**Keerthi Pothagani
Sr. Data Engineer**

**Email: anjalimanoja123@gmail.com** **Phone: (707)-886-6157**

**Professional Summary:**

* Over 9+ years of extensive IT experience as a Data Engineer with expertise in designing data-intensive applications using the Hadoop Ecosystem and Big Data Analytical, Cloud Data engineering (AWS, Azure), Data Visualization, Data Warehouse, Reporting, and Data Quality solutions.
* Hands-on expertise with data engineering stack, including Python, SQL, R, and worked with databases like Oracle, MySQL, SQL Server, Snowflake, Mongo DB, Cassandra, and writing ETLs.
* Extensive experience with the Hadoop ecosystem, including solid knowledge of Big Data technologies such as HDFS, Spark, YARN, Kafka, Map Reduce, Apache Cassandra, HBase, Zookeeper, Hive, Oozie, Impala, Pig, and Flume.
* Hands on experience inGCP, Big Query, GCS bucket, G - cloud function, cloud dataflow, Data Proc, Stack driver.
* Architect and implement ETL and data movement solutions using Azure Data Factory, SSIS
* Extract Transform and Load data from Sources Systems to Azure Data Storage services using Azure Data Factory and HDInsight.
* Experienced in implementation of Lakehouse architecture on Azure using Azure Data Lake, Delta Lake, Delta Tables, and Databricks.
* Worked on building ETL processes to load data from multiple data sources to HDFS, perform structural modifications using Map Reduce and Hive and analyse data using visualization/reporting tools.
* Databricks job configuration, Refactoring of ETL Databricks notebooks
* Hands on experience on architecting the ETL transformation layers and writing spark jobs to do the processing.
* Knowledge of Spark Context, Spark-SQL, Data frame API, Spark Streaming, and Pair RDDs, worked extensively on PySpark to increase the efficiency and optimization of existing Hadoop approaches.
* Proficiency in multiple databases like MongoDB, Cassandra, MySQL, ORACLE, and MS SQL Server.
* Responsible to develop EEIM Application as Apache Maven project and commit to code to GIT.
* Worked with Jira, Bit Bucket and source control systems like GiT and SVN and development tools like Jenkins, Artifactory.
* Designed one-time load strategy for moving large databases to Azure SQL DWH.
* Designed and developed logical and physical data models that utilize concepts such as Star Schema, Snowflake Schema, and Slowly Changing Dimensions.
* Expertise in using Airflow and Oozie to create, debug, schedule, and monitor ETL jobs. Experience with Partitions and bucketing concepts in Hive and designed both Managed and External tables in Hive to optimize performance.
* Very keen in knowing newer techno stack that Google Cloud platform (GCP) adds.
* Expert knowledge of analytics with Big-Data and Deployment tools like MLlib, Databricks, AWS Sage maker, Docker, and Tensor Flow Serving.
* Hands-on experience on AWS components EMR, EC2, S3, RDS, IAM, Auto Scaling, Cloud Watch, SNS, Athena, Glue, Kinesis, Lambda, Redshift, Dynamo DB to ensure a secure zone for an organization in AWS public cloud.
* Experienced in building Snow pipe and In-depth knowledge of Data Sharing in Snowflake and Snowflake Database, Schema, and Table structures.
* Hands-on experience with Snowflake utilities, Snow SQL, Snow Pipe, and Big Data model techniques using Python. Storage integrations and external stages for data moments to/from - S3 / Blob to Snowflake.
* Extensively used Data loading mechanisms for semi-structured (JSON, Parquet) and structured formats (CSV, flat files) using copy and Snow pipe (SQS). Developed stream & tasks to implement the CDC in Snowflake.
* Good understanding of the Fact/Dimension data warehouse design model, including star, snowflake design methods.
* Strong Experience in working with ETL Informatica, which includes components Informatica Power Center Designer, Workflow manager, Workflow monitor, Informatica server and Repository Manager.
* Good understanding of Spark Architecture with Databricks, Structured Streaming. Setting Up AWS and Microsoft Azure with Databricks, Databricks Workspace for Business Analytics, Manage Clusters in Databricks, Managing the Machine Learning Lifecycle.
* Can work parallelly in both GCP and Azure Clouds coherently.
* 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.
* Proven experience deploying software development solutions for a wide range of high-end clients, including Big Data Processing, Ingestion, Analytics, and Cloud Migration from On-Premises to AWS Cloud.
* Experience in domains like IT, Supply-chain, Retail, Healthcare, etc., with in-depth knowledge of industry workings and challenges. Proven track record of ability, adaptability, creativity, and innovation, along with the demonstration of very strong technical and managerial skills and while successfully leading teams to strict project deadlines.
* Set up dynamic inputs using Alteryx for ETL processes to bring data from multiple sources.

