Tools and Frameworks for the Machine Learning Pipeline

Your Company Name

February 22, 2025

Contents

1 Data Ingestion and Streaming

- **Apache Kafka or AWS Kinesis:** For real-time streaming of user interactions and review data into the processing pipeline.

2 Data Storage

2.1 Raw Data

- **AWS S3 or Google Cloud Storage:** To store raw JSON or CSV files from ingested data.

2.2 Structured Data

- **PostgreSQL:** For storing user profiles, metadata, and other relational data.

2.3 Unstructured Data

- **MongoDB:** For flexible storage of review text, comments, and session logs.

3 Data Processing and Preprocessing

- **Apache Spark:** For large-scale data processing and feature engineering (batch processing).
- **Python Libraries:**
  - **Pandas and NumPy:** For data manipulation.
  - **NLTK or spaCy:** For natural language processing and text preprocessing.
- **Jupyter Notebooks:** For exploratory data analysis and iterative development.

4 Machine Learning Model Development

- **Deep Learning Frameworks:**
  - **TensorFlow and/or PyTorch:** For building and training deep learning models.
- **Classical Machine Learning:**
  - **scikit-learn:** For collaborative filtering, clustering, and other classical ML approaches.
- **Pre-trained Models and NLP Libraries:**
  - **Hugging Face Transformers:** For leveraging pre-trained models such as BERT or RoBERTa for sentiment analysis.

5 Model Deployment and Serving

- **Docker:** For containerizing ML models and ensuring consistency across environments.
- **Kubernetes:** For orchestrating, scaling, and managing containerized applications.
- **Managed ML Platforms:**
  - **AWS SageMaker or Google AI Platform:** For managed model training and hosting inference endpoints.

6 ML Pipeline Orchestration and Monitoring

- **Apache Airflow:** For scheduling and orchestrating data ingestion, processing, and model retraining pipelines.

- **MLFlow:** For tracking experiments, model versioning, and deployment.
- **Monitoring Tools:**
  - **Prometheus and Grafana:** For real-time monitoring and visualization of model performance and system metrics.