**Sindhura Reddy**

**Email: sindhuradataengineer@gmail.com PH:813-438-2039.**

**Tampa, FL 33637**

**Visa Status - Green Card**

**Sr. Data Engineer**

**Professional Summary**

* Having 9+ years of professional IT experience which includes in Big Data ecosystem and Python technologies.
* Expertise in Hadoop architecture and various components such as HDFS, YARN, Hive, Pig, High Availability, Job Tracker, Task Tracker, Name Node, Data Node, Apache, Cassandra, and MapReduce programming paradigm.
* Experience in Hadoop cluster using Cloudera’s CDH, Hortonworks HDP.
* Worked on Airflow 1.8(Python2) and Airflow 1.9(Python3) for orchestration and am familiar with building custom Airflow operators and orchestrating workflows with dependencies involving multi-clouds.
* Experience in using various tools like Sqoop, Flume, Kafka, and Pig to ingest structured, semi-structured, and unstructured data into the cluster.
* Having hands-on experience in versioning using bit bucket.
* Good understanding of Cloud Based technologies such as AWS, GCP, and Azure.
* Proficient with Apache Spark ecosystems such as Spark and Spark Streaming using Scala and Python. · Developed highly optimized Spark applications to perform various data cleansing, validation, transformation, and summarization activities according to the requirement.
* Hands-on Experience in Spark architecture and its integrations like Spark SQL, Data Frames, and Datasets APIs.
* Extensive experience in migrating on-premise Hadoop platforms to cloud solutions using AWS and GCP.
* Profound knowledge of Root Cause Analysis, Metadata Analysis, Tableau Dashboard, Tableau Calculated Fields, SQL, and Tableau Blend for code automation, customized reports, code re-usability, and Ad-hoc reports.
* Experience in RDD architecture and implementing spark operations on RDD also optimizing transformations and actions in Spark.
* Hands-on experience working on statistical procedures in R like ggplot, shiny apps, and dashboard building.
* Expertise in creating and customizing Splunk applications, searches, and dashboards as desired by IT teams and businesses.
* Experience collecting log data from various sources and integrating it into HDFS using Flume; staging data in HDFS for further analysis.
* Designed and developed logical and physical data models that utilize concepts such as Star Schema, Snowflake Schema and Slowly Changing Dimensions.
* Experience in Infrastructure Development and Operations involving AWS Cloud Services, EC2, EBS, VPC, RDS, SES, ELB, Auto scaling, CloudFront, Cloud Formation, Elastic Cache, API Gateway, Route 53, Cloud Watch, SNS.
* Experience in changing over existing AWS infrastructure to Serverless architecture (AWS Lambda, AWS Kinesis) through the creation of a Serverless Architecture using AWS Lambda, API gateway, Route 53, S3 buckets.
* Proficiency in SQL across several dialects (we commonly write MySQL, PostgreSQL, Redshift, SQL Server, and Oracle)
* 5+ years of experience in writing python as an ETL framework and Pyspark to process huge amounts of data daily.
* Strong experience in implementing data models and loading unstructured data using HBase, Dynamo DB, and Cassandra.
* Having hands-on experience in Application Deployment using the CICD pipeline.
* Can work parallel in both GCP and Aws Clouds coherently. · Experience in implementing Spark using Scala and Spark SQL for faster processing of data.
* Strong experience in extracting and loading data using complex business logic using Hive from different data sources and building ETL pipelines to process terabytes of data daily.
* Experienced in transporting and processing real-time event streaming using Kafka and Spark Streaming. · Hands-on experience with importing and exporting data from Relational databases to HDFS, Hive, and HBase using Sqoop.
* Experienced in processing real-time data using Kafka 0.10.1 producers and stream processors and implemented stream process using Kinesis and data landed into Data Lake S3.
* Implemented various algorithms for analytics using **Cassandra** with **Spark and Scala.**
* Strong Experience in implementing Data warehouse solutions in Confidential Redshift; Worked on various projects to migrate data from on-premise databases to Confidential Redshift, RDS, and S3.\
* Adept in **statistical programming languages** like **R** and **Python** including **Big Data** Technologies like **Hadoop 2**, **HAVE**, **HDFS**, **MapReduce**, and **Spark**.
* Expertise in building CI/CD on AWS environment using AWS Code Commit, Code Build, Code Deploy, and Code Pipeline and experience in using AWS CloudFormation, API Gateway, and AWS Lambda in automation and securing the infrastructure on AWS.
* Created Automation to create infrastructure for Kafka clusters with different instances as per components in the cluster using Terraform for creating multiple EC2 instances & attaching ephemeral or EBS volumes as per instance type in different availability zones & multiple regions in AWS
* Experienced in configuring and administering the Hadoop Cluster using major Hadoop Distributions like Apache Hadoop and Cloudera.

