# Vamsikrishna

## Sr. Data Engineer

**+1 4753557885**
vamsikrishnav032@gmail.com
https://www.linkedin.com/in/vamsi-krishna-049a452a3/


## Professional Summary:

- Overall 5+ years of experience in Analysis, Design, Development and Implementation as a Data Engineer.
- Good experience in application development primarily using Hadoop, Python and worked on data analysis.
- Experience in creating Tableau Dashboards using Stack Bars, Bar Graphs, and geographical maps.
- Good understanding of Mapper and Reducer.
- Experienced working with Hadoop Big Data technologies (HDFS and MapReduce programs), Hadoop ecosystems (HBase, Hive, pig) and NoSQL database MongoDB
- Proficient in SQL, PL/SQL and Python coding.
- Experienced the integration of various data sources like RDBMS, Spreadsheets, Text files.
- Experience in applying the latest development approaches including applications in Spark using Scala to compare the performance of Spark with Hive and SQL/Oracle.
- Experience in working with Spark and different components in spark, pyspark, Python, Scala, Git, CI CD, Jenkins, Ansible, Cloud formation templates, SQL, NOSQL, Kafka, Hive, HBase, HDFS and Map reduce.
- Good Knowledge in Amazon AWS concepts like EMR, S3, Lambda, Triggers, Glue, Step functions, Cloud watch events, Redshift, RDS, Athena, Kinesis firehose, data streams and EC2 web services which provides fast and efficient processing of Big Data.
- Strong experience in core Python, Scala, SQL, PL/SQL and Restful web services.
- Good knowledge in querying data from Cassandra for searching, grouping and sorting.
- Extensive experience in Text Analytics, developing different Statistical Machine Learning, Data Mining solutions to various business problems and generating data visualizations using R, Python.
- Experience in data preprocessing, data analysis, machine learning to get insights into structured and unstructured data.
- Experience in advanced procedures like text analytics, processing using in memory computing capabilities like Spark written in Scala.
- Proficient in programming languages like R and Python.
- Strong experience and knowledge in Data Visualization with Tableau creating: Line and scatter plots, Bar Charts, Histograms, Pie chart, Dot charts, Box plots, Time series, Error Bars, Multiple Charts types, Multiple Axes, subplots, geographical maps etc.
- Experience in designing, developing, scheduling reports/dashboards using Tableau.
- Experience in various cloud vendors like AWS, GCP and Azure.
- Expertise in performing data analysis and data profiling using complex SQL on various sources systems including Oracle
- Experienced in building and optimizing big data pipelines, architectures, and data sets.
- Experience in developing solutions to analyze large data sets efficiently

## Technical Skills:

| | |
|---|---|
| **Big Data** | Hadoop, HDFS, Hbase, Hive, MapReduce, Spark, Cassandra, Scala, Git, CI CD, Jenkins, Kafka, Ansible |

| | |
|---|---|
| **Languages** | Python (Jupyter Notebook, PyCharm IDE), R, C, C++, SAS |
| **Cloud Computing Tools** | Azure, AWS (S3, EC2, Lambda, Triggers, Glue, EMR, Step functions, Cloud watch<br>events, Redshift, RDS, Athena, Kinesis firehose, data streams) |
| **ETL tools** | TensorFlow. Data API, PySpark |
| **Modelling and Architect Tools** | Star-Schema, Snowflake-Schema Modelling, FACT and dimension tables, Pivot<br>Tables |
| **Databases** | Snowflake Cloud Database, Oracle, MS SQL Server, Teradata, MySQL, DB2 |
| **Database Tools** | SQL Server Data Tools, Visual Studio, Spotlight, SQL Server Management Studio,<br>Query Analyzer |
| **Reporting Tools** | MS Excel, Tableau, Tableau server, Tableau Reader, Power BI |
| **Machine Learning Algorithm's** | Logistic Regression, Linear Regression, Support Vector Machines, Decision Trees, K-Nearest Neighbors, Random Forests, Gradient Boosted Decision Trees,<br>Naive Bayes, K-Means Clustering and Hierarchical Clustering. |
| **Deep Learning** | CNN Convolutional Neural Networks, LSTM, GRU |

