Assignment 2: Data Pipeline with Apache Airflow

Course: Data Acquisition and Management
Instructor: Muhammad Shahin

PGDM PREDICTIVE ANALYTICS

GURDARSHAN SINGH
Due Date: February 16, 2025

Executive Summary

This report presents the implementation of a data pipeline using Apache Airflow to perform ETL operations on a remote dataset. The pipeline automates data extraction, transformation, and storage while ensuring modularity and efficiency. The DAG successfully processes traffic data by downloading, extracting, filtering, transforming, and combining files before generating the final output.

Introduction

The objective of this assignment is to design and implement an ETL pipeline using Apache Airflow. The pipeline follows a systematic approach to process traffic data hosted remotely, ensuring seamless automation and reproducibility.

Implementation

1. Setting Up the Airflow Environment

- Installed and configured Apache Airflow.
- Initialized the Airflow standalone server using airflow standalone.
- Created a DAG in the dags folder with appropriate configurations.

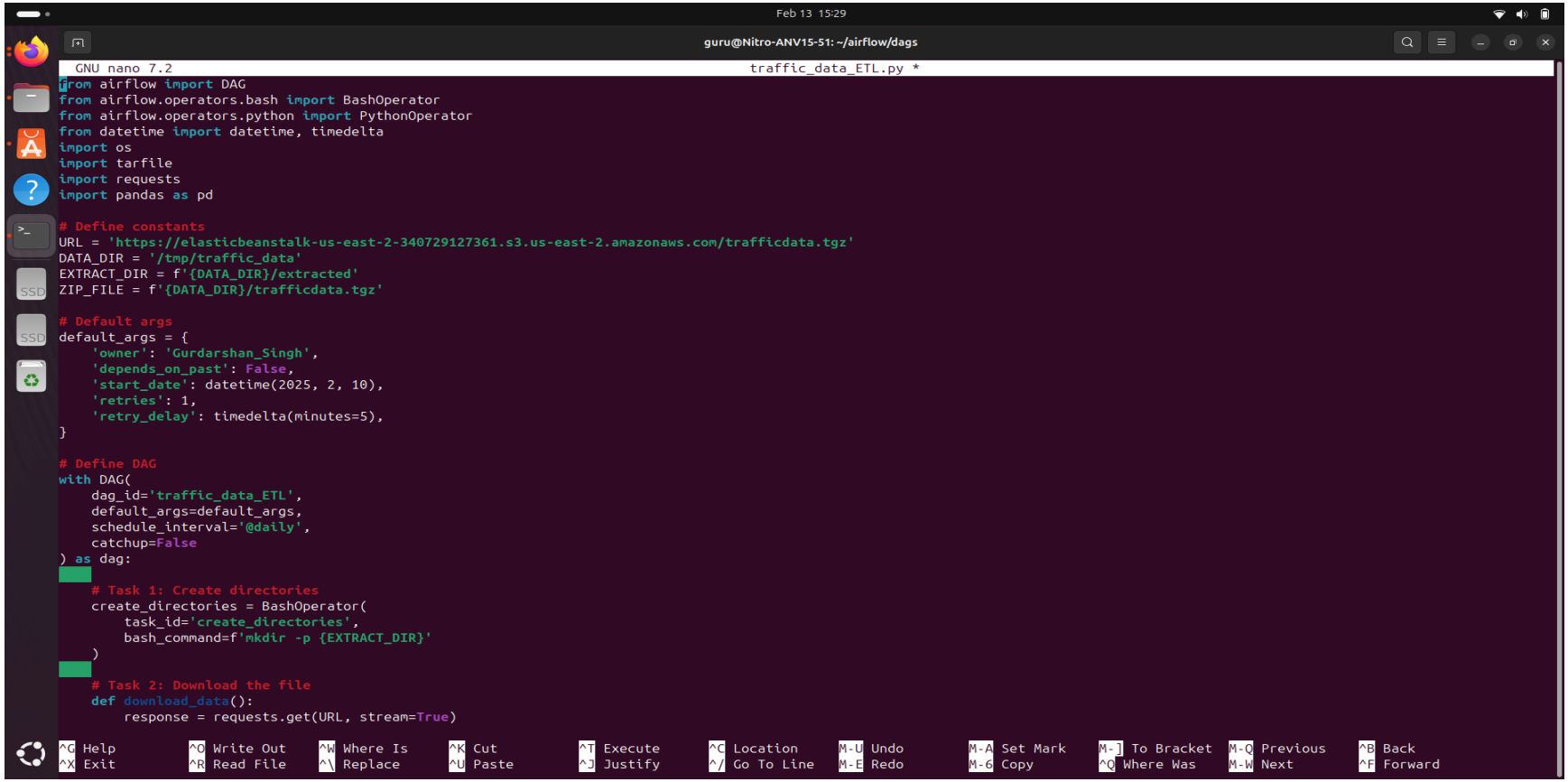
2. DAG Definition

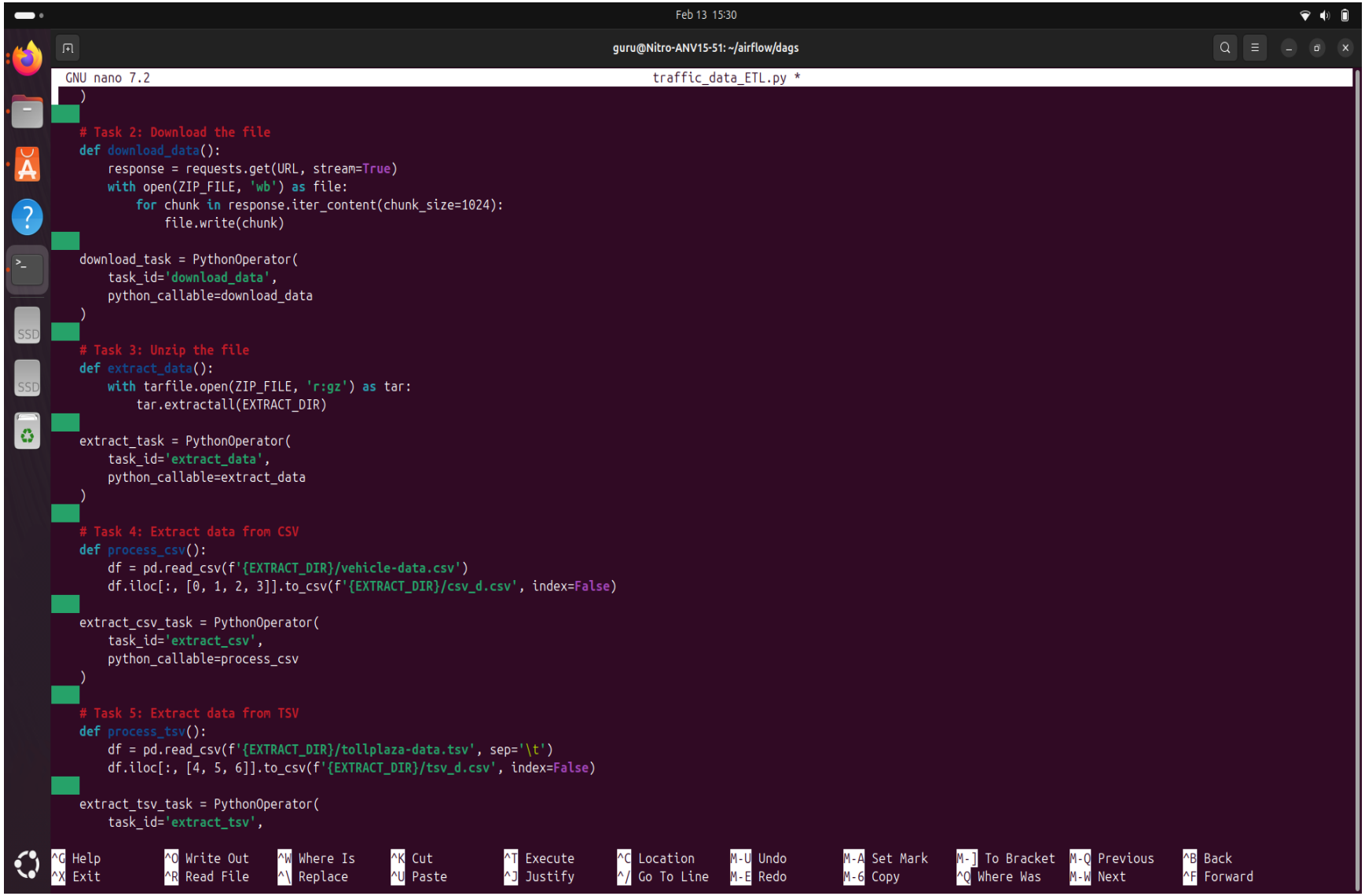
- DAG Name: traffic_data_ETL
- Owner: Gurdarshan_Singh
- Start Date: February 10, 2025
- Schedule: @daily

3. Tasks Implemented

1. **Create directories** – Ensured required directories exist for storing processed files.
2. **Download dataset** – Fetched the dataset from the remote URL.
3. **Extract files** – Decompressed the downloaded archive.
4. **Extract CSV data** – Selected specific columns from vehicle-data.csv.

5. **Extract TSV data** – Extracted relevant fields from tollplaza-data.tsv.
6. **Extract Fixed Width data** – Parsed structured fields from payment-data.txt.
7. **Combine extracted data** – Merged the extracted datasets into a single file.
8. **Transform data** – Converted vehicle type column to uppercase.
9. **Define dependencies** – Ensured optimal task execution order.




Feb 13 15:30
guru@Nitro-ANV15-51: ~/airflow/dags
GNU nano 7.2
traffic_data_ETL.py *
)
# Task 2: Download the file
def download_data():
response = requests.get(URL, stream=True)
with open(ZIP_FILE, 'wb') as file:
for chunk in response.iter_content(chunk_size=1024):
file.write(chunk)
download_task = PythonOperator(
task_id='download_data',
python_callable=download_data
)
# Task 3: Unzip the file
def extract_data():
with tarfile.open(ZIP_FILE, 'r:gz') as tar:
tar.extractall(EXTRACT_DIR)
extract_task = PythonOperator(
task_id='extract_data',
python_callable=extract_data
)
# Task 4: Extract data from CSV
def process_csv():
df = pd.read_csv(f'{EXTRACT_DIR}/vehicle-data.csv')
df.iloc[:, [0, 1, 2, 3]].to_csv(f'{EXTRACT_DIR}/csv_d.csv', index=False)
extract_csv_task = PythonOperator(
task_id='extract_csv',
python_callable=process_csv
)
# Task 5: Extract data from TSV
def process_tsv():
df = pd.read_csv(f'{EXTRACT_DIR}/tollplaza-data.tsv', sep='\t')
df.iloc[:, [4, 5, 6]].to_csv(f'{EXTRACT_DIR}/tsv_d.csv', index=False)
extract_tsv_task = PythonOperator(
task_id='extract_tsv',
^G Help ^O Write Out ^W Where Is ^K Cut ^T Execute ^C Location M-U Undo M-A Set Mark M-] To Bracket M-Q Previous ^B Back
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify ^/ Go To Line M-E Redo M-6 Copy ^Q Where Was M-W Next ^F Forward

