**Sr. Data Engineer**

**Name: Mounika**

**Ph no: 8602811711**

**Email: monikayesh2023@gmail.com**

**Professional Summary:**

* 10years of experience in designing, developing, and maintaining data pipelines and architectures using various technologies such as python, SQL, snowflake, AWS.
* Strong experience and knowledge of real-time data analytics using Spark Streaming, Kafka, and Flume
* Configured Spark streaming to get ongoing information from Kafka and store the stream information in HDFS.
* Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift
* Used various spark Transformations and Actions for cleansing the input data and used Jira for ticketing and tracking issues and Jenkins for continuous integration and continuous deployment.
* Created Data Frames and performed analysis using Spark SQL.
* Hands-on expertise in writing different RDD (Resilient Distributed Datasets) transformations and actions using Scala, Python, and Java.
* Experience as Azure Cloud Data Engineer in Microsoft Azure Cloud technologies including Azure Data Factory (ADF), Azure Data Lake Storage (ADLS), Azure Synapse Analytics (SQL Data warehouse), Azure SQL Database, Azure Analytical services, Polybase, Azure Cosmos NoSQL DB, Azure Key vaults, Azure DevOps, Azure HDInsight Big Data Technologies like Hadoop, Apache Spark and Azure Data bricks.
* Big Data - Hadoop (MapReduce & Hive), Spark (SQL, Streaming), Azure Cosmos DB, SQL Datawarehouse, Azure DMS, Azure Data Factory, AWS Redshift, Athena, Lambda, Step Function and SQL.
* Strong knowledge in Spark ecosystems such as Spark core, Spark SQL, Spark Streaming libraries.
* Very Good experience working in Azure Cloud, Azure DevOps, Azure Data Factory, Azure Data Lake Storage, Azure Synapse Analytics, Azure Analytical services, Azure Cosmos NO SQL DB, Azure HD Insight Big Data Technologies (Hadoop and Apache Spark) and Data bricks.
* Experience in designing Azure Cloud Architecture and Implementation plans for hosting complex application workloads on MS Azure.
* Performed transformations on imported data and exported back to RDBMS.
* Developed complex mappings and loaded data from various sources into the Data Warehouse, using different transformations/stages like Joiner, Transformer, Aggregator, Update Strategy, Rank, Lookup, Filter, Sorter, Source Qualifier, Stored Procedure transformation, etc.
* Implemented POC to migrate MapReduce jobs into Spark transformations using Python.
* Demonstrated automation prowess by scripting routine AWS tasks, such as snapshot creation, using Python for increased efficiency.
* Worked in agile projects, delivering end-to-end continuous integration/continuous delivery pipelines by integrating tools like Jenkins and AWS for VM provisioning.
* Implemented continuous integration and deployment (CI/CD) through Jenkins for Hadoop jobs.
* Good knowledge of Cloudera distributions and Amazon services such as Amazon S3, AWS Redshift, Lambda, Amazon EC2, Amazon SNS, Amazon SQS and Amazon EMR.
* Worked on Amazon Web Service (AWS) to integrate EMR with Spark 2 and S3 storage and Snowflake.
* Capable of using AWS utilities such as EMR, S3, and CloudWatch to run and monitor Hadoop and Spark jobs on Amazon Web Services (AWS).
* Experienced with Dimensional modeling, data migration, data cleansing, data profiling, and ETL processes for data warehouses.
* Excellent understanding of Hadoop Architecture and good exposure to Hadoop components like Hadoop MapReduce, HDFS, HBase, Hive, Sqoop, Cassandra, Kafka, and Amazon Web services (AWS) API testing, documentation, and monitoring with Postman, which easily integrates tests into your build automation.
* Understanding of AWS, Azure webservices and at least hands on experience working in projects. Knowledge of the software development life cycle, Agile methodologies, and test-driven development.
* Capable of using AWS utilities such as EMR, S3, and CloudWatch to run and monitor Hadoop and Spark jobs on Amazon Web Services (AWS).
* Designed and executed Spark SQL code to implement business logic using Python as the programming language.
* Knowledge in installing, configuring, and using Hadoop ecosystem components like Hadoop MapReduce, HDFS, HBase, Oozie, Hive, Sqoop, Zookeeper, and Flume.
* Used Apache Flume to ingest data from different sources to sinks like Avro and HDFS.
* Excellent knowledge of Kafka Architecture.
* Integrated Flume with Kafka, using Flume as both a producer and consumer (concept of FLAFKA).
* Software development involving cloud computing platforms like Amazon Web Services (AWS).
* Capable of using AWS utilities such as EMR, S3, and CloudWatch to run and monitor Hadoop and Spark jobs on Amazon Web Services (AWS).
* Strong understanding of the entire AWS Product and Service suite, primarily EC2, S3, VPC, Lambda, Redshift, Spectrum, Athena, EMR (Hadoop), and other monitoring service products, their applicable use cases, best practices, and implementation and support considerations.
* Experience in writing Infrastructure as Code (IaC) in Terraform, AWS CloudFormation.
* Created reusable Terraform modules in AWS cloud environments.
* Worked on AWS EC2, SNS, SQS EMR, and S3 to create clusters and manage data using S3.
