DATA PLATFORM

PROBLEM STATEMENT

upGrad ECommerce Pvt. Limited is a popular e-commerce website. They serve millions of active customers and have an order volume of hundreds of thousands. With such a large scale venture it is becoming increasingly difficult for them to manage their data needs.

- Various teams from analytics, marketing, call centre to business have varied data needs.
- Till now the data team used to serve these departments independently creating silos of data that are difficult to view together.

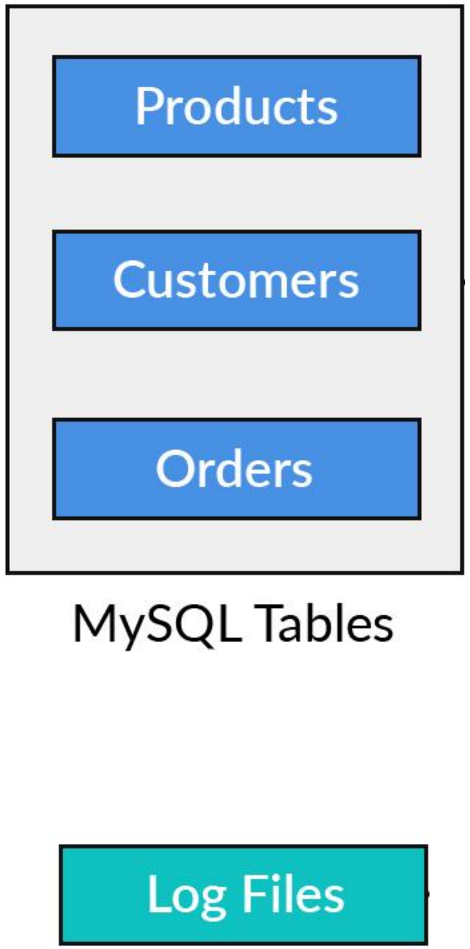
REQUIREMENTS

The data team after various considerations has shortlisted two important use cases-

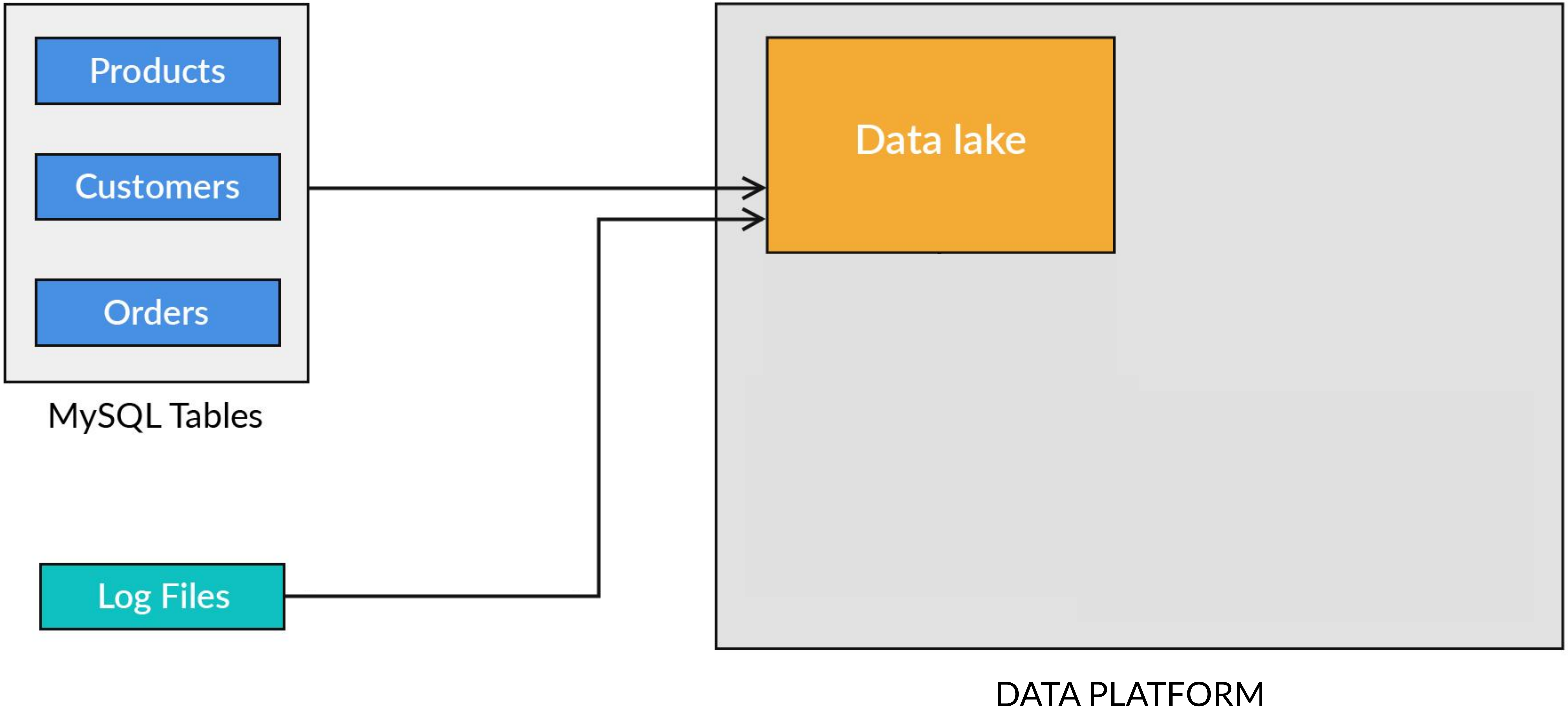
- Analytics of all the structured and unstructured data from various silos.
- A Customer 360° view to drill down on customer activity.

