**Srinivasa Rao** Plano, TX 75025 ♦ 832-410-3706 ♦ srinivas.data1464@gmail.com

**BIG DATA ENGINEER**

Big Data Enthusiastic with a strong passion to learn and invent new programming techniques and environments

that could build and leverage data lakes and achieve improved outcomes.

Professional Summary

* 9+ years of experience with handling large datasets by designing complex frameworks and algorithms using **Hadoop, Big Data, AWS services, RDBMS databases** and **Business Intelligence Tools**
* Strong experience in performing Data Cleansing, Data Wrangling and Data Masking with Bigdata ETL frameworks built using **SPARK** and **SCALA**.
* Demonstrated expertise in configuring, managing, and optimizing **Apache Spark** environments for high-volume data processing and analytics.
* Extensive experience with **MongoDB**, including database design, query optimization, and performance tuning, along with a good understanding of other NoSQL databases.
* Hands-on experience with the **STRIIM** platform for real-time data integration, streaming analytics, and ETL processes
* Deep knowledge of BI tools such as **PowerBI**, **Cognos**, and **Tableau** for data visualization and reporting, aiding in data-driven decision-making.
* Familiarity with **Azure Kubernetes** for deploying, managing, and scaling containerized applications, along with an understanding of cloud infrastructure and services.
* Hands on Experience in working with Hadoop ecosystem components like **Hive, HDFS, Pig, Sqoop, Map Reduce, Oozie.**
* Experience in building framework using **PySpark** to extract data from **PostgreSQL** database and saves output file in s3 bucket with data masking and reusability capabilities.
* Model, lift and shift custom SQL and transpose LookML into **dbt** for materializing incremental views.
* Designed, build and managed ELT data pipeline, leveraging **Airflow, python, dbt, Stitch Data** and **AWS solutions**.
* Hands on expertise with **AWS** services such as **EMR**, **EC2**, **S3**, **Redshift** and **IAM**.
* Proficient in using Bigdata tools such as Pig & Hive for data analysis, **Qlik Enterprise Manager** for data ingestion, **Airflows** for scheduling and **Zookeeper** for coordinating cluster resources.
* Experience in handling structured, unstructured and semi-structured data using various Hadoop file  
  formats like **Parquet, ORC, AVRO, dat, text, Json, CSV and deflate**.
* Experience in writing extensive **Snow-SQL** queries to do transformations on the data to be used by  
  downstream models.
* Experience in migrating on Premises ETL process (Teradata) to Cloud (**Snowflake** with AWS).
* Experience in designing complex workflows and schedule using **Airflows, Arow and Control-M.**
* Ability to work on complex **data structures,** dashboards and ad hoc reporting.
* Strong experience in writing custom **shell scripts** to handle adhoc requirements for **inbound and outbound**file transfers.
* Good knowledge on **Agile** Methodology and scrum process.
* Astute with data-driven and technology-focused approach. Communicates clearly with stakeholders and builds consensus around well-founded models. Talented in writing applications and reformulating models.

Skills

|  |  |
| --- | --- |
| **Databases:** Teradata 14, Oracle 10g, SQL Server, DB2, MS Access, Snowflake, Mongo DB, Cassandra.  **Big Data Ecosystem:** HDFS, MapReduce, PIG, HIVE, Spark, Sqoop, Oozie, Zoo Keeper. STRIIM  **ETL Tools:** SQL Server Integration Services, Talend, Data Stage, DBT, Qlik Enterprise Manager, Five Tran, Snow Pipe  **Languages:** Scala, Python, Java, SQL, PLSQL, MDX. | **Schedulers:** Arow, Control-M, Oozie, Crontab, Airflows, Azure Kubernetes services,  **Cloud Services:** AWS EMR, EC2, GCP, Simple Storage System(S3), IAM.  **Methodologies:** Agile, Waterfall.  **CI/CD Tools:** GIT-HUB, Jenkins, Puppet, Chef. |

