**VENNELA**

**Sr. DATA ENGINEER**

**Email:** vennelamde@gmail.com **Phone:** (404)857-4064

**Professional Summary:**

* Senior Data Engineer with a solid background spanning 10 years in Data Engineering. My expertise is deeply rooted in the Hadoop ecosystem, encompassing various cloud platforms (e.g., AWS, Google Cloud Platform and Microsoft Azure) and specializing in ETL processes. 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.
* **Comprehensive Ecosystem Acumen:** In-depth knowledge of the 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, QuerySurge 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.
* **Big 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.

**Databricks:**

* Hands-on with implementing transformations to scrub the incoming data and de-duplication logic.
* Created Data Frames and performed analysis using Spark SQL. Interfacing with external tables and performing ad-hoc joins and broadcast variable joins.
* Proficient in building Delta live table frameworks (DLT), implemented 4 dashboards so far using the DLT framework.
* Built real-time data processing feeds using Spark structured streaming and Kafka.

**Steaming:**

* Excellent knowledge of Kafka Architecture.
* Used Kafka for activity tracking and Log aggregation.
* Designed and developed real time streaming reads of system information and processed logs in real time using NiFi, Kafka, Spark, and HBase.
* Experience in performance optimization of streaming jobs and config parameters (manual/automated).

**SQL and NoSQL:**

* Good understanding of HBase architecture and designing tables.
* Ability to write complex SQL queries to analyze structured data.
* Integrated HBase for offset management and de-duplication.
* Extensive experience in working with NoSQL databases and their integration Dynamo DB, Cosmo DB, Mongo DB, Cassandra and HBase.

**Version Control and Build Tools:**

* Experienced in using GIT, SVN.
* Ability to deal with build tools like Apache Maven, SBT.

**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, Data Frames, Spark SQL, Spark MLlib, Spark GraphX)
**ETL Tools:** Microsoft Integration Services, Informatica Power Center, Snow SQL, 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, Query Surge, 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)

**Professional Experience:**

**Client:** Cardinal Health, REMOTE

**Location:** Dublin - OH

**Role:** Sr. Data Engineer

**Duration:** May 2023 to Present

**Responsibilities:**

* Responsible for delivering the US state reports that are regulated by FDA through monitoring the inbound data, participated in designing, implementing, and manage data solutions on Azure Cloud platform and involved in Teradata Retirement activity to replicate the data to Google Cloud.
* Build Synapse pipelines to ingest data from inbound state report data in csv format through APIs. The data is further processed through SQL pools.
* Used Synapse Spark pools for Data cleansing and data transformation for inbound feeds. The ingestion is made real-time by connecting Kafka to Synapse Analytics and processing the data through Spark streaming.
* Built Data factory ingestion pipeline to retrieve data feeds from Google AdWords API to capture the Ads response and related feeds. The pipeline runs every 1-hour and data transformation is done on Databricks.
* Experience in migrating on premise database to Azure Data Lake store using Azure data factory. Built metadata driven ingestion framework for data bricks that works in an automated fashion and runs every 15mins to capture the database events.
* Implemented Delta-live-table (DLT) framework on data bricks to build interconnected jobs for live data analytics.
* Used cloud files pattern to invoke Databricks auto-loader which takes care of auto processing of data.
* Build Streaming live tables on Databricks using DLT framework to keep the latest files readily available during the analytics script build process.
* Implemented data quality checks using data constraints on DLT scripts.
* Implemented the TBL properties (inbuilt function) to document the changes and define the layers as Databricks MEDALLION architecture.
* Implemented auto infer method for JSON data feeds landing into Databricks, this will automatically check for schema changes and add the new columns to a custom rescue column (inbuilt feature) and then built a Streaming LIVE DLT table on top of it so that the data is automatically available for downstream processing.
* Experience in implementing version control for data models, ensuring traceability and facilitating collaborative development in a team environment.
* Proven track record of optimizing data models for performance, including indexing strategies and query optimization, resulting in improved system responsiveness.
* Involved in developing Spark applications using Spark – SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Used Spark Structured streaming on Databricks for processing the claim events stored on Azure Stream Analytics. Watermarking mechanism is implemented to handle late-arriving data.
* Perform ETL operations on source systems to Azure Data Lake analytics using a combination of Azure Data factory, T-SQL, Spark SQL.
* Data ingestion to Azure services like Azure Data Lake, Azure Storage, Azure SQL, DW and processing the data in Azure Databricks.
* Created pipelines in ADF using Linked services/pipelines to extract, transform and load data from different sources like Azure SQL, Blob Storage.
* Responsible for estimating the cluster size, monitoring, and troubleshooting of Spark data bricks cluster. Used job cluster to schedule jobs on Databricks, optimized slow running spark jobs using repartition, coalesce, broadcast join and other performance tuning techniques.
* Migrated Hive and HDFS data elements using ADFs copy activity. To connect to HDFS, installed the Self-hosted integration runtime on datacenter and used WebHDFS connector to link ADF and Hadoop.
* Create and manage data storage solutions using GCP services such as Big Query, Cloud Storage and Could SQL. Design and implement data pipelines using GCP services such as Dataflow, Dataproc and Pub/Sub. Monitor and troubleshoot data pipelines and storage solutions using GCP’s Stack driver and Cloud Monitoring.
* Partner with Digital Solutions and other shared service platform teams to ensure all appropriate data sources / data flow is identified.
* Preparing and executing a strategy to migrate legacy applications to cloud-native and DevSecOps toolsets.
* Implement new integrations for solutions within the CSMP technology evolution project.
* Build detailed technical design artifacts and oversee implementation.
* Verify data from new and migrated interfaces and reports.
* Experience in optimizing MongoDB tables by creating indexes, using projections to return only necessary data, and using bulk write operations for inserting and updating multiple documents to significantly improve the performance.
* Implemented mechanisms to ingest data into MongoDB from various sources such as logs, external databases, and streaming data.
* Experience in planning and executing data migration strategies when transitioning from another database system to MongoDB.

