Vamsi V
Data Engineer
| Vamsi331v@gmail.com| (469)567-0331|

PROFESSIONAL SUMMARY:

- Diligent IT proficient looking for business with around 7+ years of work experience in designed, configured, and deployed **Amazon Web Services (AWS)** for a multitude of applications utilizing the **AWS stack** (Including **EC2, Glue, Lambda, SNS, S3, RDS, Cloud Watch, SQS, IAM),** focusing on high-availability, fault tolerance, and auto-scaling.
- Designed and developed **AWS data pipeline** to migrate data from sources like **Teradata, oracle** into Amazon **S3**
- Extensive experience in the **Data Engineering field** including Ingestion, and **Developing data models, pipeline architectures**, and providing **ETL** solutions for project models.
- Knowledge and experience in designing, creating, testing, and maintaining the complete data management from **Data Ingestion**, **Data Curation, Data processing/Transformation**.
- Hands on experience in **Data Provision** with in-depth knowledge on **Oracle Database** and **Spark APIs** like **Spark Framework-SQL, DSL, Streaming,** by dealing with various File designs like **parquet** and execution tuning of flash applications from different aspects.
- Worked with BI tools and services & Data visualization tools such as **Tableau, Amazon Quick Sight**.
- Solid programming abilities in **Python** and **Scala** to construct effective and powerful data **pipelines**.
- Experience in using and tuning relational databases (e.g: **Microsoft SQL Server, Oracle, MySQL**) and columnar databases (e.g. **Microsoft SQL Data Warehouse**).
- Expertise in end-to-end **Data Processing jobs** to analyze data using **MapReduce**, **Spark,** and **Hive**.
- Solid Experience in working with **Linux/Unix** conditions, composing Shell Scripts.
- Extensive experience with **Apache Airflow**, **Bash/Python** scripting for scheduling tasks and process automation.
- Hands-on-experience with **ETL** and **ELT** tools such as **AWS Glue**.
- Worked with **Jenkins** for **CI/CD** and **New Relic** dashboards for pipeline event logging.
- Designed, Implemented and Developed large scale solutions to solve complex problems with data from multiple areas with different types of data.
- Worked with different real time ingest services in **AWS Kinesis**
- Expertise in working on bringing in and sending out data utilizing **Apache Sqoop** from **RDBMS** to **Hadoop Platform** and **vice versa**.
- Skilled in designing and implementing **ETL architecture** and actively tuned it for better performance.
- Proficient in data processing with **Hadoop MapReduce & Apache Spark.**
- Extensive experience in understanding security requirements of **Hadoop** and **data governance**.

- Proficient in **Oracle Database, SQL, PostgreSQL, Python** programming and **DBMS** concepts.
- Experienced in software analysis**, datasets and design, development, testing,python**,implementation of **Cloud, Big Data,Spark,Hadoop** and in maintaining data pipelines in role of Data Engineer.
- Implemented conceptual, logical, physical models and meta data solutions for **Data Modeling.**

Technical Skills:

Big Data Ecosystems	HDFS, YARN, MapReduce, Spark, Hive, Airflow, StreamSets, Sqoop, HBase, Ambari, ZooKeeper, Nifi, Sentry, Ranger
Aws	EMR,EC2,EBS,RDS,S3,Athena,Glue,Lambda,SQS,DynamoDB, Redshift,ECS,Kinesis
Azure	Databricks, Data Lake, Blob Storage, Azure Data Factory, SQL Database, SQL Data Warehouse, Azure Active Directory
Scripting Language	Python, Scala, PowerShell Scripting, HiveQL.
Version Control	Git, SVN, Bitbucket
ETL Tools	Tableau, Microsoft Excel, Power BI, R, Google Data Studio,
Development Methodologies	Agile, Waterfall
Database	MySQL, Oracle, Teradata, MS SQL SERVER, PostgreSQL, DB2
NoSQL Database	DynamoDB, Redis, MongoDB
ETL Toolss	Informatica

WORK EXPERIENCE:

Client: New York Education department
Jan 2021 to Present

Role: Senior Data Engineer

Responsibilities:

- Created pipelines to consume streaming data from **AWS Kinesis** and used business logic to massage, transform, and serialize raw data.

