**Senior Data Engineer**

**Sreeja Reddy**

[**sreejareddyk13@gmail.com**](mailto:sreejareddyk13@gmail.com)

**+1 (618) 407-8660**

**PROFESSIONAL SUMMARY:**

**10+** years IT experience in Data Engineering, Data Pipeline Design, Development, and Implementation across various industries including Hospitality& Entertainment, Healthcare, Retail, Financial, etc.

* Experience in designing and building Data Management Lifecycle covering **Data Ingestion, Data integration, Data consumption, Data delivery, and integration Reporting, Analytics**, and **System-System** integration.
* Proficient in **Big Data** environment and Hands-on experience in utilizing **Hadoop** environment components for large-scale data processing including structured and semi-structured data.
* Strong experience with all phases including **Requirement Analysis, Design, Coding, Testing, Support and Documentation using Apache Spark & Scala, Python, HDFS, YARN, Sqoop, Hive, Map Reduce, KAFKA.**
* Extensive experience with Azure cloud technologies like **Azure Data Lake Storage, Azure Data Factory, Azure SQL,** **Azure Data Warehouse**, **Azure Synapse Analytical**, **Azure Analytical Services**, **Azure HDInsight** and **Databricks**.
* Hands on experience in **GCP BiqQuery, GCS Bucket, G- cloud function, Cloud data flow, pub/sub cloud shell, GSUTIL, BQ Command line utilities, dataproc**.
* Solid Knowledge of **AWS** services like **AWS EMR**, **Redshift**, **S3, EC2, Lambda, Glue** and **concepts**, configuring the servers for **auto-scaling** and elastic **load balancing**.
* Experience with monitoring the web services using **Hadoop** and **Spark** for controlling the applications and **analyzing** their operation and **performance**.
* Experienced in **Python** data manipulation for loading and extraction as well as with Python libraries such as **NumPy**, **Pandas**, **matplotlib**, **seaborn**, **sklearn** and **SciPy** for data analysis and numerical computations.
* Good knowledge and experience with NoSQL databases like **HBase, Cassandra**, and **MongoDB** and **SQL** databases like **Teradata, Oracle, PostgreSQL**, and **SQL** Server.
* Experience in the development and design of various scalable systems using **Hadoop** technologies in various environments and analyzing data using **MapReduce, Hive**, and **PIG**.
* Hands-on use of Spark and Scala to compare the performance of Spark with **Hive** and **SQL**, and **Spark** **SQL** to manipulate **Data Frames** in **Scala**.
* Strong knowledge in working with **ETL** methods for data extraction, transformation, and loading in corporate- wide **ETL** Solutions and Data **Warehouse** tools for reporting and **data analysis**.
* Hands-on experience in **designing** and implementing data engineering pipelines and analyzing data using Hadoop ecosystem tools like HDFS, **Spark**, **Sqoop**, **Hive**, **Flume**, **Kafka**, **Impala**, **PySpark**, **Oozie**, and **HBase**.
* Experience with different ETL tool environments like SSIS, Informatica, and reporting tool environments like **SQL Server Reporting Services**, **Power BI** and **Business Objects**.
* Experience in deployment of applications and scripting using the **Unix/Linux Shell** scripting.
* Solid knowledge of **Data Marts, Operational Data Store, OLAP, Dimensional Data Modeling** with **Star Schema** **Modeling**, **Snowflake** Modeling for Dimensions Tables using **Analysis Services**.
* Extensive experience with various databases like **Teradata, MongoDB, Cassandra DB, MySQL, Oracle,** and **SQL Server.**
* Experience in Creating **Teradata** SQL scripts using OLAP functions like rank and rank over to improve the **query** **performance** while pulling the **data** from **large tables**.
* Proficiency in writing complex **SQL, PL/SQL** for creating **tables, views, indexes, stored procedures**, **functions**.
* Knowledge and experience with **CI/CD** using containerization technologies like **Docker** and **Jenkins**.

