**A blue hexagon with white text

Description automatically generatedName: Anvita**

**Email:** [**Hr@seafysoft.com**](mailto:Hr@seafysoft.com)

**Phone: +19105576339**

**AWS Big Data Engineer**

***Highly skilled AWS Big Data Engineer with over 8 years of experience designing and implementing advanced solutions at the nexus of data engineering and cloud computing. Proficient in leveraging AWS services to architect and deploy scalable big data solutions that drive business growth and innovation. Specialized expertise in data processing, analytics, and machine learning on the AWS platform. Adept at utilizing technologies such as Amazon EMR, Redshift, Glue, Athena, and S3 to extract actionable insights from vast datasets. Proven ability to collaborate effectively with diverse teams to deliver high-impact projects that advance organizational objectives.***

**PROFESSIONAL SUMMARY:**

* Contributed significantly across all stages of the **Software Development Life Cycle**, specializing in Database Architecture, Logical and Physical modeling, and Data Warehouse/ETL development.
* Managed end-to-end **database lifecycle**, providing full lifecycle support for logical and physical database design, schema management, and deployment, with a keen focus on strict configuration management and effective team coordination.
* Designed and developed robust, scalable applications using **Object-Oriented Programming (OOP)** principles in **Java and .NET**, implementing design patterns like Singleton, Factory, and Observer to ensure modularity, reusability, and maintainability in enterprise-level software solutions.
* Implemented **Data Warehouse solutions** in Confidential **Redshift**, orchestrating seamless data migration from on-premise databases to Confidential Redshift, RDS, and S3, enhancing data accessibility and scalability.
* Demonstrated prowess in **Big Data analytics** and manipulation, leveraging the Hadoop ecosystem's comprehensive suite of tools including **MapReduce, HDFS, Pig, Hive, HBase, Spark, Kafka, Flume, Sqoop, and Avro**, driving actionable insights from vast and varied datasets.
* Proficient in managing diverse databases including **MongoDB, Cassandra, MySQL, Oracle, and MS SQL Server,** with expertise in handling a multitude of file formats such as **delimited files, Avro, JSON, and Parquet**.
* Orchestrated **Docker** container environments using **ECS, ALB, and Lambda** for streamlined deployment and scalability.
* Architected **Snowflake Schemas**, optimizing data organization by normalizing dimension tables and crafting Sub Dimension subsets like Demographic, empowering advanced analytics, and business intelligence capabilities.
* Developed robust data pipelines leveraging **AWS services** such as **EC2, S3, Redshift, Glue, Lambda functions, Step functions, CloudWatch, and DynamoDB, automating data processing workflows** and ensuring data integrity and reliability.
* Leveraged Azure services such as **Azure Data Factory, Azure Data Lake Storage (ADLS) Gen 2, Azure SQL Database, Azure Synapse Analytics, Azure Functions, Azure Logic Apps, Azure Monitor, and Azure Cosmos DB** to automate data processing workflows, ensuring data integrity and reliability.
* Utilized **Kubernetes** and **Docker** for **CI/CD** system runtime environments, optimizing build, test, and deployment processes for enhanced efficiency and scalability, with a track record of successfully creating and running Docker images with multiple microservices.
* Leveraged **MS SQL Server 2012/2008R2/2008** and **Oracle 11g/10g** for database management, demonstrating proficiency in ETL solutions and Analytics Applications development.
* Proficient in leveraging **SAS** utilities such as **SAS Data Integration Studio, SAS Macro, and SAS Enterprise Guide**, alongside SAP functionalities like **SAP BW/4HANA, SAP HANA, and SAP Data Services**, to architect and optimize complex data pipelines, automate ETL workflows, and streamline large-scale data integration processes, ensuring efficient data management, high-performance analytics, and seamless integration within SAP environments.
* Led initiatives within the **Global Data Assurance Committee (GDAC)** to establish and enforce comprehensive data governance frameworks, ensuring compliance with industry regulations like **GDPR, HIPAA, and CCPA**, while mitigating risks associated with data breaches and unauthorized access.
* Implemented ETL workflows using **Matillion**, leveraging its user-friendly interface and powerful data transformation capabilities to streamline data integration and enhance overall data pipeline efficiency.
* Implemented data governance frameworks using **Apache Atlas and Collibra**, enabling metadata management, lineage tracking, and data lineage visualization.
* Excelled in **Unix/Linux systems**, adeptly crafting data pipelines and scripting for seamless data processing and management.
* Designed, developed, and delivered business intelligence solutions using **SQL Server Data Tools, SQL Server Integration Services (SSIS), and Reporting Services (SSRS)**.
* Demonstrated strong SQL development skills, proficient in writing **Stored Procedures, Triggers, Views, and User Defined functions** for efficient data manipulation and retrieval.
* Utilized analytical applications such as **R, SPSS, Rattle, and Python** to derive actionable insights from data, translating findings into risk management and marketing strategies that drive value.
* Extensively employed distributed computing architectures including **AWS products, Hadoop, Python, Spark, and Azure SQL Database** to tackle big data challenges, optimizing performance and scalability.
* Developed and deployed enterprise-based applications leveraging major Hadoop ecosystem components such as **MapReduce, YARN, Hive, HBase, Flume, Sqoop, Spark MLlib, Spark GraphX, Spark SQL, and Kafka.**
* Proficient in **Spark Core, Spark SQL, Spark MLlib, Spark GraphX, and Spark Streaming**, harnessing in-memory computing capabilities to process and transform complex data effectively.
* Extensively used Sqoop to migrate data between **RDBMS, NoSQL databases, and HDFS**, demonstrating expertise in Extraction, Transformation, and Loading (ETL) processes.
* Developed customized UDFs in Python to extend **Hive and Pig Latin** functionality, facilitating seamless data processing and integration.
* Developed **RESTful APIs** and microservices using frameworks like **Spring Boot and Flask** for exposing data services and integrating with external systems.
* Demonstrated excellent communication skills and thrived in fast-paced, multitasking environments, successfully working both independently and collaboratively in cross-functional teams.
* Developed Spark applications in **Python (PySpark)** on distributed environments, loading massive CSV files with different schemas into **Hive ORC** tables.
* Developed custom machine learning models and algorithms using frameworks like **TensorFlow and PyTorch** for solving complex data analysis and prediction tasks.
* Led cross-functional teams in the design and deployment of large-scale data architectures, integrating advanced technologies such as Hadoop, Spark, and AWS to optimize data processing workflows and improve analytics capabilities across the organization.
* Implemented data virtualization solutions using tools like **Denodo and AtScale** for federated data access and integration across heterogeneous data sources and platforms.
* Championed **test-driven** development methodologies, ensuring software quality and reliability through rigorous testing approaches, and collaborated seamlessly with cross-functional teams to achieve project objectives.

