Hemanth Gopisetti

Email: kgopi115522@gmail.com **Mobile: +1 (737) 973-0566**

**Summary**

* Experienced Data Engineer With 10+ years of experience In and Microsoft SQL Server Suite of Products– T-SQL, SSIS, SSRS, Power BI with expertise in Data Mapping, Data Validation.
* Hands on Experience in implementing Data warehouse solutions in AWS Redshift; Worked on various projects to migrate data from on premise databases to AWS Redshift, RDS and S3.
* Managing Database, Azure Data Platform services (Azure Data Lake (ADLS), Data Factory (ADF), Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Databricks, NoSQL DB), SQL Server, Oracle, Data Warehouse etc. Build multiple Data Lakes.
* Experience in developing data pipelines using AWS services including EC2, S3, Redshift, Glue, Lambda functions, Step functions, CloudWatch, SNS, Dynamo DB, SQS.
* Experience in developing Flink pipelines to consume streaming data from Kafka and applied business logic to massage and transform and serialize raw data and developed common Flink module for serializing and deserializing AVRO data by applying Schema.
* Utilized Kubernetes and Docker for the runtime environment for the CI/CD system to build, test, and deploy. Experience in working on creating and running Docker images with multiple micro services.
* Experienced in developing production-ready Spark applications.
* Experience with source control systems such as Git Hub, Bit bucket, and Jenkins and in CI/CD deployments.
* creating, implementing, and documenting data models for enterprise-level applications. Experience in ETL Data pipeline, Data Visualization, Data Warehousing, and Data Lake.
* Involved in the creation of RDDs and data frames from the necessary HDFS files to transform Hive queries into different Spark actions and transformations.
* knowledge of the Big Data Ecosystem, specifically using the Hadoop framework and related technologies including HDFS, MapReduce, HIVE, PIG, HBASE, STORM, YARN, OOZIE, SQOOP, Airflow, and Zookeeper. Also includes work with Spark Core, Spark SQL, Spark Streaming, Scala, and Kafka.
* A solid background in data modeling, data pipelines, and SQL and NoSQL databases. Participating in the creation and automation of ETL pipelines from start to finish using Python and SQL
* Knowledge of version control tools like GitLab, SVN Code Commit, and Bit Bucket to arrange the versions and settings of the code.
* Using Spark Context, Spark-SQL, Data frame API, Spark Streaming, MLlib, Pair RDDs, and working specifically with PySpark and Scala, an expert with Spark.
* A solid working understanding of command-line utilities in the UNIX and Linux shell environments.
* Skilled in using Agile Methodologies, Scrum stories, sprints, daily standup meetings, and pair programming to achieve high-quality deliverables on schedule.
* Having expertise in the whole software development life cycle (SDLC), designing scalable platforms, object-oriented programming, database design, and agile approaches.
* Experience with databases like DynamoDB, S3 Buckets, MySQL, and Elastic Cache on the AWS Cloud with good knowledge of IaaS, PaaS.
* Good knowledge of NoSQL databases like MongoDB, Redis, and Apache Cassandra.
* Knowledgeable of EC2, CloudWatch, Cloud Formation, and managing security groups on AWS, as well as automating, configuring, and deploying instances on these platforms.
* Strong Python PEP Guidelines compliance and experience with Linux Bash scripting.
* skilled in creating SQL queries, stored procedures, functions, packages, tables, views, and triggers for relational databases like Oracle, DB2, MySQL, and MSSQL Server.
* Excellent understanding of SQL and working knowledge of Teradata and Oracle databases.
* Created Pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards.
* Experience in GCP’s Data proc, GCS, Cloud functions, BigQuery.
* Worked with Docker and Kubernetes on multiple cloud providers, from helping developers build and containerize their application (CI/CD) to deploying either on public or private cloud.

