ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

Big Data: Why? What? How?

K.Soniya[1]

Assistant Professor, Department of MCA, Sankara College of Science and Commerce, Coimbatore, India [1]

Abstract: Due to recent technological development, the amount of data generated by internet, social networking sites, sensor networks, healthcare applications, and many others are drastically increasing day by day. All the enormous measures of data produced from various sources in multiple formats with very high speed are referred as big data. In today's digital world, where lots of information is stored, the analysis of the databases can provide the opportunities that lead to better decisions in healthcare, business and others. This paper intends to define Big Data, its Applications and Techniques used for the analytics.

Keywords: Stock-ticker data, Internet of Things, Point-Of-Sale, Anonymization, Correlations and High-Frequency Trading.

I. INTRODUCTION

The amount of data is growing all around the world every day. Initially data has been stored in a structured format, and it was transactional data, but as more and more organizations started analyzing the data, after the advent of networking sites, the data existed outside was extremely large, also things started getting difficult in terms of the unstructured data (variety) and speed of data at which it is created (velocity). Business Intelligence analytics became the most important to analyze the data and determine the market growth, product competitiveness and it addresses three V's namely Volume, Variety, and Velocity. In this article we have defined the concept of Big Data in Section I, its applications in Section II, illustrated the life cycle in Section III, and discussed the techniques used in analysis of Big Data in Section IV, tools in Section V.

II. BIG DATA

Big Data analysis differs from traditional data analysis primarily due to the Volume, Velocity and Variety characteristics of the data being processed. Volume does not indicate alone large but a small amount of data could have sources of different types, Structured and unstructured data.

Variety refers to the nature of data from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions. Velocity of data addresses the speed at which different types of data are entered and analyzed.

III. APPLICATIONS

Big data is widely used in many applications and some of them are as follows:

BIG DATA IN AEROSPACE

Earned Value Management (EVM) is one of the most effective methods for the project management. Actual Cost and earned value are the parameters used for monitoring projects. These parameters are compared with planned value to analyze the project status. EVM covers scope, cost, time and unifies them in a common framework that allows evaluation of project health.

BIG DATA IN GOVERNANCE

Big data is used to improve many aspects of our cities and countries. For example, it allows cities to optimize traffic flows based on real time traffic information as well as social media and weather data. A number of cities are currently piloting big data analytics with the aim of turning themselves into Smart Cities, where the transport infrastructure and utility processes are all joined up. Where a bus would wait for a delayed train and where traffic signals predict traffic volumes and operate to minimize jams.

BIG DATA IN HEALTHCARE

Big data is used in Pervasive healthcare in a disease-specific manner to present a comprehensive survey. It covers the major diseases and disorders that can be quickly detected and treated. with the use of technology, such as fatal and non-fatal falls, Parkinson's disease, cardio-vascular disorders, stress, etc. Also it addresses diseases, other permanent handicaps, like blindness, motor disabilities, paralysis, etc. It provides understanding of the various aspects of pervasive healthcare with respect to different diseases.

FINANCIAL TRADING

High-Frequency Trading (HFT) is an area where big data finds a lot of use today. Big data algorithms are used to make trading decisions. Today, the majority of equity trading now takes place via data algorithms that increasingly take into account signals from social media networks and news websites to make buy and sell decisions in split seconds.

BUSINESS PROCESSES

The Big data is also increasingly used to optimize business processes. Retailers like Flipkart, Amazon are able to optimize their stock based on predictions generated from social media data, web search trends and weather forecasts. Geographic positioning and radio frequency identification sensors are used to track goods or delivery vehicles and optimize routes by integrating live traffic data, etc.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

IV. BIG DATA ANALYTICS LIFE CYCLE

A step-by-step methodology is needed to organize the activities and tasks involved in performing analysis on Big data with acquiring, processing, analyzing and repurposing data. The diagram explores a specific data analytics lifecycle that organizes and manages the tasks and activities associated with the analysis of Big Data.

