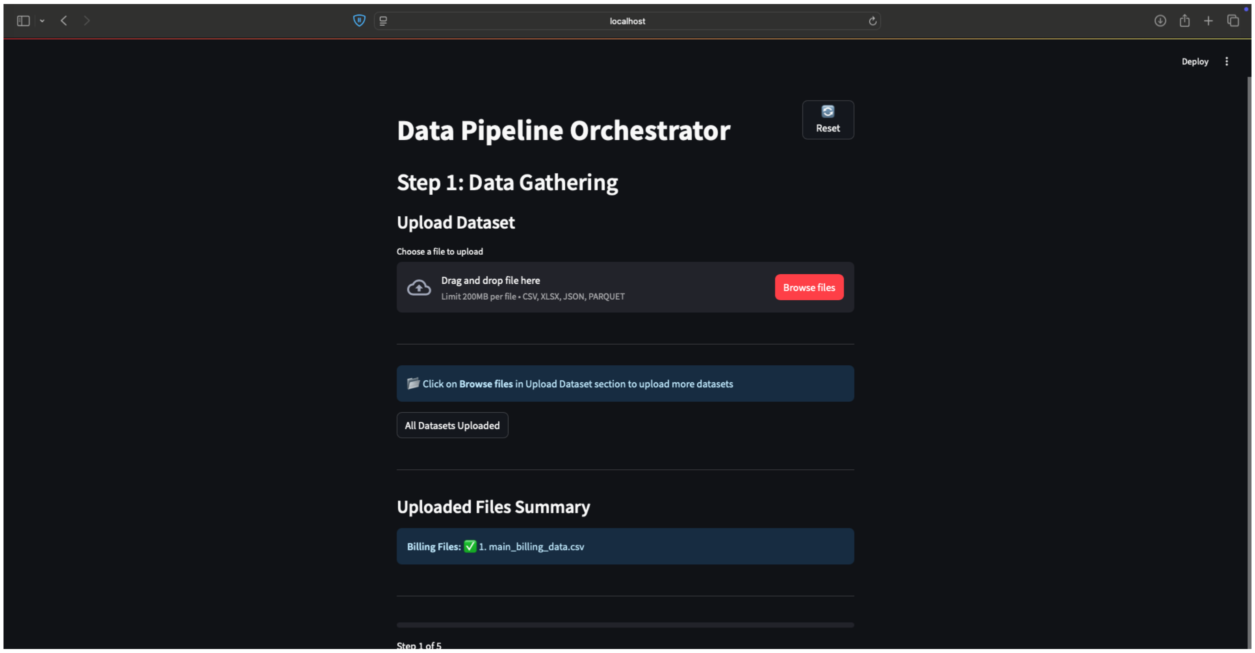
Data Prep Orchestrator Tutorial

1. Introduction

The Data Prep Orchestrator will help you unify your business data into a single dataset without code. Start by uploading your datasets one by one.



