



ESG Intelligent Agent

Document Extraction at Scale with Databricks



Why Document Extraction Matters

- Lots of useful data sits in unstructured text
 - Annual reports, invoices details, bank statements, claim forms etc. ex: PDF, web..
 - This data is hard for people to extract, input and analyse by eye and hand
- Document extraction automation must be fast, reliable, and cheap
- With the right context, Agents are particularly well suited to improve productivity
- Agent Bricks helps you create production-ready agents that can then be exposed for real-time inference or used for batch inference from, say, your ETL pipeline

Making structure from unstructured data ...



VERIFICATION OPINION DECLARATION GREENHOUSE GAS EMISSIONS

To: The Board of Directors and Shareholders of Amazon.com, Inc.

Apex Companies, LLC (Apex) was engaged to conduct an independent verification of the greenhouse gas (GHG) emissions reported by Amazon for the period stated below. This verification opinion declaration applies to the related information included within the scope of work described below.

The determination of GHG emissions is the sole responsibility of Amazon. Amazon is responsible for the preparation and fair presentation of the GHG emissions statement in accordance with the criteria. Apex's sole responsibility was to provide independent verification on the accuracy of the GHG emissions reported, and on the underlying systems and processes used to collect, analyze and review the information. Apex is responsible for expressing an opinion on the GHG emissions statement based on the verification. Verification activities applied in a limited level of verification are less extensive in nature, timing and extent than in a reasonable level of assurance verification.

Boundaries of the reporting company GHG emissions covered by the verification:

- Operational Control
- Worldwide

Types of GHGs: CO₂, N₂O, CH₄

Scope 3 (Market-based) ¹: 51,763,066 metric tons of CO₂ equivalent consisting of:

Purchased Goods & Services (Amazon corporate purchases made for Amazon's operations and services, Amazon branded products)

