AAYUSH YADAV

SENIOR DATA ENGINEER

San Jose, CA | (859) 393-2478 | Aayush.Yadav1240@gmail.com | linkedin.com/in/aayushy777

PROFESSIONAL SUMMARY

Senior Data Engineer with 7+ years building and running large-scale data platforms in banking, telecom, and healthcare. Currently lead a team of 4 at Bank of America, where our AWS pipelines process 80 million transaction records a day for risk and regulatory reporting. Deep hands-on experience with Python, SQL, Spark, Kafka, dbt, and Airflow across AWS, Azure, and Databricks. Track record of cutting cloud spend by more than $500K across three companies, passing HIPAA and SOX audits with zero findings, and building platforms that let analysts pull their own data instead of filing tickets.

TECHNICAL SKILLS

Languages	Python, SQL (T-SQL, PL/SQL, Spark SQL), PySpark, Scala, Java
Data Engineering	Apache Spark, Apache Kafka, Apache Airflow, dbt, Snowflake, Databricks, Delta Lake, Apache Iceberg
AWS	Glue, EMR, Redshift, S3, Lambda, Step Functions, Kinesis, Athena, Lake Formation, DynamoDB, CloudWatch, IAM, CloudFormation
Azure	Data Factory, Databricks, Synapse Analytics, ADLS Gen2, Azure Monitor
Databases	PostgreSQL, Oracle, SQL Server, MySQL, MongoDB
DevOps & Tooling	Terraform, Docker, Kubernetes, Jenkins, AWS CodePipeline, Git, CI/CD
BI & Governance	Power BI, Tableau, Amazon QuickSight, Collibra, dimensional and lakehouse modeling, HIPAA, SOX, BCBS 239

PROFESSIONAL EXPERIENCE

Senior Data Engineer

Mar 2024 – Present

Bank of America | Charlotte, NC (Remote)

- Lead a team of 4 engineers who own the AWS pipelines behind risk, compliance, and regulatory reporting, moving about 80 million transaction records a day.
- Rebuilt the ingestion layer on Glue, Lambda, and Step Functions to onboard 25+ internal and vendor feeds. Reporting data that used to land next day now lands within minutes, and about 20 hours a week of manual file handling went away.
- Designed the Redshift and PostgreSQL models behind risk analytics. Queries on the 8 TB warehouse come back in under a second, which kept BCBS 239 regulatory reporting on schedule through two exam cycles.
- Rolled out dbt with 200+ automated tests across 4 lines of business. Data quality scores went from 94% to 99.6%, and roughly 100 analysts now build their own reports instead of waiting on the data team.
- Tuned the heaviest Spark jobs on EMR with better partitioning, broadcast joins, and adaptive query execution. Batch runtimes dropped by half and compute spend fell about $220K a year.
- Moved all 35 production pipelines to Terraform with Jenkins and CodePipeline CI/CD, including automated tests and rollback. Deployments that used to take 2 days now ship in an afternoon.
- Built CloudWatch monitoring with anomaly detection. Pipeline uptime holds at 99.8% and failures get caught in 15 minutes instead of 2 hours.
- Mentor 3 junior engineers on Spark, data modeling, and AWS. Both of last year's new hires were shipping to production within their first month.

Technologies: *Python, SQL, Apache Spark, dbt, AWS Glue, EMR, Redshift, S3, Lambda, Step Functions, Kinesis, Athena, Lake Formation, PostgreSQL, Terraform, Jenkins, Git*

Data Engineer **Jun 2021 – Feb 2024**

T-Mobile | Overland Park, KS (Remote)

- Built streaming pipelines on Kafka and Spark Structured Streaming in Databricks that processed network telemetry and customer usage events, peaking around 900 million events a day, with curated Delta Lake tables ready within minutes.
- Owned the billing and customer activity ELT into Snowflake using dbt, around 150 tested models. Marketing and retention teams went from weekly extracts to same-day data.
- Partnered with data science on churn feature datasets covering usage, network quality, and billing history. The retention campaigns built on them were credited with protecting about $3M in annual revenue.
- Migrated 20+ legacy Informatica jobs to Databricks and Delta Lake on S3, cutting job failures by two thirds and saving about $180K a year in licensing and infrastructure.
- Introduced a schema registry and data contracts for upstream event producers, which ended a recurring class of pipeline breakages caused by unannounced payload changes.
- Carried on-call for the ingestion platform and pushed uptime from 97% to 99.5% through retry logic, dead letter queues, and alerting that pages on real symptoms instead of noise.

Technologies: *Python, Apache Kafka, Spark Structured Streaming, Databricks, Delta Lake, Snowflake, dbt, Airflow, AWS S3, Terraform, Git*

Data Engineer **Apr 2019 – May 2021**

Cardinal Health | Dublin, OH

- Built pipelines in Azure Data Factory and Databricks integrating 20+ sources across pharmacy orders, distribution, and claims data, cutting processing latency by 60% for daily operational reporting.
- Wrote and tuned complex SQL (T-SQL, Spark SQL, PL/SQL) over multi-terabyte datasets. Indexing and partition pruning work cut the nightly reporting window from 6 hours to under 3.
- Automated ingestion into Synapse that replaced roughly 25 hours a week of manual ETL and moved data freshness from T+2 to same day for 8 business units.
- Developed Python automation and internal APIs for reporting workflows, retiring 3 manual reconciliation processes for the finance team.
- Implemented retention and governance policies in Collibra for HIPAA and SOX requirements. Every audit across two years closed with zero findings.
- Worked with BI and supply chain analysts to deliver curated, analytics-ready datasets, taking ad hoc data requests from a 5-day queue to same-day turnaround.

Technologies: *Python, SQL Server, Oracle, Azure Data Factory, Azure Databricks, Synapse Analytics, ADLS Gen2, Snowflake, Collibra, Power BI*

KEY PROJECTS

Real-Time CDC Lakehouse | *Debezium, Kafka, Spark, Apache Iceberg, AWS*

- Built an end-to-end change data capture platform that streams PostgreSQL and MySQL changes through Debezium and Kafka into Iceberg tables on S3, with exactly-once writes, automatic schema evolution, and time travel queries.
- Data is query-ready in Athena within 60 seconds of the source transaction. Fully containerized, deployed with Terraform, monitored with Grafana.

AI Data Catalog Assistant | *Claude API, vector search, dbt, Python*

- Built a natural language interface over a dbt project and its lineage graph. Analysts ask questions like "where does customer_ltv come from and is it fresh today" and get back the model chain, test status, and freshness in plain English.
- Uses embeddings over model documentation and metadata with retrieval augmented generation. Cut repetitive lineage and definition questions to the data team roughly in half.

CERTIFICATIONS

- AWS Certified Data Engineer – Associate (DEA-C01)
- AWS Certified Cloud Practitioner (CLF-C02)
- Microsoft Certified: Azure Data Fundamentals (DP-900)

EDUCATION

Bachelor of Science in Computer Science, Minor in Computer Information Technology
Northern Kentucky University, Highland Heights, KY