**Professional Experience**

**Data Engineer**

**KPMG, Tampa FL May 2020 to Present**

**Responsibilities:**

* Experienced in writing Spark Applications in Scala and Python.
* Build data pipelines in airflow in GCP for ETL-related jobs using different airflow operators.
* Experience in GCP Dataproc, GCS, Cloud functions, BigQuery · Analyzed large and critical datasets using HDFS MapReduce, Kafka, Spark, HBase, Hive, Hive UDF, and Spark. · Used Kafka consumer’s API in Scala for consuming data from Kafka topics.
* Developed spark applications in Spark on the distributed environment to load a huge number of CSV files with different schema into Hive ORC tables.
* Designed SSIS Packages to transfer data from flat files, Excel SQL Server using Business Intelligence Development Studio. · Developed a POC for project migration from the on-premise Hadoop MapR system to GCP.
* Analyzed the SQL scripts and designed the solution to implement using Pyspark. · Used cloud shell SDK in GCP to configure the services Data Proc, Storage, and BigQuery.
* Coordinated with the team and Developed a framework to generate Daily Adhoc reports and Extracts from enterprise data from BigQuery.
* Developed custom aggregate functions using Spark SQL and performed interactive querying.
* Implemented Apache Airflow for authoring, scheduling, and monitoring Data Pipelines · Developed Python code to gather the data from HBase (Cornerstone) and designed the solution to implement using PySpark.
* Ran data formatting scripts in Java and created terabyte CSV files to be consumed by Hadoop MapReduce jobs. · Implemented Kafka model which pulls the latest records into Hive external tables.
* Loaded all datasets into Hive from Source CSV files using spark and Cassandra from Source CSV files using Spark/PySpark. · Develop Cloud Functions in Python to process JSON files from source and load the files to BigQuery.
* Involved in ETL, Data Integration, and Migration by writing Pig scripts.
* Exported the analyzed data to Teradata using Sqoop for visualization and to generate reports for the BI team. · Migrated the computational code in HQL to PySpark.
* Imported data into HDFS from various SQL databases and files using Sqoop and from streaming systems using Storm into Big Data Lake.
* Work related to downloading BigQuery data into pandas or Spark data frames for advanced ETL capabilities · Completed data extraction, aggregation, and analysis in HDFS by using PySpark and storing the data needed to Hive.
* Exposure to the usage of Apache Kafka develops a data pipeline of logs as a stream of messages using producers and consumers. · Sound knowledge in programming Spark using Scala.
* Populated HDFS and HBase with huge amounts of data using Apache Kafka.
* Experienced in working with various kinds of data sources such as Teradata and Oracle. Successfully loaded files to HDFS from Teradata and load loaded from HDFS to Hive and Impala.
* Installed Oozie workflow engine to run multiple Hive and Pig jobs which run independently with time and data availability.
* Involved in implementing security on Hortonworks Hadoop Cluster using Kerberos by working along with operations team to move nonsecured cluster to secured cluster.
* Created Build and Release for multiple projects in a production environment using Visual Studio Team Services
* Implemented schema extraction for Parquet and Avro file Formats in Hive/MongoDB.
* Utilized Kubernetes and Docker for the runtime environment for the CI/CD system to build, test, and deploy.
* Written Spark applications using Scala to interact with the PostgreSQL database using Spark SQL Context and accessed Hive tables using Hive Context.
* Used Spark Streaming to receive real-time data from Kafka and store the stream data to HDFS using Python and NoSQL databases such as HBase and Cassandra.
* Experience in integrating real-time data using **Kafka** and Spark **Streaming** in **Java** and **Scala** to process it coming from different source systems such as **XLSX**, **CSV**, **Database**, etc.
* Developed **PIG a**nd **Hive UDFs** in **java** for extended use of **PIG** and **Hive** and wrote **Pig Scripts** for **sorting, joining, filtering, and grouping** the data.
* Worked and troubleshot in establishing the connections between Cognos BI and AWS Athena.
* Automated builds using Maven and scheduled automated nightly builds using Jenkins. Built Jenkins pipeline to drive all microservices builds to the Docker registry and then deployed to Kubernetes.
* Involved in POC Data Extraction, aggregations, and consolidation of data within AWS Glue using PySpark.
* As a Tableau developer, supported customer service designing ETL jobs dashboards utilizing data from Redshift.
* Responsible for real-time heterogeneous replication from MySQL to Oracle DB and MySQL to AWS Redshift using the Tungsten replicator tool while performance tuning and maintaining both databases.
* Created Redshift clusters on AWS for quick accessibility for reporting needs. Designed and deployed a Spark cluster and different Big Data analytic tools including Spark, Kafka streaming, AWS, and HBase with Cloudera Distribution.

