Introduction to Big Data Analytics

Big Data Defined

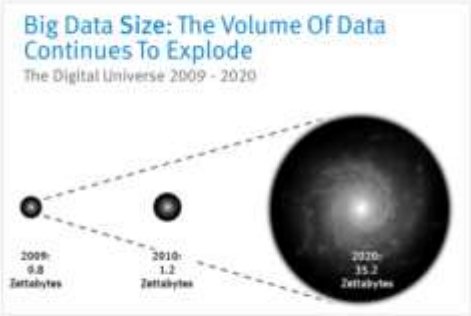
- There are multiple characteristics of big data, but 3 stand out as defining Characteristics:
  - **Huge volume of data** (for instance, tools that can manage billions of rows and billions of columns)
  - **Complexity of data types and structures**, with an increasing volume of unstructured data (80-90% of the data in existence is unstructured)….part of the Digital Shadow or "Data Exhaust"
  - **Speed or velocity of new data creation**

Question?

- What would be considered "Big Data"?
  - A. An OLAP Cube containing customer demographic information about 100,000,000 customers
  - B. Daily Log files from a web server that receives 100,000 hits per minute
  - C. Aggregated statistical data stored in a relational database table
  - D. Spreadsheets containing monthly sales data for a Global 100 corporation

Key Characteristics of Big Data

1. **Data Volume**
   - 44x increase from 2010 to 2020 (1.2zettabytes to 35.2zb)



2. **Processing Complexity**
   - Changing data structures
   - Use cases warranting additional transformations and analytical techniques

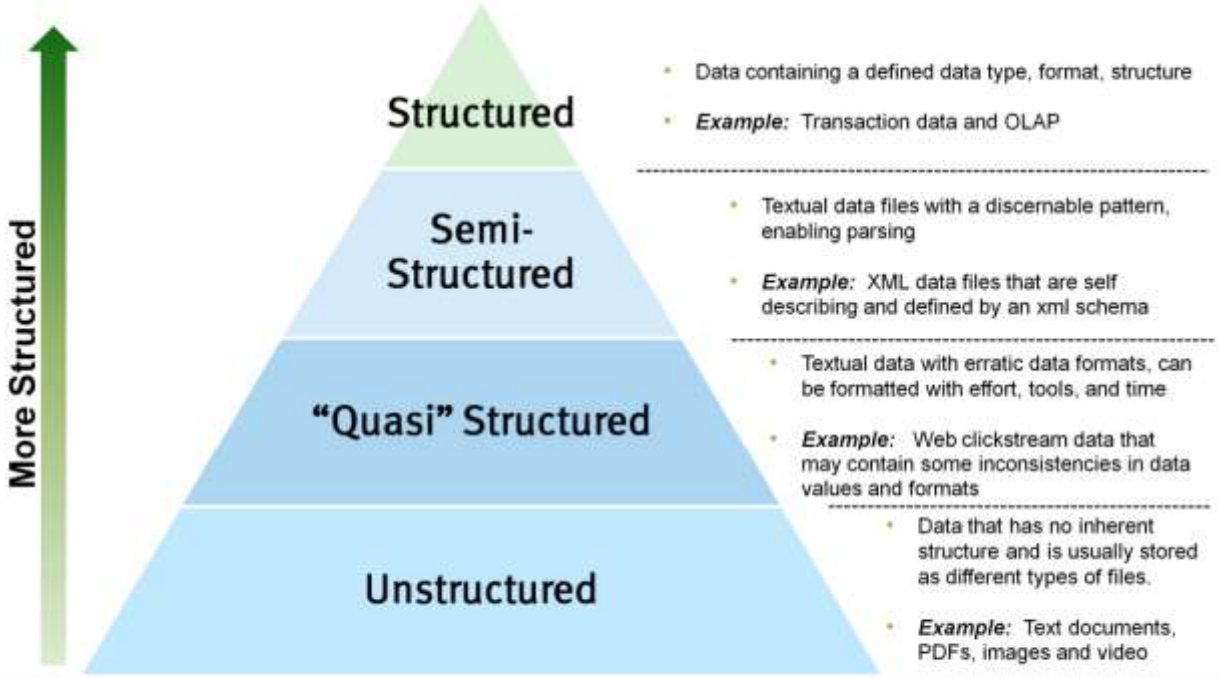
3. **Data Structure**
   - Greater variety of data structures to mine and analyze

Question?

- What are the characteristics of Big Data?
  - A. Data volume, processing complexity, and data structure variety.
  - B. Data volume, business importance, and data structure variety.
  - C. Data type, processing complexity, and data structure variety.
  - D. Data volume, processing complexity, and business importance.

Big Data Characteristics: Data Structures

Data Growth is Increasingly Unstructured



Question?

- Which data asset is an example of quasi-structured data?
  - A. Webserver log
  - B. XML data file
  - C. Database table
  - D. News article

Question?

- Which word or phrase completes the statement?

  Structured data is to OLAP data as quasi-structured data is to _____
- A. Clickstream data
- B. XML data
- C. Text documents
- D. Image files

Question?

- Which data asset is an example of semi-structured data?
  - A. XML data file
  - B. Database table
  - C. Webserver log
  - D. News article

Data Repositories, An Analyst Perspective

Data Islands
"Spreadmarts"

Isolated data marts

- Spreadsheets and low-volume DB's for recordkeeping
- Analyst dependent on data extracts

Data Warehouses

Centralized data containers in a purpose-built space

- Supports BI and reporting, but restricts robust analyses
- Analyst dependent on IT & DBAs for data access and schema changes
- Analysts must spend significant time to get extracts from multiple sources

Analytic Sandbox

Data assets gathered from multiple sources and technologies for analysis

- Enables high performance analytics using in-db processing
- Reduces costs associated with data replication into "shadow" file systems
- *"Analyst-owned" rather than "DBA owned"*

Question?

