Experiment No.6

Title: Case Study on Data Lake

Batch:B1 **Roll No.:16010420133** **Experiment No.:6**

Aim: To prepare a case study report on Data Lake.

Resources needed: Word Editor

Activities:

Prepare a report in the given section answering the following

1. Domain/Sector of the enterprise/organization
2. Name of the enterprise/organization
3. Challenges faced by enterprise/organization before using Data Lake
4. Implementation of Data Lake
    - Architecture/Framework used
    - Features
    - Usages
5. Benefits provided by Data lake
6. Your observations

DOMAIN

ecommerce Organization

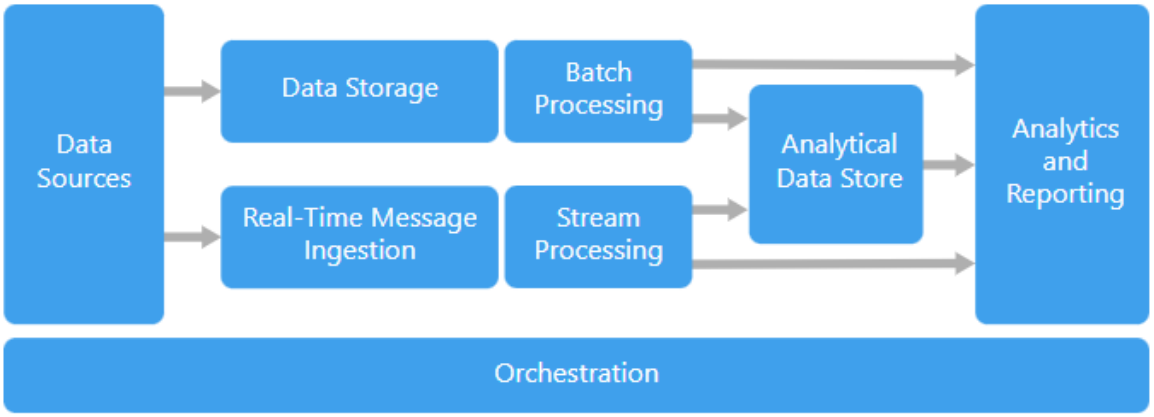
Organization Name

Flipkart

Problems faced before using big data

- Strategy
- Marketing
- retail,
- products
- logistics
- consumer experience to corporate strategy

Architecture



Big data comprises the following components:

1. Data Sources

Data sources govern Big Data architecture. It involves all those sources from where the data extraction pipeline gets built. Data Sources are the starting point of the big data pipeline.

Data arrives through multiple sources including relational databases, sensors, company servers, IoT devices, static files generated from apps such as Windows logs, third-party data providers, etc. This data can be batch data or real-time data.

Big Data architecture is designed in such a way that it handles this vast amount of data.

2. Data Storage

Data Storage is the receiving end for Big Data. Data Storage receives data of varying formats from multiple data sources and stores them. It even changes the format of the data received from data sources depending on the system requirements.

For example, Big Data architecture stores unstructured data in distributed file storage systems like HDFS or NoSQL database. It stores structured data in RDBMS.

3. Real-time Message Ingestion

We need to build a mechanism in our Big Data architecture that captures and stores real-time data that is consumed by stream processing consumers. It is simply a datastore where the new messages are dropped inside the folder.

There are a number of solutions that require the necessity of a message-based ingestion store that acts like a message buffer and supports scale based processing. They provide reliable delivery along with the other messaging queuing semantics.

It may include options like Apache Kafka, Event hubs from Azure, Apache Flume, etc.

4. Batch Processing

The architecture requires a batch processing system for filtering, aggregating, and processing data which is huge in size for advanced analytics.

These are generally long-running batch jobs that involve reading the data from the data storage, processing it, and writing outputs to the new files. The most commonly used solution for Batch Processing is Apache Hadoop.

5. Stream Processing

There is a little difference between stream processing and real-time message ingestion. Stream processing handles all streaming data which occurs in windows

or streams. It then writes the data to the output sink. It includes Apache Spark, Storm, Apache Flink, etc.

6. Analytical Data Store

After processing data, we need to bring data in one place so that we can accomplish an analysis of the entire data set. The analytical data store is important as it stores all our process data at one place making analysis comprehensive.

It is optimized mainly for analysis rather than transactions. It can be a relational database or cloud-based data warehouse depending on our needs.

7. Analytics and Reporting

After ingesting and processing data from varying data sources we require a tool for analyzing the data.

