**Nikhil Sai**

**Sr.** Data Analyst

**EMAIL:** **nikhilsairavuri27@gmail.com PHONE:** **+1 972-534-6597**

**Professional Summary:**

* Senior Data Analyst with a solid background having more than 9 years of experience in Data Analyzing.
* My expertise is deeply rooted in various platforms (e.g. AWS, Google Cloud Platform and Microsoft Azure).
* I am committed to spearheading advanced data engineering projects, with a focus on enhancing security, scalability and efficiency.
* **Programming and Scripting Expertise:** Advanced skills in Python, Scala, Java and Shell scripting within UNIX/Linux environments, facilitating powerful scripting and automation solutions.
* **Extensive Cloud Technology Expertise:** Mastery in utilizing AWS (including EMR, EC2, RDS, S3, Lambda, Glue, Redshift), Azure (spanning Data Lake, Storage, SQL, Databricks) and Google Cloud Platform, facilitating scalable and resilient cloud-based solutions.
* **Data Ingestion and Pipeline Mastery:** Proficient in architecting and executing complex data ingestion strategies, building robust data pipelines, adept in managing Hadoop infrastructures, excelling in data modeling, mining and refining ETL processes for optimal performance.
* In-depth knowledge of the Hadoop ecosystem, proficient in HDFS, MapReduce, Hive, Pig, Oozie, Flume, Cassandra and Spark technologies (including Scala integration, PySpark, RDDs, DataFrames, Spark SQL, Spark MLlib and Spark GraphX), ensuring high efficiency in big data processing and analysis.
* **Advanced ETL Methodology Implementation:** Solid background in applying ETL methodologies with tools such as Microsoft Integration Services, Informatica Power Center, SnowSQL, alongside deep understanding of OLAP and OLTP systems, enhancing data transformation and loading processes.
* **Microservices Development Using Spring Boot:** Expertise in developing lightweight, scalable Microservices with Spring Boot for real-time data processing and seamless integration, leveraging its convention-over-configuration principle for efficient application development.
* **Data Modeling Excellence:** Experienced in formulating both logical and physical data models, employing Star Schema and Snowflake Schema designs to support complex data analysis and reporting needs.
* **Database and SQL Proficiency:** High proficiency in SQL Server and NoSQL databases (such as DynamoDB and MongoDB), executing complex Oracle queries with PL/SQL and leveraging SSIS for effective data extraction, complemented by enhanced reporting through SSRS.
* **Data Visualization and Processing Tools Expertise:** Competent in utilizing data visualization tools like Tableau and Power BI and employing Talend for constructing scalable data processing pipelines, facilitating insightful data presentation and decision-making.
* **Golang-based Pipeline Development:** Skilled in developing and managing Golang-based data processing pipelines, efficiently handling voluminous data through ingestion, transformation and loading, with expertise in Snowflake data warehouse management within Azure and applying Terraform for infrastructure as code across multiple clouds.
* **API Architecture and Security:** Proficient in designing secure API endpoints, incorporating JWT, OAuth2 and API keys for robust authentication and authorization mechanisms, adept in managing various file formats (Text, Sequence, XML, JSON) for versatile data interaction.
* **Agile and Scrum Methodologies:** Strong adoption of Agile and Scrum methodologies, focusing on iterative development, collaboration and efficiency, proficient in Test-Driven Development (TDD) and leveraging CI/CD pipelines (with tools like Jenkins, Docker, Concourse and Bitbucket) for continuous integration and delivery.
* **Testing Tools Proficiency:** Proficient with leading testing tools like Apache JMeter, Query Surge and Talend Data Quality, ensuring rigorous validation of data transformations and ETL processes for accuracy and performance.
* **Network Protocols and Security:** Deep understanding of network protocols (DNS, TCP/IP, VPN), with specialized skills in configuration and troubleshooting to ensure secure and reliable data communication across networks.
* **Real-Time Data Streaming and Analytics Expertise:** Proficient in leveraging cutting-edge technologies like Apache Kafka and Apache Storm for real-time data streaming and analytics. Capable of designing and implementing high-throughput systems that facilitate immediate data processing and insights, enabling dynamic decision-making processes.
* **Data Governance and Compliance:** Expertise in establishing robust data governance frameworks to ensure data integrity, quality and compliance with global data protection regulations (such as GDPR and CCPA).
* Skilled in implementing data lifecycle management practices, metadata management and access controls to safeguard sensitive information and promote ethical data usage.
* **Machine Learning and AI Integration:** proficiency in integrating machine learning models and AI algorithms into data processing pipelines, using platforms like TensorFlow and PyTorch. Adept at developing predictive models and intelligent systems that enhance business operations, customer experiences and decision-making capabilities through actionable insights derived from large datasets.

