Big Data Management: Past, Present, and Future

Kelly Magruder
What is Big Data?

• Extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations
Need for Management Systems

• “Every two days now we create as much information as we did from the dawn of civilization up until 2003. That’s something like five exabytes of data.”
  - Google CEO Erik Schmidt in 2010 (even larger now)

• Gigabyte = $10^9 = 1,000,000,000$
  Terabyte = $10^{12} = 1,000,000,000,000$
  Petabyte = $10^{15} = 1,000,000,000,000,000$
  Exabyte = $10^{18} = 1,000,000,000,000,000,000$
Management System Goals

• The Four V’s of Big Data:
  • Volume – Scale of data
  • Variety – Different forms of data
  • Velocity – Analysis of streaming data
  • Veracity – Uncertainty and accuracy of data
First Big Data Problem and Solution

• 1880 Census took eight years to tabulate 50 million people
• The Hollerith Tabulating Machine punched cards with 80 variables and completed the 1890 Census in 6 weeks
Information Overload

- Population boom in the 1930’s
  - Census, social security numbers, new research
- Libraries adapted new storage methods to meet demand of publications
- In 1944, libraries were predicted to double every sixteen years
Centralized Computing Systems

• 1960s: Information and research had reached the point of critical magnitude (Information Scientist Derek Price)

• Organizations developed electronic automation of their inventory systems
The Relational Database

• 1970: IBM Research Lab published a paper showing how information in large databases could be easily accessed
• Modern data transactions all use structures based on relational database theory
The Relational Database

- Possible schema of Twitter’s database
- Three tables: users, tweets, following
- Each table has several attributes
Transitioning to Modern Data Management

• Mid 1990s: Content rapidly generated in the wake of the world-wide Web’s impact
• Business Intelligence storage transitioned from paper to digital to reduce cost
MapReduce Programming Models

- Google pioneered divide-and-conquer parallelism based on hashing contiguous bytes
- Apache Hadoop is open source, cheap, and scalable, and remains popular today
- Community developed high-level declarative languages for queries and data analysis (Pig, Jaql, Hive)
Parallel Database Systems

• Instead of hashed abstractions, the core file system is relational tables and is more object-oriented

• Users only able to access the outermost layer of SQL as most commercial options are closed-source
Problems in Big Data Management

• Low-latency
• High scalability grow / shrink elastically with business demand
• Support push, pull, transactional, and analytic based workloads
• Cost-effective
• Fault and failure tolerant
The NoSQL Movement

- **Not Only SQL** departs from the relational model
- Key-value, graph, or document data structures
- Open source, more versatile, scales better in some cases

- **Key-Value Stores**
  - Dynamo (Amazon), Voldemort (LinkedIn)

- **Big Table Clones**
  - BigTable (Google), Cassandra (Facebook)

- **Document Databases**
  - CouchOne, MongoDB

- **Graph Databases**
  - FlockDB (Twitter), AllegroGraph
NoSQL

Writes are fully parallel and is not blocked by Node communication.

Node 1 → Node 2
Read Entity 1

Node 2 → Client 2
Write 20 Entities

Client 1 → Node 1

RDBMS

Replication, locking and consistency check during two phase commit

Node 1 → Node 2
Read Entity 1

Node 2 → Client 2
Write 20 Entities

Client 1 → Node 1
Future of Big Data Management

Ultimately, depends on business preference, capabilities for implementation, and specific data management needs

• NoSQL
• Open Source
• High-level query languages
Thank you!

Questions?