

Milan Barot

7milan16@gmail.com ❖ (917) 497-6979 ❖ DMV, VA ❖ [linkedin.com/in/7mb16](https://www.linkedin.com/in/7mb16) ❖ <https://70milan.github.io/>

SUMMARY

Data Engineer with 6+ years of experience specializing in automated ETL pipelines and healthcare interoperability. I have a strong track record of migrating legacy systems to the cloud and delivering high-quality data products in Agile environments. Currently, I am focused on the intersection of traditional data engineering and LLM-driven development to build more intelligent, low-latency applications.

TECHNICAL SKILLS

Business Intelligence Tools	Power BI, SSRS, SSIS
Programming Languages	SQL, Python (pandas, pyspark, matplotlib, etc.), JavaScript, CSS, HTML
Databases/Data Warehouse	MS SQL, PostgreSQL
Big Data Tech	Apache Spark
Web Frameworks & APIs	FastAPI, React, Electron, RESTful APIs, OAuth
Cloud Computing Frameworks	Azure Cloud services (Data Factory), Amazon Web Services (AWS), Databricks
Workflow Tools & Integration Services	GIT, Jenkins, Dagster, Uvicorn, Airflow, Jira, ActiveBatch, Confluence, MS Office Suite
AI & Large Language Models	OpenAI API, GPT3.5t, 4+, 5 (Mini & Nano), GPT-4o, GPT-3.5, Whisper (Speech-to-Text)

TECHNICAL PROJECTS

RealTime_Context_Engine (AI-Powered Windows Runtime Host) https://github.com/70milan/job_bandit

- Developed a **dual-process desktop application** utilizing an **Electron** frontend and a **Python/FastAPI** backend to provide **real-time** interview guidance.
- Integrated **OpenAI GPT models** via **low-latency streaming (SSE)** to deliver generative feedback with a **<200ms TTFT** (Time To First Token).
- Built a custom **hardware-locked licensing system** using **RSA digital signatures** and **machine-specific HWIDs** to secure distribution.
- Engineered a high-performance **screen capture and audio processing pipeline** for real-time context analysis during live sessions.
- Automated a full **distribution pipeline** using **PyInstaller** and **Electron Builder**, incorporating **code signing** and an **automated update system**.

Tech Stack: Python (FastAPI/Uvicorn), JavaScript (Electron/Node.js), OpenAI API (GPT-4o/5), RSA Cryptography, SSE (Server-Sent Events), PyInstaller, Electron Builder.

YouTube to Spotify Automated ETL Pipeline https://github.com/70milan/YouTube_Spotify_Synchronization_Dagster

- Engineered an automated ETL pipeline using **Dagster** to seamlessly extract, transform, and synchronize music playlists between the **YouTube Data API** and **Spotify Web API**.
- Architected a Change Data Capture (**CDC**) workflow, utilizing **AWS S3** to store intermediary staging files and final datasets, ensuring efficient delta processing and minimizing redundant API calls.
- Managed secure user sessions through complex **OAuth 2.0** authentication flows.
- Developed transformation logic using python and pandas to parse video metadata, query the Spotify database, and incrementally load corresponding tracks into automated playlists.

Tech Stack: Python, Pandas, Dagster, AWS S3, RESTful APIs, OAuth 2.0

Spotify User Activity & Trends Pipeline (Databricks & Delta Lake) https://github.com/70milan/sp_pyspark_etl

- Architected an end-to-end Medallion Architecture** (Bronze, Silver, Gold) on **Databricks** to ingest and analyze multi-year Spotify user data trends.
- Engineered a local-to-cloud ingestion bridge** using Python scripts to automate the secure upload of user datasets to **AWS S3**, subsequently landing raw data into **Unity Catalog Volumes**.
- Developed complex PySpark and Spark SQL transformations** in the Silver layer to flatten nested JSON structures and normalize data into relational schemas (Artist, Metadata, Genre, etc.).

- **Implemented Change Data Capture (CDC)** logic using **Delta Lake** features to ensure incremental loading, significantly reducing compute costs by processing only updated or new records.
- **Orchestrated Gold-layer aggregations** to derive business value, such as listening trend evolution and artist affinity scoring, stored in optimized Delta tables for downstream reporting.

***Tech Stack:** Databricks, PySpark, PySpark SQL, Delta Lake, Unity Catalog, AWS S3, Python.*

WORK EXPERIENCE

AmeriBen (UST Global | Implementation partner), remote (Health insurance)

SQL Server Developer/ Data Engineer January 2026 – Current

- **Translated complex healthcare business requirements into robust SQL logic**, updating core database objects to manage the extraction of claims, accumulators, and provider data.
- **Streamlined enterprise data workflows** by identifying and resolving bottlenecks in existing SQL Server extraction routines, improving the reliability of daily payment batches.

***Tech Stack:** SQL Server 2022, Azure SQL*

Blue Cross Blue Shield Arizona (Mastek | Implementation partner), Phoenix, AZ (Health insurance)

