**SANGEETHA**

**SR DATA ENGINEER (AWS)**

**EmailID:** [**sangeethaaya21@gmail.com**](mailto:sangeethaaya21@gmail.com) **LINKEDIN:**

**Phone: +1 551-229-6098** www.linkedin.com/in/sangeetha-narayanasamy-san

**PROFESSIONAL SUMMARY:**

With over 8+ years of professional experience, I specialise in roles encompassing Data Engineering and Python Development, crafting robust data-intensive applications. My expertise spans Cloud Data Engineering, Big Data Analytics, Data Warehousing, and implementing Data Visualization, Reporting, and Data Quality solutions.

* Good knowledge of AWS technologies including S3, EC2, SQS, RDS, EMR, Kinesis, Lambda, Event Bridge, Glue, DynamoDB, Elasticsearch, Service Catalog, CloudWatch, and IAM.
* Hands-on experience in migrating data from AWS S3 to Snowflake using Snowpipe, focusing on data governance and quality.
* Skilled in NoSQL databases such as DynamoDB, Cosmos DB, MongoDB, and HBase, integrating them for real-time data processing via Rest API.
* Well-versed in Spark architecture and components, adept at utilizing Spark Core, Spark SQL, and Spark Streaming for interactive, batch, and stream processing.
* Proficient in Python scripting for statistical functions using NumPy, visualization with Matplotlib, and data organization with Pandas.
* Experienced with a multitude of Hadoop ecosystem components including Hadoop, Hive, Pig, Sqoop, HBase, Cassandra, Spark, Spark Streaming, Spark SQL, Oozie, Zookeeper, MapReduce, Yarn, and Scala.
* Expertise in developing data pipelines using tools like AWS Glue and Apache Airflow for data ingestion, transformation, and loading.
* Skilled in Dimensional Data Modeling with tools like ER/Studio and Erwin, specializing in Star Schema and Snowflake modeling.
* Proficient in SDLC, Agile methodologies (Scrum, Kanban), and Waterfall methodologies, with strong communication, interpersonal, and problem-solving skills.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Hadoop/Big Data Technologies** | Hadoop, Map Reduce, Sqoop, Hive, Oozie, Spark, PySpark, Zookeeper, Kafka, Flume. |
| **Cloud Computing** | AWS - AWS EMR, EC2, Glue, S3, Redshift, Athena, Lambda Functions, Step Functions, Data Bricks (ADBX), Data Lake (ADLS), Cosmos DB, DevOps, Cloud AD, Blob Storage, Data Factory. |
| **ETL Tools** | Talend, Informatica, IBM Infosphere DataStage, SSIS. |
| **NO SQL Database** | HBase, Cassandra, Dynamo DB, Mongo DB. |
| **Databases** | Oracle, MY SQL, Teradata, Neo4j, PL/SQL, NoSQL |
| **Hadoop Distribution** | Horton Works, Cloudera. |
| **Programming & Scripting** | Python, Java, Scala, SQL, PowerShell, Shell Scripting, R. |
| **IDE** | PyCharm, Visual Studio Code, SSMS, Data Studio, IntelliJ. |
| **Monitoring and Reporting** | Tableau, Power Bi, Qlik |
| **Version Control** | GIT, GitHub. |
| **Operating Systems** | Linux, Unix, Mac OS-X, Windows. |
| **Others** | Terraform, H3, Docker, Kubernetes. |

