

Himanshu Singh

+91 9129491699 | dev88himanshu@gmail.com

[LinkedIn](#) | [Portfolio](#)

PROFESSIONAL SUMMARY

Azure Certified Data Engineer with 4+ years of experience designing and optimizing data pipelines, lakehouse architectures, and data warehouse solutions across Azure and on-premise Hadoop ecosystems. Skilled in ADF, Databricks, Spark, Hive, SQL, and Python, delivering up to 60% performance improvements and ensuring data reliability at scale for Fortune 100 Healthcare and Retail (Pharmacy) clients.

TECHNICAL SKILLS

Cloud & Data Engineering:	ADF, Databricks, Azure Synapse, ADLS/Blob Storage, Data Lakehouse
Big Data Ecosystem:	Spark (PySpark, RDD, Streaming), Hadoop (Hive, Impala, Sqoop, MapReduce, YARN)
Pipelines & Warehousing:	ETL/ELT, Data Modelling, Data Warehousing OLAP/OLTP
Programming:	Python, SQL, Shell Scripting
Orchestration & Workflow:	Tivoli Workload Scheduler (TWS), Apache Airflow, Azure DevOps
Developer Tools:	ServiceNow, Git, WinSCP, PuTTY
GenAI:	LangChain, LangGraph, Streamlit, ChromaDB
Familiar:	Django (Rest API), Oracle, MySQL, PostgreSQL

EXPERIENCE

TATA CONSULTANCY SERVICES

Data Engineer / AI CoE SPOC

KAISER PERMANENTE (Mar 2023 – Present)

Hyderabad, India

- Delivered 20TB+ of healthcare data into Azure Data Lake Storage (ADLS) by designing and optimizing enterprise-scale data pipelines using Azure Data Factory (ADF) and Databricks Delta Lake, enabling faster analytics and reporting.
- Increased data processing efficiency by 60% by automating ETL/ELT workflows across ADF, ADLS, Databricks, and PySpark, ensuring near real-time data availability.
- Improved data accuracy by 100% through implementing robust validation and quality frameworks, delivering highly reliable and actionable business insights.
- Enabled seamless enterprise-wide analytics by integrating on-premise, third-party, and Hadoop data sources (Oracle, Taleo, JSON, Avro, CSV, Parquet) into ADLS via ADF connectors.
- Reduced Spark job runtime by 20% for 2M+ daily records by implementing Spark performance tuning techniques (caching, partitioning, broadcast joins) across Hadoop environments.
- Optimized on-premise Hadoop clusters (HDFS, Hive, Spark) by fine-tuning resource allocation and data partitioning strategies, improving large-scale batch processing performance by 35% and ensuring smooth integration with Azure cloud migration pipelines.
- Serving as SPOC (Single Point of Contact) for the AI CoE (Center of Excellence), building and integrating AI agents to automate workflows, enhance data-driven decision-making, and drive AI innovation across project teams.

TATA CONSULTANCY SERVICES

Data Engineer

WALGREENS (Oct 2021 – Feb 2023)

Noida, India

- Accelerated large-scale text processing by 90% (2 hours → 15 minutes) by developing and optimizing MapReduce programs in the Hadoop ecosystem, significantly improving ETL throughput.
- Enhanced operational efficiency by 50% by automating data workflows using Python, HiveQL, and Shell scripting, reducing manual intervention and improving reliability.
- Improved analytical data trustworthiness by 30% through data quality, consistency, and reconciliation checks using Hive, Impala, and Spark.
- Streamlined ETL orchestration and reduced manual efforts by 40% by automating complex data pipelines with Python, HiveQL, and Shell, achieving a 60% improvement in throughput.
- Supported cloud migration initiatives by managing data ingestion and batch processing in Hadoop Distributed File System (HDFS), laying the foundation for Azure-based architectures.
- Delivered enterprise-wide reporting solutions by collaborating with business teams and processing large-scale datasets across distributed ETL systems.

OPEN SOURCE CONTRIBUTIONS

Liu Embeddings

[PyPI] [Docs]

Developed and published an open-source Python library, **liuembeddings**, which combines **text embeddings** and **vector storage** for efficient local NLP retrieval tasks. Built on **Hugging Face Transformers** and **ChromaDB**, it offers a **free, fully local, and privacy-friendly** alternative to cloud-based embedding services ideal for **AI projects** requiring fast and cost-efficient **embedding, storage, and retrieval** workflows.

DbMeta

[PyPI] [Git]

Developed and published an open-source, zero-dependency Python engine, **DbMeta**, designed for parsing and analyzing batch job metadata from **Hadoop, Spark, Databricks** by using folder-based job-run structures. Automatically detects **metadata** files, extracts structured metadata, and converts them into **queryable tables** for efficient monitoring and analytics.

CERTIFICATIONS

Generative AI (Udemy) | 2025

[View Certificate](#)

Databricks Fundamentals | 2025

[View Certificate](#)

Azure Fundamentals (TCS Digital) | 2025

[View Certificate](#)

Python Programming (TCS Digital) | 2022

[View Certificate](#)

Microsoft Certified: Azure Data Fundamentals (DP-900) | Expected 2025



PROJECTS

LeetCode SQL to PySpark (Top 50 Questions)

[View on GitHub](#)

DataLemur SQL to PySpark

[View on GitHub](#)

DataLemur SQL to PySpark (RDD Implementation)

[View on GitHub](#)

Databricks Ecommerce analytics

[View on GitHub](#)

Databricks ETL

[View on GitHub](#)

AI Chatbot

[View on GitHub](#)

The Analysis Bot

[View on GitHub](#)

EDUCATION

NIET

Greater Noida, UP

Bachelor of Technology in Computer Science & Engineering

Aug 2017 – Sep 2021

ABC Public School

Gorakhpur, UP

Senior Secondary School

Apr 2015 – Mar 2017