**Sri Charana**

**Sr.Data Engineer**

Email id: sricharana0928@gmail.com

Contact no: +1 913-608-7714

LinkedIn URL: http://linkedin.com/in/sri-charana-50a650269

**PROFESSIONAL SUMMARY**

* Around 8+ years of IT experience with Data Engineer/Analyst and coding with analytical programming using SQL, Python, Snowflake and AWS.
* Good working knowledge in multi-tiered distributed environment, good understanding of Software Development Lifecycle (SDLC)-Agile and Waterfall Methodologies
* Experience in implementing various Big Data Analytics, Cloud Data Engineering, and Data Warehouse/Data Mart, Data Visualization, Reporting, Data Quality, and Data virtualization solutions.
* Have a proven track record of working as a Data Engineer on Amazon cloud services, Big data/Hadoop Applications, and product development.
* Well-versed with big data on AWS cloud services i.e., EC2, S3, Glue, Dynamo DB, and Redshift.
* Experience in job/workflow scheduling and monitoring tools like Oozie, AWS Data pipeline & Autosys.
* Defined and deployed monitoring, metrics, and logging systems on AWS.
* Experience working on creating and running Docker images with multiple micro-services.
* Good experience in deploying, managing, and developing with MongoDB clusters.
* Experienced in using distributed computing architectures such as AWS products (EC2, Redshift, and EMR, Elastic search, Athena and Lambda), Hadoop, Python, Spark and effective use of Map Reduce, SQL and to solve big data type problems.
* Experience in job/workflow scheduling and monitoring tools like Oozie, AWS Data pipeline & Autopsy’s.
* Defined and deployed monitoring, metrics, and logging systems on AWS.
* Docker container orchestration using ECS, ALB, and lambda.
* Proficiency in multiple databases like MongoDB, MySQL, Oracle, and MS SQL Server.
* Performed SQL and PL/SQL tuning and Application tuning using various tools like EXPLAIN PLAN, SQL TRACE, TKPROF, and autocracy.
* Extensive hands on experience tuning spark Jobs.
* Experienced in working with structured data using HiveQL, and optimizing Hive queries.
* Strong SQL development skills including writing Stored Procedures, Triggers, Views, and User-Defined Functions.
* Designed and Developed Data Warehouses using Star Schema, Snow Flake Schema depending upon business needs.
* Expert in developing SSIS/DTS Packages to extract, transform and load (ETL) data into data warehouse/data marts from heterogeneous sources.
* Good understanding of software development methodologies, including Agile (Scrum).
* Expertise in the development of various reports, dashboards using various Tableau Visualizations.
* Hands-on experience with different programming languages such as Python.
* IT professional with around 8+ years of experience as a Big Data Engineer with expertise in designing data intensive applications using Hadoop Ecosystem, Big Data Analytical, Cloud Data engineering, Data Warehouse / Data Mart, Data Visualization, Reporting, and Data Quality solutions.
* Experience in using different Hadoop ecosystem components such as HDFS, YARN, Map Reduce, Spark, Sqoop, Hive, Impala, and HBase, Kafka, and Crontab tools.
* Design, develop, test, implement and support Data Warehousing ETL using Talend.
* Involved in Development and Created PL/SQL stored procedures, functions.
* Experience in developing ETL applications on large volumes of data using different tools: Map Reduce, Spark,Pyspark, and Spark-SQL.

**EDUCATION**

Bachelors in computer science and Engineering-2014.

