



Chameleon Cloud Tutorial

Setting up Hadoop Sandbox on Chameleon Bare Metal Servers

Hadoop - Implementation of Map Reduce (Python)

Objective

In this tutorial, we will discuss about the Map and Reduce program, its implementation and advantages over the other processes.

Prerequisites

The following prerequisites are expected for successful completion of this tutorial:

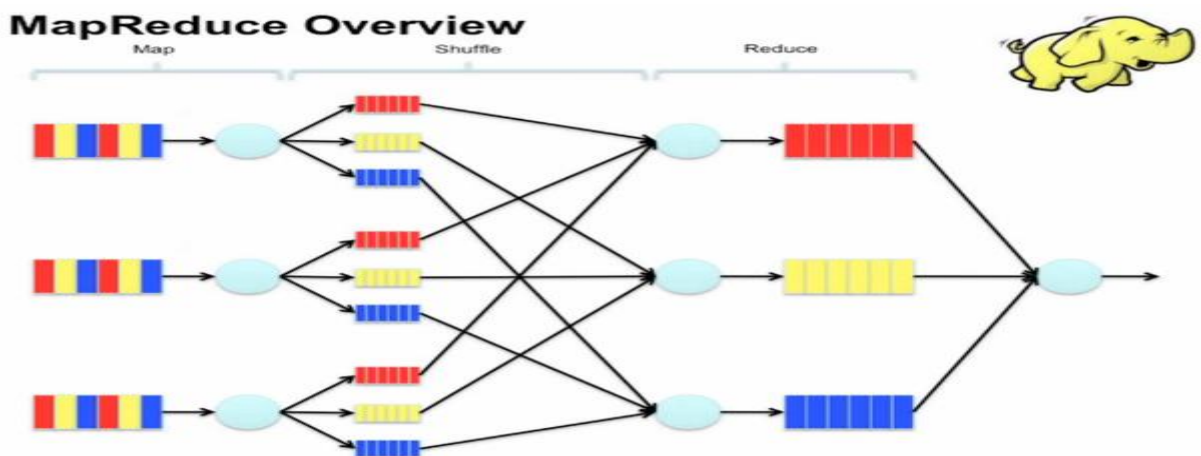
- Chameleon Cloud account (<http://chameleoncloud.org/user/register/>)
- SSH client (Windows users: download PuTTY from here: <http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html>)
- A basic knowledge of Linux.
- Installation of Hadoop and Map reduce.
- Single node set up.

MapReduce

- MapReduce is a software framework used to process vast amount of data.
- MapReduce usually splits the input data set into chunks, map task process the data in parallel manner.
- The output of the Map is fed as Input to the Reducer.
- Both input and output files are stored in system.
- MapReduce programs works exclusively on key value pair.

(input) $\langle k1, v1 \rangle \rightarrow \text{map} \rightarrow \langle k2, v2 \rangle \rightarrow \text{combine} \rightarrow \langle k2, v2 \rangle \rightarrow \text{reduce} \rightarrow \langle k3, v3 \rangle$ (output)

MapReduce flow diagram



Implementation of MapReduce

- Login to the system with user id 'CC' and enter to establish the connection.

Login as 'CC'

129.114.34.118 - PuTTY

```
login as: cc
```

Enter to establish connection(Authenticate with the public key)

cc@karthi-khj059-n01:~

```
login as: cc
Authenticating with public key "rsa-key-20150615"
Last login: Mon Jun 29 11:19:55 2015 from cpe-72-191-37-37.satx.res.rr.com
[cc@karthi-khj059-n01 ~]$
```

- Change the directory to Hadoop

cc@karthi-khj059-n01:~

```
login as: cc
Authenticating with public key "rsa-key-20150615"
Last login: Mon Jun 29 11:47:04 2015 from cpe-72-191-37-37.satx.res.rr.com
[cc@karthi-khj059-n01 ~]$ cd hadoop
```

cc@karthi-khj059-n01:~/hadoop

```
login as: cc
Authenticating with public key "rsa-key-20150615"
Last login: Mon Jun 29 11:47:04 2015 from cpe-72-191-37-37.satx.res.rr.com
[cc@karthi-khj059-n01 ~]$ cd hadoop
[cc@karthi-khj059-n01 hadoop]$
```

- Create the file 'K'(Mapper code) with py extension.

```
cc@karthi-khj059-n01:~/hadoop
login as: cc
Authenticating with public key "rsa-key-20150615"
Last login: Mon Jun 29 11:53:56 2015 from cpe-72-191-37-37.satx.res.rr.com
[cc@karthi-khj059-n01 ~]$ cd hadoop
[cc@karthi-khj059-n01 hadoop]$ vi k.py
```

Input the mapper code in K.py file

```
cc@karthi-khj059-n01:~/hadoop
#!/usr/bin/env python

import sys

for line in sys.stdin:
    words = line.split()
    for word in words:
        print "%s\t%d" % (word,1)
```

- Create the file 'A'(Reducer code) with py extension.

```
cc@karthi-khj059-n01:~/hadoop
login as: cc
Authenticating with public key "rsa-key-20150615"
Last login: Mon Jun 29 11:53:56 2015 from cpe-72-191-37-37.satx.res.rr.com
[cc@karthi-khj059-n01 ~]$ cd hadoop
[cc@karthi-khj059-n01 hadoop]$ vi k.py
[cc@karthi-khj059-n01 hadoop]$ vi a.py
```