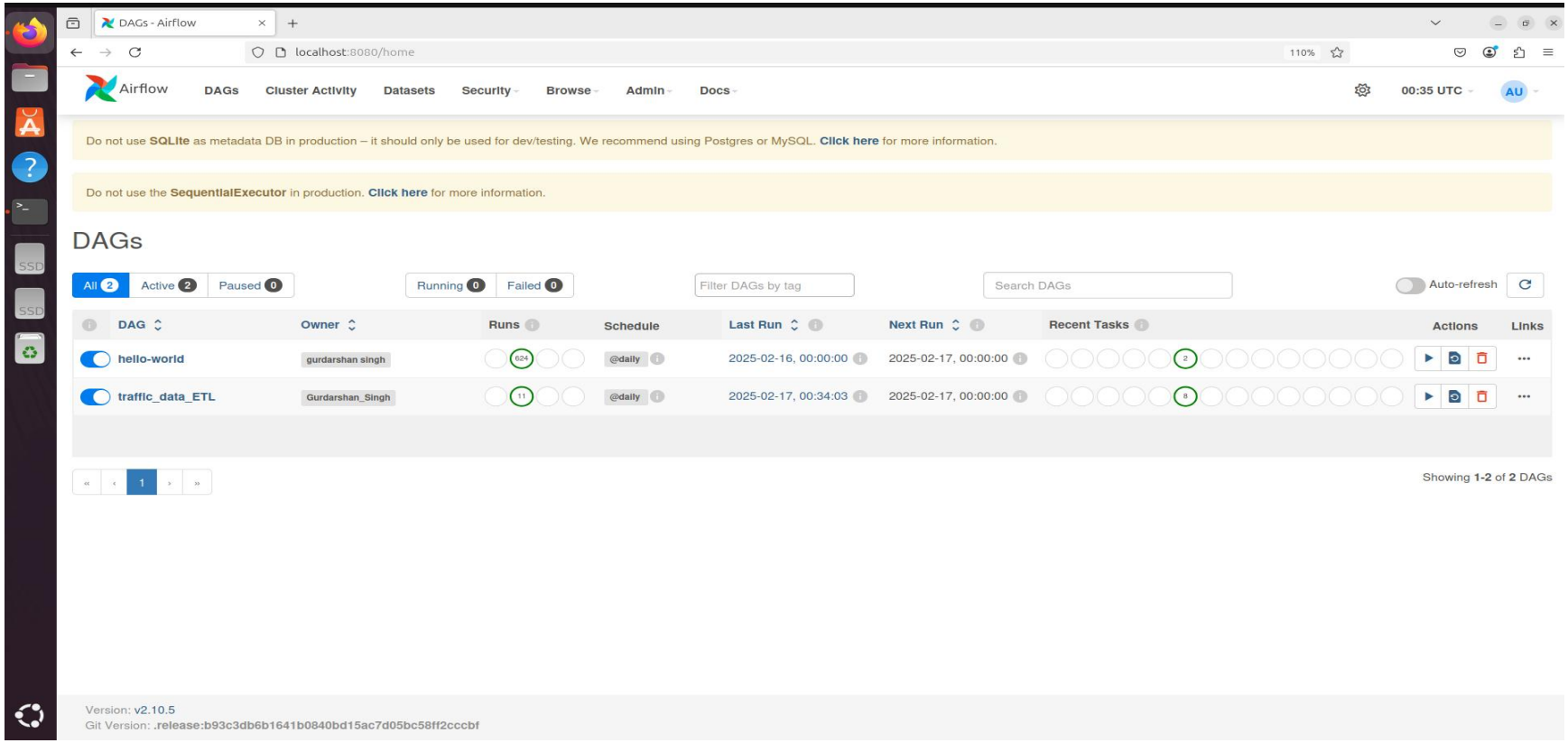

Feb 13 15:30
guru@Nitro-ANV15-51: ~/airflow/dags
GNU nano 7.2
traffic_data_ETL.py *
# Task 5: Extract data from TSV
def process_tsv():
df = pd.read_csv(f'{EXTRACT_DIR}/tollplaza-data.tsv', sep='\t')
df.iloc[:, [4, 5, 6]].to_csv(f'{EXTRACT_DIR}/tsv_d.csv', index=False)
extract_tsv_task = PythonOperator(
task_id='extract_tsv',
python_callable=process_tsv
)
# Task 6: Extract data from fixed-width file
def process_fixed_width():
colspecs = [(0, 2), (3, 5)] # Adjust based on file format
df = pd.read_fwf(f'{EXTRACT_DIR}/payment-data.txt', colspecs=colspecs)
df.to_csv(f'{EXTRACT_DIR}/fixed_width_d.csv', index=False)
extract_fixed_width_task = PythonOperator(
task_id='extract_fixed_width',
python_callable=process_fixed_width
)
# Task 7: Combine extracted files
combine_data = BashOperator(
task_id='combine_data',
bash_command=f'paste -d "," {EXTRACT_DIR}/csv_d.csv {EXTRACT_DIR}/tsv_d.csv {EXTRACT_DIR}/fixed_width_d.csv > {EXTRACT_DIR}/combined_data.csv'
)
# Task 8: Transform data
def transform_data():
df = pd.read_csv(f'{EXTRACT_DIR}/combined_data.csv')
df.iloc[:, 3] = df.iloc[:, 3].str.upper()
df.to_csv(f'{EXTRACT_DIR}/transformed_data.csv', index=False)
transform_task = PythonOperator(
task_id='transform_data',
python_callable=transform_data
)
Set task dependencies
create_directories >> download_task >> extract_task
extract_task >> [extract_csv_task, extract_tsv_task, extract_fixed_width_task] >> combine_data >> transform_task
^G Help ^O Write Out ^W Where Is ^K Cut ^T Execute ^C Location M-U Undo M-A Set Mark M-] To Bracket M-Q Previous ^B Back
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify ^/ Go To Line M-E Redo M-6 Copy ^Q Where Was M-W Next ^F Forward

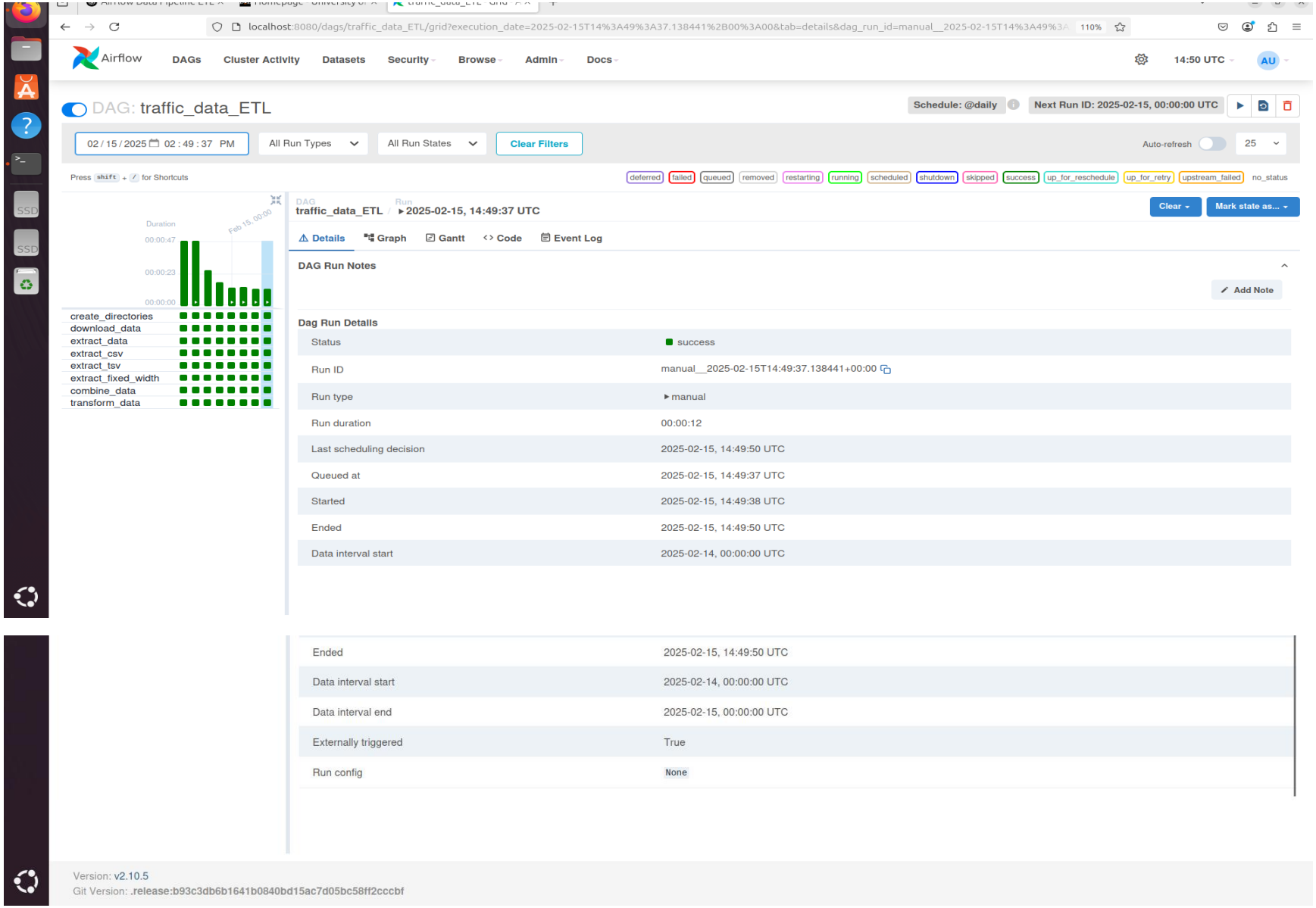
4. Airflow DAG Execution

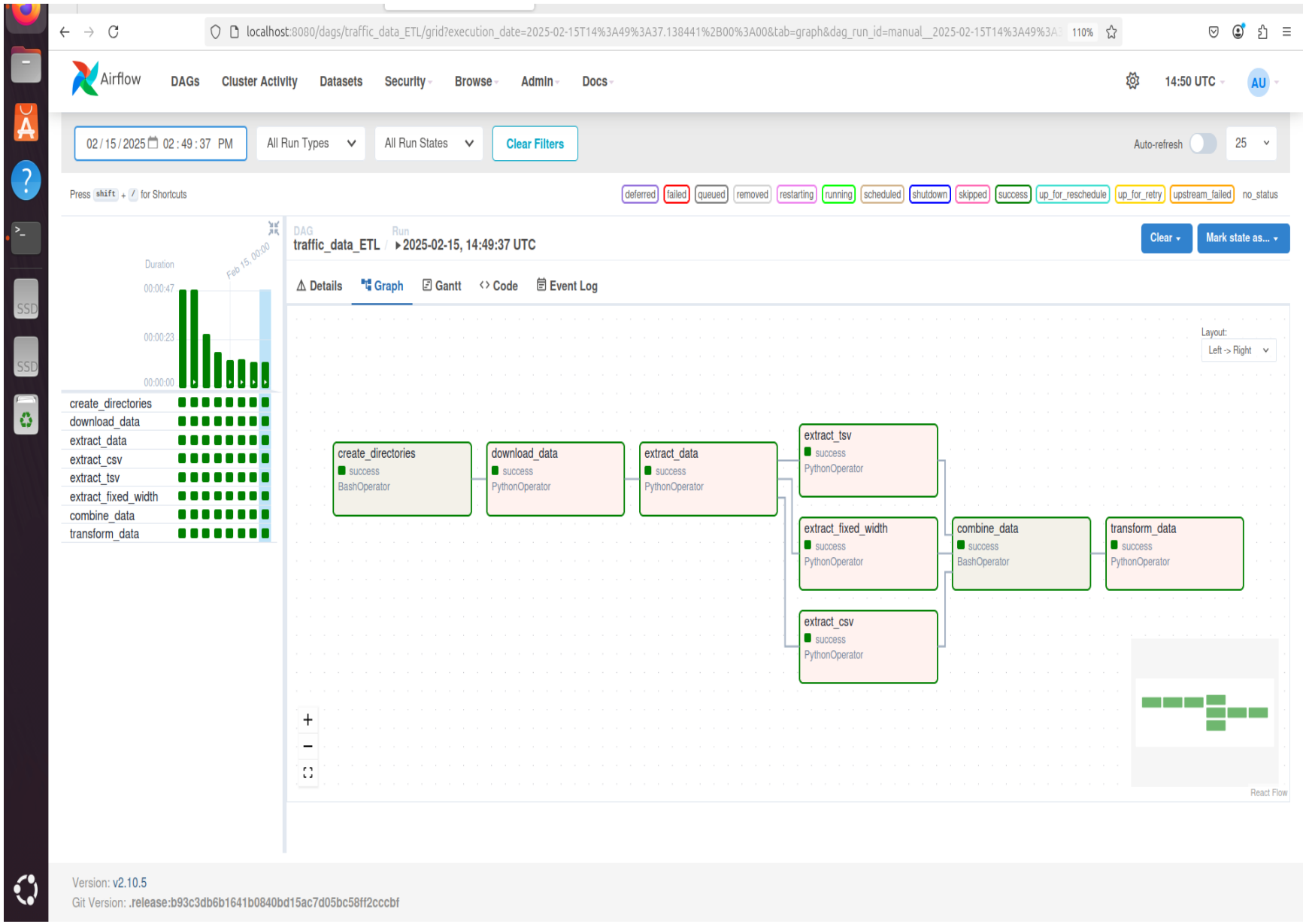
- DAG successfully appeared in the Airflow UI at http://localhost:8080
- DAG was activated and manually triggered
- Execution logs confirmed successful task completion