**Technical Skills:** **Languages:** Python (SpaCy, Pandas, NumPy, Sci-kit learn, etc.), R (dplyr, car, zoo, etc.), SQL, JavaScript, C#, Scala, Scala HQL, XPath, PySpark, PL/SQL, Shell script, Perl script.
**Database**: Oracle, My SQL, SQL Server, Mongo DB, Snowflake, Cassandra, No SQL, Alteryx, AWS Redshift, AWS Athena, Dynamo DB, Postgre SQL, Teradata, Cosmos.
**Database Modelling**: Dimension Modelling, ER Modelling, Star Schema Modelling, Snowflake Modelling.
**Operating Systems:** Windows, Ubuntu Linux, Mac OS.
**Cloud:** AWS, Azure, GCP, Amazon AWS, EC2, EC3, MS Azure, Azure SQL Database, Azure SQL Data Warehouse, Azure Analysis Services, Azure Data Lake, Data Factory
**Data Warehousing/BI:** Informatica Power Centre, Power Exchange, IDQ, Ambari view, consumption framework
**Big Data Technologies**: Apache Hadoop, Spark, Spark SQL, MLlib, Databricks, AWS Sage maker, Tensor Flow Serving, HDFS, Map Reduce, MR Unit, YARN, Hive, PIG, HBase, Impala, Zookeeper, Sqoop, Oozie, Apache Cassandra, Scala, Flume, NiFi, Kafka, Zookeeper, Yarn, Apache Spark, Mahout, Spark MLIib.
**Visualization Tools:** Tableau, Power BI, Excel, VBA, etc.
**Development Tools:** PyCharm, Eclipse, Visual Studio, SQL\*Plus, SQL Developer, TOAD, SQL Navigator, Query Analyzer, SQL Server Management Studio, SQL Assistance, Eclipse, Postman, Jupiter Notebook, ND4J, Scikit-learn, Shogun, MLlib, H2O, Cloudera Oryx, Go Learn, Apache Singa.
**ETL Tools**: MS SQL Server Analysis Services (MSOLAP, SSAS), Integration Services (SSIS),Reporting Services (SSRS), Performance Point Server (PPS), Oracle 9i OLAP, MS Office Web Components (OWC11), Informatica, Sqoop, TDCH, Manual, etc.
**Version Controller:** Tortoise HG, Microsoft TFS, SVN, GIT, Git Hub Mload, Fast Export.

**Professional Experience:**
**Amazon, Seattle, WA                                                                        November 2022- Present.
Role: Sr Data Engineer
Responsibilities:**

