

e-PGPathshala

Subject : Computer Science

Paper: Data Analytics

**Module: Big Data approach to Analytics & Web
– the actual big data**

Module No: CS/DA/2

Quadrant 1 – e-text

1.1 Introduction

The second module on data analytics is about big data approach to analytics and provides an overview of the data available over the web, which actually is big enough in terms of varieties, uncertainties and velocity, thereby becoming the actual big data.

1.2 Learning Objectives

- To understand how big data is different from traditional data environment
- To learn big data approach to analytics
- To know about potential big data domains for data analytics
- To know the importance of data offered by web
- To understand research on web data and its uses

1.3 How big Data is different?

The following are the factors that differentiate big data with respect to traditional data:

- Majority of the data that is voluminous are typically generated automatically by machine. For example, the data indexed by crawlers of various search engines.
- Big data is typically an entirely new source of data
- Not designed to be friendly
- Can be messy and ugly(junk filled data)
- No standards

1.3.1 The Big Data Approach to Analytics is Different

Traditionally, the business expected that data would be used to answer questions about what to do and when to do it. Data was often integrated as fields into general-purpose business applications. With the advent of big data, this is changing. Now, the developments of applications are being designed specifically to take advantage of the unique characteristics of big data.

Traditional Analytics	Big Data Analytics
It is structured and repeatable in nature	Iterative and exploratory in nature
Structure is built to store data	Data itself is a structure
Business users determine the questions which shall be answered by building systems by IT experts	IT team and data experts delivers the data on flexible platform for any exploration and querying by the business users

Big Data offers major improvements over its predecessor in analytics, traditional business intelligence (BI) which has always been top-down, putting data in the hands of executives and managers who are looking to track their businesses on the big-picture level. Big Data, on the other hand, is bottom-up. When properly controlled and exploited, it empowers business end-users in the trenches, enabling them to carry out in-depth analysis to inform real-time decision-making.

While the scope of traditional BI is limited to structured data that can be stuffed into columns and rows on a data warehouse, the fact is that over 90% of today's data is unstructured. BI could never have anticipated the multitude of images, MP3 files, videos and social media snippets that companies would contend with in the Big Data era, but that's the reality of business today. Traditional BI succumbs before forward-looking businesses that are desperate to tame and gain competitive advantage from unstructured business data floating around within and beyond their enterprise. Thus big data analytics provides a platform for the business users who can reconfigure and reshape the data as per their convenience. Figure 1 (a) illustrates that in a traditional analytics we start with a hypothesis, perform capacity constrained down sampling and test the hypothesis against selected data. But, in big data analytics we start exploring the correlations among the data. Since the volume and rate at which data arrives are too high correlations among the data are exploited for analyzing the data when it arrives as illustrated in figure 1 (b).

