**Srikanth D**

**Phone: 571-410-0115**

**Email: Madhu.g@daacsllc.com**

**Lead Data Engineer/Manager**

*Data science and Engineering lead professional, working knowledge in advanced ETL, SQL/NoSQL Databases, cloud infrastructure, Data analysis, solution architecture software development, mining meaningful insights from huge datasets. Proven experience, multiple certifications in Data Engineering, ETL, programming and Machine Learning. Ability to solve business problems, implement ETL/ELT strategies applying advance Data Engineering techniques and programming. Throughout my career, I have worked and met with C-level executives to discuss key corporate initiatives such as corporate mergers, acquisitions ranging from several thousands to million dollars deals.*

**PROFESSIONAL SUMMARY:**

Over **9+ years** of IT experience in all phases of diverse Data Engineering projects specializing in Finance Banking, HealthCare, Technology, Retail & Supply chain, E-commerce domains involving Core Big Data Engineering, handling high-data-volume, design, develop and engineering, SDLC and data warehousing.

Proven expertise in various SDLC stages (Software Development life cycle analysis, Requirements gathering, Designing) with hands on experience to design and build advanced Big Data solution stack involving Hortonworks, Python, NoSQL, Kafka, HiveQL, HBase, Sqoop, TEZ, Spark, SQL, and Hadoop ecosystem.

Expertise in writing/documenting Technical Design Document (TDD), STTM, BRDs, Functional Specification Document (FSD), Test Plans, GAP Analysis in E2E data processing pipeline, reporting tools like Cognos.

Proficient in transforming business requirements into analytical models, providing ETL solutions for any type of business model, designing algorithms, building models, developing data mining and reporting solutions that scale across a massive volume of structured and unstructured data, to develop functional code for it.

Experience functioning in different roles as Senior Data Engineer/Lead, AWS Big Data Engineer, ML Engineer, Python Data Engineer, Developer with over **5 years** of experience in Big Data/Hadoop solutions on cloud with both stream and batch processing, Cloudera ecosystem (HDFS, YARN, Hive, SQOOP, FLUME, HBASE, Oozie, Kafka, Pig), Data pipeline, Data analysis & processing with Hive, SQL, IMPALA, SPARK, PYSPARK.

Worked using different file formats (Avro, Parquet, Text Files, ORC, JSON) and Hadoop performance tuning, Spark query optimization, customization, DQ scripts, triggers, Data security, Compliance and Governance.

Extensive Shell/Python scripting experience for Scheduling and Process Automation, develop bash shell scripts to setup batch jobs on Unix Solaris 10 server, job scheduler scripts for data migration using UNIX. To write Automation scripts using tools like Jenkins, Autosys JILs, Ansible and Terraform.

Worked on Snowflake DW, ETL development using Snowflake bulk loading, Snowpipe, unloading, query performance tuning, time travel, create snowflake schemas, ER modeling, star and snowflake schemas.

Integrated data to/from on-prem database and cloud-based database solutions using Informatica ICS.

Informatica performance tuning like source, target, mapping, fine-tuned transformations, and sessions.

Hands on experience in configuring and usage of AWS S3, EC2, EMR, RDS, Kinesis, Glue. Apache Spark, Pyspark for pipelining, processing and model deployment on cloud. Hands on ETL with Azure, GCP as well.

Strong background and hands on experience of Microsoft SQL server, used complex SQL queries for data integration, manipulation techniques to implement approaches to performance and improvement.

Implemented cluster services using Kubernetes and Docker to manage local deployments in Kubernetes by building self-hosted Kubernetes cluster using Terraform and Ansible and deploying application containers.

Agile, Communicative, Performance-focused, and goal-oriented professional, offering dynamic daily duties and tasks with the ability to support multiple teams, cross channels and deliver data and business solutions.

Worked with business owners and CIO office to develop data monetization strategy with focus on cost reduction, new revenue generation, improving operational efficiencies. Established Enterprise Data Management (EDM), Data Governance COE, Data Analytics COE.

