

Master

サンプル処理の実行 多くのサンプルジョブ
(例) pi (モンテカルロ法で円周率を算出)
wordcount
(ファイル内で単語の出現回数のカウント)
grep (文字列による行の抽出)

以降、ジョブはNamenodeから実行

(1) piを利用して円周率を算出

コマンド

```
# cd /usr/local/hadoop
# hadoop jar /usr/local/hadoop/hadoop-examples-1.0.4.jar \
> pi 引数1 引数2
```

利用法: pi Map数 Map毎のサンプル数

最終行に円周率の結果が出力されます。

注意: Map数 Map毎のサンプル数を変更しながら複数回試します。

```
pi 4 2000
pi 8 2000
pi 10 10000 ...等
```

MapReduceのジョブ進捗状況をWebから確認します。
<http://Jobtrackerのホスト名:50030/>

```
# hadoop jar /usr/local/hadoop/hadoop-examples-1.0.4.jar pi 8 2000
Number of Maps = 8
Samples per Map = 2000
Wrote input for Map #0
Wrote input for Map #1
Wrote input for Map #2
Wrote input for Map #3
Wrote input for Map #4
Wrote input for Map #5
Wrote input for Map #6
Wrote input for Map #7
Starting Job
10/10/20 15:11:06 INFO mapred.FileInputFormat: Total input paths to process : 8
10/10/20 15:11:07 INFO mapred.JobClient: Running job: job_201010201506_0001
10/10/20 15:11:08 INFO mapred.JobClient: map 0% reduce 0%
10/10/20 15:11:16 INFO mapred.JobClient: map 25% reduce 0%
10/10/20 15:11:22 INFO mapred.JobClient: map 75% reduce 0%
10/10/20 15:11:28 INFO mapred.JobClient: map 100% reduce 25%
10/10/20 15:11:37 INFO mapred.JobClient: map 100% reduce 100%
10/10/20 15:11:39 INFO mapred.JobClient: Job complete: job_201010201506_0001
10/10/20 15:11:39 INFO mapred.JobClient: Counters: 18
10/10/20 15:11:39 INFO mapred.JobClient:   Job Counters
10/10/20 15:11:39 INFO mapred.JobClient:     Launched reduce tasks=1
10/10/20 15:11:39 INFO mapred.JobClient:     Launched map tasks=8
10/10/20 15:11:39 INFO mapred.JobClient:     Data-local map tasks=8
10/10/20 15:11:39 INFO mapred.JobClient:   FileSystemCounters
10/10/20 15:11:39 INFO mapred.JobClient:     FILE_BYTES_READ=182
10/10/20 15:11:39 INFO mapred.JobClient:     HDFS_BYTES_READ=944
10/10/20 15:11:39 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=662
10/10/20 15:11:39 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=215
10/10/20 15:11:39 INFO mapred.JobClient:   Map-Reduce Framework
10/10/20 15:11:39 INFO mapred.JobClient:     Reduce input groups=16
10/10/20 15:11:39 INFO mapred.JobClient:     Combine output records=0
10/10/20 15:11:39 INFO mapred.JobClient:     Map input records=8
10/10/20 15:11:39 INFO mapred.JobClient:     Reduce shuffle bytes=224
10/10/20 15:11:39 INFO mapred.JobClient:     Reduce output records=0
10/10/20 15:11:39 INFO mapred.JobClient:     Spilled Records=32
10/10/20 15:11:39 INFO mapred.JobClient:     Map output bytes=144
10/10/20 15:11:39 INFO mapred.JobClient:     Map input bytes=192
10/10/20 15:11:39 INFO mapred.JobClient:     Combine input records=0
10/10/20 15:11:39 INFO mapred.JobClient:     Map output records=16
10/10/20 15:11:39 INFO mapred.JobClient:     Reduce input records=16
Job Finished in 33.839 seconds
Estimated value of Pi is 3.14250000000000000000
```

© Hitachi Information Academy Co., Ltd. 2014. All rights reserved.

