**Maneesha**Sr Data Engineer  **Email: smaneesha1982@gmail.com** Ph: **+1 (254)-323-4350**

**SUMMARY:**

* Experienced and Skilled in Software development, analysis with **13+ years. Now I’m working as a Sr Data Engineer**and strong expertise in **Spark, Hive, Data warehousing, Data Modeling.**Proficiency in using**Python/Scala,** **Azure**, **GCP,** **AWS** **Web Services, Snowflake.**Designing and deploying Data Visualization using **Tableau, PowerBI**.
* Hands on experience in developing and deploying enterprise-based applications using major components in **Hadoop ecosystem** like **Hadoop, YARN, Sqoop, Spark, Hive, Pig, Map Reduce, HBase, Flume, Kafka, Oozie** and **Zookeeper**.
* Automated resulting scripts and workflow using **Apache Airflow** and **shell scripting** to ensure daily execution in production.
* Experienced in building advanced model utilizing Data mining, Data Classification, **Data Science** techniques.
* Extensive Experience on working with **Hadoop Architecture** and the components of **Hadoop - Map Reduce, HDFS, Job Tracker, Task Tracker, Name Node** and**Data Node.**
* Experienced in using **Sqoop** to import data into **HDFS,** **Hive** from **RDBMS** and into **RDBMS** from **HDFS.**
* Experienced working with real time streaming applications using tools like **Flume, Kafka, and Spark Streaming**.
* Hands on experience in using **Amazon Web Services** like **EC2, EMR** and **S3.**
* Strong experience working with big data services on the cloud especially with**Cloud SQL, Cloud deployment manager**.
* Install and configure**Apache Airflow** for **S3 bucket** and **Snowflake data warehouse** and created **dags**to run the Airflow.
* Hands on experience in using **Big Query, Dataflow**for Business agilityto provide insights.
* Good Knowledge in **Apache Spark** and **SparkSQL** for batch and real-time data for faster data processing.
* Hands on **Bash scripting** experience and building data pipelines on **Unix/Linux** systems.
* Experience in designing and developing **Spark batch** streaming to validate, extract, transform, and load 3.5 million to 4 million records per day.
* Expertise in implementing **Data Vault** methodologies and utilizing **DBT Core** for efficient **data modeling**, transformation, and integration, ensuring optimized **ETL** processes and scalable data solutions within the **Azure** ecosystem.
* Hands-on experience in **designing** and **optimizing** data solutions on **Google Cloud Platform (GCP).** Proficient in **BigQuery, Dataflow, Cloud Storage**, and **Dataprep,** delivering scalable and efficient architectures for seamless data integration and analytics.
* Experience in developing **Scala scripts** to run in **Spark Cluster**.
* Experience on **Kafka and Spark integration** for real time data processing.
* Analysed large, structured datasets using **Hive's** **data warehousing** infrastructure.
* Highly experienced in writing **HiveQL** queries for both **managed and external tables** and good at **Hive partitioning, bucketing, and performing different types of joins on Hive tables**.
* Good understanding of **NoSQL** databases like **HBase, Cassandra, MongoDB**, and hands on work experience in writing applications on **NoSQL** databases like **HBase** and **Cassandra**.
* Used **Zookeeper** to provide coordination services to the cluster.
* Extensive experience working with **ETL** tools **Airflow /SSIS in OLAP**, and **OLTP**, **Delta Lake, Data Storage** environments. Well-versed with all stages of Software Development Life Cycle (**SDLC)** and Software Testing Life Cycle **(STLC).**
* Experience in using **Hadoop** distributions **Cloudera, Hortonworks**.
* Used Talend for **ETL** processing based on business needs and extensively used **Oozie workflow engine** to run multiple **Hive** and Pig jobs by **Direct Acyclic Graph (DAG)** of actions with control flows.
* Extensively used **Microsoft SSIS** and worked in the end-to-end implementation of **ETL** projects for Retail Banking and consumer goods and health care.
* Experience in Migrating data from **SQL Server 2008 to SQL Server 2012**.
* Worked on **Apache Flink** to implements the transformation on data streams for filtering, aggregating, update state.
* Proficient in creating **T-SQL** Stored procedures, Triggers, Constraints, and indexes and developed more than 350+ scripts
* Responsible for **Performance tuning, Optimization** of stored procedures.
* Adept in Process Management and software life cycle process with experience from Analysis design to implementation.
* Hands-on experience in **Azure Cloud Services** **(PaaS & IaaS), Azure Synapse Analytics, SQL Azure, Azure Data Factory (ADF), Azure Analysis services, Application Insights, Azure Monitoring, Key Vault,** and **Azure Data Lake**.
* Extensive hands-on experience in **Snowflake Query and Performance tuning**.
* Experienced on Azure platform for services **like ADLS, Azure Data Factory (ADF), Synapse analytics & Azure Databricks.**
* Adept in interpersonal and problem-solving skills with strong communication and coordination skills.
* Experienced working on projects which involved **SCRUM** and **AGILE** methodologies.

