

CS 639: Data Management for Data Science - P4

Badri Narayanan Murali Krishnan
Net ID: bmuralikrish

April 5, 2025

Section 1: Setup environment (15 points in total)

Q1: Register Airbyte Account (0.5 points)

Answer

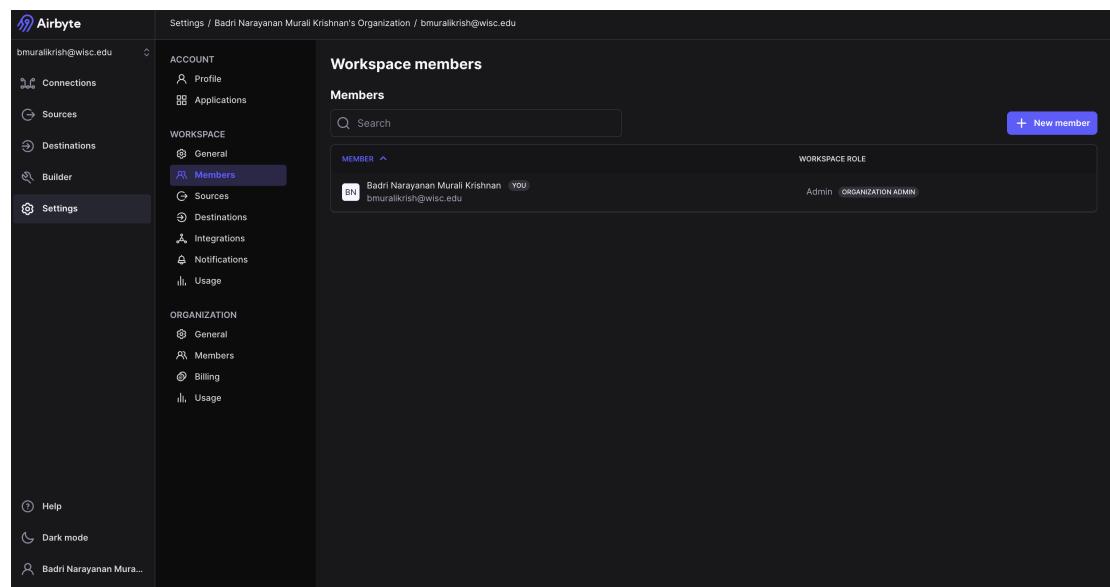
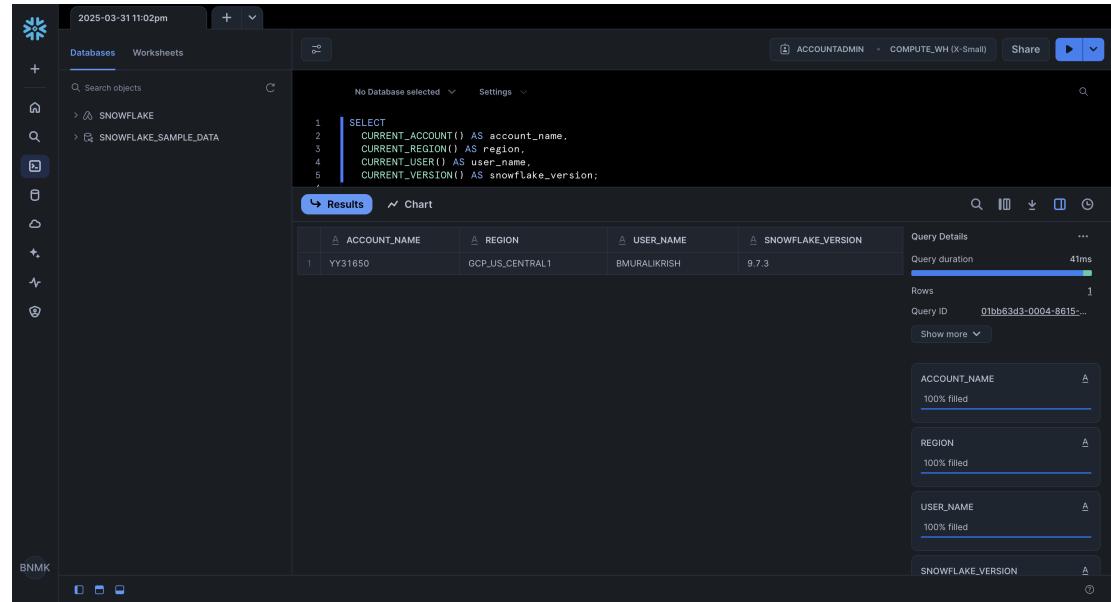


Figure 1: Registered Airbyte account.

