E-Commerce Data Processing

Deep Patel

October 2024

Big Data Problem

- Handling large Volumes of data in Various form at high Velocity
- User Interaction log.
  - Approx.. Daily Log Size:
  - 1,000,000 users × 100 interactions/user = 100,000,000 entries/day
  - Total daily log size: 100,000,000 entries × 80 bytes/entry = 8,000,000,000 bytes
  - 8,000,000,000 bytes ÷ $1{,}024^2$ = 7,629 MB/day (approximately 7.63 GB/day)

Business Use Case

- Customer Activity by Region and Loyalty Tier
- Top Products by Category and Region
- Interaction Trend Over Time
- Historical Purchase Trends
- Customer Age and Interaction Analysis

Tech Stack

Apache Spark for batch processing.

Hive for metadata management.

HDFS for storage.

PostgreSQL for source data.

matplotlib, seaborn (visualization)

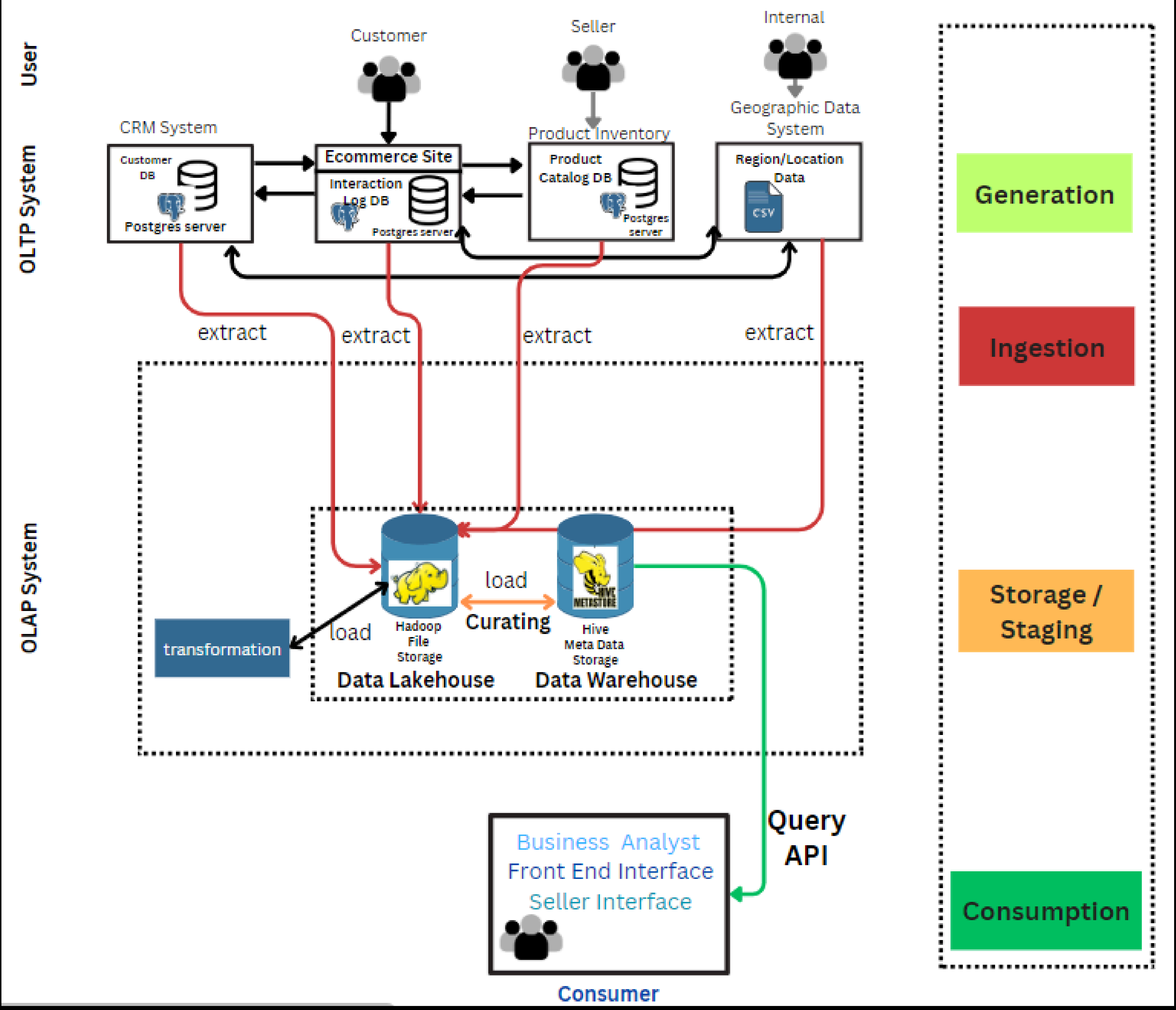
Key Integrations:

Batch processing

Query-based CDC

- Full and Incremental Loads.