Data Engineer Sept 2022 – Oct 2025

- Designed and implemented end-to-end **ETL pipelines** using **SSIS** and **SQL Server** to process eligibility, claims, accumulator, and value-based care (VBC) files, ensuring compliance with business requirements and vendor-specific standards. Collaborated with vendors such as OptumRx, HealthEquity, Equality Health, and others to deliver high-quality data solutions.
- Oversaw a small team of offshore resources by recruiting and onboarding team members as required for project execution. Ensured tasks were completed on time by maintaining close collaboration with both onshore and offshore teams
- Utilized **Azure Data Factory (ADF)** to design and orchestrate scalable **data integration pipelines** for migrating legacy member and claims data to **HealthRules Payer (HRP)** on **Azure Cloud**.
- Leveraged **Azure SQL Database** and **Azure Data Lake Storage** for staging, transformation, and persistent storage of high-volume healthcare data. Configured **ADF linked services, pipelines, and triggers** for automated, scheduled data ingestion from on-premises SQL Server to Azure.
- **Rebuilt critical ETL pipelines** using **SSIS** and **SQL Server** after the **CHC system hack (2024)**, restoring high-priority data workflows under tight deadlines with minimal downtime.
- Developed **complex data extracts** that combined on-prem warehouse data with select **cloud-based sources**, ensuring consistency, accuracy, and seamless integration across hybrid environments reducing execution time by 25 % on an average across the board.
- Collaborated closely with **Deloitte** on the HealthEquity project to develop **eligibility and claims files** and automated it end to end. Responsibilities included gathering business requirements, creating technical documentation, optimizing ETL pipelines, conducting performance testing, and ensuring smooth deployment into production environments.
- Automated data workflows by integrating **SQL Server Agent jobs** with **ActiveBatch**, setting up FTP connectivity with strict naming conventions and cadence-driven schedules for file delivery to vendors.
- Developed a consolidated **EDI framework for 834 and 837 enrollments, claims and dental**, using three core SSIS packages to automate the generation and delivery of over **35 distinct file types** to **CMS**. This architecture ensured secure data exchange and audit readiness while significantly reducing the manual maintenance and operational overhead previously required.
- Provided active production support for mission-critical **CMS data exchanges**, conducting over 25 successful production deployments and managing emergency releases to maintain 100% compliance with federal reporting deadlines.
- Led the rapid resolution of critical production failures, ensuring zero data loss during **high-volume enrollment** periods where direct connectivity with CMS was required
- Partnered with business stakeholders to gather requirements for new projects, conducting detailed assessments of complexity, technical effort, development timelines, and warranty periods. Worked with Mastek to draft **SOWs** for approval by Blue Cross Blue Shield, ensuring alignment with business objectives.
- Participated in Agile sprint meetings to track progress on deliverables, resolve blockers, and ensure timely completion of project milestones across multiple initiatives.

***Tech Stack:** Azure Data Factory (ADF), Azure SQL Database, Azure Data Lake Storage, SQL Server, SSIS, SSRS, ActiveBatch, SQL Server Agent, FTP, EDI (834, 837), HealthRules Payer (HRP), Agile +4*

CBRE, Dallas, TX (Commercial real estate services and investments)

Data Engineer Feb 2021 - Sept 2022

- Played a key role on the **Data Engineering & Reporting team** during the COVID-19 pandemic, delivering real-time data pipelines and analytics that supported CBRE's **Advisory business unit** across global commercial real estate portfolios.
- Designed, developed, and maintained robust **ETL pipelines** using **Python, SQL, and PostgreSQL** via CBRE's internal ETL framework (**Lymbic**). Translated complex business logic from real estate analysts into SQL/Python transformations for property usage, tenant risk, and leasing performance.
- Ingested diverse external data sources such as **Google Mobility Reports** and **BREAM energy efficiency assessments**, transforming them into structured datasets in **Parquet** and **CSV** formats. These outputs fed into **PowerBI dashboards** used by leadership to assess region-wise COVID impacts on commercial spaces.
- Applied data profiling and transformation libraries such as **Pandas, PyArrow** for efficient in-memory processing of large CSV/Parquet files, enabling faster QA cycles and stakeholder feedback loops.
- Supported **Airflow DAG** orchestration workflows — monitoring job execution, rerunning failed tasks, and assisting engineers with job recovery — ensuring daily and weekly batch pipelines remained consistent and reliable during volatile COVID data conditions.
- Worked in an **Agile delivery environment**, actively collaborating with analysts, product managers, and backend engineers in weekly sprint meetings to prioritize features, track bugs, and refine pipeline outputs based on stakeholder needs.
- Enabled high-impact real estate decisions by producing clean, trusted data for dashboards that revealed critical trends such as **building underutilization, shifting tenant behaviors, and geographic leasing risk**. These insights directly influenced CBRE's portfolio restructuring strategies and client advisory recommendations during the pandemic.

Tech Stack: Python, SQL, PostgreSQL, Apache Airflow, Power BI, Pandas, PyArrow, Parquet, Agile

AmNet Services Inc., Fairfield, NJ (Telecommunications)

ETL/SQL Server Developer Oct 2019 - Aug 2020

- Developed SSIS packages to automate extraction and ingestion of network test data from Excel, flat files, and MS Access sources into a SQL Server staging environment.
- Designed and implemented a dimensional data warehouse model using **ERwin Data Modeler** to support analytical queries on network KPIs.
- Wrote complex **T-SQL stored procedures** for **data cleansing, aggregation, and validation** to ensure accuracy and consistency of telecom metrics.
- Created and maintained **SSRS reports** and **PowerBI dashboards** to visualize signal strength, call drop rates, and E911 test call success metrics.
- Automated **SQL Server Agent** jobs to schedule SSIS package executions and report generation, ensuring daily data processing without manual intervention.
- Collaborated with radio engineers, business analysts, and compliance teams in Agile sprints to gather requirements and iteratively improve data solutions.
- Conducted rigorous testing of ETL workflows and reports to guarantee data quality and system performance at scale.

Tech Stack: Python, SQL, PostgreSQL, Apache Airflow, Power BI, Pandas, PyArrow, Parquet, Agile

EDUCATION

Bachelor of Science in Information Technology in Software Engineering March 2022

Washington University of Science & Technology, Vienna, VA

Bachelors in Applied Science in Computer Information Systems Dec 2019

The City University of New York, New York, New York

CERTIFICATIONS

Azure Data Fundamentals (DP-900) May 2024