Hadoop - Map Reduce Application Development: (start) Summary:

1. Download & Install :
   VMware fusion (Mac)
   Cloudera quick start vm (CentOS 64)
   keka (unzip)

login Hue: username: cloudera; password: cloudera;

2. Upload datafile to hadoop filesystem:

Method 1: commend line: (recommend)

```
/ ** Test to access cloudera
    * get cloudera ip address
    * use ssh to access cloudera
    * exit cloudera
**/
[cloudera@quickstart ~]$ ifconfig -→ inet addr:192.168.180.132
Angelas-MBP-2:~ Angela_Crabby$ ssh cloudera@192.168.180.132 →
[cloudera@quickstart ~]$ exit

/ ** Copy data file from Angela_Crabby to hadoop filesystem
   * copy file: deckofcards.txt to cloudera, path: /home/cloudera
   * put the file to hadoop HDFS
   * check hadoop HDFS file
**/
Angelas-MBP-2:Desktop Angela_Crabby$ scp deckofcards.txt
cloudera@192.168.180.132:/home/cloudera
[cloudera@quickstart ~]$ hadoop fs -put deckofcards.txt
/user/cloudera
[cloudera@quickstart ~]$ hadoop fs -ls

method 2: UI
Angela_Crabby browser: http://192.168.180.132:8888/  -> find file
Browser  → upload → find the file you want to upload
Find the upload file from Namenode UI:
http://192.168.180.132:50070/  → utilities → browser the file
system → user → cloudera  → click file and check the block ID
```

3. Download & Install:
   Java JDK;
   Spring Tool Suite

```
/ ** Add <repositories> & <dependencies> in porm.xml
   * in JAVA IDE, spring tool suite ,  new a Maven Project,  find: porm.xml
   * google: "maven repository for cloudera"
   * check hadoop MapReduce version and modify dependency
** /
```

```
<repositories>
    <repository>
      <id>cloudera</id>
      <url>https://repository.cloudera.com/artifactory/cloudera-
repos/</url>
    </repository>
  </repositories>
```

// Check hadoop MapReduce version and modify dependency:

```
[cloudera@quickstart ~]$ find /usr -name "hadoop-mapreduce-
client-common*.jar"

<dependency>
     <groupId>org.apache.hadoop</groupId>
     <artifactId>hadoop-common</artifactId>
     <version>2.6.0-cdh5.5.0</version>
   </dependency>

   <dependency>
     <groupId>org.apache.hadoop</groupId>
     <artifactId>hadoop-yarn-common</artifactId>
     <version>2.6.0-cdh5.5.0</version>
   </dependency>

   <dependency>
     <groupId>org.apache.hadoop</groupId>
     <artifactId>hadoop-mapreduce-client-common</artifactId>
     <version>2.6.0-cdh5.5.0</version>
   </dependency>

   <dependency>
     <groupId>org.apache.hadoop</groupId>
     <artifactId>hadoop-mapreduce-client-core</artifactId>
     <version>2.6.0-cdh5.5.0</version>
   </dependency>
```

4. MapReduce Program

* check input data： `Angelas-MBP-2:Desktop Angela_Crabby$ view deckofcards.txt`

* three critical programs:

(1) map function: input: deckofcards.txt; output: <"count", 1> src: -- mappers/ RecordMapper.java

(2) reduce function: input: <"count", {1, 1, 1, 1, ... }>; output: 52; src: -- reducers/ NoKeyRecordCountReducer.java

(3) job configuration; src: -- drivers/ RowCount.java

* export MapReduce Program into a JAR package

After finished modifying these three functions, in spring tool suits: File → export → java → JAR File → select project: Cards, and give the name of JAR file → finish

// check the jar file we want to upload:

```
Angelas-MBP-2:workspace-sts-3.7.2.RELEASE Angela_Crabby$ jar tvf
StartMapReducer.jar
    25 Fri Apr 08 16:40:44 EDT 2016 META-INF/MANIFEST.MF
   534 Wed Apr 06 21:34:44 EDT 2016 .project
  2355 Fri Apr 08 12:50:34 EDT 2016
lab/cards/reducers/NoKeyRecordCountReducer.class
  1606 Thu Apr 07 12:20:14 EDT 2016 pom.xml
   996 Wed Apr 06 21:34:46 EDT 2016 .classpath
   533 Thu Apr 07 12:21:02 EDT 2016 lab/cards/App.class
  1905 Fri Apr 08 12:31:56 EDT 2016
lab/cards/mappers/RecordMapper.class
  2183 Thu Apr 07 22:47:48 EDT 2016
lab/cards/drivers/RowCount.class
```

5. Compile and Execute map reduce function:

* copy StartMapReducer.jar to cloudera

```
Angelas-MBP-2:workspace-sts-3.7.2.RELEASE Angela_Crabby$ scp
StartMapReducer.jar cloudera@192.168.180.138:/home/cloudera
```

* check whether the jar file copy to cloudera successfully

```
Angelas-MBP-2:~ Angela_Crabby$ ssh cloudera@192.168.180.138
[cloudera@quickstart ~]$ ls -ltr
```

* upload input file and check HDFS has input file: deckofcards.txt

```
[cloudera@quickstart ~]$ hadoop fs -put deckofcards.txt
/user/cloudera
[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera
```

* run the map reduce program:

// jar package: StartMapReducer.jar;
// target job configure class: lab.cards.drivers.RowCount;
// input file: /user/cloudera/deckofcards.txt;
// output file: /user/cloudera/output

```
[cloudera@quickstart ~]$ hadoop jar StartMapReducer.jar
lab.cards.drivers.RowCount /user/cloudera/deckofcards* /user/cloudera/output
```

6. see the output of map reduce program:

```
[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera/output
```

```
   // this is the number of reducer, if only one, then it has one
part-r-00000
/user/cloudera/output/part-r-00000
```

* cat the output file:

```
[cloudera@quickstart ~]$ hadoop fs -cat
/user/cloudera/output/part-r-00000
output: 52
```