**Environment**: Spark RDD. Python, Hadoop, AWS Glue, Apache Kafka, Amazon S3, SQL, Spark, AWS cloud, GCP, Bigquery, Kafka, SQL, Cloudera, Cassandra, Tableau, NoSQL, redshift, ETL, and Spark 1.6 / 2.0 (PySpark, MLlib, EMR, EC2, and Amazon RDS). Data lake, Cloudera Stack, HBase, Hive, Impala, Pig, NiFi, Spark, Spark Streaming, ElasticSearch, Teradata, Logstash, Kibana, JAX-RS, Spring, Ubuntu, and Solar.

**Sr. Data Engineer**

**Walmart, Sunnyvale CA June 2018 to April 2020**

**Responsibilities:**

* Developed real-time data processing applications by using Scala and Python and implemented Apache Spark Streaming from various streaming sources like Kafka and JMS.
* Knowledge of PySpark and used Hive to analyze sensor data and cluster users based on their behavior in the events.
* Created BigQuery authorized views for row-level security or exposing the data to other teams.
* Experienced in writing live Real-time Processing and core jobs using Spark Streaming with Kafka as a data pipeline system.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs and Scala.
* Experience in building and architecting multiple Data pipelines, end-to-end ETL process for Data ingestion and transformation in GCP, and coordinating tasks among the team.
* Experience in building and architecting multiple Data pipelines, end-to-end ETL, and ELT processes for Data ingestion and transformation in GCP.
* Build a program with Python and apache beam and execute it in cloud Dataflow to run Data validation between the raw source files and BigQuery tables.
* Involved in loading data from Linux file systems, servers, and web services using Kafka producers and partitions.
* Applied Kafka custom encoders for custom input format to load data into Kafka Partitions.
* Implement POC with Hadoop. Extract data with Spark into HDFS.
* Used Spark SQL with Scala for creating data frames and performed transformations on data frames.
* Used Spark API over Cloudera Hadoop YARN to perform analytics on data in Hive.
* Developed code to read data stream from Kafka and send it to respective bolts through the respective stream.
* Creating Spark clusters and configuring high concurrency clusters using Azure Databricks to speed up the preparation of high-quality data.
* Developed Spark applications using Scala for easy Hadoop transitions.
* Optimized the code using PySpark for better performance.
* Used cloud shell SDK in GCP to configure the services Data Proc, Storage, BigQuery · Worked on Spark streaming using Apache Kafka for real-time data processing.
* Experienced in optimizing Hive queries and joins to handle different data sets.
* Involved in ETL, Data Integration, and Migration by writing Pig scripts.
* Developed MapReduce programs to cleanse the data in HDFS obtained from heterogeneous data sources.
* Designed and implemented MongoDB and associated RESTful web service.
* Involved in writing test cases and implementing test classes using MR Unit and mocking frameworks.
* Developed Shell, Perl, and Python scripts to automate and provide Control flow to Pig scripts.
* Used Talend tool to create workflows for processing data from multiple source systems.

