Hairvana
Cloud Data Pipeline

Shuhan Chang
Chemseddine Nadour
Cyprien Cazenave
Mathieu Chabirand
Vanessa Gunathasan

Agenda

- [ ] Introduction
- [ ] Cloud Data Pipeline
- [ ] Business Impact
- [ ] Conclusion

[Presenter's First Name]

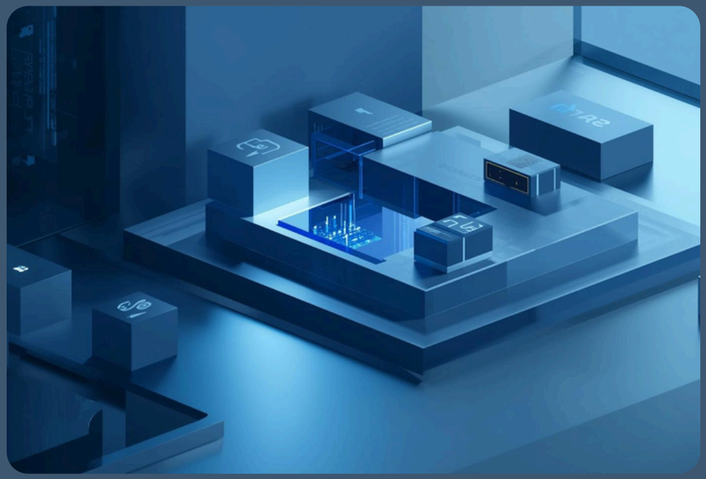
Data Ingestion and Storage

Sources

Integrating diverse data sources enhances insights for Hairvana's haircare solutions.

Data Lake

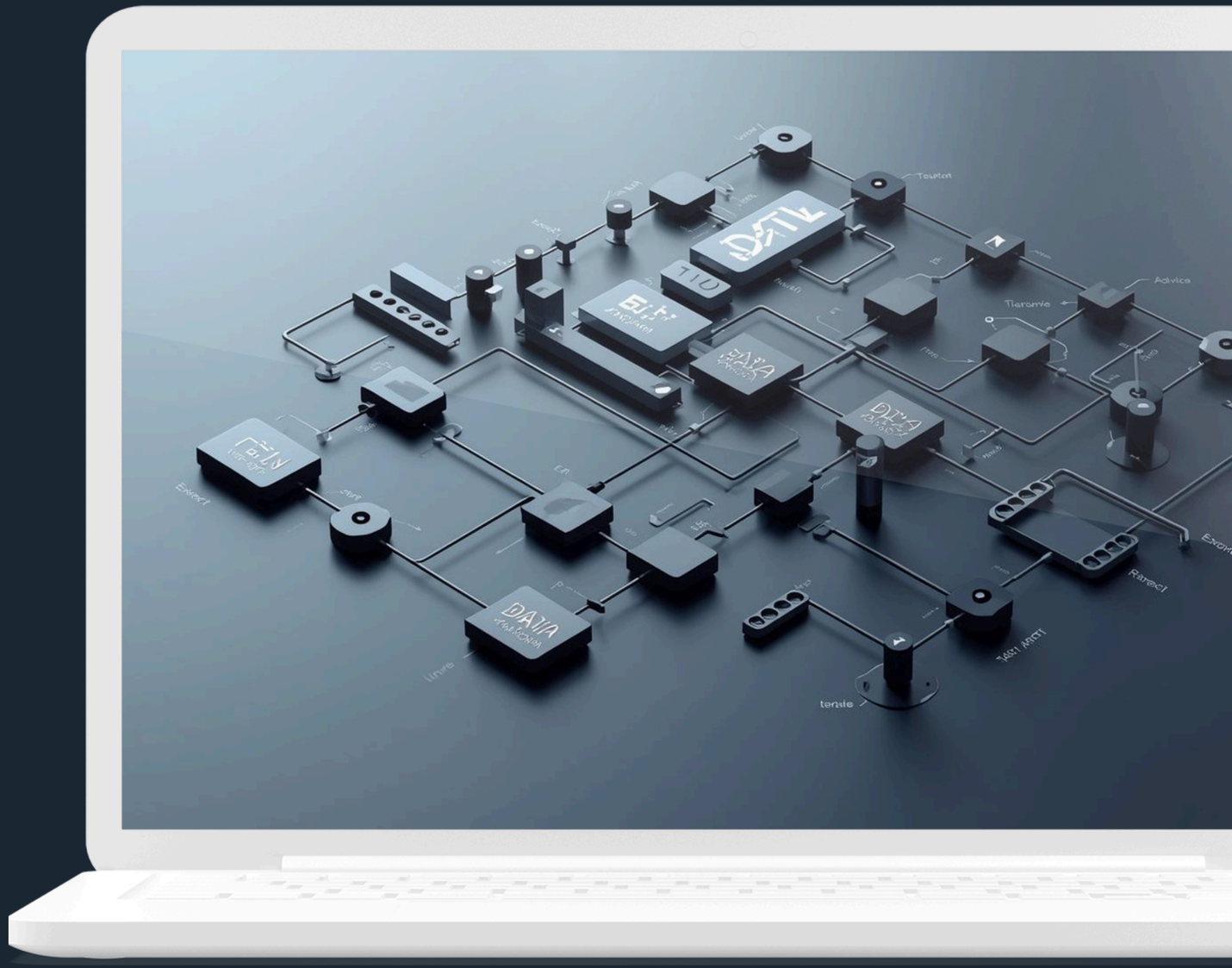
Centralized storage streamlines access and management of vast customer data.