**Environment:** Azure Data factory, Azure Databricks, Azure Synapse, Azure DW, Spark, pgAdmin 4(PostgreSQL), DBeaver, Big Query, Cloud Storage, Cloud SQL, Dataflow, Dataproc, Pub/Sub, IAM, Python, Looker, MongoDB, SharePoint.

**Client:** Cigna Insurance services

**Location:** Bloomfield, CT

**Role:** Sr. Data Engineer

**Duration:** October 2021 to April 2023

**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.

**Client:** Staples, REMOTE

**Location:** Framingham-MA

**Role:** Data Engineer

**Duration:** October 2020 to September 2021

**Project Description:** Responsible for the development efforts and performance of Full Data Pipeline (FDPL) project, this is a migration project of the Staples Retail Business. Full Data Pipeline (FDPL) is a configurable & metadata driven data ingestion, integration & delivery Framework (FW) including data driven caching strategy, Adhoc data access, metadata driven extract.

**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-premise systems (xml,CSV,XLS files) to DataLake 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, Adhoc 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 SnowSQL and External storage integration. Used COPY INTO command for ingesting large files from Azure Blob to Snowflake stage and automated the ingestion process using SnowSQL 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 databricks job execution in case of master job failue.
* Used Python SDK to remotely connect to Databricks and Azure Datafactory 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 Snow Spark to convert the pyspark jobs into Snowflake equivalent Snow Spark 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 Appservices, Azure Datahub.

**Client:** PepsiCo

**Location:** Hyderabad, Telangana, India

**Role:** Data Engineer

**Duration:** November 2018 to September 2020

**Responsibilities:**

* Build Databricks job to implement data quality checks and transform the sales data that is received from front end applications.
* Build data bricks jobs to read from real-time feeds coming from online sales and process them through Spark structured streaming.
* Built ETL jobs to read from SFTP accounts and extract files to Databricks dbfs storage and process them through pyspark.
* Built data factory jobs using the auto-loader interface, built various ingestion pipelines for oracle, salesforce and land the data into Azure Data Lake storage Gen1(ADLS Gen1).
* Built interdependent jobs for capturing the KPI and building metrics for sales funnels and then redirect the output to power BI dashboards.
* Built streaming jobs using Spark Streaming which transfer data from transaction severs had been loaded into Cassandra tables in real-time workloads. The tables are further used by machine learning models to verify transaction authenticity.
* Written Spark jobs to convert the Oracle materialized views to generate the datasets faster for the tableau Dashboards. The ETL for these jobs run on EMR which trigger downstream jobs on successful execution using Step-functions. Developed the end-to-end automation framework to implement this functionality.
* Build NiFi pipelines to transform the raw data from JSON format into separate datasets as per their category. Installed NiFi platform was installed on EC2 and redeployed the workflows from On-prem NiFi to EC2 NiFi instance.
* Migrated historical data to S3 and developed a reliable mechanism for processing the incremental updates. The data has been migrated from Hadoop cluster and used DIST CP command to migrate large datasets to AWS S3.
* The raw datasets of transactional vertical are merged based on business criteria using Spark to create a master dataset for building Dashboard. Used Python UDFs to implement business logic and the UDF are embedded into MAP transformation using pySpark.
* Developed Spark jobs using Scala on top of Yarn for interactive and Batch Analysis. Developed UNIX shell scripts to load many files into HDFS from Linux File System.
* Experience in querying data using Spark SQL for faster processing of the data sets. Offloaded data from EDW into Hadoop Cluster using Sqoop. Developed Sqoop scripts for importing and exporting data into HDFS and Hive
* Configured CRON jobs to run RAW layer data extraction jobs on a fixed schedule. Also used Oozie to schedule Sqoop and Hive jobs to build the data model on Hive.
* Built Sqoop jobs to load the data from Teradata and Oracle into HDFS. The jobs load the data incrementally using last updated timestamp.
* Built python jobs to read data from mainframe systems in fixed-width format and land on HDFS. Built PIG jobs which will read the fixed-width files and re-format data as per the defined schema and store in HDFS.
* Worked on HCatalog to manage the schema between PIG and HIVE frameworks while working with mainframe fixed-width data.
* Designed appropriate Partitioning/Bucketing schema in HIVE for efficient data access during analysis and designed a data warehouse using Hive external tables and created Hive queries for analysis.
* Developed multiple spark batch jobs using Spark SQL and performed transformations using many APIs and updated master data in Cassandra database as per the business requirement.
* Migrated the Oracle database (2 datamarts) using AWS Database migration service. And loaded it into Redshift. Optimized the EMR workloads for different type of data loads by choosing right compression, cluster type, instance type storage type and EMRFS to analyze data with low cost and high scalability.