Data Pipeline Orchestrator

Step 1: Data Gathering

Upload Dataset

Choose a file to upload

Drag and drop file here
Limit 200MB per file • CSV, XLSX, JSON, PARQUET

📁 Click on **Browse files** in Upload Dataset section to upload more datasets

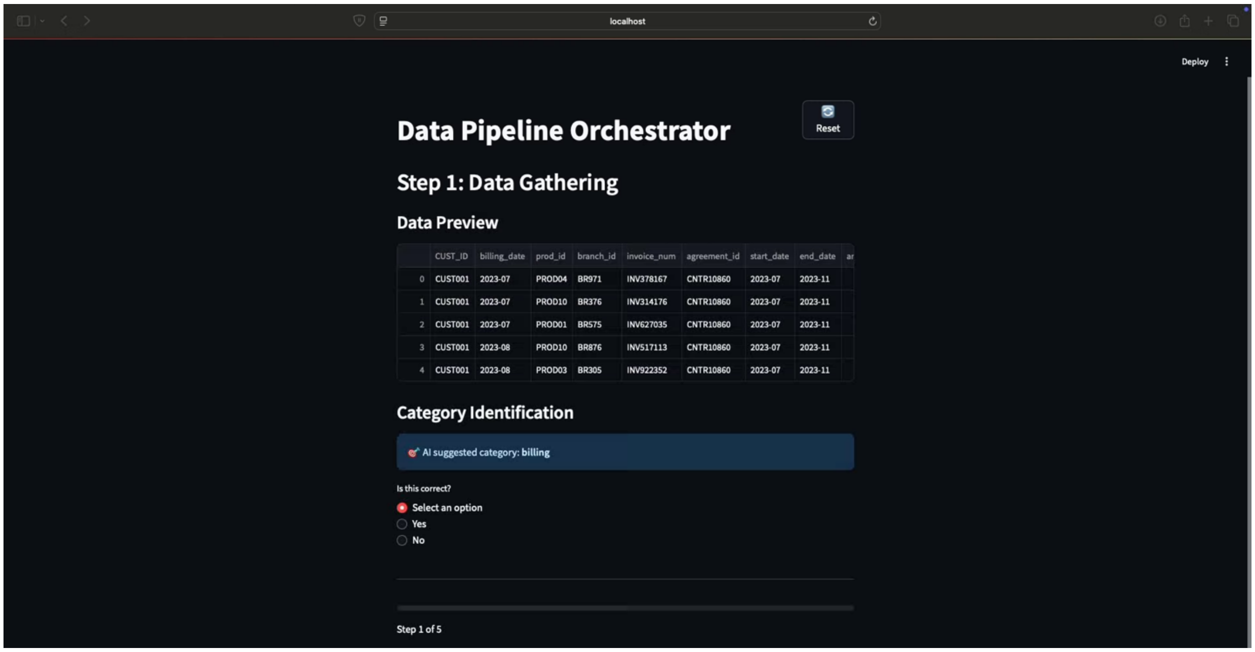
Uploaded Files Summary

Billing Files: ✅ 1. main_billing_data.csv

Step 1 of 5

2. Confirming Dataset Type

Upon upload, the Orchestrator will categorize your data so it can (in the background) prepare to map your data to an internal schema. If you agree that your data fits in the category suggested by the AI, select yes. Otherwise, saying no will allow you to manually select the category the data belongs to. Repeat this step for each dataset you upload.



Data Pipeline Orchestrator

Step 1: Data Gathering

Data Preview

	CUST_ID	billing_date	prod_id	branch_id	invoice_num	agreement_id	start_date	end_date	ar
0	CUST001	2023-07	PROD04	BR971	INV378167	CNTR10860	2023-07	2023-11	
1	CUST001	2023-07	PROD10	BR376	INV314176	CNTR10860	2023-07	2023-11	
2	CUST001	2023-07	PROD01	BR575	INV627035	CNTR10860	2023-07	2023-11	
3	CUST001	2023-08	PROD10	BR876	INV517113	CNTR10860	2023-07	2023-11	
4	CUST001	2023-08	PROD03	BR305	INV922352	CNTR10860	2023-07	2023-11	

Category Identification

🎯 AI suggested category: **billing**

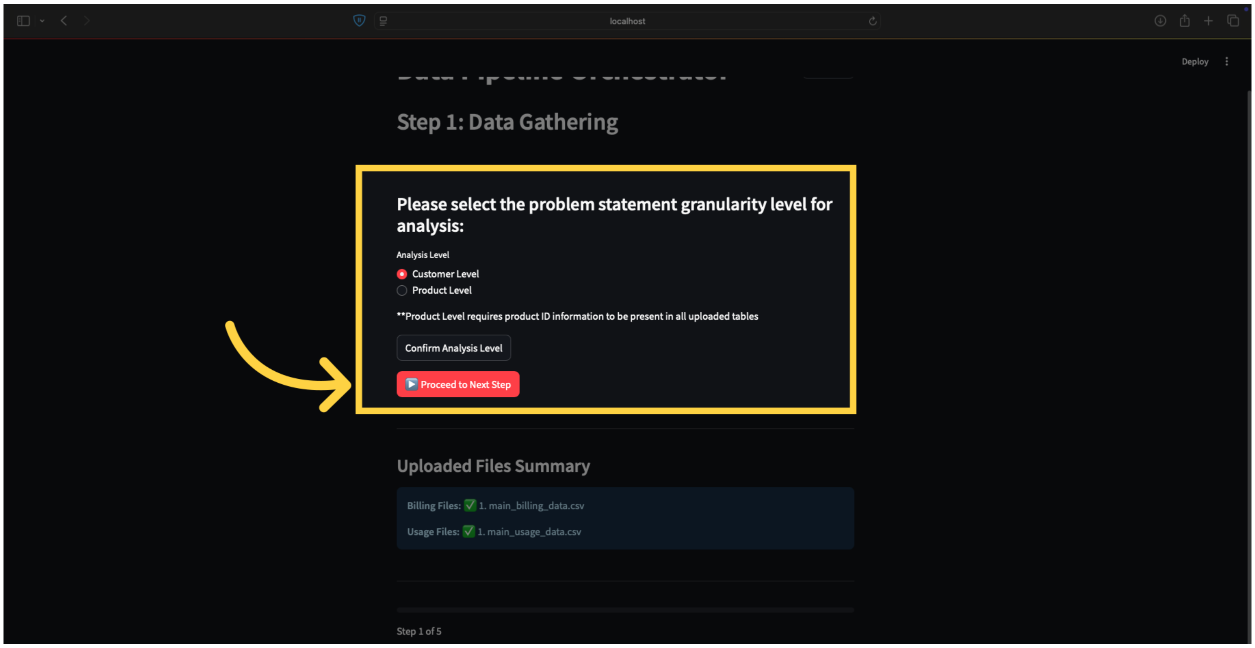
Is this correct?

● Select an option
○ Yes
○ No

Step 1 of 5

3. Unified Data Level

There are two levels at which your data can be transformed: Product and Customer level. Transforming at the Product level would lead to your data being unified at the ID/Date/Product level while the Customer level would lead to unification at the ID/Date level. Select the best level for your use case.



