# Medallion Architecture Capstone Project: The Canadian Retail Sector (2017-2025)

COMP 3839: Data Warehousing

Student Name: Vedang Sunilkumar Sharma

Student ID: A01388507

Set: NA

# Business Case & Pipeline Design:

**Business Scenario**

I'm assuming the role of a private equity firm seeking newer market opportunities in the Canadian retail sector. Sectors with strong historical performance and high growth potential are our main interests. Upon extensive research, the **Monthly Retail Trade Sales by Province and Territory** (2017 to 2025) was elected as our primary dataset. The chosen dataset not only includes both traditional and e-commerce retail sales but also breaks data down by sector and region.

This particular dataset was generously provided by **Statistics Canada** and is also updated every month. However, for our project, we anticipate that data ingestion will occur quarterly, as quarterly financial reports are a common benchmark in the financial industry.

# Goal of the Data Lakehouse:

Our rationale behind creating a **Medallion Architecture Data Lakehouse** is to derive practical, and valuable insights to help drive important business decisions. Our pipeline will be an ELT pipeline which transforms raw retail data into clean and actionable insights. Our firm will use this to:

- Identify **top-performing** and **undervalued** retail segments
- Compare **regional trends** across provinces and cities
- Analyze the growth of **e-commerce vs. traditional retail**
- Enable both real-time and historical **sector-level insights**

## Users of the Gold Layer:

The outputs of the **Gold Layer** (visualizations, summary tables, and trend reports) will be used by:

- **Investment analysts** to identify high-potential segments
- **Internal management** to evaluate the firm's strategic direction
- **Clients and stakeholders** to understand where capital is best allocated

The Gold Layer abstracts away the complexity of raw data and presents concise, business-friendly insights.

## How the Dataset Supports the Business Goal:

This dataset aligns directly with our private equity use case. It includes:

- Historical sales performance across regions and cities and even contains National aggregates
- Trends within retail segments
- E-commerce and in-store sales breakdowns

We designed the pipeline to split the data into three perspectives—National, Provincial, and City-level—so insights can be tailored for strategic regional planning.

## Additional Datasets That Could Enhance the Pipeline

To enhance future insights, we could integrate:

- Macroeconomic indicators (GDP, inflation, unemployment)
- Public company retail stock performance
- Average income and Public Demographics
- Retail real estate pricing or leasing trends

These datasets would enable correlation analysis and more predictive modeling

## Data Quality Challenges Addressed

In the silver layer, vigorous data cleaning will be performed for the sake of ensuring accuracy, readability and usability. Please find the key data transformations below:

**Getting rid of NULLs in VALUE column:**

The field called VALUE represents the overall sales figures and is the backbone of our analysis. Therefore, rows containing nulls were filtered out, as they would tinker with our final reports and visualizations.

**Dropping unnecessary columns:**

We plan to remove fields such as STATUS, SYMBOL, TERMINATED, DECIMALS, DGUID, UOM_ID, SCALAR_ID, and COORDINATE. The reason being, these fields had little to no relevance to our retail performance analysis and were just noise.

**Changing key field names for Business Readability:**

We renamed headers like:

- REF_DATE -> Report_Month
- North American Industry Classification System (NAICS) -> Retail_Category (simplified name for end-users)
- GEO -> Region (To better elucidate the granularity: provincial, national, or city)

**Cleaning the newly created REGION column:**

Upon careful inspection of the dataset, we identified three types of regions. They were – Canada, province names like Alberta, Ontario, and lastly, city-region names like Vancouver, British Columbia. Using this logic, we separate the dataset into three logical, cleaned datasets for our gold layer:

- **National_df #** Only contains “Canada”
- **Provincial_df #** Only contains provincial wide data.
- **City_df #** Only contains city-level data.

**Simplifying Category Labels:**

In the Retail_Category column, some entries had cluttered formats such as [44-45] Retail Trade. We stripped the numeric prefixes and kept only the readable sector names to avoid confusion in dashboard visuals.

