

# Enron Email Analysis - Big Data Cloud Project

## Abstract

The objective of this project is to solve a Big Data problem on the Cloud. This includes data collection, implementation, and performance analysis of an end-to-end application to analyze the Enron Email Dataset. The analysis identifies key email senders, communication patterns, and provides insight into possible fraudulent activities.

## 1. Problem Description

The Enron Email Dataset contains emails from over 158 employees, which were made public during the investigation into the Enron scandal. The dataset is large, unstructured, and difficult to process with conventional tools. The problem is to extract valuable insights, such as top senders, recipients, and potential indicators of fraudulent activity.

## 2. Need for Big Data and Cloud

The dataset size is around 1.7 GB and consists of over 600,000 emails. Handling and processing such large datasets require Big Data tools and Cloud infrastructure. Google Cloud Dataproc and Hadoop were chosen to manage the distributed storage and parallel processing. Without such infrastructure, it would be infeasible to process this dataset on a single machine.

## 3. Data Description

- **Source:** The Enron Email Dataset is available on Kaggle.
- **Size:** Approximately 1.7 GB with over 600,000 emails.
- **Structure:** Each email contains metadata like 'From', 'To', 'Date', and 'Message'.
- **Storage:** The dataset is stored in Google Cloud Storage for distributed access.

## 4. Description of the Application

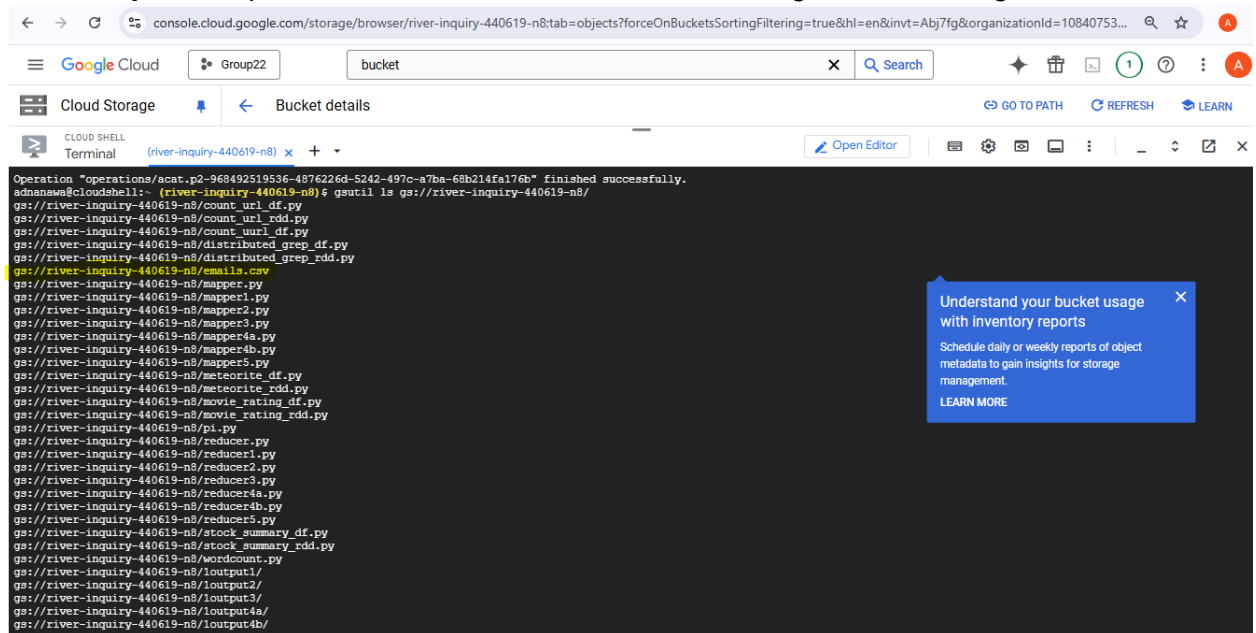
- **Programming Models:** Spark and Hadoop MapReduce.
- **Platform:** Google Cloud Dataproc.
- **Infrastructure:** The cluster is set up with 1 master node and 2 or 4 worker nodes for performance analysis.