* Implementing solutions utilizing Advanced AWS Components: EMR, EC2, etc. integrated with Big Data/Hadoop Distribution Frameworks: Hadoop YARN, Map Reduce, Spark, Hive, etc.
* Using AWS Athena extensively to ingest structured data from S3 into multiple systems, including Red Shift, and to generate reports.
* Creating on-demand tables on S3 files using Lambda Functions and AWS Glue using Python and PySpark.
* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Experience in developing, support and maintenance for the ETL (Extract, Transform and Load) processes using Informatica.
* Experience in developing very complex mappings, reusable transformations, sessions, and workflows using Informatica ETL tool to extract data from various sources and load into targets.
* Performing end-to-end Architecture and implementation assessment of various AWS services like Amazon EMR, Redshift, S3, Athena, Glue, and Kinesis.
* Creating AWS RDS (Relational database services) to work as Hive meta store and could combine EMR cluster’s metadata into a single RDS, which avoids the data loss even by terminating the EMR.
* Involving in code migration of quality monitoring tool from AWS EC2 to AWS Lambda and built logical datasets to administer quality monitoring on snowflake warehouses.
* Installed and configured Apache Airflow for S3 bucket and Snowflake data warehouse and created to run the Airflow.
* Load the data into Spark RDD and performed in-memory data computation to generate the output response.
* Creating 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.
* Designing and developing ETL process using Informatica 10.4 tool to load data from wide range of sources such as Oracle, flat files, salesforce, and AWS cloud.
* Created and modified several database objects such as Tables, Views, Indexes, Constraints, Stored procedures, Packages, Functions and Triggers using SQL and PL/SQL.
* Hands-on experience wif building Azure notebooks, dbutils functions using Visual Studio Code and creating deployment using GIT.
* Experience working on Version control tools like SVN and GIT revision control systems such as GitHub and JIRA to track issues.
* Wrote Python scripts to parse XML documents and load the data in database.
* Extracting and uploading data into AWS S3 buckets using Informatica AWS plugin.
* Automated resulting scripts and workflow using Apache Airflow and shell scripting to ensure daily execution in production. Queried both Managed and External tables created by Hive using Impala.
* Monitored and controlled Local disk storage and Log files using Amazon Cloud Watch. Played a key role in dynamic partitioning and Bucketing of the data stored in Hive Metadata.
* Used cloud shell SDK in GCP to configure the services Data Proc, Storage, BigQuery
* Created azure data factory (ADF pipelines) using Azure blob.
* Involved in optimized Delta tables by using Optimize command and applying Z-Order Clustering.
* Implemented SCD type1, SCDType2 logic using Azure data bricks for Delta tables.
* Performed ETL on data from different source systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics.
* Involving with extraction of large volumes of data and analysis of complex business logics to drive business -oriented insights and recommending/proposing new solutions to the business in Excel Report.
* Experienced in performance tuning of Spark Applications for setting right Batch Interval time, correct level of Parallelism and memory tuning.
Encoded and decoded JSON objects using PySpark to create and modify the data frames in Apache Spark.
* Created Build and Release for multiple projects (modules) in production environment using Visual Studio Team Services (VSTS).
* Developed ETL jobs to automate the real time data retrieval from Salesforce.com, suggest best methods for data replication from Salesforce.com.
* Used AWS data pipeline for Data Extraction, Transformation and Loading from homogeneous or heterogeneous data sources and built various graphs for business decision-making using Python mat plot library.
* Perform Data Cleaning, features scaling, features engineering using pandas and NumPy packages in Python.
* Developed PySpark and Spark SQL code to process the data in Apache Spark on Amazon EMR to perform the necessary transformations based on the STMs developed.
* Designed, developed, and managed Power BI, Tableau, Qlik View, Qlik Sense Apps including Dashboard, Reports, Storytelling.
* Created a new Power BI reports dashboard with 13 pages according to the design spec in two weeks beating the tight timeline.
* Deployed automation to production for updates the company holiday schedule based on company’s holiday policy which need to be updated yearly.
* Used Informatica Power Centre for extraction, transformation, and loading (ETL) of data in the data warehouse.
* Loading data into Snowflake tables from internal stage using Snow SQL. Prepared data warehouse using Star/Snowflake schema concepts in Snowflake using Snow SQL.
* Prepared Tableau reports and dashboards with calculated fields, parameters, sets, groups, or bins and published them on the server.
* Designed and implemented ETL pipelines on S3 parquet files on Data Lake using AWS Glue. Designed AWS architecture, Cloud migration, Dynamo DB and event processing using Lambda function.
* Experience in managing and securing the Custom AMI's, AWS account access using IAM.
* Managed storage in AWS using Elastic Block Storage, S3, created Volumes and configured Snapshots.
* Experience configuring AWS S3 and their lifecycle policies and backup files and archive files in Amazon Glacier.
* Experience in creating and maintaining the databases in AWS using RDS.
* Created monitors, alarms, notifications, and logs for Lambda functions, Glue Jobs, EC2 hosts using Cloud Watch and used for the data transformation, validate and data cleansing.
* Experience in Building and Managing Hadoop EMR clusters on AWS.
* Used AWS Beanstalk for deploying and scaling web applications and services developed with Java. Developed Scripts for AWS Orchestration. Designed tool API and Map Reduce job workflow using AWS EMR and S3.
* Used Spark -Streaming APIs to perform necessary transformations and actions on the fly for building the common learner data model which gets the data from Kinesis in near real-time.
* Worked with Snowflake cloud data warehouse and AWS S3 bucket for integrating data from multiple source system which include loading nested JSON formatted data into snowflake table.
* Used AWS glue catalogue with crawler to get the data from S3 and perform SQL query operations and JSON schema to define table and column mapping from S3 data to Redshift.
* Worked on EMR Security Configurations, to store the self-signed certificates as well as KMS keys created into it. This makes it spin up a cluster in an easy manner without modifying permissions after the call.
* Worked with Cloudera 5.12.x and its different components.
* Installation and setup of multi node Cloudera cluster on AWS cloud.
* Created Redshift clusters on AWS for quick accessibility for reporting needs. Designed and deployed a Spark clusterand different Big Data analytic tools including Spark, Kafka streaming, AWS and HBase with Cloudera Distribution.
* Involved in importing real-time data using Kafka and implemented Oozie jobs for daily imports.
* In Tableau development environment, supported customer service designing ETL jobs, dash boards utilizing data from Redshift.
* Optimizing and tuning the Redshift environment, enabling queries to perform up to 100x faster for Tableau and SAS Visual Analytics.
* Performed partitioning and Bucketing concepts in Apache Hive database, which improves the retrieval speed when someone performs a query.

