**SRISHANTH
Email:** srishanthm0809@gmail.com **PH: (682) 422-7455
Sr. Data Engineer**

**PROFESSIONAL SUMMARY:**

* A result-oriented Professional with 8+ years of experience in Big Data development along with Data administration and proposing effective solutions through an analytical approach with a track record of building large-scale systems using Big Data technologies
* Excellence in using Use Apache Hadoop to work with Big Data and analyze large data sets
* Hands-on experience in ecosystems like Hive, Sqoop, MapReduce, Flume, and Oozie.
* Strong knowledge of Hive's analytical functions, and the ability to extend Hive functionality by writing custom UDFs.
* Work with Data Lakes and Big Data ecosystems (Hadoop, Spark, Hortonworks, Cloudera).
* Track record of results in an Agile methodology using data-driven analytics.
* Load and transform large sets of structured, semi-structured, and unstructured data working with data on Amazon Redshift, Apache Cassandra, and HDFS in Hadoop Data Lake.
* Experience handling XML files as well as Avro and Parquet in various Big Data contexts
* Performance tuning at source, Target, and Data Stage job levels using Indexes, Hints, and Partitioning in DB2, ORACLE.
* Skilled with BI tools like Tableau and PowerBI, data interpretation, modeling, data analysis, and reporting with the ability to assist in directing planning based on insights.
* Skilled in HDFS, Spark, Hive, Sqoop, HBase, Flume, Oozie, and Zookeeper
* Add value to Agile/Scrum processes such as Sprint Planning, Backlog, Sprint Retrospective, and Requirements Gathering and provide planning and documentation for projects
* Skilled at writing SQL queries, stored Procedures, Triggers, Cursors, and Packages.
* Apply in-depth understanding/knowledge of Hadoop architectures and various components such as HDFS, MapReduce, Spark, and Hive.
* Create Spark Core ETL processes to automate using a workflow scheduler.
* Use Apache Hadoop to work with Big Data and analyze large data sets efficiently.
* Transformational Strategist with expertise in transforming business concepts and needs into mathematical models, designing algorithms, and deploying custom business intelligence software solutions; knowledge of building models with deep learning frameworks such as TensorFlow, PyTorch, and Keras
* Gained experience in applying techniques to live data streams from big data sources using Spark and Scala;possess cloud platform experience using Azure, AWS, and GCP.
* Compiling the data,including internal and external data sources, leveraging new data collection processes such as geo-location information.
* Communicator with the ability to perform at a high level, meet deadlines, and adaptable to ever-changing priorities

