

DATA 226 Assignment 7

Question 1: Import **two tables in your SnowflakeLinks to an external site.** as an ETL DAG in your Airflow

The screenshot shows the Airflow web interface with the DAG list view. The top navigation bar includes the Airflow logo and links to DAGs, Cluster Activity, Datasets, Browse, Admin, Docs, and Composer. The current time is 23:54 UTC. The DAG list table has columns for DAG, Owner, Runs, Schedule, Last Run, Next Run, Recent Tasks, Actions, and Links. Two DAGs are listed: 'airflow_monitoring' (owner: airflow) and 'import_snowflake_tables' (owner: Lab1). The 'import_snowflake_tables' DAG is selected, and its details are shown below the table. The details view shows the DAG's schedule as '*/10 * * * *', its last run as '2024-10-23, 23:40:00', and its next run as '2024-10-23, 23:50:00'. The recent tasks section shows a single task 'load_data' in a 'success' state.

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
airflow_monitoring	airflow	1	*/10 * * * *	2024-10-23, 23:40:00	2024-10-23, 23:50:00	load_data	[Play] [Stop]	...
import_snowflake_tables	Lab1	1	None	2024-10-23, 23:50:26		load_data	[Play] [Stop]	...

The screenshot shows the Airflow web interface with the DAG details view for 'import_snowflake_tables'. The top navigation bar includes the Airflow logo and links to DAGs, Cluster Activity, Datasets, Browse, Admin, Docs, and Composer. The current time is 02:11 UTC. The DAG details view shows the DAG's schedule as 'None', its next run ID as 'None', and its last run as '2024-10-23, 23:50:26 UTC'. The task graph is displayed in the center, showing three tasks: 'create_tables', 'set_stage', and 'load_data'. The 'load_data' task is highlighted in blue, indicating it is the current task. The task graph shows the following flow: 'create_tables' (success) -> 'set_stage' (success) -> 'load_data' (success). The task details for 'load_data' are shown on the right, including its state as 'success' and its duration as '00:00:00'.

Version: 2.9.3+composer
Environment Name: airflow2

DAG: import_snowflake_tables

Schedule: None | Next Run ID: None

10/24/2024 | 02:10:55 AM | 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

Clear task | Mark state as... | Filter DAG by task

import_snowflake_tables | Run | Task | 2024-10-23, 23:50:26 UTC / load_data

Details | Graph | Gantt | Code | Audit Log | Logs | XCom | Task Duration

Layout: Left -> Right

create_tables | set_stage | load_data

success | success | success

@task | @task | @task

React Flow

Question 2:

Create an ELT DAG in your Airflow to create a JOINED table of the two

Airflow DAGs Cluster Activity Datasets Browse Admin Docs Composer

02:58 UTC S-

airflow2

All 3 Active 3 Paused 0 Running 0 Failed 0

Filter DAGs by tag Search DAGs Auto-refresh

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
airflow_monitoring	airflow	150	* / 10 * * * *	2024-10-24, 02:40:00	2024-10-24, 02:50:00	1		
elt_session_summary_dag	Lab1	1	None	2024-10-24, 02:57:27		5		
import_snowflake_tables	Lab1	1	None	2024-10-23, 23:50:26		3		

Showing 1-3 of 3 DAGs

Version: 2.9.3+composer Environment Name: airflow2

Airflow DAGs Cluster Activity Datasets Browse Admin Docs Composer

02:58 UTC S-

DAG: elt_session_summary_dag

Schedule: None Next Run ID: None

10/24/2024 02:58:23 AM 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

elt_session_summary_dag / 2024-10-24, 02:57:27 UTC / create_session_summary

Clear task Mark state as... Filter DAG by task

Details Graph Gantt Code Audit Log Logs XCom Task Duration

Layout: Left -> Right

create_tables success @task

set_stage success @task

load_data success @task

create_session_summary success @task

React Flow

Snowflake:

The screenshot shows the Snowflake SQL Editor interface. On the left, a sidebar displays a tree view of the database structure, including schemas like ALPHA_VABTAGE_STOCKS, DEV, ANALYTICS, INFORMATION_SCHEMA, PUBLIC, RAW_DATA, and SNOWFLAKE. The main editor area displays a SQL script for creating a warehouse, database, schema, and tables. The script is as follows:

```
1 CREATE OR REPLACE WAREHOUSE lab1;
2 USE WAREHOUSE lab1;
3
4 CREATE OR REPLACE DATABASE dev;
5 USE DATABASE dev;
6
7 CREATE OR REPLACE SCHEMA raw_data;
8 USE SCHEMA raw_data;
9
10 CREATE OR REPLACE SCHEMA analytics;
11 USE SCHEMA analytics;
12
13 CREATE TABLE IF NOT EXISTS dev.raw_data.user_session_channel (
14     userId int not NULL,
15     sessionId varchar(32) primary key,
16     channel varchar(32) default 'direct'
17 );
18
19 CREATE TABLE IF NOT EXISTS dev.raw_data.session_timestamp (
20     sessionId varchar(32) primary key,
21     ts timestamp
22 );
23
24 CREATE OR REPLACE STAGE dev.raw_data.blob_stage
25     url = 's3://s3-geospatial/readonly/'
26     file_format = (type = csv, skip_header = 1, field_optionally_enclosed_by = '');
27
28
29
30
31
32
33
34
35
```