3-6 . Map/Reduceジョブの投入 (2/7)

(2) wordcount (単語の出現回数のカウント) (1/2)

Master

マスター・ノード(pc88xxa)にて、コマンド

```
hadoop jar /usr/local/hadoop/hadoop-examples-1.0.4.jar wordcount 引数1 引数2
```

演習

wordcount 入力ファイル名 出力ディレクトリ

```
# ls /mnt/cloud/pg* (ファイルはありますか?)
```

無い場合は、Dom-0側からデータを取得します。

```
# scp pc8801:/cloud/pg* /mnt/cloud
```

入力サンプル・ファイル (書籍をテキスト化したものです)
pg1661.txt(577KB) シャーロックホームズの冒険(コナン Doyle)
pg5000.txt(1,391KB) レオナルド・ダヴィンチのノート
pg20417.txt(659KB) 科学概要(J.アーサー・トムソン)

```
# more /mnt/cloud/pg1661.txt (内容の確認)
```

コマンド例 (先にファイルをHDFSへ配備しています。)

```
# hadoop fs -put /mnt/cloud/pg1661.txt /user/root/pg1661.txt
# hadoop fs -ls
# hadoop jar /usr/local/hadoop/hadoop-examples-1.0.4.jar wordcount pg1661.txt results
```

Master

(2) wordcount (単語の出現回数のカウント) (2/2)

```
[root@pc8817 hadoop]# hadoop fs -put pg1661.txt /user/root/pg1661.txt

[root@pc8817 hadoop]# hadoop jar /usr/local/hadoop/hadoop-examples-1.0.4.jar \
> wordcount pg1661.txt /user/root/results
11/02/15 10:36:15 INFO input.FileInputFormat: Total input paths to process : 1
11/02/15 10:36:15 INFO mapred.JobClient: Running job: job_201102151030_0002
11/02/15 10:36:16 INFO mapred.JobClient: map 0% reduce 0%
11/02/15 10:36:23 INFO mapred.JobClient: map 100% reduce 0%
11/02/15 10:36:36 INFO mapred.JobClient: map 100% reduce 100%
11/02/15 10:36:38 INFO mapred.JobClient: Job complete: job_201102151030_0002
11/02/15 10:36:38 INFO mapred.JobClient: Counters: 17
11/02/15 10:36:38 INFO mapred.JobClient:   Job Counters
11/02/15 10:36:38 INFO mapred.JobClient:     Launched reduce tasks=1
11/02/15 10:36:38 INFO mapred.JobClient:     Launched map tasks=1
11/02/15 10:36:38 INFO mapred.JobClient:     Data-local map tasks=1
11/02/15 10:36:38 INFO mapred.JobClient:   FileSystemCounters
11/02/15 10:36:38 INFO mapred.JobClient:     FILE_BYTES_READ=218292
11/02/15 10:36:38 INFO mapred.JobClient:     HDFS_BYTES_READ=590077
11/02/15 10:36:38 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=436616
11/02/15 10:36:38 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=158415
11/02/15 10:36:38 INFO mapred.JobClient:   Map-Reduce Framework
11/02/15 10:36:38 INFO mapred.JobClient:     Reduce input groups=15273
11/02/15 10:36:38 INFO mapred.JobClient:     Combine output records=15273
11/02/15 10:36:38 INFO mapred.JobClient:     Map input records=13014
11/02/15 10:36:38 INFO mapred.JobClient:     Reduce shuffle bytes=0
11/02/15 10:36:38 INFO mapred.JobClient:     Reduce output records=15273
11/02/15 10:36:38 INFO mapred.JobClient:     Spilled Records=30546
11/02/15 10:36:38 INFO mapred.JobClient:     Map output bytes=1001185
11/02/15 10:36:38 INFO mapred.JobClient:     Combine input records=106836
11/02/15 10:36:38 INFO mapred.JobClient:     Map output records=106836
11/02/15 10:36:38 INFO mapred.JobClient:     Reduce input records=15273
```

```
[root@pc8817 hadoop]# hadoop fs -ls /user/root/results
Found 2 items
drwxr-xr-x - root supergroup          0 2011-02-15 10:36
/user/root/results/_logs
-rw-r--r-- 3 root supergroup        158415 2011-02-15 10:36
/user/root/results/part-r-00000

[root@pc8817 hadoop]# 結果ファイルをHDFSから/tmpへコピー
[root@pc8817 hadoop]# hadoop fs -get /user/root/results/part-r-00000 /tmp
[root@pc8817 hadoop]# more /tmp/part-r-00000
"A      1
"About 1
"Absolute      1
"Ah!      2
"Ah,      2
"Ample,      1
"And      10
"Arthur!      1
"As      1
"At      1
"Because      1
"Breckinridge, 1
"But      1
"But,      1
"But,      1
"Certainly      2
"Certainly,      1
"Come!      1
"Come,      1
"DEAR      1
"Dear      2
"Dearest      1
"Death,      1
"December      1
"Do      3
"Don't      1
"Entirely,      1
"For      1
"Fritz!      1
```

