

 **Lakshmi**

 **Data ENGINEER**

 **• Phone : +1 312-584-0785 Extn : 111**

 **• Email:razak@webmobilez.com**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**PROFESSIONAL SUMMARY:**

* Having 8+ years of professional experience in Information Technology and around 6 years of expertise in BIGDATA using HADOOP framework and Analysis, Design, Development, Testing, Documentation, Deployment, and Integration using SQL and Big Data technologies.
* Strong experience in programming languages like Java, Scala, and Python.
* Experienced on data architecture including data ingestion, data modelling and data mining, machine learning and advanced data processing.
* Strong experience working with Hadoop ecosystem components like HDFS,
* Map Reduce, Spark, HBase, Oozie, Hive, Sqoop, Pig, Flume and Kafka
* Good hands-on experiencing working with various Hadoop distributions mainly,
* Cloudera (CDH), Hortonworks (HDP) and Amazon EMR.
* Good understanding of Distributed Systems architecture and design principles
* Behind Parallel Computing.
* Expertise in Talend for ETL and data migration processes.
* Strong knowledge and experience with PL/SQL development and optimization.
* Proficient in managing and migrating data between Oracle and PostgreSQL databases.
* Skilled in writing and tuning complex SQL queries for performance optimization.
* Experience with JSONB data handling and optimization in PostgreSQL.
* Ability to perform database management tasks, including maintenance, backups, and performance tuning.
* Strong collaboration skills, working effectively with cross-functional teams to deliver database solutions.
* Proven ability to manage and execute data migration projects, ensuring data integrity and minimal downtime.
* Expertise in developing production-ready Spark applications utilizing Spark-Core,
* Data frames, Spark-SQL, Spark-ML and Spark-Streaming API's.
* Strong experience troubleshooting failures in spark applications and fine-tuning.
* Spark applications and hive queries for better performance.
* Worked extensively on Hive for building complex data analytical applications.
* Strong experience writing complex map-reduce jobs including development of custom Input Formats and custom Record Readers.
* Highly skilled and motivated Data Engineer with extensive experience in designing, developing, and maintaining data pipelines using Apache Airflow.
* Experience working with Azure Data Factory that include designing and implementing end - to-end.
* Worked on backing up and restoring the Azure Data Factory V2.
* Involved in creation of triggers in Azure Data Factory using schedule trigger and event trigger which will trigger pipelines based on the conditions.
* Experience in different load strategies like SCD0, SCD1 and SCD2 by implementing ETL work loads in Azure data Factory Data Flows and Data bricks Notebooks.
* Extensively worked in developing ETL program for supporting Data Extraction, transformations and loading using Azure data Factory and Data bricks
* Sound knowledge in map side join, reduce side join, shuffle & sort, distributed cache, compression techniques, multiple Hadoop Input & output formats.
* MLOps (Machine Learning Operations) is a practice that focuses on the deployment, management, and continuous improvement of machine learning models in production environments.
* MLOps bridges the gap between data scientists and IT operations to ensure the successful deployment and operation of machine learning systems.
* The roles and responsibilities of individuals involved in MLOps can vary depending on the organization and project requirements
* It's important to note that the specific responsibilities may vary depending on the organization, industry, and the maturity of the machine learning and AI initiatives.
* Machine learning models are algorithms that can learn from data and make predictions or decisions without being explicitly programmed.
* These models are a key component of machine learning and are used in various applications, including image recognition, natural language processing, recommendation systems, fraud detection, and more
* Worked with Apache NiFi to automate the data flow between the systems and
* Managed flow of information between systems.
* Good experience working with AWS Cloud services like S3, EMR, Redshift, Athena,
* Glue etc.
* Worked on building real time data workflows using Kafka, Spark streaming and
* HBase.
* Extensive knowledge on NoSQL databases like HBase, Cassandra and Mongo DB.
* Solid experience in working with csv, text, sequential, parquet, orc, Json formats of
* Data.
* Extensive experience in performing ETL on structured, semi-structured data using Pig Latin Scripts.
* Designed and implemented Hive and Pig UDF's using Java for evaluation, filtering, and storing of data.
* Extensively worked on Sqoop, Hadoop, Hive, Spark, HBase to build ETL and Data processing systems having various data sources, data targets and data formats.
* Experience in designing, developing, scheduling reports/dashboards using Tableau.
* Expertise in performing data analysis and data profiling using complex SQL on various sources systems including Oracle and Teradata.
* Experience in Google Cloud components, Google container builders and GCP client libraries and cloud SDK’s
* Working with AWS/GCP cloud using in GCP Cloud storage, Data-Proc, Data Flow, Big- Query, EMR, AWS-S3, Glacier and EC2 Instance with EMR cluster.
* Building/Maintaining Docker container clusters managed by Kubernetes Linux, Bash, GIT, Docker, on GCP.
* Development experience with RDBMS, including writing SQL queries, views, stored procedure, triggers, etc.
* Strong understanding of Software Development Lifecycle (SDLC) and various methodologies (Waterfall, Agile).

