e-PGPathshala

Subject : Computer Science

Paper: Data Analytics

Module: Introduction to Analytics and Big Data

Module No: CS/DA/1

Quadrant 1 – e-text

1.1 Introduction

Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of traditional database architectures. To gain value from this data, you must choose an alternative way to process it, i.e., effective data analytics is required. This module gives you a complete picture about big data and how analytics is carried on such a data.

1.2 Learning Outcomes

- To learn the fundamentals of data analytics
- To know the difference between data analytics and data mining
- To understand big data and its impact on analytics
- To understand characteristics of big data
- To know about applications involving big data and analytics

1.3 Data Analytics

Data Analytics (DA) is the **science of examining raw data with the purpose of drawing conclusions** about that information. The data that is captured by any data collection agent or tool or software is in its raw form, i.e., unformatted or unstructured or unclean with noises/errors or redundant or inconsistent. Hence, analytics covers a spectrum of activities starting from data collection till visualization.

The science of data analytics is generally divided into three broad categories:

(i) Exploratory Data Analysis (EDA)

(ii) Confirmatory Data Analysis (CDA)

(iii) Qualitative Data Analysis (QDA)

1.3.1 Exploratory Data Analysis (EDA)

EDA is an approach/philosophy for data analysis that employs a variety of techniques enticing the data to reveal its structural secrets, and being always ready to gain some new, often unsuspected, insight into the data. Most EDA techniques are graphical in nature.

Advantages	Disadvantages
a) Flexible ways to generate hypotheses b) More realistic statements of accuracy c) Does not require more than data can support d) Promotes deeper understanding of processes by Statistical learning	a) Usually does not provide definitive answers b) Difficult to avoid optimistic bias produced by over fitting c) Requires judgment and artistry - can't be cook booked

1.3.2 Confirmatory Data Analysis (CDA)

This approach for data analysis is aimed towards proving or disproving existing hypotheses. CDA, a deductive approach is inferential in nature. It relies heavily on probability models and hypotheses are determined at outset. CDA provides precise information in the right circumstances and backed with well-established theory and methods.

Advantages	Disadvantages
a) Provide precise information in the right circumstances b) Well-established theory and methods	a) Misleading impression of precision in less than ideal circumstances b) Analysis driven by preconceived ideas c) Difficult to notice unexpected results

1.3.3 Qualitative Data Analysis (QDA)

QDA is all about drawing conclusions from non-numerical data for analysis, which might be further used for decision making.

Advantages	Disadvantages
a) allow for a broader study, involving a greater number of subjects, and enhancing the generalisation of the results b) can allow for greater objectivity and accuracy of results. Generally, quantitative methods are designed to provide summaries of data that support generalisations about the phenomenon under study. In order to accomplish this, quantitative research usually involves few variables and many cases, and employs prescribed procedures to ensure validity and reliability c) using standards means that the research can be replicated, and then analyzed and compared with similar studies. Kruger (2003) confirms that 'quantitative methods allow us to summarize vast sources of information and facilitate comparisons across categories and over time' d) personal bias can be avoided by researchers keeping a 'distance' from participating subjects and employing subjects unknown to them	a) collect a much narrower and sometimes superficial dataset b) results are limited as they provide numerical descriptions rather than detailed narrative and generally provide less elaborate accounts of human perception c) the research is often carried out in an unnatural, artificial environment so that a level of control can be applied to the exercise. This level of control might not normally be in place in the real world yielding laboratory results as opposed to real world results d) in addition preset answers will not necessarily reflect how people really feel about a subject and in some cases might just be the closest match. e) the development of standard questions by researchers can lead to 'structural' bias and false representation, where the data actually reflects the view of them instead of the participating subject.

1.3.4 Data Analytics vs Data Mining

Data analytics is distinguished from data mining by the scope, purpose and focus of the analysis. Data miners sort through huge data sets using sophisticated software to identify undiscovered patterns and establish hidden relationships. Data analytics focuses on inference, the process of deriving a conclusion based solely on what is already known by the researcher.

1.4 Big Data Analytics

Big data analytics refers to the **process of collecting, organizing and analyzing very large sets of data ("big data") to discover patterns and other useful information**.

