**SAMARNATH REDDY**



**AWS DATA ENGINEER**

**Phone:**  682‑727‑0147 **Mail:** samarnathreddyyy@gmail.com   **LinkedIn:** <https://www.linkedin.com/in/samarnath-y>

**Technical Skills**

|  |  |
| --- | --- |
| AWS Services | S3, Redshift, RDS, DynamoDB, EMR, EC2, ECS, Lambda, Athena, Glue, Data Migration Services, Kinesis, CloudWatch, IAM, route53, VPC, SNS. Aurora. |
| Big Data Technologies | Hadoop (HDFS, YARN, Hive, Impala, HBase), Spark (SQL, Streaming), Kafka, Oozie, Airflow, Sqoop, Cassandra, HBase |
| Hadoop Distribution | Cloudera, Horton Works |
| Languages | Java, SQL, PL/SQL, Python, HiveQL, Scala, R. |
| Monitoring: | Prometheus, Grafana, ELK Stack, CloudWatch. |
| Database | Teradata, Oracle, Microsoft SQL Server, MySQL, Postgres, MongoDB, DynamoDB, RDS, S3 |
| Visualization Tools | Quick Sight, Power BI |
| Scheduling | IBM Tivoli, control-m, oozie, airflow, Docker, Control M |
| Version Control | GIT, GitHub, VSS, Bitbucket |
| Methodology | Agile, Scrum. |
| IDE &Build Tools, Design | Eclipse, Visual Studio |
| ETL Tools: | Apache NiFi, AWS Glue, Kafka Connect, Pentaho, SSIS, Informatica |

|  |
| --- |
|  |

* TECHNICAL SKILLS
* AWS Services: AWS S3, redshift, EMR, SNS, SQS, Athena, glue, CloudWatch, Kinesis, route53, IAM
* Big Data Technologies: HDFS, SQOOP, PySpark, hive, MapReduce, spark, spark streaming, HBASE
* Hadoop Distribution: Cloudera, Horton Works
* Cloudera, Horton Works
* Languages
* Java, SQL, PL/SQL, Python, HiveQL, Scala.
* Operating Systems
* Windows (XP/7/8/10), UNIX, LINUX, UBUNTU, CENTOS.
* Database
* Teradata, oracle, SQL server,
* Scheduling
* IBM Tivoli, control-m, oozie, airflow
* Version Control
* GIT, GitHub, VSS, Bitbucket
* Methodology
* Agile, Scrum.
* IDE &Build Tools, Design
* Eclipse, Visual Studio.
* Environment:
* Windows server, MS SQL Server 2014, SSIS, SSAS, SSRS, SQL Profiler, Power BI, C#, Performance Point Server, MS Office, SharePoint

**Education**

* Bachelors: Electronics and Communication Engineering, 2011, Sathyabama University
* Masters: Computer Science, 2013, University of North Texas

**Professional Summary**

A goal-oriented and data driven Data Engineer with expertise in AWS and Snowflake, seeking to apply skills in designing and optimizing scalable data solutions. Committed to delivering exceptional results through strong analytical abilities and a comprehensive understanding of data architecture. Dedicated to enabling organizations to make data-driven decisions and achieve success.