## Education:

*Masters- University of Bridgeport (computer science)*

## Work Experience:

**Client: Crossover Health, CA** **May 2023 to Present**

**Role: Sr. Data Engineer**

### Responsibilities:

- Developed complete end to end Big-data processing in Hadoop eco system.
- Provided application support during the build and test phases of the SDLC for their product.
- Recreated existing application logic and functionality in the Azure Data Lake, Data Factory, Data Bricks, SQL Database and SQL data warehouse environment.
- Worked on data migration from on-prem SQL server to Azure Synapse Analytics (DW and Azure SQL DB)
- Performed data profiling and transformation on the raw data using Python, and oracle
- Developed predictive analytics using Apache Spark Scala APIs.
- Created dimensional model for the reporting system by identifying required dimensions and facts.
- Created automated python scripts to convert the data from different sources and to generate the ETL pipelines.
- Designed and implemented database solutions in Azure SQL Data Warehouse, Azure SQL
- Developed customer cleanse functions, cleanse lists and mappings for MDM Hub.
- Worked extensively on Oracle PL/SQL, and SQL Performance Tuning.
- Worked on Python Open stack API's.
- Involved in modeling (Star Schema methodologies) in building and designing the logical data model into Dimensional Models.
- Created shared dimension tables, measures, hierarchies, levels, cubes and aggregations on MS OLAP/ OLTP/Analysis Server (SSAS).
- Managed Azure Data Lakes (ADLS) and Data Lake Analytics and an understanding of how to integrate with other Azure Services.
- Wrote MapReduce jobs to generate reports for the number of activities created on a day, dumped from the multiple sources and the output was written back to HDFS.
- Developed Spark code using Scala and Spark-SQL for faster testing and data processing.

- Used Pig as ETL tool to do transformations, joins and some pre-aggregations before storing the data into HDFS.
- Used Hive to analyze data ingested into HBase by using Hive-HBase integration and compute various metrics for reporting on the dashboard
- Worked on the creating Adhoc reports, Database Imports and Exports using SSIS
- Migrated Azure pipelines to a home-grown cloud tool. Recreated existing logic and functionality of the Azure application from Azure Data Lake, Data Factory, SQL DB.
- Developed data pipelines in home-grown tool using Spark, SQL, Python, PySpark.
- Migrated the applications of a new project's module to GCP. Recreated data lake storage access in GCP. Migrated the data from SQL db to Big Query.
- Experience in moving data between GCP and Azure. Created and accessed secret versions using GCP Secret Manager to access API's.
- Responsible for ETL and data validations. Built data pipelines and submitted the jobs in homegrown tool using dataproc clusters and airflow operators at the backend.
- Developed spark applications using PySpark and Spark-Sql for data-extraction, data validations, transformations, and aggregations from different file formats.
- Hands-on experience developing SQL scripts for automation purpose. Created Big Query authorized views for exposing the data to other teams.
- Setup GCP Firewall rules to allow or deny traffic to and from the VM's instances based on specified configuration and used GCP cloud CDN (content delivery network) to deliver content from GCP cache locations drastically improving user experience and latency.
- Worked on google cloud platform (GCP) services like compute engine, cloud load balancing, cloud storage, cloud SQL, stack driver monitoring and cloud deployment manager.
- Involved in developing DAGS using Airflow orchestration tool and monitored the weekly processes.
- Coordinated with business users to gather requirements and plan roadmaps according to deliverables.
- Performed several POCs on migrating the data pipelines and well documented in the confluence.

**Environment**: Azure, Azure Synapse, Azure Data Lake, GCP, SQL, Oracle12c, PL/SQL, Bigdata, Hadoop, Spark, Scala, APIs, Pig, Python, GCP, Kafka, HDFS, ETL, MDM, OLAP, OLTP, SSAS, T-SQL, Hive, SSRS, Tableau, Map Reduce, Scala, HBase, SSI

**Client: CorVel Corporation, CA** **Oct 2022 – April 2023**

**Role: Data Engineer**

**Responsibilities:**

- Followed Test driven development of Agile Methodology to produce high quality software.
- Designed and developed a horizontally scalable APIs using Python Flask.
- Conducted JAD sessions, wrote meeting minutes and also documented the requirements.
- Worked with cloud providers and API's for Amazon (AWS) EC2, S3, VPC with GFS storage.
- Expertly handled the stream processing and storage of data to feed into the HDFS systems using Apache Spark, Sqoop.
- Installed application on AWS EC2, configured the storage on AWS S3 buckets and worked closely with AWS EC2 infrastructure teams to troubleshoot complex issues.
- Write ETL scripts to move data from HDFS to S3 and created Hive external tables on top of this data to be utilized in Big data applications.
- Configured and maintained Amazon EMR manually and as well as through Cloud Formation scripts in Amazon AWS.
- Created scripts to sync data between local MongoDB and Postgres databases with those on AWS.
- AWS EMR to process big data across Hadoop clusters of virtual servers on Amazon Simple Storage Service (S3).
- Expertise in AWS data migration between different database platforms like Local SQL Server to Amazon RDS, EMR HIVE and experience in managing and reviewing Hadoop log files in AWS S3.
- Built and supported several AWS, multi-server environment's using Amazon EC2, EMR, EBS, Redshift and deployed the Big Data Hadoop application on AWS cloud.
- Worked extensively with importing metadata into Hive using Python and migrated existing tables and applications to work on AWS cloud (S3).

