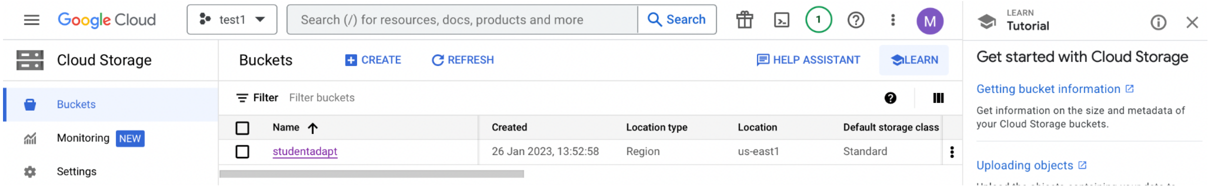
Data Engineering Final Assignment

1. Created a new project on Google Cloud Platform(GCP) project-id `test1-375701`
2. Created a bucket with the name “studentadapt” in cloud storage under the same project



□	Name	Created	Location type	Location	Default storage class
□	studentadapt	26 Jan 2023, 13:52:58	Region	us-east1	Standard

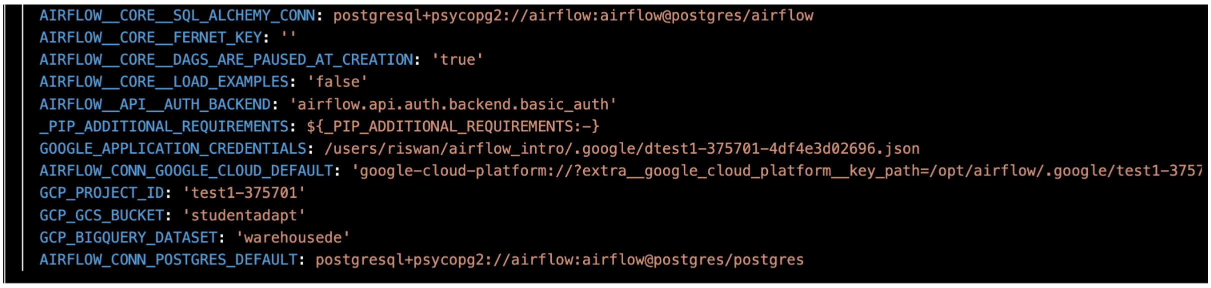
Get started with Cloud Storage

Getting bucket information

Get information on the size and metadata of your Cloud Storage buckets.

