Arpita Deshmukh

Data Engineer

arpitapdtamu@gmail.com https://arpitapd.github.io/ linkedin.com/in/arpitapd (Visa: H1B)

Summary

Data Engineer with 5+ years of experience designing and maintaining scalable ETL/ELT pipelines for healthcare analytics. Strong background in optimizing data pipelines and building analytics-ready datasets using Python, SQL, and Alteryx, including API-based data ingestion, complex query development, and dimensional modeling for reporting. Experienced in deploying data pipelines in Azure cloud environments, with additional exposure to AWS and GCP. Partner closely with business teams to translate requirements into data-backed analytics that enable improved decision-making

Skills

Programming: SQL (CTEs, window functions), Python (ETL pipelines, Data transformation, API Ingestion), PySpark
Cloud: Azure, BigQuery, GCP, Airflow (pipeline orchestration), AWS (S3, Glue, Athena, Redshift)
Data skills: Building ETL pipelines, Dimensional Modeling, Data validation
Other Tools: Alteryx (ETL), SQL Server (Relational Database), PowerBI

Work Experience

BI Data Engineer **Jan 2022 – Present**
Behavioral Health Group *Tools Used:Python, SQL, Azure, APIs, Alteryx*

- Designed ETL pipeline to Integrate 100+ tables **from multiple databases** into unified dataset, **partitioning data by year** with 25 million+ rows per partition to improve query performance
- Engineered **Python + Alteryx ETL pipeline** to ingest healthcare provider data via API, using automated **stored procedures for incremental merges and updates**; eliminating 10+ hours/week of manual reporting
- Analyzed 100 million+ records using **PySpark and SQL (CTEs, window functions, joins)** to track missed doses and follow-up calls; reducing patient no-shows by 10% in 3 months, **estimated $300K in recovered revenue**
- Evaluated **15 million+ rows** using SQL to identify missed call patterns, enabling **improved staffing decisions,** reducing missed calls from 22% to 14%
- Implemented **Row Level Security (RLS)** to protect sensitive data, enabling automated data access based on job title
- Deployed **Python ETL pipeline** including API data extraction, transformations, backing up and uploading report to HR system via SMTP using Alteryx and batch scripting, **saving $10K in labor costs**
- Automated **Python pipeline** to extract healthcare providers data, built stored procedures to perform into Azure, **saving 10 hours** of manual reporting per week

Revenue Data Analyst **Jul 2020 – Jan 2022**
Behavioral Health Group *Tools Used: SQL, Python, Azure, Alteryx, Power BI*

- Developed and maintained end-to-end automated data flows, delivering KPIs used by **5K+ business users** daily
- Optimized Alteryx ETL flows using In-DB processing, **reducing runtime by 80%** improving operational efficiency

Business Intelligence Analyst Intern **May 2019 – Aug 2019**
Strategic Materials, Inc *Tools Used: Data Modeling and Visualization in Targit*

- Partnered with C-suites to define KPIs, built dashboards to track operations-focused metrics for 50 plant locations

Projects

ELT Pipeline using GCP (Mini Project) **GCP, BigQuery, Airflow DAGs, Looker**

- Built an end-to-end cloud ELT pipeline on GCP using BigQuery and automated Airflow DAGs (workflow orchestration) to process 1 million+ records

YouTube Data Analysis in AWS **AWS S3, Glue, Athena, Lambda, IAM, Quicksight, ETL, SQL**

- Engineered a **distributed, serverless ETL** pipeline processing 2000+ daily trending videos across 10 regions
- Built efficient datasets for analysis, enabling insights on **region-specific content strategies** (60% of videos trending in Japan were Animation, compared to 15% in US)

Education

Masters in Management Information Systems **Aug 2018 – May 2020**
Texas A&M University *College Station, Texas*