**GCP Data Engineer**

**VENKATESH K**

**Email: Kandikatla05@gmail.com**

**Mobile: +1(469)-364-7949**

**PROFESSIONAL SUMMARY:**

* Around 10+ years of IT experience in Analysis, Design, Development, Implementation and Testing of Data warehousing applications using Data Modeling, Data Engineering, Data Extraction, Data Transformation, Data Loading, Data Analysis and Performance Tuning Techniques

worked extensively on AWS services like EC2, S3, EMR, Sage Maker, RDS (Aurora), Redshift, DynamoDB, and Elasticache (Memcached & Redis) & Quick Sight, Athena, Glue and other services of the AWS family.

* Experienced in Designing, Developing, Documenting, and Testing ETL jobs and mappings in Server and Parallel jobs using DataStage to populate tables in Data Warehouse and Data marts.
* Hands on experience with databases including Oracle, PL/SQL and MySQL involving stored procedures, triggers, functions, indexes, and packages.
* Experience with cloud technologies like AWS Step Functions, AWS Lambda, Quick Sight, Cloud Watch, Glue, Athena, Redshift, IAM, EMR, IAM and SNS.
* Experience in development of Big Data projects using Hadoop, Hive, HDP, Pig, Flume, Storm and MapReduce open-source tools.
* Experience in developing CI/CD (continuous integration and continuous deployment) and automation using Git and Kubernetes
* Worked in Insurance, HealthCare, and Banking & Social networking domains.
* Extensively used Python Pytest, pyodbc, NumPy, MySQL dB, sqlite3, snowflake-python-connector, and other packages.
* Experience in installation, configuration, supporting and managing Hadoop clusters.
* Experience in working with MapReduce programs using Apache Airflow Hadoop for working with Big Data.
* Experience in installation, configuration, supporting and monitoring Hadoop clusters using Apache Airflow, Cloudera distributions and AWS.
* Having experience creating Tableau dashboards to display insights from standard text- based data formats including JSON, CSV, and XML.
* Extensive experience in Hadoop development of enterprise level solutions utilizing Hadoop components such as Apache Airflow Spark, MapReduce, HDFS, Sqoop, PIG, Hive, HBase, Oozie, Flume, Nifi, Kafka, Zookeeper, and YARN.
* Experienced working on Big-Data technologies like Hadoop (MapReduce & Hive), Sqoop, HDFS and Spark streaming Kafka, and No-SQL Databases like MongoDB, Cassandra.
* Strong Knowledge of the Architecture of Spark distributed and parallel processing using Spark SQL, Spark Data frames APIs, and Spark execution framework and Py Spark
* Proficient in Data Modeling (Relational, DW), Data Migration (ETL using Data stage), shell scripting, Data Reporting.
* Experience in normalization/denormalization, data extraction, data analysis, data cleansing, data profiling, data manipulation, distributed data processing & slowly changing dimensions techniques.
* Worked on different versions of DataStage and experience in developing DataStage Enterprise Parallel jobs, Server jobs and sequence jobs.
* Have experience in following Waterfall and Agile models with the help of software like JIRA.
* Experience in Designing, Compiling, Testing, Scheduling, and running DataStage jobs.
* Designed Parallel transformation jobs for applying the business rules provided.
* Designed Sequencer Jobs to run the Parallel transformation/load jobs in a sequence order.
* Designed Runtime Column Propagation (RCP) DataStage jobs using schema file for loading multiple flat files to corresponding target database stages
* Good Knowledge in Scheduling DataStage Jobs through Tidal/Control-M/PeopleSoft process scheduler.
* Hands-on experience in Filezilla, Putty, WinSCP etc.
* Skilled in writing Technical Design Document, Unit test cases.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **ETL/ELT** | Informatica Power Center, Azure Data Factory |
| **Databases** | AWS S3, Azure Blob Storage, Azure Datalake gen1 gen2, ADLS, HDFS, MongoDB, AWS Redshift, Snowflake |
| **Languages** | SQL, PL/SQL, SPARK SQL, Shell Scripting, Python (Pandas, NumPy, Scikit Learn, OOPS, Functional Programming), SCALA, Advanced Excel, Java |
| **Cloud Technology** | GCP: Compute Engine, Google Kubernetes Engine (GKE), Cloud Datastore, Cloud Functions,Cloud Pub/Sub |
| **Tools & Utilities** | Atlassian Jira, File Zilla, WinSCP, MS-Visio, MS Office, Lucid Chart |
| **IDE & Utilities** | SQL Developer, Toad, Dai Query using Hive, Visual Studio, CI/CD, Jupyter Notebook, PyCharm, Databricks, Zeppelin |
| **Project Management & Versioning Tools** | Jira, Tortoise SVN, Bitbucket, Jenkins, Source Tree, RTC Migration, Star Team, GIT, GITHUB, |
| **SDLC** | Waterfall, Agile (Scrum) |
| **Scheduling Tool** | Autosys, Tivoli, Control-M, Tidal, PeopleSoft process scheduler. |
| **Operating Systems** | Windows, Linux, UNIX |

