**Sowmya**

**Sr. Data Engineer**

**Email Id:** [**sowmyadataengineer5657@gmail.com**](mailto:sowmyadataengineer5657@gmail.com) **Contact Details: 928-260-7569**

**Professional Summary:**

* Accumulated over 10 years of professional expertise specializing in the development of Data Systems and Business Systems, with a primary focus on Data Engineering and Data Analysis.
* Proficient in the complete Software Development Life Cycle (SDLC), actively participating in daily scrum meetings with cross-functional teams.
* Demonstrated excellence and hands-on experience in Amazon Web Services (AWS), particularly with AWS S3 and EC2.
* Expertise in building Enterprise Data Warehouses and Data Warehouse appliances from the ground up, employing both Kimball and Inman approaches.
* Skilled in data import/export using stream processing platforms like Flume and Kafka.
* Utilized SSIS and SSRS for report creation and management within organizational settings.
* Extensive experience in crafting Storm topologies for processing events from Kafka producers and emitting them into Cassandra DB.
* Hosted applications on Google Cloud Platform (GCP) using Compute Engine, App Engine, Cloud SQL, Kubernetes Engine, and Cloud Storage.
* Developed and deployed various Lambda functions in AWS, incorporating both AWS Lambda Libraries and custom Scala Libraries.
* Implemented CI/CD pipelines with Jenkins for deploying Microservices in AWS ECS, Python jobs in AWS Lambda, and containerized deployments of Java & Python.
* Proficient with Terraform for Infrastructure as Code, execution plans, change automation, and extensively used auto-scaling. Hands-on experience in Kubernetes for multi-cloud cluster management in AWS.
* Experienced with Big Data on GCP, covering services like Big Query, Pub/Sub, Datapost, and Dataflow.]
* Skilled in data modeling tools such as Erwin, Power Designer, and ER Studio.
* Strong background in big data tools, including Hadoop, HDFS, and Hive.
* Expertise in Data Migration, Data Cleansing, Transformation, Integration, Data Import, and Data Export.
* Technical and analytical skills for OLTP design and OLAP dimension modeling.
* In-depth knowledge of Snowflake Database, encompassing Schema and Table structures.
* Hands-on experience with Snowflake utilities, Snow SQL, Snow Pipe, and Big Data techniques using Python/Java.
* Migration of data warehouses and databases into Hadoop/NoSQL platforms.
* Design and development of Oracle PL/SQL and Shell Scripts, focusing on Data Conversions and Data Cleansing.
* Experience with query optimization and ETL loading on Teradata.
* In-depth understanding of Apache Spark job execution components and hands-on work with NoSQL databases, including HBase and MongoDB.
* Created dashboards using Tableau and Power BI for valuable insights based on business needs.
* Utilized Spark Data Frames API over the Cloudera platform for analytics on Hive data.
* Proficient in Python Scripting, encompassing stats functions with NumPy and visualization using Matplotlib and Pandas.
* Experience with SQOOP for importing/exporting data from RDBMS to HDFS and Hive.
* Knowledgeable about job workflow scheduling and locking tools/services like Oozie, Zookeeper, Airflow, and Apache NiFi.

**Technical Skills:**

|  |  |
| --- | --- |
| **Hadoop/Big Data Technologies** | Hadoop, Map Reduce, Sqoop, Hive, Oozie, Spark, Zookeeper and Cloudera Manager, Kafka, Flume |
| **ETL Tools** | Informatica |
| **NO SQL Database** | HBase, Cassandra, Dynamo DB, Mongo DB. |
| **Hadoop Distribution** | Horton Works, Cloudera |
| **Build Tools** | Maven |
| **Programming & Scripting** | Python, Scala, JAVA, SQL, Shell Scripting, C, C++ |
| **Databases** | Oracle, MY SQL, Teradata |
| **Machine Learning Analytics Tools** | Supervised Learning (Linear Regression, Logistic Regression, Decision Tree, Random Forest, SVM, Classification), Unsupervised Learning (Clustering, KNN, Factor Analysis, PCA), Natural Language Processing, Google Analytics Fiddler, Tableau |
| **Version Control** | Git, GitHub, SVN, CVS |
| **Operating Systems** | Linux, Unix, Mac OS-X, CentOS, Windows 10, Windows 8, Windows 7 |
| **Cloud Platform** | Google Cloud Platform (GCP), Amazon Web Services (AWS), Microsoft Azure. |
| **Cloud Technologies** | **AWS (S3, DynamoDB, Redshift, Lambda, Step Functions, RDS, Aurora, Shield, CloudFormation, CloudWatch, SNS, SQS, EC2, EMR, Kinesis Firehose, Route 53); Azure (Synapse Analytics, SQL Database, Cosmos DB, Blob Storage, Azure SQL, HDInsight, Stream Analytics, Databricks, Machine Learning Service, Data Factory, Event Hub, Azure Kubernetes Service (AKS), DevOps, Azure Resource Manager (ARM), Azure Monitor, Log Analytics)** |