- Data Acquisition and Filtering
- Data Extraction & Cleansing
- Data Aggregation
- Data Analysis
- Data Visualization

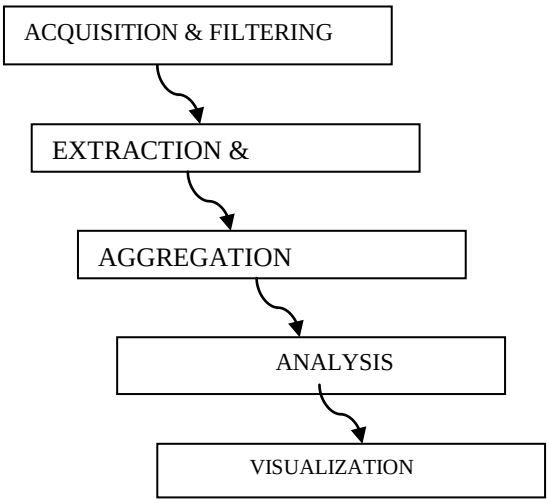

Fig 1 Big Data Analytics Lifecycle

In Acquisition & Filtering phase data is obtained from multiple sources, where in Extraction & Cleansing, the data is pre-processed. In Aggregation phase, for the protection of privacy, access restrictions as well as falsifying data techniques are used. The approaches to privacy protection are based on encryption procedures. Encryption based techniques can be further divided into Identity Based Encryption (IBE), Attribute Based Encryption (ABE) and storage path encryption. In addition, to protect the sensitive information, hybrid clouds are utilized where sensitive data are stored in private cloud. The data processing phase incorporates Privacy Preserving Data Publishing (PPDP) and knowledge extraction from the data. In PPDP, anonymization techniques such as generalization and suppression are utilized to protect the privacy of data. These mechanisms can be further divided into clustering, classification and association rule mining based techniques. While clustering and classification split the input data into various groups, association rule mining based techniques find the useful relationships and trends in the input data.Lightweight incremental algorithms should be considered that are capable of achieving robustness, high accuracy and minimum pre-processing latency. Further ahead, Internet of Things (IoT) would lead to connection of all of the things that people care about in the world due to which much more data would be produced than nowadays. Indeed, IoT is one of the major driving forces for big data analytics.

V. TECHNIQUES

A variety of Machine Learning and Data mining algorithms are available for creating valuable analytic platforms. There are 7 widely used Big Data analysis techniques and are as follows:

- Association rule learning
- Classification tree analysis
- Genetic algorithms
- Machine learning
- Regression analysis
- Sentiment analysis
- Social network analysis

ASSOCIATION RULE LEARNING

Association rule learning is a method for discovering interesting correlations between variables in large databases. It was first used by major supermarket chains to discover interesting relations between products, using data from supermarket Point-Of-Sale (POS) systems. E.g. Are people who purchase tea more or less likely to purchase carbonated drinks? Association rule learning is being used to help:

- Place products in better proximity to each other in order to increase sales.
- Extract information about visitors to websites from web server logs.
- Analyze biological data to uncover new relationships.
- Monitor system logs to detect intruders and malicious activity.
- Identify if people who buy milk and butter are more likely to buy bread.

CLASSIFICATION TREE ANLAYSIS

Statistical classification is a method of identifying categories that a new observation belongs to. It requires a training set of correctly identified observations – historical data in other words. E.g. Which categories does this document belong to? Statistical classification is being used to:

- Automatically assign documents to categories.
- Categorize organisms into groupings.
- Develop profiles of students who take online courses.

GENETIC ALGORITHMS

Genetic algorithms are inspired by the way evolution works – that is, through mechanisms such as inheritance, mutation and natural selection. These mechanisms are used to "evolve" useful solutions to problems that require optimization.

E.g. Which TV programs should we broadcast, and in what time slot, to maximize our ratings? Genetic algorithms are being used to:

- Schedule doctors for hospital emergency rooms.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

- Return combinations of the optimal materials and engineering practices required to develop fuel-efficient cars.
- Generate "artificially creative" content such as puns and jokes.

