Harshil Tyagi [Data Engineer]

(437) 662-3845 | tyagih@uwindsor.ca | Linkedin | GitHub

Experienced Data Engineer with 5 years of expertise in managing infrastructure on Cloud, collaborating across teams to ensure data availability, and implementing solutions to enrich and streamline data processing workflows

TECHNICAL SKILLS

Programming Languages: Python, C, SQL, and Scala
Cloud Platforms: AWS (EMR, EC2, Lambda, MWAA, Athena, Glue) and Azure (DataBricks)
Data Analysis and Modeling: Regression Analysis, Clustering, and Natural Language Processing (NLP)
DBMS: MongoDB, DynamoDB, MySQL, and Microsoft SQL Server
Workflow Orchestration: Docker, Kubeflow, and Apache Airflow

WORK EXPERIENCE

Involead **Delhi, India**
Senior Data Engineer Oct 2022 - Aug 2023

- Strategically implemented and managed services on **AWS** and **Azure** platforms, specifically designed to cater unique needs of American Healthcare clients
- Created an intricate linkage system utilizing **FAISS** (Facebook AI Similarity Search) to connect internal and external data sources circumventing limitations of traditional **SQL** joins
- Led project to connect Healthcare Practitioners (HCPs) in the Client's CRM with external sources (PubMed, Citeline, and Twitter). Engineered HCP-Linking infra enabling comprehensive analysis beyond internal databases
- Conducted cross-departmental training workshops on advanced data analysis techniques, empowering teams to leverage data-driven insights for strategic planning and decision-making

Zupee **Gurugram, India**
Data Engineer Oct 2021 - Oct 2022

- Designed pipelines on AWS including writing scripts to move 100+tb of data to AWS S3 utilizing **Apache HUDI**
- Implemented NoSQL database **MongoDB** with BI tool Metabase and Tableau for deprecated HBase databases
- Integrated CSV-JSON data sources through **PySpark** jobs on **AWS Glue** and AWS **MWAA(Airflow)**
- Presented POC on newly adopted technologies **Apache Airflow** and **DataBricks** and build organization level functional libraries

Infocepts **Nagpur, India**
Data Engineer Mar 2019 - Oct 2021

- Contributed to designing and delivering production grade data warehousing solutions for an American mass media and entertainment conglomerate involving migration of traditional on-premises servers to cloud(AWS)
- Streamlined operations by developing **shell and Python** scripts to automate time-consuming manual tasks, reducing processing time by 10 hrs, and developed scripts for daily ingestion and data loading using PySpark
- Led design and implementation of the AWS platform (**EC2, S3, Athena, and Glue**) for clients, working closely with cloud architects to ensure optimal performance, scalability, and security
- Developed multiple Python/PySpark utilities to migrate databases on AWS **Athena**, create schema on **Glue**, and monitor/pool spark applications **EMR clusters**

PROJECT

Semantic Search Engine Sep 2023 - Jan 2024
The project explored integration of Elasticsearch, NLTK (Natural Language Toolkit), and multiple ways of similarity matching (Cosine and Word2Vec) on academic papers, articles, and scholarly documents using CiteSeer database

EDUCATION

Master of Applied Computing Sep 2023 - Jan 2025
University of Windsor, Windsor, Ontario
Relevant Subjects: Information Retrieval Systems, Emerging Non-Traditional Databases, Adv. Database Topics

- Final semester of program requires a 4- or 8-month internship that would start in September 2024

Bachelor of Technology, Computer Science and Engineering Jun 2015 - May 2019
SRM Institute of Science and Technology, Chennai, India
Relevant Subjects: Machine Learning, Data Science and Big Data Analytics, Artificial Intelligence