**Technical Skills:**

| **Programming Languages** | Python, Scala, Java, Golang |
| --- | --- |
| **Scripting** | Shell scripting within UNIX/Linux environments |
| **Cloud Platforms** | AWS (EMR, EC2, RDS, S3, Lambda, Glue, Redshift), Azure (Data Lake, Storage, SQL, Databricks), Google Cloud Platform |
| **Big Data Technologies** | Hadoop ecosystem (HDFS, MapReduce, Hive, Pig, Oozie, Flume, Cassandra), Spark (Scala integration, PySpark, RDDs, DataFrames, Spark SQL, Spark MLlib, Spark GraphX) |
| **sETL Tools** | Microsoft Integration Services, Informatica Power Center, SnowSQL, Talend |
| **Data Modeling** | Star Schema, Snowflake Schema |
| **Databases** | SQL Server, NoSQL (DynamoDB, MongoDB), Oracle (PL/SQL) |
| **Data Visualization Tools** | Tableau, Power BI |
| **Infrastructure as Code** | Terraform |
| **API Development** | JWT, OAuth2, API keys |
| **Microservices Framework** | Spring Boot |
| **CI/CD Tools** | Jenkins, Docker, Concourse, Bitbucket |
| **Version Control Systems** | Git, SVN, Bamboo |
| **Testing Tools** | Apache JMeter, QuerySurge, Talend Data Quality |
| **Network Protocols** | DNS, TCP/IP, VPN |
| **Real-Time Data Streaming** | Apache Kafka, Apache Storm |
| **Machine Learning & AI Platforms** | TensorFlow, PyTorch |
| **Methodologies** | Agile, Scrum, Test-Driven Development (TDD) |

**Certifications:**

* Microsoft Certified Azure Data Engineer Associate.
* AWS Certified Developer Associate.
* Google Certified Data Engineer

