Introduction to Data Engineering

Prerequisites

Before we begin, it's important to understand why we need something like big data and the fundamentals of big data, which will be the focus of today's lecture. In the previous sections of this course, you may have learned about SQL and some aspects of system design. The fundamentals from those topics will be applied here, so if you need revision, please take some time to review them. Ok, so let's start

Big Data Evolution

The first question that should come to your mind is why Big Data, why did we need big data, seems like before big data came, companies still worked fine, storing data, and were able to serve all the requests, and all customers seemed to be happy, nothing seems to be missing. **Let's try to understand that with an example of how a company like Flipkart would have evolved regarding data requirements.** I am not saying Flipkart would have met its data requirements exactly like this, this is just to give you an idea.

Problem 1

Initially, **Flipkart likely stored all its data in a MySQL database**. When an order was placed, the details would be stored in MySQL, and the same database would be used to retrieve order details whenever needed. This MySQL setup handled all use cases, including storing and retrieving data.

However, as the business grew, the amount of data increased significantly. Analysts needed to run queries to understand how many orders were placed in a day, total sales for a specific period, and other metrics. These analytical queries became computationally heavy, which started to impact the performance of the production database.

To mitigate this, **a Master-Slave architecture might have been implemented**, where **the slave databases were used for running analytical queries**. However, this approach also encountered problems:

1. **Data Volume**: The data grew so large that it couldn't fit into a single MySQL database. While horizontal scaling using NoSQL databases could help with

storage, these databases weren't designed for complex analytical queries. As a result, the queries became slow or even failed.
2. **Design Limitations**: The existing systems weren't optimized for the type of queries needed for analytical purposes.

This situation highlighted the need for specialized software designed specifically for analytical use cases, capable of efficiently handling large volumes of data.

Problem 2

As industries expand, companies increasingly aim to store every possible type of data and leverage it to make informed business decisions that drive profits. This data can range from simple information like order details to more complex types like web pages, images, and files. However, **traditional databases like MySQL are not designed to handle such diverse and large-scale use cases effectively. So now it makes sense to have some specialized software that could meet the demand of industry data needs.**

With this, we can conclude that there are broadly two different use cases one transactional and one analytical and they need different systems to handle them. Now let's try to understand each of these systems.

OLTP

Flipkart, as a leading e-commerce platform, handles a massive volume of transactions every day. **These transactions include customer orders, payments, product inventory updates, and more. To manage these operations efficiently, Flipkart relies on an OLTP system.**

What is an OLTP System?

An OLTP system is designed to handle a large number of short, quick online transactions. It focuses on fast query processing, maintaining data integrity in multi-access environments, and supporting large numbers of users performing numerous transactions.

Role of OLTP in Flipkart's Order Management

1. **Order Placement:**
    - When a customer places an order on Flipkart, the OLTP system captures and stores the order details in real-time. This includes information like the customer ID, product ID, quantity, payment details, shipping address, and order status.

    - The system ensures that each order transaction is completed correctly, including processing payments and updating the inventory to reflect the stock decrease.
2. **Inventory Management:**
    - The OLTP system immediately updates the inventory as soon as an order is placed. If a product is out of stock, the system will prevent further orders for that item, ensuring accurate inventory levels and preventing over-selling.
3. **Payment Processing:**
    - The OLTP system securely handles payment transactions, ensuring that payments are processed accurately and efficiently. It interacts with multiple payment gateways to process credit card transactions, digital wallets, and other payment methods.
    - It ensures that each payment is securely recorded and associated with the correct order, enabling smooth order fulfillment.
4. **Order Tracking and Updates:**
    - The OLTP system allows customers to track their orders in real-time. As the order status changes (e.g., from "Processing" to "Shipped" to "Delivered"), the system updates the order details, ensuring that both the customer and Flipkart's logistics teams have up-to-date information.
    - Customers can also make changes to their orders (like updating the delivery address) before the order is shipped, and the system processes these changes in real-time.
5. **Data Integrity and Concurrency:**
    - Since multiple users (customers, employees, and automated systems) access and update the database simultaneously, the OLTP system ensures data integrity through ACID (Atomicity, Consistency, Isolation, Durability) properties.
    - For example, if two customers try to order the last unit of a product simultaneously, the OLTP system ensures that only one order is processed, preventing any conflict or data corruption.