- Created a cloud arrangement template in python formate to use content conveyance with cross regionreplication utilize **Amazon Virtual Private Cloud**.
- Made a Data pipeline using processor Groups and various processors in **Apache Nifi** for flat file.
- Created outer table schemas for the data being handled as the primary query engine of **EMR** with **Amazon EMR**,**S3, Athena,Glue and Kinesis.**
- **Relational Database Management system (RDBMS)** as a feature of **Proof of Concept (POC)** on Amazon **EC2.**
- Deal with **ETL** Migration services by making and sending **AWS Lambda functions** to give a serverless datapipline that can be composed to **Glue** catalog and querie from Athena.
- Applied schema to develop a single module for serializing and deserializing AVRO data by applying schema
- Maintained and developed complex **SQL** queries, views, functions and reports that qualify customer requirements on **Snowflake**.
- Built analytical warehouses in **Snowflakes** and queried data in staged files by referencing metadata columns in a staged file.
- Designed and developed Spark processes using Scala to pull data from an apply transformations to it in Snowflake.
- Migrated data from bucket to **Snowflake** by writing custom read/write **snowflake** utility function using **Airflow Snowflake Operator**.

- Implemented user provisioning, password reset, creating and mapping groups.feature Installed and configured for user provision and day to Identity administration.
- Communicate with Clients regarding suitable functional and technical areas.
- Build and Configure tasks like aggregation, ID refresh, schedule tasks, correlation, etc.
- Reusable shell scripts for **Hive**, were created. Processes for error handling, logging, and metadata management should all be standardized.
- Evaluated and implemented an **ETL tool** for processing business-critical data into **Hive** Cloud aggregated tables. Deployed and developed Bigdata apps in including **Spark, Hive, AWS Kinesis**.
- Developed a **Queryable State** by Scala to query streaming data and enriched the functionalities of the framework.
- Dealt with **Spark** Data sources, **Spark Data frames**, **Spark SQL** and Streaming utilizing **Scala**.
- Worked broadly on Components like **Databrick**, Virtual machine, Blob storage.
- Used **ETL** Streaming for pipelined **S3** engine to process data streams to deploy of flexible windows.
- Aptitude in utilizing different file formats designs like Text records, CSV, Parquet.
- Experience in custom register capacities utilizing **Spark SQL** and performed intuitive querying.
- Experienced for the masking and encrypting the sensitive data in the fly
- Experience in the maintain and creating the **AirDAG's** using the **Apache Airflow**

Client: Kaiser Permanent Organization
Oct 2019 to Dec 2020
Role: Data Engineer

Responsibilities:

- Install and configure **Apache Airflow** for **Azure Blob storage** data warehouse and created **Dags** to run the **Airflow**.
- Build migration of serveal databases and Applications and Web Servers from onpremises to the **Azure Cloud**.
- Utilized copy Activity in the Azure Data Factory to duplicate data among datastores found on-premises and the cloud.
- Made python notepads on the **Azure Databricks** for handling the datasets and the load them into **Azure SQL data bases**.
- Designed and developed **Flink pipelines** to consume streaming data from **DataLakes** and applied business logic to massage and transform and serialize raw data.
- Developed common **Flink** module for serializing and deserializing AVRO data by applying schema.
- Maintained and developed complex **SQL** queries, views, functions and reports that qualify customer requirements on **Snowflake**.
- Perform Data Profiling to learn about behavior with various features such as traffic pattern, location, Date and Time etc.
- Implemented user provisioning, password reset, creating and mapping groups to users using Azure identity management. feature Installed and configure for user provision and day to Identity administration.
- Communicate with Clients regarding **Azure IAM IIQ** at suitable functional and technical areas.
- Build and Configure Azure IAM in-built tasks like aggregation, ID refresh, schedule tasks, correlation, etc.
- Have done customizations in business process/workflow, reports, in IIQ console to add new commands.
- Development of key modules and custom requirements in the project. Perform User Access Administration using Azure Active Directory.
- Manage User Access/Login Security to **Azure IAM** Applications.
- Coordinating with the Clients / on-site team for gathering enhancement requirements, status updates and issue handling.
- Implemented layered architecture for Hadoop to modularize design. Developed framework scripts to enable quick development. Designed reusable shell scripts for **Hive, Sqoop, Flink** .Standardize error handling, logging and metadata management processes.
- Evaluated and working on Azure Data Factory as an **ETL tool** to process business critical data into aggregated tables in **Hive Cloud**. Deployed and Development in Bigdata applications like **Spark, Hive, Flink** in **Azure cloud**.
- Developed a Queryable State for Flink by Scala to query streaming data and enriched the functionalities of the framework.
- Implemented **Spark** with Scala and utilizing Data frames and **Spark** SQL API for faster processing of data.
- Involved in ingestion, transformation, manipulation and computation of data using StreamSets, **Spark** with **Scala**.
- Involved in data ingestion into **MemSql** using **Flink pipelines** for full load and Incremental load on variety of sources like web server, **RDBMS** and **Data API's**.
- Worked on Spark Data sources, **Spark Data** frames, **Spark SQL** and Streaming using **Scala**.
- Worked extensively on Azure Components such as **Databrick**, Virtual machine, Blob storage

