**Sowmya**

**Data Engineer**

**732-807-0276**

**Professional Summary:**

* Over 8+ years of experience in Data Engineering, Data Pipeline Design, Development, and Implementation as a Sr. Data Engineer/Data Developer and Data Modeler.
* Have Extensive Experience in IT data analytics projects, Hands on experience in migrating on premise ETLs to Google Cloud Platform (GCP) using cloud native tools such as BIG query, Cloud Data Proc, Google Cloud Storage, Composer.
* Strong experience in Software Development Life Cycle (SDLC) including Requirements Analysis, Design Specification, and Testing as per Cycle in both Waterfall and Agile methodologies.
* Strong experience in writing scripts using Python API, PySpark API, and Spark API for analysing the data.
* Data Modeler with a focus on **L3 Production Support**.
* Experience in using ETL methodologies for supporting Data Extraction, Data Migration, Data Transformation and loading using Informatica Power center 9.6.1/ 9.1/8. 6.1/ 7. x/6.2, IDQ and Trillium.
* EDI Systems/Data Analyst, experienced working in fast paced environments demanding strong organizational, technical and interpersonal skills.
* Strategically design and implement BI Solutions on the Cloud platform as part of the digital transformation project for Condential separating from their previous owner-HMSHost & hosting their data independently. That includes integration with OLTP & ERP databases (NetSuite Financials, Crunchtime RMS, Oracle HCM) & legacy data warehouses
* Very keen in knowing newer techno stack that Google Cloud platform (GCP) adds.
* Experienced in building Automation Regressing Scripts for validation of ETL processes between multiple databases like Oracle, SQL Server, Hive, and Mongo DB using Python.
* Proficiency in SQL across several dialects (we commonly write MySQL, PostgreSQL, Redshift, SQL Server, and Oracle).
* Experienced in Worked on NoSQL databases - Hbase, Cassandra & MongoDB, database performance tuning & data modeling.
* Experience in Informatica Data Quality ( IDQ - Informatica developer 9.6.1/ 9.1) for cleansing and formatting customer master data.
* Experienced in big data analysis and developing data models using Hive, PIG, and Map reduce, SQL with strong data architecting skills designing data-centric solutions.
* Experience working with data modelling tools like Erwin and ER/Studio.
* Experience in implementing Azure data solutions, provisioning storage accounts, Azure Data Factory, SQL server, SQL Databases, SQL Data warehouse, Azure Data Bricks, and Azure Cosmos DB.
* Experience with Data Build Tool (Dbt)to performSchema Tests, Referential Integrity Tests, and Custom Tests the data and ensured Data Quality.
* Can work parallelly in both GCP and Azure Clouds coherently.
* Expertise in Amazon Web Services (AWS) Cloud Platform which includes services like EC2, S3, VPC, ELB, IAM, DynamoDB, Cloud Front, Cloud Watch, Route 53, Elastic Beanstalk (EBS), Auto Scaling, Security Groups, EC2 Container Service (ECS), Code Commit, Code Pipeline, Code Build, Code Deploy, Dynamo DB, Auto Scaling, Security Groups, Redshift, CloudWatch, CloudFormation, CloudTrail, Ops Works, Kinesis, IAM, SQS, SNS, SES.
* Expertise in NOSQL databases like HBase, MongoDB. Expertise in NOSQL databases like HBase, MongoDB.
* Good knowledge of Data Marts, OLAP, and Dimensional Data Modelling with Ralph Kimball Methodology (Star Schema Modelling, Snow-Flake Modelling for FACT and Dimensions Tables) using Analysis Services.
* Working knowledge of Data Build Tool (Dbt) with Snowake.
* Excellent in performing data transfer activities between SAS and various databases and data file formats like XLS, CSV, etc.
* Expertise in Python and Scala, user-defined functions (UDF) for Hive and Pig using Python.
* Experienced in development and support knowledge of Oracle, SQL, PL/SQL, and T-SQL queries.
* Hands-on experience in Architecting Legacy Data Migration projects to AWS Redshift migration and from on-premises to AWS Cloud.
* Strong development skills with Azure Data Lake, Azure Data Factory, SQL Data Warehouse Azure Blob, Azure Storage Explorer.
* Very keen in knowing newer techno stack that Google Cloud platform (GCP) adds
* Expertise in writing Hadoop Jobs to analyse data using MapReduce, Hive, Kafka, and Splunk.
* Worked on migrating MapReduce programs into Spark transformations using Spark and Scala, initially done using python (PySpark).
* Providing guidance to the development team working on PySpark as ETL platform.
* Used PySpark jobs to run on Kubernetes Cluster for faster data processing

