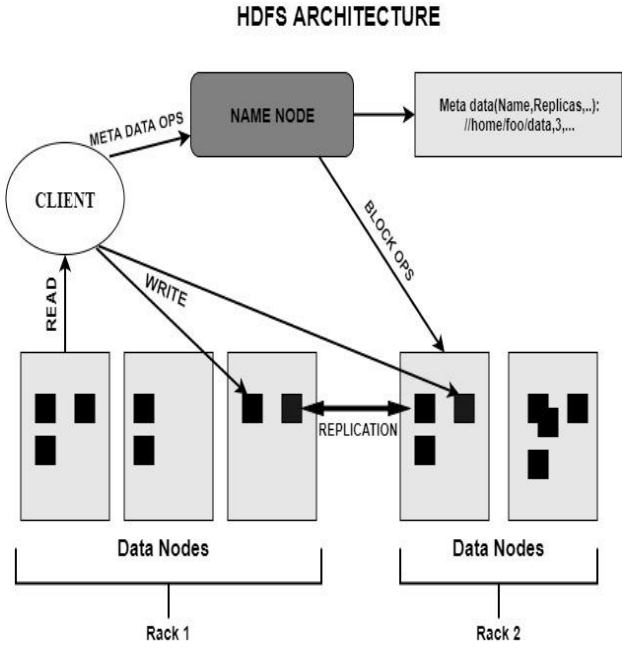

Fig.3 HDFS Architecture

Metadata: - A set of data that describes and gives information about other data.

Name node:- In Apache Hadoop HDFS Architecture name node works as a master node , which manages and controls all the existing blocks in the data node, name node is a available server which controls the clients to file accessing in HDFS

Data node:- Data node is a slave node in HDFS, the data node is not a high quality and more expensive and it is not a block server which stores the data in local system.

Block: - HDFS divides files into different blocks called data blocks which hold small data in the file system.

B. Map Reduce

The Map Reduce consists of two functions, map () and reduce (). Mapper performs the tasks of filtering and sorting and reducer performs the tasks of summarizing the result. Figure 4 shows for the distributed processing on a Computer cluster Map Reduce software framework used foe large datasets. As shown in Figure 4. Map Reduce consists of two part map and reduce phase, so it is called Map Reduce.

Input and output are in the form of key in each part of Map Reduce. Before sending to mass-produce the input value will be divided into key pair value. Every time Map Reduce produces latest key pair value. The reduce method is processed for every different key value. For each distinct key the reduce part produces one key value pair. A key value pair is the final output. Map Reduce will process in an input file manner.

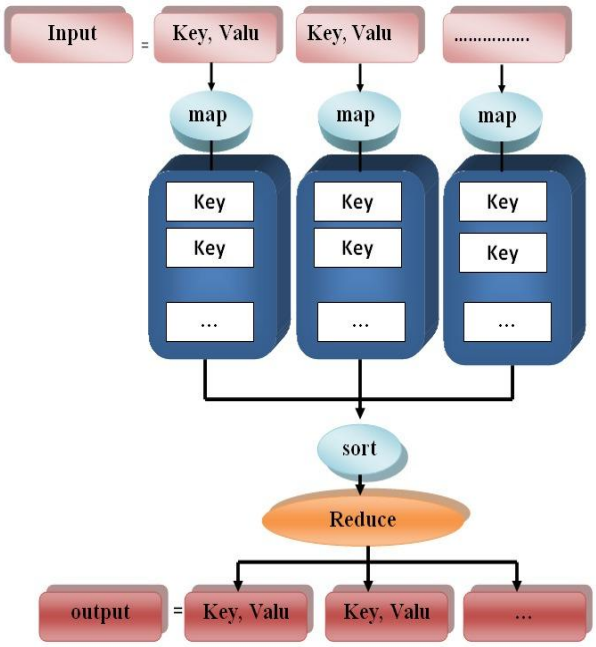

Fig.4 Map Reduce data flow with a single reduce task

To match the aggregations there are different reducers. Users will implement their own process logic by particularizing a conventional map () and reduce () function. The map will take the input as key value pairs and performs the mapping function to generate intermediate key value pairs. The Map is used for reducing intermediate key value pairs and this output is given to the reducer to generate the final output. Reducer is widely used for analysis of big information.