|  |  |
| --- | --- |
| **Big Data Ecosystems** | Hadoop, Map Reduce, HDFS, HBase, Snowflake, Spark, PySpark, Scala, Zookeeper, Hive, Pig, Sqoop, Kafka, Cassandra, Ozie, Flume, and Talend |
| **Cloud** | AWS (EC2, S3, ELB, EBS, VPC, Auto Scaling)  Azure (Azure Data Lake, Azure Data Factory (ADF), T-SQL, Azure SQL,) |
| **Programming Languages** | Java, C/C++, Scala, Python, HTML, EVB, Assembly Language |
| **Scripting Languages** | JavaScript, XML, HTML, Python and shell |
| **Databases** | NoSQL, MS SQL, MYSQL, PostgreSQL, MongoDB, CosmoDB, DocumentDB, Oracle, Microsoft SQL Server. |
| **UNIX Tools** | Apache, Yum, RPM |
| **Tools** | Eclipse, JDeveloper, JProbe, CVS, Ant, MS Visual Studio |
| **Application Servers** | Apache Tomcat, Jboss |
| **Methodologies** | Agile, UML, Design Patterns |

**TECHNICAL SKILLS:**

**EXPERIENCE:**

**NextEra Energy, FL**

***Sr. Data Engineer* Mar 2021-Present**

* Involved in designing different components of system like Sqoop, Hadoop process involves map reduce & hive, Spark, FTP integration to down systems.
* Implemented MapReduce jobs to process and analyze structured and unstructured data stored in Hadoop Distributed File System (HDFS).
* Led the design and implementation of scalable database architectures, ensuring optimal performance and reliability for mission-critical applications.
* Implemented PySpark for large-scale data processing, leveraging its capabilities to distribute data processing tasks across an AWS EMR cluster.
* Developed serverless functions using AWS Lambda for event-driven data processing and automation.
* Integrated PowerBI with data sources, ensuring real-time data updates and visualizations.
* Developed complex SQL queries and optimized database performance for AWS databases, including PostgreSQL and Aurora.
* Designed and implemented ETL jobs using AWS Glue to extract, transform, and load data from various sources into AWS data stores.
* Developed ETL’s using PySpark. Used both Dataframe API and Spark SQL API.
* Using Spark, performed various transformations and actions and the final result data is saved back to HDFS from there to target database Snowflake.
* Implemented database replication and high availability solutions on MS SQL Server, ensuring data consistency and minimizing downtime for critical applications.
* Ensured proper data governance and security measures within the AWS cloud environment
* Designed and scheduled workflows within Airflow, optimizing task dependencies and parallel processing for efficient data migration.
* Implemented Spark-based applications in Scala, facilitating the migration of data processing workflows to Databricks.
* Implemented configuration standards for MongoDB and DocumentDB, optimizing settings for performance, scalability, and security.
* Integrated Athena seamlessly with Amazon S3, allowing for efficient analysis of data stored in S3 buckets.
* Deployed and managed Apache Spark and Hadoop clusters using AWS EMR, enabling large-scale data processing and analytics.
* Create and implement new features within the lakehouse, utilizing Python for coding enhancements.
* Implemented best practices for versioning, change management, and data governance in NoSQL databases.
* Demonstrated expertise in working with Databricks Delta Lake, optimizing and tuning data pipelines for performance and efficiency.
* Utilized SnowSQL for executing SQL queries, managing data, and performing administrative tasks within the Snowflake environment.
* Managed and optimized PostgreSQL and Aurora databases, ensuring data integrity, security, and high availability.
* Designed and implemented scalable and efficient data lake architectures, integrating with data warehousing applications for comprehensive data analytics.
* Enforced security policies on AWS, including IAM configurations, encryption, and network security, to meet industry compliance standards.
* Implemented indexing strategies, query rewriting, and parallel processing to enhance the efficiency of complex SQL queries.
* Configured Spark streaming to get ongoing information from the Kafka and store the stream information to HDFS.
