Vivek Sai Chinna B

Data Engineer II

✉ viveksaichinnaburada@gmail.com — ☎ +1 (910) 885-8585 — 📍 Charlotte, NC — in viveksaichinna — viveksaichinna

SUMMARY

- 3+ years of experience in Data Engineering, Data Pipeline Design, Development and Implementation as a Data Engineer/Data Developer and Data Modeler.
- Experience in data extraction, transformation, and loading (ETL) from diverse sources into target systems using AWS services (S3, Redshift, RDS, Glue) and relational databases such as SQL Server, PostgreSQL, Oracle and applying data modeling techniques, data architecture principles, statistical methods, and delivering end-to-end data solutions.
- Proficient in Big Data technologies and the Hadoop ecosystem, with hands-on experience in HDFS, Spark (PySpark, Scala), Hive, MapReduce, Kafka, and NoSQL databases. Skilled in Python, Pytest, MySQL, Git, and AWS, with a strong foundation in Agile practices for delivering scalable, data-driven solutions Business Intelligence (BI).
- Monitor Resources and Applications using AWS Cloud Watch, including creating alarms to monitor metrics such as EBS, EC2, ELB, RDS, S3, SNS and configured notifications for the alarms generated based on events defined.
- Proficient in developing data analysis and transformation scripts using PySpark and Spark APIs in Python. Also, debugging and troubleshooting data pipelines across cloud environments.
- Strong experience in Software Development Life Cycle (SDLC) including Requirements Analysis, Design Specification and Testing as per Cycle in both Waterfall and Agile methodologies.

EDUCATION

University Of North Carolina At Charlotte — **Charlotte, NC**
Master's in Computer Science — ***August 2023 - May 2025***

PES University — **Bengaluru, India**
Bachelor's in Computer Science Engineering — ***August 2019 – May 2023***

WORK EXPERIENCE

State Farm — **Chicago, IL**
Data Engineer II — **December 2023 – Present**

- Orchestrated the deployment, automation, and maintenance of AWS cloud-based production systems using Apache Airflow, AWS Step Functions, and CloudFormation, ensuring availability, performance, and scalability.
- Established Infrastructure as Code (IaaS) using CloudFormation to provision AWS services, applying monthly critical patches via AWS Patch Manager. Implemented IAM roles and policies to ensure secure access pipeline.
- Led the migration of sensitive financial data from on-premises SQL Server to Amazon Redshift data warehouse via S3 and Lake Formation, enabling centralized, secure, and governed access to analytics-ready datasets.
- Designed and implemented scalable ETL pipelines using PySpark in AWS Glue to perform data cleansing, transformation, and feature engineering in collaboration with data science teams. Utilized AWS Glue Crawlers and the Glue Data Catalog for schema inference and metadata management.
- Built real-time ingestion pipelines from Kinesis/MSK to S3, triggering Glue jobs using Lambda and Step Functions, automating data processing and improving latency by 20%.
- Processed and transformed large Parquet datasets using PySpark and Spark SQL, creating Hive-compatible tables and leveraging partitioning, broadcast joins, and in-memory optimization techniques.
- Queried cleaned and enriched datasets via Amazon Athena for ad-hoc analysis and validation.
- Improved Redshift performance by 3x through schema optimization using a star schema. Integrated Glue Data Quality for validation rules and AWS CloudWatch for monitoring and alerting on data pipeline failures and delays.

Evoke Technologies — **Hyderabad, India**
Data Engineer — **December 2021 to July 2023**

- Actively engaged with key components within the Hadoop Ecosystem including Spark, HDFS, HIVE, HBase, Zookeeper, Sqoop, and Oozie.
- Developed Sqoop jobs to seamlessly ingest data from diverse systems of records into the Enterprise Data Lake.
- Created Spark jobs in PySpark and SparkSQL to operate on Hive tables, generating transformed datasets for downstream utilization.
- Installed and configured Hadoop MapReduce, HDFS, and created multiple MapReduce jobs in Java and Scala for data cleaning and preprocessing.
- Designed ETL jobs using Spark-Scala to migrate data from Oracle to new Cassandra tables.
- Leveraged Spark-Scala (RDDs, DataFrames, Spark SQL) and Spark-Cassandra for tasks such as data migration and business report generation.
- Employed Data Build Tool for transformations in the ETL process, alongside AWS Lambda and SQS. Extensively worked with AWS services such as EC2, S3, VPC, Appflow, ELB, Auto Scaling Groups, Route 53, IAM, CloudTrail, CloudWatch, CloudFormation, CloudFront, SNS, and RDS.
- Integrated CI/CD pipelines using Jenkins for automated deployment of Spark and ETL jobs.