**Technical Skills:**

* Big Data Technologies: Hadoop, Map Reduce, Nifi, HBase, Hive, Pig, Sqoop, Kafka, Oozie.
* Methodologies: RAD, JAD, System Development Life Cycle (SDLC), Agile, Waterfall
* Cloud Platforms: AWS (EC2, S3, Redshift, Glue), Azure (Azure Data Factory, Azure Databricks), GCP
* ETL Tools: SSIS, Informatica PowerCenter, Apache NiFi, Talend
* Programming Languages: Python, SQL, PL/SQL, Scala, Java, Unix
* Libraries: Pandas, NumPy, Matplotlib, SciPy, Scrapy, TensorFlow, PyTorch, Scikit-learn, NLTK, Plotly, Keras
* Integrated RAG models with LLMs (e.g., OpenAI GPT, LLaMA, Mistral) using frameworks like LangChain or LLamaIndex.
* ETL Tools: Snowake, Data Build Tool (Dbt), Informatica
* Cleansing Tools: Informatica Data Quality ( IDQ 9.6.1/9.5 ), Trillium
* Streaming Technologies: Amazon Kinesis, Apache Spark, Apache Kafka
* Data Visualization: Microsoft Excel, Power BI, Tableau, IBM Cognos, QlikView, QuickSight, Seaborn
* Databases: SQL Server, PostgreSQL, MySQL, Oracle, Snowflake, DynamoDB, MongoDB
* Data Warehousing: Amazon (DynamoDB, RDS, Athena), Azure (Synapse, Blob, Data Lake), BigQuery, Teradata, Snowflake
* DevOps Tools: Git, Jenkins, Docker, Kubernetes
* Pipelines: Apache Airflow, AWS Step Function, Luigi, Oozie

