

LLM Workshop - Session 1: Getting Started Guide

Revision 2

Last updated: Jul 19, 2024

Databricks Marketplace - Models

[Marketplace Model Selection](#)

[Model Serving](#)

[Testing the Model](#)

Databricks Playground - Comparing Models

[Choosing the model \(Llama 3 instruct model that was served in previous step\)](#)

Databricks Data Rooms

[Setting up data for data rooms](#)

[Uploading the csv data file to unity catalog](#)

[Creating the Gene Space](#)

Databricks Marketplace - Models

Marketplace Model Selection

- Click on marketplace on the navigation bar



databricks

New

Workspace

Recents

Catalog

Workflows

Compute

SQL

SQL Editor

Queries

Dashboards

Alerts

Query History

SQL Warehouses

Genie Spaces

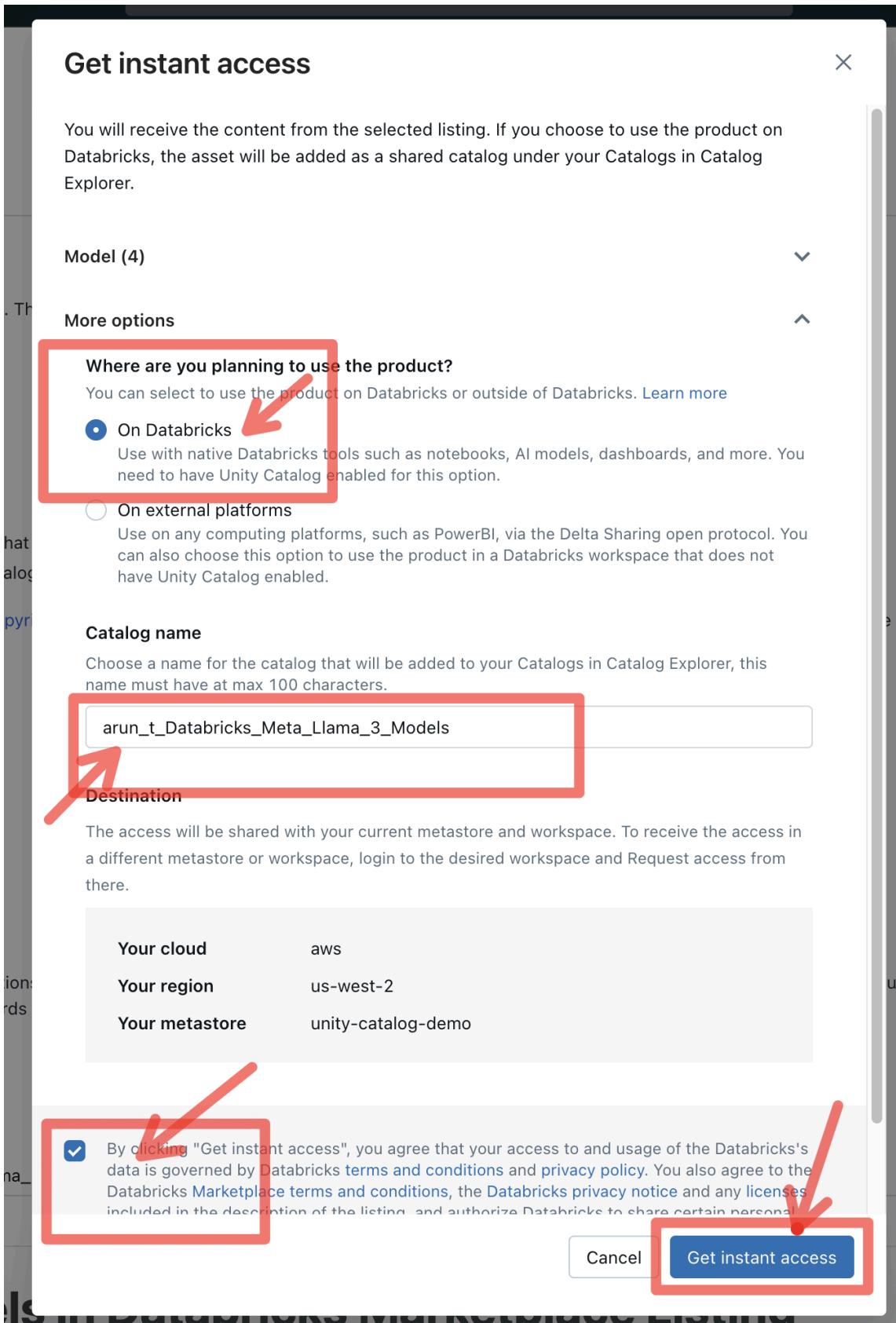
- Select filter to “models” in product

- Choose Meta Llama 3 models

- Click on Meta Llama 3 Models to open it
- Click on Get Instant Access.

