Northeastern University

CS6020: Collecting, Storing, and Retrieving Information Systems

Introduction & Essential Concepts

Lesson 4

PLANNING A BIG DATA PROJECT

Lesson Objectives

- After completing this lesson, you are able to:
  - outline the planning of a big data project
  - express the major issues relating to governance and privacy

Planning a Big Data Project

- Planning a big data project requires that a data scientist or data analyst understand:
  - Objectives
  - Data
  - Processes
  - Infrastructure
  - Analytics

Understanding “Objectives”

- What is the purpose of the data project?
- How is the data going to be used?
- What is the business or organizational value of the data project?

Understanding "Data"

- What data needs to be collected?
- Where will the data come from?
  - Internal systems?
  - Social networks?
  - External data sources?
- What is the structure of the data?
  - Quantitative or qualitative?
- What is the quality of the data?

Understanding “Processes”

- How will it be collected?
- Who is involved in collection of the data?
- How will the data be cleaned?
- How will the data be loaded and transferred?

Understanding “Infrastructure”

- Where will the data be stored?
- What database or data store will be needed based on the volume, complexity, type, and required access of the data?
- What hardware is needed to support responsive access to the data?
- Who will manage the data store?
- Who will supply the data store?

Understanding "Analytics"

- How will the data be presented?
  - Tables?
  - Visualizations?
- What predictive models will be built?
- How will the data from different sources be combined?
- What skills are needed to do the analysis?
- What programs or applications need to be built or purchased?

Governance and Privacy

- Organizations must be transparent in how they manage personal data and how they use it.
- Government regulations may limit which data can be collected and how that data can be stored, transferred, or used.
- Organizations must protect private data and not allow persons to be “identifiable”.

Summary

- In this lesson you learned that:
  - planning a big data project requires an understanding of the objectives, the characteristics of the data, the process of collecting the data, how the data will be stored, and how the data will be presented and analyzed
  - data that contains identifiable attributes must be kept confidential

Summary, Review, & Questions…