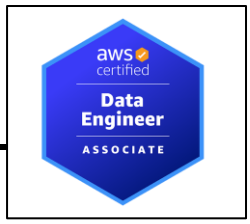
Deepika
Sr Data Engineer
Email: sreedhanasai@gmail.com **Mobile:** (910) 518-5105



PROFESSIONAL SUMMARY:

- Accomplished Senior Data Engineer with over 8 years of experience in the IT domain, delivering impactful and scalable solutions across multiple industry sectors.
- Skilled in building and maintaining advanced data processing systems using modern scripting and programming languages, enhancing automation, diagnostics, and performance.
- Designed and maintained resilient data lakes, pipelines, and scalable data frameworks to support both operational tasks and complex analytics workloads.
- Broad experience in creating, implementing, and enhancing cloud-based data infrastructures with seamless integration and high system availability.
- Strong background in the design, tuning, and management of relational and NoSQL databases for both transactional and large-scale data applications.
- Proven ability to engineer and fine-tune scalable data warehouse systems, including user-friendly interfaces for on-demand reporting and data analysis.
- Delivered critical business intelligence through dashboards, reporting tools, and analytical platforms, enabling data-driven strategic decisions.
- Proficient in automating infrastructure provisioning, scaling cloud resources, and monitoring production workflows to drive operational excellence.
- Deep understanding of data lifecycle, governance policies, and data quality frameworks, ensuring regulatory compliance and adherence to business standards.
- Skilled in managing various data formats and sources, supporting seamless data integration and optimized storage for structured and unstructured content.
- Experienced in schema design and dimensional modeling for efficient querying and performance in analytical environments.
- Strong hands-on expertise in deploying and tuning real-time and batch data pipelines in production with high availability and low latency.
- Solid knowledge of core networking technologies such as TCP/IP, DNS, and VPN, supporting secure and robust enterprise communications.
- Familiar with both Agile and Waterfall development methodologies, with practical experience in TDD, BDD, and continuous quality assurance across the software lifecycle.

TECHNICAL EXPERTISE:

Programming & Scripting	Python, Java, Scala, SQL, Bash, T-SQL, PL/SQL
Big Data & Processing Tools	Apache Spark, Hadoop, Hive, HDFS, MapReduce, Delta Lake, Presto, Pig, Kafka, Flink
API Development & Integration	REST APIs, GraphQL, Spring Boot (Java), Flask, FastAPI (Python)
Cloud Platforms	AWS, GCP, Databricks
Databases	PostgreSQL, MySQL, Oracle, SQL Server, MongoDB, Cassandra, DynamoDB, HBase
Data Warehousing & Modeling	Snowflake, Redshift, Star Schema, Snowflake Schema
ETL & Workflow Orchestration	SSIS, Informatica, IBM DataStage, Talend, Databricks
Reporting & BI Tools	Tableau, Power BI, AWS QuickSight, Looker, SSRS
Streaming & Real-time Process	Apache Kafka, AWS Kinesis, Spark Streaming
DevOps & CI/CD	Jenkins, GitLab CI/CD, AWS CodePipeline, Docker, Kubernetes, Terraform, CloudFormation
File Formats	JSON, XML, Avro, Parquet

PROFESSIONAL EXPERIENCE:

PNC Bank, Cleveland, OH **Jun 2024 to Current**
Senior Data Engineer

Responsibilities:

- Partnered with product managers and business stakeholders to capture data-driven product requirements, aligning deliverables with organizational objectives and PCI DSS compliance guidelines.
- Engineered end-to-end ETL pipelines utilizing AWS Glue, PySpark, Databricks, and Amazon EMR, employing RDDs and DataFrames to efficiently transform complex formats such as Parquet, Avro, and JSON.
- Architected and administered Apache Kafka topics with carefully optimized partitioning and replication strategies; integrated Kafka with AWS Kinesis and PySpark to enable fault-tolerant, low-latency stream processing for credit card transaction validation and fraud detection workflows.
- Automated data ingestion from heterogeneous databases including Oracle, Cassandra, and MongoDB into cloud-native platforms using Python frameworks, implementing salting and sharding techniques to ensure ACID compliance and achieve millisecond response times on queries.
- Developed batch processing pipelines for student loan datasets by replicating Oracle databases into an S3-backed data lake via AWS DMS, followed by high-throughput analytics in Amazon Redshift; applied rigorous data profiling to boost data quality and reliability.
- Strengthened data security by deploying fine-grained AWS IAM roles, KMS encryption for data at rest and in transit, and enforced private network access through VPC endpoints.
- Crafted and refined complex PL/SQL and PySpark workflows incorporating multi-stage data transformations, windowing functions, and recursive queries to optimize processing logic.
- Led the creation of a full-stack user interface using React.js for data warehouse interactions, enabling users to access both real-time transaction data and historical analytics.
- Designed RESTful APIs with Spring Boot and Flask to expose warehouse data securely, facilitating integration with internal banking platforms and third-party fraud detection systems.
- Developed dynamic data entry forms to support transaction validation workflows, allowing risk analysts to manually flag or adjust suspicious transactions prior to approval.
- Built event-driven real-time notification systems using Kafka and AWS SNS to alert financial teams on transaction anomalies, fraud indicators, and high-risk events, optimizing resource usage through dynamic partitioning and in-memory processing techniques.
- Integrated Java-based backend APIs with Node.js services to ensure smooth data flow between systems, utilizing RESTful and GraphQL endpoints for efficient retrieval of transaction and fraud detection data.
- Enhanced backend throughput with multi-threaded Java services for data ingestion, working alongside Node.js streams to process large transaction logs and minimize API latency.
- Implemented OAuth 2.0 authentication for external API consumers, integrating with AWS Cognito and IAM policies to enforce fine-grained access control.
- Designed and optimized Hadoop MapReduce jobs for large-scale batch processing of student loan data on AWS EMR, employing custom mappers and reducers and integrating Hive and HDFS for efficient storage and retrieval.
- Improved Hadoop workflows by applying Apache Spark optimizations such as partition pruning, dynamic resource allocation, and in-memory caching for data transformations within HDFS and HiveQL.
- Utilized Google Cloud Pub/Sub for real-time messaging combined with Dataflow for stream processing and BigQuery for analytics, building low-latency, fault-tolerant pipelines featuring message deduplication, auto-scaling, and schema evolution.
- Designed partitioned and clustered tables to efficiently handle multi-terabyte queries while maintaining detailed data lineage via a lineage graph.
- Configured Control-M job scheduling and triggered post-batch ETL processes using AWS Lambda to orchestrate end-to-end data curation pipelines with minimal latency.
- Developed comprehensive monitoring and alerting frameworks leveraging AWS CloudWatch, GCP Monitoring, and SNS, implementing event-driven triggers and anomaly detection for proactive incident management.
- Executed migration of legacy Oracle, MongoDB, and Cassandra databases to AWS RDS, DynamoDB, and BigQuery using AWS DMS and Schema Conversion Tool for seamless schema and data transformation.
- Built CI/CD pipelines incorporating Git and Jenkins, integrating automated linting, unit testing, and artifact versioning for reliable deployments with rollback capabilities.
- Provisioned cloud infrastructure through Terraform and AWS CloudFormation, configuring ECS, VPCs, IAM roles, and Redshift clusters to support production-grade data ecosystems.

- Deployed containerized workloads using Kubernetes on AWS EKS, enabling horizontal scaling, resource quota enforcement, and self-healing capabilities for data processing applications.
- Facilitated Agile sprint planning and backlog refinement sessions, ensuring effective collaboration among data engineering, analytics, and DevOps teams to prioritize pipeline enhancements.
- Conducted proactive cost analysis and optimization of data pipelines using AWS Cost Explorer and Databricks cluster tuning, driving reductions in cloud expenditure.
- Applied cost-saving strategies through the use of spot instances and auto-scaling policies to optimize resource allocation and minimize operational costs.

Environment:
AWS (Glue, Lambda, Kinesis, Redshift, CloudFormation, S3, DynamoDB, RDS, EMR, EKS, CloudWatch, SNS, SQS, IAM, DMS, VPC), Google Cloud Platform (Dataflow, BigQuery, Monitoring), Databricks, Apache Kafka, PySpark, Terraform, Control-M, Git, Jenkins, Docker, Kubernetes, Oracle, MongoDB, Cassandra, Node.js, Java (Spring Boot), React.js, WebSockets, Python (Pytest, Unittest), Bash, SQL (PL/SQL), Hive, HDFS, Spot Instances, Data Formats (JSON, Parquet).