**Technical Skills:**

|  |  |
| --- | --- |
| **Cloud Technologies** | AWS, Azure, Google cloud platform (GCP), Cloud Services (PaaS & IaaS), Active Directory, Application Insights, Azure Monitoring, Azure Search, Data Factory, Key Vault and SQL Azure, Azure Devops, Azure Analysis services, Azure Synapse Analytics (DW), Azure Data Lake, AWS Lambda |
| **IDE’s** | IntelliJ, Eclipse, Spyder, Jupyter |
| **Ensemble and Stacking** | Averaged Ensembles, Weighted Averaging, Base Learning, Meta Learning, Majority Voting, Stacked Ensemble, Auto ML – Scikit-Learn, MLjar, etc.  |
| **Databases** | Redshift, Oracle, MySQL, DB2, MS-SQL Server, HBASE, NOSQL |
| **Big Data Ecosystem** | HDFS, MapReduce, HBase, Pig, Hive, Sqoop, Kafka, Flume, Cassandra, Impala, Oozie, Zookeeper, MapR, Amazon Web Services (AWS), EMR |
| **Programming / Query Languages** | Java, SQL, Python, Frameworks (Pandas, NumPy, SciPy, Scikit-Learn, Seaborn, Matplotlib, NLTK), NoSQL, PySpark, PySpark SQL, PL/SQL, Linux shell scripts, Scala. |
| **Data Engineer/Big Data Tools / Cloud / Visualization / Other Tools** | Databricks, Hadoop Distributed File System (HDFS), Hive, Pig, Sqoop, MapReduce, Spring Boot, Flume, YARN, Hortonworks, Cloudera, Mahout, MLlib, Oozie, Zookeeper, etc. AWS, Azure Databricks, Azure Data Explorer, Azure HDInsight, GCP, Google Shell, Linux, PuTTY, Bash Shell, Unix, etc., Tableau, Power BI, SAS, Web Intelligence, Crystal Reports, Dashboard Design. |

**Work Experience:**

**Sketchers, TX, USA Jan 2024– Present**

**Sr. Data Engineer**

**Responsibilities**

* Build the Logical and Physical data model for snowflake as per the changes required and Evaluated Snowflake Design considerations for any change in the application.
* Spearheaded initiatives using AWS Glue, AWS Step Functions, and AWS Lambda for efficient data processing and orchestration.
* Involved in creating IAM Users, Groups, Roles, Identity Providers and defining Policies and applying to IAM Users and Groups. Experience creating EC2 AMI’s and migrating it to different regions
* Designed and developed ETL pipelines using AWS Glue, AWS Step Functions, AWS Lambda, and EMR for efficient data processing and orchestration. Ensured adherence to CI/CD practices.
* Designed and implemented comprehensive ETL pipelines using AWS Glue, optimizing data flow and transformation processes to support data integration across multiple sources.
* Wrote MapReduce and Spark jobs to process the data, created data models and implemented data warehousing solutions using Hive, HBase, and Redshift to store and manage large amounts of structured, unstructured data.
* Ensured data quality and consistency by developing robust ETL scripts in Python, automating data ingestion and transformation tasks. Worked on Docker based containers for using Airflow
* Created Python scripts to automate AWS infrastructure provisioning tasks such as creating S3 buckets in AWS, create EMR cluster, add steps to running EMR cluster to exe Spark jobs
* Migrating on premise ETLs to GCP using cloud native tools such as Big query, Cloud Data Proc, Google Cloud Storage, Composer.
* Involved in the development of real time streaming applications using PySpark, Apache Flink, Kafka, Hive on distributed Hadoop Cluster.
* Compared Self hosted Hadoop with respect to GCPs Data proc and explored Big Table (managed HBase) use cases, performance evolution.
* Designed and implemented highly performant data ingestion pipelines from multiple sources using Apache Spark
* Implemented advanced analytics use-cases using SAS programming techniques for predictive modeling and data analysis, such as fraud detection, customer segmentation, and churn prediction.
* Collaborated with stakeholders to understand business requirements and implemented solutions leveraging SAS technologies to meet business objectives.
* Have deployed Prometheus with Grafana to monitor the Kubernetes cluster and configured alerts firing when different conditions met.
* Working within an Agile delivery to deliver proof of concept and production implementation in iterative sprints.
* Designed and implemented Incremental framework to handle incremental load process effectively in Airflow.
* Worked with Docker and Kubernetes on multiple cloud providers, from helping developers build and containerize their application (CI/CD) to deploying either on public or private cloud.
* Implemented Azure Cloud Infrastructure utilizing ARM templates. Created and managed Azure Web-Apps and provided access permission to Azure AD users
* Developed PySpark script to setup the data pipeline and worked on developing ETL streams using Databricks
* worked with socket API for programming distributed applications on the Internet worked with HTML, CSS, XML, JavaScript, jQuery and Ajax while integrating with spring and hibernate Designed and created automation workflows and execution.
* written Terraform scripts to automate AWS services which include ELB, CloudFront distribution, RDS, EC2, database security groups, Route 53, VPC, Subnets, Security Groups, and S3 Bucket and converted existing AWS infrastructure to AWS Lambda deployed via Terraform.
* Worked extensively on Java/J2EE for developing Web Applications by using Spring/Spring MVC, JDBC, JSP, Servlets, Hibernate, XML, and Web Services. Integrated Hibernate and Spring MVC technologies.
* Developed python code for different tasks, dependencies, SLA watcher and time sensor for each job for workflow management and automation using Airflow tool.
* Installed KAFKA on Hadoop cluster and configure producer and consumer coding part in java to establish connection from twitter source to HDFS.
* Developed JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the SQL Activity.
* Created pipelines in Azure using Azure Data Factory v2 to get the data from disparate source systems by using different Azure Activities like Move &Transform, Copy, filter, for each, Databricks etc.