**Environment**: Apache Hadoop 0.20.203, GCP, Python, AWS, Hbase, Cassandra, MySQL, SQL, Kafka, Spark, Redshift,

Snowflake, Data Modeling, BigQuery, Kafka, Scala, Cloudera Manager (CDH4), HDFS, Eclipse, Hive, Pig, Sqoop, Oozie and SQL, Oracle 11g, Apache Spark, JMS, Azure SQL Data Warehouse, HDFS, Spark API, Shell.

**Python Developer/Data Engineer**

**Charter, MO Feb 2017 to May 2018**

**Responsibilities:**

* Integrated data from various system source flat files, CSV, and Oracle using PySpark. Created efficient documentation for process/procedure rules.
* Implemented data pipeline for processing data objects into S3 data bucket with security policies and ensuring uniformity of data.Developed test scripts using python and postgreSQL for the data validation.
* Written the python scripts to load the data in to the tables.
* Used Autosys and MoveIT tools to place the file from on premise to AWS-S3 buckets.

Written the jil scripts to autosys jobs and scheduled jobs on every business day.

Used celery scheduler to places the file into tables from AWS-S3.

* Used AWS Athena to query the data on top s3 bucket (files in inbound) which helps the business people for data analysis.
* Written postgreSQL sequels according to the business transformations to place the data in to respective tables.
* Created the splunk dashboard for run time errors directed from AWS cloud using kinesis.
* Involved in software development life cycle (SDLC) of tracking the requirements, gathering, analysis, detailed design, development, system testing and user acceptance testing.
* Actively involved in developing the methods for Create, Read, Update and Delete
* (CRUD) in Active Records.
* Developed entire frontend and backend modules using Python on Django Web
* Framework by implementing MVC architecture.
* Rewrote existing python/Flask module to deliver certain format of data.
* Created script in python for calling REST APIs.
* Deployed Django web application in Apache webserver and carpathia cloud web deployment.
* Installed Kafka manager for consumer lags and for monitoring kafka metrics also this has been used for adding topics, Partitions etc.
* · Installed Kafka manager for consumer lags and for monitoring Kafka metrics also this has been used for adding topics, Partitions etc.
* Involved in designing user interactive web pages as the frontend part of the web application using various web technologies like HTML, JavaScript, JQuery and implement CSS for better appearance and feel.
* Used Celery as task queue and RabbitMQ, Redis as messaging broker to execute asynchronous tasks.
* Worked on MongoDB database concepts such as locking, transactions, indexes,
* Sharding, replication, schema design.
* Managed datasets using Panda data frames and MySQL, queried MYSQL database queries from python using python-MySQL connector and MySQLdb package to retrieve information.
* Worked with external vendors or partners to onboard external data into Target S3 buckets.
* Worked on Oozie to develop workflows to automate the ETL data pipeline.
* Perform Data Cleaning, features scaling and features engineering using pandas and NumPy packages in python.
* Successfully set up a no authentication Kafka listener in parallel with Kerberos (SASL) Listener. Also, I tested a non-authenticated user (an Anonymous user) in parallel with the Kerberos user.
* Advanced knowledge of Confidential Redshift and MPP database concepts.
* Migrated on-premise database structure to Confidential Redshift data warehouse
* Implemented Workload Management (WML) in Redshift to prioritize basic dashboard queries over more complex longer-running ad hoc queries. This allowed for a more reliable and faster reporting interface, giving sub-second query responses for basic queries.
* Day-to-day responsibility includes developing ETL Pipelines in and out of the data warehouse and developing major regulatory and financial reports using advanced SQL queries in snowflake.
* Stage the API or Kafka Data(in JSON file format) into Snowflake DB by Flattening the same for different functional services.
* Worked on continuous Integration tools Jenkins and automated jar files at the end of the day.
* Heavily involved in testing Snowflake to understand the best possible way to use the cloud resources.
* Worked extensively on **AWS** Components such as **Airflow**, Elastic Map Reduce (EMR), Athena, and **Snowflake**.
* Implemented various Data Modeling techniques for **Cassandra.**
* Possess good knowledge in creating and launching EC2 instances using AMIs of Linux, Ubuntu, RHEL, and Windows and wrote shell scripts to the bootstrap instances.
* Design an ELK system to monitor and search enterprise alerts. Installed, configured, and managed the ELK Stack for Log management within EC2 / Elastic Load balancer for Elastic Search.
* Loaded data into redshift tables from s3, DynamoDB, EMR, and EC2 using copy utility.
* Worked on MongoDB database concepts such as locking, transactions, indexes, Sharding, replication, and schema design. Created multiple databases with shared collections and chose shard key based on the requirements. Experience in managing MongoDB environment from availability, performance, and scalability perspectives.
* Applied write concern for the level of acknowledgment while MongoDB writes operations to avoid rollback.
* Integrated with UI layer using HTML, Ajax, JavaScript
* Created multi-tier java-based multiple web services to read data from MongoDB.
* Lead role in NoSQL column family design, client access software, and Cassandra tuning; during migration from Oracle-based data stores.
* CAT EDH team's goal is to get rid of all their Teradata stacks and have all data on Amazon’s S3 storage platform.
* Used AWS Redshift, S3, Spectrum, and Athena services to query large amounts of data stored on S3 to create a Virtual Data Lake without having to go through the ETL process.
* Automate **AWS** infrastructure through infrastructure as code by writing various Terraform modules and scripts by creating AWS IAM users, groups, roles, policies, custom policies, AWS Glue, Crawlers, Redshift clusters, snapshots of clusters, EC2, and S3 buckets.
* Published Power BI Reports in the required originations and Made Power BI Dashboards available in Web clients and mobile apps
* Create ETL Mappings for the Operational dashboard for various KPIs and Business Metrics, allowing powerful drill down for Detail reports to understand the data at a very detailed level.
* Used Informatica Power Center Workflow manager to create sessions, workflows, and batches to run with the logic embedded in the mapping.