* Strong experience in Unix and Shell Scripting. Experience on Source control repositories like GIT.
* Extensive experience in designing and implementing continuous integration, continuous delivery, continuous deployment through Jenkins.
* Installed and configured Apache Airflow for workflow management and created workflows in Python.
* Skilled in designing and orchestrating complex data pipelines and workflows for improved data processing efficiency.
* Experienced in defining task dependencies, handling retries, and managing task execution order in Airflow DAGs.
* Capable of creating custom operators for tailored workflows, enhancing functionality.
* Proficient in scheduling intervals, monitoring task execution, and troubleshooting within Airflow.
* Expertise in developing and maintaining Directed Acyclic Graphs (DAGs) for workflow management.
* Familiar with integrating Airflow with cloud services for seamless data pipeline orchestration.
* Skilled in using the Airflow web UI and CLI for workflow management.
* Proficient in dynamically generating scalable workflows with Airflow's templating and macro features.
* Developing ETL pipelines in and out of the data warehouse using a combination of Python and Snow SQL
* I have expertise in Python and Scala, user-defined functions (UDF) for Hive and Pig using Python.
* Good experience in working with cloud environments like Amazon Web Services (AWS) EC2 and S3.
* Experience in Implementing Continuous Delivery pipelines with Maven, Ant, Jenkins, and AWS.
* Configured, supported, and maintained all networks, firewall, storage, load balancers, operating systems, and software in AWS EC2.
* Experience with using PostgreSQL in a cloud environment, such as AWS, Azure
* Experience in AWS EC2, configuring the servers for Auto scaling and Elastic load balancing.
* Developed Spark Applications that can handle data from various RDBMS (MySQL, Oracle Database) and Streaming sources.
* Actively crafted user-defined functions (UDFs) in Map-Reduce and Python for Pig and Hive to enhance data processing and analysis.
* Ensured data integrity by conducting comprehensive integrity checks using Hive queries, Hadoop, and Spark.
* Leveraged Scala to implement machine learning algorithms within Spark using Python, enhancing data-driven insights.
* Hands-on experience in developing and deploying enterprise-based applications using major Hadoop ecosystem components like MapReduce, YARN, Hive, HBase, Flume, Sqoop, Spark MLlib, Spark GraphX, Spark SQL, and Kafka. O Adept at configuring and installing Hadoop/Spark Ecosystem Components.
* Proficient with Spark Core, Spark SQL, Spark MLlib, Spark GraphX, and Spark Streaming for processing and transforming complex data using in-memory computing capabilities written in Scala. Worked with Spark to improve the efficiency of existing algorithms using Spark Context, Spark SQL, Spark MLlib, Data Frame, Pair RDD, and Spark YARN.
* Worked in building ETL pipeline for data ingestion, data transformation, and data validation on cloud service AWS, working along with data steward under data compliance.
* Worked on scheduling all jobs using Airflow scripts using Python and added different tasks to DAG, and Lambda.
* Used Pyspark for extracting, filtering, and transforming the Data in data pipelines.
* Skilled in monitoring servers using Nagios, Cloud watch, and using ELK Stack Elasticsearch Kibana
* Used Data Build Tool for transformations in ETL process, AWS lambda, AWS SQS
* Involved in designing different components of systems like Sqoop, Hadoop process involves map reduce & hive, Spark, FTP integration to down systems.
* Have written hive and spark queries using optimized ways like using window functions, and customizing Hadoop shuffle & sort parameters.
* Developed ETLs using PySpark. Used both Dataframe API and Spark SQL API.
* Using Spark, performed various transformations and actions, and the result data is saved back to HDFS from there to target database Snowflake
* Strong experience and knowledge of real-time data analytics using Spark Streaming, Kafka, and Flume
* Configured Spark streaming to get ongoing information from Kafka and store the stream information in HDFS.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **BigData Technologies** | Hadoop, MapReduce, Pig, Hive, YARN, Kafka, Flume, Sqoop, Impala, Oozie, Zookeeper, Spark, Storm, Drill, Ambari, Mahout, Cassandra, Avro, and Parquet. |
| **Programming Languages** | Python, Scala, Java, SQL, PL/SQL. |
| **Cloud Services** | Amazon EC2, AWS, AWS S3, AWS Lambda, AWS Glue, AWS EMR, IAM, Cloudwatch, Redshift |
| **Databases/RDBMS** | Oracle 11g/10g, DB2, MS-SQL Server, MySQL |
| **Scripting/Web Languages** | JavaScript, HTML5, CSS3, XML, J query, Angular, Terraform |
| **Operating Systems** | Windows, UNIX, Linux, Mac OS. |
| **Software Life Cycles** | SDLC, Waterfall and Agile models |
| **Webservices** | SOAP, REST Web Services |
| **Utilities/Tools** | Eclipse, Tomcat, ANT, Maven, Automation, PyCharm |
| **Orchestration** | Cron, Oozie, Apache Airflow |
| **DevOps Tools** | git, Azure DevOps, CI/CD, TFS, K8 |
| **Reporting Tools** | Tableau, PowerBI |
| **App/Web servers** | WebLogic, Tomcat |

