Web Mining Tutorial

Hadoop, Hive and Pig

Ajay Dubey, Romil Bansal, Jayant Gupta

August 18, 2013

Doubt Clarification

- Text Indexing and Crawling
- Relevance Ranking
- Similarity Search
- Link Analysis Algorithms
- LSI and EM

Apache Hadoop Installation

Pre-requisites

- Install Java 1.6 or above
- Download Hadoop Version **1.2.1**
- For Single Node Cluster
  - http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/
- For Multi Node Cluster
  - http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-multi-node-cluster/

Apache Hadoop Configurations

In configurations folder (HADOOP-HOME/conf) make changes in

- hadoop-env.sh
- core-site.xml
- mapred-site.xml
- hdfs-site.xml

Apache Hadoop Configurations cont.

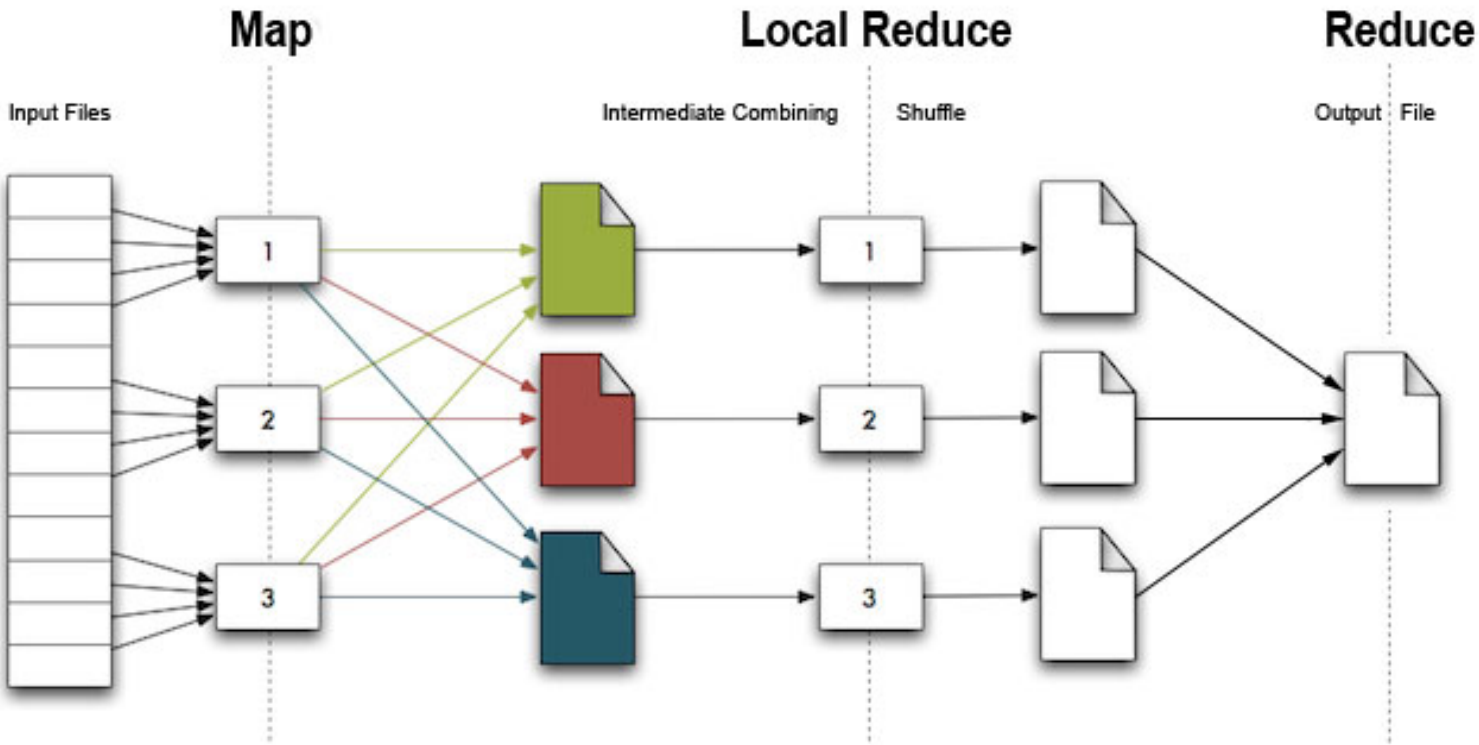
For multi node cluster setup

- make changes in folder (HADOOP-HOME/conf)
  - masters
  - slaves
- in /etc folder
  - hosts file must contain IP addresses of masters and slave machines
- ssh access from master to slave without password