Project/Product Development team in Agile-Scrum/Kanban, to track and resolve blockers for team delivery.

**TOOLS AND TECHNOLOGIES:**

|  |  |
| --- | --- |
| **Big Data Technologies** | Apache Spark and Scala, Hadoop Ecosystem Components – HDFS, Hive, HQL, Sqoop, Apache AirFlow, Splunk, YARN, Impala, Kafka, Flume, Map Reduce, Flink, Pig, Cloudera Hadoop Distribution (3, 4, 5.8.2), Hortonworks |
| **Cloud Platforms** | Microsoft Azure (Data Lake, Data Factory, Databricks, ASQL, Explorer), AWS (EC2, Lambda, Stepfunction, EMR, ELB, S3, Athena, Glue, Redshift), GCP (BigQuery, Dataproc, Composer, Dataflow, Dataprep, Pub/Sub) |
| **Programming/Query Languages** | Scala, SQL, MemSQL, PLSQL, NoSQL, Java, Python, Shell (bash, shell) and R |
| **Databases** | Oracle 10g/9i/8 11g, MySQL, DB2, PostgreSQL, Teradata, Netezza, Snowflake |
| **NoSQL Databases** | MongoDB, HBase, Cassandra, DynamoDB, Redis |
| **Data Visualization** | Tableau, R Shiny, QlikView, AWS Quciksight, Excel, MS PowerBI |
| **Automation and CI/CD tools** | GIT, SVN, Jenkins, Ansible, JFrog Artifactory, Terraform, BitBucket |
| **ETL and Testing** | Nifi, SSIS, Talend, Informatica 10.2/9.6, custom ETL script, SonarQube |

**CERTIFICATIONS:**

* Big Data and Hadoop Essentials - <https://ude.my/UC-O53I89TM>
* Complete Python Bootcamp - [ude.my/UC-ff4c0a38-d656-4003-b3b1-d9f61ea50390](https://ude.my/UC-ff4c0a38-d656-4003-b3b1-d9f61ea50390)
* Master SQL - <https://www.ude.my/UC-a8f6e3ec-d8ee-4c9c-be71-09d492bdf782>
* Complete Data Science Training - <https://ude.my/UC-9eb4b735-ef1c-4c3e-9f23-329fafd46918>
* Advanced Databases - <https://ude.my/UC-XWZEFBRZ>
* Spark and Python for BigData with Pyspark - <https://ude.my/UC-CS0O1160>
* Informatica ETL - <https://ude.my/UC-62fc68a6-fdd8-45ec-9c94-707d9bed5141>
* ETL using SSIS - <https://www.udemy.com/certificate/UC-D07LPEHH/>

**PROFESSIONAL EXPERIENCE:**

**Role: Group Manager in Data Analytics/Lead Data Engineer**

**Emp/Client: Evalueserve, Raleigh, NC Aug 2021 – Present**

**Scope:**  To support all phases of data engineering and code validation using intensively Pyspark, HQL, SQL for many purposes in the Data Analytics unit of internal client Vanguard and Microsoft. Data migration to cloud infra (AWS).

