ISSN: 2277-9655
Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Study on Causes and Issues of Big-Data

Balaji Prabhu B V[*1], Arpitha S[2]
[*1,2] Department of ISE, KIT, Tiptur, India
balajitiptur@gmail.com

Abstract

An increasing amount of data is becoming available on the internet. Each and every one of us is constantly producing and releasing data about ourselves. We do this either by moving around passively - our behavior being registered by cameras or card usage - or by logging onto our PCs and surfing the net. The volumes of data make up what has been designated 'Big Data' -- where data about individuals, groups and periods of time are combined into bigger groups or longer periods of time. The amount of data in our world has been exploding. Companies capture trillions of bytes of information about their customers, suppliers, and operations, and millions of networked sensors are being embedded in the physical world in devices such as mobile phones and automobiles, sensing, creating, and communicating data. Multimedia and individuals with smart phones and on social network sites will continue to fuel exponential growth. Big data—large pools of data that can be captured, communicated, aggregated, stored, and analyzed—is now part of every sector and function of the global economy. Like other essential factors of production such as hard assets and human capital, it is increasingly the case that much of modern economic activity, innovation, and growth simply couldn't take place without data.

Keywords: Internet, Big Data, Sensors, Multimedia, Social Network.

Introduction

The data universe is expanding rapidly. Organizations of all sizes face the challenge of collecting and storing massive amounts of data to address the requirements of large unstructured repositories of primary data. Digital technologies are moving to denser media, photos have gone digital, video and medical imaging systems are using higher resolution, and advanced analytics require significantly more storage. Retaining information is critical for ongoing business operations. Data have become torrent flowing into every area of the global economy [1]. Companies churn out a burgeoning volume of transactional data, capturing trillions of bytes of information about their customers, suppliers, and operations. Millions of networked sensors are being embedded in the physical world in devices such as mobile phones, smart energy meters, automobiles, and industrial machines that sense, create, and communicate data in the age of internet of things[2]. Indeed, as companies and organizations go about their business and interact with individuals, they are generating a tremendous amount of digital exhaust data, i.e data that are created as by-product of other activities. Social media sites, Smartphone's, and other consumer devices including PCs and laptops have allowed billions of individual around the world to contribute to the amount of big data available. And the growing volume of multimedia contents has played a major role in the exponential growing in the amount of big data. Each second of high-definition video, for example, generates more than 2,000 times as many bytes as required to store a single page of text. In a digitized word, consumers going about their day-communicating, browsing, buying, sharing, searching—create their own enormous trails of data. An increasing amount of data is becoming available on the internet. Each and every one of us is constantly producing and releasing data about ourselves. The volumes of data make up what has been designated 'Big Data' where data about individuals, groups and periods of time are combined into bigger groups or longer periods of time.

Big Data

Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, and analyze. This definition is intentionally subjective and incorporates a moving definition of how big a database needs to be in order to be considered big data.

The global data supply reached 2.8 zettabytes (ZB) in 2012 - or 2.8 trillion GB - but just 0.5% of this is used for analysis [3]. Volumes of data are projected to reach 40ZB by 2020 or 5,247 GB per person, with emerging economies accounting for an increasingly large proportion of the world's total. The report also contained a warning on data security, with levels of protection

shown to be lagging behind the expansion in volume. In 2012 less than a fifth of the world's data was protected, despite 35% requiring such measures.

The study, carried out by the International Data Corporation (IDC) and sponsored by big data specialists EMC, is the sixth annual audit of global data inventories, and includes all such material gathered, created and replicated to date. IDC estimated that almost a quarter of data currently held could yield useful insights if properly tagged and analyzed, but this potential is still a long way from being achieved. Just 3% of all data is currently tagged and ready for manipulation, and only one sixth of this - 0.5% - is used for analysis. The gulf between availability and exploitation represents a significant opportunity for businesses worldwide, with global revenues surrounding the collection, storage, and analysis of big data set to reach $16.9bn in 2015 - a fivefold increase since 2010. "As the volume and complexity of data barraging businesses from all angles increases, IT organizations have a choice: they can either succumb to information-overload paralysis, or they can take steps to harness the tremendous potential teeming within all of those data streams", said Jeremy Burton, Executive Vice President, Product Operations and Marketing for EMC. The composition of data with analysis potential is also projected to change, with consumer images and voice calls predicted to disappear almost completely by 2020 in proportional terms, while the share generated from medical uses is set to increase fourfold over the same period. In 2012 just over a third of all data required some form of protection, but with companies and the public sector generating and holding increasing amounts of personal information, this proportion is expected to exceed 40% by 2020.