**PROFESSIONAL EXPERIENCE:**

**Client: Ceva Logistics, Remote Sep 2023 - Present**

**Role: SR Data Engineer**

**Responsibilities:**

* Used Python programming and Django for the backend development, Bootstrap and Angular for frontend connectivity and MongoDB for database.
* Fine-tuned Spark applications and provided production support for various pipelines.
* Utilized AWS CloudTrail for comprehensive logging and monitoring of API activity within AWS infrastructure, ensuring detailed visibility into user actions and service usage.
* Led end-to-end data engineering pipelines including ingestion, transformation, and analysis on AWS cloud services.
* Designed and implemented scalable ETL processes using C# to integrate data from diverse sources into centralized data warehouses like Microsoft SQL Server, Oracle, MongoDB.
* Architected and implemented a centralized Data Lake on AWS utilizing core services such as S3, EMR, Redshift, and Athena.
* Successfully migrated datasets and ETL workloads from on-premises environments to AWS cloud services.
* Developed Spark applications and Hive scripts to generate analytical datasets crucial for digital marketing initiatives.
* Automated infrastructure setup, including the provisioning and termination of EMR clusters, ensuring efficient resource management.
* Created Hive external tables on S3 datasets and developed Hive scripts for generating aggregated datasets for downstream analysis.
* Designed and implemented real-time streaming pipelines using Kafka, Spark Streaming, and Redshift for immediate data processing needs.
* Implemented AWS CloudTrail for proactive security monitoring, compliance auditing, and detection of unauthorized activities within AWS resources.

**Environment**: Python, PySpark, Django, NumPy, SciPy, Matplotlib, MySQL, Pandas, REST API framework, Jenkins, Git, Javascript.

**Client: L&T Mindtree, Banglore, India Jan 2019 – Aug 2023**

**Role: Sr. Data Engineer**

**Responsibilities:**

* Designed and implemented highly scalable data architectures on AWS, utilizing services such as S3, Glue, EMR, Redshift, and Lambda to optimize data storage, processing, and analytics workflows.
* Worked on migrating Oracle E-business suite to AWS.
* Developed and implemented PySpark applications for distributed data processing, leveraging Apache Spark to handle large-scale data sets efficiently.
* Designed and optimized PySpark jobs to process structured, semi-structured, and unstructured data formats (JSON, Parquet, CSV, Avro) from diverse sources.
* Developed and maintained ETL pipelines using AWS Glue and PySpark for efficient data extraction, transformation, and loading into data lakes or warehouses.
* Developed PySpark applications for streaming data from sources like Apache Kafka, AWS Kinesis, and Apache Flume, enabling real-time analytics.
* Managed Amazon RDS instances for hosting relational databases such as MySQL, PostgreSQL, Oracle, SQL Server, and MariaDB, ensuring scalability and performance.
* Designed and implemented Amazon VPCs for secure network environments, including IP address ranges, subnets, route tables, and security groups.
* Established VPC peering connections and configured Transit Gateway for seamless communication between AWS resources and on-premises networks.
* Configured Amazon RDS parameters, instance types, storage options, backups, and maintenance windows for optimized database performance and availability.
* Developed automated scripts for processing CloudTrail logs and setting up alerts for critical events, ensuring timely responses.
* Implemented read replicas in Amazon RDS for distributing read traffic and improving database performance.
* Leveraged AWS services like Athena and Redshift Spectrum for ad-hoc querying and analysis of data stored in S3.
* Implemented real-time data processing solutions using AWS Kinesis for ingestion and analysis of high-velocity data streams.
* Utilized AWS Lambda for serverless computing to build event-driven data processing workflows.
* Created and managed AWS SNS topics for facilitating communication and message distribution among AWS components.
* Implemented message queuing systems using Amazon SQS for decoupling and scaling distributed systems.
* Designed serverless workflows with AWS Step Functions to automate complex business processes.
* Integrated Step Functions with AWS Lambda, S3, DynamoDB, and other services for seamless orchestration.
* Implemented comprehensive monitoring and alerting using AWS CloudWatch to proactively detect operational issues.
* Developed CI/CD pipelines for automated deployment and version control of data engineering workflows using Jenkins and Git.
* Designed infrastructure-as-code (IaC) solutions with Terraform to provision and manage cloud infrastructure resources on AWS.
* Implemented Terraform configurations for defining cloud resources like virtual machines, networking components, and storage buckets.
* Utilized Terraform state management features for tracking and maintaining infrastructure state across distributed teams.
* Conducted performance tuning and optimization of Spark jobs using techniques like partitioning, caching, and broadcast joins.
* Implemented data governance policies and data quality checks to ensure compliance and maintain data integrity.
* Contributed to Agile processes including Scrum meetings, sprint planning, and backlog refinement sessions.