**Environment**: PySpark, Python, AWS Glue, SAS, Lambda, Athena, Teradata, Airflow, S3, Boto3, Databricks, IAM, SNS, Azure, Json, GCP, snowflake, Terraform, Java, MongoDB, ADF

**NC Fast, Raleigh NC, USA July 2021 – Dec 2023**

**Sr. Data Engineer**

* Developed methods for ingesting data from different sources and processing Data-at-Rest utilizing Big Data tools like Hadoop, Map Reduce Frameworks, HBase, and Hive.
* Worked on workflow engine called Oozie was created to handle numerous Hive, Pig, Tealeaf, Mongo DB, Git, Sqoop, and Spark processes.
* Utilizing Informatica Designed, a variety of Mappings were created using the collection of all Sources, Targets, and Transformations.
* Worked on Snowflake Schemas, Data Warehousing and processed batch and streaming data load pipeline using Snow Pipe and Matillion from data lake.
* Effectively integrated Apache Airflow with diverse data sources and platforms, including leading cloud services like AWS, GCP as well as various databases.
* Involved in extracting and uploading data into Azure data lake using Informatica aws plugin and created Informatica Mappings and workflows to Run Informatica MDM stage, Base Objects and Match Merge Jobs.
* Use HDFS and Sqoop command to handle extensive data input. Gather "partitioned" data in a variety of storage formats, such as text, JSON, Parquet, etc. engaged in the loading of data from the LINUX file system to the HDFS.
* Configuring and integrating the required AWS services in accordance with the business requirement to start Infrastructure as a code (Iaas) in the AWS cloud platform from scratch.
* Hands of experience in GCP, Big Query, GCS Bucket, G-Cloud function, cloud dataflow, pub/suB cloud shell, GSUTIL, BQ command line utilities, Data Proc, Stack driver
* Worked with Core Java, Spring, Spring MVC, and RESTful web services.
* Designed and implemented ETL processes using Databricks to ensure accurate and efficient data processing, transformation, and loading into various data stores.
* Participating in all phases of the project and its scope, reference data for MDM was used to develop a Data Dictionary and a Mapping from Sources to the Target in the MDM Data Model.
* Creating a data pipeline using Spark Scala, and Apache Kafka to ingest data from a CSL source and store it in an HDFS secured folder.
* Designing and deploying multi-tier applications with an emphasis on high availability, fault tolerance, and auto scaling on AWS Cloud Formation utilizing all of the AWS services (EC2, AWS Glue, Athena, Lambda, S3, RDS, Dynamo DB, SNS, SQS, IAM, etc.).
* Designing and deploying several applications that make use of practically all AWS services, with an emphasis on high availability, fault tolerance, and auto-scaling in AWS Cloud Formation, including EC2, RedShift, S3, RDS, Dynamo DB, SNS, and SQS.
* Have good experience working with Azure BLOB and Data Lake storage and loading data into Azure SQL Synapse analytics (DW).
* Custom Kafka producers and consumers have been developed for a variety of publishing and subscribing to Kafka topics and worked on Azure Synapse Analyticsfor implementing Pyspark Notebooks.
* Using sophisticated SQL scripting, ETL tools, Python, Shell scripting, and scheduling tools, create and deploy several ETL systems with different data sources. XML, Web feeds, and file processing utilizing Python, Unix, and SQL for data profiling and manipulation.
* Data migration between many Teradata servers was done to assist the development team.
* Architecting and creating serverless web applications with AWS Lambda, API Gateway, Dynamo DB, and Security Token Service (STS).
* Have worked on Apache Flink to implement the transformation on data streams for filtering, aggregating, update state and used Flink Streaming for pipelined Flink engine to process data streams to deploy new API.
* proficient in coding using Python and Java script.
* Create data models for the apps, views, tables, and other database objects that contain metadata.
* Built automated ETL pipelines that can process a wide range of data sources.
* Building Python-based database models, APIs, and views to create an interactive web-based solution. Coordinating the development of the team using the GIT version control tool.
* Utilized Terraform scripts from Jenkins, involved in provisioning AWS infrastructure.