**Professional Experience:  
CNP Energy, Houston, TX          April 2023 to Present  
Role: Sr Data Engineer**

**Responsibilities:**

* Spearheaded the development of Python-based Spark applications, leading the execution of an Apache Spark data processing project to efficiently manage data from diverse relational databases (RDBMS) and streaming sources.
* Played a pivotal role in constructing scalable distributed data solutions using Hadoop, optimizing existing algorithms, and enhancing performance in collaboration with Spark.
* Responsible for sending quality data thru secure channel to downstream systems using role base access control and Stream sets
* Design, develop, configure, and troubleshoot APIs and policies using APIGEE
* Knowledge and strong deployment experience in Hadoop and Big Data ecosystems- HDFS, Map Reduce, Spark, Pig, Sqoop, Hive, Oozie, and Kafka.
* Create Pyspark frame to bring data from DB2 to Amazon S3.
* Engineered data pipelines on Google Cloud Platform (GCP) using Airflow for ETL tasks, employing various Airflow operators to facilitate seamless workflow management.
* Sound knowledge and Hands-on-experience with - MapR, Ansible, Presto, Amazon Kinesis, Storm, Flink, Stream Sets, Star Schema, Snowflake Schema, ER Modeling and Talend.  Used Bulk Collections for better performance and easy retrieval of data, by reducing context switching between SQL and PL/SQL engines.
* Worked with data visualization tools including Tableau, Adobe Analytics Omniture, and matplotlib
* Good understanding on Informatica tools like Snaplogic, Informatica, Ab initio and helped ETL Teams with mapping documents and transformations.
* Engineered NiFi dataflows for ingesting data from Kafka, executing transformations, storing in HDFS, and triggering Spark streaming jobs.
* Implement One time Data Migration of Multistate level data from SQL server to Snowflake by using Python and Snow SQL
* Utilized workday REST and SOAP API's, CSV files extracted to on - premise FTP server, and Snaplogic workday snap to collect reports.
* Experience in creating fact and dimensional model in MS SQL server and Snowflake Database utilizing Cloud base Matillion ETL tool.
* Conceptual and Logical IDA modeling for Governance, Risk, and Compliance areas: ISO 2, PCI DSS; Domain-specific metrics / measurement analysis; EA Modeling Analysis in Mega; Semantic Web & Visual Ontology Modeling; Canonical Data Services Modeling Analysis, RDF for Healthcare Exchanges; OWL; TOGAF; IaaS Magic Draw Modeling.
* Developed the SQL scripts and Procedures for the business rules using Unix Shell and NZSQL for Netezza.
* Continual monitoring of Data Center Support queues, troubleshoot and diagnose hardware problems.
* Experienced Hadoop Developer, have a strong background with file distribution systems in a big - data arena. Understands the complex processing needs of big data and have experience developing codes and modules to address those needs.
* Working in ETL methodology for supporting Data Analysis, Extraction, Transformations and Loading, in a corporate-wide-ETL Solution using Ab Initio.
* Implement requirements of the API layer like security, custom analytics, throttling, caching, logging, monetization, request and response modifications, rate limiting, developer portal using Apigee
* Worked extensively on the Google Cloud Platform (GCP) across various big data products such as Big Query, Cloud Dataproc, Google Cloud Storage, and Composer (Airflow as a service).
* Worked extensively on Spark RDD, Data Frame API, Data Set API, Data Source API, Spark SQL, and Spark Streaming, utilizing Spark Streaming APIs for real-time transformations and actions.
* Developed a Python Kafka consumer API for efficient data consumption from Kafka topics, especially in processing Extensible Markup Language (XML) messages using Kafka and Spark Streaming for capturing User Interface (UI) updates.
* Wrote python and shell scripts to perform operations or tasks that are currently not supported by snaplogic.
* Install and configured SQL Server 2014/2019, Visual Studio 2019 and configure Matillion ETL
* FIBO; canonical modeling; RESTful Services, JSON; NoSQL; Big Data; Hadoop; MarkLogic; Thought leader to engage world’s top visual ontology modeler to teach Semantic Web and Ontology Master Class at world class global healthcare leader’s formal education and training facility. RDF, OWL, SKOS, Turtle, Graph Databases, Franz, TopQuadrant, MarkLogic, SmartLogic, MongoDB, Hortonworks.
* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Experience in working with various CDC tools like Oracle Golden Gate, StreamSets and Striim
* Implemented a pre-processing job with Spark Data Frames to flatten JSON documents into flat files, showcasing adeptness in data transformation.
* Experience in developing Map Reduce Programs using Apache Hadoop for analyzing the big data as per the requirement.
* Extract data from various source systems like SFTP server, Google Drive, Email server and RQ4 (POS) system, using Snaplogic. store in AWS using S3 storage and process it using EMR, Spark and python.
* Used different AWS Data Migration Services and Schema Conversion Tool along with Matillion ETL tool
* Experienced in designing Star Schema, Snowflake schema for Data Warehouse, by using tools like Erwin data modeler, Power Designer and Embarcadero E-R Studio.
* Mentor and guide analyst on building purposeful analytics tables in dbt for cleaner schemas.
* Configured Snow pipe to extract data from Google Cloud buckets into Snowflake tables, demonstrating expertise in cloud-based data integration.
* Demonstrated a solid understanding of Cassandra architecture, replication strategy, gossip, snitches, etc.
* Utilized Hive QL for analysing partitioned and bucketed data, executing queries on Parquet tables.
* Employed Apache Kafka to aggregate web log data from multiple servers for downstream systems, actively contributing to the implementation of Kafka security measures and performance enhancements.
* Developed Oozie coordinators for scheduling Hive scripts, contributing to the creation of efficient data pipelines.
* Provide guidance to development team working on PySpark as ETL platform
* Perform regular health checks and monitor performance of the Apigee environment
* Involved in daily operational activities in order to troubleshoot ad-hoc production and data issues and enhancement of infrastructure in the space of Big Data and AWS cloud to provide better solutions to delegate the existing issues.
* Managed Snaplogic servers, pipelines and scheduled/triggered tasks.
* Assessed the Netezza environment for implementation of the ELT solutions.
* Expertise in all components such as transform components, database components, sort, partition/de-partition in the GDE of Ab Initio for creating, executing, testing and maintaining graphs in Ab Initio and also experience with Ab Initio Co-operating System in application tuning and debugging strategies.
* Optimize the Pyspark jobs to run on Kubernetes Cluster for faster data processing
* Performing Data Center walkthrough verifying alerts for PDUs, HVACs and RPPs.
* Experience in GCP Dataproc, GCS, Cloud functions, BigQuery.
* Experienced in developing Adobe Analytics reports, tagging strategy as well as creating Splunk reports to measure response times/load time
* Involved in Big data requirement analysis, develop and design solutions for ETL and Business Intelligence platform.
* Netezza Architecture. Vetted IBM MDM (former) Initiate team’s product. Vetted Palantir approach by bringing in one of the world’s top Semantic Web Ontology Leaders to train 40 architects in Dynamic Ontology Concepts
* Designed, build and managed ELT data pipeline, leveraging Airflow, python, GCP, Stitch Data and GCP solutions.
* Create and developed data load and scheduler process for ETL jobs using Matillion ETL package
* Provided extensive Production Support for Data Warehouse for internal and external data flows to Netezza, Oracle DBMS from ETL servers via remote servers.
* Convert web services from SOAP to REST or vice-versa, using Apigee API management platform
* Model, lift and shift custom SQL and transpose LookML into dbt for materializing incremental views.
* Utilize Matillion and AWS Redshift for DW enactment of Epic Data Warehousing concept in Snowflake database
* Trained 2 employees on how to use Snaplogic integration tool, Amazon Redshift and related technologies, assigned tasks and managed one.
* Perform data masking and ETL process using S3, Informatica cloud, Informatica Power Center and Informatica Test Data Management to support Snowflake Data warehousing solution in the cloud.
* Expertise in design, development and implementation of Enterprise Data Warehouse solutions using Mediation Zone Digital Route and Talend ETL Big Data Integration suite version 6.2
* Conducted cluster testing of HDFS, Hive, Pig, and MapReduce, ensuring the smooth onboarding of new users to the cluster.