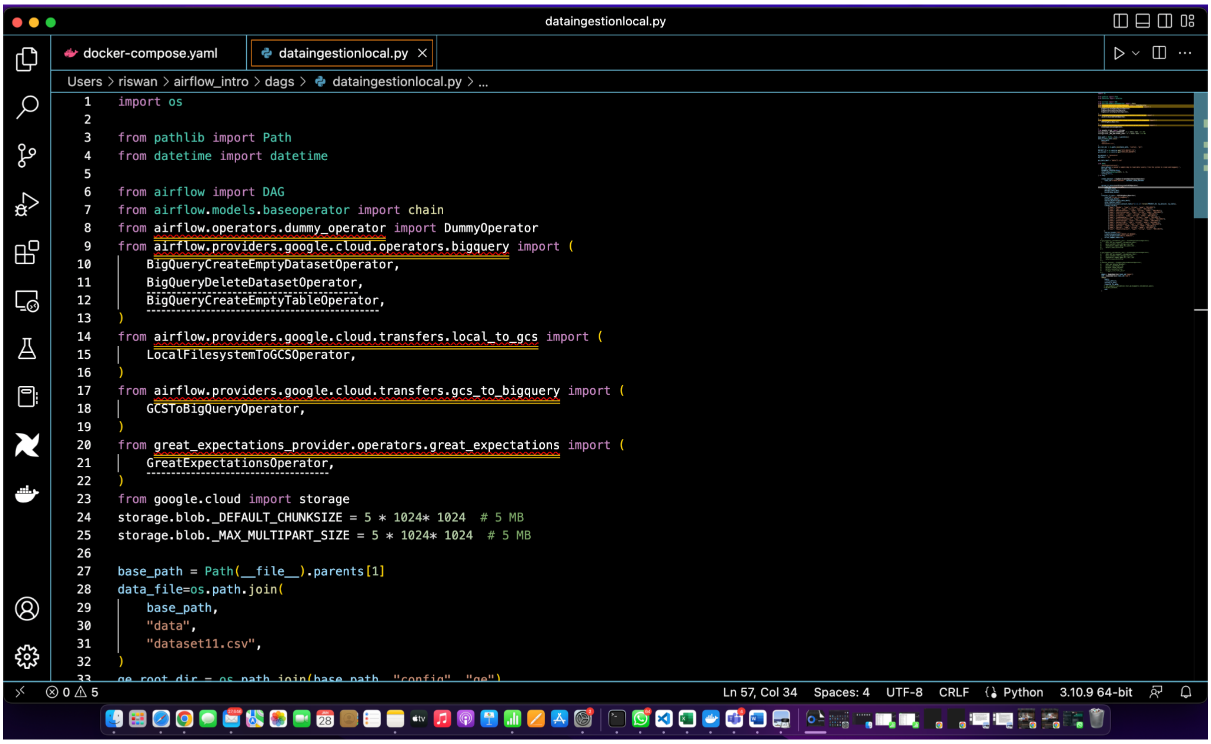
Uploading objects

3. Created a new dataset table in bigquery and named it “dataset11”
4. Created and downloaded service account key .json file and saved it locally in the airflow docker folder
5. Updated docker-compose.yaml file environment variables.



```
AIRFLOW__CORE__SQL_ALCHEMY_CONN: postgresql+psycopg2://airflow:airflow@postgres/airflow
AIRFLOW__CORE__FERNET_KEY: ''
AIRFLOW__CORE__DAGS_ARE_PAUSED_AT_CREATION: 'true'
AIRFLOW__CORE__LOAD_EXAMPLES: 'false'
AIRFLOW__API__AUTH_BACKEND: 'airflow.api.auth.backend.basic_auth'
_PIP_ADDITIONAL_REQUIREMENTS: ${_PIP_ADDITIONAL_REQUIREMENTS:-}
GOOGLE_APPLICATION_CREDENTIALS: /users/riswan/airflow_intro/.google/dtest1-375701-4df4e3d02696.json
AIRFLOW_CONN_GOOGLE_CLOUD_DEFAULT: 'google-cloud-platform://?extra__google_cloud_platform__key_path=/opt/airflow/.google/test1-3757
GCP_PROJECT_ID: 'test1-375701'
GCP_GCS_BUCKET: 'studentadapt'
GCP_BIGQUERY_DATASET: 'warehousede'
AIRFLOW_CONN_POSTGRES_DEFAULT: postgresql+psycopg2://airflow:airflow@postgres/postgres
```

6. Created a DAG python file under the dags folder named as `dataingestionlocal.py` (the dag file is present in airflow_intro>dags)



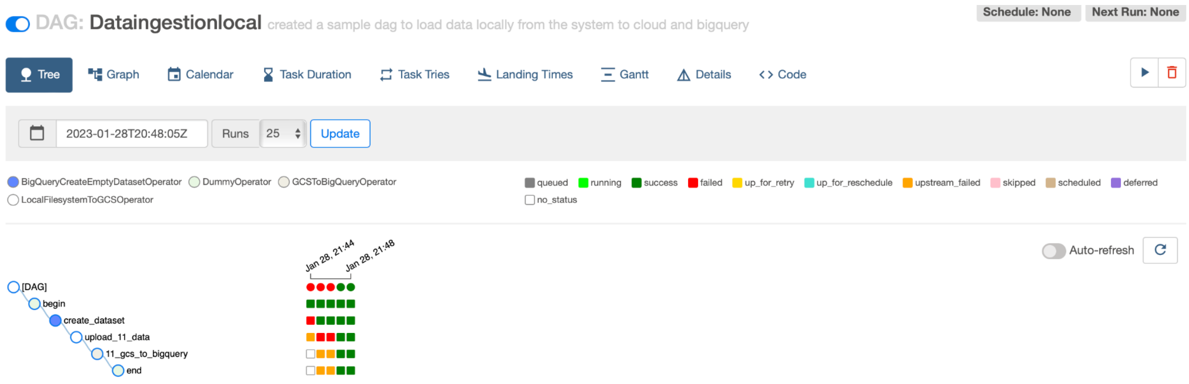
```python
import os

from pathlib import Path
from datetime import datetime

from airflow import DAG
from airflow.models.baseoperator import chain
from airflow.operators.dummy_operator import DummyOperator
from airflow.providers.google.cloud.operators.bigquery import (
    BigQueryCreateEmptyDatasetOperator,
    BigQueryDeleteDatasetOperator,
    BigQueryCreateEmptyTableOperator,
)
from airflow.providers.google.cloud.transfers.local_to_gcs import (
    LocalFilesystemToGCSOperator,
)
from airflow.providers.google.cloud.transfers.gcs_to_bigquery import (
    GCSToBigQueryOperator,
)
from great_expectations_provider.operators.great_expectations import (
    GreatExpectationsOperator,
)
from google.cloud import storage
storage.blob._DEFAULT_CHUNKSIZE = 5 * 1024* 1024  # 5 MB
storage.blob._MAX_MULTIPART_SIZE = 5 * 1024* 1024  # 5 MB

base_path = Path(__file__).parents[1]
data_file=os.path.join(
    base_path,
    "data",
    "dataset11.csv",
)
ge_root_dir = os.path.join(base_path, "config", "ge")
```