**Professional Experience:**

**Client: Pet Smart, CT June 2023 – Till Date**

**Role: Sr. Data Engineer**

**Responsibilities:**

* Developed ELT jobs using Apache beam to load data into Big Query tables.
* Process and load bound and unbound Data from Google pub/sub topic to Big query using cloud Dataflow with Python.
* Devised simple and complex SQL scripts to check and validate Dataflow in various applications.
* Developed and Demonstrated the POC, to migrate on-prem workload to Google Cloud Platform using GCS, Big Query, Cloud SQL and Cloud DataProc.
* Identified and documented strategies, tools and phases in migration to Google Cloud Platform.
* Documented the inventory of modules, infrastructure, storage, components of existing On-Prem data warehouse for analysis and identifying the suitable technologies/strategies required for Google Cloud Migration.
* Experience in writing and deploying cloud functions on AWS Lambda.
* Proficient in designing and implementing complex Spark SQL-based data processing pipelines that involve ETL, and data warehousing using Spark Data Frames and Spark SQL.
* Worked with application development teams to implement serverless architectures and event-driven computing using AWS Lambda and AWS API Gateway.
* Skilled in creating and managing AWS SNS topics and subscriptions, enabling pub/sub messaging for real- time data processing and notifications.
* Created ADF pipelines to migrate RAW data to Data Lake.
* Prepared the source-target mapping document.
* Automated the job pipeline status alerts to webex space.
* Automated the CRQ (change Request) process for production deployments.
* Created Linked Services, data sets. Implemented Copy activity, Pipeline, Get Meta data, If Condition, Lookup, Set.
* Extensive experience in building and maintaining data pipelines on AWS Databricks using **Python** and SQL.
* Leveraged **TDD** practices to maintain code quality, reduce bugs, and improve the efficiency of data processing workflows.
* Design and Develop ETL Processes in **AWS Glue** to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into **AWS Redshift**.
* Designed and implemented data warehousing solutions using **AWS Redshift**, including complex data modeling, tuning, and optimization.
* Integrated AWS ECR seamlessly with AWS ECS and EKS (Elastic Kubernetes Service) for seamless deployment and scaling of containerized data services
* Variable, Filter, For Each pipeline Activities to convert data into required file format.
* Implemented activities Copy activity, Execute Pipeline, Get Meta data, If Condition, Lookup, Set Variable, Filter, For Each pipeline Activities for On-cloud ETL Processing.
* Raw Data processed using Databricks, PySpark and populated into snowflake target tables.
* Connect to the snowflake to pull the data to power BI. Developed and published power BI reports.
* Designed and implemented scalable, fault-tolerant, and highly available data architectures using AWS services such as Elastic Load Balancing, Auto Scaling, and **CloudFormation**.
* Hands-on experience in building ETL pipelines using **AWS Glue** using AWS SDKs.
* Experience in setting up **AWS CloudWatch** metrics, alarms, and dashboards to monitor and visualize data engineering workflows and infrastructure.
* Skilled in writing and managing DAGs in **Airflow**, representing data workflows as code, and enabling modularity and reusability.
* Experience with real-time streaming data processing using **AWS Databricks** Streaming and integrating with AWS services like Kinesis or **Kafka.**
* Integrated **Terraform** with other DevOps tools, such as Ansible and Jenkins, to automate infrastructure deployment pipelines and streamline continuous integration and continuous deployment (CI/CD) processes.
* Expertise in designing and implementing scalable and fault-tolerant applications using Amazon DynamoDB, a fully managed NoSQL database service in AWS.
* Design, development and implementation of performing ETL pipelines using python API (pySpark) of Apache Spark.
* Experience in solving the priority issues and involving in SOC calls while there is any production issues.
* Involving with different teams and back tracking the flows and experience in solving the critical issues.
* Familiarity with Oozie's integration with other AWS services, such as **AWS S3**, **AWS Glue**, and **AWS Lambda**, to build serverless data processing pipelines.
* Strong understanding of data security and access control in **Athena**, including AWS Identity and Access Management (IAM) roles and policies
* Developed cloud functions to trigger the cloud composer to spin up the DataProc cluster.
* Process and load bound and unbound Data from Google pub/sub topic to Big query using cloud Dataflow with Python.
* Analyzing the different databases (Teradata and Big Query) from which the data is loading into the multiple reports and fixing the issues in the reports if any.
* Involving with different teams and back tracking the flows and experience in solving the critical issues.
* Troubleshooting production issues under client defined SLA’s.
* Having experience in creating the Priority Incidents, Change Requests and Service Requests in Service Now. Experience in creating the Jira’s.

