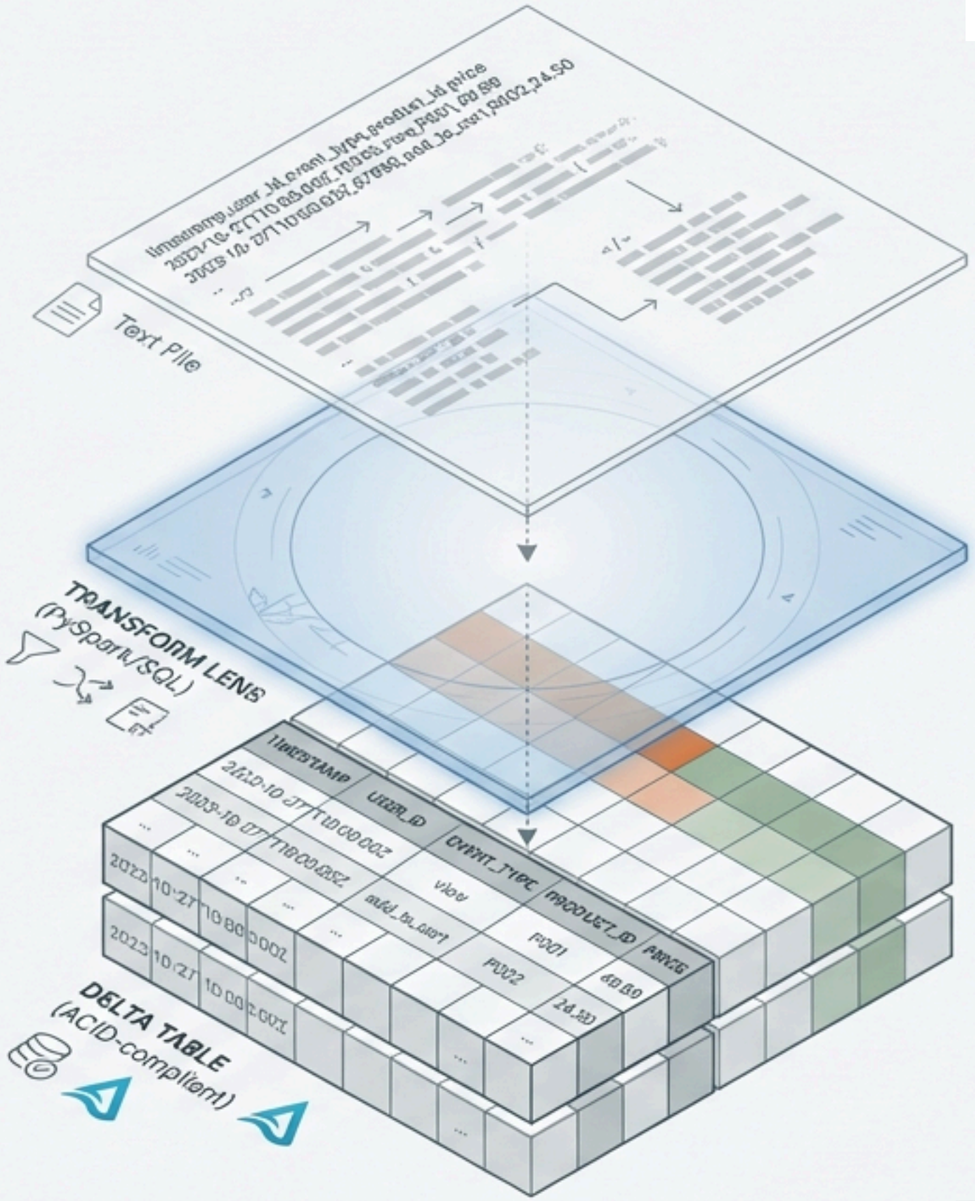
Building Lakehouse Architecture

A Databricks Delta Walkthrough

From raw CSV ingestion to ACID-compliant merge operations.

This workflow transforms eCommerce behavior logs into governed, high-performance Delta tables using PySpark and SQL.



The Objective: Structuring Raw eCommerce Logs

The Context

Core Challenge

Handling 2019 eCommerce data (October/November datasets) containing high-volume user behavior logs.

The Goal

- Schema enforcement
- High-performance querying
- Data consistency (handling duplicates)

The Data Profile

Data Card

Dataset Fields & Specs

```
event_time (Timestamp)
event_type (view / purchase)
product_id (Integer)
category_code (String)
brand (String)
price (Double)
```

2019 Activity

Approx. 10M+ records

Ingestion: Moving from Raw CSV to Delta Format

The Insight

Intelligent Read

The workflow begins by reading **raw files** from the volume path. The argument "inferSchema=True" **automatically detects types** (Integers, Timestamps), eliminating manual DDL.