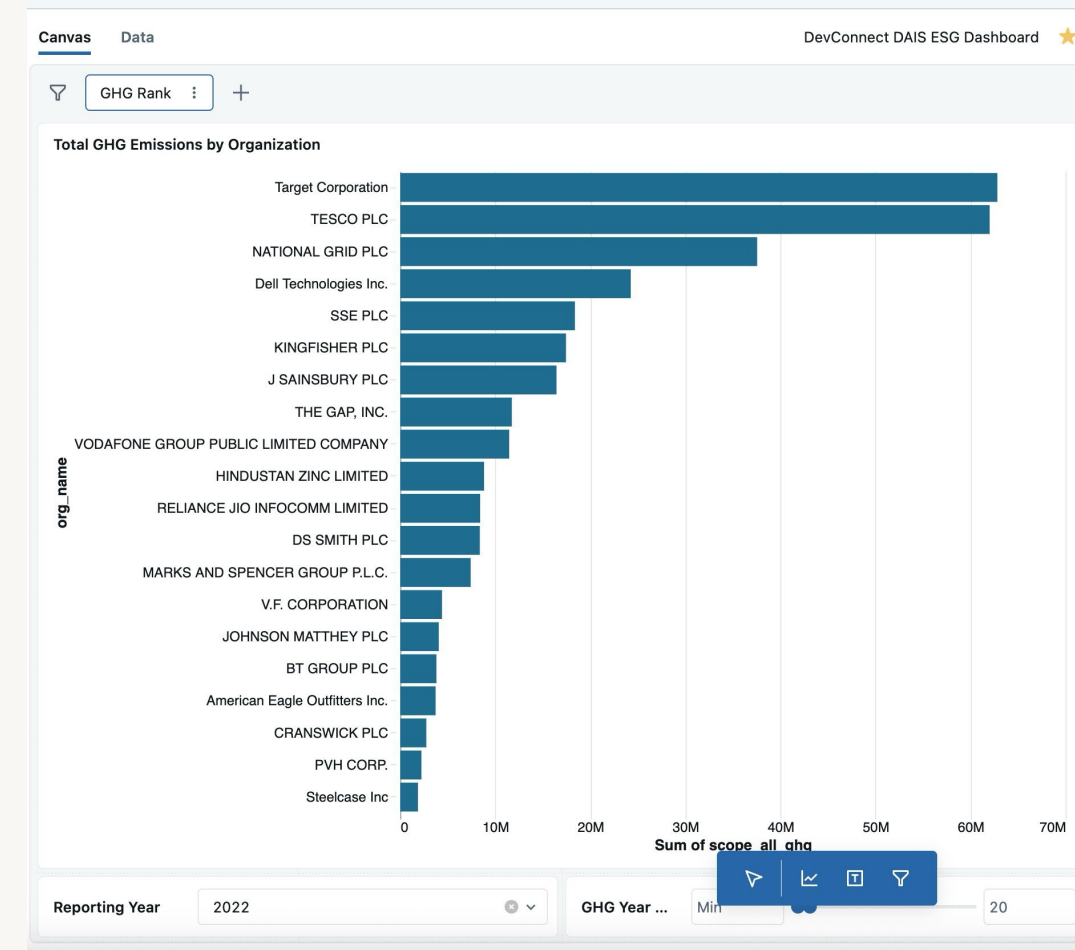
Capital Goods

Fuel-and Energy-Related Activities²

Upstream Transportation and Distribution

Business Travel

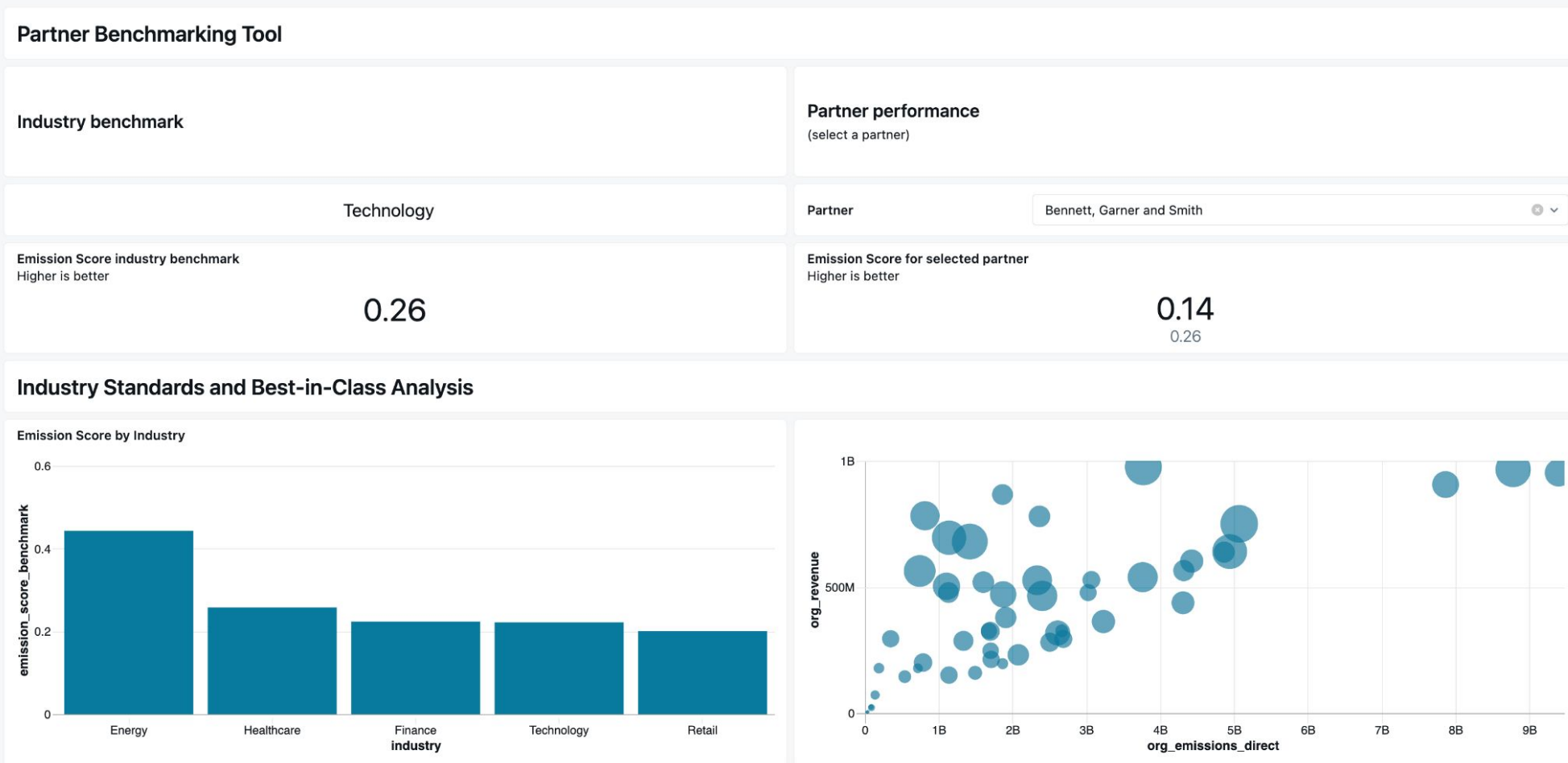
Document
extraction with
an Agent



The use case

- Your organization wants to reduce its direct and indirect greenhouse gas emissions
- To help reduce indirect (aka scope 3) emissions, you are responsible to enable the evaluation of your org's suppliers and business partners in terms of their sustainability performances by benchmarking them against their respective industry's standards.
- Proposed metric: "Emission score" = ratio of a business partner's revenue, expressed in millions USD and its direct emissions (Scope 1) expressed in tons of CO2 emitted on a yearly basis.
- End product: a business-facing interface that allows the benchmarking of your suppliers and business partners against industry standards.

Example outcome



What we provide you

- 2 data sources:
 - **CSV data files** containing the emissions and revenue for public companies in several sectors. This will allow you to compute benchmarks against which to assess your suppliers / business partners.
 - **PDFs** provided by your business partners that report on their own scope 3 emissions
- Cluster & GPUs for your team

What we provide you

PDFs and CSVs

/Volumes/workspace/esg/pdf_documents

The screenshot shows the Databricks Catalog Explorer interface. The left sidebar displays a tree view of the catalog structure, with 'workspace' > 'esg' > 'pdf_documents' selected. The main panel shows the 'Overview' tab for the 'pdf_documents' directory. It includes a search bar, a filter for files and directories at this level, and a table listing the contents.

