Information Management
MIS 381N

ONLINE LEARNING BEHAVIOR & PATTERNS
WITH
SNOWFLAKE

Team ANS: Abhiroop Kumar, Nikhil Kumar, Simoni Dalal

UNIVERSITY OF TEXAS AT AUSTIN | MCCOMBS SCHOOL OF BUSINESS

Table of Contents

Brief Intro About Data

1. **Dataset Overview:** Kaggle dataset provides metadata on DataCamp courses and learning tracks, useful for analyzing learning patterns, course structures, and optimizing user learning journeys.

2. **Main Files:**
   - courses.csv: Core dataset with 23 columns (e.g., title, difficulty, technology, instructors).
   - all_tracks.csv: Overall info on learning tracks with 18 columns (track title, course count).

3. **Mapping Files:**
   - technology_mapping.csv: Maps technology IDs to names.
   - topic_mapping.csv: Maps topic IDs to subject areas.

Bronze Layer

Bronze Schema Creation

Created a BRONZE schema as the container for raw CSV data, preserving original data format with VARCHAR columns.

Data Loading Infrastructure

Established staging area and file format definitions to facilitate CSV file ingestion into the database.

Data Import Execution

Used COPY INTO commands to load data from staged CSV files into corresponding BRONZE tables efficiently.

Validation & Foundation

Checked for errors, row counts, and column structure to ensure data accuracy for next stages.

Silver Layer - Key Steps

Create Schema

Set up a new SILVER schema.

Design ERD

Used STAR/Snowflake architecture.

Build Tables

17 tables
[2 Fact, 10 Dimension, 4 Bridge, 1 Audit]
+ 13 sequences

Surrogate Keys

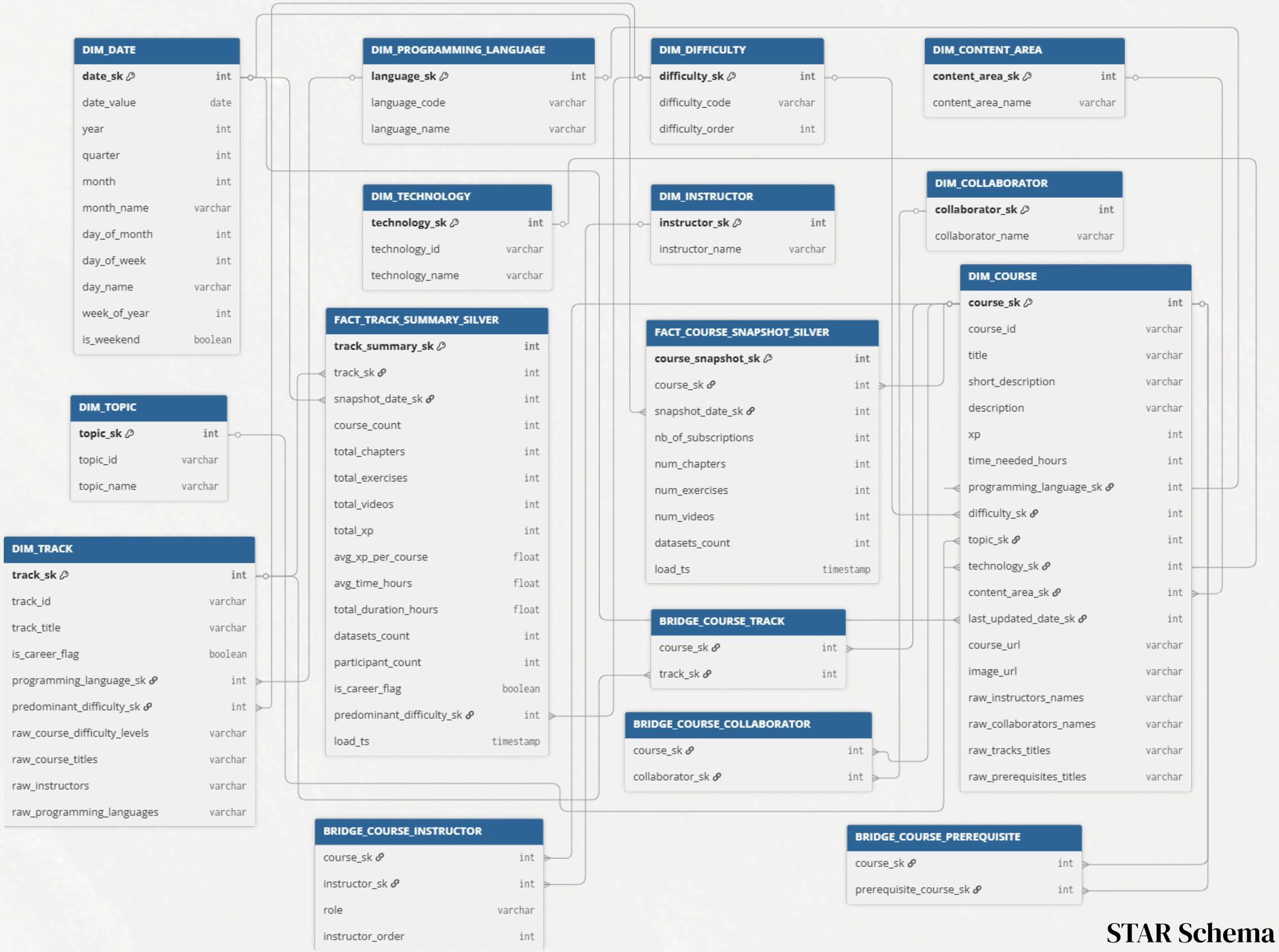
Replaced Bronze PKs with SKs for ID differentiation.