Q2: Register Snowflake Account (0.5 points)

Answer



The screenshot shows the Snowflake UI interface. On the left, there's a sidebar with icons for databases, worksheets, and other database objects. The main area has tabs for 'Databases' and 'Worksheets'. A search bar at the top says 'No Database selected'. Below it is a query editor with the following SQL code:

```
1 SELECT
2     CURRENT_ACCOUNT() AS account_name,
3     CURRENT_REGION() AS region,
4     CURRENT_USER() AS user_name,
5     CURRENT_VERSION() AS snowflake_version;
```

The results pane shows a single row of data:

ACCOUNT_NAME	REGION	USER_NAME	SNOWFLAKE_VERSION
YY31650	GCP.US_CENTRAL1	BMURALIKRISH	9.7.3

On the right, there's a 'Query Details' panel with information like 'Query duration: 41ms', 'Rows: 1', and 'Query ID: 01bb63d3-0004-8615...'. Below the results table, there are four small cards showing the values for each column: ACCOUNT_NAME (100% filled), REGION (100% filled), USER_NAME (100% filled), and SNOWFLAKE_VERSION (100% filled).

Figure 2: Registered Snowflake account.

Q3: Install DBT using Miniconda (0.5 points)

Answer

Solution has been submitted as part of p4.ipynb.

Q4: Snowflake configuration (6.5 points - 0.5 point for each step)

Answer

```
1 -- 1. Switch role
2 USE ROLE accountadmin;
3
4 -- 2. Create warehouse
5 CREATE WAREHOUSE IF NOT EXISTS P4_WAREHOUSE
6     WITH WAREHOUSE_SIZE = 'XSMALL'
7     AUTO_SUSPEND = 60
8     AUTO_RESUME = TRUE
9     INITIALLY_SUSPENDED = TRUE;
```

```

11   -- 3. Create databases
12   CREATE DATABASE IF NOT EXISTS SURVEY_DATABASE;
13   CREATE DATABASE IF NOT EXISTS STOCK_DB;
14
15   -- 4. Create role and grant to sysadmin
16   CREATE ROLE IF NOT EXISTS P4_ROLE;
17   GRANT ROLE P4_ROLE TO ROLE SYSADMIN;
18
19   -- 5. Grant usage on warehouse
20   GRANT USAGE ON WAREHOUSE P4_WAREHOUSE TO ROLE P4_ROLE;
21
22   -- 6. Grant role to your user
23   GRANT ROLE P4_ROLE TO USER BMURALIKRISH;
24
25   -- 7. Grant ownership of DBs
26   GRANT OWNERSHIP ON DATABASE SURVEY_DATABASE TO ROLE P4_ROLE;
27   GRANT OWNERSHIP ON DATABASE STOCK_DB TO ROLE P4_ROLE;
28
29   -- 8. Switch to P4_ROLE
30   USE ROLE P4_ROLE;
31
32   -- 9. Create schemas
33   USE DATABASE SURVEY_DATABASE;
34   CREATE SCHEMA IF NOT EXISTS SURVEY_SCHEMA;
35
36   USE DATABASE STOCK_DB;
37   CREATE SCHEMA IF NOT EXISTS STOCK_SCHEMA;
38
39   -- 11. Grant ownership on schemas
40   GRANT OWNERSHIP ON SCHEMA SURVEY_DATABASE.SURVEY_SCHEMA TO
41     → ROLE P4_ROLE;
42   GRANT OWNERSHIP ON SCHEMA STOCK_DB.STOCK_SCHEMA TO ROLE
43     → P4_ROLE;

```

Q5: Airbyte configuration for survey (1.5 points)