Dimensional Modeling (Star Schema).



User

Customer

Seller

Internal

OLTP System

CRM System

Customer DB

Postgres server

Ecommerce Site

Interaction Log DB

Postgres server

Product Inventory

Product Catalog DB

Postgres server

Geographic Data System

Region/Location Data

CSV

extract

extract

extract

extract

OLAP System

transformation

load

Hadoop File Storage

load

Curating

Hive Meta Data Storage

Data Lakehouse

Data Warehouse

Query API

Business Analyst

Front End Interface

Seller Interface

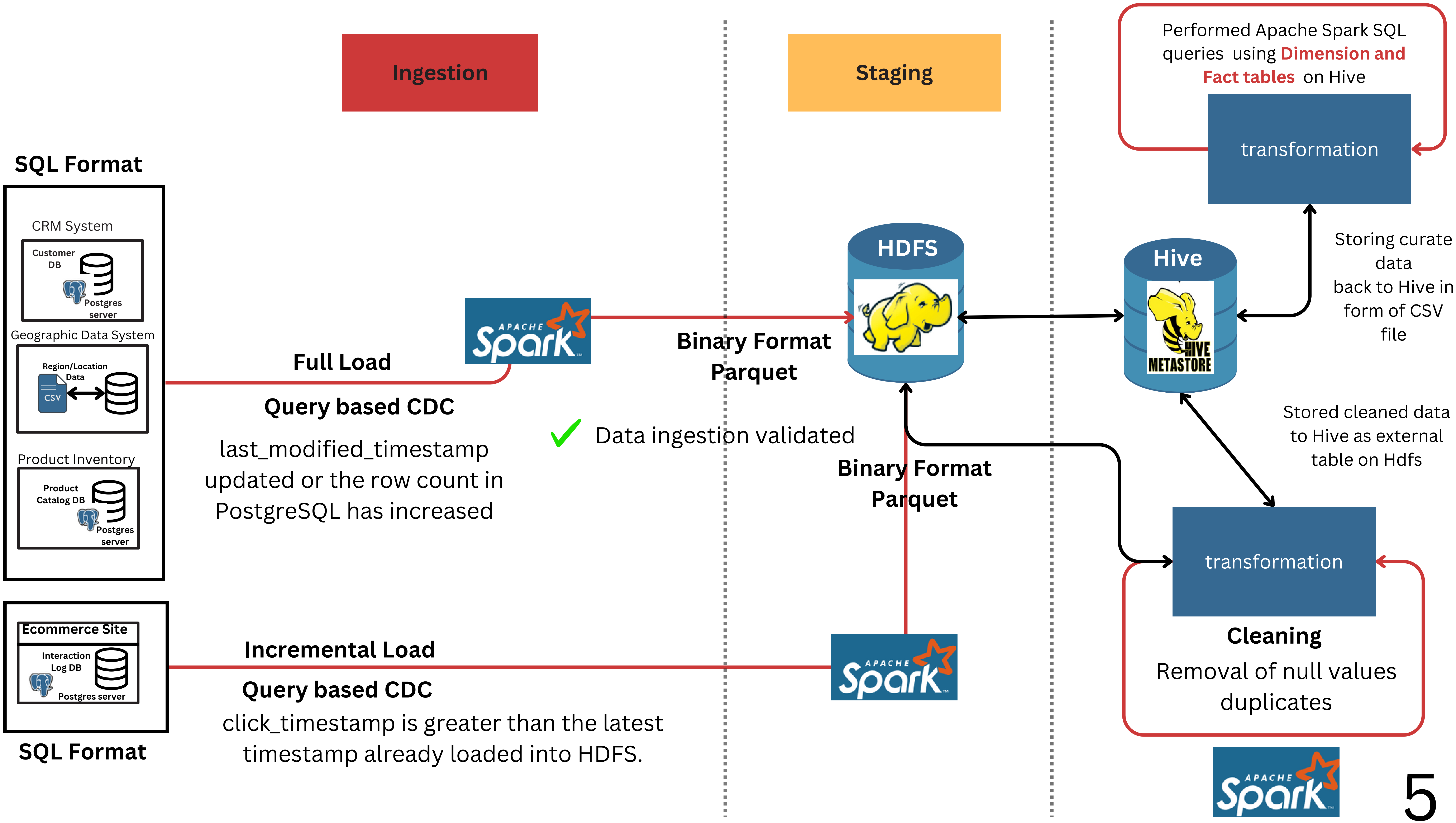
Consumer

Generation

Ingestion

Storage / Staging

Consumption