* With over 10 years of industry expertise, I bring extensive experience as an IT professional, including 6 years dedicated to Data Engineering and an additional 4 years specializing in Data Warehousing.
* Extensive experience in Data Engineering field including Ingestion, Datawarehouse, Reporting, and Analytics.
* Expertise in AWS S3 for scalable and cost-effective data storage and retrieval. Skilled in utilizing AWS EMR for big data processing, including technologies like Hadoop, Spark, Hive, MapReduce, and PySpark.
* Proficient in integrating AWS SNS and SQS for real-time event processing and messaging.
* Designed and Developed ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, Parquet/Text Files into AWS Redshift.
* I have extensive expertise in utilizing AWS CloudWatch to monitor and oversee AWS resources, establish alert systems, and gather performance data.
* Proficient in managing user access and permissions to AWS resources using IAM.
* Encapsulated data processing applications, ETL (Extract, Transform, Load) workflows, and data pipelines into Docker containers and deployed across AWS environments, ensuring that the application works the same way in development, testing, and production.
* Developed REST APIs that adhere to RESTful principles, including stateless communication, standard HTTP methods (GET, POST, PUT, DELETE), and proper status codes for clear responses.
* Experienced in partitioning strategies and multi-cluster warehouses in Snowflake to ensure optimal query performance and scalability.
* Skilled in designing roles, views, and implementing performance tuning techniques to enhance Snowflake system performance.
* Proficient in utilizing virtual warehouses, caching, and Snowpipe for real-time data ingestion and processing in Snowflake.
* I have extensive expertise in Snowflake's time travel capability, enabling me to conduct meticulous audits and in-depth analysis of historical data.
* Extensive experience in leveraging window functions, Snowflake arrays, regular expressions, and JSON parsing for advanced data analysis and manipulation.
* Highly proficient in Snowflake scripting to automate ETL processes, data transformations, and data pipelines.
* Created and maintained several types of Snowflake tables, including transitory, temporary, and persistent tables, to meet unique data storage and processing needs.
* Implemented regular expressions in Snowflake for seamless pattern matching and data extraction tasks.
* Worked on Dimensional Data modeling in Star and Snowflake schemas and Slowly Changing Dimensions (SCD).
* Extensive experience in working with HDFS, Sqoop, PySpark, Hive, MapReduce, and HBase for big data processing and analytics.
* Proficiently developed real-time analytics solutions and event-driven applications using Apache Flink. These solutions enable immediate insights, facilitate fraud detection, clickstream analysis, and process IOT data.
* Used Kubernetes to design, manage, and orchestrate data processing pipelines as a series of interconnected containers or Pods.
* Implemented a data pipeline using Kafka, Spark, and Hive for ingestion, transformation, and analysis of data.
* Developed custom scripts and tools using Oracle's PL/SQL language to automate data validation, cleansing, and transformation processes.
* Implemented Spark Streaming for real-time data processing and advanced analytics.
* Leveraged HBASE integration with Hive to build HBASE tables in the Analytics Zone.
* Demonstrated expertise in scheduling and job automation using IBM Tivoli, Control-M, Oozie, and Airflow, for the execution of data processing and ETL pipelines.
* Proficient in version control systems like Git, GitLab, and Bitbucket for code repository management and collaboration.