BCBS NC, Durham, NC. **Aug 2021 to May 2024**
Data Engineer

Responsibilities:
- Collaborated with cross-functional teams to gather and assess data requirements related to claims processing, policy management, and predictive analytics, ensuring full compliance with HIPAA through the application of robust security protocols.
- Architected and fine-tuned Snowflake data warehouse solutions, utilizing multi-cluster compute resources, clustering keys, and materialized views to significantly improve query speed and scalability for claims and policy datasets.
- Built comprehensive ETL pipelines and workflows leveraging AWS Glue, Amazon Redshift, and Snowflake, incorporating Dynamic Data Masking (DDM), Secure Views, and Row Access Policies to tightly control access to sensitive healthcare information.
- Automated real-time data ingestion and transformation processes using PySpark along with Snowflake Streams and Tasks, supporting continuous data updates and integration for healthcare analytics.
- Employed Snowflake's Time Travel and Fail-safe capabilities to enable data recovery and ensure regulatory compliance, providing a robust disaster recovery framework.
- Designed Slowly Changing Dimensions (SCD) Types 1 and 2 using AWS Glue and Snowflake to accurately track historical data changes for trend analysis and comprehensive reporting.
- Implemented tokenization in Snowflake via Secure Functions to replace sensitive data with unique tokens, safeguarding test and staging environments while preserving analytical accuracy.
- Developed PySpark ETL workflows within Databricks integrated with Snowflake, utilizing Unity Catalog to enforce secure and compliant data governance over high-volume transactional datasets.
- Enhanced security and query performance in Amazon Redshift by optimizing distribution and sort keys alongside workload management queues, enabling seamless interoperability with Snowflake in hybrid cloud architectures.
- Designed and launched a cloud-based reporting platform on AWS, empowering healthcare analysts with real-time access to claims data and the ability to execute custom SQL queries through a user-friendly web interface.
- Created a Vue.js frontend coupled with a Flask API backend to provide real-time tracking of claims status, including approvals, denials, and fraud notifications.
- Containerized ETL workflows using Docker and deployed them on Amazon ECS with Fargate, delivering scalable and secure healthcare data processing environments.
- Integrated Amazon SQS, SNS, and AWS Lambda to build a reliable, event-driven infrastructure that supports smooth communication and data flow across distributed systems.
- Managed complex data orchestration using Apache Airflow, developing DAGs for scheduling, monitoring, and handling dependencies within ETL pipelines across Snowflake and AWS platforms.
- Implemented advanced monitoring and observability solutions using Datadog to track performance metrics, analyze logs, and enable end-to-end traceability across multi-cloud environments.
- Tuned Snowflake warehouses for optimized query execution and cost control, applying Resource Monitors to enforce budget limits and avoid overspending.
- Executed sophisticated Spark transformations in PySpark for data cleansing, enrichment, and validation, ensuring high data quality and readiness for downstream business intelligence.
- Applied statistical and predictive modeling techniques to identify fraudulent claim patterns, reducing detection time and improving fraud prevention efforts.

- Leveraged time-series forecasting on healthcare data to anticipate claim volumes, enhancing resource planning for processing operations.
- Designed backup and disaster recovery protocols by combining AWS Backup with Snowflake data replication, meeting stringent data retention and availability standards.
- Streamlined continuous integration and delivery pipelines using AWS CodePipeline and CodeBuild, automating testing, deployments, and infrastructure updates while maintaining healthcare security compliance.

Environment:
HIPAA Compliance, Snowflake, AWS (Glue, Redshift, Lambda, S3, RDS, Spark, ECS, Fargate, Athena, CloudWatch, SQS, SNS, CodePipeline, CodeBuild, Backup), PySpark, SnowSQL, Databricks, Docker, Apache Airflow, Jenkins, Power BI, Datadog, Python (Pytest, Unittest), SQL, REST APIs, JSON, Kubernetes, Spot Instances, HDFS, Hive, Data Formats (JSON, Parquet), Git, Bash, React.js, Flask, FastAPI.