**PROFESSIONAL TOOLS:**

|  |  |
| --- | --- |
| **Hadoop Components**  **Big Data Ecosystem** | HDFS, Hue, MapReduce, Pig, Hive, HCatalog, HBase, Sqoop, Impala, Zookeeper, Flume, Kafka, Yarn, Cloudera Manager, Kerberos, Informatica, PySparkAirflow, Kafka Snowflake. |
| **Languages** | Scala, Python, SQL, Python, Hive QL |
| **IDE Tools** | Eclipse, IntelliJ, PyCharm. |
| **Cloud platform** | AWS (Lambda, DynamoDB, RDS), MS Azure (Azure Databricks, ADF, Azure Data Explorer, Azure HDInsight, ADLS), GCP |
| **Cloud Services** | AWS Glue, S3, RedShift, Data Bricks, EC2, S3, EMR, Dynamo DB, Data Lake, AWS Lambda, cloud watch, Azure Data Factory, Azure Data Lake Storage, Salesforce, Azure Synapse Analytics, Azure Analytical Services, HDInsight, Azure SQL Datawarehouse. |
| **Reporting and ETL Tools** | Tableau, MuleSoft, Power BI, Talend, AWS GLUE. |
| **Databases** | Oracle, SQL Server, MySQL, MS Access, NoSQL Database (HBase, Cassandra, MongoDB) |
| **Containerization** | Docker, Kubernetes |
| **CI/CD Tools** | Jenkins, Bamboo, GitLab  |
| **Software Methodologies** | Agile, Scrum, Waterfall |

