

# ML1020 – Assignment 2

## Nvidia Labs and Airflow Wordcount

Submitted by: Michael Vasiliou

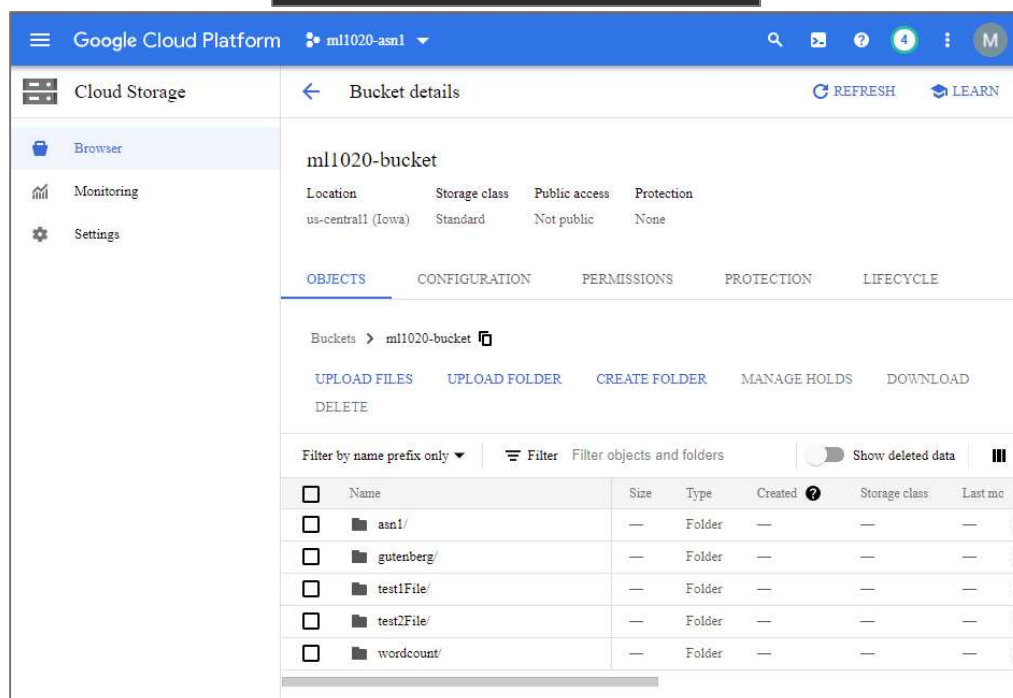
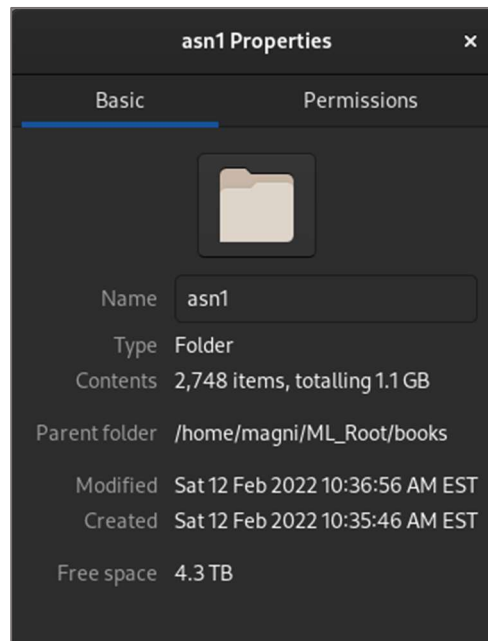
### Nvidia Labs Completion Certificate



## Deploy Wordcount using Airflow in GCP

### Dataset

For the large dataset I used python to download books from Project Gutenberg. The download script is included at the end of the PDF. The final set of books used for the wordcount consisted of 2,748 files totaling over 1GB in size.