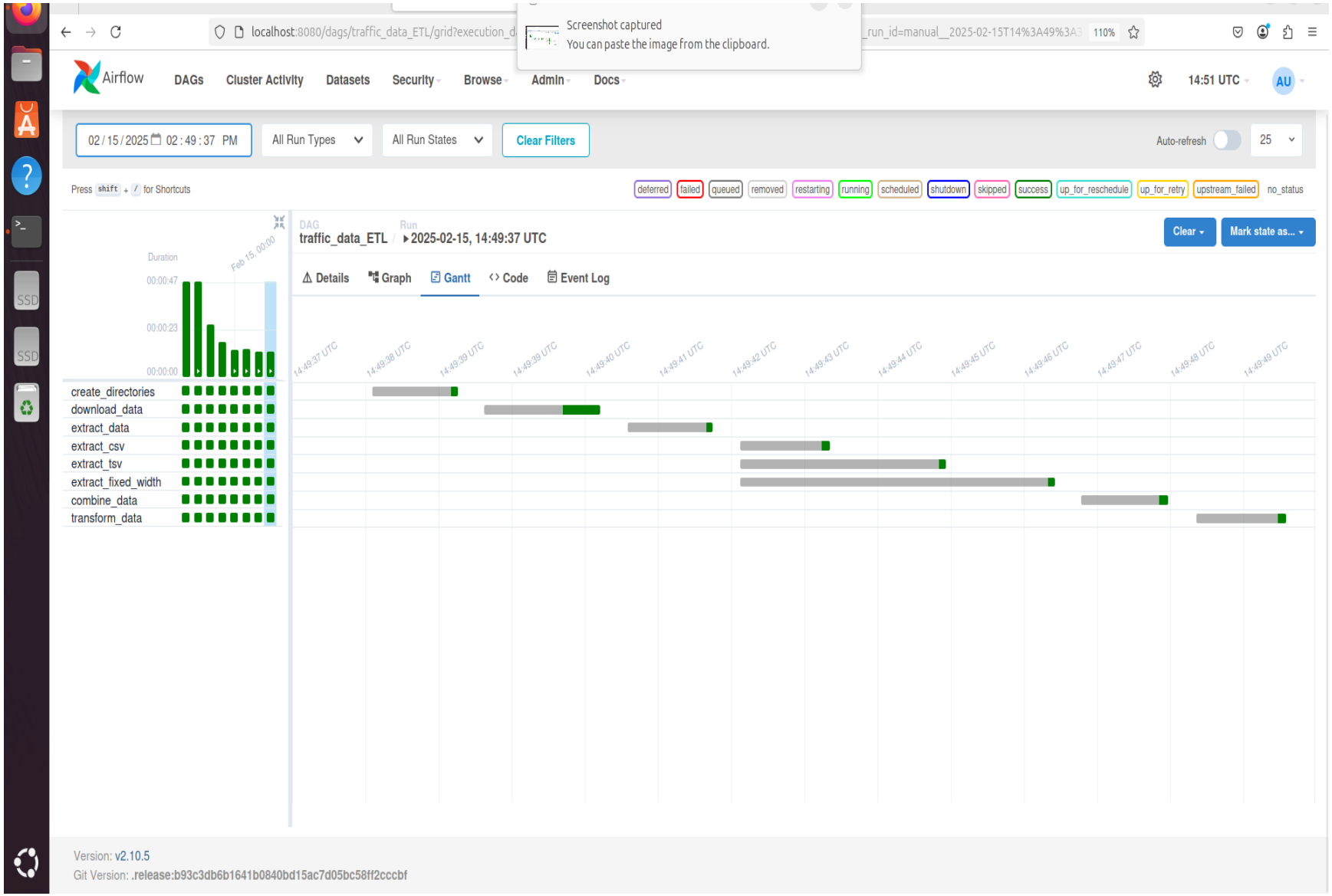
Screenshots

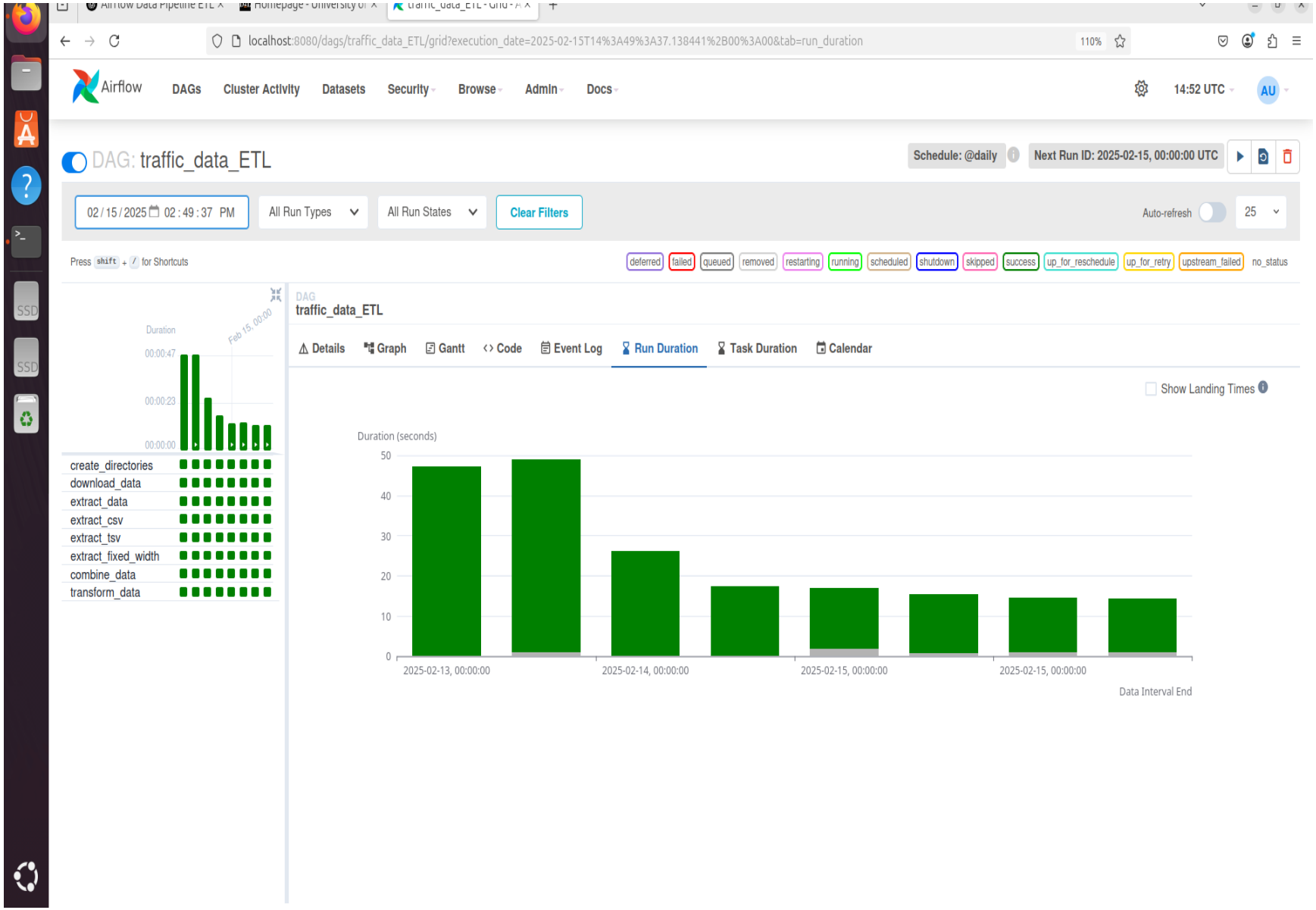
(Include screenshots of DAG Graph View, Task Logs, Gantt Chart, and Data Output)