**Environment:** Spark, Spark-Streaming, Spark SQL, GCP, Stream sets, Data Proc, Map R, HDFS, Hive, Pig, Apache Kafka, Sqoop, Python, PySpark, Shell scripting, Linux, MySQL Oracle Enterprise DB, Big Data, SOLR, Jenkins, Eclipse, Oracle, Git, Oozie, Tableau, MySQL, Soap, Cassandra & Agile Methodologies.

**Homesite insurance Boston, MA April 2021 - March 2023**

**Role: Sr Data Engineer**

**Responsibilities:**

* Designed large-scale data processing using Azure Synapse Analytics and advanced SQL techniques.
* Achieved seamless integration and migration from legacy systems to SQL Server, Oracle, and Teradata.
* Experience in moving data between GCP and Azure using Azure Data Factory.
* Real time experience in Hadoop/Big Data related technology experience in Storage, Querying, Processing and analysis of data.
* Worked directly with Snaplogic support team to debug and fix issues and wrote python and shell scripts that run using script snap to temporarily fix issues or missing functionalities like unzipping files and managing ftp files.
* Maintained all Microsoft Data center security/safety protocols in the environment.
* Developed enterprise data models, capturing relationships between, policies, policyholders, claims, and risks and collaborated with stakeholders to ensure data structures met business objectives.
* Implemented and managed database solutions with Azure SQL Database and Azure Cosmos DB, utilizing T-SQL for structured querying, indexing, and targeted database performance optimizations.
* Integrated PySpark with Azure Blob Storage, enabling efficient data storage and retrieval processes.
* Developed Python-driven solutions, utilizing azure tools and services, to effectively process and transform complex data, ensuring smooth integration and scalability within the Azure cloud ecosystem.
* Stage the API or Kafka Data (in JSON file format) into Snowflake DB by flattening the same for different functional services.
* Responsible for determining the bottlenecks and fixing the bottlenecks with performance tuning using Netezza Database.
* Agile project Scrum team member as T-shaped skill of Technical Data Analyst / ETL developer in Big Data apps.
* Utilize DBT to cofnvert raw, unstructured data into structured datasets, enabling efficient analysis and reporting.
* Design and help in development of Snaplogic pipelines to extract the processed data from the Data Lake to staging server, then using Informatica for data processing from staging server to within Teradata in multiple stages within the EDW.
* Created a Corporate Product Information Ontology using the Florida Institute for Machine & Human Cognition (IMHC) CmapTools Knowledge Modeling Kit.
* Built ontology-backed time series analysis and monitoring product for Palantir Foundry using
* Making automates the development of Ab initio graphs and functions utilizing the Meta data from EME.
* Optimization of Hive queries using best practices and right parameters and using technologies like Hadoop, YARN, Python, PySpark.
* Debug Apigee EDGE components
* Performed Data Migration to GCP
* Utilized Apache Spark with Python to develop and execute Big Data Analytics and Machine learning applications, executed machine learning use cases under Spark ML and Mllib.
* Hands on Experience in implementing Error Handling, Data Reconciliation, Audit/logging monitoring, Job scheduling using StreamSets and Python
* Understanding of structured data sets, data pipelines, ETL tools, data reduction, transformation and aggregation technique, Knowledge of tools such as DBT, Data Stage
* Evaluated suitability of Hadoop and its ecosystem to the above project and implementing / validating with various proof of concept (POC) applications to eventually adopt them to benefit from the Big Data Hadoop initiative.
* Performing transformations of source data with transform components in Ab initio
* Debug Apigee EDGE components design and development of REST API platform using APIGEE, converting web services from SOAP to REST or vice-versa
* Utilized Azure Data Factory and Hive to create ETL and ELT pipelines, orchestrating data movement between diverse data platforms such as SQL Server, Oracle (PL/SQL), and Teradata.
* Establish best DBT processes to improve performance, scalability, and reliability.
* Worked with Kafka to integrate data from multiple topics to database. Manage RESTful API, integrate with Stream sets to move data.
* Converting old oracle SQL/PL- SQL, Microsoft SQL server/ T- SQL written SQL to run on big data platforms using PYSPARK, SPARKSQL and HIVE with SAS and Python as programming platforms
* Perform testing used to built-in plug-in and 3rd party debugging tools such as Adobe debugger
* Utilized Apache Airflow in the GCP Composer environment to construct robust data pipelines, incorporating operators such as Bash, Hadoop, Python callable, and branching operators.