**Experience**

**Vanguard | PA(Remote) |*Senior Data Engineer* December 2024 - Present**

* Designed AWS architecture, Cloud migration, AWS EMR, DynamoDB, Redshift and event processing using lambda function.
* Implemented usage of Amazon EMR for processing Big Data across a Hadoop Cluster of virtual servers on Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3).
* Utilized AWS services with focus on big data analytics, enterprise data warehouse and business intelligence solutions to ensure optimal architecture, scalability, flexibility.
* GCP Build data pipelines in airow in GCP for ETL related jobs using dierent airow operators.
* Experienced in Importing and exporting data into HDFS and Hive using Sqoop.
* Participated in development/implementation of Cloudera Hadoop environment.
* Experienced in running query-using Impala and used BI tools to run ad-hoc queries directly on Hadoop.
* Built scalable **ETL/ELT data pipelines** using Spark, Airflow, and Python, integrated with Privacera policy enforcement points (PEPs).
* Provided **L3 Production Support** for **ETL jobs** using **Talend**, **Sqoop**, and **Hive**, ensuring timely identification and resolution of performance-related issues in the production environment
* Facilitated EDI testing with trading partners to implement new initiatives, transaction sets, version upgrades and alterations to current transaction sets
* Nvolved in comparison testing as part of the data migration from the legacy applications into Oracle RMS environment
* Tuned RAG model parameters (e.g., top\_k, chunk overlap, retrieval temperature) to optimize output.
* Revised our Data Architecture standardization in per the new platform the Data Lake.
* Experience in GCP Dataproc, GCS, Cloud functions, BigQuery.
* Ensured the correct translation was generated by Condential, and that the les were accepted correctly into the ERP application .
* Implemented **Privacera** to enforce centralized data access policies across cloud data platforms (AWS, Azure, GCP, Databricks)
* Implemented EDI maps for transactions 820, 834, 837, 835, 270 and generated functional acknowledgment 999 using the X12 standard, in versions ranging from 4010 through 5010
* Using Bash and Python included Boto3 to supplement automation provided by Ansible and Terraform for tasks such as encrypting EBS volumes backing AMIs.
* Imported mapplets and mappings from Informatica developer (IDQ) to Power Center.
* Involved in using Terraform migrate legacy and monolithic systems to Amazon Web Services.
* Wrote Lambda function code and set Cloud Watch Event as trigger with Cron job Expression.
* Created an image classication model that compares actual and expected solar plant performance and identies the driver of any deviation in those two metrics. While it could be handled as a time-series problem, the values are processed into a graph and identied using computer vision AI that is more than 93% accurate. The project is expected to save the business over $500,000 annually.
* Experienced in onboarding new datasets into Privacera and defining access zones, **resource sets**, and **data classifications**.
* Monitored and evaluated RAG pipeline performance with custom metrics for retrieval precision, recall, and generation quality.
* Provided **L3 Production Support** for AWS Redshift and **Azure Data Lake** environments, ensuring **high system uptime** and effective troubleshooting of data pipeline issues
* Developed EDI maps using the Windows Condential tool
* Worked on creating and running Docker images with multiple micro – services and Docker container orchestration using ECS, ALB and lambda. Used AWS services EC2 and S3 for small data sets processing and storage.
* Created SQL tables with referential integrity, constraints and developed queries using SQL, SQL\*PLUS and PL/SQL.
* Architected and led a team to Intergrade Investment data to Treasury Data Lake.
* Completed Condential EDI Condential Integrator training to dene data mapping rule, setting up FTP connections and trading partner proles to align all the X12 transaction sets.
* Curate data sourced out of Lake in to Databricks in different phases environments and perform Delta strong data engineering experience in Spark and Azure Databricks, running notebooks using ADF
* Developed and implemented Historical and Incremental Loads using Databricks & Delta Lake run using ADF pipelines.
* EDI Systems/Data Analyst, experienced working in fast paced environments demanding strong organizational, technical and interpersonal skills.
* Redesigned Star schema for Short Term funding, long Term funding, derivatives, Investment, Third Party Debt, Collateral data per Data lake standards.
* Created sanity checks and alert for monitoring data quality
* Migration large data sets to Databricks (Spark), create and administer cluster, load data, configure data pipelines, loading data from ADLS Gen2 to Databricks using ADF pipelines
* Extensive hands-on experience of writing notebooks in data bricks using python/Spark SQL for complex data aggregations, transformations, schema operations. Good familiarity with Databricks delta and data frames concepts
* Developed Spark scripts by writing custom RDDs in Scala for data transformations and perform actions on RDDs.
* Experienced in writing Spark Applications in Scala and Python (PySpark).
* Imported Avro files using Apache Kafka and did some analytics using Sparking Scala.
* Created Pyspark frame to bring data from RDBMS to Amazon S3.
* 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.