**Professional experience:**

**Client: Comcast, PA| OCT 2023- Present**

**Roles: GCP Data engineer**

**Responsibilities:**

* Engineered and mastered Informatica platforms for Cloud Services, Big Data Management, Master Data Management, Data Integration and Data Quality.
* Wrote python scripts to parse XML documents and load the data in database and developed web - based applications using Python, CSS and HTML.
* Utilize Ansible to automate the provisioning, configuration, and management of infrastructure components across Google Cloud Platform (GCP), including compute instances, networks, and storage resources.
* Define and enforce consistent configurations across GCP environments using Ansible playbooks, roles, and modules, ensuring scalability, reliability, and security of infrastructure deployments.
* Integrate Ansible with GCP services such as Google Compute Engine (GCE), Google Kubernetes Engine (GKE), Cloud Storage, Cloud SQL, and Pub/Sub for streamlined management and orchestration of cloud resources.
* Implement Ansible within CI/CD pipelines (e.g., Jenkins, GitLab CI/CD) to automate testing, deployment, and rollback processes for data engineering solutions on GCP.
* Worked on applications and developed them with XML, JSON, XSL (PHP, Django, Python and Rails).
* Developed an e-commerce platform with a Java-based backend for handling orders, payments, and user authentication. Implemented a microservices architecture for a fintech application, using Java.
* Developed a scalable data warehouse solution on GCP using BigQuery, enabling efficient querying and analysis of petabyte-scale datasets.
* To ensure successful creation of the infrastructure required for optimal extraction, transformation, and loading of data from a wide variety of data sources using Spark, SQL, HDFS, Hive, MapReduce, Druid, Python, Unix, Hue and Shell Scripting.
* Experienced Learner data model which gets the data from Kafka in real time and persist it to Cassandra.
* Developed Kafka consumer API in python for consuming data from Kafka topics.
* Consumed Extensible Mark-up Language (XML) messages using Kafka and processed the xml file using Spark Streaming to capture User Interface (UI) updates.
* Design, implement, and maintain infrastructure components using Terraform to ensure scalability, security, and reliability on Google Cloud Platform (GCP).
* Provision and manage compute resources such as virtual machines (Compute Engine), Kubernetes clusters (Google Kubernetes Engine - GKE), and managed instance groups to support data processing and analytics workloads.
* Configure and optimize network resources (VPC, VPN, CDN) to ensure secure and efficient data transmission and access.
* Used Apache Airflow Kafka to aggregate web log data from multiple servers and make them available in downstream systems for Data analysis and engineering type of roles.
* Worked in Implementing Kafka Security and boosting its performance.
* ister and optimize data storage solutions, including Google Cloud Storage (GCS) for object storage, Cloud SQL for relational databases, Bigtable for NoSQL databases, and Firestore for scalable document databases.
* Implement data lifecycle management strategies, including backup, replication, and archival policies to ensure data availability, integrity, and compliance.
* Use Helm charts to package, version, and deploy Kubernetes applications and microservices on Google Kubernetes Engine (GKE) and other Kubernetes clusters on GCP.
* Define and manage configuration settings, including environment-specific configurations, using Helm values files to streamline application deployments and updates.
* Collaborate with DevOps and development teams to orchestrate infrastructure deployments using Helm, ensuring consistency and reliability across different environments (e.g., dev, test, prod).
* Implement and manage Helm releases, including rollbacks and upgrades, to maintain application availability and performance while adhering to release management best practices.
* Define and manage GCP resources such as virtual machines, storage buckets, networks, and IAM policies using Terraform configurations.
* Develop Terraform modules and templates to automate deployment and configuration tasks, promoting consistency and repeatability across environments (dev, test, prod).
* Created a real-time chat application with Java for the server-side components using WebSockets and Spring Boot.
* Implemented a real-time data processing pipeline with GCP Pub/Sub and Dataflow, achieving low-latency data ingestion and transformation.
* Architected a microservices-based application on GCP, utilizing Google Kubernetes Engine (GKE) for orchestration and management of containerized services.
* Designed a cost-effective backup and disaster recovery strategy using GCP Cloud Storage and Cloud SQL, ensuring data availability and resilience.
* Migrated on-premises infrastructure to GCP, leveraging Compute Engine and Cloud VPN to seamlessly integrate with existing systems.
* Implement algorithms to anonymize PHI within healthcare datasets while preserving data utility for analysis.
* Developed a machine learning model on GCP AI Platform, using AutoML to automate model training and hyperparameter tuning.
* Implemented an end-to-end CI/CD pipeline on GCP using Cloud Build, Cloud Source Repositories, and Cloud Run for continuous integration and deployment.
* Designed a data lake architecture on GCP using Cloud Storage and BigQuery, enabling unified data storage and analytics.
* Maintain Docker images and Docker files to standardize development and production environments, ensuring reproducibility and minimizing configuration drift in data workflows.
* Optimize Docker container configurations and resource allocation (e.g., CPU, memory) for data-intensive workloads to achieve optimal performance and cost efficiency on GCP.
* Implement Docker security best practices (e.g., image scanning, secrets management) and ensure compliance with data protection regulations (e.g., GDPR, CCPA) in containerized environments.
* Utilize Docker logging mechanisms and integrate with monitoring tools (e.g., Prometheus, Stackdriver) to monitor containerized applications and troubleshoot performance issues on GCP.
* Worked closely with the SME to get an understanding of the business requirements.
* Built an IoT data processing pipeline with Python for data collection from sensors and Java for backend processing and storage.
* Implement CI/CD pipelines using Google Cloud Build and Kubernetes Engine for automated deployment and scaling of data processing and analytics workflows.
* Developed a financial trading platform using Java for the trading engine and transaction management.
* Configured and managed a secure GCP environment using Identity and Access Management (IAM) policies and VPC Service Controls.
* Used PySpark and Spark to implement data quality checks, data transformation, and data validation processes.
* Implemented a real-time analytics platform using GCP BigQuery and Data Studio, providing interactive dashboards and reports for business insights.
* Developed serverless applications on GCP using Cloud Functions and Cloud Firestore, reducing infrastructure management overhead.
* Implemented a data governance framework on GCP using Data Catalog and Cloud DLP, ensuring data compliance and security.
* Python-based analytics module to predict patient no-shows and optimize scheduling.
* Performed data quality checks to ensure high quality of data.
* Working on data validation and data profiling to ensure the accuracy of the data between the warehouse and source systems.

**Environment:** Python, CSS and HTML, XML, JSON, X, MySQL, Postgre SQL, EC2, SQS, SNS, IAM, S3, and DynamoDB Data-Integration, Data Mapping, Data Profiling, Data Lake, AWS Athena, PySpark, Spark