* Utilize troubleshooting methodology to assist various IT teams in analyzing problems and resolving Data Center issues.
* Created Netezza Sql scripts to test the table loaded correctly
* Developed spark applications in python (PySpark) on distributed environment to load huge number of CSV files with different schema in to Hive ORC tables.
* Interest and experience in cloud data platforms like Snowflake for scalable and efficient data processing.
* Employed Azure Stream Analytics for real-time data processing, monitoring, and analytics.
* Deployed and managed Hadoop clusters using Azure HDInsight, integrating Hive for querying, and managing large insurance datasets, further enhancing data analytics capabilities.
* Utilized Azure Data bricks for large-scale data engineering and analytics tasks, leveraging the collaborative environment of Databricks notebooks for code development in Python, SQL, and Scala.
* Integrated Azure Databricks with Azure Data Factory for complex ETL operations, allowing for streamlined data transformations and enrichments necessary for insurance analytics.
* Orchestrated secure delta lakes using Azure Databricks platform, ensuring timely and consistent data availability for in-dept analysis and processing in critical downstream analytics processes.
* Established and optimized CI/CD pipelines using Azure DevOps, streamlining automated testing and deployment of Python and Hive scripts related to insurance data solutions.
* Have Extensive Experience in IT data analytics projects, Hands on experience in migrating on premise ETLs to Google Cloud Platform (GCP) using cloud native tools such as BIG query, Cloud Data Proc, Google Cloud Storage, and Composer.
* Design, develop, configure, and troubleshoot APIs and policies using APIGEE
* Automated feature engineering mechanisms using Python scripts, deploying them on Google Cloud Platform (GCP) and Big Query.
* Fix Teradata utility failures and ability to handle other AWS, Snaplogic, Informatica and Teradata related errors by making necessary code overrides.
* Creating Applications on Splunk to analyze the Big Data.
* Created data sharing between two snowflake accounts.
* Developing Ab Initio graphs using GDE (Graphical Development Environment) to extract the data from DB2 tables & mainframe Files (legacy source system) to Oracle Tables (target database fields).
* Experienced in Performance Engineering and testing using Adobe Analytics and Splunk.
* Effectively utilized Azure Monitor and Azure Log Analytics for real-time monitoring, guaranteeing optimal performance of SQL, Hive, and various Python-driven tasks, maximizing efficiency.
* Used cloud shell SDK in GCP to configure the services Data Proc, Storage, BigQuery
* Orchestrated containerized applications using Azure Kubernetes Service (AKS), ensuring Python applications and Hive tasks are scalable and easily manageable even under heavy loads.
* Implemented Azure Policy and Azure Blueprints to ensure SQL and Hive-driven solutions continuously aligned with and adhered to organizational compliance norms and standards.
* Strong understanding of Data Modelling (Relational, dimensional, Star and Snowflake Schema), Data analysis, Palantir Foundry, implementations of Data warehousing using Windows and UNIX.
* Leveraged Azure Machine Learning Service and Python for predictive analytics, enhancing understanding and prediction of insurance trends and claims, ensuring informed approach to business strategy.
* Applied Azure Resource Manager (ARM) templates and utilized Terraform techniques, ensuring robust and consistent deployment of SQL and Hive solutions across multiple environments.
* Worked on analyzing Hadoop stack and different big data analytic tools including Pig and Hive, HBase database and Sqoop.
* Worked on reading and writing multiple data formats like JSON, ORC, Parquet on HDFS using PySpark.
* Design, develop, and maintain scalable data models and transformations using DBT in conjunction with Snowflake, ensure the effective
* Maintained the Hadoop cluster on GCP, utilizing Google Cloud Storage, Big Query, and Data proc.
* Configured GCP services (Data Proc, Storage, Big Query) using Cloud Shell SDK in GCP, showcasing proficiency in cloud infrastructure management.
* Create custom policies in Java, JavaScript, in Apigee
* Leveraged GCP for Cloud Functions, event-based triggering, Cloud Monitoring, and Alerting, and implemented G-cloud functions with Python for loading data into Big Query from newly arrived CSV files in GCS buckets.
* Experience working on Vagrant boxes to setup a local Kafka and StreamSets pipelines
* Incorporated Azure Event Hub and Azure Service Bus for event-driven programming capabilities using SQL triggers and Python-driven events for better data responsiveness.
* Developed interactive dashboards and reports using Power BI, sourcing data from multiple databases like Oracle and Teradata, offering stakeholders real-time insights and analytics.

