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

A Survey of Characteristics and Emerging Technologies in Big Data

Dr. S. Krishnaveni[1], R. Divya[2], V. Attchara[3]

[1,2,3]Assistant Professor, Department of Computer Applications, Pioneer College of Arts and Science, Coimbatore, India

Abstract: Big data make values for business and research, but create significant quarrel in terms of networking, storage, management, analytics and ethics of data. The 3Vs have been extended to other corresponding characteristics of big data: Volume: big data doesn't sample; it just views and follows what happens. Velocity: big data is frequently available in real-time. Variety: big data depicts from multimedia like Audio, Video, Text and Animation; plus it entries missing pieces through data mixture. Engineers, computer scientists, statisticians and social scientists are needed to tackle, discover and understand big data for Multidisciplinary collaborations. This survey presents an overview of big data initiatives, technologies and Characteristics of Big Data.

Keywords: Three V's of Big Data, Emerging Technologies.

I. Introduction

A growing number of data sources – such as commercial media – and an increasing number of media-rich data types – such as X-rays and video – are fueling the faces associated with big data at companies that might in no way have thought of themselves as big data customers [1]. Big data technologies are important in providing more enormous analysis, this may lead to more actual decision-making resulting in greater prepared efficiencies, cost reductions, and condensed risks for the business. To attach the power of big data, you would need a communications that can direct and process massive volumes of structured and unstructured data in real time and can keep data privacy and security.

II. Three V's of Big Data

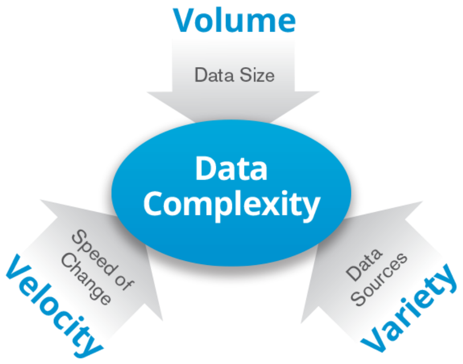

Figure 1: Three V's of Big Data

2.1 VOLUME

Volume refers to the quantity of data, mixture refers to the quantity of types of data and Volume is the V most related with big data because, volume can be huge. What we're talking about here is quantities of data that make almost inconceivable proportions. Consider our new world of linked apps [8]. Each one is carrying a Smartphone. Let's seem at a simple example, a to-do list app. More and more sellers are organizing app data in the cloud, so users can access their to-do lists across devices. While many apps use a freemium model, where a free version is used as a loss-leader for a premium version, SaaS-based application trade lean to have a lot of data to store.

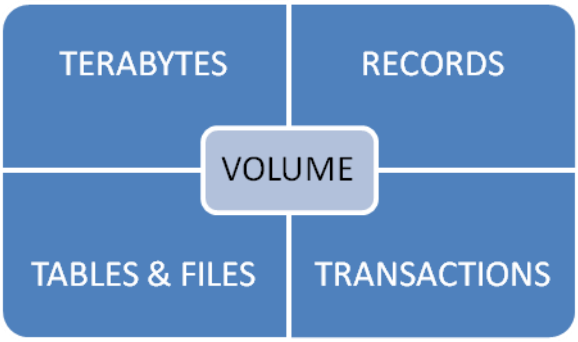

Figure 2: Steps in Volume

2.2 VELOCITY

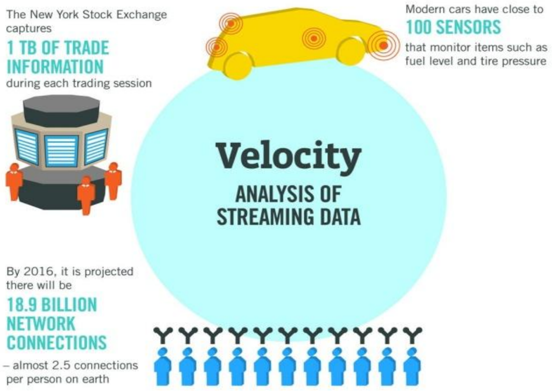
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Figure 3: Analysis of Streaming Data

Big Data Velocity compact with the rate at business processes, machines, networks and human interaction with things like social media sites, mobile devices, etc. The flow of data is huge and unbroken. This real-time data can help researchers and businesses create valuable resolutions that provide planned competitive advantages and ROI if you are able to hold the velocity. Internal propose that sample data can help deal with issues like volume and velocity. Velocity is evaluated of how fast the data is coming in. Analysis of Streaming Data Showed in Figure 3[16].

For example, let's say you're running a presidential battle and you want to know how the people "out there" are feeling about your runner right now. How would you do it? One way would be to certify some to grab a steady stream of tweets, and subject them to emotional analysis.

