What is Big Data and Machine Learning

1. What is Big Data?

Big Data refers to vast amounts of structured and unstructured data that are difficult to process using traditional database tools. It's crucial for industries that manage data from various sources like social media, web pages, and cloud services.

🔍 Challenges in Big Data:

- Capturing
- Curating
- Storing
- Searching
- Sharing
- Transferring
- Analyzing
- Visualizing

---

2. What is Machine Learning (ML)?

Machine Learning is a subset of AI where machines learn from past data and improve performance without being explicitly programmed.

🌟 Applications of ML:

- Image & speech recognition
- Healthcare
- Banking & finance
- Self-driving cars
- Virtual assistants
- Education
- Marketing & trading

3. Difference Between Big Data and Machine Learning:

Machine Learning	Big Data
Learns from data to make predictions	Deals with large datasets
Types: Supervised, Unsupervised, etc.	Types: Structured, Unstructured, Semi-structured
Uses: Scikit-learn, TensorFlow, etc.	Uses: Hadoop, MongoDB, etc.
Less complex data	High-dimensional, complex data
Less human intervention needed	Often requires human intervention

---

4. Big Data + Machine Learning

- ML helps in **analyzing** and **making sense** of Big Data.
- Big Data provides the **huge and diverse datasets** that ML needs.
- Together, they are used in companies like **Google, Amazon, Netflix**, etc.

---

5. How ML Helps in Big Data

- **Segments** and organizes the data.
- Finds **patterns** and insights.
- Helps in **fast decision making**.
- Converts raw data into **useful information**.

Recommender Systems in BigData

1. Recommender Systems Overview

A recommender system helps suggest items to users based on their preferences and behavior. It predicts what users may like and recommends top items.

---

2. Types of Recommender Systems

1. 📚 Content-Based Filtering

- Uses item **attributes/tags** (like genre, brand, etc.).
- Recommends similar items based on user's previous preferences.
- Example: If a user rates **action movies** highly, suggest more action movies.

2. 👥 Collaborative Filtering

- Based on **user-item interactions**.
- Uses patterns from multiple users.

a. User-User Filtering:

- Finds users similar to the target user.
- Recommends items liked by similar users.

b. Item-Item Filtering:

- Finds items rated similarly by users.
- Recommends similar items based on **user behavior**, not item tags.

3. 🔄 Hybrid Filtering

- Combines **Content-Based** and **Collaborative Filtering**.
- Methods:
    1. Combine both outputs.
    2. Apply one after the other (e.g., collaborative → content-based).

🧠 Cosine Similarity Formula

$$\text{sim}(A, B) = \frac{A \cdot B}{||A|| \cdot ||B||}$$

- Measures similarity between two users or items.

---

3. Types of Feedback

Type	Description
Explicit	Ratings, likes, reviews from users
Implicit	Clicks, views, browsing behavior

---

4. Big Data Tools for Recommender Systems

◆ Data Storage & Processing

- **Hadoop** – Large-scale distributed storage
- **Apache Spark** – Fast in-memory processing

◆ Recommendation Frameworks

- **Apache Mahout** – Scalable machine learning
- **TensorFlow Recommenders (TFRS)** – Deep learning for recommendations
- **Surprise (Python)** – For collaborative filtering

◆ Databases

- **MongoDB**, **DynamoDB** – Handle large unstructured interaction data

◆ Visualization Tools

- **Power BI**, **Tableau**, or Python libraries (matplotlib, seaborn, plotly)

---

5. Applications

Domain	Example Use
E-commerce	Product recommendations (Amazon)
Entertainment	Movie/music suggestions (Netflix, Spotify)
Education	Course suggestions (Coursera, Khan Academy)
Social Media	Friends, pages, posts (Facebook, Instagram)
Healthcare	Personalized treatment or drug suggestions

---

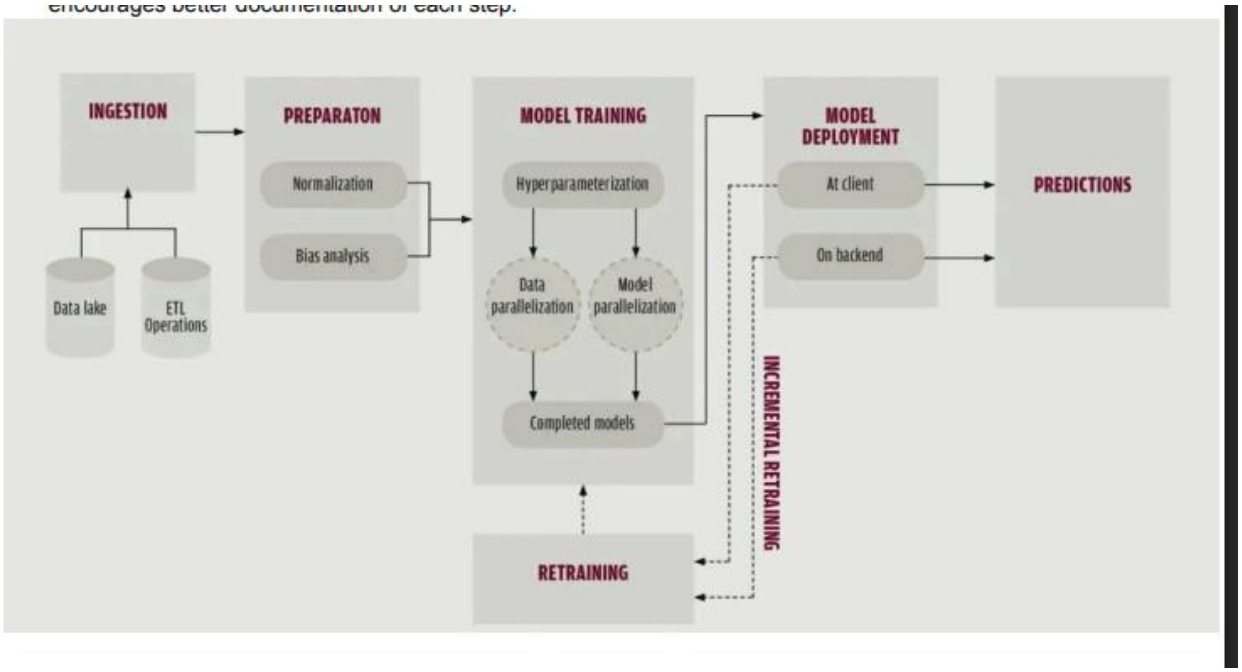
6. Challenges in Big Data Recommender Systems

1. **Cold Start Problem** – No past data for new users/items
2. **Data Sparsity** – Not enough user-item interactions
3. **Scalability** – Handling billions of users/items
4. **Bias and Fairness** – Avoiding skewed suggestions
5. **Real-Time Processing** – Recommending instantly as users interact

User-based collaborative filtering

- **Nummericals**

Graph analytics (Quiz)



ETL (Extract, Transform, Load) operations are performed on the data to extract relevant information, transform it into a suitable format, and load it into a database or data warehouse.

The image shows a machine learning (ML) pipeline workflow, which is a series of steps to build, train, and deploy ML models. Here's a simplified breakdown:

1. Ingestion: Collect and store data.

2. Preparation: Clean and prepare data for use.

3. Model Training: Train the ML model using prepared data.

4. Model Deployment: Deploy the trained model to a production environment.

5. Predictions: Use the deployed model to make predictions on new data.

6. Retraining: Retrain the model on new data to improve its performance over time.

These steps work together to create a continuous cycle of improvement, ensuring that the ML model remains accurate and effective.