Ingestion

Staging

Performed Apache Spark SQL queries using **Dimension and Fact tables** on Hive

transformation

SQL Format

Full Load

Query based CDC

last_modified_timestamp updated or the row count in PostgreSQL has increased

Binary Format Parquet

✓ Data ingestion validated

Binary Format Parquet

Storing curate data back to Hive in form of CSV file

Stored cleaned data to Hive as external table on Hdfs

transformation

Cleaning

Removal of null values duplicates

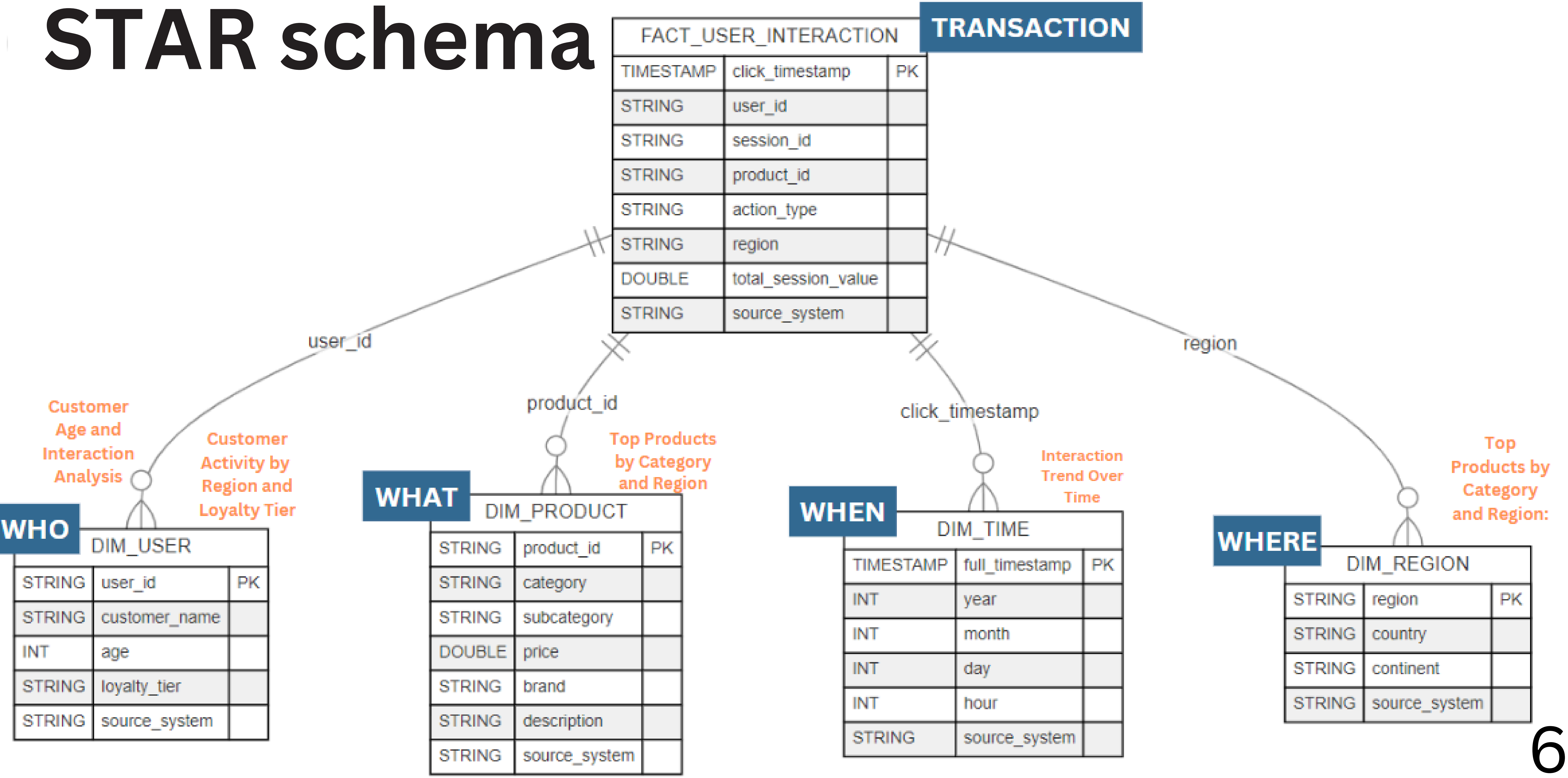
Incremental Load

Query based CDC

click_timestamp is greater than the latest timestamp already loaded into HDFS.

