

Bishwajit Prasad Singh

Aspiring Data Engineer | Azure Data Factory • Databricks • PySpark • Delta Lake • Real-Time ETL • SQL
bishwajit.1804@gmail.com | +91-8882218924 | GitHub: github.com/bishwajitSingh123 | LinkedIn:
linkedin.com/in/bishwajitsingh | Portfolio: bishwajitsingh123.github.io

Professional Summary

Aspiring Data Engineer with hands-on experience in building scalable ETL pipelines using Azure Data Factory, Databricks, and PySpark. Proficient in Python, SQL, and real-time data streaming architectures. Skilled in Delta Lake, NLP-driven applications, and deploying AI-powered tools with Streamlit. Familiar with AWS services and database systems (RDBMS, NoSQL). Strong foundation in Data Structures, Algorithms, and OOP. A collaborative team player and a quick learner with a problem-solving mindset.

Technical Skills

- Cloud Platforms: Azure (ADF, Databricks, Event Hub, Stream Analytics, Data Lake Gen2, Key Vault, Azure DevOps), AWS (S3, EMR, Glue, Redshift, RDS)
- Data Engineering: ETL/ELT (Batch & Real-time), Bronze–Silver–Gold Architecture, Delta Lake (Managed & External Tables), Streaming Pipelines, REST API Ingestion, JSON & Parquet Formats, CI/CD (Azure DevOps)
- Programming Languages: Python (Django, Flask, Pandas, NumPy), PySpark (RDD & DataFrame), SQL (Joins, Aggregations, Window Functions), Shell Scripting, OOP, DSA
- NLP & ML Tools: TF-IDF, SpaCy, NLTK, Cosine Similarity, PDF Parsing (PyMuPDF), Streamlit
- Tools & Platforms: Git, GitHub, Visual Studio Code, Jupyter Notebook, Postman, Power BI, Azure Portal

Projects

♦ AI Resume Matcher Suite | [GitHub](#)

Tech Stack: Python, Streamlit, TF-IDF, SpaCy, Azure Databricks, Cosine Similarity

- Developed 2 intelligent NLP-based tools to match resumes with job descriptions using TF-IDF and cosine similarity.
- Extracted top resume skills with over 90% accuracy, added filters (match %, location), and export options.
- Reduced manual screening workload by 70% with automated matching engine and intuitive Bhagwat-themed UI.

♦ Olympics Data Lakehouse | [GitHub](#)

Tech Stack: Azure Data Factory, Azure Databricks, PySpark, Delta Lake, CI/CD, Azure DevOps

- Built a scalable Lakehouse architecture to process Olympics data using ADF pipelines and PySpark jobs.
- Automated ingestion & transformation using parameterized pipelines and deployed Gold Layer via CI/CD.
- Enabled seamless incremental data updates through Delta Lake optimization.

- ◆ NYC Yellow Taxi Streaming Pipeline | [GitHub](#)

Tech Stack: Azure Event Hub, Stream Analytics, Azure Databricks, Delta Lake

- Engineered a real-time streaming pipeline from ingestion to analytics-ready layers.
- Processed taxi trip data live into Bronze → Silver → Gold Delta tables for real-time insights.

- ◆ AdventureWorks Data Warehouse | [GitHub](#)

Tech Stack: SQL Server, SSMS, ETL, Star Schema

- Designed a robust data warehouse from OLTP data to improve reporting and business decision-making.
- Modeled fact/dimension tables, applied SQL-based ETL, and built star schema for BI consumption.
- Improved query performance by 50% with indexed tables and optimized joins.

Education

B.Tech in Computer Science Engineering – NSHM Knowledge Campus, Durgapur, WB | 2023

CGPA: 8.4

Certifications

- Microsoft Certified: Azure Fundamentals (AZ-900)
- Microsoft Certified: Azure Data Fundamentals (DP-900)
- Microsoft Certified: Azure Data Engineer Associate (In Progress)