|  |  |
| --- | --- |
| **Big Data Ecosystem** | HDFS, MapReduce, HBase, Pig, Hive, Sqoop, Kafka,  Flume, Cassandra, Impala, Oozie, Zookeeper, Map R, Amazon Web Services (AWS), EMR |
| **Machine Learning Classification Algorithms** | Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor (KNN), Gradient Boosting Classifier, Extreme Gradient Boosting Classifier, Support Vector Machine (SVM), Artificial Neural Networks (ANN), Naïve Bayes Classifier, Extra Trees Classifier, Stochastic Gradient Descent, etc. |
| **Cloud Technologies** | AWS, Azure, Google cloud platform (GCP) |
| **IDE’s** | IntelliJ, Eclipse, Spyder, Jupyter |
| **Ensemble and Stacking** | Averaged Ensembles, Weighted Averaging, Base Learning, Meta Learning, Majority Voting, Stacked Ensemble, Auto ML – Scikit-Learn, MLjar, etc. |
| **Databases** | Redshift, Oracle, MySQL, DB2, MS-SQL Server, HBASE |
| **Programming / Query Languages** | Java, .NET, SQL, Python Programming (Pandas, NumPy, SciPy, Scikit-Learn, Seaborn, Matplotlib, NLTK), NoSQL, PySpark, PySpark SQL, SAS, R Programming (Caret, Glmnet, XGBoost, rpart, ggplot2, sqldf), RStudio, PL/SQL, Linux shell scripts, Scala. |
| **Data Engineer/Big Data Tools / Cloud / Visualization / Other Tools** | Databricks, Hadoop Distributed File System (HDFS), Hive, Pig, Sqoop, MapReduce, Spring Boot, Flume, YARN, Hortonworks, Cloudera, Mahout, MLlib, Oozie, Zookeeper, etc. AWS, Azure Databricks, Azure Data Explorer, Azure HDInsight, Salesforce, GCP, Google Shell, Linux, PuTTY, Bash Shell, Unix, etc., Tableau, Power BI, SAS, We Intelligence, Crystal Reports, Dashboard Design. |