To solve these use cases the team has decided to build out a data-platform.

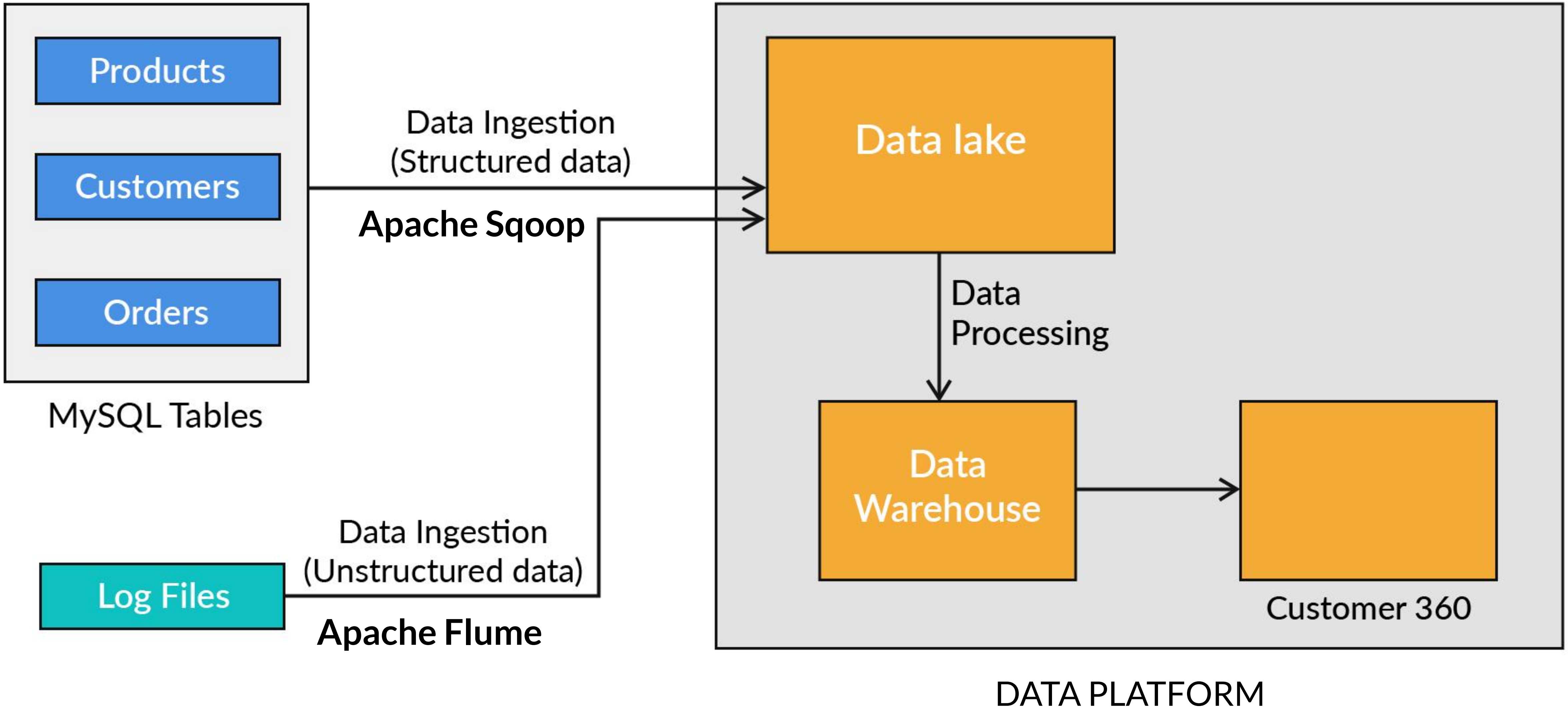
DATA PLATFORM