© Hitachi Information Academy Co., Ltd. 2014. All rights reserved.

3-6 . Map/Reduceジョブの投入 (4/7)

Master

(3) grep (文字列による行の抽出)

コマンド

```
hadoop jar /usr/local/hadoop/hadoop-examples-1.0.4.jar \
> grep 引数1 引数2 引数3
grep 入力ファイル名 出力ディレクトリ 検索文字列
```

(例) シャーロックホームズのファイルから、
文字列“where”にマッチする回数をレポート

```
# hadoop jar /usr/local/hadoop/hadoop-examples-1.0.4.jar \
> grep /user/root/pg1661.txt /user/root/results01 where
(正規表現にマッチした文字列と生起頻度をoutputファイルに出力
出力ファイル名 “part-XXXXX”)
```

```
[root@pc8817 hadoop]# hadoop jar /usr/local/hadoop/hadoop-examples-1.0.4.jar \
grep pg1661.txt /user/root/results01 where
11/02/15 10:42:07 INFO mapred.FileInputFormat: Total input paths to process : 1
11/02/15 10:42:07 INFO mapred.JobClient: Running job: job_201102151030_0005
11/02/15 10:42:08 INFO mapred.JobClient: map 0% reduce 0%
11/02/15 10:42:19 INFO mapred.JobClient: map 100% reduce 0%
11/02/15 10:42:31 INFO mapred.JobClient: map 100% reduce 100%
11/02/15 10:42:33 INFO mapred.JobClient: Job complete: job_201102151030_0005
11/02/15 10:42:33 INFO mapred.JobClient: Counters: 18
11/02/15 10:42:33 INFO mapred.JobClient:   Job Counters
11/02/15 10:42:33 INFO mapred.JobClient:     Launched reduce tasks=1
11/02/15 10:42:33 INFO mapred.JobClient:     Launched map tasks=2
11/02/15 10:42:33 INFO mapred.JobClient:     Data-local map tasks=2
11/02/15 10:42:33 INFO mapred.JobClient:   FileSystemCounters
11/02/15 10:42:33 INFO mapred.JobClient:     FILE_BYTES_READ=38
11/02/15 10:42:33 INFO mapred.JobClient:     HDFS_BYTES_READ=594048
11/02/15 10:42:33 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=146
11/02/15 10:42:33 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=108
11/02/15 10:42:33 INFO mapred.JobClient:   Map-Reduce Framework
11/02/15 10:42:33 INFO mapred.JobClient:     Reduce input groups=1
11/02/15 10:42:33 INFO mapred.JobClient:     Combine output records=2
11/02/15 10:42:33 INFO mapred.JobClient:     Map input records=13014 (中略)
```

```
11/02/15 10:42:33 WARN mapred.JobClient: Use GenericOptionsParser for
11/02/15 10:42:33 INFO mapred.FileInputFormat: Total input paths to
11/02/15 10:42:33 INFO mapred.JobClient: Running job: job_201102151030_0006