**Professional Experience:**

**Client: Humana, Louisville, KY June 2023 to Present**

**Role: Sr. Data Analyst**

**Responsibilities:**

* Participated in the entire Software Development Life Cycle (SDLC), from gathering requirements to deploying solutions, particularly within the healthcare domain, harnessing Azure's PaaS solutions.
* Utilized advanced Python libraries, such as Pandas and NumPy, for complex data manipulation, transformation and analysis projects within the Azure ecosystem.
* Designed and oversaw data pipelines for transitioning patient data to Azure Synapse Analytics, guaranteeing smooth data transfer from diverse sources such as clinical and administrative systems.
* Oversaw and rectified challenges in daily data pipelines and operations, including the validation and examination of incoming patient data sets.
* Proficiency with Snowflake’s cloud data platform and its features, such as snow pipe, time travel and zero-copy cloning.
* Directed the migration from legacy systems to Azure, applying Databricks Spark notebooks and Python for data transformation and loading into Azure SQL through Azure Data Factory pipelines.
* Constructed Azure Data Factory pipelines for extracting, transforming and loading (ETL) data from various sources, including Azure SQL, Azure Data Lake, Hadoop and legacy systems.
* Administered Hive tables and Hadoop processes, migrating them to the Azure cloud for enhanced data analysis and management in healthcare informatics.
* Established CI/CD pipelines within Azure workspaces, focusing on automation and the streamlined deployment process for healthcare applications.
* Implemented version control and configuration management tools, embedding DevOps methodologies into healthcare data management workflows.
* Managed extensive patient and research data volumes, concentrating on ETL processes to maintain data accuracy and efficiency.
* Experience with snowflake cloud data warehouse for integrating data from multiple source systems, including loading nested JSON formatted data into snowflake tables.
* Created snow pipes for continuous data loads.
* Oversaw data ingestion and processing using Azure components, augmenting data visualization and reporting for strategic healthcare decisions.
* Designed and maintained high-performance databases (relational and NoSQL), optimizing database structures for enhanced performance and integrity in healthcare applications.
* Collaborated with healthcare BI teams, formulating complex queries for integration with Power BI, enabling sophisticated data visualizations and reports for patient care and research.
* Sustained large healthcare data warehouses (Amazon Redshift, Google Big Query and Azure SQL Data Warehouse), facilitating advanced analytics on both structured and unstructured medical data.
* Engineered ETL pipelines with Azure Data Factory and implemented star-schema for data warehouse architecture to enhance data storage and accessibility in healthcare settings.
* Developed Spark applications using PySpark for efficient data extraction, transformation and aggregation related to patient and treatment data.
* Employed Apache Airflow and Apache NiFi for the management of healthcare data pipelines and the effective ingestion, transformation and distribution of clinical data.
* Crafted custom functions in Scala and PySpark for bespoke data processing operations in medical research and patient care.
* Managed SSIS packages and integrated automation tools (Jenkins, Artifactory, SonarQube, Chef and Puppet) for comprehensive CI/CD operations in healthcare IT.
* Devised audit, balance and control frameworks to ensure data integrity and security, utilizing SQL DB audit tables to verify accuracy during ETL processes in a healthcare context.
* Created multi-dimensional structures, such as cubes and dimensions, with SQL Server Analysis Services (SSAS) for in-depth analysis of healthcare data.

**Environment:** Azure PaaS, Python, Pandas, NumPy, Azure Synapse Analytics, Stream Sets, Databricks Spark, Azure SQL, Azure Data Factory, Hive, Hadoop, CI/CD pipelines, Version control and configuration management tools, Azure components, Power BI, Snowflake, Amazon Redshift, Google Big Query, Azure SQL Data Warehouse, PySpark, Apache Airflow, Apache NiFi, Scala, SSIS, Jenkins, Artifactory, SonarQube, Chef, Puppet, SQL DB audit tables, SQL Server Analysis Services (SSAS).