- Input the Reducer code in A.py file

```
cc@karthi-khj059-n01:~/hadoop
#!/usr/bin/python

import sys

def output(previous_key,total):
    if previous_key is not None:
        print "%s was found %d times" % (previous_key, total)

previous_key = None
total = 0

for line in sys.stdin:
    key, value = line.split("\t", 1)
    if key != previous_key:
        output(previous_key, total)
        previous_key = key
        total = 0
    total += int(value)

output(previous_key, total)
```

- Make sure the file has the execution permission

Chmod +x /path of the file

```
cc@karthi-khj059-n01:~/hadoop
login as: cc
Authenticating with public key "rsa-key-20150615"
Last login: Mon Jun 29 11:57:07 2015 from cpe-72-191-37-37.satx.res.rr.com
[cc@karthi-khj059-n01 ~]$ cd hadoop
[cc@karthi-khj059-n01 hadoop]$ vi k.py
[cc@karthi-khj059-n01 hadoop]$ vi a.py
[cc@karthi-khj059-n01 hadoop]$ chmod +x /home/cc/hadoop/k.py
```

```
cc@karthi-khj059-n01:~/hadoop
login as: cc
Authenticating with public key "rsa-key-20150615"
Last login: Mon Jun 29 11:57:07 2015 from cpe-72-191-37-37.satx.res.rr.com
[cc@karthi-khj059-n01 ~]$ cd hadoop
[cc@karthi-khj059-n01 hadoop]$ vi k.py
[cc@karthi-khj059-n01 hadoop]$ vi a.py
[cc@karthi-khj059-n01 hadoop]$ chmod +x /home/cc/hadoop/k.py
[cc@karthi-khj059-n01 hadoop]$ chmod +x /home/cc/hadoop/a.py
```

- Copy the text data to an input file(t.txt)

```
cc@karthi-khj059-n01:~/hadoop
login as: cc
Authenticating with public key "rsa-key-20150615"
Last login: Mon Jun 29 12:13:11 2015 from cpe-72-191-37-37.satx.res.rr.com
[cc@karthi-khj059-n01 ~]$ cd hadoop
[cc@karthi-khj059-n01 hadoop]$ vi k.py
[cc@karthi-khj059-n01 hadoop]$ vi a.py
[cc@karthi-khj059-n01 hadoop]$ vi t.txt
```

```
cc@karthi-khj059-n01:~/hadoop
IS6353-SECURITY INCIDENT RESPONSE

Security Intrusion and Detection Process based on Cuckoo's Egg.

Karthi Amaravati @ 01521658

6/23/2015

□
Table of contents
S No          Page No          Contents
1              3          Introduction
2              3              Goal
3              3 Components of Incidence Response
```

- Run the Map program for the input text file(where the whole data is divided into chunks) and key value pair is formed.

```
cc@karthi-khj059-n01:~/hadoop
login as: cc
Authenticating with public key "rsa-key-20150615"
Last login: Mon Jun 29 12:13:11 2015 from cpe-72-191-37-37.satx.res.rr.com
[cc@karthi-khj059-n01 ~]$ cd hadoop
[cc@karthi-khj059-n01 hadoop]$ vi k.py
[cc@karthi-khj059-n01 hadoop]$ vi a.py
[cc@karthi-khj059-n01 hadoop]$ vi t.txt
[cc@karthi-khj059-n01 hadoop]$ vi t.txt
[cc@karthi-khj059-n01 hadoop]$ cat t.txt | /home/cc/hadoop/k.py
```

- Output of an Map program

```
cc@karthi-khj059-n01:~/hadoop
Edition.      1
22.           1
Page          1
no.           1
363,          1
The           1
Cuckoo's      1
Egg:          1
tracking      1
a            1
spy          1
through      1
the          1
maze         1
of           1
computer      1
espionage     1
by           1
Cliff        1
Stoll,       1
September    1
2005         1
Edition.      1
[cc@karthi-khj059-n01 hadoop]$
```

- Execute Map and Reduce program together.

```

cc@karthi-khj059-n01:~/hadoop
login as: cc
Authenticating with public key "rsa-key-20150615"
Last login: Mon Jun 29 12:21:48 2015 from cpe-72-191-37-37.satx.res.rr.com
[cc@karthi-khj059-n01 ~]$ cd hadoop
[cc@karthi-khj059-n01 hadoop]$ vi k.py
[cc@karthi-khj059-n01 hadoop]$ vi a.py
[cc@karthi-khj059-n01 hadoop]$ vi t.txt
[cc@karthi-khj059-n01 hadoop]$ cat t.txt | /home/cc/hadoop/k.py | sort | /home/cc/h
adoop/a.py

```

- Output of an Map and Reduce program

```

cc@karthi-khj059-n01:~/hadoop
[cc@karthi-khj059-n01 ~]$ cd hadoop
[cc@karthi-khj059-n01 hadoop]$ vi k.py
[cc@karthi-khj059-n01 hadoop]$ vi a.py
[cc@karthi-khj059-n01 hadoop]$ vi t.txt
[cc@karthi-khj059-n01 hadoop]$ cat t.txt | /home/cc/hadoop/k.py | sort | /home/cc/h
adoop/a.py
01521658 was found 1 times
, was found 3 times
. was found 1 times
@ was found 1 times
□ was found 2 times
100th was found 1 times
10. was found 1 times
10 was found 4 times
101, was found 1 times
108, was found 1 times
10-Firing was found 1 times
1, was found 1 times
1. was found 3 times
1 was found 3 times
11- was found 1 times
11. was found 1 times
11 was found 5 times
111, was found 1 times

```

Advantages

1. Distribute data and computation, which reduces network overload.
2. Tasks are independent due to this it can easily handle partial failures.
3. Linear scaling in the ideal case.
4. Simple programming model.
5. Flat scalability.