**TECHNICAL SKILLS**

| **Big Data Ecosystem** | HDFS, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Zookeeper, Kafka, Cassandra, Apache Spark, Spark Streaming, HBase, Flume, Impala |
| --- | --- |
| **Hadoop Distribution** | Cloudera CDH, Horton Works HDP, Apache, AWS |
| **Languages** | Shell scripting, SQL, PL/SQL, Python, R, PySpark, Pig, Hive QL, Scala, Regular Expressions |
| **Web Technologies** | HTML, JavaScript, Restful, SOAP |
| **Operating Systems** | Windows (XP/7/8/10), UNIX, LINUX, UBUNTU, CENTOS. |
| **Version Control** | GIT, GIT HUB |
| **IDE & Tools, Design** | Eclipse, Visual Studio, Net Beans, Junit, CI/CD, SQL Developer, MySQL, SQL Developer, Workbench, Tableau |
| **Databases** | Oracle, SQL Server, MySQL, Cassandra, Teradata, PostgreSQL, MS Access, Snowflake, NoSQL Database (HBase, MongoDB). |
| **Operating Systems** | Windows 98, 2000, XP, Windows 7,10, Mac OS, Unix, Linux |
| **Cloud Technologies** | Amazon Web Services (AWS) |
| **Data Engineer/Big Data Tools / Cloud / Visualization / Other Tools** | Databricks, Hadoop Distributed File System (HDFS), Hive, Pig, Sqoop, MapReduce, Flume, YARN, HBase, Hortonworks, Cloudera, MLlib, Oozie, Zookeeper, etc. AWS S3, EC2, Athena, Glue, kinesisAWS, Salesforce, Linux, Bash Shell, Unix, etc., Power BI, SAS, Crystal Reports, Dashboard Design, Palantir, BW, Terraform, Jenkins, Informatica, Flink, Celonis, Druid. |

***Client: Verizon, Dallas, TX Feb 2022 to Current***

***Role: Sr.AWS Data Engineer***

***Responsibilities:***

* Used AWS Athena extensively to ingest structured data from S3 into other systems such as RedShift or to produce reports.
* The Spark-Streaming APIs were used to conduct on-the-fly transformations and actions for creating the common learner data model, which receives data from Kinesis in near real time.
* Performed end- to-end Architecture &amp; implementation assessment of various AWS services like Amazon EMR, Redshift, S3, Athena, Glue and Kinesis.
* Develop and maintain Angular-based front-end applications to visualize and interact with data stored in AWS data services such as S3, Redshift, Athena, and Glue.
* AWS CI/CD Data pipeline and AWS Data Lake using EC2, AWS Glue, AWS Lambda.
* Experience in GCP Dataproc, GCS, Cloud functions, BigQuery.
* Experience in moving data between GCP and Azure using Azure Data Factory.
* Utilized Kubernetes and Docker for the runtime environment for the CI/CD system to build, test,

and deploy.

* Expertise in implementing DevOps culture through CI/CD tools like Repos, Code Deploy, Code

Pipeline, GitHub.

* Experience in building power bi reports on Azure Analysis services for better performance.
* Used cloud shell SDK in GCP to configure the services Data Proc, Storage, BigQuery
* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Experience in developing Spark applications using Spark-SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats for Analyzing Transforming the data to uncover insights into the customer usage patterns.
* Implemented solutions for ingesting data from various sources and processing the Data-at-Rest utilizing Big Data through Hadoop, MapReduce, Pig and Hive.
* Provided the technical support for debugging, code fix, platform issues, missing data points, unreliable data source connections and big data transit issues.
* Having proficient experience in various Big Data technologies like Hadoop, Apache NiFi, Hive Query Language, HBase NoSQL database, Sqoop, Spark, Scala, OOZIE and Pig, UNIX shell Scripting technologies.
* Experience in working with different & complex datasets, like Flat files, JSONs, XML files and Databases, in combination big data technologies
* Experience in working with different & complex datasets, like Flat files, JSONs, XML files and Databases, in combination big data technologies
* Strong knowledge of MemSQL architecture, including sharding, replication, and distributed query processing.
* Demonstrated ability to integrate MemSQL with other AWS services, such as Amazon S3, Amazon Redshift, and Amazon Kinesis, for seamless data ingestion and analytics.
* Familiarity with MemSQL tools and utilities for administration, monitoring, and automation, such as MemSQL Ops and MemSQL Studio.
* Expertise in connecting and integrating Celonis with AWS services such as S3, Redshift, and EC2 to collect, process, and analyze data from various sources.
* Proficiency in configuring Celonis data connectors, ETL processes, and data models to ensure data accuracy and completeness.
* Design and develop interactive dashboards and reports in Angular for data analysis, insights, and monitoring of AWS data pipelines, data lakes, and data warehousing solutions.
* Knowledge of best practices and methodologies in process mining and data engineering, including Agile, DevOps, and CI/CD.
* Extract Transform and Load data from sources Systems to Azure Data Storage services using a combination of Azure Data factory, T-5QL, Spark SQL, and U-SQL Azure
* Data Lake Analytics. Data ingestion to one or more Azure services (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in Azure Databricks
* Designed and implemented a high-availability, fault-tolerant data platform using AWS and Cassandra, resulting in improvement in query performance
* Managed and automated the deployment of AWS resources and services using Terraform and Ansible, resulting in improved scalability and availability of the system.
* Utilized Informatica to integrate and transform data from various sources, including databases, flat files, and web services, resulting in a more efficient and streamlined data pipeline.
* Developed and maintained ETL workflows using Informatica PowerCenter, ensuring accurate and timely data processing and transformation.
* Developed and maintained data pipelines using Flink, AWS Lambda, and AWS Kinesis, ensuring accurate and timely data processing and transformation.
* Developed ETL data pipelines using Spark, Spark streaming and Scala.
* Loaded data from RDBMS to Hadoop using Sqoop.
* Responsible for estimating the cluster size, monitoring, and troubleshooting of the Spark databricks cluster
* Hive as the primary query engine of EMR, we have built external table schemas for the data being processed.
* AWS RDS (Relational database services) was created to serve as a Hive Meta store, and it was possible to integrate the Meta data from 20 EMR clusters into a single RDS, avoiding data loss even if the EMR was terminated.
* Utilized Kubernetes and Docker for the runtime environment for the CI/CD system to build, test,
* and deploy.
* Implemented data ingestion from various source systems using sqoop and PySpark.
* Hands on experience implementing Spark and Hive jobs performance tuning.
* Created continuous integration and continuous delivery (CI/CD) pipeline on AWS that helps to

