CS 329E: Team1

Francis Tang and Tarun Chinta
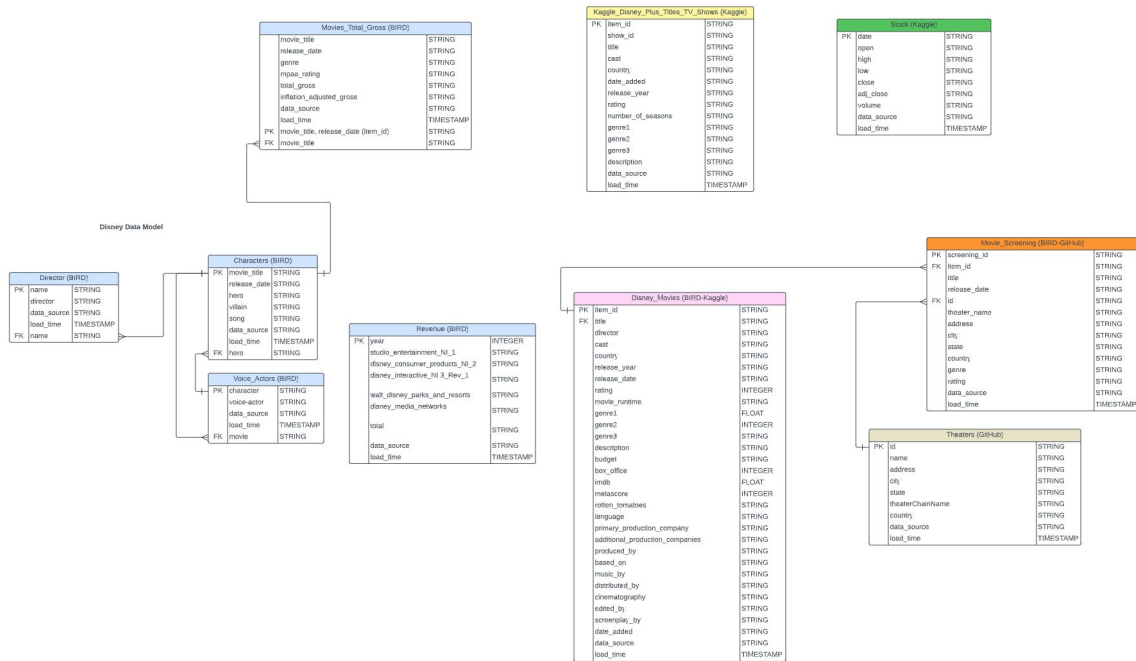
Table of Contents

1. Final ERD
2. Project Overview
3. Small Challenges
4. Airflow Challenges
5. Live Demo

ERD for Final Project

Movies_Total_Gross (BIRD)

	movie_title	STRING
	release_date	STRING
	genre	STRING
	mpaa_rating	STRING
	total_gross	STRING
	inflation_adjusted_gross	STRING
	data_source	STRING
	load_time	TIMESTAMP
PK	movie_title, release_date (item_id)	STRING
FK	movie_title	STRING

Kaggle_Disney_Plus_Titles_TV_Shows (Kaggle)

PK	item_id	STRING
	show_id	STRING
	title	STRING
	cast	STRING
	country	STRING
	date_added	STRING
	release_year	STRING
	rating	STRING
	number_of_seasons	STRING
	genre1	STRING
	genre2	STRING
	genre3	STRING
	description	STRING
	data_source	STRING
	load_time	TIMESTAMP

Stock (Kaggle)

PK	date	STRING
	open	STRING
	high	STRING
	low	STRING
	close	STRING
	adj_close	STRING
	volume	STRING
	data_source	STRING
	load_time	TIMESTAMP

Disney Data Model

Director (BIRD)

PK	name	STRING
	director	STRING
	data_source	STRING
	load_time	TIMESTAMP
FK	name	STRING

Characters (BIRD)