- In the pop up screen choose

- “On Databricks” to where are you planning to use the product
- Change catalog name to different name (add your name to the front)
arun_databricks_meta_llama_3_model
- Click on the check box to accept the agreement
- Click on get instant access



- Now you should see the option to “open”
- Also see the catalog name and install date below installation details

Databricks Marketplace > Meta Llama 3 Models

Meta Llama 3 Models

Overview

The Llama_3 models are large language models (LLMs) developed by Meta AI. The models provided in this listing are

- meta_llama_3_70b
- meta_llama_3_8b
- meta_llama_3_70b_instruct
- meta_llama_3_8b_instruct

The models are packaged using Mlflow's transformers flavor.

- meta_llama_3_70b, meta_llama_3_8b: pretrained generative text model that can be used for further fine-tuning on specific applications.
- meta_llama_3_70b_instruct, meta_llama_3_8b_instruct: fine-tuned for dialogue-based use cases.

Llama_3 models are licensed under the Meta Llama 3 Community License, Copyright © Meta Platforms, Inc. All Rights Reserved. By installing this listing, you acknowledge and agree to the license.

Use cases

- meta_llama_3_70b, meta_llama_3_8b:
 - Fine-tuning on specific applications.
- meta_llama_3_70b_instruct, meta_llama_3_8b_instruct:
 - Generating responses to text-based instructions

Installation details

Installed on May 6, 2024
Catalog arun_t_databricks_meta_llama...

Details

Last updated	Apr 26, 2024
Pricing	Free
Access	Instantly available
Categories	Other
Model tasks	Text generation
Visibility	Public

About this provider

Model Serving

Open the model by navigating to

1. Click on Catalog on the left menu and filter to the catalog you created. In this case type in arun_t_databricks_meta_llama
2. Click on Models and select “meta_llama_3_8b_instruct”
3. Click on serve the model

New

Workspace

Catalog **arun_t_databricks_meta_llama**

Workflows

Compute

SQL

SQL Editor

Queries

Dashboards

Alerts

Query History

SQL Warehouses

Genie Spaces

Catalog Explorer unity-catalog-demo

Catalog arun_t_databricks_meta_llama

Shared

arun_t_databricks_meta_llama_models

Information schema

Tables

meta_llama_3_70b

meta_llama_3_70b_instruct

meta_llama_3_8b

meta_llama_3_8b_instruct

Overview Details Permissions

Description

meta_llama_3_8b_instruct - preview

The meta_llama_3_8b_instruct model offered in Databricks Marketplace is a text generation model released by Meta AI. It is an Mlflow model that packages Hugging Face's implementation for the meta_llama_3_8b_instruct model using the transformers flavor in Mlflow.

- It has 8 Billion parameters. While it offers the fastest processing speed, it may have lower quality compared to other models in the model family.
- It is fine-tuned specifically for dialogue use cases.

Input: Request that describes the conversation containing the text of instructions, where the messages field must alternate between user and assistant roles, ending with a user message. (AWS/Azure)

Output: Chat completion object that provides the next assistant message containing the generated response text in the conversation(AWS/Azure)

For details about the `meta_llama_3_8b_instruct` model, please visit the Hugging Face model card.

Usage

+ Add Browse DBFS DAB Endpoint of ... Serverless S

About this model

Owner: arunachalam thiagarajan Tags: Add tags

Serve this model

4. In the popup window type the name “arun_t_demo_llama3”
5. Choose the provisioned throughput “up to 7200” tokens. Change the value by selecting the dropdown based on the needs
6. Check the box “scale to zero”

Serving endpoints >
Create serving endpoint

General

Name
Endpoint name cannot be changed after creation.
arun_t_demo_llama3

URL preview: https://e2-demo-field-eng.cloud.databricks.com/serving-endpoints/arun_t_demo_llama3/invocations

Served entities

Entity details

Entity arun_t_databricks_meta_llama3 Version 1 Traffic (%) 100

Provisioned Throughput

Provisioned Throughput provides optimized inference for Foundation Models with performance guarantees for production workloads. Learn more about license requirements.

Up to 7200 tokens/second

Endpoint scales from 0 tokens/second to 7200 tokens/second. [Modify](#)

Advanced configuration

Scale to zero
Enabling scale-to-zero will lead to increased latency on the first request. Not recommended for latency-sensitive applications.

At least 0 tokens/second

Served entity name meta_llama_3_8b_instruct-1

+ Add served entity

Summary

Served entities arun_t_databricks_meta_llama3 0-7200 tokens/second 0-212 DBU

Tags Not configured

Inference tables Not enabled

7. Choose inference table and choose a schema to enable to inference tables. This step is optional

Provisioned Throughput

Up to: 7200 tokens/second

Advanced configuration

Scale to zero

Served entity name: meta_llama_3_Bb_instruct-1

Route optimization

Enable route optimization

Tags

Inference tables

Optional. Required for monitoring and diagnostics. You can configure inference tables later.

