CASE STUDY OF BIG DATA IN FLIPKART

Submitted by :

Kiran prasannan

Mvoc-MPAD

DDUKK CUSAT

INTRODUCTION

Big data is data that contains greater variety, arriving in increasing volumes and with more velocity. This is also known as the three Vs. Put simply, big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can't manage them. Flipkart Data Platform is a service-oriented architecture that is capable of computing batch data as well as streaming data. This platform comprises of various micro-services that promote user experience through efficient product listings, optimization of prices, maintaining various types of data domains

What is Big Data

big data is data that contains greater variety, arriving in increasing volumes and with more velocity. This is also known as the three Vs.

Put simply, big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can't manage them. But these massive volumes of data can be used to address business problems you wouldn't have been able to tackle before.

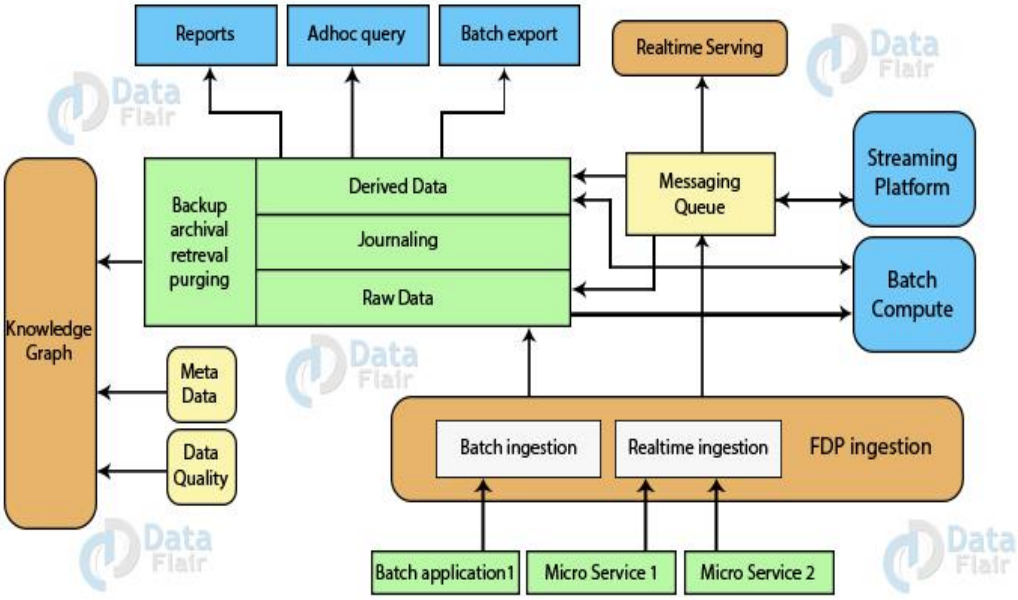
The three Vs of big data

Volume	The amount of data matters. With big data, you'll have to process high volumes of low-density, unstructured data. This can be data of unknown value, such as Twitter data feeds, clickstreams on a web page or a mobile app, or sensor-enabled equipment. For some organizations, this might be tens of terabytes of data. For others, it may be hundreds of petabytes.
Velocity	Velocity is the fast rate at which data is received and (perhaps) acted on. Normally, the highest velocity of data streams directly into memory versus being written to disk. Some internet-enabled smart products operate in real time or near real time and will require real-time evaluation and action.
Variety	Variety refers to the many types of data that are available. Traditional data types were structured and fit neatly in a relational database. With the rise of big data, data comes in new unstructured data types. Unstructured and semistructured data types, such as text, audio, and video, require additional preprocessing to derive meaning and support metadata.

Big Data at Flipkart

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

Architecture of Flipkart Data Platform



How Big Data is helping Flipkart

1. FPD Ingestion System

A Big Data Ingestion System is the first place where all the variables start their journey into the data system. It is a process that involves the import and storage of data in a database.

This data can either be taken in the form of batches or real-time streams. Simply speaking, batch consists of a collection of data points that are grouped in a specific time interval. On the contrary, streaming data has to deal with a continuous flow of data.

Batch Data has greater latency than streaming data which is less than sub-seconds. There are three ways in which ingestion can be performed –

- **Specter –** This is a Java library that is used for sending the draft to Kafka.
- **Dart Service –** This is a REST service which allows the payload to be sent over HTTP.
- **File Ingestor –** With this, we can make use of the CLI tool to dump data into the ***HDFS***.

Then, the user creates a schema for which the corresponding Kafka topic is created. Using Specter, data is then ingested into the FDP. The payload in the HDFS file is stored in the form of HIVE tables.

