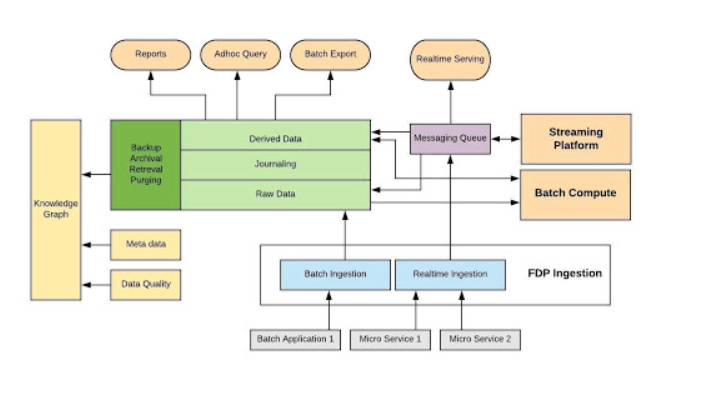
**BIG DATA CASE STUDY ON FLIPKART**

Flipkart Data Platform has an architecture that is service oriented and facilitates manipulation of batch data as well as streaming data. This platform contains various micro-services that facilitates user experience through efficient product listings, optimization of prices, maintaining various types of data domains – Redis, HBase etc. This FDP is capable of storing 35 Peta Bytes of data and is capable of managing 800+ Hadoop nodes on the server.

**The Architecture of Flipkart Data Platform:**



**The process of how Flipkart works on Big Data:**

Flipkart Data Platform is a service-oriented architecture that facilitates manipulation of batch data as well as streaming data. The platform uses microservices for better user experience with product listings, price optimization, and diverse data management (Redis, HBase, SQL). It has the storage capacity of 35 PB and can manages over 800 Hadoop nodes on its server.

Let's delve into the comprehensive procedure of how Flipkart leverages Big Data.

**1. FPD Ingestion System**

A Big Data Ingestion System serves as the initial gateway for variables to enter the information system, involving the importing and storing of data into a database. This data can be received in real-time streams. There are 3 ways during which ingestion are often done –

Specter – this is often a Java library that’s used for sending the draft to Kafka.

• Dart Service – a REST service that permits the payload to be sent over HTTP.

• File Ingestor – we can use the CLI tool to dump data into the HDFS.

The user defines a schema, which creates a corresponding Kafka topic. Using Specter, data is ingested into the FDP, and the payload in the HDFS file is stored as HIVE tables.

### 2. Batch Compute

Batch Compute in the big data ecosystem processes data collected over time in batches, ideal for large-scale transaction processing. It's efficient for end-of-day processing of voluminous data.

### 3. Streaming Platform

The streaming platform process the data that is generated in sub-seconds. Apache Flink is a leading real-time streaming platform known used for the purpose.

### 4. Messaging Queue

### A Messaging Queue facilitates temporary storage for messages when the destination is unavailable. Messages can be plain text, byte arrays with headers, or commands. It consists of Producers, who generate and deliver messages, and Consumers, the endpoints where messages are processed.

### 5. Real-time Serving

### After retrieving messages from the Messaging Queue, the real-time serving system acts as a consumer. This platform enables users to gather real-time knowledge and access data through everchanging pipelines.

### 6. Data Lake

From the data lake, data is transferred to three main routes:

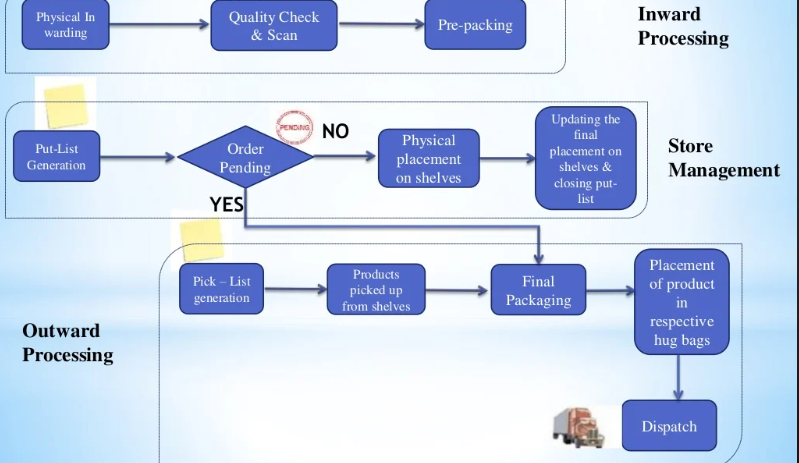
**Reports**: Generated mainly from batch data, these reports offer comprehensive insights into website logs, daily readings, and other metrics. They help companies like Flipkart quantify market needs effectively.

