Pushkar Patil

+1 (857) 313-0289 | pushkar.patil1269@gmail.com | Linkedin | Portfolio

Summary

Data engineer with 3+ years of experience building high-throughput pipelines and analytics solutions across AWS, Azure and GCP. Developed geospatial search and ETL workflows using PySpark, Airflow, and Delta Lake. Created real-time BigQuery data marts for environmental insights and Snowflake models for finance, marketing and real estate intelligence.

Education

Northeastern University **Sep 2022 – Dec 2024**
Master of Science in Information Systems *Boston, MA*

Mumbai University **Jun 2017 – Oct 2020**
Bachelor of Science in Computer Science *Mumbai, IN*

Experience

Students Landing **Feb 2025 – Present**
Data Engineer *Boston, US*

- Pioneered geospatial property search for an international student housing aggregator, integrating **Elasticsearch** with **OpenStreetMap** to enable location-based filtering and interactive mapping, increasing property discoverability by 25%
- Orchestrated ETL pipeline for nested JSON from **DynamoDB** via **PySpark, AWS Glue, Delta Lake** in Medallion architecture (Bronze to Silver), flattening, enforcing business logic, and expediting workflows, cutting processing time by 20%
- Enabled strategic real estate investments by leveraging comprehensive market insights, property data, and predictive analytics

Experiential AI **Jan 2024 – Aug 2024**
Data Analyst/Engineer *Boston, US*

- Established hyper local air quality pipeline integrating sensor feeds, leveraging high-throughput streaming with **Flink Clusters(Data Proc, Cloud Dataflow)** and **GCP Pub/Sub**, **BigQuery** transformations to deliver **1M+** records per minute
- Optimized data delivery analytics by structuring scalable **BigQuery** data marts for research teams, deploying **Spark SQL** for high-throughput batch processing to accelerate insights for **10M+** users
- Delivered air quality insights to government stakeholders via real-time **Looker** dashboards and historical **KPI** visualizations, securing executive buy-in and saving **$10K** per month

Asfaliea **Dec 2020 – Aug 2022**
Data Engineer I *Mumbai, IN*

- Constructed **CDC** pipelines with **Airflow DAGS** and **AWS S3** to ingest server logs, application events, and page-visits data, delivering sales funnel insights that informed e-commerce strategies, reducing manual intervention by **25%**
- Architected scalable, event-driven **ETL** pipeline using **AWS (Glue, SQS, SNS, Lambda)** with **Spark**, integrating with **Snowflake(SnowPipe, Streams)** to process guest data and optimize **marketing campaigns**, boosting ingestion by **15%**
- Led **AWS migration**, implementing unit testing with **Pytest** while collaborating with cross-functional teams to fine-tune data solutions

SpanishBOLO **Jan 2020 – Sept 2020**
SQL Developer *Hyderabad, IN*

- Architected database infrastructure by redefining Schema, Partitions, UDLs and triggers to improve query performance by **10%**
- Engineered **SQL** queries with window functions and indexing and deployed **Pulumi** for analysis, cutting costs by **$2k** per month
- Revitalized cloud operations with **AWS Redshift, S3,** and **Athena** for ETL-free analytics, enforced **Cognito** for secure access, implemented data governance, leveraging **Tableau** for sales projections, reducing overhead by **12%**

Projects

LogiFlow: Logistics Data Pipeline | *GCP, Pulumi, FastAPI, Apache Spark, Docker, Airflow*

- Engineered logistics analytics pipeline on GCP using Python, Dataflow, and BigQuery, increasing throughput by **40%**
- Automated workflows for supply chain data with Mage.ai, Cloud Scheduler, and Dataflow, improving analytics accuracy by **35%**

Fraud Detection System | *Spark, Databricks, Hadoop, Pulumi, Node.js, MongoDB, LLM*

- Programmed an application by conceptualizing a Hadoop, Spark, REST API architecture to enable large-scale anomaly analysis
- Formulated an ML-driven workflow with Node.js, and MongoDB, reducing detection time by **80%**

PodcastGPT: RAG-Based AI for Intelligent Podcast Search | *FAISS, OpenAI API, LangChain, FastAPI*

- Innovated PodcastGPT by integrating FAISS, OpenAI GPT, and LangChain, and incorporated GitHub Actions for CI/CD to accelerate query retrieval speed by **15%**
- Orchestrated a scalable API using FastAPI, Docker, AWS Lambda, and Step Functions to optimize podcast search efficiency

Technical Skills

Languages: Python, Java, Scala
ML/Libraries: NumPy, Pandas, Scikit-learn, TensorFlow, Matplotlib, Seaborn
Databases & Cloud: PostgreSQL, MySQL, Oracle DB, MongoDB, Firebase, Cassandra; AWS (Glue, Lambda, S3, Kinesis, Cognito, Elasticsearch, DynamoDB, RDS, Athena), Apache (Airflow, Kafka, Spark, Hadoop, NiFi, Flink), GCP (BigQuery, Pub/Sub, CloudComposer, Dataflow), Snowflake, Oracle Cloud, Docker, Kubernetes, Terraform, Git, Databricks
Generative AI: LangChain, FAISS, OpenAI API, RAG