Name	Size	Last modified
Apex Education Agency_ 2025 Environmental,	321.11 KB	1 day ago
Data Nexus Hub.pdf	388.49 KB	1 day ago
GenomeX Biotechnologies.pdf	380.19 KB	1 day ago
Global Advisory Partners ESG Report 2025.pdf	342.44 KB	1 day ago
Global Freight Solutions Inc. Environmental, So	343.48 KB	1 day ago
GlobalConnect Telecom ESG Report 2024.pdf	293.87 KB	1 day ago
Greenfield Agricultural Enterprises_ ESG Repoi	354.50 KB	1 day ago
Helix Pharmaceuticals ESG Report 2024.pdf	304.27 KB	1 day ago
Heritage Foods Inc..pdf	381.80 KB	1 day ago
Heritage Mutual Insurance Group_ Environmen	349.68 KB	1 day ago
Heritage Properties Inc. ESG Report 2025.pdf	331.58 KB	1 day ago
Heritage Utilities Group ESG Report 2025.pdf	353.82 KB	1 day ago
HexaChem Industries ESG Report 2024.pdf	398.43 KB	1 day ago

/Volumes/workspace/esg/emissions_and_revenue_data

The screenshot shows the Databricks Catalog Explorer interface. The left sidebar displays a tree view of the catalog structure, with 'workspace' > 'esg' > 'emissions_and_revenue_data' selected. The main panel shows the 'Overview' tab for the 'emissions_and_revenue_data' directory. It includes a search bar, a filter for files and directories at this level, and a table listing the contents.

Name	Size	Last modified
companies-annual-revenue.csv	12.99 KB	6 hours ago
emissions-data-france.csv	17.25 KB	1 day ago
emissions-data-germany.csv	11.86 KB	1 day ago
emissions-data-other-1.csv	23.78 KB	1 day ago
emissions-data-uk.csv	28.46 KB	1 day ago
emissions-data-us-1.csv	23.83 KB	1 day ago
emissions-data-us-2.csv	23.89 KB	1 day ago
emissions-data-us-3.csv	6.22 KB	1 day ago

What you will build (suggestion)