Progressive Insurance, Cleveland, OH **Mar 2020 to Jul 2021**
Data Engineer

Responsibilities:
- Led the transition of ETL workflows from PostgreSQL to AWS infrastructure, utilizing EMR, Redshift, and S3 for scalable data processing, while integrating Snowflake to enable advanced cloud-native analytics.
- Leveraged AWS Glue and Snowflake's Data Cloud to automate schema discovery and ETL processes, substantially accelerating insights for critical insurance functions.
- Designed and implemented automated ETL pipelines using Apache Spark, Python, AWS Step Functions, and Snowflake Streams, ensuring efficient data ingestion, transformation, and real-time synchronization across multiple platforms.
- Integrated a variety of data sources including Hive, JSON files, and Apache NiFi into unified workflows; optimized log management with Kafka, S3, and Snowflake to support real-time analytics.
- Enhanced claims processing and fraud detection by setting up SNS and SQS for instant event notifications, while utilizing Snowflake's Time Travel feature for historical data auditing and regulatory compliance.
- Developed secure data sharing mechanisms through Snowflake Secure Data Sharing, promoting cross-department collaboration and strict adherence to governance standards.
- Conducted comprehensive evaluations of PostgreSQL, AWS Redshift, and Snowflake to boost query performance and scalability, facilitating efficient handling of large-scale insurance datasets.
- Enabled analysts with ad hoc query capabilities via Presto and Snowflake's SQL engine, delivering faster insights into risk evaluation and operational trends.
- Created dynamic dashboards in Tableau, integrating with Redshift, Snowflake, and S3, providing real-time reporting and actionable intelligence for insurance operations.
- Built and scaled an internal analytics interface to furnish underwriters and actuaries with real-time risk assessments based on policyholder data and external market indicators.
- Engineered an automated, self-service reporting platform using React.js, AWS Athena, and QuickSight, empowering non-technical users to generate complex SQL queries through an intuitive interface.
- Utilized advanced Tableau functionalities such as calculated fields, parameter controls, and data blending, enabling stakeholders to conduct detailed risk analyses, monitor KPIs, and make informed decisions with enhanced operational transparency.
- Orchestrated scalable, automated data pipelines using AWS Managed Workflows for Apache Airflow (MWAA) and Snowflake Tasks, improving operational efficiency and reliability.
- Fostered improved collaboration and streamlined deployments by managing ETL codebases in GitHub and Bitbucket repositories and implementing CI/CD pipelines, minimizing downtime during iterative development cycles.

Environment:
AWS (Redshift, EMR, S3, SNS/SQS, DynamoDB, Glue, Step Functions, MWAA), Snowflake (Data Cloud, Secure Data Sharing, Streams, Time Travel), PostgreSQL, Apache Airflow, Apache NiFi, Apache Spark, Hive, Presto, Apache Parquet, Kafka, Tableau, GitHub, Bitbucket, Vue.js, Flask, FastAPI, JIRA, Confluence.

Visteon, Chennai, India **May 2015 to Dec 2018**
ETL Developer

Responsibilities:

- Designed and executed comprehensive ETL workflows using SSIS and IBM DataStage to extract, transform, and load data from multiple sources including flat files, Excel spreadsheets, IBM DB2, and Microsoft SQL Server, ensuring seamless data integration.
- Crafted and fine-tuned complex T-SQL queries, stored procedures, and triggers to implement business logic, facilitate data transformations, and support reporting requirements.
- Managed SQL Server upgrades from 2014 to 2016, employing in-place upgrade methods and backup/restore techniques to guarantee high availability, maintain data integrity, and prevent data loss.
- Improved ETL pipeline efficiency by optimizing SSIS and DataStage processes, enhancing memory utilization, and leveraging parallel processing for handling large-scale data workflows.
- Simplified migration processes across development, QA, and production environments by configuring SSIS package settings and implementing event-driven error handling for proactive issue resolution.
- Migrated legacy ETL pipelines to Google Cloud Platform (GCP), leveraging its scalable infrastructure to boost system performance and simplify maintenance.
- Integrated ETL results with MicroStrategy to deliver real-time analytics via interactive dashboards and reports, empowering data-driven decision-making.
- Utilized Git for source control and repository management, promoting collaborative development and robust versioning across ETL and database projects.
- Led daily stand-ups and sprint retrospectives, using JIRA Kanban boards to monitor ETL development progress and enhance team productivity in a fast-paced, cross-functional setting.

Environment:
Google Cloud Platform (GCP), Microsoft SQL Server 2014/2016, IBM DB2, SSIS, IBM DataStage, T-SQL, Python, Git, Visual Studio, MicroStrategy, JSON, CSV, Excel.

EDUCATION:

- Master in Data Science, Pace University, NY