Populate & Transform

Inserted data using queries and functions (e.g., TO_DATE, TRIM, LOWER).

Validate

Ran checks for clean, accurate data insertion.



STAR Schema - Silver Layer

Gold Layer

Created a new GOLD schema with 3 dynamic tables to answer key business questions, ensuring near real-time data freshness from SILVER. Added an Audit Table for log tracking.

Use Case	Business Question	Required Silver Tables
1 - Programming Language Instructional Effort	Identify which programming languages require the most instructional effort (time, chapters, exercises, videos) across all courses.	FACT_COURSE_SNAPSHOT_SILVER, DIM_PROGRAMMING_LANGUAGE, DIM_DIFFICULTY, DIM_COURSE.
2 - Track-Level Summary	For each career or skill track, calculate total learning content (time, chapters, exercises, videos) and number of courses per track.	BRIDGE_COURSE_TRACK, DIM_TRACK, FACT_COURSE_SNAPSHOT_SILVER, DIM_COURSE.
3 - Course Difficulty Distribution & Curriculum Depth	Analyze course catalog distribution across difficulty levels (Beginner, Intermediate, Advanced) and identify which level contributes the most learning content (chapters, exercises, videos).	FACT_COURSE_SNAPSHOT_SILVER, DIM_COURSE, DIM_DIFFICULTY.

Streamlit Dashboard

Our approach blends strategic insight, creative execution, and data-driven optimization - tailored specifically for growing businesses.

