Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

A Smart Start on Big Data

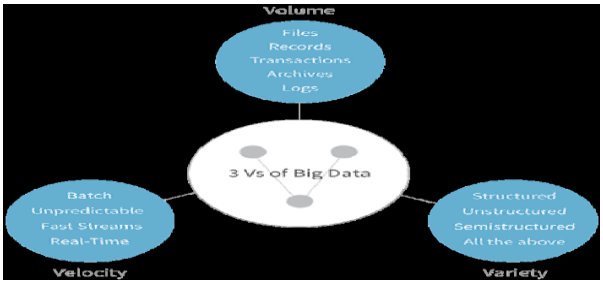
Ayasha,
Assistant Professor Department of Computer Science MGR College, Hosur, Tamil Nadu, India

Abstract: *Today the term big data draws a lot of attention, but behind the propaganda there's a simple story. For decades, companies have been making business decisions based on transactional data stored in relational databases. Beyond that critical data, however, is a potential treasure trove of non-traditional, less structured data: web logs, social media, email, sensors, and photographs that can be mined for useful information. Decreases in the cost of both storage and compute power have made it feasible to collect this data - which would have been thrown away only a few years ago. As a result, more and more companies are looking to include non-traditional yet potentially very valuable data with their traditional enterprise data in their business intelligence analysis. This paper presents an overview of big data's content, scope, methods, advantages and challenges.*

I. INTRODUCTION

Today is the era of Google. The thing which is unknown for us, we Google it. And in fractions of seconds we get the number of links as a result. This would be the better example for the processing of Big Data. This Big Data is not any different thing than out regular term data. Just big is a keyword used with the data to identify the collected datasets due to their large size and complexity? We cannot manage them with our current methodologies or data mining software tools. Another example, the first strike of Anna Hajare triggered number of tweets within 2 hours. Among all these tweets, the special comments that generated the most discussions actually revealed the public interests. Such online discussions provide a new means to sense the public interests and generate feedback in real-time, and are mostly appealing compared to generic media, such as radio or TV broadcasting. This example demonstrates the rise of Big Data applications. The data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within a tolerable time.

A. Defining Big Data



Volume: Machine-generated data is produced in much larger quantities than non-traditional data. For instance, a single jet engine can generate 10TB of data in 30 minutes. With more than 25,000 airline flights per day, the daily volume of just this single data source runs into the Petabytes. Smart meters and heavy industrial equipment like oil refineries and drilling rigs generate similar data volumes, compounding the problem.

Velocity: Social media data streams – while not as massive as machine-generated data – produce a large influx of opinions and relationships valuable to customer relationship management. Even at 140 characters per tweet, the high velocity (or frequency) of Twitter data ensures large volumes (over 8 TB per day).

Value: The economic value of different data varies significantly. Typically there is good information hidden amongst a larger body of non-traditional data; the challenge is identifying what is valuable and then transforming and extracting that data for analysis.
To make the most of big data, enterprises must evolve their IT infrastructures to handle these new high-volume, high-velocity, high-variety sources of data and integrate them with the pre-existing enterprise data to be analyzed.

We consider two additional dimensions when thinking about big data:

Variability: In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Is something trending in social media? Daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data involved.

Complexity: Today's data comes from multiple sources. And it is still an undertaking to link, match, cleanse and transform data across systems. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control.

B. Why big data should matter to you?

The real issue is not that you are acquiring large amounts of data. It's what you do with the data that counts. The hopeful vision is that organizations will be able to take data from any source, harness relevant data and analyze it to find answers that enable

1) Cost reductions,
2) Time reductions,
3) New product development and optimized offerings, and
4) Smarter business decision making. For instance, by combining big data and high-powered analytics,

It is possible to:

- Determine root causes of failures, issues and defects in near-real time, potentially saving billions of dollars annually.
- Optimize routes for many thousands of package delivery vehicles while they are on the road.
- Analyze millions of SKUs to determine prices that maximize profit and clear inventory.
- Generate retail coupons at the point of sale based on the customer's current and past purchases.

- Send tailored recommendations to mobile devices while customers are in the right area to take advantage of offers.
- Recalculate entire risk portfolios in minutes.
- Quickly identify customers who matter the most.
- Use click stream analysis and data mining to detect fraudulent behavior.

II. BIG DATA AND DATA MINING

The Big Data is nothing but a data, available at heterogeneous, autonomous sources, in extreme large amount, which get updated in fractions of seconds. For example, the data stored at the server of Face book, as most of us, daily use the Face book; we upload various types of information, upload photos. All the data get stored at the data warehouses at the Server of Face book. This data is nothing but the big data, which is so called due to its complexity. Also another example is storage of photos at Flicker. These are the good real-time examples of the Big Data. Another best example of big data would be, the readings taken from an electronic microscope of the universe. Now the term Data Mining, Finding for the exact useful information or knowledge from the collected data, for future actions, is nothing but the data mining. So, collectively, the term Big Data Mining is a close up view, with lots of detail information of a Big Data with lots of information

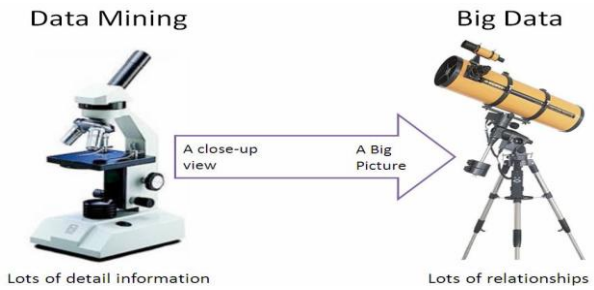


Figure 1: Data Mining with Big Data

Is Big Data a Volume or a Technology?