Answer

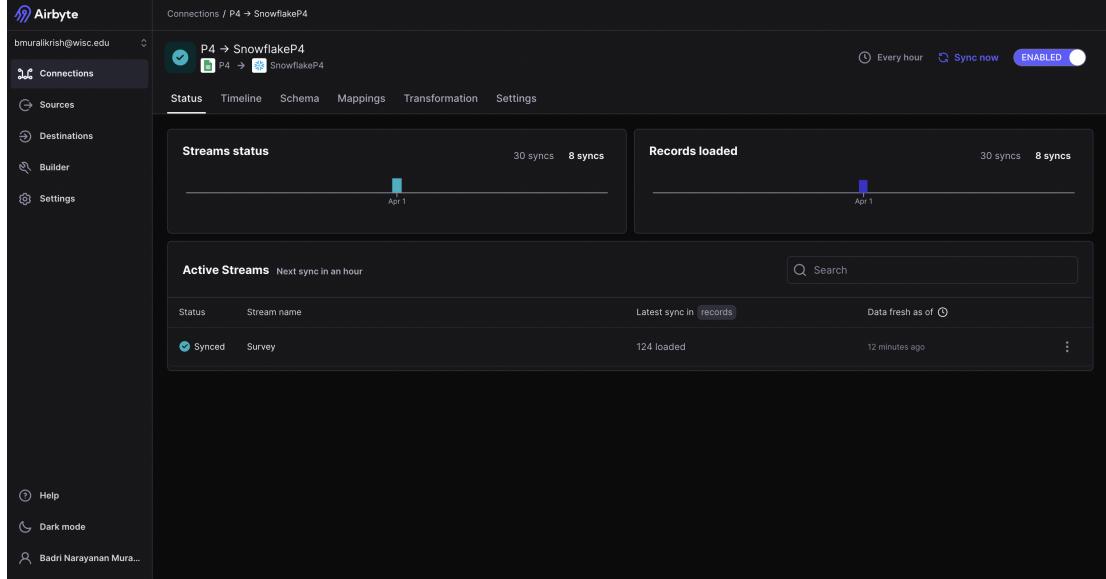
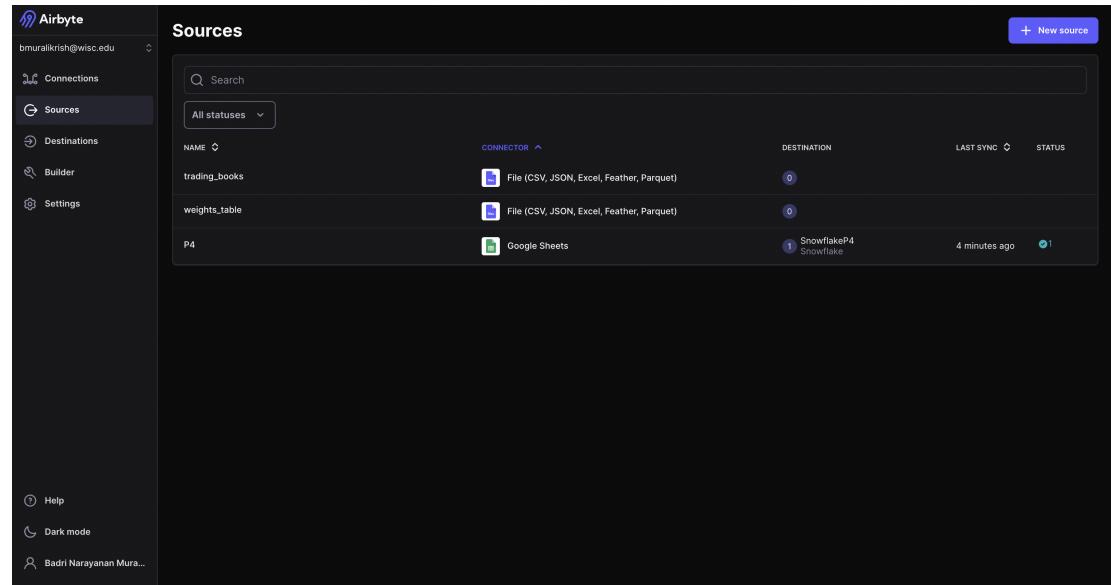


Figure 3: Airbyte configuration for the survey.

Q6: Airbyte configuration for trading_books.csv and weights_table.csv files (0.5 points)

Answer



The screenshot shows the Airbyte web application interface. On the left is a dark sidebar with navigation links: Connections, Sources (which is highlighted), Destinations, Builder, Settings, Help, Dark mode, and a user profile for Badri Narayanan Mura... At the top right is a blue button labeled '+ New source'. The main area is titled 'Sources' and contains a search bar and a dropdown menu set to 'All statuses'. Below this is a table with three rows:

NAME	CONNECTOR	DESTINATION	LAST SYNC	STATUS
trading_books	File (CSV, JSON, Excel, Feather, Parquet)	0		
weights_table	File (CSV, JSON, Excel, Feather, Parquet)	0		
P4	Google Sheets	1 SnowflakeP4 Snowflake	4 minutes ago	1

Figure 4: Airbyte configuration for trading_books.csv and weights_table.csv files.

Q7: Connect Airbyte to Snowflake to load trading_books and weights_table (0.5 points)

Answer

 Airbyte

bmuralikrish@wisc.edu

 Connections

 Sources

 Destinations

 Builder

 Settings

 Help

 Dark mode

 Badri Narayanan Mura...