**Environment**: PySpark, Python, AWS Glue, SAS, Lambda, Athena, Teradata, Airflow, S3, Boto3, Databricks, IAM, SNS, Azure, Json, GCP, snowflake, GCP, Java, MongoDB

**Kajeet Inc, McLean, VA Jan 2021 – July 2021**

**Sr Data Engineer**

**Responsibilities:**

* Worked on data pre-processing and cleaning the data to perform feature engineering and performed data imputation techniques for the missing values in the dataset using Python.
* Involved in Migrating Objects from Teradata to Snowflake, Created Snow pipe for continuous data load and scheduled different Snowflake jobs using NiFi.
* Created Batch & Structured streaming Azure Data Factory (ADF) pipelines using Databricks Spark, Delta tables.
* creating python topology script to generate cloud formation template for creating the EMR cluster in AWS.
* Worked on Batch and real-times park jobs were created using Scala.
* Designed and Configured Azure Cloud relational servers and databases analyzing current and future business requirements, worked on migration of data from On-prem SQL server to Cloud databases (Azure Synapse Analytics (DW) & Azure SQL DB).
* Used cloud components and connectors to make API calls for accessing data from cloud storage Amazon S3 in Talend open Studio and have created mappings in TALEND.
* Stored and Wrote a Python program that checks a Linux directory for incoming XML files, upload all new files to a Google Cloud Platform storage location before the data is parsed and loaded into a Big Query table
* Designed and developed scalable and cost-effective architecture in AWS Big Data services for data life cycle of collection, ingestion, storage, processing, and visualization
* Responsible for estimating the cluster size, monitoring, and troubleshooting of the Spark Databricks cluster
* Involved in creating End-to-End data pipeline within distributed environment using the big data tools, Spark framework and Tableau for data visualization.
* Extensively used AWS S3, EC2 and EMR instances to deploy and test the applications in various environments (DEV, QA, PROD).
* Worked in developing Spark programs in Scala to perform Data Transformations and actions like map, filters, collect, joins, pre-aggregations and creating Datasets, Data frames and writing spark SQL queries
* Used Bulk Collections for better performance and easy retrieval of data, by reducing context switching between SQL and PL/SQL engines.
* Created PL/SQL scripts to extract the data from the operational database into simple flat text files using UTL\_FILE package.
* Developed the application using spark advanced features like Broadcast variables, Accumulators and Controllable Partitioning
* Involved in understanding the business and technical requirements and implementing solution based on the Hadoop ecosystem
* Created Spark applications using RDD, Data frames for data preprocessing for greater performance and developed a python script to validate the files in one S3 with same daily files in another bucket.
* Written Post processing shell scripts to automate the file transfer workflow and managed deployments in EKS managed Kubernetes, setup multi nodes cluster and deployed containerized applications.
* Automated the workflow by scheduling the jobs using arow job scheduler.
* Experienced in ELT process that is key to the data warehouse layer and connected to snowflake, loaded the Continuous data transfers through Snow pipe and written Snow SQL queries to analyze the data.
* Responsible for estimating the cluster size, monitoring, and troubleshooting of the Spark data bricks cluster.

