Int. J. of Intelligent Computing and Applied Sciences

A SURVEY ON BIG DATA ANALYTICS

Surajit Mohanty*, Rajeev Agarwal, Shekharesh Barik

***Corresponding author: mohanty. surajit@gmail.com**

Abstract: Big data is certainly one of the biggest buzz phrases in IT today. Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. The process of research into massive amounts of data to reveal hidden patterns and secret correlations named as big data analytics. These useful information for companies or organizations with the help of gaining richer and deeper insights and getting an advantage over the competition. For this reason, big data implementations need to be analyzed and executed as accurately as possible. This paper presents an overview of what is big data, its characteristics and platforms and tools for big data analytics.

Keywords: Big data, Data Analytics

1. Introduction

People and devices are constantly generating data. While streaming a video, playing the latest game with friends, or making in-app purchases, user activity generates data about their needs and preferences, as well as the quality of their experiences. Even when they put their devices in their pockets, the network is generating location and other data that keeps services running and ready to use.

In 2014, estimates put worldwide data generation at a staggering 7ZB [1], and by 2018 each smart phone is expected to generate 2GB of data every month [2]. At the same time, the big data technology and services market is expected to grow at a 40 percent compound annual growth rate (CAGR) – about seven times the rate of the overall ICT market – with revenues expected to reach USD 16.9 billion in 2015 [3]. Clearly, the age of big data has begun.Big data was a serious problem just a few years ago. When data volumes started skyrocketing in the early 2000s, storage and CPU technologies were overwhelmed by the numerous terabytes of big data to the point that IT faced a data scalability crisis. Then we were once again snatched from the jaws of defeat by Moore's law. Storage and CPUs not only developed greater capacity, speed, and intelligence; they also fell in price. Enterprises went from being unable to afford or manage big data to lavishing budgets on its collection and analysis.Today, enterprises are exploring big data to discover facts they didn't know before. This is an important task right now because the recent economic recession forced deep changes into most businesses, especially those that depend on mass consumers. Using advanced analytics, businesses can study big data to understand the current state of the business and track still evolving aspects such as customer behavior.

Big data analytics is where advanced analytic techniques operate on big data sets. Hence, big data analytics is really about two things—big data and analytics—plus how the two have teamed up to create one of the most profound trends in business intelligence (BI) today.

2. Big Data

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time[4]. Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many peta-bytes of data. Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale[5].

In a 2001 research report[6] and related lectures, META Group (now Gartner) analyst Doug Laney defined data growth challenges and opportunities as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources). Gartner, and now much of the industry, continue to use this "3Vs" model for describing big data [7]. In 2012, Gartner updated its definition as follows: "Big data is high volume, high velocity, and/or high variety information assets that require

new forms of processing to enable enhanced decision making, insight discovery and process optimization [8]."Additionally, a new V "Veracity" is added by some organizations to describe it[9].

Gartner's definition of the 3Vs is still widely used, and in agreement with a consensual definition that states that "Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value [10]".The 3Vs have been expanded to other complementary characteristics of big data [11]:

- Volume: big data doesn't sample. It just observes and tracks what happens
- Velocity: big data is often available in real-time
- Variety: big data draws from text, images, audio, video; plus it completes missing pieces through data fusion
- Machine Learning: big data often doesn't ask why and simply detects patterns
- Digital footprint: big data is often a cost-free byproduct of digital interaction

The growing maturity of the concept fosters a more sound difference between big data and Business Intelligence, regarding data and their use:

1. Business Intelligence uses descriptive statistics with data with high information density to measure things, detect trends etc.
2. Big data uses inductive statistics and concepts from nonlinear system identification[12] to infer laws (regressions, nonlinear relationships, and causal effects) from large sets of data with low information density[13] to reveal relationships, dependencies and perform predictions of outcomes and behaviors.