**Environment:** Cloud SQL, Big Query, ADF, DBT, Pyspark, Snowflake, Databricks, Cloud Data Proc, GCS, Cloud SQL, Power BI, Cloud Composer, Informatica Power Center 10.1, Talend 6.4 for Big Data, Hadoop, Hive, Teradata, SAS, Teradata, Spark, Python, Java, SQL Server, Service Now, Confluence.

**Client: Qualcomm, India Aug 2021 – Feb 2023 Role: Big Data Developer**

**Responsibilities:**

* To analyze the root cause for a problem raised, and provide quick solution as soon as possible.
* Responsible for importing data to HDFS using Sqoop from different RDBMS servers and exporting data using Sqoop to the RDBMS servers after aggregations for other ETL operations.
* Experienced with batch processing of data sources using Apache Spark and Elastic search.
* Experienced in implementing Spark RDD transformations, actions to implement business analysis.
* Work with the Analysts, principal architect to understand the BRD and Prepared Technical Design Documents Involved in developing pyspark code for Data transformations.
* Migrated Hive QL queries on structured into Spark QL to improve performance.
* Built data ingestion process (DIP) from sources to HDFS using pyspark.
* Data Pipeline Process (DPP) is built using pyspark to transform and populate processed data into target tables.
* Involved in creating a Power BI data model.
* Implemented row level security on data in Power BI.
* Experience in exporting the data from snowflake to power BI.
* Developed visual reports, dashboards and KPI scorecards using Power BI desktop.
* Weekly meetings with technical collaborators and active participation in code review sessions with Team.
* Optimized Map Reduce Jobs to use HDFS efficiently by using various compression mechanisms.
* Worked on partitioning HIVE tables and running the scripts in parallel to reduce run-time of the scripts.
* Experienced in creating data pipeline integrating Kafka with spark streaming application used Scala for writing applications.
* Prepared the test cases and captured the test results.
* Involved in Production deployment and post-production support.
* Used spark SQL for reading data from external sources and processes the data using Scala computation framework.
* Experienced in querying data using Spark SQL on top of Spark engine for faster data sets processing.
* Used Hive QL to analyze the partitioned and bucketed data, Executed Hive queries on Parquet
* Tables stored in Hive to perform data analysis to meet the business specification logic.
* Used Zookeeper to coordinate the servers in clusters and to maintain the data consistency.
* Created Hive tables, loaded data and wrote Hive queries that run within the map.
* Developed Sqoop and Kafka Jobs to load data from RDBMS, External Systems into HDFS and HIVE.
* Collected and aggregated large amounts of web log data from different sources such as web servers, mobile using Apache Flume and stored the data into HDFS/Cassandra for analysis.
* Experienced in Analyzing Cassandra database and compare it with other open-source NoSQL databases to find which one of them better suites the current requirements.
* Worked on SparkSQL, Created Data frames by loading data from hive tables and created prep data and stored in AWS S3.
* Load the data into Spark RDD and do in memory data Computation to generate the Output response.
* Implemented and extracted the data from hive using Spark.
* Developed Spark jobs using Scala on top of Yarn/MRv2 for interactive and Batch Analysis.
* Experience in writing Apache Spark streaming API on Big Data distribution in the active cluster environment.
* Used Spark SQL to process the huge amount of structured data.
* Import the data from different sources like Cassandra into Spark RDD Using Spark Streaming.
* Larger sized Batch and S­­­­tream processing using Spark.
* Developed automated processes for flattening the upstream data from Cassandra which in JSON format. Used Hive UDFs to flatten the JSON Data.
* Created Partitioning, Bucketing, Map side Join, Parallel execution for optimizing the hive queries.

