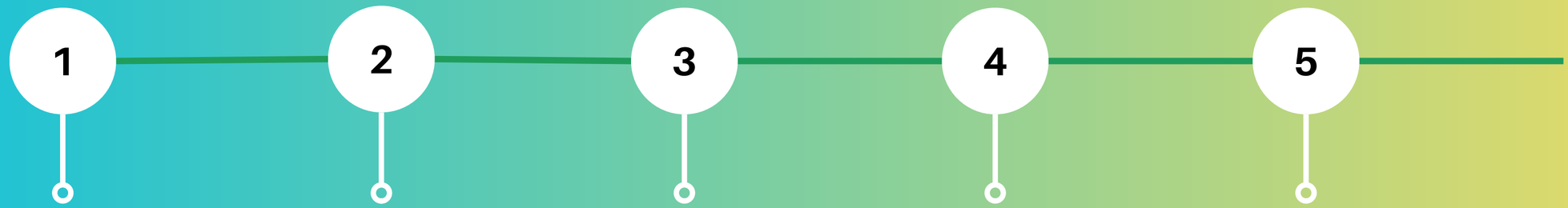
# A UNIFIED DATA PLATFORM FOR Q COMPANY'S RETAIL OPERATIONS

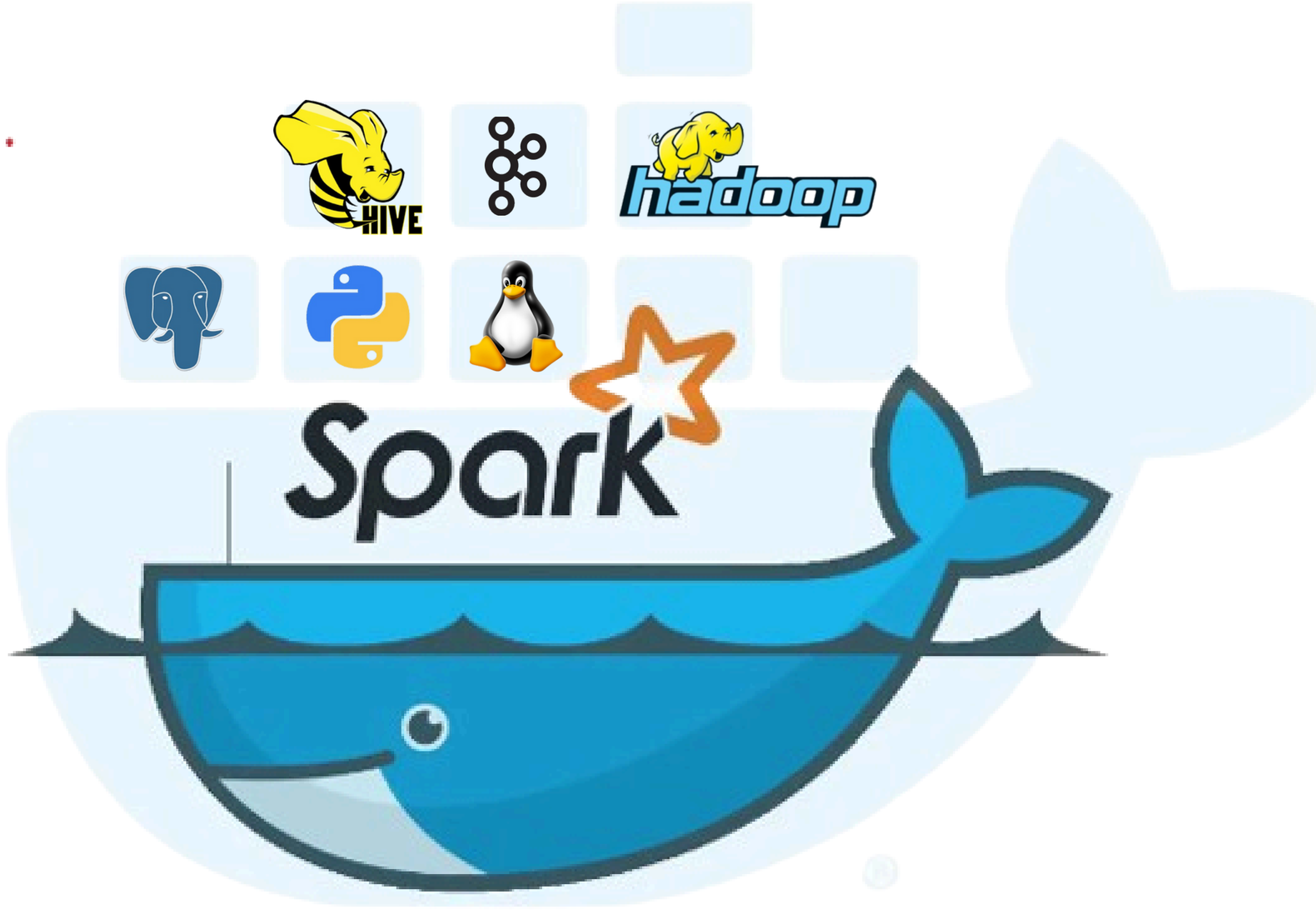
## HARNESSING THE POWER OF APACHE SPARK

# AGENDA

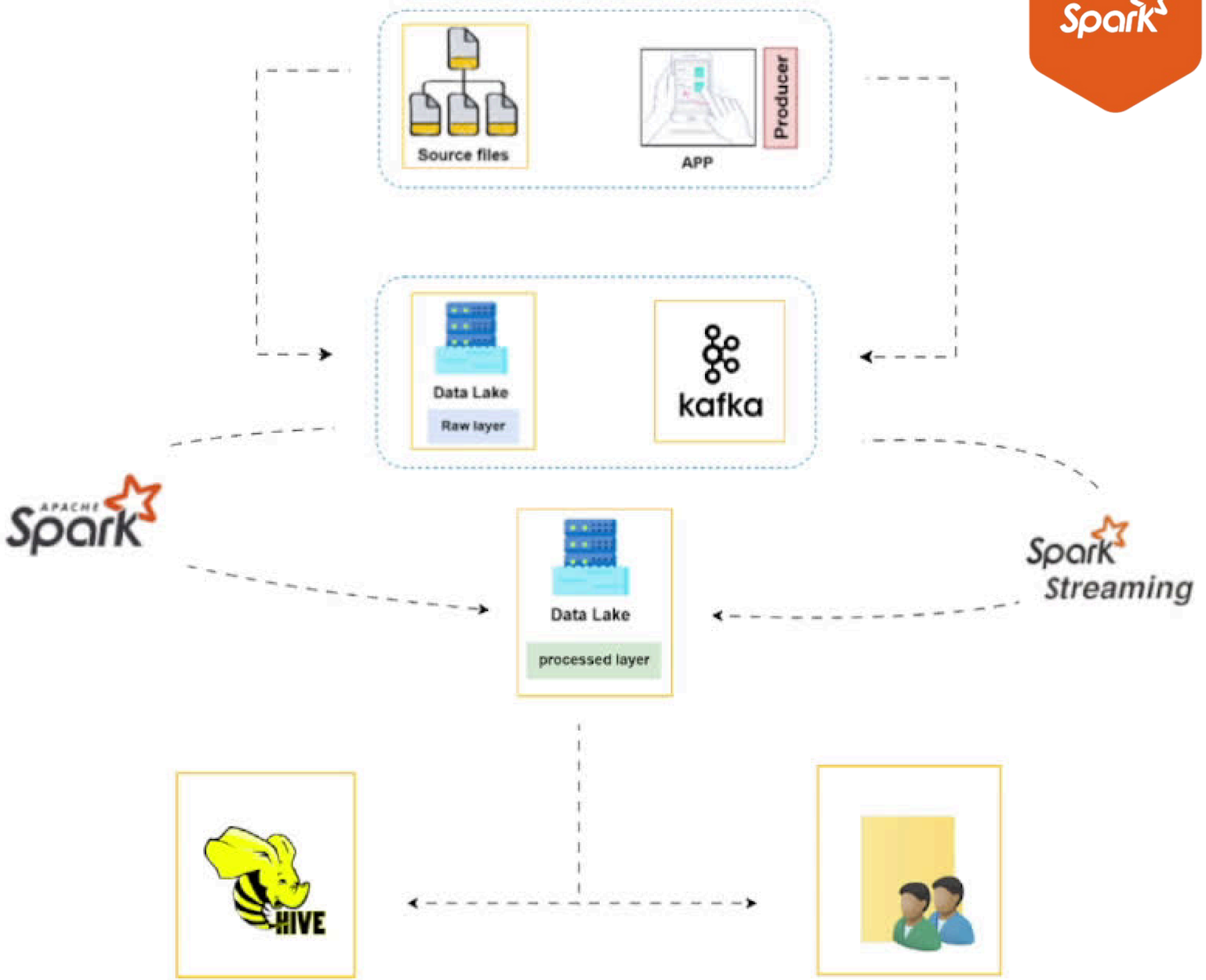



1. PROJECT OVERVIEW
2. INTRODUCTION
3. BATCH PROCESSING