**TECHNICAL SKILLS:**

**PROFESSIONAL EXPERIENCE:**

**Role: Lead/Sr AWS Big Data Engineer April 2023 - Present**

**Client: Brivo, Bethesda, Maryland.**

* Developed **Python** scripts for data extraction, transformation, and loading ETL processes, improving efficiency and accuracy.
* Utilized **HDFS (Hadoop Distributed File System)** for distributed storage and processing of large datasets, ensuring fault tolerance and scalability in data processing workflows.
* Automated all the jobs starting from pulling the Data from different Data Sources like **MySOL** and pushing the result dataset to **Hadoop Distributed File System** and running **MapReduce** jobs and **PIG**/**Hive** using **Oozie** (Workflow management).
* Validated the test data in DB2 tables on Mainframes and on **Teradata** using **SQL queries**.
* Performed query optimization and indexing in **SQL databases** to enhance performance and reduce query execution time.
* Built and maintained **ETL** processes using **SQL scripts**, ensuring accurate and timely data migration between systems.
* Created automated **SQL-based** reports and dashboards for business intelligence, providing stakeholders with real-time data insights.
* Processed and exchanged data in **JSON (JavaScript Object Notation**) format for interoperability and compatibility with various systems and services, ensuring seamless integration and data interchange.
* Implemented data streaming solutions using **Apache Kafka** for real-time data ingestion, processing, and event-driven architecture, enabling scalable and fault-tolerant data pipelines.
* Deployed and managed **Apache Spark** clusters on **Databricks** for distributed data processing, providing high performance and scalability for big data analytics tasks.
* Worked on Dimensional and Relational Data Modelling using Star and Snowflake Schemas, **OLTP/OLAP** system, Conceptual, Logical and Physical data modelling using **Erwin**.
* Conducted data profiling and wrangling tasks on diverse data formats including **XML, Web feeds, and files** using Python, Unix, and SQL, ensuring data integrity and quality throughout the ETL process.
* Implemented **Sqoop** for incremental data processing, reading from **DB2** and loading into **Hive tables**, optimizing data handling and storage efficiency.
* Connected **Hive server** to **Tableau** for generating interactive reports, enabling stakeholders to make informed decisions based on real-time data insights.
* Automated data ingestion processes using **AWS Glue crawlers**, reducing manual intervention, and increasing data catalog accuracy.
* Orchestrated data flows and transformations using **Apache NiFi**, facilitating data routing, transformation, and enrichment with built-in processors and intuitive graphical interface.
* Utilized **Databricks Delta Lake** to implement **ACID transactions** and schema enforcement, enhancing the reliability and consistency of big data workflows and facilitating efficient data versioning and auditing.
* Leveraged **Databricks MLflow** for tracking experiments, managing models, and facilitating the deployment of machine learning models into production environments.
* Implemented custom Python and SQL scripts within **Matillion** to perform advanced data transformations, enhancing the flexibility and functionality of ETL jobs.
* Employed **Matillion's metadata management features** to track data lineage and ensure transparency and compliance with data governance policies.
* Designed and optimized **HVR ETL** replication pipelines to enable low-latency, real-time data replication from on-premises Microsoft SQL Server and Oracle databases into Snowflake, leveraging advanced **Change Data Capture (CDC)** techniques to ensure data integrity and synchronization.
* Designed, developed, and optimized complex ETL workflows using **IBM DataStage**, employing parallel processing techniques and advanced data transformation logic to handle large-scale data integration tasks efficiently and meet complex business requirements.
* Developed and optimized intricate ETL workflows within **SAS Data Integration Studio**, leveraging advanced transformation logic, parallel processing, and optimization techniques to handle large-scale, multi-source data integration tasks, ensuring minimal latency and high throughput.
* Leveraged **SAS Macro language** to automate complex, repetitive ETL processes, incorporating dynamic code generation, parameterization, and conditional logic to enhance the scalability, reusability, and efficiency of data processing pipelines, thereby reducing runtime and resource consumption.
* Optimized **Spark jobs** in **Databricks** by tuning configurations, partitioning data, and leveraging advanced Spark features, improving performance and resource utilization.
