International Journal of Advance Research in Science and Engineering
Vol. No.4, Special Issue (01), September 2015
www.ijarse.com
IJARSE
ISSN 2319 - 8354

BIG DATA: BIG BOOST TO BIG TECH

Ms. Tosha Joshi

Department of Computer Applications, Christ College, Rajkot, Gujarat (India)

ABSTRACT

Data formation is occurring at a record rate. A staggering 2.9 billion by a billion bytes of data are created every day. According to IDC, the world's digital output soared from 180 Exabytes in 2006 to about 1,800 Exabytes in 2011. (One Exabyte equals one billion gigabytes.) Data volume is expected to reach 35,000 Exabytes by 2020, representing a 20-fold increase in the next 10 years. Typically complex and unstructured, big data presents a major technological challenge for companies that capture, analyse and exploit the information. Much of this data explosion is the result of a histrionic increase in devices located at the boundary of the network including embedded sensors, smartphones, and tablet computers. All of this data creates new opportunities to "extract more value" in healthcare, search, surveillance, finance, and many other areas. We are arriving in the age of "Big Data." organizations that are best able to make real-time business decisions using Big Data solutions will succeed, while those that are unable to hold and make use of this shift will increasingly find themselves at a competitive disadvantage in the market and face potential failure. This paper will mainly focus on the benefits, big lift to big technologies and also what user gets what he or she needs, and whether IT has the freedom to manage the aggregate of all big data services to best optimize their environment!

Keywords: ***Big Data, Dirty Data, Data Mining, Data Quality, Industry Vertical,***

I. INTRODUCTION

Big Data is a phrase that booms across all corners of the business. Big Data technologies describe a new generation of technologies and architectures, designed so organizations can economically extract value from large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis. Big data has arisen because digitalisation has accelerated the growth in data across every organisation, industry and economy. This world of Big Data requires a shift in computing architecture so that customers can handle both the data storage requirements and the heavy server processing required to analyse large volumes of data economically. Massive companies like Amazon and Wal-Mart as well as bodies such as the U.S. government and NASA are using Big Data to meet their business and/or strategic objectives. Big Data can also play a role for small or medium-sized companies and organizations that recognize the possibilities to capitalize upon the gains. Big Data in the enterprise should not live in a vacuum. It materializes from dozens of databases, applications, and external sources.

A) What is big data really?

A big data is your data. Any big data expert define as "The Three V's "– "volume, velocity and variety". Con-

cepts originally coined by Doug Laney in 2001 to refer the challenges of data management. In short, it's a lot of data formed very quickly in many different systems. This could involve customer transactional histories, production databases, web traffic logs, online videos, social media interactions, and so forth.

B) What is unique about big data?

Big Data is special because it represents both significant information - which can open new doors - and the way this information is analysed to help open those doors. The analysis goes hand-in-hand with the information, so in this sense "Big Data" represents a noun - "the data" - and a verb - "combing the data to find value."

II. HOW BIG DATA IS GOING TO FEED NINE BILLION PEOPLE BY 2020?

The bases of data growth that are driving Big Data technology investments vary widely. Some represent entirely new data sources, while others are a change in the "determination" of existing data produced.

Industry Digitization

- Media / Entertainment.
- Life Sciences
- Healthcare
- Transportation, logistics, retail and telecommunications.
- Video surveillance

Social media solutions like Facebook and Twitter are the newest new data sources. A number of new businesses are now building Big Data environments, based on scale-out clusters using power-efficient multicore processors like the AMD Opteron 4000 and 6000 Series platforms that leverage consumers' nearly continuous streams of data about themselves (e.g., likes, locations, opinions).

III. BIG BOOST TO TECH

Big data is a boon to market-research firms. The internet has widened the ways to reach and market to an audience because every page viewed and every click made are recorded. Marketing teams are pouring through the huge amounts of data available from search and social media leaders such as Google and Facebook. Large online businesses such as Amazon.com already use the customer data to suggest other products. However, the set-up costs and necessary infrastructure are problematic for smaller companies. An alternative solution is to outsource the data analysis and targeted marketing to specialists.

- Address critical IT skills
- Diminish the value of legacy system
- Create less noise
- Work every time
- Improve your data quality
- Validate current ROI metrics

IV. BIG BOOST TO MANUFACTURING

What has given Wal-Mart an edge is the way it pioneered the expansion of its electronic-data-interchange systems to connect suppliers. Suppliers can now view a link that shows when restocking is required rather than wait for an order from Wal-Mart. These vendor-managed-inventory techniques have been around since the 1980s. It's just that Wal-Mart has deployed it on a massive scale.

Since the start of the computer age, manufacturers have used data to help drive production quality and efficiency. Big data allows designers and manufacturers to share data quickly and cheaply and create simulations that test designs. The aerospace and car industries use big data for this purpose. Big data can help improve business management for it can help maximize cash flows.

V. WHERE ARE WE WITH BIG DATA?

A. *Data Quality:*

The situation is characteristic of data that is not "clean enough" to fully populate the requested fields for an analytics query. The value proposition provided to companies that have difficulty cleaning all of their data is that at least they can begin to receive some value from big data that will benefit them. In other words, clients are not in a 'yes or no' situation when it comes to having all of their data clean as a prerequisite before they can start using big data analytics.