automate steps in software delivery process

* Involved in the development of a shell script that collects and stores logs created by users in AWS S3 (Simple storage service) buckets. This contains a record of all user actions and is a good indicator of security to detect cluster termination and safeguard data integrity.
* Partitioning and bucketing ideas were implemented in the Apache Hive database, which increases query retrieval performance.
* Using AWS Glue, I designed and deployed ETL pipelines on S3 parquet files in a data lake.
* Created a cloud formation template in JSON format to leverage content delivery with cross-region replication through Amazon Virtual Private Cloud
* AWS Code Commit Repository was used to save programming logics and scripts and then replicate them to new clusters.
* Used the Multi-node Redshift technology to implement Columnar Data Storage, Advanced Compression, and Massive Parallel Processing
* Worked with Snowflake cloud data warehouse and AWS S3 bucket for integrating data from multiple source system which include loading nested JSON formatted data into snowflake table.
* Worked on the code transfer of a quality monitoring program from AWS EC2 to AWS Lambda, as well as the creation of logical datasets to administrate quality monitoring on snowflake warehouses.

**Environment:** Amazon Web Services, Elastic Map Reduce cluster, EC2s, Cloud Formation, Amazon S3, Amazon Redshift, Dynamo DB, Cloud Watch, Hive, Scala, Python, HBase, Apache Spark, Spark SQL, Shell Scripting, Tableau, Cloudera.