2.1. When data becomes big?

Big data means data that cannot be handled and processed in a straightforward manner. A spreadsheet fits in memory; it is reasonably quick to determine if the data is clean, the values are reasonable, and the results can be computed rapidly. In contrast, a big dataset won't fit in memory, so it will be hard to check whether it is clean. Computations will take a long time. New data may well be constantly streaming in, so the processing system needs to make decisions about which part of the stream to capture as shown in fig 1. The dataset may consist of images, natural language text, or heterogeneous data, so it will be hard to predict where the database join keys are. Finally, a big dataset will probably be so large as to not fit on a single hard drive; as a result, it will be stored on several different disks, and will be processed on a number of cores. Queries will have to be distributed and written to work across network.

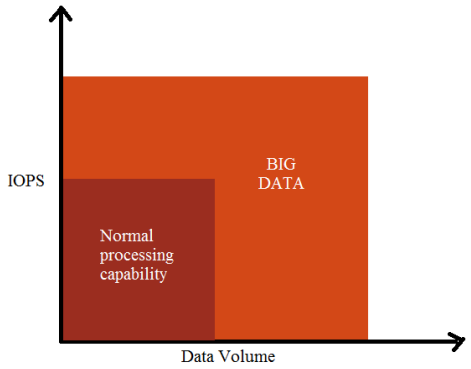

Fig. 1: IOPS(Input Output per Second)vs data volume

2.2 Characteristics

Why is Big Data different from any other data that we have dealt with in the past? There are "four V's" that characterize this data: Volume, Velocity, Variety, and Veracity:

2.2.1 Volume:

Most organizations were already struggling with the increasing size of their databases as the Big Data tsunami hit the data stores. According to *Fortune* magazine, we created 5 Exabyte of digital data in recorded time until 2003.

In 2011, the same amount of data was created in two days. By 2013, that time period is expected to shrink to just 10 minutes[14].

A decade ago, organizations typically counted their data storage for analytics infrastructure in terabytes. They have now graduated to applications requiring storage in petabytes. This data is straining the analytics infrastructure in a number of industries. For a communications service provider (CSP) with 100 million customers, the daily location data could amount to about 50 terabytes, which, if stored for 100 days, would occupy about 5 petabytes. Cable Company discard most of their network data at the end of the day because they lack the capacity to store it. However, regulators have asked most CSPs and cable operators to store call detail records and associated usage data. For a 100-million-subscriber CSP, the CDRs could easily exceed 5 billion records a day.

2.2.2 Velocity*:*

There are two aspects to velocity, one representing the throughput of data and the other representing latency. Let us start with throughput, which represents the data moving in the pipes. The amount of global mobile data is growing at a 78 percent compounded growth rate and is expected to reach 10.8 Exabyte per month in 20164 as consumers share more pictures and videos. To analyze this data, the corporate analytics infrastructure is seeking bigger pipes and massively parallel processing.

Latency is the other measure of velocity. Analytics used to be a "store and report" environment where reporting typically contained data as of yesterday - popularly represented as "D-1." Now, the analytics is increasingly being embedded in business processes using data-in-motion with reduced latency. For example, Turn (*www.turn.com*) is conducting its analytics in 10 milliseconds to place advertisements in online advertising platforms.

2.2.3 Variety:

In the 1990s, as Data Warehouse technology was rapidly introduced, the initial push was to create meta-models to represent all the data in one standard format. The data was compiled from a variety of sources and transformed using ETL (*E*xtract, *T*ransform, *L*oad) or ELT (*E*xtract the data and *L*oad it in the warehouse, then *T*ransform it inside the warehouse). The basic premise was narrow variety and structured content. Big Data has significantly expanded our horizons, enabled by new data integration and analytics technologies. A number of call center analytics solutions are seeking analysis of call center conversations and their correlation with emails, trouble tickets, and social media blogs. The source data includes unstructured text, sound, and video in addition to structured data. A number of applications are gathering data from emails, documents, or blogs.

2.2.4 Veracity:

Unlike carefully governed internal data, most Big Data comes from sources outside our control and therefore suffers from significant correctness or accuracy problems. Veracity represents both the credibility of the data source as well as the suitability of the data for the target audience.