**PROFESSIONAL EXPERIENCE:**

**Client Company: FORD Jan 2022– Present**

**Role: Sr. Big Data Developer/Machine Learning Engineer**

**Responsibilities**:

* Evaluating client needs and translating their business requirement to functional specifications thereby onboarding them onto the Hadoop ecosystem.
* Involved in developing roadmap for migration of enterprise data from multiple data sources like SQL Server, provider databases into S3 which serves as a centralized datahub across the organization.
* Loaded and transformed large sets of structured and semi structured data from various downstream systems.
* Developed ETL pipelines using Spark and Hive for performing various business specific transformations.
* Building data applications and automating the pipelines in Spark for bulk loads as well as Incremental Loads of various Datasets.
* Seeking a challenging position to leverage my expertise in data engineering and Airflow to contribute to the success of an innovative organization.
* Designed, developed, and maintained scalable data pipelines using Apache Airflow, ensuring reliable and efficient data processing and workflow orchestration.
* Create Source to Target Mapping for the Azure Data Factory Pipeline Development. Design pipelines to process data from various sources to target databases. Create SQL Server Configurations, Performance Tuning of stored procedures.
* Implemented Azure Data Factory pipelines using Control tables as metadata driven pipeline which has details of source and target.
* Developed Parameterized Azure data Factory pipeline to load complete data from oracle to storage account which is Raw layer.
* Implemented Azure data Factory pipelines to control the load either full or incremental by using orascan values from Oracle tables to ADLS Gen2 folder Raw layer.
* Handling multiple files loading from Raw layer to Refined layer in ADLS Gen2 in the form of delta files using meta driven Azure Data Factory pipelines.
* Building Data Ware housing solutions using different implementations using Data Flows in Azure Data Factory.
* Collaborate with stakeholders to understand business problems and identify opportunities for applying machine learning and AI techniques.
* Define project objectives, success criteria, and key performance indicators (KPIs).
* Design machine learning and AI solutions that address the identified problems and align with business goals.
* Identify relevant data sources and work with data engineers to gather and preprocess data for model training and evaluation.
* Perform data cleaning, transformation, feature engineering, and data augmentation to prepare high-quality datasets for training and testing.
* Select appropriate machine learning algorithms, models, and frameworks based on the problem domain and data characteristics.
* Develop and implement machine learning pipelines and workflows.
* Train and fine-tune machine learning models using suitable techniques such as hyperparameter optimization, cross-validation, and regularization.
* Evaluate the performance of trained models using appropriate evaluation metrics.
* Validate models for accuracy, precision, recall, F1-score, or other relevant metrics.
* Conduct A/B testing, validation, and statistical analysis to assess the model's effectiveness.
* We use model is used for predicting continuous numerical values based on input features. It assumes a linear relationship between the input variables and the target variable.
* As  model is used for binary classification problems. It predicts the probability of an instance belonging to a particular class based on input features.
* Managed data migration projects utilizing Talend for ETL processes, ensuring seamless transition and data integrity.
* Migrated data between Oracle and PostgreSQL databases, focusing on minimizing downtime and ensuring data accuracy.
* Developed and executed migration scripts and routines to facilitate the movement of data between heterogeneous systems.
* Decision trees are versatile models that can be used for both classification and regression tasks. They create a flowchart-like structure to make decisions based on feature values.
* Random Forest is an ensemble model that combines multiple decision trees to make predictions. It improves accuracy by reducing overfitting and handling complex relationships in the data.
* SVM is a powerful model for both classification and regression tasks. It finds an optimal hyperplane that separates data points of different classes or predicts continuous values.
* Design and implement infrastructure and systems to support the deployment and monitoring of machine learning models.
* Develop automation scripts and pipelines for model training, validation, deployment, and scaling.
* Collaborate with data scientists and software engineers to ensure the integration of machine learning models into production systems.
* Implement model versioning, tracking, and performance monitoring.
* Monitor and troubleshoot production machine learning systems, ensuring high availability and performance.
* Implement security and compliance measures for machine learning systems.
* Stay up to date with the latest tools and technologies in the MLOps field.
* Develop and maintain data pipelines to support the training and deployment of machine learning models.
* Design and optimize data storage and processing systems to handle large-scale datasets.
* Ensure data quality, data integrity, and data privacy in the machine learning pipeline.
* Collaborate with MLOps engineers and data scientists to design data architectures that support model deployment and monitoring.
* Implement data versioning and lineage tracking to ensure reproducibility and auditability.
* After implementing all data flows created Azure data factory pipelines to load data to Final destination EDW layer.
* Created notebooks in Azure data bricks for doing transformations like adding headers to the files which doesn’t contain headers, splitting file data header, detail and trailer separately finally loading to snowflake tables, removing last empty and unnecessary records from source files.
