Big Data Analytics

Unit I

Introduction to Big Data

Today’s Topics

- Big Data Overview
  - What is Big Data?
  - Data -> Big Data
  - Definition of Big Data
  - Data Structure / Types of Big Data
  - Growth of Data - Tools and Techniques

Big Data Overview

What is Big Data?

The term Big Data refers to a huge volume of data that can not be stored, processed by any traditional data storage or processing units.

Big Data is generated at a very large scale and it is being used by many multinational companies to process and analyse in order to uncover insights and improve the business of many organisations.

What is Big Data?

Data is created constantly, and at an ever-increasing rate

Data Storage Units Chart: From Smallest to Largest

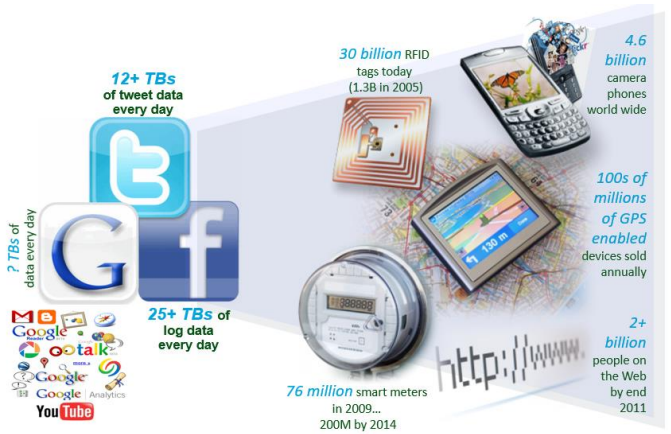
Unit	Shortened	Capacity
Bit	b	1 or 0 (on or off)
Byte	B	8 bits
Kilobyte	KB	1024 bytes
Megabyte	MB	1024 kilobytes
Gigabyte	GB	1024 megabytes
Terabyte	TB	1024 gigabytes
Petabyte	PB	1024 terabytes
Exabyte	EB	1024 petabytes
Zettabyte	ZB	1024 exabytes
Yottabyte	YB	1024 zettabytes

Why Big Data?

- With the development and increase of apps and social media and people and businesses moving online, there's been a huge increase in data.
- If we look at only social media platforms, they interest and attract over a million users daily, scaling up data more than ever before.

How exactly is this huge amount of data handled and how is it processed and stored?

This is where Big Data comes into play ….



Why Big Data?

- Big Data is creating significant new opportunities for organizations to derive new value and create competitive advantage from their most valuable asset: information.
- For businesses, Big Data helps drive efficiency, quality, and personalized products and services, producing improved levels of customer satisfaction and profit
- Big Data Analytics provide new age tech like machine learning, mining, statistics and more
- For scientific efforts, Big Data analytics enable new avenues of investigation with potentially richer results and deeper insights than previously available.
- In many cases, Big Data analytics integrate structured and unstructured data with real time feeds and queries, opening new paths to innovation and insight.

Data → Big Data

- Devices and sensors automatically generate diagnostic information that needs to be stored and processed in real time.
- Merely keeping up with this huge influx of data is difficult, but substantially more challenging is analyzing vast amounts of it, especially when it does not conform to traditional notions of data structure, to identify meaningful patterns and extract useful information.
- These challenges of the data deluge present the opportunity to transform business, government, science, and everyday life.

Data → Big Data

Several industries have led the way in developing their ability to gather and exploit data:

- Credit card companies monitor every purchase their customers make and can identify fraudulent purchases with a high degree of accuracy using rules derived by processing billions of transactions.
- Mobile phone companies analyze subscribers' calling patterns to determine, for example, whether a caller's frequent contacts are on a rival network. If that rival network is offering an attractive promotion that might cause the subscriber to defect, the mobile phone company can proactively offer the subscriber an incentive to remain in her contract.
- For companies such as Linked In and Facebook, data itself is their primary product. The valuations of these companies are heavily derived from the data they gather and host, which contains more and more intrinsic value as the data grows.

