SESSION 3

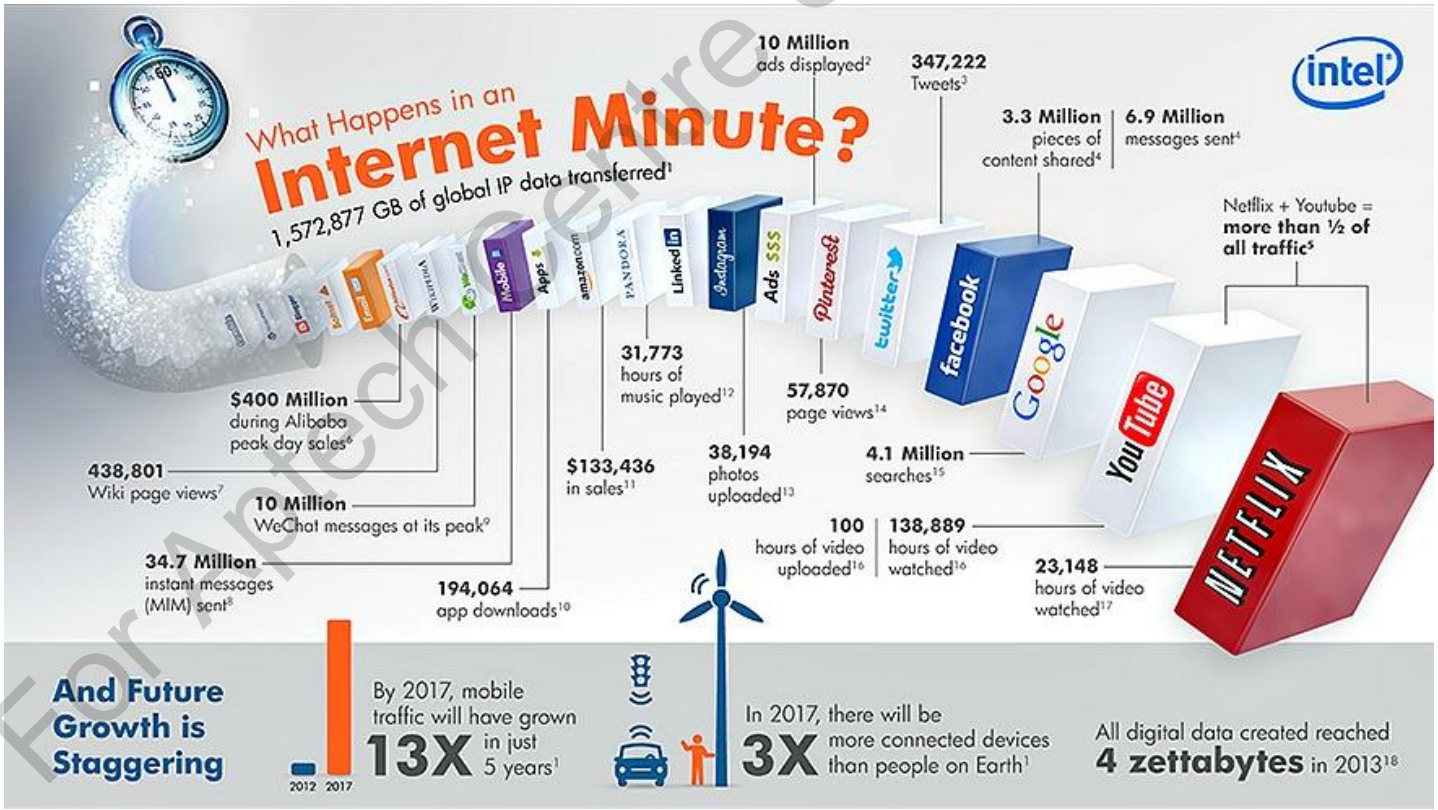
Sources of Big Data and Data Management

Objectives

- Explain the various data sources for Big Data
- Describe data management using Big Data

Sources of Big Data in General (1-3)

- Data is growing at an exponential rate. For example, consider the data produced by the Internet.



