ETL DAG's

Outline the DAGs (Directed Acyclic Graphs) required for the ETL pipeline to perform attrition prediction. The pipeline will consist of multiple DAGs, each focusing on different stages of the ETL process. Here's an overview of the DAGs and the instructions for each step:

DAG 1: Data Extraction

Purpose: Extract data from different sources (HR database, survey files in S3, and attendance records).

1. **Task 1: Extract HR Data**
   - Extract employee demographic and job data from the HR PostgreSQL database.
   - Schedule: Daily.
   - Steps: Connect to the PostgreSQL database, run SQL queries to fetch the data, and save it to a staging area (e.g., S3 or local file system).
2. **Task 2: Extract Survey Data**
   - Extract employee satisfaction survey data from AWS S3.
   - Schedule: Daily.
   - Steps: Connect to S3, download the survey files, and save them to a staging area.
3. **Task 3: Extract Attendance Records**
   - Extract attendance records from AWS S3.
   - Schedule: Daily.
   - Steps: Connect to S3, download the attendance files, and save them to a staging area.

DAG 2: Data Cleaning and Transformation

Purpose: Clean and transform the extracted data to prepare it for analysis.

1. **Task 1: Clean HR Data**

- Clean the extracted HR data (handle missing values, correct data types, remove duplicates).
- Schedule: Daily, after data extraction.
- Steps: Apply data cleaning procedures and save the cleaned data to a staging area.

2. **Task 2: Clean Survey Data**
   - Clean the extracted survey data (handle missing values, correct data types, remove duplicates).
   - Schedule: Daily, after data extraction.
   - Steps: Apply data cleaning procedures and save the cleaned data to a staging area.
3. **Task 3: Clean Attendance Data**
   - Clean the extracted attendance data (handle missing values, correct data types, remove duplicates).
   - Schedule: Daily, after data extraction.
   - Steps: Apply data cleaning procedures and save the cleaned data to a staging area.
4. **Task 4: Transform Data**
   - Transform and integrate the cleaned data from different sources.
   - Schedule: Daily, after data cleaning.
   - Steps: Join datasets on common keys (e.g., employee ID), create new features (e.g., tenure, average performance score), and save the transformed data.

DAG 3: Data Loading

Purpose: Load the cleaned and transformed data into a data warehouse for further analysis.

1. **Task 1: Load Data into Data Warehouse**
   - Load the transformed data into a PostgreSQL or Cassandra database.
   - Schedule: Daily, after data transformation.
   - Steps: Connect to the data warehouse, upload the data, and ensure the data integrity.

DAG 4: Data Analysis and Model Training

Purpose: Perform exploratory data analysis (EDA) and train the attrition prediction model.

1. **Task 1: Exploratory Data Analysis**
   - Perform EDA on the loaded data to understand patterns and trends.
   - Schedule: Weekly.
   - Steps: Use SQL and Python (e.g., pandas, matplotlib) to visualize data, identify trends, and create summary statistics.
2. **Task 2: Feature Engineering**
   - Create new features from the existing data that might help in prediction.
   - Schedule: Weekly, after EDA.
   - Steps: Implement feature engineering techniques, such as creating binary indicators, aggregating data, etc.
3. **Task 3: Train Predictive Model**
   - Train a machine learning model to predict employee attrition.
   - Schedule: Weekly, after feature engineering.
   - Steps: Split the data into training and testing sets, train the model using Spark MLlib or another ML framework, and evaluate model performance.

DAG 5: Model Deployment and Prediction

Purpose: Deploy the trained model and use it to make predictions on new data.

1. **Task 1: Deploy Model**
   - Deploy the trained model as a REST API using Flask or another web framework.
   - Schedule: As needed, after model training.
   - Steps: Save the trained model, create an API endpoint, and deploy the service.
2. **Task 2: Predict Attrition**
   - Use the deployed model to predict employee attrition.
   - Schedule: Daily.
   -

Steps: Collect new data, preprocess it using the same transformations as the training data, and send it to the model API for predictions.

DAG 6: Monitoring and Maintenance

Purpose: Monitor the pipeline and model performance, and retrain the model periodically.

1. **Task 1: Monitor Pipeline**
   - Monitor the ETL pipeline to ensure smooth operation and handle failures.
   - Schedule: Continuous.
   - Steps: Set up alerts for task failures, check logs, and ensure data is correctly processed.
2. **Task 2: Monitor Model Performance**
   - Track the performance of the deployed model over time.
   - Schedule: Monthly.
   - Steps: Evaluate model predictions against actual outcomes, calculate performance metrics, and identify the need for retraining.
3. **Task 3: Retrain Model**
   - Retrain the model with new data to maintain prediction accuracy.
   - Schedule: Monthly or as needed based on performance.
   - Steps: Collect new data, preprocess it, train a new model, and deploy it.

Instructions for Setting Up Each DAG

1. **Define the DAG Structure**: Define the dependencies and schedule for each task within the DAG.
2. **Create Task Operators**: Use appropriate Airflow operators (e.g., PythonOperator, PostgresOperator, S3ToS3Operator, SparkSubmitOperator) to define each task.
3. **Implement Task Functions**: Write the Python functions or SQL queries that each operator will execute.
4. **Configure Airflow**: Ensure that Airflow connections to your databases, S3 buckets, and other services are properly configured.
5. **Test DAGs**: Test each DAG individually to ensure that it works as expected before running them in production.
6. **Monitor and Debug**: Use the Airflow UI to monitor DAG runs, view logs, and debug any issues that arise.

By following these steps and setting up the DAGs as described, you'll be able to build a robust ETL pipeline for attrition prediction, enabling HR departments to proactively manage employee retention.

Previous
Project: Employee Attrition Prediction Pipeline

Last updated 3 months ago