Bala Venu Valeti

Data Engineer

TX | (469) 922-3927 | balavenuvaleti@gmail.com

Summary

- Data Engineer with 5+ years of expertise in designing, developing, and optimizing end-to-end data pipelines within the realms of healthcare and financial services.
- Proficient in extracting, transforming, and loading (ETL) large volumes of data from various sources such as HDFS, MongoDB, and AWS S3 using tools like Apache Nifi, PySpark, and AWS Glue.
- Skilled in utilizing cutting-edge technologies like Hadoop, MapReduce, and Apache Spark for distributed computing and big data processing, ensuring scalability and performance.
- Adept at data visualization and reporting using tools like Tableau and Power BI, enabling stakeholders to derive meaningful insights from complex datasets.
- Proficient in database management systems including SQL Server, MySQL, PostgreSQL, and MongoDB, with a strong emphasis on data modeling, normalization, and optimization.
- Strong soft skills encompassing time management, leadership, problem-solving, negotiation, decision-making, and effective documentation and presentation
- Proficient in collecting, organizing, and analyzing large volumes of banking data.
- Developed and maintained data pipelines to ensure timely and accurate data extraction, transformation, and loading (ETL) processes.
- Implemented data quality checks and validation routines to ensure the integrity of banking data.
- Collaborated with cross-functional teams, including data scientists, business analysts, and IT professionals, to understand data requirements and deliver data solutions.

Technical Skills

Methodologies	SDLC, Agile, Waterfall
Language	Python, SQL, R
Packages	NumPy, Pandas, Matplotlib, SciPy, Scikit-learn, TensorFlow, Seaborn
Visualization Tools	Tableau, Power BI, Advanced Excel
IDEs	Visual Studio Code, PyCharm, Jupyter Notebook
Database Management	MySQL, PostgreSQL, MongoDB, SQL Server, Oracle
Big Data Technologies	Hadoop, MapReduce, HDFS, Sqoop, Hive, NIFI, Kafka, Apache Spark
Cloud Technologies	Amazon Web Services (AWS), GCP
Machine Learning Algorithms	Supervised Learning (Linear Regression, Logistic Regression, Decision Tree, Random Forest, SVM, Classification), Unsupervised Learning (Clustering, KNN, Factor Analysis, PCA)
Other Technical Skills	Data Management, Marketing Analytics, Information Technology Strategy, Jira, Digital Innovation, ETL/ELT Process Innovation & Management, SSIS, SSRS, SSAS, Kubernetes, Snowflake, Informatica, Data Warehouse, Data Architecture Design, Data Security, Data Management, Data Modeling, Data Integration, Git, GitHub, SVN
Operating System	Windows, Linux, Mac OS

Education

Master of Science in Business Analytics | The University of Texas at Dallas, USA
Bachelor of Technology in Computer Science and Engineering | Sreenidhi Institute of Science and Technology, India

Certification

AWS Associate Cloud Practitioner
Incorta 4 essential for Business and Analysts

Experience

American Express, TX | Aug 2022 – Present (Data Engineer)

- Extracted data from HDFS, including customer behavior, sales and revenue data.
- Transferred the data to AWS S3 using Apache Nifi, which is an open-source data integration tool that enables powerful and scalable dataflows.
- Used PySpark to process and transform the data, which is a distributed computing framework for big data processing with Python API.
- Loaded the transformed data into AWS RedShift data warehousing to analyze the data and validated and cleaned the data using Python scripts before storing it in S3.
- Scheduled the pipeline using Apache Oozie, which is a workflow scheduler system to manage Apache Hadoop jobs.
- Developed visualizations and dashboards using Tableau for reporting and business intelligence purposes.
- Analyzed the data in HDFS using Apache Hive, which is a data warehouse software that facilitates querying and managing large datasets.
- Used GitHub as a version control system for managing code changes.
- Developed data visualization dashboards using Tableau and Power BI, providing stakeholders with actionable insights for making informed business decisions.
- Collaborated with data scientists to integrate machine learning models into data pipelines, enabling predictive analytics and fraud detection capabilities.

NextGen Healthcare, India | March 2018 – July 2021 (Data Engineer)

- Developed and implemented data warehouse, data processing algorithms, ensuring optimal data delivery architecture is consistent throughout Pharmacy Benefits Management (PBM) services projects following Agile framework.
- Designed, built, and maintained efficient, reusable, and reliable data pipelines using pySpark, Hadoop and other new technologies.
- Utilized Parquet file and global tables for efficient storage and performance, working with MongoDB NoSQL databases.
- Supported existing Snowflake and AWS cloud Data Management implementations, data ingestion – transformation, addressing performance optimization, and implement enhancements as in building tables/columns, Historical views, Raw Data Views, Materialized views.
- Participated in the code review process, ensuring code quality using Pytest and sharing knowledge with the team.
- Analyzed transformed data to draw insights and develop strategies in Power BI.
- Extracted real-time data integration with Kafka and Spark Streaming, converting them to RDDs and processing data as Data Frames in HDFS.
- Documented data model, including normalization /de normalize entities, metadata, and constraints, essential for future reference and maintaining data governance.
- Worked with Python NumPy, SciPy, Pandas, Matplot, Stats packages to perform dataset manipulation, data mapping, data cleansing and feature engineering. Built and analyzed datasets using R and Python.