## 5. Software Design

- **Architectural Design:** Data is stored on Google Cloud Storage and processed on Google Cloud Dataproc using PySpark.
- **Code Baseline:** PySpark script `email_analysis.py` was written to extract relevant information from the emails, such as the top senders, top recipients, and email frequencies.
- **Dependencies:** PySpark, Google Cloud SDK, Hadoop, and necessary Google Cloud libraries.

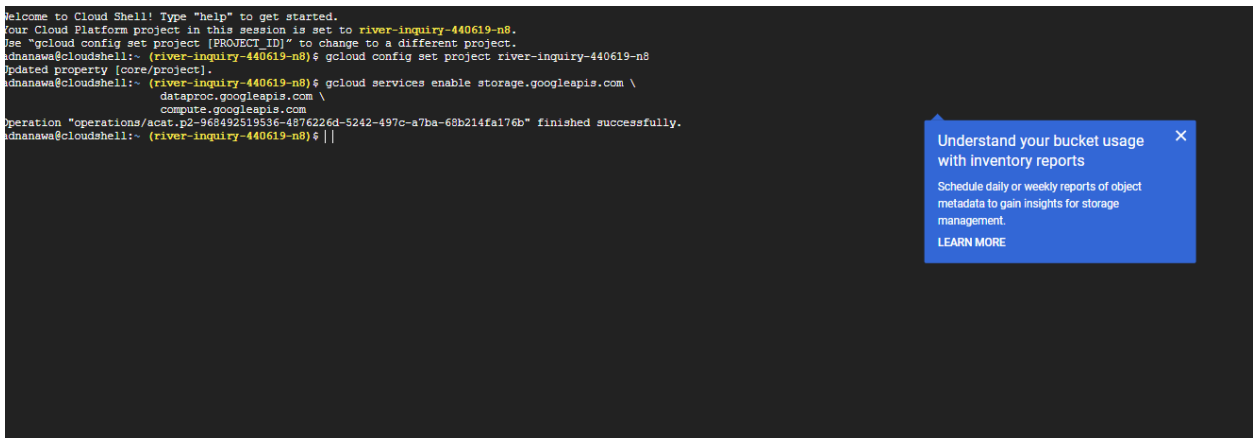
## 6. Usage

### 1. Data Upload: Upload the Enron Email Dataset to the Google Cloud Storage bucket.



### 2.Enable Services:

gcloud services enable compute.googleapis.com dataproc.googleapis.com  
storage.googleapis.com



### 3 Upload your Enron email file: (assume it's named "emails.csv")

gsutil cp /path/to/your/emails.csv gs://river-inquiry-440619-n8/input/emails.csv

Note: If you don't have the "emails.csv" file, download it from Kaggle and place it on your local machine.