While the term may seem to reference the volume of data, that isn't always the case. The term big data, especially when used by vendors, may refer to the technology (which includes tools and processes) that an organization requires handling the large amounts of data and storage facilities. The term big data is believed to have originated with Web search companies who needed to query very large distributed aggregations of loosely-structured data.

An Example of Big Data

An example of big data might be petabytes (1,024 terabytes) or exabytes (1,024 petabytes) of data consisting of billions to trillions of records of millions of people—all from different sources (e.g. Web, sales, customer contact center, social media, mobile data and so on). The data is typically loosely structured data that is often incomplete and inaccessible.

A. Building a Big Data Platform

1 Infrastructure Requirements

The requirements in a big data infrastructure span
*) Data acquisition,
*) Data organization
*) Data analysis.

Platform Infrastructure:

The big data "platform" is typically the collection of functions that comprise high-performance processing of big data. The platform includes capabilities to integrate, manage, and apply sophisticated computational processing to the data. Typically, big data platforms include a Hadoop (or similar open-source project) foundation. Hadoop was designed and built to optimize complex manipulation of large amounts of data while vastly exceeding the price/performance of traditional databases. Hadoop is a unified storage and processing environment that is highly scalable to large and complex data volumes.

Hadoop:

Hadoop was created by Doug Cutting and Mike Cafarella in 2005. Doug Cutting, who was working at Yahoo! at thetime, named it after his son's toy elephant. It was originally Developed to support distribution for the Nutch searchengine project. Hadoop is open-source software thatenables reliable, scalable, distributed computing on clusters of inexpensive servers Hadoop is:

***Reliable*:** The software is fault to learnt, it expects and handles hardware and software failures
***Scalable*:** Designed for massive scale of processors, memory, and local attached storage

***Distributed*:** Handles replication. Offers massively parallel programming model, Map Reduce Hadoop is an Open Source implementation of a large-scale batch processing system.

Hadoop is particularly useful when:

- Unstructured data needs to be turned into structured data
- Queries can't be reasonably expressed using SQL
- Heavily recursive algorithms
- Complex but parallelizable algorithms needed such as geo-spatial analysis or genome sequencing
- Machine learning
- Data sets are too large to fit into database RAM, discs, or require too many cores (10's of TB up to PB)
- Data value does not justify expense of constant real-time availability, such as archives or special interest info, which can be moved to Hadoop and remain available at lower cost
- Results are not needed in real time
- Fault tolerance is critical
- Significant custom coding would be required to handle job scheduling

Note that in the example the data sources themselves are heterogeneous, involving more diverse unstructured and semi-structured data sets like emails, web logs, or images. These data sources are increasingly likely to be found outside of the company's firewall. The big companies adopting production-

class big data environments need faster and lower-cost ways to process large amounts of atypical data. Think of the computing horsepower needed by energy companies to process data streaming from smart meters, or by retailers tracking in-store smart phone navigation paths, or LinkedIn's reconciliation of millions of colleague recommendations.

III. APPLICATIONS

Big data has increased the demand of information management specialists in that Software AG, Oracle Corporation, IBM, Microsoft, SAP, EMC, HP and Dell have spent more than $15 billion on software firms specializing in data management and analytics. In 2010, this industry was worth more than $100 billion and was growing at almost 10 percent a year: about twice as fast as the software business as a whole.

Developed economies increasingly use data-intensive technologies. There are 4.6 billion mobile-phone subscriptions worldwide, and between 1 billion and 2 billion people accessing the internet. Between 1990 and 2005, more than 1 billion people worldwide entered the middle class, which means more people become more literate, which in turn leads to information growth. The world's effective capacity to exchange information through telecommunication networks was 281 petabytes in 1986, 471 petabytes in 1993, 2.2 exabytes in 2000, 65 exabytes in 2007and predictions put the amount of internet traffic at 667 exabytes annually by 2014. According to one estimate, one third of the globally stored information is in the form of alphanumeric text and still image data, which is the format most useful for most big data applications. This also shows the potential of yet unused data (i.e. in the form of video and audio content).

While many vendors offer off-the-shelf solutions for Big Data, experts recommend the development of in-house solutions custom-tailored to solve the company's problem at hand if the company has sufficient technical capabilities.

IV. 5 REASONS TO MOVE TO BIG DATA

1 Reason Why It Won't Be Easy You'll Manage Data Better

Many of today's data processing platforms let data scientists analyze, collect and sift through various types of data. While it does take some technical know-how to define how the data is collected and stored, many of today's big data and business intelligence tools let users sit in the driver's seat and work with data without going through too many complicated technical steps.