4. DWH MODEL
5. STREAMING PROCESSING

# Project Overview



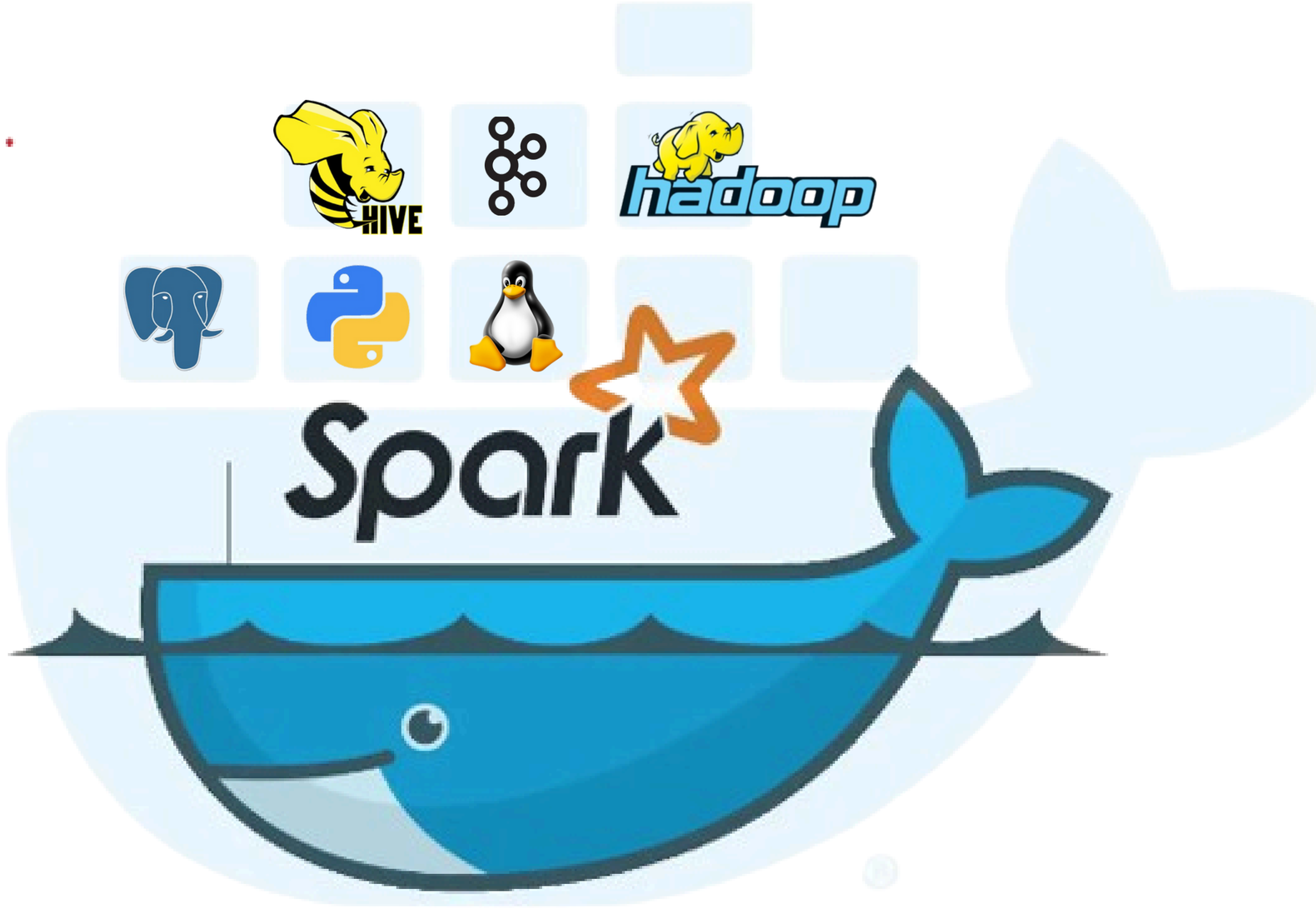
# DATA PIPLINE



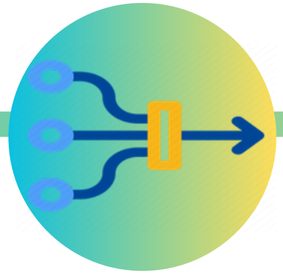
# Introduction

Q Company's innovative Retail Data Platform revolutionizes retail data management by utilizing Apache Spark and Kafka. The platform combines robust batch and streaming data processing to handle diverse data sources, providing real-time insights and enhancing decision-making agility. This initiative optimizes business processes, drives strategic growth, and delivers exceptional customer experiences in the dynamic retail landscape.

# Batch Processing



# Data Ingestion



Regular files containing information about branches, sales agents, and transactions are ingested into the local file system and then transferred to a data lake as raw files using Python scripts.