**Environment:** Spark RDD. AWS Glue, Apache Kafka, Amazon S3, SQL, Spark , GCP, AWS cloud, ETL, GIT, NumPy, SciPy, pandas, Scikit-learn, Seaborn, NLTK) and Spark 1.6 / 2.0 (PySpark , MLlib, EMR, EC2, and Amazon RDS. data lake, Python, Cloudera Stack, HBase, Hive, Impala, Pig, NiFi, Spark , Spark Streaming, Elastic search, Log stash, Kibana, JAX-RS, Spring, Hibernate, Apache Airflow, Oozie, REST ful API, JSON,JAXB, XML, WSDL, MySQL, Cassandra, Mongo DB, HDFS, ELK/Splunk, Athena, Azure, tableau, Red shift, Scala, snowflake, java, Jenkins, Snow SQL.

**Papa John’s Intl, Louisville, KY                                               May 2021-October 2022
Role: Data Engineer**

**Responsibilities:**

* Designed and set up Enterprise Data Lake to provide support for various use cases, including Storing, processing, Analytics, and Reporting of voluminous, rapidly changing data by using various AWS Services.
* Used various AWS services, including S3, EC2, AWS Glue, Athena, Red Shift, EMR, SNS, SQS, DMS, and Kinesis.
* Extracted data from multiple source systems S3, Redshift, and RDS and Created multiple tables/databases by creating Glue Crawlers.
* Used AWS data pipeline for Data Extraction, Transformation, and Loading from homogeneous or heterogeneous data sources and built various graphs for business decision-making using the Python mat plot library.
* Implemented solutions utilizing Advanced AWS Components: EMR, EC2, etc., integrated with Big Data/Hadoop Distribution Frameworks: Hadoop YARN, Map Reduce, Spark, Hive, etc.
* Used AWS Athena extensively to ingest structured data from S3 into multiple systems, including Red Shift, and to generate reports.
* Exposure on IAM roles in GCP.
* Created on-demand tables on S3 files using Lambda Functions and AWS Glue using Python and PySpark.
* Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in InAzure Databricks.
* Performed end-to-end Architecture and implementation assessment of various AWS services like Amazon EMR, Redshift, S3, Athena, Glue, and Kinesis.
* Created AWS RDS (Relational database services) to work as Hive meta store and could combine EMR cluster's metadata into a single RDS, which avoids data loss even by terminating the EMR.
* Involved in code migration of quality monitoring tool from AWS EC2 to AWS Lambda and built logical datasets to administer quality monitoring on snowflake warehouses.
* Install and configure Apache Airflow for the S3 bucket and Snowflake data warehouse and create dags to run the Airflow.
* Loaded the data into Spark RDD and performed in-memory data computation to generate the output response.
* Extracting and uploading data into AWS S3 buckets using the Informatica AWS plugin.
* Automated resulting scripts and workflow using Apache Airflow and shell scripting to ensure daily execution in production. Queried both Managed and External tables created by Hive using Impala.
* Monitored and controlled Local disk storage and Log files using Amazon Cloud Watch. Played a key role in dynamic partitioning and bucketing the data stored in Hive Metadata.
* Involved with extracting large volumes of data and analysing complex business logic to derive business-oriented insights and recommend/propose new solutions to the business in Excel Reports.
* Experienced in performance tuning of Spark Applications for setting the right Batch Interval time, the correct level of Parallelism, and memory tuning.
* Encoded and decoded JSON objects using PySpark to create and modify the data frames in Apache Spark.