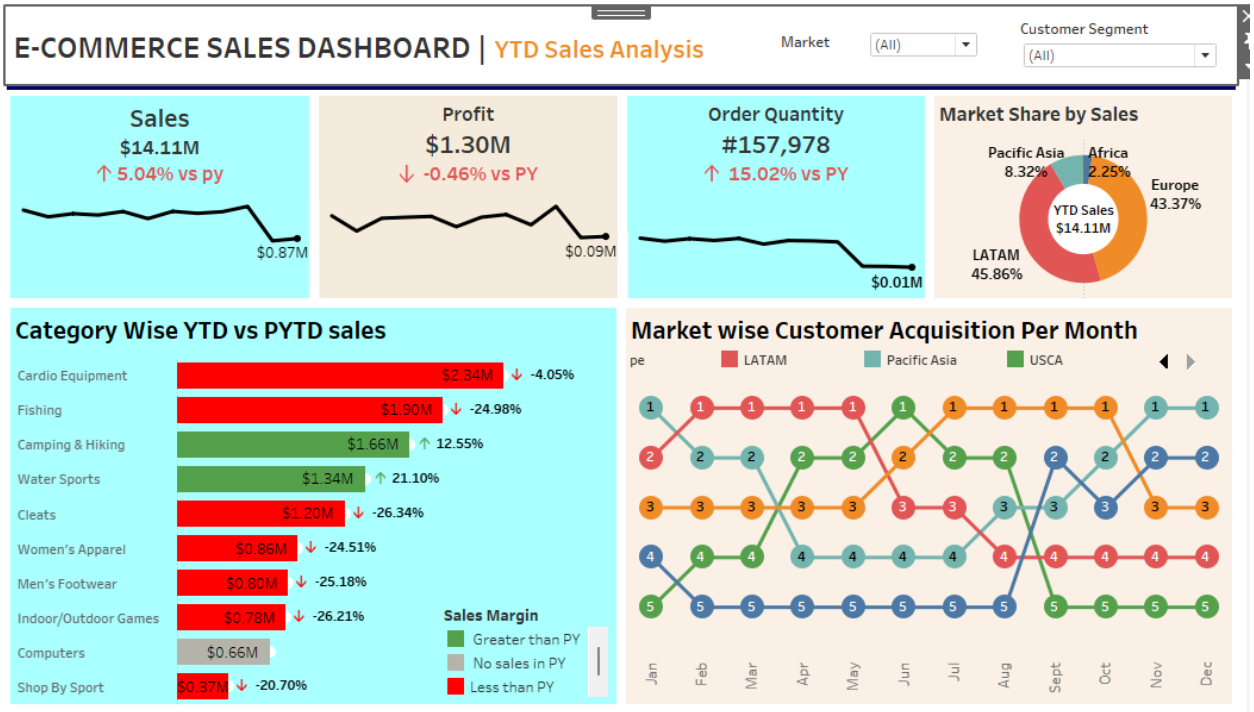
OLAP (Online Analytical Processing)

Flipkart, like any large e-commerce platform, not only needs to process transactions efficiently but also requires deep insights into its operations to make informed business decisions. This is where OLAP (Online Analytical Processing) systems come into play

What is an OLAP System?

An OLAP system is designed for complex queries and data analysis, providing insights from large volumes of data. Unlike OLTP systems, which handle day-to-day transactional data, OLAP systems are optimized for querying and reporting, often using aggregated historical data to help businesses identify trends, patterns, and insights.

Role of OLAP in Flipkart's Order Analysis



E-commerce dashboard powered via

1. **Sales Trend Analysis:**
    - This analysis helps Flipkart identify peak sales periods, forecast demand, and adjust marketing strategies accordingly.
2. **Customer Behavior Analysis:**
    - For instance, Flipkart can analyze which age group prefers to buy electronics during a sale, helping the company to tailor marketing campaigns and product recommendations to different customer segments.
3. **Product Performance Analysis:**
    - It can identify underperforming products, allowing Flipkart to take corrective actions such as running promotions or discontinuing certain products.
4. **Profitability Analysis:**

    - This analysis can highlight the most profitable segments, enabling Flipkart to focus its resources on high-margin products or regions.
5. **Supply Chain Optimization:**
    - Flipkart's OLAP system can analyze data related to order fulfillment times, shipping costs, and supplier performance. For example, it can identify delays in specific regions or with certain suppliers, allowing Flipkart to optimize its supply chain.
6. **Personalization and Recommendations:**
    - The OLAP system can be leveraged to analyze browsing and purchasing data to improve the accuracy of product recommendations. For example, by analyzing what products are frequently bought together, Flipkart can offer better bundle deals or upsell opportunities.