11/02/15 10:42:34 INFO mapred.JobClient: map 0% reduce 0%
11/02/15 10:42:43 INFO mapred.JobClient: map 100% reduce 0%
11/02/15 10:42:55 INFO mapred.JobClient: map 100% reduce 100%
11/02/15 10:42:57 INFO mapred.JobClient: Job complete: job_201102151030_0006
11/02/15 10:42:57 INFO mapred.JobClient: Counters: 18
11/02/15 10:42:57 INFO mapred.JobClient:   Job Counters
11/02/15 10:42:57 INFO mapred.JobClient:     Launched reduce tasks=1
11/02/15 10:42:57 INFO mapred.JobClient:     Launched map tasks=1
11/02/15 10:42:57 INFO mapred.JobClient:     Data-local map tasks=1
11/02/15 10:42:57 INFO mapred.JobClient:   FileSystemCounters
11/02/15 10:42:57 INFO mapred.JobClient:     FILE_BYTES_READ=22
11/02/15 10:42:57 INFO mapred.JobClient:     HDFS_BYTES_READ=108
11/02/15 10:42:57 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=76
11/02/15 10:42:57 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=10
11/02/15 10:42:57 INFO mapred.JobClient:   Map-Reduce Framework
11/02/15 10:42:57 INFO mapred.JobClient:     Reduce input groups=1
11/02/15 10:42:57 INFO mapred.JobClient:     Combine output records=0
11/02/15 10:42:57 INFO mapred.JobClient:     Map input records=1
11/02/15 10:42:57 INFO mapred.JobClient:     Reduce shuffle bytes=0
11/02/15 10:42:57 INFO mapred.JobClient:     Reduce output records=1
11/02/15 10:42:57 INFO mapred.JobClient:     Spilled Records=2
11/02/15 10:42:57 INFO mapred.JobClient:     Map output bytes=14
11/02/15 10:42:57 INFO mapred.JobClient:     Map input bytes=22
11/02/15 10:42:57 INFO mapred.JobClient:     Combine input records=0
11/02/15 10:42:57 INFO mapred.JobClient:     Map output records=1
11/02/15 10:42:57 INFO mapred.JobClient:     Reduce input records=1
```

```
[root@pc8817 hadoop]# hadoop fs -ls results01
Found 2 items
drwxr-xr-x - root supergroup          0 2011-02-15 10:42
/user/root/results01/_logs
-rw-r--r-- 3 root supergroup        10 2011-02-15 10:42
/user/root/results01/part-00000
```

```
[root@pc8817 hadoop]# hadoop fs -get /user/root/results01/part-00000 /tmp
[root@pc8817 hadoop]#
[root@pc8817 hadoop]# more /tmp/part-00000
116 where 結果・・・116回出現
0
```

© Hitachi Information Academy Co., Ltd. 2014. All rights reserved.