**Cigna Insurance services Bloomfield, CT March 2022to May 2023**

**Role: Sr. Data Analyst**

**Responsibilities:**

* Directed and orchestrated complex data engineering initiatives from inception through to maturity in the insurance sector, adeptly navigating through planning, strategy, execution and maintenance phases while seamlessly integrating Agile and Waterfall methodologies to ensure flexibility and rigor in project management.
* Engineered and maintained robust multi-node clusters on AWS, leveraging EC2 instances for scalable computing. Implemented comprehensive monitoring and alerting systems utilizing CloudWatch and CloudTrail, ensuring operational excellence and security across EBS, EC2, ELB, RDS, S3 and SNS services, while enforcing best practices in data security within S3 buckets.
* Spearheaded the modernization of legacy insurance claims database systems and Informatica ETL processes to AWS Cloud, Redshift and Snowflake platforms, innovating with asynchronous task management tools like Celery, RabbitMQ and Redis to enhance performance and scalability.
* Advanced AWS DynamoDB functionalities by integrating with Lambda for serverless operations and developing Spark scripts for efficient AWS Glue jobs and EMR processing, automating workflows with Python to achieve operational efficiency and reliability.
* Mastered data transfer operations with Sqoop, facilitating seamless data exchange between Snowflake, Oracle and DB2 systems in the insurance sector. Refined database management capabilities with advanced SQL and PL/SQL scripting, deepening proficiency in Snowflake's database architecture.
* Experience building snow pipe and developing transformation logic using snow pipe.
* Employed Big Query for efficient data warehousing solutions, optimizing data integration, loading and transformation processes from Google Cloud Storage, Cloud Pub/Sub and heterogeneous external databases, ensuring data fluidity and accessibility.
* Leveraged a comprehensive suite of technologies including Spark, PySpark, Hive, Hadoop and Scala for a broad spectrum of data-related tasks from analytics, ingestion and integrity verification to the management of diverse data formats such as JSON, CSV, Parquet and Avro in the insurance analytics domain.
* Architected and administered real-time data streaming infrastructures using Apache Kafka, facilitating immediate data processing and insights, essential for real-time decision-making and analytics in insurance.
* Pioneered automation in data collection from a variety of sources such as APIs, AWS S3, Teradata and Redshift, utilizing PySpark and Scala. Implemented Oozie workflows for strategic job orchestration, enhancing the software development lifecycle's efficiency.
* Created compelling and interactive dashboards and reports with Power BI, transforming raw insurance data into actionable insights and decision-support tools.
* Good understanding of snowflake cloud technology and experience working with snowflake clone and time travel.
* Conceptualized and evaluated advanced dimensional data models, employing Star and Snowflake schemas and integrating industry-standard methodologies advocated by Ralph Kimball and Bill Inmon to optimize data warehouse design and functionality in insurance.
* Developed and instituted sophisticated logging, monitoring and error management frameworks within REST APIs, enhancing system reliability and operational transparency.
* Executed the deployment of microservices architecture on Kubernetes clusters, leveraging Jenkins for robust CI/CD pipelines and utilized Jira for effective project management and issue tracking, facilitating agile and efficient development cycles in the insurance domain.
* Demonstrated version control using Git, ensuring meticulous code management practices and fostering a culture of transparency and collaboration in the development workflow.
* Applied advanced testing strategies and tools, including Apache JMeter, to rigorously validate the accuracy, performance and integrity of ETL processes and data migrations, ensuring the highest data quality standards in insurance data handling.

**Environment:** Agile, Waterfall, AWS, EC2, CloudWatch, CloudTrail, EBS, ELB, RDS, S3, SNS, Informatica ETL, Redshift, Snowflake, Celery, RabbitMQ, Redis, AWS DynamoDB, Lambda, Spark, AWS Glue, EMR, Python, Sqoop, Oracle, DB2, SQL, PL/SQL, Big Query, Google Cloud Storage, Cloud Pub/Sub, Spark, PySpark, Hive, Hadoop, Scala, JSON, CSV, Parquet, Avro, Apache Kafka, APIs, Teradata, Oozie, Power BI, Star Schema, Snowflake Schema, REST APIs, Kubernetes, Jenkins, Jira, Git, Apache JMeter.

