

Saurav Karki

+977-9867148963 | sauravkarki386@gmail.com | [linkedin.com/in/saurav-karki-3091601b2](https://www.linkedin.com/in/saurav-karki-3091601b2) | github.com/saurab86

Data Engineer with over 4 years of experience in cloud services and big data technologies such as AWS, Azure, Spark and RDBMs. Associated with US Health care domain to create, maintain and monitor healthcare based product. Highly proficient in programming languages such as Python, Java, C++, PySpark and SQL, and have a deep understanding of data warehousing, relational database systems, analyzing and transforming data, maintaining ETL/ELT workflows and also have expertise in BI tools(Superset and Looker studio), web scraping and scraping tools.

EDUCATION

Nepal Engineering College, Pokhara University

Bachelor of Engineering in Computer Science

Changunarayan, Nepal

Nov. 2016 – July 2021

EXPERIENCE

Data Engineer-II, Cedar Gate Technologies

Lalitpur, Nepal

April 2025 – April 2026

- Created ETL pipeline to extract the data from the source and implement data by using Cedar Gate Technologies in-house tool.
- Transformed the extracted data by using different config module of built in tool of Cedar Gate to dump data in client specific required format.
- Prior to release, the eligibility, pharmacy, and medical claim data were examined in order to identify and address any issues with the data in accordance with US health care standards.
- Worked together with the Data QA teams to create the analysis report and address any issues that the teams brought to light.
- Guide and help the junior engineers to ensure the task running smoothly as well as groom them to grow technically and professionally.

Data Engineer, Fusemachines Nepal

Kathmandu, Nepal

August 2022 – 2025 March

- Designed, built, and maintained ETL solutions to extract, transform, and load data from various sources into a data warehouse or data lake.
- Created different metrics by using SQL queries which then used to visualize data in Visualization tools - Apache Superset.
- Designed Data masking pipeline to encrypt sensitive data from prod server while migrating data to dev server.
- Used AWS Cloud services such as s3 bucket, Athena, EC2 and Glue job for transforming data with PySpark.
- Used Azure Synapse to automate the ETL pipeline from the client data source and use PySpark pool for the data transformation.
- Automated the whole pipeline and monitoring process while running the pipeline.

Data Engineer Trainee, Fusemachines Nepal

Kathmandu, Nepal

August 2022 – Jan 2023

- Collaborated with mentors to perform ETL, data analysis, cleaning, and transformation using SQL, Python, PySprk, AWS(S3, Glue Gob, Athena, RDS, EC2) and Looker during a structured trainee program.
- Gained hands-on experience with modern data tools including Snowflake, Databricks, Azure Synapse, Superset for ETL using data warehousing and visualization tasks.

PROJECTS

Internal Analytics | *PySpark, AWS (EC2, Glue, Athena), Airbyte, Airflow, Superset, Docker*

- Loaded and transformed data from multiple sources as per business requirements and stored on s3 bucket in medallion architecture.
- Created business logic based metrics using SQL queries and visualize them by using Superset for actionable insight
- Hosted and integrated cloud tools (Airflow, Airbyte, Superset) to build a scalable data infrastructure.
- Deployed the development environment to production and troubleshooted issues faced during the transition.
- Reduced data processing time by 30% by optimizing PySpark jobs and AWS Glue workflows.
- Conducted tool research and resolved issues related to data ingestion and transformation pipelines.

Heathcare Benefit Analytics | *AWS, In-House ETL Tools, SQL, Python, Redshift*

- Designed and implemented in-house ETL pipelines using Cedar Gate Technologies tools, including File Mapper and Operations Manager, to extract data from multiple healthcare source systems.
- Configured and executed data transformations using built-in configuration modules to standardize data and load eligibility, pharmacy, medical, and other data-type claims data into client-specific required formats.
- Performed pre-release data validation and reconciliation on eligibility, pharmacy, and medical claims datasets to ensure data accuracy, integrity, and compliance with U.S. healthcare standards.
- Collaborated closely with Data QA teams to analyze data discrepancies, prepare analytical reports, and resolve issues identified during validation cycles.

Find-A-Play-ETL | *Beautiful Soup, Azure (Azure Database, ADLS Gen2, Azure Synapse),PySpak, Pandas, Docker*

- Created an automated pipeline to scrape and collect data from different plays and drama licensing houses of the United States.
- Transformed and cleaned the crawled data through Pyspark and merged it to standard format with client's data.
- Integrated the slack API integration for the success and failure report of the pipeline.
- Made the processed data accessible to the ML team through Azure synapse for building a drama recommendation system, and to the full-stack team for creating the "Find a Play" APIs.

CERTIFICATIONS

Apache Airflow DAG Authoring

Adavance SQL

Python

Pandas

TECHNICAL SKILLS

Languages: Python, C/C++, Flutter

Cloud Deployment: AWS(EC2, RDS, S3, lambda function, AWS Glue, Athena), Azure Synapse, ADLS Gen2, ADF, Databricks

Database: Azure Database, MySQL, PostgresSQL,MS SQL, MongoDB, Big Queries

Libraries: PySpark, Pandas

Developer Tools: Git, Docker, Databricks, VS Code, Visual Studio, PyCharm

Data Warehousing: RedShift, Snowflake

Data Workflow Management tool : Airflow

Visualization tool: Apache Superset, Looker Studio

Web Scraping: Selenium, BeautifulSoup

Framework: Flask