3. Why put big data and analytics together now?

Big data provides gigantic statistical samples, which enhance analytic tool results. Most tools designed for data mining or statistical analysis tend to be optimized for large data sets. In fact, the general rule is that the larger the data sample, the more accurate are the statistics and other products of the analysis. Instead of using mining and statistical tools, many users generate or hand-code complex SQL, which parses big data in search of just the right customer segment, churn profile, or excessive operational cost. The newest generation of data visualization tools and in-database analytic functions likewise operate on big data.

Analytic tools and databases can now handle big data. They can also execute big queries and parse tables in record time. Recent generations of vendor tools and platforms have lifted us onto a new plateau of performance that is very compelling for applications involving big data.

The economics of analytics is now more embraceable than ever. This is due to a precipitous drop in the cost of data storage and processing bandwidth. The fact that tools and platforms for big data analytics are relatively affordable is significant because big data is not just for big business.

Many small-to-midsize businesses (especially those deep into digital processes for sales, customer interactions, or supply chain) also need to manage and leverage big data.

There's a lot to learn from messy data, as long as it's big. Most modern tools and techniques for advanced analytics and big data are very tolerant of raw source data, with its transactional schema, non-standard data, and poor-quality data. That's a good thing, because discovery and predictive analytics depend on lots of details—even questionable data. For example, analytic applications for fraud detection often depend on outliers and non-standard data as indications of fraud. So, be careful:

Big data is a special asset that merits leverage. That's the real point of big data analytics. The new technologies and new best practices are fascinating, even mesmerizing, and there's a certain macho coolness to working with dozens of terabytes. But don't do it for the technology. Put big data and discovery analytics together for the new insights they give the business.

Analytics based on large data samples reveals and leverages business change. The recession has accelerated the already quickening pace of business. The recovery, though welcome, brings even more change. In fact, the average business has changed beyond all recognition because of the recent economic recession and recovery. The change has not gone unnoticed. Businesspeople now share a wholesale recognition that they must explore change just to understand the new state of the business.

Analytic platforms today handle big data better than ever. Big data is an enterprise asset that yields actionable business insights. Preparation of big data for advanced analytics rarely follows the same best practices we associate with mainstream data warehousing, reporting, and OLAP. Even more compelling, however, is the prospect of discovering problems that need fixing (such as new forms of customer churn and competitive pressure) and opportunities that merit leverage (such as new customer segments and sales prospects).

4. Techniques For Analyzing Big Data

When you use SQL queries to look up financial numbers or OLAP tools to generate sales forecasts, you generally know what kind of data you have and what it can tell you. Revenue, geography and time all relate to each other in predictable ways. You don't necessarily know what the answers are but you do know how the various elements of the data set relate to each other. BI users often run standard reports from structured databases that have been carefully modeled to leverage these relationships.

Big data analysis involves making "sense" out of large volumes of varied data that in its raw form lacks a data model to define what each element means in the context of the others. There are several new issues you should consider as you embark on this new type of analysis:

4.1 Discovery

In many cases you don't really know what you have and how different data sets relate to each other. You must figure it out through a process of exploration and discovery.

4.2 Flexible Capacity

Because of the iterative nature of big data analysis, be prepared to spend more time and utilize more resources to solve problems.

4.3 Mining and Predicting

Big data analysis is not black and white. You don't always know how the various data elements relate to each other. As you mine the data to discover patterns and relationships, predictive analytics can yield the insights that you seek.

4.4 Decision Management

Consider the transaction volume and velocity. If you are using big data analytics to drive many operational decisions (such as personalizing a web site or prompting call center agents about the habits and activities of consumers) then you need to consider how to automate and optimize the implementation of all those actions.

For example you may have no idea whether or not social data sheds light on sales trends. The challenge comes with figuring out which data elements relate to which other data elements, and in what capacity. The process of discovery not only involves exploring the data to understand how you can use it but also determining how it relates to your traditional enterprise data.

