

SNEHANGSU DE

DATA ENGINEER



+91-888-679-5859



desnehangsu@gmail.com



<https://snehangsude.github.io>

PROFILE

Results-driven Data Engineer with a compelling record of consistently achieving and surpassing data-driven objectives. With over 6 years of professional experience, I have made significant contributions in open-source projects like Shikshalokam in developing & optimizing robust data plumbing systems and ETL processes. Most notably, I have automated and optimized batch processes to reduce manual workload while increasing runtime efficiency by 34%. Seeking to improve my skills over cloud technologies while contributing towards efficient and optimized data pipelines.

SKILLS

MAJOR

- Python
- BASH
- SQL
- DBT
- GCP
- Oracle
- Apache Spark / PySpark
- Apache Kafka
- Apache Flink
- CassandraDB
- Terraform
- Docker

SECONDARY

- BigQuery
- Mage
- MongoDB
- Apache Druid
- AWS Lambda
- AWS Redshift
- Git / GitHub
- PowerBI
- Looker
- Selenium
- DLT

WORK EXPERIENCE

DATA ENGINEER

Cloudcraftz Solutions • 2022 - Present (2 years)

Shikshalokam

- Maintained ETL data pipeline up-time of 97.8% while ingesting streaming and transactional data across 8 different data sources using **PySpark**, **Druid**, **Storage**, and **Python**.
- Orchestrated the data migration from **Apache Druid** to **CassandraDB**, transitioning to event-based data capture increasing performance by 17%.
- Utilised **Apache Flink**, and **Kafka** for real-time event pre-processing, enabling dynamic Cassandra-based reporting to be used in BI tools like **Apache Superset**.
- Developed cloud-agnostic data storage architecture, automated data-pipeline logs, and proposed **MongoDB** schema improvements for enhanced user data validation for support team usability.

HSBC

- Developed an internal dashboard using **ETL** and **PySpark** for data processing from global sources, stored in **BigQuery**, and visualized with **Looker**.
- Enhanced business by organizing exact **geo-locations** from user addresses during onboarding and card issuance, focusing on credit card-only users.
- Executed **SQL** scripts to supply pertinent data to stakeholders for spontaneous inquiries.

Udaan

- Scraped data using **Selenium** from reference websites storing them in a **Data Lake**, using **Redshift** as a **Data Warehouse** for analysing transactional data.
- Designed the **database schema** and helped create a **Kimball data modelling architecture** for vendor deals commodities, prices, purchases allowing for a systemic commodity procurement.
- Analysed and provided daily, weekly reports for daily commodities prices to facilitate price prediction for a day and a week to ensure sustainability and data governance in inventory model as well as JIT model.

DATA SERVICES

Concentrix • 2019 - 2021 (2 years)

Intuit

- Handled complicated data issues by restoring damaged corporate files for Quick-books Enterprise & Premier products utilising **MySQL** commands (foreign keys, inner joins) with a satisfaction score of 89.3%.
- Engaged in technical support to address user-reported database challenges, employing a customer-focused approach. Effectively communicated solutions to non-technical users, showcasing the ability to translate complex data issues into user-friendly explanations.

CUSTOMER OPERATIONS

Amazon • 2016 - 2019 (3 years)

- Managed customer relations, participated in the pilot team for Concession Abuse Prevention, helping create business protocols and policies based on live customer queries.
- Served as a Subject Matter Expert (SME) in helping new hires understand Concession Abuse, along with helping team managers with sales queries data collection and analysis allowing for a streamline process for better business-oriented results.

CERTIFICATIONS

- Data Structures and Algorithms - **Jovian**
- Get Started with Databricks - **Databricks**
- Introduction to Docker - **KodeKloud**
- Python Certified - **Cutshort**

EDUCATION

BACHELOR OF TECHNOLOGY

JIS College of Engineering : **2012 -2016**

PROJECTS

Audible Scraper

- This is a project that scraped Audio-books data for all possible categories present in audible.in. This is done using both **Selenium** and **Beautiful-soup** libraries of Python. I have used Selenium to manually navigate pages and BeautifulSoup to scrape the data from the navigated web-pages.

NYC Taxi Trips: ETL to Dashboard

- This is a project that gathers data from NYC Taxi Trips open data for the year of 2019-2023. To see the rise & change of rides from pre and post-covid era. I've used: **dlt** to extract the data. **Mage** for Orchestration. **Bigquery** as Data Warehouse. **DBT** for generating analytical data. And finally, **Looker** for Dashboarding.

Beyond The Bias (InProgress)

- This is a project that scrapes RSS news feeds to generate news summaries while also gathering detailed news from actual pages. This is a stream-based project focusing on **Apache Kafka** and **Apache Flink** to extract the data. Store in **CosmosDB** and then using LLMs to gather similar news and then summarize them.