Hadoop Coding

Tutorial

Before running the code

- make sure that HDFS and MapReduce is running

- you do that from cloudera manager web service (see next slide)

quickstart.cloudera:7180/cmf/home

Search

Cloudera | Hue | Hadoop | HBase | Impala | Spark | Solr | Oozie | Cloudera Manager | Getting Started

Home

Status | All Health Issues 2 | Configuration 12 | All Recent Commands | Try Cloudera Enterprise for 60 Days | Add Cluster

Cloudera Quick... (CDH 5.13.0, Packages)

- Hosts 1
- HBase
- HDFS 2 2
- Hive
- Hue
- Impala
- Key-Value S...
- Oozie
- Solr
- Spark
- Sqoop 1 Cli...
- Sqoop 2
- YARN (MR2 ... 4

HDFS Actions

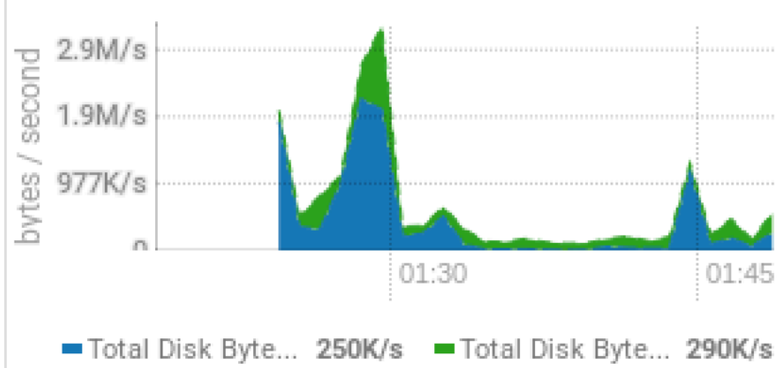
- Start
- Stop
- Restart
- Rolling Restart
- Instances
- Configuration
- Add Role Instances

Charts

30m 1h 2h 6h 12h 1d 7d 30d

Cluster CPU

Cluster Disk IO



Cluster Network IO

create input and output locations in HDFS

- Use the following commands to create the input directory `/user/cloudera/wordcount/input` in HDFS:

```
$ sudo su hdfs

$ hadoop fs -mkdir /user/cloudera

$ hadoop fs -chown cloudera /user/cloudera

$ exit

$ sudo su cloudera
```

only once
the first time you access hfs
using cloudera user

```
$ hadoop fs -mkdir /user/cloudera/wordcount /user/cloudera/wordcount/input
```

Create sample text files to use as input

- create sample input files and move them to the `/user/cloudera/wordcount/input` directory in HDFS.
- You can use any files you choose; for convenience, the following shell commands create a few small input files for illustrative purposes

```
$ echo "Hadoop is an elephant" > file0

$ echo "Hadoop is as yellow as can be" > file1

$ echo "Oh what a yellow fellow is Hadoop" > file2

$ hadoop fs -put file* /user/cloudera/wordcount/input
```

Compile the WordCount class

- you need to create the java class WordCount.java

- then build the project : WordCount.java plus all dependencies

WordCount Code

- you can get the code from course page on LMS or from the lectures slide

- https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#Source_Code

Compile the Code

- build and dependencies

```
$ mkdir -p build

$ javac -cp /usr/lib/hadoop/*:/usr/lib/hadoop-mapreduce/* WordCount.java -d build -Xlint
```

Create and run application

- *Create a JAR file for the WordCount application.*

$ jar -cvf wordcount.jar -C build/ .

- *Run the WordCount application from the JAR file, passing the paths to the input and output directories in HDFS.*

$ hadoop jar wordcount.jar org.myorg.WordCount /user/cloudera/wordcount/input /user/cloudera/wordcount/output

- *note class name including packages if it exists inside packages*

-

Job tracking

- on master node you can track the progress of your job
- also on the terminal from where you run the application

quickstart.cloudera:19888/jobhistory/job/job_1569671520570_0001
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

job_1569671520570_0001

- Application
- Job
  - Overview
  - Counters
  - Configuration
  - Map tasks
  - Reduce tasks
- Tools

Job Overview	
Job Name:	word count
User Name:	cloudera
Queue:	root.users.cloudera
State:	SUCCEEDED
Uberized:	false
Submitted:	Sat Sep 28 05:49:10 PDT 2019
Started:	Sat Sep 28 05:50:03 PDT 2019
Finished:	Sat Sep 28 05:51:21 PDT 2019
Elapsed:	1mins, 18sec
Diagnostics:	
Average Map Time	36sec
Average Shuffle Time	4sec
Average Merge Time	0sec
Average Reduce Time	1sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Sat Sep 28 05:49:37 PDT 2019	quickstart.cloudera:8042	logs

Task Type	Total	Complete
Map	3	3
Reduce	1	1

Attempt Type	Failed	Killed	Successful
Maps	0	0	3

cloudera@quickstart:~/wordcount

File Edit View Search Terminal Help

```
[cloudera@quickstart wordcount]$ hadoop jar wordcount.jar WordCount /user/cloudera/wordcount/input /user/cloudera/wordcount/o
utput
19/09/28 05:49:06 INFO client.RMProxy: Connecting to ResourceManager at quickstart.cloudera/127.0.0.1:8032
19/09/28 05:49:08 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool in
terface and execute your application with ToolRunner to remedy this.
19/09/28 05:49:09 INFO input.FileInputFormat: Total input paths to process : 3
19/09/28 05:49:09 INFO mapreduce.JobSubmitter: number of splits:3
19/09/28 05:49:10 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1569671520570_0001
19/09/28 05:49:11 INFO impl.YarnClientImpl: Submitted application application_1569671520570_0001
19/09/28 05:49:12 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_15696715205
70_0001/
19/09/28 05:49:12 INFO mapreduce.Job: Running job: job_1569671520570_0001
19/09/28 05:50:05 INFO mapreduce.Job: Job job_1569671520570_0001 running in uber mode : false
19/09/28 05:50:05 INFO mapreduce.Job:  map 0% reduce 0%
19/09/28 05:51:05 INFO mapreduce.Job:  map 67% reduce 0%
19/09/28 05:51:14 INFO mapreduce.Job:  map 100% reduce 0%
19/09/28 05:51:23 INFO mapreduce.Job:  map 100% reduce 100%
19/09/28 05:51:24 INFO mapreduce.Job: Job job_1569671520570_0001 completed successfully
19/09/28 05:51:24 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=147
                FILE: Number of bytes written=589263
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=481
                HDFS: Number of bytes written=88
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=3
                Launched reduce tasks=1
                Data-local map tasks=3
                Total time spent by all maps in occupied slots (ms)=56239104
                Total time spent by all reduces in occupied slots (ms)=3208704
                Total time spent by all map tasks (ms)=109842
                Total time spent by all reduce tasks (ms)=6267
                Total vcore-milliseconds taken by all map tasks=109842
```

View Job Output

```
$ hadoop fs -cat /user/cloudera/wordcount/output/*

Hadoop    3

Oh        1

a         1

an        1

as        2

be        1

can       1

elephant  1

fellow    1

is        3

what      1

yellow    2
```

References

- https://docs.cloudera.com/documentation/other/tutorial/CDH5/topics/ht_usage.html

- https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#Source_Code