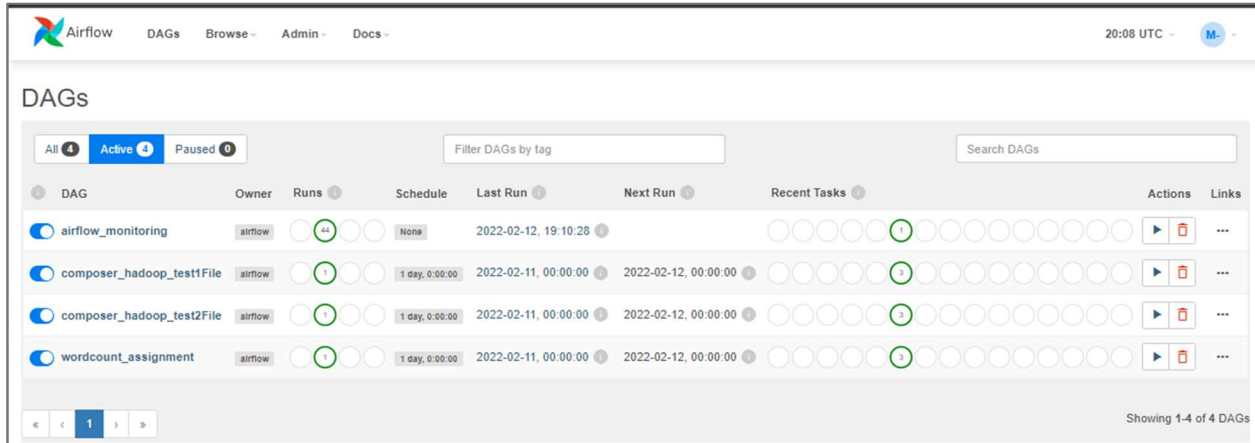





















## Airflow deployment

Airflow DAGs Browse Admin Docs 20:08 UTC M

### DAGs

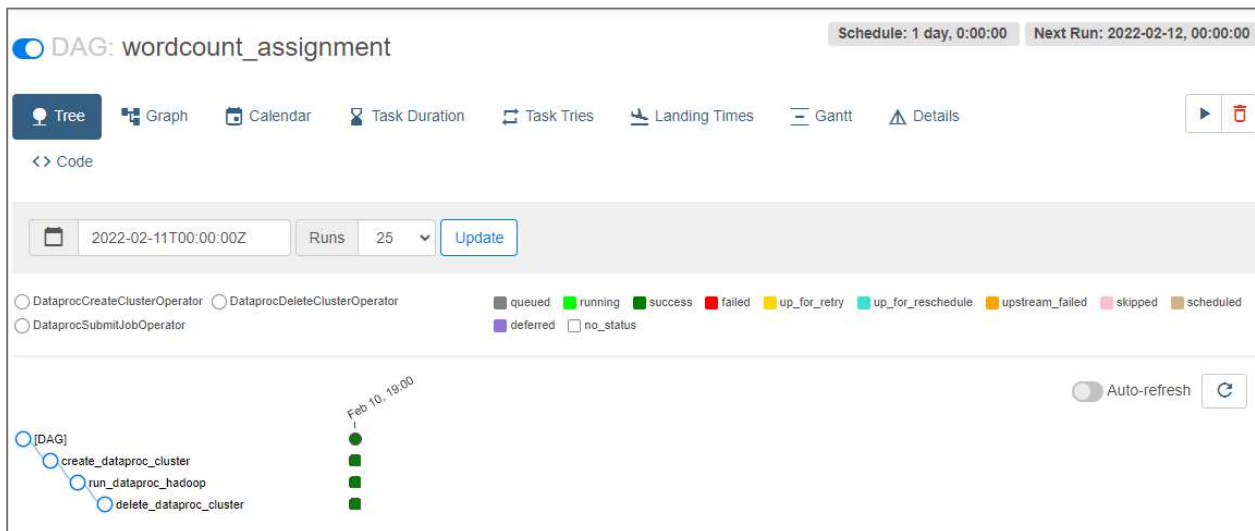







All 4 Active 4 Paused 0 Filter DAGs by tag Search DAGs

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
 airflow_monitoring	airflow		None	2022-02-12, 19:10:28			 	...
 composer_hadoop_test1File	airflow		1 day, 0:00:00	2022-02-11, 00:00:00	2022-02-12, 00:00:00		 	...
 composer_hadoop_test2File	airflow		1 day, 0:00:00	2022-02-11, 00:00:00	2022-02-12, 00:00:00		 	...
 wordcount_assignment	airflow		1 day, 0:00:00	2022-02-11, 00:00:00	2022-02-12, 00:00:00		 	...


< > 1 > Showing 1-4 of 4 DAGs

### DAG: wordcount\_assignment

Schedule: 1 day, 0:00:00 Next Run: 2022-02-12, 00:00:00


 Tree  Graph  Calendar  Task Duration  Task Tries  Landing Times  Gantt  Details


<> Code

 2022-02-11T00:00:00Z Runs 25 Update

☐ DataprocCreateClusterOperator ☐ DataprocDeleteClusterOperator ☐ DataprocSubmitJobOperator

☐ queued ☐ running ☐ success ☐ failed ☐ up\_for\_retry ☐ up\_for\_reschedule ☐ upstream\_failed ☐ skipped ☐ scheduled ☐ deferred ☐ no\_status



Auto-refresh 

Airflow DAGs Browse Admin Docs 19:02 UTC M-

**DAG: wordcount\_assignment** success Schedule: 1 day, 0:00:00 Next Run: 2022-02-12, 00:00:00

Tree Graph **Calendar** Task Duration Task Tries Landing Times Gantt Details

<> Code

2022-02-11T00:00:01Z Runs 25 Run scheduled\_\_2022-02-11T00:00:00+00:00 Find Task...