Data → Big Data

Attributes stand out as defining Big Data Characteristics

- **Huge volume of data**
  - Rather than thousands or millions of rows, Big Data can be billions of rows and millions of columns.
- **Complexity of data types and structures**
  - Big Data reflects the variety of new data sources, formats, and structures, including digital traces being left on the web and other digital repositories for subsequent analysis.
- **Speed of new data creation and growth**
  - Big Data can describe high velocity data, with rapid data ingestion and near real time analysis.

Big Data Definition

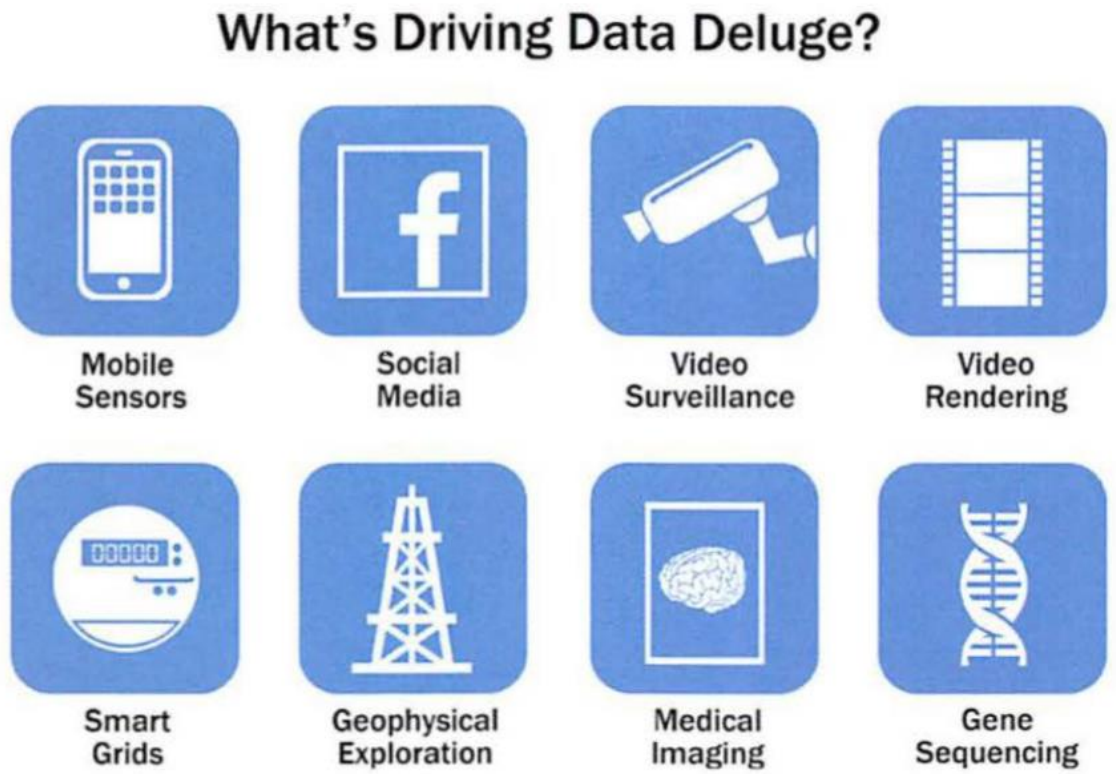
- No single standard definition…

"***Big Data***" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to enable insights that unlock new sources of business value.

Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

Big Data Definition…

- The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.



Data Structures / Types of Big Data

- Big data can come in multiple forms, including structured, Semi-structured and non-structured data such as financial data, text files, multimedia files, and genetic mappings.

- Contrary to much of the traditional data analysis performed by organizations, most of the Big Data is unstructured or semi-structured in nature, which requires different techniques and tools to process and analyze.

