**Sai Gopi Thulluru**

**Phone:2602094395**

**Email: gopirajyahoo@gmail.com**

**LinkedIn:**[**linkedin.com/in/sai-gopi-thulluru-3045762b7**](https://www.linkedin.com/in/sai-gopi-thulluru-3045762b7)

**PROFESSIONAL SUMMARY:**

* 10+ years of IT experience with 5+ years in Data Engineering and 5 years in Data Warehousing across healthcare, finance, telecom, and technology domains.
* Specialized in designing scalable data pipelines using Azure Data Factory, GCP Dataflow, and AWS Glue for hybrid cloud environments.
* Expert in distributed processing using Apache Spark (PySpark/Scala), optimizing jobs with caching, partitioning, and AQE.
* Skilled in Snowflake data engineering including schema design, clustering, Time Travel, cloning, and workload isolation.
* Proficient in data modeling (Star, Snowflake, Data Vault 2.0) supporting BI and advanced analytics.
* Hands-on with real-time streaming using Kafka, Spark Streaming, AWS Kinesis, and GCP Pub/Sub for high-velocity pipelines.
* Strong expertise in SQL performance tuning, partitioning, and indexing to reduce query latency and compute costs.
* Experienced in ETL/ELT frameworks across Informatica, dbt, Airflow, Oozie, and ADF for batch and streaming.
* Implemented data governance, lineage, and cataloging using Azure Purview, Apache Atlas, and GCP Data Catalog.
* Skilled in data quality frameworks (Great Expectations, PyDeequ) for automated validations and reconciliation.
* Designed and deployed CI/CD pipelines in Jenkins, Azure DevOps, GitHub Actions for automated deployments.
* Built infrastructure with Terraform, CloudFormation, and ARM templates for cloud IaC provisioning.
* Worked with Kubernetes (AKS/EKS/GKE) and Docker for containerized data workloads.
* Proficient in multi-cloud architectures integrating AWS, Azure, and GCP services for cost-efficient solutions.
* Hands-on experience in Python, Scala, SQL, and Shell scripting for orchestration and automation.
* Integrated machine learning workflows into pipelines using Data bricks ML, Azure ML, and GCP Vertex AI.
* Strong in BI/Visualization with Power BI, Tableau, and Looker integrated with Snowflake and Synapse.
* Applied data security and compliance via IAM, RBAC, data masking, and encryption meeting HIPAA/GDPR.
* Designed data lakes and lake houses in S3, ADLS, and GCS with raw/curated/trusted zones.
* Managed Hadoop ecosystems (Hive, Pig, Sqoop, MapReduce, HBase) for legacy big data workloads.
* Hands-on with workflow orchestration in Airflow/ADF including scheduling, retries, SLAs, and monitoring.
* Agile practitioner with JIRA/Scrum/Safe experience, working across cross-functional global teams.
* Proven record of mentoring data engineers and leading performance optimization across pipelines.
* Experienced in building end-to-end data migration strategies and executing large-scale data movement across on-premise and cloud platforms, minimizing downtime and data loss.
* Developed and operationalized data observability frameworks leveraging tools like Monte Carlo and Datafold to proactively monitor data reliability.
* Implemented robust data versioning and reproducibility practices using tools such as DVC and MLflow for auditability and governance.
* Expertise in implementing CDC (Change Data Capture) solutions using Debezium, AWS DMS, and Oracle GoldenGate for real-time analytics and system integration.
* Optimized storage costs and retrieval performance by architecting tiered data storage solutions combining hot, warm, and cold layers across cloud providers.
* Led data privacy audits and remediation initiatives, working closely with legal and compliance teams to ensure regulatory adherence and data ethics.
* Built custom connectors and integrations with third-party APIs for seamless data acquisition and enrichment in heterogeneous environments.
* Advanced troubleshooting and resolution skills for complex distributed system issues impacting data availability and consistency in mission-critical pipelines.
* Drove adoption of data mesh and data product thinking, enabling decentralized data ownership and increased agility for business domains.

