Suman Khatua

Senior Data Engineer

+91 8250507926 skhatua19@gmail.com Kolkata, IN https://github.com/sumankhatua
https://www.linkedin.com/in/suman-khatua-254a63168/ https://www.credly.com/users/suman-khatua

SUMMARY

3+ years experienced Data engineer with a demonstrated of working in one of the largest firm globally. Skilled in MySQL, Python, Scala, Hadoop, Hbase, MongoDb, Extract, Transformation & Load (ETL with PySpark), AWS cloud services (Elastic MapReduce, DynamoDB, Redshift Warehousing, Sqoop, RDS, Databricks on AWS cloud). Strong engineering professional with a Bachelor of Technology focused in Electronics and Communication Engineering from Cochin University of Science and Technology.

KEY SKILLS

• Python programming • Data Analysis • MySQL • Amazon Web Services • NoSQL
• Scala programming • Leadership & Training • ETL pipeline • Machine Learning • Data Visualization

TECHNICAL SKILLS

Tools: Jupyter, Pandas, Matplotlib, Seaborn, Python, SQL
Languages: Python(Advance) , Scala, MySQL, NoSQL
Big Data Tools: Spark, Hive, Mapreduce, Hadoop, Hbase, Apache, Sqoop, Flume, Databricks, Data Lake
Cloud: Amazon AWS, Microsoft Azure
Database: MySQL, PostgreSQL, MongoDB, HBase, Redshift, RDS, CosmosDB
MS Excel (Very Basic)

EDUCATION

Post Graduate Diploma in Data Engineering — Nov '21 - Present
IIIT Bangalore — Bengaluru, IN

- **Course Modules:**
  - Data Management and Relational Modelling |Introduction to Cloud Computing & AWS Setup | Introduction to Hadoop and MapReduce Programming
  - NoSQL Databases and Apache HBase | Data Warehousing and ETL | NoSQL Databases and MongoDB | Data Ingestion with Apache Sqoop and Apache Flume
  - Hive and Querying | Amazon Redshift | Introduction to Apache Spark | Spark using Scala

Bachelor of Technology in Electronics and Communication Engineering — Apr '15 - May '19
SOE, Cochin University of Science and Technology — New Delhi, IN

- Secured 75%
- Ranked among top 5% in ECE Batch.

KEY PROJECTS

ETL and Data Analysis | Tech Stack: PySpark

- Extracting the **transactional data** from a given MySQL RDS server to HDFS(EC2) instance using **Sqoop**.
- Transforming the transactional data according to the given target schema using **PySpark**.
- This transformed data is to be loaded to an **S3 bucket**.
- Creating the **Redshift** tables according to the given schema.
- Loading the data from Amazon S3 to Redshift tables.
- Performing the analysis queries.

Domain: Retail | Tech Stack: R | Apr '20

- Objective: Predict the demand of bikes with the available independent variables
- Solution: Developed a **multiple regression** model to predict the demand of shared bikes and providing them a **visualization** if there is any multicollinearity or significance discrepancy using automated python tool(RFE) for course Tuning and manual tools like p-Value or VIF for fine tuning.
- Key Achievement: Developed a model with an **R2 score of 81%** in test data.

PROFESSIONAL EXPERIENCE

Data Engineer (Senior Systems Engineer)
May '22 - Present

Infosys
Bhubaneswar, IN

Building AWS Lambda and Step Functions

- Development and enhancement of the infrastructure required for extraction, optimal transformation, and loading of data from a wide variety of data source applications using AWS data services, Spark SQL, Apache Airflow.
- Build scripts in Scala for generating data flow and templates for the data extraction interface and load the metadata details in a MongoDB.
- Building multiple AWS Lambda functions to convert the generated data in MongoDB to responsible for triggering AWS Step function to execute extraction job and ingesting data into AWS Data lake.

Data Processing & Transformation

- Develop and enhance the data transformation using Spark SQL for integrating multiple data source result sets and apply core transformation logic to produce required datasets for the BI tool to interpret data sets for business insights.
- Orchestrating and scheduling the complex data pipeline using Apache Airflow workflow engine.

MongoDB NoSQL Data converter

- Writing python Script to convert local data to the user defined data model from CSV and Text file.
- Populating workflow and pipeline info from metadata to the converted data.

Data Engineer (System Engineer)
Jul '19 - May '22

Tata Consultancy Services
Bengaluru, IN

Building ETL Pipelines

- Built **ETL pipeline** with Azure Data factory to load data from different on-prem servers to raw layer of **azure Data Lake**. Then cleansing data with Azure Databricks to move the data to cleanse layer for ADRM Data modelling. Applied transformations like scd1 and scd2 on the cleansed data and breaking source tables into different entities in 3NF form as per ADRM model to make the data available in stage layer.
- Developed daily pipelines to migrate millions of data insights from the stage layer ADRM data to Azure **Cosmos DB** using spark-cosmos connector. proper **incremental** logic, **transformation** based on the modelling, optimization and partitioning done in spark to move the data to Cosmos DB effeciently. This helped our client to reduce cost by 40%.
- Implemented near real time batch pipeline in Azure Data Factory and Spark which will load one **transactional** table incrementally to CosmosDB after doing proper cleansing and transformations.
- Used Databricks Autoloader as streaming source for a requirement where raw xml files are placed in Azure Data Lake. Processed and transformed xml files into structured data and loading into Spark Data table. Finally exposing the data to CosmosDB.

CERTIFICATIONS

- **AWS certified Cloud Practitioner** | AWS
- **3X Microsoft Azure Certified** | AZURE
- **Infosys Machine Learning certified** | Infosys
- **Python Data Science certification** | LinkedIn
- **Advance SQL for Data Science** | Udemy

ADDITIONAL INFORMATION

- **Languages: English (Indian)**
- **5 star Gold badge in SQL on HackerRank platform.**
- **5 star Gold badge in Python on HackerRank platform.**
- **2 star Bronze badge in Problem Solving on HackerRank platform**