- A Information Extraction Agent (using Agent Bricks)
 - A Document Intelligence Agent that can extract the relevant data points from unstructured text
- A Job, scheduled Notebook or Lakeflow Declarative Pipeline
 - To productionize the Agent and automate the document extraction process
 - To combine with data from other companies in CSV format
- An AI/BI Dashboard or a Databricks App
 - To benchmark your business partner's performance versus others
- A Metric View (optional)
 - To standardize the definition of a KPI across the organization

Suggested high-level project plan

► Tip 1: Start building the dataset for the user interface, and then split the work for 2 and 3

► Tip 2: TBD

1. Data preparation and information extraction

PDF Loading (read raw data into a Delta table)

Information extraction using Agent Bricks (text data to structured data)

Data consolidation (join data from the different data sources and create a unified gold table)

2. User interface development

Create a Semantic layer with Unity Catalog Metric View

Option 1: Create an user-facing AI/BI Dashboard / AI/BI Genie Space

Option 2: Create a user-facing Databricks App

3. (Bonus) Productionisation using Lakeflow Declarative Pipelines

Leverage Declarative Pipelines to productionize the work done in step 1

Benefits:

- out-of-the-box incremental load
- data quality monitoring and enforcement framework
- near-real-time one click away

Get Started Here

- Access the [Workspace](#)
- Explore the data in the schema [workspace.esg](#)
- (Optional) Have a look at the reference slides for step-by-step instructions and helpful tips
- (Optional, in case you get stuck) Clone the [Git Repository](#), and use the notebook in the folder `esg/backup`
 - Use the backup folder only if you get stuck, and speak with your SAs to help you

Reference Slides

Step 1: PDF loading and Information Extraction

- Start by loading the PDFs into a Delta Table
 - How to do it:
 - Use the button “Use PDFs in Agent Bricks” in the Agents Tab to load the PDFs
 - Outcome:
 - A table with at least one column containing the PDF text as string
- Use Agent Bricks to extract relevant information from the PDFs
 - How to do it:
 - Use Agent Bricks’ Information Extraction, for instructions see following slides
 - Outcome:
 - An agent that extracts information such as Scope 3 Emissions, Company Name, Annual Revenue, Number of employees, etc.
 - Think about what information could be interesting to display in a Dashboard / App

Step 2 : Data processing

- Using the Information Extraction Agent you have just created, run batch inference on the table containing the PDF texts
 - How to do it:
 - Tip : from the Agent UI, deploy a ready to use Declarative pipeline to run batch inference on all PDFs
 - Add a post-processing step to handle edge cases and apply constraints on the extracted data
 - Outcome:
 - A reusable pipeline that handles extraction and postprocessing
 - A structured Delta Table with clean data extracted from PDFs
- Extract, clean and load the CSV data containing the emissions and revenue of public companies
 - How to do it:
 - Leverage Lakeflow Declarative Pipelines and/or Notebooks to extract structured data from the CSV
 - Outcome:
 - A reusable pipeline OR notebook that handles extraction and postprocessing
 - A structured Delta Table with clean data extracted from CSVs

Agent Bricks Workflow

Documentation

1. Load your PDFs into a Unity Catalog table
2. Go to Agent Bricks and select “Information Extraction”
3. Specify the information you want to extract
4. Start the agent
5. Leverage the agent endpoint for batch inference on the data

Instructions

Use PDFs in Agent Bricks

In order to use PDFs in Information Extraction and Custom LLM, we will automatically kick off a workflow using `ai_parse_document` to convert your PDFs into markdown, stored in a table. You do not need to do this for Knowledge Assistant, which supports PDFs directly.

Create workflow All workflows

Complete the form to import PDFs into a UC table.

Select folder with PDFs

Select folder containing PDF files for processing

Browse

Select destination table

Converted markdown will be stored in this table.

Choose the schema for the destination table

Browse

Destination table name

How long will it take?