Currently the three main reasons for data to require protection are privacy (simple contact details such as an email address), custody (information whose leak could lead to identity theft) and confidentiality (private documents, contact lists), but high security data, such as bank details and medical records, is expected to overtake confidential information in volume over the next eight years. The report also highlights the concern that emerging markets, which will account for almost two thirds of the world's data by the end of this decade, typically have lower rates of data protection than the global average. In 2012, 53% of the world's data classified as requiring protection of some kind was found to have such measures in place, compared to 44% for India. Outside of Western Europe (59%), the US (58%), China (48%), and India, the rate is 49%. According to a study “India has 55.48 crore mobile users as per our India Mobile Landscape (IML) . More than 29.8 crore, about 54 per cent, of these device owners are in rural areas as compared to 25.6 crore in cities and towns,”. IML study finds that there are 14.32 crore internet users in the country[4]. "The number of unique Internet users in India, who access Internet from their desktop or laptop, smart TV or mobile data connections together stand at around 94.7 million. But when one adds the number of users who also access Internet through operators portals such as Airtel Live and Reliance R World, the number goes up to 143.2 million[5]". The study found 2.38 crore individuals access Internet from their mobile phones using a data connection such as GPRS or 3G. Out of this, 93 lakh access Internet only through mobile phones and around 77 per cent of these users are in rural area**s.**

Causes for Big Data

A. Big Data in Today’s Business and Technology Environment

- 2.7 Zetabytes of data exist in the digital universe today.
- 235 Terabytes of data has been collected by the U.S. Library of Congress in April 2011.
- IDC Estimates that by 2020, business transactions on the internet- business-to-business and business-to-consumer – will reach 450 billion per day.
- Facebook stores, accesses, and analyzes 30+ Petabytes of user generated data.
- Akamai analyzes 75 million events per day to better target advertisements.
- Walmart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data.
- More than 5 billion people are calling, texting, tweeting and browsing on mobile phones worldwide.
- The largest AT&T database boasts titles including the largest volume of data in one unique database (312 terabytes) and the second largest number of rows in a unique database (1.9 trillion), which comprises AT&T’s extensive calling records.

B. The Rapid Growth of Unstructured Data

- YouTube users upload 48 hours of new video every minute of the day.
- 571 new websites are created every minute of the day.
- Brands and organizations on Facebook receive 34,722 Likes every minute of the day.
- 100 terabytes of data uploaded daily to Facebook.

ISSN: 2277-9655
Impact Factor: 1.852

- According to Twitter's own research in early 2012, it sees roughly 175 million tweets every day, and has more than 465 million accounts.
- 30 Billion pieces of content shared on Facebook every month.
- Data production will be 44 times greater in 2020 than it was in 2009.

C. Data Stored in Goggle

Google search crawler uses 850 TB of information (1 TB = 1024 GB), so that's the amount of raw data from the web. Google Analytics uses 220 TB stored in two tables: 200 TB for the raw data and 20 TB for the summaries. Google Earth uses 70.5 TB: 70 TB for the raw imagery and 500 GB for the index data. The second table "is relatively small (˜500 GB), but it must serve tens of thousands of queries per second per datacenter with low latency[4]". Google Base uses 2 TB and Orkut only 9 TB of data. If we take into account that all this information is compressed (for example, the crawled data has compression rate of 11%, so 800 TB become 88 TB), Google uses for all the services mentioned before 220 TB. It's also interesting to note that the size of the raw imagery from Google Earth is almost equal to the size of the compressed web pages crawled by Google [6].

Issues with Big Data

A. Big Data & Real Business Issues

- According to estimates, the volume of business data worldwide, across all companies, doubles every 1.2 years [7].
- Poor data can cost businesses 20%–35% of their operating revenue [8].
- Bad data or poor data quality costs US businesses $600 billion annually [8].
- According to execs, the influx of data is putting a strain on IT infrastructure. 55 percent of respondents reporting a slowdown of IT systems and 47 percent citing data security problems, according to a global survey from Avanade [9].
- Executives at small companies (fewer than 1,000 employees) are nearly 10 percent more likely to view data as a strategic differentiator than their counterparts at large enterprises [9].
- Three-quarters of decision-makers (76 per cent) surveyed anticipate significant impacts in the domain of storage systems as a result of the "Big Data" phenomenon [10].
- A quarter of decision-makers surveyed predict that data volumes in their companies will rise by more than 60 per cent by the end of 2014, with the average of all respondents anticipating a growth of no less than 42 per cent [10].