Enable inference tables

arun_t_demo Table prefix (optional)

Table name: arun_t_demo.model demo_data ima3_payload

information_schema model_inference_schema

Summary

Served entities: arun_t_databricks_meta_ll... < 100% 0-7200 tokens/second 0-212 DBU

Tags: Not configured

Inference tables: ✓ Enabled

8. Click on create at the bottom right corner of the page

Provisioned Throughput

Up to: 7200 tokens/second

Advanced configuration

Scale to zero

Served entity name: meta_llama_3_Bb_instruct-1

Route optimization

Enable route optimization

Tags

Inference tables

Optional. Required for monitoring and diagnostics. You can configure inference tables later.

Enable inference tables

arun_t_demo Table prefix (optional)

Table name: arun_t_demo.model demo_data ima3_payload

information_schema model_inference_schema

Summary

Served entities: arun_t_databricks_meta_ll... < 100% 0-7200 tokens/second 0-212 DBU

Tags: Not configured

Inference tables: ✓ Enabled

Create

9. End point is now being created and it will take few min to complete

Serving endpoints >
arun_t_demo_llama3

Serving endpoint state: Not ready (Updating)
Created by: arun.thiagarajan@dataricks.com
URL: https://e2-demo-field-eng.cloud.databricks.com/serving-endpoints/arun_t_demo_llama3/invocations
Tags: /

Inference table: Not ready
Route optimization: Not enabled

Pending configuration

Entity	Version	Name	State	Compute	Traffic (%)
arun_t_databricks_meta_llama_3_models.models.meta_llama_3_8b_instruct	Version 1	meta_llama_3_8b_instruct-1	<input type="radio"/> Creating	0-7200 tokens/second (0-212 DBU)	100

Enable inference tables Table name: arun_t_demo.model_inference_schema.arun_t_demo_llama3_payload

Metrics Events Logs

Timestamp	Event type	Served entity name	Message
2024-05-06 22:43:22	SERVED_MODEL_CREATION_EVENT	meta_llama_3_8b_instruct-1	Served entity created for served entity 'meta_llama_3_8b_instruct-1', config version 1.
2024-05-06 22:43:19	ENDPOINT_UPDATE_EVENT		Endpoint updated by arun.thiagarajan@dataricks.com.
2024-05-06 22:43:19	ENDPOINT_CREATION_EVENT		Endpoint created by arun.thiagarajan@dataricks.com.

10. Model should be ready after it is deployed

Serving endpoints >
arun_t_demo_llama3

Serving endpoint state: Ready
Created by: arun.thiagarajan@dataricks.com

Inference table: Enabled
arun_t_demo.model_inference_schema.arun_t_demo_llama3_payload
Route optimization: Not enabled

Active configuration

Entity	Version	Name	State	Compute	Traffic (%)
arun_t_databricks_meta_llama_3_models.models.meta_llama_3_8b_instruct	Version 1	meta_llama_3_8b_instruct-1	<input checked="" type="radio"/> Ready	0-7200 tokens/second (0-212 DBU)	100

Metrics Events Logs

Timestamp	Event type	Served entity name	Message
2024-05-06 22:54:44	ENDPOINT_EVENT		Endpoint 'arun_t_demo_llama3' entered READY state.
2024-05-06 22:54:43	ENDPOINT_UPDATE_EVENT		Endpoint update succeeded for endpoint 'arun_t_demo_llama3', config version 1.
2024-05-06 22:54:43	SERVED_MODEL_SERVICE_EVENT	meta_llama_3_8b_instruct-1	Served entity creation succeeded for served entity 'meta_llama_3_8b_instruct-1', config version 1.
2024-05-06 22:53:28	SERVED_MODEL_SERVICE_EVENT		Served entity 'meta_llama_3_8b_instruct-1' entered DEPLOYMENT_READY state.
2024-05-06 22:43:22	SERVED_MODEL_CREATION_EVENT	meta_llama_3_8b_instruct-1	Served entity created for served entity 'meta_llama_3_8b_instruct-1', config version 1.
2024-05-06 22:43:19	ENDPOINT_UPDATE_EVENT		Endpoint updated by arun.thiagarajan@dataricks.com.
2024-05-06 22:43:19	ENDPOINT_CREATION_EVENT		Endpoint created by arun.thiagarajan@dataricks.com.

Testing the Model

- Click on query end point
- In the pop up window select send request
- Now you can see the response from the end point

The screenshot shows the Databricks Query endpoint interface. At the top right, there are buttons for 'Permissions', 'Edit endpoint', and 'Query endpoint' (which is highlighted with a red box and a red arrow). Below these are tabs for 'Browser', 'Curl', 'Python', and 'SQL'. A 'Request' section contains a JSON message:

```

{
  "messages": [
    {
      "role": "user",
      "content": "Hello!"
    },
    {
      "role": "assistant",
      "content": "Hello! How can I assist you today?"
    },
    {
      "role": "user",
      "content": "What is Databricks?"
    }
  ]
}

