Haseeb Khan

Mobile:+91-9945023052 Email: haseebkhan196hk@gmail.com

Professional Summary:

Experienced Data Engineer with expertise in building scalable data infrastructure, designing and implementing ETL pipelines, and optimizing data workflows. Proficient in SQL, Python, PySpark, and cloud technologies such as Azure Data Factory, Azure Databricks, and Azure Storage Accounts.Skilled in end-to-end data engineering, from ingestion to transformation and orchestration, using Azure services.

Skills Summary

Languages: Python, SQL, Pyspark
Data Warehousing: Snowflake, Databricks
Version Control: Git, GitHub
Big Data Technologies: Apache Hadoop, Apache Spark
Tools: Databricks, Azure Data Factory (ADF).
Databases: MySQL, PostgreSQL
Soft Skills: Communication, Adaptability, Problem-solving, Time management

Education

Cambridge Institute of Technology. Bangalore, India
Bachelor of Engineering (B.E) in Computer Science Engineering 2018 - 2022

Experience

Wissen Technology

Dec-2024-present

Role: Data Engineer

Tech Stack: Azure Databricks, PySpark, SQL, Azure Data Factory (ADF)

Responsibilities

- Extracted, transformed, and loaded data from multiple sources (e.g., APIs, SQL databases) to build comprehensive datasets for reporting
- Processed data-loaded files in **Azure Databricks**, applying **Spark SQL** transformations, with execution managed through ADF pipelines. Proficient in SQL for querying, analyzing, and managing large datasets.
- Proficient in writing complex queries and managing relational databases such as **SQL Server and MySQL**
- Partnered with stakeholders to gather business requirements and deliver end-to-end BI solutions in collaboration with analysts and IT teams

Kozlite Technology

May 2024 - Dec-2024.

Role: Data Integration Engineer

Client: Uber

Project: Revamping Critical Analytical Data Flows for Uber's Data Intelligence

Tech Stack: Azure Databricks, PySpark, Delta Lake, Azure Data Factory (ADF)

Responsibilities

- Collaborated with the Source team to extract and load data into Azure Data Lake Storage (ADLS), creating linked services for source and target connectivity based on specific business requirements.
- Engineered scalable data pipelines to extract, transform, and load (ETL) data from multiple sources into **Azure Data Lake** using **Databricks**, improving **data integration** efficiency by 25%.
- Designed and developed pipelines and datasets deployed within **Azure Data Factory (ADF),** ensuring efficient data movement and transformation. These pipelines are triggered dynamically based on specific load operations, automating the entire **data integration** process from extraction to loading.
- Designing and developing data pipelines, leveraging activities like Move and Transform to orchestrate complex **ETL/ELT processes.**
- Utilized **Databricks utilities (dbutils)** to streamline data workflows, enhancing efficiency and automation throughout the data pipeline process.

- Implemented **SCD Type 2** and **SCD Type 3** logic using Mapping Data Flows, ensuring accurate **historical data tracking** and versioning.
- I have configured **email alerts** for pipeline failures and performance monitoring, ensuring proactive resolution of issues in ADF, and I have used the **Azure logic app** for custom email alerts.
- Collaborate with cross-functional teams to understand business requirements and translate them into analytical solutions.

Byju's Think and Learn Pvt Limited

Aug 2022 - May 2024

Role: Data Engineer

Project: Scalable Data Pipeline for Personalized Learning Insights in BYJU's

Tech Stack: Azure Databricks, **PySpark, SQL**

Responsibilities

- Developed pipelines that can extract data from various sources and merge them into single source datasets in **Data Lake** using **Databricks.**
- Worked on medallic architecture to enhance **data quality** and ensure consistency.
- Developed multithreaded scripts for data **extraction, transformation, and loading (ETL)**.
- Worked on different mode options for **reading/writing a data frame** based on the requirement.
- Built complex data ingestion/processing frameworks using **Azure Databricks/ Pyspark.**
- Optimized data pipelines by **repartitioning**, **caching**, and utilizing **broadcast joins** to improve performance on large datasets, reducing job runtimes by up to **30%**.
- Worked on different file formats and used file compression techniques such as **Gzip** and **Snappy.**
- Monitored the data flow and set up automated alerts to ensure continuous data availability and reliability.

Nikveen Infotech

Data Analyst Intern — Nov 2021 - Jul 2022

Leveraged SQL to extract and analyze extensive datasets from relational databases, generating actionable insights and optimizing business performance. Proficient in crafting intricate SQL queries, including joins, subqueries, window functions, and aggregations, to address complex business inquiries. Conducted rigorous data cleaning processes, identifying and rectifying duplicates, null values, and outliers to ensure data integrity

Compsoft Technology

Full Stack Developer Intern — Sep 2021 - Oct 2021

We developed a client-facing web application that utilized Python capabilities to streamline data processing workflows, construct intuitive user interfaces, and conduct comprehensive analyses of extensive datasets. For example, I developed a Python script that automated extracting relevant data from a large CSV file, significantly reducing manual effort and improving data accuracy.

Achievements

Databricks Certified Data Engineer Associate from **Azure Databricks.**
Recognized as **Best Performer of the Month** at Byju's Think & Learn Private Limited.
Awarded **CSI Accredited Student** status by the Computer Society of India.