**AWS Data Engineer / Cloud Engineer**

**Vinay Gaddam**

**gaddamvinaykumarreddy7@gmail.com**

**+1-(848) 249-6977**

**Professional Summary:**

* 9+ years of Experience working as a AWS Data Engineer / Cloud Engineer . Expertise in Python, ETL, Scala, SQL, PL/SQL, and T-SQL, with a strong focus on data processing and analysis.
* Skilled in developing web applications using Django, HTML5, CSS3, AJAX, JSON, JavaScript, and XML, creating intuitive user interfaces and interactive experiences.
* experience in Information Technology with a strong background in Analyzing, Designing, Developing, Testing and Implementation of Data warehousing applications in various verticals such as Health care, Retail and AWS platform. Hands on experience on building cloud DWH like AWS S3 - Redshift.
* Extensive Experience with Hadoop distributions like Cloudera and MapR, leveraging the power of big data technologies for large-scale data processing and storage.
* Proficient in working with various database technologies, including MongoDB, Oracle, MySQL, Hadoop HDFS, and Cassandra, ensuring efficient data storage and retrieval.
* Well-versed in cloud platforms such as AWS, Azure and GCP, utilizing their services for scalable and reliable data solutions.
* Skilled in Data Migration, Data Profiling, Data Ingestion, Data Cleansing, Transformation, Data Import, and Data Export using multiple ETL tools such as Informatica Power Centre.
* Experience in job/workflow scheduling and monitoring tools like Oozie, AWS Data pipeline & Autosys
* Expertise in big data processing frameworks, including PySpark, Spark SQL, Hive, Pig, MapReduce, Flume, Impala, Oozie, and Zookeeper, enabling distributed data processing and analysis.
* Competent in version control and collaboration tools like GIT, GitLab, and GitHub, ensuring efficient code management and seamless teamwork.
* Played a key role in migrating Cassandra, Hadoop cluster on AWS and defined different read/write strategies
* Used Matillion inbuilt Scheduler for scheduling simple intervals for Matillion ETL jobs.
* Experienced in managing Azure Data Lakes (ADLS) and Data Lake Analytics and an understanding of how to integrate with other Azure Services.
* Comprehensive Experience in implementing CI/CD practices using Jenkins, CI/CD pipelines, Maven, Docker, and Kubernetes, automating the build, testing, and deployment processes.
* Familiarity with data warehousing technologies like RedShift and Big Query, designing and optimizing data storage for efficient querying and analytics.
* Good knowledge of data visualization using Power BI and Tableau, creating visually appealing and insightful reports and dashboards for data-driven decision-making.
* Expert in configuring and managing HTTP servers like Nginx, ensuring reliable web application deployment and performance.
* Implemented object-oriented Python programming, Multithreading, Django, MySQL, exception handling, and Collections in various projects.
* Designed and developed applications using the Django MVC Framework, following the MVC design pattern.
* Developed consumer-based features and applications using Python, Django, HTML, Behaviour-Driven Development (BDD), and pair-based programming.
* Experience in connecting Cognos and Tableau to Various data sources, including database ,data warehouses, and other data repositories
* Performed various grid operations using Pandas, Flask, and SQL Alchemy.
* Developed Python scripts for parsing XML documents and loading the data into a database.
* queried and managed MySQL database from Python using Python-MySQL connector package and pandas data frames.
* Developed software in Python using libraries such as Beautiful Soup, NumPy, SciPy, Matplotlib, Panda’s data frame, network, urllib2, and MySQL dB for database connectivity. Worked with IDEs like Sublime Text, Spyder, and PyCharm.
* Practical knowledge of logging, monitoring, and alerting using the ELK stack (Elasticsearch, Logstash, and Kibana), enabling real-time data analysis and troubleshooting.
* Strong understanding of Agile methodologies and Experience working with project management tools like JIRA, facilitating efficient project planning and execution.
* Profound Experience in using PyCharm as an IDE for Python development and debugging.
* Demonstrated expertise in designing and implementing data engineering solutions involving data extraction, transformation, and loading (ETL) processes.
* Experienced in working with various data formats and protocols, including CSV, XML, JSON, and RESTful APIs, ensuring seamless data integration and exchange.
* Skilled in troubleshooting and resolving production issues, ensuring the stability and performance of data applications.
* Strong knowledge of data security and privacy practices, ensuring compliance with industry standards and regulations.
* Proven ability to collaborate effectively in cross-functional teams, delivering high-quality solutions within specified timeframes.

