Abhay Singh

Data Engineer

+19309044335 | abhaysaikap@gmail.com | Bloomington, IN | LinkedIn | GitHub | Portfolio

PROFESSIONAL EXPERIENCE

Data Engineer, **O'Neill School Of Public And Environmental Affairs | Indiana, USA** **May 2025 - Present**

- Engineered real-time tracking of 500+ toxic leakage incidents by productionizing a Flask and React-based web search tool, an ArcGIS-powered geospatial platform, and interactive visualizations.
- Deployed accurate and cost-efficient environmental toxicology insights to communities by building a GenAI-powered chatbot using a two-tier RAG architecture containerized via Docker on AWS.
- Automated environmental risk KPI monitoring across 92 Indiana counties by orchestrating event-triggered ETL workflows in Python using Airflow and PostgreSQL to process 50,000+ EPA TRI emission records monthly.

Senior Data Engineer, **AP Moller Maersk | Bangalore, India** **Apr 2022 - Jun 2024**

- Minimized shipment delays by 45% by developing Kafka and PySpark micro-batch streaming pipelines that processed 400–500 GB/day of IoT data from 60+ terminals, enabling data-driven insights on shipment delays.
- Streamlined manual data validation effort by 88% by building a data quality framework in Python and SQL on Databricks, deployed via CI/CD pipelines on Azure DevOps with automated validation and alerting workflows.
- Boosted routing efficiency by 74% and saved $100K–200K annually by deploying a CUDA-optimized ensemble (RF, XGBoost, CNN) as a GPU microservice on AWS ECS for large-scale inference.
- Reduced vessel travel time by 4% and redundant crane moves by 6% across global trade hubs by applying K-Means clustering and spatial optimization to streamline container allocation and stacking.

Data Engineer 2, **Microsoft (Contract) | Hyderabad, India** **Apr 2019 - Apr 2022**

- Expedited ETL runtimes from 20–22 hrs to 5–6 hrs by building concurrent Azure Data Factory and Databricks pipelines processing terabytes of data across 10+ metrics, ensuring SLA compliance.
- Refactored report performance by 97% and cut OLAP load times from 130–140s to 2–3s by building pre-aggregated fact tables, optimizing DAX measures and enhanced data warehousing for Power BI reporting.
- Cut Spark refresh runtime by 66% (3 hrs to 1 hr) by profiling long-running stages and eliminating shuffle-heavy joins using partition and caching optimizations guided by Spark execution plans.
- Addressed 200+ critical support requests for Microsoft's Finance team, ensuring the accuracy and availability of high-impact business data used in executive reporting, audits, and operational planning.

EDUCATION

Indiana University, Bloomington, USA - *Master of Science*, *Data Science* **Aug 2024 - May 2026**

PROJECTS

ServiceNow Idea Portal AI Agent GitHub

- Cut idea evaluation time from 3 hrs to 30 mins by designing and deploying an AI-driven platform for ServiceNow, combining Random Forest based metric weighting with GPT-4o and LangChain, built using the ReAct framework and hosted via Streamlit UI.

PoseNet: Ape Pose Detection GitHub

- Attained 93.4% PCK@0.2 accuracy in primate pose estimation by fine-tuning ViTPose and ResNeXt models using the MMPose library on 70,000+ annotated images; developed a CUDA-enabled Hugging Face training pipeline with data scaling.

Ask DocAI GitHub

- Architected a document-aware bot using LLaMA 3.3 70B (Groq) and conversational memory to query domain documents; reduced info retrieval time by 80%. Leveraged BERT embeddings and knowledge-graph retrieval via the Leiden algorithm for global context.

SKILLS

Programming Languages : C++, Python, Java, Scala, SQL, Go, R, Javascript

Data Processing : PySpark, TensorFlow, PyTorch, Pandas, NumPy, Scikit-Learn, CUDA, Hadoop, DBT, Airflow

Databases : MySQL, PostgreSQL, SQL Server, MongoDB, Cassandra, Redis, Neo4J

Cloud Technologies : Azure Data Factory, Databricks, Snowflake, ADLS Gen2, Redshift, AWS (EC2, S3), BigQuery

Visualization : Power BI, Tableau, Looker, Matplotlib, Seaborn, Streamlit, Dash

DevOps : Docker, Azure DevOps, GitHub Actions, CI/CD

CERTIFICATIONS

MCSE: Data Management and Analytics, **Microsoft**

MCSA: SQL Database Development, **Microsoft**

Azure Architect Solutions Expert, **Microsoft**

Fabric Data Engineer Associate, **Microsoft**