**Infodat International, Hyderabad, India Oct 2017 – Nov 2020**

**Big Data Engineer**

**Responsibilities:**

* Partnered with stakeholders to gather data flow requirements and analyzed data using Google Analytics and Excel. Collaborated with senior members to assess and prioritize cyber risks, ensuring data security and integrity.
* Created and published Test Plan, Test Strategy documents on SharePoint for project team members.
* Normalized the tables/relationships to arrive at effective relational schemas without any redundancies.
* Participated in daily stand-up meetings and effectively communicated progress, issues, and resolutions with the team.
* Built and managed ETL pipelines using AWS Glue, automating data extraction, transformation, and loading processes to streamline data operations.
* Managing Database, Azure Data Platform services (Azure Data Lake (ADLS), Data Factory (ADF), Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Databricks, NoSQL DB), SQL Server, Oracle, Data Warehouse etc.
* Enhanced data security and integrity within ETL workflows, ensuring compliance with industry standards and best practices.
* Collaborated with senior members in the assessment and prioritization of identified cyber risks.
* Analyzed and documented existing and future processes related to the procurement of vendors.
* Involved in updating the project scope and objective, and gathered business requirements by interacting with Business users, Business owners and stakeholders including the Managers, Traders, Risk Managers, and Compliance team.
* created reusable Azure Data Factory pipeline to pull the data from Teradata and load it to ADLS in incremental fashion/truncate and reload.
* Prepared data for Request for Proposal (RFP), drafted the RFP and coordinated with facility maintenance.
* Partnered with different work locations to gather data flow, analyzed the data using Google Analytics and Excel
* Designed and created automation workflows for data pipeline development, leveraging AWS services.
* Maintained data integrity of databases, made changes as required enhancing accuracy and access.
* Organized Joint Application Development (JAD) sessions with data steward and ETL teams while having walkthroughs for mapping documents.
* Designed and developed data ingestion pipelines using Sqoop and Kafka to collect and stream data from various sources.
* wrote scripts using Python API, and Spark API for analyzing the data Hands-On experience with Spark Core, Spark SQL, Spark Streaming, and creating the Data Frames handle in SPARK with Scala.
* Created and managed Azure Web-Apps and provided access permission to Azure AD users.
* Created reports with Tableau dashboards using scatter plots, Histograms, Line charts.
* Conducted stand-up meetings daily to ensure that the project was flowing in the right direction with respect to progress, issues, tasks, and resolutions.
* Acted as a liaison between several teams within the business and technology group to implement different types of data strategies, Data Cleansing rules, Data Profiling, and validity, built data flows
* Extensively performed data validation testing and data analysis using SQL (DML) queries.
* Assisted development & QA in unit testing, functional & regression testing

**Environment:** Python, AWS Services, Hadoop, Hive, Serdes, HBase, HDFS, Pig, Apache Nifi, MapReduce, Sqoop, Spark, PySpark, Spark SQL, Scala, Kafka, Airflow, Snowflake, ETL, SQL, Avro, Parquet, RC File, Unix Shell Scripting, Tableau.

**Verity Knowledge Solutions Pvt Ltd, Hyd India June 2014 – Oct 2017**

**SQL Developer/ SSIS Developer (Data Engineer)**

**Responsibilities:**

* Designed the Functional Requirements and Mapping Technical Specifications based on Business Requirements.
* Creating Database objects – Tables, Views, and Functions and developing procedures for implementing application functionality at the database side for performance improvement.
* Performing query plans and making sure each and every query is using appropriate useful Indexes.
* Creating and Modifying T-SQL Stored Procedures/Triggers for validating the integrity and used Performance Monitor and SQL Profiler for Monitoring memory, processor, SQL Queries etc., rebuilding Indexes and Tables as part of performance tuning exercise.
* Deployed the created SSIS packages in Development and Testing environments, wrote/utilized T-SQL objects Complex joins, correlated sub queries, CTEs to match (data matching) current and historical data for reporting.
* Provided source to target mappings to perform initial, full and incremental loads into the target data mart.
* Studied the existing OLTP system(s) and Created facts, dimensions and star schema representation for the data mart and OLAP model based on Dimension and FACTS for efficient loads of data based on Star Schema structure on levels of reports.
* Converted the data mart from Logical design to Physical design, defined data types, Constraints, Indexes, generated Schema in the Database.
* Analyzed data and created new data match columns/keys to identify and match data from different departments to reconcile cash transactions. Improved the reconciliation process.
* Utilized many system functions and user functions to cleanse and transform data as per the ad-hoc reporting needs.
* Built packages and handled slowly changing dimensions to maintain the historical data and modified existing SSIS packages to meet the change requests from the users.
* Used SSIS to create ETL Packages to validate, extract, transform and load data to data warehouse databases, data mart databases.
* Generated packages using different Transformations like Lookups, Derived Column, Merge Join, Fuzzy Lookup, For Loop, For Each Loop, Conditional Split, Union all, Script component etc.
* Experience in providing Logging, Error handling by using Event Handler, and Custom Logging for SSIS Packages.

**Environment:** SQL Server 2012/2008 R2, Tableau Desktop, SSAS, SSIS, SSRS, T-SQL, TFS, Epic Clarity, Lawson.

**Education:**

# University of North Texas at Denton, Texas Master’s in Business Analytics-GPA 3.8/4.0 May 2022

# (Awarded ITDS Outstanding New Master's Student Scholarship)

**Course Work:** Foundations of Database Systems, Data warehousing, Predictive Analytics, Big Data Analytics, Data Mining, Information Systems Development

# B.E. in Electronics &Communication Eng, India. GPA: 3.5/4.0 April 2012

# Course Work: Computer Prog. & Numerical Methods, Network Theory, Material Science, Electronic Devices & Circuits, Computer Architecture & Organization, Analog Communications, Digital Communication & Signal Processing, Computer Networks Engineering, Data Communications.