New types of inquiry entail not only what happened, but why. For example, a key metric for many companies is customer churn. It's fairly easy to quantify churn. But why does it happen? Studying call data records, customer support inquiries, social media commentary, and other customer feedback can all help explain why customers defects. Similar approaches can be used with other types of data and in other situations. Why did sales fall in a given store? Why do certain patients survive longer than others? The trick is to find the right data, discover the hidden relationships, and analyze it correctly.

5. Big Data Use Cases

This section includes a few use cases that demonstrate the potential of big data analytics within various business domains.

5.1 Machine-Generated Data

As the "Internet of Things" grows steadily each year, researchers predict that the amount of data generated by machines will one day outstrip the amount of data generated by humans. Machina Research, a UK-based research firm, believes there will be 12.5 billion "smart" connected devices excluding phones, PCs and tablets in the world in 2020, up from 1.3 billion today. Equipment sensors are prevalent in heavy machinery, automobiles, assembly lines, electric grids, computer equipment, and many other domains. And that's just the beginning, as more and more devices are manufactured with sensors that monitor their own operation and log the results for troubleshooting and analysis. For example, manufacturing companies commonly embed sensors in their machinery to monitor usage patterns, predict maintenance problems, and enhance build quality. Even consumer devices such as bicycles, washing machines, and thermostats are part of this machine-to-machine (M2M) communications phenomenon.

Studying these data streams allows them to improve their products and devise more accurate service cycles. Electronic sensors not only monitor mechanical and atmospheric conditions, but also the biometrics of the human body. In health care there is a huge opportunity not only to improve patient outcomes but also to monitor trends in health care diagnoses, treatments, and claims to make better clinical and administrative decisions. The opportunities become even more compelling once data is analyzed in aggregate form. If a thousand sensors reveal a pattern of equipment failure, or a thousand cardiac monitors show a correlation between biometric levels and adverse reactions, then we can begin to turn trends into predictions and ultimately use big data to take corrective or preemptive action.

5.2 Online Reservations

If you were running an online travel booking website, there are lots of interesting things you could do with your data to better understand your users. For example, when consumers book air travel, does the time that they booked a ticket have any bearing on how much money they spent? Perhaps holiday bargain seekers log on at night, while corporate travelers book flights early in the morning. What are the margins associated with each type of travel, and how do you discover the patterns of usage?

You might start by sorting through log files to determine when people started, ended, or completed a booking. You could also examine several related factors. For example, did they sort by price or by travel duration? Did they express airline preferences? Did each type of buyer prefer flights during the day or at night? How many different flight options did they consider? How many visits to your site did they make before booking, and how long did they spend contemplating their purchases?

Answering these questions requires comparing and analyzing lots of web log data that is constantly being generated. Most of that information is not very important in isolation, but when you analyze it in aggregate you can beg into see the patterns and discern important trends. Using HDFS to acquire the original data and MapReduce to process it enables you to correlate variables such as time of login, number of mouse clicks, duration of each session, and which queues or pages preceded a purchase. Then you can add this answer set to your data warehouse for additional analysis.

5.3 Multi-Channel Marketing And Sentiment Analysis

Today's retailers must contend with a multitude of overlapping touch-points including social, digital, direct, in store, mobile, and call center. Market leaders gain insight by analyzing transaction histories and web-

behavior, as well as by concatenating data from external environments such as social media, demographics, and finance.

Forward looking companies combine social media feeds, customer demographic information, psychographic data (values, attitudes, interests, or lifestyles), purchase data, and network usage data to paint a complete picture of each customer's behavior, likes, and dislikes. Harnessing this information helps retailers to understand each potential buyer as a "market of one" and to present personalized, tailored offerings to individual customers. To achieve this level of personalization, retailers must find answers hidden in massive amounts of data about customers, spending histories, inventory, pricing, marketing campaigns, and other promotions. By analyzing this data they can better understand the factors that trigger desired behavior in various segments and channels. The data also reveals the factors that impact customer loyalty and retention, such as ease of use, value for money, and the effect of customer rewards programs. Customer churn is a major problem with retailers and the right analytic solution can help them uncover the reasons behind the churn. By examining the records about customers who have defected, you can detect patterns and then search for the early signs of those same patterns in current customers. Customer interactions can be captured, aggregated, analyzed, and correlated with other KPIs like Net Promoter Scores, to develop insights into customer behavior. For example, analyzing Twitter feeds and Facebook posts can reveal quality of service issues within specific regions or customer groups.