**Environment: Azure Synapse Analytics, SQL, Oracle, GCP, Teradata, Azure SQL Database, SQL Server, T-SQL, Azure Cosmos DB, PySpark, Azure Blob Storage, Python, Azure Data Factory, Hive, Azure Stream Analytics, Hadoop, Azure HDInsight, Stream sets, Azure Data bricks, Scala, Azure DevOps, CI/CD, Azure Monitor, Big Data, Azure Log Analytics, Azure Kubernetes Service (AKS), Azure Policy, Azure Blueprints, Azure Machine Learning Service, Azure Resource Manager (ARM), Terraform, Azure Event Hub, Azure Service Bus, Power BI.**

**Meta Platforms, New York City, NY May 2020 to March 2021  
Role: Sr. Data Engineer**

**Responsibilities:**

* Established a robust Continuous Delivery pipeline by seamlessly integrating Docker and GitHub.
* Developed and deployed Spark and Scala solutions on a Hadoop cluster within the Google Cloud Platform (GCP) environment, showcasing expertise in cloud-based data processing.
* Configured Snow pipe to extract data from Google Cloud buckets into Snowflake tables, demonstrating expertise in cloud-based data integration.
* Demonstrate substantial depth of knowledge and experience in a specific area of Big Data and development.
* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Migrate legacy transformation code into modular DBT data models.
* Worked on Dimensional Data modelling in Star and Snowflake schemas and Slowly Changing Dimensions (SCD).
* Responsible for sending quality data thru secure channel to downstream systems using role base access control and Stream sets.
* Developed Spark/Scala, Python for regular expression (regex) project in the Hadoop/Hive environment with Linux/Windows for big data resources.
* Analysed the SQL scripts and designed it by using PySpark SQL for faster performance.
* Leveraged Google Cloud Functions with Python to efficiently load data into Big Query, specifically handling incoming CSV files stored in Google Cloud Storage (GCS) buckets.
* Exhibited proficiency in managing various Azure and AWS cloud services such as HDInsight, Databricks, Data Lake, Blob storage, Data Factory, Synapse, SQL DB, SQL DWH, EC2, and Lambda to ensure efficient resource utilization and data solutions.
* Used cloud shell SDK in GCP to configure the services Data Proc, Storage, BigQuery
* Processed and loaded both bound and unbound data from Google Pub/Subtopics to Big Query utilizing Cloud Dataflow with Python, demonstrating proficiency in event-driven data processing.
* Applied Spark and Scala APIs hands-on to conduct performance comparisons with Hive and SQL, utilizing Spark SQL for effective manipulation of Data Frames in Scala.
* Redesigned the Views in snowflake to increase the performance.
* Devised and executed simple and complex SQL scripts to verify Dataflow in various applications.
* Conducted comprehensive Data Analysis, Migration, Cleansing, Transformation, Integration, Import, and Export tasks using Python.
* Mastered major Hadoop distributes like Horton Works and Cloudera numerous Open-Source projects and prototype various applications that utilize modern Big Data tools.
* Worked on expertise with big data technologies (HBASE, HIVE, MAPR PIG and Talend).
* Intensively used Python, JSON & Groovy scripts coding to deploy the Streamsets pipelines into the server.
* Unit tested the data between Redshift and Snowflake.
* Manage Data Storage and processing pipelines in GCP for serving AI and ML services in Production, development and testing using SQL, Spark, Python and AI
* Architected multiple Directed Acyclic Graphs (DAGs) for automating ETL pipelines, demonstrating proficiency in workflow management.
* Hadoop, Cloudera CDH 4.5, HDFS, PIG Scripting, Hive, Map Reduce, Sqoop, Flume, Oozie, Spark, Autopsy’s, Unix scripting, Tableau, Talend Big data ETL.
* Proposed and implemented new DBT design in the client's ecosystem.
* Validating the data from SQL Server to Snowflake to make sure it has Apple to Apple match.
* Implemented monitoring solutions in Ansible, Terraform, Docker, and Jenkins, ensuring robust system health and performance tracking.
* Automated Datadog Dashboards through Terraform Scripts, streamlining visualization and reporting.
* Hands-on experience in architecting ETL transformation layers and writing Spark jobs for efficient data processing.
* Gathered and processed raw data at scale using various methods such as scripting, web scraping, API calls, SQL queries, and application development.
* Proficient in fact-dimensional modelling (Star schema, Snowflake schema), transactional modelling, and Slowly Changing Dimension (SCD) techniques.
* model development, and real-time streaming analytics, to derive actionable insights from large and diverse datasets.
* Implemented Hadoop clusters for processing big data pipelines using Amazon EMR and Cloud era whereas it depended on Apache Spark for fast processing and for the integration of APIs
* Implemented Apache Airflow for authoring, scheduling, and monitoring Data Pipelines, ensuring streamlined workflow management.
* Proficient in machine learning techniques such as Decision Trees, Linear/Logistic Regressors, and statistical modelling.
* Designed, build and managed ELT data pipeline, leveraging Airflow, python, dbt, Stitch Data and GCP solutions.
* Integrated data from Cloudera Big data stack, Hadoop, Hive, Hbase, and MongoDB. Build Stream sets pipeline to accommodate change.
* Develop stored procedures/views in Snowflake and use in Talend for loading Dimensions and Facts.
* Developed DBT deployment process for Starburst Integrations.
* Experience in GCP Dataproc, GCS, Cloud functions, BigQuery, Azure Data Factory Data Bricks.
* Experience in building efficient pipelines for moving data between GCP and Azure using Azure Data Factory.
* Implemented machine learning back-end pipelines with Pandas and NumPy, showcasing proficiency in data manipulation and analysis.

**Environment:** GCP, Big Query, GCS Bucket, G-Cloud Function, Apache Beam, Cloud Dataflow, Cloud Shell, Gsutil, Docker, Kubernetes, AWS, Apache Airflow, Python, Pandas, Dtabricks, Matplotlib, seaborn library, text mining, NumPy, Scikit-learn, Heat maps, Bar charts, Line charts, ETL workflows, Big data, linear regression, multivariate regression, Python, Scala, Spark.