Layout Left > Right Update

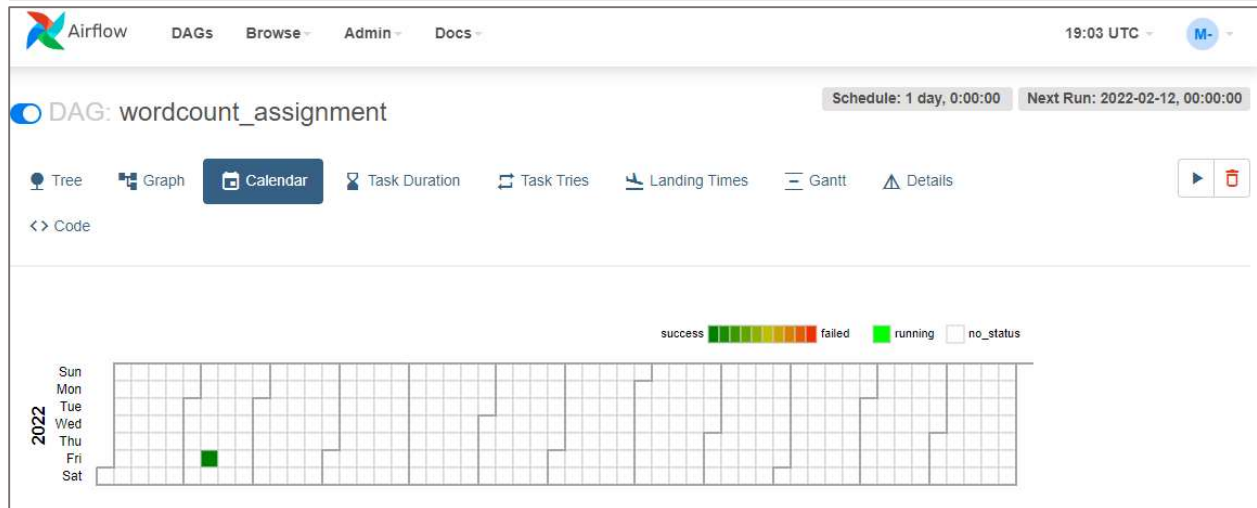
DataprocCreateClusterOperator DataprocDeleteClusterOperator  
DataprocSubmitJobOperator