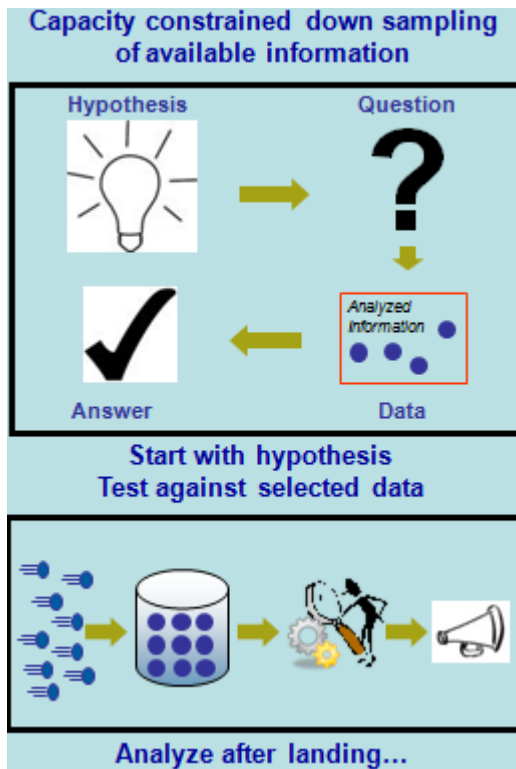


Figure 1 (a) Traditional Analytics

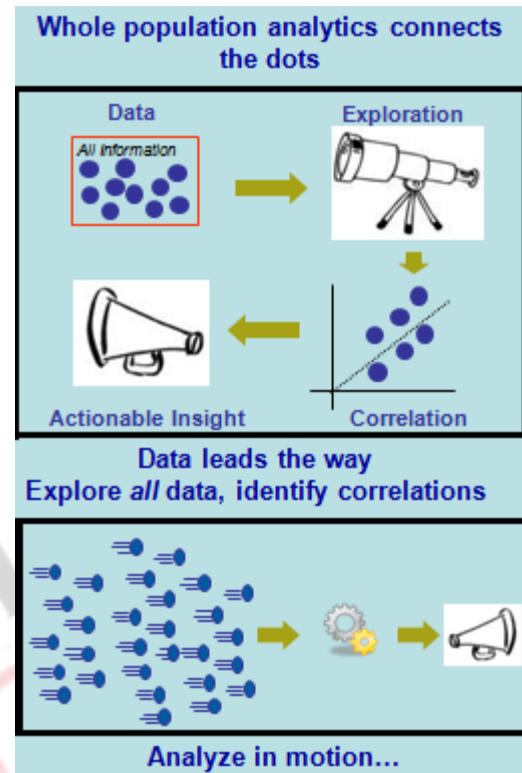


Figure 1 (b) Big Data Analytics

1.4 Big data: This is just the beginning

More people access information online, and deliver information into the digital world. In order to capitalize on the opportunities of this new era, we must find innovative ways to turn big data into business results. The vast quantities of information that are now available to us are changing the way businesses operate and more research into the domain. Data is just a means to an end. At its essence, the massive growth in digital information provides a new and powerful tool to drive business and add more value to research. Analyzing internal and external data can deliver insights into what your customers are buying, and their broader purchasing habits and interests. This information can reveal opportunities to offer complementary products or services that are targeted to individual customer needs. There are also significant opportunities for businesses to acquire new customers. By analyzing external data to find where prospective clients are, businesses can find a way to be there too.

Data can help you understand market opportunity and how to segment that opportunity. More information means more insights. The big data era offers plenty of scope for enhancing the relevance and value of your business offering. However to truly gain the advantage in this new operating environment, a forward-thinking, innovative approach is absolutely a prerequisite.

The graph in figure 2 provides a clear picture about the growth of data and hence the level of uncertainty

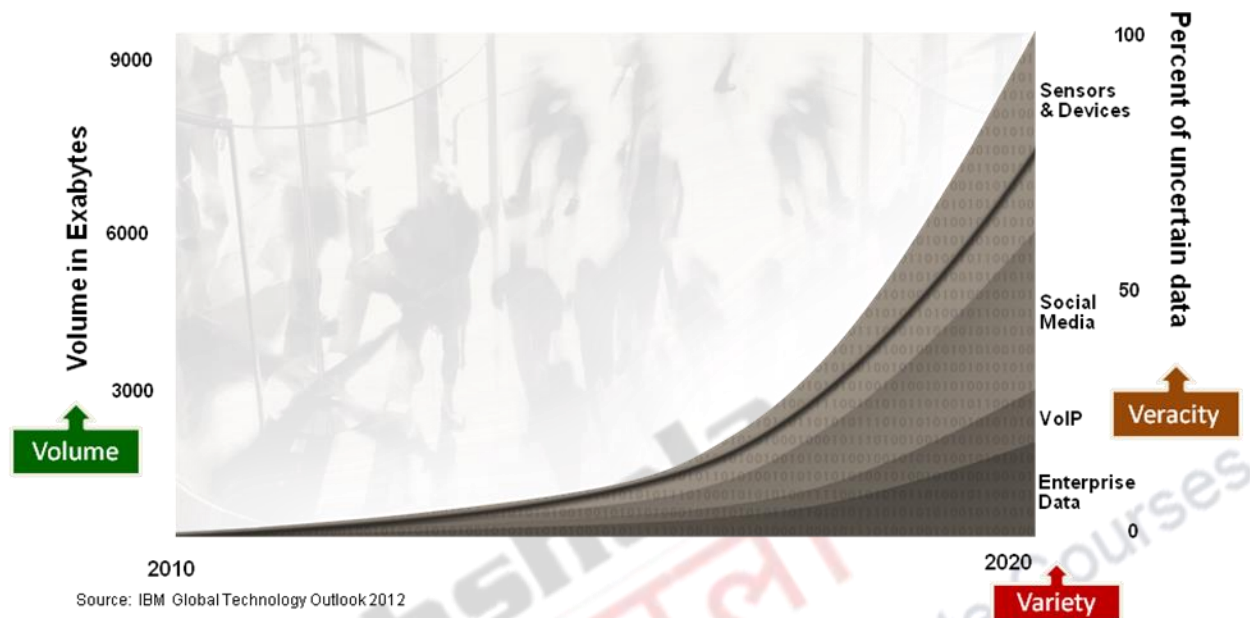


Figure 2. Data Growth – Volume vs Variety vs Veracity

1.5 What to do with these data?

The data that is getting collected are used for various purposes like:

1.5.1 Aggregation and Statistics

Data aggregation is any process in which information is gathered and expressed in a summary form, for purposes such as statistical analysis. A common aggregation purpose is to get more information about particular groups based on specific variables such as age, profession, or income. Data aggregation is a common phenomenon for multidimensional data analysis in data warehouses and OLAP operations.