**Ad hoc Query**: Data analysts use business intelligence tools to generate specific queries from the data lake for targeted analysis.

**Batch Export**: This component exports large volumes of data from the data lake in various formats for further processing.

**Knowledge Graphs**: Utilizing metadata, knowledge graphs create interconnected networks of entities to derive insights and relationships efficiently. Tools like Apache Spark’s GraphX library are commonly used for graph construction.Top of Form

**Management System in Flipkart:**



**Frontend Technologies:**

Flipkart employs a variety of frontend technologies to enhance user experience:

1**. React-Native**: Used for developing mobile apps across Android, iOS, and mWeb platforms. It ensures performance efficiency, supports native components, and enables over-the-air updates.

2. **Proteus:** Developed internally for Android, it facilitates OTA layout updates and data bindings, accelerating experimentation and bug fixes in mobile app development.

3. **Viewport Meta:** Ensures Flipkart's website responsiveness across diverse screen types with various resolutions and sizes, optimizing user’s experience on smartphones, tablets, and desktops.

4. **Apple Mobile Web App Capable**: Enhances iOS user experience by allowing the Flipkart website to function like a native app from the home screen, improving accessibility and usability.

These technologies collectively contribute to a seamless and engaging shopping experience on Flipkart's platforms.

**Backend Technologies:**

Flipkart's backend infrastructure is fortified with several key technologies:

1. **Nginx:** A lightweight, open-source web server that optimizes performance under high traffic, serving as Flipkart's core web server to handle multiple user connections seamlessly.

2. **Kafka**: Apache Kafka is pivotal for Flipkart's messaging queue system, ensuring real-time data feeds and efficient communication across its tech stack, supporting reliable data processing.

3. **Dropwizard**: This Java-based framework simplifies development of RESTful web services, empowering Flipkart to build scalable backend services efficiently.

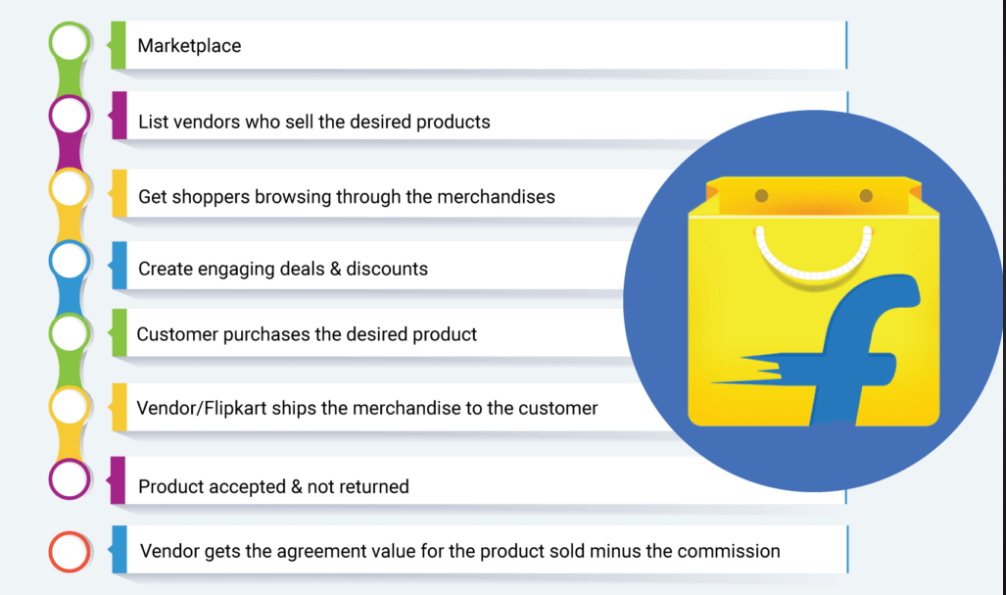
4. **Hadoop Distributed File System (HDFS):** HDFS manages Flipkart's extensive data, providing high throughput access and enabling robust data analysis for improving user experience.

5. **Quartz**: An open-source job scheduling library that aids Flipkart in managing and scheduling tasks across its platform, enhancing operational efficiency.

