Name:Pratheeban Panchalingam **Reg No:984636**

Steps to set up a single node cluster and optionally an eclipse development environment to create and test map reduce programs.

Development Environment Setup

1. Download Cloudera Quickstart vm 5.5 64 bit from bellow url.
   http://www.cloudera.com/content/www/en-us/downloads/quickstart_vms/5-5.html
2. Unzip cloudera-quickstart-vm-5.5.0-0-vmware.zip file into hard drive.
3. Download VMware 12 Player from bellow url and install it.
   https://www.vmware.com/products/player
4. Open the Cloudera Virtual Machine by using VMware Player.
5. The Coudera contains following software. (we don't need to install separately)
   - Cent OS 6.4
   - Eclipse Luna
   - Apache Hadoop 2.6
   - Java 7

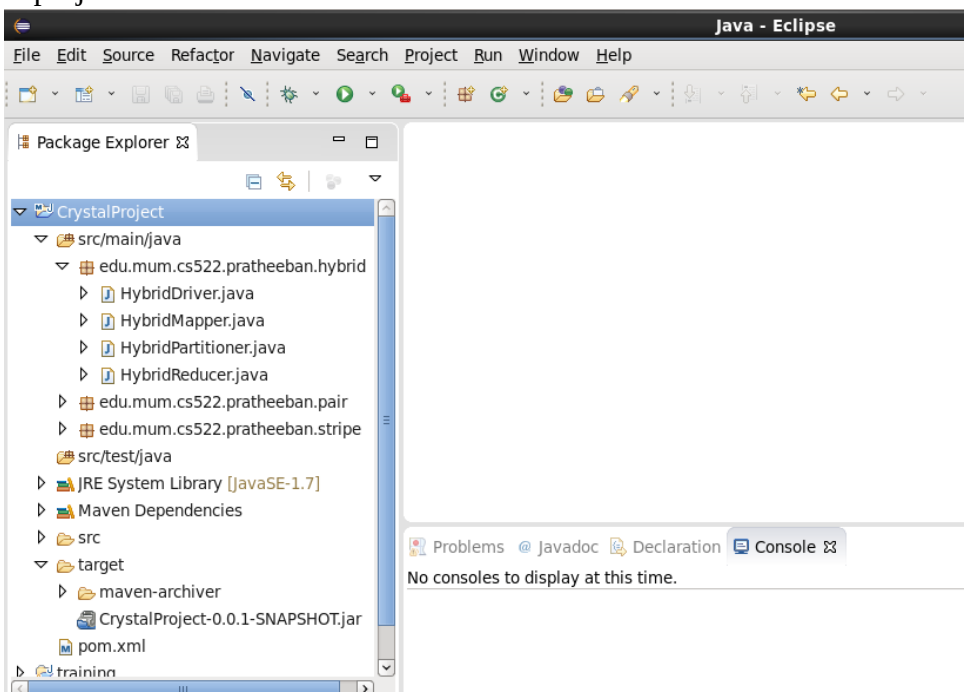
Map Reduce Programs Implementation

1. Create a new maven project in eclipse.
2. Configure pom.xml for map reduce development.

```
<dependency>
        <groupId>org.apache.hadoop</groupId>
        <artifactId>hadoop-client</artifactId>
        <version> 2.6.0-mr1-cdh5.5.0 </version>
</dependency>
<repository>
        <id>cloudera-repo</id>
        <url>http://repository.cloudera.com/artifactory/cloudera-repos/</url>
</repository>
```

3. Create following packages in Eclipse for each problem
   I. edu.mum.cs522.pratheeban.pair
   II. edu.mum.cs522.pratheeban.stripe
   III. edu.mum.cs522.pratheeban.hybrid
4. Create and implement the Mapper, Reducer, Partitioner and Driver classes with necessary extends and methods for each problem
   Example
   I. PairMapper class - Extend hadoop.mapreduce.Mapper class and override setup(), map(), cleanup() methods
   II. PairReducer class - Extend hadoop.mapreduce.Reducer class andoverride reduce() method
   III. PairPartitioner class - Extend hadoop.mapreduce.Partitioner class and override getPartiton() method
   IV. PairDriver class - Extend org.apache.hadoop.conf.Configured class