**Removing Vector and UOM columns:**

We are of the opinion that UOM and Vector fields are constant, and therefore do not influence our analysis, and therefore have decided to get rid of them.

**Removing Low-Value Fields like Adjustments and Scalar Factor**:
Adjustments often repeated "Seasonally adjusted" and added no meaningful variability. Similarly, SCALAR_FACTOR was always "millions", and its value was already factored into the VALUE field. These were dropped for clarity.

**Cleaned Silver Layer table:**

Upon making all the transformations, we will be left with a clean table ready for analysis. It contains fields like:

- Report Month
- Retail Category
- Region
- VALUE # renamed to **Sales**
- Ingested_date # added for pipeline auditing and data lineage purposes

### Gold Layer

We created three Gold Layer tables:

- retail_sales_national: Shows Canada-wide trends. Used for visualizing national sales, comparing e-commerce vs. traditional retail, and identifying top-performing sectors.
- retail_sales_provincial: Focuses on provinces. Helps spot high-performing regions and trends across provinces.
- retail_sales_city: Zooms into cities. Useful for identifying top cities and local sector strength.

Each table powers clean visuals like monthly sales trend line charts, bar graphs for top sectors, and tables showing average sales by region. Filters by year, help users explore and compare segment performance by year easily to guide smart investment decisions.

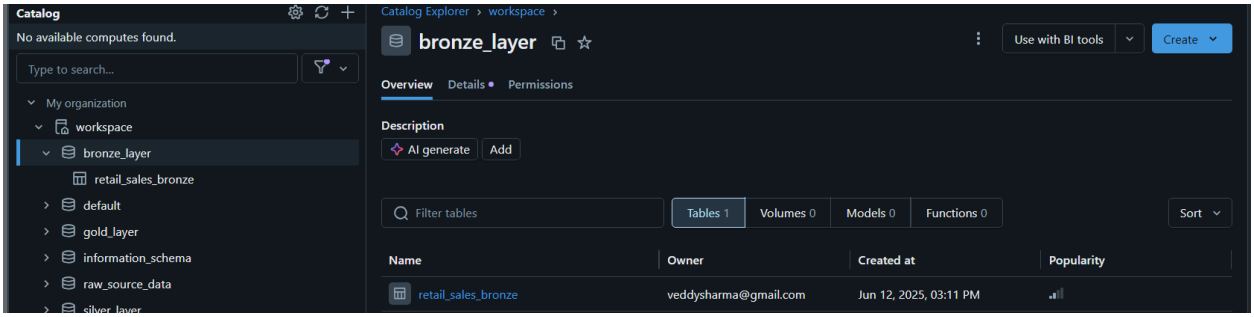
# 2. Pipeline Implementation:

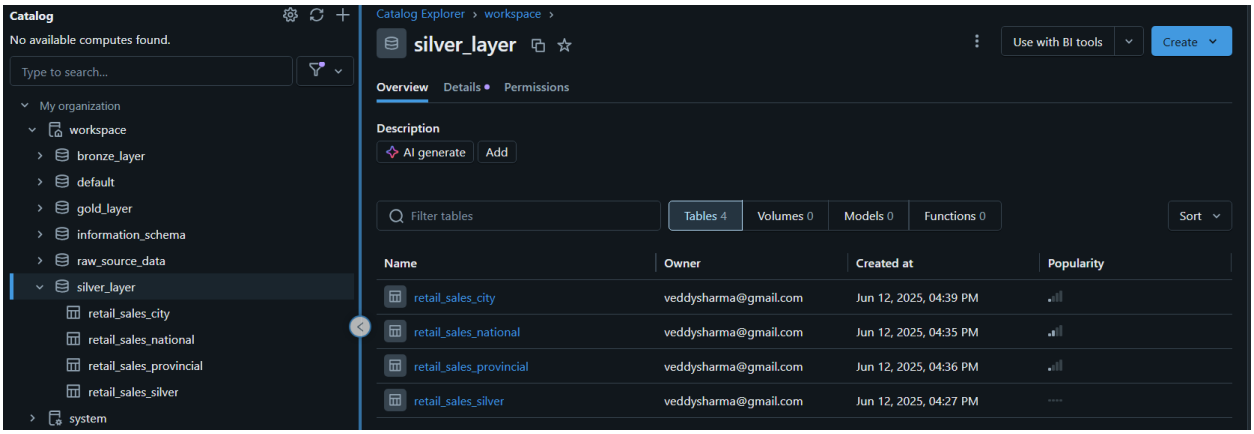
### Submitted Source Code:

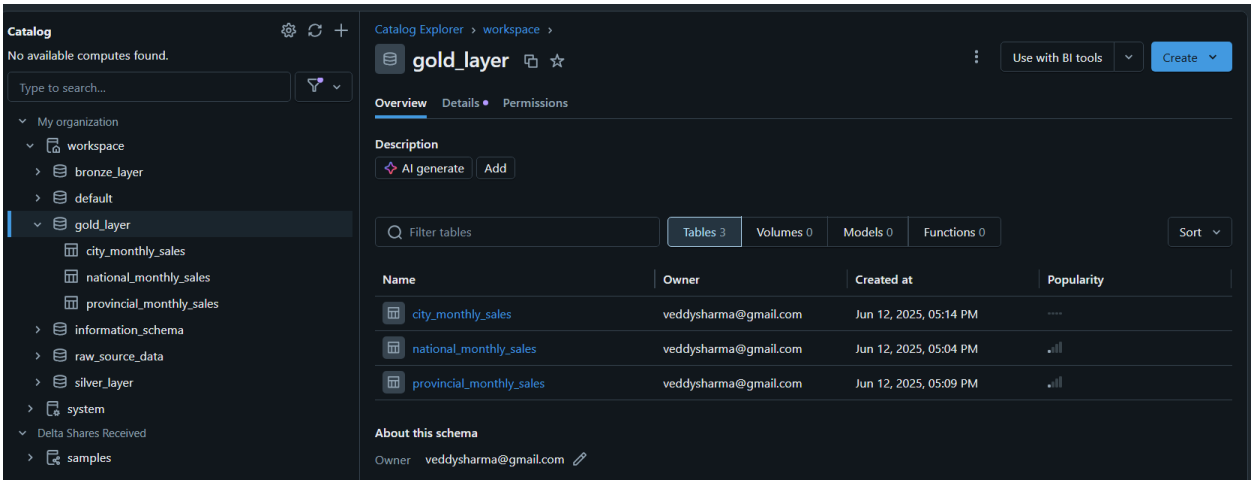
Please find all the required executed notebooks in the Learning Hub Assignment folder, containing 2 different files for the Bronze, Silver, and Gold layers.

### Catalog Screenshots:

Below are the catalog screenshots displaying the Schemas, and the tables created along the way. Please find them below:




Catalog
No available computes found.
Type to search...
My organization
workspace
bronze_layer
default
gold_layer
information_schema
raw_source_data
silver_layer
retail_sales_city
retail_sales_national
retail_sales_provincial
retail_sales_silver
system
Catalog Explorer › workspace ›
silver_layer
Use with BI tools
Create
Overview
Details
Permissions
Description
AI generate
Add
Filter tables
Tables 4
Volumes 0
Models 0
Functions 0
Sort
Name
Owner
Created at
Popularity
retail_sales_city
veddysharma@gmail.com
Jun 12, 2025, 04:39 PM
retail_sales_national
veddysharma@gmail.com
Jun 12, 2025, 04:35 PM
retail_sales_provincial
veddysharma@gmail.com
Jun 12, 2025, 04:36 PM
retail_sales_silver
veddysharma@gmail.com
Jun 12, 2025, 04:27 PM