**Client: 7-Eleven, Dallas, Tx | Aug 2022- Sept 2023**

**Role: Azure Data Engineer**

**Responsibilities:**

* Involved in planning and developing roadmaps and deliverables to advance the migration of existing SSIS on-premises systems/applications to Azure cloud.
* Used Azure architecture decision-making Architect and implemented ETL and data movement solutions using Azure Data Factory (ADF), and SSIS.
* Design and implement database solutions in Azure Synapse (SQL Data Warehouse), and Azure SQL.
* Implemented ad-hoc analysis solutions using Azure Data Lake Analytics and HDInsight.
* Architect and implemented ETL and data movement solutions using Azure Data Factory, SSIS create and run SSIS Package ADF V2 Azure-SSIS IR.
* Configuring SQL Azure firewall for a security mechanism.
* Created Azure SQL database, performed monitoring and restoring of Azure SQL database.
* Performed migration of Microsoft SQL server to Azure SQL database.
* Managed Azure Data Lakes (ADLS) and Data Lake Analytics to integrate with other Azure Services.
* Used various sources to pull data into Power BI such as SQL Server, Excel, SQL Azure, etc.
* Created POWER BI Visualizations and Dashboards as per the requirements.
* Utilizing Cosmos DB (NoSQL) (AWS equivalent = DynamoDB) for simple data entry, manipulation, and transformation; scaling and partitioning for growth.
* Azure cloud migration is the process of understanding the Azure cloud platform to deliver data needs and provide PAAS solutions using storage and Spark ecosystems.
* Configured AWS Cloud architecture (configured IAM user and group policies and S3 bucket storage for the company and backup/migration of company website to AWS with web hosting on EC2. Training / support provided for AWS certification preparation)
* Develop Azure Data Factory pipelines for Data Flow Orchestration across Azure Data Lake zones
* Configuring and managing company AWS cloud system using services such as IAM and S3; assisting Alliance Lead in getting company Amazon Partner Network (APN) certified
* Web development of company website (and joint venture partnership websites) in line with company business development and marketing plans; planned migration from WordPress to Drupal Executed Power Shell Scripts to move Data from Lower to Higher Environment.