For this, there are many data analytics and visualization tools that analyze the data and generate reports or a dashboard. Companies use these reports for making data-driven decisions.

8. Orchestration

Moving data through these systems requires orchestration in some form of automation.

Ingesting data, transforming the data, moving data in batches and stream processes, then loading it to an analytical data store, and then analyzing it to derive insights must be in a repeatable workflow. This allows us to continuously gain insights from our big data.

Features

1. Velocity

Volume refers to the amount of data that you have. We measure the volume of our data in Gigabytes, Zettabytes (ZB), and Yottabytes (YB). According to the industry trends, the volume of data will rise substantially in the coming years.

2. Volume

Velocity refers to the speed of data processing. High velocity is crucial for the performance of any big data process. It consists of the rate of change, activity bursts, and the linking of incoming data sets.

3. Value

Value refers to the benefits that your organization derives from the data. Does it match your organization's goals? Does it help your organization enhance itself? It's among the most important big data core characteristics.

4. Variety

Variety refers to the different types of big data. It is among the biggest issues faced by the big data industry as it affects performance. It's vital to manage the variety of your data properly by organizing it. Variety is the various types of data that you gather from different kinds of sources.

5. Veracity

Veracity refers to the accuracy of your data. It is among the most important Big Data characteristics as low veracity can greatly damage the accuracy of your results.

6. Validity

How valid and relevant is the data to be used for the intended purpose.

7. Volatility

Big data is constantly changing. The data you gathered from a source a day ago might be different from what you found today. This is called variability of data, and it affects your data homogenization.

8. Visualization

Visualization refers to showing your big data-generated insights through visual representations such as charts and graphs. It has become prevalent recently as big data professionals regularly share their insights with non-technical audiences.

Usage

Here is the list of the top 10 industries using big data applications:

1. Banking and Securities
2. Communications, Media and Entertainment
3. Healthcare Providers
4. Education
5. Manufacturing and Natural Resources
6. Government
7. Insurance
8. Retail and Wholesale trade
9. Transportation
10. Energy and Utilities

Benefits

- Using big data cuts your costs. ...
- Using big data increases your efficiency. ...
- Using big data improves your pricing. ...
- You can compete with big businesses. ...
- Allows you to focus on local preferences. ...
- Using big data helps you increase sales and loyalty.
- Using big data ensures you hire the right employees.

1. My observations : Although the term "big data" has never been precise, the concept of "big data"—which really is "all" data—and the value of its analysis, is still extremely valid, and new sources of large-scale, unstructured data continue to emerge.
2. Data products, initially offered just by native digital companies (e.g., Google, Facebook, and LinkedIn) have become important business objectives for companies across many industry sectors.
3. While the skills to deal with big data are becoming much more widely distributed, management awareness and understanding of the business potential of big data remain in short supply, and the technology may be outpacing the ability of organizations to deploy and manage it effectively.

Questions:

Explain how Data Lake differs from traditional analytical systems.

Characteristics	Data Warehouse	Data Lake
Data	Relational from transactional systems, operational databases, and line of business applications	Non-relational and relational from IoT devices, web sites, mobile apps, social media, and corporate applications
Schema	Designed prior to the DW implementation (schema-on-write)	Written at the time of analysis (schema-on-read)
Price/Performance	Fastest query results using higher cost storage	Query results getting faster using low-cost storage
Data Quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (ie. raw data)
Users	Business analysts	Data scientists, Data developers, and Business analysts (using curated data)
Analytics	Batch reporting, BI and visualizations	Machine Learning, Predictive analytics, data discovery and profiling

---

Outcomes: CO3 Understanding of Data warehousing and data modelling .

Conclusion: (Conclusion to be based on outcomes achieved)

Made a report on big data .

Grade: AA / AB / BB / BC / CC / CD /DD

Signature of faculty in-charge with date

References:

https://docs.microsoft.com/en-us/azure/architecture/guide/architecture-styles/big-data#:~:text=A%20big%20data%20architecture%20is,big%20data%20sources%20at%20rest.

https://www.babson.edu/academics/executive-education/babson-insight/analytics-and-big-data/10-observations-on-big-data-in-late-2015/

https://www.simplilearn.com/tutorials/big-data-tutorial/big-data-applications

https://www.analytixlabs.co.in/blog/characteristics-of-big-data/

https://techvidvan.com/tutorials/big-data-at-flipkart/

https://www.guru99.com/what-is-big-data.html

https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/?nc=sn&loc=2