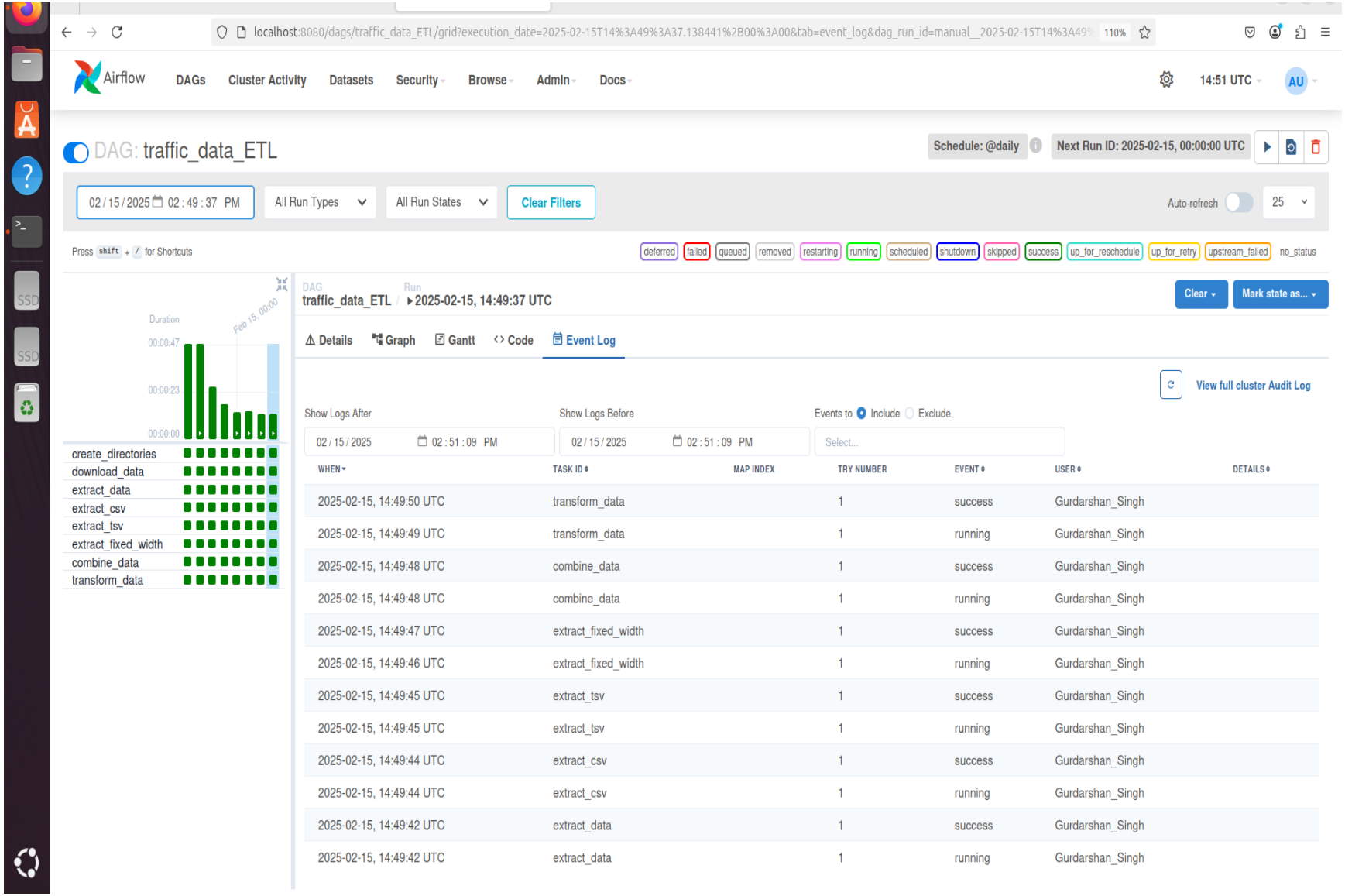


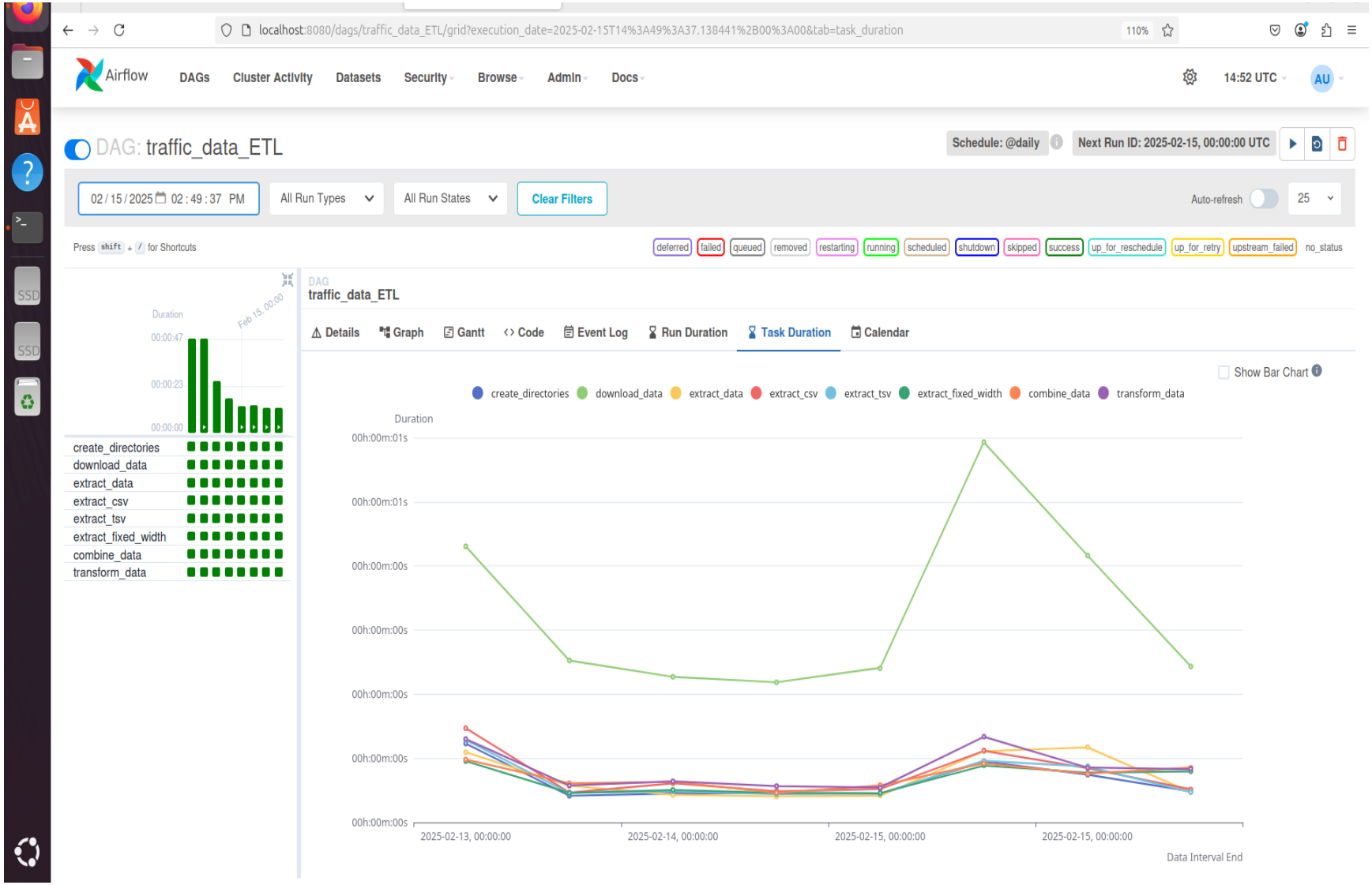

localhost:8080/dags/traffic_data_ETL/grid?execution_date=2025-02-15T14%3A49%3A37.138441%2B00%3A00&tab=details&dag_run_id=manual__2025-02-15T14%3A49%3A
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
14:50 UTC
AU
DAG: traffic_data_ETL
Schedule: @daily
Next Run ID: 2025-02-15, 00:00:00 UTC
02 / 15 / 2025 02 : 49 : 37 PM
All Run Types
All Run States
Clear Filters
Auto-refresh
25
Press shift + / for Shortcuts
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Duration
00:00:47
00:00:23
00:00:00
Feb 15, 00:00
create_directories
download_data
extract_data
extract_csv
extract_tsv
extract_fixed_width
combine_data
transform_data
DAG
traffic_data_ETL
Run
▸2025-02-15, 14:49:37 UTC
Clear
Mark state as...
Details
Graph
Gantt
Code
Event Log
DAG Run Notes
Add Note
Dag Run Details
Status
success
Run ID
manual__2025-02-15T14:49:37.138441+00:00
Run type
▸manual
Run duration
00:00:12
Last scheduling decision
2025-02-15, 14:49:50 UTC
Queued at
2025-02-15, 14:49:37 UTC
Started
2025-02-15, 14:49:38 UTC
Ended
2025-02-15, 14:49:50 UTC
Data interval start
2025-02-14, 00:00:00 UTC
Ended
2025-02-15, 14:49:50 UTC
Data interval start
2025-02-14, 00:00:00 UTC
Data interval end
2025-02-15, 00:00:00 UTC
Externally triggered
True
Run config
None
Version: v2.10.5
Git Version: .release:b93c3db6b1641b0840bd15ac7d05bc58ff2cccbf

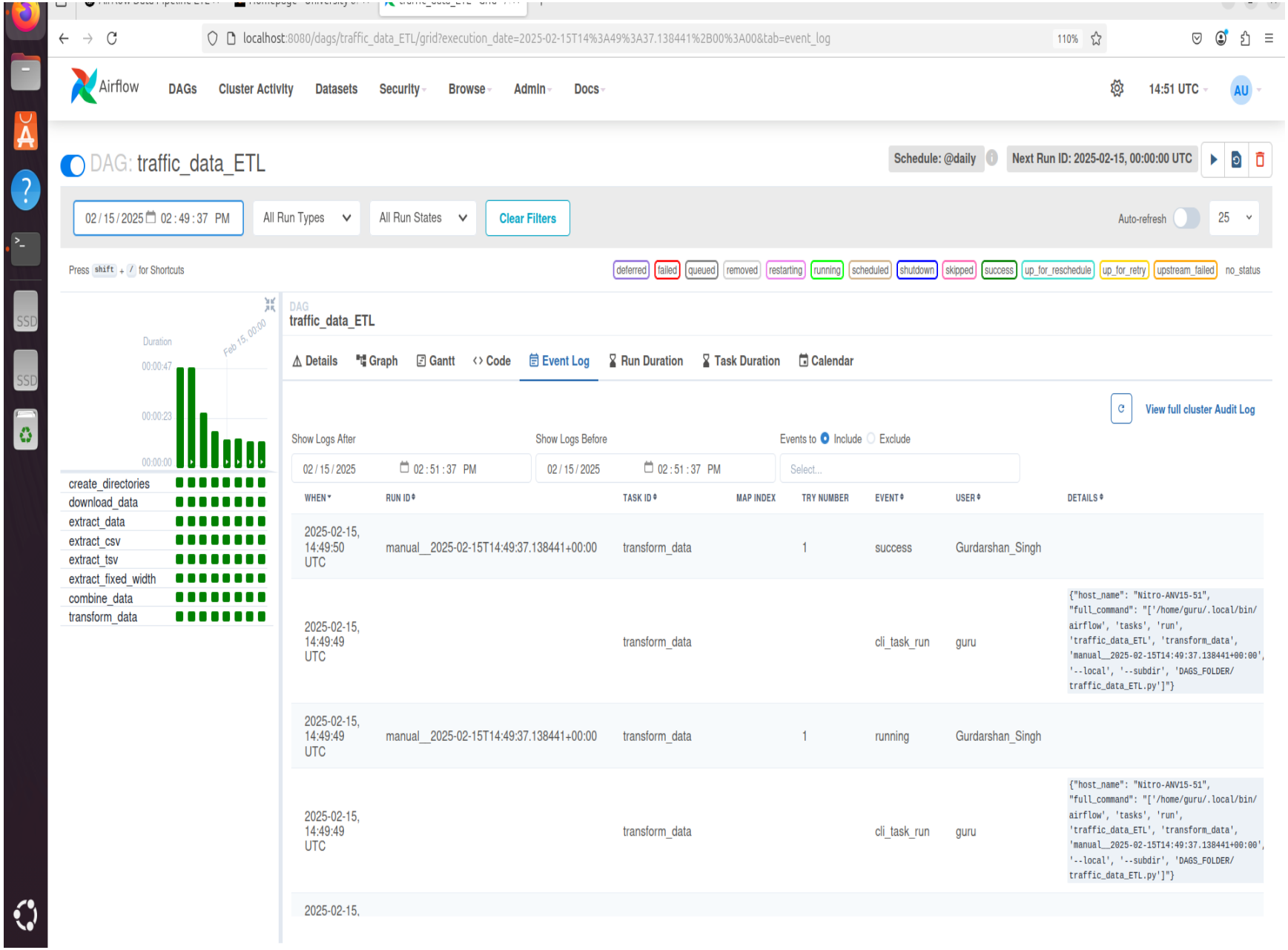

localhost:8080/dags/traffic_data_ETL/grid?execution_date=2025-02-15T14%3A49%3A37.138441%2B00%3A00&tab=graph&dag_run_id=manual__2025-02-15T14%3A49%3A
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
14:50 UTC
AU
02 / 15 / 2025 02 : 49 : 37 PM
All Run Types
All Run States
Clear Filters
Auto-refresh
25
Press shift + / for Shortcuts
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Duration
00:00:47
00:00:23
00:00:00
Feb 15, 00:00
create_directories
download_data
extract_data
extract_csv
extract_tsv
extract_fixed_width
combine_data
transform_data
DAG
traffic_data_ETL
Run
2025-02-15, 14:49:37 UTC
Clear
Mark state as...
Details
Graph
Gantt
Code
Event Log
Layout:
Left -> Right
create_directories
success
BashOperator
download_data
success
PythonOperator
extract_data
success
PythonOperator
extract_tsv
success
PythonOperator
extract_fixed_width
success
PythonOperator
extract_csv
success
PythonOperator
combine_data
success
BashOperator
transform_data
success
PythonOperator
React Flow
Version: v2.10.5
Git Version: .release:b93c3db6b1641b0840bd15ac7d05bc58ff2cccbf

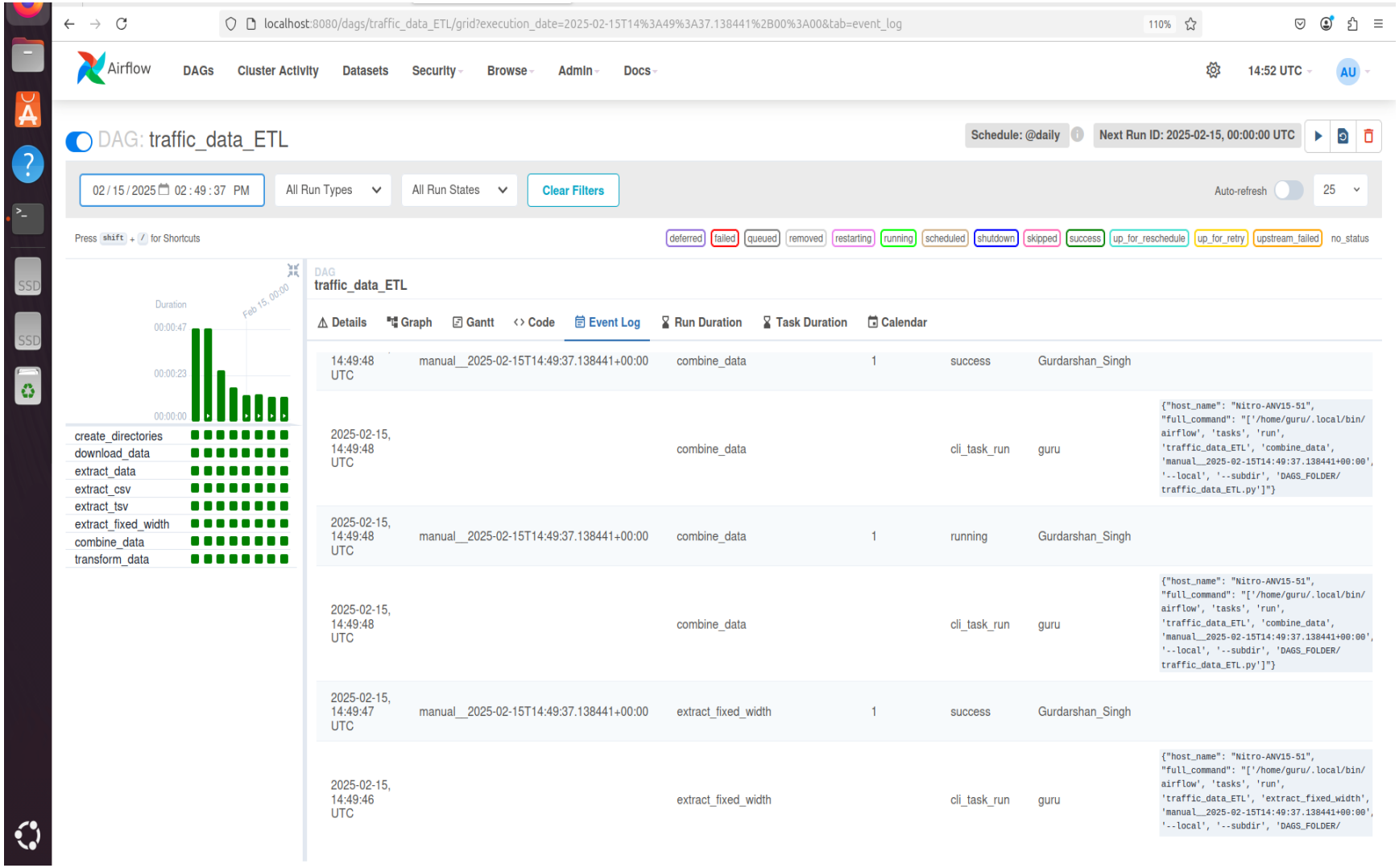

Screenshot captured
You can paste the image from the clipboard.
localhost:8080/dags/traffic_data_ETL/grid?execution_d
run_id=manual__2025-02-15T14%3A49%3A3
110%
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
14:51 UTC
AU
02 / 15 / 2025 02 : 49 : 37 PM
All Run Types
All Run States
Clear Filters
Auto-refresh
25
Press shift + / for Shortcuts
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Duration
00:00:47
00:00:23
00:00:00
Feb 15, 00:00
DAG
traffic_data_ETL
Run
2025-02-15, 14:49:37 UTC
Clear
Mark state as...
Details
Graph
Gantt
Code
Event Log
14:49:37 UTC
14:49:38 UTC
14:49:39 UTC
14:49:39 UTC
14:49:40 UTC
14:49:41 UTC
14:49:42 UTC
14:49:43 UTC
14:49:44 UTC
14:49:45 UTC
14:49:46 UTC
14:49:47 UTC
14:49:48 UTC
14:49:49 UTC
create_directories
download_data
extract_data
extract_csv
extract_tsv
extract_fixed_width
combine_data
transform_data
Version: v2.10.5
Git Version: .release:b93c3db6b1641b0840bd15ac7d05bc58ff2cccbf

