International Journal of Advanced Technology in Engineering and Science
Vol. No.4, Issue No. 10, October 2016
www.ijates.com
ijates
ISSN 2348 - 7550

APPLICATIONS AND FUTURE USE OF BIG DATA TOOLS IN NEW FIELDS

Harshita Khangarot[1], Avinav Baruah[2]

[1,2]Student, Department of computer science, JK Lakshmipat University, Jaipur

ABSTRACT

Big data is the field that is playing vital role in almost every application most probably in business, organizations, health and many more. The paper presents the scope of big data in various fields. One of them is the insurance industry which runs on data, success of any model depends on analyzing the data to make profitable decisions. Based on data of insurance of different countries comparison is made in various claims. The big data tools hadoop and spark used for analyzing the data making future perceptions and its comparison.

Keywords*: *Analyst, BDA, Dataset, Hadoop, Spark.

I. INTRODUCTION

The world without data seems to be the world without information and knowledge. Data available in excess is creating large problems to us. Big data is a field that has been attracting more and more researcher towards it varying from business to administration. It refers to data set that is complex enough and can be mined for further knowledge. Process of extracting useful information is known as big data analytic. They help in making decisions by analyzing large amount of data. Data available can be structured that can be analyzed easily or unstructured that is complex like review data.

II. FIELDS

The job of data analyst depends on data on which they have to work. The data of one field can be applicable for the development of other field.

BDA Big data analytics is a software application which analysis massive data. Such applications are firstly developed taking small data sets in pseudo-cloud and then deploying at large cloud with large data. These type of data includes running data from traffic, games, stocks or from rapidly growing systems. CF4BDA is a conceptual framework that help researchers to identify various opportunities in well structured manner and implementing it in cloud.[1]

Business: Data is the major part based on which decisions are made in business analysis. Analysts review purchase data and helps customers to decide what kinds of improvements required to meet their customers' needs. Big data helps businesses efficient, detect errors by studying real time data. Here is an example of retail industry, a brief demonstration for various functions of Big Data in many commercial activities. The following statistics 267 million transactions are made every day in Wal-Mart's 6000 worldwide stores. [20]

E-Commerce: Data analyst helps a company improving customer service by studying how they feel about products of company through reviews, comments, and suggestions. Predictive modeling techniques are used by many commercial websites for suggesting similar options when users want to browse products. Trends in purchasing or traffic of websites are also searched by data analyst.

Finance: Transaction data related to credit or debit card, account data and market data are examples of financial big data. Analysis of transaction data can be used to detect fraud activities. Investment portfolios are monitored and altered by them so that risk can be compensated and price changes.

Government: Vast data about the constituents of government is collected, but policies and security concerns or policies do not allow them from sharing such data. However, using big data can serve for improving policy decisions. For example, use of data to fill tax or any other forms for government constituents. Data analysts studies constituent's levels by monitoring social sites.

Science: Large amount of datasets are produced in various fields of science. For example, a zoologist studies about some new species and their properties about how they do various activities and their features. Data Analyst records all the data related to it and makes decisions to it. It is possible that the location varies from where the research is done and where the analyst does its activity. Various other areas of science, includes physics, chemistry, and biology, where work is done based on logics and analysis on huge datasets.

Social networking: Data analyst specializes on the data collected on social websites. They gather large data which includes comments, pictures, and videos from sites. Sorting such data, the analysts studies the preferences made by users that help in creating advertisements and better services to the users. And as networking is growing continuously, the work starts on finding new ways to use such a large data. A huge amount of data from social websites and other GPS services is personal location data. Even many nonhuman objects, such as packages or shipping containers, have location data that are used for tracking. These help in making better business related decisions and for effective advertisements.

Telecommunications: As the use of smart phones is increasing, the amount of big data in the field of telecommunications has increased very vast. Using smart phones, users preferences through their various actions and can be used for tracking user location through GPS data. This processing allows analysts who are working for telecommunications companies to improve their services to their respective customers' preferences, based on the usage of phone. Analysts also minimize drop calls and many other problems by studying large amount of datasets.