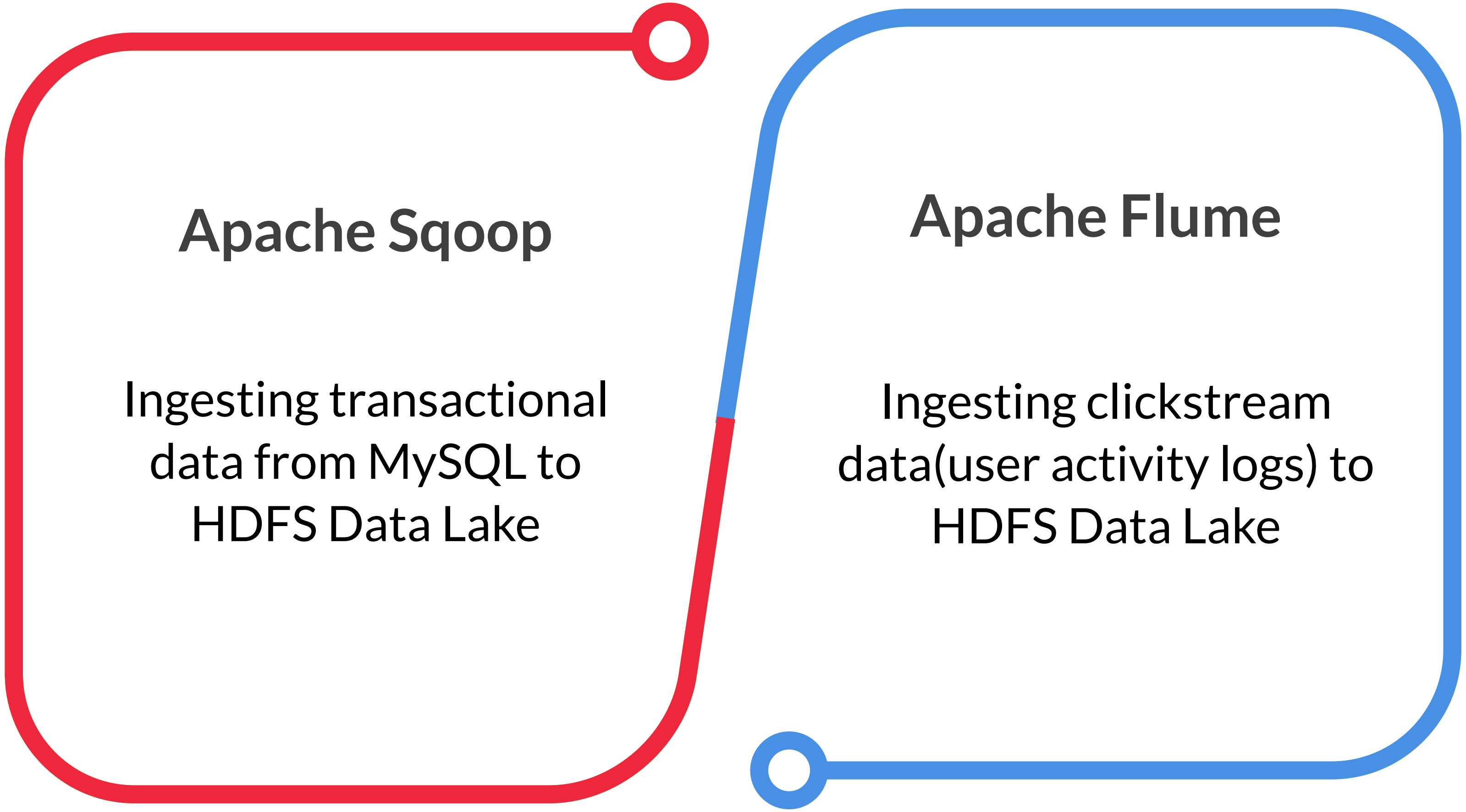
DATA PLATFORM



DATA INGESTION



DATA INGESTION



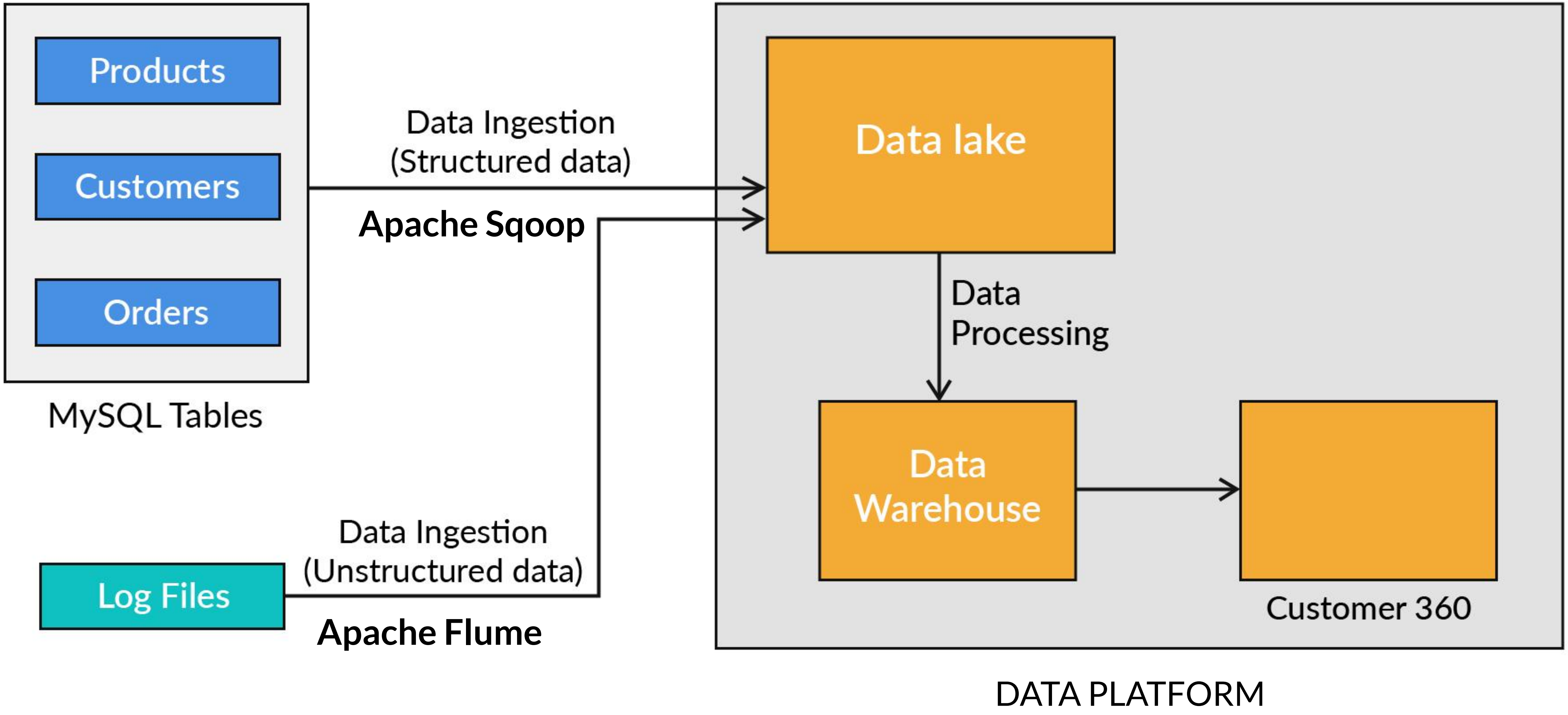