**Responsibilities:**

* Created and maintained optimal data pipeline architecture over AWS infrastructure, assembled large, complex datasets that meet functional/non-functional business requirements to process and transform.
* Built the client Vanguard data migration infrastructure and pipelines for optimal extraction, transformation and loading of data from a wide variety of sources using SQL, Pyspark, Python, Scala and AWS technologies stack. Built analytical tools that utilize this pipeline to provide actionable insights into customer acquisition, operational efficiency and other key business metrics.
* Create continuous data ingestion pipelines using Snowpipe, Snowflake streams and snowflake tasks, to integrate AWS S3 buckets with Snowflake DW using Snowpipe, also to configure CDC logic.
* Responsible of manual ingest process using Pyspark and SQL as main tools, Tableau for dashboards. Successful completion of various business and functional training modules, KT sessions to different teams.
* Responsible to lead and support 2 teams and an offshore team, track progress, providing insights conducting KT sessions and to train the team on the process, help setup dev environment, also overseeing the tasks of less experienced programmers, access issues and system troubleshooting supports.
* Data Visualization: Develop code to create, implement datasets, tables at Vanguard for KPI dashboards to analyze Vanguard investors status using SQL, helping create new views, merging data as needed.
* Data Integration: Ingestion of raw data into AWS S3, transform the same data using Pyspark and further create dataframes, Hive tables and data pipelines using Python, SQL, HQL to maintain and automate whole process to create updated dashboards for Data Governance.
* Integrate Apache Airflow with AWS (MWAA) to monitor multi-stage data pipeline workflows with data processing tasks using Spark, Hive. Create managed workflow, workflow code and monitor the DAGs.
* Data Architecture: Was assigned tasks of new data ingest and data transformation, checking the structure of the new data, how to extract, load and use it.
* Help code and create reporting tables on the ADDI (Microsoft initiative for Alzheimer's disease) and maintain those tables with updated data and visualizations on the ADDI website using Python, Flask.
* Regular code change management for updating the ADDI website dashboards, Tableau viz etc.
* End to end pipelines management: Checkout code, Build App, analyze code, Deploy App (Dev/QA), Test App (Dev/QA), Create release candidate, Deploy App (Prod), complete release.
* Creating, Debugging, Scheduling and Monitoring jobs using Airflow. Experienced with creating DAGs, debugging the DAGs, using most common Operators in Airflow - Python Operator, Bash Operator.
* Making tools, and the environment available for Data scientists to carry their research work over ADDI VMs, was involved in Data science discussions to provide ETL and data engineering solutions.

**Environment:** Spark, CI/CD, MySQL, Python, Hadoop, Hive, Tableau, PowerBI, Docker, AWS, Linux, Shell (bash, shell), cron schedule, CI/CD Cloudbees Jenkins, Job flow Orchestration, Splunk, Airflow, Ubuntu.

**Role: Lead Data Engineer / AWS Big Data Engineer**

**Emp/Client: Capital One Bank, Richmond, VA July 2020 – Aug 2021**

**Scope:** Under the project named ‘Prometheus’ the mission at Capital One is to innovate Core Banking Abstraction layer backed by a Modern Financial Core so the Lines of Business can “cross the canyon” by decoupling surrounds from financial cores and migrating financial account processing to the modern core.

**Responsibilities:**

PFC Architecture idea is Data Acquisition from concerned departments like Billing, Finance hub, Payments Exchange, Auto finance are fed to the Core processing Abstraction layer (CPAL) which is then fed to Adapters and the command processors and then is streamed onto the Purple Kafka (Loopback/Forwarder) and then the shards are shared onto the Event processors where the PFC (Prometheus Financial Core) lies.

Current roles and responsibilities involve helping the team create, maintain and optimizing their data delivery and extraction from multiple data sources into their data warehouse and onto AWS.

Upgrading Docker image to Spring Boot from 2.3 to 2.4, Logging EDS API via Splunk, Create and implement local dev strategy for Event Driven Scheduler (EDS), Create par approved pipeline for EDS API which includes DynamoDB tables, design from scratch, tables design, design document data model, capacity planning.

Experience in building and architecting multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in local setup and AWS and coordinate task among the team.

Managed source code repository, builds, and release configurations, processes, and tools to support daily development, test, and production build and software deployment operations.

Source code admin - Manage and configure GitHub commits, pull requests, resolve issue regarding source code management, manage branching and merging, code freeze process.

Creating pipelines and deploy services to the new ECS clusters. Verify infrastructure provisioned in the console output of Jenkins and in AWS console. Add Jenkins webhooks and configuring of the Jenkins file, Gears file and the Bogie files for deployments to Dev, QA, Pre-Prod and Prod.