Now that we have understood all about these systems, let's check if we have understood them properly by identifying which system you would use in these scenarios

1. A hospital uses an electronic health record (EHR) system where doctors and nurses enter patient information, such as medical history, prescriptions, and lab test results, during each patient visit. This information needs to be updated immediately and accessed by multiple healthcare professionals in real-time to ensure accurate patient care.
    a. Right, you will use OLTP
2. The hospital's management wants to analyze patient data from the last five years to identify trends in chronic disease diagnoses, treatment outcomes, and readmission rates. This analysis will help the hospital improve its long-term care strategies and optimize resource allocation for specific health issues.
    a. Right, you will use OLAP

Cool, now lets move forward.

Now let's deep dive into what defines big data. To understand that let's take a look at 6V's
Of Big Data.

6 V’s of Big Data

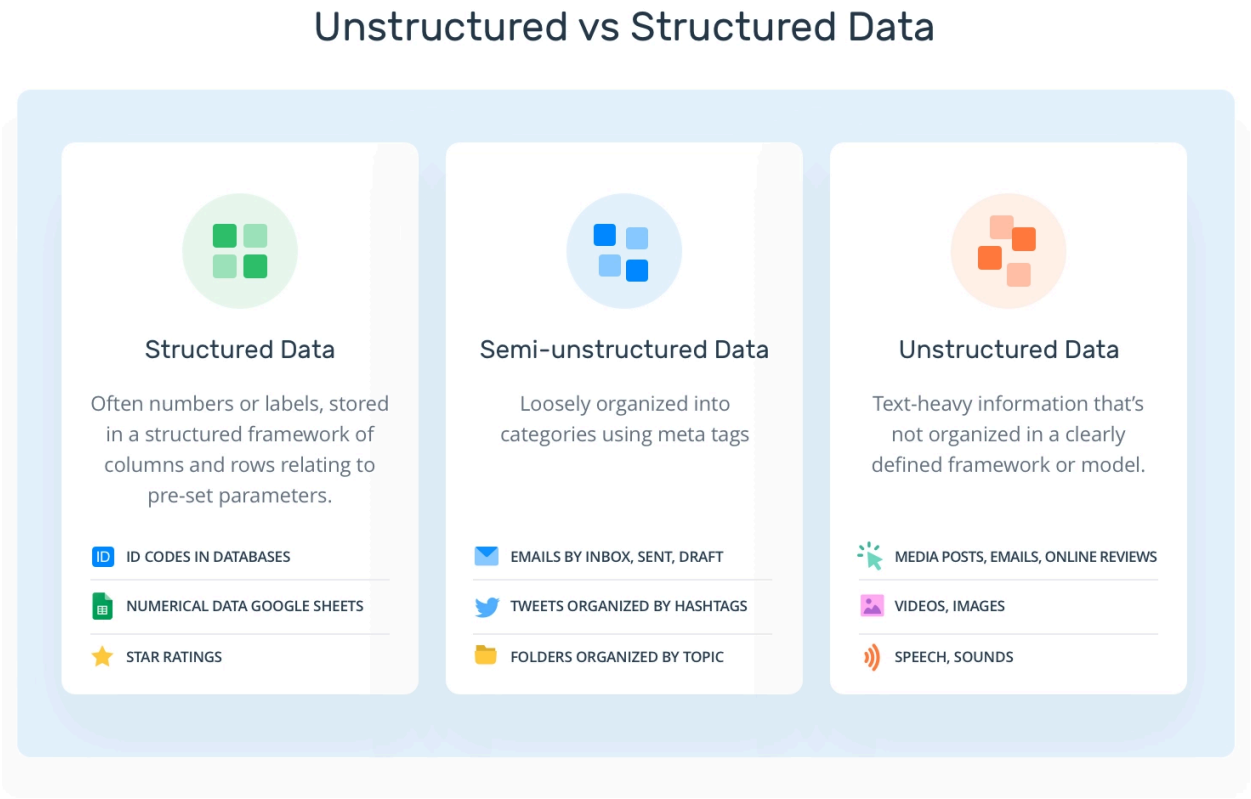
Volume

- Flipkart generates vast amounts of data every day, including transactional data (orders, payments), user interaction data (browsing history, search queries), and logistical data (shipping, delivery).
- For example, during major sales events like the Big Billion Days, the volume of data spikes dramatically as millions of customers interact with the platform simultaneously. The amount of adat being generated can span around **100 TB per day**.
- Managing and storing this massive volume of data requires robust infrastructure, including distributed storage systems and scalable databases.

Variety

Every company wants every data they can get, so there is so much variety of data. This data can be broadly categorized into 3 types

a. Structured
b. Semi-Structured
c. Unstructured



Structured data
It is the data that follows a rigid format and can be organized neatly into rows and columns as structured data. It has a well-defined structure and can be stored in well-defined schemas such as relational databases.

Ex: customer information, product catalogs, and transaction records

Semi-Structured data
Let's take another use case of sending an email when you booked the flight tickets, primary attributes are like,
email (sender_email, receiver_emails, subject, body, attachments)
Again seems like well-structured data, but is it?
No, actually it's **semi-structured**, because of the different nature of parameters in the schema (it's not that rigid!)
e.g, The 'receiver_emails' column is very much possibly a JSON, like,

```
{
  "direct": "harshit@scaler.com",
```

```
  "cc": "mudit@scaler.com",
  "bcc":[
    "anshuman@scaler.com",
    "abhimanyu@scaler.com"
  ]
}
```

Apart from this, you would have file attachments as well. It can be a *.pdf, .png* etc.

