Wikipedia Hourly Pageviews Analytics

Presented by Monica Ghavalyan, Associate Data Engineer at SADA

SADA™
An Insight company

Project Objective

The aim is to build an automated, daily batch pipeline that ingests hourly pageview data from Wikipedia, processes it to find the top 10 most viewed articles for the English Wikipedia (en.wikipedia.org), and loads the aggregated results into BigQuery for analysis.

Used Services and Software

- **Spark**: Distributed computing engine for large-scale data processing across clusters
- **Dataproc**: Google Cloud's managed service for running Apache Spark and Hadoop clusters
- **Airflow**: Open-source platform for orchestrating and scheduling complex data workflows
- **Composer**: Google Cloud's fully managed Apache Airflow service for workflow orchestration

Setup Environment

Cloud Storage Buckets

- project4-landing-zone - Raw Wikipedia data storage
- project4-processed-zone - Spark job output storage
- project4-dataproc-staging - Dataproc temporary files

Cloud Composer

- Environment: my-composer
- DAG location: gs://us-central1-my-composer-82fad1a3-bucket/dags/
- Scripts location: gs://us-central1-my-composer-82fad1a3-bucket/scripts/

BigQuery

- Dataset: wikipedia_analytics
- Table: top_en_articles_daily

Code Files

- wikipedia_pipeline_dag.py - Airflow workflow orchestration
- process_wiki_views.py - PySpark data processing script

Dataproc

- Ephemeral cluster: wikipedia-processing-cluster
- Purpose: Distributed data processing

PySpark Data Processing Script

Main Functionality

- Reads 24 hourly compressed Wikipedia pageview files (.gz format)
- Filters data for English Wikipedia (desktop and mobile)
- Aggregates view counts by article title across 24-hour period
- Identifies top 10 most viewed articles
- Outputs results as Parquet format for BigQuery

Data Schema & Types

- **Input Schema**:
  - domain_code (StringType) - Language/project identifier
  - page_title (StringType) - Article name
  - count_views (LongType) - Hourly view count
  - total_response_size (IntegerType) - Unused field

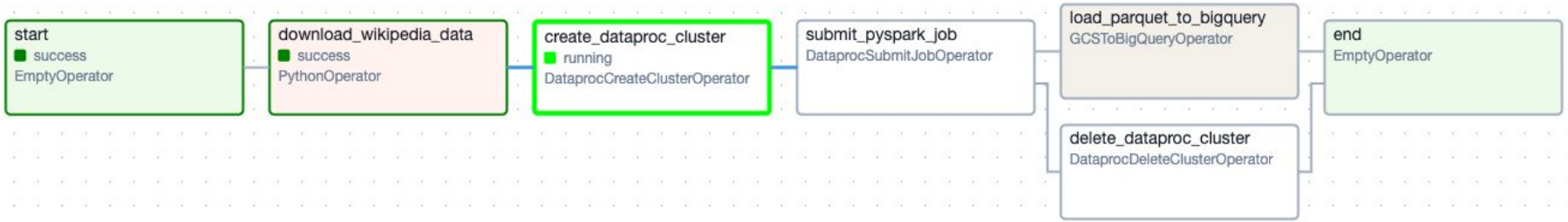
Key Transformations

- **Filter**: Keep only "en" and "en.m" domain codes
- **GroupBy**: Aggregate by article title using groupBy("page_title")
- **Sum**: Total views with sum("count_views").alias("total_views")
- **Sort**: Descending order by view count using orderBy(desc())
- **Limit**: Top 10 results using .limit(10)
- **Add Column**: Processing date for data lineage

Output Format

- **File Type**: Parquet (columnar, compressed)
- **Write Mode**: Overwrite existing data
- **Final Columns**: article_title, total_views, processing_date

Airflow DAG Tasks



Airflow DAG Tasks

Task 1: Download Wikipedia Data

- **Operator Used**: `PythonOperator`
- **Functionality**:
  - Executes a Python function that loops 24 times to download the previous day's hourly data files.
  - Uses `curl` and `gsutil` shell commands to stream each compressed file directly to the raw data storage bucket (`project4-landing-zone`).

Airflow DAG Tasks

Task 2 & 5: Ephemeral Cluster Management

- **Operators Used**: `DataprocCreateClusterOperator` & `DataprocDeleteClusterOperator`
- **Functionality**:
  - **Create**: Provisions a temporary `wikipedia-processing-cluster` for the Spark job. This is known as an ephemeral cluster.
  - **Delete**: Destroys the cluster after processing is complete to manage costs effectively. This task runs even if the Spark job fails.