Step 1: Data Gathering

Please select the problem statement granularity level for analysis:

Analysis Level

● Customer Level
○ Product Level

**Product Level requires product ID information to be present in all uploaded tables

Uploaded Files Summary

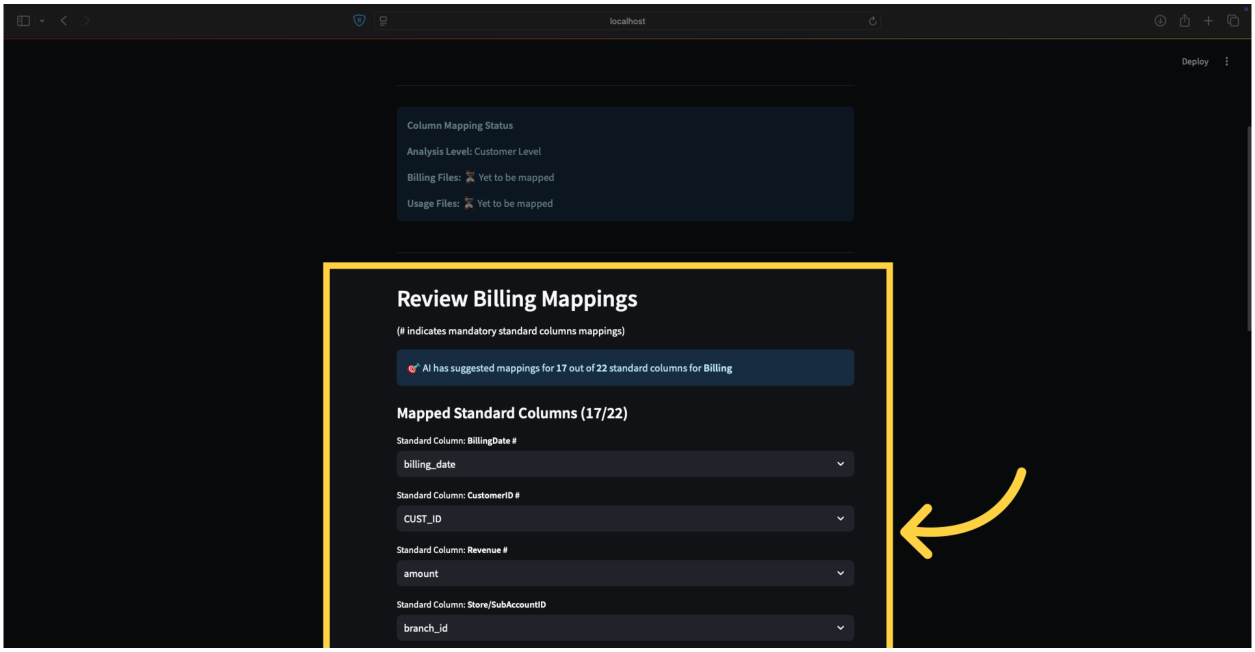
Billing Files: ✓ 1. main_billing_data.csv

Usage Files: ✓ 1. main_usage_data.csv

Step 1 of 5

4. Mandatory Mappings

After selecting the target level, you will be asked to map your data to the agent's internal schema. This mapping allows the Orchestrator to leverage different strategies for cleaning + transforming your data in later steps. The Orchestrator will suggest some mappings for you, but if you want to change them, click on the respective dropdown. At minimum, your data must map to the mandatory fields (denoted by # next to the column name).



Column Mapping Status

Analysis Level: Customer Level

Billing Files: ⌛ Yet to be mapped

Usage Files: ⌛ Yet to be mapped

Review Billing Mappings

(# indicates mandatory standard columns mappings)

🎯 AI has suggested mappings for **17** out of **22** standard columns for **Billing**

Mapped Standard Columns (17/22)

Standard Column: **BillingDate #**
billing_date

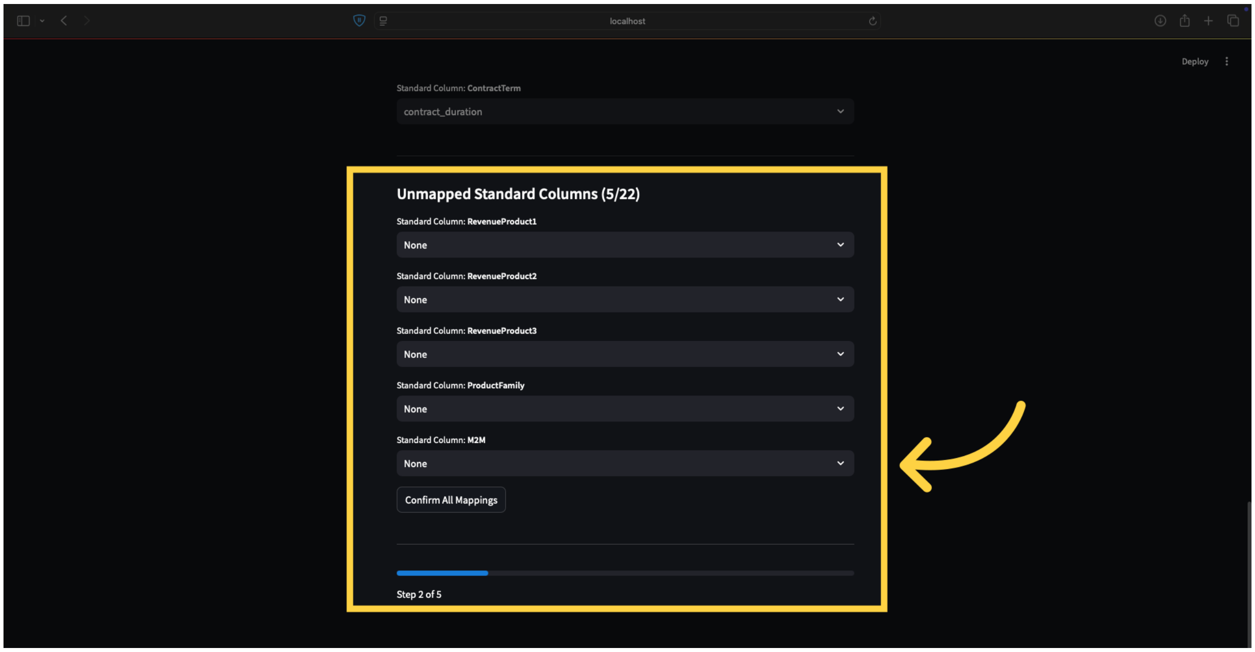
Standard Column: **CustomerID #**
CUST_ID