Key Steps

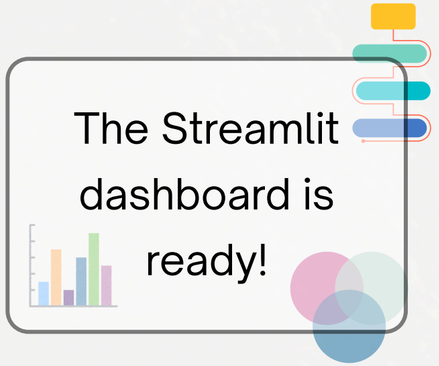
Create a new stage in GOLD schema

Create a Streamlit python script that visualizes Gold table result-set data

Upload the Python file to the new stage

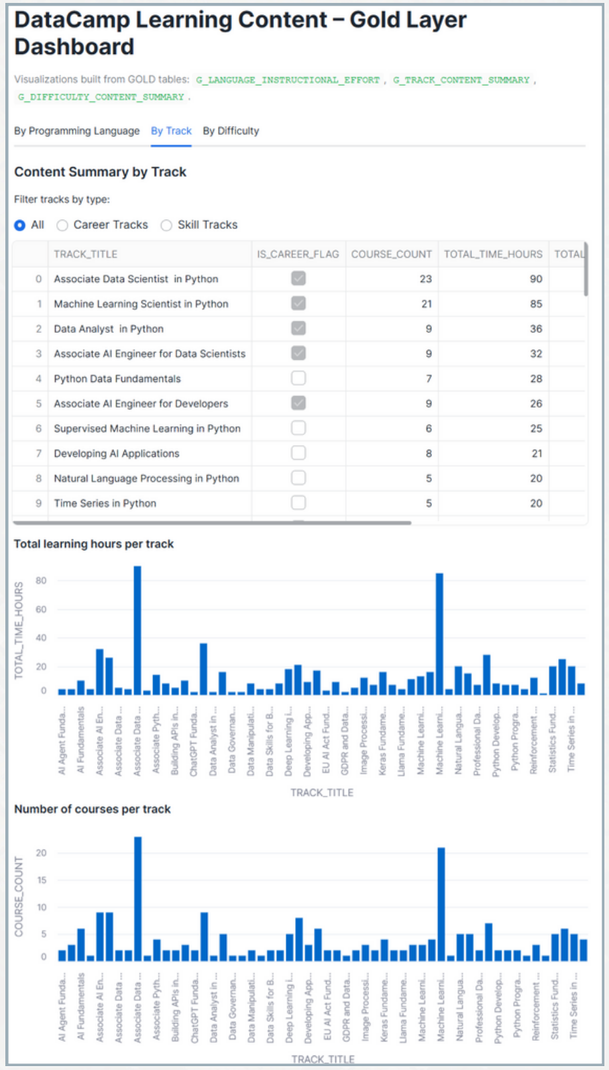
In SnowSQL, create a Streamlit object pointing to the stage & script

The Streamlit dashboard is ready!

