Chaitanya Durgesh Nynavarapu
A20561894

ASSIGNMENT-2

BIG DATA (CSP 554)

1. A Comprehensive Performance Analysis of Apache Hadoop and Apache Spark for Large-Scale Data Sets Using HiBench

This study addresses a critical question in big data processing: Can optimized parameter selection improve cluster performance for large datasets in Hadoop and Spark frameworks? The researchers hypothesized that careful tuning of system parameters could significantly enhance performance beyond default settings.

Methodology

- Workloads: WordCount and TeraSort
- Performance metrics: execution time, throughput, and speedup
- 18 key parameters tuned using trial-and-error approach
- Real cluster deployment with datasets up to 600 GB

Key Findings

1. Performance heavily depends on input data size and parameter selection
2. Spark outperforms Hadoop for smaller datasets, while Hadoop shows better performance for larger datasets (300 GB and 600 GB)
3. Increasing task parallelism improves CPU utilization and reduces job execution time
4. Memory allocation and serialization significantly impact performance

Critical Analysis

The study's strength lies in its comprehensive approach, using larger datasets and more parameters than previous research. However, the use of only two workloads limits the generalizability of the findings. The trial-and-error approach, while practical, may not capture the full complexity of parameter interactions.The results convincingly demonstrate the impact of parameter tuning on performance, supporting the authors' hypothesis. However, the study could benefit from a more systematic exploration of parameter combinations to provide stronger guidelines for optimization.

Significance and Future Work

This research fills a crucial gap by providing insights into large-scale data processing optimization. It offers practical recommendations for system administrators and researchers, potentially improving efficiency in big data applications across industries.

Future work could extend this research by:

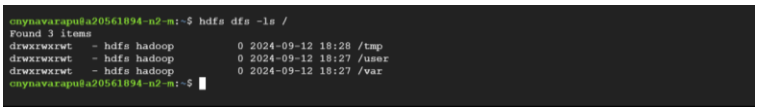
1. Incorporating a wider range of workloads to increase generalizability
2. Developing machine learning models to predict optimal parameter settings
3. Investigating the impact of hardware configurations on parameter optimization

Challenges and Limitations

- Complex interplay between parameters makes optimization difficult
- Performance gains from parameter tuning vary based on workload and data size
- The study doesn't address the long-term stability of optimized configurations

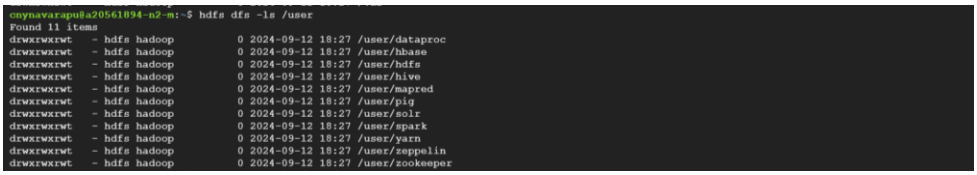
In conclusion, while this study provides valuable insights into Hadoop and Spark performance optimization, it also highlights the need for more sophisticated approaches to parameter tuning in big data frameworks. The research contributes significantly to the field by demonstrating the importance of tailored configurations for large-scale data processing, paving the way for more efficient and cost-effective big data solutions.

2. HDFS command to list the files and directories



- $ hdfs dfs -ls /

3. Command to list the files and directories under HDFS directory



- $ hdfs dfs -ls /user

4. Command to create HDFS Directory

- $ hdfs dfs -mkdir /user/csp554

5. Command to create HDFS Directory

- $ hdfs dfs -mkdir /user/csp554-5

6. Command that copies a given local file to the given hdfs directory

- $ hdfs dfs -copyFromLocal /home/hadoop/Chaitanya.txt /user/csp554/

7. Copy a file from one hdfs directory to another hdfs directory

```
cnynavarapu@a20561894-n2-m:~$ hdfs dfs -cp /user/csp554/Chaitanya.txt /user/csp554-5/
cnynavarapu@a20561894-n2-m:~$
```

- $ hdfs dfs -cp /user/csp554/Chaitanya.txt /user/csp554-5/

8. Copying the object myid.txt you uploaded to a GCS bucket into the Hadoop master node Linux file system.

```
cnynavarapu@a20561894-n2-m:~$ gsutil cp gs://a20561894/A20561894.txt /home/hadoop/A20561894.txt
Copying gs://a20561894/A20561894.txt...
/ [1 files][   21.0 B/   21.0 B]
Operation completed over 1 objects/21.0 B.
cnynavarapu@a20561894-n2-m:~$ ls /home/hadoop/
A20561894.txt  Chaitanya.txt
cnynavarapu@a20561894-n2-m:~$
```

