**MOULITH**

**SR. DATA ENGINEER**

**PROFESSIONAL SUMMARY:**

* Over 9 + years of IT experience and Technical proficiency in the Data Warehousing involving Business Requirements Analysis, Application Design, Data Modeling, Development, Testing and Documentation.
* Presented Data to CEO of Apple Tim Cook
* Experience as Azure Data Engineer in Azure Cloud, Azure Data Factory, Azure Data Lake Storage, Azure Synapse Analytics, Azure Analytical services, Azure Cosmos, NoSQL DB, Azure HDInsight, Big Data Technologies (Hadoop and Apache Spark) and Data bricks.
* Worked on designing Poc's for implementing various ETL Process.
* Strong expertise in building scalable applications using various programming languages (Java, Scala and Python).
* Exposed to all aspects of Software Development Life Cycle (SDLC) such as Analysis, Planning, Developing, Testing and implementing and Post - production analysis of the projects and methodologies such as Agile, SCRUM and waterfall.
* Developed and maintained efficient data models for both transactional and analytical data systems, ensuring optimal data organization and retrieval.
* Strong understanding of Distributed systems design, HDFS architecture, internal working details of MapReduce and Spark processing frameworks.
* Solid experience developing Spark Applications for performing highly scalable data transformations using RDD, Dataframe, Spark-SQL and Spark Streaming.
* Strong Experience working with various configurations of Spark like broadcast thresholds, increasing shuffle partitions, caching, repartitioning etc., to improve the performance of the jobs.
* Worked on Spark Streaming and Structured Spark streaming including Kafka for real time data processing.
* Experience in using the cloud services like Amazon EMR, S3, EC2, Red shift and Athena.
* Experience with different cloud-based storage systems like S3, Azure Blob Storage, Azure DataLake Storage Gen 1 & Gen2.
* Solid experience in using the various file formats like CSV, TSV, Parquet, ORC, JSON and AVRO.
* Strong Experience in working with Databases like Oracle, and MySQL, Teradata and proficiency in writing complex SQL queries.
* Developed and maintained efficient data models for both transactional and analytical data systems, ensuring optimal data organization and retrieval.
* Collaborated seamlessly with cross-functional teams, including data engineers, developers, data scientists, and business analysts, to align data solutions with specific business needs and objectives.
* Experienced working with JIRA for project management, GIT for source code management, JENKINS for continuous integration and Crucible for code reviews.
* Experience working with Data Lake is a system or repository of data stored in its natural/raw format, usually object blobs or files.
* Working with Data Lake is usually a single store of all enterprise data including raw copies of source system data and transformed data used for tasks such as reporting, visualization, advanced analytics, and machine learning.
* Developed a detailed project plan and helped manage the data conversion migration from the legacy system to the target snowflake database.
* Experienced in building Automation Regressing Scripts for validation of ETL process between multiple databases like Oracle, SQL Server, Hive, and Mongo DB using Python.
* Proficiency in SQL across several dialects (we commonly write MySQL, PostgreSQL, Redshift, SQL Server, and Oracle)
* Have Extensive Experience in IT data analytics projects, Hands on experience in migrating on premise ETLs to Google Cloud Platform (GCP) using cloud native tools such as BIG query, Cloud Data Proc, Google Cloud Storage, Composer.

**TECHNICAL SKILL-SET**

|  |  |
| --- | --- |
| **Big Data Ecosystem** | HDFS, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Zookeeper, Kafka, Cassandra, Apache Spark, Spark Streaming, HBase, Flume, Impala |
| **Languages** | Shell scripting, SQL, PL/SQL, Python, R, PySpark, Pig, Hive QL, Scala, |
| **Operating Systems** | Windows (XP/7/8/10), UNIX, LINUX, UBUNTU, CENTOS. |
| **Version Control** | GITHUB, GITLAB, BITBUCKET, SVN |
| **IDE & Tools, Design** | Eclipse, Visual Studio, Net Beans, Junit, CI/CD, SQL Developer, MySQL, SQL Developer, Workbench, Tableau, VsCode |
| **Databases** | Oracle, SQL Server, MySQL, Cassandra, Teradata, PostgreSQL, MS Access, Snowflake, NoSQL Database (HBase, MongoDB). |
| **Operating Systems** | Windows 98, 2000, XP, Windows 7,10, Mac OS, Unix, Linux |
| **Cloud Technologies** | MS Azure, Amazon Web Services (AWS), ACS (Apple cloud services) |
| **Data Engineer/Big Data Tools / Cloud / Visualization / Other Tools** | Databricks, Hadoop Distributed File System (HDFS), Hive, Pig, Sqoop, MapReduce, Spring Boot, Flume, YARN, MLlib, Oozie, Zookeeper, etc. AWS, EC2, EMR, S3, Glue, Redshift, Azure Databricks, Azure Data Explorer, Azure HDInsight, KeyVault, Salesforce, Linux, Bash Shell, Unix, etc., Tableau, Power BI, SAS, Dashboard Design, Power BI, Whisper |

