| Sai Kiran Adepu  Mobile (+1 (469) 846-8345)  Email: [saisri@mastroservices.com](mailto:saisri@mastroservices.com)  Linked In: [SaiKiran Adepu](https://www.linkedin.com/in/saikiran-adepu-4612391a1?utm_source=share&utm_campaign=share_via&utm_content=profile&utm_medium=ios_app) |  |
| --- | --- |
| Career Objectives | |
| * Dedicated IT professional with 9 years of experience in the analysis, design, development, and optimization of large-scale data processing solutions. Seeking a challenging role where my expertise in big data technologies, including Hadoop, Spark, and cloud platforms like AWS, Azure, can be leveraged to drive innovation and efficiency. Proven track record of success in deploying and managing Cloudera and Hortonworks distributions, optimizing data workflows, and implementing real-time data processing pipelines. Eager to contribute my skills in data analysis, optimization, and workflow automation to a dynamic team, fostering continuous improvement and achieving organizational objectives.  |  | | --- | | Skills | | * **Big Data Technologies:** Hadoop (HDFS, MapReduce, HBase), Spark, Hive, Sqoop. * **Cloudera Distribution:** Proficient in CDH 6.x versions. * **ETL Operations:** Implemented data import/export operations between HDFS and RDBMS using Sqoop. Contributing to NIFI data pipeline projects designed to efficiently handle and process vast datasets, with a specialized focus on configuring Lookups for rigorous data validation and integrity checks. Developed Spark/Scala applications for executing intricate ETL operations with precision, ensuring reliable data movement across platforms like Snowflake, Azure Delta Lake, blob storage, and external dashboards. * **Data Processing / Validation:** Developed analytical components using Spark, Scala, Python, Apache NIFI and Spark Streaming. Utilized AWS Services like AWS Glue, EC2, S3, Databricks notebooks for validation, ensuring the integrity of data inbound and outbound from external sources like Amperity. * **Programming/Scripting:** Spark, Python, PySpark, Unix Shell, Hive, HQL, SQL, HDFS. * **Cloud Platforms:** Extensive experience in AWS including Glue, EC2, EMR * **Testing Support:** Active participation in test case design, testing, and documentation, contributing to the seamless integration of enhancements and optimizations in large-scale data environments. * **Environments:** Unix Shell, Hive, HQL, SQL, Apache NIFI, HDFS, Linux, Sqoop, Python, Boto3, Pandas, NumPy, Ambari/Yarn, Cloudera, Spark, Spark SQL, Kubernetes, VS Code, AWS Glue, S3, EMR, EC2, RedShift, Lambda, CloudFormation, Athena, CloudWatch, IAM, Azure Data Lake, Azure Data Factory, Data Bricks, SQL Server, Hadoop, Snowflake, GitHub Actions, Bit Bucket, Jira. | | |
|  | |
| Education: | |
| **Masters in Software Engineering- Stratford University, 2017.**  **Bachelor of Technology- JNTU Hyderabad, 2014.** | |
|  | |
| Relevant Experience | |
| **Sr. Cloud Engineer (ETL)**  Vanguard, Malvern, PA  July 2024 – Till Date   * Designed and implemented ETL pipelines using AWS Glue to run SQL queries on S3 data sources, transforming and writing processed data to target S3 buckets and Collibra Data Market place. * Utilized Troposphere to programmatically generate CloudFormation templates (CFT) for AWS resources, enabling repeatable and automated infrastructure provisioning. * Utilized AWS Athena to perform data validation on datasets stored in S3, ensuring data accuracy and consistency for downstream processes. * Optimized data validation processes in S3 by leveraging AWS Athena’s policy limits to ensure efficient query execution and cost management. * Configured and monitored EMR clusters for large-scale data processing, implementing Spark and Hadoop jobs to streamline big data workloads. * Automated IAM policy generation and deployment using Python scripts, reducing manual errors and improving efficiency. * Supported the RDW migration from Db2 to AWS Redshift and S3 by assisting in data extraction, transformation, and validation to ensure successful transitions. * Implemented AWS Event Bridge (formerly CloudWatch Events) to create event-driven architectures, automating responses to detect anomalies or any job failures and triggering AWS Lambda functions / SNS for real-time actions. * Integrated GitHub and Bitbucket repositories into automated pipelines to ensure consistent code quality and streamlined version control.   **Sr. Cloud Data Engineer (ETL)**  BlueCross BlueShield Association, Chicago, IL  May 2021 – July 2024   * Writing and optimizing complex SQL queries to extract valuable insights from massive datasets, ensuring accurate and timely reporting. * Skilled in identifying and rectifying data anomalies, errors, and inconsistencies through data profiling and cleansing techniques having implemented in SQL/ Python Scripts for Rule Engine performs defined rules on Ingested Data. * Implemented security best practices and compliance frameworks to protect cloud resources and data, including network security, identity and access management (IAM), and data encryption. * Implemented and maintained CI/CD pipelines using Jenkins to automate the build, test, and deployment processes of data engineering applications. * Running Automated Test scripts to validate the Data integrity from configured Hadoop Data Storages. * Familiar with best practices for network