While traditional segmentation strategies grouped customers based on channel-specific purchase cycles, value is increasingly defined by how well a company can manage interactions across any channel including mobile, web, call center, IVR, dealers, and retail outlets. Sentiment data can tell you if a particular individual likes or doesn't like your company and product. When you combine this information with other e-business data, you can also tell if they are a big spending customer, a regular customer, or not yet a customer. You can also see if they are influencing other people in your customer database.

When you combine all this data and analyze it appropriately you can uncover hidden relationships that you would otherwise not be aware of. You can determine behavior patterns and even predict what others might do in a similar situation.

6. Big Data Analysis Requirements

In the previous section, Techniques for Analyzing Big Data, we discussed some of methods you can use to find meaning and discover hidden relationships in big data. Here are three significant requirements for conducting these inquiries in an expedient way:

- Minimize data movement
- Use existing skills
- Attend to data security

Minimizing data movement is all about conserving computing resources. In traditional analysis scenarios, data is brought to the computer, processed, and then sent to the next destination. For example, production data might be extracted from e-business systems, transformed into a relational data type, and loaded into an operational data store structured for reporting. But as the volume of data grows, this type of ETL architecture becomes increasingly less efficient. There's just too much data to move around. It makes more sense to store and process the data in the same place.

With new data and new data sources comes the need to acquire new skills. Sometimes the existing skill set will determine where analysis can and should be done. When the requisite skills are lacking, a combination of training, hiring and new tools will address the problem. Since most organizations have more people who can analyze data using SQL than using MapReduce, it is important to be able to support both types of processing.

Data security is essential for many corporate applications. Data warehouse users are accustomed not only to carefully defined metrics and dimensions and attributes, but also to a reliable set of administration policies and security controls. These rigorous processes are often lacking with unstructured data sources and open source analysis tools. Pay attention to the security and data governance requirements of each analysis project and make sure that the tools you are using can accommodate those requirements.

6.1 Big Data Analysis Platforms And Tools

6.1.1 Hadoop: Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines, or racks of machines) are commonplace and thus should be automatically handled in software by the framework[15]. Operating System: Windows, Linux, OS X.

6.1.2 Mapreduce: Originally developed by Google, the MapReduce website describe it as "a programming model and software framework for writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes." It's used by Hadoop, as well as many other data processing applications. Operating System: OS Independent.

6.1.3 Gridgain: GridGrain offers an alternative to Hadoop's MapReduce that is compatible with the Hadoop Distributed File System. It offers in-memory processing for fast analysis of real-time data. You can download the open source version from GitHub or purchase a commercially supported version from the link above. Operating System: Windows, Linux, OS X.

6.1.4 HPCC: Developed by LexisNexis Risk Solutions, HPCC is short for "high performance computing cluster." It claims to offer superior performance to Hadoop. Both free community versions and paid enterprise versions are available. Operating System: Linux.

6.1.5 Storm

Now owned by Twitter, Storm offers distributed real-time computation capabilities and is often described as the "Hadoop of real time." It's highly scalable, robust, fault-tolerant and works with nearly all programming languages. Operating System: Linux.

6.1.6 Cassandra

Originally developed by Facebook, this NoSQL database is now managed by the Apache Foundation. It's used by many organizations with large, active datasets, including Netflix, Twitter, Urban Airship, Constant Contact, Reddit, Cisco and Digg. Commercial support and services are available throughthird-party vendors. Operating System: OS Independent.

6.1.3 Hbase

Another Apache project, HBase is the non-relational data store for Hadoop. Features include linear and modular scalability, strictly consistent reads and writes, automatic failover support and much more. Operating System: OS Independent.