**TECHNICAL SKILLS:**

| **Category** | **Technologies & Tools** |
| --- | --- |
| Programming | Python, Scala, SQL, Shell Scripting |
| Cloud | AWS (Glue, S3, IAM, EC2, Kinesis, Redshift), Azure (ADF, Purview, ADLS), GCP (Dataflow, Pub/Sub, Data Catalog) |
| Data Processing | Apache Spark (PySpark/Scala), Apache Kafka, Spark Streaming, AWS Kinesis, GCP Dataflow, dbt |
| Databases | Snowflake, PostgreSQL, MySQL, MariaDB, Oracle, MongoDB, Cassandra, DynamoDB, HBase, Redis |
| ETL/ELT | Informatica, dbt, Airflow, Oozie, AWS Glue, Azure Data Factory |
| Data Orchestration | Apache Airflow, Control-M |
| Infrastructure as Code | Terraform, AWS CloudFormation, Azure ARM Templates |
| DevOps/CI/CD | Jenkins, GitHub Actions, Azure DevOps, GitLab CI/CD |
| Containerization | Docker, Kubernetes (EKS, AKS, GKE) |
| Data Governance | Azure Purview, Apache Atlas, GCP Data Catalog |
| Data Quality | Great Expectations, PyDeequ |
| Visualization/BI | Power BI, Tableau, Looker |
| Machine Learning | Databricks ML, Azure ML, GCP Vertex AI, MLflow |
| Big Data | Hadoop, Hive, Pig, Sqoop, MapReduce |
| Security/Compliance | IAM, RBAC, TLS/SSL, Data Masking, Encryption (at-rest/in-transit) |
| Workflow & Monitoring | Prometheus, Grafana, AppDynamics, Clou |

**CERTIFICATIONS:**

**Microsoft Certified: Azure Data Engineer Associate**

**PROFESSIONAL EXPERIENCE:**

**Client: Deloitte, Mechanicsburg Jun 2023 – Till Now**

**Role: Azure Snowflake data engineer**

**Responsibilities:**

* Designed enterprise-grade data pipelines in ADF & Glue, integrating RDBMS, APIs, and flat files into Snowflake and cloud data lakes.
* Developed Spark workflows in Databricks for distributed batch/stream processing, reducing transformation time by 35%.
* Implemented real-time ingestion with Kafka, Spark Streaming, and AWS Kinesis to support live reporting dashboards.
* Structured multi-zone data lakes (raw, curated, trusted) in ADLS & S3 with governance and retention policies.
* Created Snowflake schemas, clustering, and materialized views, optimizing query execution by 40%.
* Automated data quality checks using PyDeequ and Great Expectations, ensuring consistent and accurate datasets.
* Applied CDC (Change Data Capture) methods for incremental loading from SQL Server and Oracle into Snowflake.
* Optimized Spark job execution via partitioning, caching, and broadcast joins for large-scale analytics workloads.
* Built ETL/ELT pipelines using Informatica & dbt, enabling modular SQL transformations.
* Automated CI/CD deployments with Jenkins & Azure DevOps, integrating version control with GitHub repos.
* Integrated Power BI with Snowflake for dashboards, implementing row-level security (RLS) for secure reporting.
* Created Python scripts for data validation, API integration, and automation of reporting workflows.
* Applied Snowflake query tuning techniques (result caching, clustering keys) for cost and performance efficiency.
* Configured event-driven triggers in ADF to enable near real-time updates and reduce pipeline latency.
* Established metadata lineage using Purview/Atlas for audit and compliance requirements.
* Worked with business stakeholders and analysts to define KPIs, building interactive BI dashboards.
* Conducted capacity planning for Snowflake warehouses, implementing auto-scaling policies.
* Collaborated with DevOps teams to containerize pipelines with Docker & Kubernetes (AKS).
* Participated in full Agile lifecycle: sprint planning, backlog refinement, and PI planning.
* Mentored junior developers on Snowflake optimization and Spark programming best practices.
* Delivered cross-cloud integrations across AWS S3, Azure Blob, and GCP BigQuery for client analytics platforms.

**Environment**: Azure Databricks, Data Factory, Snowflake, Logic Apps, Functional App, Snowflake, MS SQL, Oracle, HDFS, MapReduce, YARN, Spark, Hive, SQL, Python, Scala, PySpark, shall scripting, GIT, JIRA, Jenkins, kaka, ADF Pipeline, Power Bi.