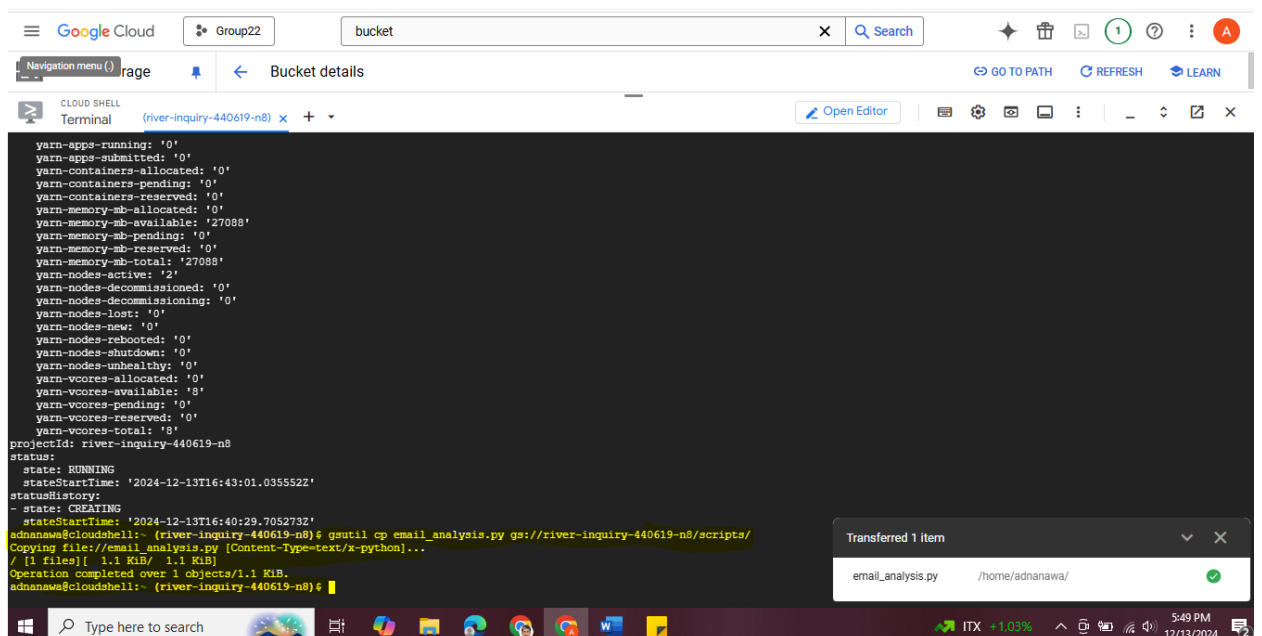
**3.Setup:** Create a Dataproc cluster with the 2 nodes using the following command:

gcloud dataproc clusters create enron-cluster --region=europe-southwest1 --num-workers=2

```
adnanawa@cloudshell:~ (river-inquiry-440619-n8) $ gcloud dataproc clusters create enron-cluster \
--region=europe-southwest1 \
--master-machine-type=e2-standard-4 \
--master-boot-disk-size=50 \
--worker-machine-type=e2-standard-4 \
--worker-boot-disk-size=50 \
--num-workers=2 \
--image-version=2.0-debian10 \
--enable-component-gateway
Waiting on operation [projects/river-inquiry-440619-n8/regions/europe-southwest1/operations/989893b4-b2e3-3a9f-8693-816f3129c4c].
Waiting for cluster creation operation...
WARNING: For PD-Standard without local SSDs, we strongly recommend provisioning 1TB or larger to ensure consistently high I/O performance. See https://cloud.google.com/compute/docs/disks/performance for information on disk I/O performance.
WARNING: The firewall rules for specified network or subnetwork would allow ingress traffic from 0.0.0.0/0, which could be a security risk.
WARNING: The specified custom staging bucket 'dataproc-staging-europe-southwest1-968492519536-ixp1sti' is not using uniform bucket level access IAM configuration. It is recommended to update bucket to enable the same. See https://cloud.google.com/storage/docs/uniform-bucket-level-access.
Waiting for cluster creation operation...done.
Created [https://dataproc.googleapis.com/v1/projects/river-inquiry-440619-n8/regions/europe-southwest1/clusters/enron-cluster] Cluster placed in zone [europe-southwest1-c].
adnanawa@cloudshell:~ (river-inquiry-440619-n8) $ gcloud dataproc clusters describe enron-cluster --region=europe-southwest1
clusterName: enron-cluster
clusterDuid: eae90da4-a02a-49ee-8fa7-b0ebafd81f26
config:
  configBucket: dataproc-staging-europe-southwest1-968492519536-ixp1sti
  endpointConfig:
    enableHttpPortAccess: true
  httpPorts:
    HDFS NameNode: https://ykudon2i4nhehpdh5o3nix6fgq-dot-europe-southwest1.dataproc.googleusercontent.com/hdfs/dfshealth.html
    HiveServer2 (enron-cluster-m): https://ykudon2i4nhehpdh5o3nix6fgq-dot-europe-southwest1.dataproc.googleusercontent.com/hiveserver2ui/enron-cluster-m?host=enron-cluster-m
    MapReduce Job History: https://ykudon2i4nhehpdh5o3nix6fgq-dot-europe-southwest1.dataproc.googleusercontent.com/jobhistory/
```