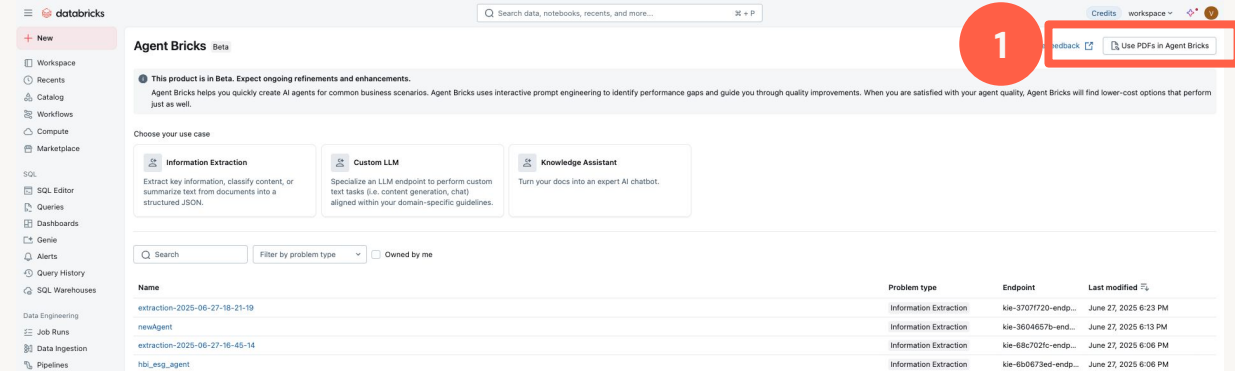
On average the import will take ~10 seconds per page. You can also track the import process under Workflows.

Select active SQL warehouse

Serverless Starter Warehouse

Cancel

Start Import



1

Click on “Use PDFs in Agent Bricks”

2

Select Folder Containing PDFs

3

Select the destination schema and give the table a name

4

Select the compute

5

Start the processing and wait for the job to finish

Instructions

The screenshot shows the Databricks Agent Bricks interface. On the left sidebar, the 'Agents' menu item is highlighted with a red box and a red circle containing the number '1'. In the main content area, the 'Information Extraction' use case is highlighted with a red box and a red circle containing the number '2'. Below the use case selection, there is a table of agents.

Agent Bricks Beta

1 This product is in Beta. Expect ongoing refinements and enhancements.

Agent Bricks helps you quickly create AI agents for common business scenarios. Agent Bricks uses interactive prompt engineering to identify performance gaps and guide you through quality improvements to find lower-cost options that perform just as well.

Choose your use case

2

Information Extraction

Extract key information, classify content, or summarize text from documents into a structured JSON.

Custom LLM

Specialize an LLM endpoint to perform custom text tasks (i.e. content generation, chat) aligned within your domain-specific guidelines.

Filter agents

Name	Problem type
custom-llm-sentiment	Custom LLM
medical-extraction-kie-demo	Information Extraction
extraction-2025-06-11-13-41-48	Information Extraction
custom-llm-2025-06-11-15-13-33	Custom LLM
custom-llm-2025-06-11-13-42-47	Custom LLM
zhl-extraction-txt	Information Extraction
eliou-custom-llm-siem	Custom LLM
new-specialization-2025-06-10-09-35-59	Custom LLM
maintenance_report_turbine_status	Custom LLM
new-specialization-2025-06-10-13-28-39	Custom LLM
model-specialization	Custom LLM

1 Select "Agents" in the sidebar menu

2 Select "Information Extraction"

Instructions

The screenshot shows the 'Information Extraction' configuration page in Databricks. It includes tabs for 'Configure', 'Build', 'Review', and 'Use'. A beta notice states: 'This product is in Beta. Expect ongoing refinements and enhancements.' The 'Overview' section describes the tool's purpose: 'Convert text documents into structured JSON by automatically detecting and classifying key information elements such as names, dates, and amounts. Ideal for processing legal agreements, HR documents, supplier contracts, financial statements, and regulatory filings. Agent Bricks may use endpoints hosted on Databricks Inc. [Documentation & License](#)'.

The 'Source documents' section has a text input field containing 'vs_test_demos.hackathon.document_texts' and a 'Browse' button. A red circle with the number '1' is placed over the 'Browse' button.

The 'Select the column containing your text data' section has a dropdown menu with 'text' selected. A red circle with the number '2' is placed over the dropdown.

The 'Sample output' section contains a text area with a JSON schema example for Amazon's carbon footprint. A 'Generate example' button is to the right. A red circle with the number '3' is placed over the 'Generate example' button.

The 'Name' section has a text input field containing 'new-extraction-2025-06-12-15-41-49'. A red circle with the number '4' is placed over the input field.