**Client: Government of Washington Dc, DC Apr 2021 – May 2023**

**Role: Azure Snowflake data engineer**

**Responsibilities:**

* Designed batch and streaming pipelines in ADF & Databricks, integrating Event Hubs, Functions, and Snowflake.
* Built ETL jobs in AWS Glue & GCP Dataflow, processing semi-structured data (JSON, Parquet, Avro).
* Developed machine learning workflows in Azure ML + Snowflake for predictive analytics and anomaly detection.
* Automated infrastructure with Terraform and GCP Deployment Manager, enabling IaC consistency.
* Created Snowflake stored procedures, UDFs, and secure views, encapsulating business logic.
* Implemented CDC replication from on-prem to Snowflake using ADF and Debezium.
* Integrated data cataloging solutions (Azure Purview, GCP Data Catalog) for lineage and governance.
* Optimized Spark configurations and partitioning, improving batch job performance by 30%.
* Designed row-level and column-level security in Snowflake for compliance with HIPAA/GDPR.
* Built Power BI dashboards for operations, finance, and public reporting with RLS and KPIs.
* Automated pipeline monitoring and alerting using Azure Monitor, QPM, and custom logging frameworks.
* Integrated Athena & Redshift for ad-hoc queries on AWS-hosted datasets.
* Processed large HDFS/Dataproc workloads, migrating legacy Hadoop jobs to Spark clusters.
* Collaborated with data scientists to prepare training datasets for ML experiments.
* Created data archiving policies using Snowflake Time Travel & Blob Storage retention rules.
* Established DevOps CI/CD pipelines using GitHub Actions and Cloud Build for versioned deployments.
* Developed unit tests and regression checks in PySpark to validate transformations.
* Conducted cost optimization in Snowflake via warehouse auto-suspend and query caching.
* Actively participated in Agile ceremonies, managing sprint deliverables and stakeholder reporting.
* Mentored internal staff on hybrid cloud integration between on-prem, AWS, Azure, and GCP platforms.
* Delivered scalable BI solutions with Power BI, Synapse, and Snowflake integration.

**Environment**: Azure Databricks, Data Factory, Logic Apps, Snowflake, Functional App, Snowflake, MS SQL, Oracle, HDFS, MapReduce, YARN, Spark, Hive, SQL, Python, Scala, PySpark, shall scripting, GIT, JIRA, Jenkins, kafka, ADF Pipeline, Power Bi.

**Client: Ameritas Apr 2020 – Mar 2021**

**Role: Big data Developer**

**Responsibilities:**

* Built and maintained data ingestion pipelines using Sqoop, Flume, and Kafka to process structured and semi-structured customer behavior data.
* Designed batch and streaming ETL workflows in Hive and Spark, enabling large-scale transformations with high reliability and low latency.
* Developed PySpark and Scala scripts for complex aggregations, cleansing, and enrichment of insurance datasets.
* Leveraged AWS Glue crawlers to catalog evolving schemas in S3, supporting automated discovery and data lake governance.
* Orchestrated multi-cloud jobs in Apache Airflow/Composer, coordinating workflows across AWS S3, GCP BigQuery, and Snowflake.
* Implemented data lake zoning (raw, curated, trusted) in Hadoop and AWS S3, enforcing governance and retention strategies.
* Migrated on-premises Oracle/Sybase workloads into Snowflake, implementing performance tuning and benchmarking.
* Built Python-based validation frameworks applying schema checks, null validations, and cross-system reconciliations.
* Applied data masking and RBAC policies in Snowflake, ensuring compliance with HIPAA and GDPR standards.
* Developed CI/CD pipelines in AWS Code Pipeline & Code Build, enabling automated deployments with zero downtime.
* Integrated business rules into Hive queries and automated transformations to meet insurance regulatory requirements.
* Tuned Spark jobs and Hive queries with bucketing, partitioning, and parallelism to reduce runtime by 30%.
* Designed reusable Oozie workflows to orchestrate Sqoop, Spark, and Hive jobs for batch pipelines.
* Created audit logging frameworks to track pipeline executions, ensuring traceability and compliance audits.
* Collaborated with analysts to deliver Power BI dashboards showing customer behavior KPIs and financial trends.
* Migrated legacy Excel and SSRS reports into Power BI, improving interactivity and executive reporting adoption.
* Enhanced Power BI performance using aggregations, DAX optimizations, and optimized data models.
* Partnered with DevOps teams to containerize data pipelines using Docker for portable deployments.
* Implemented real-time monitoring and alerting with custom Python loggers and AWS CloudWatch metrics.
* Worked in Agile Scrum teams, managing sprint deliverables with JIRA and conducting daily stand-ups.
* Used GitHub version control for collaborative development and branching strategy enforcement.
* Mentored junior developers on Spark optimization, Hive design patterns, and AWS Glue integration best practices.

