# HARIKRISHNA REDDY | DATA ENGINEER

TX • (479) 345-0743 • harikrishnareddy0408@gmail.com
Linkedin: https://www.linkedin.com/in/harikrishnareddy9

## SUMMARY

- Experienced Data Engineer with a Master's degree in Computer Science and 4+ years of experience in data driven solutions.
- Designed and deployed a highly scalable data processing system that processed over a million records per day on Azure cloud.
- Extensive experience in applying MLOps practices to ensure the scalability and maintainability of machine learning models.
- Proficient in multiple programming languages, including Python, SQL, Java, and JavaScript.
- Skilled in extracting data from BigQuery into PySpark DataFrames for performing complex data manipulations and transformations.
- Proficient in building and maintaining RESTful APIs using Python frameworks including FASTAPI and Flask
- Achieved a 50% reduction in processing time and a 30% improvement in data quality for the data processing system.
- Familiar with Agile/Scrum software development methodologies and database/SQL
- Skilled in analyzing data and deriving insights to make informed engineering decisions.
- Implemented data analysis and security processes resulting in 30% increased efficiency.
- Knowledge of Dataflow's scalable and distributed processing capabilities for handling large volumes of data.
- Automated data-related tasks through custom scripts and programs, reducing data processing time by 40%.
- Experienced in utilizing data analytics tools such as Tableau for data visualization and analysis.
- Skilled in Python and Pyspark programming languages, with expertise in data processing, analysis, and visualization.
- Experienced in ETL development, data modeling, and data warehousing using Pyspark, PySpark SQL, and PySpark DataFrame.
- Expertise in statistical modeling, predictive analytics, and data visualization, with proven success in driving business decisions in supply chain and operations through insightful dashboards and robust data-driven strategies.
- Experienced in object-relational mapping (ORM) technologies, including SQLAlchemy and Django ORM, enhancing data manipulation efficiency and database management.
- Proficiency in designing and developing data processing workflows in Dataflow using Apache Beam SDK and programming languages like Python or Java.
- Strong interest in learning new data science tools and techniques to enhance problem-solving capabilities.
- Strong knowledge in cloud storage and retrieval using services such as Amazon S3 and Azure Blob Storage, as well as inmemory data processing with Apache Spark and Apache Ignite.
- Skilled in data modeling, data warehousing, data integration, and SQL scripting.
- Proficient in integrating BigQuery with other GCP services like Cloud Storage, Dataflow, and Data Studio for end-to-end data analytics pipelines.
- Excellent SQL programming skills and developed Stored Procedures, Triggers, Functions, Packages using SQL, PL/SQL.
- Deep knowledge of Hadoop ecosystem tools such as HDFS, MapReduce, Spark, Yarn, and Sqoop for designing and developing Big Data Analytics solutions.
- Experienced in designing and implementing data lakehouse solutions that allow for the storage, management, and processing of structured and unstructured data.
- Developed automation scripts and programs to streamline data-related tasks, reducing processing time and enhancing efficiency within the data lakehouse.
- A quick learner with strong attention to detail and effective communication skills, capable of solving complex problems.

