**Srinivas Raju Upendram**

**Sr. Data Engineer**

**AWS, Snowflake, Apache**

Phone: 945 218 2002

Email: srinivasraju.0020@gmail.com

**PROFESSIONAL SUMMARY:**

* About a decade of IT 11+experience in software analysis, design, development, testing and implementation of Big Data, Hadoop, SQL, No SQL technologies.
* Experience in designing, developing, and optimizing scalable data solutions using Big Data, Hadoop, and NoSQL technologies.
* Proficient in Apache Spark (batch and streaming), Scala, PySpark, and Spark SQL for high-performance data processing and transformation.
* Extensive experience with Hadoop Ecosystem tools: MapReduce, Hive, Pig, HBase, Kafka, Flume, Sqoop, Oozie, and Zookeeper.
* Specialized in building ETL pipelines and CDC (Change Data Capture) integration with Debezium.
* Expertise in ensuring efficient data movement across heterogeneous data sources: AWS S3, Azure Blob Storage, and SQL/NoSQL databases (SQL Server, MongoDB, Cassandra, DynamoDB).
* Hands-on experience with Databricks, Snowflake (SnowSQL, Snowflake SQL), and AWS services (EMR, S3, Glue, Lambda, Step Functions) for seamless data processing and cloud-based solutions.
* Data Warehousing and Dimensional Modeling (Star and Snowflake schemas); skilled in OLAP and using Tableau and Cognos for insightful reporting and dashboards.
* Expertise in Data Preprocessing, Machine Learning, and Data Analysis using Python, R, SQL, and Microsoft Excel for data mining, cleansing, and transformation.
* Proficient in managing complex workflows with Apache Airflow, Oozie, and containerization technologies like Docker and Kubernetes.
* Proven ability to optimize data pipeline performance at scale and ensure efficient data hydration in data lakes.
* Strong foundation in security best practices, performance tuning, and data quality processes using Apache Hudi and Apache Griffin.
* Experienced in infrastructure automation and configuring Cloudera Hadoop platforms and supporting data governance frameworks.
* Passionate about delivering high-quality, clean, and query-able datasets for advanced analytics, decision-making, and business intelligence.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Bigdata Ecosystem** | Hadoop, MapReduce, Pig, Spark, Hive/Impala, YARN, Kafka, Flume, Sqoop, Zookeeper, Airflow, MongoDB, Cassandra, HBase, and Storm. |
| **Hadoop Distribution** | Cloudera distribution, Horton works. |
| **Programming Languages** | Python (Pandas, NumPy, SciPy, Scikit-Learn, Seaborn, Matplotlib), Oracle, T-SQL, PL/SQL, Scala, Spark, Hibernate, JDBC, HTML, CSS, Java, Javascript |
| **Scripting Languages** | Python, Azure Power Shell |
| **Databases** | Oracle 11g/10g/9i, HBase, SQL, Cassandra, MongoDB, MySQL, Redshift, ADLS |
| **Cloud** | AWS, Azure, GCP |
| **ETL, Scheduling Tools** | Snowflake, Informatica, Databricks,Talend, Control-M, Active Batch, Zena |
| **All Other Tools** | GIT, Subversion, Tableau, Power BI, SSMS |

**Professional Experience:**

**Tredence, San Jose Jun 2022 to Present**

**Sr. Data Engineer /ETL Specialist**

**Responsibilities:**

* Designed, developed, and supported scalable data solutions and APIs, enabling rapid delivery of business capabilities.
* Collaborated with IT application teams, enterprise architecture, infrastructure, information security, and LOB stakeholders to translate business and technical strategies into data-driven solutions.
* Acted as a technical expert, resolving issues related to system performance, integration, security, and application design in complex data solutions.
* Conducted research on emerging trends in banking, data engineering, data security, and architecture to implement cutting-edge data solutions.
* Developed CI/CD pipelines, self-service build tools, and automated deployment processes to enhance operational efficiency.
* Evaluated and recommended software products, providing documented analysis for stakeholders to optimize data platform performance.
* Provided on-call troubleshooting support for critical incidents, ensuring seamless operations for data platforms and quick resolution of issues.
* Contributed to internal projects, including legacy system replacement, data monitoring improvements, analytics enhancements, and tool development.
* Delivered technical guidance and mentorship to junior engineers, promoting best practices in data engineering and application design.
* Built reusable C# components for data validation, cleansing, and enrichment, improving data quality by 35%
* Managed and prioritized multiple assignments, delivering high-quality data solutions within deadlines.
* Developed Perl scripts to automate ETL workflows, handling extraction, transformation, and loading of data across heterogeneous systems.
* Developed robust ETL pipelines using C# and .NET, automating data ingestion from various APIs and flat files into SQL Server and Azure Data Lake.
* Designed and implemented scalable ETL pipelines using Java, handling structured and unstructured data across distributed systems.
* Built reusable Perl modules to standardize data validation, cleansing, and transformation across data pipelines.
* Developed cloud-based data solutions on Snowflake, ensuring efficient data storage, processing, and security for large-scale applications.
* Implemented application and data security best practices, ensuring compliance with industry standards for data management.
* Leveraged big data technologies (e.g., Hadoop, Spark), metadata management, ETL tools, and data warehousing to support large-scale data platforms.
* Designed and developed REST APIs for seamless data integration and retrieval across various systems.
* Involved in file movements between HDFS and AWS S3 and extensively worked with S3 bucket in AWS.
* Enabled speedy reviews and first mover advantages by using Oozie to automate data loading into the Hadoop Distributed File System and PIG to pre-process the data.
* Extensive experience in working with AWS cloud Platform (EC2, S3, EMR, Redshift, Lambda and Glue).
* Utilized Java Streams, Collections, and multithreading to efficiently process large datasets and optimize data flow operations
* Worked on a POC on Spark and Scala parallel processing. Real streaming the data using Spark with Kafka.
* Used No SQL DB, Querying data from MongoDB and use them as input for machine learning models.
* Supported MapReduce Programs running on cluster. Cluster monitoring, maintenance and troubleshooting.

