**Surya Mithra Reddy Ram**

**Senior Data Engineer**

**Contact :** ( 832)-786-8694

**Email:** rsmithrar@gmail.com

**PROFESSIONAL SUMMARY**

* GCP Architect with 12+ years of cloud experience, including leading enterprise migrations from on-prem to GCP. Skilled in secure, cloud-native design using Terraform, GKE, App Engine, and Cloud Functions.
* Proficient in programming languages such as Python, PySpark, SQL, and Scala, enabling complex data transformations and real-time data processing.
* Expert in designing cloud infrastructure on platforms like AWS and GCP, leveraging services such as EC2, S3, RDS, BigQuery, and Cloud Pub/Sub.
* Specialized in architecting secure, scalable GCP infrastructure using Compute Engine, GKE, and Cloud Storage, ensuring enterprise-grade reliability and compliance.
* Extensive experience with data processing frameworks and tools like Spark, Dataproc, Google Cloud Dataflow, AWS EMR, and AWS Lambda, creating scalable and efficient data pipelines.
* Skilled in data warehousing techniques such as partitioning, clustering, and denormalization to optimize storage and query performance.
* Skilled in stakeholder engagement and architecture review boards, translating business needs into scalable GCP solutions across compute, storage, IAM, and network.
* Proven ability to lead consulting engagements, define cloud strategy, and mentor teams. Experienced in DevSecOps practices, Agile delivery, and aligning cloud solutions with business goals
* Led client-facing GCP architecture consulting engagements, assessing on-prem infrastructure and creating cloud-native migration strategies using Terraform, GKE, and App Engine.
* Proficient in ETL tools such as Informatica PowerCenter, Talend, AWS Glue, and SAP BODS to streamline data integration and transformation.
* Expertise in workflow orchestration tools like Apache Airflow and Google Cloud Composer to automate and monitor data pipelines for peak performance.
* Applied machine learning techniques for building classification and recommendation models using Python, supporting real-time, data-driven decision-making.
* Developed interactive dashboards and reports using tools like Power BI, Google Data Studio, and Looker Studio to drive informed decision-making for stakeholders.
* Extensive experience with monitoring tools like AWS CloudWatch, AWS CloudTrail, and Google Stackdriver, ensuring real-time logging and pipeline reliability.
* Hands-on experience with version control systems, including Git, for effective code management, collaboration, and CI/CD integration.
* Skilled in building CI/CD pipelines using tools like Jenkins, Terraform, and Google Cloud Build for seamless deployment of production-ready data pipelines.
* Proficient in real-time data ingestion and processing using tools such as Google Pub/Sub, SLT, and HDFS, ensuring timely data availability for analytics and reporting.
* Familiar with regulatory compliance frameworks such as PHI, HL7, FHIR, and HIPAA, ensuring secure and compliant data handling.
* Adept at Agile methodologies, collaborating with cross-functional teams, and using tools like Jira for task management and sprint planning to deliver high-quality data solutions.

**CERTIFICATIONS**

* Certified Google Professional Data Engineer
* Certified AWS Solutions Architect Associate

**EDUCATION**

* Bachelor of Technology in Computer Science and Engineering, JNTU Anantapur, India

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Cloud Platforms** | GCP, AWS, Azure, Snowflake, Databricks |
| **Programming Languages** | Python, PySpark, Scala, SQL, R, Bash, Shell. |
| **Data Processing** | BigQuery, S3, Spark, Apache Kafka, Dataproc, Google Cloud Dataflow, EMR, Lambda, Databricks, Hadoop, Hive, MapReduce, Talend, AWS Glue, BODS. |
| **Data Warehouses** | Google BigQuery, Cloud Spanner, Redshift, Teradata, Oracle, SAP HANA, SAP ASE, MySQL, Amazon RDS. |
| **ETL and Tools** | Informatica PowerCenter, Pentaho, SSIS, SSRS, BODS, and SLT Replication. |
| **Data Visualization** | Power BI, Advanced Excel, Google Data Studio, Looker Studio. |
| **Workflow Orchestration** | Apache Airflow, Google Cloud Composer, AWS Step Functions, and Terraform. |
| **Version Control** | GitHub. |
| **CI/CD Tools** | Jenkins, Cloud Build |