Table-I: Technologies for Big Data

SL	Data analysis	Description
1	Hadoop [34]	Hadoop is an open source framework which controls storing of Big Data in HDFS and processes of big data using Map Reduce with a parallel Distributed algorithm running on clustered systems.
2	Hive[35]	Apache Hive is a framework for querying of Big Data in Hadoop Platform.
3	Pig [36]	Apache pig is a good platform to run programs on Hadoop platform..
4	Wibidata [37]	It is a software system company that developed huge information applications for enterprises to change their consumer experiences.

5	Platfora [38]	Platfora works with Hadoop to assist in data analysis and image sharing.
6	Rapid Miner [39]	Rapid Miner is a software platform which provides an integrated environment for data preparation for deep learning, machine learning, text mining, and predictive analytics.
7	Scoop [40]	Scoop is used to transfer the bulk of data effectively between Hadoop and databases.
8	Zookeeper [41]	Zookeeper is used to provide a distributed configuration service for large distributed systems
9	Mahout [42]	Mahout is a distributed linear algebra framework to data scientist for implementing their own algorithm.
10	Hbase [43]	Hbase is a column-oriented data management system that runs on Hadoop Distributed organization. Which stores spares data sets commonly used in many use cases.
11	Sky tree [44]	Sky tree is a software platform for developing and testing advanced analytics solutions for big data.
12	NoSQL [45]	NoSQL Database is used to store unstructured data and MongoDB, Redis are based on NoSQL.
13	Spark [46]	Spark is an engine for processing big data within Hadoop and s upto one hundred times faster than the standard Hadoop engine.
14	R [47]	R is a programming language as well as a software environment used for working with statistics for big data analytics.
15	Data Lakes [48]	Data Lakes are generally used to access huge data which stores the data as a result they are known as data lakes. It is bit different from the data warehouse.
16	Scala [49]	scala is a general-purpose programming language which provides built in API and libraries for Big data analytics tools which used to store and access huge amount of data. .

V. BIG DATA ANALYSIS:

The Big Data Analysis having the following components as shown in Figure 5. Many of usably focus simply on the analysis section.

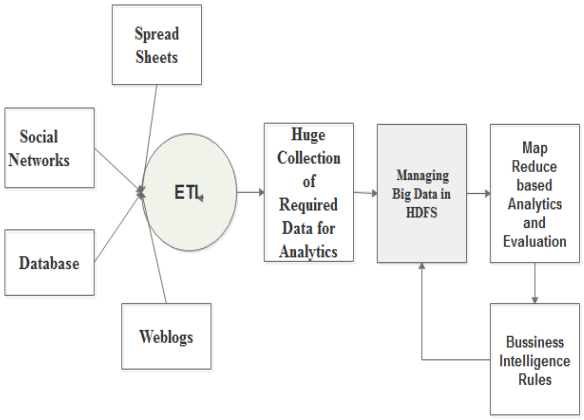

Fig.5 Structure of Big Data Analysis

Input Source:

Social Network is the platform to share the ideas and interest with friends and to make new friends for sharing their ideas. This holds n number of people information in HDFS Server.

Database: Database is a collection of all different variety of data from different sources for target process.

Weblog is a place where we can obtain the information on particular topic. A weblog gives a link to the other website fir relevant information**.**

Spread Sheets is a structured file consists of rows and columns which help to sort the data and easily calculate all numerical values using much mathematical formula.

ETL-Extraction, Transform, Load

Extraction is the process of collecting structured and unstructured data from different sources.

Transformation is the process of converting the extracted data into required data, so that it can be placed in another database. Transformation is occurred by using rules or lookup tables.

Loading is the process of loading the data into the target database.

Huge Collection of Required Data for Analytics

Collecting multiple data from different sources that should be converted into required form for big data analytic. Collection of data is placed in HDFS for managing big data.