2. You'll Benefit from Speed, Capacity and Scalability of Cloud Storage

Organizations that want to utilize substantially large data sets should consider third-party cloud service providers, which can provide both the storage and the computing power necessary crunch data for a specific period.

Cloud storage presents two clear advantages. One, it lets companies analyze massive data sets without making a significant capital investment in hardware to host the data internally. Two, as internal IT departments recognize that big data hosting platforms require new skills and training, they find that a hosted model tends to abstract that complexity, enabling more immediate deployment of big data technology. This also lets developers build a sandbox environment that's preconfigured and ready to go without having to set up the necessary configurations from scratch.

3. Your End Users Can Visualize Data

While the business intelligence software market is relatively mature, a big data initiative is going to require next-level data visualization tools, which present BI data in easy-to-read charts, graphs and slideshows. Due to the vast quantities of data being examined, these applications must be able to offer processing engines that let end users query and manipulate information quickly—even in real time in some cases. Applications will also need adaptors that can connect to external sources for additional data sets.

4. Your Company Can Find New Business Opportunities

As big data analytics tools continue to mature, more users are realizing the competitive advantage to being a data-driven enterprise. The 2012 presidential election demonstrated this. Campaign managers in both the Democratic and Republican parties saw a critical need for information on voters and their specific interests; taking this info and addressing an issue through a customized email or flyer meant the potential to gain or sway a vote.Finally, big data use cases in about in retail, where the focus is on gaining insights by studying consumer behavior in online stores or physical shopping centers.

5. Your Data Analysis Methods, Capabilities Will Evolve

Data is no longer simply numbers in a database. Text, audio and video files can also provide valuable insight; the right tools can even recognize specific patterns based on predefined criteria. Much of this happens using natural language processing tools, which can prove vital to text mining, sentiment analysis, clinical language and name entity recognition efforts.

One example that highlights the use of audio analysis and big data comes from Matter Sight. This call center tool can match incoming caller to the appropriate customer agent by using predictive behavioral routing and other analytics technology. Matter Sight performs audio analysis to identify and score the calls based on specific criteria and then match customers with the best department to ensure the best experience. These advanced capabilities highlight some of the advancements we continue to see in unstructured data analysis and Big Data capabilities.

The Big Data Challenge: You'll Need New People

In addition to buying the right software, recruiting the right talent ranks among the most important investments an organization can make in its big data initiative. Having the right people in place will ensure that the right questions are

asked—and that the right insights are extracted from the data that's available. Keep in mind that data scientists, as many refer to those working with big data, are in short supply and are being quickly snapped up by top firms.

Every CIO wants to keep his finger on the pulse of innovations that can transform his company, enhance existing business models and identify potential revenue sources. Enabling this business transformation means adopting the right tools, hiring the right people and—most of all—convincing executive leadership to embrace new models for using existing and brand-new data.A successful big data initiative, then, can require a significant cultural transformation that's driven by the IT department. you need to rise to the challenge.

CONCLUSION

Big Data is going to continue growing during the next years, and each data scientist will have to manage much more Amount of data every year. This data is going to be more diverse, larger, and faster. We discussed some insights about the topic, and what we consider are the main concerns and the main challenges for the future. Big Data is becoming the new Final Frontier for scientific data research and for business applications. We are at the beginning of a new era where Big Data mining will help us to discover knowledge that no one has discovered before. Everybody is warmly invited top articipate in this intrepid journey.

FURURE REFERENCE

It is no secret that big data is influencing the IT industry like few technologies or trends have done so before. If analyzed effectively, these massive information caches can help companies improve their decision-making and take their businesses to another level. However, managing big data is a difficult endeavor, according to a recent report by Microsoft.
"Big data absolutely has the potential to change the way governments, organizations, and academic institutions conduct business and make discoveries, and its likely to change how everyone lives their day-to-day lives,"

References

[1] V. R. Borkar and M. J. Carey, “Big data technologies circa 2012,” in *COMAD*, 2012, pp. 12–14.
[2] R. Ramakrishnan, “Big data in 10 years,” in *IPDPS*, 2013, p. 887.
[3] V. R. Borkar, M. J. Carey, and C. Li, “Big data platforms: what’s next?” *ACM Crossroads*, vol. 19, no. 1, pp. 44–49, 2012.
[4] A Sustainable Future. Computing Community Consortium. Summer 2011.
[5] Using Data for Systemic Financial Risk Management. Mark Flood, H V Jagadish, Albert Kyle, Frank Olken, and Louiqa Raschid. Proc. Fifth Biennial Conf. Innovative Data Systems Research, Jan. 2011.
[6] The Emerging Big returns on big data,TCS-Big-Data-Global-Trend-Study-2013,mar 21,2013
[7] R. Ramakrishnan, “Cap and cloud data management,” *IEEE Computer*, vol. 45, no. 2, pp. 43–49, 2012.
[8] J. Shen, J. Fang, H. Sips, and A. L. Varbanescu, “Performance Traps in OpenCL for CPUs,” in *PDP 2013*, February 2013, pp. 38–45.