Standard Column: **Revenue #**
amount

Standard Column: **Store/SubAccountID**
branch_id

5. Optional Mappings

In addition to the suggested mappings, you can map other fields that the Orchestrator did not recognize. This step is optional, but any additional fields that are mapped will help the Orchestrator understand your data for later steps.



Standard Column: **ContractTerm**

contract_duration

Unmapped Standard Columns (5/22)

Standard Column: **RevenueProduct1**

None

Standard Column: **RevenueProduct2**

None

Standard Column: **RevenueProduct3**

None

Standard Column: **ProductFamily**

None

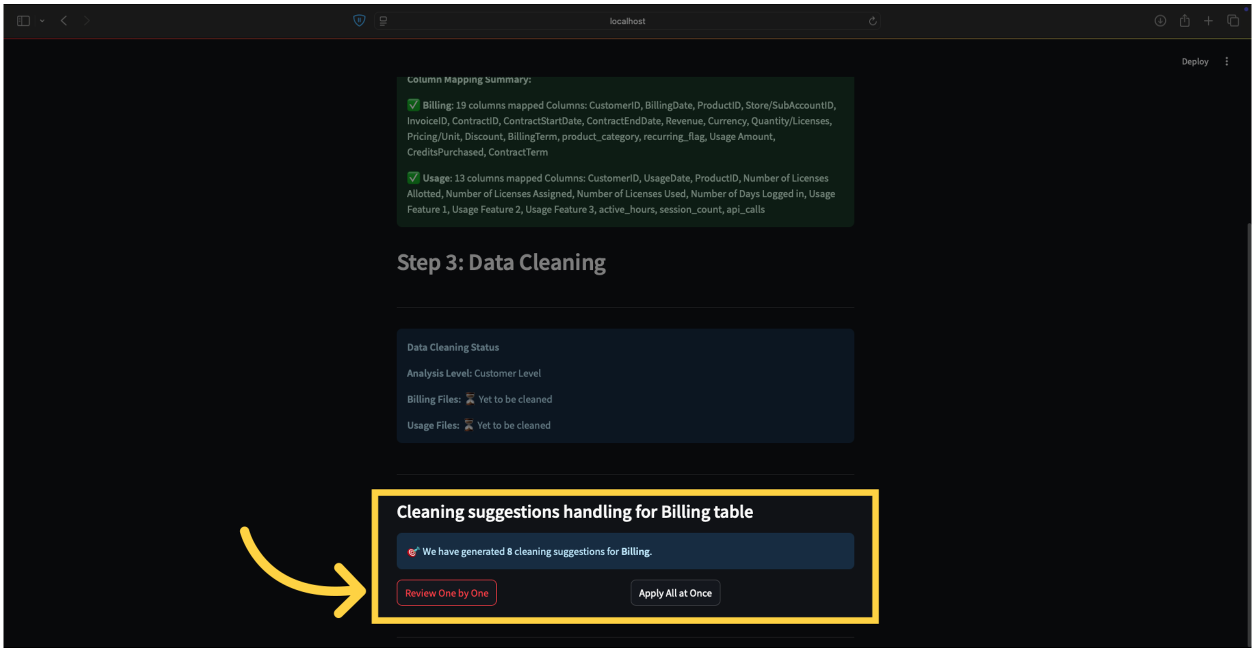
Standard Column: **M2M**

None

Step 2 of 5

6. Data Cleaning

The Orchestrator will display cleaning suggestions and the reasons it wants to perform them on your data. If you feel that the suggestion is both necessary and beneficial to your data, you can apply it. If you don't feel that you need to apply a given suggestion, you can always skip it. You can always apply the Orchestrator's suggestions at once by clicking the Apply Remaining button.