* Created Build and Release for multiple projects (modules) in the production environment using Visual Studio Team Services (VSTS).
* Used AWS data pipeline for Data Extraction, Transformation, and Loading from homogeneous or heterogeneous data sources and built various graphs for business decision-making using the Python mat plot library.
* Developed PySpark and Spark SQL code to process the data in Apache Spark on Amazon EMR to perform the necessary transformations based on the STMs developed.
* Used Informatica Power Centre for data extraction, transformation, and loading (ETL) in the data warehouse.
* Created GIT repositories and specified branching strategies dat best fitted the needs of the client.
* Prepared data warehouse using Star/Snowflake schema concepts in Snowflake using Snow SQL.
* Build and test the ETL process using Informatica and Python loading data into Oracle Exadata.
* Developed Spark API to import data into HDFS from MySQL, SQL Server, Oracle and created Hive tables.
* Developed Sqoop jobs to import data in Avro file format from Oracle database and created Hive tables on top of it.
* Designed and implemented ETL pipelines on S3 parquet files on data lake using AWS Glue.
* Designed AWS architecture, Cloud migration, Dynamo DB, and event processing using the Lambda function.
* Experience managing and securing Custom AMI's AWS account access using IAM.
* Managed storage in AWS using Elastic Block Storage, S3, created Volumes, and configured Snapshots.
* Experience configuring AWS S3 and its lifecycle policies and backup and archive files in Amazon Glacier.
* Experience in creating and maintaining the databases in AWS using RDS.
* Experience in Building and Managing Hadoop EMR clusters on AWS.
* Used AWS Beanstalk for deploying and scaling web applications and services developed with Java.
* Designed tool API and Map Reduce job workflow using AWS EMR and S3.
* Used Spark -Streaming APIs to perform necessary transformations and actions on the fly for building the common learner data model, which gets the data from Kinesis in near real-time.
* Worked with Snowflake cloud data warehouse and AWS S3 bucket to integrate data from multiple source systems, including loading nested JSON formatted data into snowflake table.
* Used AWS glue catalogue with crawler to get the data from S3 and perform SQL query operations and JSON schema to define table and column mapping from S3 data to Redshift.
* Involved in importing real-time data using Kafka and implemented Oozie jobs for daily imports.
* In Tableau development environment, supported customer service designing ETL jobs, dash boards utilizing data from Redshift.
* Optimizing and tuning the Redshift environment, enabling queries to perform up to 100x faster for Tableau and SAS Visual Analytics.
* Performed partitioning and Bucketing concepts in Apache Hive database, which improves the retrieval speed when someone performs a query.

**Environment:** Spark RDD, AWS Glue, Apache Kafka, Amazon S3, SQL, Spark, AWS cloud, ETL, NumPy, SciPy, pandas, Scikit-learn, Seaborn, NLTK) and Spark 1.6 / 2.0 (PySpark, MLlib, EMR, EC2, and Amazon RDS. Datalake, Python, Cloudera Stack, HBase, GCP, Hive, Impala, Pig, NiFi, Spark, Spark Streaming, Elastic search, Log stash, Kibana, JAX-RS, Spring, Hibernate, Apache Airflow, Oozie, Restful API, JSON, JAXB, XML, WSDL, MySQL, Cassandra, Mango DB, HDFS, ELK/Splunk, Athena, Azure, Tableau, Redshift, Scala, Snowflake,GIT, Jenkins, Snow SQL, JIRA, Alteryx.