PK	movie_title	STRING
	release_date	STRING
	hero	STRING
	villain	STRING
	song	STRING
	data_source	STRING
	load_time	TIMESTAMP
FK	hero	STRING

Voice_Actors (BIRD)

PK	character	STRING
	voice-actor	STRING
	data_source	STRING
	load_time	TIMESTAMP
FK	movie	STRING

Revenue (BIRD)

PK	year	INTEGER
	studio_entertainment_NI_1	STRING
	disney_consumer_products_NI_2	STRING
	disney_interactive_NI 3_Rev_1	STRING
	walt_disney_parks_and_resorts	STRING
	disney_media_networks	STRING
	total	STRING
	data_source	STRING
	load_time	TIMESTAMP

Disney_Movies (BIRD-Kaggle)

PK	item_id	STRING
FK	title	STRING
	director	STRING
	cast	STRING
	country	STRING
	release_year	STRING
	release_date	STRING
	rating	INTEGER
	movie_runtime	FLOAT
	genre1	INTEGER
	genre2	STRING
	genre3	STRING
	description	STRING
	budget	INTEGER
	box_office	FLOAT
	imdb	INTEGER
	metascore	INTEGER
	rotten_tomatoes	STRING
	language	STRING
	primary_production_company	STRING
	additional_production_companies	STRING
	produced_by	STRING
	based_on	STRING
	music_by	STRING
	distributed_by	STRING
	cinematography	STRING
	edited_by	STRING
	screenplay_by	STRING
	date_added	STRING
	data_source	STRING
	load_time	TIMESTAMP

Movie_Screening (BIRD-GitHub)

PK	screening_id	STRING
FK	item_id	STRING
	title	STRING
	release_date	STRING
FK	id	STRING
	theater_name	STRING
	address	STRING
	city	STRING
	state	STRING
	country	STRING
	genre	STRING
	rating	STRING
	data_source	STRING
	load_time	TIMESTAMP

Theaters (GitHub)

PK	id	STRING
	name	STRING
	address	STRING
	city	STRING
	state	STRING
	theaterChainName	STRING
	country	STRING
	data_source	STRING
	load_time	TIMESTAMP

Option 2: Data Enrichment

- 7 columns to be enriched using the LLM
- Disney Movies: imdb, metascore, rotten_tomatoes, release_date, budget, box_office
- Movies Total Gross: mpaa_rating
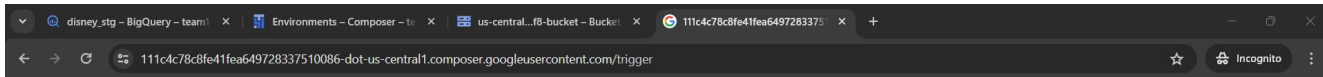- Enrichment scenarios implemented in BigQuery and ported to Airflow

Small Challenges Faced

Issues	Solutions
Type casting and formatting	Added CAST statements to cast values to proper type. Used REPLACE statements to remove unnecessary characters.
"Not found in region" error	Created multiple remote models and vertex connections in order to cover all the relevant regions.
Payment account	Ran out of credits

Airflow Challenges

- One controller did not run successfully - break down to multiple controllers

- Ran slowly - possibly due to being out of credits

- Code that worked on BigQuery runs into 'this dataset does not exist in us-central1' errors on Airflow - created larger composer environments

Ooops!

```
Something bad has happened.

Airflow is used by many users, and it is very likely that others had similar problems and you can easily find
a solution to your problem.

Consider following these steps:

  * gather the relevant information (detailed logs with errors, reproduction steps, details of your deployment)

  * find similar issues using:
    * GitHub Discussions
    * GitHub Issues
    * Stack Overflow
    * the usual search engine you use on a daily basis

  * if you run Airflow on a Managed Service, consider opening an issue using the service support channels

  * if you tried and have difficulty with diagnosing and fixing the problem yourself, consider creating a bug report.
    Make sure however, to include all relevant details and results of your investigation so far.

Python version: 3.11.5
Airflow version: 2.6.3+composer
Node: redact
-------------------------------------------------------------------------------
Error! Please contact server admin.
```

Composer   ←   Environment d...   🔗 OPEN AIRFLOW UI   📁 OPEN DAGS FOLDER   🗔 SAVE SNAPSHOT   ⬇ LOAD SNAPSHOT   ↻ REFRESH   🗑 DELETE   🎓 LEARN

Get started with Cloud Composer

ℹ This environment cannot be edited due to the errors that occurred during environment creation/update. Please investigate the logs to determine the cause, or create a new environment.