B. Dirty Data:

Dirty data is also a challenge organizationally for companies. This is because it's hard to get data squeaky clean without knowing what it really should look like within the context of business. It is here where data cleaning can become a painstakingly manual task. There is also the difficulty of recruiting a business leader with enough organizational authority to assume responsibility for this "back office" task. At the end of the day, cleaning data can be hard to justify for ROI, because you have yet to see what clean data is going to deliver analytics and what the analytics will deliver to business.

VI. PUBLIC ROLE/INDUSTRY VERTICAL

With most hype-worthy technologies, there's usually something valid beneath the smoke and mirrors, and Big Data is no exception. New technologies have provided the technical tools to perform more rapid analyses on large data sets, and everything from storage to networks have advanced to the point where we can more rapidly move, process, and manipulate these data. While that's exciting and there are certainly some interesting innovations in this area. Big data is still a relatively new phenomenon. Sure, many companies are racing to capture and exploit as much data as they can, but that doesn't mean that every company is currently in a position to begin a big data initiative. The reality is that most companies are still struggling to find value from the "small data" and while they know what big data is, most are unsure about how to harness it and put it to use. Arguably, the technology is one small portion of the promise of Big Data. Big Data has shifted IT's focus to how it provides timely and accurate reporting.

VII. DATA MINING TOOLS

Taking anonymized data from the results of the 2015 KDnuggets Data Mining Software Poll, and performed association analysis the top 20 tools. The dataset consisted of 2759 votes, each for one or more tools. Version of Apriori algorithm is used in this paper to analyze the results. There are many ways to measure how significant is associations between two nominal or binary features, eg chi-square or T-test, but we use a simple measure we call "Boost", defined as

Boost (X & Y) = pct (X & Y) / (pct (X) * pct (Y))

where pct(X) is the percent of users who selected X.

Boost (X&Y) > 1 indicates that X&Y appear together more than expected if they were independent,

Boost=1 if X & Y appear with frequency expected if they are independent, and boost < 1 if X & Y appear together less than expected (negatively correlated)

Note that this measure is symmetric: Boost (X & Y) = Boost (Y & X)

Fig. 1 shows the map for the top 10 Data Mining tools. The lift values are displayed in their respective matrix positions and the colour gradient represents the degree of association from high to low.

If Boost is > 1.2 the square is reddish, if less than 0.8, bluish, else grey.

Spark and Hadoop have the highest association with a Boost =3.31, followed by Spark and Python (Boost=2.05).

We also note strong association between Excel and SQL, and Tableau and SQL.

The lowest associations were found between SAS base and KNIME (0.51), SAS base and RapidMiner (0.52), and KNIME and RapidMiner (0.56).

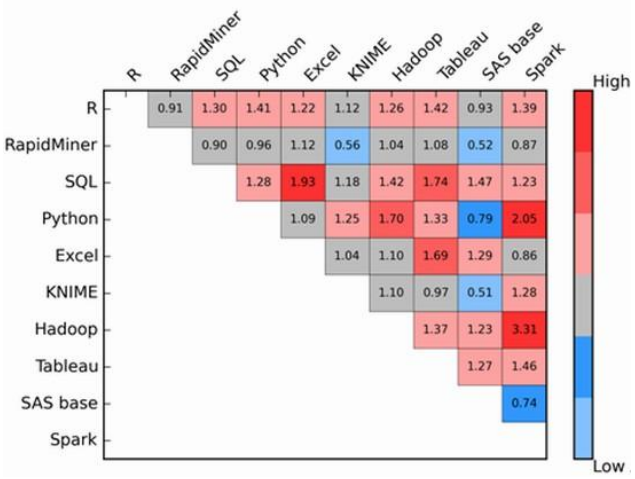

Fig 1: Association Matrix Map for Top 10 Most Popular Data Mining Tools

VIII. CONCLUSION

BIG DATA – THE ROAD AHEAD OF US

The biggest challenges does not seems to be the technology itself – as this is evolving much more rapidly than humans but rather how to make sure we have enough skills to make effective use of the technology at our disposal and make sense out of the data collected. The management of bid data, the intelligent use of large, heterogeneous data sets, is becoming increasingly important for competitions. It is affecting all sectors – industry and

academia but also the public sector. While the economy is exploring big data as new gold mine, politician are fighting over the problem of data capitalism, whereas science tackles the challenges and the likely consequences for technology, innovation and society. So provided big data is exploited in an open and transparent manner. In conclusion, the promise and potential of big data needs to be matched by a considered approach to collection, storage, licensing and use. Big data has a range of practical and commercial benefits to businesses but can be fraught with privacy and legal issues. Delivery of big data is not so far ahead of us.

REFERENCES

[1] http://www.techproresearch.com/downloads/research-big-data-trends-costs-payoffs-outcomes-staffing/
[2] http://www.ebusinessbook.nl/185
[3] iDc, 2011. Big Data: What it is and Why you Should care. iDc. [Accessed July 10, 2013].
[4] Analysis Mason, 2013. Big Data Analytics: how To generate Revenue and customer loyalty Using Real-time network Data. [online] Analysis Mason. [Accessed August 5, 2013].
[5] cloud Security Alliance Big Data Working group, 2013. Expanded Top Ten Big Data Security and Privacy challenges. [online] cloud Security Alliance Big Data Working group. [Accessed August 5, 2013].
[6] www.i-zdnet.com
[7] www.trendmicro.com\Big data
[8] www.techtarget.com
[9] http://www.fidelity.com.au/insights-centre/investment-articles/big-data-is-a-huge-opportunity