**Work History**

AWS Data Engineer  **July 2022 - Current**

Credit Suisse**, Madison Avenue, New York**

* Proficient in designing and implementing real-time data streaming solutions using AWS Kinesis, facilitating the processing of high-volume data streams for immediate insights and analytics.
* Extensive experience in leveraging AWS S3 for scalable and secure data storage, managing data lakes, and optimizing data retrieval performance.
* Skilled in utilizing AWS DynamoDB to develop NoSQL database solutions, providing low-latency, high-performance storage for dynamic and rapidly changing datasets.
* Designed and Developed ETL workflows using AWS Glue to seamlessly extract, transform, and load data from various sources into AWS Redshift and PostgreSQL, ensuring data accuracy and consistency across multiple data stores.
* Proficiently integrated AWS SNS and SQS for real-time event-driven architectures, enabling efficient communication and data sharing between different components of complex systems.
* Implemented robust user access management and permissions to AWS resources using IAM, ensuring secure and controlled access to sensitive data and services.
* Proficient in managing and optimizing data warehousing solutions using AWS Redshift, including designing data models, setting up performance tuning, and ensuring efficient query execution for analytics and reporting.
* Extensive experience in utilizing AWS CloudWatch to monitor and manage AWS resources, creating custom dashboards, setting up alarms, and analyzing performance metrics to ensure system health and reliability.
* Designed and implemented RESTful APIs using AWS API Gateway, enabling seamless and secure communication between applications and services, and facilitating easy integration with external partners and clients.
* Proficient in deploying and managing Docker containers using Amazon ECS, optimizing resource allocation, scaling applications, and ensuring high availability.
* Experience in creating ECS clusters, defining task definitions, setting up auto-scaling policies, and integrating with AWS load balancers for seamless application deployment and management.
* Demonstrated expertise in automating ECS deployment processes using AWS CloudFormation, Terraform, or similar infrastructure as code (IaC) tools, streamlining provisioning and ensuring consistency across environments.
* Implemented CI/CD pipelines with tools like AWS Code Pipeline and AWS Code Build to automate container builds, testing, and deployment to ECS, enhancing development efficiency and accelerating time to market.
* Successfully integrated Snowflake with streaming technologies like Apache Kafka through Databricks to create hybrid data pipelines that handle both real-time streaming data and batch data processing.
* Utilized Snowflake's time travel features in conjunction with Databricks to track and analyze historical streaming data, enabling thorough audits and comprehensive analysis of data changes over time.
* Extensive experience in designing and implementing data processing workflows using Hadoop ecosystem technologies, including HDFS, MapReduce, Hive, and Sqoop, orchestrated and optimized through Databricks, to efficiently process and analyze large datasets.
* Proficient in leveraging Databricks for Apache Spark-based big data processing, utilizing its advanced analytics capabilities, including data transformations, machine learning, and real-time processing, to derive actionable insights.
* Successfully developed Spark Streaming applications for real-time data processing, enabling immediate data analysis and decision-making in dynamic environments.
* Skilled in Python programming for data processing, transformation, and integration, utilizing its libraries and frameworks to manipulate and analyze datasets efficiently.
* Proficient in utilizing Python scripting to automate ETL processes, data transformations, and data pipelines, enhancing operational efficiency and reducing manual intervention.
* Successfully developed data processing solutions using PySpark, combining the power of Python and Spark to perform complex data transformations and analytics.
* Proficient in designing and implementing automated job scheduling workflows using tools such as IBM Tivoli, Control-M, Oozie, and Apache Airflow, optimizing resource utilization and ensuring seamless execution of complex data processing pipelines.
* Collaborated with cross-functional teams to gather requirements, design automation workflows, and establish proactive monitoring and error-handling mechanisms, enhancing operational efficiency and minimizing manual intervention.
* Proficient in utilizing version control systems such as Git, GitLab, and Bitbucket to manage code repositories, track changes, collaborate with team members, and ensure code integrity throughout development projects.

**Environment:** AWS, AWS S3, redshift, EMR, SNS, SQS, Athena, glue, CloudWatch, Kinesis, route53, IAM, ECS, Sqoop, MYSQL, HDFS, Apache Spark, Hive, Cloudera, Kafka, Zookeeper, Oozie, PySpark, Ambari, JIRA, IBM Tivoli, control-m, OOZIE, airflow, Teradata, oracle, SQL, Docker, Flink.