```

A 'Send request' button is highlighted with a red box and a red arrow. To the right, a 'Response' section shows the generated completion:

```

from meta_llama_3_8b_instruct-1
[{"id": "chatcmp-6e851de8fd83cf0db3af893be21869bf", "object": "chat.completion", "created": 1715054266, "choices": [{"index": 0, "message": "Databricks is a fast-growing cloud-based data engineering platform that provides a collaborative workspace for data engineering, data science, and data analytics. It was founded by the original creators of Apache Spark and is backed by AWS, Microsoft, and other leading technology companies.", "role": "assistant", "content": "Databricks is a fast-growing cloud-based data engineering platform that provides a collaborative workspace for data engineering, data science, and data analytics. It was founded by the original creators of Apache Spark and is backed by AWS, Microsoft, and other leading technology companies."}], "model": "meta_llama_3_8b_instruct-1", "usage": {"prompt_tokens": 10, "completion_tokens": 10, "total_tokens": 20}, "logprobs": 0, "stop": null, "finish_reason": "length"}, {"id": "chatcmp-6e851de8fd83cf0db3af893be21869bf", "object": "chat.completion", "created": 1715054266, "choices": [{"index": 0, "message": "Databricks is a fast-growing cloud-based data engineering platform that provides a collaborative workspace for data engineering, data science, and data analytics. It was founded by the original creators of Apache Spark and is backed by AWS, Microsoft, and other leading technology companies.", "role": "assistant", "content": "Databricks is a fast-growing cloud-based data engineering platform that provides a collaborative workspace for data engineering, data science, and data analytics. It was founded by the original creators of Apache Spark and is backed by AWS, Microsoft, and other leading technology companies."}], "model": "meta_llama_3_8b_instruct-1", "usage": {"prompt_tokens": 10, "completion_tokens": 10, "total_tokens": 20}, "logprobs": 0, "stop": null, "finish_reason": "length"}]

```

The interface also includes a 'Metrics' tab with event logs and a 'Events' tab. The 'Events' tab shows log entries like 'Endpoint created' and 'Endpoint updated'.

Databricks Playground - Comparing Models

Choosing the model (Llama 3 instruct model that was served in previous step)

- Click on Playground on the menu from left side navigation
- Click on the “+” sign to add another window to the right side
- Click on the drop down box “DBRX Instruct” and change that to “arun_t_databricks_meta_llama_3” this was the model which we served in the previous step

The screenshot shows the Databricks Playground interface. On the left, the navigation sidebar includes options like Workspace, Recents, Catalog, Workflows, Compute, SQL, Data Engineering, Machine Learning, and Marketplace. The main area has tabs for 'Playground' and 'Preview'. Two AI models are displayed side-by-side:

- DBRX Instruct**: Shows a system prompt: "Databricks is an analytics platform that accelerates innovation by unifying data science, engineering, and business. It's built around Apache Spark and provides a collaborative workspace for data teams to work on projects. Databricks was founded by the original creators of Apache Spark and its goal is to help data teams solve the world's toughest problems." Below it, a message says "what is databricks".
- DBRXL Instruct**: Shows a system prompt: "Databricks offers a managed platform for running Apache Spark, which is optimized for performance, scalability, and ease of use. The platform provides a suite of tools and services for data processing, analytics, and machine learning, including:" followed by a list of features.

- Now ask the question “what is databricks” you can see the output from both the models displayed side by side and number of tokens used and response time

The screenshot shows the Databricks Playground interface with two AI models displayed side-by-side:

- DBRX Instruct**: Shows a system prompt: "Databricks is an analytics platform that accelerates innovation by unifying data science, engineering, and business. It's built around Apache Spark and provides a collaborative workspace for data teams to work on projects. Databricks was founded by the original creators of Apache Spark and its goal is to help data teams solve the world's toughest problems." Below it, a message says "what is databricks".
- DBRXL Instruct**: Shows a system prompt: "Databricks offers a managed platform for running Apache Spark, which is optimized for performance, scalability, and ease of use. The platform provides a suite of tools and services for data processing, analytics, and machine learning, including:" followed by a list of features.

- DBRX model returned results faster compared to the llama model which we just served.
- You can compare other models now based on your needs to choose the best for your use case