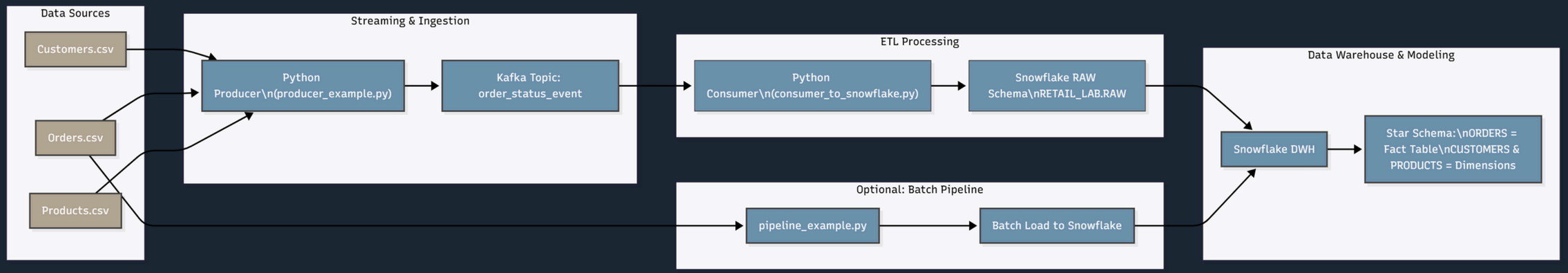
DATA PROCESSING

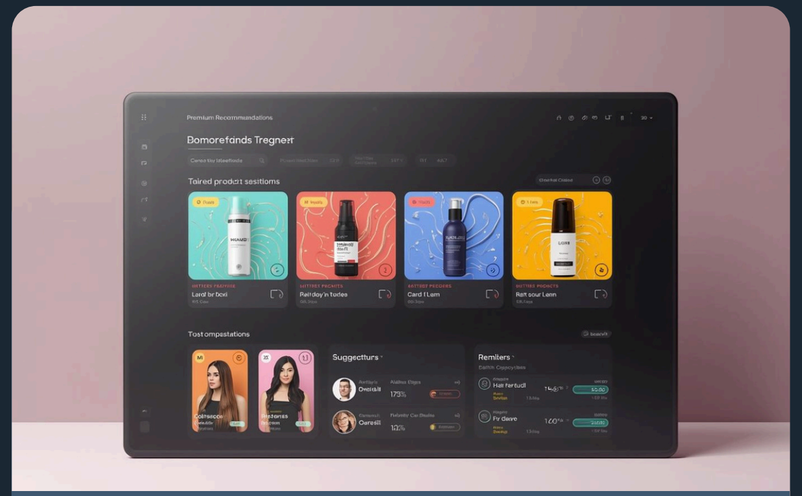
ETL and ELT Explained

This section explores the **data transformation pipeline**, detailing the processes of ETL and ELT that enable Hairvana to derive actionable insights from vast amounts of data.



