BEFORE THE WORKSHOP

How much do you know about Hadoop?

Why do you want to learn Hadoop?
WHAT IS HADOOP AND WHY?

The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing.

What’s in the project? Why?
SCHEDULE

Big data problems

Introduction to Hadoop

Cloudera tutorial
BIG DATA PROBLEMS

Big data is generated by everything around us

High velocity, volume and variety

- every digital process: systems, sensors, mobile devices
- online behaviors: shopping, social media
- scientific researches: DNA, simulation, health
BIG DATA PROBLEMS

Google:
Grew from processing 100TB a day with MapReduce in 2004 to 20 PB a day in 2008
Search Index is 100+ PB (05/2014) (1 PB = 10^{15} bytes)
3.5 billion searches per day

Facebook:
300 PB data in Hive + 600 TB/day (04/2014)
BIG DATA PROBLEMS

Why Big Data?

source: Wikipedia (Everest)
BIG DATA PROBLEMS

Science
• Data-intensive e-Science

Engineering
• Data driven decisions

Commerce
• Data -> Insights -> Competitive advantages
EXAMPLE

Big data & simple algorithm VS. small data & complicated algorithm

Figure 1. Learning Curves for Confusion Set Disambiguation

[Banko and Brill, 2001]
DISCUSSION

Have you ever met with any practical big data problems?
The biggest dataset you have ever worked on?
TACKLE BIG DATA PROBLEMS

"Worker"

Result
TACKLE BIG DATA PROBLEMS

Challenges ??
TACKLE BIG DATA PROBLEMS

How do we assign work units to workers?
How many workers should be involved?
What if workers need to share partial results?
How to aggregate partial results?
How do we know all the workers have finished?
What if workers die?
TACKLE BIG DATA PROBLEMS

Parallelization problems arise from:

• Communication between workers (e.g., to exchange state)
• Access to shared resources (e.g. data)
TACKLE BIG DATA PROBLEMS

Managing multiple workers is difficult because:

• We don’t know the order in which workers run
• We don’t know when workers interrupt each other
• We don’t know when the workers need to communicate
• We don’t know the order in which workers access
• ...
TACKLE BIG DATA PROBLEMS

Move computation to data!
This data center has over 115000 square feet of space
(source: socialpositives.com)
Google Data Centers

**Americas**
- Berkeley County, South Carolina
- Council Bluffs, Iowa
- Douglas County, Georgia
- Jackson County, Alabama
- Lenoir, North Carolina
- Mayes County, Oklahoma
- Montgomery County, Tennessee
- Quillcura, Chile
- The Dalles, Oregon

Mayes County, Oklahoma
SCHEDULE

Big data problems
Introduction to Hadoop
Cloudera and tutorial
WHAT IS HADOOP?

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.
HADOOP

Scalability

• cheap computing storage
• distribute and scale computation easily in a very cost effective manner

Reliability

• Hardware failures handled automatically

Keep all data
HADOOP BASIC MODULES

Hadoop Common
  • libraries and utilities
Hadoop Distributed File System
Hadoop YARN
  • resource management platform
Hadoop MapReduce
  • programming model that scales data across a lot of different processes
import org.apache.commons.cli.CommandLine;
import org.apache.commons.cli.CommandLineParser;
import org.apache.commons.cli.GnuParser;
import org.apache.commons.cli.HelpFormatter;
import org.apache.commons.cli.OptionBuilder;
import org.apache.commons.cli.Options;
import org.apache.commons.cli.ParseException;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.conf.Configurationured;
import org.apache.hadoop.fs.FileSystem;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.util.Tool;
import org.apache.hadoop.util.ToolRunner;
HADOOP BASIC MODULES

source: http://ksoong.org/big-data
HDFS

Hadoop Distributed File System
Distributed, scalable, and portable file system written in Java for the Hadoop framework
HDFS

- Datanode
- Namenode
- HDFS Client
- Replication
- Blocks
- Rack-awareness

**HDFS Architecture**

Metadata (Name, replicas, ...): /home/foo/data, 3, ...

Client ➔ Read ➔ Datanodes ➔ Replication ➔ Blocks ➔ Write ➔ Client

Namenode ➔ Metadata ops ➔ Namenode ➔ Block ops ➔ Datanodes ➔ Blocks ➔ Rack 1 ➔ rack 2
HDFS-NAMENODE

Managing the file system namespace:
  • holds file/directory structure, metadata, access permissions, etc.

Coordinating file operations
  • directs clients to Datanodes for reads and writes
  • no data is moved through the namenode

Maintaining overall health
  • periodic communication with the data
  • block re-replication and rebalancing
  • garbage collection
**HDFS**

Replication: blocks are replicated, the default is 3 times

Rack awareness: can recover from a rack failure or a node failure
HADOOP BASIC MODULES

source: http://ksoong.org/big-data
YARN

Separate resource management and job scheduling/monitoring

Global Resource Manager (RM)

Node Manager on each node

Application Master one for each application
YARN
HADOOP BASIC MODULES

source: http://ksoong.org/big-data
MAPREDUCE FRAMEWORK

Map tasks process data chunks
Framework sorts map output
Reduce tasks use sorted map data as input
**MAP/REDUCE- WORD COUNT**