Sources of Big Data in General (2-3)

- Big Data sources are repositories of huge amount of data
- Various sources of Big Data are:

Internal Data Streams

- It appear from various controlled channels within the organization
- Website, social media, press releases, or any company related blog

Shared Data Streams

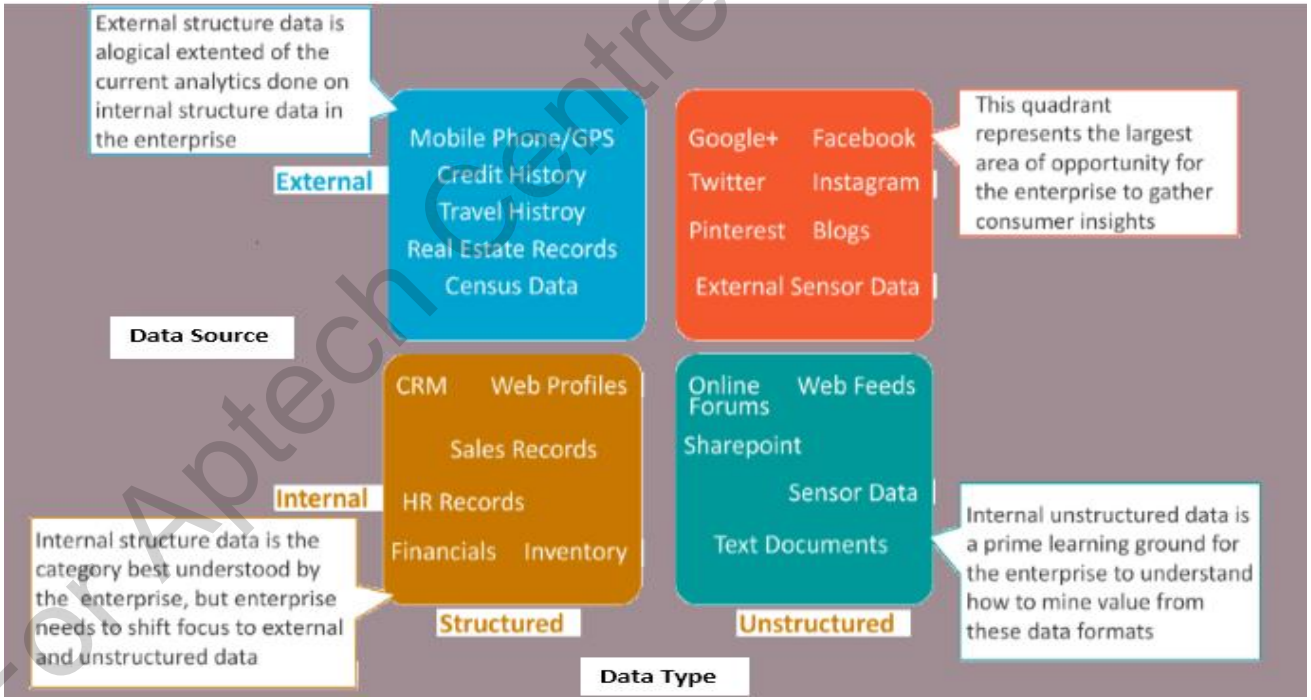
- It is semi controllable stream - collected through sources accessible by the organization and the third-party data streams
- Event, publicity sponsorships, or industry research

External Data Streams

- Processes of business and public relations regarding landscape analysis, strategy development tactics objective setting, and evaluation.
- Includes outside research, academic studies, organic social media conversations, and news