These numbers, though limited to a certain set of users, still highlights on some interesting growth trends specially the adoption of the medium in rural areas.

Despite considering only web users, the 2013 ComScore report[11] shared some interesting facts about Indians such as:

1. With 25% time being spent on social networking [12], the activity grabs the largest share of PC screen time in the country. Besides this Email and Internet messaging hold a significant time too.
2. Video consumption grows by 27% and with more than 31 million viewers watching videos on YouTube [13], the social networking video site remains the number one destination for users.
3. Facebook is the most popular social networking site in the country followed by LinkedIn and Twitter. Besides this, new social networking sites like Tumblr and Pinterest are the fastest growing networking sites in the country.

B. Security and Privacy Issues of Big-Data

Security and privacy issues are magnified by velocity, volume, and verity of big data, such as large-scale cloud infrastructure, diversity of data sources and formats, streaming nature of data acquisition and high volume inter-cloud migration

. The top ten big data specific security and privacy challenges are listed below

1. Secure computation in distributed programming frameworks
2. Secure best practice for non-relational data stores.
3. Secure data storage and transaction logs
4. End-point input validation/filtering
5. Real-time security/compliance monitoring
6. Scalable and composable privacy-preserving data mining and analytics
7. Cryptographically enforced access control and secure communication
8. Granular access control
9. Granular audit
10. Data provenance

Conclusion

Big data is changing the world. Practically everything we do can be recorded. The potential to improve public health, win elections, map the human genome, and cut down on wasteful processes is only the beginning. To borrow Cotton Inc.'s tag line, big data really is the "fabric of our lives." Whoever explores it more deeply and aggressively first will have that much greater an insight into its commercial, social, and

ISSN: 2277-9655
Impact Factor: 1.852

scientific potential and will be able to make decisions that change the course of our lives. Whoever hesitates will be left behind.

If big data is to provide its promised value, it can't wait for experts. It can't take weeks to generate a report, when data from the Web and social media is constantly changing. We can't spend all of our time and effort capturing, storing, and cleansing data, without thinking about that critical "last mile," where the user interrogates the data. And we can't neglect that much of the value of unstructured big data from new sources will come from correlation with the standardized, structured enterprise data businesses have carefully been collecting and managing for decades. If big data is truly to be liberated from bottlenecks, it must be exposed and explored in an intuitive, user-friendly way. Sophisticated, yet easy-to-use methods are required to harness big data's full potential, for every user in every organization. CITO Research has determined that Business Discovery, especially its ability to simultaneously query real-time and historical databases, will play a major role in delivering big data in a way that is useful to everyone.

References

[1] "A special report on managing information: Data, data everywhere," The Economist, February 25, 2010; and special issue on "Dealing with data," Science, February 11, 2011.

[2] Michael Chui, Markus Löffler, and Roger Roberts, "The Internet of Things," McKinsey Quarterly, March 2010.

[3] New IDC Digital Universe study, "Extracting Value from Chaos" (sponsored by EMC)

[4] Study by research firm Juxt: IML 2013 study

[5] http://ibnlive.in.com/news/india-has-5548-crore-mobile-owners-1432-crore-internet-users/420444 11.html?utm_source=ref_article

[6] Paper about BigTable.

[7] eBAY study: How to build trust and improve the shopping experience -May 08, 2012

*[8] Fathom Blog :News & analysis on digital marketing & analytics- "Big Data" Facts and Statistics That Will Shock You By **Chad Luckie** | May 8, 2012*

[9] Global Survey: The Business Impact of Big Data November 2010

[10]"Big Data"survey: cloud computing and collaboration drive up data volume I German enterprises-11 July 2012

[11]Indian Mobile Landsape 2013- by PRASANT NAIDU on SEPTEMBER 11, 2013.

[12]State Of The Media: The Social Media Report Q3 2011.

[13]India Digital Future in Focus 2013 :Key Insights and Digital Trends Shaping the Indian Online Space.