**Databases :** Teradata 14, Oracle 10g, SQL Server, DB2, MS Access, Snowflake, Mongo DB, Cassandra.

**Big Data Ecosystem :** HDFS, MapReduce, PIG, HIVE, Spark, Sqoop, Oozie, Zoo Keeper.

**ETL Tools:** Sql Server Integration Services, Talend, Data Stage, DBT, Qlik Enterprise Manager, Five Tran, Snow Pipe

**Languages :** Scala, Python, Java, SQL, PLSQL, MDX.

**Schedulers :** Arow, Control-M, Oozie, Crontab,Airflows

**Cloud Services :** AWS EMR, EC2, Simple Storage System(S3), IAM.

**Methodologies :** Agile, Waterfall.

**CI/CD Tools :** GIT-HUB, Jenkins, Puppet, Chef.

Work History

Big Data Developer, 01/2023 - Current

LGL Technologies Inc / Wex Inc– Dallas, TX

* Working on building data pipelines for transforming the streaming data from internal and external sources and loading it into the AWS S3 data lake and then to snowflake data warehouse
* Working on creating a framework that detects and masks the Customer's Non-public Personal Information (NPI data) and Payment Card Industry (PCI data) that resides in AWS S3 and Snowflake warehouse
* Led the configuration and administration of Apache Spark environments, optimizing performance and scalability for large-scale data processing.
* Designed and implemented efficient queries and stored procedures in SQL Server and Oracle, achieving significant improvements in data retrieval times.
* Developed and maintained robust data pipelines using MongoDB and the STRIIM platform, ensuring real-time data accuracy and availability.
* Utilized Python to automate data processing tasks, enhancing operational efficiency and reducing manual errors.
* Worked on building Data pipelines, end to end ETL process for ingesting data in GCP.
* Created insightful visualizations and dashboards with Power BI, Cognos, and Tableau, facilitating data-driven decision-making across the organization
* Collaborated closely with cross-functional teams to develop and debug solutions, adhering to best practices and design documents.
* Co-ordinated with in configuration (Source & Aiven Connectors) setup of Kafka for real-time streams into S3 & Snowflake
* Design, develop, and maintain scalable data models and transformations using DBT in conjunction with Snowflake, ensure the effective transformation and load data from diverse sources into data warehouse or data lake
* Implement and manage data models in DBT, guarantee accurate data transformation and alignment with business needs
* Utilize DBT to convert raw, unstructured data into structured datasets, enabling efficient analysis and reporting
* Write and optimize SQL queries within DBT to enhance data transformation processes and improve overall performance.
* Establish best DBT processes to improve performance, scalability, and reliability.
* Built Snow pipe pipelines for continuous data load to AWS S3 and Snowflake Datawarehouse
* Working on validating the data between SDP (Streaming Data Platform) and AWS S3 to check whether there are any data gaps (Data missing & Data mismatch)
* Worked on the ETL jobs (FiveTran,qlik enterprise manager) to migrate data from on premise to AWS cloud S3 and snowflake by generating JSON and CSV files to support Catalog API integration.
* Partitioned the fact tables and materialized views to enhance the performance.
* Extensively used bulk collection in PL/SQL objects for improving the performing.
* Understanding of structured data sets, data pipelines, ETL tools, data reduction, transformation and aggregation technique, Knowledge of tools such as DBT
* Experience on DBT provides a few benefits for data engineering teams. It allows data engineers to write modular, reusable code using SQL, which can be version controlled and tested like any other software code. DBT also provides several built-in features for data modeling, such as automatic type inference, schema management, and data lineage tracking
* Working with PySpark for encrypting sensitive Data residing in the history and the ongoing data set files
* Created a wrapper shell script for each of the framework developed in PySpark and provided it as an input to the Airflows jobs.
* Experience in creating Data Factory Pipelines, custom Azure development, deployment, troubleshoot data load / extraction using ADF.
* Created a shell script to load the historical data which resides in S3 buckets to snowflake.
* Setup the replication and clone for large tables by splitting into multiple tables based on the partitions to migrate data into Snowflake data warehouse
* Supported the migration of databases to Azure Kubernetes, optimizing for high availability and resilience.
* Used Maximized Warehouse cluster while running the queries and tested with different queries each time using multiple warehouse sizes

**Environment:** Scala, Java, Snow SQL, Python, JSON, Snowflake, SQL, Airflows, Qlik enterprise manager, Apache Spark, Five Tran, Snow pipe, Shell Scripting, GIT, GCP, AWS services (S3, EMR,EC2 &IAM), Nebula, Exchange, Agile. Azure Kubernetes,Spark