Politics: Politicians rely on polling data and ratings of approval, which were traditionally numerical. Now, however, analysts gather public sentiment data from comments on social networking and other websites.

Smart meters: Smart meters are installed on different kinds of equipment, such as cars and electric meters. The performance of equipment is transmitted by meters. Data analysts examine this data to determine the cause of any malfunctions and help prevent future ones.

Agriculture: A biotechnology firm uses sensor data to optimize crop efficiency. Test crops are planted and simulations made to measure the reaction of plants and changes according to various conditions. Its data constantly adjusts to various change in attributes, including temperature, water levels, soil composition, growth and sequencing of genes of each plant in the test bed. These simulations allow it to discover the optimal environmental conditions for specific gene types.

Learning: Variety of ways possible through which learning can be analyzed. Performance of student is predicted by analyzing the interaction with others. Attrition risk is detected at the right time and retains the student. Visual reports of educational data created for easily identifying trends. Intelligent feedback provided which eventually improves the performance. Recommending courses according to the interest ensuring no misguiding and planning schedules [19].

Scientific research: Scientific data in the field of bioinformatics, meteorology, astronomy based on discovery, probe the knowledge for simulation. For instances, large number of images of universe are generated by a sophisticated telescope is regarded as digital camera. Data collected by images is in a very huge amount. Astronomer utilizes such facilities for computing and advanced analysis methods to investigate the universe from where it has been origin.

Hospital queuing: Effective queue management in hospital and crowd control are the major issues in hospitals. To predict the waiting time for each Patient Treatment Time Prediction algorithm is proposed which saves patients time as well as frustration among them [5]. Here HQR calculates the efficient plans for the patients.

Remote Sensing: The devices which observes our plant from different perspective for making our lives easier. The applications where such data are used from urban planning, climate or hazard.

III. BIG DATA IN LIFE INSURANCE PLANNING

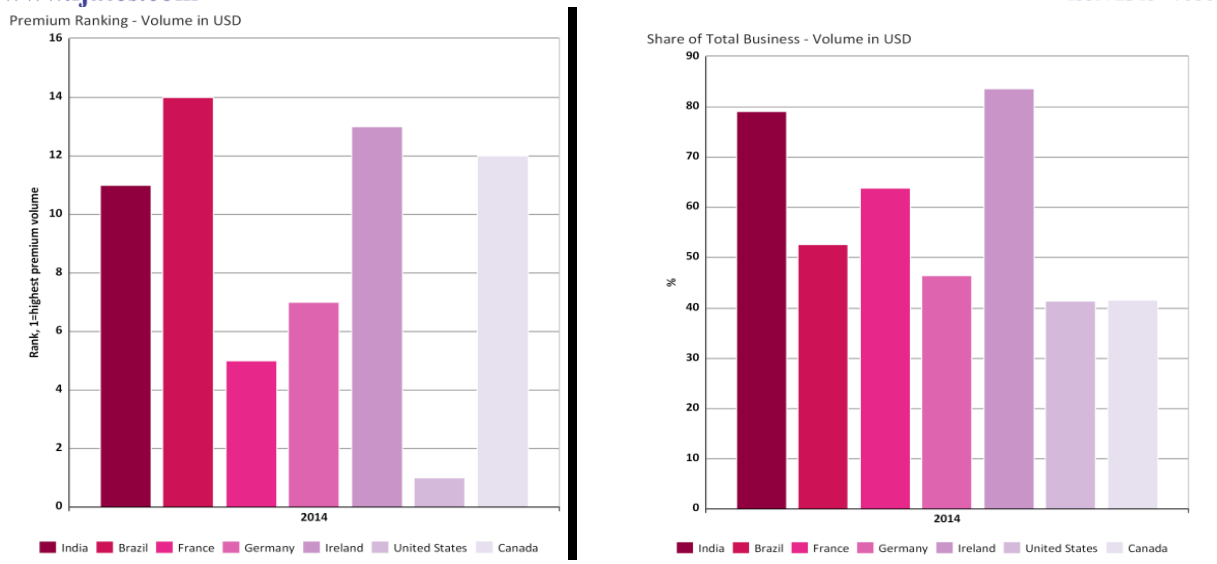
Big data has attracted every sector of industry even insurers. In this area 25% are planning, 25% are collecting data and 34% has started analyzing results using big data but much more work is needed to fully involve big data in this field. From the research it has been expected that two years from now, life insurers will use the big data and predictive analysis for customer relations, improve internal performance.