**PROFESSIONAL EXPERIENCE:**

**Apple , Cupertino, CA May 2022 to present**

**Sr. Data Engineer**

**Responsibilities:**

* Worked on design and implementation of data pipelines using Apache Airflow, ensuring data availability and reliability for analytics and reporting.
* Proficient in leveraging Apache Airflow DAGs to perform data backfilling, enabling historical data processing and analysis by rerunning tasks for specific time periods, thus enhancing data completeness and accuracy.
* Scheduled and automated tasks within DAGs to optimize data processing and resource allocation.
* Optimized DAG execution by configuring parallelism, concurrency, and worker settings.
* Conducted routine maintenance tasks, including DAG updates and library upgrades.
* Successfully developed and implemented customized NoSQL database solutions, aligning them with project-specific requirements and performance benchmarks
* Identified and resolved performance bottlenecks, ensuring data systems operated at peak speed and efficiency.
* Conducted performance tuning for DAGs and Airflow components to address bottlenecks and diagnosed and resolved issues related to DAG execution and task failures.
* Designed and optimized graph database structures for large clusters of nodes, ensuring exceptional performance and scalability in complex data relationships.
* Designed strategies for scaling Airflow infrastructure to accommodate increased workloads.
* Led the design and implementation of data pipelines using Apache Airflow, ensuring data availability and reliability for analytics and reporting.
* Utilized Druid for real-time analytics, enabling faster decision-making through interactive dashboards and visualizations.
* Proficiently managed code changes, actively contributed to pull request (PR) discussions, and conducted comprehensive code reviews, ensuring codebase integrity, adherence to coding standards, and collaborative development for enhanced software quality.
* Developed custom Python scripts and ETL processes to transform and load data from various sources into the data warehouse and managed data storage and retrieval using Amazon S3, HDFS, and optimized data access for performance and cost-efficiency.
* Administered MySQL and Cassandra databases, optimizing data models and query performance. Integrated reporting tools (e.g., Tableau, Power BI) to provide stakeholders with actionable insights and wrote complex SQL queries for data extraction, transformation, and analysis on Pivot UI(Reporting tool).
* Implemented CI/CD pipelines using Jenkins for automated testing, deployment, and monitoring.
* Orchestrated containerized applications using Kubernetes, ensuring scalability and high availability.
* Migrated projects, successfully transferring data between different systems with minimal downtime.
* Utilized Druid for real-time analytics, enabling faster decision-making through interactive dashboards and visualizations and developed custom Python scripts and ETL processes to transform and load data from various sources into the data warehouse.
* Managed data storage and retrieval using Amazon S3, HDFS, and optimized data access for performance and cost-efficiency.
* Proficient in shell scripting (Bash) to automate repetitive tasks, streamline workflows, and enhance system administration, resulting in increased efficiency and reduced manual intervention.
* Administered MySQL and NoSQL(Cassandra) databases, optimizing data models and query performance.
* Designed and maintained RESTful APIs for data consumption and integration with external systems.
* Integrated reporting tools (e.g., Tableau, Power BI) to provide stakeholders with actionable insights.
* Implemented CI/CD pipelines using Jenkins for automated testing, deployment, and monitoring.
* Orchestrated containerized applications using Kubernetes, ensuring scalability and high availability.
* Led data migration projects, successfully transferring data between different systems with minimal downtime.