* Leveraged **Fivetran’s** automated data extraction and load capabilities to enhance the efficiency of data ingestion workflows, reducing manual intervention and ensuring seamless integration with Snowflake for large-scale, high-volume datasets.
* Enhanced data quality by implementing data validation and transformation rules within **AWS Glue** ETL jobs.
* Employed **Amazon S3** lifecycle policies to manage data retention and reduce storage costs by archiving infrequently accessed data.
* Designed and implemented custom monitoring and logging solutions using **Amazon CloudWatch** and **AWS CloudTrail** to track pipeline performance and troubleshoot issues effectively.
* Deployed infrastructure as code using **AWS CloudFormation** and **Terraform**, ensuring reproducibility and consistency across development and production environments.
* Leveraged **AWS IAM** roles and policies to enforce fine-grained access control, ensuring secure access to sensitive data and services.
* Optimized **Redshift** cluster performance by using distribution keys, sort keys, and query optimization techniques, resulting in faster query execution times.
* Conducted data partitioning and compression in **Amazon S3** and **Amazon Redshift** to improve storage efficiency and query performance.
* Integrated third-party data sources and **APIs** into the data pipeline, enriching the data set and expanding analytics capabilities.
* Designed and implemented a real-time data processing pipeline using **Amazon Kinesis**, **AWS Lambda**, and **Amazon DynamoDB**, achieving sub-second latency for data ingestion and processing.
* Developed a scalable data lake solution on Amazon S3, utilizing **AWS Glue** for data cataloging and ETL operations, and **Amazon Athena** for querying large datasets with SQL.
* Integrated **Aurora** seamlessly with AWS services such as Lambda for serverless computing and AWS Glue for ETL operations, enhancing overall data pipeline efficiency and reliability.
* Orchestrated batch ETL workflows using AWS Data Pipeline and processed large datasets with **Amazon EMR (Hadoop/Spark)**, significantly improving data processing efficiency.
* Built a serverless data processing pipeline with **AWS Lambda** and **AWS Step Functions**, reducing infrastructure management overhead and improving scalability.
* Established a robust data warehousing solution with **Amazon Redshift**, loading and transforming data using AWS Glue, and optimizing complex queries for business analytics.
* Created interactive dashboards and visualizations in Amazon **QuickSight**, providing real-time and historical insights to stakeholders.
* Implemented data storage and management strategies in **Amazon S3** for raw, processed, and final datasets, ensuring data integrity and availability.
* Configured **Amazon SNS** to send notifications and alerts for real-time monitoring of data pipeline statuses and performance.
* Employed **AWS EKS (Elastic Kubernetes Service**) to orchestrate and manage containerized applications, ensuring high availability, scalability, and efficient resource utilization. Configured EKS clusters with advanced networking (CNI) plugins, IAM roles for service accounts, and integrated with **AWS Fargate** for serverless compute.
* Developed and deployed machine learning models using **AWS SageMaker** to provide predictive analytics capabilities integrated into the data pipeline.
* Managed and optimized relational databases such as **Amazon RDS**, ensuring high availability, performance, and data consistency.
* Integrated **Artificial Intelligence** driven analytics into ETL processes, leveraging machine learning models for anomaly detection, predictive maintenance, and data quality enhancement, ensuring proactive insights and improved decision-making.
* Implemented **Continuous Integration and Continuous Deployment (CI/CD) pipelines** using tools like **Jenkins**, **AWS CodePipeline**, and **GitLab CI**, enabling automated testing and deployment of data pipeline updates.
* Developed and managed data integration workflows using **SnapLogic**, ensuring seamless and efficient data movement across systems.
* Containerized data pipeline components and dependencies using **Docker**, improving portability and reproducibility across different environments.
* Orchestrated containerized applications using **Kubernetes** for efficient deployment, scaling, and management of data processing workloads.
* Developed automated test scripts for continuous testing within **CI/CD pipelines**, reducing deployment risks and ensuring consistent application quality.
* Collaborated with cross-functional teams as part of the **Global Data Assurance Committee (GDAC)** to audit and optimize data handling processes, enhancing data integrity, privacy, and security across enterprise systems, and driving continuous improvements in data assurance practices.