Creating and managing Sphinx AD groups for centralized PAM (Privileged Access Management). Sphinx IDs are created with AD groups for QA and Dev, onboarding to Sphinx.

Test Data Readiness source files (Account and transaction) production activity already existing in the One Lake QA, Reference data requirement understanding and the gathering. Tagging logic by setting up the DynamoDB for different scenarios including calculation logic setup.

Tag Prometheus data with accounting/reporting identifiers, tagging account level data with BU/GL/DID once a day. Built enrichment engine to help the process and publishing enriched data to One Lake.

Enrichment Pyspark Application, working on the tagging logic setup, script execution and validation. Developed sample Pyspark script in local to do tagging for BU, GL, DID for both accounts & transaction files.

Enrichment result validation by creating Presto tables for validation based on the test enriched file using Exchange user spaces (Collaboration space).

One pipeline setup and management for managed pipelines work, distinguishing components that make up the One Pipeline, deploy infra and application via managed One Pipeline.

Created CloudFormation templates to build out new server or application stack with RDS to serve as a Hive metastore, made possible to integrate metadata from 20 EMR clusters to single RDS, avoiding data loss even if the EMR was terminated. Static and Dynamic partitioning including buckets for efficiency, Autostaging.

Design Data pipeline flows using Airflow, scheduling jobs programmatically (DAG) in Airflow workflow engine and providing support for the scheduled jobs. Manipulate and process data with tools like Pyspark.

**Environment**: AWS DynamoDB, IAM, Lambda, RDS, EMR, EC2, Redshift, Glue, Snowflake, Jenkins, Docker, Maven, IntelliJ Idea IDE, Python, Tableau, Data Modeling, Splunk logs processing, Spark, Kafka, PostgreSQL, MySQL, Pyspark, Apache Airflow, GitHub, SparkSQL, Airflow DAGs, Informatica ETL.

**Role: Senior Data Engineer / ML Engineer**

**Emp/Client: Cigna, Nashville, TN Sept 2019 – July 2020**

**Scope:** Accountable for processing Cigna PPO, HMO and RX data to produce analyses, reports and applications to enhance the development and analysis of healthcare cost data, and trends. Help and develop data engineering solutions, analysis and ML algorithms, to maintain and constantly update Cigna directory of doctors and services.

**Responsibilities:**

Design, Implement and manage migration services, strategies for traditional systems on Azure (Lift and Shift/Azure Migrate, third-party tools). Hands on implementations and work performed using Azure Suite: Azure SQL DB, ADLS, ADF V2, Azure SQL Datawarehouse, Service Bus, Azure blob, Data platform services.

Data Acquisition of concerned medical cost data involving Data ingestion from and to sources like Azure Cloud Storage, web scraping of HTML and XML files using beautiful soup, Scrapy etc. to Azure blob.

Create ADF pipeline for patients flow in Greater Los Angeles Area, ranked top 10 critical medical institutions to maximize investment influence for controlling hospital-onset infection, analyze relationship b/w health care policy change and sepsis admission rate, decreased incidence rate of hospital-onset infection by 3.2%.

Build ETL solutions involving Databricks by executing code in Notebooks against data in Data Lake using Delta Lake and load into Azure DW. Extensive work on processing, transforming big data using Databricks.

Azure Container Service (ACS) with Containerized deployments, Azure Container Instances and AKS managing compute capacities by handling ACI and AKS side by side to handle spikes in demand.

Build and maintain Docker container clusters managed by ACI, Azure Kubernetes Service. Containerize data wrangling jobs in Docker containers by using Git and Azure Devops for version controlling.

Develop Spark apps use Spark-SQL in Databricks for extraction, transform, and aggregation from multiple formats to analyze and transform data to draw insights of customer usage reports in Azure Analysis services.