* Ensuring version control and documentation best practices are followed within DBT.
* Implemented data models and schemas to guarantee the accuracy, consistency, and reliability of the data stored in Databricks Delta Lake.
* Integrated Kinesis streams with AWS Lambda for event-driven, serverless data processing.
* Implemented indexing strategies in MongoDB to improve query performance on large datasets.
* Utilized AWS EMR clusters for distributed computing, enhancing the scalability of Spark jobs.
* Utilized version control systems like Bitbucket or GitLab for managing and tracking changes to code and infrastructure.
* Devised and implemented partitioning strategies in Oracle databases, optimizing data distribution and retrieval to enhance performance for large datasets
* Used Jira for ticketing and tracking issues and Jenkins for continuous integration and continuous deployment.
* Enforced standards and best practices around data catalog, data governance efforts.
* Created Datastage jobs using different stages like Transformer, Aggregator, Sort, Join, Merge, Lookup, Data Set, Funnel, Remove Duplicates, Copy, Modify, Filter, Change Data Capture, Change Apply, Sample, Surrogate Key, Column Generator, Row Generator, Etc
* Expertise in Creating, Debugging, Scheduling and Monitoring jobs using Airflow for ETL batch processing to load into Snowflake for analytical processes.
* Writing and optimizing SQL queries within DBT for efficient data transformation.
* Involved in file movements between HDFS and AWS S3 and extensively worked with S3 bucket in AWS.
* Transformed the data using AWS Glue dynamic frames with PySpark, cataloged the transformed the data using Crawlers and scheduled the job and crawler using workflow feature
* Developing and maintaining DBT models for transforming raw data into business-ready datasets.
* Implemented robust data security and privacy measures on Databricks to ensure compliance with industry standards and regulations.
* Worked on scheduling all jobs using Airflow scripts using python added different tasks to DAG, LAMBDA.
* Used Pyspark for extract, filtering and transforming the Data in data pipelines.
* Skilled in monitoring servers using Nagios, Cloud watch and using ELK Stack Elasticsearch Kibana.
* Implemented SnowPipe for real-time data ingestion, automating the loading of data into Snowflake from various sources.
* Managed Cloudera Manager for monitoring, troubleshooting, and optimizing Hadoop clusters.
* Implemented version control using Git for efficient collaboration and integrated DevOps practices and tools for automated testing and deployment.
* Implemented and optimized Spark applications on AWS EMR for large-scale data processing and analysis.
* Developed serverless functions using AWS Lambda for event-driven data processing and automation.
* Implemented effective branching strategies in GitHub to support parallel development efforts and feature isolation.
* Experience in Developing Spark applications using Spark - SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Responsible for estimating the cluster size, monitoring and troubleshooting of the Spark databricks cluster.
* Set up automated build processes in the CI/CD pipeline to ensure consistent and reproducible execution of data pipelines.
* Developed major operational data reports using advanced SQL queries in Snowflake.
* Led the development and execution of disaster recovery plans for databases, minimizing downtime and ensuring quick recovery in case of catastrophic events.
* Adapted to changing requirements and priorities in Agile projects, ensuring the timely delivery of high-quality data solutions.
* Created a library of custom functions and modules for data transformations, enhancing code maintainability and scalability.
* Collaborated with the development team to establish coding standards and best practices for creating reusable components.
* Created Unix Shell scripts to automate the data load processes to the target Data Warehouse.
* Responsible for implementing monitoring solutions in Ansible, Terraform, Docker, and Jenkins.

