**Sandhyarani Timmapuram**

**Lead Data Engineer/Snowflake Developer**

**SUMMARY:**

Experienced Data Engineer with a decade of expertise in data applications, databases, and cloud technologies. Skilled in Agile methodologies, data analysis, and Data pipelines/ETL workflows. Proficient in SQL, Python, and visualization tools for predictive analytics and data cleaning.

**EXPERIENCE HIGHLIGHTS:**

* A skilled professional with 10 years of experience in Data Engineering, showcasing proficiency in AWS, Azure, GCP and Snowflake cloud technologies, with a focus on both data engineering and Business Intelligence (BI)
* Strong hands-on experience in Scrum/Agile and Waterfall project execution.
* Expertise in R, SAS, Python, and visualization tools like Tableau and Power BI.
* Extensive knowledge of NoSQL and SQL databases, including HBase, Cassandra, MongoDB, Neo4J, Dynamo DB, MYSQL, Teradata, and more.
* Skilled in CI/CD tools like Git, Bitbucket.
* Experienced in data warehousing, ETL, and Business Intelligence.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| Big Data Technologies | Hadoop, HDFS, Sqoop, Hive, Spark SQL, HBase, Apache Spark, Apache Kafka, |
| Data Modeling tools | Erwin Data Modeler, ER Studio v17 |
| Python Libraries | BeautifulSoup, , NumPy, SciPy, Pandas, Matplotlib, PyTables, Pyspark, OpenCV, Django, Flask, urllib, BeautifulSoup, Scrapy, Pytest, pymongo, sqlalchemy, pymysql |
| Databases | Microsoft SQL Server, Oracle, MYSQL, No SQL, Cassandra, PostgreSQL, MongoDB, Cosmos DB, Neo4J, Graph DB, Teradata, Big Data, DynamoDB |
| Version Controls | Git, GitHub, Bitbucket, SVN, CVS, SDLC Agile methodologies, SCRUM. |
| IDEs / Tools | PyCharm, Spyder, Jupyter Notebook, Visual Studio NetBeans, CMDER, Sublime Text, other tools like Putty. |
| ETL and Reporting tools | Tableau, Power BI, OBIEE, Looker, Matillion Alteryx, Sisence, |
| Cloud Environment | AWS (EC2, S3, Glue, Route53, Lambda, Snowflake, and Redshift) Kubernetes, |
| Deployment/Delivery Tools | Jenkins, Heroku, Ansible, Chef, Puppet  |
| Azure platform | Azure SQL, SQL Server, Azure Synapse, Azure Analytic service, Databricks, Azure Cosmos DB, Azure Event hub, App Services, |
| Programming Languages | Python, SAS, C, Java, J2EE |
| Domains | Banking, Retail, Financial, Insurance, Healthcare & Pharma |