**Dish Network, Englewood, CO June 2021 to Februarys 2022**

**Role: Sr. Data Analyst**

**Responsibilities:**

* Performing ETL on large amounts of data to make it available for observability, analysis and create data pipelines and Datawarehouse
* Understanding features and business requirements by involving in meetings with Business team and Data Analyst Team.
* Fetching data from different source systems using SQL scripts, Pyspark and Python scripts
* Worked on writing and reading data from csv and excel file formats in python.
* Manage large datasets using Python panda’s library and MYSQL.
* Implementation and optimization of Snowflake SQL queries, ensuring efficient data processing and retrieval
* Use of SQL window functions for advanced data analysis and partitioning within result sets.
* Utilizing SQL aggregate functions to perform calculations on groups of rows, enabling comprehensive data summarization and reporting
* Configured and optimized AWS SQS queues to efficiently manage message throughput, ensuring timely processing and delivery of critical data across distributed systems.
* Designed and implemented EKS clusters tailored to specific workload requirements
* Use of AWS Code Build for continuous integration and deployment pipelines.
* Implemented and maintained Pyspark clusters on AWS EMR, optimizing performance and scalability for large-scale data processing.
* Developed and deployed MapReduce programs to process and analyze massive datasets, enhancing data insights and decision-making capabilities.
* Demonstrated proficiency in working with various relational database management systems (RDBMS) like AWS Redshift, Athena, RDS, DynamoDB, Postgres, Oracle, and SQL Server
* Automation of data update to various google sheets, tableau dashboards and Power BI dashboards as required.
* Use of different python libraries NumPy, SciPy, PyTables, matplotlib, Panda’s data frame and urllib2 in various tasks.
* Leveraged AWS Athena to perform querying of data stored in Amazon S3, developing, and executing complex SQL queries in Python to extract data.
* Developed Python scripts for PDF extraction, utilizing libraries to parse and retrieve text content from PDF documents.
* Developed and maintained SQL scripts for data manipulation, validation, and reporting, ensuring data integrity and consistency.
* Automation of manual data related tasks including data transformations using Python.
* Automated web interactions using Python Selenium, improving efficiency by executing repetitive tasks such as data downloads, data scraping.
* Implemented and optimized data pipelines using Python and Snowflake to enable efficient data processing and analytics.
* Implemented Event Bridge for building event-driven architectures, enabling seamless integration and real-time event processing.
* Utilized Glue Data Catalog to create and manage metadata for various data sources, facilitating efficient data discovery and ETL processes.
* Managed sensitive information and credentials securely using Secrets Manager, ensuring robust security practices within AWS infrastructure.
* Demonstrated proficiency in Identity and Access Management (IAM), including the creation and management of IAM users, roles, and policies to enforce least privilege access and ensure compliance.
* Leveraged AWS Cloud Development Kit (CDK) to provision and manage cloud resources using familiar programming languages, enhancing automation and infrastructure-as-code practices.
* Implemented ETL processes using PySpark to process large datasets, enhancing data quality and reducing processing time
* Using packages such as beautiful soup to parse XML and HTML data
* Implemented batch job automation for scheduled execution of Python scripts ensuring timely and accurate processing of critical tasks.
* Performing CRUD operation using REST APIs to get data and update data using secured method such as OAuth2, tokenization.
* Implemented ETL processes using AWS such as Glue, Lambda, EC2, CloudWatch to extract, transform, and load data from various sources into the data lake.
* Implemented automated PDF text extraction utilizing AWS Tex tract and Boto3 in Python and converting tabular content into Pandas Data frames, facilitating seamless analysis and reporting processes.
* Managed S3 storage infrastructure by developing and implementing automated solutions in Python using Boto3, ensuring efficient data upload, download.
* Test and troubleshooting REST APIs using Postman.
* Connect to file transfer systems such WinSCP, SSH to import/export data using python.
* Utilizing GitLab for version control, collaborating with cross-functional teams to manage source code repositories, track changes.

**Environment:** Python, Pandas, NumPy, selenium, Beautiful Soup, SQL, MYSQL, Snowflake, Redshift, Athena, RDS, Glue, Lambda, EC2, S3, XML, ETL, Git, AWS, Tex tract, S3, Lambda, Glue Job, REST API, Postman, WinSCP, SSH and Windows.