SQL Format

STAR schema



FACT_USER_INTERACTION

TIMESTAMP	click_timestamp	PK
STRING	user_id	
STRING	session_id	
STRING	product_id	
STRING	action_type	
STRING	region	
DOUBLE	total_session_value	
STRING	source_system	

Customer Age and Interaction Analysis

Customer Activity by Region and Loyalty Tier

DIM_USER

STRING	user_id	PK
STRING	customer_name	
INT	age	
STRING	loyalty_tier	
STRING	source_system	

Top Products by Category and Region

DIM_PRODUCT

STRING	product_id	PK
STRING	category	
STRING	subcategory	
DOUBLE	price	
STRING	brand	
STRING	description	
STRING	source_system	

Interaction Trend Over Time

DIM_TIME

TIMESTAMP	full_timestamp	PK
INT	year	
INT	month	
INT	day	
INT	hour	
STRING	source_system	

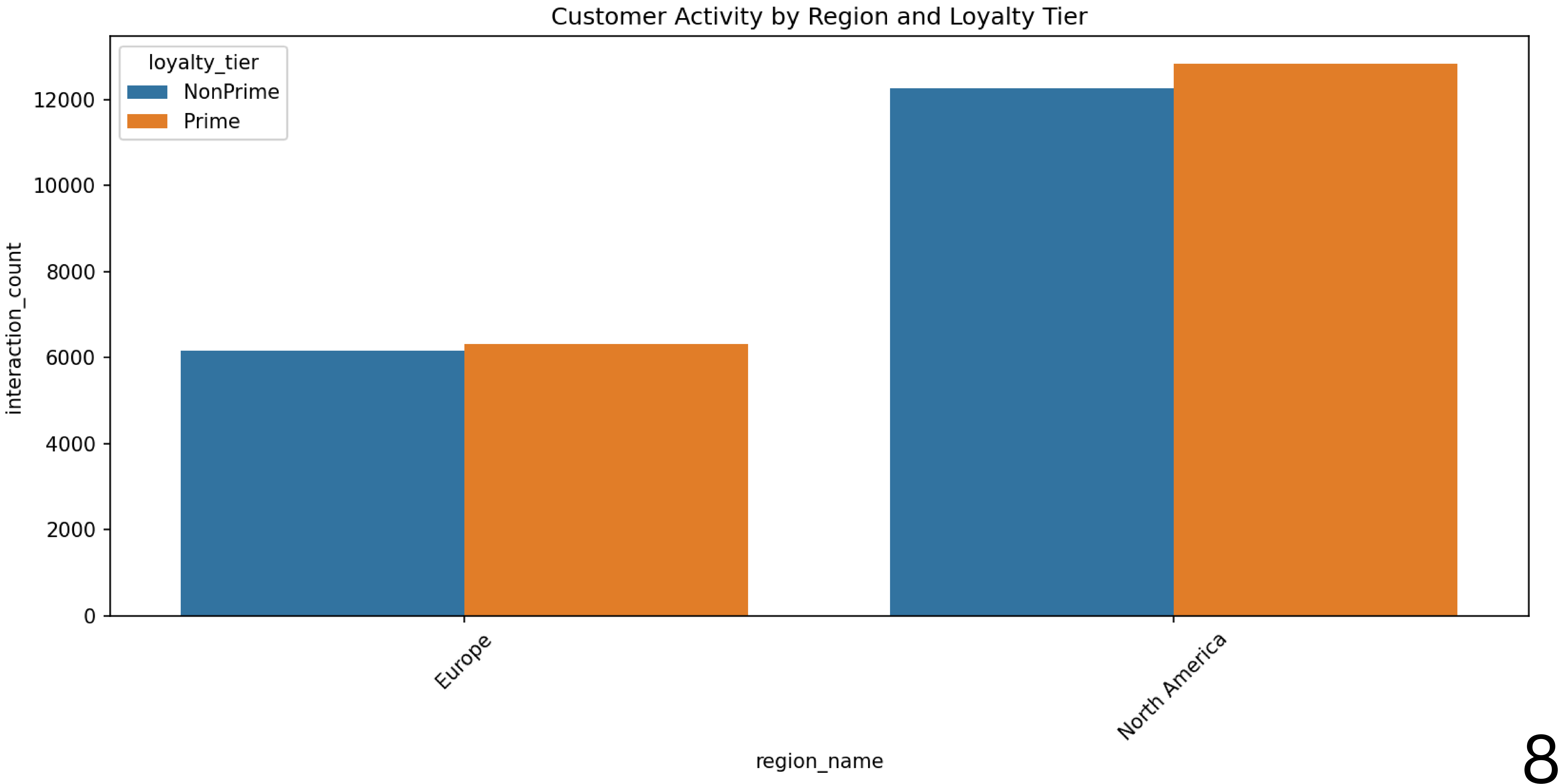
Top Products by Category and Region:

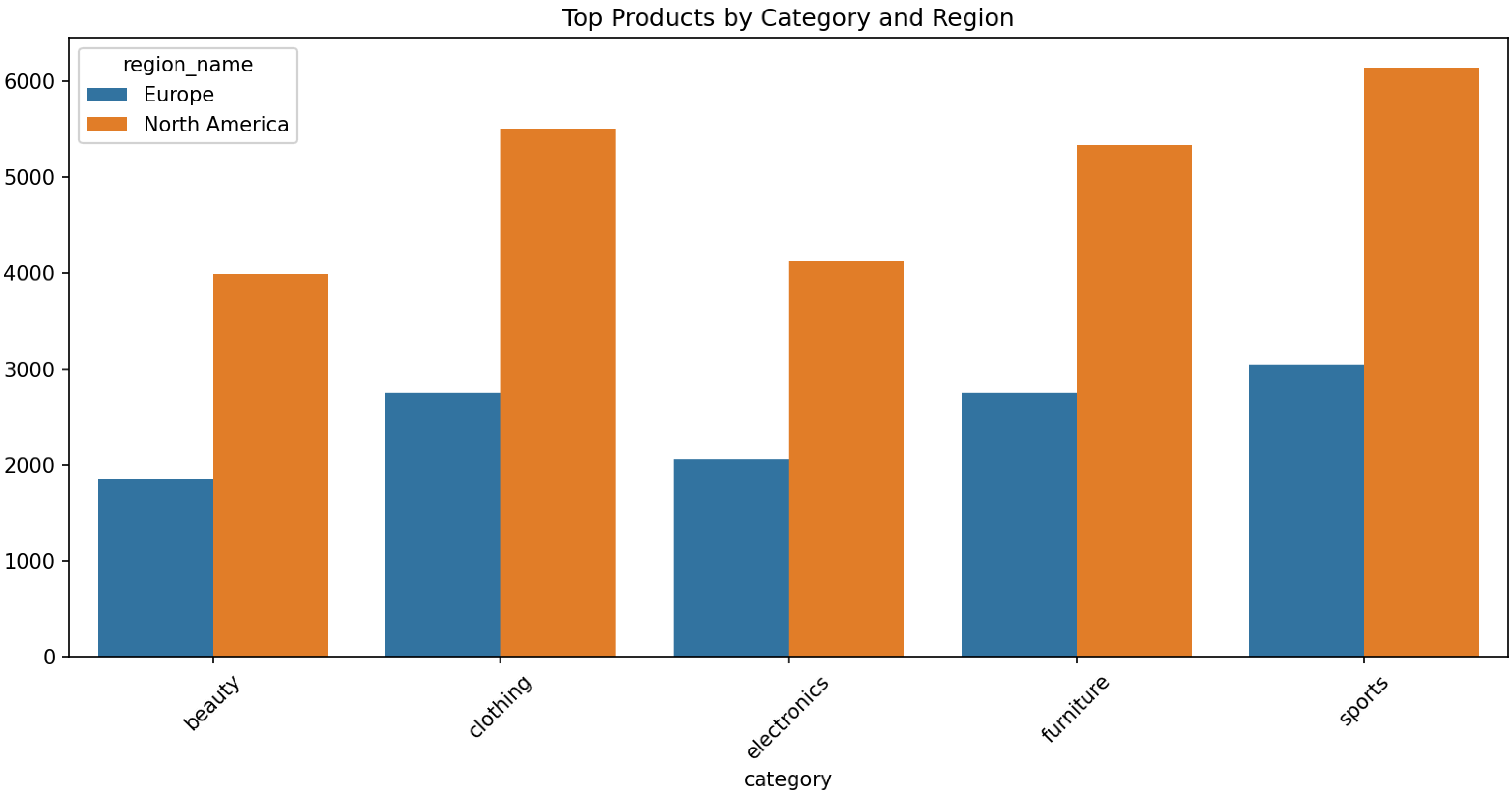
DIM_REGION

STRING	region	PK
STRING	country	
STRING	continent	
STRING	source_system	

Analysis

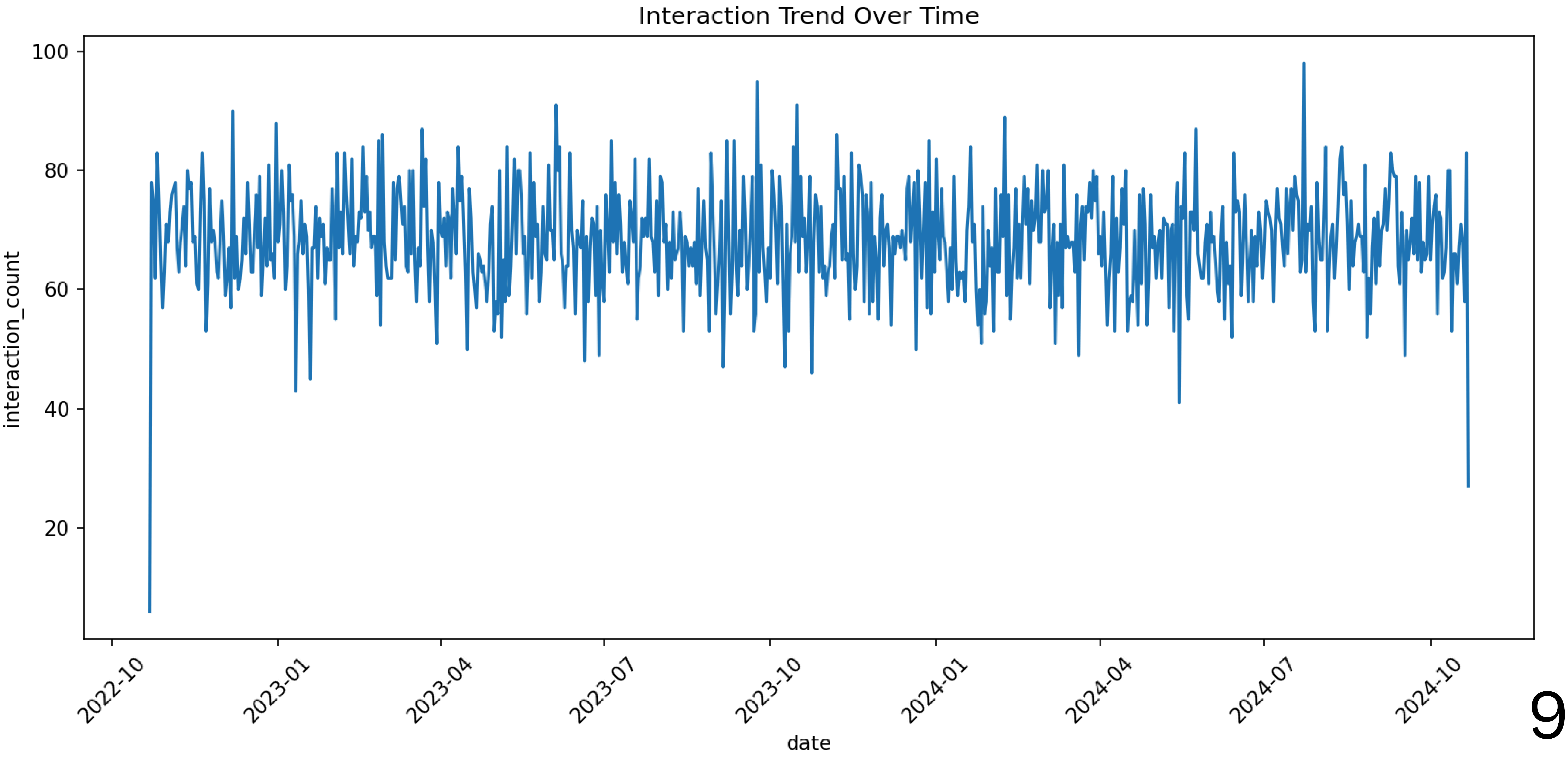
- Customer Activity by Region and Loyalty Tier
- Top Products by Category and Region
- Interaction Trend Over Time
- Historical Purchase Trends
- Customer Age and Interaction Analysis