Definition: Semi-structured data is a mix of data that has consistent characteristics but data that doesn't conform to a rigid structure. It contains tags and elements, or metadata, which is used to group and organize it.

Unstructured Data

- Other types of data in our example could be GPS (location) data, the current latitude and longitude along with other details in native or raw formats are going to be unstructured data.
- More examples of unstructured data could be content like
    - photos
    - videos
    - text files,
    - PDFs
    - social media posts

Definition: Data that does not have an easily identifiable structure and, therefore, cannot be organized in a mainstream relational database in the form of rows and columns can be termed as unstructured data. It does not follow any particular format, sequence, semantics, or rules.

Being able to leverage the potential of variable data sources and types, structured, semi-structured and unstructured. Integrating diverse data into a manageable structure is key to a robust big data opportunity

Velocity

How quickly we can ingest and process data is also known as velocity.

Types of Data processing based on velocity

Data at Flipkart is generated at a high velocity, especially during peak shopping periods. This includes real-time data from customer clicks, purchases, and social media interactions. To alayize data like these we need to process the data in real time and these are the examples of **Stream Processing** (a type of Velocity).

- **Definition: Stream processing allows you to feed data into analytics tools as soon as they get generated and get instant analytics results.**
- Tech: There are multiple open-source stream processing platforms such as Apache Kafka, Apache Flink, Apache Storm, etc.
- **Other Use Cases**
    1. Social media sentiment analysis (are the people happy-sad about the news, budget etc.)
    2. Log monitoring (analyze the logs generated by a system and send an alert if it encounters words like FAIL, FAILURE, FAILED etc. It will help bring the issue to our attention, in real-time and may prevent system outage)
    3. Fraud detection
       If you stream-process transaction data, you can detect anomalies that signal fraud in real-time, then stop fraudulent transactions before they are completed. Savior, isn't it?

The other type of velocity is **Batch Processing.**

Example: Let us say your task is to analyze transactional data of a major financial firm over a day, every day so that you can understand any pattern which can help in business/revenue.

This data contains millions of records for a day that can be stored as a file or record etc. This particular file will undergo processing at the end of the day for various analysis that the firm wants to do. It will take a large amount of time for that file to be processed. This is a common example of Batch Processing.

