

# ARPITA DESHMUKH

## Data Engineer

[✉️](mailto:arpitapdtamu@gmail.com) [🔗](https://arpitapd.github.io/) [LinkedIn](https://linkedin.com/in/arpitapd) (Visa: H1B) Dallas, TX

### Summary

Data Engineer with 5+ years of experience handling End-to-End healthcare data projects, designing and maintaining scalable ETL/ELT pipelines. Strong background in optimizing data pipelines using Python, SQL, and Alteryx, including API-based data ingestion, complex query development, and dimensional modeling for BI 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, Stored Procedures), Python (ETL pipelines, Data transformation, API Ingestion), PySpark

**Cloud:** Azure, BigQuery, GCP, Airflow (pipeline orchestration), AWS (S3, Glue, Athena, Redshift)

**Other Tools:** Alteryx (ETL), SQL Server (Relational Database), PowerBI, Looker, Git

**Data skills:** Building ETL pipelines, Dimensional Modeling, Data validation

### Work Experience

**BI Data Engineer** Jan 2022 – Present

*Behavioral Health Group*

*Tools Used: Python, SQL, Azure, APIs, Alteryx*

- Designed ETL pipeline to Integrate **100+ source tables** into unified dataset, **partitioning data by year** with 25 million+ rows per partition to improve query performance
- Engineered **Python + Alteryx ETL pipeline** to ingest 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 support operational analytics; reducing patient no-shows by 10% in 3 months, **estimated \$300K in recovered revenue**
- Evaluated **15 million+ rows** using SQL and PowerBI 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

**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 10 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*