Our Pipeline Architecture



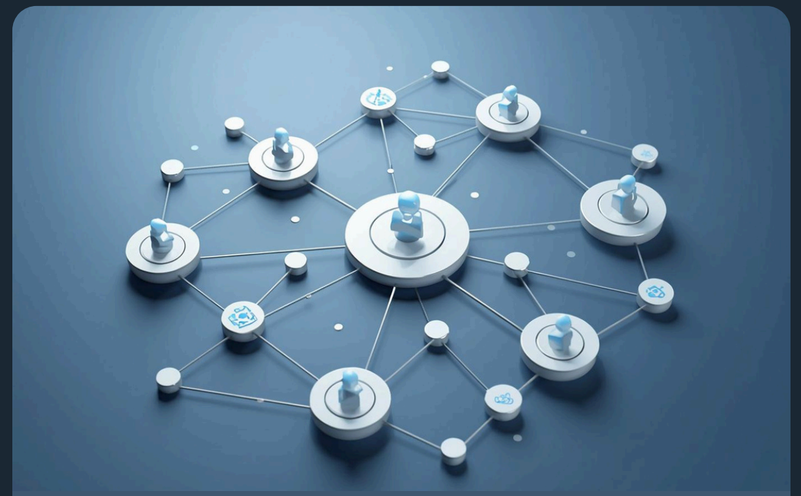


Tailored Recommendations for Individual Customers

Discover Your Perfect Match

Streamlined Inventory for Enhanced Efficiency

Optimize Your Stock Levels



Data-Driven Experience for Every Customer

Elevate Your Service Today

Real-time Insights for Premium Haircare

Driving Marketing Efficiency

Leverage real-time data to tailor marketing strategies and optimize campaign performance for maximum impact.

Enhancing Product Development

Utilize data-driven insights to refine product offerings and develop new formulations that meet customer needs.

Spotting Future Trends

Implement predictive analytics to anticipate market shifts and adapt strategies for emerging customer preferences.

Data-Driven Growth

Foster a culture of data-driven decision making to continuously improve operations and boost overall business performance.

Data Structures

Data Structures

After processing, raw data is organized into clean, structured tables within our Snowflake data warehouse. This makes it easy to query and analyze. Select a data model below to see its structure.

ORDERS | CUSTOMERS | PRODUCTS

ORDERS

This is the main fact table containing all transactional data. It is optimized for analytical queries on sales performance and trends.

Column Name	Data Type	Description
ORDER_ID	INTEGER	Unique identifier for each order.
CUSTOMER_ID	INTEGER	Foreign key to the Customers table.
PRODUCT_ID	INTEGER	Foreign key to the Products table.
QUANTITY	INTEGER	Number of units sold.
ORDER_TOTAL	DECIMAL(10,2)	Total value of the order.
ORDER_STATUS	VARCHAR	Current status of the order (e.g., Shipped, Delivered).
SOLD_AT	TIMESTAMP	Date and time of the transaction.
REGION	VARCHAR	Market region where the order occurred.

CUSTOMERS

This dimension table holds all information about our customers, allowing for segmentation and behavioral analysis.

Column Name	Data Type	Description
CUSTOMER_ID	INTEGER	Unique identifier for each customer.
NAME	VARCHAR	Full name of the customer.
EMAIL	VARCHAR	Customer's contact email.
JOIN_DATE	DATE	Date the customer first registered.
ADDRESS	VARCHAR	Street address.
CITY	VARCHAR	City of residence.
COUNTRY	VARCHAR	Country of residence.
POSTAL_CODE	VARCHAR	Postal code.
REGION	VARCHAR	Customer's market region (e.g. Europe, Asia, Americas).