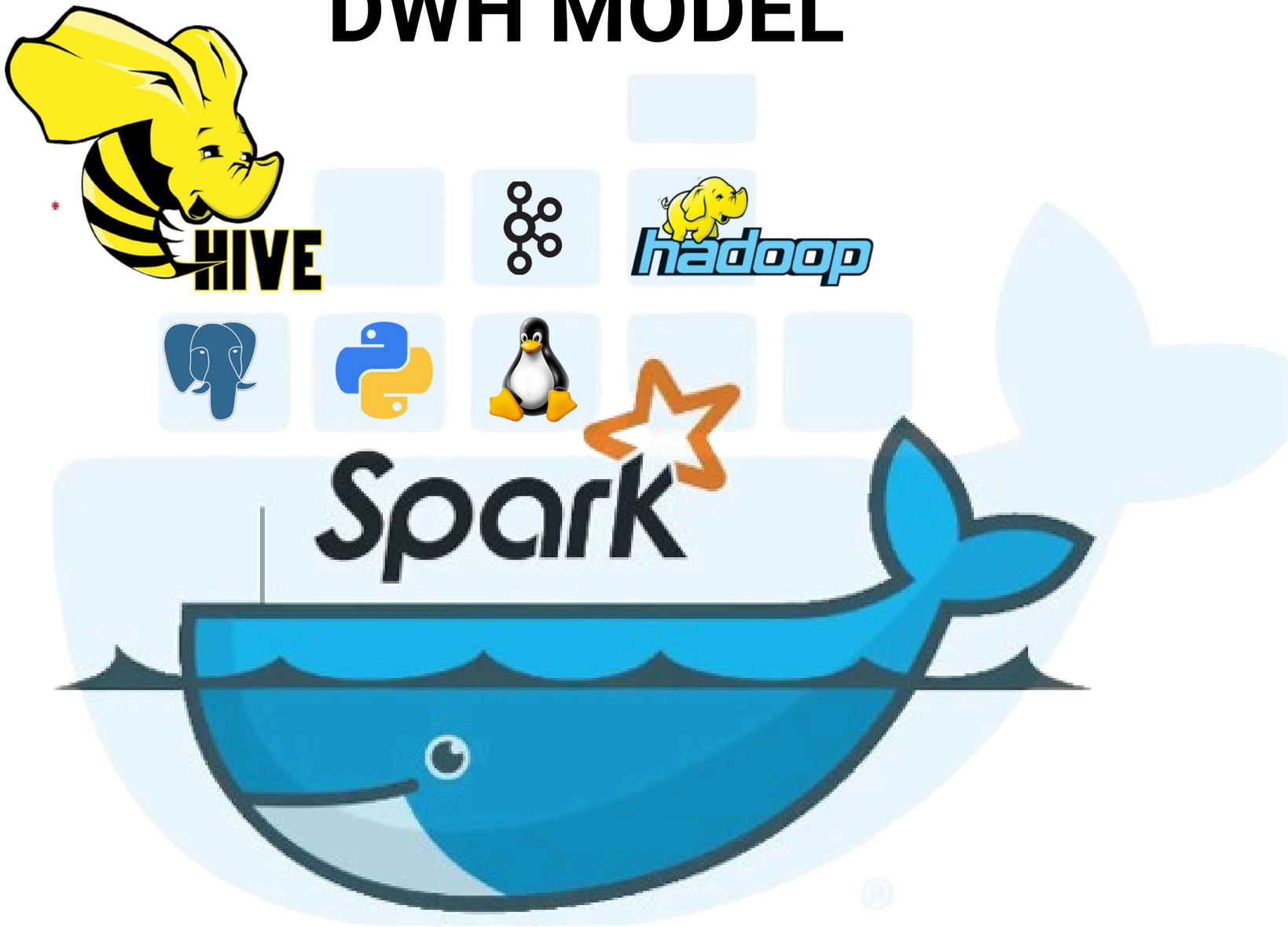
# Batch Processing

Apache Spark processes these raw files to clean and transform the data, ensuring accuracy and consistency. This processed data is then stored in Hive tables within a Data Warehouse (DWH) to facilitate efficient querying and reporting.

