Sourab Saklecha

San Jose, CA (Open to Relocation)
Phone: (408) 581-4470 | Email: saklechasourab@gmail.com
LinkedIn: linkedin.com/in/sourabsaklecha/ | Portfolio: ssaklecha.github.io/

Summary

Data Engineer with 3 years of experience designing cloud-based data platforms, building scalable ETL pipelines, and automating data workflows. Proficient in SQL, Python, and big data distributed processing with Hadoop and Spark across AWS, Azure, and GCP. Demonstrated success by reducing data infrastructure costs, improving data reliability, and enabling self-service analytics through modern tools like dbt, Power BI, and Databricks. Adept at collaborating with cross-functional teams to drive data-driven decision-making.

Work Experience

Data Engineer II **Feb 2025 – Present**
WinWin Labs, Plymouth, MI

- Built a reusable, **cloud-native analytics platform**, based on **Kimball methodology,** using GA4, **BigQuery**, and **CloudSQL** that unified marketing and product data for a gaming platform—driving 80% self-service reporting.
- Engineered **cost-optimized pipelines** and data models with **partitioned** GA4 event tables, **clustered** user-level data, and **daily rollups**—reducing BigQuery test workload costs by ~20% and improving dashboard speeds.
- Delivered **dashboards** in **Looker Studio** on 10M+ events, uncovering insights into acquisition funnels and user behavior—now **used as a POC** to win new consulting clients.

Data Engineer I **Jul 2022 – Jan 2023**
Core Molding Technologies Inc., Columbus, OH (Remote)

- Spearheaded the deployment of an **enterprise data warehouse** using a **snowflake schema** in **Azure Synapse**, integrating sources on on-premises, **ADLS, ADB, REST APIs, and ERP** systems using **Azure Data Factory**.
- Built and orchestrated **15+ batch pipelines** in **ADF**, integrating **dbt** for transformation versioning, staging-to-prod workflows, and **automated testing**—enhancing pipeline reliability and deployment efficiency.
- **Reduced storage costs** by ~20% with use of **columnstore indexing,** type optimization, and **partitioning**. Improved **Power BI** dashboard load times by 35% (9+ secs) by **tuning query plans** and performance tuning.
- **Migrated** legacy reports from **Cognos to Power BI**, enabling self-service analytics across Sales and Finance.

Data Engineer I **Apr 2022 – Jul 2022**
JoulesToWatts Pvt Ltd., Bangalore, India

- **Migrated on-premises BI** system to **AWS**. Built scalable ETL pipelines using **AWS Glue** and **PySpark** for HR data from **S3** to **AWS Redshift** OLAP models—resulting in 40% faster dashboard load times.
- Engineered **automated data validation** across the DE lifecycle checks using **AWS Lambda** and **dbt** to **detect schema mismatches**— saving 9 hours of daily manual effort for 6 teams**.**
- Led cross-team efforts to implement **Role Based Access Controls** at the data/reporting level using **AWS IAM** roles and policies—cutting downstream unauthorized access incidents by ~75% and aligning with audit controls.

Jr. Data Engineer **Oct 2020 – Apr 2022**
Santander Bank, Boston, MA (Remote)

- **Migrated** 10+ TB of operational data from on-prem SQL Servers to Azure using 40+ **parameterized ETL pipelines** (**Azure Data Factory + Databricks**), achieving zero downtime during business.
- Designed and deployed **CI/CD pipelines** via **Azure DevOps** to automate **ADF** and **Databricks** notebook deployments across environments, reducing release time by 40% and improving deployment consistency.
- Developed **automated validation workflows** on ADF using **Python** with **hash checks** and **sampling logic**, achieving 99.9%+ data accuracy post-migration and streamlining compliance checks for internal audits.
- Leveraged **Azure Monitor** for alerts and **Power BI** to deliver near **real-time dashboards** on pipeline health and migration KPIs, enabling early bottleneck detection and improving SLA compliance.

Data Analyst Intern **Sep 2019 – Mar 2020**
Infosys Ltd., Mysore, India

- **Cleaned and transformed marketing data** from GA, Mailchimp, and Google Ads using **Power Query**; built **automated Power BI dashboards** and 20+ ad-hoc reports with **DAX** and **Pivot Tables**—cutting prep time by 80% and driving campaign strategy through insights on CTRs and conversions.

Education

- **MS in Data Analytics**, San Jose State University, San Jose, CA, USA. **Jan 2023 – Dec 2024**
- **BE in Computer Science**, VTU, India. **Aug 2016 – Oct 2020**

Technical Skills

- **Programming Languages:** SQL (T-SQL), Python (Pandas, PySpark), Java, Scala.
- **Databases/Storage:** MySQL, PostgreSQL, MongoDB (NoSQL), Cassandra.
- **Big Data Processing:** Apache (Spark, Airflow, Kafka, Iceberg), Hadoop (HDFS, Hive, Pig, MapReduce).
- **ETL:** Power Query, SSIS, PySpark.
- **Data Warehousing:** AWS Redshift, Google BigQuery, Snowflake.
- **AWS:** S3, Glue, RDS, Athena, Lambda, EMR, Aurora.
- **Microsoft Azure**: Blob, ADLS, Azure Data Factory, Databricks, Azure Synapse Analytics, Microsoft Fabric.
- **Others**: JIRA, Git.

Projects

Optimizing Ride-Hailing—Big Data Analytics (GCP, BigQuery, Microsoft PowerBI) (***GitHub)***

- Developed an automated cloud ETL pipeline to process 84.5M records from NYC taxi rides using GCP (Storage), Mage (ETL), and BigQuery (Warehousing).
- Achieved high throughput (200k+ records per min) through efficient orchestration and scalable architecture for Parquet files.

Real-Time Stock Market Data Pipeline—Data Engineering (Apache Kafka, Python, AWS) ***(GitHub)***

- Built a real-time pipeline that ingests high volumes of stock market data (handling 100,000+ records per minute) using Kafka.
- Employed AWS S3 for scalable storage, Glue for automated schema inference, and Athena for interactive SQL-based analytics.

Subscription Churn Rate Prediction for OTT Platforms—Machine Learning (Python, SQL, Tableau) (***GitHub)***

- Achieved 97% accuracy in predicting potential subscription churn with the Random Forest predictive modeling.
- Conducted data transformation and feature engineering to enhance the performance of SVM model by 12%.

Certifications

- Microsoft Certified: Fabrics Data Engineer Associate (DP-700) – *In Progress, May 2025*
- Microsoft Certified: Power BI Data Analyst (PL-300)
- Microsoft Certified: Azure Data Fundamentals (AZ-900)
- NPTEL Certified: Programming, Data Structures, and Algorithms using Python (From IIT, Madras)