Airflow DAG Tasks

Task 3: Submit PySpark Job

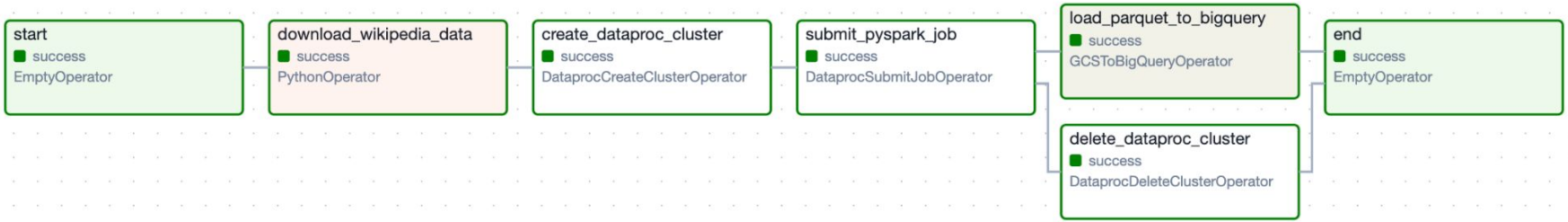
- **Operator Used**: `DataprocSubmitJobOperator`
- **Functionality**:
  - Submits the `process_wiki_views.py` script to the Dataproc cluster for distributed processing.
  - Passes critical arguments to the script, including the input/output GCS paths and the specific date to be processed.

Airflow DAG Tasks

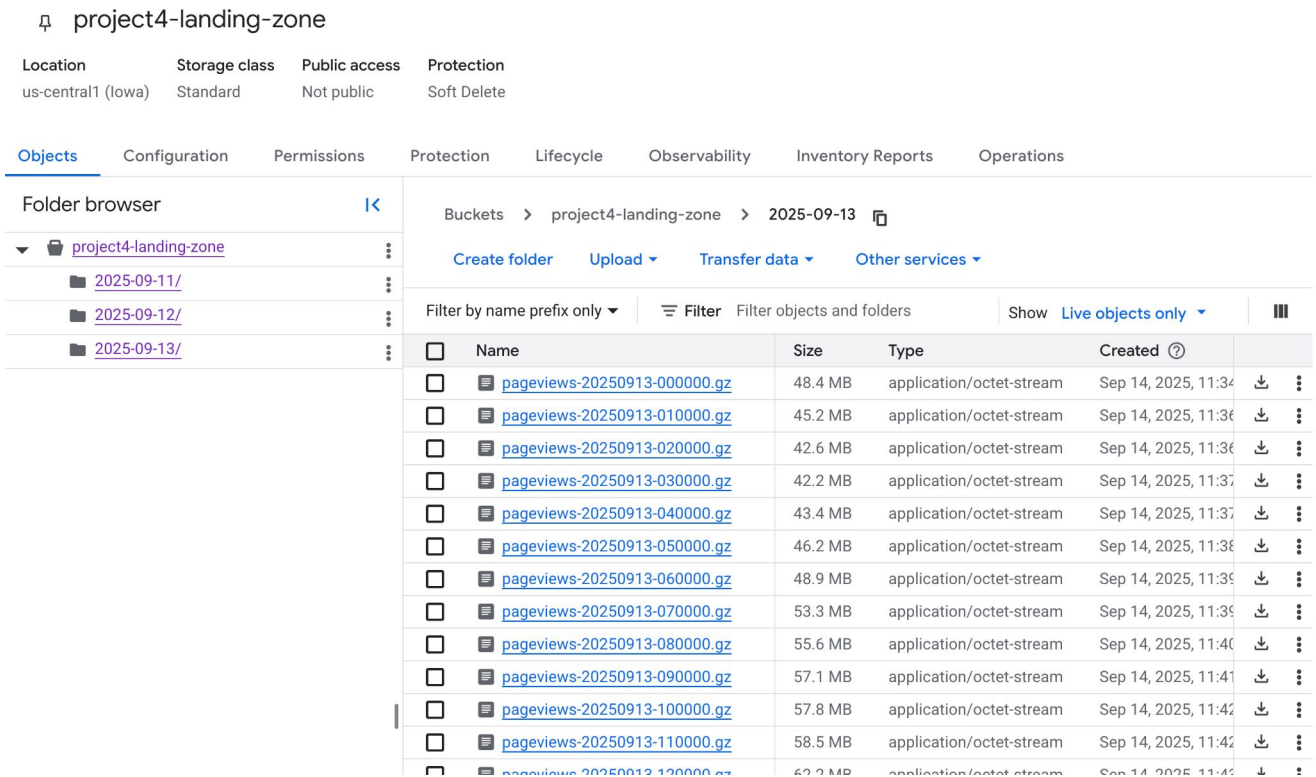
Task 4: Load Data to BigQuery

- **Operator Used**: `GCSToBigQueryOperator`
- **Functionality**:
  - Loads the final Parquet file from the processed data bucket (`project4-processed-zone`) into BigQuery.
  - Appends the top 10 articles to the `top_en_articles_daily` table within the `wikipedia_analytics` dataset.