AWS Snowflake Data Engineer **Jan 2018 – July 2022**

Visa USA Inc, **Ashburn, Virginia**

* Solid experience and understanding of implementing large-scale Data Warehousing Programs and E2E Data Integration Solutions on Snowflake Cloud, AWS Redshift, Informatica Intelligent Cloud Services (IICS - CDI), and Informatica PowerCenter integrated with multiple relational databases (MySQL, Teradata, Oracle, Sybase, SQL Server, DB2).
* Proficient in optimizing Flink jobs and Snowflake queries to achieve maximum performance and efficiency while integrating Flink with Snowflake, ensuring data integrity and accuracy.
* Defined robust roles and access privileges within Snowflake to enforce strict data security and governance protocols.
* Created and maintained various types of Snowflake tables, including transitory, temporary, and persistent tables, to meet unique data storage and processing needs.
* Developed complex SnowSQL queries to extract, transform, and load data from various sources into Snowflake.
* Optimized Snowflake warehouses by selecting appropriate sizes and configurations to achieve optimal performance and cost efficiency.
* Worked on Dimensional Data modeling in Star and Snowflake schemas and Slowly Changing Dimensions (SCD).
* Implemented regular expressions in Snowflake for seamless pattern matching and data extraction tasks.
* Implemented Snowflake caching mechanisms, reducing data transfer costs, and improving query performance.
* Developed and implemented Snowflake scripting solutions to automate critical data pipelines, ETL processes, and data transformations.
* Enhanced data by combining real-time Kinesis data with valuable reference information from various sources like databases and external APIs using Flink's powerful join capabilities.
* Experienced with Teradata utilities Fast Load, Multi Load, BTEQ scripting, Fast Export, SQL Assistant, and Tuning of Teradata Queries using Explain plan.
* Developed ETL processes in AWS Glue to migrate Campaign data from external sources like S3, Parquet/Text Files into AWS Redshift.
* Configured and fine-tuned Redshift clusters to achieve high-performance data processing and streamlined querying.
* Integrated AWS SNS and SQS for real-time event processing and efficient messaging.
* Implemented AWS Athena for ad-hoc data analysis and querying on data stored in AWS S3.
* Designed and implemented data streaming solutions using AWS Kinesis, enabling real-time data processing and analysis.
* Effectively managed DNS configurations and routing using AWS Route53, ensuring efficient deployment of applications and services.
* Implemented robust IAM policies and roles to ensure secure user access and permissions for AWS resources.
* Used Scala and Apache Spark to perform aggregations on large datasets, saving outcomes in the Hive data warehouse for further analysis.
* Developed custom scripts and tools using Oracle's PL/SQL language to automate data validation, cleansing, and transformation processes.
* Implemented a comprehensive data pipeline using Kafka, Databricks, Spark, and Hive for seamless ingestion, transformation, and analysis of data.
* Leveraged HBASE integration with Databricks and Hive to construct HBASE tables in the Analytics Zone, optimizing data storage and retrieval.
* Integrated Spark Streaming seamlessly through Databricks for real-time data processing and advanced analytics, enabling timely insights.
* Proficient in monitoring and optimizing ECS container performance using AWS CloudWatch, custom metrics, and logs analysis, effectively identifying, and addressing bottlenecks and improving application responsiveness.
* Experience in configuring alarms, setting up dashboards, and utilizing AWS CloudTrail for auditing ECS actions, ensuring compliance with security standards and best practices.
* Designed and developed database solutions using Teradata, Oracle, and SQL Server, including schema design and optimization, stored procedures, triggers, and cursors.
* Proficient in utilizing version control systems such as Git, GitLab, and VSS for efficient code repository management and collaborative development processes.

**Environment:** AWS, AWS S3, redshift, EMR, SNS, ECS, Athena, glue, CloudWatch, Kinesis, route53, IAM, Sqoop, MYSQL, HDFS, Apache Spark, Hive, Cloudera, Kafka, Zookeeper, Oozie, PySpark, Ambari, JIRA, IBM Tivoli, control-m, OOZIE, airflow, Teradata, oracle, SQL, Apache Flink.

Hadoop Developer  **June 2016 – Dec 2017**

Cigna healthcare, Nashville, TN

* Experience in importing and exporting data using Sqoop from HDFS to Relational Database Systems and vice versa.
* Worked extensively with Data Lakes and big data ecosystems, including Hadoop, Spark, Hortonworks, and Cloudera.
* Expertise in writing Spark RDD transformations, actions, Data Frame's, Persistence (Caching), Accumulators, Broadcast Variables, Case classes for the required input data and performed the data transformation using Spark-core.
* Utilized Kafka and Spark Streaming to process streaming data for specific use cases.
* Wrote Hive queries and used Hive QL to simulate MapReduce functionalities for data analysis and processing.
* Processed HDFS data and created external tables using Hive and developed scripts to ingest and repair tables that can be reused across the project.
* Imported data from MySQL to HDFS on a regular basis using Sqoop for efficient data loading.
* Analyzed the source data and handled efficiently by modifying the data types.
* Used excel sheet, flat files, CSV files to generate PowerBI ad-hoc reports.
* Processed the feeds to Kafka streaming API and processed and stored streaming data into Cassandra.
* Prepared an ETL framework with the help of Sqoop, pig and hive to be able to frequently bring in data from the source and make it available for consumption.
* Analyzed the SQL scripts and designed the solution to implement using PySpark.
* Created Airflow DAGs to schedule the Ingestions, ETL jobs and various business reports.
* Extracted the data from other data sources into HDFS using Sqoop.
* Developed data pipelines using Flume and Sqoop to ingest customer behavioral data into HDFS for analysis.
* Implemented CI/CD pipelines for building and deploying projects in the Hadoop environment by using docker.
* Handled importing of data from various data sources, performed transformations using Hive, MapReduce, loaded data into HDFS.
* Extracted the data from MySQL into HDFS using Sqoop.
* Utilized the Oozie workflow engine for job scheduling in Hadoop.
* Implemented automation for deployments by using YAML scripts for massive builds and releases.
* Implemented Data classification algorithms using MapReduce design patterns.
* Extensively worked on creating combiners, Partitioning, distributed cache to improve the performance of MapReduce jobs.
* Worked on GIT to maintain source code in Git and GitHub repositories.