Column Mapping Summary:

✅ **Billing:** 19 columns mapped Columns: CustomerID, BillingDate, ProductID, Store/SubAccountID, InvoiceID, ContractID, ContractStartDate, ContractEndDate, Revenue, Currency, Quantity/Licenses, Pricing/Unit, Discount, BillingTerm, product_category, recurring_flag, Usage Amount, CreditsPurchased, ContractTerm

✅ **Usage:** 13 columns mapped Columns: CustomerID, UsageDate, ProductID, Number of Licenses Allotted, Number of Licenses Assigned, Number of Licenses Used, Number of Days Logged in, Usage Feature 1, Usage Feature 2, Usage Feature 3, active_hours, session_count, api_calls

Step 3: Data Cleaning

Data Cleaning Status

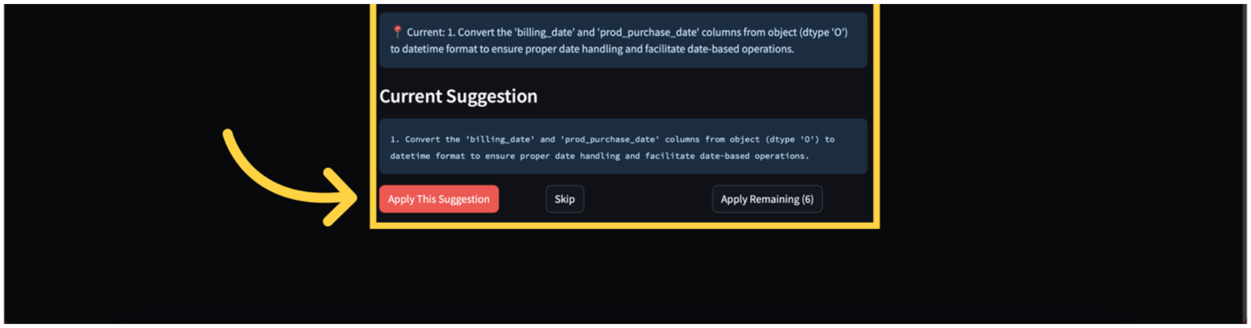
Analysis Level: Customer Level

Billing Files: ⌛ Yet to be cleaned

Usage Files: ⌛ Yet to be cleaned

Cleaning suggestions handling for Billing table

🎯 We have generated 8 cleaning suggestions for **Billing**.



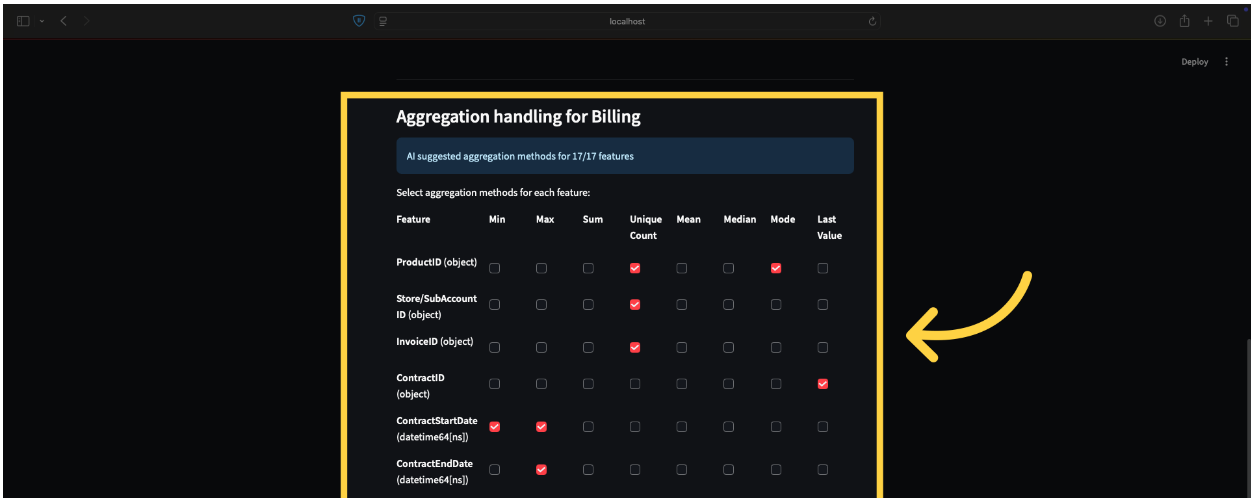
📍 Current: 1. Convert the 'billing_date' and 'prod_purchase_date' columns from object (dtype 'O') to datetime format to ensure proper date handling and facilitate date-based operations.

Current Suggestion

```
1. Convert the 'billing_date' and 'prod_purchase_date' columns from object (dtype 'O') to
datetime format to ensure proper date handling and facilitate date-based operations.
```

7. Aggregations

The Aggregation agent will provide you with a set of recommended methods to aggregate your data. You can manually tune them by clicking on their respective text boxes. If you want to see why the agent chose what it did, click on the Show Aggregation Explanations button.