**Environment:** Red Hat Enterprise Linux, Hadoop, Hive, Talent, HDP, PySpark, Spark3, AWS, Athena, EMR, PL/SQL, Hadoop, Map Reduce, HDFS, DBT, Hive, Shell Script, SQOOP, Python, SQL, PostgreSQL, MongoDB, DocumentDB, spark, airflow, snowflake, GIT, GitHub, Jenkins.

**Sentara Health care, VA**

***Sr. Data Engineer*  Aug 2019 -Feb 2021**

* Worked on building centralized Data Lake on AWS Cloud utilizing primary services like S3, EMR, Redshift and Athena.
* Worked on migrating datasets and ETL workloads from On-prem (MapR Cluster) to AWS Cloud.
* Applied healthcare domain expertise, incorporating standards such as HIPAA and HITRUST into data engineering processes to ensure compliance and data security.
* Leveraged Spark and Scala to handle large-scale data processing tasks, ensuring the efficiency of data transformations and analytics during migration.
* Implemented Master Data Management strategies, particularly in managing Master Patient Index (MPI), ensuring accuracy and consistency of patient data across the healthcare ecosystem.
* Built series of Spark Applications and Hive scripts to produce various analytical datasets needed for digital marketing teams.
* Worked extensively on PYSPARK to build Big Data flow.
* Spearheaded enhancements to DocumentDB partitioning strategies, optimizing data distribution and retrieval for improved performance and scalability.
* Ensured the efficient orchestration of complex data workflows within DataStage, optimizing the flow of data through various stages.
* Leveraged MySQL Workbench for database design and modeling, streamlining the process of creating and visualizing complex database schemas.
* Build, test, and deploy predictive models through batch & API in an AWS ecosystem (Sagemaker).
* This entire pipeline was orchestrated using state machine from AWS Step Functions.
* Work on projects involving lake house architecture, ensuring scalability and flexibility.
* Worked on full spectrum of data engineering pipelines: data ingestion, data transformations and data analysis/consumption.
* Worked on automating the infrastructure setup, launching and termination EMR clusters etc.
* Implemented effective partitioning strategies in Athena to enhance data retrieval efficiency for large datasets.
* Created Hive external tables on top of datasets loaded in S3 buckets and created various hive scripts to produce series of aggregated datasets for downstream analysis. Worked on data using Sqoop from HDFS to Relational Database Systems and vice-versa. Maintaining and troubleshooting.
* Implemented and managed relational databases on AWS RDS and Aurora, optimizing configurations for performance and cost-effectiveness.
* Designed and implemented MongoDB schemas for optimal performance and flexibility.
* Extracted data from Oracle ERP using AWS data services like DMS or other relevant methods.
* Explored with Spark to improve the performance and optimization of the existing algorithms in Hadoop using Spark context, Spark-SQL, Data Frame, pair RDD.
* Developed and optimized PL/SQL stored procedures and functions in Oracle, improving data processing efficiency and supporting complex business logic.
* Deploy, configure, and manage PostgreSQL and MySQL databases on GCP Cloud SQL, ensuring optimal performance and adherence to best practices.
* Ensured Power Query scripts adhered to data governance policies, implementing data privacy and security measures as required.
* Performed data comparison between Streaming Data Platform’s real-time data with AWS S3 data and wrote complex SQL queries to analyze data present in Snowflake.
* Automated schema evolution on Delta Lake, ensuring backward compatibility.
* Created Hive Tables, loaded claims data from Oracle using Sqoop and loaded the processed data into target database.
* Performed the migration of Hive and MapReduce Jobs from on - premise MapR to AWS cloud using EMR and Qubole.
* Involved in designing and deploying multi-tier applications using all the AWS services like (EC2, Route53, S3, RDS, Dynamo DB, SNS, SQS, IAM) focusing on high-availability, fault tolerance, and auto-scaling in AWS CloudFormation
* Developed data validation tools / utility functions in PYSPARK.
* Developed Apache flows dealing with various kinds of data formats such as XML, JSON, and Avro.
* Worked on importing data from HDFS to MYSQL database and vice-versa using SQOOP.
* Configured Hive meta store with MySQL, which stores the metadata for Hive tables.
* Performed data analytics in Hive and then exported those metrics back to Oracle Database using Sqoop.
* Upgraded the Hadoop Cluster from CDH3 to CDH4, setting up high availability Cluster and integrating Hive with existing applications.
* Developed data mapping strategies to extract specific information from REST API responses.
* Integrated ADF activities with Azure SQL Database for seamless data retrieval processes.
* Implemented Git-driven CI for Databricks notebooks, ensuring code quality.
* Developed data warehouse model in Snowflake for 100 datasets using WhereScape.
* Created data sharing between two snowflake accounts.
* Worked on NoSQL support enterprise production and loading data into HBase using Impala and Sqoop.
* Performed multiple MapReduce jobs in Pig and Hive for data cleaning and pre-processing.
* Build Hadoop solutions for big data problems using MR1 and MR2 in YARN.
* Handled importing of data from various data sources, performed transformations using Hive, PIG, and loaded data into HDFS.
* Redesigned the views in Snowflake to increase the performance Enabling CI/CD for DevOps process.
* Implemented the AWS lambda functions using Python to improve the performance of the file upload and merge functionality to AWS S3 buckets within the Amazon Cloud environment.
* Loaded data into S3 buckets using AWS Glue and PySpark and was involved in filtering data stored in S3 buckets using Elasticsearch and loaded data into Snowflake.
* Created an automated process of loading data from ORACLE Database to HIVE and done analysis and reporting
* Data Extraction, aggregations and consolidation of Adobe data within AWS Glue using PySpark.
* Developed several business services using Java RESTful Web Services using Spring MVC framework.