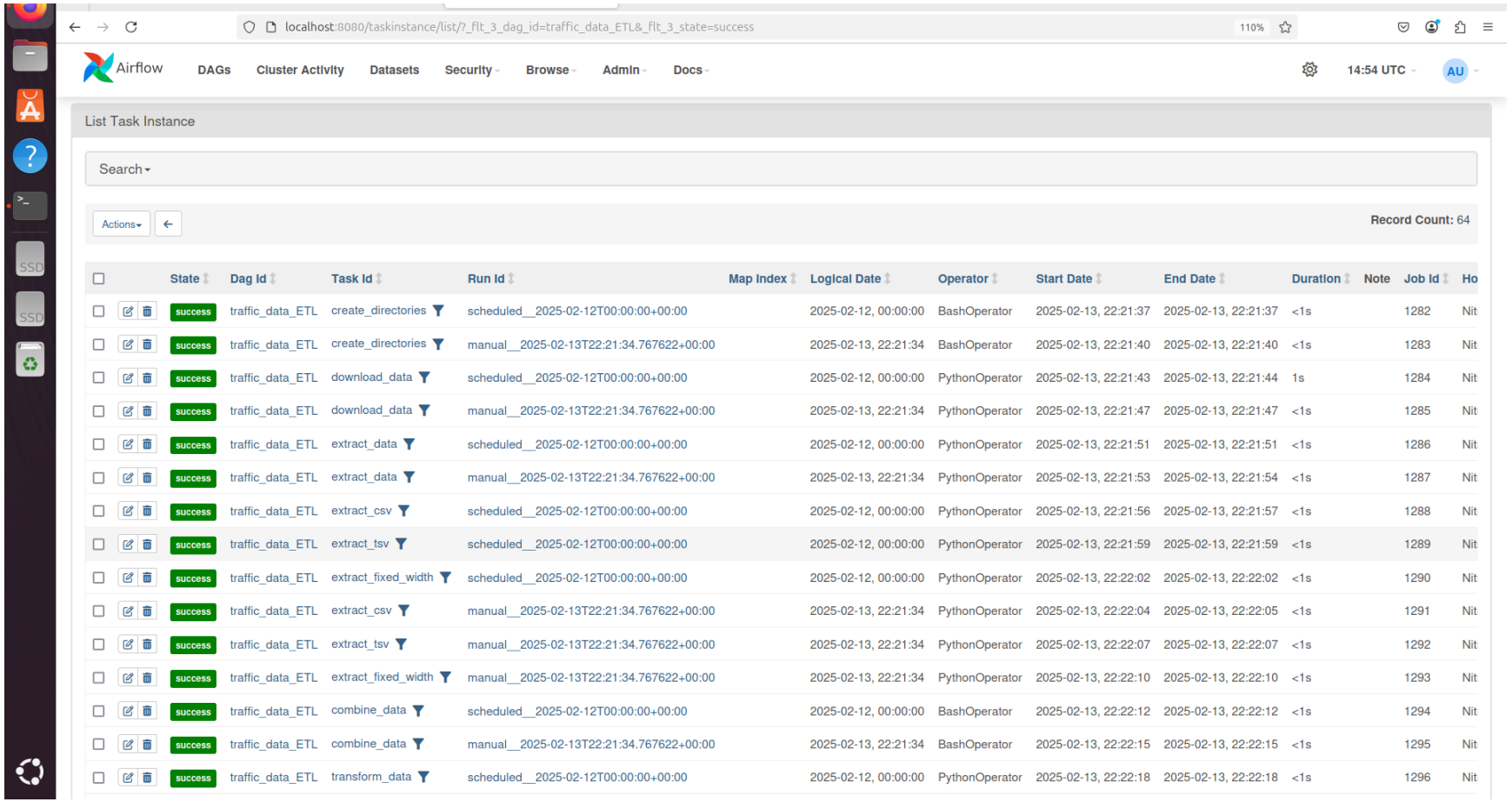

localhost:8080/dags/traffic_data_ETL/grid?execution_date=2025-02-15T14%3A49%3A37.138441%2B00%3A00&tab=run_duration
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
14:52 UTC
AU
DAG: traffic_data_ETL
Schedule: @daily
Next Run ID: 2025-02-15, 00:00:00 UTC
02 / 15 / 2025 02 : 49 : 37 PM
All Run Types
All Run States
Clear Filters
Auto-refresh
25
Press shift + / for Shortcuts
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Duration
Feb 15, 00:00
00:00:47
00:00:23
00:00:00
create_directories
download_data
extract_data
extract_csv
extract_tsv
extract_fixed_width
combine_data
transform_data
DAG
traffic_data_ETL
Details
Graph
Gantt
Code
Event Log
Run Duration
Task Duration
Calendar
Show Landing Times
Duration (seconds)
50
40
30
20
10
0
2025-02-13, 00:00:00
2025-02-14, 00:00:00
2025-02-15, 00:00:00
2025-02-15, 00:00:00
Data Interval End


localhost:8080/dags/traffic_data_ETL/grid?execution_date=2025-02-15T14%3A49%3A37.138441%2B00%3A00&tab=event_log&dag_run_id=manual__2025-02-15T14%3A49
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
14:51 UTC
AU
DAG: traffic_data_ETL
Schedule: @daily
Next Run ID: 2025-02-15, 00:00:00 UTC
02 / 15 / 2025 02 : 49 : 37 PM
All Run Types
All Run States
Clear Filters
Auto-refresh
25
Press shift + / for Shortcuts
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Duration
00:00:47
00:00:23
00:00:00
Feb 15, 00:00
create_directories
download_data
extract_data
extract_csv
extract_tsv
extract_fixed_width
combine_data
transform_data
DAG
traffic_data_ETL
Run
2025-02-15, 14:49:37 UTC
Clear
Mark state as...
Details
Graph
Gantt
Code
Event Log
View full cluster Audit Log
Show Logs After
02 / 15 / 2025 02 : 51 : 09 PM
Show Logs Before
02 / 15 / 2025 02 : 51 : 09 PM
Events to Include Exclude
Select...
WHEN
TASK ID
MAP INDEX
TRY NUMBER
EVENT
USER
DETAILS
2025-02-15, 14:49:50 UTC transform_data 1 success Gurdarshan_Singh
2025-02-15, 14:49:49 UTC transform_data 1 running Gurdarshan_Singh
2025-02-15, 14:49:48 UTC combine_data 1 success Gurdarshan_Singh
2025-02-15, 14:49:48 UTC combine_data 1 running Gurdarshan_Singh
2025-02-15, 14:49:47 UTC extract_fixed_width 1 success Gurdarshan_Singh
2025-02-15, 14:49:46 UTC extract_fixed_width 1 running Gurdarshan_Singh
2025-02-15, 14:49:45 UTC extract_tsv 1 success Gurdarshan_Singh
2025-02-15, 14:49:45 UTC extract_tsv 1 running Gurdarshan_Singh
2025-02-15, 14:49:44 UTC extract_csv 1 success Gurdarshan_Singh
2025-02-15, 14:49:44 UTC extract_csv 1 running Gurdarshan_Singh
2025-02-15, 14:49:42 UTC extract_data 1 success Gurdarshan_Singh
2025-02-15, 14:49:42 UTC extract_data 1 running Gurdarshan_Singh

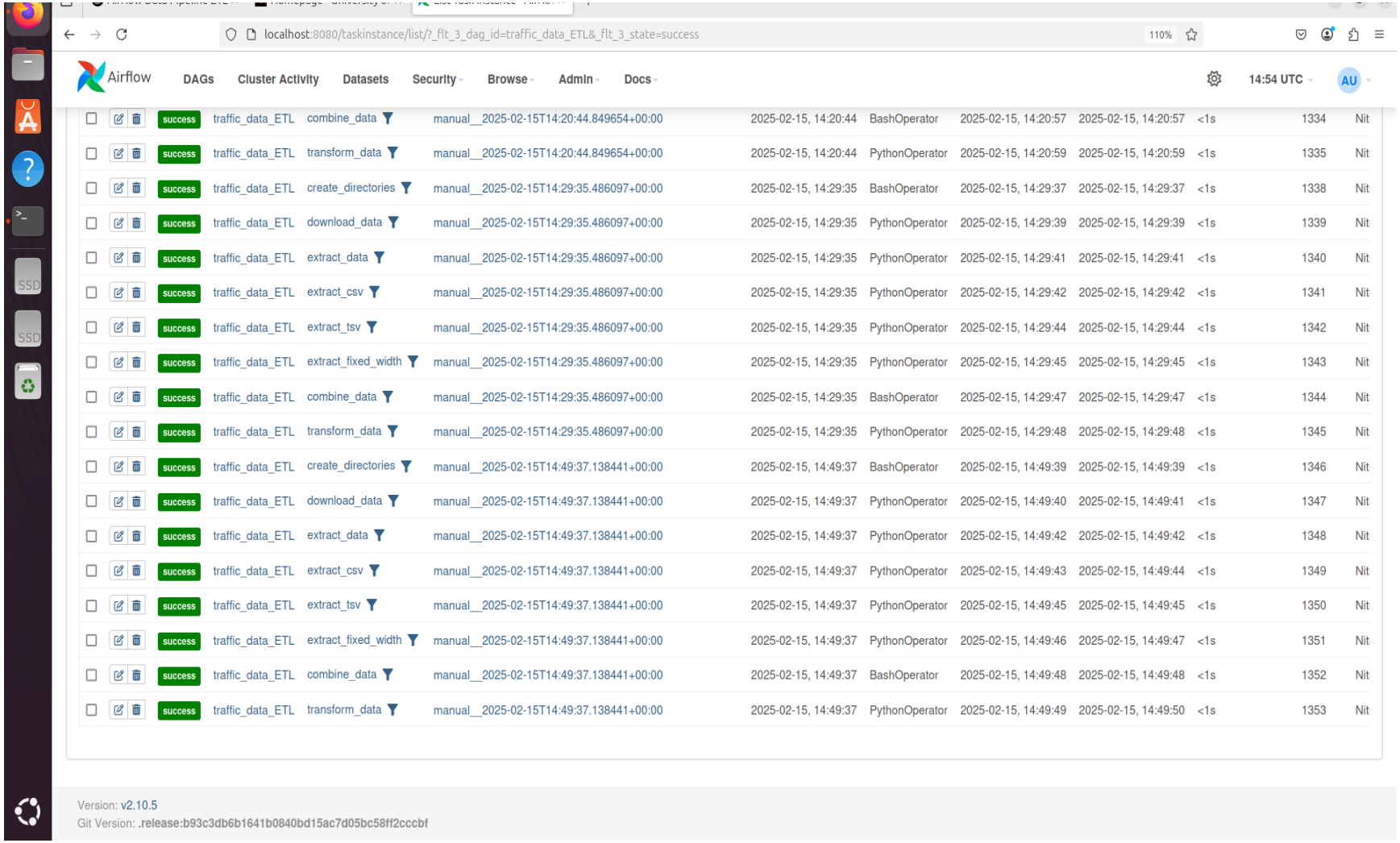

localhost:8080/dags/traffic_data_ETL/grid?execution_date=2025-02-15T14%3A49%3A37.138441%2B00%3A00&tab=task_duration
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
14:52 UTC
AU
DAG: traffic_data_ETL
Schedule: @daily
Next Run ID: 2025-02-15, 00:00:00 UTC
02 / 15 / 2025 02 : 49 : 37 PM
All Run Types
All Run States
Clear Filters
Auto-refresh
25
Press shift + / for Shortcuts
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Duration
Feb 15, 00:00
00:00:47
00:00:23
00:00:00
create_directories
download_data
extract_data
extract_csv
extract_tsv
extract_fixed_width
combine_data
transform_data
DAG
traffic_data_ETL
Details
Graph
Gantt
Code
Event Log
Run Duration
Task Duration
Calendar
Show Bar Chart
create_directories
download_data
extract_data
extract_csv
extract_tsv
extract_fixed_width
combine_data
transform_data
Duration
00h:00m:01s
00h:00m:01s
00h:00m:00s
00h:00m:00s
00h:00m:00s
00h:00m:00s
2025-02-13, 00:00:00
2025-02-14, 00:00:00
2025-02-15, 00:00:00
2025-02-15, 00:00:00
Data Interval End