7. Initialised docker-compose up yaml file (inside folder airflow_intro>)
8. Made sure that all the docker containers in airflow_intro folder is running
9. Opened and logged in airflow user interface
10. Triggered the DAG dataingestionlocal
11. Result show as follows



DAG: Dataingestionlocal created a sample dag to load data locally from the system to cloud and bigquery

Schedule: None Next Run: None

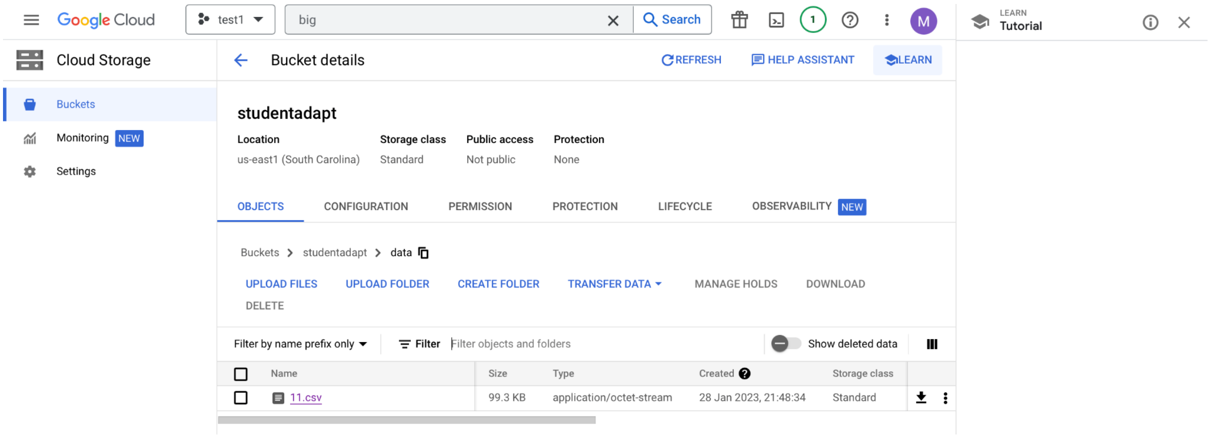
2023-01-28T20:48:05Z Runs 25

● BigQueryCreateEmptyDatasetOperator ○ DummyOperator ○ GCSToBigQueryOperator ○ LocalFilesystemToGCSOperator

[DAG]	🟥	🟥	🟥	🟩	🟩
begin	🟩	🟩	🟩	🟩	🟩
create_dataset	🟥	🟩	🟩	🟩	🟩
upload_11_data	🟧	🟥	🟥	🟩	🟩
11_gcs_to_bigquery	□	🟧	🟧	🟩	🟩
end	□	🟧	🟧	🟩	🟩

12. It shows that the process has begun, data set created in cloud storage, the data uploaded to cloud storage, the data transferred from cloud storage to bigquery.