2.3 VARIETY

Variety refers to the many sources and types of data both planned and unplanned. We used to store data from sources like spreadsheets and databases. Data comes in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc in now a days. This variety of unordered data constructs problems for storage, mining and analyzing data. Jeff Veis, VP Solutions at HP Autonomy obtainable how HP is helping organizations convention with large challenges including data variety[8].

For example, email messages. An authorized discovery process may require sifting through thousands to millions of email messages in a group. Not one of those messages is going to be accurately like another. Each one will consist of a sender's email address, a destination, plus a time beat. Each message will have human-written text and possibly attachments.

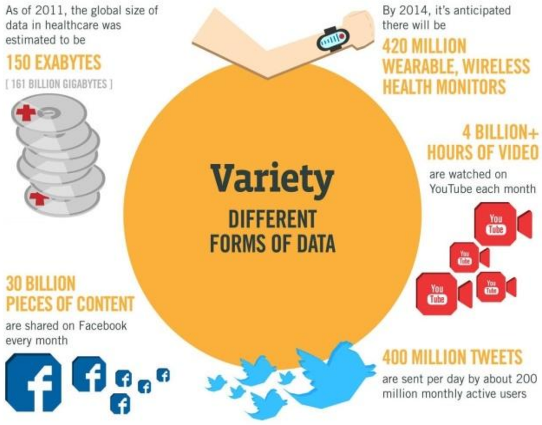
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
Variety
DIFFERENT FORMS OF DATA
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Figure 4: Variety Different forms of Data

Photos, videos, audio recordings, email messages, documents, books, presentations, tweets and ECG strips are all data, but they're generally unstructured, and extremely varied. All that data diversity makes up the variety vector of big data and different forms of data showed in figure 4[16].

III.EMERGING TECHNOLOGIES FOR BIG DATA

A. Column-oriented databases

The Conventional, row-oriented databases are exceptional for online business processing with high update speeds, but they descend short on query presentation as the data volumes raise and as data turn into more unstructured. Column-oriented databases store data with a center on columns, instead of rows, allowing for vast data compression and very fast query times. The disadvantage to these databases is that they will typically allow batch updates, having a much slower update time than conventional models [7].

B. Schema-less databases, or NoSQL databases

There are quite a few database kinds that well into this category, such as key-value stores and document stores, which focus on the storage and recovery of large volumes of unstructured, semi-structured, or even structured data. They achieve performance achieves by doing away with some (or all) of the limitations traditionally linked with conventional databases, such as read-write consistency, in exchange for scalability and distributed processing[7].

C. MapReduce

This is a programming pattern that allows for huge job execution scalability against thousands of servers or clusters of servers[7]. Any MapReduce performance consists of two tasks:

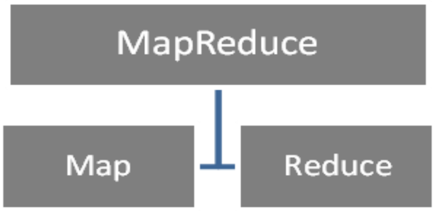

Figure 5: Two tasks of MapReduce

a. The "Map" task, where an input dataset is transformed into a different set of key/value pairs, or tuples.
b. The "Reduce" task, where various outputs of the "Map" task are shared to form a condensed set of tuples.

D. Hadoop

Hadoop is by future the most popular implementation of MapReduce, being an totally open source platform for managing Big Data. It is stretchy enough to be able to work with multiple data sources, either cumulating multiple sources of data in structure to do huge scale processing, or even reading data from a database in order to run processor-intensive machine learning jobs [4]. It has several applications, but one of the top use cases is for big volumes of constantly changing data, such as location-based data from weather or traffic sensors, web-based or social media data, or machine-to-machine transactional data.

E. Hive

Hive is a "SQL-like" connection that allows traditional BI applications to run queries next to a Hadoop cluster. It was developed initially by Facebook, but has been ready open source for some time now, and it's a higher-level concept of the Hadoop framework that allows anyone to make queries against data stored in a Hadoop cluster just as if they were control a conventional data store. It amplifies the reach of Hadoop, making it more familiar for BI users [6].

F. PIG

PIG is a new bridge that tries to bring Hadoop faster to the realities of developers and business users, like to Hive. Unlike Hive, however, PIG consists of a "Perl-like" language that allows for query execution over data stored on a Hadoop cluster, as an alternative of a "SQL-like" language. PIG was developed by Yahoo!, and, just like Hive, has also been made fully open source [7].

G. WibiData

WibiData is a mixture of web analytics with Hadoop, being made on top of HBase, which is itself a database layer on top of Hadoop. It agree to web sites to improved explore and work with their user data, permiting real-time responses to user manners, such as serving customized content, recommendations and decisions.