**PROFESSIONAL EXPERIENCE**

**Client: Tennessee Farmers Insurance Company, Nashville, TN Mar 2024 – Present**

**Role: Senior Data Engineer**

 **Responsibilities:**

* Executed seamless data migration from Oracle and SAP ASE databases to BigQuery by staging in Google Cloud Storage (GCS) via BODS, aligning with ECC structure requirements.
* Designed hybrid cloud solutions to optimize batch and real-time ingestion, leveraging Cloud Dataflow for Google Cloud and AWS Glue for AWS to standardize S/4 HANA and ECC structures with advanced normalization techniques.
* Engineered and managed data ingestion pipelines using BODS and SLT, deployed via Terraform, to efficiently move data from Oracle, SAP ASE, and SAP HANA into Google Cloud.
* Automated pipelines for optimized real-time and batch data integration by utilizing Cloud Dataflow, Dataproc, and Apache Airflow.
* Developed cost-effective, multi-environment GCP infrastructure using Terraform and network blueprints, ensuring compliance, modularity, and rapid deployment across dev, test, and prod tiers
* Designed GKE clusters with node pools, autoscaling, and custom workload identity for efficient, secure microservices hosting and orchestration.
* Led cloud migration from Oracle and SAP ASE to BigQuery using BODS and Dataflow, ensuring zero downtime and full audit traceability across batch and real-time pipelines.
* Defined GCP application migration strategy for SAP-based on-prem workloads, evaluating rehosting vs. replatforming options and aligning target architecture with business availability requirements
* Applied Spark and Python to clean and transform historical and real-time datasets, ensuring high data accuracy and readiness for analytics in both BigQuery and AWS Redshift.
* Automated workflows using Airflow and shell scripting to significantly reduce processing times for daily data integrations.
* Implemented end-to-end GCP network design including VPC, subnets, and firewall rules to support secure multi-region deployments for enterprise workloads.
* Built GCP Dataflow and Pub/Sub pipelines to process IoT sensor and transactional data in real time, supporting alerts, fraud detection, and KPI dashboards.
* Architected CI/CD frameworks integrating Cloud Build, Terraform, and Git for secure, automated deployment of infrastructure and data solutions.
* Defined IAM strategy and implemented org policies, VPC Service Controls, and KMS encryption to enforce enterprise security standards on GCP.
* Designed and maintained infrastructure using Terraform to provision GCP resources such as BigQuery, Dataflow, and Pub/Sub, enabling reproducible environments and streamlined deployments.
* Wrote custom stored procedures and transformations in Python, Spark, and Scala for handling complex data cleaning and migration needs.
* Enhanced query performance in BigQuery with advanced partitioning, clustering, and denormalization techniques, complemented by automation through shell scripts.
* Participated in architecture review boards to validate VPC design, IAM hierarchy, and cluster sizing for scalable GKE workloads hosting batch and streaming pipelines.
* Monitored pipelines in BigQuery, Dataproc, and Airflow environments using Cloud Monitoring, ensuring stable operations and issue resolution.
* Managed version-controlled data pipelines and collaborated across Agile teams using GitHub and Jira to ensure sprint-based delivery of projects.

**Tech Stack:** Apache Spark, Google Cloud Pub/Sub, Google Cloud Storage, Google Dataflow, Cloud Functions, Airflow, PySpark, BigQuery, Python, Stack driver, Firestore, MySQL, Google Data Studio, Big Query ML, Jira, Agile, Cloud Build, Terraform.