**Global Atlantic financial group, Indianapolis, IN                          June 2019 - December 2020
Role: AWS Data Engineer
Responsibilities:**

* Exploring with Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark SQL, Data Frame, and Spark Yarn.
* Involved in file movements between HDFS and AWS S3 and extensively worked with S3 bucket in AWS and converted all Hadoop jobs to run in EMR by configuring the cluster according to the data size.
* Wrote Spark applications for Data validation, cleansing, transformations, and custom aggregations and imported data from different sources into Spark RDD for processing and developed custom aggregate functions using Spark SQL and performed interactive querying.
* Worked on data pipeline creation to convert incoming data to a common format, prepare data for analysis and visualization, migrate between databases, share data processing logic across web apps, batch jobs, and APIs, consume large XML, CSV, and fixed-width files and created data pipelines in Kafka to replace batch jobs with real-time data.
* Involved in converting Hive/SQL queries into Spark Transformations using Spark.
* RDDs and Scala and involved in using SQOOP for importing and exporting data between RDBMS and HDFS.
* Collected data using Spark Streaming from AWS S3 bucket in near-real-time and performs necessary Transformations and aggregation on the fly to build the common learner data model and persistence the data in HDFS.
* Created AWS Glue job for archiving data from Redshift tables to S3 (online to cold storage) as per data retention requirements and involved in managing S3 data layers and databases including Redshift and Postgres.
* Processed the web server logs by developing multi-hop flume agents by using Avro.
* Sink and loaded into Mongo DB for further analysis and worked on Mongo DB No SQL data modelling, tuning, disaster recovery and backup.
* Used Airflow users to schedule and run Data Pipelines using the flexible Python Operators and framework and to implement the pipelines allows users to streamline various business processes.
* Developed a Python Script to load the CSV files into the S3 buckets and created AWS S3 buckets, performed folder management in each bucket, managed logs, and objects within each bucket.
* Worked with different file formats like JSON, AVRO and parquet and compression techniques like snappy and developed Python code for different tasks, dependencies, SLA watcher and time sensor for each job for workflow management and automation using Airflow tool.
* Developed shell scripts for dynamic partitions adding to Hive stage table, verifying JSON schema change of source files, and verifying duplicate files in source location.
* Worked with importing metadata into Hive using Python and migrated existing tables and applications to work on AWS cloud (S3).
* Integrated Hadoop into traditional ETL, accelerating the extraction, transformation, and loading of massive structured and unstructured data.
* Involved with writing scripts in Oracle, SQL Server and Netezza databases to extract data for reporting and analysis and worked in importing and cleansing of data from various sources like DB2, Oracle, flat files onto SQL Server with high volume data.
* Container management using Docker by writing Docker files and set up the automated build on Docker HUB and installed and configured Kubernetes.
* Worked extensively with importing metadata into Hive and migrated existing tables and applications to work on Hive and AWS cloud and making the data available in Athena and Snowflake.
* Extensively used Stash Git-Bucket for Code Control and Worked on AWS Components such as Airflow, Elastic Map Reduce (EMR), Athena and Snowflake.

**Environment:** Spark, AWS, EC2, EMR, Hive, SQL Workbench, Tableau, Kibana, Sqoop, Spark SQL, Spark Streaming, Scala, Python, Hadoop (Cloudera Stack), Informatica, Jenkins, Docker, Hue, Spark, Netezza, Kafka,HBase, HDFS, Hive, Pig, Sqoop, Oracle, ETL, AWS S3, AWS Glue, GIT, Grafana.

