**Flipkart Data Platform Overview**

Flipkart processes **terabytes of data daily** from a **vast network of sources**, including **customer interactions**, **product data**, and **internal systems**. Its **Data Platform (FDP)** handles data from **thousands of sources**, notably over **52,000 serviceability data sources**, and supports **millions of keys** tied to **all India PIN codes**.

* **Diverse Data Sources**: Includes customer feedback, search queries, browsing and purchase history.
* **Microservices Architecture**: Thousands of applications and services generate and consume data.
* **MySQL Clusters**: Used for critical operations like order management and logistics.
* **Streaming (fStream)**: Ingests data in real-time, integrated with Kafka.

**Storage & Processing**: Petabyte-scale infrastructure using Hadoop and other big data tools.

**How Flipkart stores and process massive amount of data**

Flipkart stores and processes its massive volumes of data using a robust, scalable, and distributed big data architecture. Here's how it works:

### 🔹 ****1. Data Ingestion****

Flipkart collects **terabytes of data daily** from:

**User interactions** (search, click, purchase history)

**Backend systems** (order management, logistics, inventory)

**Microservices** (thousands of services producing data)

**Streaming platforms** (real-time events via Kafka/fStream)

### 🔹 ****2. Data Storage****

Flipkart uses a **hybrid data storage architecture**:

**Hadoop Distributed File System (HDFS)**: Stores **petabytes** of structured and unstructured data.

**MySQL Clusters**: Used for high-speed transactional systems (e.g., orders, inventory, supply chain).

**Data Lakes**: Centralized storage to hold raw data for various processing needs.

### 🔹 ****3. Data Processing****

#### Batch Processing:

Uses **Apache Hadoop**, **Hive**, and **Spark** for large-scale batch jobs like analytics, reporting, and model training.

#### Real-time Processing:

**fStream** (Flipkart’s streaming platform) + **Apache Kafka**: Processes real-time data like order updates, user behavior, etc.

### 🔹 ****4. Data Pipeline & Orchestration****

**Flipkart Data Platform (FDP)** manages ingestion, processing, and orchestration of data pipelines.

Handles data from **over 52,000 serviceability source hashes** and **millions of service keys** (like PIN codes).

Uses internal tools and scheduling systems to run ETL jobs reliably.