Destinations

+ New destination

Search

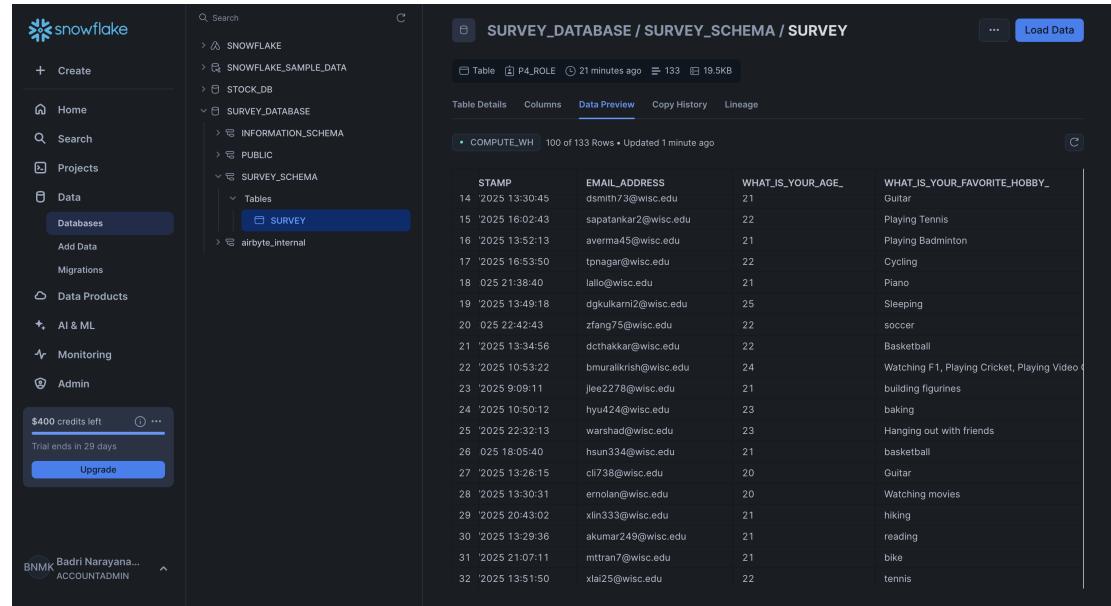
All statuses ▾

NAME	CONNECTOR	SOURCE	LAST SYNC	STATUS
SnowflakeP4	 Snowflake	P4 Google Sheets	35 minutes ago	 1
Snowflake	 Snowflake	2 sources File (CSV, JSON, Excel, Fasta...)	35 seconds ago	 1

Figure 5: Airbyte to Snowflake connection to load trading_books table and weights_table.

Q8: Data Preview on Snowflake (0.5 points)

Answer



The screenshot shows the Snowflake web interface. On the left, the sidebar navigation includes Home, Search, Projects, Data (with Databases selected), Add Data, Migrations, Data Products, AI & ML, Monitoring, and Admin. A message at the bottom indicates '\$400 credits left' and 'Trial ends in 29 days'. The main panel displays the 'SURVEY_DATABASE / SURVEY_SCHEMA / SURVEY' table. The table has columns: STAMP, EMAIL_ADDRESS, WHAT_IS_YOUR_AGE_, and WHAT_IS_YOUR_FAVORITE_HOBBY_. The preview shows 100 of 133 rows. The first few rows of data are:

STAMP	EMAIL_ADDRESS	WHAT_IS_YOUR_AGE_	WHAT_IS_YOUR_FAVORITE_HOBBY_
14 '2025 13:30:45	dsmith73@wisc.edu	21	Guitar
15 '2025 18:02:43	sapatankar2@wisc.edu	22	Playing Tennis
16 '2025 13:52:13	averma45@wisc.edu	21	Playing Badminton
17 '2025 16:53:50	tpnagar@wisc.edu	22	Cycling
18 025 21:38:40	lalli@wisc.edu	21	Piano
19 '2025 13:49:18	dgkulkarni2@wisc.edu	25	Sleeping
20 025 22:42:43	zfang75@wisc.edu	22	soccer
21 '2025 13:34:56	dcthakkar@wisc.edu	22	Basketball
22 '2025 10:53:22	bmuralikrish@wisc.edu	24	Watching F1, Playing Cricket, Playing Video Games
23 '2025 9:09:11	jlee2278@wisc.edu	21	building figurines
24 '2025 10:50:12	hyu424@wisc.edu	23	baking
25 '2025 22:32:13	warshad@wisc.edu	23	Hanging out with friends
26 025 18:05:40	hsun34@wisc.edu	21	basketball
27 '2025 13:26:15	cl738@wisc.edu	20	Guitar
28 '2025 13:30:31	ernolari@wisc.edu	20	Watching movies
29 '2025 20:43:02	xlin333@wisc.edu	21	hiking
30 '2025 13:29:36	akumar249@wisc.edu	21	reading
31 '2025 21:07:11	mttran7@wisc.edu	21	bike
32 '2025 13:51:50	xlaiz5@wisc.edu	22	tennis

Figure 6: Data Preview from Snowflake for Survey Table.