**Environment:** Hadoop, MapReduce, HDFS, Pig, Hive, PySpark, Kafka, ADF, YARN, Scala, SQL, Git, Azure, AWS

**Client: AGL -INDIA | OCT 2020 – FEB 2022**

**Role: Big Data Engineer**

**Responsibilities:**

* Extracted data from various sources such as databases, APIs, and flafiles.
* Transformed the data to fit the structure and format of the target repository.
* Loaded and transformed large sets of structured, semi structured, and unstructured data using Hadoop/Big Data concepts.
* Document Terraform code, infrastructure architecture, and deployment processes to facilitate knowledge transfer and support ongoing maintenance and enhancements.
* Stay updated with Terraform releases and GCP advancements to incorporate new features and best practices into infrastructure automation workflows.
* Worked from Scratch in Configurations of Kafka such as Managers and Brokers.
* Used Apache Airflow Kafka to aggregate web log data from multiple servers and make them available in downstream systems for Data analysis and engineering type of roles.
* Worked in Implementing Kafka Security and boosting its performance.
* Document Ansible playbooks, automation workflows, and infrastructure configurations to facilitate collaboration, knowledge transfer, and support for DevOps and development teams.
* Use Ansible for troubleshooting infrastructure issues, performing configuration audits, and implementing remediation actions to ensure the stability and availability of data engineering solutions on GCP.
* Stay updated with Ansible releases, GCP updates, and industry best practices to enhance automation capabilities, optimize workflows, and drive efficiency in infrastructure management.
* Loaded the data into the target repository, such as a Data warehouse or Data Lake.
* Ensure compliance with healthcare regulations (e.g., HIPAA) during PHI anonymization and vector storage.
* Implemented a log analysis solution on GCP using Cloud Logging and BigQuery, enabling real-time monitoring and troubleshooting.
* Utilize AI APIs (e.g., Vision, Natural Language Processing) to integrate AI capabilities into data-driven applications and solutions.
* Document Docker configurations, Dockerfiles, and deployment processes to facilitate knowledge transfer, collaboration, and support for development and operations teams.
* Stay updated with Docker releases, best practices, and GCP advancements to enhance containerization strategies, deployment automation, and infrastructure reliability.
* Document Helm charts, deployment processes, and infrastructure configurations to facilitate collaboration, knowledge transfer, and support for DevOps and development teams.
* Stay updated with Helm releases, Kubernetes best practices, and GCP advancements to enhance application deployment automation, scalability, and reliability.
* Use vector databases to quickly retrieve and analyze transaction vectors for suspicious activity detection.
* Ensured data integrity and accuracy throughout the ETL process.
* Created and maintained documentation for ETL processes.
* Store patient records securely in Snowflake's encrypted data warehouse.
* Developed a scalable IoT solution on GCP using Cloud IoT Core and BigQuery, enabling real-time data collection and analysis from connected devices.
* Installed and Configured Apache Airflow Hadoop clusters for application development and Hadoop tools.
* Monitored and trouble shooted ETL processes to ensure they run efficiently and without errors
* Configured and managed a secure networking environment on GCP using VPC, Cloud Armor, and Cloud NAT, ensuring robust security and performance.
* Used Spark API over Hortonworks Hadoop YARN to perform analytics on data in Hive.
* Implemented a multi-cloud strategy using GCP Interconnect and Cloud VPN, enabling seamless connectivity between GCP and other cloud providers.
* Designed and implemented a serverless data pipeline on GCP using Cloud Dataflow and Cloud Storage, automating data processing and analytics workflows.
* Built a secure voting system using Java for the blockchain backend to ensure integrity and transparency, and Python for the user interface.
* Utilize Google Cloud Bigtable or Cloud Spanner for structured data storage and schema management.
* Transformed and cleaned data using Py Spark libraries such as pyspark.sql.functions and pyspark.sql.types to fit the structure and format of the target repository