Develop ETL pipelines in and out of the data warehouse using a combination of Python and Snowflake’s SnowSQL

* Extract Transform and Load data from sources Systems to Azure Data Storage services using a combination of Azure Data factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics. Data ingestion to one or more Azure services (Azure Data Lake, Azure Storage, Azure SQL, Azure DW)and processing the data in Azure Databricks.
* Implemented Security roles on SSAS cubes and Power BI services by implementing authentication and authorization methods.
* As a BI Developer, I was involved in developing different reports using Microsoft BI tools like SSIS and SSRS and developed Power BI dashboards that would help businesses operate smoothly.
* Used APIs with Python to extract and push prediction outputs in EMR.
* Created continuous integration and continuous delivery (CI/CD) pipeline on AWS that helps to automate steps in the software delivery process.
* Upgraded from Cloudera manager and CDH from 5.7.1 to CDH 5.11.2 worked with Cloudera to set up SSO login and worked on data migration in HIVE. worked on Importing and exporting data into HDFS and Hive using Sqoop.

**Environment**: CSV, Oracle, PySpark, Spark RDD, S3. JSON, Parquet, ETL, Hive, UNIX, HDFS, Power BI, SQL Server. DevOps, Microsoft Azure, Kerberos, **Scala,** Python, AWS, Kafka, Redshift, **Airflow, Snowflake, Cassandra,** Map Reduce (EMR), AWS Athena, Linux, Ubuntu, EC2, MongoDB, JavaScript, Ajax, CAT, AWS Glue, Crawlers, Power Center SnowSQL, Cloudera, Jenkins. data bricks.