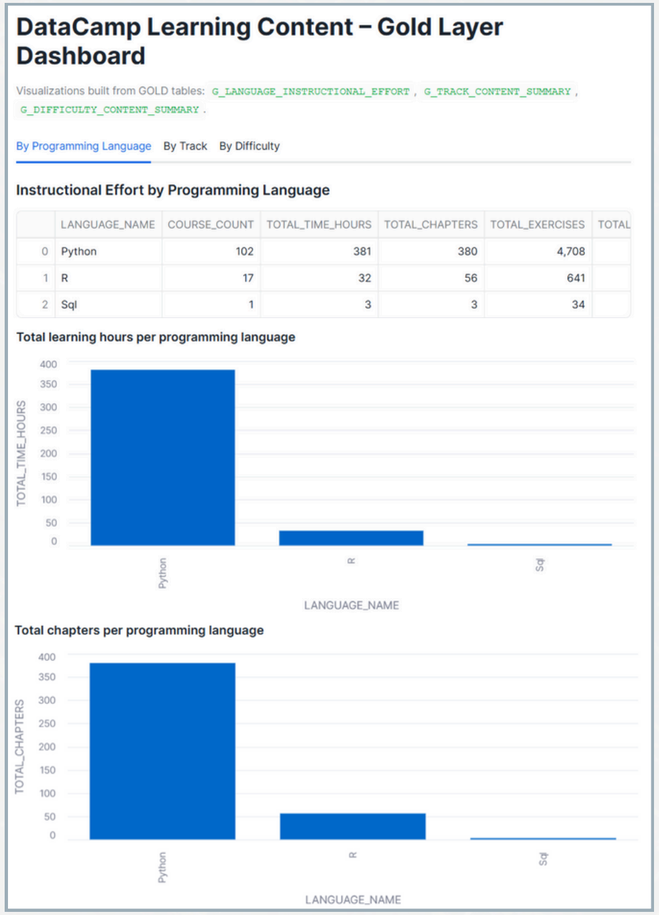


Visualization - 3 Use Cases

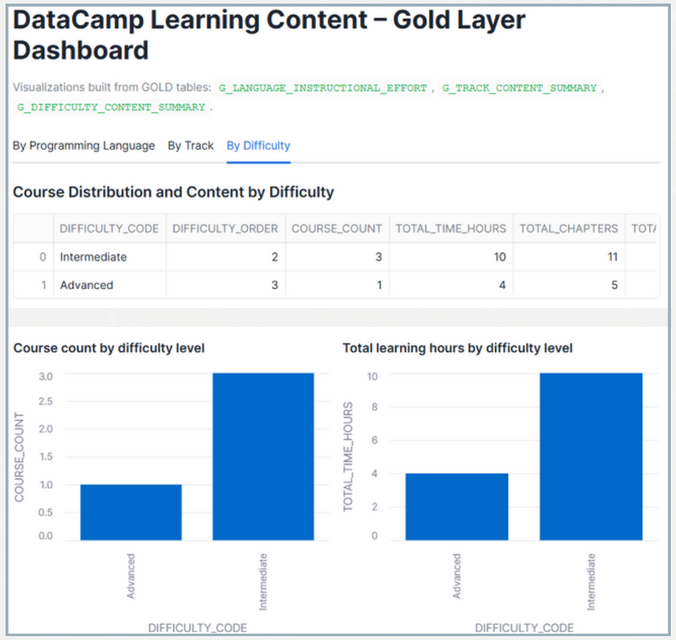
By Track



By Programming Language



By Difficulty



Incremental Load Workflow

Pre-load validation

Silver & Gold Audit Log SPs

Bronze Layer

- Create 4 new CSVs
- Create New Bronze stage
- Create 4 snowpipes
- Upload CSVs → Stage
- Auto-ingest via Snowpipe
- Validate Bronze loads

Silver Layer

- SP1 → Update BTs
- SP2→ Update DTs
- SP3→ Update FTs
- Wrapper SP → Run all 3
- Validate Silver loads

Gold Layer

- No manual step required
- Gold dynamic tables auto-refresh

Incremental Data Verification

Call the Audit Log stored procedures again and view the audit tables. Validate the updated count of rows and change in aggregate values across all the tables.

Incremental load

SILVER_LOAD_AUDIT

	AUDIT_ID	LOAD_TS	SILVER_DIM_COURSE_ROW_COUNT	SILVER_FACT_COURSE_ROWS	SILVER_DIM_TRACK_ROW_COUNT	SILVER_FACT_TRACK_ROWS
	1 – 2	1/12/2025 – 1/12/2025	116 – 120	116 – 236	115 – 117	115 – 232
1	1	2025-12-02 22:41:05.575	116	116	115	115
2	2	2025-12-02 22:54:01.515	120	236	117	232

GOLD_LOAD_AUDIT

	AUDIT_ID	LOAD_TS	GOLD_LANGUAGE_TOTAL_COURSES	GOLD_TRACK_TOTAL_COURSES	GOLD_DIFFICULTY_TOTAL_SUBS
	1 – 2	1/12/2025 – 1/12/2025	116 – 120	214 – 219	16083645 – 16126245
1	1	2025-12-02 22:44:33.567	116	214	16083645
2	2	2025-12-02 22:54:03.508	120	219	16126245

AI SQL Functions

AI in Snowflake

Use Case 1: AI_COMPLETE

Predicts course ***title*** using external LLM (claude-3-7-sonnet)

Added column ***ai_predicted_title***, filled via AI_COMPLETE

Use Case 2: AI_SIMILARITY

Calculates semantic similarity between ***title*** and ***ai_predicted_title***

Added column ***ai_title_similarity,*** updated using AI_SIMILARITY

- **34% of titles have perfect similarity (1.0)**
- **55% ⩾ 0.90 → very high accuracy; only 1 value < 0.40**