**Environment**: Hadoop, Hive, spark, PySpark, Sqoop, Spark SQL, Cassandra, YAML, ETL, Cloudera,

Kafka, Zookeeper, Oozie, Docker.

Data Warehouse Developer **Sept 2015 – May 2016**

Aetna Inc, Hartford, CT

* Experience in loading various data sources like Oracle, SQL Server, Teradata, DB2 and Flat Files into DataMart's.
* Experience in requirement gathering and documenting.
* Extensively worked with mappings using different transformations like Filter, Joiner, Router, Source Qualifier, Expression, Union, Update Strategy, Unconnected / Connected Lookup, Aggregator and SCD Type-2.
* Interacted with functional and end users to gather requirements for the core reporting system, understand the exceptional features users expect from ETL and Reporting systems, and successfully implement business logic.
* Utilizing TFS for source controlling and tracking environment-specific script deployments.
* Exporting current data models from Erwin to PDF format and publishing them on SharePoint for user access.
* Developing, administering, and managing databases such as Consolidated Data Store, Reference Database, and Actuarial Data Mart.
* Writing triggers, stored procedures, and functions using Transact-SQL (T-SQL) and maintaining physical database structures.
* Deploying scripts in different environments based on Configuration Management and Playbook requirements.
* Creating and managing files and file groups, establishing table/index associations, and optimizing query and performance tuning.
* Tracking and closing defects using Quality Center for effective issue management.
* Maintaining users, roles, and permissions within the SQL Server environment.

**Environment:** SQL Server 2008/2012 Enterprise Edition, SSRS, SSIS, T-SQL, Windows Server 2003, Performance Point Server 2007, Oracle 10g, visual Studio 2010.

Data Warehouse Developer **Jan 2013– Aug 2015**

Charter Communications, **Negaunee, MI**

* Experienced using SSIS to create packages to transfer data between ORACLE, MS ACCESS, FLAT FILES, Excel files, SQL SERVER 2008R2.
* Expertise in creating ETL packages that used SSIS to extract data from heterogeneous databases, transform it, and load it into data marts.
* Had extensive experience in deploying SSIS packages to production environments, using different types of package configurations to export different package properties, and isolating package environments.
* Created and maintained a server inventory and performance inventory database.
* Engaged in Agile Scrum methodology in daily standup meetings, had extensive knowledge of using Visual SourceSafe in Visual Studio 2010, tracked projects in Trello.
* Experienced in creating, managing, and delivering both paper-based and interactive web-based reports using SQL Server Reporting Services (SSRS).
* Created drop-down menu options to generate drill-through and drill-down reports, sorted data and defined subtotals in Power BI.
* Experienced in creating cubes and dimensions using various architectures and data sources for business intelligence and creating MDX scripts.
* Created a data warehouse to develop data marts used to feed downstream reports, utilized user access tools that allowed users to create ad-hoc reports and run queries to analyze data in proposed cubes.
* Deployed SSIS packages and created jobs to efficiently execute packages.
* Participated in creating SSIS jobs to automate report generation and cube update packages.
* Developed stored procedures and triggers to enable consistent data entry into the database.
* Used Snowflake to share data externally and quickly set up data sharing without pipeline transfers or development.

**Environment:** Windows server, MS SQL Server 2014, SSIS, SSAS, SSRS, SQL Profiler, Power BI, C#, Performance Point Server, MS Office, SharePoint.