Algorithms and Systems on big data management

Lecturer: Jiaheng Lu

Fall 2016
We are in the era of big data

- Lots of data is being collected
  - Web data, e-commerce
  - Bank/Credit Card transactions
  - Social Network
  - Scientific data
Data sizes

- Byte (B)
- Kilobyte (KB)
- Megabyte (MB)
- Gigabyte (GB)
- Terabyte (TB)
- Petabyte (PB)
- Exabyte (EB)
- Zettabyte (ZB)
- Yottabyte (YB)
How much data?

- Google processes 100 PB a day, 3 millions servers
- Facebook has 300 PB of user data + 500 TB/day
- Youtube has 1000 PB video storage
- SMS messages 6.1 TB per year
- US Credit card: 1.4B cards, 20B transaction/year

- In 2009, total data is about 1ZB, in 2020, it is estimated to be 35ZB.
Type of Data

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Image data, audio data, video data
- Graph Data
  - Social Network, Semantic Web (RDF), …
- XML and JSON data
Four V’s
• Watch two videos about big data
Outline

• About the course
• Practical information and requirement
• Course topics
• Our schedule
At the end of the course

- You should be able to tell what these terms stand for!
  And more…

Hadoop Mapreduce

Bigtable

NOSQL databases

Count-min, FM sketch

Volume, Velocity, Variety

Data streaming

XML and JSON

Data model
After this course

- Students are expected to
  - Select a data model to suit the characteristics of your data;
  - Understand sketch techniques to handle streaming data, including Count-Min, Count-Sketch and FM Sketch;
  - Understand GFS, MapReduce and Bigtable technology
  - Hands-on experience on Hadoop MapReduce
Workflow of this course

Exercise 1
15 November

Exercise 2
29 November

Exercise 3 (one programming task)
13 December

Study group 1
9 November

Study group 2
23 November

Final Examination
Topics of this course

- Introduction to gig data and data models (2 weeks)
- NoSQL databases (1 week)
- Sketches algorithms (2 weeks)
- GFS, Mapreduce and Bigtable (2 weeks)
Introduction to big data engineering

- Five steps in Big data engineering
  - 1. Acquire data
  - 2. Prepare data
  - 3. Analyze data
  - 4. Report data
  - 5. Act
Step 1 Acquire data

- Identify data set
- Retrieve data or buy data
Step 2. Prepare data

- 2.1 Explore data
  - Understand the nature of data
  - Preliminary analysis

- 2.2 Preprocess data
  - Clean, integrate and package
Step 3. Analyze data

- Select analytical techniques
- Build model
Step 4. Communication results

- Visualization and summary
Step 5. Apply results

- The above five steps are iterative process
Key technical challenges for big data

1. Enable scalability
   - Commodity hardware is cheap

2. Handle fault tolerance
   - Be ready, crashes happen

3. Optimize performance

4. Provide values
What is Hadoop?

- Apache top level project, open-source implementation of frameworks for reliable, scalable, distributed computing and data storage.
Hadoop’s Developers

2005: Doug Cutting and Michael J. Cafarella developed Hadoop to support distribution for the Nutch search engine project.

The project was funded by Yahoo.

2006: Yahoo gave the project to Apache Software Foundation.
Google Origins

The Google File System
Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung
Google

MapReduce: Simplified Data Processing on Large Clusters
Jeffrey Dean and Sanjay Ghemawat
jeff@google.com, sanjay@google.com
Google, Inc.

Bigtable: A Distributed Storage System for Structured Data
Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Holtz, Deborah A. Wobson, Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber
Bigtable@Google.com
Google, Inc.

Abstract
Bigtable is a distributed storage system for managing structured data that is designed to scale to a very large number of data across thousands of commodity servers. Many projects in Google store data in Bigtable following web indices, Google Earth, and Google Fi
et. These applications place very different demands on scalability, both in terms of data size and number of concurrent clients. Bigtable is optimized for applications that require high throughput. The system provides scalable, durable, and consistent access to structured data. It uses a row-oriented key-value store to store and manage its data. Data is stored in table columns and accessed through a simple, powerful interface that includes methods to search, update, and scan the stored data. The system is designed to handle a large number of clients and to maintain high performance under heavy load. The Bigtable system is implemented in C++ and is distributed under the Apache License. The source code is available at http://www.apache.org/licenses/.

www.helsinki.fi
Some Hadoop Milestones

- **2008 - Hadoop Wins Terabyte Sort Benchmark** (sorted 1 terabyte of data in 209 seconds, compared to previous record of 297 seconds)

- **2010 - Hadoop's Hbase, Hive and Pig subprojects completed**, adding more computational power to Hadoop framework

- **2013 - Hadoop 1.1.2 and Hadoop 2.0.3 alpha.**
  - Ambari, Cassandra, Mahout have been added

- **2016 - Hadoop 3.0.0 Alpha-1**
Hadoop File System was developed using distributed file system design. It is run on commodity hardware. Unlike other distributed systems, HDFS is highly fault tolerant and designed using low-cost hardware.
YARN is the architectural center of Hadoop that allows multiple data processing engines such as interactive SQL, real-time streaming, data science and batch processing to handle data stored in a single platform.
Apache HBase™

Apache HBase™ is the Hadoop database, a distributed, scalable, big data store.
The **Apache Mahout™** project's goal is to build an environment for quickly creating scalable performant machine learning applications.
Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs.
The Apache Hive™ data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL.
Apache Spark™ is a fast and general engine for large-scale data processing.
Oozie is a workflow scheduler system to manage Apache Hadoop jobs. Oozie Workflow jobs are Directed Acyclical Graphs (DAGs) of actions.
ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services.
Apache Sqoop is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases.
Apache™ Storm adds reliable real-time data processing capabilities to Enterprise Hadoop. Storm on YARN is powerful for scenarios requiring real-time analytics, machine learning and continuous monitoring of operations.
The **Apache Ambari** project is aimed at making Hadoop management simpler by developing software for provisioning, managing, and monitoring Apache Hadoop clusters.
Applications of big data management

- Big data management on gaming industry
  - 7 minutes Video

- Big data management on flight industry
  - 13 minutes video
Wrap-up

- We are in the era of big data
- 4Vs of Big data: Volume, Variety, Velocity and Veracity
- There are five steps in big data engineering, including Acquire data, Prepare data, Analyze data, Report data and Act
- Hadoop is an open-source implementation of frameworks for reliable, scalable, distributed computing and big data storage.