MACHINE LEARNING

Machine learning includes software that can learn from data. It gives computers the ability to learn without being explicitly programmed, and is focused on making predictions based on known properties learned from sets of "training data."

E.g. Which movies from our catalogue would this customer most likely want to watch next, based on their viewing history?

Machine learning is being used to help:

- Distinguish Between Spam And Non-Spam Email Messages
- Learn User Preferences And Make Recommendations Based On This Information.
- Determine the best content for engaging prospective customers.
- Determine the probability of winning a case, and setting legal billing rates.

REGRESSION ANALYSIS

At a basic level, regression analysis involves manipulating some independent variable (i.e. background music) to see how it influences a dependent variable (i.e. time spent in store). It describes how the value of a dependent variable changes when the independent variable is varied. It works best with continuous quantitative data like weight, speed or age.

E.g. How does your age affect the kind of car you buy?

Regression analysis is being used to determine how:

- Levels of customer satisfaction affect customer loyalty.
- The number of supports calls received may be influenced by the weather forecast given the previous day.
- Neighborhood and size affect the listing price of houses
- To find the love of your life via online dating sites.

SENTIMENT ANALYSIS

Sentiment analysis helps researchers determine the sentiments of speakers or writers with respect to a topic.

E.g. how well our new is return policy being received?

Sentiment analysis is being used to help:

- Improve service at a hotel chain by analyzing guest comments.
- Customize incentives and services to address what customers are really asking for.
- Determine what consumers really think based on opinions from social media.

SOCIAL NETWORK ANALYSIS

Social network analysis is a technique that was first used in the telecommunications industry, and then quickly adopted by sociologists to study interpersonal relationships. It is now being applied to analyze the relationships between people in many fields and commercial activities. Nodes represent individuals within a network, while ties represent the relationships between the individuals. Social network analysis is being used to:

- See how people from different populations form ties with outsiders
- Find the importance or influence of a particular individual within a group
- Find the minimum number of direct ties required to connect two individuals
- Understand the social structure of a customer base whether our business wants to discover interesting correlations categorize people into groups, optimally schedule resources, or set billing rates, a basic understanding of the seven techniques mentioned above can help Big Data work for us.

VI. TOOLS

Variety of big data products are available, many products comes with a combination of infrastructure, visualization, and analytical capabilities. Some of them are listed below:

- Spark
- Hive
- Hadoop
- Hbase
- Impala
- MapReduce
- R Programming
- Statistics
- Tera data
- Zookeeper

Apache Spark is a lightning-fast cluster computing designed for fast computation. Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Impala is the open source, native analytic database for Apache Hadoop. MapReduce is a programming paradigm that runs in the background of Hadoop to provide scalability and easy data-processing solutions. Teradata is a popular Relational Database Management System (RDBMS) suitable for large data warehousing applications. It is capable of handling large volumes of data and is highly scalable. ZooKeeper is a distributed co-ordination service to manage large set of hosts.

VII. CONCLUSION

Big data is analyzed for bits of knowledge that leads to better decisions and strategic moves for overpowering businesses and is more than simply a matter of size; it is an opportunity to find insights in new and emerging types of data and content, to make businesses more agile and to answer questions that were previously considered beyond

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

reach. In this paper, we have presented the concepts, aspects and where it is widely used and then discussed how analytics are done with the techniques that are sufficient for big data processing.