**Environment:** AWS, Hadoop, Python**,** Spark, HDFS, Java, Hive, C#, Pig, HBase, Tableau, Big Data, Apache Storm, Oozie, Sqoop, Kafka, Flume, Zookeeper, MapReduce, Mongo db, Scala, Linux, No SQL, MySQL, Oracle, Git, Jenkins.

**KPMG, WA Aug 2020 to May 2022**

**Sr.Data Engineer**

**Responsibilities:**

* Designed & Implemented complex applications & distributed systems into Cloud Infrastructure GCP, Azure.
* Used to manage Git lab and Bit Bucket account for providing access to developers and storing source code.
* Led the design, development, and implementation of scalable data engineering solutions, ensuring efficient data processing and transformation across multiple platforms.
* Collaborated with cross-functional teams (data scientists, analysts, business stakeholders) to gather requirements and translate business needs into technical solutions.
* Used C# with AWS SDKs (S3, Kinesis, DynamoDB) to automate data ingestion and metadata tracking workflows.
* Engineered and optimized ETL pipelines using Apache Spark, PySpark, and SQL for data extraction, transformation, and loading from diverse sources into data lakes and data warehouses.
* Developed and managed cloud-based data platforms on AWS and Azure, utilizing services like S3, Glue, Redshift, and SQL Server to support seamless data storage, processing, and analytics.
* Implemented real-time data streaming and Change Data Capture (CDC) using Kafka, ensuring timely and accurate updates for transactional systems.
* Conducted data integration and transformation tasks from multiple sources such as RDBMS, NoSQL databases, and cloud storage systems, ensuring smooth data flow across platforms.
* Built real-time data processing applications in Java using Apache Kafka, Flink, and Spark Streaming to support low-latency analytics.
* Built and maintained data models, focusing on dimensional modeling (Star and Snowflake schemas) to support business intelligence and reporting.
* Designed and developed data pipelines to integrate large-scale datasets into business intelligence tools such as Tableau, Power BI, and Cognos, delivering actionable insights.
* Integrated C# applications with Kafka and RabbitMQ for near real-time data processing and event-driven architecture.
* Led data security initiatives by implementing best practices in data encryption, access control, and compliance across all data platforms and systems.
* Improved data quality through validation, cleansing, and transformation techniques, ensuring reliable and consistent datasets for reporting and analytics.
* Designed Perl utilities to interface with APIs and FTP/SFTP servers for automated data retrieval and ingestion.
* Optimized query performance and data processing efficiency, reducing data processing time and ensuring high availability and fault tolerance of data systems.
* Developed scalable C# utilities to process millions of rows from distributed sources and load into Hadoop-based systems using REST APIs.
* Worked closely with the DevOps team to integrate CI/CD pipelines for automating the deployment of data solutions, ensuring faster release cycles and minimal downtime.
* Documented and transitioned legacy Perl data workflows as part of larger system upgrades or platform migrations.
* Developed custom Kafka producers and consumers in Java to facilitate streaming ingestion and downstream microservice communication.
* Mentored junior team members, providing guidance on best practices for data engineering, coding standards, and problem-solving techniques.
* Participated in client-facing meetings, presenting data solutions and providing insights into how data engineering could drive business outcomes.

**Environment:** Azure, GCP, Python**,** Hadoop, Spark, ETL, Java ,C#, PySpark, HDFS, Hive, Pig, HBase, Big Data, Oozie, Sqoop, Kafka, Flume, Zookeeper, MapReduce, MongoDB, Scala, No SQL, MySQL, Mongo db, Oracle, Git.