## EDUCATION

### Master of Science in Computer Science | Aug 2021 - Dec 2022

University of Houston Clear Lake, TX, USA

### Bachelor of Science in Computer Science | Aug 2015 - July 2019

Sarvepalli Radhakrishnan University, Bhopal, India

## SKILLS

| | |
|---|---|
| Programming languages: | Python, Java, Scala, R, SQL, JavaScript |
| Data processing tools: | Hadoop, HDFS, MapReduce, Hive, Spark, Oozie, Sqoop, Flume, Pig |
| Data & Analytics Techniques: | Document analysis, Data Profiling, Data Integration, Data Modeling, Data Mining, Data Visualization, Market analysis, GAP analysis, SWOT analysis, MOST analysis, Cost-benefit analysis, Feasibility Analysis, Big Data Analytics, Data Analysis, Data Migration |
| Databases & Cloud: | Cassandra, HBase, Relational databases (MySQL), AWS (S3, Ec2, lambda, Apache airflow, Cloud Watch, IAM, Cloud Formation), Azure (Data Factory, Databricks, Stream Analytics), GCP (BigQuery, Dataflow, Cloud Storage, Pub/Sub, Dataproc, Data Studio, Composer) |
| Data visualization and reporting: | Tableau, Talend |
| Other Tools: | Jenkins, Maven, Git/Bitbucket, Unix/Linux platforms, Cloudera, Hortonworks, MapR, EMR, ETL |
| Environment: | Agile, SDLC, CI/CD |

## PROFESSIONAL EXPERIENCE

AdventHealth Georgia, US

Role: Data Analyst Nov 2023 - Present

- Developed decision support tools for financial reporting, optimizing revenue cycle management.
- Utilized SQL, Python, and Tableau to create reports and databases, aiding strategic decision-making across departments.
- Managed end-to-end data science workflows from data gathering to model deployment
- Innovated in report development and ad hoc analyses, improving business strategies with data-driven insights.
- Maintained data integrity across large datasets, ensuring accuracy and reliability in reporting.
- Served as a technical expert on software applications, enhancing departmental data analysis capabilities.
- Completed special projects on deadline, contributing to departmental goals and operational efficiencies.
- Automated reporting processes and implemented data quality checks, streamlining data availability for decision-makers.
- Facilitated cross-departmental collaboration to translate data needs into actionable insights and tailored reports.
- Utilized advanced statistical techniques and machine learning models to analyze large datasets, identifying key patterns that informed business strategy and operational improvements in healthcare logistics.

Intelligenie LLC, US

Role: Data Engineer Feb 2023 – Nov 2023

- Optimized PySpark jobs for efficient data processing and parallel execution on BigQuery datasets, utilized techniques like partitioning and caching.
- Performing statistical data analysis and data visualization using Python and Imported the claims data into Python using Pandas libraries and performed various data analysis
- Written PySpark code to transform and manipulate data within
- Developed and productionized machine learning models that significantly improved operational efficiencies, employing Python and MLOps techniques to ensure optimal performance and scalability.
- Utilized Google BigQuery for high-speed SQL-like querying and analysis of large datasets, optimizing data retrieval and reporting. Dataflow pipelines, leveraging Spark's data processing capabilities.
- Developing containment scripts for data reconciliation using SQL and Python
- Designed and implemented robust data cleansing and transformation processes to prepare both internal and external datasets for complex analytics, enhancing model accuracy and insights.
- Utilized Apache Airflow to automate and schedule complex data pipelines, improving operational efficiency and ensuring on-time data processing.
- Developed a RESTful service using FASTAPI to streamline data transactions, achieving a 20% improvement in request handling efficiency.
- Implemented data orchestration tools to integrate data from diverse sources, ensuring data consistency and reliability. Automated workflows for data quality checks and error handling.

- Presented Dashboards to Higher Management for more Insights using Tableau
- Utilized MySQL and MS Excel to review and analyze system data to identify trends
- Worked with internal client team to structure, analyze and interpret data requests
- Collaborating with Senior Data Scientists for understanding of data
- Prepared and presented Business Requirement Document (BRD), System Requirement Specification (SRS) and Functional Requirement Document (FRD)
- Utilized SQLAlchemy ORM for database interactions in a data integration project, improving data retrieval processes and system stability.
- Converted charts into Crosstabs for further underlying data analysis in MS Excel
- Developed and completed action plans, implemented control panel structure, conducted audits and provided technical solutions for a diverse array of issues