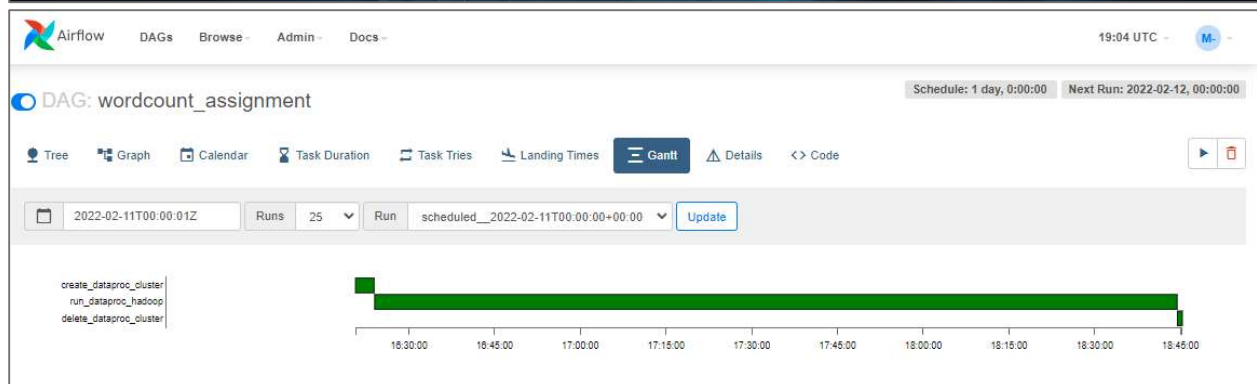
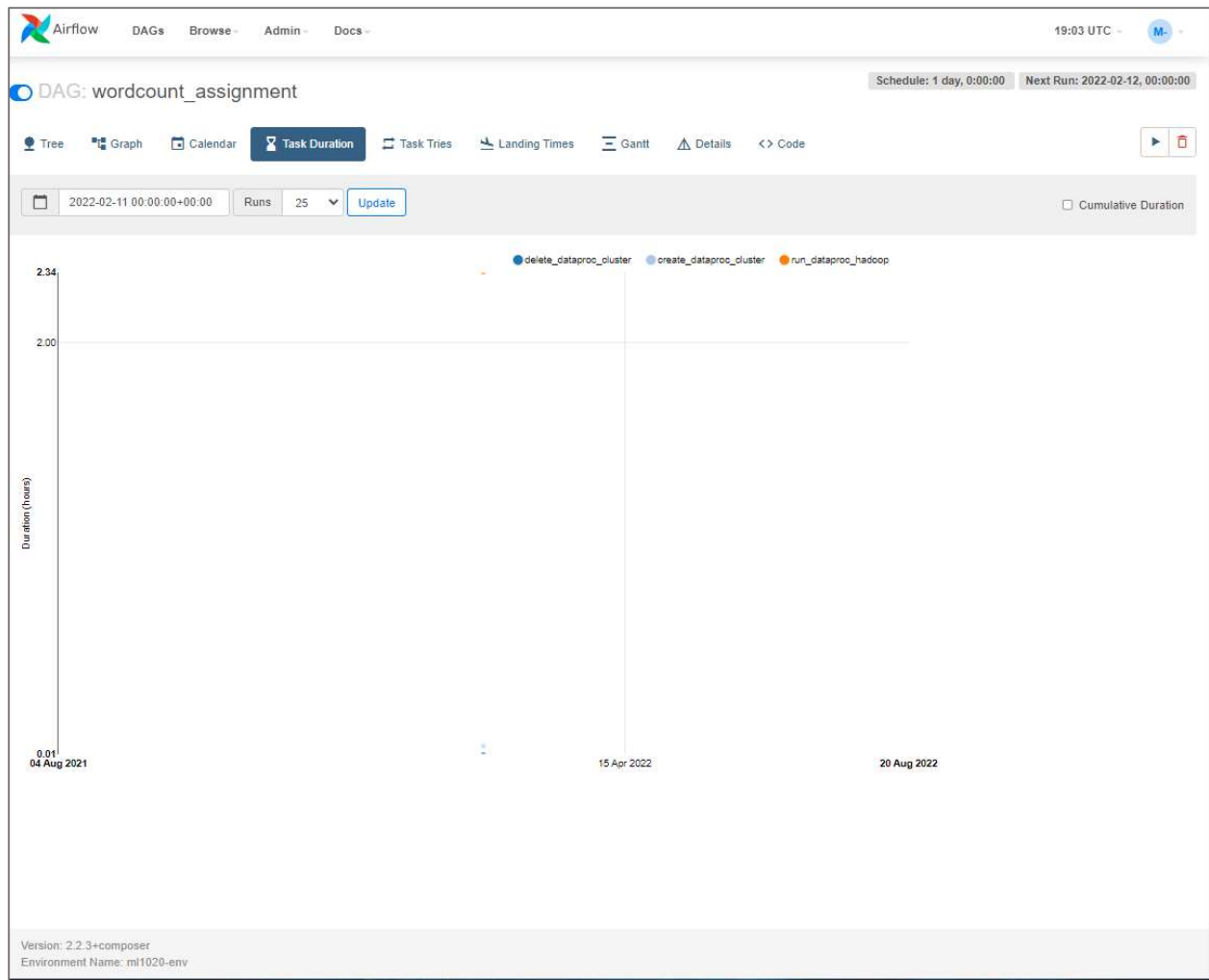
queued running success failed up\_for\_retry up\_for\_reschedule upstream\_failed skipped scheduled deferred  
no\_status

Auto-refresh

```

graph LR
    A[create_dataproc_cluster] --> B[run_dataproc_hadoop]
    B --> C[delete_dataproc_cluster]
  
```





[DAGs](#)
[Browse](#)
[Admin](#)
[Docs](#)

19:04 UTC

DAG: wordcount\_assignment

Schedule: 1 day, 0:00:00
Next Run: 2022-02-12, 00:00:00

Tree
 Graph
 Calendar
 Task Duration
 Task Tries
 Landing Times
 Gantt
 Details
 Code

DAG Details

success 3

Schedule Interval	1 day, 0:00:00
Catchup	True
Start Date	None
End Date	None
Max Active Runs	0 / 15
Concurrency	15
Default Args	{'email_on_failure': False, 'email_on_retry': False, 'location': 'us-central1', 'project_id': 'ml1020-asnt1', 'retries': 1, 'retry_delay': datetime.timedelta(seconds=300), 'start_date': DateTime(2022, 2, 11, 0, 0, 0, tzinfo=Timezone('UTC'))}
Tasks Count	3
Task IDs	['create_dataproc_cluster', 'run_dataproc_hadoop', 'delete_dataproc_cluster']
Relative file location	hadoop_assignment.py
Owner	airflow
DAG Run Timeout	None
Tags	None

## Wordcount DAG

The wordcount DAG I used was based on a tutorial from google :  
<https://cloud.google.com/composer/docs/tutorials/hadoop-wordcount-job>