Chapter 1

Assignment-01

Objective

Upon completion of this assignment you will be able to;

- Understand and modify MapReduce programs
- Understand the MapReduce process and the major variables available for optimization
- Understand the reason for error handling in schemaless data
- Be able to explain the effect and the reason for the impact of MapReduce Options

Outcomes

You will have worked through a matrix of MapReduce options run against the sample MaxTemperature.java code. You will be able to explain the outcome of these results from the textbook and in your own words. You will use additional charts and debug output to describe the output.

Assignment Pre-requisits

0. If not already done - do the steps below (do not use your existing Vagrant box, make a second one)
1. Create new Vagrant box
2. Clone my sample code at: https://github.com/illinoistech-itm/jhajek
    - Needed setup files under the ITMD-521 > config folder
3. Install hadoop 2.8.5 and configure it so you recieve the same output to the hadoop version command
4. Test your connection by running the `hadoop fs -ls /user/controller/ncdc` command

Assignment Core

Based on MaxTemperature sample code we will modify various application and MapReduce based settings:

Deliverable In GitHub create a folder named **Week-09** and then under that folder create a folder for each experiment and name it: experiement-01-60 (where XX is the number of the bullet point list below).

Place all source code for each experiement (*.java files and Readme.md conatining the set parameters.) based on the MaxTemperature source code provided in chapter 02 of the hadoop-book source code into those folders. You are to submit the URL to your GitHub account ot Blackboard. It will be due: 11:59 PM ~~04/04/19~~ 04/07/19.

In the Week-09 folder, provide a single **Readme.md** report that performs the analysis of each group of 4 tests per dataset and the impact of the feature optimazations based on execution time per job. Seperate each report with an Hortizontal Rule (or three — in markdown) and an h2 as the title. See https://github.com/illinoistech-itm/jhajek/tree/master/itmd-521/template for the template to use. Explain how each optimzation increased perforance or negatively impacted performance. Use References from the text book with page numbers (The term, "it is faster or slower," is not valid).

There will be three datasets used in this assignment. If your school Anumber is odd use (1990.txt, 80.txt, 80-90.txt) if even use (~~1970.txt~~1982.txt, 60-70.txt, 200.txt).

Matrix Assignment

For *each* dataset, run the MaxTemperature Job with these parameters:

1. Number of Reducers 1, Intermediate Compression off, Combiner off
2. Number of Reducers 2, Intermediate Compression off, Combiner off
3. Number of Reducers 4, Intermediate Compression off, Combiner off
4. Number of Reducers 8, Intermediate Compression off, Combiner off
5. Number of Reducers 1, Intermediate Compression On (Snappy), Combiner off
6. Number of Reducers 2, Intermediate Compression On (Snappy), Combiner off
7. Number of Reducers 4, Intermediate Compression On (Snappy), Combiner off
8. Number of Reducers 8, Intermediate Compression On (Snappy), Combiner off
9. Number of Reducers 1, Intermediate Compression On (Snappy), Combiner on
10. Number of Reducers 2, Intermediate Compression On (Snappy), Combiner on
11. Number of Reducers 4, Intermediate Compression On (Snappy), Combiner on
12. Number of Reducers 8, Intermediate Compression On (Snappy), Combiner on
13. Number of Reducers 1, Intermediate Compression On (Gzip), Combiner off
14. Number of Reducers 2, Intermediate Compression On, (Gzip) Combiner off
15. Number of Reducers 4, Intermediate Compression On, (Gzip) Combiner off
16. Number of Reducers 8, Intermediate Compression On, (Gzip) Combiner off
17. Number of Reducers 1, Intermediate Compression On (Gzip), Combiner on
18. Number of Reducers 2, Intermediate Compression On, (Gzip) Combiner on
19. Number of Reducers 4, Intermediate Compression On, (Gzip) Combiner on

20. Number of Reducers 8, Intermediate Compression On, (Gzip) Combiner on

Report Structure

In the report and analysis phase you will aggregate the execution time results accross sets of 4 and then accross the 3 data types. Grouping the results to derive a conclusion based on our textbook. You will need to analyze the outcomes of the groups of 4 tasks per data set, but will not be graded on them as you will be graded on the 3 dataset comparison.

You will have 5 analysis paragraphs and 1 conclusion paragraph makeing overall technical observations and conclusions. 4 point scale per item, 1 point for you name. Totla of 25 points.

References

https://hadoop.apache.org/docs/current/api/org/apache/hadoop/mapreduce/Job.html

http://hadoop.apache.org/docs/r2.8.5/hadoop-mapreduce-client/hadoop-mapreduce-client-core/mapred-default.xml.

Example code

```
// to set the number of reducers in the Job object, 1 is default
// https://hadoop.apache.org/docs/r2.8.5/api/org/apache/hadoop/mapreduce/Job.html#setN
job.setNumReduceTasks(2);

// Turm Intermediate Compression on in Job Object
// http://bigdatums.net/2016/11/17/compressing-intermediate-map-output-in-hadoop/
//turn on intermediate (map output) compression

import org.apache.hadoop.conf.*;
import org.apache.hadoop.io.compress.CompressionCodec;
import org.apache.hadoop.io.compress.GzipCodec;
import org.apache.hadoop.io.compress.SnappyCodec;

job.getConfiguration().setBoolean("mapreduce.map.output.compress", true);
job.getConfiguration().setClass("mapreduce.map.output.compress.codec",
GzipCodec.class, CompressionCodec.class);

// or
job.getConfiguration().setBoolean("mapreduce.map.output.compress", true);
job.getConfiguration().setClass("mapreduce.map.output.compress.codec",
SnappyCodec.class, CompressionCodec.class);

// set each job name with parameters and your initials in the job object
job.setJobName("use-this-string");
```

```
// to turn on combiner -- see the source code in MaxTemperatureWithCombiner.java
```

Access to Grafana

To view our hadoop cluster metrics

https://192.168.1.4:3000 - note this is a self-signed certificate and it will warn you but you can accept it.

viewer

itmd521viewer