Halix Tech Solutions, India

Role: Data Engineer Jan 2018 – Jun 2021

- Improve efficiency and accuracy by evaluating model in Python and R
- Extensive experience in working with AWS cloud Platform (S3, Ec2, lambda, Apache airflow, Cloud Watch, IAM, Fsx, and Cloud Formation).
- Created and analyzed business requirements to compose functional and implementable technical data solutions.
- Identified integration impact, data flows and data stewardship.
- Developed interactive visualizations and reports using Tableau to present data-driven insights to stakeholders.
- Generated, wrote and run SQL script to implement the DB changes including table update, addition or update of indexes, creation of views and store procedures.
- Reviewed and revised data models for soundness of data structures and adherence to client standards.
- Performed Extraction, Transformation and Loading (ETL) using Informatica power center.
- Define the ETL mapping specification and Design the ETL process to source the data from sources and load it into DWH tables.
- Developed and maintained advanced predictive models that significantly enhanced decision-making processes, focusing on high performance, reliability, and maintainability in a production environment.
- Implemented secure and scalable APIs using FASTAPI framework to facilitate real-time data processing and enhance application responsiveness.
- Managed AWS cloud infrastructure to deploy and maintain scalable machine learning models, utilizing services such as EC2, Lambda, Sagemaker, and S3 for data storage, processing, and retrieval, aligning with best practices in software engineering and MLOps
- Creating complex SQL queries and scripts to extract and aggregate data to validate the accuracy of the data and Business requirement gathering and translating them into clear and concise specifications and queries.
- Leveraged Azure Data Factory to orchestrate and automate data pipelines, facilitating efficient data integration, transformation, and loading processes.
- Managed and optimized data warehousing using Azure SQL Data Warehouse to ensure high-performance querying and reporting capabilities.
- Conducted thorough data cleansing and transformation tasks to prepare large, complex datasets for analysis, ensuring data quality and suitability for machine learning applications.
- Ensured data accuracy and consistency in Tableau visualizations by performing data validation and verification.
- Analyze and Prepare data, identify the patterns on dataset by applying historical models
- Wrote MySQL queries from scratch and created views on MySQL for Tableau

## PROJECTS

Machine Learning Project

Supply Chain Logistics Optimization Using Predictive Analytics

- Completed as part of a Machine Learning course, this project focused on applying predictive analytics to optimize supply chain logistics for a simulated logistics company.
- Developed forecasting models using Python to predict demand and inventory levels, demonstrating a potential 25% reduction in operational costs through enhanced decision-making.

- Leveraged Apache Spark for handling large-scale data simulations and integrated with AWS cloud services like S3 and Redshift to showcase the application of scalable analytics in supply chain management.
- Created visualizations and dynamic dashboards in Tableau to effectively communicate the impacts of predictive analytics on supply chain operations to academic peers and instructors.

Big Data Analytics Project
Similar Hospital Identification using Locality Sensitive Hashing (LSH)

- As a Team Lead researched backend, user experience and communicated project's vision and objectives
- Implemented a solution to identify similar hospitals in terms of their impact of COVID-19 using Spark and Python.
- Extracted binary features of hospitals and converted each row to an array representation stored on Spark RDD.
- Utilized MinHash algorithm to convert the set representation of each hospital into 100 dimensions.
- Calculated the Jaccard distance for each hospital with all other hospitals and used random hash functions to divide the signature matrix into bands.
- Applied Locality Sensitive Hashing (LSH) to determine similar hospitals based on the computed Jaccard distance.

Big Data Analytics Project
Web Scraping and Analysis of Solar Flares using Python

- Scraped data of top 50 solar flares from a NASA website using python, beautiful-soup,requests, pandas, and numpy.
- Analyzed the data to extract valuable insights and information.

Human Performance Assessment and Prediction in Collaborative Learning Environments

- Human Performance Assessment and Prediction in Collaborative Learning Environments
- Developed a framework for efficient assessment and prediction of human performance in Collaborative Learning Environments (CLE) using Machine Learning.
- Integrated techniques from Computational Psychometrics (CP) and Deep Learning models, including Convolution Neural Networks (CNNs) for feature extraction, skill identification, and pattern recognition.