* File configuration issues like quote character, zip or deflate functionality issues in copy activity, issues in spark SQL code while executing notebooks in Azure data bricks.
* Worked on building real time pipelines using Kafka and Spark Streaming.
* Worked closely with our data scientist team’s and business consumers to shape the datasets as per the requirements. Data transformation: Apply transformations to raw data, including data cleansing, aggregation, normalization, and enrichment, to prepare it for analysis and storage.
* Automated the data pipeline to ETL all the Datasets along with full loads and incremental loads of data.
* Utilized AWS services like EMR, S3, Glue Meta store and Athena extensively for building the data applications.
* As a GCP (Google Cloud Platform) Data Engineer, your role involves designing, building, and maintaining data infrastructure and solutions to enable efficient and reliable data processing. Here are some key responsibilities and their detailed explanations
* Worked on building input adapters for data dumps from FTP Servers using Apache spark.
* Libraries and frameworks: Working with well-known Scala frameworks and libraries to speed up development and exploit existing functionality. Scala frameworks that are regularly utilized include:
* Wrote spark applications to perform operations like data inspection, cleaning, load and transforms the large sets of structured and semi-structured data.
* Big Data refers to large and complex datasets that exceed the processing capabilities of traditional data management systems.
* The Data Stack enables organizations to handle data at scale, derive insights, and make informed decisions.
* These concepts are fundamental in the realm of data engineering, data analytics, and machine learning, providing the foundation for building robust and scalable data-driven solutions.
* Snowflake supports various data types and provides built-in functionalities for data integration, transformation, and analytics. It also offers features like automatic scaling, concurrency control, and data sharing, making it popular for modern data warehousing and analytics use cases.
* Developed Spark application with Scala and Spark-SQL for testing and processing of data.
* Reporting the spark job stats, monitoring, and running data quality checks are made available for each Datasets.
* Big Data is characterized by the three V's: volume (large amounts of data), velocity (high data ingestion and processing speed), and variety (diverse data types and sources)
* Data Pipeline Design and Implementation: Data Engineers are responsible for designing and implementing data pipelines that extract, transform, and load (ETL) data from various sources into target data systems.
* Managing and analyzing Big Data often requires specialized tools and technologies such as Apache Hadoop, Apache Spark, or cloud-based platforms like Google BigQuery or Amazon Redshift.
* Developed Business logic using python.
* Big Data is widely used in applications such as predictive analytics, machine learning, and business intelligence.
* SQL and programming languages: Write efficient SQL queries and code in languages like Python, Java, or Scala to perform complex data manipulations, calculations, and aggregations
* Writing and performing unit tests, integration tests, and automated tests to assure the quality and dependability of software. This involves the use of testing frameworks such as Scala Test or Specs2.
* Worked on python scripts to parse the XML documents and load the data in database.
* Experienced in python packages such as NumPy, matplotlib, pandas, SciPy.
* Optimizing data structures: Employ techniques like indexing, partitioning, and clustering to enhance data querying and processing performance.
* Spark is a large data analytics distributed data processing framework.
* CI/CD stands for continuous integration and deployment.
* ML pipelines typically involve a combination of data processing, statistical modeling, and algorithmic techniques. They help streamline the end-to-end ML workflow and enable reproducibility
* Understanding business requirements: Collaborate with stakeholders to understand their data requirements, such as the types of data needed, frequency of updates, and desired output formats
* Integrating Scala code into CI/CD pipelines, as well as automating build, testing, and deployment procedures with tools like as Jenkins, GitLab CI, and Travis CI.
* Creating data pipelines AS a data engineer design and oversees the management of data pipelines that transfer data from multiple sources to the intended systems. Working with tools like GCP Dataflow, Apache Beam, and Apache Spark is required for this.
* Data ingestion: Identify and configure data ingestion mechanisms, such as real-time streaming or batch processing, to collect data from different sources like databases, APIs, or log files
* Writing clean, efficient, and maintainable code in Scala to create software applications or components. Understanding requirements, developing software architecture, and implementing features using proper design patterns and best practices are all part of this process.
* Building and maintaining databases: The data engineer builds and maintains databases utilizing solutions like Cloud Bigtable, Cloud Spanner, and Google Cloud SQL.
* Storing Data Files in Google Cloud S3 Buckets daily basis. Using Data Proc, Big Query to develop and maintain GCP cloud base solution.
* A Machine Learning (ML) pipeline is a series of interconnected steps that automate the process of building, training, evaluating, and deploying machine learning models.
* Used cloud shell SDK in GCP to configure the services Data Proc, Cloud Storage, Big Query, Data Flow, Cloud Composer services.
* Developing and deploying data processing applications: You will need to develop and deploy data processing applications that can handle large-scale data processing tasks.
* Worked on google cloud platform (GCP) services like compute engine, cloud load balancing, cloud storage, cloud SQL, stack driver monitoring and cloud deployment manager.

