

Amita Sahoo

+1 669-302-0023 | amita.sahoo173@gmail.com | linkedin.com/in/amitasahoo | San Jose, CA

SUMMARY

Senior Data Analyst | Analytics Engineer | GenAI-Driven Insights

Results-driven Senior Data Analyst and Analytics Engineer with 7+ years of experience in data analysis, statistical analysis, data visualization, and self-service analytics across technology, financial services, and marketing domains. Expertise in SQL, Python, Tableau, Power BI, Snowflake, dbt, and Databricks delivering \$1.5M+ in measurable cost savings across 45M+ regulated records. Proven track record in stakeholder management, executive reporting, A/B testing, cohort analysis, and dashboard development. Integrates GenAI workflows including RAG, LangChain, and OpenAI APIs to drive AI-powered business intelligence in Agile, cloud-native environments.

SKILLS

Analytics & Programming: Python (Pandas, NumPy, Scikit-learn, SciPy), SQL (Joins, CTEs, Window Functions), R, Scala, Statistical Analysis, Exploratory Data Analysis (EDA), A/B Testing, Hypothesis Testing, Cohort Analysis, Funnel Analysis, Forecasting & Trend Analysis, Customer Segmentation, CLV Modeling, Feature Engineering, Jupyter Notebooks

Data Engineering & Pipelines: Apache Spark (PySpark), Apache Airflow, Apache Kafka, Databricks, dbt (Core & Cloud), Delta Lake, Apache Iceberg, ETL/ELT Design, Batch & Streaming Processing, Change Data Capture (CDC), Azure Data Factory (ADF), Terraform (IaC), Data Cleaning & Preprocessing

Databases & Data Warehousing: Snowflake, Google BigQuery, Amazon Redshift, PostgreSQL, Microsoft SQL Server, Azure Synapse Analytics, Cosmos DB, Dimensional Data Modeling, Star Schema, Slowly Changing Dimensions (SCD), Data Contracts, Data Quality & Validation

BI & Data Visualization: Tableau, Power BI, Looker, Looker Studio, Microsoft Excel (Advanced, Pivot Tables), KPI Dashboard Development, Executive Reporting, Data Storytelling, ROI Analysis

Applied AI & GenAI: Retrieval-Augmented Generation (RAG), LangChain, LlamaIndex, LLM Automation, Agentic Workflows, Prompt Engineering, OpenAI APIs, Azure OpenAI, Microsoft Copilot for Data, Vector Databases (Pinecone, ChromaDB)

Cloud Platforms & DevOps: Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP), AWS Glue, CI/CD, Docker, FastAPI, AWS EKS, Git, GitHub

Data Governance & Compliance: Data Observability, Great Expectations, Monte Carlo, Data Lineage, Metadata Management, PII Handling, HIPAA Compliance, Data Mesh, Data Contracts

Soft Skills & Methodologies: Stakeholder Management, Executive Communication, Cross-functional Leadership, Requirement Gathering, Presentation Skills, Agile (Scrum), Waterfall, SDLC, Jira, Confluence

EXPERIENCE

Uber May 2025 – Present
Data Analyst / Data Engineer San Francisco, CA

- Performed geospatial cohort analysis and exploratory data analysis (EDA) using Python (H3) and AWS S3 to identify high-churn driver zones, designing AI-assisted incentive structures that contributed to reactivating **6,500 dormant drivers** as part of a **\$680,000** West Coast reactivation initiative.
- Architected automated executive reporting pipelines in SQL, Airflow, and Snowflake with dbt transformations, eliminating **5 days** of manual reporting overhead and cutting delivery to **2 hours** per monthly review cycle.
- Executed A/B testing and hypothesis testing in R to evaluate a Generative AI chatbot, refining LLM sentiment models that diverted **45,000 support tickets** and saved **\$180,000** in annual support costs.
- Built real-time marketplace KPI dashboards in Tableau and Michelangelo, surfacing supply-demand imbalances and enabling data-driven decisions that shielded an estimated **\$315,000** in gross bookings.
- Quantified driver reactivation ROI through cohort analysis and CLV modeling in Python and Snowflake, securing **\$800,000** in incremental incentive budget through executive stakeholder presentations.
- Deployed an LLM-powered agentic workflow via OpenAI APIs and LangChain to classify **12,000+ monthly support requests**, compressing triage time to **15 minutes** and unlocking 2 FTEs for strategic analysis.
- Presented monthly marketplace health metrics and driver supply forecasts to the **VP of Marketplace Analytics**, directly informing Q3 expansion strategy across 12 metro markets.
- Mentored and coached 6 junior analysts on data modeling and experimentation best practices, accelerating onboarding by **3 weeks per analyst** across 40 global growth campaigns.

