Introduction to Data Science

CS 5665
Utah State University
Department of Computer Science
Instructor: Prof. Kyumin Lee

MapReduce

Example 2: Language model

- Statistical machine translation
  - Need to count number of times every 5-word sequence occurs in a large collection of documents
- Solution
  - Map(doc_id, document) => [(5-word seq, count),…]
  - Reduce(5-word seq, [count1, …]) => (5-word seq, sum([count1, …]))

Example 3: Reverse Web-Link Graph

- Determine in-coming links (Page rank)
- Solution:
  - Map(src_page_url, page_html) => [(link1, src_page_url), …]
  - Reduce(link, [src_page_url1, …]) => (link, [src_page_url1, …])

Introduction to Data Science

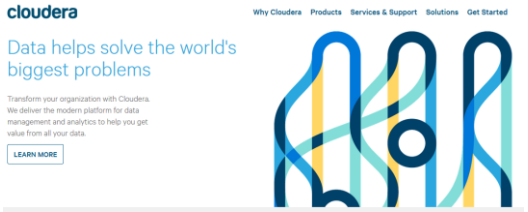
CS 5665
Utah State University
Department of Computer Science
Instructor: Prof. Kyumin Lee

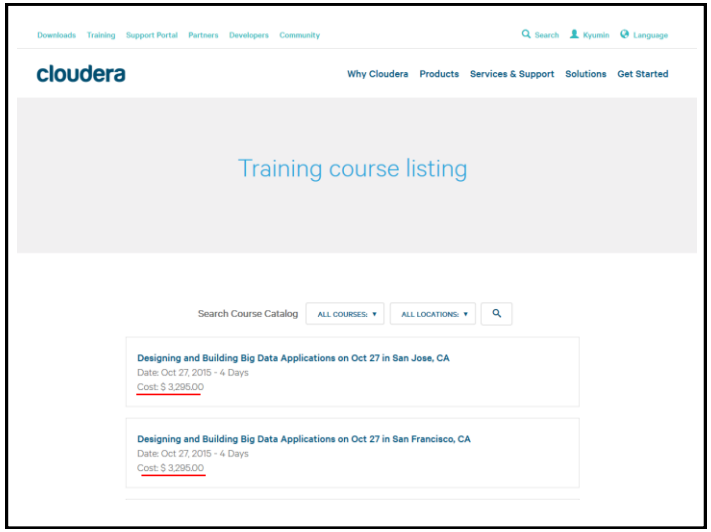
Some Slides were adapted from Cloudera
DO NOT SHARE THE SLIDES NOR UPLOAD THEM IN PUBLIC

cloudera ACADEMIC PARTNER

Cloudera

- **Cloudera Inc.** is an American-based software company that provides Apache Hadoop-based software, support and services, and training to business customers.

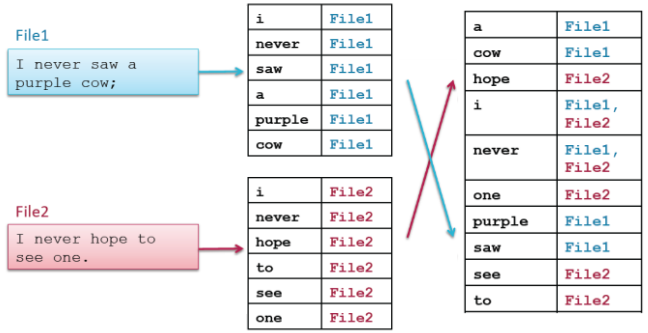




Example: Indexing

- Assume the input is a set of files containing lines of text
- Mapper:
  - For each word in the line, emit (*word*, *filename*)
- Reducer:
  - Collect together all values for a given key (i.e., all filenames for a particular word)
  - Emit (*word*, *filename_list*)

Inverted Index: Dataflow



Example: A Chain of Three MapReduce Jobs

- Given one million tweets, find similar pairs (say, similarity is over 0.7)
  - 1,000,000 * 999,999 / 2 = 499,999,500,000 pairs

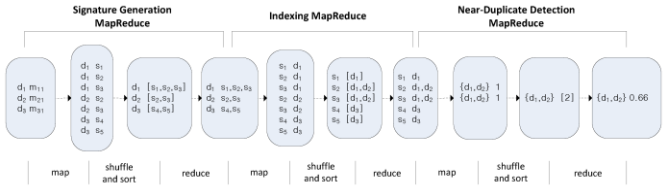
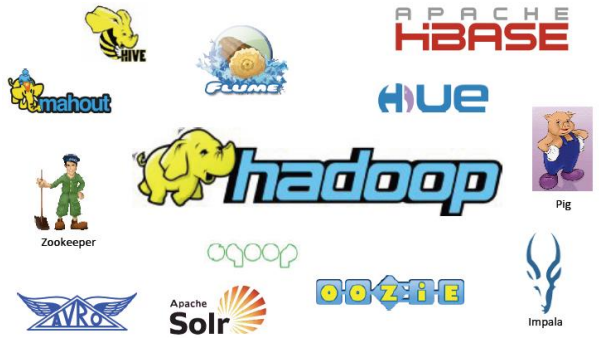


