



Distributed Systems Assignment

Track 1 - MapReduce

By - Abhishek Agrawal



Index

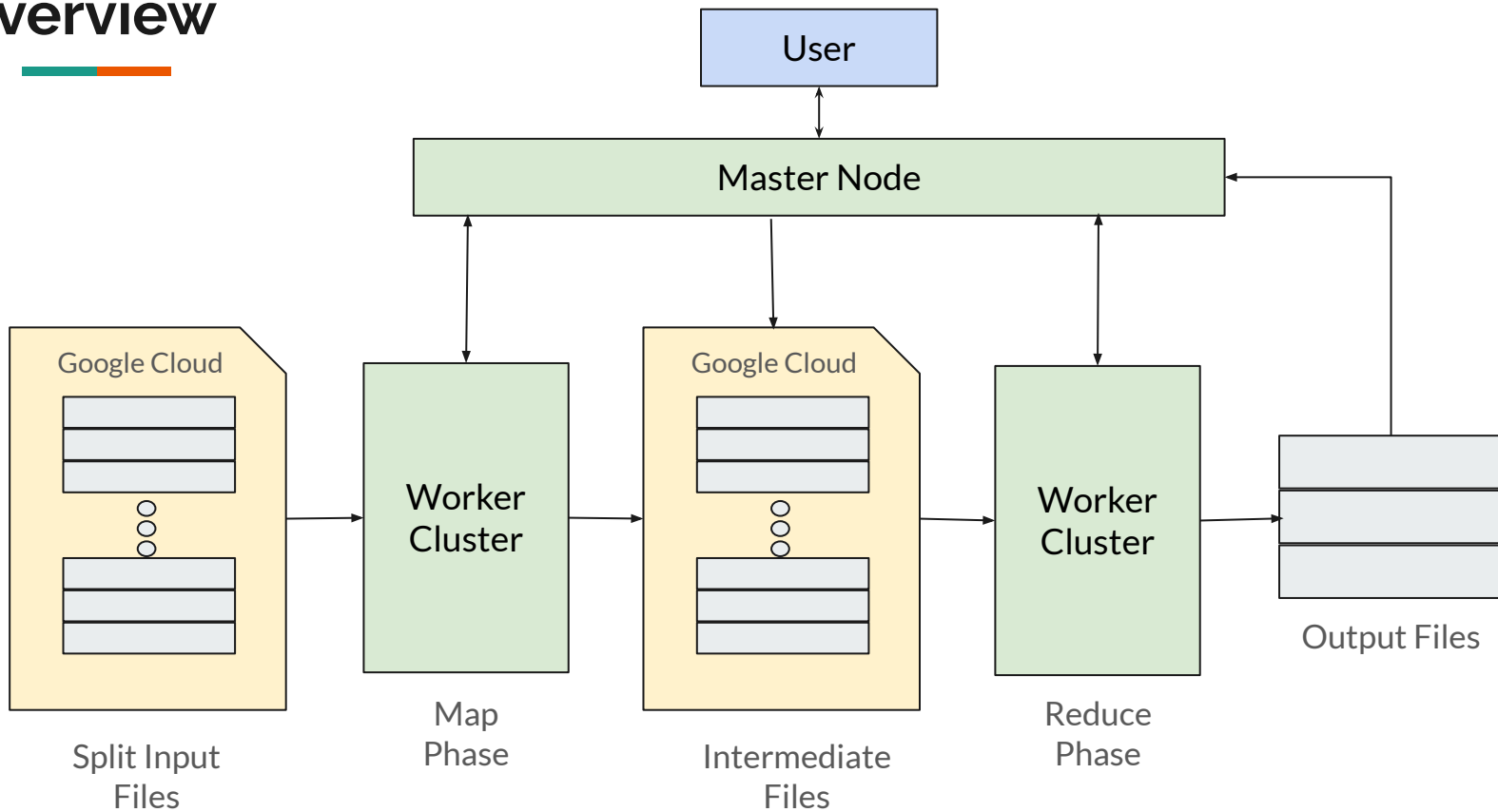
- 1) Problem Statement
- 2) Overview
- 3) Workflow
- 4) Component 1 - GCS
- 5) Component 2 - Master Node
- 6) Component 3 - Worker Node
- 7) Demo
- 8) Improvements
- 9) References