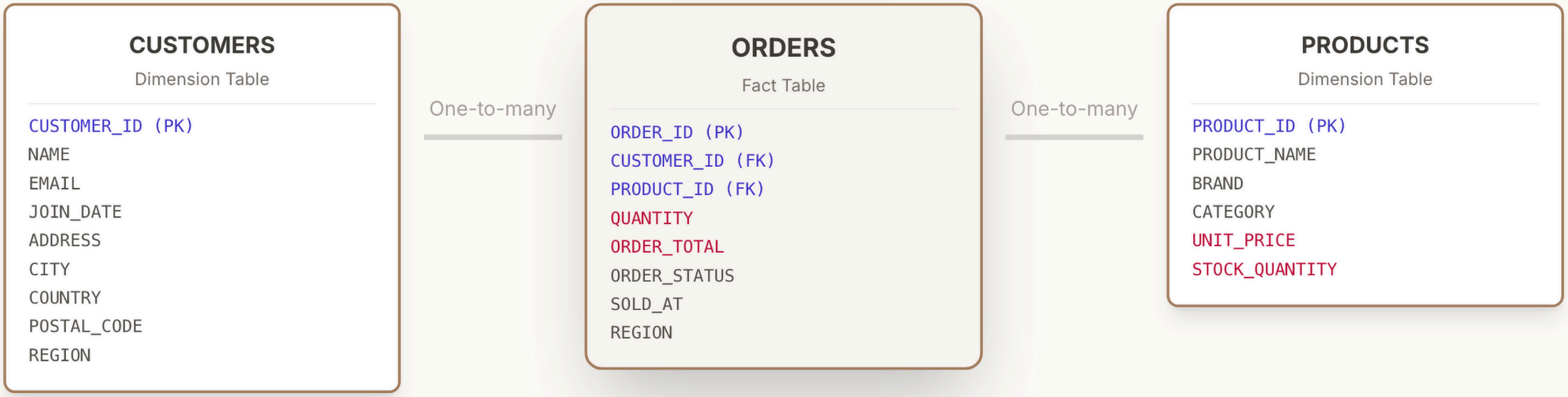
PRODUCTS

This dimension table contains details for every product in our catalog, used for inventory and sales analysis.

Column Name	Data Type	Description
PRODUCT_ID	INTEGER	Unique identifier for each product.
PRODUCT_NAME	VARCHAR	Name of the product.
BRAND	VARCHAR	Brand name of the product.
CATEGORY	VARCHAR	Product category (e.g., Shampoo, Serum).
UNIT_PRICE	DECIMAL(10,2)	Retail price of one unit.
STOCK_QUANTITY	INTEGER	Current number of units in stock.