**KPMG, Hyderabad, Telangana, India          November 2017 to March 2020  
Role: Data Engineer**

**Responsibilities:**

* Extensive hands-on exposure to the AWS cloud platform, showcasing proficiency in key services such as EC2, S3, EMR, Redshift, Lambda, and Glue.
* Demonstrated expertise in Spark RDD, Data Frame API, Data Set API, Data Source API, Spark SQL, Spark Streaming, SQL, and MongoDB within the AWS ecosystem.
* Developed and deployed data pipelines in cloud environments, with a particular focus on AWS, demonstrating a strong understanding of core components like EC2 and S3.
* Implemented Spark applications using Python and R for Apache Spark data processing projects, effectively managing data from diverse relational databases (RDBMS) and streaming sources.
* Developed and deployed Spark and Scala solutions on a Hadoop cluster within the Google Cloud Platform (GCP) environment, showcasing expertise in cloud-based data processing.
* Proficient in various Google Cloud components, Google Container Builders, GCP client libraries, and Cloud SDKs.
* Experience in GCP Dataproc, GCS, Cloud functions, BigQuery, Azure Data Factory DataBricks.
* Leveraged Apache Spark Data Frames, Spark-SQL, and Spark MLlib extensively, designing and implementing proof-of-concepts (POCs) with Scala, Spark SQL, and MLlib libraries.
* Specialized in data integration, employing traditional ETL tools and methodologies to ingest, transform, and integrate structured data into a scalable data warehouse platform.
* Designed and deployed multi-tier applications on AWS, utilizing services like EC2, Route53, S3, RDS, DynamoDB, SNS, SQS, and IAM, with a strong emphasis on high availability, fault tolerance, and auto-scaling using AWS Cloud Formation.
* Utilized Azure Data bricks for big data processing and analytics, including data exploration, machine learning model development, and real-time streaming analytics, to derive actionable insights from large and diverse datasets.
* Proficient in Python and Scala, developing user-defined functions (UDFs) for Hive and Pig using Python.
* Extracted data from SQL Server, Teradata, Amazon S3 buckets, and internal SFTP, loading them into the data warehouse AWS S3 bucket.
* Developed PySpark POCs and deployed them on the Yarn Cluster, conducting performance comparisons between Spark, Hive, and SQL/Teradata.
* Created Spark jobs for data processing, including instance and cluster creation, and loaded the data into AWS S3 buckets, creating DataMart’s.
* Stored data efficiently in GCP Big Query Target data warehouse, catering to diverse business teams with tailored use cases.
* Utilized AWS EMR for processing and transforming data, collaborating with the Data Science team based on specific business requirements.
* Designed and implemented ETL processes in AWS Glue to migrate campaign data from external sources (S3, ORC/Parquet/Text Files) into AWS Redshift.
* Worked on both batch processing and Realtime data processing on Spark Streaming using the Lambda architecture.
* Demonstrated hands-on experience in GCP, covering services such as Big Query, GCS bucket, G-cloud function, Cloud Dataflow, Pub/Sub, Cloud Shell, GSUTIL, BQ command-line utilities, Data Proc, and Stack driver.
* Developed Spark applications for cleaning and validating ingested data into the AWS cloud.
* Executed simple to complex MapReduce jobs using Java for data processing and validation.
* Demonstrated hands-on experience in GCP, covering services such as Big Query, GCS bucket, G-cloud function, Cloud Dataflow, Pub/Sub, Cloud Shell, GSUTIL, BQ command-line utilities, Data Proc, and Stack driver.
* Created scripts to load data to Hive from HDFS and ingested data into the Data Warehouse using various data loading techniques. Utilized Spark Streaming APIs for real-time transformations and actions.
* Developed pre-processing jobs using Spark Data Frames to flatten JSON documents into flat files.
* Loaded D-Stream data into Spark RDD, performed in-memory data computations to generate output responses.
* Involved in building ETL to Kubernetes with Apache Airflow and Spark in GCP, emphasizing workflow automation.
* Implemented Kubernetes for the runtime environment of the CI/CD system, covering building, testing, and deployment.
* Collaborated with the DevOps team to implement NiFi Pipeline on EC2 nodes integrated with Spark, Kafka, Postgres, running on other instances using SSL handshakes in QA and Production Environments.
* Built Informatica mappings, sessions, and workflows, managing code changes through version control in Informatica.

**Environment:** Spark, Spark-Streaming, Spark SQL, AWS EMR, Scala, MapReduce, HDFS, Hive, Pig, Apache Kafka, Sqoop, Python, PySpark, Shell scripting, Linux, DataBricks, MySQL Oracle Enterprise DB, Big query, SOLR, Jenkins, Eclipse, Dataflow, Oracle, Git, Oozie, Tableau, MySQL, GCP, Soap, Cassandra and Agile Methodologies.  
  