The screenshot shows the Snowflake web interface. On the left, the navigation sidebar includes 'Create', 'Home', 'Search', 'Projects', 'Data' (selected), 'Databases', 'Add Data', 'Migrations', 'Data Products', 'AI & ML', 'Monitoring', and 'Admin'. A message at the bottom indicates '\$400 credits left' and 'Trial ends in 29 days' with a 'Upgrade' button. The main area displays the 'STOCK_DB / STOCK_SCHEMA / TRADING_BOOKS' table. The 'Data Preview' tab is selected, showing 30 rows of data. The columns are PRICE, TICKER, QUANTITY, TRADE_ID, TRADE_DATE, and TRADE_TYPE. The data includes various stock trades like AAPL, MSFT, and GOOGL.

	PRICE	TICKER	QUANTITY	TRADE_ID	TRADE_DATE	TRADE_TYPE
1	178.25	AAPL	100	1002	2024-03-01	SELL
2	1.085	EUR/USD	100000	1003	2024-03-01	BUY
3	1.092	EUR/USD	100000	1004	2024-03-01	SELL
4	405.75	MSFT	50	1005	2024-03-01	BUY
5	408.5	MSFT	50	1006	2024-03-01	SELL
6	1.265	GBP/USD	75000	1007	2024-03-01	BUY
7	1.272	GBP/USD	75000	1008	2024-03-01	SELL
8	141.25	GOOGL	75	1009	2024-03-01	BUY
9	142.5	GOOGL	75	1010	2024-03-01	SELL
10	178.25	AAPL	150	1011	2024-03-04	BUY
11	179.5	AAPL	150	1012	2024-03-04	SELL
12	1.088	EUR/USD	125000	1013	2024-03-04	BUY
13	1.095	EUR/USD	125000	1014	2024-03-04	SELL
14	407.25	MSFT	75	1015	2024-03-04	BUY
15	410.75	MSFT	75	1016	2024-03-04	SELL
16	1.268	GBP/USD	100000	1017	2024-03-04	BUY
17	1.275	GBP/USD	100000	1018	2024-03-04	SELL
18	142.25	GOOGL	100	1019	2024-03-04	BUY
19						

Figure 7: Data Preview from Snowflake for Trading Books Table.

The screenshot shows the Snowflake web interface. The left sidebar is identical to Figure 7. The main area displays the 'STOCK_DB / STOCK_SCHEMA / WEIGHTS_TABLE' table. The 'Data Preview' tab is selected, showing 2 rows of data. The columns are _AIRBYTE_GENERATION_ID, DESK, REGION, and TARGET_ALLOCATION. The data shows two entries for id: 32344071.

	_AIRBYTE_GENERATION_ID	DESK	REGION	TARGET_ALLOCATION	
1	{ "id": 32344071 }	1	FX Desk	Europe	0.25
2	{ "id": 32344071 }	1	Equity Desk	North America	0.75

Figure 8: Data Preview from Snowflake for Weights Table.

Q9: Use dbt to transform the column names to single word or two-word names with “_” separator and create a view named survey_transformed. (4 points)

Answer

Solution has been submitted as part of p4.ipynb.

The screenshot shows the Snowflake web interface. On the left, the navigation sidebar is visible with sections like Home, Search, Projects, Data (selected), Databases, Add Data, Migrations, Data Products, AI & ML, Monitoring, and Admin. A message at the bottom indicates '\$397 credits left' and 'Trial ends in 29 days'. The user is identified as 'BNMK Badri Narayana... ACCOUNTADMIN'. The main panel shows a database named 'SURVEY_DATABASE / SURVEY_SCHEMA / TRANSFORM_SURVEY'. The 'Data Preview' tab is selected, displaying a table with columns: 'TIMESTAMP', 'EMAIL_ADDRESS', 'AGE', and 'FAVORITE_Hobby'. The data preview shows 55 rows of transformed survey data. The first few rows are:

	TIMESTAMP	EMAIL_ADDRESS	AGE	FAVORITE_Hobby
38	4/2/2025 13:38:14	srgarlapati@wisc.edu	20	Reading
39	3/30/2025 20:01:34	meklia@wisc.edu	22	driving
40	3/21/2025 13:50:59	srathi3@wisc.edu	21	Baking and cooking
41	3/19/2025 13:29:07	pctrophy@wisc.edu	22	Rock Climbing
42	3/21/2025 14:54:32	qxiang8@wisc.edu	22	eat
43	3/19/2025 20:25:44	zwei232@wisc.edu	19	Travel
44	3/31/2025 13:15:05	mducha@wisc.edu	20	playing video games
45	3/19/2025 13:29:59	dkim756@wisc.edu	19	Listening to music
46	3/19/2025 13:30:55	zchehime@wisc.edu	21	Playing video games with my friends online
47	3/19/2025 13:30:08	vphilavong@wisc.edu	20	Video Games
48	3/21/2025 13:50:15	mdhanif@wisc.edu	21	Playing video games
49	3/23/2025 13:20:03	ychaudhari@wisc.edu	25	Playing CS2 and Soccer
50	3/21/2025 13:50:56	binnmohdizanji@wisc.edu	21	Sports
51	4/2/2025 13:27:15	aldosery@wisc.edu	21	working out
52	3/19/2025 13:25:50	zhou658@wisc.edu	20	Travel
53	3/30/2025 12:58:31	niekro@wisc.edu	19	Baking
54	3/22/2025 17:09:04	bhlee8@wisc.edu	23	Eating
55	3/30/2025 14:02:12	atewart3@wisc.edu	21	Playing volleyball

Figure 9: Data Preview from Snowflake for the Transformed Survey Table.

Section 2: Custom student dataset analysis (7 points in total)

Q10: How many students are in each academic standing? (1 point)

Answer

The SQL query and the Pie chart have been submitted as part of p4.ipynb.

Q11: What is the average age of students in each academic standing? (1 point)

Answer

The SQL query and the Bar plot have been submitted as part of p4.ipynb.

Q12: What percentage of students would choose to sleep if they had an extra hour in their day? (1 point)

Answer

The SQL query has been submitted as part of p4.ipynb.

Q13: What is the academic standing and course selection factor of the student with email ‘bmuralikrish@wisc.edu’ (1 point)

Answer

The SQL query has been submitted as part of p4.ipynb.

Q14: How many students have a secondary major? (1 point)

Answer

The SQL query has been submitted as part of p4.ipynb.

Q15: What are the top 5 most common ZIP codes among students (1 point)

Answer

The SQL query has been submitted as part of p4.ipynb.

Q16: How do Night Owls prefer to spend an extra hour in their day? (1 point)

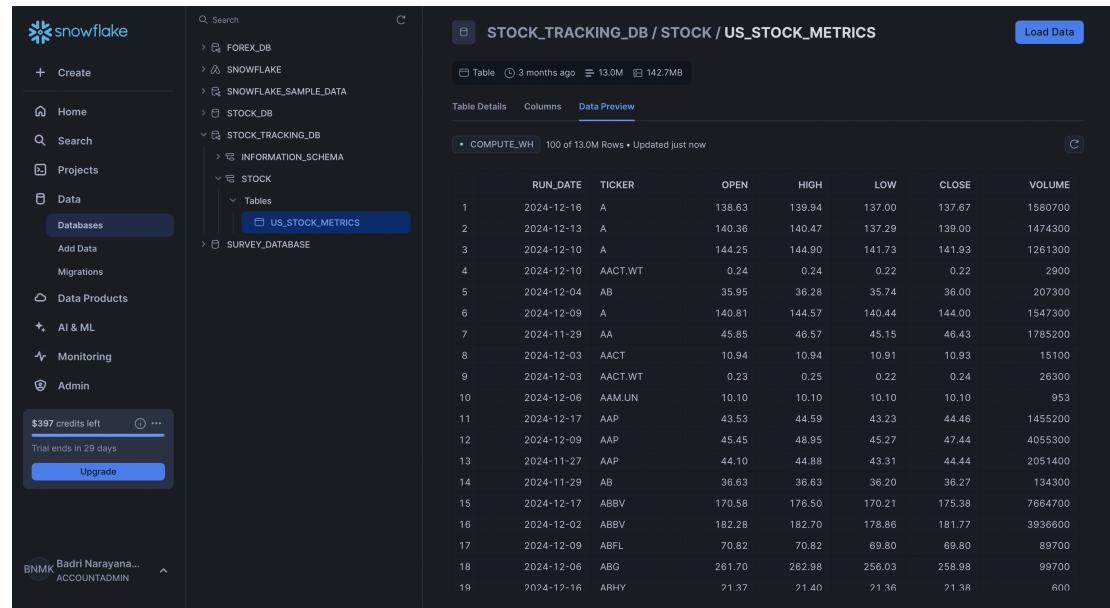
Answer

The SQL query has been submitted as part of p4.ipynb.