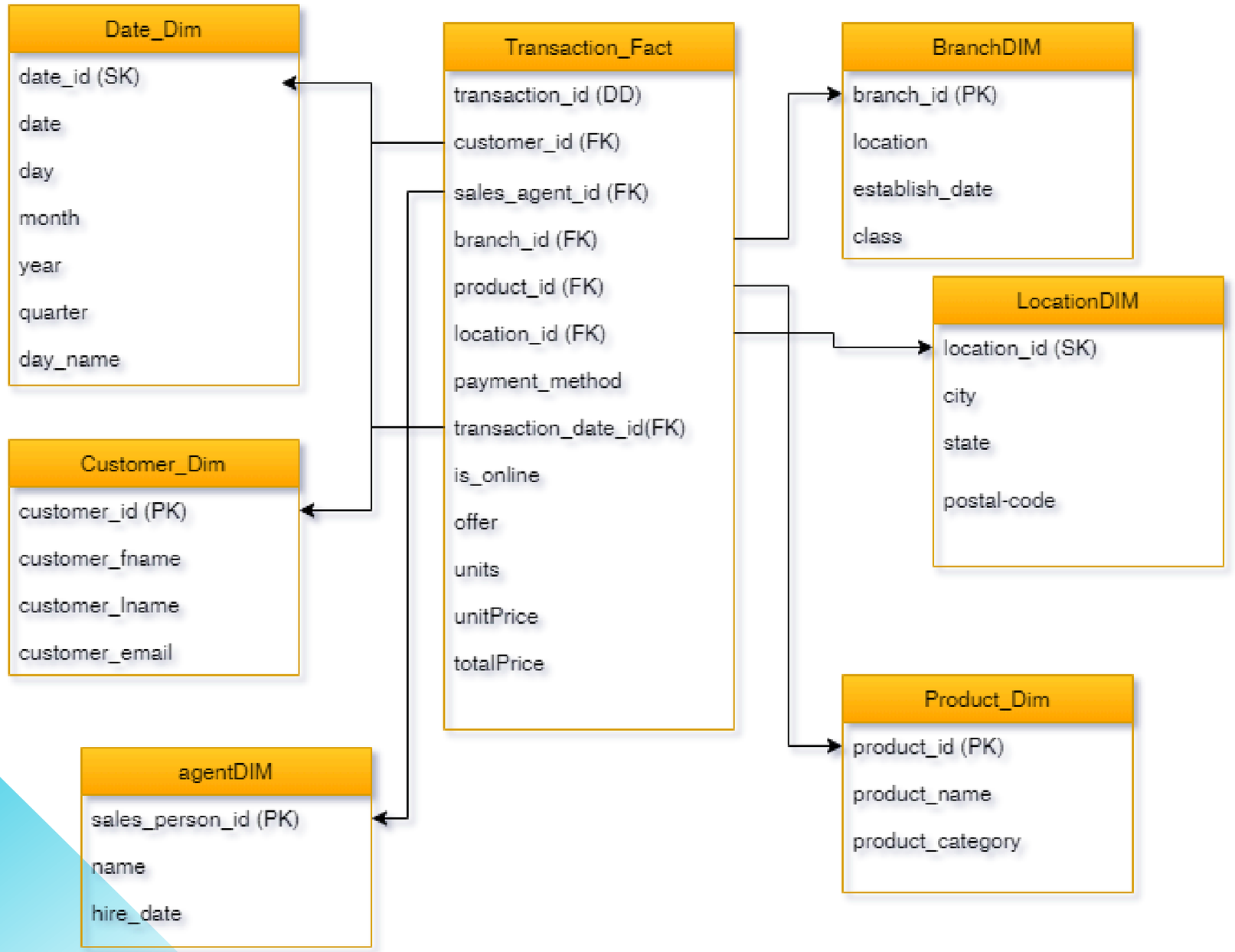
# Business Insights

The platform generates insights such as most selling products, customer offer redemptions, and performance metrics for marketing campaigns.

# DWH MODEL



# DWH MODEL


Date_Dim
date_id (SK)
date
day
month
year
quarter
day_name
Transaction_Fact
transaction_id (DD)
customer_id (FK)
sales_agent_id (FK)
branch_id (FK)
product_id (FK)
location_id (FK)
payment_method
transaction_date_id(FK)
is_online
offer
units
unitPrice
totalPrice
BranchDIM
branch_id (PK)
location
establish_date
class
LocationDIM
location_id (SK)
city
state
postal-code
Customer_Dim
customer_id (PK)
customer_fname
customer_lname
customer_email
agentDIM
sales_person_id (PK)
name
hire_date
Product_Dim
product_id (PK)
product_name
product_category