**WebSoc Technologies, Hyderabad, Telangana, India November 2015 to October 2017**

**Role: Data Engineer**

**Responsibilities:**

* Operated effectively within an Agile environment, utilizing the Rally tool to manage user stories and tasks for streamlined project management.
* Implemented ad-hoc analysis solutions through the utilization of Data Lake Analytics/Store and HDInsight, showcasing a versatile approach to data analytics.
* Established Apache Sentry to exert granular control over access to Hive tables, particularly focusing on group- level access management.
* Demonstrated proficiency in MapReduce programming using Java, PIG Latin Scripting, and Distributed Application and HDFS, showcasing a comprehensive skill set in distributed data processing.
* Leveraged Tidal enterprise scheduler and Oozie Operational Services for seamless coordination of clusters and scheduling workflows, ensuring efficient cluster management.
* Integrated various azure services, such as Azure Data Factory, Azure Data Lake, Azure Data Warehouse, Azure Active Directory, Azure SQL Database, and Web Apps, with AWS services to harness the advantages of both cloud platforms and effectively process unstructured data.
* Designed and implemented Kafka clusters by configuring Topics across all environments, ensuring a robust and scalable messaging system.
* Developed multiple Tableau dashboards catering to diverse business needs, emphasizing the ability to translate complex data into actionable insights.
* Implemented Partitioning, Dynamic Partitions, and Buckets in HIVE to optimize data access, showcasing a strategic approach to data storage and retrieval.
* Architected and implemented medium to large-scale BI solutions on Azure, leveraging Azure Data Platform services, including Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Databricks, and NoSQL DB.
* Utilized AVRO format for entire data ingestion to enhance operational speed and reduce storage space consumption.
* Designed SSIS Packages for ETL operations, extracting, transferring, and loading existing data into SQL Server from different environments for SSAS cubes (OLAP).
* Ingested data into one or more Azure Services (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processed the data in Azure Databricks, showcasing proficiency in Azure data processing tools.
* Developed visualizations and dashboards using PowerBI, ensuring clear and effective communication of analytical findings.
* Implemented Composite server for data virtualization needs and created multiple views for restricted data access using a REST API, showcasing versatility in data access and management.
* Exported analysed data to relational databases using Sqoop for visualization and report generation for the BI team using Tableau, highlighting a comprehensive approach to data reporting.
* Developed Apache Spark applications for data processing from various streaming sources, showcasing expertise in real-time data processing.
* Developed data pipelines using Spark, Hive, Pig, Python, Impala, and HBase to ingest and process customer data, showcasing adaptability in selecting appropriate tools for specific data processing tasks.
* Converted Hive/SQL queries into Spark transformations using Spark RDDs, Python, and Scala, ensuring efficient data processing in Spark.
* Queried and analysed data from Cassandra for quick searching, sorting, and grouping through CQL, showcasing expertise in NoSQL data processing.
* Joined various tables in Cassandra using Spark and Scala, running analytics on top of them, demonstrating proficiency in leveraging Spark for complex data operations.

**Environment:** MapR, Map Reduce, HDFS, Hive, pig, Impala, Kafka, Cassandra, Spark, Scala, Azure (SQL, Databricks, Datalake, Data Storage, HDInsight), Java, SQL, Tableau, PIG, Zookeeper, Sqoop, Kafka, Teradata, Power BI.

**Arisoft Technologies, Hyderabad, Telangana, India July 2014 to October 2015**

**Role: Data Analyst**

**Responsibilities:**

* Collaborated with Business Analysts and SMEs to gather business requirements and identify actionable items, aligning with strategic goals.
* Implemented robust data cleansing processes with the ETL development team, leveraging Apache Pig for data processing.
* Stored data securely in AWS S3 and AWS Redshift using AWS EC2 instances for efficient data management.
* Used PySpark and Pandas for complex calculations and analysis of stocks in the data warehouse.
* Optimized Hadoop with Spark, Spark-SQL, PostgreSQL, Data Frames, OpenShift, Talend, and pair RDDs for efficient data processing.
* Developed custom-built connectors using Spark, Sqoop, and Oozie, enabling efficient data ingestion and processing from Relational Database Management Systems (RDBMS) to Azure Data Lake.
* Effectively utilized Azure Synapse to manage processing workloads and serve data for Business Intelligence (BI) and predictive analytics, ensuring efficient data analysis and insights.
* Integrated Kafka and Spark for real-time data processing, harnessing the power of real-time data analytics and improving data handling and processing capabilities.
* Developed Spark SQL scripts and Scala shell commands to efficiently process and analyze data, leveraging Spark's distributed computing capabilities.
* Exhibited proficiency in managing various Azure and AWS cloud services such as HDInsight, Databricks, Data Lake, Blob storage, Data Factory, Synapse, SQL DB, SQL DWH, EC2, and Lambda to ensure efficient resource utilization and data solutions.
* Designed and optimized data pipelines within Azure Data Factory and Spark, effectively connecting Azure SQL, Blob storage, and Azure SQL Data warehouse, enhancing data processing workflows.
* Exhibited expertise in building data lake and cloud data platform solutions in the financial services industry, showcasing domain-specific knowledge.
* Effectively integrated RESTful Web Services (REST APIs) into data pipelines, enhancing data flow between system components.
* Utilized API Gateways like APIGEE to bolster web service endpoint security, ensuring safe communication between components.
* Demonstrated expertise in concurrency, parallelism, and data processing optimization, contributing to efficient workflows and system performance.
* Guided data pipeline and ML development cycles, employing ML knowledge to design and deploy data-driven solutions for enhanced decision-making.
* Showcased expertise in working with NoSQL database technologies, including MongoDB, Bigtable, and DynamoDB.
* Designed, developed, and maintained effective build and deployment pipelines using tools such as Visual Studio, PyCharm, Git/Bitbucket/Bamboo, Maven, Jenkins, and Nexus.

**Environment:** Apache Pig, AWS S3, AWS EC2, AWS Redshift, Python, PySpark, Pandas, Spark, Spark-SQL, PostgreSQL, Data Frames, OpenShift, Talend, Pair RDDs, AWS Lambda, Java, ELK (Elasticsearch, Logstash, Kibana), Apache Kafka, Zookeeper.