Apache Sqoop

Ingesting transactional data from MySQL to HDFS Data Lake

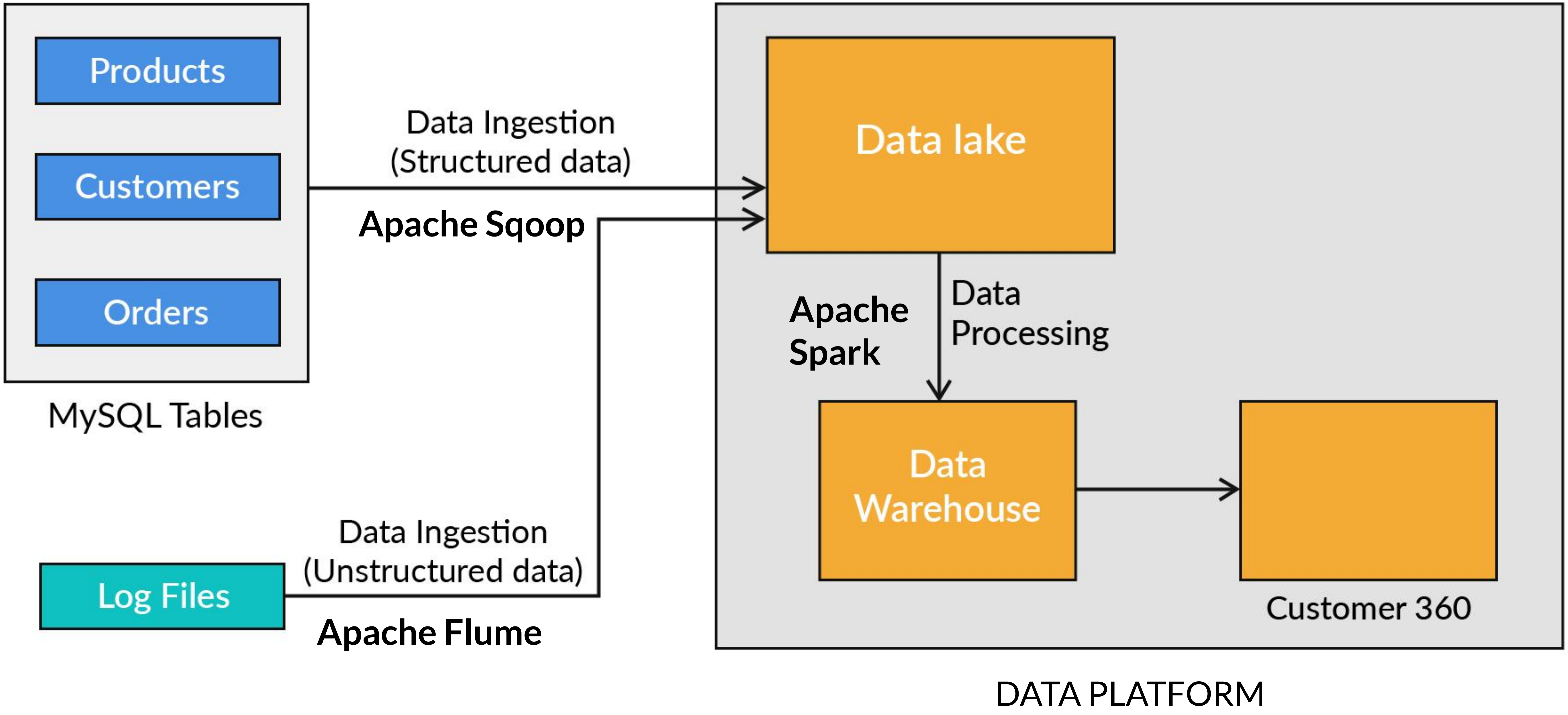
Apache Flume