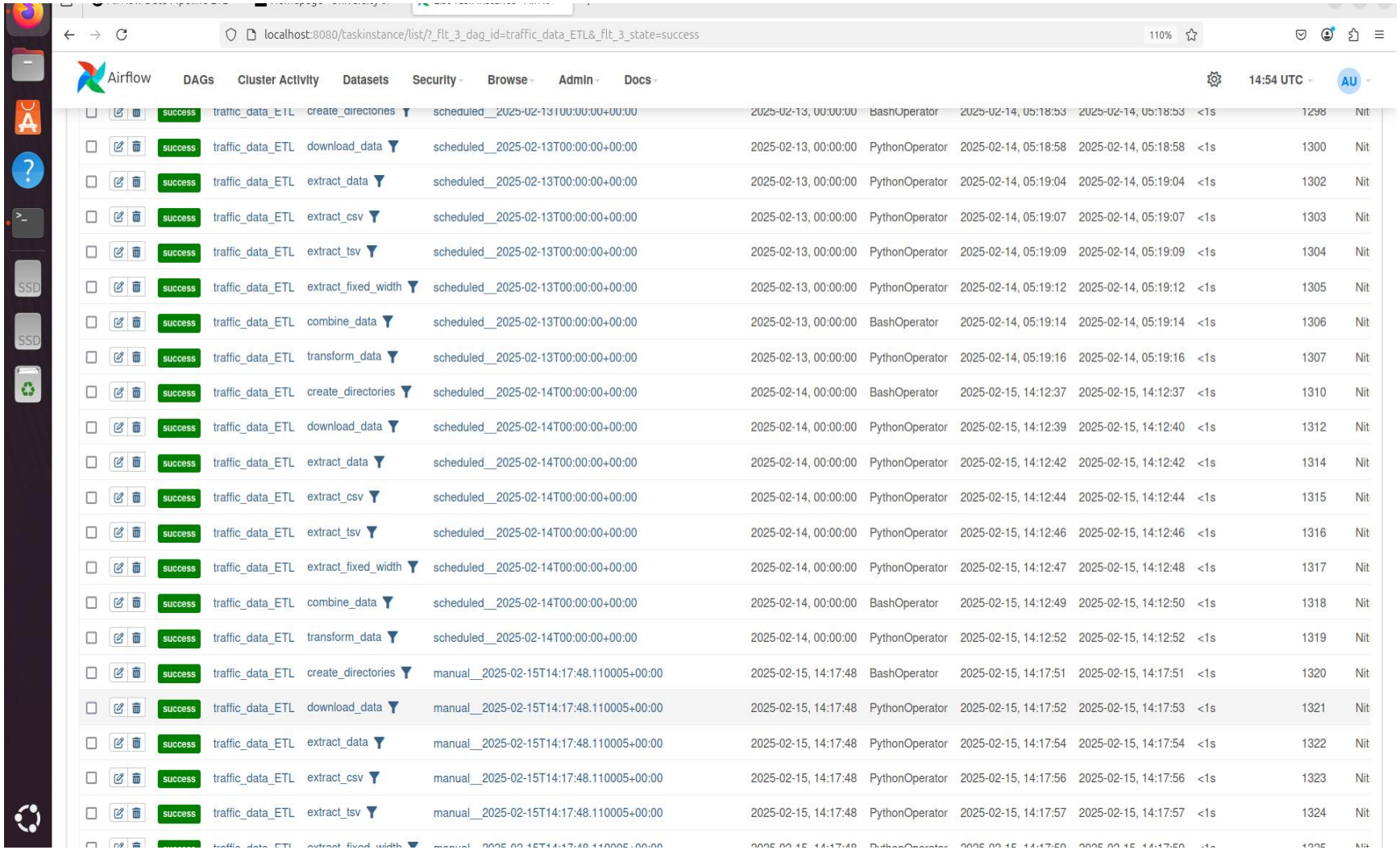

localhost:8080/dags/traffic_data_ETL/grid?execution_date=2025-02-15T14%3A49%3A37.138441%2B00%3A00&tab=event_log
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
14:51 UTC
AU
DAG: traffic_data_ETL
Schedule: @daily
Next Run ID: 2025-02-15, 00:00:00 UTC
02 / 15 / 2025 02 : 49 : 37 PM
All Run Types
All Run States
Clear Filters
Auto-refresh
25
Press shift + / for Shortcuts
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Duration
00:00:47
00:00:23
00:00:00
Feb 15, 00:00
create_directories
download_data
extract_data
extract_csv
extract_tsv
extract_fixed_width
combine_data
transform_data
DAG
traffic_data_ETL
Details
Graph
Gantt
Code
Event Log
Run Duration
Task Duration
Calendar
View full cluster Audit Log
Show Logs After
02 / 15 / 2025 02 : 51 : 37 PM
Show Logs Before
02 / 15 / 2025 02 : 51 : 37 PM
Events to Include Exclude
Select...
WHEN
RUN ID
TASK ID
MAP INDEX
TRY NUMBER
EVENT
USER
DETAILS
2025-02-15, 14:49:50 UTC
manual__2025-02-15T14:49:37.138441+00:00
transform_data
1
success
Gurdarshan_Singh
2025-02-15, 14:49:49 UTC
transform_data
cli_task_run
guru
{"host_name": "Nitro-ANV15-51", "full_command": "['/home/guru/.local/bin/airflow', 'tasks', 'run', 'traffic_data_ETL', 'transform_data', 'manual__2025-02-15T14:49:37.138441+00:00', '--local', '--subdir', 'DAGS_FOLDER/traffic_data_ETL.py']"}
2025-02-15, 14:49:49 UTC
manual__2025-02-15T14:49:37.138441+00:00
transform_data
1
running
Gurdarshan_Singh
2025-02-15, 14:49:49 UTC
transform_data
cli_task_run
guru
{"host_name": "Nitro-ANV15-51", "full_command": "['/home/guru/.local/bin/airflow', 'tasks', 'run', 'traffic_data_ETL', 'transform_data', 'manual__2025-02-15T14:49:37.138441+00:00', '--local', '--subdir', 'DAGS_FOLDER/traffic_data_ETL.py']"}
2025-02-15,


localhost:8080/dags/traffic_data_ETL/grid?execution_date=2025-02-15T14%3A49%3A37.138441%2B00%3A00&tab=event_log
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
14:52 UTC
AU
DAG: traffic_data_ETL
Schedule: @daily
Next Run ID: 2025-02-15, 00:00:00 UTC
02 / 15 / 2025 02 : 49 : 37 PM
All Run Types
All Run States
Clear Filters
Auto-refresh
25
Press shift + / for Shortcuts
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Duration
00:00:47
00:00:23
00:00:00
create_directories
download_data
extract_data
extract_csv
extract_tsv
extract_fixed_width
combine_data
transform_data
DAG
traffic_data_ETL
Details
Graph
Gantt
Code
Event Log
Run Duration
Task Duration
Calendar
14:49:48 UTC
manual__2025-02-15T14:49:37.138441+00:00
combine_data
1
success
Gurdarshan_Singh
2025-02-15, 14:49:48 UTC
combine_data
cli_task_run
guru
{"host_name": "Nitro-ANV15-51", "full_command": "['/home/guru/.local/bin/airflow', 'tasks', 'run', 'traffic_data_ETL', 'combine_data', 'manual__2025-02-15T14:49:37.138441+00:00', '--local', '--subdir', 'DAGS_FOLDER/traffic_data_ETL.py']"}
2025-02-15, 14:49:48 UTC
manual__2025-02-15T14:49:37.138441+00:00
combine_data
1
running
Gurdarshan_Singh
2025-02-15, 14:49:48 UTC
combine_data
cli_task_run
guru
{"host_name": "Nitro-ANV15-51", "full_command": "['/home/guru/.local/bin/airflow', 'tasks', 'run', 'traffic_data_ETL', 'combine_data', 'manual__2025-02-15T14:49:37.138441+00:00', '--local', '--subdir', 'DAGS_FOLDER/traffic_data_ETL.py']"}
2025-02-15, 14:49:47 UTC
manual__2025-02-15T14:49:37.138441+00:00
extract_fixed_width
1
success
Gurdarshan_Singh
2025-02-15, 14:49:46 UTC
extract_fixed_width
cli_task_run
guru
{"host_name": "Nitro-ANV15-51", "full_command": "['/home/guru/.local/bin/airflow', 'tasks', 'run', 'traffic_data_ETL', 'extract_fixed_width', 'manual__2025-02-15T14:49:37.138441+00:00', '--local', '--subdir', 'DAGS_FOLDER/