With big data analytics, data scientists and others can analyze huge volumes of data that conventional analytics and business intelligence solutions can't touch. Consider that your organization could accumulate billions of rows of data with hundreds of millions of data combinations in multiple data stores and abundant formats. High-performance analytics is necessary to process that much data in order to figure out what's important and what isn't.

Analyzing big data allows analysts, researchers, and business users to make better and faster decisions using data that was previously inaccessible or unusable. Using advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics, and natural language processing, businesses can analyze previously untapped data sources independent or together with their existing enterprise data to gain new insights resulting in significantly better and faster decisions.

1.5 Introduction to Big Data

The recent hot IT buzzword, big data has become viable as cost-effective approaches have emerged to tame the volume, velocity and variability of massive data. Within this data lie valuable patterns and information, previously hidden because of the amount of work required to extract them. Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn’t fit the structures of traditional database architectures. To gain value from this data, you must choose an alternative way to process it.

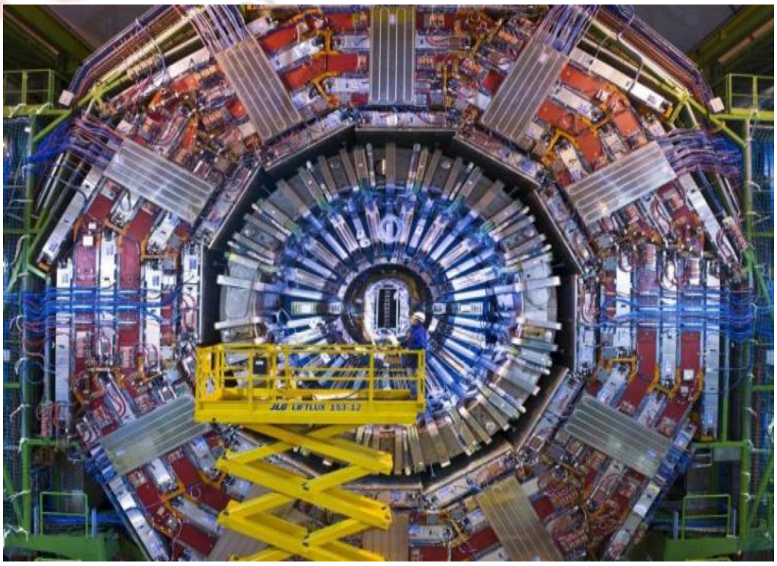
1.5.1 What does big data look like?

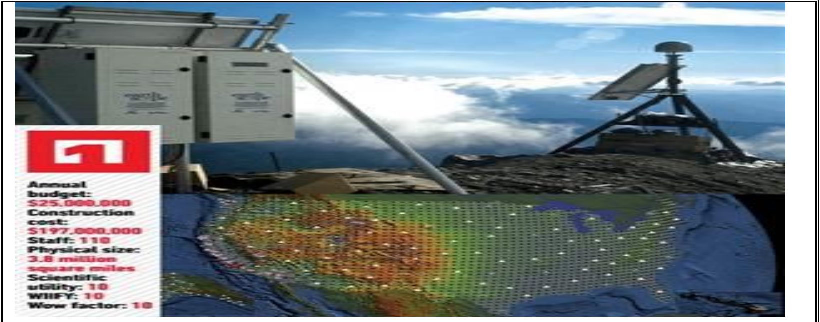
As a catch-all term, “big data” can be pretty nebulous, in the same way that the term “cloud” covers diverse technologies. Input data to big data systems could go on from social networks, web server logs, traffic flow sensors, satellite imagery, broadcast audio streams, banking transactions, MP3s of music, the content of web pages, scans of government documents, GPS trails, telemetry from automobiles, financial market data, etc.