Results - Succeeded DAG

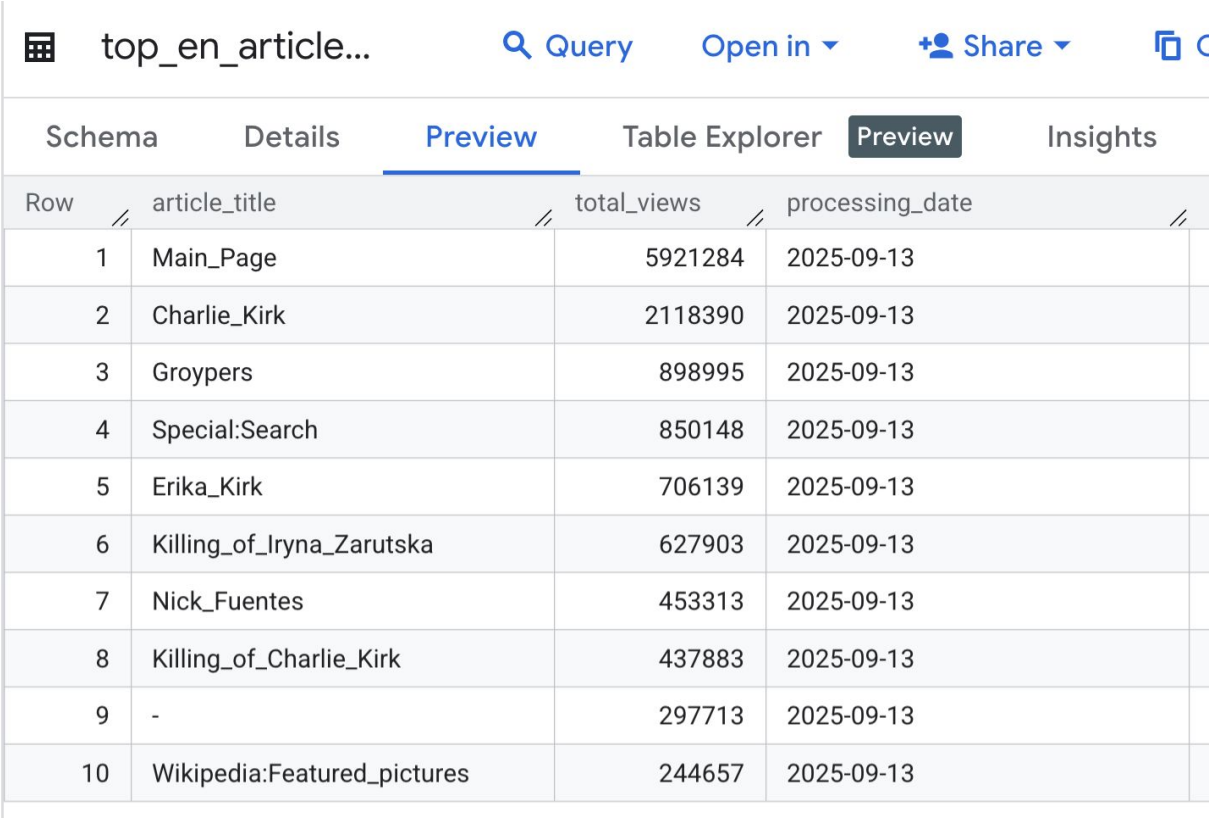


Results - Downloads



SADA™
An Insight company

Results - BigQuery Table



Row	article_title	total_views	processing_date
1	Main_Page	5921284	2025-09-13
2	Charlie_Kirk	2118390	2025-09-13
3	Groypers	898995	2025-09-13
4	Special:Search	850148	2025-09-13
5	Erika_Kirk	706139	2025-09-13
6	Killing_of_Iryna_Zarutska	627903	2025-09-13
7	Nick_Fuentes	453313	2025-09-13
8	Killing_of_Charlie_Kirk	437883	2025-09-13
9	-	297713	2025-09-13
10	Wikipedia:Featured_pictures	244657	2025-09-13

Thank You!

Questions?

SADA™
An Insight company