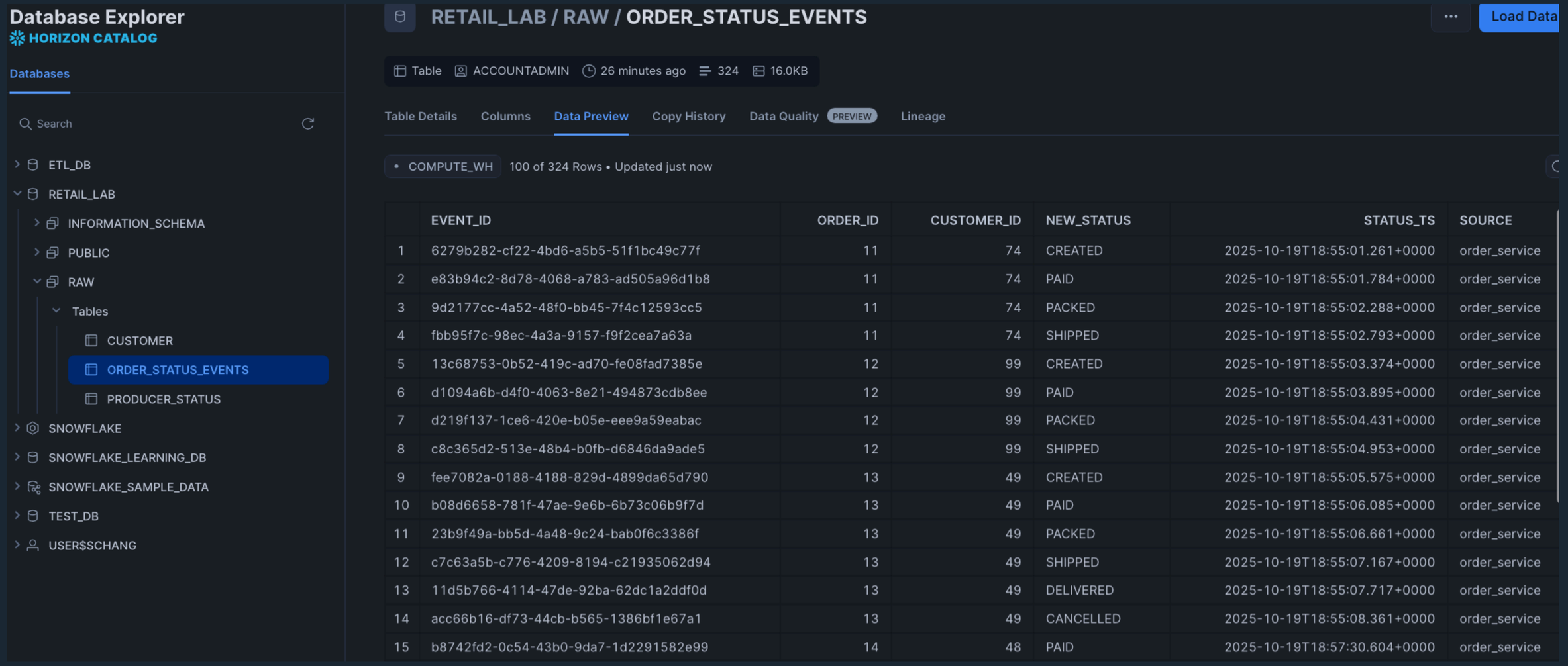
Star Model

We model our data using a star schema. This design features a central "fact" table (`ORDERS`) containing quantitative data, connected to descriptive "dimension" tables. This structure simplifies queries and improves performance for reporting and analytics.



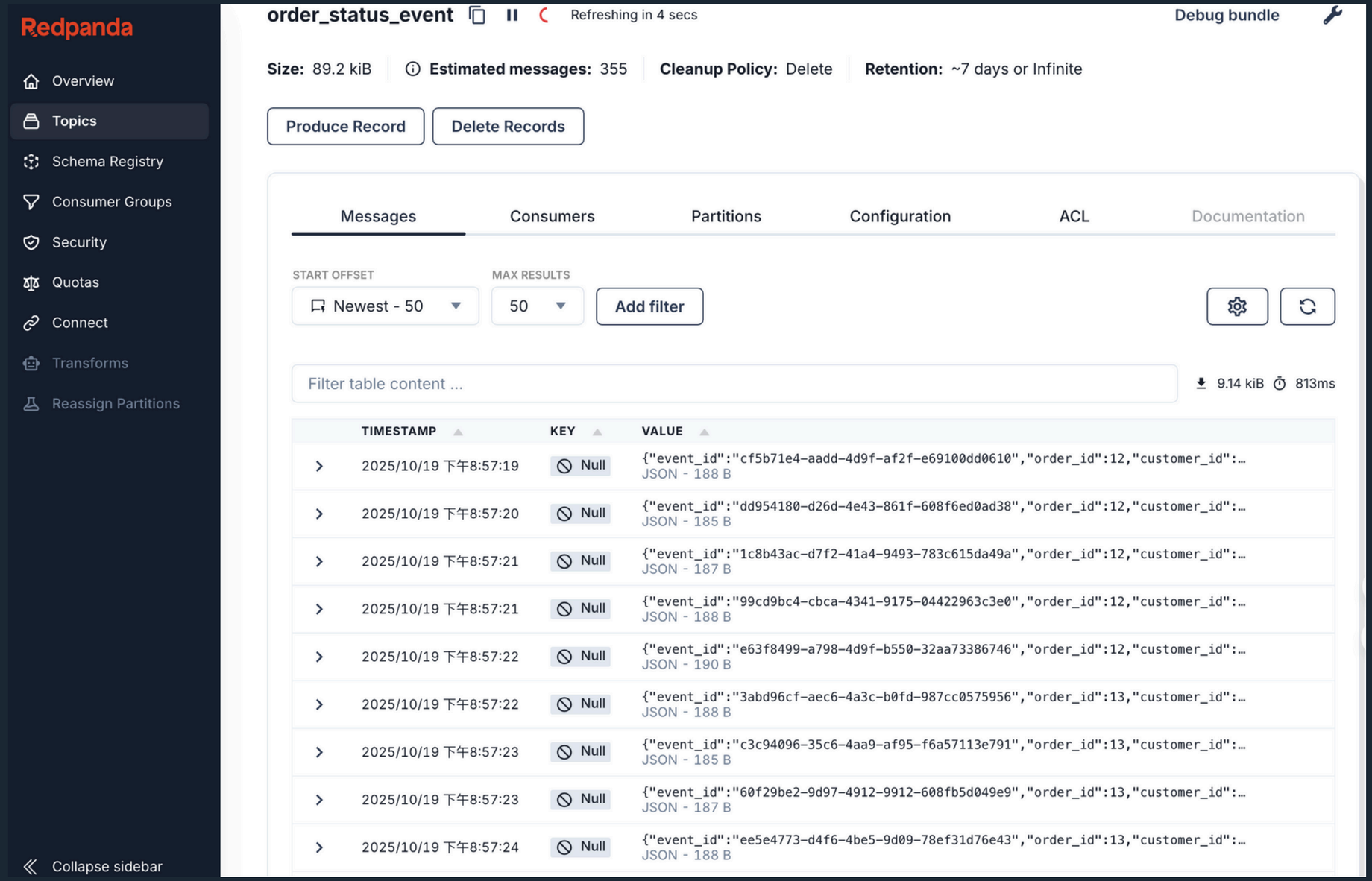
Git hub
https://github.com/shuhanchang12/ETL_ecommerce