Ingesting clickstream data(user activity logs) to HDFS Data Lake

DATA INGESTION



DATA PROCESSING

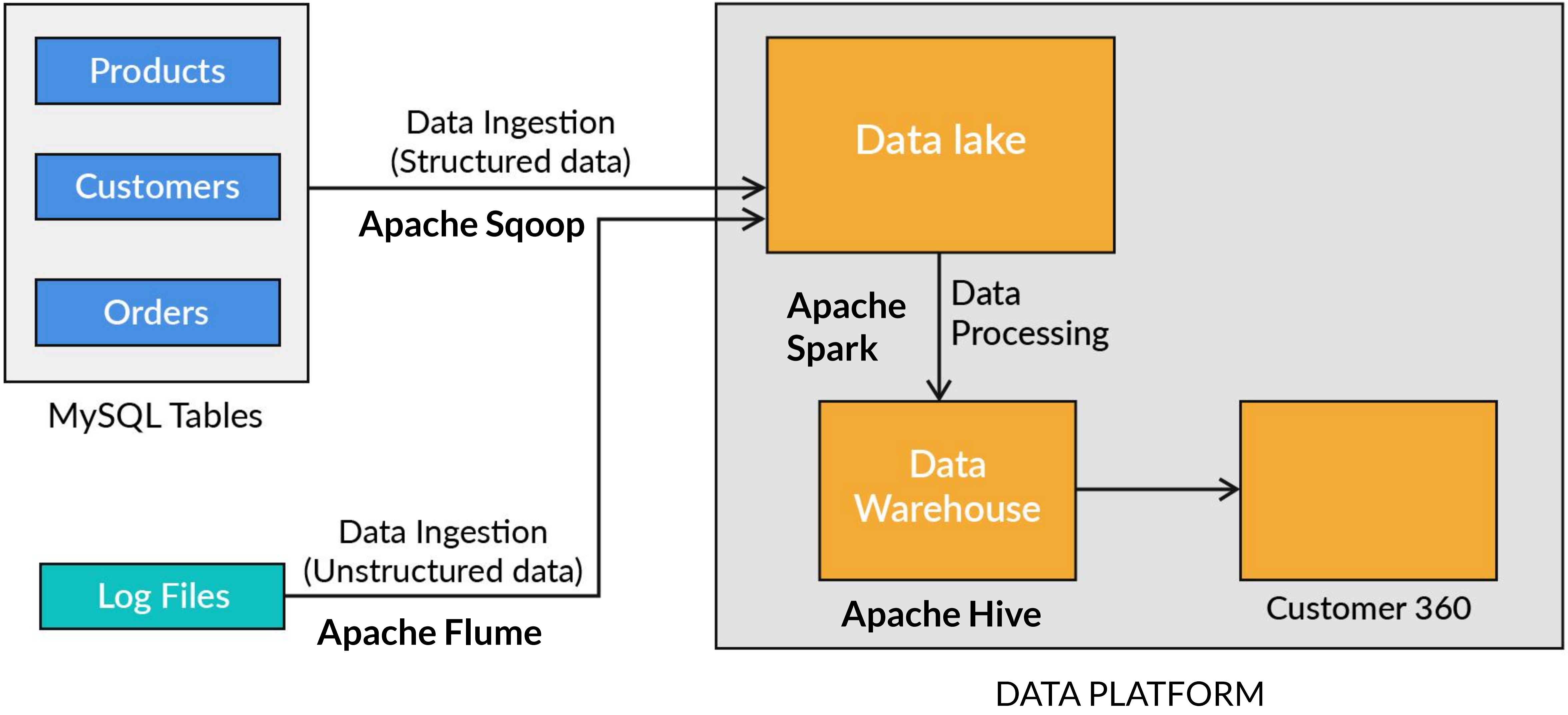


DATA PROCESSING

Apache Spark

- Processing Log File
- Loading the data into HDFS
- This processed data will be used to create Hive tables

