

SWAPNIL BHAKARE

Data Engineer | AWS | PySpark | ETL Pipelines | Big Data

Pune, Maharashtra, India | +91 7030001343 | swapnilbhakare19@gmail.com | linkedin.com/in/swapnil-bhakare

About Me

Data Engineer with 4+ years of experience, mostly working on AWS. I spend most of my time building and fixing ETL pipelines with PySpark and Glue, keeping S3 data lakes clean, and making sure the data reaching BI teams actually makes sense. I've dealt with incremental loads, data quality issues, and Spark performance problems. Before this role, I was at TCS working on an on-prem to AWS migration and managing Airflow workflows for batch scheduling – picked up two awards there. I've worked as part of an 8-member data engineering team and can hit the ground running once I understand what's needed.

Technical Skills

AWS Services	Glue, S3, Redshift, EMR, Athena, Lambda, DMS, Step Functions, EventBridge, SNS, SQS, KMS, Secrets Manager, EC2, CloudWatch, IAM
Programming	Python, PySpark, SQL, Apache Spark, Databricks
Data Engineering	ETL Design, Incremental Loads, Data Validation, Schema Checks, Source-to-Target Reconciliation, Deduplication, Backfill & Reprocessing
Orchestration	Apache Airflow, AWS Step Functions, EventBridge
Storage	Parquet (Snappy), Delta Lake, HDFS, Kafka, Hadoop
Tools	Git, GitHub, Linux, CloudWatch Logs, Jira, Agile (Scrum)
Databases	Redshift, Athena SQL, Microsoft SQL Server

Work Experience

Data Engineer

Newgen Technomate LLP

Jan 2024 – Present

Remote

- Built and maintained **PySpark-based Glue ETL jobs** from S3 – handling transformations, business logic, and validations before writing output as **Parquet with Snappy compression**.
- Set up **incremental load logic** using timestamps and primary keys so pipelines only pick up new or changed records – this cut full reloads entirely and brought pipeline run time down by roughly **40%**.
- Wrote **backfill and reprocessing logic** for partitioned S3 datasets so the team could fix historical data issues without touching live pipelines – handled corrections going back **12+ months**.
- Put in place **data quality checks** at the staging layer – null validation, schema checks, row count matching, and source-to-target reconciliation. Data accuracy consistently stayed at **99%+**.
- Used **CloudWatch** to monitor Glue jobs daily, tracked failures and slow runs, and fixed performance issues by tuning Spark partitions and file sizes – cut average job runtime by about **30%**.

Programmer – Data Engineering & Support

Tata Consultancy Services (TCS)

Jan 2022 – Dec 2023

Pune, India

- Was part of an **on-prem to AWS migration project** – helped build and tune Glue ETL jobs and PySpark scripts to move data across S3, Redshift, Athena, and EMR.
- Wrote and maintained **Airflow DAGs** to schedule and manage batch pipelines – set up retries, alerts, and dependency handling so failures didn't go unnoticed.
- Used **Athena SQL** for data checks and validation queries; also kept an eye on pipeline runs daily and fixed issues by going through logs and traces.
- Handled production issues on live pipelines – log analysis, access fixes, and incident resolution to keep data workflows stable on Linux environments.

Projects

Retail Data Analytics Pipeline

AWS Glue · S3 · Redshift · Athena · DMS · Step Functions · Lambda · PySpark

- Built a full data pipeline for retail data – starting from raw ingestion via **AWS DMS**, through Glue-based PySpark transformations, to clean output in S3 ready for Redshift and Athena queries.
- Switched storage from CSV to **Parquet with Snappy compression** – storage costs dropped by around **60%** and query times improved noticeably on the Athena side.
- Added incremental logic so only new records get processed each run; used **Step Functions and Lambda** to orchestrate the workflow and handle scheduling.

Education

Bachelor of Computer Applications (B.C.A) – 75.89%

Dr. Babasaheb Ambedkar Marathwada University

June 2017 – April 2020

Aurangabad, India

Awards

- **Xcelerate Victor Award – TCS (2023)**: Given for taking ownership on a key delivery that had high business impact.
- **Xcelerate Warrior Award – TCS (2022)**: Recognized for picking up new AWS and data engineering skills quickly and applying them on the job.

Languages: English | Hindi | Marathi