Screenshots of snowflake tables to validate that you ingested the data.

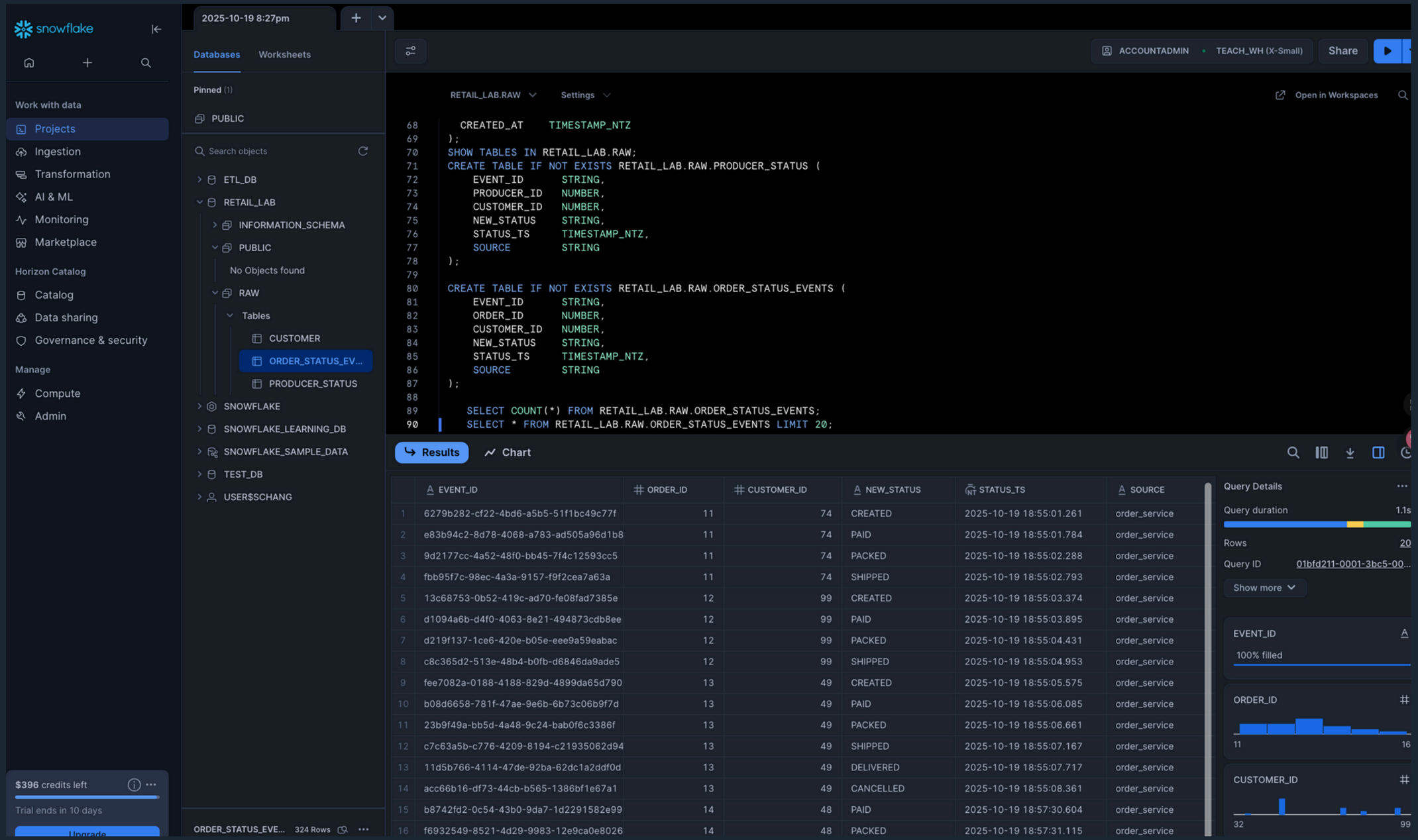


	EVENT_ID	ORDER_ID	CUSTOMER_ID	NEW_STATUS	STATUS_TS	SOURCE
1	6279b282-cf22-4bd6-a5b5-51f1bc49c77f	11	74	CREATED	2025-10-19T18:55:01.261+0000	order_service
2	e83b94c2-8d78-4068-a783-ad505a96d1b8	11	74	PAID	2025-10-19T18:55:01.784+0000	order_service
3	9d2177cc-4a52-48f0-bb45-7f4c12593cc5	11	74	PACKED	2025-10-19T18:55:02.288+0000	order_service
4	fbb95f7c-98ec-4a3a-9157-f9f2cea7a63a	11	74	SHIPPED	2025-10-19T18:55:02.793+0000	order_service
5	13c68753-0b52-419c-ad70-fe08fad7385e	12	99	CREATED	2025-10-19T18:55:03.374+0000	order_service
6	d1094a6b-d4f0-4063-8e21-494873cdb8ee	12	99	PAID	2025-10-19T18:55:03.895+0000	order_service
7	d219f137-1ce6-420e-b05e-eee9a59eabac	12	99	PACKED	2025-10-19T18:55:04.431+0000	order_service
8	c8c365d2-513e-48b4-b0fb-d6846da9ade5	12	99	SHIPPED	2025-10-19T18:55:04.953+0000	order_service
9	fee7082a-0188-4188-829d-4899da65d790	13	49	CREATED	2025-10-19T18:55:05.575+0000	order_service
10	b08d6658-781f-47ae-9e6b-6b73c06b9f7d	13	49	PAID	2025-10-19T18:55:06.085+0000	order_service
11	23b9f49a-bb5d-4a48-9c24-bab0f6c3386f	13	49	PACKED	2025-10-19T18:55:06.661+0000	order_service
12	c7c63a5b-c776-4209-8194-c21935062d94	13	49	SHIPPED	2025-10-19T18:55:07.167+0000	order_service
13	11d5b766-4114-47de-92ba-62dc1a2ddf0d	13	49	DELIVERED	2025-10-19T18:55:07.717+0000	order_service
14	acc66b16-df73-44cb-b565-1386bf1e67a1	13	49	CANCELLED	2025-10-19T18:55:08.361+0000	order_service
15	b8742fd2-0c54-43b0-9da7-1d2291582e99	14	48	PAID	2025-10-19T18:57:30.604+0000	order_service

