Exercise 10: Geo-Distributed Hadoop with Amazon Web Services (cont.)

TECHNISCHE UNIVERSITÄT DARMSTADT

Concepts and Technologies for Distributed Systems and Big Data Processing – SS 2016

Task 1 Prepare two Small Hadoop Cluster

Setup two identical small Hadoop clusters in two distinct data centers, *EU (Ireland)* and *US West (Oregon)*. Use 4 machines (*t2.large*) for each cluster like in the previous exercise. Download data_1 (`https://www.dropbox.com/s/d2xme9j7wiv6koa/data1.zip?dl=0`) to the 1st cluster (cluster_1), and data_2 (`https://www.dropbox.com/s/9r3bwqm5bbarit6/data2.zip?dl=0`) to the 2nd cluster (cluster_2).

Test each setup with a small (< 1GB) MR word count example.

Task 2 "Old-fashioned:" Copy all data to 1 datacenter, perform job

Execute the experiment as follows:

1) Copy data_1 from cluster_1 to cluster_2 and measure the transfer time (t_1)
2) Extract and concatenate data_1 and data_2
3) Use cluster_2 to run MR word count on the aggregated input files and measure the time (t_2)
4) Repeat the experiment twice and take the average of the run time ($t_1 + t_2$)

Task 3 Perform mapping and reducing in respective datacenters

Prepare the experiment: Write a simple script for merging the results of two MR word count outputs.

Execute the experiment as follows:

1) Extract data_1 and data_2 in respective datacenters
2) Use cluster_1 to run MR word count on data_1 (and measure the time t_1)
3) Use cluster_2 to run MR word count on data_2 (and measure the time t_2)
4) Copy the results of cluster_2 to cluster_1 (and measure the time t_3)
5) Aggregate both results by using your script (and measure the time t_4)
6) Repeat the experiment twice and take the average of the run time ($t_1 + t_2 + t_3 + t_4$)

Task 4 Perform mapping in respective datacenters, allocate all reducers in 1 datacenter

Prepare the experiment: Modify your word count experiment in order to perform mapping in both clusters (cluster_1 and cluster_2), but allocate all reducers in cluster_1.

Execute the experiment as follows:

1) Extract data_1 and data_2 in respective datacenters
2) Start your experiment timer and ...
3) Perform mapping of data_1 and data_2 simultaneously in cluster_1 and cluster_2
4) Allocate all reducers in cluster_1 and process the intermediate results
5) Repeat the experiment twice and take the average of the run time

Task 5 Discussion

Compare your results of all three setups (Task 2, Task 3, and Task 4).

Task 6 [Bonus] How to Efficiently Perform Word Count on the Given Data Set?

Describe your suggested execution path for these two clusters and the given data sets. Discuss your ideas with other students. If you like, prepare your experiment and measure the performance improvements. Which strategy performs best? Submit your best results (+ execution path).