# ETL Pipline on HR Attrition Analysis

## Problem Statement

The goal of this project is to predict employee attrition in an organization by building an end-to-end ETL pipeline. Employee attrition can significantly impact a company's operations, resulting in increased recruitment and training costs, loss of organizational knowledge, and decreased productivity. By predicting which employees are likely to leave, HR departments can take proactive measures to retain valuable talent, improving overall employee satisfaction and reducing turnover rates.

### Step 1: Data Extraction

- **Extract HR Data**: Schedule and extract employee data from a PostgreSQL database using Apache Airflow.
- **Extract Survey Data**: Extract survey data stored in AWS S3 buckets.
- **Extract Attendance Records**: Extract attendance data from a CSV file stored in AWS S3.
- **Extract Performance Metrics**: Extract performance data from a Kafka stream.

### Step 2: Data Transformation

- **Data Cleaning**: Use Apache Spark to clean the data (handle missing values, correct data types, etc.).
- **Data Integration**: Merge data from various sources using Spark.
- **Feature Engineering**: Create new features that might be indicative of attrition (e.g., employee tenure, average performance score).

### Step 3: Data Loading

- **Load into Data Warehouse**: Load the cleaned and transformed data into a PostgreSQL or Cassandra database for further analysis.

## Step 4: Data Analysis and Modeling

- **Exploratory Data Analysis (EDA)**: Use SQL and Python (e.g., pandas, matplotlib) to understand the data and identify trends.
- **Model Training**: Use Spark MLlib or another ML framework to train a predictive model for attrition.

## Step 5: Model Deployment

- **Deploy Model**: Deploy the trained model using a REST API (e.g., Flask) that HR systems can query to get attrition predictions.

## Step 6: Monitoring and Maintenance

- **Pipeline Monitoring**: Use Airflow's monitoring capabilities to ensure that the pipeline runs smoothly.
- **Model Monitoring**: Track model performance over time and retrain as necessary.

By following this guide, you will be able to build an end-to-end ETL pipeline for predicting employee attrition, enabling HR departments to make data-driven decisions to retain their workforce.

Previous
Building ML Models

Next
Project: Employee Attrition Prediction Pipeline

Last updated 3 months ago