REFERENCES

[1] G. Noseworthy, Infographic: Managing the Big Flood of Big Data in Digital Marketing, 2012.
[2] Elena Geanina ULARU, Florina Camelia PUICAN, Anca APOSTU, Manole VELICANU: Perspectives on Big Data and Big Data Analytics.
[3] Napoleon D, Praneesh M, Subramanian MS, Sathya S (2012) Manhattan distance based affinity propagation technique for clustering in remote sensing images. Int J Adv Res Comput Sci Softw Eng (IJARCSSE) 2(3):326–330
[4] C-W, Lai C-F, Chao H-C, Vasilakos AV. Big data analytics: a survey.
[5] Praneesh, M., and Jaya R. Kumar. "Article: Novel Approach for Color based Comic Image Segmentation for Extraction of Text using ModifyFuzzy Possiblistic C-Means Clustering Algorithm." IJCA Special Issue on Information Processing and Remote Computing IPRC (1) (2012):16-18
[6] Sedayao J, Bhardwaj R. Making big data, privacy, and anonymization work together in the enterprise: experiences and issues. Big Data Congress; 2014.
[7] Qin Y, et al. When things matter: a survey on data-centric internet of things. J Netw Comp Appl. 2016.
[8] Napoleon, D., and M. Praneesh. "Detection of Brain Tumor using Kernel Induced Possiblistic C-Means Clustering." International Journal of Computer & organization Trends (IJCOT) i 1.3 (2013): 436-438.
[9] Liu S. Exploring the future of computing. IT Prof. 2011.
[10] Oracle Big Data for the Enterprise, 2012.
[11] http://www.zdnet.com/topic/the-power-of-iot-and-big-data/
[12] http://www.datamation.com/applications/why-big-data-and-the-internet-of-things-are-a-perfect-match.html
[13] http://www.forbes.com/sites/louiscolumbus/2016/10/02/2016-internet-of-things-iot-big-data-business-intelligence-update/
[14] Napoleon, D., Praneesh, M., Sathya, S., & SivaSubramani, M. (2012). An Efficient Numerical Method for the Prediction of Clusters Using K-Means Clustering Algorithm with Bisection Method. In Global Trends in Information Systems and Software Applications (pp. 256-266). Springer Berlin Heidelberg.
[15] Agrawal D., Bernstein P., Bertino E., Davidson S., Dayal U., Franklin M., ... Widom J. (2012). Challenges and Opportunities with Big Data: A white paper prepared for the Computing Community Consortium committee of the Computing Research Association. http://cra.org/ccc/resources/ccc-led-whitepapers/
[16] http://strata.oreilly.com/2010/01/roger-magoulas-on-big-data.html
[17] http://www.eol.ucar.edu/projects/bomex/
[18] http://www.eol.ucar.edu/projects/bomex/images/DataAcquisitionSystem.jpg
[19] http://www.bis.gov.uk/assets/biscore/science/docs/i/11-p123-international-comparative-performance-uk-research-base-2011.pdf
[20] Lane, J. (2010) “Let’s make science metrics more scientific”, Nature 464, 488–489.
[21] Furman, J. L., Murray, F. & Stern, S. (2012) “Growing Stem Cells: The Impact of Federal Funding Policy on the U.S. Scientific Frontier”, J. Pol. Anal. Manage. 31, 661–705.
[22] File: Experiment Pasteur English.jpg - Wikipedia, the free encyclopedia.http://en.wikipedia.org/wiki/File:Experiment_Pasteur_English.jpg
[23] Gertler, P.J., Martinez, S., Premand, P., Rawlings, L.B., & Vermeersch, C.M.J. (World Bank 2011) “Impact Evaluation in Practice”,athttp://siteresources.worldbank.org/EXTHDOFFICE/Resources/5485726129545562862O/Impact_Evaluation_in_Practice.pdf
[24] Journal of Policy Analysis and Management - Volume 31, Issue 3 - Summer 2012 - Wiley Online Library. At <http://onlinelibrary.wiley.com/doi/10.1002/pam.2012.31.issue-3/issuetoc>
[25] Lane, J. & Black, D. (2012) “Overview of the Science of Science Policy Symposium”, J. Pol. Anal. Manage. 31, 598–600.
[26] NAS CNSTAT SciSIP Principal Investigator Conference. (2012).at <http://www7.nationalacademies.org/cnstat/SciSIP%20Invitation.pdf>
[27] Largent, M. A. & Lane, J. I. (2012) “Star Metrics and the Science of Science Policy”, Review of Policy Research 29, 431–438.
[28] Barajas J, Akella R, Holtan M, Flores A (2016) Experimental designs and estimation for online display advertising attribution in marketplaces. Marketing Sci. 35(3):000–000.