**PROFESSIONAL EXPERIENCE:**

**Sr. Data Engineer
Nationwide, Columbus, OH March 2022 to Present
Responsibilities:**

* Designed and developed web app BI for performance analytics.
* Designed Python-based notebooks for automated weekly, monthly, and quarterly reporting E.T.L
* Present weekly updates to managers and key stakeholders to preview the user interface designs and analytical results of stress analysis findings, etc.
* Implemented an international digitalization methodology in the manufacturing processes. Also, I develop the training system and material which was used to instruct the personnel of the company on the correct implementation of the system.
* Gathering of historical data of brother companies for its analysis and interpretation by creating a model which provided easy understanding statistics and graphs for the managers to use at their meetings and to expose to the auditors and visits of the plant
* Present data using PowerPoint, Tableau, and Excel for data work and charts.
* This project involves implementing Azure IoT to connect and manage IoT devices.
* Worked in Git development environment.
* Used Python to create a semi-automated conversion process to generate a raw archive-linked data file.
* Used Oozie workflow engine to manage interdependent Hadoop jobs and automate several types of Hadoop jobs such as Hive and Sqoop
* Produced scripts for doing transformations using Scala
* Developed and implemented Hive custom UDFs involving date functions
* Created a benchmark between Hive and HBase for fast ingestion
* Configured Spark Streaming to receive real-time data from Apache Kafka and store the streamed data to HDFS using Scala.
* Configured and managed Apache Kafka clusters for reliable and scalable message queuing and streaming.
* Fine-tuned Kafka broker configurations for optimal resource utilization and performance under varying workloads.
* Designed Kafka topics with optimal partitioning strategies to distribute data evenly across Kafka brokers.
* Utilized partitioning techniques such as key-based partitioning to ensure data locality and balanced workload distribution.
* Developed RESTful APIs for seamless integration with external systems and applications to extract real-time data.
* Utilized industry-standard authentication and authorization mechanisms such as OAuth2 for secure API access.
* Implemented Apache Airflow workflows to automate and orchestrate complex data pipelines and ETL tasks.
* Designed DAGs (Directed Acyclic Graphs) in Apache Airflow to schedule and monitor tasks for cloud service automation.
* Led the development of proof-of-concept (POC) projects to validate the feasibility and effectiveness of new technologies and solutions.
* Collaborated with cross-functional teams to gather requirements and define POC objectives and success criteria.
* Designed scalable and resilient cloud architectures on AWS, utilizing services such as EC2, S3, Lambda, and RDS.
* Implemented Infrastructure as Code (IaC) using AWS CloudFormation to automate deployment and management tasks.
* Implemented real-time data processing pipelines using Apache Storm for efficient stream processing.
* Leveraged Apache Spark to analyze large-scale data streams in real-time, enhancing data-driven decision-making processes.
* Integrated Google Cloud Platform (GCP) AI services such as Vision API, Natural Language API, and AutoML into applications to enable advanced machine learning and AI capabilities.
* Developed custom machine learning models using Google Cloud AI Platform, TensorFlow, and other GCP tools to address business challenges and opportunities.
* Utilized Impala for interactive SQL querying and analysis of data stored in Hadoop clusters, providing real-time insights and decision support.
* Developed complex SQL queries and analytical functions in Impala to perform advanced data analysis and reporting tasks.
* Installed, configured, and maintained Cloudera Hadoop clusters to support distributed data processing and analysis.
* Developed and optimized MapReduce jobs and Hive queries to perform batch processing and analytics on large datasets stored in Hadoop Distributed File System (HDFS).
* Deployed and configured AWS Redshift clusters to store and analyze large volumes of structured data efficiently.
* Optimized Redshift performance by implementing distribution keys, sort keys, and compression techniques to improve query execution times.
* Led Agile development teams in sprint planning, daily stand-ups, and retrospective meetings, fostering collaboration and transparency throughout the project lifecycle.
* Implemented Agile practices such as Scrum and Kanban to prioritize tasks, manage workloads, and adapt to changing requirements effectively.
* Integrated various SaaS applications and APIs to streamline business processes and improve data interoperability.
* Developed custom connectors and middleware for seamless integration between disparate systems, enabling real-time data exchange and synchronization.
* Designed and implemented Hive data models to organize structured and semi-structured data effectively, facilitating efficient querying and analysis.
* Optimized Hive schemas and partitioning strategies to enhance query performance and reduce data retrieval times.