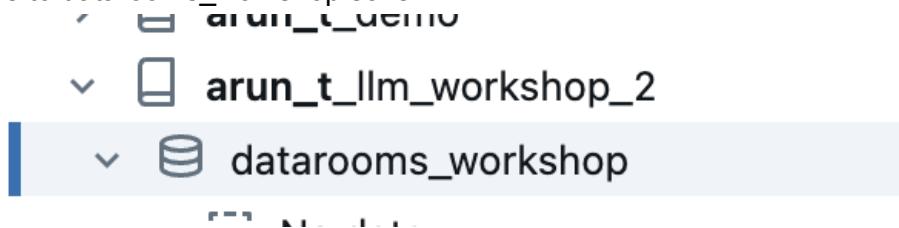
Databricks Data Rooms

Setting up data for data rooms

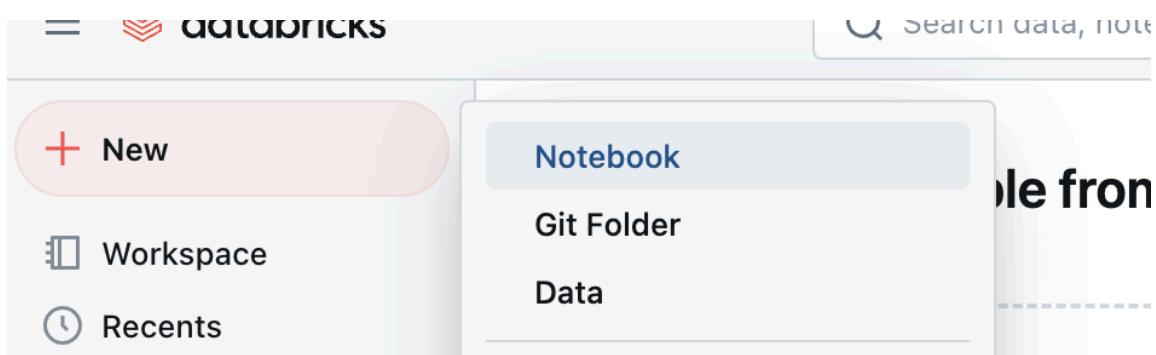
- You can choose to use your data stored in databricks unity catalog or download the file for this workshop using the link below
<https://drive.google.com/file/d/1j8nBBISTtQpFAdFtJlnrg9CREc2IQuOJ/view?usp=sharing>
- Download the csv file

Uploading the csv data file to unity catalog

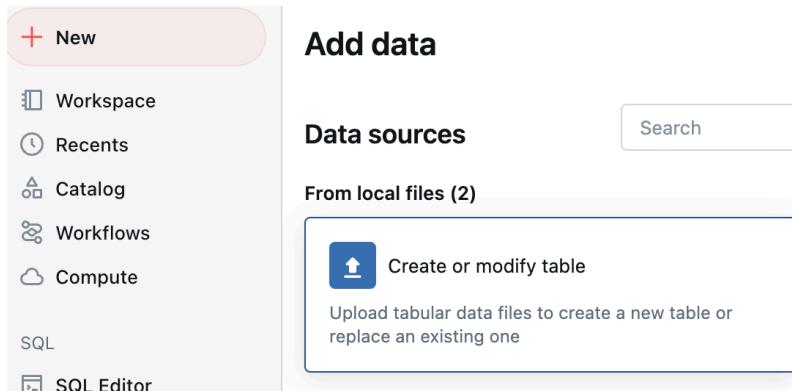
- Use any catalog & schema which you have access to upload this file. I will be upload the file to datarooms_workshop schema



- Select New->Data



- Choose create or modify table



- Select Browse and select the file

Add data >

Create or modify table from file upload

Drop one or more files here, or [browse](#)

Maximum of 10 files and total upload size of 2GB

Requires a SQL warehouse or a cluster with Databricks Runtime 10.3 and above

Supported file formats: `.csv, .tsv, .tab, .json, .jsonl, .avro, .parquet, .txt, or .xml`

- You will get uploaded.
- Choose your catalog and schema

RICKS

Add data >

Create or modify table from file upload

Retail_4k.csv uploaded 42.87MB

Preview mode Catalog Schema Table name

Advanced attributes

datarooms_workshop

retail_4_k

Previewing 50 rows, 11 columns

- Click on create table at the bottom of the screen

Sportswear	THOMAS	HENDERSON	Purchase	84.8
Mens Sportswear	WILLIAMS	HERNANDEZ	Purchase	16.7

[Cancel](#) [Create table](#)

- Your table will be created and select the sample data to review data that was uploaded

	Department	Manager	Salesperson	Transaction Type	1.2 Sale Amount	Sale Date	Store Number	Country	City
1	Mens Furnishings	BRANCH	WEBB	Purchase	48.49	2011-06-01	281a	NIA	
2	Sleepwear	VAUGHAN	MCCORMICK	Purchase	51.98	2011-06-01	137a	NIA	
3	Womens Clothing	PETERSEN	CARPENTER	Purchase	2.09	2011-06-01	211a	NIA	
4	Cosmetics	CLARK	RUSSELL	Purchase	53.66	2011-06-01	579a	NIA	
5	Mens Furnishings	HARDIN	TURNER	Purchase	60.6	2011-06-01	582a	NIA	
6	Cosmetics	SELLERS	VAUGHAN	Purchase	61.77	2011-06-01	575a	NIA	
7	Womens Shoes	ROBBINS	TAYLOR	Purchase	14.87	2011-06-01	577a	United States	Alabama
...

Creating the Gene Space

- Click on create Gene space on from the drop down on the right corner. Screen capture shown below for reference

- Provide a title for your gene space and choose a default warehouse and click on save

New

Create a new space by giving it a name and define which tables make sense for the use case.