**Educational Details**

Bachelors from JNTU 2013

**Technical Skills:**

|  |  |
| --- | --- |
| **Big Data Technologies** | **Hadoop, MapReduce, Spark, HDFS, Sqoop, YARN, Oozie, Hive,**  **Impala, Zookeeper, Apache Flume, Apache Airflow, Cloudera,**  **HBase** |
| **Programming Languages** | **Python, Java, PL/SQL, SQL, Scala, PowerShell, C, C++, T-SQL** |
| **Cloud Services** | **Azure Data Lake Storage Gen 2, Azure Data Factory, Blob storage,**  **Azure SQL DB, Databricks, Azure Event Hubs, AWS RDS,**  **Amazon SQS, Amazon S3, AWS EMR, AWS S3, Redshift, Glue,**  **Lambda, AWS SNS, BiqQuery, GCS Bucket, G- cloud function, Data flow, pub/sub cloud shell** |
| **Databases** | **MySQL, SQL Server, Oracle, MS Access, Teradata, and Snowflake** |
| **NoSQL Data Bases** | **MongoDB, DynamoDB, Cassandra DB, HBase** |
| **Monitoring tool** | **Apache Airflow** |
| **Visualization & ETL tools** | **Tableau, Power BI, Informatica, Talend, SSIS, and SSRS** |
| **Version Control & Containerization tools** | **GitHub, Bitbucket, Docker, Kubernetes** |
| **Operating Systems** | **Unix, Linux, Windows, Mac OS** |

**PROFESSIONAL EXPERIENCE:**

**Client: Humana, KY**

**Role: Sr Data Engineer**

**Duration: August 2022 to Till Date**

**Roles & Responsibilities:**

* Worked in complete Software Development Life Cycle (SDLC) process by analyzing business requirements and understanding the functional workflow of information from source systems to destination systems.
* Utilizing analytical, statistical and programming skills to collect, analyze and interpret large data sets to develop data-driven and technical solutions to difficult business problems using tools such as **SQL**, and **Python**.
* Worked on designing **AWS EC2** instance architecture to meet high availability application architecture and security parameters.
* Created **AWS S3 buckets** and also managed policies for S3 buckets and Utilized S3 buckets and Glacier for storage and backup.
* Worked on creating **Azure Data Factory(ADF)** and managing policies for Data Factory and Utilized **Blob storage** for storage and backup on Azure.
* Worked on **Hadoop** cluster and data querying tools to store and retrieve data from the stored databases.
* Worked with different file formats like Parquet files and also Impala using **PySpark** for accessing the data and performed **Spark Streaming** with RDDs and Data Frames.
* Hands on experience on Google Cloud Platform (GCP) in all the bigdata products **BigQuery**, **Cloud Data Proc**, **Google Storage**, **Composer** (Air Flow as a service).
* Performed the aggregation of log data from different servers and used them in downstream systems for analytics using **Apache Kafka**.
* Worked on designing and developing the **SSIS** Packages to import and export data from MS Excel, SQL Server, and Flat files.
* Worked on Data Integration for extracting, transforming, and loading processes for the designed packages.
* Designed and deployed automated **ETL workflows** using **AWS lambda**, organized and cleansed the data in S3 buckets using **AWS Glue**, and processed the data using **Amazon Redshift**.
* Project experience on Terraform and Airflow.
* Experience in using stack driver service/ dataproc clusters in GCP for accessing logs for debugging.
* Worked within the **ETL architecture** enhancements to increase the performance using query optimizer.
* Implemented the data that is extracted using **Spark**, **Hive**, and large data sets using **HDFS.**
* Worked on Streaming data transfer, data from different data sources into **HDFS**.