(1) syslog

/var/log/messages

(例) #tail -n 500 /var/log/messages | grep ERROR

(2)hadoop

Hadoopログ WebUIから取得できます。

リンク:logまたはNamenode Logs を
クリック。

次に、各々のファイル名をクリックすると
ログ内容をエディタで確認できます。

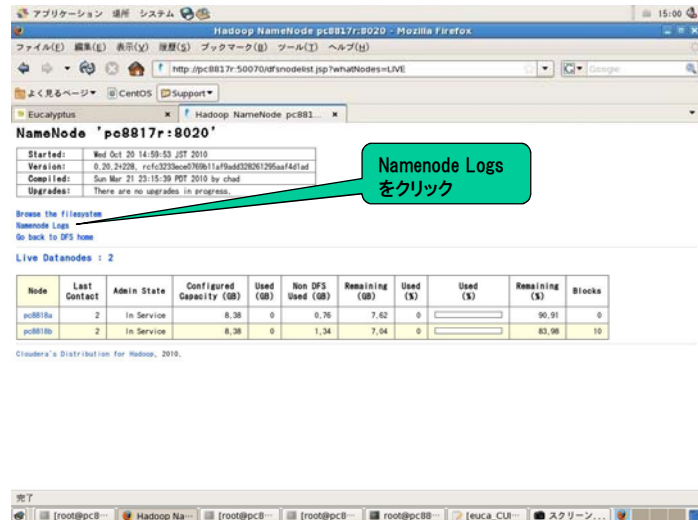
コマンド例:

Namenode / Jobtracker

```
# cd /usr/local/hadoop/logs
# tail -n 500 hadoop-root-namenode-pc8801a.log ¥
> | grep ERROR
# tail -n 500 hadoop-root-jobtracker-pc8801a.log ¥
> | grep ERROR
```

Datanode / Tasktracker

```
# cd /usr/local/hadoop/logs
# tail -n 500 hadoop-root-datanode-pc8801b.log
# tail -n 500 hadoop-root-tasktracker-pc8801b.log
```



© Hitachi Information Academy Co., Ltd. 2014. All rights reserved.

64

4-2. Map/Reduceの問題

(1) MapReduceジョブが異常終了する場合

一時ファイルの削除:

ディレクトリのリスト #hadoop fs -ls (MapReduceの一時ファイルは、恐らく "/user/ユーザ名" にあり。)

一時ファイルの削除 #hadoop fs -rmr 対象ディレクトリ名 (例: "/user/root/PiEstimator_TMP_3_141592654")

(2) エラーが解消しない、またはdatanodeが見えない場合

(エラー: "DataStreamer Exception: ...be replicated 0 nodes, instead of 1")

1. 停止 # stop-all.sh
2. 各マシンにログインしてcacheの削除(それぞれ実行) # rm -fr /usr/local/hadoop/cache*
3. ファイルシステムのフォーマット # hadoop namenode -format
4. 始動 # start-all.sh

(3) ファイルシステム破壊の可能性がある場合

ファイルシステムの整合性チェック

ダメージが大きい場合、ファイルシステムのフォーマット

```
# hadoop fsck /
# stop-all.sh
# hadoop namenode -format
# start-all.sh
```

(4) メモリ(ヒープサイズ)不足で起動しない場合

MapReduceはメモリを大量に消費します。

・仮想サーバインスタンスを起動(ec2run)する時のTypeを確認、実行できる最大RAMサイズで起動してみます。

・hadoopコマンド 実行時のオプションでヒープサイズの指定ができます。(JVM起動エラーの場合)

(例) # hadoop -Xms1000m (1GB確保)

(5) Reduce数の調整 (プログラムのコーディングの時点) ... 演習では非該当

ジョブの終了を早めるためには、Reduce数が1では非効率です。

最大スロット数(デフォルト: 2 x ノード数)よりやや小さめに設定します。

Reduce数 = 0.95X スロット総数(Reduce Task Capacity) = (ノード毎の最大スロット数 x ノード数)

(ノード毎の最大スロット数="mapred.tasktracker.reduce.tasks.maximum" (デフォルト) の設定ファイル: "mapred-site.xml" 参考: 4-3.(3)e.)

これにより、全Reduceジョブは、要求時に即、起動します。

Reduce数 = 1.75Xスロット総数 ... Reduce処理終了時ノードは次のReduceを処理できるが、負荷が増大する。

(参考: マニュアル "MapReduce Tutorial")

各サービスは起動していますか?

Master側
jps
Namenode
Jobtracker

Slave側
jps
Datanode
Tasktracker

© Hitachi Information Academy Co., Ltd. 2014. All rights reserved.

65