2. Batch Compute

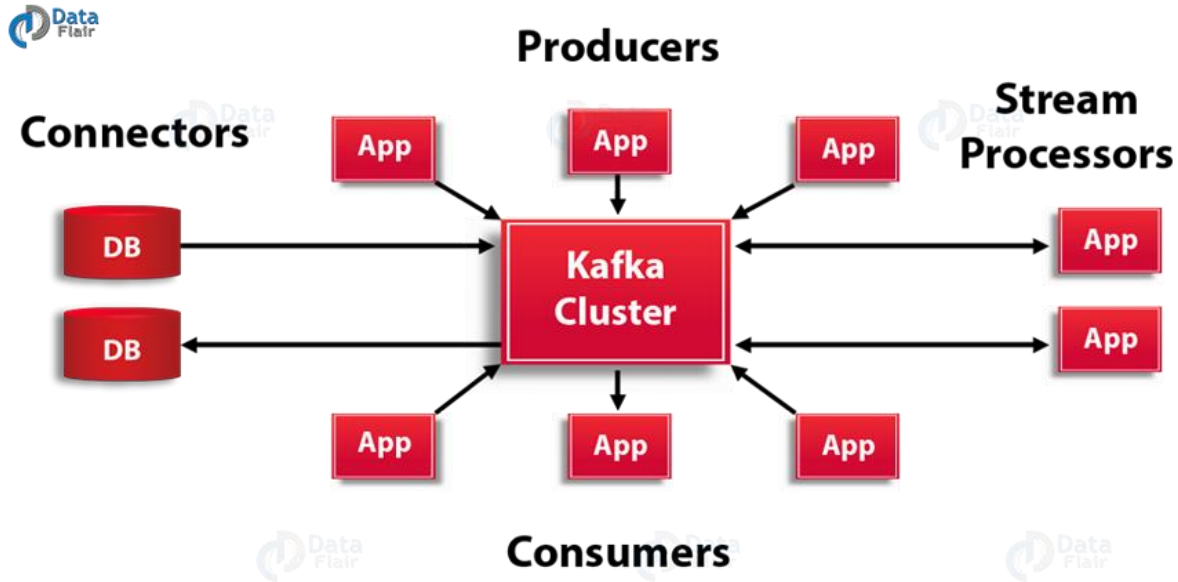
This part of the big data ecosystem is used for computing and processing data that is present in batches.Batch Compute is an efficient method for processing large scale data that is present in the form of transactions that are collected over a period of time. These batches can be computed at the end of the day when the data is collected in large volumes, only to be processed once.This is the time you need to explore Big Data as much as possible.

3. Streaming Platform

The streaming platforms process the data that is generated in sub-seconds. *Apache Flink* is one of the most popular real-time streaming platforms that are used to produce fast-paced analytical results**.** It provides a distributed, fault-tolerant and scalable data streaming capabilities that can be used by the industries to process a million transactions at one time without any latency.

4. Messaging Queue

A ***Messaging Queue*** acts like a buffer or a temporary storage system for messages when the destination is busy or not connected. The message can be in the form of a plain message, a byte array consisting of headers or a prompt that commands the messaging queue to process a task. There are two components in the Messaging Queue Architecture – Producer and Consumer. A Producer generates the messages and delivers them to the messaging queue. A Consumer is the end destination of the message where the message is processed



5. Real-time Serving

After the messages are retrieved from the Messaging Queue, the real-time serving system acts as a consumer for the messaging queue. With the help of this real-time serving platform, users can gather real-time insights from the data platform.Furthermore, with the help of real-time serving, the users can access the data through dynamic pipelines.

6. Data Lake

The core component of this architecture is the data storage platform. This is a ***Hadoop platform*** that stores raw data, journaled data as well as derived data. Using this, the data is stored in the form of a backup, archive that can be retrieved or purged according to the requirements.

The raw data is mostly used by the data scientists who use the insights from the original data to make decisions and develop data products. The data is present in the form of batches or real-time streams. The real-time data is in the form of click streams, summarized reports of user data, product insights, reviews, etc

7. Knowledge Graphs

Knowledge graphs represent an inter-linked network of real-world entities or objects through which we can extract information to process it in an efficient manner. This knowledge graph takes input from the meta-data. This metadata is beneficial for understanding the underlying semantics which is used for deriving newer facts.

The knowledge graph also makes use of various machine learning tools and libraries to gain insights and understand the relationships between the objects. One of the most popular tools that are used for building graph is Apache Spark's GraphX library.

Conclusion

Big Data provides some astonishing benefits to all kinds of businesses across the globe. From the education sector to the healthcare industry, almost every industry is now bound to Big Data Analytics in some or the other way.

Big Data is helping Flipkart to offer the best services all around the World. In this article, we looked at the ingenious big data platform that is designed by Flipkart to handle large scale data transactions. We also understood how Flipkart makes use of various big data components to deliver dynamic results to the user. We also had a look at how the Big Data Platform is capable of processing large scale data queries that allow it to produce results.