Title

Retail Data - Workshop data

Description

Describe what data is available in this space and what type of questions users can ask.

Default warehouse

[Running] Serverless Shared Endpoint

Tables

Choose tables to use for answering questions in the space. It is best to keep the scope for each space as small as possible. Data access is governed with the viewers Unity Catalog permissions.

Catalog

Schema

Table

arun_t_llm_workshop_2

datarooms_workshop

retail_4_k

arun_t_llm_workshop_2.datarooms_workshop.retail_4_k

[Remove](#)

Sample questions

Sample questions will be presented in new chat windows for users to ask the Space.

E.g. What is our annual revenue?

+

[Cancel](#)

[Save](#)

- Gene space is now ready

The screenshot shows a Genie spaces interface. At the top, there's a header with "Genie spaces >" and the title "Retail Data - Workshop data". A "Share" button is located in the top right corner. Below the header, it says "No description". On the left side, there's a sidebar with icons for "New chat" (+ New chat), "Your conversations with Genie will be saved here.", and other navigation options. The main area contains a "Retail Data - Workshop data" section with a blue star icon. Below this, there are two buttons: "Give me 5 example questions" and "Explain the data set". At the bottom, there's a text input field with "Ask your question..." placeholder text and a send arrow icon.

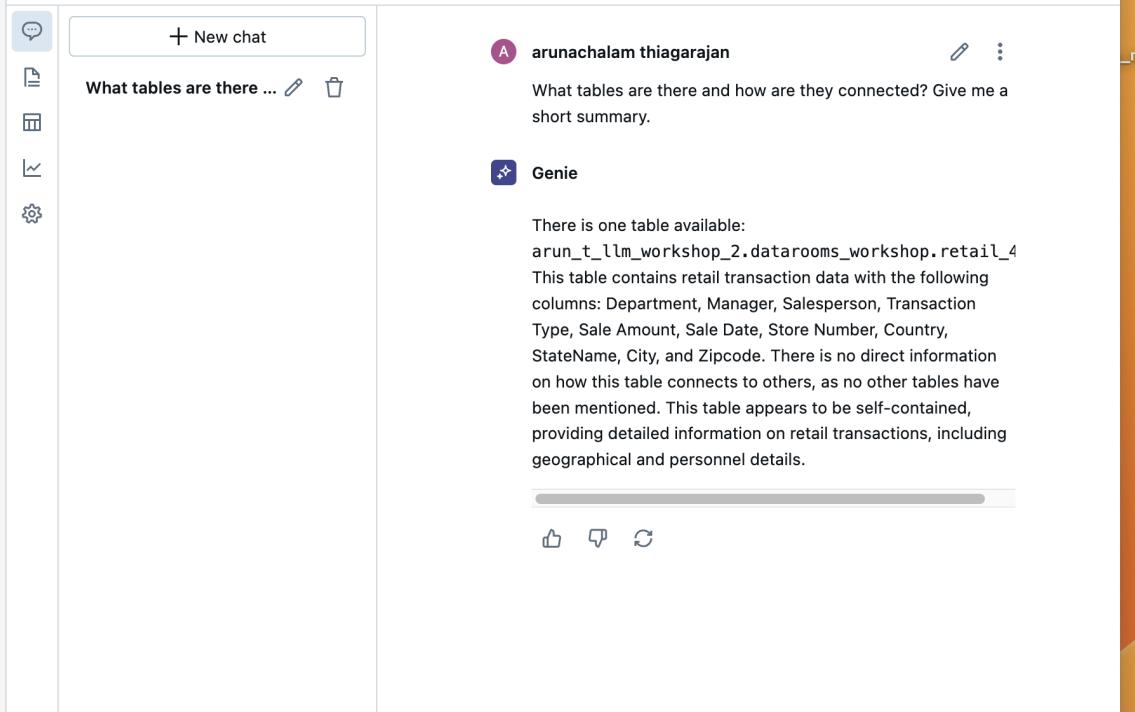
- Click on explain the data set to get details about the table. Sample shown below

Genie spaces >

Retail Data - Workshop data

No description

Share



+ New chat

What tables are there ...  

A arunachalam thiagarajan

What tables are there and how are they connected? Give me a short summary.

Genie

There is one table available:
arun_t_llm_workshop_2.datarooms_workshop.retail_4
This table contains retail transaction data with the following columns: Department, Manager, Salesperson, Transaction Type, Sale Amount, Sale Date, Store Number, Country, StateName, City, and Zipcode. There is no direct information on how this table connects to others, as no other tables have been mentioned. This table appears to be self-contained, providing detailed information on retail transactions, including geographical and personnel details.