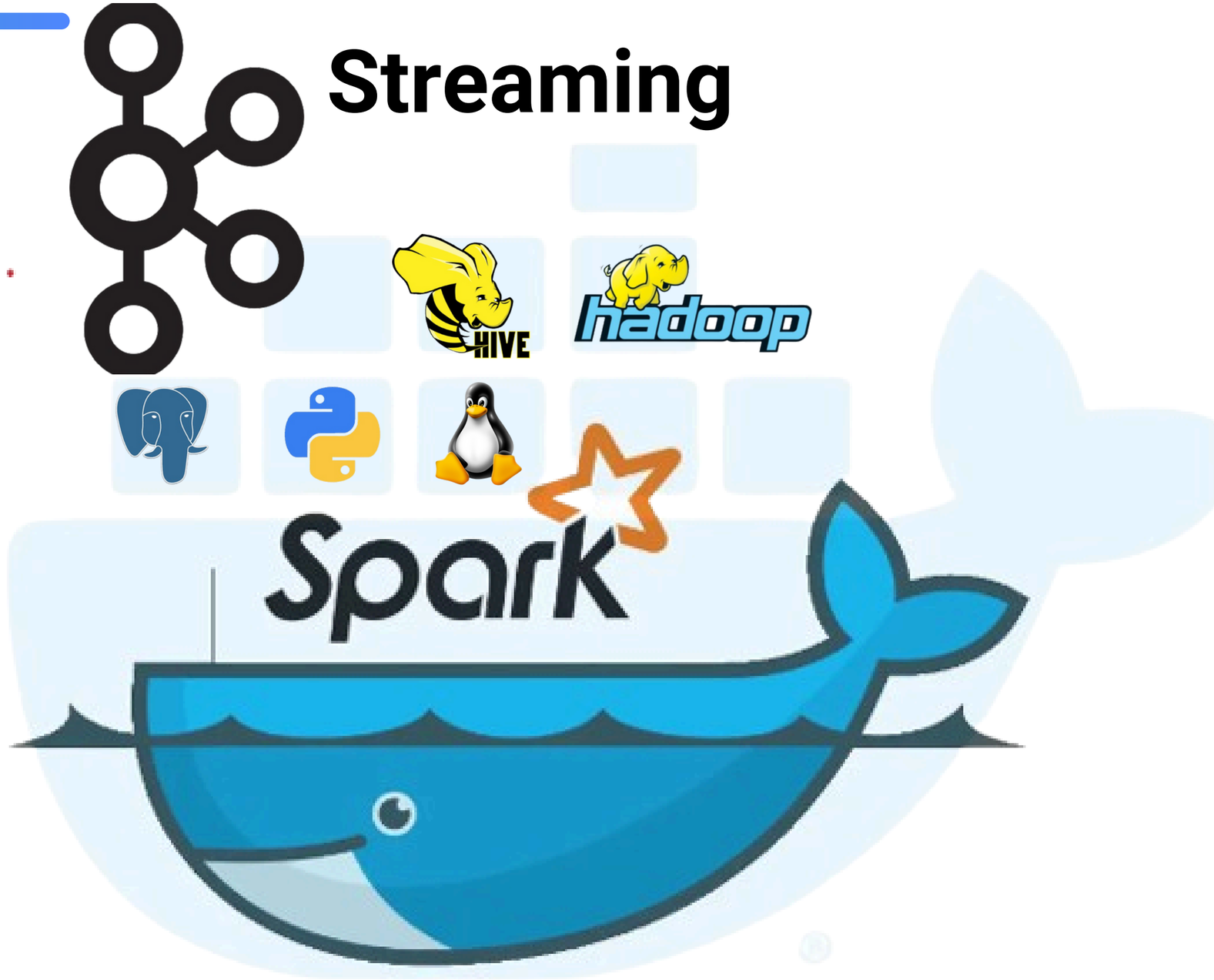
# DWH MODEL

/user/hive/warehouse/transactionsdb.db Go!

