JAISHREE SREEKUMAR

1. 597-0128 • [sreejaishree21@gmail.com](mailto:sreejaishree21@gmail.com" \t "_blank)

**PROFESSIONAL SUMMARY:**

* Around 11+ years of experience as a Data Engineer and extensively worked with designing, developing, and implementing Data models for enterprise-level applications and BI solutions.
* 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.
* Experience in application of various data sources like Oracle SE2, SQL Server, Flat Files and Unstructured files into a data warehouse.
* Extensive experience with MSFT 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.
* Experience in Extraction, Transformation and Loading (ETL) data from various sources into Data Warehouses, as well as data processing like collecting, aggregating and moving data from various sources using Apache Flume, Kafka, PowerBI and Microsoft SSIS.
* 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.
* Excellent understanding of best practices of Enterprise Data Warehouse and involved in Full life cycle development of Data Warehousing.
* 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, 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.

**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 |  | MSFT Azure Data Lake Storage Gen 2, Azure Data Factory,Data warehouse, 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, Informatica, Talend, SSIS, and SSRS |
| Version Control & Containerization tools |  | GitHub, Bitbucket, Docker, Kubernetes |
| Operating Systems |  | Unix, Linux, Windows, Mac OS |

**WORK EXPERIENCE**

# Clent: Walmart Rogers, AR Oct 2021 - present

**Lead Data Engineer**

* Responsible for implementation, administration and management of Hadoop infrastructures and worked on Azure cloud services to do machine learning on big data.
* Improved the performance of end of the day batch process by introducing objects features of Oracle and reducing the querying on the entity tables.
* Developed Python based API (RESTful Web Service) to track revenue and perform revenue analysis.
* Worked on Data extraction (Interface) into flat files using UTL File and Data Load from flat files using SQL\*Loader and Unix Shell Script.
* Monitoring and working on Platform incidents, Alerts and Service requests. Analyzing the windows azure platform issues and reporting them to windows azure developers.
* The data stored in Azure Synapse is then processed using Azure HDInsight (Spark) or Power BI.
* Developed a detailed project plan and helped manage the data conversion migration from the legacy system to the target Snowflake database.
* Performing testing activities like running the Jobs, Extracting the data using necessary queries from database transform, and upload into the Data warehouse servers.
* Developed Procedures in SAP Hana to load data for Product interface.
* Design, develop, and test dimensional data models using Star and Snowflake schema methodologies under the Kimball method.
* Developed data pipeline using Spark, Hive, Pig, python, Impala, and Hbase to ingest customer
* Integrated Apache Storm with Kafka to perform web analytics. Uploaded click stream data from Kafka to Hdfs, Hbase and Hive by integrating with Storm
* Developed automated workflows for monitoring the landing zone for the files and ingestion into HDFS in Bedrock Tool and Abinitio.
* Involved in converting Hive/SQL queries into spark transformation using spark RDDs, Python and Scala.
* Designed Reporting Data warehouse based on User Requirements
* Performed Hive incremental updates, merge, partitioning, bucketing, windowed operations, efficient and effective joins for faster data operations.
* Designed Packages to extract, transfer, load (ETL) existing data into SQL Server from different environments for the SSAS cubes (OLAP)
* Tracking the issues that are raised in JIRA, doing SLA management to make sure that SLA is not missed during UAT and postproduction support.
* Used Spark SQL to perform complex data manipulations, and to work with large amounts of structured and semi- structured data stored in a cluster using Data frames/ Datasets.
* Developed Spark Python modules for machine learning & predictive analytics in Hadoop.
* Evaluation of Hadoop infrastructure requirements and design/deploy solutions (high availability, big data clusters and involved in cluster monitoring and troubleshooting Hadoop issues.
* Worked with application teams to install OSs and Hadoop updates, patches, version upgrades as required.
* Used to implement the Slowly Changing Transformation, to maintain Historically Data in Data warehouse.
* Experienced in handling data from different datasets for gathering and pre-processing.
* Imported and exported data from Teradata to HDFS and vice-versa.
* Analyze data and provide support to business users during UAT phase and post go-live.
* Monitor long running critical jobs and provide solutions for reducing the time (performance tuning).
* Performing testing activities like running the Jobs, Extracting the data using necessary queries from database transform, and upload into the Data warehouse servers.
* Strong understanding of Hadoop eco systems such as HDFS, MapReduce and HBase.
* Utilized Amazon Web Services to perform big data analytics.
* Actively participated in software development lifecycle (scope, design, implement, deploy, test), including design and code reviews.