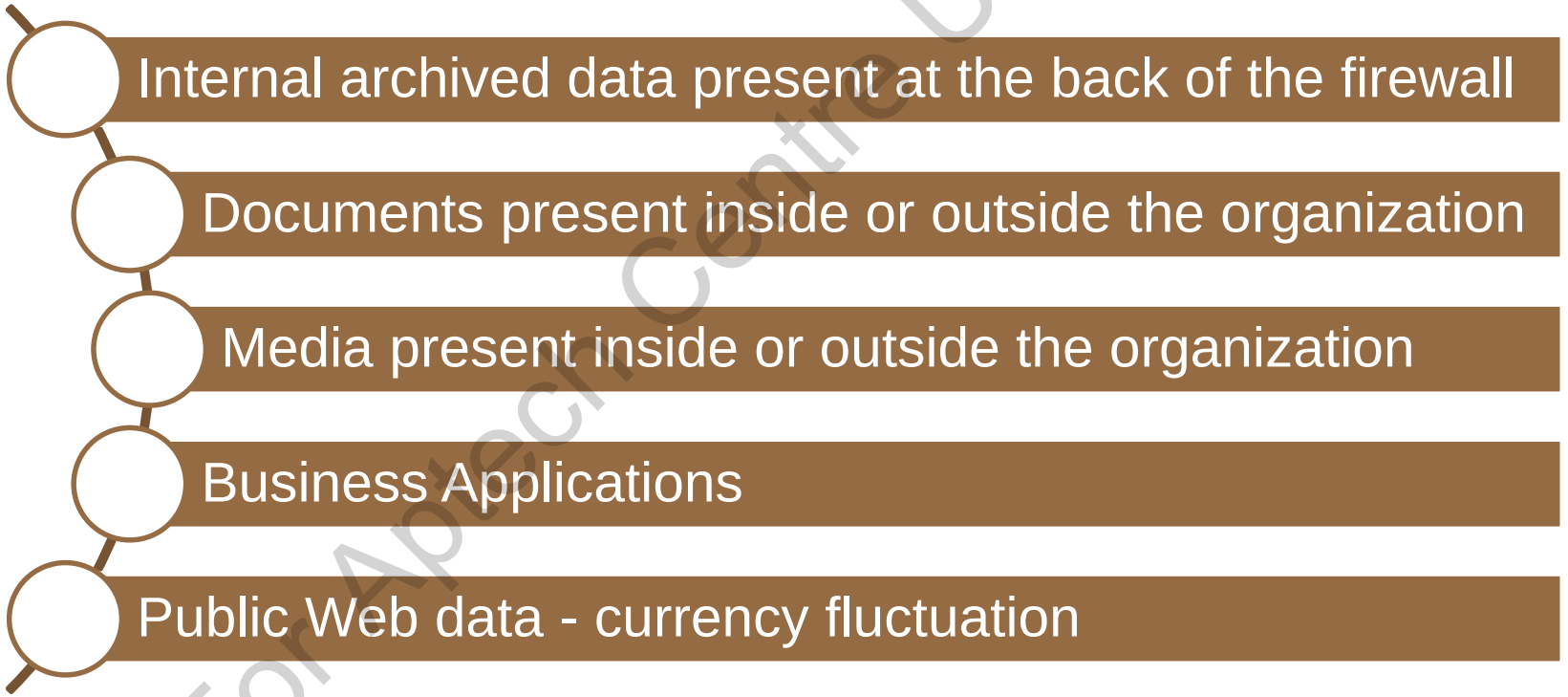
Sources of Big Data in General (3-3)

- Though these data streams are concrete, sometimes the data can be moved from one type of data stream to another type within the organization.



Big Data Sources (1-2)

- Specific sources of Big Data include:



Big Data Sources (2-2)

- Top ten Big Data source types:

Social Network Profiles	Social influencers	Activity Generated Data	SaaS and Cloud Applications	Public
Hadoop MapReduce Application Results	Data Warehouse Appliances	Columnar/No SQL Data Sources	Network and in-stream Monitoring Technologies	Legacy Documents

Big Data Websites (1-2)

- Some Websites that generate Big Data are:

Data.gov	U.S. Census Bureau	Socrata	European Union Open Data Portal	Data.Gov.UK
Canada Open Data	Datacatalogs.org	CIA World Factbook	Healthdata.gov	NHS Digital
UNICEF	WHO	Amazon Web Services	Facebook for Developers	Face.com
		UCLA		

Big Data Websites (2-2)

- Some more Websites that generate Big Data include:

Data Market	Google Public data explorer	Junar	Buzzdata	Gapminder
Google Trends	Google Finance	Google Books	National Climatic Data Center	DBPedia
New York Times	Freebase	Million Song Data Set	UCI Machine Learning Repository	

Big Data Management (1-2)

- Data management is the process of organizing, administrating, and governing large volumes of data. Data can be structured or unstructured data.
- Effective Big Data management helps organizations to gather information.
- Three challenges of Big data are – **Storing**, **Processing**, and **Managing**.