At the bottom right, there is a 'Create agent' button. A red circle with the number '5' is placed over the button.

1

Browse the data and select the table with the pdf data

2

Select the column containing the text

3

Fill the sample output with the information you want to extract

4

Give the agent a name

5

Click "Create agent" to launch it

Instructions

Build your agent

This product is in Beta. Expect ongoing refinements and enhancements.

Recommendation

scope_3_emissions

Specify the methodology used for calculating Scope 3 emissions and ensure it covers all relevant categories as defined by the GHG Protocol.

Dismiss

1

Review results

Review sample agent inputs and outputs to validate performance. Adjust model instructions or schema definitions based on your analysis to improve accuracy.

2 of 5

Model input	Text	Model output	JSON
Scope 3 GHG Reporting 2023	<p>Scope 3 emissions cover all upstream and downstream activities along our value chain. 2023 marks the first year we are disclosing our Group-wide Scope 3 emissions in accordance with the criteria set out on pages 8 and 9 of the publication "A Corporate Accounting and Reporting Standard - Revised Edition" of the "Greenhouse Gas Protocol" initiative (World Business Council of Sustainable Development / World Resources Institute). After closely examining all 15 Scope 3 emission categories, we have identified 10 categories as relevant; the remaining categories are not reported as they are either already covered in the public reporting of Scope 1 and 2 or are not applicable to our business model. Group-wide Scope 3 emissions (t CO2 equivalents in thou.)¹</p> <p>Categories</p> <p>2023</p> <p>Upstream emissions</p>	<pre>1 { 2 "company_name": "Fresenius SE & Co. KGaA", 3 "scope_3_emissions": "3662" 4 }</pre>	

2

Agent configuration

Instructions

Add instructions for how your agent should respond that apply to all extracted elements.

Enter instructions that apply to all extracted elements

Schema definition

Simulate results and keep iterating on your schema definition until you are happy with the evaluation results.

Search

Edit JSON

+ Add new field

company_name

string

The full legal name of the company as it appears in official registration documents.

scope_3_emissions

string

The total greenhouse gas emissions in metric tons of CO2 equivalent, not directly caused by the company's operations but resulting from the use of the products it sells, following the GHG Protocol's Scope 3 categorization.