**EY | TX | *Data Engineer* November 2022 - November 2024**

* Involved in gathering business requirements, logical modeling, physical database design, data sourcing and data transformation, data loading, SQL, and performance tuning.
* Used SSIS to populate data from various data sources, creating packages for different data loading operations for applications.
* Built ETL solutions using Databricks by executing code in Notebooks against data in Data Lake and Delta Lake and loading data into Azure DW following the bronze, silver, and gold layer architecture.
* Experience in developing Spark applications using Spark-SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats.
* Extracted data from various sources like SQL Server 2016, CSV, Microsoft Excel and Text file from Client servers.
* Developed and executed a migration strategy to move Data Warehouse from an Oracle platform to AWS Redshift.
* Built S3 buckets and managed policies for S3 buckets and used S3 bucket and Glacier for storage and backup on AWS.
* Architected and designed a future state solution for treasury data warehouse in Data Lake
* Leverage Image, text and numeric data within Condential using AI tools to deliver results
* Developed Spark scripts using Python on AWS EMR for Data Aggregation, Validation and Adhoc querying.
* Worked with **cloud-native data lakes** and **data warehouses**, ensuring secure data democratization via Privacera.
* Successfully migrated many Treasury Data domain from Oracle platform to Data Lake
* Performed data analytics on DataLake using pyspark on databricks platform
* Involved in creation/review of functional requirement specifications and supporting documents for business systems, experience in database design process and data modeling process.
* Designed and documented the entire Architecture of Power BI POC.
* Implementation and delivery of MSBI platform solutions to develop and deploy ETL, analytical, reporting and scorecard / dashboards on SQL Server using SSIS, SSRS.
* Monitor the activities of data quality, data proling and Meta data
* Extensively worked with SSIS tool suite, designed and created mapping using various SSIS transformations like OLEDB command, Conditional Split, Lookup, Aggregator, Multicast and Derived Column.
* Load data into Amazon Redshift and use AWS Cloud Watch to collect and monitor AWS RDS instances within Confidential.
* Integrated Privacera with **Apache Ranger**, **Active Directory/LDAP**, and **SSO** for enterprise-wide identity and access management.
* Worked extensively on SQL, PL/SQL, and UNIX shell scripting.
* Expertise in creating PL/ SQL Procedures, Functions, Triggers and cursors.
* Loading data in No SQL database (HBase, Cassandra)
* Used Teradata utilities fastload, multiload, tpump to load data.
* Utilized Power Query in Power BI to Pivot and Un-pivot the data model for data cleansing.
* Environment: MS SQL Server 2016, ETL, SSIS, SSRS, SSMS, Cassandra, AWS Redshift, AWS S3, Oracle 12c, Oracle Enterprise Linux, Teradata, Databricks, Jenkins, PowerBI,
* Developed the batch scripts to fetch the data from AWS S3 storage and do required transformations in Scala using Spark framework.
* Implemented Spark using Scala and Spark SQL for faster testing and processing of data.

**TCS (Pfizer) | Hyd, India| Data Engineer August 2020 - December 2021**

* Performed Data Analysis, Data Migration, Data Cleansing, Transformation, Integration, Data Import, and Data Export.
* Worked on the Design, Development, and Documentation of the ETL strategy to populate the data from the various source systems using the Talend ETL tool into the Data Warehouse.
* Devised PL/SQL Stored Procedures, Functions, Triggers, Views and packages. Made use of Indexing, Aggregation, and Materialized views to optimize query performance.
* Developed logistic regression models (using R programming and Python) to predict subscription response rate based on customer variables like past transactions, response to prior mailings, promotions, demographics, interests and hobbies, etc.
* Created Tableau dashboards/reports for data visualization, Reporting, and Analysis and presented them to Business.
* Worked to land data from heterogeneous sources (RMS, QuickBooks, and Tempest, Spreadsheets etc.) to given destination with appropriate transformation/Business rules applied.
* Transferred the result from Hive to Oracle using Sqoop thereby allowing the downstream to use the consistent data.
* Transferred data from DataBases to HDFS using Sqoop.
* Used Flume to stream through the log data from various sources.
* Stored the data in the tabular formats using the Hive tables.
* Worked with senior management to plan, define, and clarify dashboard goals, objectives, and requirements.
* Responsible for daily communications to management and internal organizations regarding the status of all assigned projects and tasks.
* Analyzed data using Hadoop components Hive and Pig.
* Involved in running Hadoop streaming jobs to process terabytes of data.
* Gained experience in managing and reviewing Hadoop Clusters.

**Reliant Vision | Hyd, India| BI Analyst March 2016 - July 2020**

* Collaborated with stakeholders to comprehend company objectives, metrics, processes, and requirements.
* Created of comprehensive solution(s) to meet business requirements utilizing.
* To design/develop an efficient overall technological solution, identify and resolve conflicts across numerous use cases.
* Reviewed of new releases to determine their impact and potential for technological and business process enhancements.
* Managing and supporting Microsoft Oce 365 compliance controls such as DLP and RMS.
* Involved in requirements gathering, documentation, and solution design standards and best practices.
* Participated in the facilitation, definition, and recording of requirements, as well as the analysis of work/data flow, acceptance criteria, and test preparation.
* Estimated and assisted with application configuration and testing on a hands-on basis.
* To developed, built, and delivered a solution, collaborate with other technical and non-technical team members as well as vendors in a technology-neutral language.
* Understand the business requirements and create problem statement, gather data, analyzing and validate the data using Excel and creating repots and dashboards in Excel and Tableau