* Utilized Pandas to manipulate and clean large datasets efficiently, enhancing data quality and preparing it for analysis.
* Implemented Spark for distributed data processing, improving scalability and performance of data pipelines across clusters.
* Developed unit tests using PyTest framework to validate the functionality of Python code modules and ensure robustness and correctness.
* Implemented test fixtures and parametrized test cases in PyTest to efficiently organize and execute test suites across different scenarios and inputs.
* Developed Python scripts to collect data from RESTful APIs using libraries like Requests or aiohttp, enabling integration with external systems and services.
* Implemented authentication mechanisms such as OAuth 2.0 or API keys to securely access and retrieve data from REST APIs while adhering to security best practices.
* Designed and implemented serverless data processing solutions on Google Cloud Platform (GCP) using services like Cloud Functions, Cloud Pub/Sub, and Cloud Dataflow.
* Leveraged serverless architecture to auto-scale resources dynamically based on workload demands, optimizing cost efficiency and performance.
* Configured rack awareness in Apache Kafka to distribute replicas across multiple racks for fault tolerance and resilience against rack-level failures.
* Leveraged rack-awareness to improve data locality and reduce cross-rack network traffic, enhancing overall system performance and throughput.
* Utilized Amazon Managed Streaming for Apache Kafka (MSK) to deploy and manage Kafka clusters in AWS, leveraging managed services for seamless operations.
* Integrated AWS MSK with other AWS services such as Amazon S3 and Amazon Redshift to build end-to-end data pipelines for analytics and reporting.
* Monitored Kafka clusters using tools like Kafka Manager and Prometheus to ensure optimal performance and reliability.
* Tuned Kafka configurations such as topic partitions, replication factors, and retention policies to meet specific throughput and durability requirements.
* Developed complex DAGs (Directed Acyclic Graphs) in Apache Airflow to automate data workflows, including data ingestion, transformation, and model training.
* Configured Airflow to schedule and monitor ETL (Extract, Transform, Load) jobs, ensuring timely execution and adherence to SLAs (Service Level Agreements).
* Leveraged Spark Streaming with Kafka to develop real-time data processing pipelines for analyzing user behavior on an e-commerce platform.
* Integrated Spark Streaming with Kafka to ingest large volumes of clickstream data, enabling timely insights for marketing campaigns and product recommendations.
* Developed custom User Defined Functions (UDFs) in Hive using Java or Scala to extend the functionality of Hive SQL for specialized data processing requirements.
* Optimized Hive UDFs for performance by implementing them in native code or leveraging in-memory caching mechanisms to reduce computational overhead.
* Developed Scala scripts for data transformation tasks using libraries like Apache Spark, leveraging its distributed processing capabilities for large-scale data processing.
* Implemented complex data transformations and aggregations using Scala functions and Spark SQL, optimizing performance and resource utilization for efficient data processing.
* Designed and orchestrated complex data workflows using Apache Oozie, scheduling and coordinating various data processing tasks such as Hive queries, MapReduce jobs, and Spark jobs.
* Integrated Oozie workflows with data ingestion pipelines, ensuring seamless data movement and transformation across different stages of the data processing pipeline.
* Maintained version control of codebase using Git, ensuring collaborative development and tracking changes across distributed teams.
* Implemented branching strategies such as GitFlow to manage feature development, bug fixes, and releases in an organized manner.
* Designed and implemented end-to-end IoT solutions on Azure IoT Hub, including device provisioning, telemetry ingestion, and command and control functionalities.
* Leveraged Azure Stream Analytics to process real-time streaming data from IoT devices, enabling actionable insights and predictive analytics.
* Implemented data visualization dashboards using Python libraries such as Flask and Django, integrating with BI tools like Power BI and Tableau for dynamic reporting.
* Developed RESTful APIs using Python frameworks like Flask to facilitate seamless data exchange between web applications and BI platforms.
* Hands-on experience in Spark and Spark Structured Streaming, using DataFrames to process a constant stream of data
* Deployed and managed multipe applications on GCP, including Spark processing on Cloud Dataproc and data storage and management in BigQuery
* Used Kafka on publish-subscribe messaging as a distributed commit log
* Ran data processing jobs from a Kubernetes cluster
* Created Airflow Scheduling scripts in Python to automate data pipeline and data transfer
* Programmed Flume and HiveQL scripts to move the data into the database and process it
* Implemented AWS Fully Managed Kafka streaming to send data streams from the company APIs to Spark cluster in AWS Databricks