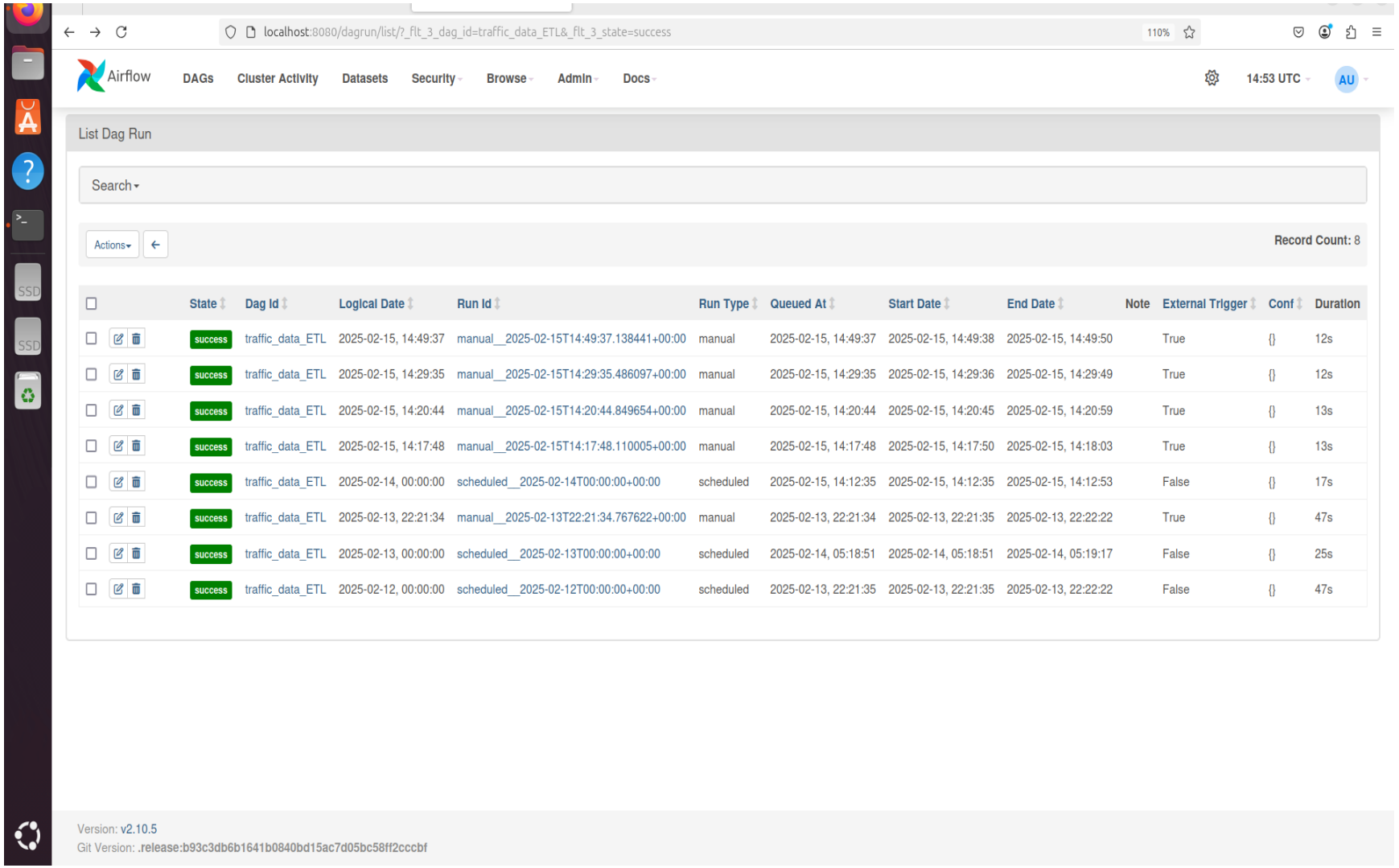


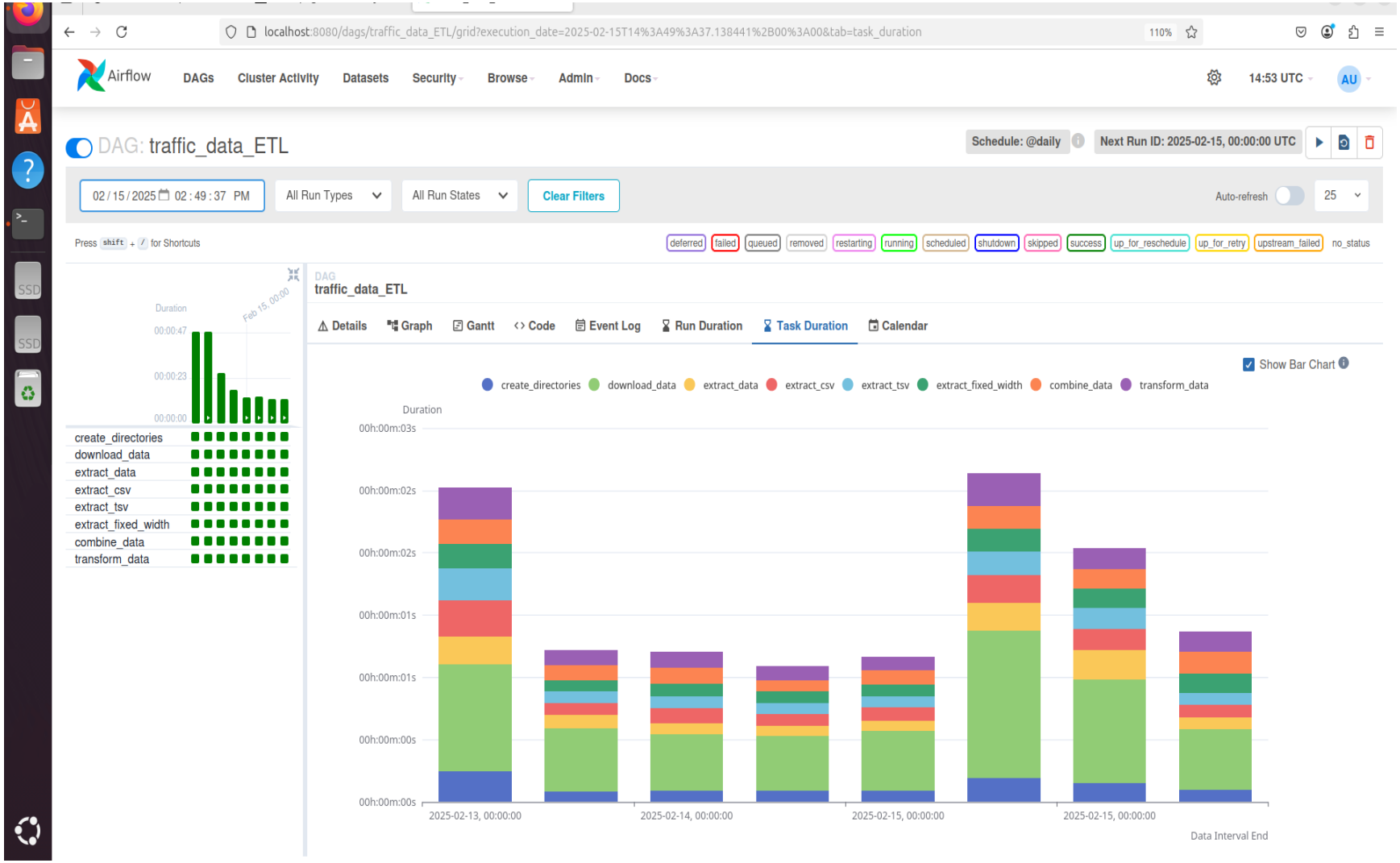
	State	Dag Id	Task Id	Run Id	Map Index	Logical Date	Operator	Start Date	End Date	Duration	Note	Job Id	Ho
	success	traffic_data_ETL	create_directories	scheduled__2025-02-12T00:00:00+00:00		2025-02-12, 00:00:00	BashOperator	2025-02-13, 22:21:37	2025-02-13, 22:21:37	<1s		1282	Nit
	success	traffic_data_ETL	create_directories	manual__2025-02-13T22:21:34.767622+00:00		2025-02-13, 22:21:34	BashOperator	2025-02-13, 22:21:40	2025-02-13, 22:21:40	<1s		1283	Nit
	success	traffic_data_ETL	download_data	scheduled__2025-02-12T00:00:00+00:00		2025-02-12, 00:00:00	PythonOperator	2025-02-13, 22:21:43	2025-02-13, 22:21:44	1s		1284	Nit
	success	traffic_data_ETL	download_data	manual__2025-02-13T22:21:34.767622+00:00		2025-02-13, 22:21:34	PythonOperator	2025-02-13, 22:21:47	2025-02-13, 22:21:47	<1s		1285	Nit
	success	traffic_data_ETL	extract_data	scheduled__2025-02-12T00:00:00+00:00		2025-02-12, 00:00:00	PythonOperator	2025-02-13, 22:21:51	2025-02-13, 22:21:51	<1s		1286	Nit
	success	traffic_data_ETL	extract_data	manual__2025-02-13T22:21:34.767622+00:00		2025-02-13, 22:21:34	PythonOperator	2025-02-13, 22:21:53	2025-02-13, 22:21:54	<1s		1287	Nit
	success	traffic_data_ETL	extract_csv	scheduled__2025-02-12T00:00:00+00:00		2025-02-12, 00:00:00	PythonOperator	2025-02-13, 22:21:56	2025-02-13, 22:21:57	<1s		1288	Nit
	success	traffic_data_ETL	extract_tsv	scheduled__2025-02-12T00:00:00+00:00		2025-02-12, 00:00:00	PythonOperator	2025-02-13, 22:21:59	2025-02-13, 22:21:59	<1s		1289	Nit
	success	traffic_data_ETL	extract_fixed_width	scheduled__2025-02-12T00:00:00+00:00		2025-02-12, 00:00:00	PythonOperator	2025-02-13, 22:22:02	2025-02-13, 22:22:02	<1s		1290	Nit
	success	traffic_data_ETL	extract_csv	manual__2025-02-13T22:21:34.767622+00:00		2025-02-13, 22:21:34	PythonOperator	2025-02-13, 22:22:04	2025-02-13, 22:22:05	<1s		1291	Nit
	success	traffic_data_ETL	extract_tsv	manual__2025-02-13T22:21:34.767622+00:00		2025-02-13, 22:21:34	PythonOperator	2025-02-13, 22:22:07	2025-02-13, 22:22:07	<1s		1292	Nit
	success	traffic_data_ETL	extract_fixed_width	manual__2025-02-13T22:21:34.767622+00:00		2025-02-13, 22:21:34	PythonOperator	2025-02-13, 22:22:10	2025-02-13, 22:22:10	<1s		1293	Nit
	success	traffic_data_ETL	combine_data	scheduled__2025-02-12T00:00:00+00:00		2025-02-12, 00:00:00	BashOperator	2025-02-13, 22:22:12	2025-02-13, 22:22:12	<1s		1294	Nit
	success	traffic_data_ETL	combine_data	manual__2025-02-13T22:21:34.767622+00:00		2025-02-13, 22:21:34	BashOperator	2025-02-13, 22:22:15	2025-02-13, 22:22:15	<1s		1295	Nit
	success	traffic_data_ETL	transform_data	scheduled__2025-02-12T00:00:00+00:00		2025-02-12, 00:00:00	PythonOperator	2025-02-13, 22:22:18	2025-02-13, 22:22:18	<1s		1296	Nit


localhost:8080/taskinstance/list/?_flt_3_dag_id=traffic_data_ETL&_flt_3_state=success
110%
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
14:54 UTC
AU
success traffic_data_ETL combine_data manual__2025-02-15T14:20:44.849654+00:00 2025-02-15, 14:20:44 BashOperator 2025-02-15, 14:20:57 2025-02-15, 14:20:57 <1s 1334 Nit
success traffic_data_ETL transform_data manual__2025-02-15T14:20:44.849654+00:00 2025-02-15, 14:20:44 PythonOperator 2025-02-15, 14:20:59 2025-02-15, 14:20:59 <1s 1335 Nit
success traffic_data_ETL create_directories manual__2025-02-15T14:29:35.486097+00:00 2025-02-15, 14:29:35 BashOperator 2025-02-15, 14:29:37 2025-02-15, 14:29:37 <1s 1338 Nit
success traffic_data_ETL download_data manual__2025-02-15T14:29:35.486097+00:00 2025-02-15, 14:29:35 PythonOperator 2025-02-15, 14:29:39 2025-02-15, 14:29:39 <1s 1339 Nit
success traffic_data_ETL extract_data manual__2025-02-15T14:29:35.486097+00:00 2025-02-15, 14:29:35 PythonOperator 2025-02-15, 14:29:41 2025-02-15, 14:29:41 <1s 1340 Nit
success traffic_data_ETL extract_csv manual__2025-02-15T14:29:35.486097+00:00 2025-02-15, 14:29:35 PythonOperator 2025-02-15, 14:29:42 2025-02-15, 14:29:42 <1s 1341 Nit
success traffic_data_ETL extract_tsv manual__2025-02-15T14:29:35.486097+00:00 2025-02-15, 14:29:35 PythonOperator 2025-02-15, 14:29:44 2025-02-15, 14:29:44 <1s 1342 Nit
success traffic_data_ETL extract_fixed_width manual__2025-02-15T14:29:35.486097+00:00 2025-02-15, 14:29:35 PythonOperator 2025-02-15, 14:29:45 2025-02-15, 14:29:45 <1s 1343 Nit
success traffic_data_ETL combine_data manual__2025-02-15T14:29:35.486097+00:00 2025-02-15, 14:29:35 BashOperator 2025-02-15, 14:29:47 2025-02-15, 14:29:47 <1s 1344 Nit
success traffic_data_ETL transform_data manual__2025-02-15T14:29:35.486097+00:00 2025-02-15, 14:29:35 PythonOperator 2025-02-15, 14:29:48 2025-02-15, 14:29:48 <1s 1345 Nit
success traffic_data_ETL create_directories manual__2025-02-15T14:49:37.138441+00:00 2025-02-15, 14:49:37 BashOperator 2025-02-15, 14:49:39 2025-02-15, 14:49:39 <1s 1346 Nit
success traffic_data_ETL download_data manual__2025-02-15T14:49:37.138441+00:00 2025-02-15, 14:49:37 PythonOperator 2025-02-15, 14:49:40 2025-02-15, 14:49:41 <1s 1347 Nit
success traffic_data_ETL extract_data manual__2025-02-15T14:49:37.138441+00:00 2025-02-15, 14:49:37 PythonOperator 2025-02-15, 14:49:42 2025-02-15, 14:49:42 <1s 1348 Nit
success traffic_data_ETL extract_csv manual__2025-02-15T14:49:37.138441+00:00 2025-02-15, 14:49:37 PythonOperator 2025-02-15, 14:49:43 2025-02-15, 14:49:44 <1s 1349 Nit
success traffic_data_ETL extract_tsv manual__2025-02-15T14:49:37.138441+00:00 2025-02-15, 14:49:37 PythonOperator 2025-02-15, 14:49:45 2025-02-15, 14:49:45 <1s 1350 Nit
success traffic_data_ETL extract_fixed_width manual__2025-02-15T14:49:37.138441+00:00 2025-02-15, 14:49:37 PythonOperator 2025-02-15, 14:49:46 2025-02-15, 14:49:47 <1s 1351 Nit
success traffic_data_ETL combine_data manual__2025-02-15T14:49:37.138441+00:00 2025-02-15, 14:49:37 BashOperator 2025-02-15, 14:49:48 2025-02-15, 14:49:48 <1s 1352 Nit
success traffic_data_ETL transform_data manual__2025-02-15T14:49:37.138441+00:00 2025-02-15, 14:49:37 PythonOperator 2025-02-15, 14:49:49 2025-02-15, 14:49:50 <1s 1353 Nit
Version: v2.10.5
Git Version: .release:b93c3db6b1641b0840bd15ac7d05bc58ff2cccbf