The screenshot shows the Snowflake Data Preview interface for the table `DEV / ANALYTICS / SESSION_SUMMARY`. The table is owned by `ACCOUNTADMIN` and was updated just now. It contains 101.5K rows and is 3.9MB in size. The table details are as follows:

	USERID	SESSIONID	CHANNEL	TS
1	779	7cdace91c487558e27ce54df7cdb299c	Instagram	2019-05-01T00:13:11.783Z
2	230	94f192dee568b018e0acf31e1f99a2d9	Naver	2019-05-01T00:49:46.073Z
3	369	7ed2d3454c5eas71148b11d0c25104ff	Youtube	2019-05-01T10:18:43.212Z
4	248	f1daf122cde863010844459363cd31db	Naver	2019-05-01T13:10:56.413Z
5	676	fd0efcca272f704a760c3b61dcc70fd0	Instagram	2019-05-01T13:45:19.793Z
6	40	8804f94e16ba5b680e239a55a08f7d2	Youtube	2019-05-01T14:23:07.66Z
7	468	c5f441cd5f43eb2f2c024e1f8b5d00cd	Instagram	2019-05-01T15:03:54.65Z
8	69	d5fcc35c94879a4afad61cacca56192c	Facebook	2019-05-01T15:13:16.14Z
9	420	3d191ef6e236bd1b9bdb9ff4743c47fe	Youtube	2019-05-01T15:33:58.197Z
10	572	c17028c9b6e0c5deaad29665d582284a	Organic	2019-05-01T15:59:57.49Z
11	253	cd0b43eac0392accf3624b7372dec36e	Facebook	2019-05-01T16:33:03.463Z
12	87	0a4bbceda17a6253386bc9eb45240e25	Youtube	2019-05-01T17:10:29.417Z
13	277	c67ba7c4c5c0cd4cc3e3a7146fe5c015	Naver	2019-05-01T17:42:13.58Z
14	97	d8a3a3c3234392b0add43c5f9c05a246	Organic	2019-05-01T18:18:00.407Z
15	345	63dfdeb1ff9ff09ecc3f05d2d7221ffa	Google	2019-05-01T18:35:48.62Z
16	187	dfbfa7ddcfffefb581f50edcf9a0204bb	Organic	2019-05-01T19:37:45.037Z
17	788	8d3215ae97598264ad6529613774a038	Organic	2019-05-01T21:21:12.953Z
18	757	636efd4f9aeb5781e9ea815cdd633e52	Naver	2019-05-01T22:24:04.87Z
19	765	4c4c937b67cc8d785cea1e42ccea185c	Organic	2019-05-01T23:50:38Z

https://app.snowflake.com/nsmusq/ba86308/#/data/databases/DEV/schemas/ANALYTICS/table/SESSION_SUMMARY/data-preview

Question 3:

Set up your Preset account or [Docker Superset environment](#)

Used Preset

Superset interface showing the Datasets tab. The top navigation bar includes Home, Dashboards, Charts, Datasets (active), and SQL. The Datasets tab displays a table of datasets with columns: Name, Type, Database, Schema, Owners, Last modified, and Actions. A dataset named 'session_summary' is listed with Type 'Physical', Database 'Snowflake', Schema 'analytics', and Last modified '40 minutes ago'.

Name	Type	Database	Schema	Owners	Last modified	Actions
session_summary	Physical	Snowflake	analytics	SG	40 minutes ago	

Question 4:

Create your WAU chart

Superset interface showing the Charts tab. The top navigation bar includes Home, Dashboards, Charts (active), Datasets, and SQL. The Charts tab displays a line chart titled 'Add the name of the chart'. The chart shows the Weekly Active Users (WAU) over time, with the Y-axis ranging from 0 to 700 and the X-axis showing months from May to November. The chart is configured with the following settings:

- Chart Source: analytics.session_summary
- Metrics: f(x) COUNT(*)
- Columns: TS, USERID, SESSIONID, CHANNEL
- Dimensions: None
- Contribution Mode: None
- Filters: TS (No filter)
- Series Limit: None
- Sort By: None
- Row Limit: 10000
- Truncate Metric: ☒
- Show Empty Columns: ☒

The chart displays a line series for WAU, showing a steady increase from May to November, peaking around 700 in October. The chart is titled 'Add the name of the chart' and includes a 'SAVE' button. The chart is also labeled '31 rows' and '00:00:02.19'.