Optimized Write

Data is immediately **serialized** into the '.format("delta")'. This converts inefficient row-based CSVs into **columnar storage optimized for analytics**.

The Action

```
%python
# 1. Convert CSV to Delta format
df_oct = spark.read.csv(
    "/Volumes/workspace/ecommerce/ecommerce_data/2019-Oct.csv",
    header=True,
    inferSchema=True
)

df_oct.write.format("delta").mode("overwrite").save(
    "/Volumes/workspace/ecommerce/ecommerce_data/delta/2019-Oct"
)

df_nov = spark.read.csv(
    "/Volumes/workspace/ecommerce/ecommerce_data/2019-Nov.csv",
    header=True,
    inferSchema=True
)

df_nov.write.format("delta").mode("overwrite").save(
    "/Volumes/workspace/ecommerce/ecommerce_data/delta/2019-Nov"
)

display(df_oct)
display(df_nov)
```

Inspection: Verifying the eCommerce Dataset

Validating the in-memory dataframe structure before commitment.

event_time	event_type	product_id	category_id	category_code	brand	price
2019-10-01 00:00:00 UTC	view	1480613	2053013556109286677	computers.desktop	pulser	908.62
2019-10-01 00:00:05 UTC	view	17300353	2053013553853497655	null	creed	380.96
2019-10-01 00:00:08 UTC	view	31500053	2053013553853497655	null	luminarc	41.16
2019-10-01 00:00:10 UTC	view	28719074	2053013565480109009	apparel.shoes.keds	baden	102.71
2019-10-01 00:00:11 UTC	view	1004545	2053013555631882655	electronics.smartphone	huawei	566.01
2019-10-01 00:00:11 UTC	view	2900536	2053013554776244510	appliances.kitchen.microwave	elenberg	51.46
2019-10-01 00:00:13 UTC	view	3900746	2053013552326770905	appliances.environment.water_heater	haier	102.38
2019-10-01 00:00:15 UTC	view	44600062	2103807459595387724	null	shiseido	35.79
2019-10-01 00:00:16 UTC	view	13500240	2053013557099889147	furniture.bedroom.bed	brw	93.18
2019-10-01 00:00:17 UTC	view	23100006	2053013561638126333	null	null	357.79

Definition: Formalizing Data into Managed Tables

Method A: PySpark API

```
df_oct.write.mode("overwrite").saveAsTable("default.ecommerce_2019_oct")
```

Directly registers the dataframe to the Hive metastore.