* Created ETL Mapping with **Talend** Integration Suite to pull data from Source, apply transformations, andload data into the target database.
* Worked with **Azure Blob Storage** and developed the framework for the implementation of the huge volume of data and the system files.
* Worked on scripting with **Python** in **Spark** for transforming the data from various files like Text files, CSV and JSON.
* Loaded the data from different relational databases like **MySQL** and **Teradata** using **Sqoop** jobs.
* Worked on processing the data and testing using **Spark SQL** and on real-time processing by Spark Streaming and **Kafka** using **Python**.
* Scripted using **Python** and **PowerShell** for setting up baselines, branching, merging, and automation processes across the process using **GIT**.
* Worked with the implementation of the ETL architecture for enhancing the data and optimized workflows by building DAGs in Apache Airflow to schedule the ETL jobs and additional components in Apache Airflow like Pool, Executors, and multi-node functionality.
* Used various Transformations in SSIS Dataflow, Control Flow using for loop Containers and Fuzzy.
* Worked on creating SSIS packages for Data Conversion using data conversion transformation and producing the advanced extensible reports using SQL Server Reporting Services.
* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Deploy all the AWS components with Terraform (both CLI and using TFE)
* Experience in building and architecting multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in GCP and coordinate task among the team.
* Implemented a distributed stream processing platform with low latency and seamless integration, with data and analytics services inside and outside **Azure** to build your complete big data pipeline.
* Developed and deployed the outcome using **Spark** and **Scala** code in Hadoop cluster running on GCP.
* Experience in working with product teams to create various store level metrics and supporting data pipelines written in GCP big data stack.
* Implemented continuous integration & deployment (CI/CD) through Jenkins for automating terraform scripts for Hadoop jobs and Managed Hadoop clusters using Cloudera.
* Used Cloud shell SDK in GCP to configure the services **Data Proc**, **Storage**, **BigQuery**.
* Demonstrated expertise in SharePoint for collaborative document management, including creating and managing document libraries, lists, and permissions.
* Developed custom SharePoint workflows to automate business processes, increasing efficiency and productivity.
* Extensive experience in creating complex Excel formulas and functions to perform data analysis, calculations, and reporting, including VLOOKUP, HLOOKUP, SUMIF, COUNTIF, and more.
* Proficient in writing VBA macros to automate repetitive tasks, enhance data processing, and create custom Excel solutions, improving data accuracy and efficiency.
* Integrated data from SharePoint into Excel workbooks, allowing for real-time data updates and collaborative data analysis.
* Designed automated processes to extract data from SharePoint lists and libraries into Excel for reporting and analysis.
* Developed and optimized Oracle PL/SQL procedures and functions for data extraction, transformation, and loading (ETL) processes, ensuring efficient data processing in the Oracle database.
* Implemented complex queries and subqueries in PL/SQL to retrieve and manipulate data from Oracle databases, contributing to the creation of comprehensive data solution.