Section 3: Stock and foreign exchange (fx) dataset analysis (7 points in total)

Q17: Ingest data directly from Snowflake Marketplace (0.5 points)

Answer



The screenshot shows the Snowflake web interface. On the left, the sidebar navigation includes Home, Search, Projects, Data (with Databases selected), Add Data, Migrations, Data Products, AI & ML, Monitoring, Admin, and a trial status message (\$397 credits left, trial ends in 29 days). The main content area displays the 'STOCK_TRACKING_DB / STOCK / US_STOCK_METRICS' table. The table has 19 rows and 8 columns: RUN_DATE, TICKER, OPEN, HIGH, LOW, CLOSE, and VOLUME. The data preview shows various stock metrics for different dates and tickers like A, AB, AACT, AAP, etc.

	RUN_DATE	TICKER	OPEN	HIGH	LOW	CLOSE	VOLUME
1	2024-12-16	A	138.63	139.94	137.00	137.67	1580700
2	2024-12-13	A	140.36	140.47	137.29	139.00	1474300
3	2024-12-10	A	144.25	144.90	141.73	141.93	1261300
4	2024-12-10	AACT.WT	0.24	0.24	0.22	0.22	2900
5	2024-12-04	AB	35.95	36.28	35.74	36.00	207300
6	2024-12-09	A	140.81	144.57	140.44	144.00	1547300
7	2024-11-29	AA	45.85	46.57	45.15	46.43	1785200
8	2024-12-03	AACT	10.94	10.94	10.91	10.93	15100
9	2024-12-03	AACT.WT	0.23	0.25	0.22	0.24	26300
10	2024-12-06	AAM.UN	10.10	10.10	10.10	10.10	953
11	2024-12-17	AAP	43.53	44.59	43.23	44.46	1455200
12	2024-12-09	AAP	45.45	48.95	45.27	47.44	4055300
13	2024-11-27	AAP	44.10	44.88	43.31	44.44	2051400
14	2024-11-29	AB	36.63	36.63	36.20	36.27	134300
15	2024-12-17	ABBV	170.58	176.50	170.21	175.38	7664700
16	2024-12-02	ABBV	182.28	182.70	178.86	181.77	3936600
17	2024-12-09	ABFL	70.82	70.82	69.80	69.80	89700
18	2024-12-06	ABG	261.70	262.98	256.03	258.98	99700
19	2024-12-16	ARHY	21.37	21.40	21.36	21.38	600

Figure 10: Data Preview from Snowflake for US Stock Metrics Table.

	RUN_DATE	OPEN	HIGH	LOW	CLOSE	CURRENCY_PAIR_NAME
1	2025-01-08	281.85818	282.03400	281.75354	281.97794	AED/ARS
2	2025-01-08	0.43580	0.43595	0.43580	0.43817	AED/AUD
3	2025-01-08	0.54996	0.55294	0.54996	0.55032	AED/BBD
4	2025-01-08	0.10263	0.10275	0.10251	0.10272	AED/BHD
5	2025-01-08	1.66210	1.67589	1.66039	1.66384	AED/BRD
6	2025-01-08	0.39033	0.39198	0.39033	0.39148	AED/CAD
7	2025-01-08	0.24708	0.24848	0.24708	0.24804	AED/CHF
8	2025-01-08	273.53796	275.91394	272.69125	272.71576	AED/CLP
9	2025-01-08	1.99463	1.99821	1.99419	1.99504	AED/CNY
10	2025-01-08	1.96169	1.97700	1.96144	1.97035	AED/DKK
11	2025-01-08	0.26296	0.26496	0.26290	0.26406	AED/EUR
12	2025-01-08	4404.16553	4426.67041	4396.75977	4421.22266	AED/IDR
13	2025-01-08	0.99267	1.00188	0.99140	0.99829	AED/ILS
14	2025-01-08	38.20038	38.44829	38.19525	38.32017	AED/ISK
15	2025-01-08	42.60277	42.83421	42.60277	42.68482	AED/JMD
16	2025-01-08	0.19314	0.19363	0.19277	0.19313	AED/JOD
17	2025-01-08	35.25728	35.40974	35.12115	35.25728	AED/KES
18	2025-01-08	24495.40430	24523.66211	24394.22656	24394.22656	AED/LBP
19	2025-01-08	80.48778	80.92347	80.48778	80.67178	AED/IKR

Figure 11: Data Preview from Snowflake for FOREX Table.