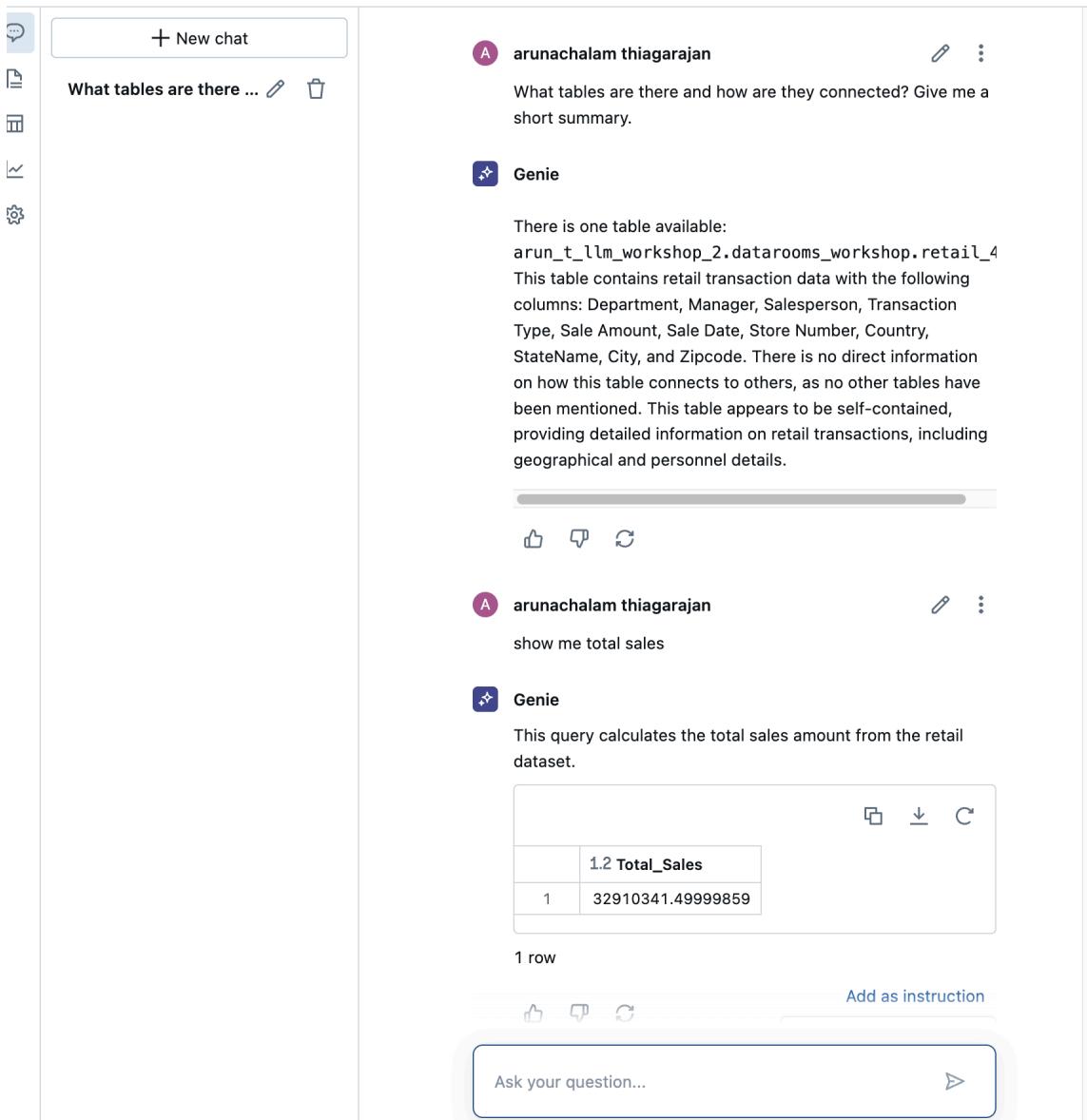
Like  Dislike  Refresh 

- Now you can ask the questions such as “show me total sale”. Gene will generate the results

Retail Data - Workshop data

Share

No description



The screenshot shows a chat interface with a sidebar on the left containing icons for New chat, Chat history, Recent files, and Settings.

User (A): arunachalam thiagarajan
What tables are there and how are they connected? Give me a short summary.

Genie: There is one table available:
`arun_t_llm_workshop_2.datarooms_workshop.retail_4`
This table contains retail transaction data with the following columns: Department, Manager, Salesperson, Transaction Type, Sale Amount, Sale Date, Store Number, Country, StateName, City, and Zipcode. There is no direct information on how this table connects to others, as no other tables have been mentioned. This table appears to be self-contained, providing detailed information on retail transactions, including geographical and personnel details.

User (A): arunachalam thiagarajan
show me total sales

Genie: This query calculates the total sales amount from the retail dataset.

	1.2 Total_Sales
1	32910341.49999859

1 row

Add as instruction

Ask your question... 

- Continue the interaction with Gene to ask for more information such as break it down by year
- Sales is now displayed by year

dataset.

	1.2 Total_Sales
1	32910341.49999859

1 row



Add as instruction

> Show generated code



arunachalam thiagarajan



break it down by year



Genie

This query calculates the total sales for each year from the retail dataset, and presents the results in ascending order based on the sale year.

