Developing a MapReduce Application

Oguzhan Gencoglu

Outline

Outline

What is MapReduce

MapReduce is a software framework for processing (large) data sets in a distributed fashion over several machines.

Core idea

$<$ **key, value** $>$ pairs

What is MapReduce

MapReduce is a software framework for processing (large) data sets in a distributed fashion over several machines.

Core idea

$<$ **key, value** $>$ pairs

- Almost all data can be mapped into key, value pairs.

What is MapReduce

MapReduce is a software framework for processing (large) data sets in a distributed fashion over several machines.

Core idea

$<$ **key, value** $>$ pairs

- Almost all data can be mapped into key, value pairs.
- Keys and values may be of any type.

Outline

MapReduce Workflow

- Write your map and reduce functions

MapReduce Workflow

- Write your map and reduce functions
- Test with a small subset of data

MapReduce Workflow

- Write your map and reduce functions
- Test with a small subset of data
- If it fails use your IDE's debugger to find the problem

MapReduce Workflow

- Write your map and reduce functions
- Test with a small subset of data
- If it fails use your IDE's debugger to find the problem
- Run on full dataset

MapReduce Workflow

- Write your map and reduce functions
- Test with a small subset of data
- If it fails use your IDE's debugger to find the problem
- Run on full dataset
- If it fails Hadoop provides some debugging tools

MapReduce Workflow

- Write your map and reduce functions
- Test with a small subset of data
- If it fails use your IDE's debugger to find the problem
- Run on full dataset
- If it fails Hadoop provides some debugging tools
  - e.g. IsolationRunner : runs a task over the same input which it failed.

MapReduce Workflow

- Write your map and reduce functions
- Test with a small subset of data
- If it fails use your IDE's debugger to find the problem
- Run on full dataset
- If it fails Hadoop provides some debugging tools
  - e.g. IsolationRunner : runs a task over the same input which it failed.
- Do profiling to tune the performance

Outline

Hadoop Default Ports

- Handful of ports over TCP.
- Some used by Hadoop itself (to schedule jobs, replicate blocks, etc.).
- Some are directly for users (either via an interposed Java client or via plain old HTTP)

	Daemon	Default Port	Configuration Parameter
HDFS	Namenode	50070	dfs.http.address
	Datanodes	50075	dfs.datanode.http.address
	Secondarynamenode	50090	dfs.secondary.http.address
	Backup/Checkpoint node[?]	50105	dfs.backup.http.address
MR	Jobracker	50030	mapred.job.tracker.http.address
	Tasktrackers	50060	mapred.task.tracker.http.address

[?] Replaces secondarynamenode in 0.21.

Outline

Word Count

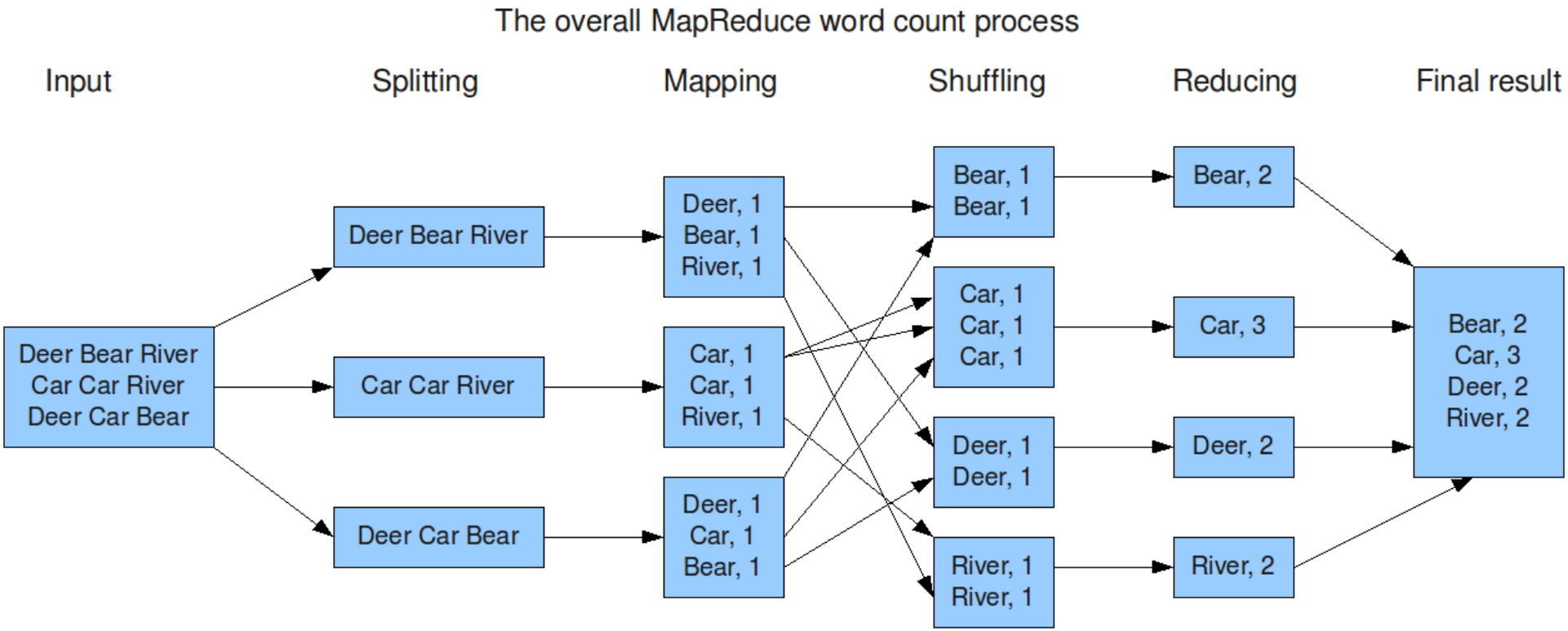
Task: Counting the word occurances (frequencies) in a text file (or set of files).