3.1 Barriers and challenges to using big data

Biggest challenge for life insurers is infrastructure limitations (71%), financial constraints (54%), lack of knowledge and expertise (34%).

3.2 Ways to improve claims made in insurance processing using big data

Fraud- It is estimated that about 10% claims are fraud. Most of the work is rule based which can be easily be transformed according to the requirement .Predictive analysis using database search and exception identifying fraud. **Subrogation-** Often subrogation is neglected in large datasets. Using this opportunity maximizes loss recovery and minimizes losses. **Settlement-** When a claim is made, for settling the claim payment is made to whoever holding the policy. But such settling can lead to cost if overpay. By analyzing the claims made earlier, instant payouts can be optimized or claim cycle can be shortened. **Loss reserve-** Prediction of size and duration of claim reported is nearly impossible. Loss reserve can be calculated by comparing the similar claims. Using updated data, analyst can make decision to understand the future claims. **Litigation-** Insurers uses analytics to calculate litigation score to predict the claims that result in one or other form of litigation. Such results can lead to settle the claims that are creating problems.



(A) **(B)**

Fig –Source:https://knoema.com/WLDINS2016/world-insurance-in-2014-back-to-life (a)premium ranking in life insurance made in various countries including India, result is in USD (b)share of total business in insurance, result is in percentage

3.3 Future use

For big data analysis top future uses of life insurers will include transforming business models, expanding customer relationships, enhancing customer value proposition and improving performance management. Survey says that all these factors will triple double its percentage by 2017[Willis tower Watson's 2015 North America Life Insurance CFO survey on big data and predictive analysis].

IV. BIG DATA TOOL- HADOOP

Hadoop is an open-source software framework to store data and run applications on clusters of commodity hardware. It possess capabilities such as large storage of any kind of data, immense processing power and handling of virtually unlimited concurrent tasks or jobs. It is a free Java-based programming framework that processes large data sets in a distributed computing environment. It is included in the Apache project which is sponsored by the Apache Software Foundation. Hadoop was created by Doug Cutting, who termed the framework after his child's stuffed toy elephant. Hadoop was influenced by Google's Map-Reduce , a software framework where an application is divided into various small parts. These parts are also called fragments or blocks. Each of these parts can be made to run on any node in the cluster. The Apache Hadoop ecosystem is comprised of the Hadoop kernel, Hadoop distributed file system(HDFS), Map-Reduce and a handful of projects which include Apache Hive, HBase and Zookeeper. Hadoop enables applications to be run on systems consisting of thousands of nodes and dealing in thousands of terabytes. Hadoop distributed file system(HDFS) permits the system to continuously operate during a node failure, by speeding up rapid data transfer rates amidst nodes. This leads to reducing the risk of catastrophic system failure, although a notable number of nodes are not functioning. Google, Yahoo, IBM etc make use of the Hadoop framework, which includes applications

comprising of search engines and advertising. Hadoop prefers the operating systems – Windows and Linux. But it has the ability to work with BSD and OS X.

4.1 Features of Hadoop

- It has the capability to store and process huge amounts of any kind of data, at a fast rate.
- Computing power: Hadoop has the ability to process big data quickly.
- Fault tolerance: Data and application processing are safeguarded against hardware failure.
- Flexibility: Unlike traditional relational databases, it is not required to pre-process data before storing it. One can store as much data as one wants and decide how to use it later.
- Low cost: The open-source framework is available for free and makes use of commodity hardware for the storage of large amount of data.
- Scalability: One can easily grow one's system to handle more data simply by adding nodes.

4.2 SPARK

Another open source big data processing framework is Apache Spark. It possesses advantages like speed, ease of use, and sophisticated analytics.