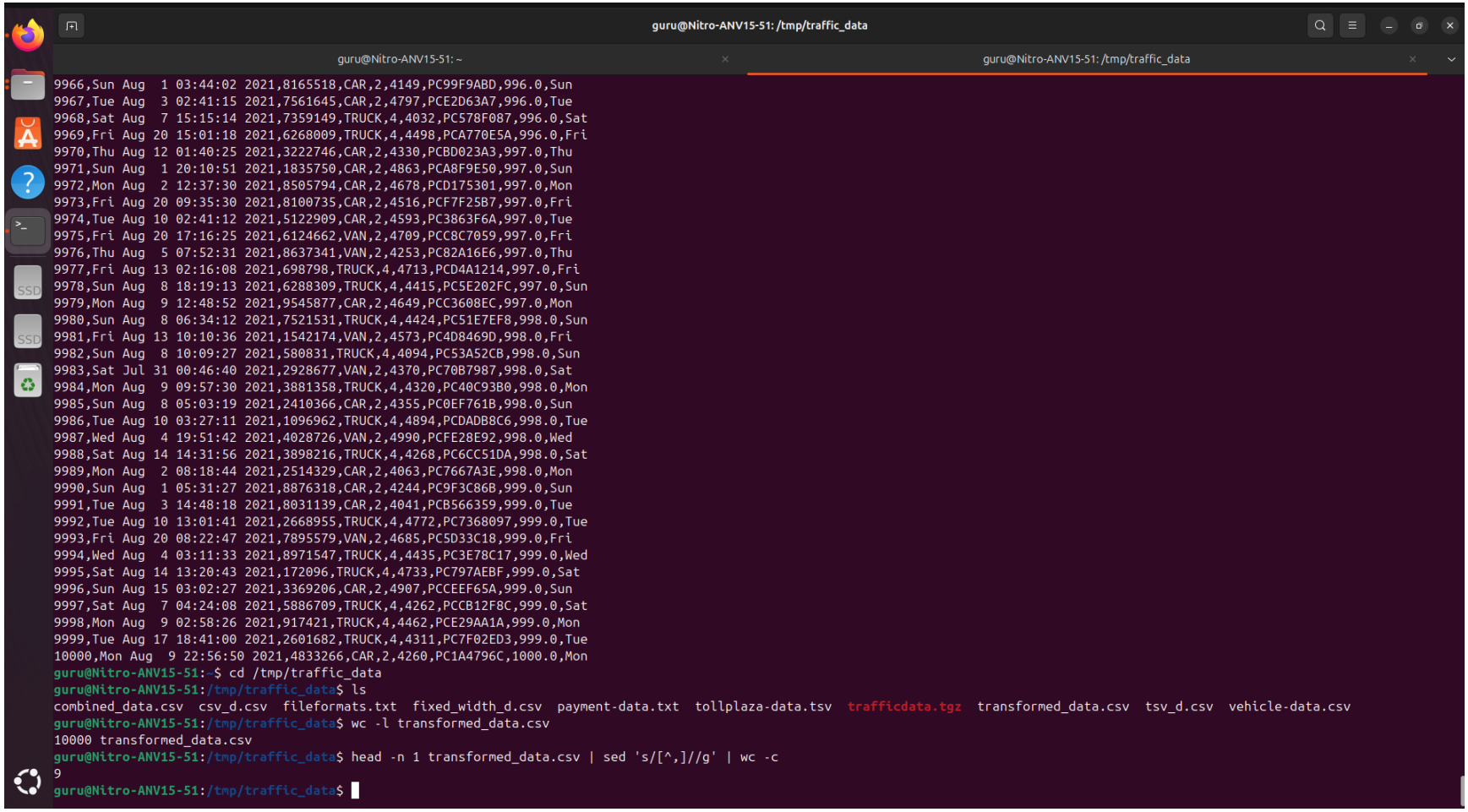
success	traffic_data_ETL	create_directories	scheduled__2025-02-13T00:00:00+00:00	2025-02-13, 00:00:00	BashOperator	2025-02-14, 05:18:53	2025-02-14, 05:18:53	<1s	1298	Nit
success	traffic_data_ETL	download_data	scheduled__2025-02-13T00:00:00+00:00	2025-02-13, 00:00:00	PythonOperator	2025-02-14, 05:18:58	2025-02-14, 05:18:58	<1s	1300	Nit
success	traffic_data_ETL	extract_data	scheduled__2025-02-13T00:00:00+00:00	2025-02-13, 00:00:00	PythonOperator	2025-02-14, 05:19:04	2025-02-14, 05:19:04	<1s	1302	Nit
success	traffic_data_ETL	extract_csv	scheduled__2025-02-13T00:00:00+00:00	2025-02-13, 00:00:00	PythonOperator	2025-02-14, 05:19:07	2025-02-14, 05:19:07	<1s	1303	Nit
success	traffic_data_ETL	extract_tsv	scheduled__2025-02-13T00:00:00+00:00	2025-02-13, 00:00:00	PythonOperator	2025-02-14, 05:19:09	2025-02-14, 05:19:09	<1s	1304	Nit
success	traffic_data_ETL	extract_fixed_width	scheduled__2025-02-13T00:00:00+00:00	2025-02-13, 00:00:00	PythonOperator	2025-02-14, 05:19:12	2025-02-14, 05:19:12	<1s	1305	Nit
success	traffic_data_ETL	combine_data	scheduled__2025-02-13T00:00:00+00:00	2025-02-13, 00:00:00	BashOperator	2025-02-14, 05:19:14	2025-02-14, 05:19:14	<1s	1306	Nit
success	traffic_data_ETL	transform_data	scheduled__2025-02-13T00:00:00+00:00	2025-02-13, 00:00:00	PythonOperator	2025-02-14, 05:19:16	2025-02-14, 05:19:16	<1s	1307	Nit
success	traffic_data_ETL	create_directories	scheduled__2025-02-14T00:00:00+00:00	2025-02-14, 00:00:00	BashOperator	2025-02-15, 14:12:37	2025-02-15, 14:12:37	<1s	1310	Nit
success	traffic_data_ETL	download_data	scheduled__2025-02-14T00:00:00+00:00	2025-02-14, 00:00:00	PythonOperator	2025-02-15, 14:12:39	2025-02-15, 14:12:40	<1s	1312	Nit
success	traffic_data_ETL	extract_data	scheduled__2025-02-14T00:00:00+00:00	2025-02-14, 00:00:00	PythonOperator	2025-02-15, 14:12:42	2025-02-15, 14:12:42	<1s	1314	Nit
success	traffic_data_ETL	extract_csv	scheduled__2025-02-14T00:00:00+00:00	2025-02-14, 00:00:00	PythonOperator	2025-02-15, 14:12:44	2025-02-15, 14:12:44	<1s	1315	Nit
success	traffic_data_ETL	extract_tsv	scheduled__2025-02-14T00:00:00+00:00	2025-02-14, 00:00:00	PythonOperator	2025-02-15, 14:12:46	2025-02-15, 14:12:46	<1s	1316	Nit
success	traffic_data_ETL	extract_fixed_width	scheduled__2025-02-14T00:00:00+00:00	2025-02-14, 00:00:00	PythonOperator	2025-02-15, 14:12:47	2025-02-15, 14:12:48	<1s	1317	Nit
success	traffic_data_ETL	combine_data	scheduled__2025-02-14T00:00:00+00:00	2025-02-14, 00:00:00	BashOperator	2025-02-15, 14:12:49	2025-02-15, 14:12:50	<1s	1318	Nit
success	traffic_data_ETL	transform_data	scheduled__2025-02-14T00:00:00+00:00	2025-02-14, 00:00:00	PythonOperator	2025-02-15, 14:12:52	2025-02-15, 14:12:52	<1s	1319	Nit
success	traffic_data_ETL	create_directories	manual__2025-02-15T14:17:48.110005+00:00	2025-02-15, 14:17:48	BashOperator	2025-02-15, 14:17:51	2025-02-15, 14:17:51	<1s	1320	Nit
success	traffic_data_ETL	download_data	manual__2025-02-15T14:17:48.110005+00:00	2025-02-15, 14:17:48	PythonOperator	2025-02-15, 14:17:52	2025-02-15, 14:17:53	<1s	1321	Nit
success	traffic_data_ETL	extract_data	manual__2025-02-15T14:17:48.110005+00:00	2025-02-15, 14:17:48	PythonOperator	2025-02-15, 14:17:54	2025-02-15, 14:17:54	<1s	1322	Nit
success	traffic_data_ETL	extract_csv	manual__2025-02-15T14:17:48.110005+00:00	2025-02-15, 14:17:48	PythonOperator	2025-02-15, 14:17:56	2025-02-15, 14:17:56	<1s	1323	Nit
success	traffic_data_ETL	extract_tsv	manual__2025-02-15T14:17:48.110005+00:00	2025-02-15, 14:17:48	PythonOperator	2025-02-15, 14:17:57	2025-02-15, 14:17:57	<1s	1324	Nit


localhost:8080/dagrun/list/?_flt_3_dag_id=traffic_data_ETL&_flt_3_state=success
110%
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
14:53 UTC
AU
List Dag Run
Search
Actions
Record Count: 8
State
Dag Id
Logical Date
Run Id
Run Type
Queued At
Start Date
End Date
Note
External Trigger
Conf
Duration
success traffic_data_ETL 2025-02-15, 14:49:37 manual__2025-02-15T14:49:37.138441+00:00 manual 2025-02-15, 14:49:37 2025-02-15, 14:49:38 2025-02-15, 14:49:50 True {} 12s
success traffic_data_ETL 2025-02-15, 14:29:35 manual__2025-02-15T14:29:35.486097+00:00 manual 2025-02-15, 14:29:35 2025-02-15, 14:29:36 2025-02-15, 14:29:49 True {} 12s
success traffic_data_ETL 2025-02-15, 14:20:44 manual__2025-02-15T14:20:44.849654+00:00 manual 2025-02-15, 14:20:44 2025-02-15, 14:20:45 2025-02-15, 14:20:59 True {} 13s
success traffic_data_ETL 2025-02-15, 14:17:48 manual__2025-02-15T14:17:48.110005+00:00 manual 2025-02-15, 14:17:48 2025-02-15, 14:17:50 2025-02-15, 14:18:03 True {} 13s
success traffic_data_ETL 2025-02-14, 00:00:00 scheduled__2025-02-14T00:00:00+00:00 scheduled 2025-02-15, 14:12:35 2025-02-15, 14:12:35 2025-02-15, 14:12:53 False {} 17s
success traffic_data_ETL 2025-02-13, 22:21:34 manual__2025-02-13T22:21:34.767622+00:00 manual 2025-02-13, 22:21:34 2025-02-13, 22:21:35 2025-02-13, 22:22:22 True {} 47s
success traffic_data_ETL 2025-02-13, 00:00:00 scheduled__2025-02-13T00:00:00+00:00 scheduled 2025-02-14, 05:18:51 2025-02-14, 05:18:51 2025-02-14, 05:19:17 False {} 25s
success traffic_data_ETL 2025-02-12, 00:00:00 scheduled__2025-02-12T00:00:00+00:00 scheduled 2025-02-13, 22:21:35 2025-02-13, 22:21:35 2025-02-13, 22:22:22 False {} 47s
Version: v2.10.5
Git Version: .release:b93c3db6b1641b0840bd15ac7d05bc58ff2cccbf


localhost:8080/dags/traffic_data_ETL/grid?execution_date=2025-02-15T14%3A49%3A37.138441%2B00%3A00&tab=task_duration
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
14:53 UTC
AU
DAG: traffic_data_ETL
Schedule: @daily
Next Run ID: 2025-02-15, 00:00:00 UTC
02 / 15 / 2025 02 : 49 : 37 PM
All Run Types
All Run States
Clear Filters
Auto-refresh
25
Press shift + / for Shortcuts
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Duration
00:00:47
00:00:23
00:00:00
Feb 15, 00:00:00
create_directories
download_data
extract_data
extract_csv
extract_tsv
extract_fixed_width
combine_data
transform_data
DAG
traffic_data_ETL
Details
Graph
Gantt
Code
Event Log
Run Duration
Task Duration
Calendar
Show Bar Chart
create_directories
download_data
extract_data
extract_csv
extract_tsv
extract_fixed_width
combine_data
transform_data
Duration
00h:00m:03s
00h:00m:02s
00h:00m:02s
00h:00m:01s
00h:00m:01s
00h:00m:00s
00h:00m:00s
2025-02-13, 00:00:00
2025-02-14, 00:00:00
2025-02-15, 00:00:00
2025-02-15, 00:00:00
Data Interval End

Results

- The extracted and transformed data was saved in the /tmp/traffic_data directory.
- The final processed file, transformed.csv, contained the expected transformed data.



Observations

1. DAG Structure & Task Execution

- The DAG consists of several tasks, each responsible for different stages of the ETL process:
    - **create_directories**: Ensures necessary directories are available before data extraction.
    - **download_data**: Fetches raw traffic data, possibly from an API or external source.
    - **extract_data**: Extracts relevant portions from the downloaded data.
    - **extract_csv, extract_tsv, extract_fixed_width**: Handle extraction from different file formats.
    - **combine_data**: Merges data from multiple sources into a unified dataset.
    - **transform_data**: Applies necessary data cleaning and processing before storage or analysis.
- The execution follows a structured workflow where tasks are executed in a sequence, ensuring data integrity and correctness.

2. Task Execution Status

- All tasks in the DAG have been successfully executed, as indicated by the green status markers.
- No failed or pending tasks were observed, suggesting that the ETL pipeline is functioning without errors.
- The DAG execution timestamps indicate consistent processing, with runs scheduled daily.

3. Task Duration Analysis

- The **Task Duration Chart** highlights the execution time for individual tasks over multiple DAG runs.
- **download_data** consistently records the highest execution time, indicating that fetching raw data is a time-consuming process.
- **combine_data** also takes a relatively longer time compared to other tasks, which is expected as it integrates multiple data sources.

- Other tasks, such as **create_directories** and **extract_data**, execute within milliseconds to seconds, suggesting efficient performance.

4. Execution Trends Over Time

- The **Run Duration Chart** provides insights into total DAG execution time across multiple days (February 13–15, 2025).
- The DAG run on **February 13, 2025**, took the longest time (~50 seconds), while subsequent runs show reduced execution times.
- The decreasing execution time trend indicates potential optimizations in data handling or reduced processing loads.
- Recent DAG executions have stabilized, maintaining a consistent duration across multiple runs.

5. Gantt Chart Observations

- The **Gantt Chart View** provides a timeline representation of task execution.
- Tasks are executed sequentially, with minimal idle time between them, suggesting efficient scheduling.
- Green bars in the chart confirm successful completion, while gray bars indicate queued tasks waiting for execution.
- The **combine_data** task appears to have the longest execution span, aligning with findings from the Task Duration analysis.

6. Airflow Monitoring & Interface Features

- The Airflow UI provides multiple visualization options such as:
    - **Graph View**: Shows task dependencies and execution flow.
    - **Gantt Chart View**: Displays task execution duration and scheduling gaps.
    - **Run Duration View**: Tracks execution time trends over multiple DAG runs.
    - **Task Duration View**: Highlights individual task performance and bottlenecks.
- The system actively monitors DAG runs, ensuring any performance issues or failures are easily identifiable.

Conclusion

The **traffic_data_ETL** pipeline in Apache Airflow executes successfully with structured task dependencies and efficient scheduling. The **download_data** and **combine_data** tasks require the longest execution time, making them key areas to monitor for performance improvements. The DAG's execution time has decreased over multiple runs, suggesting optimizations or variations in data volume. The monitoring capabilities in Airflow allow detailed tracking of task performance, ensuring smooth operation of the ETL workflow.