**Data Engineer**

**ViacomCBS, NYC Jan 2017 to Dec 2017**

**Responsibilities:**

* Involved in Requirement Gathering, Design, and Deployment of the application using Scrum (Agile) as a Development methodology.
* Wrote Spark codes in Java to run a sorting application on the data stored on AWS.
* Hands-on experience in writing Java-based Kafka and Spark streaming programs
* Used Spark Streaming to divide streaming data into batches as an input to the Spark engine for batch processing.
* Designed batch processing jobs using Apache Spark to increase speed compared to that MapReduce jobs.
* Wrote Spark applications for data validation, cleansing, transformation, and custom aggregation.
* Automated all the jobs for pulling data from the FTP server to load data into Hive tables, using Oozie workflows.
* Involved in scheduling Oozie workflow engine to run multiple HiveQL, Sqoop, and Pig jobs.
* Designed HBase row key and data modeling to insert to HBase tables using concepts of lookup tables and staging tables.
* Developed data pipeline using Sqoop to ingest customer behavioral data and purchase histories into HDFS for analysis.
* Collected the JSON data from HTTP Source and developed Spark APIs that help to do inserts and updates in Hive tables.
* Developed shell scripts for running Hive scripts in Hive and Impala.
* Designed new VPCs for the secure access POCs, and architected and implemented multiple Amazon Glue/Spark pipelines with 100s DPUs.
* Involved in importing the real-time data to Hadoop using Kafka and implemented the Oozie job for daily imports.
* Used Azure Data Factory as an orchestration tool for integrating data from upstream to downstream systems.
* Automated jobs using different triggers (Event, Scheduled, and Tumbling) in ADF.
* Used Cosmos DB for storing catalog data and event sourcing in order processing pipelines.
* Designed and developed user-defined functions, stored procedures, and triggers for Cosmos DB
* Created DA specs and Mapping Data flow and provided the details to the developer along with HLDs.
* Implemented AWS Elastic Container Service (ECS) scheduler to automate application deployment in the cloud using Docker Automation techniques.
* ETL Restarting capability for a date or date range or from the point of failure or from the beginning.
* Extract data from different social analytic sites (Facebook, Google+, Twitter, YouTube, Ooyala, iTunes, Google Analytics, Sony games) through Pentaho DI (ETL)
* Conducted statistical analysis on healthcare data using python and various tools.
* Responsible for building scalable distributed data solutions using Hadoop.
* Develop a big data web application using Agile methodology in Scala as Scala has the capability of combining functional and object-oriented programming.
* Responsible for Designing Logical and Physical data modeling for various data sources on Confidential
* Designed and Developed ETL jobs to extract data from Salesforce replica and load it in data mart in Redshift.
* Integrated **Cassandra** as a distributed persistent metadata store to provide metadata resolution for network entities on the network
* Queried and analyzed data from **DataStax** **Cassandra** for quick searching, sorting, and grouping.
* Data Migration from existing Teradata Systems to Hortonworks HDInsight cluster on Azure
* Monitor Resources and Applications using AWS Cloud Watch, including creating alarms to monitor metrics such as EBS, EC2, ELB, RDS, S3, and SNS, and configured notifications for the alarms generated based on events defined
* Used Confidential EMR to create spark clusters and EC2 instances and imported data stored In S3
* All the jobs are integrated using complex Mappings and Workflows using Informatica power center designer and workflow manager.
* Created and provisioned different **Databricks** clusters, notebooks, jobs, and autoscaling.
* Created several **Databricks Spark** jobs with **PySpark** to perform several tables-to-table operations.
* Implemented **UDFS, UDAFS, and UDTFS** in java for the hive to process the data that can’t be performed using Hive's inbuilt functions
* Manage MongoDB upgrades from 2.6 to 3.2, including hardware migrations.
* Setting up DevOps pipelines for CI/CD on GIT, Jenkins, Nexus repository
* I have integrated product data feeds from Kafka to the Spark processing system and stored the order details in the PostgreSQL database.

**Environment**: Agile, AWS, Spark, Hive, Oozie, MapReduce, FTP, HiveQL, Sqoop, Pig, Amazon S3, HDFS, Cosmos, ETL scripts, MYSQL, Kafka, **Cassandra,** T-SQL, Spark SQL, and U-SQL. Azure Data Factory. Teradata, Cloud. EBS, EC2, ELB, RDS, S3, SNS, S3, Azure Data bricks, informatica,redshift,python,PySpark,scala,MongoDB,java, Jenkins,data bricks.