4.3 Features of Spark

- In case of Spark, less expensive shuffles are involved in the data processing. In-memory data storage and near real-time processing are possible here, which leads to faster performance.
- Also, lazy evaluation of big data queries are done in Spark. As a result, the various steps involved in data processing workflows are optimized. Higher level API is used in Spark for improving developer productivity. Also, a consistent architect model is involved for handling big data solutions.
- Spark uses an approach wherein intermediate results gets stored in memory instead of getting stored in disk. This approach is very handy especially in the case where one needs to work on the same dataset multiple times.
- Spark stores data in memory as much as possible. Next the data is stored in disk. It stores a portion of a data set in memory. The remaining amount of data is stored on the disk. One has to take into account data and use cases, while assessing the memory requirements.

4.4 Advantages of Spark

- A unified, comprehensive framework is used in Spark, so as to handle big data processing requirements.
- Using Spark, applications in Hadoop are made to run quicker both in memory and disk.
- Applications can be written in Python, Java, Scala etc. with the help of Spark.
- Various concepts such as machine learning, streaming data, SQL queries and graph data processing are supported in Spark.
- Runs everywhere: Spark runs on various platforms including Hadoop, Mesos, standalone, cloud. Moreover it can access diverse data sources, which includes HDFS, Cassandra, HBase and S3.

V. COMPARISON OF HADOOP AND SPARK

ASPECT	HADOOP	SPARK
Difficulty:	It is difficult to program and requires abstractions.	It is easy to program and does not need any abstractions.
Interactive mode:	It has no in-built interactive mode with the exception of Pig and Hive.	It has interactive mode.
Streaming:	It generate reports that are used to find answers to historical queries.	It is used to perform streaming, batch processing and machine learning, all in the same cluster.
Performance:	It does not make use of the memory of the Hadoop cluster to the maximum.	It can process batch processing jobs almost 10 to 100 times faster than Hadoop.
Latency:	It is completely disk oriented.	It assures lower latency computations. It does so by caching the partial results across its memory of distributed workers.
Ease of coding:	It is both a complex and lengthy process to write Hadoop pipelines.	It is always compact to write Spark code than to write Hadoop code.
Graph processing:	The data has to be read from the disk and written back every time, which leads to increasing of the latency, and thus Hadoop is slower.	It consists of a Graph computation library GraphX, which is very fast, and thus Spark is faster.
Hardware requirements:		
➢ Cores	4	8-16
➢ Memory	24 GB	8 GB to hundreds of gigabytes
➢ Disks	4-6 one-TB disks	4-8
➢ Network	1 GB Ethernet all-to-all	10 GB or more
Failure tolerance:	Both display good failure tolerance property, but Hadoop is a little more tolerant as compared to Spark.	Spark is slightly less tolerant as compared to Hadoop.
Security:	It possesses more security.	It possesses less security.

VI. CONCLUSION

Big data is the area which has attracted many new fields. Data is collected from sources and analyzed for making future decisions which may help in one way or the other. Many applications of big data are listed in paper. Insurance planning is one field where big data analysis is not used at that extend and is increasing day by day. Many tools can be used for analyzing huge data, such as hadoop or spark. Comparison between both tools

is being listed and concluding spark is better tool for analyzing. But it consumes lot of memory and issues related to its consumption.

REFERENCES

[1] Qinghua Lu, Zheng Li, Maria Kihl, Liming Zhu, And Weishan Zhang "CF4BDA: A Conceptual Framework for Big Data Analytics Applications in the Cloud", IEEE Access, October 27, 2015.

[2] Sihai Zhang, (Member, Ieee), Dandan Yin, Yanqin Zhang, And Wuyang Zhou, "Computing on Base Station Behavior Using Erlang Measurement and Call Detail Record", Ieee Transactions On Emerging Topics In Computing, September, 2015.

[3] Marco Viceconti, Peter Hunter, and Rod Hose, "Big Data, Big Knowledge: Big Data for Personalized Healthcare", Ieee Journal Of Biomedical And Health Informatics, Vol. 19, No. 4, July 2015.