**Environment:** Hadoop, McQueen(AWS S3), Apache Airflow, Druid, Mysql, Rio(CICD), Table plus, Postman, Cyberduck, Kubernetes, Python 3.9, git, VS code, ACS, AWS, Jenkins, Rio, Pivot UI, Radar, Slack, Data Platform, API, Data Migration, NoSQL (Cassandra)db

**Charter Communications, Negaunee, MI August 2020 to May 2022**

**Sr. Data Engineer**

**Responsibilities:**

* Meetings with business/user groups to understand the business process, gather requirements, analyze, design, development and implementation according to client requirement.
* Developed JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the SQL Activity.
* 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.
* Ingested data from RDBMS and performed data transformations, and then export the transformed data to Cassandra as per the business requirement.
* Scalable metadata handling, Streaming and batch unification are offered by Delta Lake.
* Optimized apache spark clusters using Delta lake
* Used spark sql to load data and created schema RDD on top of that which loads into hive tables and handled structured using spark sql
* Involved in converting the hql’s in to spark transformations using spark RDD with support of python and Scala
* Developed JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) in batches through Azure Databricks Workspace.
* Worked on creating tabular models on Azure analysis services for meeting business reporting requirements.
* Worked with Azure BLOB and Data lake storage and loading data into Azure SQL Synapse analytics (DW).
* Created on-demand tables on S3 files using Lambda Functions and AWS Glue using Python and PySpark.
* Build an ETL which utilizes spark jar inside which executes the business analytical model
* Optimized existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames and Pair RDD's.
* Setting up and installing Azure Databricks, Azure Data Factory, Azure Data Lake, Delta Lake
* Scheduled and automated workflows using Airflow to map the data from source to destination
* Involved in the data support team as role of bug fixes, schedule change, memory tuning, schema changes loading the historic data.
* Developed code using: Apache Spark and Scala, IntelliJ, NoSQL databases (Cassandra), Jenkins, Docker pipelines, GITHUB, Kubernetes, HDFS file System, Hive, Kafka for streaming Real time streaming data, Kibana for monitor logs et
* Worked on implementation of some check points like hive count check, Sqoop records check, done file create check, done file check and touch file lookup.
* 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 Delta Lakes for time travelling as Data versioning enables rollbacks, full historical audit trails, and reproducible machine learning experiments.
* Used Azure Databricks for fast, easy and collaborative spark-based platform on Azure.
* Used Databricks to integrate easily with the whole Microsoft stack.
* Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Databricks.
* Used Azure Data Catalog which helps in organizing and to get more value from their existing investments.
* Experienced in performance tuning of Spark Applications for setting right Batch Interval time, correct level of Parallelism and memory tuning.
* Created Partitioned and Bucketed Hive tables in Parquet File Formats with Snappy compression and then loaded data into Parquet hive tables from Avro hive tables.
* Used Azure Data Factory, SQL API and MongoDB API and integrated data from MongoDB, MS SQL, and cloud (Blob, Azure SQL DB, cosmos DB)
* Involved in designing and developing tables in HBase and storing aggregated data from Hive Table.
* Analysed the sql scripts and designed it by using PySpark SQL for faster performance.

**Environment:** Hadoop 2.x, Hive v2.3.1, Spark v2.1.3, Databricks, Lambda, Glue, Azure data grid, Azure Synapse analytics, Azure data catalog, Service bus ADF, Delta lake, Blob, cosmos DB, Python, PySpark, Java, Scala, SQL, Sqoop v1.4.6, Kafka, Airflow v1.9.0, Oozie, HBase, Oracle, Teradata, Cassandra, Tableau, Maven, Git, Jira.

