MGS:655 – Distributed Computing and Big Data Technologies
Trigram Count Project using MapReduce Hadoop program

1)Team Composition and Contribution

Handled the Trigram Count Project alone, done the configuration changes on Apache Hadoop 2.7.3 setup on my Mac system to connect to Center for Computational Research (CCR) cluster. Downloaded OCR text file from the given "Library of Congress" Chronicling America-Historic American Newspapers. Coded the mapper and reducer java program on Eclipse Mars IDE. Finally ran the executable for the taken inputs and collected the output trigram. Followed by report on comparing the mapper and reducer numbers ran and performance impact on trigram output.

2)Connection to CCR Cluster

- Downloaded the Apache Hadoop 2.7.3 from hadoop.apache.org, changed the configuration files. Core-site.xml to listen on hdfs://euca-199-109-194-186.cloud.ccr-cbls-2.ccr.buffalo.edu:9000 for namenode which is a metadata for HDFS, hdfs-site.xml to "clients should use datanode hostnames when connecting to datanodes" condition, mapred-site.xml to run on CCR cluster with mapreduce job history on euca-199-109-192-91.cloud.ccr-cbls-2.ccr.buffalo.edu:10020, yarn-site.xml to tell NodeManagers that there will be an auxiliary service to shuffle the mapper and reducer and the node manager sits on CCR cluster.
- Created the directory in input folder, and observed in the CCR utilities Browse file system for the presence of the set up paths.
- Run the pre-existing Word Count program on downloaded 1 text file for checking the working state of cluster on the cluster and not on the local machine by jps command
- Set up Eclipse IDE-Mars version on MAC OS. Write mapper and reducer code for trigram
- Create jar file. Run on the dataset from the terminal. Observe the mapper, reducer and output and report.

3)Data sets

Wrote script to download 45 OCR files from the given site. Selected 45 files as my UB personal number ends with 45. Selected the data set from 1920-07-03, 1920-08-03 (yyyy-dd-mm) which has 18 articles each, and 1920-09-03 13 articles. Obtained from script by introducing a loop to identify the date and pages. Chose this data set as the data has huge English writings. Constructed a single file, dual and triple by concatenating via script to see the mapper behavior and duplicated 2 files to observe reducer behavior.

4)Design Issues

Designed the Java project by taking Hadoop core 1.2.4 jar library as an external reference.

Created the Jar file as executable, wrote the automated script to run the executable from fetching the file from internet, Executing the jar file and taking output file from hdfs system.

Case 1) Designed one Mapper with one, two, three reducer. Written a code to introduce required reducers

Observation: As all the 45 files were concatenated, the output has the continuous trigram

Case 2) Designed two mapper, three mapper with one , two, three Reducer

Observation: As 45 files divided into two files and assigned two mappers first then divided into three files and assigned three mappers.

The mapper output differs as there is no continuation in the file, if we read separated file context to the mapper for running the Trigram. So to have consistency of output, ran Trigram with below design.

- Designed Mapper to read more than one text file from the DFS location (to act for multi mapper) and to read each text line (so that multi mapper can be acted, without having output differences). Used single file split to needs, which are downloaded from url to hdfs path directly via script
- Designed script with all the mapper, reducer cases to be run from mac terminal via shell script

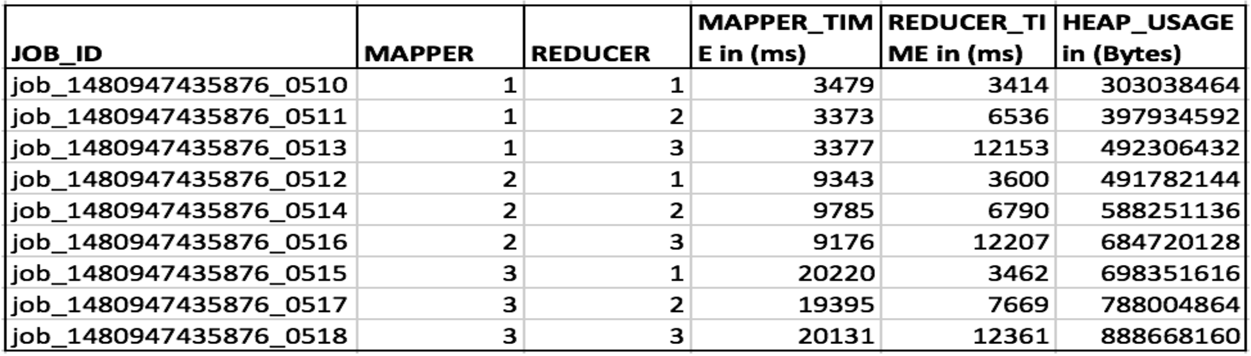
- Placed checking for input and output directory input provided in the command line
- Created own java package file
- Added script to delete the directory in hdfs before and togather more than one reducer file output to a single file after execution, and to count unique mapper.

5)Empirical Results and Discussions on Trigram

Initially as a part of Hadoop set up, taken the inbuilt wordcount jar and ran one text file newspaper. Observed one mapper and one reduce with mapper task time 2565 ms and reducer task time 2674 ms. size of the block is 128MB, as the input size is directly proportional to the overall usage of blocks. In order to reduce the cost of overhead for namenode kept block size as 128MB. Ran Bigram code which was thought in the class

General observation:

Top trigram: one of the, the New York, the United States, League of Nations, out of the, member of the It is very common to have “the New York” and “the United States” in this dataset as the observed paper is of New York state and USA country. As observed in the newspaper generally has “one of the” trigram. Trigram is useful in identifying the great scholars of our past, who contributed veraciously to literature.
Unique trigram: 150515 using grep on output file



JOB_ID	MAPPER	REDUCER	MAPPER_TIME in (ms)	REDUCER_TIME in (ms)	HEAP_USAGE in (Bytes)
job_1480947435876_0510	1	1	3479	3414	303038464
job_1480947435876_0511	1	2	3373	6536	397934592
job_1480947435876_0513	1	3	3377	12153	492306432
job_1480947435876_0512	2	1	9343	3600	491782144
job_1480947435876_0514	2	2	9785	6790	588251136
job_1480947435876_0516	2	3	9176	12207	684720128
job_1480947435876_0515	3	1	20220	3462	698351616
job_1480947435876_0517	3	2	19395	7669	788004864
job_1480947435876_0518	3	3	20131	12361	888668160

Performance Observation: Used Hadoop core jar file to capture mapper and reducer timing.

1) Outputs were similar as done Trigram on line basis, only had shuffled trigram
2) From the different cases showed above, mapper time increases as mappers are increased
3) From the different cases showed above, reducer time increases as reducers are increased
4) CPU time is more in for increased mappers and increased reducer case
5) Heap usage increases as the mapper and reducers are increased

As the file used is comparatively less, It can be handled per mapper, reducer basis. In the industry the memory required for computation is more. For the better performance of trigram use multiple reducers.

Practical Experiences

Challenges encountered: Initially for local machine set up of Hadoop JAVA 8 jdk and jre were used, as the CCR runs on JAVA 7, compiled on Major version 52 for JAVA 8 running on Minor version JAVA 7 observed. Work around, on mac set up the java 7 and recompile and replace every link to java 8. In mac JAVA_HOME pointed to /user/lib directory, as Hadoop required java from /bin directory. Mac has System Integrity Protection was not able to create soft link in /bin directory for java 1.8-jdk. Changed the booting restore of mac and granted the write access and worked ahead. To analyze the multiple reducer event cluster was very slow due to many users sometimes.
Industrial need of trigram to analyze for complete document should be achieve very carefully, by mapping within taken block size allocated to mappers