Managing Big Data in HDFS

Collection of different types of data from different sources like spread sheets, Social Networks, Database and Weblogs, that information should be converted into structured form using ETL tools and it is stored in HDFS for further operations.

Map Reduce based Analytics and Evaluation

Map Reduce performs two functions Map and Reduce, map function splits the data into different chunks. The reducer combines the splited data and put it in the order and it is analyzed the output of map reduced.

Business Intelligence Rules

The generated Business Intelligence Rules by evaluating the map reduce operation and rules are stored in HDFS**.**

VI. CHALLENGES IN BIG DATA ANALYSIS:

A. Storage:

The storing and analysis of huge data is a challenging for bi g data, because the huge volume of data is produced from all the sources. To perform cleaning or preprocessing of data for smaller size is easy but for large volume of data is a tedious task, so it requires a good techniques and good management software tools to handle large data.

B. Data representation:

The data produced over internet will be in unstructured form, so to convert unstructured data into structured data is a skill full job and all the user will not understand the representation of computer language, so it has to be represented in an visualization form so that all user will understand and should help the users in the easier manner.

C. Heterogeneity and Incompleteness

We must provide homogeneous data rather than providing heterogeneous data, algorithm s expects homogeneous data and data must be structured in proper manner which help algorithm to analyze the data. It does not matter how much we clean the data and provide error correction some incompleteness and some error will also exists. This error must be cleaned by data analysis

D. Scale

As the name suggests "big data" main challenge is to handle the large volume of data in big data and management of huge data. We must provide correct tools to handle the large volumes of data, and data scaling must be done in efficient way which improves the performance and fastness or speed.

E. Timeliness

The main challenge in big data is the time, how much time is required for examining the huge data. The time required to analyze the data depends on the size of the datasets. We must provide the size of dataset in such a way that time required to analyze the dataset should be less. We must try to provide the result in small time rather than taking more time. Datasets must be structured in such a way that it takes consume less time to examine the data.

F. Privacy

The main challenge in big data is privacy of data. It is a topic of great concern we must provide the security to data so that it must be accessed by authorized user. There is various factors related to data security, We must provide data sharing to authorized user rather than to all. The more privacy enables the more security to the data.

G. Reducing highly associated columns

To reduce or eliminate the highly interrelated columns, which may create confusion during interpretation of big data. so We must eliminate the such columns from datasets which are not related to each other, which reduces the redundancy and analyzes is become simple. Reducing the associated columns will decrease the size of datasets and complexity will also decrease the time taken in analyzing.

D.**Human Collaboration**

Human collaboration, many experts are required to analyze the data, these men will be in different areas and they will provide different inputs to the program. We must design a program in such a way that it should be able to take inputs from different experts and program must be able to support their collaboration.

E. **Dimensionality Reduction via Tree Ensembles:**

To reduce the dimensionalities which reduce the complexity in interpret of data by generating the tree structure of data rather than creating columns which will help in analyzing the data in simple manner. Large amount of data sometime produce worst performance by reducing the dimensionality. And it reduces the size of data and will increase the performance.

VII. ISSUES OF BIG DATA

A. Transport Issues:

Nowadays all the transportation face a lot of trouble to design a reliable, efficient, sustainable and safe transportation system. Increases in population which lead more complex, the transformation behavior and users preferences are major drawback of big data.

B. Management Issues:

Generally the most common issues are faced is managing the big data and the tools which are required to process such huge data with accuracy and efficiency is not good. The complexities which are related with the data associated, less number of peoples which have the knowledge of big data, failure of adding big data into analytics so as to have a better execution environment and digital marketing strategies.

C. Processing Issues:

Due to rapid growth of emerging application there are more generation of data which needs to be processed efficiently and quickly. Processing these data is also quite critical. The main processing problem is the cloud data management and big data is processing mechanism in big data models.

D. Scalability Issues:

When it comes to big data storing, managing and processing is not an easy job. Scalability is essential because it contributes to competitiveness, efficiency, reputation and quality.

E. Query Optimization Issues:

It is a part of the query process in which each database system have different systems and we have to decide the least expected answers to provide solution, but the problem is fetching all the queries and finding the best which will be time consuming.

