CSE 444: Database Internals

Lectures 25
MapReduce

References

- MapReduce: Simplified Data Processing on Large Clusters. Jeffrey Dean and Sanjay Ghemawat. OSDI'04.

Outline

- Review high-level MR ideas from 344
- Discuss implementation in greater detail

Map Reduce Review

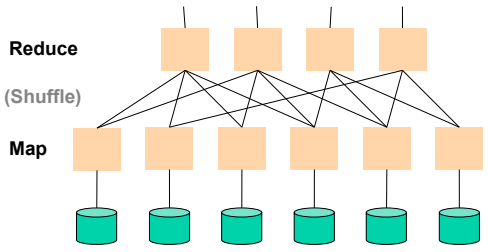
- Google: [Dean 2004]
- Open source implementation: Hadoop
- MapReduce = high-level programming model and implementation for large-scale parallel data processing

MapReduce Motivation

- Not designed to be a DBMS
- Designed to simplify task of writing parallel programs
  - A simple programming model that applies to many large-scale computing problems
- Hides messy details in MapReduce runtime library:
  - Automatic parallelization
  - Load balancing
  - Network and disk transfer optimizations
  - Handling of machine failures
  - Robustness
  - **Improvements to core library benefit all users of library!**

content in part from: Jeff Dean

Observation: Your favorite parallel algorithm…



Typical Problems Solved by MR

- Read a lot of data
- Map: extract something you care about from each record
- Shuffle and Sort
- Reduce: aggregate, summarize, filter, transform
- Write the results

Outline stays the same,
map and reduce change to fit the problem

slide source: Jeff Dean

MapReduce Programming Model

- Input & Output: each a set of key/value pairs
- Programmer specifies two functions

map (in_key, in_value) -> list(out_key, intermediate_value)

- Processes input key/value pair
- Produces set of intermediate pairs

reduce (out_key, list(intermediate_value)) -> list(out_value)

- Combines all intermediate values for a particular key
- Produces a set of merged output values (usually just one)

slide source: Google, Inc.

Example: What does this do?

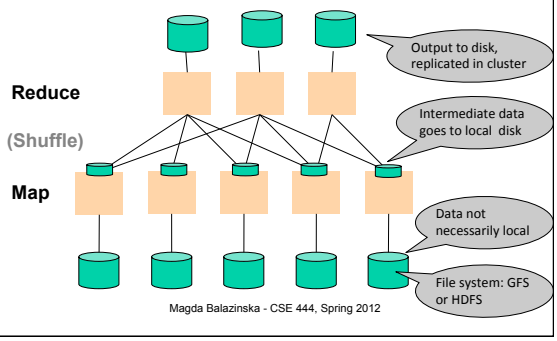
```
map(String input_key, String input_value):
  // input_key: document name
  // input_value: document contents
  for each word w in input_value:
     EmitIntermediate(w, 1);

reduce(String output_key, Iterator intermediate_values):
  // output_key: word
  // output_values: ????
  int result = 0;
  for each v in intermediate_values:
     result += v;
  Emit(result);
```


slide source: Google, Inc.

Parallel DBMS vs MapReduce

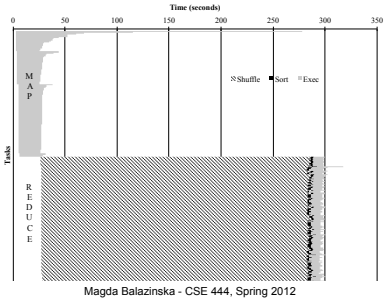
- Parallel DBMS
  - Relational data model and schema
  - Declarative query language: SQL
  - Many pre-defined operators: relational algebra
  - Can easily combine operators into complex queries
  - Query optimization
  - Can do more than just run queries: Data management
    - Updates and transactions, constraints, security, etc.
- MapReduce
  - Data model is a file with key-value pairs!
  - No need to "load data" before processing it
  - Easy to write user-defined operators