Ineffective data aggregation is currently a major component that limits query performance. And, with up to 90 percent of all reports containing aggregate information, it becomes clear why proactively implementing an aggregation solution can generate significant performance benefits, opening up the opportunity for companies to enhance their organizations' analysis and reporting capabilities.

Aggregates are used in dimensional models of the data warehouse to produce dramatic positive effects on the time it takes to query large sets of data. At the simplest form an aggregate is a simple summary table that can be derived by performing a Group by SQL query. A more common use of aggregates is to take a dimension and change the

granularity of this dimension. When changing the granularity of the dimension the fact table has to be partially summarized to fit the new grain of the new dimension, thus creating new dimensional and fact tables, fitting this new level of grain. Aggregates are sometimes referred to as pre-calculated summary data, since aggregations are usually precomputed, partially summarized data that are stored in new aggregated tables.

1.5.2 Indexing, Searching, and Querying

The term “big data” is defined as a huge amount of digital information, so big and so complex that normal database technology cannot process it. Accessing and searching such voluminous data requires choosing appropriate indexing techniques. Indexing based on keys is suitable for keyword based search and pattern matching applications.

Big data indexing requirements are as follows:

- i) Speed of search - Search over billions – trillions data values in seconds
- ii) Multi-variable queries - Be efficient for combining results from individual variable search results
- iii) Size of index - Index size should be a fraction of original data
- iv) Parallelism - Should be easily partitioned into pieces for parallel processing
- v) Speed of index generation - For in situ processing, index should be built at the rate of data generation

1.5.3 Knowledge discovery

Data from sources including online activities (social networking, social media), telecommunications (mobile computing, call statistics), scientific activities (simulations, experiments, environmental sensors), and the collation of traditional sources (forms, surveys). Knowledge discovery by applying various data mining and statistical modeling techniques on such data has become strategically important for large business enterprises, government organizations, and research institutions.

1.6 Web – The Actual Big Data

With the advent of technologies such as mobile devices, video customer support, online communities, social media platforms, and more, the various touch points with which customers may interact has proliferated, which can make the task of aggregating the data from these various interactions more difficult to achieve.

1.6.1 360-Degree View

Companies now use an increasing array of tools to develop this 360-degree view (figure 3), including social media listening tools to gather what customers are saying on sites like Facebook and Twitter, predictive analytics tools to determine what customers may research or purchase next, customer relationship management suites and marketing automation software. The 360-degree customer view is the idea, sometimes considered unattainable, that companies can get a complete view of customers by aggregating data from the various touch points that a customer may use to contact a company to purchase products and receive service and support.



Figure 3. 360-degree view of customers

The 360-degree view of customers also often requires a big data analytics strategy to marry structured data, or data that can reside in the rows and columns of a database, with unstructured data as it resides on social media platforms and so forth is becoming increasingly important. Many companies are trying to develop to combine these sources of data and to analyze them in a central location.

Unfortunately the data collected by organizations is uncertain because of missing data. About 2% of browsing sessions complete a purchase and information is missing on more than 98% of web sessions which might be avoided unless otherwise if only transactions are tracked. For every purchase transaction, there might be dozens or hundreds of specific actions and such information needs to be collected and analyzed.

Data that should be collected must be detailed event history from any customer touch point like Web sites, Kiosks, Mobile apps, Social media, etc. Such data helps in capturing the following behaviors:

Purchases	Requesting help
Product views	Forwarding a link
Shopping basket additions	Posting a comment
Watching a video	Registering for a webinar
Accessing a download	Executing a search
Reading / writing a review	And many more!

1.6.2 What Web Data Reveals

The web data reveals the shopping behavior of customers when the following points are analyzed:

1. How customers come to a site to begin shopping
2. What search engine do they use?
3. What specific search terms are entered?
4. Do they use a bookmark they created previously?
5. Who took advantage of any other information?
6. Which products were added/later removed to a wish list or basket?


By examining all the products the customer explore by asking the following questions:

1. Who looked at a product landing page?
2. Who drilled down further?
3. Who looked at detailed product specifications?
4. Who looked at shipping information?

Such analysis helps organizations understand how customers utilize the research content which can lead to tremendous insights into:

1. How to interact with each individual customer
2. How different aspects of the site do or do not add value

By parsing the questions and comments via online help, it is possible to get a feel for what each specific customer is asking about. Moreover, Web data enables to segment customers based upon typical browsing patterns and research on feedback behaviors shall reveal detailed feedback on products and services. Also, by using text mining, we can understand the tone, intent and topic of interest of customers.

