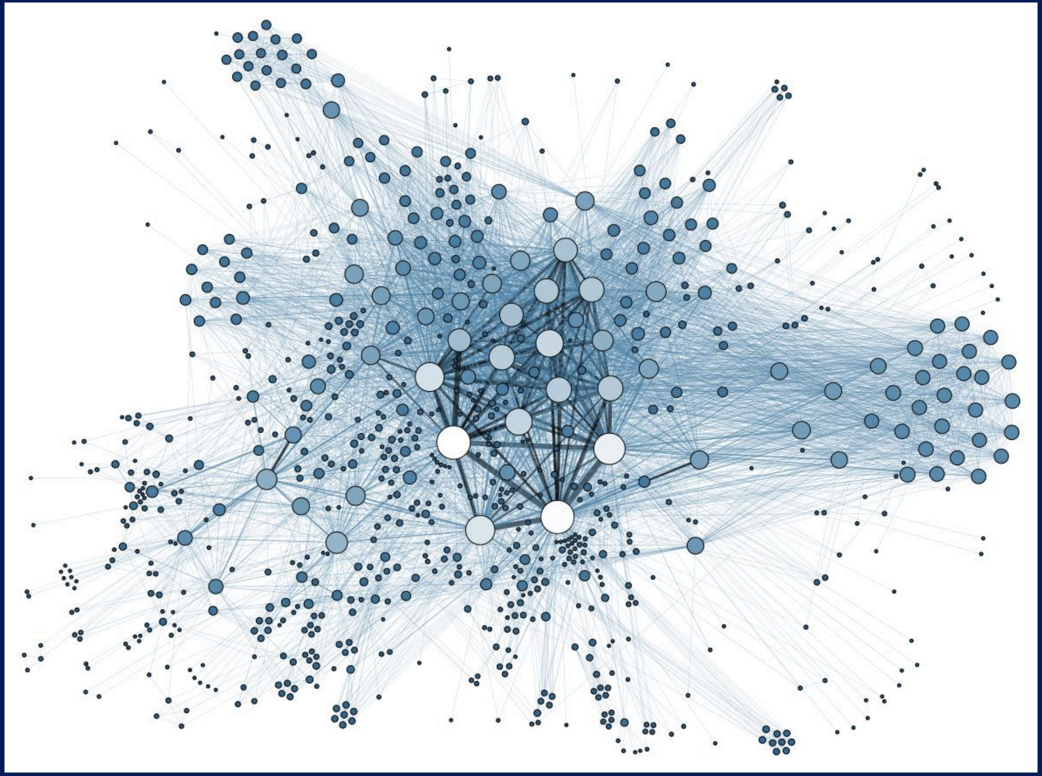
Summary of Big Data Modeling and Management



After this video you will be able to..

- Recall why big data modeling and management is essential in preparing to gain insights from your data
- Summarize different kinds of data models
- Describe streaming data and the different challenges it presents
- Explain the differences between a DBMS and a BDMS

Big Data Modeling and Management

- Data modeling tells you
  - How your data is structured
  - What operations can be done on the data
  - What constraints apply to the data

data model <
structure
operations
constraints

- Database Management Systems
  - Typically handle many low-level details of data storage, manipulation, retrieval, transactional updates, failure and security
  - Relieves a user to focus on higher level operations like querying and analysis

Different Data Models

- ① Relational Data
  - Where data look like tables (relations)
- ② Semi-structured Data * Tree
  - Document data, XML and JSON (embede
- ③ Graph Data nodes — entities
  - Social Networks, email networks edges — relation
- ④ Text Data
  - Articles, reports * primary in search engines

Streaming Data

stream → infinite data source

- An infinite flow of data coming from a data source
  - Sensor data from instruments
  - Stock price data
- Data rates vary – can be too fast and too large to store

  * needs different kind of management system
- Often processed in memory — in chunks : windows
- May need to be processed immediately
  - Inform whenever 3 tech stocks go up by 3% within a 30 second span
  - Used for event detection and prediction

typical type of query : alerts or notifications

DBMS and BDMS

- BDMS (with different data models and different capabilities
  - Designed for parallel and distributed processing
    - Data-partitioned parallelism : the process of segmenting the data into multiple machines
      (parallel data retrieval and operations)
  - May not always guarantee consistency for every update
    not at every moment, sooner or later
    - More likely to guarantee eventual consistency
  - Often built-on Hadoop
    - Offer Map-reduce style computation
    - Utilizes replication natively offered by HDFS