Technical Skills:

Programming Languages: Python, Scala, SQL, PL/SQL T-SQL

Framework and Web Technologies: Django, HTML5, CSS3, AJAX, JSON, JavaScript, and XML

Hadoop Distribution: Cloudera, MapR

Database Technologies: MongoDB, Oracle, MySQL, Hadoop HDFS, Cassandra

Cloud Platforms: AWS, GCP, Azure

Big Data: PySpark, Spark SQL, Hive, Pig, MapReduce, Flume, Impala, Oozie, Zookeeper

Version Control and Collaboration : GIT, GitLab, GitHub

CI/CD: Jenkins, CI/CD pipelines, Maven, Docker, Kubernetes

Data Warehouses: RedShift, Big Query

Data Visualization: Power BI, Tableau, Snowflake, Cognos

Server: HTTP Server, Nginx

Logging, Monitoring, and Alerting: ELK stack (Elasticsearch, Logstash, and Kibana)

Others: Agile, JIRA, PyCharm

**Experience:**

**Union Bank, Charlotte, NC Oct 2022 - Present**

**Senior Data Engineer**

**Responsibilities:**

* Written scripts from scratch to create AWS infrastructure using languages such as BASH and Python, created Lambda functions to upload code and to check changes in S3, DynamoDB table.
* Worked on implementing Data warehouse solutions in AWS Redshift, worked on various projects to migrate data from one database toWS Redshift, RDS, ELB, EMR, Dynamo DB and S3
* Managed AWS EC2 instances utilizing S3 and Glacier for our data archiving and long-term backup and UAT environments as well as infrastructure servers for GIT.
* Backing up AWSPost GREtoS3on daily job run onEMRusing Data Frames.
* Create dashboards on snowflake cost model, usage inQlikView. Created program inpythonto handle PL/SQL functions like cursors and loops which are not supported by snowflake.
* Developed strategy for cloud migration and implementation of best practices using AWS services like database migration service, AWS server migration service from On-Premises to cloud.
* Implementation of new tools such as Kubernetes with Docker to assist with auto-scaling and continuous integration (CI) and upload a Docker image to the registry so the service is deployable through Kubernetes. Use the Kubernetes dashboard to monitor and manage the services.
* Improvised a python module that de-normalizes data from RDBMS to JSONs and saved 35 hours as part of the migration.
* Promoted best practices, prepared and implemented Cognos BI Standard Operating Procedures (SOPs) to streamline the reporting process.
* Strengthened Cognos Framework Manager modeling experience due to the utilization of Cognos BI Best Practices.
* Created multiple Glue ETL jobs in Glue Studio and then processed the data by using different transformations and then loaded into S3, Redshift and RDS.
* Responsible for Setup and build AWS infrastructure using resources VPC, EC2, S3, Dynamo DB, IAM, EBS, Route53, SNS, SES, SQS, Cloud Watch, Cloud Trail, Security Group, Auto scaling and RDS using Cloud Formation templates.
* Created job chains with Jenkins Job Builder, Parameterized Triggers, and target host deployments. Utilized many Jenkins plugins and Jenkins API.
* Backing up AWSPost GREtoS3on daily job run on EMRusing Data Frames.
* Automated tasks of extracting metadata and lineage from tools using Python scripts and saved 70+ hours' manual efforts.
* Experience in developing very complex mappings, reusable transformations, sessions, and workflows using Informatica ETL tool to extract data from various sources and load into targets.
* Experience in Developing Spark applications using Spark - SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats.
* Created multiple Recipes in Glue Data Brew and then used in various Glue ETL Jobs.
* Design and Develop ETL Processes in AWS Glue to migrate data from external sources like S3, Parquet/Text Files into AWS Redshift.
* Profile structured, unstructured, and semi-structured data across various sources to identify patterns in data and Implement data quality metrics using necessary queries or python scripts based on source.
* Worked on AWS Data pipeline to configure data loads from S3 to into Redshift.
* Using AWS Redshift, me Extracted, transformed, and loaded data from various heterogeneous data sources and destinations.
* Developed CI/CD pipelines to automate build and deploy to Dev, QA, and production environments.
* Created multiple scripts to automate ETL/ ELT process using PySpark from multiple sources.
* Worked on Spark for improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark SQL, and Spark Yarn.
* Environment: Spark, Hive, Pig, Spark SQL, Spark Streaming, HBase, Sqoop, Kafka, AWS EC2, S3, Cloudera, Scala IDE (Eclipse), Scala, Linux Shell Scripting, HDFS, Python, Snowflake, QlikView, Jon, OpenShift, AWS Glaciered, ETL-Informatica, Oracle, Red Hat Linux, Tableau, Cognos.