1.5.2 Quantum of data that we process

- The **Internet Archive** surpassed 15 petabytes as of May 2014
- **Google** processes 30 PB a day- "MapReduce“ - Portal.acm.org
- **AT&T** transfers about 30 petabytes of data through its networks each day - "AT&T- News Room“

- The **German Climate Computing Centre (DKRZ)** has a storage capacity of 60 petabytes of climate data
- **Wayback Machine** has 3 PB + 100 TB/month
- **Facebook** has 2.5 PB of user data + 15 TB/day
- As of January 2013, Facebook users had uploaded over 240 billion photos. For each uploaded photo, Facebook generates and stores four images of different sizes, which translated to a total of 960 billion images and an estimated 357 petabytes of storage
- **eBay** has 6.5 PB of user data + 50 TB/day
- **CERN's Large Hydron Collider (LHC)** generates 30 PB a year
- Movie **Avatar** is reported to have taken over 1 petabyte of local storage for the rendering of the 3D CGI effects
- **Teradata** Database 12 has a capacity of 50 petabytes of compressed data

<table>
<tr><td></td><td><b>Do you Know?</b></td></tr>
<tr><td colspan="2"><b>Computing in CERN</b></td></tr>
<tr><td>In CERN's Large Hydron Collider (LHC) approximately 600 million times per second, particles collide within the Large Hadron Collider (LHC). Electronic circuits record and send the data to the CERN Data Centre (DC) and approximately 30 petabytes of data are produced annually. The Data Centre processes about one petabyte of data every day which is equivalent to around 210,000 DVDs. The centre hosts 11,000 servers with 100,000 processor cores.</td><td>
Source: http://home.web.cern.ch/about/experiments/cms</td></tr>
</table>

<table>
<tr><th colspan="2">Earthscope</th></tr>
<tr><td>The <b>Earthscope</b> is the world's largest science project, designed to track North America's geological evolution. Earthscope records 67 terabytes of data.</td><td>
1
Annual budget: $25,000,000
Construction cost: $197,000,000
Staff: 110
Physical size: 3.8 million square miles
Scientific utility: 10
WIIFY: 10
Wow factor: 10
Source: http://www.earthscope.org/</td></tr>
</table>

1.5.3 What is Big Data?

> “Big data **exceeds the reach of commonly used hardware environments and software tools to capture, manage and process** it within a tolerable elapsed time for its user population.”
>
> **- Gartner’s Merv Adrian in Q1, 2011 Teradata Magazine article**
>
> “Big data” refers to **datasets whose size is beyond the ability of typical database software tools** to capture, store, manage, and analyze.”
>
> **- McKinsey Global Institute - Big data: The next frontier for innovation, competition, and productivity , 2011.**

1.5.3.1 Characteristics

Big data can be described by four major characteristics as shown in figure1:

i) Volume

ii) Variety

iii) Velocity

iv) Veracity

Volume

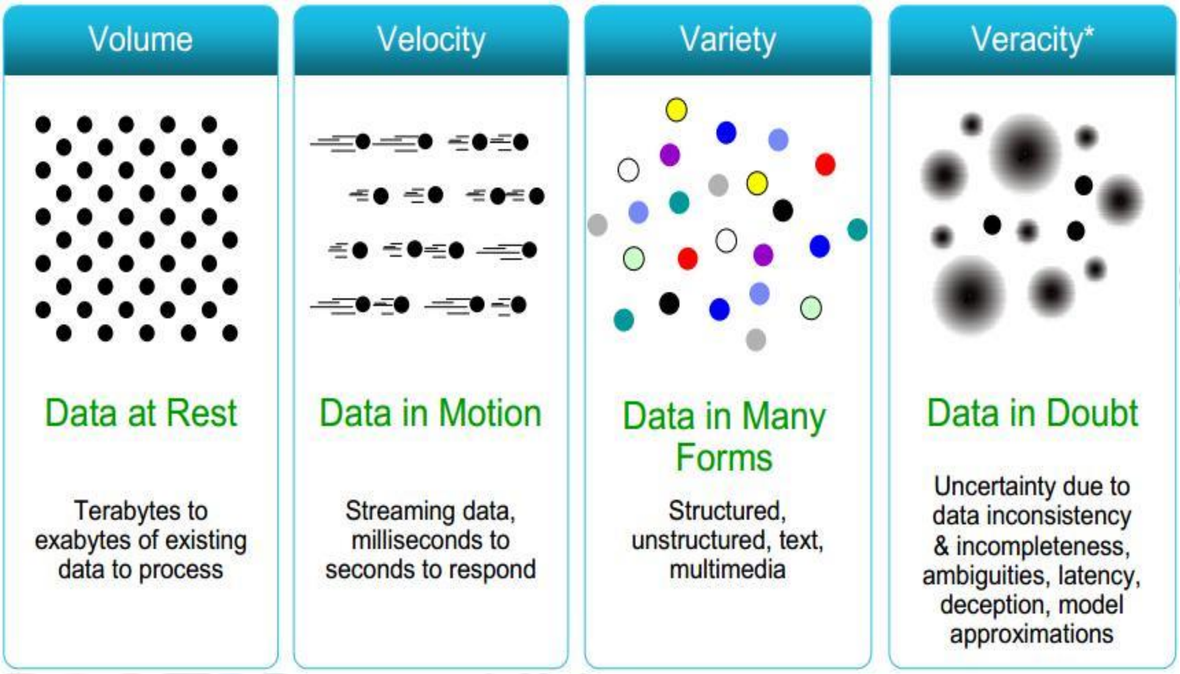
The quantity of generated data is important in this context. The size of the data determines the value and potential of the data under consideration, and whether it can actually be considered big data or not. The name ‘big data’ itself contains a term related to size, and hence the characteristic.