❗ CREATE operation on this environment failed 3 minutes ago with the following error message:
Your environment could not complete its creation process because it could not successfully initialize the Airflow database. This can happen when the GKE cluster is unable to reach the SQL database over the network.

The issue may be caused by missing IAM roles in the following Service Accounts:
 - service-341419945798@cloudcomposer-accounts.iam.gserviceaccount.com in project 341419945798 is missing role roles/composer.ServiceAgentV2Ext

The list of missing roles is generated without checking individual permissions in IAM custom roles. If any of the Service Accounts above uses custom IAM roles, its permissions may be sufficient and a corresponding warning may be ignored.

LEARN MORE ABOUT TROUBLESHOOTING ENVIRONMENT CREATION 🔗

❗ disney2   This environment has errors

MONITORING | LOGS | DAGS | ENVIRONMENT CONFIGURATION | AIRFLOW CONFIGURATION OVERRIDES | ENVIRONMENT VARIABLES | LABELS | PYPI PACKAGES

6 HOURS ▾   🕐 🔍 |◁   View Alerting Policies

Overview

DAG Statistics

Schedulers

Workers

Environment overview

Environment health (Airflow Monitoring DAG) ❓

Scheduler heartbeat ❓

Web server health ❓

Manage DAGs

📋 Add and update DAGs 🔗
Help document
Learn how to add and update DAGs in your environment.

📋 View DAGs, DAG runs, and tasks 🔗
Help document
Learn how to get information about DAGs and DAG runs your environment.

📋 Trigger DAGs 🔗
Help document
Learn different ways to trigger DAGs.

Monitor environments

📋 Use the monitoring dashboard 🔗
Help document
Learn how to use the monitoring dashboard.

📋 Monitor environment health and performance with key metrics 🔗
Help document
Monitor the overall Cloud Composer environment health and performance with key metrics in the monitoring dashboard.

📋 Monitor environments with Cloud Monitoring 🔗

❗ CREATE operation on this environment failed 3 minutes ago with the following error message:
CREATE operation failed. Could not trigger cluster creation: Insufficient regional quota to satisfy request: resource "CPUS": request requires '1.0' and is short '1.0'. project has a quota of '24.0' with '0.0' available. View and manage quotas at https://console.cloud.google.com/iam-admin/quotas?usage=USED&project=team1-final.

The issue may be caused by missing IAM roles in the following Service Accounts:
 - service-341419945798@cloudcomposer-accounts.iam.gserviceaccount.com in project 341419945798 is missing role roles/composer.ServiceAgentV2Ext

The list of missing roles is generated without checking individual permissions in IAM custom roles. If any of the Service Accounts above uses custom IAM roles, its permissions may be sufficient and a corresponding warning may be ignored.

LEARN MORE ABOUT TROUBLESHOOTING ENVIRONMENT CREATION 🔗

❗ disney4   This environment has errors

MONITORING | LOGS | DAGS | ENVIRONMENT CONFIGURATION | AIRFLOW CONFIGURATION OVERRIDES | ENVIRONMENT VARIABLES | LABELS | PYPI PACKAGES

6 HOURS ▾   🕐 🔍 |◁   View Alerting Policies

Overview

DAG Statistics

Schedulers

Workers

Environment overview

Environment health (Airflow Monitoring DAG) ❓

Scheduler heartbeat ❓

Web server health ❓

Let's look at the Demo!