**Sr. Data Engineer
AgFirst columbia, SC January 2020 to February 2022
Responsibilities:**

* Extracted data from different databases and scheduled Oozie workflows to execute the task daily
* Worked with Amazon Web Services (AWS) and was involved in ETL, Data Integration, and Migration
* Documented the requirements including the available code which should be implemented using Spark, Amazon DynamoDB, Redshift, and Elastic Search
* Managed version control setup for the platform using Git.
* Perform analysis of user profiles and current application entitlements based on user profiles, organization, departments, and groups.
* Built a recommending system to auto-provisioning applications and platform access to new employees/contractors so they are productive as soon as they on-boarded
* Worked in setting up IoT hubs, configuring device connectivity, and implementing data processing and analysis.
* Selected and built dashboards for internal usage
* Developed multiple Spark Streaming and batch Spark jobs using Python on AWS.
* Implemented advanced procedures of feature engineering for the data science team using in-memory computing capabilities like Apache Spark written in Python
* Implemented serverless architecture on GCP. The architecture included using Google Cloud Functions, Cloud Run, and Cloud Scheduler to build scalable project
* Implemented Rack Awareness in the Production Environment.
* Collected data using REST API, built HTTPS connection with client-server, sent GET request, and collected response in Kafka Producer
* Implemented Hadoop ecosystem tools such as HBase, Impala, and Hue to enable real-time querying and analysis of data in CDH clusters.
* Conducted performance tuning and troubleshooting of CDH components to optimize cluster performance and reliability.
* Designed and implemented data loading pipelines using AWS Glue and other ETL tools to ingest data into Redshift from various sources.
* Implemented data encryption and access controls to ensure the security and compliance of sensitive data stored in Redshift.
* Utilized Agile project management tools such as Jira and Trello to track progress, monitor team velocity, and facilitate continuous improvement.
* Facilitated cross-functional communication and coordination between development, QA, and product management teams to ensure alignment with Agile principles and objectives.
* Conducted API discovery and evaluation to identify suitable integration solutions, considering factors such as security, scalability, and performance.
* Worked closely with vendors and third-party providers to troubleshoot integration issues and ensure seamless connectivity between systems.
* Conducted data profiling and analysis to identify patterns and relationships, informing the design of Hive data models for optimal storage and retrieval.
* Collaborated with cross-functional teams to define Hive table structures and metadata requirements, ensuring alignment with business objectives and analytical needs.
* Integrated Pandas and Spark seamlessly in data processing workflows, leveraging the strengths of each for optimal results.
* Developed custom functions and algorithms in Pandas for complex data transformations, ensuring accuracy and reliability in processing.
* Utilized mocking and patching techniques in PyTest to isolate code under test from external dependencies and facilitate faster and more focused unit testing.
* Integrated PyTest with continuous integration pipelines using tools like Jenkins or GitLab CI to automate test execution and regression testing.
* Designed data models and schema mappings to transform raw JSON or XML responses from REST APIs into structured data formats suitable for downstream processing.
* Handled pagination and rate limiting in REST API data collection scripts to efficiently fetch large datasets without exceeding API usage limits or encountering errors.
* Integrated custom UDFs into Hive queries and scripts to perform complex calculations, text processing, or data transformations that are not natively supported by Hive.
* Tested and validated custom UDFs to ensure compatibility with different versions of Hive and compatibility with existing data processing workflows.
* Optimized Spark Streaming jobs to handle high throughput and low latency requirements in a distributed environment, ensuring smooth data processing.
* Implemented fault-tolerant mechanisms using Spark Streaming and Kafka to handle data inconsistencies and ensure data integrity across multiple microservices.
* Orchestrated dependencies between tasks in Airflow DAGs to streamline data pipelines and optimize resource utilization.
* Integrated Airflow with external systems and services to trigger workflows based on events or changes in data sources, enhancing automation capabilities.
* Implemented monitoring alerts and thresholds to proactively identify and address issues such as under-replicated partitions or high consumer lag.
* Conducted capacity planning for Kafka clusters based on projected data growth and usage patterns, scaling resources as needed to accommodate increased load.
* Implemented security best practices for AWS MSK, including encryption at rest and in transit, IAM (Identity and Access Management) policies, and VPC (Virtual Private Cloud) configuration.
* Leveraged AWS CloudFormation and AWS CLI to automate the provisioning and configuration of AWS MSK clusters, ensuring consistency and repeatability.
* Implemented custom partition assignment strategies in Kafka based on rack awareness to ensure even distribution of partitions across racks while minimizing data movement during rebalancing.
* Conducted rack failure simulations and failover tests to validate the effectiveness of rack awareness configurations and ensure high availability of Kafka clusters.
* Orchestrated serverless workflows using Cloud Composer (based on Apache Airflow) for scheduling and monitoring data pipelines on GCP.
* Integrated serverless components with GCP's IAM and networking services to enforce security policies and ensure data privacy and compliance.
* Integrated Scala scripts with existing data pipelines, orchestrating data ingestion, transformation, and storage operations in a unified workflow.
* Leveraged Scala's functional programming capabilities to write concise and expressive code for data manipulation and transformation tasks.
* Implemented error handling and retry mechanisms within Oozie workflows to handle transient failures and ensure data reliability and consistency.
* Monitored and troubleshooted Oozie workflow executions, identifying performance bottlenecks and optimizing resource utilization for efficient job scheduling and execution.
* Utilized Git submodules for managing dependencies and integrating external libraries into the project repository.
* Implemented continuous integration and deployment pipelines using Git hooks and CI/CD tools like Jenkins, automating build, test, and deployment processes.
* Integrated Azure Functions with IoT workflows to automate device management tasks and trigger alerts based on predefined conditions.
* Implemented Azure IoT Edge deployments for edge computing scenarios, enabling local data processing and reducing latency for critical IoT applications.
* Utilized Pandas and NumPy libraries for data manipulation and analysis within web applications, ensuring efficient business intelligence processes.
* Integrated authentication and authorization mechanisms using Python frameworks like Flask-Security to secure access to sensitive business intelligence data.
* Integrated Impala with business intelligence tools such as Tableau and Power BI to visualize and explore data insights effectively.
* Optimized Impala query performance by tuning query execution plans, leveraging query hints, and optimizing table partitioning strategies.
* Implemented pipelines for data preprocessing, feature engineering, and model training using GCP services like Dataflow and BigQuery ML.
* Deployed and scaled machine learning models in production environments on GCP, leveraging services like AI Platform Prediction and Kubernetes Engine.
* Optimized Apache Storm topologies for increased throughput and low-latency processing of streaming data.
* Integrated Apache Spark with external data sources to ingest, process, and analyze data streams seamlessly.
* Orchestrated CI/CD pipelines on AWS using tools like AWS CodePipeline and CodeDeploy for efficient DevOps practices.
* Provided operational support for AWS cloud infrastructure, including monitoring, alerting, and incident response.
* Built comprehensive roadmaps outlining key milestones, deliverables, and resource requirements for project execution.
* Conducted thorough market research and feasibility studies to inform POC development and roadmap construction.
* Integrated Apache Airflow with cloud services such as AWS S3, EC2, and Lambda for seamless workflow execution.
* Developed custom plugins and operators in Apache Airflow to extend functionality and accommodate specific use cases.
* Implemented data serialization techniques like JSON and XML for efficient data transmission over RESTful APIs.
* Conducted thorough API testing using tools like Postman to ensure reliability and interoperability.
* Monitored and adjusted partition assignments dynamically to accommodate changes in data volume and processing requirements.
* Implemented partitioning policies considering factors like data skew, throughput requirements, and fault tolerance.
* Implemented security measures such as SSL encryption and authentication mechanisms to secure Kafka clusters.
* Monitored Kafka cluster health and performance metrics using tools like Kafka Manager and Prometheus.
* Installed and configured Kafka cluster and monitored the cluster
* Architected a lightweight Kafka broker and integrated Kafka with Spark for real-time data processing.
* Received real-time data from Apache Kafka and stored the streamed data to HDFS using Scala.
* Wrote Unit tests for all code using PyTest for Python
* Used Python Boto3 for interacting with AWS resources programmatically
* Used the Pandas library and Spark in python for data cleansing, validation, processing, and analysis.
* Created Hive external tables and designed data models in Apache Hive
* Implemented and cleans datasets for network accesses based on user profiles