**4.Script Upload:** Upload the email\_analysis.py script to the Cloud Storage bucket using the command:

gsutil cp email\_analysis.py gs://river-inquiry-440619-n8/scripts/



## 5.Run the Job: Run the email analysis on the Dataproc cluster:

gcloud dataproc jobs submit pyspark gs://river-inquiry-440619n8/scripts/email\_analysis.py\--cluster=enron-cluster \--region=europe-southwest1

```
CLOUD SHELL
Terminal (river-inquiry-440619-n8) x +
Open Editor

24/12/14 00:03:52 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find 'resource-types.xml'.
24/12/14 00:03:52 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1734108133577_0011
24/12/14 00:03:53 INFO org.apache.hadoop.yarn.client.RMPProxy: Connecting to ResourceManager at enron-cluster-m/10.204.0.62:8030
24/12/14 00:03:55 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException
; verified object already exists with desired state.
Total number of valid emails: 32303
+-----+
| From|count|
+-----+
| Kay.Mann@enron.com| 1482|
| Matthew.Lenhart@e...| 870|
| Jeff.Dasovich@enr...| 637|
| Vince.J.Kaminski@...| 566|
| Eric.Bass@enron.com| 436|
| Sara.Shackleton@e...| 348|
| Tana.Jones@enron.com| 345|
| Chris.Germany@enr...| 304|
| erwillam@hotmail.com| 290|
| Kim.Ward@enron.com| 286|
+-----+
only showing top 10 rows

+-----+-----+
| From|To|count|
+-----+-----+
| Kay.Mann@enron.com| Kay.Mann@enron.com| 1482|
| Matthew.Lenhart@e...|Matthew.Lenhart@e...| 870|
| Jeff.Dasovich@enr...|Jeff.Dasovich@enr...| 637|
| Vince.J.Kaminski@...|Vince.J.Kaminski@...| 566|
| Eric.Bass@enron.com| Eric.Bass@enron.com| 436|
| Sara.Shackleton@e...|Sara.Shackleton@e...| 348|
| Tana.Jones@enron.com|Tana.Jones@enron.com| 345|
+-----+-----+

Job [c5d27ac698444380830249243bdae6d] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-europe-southwest1-968492519536-1xpplst1/google-cloud-dataproc-metainfo/eae90da4-a02a-49ee-8fa7-b0ebafd81f26/jobs/c5d27ac698444380830249243bdae6d/
driverOutputResourceUri: gs://dataproc-staging-europe-southwest1-968492519536-1xpplst1/google-cloud-dataproc-metainfo/eae90da4-a02a-49ee-8fa7-b0ebafd81f26/jobs/c5d27ac698444380830249243bdae6d/driveroutput
jobUuid: 654ff3a8-031b-30c4-9f37-ed0c8f82e1e1
placement:
  clusterName: enron-cluster
  clusterUuid: eae90da4-a02a-49ee-8fa7-b0ebafd81f26
pysparkJob:
  mainPythonFileUri: gs://river-inquiry-440619-n8/scripts/email_analysis.py
reference:
  jobId: c5d27ac698444380830249243bdae6d
  projectId: river-inquiry-440619-n8
status:
  state: DONE
  stateStartTime: '2024-12-14T00:06:59.0938302'
statusHistory:
- state: PENDING
  stateStartTime: '2024-12-14T00:03:43.3910332'
- state: SETUP_DONE
  stateStartTime: '2024-12-14T00:03:43.4114012'
- details: Agent reported job success
  state: RUNNING
  stateStartTime: '2024-12-14T00:03:43.5890652'
yarnApplications:
- name: Enron Email Analysis - Multiline Support
  progress: 1.0
  state: FINISHED
  trackingUri: http://enron-cluster-m:8080/proxy/application_1734108133577_0011/
adnanawa@cloudshell:~ (river-inquiry-440619-n8) $
```