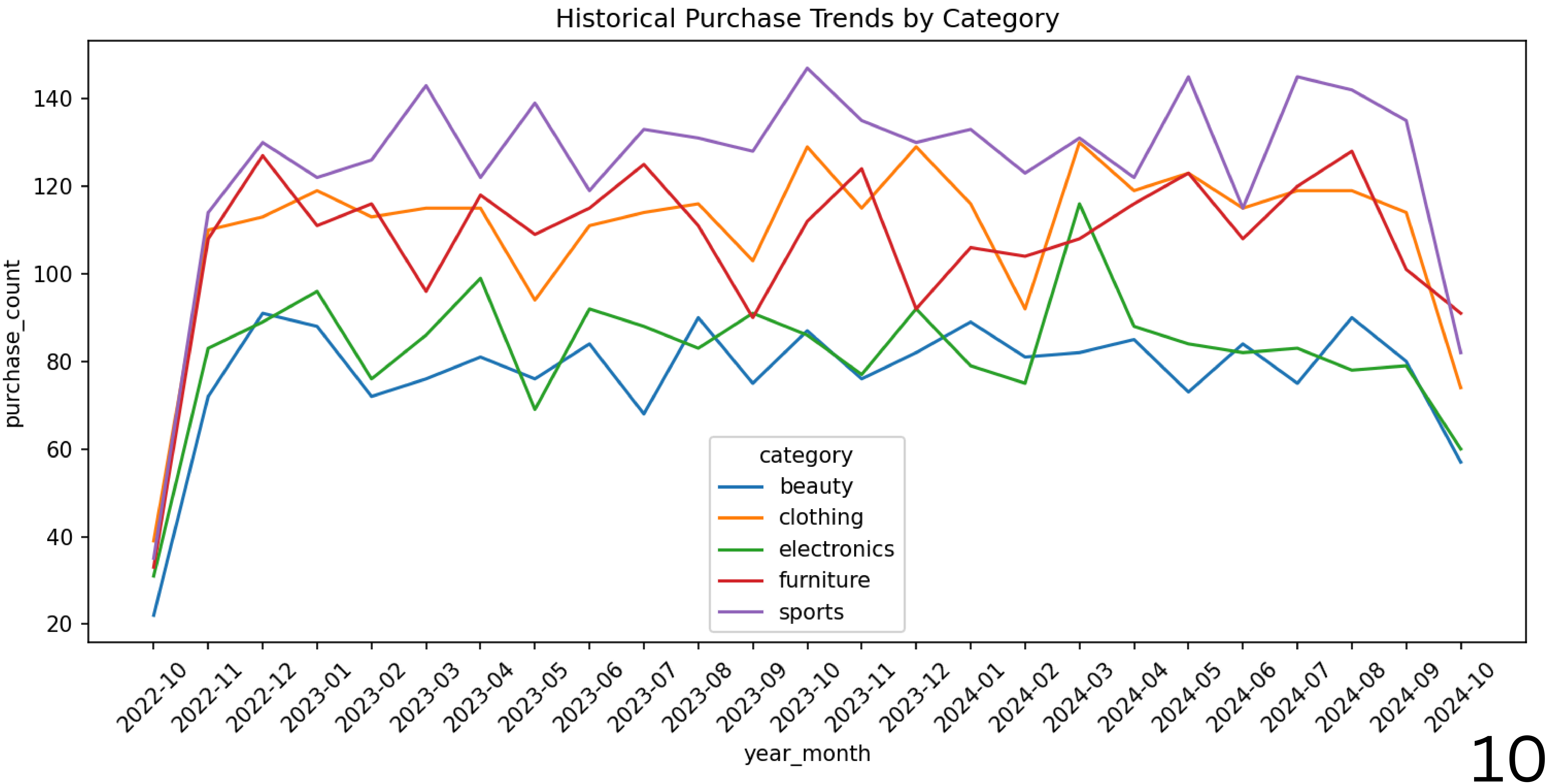

Customer Activity by Region and Loyalty Tier
loyalty_tier
NonPrime
Prime
interaction_count
0
2000
4000
6000
8000
10000
12000
Europe
North America
region_name

