Gaurav Sahadev

⚲ Kolkata   ✉ gauravsahadev@outlook.com   ☎ +91-7979788681   🌐 https://gauravsahadev.github.io/

Work Experience

Cognizant   Kolkata, West Bengal

Associate (Data Engineer)   *January 2020 - Present (7 months)*

- Have worked on applications to load incremental, Historical data from multiple data sources daily through JDBC to Raw layer of Azure Blob Storage, transform and load data to successive layers(to unified, refinement) through databricks.
- Have worked on failure of ingestion jobs , do data validation, analyse data issues using databricks. Identified causes of issues within applications and reported to respective developers and Business POC.
- Have worked on an application which ingests full/incremental data(retail related) daily from multiple databases using azure data factory and load into raw layer, used spark data frame to transform the data and load into output layer.

Programmer Analyst   *December 2020 - December 2022 (2 years)*

- Involved in Analysis, Reverse Engineering of the existing Workflows and Mappings in Informatica PowerCenter.
- Designed and developed Azure Data Factory pipeline and databricks notebook using existing Informatica PC logic.
- Implemented SCD Type 2 from staging to warehouse Tables using Azure Data factory.
- Conducted unit testing and data validation to ensure accuracy and consistency of data across different systems.

Education

Siliguri Institute of Technology   Siliguri

Bachelor of Technology Computer Science   August 2016 - July 2020

Skills

Programming Languages:	Java, Python, Shell, SQL
Architectures & Concepts:	ETL (Extract Transform Load), Data structures & algorithms
Databases:	MySQL, Oracle, & Azure SQL Database
Frameworks:	Databricks & Cloudera Data Platform
Other Technical Skills:	Data wrangling, Data Analysis using Python, PySpark, Hive, Hadoop, HDFS
Azure Services:	Azure Data Factory, Azure Data Lake Storage, Azure Databricks

Personal Projects

Real-Time ETL Pipeline:- *PySpark, Databricks, Azure Data Lake Storage, Azure SQL database*
- Built a Real-Time ETL Pipeline using PySpark and Azure Data Lake Storage.
- Used JDBC connection string to fetch data from the Azure SQL database and created PySpark Dataframe.
- Cleaned the data to remove duplicate and null values from dimension and fact tables.
- Applied transformations on selected fields of Dataframe exported the cleansed data to the ADLS container in parquet format.

Achievements

- Explore ML Facilitator (Speaker) - *Google*
- Software Innovator program (Speaker) - *Intel*

Certifications

- **Azure Data Engineer (DP 203)** - Microsoft
- **Azure Data Fundamentals (DP 900)** - Microsoft
- **Azure Fundamentals (AZ 900)** - Microsoft