7. Conclusion

Organizations in every industry are trying to make sense of the massive influx of big data, as well as to develop analytic platforms that can synthesize traditional structured data with semi-structured and unstructured sources of information.

When properly captured and analyzed, big data can provide unique insights into market trends, equipment failures, buying patterns, maintenance cycles and many other business issues, lowering costs, and enabling more targeted business decisions.

To obtain value from big data, a cohesive set of solutions is required for capturing, processing, and analyzing the data, from acquiring the data and discovering new insights to making repeatable decisions and scaling the associated information systems.

References

[1] IDC,2011. Big Data: What It Is and Why You Should Care. IDC. Available at:http://sites.amd.com/us/Document/IDC_AMD_Big_Data_Whitepaper.pdf

[2] Ericsson, 2013. Ericsson Mobility Report. [online] Ericsson. Available at: http://www.ericsson.com/res/docs/2013/ericsson-mobility-report-june-2013.pdf

[3] IDC Press Release, 2012. Worldwide Big Data Technology and Services 2012-2016 Forecast. [online] IDC. Available at: http://www.idc.com/getdoc.jsp?containerId=prUS23355112

[4] Snijders, C.; Matzat, U.; Reips, U.-D. (2012). "'Big Data': Big gaps of knowledge in the field of Internet". International Journal of Internet Science 7: 1–5.

[5] Ibrahim; TargioHashem, Abaker; Yaqoob, Ibrar; BadrulAnuar, Nor; Mokhtar, Salimah; Gani, Abdullah; Ullah Khan, Samee (2015). "big data" on cloud computing: Review and open research issues". Information Systems 47: 98–115.doi:10.1016/j.is.2014.07.006

[6] Laney, Douglas. "3D Data Management: Controlling Data Volume,Velocity and Variety" (PDF). Gartner.Retrieved 6 February 2001.

[7] Beyer, Mark. "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data". Gartner. Archived from the original on 10 July 2011.Retrieved 13 July 2011.

[8] Laney, Douglas. "The Importance of 'Big Data': A Definition". Gartner.Retrieved21 June 2012.

[9] "What is Big Data?". Villanova University.

[10] De Mauro, Andrea; Greco, Marco; Grimaldi, Michele (2015). "What is big data? A consensual definition and a review of key research topics". AIP Conference Proceedings 1644: 97–104. doi:10.1063/1.4907823.

[11] Hilbert, M. Big Data for Development: A Review of Promises and Challenges. Development Policy Review accessible at martinhilbert.net/big-data-for-development

[12] Delort P., Big data Paris 2013 http://www.andsi.fr/tag/dsi-big-data/

[13] Delort P., Big Data car Low-Density Data ? La faibledensité en information commefacteur discriminanthttp://lecercle.lesechos.fr/entrepreneur/tendances-innovation/221169222/big-data-low-density-data-faible-densite-information-com

[14] "What Data Says About Us," Fortune, September 24, 2012, p. 163.

[15] "Welcome to Apache™ Hadoop®!". *hadoop.apache.org*. Retrieved2015-09-20.

Biographical Notes

Mr. Surajit Mohanty, Assistant professor in Dept. of CSE, DRIEMS, Tangi, Cuttack. He has 11 year industry and teaching experience. He completed his M-tech in the year 2010.His area of interest is in SAP, ERP, Data mining and BIG data analysis. He has published 12 paper in national and international journal.

Mr. Rajeev Agarwal, Associate professor in the Dept. of CSE, DRIEMS, Tangi, Cuttack. He has 14 years of teaching experience. He completed his M-Tech in the year 2010. His area of interest is in computer Architecture, Data mining. He has published 6 papers in national and international journals.

Mr. Shekharesh Barik, Assistant professor in the Dept. of CSE, DRIEMS, Tangi, Cuttack. He has 12 years of industry and teaching experience. He completed his M-Tech in the year 2010. His area of interest is in image processing, Algorithm Analysis and Design, Data mining. He has published 4 papers in national and international journals.