**PROFESSIONAL EXPERIENCE:**

**Lead Data engineer | Snowflake Oct 2022 – Present
American Airlines Fort Worth, TX**

* Working on migrating data to the cloud Snowflake from the legacy data warehouses and developing the infrastructure.
* Loading data into snowflake tables from the internal stage using snowsql.
* Understanding and implementing a robust Data Governance framework within the organization, ensuring data is managed, governed, and used effectively.
* Integrating Druid with other components of the data ecosystem, such as Apache Kafka, Apache Hadoop, and data visualization tools.
* Integrated ThoughtSpot with various data sources and systems, ensuring seamless data flow for analysis.
* Experience in building multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in GCP and coordinate task among the team.
* Created detailed documentation for Grafana and Prometheus configurations, making it easier for team members to understand and maintain the setup. Conducted knowledge-sharing sessions to educate colleagues on Grafana and Prometheus best practices.
* Worked closely with DevOps teams to deploy and manage Flink clusters in production environments.Conducted performance tuning and optimization of Prometheus instances to ensure efficient metric collection and storage.
* Develop ELT data pipeline to migrate applications using DBT and Snowflake framework
* Utilized Kubernetes and Docker for the runtime environment for the CI/CD system.
* Performed data quality issue analysis using Snow SQL by building analytical warehouses on Snowflake.
* Created an automation process for Distribution group which receive Inventory and sales data send activation report using Talend and Redshift.
* Worked on migration of legacy Data Warehouse from Oracle/SQL Server/Netezza to central Data Lake Formation/Snowflake using Talend and Informatica.
* Experienced in test automation to ensure data quality, and proficient in setting up continuous integration and deployment (CI/CD) processes using tools like Docker, GitHub, Jenkins, and Unix/Linux shell scripts.
* Build data pipelines using Informatica IICS technologies and load into Snowflake warehouse from disparate source systems.
* Proficiently managed Databricks workspaces, including creating notebooks, folders, and organizing resources to streamline collaboration among different teams and projects
* Develop and schedule ETL workflows using Azure Data Factory (ADF) and Spark-SQL in Azure Databricks.
* Extensively used Databricks spark, Pyspark, Delta lake with python Synapse and Jupyter notebooks for Data Analytics.
* Performed data analysis, developed logical and physical data models, utilize Data Vault, suggested changes in design & architecture.
* Responsibility for implementing and maintaining Data Quality tools to ensure data accuracy, consistency, and reliability.
* Automate Datadog Dashboards with the stack through Terraform Scripts.
* Implemented deployment strategies for Flink applications, including containerization and orchestration (e.g., Kubernetes).Developed and designed data ingestion using Cassandra, Apache Nifi/Kafka.
* Building and maintaining data warehouses or data lakes to store historical transaction data is a common responsibility. This data serves as the foundation for AML analytics.
* Design and configure database, back-end applications and programs. Managed large datasets using Pandas data frames and SQL.
* Used Jenkins pipelines to drive micro services build Docker registry and then deployed to Kubernetes, Created pods and managed using Kubernetes.
* Performed ETL to move the data from source system to destination systems and worked on the Datawarehouse.
* Work with similar Microsoft on-prem data platforms, specifically SQL Server and related technologies such as SSIS, SSRS, and SSAS
* Build, optimize, and manage Spark clusters in Azure Databricks.
* Experience managing Azure Data Lakes (ADLS) and Data Lake Analytics and an understanding .Utilized Java-based tools and technologies, such as Apache Kafka and Apache Spark, for real-time data streaming and batch processing.
* Recreating existing application logic and functionality in the Azure Data Lake Formation, Neo4J (Graph DB), Cosmos DB, Data Factory, SQL Database and SQL Datawarehouse environment. experience in DWH/BI project implementation using Azure ADF
* Proven track record in optimizing ThoughtSpot queries and performance to ensure fast and efficient data retrieval. Develop ELT data pipeline to migrate applications using DBT and Snowflake framework
* Facilitated the transition from OBIEE to OAS, ensuring a seamless upgrade process.
* Extensive experience in setting up, configuring, and managing Prometheus monitoring solutions.
* Designed and implemented data pipelines using Falcon, automating data workflows and ensuring data reliability and consistency.
* Implement ETL and data movement solutions using Azure Data Factory, SSIS/SSRS.
* Utilized Falcon for data transformation tasks, including data cleansing, enrichment, and aggregation, resulting in improved data quality.
* Implementation of Data Vault Modelling and high dimensional data modeling techniques
* Designed Star and Snowflake Data Models for Enterprise Data Warehouse using ER Studio.
* Experienced in MLOps Concepts (Deploying, Monitoring, Maintenance) using Databricks, Google Cloud AI platform.
* Expertise in data cataloging, particularly with tools like Alation or Collibra, to enable effective data discovery and metadata management.
* Good knowledge in using apache Nifi to automate the data movement between two different Hadoop systems.
* Created the Materialized View to extract information from multiple data sources in the OBIEE administration.
* Work related to downloading BigQuery Databricks into pandas or Spark data frames for advanced ETL capabilities.
* As a Snowflake DBA I have managed all the user access to the platform and metadata changes to the platform.
* Implemented MLOps based approaches for training thereby capture metrics and for tracking models purpose and baselining the parameters.
* Skilled in using Big Data tools and technologies such as BigQuery, Databricks, Snowflake, Spark, and Kafka Streams for efficient data processing.
* Implemented airflow for workflow automation and scheduling tasks and created DAGs tasks.
* Designed and developed the ELT jobs using DBT to achieve the best performance.
* Worked on GCP POC to migrate data and applications from On-Prem to Google Cloud.
* Control access to the Oracle Analytics Server (OAS) application and ensure that operations are secure and efficient.
* Collaborated with Quality Engineers to ensure data quality standards.
* Wrote several SQL scripts to enhance the reporting flexibility and tuned the Cognos reports for better performance.
* Every project and initiative was completed with a high level of communication, collaboration and organizing of cross-functional project teams, using elements from OOD&D, DevOps/DataOps, SDLC and Agile approaches.