F. Transformation Issues:

While the data which are present in the real world consist of various columns which are not useful and the data which are not use full must be removed. So that it can be processed easily. The main issue with this is the time taken to get the best columns of data for a better performance.

G. Data Privacy Issues:

As the data gets bigger the chances of getting affected by some kind of virus or hacking gets more and more serious. Protecting the data requires a lot of money and hence increases the scenario.

H. Data Provenance Issues:

Truthful records when analyzing the statistics using huge data Tools. Integrity and authenticity are the two parameters that we have to think about when we prefer to analyze the data.

VIII. VARIOUS STUDY ON BIG DATA ANALYTICS

TABLE II Strength and drawbacks of Big Data Analytics

Authors/Year of publication	Domain tested	Technique/Algorithm employed	Advantages	Disadvantages/Remarks
A.Alexandrov *et al., [1]*	Big data Analytics is one of the stratosphere platform	Task scheduling algorithm using Thermal-aware techniques	The total energy consumption is to be effectively reduced and it improves the energy efficiency of thee data center	Lightning is one of the additional energy consumption so this model perform task migration, energy consumption
Rami Sellami *et al.,[2]*	Evaluation over relational and no sql data store and cloud environment are the complex queries optimization	the optimal execution plan algorithm	Heterogeneity is covered the unified data model and between relational and nosql data stores	When evaluated a native access cost is to be comparably overhead
Wu.DongyaoWu *et al.,[3]*	HDM: Big data processing is a accomposable framework	Data-flow optimization is employed in word count and physical plan and logical plan algorithm are implemented	By reinforce the composition registering and loading user defined function Apache pig[8] gives few reusability	Experiment of this section are not supposed to show essentially better framework
Jun Wanget *et al.,[4]*	To storage, distribution of sub data sets are unvellingly to speed up the bigdata analytics	For implementing the balanced computing over a sub datasets is one of the distributed aware algorithm	Using elastic map, to achieve balanced and efficient computing its to use dot net enables sub-dataset	Overall performance is to be degraded because of longer execution
Y.Zhu *et al.,[5]*	In distributed file system provides a scalable metadata lookup services in datacenter	Based on metadata ids utilizing a SDN technique to form network packets using metaflow implementation	Ties used in real world application meta data flow is used to increase the system throughput .	For system performance and scalability it's very hard to limit the size
Zhenhua Chen *et al.,[6]*	GPU-accelerated high throughput process online-stream-data	To implement storm,GPU programming and JCUDA	Programming model its easy to preserve the G-storm nicely integrate GPU's	Its take a time for each execution at least 10 min
Zhikui Chen *et al.,[7]*	Experimental feature selection for efficient economic in big data analysis	It is used as selection feature method and also implement parallelization of distance matrix calculation algorithms	To construct economic model its to give the opportunity for high dimension and volume.	Having only few economic factors for model construction to provide traditional econometric methods.
Hagras T.Janecek J *et al.,[8]*	Dependent task scheduling algorithm in distributed system	DAG-based scheduling algorithms	Proposed system that suggest a information to a user depends upon the explanation	User only to provide recommendation and it is only apart of the algorithm
S.Hoch Reiter J et.al.,[9]	Dynamic scheduling of manufacturing process is the time series forecasting	Decision making algorithms	Sequencing and scheduling is the operation of recursive neural network(RNN)	Recursive neural networks can optimize resource utilization and energy consumption

Huazhongliu *et al.,[10]*	DVFS and thermal-aware is enabled big data scheduling	Thermal aware and DVFS enabled technique is the algorithm based task scheduling	It is constructively reducing the total energy consumption and energy efficiency of the data center also implemented	It requires additional consumption such as lightning and also model takes the task migration, energy consumption
Xuhong Zhang *et al.,[11]*	To storage, distribution of sub-dataset is unvellingly to speed up bigdata analytics	Sub-datasets are implemented by using distribution aware algorithm for balanced computing	To achieve balanced efficient computing it using dot net enables sub-datsets analysis using elastic map	To degrading overall performance it needed longer execution time