**Client: Starbucks, Seattle, WA Aug 2023 – Feb 2024**

**Role: GCP Data Engineer**

**Responsibilities:**

* Engineered robust ELT pipelines integrating FiveTran, Composer, and GCS, following Google’s dataflow design patterns to improve ingestion resilience and auditability.
* Captured real-time IoT data via Google Pub/Sub and automated pipelines for timely ingestion and processing with Airflow orchestration.
* Designed staging processes in GCS for real-time data, applying scalable transformations using Dataproc and Python.
* Utilized Docker containers to package and deploy Spark and Python-based data jobs, orchestrated on GKE for scalable, isolated, and consistent runtime environments across dev and prod.
* Utilized Docker containers to package and deploy Spark and Python-based data jobs, orchestrated on GKE for scalable, isolated, and consistent runtime environments across dev and prod.
* Developed Dataflow pipelines to process real-time data into BigQuery and Spanner, employing advanced partitioning and clustering to enhance analytics capabilities.
* Orchestrated ETL pipelines across platforms using Google Cloud Composer, integrating data with Python and Spark transformations.
* Containerized ML workloads and microservices with Docker and deployed to Istio-enabled GKE clusters with ingress controls, autoscaling, and service discovery
* Built reusable Terraform modules for provisioning GCP services, enabling consistent environment creation and reducing provisioning time across multiple projects.
* Implemented end-to-end GCP network design including VPC, subnets, and firewall rules to support secure multi-region deployments for enterprise workloads.
* Wrote advanced SQL queries in BigQuery involving window functions, CTEs, and nested queries to enable detailed analytics and optimize business reporting workflows.
* Built event-driven pipelines using Google Cloud Functions to trigger data workflows upon object uploads or metadata changes, ensuring dynamic and responsive processing.
* Designed modular ELT workflows using FiveTran, Cloud Composer, and GCS to automate ingestion and transformation of structured and semi-structured data from multiple POS and telemetry sources.
* Integrated GCP Cloud Monitoring and Logging for proactive alerting and visualization of system metrics, enhancing system uptime and incident response
* Implemented IAM policies with the principle of least privilege, hardened GCP resources based on NIST and CIS benchmarks for enhanced cloud security posture.
* Automated ETL pipelines with Apache Airflow, enhanced scheduling with cron jobs, and monitored workflows with comprehensive logging mechanisms.
* Created Python DAGs in Airflow to orchestrate end-to-end data pipelines and integrated real-time datasets via Dataproc.
* Used Cloud Composer to orchestrate cross-platform ETL pipelines, integrating Dataflow and BigQuery tasks with automated retries, alerts, and dependency tracking.
* Streamlined CI/CD processes using Jenkins, ensuring automated deployment of data pipelines to production.

Delivered App Engine PoCs for cloud-native CRM components, enabling real-time customer analytics and seamless scaling with minimal infrastructure ops

**Tech Stack:** Snowflake, Google Pub/Sub, Google Cloud Storage, Google Cloud Composer, Dataflow, Dataproc, BigQuery, SQL, Python, Machine Learning, Ansible, Data Governance and IAM, Apache Airflow, Cron Jobs, Python DAGs, Git, Jenkins, Apache Airflow, CI/CD Pipelines, Agile Methodology.

**Client: CVS, Woonsocket, RI Dec 2021-Aug 2022**

**Role: GCP Data Engineer**

**Responsibilities:**

* Migrated 30+ TB of healthcare data from Teradata to BigQuery using Talend and Dataflow, ensuring HIPAA and HL7 compliance with detailed lineage tracking.
* Maintained architecture diagrams and runbooks for GCP deployments, enabling clear knowledge transfer and streamlined onboarding for DevOps and SRE teams.
* Designed data mirroring techniques, staging PHI-compliant healthcare data securely in Google Cloud Storage.
* Led infrastructure automation using modular Terraform scripts, provisioning scalable GCP network architecture, IAM policies, and Cloud Storage buckets.
* Developed Python and Spark scripts on Dataproc and Cloud Functions to transform healthcare data for analytics in BigQuery.
* Architected real-time ingestion strategy using Pub/Sub, Dataflow, and BigQuery, enabling low-latency CRM analytics for marketing teams with automated SLA monitoring via Cloud Monitoring
* Designed CI/CD pipelines using GitHub Actions and GCP Cloud Build to automate deployments across staging and production environments with rollback strategies.
* Created and maintained table schemas in BigQuery, ensuring seamless integration with downstream systems.
* Designed GCS-based staging layers for both structured and unstructured data, optimizing lifecycle rules and access policies for efficient data lake operations.
* Captured and processed real-time healthcare data via APIs, storing it in efficient formats for scalability and analytics.
* Explored Ansible for lightweight infrastructure automation and configuration management, complementing Terraform for end-to-end provisioning and deployment workflows.
* Created reusable Terraform modules for PHI-compliant cloud landing zones, empowering teams to self-serve new projects within HIPAA-aligned boundaries.
* Configured and scaled GKE clusters for high-availability microservices using Istio for service mesh and traffic management in production environments.
* Built ELT pipelines with audit tracking to monitor and log data migrations and transformations for compliance purposes.
* Automated batch processing with Cloud Composer, Dataflow, Dataproc, and Cloud Functions to handle large volumes of healthcare data.
* Monitored pipelines using Cloud Logging and Cloud Monitoring, ensuring real-time tracking of healthcare data workflows.
* Managed metadata using Cloud SQL for tracking schema versions, data sources, and transformation logic.