**Environment**: Python 3.6, Java/J2EE, .Net, C#, ETL, Docker, Snowflake DB, Snowflake DBA, GCP, GCS, MySQL, MongoDB, Control M, PySpark, Databricks, Apache Pyspark, Spark, Scala, Jenkins, Jira, Linux, NumPy, Pandas, Alteryx, PostgreSQL, Spark SQL, AWS, Glue, S3, Docker, Kubernetes, Redshift, Sagemaker, Nifi, DBT, Snowflake, Data Vault, OBIEE, OAS, Informatica IICS, Redshift, Oracle DB, TDM, Teradata, CosmosDB, DB2, PL/SQL, Cognos, SSIS, T-SQL.

**Sr. Data engineer | Snowflake April 2019 – Feb 2020**

**Optum Eden prairie, MN (USA project), India**

* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Experience in defining data stewardship roles and responsibilities and facilitating data stewardship activities within the organization.
* Working on enhancements which comes along the way in the sprints from scratch (GCS storage layer to DW Bigquery layer)
* Maintained proficiency in Oracle/PL-SQL and Informatica, ensuring data processing excellence.Adapted to reporting tools such as OBIEE and Power BI, delivering actionable insights.
* Utilized Redshift extensively to design and maintain data infrastructure, ensuring optimal performance.
* Experience in building Power BI reports on Azure Synapse Analysis services for better performance.
* Used cloud shell SDK in GCP to configure the services Data Proc, Storage, GCP BigQuery snapshots/copies of several ORACLE database. And worked on reverting, taking bookmarks, reinstating those sandbox databases.
* Configured Postgres SQL database parameters like connection settings, log settings and query optimization settings with the help of terraform.
* Created data mappings in Informatica ETL tool to extract data from different source files, transform the data using Filter, Lookup, Aggregator, Expression, Transformations and then loaded into data warehouse.
* Configured applications on Informatica cloud (IICS), lift & shift of PowerCenter repositories into cloud.
* Demonstrated expertise in Google Cloud Platform, including BigQuery, Compute Engine, and Google Cloud Storage (GCS).
* Managed user provisioning and access controls within the Databricks environment to enhance data security and compliance. Utilizing Spark's Structured Streaming API to simplify and streamline the development of real-time applications
* Built data workflows with ETL, SSIS, BI, DW, AWS EMR, Data Lake Formation, Redshift, Hadoop, Spark, Spark, SQL, Scala, and Python.
* Integrating Data Governance and Data Quality tools into existing data pipelines and workflows, ensuring seamless data management processes.
* Specify the importance of capturing and visualizing data lineage using the selected tools to track data flow and transformations.
* Integrated Apache Airflow to mouniter multi-stage ML floworks with tasks running on Amazon Sagemaker.
* Utilized bigdata tools for MLOps like GCP, Big Query, DataProc for streaming Data Lakes. AutoML for automating the model building process.
* Developed and maintained data pipelines using Java, ensuring efficient and reliable data extraction, transformation, and loading (ETL) processes.
* Developed and implemented Apache NIFI across various environments, written QA scripts in Python.
* Work related to downloading BigQuery Databricks into pandas or Spark data frames for advanced ETL capabilities
* Used Erwin Data Modeler (Erwin) for Data Modeling (data requirements analysis, database design etc.) of custom developed information systems, including databases of transactional systems and data marts.
* Developed Python-based API (RESTful Web Service) to track revenue and perform revenue analysis.
* Designed and Co-ordinated with Data Science team in implementing Advanced Analytical Models in Hadoop Cluster over large Datasets. Wrote scripts in Hive SQL and MongoDB database for creating complex tables with high performance metrics like partitioning, clustering and skewing
* Work related to downloading BigQuery Databricks into pandas or Spark data frames for advanced ETL capabilities.
* Worked closely with cross-functional teams to understand data requirements and ensure Falcon pipelines met business needs effectively.
* Proficient in configuring and managing Databricks clusters to ensure optimal performance, scalability, and resource utilization for various workloads
* Experience with MLOps/Azure CI/CD pipeline in Azure cloud process for collaboration and increasing the pace of model development and deployment via monitoring, validation, and governance of machine learning models
* Worked with google data catalog and other google cloud API's for monitoring, query and billing related analysis for BigQuery usage.
* Ability to troubleshoot issues related to Data Governance and Data Quality tools and implement solutions to ensure data integrity.
* Able to quickly visualize data and create dashboards using tools such as Power BI, Tableau, Alteryx.
* Working with CI/CD pipelines to integrate C# code and align with DevOps & DataOps practices using build and release pipelines.
* Responsible for estimating the cluster size, monitoring, and troubleshooting the Spark databricks cluster.
* 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.
* Worked on ETL tool Informatica, Oracle Database and PL/SQL, Python, and Shell Scripts.