connectivity, security, Maintaining IAM Role, and compliance to maintain data confidentiality and integrity during transit and at rest in AWS. * Configuring and optimizing NIFI components such as processors, controllers, and connectors to efficiently handle flow file requirements in Data Pipeline. * Designed, developed, and maintained automated test frameworks leveraging AWS tools such as AWS Device Farm, Lambda, and Selenium for web application testing. * Capable of integrating NIFI with other big data technologies such as Apache Spark, and Apache Hive, AWS S3 for advanced data processing and analytics. * Led the implementation of ETL processes using AWS Glue, improving data processing efficiency. * Experienced in defining and implementing data transformation logic using Apache Spark within AWS Glue ETL scripts to cleanse, enrich, and aggregate raw data. * Proficient in setting up and managing Amazon Kinesis streams for real-time data processing. * Experience with configuring Kinesis Data Firehose for data delivery to various AWS services like S3, Redshift. * Optimized DynamoDB tables and indexes for performance, ensuring efficient query processing of large Data sets from Hive / Hadoop Database. * Leveraging Sqoop to seamlessly import data from MySQL to HDFS on a recurring basis, establishing a robust foundation for large-scale data integration. * Integrated AWS Glue with AWS services such as Amazon Athena, Amazon EMR, AWS Lambda, and AWS Step Functions to build end-to-end data processing workflows. * Optimizing Hive Queries using techniques and parameters of PySpark * Creating and managing Hive tables proficiently, employing Hive QL to optimize data structures for efficient querying and processing. * Developed Spark/Scala applications using PySpark to execute complex ETL operations with precision, contributing to improved data processing efficiency. * Playing a crucial role in troubleshooting production issues related to large-scale data processing, ensuring continuous system stability and performance. * Active participation in test case design, testing, and documentation, contributing to the seamless integration of enhancements and optimizations in large-scale data environments. * Contributing to NIFI data pipeline projects designed to efficiently handle and process vast datasets, with a specialized focus on configuring Lookups for rigorous data validation and integrity checks. * Expertise in managing diverse file formats, including JSON, AVRO, and parquet, and implementing advanced compression techniques such as snappy within the NIFI ecosystem to streamline the processing of extensive data sets.   **BIG DATA / PROD ENGINEER**  Deloitte Inc, Princeton, NJ  Dec 2020 – May 2021   * Engineered the backend codebase for Prep Orch Micro Services, incorporating proficient handling of diverse datasets including CSV, Parquet, JSON, and Text Files. * Developed Micro Services with a focus on loading logs metadata into the NoSQL database, MongoDB. * Successfully interfaced with external applications, executing Spark transformations and actions to enhance overall system functionality. * Demonstrated expertise in building actions and test cases utilizing Scala, contributing to robust software testing methodologies. * Applied knowledge in resolving production issues by analyzing SJS logs during job execution. * Utilized Kubernetes for deployment orchestration, analyzing logs from pods and containers to troubleshoot deployment issues within the Spark ecosystem. * Applied advanced techniques for handling large datasets, including partitioning, leveraging Spark in-memory capabilities, implementing broadcasts, and optimizing joins and transformations during the data ingestion process. * Optimized Spark code for aggregating, grouping, and executing data mining tasks, enhancing overall performance within the Spark framework. * Possessed a strong foundation in Azure Cloud, with exposure to ADF (Azure Data Factory), ADLS (Azure Data Lake Storage), Azure DevOps (VSTS), and portal services. * Proficiently developed and deployed Shell Scripts for automation, notification, and monitoring purposes. * Conducted performance tuning on Spark applications, ensuring optimal execution and resource utilization. * Facilitated the import of data from diverse sources, executed transformations using Spark, and loaded data into ADLS. * Leveraged Apache Spark SQL and data frame functions for complex semi-structured data transformations and aggregations. * Demonstrated hands-on experience in creating RDDs, implementing Data Frame transformations, and executing actions while developing Spark applications.   **DATA ENGINEER**  WILLIAMS SONOMA, SF, CA  Oct 2020 – Dec 2020   * Architecting, constructing, and deploying highly efficient and dependable data pipelines, orchestrating the seamless movement of data across diverse platforms such as Data Warehouse, online caches, and real-time systems. * Demonstrated expertise in utilizing Azure Function Apps and App Services to enhance data processing capabilities. * Proficient in working with Azure Databricks, employing scripting languages like Scala, Shell, and PowerShell to facilitate robust ETL operations. * Utilized Azure Databricks notebooks for validation purposes, ensuring the integrity of data inbound and outbound from external sources such as Amperity. * Successfully engineered reliable data pipelines for data movement across multiple platforms, including Snowflake, Azure Delta Lake, blob storage, and external dashboards. * Implemented Snowpipe to enable near-real-time data ingestion from cloud storage platforms (Azure Data Lake Storage) into Snowflake data warehouse, facilitating continuous data loading processes. * Optimized Snowpipe configurations to efficiently handle varying data ingestion rates and ensure timely loading of data into Snowflake tables, leveraging automatic scaling capabilities. * Implemented data pipelines and ETL processes using Snowpark, facilitating the efficient ingestion, transformation, and loading of large volumes of data into Snowflake data warehouse for analysis and reporting purposes. * Developed PySpark frames to extract data from Db2 and load it into Delta Lake on Azure, enabling efficient storage and analytics with optimized data formats and schema management. * Build robust code for Data cleaning / Validation using PySpark & Spark SQL Lib in Databricks Notebooks. * Developed Spark/Scala applications using PySpark to execute complex ETL operations with precision, contributing to improved data processing efficiency. * Developed Spark/Scala applications to execute intricate ETL operations with precision. * Specialized in source-to-target mapping and streaming for diverse data transfers through API or Azure Data Factory (ADF) Pipelines, adept at troubleshooting and implementing various logics based on unique requirements. * Proficient in crafting complex SQL queries to drive thorough analysis and derive actionable insights from the data. * Collaborated with cross-functional teams to troubleshoot and resolve production incidents, conducting root cause analysis (RCA) and implementing preventive measures to mitigate future issues. * Implemented Pipelines using Spark Jars and notebook activities, contributing to a streamlined and optimized data processing environment.   **SPARK DEVELOPER**  TMW Systems, Dallas TX  Jan 2018 to Sep 2020   * Developed framework to encrypt sensitive data (SSN, Account number ...etc.) in all kinds of datasets and moved datasets one S3 bucket to another. * Processed datasets like Text, Parquet, Avro, Fixed Width, Zip, JSON and XML. * Developed framework to check data quality of datasets, schema defined in cloud. worked on Amazon Web service (AWS) to integrate EMR with Spark 2 and S3 storage and Snowflake. * Configured Spark streaming to receive real time data from the Kafka and store the stream data into AWS S3 using Scala. * Experienced in handling large datasets using Partitions, Spark in Memory capabilities, Broadcasts in Spark, Effective & efficient Joins, Transformations and other during ingestion process itself. * Enhanced and optimized product Spark code to aggregate, group and run data mining tasks using the Spark framework. * Developed a Data CI/CD pipeline using Data controller UI and Control Hub. * Configured connections and variables within Airflow to securely manage credentials and configurations for different environments. * Worked on migrating Map Reduce programs into Spark transformations using Spark and Used File Broker to schedule workflows to run Spark jobs to transform data on a persistent schedule. * Experience developing, deploying Shell Scripts for automation/notification/monitoring. * Extensively used Apache Kafka, Apache Spark, HDFS and Apache Impala to build a near real time data pipelines that get, transform, store, and analyze click stream data to provide a better personalized user experience. * Spark Streaming collects this data from Kafka in near-real-time and performs necessary transformations and aggregation on the fly to build the common learner data model and persists the data in Cassandra cluster. * Worked with Apache Spark SQL and data frame functions to perform data transformations and aggregations on complex semi structured data. * Hands on experience in creating RDDs, transformations and actions while implementing Spark applications.   **HADOOP DEVELOPER**  GENERIX GROUP, HYDERABAD.  May 2014 to Dec 2015   * Importing and exporting data into HDFS and Hive using Sqoop. * Used Bash Shell Scripting, Sqoop, AVRO, Hive, Pig, Java, Map/Reduce daily to develop ETL, batch processing, and data storage functionality. * Exploited Hadoop MySQL-Connector to store Map Reduce results in RDBMS. * Analyzed large amounts of data sets to determine optimal way to aggregate and report on it. * Worked on loading all tables from the reference source database schema through Sqoop. * Worked on Designing and Developing ETL Workflows using Java for processing data in HDFS/HBase using Oozie. * Experience in Oozie and workflow scheduler to manage Hadoop jobs by Direct Acyclic Graph (DAG) of actions with control flows. * Working on extracting files from MySQL through Sqoop and placed in HDFS and processed. * Created several Hive tables, loaded with data and wrote Hive Queries to run internally in MapReduce. * Experienced in managing and reviewing Hadoop log files. * Involved in loading and transforming large sets of structured, semi structured, and unstructured data from relational databases into HDFS using Sqoop imports. * Working on extracting files from MySQL through Sqoop and placed in HDFS and processed. * Created several Hive tables, loaded with data and wrote Hive Queries to run internally in MapReduce. | |
|  | |