- Which word or phrase completes the statement?

  A spreadsheet is to a data island as a centralized database for reporting is to a _____?

  - A. Data Warehouse
  - B. Data Repository
  - C. Analytic Sandbox
  - D. Data Mart

Question?

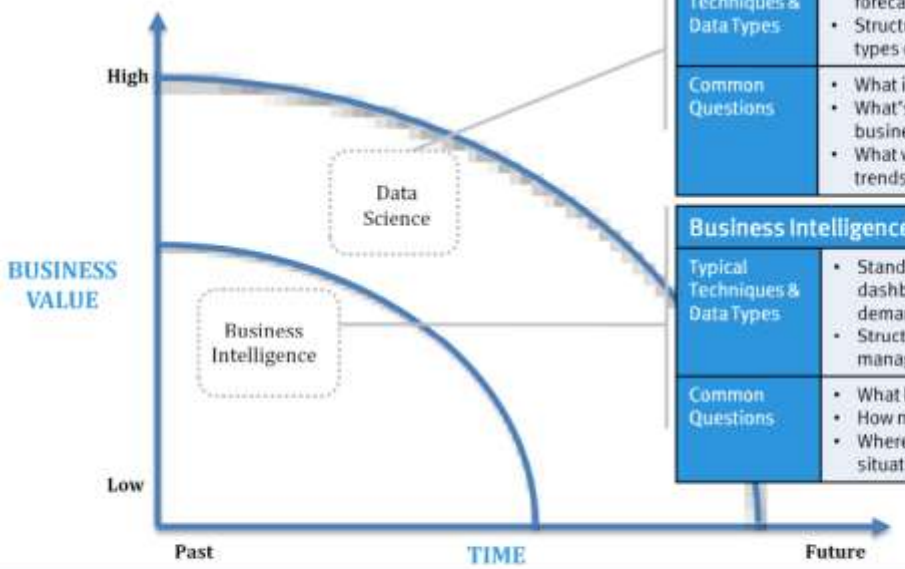
- Which word or phrase completes the statement?

  A data warehouse is to a centralized database for reporting as an analytic sandbox is to a _____?

  - A. Collection of data assets for modeling
  - B. Collection of low-volume databases
  - C. Centralized database of KPIs
  - D. Collection of data assets for ETL

Analytical Approaches for Meeting Business Drivers

Business Intelligence vs. Data Science



Predictive Analytics & Data Mining (Data Science)	
Typical Techniques & Data Types	• Optimization, predictive modeling, forecasting, statistical analysis • Structured/unstructured data, many types of sources, very large data sets
Common Questions	• What if.....? • What's the optimal scenario for our business? • What will happen next? What if these trends continue? Why is this happening?

Business Intelligence	
Typical Techniques & Data Types	• Standard and ad hoc reporting, dashboards, alerts, queries, details on demand • Structured data, traditional sources, manageable data sets
Common Questions	• What happened last quarter? • How many did we sell? • Where is the problem? In which situations?

EMC² PROVEN PROFESSIONAL

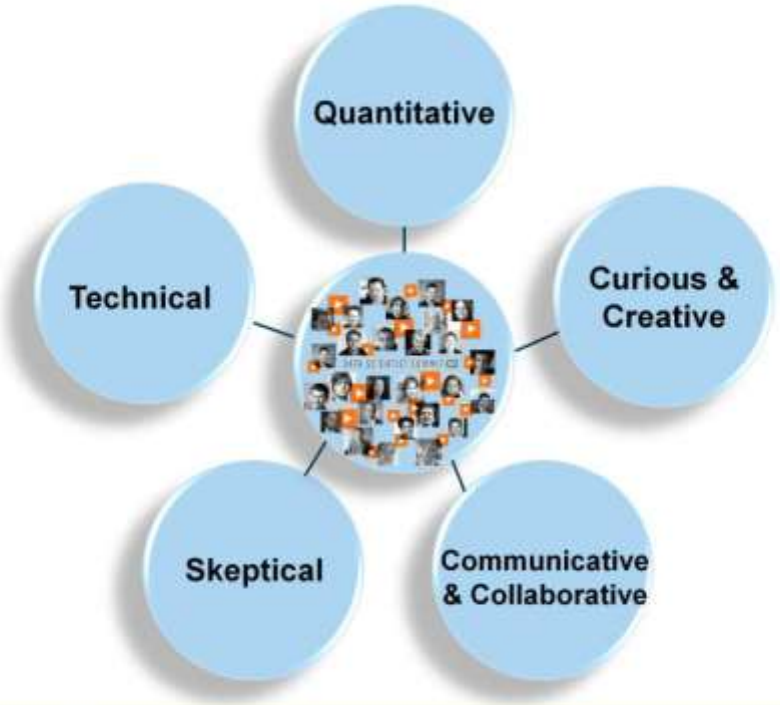
Question?

- Which word or phrase completes the statement?

  Business Intelligence is to monitoring trends as Data Science is to _____ trends.

  - A. Predicting
  - B. Discarding
  - C. Driving
  - D. Optimizing

Profile of a Data Scientist



Question?

- Which word or phrase completes the statement?

  Theater actor is to "Artistic and Expressive" as Data Scientist is to

  _____

  - A. "Communicative and Collaborative"
  - B. "Introverted and Technical"
  - C. "Logical and Steadfast"
  - D. "Independent and Intelligent"