AI SQL Columns - Bronze Layer

	TITLE	AI_PREDICTED_TITLE	AI_TITLE_SIMILARI	DESCRIPTION	SHORT_DESCRIPTION
1	Hyperparameter Tuning in Python	Hyperparameter Tuning in Python	1	Building powerful machine learning models depends heavily	Learn techniques for aut
2	Manipulating Time Series Data in Py	Manipulating Time Series Data in Py	1	In this course you'll learn the basics of manipulating time se	In this course you'll learn
3	Dimensionality Reduction in Python	Dimensionality Reduction in Python	1	High-dimensional datasets can be overwhelming and leave	Understand the concept
4	Machine Learning with PySpark	Machine Learning with PySpark	1	Spark is a powerful, general purpose tool for working with B	Learn how to make predi
5	Time Series Analysis in Python	Time Series Analysis in Python	1	From stock prices to climate data, time series data are foun	In this four-hour course,
6	Introduction to Python	Introduction to Python	1	Python is a general-purpose programming language that is	Master the basics of data
7	Extreme Gradient Boosting with XGE	Extreme Gradient Boosting with XGE	1	Do you know the basics of supervised learning and want to	Learn the fundamentals c
8	Intermediate Python	Intermediate Python	1	Learning Python is crucial for any aspiring data science prac	Level up your data scienc
9	Machine Learning for Finance in Pyt	Machine Learning for Finance in Pyt	1	Time series data is all around us; some examples are the we	Learn to model and predi
10	Data Manipulation with pandas	Data Manipulation with pandas	1	pandas is the world's most popular Python library, used for	Learn how to import and
11	Advanced Deep Learning with Keras	Advanced Deep Learning with Keras	1	This course shows you how to solve a variety of problems u	Learn how to develop de
12	Web Scraping in Python	Web Scraping in Python	1	The ability to build tools capable of retrieving and parsing in	Learn to retrieve and par
13	Biomedical Image Analysis in Python	Biomedical Image Analysis in Python	1	The field of biomedical imaging has exploded in recent year	Learn the fundamentals c
14	Model Validation in Python	Model Validation in Python	1	Machine learning models are easier to implement now more	Learn the basics of mode
15	Explainable AI in Python	Explainable AI in Python	1	Explainable AI is critical for data scientists and machine lear	Gain the essential skills u
16	Working with Categorical Data in Py	Working with Categorical Data in Py	1	Being able to understand, use, and summarize non-numeric	Learn how to manipulate
17	Reinforcement Learning from Humar	Reinforcement Learning from Humar	1	Combine the efficiency of Generative AI with the understan	Learn how to make GenA
18	Introduction to Data	Introduction to Data	1	Data is an exciting and growing field that plays a significant	Gain an introduction to d

DCAMP_COURSES_BRONZE

Implementation of Cortex Search

Why: Avoid repetitive SELECT queries for searching relevant details; use Cortex Search for faster, smarter lookups.

Setup: Create 2 Cortex Search Services in BRONZE schema for tables COURSES and ALL_TRACKS.

Search Focus: Select an appropriate column as the search focus

COURSES	ALL_TRACKS
Focus on TITLE column.	Focus on TRACK_TITLE column.

Search Filters: Include relevant columns as search filter parameters.

Execution: Run search services for related course/track titles.

Validation: Confirm search results return accurate and relevant values.

Cortex Search Results

COURSES_CORTEX_SEARCH

"python beginner"


```
[
  {
    "@scores": {
      "cosine_similarity": 0.62192774,
      "text_match": 1.4807121000000000e-07
    },
    "DIFFICULTY_LEVEL": "1",
    "PROGRAMMING_LANGUAGE": "python",
    "TITLE": "Introduction to Python"
  },
  {
    "@scores": {
      "cosine_similarity": 0.6165951,
      "text_match": 1.4807121000000000e-07
    },
    "DIFFICULTY_LEVEL": "1",
    "PROGRAMMING_LANGUAGE": "python",
    "TITLE": "Introduction to Python for Developers"
  },
  {
    "@scores": {
      "cosine_similarity": 0.5640888,
      "text_match": 1.4807121000000000e-07
    },
    "DIFFICULTY_LEVEL": "1",
    "PROGRAMMING_LANGUAGE": "python",
    "TITLE": "Intermediate Python"
  },
```