**Environment:** Python, SQL, Hadoop, MapReduce, HDFS, Pig, Hive, HBase, Spark, Kafka, Flume, Sqoop, Avro, MongoDB, Cassandra, MySQL, Oracle, MS SQL Server, Amazon RDS, DB2, Teradata, Amazon Redshift, Snowflake, Azure SQL Data Warehouse, JSON, Avro, Parquet, CSV, XML, Erwin, Star Schema, Snowflake Schema, Databricks, AWS, Azure, EC2, S3, Redshift, Glue, Lambda, Step Functions, CloudWatch, SNS, DynamoDB, SQS, EMR, ElasticSearch, IAM, CloudFormation, Terraform, CloudTrail, Athena, Kinesis, QuickSight, SageMaker, Data Pipeline, Matillion, DBT, Glue Crawlers, Microsoft SSIS, Apache NiFi, Oozie, Docker, Kubernetes, Jenkins, AWS CodePipeline, GitLab CI, Tableau, Power BI, Amazon QuickSight, Pandas, NumPy, Boto3, Apache NiFi, AWS Step Functions, Amazon Kinesis, Amazon CloudWatch, AWS CloudTrail, Sqoop, API Integrations, Python UDFs, AWS Glue ETL Jobs.

**Role: Sr AWS Big Data Engineer August 2022 – April 2023**

**Client: HTD Health, Brooklyn, NY.**

* Evaluated the suitability of Hadoop and its ecosystem for the project by implementing and validating various **proof of concept (POC**) applications, ultimately adopting them for the Big Data Hadoop initiative.
* Estimated software and hardware requirements for the NameNode and DataNodes in the **Hadoop cluster**, ensuring optimal performance and scalability.
* Optimized **Spark** jobs by tuning Spark configurations and partitioning strategies, significantly reducing execution time and resource consumption.
* Developed machine learning models using **PySpark** and integrated them into the data pipeline for predictive analytics and data-driven decision-making.
* Optimized **machine learning models** using advanced techniques in **Python and scikit-learn**, including **grid search** for hyperparameter optimization, **k-fold cross-validation** for robust model validation, and feature engineering to enhance model performance. Deployed these models into production using **Docker and Kubernetes**, leading to scalable and efficient predictive analytics systems.
* Employed container orchestration with **Kubernetes** for deploying and managing containerized data processing applications, ensuring scalability and resilience.
* Utilized **Terraform** and AWS CloudFormation for **infrastructure as code (IaC**) deployments, ensuring consistent and repeatable environment setups.
* Designed, developed, and maintained a highly scalable and efficient **Operational Data Store (ODS)** architecture, integrating data from disparate sources using **ETL** processes to provide a consolidated and consistent data view.
* Migrated existing databases from on-premises to **AWS Redshift** using AWS services like **DMS, S3, and Glue**, enhancing data accessibility and cloud integration.
* Developed PySpark code for **AWS Glue jobs and EMR**, optimizing data processing workflows for efficiency and scalability.
* Installed and configured **Hadoop MapReduce and HDFS** and developed multiple **MapReduce** jobs in **Java and Scala** for data cleaning and preprocessing.
* Configured **Aurora** clusters for optimal performance and scalability, integrating seamlessly with existing AWS services for efficient data processing and analytics workflows.
* Engineered and deployed data-driven solutions using **SAS Viya 4**, utilizing its cloud-native, multi-threaded architecture for high-performance parallel processing, enabling real-time data analytics and machine learning model deployment.
* Integrated **SAS Viya 4 with SAP HANA** for efficient data ingestion, transformation, and analysis, optimizing the workflow for handling large-scale, complex datasets within a unified analytics environment.
* Created Java **MapReduce** programs for log file analysis stored in the Hadoop cluster, facilitating comprehensive data insights.
* Implemented **Spark** with Python and Spark SQL for faster data testing and processing, improving turnaround times for data analytics.
* Developed Spark scripts using Python on **Azure HDInsight** for data aggregation and validation, verifying performance improvements over MapReduce jobs.
* Optimized **ETL/ELT** pipelines using **Azure Data Factory (ADF)** and **Azure Synapse Analytics**, integrating with **Azure Data Lake Storage (ADLS) Gen 2** and **Blob Storage** for scalable and efficient data solutions.
* Led the migration of databases from on-premises to **Azure SQL Database** and **Azure Cosmos DB**, utilizing **Azure Data Factory and Azure Synapse Analytics** to enhance data accessibility and cloud integration.
* Installed and configured **Hadoop MapReduce and HDFS**, developed **MapReduce** jobs in **Java and Scala**, and implemented **Azure Data Lake Storage (ADLS) Gen 2** for advanced data storage solutions.
* Integrated **Apache Flink** for real-time stream processing and used **Azure Event Hubs and Azure Stream Analytics** to handle high-velocity data streams efficiently.
* Regularly imported data using **Sqoop** to load data from MySQL to HDFS, ensuring up-to-date and accurate data storage.
* Converted **Hive** and **SQL** queries into Spark transformations using Spark RDDs, Python, and Scala, optimizing query performance.
* Utilized **IAM** to create new accounts, roles, groups, and policies, developing critical modules for generating Amazon Resource Numbers and integration with **S3, DynamoDB, RDS, Lambda, and SQS Queue**.
* Implemented **Matillion** for ETL processes, leveraging its cloud-native capabilities to design and automate data transformation workflows, and integrating it with cloud data warehouses like **Snowflake** for optimized data management.
* Reviewed SQL explain plans in **Snowflake** to optimize query performance and resource utilization.
* Developed and optimized **ABAP programs** for **SAP** systems, including custom reports, interfaces (IDocs, BAPIs), enhancements (User Exits, BADIs), and forms (SmartForms, SAPscript).
* Implemented advanced data extraction routines using **ABAP Data Dictionary** and ALV reporting, ensuring seamless integration with external systems.
* Integrated **ABAP programs** with **SAP modules like SD, MM, and FI** for streamlined business processes and accurate data synchronization.
* Developed ETL parsing and analytics using Python and Spark to build a structured data model in **Elasticsearch** for **API and UI** consumption.
* Implemented **ETL** processes using **IBM DataStage** to efficiently extract, transform, and load data from various sources into the data warehouse, ensuring data integrity and consistency.
* Created ETL jobs using **Spark-Scala** to migrate data from **Oracle to Cassandra** tables, ensuring data consistency and performance.
* Utilized **SnapLogic**’s data transformation tools to standardize and cleanse data, ensuring high-quality and consistent data for downstream analytics.
* Employed **DBT (Data Build Tool)** to transform data in the warehouse, enabling efficient, version-controlled transformations and facilitating modular and reusable SQL-based data transformation workflows.
* Utilized **Spark-Scala (RDDs, DataFrames, Spark SQL**) and **Spark-Cassandra-Connector APIs** for tasks such as data migration and business report generation.
* Integrated **Apache Flink** for real-time data stream processing, enhancing the capability to handle high-velocity data streams.
* Created partitions and buckets based on state for efficient processing using bucket-based **Hive joins**.
* Enforced **HIPAA** compliance through the implementation of advanced encryption algorithms, granular IAM policies, and audit trails, ensuring the secure handling of PHI across data processing and storage systems.
* Implemented **FHIR** standards for interoperable healthcare data exchange, ensuring seamless integration and secure communication between diverse healthcare systems.
* Implemented security measures to meet PCI requirements, using **VPC Public/Private subnets**, **Security Groups**, **NACLs, IAM roles and policies, VPN, WAF, Trust Advisor, and CloudTrail**, successfully passing penetration tests.
* Defined job workflows and dependencies using **Oozie**, ensuring streamlined and efficient task management.
* Played a key role in productionizing the application after thorough testing by BI analysts, ensuring a seamless transition to production.