- Experience in integrating **Spark-Memsql** connector and **DBC** connector to save the data processed in Spark to **MemSql**.
- Used Flink Streaming for pipelined **Flink** engine to process data streams to deploy new API including definition of flexible windows.
- Data pipeline with data producers streaming data into large scale clusters, events being consumed by large scale **Spark/Flink** consumers
- Expertise in using different file formats like Text files, CSV, Parquet.
- Experience in custom compute functions using **Spark SQL** and performed interactive querying.

Client: Supraja Technologies
Sep2017 to Aug2019

Role: Data Engineer

Responsibilities:

- Understanding business needs, analyzing functional specifications and map those to develop and designing **MapReduce** programs and algorithms
- Designed and implemented **MapReduce**-based large-scale parallel relation-learning system.
- Customized **Air dags** interceptors to encrypt and mask customer sensitive data as per requirement
- Recommendations using Item Based Collaborative Filtering in **Apache Spark.**
- Worked with NoSQL databases like **Hbase** in creating Hbase tables to load large sets of semi structured data coming from various sources.
- Built web portal using python, it makes a data call to the **Elastic search** and gets the row key.
- Used **Kibana**, which is an open source based browser analytics and search dashboard for **Elastic Search**.
- Used **Amazon web services (AWS)** like EC2 and S3 for small data sets.
- Performed importing data from various sources to the **Cassandra** cluster using Sqoop.
- Developed iterative algorithms using **Spark Streaming** in **Scala** for near real-time dashboards.
- Installed and configured Hadoop and Hadoop stack on a **40 node** cluster.
- Involved in customizing the partitioner in **MapReduce** in order to root Key value pairs from Mapper to Reducers in XML format according to requirement.
- Configured Airflow for efficiently collecting, aggregating and moving large amounts of log data.
- Involved in creating Hive tables, loading the data using it and in writing Hive queries to analyze the data.
- Implemented **AWS** services to provide a variety of computing and networking services to meet the needs of applications
- Involved in scheduling workflow engine to run multiple Hive.
- Designed and built the Reporting Application, which uses the **Spark SQL** to fetch and generate reports on **HBase** table data.
- Worked on batch processing of data sources using **Apache Spark, Elastic search**

- Extracted the needed data from the server into **HDFS** and Bulk Loaded the cleaned data into **HBase**.
- Used different file formats like Text files, Sequence Files, Avro, Record Columnar CRC, ORC
- Strong Experience in implementing Data warehouse solutions in **Amazon web services (AWS)** Redshift; Worked on various projects to migrate data from on premise databases to AWS Redshift, RDS and S3.
- Involved in **ETL**, Data Integration and Migration
- Responsible for creating **Hive UDF's** that helped spot market trends.
- Optimizing Hadoop **MapReduce** code, **Hive** for better scalability, reliability and performance
- Experience in storing the analyzed results back into the **Cassandra** cluster.
- Developed custom aggregate functions using **Spark SQL** and performed interactive querying