PROJECTS

Real-time CAPTCHA using hand gesture recognition for highly secure websites

- **Problem Statement:** Standard CAPTCHA systems are up to 85% vulnerable to automated attacks, risking sensitive access. A more secure, human-verifiable method was needed to protect high-risk web platforms.
- **Model Development:** Engineered a dual-layer CAPTCHA system combining Text-CAPTCHA with real-time hand gesture recognition. Trained gesture models using 100,000+ samples; improved accuracy from 82% (SVM) to 96% using CNN and MediaPipe for sub-100ms real-time processing.
- **Outcome:** Delivered a "bot-proof" authentication mechanism with 99.9% resistance to automated attacks, reducing false acceptance rate by over 90% compared to traditional CAPTCHA methods.

Air Quality Monitoring & Forecasting with Real-Time Data Pipeline & Machine Learning

- **Problem Statement:** Poor air quality causes over 7 million premature deaths annually. Accurate, real-time forecasting and trend detection are essential for public health intervention and alerting systems.
- **Data Processing:** Built a 5-stage big data pipeline using Apache Spark to process 100K+ real-time records with TCP ingestion, advanced SQL analytics, and predictive ML models (RMSE = 5.2, $R^2 = 0.85$). Integrated Spark MLlib and UDF-based AQI classification to forecast pollution levels and categorize regions.
- **Outcome:** Delivered an end-to-end solution with lesser than 100ms streaming latency, real-time alerting, and an interactive dashboard (via Plotly/Grafana), enabling 95% accurate AQI forecasting and data-driven environmental decision-making.

Word-Based Sentiment Scoring & Trend Analysis on Historical Texts (Hadoop & Hive)

- **Problem Statement:**Understanding socio-cultural sentiment shifts in the 18th–19th centuries requires analyzing unstructured historical text at scale using distributed systems and NLP techniques.
- **Data Processing:** Built a 5-stage NLP and big data pipeline using Java-based Hadoop MapReduce and Hive on a dataset of 50+ historical texts (greater than 1M words). Implemented stopword removal, lemmatization, sentiment scoring (AFINN), and a custom Hive UDF for bigram analysis.
- **Outcome:** Achieved decade-level sentiment trend mapping with 90% accuracy in polarity classification. Optimized MapReduce jobs with combiners and counters, reducing job runtime by 30%. Produced a searchable Hive database for linguistic and sentiment trends.

Chatbot for Support Ticket Automation (AWS Lex)

- **Problem Statement:**Clients frequently raised recurring issues, resulting in high ticket volumes and repetitive manual responses from support teams. A scalable solution was needed to auto-resolve repeat issues and triage new ones efficiently.
- **Data Processing:** Developed a support ticket chatbot using Amazon Lex, leveraging historical issue logs for pattern recognition. Integrated Lex with backend databases to identify and auto-respond to previously seen issues, while escalating new or complex tickets to relevant support teams.
- **Outcome:** Achieved a 40% reduction in resolution time and significantly reduced manual workload, enhancing the support process's speed and consistency.

TECHNICAL SKILLS

Programming Languages : Python, Scala, R, Java, JavaScript
Scripting : Python, Shell scripting
Big Data & ETL Tools : Hadoop, Hive, Apache Spark, PySpark, Kafka, YARN, Sqoop, Impala,Apache Airflow, Oozie, Pig, MapReduce, Zookeeper, Flume, Informatica, SSIS, Talend
Big Data Platforms : Databricks, Amazon EMR, Cloudera, Hortonworks
Cloud Services
AWS: EC2, S3, EMR, RDS, Redshift, Glue, Athena, Lambda, CloudFormation, Step Functions, Data Lake
Azure: Synapse Analytics, Data Factory, Data Lake, BLOB Storage, Azure SQL
GCP & Others: Google BigQuery, Snowflake
Infrastructure & DevOps: : Terraform, CloudFormation, Kubernetes, Docker, Jenkins, Apache Maven, SBT, Bitbucket, Git, SVN, AWS CloudWatch, CloudTrail
BI and Data Visualization : Tableau, Power BI
Relational Databases : Oracle, SQL Server, Teradata, MySQL, PostgreSQL, Netezza, PL/SQL
NoSQL Databases : Cassandra, MongoDB, HBase

Certifications, Awards & Publications

- **AWS Certified Data Engineer – Associate**, Amazon Web Services (**Certificate**)
  Issued: December 2024
- **Published Paper:** *Real-time CAPTCHA using hand gesture recognition for highly secure websites* (**Journal**)
  Taylor & Francis Group, May 2023
- **Four-Time Academic Distinction Awardee**, PES University
  Awarded: 2019-2023