Big Data Management (2-2)

- Data management can be improved in three key-areas:

Time frame: Time taken to manage and process the data is reduced

Security: Data can be secured as the management is centralized

Accuracy: The results of data analysis are accurate as all the copies of data are visible

Big Data Operations (1-5)

- Data Storage
  - Big Data can be processed synchronously or asynchronously
  - Big Data storage infrastructure must reduce latency

Big Data Operations (2-5)

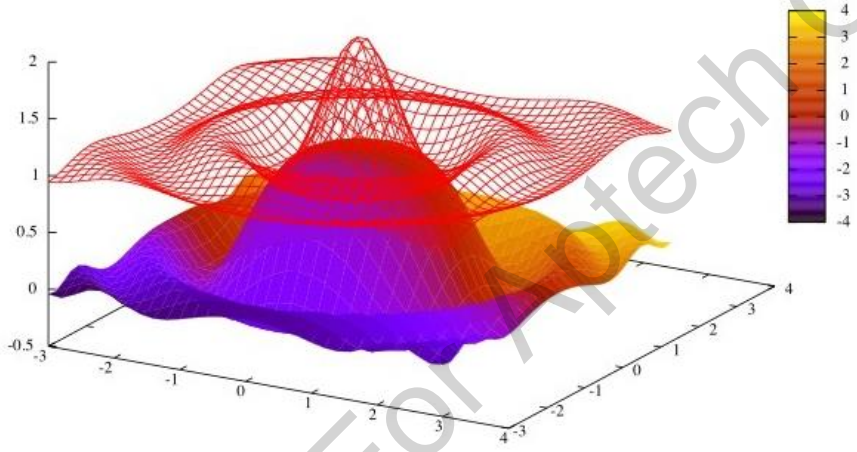
- Data Visualization
  - Is the process of representing data in the form of visuals.
  - Provides an easy and a simple way to convey complex data
  - Allows to expose the patterns, trends, and correlations
  - Displays data as:

Infographics	Dials and Gauges	Geographic Maps	Sparklines	Heat Maps	Bar and Pie Charts

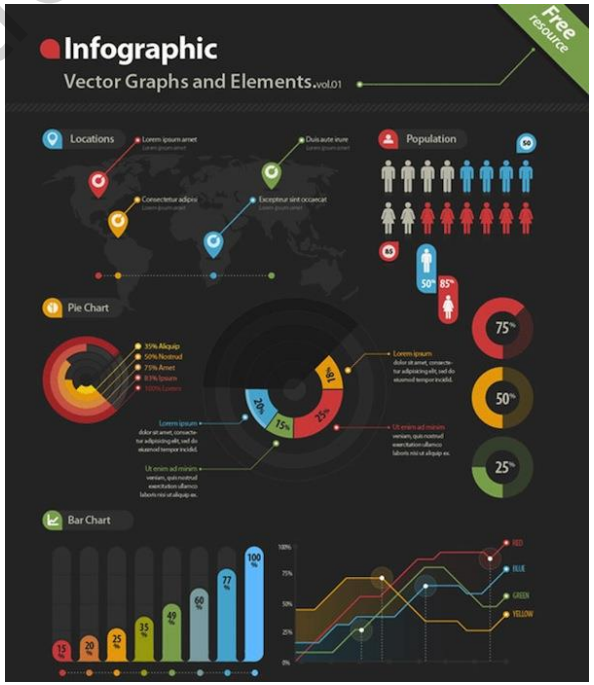
Big Data Operations (3-5)

Examples of Data Visualization

Heat maps



Infographics



Big Data Operations (4-5)

Data Language

- Programming languages are best to handle large amount of data
- Some programming languages in Big Data are R, Python, and Julia

Using Multiple Languages

- Sometimes many languages may be required to obtain necessary results
- R language is used to run large number of calculations
- Python is helpful for advanced analytics
- Julia is used by projects involve database analytics on real-time streams

Big Data Operations (5-5)

- Data Extraction is a means of collecting the unstructured data from various sources:
  - **Import.io** provides a platform used to convert Webpages or Websites into a structured, machine-readable data
  - It does not require any coding knowledge
  - It is also used to convert a Webpage into spreadsheets

Summary

- Internal data streams, shared data streams, and external data streams are the three sources of Big Data in general.
- Big Data management is the process of organizing, administrating, and governing of large amount of structured and unstructured data.
- Big Data can be processed synchronously or asynchronously.
- Data Visualization is the process of representing data in the form of visuals.
- Data Extraction is a means of collecting unstructured data from various sources.