- Distributed computing environments and massively parallel processing (MPP) architectures that enable parallelized data ingest and analysis are the preferred approach to process such complex data

Data Structures….

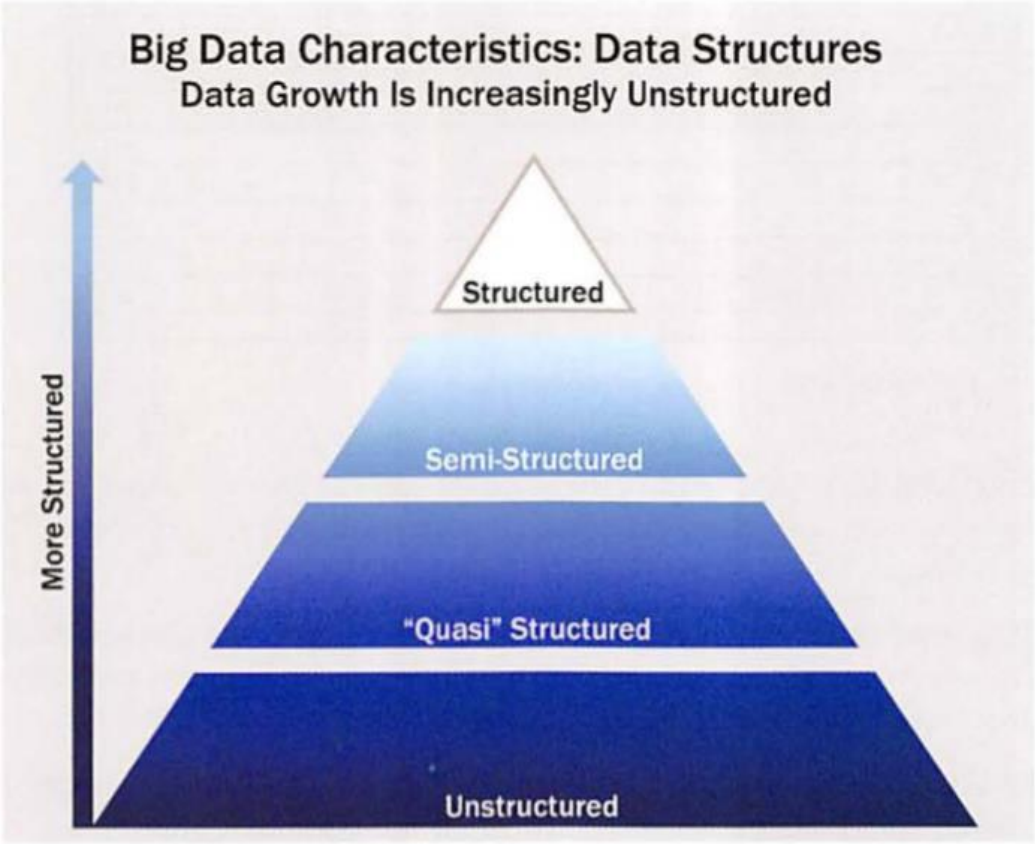


FIGURE 1-3 *Big Data Growth is increasingly unstructured*

Data Structures….

Structured data:

- Data containing a defined data type, format, and structure (that is, transaction data, online analytical processing [OLAP] data cubes, traditional RDBMS, CSV files, and even simple spreadsheets).
- The structured data is stored, processed and retrieved in a fixed format.
- Such data can be stored in Excel, database etc.

SUMMER FOOD SERVICE PROGRAM 1]				
(Data as of August 01, 2011)				
Fiscal Year	Number of Sites	Peak (July) Participation	Meals Served	Total Federal Expenditures 2]
	----------Thousands----------		--Mil.--	---Million $---
1969	1.2	99	2.2	0.3
1970	1.9	227	8.2	1.8
1971	3.2	569	29.0	8.2
1972	6.5	1,080	73.5	21.9

Data Structures….

