README

Yaxin Huang

02/01/2015

1. What's in this Folder Huang_Yaxin

1. `MapReduce_A1_v1`
   - `src`: contains source codes.
   - `build.xml`: This is the build file for Ant.
2. `MapReduce_A1_v2`: same structure as v1.
3. `MapReduce_A1_v3`: same structure as v1.
4. `MapReduce_A1_v4`: same structure as v1.
5. `pseudo-distributed-config`: The configuration directory for running pseudo-distributed mode.
6. `run_v1`: The script to run version 1.
7. `run_v2_pseudodist`: The script to run version 2 in pseudo-distributed mode.
8. `run_v2_standalone`: The script to run version 2 in standalone mode.
9. `run_v3_pseudodist`: The script to run version 3 in pseudo-distributed mode.
10. `run_v3_standalone`: The script to run version 3 in standalone mode.
11. `run_v4_pseudodist`: The script to run version 4 in pseudo-distributed mode.
12. `run_v4_standalone`: The script to run version 4 in standalone mode.
13. README.pdf

2. Before Running Any of Them ...

1. All these programs can be run in Linux environment. And you need to have hadoop binaries in your machine.
2. Make sure you've already set the `JAVA_HOME` and `HADOOP_HOME` in your ~/.bashrc:

   ```
   $ export JAVA_HOME=/usr/lib/jvm/java-6-sun
   $ export HADOOP_HOME=/home/yaxin/hadoop-x.y.z
   $ export PATH=$PATH:$HADOOP_HOME/bin
   ```

3. By doing the step two, you can start the Terminal and run:

   ```
   $ hadoop version
   ```

   to get your hadoop version.
4. Copy the input file `purchases4.txt` to directory `Huang_Yaxin`.
5. Make sure you can ssh to localhost without entering password.

3. How to Run Version 1

1. In your terminal, cd to `Huang_Yaxin`
2. Run the following command to gain permission:

```
$ chmod a+rwx ./run_v1
```

3. Run the following command to start the program:

```
$ ./run_v1
```

4. The expected running time in a machine with 8GB memory and 2.00GHz*4 CPU is 330 seconds. The result file resides in `MapReduce_A1_v1/output.txt`.

4. How to Run Version 2

Run in Standalone Mode

1. In your terminal, cd to `Huang_Yaxin`
2. Run the following command to gain permission:

```
$ chmod a+rwx ./run_v2_standalone
```

3. Run the following command to start the program:

```
$ ./run_v2_standalone
```

4. The expected running time in a machine with 8GB memory and 2.00GHz*4 CPU is 125 seconds. The result resides in `MapReduce_A1_v2/output`.

Run in Pseudo-Distributed Mode

1. Run the following command in your Terminal to ssh to localhost:

```
$ ssh localhost
```

2. Cd to `Huang_Yaxin`
3. Run the following command to gain permission:

```
$ chmod a+rwx ./run_v2_pseudodist
```

4. Run the following command to start the program:

```
$ ./run_v2_pseudodist
```

   You will need to answer "Y" during the process since the program will ask for permission to format the name node.

5. The expected running time in a machine with 8GB memory and 2.00GHz*4 CPU is 125 seconds. The result resides in `MapReduce_A1_v2/output`.

5. How to Run Version 3

Run in Standalone Mode

1. In your terminal, cd to `Huang_Yaxin`
2. Run the following command to gain permission:

   ```
   $ chmod a+rwx ./run_v3_standalone
   ```

3. Run the following command to start the program:

   ```
   $ ./run_v3_standalone
   ```

4. The expected running time in a machine with 8GB memory and 2.00GHz*4 CPU is 680 seconds. The result resides in `MapReduce_A1_v3/output`.

Run in Pseudo-Distributed Mode

1. Run the following command in your Terminal to ssh to localhost:

   ```
   $ ssh localhost
   ```

2. Cd to `Huang_Yaxin`
3. Run the following command to gain permission:

   ```
   $ chmod a+rwx ./run_v3_pseudodist
   ```

4. Run the following command to start the program:

   ```
   $ ./run_v3_pseudodist
   ```

   You will need to answer "Y" during the process since the program will ask for permission to format the name node.

5. The expected running time in a machine with 8GB memory and 2.00GHz*4 CPU is 715 seconds. The result resides in `MapReduce_A1_v3/output`.

6. How to Run Version 4

Run in Standalone Mode

1. In your terminal, cd to `Huang_Yaxin`
2. Run the following command to gain permission:

   ```
   $ chmod a+rwx ./run_v4_standalone
   ```

3. If you want to change the argument N (In this version the program will calculate the Fibonacci of N as well, but the result won't be output), you can update the last number in the 4th command in `run_v4_standalone`. The default is 5.
4. Run the following command to start the program:

   ```
   $ ./run_v4_standalone
   ```

5. The expected running time in a machine with 8GB memory and 2.00GHz*4 CPU is 200 seconds(N=5). The result resides in `MapReduce_A1_v4/output`.

Run in Pseudo-Distributed Mode

1. Run the following command in your Terminal to ssh to localhost:

   ```
   $ ssh localhost
   ```

2. Cd to `Huang_Yaxin`

3. Run the following command to gain permission:

   ```
   $ chmod a+rwx ./run_v4_pseudodist
   ```

4. If you want to change the argument N (In this version the program will calculate the Fibonacci of N as well, but the result won't be output), you can update the last number in the 4th command in `run_v4_standalone`. The default is 5.

5. Run the following command to start the program:

   ```
   $ ./run_v4_pseudodist
   ```

   You will need to answer "Y" during the process since the program will ask for permission to format the name node.

6. The expected running time in a machine with 8GB memory and 2.00GHz*4 CPU is 200 seconds(N=5). The result resides in `MapReduce_A1_v4/output`.