**State Farm, IL May 2018 to Jun 2020**

**“AWS Data Engineer**

**Responsibilities:**

* Used Spark Streaming APIs to perform transformations and actions on the fly for building common.
* Implemented Spark RDD transformations, actions to implement business analysis and Worked with Spark accumulators and broadcast variables. Installed and configured Hadoop MapReduce, HDFS, HIVE, PIG, SQOOP, Flume, OOZIE on the Hadoop cluster.
* Enhanced performance of PySpark jobs on AWS EMR clusters by optimizing job execution time and resource utilization through techniques like partitioning, caching, and broadcast variables.
* Utilized Redis data structures including sets, lists, and hashes to construct intricate data structures and algorithms, facilitating streamlined data manipulation and retrieval processes.
* Implemented Redis Sentinel for ensuring high availability and automatic failover, guaranteeing uninterrupted access to critical data and applications.
* Automated data ingestion from APIs, DB into AWS S3 buckets using Python scripts for downstream analytics.
* Designed and improved SQL queries and scripts to execute intricate data manipulations, aggregations, and joins on AWS Redshift, thereby boosting query performance and system efficiency.
* Established automated data archival and retention policies on Snowflake, employing time-based lifecycle.
* Implemented Spark jobs in Java to process batch and streaming data from various sources including S3, HDFS, and Kafka.
* Conducted performance analysis, optimization of Snowflake workloads by evaluating query execution plans, storage utilization, and resource consumption to identify areas for enhancement.
* Managed continuous integration and continuous deployment (CI/CD) pipelines to automate deployment, testing of AWS Databricks solutions, utilizing platforms such as Jenkins, GitHub Actions, and AWS CodePipeline.
* Loaded data into S3 buckets using AWS Glue and PySpark. Involved in filtering data stored in S3.
* Assisted in migrating from On-Premises Hadoop Services to cloud-based Data Analytics using AWS.
* Created monitors, alarms, notifications for EC2 hosts using AWS Cloud Watch, Athena, AWS Glue, Cloud trail, Docker, SNS and Implemented generalized solution model using AWS Sage Maker.
* Developed end to end ETL batch and streaming data integration into Hadoop (MapR), transforming data.
* Created ODI interfaces on Snowflake DB. worked on AWS Redshift for shifting all Data warehouses into one.
* Developed custom data flow using Apache Nifi to fully automate ETL process by taking scenarios into account.
* Used Hive QL to analyze the partitioned and bucketed data, Executed Hive queries on Parquet tables.
* Developed scripts, UDF's using both Spark SQL and Spark Core in Scala for Data Aggregation, queries and verified its performance over MR jobs.
* Used Apache Kafka to aggregate web log data from multiple servers and make them available in Downstream systems for Data analysis and engineering type of roles.
* Installed and configured Hive and written Hive UDFs and Used Map Reduce and Junit for unit testing.

**Environment:** Python, AWS,Hadoop, ETL, Java, MapReduce, Spark, PySpark, HDFS, Hive, Pig, HBase, Big Data, Oozie, Sqoop, Scala, Kafka, Flume, Zookeeper, MapReduce, Spark SQL, Cassandra, Tableau, Scala, Unix, REST.