**Environment:** Hadoop, HDFS, Map Reduce, RDBMS/DB, MySQL, CSV, Java, J2EE

# Client: Hcentive Lakewood, CO Jan 2017 - Sep 2021

**Lead Cloud Data engineer/ ETL Developer**

* Responsible for provisioning and setting up the data platform technologies that are on-premises and in the MSFT Azure cloud.
* Designed several DAGs (Directed Acyclic Graph) for automating ETL pipelines
* Monitoring and working on Platform incidents, Alerts and Service requests. Analyzing the windows azure platform issues and reporting them to windows azure developers.
* Experienced in fact dimensional modeling (Star schema, Snowflake schema), transactional modeling and SCD (Slowly changing dimension)
* Performed Data Migration to GCP
* Selecting the appropriate MSFT Azure service based on compute, data or security requirements and leveraging Azure SDKs to interact with Azure services from your application.
* Implemented the migration of data from multiple on-premises servers into cloud using Data Factory service in Azure and Data Migration Assistant application.
* Designed and implemented configurable data delivery pipeline for scheduled updates to customer facing data stores built with Python
* Implemented a Continuous Delivery pipeline with Docker, and Git Hub and AWS
* Worked in building the Data Ingestion pipelines in MSFT Azure using Data Factory and ingest the data into Azure Data Lake Storage Gen 2 in either batches or as a stream.
* Built reporting data warehouse from ERP system using Order Management, Invoice & Service contracts modules.
* Experienced in ETL concepts, building ETL solutions and Data modeling
* Experience in fact dimensional modeling (Star schema, Snowflake schema), transactional modeling and SCD (Slowly changing dimension)
* Format the data collected in Data Lake into relational format using Azure PolyBase Service and store that data in Azure Synapse for further processing on the data.
* Worked with Power BI to create visualizations on data based on the customer needs.
* The data stored in Azure Synapse is then processed using Azure HDInsight (Spark) or Power BI.
* Worked on architecting the ETL transformation layers and writing Spark jobs to do the processing.
* Performed data extraction, transformation, loading, and integration in data warehouse , operational data stores and master data management
* Also used Spark scripts for data streaming and transforming data in less time using the Spark clusters in Azure Databricks.
* Acted as SME for Data Warehouse related processes.
* Provisioning the required servers based the latency and throughput requirements.
* Identifying, escalating, diagnosing and mitigating service impacting events on Microsoft Azure Core components as per defined SLAs and Diagnose incidents using the Microsoft Azure Engineering System.
* Responsible for provisioning of Azure services on the cloud to build scalable production systems load balancers.
* Loaded application analytics data into data warehouse in regular intervals of time
* Worked on maintaining high availability by providing the servers closest to the users accessing the data and Disaster recovery by correctly defining the RPO (Recovery Point Objective) and RTO (Recovery Time Objective).
* Worked with Oracle SQL Server for analytical purposes.
* Wrote SQL queries to identify and validate data inconsistencies in data warehouse against source system.
* Analyzed the system for new enhancements/functionalities and perform Impact analysis of the application for implementing
* Built performant, scalable ETL processes to load, cleanse and validate data
* Configured Oracle databases such that they can withstand attacks and the loss of information.
* Formulating data dictionaries in Oracle that are congruent with task specifications.
* Used Azure Resource Manager (ARM) to deploy, update, or delete all the resources for your solution in a single, coordinated operation.
* Implement and manage CI/CD pipelines.
* Developing, Managing, and Operating the scalable, highly available, and fault tolerable systems on Azure. **Environment:** Azure Cosmos DB, Azure SQL Database, Azure Data Factory, Data Migration Assistant, SQL Data Explorer, Azure HDInsight, Azure Synapse Analytics, Azure PolyBase, Azure Active Directory, Azure Data Lake Storage Gen2, Azure Power BI, Azure Data Warehouse.

# Client: Foray Software Pvt Ltd. India Nov 2011 - Aug 2015 Data Engineer

* Worked with Data Warehouse team in developing Dimensional Model and analyzing the ER-Diagrams
* Identified and analyze stakeholders and subject areas.
* Participated in Business Analysis, talking to business Users and determining the entities and attributes for Data Model.
* Experience in building and architecting multiple Data pipelines, end to end ETL
* Implemented reporting Data Warehouse with online transaction system data.
* Identified and determined physical attributes and their relationships through cross-analysis of functional areas.
* Identified and analyzed source data coming from Oracle, SQL server and flat files.
* Extensively used ERWIN to design and restructure Logical and Physical Data Models.
* Evaluated and enhanced current data model as per the requirements.
* Performed forward and reverse engineering, applying DDLs to database in restructuring the existing data Model using ERWIN.
* Performed data engineering functions: data extract, transformation, loading, and integration in support of enterprise data infrastructures - data warehouse, operational data stores and master data management
* Developed and maintained data warehouse for PSN project.
* Participated in the development of an enterprise warehouse giving mainstream users the ability to access, analyze and share information in the client’s database.
* Designed ETL specification documents to load the data in target using various transformations according to the business requirements.
* Used Informatica- Power center for extracting, transforming, and loading.
* Performed Data profiling, Validation, and Integration.
* Created materialized views to improve performance and tuned the database design.
* Involved in Data migration and Data distribution testing.
* Worked closely with Business Analyst and report developers in writing the source to target specifications for Data warehouse tables based on the business requirement needs.
* Performed testing, knowledge transfer and mentored other team members.