Fig. Logical data flow of the three MapReduce jobs for identifying correlated messages

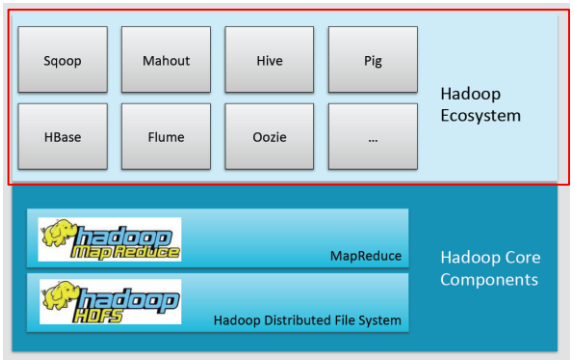
K. Lee, J. Caverlee, Z. Cheng, D. Z. Sui. Campaign Extraction from Social Media. ACM Transactions on Intelligent Systems and Technology (ACM TIST), Vol. 5, No. 1, January 2014.

The Hadoop Ecosystem

A Large (and Growing) Ecosystem



Hadoop Components



The Hadoop Ecosystem

- Ecosystem projects may be
  - Built on HDFS and MapReduce
  - Built on just HDFS
  - Designed to integrate with or support Hadoop
- Most are Apache projects or Apache Incubator projects
  - Some others are not managed by the Apache Software Foundation
  - These are open hosted on GitHub or a similar repository
- Following is an introduction to some of the most significant projects

[Data Storage] HBase

- HBase is the Hadoop database
- A 'NoSQL' datastore
- Can store massive amounts of data
  - Petabytes+
- High write throughput
  - Scales to hundreds of thousands of inserts per second
- Handles sparse data well
  - No wasted spaces for empty columns in a row
- Limited access model
  - Optimized for lookup of a row by key rather than full queries
  - No transactions: single row operations only
  - Only one column (the 'row key') is indexed

When To Use HBase

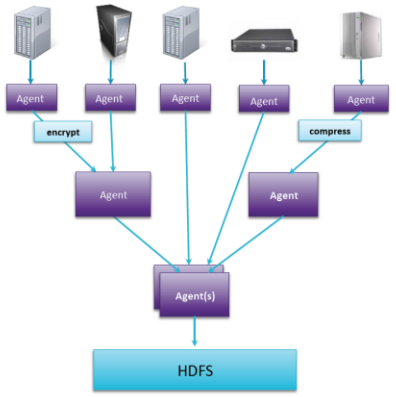
- Use plain HDFS if…
  - You only append to your dataset (no random write)
  - You usually read the whole dataset (no random read)
- Use HBase if…
  - You need random write and/or read
  - You do thousands of operations per second on TB+ of data
- Use an RDBMS if…
  - Your data fits on one big node
  - You need full transaction support
  - You need real-time query capabilities

[Data Integration]
Flume: Real-time Data Import

- What is Flume?
  - A service to move large amounts of data in real time
  - Example: storing log files in HDFS
- Flume imports data into HDFS as it is generated
  - Instead of batch-processing it later
  - For example, log files from a Web server
- Flume is
  - Distributed
  - Reliable and available
  - Horizontally scalable
  - Extensible

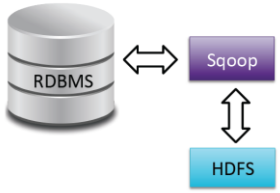
Flume: High-Level Overview

- Collect data as it is produced
  - Files, syslogs, stdout or custom source
- Process in place
  - e.g., encrypt, compress
- Pre-process data before storing
  - e.g., transform, scrub, enrich
- Write in parallel
  - Scalable throughput
- Store in any format
  - Text, compressed, binary, or custom sink



[Data Integration]
Sqoop: Exchanging Data With RDBMSs

- Sqoop transfers data between RDBMSs and HDFS
  - Does this very efficiently via a Map-only MapReduce job
  - Supports JDBC, ODBC, and several specific databases
  - "Sqoop" = "SQL to Hadoop"



[Data Processing]
Apache Spark

- Apache Spark is a fast, general engine for large-scale data processing on a cluster
- Originally developed UC Berkeley's AMPLab
- Open source Apache project
- Provides several benefits over MapReduce
  - Faster
  - Better suited for iterative algorithms
    - Can hold intermediate data in RAM, resulting in much better performance
  - Easier API
    - Supports Python, Scala, Java
  - Supports real-time streaming data processing