**Environment:** Hadoop ecosystem, AWS (Amazon Web Services), Kubernetes, Terraform, Azure HDInsight, Oracle, Cassandra, Elasticsearch, Apache Flink, VPC (Virtual Private Cloud), VPN (Virtual Private Network), WAF (Web Application Firewall), Trust Advisor, CloudTrail.

**Role: Data Engineer January 2020 – July 2022**

**Client: Pitney Bowes, Charlotte, NC**

* Deployed and managed multiple applications leveraging a comprehensive suite of AWS services including **EC2, Route53, S3, RDS, DynamoDB, SNS, SQS, and IAM**, focusing on high availability, fault tolerance, and auto-scaling using AWS CloudFormation.
* Implemented **Infrastructure as Code (IaC)** principles to automate the provisioning and configuration of AWS resources, leveraging AWS CloudFormation templates and AWS CLI commands.
* Leveraged **EC2 instances, S3 buckets, RDS databases, EBS volumes, Elastic Load Balancers, Auto Scaling groups, VPCs, and CloudFormation** stacks to architect scalable, resilient, and performant systems on AWS.
* Employed **AWS WAF (Web Application Firewall)** to protect web applications from common security threats, such as SQL injection and **cross-site scripting (XSS**), by filtering and monitoring HTTP traffic.
* Implemented **AWS Shield** to mitigate **distributed denial of service (DDoS)** attacks, safeguarding applications and services against malicious traffic and ensuring uninterrupted availability.
* Utilized **AWS Step Functions** to orchestrate complex workflows and automate business processes, coordinating the execution of multiple AWS services in a serverless environment.
* Developed **Ansible scripts and YAML configurations** for automated configuration management and deployment of software applications on AWS infrastructure.
* Engineered **Chef Recipes, Puppet Manifests, Salt Stack pillars, and states** to define and manage infrastructure configurations, enabling consistent and reproducible deployments.
* Implemented **Gradle build scripts** to compile Java applications, run unit tests, and package artifacts into **deployable EAR files**, facilitating version-controlled software releases.
* Orchestrated deployment workflows using custom **Ansible Playbooks**, automating server provisioning, software installations, and configuration updates across distributed environments.
* Established a Continuous Delivery pipeline with **Docker containers, Jenkins CI/CD server, GitHub version control, and AWS AMIs**, enabling automated build, test, and deployment processes.
* Configured **Azure services** equivalent to AWS services being migrated, such as **Azure VMs, Azure DNS, Azure Blob Storage, Azure SQL Database, Cosmos DB, Azure Notification Hubs, Azure Service Bus, and Azure Active Directory (AAD).**
* Implemented **Azure Resource Manager (ARM) templates** and **Azure CLI scripts** to automate the provisioning and configuration of Azure resources, adhering to Infrastructure as Code (IaC) principles.
* Utilized **Azure DDoS** Protection to safeguard applications and services against **distributed denial of service (DDoS)** attacks, ensuring uninterrupted availability.
* Orchestrated complex workflows and automated business processes using **Azure Logic Apps** and **Azure Functions**, replacing **AWS Step Functions** in a serverless environment.
* Employed **Talend Big Data components** to process large datasets efficiently, leveraging **Azure HDInsight, Azure Databricks, and Azure Synapse Analytics** to optimize data workflows.
* Configured and maintained **SAP BW/4HANA** data warehouse solution to optimize data processing and analytics capabilities, ensuring seamless integration with existing AWS infrastructure and data workflows.
* Employed **Talend Big Data components** to process large datasets efficiently, leveraging technologies like Hadoop, Spark, and HDFS to optimize data workflows.
* Automated data integration workflows using **Talend Management Console (TMC)** to schedule, monitor, and manage **ETL and ELT jobs**, ensuring reliable and timely data processing and integration across diverse data sources.
* Utilized **Git and GitHub** for version control management, employing branching strategies, tagging releases, and performing code reviews to ensure codebase integrity and collaboration.
* Conducted version upgrades and migration of **Jenkins** CI/CD infrastructure, ensuring minimal disruption to development workflows and maximizing system performance and reliability.