**Environment:** Java, Python, Databricks, Apache Spark, Kafka, (NumPy, Pandas, Matplotlib), Snowflake DBA, Databricks, GCP, DBT, Nifi. ETL, Matillion, Control-M, Autosys, Tableau, Informatica, Azure Devops, DataOps, MSOffice, MSSQL Server, MSBI Stack, Linux, GIT Hub, Jira, AML.

**Sr. Data Engineer | Snowflake | Data Modeler Aug 2016 – April 2019**

**Abbie LabCorp Karnataka, India**

* Involved in various phases of Software Development Life Cycle (SDLC), including requirement gathering, modeling, analysis, architecture design, prototyping, developing and testing
* Documenting Spark Streaming applications, configurations, and best practices for the development team and stakeholders
* Used Sqoop to load data for creation of RDD’s, Datasets and Data frames in Spark SQL.
* Used Qlik Replicate to manage the Change Data Capture (CDC) and automate the data loading into HDFS.
* Data ingestion from Netezza to HDFS using automated Sqoop scripts.
* Used Spring Kafka API calls to process the messages smoothly on Kafka Cluster setup
* Created Airflow Scheduling scripts in Python.
* Worked on end-to-end machine learning workflow, written python code for gathering the data from  Azure Synapse, snowflake, data preprocessing, feature extraction, feature engineering, modeling, evaluating the model, deployment. Written python code for exploratory data analysis using Scikit-learn machine learning python packages- NumPy, Pandas, Matplotlib, Seaborn, stats models, pandas profiling.
* Stayed up-to-date with industry trends and emerging technologies in data engineering and Java development.
* Cloud computing implementation experience using HDInsight, Azure Data Lake (COSMOS), Azure Data Factory, Azure Machine Learning & PowerShell scripting.
* Developed Spark scripts by using Python in PySpark shell command in development.
* Creating Databricks notebooks using SQL, Python and automated notebooks using jobs.
* Expertise in Creating, Debugging, Scheduling and Monitoring jobs using Airflow and Oozie.
* Responsible for designing and implementing the data pipeline using Big Data tools including Oozie, Airflow.
* Used several python libraries like Python, numPY and matPlotLib.
* Ensured data security and compliance with industry standards and regulations while using Falcon for data processing.
* Build validation module to summarize the data between source and target using pyspark
* Utilized PL/SQL for advanced data querying and optimization tasks.
* Worked on various machine learning algorithms like Linear regression, logistic regression, Decision trees, random forests, K- means clustering, Support vector machines, XGBoosting on client requirements.
* Creating pipelines, data flows and complex data transformations and manipulations using ADF and PySpark with Databricks.
* Skilled in using collections in Python for manipulating and looping through different user defined objects.
* Created Services to consume REST API's and to communicate between components using Dependency Injection provided by Spring Framework.
* Used Hive for creation of ORC formatted tables and used ADF for data orchestration to Azure database. Data copied from ADLS to Azure SQL database using ADF pipelines invoked using PowerShell scripting.
* Developed SQOOP queries to export HIVE tables to Azure SQL database scheduled using CRON tab
* Utilize Kubernetes, Azure Databricks, and other Azure services.
* Used Pandas API to put the data as time series and tabular format for east timestamp data manipulation and retrieval.
* Overcame challenges like data migration from MySQL to MongoDB.
* Manage end to end complex data migration, conversion, and data modeling (using Alteryx, SQL), and create Data visualization using tableau to develop high quality dashboards.
* Usage of tools like Mongo Vue and NoSQL Manager for migration of data between databases without any kind of data loss.
* Configured OBIEE Metadata Objects, including repository, tables, schemas, variables, and reports. Designed and developed Informatica ETLs Views, Materialized Views, and modeling in OBIEE.
* Optimizing of existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames and Pair RDD's.
* Responsible for managing MongoDB environment with high availability, performance and scalability perspectives.
* Developed ADF pipelines to move data from on - premise source systems to COSMOS, from COSMOS (with data transformation) to Azure Warehouse (staging), from Azure warehouse to Azure ML (for scoring) and appending scores back to the data in Azure warehouse.
* Involved in automating end-2-end model refresh process using COSMOS, Azure Data Factory (ADF) pipelines, Azure ML and PowerShell scripting.
* Creation, configuration and monitoring Shards sets. Analysis of the data to be shared, choosing a shard Key to distribute data evenly. Architecture and Capacity planning for MongoDB clusters. Implemented scripts for mongo DB import, export, dump and restore.
* Deliver end-to-end Data Archive and Data Subset solutions using Informatica Power Center, ILM and TDM.
* Experience in python, Jupyter, Scientific computing stack (numpy, scipy, pandasand matplotlib).
* Created DAG interface, abstract class and concrete classes to interact with persistence entities.
* Worked on container-based technologies like Kubernetes and OpenShift. Implemented APIs to solve technology problems using NoSQL and graph databases like Neo4j and MongoDB. Experience in using collections in Python for manipulating and looping through different user defined objects.
* Design and develop data solutions in Azure Data Lake, Data Factory, and SQL Database environment.
* Worked on migrating Map Reduce programs into Spark transformations using Spark and Scala.
* Experienced in developing Web Services with Python programming language.
* Designed, Developed extensive additions to existing Struts, Java, J2EE Web Application utilizing Service Oriented Architecture (SOA) techniques.
* Experienced in using Docker tools for end-to-end integration testing.
* Used Apache Camel jobs to create data pipelines between multiple Kafka messaging systems.
* Experienced in version control tools like GIT and ticket tracking platforms like JIRA.