9. Copying the same object myid.txt you created in an GCS bucket into HDFS into the directory /users/csp554 and listing the files and directories under the hdfs directory.

```
cnynavarapu@a20561894-n2-m:~$ hdfs dfs -cp gs://a20561894/A20561894.txt /user/csp554-5
2024-09-12 21:19:31,807 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-09-12 21:19:32,033 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-09-12 21:19:32,033 INFO impl.MetricsSystemImpl: google-hadoop-file-system metrics system started
Sep 12, 2024 9:19:33 PM com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.util.RequestTracker stopTracking
INFO: Detected high latency for [url=https://storage.googleapis.com/storage/v1/b/a20561894/o/A20561894.txt?fields=bucket,name,timeCreated,updated,generation,metageneration,size,contentType,contentEncoding,md5Hash,crc32c,metadata; invocationId=gl-java/11.0.20 gdcl/2.1.1 linux/6.1.0 gccl-invocation-id/0a212e19-5df4-4bb1-ad97-b8ae7134b416], durationMs=415; method=GET [CONTEXT ratelimit_period="10 SECONDS" ]
Sep 12, 2024 9:19:33 PM com.google.cloud.hadoop.fs.gcs.GhfsGlobalStorageStatistics trackDuration
INFO: periodic connector metrics: {gcs_api_time=821, gcs_api_total_request_count=2, gcs_connector_time=1118, gcs_list_file_request=1, gcs_list_file_request_max=407, gcs_list_file_request_mean=407, gcs_list_file_request_min=407, gcs_metadata_request=1, gcs_metadata_request_max=414, gcs_metadata_request_mean=414, gcs_metadata_request_min=414, gs_filesystem_create=2, gs_filesystem_initialize=1, op_get_file_status=1, op_get_file_status_max=1118, op_get_file_status_mean=1118, op_get_file_status_min=1118, op_glob_status=1, uptimeSeconds=3} [CONTEXT ratelimit_period="5 MINUTES" ]
cnynavarapu@a20561894-n2-m:~$ hdfs dfs -ls /user/csp554-5
Found 2 items
-rw-r--r--   2 cnynavarapu hadoop         21 2024-09-12 21:19 /user/csp554-5/A20561894.txt
-rw-r--r--   2 cnynavarapu hadoop         23 2024-09-12 21:17 /user/csp554-5/Chaitanya.txt
cnynavarapu@a20561894-n2-m:~$
```

- $ hdfs dfs -cp gs://a20561894/A20561894.txt /user/csp554-5
- $ hdfs dfs -ls /user/csp554-5

10. Executing a command to show the contents of the myid.txt file in the hdfs directory

```
cnynavarapu@a20561894-n2-m:~$ hdfs dfs -cat /user/csp554-5/A20561894.txt
this is the id file

cnynavarapu@a20561894-n2-m:~$ hdfs dfs -rm /user/csp554-5/A20561894.txt
Deleted /user/csp554-5/A20561894.txt
cnynavarapu@a20561894-n2-m:~$ hdfs dfs -ls /user/csp554-5
Found 1 items
-rw-r--r--   2 cnynavarapu hadoop         23 2024-09-12 21:17 /user/csp554-5/Chaitanya.txt
cnynavarapu@a20561894-n2-m:~$
```

- $ hdfs dfs -cat /user/csp554-5/A20561894.txt

11. Execute a command to remove the myid.txt file in the hdfs directory /user/csp554-5

```
cnynavarapu@a20561894-n2-m:~$ hdfs dfs -rm /user/csp554-5/A20561894.txt
Deleted /user/csp554-5/A20561894.txt
cnynavarapu@a20561894-n2-m:~$ hdfs dfs -ls /user/csp554-5
Found 1 items
-rw-r--r--   2 cnynavarapu hadoop         23 2024-09-12 21:17 /user/csp554-5/Chaitanya.txt
cnynavarapu@a20561894-n2-m:~$
```

- $ hdfs dfs -rm /user/csp554-5/A20561894.txt
- $ hdfs dfs -ls /user/csp554-5