H. PLATFORA

Perhaps the supreme limitation of Hadoop is that it is a very low-level realization of MapReduce, need wide developer knowledge to operate. Between preparing, testing and running jobs, a full cycle can take hours, reducing the interactivity that users liked with conventional databases. PLATFORA is a platform that revolved user's queries into Hadoop jobs mechanically, thus creating a concept layer that anyone can use to simplify and arrange datasets stored in Hadoop.

I. SkyTree

SkyTree is a high-performance machine learning and data analytics platform focused specifically on handling Big Data. Machine learning, in turn, is an necessary part of Big Data, since the massive data volumes create manual exploration, or even straight automated study methods impractical or too expensive[7].

IV. CONCLUSION

In this paper, we presented the Characteristics of Big Data such as Volume, Variety, and Velocity. We also present the Emerging technology of Big Data. The accessibility of Big Data, low-cost commodity hardware, and novel information management and logical software has formed a unique instant in the history of data analysis. The meeting of these trends means that we have the potential required to analyze astonishing data sets rapidly and cost-effectively for the first time in history. The big data technologies have been and continue to be developed.

REFERENCES

[1] Harati A, Lopez S, Obeid I, Picone J, Jacobson M, Tobochnik S. The TUH EEG CORPUS: A big data resource for automated eeg interpretation. In: Proceeding of the IEEE Signal Processing in Medicine and Biology Symposium, 2014. pp 1–5.

[2] Inside Big Data- Available [Online] :http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity.

[3] Apache Mahout, February 2, 2015. [Online]. Available: http://mahout.apache.org/.

[4] Essa YM, Attiya G, El-Sayed A. Mobile agent based new framework for improving big data analysis. In: Proceedings of the International Conference on Cloud Computing and Big Data. 2013, pp 381–386

[5] Katal A, Wazid M, Goudar R. Big data: issues, challenges, tools and good practices. In: Proceedings of the International Conference on Contemporary Computing, 2013. pp 404–409.

[6] Sobia and Love Arora Technologies to Handle Big Data: Proceedings of International Conference on Communication, Computing and System (ICCCS), 2014.

[7] International Conference on Consumer Electronics(ICCE) 2013, Dr. Sawant Kaur Key note. Available [Online]: http:// www. techrepublic.com/blogs/big-data-analytics/10-emerging-technologies-for-big-data

[8] Mr.Mohammed Raziuddin & Prof T.Venkata Ramana : Literature Survey Big Data: Proceedings of IJAEGT, Vol 3 Iss 4 April 2015.

[9] Big data and analytics—an IDC four pillar research area, IDC, Tech. Rep. 2013. [Online]. Available: http://www.idc.com/prodserv /FourPillars/bigData/index.jsp.

[10] Big Data Survey Research Brief", Tech. Rep.SAS, 2013.

[11] Elena Geanina ULARU, Florina Camelia PUICAN, Anca APOSTU, Manole Velicanu," Perspectives on Big Data and Big Data Analytics", Database Systems Journal vol. III, no. 4/2012.

[12] Srinivasan,"SOA and WOA Article, Traditional vs. Big Data Analytics, "Why big data analytics is important to enterprises", [Online]. Available: http://soa.sys-con.com/node/1968472.

[13] Dhruba Borthakur, The Hadoop Distributed File System: Architecture and Design. [Online] Available: https://hadoop.apache.org

[14] Available:http://www.mckinsey.com NESSI-Big Data White Paper," Big Data –a new world of opportunities" December 2012.

[15] Cormode G, Duffield N. Sampling for big data: a tutorial. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014. pp 1975–1975.

[16] Big data Images Available [Online]:http://www.pinterest.com/big-data-analyticas.

[17] Characteristics of Big Data Images Available [Online]: http://www. mytechnology.com/IT/blogs/7151/the-four-vs-of-big-data

BIOGRAPHY

Dr. S. Krishnaveni completed MCA., M.Phil., Ph.D., in Computer Science and currently working as an Assistant Professor, Dept. of Computer Applications in Pioneer College of Arts and Science. Four years of experience in teaching and published fourteen papers in International Journals and also presented Ten papers in various National and International conferences. Area of research are Data mining and warehousing, Grid computing, Mobile computing, Cloud computing, Bioinformatics and Computer Network.

Ms.R.Divya pursuing M.Phil in Computer Science and currently working as an Assistant Professor, Dept. of Computer Applications in Pioneer College of Arts and Science. Three years of experience in teaching and presented papers in various National and International conferences. Editorial Members in College Magazine and our Proceedings. Area of research is Data mining and warehousing, Computer Networks, and Cloud Computing.

Ms.V.Attchara completed M.Sc in Computer Science and currently working as an Assistant Professor, Dept. of Computer Applications in Pioneer College of Arts and Science. Two years of experience in teaching and presented papers in various National and International conferences. Area of research is Data mining and warehousing, Computer Networks, Grid Computing.