Q18: Configure `/ .dbt/profiles.yml` for STOCK_DB (0.5 points)

Answer

Solution has been submitted as part of `p4.ipynb`.

Q19: Update `models/schema.yml` file to include new sources for STOCK_DB tables (0.5 points)

Answer

Solution has been submitted as part of `p4.ipynb`.

Q20: Create `dbt_project.yml` file in your dbt project directory — `p4_data_pipeline` (0.5 points)

Answer

Solution has been submitted as part of `p4.ipynb`.

Q21: Creating staging tables for FX and Stock (1 point)

Answer

Solution has been submitted as part of `p4.ipynb` and the required files have been added to the submission repository.

Q22: Create staging table for trading pairs (0.5 points)

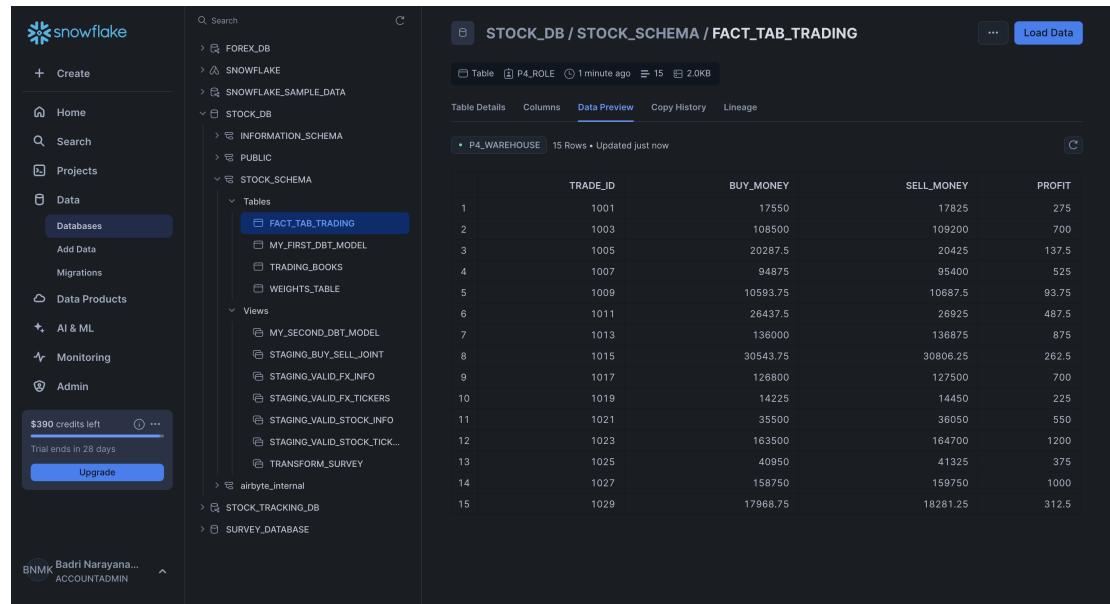
Answer

Solution has been submitted as part of p4.ipynb and the required file has been added to the submission repository.

Q23: Create a fact Table for trading data (1.5 points)

Answer

Solution has been submitted as part of p4.ipynb and the required file has been added to the submission repository.



	TRADE_ID	BUY MONEY	SELL MONEY	PROFIT
1	1001	17550	17825	275
2	1003	108500	109200	700
3	1005	20287.5	20425	137.5
4	1007	94875	95400	525
5	1009	10593.75	10687.5	93.75
6	1011	28437.5	28925	487.5
7	1013	136000	136875	875
8	1015	30543.75	30806.25	262.5
9	1017	126800	127500	700
10	1019	14225	14450	225
11	1021	35500	36050	550
12	1023	163500	164700	1200
13	1025	40950	41325	375
14	1027	158750	159750	1000
15	1029	17968.75	18281.25	312.5

Figure 12: Data Preview from Snowflake that shows the views and tables.

Q24: Compute total profit by desk (1 point)

Answer

Solution has been submitted as part of p4.ipynb and the required SQL query has been executed.

Q25: Compute profit rate by desk (1 point)

Answer

Solution has been submitted as part of p4.ipynb and the required SQL query has been executed.