**Environment:** AWS, S3, Sqoop, Kafka, Spark, Spark SQL, Hive, LINUX, Oozie, Java, Scala, Eclipse, Tableau, UNIX Shell Scripting, Putty.

**Client:** HSBC

**Location:** Hyderabad, Telangana, India

**Role:**  Data Engineer

**Duration:** July 2016 to October 2018

**Responsibilities:**

* Developed and maintained data orchestration workflows with AWS Step Functions for ETL tasks.
* Wrote SQL scripts for data migration and handled data discrepancies, including data migration from Teradata SQL to Snowflake.
* Performed structural modifications using MapReduce and Hive, and analyzed data with Tableau.
* Designed and executed data solutions on AWS, utilizing services like AWS Lambda and Amazon Redshift.
* Migrated an on-premises application to AWS, using services like EC2 and S3 for data processing and storage, and maintained a Hadoop cluster on AWS EMR.
* Building data pipeline ETLs for data movement to S3, then to Redshift.
* Developed MapReduce jobs to enhance data quality and accuracy.
* Used Spark SQL for loading JSON data, creating Schema RDD, and handling structured data.
* worked with the Hadoop ecosystem, implemented Spark with Scala, and utilized Data Frames and Spark SQL for faster data processing.
* Applied Amazon Kinesis for real-time data streaming and processing for timely analysis.
* Converted Hive/SQL queries into Spark transformations using Spark RDDs and PySpark.
* Proficient in managing and configuring YARN clusters for efficient resource allocation and job scheduling.
* Expertise in deploying and administering HDFS clusters for distributed storage and data processing.
* Created Hive tables, managed HDFS clusters, and improved storage utilization.
* Designed and optimized data models in Amazon Redshift for efficient querying and high performance.
* AWS CloudWatch for monitoring data pipelines' performance.
* Applied AWS Lambda functions for automating data movement and ETL processes.
* Imported required tables from RDBMS to HDFS using Sqoop and used PySpark RDDs to get real time streaming of data into HBase.
* Created Business Logic using Python to create Planning and Tracking functions.
* Used Amazon Web Services (AWS) such EC2, S3, Lambda for enhanced productivity of capacity and quick access.
* Implemented Spark using PySpark libraries for faster testing and processing of data.
* Extensively utilized Python frameworks like Django, Flask, PyUnit and libraries like matplotlib.
* Developed tools using Python, Shell scripting, XML to automate some of the menial tasks.
* Added support for Amazon AWS S3 and RDS to host static/media files and the database into Amazon Cloud.
* Orchestrated real-time data streaming pipelines with Amazon Kinesis and FireHose, enabling timely insights and actions on streaming data.
* Implemented multi-factor authentication (MFA) and IAM role session policies to enhance authentication security and prevent unauthorized access to AWS resources.
* Used PySpark-SQL to load JSON data and create schema RDD, Data Frames and loaded it into Hive Tables and handled structured data using Spark-SQL.
* Created PyUnit test scripts and used it for unit testing.
* Stored the data in the form of JSON structure-based documents, stored in a collection using MongoDB. Worked on creating the Docker containers and Docker consoles for managing the application life cycle.
* Development of Python APIs to dump the array structures in the Processor at the failure point for debugging.
* Collaborated within a team using an agile development workflow and widely accepted collaboration practices using Git.
* Used JIRA for Bug tracking and issue tracking.

**Environment:** Python, Pandas, Microsoft, SQL server, Snowflake, MongoDB, GitHub, Jenkins, Django, HTML5, CSS, Bootstrap, JSON, JavaScript, AJAX, RESTful, MongoDB, MySQL, SQLite, Docker, AWS (EC2, S3), PyUnit, Jenkins,

**Client:** HP

**Location:** Hyderabad, Telangana, India

**Role:** Data Engineer

**Duration:** June 2015 to June 2016

**Responsibilities:**

* 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