$$< \textbf{word, count} > \textbf{ as } < \textbf{key, value} > \textbf{ pair}$$

Mapper: Emits $<$ word, 1 $>$ for each word (no counting at this part).

Shuffle in between: pairs with same keys grouped together and passed to a single machine.

Reducer: Sums up the values (1s) with the same key value.


The overall MapReduce word count process
Input
Splitting
Mapping
Shuffling
Reducing
Final result
Deer Bear River
Car Car River
Deer Car Bear
Deer Bear River
Car Car River
Deer Car Bear
Deer, 1
Bear, 1
River, 1
Car, 1
Car, 1
River, 1
Deer, 1
Car, 1
Bear, 1
Bear, 1
Bear, 1
Car, 1
Car, 1
Car, 1
Deer, 1
Deer, 1
River, 1
River, 1
Bear, 2
Car, 3
Deer, 2
River, 2
Bear, 2
Car, 3
Deer, 2
River, 2

Outline

Job Tracker

130 Hadoop Map/Reduce Administration

State: RUNNING
Started: Mon Nov 17 22:41:46 PST 2014
Version: 0.18.0, r686010
Compiled: Thu Aug 14 19:48:33 UTC 2008 by hadoopqa
Identifier: 201411172241

Cluster Summary

Maps	Reduces	Total Submissions	Nodes	Map Task Capacity	Reduce Task Capacity	Avg. Tasks/Node
0	0	3	1	2	2	4.00

Running Jobs

Running Jobs
none

Completed Jobs

Completed Jobs								
Jobid	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Reduces Completed
job_201411172241_0003	hadoop-user	streamjob16751.jar	100.00%	2	2	100.00%	1	1
job_201411172241_0004	hadoop-user	streamjob28967.jar	100.00%	2	2	100.00%	1	1

Failed Jobs

Failed Jobs								
Jobid	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Reduces Completed
job_201411172241_0001	hadoop-user	streamjob64235.jar	100.00%	2	2	100.00%	1	0

Local logs

Log directory, Job Tracker History

Hadoop, 2014.

Tasks

Hadoop map task list for job_200904110811_0003 on ip-10-250-110-47

Completed Tasks

Task	Complete	Status	Start Time	Finish Time	Errors	Counters
task_200904110811_0003_m_000043	100.00%	hdfs://ip-10-250-110-47.ec2.internal/user/root/input/ncdc/all/1949.gz:0+220338475	11-Apr-2009 09:00:06	11-Apr-2009 09:01:25 (1mins, 18sec)		10
task_200904110811_0003_m_000044	100.00%	Detected possibly corrupt record: see logs.	11-Apr-2009 09:00:06	11-Apr-2009 09:01:28 (1mins, 21sec)		11
task_200904110811_0003_m_000045	100.00%	hdfs://ip-10-250-110-47.ec2.internal/user/root/input/ncdc/all/1970.gz:0+208374610	11-Apr-2009 09:00:06	11-Apr-2009 09:01:28 (1mins, 21sec)		10

Name Node

NameNode '130.230.16.37:9000'

Started: Tue Nov 18 18:09:31 PST 2014
Version: 0.18.0, r686010
Compiled: Thu Aug 14 19:48:33 UTC 2008 by hadoopqa
Upgrades: There are no upgrades in progress.

Browse the filesystem

Cluster Summary

25 files and directories, 28 blocks = 53 total. Heap Size is 5.98 MB / 992.31 MB (0%)

Capacity	:	23.73 GB
DFS Remaining	:	21.42 GB
DFS Used	:	529.41 KB
DFS Used%	:	0 %
Live Nodes	:	1
Dead Nodes	:	0

Live Datanodes : 1

Node	Last Contact	Admin State	Size (GB)	Used (%)	Used (%)	Remaining (GB)	Blocks
hadoop-desk	2	In Service	23.73	0		21.42	28

Dead Datanodes : 0

Outline

Key Points

- Test mapper and reducer outside hadoop.

Key Points

- Test mapper and reducer outside hadoop.
- Copy your MapReduce function and files to DFS.

Key Points

- Test mapper and reducer outside hadoop.
- Copy your MapReduce function and files to DFS.
- Test mapper and reducer with hadoop using a small portion of the data.

Key Points

- Test mapper and reducer outside hadoop.
- Copy your MapReduce function and files to DFS.
- Test mapper and reducer with hadoop using a small portion of the data.
- Track the jobs, debug, do profiling

Questions/Comments