**Highmark Health, Pittsburgh, PA                                                             October 2018 - May 2019
Role: Data Analyst
Responsibilities:**

* Involved in writing complex data queries using advanced SQL and Database concepts.
* Generated surveys and different reports. Developed SQL programs for quality checks and macros for standard reports and validations using different KPI analytical skills.
* Experienced in analysing enrolment data (Electronic Medical System and Electronic data interfaces).
* Created and executed claims processing procedures that were organized and resource efficient.
* Efficiently and independently performed HealthCare Claim (MEDICAID AND MEDICARE) analysis and created reports included as federal reports, financial arrangements, inventive agreements, therapeutic value, advantage structure, benefit design, healthcare reform, and health systems.
* Prepared summary statistics (mean, median, mode, standard deviation, minimum, maximum, sum, etc.) of quantitative variables within each data set.
* By using HL7 integrate, store, and share data electronically daily.
* Conducted hands-on data transfer, such as coding of patient identification algorithms and cost and utilization outcomes. Produced code that is logically organized and well-documented.
* Created data analysis and business analysis reports in various formats (RTF, PDF, HTML, etc.)
* Performed Data Collection, Data Cleaning, Data Visualization, and Feature Engineering using Python libraries such as Pandas and Numpy, matplotlib, and sea born.
* Applied Elbow method to select optimum clusters for K-means algorithm.
* Optimized SQL queries for transforming raw data into MySQL with Informatica to prepare structured data for machine learning. Used Tableau for data visualization and interactive statistical analysis.
* Worked with Business Analysts to understand the user requirements, layout, and look of the interactive dashboard.

**Environment:** AWS, Python, Pandas, Numpy, SciPy, Apache, ggplot, Plotly, Python matplotlib, Tableau, ggplot, Bash, Shell, SQL server, Mango DB.

**ADP, Hyderabad, Telangana, India                                          January 2017 – October 2018
Role: Data Engineer
Responsibilities:**

* Followed the Agile Methodology (Scrum) to fulfil client expectations, timelines with quality deliverables.
* Created Spark jobs by writing RDDs in Python and created data frames in Spark SQL to perform data analysis and stored in Azure Data Lake.
* Configured Spark Streaming to receive real-time data from the Apache Kafka and store the stream data to HDF Susing Scala.
* Developed Spark Applications by using Kafka and Implemented Apache Spark data processing project to handle data from various RDBMS and Streaming sources.
* Designed the Airflow scheduler for persistent service in an Airflow production environment.
* Created various data pipelines using Spark, Scala, and Spark SQL for faster processing of data.
* Designed batch processing jobs using Apache Spark to increase speed compared to that of Map Reduce jobs.
* Written Spark-SQL and embedded the SQL in SCALA files to generate jar files for submission onto the Hadoop cluster. Developed data pipeline using Flume to ingest data and customer histories into HDFS for analysis.
* Executing Spark SQL operations on JSON, transforming the data into a tabular structure using data frames, and storing and writing the data to Hive and HDFS.
* Worked with HIVE data warehouse infrastructure-creating tables, data distribution by implementing partitioning and bucketing, writing, and optimizing the HQL queries.
* Created Hive tables as per requirement were Internal or External tables defined with appropriate static, dynamic partitions, and bucketing, intended for efficiency.
* Used Hive as an ETL tool for event joins, filters, transformations, and pre-aggregations.
* Involved in moving all log files generated from various sources to HDFS for further processing through Kafka.
* Extracting real-time data using Kafka and Spark streaming by Creating DStreams and converting them into RDD, processing kit, and stored it into.
* Used Spark SQL for Scala interface that automatically converts RDD case classes to schema RDD.
* Extracted source data from Sequential files, XML files, CSV files, transformed and loaded it into the target Data warehouse. Solid understanding of No SQL Database (Mongo DB and Cassandra).
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Spark SQL and Scala extracted large datasets from Cassandra and Oracle servers into HDFS and vice versa using Sqoop.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs and Scala.
* Involved in Migrating the platform from Cloudera to EMR platform.
* Developed analytical component using Scala, Spark, and Spark Streaming.
* Used Hook for writing the low-level code that hits their API or uses special libraries and for the building blocks that Operators are built out of.
* Worked on developing ETL processes to load data from multiple data sources to HDFS using FLUME and performed structural modifications using HIVE.
* Provided technical solutions on MS Azure HD Insight, Hive, HBase, Mongo DB, Telerik, Power BI, Spot Fire, Tableau and Azure SQL Data Warehouse Data Migration Techniques using BCP, Azure Data Factory, and Fraud prediction using Azure Machine Learning.