IX. PERFORMANCE EVAUTION

The Following Datasets have been used to do Some Performance Evaluation in Big Data such as Naïve Bayes Algorithm, Linear Regression, Generalized additive model, Logistic Regression and Multiclass Logistic Regression

Sl	Datasets	Description
1	GDP	Current US Dollars represents the regional, country and world GDP. Regional includes collection of nations. Ex. Canada & Europe. The data is received from World Bank https://datahub.io/core/gdp
2	World cities	Collection of main metropolitan cities in the world crosses above 15, 000 in habitants, and each metropolis is connected with its own country and with the sub country to limit different verities the name of the state indicates sub country. https://github.com/datasets/world-cities
3	Country codes	Detailed country code list which includes codes like ISO3166 and dialing code like ITU and Currency code like 4217. .https://github.com/datasets/country-codes
4	Cash Surplus deficit	Data Repository is a large database infrastructure gives cash Surplus or Deficit in GDP from 1990 to 2013. Original data received from World Bank. https://datahub.io/core/cash-surplus-deficit
5	Geospatial	This is a listing of GIS records sources (including some geoportals) that furnish information sets that can be used in geographic data structures (GIS) and sptial databases for functions of geospatial evaluation and cartographic mapping. .https://catalog.data.gov/dataset?metadata_type=geospatial
6	Wikipedia Edits	User name, article name and date further changes edited by log of 1000 wikipedia. https://snap.stanford.edu/data/wiki-meta.html
7	Google public data directory	The Google public facts Explorer makes giant datasets effortless to explore visualize and communicate. As the maps and charts changes over time, https://www.google.com/publicdata/directory
8	Google datasets	Big query tool is a data set provided by Google which includes names of baby and also Git hub public data infrastructure ,New Hackers provides various stories and comments etc. https://ai.google/tools/datasets/
9	Quandl	Financial, economic substitute data can be obtained from several souces through their website/API or with tools which are directly integrated it has been divided into a premium. https://www.quandl.com/
10	Chars74k	Chars are further led of enormous amount of data available , in terms of written digits this data infrastructure includes recognition of characters in normal state of 74,000 images. http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/
11	Dexter	Dexter is a textual content classification hassle in a big of phrase representation. https://archive.ics.uci.edu/ml/datasets/dexter
12	Nomao	Nomao collects **information** about places (name,phone,localization)from many sources. deduplication consists of identifying what data refer to the same region .https://archive.ics.uci.edu/ml/datasets/Nomao

13	Euribor	Percentage of interest which is in fundamental between banks in the European Union interbank market and also used as a source of information for applying rate of interest on other loans. https://github.com/datasets/euribor
14	OpenAQ	Real time quality of air data infrastructure obtained from all around the world, can be received by open –source project. https://openaq.org/#/?_k=0kws2i
15	ImageNet	Image of big data infrastructure that is came to standardized accordingly to hierarchy of WorldNet. Approximately 100,000 phrases and ImageNet are included in World Net. It has been provided with 1000 images all around an average to explain each phase http://www.image-net.org/

X. CONCLUSION

In this paper, it is in sighted the characteristics, application, tools of big data, Technologies, Bigdata analytics, challenges and issues in Bigdata. The huge facts and it's a variety of ideas consists of huge records analytics, large data analytics techniques, statistics visualization and big records analysis algorithm are studied. The survey also offers an outline of the viable prospects of huge facts search environment In, recent years records area unit generated The parameter estimation is considered because the loss characteristic minimization over the coaching info set on supervised learning. The drawback of the linear regression mannequin for records analysis is that the outcome is no longer linear for all the income, the data isn't adequate decides the coefficient of the model.