	1 ² 3 Sale_Year	1.2 Total_Sales
1	2011	14704496.16000043
2	2012	18205845.34000013

2 rows



Add as instruction

> Show generated code

Quick actions

Auto visualize

Bar chart

Line chart

Pie chart

- You can break it down by department to show the sales numbers

[Show generated code](#)

A arunachalam thiagarajan



add department



Genie

This query calculates the total sales for each department, grouped by the year of the sale.

	¹ ₂ Sale_Year	^B _C Department	1.2 To
1	2011	Bedding	1419
2	2011	Casual	102
3	2011	Cosmetics	1374
4	2011	Dresses	2358
5	2011	Fashion Accessories	329
6	2011	Fine Jewelry	3986
7	2011	Fragrances	2889
8	2011	Home Furnishings	343
9	2011	Housewares	668
10	2011	Intimate	5169

50 rows

[Add as instruction](#)



[Show generated code](#)

Quick actions

Auto visualize

Bar chart

Line chart

Pie chart

Ask your question...



- You can create visualizations by selecting auto visualization or by choose bar chart or line chart

50 rows



Add as instruction

> Show generated code



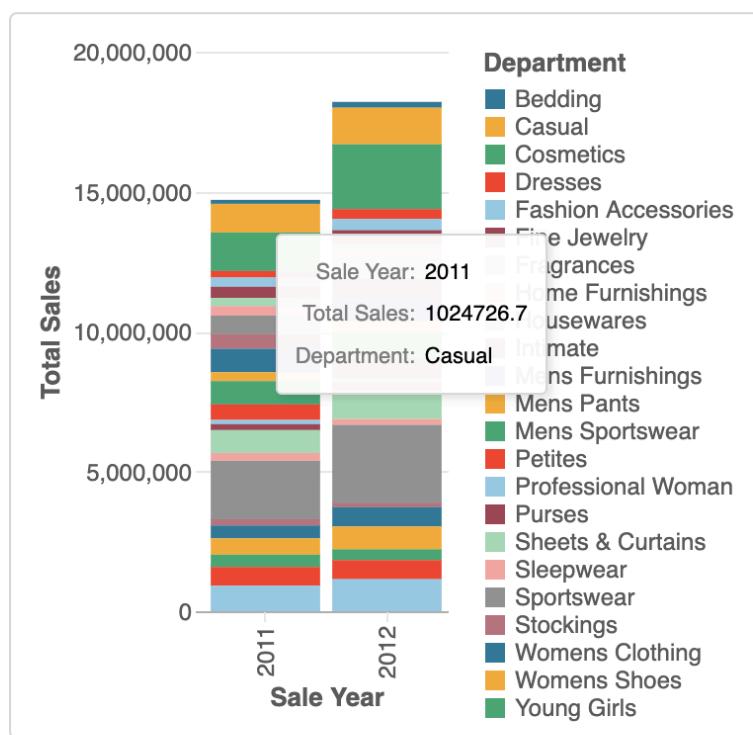
arunachalam thiagarajan



Visualize



Genie



> Show generated code

Ask your question...



- Click on the table icon to see the tables connected to this gene space

Genie spaces >

Retail Data - Workshop data

No description



Data



The tables power the space. Select the tables that you want to give as context to the space in the [edit form](#).



retail_4_k

Created by the file upload UI

5

[Open in Catalog Explorer](#) 



Department

No comment available



a

V



G

Manager

No comment available

Salesperson

No comment available

Transaction Type

No comment available

1.2 Sale Amount

No comment available

Sale Date

No comment available

Store Number

No comment available

Country

No comment available

StateName

No comment available

City

No comment available

- Click on monitoring to see the query history

The screenshot shows the Genie spaces interface for the "Retail Data - Workshop data" space. The left sidebar has icons for Home, Genie spaces, Recent, and Help. The main area shows the title "Retail Data - Workshop data" with a pencil icon, a "Share" button, and a "No description" note. A sidebar on the left has icons for Chat, File, List, and a gear. The "Monitoring" section is selected, displaying the message: "Monitor the activity within your space. Understand the errors, feedback, and results the space is returning for your users". Below this is a "Messages feed" section with filters for "Last 30 days", "Rating", "User", and "Status". A table lists five messages from a user named "arunachalam ...".

Status	Question	Rating	User	Created
Co...	Visualize		arunachalam ...	2024-07-...
Co...	add department		arunachalam ...	2024-07-...
Co...	break it down by year		arunachalam ...	2024-07-...
Co...	show me total sales		arunachalam ...	2024-07-...
Co...	What tables are there and how are they connected? Give me ...		arunachalam ...	2024-07-...

AI Function

<https://docs.databricks.com/en/large-language-models/ai-functions.html>

AI Demo Link

https://www.databricks.com/resources/demos/tutorials/data-warehouse/query-llm-with-dbsql?itm_data=demo_center

<https://notebooks.databricks.com/demos/sql-ai-functions/index.html>