**Environment**: AWS EMR, S3, Spark, Azure Data factory, Scala, Hive, MapReduce, Sqoop, ETL, Java, GCP

**Client Company: Accenture, Hyderabad** **Dec 2018 – Dec 2021**

**Role: Data Engineer/Machine Learning Engineer**

**Responsibilities**:

* Involved in complete Big Data flow of the application starting from data ingestion upstream to HDFS, processing the data in HDFS and analyzing the data and involved.
* Involved in developing roadmap for migration of enterprise data from multiple data sources like SQL Server, provider databases into S3 which serves as a centralized datahub across the organization.
* Loaded and transformed large sets of structured and semi structured data from various downstream systems.
* Developed ETL pipelines using Spark and Hive for performing various business specific transformations.
* Building data applications and automating the pipelines in Spark for bulk loads as well as Incremental Loads of various Datasets.
* Worked on building real time pipelines using Kafka and Spark Streaming.
* Worked closely with our data scientist team’s and business consumers to shape the datasets as per the requirements.
* Recreating existing application logic and functionality in the Azure Data Lake, Data Factory, SQL Database and SQL Datawarehouse environment. experience in DWH/BI project implementation using Azure DF and data bricks.
* Experience in Azure Cloud Computing services, such as Azure Data Factory, Azure Storage accounts, Synapse, Key vaults, ADLS Gen2, Azure App insights, Azure data bricks, Arm templates, Azure devops, IR’s, Azure data flows.
* Play Framework is a web development framework that allows you to create scalable and reactive web applications.
* Automated the data pipeline to ETL all the Datasets along with full loads and incremental loads of data.
* Utilized AWS services like EMR, S3, Glue Meta store and Athena extensively for building the data applications.
* Worked on building input adapters for data dumps from FTP Servers using Apache spark.
* Spark is a large data analytics distributed data processing framework.
* Wrote spark applications to perform operations like data inspection, cleaning, load and transforms the large sets of structured and semi-structured data.
* Developed Spark application with Scala and Spark-SQL for testing and processing of data.
* Converting all types of files to parquet format in azure data factory by changing dataset and modifying mapping details.
* Automation of execution of pipelines in azure data factory by creating 0kb trigger files and using even triggers.
* Created schedule triggers in Azure data factory and implemented parameterized pipelines for automation process.
* Created Azure Data Factory pipelines to load data from multiple files to final processed layer.
* Orchestrated a Azure Data Factory frame work of pipeline to handle different files format files like excel with header, text files with different delimiters and zip files using all functionalities of copy activity like Zip, quote character and header options with the help of metadata table to send all details as parameters.
* Maintained ingestion history table to load all log details like pipeline name, runid, activity outputs by running select query on system variables in Azure data factory.
* Deploy trained models into production environments, considering factors like scalability, latency, and reliability.
* Implement model monitoring and performance tracking to ensure models remain effective and produce reliable predictions.
* Monitor and analyze model outputs, error rates, and data drift to identify issues and update models as needed.
* Collaborate with cross-functional teams, including data engineers, data scientists, software developers, and business stakeholders.
* Communicate findings, insights, and model performance to non-technical stakeholders in a clear and understandable manner.
* Present results, reports, and recommendations to management and provide actionable insights for business decision-making.
* Ensure compliance with data privacy and ethical guidelines in handling sensitive data and model outputs.
* Address biases and fairness concerns in machine learning models, promoting transparency and responsible AI practices.
* ous Learning and Improvement:
* Stay updated with the latest advancements in machine learning, AI algorithms, and tools.
* Explore new techniques, methodologies, and research papers to improve model performance and solve complex problems.
* Participate in knowledge sharing, attend conferences, and contribute to the machine learning and AI community.
* Naive Bayes is a probabilistic model based on Bayes' theorem. It assumes independence between input features and calculates the probability of an instance belonging to a certain class.
* Neural networks are a class of models inspired by the human brain. They consist of interconnected layers of artificial neurons and are capable of learning complex patterns and relationships in data.
* GBM is an ensemble model that combines weak predictive models (typically decision trees) in a sequential manner to improve prediction accuracy.
* KNN is a simple yet effective model that predicts the class of an instance based on the classes of its nearest neighbors in the feature space.
* Deep learning models, such as Convolutional Neural Networks (CNNs) for image recognition and Recurrent Neural Networks (RNNs) for sequence data, are specifically designed to handle complex patterns and large-scale datasets.
* Develop and validate machine learning models using appropriate algorithms and techniques.
* Collaborate with MLOps engineers and data engineers to deploy models into production environments.
* Design and implement model monitoring and performance evaluation frameworks.
* Continuously improve models based on feedback and real-world performance.
* Collaborate with stakeholders to understand business requirements and translate them into machine learning solutions.
* Conduct experiments and perform A/B testing to evaluate model performance and make data-driven decisions.
* Provide infrastructure support and ensure the smooth deployment and operation of machine learning systems.
* Implement automation and orchestration processes for the deployment and scaling of machine learning models.
* Collaborate with MLOps engineers and data engineers to ensure system reliability, scalability, and availability.
* Implement CI/CD pipelines for machine learning model deployment and version control.
* Monitor and optimize system performance, including resource allocation, network configurations, and security
* Collaborate with MLOps engineers, data scientists, and software engineers to define business requirements and success criteria.
* Provide domain expertise and guidance on the use of machine learning models.
* Validate and evaluate model performance against business objectives.
* Participate in the feedback loop for model improvement and iteration.
* Working with cross-functional teams that include other engineers, designers, and product managers. Participating in code reviews, providing technical help, and ensuring efficient communication and collaboration are all part of the job.
* Reporting the spark job stats, monitoring, and running data quality checks are made available for each Datasets.
* Ensuring data quality, reliability, and integrity across the data pipeline and maintaining a robust data governance framework
* Develop and maintain Terraform modules, templates, and scripts to automate infrastructure deployment and configuration.
* To guarantee that data processing systems satisfy their needs and requirements, you will need to work collaboratively with other teams, including data analysts, data scientists, and business stakeholders.