**Environment**: Map Reduce, HDFS, AWS S3, Spring Boot, Microservices, AWS, Pyspark, Hive, Unix, Pig, SQL, Sqoop, Oozie, Shell scripting, Cron Jobs, Snowflake, Power BI, Apache Kafka, J2EE.

**Client: British American Tobacco, India** **Oct 2020 – June 2021**

**Role: Data Engineer**

**Responsibilities**

* Worked with ecommerce, Sales & Ops, Manufacturing, customer support team to implement projects like Data warehouse, Data Engineering, Data integration automation, process design, API enablement, Analytics, Data quality etc.
* Developed data pipelines to implement enterprise data warehouse in Google cloud.
* Developed ingestion layer in google data storage for manufacturing team to process daily 200GB data.
* Configuration of Azure Cloud services that includes Azure Blob, Azure Sql, Azure Data Lake.
* Creating Pipeline Azure, date moving between Azure SQL to Azure Blob and Azure Data warehouse.
* Developed Pyspark scripts using Data frames/Spark SQL and RDD in Spark for Data Aggregation, queries.
* Created various Google Cloud Terraform modules for reusability into different projects, some of the modules include Compute Engine, Compute Templates, CloudSQL, Big Query, VPC, Pub/Sub, External, and Internal Load Balancers.
* Deployed Cloud Functions to trigger services like Pub/Sub, Cloud Storage, and Data store for real-time processing of messages on Pub/Sub or process files on Cloud Storage when a Compute Engine modifies or writes to it and uploads the data and messages to Data store.
* Made use of Cloud Functions with Python to update the Instance Templates and Managed Instance Groups (MIG) to accommodate the new image versions and patching support for the existing Compute Engine Instances.
* complex SQL views, stored procs in Azure SQL DW and Hyperscale
* Designed and developed a new solution to process the NRT data by using Azure stream analytics, Azure Event Hub and Service Bus Queue.
* Created Linked service to land the data from SFTP location to Azure Data Lake.
* Created numerous pipelines in Azure using Azure Data Factory v2 to get the data from disparate source systems by using different Azure Activities like Move &Transform, Copy, filter, for each, Databricks etc.
* Configured Server less VPC Access Connector for server less environments (Cloud Functions and App Engine) to access an Internal Load Balancer which has backend service on a Managed Instance Group without exposing to the public internet.
* Created a workflow to move Cloud Asset Inventory to Big Query by creating cron jobs for scheduled exports using App Engine and importing into Big Query using Dataflow templates to better understand the utilization of Google resources.
* Integrated Azure Active Directory authentication to every Cosmos DB request sent and demoed feature to Stakeholders.
* Managed GKE and Compute Engine version upgrades and patching regularly to keep the infrastructure up to date and maintain the applications without downtime during regular hours.
* Created Kubernetes YAML configuration files to deploy Services, Ingress for the pods, and created Gitlab CI templates to deploy APIs on the Kubernetes Clusters.
* Improved daily jobs performance using data cleaning, query optimization and table partitioning.
* Created an automation process for Distribution group which receive Inventory and sales data send activation report using Talend and Big query.
* Experience in Extraction, Transformation and Loading of Data from different Heterogeneous Origin systems likes Complex JSON, XML, Flat Files, Excel, Oracle, MySQL and SQL Server, Sales force Cloud, API endpoint.
* Created a mechanism to import third party vendor orders and distributor information data using API endpoint extraction.
* Create a process to extract email attachments and send required information from Big Query.

**Environment:** Kubernetes, Gitlab, XML, Azure, Oracle, MYSQL, Excel, Spark, API’S, JSON

**Client: Dell, India Mar 2015 – Sept 2020**

**Role: Software Developer**

**Responsibilities:**

* Prepared high level design and detailed data flow documents
* Worked with Sqoop to ingest data from finance cube.
* Responsible to manage data from multiple sources.
* Process the data using Spark SQL & Data Frames as per the business requirement.
* Load the processed data into final target tables.
* Coordinate with both the on-site and off-shore team.
* Responsible for overseeing the Quality procedures related to the project.
* Create test scenarios and capture test results.

**Environment:** Hadoop Ecosystem, Sql Server, Git, Unix, Tableau