Show 25 entries Search:

| Permission | Owner | Group | Size | Last Modified | Replication | Block Size | Name |
|---|---|---|---|---|---|---|---|
| drwxr-xr-x | itversity | supergroup | 0 B | Jul 06 01:44 | 0 | 0 B | agentdim |
| drwxr-xr-x | itversity | supergroup | 0 B | Jul 06 01:47 | 0 | 0 B | branchsdim |
| drwxr-xr-x | itversity | supergroup | 0 B | Jul 06 11:12 | 0 | 0 B | customer1_dim |
| drwxr-xr-x | itversity | supergroup | 0 B | Jul 06 11:17 | 0 | 0 B | date1_dim |
| drwxr-xr-x | itversity | supergroup | 0 B | Jul 06 11:16 | 0 | 0 B | location1_dim |
| drwxr-xr-x | itversity | supergroup | 0 B | Jul 06 11:14 | 0 | 0 B | product1_dim |
| drwxr-xr-x | itversity | supergroup | 0 B | Jul 06 11:19 | 0 | 0 B | transaction_facttable |

# Streaming

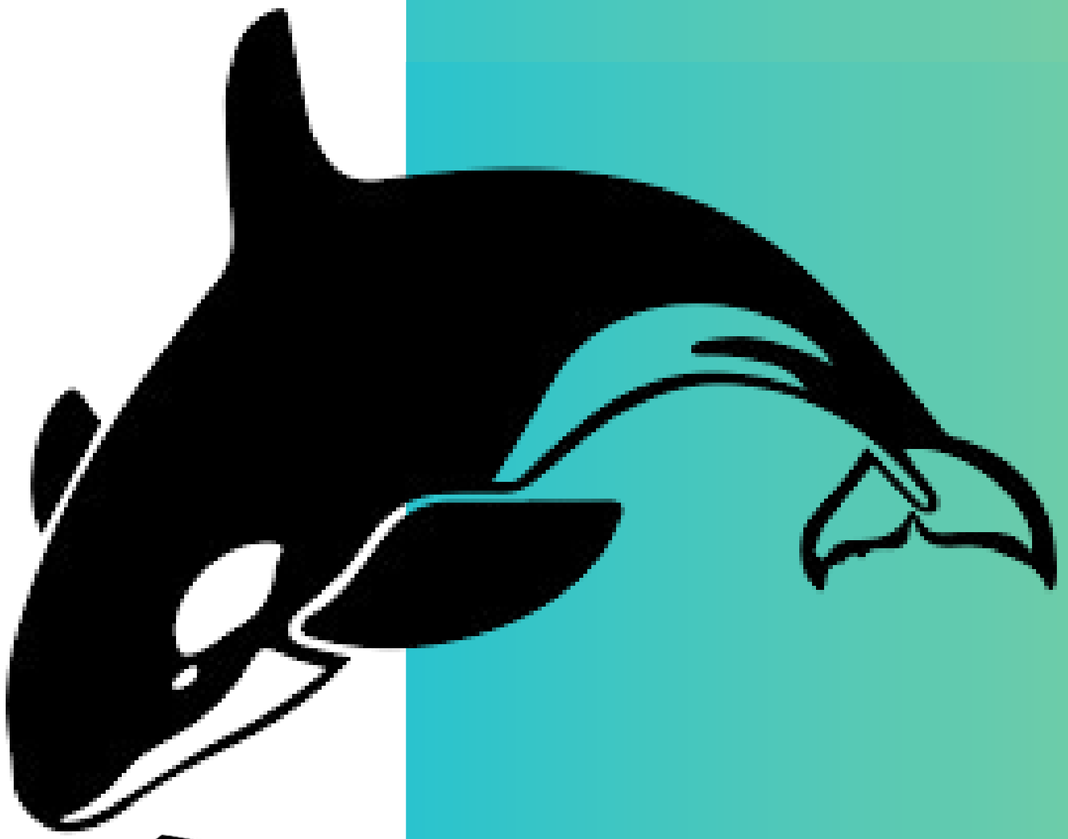


Streaming Processing

01 Kafka Integration

02 Spark Streaming

03 Business Reports



# Areas of Development

## writing the data to Hbase

writing the data to hbase instead of hive or any other file formant will help us to avoid manual compaction

# Our Team

Marina Fawsy

Mariam Maged

Malak Sherif

Ahmed Farid