* Created and maintained PySpark scripts and jobs that automate the ETL process
* Optimizing PySpark jobs for performance and scalability
* Monitored and troubleshooter PySpark jobs to ensure they run efficiently and without errors.
* using Python for device control and data collection from sensors, and Java for backend services to manage user profiles, schedules, and remote access.
* Continuously learned and stayed up to date with new PySpark libraries and features to improve the ETL process
* Familiar with Big Data Eco-systems and distributed computing concepts like HDFS, YARN, and Mesos.

**Environment:** Hadoop/Big Data, Data warehouse, Data Lake, ETL, pyspark, Sql Functions, HDFS, YARN, Mesos.

**Client: Schlumberger, INDIA | JUN 2017– Sept 2020**

**Role: Hadoop/Big Data Engineer**

**Responsibilities:**

* Used Hive Queries in Spark-SQL for analysis and processing the data.
* Responsible for handling different data formats like Avro, Parquet and ORC formats
* Worked on Import & Export of data using ETL tool Sqoop from MySQL to HDFS using Teradata studio and DBeaver
* Hands on experience in installation, configuration, supporting and managing Hadoop Clusters
* Implemented Optimized Map Joins to get data from different sources to perform cleaning operations before applying the algorithms.
* Used Spark API over Cloudera Hadoop YARN to perform analytics on data in Hive.
* Enhanced and optimized product Spark code to aggregate, group and run data mining tasks using the Spark framework and handled Json Data
* Implemented Spark Scripts using Scala, Spark SQL to access hive tables into Spark for faster processing of data.
* Involved in Developing a Restful service using Python Flask framework.
* Used Python modules such as requests, urllib, urllib2 for web crawling.
* Experienced in managing and reviewing Hadoop log files.
* Involved in business analysis and technical design sessions with business and technical staff to develop requirements document and ETL design specifications.
* Use Snowflake's SQL capabilities to transform raw market data into actionable insights.
* Perform complex queries and aggregations on financial data using Snowflake's built-in functions and SQL.
* Visualize financial analytics using Snowflake's integration with BI tools like Tableau or Looker.
* Snowflake's elastic scaling allows handling large volumes of financial data and concurrent user queries effectively.
* Created and monitored sessions using workflow manager and workflow monitor.
* Involved in loading data from UNIX file system to HDFS.
* Responsible for design & development of Spark SQL Scripts based on Functional Specifications
* Design and develop extract, transform, and load (ETL) mappings, procedures, and schedules, following the standard development lifecycle.
* Defined job flows and developed simple to complex Map Reduce jobs as per the requirement. Optimized Map/Reduce Jobs to use HDFS efficiently by using various compression mechanisms.
* Worked on Informatica Source Analyzer, Mapping Designer & Mapplet, and Transformations
* Developed end to end ETL batch and streaming data integration into Hadoop (MapR), transforming data.
* Created highly optimized SQL queries for MapReduce jobs, seamlessly matching the query to the appropriate Hive +table configuration to generate efficient report.
* Work closely with Quality Assurance, Operations and Production support group to devise the test plans, answer questions and solve any data or processing issues.
* Worked on large-scale Hadoop YARN cluster for distributed data processing and analysis using Data Bricks Connectors, Spark core, Spark SQL, Sqoop, Hive and NoSQL databases.
* Worked in writing Spark Sql scripts for optimizing the query performance.
* Concerned and well-informed on Hadoop Components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node, YARN and Map Reduce programming.
* Implemented Hive UDF's and did performance tuning for better results.
* Tuned, and developed SQL on HiveQL, Drill and Spark SQL.
* Experience in using Sqoop to import and export the data from Oracle DB into HDFS and HIVE
* Developed Spark code using Spark RDD and Spark-SQL/Streaming for faster processing of data.
* Implemented Partitioning, Dynamic Partitions and Buckets in HIVE for efficient data access.