**Map(String docid, String text):**
for each word w in text:
   Emit(w, 1);

**Reduce(String term, Iterator<Int> values):**
int sum = 0;
for each v in values:
   sum += v;
   Emit(term, value);

---

**Diagram:**

- **Map** operations:
  - k1 v1
  - k2 v2
  - k3 v3
  - k4 v4
  - k5 v5
  - k6 v6

- **Reduce** operations:
  - a 1 b 2
  - c 3 c 6
  - a 5 c 2
  - b 7 c 8

**Shuffle and Sort:** aggregate values by keys

- a 1 5
- b 2 7
- c 2 3 6 8

- **Reduce** operations:
  - \( r_1 s_1 \)
  - \( r_2 s_2 \)
  - \( r_3 s_3 \)
MAP/REDUCE

Shuffle and Sort: aggregate values by keys

reduce

\( r_1 \quad s_1 \)

\( r_2 \quad s_2 \)

\( r_3 \quad s_3 \)
MAP/REDUCE

Partitioner:

- Often a simple hash of the key
- Divides up key space for parallel reduce operations

Combiner:

- In the map side, spills merged in a single, partitioned file, reduce network traffic
1. Given a large amount of text, calculate the co-occurrence of two words in a line. Write the pseudo code of the mapper and reducer.

2. Given a large amount of shopping records \((u, p)\): user spend \(p\) in one shopping. Write the pseudo code of the mapper and reducer to calculate the average spend for each user.
HADOOP WORKFLOW

1. Load data into HDFS

2. Develop code locally

3. Submit MapReduce job
   3a. Go back to Step 2

You

Hadoop Cluster

4. Retrieve data from HDFS
HADOOP BASIC MODULES

source: http://ksoong.org/big-data
HADOOP ECOSYSTEM

source: http://ksoong.org/big-data
HADOOP ECOSYSTEM

HBase:

• Column-oriented database management system
• A scalable data warehouse with support for large table
• Based on Google Big Table, not a relational DBMS (primary key, foreign key, no redundancy etc..)
HADOOP ECOSYSTEM

Hbase: NoSQL => Not only SQL

HOW TO WRITE A CV

DO YOU HAVE ANY EXPERTISE IN SQL?

NO

DOESN'T MATTER. WRITE: "EXPERT IN NO SQL."

Leverage the NoSQL boom
HADOOP ECOSYSTEM

HBase

Map indexed by a row key, column key, and a timestamp

- (row:string, column:string, time:int64) → uninterpreted byte array

Supports lookups, inserts, deletes

- Single row transactions only
HADOOP ECOSYSTEM

source: http://ksoong.org/big-data
HADOOP ECOSYSTEM

Hive:

• a data warehouse infrastructure that provides data summarization and ad hoc querying
• project structure onto the data and query the data using SQL-like language called HiveQL
HADOOP ECOSYSTEM

Hive Syntax

```sql
SELECT [ALL | DISTINCT] select_expr, select_expr, ...
FROM table_reference
[WHERE where_condition]
[GROUP BY col_list]
[HAVING having_condition]
[CLUSTER BY col_list | [DISTRIBUTE BY col_list] [SORT BY col_list]]
[LIMIT number]
;
```
HADOOP ECOSYSTEM

source: http://ksoong.org/big-data
HADOOP ECOSYSTEM

Pig: a high-level data-flow language and execution framework for parallel computation, on top of Hadoop MapReduce

```
A = LOAD '/shared/tweets2011.txt' USING PigStorage('	') AS (uid:chararray, name:chararray, date:chararray, text:chararray);
B = FOREACH A GENERATE ToDate(date, 'EEE MMM dd HH:mm:ss Z YYYY','Etc/GMT') AS date;
C = FOREACH B GENERATE ToString(date, 'MM/dd HH') AS date;
D = group C by date;
E = foreach D generate group as term, COUNT(C) as count;
store E into 'hourly-counts-pig-all';
```
HADOOP ECOSYSTEM

source: http://ksoong.org/big-data
HADOOP ECOSYSTEM

Oozie:

Workflow scheduler system to manage Apache Hadoop Jobs

Supports MapReduce, Pig, Apache Hive, Sqoop, etc.
HADOOP ECOSYSTEM

source: http://ksoong.org/big-data
HADOOP ECOSYSTEM

Zookeeper:

• Maintaining configuration information
• Naming services
• Providing distributed synchronization
• Providing group services
HADOOP ECOSYSTEM

source: http://ksoong.org/big-data
HADOOP ECOSYSTEM

Sqoop: Tool designed for efficiently transferring data between Hadoop and structured data stores such as relational databases

details in Cloudera hands on...
HADOOP ECOSYSTEM

source: http://ksoong.org/big-data
HADOOP ECOSYSTEM

Flume:

• Distributed, reliable and available service for efficiently collecting, aggregating, and moving large amounts of log data
HADOOP ECOSYSTEM

Spark:
a fast and general compute engine for Hadoop data
Wide range of applications – machine learning, graphic analytics, stream processing

more details in later parts…
REFERENCES


3. Coursera: Hadoop Platform and Application Framework by University of California, San Diego

SCHEDULE

Big data problems
Introduction to Hadoop
Cloudera and tutorial
Install Virtual box and Cloudera

Any Question?