**Environment**: Sqoop, MYSQL, HDFS, Apache Spark Scala, Hive Hadoop, Cloudera, HBASE, Kafka, MapReduce, Zookeeper, Oozie, Data Pipelines, RDBMS, AWS, EC2, Python, PySpark, shall script, Ambari, JIRA.

**Client: Charter Communications, Englewood, CO May 2019 – Mar 2020**

**Role: Big data Developer**

**Responsibilities:**

* Developed a scalable ETL framework using Sqoop, Pig, and Hive to integrate structured, semi-structured, and unstructured data from multiple sources.
* Built and maintained Spark ETL pipelines in Scala and PySpark, processing terabytes of telecom data with improved reliability and throughput.
* Designed Hive external tables and optimized schemas for high-performance querying in enterprise-scale data warehouses.
* Utilized AWS EMR clusters to run Spark jobs, achieving significant performance improvements in batch transformations.
* Built real-time streaming applications using Kafka and Spark Streaming for customer usage analytics and operational insights.
* Leveraged AWS Athena to create ad-hoc analytics dashboards querying data directly from S3 with reduced infrastructure costs.
* Applied MapReduce design patterns for classification algorithms, improving data enrichment workflows.
* Automated metadata lineage tracking by integrating Glue Catalog and Collibra, ensuring compliance and governance.
* Created reusable Informatica mappings for ETL jobs, reducing development time and enhancing maintainability.
* Implemented partitioning, bucketing, and data compression in Hive to optimize query performance.
* Developed YAML-based deployment pipelines to streamline code migration across development and production environments.
* Designed Spark-Cassandra connectors for real-time reporting and historical analysis on distributed NoSQL systems.
* Built error-handling and retry frameworks within Oozie workflows to ensure high pipeline availability.
* Collaborated with cross-functional teams to design data lake zones (raw, cleansed, aggregated), improving reusability and trust.
* Conducted root cause analysis on slow SQL/Hive queries, optimizing join strategies and execution plans.
* Developed HDFS ingestion frameworks for CSV, JSON, and XML data sources, ensuring flexible processing pipelines.
* Implemented real-time monitoring dashboards with Spark job logs, CloudWatch, and custom Python scripts.
* Used Git and GitHub for version control, maintaining branching strategies and collaborative workflows.
* Partnered with BI teams to design Power BI and Tableau dashboards integrating Hive/Spark data for customer analytics.
* Participated in Agile ceremonies (sprint planning, daily scrums, retrospectives), delivering milestones on time.
* Mentored junior developers on Hive optimization, Spark tuning, and best practices in cloud-native big data systems.

**Environment**: Hadoop, Hive, spark, PySpark, Sqoop, Spark SQL, Shell script, Cassandra, YAML, ETL.

**Client: Bank of America, Plano, TX May 2018 – Apr 2019**

**Role: Big data Developer**

**Responsibilities:**

* Designed and developed SSIS ETL workflows to migrate financial data from SQL Server, Access, and Excel into enterprise warehouses.
* Built and optimized stored procedures, triggers, functions, and indexed views to improve SQL Server performance.
* Applied dimensional modeling (Star, Snowflake, SCD Type 1 & 2) for risk, credit, and financial reporting.
* Tuned long-running Hive queries and MapReduce jobs by applying partitioning, bucketing, and resource optimization.
* Created and deployed SSAS cubes, dimensions, KPIs, and aggregations, supporting advanced OLAP reporting.
* Developed SSRS reports, dashboards, and scorecards with drill-down and drill-through for executive analytics.
* Leveraged PySpark and Hadoop (YARN, Ambari) to ingest and process large-scale financial transaction datasets.
* Integrated Kafka streaming pipelines for near real-time ingestion of high-volume customer data.
* Implemented data security controls with encryption, masking, and role-based access to meet compliance standards.
* Automated JIRA workflows for sprint tracking, SLA alerts, and project milestone reporting.
* Collaborated with QA and BI teams to ensure data quality validation and traceability across ETL pipelines.
* Participated in Agile ceremonies, delivering enhancements and bug fixes within sprint cycles.
* Documented ETL workflows, data dictionaries, and job dependencies for compliance and knowledge transfer.
* Monitored cluster health and Spark job performance, resolving memory and execution bottlenecks.
* Mentored junior developers on SQL tuning, SSIS pipelines, and Hadoop ecosystem best practices.

