Atharva Dumbre

Data Engineer

Harrison, NJ | atharvadumbre001@gmail.com
| +1 (862) 381 9846 | LinkedIn

aws certified Cloud Practitioner FOUNDATIONAL

ORACLE Certified Professional Oracle Cloud Infrastructure Generative AI 2024

TensorFlow Certified Developer

NVIDIA. INTRODUCTION TO AI IN THE DATA CENTER

PROFESSIONAL SUMMARY

Data Engineer with 3 years of experience in architecting and deploying robust data pipelines and infrastructure. Proficient in Python, SQL, and Scala, with extensive hands-on experience in big data technologies such as Apache Spark, Hadoop, and Kafka. Skilled in leveraging cloud platforms (AWS, Azure, GCP) and data warehousing solutions to drive efficiency and scalability. Adept at delivering high-performance solutions that optimize data processes and enable data-driven decision-making for business impact.

SKILLS

Methodologies:	SDLC, Agile, Waterfall
Programming Language:	R, Python, SQL, SAS
IDE's:	PyCharm, Jupyter Notebook, Visual Studio Code
Big Data Ecosystem:	Hadoop, MapReduce, Hive, Apache Spark, Pig, Sqoop, Pyspark, Snowflake
ETL Tools:	SSIS, Apache NiFi, Apache Kafka, Talend, Apache Airflow, Informatica
Packages:	NumPy, Pandas, Matplotlib, SciPy, Scikit-learn, Seaborn, TensorFlow, ggplot2
ML Algorithm:	Linear Regression, Logistic Regression, Decision Trees, Supervised Learning, Unsupervised Learning, K Means, SVM, Random Forests, Naive Bayes, KNN, CNN
Visualization Tools:	Tableau, Power BI, Microsoft Excel
Cloud Technologies:	AWS, Azure, GCP, Databricks
Database:	MongoDB, MySQL, SQL Server, PostgreSQL, HBase, Cassandra, DynamoDB
Version Control:	Git, GitHub, GitLab
Other Skills:	Data Cleaning, Data Wrangling, Communication Skills, Presentation Skills, Problem-Solving
Operating Systems:	Windows, Linux

EXPERIENCE

PTC, USA | Data Engineer **June 2024 - Present**

- Directed Agile, cross-functional teams in designing and implementing data pipelines, increasing project delivery speed by 30% and reducing defects by 20%.
- Engineered and optimized ETL pipelines using Python, SQL, SSIS, and Talend, boosting data processing efficiency by 35%.
- Conducted performance tuning and optimization of ML algorithms, ensuring effective resource utilization and improved processing times for large-scale data applications.
- Developed data processing frameworks using Hadoop, MapReduce, and Apache, reducing processing times for large-scale datasets; Enhanced query performance by 30% through the use of Hive and Apache Spark for large dataset processing.
- Developed and deployed machine learning models using algorithms such as Linear Regression, Logistic Regression, and Decision Trees to drive predictive analytics and enhance data-driven decision-making processes.
- Designed and deployed scalable data storage and processing solutions on AWS and GCP, reducing infrastructure costs by 25% and created interactive, real-time dashboards using Tableau, delivering actionable insights to business stakeholders.
- Implemented data security and compliance protocols in AWS and GCP cloud environments, ensuring industry standards were met; Managed and optimized data warehouses using MySQL and PostgreSQL, improving data retrieval speed and storage efficiency.
- Collaborated with cross-functional teams to integrate data pipelines and optimize workflows, utilizing GitHub for version control.

New Jersey Institute of Technology | Data Engineer Research Assistant & Teaching Assistant **Sep 2023 - May 2024**

- Helped to increase the CS656 (CS656: Internet & Higher Layer Protocols) pass rate to 85% by tutoring 100+ students, leading a team of graders, & Assisted the research team in implementing over 3 reinforcement learning algorithms in Python using PyTorch for traffic light cycle control.
- Accomplished a 7% reduction in average vehicle waiting times compared to fixed-time methods in simulated traffic intersection environments.
- Mentored students in utilizing AWS for research projects, helping them design and implement data engineering solutions using services like S3, Redshift, Glue, and Lambda.
- Assisted faculty in developing course materials for cloud computing and big data analytics courses, focusing on AWS and its services for practical, hands-on learning.
- Built data pipelines on Databricks using Apache Spark, leveraging its in-built integration with AWS S3 for data ingestion and processing. Developed machine learning pipelines using Databricks' built-in MLflow for model training and tracking.
- Integrated dueling DQN, double Q-learning, & prioritized experience replay algorithms into optimization model, experimenting for 1500 epochs. Reviewed and summarized 25+ research papers with PhD students, contributing to a 20% increase in publication acceptance rates.
- Architected and deployed scalable data processing workflows using AWS services such as S3, Lambda, and Glue to support academic research projects, ensuring high availability and fault tolerance.

- Utilized Amazon RDS and DynamoDB for managing and storing structured and unstructured data, creating optimized schemas for various research datasets.
- Achieved 99% annotation accuracy for over 300 images of simulated traffic intersections using Roboflow resulting in improved dataset quality; Conducted in-depth research on big data algorithms and demonstrated their practical use in real-time data analysis.
- Configured, installed, and troubleshooted CARLA & SUMO traffic light cycle simulation environments for over 10 researcher systems.