**Client: Staples, Framingham, MA November 2020 to May 2021**

**Role: Data Analyst**

**Responsibilities:**

* Full Data Pipeline (FDPL) -Building frameworks which are configurable, metadata driven & customized.
* Developed ETL pipeline using Databricks to build SILVER/GOLD layer tables using PySpark.
* Experience sizing clusters for development and GIT integration with Azure DevOps.
* Data Ingestion to ingest data from various data sources using Data factory and Azure webapps.
* Migrated the Databricks ETL jobs to Azure Synapse Spark pools.
* The ETL deals with structured data coming from Oracle source system will be ingested through Synapse pipelines using the Oracle SQL connector.
* Designed a star schema for customer analytics, resulting in a 20% reduction in query response time.
* Implemented snowflake schema for a financial reporting system enhancing data integrity and ease of maintenance.
* Collaborated with marketing teams to align dimensional models with campaign tracking requirements, enabling more accurate performance analysis.
* Built Synapse spark ETL notebooks to standardize the data coming from vendor specific SFTP accounts and write the resultant data into Azure blob storage.
* Used Azure Synapse for dedicated SQL database to build data model using Fact and dimensional tables for KPI using 24-hour delta records.
* Built Synapse Dataflows for standard KPI model building by reading the data from shared data sources (internal data teams/pods) and load the data into SQL Pools (MPP). The data will then be pulled into Power BI for dashboarding.
* Migrated the legacy workflows from batch jobs on-premises systems (xml, CSV, XLS files) to Data Lake using Synapse pipelines.
* Data Integration to apply business rules & make data available to different consumers using Databricks spark. Data delivery FW for data-driven caching, Ad hoc data access, Vendor & API integration
* Building a data-driven Caching Layer for data delivery of Sales Dashboard using Power BI and Snowflake. Building a standardized automated Vendor Integration Model (recognized as a standard template by Eng. org).
* Migrated two Data marts from Teradata into Snowflake using Snow SQL and External storage integration. Used COPY INTO command for ingesting large files from Azure Blob to Snowflake stage and automated the ingestion process using Snow SQL commands with shell scripts.
* Conduct code reviews daily. Provide low level architecture design for the Azure pipelines. Interact with SLT to gather additional requirements (stretch goals) and provide demos to cross-functional teams.
* Developed workflows suing Databricks Delta live tables and used MERGE SQL to perform Change data capture for implementing the SCD type-2 tables. Used Z-optimize for data compaction and VACCUM commands for maintaining the lifecycle for datasets.
* Created Mount point on Databricks to connect with blob storage to retrieve data and perform data analysis using Pyspark on Databricks clusters.
* Enabled GIT on Databricks for versioning and used widgets for setting the parameters in the script. Implemented sub-routines using remote data bricks job execution in case of master job failure.
* Used Python SDK to remotely connect to Databricks and Azure Data factory to run jobs/pipelines. Optimized the pyspark jobs on Databricks using memory fraction/ storage fraction limit changes and other custom Spark configurations.
* Implemented streaming pipeline on Clickstream data by connecting Databricks with Azure Event-hubs and with Azure Stream Analytics.
* Migrated Apache Hive tables/models from Hadoop to Databricks on Azure, implemented access policies to restrict user access on Table and Schema level for all Databricks tables. Currently working on Snowpark to convert the pyspark jobs into Snowflake equivalent Snowpark jobs.
* Built Shell scripts which load the data from SFTP and land in HDFS as raw data. Build python jobs to run data quality checks using DBT and using great Expectations (this is a python package name).

**Environment:** Azure Data factory, Azure Databricks, spark, Kafka, Log Analytics, Azure DevOps (git &CICD), HDP, hive, Sqoop, Oracle Goldengate, Teradata, Google Campaign Manager, python, shell scripting, snowflake, Azure Webapps, Azure App services, Azure Datahub.

**UBISOFT, Hyderabad, Telangana, India April 2017 to October 2019**

**Role: Data Analyst**

**Responsibilities:**

* Designed and managed advanced ETL pipelines utilizing Python, Spark and PySpark, incorporating Airflow within the Google Cloud Platform (GCP) to optimize data collection and ensure seamless synchronization between online sales databases and inventory systems.
* Led the modernization of existing Hadoop infrastructures to GCP, leveraging the capabilities of Cloud Storage, Dataproc, Dataflow and Big Query.
* Conducted thorough comparisons of on-premises Hadoop versus GCP's Dataproc and assessed Bigtable’s performance improvements within GCP for handling high-volume retail transactions.
* Orchestrated comprehensive data warehousing and integration endeavors on platforms like Snowflake Cloud and Informatica Intelligent Cloud Services (IICS).
* Crafted complex Informatica ETL workflows and mappings to consolidate data from various relational databases, transitioning reporting and analytical processes from Oracle to Big Query using Power BI for retail sales analysis.
* Innovated data model structures in Neptune to enhance data insertion efficiency and utilized the Gremlin query language for sophisticated data queries.
* Implemented Grafana for developing dashboards that visualize real-time metrics extracted from the Cassandra database for inventory tracking.
* Created custom User-Defined Functions (UDFs) in Pig and Hive, integrating Pig Latin and HiveQL with Python for enriched data processing capabilities and deployed solutions on application servers such as Glassfish and CGI to facilitate system compatibility and integration across retail chains.
* Engineered and maintained high-efficiency database systems on SQL Server and Cassandra, establishing database architectures aimed at reducing data redundancy and optimizing query performance for real-time retail data access.
* Utilized Terraform Cloud and Terraform Enterprise for enhanced collaboration, state management and secure deployment of Terraform configurations, streamlining infrastructure as code practices in retail environments.
* Developed RESTful APIs with Golang, extending data processing services to a wider range of applications and enhancing interoperability across software ecosystems in the retail sector.
* Deployed and maintained scalable Kafka clusters, designed to facilitate real-time data streaming with high availability, fault tolerance and optimal performance, integrating seamlessly with a variety of data sources and applications within the retail industry.
* Directed integration testing efforts across multiple data platforms and applications, ensuring seamless data flow and interoperability among disparate systems, including databases, APIs and external tools in the retail domain.
* Implemented machine learning models within data pipelines using TensorFlow and Scikit-Learn, integrating predictive analytics and AI-driven insights into decision-making processes.
* Adopted streaming analytics with Apache Flink for real-time data processing, complementing existing Kafka implementations.