Built efficient ETL packages, using both SSIS and ADF2 pipelines to process Fact and Dimension tables with highly complex transforms, SCD type 1/type 2, built workflow using Python and SDF2 to automate dataflow.

Develop BI dashboards, analytical reports and data viz. using MS PowerBI by creating multiple measures using DAX expressions for sales user groups, operations, and finance needs. Azure Data Lake analytics U-SQL jobs to transform data in Azure Data Lake store to move from Raw to Stage and then to Curated zones.

**Environment**: Azure Databricks, MS Azure stack (Azure Analysis, blob, AKS, ACI, Data Lake ADLS, ADF, AML), Azure Devops, Python, Data Modeling, SSIS ETL, Docker, Kubernetes, MySQL, Autosys, Agile, Git, Power BI

**Role: Data Engineer/Cloud Developer**

**Emp/Client: Walmart, Bentonville, AR Aug 2018 – Sept 2019**

**Scope:** Improving effectiveness of supply chain operations in Walmart by extensive research using big data technologies and building custom hybrid ETL models using GCP and Azure.

**Responsibilities:**

* Collaborated with business leaders from Walmart Global sourcing for defining product requirements, scope, vision, and roadmap for forecasting and provided actionable insights relating to the project.
* Implemented end to end high performing systems, integrations that are production ready to commit platform at an enterprise level, involving production rollouts, support, engineering and operational excellence.
* To build and maintain the backend software services, GCP and Azure services that integrate with Walmart owned systems and 3rd party software. Utilize REST APIs, serverless, and event driven microservices.
* To create and to deploy large-scale, containerized apps using Docker, Kubernetes in public clouds like both GCP and Azure to support and build new hybrid-cloud structure to switch seamlessly between GCP and Azure.
* Help, support and Utilize Spring Boot or Spring Framework in Java to design implement low-latency REST APIs involving Azure SQL Server and Cassandra as storage systems.
* Improved and contributed towards Advanced Data Management: Proven proficiency in managing and manipulating huge datasets in the order of terabytes (TB), including advanced knowledge of distributed file systems, data partitioning, storage, retrieval and optimization techniques for handling large-scale data daily.
* Design to code, analyze, and test deployment of cloud technical solutions to Data Governance requirements for Data Lake: Technical Data Catalog, Business Data Catalog, Data Lineage, Data Quality& Access control, Data classification etc. Regulatory and ethical requirements of Data quality framework and metrics.
* Wrote code, develop required cloud solutions both on GCP and Azure, create test cases to review and validate the proposed solution. Maintain code documentation, playbook maintain, and provide timely progress.
* Worked extensively on Azure Data Factory activities such as Lookups, Stored procedures, if condition, for each, set variable, append variable, Get Metadata, Filter and wait. Also, implemented the ADF triggers and scheduled the pipeline, also monitoring them and configured alerts to get notification on failures.
* Performed Root cause analysis on internal and external data and processes to answer all the business specific questions identifying and suggesting improvements based off the results.
* Work with stakeholders, Sam’s Club teams including the Executive, Product, Data and Design teams to discuss and assist data-related technical issues and support their data infrastructure needs, keep Walmart data separate and secure across national boundaries through various data centers and GCP regions.
* Responsible for defining key business problems to be solved while developing, maintaining relationships with stakeholders, SMEs, and cross-functional teams.

**Environment**: Google Cloud Platform, Apache Airflow, Python 3, A/B Test, Risk Analysis, Hadoop Ecosystem, MapReduce, HDFS, Tableau, Spark, Hive, HQL, HBase, Cassandra, Snowflake, Azure ADF, Azure SQL, Spring Boot, Java, Docker, Kubernetes, Tableau, Agile, Scrum, Hybrid-cloud, microservices.