**Environment:** AWS CloudFormation, Ansible, Chef, Puppet, Salt Stack, Gradle, GitHub, Docker, Jenkins, AWS AMIs, Git, AWS WAF, AWS Shield, AWS Step Functions.

**Role: Data Analyst April 2018 – December 2019 Client: AppiVa Software, Bengaluru, India**

* Developed applications with scalability as a primary focus, ensuring they can easily adapt to future modifications and handle increasing data volumes seamlessly.
* Engineered cloud-based solutions to optimize big data processing performance, handling high volumes of data to provide superior customer support. Utilized technologies like **Hive, Spark, Scoop, NIFI, Python, Bash scripting, and Apache Airflow** for job scheduling in GCP environments.
* Successfully migrated **Oracle SQL Enterprise License (EL)** to run on Google Cloud Platform using **Cloud Dataproc** and **BigQuery**, with Cloud **Pub/Sub** for job triggering, enhancing scalability and efficiency.
* Leveraged **Presto, Hive, Spark-SQL, and BigQuery** with Python client libraries to develop high-performance analytics programs, optimizing data processing and analysis workflows.
* Leveraged machine learning interpretability techniques such as **SHAP (SHapley Additive exPlanations)** values to understand the underlying factors driving recommendation outcomes.
* Conducted **cohort analysis** to segment users based on behavior or demographics, enabling targeted recommendation strategies for different user groups.
* Employed **Alteryx** to streamline data preparation, blending, and analytics workflows, enhancing the efficiency and accuracy of data processing tasks.
* Demonstrated proficiency in utilizing all major big data services offered by **Google Cloud Platform**, ensuring comprehensive utilization of cloud resources for efficient data processing and analysis.
* Leveraged **Apache Airflow** within **GCP** Composer environment to construct robust data pipelines, employing various Airflow operators such as **Bash, Hadoop, Python callable**, and branching operators for streamlined workflow orchestration.
* Developed monitoring reports to track data loads into GCP, enhancing reliability and performance at the site level, and facilitating proactive maintenance and troubleshooting.
* Implemented serverless data processing pipelines using **Google Cloud Dataflow** to handle complex data transformations and enrichment tasks, ensuring scalability and cost-efficiency.
* Utilized **Google Cloud AutoML** for automating the development of machine learning models, accelerating the iteration and improvement of recommendation algorithms.
* Deployed recommendation models using **TensorFlow** Serving on **Google Kubernetes Engine (GKE**) to provide scalable and low-latency model serving capabilities, enabling real-time inference for personalized recommendations.
* Implemented data replication and integration solutions using **Attunity (Qlik Replicate)** to ensure real-time data synchronization across heterogeneous systems, enhancing data availability and consistency for analytics and reporting.
* Utilized **Informatica PowerCenter** to design and implement complex ETL workflows, ensuring data accuracy and integrity across various systems and platforms.
* Employed **Informatica Data Quality (IDQ)** tools to enhance data governance, implementing data profiling, cleansing, and validation processes to ensure high data quality standards.
* Leveraged **Informatica Big Data Management (BDM)** to optimize big data processing and integration tasks, ensuring efficient handling of large data volumes in a cloud environment.
* Integrated **Informatica with GCP services such as BigQuery and Cloud Storage** to enhance data processing and analytics capabilities, enabling seamless data flow across hybrid environments.
* Engineered and optimized **RESTful APIs** using **Django**, ensuring efficient data serialization and deserialization with **Django Rest Framework (DRF)**. Leveraged Django's **ORM** for complex query optimization and database schema management, resulting in improved application performance and reduced latency.
* Collaborated with **FinCPDRes** service delivery partners to integrate business-embedded technology solutions, ensuring seamless alignment with enterprise architecture and service-oriented design principles.
* Integrated **Google Analytics** with data pipelines to capture and analyze website traffic data, enabling insights into user behavior, engagement metrics, and conversion funnels for data-driven decision-making and optimization of digital marketing strategies.
* Developed and deployed interactive **Power BI dashboards**, transforming complex datasets into actionable insights by leveraging advanced **DAX expressions**, data modeling, and custom visuals.