**Environment:** Cloudera, Horton Works distribution, HDFS, Spark, Hive, Map Reduce, Python, Agile Methodology, Hue, Sqoop, Putty, HaaS (Hadoop as a Service), Spark, SPARK SQL, Maven, Java, Scala, SQL and Linux, Toad, YARN, Agile Methodology, MongoDB, ETL, Codecloud.

**Client: Royal Dutch shell, Hyderabad, India | June 2014 – May 2017**

**Role: ETL/Informatica, PL/SQL developer**

**Responsibilities:**

* Designed, developed and maintained BI solutions using SQL, including data warehouse and data mart structures.
* Created and maintained SQL-based ETL (Extract, Transform, Load) processes to extract, clean and load data into the data warehouse.
* Created and maintained SQL-based reports and analytics using tools like SSRS (SQL Server Reporting Services), Power BI and Tableau
* Collaborated with stakeholders and other teams to understand business requirements and design BI solutions that meet their needs.
* Extracted, loaded data from a JSON file into PostgreSQL using python pandas library.
* Developed SSIS packages to extract, Transform and Load (ETL) data into the data warehouse database from heterogeneous databases/data sources.
* Create new SSIS Packages to export flat files.
* Uses different data Transformations like Lookup, Derived Column, Conditional Split, data conversation while creating the SSIS packages.
* Write Stored procedures, Triggers, Functions and T-SQL Queries to capture updated and deleted data.
* Created Ad-Hoc Reports, Sub Reports and summary reports using SSRS.
* Export/import data from CSV files, Text files and Excel Spreadsheets by creating SSIS Packages.
* Experience in reporting services based on existing XML documents.
* Scheduled Jobs and alerts using SQL Server Agent.
* Use Selenium Grid to execute tests across multiple browsers (Chrome, Firefox, Safari) to ensure compatibility.
* Integrate Selenium tests with continuous integration tools like Jenkins for automated builds and deployments.
* Creating, maintaining the users, roles and granting privileges.
* Fine tuning the Database Settings as well as the server Settings.
* Created data models and designed database schemas to support reporting and analytics.
* Optimized SQL queries for performance and scalability
* Trouble shooted and debugged BI issues, identifying, and resolving data and performance issues.
* Worked with other BI developers and IT teams to design, develop and implement security, backup and recovery procedures.
* Continuously improved the BI development process by researching and experimenting with new tools and technologies.
* Provided technical guidance and mentorship to other team members, and act as a subject matter expert on BI development using SQL.

**Environment:** BI, SQL, SQL-based ETL, SSRS, Power BI and Tableau, pandas library, SSIS Packages, T-SQL Queries, SSRS, CSV files, Text files and Excel Spreadsheets, SQL Server Agent, Database Settings, SQL queries.