Spark vs Hadoop MapReduce

- MapReduce
  - Widely used, huge investment already made
  - Supports and supported by many complementary tools
  - Mature, well-tested
- Spark
  - Flexible
  - Elegant
  - Fast
  - Supports real-time streaming data processing
- MapReduce is still the dominant technology – But losing ground to Spark fast

[Data Analysis] Hive and Pig: High Level Data Languages

- The motivation: MapReduce is powerful but hard to master
- The solution: Hive and Pig
  - Languages for querying and manipulating data
  - Leverage existing skillsets
    - Data analysts who use SQL
    - Programmers who use scripting languages
  - Open source Apache projects
    - Hive initially developed at Facebook
    - Pig Initially developed at Yahoo!
- Interpreter runs on a client machine
  - Turns queries into MapReduce jobs
  - Submits jobs to the cluster

Hive

- What is Hive?
  - HiveQL: An SQL-like interface to Hadoop

```
SELECT * FROM purchases WHERE price > 10000 ORDER BY
storeid
```

Pig

- What is Pig?
  - Pig Latin: A dataflow language for transforming large data sets

```
purchases = LOAD "/user/dave/purchases" AS (itemID,
                  price, storeID, purchaserID);
bigticket = FILTER purchases BY price > 10000;
...
```

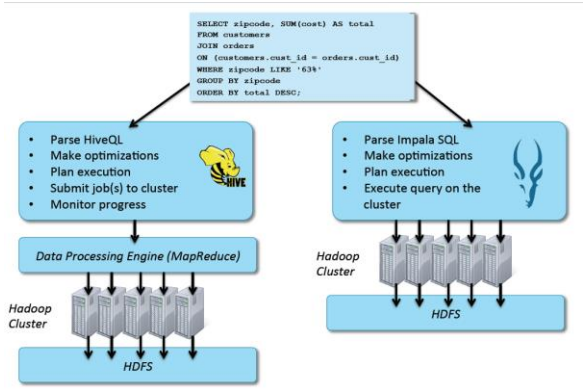
[Data Analysis] Impala: High Performance Queries

- High-performance SQL engine for vast amounts of data
  - Similar query language to HiveQL
  - 10 to 50+ times faster than Hive, Pig, or MapReduce
- Impala runs on Hadoop clusters
  - Data stored in HDFS
  - Does not use MapReduce
- Developed by Cloudera
  - 100% open source, released under the Apache so]ware license

What's the Difference between Hive and Impala?

- Hive has more features
  - E.g. Complex data types (arrays, maps) and full support for windowing analytics
  - Highly extensible
  - Commonly used for batch processing
- Impala is much faster
  - Specialized SQL engine offers 5x to 50x better performance
  - Ideal for interactive queries and data analysis
  - More features being added over time

High-Level Overview



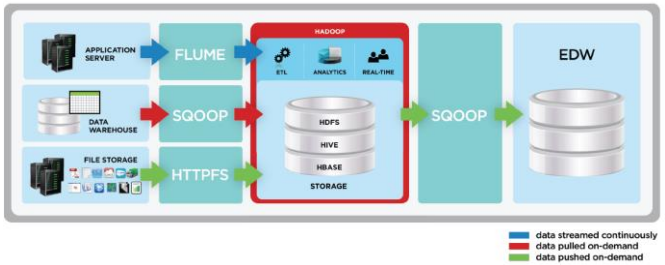
[Machine Learning] Mahout

- Mahout is a Machine Learning library written in Java
- Used for
  - Collaborative filtering (recommendations)
  - Clustering (finding naturally occurring "groupings" in data)
  - Classification (determining whether new data fits a category)
- Why use Hadoop for Machine Learning?
  - "It's not who has the best algorithms that wins. It's who has the most data."

Hue: The UI for Hadoop

- Hue = Hadoop User Experience
- Hue provides a Web front-end to a Hadoop
  - Upload and browse data
  - Query tables in Impala and Hive
  - Run Spark and Pig jobs and workflows
  - Search
  - And much more
- Makes Hadoop easier to use
- Hue is 100% open-source
- Created by Cloudera
  - Open source, released under Apache license

A Typical Data Center With Hadoop



Where to Run a MapReduce Application?

Where to Run Your MapReduce Applications?

- In your local machine or Hadoop cluster
- In a Cloud Computing Platform

Hadoop Installation

- In your local computer
  - Apache hadoop
    - https://hadoop.apache.org/
    - Time-consuming…
- Cloudera QuickStarts VM (Virtual Machine)
  - http://www.cloudera.com/content/www/en-us/downloads.html
  - Hadoop is already set up and the VM contains various hadoop ecosystem components
  - Demo

Midterm (Oct. 27)

- The midterm exam is closed book.
- You may bring one standard 8.5" by 11" piece of paper with any notes you think appropriate or significant (front and back).
- No electronic devices allowed.