**Nationwide Financial, OH Feb 2017 – Apr 2018**

**Sr. Data Engineer**

**Responsibilities:**

* Created Pipelines in ADF using Linked Services, Datasets to Extract, Transform load data from different sources like Azure SQL, Blob storage, Azure Synapse, Azure SQL Data warehouse, write-back tool.
* Migrating on-prem ETLs from MS SQL server to Azure Cloud using Azure Data Factory and Databricks
* Involved in Creating, Debugging, Scheduling and Monitoring jobs using Airflow for ETL batch processing to load into Snowflake for analytical processes.
* Developed custom data ingestion pipelines using Azure Data Factory (ADF) to extract data from Hadoop HDFS, transform it using Azure Databricks (ADB), and load it into Azure Blob Storage and ADLS Gen2.
* Proficient in integrating PySpark with Azure Data Factory to orchestrate comprehensive data workflows, manage PySpark job scheduling, and oversee job execution monitoring.
* Enhanced HiveQL queries and scripts to transform, aggregate, and analyze extensive datasets in Hadoop HDFS, resulting in improved query performance and reduced execution times.
* Skilled in crafting and executing data transformation and cleansing procedures using Apache Spark and Azure Databricks to prepare data for subsequent analytics and reporting tasks.
* Developed sophisticated MapReduce jobs in Java and Scala to process and analyze vast datasets in Hadoop Distributed File System (HDFS), achieving notable enhancements in performance.
* Orchestrated real-time data processing workflows by integrating Apache Kafka with Hadoop clusters, enabling seamless ingestion and processing of streaming data for timely insights and analytics.
* Used T-SQL functions and stored procedures to encapsulate business logic and data transformation regulations, facilitating efficient processing within PySpark pipelines.
* Optimized Hadoop clusters for resource management, fine-tuning YARN, configuring Hadoop schedulers.
* Managed data pipelines with Apache Oozie to orchestrate intricate workflows across various Hadoop ecosystem components, ensuring consistent and reliable data processing.
* Developed MapReduce algorithms in Scala to execute graph analysis, pattern recognition, machine learning.
* Established data ingestion pipelines with Apache Kafka and Apache NiFi to stream data into Hadoop clusters, enabling real-time data processing and analytics.
* Utilized Kafka Streams interactive queries to enable real-time analytics and dashboard applications, empowering users to directly query and analyze streaming data within Hadoop ecosystems.
* Employed Kafka Connect JDBC and HDFS connectors to bring in data from external databases and files into Kafka topics, facilitating a seamless integration of diverse data sources with Hadoop ecosystems.
* Played major role in Migration of DAGs from legacy Airflow to Managed Airflow platform
* Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure Data Lake Analytics, Azure SQL, Azure DW) and processing the data in Azure Databricks.
* Automated process in Azure cloud which can ingest data daily from web service and load into Azure SQL DB.
* Using apache NiFi to automate the data movement between Hadoop systems
* Configured Snowflake Spark Streaming to receive real time data from Kafka to store the stream data to HDFS.
* Developed Streaming pipelines using Azure Event Hubs and Stream Analytics to analyze data for efficiency and open table counts for data coming in from IOT enabled poker and other pit tables.
* To meet specific business requirements wrote UDF’s in Scala and Store procedures.
* Involved in using ScalaTest Funsuite Framework for developing Unit Tests cases and Integration testing.

**Environment:** Azure, Hadoop, HDFS, Hive, MapReduce, Python, Pyspark, Spark, Kafka, Cosmos, ETL, YARN, Scala, Power BI, SQL, Git.

**Cooper Technologies, India Jul 2013 to Nov 2016**

**SQL Developer/ linux Administrator**

**Responsibilities:**

* Installation, Configuration & Upgrade of Solaris and Linux operating system.
* Installed & configured Sun Fire 210, V240, V440, V490, V880, 1280 & HP Proliant DL- 320/360/380/560/580 L Servers.
* Used to import and export data from various CSV files, Flat files, Excel spread sheets and SQL Server
* Designed and developed several types of reports like matrix, tabular, chart reports using SSRS
* Expertise in designing and scheduling complex SSIS packages for transferring data from multiple data sources to SQL Server
* Problems & performance issues; deployed latest patches for Sun, Linux and Application servers, Performed Solaris and Red Hat Linux Kernel Tuning for TCP stack.
* Installation Red Hat Linux 8, 9.0 on Dell, Proliant, Enterprise Server 4.0.
* Installation of Web sphere, upgraded to service pack updates, installed IBM patches, configuring and creation new admin & managed servers, start & stop Web sphere server.
* Linux kernel memory upgrades and swaps area. Red hat Linux Kickstart Installation Sun Solaris Jump Start Installation. Configuring DNS, IDNS, DHCP, NIS, NFS in Sun Solaris 8/9 & other Network services.
* Written shell scripts and Perl scripting to take backup oracle database.

**Environment**: Sun Solaris 8/9, Red Hat Linux 8,9.0, Samba, Sun fire (210, V240, V440, V490, V880, 1280) Servers, HP DL-320/360/380/560/580L servers, SQL Server, JBoss, EMC storage devices, Tivoli Storage Manager 5.x, web sphere, Oracle 8/9i/RAC, Apache 2.x/3.x

**CERTIFICATIONS:**

* **Microsoft Azure Data Engineer Associate**
* **Snowflake Data Engineer Core**
* **Databricks Lakehouse Data Enginee Associate**

**KEY ACHIEVEMENTS:**

**Detection of potholes in road surface**

* Collected images of potholes and perform data labeling and data cleaning using Microsoft Vott.
* Trained the labeled images on SSD Inception V2 Coco model using TensorFlow and Google Colab.
* Tested the model on video of potholes on the trained model using Python, TensorFlow and OpenCV.
* Speed Analysis using various Frames per Seconds of video to detect maximum speed, detect all potholes.

**Image Recognition using Convolutional Neural Networks**

* Goal is to detect Logos using brand logo recognition system using Computer Vision.
* Flickr27 Logo Dataset is collected from Kaggle which has 27 classes.
* Performed Image Cropping and Image Argumentation for dataset prepossessing.
* Trained with accuracy of 94% by implementing Convolutional Neural Networks using TensorFlow in Python.
* Tested the model of test images to get output which is displayed in 2 formats.

**EDUCATION:**

B.Tech in computer science from CVR College of engineeeirng