	Case Studies
A) eBay's Customer Journey	
Source: http://www.computerweekly.com/news/2240219736/Case-Study-How-big-data-powers-the-eBay-customer-journey	
<p>As a marketplace, eBay's primary business involves being successful from a buyer's and a seller's perspective. The company is using analytics to help it understand its customers better. eBay tries to adopt the working principle of a small store, "engaging the customer is key, helping them with search and recommendations, understanding their preferences and learning from existing customers".</p> <p>Web metrics data is the raw material at eBay's disposal. The auction site generates a huge amount of web analytics, which is described as "the customer journey data", which tells, what people do on eBay and how they use the site.</p> <p>The web offers the same experience [as a local shop], and provide customers with comparisons," and help learn customers' intentions." All this insight drives technology changes at eBay.</p> <p>The challenge for eBay is that web analytics is like having a video camera mounted on the head of every customer going into a supermarket. Recording everything every customer does generates 100 million hours of customer interaction [per month], creating an unmanageable amount of customer data. Understanding the customers, learning from customers and apply data science techniques to get more</p>	

data and new types of data, becomes vital for retaining the customers.

The eBay site has 100 million customers who list items in 30,000 categories. In terms of transactions, the site processes thousands of dollars per second. The big data challenge for eBay is that asking a simple business question such as "What were the top items that showed up in searches yesterday?" involves processing five billion page views. But eBay needs to do more than ask simple questions like sentiment analysis, network analysis and image analysis, all of which cannot be run in a traditional transactional database.

The company has split its data analytics across three platforms, the first of which is a traditional enterprise data warehouse from Teradata. This core transactional system is extremely reliable, and every day eBay processes 50TB of data, accessed by 7,000 analysts with 700 [concurrent users]." In 2002, eBay built a 13TB Teradata enterprise data warehouse, which effectively provides a massive parallel relational database. This has now grown to 14PB, with the system built on hundreds of thousands of nodes. The enterprise data warehouse gives tremendous performance on standard structured queries, but it is unable to meet eBay's needs for storage and processing flexibility.

To address this issue, eBay started its second data initiative. Seven years ago, the company began a project to store all its customer data. The auction site needed a product that could handle hundreds of petabytes of raw customer journey data, but would be easy to maintain by a team of five people, yet could be accessed easily by analysts. The company worked with Teradata to develop a custom appliance built with several hundred user-defined functions. The system was built on commodity hardware, with proprietary software to process all the customer journey data and store it cheaply.

The end result is a custom data warehouse called Singularity. The system eBay has developed can run ad-hoc queries in 32 seconds. Along with the enterprise data warehouse and Singularity, eBay is also using Hadoop, which completes the third side of its data analytics triangle. The auction site has built two 20,000-node Hadoop clusters with 80PB of capacity. These work alongside the Teradata data warehouse and Singularity custom data analytics appliance to give eBay the tools it needs to use data analysis to follow the customer journey.

B) Asean organisation's enterprise-wide approach to analytics for drawing on customer insight to maximize the business value of data

Source:<http://www.computerweekly.com/feature/Analytics-helps-Asean->

organisations-read-between-the-lines

In 2013, 96% of Asean business leaders were committed to the adoption of analytics or fact-based decision-making, sur-passing their counterparts in the UK (86%) and the US (85%).

However, close to 90% of Asean companies indicated that analytics is not yet deeply ingrained as an inte-grated, enterprise-wide approach. Nine days was all it took the National Library Board (NLB) of Singapore to generate in excess of 130 million “related articles” for more than 1.5 million newspaper articles in the NewspaperSG collection of historic Singapore newspapers.

The NLB was able to automatically identify the related articles using a Hadoop cluster through the text analytics (TA) component of its big data programme. Had it done this manually, it would have taken 992 man-years, assuming it takes a person who works 42 hours per week one minute to identify each of the 130 million relationships.

The NLB consists of one national library, one national archive and 26 public libraries. According to available figures from 2010 – now probably modest compared with today – two million members visited the libraries 36 mil-lion times to borrow from a collection of 1.5 million titles and 8.5 million items. In that year alone, the libraries processed more than 33 million loans and had more than 8.1 million digital user visits and 47.4 million e-retrievals.

NLB’s big data architecture consists of its enterprise data warehouse; data marts; extract, transform and load (ETL) tool; in-memory dashboards (including collection, patron and production dashboards); advanced analytics (including recommendation engine, collection planning, demand analysis and geospatial ana-lytics); content analytics; and text analytics.

Strategically, the NLB big data programme analyses and optimises the network of libraries nationwide, examines the use of these libraries and how well they are serving the community, and looks at how new libraries affect existing ones.

Summary

- Big data has lots of uncertainty and no standards available
- Big data approach to analytics provides a new dimension to analytics
- Many potential applications of Big Data Web data offers many useful information

- The web proves be the original big data
 - Quantum of data
 - Variety
 - uncertainty

