

SUMMARY

Dynamic Data Engineer with 3+ years of experience across the full data engineering stack and modern data architectures. Skilled in building end-to-end ETL pipelines using Python, Scala, SQL, and big data tools like Apache Spark, Hadoop, and Snowflake. Proficient in cloud platforms (AWS, Azure, GCP), data workflow orchestration (Apache Airflow, Kafka), and visualization tools (Tableau, Power BI). Strong expertise in machine learning libraries (scikit-learn, TensorFlow) and data analysis (Pandas, NumPy). Experienced in Agile and Waterfall methodologies, optimizing database performance (PostgreSQL, MySQL, Oracle), and implementing data governance frameworks to deliver scalable, business-focused solutions.

SKILLS

<b>Methodologies:</b> SDLC, Agile, Waterfall	<b>Packages:</b> NumPy, Pandas, Matplotlib, SciPy, Scikit-learn, Seaborn, TensorFlow
<b>Programming Language:</b> Python, R, SQL, Scala	<b>Databases:</b> MySQL, SQL Server, PostgreSQL, Oracle
<b>AI/ML Tools:</b> TensorFlow, PyTorch, Keras, Scikit-learn, OpenAI API, LLM, CNNs	<b>Version Control Tools:</b> Git, GitHub, GitLab
<b>Big Data Ecosystem:</b> Hadoop, MapReduce, Hive, Apache Spark, Pig, Sqoop, Pyspark, Snowflake, HDFS	<b>Other Skills:</b> Data Cleaning, Data Wrangling, Critical Thinking, Communication Skills, Presentation Skills, Problem-Solving, Data Management
<b>ETL Tools:</b> SSIS, Apache NiFi, Apache Kafka, Talend, Apache Airflow, Informatica	<b>Operating Systems:</b> Windows, Linux, Mac
<b>Cloud Technologies:</b> AWS, Azure, GCP, DataBricks	
<b>Reporting Tools:</b> Tableau, Power BI, SSRS	

EXPERIENCE

<b>Pfizer, USA   Data Engineer</b>	<b>Jul 2024 - Present</b>
<ul style="list-style-type: none"><li>Designed and implemented scalable ETL pipelines using Apache Airflow, Kafka, and Talend, enabling seamless data integration across on-premises and AWS cloud platforms.</li><li>Developed distributed data processing workflows using Apache Spark and Pyspark to analyze multi-terabyte datasets, reducing processing time by 40%.</li><li>Deployed and managed Snowflake data warehouses, worked on pre-trained LLM's with data scientists such as GPT-3 and BERT to automate and enhance natural language processing (NLP) tasks.</li><li>Engineered near real-time analytics dashboards in Tableau by integrating with Azure Data Lake, enhancing decision-making speed for cross-functional teams.</li><li>Played an integral role in Agile project development, leveraging Jenkins-driven CI/CD pipelines to streamline deployments and integrate new features seamlessly across environments, improving the efficiency of development cycles.</li></ul>	
<b>Metasystems, India   Data Engineer</b>	<b>Jan 2020 - Dec 2022</b>
<ul style="list-style-type: none"><li>Spearheaded the migration of on-premises Hadoop ecosystems to Google Cloud Platform (GCP), leveraging BigQuery for advanced analytics, achieving 50% faster query execution.</li><li>Led the migration of legacy data warehouse to modern Azure Synapse Analytics platform, implementing CI/CD pipelines with GitHub Actions, resulting in 60% faster deployment cycles and improved data quality.</li><li>Built data pipelines using Apache NiFi and Sqoop to ingest structured and unstructured data into HDFS,</li><li>Worked with algorithms like decision trees, random forests, and SVM for improving data analysis, while engineering on HDFS, optimizing big data storage and retrieval with Hive and MapReduce to enable advanced analytics.</li><li>Conducted performance tuning and query optimization for complex SQL scripts on Oracle and SQL Server, achieving a 20% improvement in processing speeds.</li></ul>	

PROJECTS

<b>Software and User Groupification using Clustering   Honeywell</b>	<b>Aug 2024 – Dec 2024</b>
<ul style="list-style-type: none"><li>Processed and cleaned over 15 million data points using Python, improving data quality by 93%.</li><li>Applying clustering techniques (K-Means, Hierarchical Clustering) to form groups and find different patterns and trends.</li><li>Generated results using Tableau to find similar patterns which provide in-depth insights on every cluster.</li><li>Collaborated with cross-functional teams to investigate root causes of defects, review and qualify objective data.</li></ul>	
<b>Building a Predictive and Summarizing Model</b>	<b>Aug 2023- Dec 2023</b>
<ul style="list-style-type: none"><li>Processed datasets for 50+ high schools in Maricopa County, improving accuracy in predictions by 10%.</li><li>Employed advanced data cleansing for Exploratory Data Analysis (EDA), including duplication, inconsistency resolution, and error correction, utilizing algorithms for validation, standardization, and normalization.</li><li>Produced a refined dataset, optimized for EDA, and transformed into a summarization model for enhanced insights.</li></ul>	
<b>Data Exploration into Movie Dynamics</b>	<b>Aug 2023 – Dec 2023</b>
<ul style="list-style-type: none"><li>Generated insights from 100,000+ movie records, designing dynamic visualizations that resulted in a 15% increase in engagement.</li><li>Incorporating Python, SQL, and various data visualization tools to generate dynamic visualizations and predictive models.</li><li>Executed comprehensive exploratory data analysis (EDA) using statistical techniques like correlation analysis.</li></ul>	
<b>Sign Language to Speech Recognition using Data Analytics and Machine Learning</b>	<b>Nov 2021 – March 2022</b>
<ul style="list-style-type: none"><li>Utilized TensorFlow and scikit-learn along with architectures like CNNs to build machine learning models for sign recognition.</li><li>Trained the model using Jupyter notebooks in the Anaconda environment and evaluated the system's performance using metrics like accuracy, precision, recall, and F1-score.</li></ul>	
<b>Emotion Classification Using Machine Learning and Data Analytics</b>	<b>May 2021 – Aug 2021</b>
<ul style="list-style-type: none"><li>Effectively utilized CNN for training and testing, proposing a model using OpenCV with 98 %accuracy.</li><li>Leveraged VGG16, Keras, Python, and Jupyter Notebook for classification and implemented it through a UML diagram.</li><li>Achieved 98% model accuracy using CNN for emotion detection, improving recognition speed by 20%.</li></ul>	

EDUCATION

<b>Masters in Information Technology</b>   Arizona State University, Tempe, Arizona   GPA: 4	<b>Dec 2024</b>
<b>Bachelors in Computer Science and Engineering</b>   Vidya Jyothi Institute of Technology, Hyderabad   GPA: 3.8	<b>Aug 2022</b>

ACHIEVEMENTS

<b>Google Project Management: “Foundations of Project Management” by Coursera</b>	<b>Aug 2024</b>
<b>Programming Essentials in Python, CISCO</b>	<b>April 2020</b>
<b>Programming Essentials in Networking” by CISCO</b>	<b>Dec 2021</b>