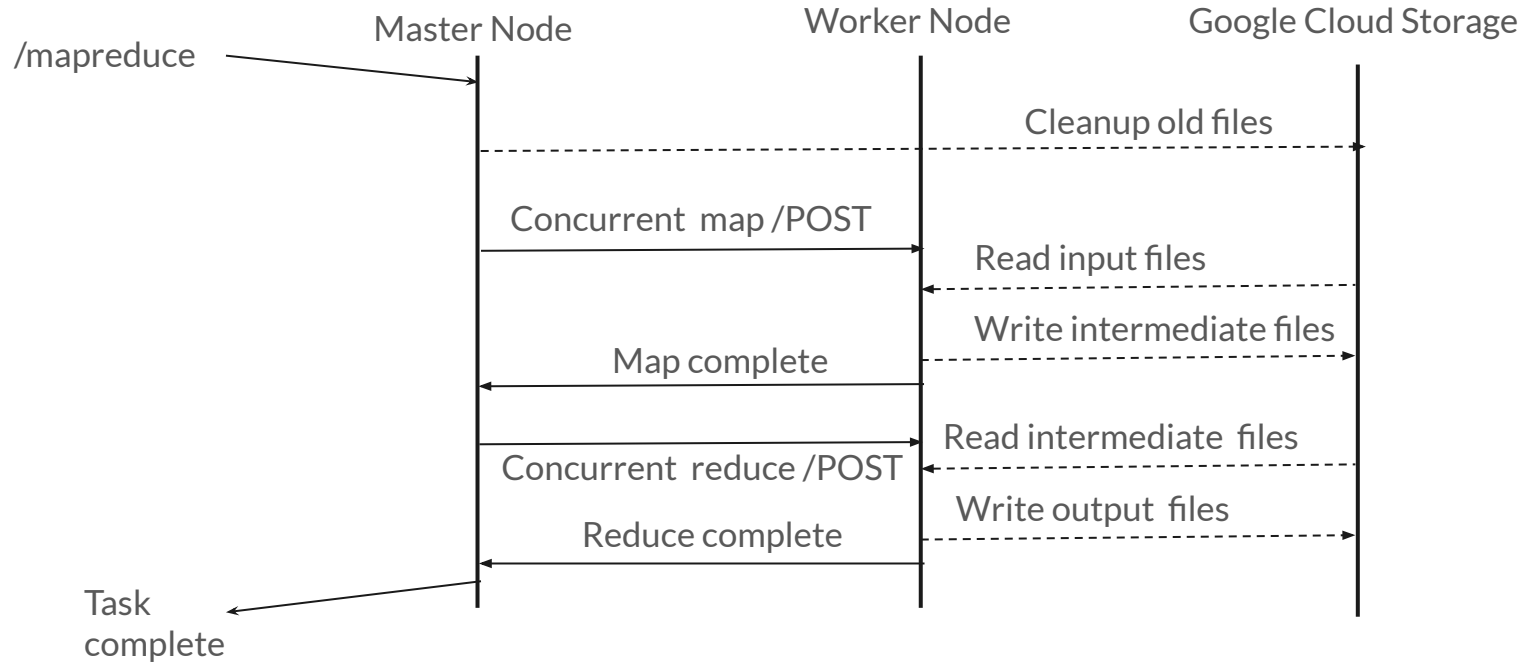
Problem Statement

- 1) Count Frequency of Words in given data
- 2) Employ Distributed Architecture - Map Reduce

Overview







Workflow



Google Cloud Storage

Component 1

- 1) Distributed File System
- 2) Store Unstructured Data
- 3) Benefits -
 - a) Object Lifecycle Management
 - b) Read/Write Lock
 - c) Powerful APIs
- 4) Cons -
 - a) Management to be done by User

Name
 <u>customer_trends/</u>
 <u>generate_trends.py</u>
 <u>intermediate_files/</u>
 <u>output_files/</u>



Master Node

Built in GOLANG-

- 1) GO Routines
- 2) Wait Groups
- 3) Light Weight
- 4) Easy to Learn

GOLANG HTTP Server -

- 1) /mapreduce
 - a) Starts the workflow
 - b) Returns the steps
- 2) /output
 - a) Retrieves output files
 - b) Formats to send
- 3) /cleanup
 - a) Manual remove Intermediate Files before next run



Worker Node

- 1) Also Built in GO
- 2) Stateless (Almost)

- 1) /map
 - a) Reads Input Files from GCS
 - b) Builds a word-map
 - c) Hashes to 2 files
 - d) Compress and Encode as gob
 - e) Writes files as -
0_bucket_uuid.gob
- 2) /reduce
 - a) Reads Intermediate files from one bucket
 - b) Decompresses each file
 - c) Aggregate output of each file
 - d) Format and Write to Output

Demo



Further Improvements

- 1) Fault Tolerance in nodes
- 2) Deployment to GCP - VM
- 3) Inter-Node Communication
- 4) Service Discovery
- 5) Config Driven
- 6) Evaluation and Statistics



References

- [1] J. Dean and S. Ghemawat, “Mapreduce: Simplified data processing on large clusters,” in OSDI’04: Sixth Symposium on Operating System Design and Implementation, (San Francisco, CA), pp. 137–150, 2004.
- [2] <https://go.dev/learn/>
- [3] <https://pkg.go.dev/encoding/gob>
- [4] <https://cloud.google.com/storage/docs/introduction>