Variety

The type of content, and an essential fact that data analysts must know. This helps people who are associated with and analyze the data to effectively use the data to their advantage and thus uphold its importance.

Velocity

In this context, the speed at which the data is generated and processed to meet the demands and the challenges that lie in the path of growth and development.


Volume
Data at Rest
Terabytes to exabytes of existing data to process
Velocity
Data in Motion
Streaming data, milliseconds to seconds to respond
Variety
Data in Many Forms
Structured, unstructured, text, multimedia
Veracity*
Data in Doubt
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Source: http://www.datasciencecentral.com/profiles/blogs/data-veracity

Figure 1. Characteristics of Big Data

Veracity

The quality of captured data, which can vary greatly. Accurate analysis depends on the veracity of source data.

Other features that might also be used for characterizing bid data are as follows:

Variability

The inconsistency the data can show at times—-which can hamper the process of handling and managing the data effectively.

Complexity

Data management can be very complex, especially when large volumes of data come from multiple sources. Data must be linked, connected, and correlated so users can grasp the information the data is supposed to convey.

1.5.3.2 How big data is different?

Typically the following points may be used by the readers to understand how big data differs from that of a traditional data which is organized in single or multiple storages:

- Generated automatically by machine
- 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

1.6 What to do with these data?

Having understood the context of big data and its features, ultimately the next question that arises in the readers' mind is, "what can be done with such a data?". Big data shall lead the users to perform the following tasks:

- Aggregation and Statistics
  - Data warehouse and OLAP
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)
- Knowledge discovery
  - Data Mining
  - Statistical Modeling

1.7 Big Data Use Cases

A big data use case can help you solve a specific business challenge by using patterns or examples of big data technology solutions. Few domains which might produce big data and the challenges in them are summarized below. The new form of data that might emerge from such domains and possible analytics are also highlighted in the table 1.

1.7.1 Healthcare

Making use of the petabytes of patient data that healthcare organizations possess requires extracting it from legacy systems, normalizing it and then building applications that can make sense of it in the following ways:

1. Develop evidence-based Medicine
2. 'Domesticate' Data for Better Public Health Reporting, Research
3. Use Free Public Health Data For Informed Strategic Planning

Table 1. Big Data Use Cases

Application Domain	**New Data**	**Solution**
Healthcare	Patient Monitoring Data	Preventive health care, Disease Drug Analysis
Manufacturing	Sensor data	Automated Diagnosis, support
Retail	Social media data	Customer and market segmentation, sentiment analysis
Public sector	Surveys	Personalized services

1.7.2 Manufacturing

Analyzing big data use cases in the manufacturing industry can reduce processing flaws, improve production quality, increase efficiency, and save time and money. It also leads to the following:

1. Improving Manufacturing Process
2. Custom Product Design
3. Better Quality Assurance
4. Managing Supply Chain Risk

1.7.3 Location based services

Mobile location-based services and corresponding big data can help developing patterns of usage that lets you not only provide a wider variety of services but also aids predictive analytics to understand who your subscribers are, how they use your services and what will entice them to stay with your service. Telecommunications companies can focus marketing efforts on the right customers by using Behaviour-based Customer Insight for Telecommunications.