***Client: Nevro, Atlanta Aug 2020 to Jan 2022***

***Role: AWS Data Engineer***

***Responsibilities:***

* Developed ETL data pipelines using Spark, Spark streaming and Scala.
* Loaded data from RDBMS to Hadoop using Sqoop.
* Worked collaboratively to manage build outs of large data clusters and real time streaming with Spark.
* Responsible for loading Data pipelines from web servers using Sqoop, Kafka and Spark Streaming API.
* Used Spark for interactive queries, processing of streaming data and integration with popular NoSQL database for huge volume of data.
* Implemented Spark using Scala and SparkSQL for faster testing and processing of data.
* Data Processing: Processed data using Map Reduce and Yarn. Worked on Kafka as a proof of concept for log processing.
* Monitoring the Hive Meta store and the cluster nodes with the help of Hue.
* Created AWS EC2 instances and used JIT servers.
* Extracted data from HDFS using Hive, presto and performed data analysis using Spark with Scala, PySpark and feature selection and created nonparametric models inSpark.
* Involved in source to target implementation of ETL pipelines using Python and SQL for high volume analytics and loaded to HDFS.
* Involved in queue extraction Dstreams from near real-time data using Kafka & Spark Streaming with Scala and also have development experience using presto with EMR, Spark- Core, Context, DataFrame API, RDD and Spark SQL with build tools such as Maven & SBT
* Building servers using AWS, importing volumes, launching EC2, creating security groups, auto-scaling, load balancers, Route 53, SES and SNS in the defined virtual private connection.
* Used IAM to create new accounts, roles, and groups.
* Configured AWS Cloud Watch to monitor AWS resources, including creating AWS customized Scripts to monitor various application and system & Instance metrics.
* Configuring IAM roles for EC2 instances and assigns them policies granting specific level access to S3 buckets.
* Using Cloud Watch service, created alarms for monitoring the EC2 server’s performance like CPU Utilization, disk usage etc.
* Data Integrity checks have been handled using hive queries, Hadoop, and Spark.
* Performed end- to-end Architecture &amp; implementation assessment of various AWS services like Amazon EMR, Redshift, S3, Athena, Glue and Kinesis.
* Worked on performing transformations & actions on RDDs and Spark Streaming data with Scala.
* Defined job flows and developed simple to complex Map Reduce jobs as per the requirement.
* Using AWS Glue, I designed and deployed ETL pipelines on S3 parquet files in a data lake.
* Optimized Map/Reduce Jobs to use HDFS efficiently by using various compression mechanisms.
* Responsible in handling Streaming data from web server console logs
* Installed Oozie workflow engine to run multiple Hive and Pig Jobs.
* Developed PIG Latin scripts for the analysis of semi structured data.
* Used Hive and created Hive tables and involved in data loading and writing Hive UDFs.
* Used Sqoop to import data into HDFS and Hive from other data systems.
* Installed and configured Hive, Pig, Sqoop, Flume and Oozie on the Hadoop cluster.
* Worked on developing ETL processes (Data Stage Open Studio) to load data from multiple data sources to HDFS using FLUME and SQOOP, and performed structural modifications using Map Reduce, HIVE.
* Create PySpark frame to bring data from DB2 to Amazon S3.
* Optimize the PySpark jobs to run on Kubernetes Cluster for faster data processing.
* Optimization of Hive queries using best practices and right parameters and using technologies like Hadoop, YARN, Python, PySpark.
* Experience with Snowflake Multi-Cluster Warehouses.
* Experience with Snowflake Virtual Warehouses.
* Incredibly good understanding of Partitions, bucketing concepts in Hive and designed both Managed and External tables in Hive to optimize performance.

**Environment**: Agile, ODS, OLTP, ETL, HDFS, Kafka, AWS, S3, EC2, Python, XML, SQL, Talend, Snowflake IAM, Redshift, Glue, Lambda, MS SQL, MongoDB, Ambari, PowerBI, Git, Spark, Hive, Scala, PySpark.

***Client: Coca Cola, Atlanta, GA Apr 2018 to Jul 2020***

***Role: AWS Data Engineer***

***Responsibilities:***

* Used AWS Athena extensively to ingest structured data from S3 into other systems such as RedShift or to produce reports.
* The Spark-Streaming APIs were used to conduct on-the-fly transformations and actions for creating the common learner data model, which receives data from Kinesis in near real time.
* Performed end- to-end Architecture &amp; implementation assessment of various AWS services like Amazon EMR, Redshift, S3, Athena, Glue and Kinesis.
* Hive As the primary query engine of EMR, we have built external table schemas for the data being processed.
* Involved in the development of a shell script that collects and stores logs created by users in AWS S3 (Simple storage service) buckets. This contains a record of all user actions and is a good indicator of security to detect cluster termination and safeguard data integrity.
* Partitioning and bucketing ideas were implemented in the Apache Hive database, which increases query retrieval performance.
* Implemented data ingestion from various source systems using Sqoop and PySpark.
* Hands on experience implementing Spark and Hive jobs performance tuning.
* Using AWS Glue, I designed and deployed ETL pipelines on S3 parquet files in a data lake.
* Created a cloud formation template in JSON format to leverage content delivery with cross-region replication through Amazon Virtual Private Cloud
* AWS Code Commit Repository was used to save programming logics and scripts and then replicate them to new clusters.
* Used the Multi-node Redshift technology to implement Columnar Data Storage, Advanced Compression, and Massive Parallel Processing
* Worked with Snowflake cloud data warehouse and AWS S3 bucket for integrating data from multiple source system, which include loading nested JSON formatted data into snowflake table.
* Worked on the code transfer of a quality-monitoring program from AWS EC2 to AWS Lambda, as well as the creation of logical datasets to administrate quality monitoring on snowflake warehouses.