REFERENCES

1. A. Alexandrov"The Stratosphere platform for big data analytics,"VLDB J., vol. 23, no. 6, pp. 939–964, 2014.
2. R. Sellami, S. Bhiri, and B. Defude"Supporting multi data stores applications in cloud environments," IEEE Trans. Services Comput., vol. 9, no. 1, pp. 59–71, Jan./Feb. 2016.
3. D. Wu, S. Sakr, L. Zhu, and Q. Lu "Composable and efficient functional big data processing framework," in Proc. IEEE Int. Conf. Big Data, 2015, pp. 279–286.
4. S. Ji, W. Li, S. Yang, P. Mittal, and R. Beyah "On the relative de-anonymizability of graph data: Quantification and evaluation," in IEEE INFOCOM 2016—The 35th Ann. IEEE Int. Conf. Comput. Commun., 10-14 Apr. 2016, doi: 10.1109/ INFOCOM.2016.7524585.
5. Y. Zhu, H. Jiang, J. Wang, and F. Xian "HBA: Distributed metadata management for large cluster-based storage systems," IEEE Trans. Parallel Distrib. Syst., vol. 19, no. 6, pp. 750–763, Jun. 2008.
6. L. Chen "Optimizing Map Reduce for GPUs with effective shared memory usage," in Proc. 21st Int. Symp. High-Performance Parallel Distrib.Comput., 2012, pp. 199–210.
7. L. Zhao, Z. Chen, Z. Yang, and Y. Hu, "A hybrid method for incomplete data imputation," in Proc. 17th IEEE Int. Conf. High Performance Comput. Commun., 2015, pp. 1725–1730.
8. Hagras T, Janecek J. A simple scheduling heuristic for heterogeneous computing environments[C]//Proceedings of the Second international conference on Parallel and distributed computing. IEEE Computer Society, 2003: 104-110.
9. S. Hoch Reiter, J."Long short-term memory.Neural Computation, 9(8), p. 1735-1780, 1997.
10. A. Sheth"Transforming big data into smart data: Deriving value via harnessing volume, variety, and velocity using semantic techniques and technologies," in Proc. 30th IEEE Int. Conf. Data Eng.,2014, Art. no. 2.
11. Xuhong Zhang World economic forum, "Big data, big ImpactNew possibilities for international development," 2012.[Online]. Available: http:// www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf
12. Liang Zhang"Big data across the Federal government", 2014.[Online]. Available:http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final_1.pdf
13. H. Giersch "Urban Agglomeration and Economic Growth". Berlin, Germany: Springer, 2012.
14. R. B. EkelundJr and R. F. Hbert "A History of Economic Theory and Method". Long Grove, IL, USA: Waveland Press, 2013.
15. B. Liddle"The energy, economic growth, urbanization nexus across development: Evidence from heterogeneous panel estimates robust to cross-sectional dependence," Energy J., vol. 34, no. 2, pp. 223–244, 2013.
16. S. Ghosh and K. Kanjila"Long-term equilibrium relationship between urbanization, energy consumption and economic activity: Empirical evidence from India," Energy, vol. 66, no. 3, pp. 24–331, 2014.
17. S. H. Law and N. Singh "Does too much finance harm economic growth?" J. Banking Finance, vol. 41, no. 4, pp. 36–44, 2014.
18. D. Baglan and E. Yoldas "Non-linearity in the inflation-growth relationship in developing economies: Evidence from a semiparametric panel model," Econ. Lett., vol. 125, no. 1, pp. 93–96, 2014.
19. Q. Ashraf and O. Galor"The 'Out of Africa' hypothesis, human genetic diversity, and comparative economic development," Amer. Econ. Review, vol. 103, no. 1, pp. 1–46, 2013.
20. V. Bol_on-Canedo, N. S_anchez-Marono, and A. Alonso-Betanzos "A review of feature selection methods on synthetic data," Knowl.Inf. Syst., vol. 34, no. 3, pp. 483–519, 2013.
21. S. Alelyani, J. Tang, and H. Liu "Feature selection for clustering: A review," Data Clustering: Algorithms Appl., Florida, USA: CRC Press, 2013.
22. M. A. Hall "Correlation-based feature selection for machine learning," Ph.D. dissertation, Dept. Comput. Sci., Univ. Waikato, Hamilton, New Zealand, 1999.
23. M. Dash and H. Liu"Consistency-based search in feature selection," Artificial Intell., vol. 151, no. 1, pp. 155–176, 2003.
24. M. A. Hall and L. A. Smith "Practical feature subset selection for machine learning," in Proc. 21st Australian Comput. Sci. Conf., 1998, pp. 181–191.
25. L. Beretta and A. Santaniello "Implementing ReliefF filters to extract meaningful features from genetic lifetime datasets," J. Biomed.Informat., vol. 44, no. 2, pp. 361–369, 2011.
26. Q. Gu, Z. Li, and J. Han "Generalized Fisher score for feature selection," in Proc. 27th Conf. Annu. Conf. Uncertainty Artif.Intell., 2011, pp. 266–273.
27. H. Peng, F. Long, and C. Ding "Feature selection based on mutual information criteria of max-dependency, max-relevance, and minredundancy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
28. I. H. Witten and E. FrankData Mining: Practical Machine Learning Tools and Techniques. SanMateo,CA, USA:Morgan Kaufmann, 2005.
29. J. G. Dy and C. E. Brodley"Feature subset selection and order identification for unsupervised learning," in Proc. Int. Conf. Mach. Learn., 2000, pp. 247–254.
30. P. S. Bradley and O. L. Mangasarian "Feature selection via concave minimization and support vector machines," in Proc. Int. Conf. Mach. Learn., 1988, pp. 82–90.
31. A. Rakotomamonjy "Variable selection using SVM based criteria," J. Mach. Learn. Res., vol. 3, no. 3, pp. 1357–1370, 2003.
32. M. Mejła-Lavalle, E. Sucar, and G. Arroyo"Feature selection with a perceptron neural net," in Proc. Int. Workshop Feature Selection Data Mining, 2006, pp. 131–135.
33. G. C. Cawley, N. L. C. Talbot, and M. Girolami "Sparse multinomial logistic regression via Bayesian L1 regularisation," in Proc. Advances Neural Inf. Process. Syst., 2007, pp. 209–216.