5. The project structure is shown in below

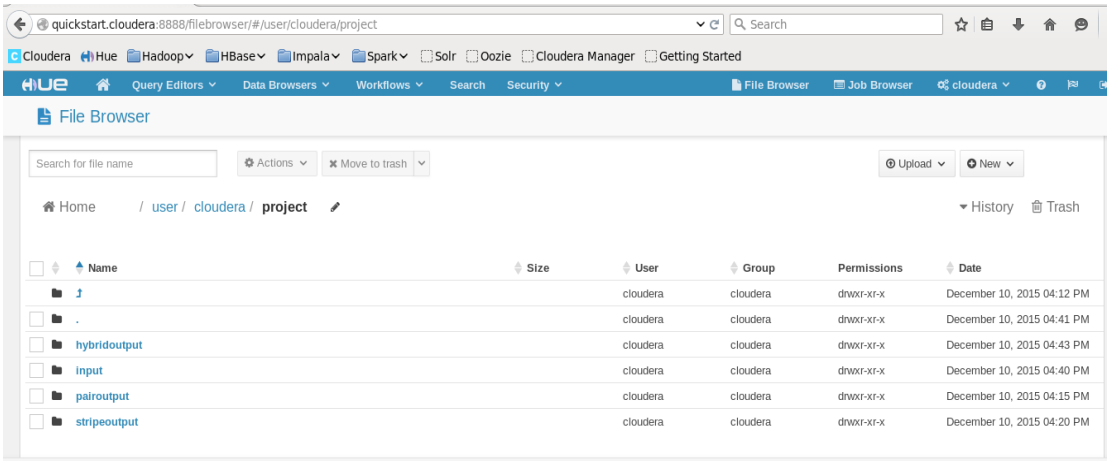


Compile and execute Map Reduce Programs

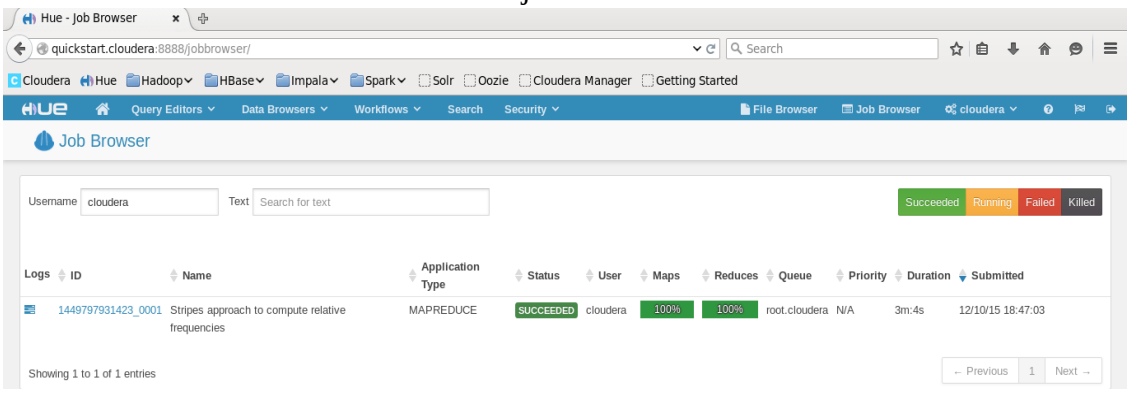
1. Build maven project(Right click on project in Eclipse and select Run As -> Maven build).
2. Copy the jar file in the target directory in Eclipse project into directory or go to file in Eclipse and select Export-->java-->jar and then click on next and then select the project and give the name and then click on Finish.
3. Create an input file with given input (input.txt).
4. Creates a input directory named path in HDFS.
   hadoop fs -mkdir -p /user/cloudera/project/input
5. Copy the input file from the local file system to directory within the HDFS.
   hadoop fs -put input.txt /user/cloudera/project/input/
6. Execute the jar file with created HDFS input by using Linux command in Terminal
   1. Execute Pair Program
      hadoop jar CrystalProject edu.mum.cs522.pratheeban.pair.PairDriver /user/cloudera/project/input/input.txt /user/cloudera/project/pairoutput
   2. Execute Stripe Program
      hadoop jar CrystalProject edu.mum.cs522.pratheeban.stripe.StripeDriver /user/cloudera/project/input/input.txt /user/cloudera/project/stripeoutput
   3. Execute Pair Program
      hadoop jar CrystalProject edu.mum.cs522.pratheeban.hybrid.HybridDriver /user/cloudera/project/input/input.txt /user/cloudera/project/hybridoutput

Monitor job and View the output

1. The output can be viewed in Terminal
   hadoop fs -cat /user/cloudera/project/pairoutput/*
   hadoop fs -cat /user/cloudera/project/stripeoutput/*
   hadoop fs -cat /user/cloudera/project/hybridoutput/*
2. The output can also be viewed in Hue
   1. Go to the browser and click on Hue in browser
   2. Give username as cloudera and password as cloudera
   3. Click on File Browser and then go to the directory and open the file



3. Click on Job Browser in Hue to monitor job status.



4. Click on job id to view more information about job.