

Vishesh Paka

Clearwater, FL |  |  | [GitHub](#)

EDUCATION

Masters in Data Analytics Engineering (GPA: 4.0/4.0), Northeastern University, Boston, MA **May 2024**

Courses: Database Management & Design, Data Warehousing, Data Mining, Cloud Computing (AWS)

Bachelors in Computer Science (GPA: 3.82/4.0), VNR VJIET, Hyderabad, India **Aug 2021**

Courses: Data Structures & Algorithms, Probability Statistics, Big Data & Predictive Analytics, Artificial Intelligence, NLP

TECHNICAL SKILLS

Programming Languages:	Python (Pandas, NumPy, Scikit-learn, SciPy, TensorFlow, PyTorch), R, C, C++, Java
Databases:	SQL (MySQL, Oracle, PostgreSQL), NoSQL (MongoDB), AWS (RDS, Redshift, DynamoDB)
ETL & Data Visualization Tools:	Informatica, SSIS, Snowflake, DBT, Tableau, Power BI, Alteryx, AWS QuickSight, Excel
Cloud & Big Data Technologies:	Apache Spark, Kafka, Airflow, Hive, Hadoop, Flink, AWS (S3, Glue, EMR, Lambda, Athena) Azure (Databricks, Data Factory, Synapse Analytics, ADLS, Delta Lake)
Tools & Specializations:	Git, Jira, Confluence, CI/CD pipelines, Kubernetes, Docker, Hypothesis Testing, A/B testing

CERTIFICATIONS

- AWS Certified Data Engineer Associate ([Link](#)), Certified in Machine Learning from STANFORD UNIVERSITY ([Link](#))
- Certified in Python for Data Science and AI & ML from IBM ([Link](#))

PROFESSIONAL EXPERIENCE

Accenture, Bangalore, India **Sep 2021- Sep 2022**

Data Engineer

- Engineered end-to-end ETL data pipelines using Informatica for extracting, transforming, and loading diverse datasets (CSV, JSON, flat files, Parquet) into a Snowflake data warehouse, resulting in a 40% reduction in data processing time
- Implemented complex data mapping and transformation logic to standardize data across multiple sources, reducing data inconsistency by 30% and improving overall data integrity and processing efficiency
- Architected an Enterprise Data Lake on AWS S3 integrated with Glue and Athena, significantly improving data accessibility, query performance, and supporting use cases in analytics, processing, storage, and reporting of high-volume data
- Collaborated with 5+ cross-functional teams (Data Science, Analytics, and IT) to design and deploy scalable data solutions, increasing data integration efficiency by 25% and reducing integration errors by 20%
- Implemented Slowly Changing Dimension (SCD) Type 2 tables in Snowflake, enabling historical data tracking and improving data versioning accuracy by 40%
- Designed and optimized 5 interactive Tableau dashboards, empowering stakeholders with real-time analytics, improving data-driven decision-making and boosting adoption by 15%

Cognizant, Hyderabad, India

Jan 2021 - Aug 2021

Analytics Engineer

- Designed and developed SQL based ETL workflows to support data warehousing and analytics, improving data accessibility by 30% through optimized queries and indexing strategies
- Automated data pipelines with Python and SQL, ensuring 99% data accuracy and streamlining data validation processes
- Implemented Change Data Capture (CDC) mechanisms using SQL and AWS Glue to efficiently track and process incremental data updates, reducing redundant processing by 50% and optimizing ETL performance
- Developed and deployed real-time Tableau dashboards, increasing reporting efficiency by 15%, and empowering business users with faster insights into operational data

PROJECTS

AWS-Powered YouTube Data Pipeline and Analytics **Aug 2024**

- Engineered a cloud-based ETL pipeline using AWS Glue, Lambda, and Athena to ingest and process structured and semi-structured YouTube video data, reducing data processing time by 40% and enabling scalable storage in the S3 data lake
- Crafted interactive dashboards using AWS QuickSight, leveraging PySpark partitioned data by region and category to visualize key video trends, enhancing reporting capabilities and improving data-driven insights by 30% for stakeholders

Streamlining Stock Market Data: A Real-Time Kafka and AWS Glue Pipeline **Jun 2024**

- Orchestrated a data pipeline using Kafka for seamless ingestion and streaming into AWS S3, reducing data latency by 25%, while utilizing Kafka Producer for dynamic data generation and Kafka Consumer for efficient data consumption
- Automated schema creation with AWS Glue Crawler and Data Catalog, and leveraged Athena for optimized querying, achieving a 40% boost in analytics performance and enabling scalable data processing

Store Management System **Nov 2022**

- Engaged in data modeling by designing an Entity-Relationship (ER) diagram to capture 10 key entities, 12 relationships, and over 50 attributes in the system and implemented a relational database using MySQL Workbench
- Optimized the database design by implementing a Star Schema, enhancing query performance by 40% and improving data retrieval speed for analytical processing. This streamlined approach boosted reporting efficiency and system scalability

Analysis on Global Carbon Emissions **Oct 2022**

- Designed a dynamic Google website, showcasing interactive visualizations and analysis of [global carbon emissions](#) data leveraging Tableau, Flourish, and Datawrapper, to create a visually striking and information-rich website