**Silicon Valley Bank, CA January 2019 to July 2020**

**Data Engineer**

**Responsibilities:**

* Worked on scalable distributed data system using Hadoop ecosystem in AWS EMR and MapR (MapR data platform).
* Developed Simple to complex Map/reduce streaming jobs using Python, Hive and Pig.
* Used various compression mechanisms to optimize Map/Reduce Jobs to use HDFS efficiently.
* Used ETL component Sqoop to extract the data from MySQL and load data into HDFS.
* Created map reduce jobs that can perform entire ETL process
* Wrote Hive queries and Pig scripts to study customer behavior by analyzing the data.
* Loaded data into Hive tables from Hadoop Distributed File System (HDFS) to provide SQL-like access on Hadoop data.
* Used Unix scripting and good hands on shell scripting.
* Wrote Python scripts to process semi-structured data in formats like JSON.
* Involved in loading and transforming of large sets of structured, semi structured and unstructured data.
* Troubleshooting and finding the bugs in the Hadoop applications and to clear off all the bugs took help from the testing team.
* Designed Column families in Cassandra and Ingested data from RDBMS, performed data transformations, and then export the transformed data to Cassandra as per the business requirement.
* Involved in file movements between HDFS and AWS S3 and extensively worked with S3 bucket in AWS.
* Load data into Amazon Redshift and use AWS Cloud Watch to collect and monitor AWS RDS instances within Confidential.
* Used Python API by developing Kafka producer, consumer for writing Avro Schemes.
* Developed and executed a migration strategy to move Data Warehouse from an Oracle platform to AWS Redshift.
* Involved in performing importing data from various sources to the Cassandra cluster using Sqoop.
* Worked on creating data models for Cassandra from Existing Oracle data model.
* Developed the Pyspark code for AWS Glue jobs and for EMR.
* Worked on data ingestions techniques for batch and stream processing using AWS Batch, AWS Kinesis, AWS Data Pipeline
* Worked with AWS Databases such as RDS(Aurora), Redshift, DynamoDB and Elastic Cache
* Developed Scala scripts using both Data frames/SQL and RDD/MapReduce in z for Data Aggregation, queries and writing data back into the OLTP system through Sqoop.
* Used Spark API over Cloudera Hadoop Yarn to perform analytics on data in Hive.
* Designed and implemented an ETL framework using Java to load data from multiple sources into Hive.
* Created multi - tier java based multiple web services to read data from MongoDB.
* Created Hive tables for loading and analysing data.
* Created various reports using Tableau based on requirements with the BI team.
* Created Snow pipe for continuous data load from staged data residing on cloud gateway servers.
* Working with both Maximized and Auto-scale functionality while running the multi-cluster warehouses.
* Shared sample data using grant access to customer for UAT/BAT.
* Used Snowflake time travel feature to access historical data.
* Heavily involved in testing Snowflake to understand best possible way to use the cloud resources.
* Worked on migration of jobs from Tidal to Control-M & creating new scheduled jobs in Control-M.

**Environment:** Agile, OLTP, ETL, HDFS, Kafka, AWS, S3, Talend, Redshift, Glue, Delta Lake Lambda, Cosmos DB, MongoDB, PowerBI, Azure DevOps, Hive, Scala, pyspark. Snowflake Web UI, Snow SQL, Hadoop MapR 5.2, Hive, Hue, ServiceNow, Teradata Studio, Oracle 12c, Tableau, Hadoop Yarn, Spark Core, Spark Streaming, Spark SQL.

**Global Atlantic financial group, Indianapolis, IN Nov 2017 to Dec 2019**

**Data Engineer**

**RESPONSIBILITIES:**

* Worked on development of data ingestion pipelines using ETL tool, Talend & bash scripting with big data technologies including but not limited to Hive, Impala, Kafka, and Talend.
* Supported data quality management by implementing proper data quality checks in data pipelines.
* Delivered data engineer services like data exploration, ad-hoc ingestions, subject-matter-expertise to Data scientists in using big data technologies.
* Integrated services like Bitbucket AWS Code Pipeline and AWS Elastic Beanstalk to create a deployment pipeline.
* Created S3 buckets in the AWS environment to store files, sometimes which are required to serve static content for a web application.
* Configured S3 buckets with various life cycle policies to archive the infrequently accessed data to storage classes based on requirement.
* Worked on AWS Data Pipeline to configure data loads from S3 to into Redshift
* Used JSON schema to define table and column mapping from S3 data to Redshift
* On demand, secure EMR launcher with custom spark submit steps using S3 Event, SNS, KMS and Lambda function.
* Created EBS volumes for storing application files for use with EC2 instances whenever they are mounted to them.
* Designed columnar families in Cassandra and Ingested data from RDBMS, performed data transformations, and then exported the transformed data to Cassandra as per the business requirement.
* Build machine learning models to showcase Big data capabilities using Pyspark and MLlib.
* Enhancing Data Ingestion Framework by creating more robust and secure data pipelines.
* Implemented data streaming capability using Kafka and Talend for multiple data sources.
* Worked with multiple storage formats (Avro, Parquet) and databases (Hive, Impala, Kudu).
* Implemented the JILs to automate the jobs in production cluster.
* Worked with SCRUM team in delivering agreed user stories on time for every Sprint.
* Worked on analyzing and resolving the production job failures in several scenarios.
* Implemented UNIX scripts to define the use case workflow and to process the data files and automate the jobs.
* Implemented the JILs to automate the jobs in production cluster.