**Environment**: MS SQL Server 2016, Visual Studio 2017/2019, SSIS, Share point, MS Access, Team Foundation server, Git.

**Client: Zillow Group, Seattle, WA May 2016 – Apr 2018**

**Role: Data warehouse Developer**

**Responsibilities:**

* Designed and developed stored procedures, triggers, indexes, and UDFs to optimize SQL Server database performance.
* Built ETL workflows in SSIS to transfer data from Excel, Access, and flat files into SQL Server, ensuring clean and structured data.
* Implemented DTS and SSIS packages for automated imports, exports, and complex data transformations.
* Developed derived columns, lookups, and data conversions in SSIS to handle business-specific requirements.
* Created ETL logging and auditing frameworks at task and package levels to track processed records.
* Wrote and tuned T-SQL queries and scripts to support business analytics and reporting teams.
* Designed and deployed Power BI and SSRS reports with drill-down, summary, OLAP, and cross-tab functionality.
* Implemented data modeling (logical and physical) for enterprise reporting databases.
* Supported and maintained production SQL Server databases, resolving query and reporting issues.
* Built conditional, parameterized, and ad hoc reports to support finance and operational stakeholders.
* Provided application and database support including 24/7 production monitoring with custom SQL jobs.
* Utilized Erwin and TOAD for data modeling and schema design across enterprise data warehouse projects.
* Participated in Agile development cycles, collaborating with analysts and developers to deliver reporting enhancements.
* Documented ETL processes, report logic, and database schemas for compliance and knowledge sharing.

**Environment**: IBM WebSphere DataStage EE/7.0/6.0 (Manager, Designer, Director, Administrator), Ascential Profile Stage 6.0, Ascential QualityStage 6.0, Erwin, TOAD, Autosys, Oracle 9i, PL/SQL, SQL, UNIX Shell Scripts, Sun Solaris, Windows 2000.

**Client: JP MORGAN CHASE, West Haven, CT Feb 2013 – Apr 2016**

**Role: Data warehouse Developer**

**Responsibilities:**

* Designed and optimized stored procedures, triggers, and indexed views to enhance SQL Server performance and scalability.
* Developed ETL pipelines in SSIS to migrate and transform data from SQL Server, Access, and Excel into enterprise warehouses.
* Applied dimensional modeling (fact, dimension, SCD Type 1 & 2) for reporting and analytics in financial systems.
* Built and deployed SSAS cubes, dimensions, KPIs, and measures, supporting OLAP and advanced reporting needs.
* Created SSRS reports, dashboards, and scorecards with drill-through, drill-down, and cascading parameters.
* Implemented error handling, checkpoints, and logging in SSIS to ensure reliable and traceable ETL execution.
* Tuned T-SQL queries and SSIS data flows to improve execution plans and optimize ETL run times.
* Designed data marts and star/snowflake schemas, supporting BI and ad hoc reporting.
* Applied data governance and security policies, including row-level security and role-based access.
* Participated in Agile development cycles, attending sprint planning, reviews, and daily stand-ups.
* Collaborated with business analysts and QA teams to validate data quality and reporting accuracy.
* Provided production support for SQL Server databases, troubleshooting ETL failures and performance issues.
* Documented ETL workflows, cube designs, and reporting logic for knowledge sharing and compliance audits.
* Mentored junior developers on SSIS, SSAS, and best practices in data modeling and query tuning.

**Environment**: MS SQL Server 2016, Visual Studio 2017/2019, SSIS, Share point, MS Access, Team Foundation server, Git.