[4] S. M. Riazul Islam, (Member, Ieee), Daehan Kwak, Md. Humaun Kabir, Mahmud Hossain, And Kyung-Sup Kwak, (Member, Ieee), "The Internet Of Things For Health Care:A Comprehensive Survey", IEEE Access , June 4, 2015.

[5] Jianguo Chen, (Student Member, Ieee), Kenli Li, (Senior Member, Ieee),Zhuo Tang1, (Member, Ieee), Kashif Bilal, And Keqin Li, (Fellow, Ieee), "A Parallel Patient Treatment Time Prediction Algorithm and Its Applications in Hospital Queuing-Recommendation in a Big Data Environment", IEEE Access, May 9, 2016.

[6] Mingmin Chi, Member, IEEE, Antonio Plaza, Fellow, IEEE, J´on Atli Benediktsson, Fellow, IEEE, Zhongyi Sun, Jinsheng Shen, and Yangyong Zhu, "Big Data for Remote Sensing: Challenges and Opportunities ",The 462th Xiangshan Science Conference, Beijing, China, May 29-31, 2013.

[7] Yaxiong Zhao_, Jie Wu, and Cong Liu," Dache: A Data Aware Caching for Big-Data Applications Using the MapReduce Framework", Tsinghua Science And Technology Issn 1007-0214 05/10 pp39-50 Volume 19, Number 1, February 2014.

[8] Xue-Wen Chen, (Senior Member, IEEE), And Xiaotong Lin, "Big Data Deep Learning: Challenges and Perspectives", IEEE Access, May 28, 2014.

[9] Lei Xu, Chunxiao Jiang, (Member, IEEE), Jian Wang, (Member, IEEE), Jian Yuan, (Member, IEEE), And Yong Ren, (Member, IEEE)," Information Security in Big Data: Privacy and Data Mining", IEEE Access, October 20, 2014.

[10] Yanhao Huang and Xiaoxin Zhou, Fellow, IEEE, Fellow, CSEE, "Knowledge Model for Electric Power Big Data Based on Ontology and Semantic Web", Csee Journal Of Power And Energy Systems, Vol. I, No. I, March 2015.

[11] Anton Akusok, Kaj-Mikael Björk, Yoan Miche, And Amaury Lendasse, "High-Performance Extreme Learning Machines: A Complete Toolbox for Big Data Applications", IEEE Access, July 17, 2015.

[12] Sergio Pissanetzky, (Life Member, IEEE), "On the Future of Information: Reunification, Computability, Adaptation, Cybersecurity, Semantics", IEEE Access, April 1, 2016.

[13] Shui Yu, (Senior Member, IEEE), "Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data", IEEE Access, June 24, 2016.

[14] Dilpreet Singh and Chandan K Reddy, "A survey on platforms for big data analytics", Springer Open General, Journal of Big Data 2014.

[15] Thomas H. Davenport Jill Dyché, "Big Data in Big Companies", International Institution for Analytics, May 2013.

[16] Nada Elgendy and Ahmed Elragal, "Big Data Analytics: A Literature Review Paper", Springer International Publishing Switzerland 2014.

[17] Weiyi Shangy, Zhen Ming Jiangy, Hadi Hemmatiy, Bram Adamsz, Ahmed E. Hassany, Patrick Martin, "Assisting Developers of Big Data Analytics Applications When Deploying on Hadoop Clouds", ICSE 2013, San Francisco, CA, USA.

[18] Kuchipudi Sravanthi, Tatireddy Subba Reddy, "Applications of Big data in Various Fields", International Journal of Computer Science and Information Technologies, Vol. 6 (5) , 2015.

[19] Katrina Sin and Loganathan Muthu, "Application Of Big Data In Education Data Mining And Learning Analytics – A Literature Review", Ictact Journal On Soft Computing: Special Issue On Soft Computing Models For Big Data, July 2015, Volume: 05, Issue: 04.

[20] C.L. Philip Chen, Chun-Yang Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data", Information Sciences 275 (2014), Elsevier Inc.