**Environment:** Python, Git, Oozie, Talend, Agile Methodology AWS (EC2, S3, EBS, ELB, RDS, SNS, SQS, VPC, Cloud formation, CloudWatch, ELK Stack), Bitbucket, Ansible, Python, Shell Scripting, PowerShell, ETL, AWS Glue, Jira, , Docker, Web Logic, Maven, Web sphere, Unix/Linux, CodeDeploy, CodePieline, CodeBuild, CodeCommit, Splunk.

**Couth InfoTech Pvt. Ltd, Hyderabad, India Oct 2016 to July 2017**

**Data Analyst/Engineer**

**RESPONSIBILITIES:**

* Research and recommend suitable technology stack for Hadoop migration considering current enterprise architecture.
* Responsible for building scalable distributed data solutions using Hadoop.
* Load and transforming of large sets of structured, semi-structured and unstructured data.
* Developed Spark jobs and Hive Jobs to summarize and transform data.
* Involved in converting Hive/SQL queries into Spark transformations using Spark data frames, Scala and Python.
* Developed Spark scripts for data analysis in both python and Scala.
* Wrote Scala scripts to make spark streaming work with Kafka as part of spark Kafka integration efforts.
* Built on-premise data pipelines using Kafka and spark for real-time data analysis.
* Created reports in TABLEAU for visualization of the data sets created and tested Spark SQL connectors.
* Implemented Hive complex UDF's to execute business logic with Hive Queries.
* Developed a different kind of custom filters and handled pre-defined filters on HBase data using API.
* Implemented Spark using Scala and utilizing Data frames and Spark SQL API for faster processing of data.
* Handled importing data from different data sources into HDFS using Sqoop and performing transformations using Hive and then loading data into HDFS.
* Exporting of a result set from HIVE to MySQL using Sqoop export tool for further processing.
* Collecting and aggregating large amounts of log data and staging data in HDFS for further analysis.
* Managing and reviewing Hadoop Log files.
* Used Sqoop to transfer data between relational databases and Hadoop.
* Worked on HDFS to store and access huge datasets within Hadoop.

**Environment:** HDFS, Sqoop, Pig, Hive, Oozie, Kafka, flume, Java, Git, Tableau.

**Careator Technologies Pvt Ltd Hyderabad, India June 2014 to Sep 2016**

**Data Analyst**

**RESPONSIBILITIES:**

* Processed data received from vendors and loading them into the database. The process was carried out on weekly basis and reports were delivered on a bi-weekly basis. The extracted data had to be checked for integrity.
* Documented data cleansing and data profiling and wrote SQLscripts to meet the business requirement.
* Generated weekly, bi-weekly reports and sent to client business team using business objects.
* Data Cleaning, merging and exporting the dataset was done in Tableau Prep.
* Ability to manage multiple projects simultaneously tracking them towards varying timelines effectively through a combination of business and technical skills.
* Good Understanding of clinical practice management, medical and laboratory billing and insurance claim with processing with process flow diagrams.
* Assisted QA team in creating test scenarios that cover a day in a life of the patient for Inpatient and Ambulatory workflows.
* Analyzed the source data and handled efficiently by modifying the data types.
* Used excel sheet, flat files, CSV files to generated Tableau adhoc reports.
* Generated context filters and used performance actions while handling huge volume of data.
* Generated tableau dashboards for sales with forecast and reference lines.
* Generated tableau dashboards with combination charts for clear understanding.

**Environment**: SQL Developer, data loading, Tableau Desktop 8.2/8.1/8.0, Tableau server / Administrator