Evolution of Information Systems

Yanlei Diao
University of Massachusetts Amherst
Information Technology has been evolving rapidly
In 50 years, computers will be intelligent, with a storage capacity of about $10^9$.

--- Alan Turing, 1950
When processing, storage, and transmission cost micro-dollars, then the only real value is the data and its organization.

-- James Gray, 1998 Turing Award Speech
Era of BIG DATA

- 1950: Turing’s Vision
- 1970: Codd: Relational Model & DBMS
- 1998: Gray: Turing Award

New Millennium
Era of BIG DATA

- Sensors
- Smart devices
- Sequencers
- Large simulations
- Social software
- ...

New Millennium

1950
Turing's Vision

1970
Codd: Relational Model & DBMS

1998
Gray: Turing Award

2003
Cluster Computing

2006
Cloud Computing: Hadoop

2009
Spark

- Cluster computing
- Cloud computing
- Open source

Infra-structure
Objectives of This Course

Fundamental **Design Principles** and **Key Algorithms** for Influential Software Systems for Big Data Analysis
Main Topics

1. Relational Database Management Systems (RDBMS)
   Apps: banking, ticketing, electronic commerce…

- Data Modeling and Query Languages
  - Relational model, relational algebra/calculus, SQL, theory
- Query Processing
  - Algorithms for sorting, selection, join, group by-aggregation.
- Query optimization
  - System R style, multi-objective, parametric optimization
- Transaction processing: consistency and recovery
Main Topics (cont’d)

2. Data Warehouses

Apps: retail management, fraud detection, decision analytics…

- Online Analytic Processing (OLAP)
  - Large queries over (read-only) historical data
  - OLAP operations
  - New query processing and optimization techniques
  - Data mining algorithms
Main Topics (cont’d)

3. Parallel Databases & MapReduce Systems

Apps: retail management, web data analysis, social networks…

- Parallel databases (90’s)
  - Programming model: SQL
  - Data-parallel algorithms for relational operators

- MapReduce systems (circa 2003-2004)
  - Programming model: functional
  - MapReduce under data parallelism
  - Scheduling and fault tolerance

- Distributed storage
  - HDFS, key-value stores, column stores
Main Topics (cont’d)

4. Systems for Iterative Computation and Machine Learning
   Apps: machine learning, graph analytics…

- Spark (2009-)
  - Programming model: RDDs and data frames
  - Innovation in fault tolerance
  - Unified programming interface for SQL, ML, graphs, and stream workloads
Three Forms of Parallelism

Data Parallelism

Pipeline Parallelism

Model Parallelism