**Environment**: AWS EMR, ADF, S3, Spark, Scala, Hive, MapReduce, Sqoop, ETL, Java, GCP, Terraform

**Client** **: IBM, Hyderabad Apr 2016 - Nov 2018**

**Role: Data Engineer/Hadoop Developer**

**Responsibilities:**

* Designed end to end scalable architecture to solve business problems using various Azure Components like HDInsight, Data Factory, Data Lake, Storage and Machine Learning Studio.
* Developed JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the SQL Activity.
* Implementing data validation and data cleansing methods will be your responsibility for assuring the accuracy of the data.
* Monitoring and problem-solving: You oversee keeping an eye out for performance problems in Snowflake clusters and data processing applications, as well as solving problems when they come up.
* Implementing security and compliance guidelines for Snowflake clusters and data processing apps will be your responsibility. This could entail handling encryption keys, establishing role-based access control (RBAC), and making sure that data protection laws are followed.
* Working together to ensure that data processing applications satisfy their objectives and expectations, you will be expected to work together with other teams, including data analysts, data scientists, and business stakeholders.
* It involves collecting and preprocessing data, feature engineering, model training and selection, model evaluation, and finally, deploying the trained model into production. ML pipelines typically involve a combination of data processing
* Managing and optimizing data infrastructure: You will need to manage and optimize the data infrastructure to ensure that it is scalable, reliable, and cost-effective.
* Data Stack refers to the collection of technologies, tools, and frameworks used for building and managing data-driven solutions. It includes components for data storage, processing, integration, analysis, and visualization.
* Created parameterized pipelines using parameters and variables in azure data factory which are used for incremental, UPSERT and truncate-Load pipelines load strategies.
* Orchestrating Azure data factory pipelines for migrating data from different files with format types like parquet, ORC, JSON, CSV, and fixed width files.
* Using web activity for calling URLs of API’s and azure functions into azure data factory while executing pipelines.
* Deploying Azure data factory code and SQL scripts for inserting metadata using Azure DevOps from one environment to another.
* Managed multiple activities in Azure data factory.
* Designing and deploying Azure Solutions using Azure Data Factory, Synapse, Azure Databricks, Azure Data flows, Storage accounts.
* In overall, the goal as a Snowflake developer is to build scalable, effective, and secure data processing applications that offer business users accurate and dependable data.
* You should be well-versed in big data technologies, programming languages, and Snowflake data warehousing components.
* As we should also have experience with data architecture, ETL procedures, and data quality.
* Designing, creating, testing, and maintaining Hadoop-based data processing applications is the core responsibility of a Hadoop developer.
* Elastic Data Warehouse, which separates compute and storage resources, allowing independent scaling of each component.
* The Data Stack can vary based on specific requirements, but commonly includes components like databases (relational, NoSQL, or data warehouses), data processing frameworks (Hadoop, Spark, or Flink), data integration tools (ETL/ELT), data visualization platforms (Tableau, Power BI), and other specialized tools for machine learning, natural language processing, or real-time analytics.
* Disaster recovery and backup: Implement data backup and disaster recovery mechanisms to ensure data resilience and business continuity in case of failures or outages