Catalog
No available computes found.
Type to search...
My organization
workspace
bronze_layer
default
gold_layer
city_monthly_sales
national_monthly_sales
provincial_monthly_sales
information_schema
raw_source_data
silver_layer
system
Delta Shares Received
samples
Catalog Explorer › workspace ›
gold_layer
Use with BI tools
Create
Overview
Details
Permissions
Description
AI generate
Add
Filter tables
Tables 3
Volumes 0
Models 0
Functions 0
Sort
Name
Owner
Created at
Popularity
city_monthly_sales
veddysharma@gmail.com
Jun 12, 2025, 05:14 PM
national_monthly_sales
veddysharma@gmail.com
Jun 12, 2025, 05:04 PM
provincial_monthly_sales
veddysharma@gmail.com
Jun 12, 2025, 05:09 PM
About this schema
Owner
veddysharma@gmail.com

# 3. Dashboard Implementation:


### Canada-Wide Retail Sales Trend (2017–2025)

This line chart is designed for internal analysts and management to understand national retail trends over time. It helps them identify market fluctuations, overall volatility, and post COVID recovery phases. The line format was chosen because it clearly shows how sales evolve month by month across several years.

**Traditional vs E-Commerce Sales Proportion**

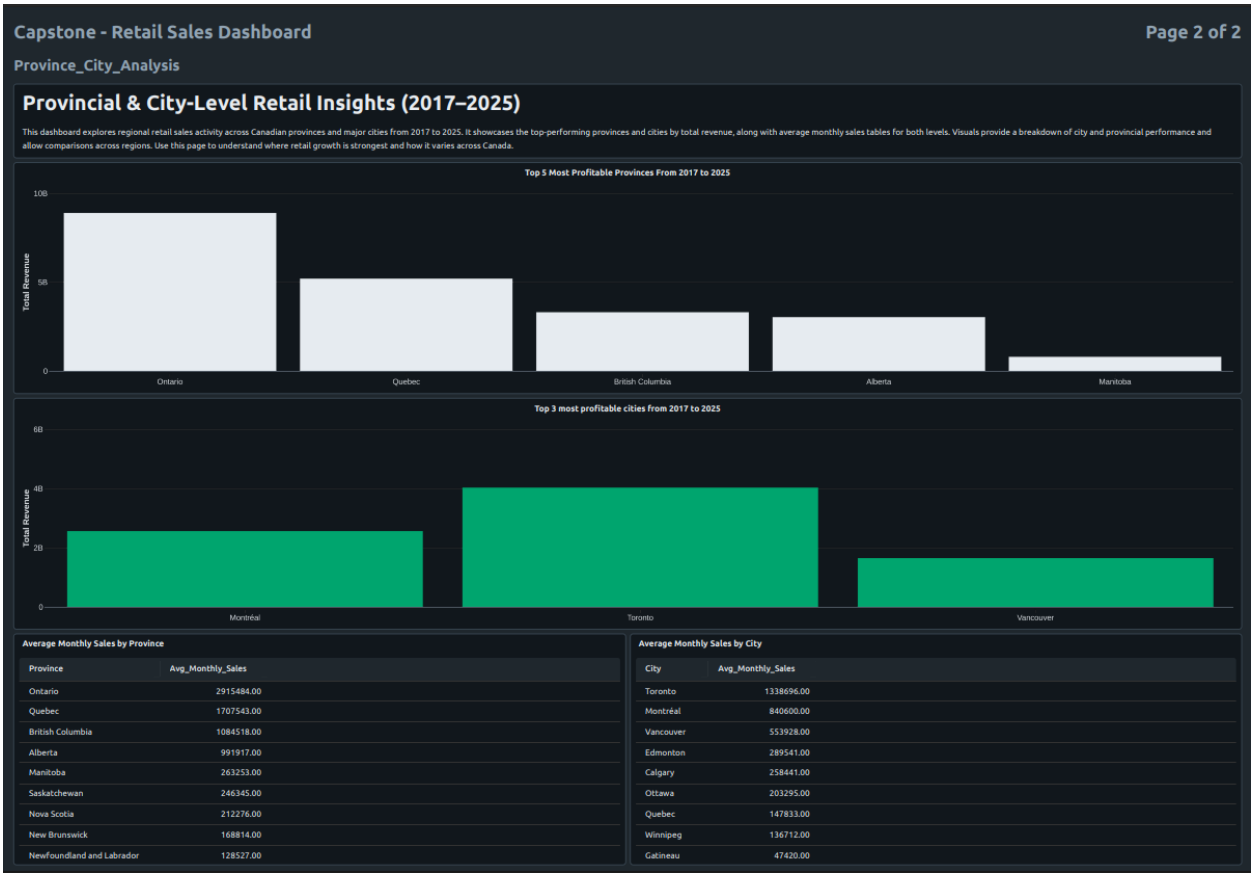
This donut chart targets strategy and digital teams, offering a quick comparison between traditional and e-commerce sales. It shows the dominance of brick-and-mortar retail and elucidates and clearly outlines that how small the e-commerce share still is. The donut format makes proportions immediately clear.