**Environment:** Python, Java/J2EE, AWS, Glue, Kafka, Azure HDInsight, Cosmos, Databricks, Data Lake, Cosmos DB, MySQL, Azure SQL, Snowflake, MongoDB, Cassandra, Teradata, Azure AD, Git, Blob Storage, Data Factory, Data Modeler, Scala, Hadoop (HDFS, MapReduce, Yarn), Spark, PySpark, Airflow, Hive, Sqoop, HBase, Cassandra, Oozie, Talend, Power BI, Tableau

**Big Data Engineer Sep 2014 – Jul 2016**

**AEGON Hyderabad, India**

* Working extensively on advanced features of PL/SQL like records and Tables in Amazon RDS for Oracle. Designed, supported and continuously enhanced the project code base, continuous integration pipeline, etc.
* Worked on end to end data migration and data integration projects within the company which are based on technical design in AWS/Athena/Sagemaker/Jupyter Lab, Hadoop and Big Data.
* Collaborated with Snowflake Data Warehouse implementation teams, BI Administrators, Developers and Analysts for successful development of BI reporting and analytic solutions.
* Designed, developed, implemented and supported reporting and analytics applications leveraging tools such as Cognos, Tableau, SPSS, R, Python NodeJS to meet client objectives and requirements.
* Worked with cutting-edge AWS cloud technologies to deliver next-Gen Cloud solutions
* Worked on designing, developing, and deploying modular AWS Cloud-based systems.
* Worked on developing and maintaining AWS cloud solutions in accordance with EMR for running big data frameworks, such as Apache Hadoop and Apache Spark, on AWS to process and analyze vast amounts of data.
* Interacted with clients, providing cloud support, and making recommendations based on client needs.
* Gathered and translated requirements from clients for building insightful, compelling reports and dash boards in Tableau that helps end users in processes and procedures for cross-application and cross-function control reporting using Tableau BI tool.
* Worked on Jupyter Notebook for data cleaning, building visualizations, creating machine learning models and other data manipulating related tasks.
* Create notebooks using AWS Databricks, Python, and Spark for data processing.
* Worked on defining BI standards, guidelines and best practices related to the AWS/Athena/Sagemaker/Jupyter cloud platform for clients, client services, and technical teams.
* Developed interactive Solutions in using different platforms like Tableau, Alteryx, OBIEE, Power BI, Looker.
* Maintained accurate and complete technical documents on confluence and Jira for the data Migration from legacy system to AWS Datalake platform.
* Deployed Kubernetes in AWS using Kinesis. Setup cluster, replicator and deployed multiple containers in a pod environment.
* Integrating data from various sources, including Kafka, Flume, Kinesis, and other message queues or data streams.
* Developed Spark Scripting and PySpark programs using Scala API’S, Pandas, NumPy, Python on Azure HDInsigt for Data Aggregation, Validation to compare the performance of Spark with Hive and SQL and implemented Spark using Scala and Sparks SQL for Faster Testing and Processing of data.
* Developed analytical component using Scala, PySpark and Spark Streaming.
* Used Apache Airflow for Growth analytics to compute metrics around guest and host engagement as well as growth accounting.
* Used Apache Airflow for Sessionization to compute clickstream and time spent datasets and for Data infrastructure maintenance work, database scrapes, folder cleanup, applying data retention policies, etc.
* Built data pipeline using Azure Service like Data Factory to load the data from Legacy SQL server to Azure Data Base using Data Factories, API Gateway Services, SSIS Packages, Talend Jobs, custom .Net and Python codes.
* Worked with DataOps and DevOps team to develop and deploy a new Azure service, involving architectural design and implementation of Microsoft's Azure Container Service.
* Worked on NoSQL / SQL, Microservice, RESTful API development for interactive queries, processing of streaming data and integration with popular NoSQL database for huge volume of data.
* Defined subsets of test data from the production environments and used Informatica TDM Self-service portal for data sets testing, finding the missing data, and visualizing test data coverage.