**Environment:** Google Cloud Platform (GCP), Google Analytics, Apache Airflow, Presto, Hive, Spark, BigQuery, Cloud Dataproc, Cloud Pub/Sub, Google Kubernetes Engine (GKE), Google Cloud Dataflow, Google Cloud AutoML, Alteryx.

**Role: Data Analyst October 2016 – March 2018**

**Client: StoneX Group, Bengaluru, India**

* Collaborated with cross-functional teams including data engineers and operations to implement the **ETL process**, optimizing **SQL** queries for efficient data extraction in alignment with analytical requirements.
* Queried and retrieved data from **Oracle database** servers to obtain the dataset for analysis.
* Designed Data warehouse schemas like **star schema and snowflake** by using **ERWIN.**
* Utilized **T-SQL** for writing complex queries, stored procedures, and triggers, to enhance data retrieval and manipulation processes within **SQL Server databases**.
* Executed **complex SQL queries** for data manipulation, extraction, and reporting, optimizing query performance and ensuring data integrity across various relational databases.
* Developed and managed **PostgreSQL databases,** implementing advanced features such as indexing, partitioning, and replication to ensure high availability and performance.
* Utilized Pandas for preprocessing, including handling missing data, and balanced the dataset through **Over-sampling and Under-sampling** techniques.
* Employed **PCA** and other feature engineering methods, along with **Scikit-learn** preprocessing techniques, to reduce high-dimensional data derived from patient visit history and comorbidity flags.
* Conducted data exploration using correlation analysis and graphical techniques in **Matplotlib and Seaborn** to gain insights into patient admission and discharge patterns.
* Integrated **Django** with advanced third-party tools like **Celery** for distributed task processing and **Redis** for in-memory data caching, enhancing application scalability and fault tolerance. Utilized **Django's** middleware for security hardening, including **CSRF protection**, **XSS mitigation**, and robust user authentication mechanisms.
* Developed advanced visualizations and custom plots using **ggplot** in R, incorporating layered graphics, aesthetic mappings, and statistical transformations to effectively communicate complex data insights and trends.
* Experimented with various predictive models such as **Logistic Regression, Support Vector Machine (SVM), Gradient Boosting, and Random Forest** using Python's Scikit-learn library to predict patient readmission.
* Implemented Cross-validation and statistical tests like **ANOVA and Chi-square** to assess the significance of the models.
* Implemented **Informatica** PowerCenter to design and manage complex ETL workflows, optimizing data integration processes and ensuring data quality across multiple sources.
* Deployed and tested the model on **AWS EC2**, optimizing algorithms and parameters for improved performance.
* Established a data preprocessing pipeline to ensure consistency between training data and new incoming data.
* Deployed the model on **AWS Lambda**, collected feedback post-deployment, and iteratively retrained and adjusted parameters to enhance performance.
* Contributed to the design and implementation of secure, scalable data architectures by applying deep technical knowledge of **CoP** Enterprise data strategies, security frameworks, and platform-specific protocols.
* Skilled in creating interactive web-based dashboards and reports using **R Markdown and Shiny**, enabling stakeholders to interactively explore data insights.
* Streamlined data visualization processes by integrating **Power BI** with various data sources, optimizing performance through efficient data loading, and implementing role-based security for secure access to sensitive information.
* Designed, developed, and maintained daily and monthly summary, trending, and benchmark reports using **Tableau** Desktop.
* Employed **Agile methodology and Scrum processes** for project development.
* Implemented comprehensive unit and integration testing using frameworks like **PyTest and Unittest** to ensure code reliability and maintainability.

**Environment:** AWS EC2, S3, Oracle DB, AWS, Linux, Python (Scikit-Learn/NumPy/Pandas/Matplotlib,ggplot), Machine Learning (Logistic Regression/Support Vector Machine/Gradient Boosting/Random Forest), Informatica, Tableau.

**EDUCATION:**

**University:** University of Missouri Kansas City

**Degree:** Master’s in computer science