**Tech Stack:** Talend, Google Cloud Dataflow, Google BigQuery, Teradata, Oracle, HL7 & FHIR Standards, Google Cloud Storage (GCS), PHI Compliance, Python, Data Governance and IAM, Apache Spark, Google Cloud Dataproc, Google Cloud Functions, Cloud Data Fusion, Ansible, Salesforce CRM, Cloud Composer, Cloud SQL, Git, Jira, Agile Methodology.

**Client: Nextera, Juno Beach, FL Nov 2018 – Aug 2021**

**Role: Data Engineer**

**Responsibilities:**

* Built a framework to accommodate Full loads & Incremental loads across AWS and GCP environments.
* Worked on Data Analysis, validations, and audit framework implementation.
* Authored technical design documents and participated in architecture review boards to validate cloud solution designs before deployment.
* Responsible for collecting client requirements for specific data sources and designing solutions accordingly.
* Developed & implemented solutions using Apache Spark and Python API to load data from AWS S3 and Google Cloud Storage (GCS).
* Created dimension and fact tables and loaded transformed data into AWS Redshift and Google BigQuery.
* Applied end-to-end unit testing and documented results for each requirement story, reviewing them with the test lead before production deployment.
* Worked in Agile methodology to meet deadlines for full ELT cycle requirements.
* Collaborated closely with Business users, interacting with ETL developers, Project Managers, and QA teams.
* Created different KPIs using calculated key figures and parameters.
* Automated manual processes using Python scripts to improve efficiency and save time.
* Optimized performance and processing for existing data flows.
* Responsible for the documentation, design, development, and architecture of visualization reports.
* Handled the installation, configuration, and support of a multi-node setup for AWS EMR and GCP Dataproc.
* Developed automation solutions using Python to streamline processes across AWS and GCP environments.

**Tech Stack:** AWS S3, AWS Redshift, Google Cloud Storage (GCS), Google BigQuery, Apache Spark, Agile Methodology, Fact and Dimension Tables, Python Automation, KPIs, AWS EMR, Google Cloud Dataproc.

**Client: Indium Soft, Hyderabad, India**

**Role: Big Data Engineer Aug 2015– Sep 2018**

**Responsibilities:**

* Built and maintained ETL pipelines using Informatica PowerCenter to integrate sales data into Teradata while applying SCD Type 1 and Type 2 for handling historical data.
* Designed and optimized Informatica mappings for large-scale data loads using Fast Load and MultiLoad, with staging tables in Oracle DB to streamline transformations.
* Developed star schema-based data marts in Teradata, including fact and dimension tables, to support advanced reporting and analytics.
* Created ETL workflows to load data into data marts, enabling advanced business insights and KPI reporting.
* Integrated KPI logic into ETL workflows to ensure accurate business metric generation.
* Conducted extensive testing, including unit, integration, and user acceptance testing, to ensure data reliability.
* Documented ETL processes, transformations, and mappings, creating user manuals and data dictionaries for maintenance and user reference.

**Tech Stack:** Informatica PowerCenter, Teradata, Oracle DB, FastLoad, MultiLoad, Star Schema

**Client: Spurtree Technologies Inc, Bangalore, India Jun 2013 – Aug 2015**

**Role: ETL Developer**

**Responsibilities:**

* Designed and implemented customized database systems to meet client requirements.
* Authored comprehensive design specifications and documentation for database projects.
* Developed ETL pipelines using Pentaho for seamless integration of data from various sources.
* Troubleshot and optimized MapReduce jobs to resolve failures and improve performance.
* Facilitated data import/export across multiple systems and the Hadoop Distributed File System (HDFS).
* Built scalable and distributed data solutions utilizing Hadoop, Hive, MapReduce, and Spark.
* Transformed structured and semi-structured data using tools like Hive and Spark.
* Created detailed user documentation for Hadoop ecosystems and processes.Executed Hive queries to perform in-depth data analysis and validation.

**Tech Stack:** MapReduce, Pig, Hive, Hadoop, Cloudera, HBase, Sqoop