## 6.Setup: Create a Dataproc cluster with the 4 nodes using the following command:

```
CLOUD SHELL
Terminal (river-inquiry-440619-n8) x +
Open Editor

Job [48dde03dde4041d0955d1d52f63b6a6d] finished successfully.
Done: true
DriverControlFilesUri: gs://dataproc-staging-europe-southwest1-968492519536-ixpplsti/google-cloud-dataproc-metainfo/ea90da4-a02a-49ee-8fa7-b0ebafd81f26/jobs/48dde03dde4041d0955d1d52f63b6a6d/
DriverOutputResourceUri: gs://dataproc-staging-europe-southwest1-968492519536-ixpplsti/google-cloud-dataproc-metainfo/ea90da4-a02a-49ee-8fa7-b0ebafd81f26/jobs/48dde03dde4041d0955d1d52f63b6a6d/driveroutput
JobUuid: 50c3a341-3449-32b3-a521-1183baebd7f
Placement:
  clusterName: enron-cluster
  clusterUuid: ea90da4-a02a-49ee-8fa7-b0ebafd81f26
PySparkJob:
  mainPythonFileUri: gs://river-inquiry-440619-n8/scripts/email_analysis.py
Reference:
  jobId: 48dde03dde4041d0955d1d52f63b6a6d
  projectId: river-inquiry-440619-n8
Status:
  state: DONE
  stateStartTime: '2024-12-14T00:29:59.518535Z'
StatusHistory:
  state: PENDING
  stateStartTime: '2024-12-14T00:27:38.855477Z'
  state: SETUP_DONE
  stateStartTime: '2024-12-14T00:27:38.878002Z'
  details: Agent reported job success
  state: RUNNING
  stateStartTime: '2024-12-14T00:27:39.100195Z'
YarnApplications:
  name: Enron Email Analysis - Multiline Support
  progress: 1.0
  state: FINISHED
  trackingUrl: http://enron-cluster-m:8088/proxy/application_1734108133577_0012/
adnanawa@cloudshell:~ (river-inquiry-440619-n8) $
```

```
CLOUD SHELL
Terminal (river-inquiry-440619-n8) x +
Open Editor

statusHistory:
  state: PENDING
  stateStartTime: '2024-12-14T00:03:43.391033Z'
  state: SETUP_DONE
  stateStartTime: '2024-12-14T00:03:43.411401Z'
  details: Agent reported job success
  state: RUNNING
  stateStartTime: '2024-12-14T00:03:43.589065Z'
YarnApplications:
  name: Enron Email Analysis - Multiline Support
  progress: 1.0
  state: FINISHED
  trackingUrl: http://enron-cluster-m:8088/proxy/application_1734108133577_0011/
adnanawa@cloudshell:~ (river-inquiry-440619-n8) $ ^C
adnanawa@cloudshell:~ (river-inquiry-440619-n8) $ gsutil ls gs://river-inquiry-440619-n8/output/email_analysis_results/
gs://river-inquiry-440619-n8/output/email_analysis_results/
gs://river-inquiry-440619-n8/output/email_analysis_results/_SUCCESS
gs://river-inquiry-440619-n8/output/email_analysis_results/part-00000-7385a217-29b1-47b9-b726-6553eb6289d9-c000.csv
gs://river-inquiry-440619-n8/output/email_analysis_results/part-00001-7385a217-29b1-47b9-b726-6553eb6289d9-c000.csv
gs://river-inquiry-440619-n8/output/email_analysis_results/part-00002-7385a217-29b1-47b9-b726-6553eb6289d9-c000.csv
gs://river-inquiry-440619-n8/output/email_analysis_results/part-00003-7385a217-29b1-47b9-b726-6553eb6289d9-c000.csv
gs://river-inquiry-440619-n8/output/email_analysis_results/part-00004-7385a217-29b1-47b9-b726-6553eb6289d9-c000.csv
gs://river-inquiry-440619-n8/output/email_analysis_results/part-00005-7385a217-29b1-47b9-b726-6553eb6289d9-c000.csv
adnanawa@cloudshell:~ (river-inquiry-440619-n8) $ gcloud dataproc clusters update enron-cluster \
--region=europe-southwest1 \
--num-workers=4
Waiting on operation [projects/river-inquiry-440619-n8/regions/europe-southwest1/operations/c0a91c8-96ec-30ae-8ef9-51f2d619f1a7].
Waiting for cluster update operation...done.
Updated [https://dataproc.googleapis.com/v1/projects/river-inquiry-440619-n8/regions/europe-southwest1/clusters/enron-cluster].
adnanawa@cloudshell:~ (river-inquiry-440619-n8) $
adnanawa@cloudshell:~ (river-inquiry-440619-n8) $
```

