**VENKATESH BATTULA**  
United States | Email: Venkatesh.battula11@gmail.com | Tel: +1 (216) 370-8154 | [LinkedIn](https://www.linkedin.com/in/venkateshbattula) | [GitHub](https://github.com/venkyvenkatesh10)

**EDUCATION**Cleveland State University, Cleveland, OH  
Master of Science in Computer Science and Information Systems - 2022

Jawaharlal Nehru Technological University, Hyderabad, India  
Bachelor of Technology in Computer Science - 2019

**TECHNICAL SKILLS**

* Big Data Technologies: Hadoop, HDFS, Hive,MapReduce 2.0, Hive, YARN, Pig, Sqoop, Presto, PySpark, Scala.
* Programming Languages: Python, Java, Scala, C, SQL.
* Cloud Technologies:
  + AWS: S3, Lambda, Glue, CloudWatch, SNS, SQS, CloudFormation, Redshift.
  + Azure: Blob Storage, Functions, Data Factory, Monitor, Event Grid, Queue Storage, Resource Manager, Synapse Analytics.
  + GCP: Cloud Storage, Cloud Functions, DataFlow/DataPrep, Cloud Monitor, Pub/Sub, Deployment Manager, BigQuery, Dataproc, GCS.
* Script Languages: Python, PySpark, SQL, Scala, HTML, CSS.
* Databases: MS SQL Server, DBeaver, PostgreSQL, Oracle, MySQL, Kudu.
* IDEs and Tools: IntelliJ, PyCharm, Jupiter Notebook, Visual Studio.
* BI Tools: Tableau, PowerBI, Google Looker.

**PROFESSIONAL EXPERIENCE**

**Data Engineer | Verizon (Contract) | Irving, TX  
*Nov 2023 – Present***

* Designed and implemented efficient ETL pipelines using Apache NiFi to ingest raw data from multiple data centers into a centralized HDFS cluster.
* Utilized Apache Spark for large-scale data processing and transformation, optimizing Spark jobs for performance and scalability.
* Managed and maintained HDFS clusters ensuring optimal data storage and retrieval performance.
* Developed data quality checks to validate and clean incoming data, ensuring high data integrity throughout the pipeline.
* Used Pandas, Numpy, Seaborn, SciPy, Matplotlib, Scikit-learn, and NLTK in Python for developing various machine learning algorithms and utilized machine learning algorithms such as linear regression.
* Worked on proof of concept with Spark, Scala, and Kafka for real-time data streaming processing.
* Design and implement real-time data processing pipelines using Apache NiFi and Spark Streaming to ingest, process, and analyze data streams from Kafka. Ensure the pipelines are optimized for performance, scalability, and fault tolerance, enabling low-latency processing of large-scale data.
* Developed NiFi flows dealing with various data formats such as XML, JSON, and Avro.
* Automated the ETL workflow using cron jobs and integrated with Apache Airflow for complex scheduling requirements.
* Leverage HBase's key-value storage capabilities to efficiently store and retrieve large volumes of structured and semi-structured data.
* Work on designing and maintaining HBase tables with appropriate row keys and column families, ensuring fast access to data for analytical queries and real-time applications.
* Transformed processed data into meaningful visualizations using Google Looker Studio enabling data-driven decision-making.
* Conducted detailed data analysis to extract key business insights supporting strategic decision-making processes.
* Worked with NoSQL databases, particularly Cassandra, for efficient storage and retrieval of large datasets.
* Implemented CI/CD pipelines using Jenkins and Apache Airflow to streamline the deployment process and ensure continuous integration and delivery.
* Collaborated with cross-functional teams to understand data requirements and deliver accurate solutions.

**Data Engineer | Fidelity Investments (Contract) | South Lake, TX  
*Jan 2022 – Oct 2023***

* Designed and developed data pipelines to ingest raw data from Oracle databases into AWS S3 ensuring secure and efficient data transfer.
* Configured AWS Glue crawlers for automatic data detection and cleansing, maintaining high data quality and consistency.
* Experience in Amazon AWS services such as EMR, EC2, S3, CloudFormation, Redshift, Lambda, and DynamoDB, which provide fast and efficient processing of Big Data.
* Experience with API Gateway publishing using AWS API Gateway.
* Implemented ETL scripts in AWS Glue to process and transform data into hourly buckets in S3, optimizing storage and retrieval efficiency.
* Conducted further transformations to enrich datasets with additional business logic and aggregations, enhancing data value for analysis.
* Established data pipelines to load enriched data from S3 into Amazon Redshift, facilitating scalable and high-performance data warehousing.
* Tuned Redshift cluster performance to ensure efficient query execution and data retrieval, supporting business intelligence needs.
* Enabled integration of Redshift with tools like Tableau and Looker, providing real-time access to insights and analytics.
* Developed and maintained dashboards and reports, ensuring accurate and timely business intelligence outputs.
* Implemented AWS CloudWatch for comprehensive monitoring of the data pipeline, promptly addressing performance and operational issues.
* Built internal tools using Django and Flask, enhancing the capabilities of the data engineering team.
* Developed serverless data processing tasks using AWS Lambda, complementing Glue jobs and enhancing workflow flexibility.
* Ensured data security by implementing robust access controls and encryption, maintaining compliance with industry regulations.
* Worked closely with data scientists, analysts, and stakeholders to understand data requirements and deliver tailored solutions.
* Maintained thorough documentation of data pipeline processes, ETL scripts, and configurations for knowledge sharing and onboarding.
* Regularly optimized data processing workflows, leveraging new technologies and best practices to improve efficiency and performance.

**Data Engineer/Data Analyst | Defense Research & Development Laboratory (DRDL)  
*Dec 2018 – Nov 2020***

* Developed and optimized analysis software for performance analysis of Command-and-Control Systems.
* Calculated and visualized various trajectories using GNU plotting, displaying time and aircraft variations.
* Collected data from various sources, including databases, files, APIs, and external data sources.
* Employed SQL queries and ETL tools to efficiently retrieve the required data.
* Applied transformations to cleanse, enrich, and structure the data, preparing it for integration into a data warehouse.
* Used ETL tools like Talend, Informatica, and AWS Glue for transformations including data type conversions, validation, and cleansing.
* Loaded the transformed data into designated data warehouses or storage systems.
* Implemented quality checks to ensure the loaded data meets predefined standards, promptly addressing any quality issues.
* Worked closely with data analysts, data engineers, and business stakeholders to comprehend data requirements and align with the business's objectives.

**CERTIFICATIONS**

* **AWS Certified Developer** - Associate  
  Validation Number: Z1FP1X2KX2RQQJ37  
  Validate at: [AWS Verification](https://aws.amazon.com/verification)
* **Google Cloud Certified Associate Engineer**Validation Number: [Google Accredible](https://google.accredible.com/d11a4423-c645-48bf-9925-e69e0ac42654)  
  Validate at: [Google Accredible](https://google.accredible.com/d11a4423-c645-48bf-9925-e69e0ac42654)