**Environment:** Python, SQL, AWS EC2, AWS S3 buckets, Hadoop, PySpark, AWS lambda, AWS Glue, Amazon Redshift, Spark Streaming, Apache Kafka, SSIS, Informatica, ETL, Hive, HDFS, Talend, MySQL, Teradata, Sqoop, PowerShell, Terraform, GIT, Apache Airflow, GCP Dataproc, GCP Cloud functions, BigQuery, Power BI, Composer, Spark, Scala

**Client: Capital One, VA**

**Role: Senior Data Engineer**

**Duration: Nov 2021 to July 2022**

**Roles & Responsibilities:**

* Worked with business/user groups for gathering the requirements and working on the creation and development of pipelines.
* Migrated applications from **Cassandra DB** to **Azure Data Lake Storage Gen 2** using **Azure Data Factory(ADF)**, created tables, and loading and analyzed data in the Azure cloud.
* Expertise with **AWS databases** such as **RDS** (Aurora), **Redshift**, **DynamoDB**, and **Elastic Cache (Memcached & Redis)**
* Worked on developing the process and ingested the data in Azure cloud from web service and load it to **Azure SQL DB**.
* Worked with **Spark** applications in **Python** for developing the distributed environment to load high volume files using **PySpark** with different schema into PySpark Data frames and process them to reload into **Azure SQL DB** tables.
* Designed and developed the pipelines using **Databricks** and automated the pipelines for the **ETL** processes and further maintenance of the workloads in the process.
* Experience in identifying production bugs in the data using stack driver logs in GCP.
* Managed security groups on AWS, focusing on high-availability, fault-tolerance, and auto scaling using Terraform templates. Along with Continuous Integration and Continuous Deployment with AWS Lambda and AWS code pipeline.
* Worked on creating ETL packages using **SSIS** to extract data from various data sources like Access database, Excel spreadsheet, and flat files, and maintain the data using **SQL Server**.
* Developed **AWS Athena** extensively to ingest structured data from **S3** into various systems such as **RedShift** or to **generate reports**.
* Worked with **ETL operations** in **Azure Databricks** by connecting to different relational databases using **Kafka** and used **Informatica** for creating, executing, and monitoring sessions and workflows.
* Proficient in **SQLite**, **MySQL** and **SQL** **databases** with **Python**.
* Worked on automating data ingestion into the Lakehouse and transformed the data, used **Apache Spark** for leveraging the data, and stored the data in **Delta Lake**.
* Ensured data quality and integrity of the data using **Azure SQL Database** and automated ETL deployment and operationalization.
* Used Databricks, Scala, and Spark for creating the data workflows and capturing the data from Delta tables in **Delta Lakes**.
* Performed Streaming of pipelines using **Azure Event Hubs** and **Stream Analytics** to analyze the data from the data-driven workflows.
* Worked with **Delta Lakes** for consistent unification of Streaming, processed the data, and worked on ACID transactions using **Apache Spark**.
* Created **S3 buckets** and managed **S3 bucket** **policies**, as well as using **S3 buckets** and **Glacier** for storage and backup on **AWS**.
* Experience in GCP Dataproc, GCP Cloud functions, BigQuery.
* Worked with **PowerShell** scripting for maintaining and configuring the data. Automated and validated the data using **Apache Airflow**.
* Worked on optimization of **Hive** queries using best practices and right parameters and using **Hadoop, YARN, Python,** and **PySpark**.
* Used Sqoop to extract the data from Teradata into HDFS and export the patterns analyzed back to Teradata.
* Expert in building containerized apps using tools like Docker, Kubernetes and Terraform.
* Used Accumulators and Broadcast variables to tune the **Spark** applications and to monitor the created analytics and jobs.
* Tracked **Hadoop** cluster job performance and capacity planning and tuning Hadoop performance for high availability and Hadoop cluster recovery.
* Worked with **Tableau** for generating reports and created Tableau dashboards, pie charts, and heat maps according to the business requirements.
* Worked with all phases of **Software Development Life Cycle** and used **Agile methodology** for development.
* Designed and maintained SharePoint sites for team collaboration and information sharing, ensuring seamless data integration with other systems.
* Utilized Power Pivot to build data models, perform data modeling, and create interactive dashboards, enabling in-depth data analysis and visualization.
* Created and managed Pivot Tables to summarize and analyze large datasets, facilitating quick insights and data-driven decision-making.
* Utilized Oracle Database features such as materialized views and partitioning for performance optimization and data management in large-scale systems.
* Collaborated with the database team to enhance data integrity and implemented constraints and triggers using PL/SQL for data validation and consistency.