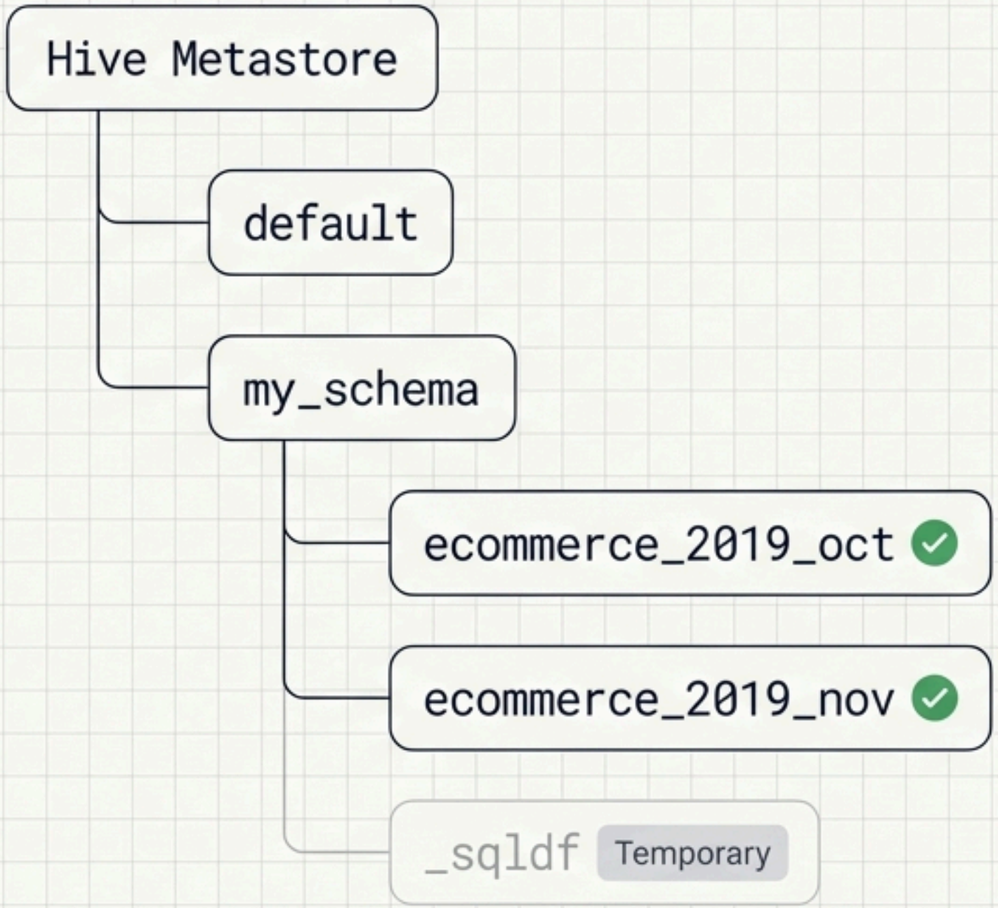
VS

Method B: Standard SQL

```
CREATE TABLE IF NOT EXISTS my_schema.ecommerce_2019_oct
USING DELTA
AS SELECT * FROM default.ecommerce_2019_oct
```

Declarative table creation ensuring idempotency and explicit format definition.

Discovery: Navigating the Data Catalog



databaseName
default
ecommerce
my_schema

database	tableName	isTemporary
my_schema	ecommerce_2019_nov	false
my_schema	ecommerce_2019_oct	false
...	_sqldf	true

Governance: Validating Schema and Types

Schema Blueprint

Field Name	Data Type	Role
event_time	timestamp	Time-series partitioning base
product_id	int	Primary Key candidate
category_id	bigint	Large integer handling
price	double	High-precision currency
brand	string	Categorical attribute

Narrative Insight

Strict Typing: This metadata layer enables Delta Lake's Schema Enforcement, automatically rejecting data that violates these definitions.

Integrity: Testing Schema Enforcement

Experiment: Code Execution

Deliberate Type Mismatch (String vs Double)

```
INSERT INTO my_schema.ecommerce_2019_oct (product_id, price)
SELECT 12345, try_cast('not_a_number' AS DOUBLE)
```