DATA WAREHOUSING

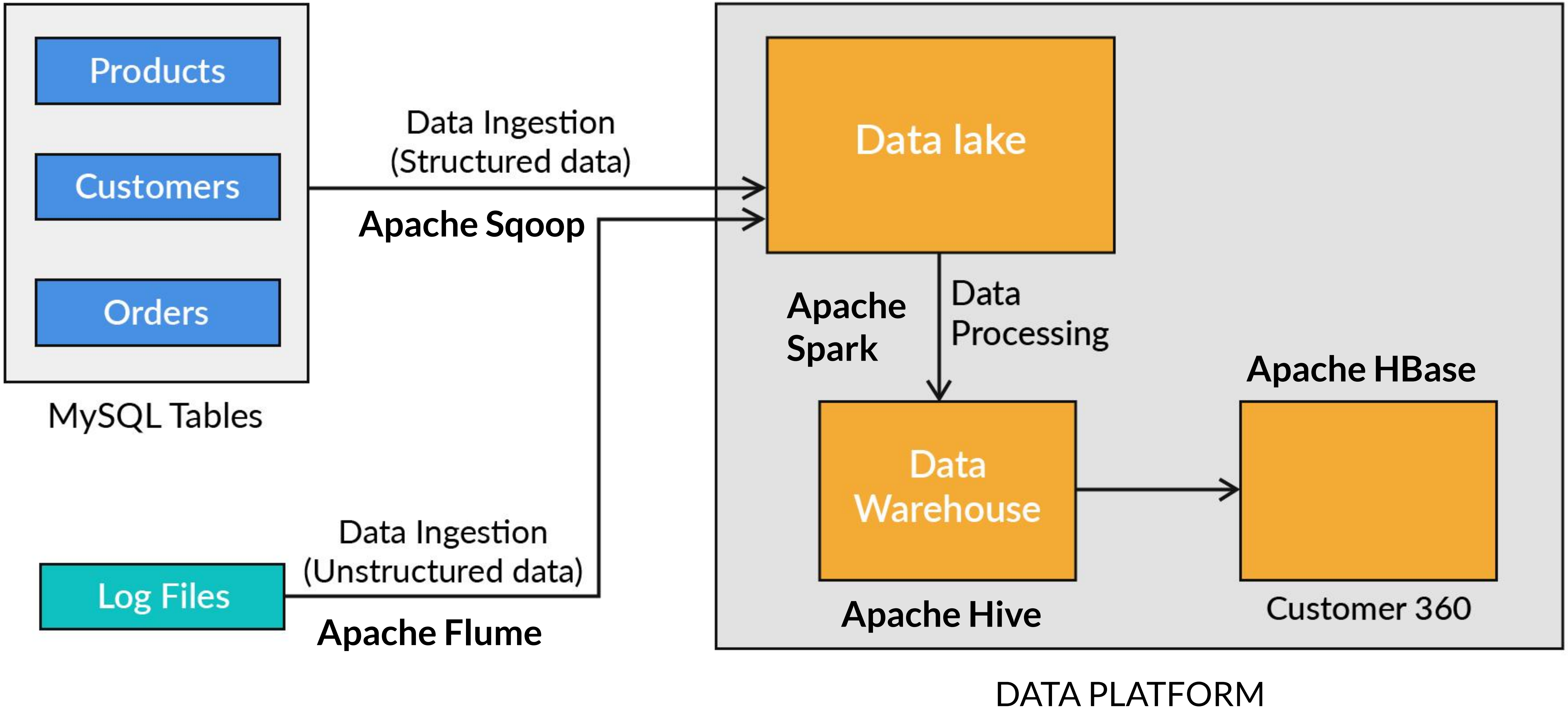


DATA WAREHOUSING

Apache Hive

- Creating table for data - Transactional + Structured Logs
- Querying the data
- SCD-2 implementation