Hadoop commands

- bin/hadoop namenode -format
- bin/start-dfs.sh
- bin/start-mapred.sh
- bin/hadoop dfs -ls
- bin/hadoop dfs -copyFromLocal ⟨local-dir⟩ ⟨hdfs-dir⟩
- bin/hadoop dfs -copyToLocal ⟨hdfs-dir⟩ ⟨local-dir⟩
- bin/hadoop jar hadoop-examples-1.2.1.jar wordcount data output

Map Reduce Architecture



Word Count Example

Map Function

```
public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {
   private final static IntWritable one = new IntWritable(1);
   private Text word = new Text();

   public void map(LongWritable key, Text value, Context context) throws IOExceptio
       String line = value.toString();
       StringTokenizer tokenizer = new StringTokenizer(line);
       while (tokenizer.hasMoreTokens()) {
           word.set(tokenizer.nextToken());
           context.write(word, one);
       }
   }
}
```

Reduce Function

```
public static class Reduce extends Reducer<Text, IntWritable, Text, In

   public void reduce(Text key, Iterable<IntWritable> values, Context
     throws IOException, InterruptedException {
       int sum = 0;
       for (IntWritable val : values) {
           sum += val.get();
       }
       context.write(key, new IntWritable(sum));
   }
}
```

Jar File Creation

Export as a jar file and use class name containing both Map and Reduce

Apache Pig

Installation

- Download Apache Pig Version **0.11.1**
- export PATH of your PIG-Home

Execution

- Local Mode
  pig -x local
- Map Reduce Mode
  pig OR pig -x mapreduce

Apache Pig

Installation

- Download Apache Pig Version **0.11.1**
- export PATH of your PIG-Home

Execution

- Local Mode
  pig -x local
- Map Reduce Mode
  pig OR pig -x mapreduce

Apache Pig cont.

Pig Latin Statements

- A LOAD statement reads data from the file system.
- A series of "transformation" statements process the data.
- A STORE statement writes output to the file system; or, a DUMP statement displays output to the screen.

Apache Hive

Installation

- Download Apache Hive Version **0.10.0**
- export PATH of your HIVE-Home

Execution (only in Map Reduce way)

- hive

Apache Hive Commands

Hive Commands

- hive: SHOW TABLES;
- hive: CREATE TABLE wordcount (words STRING, count INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY tab STORED AS TEXTFILE;
- hive: DESCRIBE wordcount;
- hive: LOAD DATA INPATH words INTO TABLE wordcount;
- hive: LOAD DATA LOCAL INPATH

Apache Hive Advanced Features

TRANSFORM: Allows user to stream data through user-defined scripts.

user: cat split.py

```
for line in sys.stdin:
for word in line.split():
print word
```

Add file split.py

```
INSERT OVERWRITE TABLE test
SELECT TRANSFORM(line)
USING   python split.py
AS word
FROM wordcount;
```

References

- Apache Hadoop Map Reduce Tutorial

  http://hadoop.apache.org/docs/stable/mapred_tutorial.html

- Apache Pig Installation

  http://pig.apache.org/docs/r0.7.0/tutorial.html

- Apache Pig Reference Manual 1

  http://pig.apache.org/docs/r0.8.1/piglatin_ref1.html

- Apache Pig Reference Manual 2

  http://pig.apache.org/docs/r0.8.1/piglatin_ref2.html

- Introduction to Hive

  http://blog.cloudera.com/wp-content/uploads/2010/01/6-IntroToHive.pdf

- Apache Hive Tutorial

  https://cwiki.apache.org/confluence/display/Hive/Tutorial

Thank You