**Environment**: Python, PySpark, Tableau, Apache Spark, PySpark, GIT, AWS services (EMR, Redshift, EC2, S3, Glue, Cloud watch, cloud trail, SNS, DynamoDB), Glue, Oracle E-business.

**Client:HDFC, Banglore, India Jan 2016 – Dec 2018**

**Role: Data engineer**

**Responsibilities:**

* Developed Spark Scala scripts and UDFs within AWS Databricks to read from AWS S3 storage and perform transformations on large datasets.
* Created scalable data ingestion pipelines on AWS EMR Spark cluster using Spark SQL within AWS Databricks. Also, collaborated with Amazon DynamoDB.
* Built a robust ETL pipeline in AWS Glue to integrate data from both on-premises (MySQL, Cassandra) and cloud (S3), applying transformations using PySpark to load enriched data into Amazon Redshift.
* Familiar with the Databricks notebook environment, including features such as cell execution, markdown documentation, and collaborative editing, facilitating efficient team collaboration and knowledge sharing.
* Configured Spark Streaming within AWS Databricks to receive real-time data from Apache Flume and store the stream data in Amazon DynamoDB, leveraging Scala for optimization and performance improvement.
* Designed pipelines in AWS Glue using linked services to extract, transform, and load data from multiple sources like Amazon RDS, S3, and Amazon Redshift.
* Implemented custom alerts using Kusto Query Language within Amazon CloudWatch.
* Extracted, transformed, and loaded data from source systems to AWS Data Storage services using a combination of AWS Glue and AWS EMR.
* Designed and implemented new AWS Accounts, AWS Glue jobs, Amazon EC2 instances, Amazon RDS instances, EMR clusters, and installed DMGs on EC2 instances to connect to on-premises servers.
* Developed Spark SQL Data Frames within AWS Databricks to apply business transformations and data cleansing operations, ensuring data quality.
* Developed Python scripts to build ETL pipeline and Directed Acyclic Graph (DAG) workflows in Apache Airflow, Apache NiFi.
* Designed custom-built input adapters using Spark and Hive to ingest and analyze data in Apache Airflow, ingesting enriched data into Snowflake.
* Leveraged Databricks notebooks to create interactive and collaborative environments for data exploration and analysis, enabling seamless integration with Spark SQL for querying and transforming large datasets.
* Managed resources and scheduling across the cluster using Amazon Elastic Kubernetes Service (EKS) for handling online and batch workloads.
* Utilized AWS CodePipeline for CI/CD, leveraging IAM for authentication and AWS IAM policies for authorization.
* Leveraged Scala's concurrency support for developing map reduces jobs, compiling the program code into bytecode for the JVM for data processing.
* Developed and implemented Amazon CloudWatch solutions for log analytics, monitoring, and real-time data analysis.
* Designed and optimized Amazon CloudWatch data ingestion pipelines using various data ingestion methods such as AWS Glue, Kinesis Data Firehose, and AWS Lambda.
* Developed complex CloudWatch Logs Insights queries to extract insights from large datasets, including time-series data.
* Actively involved in code reviews and pull requests within AWS CodeCommit, ensuring code quality and adherence to best practices.
* Configured Amazon CloudWatch cluster settings for high availability and disaster recovery, including cross-region replication and backup policies.
* Utilized CloudWatch Logs Insights queries and published data for interactive Amazon QuickSight dashboards and reporting purposes based on business requirements.

**EDUCATION DETAILS :**

* Masters in Information Systems, PACE University– USA,
* Masters in Computer Application , PSNA College- India