**Environment**: Hadoop, Zookeeper, Cassandra, PySpark, Spark3, Snowflake, Apache ETL, AWS, Athena, GCP, Hive, Pig, HDFS, Flume, Airflow, Tableau, GIT, Kafka, MapReduce, MySQL, PostgreSQL, MongoDB, DocumentDB, JSON, AVRO, Teradata, Maven, SOAP.

**S&P Global, NJ**

***Big Data Engineer* May 2016-Jul 2019**

* Used Spark SQL and Hive queries for data analysis and processing.
* Implemented Scala and Spark SQL scripts to access Hive tables in Spark for faster data processing.
* Deploy and manage relational databases in GCP using Cloud SQL, ensuring high availability and scalability for applications with structured data requirements.
* Implemented error-handling mechanisms and data validation checks within Talend jobs to maintain data integrity throughout the integration process.
* Participated in business analysis and technical design sessions to develop requirements documents and ETL design specifications.
* Wrote complex SQL scripts to optimize performance and avoid Informatica look-ups for heavy data volumes.
* Designed and developed Spark SQL scripts based on functional specifications.
* Installed, configured, supported, and managed Hadoop clusters.
* Worked on Google Cloud Platform (GCP) services like compute engine, cloud load balancing, cloud storage, cloud SQL, stack driver monitoring and cloud deployment manager.
* Orchestrated and managed complex workflows using Control-M for scheduling and automating big data processing jobs.
* Developed shell scripts to automate incremental Hive table updates for generating reports in Tableau.
* Utilized Spark API on Cloudera Hadoop YARN for data analytics in Hive.
* Facilitate data transfers to GCP Cloud Storage using the Transfer Service, managing large-scale migrations and ensuring data consistency.
* Integrated Big Data projects into CI/CD pipelines, automating the build, testing, and deployment processes for increased efficiency.
* Enhanced and optimized Spark code for aggregation, grouping, and data mining tasks, including handling JSON data.
* Designed and developed ETL mappings, procedures, and schedules following standard development practices.
* Optimized Map/Reduce jobs by leveraging HDFS efficiently through various compression mechanisms.
* Collaborated with QA, Operations, and Production support teams to devise test plans and resolve data or processing issues.
* Worked on large-scale Hadoop YARN clusters for distributed data processing and analysis using Spark core, Spark SQL, DataBricks Connectors, GCP, Sqoop, Hive, and NoSQL databases.
* Create Extract, Transform, Load (ETL) processes using GCP services such as Cloud Storage, Dataflow, or Dataprep to move and transform data between systems.
* Configured job dependencies and sequencing within AutoSys to ensure the accurate execution of complex workflows.
* Developed Spark SQL scripts for query performance optimization.
* Demonstrated proficiency in Hadoop components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node, YARN, and MapReduce programming.
* Configured and optimized Azure Cosmos DB for scalability and performance, considering request units (RU/s), partition keys, and indexing policies.
* Implemented Hive UDFs and performed performance tuning for improved results.
* Tuned and developed SQL queries on HiveQL, Drill, and Spark SQL.
* Provide guidance to development team working on PySpark as ETL platform.
* Demonstrated proficiency in working with any job scheduler, adapting skills to various platforms and technologies as needed.
* Used Sqoop for importing and exporting data between Oracle DB, HDFS, and Hive.
* Developed Spark code using RDDs and Spark SQL/Streaming for efficient data processing.
* Implemented partitioning, data modeling, dynamic partitions, and buckets in Hive for optimized data access.

