Atharva Dumbre

Data Engineer

Harrison, NJ | atharva.dumbre01@gmail.com | +1 (862) 381-9846 | LinkedIn | Github | atharvadumbre.com

PROFESSIONAL SUMMARY

Data Engineer with 4 years of experience in architecting & deploying robust data pipelines & infrastructure. Proficient in Python, SQL, & Scala, with extensive hands-on experience in big data technologies such as Apache Spark, Hadoop, & Kafka. Skilled in leveraging cloud platforms (AWS, GCP) & data warehousing solutions to drive efficiency & scalability. Adept at delivering high- performance solutions that optimize data processes & enable data-driven decision-making for business impact.

SKILLS

Methodologies:	SDLC, Agile, Waterfall
Programming Language:	Python, SQL, R, SAS
IDE's:	PyCharm, Jupyter Notebook, Visual Studio Code
Big Data & ETL Tools:	Snowflake, Pyspark, Apache Spark, Hadoop, MapReduce, Kafka, Apache Airflow, NiFi, Talend, Informatica, SSIS
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, GCP, Databricks
Databases:	MongoDB, MySQL, SQL Server, PostgreSQL, HBase, Cassandra, DynamoDB
Version Control & OS:	Git, GitHub, GitLab, Windows, Linux
Misc:	Data Cleaning, Data Wrangling, Communication Skills, Presentation Skills, Problem-Solving

EXPERIENCE

PNC Financial, USA | Data Engineer **June 2024 - Present**

- Directed Agile, cross-functional teams in designing & implementing data pipelines, increasing project delivery speed by 30% & reducing defects.
- Engineered & optimized ETL pipelines using Python, SQL, SSIS, & Talend, boosting data processing efficiency.
- Maintained & optimized big data processing workflows in Hadoop, ensuring efficient storage & processing of structured & unstructured data sets.
- Designed & implemented scalable data pipelines using Apache Spark & PySpark, processing over 10TB of data daily, reducing data processing time by 40%.
- Developed & deployed machine learning models using algorithms such as Linear Regression, Logistic Regression, & Decision Trees to drive predictive analytics & enhance data-driven decision-making processes.
- Implemented data warehousing solutions in Snowflake, supporting fast querying on multi-terabyte datasets & Set up Snowflake Virtual Warehouses to support multiple workloads simultaneously, ensuring resource efficiency & cost control.
- Industrialized ETL workflows using Apache Airflow, improving job reliability & scheduling by automating over 200 tasks per week.
- Developed & managed data warehouses using AWS Redshift & Snowflake, enhancing data accessibility & enabling seamless BI reporting; Skilled in using Jira for project tracking, ensuring tasks are completed on time & aligned with project goals.
- Collaborated with data analysts to ensure accurate data representation, utilizing Power BI & Tableau for effective data visualization; Applied Dimensional Modeling & Star Schema principles to support analytical queries & facilitate BI processes.
- Designed & deployed scalable data storage & processing solutions on AWS & GCP, reducing infrastructure costs & created interactive, real-time dashboards using Tableau, delivering actionable insights to business stakeholders.
- Spearheaded the creation of a data warehouse using Amazon Redshift, resulting in a 70% improvement in data query speed for data analysts & business users.
- Managed data pipelines across multiple cloud platforms, including AWS (S3, RDS) & GCP, ensuring 99.9% data availability & scalability for diverse business needs; Automated daily data ingestion processes for 5+ data sources, using Apache NiFi, reducing manual data processing time by 60%.

New Jersey Institute of Technology, USA | Data Engineer **Sep 2023 - May 2024**

- Helped to increase the CS656 (CS656: Internet & Higher Layer Protocols) pass rate to 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 great 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 & implement data engineering solutions using services like S3, Redshift, Glue, & Lambda.
- Assisted faculty in developing course materials for cloud computing & big data analytics courses, focusing on AWS & 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 & processing. Developed machine learning pipelines using Databricks' built-in MLflow for model training & tracking.
- Integrated dueling DQN, double Q-learning, & prioritized experience replay algorithms into optimization model, experimenting for 1500 epochs.
- Reviewed & summarized 25+ research papers with PhD students, contributing to increase in publication acceptance rates; Architected & deployed scalable data processing workflows using AWS services such as S3, Lambda, & Glue to support academic research projects, ensuring high availability & fault tolerance.
- Utilized Amazon RDS & DynamoDB for managing & storing structured & unstructured data, creating optimized schemas for various research datasets; Configured, installed, & troubleshooted CARLA & SUMO traffic light cycle simulation environments for over 10 researcher systems.
- 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 & demonstrated their practical use in real-time data analysis.