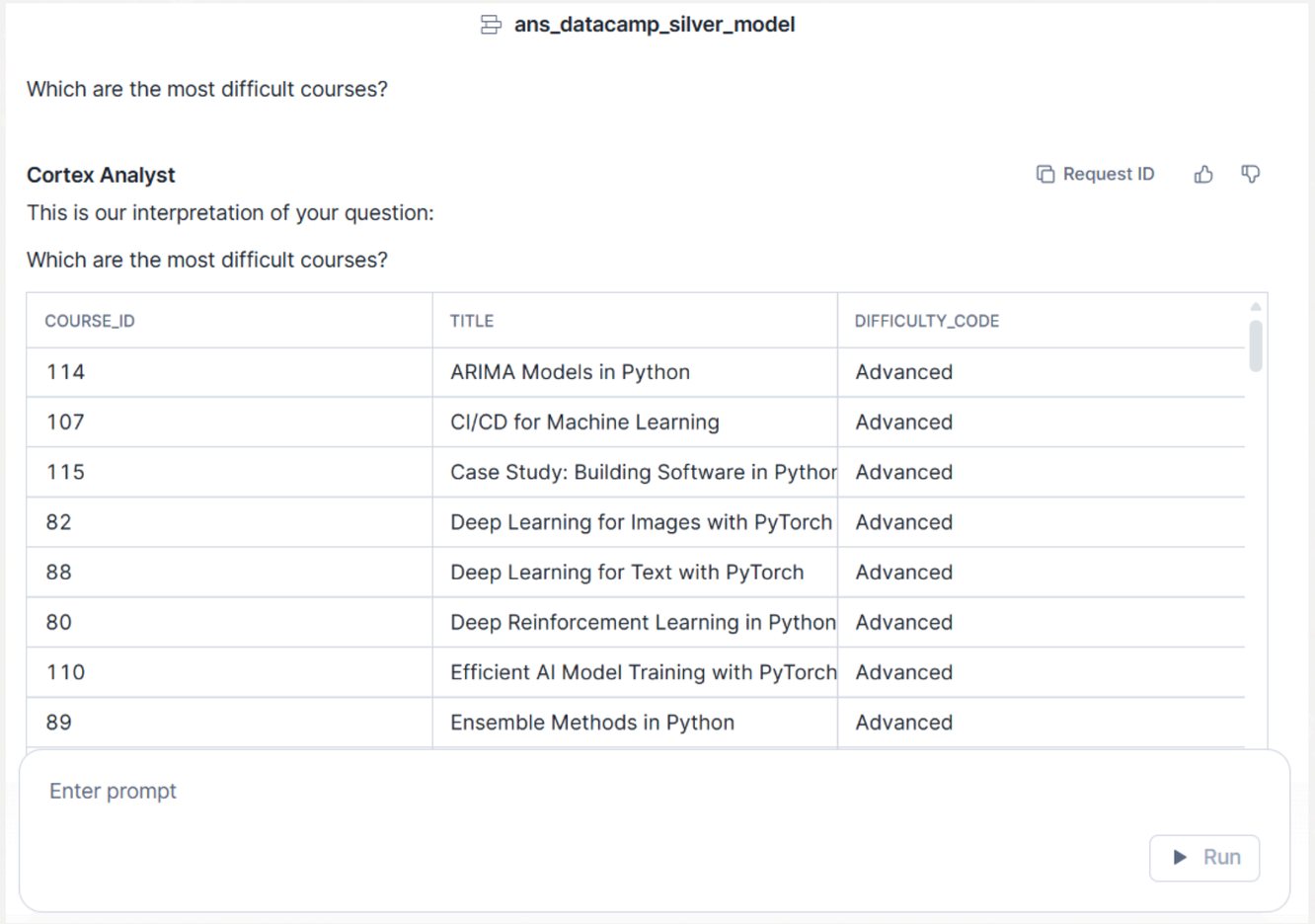

TRACKS_CORTEX_SEARCH

"machine learning"


```
[
  {
    "@scores": {
      "cosine_similarity": 0.5353421,
      "text_match": 1.6448731
    },
    "COURSE_COUNT": "12",
    "IS_CAREER": "Yes",
    "TRACK_TITLE": "Machine Learning Engineer"
  },
  {
    "@scores": {
      "cosine_similarity": 0.51772237,
      "text_match": 1.6448731
    },
    "COURSE_COUNT": "6",
    "IS_CAREER": "No",
    "TRACK_TITLE": "Supervised Machine Learning in Python"
  },
  {
    "@scores": {
      "cosine_similarity": 0.49985087,
      "text_match": 1.6448731
    },
    "COURSE_COUNT": "21",
    "IS_CAREER": "Yes",
    "TRACK_TITLE": "Machine Learning Scientist in Python"
  },
```

Cortex Analyst - Answer to Complex Queries

- Download YAML template
- Add silver schema tables + relationships
- Upload YAML to stage
- Open Cortex Analyst
- Select semantic model
- Open YAML
- Use chatbot to query the data



Meet the Team

Abhiroop Kumar

Nikhil Kumar

Simoni Dalal

The End

THANK YOU

DROPPING THE TABLES & SIGNING OFF!