3

Update agent

Use your agent

1 Review the sample results

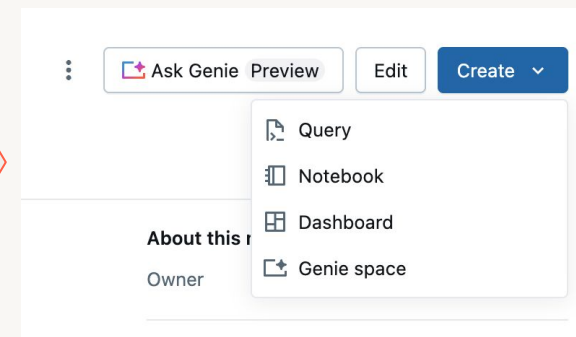
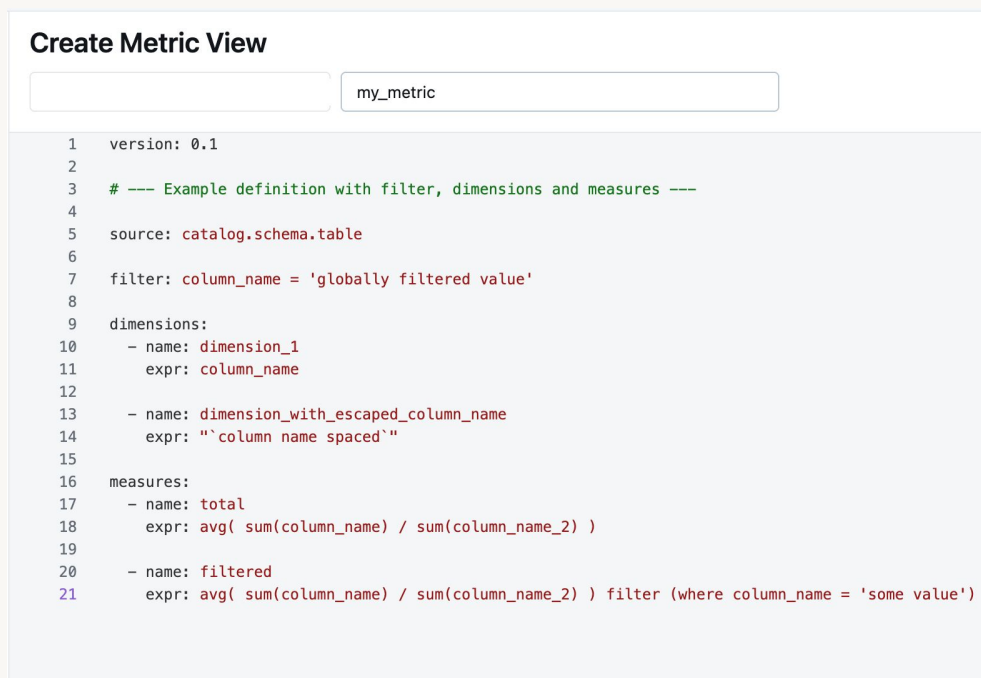
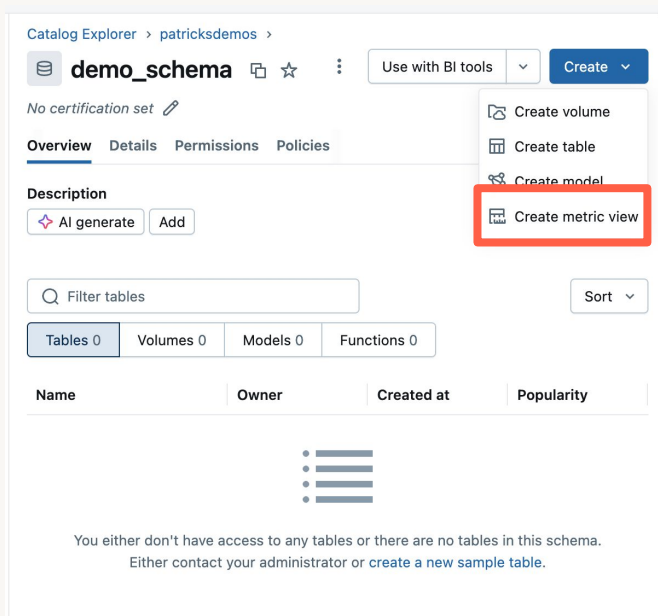
2 Add descriptions to configure your agent

3 Update or use your agent

UC Metric Views 101

Documentation

- Metric views allow to define *Measures* using common SQL aggregations, together with a set of *Dimensions* that are attributes based on which we want to enable the Measures to be dynamically analyzed.
- Once defined, the different Measures can be used in DBSQL, AI/BI Dashboards, and AI/BI Genie!



UC Metric Views pro-tip 1

Documentation

- One way to compare filtered values with unfiltered values within a same dashboard is to create window measures:

```
- name: my_measure_1  
  expr: <measure formula>
```



```
- name: my_measure_1_benchmark  
  expr: <measure formula>  
  window:  
    - order: dimension_attribute_to_be_ignored  
      range: all  
      semiadditive: last
```

- When filtering the dashboard on `dimension_attribute_to_be_ignored` (e.g. a specific company), the measure `my_measure_1` will react as anticipated. On the other hand, `my_measure_1_benchmark` will be filtered by other corresponding attributes (e.g. the industry in which the company operates) but not on the specific company itself!

UC Metric Views pro-tip 2

Documentation

- By default, Dashboard filters are applying to the whole dataset and cross-filtering (e.g. automatic filtering across visuals when clicking on a given part of a bar chart) is observed across the datasets as well.
- A way to make *Visual B* indifferent to *Dashboard Filters A* and to clicks on *Visual A* is to create *Visual B* from a different dataset, for example by cloning the initial dataset used for Filter A and Visual A and using that second dataset for Visual B.

The screenshot illustrates the process of cloning a dataset in Databricks. On the left, the 'Data' tab is active, showing a list of datasets. A red circle '1' highlights the 'Data' tab. A red circle '2' highlights the 'Run' button next to the 'emission_score_kpi_comparison' dataset. A red circle '3' highlights the 'Clone' option in the context menu. A red arrow points from the 'Clone' option to the right, where a 'Widget' configuration panel is shown. The 'Widget' panel has 'Title' checked and 'Description' unchecked. Under the 'Dataset' section, 'emission_score_kpi_overview_analy...' is selected. Below the dataset list, a 'Bar' chart type is selected.