34. https://www.sas.com/en_in/insights/big-data/hadoop.html

AUTHORS PROFILE

Chandrashekar D K is currently working as Assistant Professor in the Department of Computer Science and Engineering at S J B Institute of Technology, Bangalore, India. And pursuing the Ph.D degree in the Department of Computer Science and Engineering at S J B Institute of Technology, Bangalore, under Visvesvaraya Technological University Belgavi, India. He obtained his B.E degree in 2009 and M.Tech degree in 2014 from Visvesvaraya Technological University Belgavi, India. His research interest is in Data Mining, Big Data Analytics and Cloud Computing.

Srikantaiah K C is currently working as Professor in the Department of Computer Science and Engineering at S J B Institute of Technology, Bangalore, India. He obtained his B.E from Bangalore Institute of Technology, M.E from University Visvesvaraya College of Engineering, Bangalore in 2002 and Ph.D degree in Computer Science and Engineering from Bangalore University, Bangalore, in the year 2014. He is guiding five Ph.D students in VTU. During his 15 years of service, he has 20 research papers to his credit. He has authored a book on Web Mining Algorithms. He has awarded best paper presentation award in the conference ICIP 2011 and his name is listed in Marquis who is who in the World 2014, 2015, 2016. His research interest is in Data Mining, Web Mining, Big Data Analytics, Cloud Analytics and Semantic Web.

V**enugopal K R** is currently Vice-Chancellor of Bangalore, University. He obtained his Bachelor of Engineering from University Visvesvaraya college of Engineering. He received his master's degree in computer science and Automation from Indian Institute of Science Bangalore. He was awarded Ph.D in Economics from Bangalore University and Ph.D in Computer Science from Indian Institute of Technology, Madras. He has a distinguished academic career and has degrees in Electronics, Economics, Law, Business Finance, Public Relations, Communications, Industrial Relations, Computer Science and Journalism. He has authored and edited 57 books on Computer Science and Economics, which include Petrodollar and the World Economy, C Aptitude, Mastering C, Microprocessor Programming, Mastering C++ and Digital Circuits and Systems etc., He has filed 101 patents. During his three decades of service at UVCE he has over 550 research papers to his credit. His research interests include Computer Networks, Wireless Sensor Networks, Parallel and Distributed