HIVE-HBASE LOOKUP TABLE



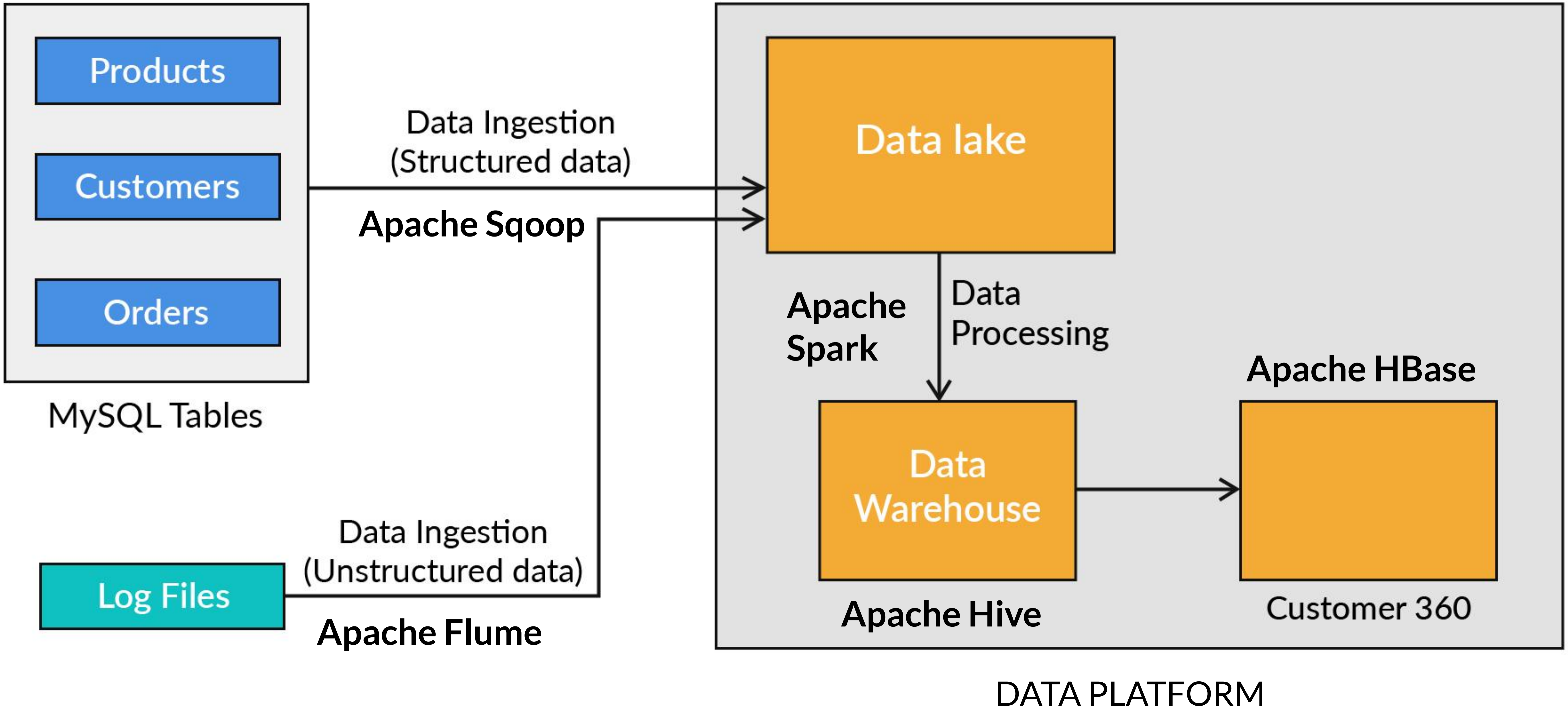
Products
Customers
Orders
MySQL Tables
Data Ingestion
(Structured data)
Apache Sqoop
Log Files
Data Ingestion
(Unstructured data)
Apache Flume
Data lake
Apache Spark
Data Processing
Data Warehouse
Apache Hive
Apache HBase
Customer 360
DATA PLATFORM

HIVE-HBASE LOOKUP TABLE

Apache HBase

- 360° view on User Activity

DATA PLATFORM



Thank You