13. The data in google cloud bucket is shown as follows



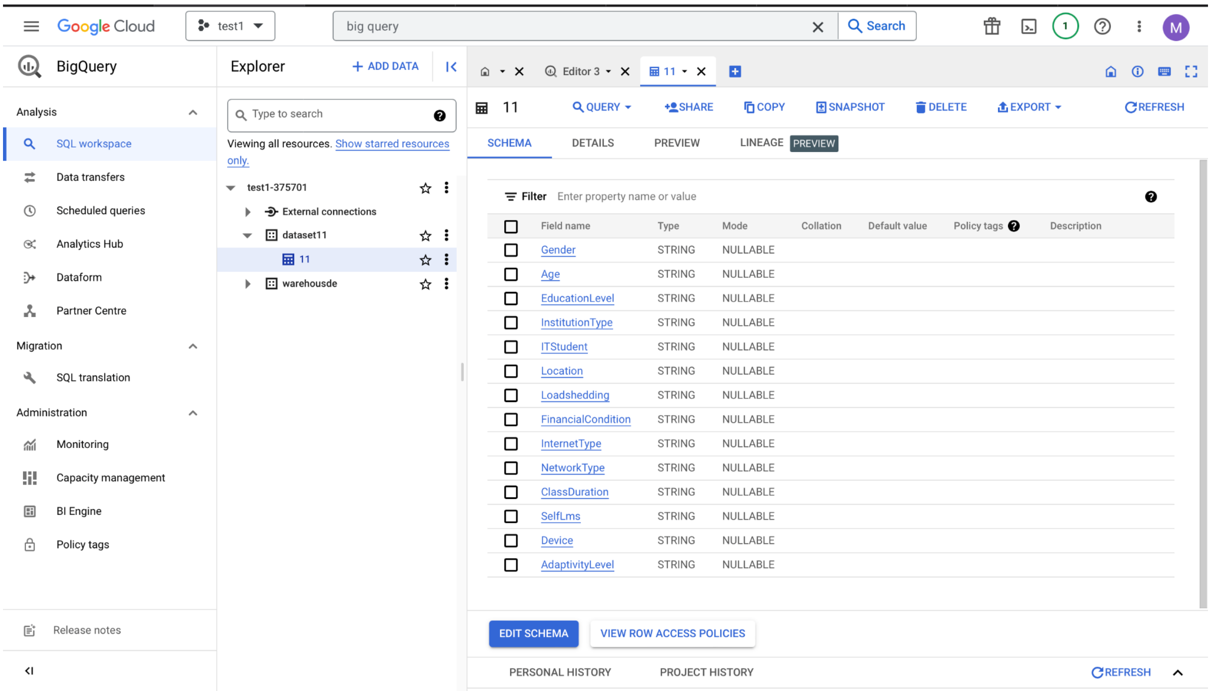
studentadapt

Location	Storage class	Public access	Protection
us-east1 (South Carolina)	Standard	Not public	None

Buckets > studentadapt > data

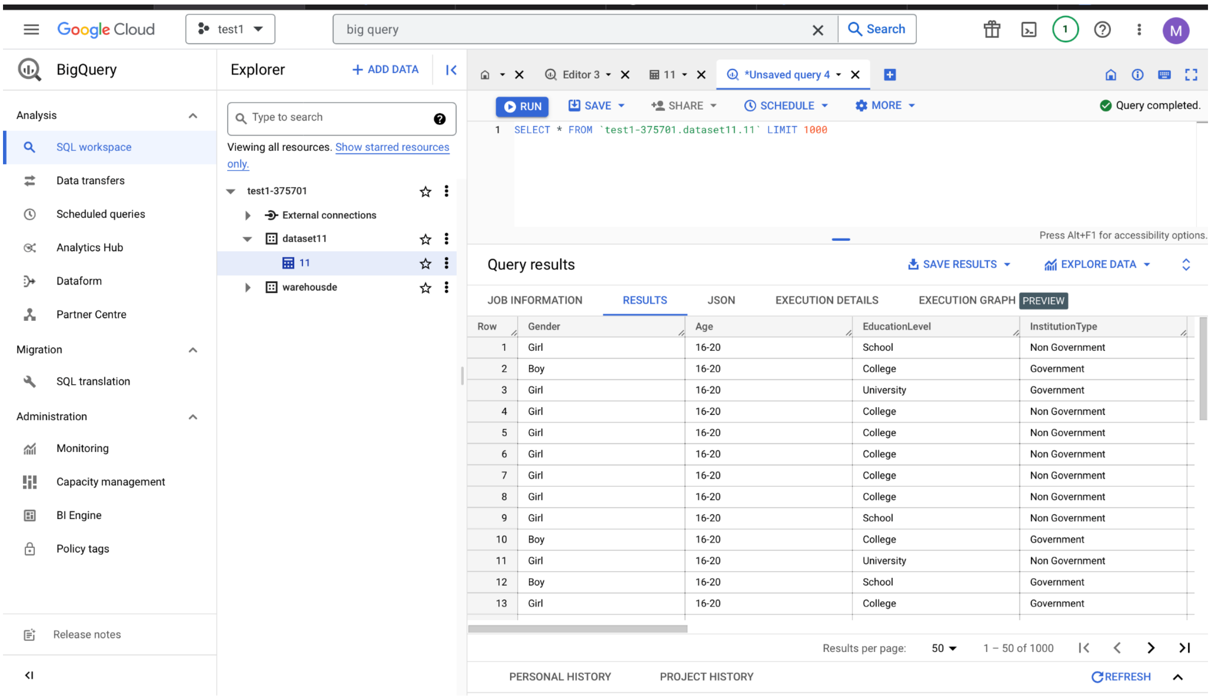
Name	Size	Type	Created	Storage class
11.csv	99.3 KB	application/octet-stream	28 Jan 2023, 21:48:34	Standard

