Northeastern University

CS6020: Collecting, Storing, and Retrieving Information Systems

Introduction & Essential Concepts

Lesson 2

CHARACTERISTIC AND CHALLENGES OF "BIG DATA"

Lesson Objectives

- After completing this lesson, you are able to:
  - provide a definition of “big data”
  - list common sources of data that result in uncommonly large unstructured data sets
  - state the characteristics of “big data” sets
  - explain the key challenges when working with “big data”

Big Data

- Big Data is a relatively new term that describes data set that are so large and complex that traditional methods of storing and processing them are not sufficient.
- Larger data sets allow for more detailed analysis and application to social sciences, biology, pharmacology, business, marketing, and more.

Volume of Data is Growing

- Each day over 2.5 quintillion bytes of data is being generated[1].
- 90% of the world's data has been generated over the past two years.
- Data from multiple sources is being integrated into single massive data sets.

1. http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

Sources of Data

- Web Behavior and Content:
  - There are nearly five billion web pages
  - Collected data includes network traffic, site and page visits, page navigation, page searches
- User Generated Content:
  - Content generated by millions of users on social media, including Facebook, Twitter, Instagram, blogs, YouTube, forums, wikis, and so forth

Some Quick Statistics

- Big Data = Big Velocity
- Every 60 seconds there are
  - Over 100,000 tweets
  - 695,000 Facebook status updates
  - 11 million instant messages
  - 700,000+ Google searches
  - 168 million+ emails sent
  - 1,820TB of data created
  - 217 new mobile web users

More Data Sources

- RFID Data:
  - Radio Frequency Identifiers
  - Tags for tracking merchandise and shipments, mobile payments, sports performance measurement, and automated toll collection
- Geo Data:
  - GPS tracking data generated by mobile devices
  - Tracking of movement of equipment, vehicles, and people

Even More Data Sources

- Environmental Data:
  - Weather conditions
  - Tidal movements
  - Seismic activity
- Organizational Transactional Data:
  - Transactional activities such as purchases, registration, manufacturing

And Some More Data Sources…

- Research Data:
  - Social science data, *e.g.*, census, polls
  - Health care data
  - Education, law and order, economic activity, agriculture, food production
  - “Big Data” such as radio telescopes, particle physics

Definition of “Big Data”

- While there is no single agreed upon definition of “Big Data”, here is one possible definition:

> Big Data is the integration of large amounts of multiple types of structured and unstructured data into a single data set that can be analyzed to gain insight and new understanding of an industry, business, the environment, medicine, disease control, science, and the human interactions and expectations.

What's Being Said About "Big Data"

To define big data in competitive terms, you must think about what it takes to compete in the business world. Big data is traditionally characterized as a rushing river: large amounts of data flowing at a rapid pace. To be competitive with customers, big data creates products which are valuable and unique. To be competitive with suppliers, big data is freely available with no obligations or constraints. To be competitive with new entrants, big data is difficult for newcomers to try. To be competitive with substitutes, big data creates products which preclude other products from satisfying the same need."

Cited from John Weathington on the TechRepublic Blog

Get more quotes from http://www.opentracker.net/article/definitions-big-data

Big Data Challenges

- The challenges of big data include:
  - Analysis
  - Collection
  - Storage
  - Curation
  - Search and retrieval
  - Sharing
  - Transfer
  - Visualization
  - Privacy

Data Storage Technologies

- Traditional data storage technologies including text files, XML, and relational databases reach their limits when used to store very large amounts of data.
- Furthermore, the data that is needed for analysis includes not only text and numeric data, but unstructured data, such as text files, video, audio, blogs, sensor data, geospatial data, among others.

Examples of "Big Data"

- The Large Hadron Collider would generate 5×10^{20} bytes per day if all of its sensors were turned on, almost 200 times more than all other data sources in the world combined.
- The Square Kilometer Array radio telescope is expected to collect 14 *exabytes* of data per day for analysis
- Walmart generates over 1 million customer transactions per hour that are curated in a multi-petabyte database for trend analysis

Big Data Analysis

- Big data requires complex analysis within relatively short time spans in order to detect trends and make decisions.
- Analysis techniques include, among many others:
  - A/B Testing
  - Visualization
  - Machine Learning
  - Time Series Analysis

Big Data Characteristics

- Big Data generally exhibits the following characteristics:
  - Very large, distributed aggregations of loosely structured data – often incomplete
  - In excess of multiple petabytes or exabytes of data
  - Billions of records about people or transactions
  - Loosely-structured and often distributed data
  - Flat schemas with few complex interrelationships
  - Time series data containing time-stamped events
  - Connections between data elements that must be probabilistically inferred through machine learning

Adapted from Wikibon.org

Information Quality

- Information (or data) quality is a measurement of the fit of information for a particular use.
- Poor information quality can be costly:
  - One study estimates that on average bad information costs businesses up to 10% of revenue
  - Another study pegs the loss at over $600 billion annually in the U.S. alone

Summary

- In this lesson you learned that:
  - big data is a somewhat nebulous term that describes data sets that are too large and complex to analyze by traditional means
  - the challenges of big data include collection, storage, retrieval, curation, analysis, transfer, and visualization
  - there are numerous sources generating extremely large and complex data sets

Summary, Review, & Questions…