**Environment**: Amazon Web Services, Elastic Map Reduce cluster, EC2s, Cloud Formation, Amazon S3, Amazon Redshift, Dynamo DB, Cloud Watch, Hive, Scala, Python, HBase, Apache Spark, Spark SQL, Shell Scripting, Tableau, Cloudera.

***Client: Sutherland global services, Hyderabad, India Sep 2016 to Mar 2018***

***Role: Data Engineer***

***Responsibilities:***

* Developed various data loading strategies and performed various transformations for analyzing the datasets by using Hortonworks Distribution for Hadoop ecosystem.
* Operated with big data components like HDFS, Spark, YARN, Hive, HBase, Druid, Sqoop, and Pig.
* Applied transformations using Databricks and Spark data analysis after cleaning the data.
* Implemented data ingestion from various source systems using Sqoop and PySpark.
* Hands on experience implementing Spark and Hive jobs performance tuning.
* Involved in source to target implementation of ETL pipelines using Python and SQL for high volume analytics and loaded to HDFS.
* Involve in creating database objects like tables, views, stored procedures, triggers, packages, and functions using T-SQL to provide structure and maintain data efficiently.
* Developed a reconciliation process to make sure the Elastic Search index document count matches source records using Python Flask specification.
* Developed data ingestion pipeline from HDFS into AWS S3 buckets using Knife.
* Created external and permanent tables in Snowflake on the AWS data.
* Extensive experience in Snowflake administration, including user management, security implementation, data ingestion, extraction, and query optimization.
* Used Flume to collect, aggregate, and store the weblog data from different sources like web servers, mobile and network devices, and pushed to HDFS.
* Migrated data between traditional RDBMS and HDFS using Sqoop.
* Ingested data into HDFS from Teradata, Oracle, and MySQL. Identified required tables and views and exported them into Hive.
* For faster access of data, used performed ad-hoc queries using Hive joins, partitioning, bucketing technique
* Responsible for collecting, scrubbing, and extracting data from various generate reports, dashboards, and analytical solutions. Helped in debugging the Tableau dashboards.
* Worked in Agile development environment and participated in daily scrum and other design related meetings.

**Environment**: Python, PySpark, Kafka, PyCharm, AWS (EMR, EC2, S3), Data Lake, Snowflake, Hive, Flume, SQL, Sqoop v1.4.4, Oozie v4.1.0, Oracle, SQL Server, Tableau, Agile Methodology, Hadoop, HDFS, Spark.

***Client: Kalstride InfoTech - Hyderabad, India Jun 2014 to Aug 2016***

***Role: Data Analyst***

***Responsibilities:***

* Performed Data Analysis using SQL queries on source systems to identify data discrepancies and determine data quality.
* Performed extensive Data Validation, Data Verification against Data Warehouse and performed debugging of the SQL-Statements and stored procedures for business scenarios.
* Designed and developed Tableau dashboards using stack bars, bar graphs, scattered plots, and Gantt charts.
* Familiar with DBMS table design, loading, Data Modeling, and experience in SQL.
* Worked on ER/Studio for Conceptual, logical, physical data modeling and for generation of DDL scripts
* Handled performance requirements for databases in OLTP and OLAP models.
* Analyzed the data which is using the maximum number of resources and made changes in the back-end code using PL/SQL stored procedures and triggers
* Performed data completeness, correctness, data transformation and data quality testing using SQL.
* Involved in designing Business Objects universes and creating reports.
* Conducted design walk through sessions with Business Intelligence team to ensure that reporting requirements are met for the business.
* Prepared complex T-SQL queries, views and stored procedures to load data into staging area.
* Wrote UNIX shell scripts to invoke all the stored procedures, parse the data and load into flat files.
* Created reports analyzing large-scale database utilizing Microsoft Excel Analytics within legacy system.

**Environment:** SQL, PL/SQL, OLAP, OLTP, UNIX, MS Excel, T-SQL.