**Big Data Engineer**

**WebSoc Technologies, Hyderabad India July 2014 to Dec 2016**

**Responsibilities:**

* Worked on analyzing data in the Hadoop cluster using big data tools including Pig, Hive, and Sqoop.
* Collected large amounts of log data using Apache Flume/Sqoop and aggregated using Pig/Hive in HDFS for further analysis.
* Worked on analyzing and writing Hadoop MapReduce jobs using Java API, Pig, and Hive.
* Developed Spark SQL applications to perform complex data operations on structured and semi-structured data stored as Parquet.
* Extensively worked on creating combiners, partitioning, and Distributed cache to improve the performance of MapReduce jobs.
* Used Pig to do data transformations, event joins filter bot traffic, and some pre-aggregations before storing the data onto HDFS.
* Developed data aggregation metrics using Data frames and Dataset APIs of Spark.
* Tuned Spark application performance by setting the correct amount of memory and no. of executors.
* Coordinated with the project management team on database development timelines and project scope.
* Worked with Impala for querying the hive meta store data.
* Creating pipelines, data flows, and complex data transformations and manipulations using ADF and PySpark with Databricks
* Created and provisioned different Databricksclusters needed for batch and continuous streaming data processing and installed the required libraries for the clusters.
* Improved performance by optimizing computing time to process the streaming data and saved cost to the company by optimizing the cluster run time.
* Responsible for analysis of requirements and designing generic and standard ETL processes to load data from different source systems
* Involved in developing and documenting the ETL (Extract, Transformation, and Load) strategy to populate the Data Warehouse from various source systems
* Coordinating and leading ETL (Extract-Transform-Load) Development strategy for implementation of the Design requirement.
* Conducted statistical analysis to validate data and interpretations using Python and R, as well as presented Research findings, and status reports and assisted with collecting user feedback to improve the processes and tools.
* Implement code in Python to retrieve and manipulate data.
* Integrated Flume with Kafka and Worked on monitoring and troubleshooting the Kafka-Flume-HDFS data pipeline for real-time data ingestion in HDFS
* Worked on AWS Data Pipeline to configure data loads from S3 to Redshift
* Used JSON schema to define table and column mapping from S3 data to Redshift
* Optimizing and tuning the Redshift environment, enabling queries to perform up to 100x faster for Tableau and SAS Visual Analytics
* Provided technical solutions on MS Azure HDInsight, Hive, HBase, MongoDB, Telerik, Power BI, Spot Fire, Tableau, Azure SQL Data Warehouse Data Migration Techniques using BCP, Azure Data Factory, and Fraud prediction using Azure Machine Learning.
* User Management - creating users, assigning roles, managing permissions for **SQL Server, MongoDB** database server.
* Designed various Jenkins jobs to continuously integrate the processes and executed CI/CD pipelines using Jenkins.
* Set up Azure Repo and Pipelines for CI/CD deployment of objects.
* Transformed batch data from several tables containing tens of thousands of records from SQL Server, MySQL, PostgreSQL, and CSV file datasets into data frames using PySpark.
* Worked with **NoSQL** databases like **HBase** in creating tables to load large sets of semi-structured data coming from source systems.
* Integrated Salesforce.com with external systems like Oracle and SAP   using SOAP API and REST API.
* Worked with Cloudera 5.12.x and its different components.
* Installation and setup of multi-node Cloudera cluster on AWS cloud.
* Hand full experience on AWS services like S3 bucket, lambda, AWS Glue, Cognito, Step Function, RedShift, and Cloud Watch using CDK
* Implemented Data Quality framework using AWS Athena, Snowflake, Airflow, and Python.
* Created ETL jobs on AWS glue to load vendor data from different sources, transformations involving data cleaning, data imputation, and data mapping and storing the results into S3 buckets. The stored data was later queried using AWS Athena.

**Environment**: Hadoop Framework, CDH 5.4.0, 5.4.1, Hive, Pig, Impala, Flume, Oozie, Kafka, Spark, HBase, Java (JDK1.7, 1.8), UNIX Shell Scripting, Windows7, Linux RedHat, Eclipse, Oracle, SAP, SQL Server, Cloudera, AWS. ETL, YARN, Spark SQL, python, redshift, Azure, MongoDB, CI/CD, HDFS. RDD, Data Bricks, R programming, Athena, Power BI, Tableau, PySpark, snowflake, AWS glue, S3, Jenkins, data bricks.