Big Data Developer, 06/2017 - 11/2021

HDFC – India

* Constructed a data pipeline to process semi-structured data by incorporating 100 million raw records from 14 data sources
* Designed the data pipeline architecture for a new product that quickly scaled from 0 to 60,000 daily users
* Integrated data from multiple third-party APIs that provided data around local language preferences, leading to customized landing pages that improved paid conversion rate by 6%.
* Ingested streaming and transactional data across 9 diverse primary data source using Spark, Redshift, S3, and Python
* Created Python library to parse and reformat data from external vendors, reducing error rate in the data pipeline by 12%.
* Automate ETL processes across billions of rows of data, which saved 45 hours of manual hours per month
* Build the Logical and Physical data model for snowflake as per the changes required
* Experience performing root cause analysis on internal and external data and processes to answer specific business questions.
* Worked on Oracle Databases, RedShift and Snowflakes
* Bulk loading from the external stage (AWS S3), internal stage to snowflake cloud using the COPY command.
* Used import and Export from the internal stage (snowflake) from the external stage  
  (AWS S3).
* Used cloud shell SDK in GCP to configure the services Data Proc, Storage, Big Query.
* Performed data quality issue analysis using Snow SQL by building analytical warehouses on Snowflake
* Experience with building processes supporting data transformation, data structures, metadata, dependency and workload management.
* Responsible for estimating the cluster size, monitoring and trouble shooting with Spark Databricks cluster.
* Designed and developed scalable solutions for storing and processing large amounts of data across multiple regions.
* Analyzed the business requirements and translate them into technical specifications that can be used by developers to implement new features or enhancements.
* I have experience on DBT provides a few benefits for data engineering teams. It allows data engineers to write modular, reusable code using SQL, which can be version controlled and tested like any other software code. DBT also provides several built-in features for data modeling, such as automatic type inference, schema management, and data lineage tracking
* Provided support during all phases of development including design, implementation, testing, deployment and maintenance of applications/services.
* Participated in cross-functional teams (e.g., infrastructure engineering) when required to ensure effective communication between groups with overlapping functionality or shared resources.
* Developed and implemented data pipelines using AWS services such as Kinesis, S3, EMR, Athena, Redshift to process petabyte-scale data in real time.
* Implemented a data warehouse using Redshift to store and analyze terabytes of raw data
* Built ETL processes in Python, Pig, and SQL to transform unstructured data into structured datasets
* Developed an automated machine learning system that reduced manual labor by 80%
* Created custom dashboards with Tableau for real-time monitoring of key business metrics
* Spearheaded the migration from on-premise servers to AWS cloud infrastructure (EC2, S3, RDS)
* Conducted data analysis to support business decision-making by extracting, cleansing, and manipulating data from various sources.
* Created data visualizations to communicate complex data sets in an easily understandable format for business users.

SQL/BI Developer, 05/2014 - 05/2017

Riosoft Technologies – India

**Responsibilities:**

* Designed, developed and maintain BI solutions using SQL, including data warehouse and data mart structures.
* Created and maintained SQL-based ETL (Extract, Transform, Load) processes to extract, clean and load data into the data warehouse.
* Collaborated with stakeholders and other teams to understand business requirements and design BI solutions that meet their needs.
* Created and maintained SQL-based reports and analytics using tools like SSRS (SQL Server Reporting Services), Power BI and Tableau Monitored installation and operations to consistently meet customer requirements.
* Created data models and design database schemas to support reporting and analytics.
* Troubleshooted and debugged BI issues, identifying and resolving data and performance issues.
* Optimized SQL queries for performance and scalability.
* Continuously improved the BI development process by researching and experimenting with new tools and technologies.
* Worked with other BI developers and IT teams to design, develop and implement security, backup and recovery procedures.
* Provided technical guidance and mentorship to other team members, and act as a subject matter expert on BI development using SQL.

**Environment:** SQL Server Business Intelligence Development Studio, PL/SQL, SQL Server, Oracle, Power BI,Tableau, MS Office, Windows

Education

Bachelor of Science: Bachelor of Engineering, 2014

Acharya Nagarjuna University - India

Master of Science: Information Technology, 12/2022

Auburn University-Montgomery - Montgomery, AL