Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

BDI

Name: Shashwat Shah **Branch:** Computer Engineering

SAP ID: 60004220126 **Batch:** C22

EXPERIMENT NO. 1

Aim: Big Data Case Study with Hadoop Ecosystem.

Theory:

1. **Take examples of organization's such as Amazon, Google, Netflix etc. Flipkart.**

2. **Study the Big Data Approach they have chosen.**

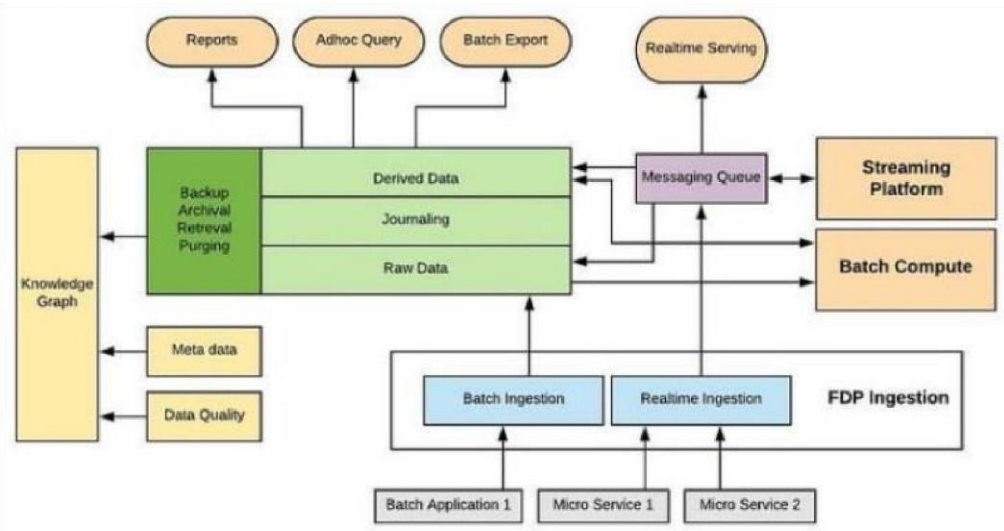
Flipkart's uses Big Data Tools and service oriented architecture (SOA) building thousands of microservices that power specific user experiences — such as search service, product listings service, pricing engine, estimating delivery dates etc. Each micro-service maintains domain data in the data store of their choice be it MySQL, HBase, Redis, Elasticsearch and more. Analytics, Insights, Data Science and even other microservice teams depend on data from multiple teams to run their business as usual. Flipkart Data Platform (FDP) provides the capabilities necessary for the teams to consume and act on this data.

FDP manages 800+ nodes Hadoop cluster to store more than 35 PB of data. They also run close to 25,000 compute pipelines on our Yarn cluster. Daily TBs of data is ingested into FDP and it also handles data spikes because of sale events. The tech stack majorly comprises of HDFS, Hive, Yarn, MR, Spark, Storm & other API services supporting the meta layer of the data.

3. **Discuss the Ecosystem Components they are using.**

Overall FDP can be broken down into following high level components.

1. Ingestion System.
2. Batch Data Processing System.
3. Real time Processing System.
4. Report Visualization.



1. Ingestion System

Each ingested payload has a fixed schema which can be created via a self serve UI. The tech stack for the system includes Messaging Queue (Kafka), Dropwizard, HDFS, Quartz, Azkaban & Hive. The user creates a schema for which a corresponding Kafka topic is created. Using Specter or Dart client the user can start ingesting the data to FDP. The ingestion system then stores the payload in HDFS files as raw Hive Tables and then journaling is run on top to make the data in the payload query able.

2. Batch Data Processing System

The ingested data which is now query able via hive queries is ready to be prepared for consumption. A user can create a Star Schema of Fact with multiple dimensions. The facts and dimensions can either be Hive tables.

3. Real time Processing System

The real time ingested data which happens via Dart or Specter clients can be directly consumed via their respective Kafka topics. Flipkart Streaming Platform (FSP) enables plugging up custom spark jobs to these topics. The streaming platform allows near real time aggregations to be built on all the ingested data.

4. Real time Processing System

Once the Analyst has created a Fact & Dimension in a columnar storage RDBMS. They can create boilerplate reports on top of the data using an in-house self serve ui. Analysts can come

and select the various metrics, filters & group by dimensions. According to the selection a set of allowed visualization charts are visible to the user.

Pros:

1. **Scalability:** Hadoop ecosystem tools like Hadoop Distributed File System (HDFS) and MapReduce are designed to scale horizontally, which means they can handle large amounts of data and can be easily expanded to meet growing data needs.

2. **Cost-effective:** Hadoop ecosystem tools are open-source, which means they are free to use and can be easily customized to meet specific business needs. This makes them an affordable solution for data management.

3. **Flexibility:** Hadoop ecosystem tools are versatile and can work with a wide range of data formats, including structured, semi-structured, and unstructured data. They can also integrate with a variety of other data processing tools and systems.

4. **Data processing:** Hadoop ecosystem tools are designed for batch processing of large amounts of data. This makes them ideal for Flipkart, which needs to process massive amounts of customer data to provide personalized recommendations, optimize supply chain management, and improve customer experience.

Cons:

1. **Complexity:** Hadoop ecosystem tools can be complex to set up and manage, requiring specialized skills and knowledge. This can add to the cost and time needed to implement and maintain the system.

2. **Performance:** Hadoop ecosystem tools are designed for batch processing, which means they may not be suitable for real-time processing of data. This can be a limitation for Flipkart, which needs to process data quickly to deliver real-time recommendations and insights to customers.

3. **Data security:** Hadoop ecosystem tools may not have the same level of security as traditional relational databases. This can be a concern for Flipkart, which needs to ensure the security of customer data.

4. **Lack of support:** Hadoop ecosystem tools are open-source, which means there is no formal support system available. This can be a challenge for Flipkart, which may need help troubleshooting issues or making updates to the system.

Conclusion: Flipkart uses big data tools and Hadoop ecosystem tools for Scalability, Cost-Effective, Flexibility, Data processing operations though this increases the complexity it is the requirement for effective management of data. The FDP system of flipkart uses multiple components of the Hadoop ecosystem to meet the increasing demands of industry.