Semi-structured data:

- Semi structured data pertains to the data containing both formats that is structured and unstructured.
- To be precise, it refers to the data that although has not been classified under a particular repository (database), yet contains vital information or tags that segregate individual elements within the data.
- Textual data files with a discernible pattern that enables parsing (such as Extensible Markup Language [XML] data files that are self-describing and defined by an XML schema).

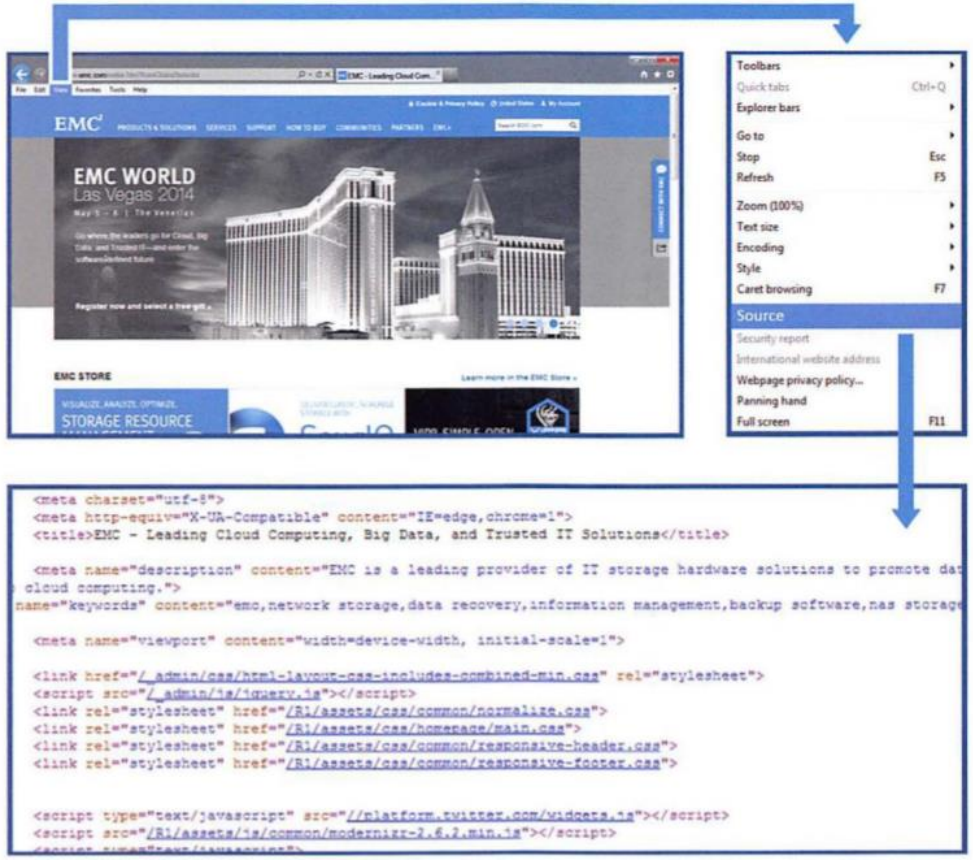


FIGURE 1-5 *Example of semi-structured data*

Data Structures….

Quasi-structured data:

- Textual data with erratic data formats that can be formatted with effort, tools, and time (for instance, web clickstream data that may contain inconsistencies in data values and formats).