Graceful Failure Handling

Result: Data Ingestion

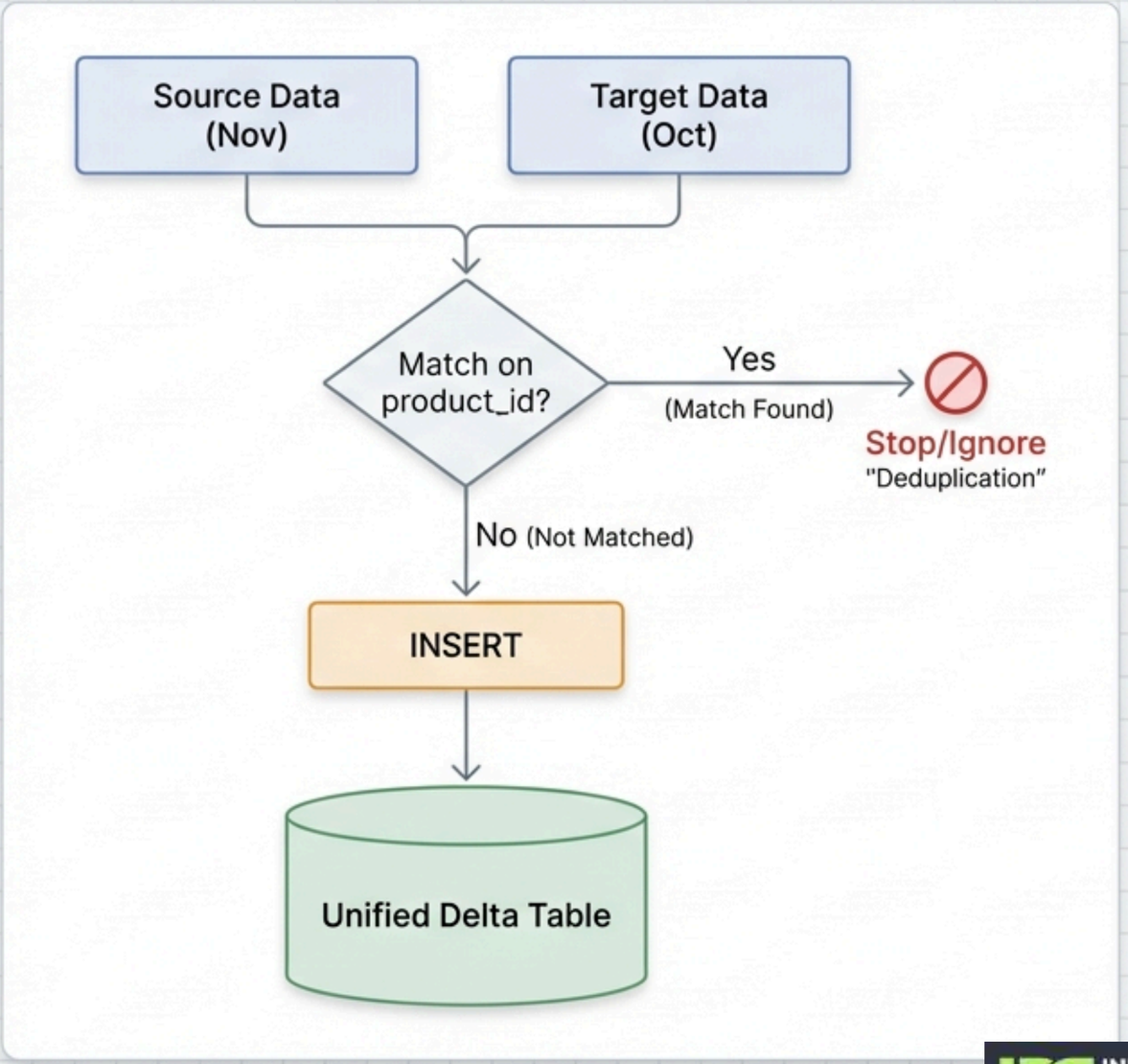
product_id	price
12345	null

Result: The system enforces the schema. Since 'not_a_number' cannot be cast to a Double, it inserts a NULL value rather than corrupting the column or breaking the pipeline.

Swiss Architectural Technical

Consistency: Implementing Conditional Merges

```
MERGE INTO ...target
  USING new_data AS source
  ON target.product_id = source.product_id
WHEN NOT MATCHED THEN INSERT *
```



Swiss Architectural Technical

Validation: Confirming the Pipeline Output

The final merged state containing unified history and new streams.

event_time	event_type	product_id	category_code	brand	price	user_id
2019-10-01T00:00:00.000+00:00	view	44600062	null	shiseido	35.79	541312140
2019-10-01T00:00:00.000+00:00	view	3900821	appliances.environment.water_heater	aqua	33.20	554748717
2019-10-01T00:00:01.000+00:00	view	17200506	furniture.living.room.sofa	null	543.10	519107250
2019-10-01T00:00:01.000+00:00	view	1307067	computers.notebook	lenovo	251.74	550050854
2019-10-01T00:00:04.000+00:00	view	1004237	electronics.smartphone	apple	1081.98	535871217
2019-10-01T00:00:05.000+00:00	view	1480613	computers.desktop	pulser	908.62	512742880
2019-10-01T00:00:08.000+00:00	view	17300353	null	creed	380.96	555447699
2019-10-01T00:00:08.000+00:00	view	31500053	null	lumincarc	41.16	550978835
2019-10-01T00:00:10.000+00:00	view	28719074	apparel.shoes.keds	baden	102.71	520571932
2019-10-01T00:00:11.000+00:00	view	1004545	electronics.smartphone	huawei	566.01	537918940
2019-10-01T00:00:11.000+00:00	view	2900536	appliances.kitchen.microwave	elenberg	51.46	555158050
2019-10-01T00:00:11.000+00:00	view	1005011	electronics.smartphone	samsung	900.64	530282093
2019-10-01T00:00:13.000+00:00	view	3900746	appliances.environment.water_heater	haier	102.38	555444559
2019-10-01T00:00:15.000+00:00	view	44600062	null	shiseido	35.79	541312140

System Status
Status: OK
Command took: 15.49s
Rows: 18,000+ (Truncated)

Swiss Architectural Technical

The Delta Advantage: Recap

Intelligent Ingestion

Automated schema inference from CSV to memory.

Unified Architecture

Interoperable Python and SQL management.

Data Governance

Explicit typing and strict schema enforcement.

Transactional Integrity

ACID-compliant merges for deduplication.

Result: A trusted Lakehouse foundation ready for analytics.