**Data Engineer**

**ARISSOFT SOLUTIONS, Hyderabad, India May 2013 to June 2014**

**Responsibilities:**

* Analyzed requirements and prepared test cases with the purpose of finding defects in software for a banking client. The test cases developed for the application ensured a defect-free application for customers.
* Implemented Data Preparation team which improved test data for testing which improved testing by 6%.
* Delivered a quality product after a month of working with a client with minimal direction from the client and acceptance from the client.
* Created new datasets from raw datasets files using PROC IMPORT, LIBNAME, INFILE, and fetched data from data warehouses; manipulated customer datasets using SQL and DATA STEP statements such as SET and MERGE. Decreased execution time by creating and utilizing multiple macros.
* Worked on SAS reporting procedures like PROC MEANS, PROC SORT, PROC TRANSPOSE, PROC PRINT, and PROC TABULATE.
* 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.
* Created several Databricks Spark jobs with PySpark to perform several tables-to-table operations.
* Importing and exporting data jobs to perform operations like copying data from HDFS and to HDFS using Sqoop and developed Spark code and Spark-SQL/Streaming for faster testing and processing of data.
* Use Kafka a publish-subscribe messaging system by creating topics using consumers and producers to ingest data into the application for Spark to process the data and create Kafka topics for application and system logs.
* Advanced knowledge of Confidential Redshift and MPP database concepts.
* Analyze escalated incidences within the Azure SQL database. Implemented test scripts to support test-driven development and continuous integration.
* Responsible for ingesting data from various source systems (RDBMS, Flat files, BigData) into Azure (Blob Storage) using the framework model.
* Played a key role in migration from Adempiere LTS customizing and optimizing the new version to match the same functionalities as the legacy System, transferring the data from PostgreSQL to an Oracle Database, and building Restful web services as well.
* Deployments of Salesforce code in different sandboxes - DevOps Engineer.
* Developed and implement of Web-Lead, Lead conversion, and Entire Sales Process of the End-to-End project with Full Lifecycle Implementation of Salesforce.
* Designed EMR/Hadoop/Spark process to convert files to parquet and query on s3 using Athena, SQS, and Redshift Spectrum.
* To help developers automatically build and deploy software into production multiple times a day safely while maintaining compliance in a highly regulated financial industry. We use tools like Atlassian Bamboo, Bitbucket, Confluence, JIRA, Jenkins, Sonar type Nexus and Nexus IQ, SonarQube, Grunt, and Maven to get the job done.
* Used Amazon EMR for map reduction jobs and test locally using Jenkins.
* Used Jenkins and pipelines which helped us drive all Microservices builds out to the Docker registry and then deployed to Kubernetes.
* Setting up of **CI/CD** pipeline using continuous integration tools such as **Cloud Bees Jenkins** and automated the entire **AWS EC2, VPC, S3, SNS, RedShift, EMR** based infrastructure using **Terraform, Chef, Python, Shell, Bash scripts** and managing security groups on **AWS** and custom monitoring using **Cloud Watch.**
* Created data pipeline and ETL process to load Data from AWS S3 to RDS and redshift.
* Develop Informatica mappings to be implemented based on client requirements and for the analytics team.
* Develop Informatica SCD Type-2 Mappings based on the requirements.
* Involved in understanding requirements and in modeling activities of the attributes identified from different source systems which are in Oracle, Teradata, and CSV FILES. Data is Staged, integrated, Validated, and finally loaded the data into Teradata Warehouse using Informatica and Teradata Utilities.

**Environment:** Windows 7,windows 8. PROC IMPORT, LIBNAME, INFILE, ca, PROC REPORT, PROC FREQ, Azure Data Factory, AWS Redshift, S3, Spectrum and Athena, **Mspark.** RDBMS, Flat files, Big Data, Spark jobs, Databricks, python, kafka, Pyspark, SQL, ETL, EC2, Jenkins.

**Education:** Bachelors of Technology in Electronics and Communication Engineering from JNTU ANANTAPUR(2009-2013).