Citus Infotech, India | Data Engineer **Aug 2020 - Jul 2022**

- Managed end-to-end data integration projects using the Waterfall methodology, ensuring comprehensive planning, execution, and on-time delivery of data pipelines.
- Designed and implemented ETL processes with Informatica, Apache NiFi, and Apache Kafka to optimize data flow across multiple systems, achieving a 30% reduction in data latency.
- Developed and maintained big data processing solutions using Pig, Sqoop, PySpark, and Databricks, enhancing data transformation processes and reducing processing times by 40%.
- Architected & implemented data warehousing solutions in Snowflake, supporting fast querying on multi-terabyte datasets and Set up Snowflake Virtual Warehouses to support multiple workloads simultaneously, ensuring resource efficiency and cost control; Planned scalable data warehousing solutions in Snowflake and Azure, improving storage and query performance for large datasets.
- Designed and optimized data pipelines for feeding training datasets into models utilizing SVM, Random Forests, Naive Bayes, and KNN, ensuring efficient data flow and reducing model training times.
- Conducted advanced statistical analyses and predictive modeling with SAS, NumPy, and Scikit-learn, facilitating data-driven decision-making and enriching business insights.
- Engineered scalable data solutions using Hadoop for batch processing large datasets, enabling management of big data and designed and executed Hadoop-based workflows to analyze terabytes of structured and semi-structured data across various data lakes.
- Orchestrated data processing workflows using Kubernetes, deploying scalable and fault-tolerant applications in containerized clusters. Deployed Kafka clusters and implemented Kafka Connect for seamless integration with various data sources like PostgreSQL and Snowflake.
- Implemented machine learning models using TensorFlow and Scikit-learn to tackle complex data challenges, resulting in a 25% improvement in predictive accuracy.
- Created interactive dashboards and visualizations with Power BI, providing actionable insights for business stakeholders and increasing data accessibility and usability by 35%.
- Managed and optimized NoSQL databases such as MongoDB and DynamoDB, ensuring flexible and high-performance data storage solutions that enhance system scalability and availability.

EDUCATION

M.S. in Computer Science - Concentration: Machine Learning | New Jersey Institute of Technology | Sep 2022 – May 2024

B.E. in Computer Engineering – Concentration: Machine Learning | University of Mumbai | Aug 2018 - May 2022

PROJECTS

Realtime Data Engineering Pipeline – Python | Apache Airflow | Kafka | Spark | Cassandra | PostgreSQL | Docker Link

- Developed an end-to-end data pipeline for real-time data ingestion, streaming with Kafka, processing with Apache Spark, and storage in Cassandra. Orchestrated tasks with Apache Airflow and containerized services using Docker for seamless scalability.

YouTube Video Trends Data Engineering & Analysis – AWS | Power BI | SQL | DAX Link

- Implemented an ETL pipeline using AWS S3, Glue, Lambda, Athena & CloudWatch to process & analyze 70,000+ YouTube Trending Videos from the USA, Canada, & Great Britain. Converted JSON to Parquet & created a dashboard with 5 key performance indicators (KPIs) & trends.

Fintech - Revenue Analysis: Hospitality Domain – Power BI | SQL | DAX Link

- Designed a Power BI dashboard for the Codebasics Resume Challenge, using DAX and SQL to process 90 days of hospitality data. Created 26 new key performance indicators (KPIs) and plotted comparison graphs to analyze trends and performance.

RAG for Pdf's – Python | Streamlit | LangChain | HuggingFace | FAISS Link

- Programmed a chatbot using Streamlit to convert PDFs into a searchable knowledge base with the FAISS vector database, LangChain, & Google's T5 LLM with average query retrieval in 4 seconds.

AI Logo Generator (creative-dalle.com) – MERN (Mongo, Express.js, React, Node.js) | OpenAI | Hostinger | Cloudinary Link

- Launched a logo generation platform with OpenAI's DALL-E 3, featuring "Surprise Me," social sharing, and keyword search. Ensured 99.9% uptime for the platform through Render.com and Hostinger.com cloud hosting, and optimized logo delivery using Cloudinary.

CERTIFICATES, ACHIEVEMENTS & RESEARCH PUBLICATION

- Oracle Generative AI Professional – Oracle | AWS Cloud Practitioner | TensorFlow Developer Certificate – Google | LLaMa for Developers | Foundations of Machine Learning & Applications Bootcamp - IIT Mandi | Neural Networks & Deep Learning | HTML, CSS, & JavaScript for Web Developers | Fundamentals of Deep Learning for Computer Vision – NVIDIA | Introduction to AI in the Data Center – NVIDIA
- Secured 1st place for "Most Creative Use of Redis Cloud" at HackNJIT Oct 2023 | Kaggle Expert (3647 Highest Rank).
- **Publication:** Classification Of Indian Sign Language Characters Utilizing Convolutional Neural Networks & Transfer Learning Models with Different Image Processing Techniques | IEEE Link