**CMS: - Centres for Medicare & Medicaid Services, Baltimore, MD Jan 2021 - Oct 2022**

**Data Engineer / Python Developer**

**Responsibilities:**

* Developed and maintained ETL pipelines using Python and Django frameworks to extract, transform, and load data from various sources into target systems, ensuring data integrity and accuracy.
* Utilized Sqoop for seamless data transfer between Hadoop and relational databases like MySQL, enabling efficient data integration and synchronization.
* Was responsible for creating reports using ETL Tool like Informatica and Cognos reporting tool based on data received from CMS Health Care System.
* Worked with MongoDB, a NoSQL database, to store and retrieve unstructured and semi-structured data, enabling flexible data modeling and storage.
* Experience with 3 years of snowflake as a DBA and a Developer, also with Matillion ETL tool
* Working with Python ORM Libraries, including Django ORM.
* Leveraged MapReduce programming model for distributed data processing, enabling parallel computation and scalability for large datasets.
* Involved in infrastructure as code, execution plans, resource graph and change automation using Terraform. Managed AWS infrastructure as code using Terraform and CloudFormation.
* Well versed with Big data on AWS cloud services i.e. EC2, S3, Glue, Athena, DynamoDB and snowflake.
* Designed and implemented scalable data solutions on AWS using various services, such as EC2, S3, and EMR, ensuring efficient data storage, processing, and analysis.
* Used Matillion tool for setting up ETL process in snowflake.
* Experience in using READ, JOIN, TRANSFORM, WRITE of Matillion ETL tool.
* Developed and maintained data pipelines using AWS Redshift data warehouse, enabling high-performance data querying and analytics for large datasets.
* Played a key role in migrating Cassandra, Hadoop cluster on AWS and defined different read/write strategies
* Experience in job/workflow scheduling and monitoring tools like Oozie, AWS Data pipeline & Autosys
* Implemented responsive web interfaces using HTML5, CSS3, AJAX, JSON, JavaScript, and XML, providing intuitive user experiences for data exploration and visualization.
* Utilized Spark and PySpark for distributed data processing and analytics, enabling high-performance data transformations and machine learning tasks.
* Comprehensive experience in BI & OLAP tools like IBM Cognos Analytics (C11)/Cognos BI 10.2.2 (Report Studio, Framework Manager, Report Studio, Workspace advanced, Analysis Studio, Query Studio, Event Studio, Cognos Connection, PowerPlay Transformer), Cognos Planning 8.3, and ETL tools Informatica, and IBM Data Manager/Decision Stream.
* Develop, Maintain and support Continuous integration framework (CL-CD) based on Jenkins.
* Provide guidance to development team working on PySpark as ETL platform and makes sure that quality standards are defined and met.
* Designed and optimized Hive tables and partitions for improved query performance, ensuring timely data processing and analysis.
* Designed and managed API system deployment using a fast HTTP server and Amazon AWS architecture.
* Leveraged SparkSQL for querying and analysing structured data within Spark, enabling efficient data retrieval and manipulation.
* Developed tools using Python, Shell scripting, and XML to automate some menial tasks.
* Orchestrated data workflows and scheduled jobs using Oozie, ensuring reliable and timely execution of data processing tasks.
* Worked with Zookeeper for distributed coordination and synchronization of tasks across a cluster, ensuring efficient and reliable data processing.
* Utilized Pig for scripting and executing data flows on Hadoop, enabling efficient data manipulation and transformation tasks.
* Implemented containerization of data applications using Docker and orchestrated their deployment and management using Kubernetes, ensuring seamless scalability, portability, and efficient utilization of resources.
* Implemented ELK (Elasticsearch, Logstash, Kibana) stack for real-time log analysis and monitoring, enabling efficient data exploration and visualization.
* Created interactive and visually appealing data visualizations using Tableau. Used Maven as a Build tool.
* Collaborated with GitLab for version control. Worked in Agile environments, utilizing JIRA for project management, tracking tasks, and maintaining an organized workflow throughout the development cycle.
* Environment: Python, Django, ETL, AWS, Redshift, Hive, MySQL, ETL, cognos,Sqoop, MongoDB, Shell Scripting, MapReduce, HTML5, CSS3, AJAX, JSON, JavaScript, XML, Spark, PySpark, SparkSQL, Oozie, Zookeeper, Pig, ELK, Docker, Kubernetes, Tableau, GitLab, JIRA.