**Big Data Engineer
Mayo Clinic Rochester MN May 2017 to December 2019
Responsibilities:**

* Optimized and integrated Hive, SQOOP, and Flume into existing ETL processes, accelerating the extraction, transformation, and loading of massive structured and unstructured data.
* Designed, developed and produced reports that connect quantitative data to insights that drive and change business
* Handling ETLs, and data analysis with billions of records.
* Optimized SQL queries embedded within Python code to retrieve and process large datasets for business analytics, ensuring high performance and scalability.
* Utilized Azure IoT Central for rapid prototyping and deployment of IoT solutions, accelerating time-to-market and reducing development overhead.
* Conducted code reviews and pull request evaluations within Git repositories, ensuring code quality, consistency, and adherence to coding standards.
* Worked with Oozie coordinator and bundle applications to manage dependencies and scheduling constraints across multiple interconnected workflows.
* Utilized Scala REPL (Read-Eval-Print Loop) for interactive development and testing of data transformation logic, ensuring correctness and reliability before deployment.
* Documented usage and parameters of custom UDFs for internal knowledge sharing and future reference by data engineers and analysts.
* Collaborated with cross-functional teams to troubleshoot and resolve performance bottlenecks in Spark Streaming with Kafka applications, improving overall system efficiency.
* Implemented error handling and retry mechanisms in Airflow to handle job failures gracefully and minimize data processing interruptions.
* Automated routine maintenance tasks for Kafka clusters, such as log rotation and broker rebalancing, to minimize manual intervention and ensure system stability.
* Optimized costs by utilizing AWS MSK's flexible pricing model based on usage, and right-sizing clusters based on workload demands.
* Documented rack awareness configurations and best practices for future reference and knowledge sharing within the team.
* Implemented CI/CD pipelines for serverless applications on GCP using tools like Cloud Build and Cloud Deployment Manager, enabling automated deployment and updates.
* Implemented error handling and logging mechanisms to capture and report exceptions during REST API data collection, ensuring data integrity and reliability.
* Leveraged PyTest's rich assertion library and plugins ecosystem to enhance test coverage and effectiveness, including assertions for data validation and error handling.
* Leveraged Spark's RDDs and DataFrames to perform advanced analytics tasks on massive datasets, achieving significant time savings and scalability.
* Implemented best practices for Hive data modeling, including schema normalization, denormalization, and optimization techniques, to improve overall system performance and usability.
* Implemented RESTful APIs and web services to enable communication between internal and external systems, facilitating data sharing and collaboration.
* Emphasized the value of iterative development and feedback loops, encouraging continuous delivery and improvement in software development processes.
* Monitored Redshift cluster performance and capacity utilization, scaling resources as needed to meet changing demand and workload requirements.
* Implemented data governance and security controls in CDH clusters to ensure compliance with regulatory requirements and protect sensitive information.
* Collaborated with data engineers and analysts to design and optimize Impala schemas and data models for efficient querying and analysis.
* Conducted performance monitoring, model evaluation, and continuous improvement of AI systems deployed on GCP, ensuring reliability and scalability.
* Conducted performance tuning and troubleshooting on Apache Storm and Apache Spark clusters to ensure high availability and reliability.
* Optimized cost and performance of AWS resources by implementing best practices and utilizing AWS Cost Explorer.
* Presented POC findings and recommendations to stakeholders, facilitating informed decision-making processes.
* Conducted performance optimization and scaling of Apache Airflow deployments for increased efficiency and reliability.
* Documented API endpoints and specifications following OpenAPI (Swagger) standards for clear communication and ease of integration.
* Conducted performance testing and benchmarking to evaluate the effectiveness of partitioning strategies and optimize Kafka cluster performance.
* Conducted routine maintenance tasks such as backups, upgrades, and scaling of Kafka clusters to ensure uninterrupted operation.
* Requirements capture with the management team to design reports to make decisions
* Predictive user analysis for advertising campaigns
* Construction and customization of integration systems using technologies such as Saas, API, and web services
* Implemented Agile Methodology for building an internal application.
* Use of knowledge databases and language ontologies.
* Performed cluster capacity and growth planning and recommended node configuration.
* Worked with both unstructured and structured data.