* Exported the analysed data to the relational databases (MySQL, Oracle) using Sqoop from HDFS on Cloudera and accessing Kafka cluster to consume data to Hadoop and analysing the data by performing Hive queries and running Pig Scripts.
* Worked on analysing business, system and data mapping requirements and developing ETL data pipelines for real-time streaming of data using Kafka, Spark and Airflow.
* Built pipelines to move hashed and un-hashed data from Azure Blob storage to Datalake and wrote AZURE POWERSHELL scripts to copy or move data from local file system to HDFS Blob storage
* Involved and supported creating code for Data Ingestion - Historical, Incremental. IT flattening, curation logic cr complex business scenarios. Developing code and performing validation to support data movement from HDFS files to DB2. Netezza databases for end to end business logic validation
* Worked on SQOOP implementation which helps in loading data from various RDBMS sources to Cloudera Hadoop systems.
* Worked on developing of Confidential Data Lake and in building Confidential Data Cube on Microsoft Azure HDINSIGHT cluster.
* Worked on Data Migration process using Azure by integrating with Git-hub repository and Jenkins and worked on Angular and React for SPA development.

**Environment:** Python, Java, Tableau, Looker, Power BI, Snowflake, Kafka, Oracle, MapReduce, HDFS, Netezza, DB Visualizer, Oozie, WinSCP, UNIX Shell Scripting, HIVE, Postman, Cloudera (Hadoop distribution), AWS, Devops Docker, Informatica power center, Test Data, Data Management, IICS JIRA, etc.

**Teradata Developer Aug 2013 – Sep 2014**

**Region Bank Hyderabad, India**

* Analyzing, designing and developing ETL strategies and processes, writing ETL specifications, Informatica development, and administration and mentoring other team members.
* Led the team in designing and planning functional, regression, GUI and back-end testing.
* Actively participated in writing and executing SQL queries for back-end testing
* Coordinated the smooth implementation of all the go-lives; extending post-implementation, and application maintenance support.
* Worked on migrating data from Teradata to AWS using Python and BI tools like Alteryx.
* Worked as a key contributor in the people development by conducting multiple knowledge.
* Developed mapping parameters and variables to support SQL override.
* Worked on Teradata and its utilities - tpump, fastload through Informatica. Also created complex Teradata Macros.
* Worked with the data from Mainframe DB2 and ADABAS using Apache Sqoop and stored it in AWS S3 to create the data lake for Apache Spark, Athena, and Redshift
* Involved in the installation and upgrade of Informatica TDM.
* Have used Application Integration, Data Integration, Data Synchronization, Mass ingestion, administration, Monitor etc. services in IICS.
* Worked and Implemented Pushdown Optimization (PDO) to optimize performance issues of complex mappings involving numerous transformations and hence degrading the performance of the session.
* Involved in Performance tuning at source, target, mappings, sessions, and system levels.
* Worked on Change Data Capture (CDC) using CHKSUM to handle any change in the data if there is no flag or date column present to represent the changed row.
* Worked on reusable code known as Tie outs to maintain the data consistency. Compared the source and target after the ETL loading is complete to validate no loss of data during the ETL process.
* Developed ETL jobs and pipelines using tools such as IBM Datastage and Apache Nifi.
* Develop ETL and ELT Process using SPARK, SCALA, HIVE and HBASE. Setting up Clusters and jobs for Azure Databricks.
* Worked on Power Exchange bulk data movement process by using Power Exchange Change Data Capture (CDC) method, Power Exchange Navigator, Power Exchange Bulk Data movement. Power Exchange CDC can retrieve updates at user-defined intervals or in near real time.
* Worked independently on the critical milestone of the project interfaces by designing a completely parameterized code to be used across the interfaces and delivered them on time in spite of several hurdles like requirement changes, business rules changes, source data issues and complex business functionality.