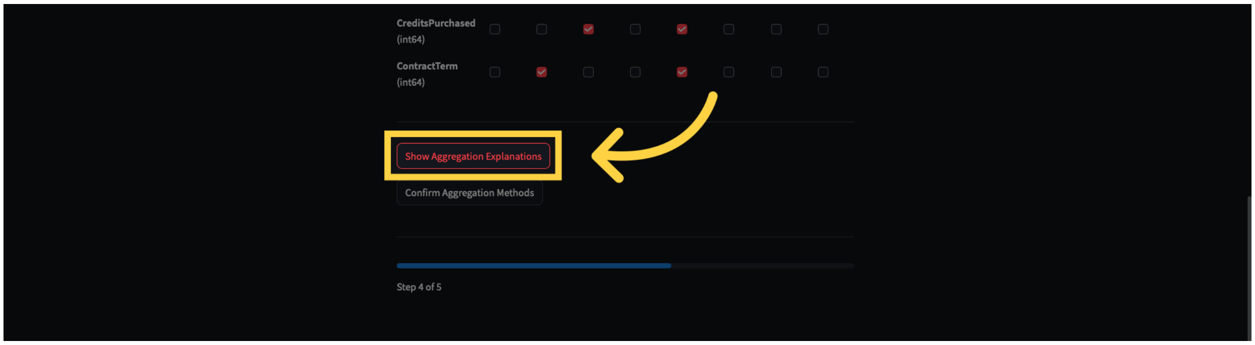


Aggregation handling for Billing

AI suggested aggregation methods for 17/17 features

Select aggregation methods for each feature:

Feature	Min	Max	Sum	Unique Count	Mean	Median	Mode	Last Value
ProductID (object)	☐	☐	☐	☑	☐	☐	☑	☐
Store/SubAccount ID (object)	☐	☐	☐	☑	☐	☐	☐	☐
InvoiceID (object)	☐	☐	☐	☑	☐	☐	☐	☐
ContractID (object)	☐	☐	☐	☐	☐	☐	☐	☑
ContractStartDate (datetime64[ns])	☑	☑	☐	☐	☐	☐	☐	☐
ContractEndDate (datetime64[ns])	☐	☑	☐	☐	☐	☐	☐	☐

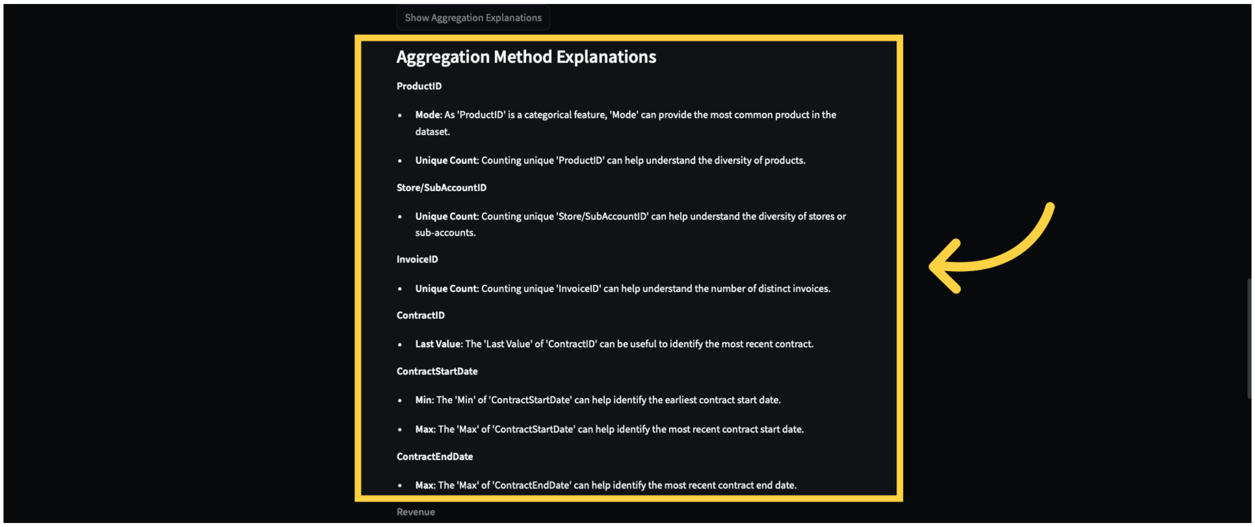


Feature	Min	Max	Sum	Unique Count	Mean	Median	Mode	Last Value
CreditsPurchased (int64)	☐	☐	☑	☐	☑	☐	☐	☐
ContractTerm (int64)	☐	☑	☐	☐	☑	☐	☐	☐

Step 4 of 5

8. Aggregation Explanations

The Orchestrator will then show, column by column, the aggregations it selected and why. If you are satisfied with the reasons, confirm them, and the Orchestrator will then aggregate your data. Otherwise, you can go back and manually edit them via the checkboxes.



Aggregation Method Explanations

ProductID

- **Mode:** As 'ProductID' is a categorical feature, 'Mode' can provide the most common product in the dataset.
- **Unique Count:** Counting unique 'ProductID' can help understand the diversity of products.

Store/SubAccountID

- **Unique Count:** Counting unique 'Store/SubAccountID' can help understand the diversity of stores or sub-accounts.

InvoiceID

- **Unique Count:** Counting unique 'InvoiceID' can help understand the number of distinct invoices.

ContractID

- **Last Value:** The 'Last Value' of 'ContractID' can be useful to identify the most recent contract.