**Environment:** Oracle 10g, Data Extract, Data infrastructures-Data warehouse, Microsoft Excel, MicroStrategy, Database design, MS Visio, MS Project, Informatica, Matlab, SAS.

# Client: Genpact, Chennai, Tamil Naidu March 2011 - Nov 2011 Data Engineer

* Worked on google cloud platform (GCP) services like compute engine, cloud load balancing, cloud storage, cloud SQL, stack driver monitoring and cloud deployment manager.
* Loading salesforce data every 15 min on incremental basis to Big Query, Google DataProc, and GCS Buckets.
* Using rest API with Python to ingest Data from external applications to BigQuery.
* Monitoring Big Query, DataProc and Cloud DataProc Jobs via stack driver for all environments.
* Developed and maintained data lineage tracking tools to ensure data quality and compliance with regulations such as GDPR and CCPA
* Evaluate Snowflake Design considerations for any change in the application.
* Experience in setting up and configuring version control systems, such as Git, and integrating them with CI/CD tools.
* Consulting on Snowflake Data Platform Solution Architecture, Design, Development, and deployment focused to bring the data driven culture across the enterprises.
* Collaborated with data governance teams to establish data quality and compliance standards and implemented data lineage tracking as part of these standards.
* Involved in Designing and Developing Enhancements of CSG using AWS API.
* Created Snowflake Schemas by normalizing the dimension tables as appropriate and creating a Sub Dimension named Demographic as a subset to the Customer Dimension.
* Created internal and external stage and transformed data during load.
* Ensured data lineage is captured at every stage of the data processing pipeline, including data ingestion, transformation, and output.
* Performed Data Analysis, Data Migration, Data Cleansing, Transformation, Integration, Data Import, and Data Export through Python
* Create firewall rules to access Google Dataproc from other machines.
* Written Kafka REST API to collect events from front end.
* Automated data lineage tracking processes using GCP tools such as Cloud Functions and Cloud Run, resulting in significant time and cost savings.
* Created projects, VPC's, Subnetwork's, GKE Clusters for environments QA3, QA9 and prod using Terraform.
* Build web data scrapers from APIs using Python to extract data from external sources.
* Worked on performance tuning of long running Spark Jobs using Data Frames, Spark-SQL, Memory tuning, Executors tuning, Spark YARN.
* Developing Spark applications using Scala and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats.
* Using Kafka and integrating with the Spark Streaming.
* Worked closely with data scientists and analysts to ensure data lineage is accurate and accessible for their analysis and reporting needs.
* Designed and created logical Data warehouse models to support strategic decisions.
* Worked on Docker Containerization for to make pipelines runs on platform dependencies.
* Extensive knowledge in programming with Resilient Distributed Datasets (RDDs)
* Defined standards and best practices for development team for building Spark data pipelines.
* Responsible for working with Business Analyst and client SMEs to gather requirements and prepared high level and low-level design documents.
* Hands on experience in Google Cloud in various Cloud services such as Big Query and domain configuration.
* Used AWS services like EC2 and S3 for small data sets.
* Created and maintained documentation for data pipeline processes and data lineage, enabling efficient troubleshooting and knowledge sharing among team members.
* Analyzing various logs that are been generating and predicting/forecasting next occurrence of event with various Python libraries.
* Used Rally tool extensively and followed strong Agile processes such as story grooming, agile planning, sizing estimations and KANBAN board updates, etc.
* Worked on Docker Containerization for to make pipelines runs on platform dependencies.
* Created the CI/CD pipelines for ARM templates, Databricks Notebook.
* Developed Hadoop ETL architecture comprising of ‘job design pattern’, logging, error handling, email notification, change data capture, slowly changing dimension, etc.
* Established best practices for ingesting and storing large amount of source system data into HIVE tables.
* Worked on building near real-time streaming pipeline using Kafka, Spark, and Hive components

**Environment:** PySpark, Python, Google Cloud Platform, YARN, GitHub, Kubernetes, REST API, Linux, Big Query, GCS Bucket, Cloud SQL, DataProc

|  |  |
| --- | --- |
| **EDUCATION** |  |
| **Anna University, Chennai, India**  **Master of Business Administration, Human Resources** | **2011** |
| **Sri Ram Engineering College**  **Bachelor of Engineering, Electrical and Electronics Engineering** | **2009** |