- **Definition: Batch processing works well in situations where you don't need real-time analytics results, and when it is more important to process large volumes of data to get more detailed insights than it is to get fast analytics results.**
- Tech: Prominently, Hadoop MapReduce is being utilized in the industry
- Use case: Payroll, Billing, Orders from customers

  To take advantage of geolocation data, perceived hypes and trends, and real-time available market and customer information, we need to process it

fast. The data platform system built for it should have such capabilities.

Value

There is data in abundance. But, not all data points may be relevant for all businesses.
For example,

- The ultimate goal of Flipkart's data strategy is to extract value from the vast amounts of data it collects. This value is realized through better decision-making, improved customer experiences, and increased operational efficiency.
- For example, by analyzing customer purchase patterns, Flipkart can offer personalized recommendations, leading to increased sales and customer satisfaction.
- Data-driven insights also help Flipkart optimize its supply chain, reduce delivery times, and manage inventory more effectively, thereby enhancing overall business performance.

Definition: Value of data means understanding the potential to create revenue or unlock opportunities through your data. If it is not valuable, then questions should be raised about why and where to store it.

Veracity

Veracity refers to the trustworthiness and accuracy of the data.

- Flipkart must ensure that the data it collects is accurate and reliable. This is critical for making informed decisions, such as inventory management, pricing strategies, and personalized recommendations.
- Issues like duplicate records, incomplete data, and false information (e.g., fake reviews) can lead to poor decision-making.
- To ensure data veracity, Flipkart employs data cleaning, validation processes

.

Definition: Identifying the relevance, correctness or accuracy of data, and applying it to the appropriate purposes.

Understanding data relevance is key to value.
In short: the truth and authenticity of the data, and what can you do with it? In a sense, it is a hygiene factor. By showing the veracity of your data, you show that you have taken a critical look at it.

Variability

Variability refers to the inconsistencies and fluctuations in data over time.

- Flipkart's data can vary significantly depending on factors like seasonality, marketing campaigns, and external events. For instance, customer behavior during festive seasons or sales events can differ drastically from regular days.
- **The challenge for Flipkart is to manage this variability and adapt its systems to handle the fluctuating data loads and changing patterns.**
- Predictive analytics and machine learning models are used to anticipate variability and adjust strategies in real-time, ensuring that Flipkart remains agile and responsive to market dynamics.

What do Data Engineers do?

Problem

Flipkart wants to enhance its recommendation system to increase customer engagement and sales. The goal is to provide more personalized product recommendations based on user behavior, past purchases, and browsing patterns.

First let's take a moment to understand what data engineers do, how they interact with other profile engineers, and what sets them apart. Data engineering is at the intersection of software engineering and data science. **Data engineers** consume data generated by applications written by **Software Developers** and build and manage the data infrastructure that **Data Scientists** use to perform their analyses. They ensure that data is accessible, clean, and ready for analysis.

Role of Data Engineers

1. Data Collection and Ingestion

- **Scenario:** To improve the recommendation system, Flipkart must collect data from multiple sources like website clicks, mobile app interactions, purchase histories, product reviews, and more.
- **Data Engineer's Role:** Data Engineers are responsible for designing and implementing data pipelines that collect this data from various sources in real time. They ensure that data from the website, mobile app, and other platforms are ingested into a centralized data repository, such as a data lake or a data warehouse.

    - **Example:** They might build ETL (Extract, Transform, Load) pipelines that capture user interaction data every time a customer views a product, adds it to the cart, or makes a purchase. This data is then transformed into a consistent format and loaded into Flipkart's data warehouse.

2. Data Storage and Management

- **Scenario:** Flipkart needs a robust system to store vast amounts of structured and unstructured data securely and efficiently.
- **Data Engineer's Role:** Data Engineers design and manage the data architecture, ensuring that the data is stored in a scalable and efficient manner. They work on database management systems, data warehousing solutions, and data lakes to handle storage needs.
    - **Example:** They might set up a distributed storage system using technologies like Hadoop or Amazon S3 to store petabytes of data. They also ensure that the data is partitioned and indexed correctly, enabling quick access and processing.

3. Data Processing and Transformation

- **Scenario:** The raw data collected from various sources is not immediately ready for analysis. It needs to be cleaned, normalized, and transformed.
- **Data Engineer's Role:** Data Engineers develop processes to clean and transform this data into a usable format. They work with big data processing frameworks like Apache Spark or Apache Flink to perform these transformations at scale.
    - **Example:** They might develop Spark jobs that clean the data by removing duplicates, handling missing values, and transforming categorical variables into numerical formats, making the data ready for machine learning models.