**Environment**: Spark SQL, Hive, Hadoop Yarn, Map Reduce, Talent, Hive QL, GCP, Sqoop, SQL Server, No SQL Databases, Python, Scala, Shell, Bash Scripting, Git.

**Wipro, INDIA**

***Python Developer* Jan 2014-Jan 2016**

* Used Django framework for application development.
* Developed user interface using, CSS, HTML, JavaScript and JQuery and Django.
* Assisted in reduction of cost and optimization of supplier selection for the CRM Applications. Also used methods for cross application use using the Class Builder and Object-Oriented programming. Generated an API Reference guide to help application developers for creating the software applications to store data using DDN S3 (compatible with Amazon S3 API) or DDN Open Stack Swift API calls.
* Ensured high quality data collection and maintaining the integrity of the data.
* Cleaned data and processed third party spending data into maneuverable deliverables within specific formats with Excel macros and python libraries. Used TDD(Test driven development) methodology.
* Used several python libraries like NumPy and Matplotlib.
* Was involved in environment, code installation as well as the SVN implementation.
* Build all database mapping classes using Django models and Cassandra.
* UsedPandas API to put the data as time series and tabular format for east timestamp data manipulation and retrieval.
* Designed and developed data management system using MySQL.
* Creating unit test/regression test framework for working/new code.
* This project also used other technologies like JQuery for java script manipulations, bootstrap for the front-end html layout.
* Designed and configured database and backend applications and programs
* Migration of data validation code from the legacy mainframe systems to ETL framework using informatica workflows and shell scripts.
* Help the Data Analysts work on informatica and Python providing the technical solutions during the development phase of the reporting projects to store data in Oracle and MongoDB.
* Worked on designing and developing the Python programs and Python API’s to exchange the data within the Qualtrics survey portal to add new eligible participants and get their federal tax election survey results for taxation and reporting needs.

**Environment:** Python, Django, Cassandra, JavaScript, RabbitMQ, SoapUI, Node.js, React, MySQL, Oracle, NumPy, Pandas API, Py Dev, Agile.

**Infosys, INDIA**

***Mainframe Developer* June 2010-Dec 2013**

* Conduct impact analysis of interface applications
* Involved in effort estimation
* Design the HLDs & LLDs
* Participate in design review meetings
* Develop application in Mainframes using JCL, COBOL Batch and CICS, DB2
* Perform unit testing to verify the functionalities of developed applications work as expected
* Work with testing team for RITs, SITs & defect fixes
* Track & fix the defects raised by testing teams
* Work with business for UAT testing
* Creating knowledge document & providing knowledge transition to support teams
* Provide post implementation support for the developed applications
* Provide solution for stability issues post implementation

**Environment:**  IBM Z/OS, MVS, JCL, COBOL, CICS, DB2, VSAM, DFSORT, JMR, Changeman, Abend-aid, File-Manager, Expeditor, SPUFI, CA7, MFT, ServiceNow, Endevor, CSF

**EDUCATION DETAILS:**

* Bachelor of Engineering in Electrical and Electronics from Visvesvaraya Technological University (VTU) – 2003.