Parallel MapReduce



Example MapReduce Execution

PageRank Application



Data Storage: GFS/HDFS

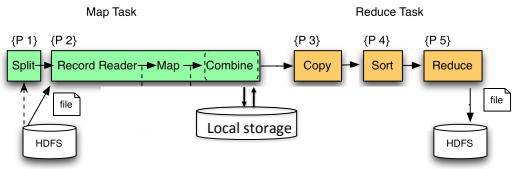
- MapReduce job input is a file
- Common implementation is to store files in a highly scalable file system such as GFS/HDFS
  - GFS: Google File System
  - HDFS: Hadoop File System
  - Each data file is split into M blocks (64MB or more)
  - Blocks are stored on random machines & replicated
  - Files are append only

MapReduce Implementation

Illustration on the white-board (please take notes):

- There is one master node
- Input file gets partitioned further into *M' splits*
  - Each split is a contiguous piece of the input file
- Master assigns *workers* (=servers) to the *M' map tasks*, keeps track of their progress
- Workers write their output to local disk
- Output of each map task is partitioned into *R regions*
- Master assigns workers to the *R reduce tasks*
- Reduce workers read regions from the map workers' local disks

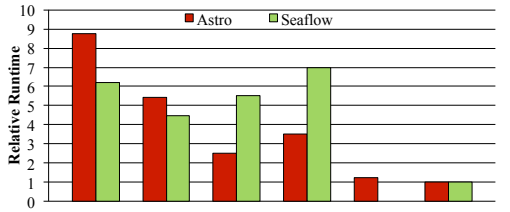
MapReduce Phases



Interesting Implementation Details

- Worker failure:
  - Master pings workers periodically,
  - If down then reassigns its split *to another* worker
  - (≠ a parallel DBMS restarts whole query)
- How many map and reduce tasks:
  - Larger is better for load balancing
  - But more tasks also add overheads
  - (≠ parallel DBMS spreads ops across all nodes)

MapReduce Granularity Illustration



Coarse	Fine	Finer	Finest	Manual	SkewReduce	
14.1	8.8	4.1	5.7	2.0	1.6	Hours
87.2	63.1	77.7	98.7	-	14.1	Minutes

Interesting Implementation Details

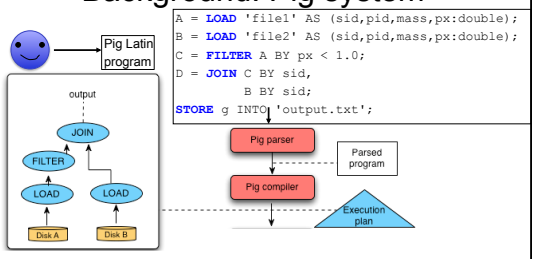
Backup tasks:

- *Straggler* = a machine that takes unusually long time to complete one of the last tasks. Eg:
  - Bad disk forces frequent correctable errors (30MB/s → 1MB/s)
  - The cluster scheduler has scheduled other tasks on that machine
- Stragglers are a main reason for slowdown
- Solution: *pre-emptive backup execution of the last few remaining in-progress tasks*

Parallel DBMS vs MapReduce

- Parallel DBMS
  - Indexing
  - Physical tuning
  - Can stream data from one operator to the next without blocking
- MapReduce
  - Can easily add nodes to the cluster (no need to even restart)
  - Uses less memory since processes one key-group at a time
  - Intra-query fault-tolerance thanks to results on disk
  - Intermediate results on disk also facilitate scheduling
  - Handles adverse conditions: e.g., stragglers
  - Arguably more scalable… but also neesd more nodes!

Background: Pig system



MapReduce State

- Lots of extensions to address limitations
  - Capabilities to write DAGs of MapReduce jobs
  - Declarative languages (see 344)
  - Ability to read from structured storage (e.g., indexes)
  - Etc.
- Most companies use both types of engines
- Increased integration of both engines