FIGURE 1-6 *Example of EMC Data Science search results*

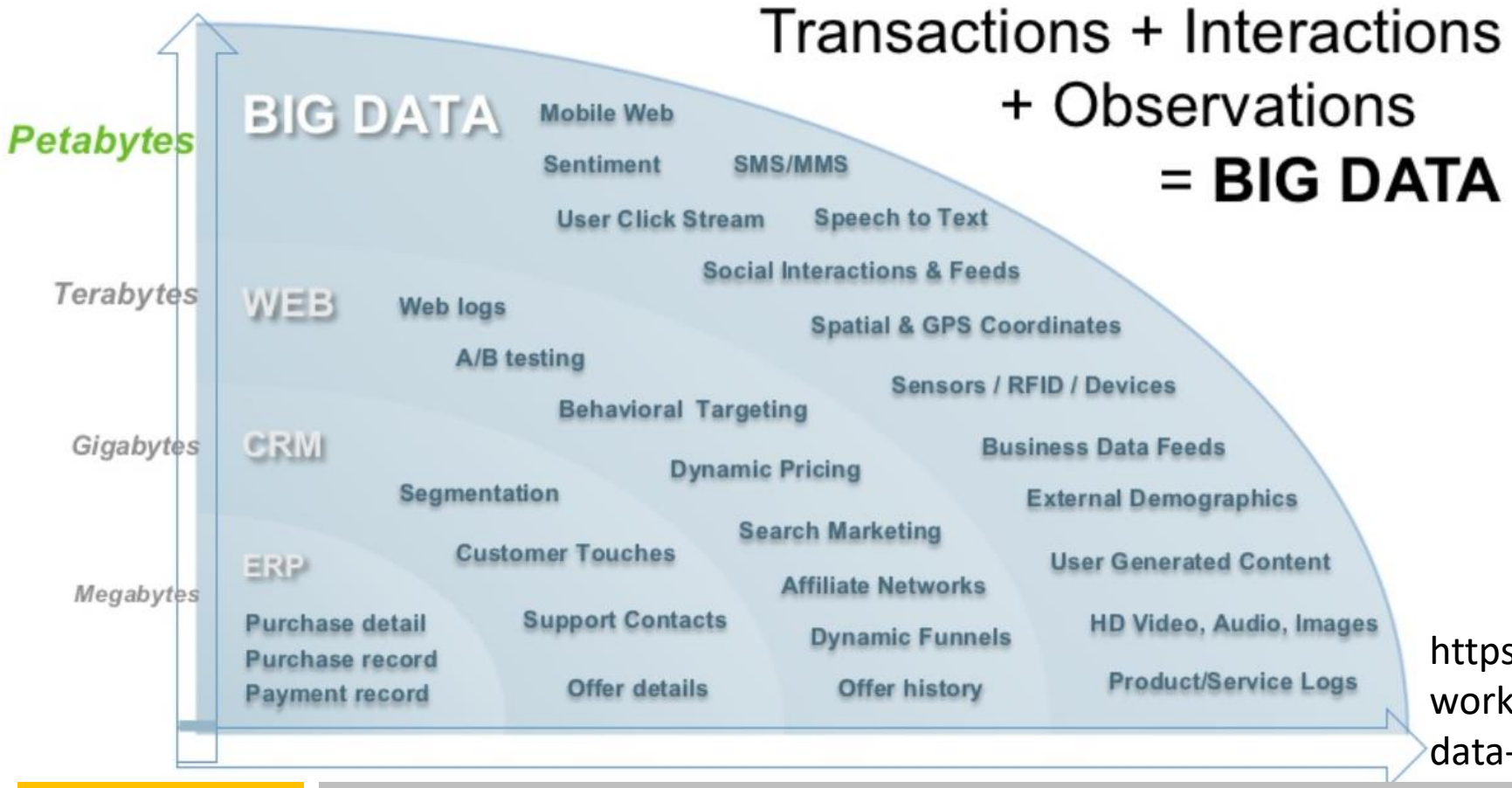
Data Structures….

Unstructured data:

- Data that has no inherent structure, which may include text documents, PDFs, images, email and video.
- Analysis of unstructured data is time-comsuming.

FIGURE 1-7 *Example of unstructured data: video about Antarctica expedition [3]*

Data → Big Data
Growth of Data – Size and Techniques



https://www.slideshare.net/hortonworks/the-next-generation-of-big-data-analytics

Thank You….

Revise the topics from Syllabus References…