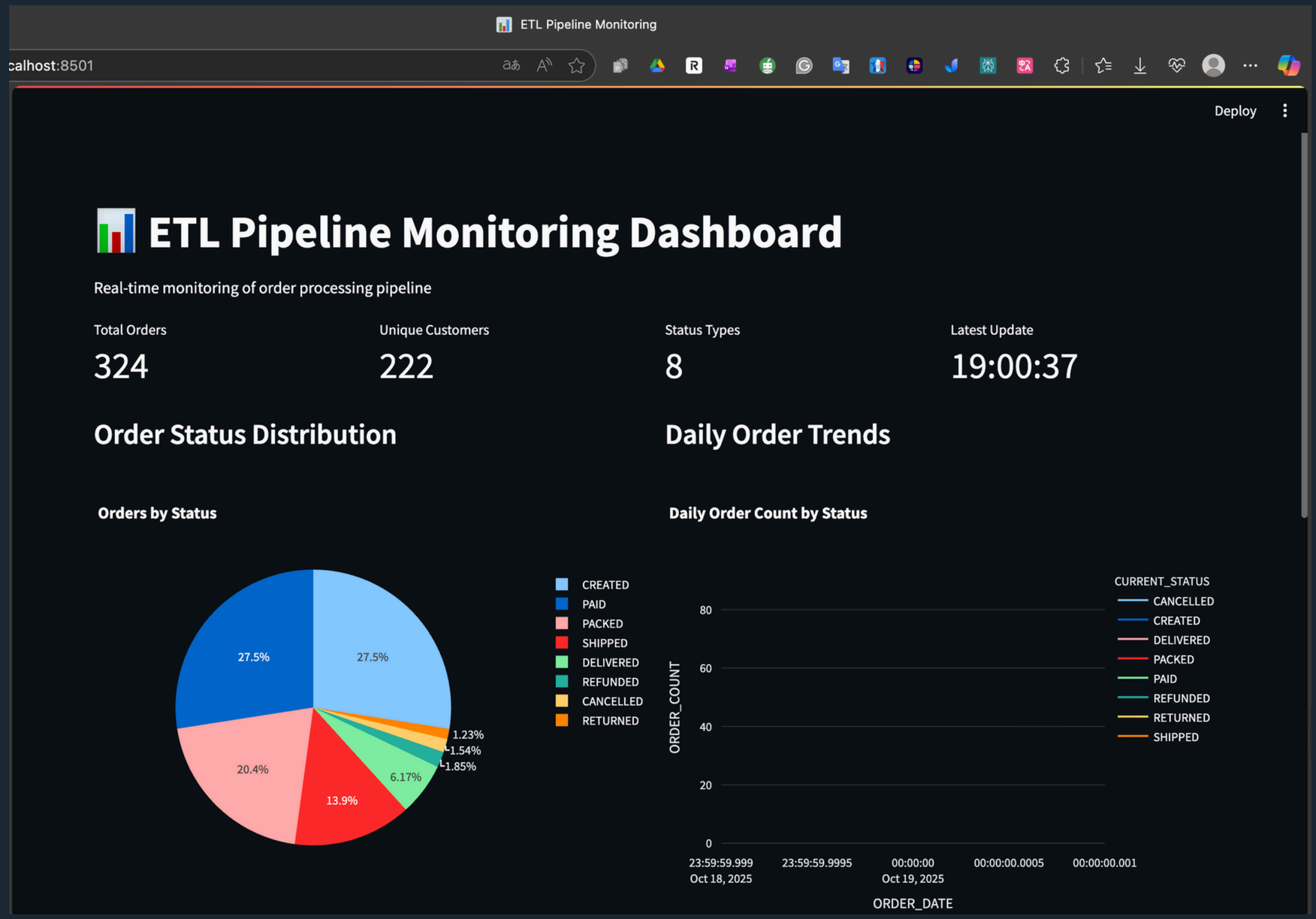
The data streaming with Redpanda-events created,



Ingestion in Snowflake



monitoring APP-Stramlit



Data Governance & Security

Protecting Hairvana's Data Assets with Enterprise-Grade Solutions

Access Control & Authentication

- Role-Based Access Control (RBAC) in Snowflake
- Multi-Factor Authentication (MFA)
- Least Privilege Principle enforced
- User access auditing
- Automatic credential rotation

Snowflake RBAC

OAuth 2.0

Encryption & Data Protection

- End-to-end encryption (TLS 1.3)
- Data at rest encrypted (AES-256)
- Sensitive data masking (PII)
- Anonymization for dev/test environments
- Secure key management (KMS)

AES-256

TLS 1.3

Data Quality & Lineage

- Automated validation of ingested data
- Complete traceability (source → destination)
- Data lineage visualized in Snowflake
- Quality checks at each pipeline stage
- Alerts for detected anomalies

Snowflake Streams

Tasks

Monitoring & Audit

- Centralized logs of all activities
- Real-time Kafka pipeline monitoring
- Automatic incident alerts
- 24/7 surveillance dashboards
- Quarterly audit reports

Snowflake Query History

Monitoring App

Hairvana complies with international data protection standards

Data Retention: 7 years for transactional data
Right to be Forgotten: Guaranteed deletion within 30 days
Backup: Daily snapshots with 90-day retention

Merci