ContractStartDate

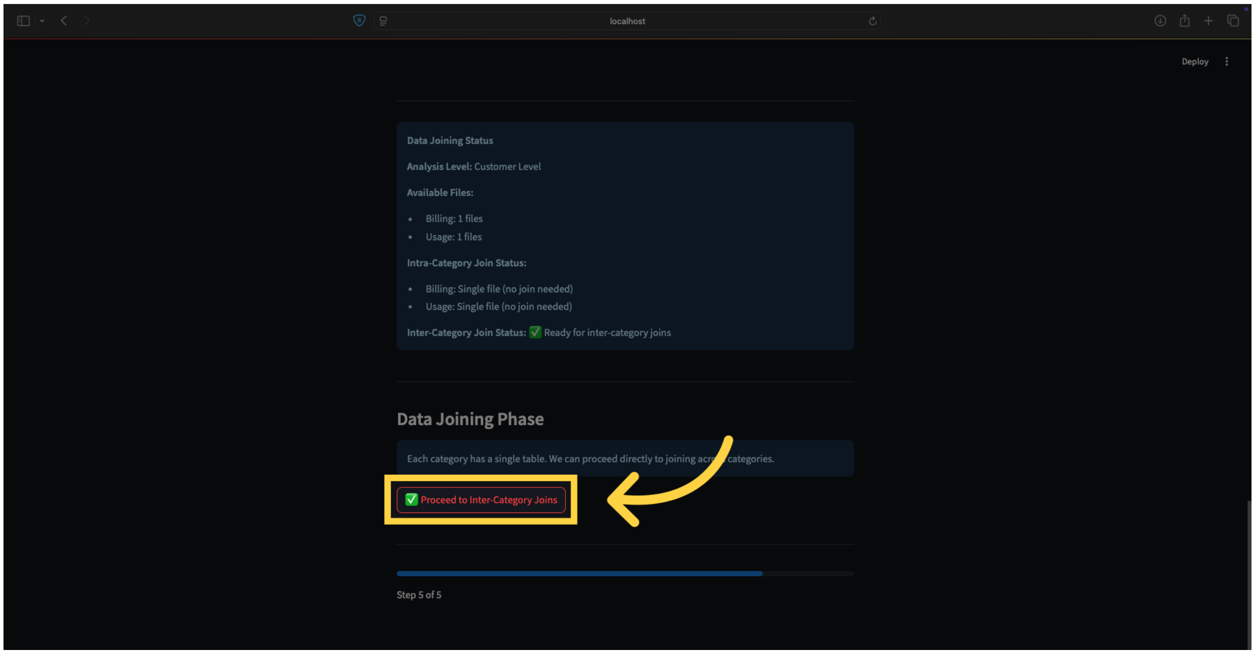
- **Min:** The 'Min' of 'ContractStartDate' can help identify the earliest contract start date.
- **Max:** The 'Max' of 'ContractStartDate' can help identify the most recent contract start date.

ContractEndDate

- **Max:** The 'Max' of 'ContractEndDate' can help identify the most recent contract end date.

9. Joining your Data

The Orchestrator will determine which columns are best for joining your datasets together based on the level you selected in Step 3. Click proceed, and your data will be joined for you.



Data Joining Status

Analysis Level: Customer Level

Available Files:

- Billing: 1 files
- Usage: 1 files

Intra-Category Join Status:

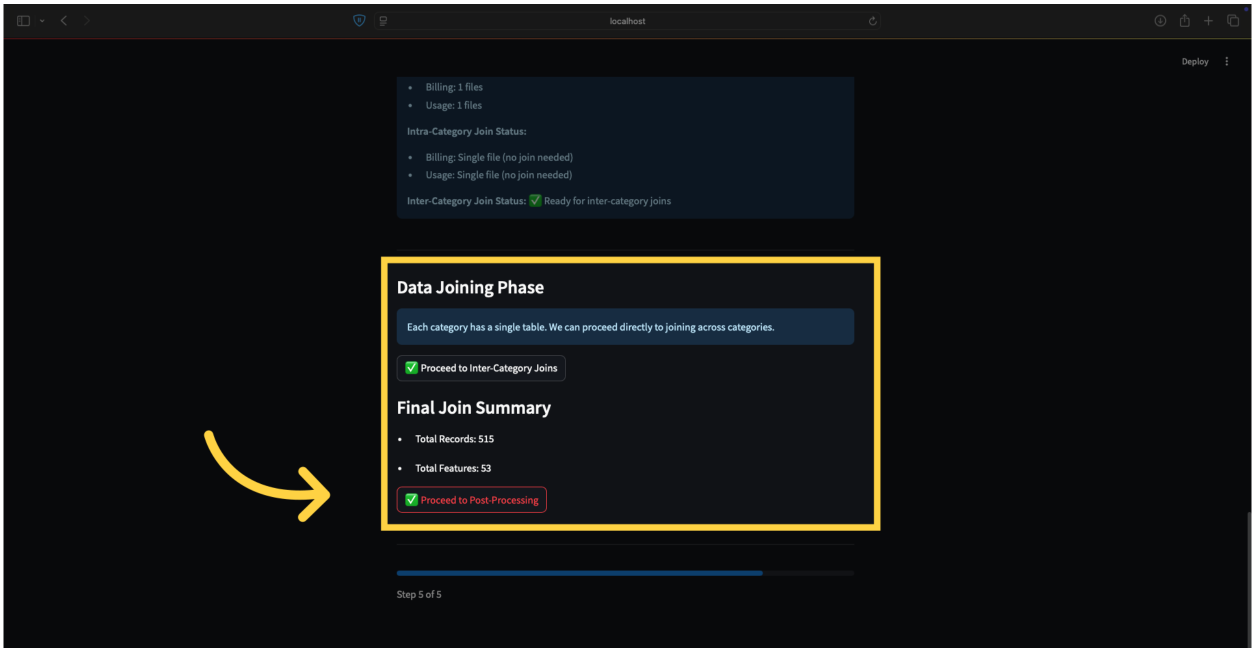
- Billing: Single file (no join needed)
- Usage: Single file (no join needed)