1.7.4 Public sector

Big data is changing the way government agencies store, manage and collect data, but they cannot rely on structured data to meet their goals. Like commercial businesses, public-sector agencies need to capture data from a variety of sources, including unstructured data such as email and social media, to turn it into actionable items. Government agencies can build a case for big data as well as develop a data and analytics strategy and platform so they become an integral part of the agency's big data strategy.

1.7.5 Retail

Retailers can use "big data" - combining data from web browsing patterns, social media, industry forecasts, existing customer records, etc. - to predict trends, prepare for demand, pinpoint customers, optimize pricing and promotions, and monitor real-time analytics and results.

Case Studies

A) **Massachusetts General Hospital**

Source: http://healthitanalytics.com/news/four-use-cases-for-healthcare-predictive-analytics-big-data

At Massachusetts General Hospital, an analytics system called QPID is helping providers ensure that they don’t miss critical patient data during admission and treatment. The system is also used to predict surgical risk, helping match patients with the right course of action that will keep them safest during their care. The system automates searches using national guidelines, and then it essentially shows the results in a dashboard with a red, yellow, or green risk indicator for the surgeon to see.”

QPID’s Q-Core clinical reasoning engine mines the entire patient record and extracts facts that are assembled against a structured patient data model. The model is used by the reasoning engine to answer complex questions about the patient’s medical history and status. Q-Core findings are delivered through QPID applications, which expedite quality reporting, registry submissions and cohort identification. Unlike standard document-based natural language processing (NLP), which can yield results that are not highly accurate, Q-Core mines the entire longitudinal patient record and extracts facts that are assembled against a structured patient data model to produce a synthesized view of the patient's medical history and status.

B) United Parcel Service (UPS) Incorporation

Source: http://www.sas.com/en_us/insights/big-data/what-is-big-data.html

UPS having begun to capture and track a variety of package movements and transactions as early as the 1980s.The company now tracks data on 16.3 million packages per day for 8.8 million customers, with an average of 39.5 million tracking requests from customers per day. The company stores more than 16 petabytes of data.

Much of its recently acquired big data, however, comes from telematics sensors in more than 46,000 vehicles. The data on UPS trucks, for example, includes their speed, direction, braking and drive train performance. The data in not only used to monitor daily performance, but to drive a major redesign of UPS drivers' route structures. This initiative, called ORION (On-Road Integration Optimization and Navigation), is arguably the world's largest operations research project. It also relies heavily on online map data, and will eventually reconfigure a driver's pickups and drop-offs in real time.

C) Tesco Plc

Source: http://dataconomy.com/tesco-pioneers-big-data/

Tesco Plc, the British supermarket chain, is currently the second most profitable retailer in the world with outlets in twelve countries. Tesco began collecting ever more data on its consumers and was one of the first companies to embrace, and learn from, Big Data analytics. Tesco realized the value of the insight it would be getting into its customers' behaviours and now receives detailed data on two-thirds of all shopping baskets. Tesco was processed the flood of data that descended upon them.

The first step, however, was to segment the customers into appropriate groups. That resulted in two things. On the one hand Tesco could actually be more targeted in its mailings of vouchers and coupons. Having broken down customers into segments, Tesco increased its reach by launching the Clubcard Plus, which had an integrated debit card. This was later replaced by a credit card but nonetheless lured customers into spending more at Tesco. Using all this data Tesco started trying to convert the non-buyers. For example, finding that recent parents were spending their money elsewhere, they launched a Baby club and ended up capturing 24% of the baby market.

Seeing that its analytics approach worked, Tesco started applying it to other fields also. One example is its optimized stock keeping system which forecasts sales by product for each store based on historical sales and weather data. Through predictive analytics Tesco managed to save 100m pounds in stock that would have otherwise expired and thus wasted. In another instance Tesco found that its management of the fridge and store temperatures was sub-optimal and thus enabled significant savings in energy costs.

Using the insights it gained from the collected data Tesco evolved from a retailer that thought it knew what the customers wanted into one that actually did know and could monitor the preferences as they changed over time. Tesco managed to break its customers down into segments it understood better and thus target its sales efforts accordingly.

Summary

The introductory module on big data shall convey the readers the following points:

- Definition of Data Analytics and its broad classification.
- The differences between Data analytics and Data Mining.
- Perspective of big data and application of analytics on it. Characteristics of Big Data and its implications in decision making.
- Real time use cases related with big data and possible application of analytics.