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

- Managed end-to-end data integration projects using Waterfall methodology, ensuring comprehensive planning, execution, & on- time delivery of data pipelines.
- Designed & implemented ETL processes with Informatica, Apache NiFi, & Apache Kafka to optimize data flow across multiple systems, achieving a reduction in data latency.
- Developed & maintained big data processing solutions using Pig, Sqoop, PySpark, & Databricks, enhancing data transformation processes & reducing processing times.
- Implemented & managed Hadoop ecosystems, including HDFS for scalable data storage & MapReduce for data processing on distributed frameworks.
- Python experience in data processing, with experience in libraries like Pandas & NumPy for data manipulation & pre-processing in ETL workflows; Managed data lifecycle policies on AWS S3 to automate data archiving & deletion, ensuring cost-effective & compliant storage management; Secured & managed access to data resources with AWS IAM, ensuring compliance & data security.
- Planned scalable data warehousing solutions in Snowflake, improving storage & query performance for large datasets. Designed & optimized data pipelines for feeding training datasets into models utilizing SVM, Random Forests, Naive Bayes, & KNN, ensuring efficient data flow & reducing model training times.
- Conducted advanced statistical analyses & predictive modeling with SAS, NumPy, & Scikit-learn, facilitating data-driven decision- making & enriching business insights.
- Engineered scalable data solutions using Hadoop for batch processing large datasets, enabling management of big data & designed & executed Hadoop-based workflows to analyze terabytes of structured & semi-structured data across various data lakes.
- Orchestrated data processing workflows using Kubernetes, deploying scalable & fault-tolerant applications in containerized clusters. Deployed Kafka clusters & implemented Kafka Connect for seamless integration with various data sources like PostgreSQL & Snowflake.
- Implemented machine learning models using TensorFlow & Scikit-learn to tackle complex data challenges, improvement in predictive accuracy.
- Created interactive dashboards & visualizations with Power BI, providing actionable insights for business stakeholders & increasing data accessibility & usability.
- Managed & optimized NoSQL databases such as MongoDB & DynamoDB, ensuring flexible & high-performance data storage solutions that enhance system scalability & availability.

EDUCATION

M.S. in Computer Science | New Jersey Institute of Technology, Newark, NJ, USA | GPA: 3.78 / 4.00 | Sep 2022 – May 2024

B.E. in Computer Engineering | University of Mumbai, Maharashtra, India | GPA: 7.88 / 10.00 | Aug 2018 - May 2022

PROJECTS

Customer data Realtime DE Pipeline – Python | Apache Airflow | Kafka | Spark | Cassandra | PostgreSQL | Docker

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

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

- 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

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

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

- 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

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

CERTIFICATES

- AWS Cloud Practitioner | Oracle Generative AI Professional – Oracle | TensorFlow Developer Certificate – Google
- Foundations of Machine Learning & Applications Bootcamp - IIT Mandi | Neural Networks & Deep Learning - NVIDIA
- Fundamentals of Deep Learning for Computer Vision – NVIDIA | HTML, CSS, & JavaScript for Web Developers – John Hopkins University

ACHIEVEMENTS & REASEARCH PUBLICATION

- Secured 1st place for "**Most Creative Use of Redis Cloud**" at HackNJIT Oct 2023.
- Achieved **Kaggle Expert** Badge (3647 Highest Rank).
- Published "Classification Of Indian Sign Language Characters Utilizing Convolutional Neural Networks & Transfer Learning Models with Different Image Processing Techniques" in IEEE conference - **Link**