**Environment:** Teradata, ETL, Informatica IICS, Databricks, SQL Server, Shiny, Mahout, HBase, HDFS, Flat files, UNIX- AIX, Windows XP, MS Word, Excel

**Data Engineer**

**KPMG, Madhapur, Hyd - India May 2011 – Jul 2013**

* Gathered all the Sales Analysis report prototypes from the business analysts belonging to different Business units.
* Worked with Master SSIS packages to execute a set of packages that load data from various sources onto the Data Warehouse on a timely basis.
* Involved in Data Extraction, Transformation and Loading (ETL) from source systems.
* Responsible with ETL design (identifying the source systems, designing source to target relationships, data cleansing, data quality, creating source specifications, ETL design documents.
* The data received from Legacy Systems of customer information were cleansed and then Transformed into staging tables and target tables in DB2.
* Used External Tables to Transform and load data from Legacy systems into Target tables.
* Use of data transformation tools such as DTS, SSIS, Informatica or Data Stage.
* Conducted Design reviews with the business analysts, content developers and DBAs.
* Designed, developed, and maintained Enterprise Data Architecture for enterprise data management including business intelligence systems, data governance, data quality, enterprise metadata tools, data modeling, data integration, operational data stores, data marts, data warehouses, and data standards.
* Incremental loading of Fact table from the source system to Staging Table done on daily basis.
* Responsible for data extraction and data ingestion from different data sources into Hadoop Data Lake by creating ETL pipelines using Pig, and Hive.
* Migrate on in-house database to AWS Cloud and also designed, built, and deployed a multitude of applications utilizing the AWS stack (Including EC2, Glue, RDS) by focusing on high-availability and auto-scaling.
* Built pipelines to move hashed and un-hashed data from XML files to Data lake.
* Extensively worked with Spark-SQL context to create data frames and datasets to preprocess the model data.
* Experienced in loading and transforming large sets of Structured, Semi-Structured and Unstructured data and analyzed them by running Hive queries and Pig scripts.
* Configured Spark Streaming to receive real time data from the Apache Kafka and store the stream data to DynamoDB using Scala.
* Involved in designing the row key in HBase to store Text and JSON as key values in HBase table and designed row key in such a way to get/scan it in a sorted order.
* Developed API for using AWS Lambda to manage the servers and run the code in the AWS.
* Wrote Junit tests and Integration test cases for those Microservices.
* Work heavily with Python, C++, Spark, SQL, Airflow, and Looker
* Used Apache Mongo DB (NoSQL) in AWS Linux instance in parallel to RDS MySQL to store and analyze and migrated an existing on-premises application to AWS and used AWS services like EC2 and S3 for small data sets.
* Developed NiFi workflow to pick up the multiple files from ftp location and move those to HDFS on daily basis.
* Used various Transformations in SSIS Dataflow, Control Flow using for loop Containers and Fuzzy Lookups and Implemented Event Handlers and Error Handling in SSIS packages.
* Involved in Cloudera Navigator access for auditing and viewing data.
* Generated document coding to create metadata names for database tables.
* Analyzed metadata and table data for comparison and confirmation.
* Adhered to document deadlines for assigned databases.
* Ran routine reports on a scheduled basis as well as ad-hocs based on key point indicators.
* Develop DataStage jobs to cleanse, transform and load data to Data Warehouse and sequencers to encapsulate the DataStage job flow.
* Designed data visualizations to analyze and communicate findings.

**Environment**: Linux, Erwin, SQL Server, Crystal Reports9.0, AWS, HTML, DTS, SSIS, Informatica, Data Stage Version 7.0, Oracle, Toad, MS Excel.

**Education:** BE (Instrumentation) – 2012 Mumbai University, M. G. Road, Fort, Mumbai-400032