Program Overview

Saturday, April 6, 2019 2:46 AM

Projects

Project 1 - Data Modeling
In this project, you'll model user activity data for a music streaming app called Sparkify. The project is done in two parts. You'll create a database and import data stored in CSV and JSON files, and model the data. You'll do this first with a relational model in Postgres, then with a NoSQL data model with Apache Cassandra. You'll design the data models to optimize queries for understanding what songs users are listening to. For PostgreSQL, you will also define Fact and Dimension tables and insert data into your new tables. For Apache Cassandra, you will model your data to help the data team at Sparkify answer queries about app usage. You will set up your Apache Cassandra database tables in ways to optimize writes of transactional data on user sessions.

Project 2 - Cloud Data Warehousing
In this project, you'll move to the cloud as you work with larger amounts of data. You are tasked with building an ELT pipeline that extracts Sparkify's data from S3, Amazon's popular storage system. From there, you'll stage the data in Amazon Redshift and transform it into a set of fact and dimensional tables for the Sparkify analytics team to continue finding insights in what songs their users are listening to.

Project 3 - Data Lakes with Apache Spark
In this project, you'll build an ETL pipeline for a data lake. The data resides in S3, in a directory of JSON logs on user activity on the app, as well as a directory with JSON metadata on the songs in the app. You will load data from S3, process the data into analytics tables using Spark, and load them back into S3. You'll deploy this Spark process on a cluster using AWS.

Project 4 - Data Pipelines with Apache Airflow
In this project, you'll continue your work on Sparkify's data infrastructure by creating and automating a set of data pipelines. You'll use the up-and-coming tool Apache Airflow, developed and open-sourced by Airbnb and the Apache Foundation. You'll configure and schedule data pipelines with Airflow, setting dependencies, triggers, and quality checks as you would in a production setting.

Project 5 - Data Engineering Capstone
The capstone project is an opportunity for you to combine what you've learned throughout the program into a more self-driven project. In this project, you'll define the scope of the project and the data you'll be working with. We'll provide guidelines, suggestions, tips, and resources to help you be successful, but your project will be unique to you. You'll gather data from several different data sources; transform, combine, and summarize it; and create a clean database for others to analyze.

Prerequisites

- Python
- SQL
- Command Line / Linux
- Git Hub

All have supplemental courses if needed (git skills needed some work)

Program Structure

- One project per month give or take (Later projects are generally more open ended so you can spend as much time as you want)
- Weekly discussion sections
- Focus on the community on Slack

Profile / Career Services

- Resume, Linked In and Git hub reviews included
- Afterwards those are amiable at a cost (essentially linked in premium)
- Not a big thing for me, given that I'm not activley looking for a new role but a good way to get feedback.

Keys to Success

- Be an active Lerner (engage with the content, take notes)
- Write your own code, practice a lot. Don't skip curriculum.
- Go to the classroom at least 2x a week.
- Utilize your mentor
- Get better at googling / solving problems