**Role: Data Engineer**

**Emp/Client: Knoah Solutions, Hyderabad, India Dec 2015 – Nov 2017**

**Responsibilities:**

* Develop an ETL solution platform end-to-end, involved in all the requirement gathering, analyzing all the phases of this project, documentation of FRDs and BRDs.
* Wrote complex Spark SQL queries for data analysis, cleanse and transform, involving Hive and PySQL to meet business requirement and end goals. Teradata utilities like BTEQ, Fastload, Multiload, macros in SQL assistant.
* Designed tables in MYSQL, Hive using the Sqoop jobs to process data for import, export to HDFS. Processed large datasets of all different forms and formats (structured, semi and unstructured).
* Worked on creating, testing the Rest APIs using python with Flask framework for the integration of Textfiles, Spreadsheets, FTP server files, JDBC, RDBMS.
* Create and test various transformations considering the business requirement and logics like Router, Filter, Normalizer, Lookup, Stored procedures, Aggregators, Joiner etc.
* Worked extensively on Informatica Power Center mappings using different transformations such as Update Strategy, Expression. Informatica Power exchange for metadata analysis.
* Automation of the code reviewing and unit testing using Python to deliver high quality code maintenance.
* Optimizing the transformations Creating, Designing SQL queries sources, targets and to tune the objects (views, indexes, tables, PPI, UPI) to the dimensional models. Database triggers, stored procedures and use of user exits to access foreign functions.
* Source code admin - Manage and configure GitHub commits, pull requests, resolve issue regarding source code management, manage branching and merging, code freeze process.
* Perform fixes and troubleshooting Data Quality and governance issues using DQ Analyzer and preprocessing.
* Exposure to end-clients, Develop team communication and technical info for daily offshore calls and meetings.

**Environment**: Oracle 9i SQL, Windows XP/Vista, Teradata, ETL, Informatica Power Center 8.1, PLSQL, Hadoop, Hive, Sqoop, Flask, Python Unix Shell Scripting, XML, HTML, Java, Github, Git.

**Role: Junior Python Data Engineer/Analyst**

**Emp/Client: Information and Telecom Systems (India) PVT.LTD, Hyd. India Mar 2013 - Oct 2015**

**Responsibilities:**

Implemented end-to-end systems for Data Analytics, Data Automation and customized visualization tools using Python, Hadoop, Tableau and SQL.

Worked on csv, json, excel different types of files for the Data cleaning and Data analysis.

Managed large datasets using pandas data frames and Spark framework. Hands on data mining on large scale DBs like DB2 and Teradata. Maintain the DBs, summary tables, create them and to optimize.

Performed data visualization with Tableau and generated dashboards to present the findings.

Worked independently and collaboratively throughout the project lifecycle including data extraction/preparation, design and implementation of scalable machine learning analysis and solutions, and documentation of results. Collab with Data Scientists, Data Architects to understand and capture needs.

* Worked closely with developers, end users and administrator in designing, creating databases and objects.
* Developed queries to extract complex data from different tables in remote databases using joins database links and formatted the results into reports and kept log.
* Expertise with DDL and DML statements RDBM’S data dictionaries and normal forms, Data warehousing.
* Work with end users to determine data requirements for reporting requests, data mining requests, the development and delivery of both adhoc/recurring reports and assisting with testing and implementation.
* Created multiple custom SQL queries in Teradata SQL Workbench to prepare the right data sets for Tableau dashboards. Queries involved retrieving data from multiple tables using various join conditions that enabled to utilize efficiently optimized data extracts for Tableau workbooks.

**Environment:** Python, Hive, Spark, MySQL, PostgreSQL, Matplotlib, Seaborn, Tableau, MS SQL Server, Teradata, ETL, IBM DB2, MS Access, MapReduce, FTP servers, HDFS, Linux shell scripts.

**EDUCATION**:

**Phd. in Information Technology** August2022 – 2025

Cumberlands University **Williamsburg, KY**

**M.S. in Data Science** Jan 2018 - Dec 2019

University of Bridgeport **Bridgeport, CT**

**Bachelor of Technology**

Mewar University **Rajasthan, India**