4. Ensuring Data Quality and Consistency

- **Scenario:** The recommendation system requires high-quality, consistent data to function effectively.
- **Data Engineer's Role:** Data Engineers implement data validation and monitoring systems to ensure that the data being fed into the recommendation engine is accurate and consistent.
    - **Example:** They might set up automated checks to validate data consistency across different data sources, ensuring that a customer's purchase history is correctly reflected in all related datasets.

Role of Data Scientists

1. Data Exploration and Analysis

- **Scenario:** With the cleaned and processed data available, Flipkart needs to understand customer behavior patterns to build an effective recommendation system.
- **Data Scientist's Role:** Data Scientists explore and analyze the data to extract meaningful insights. They use statistical techniques and data visualization tools to understand patterns, trends, and relationships in the data.
    - **Example:** A Data Scientist might analyze browsing patterns to identify which products are often viewed together or which types of products are purchased by similar customer segments. They might use clustering algorithms to segment customers based on their shopping behavior.

2. Building Predictive Models

- **Scenario:** The core of the recommendation system relies on predictive models that can suggest products to customers based on their behavior.
- **Data Scientist's Role:** Data Scientists develop and train machine learning models to predict which products a customer is likely to buy next. They experiment with different algorithms, such as collaborative filtering, content-based filtering, or hybrid models, to find the best approach.
    - **Example:** They might build a collaborative filtering model that recommends products based on the purchase histories of similar customers or a content-based model that recommends products similar to those a customer has shown interest in.

3. Model Evaluation and Optimization

- **Scenario:** To ensure the recommendation system performs well in a real-world setting, the models need to be rigorously tested and optimized.
- **Data Scientist's Role:** Data Scientists evaluate the performance of their models using metrics like precision, recall, and F1 score. They perform A/B testing to compare different model versions and select the one that provides the best recommendations.
    - **Example:** A Data Scientist might test the recommendation model on a subset of users to compare the engagement rates (click-through rate, conversion rate) against a control group. Based on the results, they might fine-tune the model's parameters or features.

4. Deploying and Monitoring Models

- **Scenario:** Once the models are ready, they need to be deployed into Flipkart's production environment to start generating recommendations in real-time.
- **Data Scientist's Role:** In collaboration with Data Engineers, Data Scientists deploy the trained models into production. They also set up monitoring to track the model's performance over time and ensure it continues to deliver accurate recommendations.
    - **Example:** The deployed model might run in a microservices architecture where it listens to customer interactions in real-time and generates product recommendations. The Data Scientist monitors the model's accuracy and makes updates as needed based on new data.

Collaboration Between Data Engineers and Data Scientists

In a project like enhancing Flipkart's recommendation system, Data Engineers and Data Scientists must work closely together:

- **Data Pipeline Design:** Data Engineers work with Data Scientists to understand the data requirements for building models and ensure the data is available in the right format.
- **Model Deployment:** Once the models are developed, Data Engineers help deploy them into production, ensuring they integrate seamlessly with Flipkart's existing infrastructure.
- **Ongoing Optimization:** Data Engineers might create automated pipelines that regularly update the model's training data, while Data Scientists continue to refine the models based on real-world performance.

Skillset required to become a Data Engineer

Now we have understood what Big data is and what Data Engineer does to handle this data, now let's take a look at what skill sets are required to play with all these data and related technologies

Programming Skills: Proficiency in languages like Python, Java, or Scala.
SQL: Strong skills in SQL for querying and manipulating relational databases.
Data Warehousing: Experience with data warehousing solutions
Big Data Technologies: Knowledge of big data frameworks like Apache Hadoop, Apache Hive, Apache Spark, or Kafka for handling large-scale data processing.
ETL Tools: Familiarity with ETL (Extract, Transform, Load) tools and frameworks
Cloud Platforms: Experience with cloud services from providers like AWS, Google Cloud for data storage and processing.
Database Management: Basic Understanding of both SQL (MySQL, PostgreSQL) and NoSQL (MongoDB, Cassandra) databases.

Data Modeling: Skills in designing data models and schemas to optimize data storage and retrieval.