**Top 3 Performing Retail Sectors Nationwide**

This bar chart is useful for investment analysts and business leads looking to identify which sectors drive the most revenue. It supports targeting high-performing areas for future investment. A bar chart was used because it makes total sales comparisons across categories simple and visually obvious.

**Average Monthly Sales by Retail Segment**

This table is intended for finance teams and decision-makers who need precise, numerical insights. It shows the average monthly sales across top sectors, helping assess long-term performance and consistency. A table was chosen to present exact figures that would be hard to interpret through a visual chart alone.




### Top 5 Most Profitable Provinces (2017–2025)

This bar chart is designed for regional investment teams to identify where retail growth is most concentrated across Canada. It supports allocation of resources to high-performing provinces. A bar chart was selected to clearly show revenue gaps between regions.

### Top 3 Most Profitable Cities (2017–2025)

City-level executives and urban planners can use this chart to benchmark performance across major metropolitan areas. The visual highlights revenue leaders like Toronto and Montréal. Bar format ensures clean city-by-city comparison.

### Average Monthly Sales by Province

This table is useful for internal finance and regional strategists who require specific monthly performance data. It helps compare sustained average sales over time. A table format allows for easy extraction of exact figures.

### Average Monthly Sales by City

This supports city-level sales and investment teams in tracking average profitability per urban center. The table format ensures stakeholders can access numeric values precisely, without visual ambiguity.

# 4. Pipeline Analysis:

### 1. Is a Data Lakehouse Appropriate?

In all honestly, a traditional Data Warehouse would be a better fit for this project. In this project, we are primarily working with structured data in CSV format, and we're also ingesting newer data every quarter. Due to a lack of unstructured data and even a lack of involvement of streaming data, a data warehouse would be more suitable. The pipeline created is relatively simple, and most of work occurs in the silver layer. As a result, we think that a traditional data warehouse would be more appropriate for this project

### 2. Frequency & Business Logic for Data Ingest

The primary dataset used is updated monthly by Stats Canada. However, for private equity, we would ingest it quarterly, since most investment decisions are based on quarterly trends. Moving on, if we were to add company specific data, demographics data, or consumer preferences data, we would still ingest them quarterly. As per the business logic, it is straightforward. Bring in newer data, clean it, and update the gold tables.

### 3. Handling Batch vs Streaming Data in Databricks

- Batch data would be ingested on a schedule using Databricks jobs or workflows. You read the files, clean them, and update the tables — usually on a weekly or monthly schedule.
- Streaming data (like real-time POS transactions or web traffic) would use Structured Streaming in Databricks. It would continuously ingest data using sources like Kafka or Auto Loader, write it into the bronze layer in micro-batches, and then trigger updates to silver/gold layers with incremental logic.