**American Eagle, Atlanta, GA Dec 2018 – Jan 2021**

**Data Engineer**

**Responsibilities:**

* Provided the architectural leadership in shaping strategic, business technology projects, with an emphasis on application architecture.
* Utilized domain knowledge and application portfolio knowledge to play a key role in defining the future state of large, business technology programs.
* Created numerous Orchestration and transformation job as part of building data pipelines using Matillion.
* Participated in all phases of data mining, data collection, data cleaning, developing models, validation, and visualization and performed Gap analysis.
* Experience in Metadata Migration from Cognos ReportNet/Cognos 10 BI to Cognos Analytics and troubleshoot issues for Business Partners, End Users and provide accurate data for day - to-day decision-making.
* Designing and building multi-terabyte, full end-to-end Data Warehouse infrastructure from the ground up on Confidential Redshift for large scale data handling Millions of records every day.
* Worked on Big data on AWS cloud services i.e., EC2, S3, EMR and DynamoDB.
* Managed security groups on AWS, focusing on high-availability, fault-tolerance, and auto scaling using Terraform templates. Along with Continuous Integration and Continuous Deployment with AWS Lambda and AWS code pipeline.
* Developed SSRS reports, SSIS packages to Extract, Transform and Load data from various source systems.
* Implementing and Managing ETL solutions and automating operational processes.
* Optimizing and tuning the Redshift environment, enabling queries to perform up to 100x faster for Tableau and SAS Visual Analytics.
* Defined facts, dimensions and designed the data marts using the Ralph Kimball's Dimensional Data Mart modelling methodology using Erwin.
* Created ad hoc queries and reports to support business decisions SQL Server Reporting Services ( SSRS).
* Analyzer the existing application programs and tune SQL queries using execution plan, query analyzer, SQL Profiler and database engine tuning advisor to enhance performance.
* and generate Data Models using ERwin and subsequent deployment to Enterprise Data Warehouse.
* Wrote various data normalization jobs for new data ingested into Redshift .
* Created various complex SSIS/ETL packages to Extract, Transform and Load data
* Advanced knowledge on Confidential Redshift and MPP database concepts.
* Migrated on premise database structure to Confidential Redshift data warehouse.
* Was responsible for ETL and data validation using SQL Server Integration Services.
* Defined and deployed monitoring, metrics, and logging systems on AWS.
* Implemented Work Load Management (WML) in Redshift to prioritize basic dashboard queries over more complex longer-running adhoc queries. This allowed for a more reliable and faster reporting interface, giving sub-second query response for basic queries.
* Experience in building sophisticated metadata models from various data sources using Cognos Framework Manager and used Cognos Transformer to build Multidimensional Cubes. Maintaining and Support Analyst Models and Contributor Applications using the Cognos Planning environment.
* Worked publishing interactive data visualizations dashboards, reports /workbooks on Tableau and SAS Visual Analytics.
* Used Hive SQL, Presto SQL and Spark SQL for ETL jobs and using the right technology for the job to get done.
* Designed and implemented system architecture for Amazon EC2 based cloud-hosted solution for the client.
* Tested Complex ETL Mappings and Sessions based on business user requirements and business rules to load data from source flat files and RDBMS tables to target tables.
* Environment: SQL Server, Erwin, Oracle, Redshift,cognos,IAM, RDS, NOSQL, Snowflake Schema, MySQL, PostgreSQL, AWS Sage Maker