**Environment:** Python, SQL, Cassandra DB, Azure Data Lake Storage Gen 2, Azure Data Factory, Azure SQL DB, Spark, Databricks, SSIS, SQL Server, Kafka, Informatica, Apache Spark, Delta Lake, Azure Event Hubs, Stream Analytics, Terraform , Azure Blob Storage, PowerShell, Apache Airflow, Hadoop, YARN, PySpark, Hive, Teradata, Sqoop, HDFS, Spark, Agile, Oracle PL/SQL

**Client: Burger King, Miami**

**Role: Data Engineer**

**Duration: Jan 2020 to Oct 2021**

**Roles & Responsibilities:**

* Involved in various phases of Software Development Life Cycle (**SDLC**) as requirement **gathering, data modeling, analysis, design** & **development** for the project.
* Created the infrastructure needed for optimal data **extraction**, **transformation**, and **loading** from a wide range of data sources.
* Created optimal **data** **pipeline** architecture. In the **Hadoop** environment with **Linux** for big data resources, developed **Spark/Scala**, **Python** for regular expression (regex) project.
* Data sources are **extracted**, **transformed,** and **loaded** to generate **CSV** data files with **Python** programming and **SQL** queries.
* Involved in developing **Pig** **Scripts** for change **data capture** and **delta** **record** **processing** between newly arrived data and already existing data in **HDFS**.
* Experience in **GCP Dataproc**, **GCS**, **Cloud functions, Data prep, Data Studio** and **Big Query**.
* Designed and developed **Security Framework** to provide fine grained access to objects in **AWS S3** using **AWS Lambda**, **DynamoDB**.
* Responsible for loading data from the internal server and the **Snowflake** data warehouse into **S3 buckets.**
* Experience in **building** and **architecting** multiple Data pipelines, end to end **ETL** and **ELT** process for **Data ingestion** and **transformation** in **AWS** and coordinate task among the team.
* Implementing and maintaining (CI/CD) pipelines to automate the deployment of data pipeline and data warehouse changes using Azure DevOps.
* Created and written aggregation logic on **Snowflake** **Datawarehouse** tables.
* Recreated and maintained existing Access **Database** artifacts in **Snowflake**.
* Consumed **Kafka** messages and curated using **Python** send the data into multiple targets **Redshift, Athena** and **S3 buckets**.
* Used **AWS** **Quick sight** for visualization.
* Used Python libraries like **NumPy**, **Pandas**, **Scify**, **matplotlib**, **seaborn** and **sklearn**.
* Used Data science models **Linear regression**, **logistic Regression**, **Kneighbors Classifier**, Random Forest Classifier, Dummy Classifier, **ARIMA, SARIMA** to predict and used the data insights to make business decisions.
* Worked with Data Science, Marketing and Sales team to develop the Data Pipelines as per there need.
* Designed, built, and maintained data integration programs in a Hadoop and RDBMS environment, worked with both traditional and non-traditional source systems, as well as **RDBMS** and **NoSQL** data stores for data access and analysis.
* Used the **Spark** **API** to analyze Hive data in conjunction with the **EMR** Cluster Hadoop Yarn.
* **AWS Cloud Formation** templates were designed to create **VPCs, subnets** and **NAT** to ensure the successful deployment of Web applications and **database templates.**
* Experience in building **Power BI** reports by connecting to GCP Data sources.
* Migrated Hive and MapReduce jobs from on-premises MapR to **AWS** cloud using **EMR** and **Quble.**
* Worked with **Spark** applications in **Python** for developing the distributed environment to load high volume files using PySpark with different schema into **PySpark** Data frames and process them to reload into **SQL DB tables**.
* Installed **Kafka** on Hadoop cluster and configured **producer** and **consumer** in java to establish connection from source to **HDFS** with popular hash tags.
* Utilized Azure Monitor, Azure Log Analytics, and Application Insights to monitor and diagnose issues, analyze logs and metrics, and improve the overall health and performance of Azure environments and applications.
* Worked on creating **ETL** packages using **SSIS** to extract data from various data sources like Access **database**, **Excel** **spreadsheet**, and **flat files**, and maintain the data using **SQL Server**.
* Worked on automating data **ingestion** into the **Lakehouse** and transformed the data, used **Apache Spark** for leveraging the data, and stored the data in **Delta Lake**.
* Ensured **data quality** and integrity of the data using **SQL Database** and automated **ETL** deployment and operationalization.
* Analyze Logs for **debugging** & **performance** tuning.
* Writing **Maven** build **script**, **Jenkins** & **Chef scripts**.
* Led data migration projects, transferring data between SharePoint and Excel, ensuring data integrity, and maintaining data relationships during the migration.
* Worked on data migration projects, employing PL/SQL scripts to transfer data between Oracle databases and ensuring data accuracy and completeness during the migration process.
* Integrated **Azure Data Factory(ADF)** with other Azure services, such as Azure Blob Storage and Azure SQL Database, to create comprehensive data solutions in the Azure cloud environment

**Environment**: Python, SQL, Hadoop, Pig Scripts, HDFS, AWS S3, Lambda, Dynamo DB, Snowflake, Redshift, Athena, Kafka, Quick sight, EMR, RDS, Elastic Cache, Jenkins**,** Oracle PL/SQL,ADF

