Ashima Arora

ashimaarora63@gmail.com | (929) 369-9623 | New Jersey, USA | LinkedIn | Portfolio | Github

SUMMARY

Data Engineer with **4+ years of experience** in designing, developing, and optimizing **scalable data pipelines** and **ETL processes**. Proficient in **Hadoop, Spark (PySpark), SQL, and Python.** Experienced in **data architecture, cloud integration, big data processing** and **ETL automation,** with a strong understanding of **software development lifecycle (SDLC)** and **agile methodologies.** Skilled in **batch** and **real-time data processing, data modeling,** and **performance optimization,** while maintaining **clear technical documentation** and **process standards** to support **cross-functional teams** and enable data-driven decision-making.

TECHNICAL SKILLS

Data Processing & Analytics: SQL, NoSQL databases, Apache Hadoop, Apache Hive, Apache Spark, Apache Airflow, Kafka, Tableau
Machine Learning: PyTorch, TensorFlow, Keras, Scikit-Learn, Pandas, NumPy
Programming: Python (Proficient), Shell Scripting (Proficient), C++ (Intermediate), Java (Intermediate)
Cloud Computing: Google Cloud Platform, Kubernetes, Docker, Cloud Storage, BigQuery, DataProc, Compute Engine
DevOps: Jenkins, UrbanCode Deploy, CI/CD Pipelines, Git, GitHub, Monitoring Tools
Collaboration & Process: Agile methodologies, Software development lifecycle, Technical Documentation and Runbook Creation, Process Standardization, Knowledge-Sharing and Mentorship, Best Practices Development
Others: Linux, Confluence**,** JIRA, API Integration and Development, Data Modeling, Data Warehousing, Data Migration, Database Programming, Data Quality Assurance

EDUCATION

- *The University of Texas at Austin* **- Master of Science** in **Computer Science**, *Aug 2022 - 2024; Focus* in **Machine Learning & AI**
- *CUNY Queens College* **- Bachelor of Science** in **Computer Science**, *Aug 2017 - June 2021*

WORK EXPERIENCE

Data Engineer | Wells Fargo | June 2021 to Present

- Engineered 8 scalable data pipelines using Hadoop, Hive, and Spark (PySpark), supporting **batch and near real-time data processing** of up to **10 TB of data monthly** and improving **data pipeline efficiency by 25%**.
- Execute **data integration strategies** to unify over 10 disparate internal and external data sources, enabling a **60% improvement in data accessibility** and analytical consistency.
- Co-architect **cloud based data solutions** to improve data accessibility and enable scalable analytics solutions.
- Built and maintained robust **automated data validation and testing frameworks**, while **writing operational runbooks** to guide incident response, significantly improving data quality and **reducing pipeline-related incidents** by **50%**, and **shortened mean time to resolution (MTTR) by 30%.**
- Streamlined and **automated complex ETL workflows** using **Python and Autosys**, **increasing operational efficiency by 40%** by reducing manual effort and improving data reliability and accuracy.
- **Integrated Apache Kafka** to enable high-throughput, **real-time data ingestion**, **increasing data processing speed by 35%** and **maintaining integration documentation** to support **production stability** and **knowledge transfer** across the team.
- Lead a team of 5 engineers through **Agile sprint cycles**, facilitating **cross-functional collaboration** with product managers, analysts, and DevOps to deliver reliable data pipeline and visualization solutions on time.
- Ensured compliance with industry and organizational data standards by **implementing robust risk mitigation strategies** across data workflows, resulting in a **66% reduction in audit findings** and minimizing exposure to potential data breaches.
- Spearheaded performance optimization initiatives, **reduced product average feature time to market by 40%**, while continuing to keep the total production defect rates below organization benchmarks.

Tutor | Varsity Tutors LLC | January 2020 to April 2021

- Mentored a group of 6 students by **fostering effective communication and collaboration**, leveraging structured learning plans and tailored instruction to address individual skill gaps.
- Utilized productivity and monitoring tools to track progress, optimize study schedules, and deliver measurable performance improvements.
- Achieved 15% increase in student scores and maintained a 5.0/5.0 tutor rating for professionalism and effectiveness.

Research Intern | Research Foundation of CUNY | August 2019 to November 2019

- **Designed a feedback pipeline** to capture and categorize user input (bug reports, feature requests) from testers and panelists, storing data in **Excel/ SQL for structured analysis** and **maintaining clear documentation of findings** for engineering teams.
- Performed **exploratory data analysis (EDA)** on feedback trends to identify high-impact issues and **standardizing prioritization criteria for feature development, reducing time-to-fix for critical bugs by 30%**.
- **Streamlined participant recruitment tracking and built dashboards** to monitor response rates, increasing participation by **35%**.
- Collaborated with engineering teams to translate insights into updated product specifications, improving release velocity and overall user satisfaction.

KEY PROJECTS

Object Detector and Auto-Pilot Driving - *Python, PyTorch, Fully Convolutional Neural Network (CNN)*

- Developed an **object detection program** for **auto-pilot driving in a simulated environment, using Deep Learning** and **Convolution Neural Networks,** achieving **over 90% prediction accuracy** and enabling **reliable real-time decision-making.**

Sentiment Analysis for Amazon's Product Reviews - *Python, Natural Language Toolkit (NLTK)*

- Built a **Natural Language Processing (NLP)** pipeline to preprocess and **classify Amazon Alexa reviews** by **sentiment** using **Random Forest and XGBoost**, achieving **94% test accuracy and** generating **data driven insights from customer feedback**.

CERTIFICATIONS

Google Cloud Certified – Associate Cloud Engineer