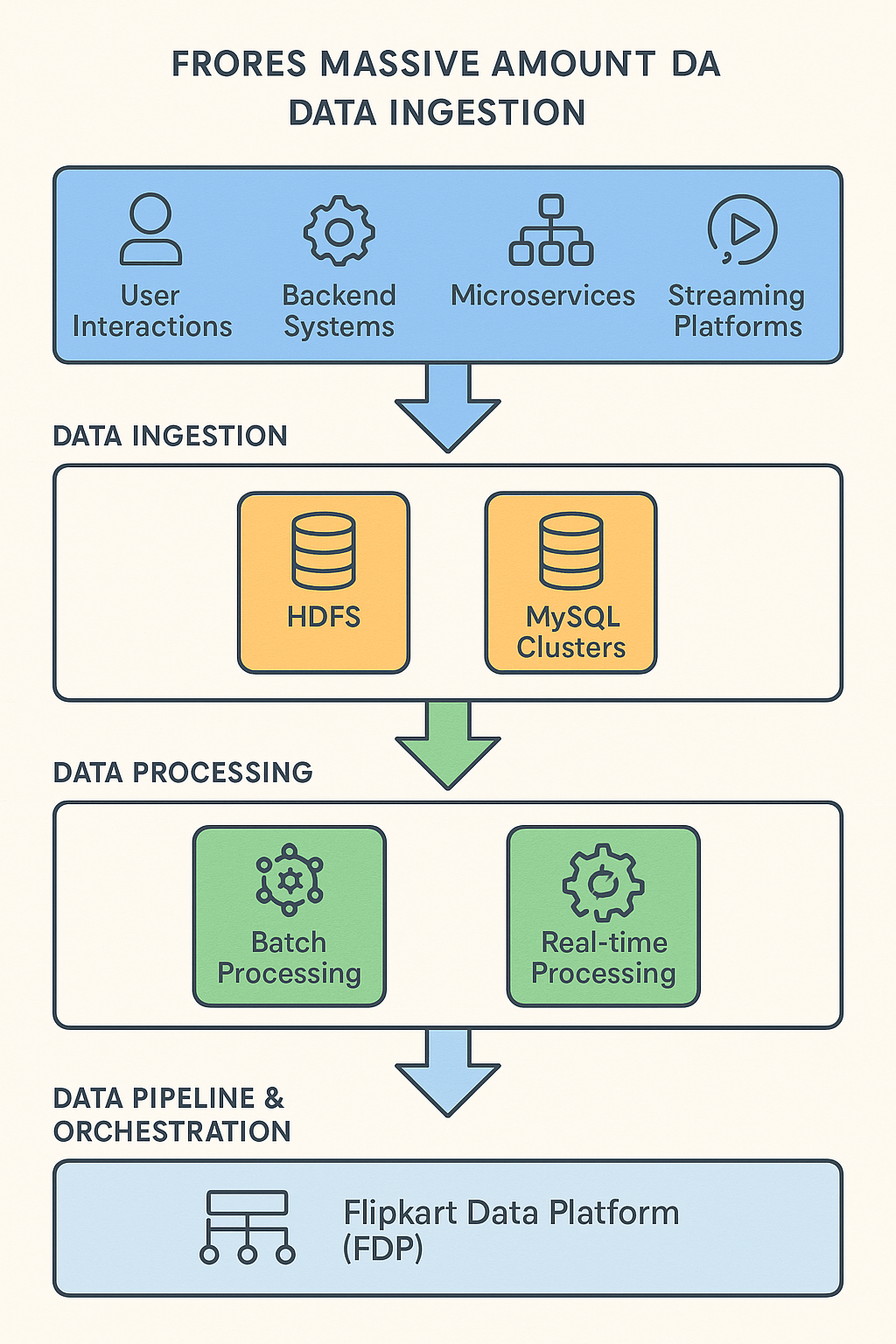
### 🔹 ****5. Scalability & Reliability****

The system is **fault-tolerant**, **scalable**, and **highly distributed**.

Uses **replication**, **partitioning**, and **caching** for speed and reliability.

### Summary Table:

| **Component** | **Technology Used** | **Purpose** |
| --- | --- | --- |
| Data Ingestion | Kafka, fStream | Collect data from services in real-time |
| Storage | HDFS, MySQL, Data Lakes | Store structured/unstructured data |
| Batch Processing | Hadoop, Hive, Spark | Large-scale analytics |
| Real-time Processing | Kafka + fStream | Real-time data handling |
| Pipeline Management | Flipkart Data Platform (FDP) | Manage, orchestrate, monitor pipelines |

  
  
  
Flipkart uses various **data formats** to handle the massive volumes of data flowing through its platform. The formats depend on the **data type**, **use case**, and **stage in the data pipeline** (e.g., ingestion, processing, or storage). Here's a breakdown:

### 📊 ****Flipkart Data Types and Size Overview****

| **Data Type** | **Formats Used** | **Where It's Used** | **Estimated Size** |
| --- | --- | --- | --- |
| **Structured Data** | CSV, SQL Tables, TSV | MySQL (orders, payments, inventory, logistics) | **Several TB per cluster** |
| **Semi-Structured** | JSON, XML, Avro | APIs, Kafka streams, microservices communication | **10–20+ TB/day** |
| **Unstructured** | Text, Images, Videos, Binary | Reviews, user uploads, product images | **Petabytes (long-term storage)** |
| **Columnar** | Parquet, ORC | Big data analytics (Hive, Spark, Presto) | **Petabytes in data lake** |
| **Streaming Data** | Kafka + Protobuf, JSON | Real-time clickstreams, order events (fStream) | **Millions of events/sec** |
| **Logs/Events** | JSON, Avro, Thrift, Text | Service logs, app telemetry | **10–100+ TB/day** |
| **Serviceability** | Key-value, Hash Maps (compressed) | PIN code reachability, delivery availability | **Millions of keys**, **TB scale** |