**Data Engineer Oct 2015 – Nov 2018**

**Digiapt, Bangalore, India**

**Responsibilities:**

* Designed and set up Enterprise Data Lake to provide support for various use cases including Analytics, processing, storing, and Reporting of voluminous, rapidly changing data.
* Responsible for maintaining quality reference data in source by performing operations such as cleaning, transformation,
* and ensuring Integrity in a relational environment by working closely with the stakeholders & solution architect.
* Designed and developed Security Framework to provide fine-grained access to objects in AWS S3 using AWS Lambda and DynamoDB.
* communication on cluster and testing of HDFS, Hive, Pig, and Map Reduce to access cluster for new users.
* Performed end-to-end Architecture & implementation assessment of various AWS services like Amazon EMR, Redshift, and S3.
* Implemented the machine learning algorithms using python to predict the quantity a user might want to order for a specific item so we can automatically suggest using kinesis firehose and S3 data lake.
* Used AWS EMR to transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB.
* Installed Oozie workflow engine to run multiple Hive and MR jobs which run independently with time and data availability.
* Creating Lambda functions with Boto3 to deregister unused AMIs in all application regions to reduce the cost of EC2 resources.
* Importing & exporting databases using SQL Server Integrations Services (SSIS) and Data Transformation Services (DTS Packages).
* Used FLUME to export the application server logs into HDFS.
* Developed a reusable framework to be leveraged for future migrations that automate ETL from RDBMS systems to the Data Lake utilizing Spark Data Sources and Hive data objects.
* Conducted Data blending, Data preparation using Alteryx and SQL for Tableau consumption, and publishing data sources to the Tableau server.
* Developed Kibana Dashboards based on the Log stash data and Integrated different sources and target systems into Elastic search for near real-time log analysis of monitoring End to End transactions.
* Implemented AWS Step Functions to automate and orchestrate the Amazon Sage Maker-related tasks such as publishing data to S3, training the ML model, and deploying it for prediction.
* Integrated Apache Airflow with AWS to monitor multi-stage ML workflows with the tasks running on Amazon Sage Maker.
* Environment: Python, AWS, EC2, S3, Lambda, DynamoDB, Oozie, AMIs, SSIS, DTS Packages, FLUME, Pig, Hive, HBase, Sqoop, SQL, DB2, Oracle, MongoDB, Cassandra, Map Reduce, Kafka, Scala, XML, Teradata, ETL, OLTP, OLAP, HDFS, ETL, Alteryx, Tableau, Kibana, Apache Airflow, Amazon Sage Maker.

**Python developer June 2014 – Sep 2015**

**Simform, Ahmedabad, Gujarat, India**

**Responsibilities**:

* Worked in stages such as analysis and design, development, testing, and debugging.
* Developed Views and templates with Python and Django’s view controller and templating language to create a user-friendly website interface.
* Designed and developed the UI of the website using HTML, AJAX, CSS, and JavaScript.
* Worked on frontend frameworks like CSS and Bootstrap to develop web applications.
* Worked on CSV files while trying to get input from the MySQL database.
* Worked with a web server like Nginx.
* Used JavaScript and XML to update a portion of a webpage.
* Developed automated testing framework for command-line-based tests on Linux.
* Used PyCharm in various integrated development environments.
* Used GitHub for version control.
* Involved in developing web services, which were designed to receive client requests.
* Troubleshooting and coding corrections for production issues discovered in the application.
* Worked in developing applications, especially in the UNIX environment, and was familiar with all its commands.
* Environment: Python, Django, HTML, CSS, JavaScript, Jquery, SOAP, MySQL, XML, JSON, Nginx, GIT, PyCharm, Agile.

**Education Details**:

* Bachelors in computer science and Engineering, Lovely professional university