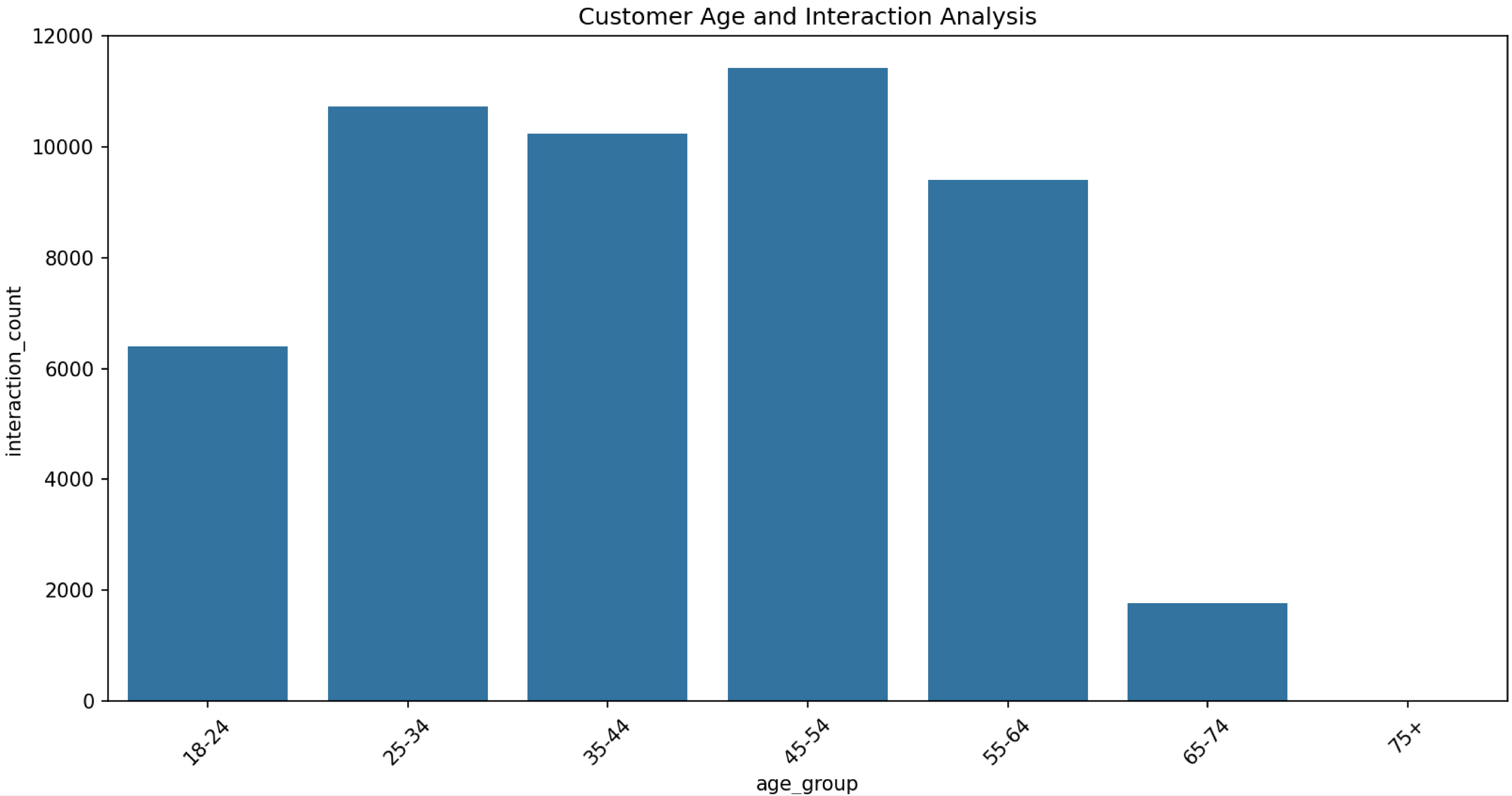


Top Products by Category and Region
region_name
Europe
North America
0
1000
2000
3000
4000
5000
6000
beauty
clothing
electronics
furniture
sports
category

Interaction Trend Over Time (2023-01-01 to 2024-10-15)




Historical Purchase Trends by Category
purchase_count
20
40
60
80
100
120
140
year_month
2022-10
2022-11
2022-12
2023-01
2023-02
2023-03
2023-04
2023-05
2023-06
2023-07
2023-08
2023-09
2023-10
2023-11
2023-12
2024-01
2024-02
2024-03
2024-04
2024-05
2024-06
2024-07
2024-08
2024-09
2024-10
category
beauty
clothing
electronics
furniture
sports


Customer Age and Interaction Analysis
12000
10000
8000
6000
4000
2000
0
interaction_count
18-24
25-34
35-44
45-54
55-64
65-74
75+
age_group

Challenges:

- Performing Incremental load and full load using CSV file source - Solution: Use Postgres in Data Source Layer
- Encounter issues during the loading clean data to my dimension and fact table
- Permission relation issue in file path (Hdfs file path)
- Transformation (Time stamp format mis-match can give invalid result in extraction)

Future Extention

- Visualization - through power BI or tableau
- Kafka - (Real-time capturing processing user
  Interaction log)
- Automate the data pipeline using Airflow transformation between layers
- Docker
- Airflow
- HBase (NoSQL) - Storing JSON based curated data

QA

Thank you