14. Data reflected in BigQuery:



Field name	Type	Mode	Collation	Default value	Policy tags	Description
Gender	STRING	NULLABLE				
Age	STRING	NULLABLE				
EducationLevel	STRING	NULLABLE				
InstitutionType	STRING	NULLABLE				
ITStudent	STRING	NULLABLE				
Location	STRING	NULLABLE				
Loadshedding	STRING	NULLABLE				
FinancialCondition	STRING	NULLABLE				
InternetType	STRING	NULLABLE				
NetworkType	STRING	NULLABLE				
ClassDuration	STRING	NULLABLE				
SelfLms	STRING	NULLABLE				
Device	STRING	NULLABLE				
AdaptivityLevel	STRING	NULLABLE				

15. Running a sample query to make sure the data has reached the datawarehouse.



Query completed.

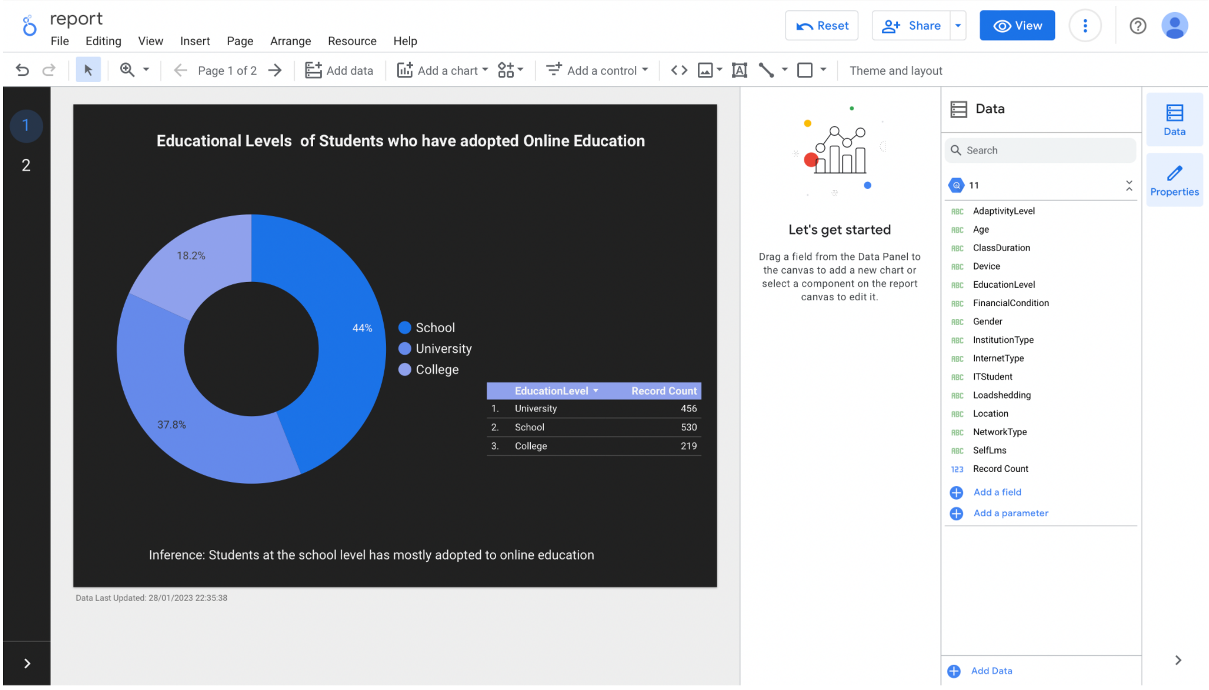
```
SELECT * FROM `test1-375701.dataset11.11` LIMIT 1000
```