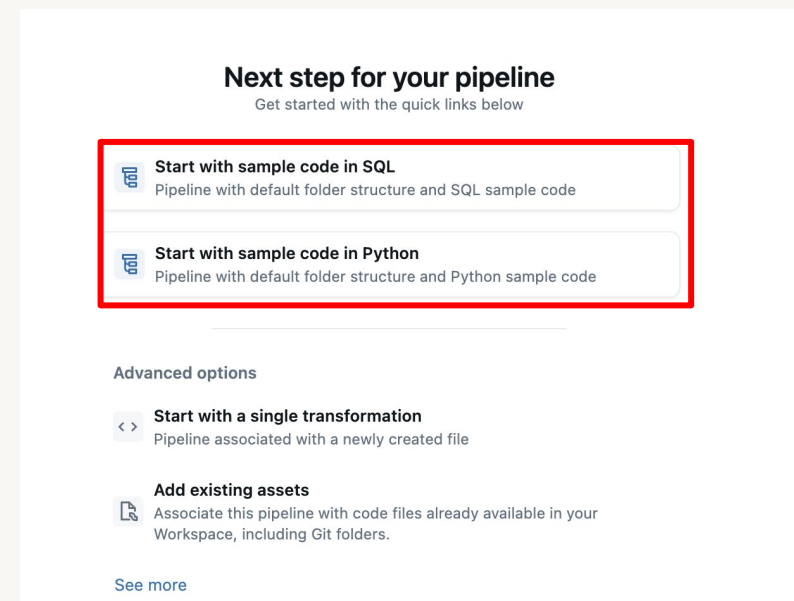
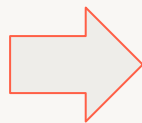
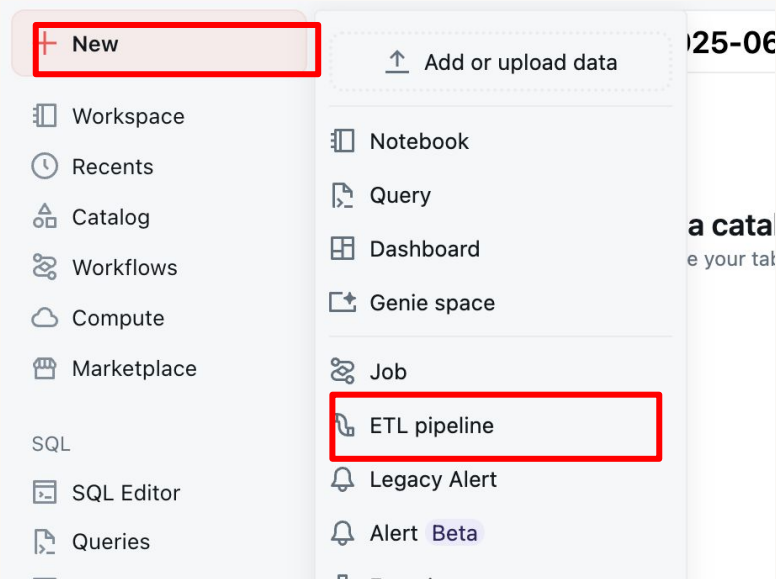
	organization
3	Retail
4	Technology
5	Technology
6	Technology
7	Finance

Declarative Pipelines 101

Documentation

- Lakeflow Declarative Pipelines is a framework for creating batch and streaming data pipelines in SQL and Python
- A common use for Lakeflow Declarative Pipelines is as follows : read data from source systems, transform that data based on requirements, such as data quality checks, and write the data to a target system, such as a data warehouse or a data lake.

QuickStart a Declarative Pipeline



Databricks Apps useful resources

- Getting started: [Documentation](#), [Medium blog post](#)
- Many useful code snippets are available on the [Databricks cookbook portal](#):
 - Accelerators for the Streamlit and Dash frameworks
 - Many use cases from interaction with Tables and Model Endpoints to dashboard embedding and API hosting.