- Worked on AWS EC2, IAM, S3, LAMBDA, EBS, Elastic Load balancer (ELB), auto scaling group services.
- Developed Spark programs and Spark-SQL/Streaming for faster testing and processing of data.
- Transformed the On-premise Hadoop jobs to run on AWS EMR.
- Used Jenkins pipelines to drive all microservices builds out to the Docker registry and then deployed to Kubernetes.
- Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
- Worked extensively with importing metadata into Hive using Python and migrated existing tables and applications to work on AWS cloud (S3).
- AWS EMR to process big data across Hadoop clusters of virtual servers on **Amazon Simple Storage Service (S3)**.
- Worked with Data ingestion, querying, processing and analysis of big data.
- Performed tuned and optimized various complex SQL queries.
- Developed normalized Logical and Physical database models to design OLTP system.
- Extensively involved in creating PL/SQL objects i.e. Procedures, Functions, and Packages.
- Performed bug verification, release testing and provided support for Oracle based applications.
- Used Model Mart of Erwin for effective model management of sharing, dividing and reusing model information and design for productivity improvement
- Extensively used Hive optimization techniques like partitioning, bucketing, Map Join and parallel execution.
- Wrote, tested and implemented Teradata Fastload, Multiload, DML and DDL.
- Used various OLAP operations like slice / dice, drill down and roll up as per business requirements.
- Wrote SQL queries, stored procedures, views, triggers, T-SQL and DTS/SSIS.
- Handled importing of data from various data sources, performed data control checks using Spark and loaded data into HDFS.
- Designed SSRS reports with sub reports, dynamic sorting, defining data source and subtotals for the report.
- Gathered SSRS reports requirements and created in Tableau.
- Designed and developed Map Reduce jobs to process data coming in different file formats like XML.

**Environment**: SQL, PL/SQL, Kafka1.1, AWS, API's, Agile, ETL, HDFS, OLAP, HDFS, T-SQL, SSIS, Teradata, Hive, SSRS, Sqoop, Tableau, Map Reduce, XML.

**Client: Ethon Healthcare Solutions-India** **June 2018-Nov 2021**

**Role: Data Engineer**

**Responsibilities:**

- Used custom developed PySpark scripts to pre-process, transform data and map to tables inside the CIF (Inmon Corporate Information Factory) data warehouse.
- Developed shell scripts of Sqoop jobs for loading periodic incremental imports of structured data from various RDBMS to S3 and used Kafka to ingest real-time website traffic data to HDFS.
- As part of reverse engineering discussed issues/complex code to be resolved and translated them into Informatica logic and prepared ETL design documents.
- Experienced working with team, lead, developers, interfaced with business analysts, coordinated with management and understand the end user experience.
- Used Informatica Designer to create complex mappings using different transformations to move data to a Data Warehouse.
- Developed mappings in Informatica to load the data from various sources into the Data Warehouse using different transformations like Source Qualifier, Expression, Lookup, aggregate, Update Strategy and Joiner.
- Scheduling the sessions to extract, transform and load data into the warehouse database on Business requirements using scheduling tools.
- Extracted (Flat files, mainframe files), Transformed and Loaded data into the landing area and then into staging area followed by integration and semantic layer of Data Warehouse (Teradata) using Informatica mappings and complex transformations (Aggregator, Joiner, Lookup, Update Strategy, Source Qualifier,

Filter, Router and Expression Optimized the existing ETL pipelines by tuning SQL queries and data partition techniques.

- Created independent data marts from existing data warehouses as per the application requirement and updated them on bi-weekly basis.
- Decreased the Azure billing by pivoting from using Redshift storage to Hive tables for unpaid services and implemented various techniques like Partitioning and Bucketing over hive tables to improve the query performance.
- Used Presto distributed query engine over hive tables for its high performance and low cost

**Environment**: ER/ Studio, SQL, Python, APIs, OLAP, OLTP, PL/SQL, Oracle, Teradata, BI, Tableau, ETL, SSIS, SSAS, SSRS, T-SQL, Redshift.