L2: Understanding MapReduce Fundamentals (MapReduce 原理)

Feng Li

feng.li@cufe.edu.cn

School of Statistics and Mathematics

Central University of Finance and Economics

Revised on November 30, 2016

Today we are going to learn...
(本节知识要点)

1. **Hadoop Streaming（Hadoop 流）**
2. **Other Hadoop API（其他 Hadoop 接口）**
3. **Exercises**

Hadoop Streaming
(Hadoop 流)

- Hadoop provides an API to MapReduce that allows you to write your map and reduce functions in languages other than Java.
- Hadoop Streaming uses Unix standard streams as the interface between Hadoop and your program, so you can use any language that can read standard input and write to standard output to write your MapReduce program.
- The complete Hadoop Streaming Documentation can be found from Hadoop Installation directory ``share/doc/hadoop/hadoop-mapreduce-client/hadoop-mapreduce-client-core/HadoopStreaming.html''

A toy word count example (运行一个词频统计的示例) I

- mapper: `/usr/bin/cat`

  concatenate files and print on the standard output.
- reducer: `/bin/wc`

  print newline, word, and byte counts for each file, and a total line if more than one file is specified.
- test file: any text documents in HDFS
- The syntax

A toy word count example
(运行一个词频统计的示例) II

```
$ ~/hadoop/bin/hadoop jar \
  ~/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.5.2.jar \
  -input /stocks.txt \
  -output wcoutfile \
  -mapper "/bin/cat" \
  -reducer "/usr/bin/wc" \
```

- **NOTE**
  - The backslash at the end of each line allow you to break a long command into many lines.
  - Quotation marks should be used if the file is not at HDFS, e.g. the `mapper` and `reducer` options.

Streaming and map-only R
(运行一个只用 Mapper 的实例)

- Just like with regular MapReduce, you can have a map-only job in Streaming and R.
- Map-only jobs make sense in situations where you don't care to join or group your data together in the reducer.

Streaming with Python
(将 Python 代码在 Hadoop 上运行)

- In principle, Streaming can be easily applied to Python version MapReduce.
- You only need to make the input and output be standard in Linux.
- Use `#!/usr/bin/python` at the beginning of your Python script.

Other Hadoop API
(其他 Hadoop 接口） I

- Hadoop with Java MapReduce
  - Hadoop is written in Java. There are rich classes of Java MapReduce modules ready to use.
  - You need `javac` (in JDK) and `hadoop-mapreduce-client-core-xxx.jar` to comile your jar files

```
$ javac -classpath \
  ~/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-client-core-2.5.
   -d FirstJar\
$ jar -cvf FirstJar.jar -C FirstJar/
```

  - The Java version syntax is

```
$ hadoop/bin/hadoop jar FirstJar.jar [mainClass] input output
```

Other Hadoop API
(其他 Hadoop 接口) II

- Hadoop Pipe
  - Hadoop Pipes is the name of the C++ interface to Hadoop MapReduce.
  - Pipes uses sockets as the channel over which the tasktracker communicates with the process running the C++ map or reduce function.
  - The application links against the Hadoop C++ library, which is a thin wrapper for communicating with the tasktracker child process.
  - You have to compile and link your C++ program before send it to Hadoop
  - The Hadoop Pipe syntax

```
$ hadoop pipes \
  -D hadoop.pipes.java.recordreader=true \
  -D hadoop.pipes.java.recordwriter=true \
  -input sample.txt \
  -output output \
  -program myCPProgram
```

Assignment (II)

- Find a real problem and solve it with Python and Hadoop streaming.
- Write down a python version of word count with Hadoop streaming.