**Client: Avon Technologies, India**

**Role: Data Engineer**

**Duration: Apr 2017 to Dec 2019**

**Roles & Responsibilities:**

* Experience in building and architecting multiple data pipelines end to end ETL and ELT for data ingestion and transformation in GCP and coordinate task among them.
* Implemented and Managed **ET** solutions and automating operational processes
* Design and develop ET integration patterns using **Python** on **Spark.**
* Develop framework for converting existing PowerCenter mappings and to **PySpark (**Python and Spark) Jobs.
* Build data pipelines in airflow in **GCP** for ET related jobs using different airflow operators.
* Utilized Azure Monitor, Azure Log Analytics, and Application Insights to monitor and diagnose issues, analyze logs and metrics, and improve the overall health and performance of Azure environments and applications.
* Loaded the Data to AWS using Databricks clusters and process it as per the business requirements
* Used **Stitch ETL** tools to integrate data into the central data warehouse.
* Experience in using G-cloud function with python to load data into BiqQuery for on arrival CSV files to GCS Bucket.
* Experience in loading bound and unbound data from Google subtopic to Big Query using cloud data flow with python.
* Used Rest API with python to ingest data from and other sites to BiqQuery.
* Implemented **Spark RDD** transformations to map business analysis and apply actions on top of Transformations.
* Design **star schema** in Big Query.
* Used Azure data platform and tools,as well as enterprice ETL tool Talend,Leveraging its DQ, DI and Data Catalogue features.
* Developing and maintaining the data warehouse architecture, including data modeling, schema design, and query optimization, using technologies such as Azure Synapse Analytics.
* Launched multi-node kubernetes cluster in Google Kubernetes Engine (GKE) and migrated the dockerized application from AWS to GCP.
* Worked on creating various types of indexes on different collections to get good performance in **Mongo** database.
* Monitoring Big query, **Dataproc** and **cloud Data** flow jobs via Stack driver for all the environments.
* Used **Agile** for the continuous model deployment.
* Developed API for using a Lambda to manage the servers and run the code in the AWS.
* Worked with **Google data catalog** and other google cloud APIs for monitoring, query and billing related analysis for big query usage.
* Knowledge about **cloud data flow** and Apache beam.
* Used **Snowflake** for the Data Storage, processing which is easier and faster to use.
* Write a Python program to maintain raw file archival in GCS bucket.
* Worked with google data catalog and other google cloud APl's for monitoring, query, and billing related analysis for **Big Query** usage.
* Used Airflow to manage task scheduling, progress, and success status using **DAG** graphs.
* Created **Big Query** authorized views for row level security or exposing the data to other teams.
* Integrated services like **GitHub**, **Jenkins** to create a deployment pipeline.
* Implemented new projects builds framework using Jenkins as build framework tools.
* Proficient in data transformation and cleansing to ensure data accuracy during migration processes.
* Implemented parameterization and dynamic configurations in **Azure Data Factory(ADF**) pipelines to enhance reusability and flexibility in handling different data scenarios

**Environment:** T-SQL, PL/SQL, Google Cloud, Python, Big query, Dataflow, Dataproc, Dataprep, Data Studio, BigtableStitch ETL, PySpark, Snowflake, MySQL, Airflow, Shell Scripts, Mongo DB, GIT, Apache, Spark, Docker, Oracle PL/SQL**,**ADF

**Client: Netxcell, India**

**Role: Data Analyst / Data Engineer**

**Duration: June 2014 to March 2017**

**Roles & Responsibilities:**

* Installed, configured, and maintained **Apache Hadoop** clusters for the development of applications in accordance with the specifications.
* Create ETL data pipelines by combining technologies such as **Hive**, **Spark SQL** and **PySpark**.
* Created Spark programs using **Scala** and **Batch** processing using functional programming techniques.
* Added data to Power **BI** from a range of sources, including **SQL**, **Excel**, **Oracle**.
* Writing **Spark** Core Programs to process and clean data before loading it into **Hive** or **HBase** to be processed further.
* Utilization of tools for data transformation such as **Data Stage**, **SSIS**, **Informatica**, or **DTS**.
* Proficient in using UML for **Use Cases, Activity Diagrams, Sequence Diagrams, Data Flow Diagrams, Collaboration Diagrams** and **Class**.
* In charge of building **ETL** pipelines with **Pig** and **Hive** to extract data from various data sources and import it into the **Hadoop Data Lake**.
* Worked with several data types, including **JSON** and **XML**, and ran **Python** machine learning algorithms.
* Created reusable items, such as **PL/SQL** program units and libraries, database functions and procedures, and database triggers that the team could utilize to meet business rules.
* Used **SQL Server Integrations Services (SSIS)** to extract, manipulate, and load data from a variety of sources into the target system.
* Created **data mapping**, **transformation**, and **cleaning** **rules** for **OLTP** and **OLAP** data management.
* Used Tableau for the data visualization during the quick model construction process in **Python**. These models are then put into practice in **SAS**, where they are connected to **MSSQL** databases and have timely update schedules.
* Created numerous data frames and datasets using the **Spark-SQL** context to pre-process the model data.
* Worked on designing the **HBase** row key to store Text and **JSON** as key values in the database and to get/scan it in sorted order.
* In charge of **ETL** design (identifying the source systems, designing source to target relationships, data cleansing, data quality, creating source specifications, **and ETL** design documents).
* Strong Knowledge/experience in creating **Jenkins** CI pipelines.
* Implemented querying using **Airflow, presto** as well as reporting in **PySpark, Zeppelin** and **Jupyter**.
* Installed and set up Airflow for managing workflows and built workflows in Python.

**Environment: Hive, Spark SQL, PySpark, Oracle, Hive, HBase, Data Stage, SSIS, Informatica, Pig, Jenkins, Airflow, Presto, Zeppelin, Jupyter.**