**Hadoop Data Engineer
Brio Technologies Private Limited Hyd India November 2015 to February 2017
Responsibilities:**

* Used AWS Redshift for cloud data storage.
* Used different file formats such as Text files, Sequence Files, and Avro for data processing in the Hadoop system.
* Loaded data from various data sources into Hadoop Distributed File System (HDFS). using Kafka.
* Integrated Kafka with Spark Streaming for real-time data processing in Hadoop.
* Used Cloudera (CDH) distribution of Hadoop to store and process data
* Used image files to create instances containing Hadoop installed and running.
* Streamed analyzed data to Hive Tables using Sqoop, making it available for data visualization.
* Tuning and operating Spark and its related technologies like Spark SQL and Spark Streaming.

**Cloud Engineer
Maisa Solutions Private Limited Hyderabad, India July 2014 to October 2015
Responsibilities:**

* Set up and implemented Kafka brokers to write data to topics and utilize its fault tolerance mechanism
* Created and managed Topic creation inside Kafka
* Used Azure Functions, Logic Apps, and Event Grid to build scalable and event-driven applications
* Configured a full Kafka cluster with the multi-broker system for high availability
* Used Spark Streaming to consume from Kafka topics and transform the data processed.
* Presented finding to the multiple teams using PowerPoint, Tableau, and Excel
* Involved in the complete Big Data flow of the application starting from data ingestion from upstream to HDFS
* Used Impala where possible to achieve faster results compared to Hive during data Analysis.
* Implemented GCP AI to build intelligent application that includes natural language processing, computer vision, and machine learning
* Worked on Hadoop streaming jobs to process terabytes of XML format data.
* Real-Time/Stream processing Apache Storm, Apache Spark
* Architecting and Data Engineering for AWS cloud services including AWS Cloud services planning, designing, and DevOps support like IAM user, group, roles & policy management
* Proposed a working POC and constructed the roadmap for the prediction pipeline
* Designed jobs using DB2 UDB, ODBC,.Net, Join, Merge, Lookup, Remove duplicate, Copy, Filter, Funnel, Dataset, Lookup file set, Change data capture, Modify, Row merger, Aggregator and Peek, and Row generator stages
* Involved in loading data from the UNIX file system to HDFS
* Wrote Spark SQL queries and optimized the Spark queries in Spark SQL
* Used a REST API to extract real-time financial data of bitcoin, and alt-coin prices every minute
* Optimized data storage in Kafka Brokers within the Kafka cluster by partitioning Kafka Topics
* Created modules for Apache airflow to call different services in the cloud including EMR, S3, Athena, Crawlers, Lambda functions, and Glue jobs.
* Set up and implemented Kafka brokers to write data to topics and utilize its fault tolerance mechanism.
* Configured a full Kafka cluster with the multi-broker system for high availability
* Used Spark Streaming to consume from Kafka topics and transform the data processed.