* Data architecture design and implementation: You oversee coming up with scalable, effective, and economical data architecture designs and implementations.
* This could entail deciding on the best Hadoop distribution, creating data models, and creating data flows.
* Creating applications for data processing: You oversee creating Hadoop-based programs that extract, transform, and load (ETL) data from a variety of sources into the required target data store.
* This could entail creating unique MapReduce jobs in Java or using technologies like Pig, Hive, or Spark.
* Building and maintaining the data infrastructure: You oversee setting up and maintaining Hadoop clusters, configuring the network, and installing the data storage systems.
* Working with big data technologies: I ought to be familiar with tools like Hadoop, Spark, and Hive. To ensure optimum performance, you will oversee configuring and tweaking these technologies.
* Monitoring and problem-solving: You oversee keeping an eye on the performance of Hadoop clusters and the applications used for data processing, as well as solving problems when they come up.
* Developed Spark applications using Scala and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Written multiple Hive UDFS using Core Java and OOPS concepts and spark functions within Python programs.
* Snowflake is a cloud-based data warehousing platform that provides a scalable and fully-managed solution for storing and analyzing large volumes of data. It offers a unique architecture called the Snowflake
* Written multiple MapReduce programs in Java for data extraction, transformation and aggregation from multiple file formats including XML, JSON, CSV, and other compressed file formats.
* Collaborated with Business Analysts, SMEs across departments to gather business requirements, and identify workable items for further development.
* Partnered with ETL developers to ensure that data is well cleaned, and the data warehouse is up to date for reporting purpose by Pig.
* Selected and generated data into csv files and stored them into AWS S3 by using AWS EC2 and then structured and stored in AWS Redshift.
* Deployed Machine Learning models for item-item similarity on Amazon Sage Maker (AWS).
* Deploy services on AWS and utilized step function to trigger the data pipelines.
* Collecting and aggregating large amounts of log data using Apache Flume and staging data in HDFS for further analysis.
* Creation of catalog database tables for Athena for fast querying S3 data and setting up for query performance at Athena without moving data into Redshift.
* Used the AWS Sage Maker to quickly build, train and deploy the machine learning models.
* Created plugins to extract data from multiple sources like Apache Kafka, Database and Messaging Queues.
* Experience in DWH/BI project implementation using Azure Data Factory.
* Developing parser and loader MapReduce application to retrieve data from HDFS and store to HBase and Hive.
* Configured, designed, implemented, and monitored Kafka cluster and connectors.
* Responsible for ingesting large volumes of IOT data to Kafka.
* Wrote Kafka producers to stream the data from external rest APIs to Kafka topics.
* Worked with teams to use KSQL for real-time analytics.
* Worked with multiplexing, replicating and consolidation in Flume.
* Used OOZIE operational Services for batch processing and scheduling workflows dynamically.

**Environment:** Spark (PySpark, Spark SQL, Spark Streaming, Spark MLlib), Kafka, Python 3.x (Scikit-learn, NumPy, Pandas), Tableau 10.1, GitHub, AWS EMR/EC2/S3/Redshift/Glue, and Pig, Ubuntu 16.04, Hadoop 2.0, Spark (PySpark, Spark streaming, Spark SQL, Spark MLlib), NiFi, Jenkins, Pig 0.15, Python 3.x(Nltk, Pandas), Tableau 10.3, GitHub, AWS Redshift.