**7.View Results:** The output files can be found in the following Google Cloud Storage path:

```
gsutil ls gs://river-inquiry-440619-n8/output/email_analysis_results/
```

```
gsutil cat gs://river-inquiry-440619-n8/output/email_analysis_results/part-00000-a71a002e-9821-4896-a6d3-8175a7259bf5-c000.csv | head -n 10
```

☰

Google Cloud

Group22

buck

X

Search

🔦

📁

📄

🔔

?

⋮

🔴A

☰

Cloud Storage

📌

GO TO PATH

REFRESH

LEARN

☰

Overview

📁

Buckets

📈

Monitoring

⚙️

Settings

🛒

Marketplace

📄

Release Notes

⏪

OBJECTS

CONFIGURATION

PERMISSIONS

PROTECTION

LIFECYCLE

OBSERVABILITY

INVENTORY REPORTS

OPERATIONS

🔗

Buckets > river-inquiry-440619-n8 > output > email\_analysis\_results

CREATE FOLDER

UPLOAD

TRANSFER DATA

OTHER SERVICES

Filter by name prefix only



















Filter

Filter objects and folders

Show

Live objects only

⋮

<input type="checkbox"/>	Name	Size	Type	Created	Storage class	Last modified	
<input type="checkbox"/>	 <a href="#">_SUCCESS</a>	0 B	application/octet-stream	Dec 14, 2024, 1:29:57 AM	Standard	Dec 14, 2024, 1:29	 ⋮
<input type="checkbox"/>	 <a href="#">part-0000-a71a002e-9821-4896...</a>	15.8 KB	application/octet-stream	Dec 14, 2024, 1:29:56 AM	Standard	Dec 14, 2024, 1:29	 ⋮
<input type="checkbox"/>	 <a href="#">part-00001-a71a002e-9821-4896...</a>	19.2 KB	application/octet-stream	Dec 14, 2024, 1:29:56 AM	Standard	Dec 14, 2024, 1:29	 ⋮
<input type="checkbox"/>	 <a href="#">part-00002-a71a002e-9821-4896...</a>	17.1 KB	application/octet-stream	Dec 14, 2024, 1:29:56 AM	Standard	Dec 14, 2024, 1:29	 ⋮
<input type="checkbox"/>	 <a href="#">part-00003-a71a002e-9821-4896...</a>	10.9 KB	application/octet-stream	Dec 14, 2024, 1:29:56 AM	Standard	Dec 14, 2024, 1:29	 ⋮
<input type="checkbox"/>	 <a href="#">part-00004-a71a002e-9821-4896...</a>	24.4 KB	application/octet-stream	Dec 14, 2024, 1:29:56 AM	Standard	Dec 14, 2024, 1:29	 ⋮
<input type="checkbox"/>	 <a href="#">part-00005-a71a002e-9821-4896...</a>	37 KB	application/octet-stream	Dec 14, 2024, 1:29:56 AM	Standard	Dec 14, 2024, 1:29	 ⋮
<input type="checkbox"/>	 <a href="#">part-00006-a71a002e-9821-4896...</a>	74.9 KB	application/octet-stream	Dec 14, 2024, 1:29:56 AM	Standard	Dec 14, 2024, 1:29	 ⋮
<input type="checkbox"/>	 <a href="#">part-00007-a71a002e-9821-4896...</a>	165.9 KB	application/octet-stream	Dec 14, 2024, 1:29:56 AM	Standard	Dec 14, 2024, 1:29	 ⋮

- Key insights were extracted, such as identifying top senders and recipients from the Enron Email Dataset.
- Performance improved as the cluster size increased from 2 to 4 worker nodes.
- Lessons learned include the importance of cloud resources, distributed processing, and resource optimization.

## 10. References

- Enron Email Dataset from Kaggle  
<https://www.kaggle.com/datasets/wcukierski/enron-email-dataset>