**Environment:** Hadoop, Hive, Kafka, Snowflake, Spark, Scala, HBase, Cassandra, JSON, XML, UNIX Shell Scripting, Cloudera, Map Reduce, Power BI, ETL, MySQL, No SQL.

**Intuit, Hyderabad, Telangana, India.                               April 2015 - December 2016**

**Role: Data Engineer
Responsibilities:**

* Collaborated with business users/product owners/developers to contribute to analysing functional requirements.
* Implemented Spark SQL queries that combine Hive queries with Python programmatic data manipulations supported by RDDs and data frames.
* Used Kafka Streams to Configure Spark streaming to get information and store it in HDFS.
* Extract Real-time feed using Spark Streaming and convert it to RDD, process data in the form of a Data Frame, and save the data in HDFS. Developing Spark scripts and UDFS using Spark SQL query for data aggregation, querying, and writing data back into RDBMS through Sqoop.
* Installed and configured Hadoop Map Reduce HDFS. Developed multiple Map Reduce jobs in java for data cleaning and pre-processing.
* Installed and configured Pig and wrote Pig Latin scripts. Wrote Map Reduce job using Pig Latin.
* In Airflow, a (DAG) or a Directed Acyclic Graph, I used a collection of all the tasks you want to run and organized in a way that reflects their relationships and dependencies.
* Worked on analysing Hadoop clusters using different big data analytic tools, including HBase database and Sqoop.
* Worked on importing and exporting data from Oracle and DB2 into HDFS and HIVE using Sqoop for analysis, visualization, and generating reports.
* Creating and inserting data into Hive tables for dynamically inserting data into data tables using partitioning and bucketing for EDW tables and historical metrics.
* Experienced in handling large datasets using Partitions, Spark in Memory capabilities, Broadcasts in Spark,
Effective & efficient Joins, Transformations, and others during the ingestion process.
* Created ETL packages with different data sources (SQL Server, Oracle, Flat files, Excel, DB2, and Teradata) and loaded the data into target tables by performing various SSIS transformations.
* Designed and developed data integration programs in a Hadoop environment with No SQL data store Cassandra for data access and analysis.
* Created partitions, bucketing across the state in Hive to handle structured data using Elastic search.
* Performed Sqooping for various file transfers through the HBase tables for data processing to several No SQL DBs - Cassandra, Mongo DB.

**Environment:** Hadoop, Map Reduce, HDFS, Hive, Python, Kafka, HBase, Sqoop, No SQL, Spark 1.9, PL/SQL, Oracle, Cassandra, Mongo DB, ETL, MySQL, Python, SQL, Tableau, AWS, Data Visualization, ETL.
  **CSC, Hyderabad, Telangana, India                                            August 2014 - March 2015
Designation: Data Analyst
Responsibilities:**

* Creating and inserting data into Hive tables for dynamically inserting data into data tables using partitioning and bucketing for EDW tables and historical metrics.
* Gathered requirements, analysed, and wrote the layout documents.
* Involved in complete Agile Requirement Analysis, Development, System, and Integration Testing.
* Built various graphs for business decision making using Packages like NumPy, Pandas, Matplotlib, SciPy ggplot2
for Numerical analysis.
* Involve in records mining, transformation and loading from the supply structures to the goal system.
* Working with various Integrated Development Environments (IDEs) like Visual Studio Code and PyCharm.
* Create information fashions for AWS Red shift and Hive from dimensional information fashions.
* Documental statistics mapping and transformation techniques with inside the Functional Design files.
* Enterprise requirements Performed Data Profiling and Data Quality checks.
* Worked on exporting reviews in a couple of codecs which include MS Word, Excel, CSV, and PDF.

**Environment:** Agile, NumPy, Pandas, Matplotlib, SciPy, ggplot2, AWS Red shift and Hive.