6. **Azkaban:** Flipkart utilizes Azkaban for orchestrating complex workflows, automating tasks to streamline operations and resource management effectively.

7. **Hive**: Apache Hive supports Flipkart's data warehousing needs by facilitating querying and management of large datasets stored in HDFS, empowering informed decision-making and optimizing platform performance.

**Working Process:**



**Brief Definitions**:

Batch Processing refers to handling large data sets within a scheduled period, like credit card transactions and bill generation. Stream Processing deals with continuous data streams as they occur.

**Hadoop Nodes**

Hadoop is a Java-based open-source framework that facilitates storing and processing of large datasets across distributed computing clusters, leveraging commodity hardware.

**Ingestion**

Apache Kafka is a distributed platform that facilitates event streaming for building high-performance data pipelines, streaming analytics, and integrating data in real-time.

**REST**

Representational State Transfer (REST) is an architectural style for constructing networked applications, facilitating flexible access to web services.

**CLI Tools**

Command Line Interface (CLI) tools execute operating system functions via text input, enhancing automation and control.

**HDFS (Hadoop Distributed File System)**

HDFS is a software system used for storage purpose in Hadoop clusters, designed for reliable data storage across commodity hardware, using large blocks rather than small ones.

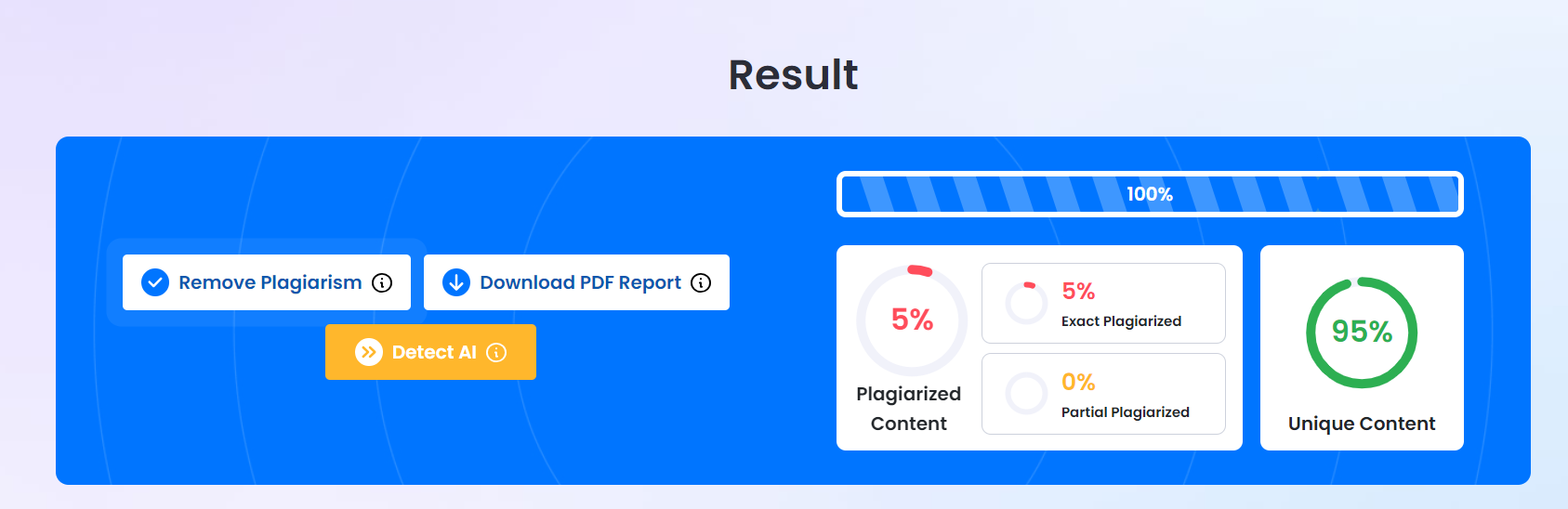
**Hive Table**

Apache Hive enables complex data analysis on Hadoop using SQL-like queries, providing a data warehousing solution within the Hadoop ecosystem.

**Apache Flink**

Apache Flink is an open-source framework which processes streams capable of handling large-scale data in real-time, supporting batch, stream, graph, and iterative processing.

Plagiarism Report:



References:

<https://data-flair.training/blogs/big-data-at-flipkart/>

<https://appscrip.com/blog/inside-flipkart-tech-stack-and-infrastructure/>