Environment: Python, Spark, PySpark, Airflow, Google Cloud Platform (GCP), Cloud Storage, Dataproc, Dataflow, Big Query, Snowflake Cloud, Informatica Intelligent Cloud Services (IICS), Oracle, Power BI, Neptune, Gremlin, Grafana, Cassandra, Pig, Hive, Glassfish, CGI, SQL Server, Terraform Cloud, Terraform Enterprise, Golang, Kafka, TensorFlow, Scikit-Learn, Apache Flink.

**HCL Technologies, Hyderabad, Telangana, India June 2015 to May 2017**

**Role: Data Analyst**

* Implemented data pipelines on AWS Glue to effectively extract, transform, and load various datasets for Chevron's analytics, improving operational and decision-making insights.
* Responsible for the execution of big data analytics, predictive analytics, and machine learning initiatives.
* Built real-time data pipelines by developing Kafka producers and Spark Streaming applications for processing large-scale data from oil and gas operations.
* Monitored Spark jobs using the UI interface (Name Node Manager, Resource Manager ETS) in AWS.
* Utilized AWS services with a focus on big data architecture, analytics, enterprise data warehouses, and business intelligence solutions.
* Experience in AWS services like EC2, EMR, DynamoDB, Athena, and Redshift
* Automated data workflows using Python and Apache Airflow, resulting in increased efficiency and reduced manual errors.
* Developed Spark SQL scripts using PySpark to perform transformations and actions on Data Frames and Data Sets in Spark for faster data processing.
* Created data pipelines for extracting, transforming, and loading data from various sources, including internal and external APIs.
* Conducted performance tuning and optimization of SQL queries on AWS Redshift to enhance data processing efficiency.
* Developed Spark scripts using Python on AWS EMR for data aggregation, cleansing, and mining.
* Developed and maintained data orchestration workflows using AWS Step Functions to manage complex ETL tasks and dependencies.
* Worked together with data scientists to enable real-time model inference through SQS-triggered Lambda functions and to run scripts in response to events in DynamoDB and S3.
* Collaborated with cross-functional teams to understand business requirements and translate them into actionable Tableau visualizations.
* Proficient in using Python for DynamoDB interactions, including Boto3 library for seamless integration
* Implemented CRUD operations on DynamoDB tables using Python scripts, ensuring data consistency.
* Generating reports using Python as per the business requirement and create visualization.
* Participate in the design, build and deployment of NoSQL implementations like MongoDB.
* Added support for Amazon AWS S3 and RDS to host static/media files and the database into Amazon Cloud.
* Extensive code reviewing using GitHub pull requests, improved code quality, also conducted meetings among Team.
* Managed and processed large datasets using Hadoop MapReduce, improving data processing efficiency
* Developed scripts to migrate data from proprietary database to Postgres SQL.
* Followed Agile Methodologies and SCRUM Process.

**Environment:** Python, Django, HTML5, CSS, Bootstrap, jQuery, JSON, JavaScript, PostgreSQL, MongoDB, Ansible, MySQL, Google Cloud, Amazon AWS S3, Bugzilla, JIRA, Hadoop, Hive, Apache Airflow