Client: Verinon Technologies
March 2016 to Aug 2017

Role: Data Analyst

Responsibilities:

- Migrated the Django database from SQLite to **MySQL** to PostgreSQL with complete data integrity and designed, developed and deployed CSV Parsing using the big data approach on **AWS EC2**.
- Developed tools using Python, **Shell scripting, XML** to automate some of the menial tasks. Interfacing with supervisors, artists, systems administrators and production to ensure production deadlines are met. Developed frontend and backend modules using **Python** on Django including Tasty Pie Web Framework using Git.
- Supported various client projects and internal efforts involving **AWS Engineering**, managing the Innovation Center specifically the Big Data Platform & Analytics tools-platforms available on Hortonworks.
- Administered and monitored multi–Data center Cassandra cluster based on the understanding of the Cassandra Architecture.
- Created automated archive process to remove unused tables to ensure optimal database speed. Implemented 3rd party data transformation process **using Redshift, Lambda S3, Kinesis & EDI** Exchange software reducing integration time by a factor of 10.
- Involved in general application development using Python, HTML/CSS with strong integration with Cloud Technologies.
- Involved in the migration from Sqlite3 to Apache Cassandra database. Cassandra data model designing, implementation, maintaining and monitoring.

- Configured various big data workflows to run on top of Hadoop and these workflows comprise of heterogeneous jobs like MapReduce and Involve in evaluating existing server and virtualization environments for needed and useful upgrade opportunities.
- Designed Spark based real-time data ingestion and real-time analytics, Wrote data to synthesize alarms using Scala also used Spark-SQL to Load data and create SchemaRDD and loaded it into Hive Tables and handled Structured data using Spark SQL.
- Built Single Page Applications (SPA), Responsive Web Design (RWD) UI, Rich Restful Service Applications, and HTML Wireframes using HTML5 Grid Structures/Layouts, CSS3 Media Queries, Ajax,and Bootstrap.
- Built a new CI pipeline. Testing and deployment automation with Docker, Jenkins and Puppet. Utilized continuous integration and automated deployments with Jenkins and Docker.
- Involved in development of Python APIs to dump the array structures in the Processor at the failure point for debugging, used Django APIs for database access.
- Used ETL warehouse IBM DataStage for filtering and visualize the raw data, Created Server instances on AWS and installed Swagger for deploying Microservices.
- Used **Amazon Web Services (AWS)** for improved efficiency of storage and fast access and Working on Development & testing of many features for dashboard using Python, Bootstrap.
- Analyzed data and identify leading SaaS, PaaS or IaaS solutions for clients. Involved in front end and utilized Bootstrap for page design and Using the advanced python packages like NumPy, SciPy for various sophisticated numerical and scientific calculations.
- Developed ETL (Extraction, Transformation and Loading) procedures and Data Conversion Scripts using Pre-Stage, Stage, Pre-Target and Target tables.

Client:Esoft Consulting
March 2015 to Feb 2016

Role: SQL Developer

Responsibilities:

- Created for building-scalable distribution in the data solution using the **Hadoop.**
- Expericed in the bring inside the data from **MS SQL Server, MySQL** and **Teradata** into **HDFS** using the **Sqoop.**
- Played a Important role in the dynamic partitioning and the Bucket of the data stored in the **Hive Metadata.**
- Composed Hive QL queries for coordinating various tables for the make views and produce the outcomes in the set.
- Gathered the log data from **Web Servers** and coordinated into **HDFS.**
- Experienced and changing of huge arrangements of organized and unstructured data

- Utilized MapReduce programs for data cleaning and changes and load the result into the **Hive** tables in various record designs.
- Made information pipelines for various occasions to stack the information from **DynamoDB** to **AWS S3** pail and afterward into HDFS location.
- Engaged with stacking data into **HBase NoSQL database.**
- Assembling, Managing and booking workflows for start to finish work handling.
- Dealt with expanding Hive.
- core functionality by writing custom UDFs utilizing python.
- Examining of Large volumes of organized data utilizing **SparkSQL.**
- Moved HiveQL queries into **SparkSQL** to further develop execution.