Inter-Category Join Status: ✅ Ready for inter-category joins

Data Joining Phase

Each category has a single table. We can proceed directly to joining across categories.

Step 5 of 5



- Billing: 1 files
- Usage: 1 files

Intra-Category Join Status:

- Billing: Single file (no join needed)
- Usage: Single file (no join needed)

Inter-Category Join Status: ✅ Ready for inter-category joins

Data Joining Phase

Each category has a single table. We can proceed directly to joining across categories.

✅ Proceed to Inter-Category Joins

Final Join Summary

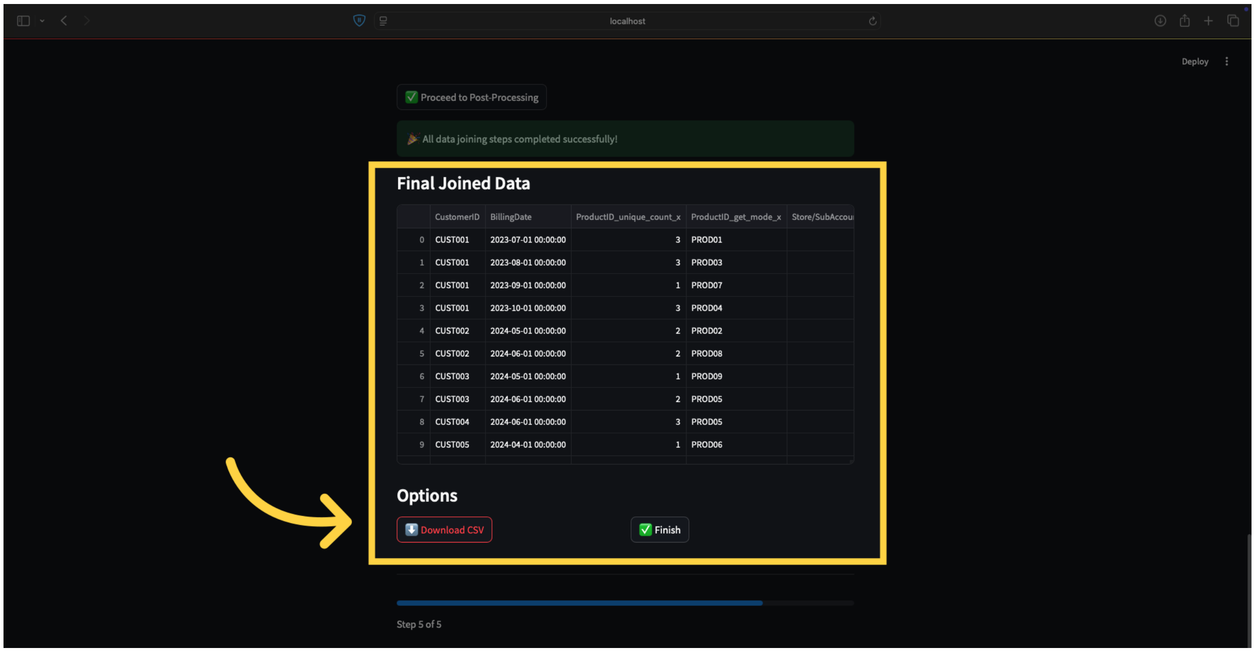
- Total Records: 515
- Total Features: 53

✅ Proceed to Post-Processing

Step 5 of 5

10. Final

Download your data, and you have finished your end-to-end data unification!



🎉 All data joining steps completed successfully!

Final Joined Data

	CustomerID	BillingDate	ProductID_unique_count_x	ProductID_get_mode_x	Store/SubAccou
0	CUST001	2023-07-01 00:00:00	3	PROD01	
1	CUST001	2023-08-01 00:00:00	3	PROD03	
2	CUST001	2023-09-01 00:00:00	1	PROD07	
3	CUST001	2023-10-01 00:00:00	3	PROD04	
4	CUST002	2024-05-01 00:00:00	2	PROD02	
5	CUST002	2024-06-01 00:00:00	2	PROD08	
6	CUST003	2024-05-01 00:00:00	1	PROD09	
7	CUST003	2024-06-01 00:00:00	2	PROD05	
8	CUST004	2024-06-01 00:00:00	3	PROD05	
9	CUST005	2024-04-01 00:00:00	1	PROD06	

Options

Step 5 of 5