Epsilon Sep 2024 – Apr 2025
Data Analyst Chicago, IL

- Engineered ETL/ELT pipelines using PySpark and AWS Glue, reducing daily data processing latency by **4 hours** and enabling same-day campaign reporting across 3 enterprise clients.
- Developed executive Tableau dashboards using advanced SQL (window functions, CTEs) and funnel analysis, identifying **14 underperforming segments** and surfacing **\$215,000** in monthly optimization opportunities across a **\$3M+ annual campaign budget**.
- Implemented Great Expectations for automated data quality monitoring within the Snowflake and dbt pipeline, enforcing data validation checks across 6 upstream sources and reducing pipeline failures by **28 per month**.
- Applied machine learning-based customer segmentation (K-Means clustering) in Python and Databricks to profile **8M+ consumer records**, reducing cost per acquisition by **\$18 per lead**.

- Validated and cleansed **28M+ consumer records** in Snowflake and dbt, resolving 85 critical data quality discrepancies with zero defects at go-live.
- Delivered weekly campaign performance reports and data-driven recommendations to **C-suite stakeholders** across 3 Fortune 500 accounts, improving client retention and satisfaction.
- Automated identity resolution workflows via Databricks and Python, eliminating **12 weekly manual reporting hours** and redirecting analyst capacity toward strategic analysis.

Tata Consultancy Services (TCS)

Jan 2021 – Aug 2022

Senior Data Analyst

India

- Designed and optimized ETL pipelines using PySpark and Azure Data Factory across 14 reporting suites, saving **60+ hours** of monthly rework with zero post-deployment failures.
- Built a centralized SQL Server data warehouse with star schema and SCD Type-2 architecture for dimensional data modeling, trimming query execution time by **45 minutes per session** for 5 cross-functional teams.
- Delivered 12 Power BI KPI dashboards surfacing **400 underperforming logistics hubs** and informing **\$500,000** in operational restructuring decisions.
- Developed customer segmentation models using K-Means clustering and RFM analysis in Python and Scikit-learn, recovering **\$1.1M** in at-risk annual contract value across 3 client accounts.
- Built predictive churn models using regression and time-series analysis in R and Scikit-learn, retaining **8,500 SMB enterprise clients** worth an estimated **\$1.8M** in annual recurring revenue.
- Migrated **2,000,000+ records** to Microsoft Azure cloud using optimized Python delta load scripts, achieving zero downtime during production cutover.

KPMG

Mar 2017 – Sep 2020

Data Analyst, Financial Services & Tax Compliance

India

- Automated data extraction and transformation via Alteryx and SQL ETL workflows, eliminating **140 hours** of manual monthly reporting and accelerating audit delivery by **18 business days**.
- Designed and published executive Tableau dashboards for global tax compliance tracking, uncovering **\$1.2M** in under-reported assets across 4 business units.
- Applied statistical analysis including time-series regression and cohort churn modeling in R, delivering retention forecasts that secured **14 high-value accounts** during the fiscal year.
- Built rule-based anomaly detection scripts in SQL and Excel VBA to flag irregular journal entries, surfacing **\$320,000** in reconciliation discrepancies before annual audit submission.
- Established data governance, data lineage, and PII handling frameworks in Microsoft Azure, reducing data entry errors by **450 incidents annually** across 4 departments.
- Presented quarterly audit findings and compliance risk summaries to **senior partners and client CFOs**, driving remediation decisions across 4 regulated business units.
- Led an Agile Data Excellence initiative mentoring 5 junior analysts on SQL, reporting best practices, and data quality, compressing annual audit delivery by **18 business days**.

PROJECTS

AirBnB CDC Ingestion Pipeline | Azure Data Factory, Cosmos DB, ADLS, Azure Synapse, n8n, OpenAI, LangChain

- Orchestrated a near-real-time CDC pipeline ingesting 5+ source entities at hourly refresh cycles with SCD Type-1 fact and 4+ dimension tables in Azure Synapse, integrating LLM automation via n8n and OpenAI to auto-generate leadership-ready reports and save **8 hours per week**.

E-Commerce GenAI Product Assistant | LangChain, LlamaIndex, RAG, Pinecone, FastAPI, Docker, AWS EKS, CI/CD

- Architected a production-grade GenAI assistant using Retrieval-Augmented Generation (RAG), LangChain, and Pinecone for multi-turn product queries at **sub-2-second latency**, deployed via FastAPI, Docker, and AWS EKS with full CI/CD pipeline, cutting analyst resolution time by **4 hours per week**.

BookMyShow Real-Time Streaming Pipeline | Azure Event Hubs, Azure Stream Analytics, Synapse Analytics, JavaScript UDFs

- Engineered a real-time booking and payment streaming pipeline using Azure Event Hubs, Stream Analytics, and JavaScript UDFs, generating CXO-ready insight emails from live Synapse Analytics metrics within **60-second latency windows**.

EDUCATION

Florida Institute of Technology

Aug 2022 – May 2024

Master of Science (M.S.) in Computer Engineering

Melbourne, FL

- **Relevant Coursework:** Machine Learning, Data Mining, Cloud Computing, Statistical Analysis, Big Data Systems, AI & Deep Learning