Row	Gender	Age	EducationLevel	InstitutionType
1	Girl	16-20	School	Non Government
2	Boy	16-20	College	Government
3	Girl	16-20	University	Government
4	Girl	16-20	College	Non Government
5	Girl	16-20	College	Non Government
6	Girl	16-20	College	Non Government
7	Girl	16-20	College	Non Government
8	Girl	16-20	College	Non Government
9	Girl	16-20	School	Non Government
10	Boy	16-20	College	Government
11	Girl	16-20	University	Non Government
12	Boy	16-20	School	Government
13	Girl	16-20	College	Government

Results per page: 50 1 – 50 of 1000

16. Created a blank report in Google Looker Studio and connected it with BigQuery using in-built function of google connectors.
17. The project name test1 was selected as shown in bigquery and the dataset was loaded.
18. Created 2 graphs
19. Graph1 shows quantitative educational levels of students who have adopted online education. Graph was returned as

below:



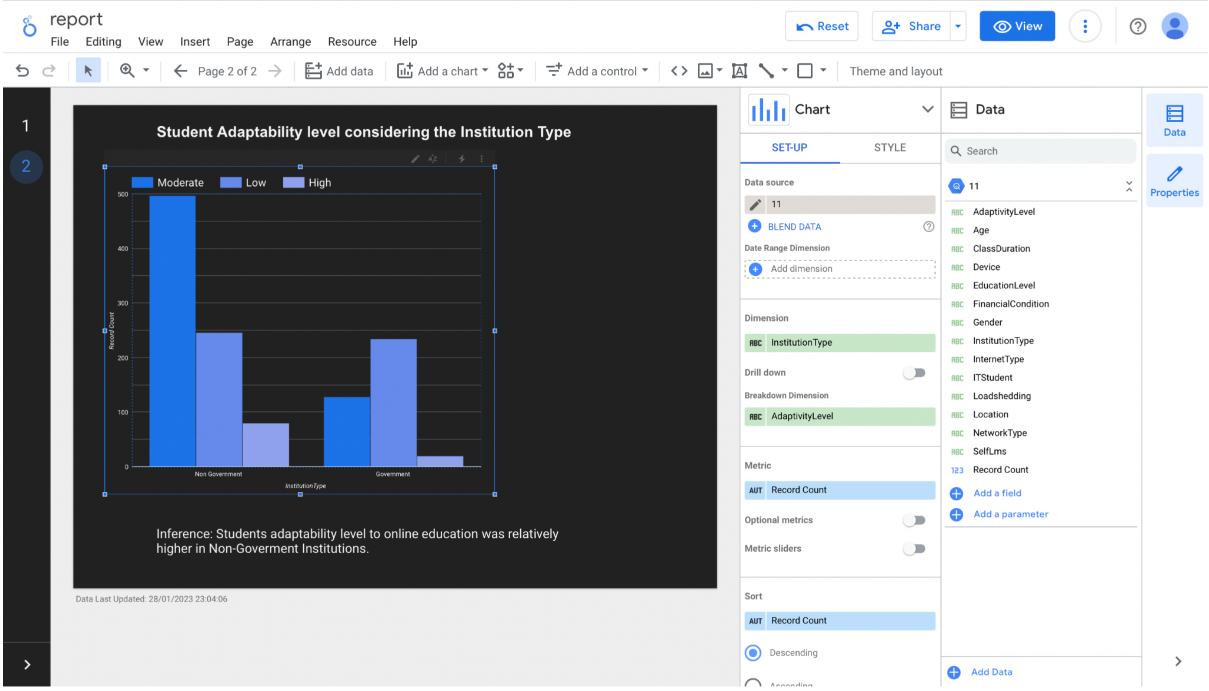
Educational Levels of Students who have adopted Online Education

	EducationLevel ▾	Record Count
1.	University	456
2.	School	530
3.	College	219

Inference: Students at the school level has mostly adopted to online education

Data Last Updated: 28/01/2023 22:35:38

20. Graph2 shows categorical data of students' adaptability to online education based on the type of institution.



Student Adaptability level considering the Institution Type

Inference: Students adaptability level to online education was relatively higher in Non-Goverment Institutions.

Data Last Updated: 28/01/2023 23:04:06

Data source: 11

BLEND DATA

Date Range Dimension: Add dimension

Dimension: InstitutionType

Drill down

Breakdown Dimension: AdaptivityLevel

Metric: Record Count

Optional metrics

Metric sliders

Sort: Record Count

● Descending

○ Ascending