Northeastern University

CS6020: Collecting, Storing, and Retrieving Information Systems

Introduction & Essential Concepts

Lesson 3

THE SIX V'S OF BIG DATA

Lesson Objectives

- After completing this lesson, you are able to:
  - provide a definition of the Six V's of volume, variety, velocity, veracity, validity, volatility
  - explain the Six V's and how they influence the analysis of data sets

The 6 *V's* of Big Data

- IT, business, and data scientists need to understand the 6 V's based on Gartner's original 3 V's (Normadeau, 2013):
  - Volume
  - Variety
  - Velocity
  - Veracity
  - Validity
  - Volatility

Volume

- Volume is one of the core defining attributes of “big data”.
- Big Data implies enormous amounts of structured and unstructured data that is generated by social and sensor networks, transaction and search history, and manual data collection.

Variety

- Data comes from a variety of sources and contains both structured and unstructured data.
- Data types are not restricted to simply numbers and short text fields, but also include images, emails, text messages, web pages, blog entries, documents, audio, video, and time series.

Velocity

- The flow of data that needs to be stored and analyzed is continuous.
- Human interactions, business processes, machines, and networks generate data continuously and in enormous quantity.
- The data is generally analyzed in real-time to gain a strategic advantage.
- Sampling can help mitigate some of the problems with large data volume and velocity.

Veracity

- Data veracity characterizes the inherent noise, biases, abnormalities, and mistakes present in virtually all data streams.
- “Dirty” data presents a significant risk as analyzes are incorrect when based on “bad” data.
- Data must be cleaned in real-time and processes must be established to keep “dirty data” from accumulating.

Validity

- While the data may not be “dirty”, biased, or abnormal and it may not be valid for the intended use.
- Valid data for the intended use is essential to making decisions based on the data.

Volatility

- Data changes over time and volatility characterizes the degree to which the data changes over time.
- Decisions and analyses are based on data that has an “expiration date”.
- Data scientists must define at what point in time a data stream is no longer relevant and cannot be used to make a decision.

Learning Checkpoint

- *Concepts Pharma has built a data repository in which it collects self-reported eating habits of clinical trial participants through a mobile app. The translational medicine group is using the data to determine if the drug in trial is causing digestive issues when taken with certain food groups. Which of the V's should be of most concern to them?*

  A. Veracity

  B. Volume

  C. Volatility

  D. Velocity

  E. Variety

Learning Checkpoint

- *Concepts Pharma has built a data repository that collects self-reported eating habits of clinical trial participants through a mobile habit. The translation medicine group is using the data to determine if the drug in trial is causing digestive issues when taken with certain food groups. Which of the V's should be of most concern to them?*

  A. Veracity

  B. Volume

  C. Volatility

  D. Velocity

  E. Variety

Summary

- In this lesson you learned that:
  - data scientists need to understand the Six V's of big data to ensure sound analysis and decision making
  - The Six V's are based on the original Three V's proposed by Gartner

Summary, Review, & Questions…