**Name: Maqsood**

**Sr. Data Engineer**

**PROFESSIONAL SUMMARY**

Accomplished **Senior Data Engineer** with a track record of **driving innovative** solutions for **big data challenges** for over **9+ years**.

Proven expertise in **designing** and implementing **end-to-end ETL pipelines**, leveraging **advanced technologies** and **cloud platforms**. Skilled in translating complex business needs into scalable **Big Data solutions**, **crafting strategies**, and **delivering results**. Proficient in a wide range of tools and technologies, including **Hadoop**, **Spark**, **AWS**, **GCP**, and **data quality validation**. Adept at **streamlining data** **processes**, **optimizing performance,** and ensuring **data accuracy**. Strong team collaborator, mentor, and problem solver, committed to delivering excellence in data engineering.

**Experience Highlights:**

* **AWS certified professional with 9+** years of experience in **Data Engineering,** **Big Data**/**ETL solution** development, **Data Visualization** & **Reporting**
* Extensively worked on **AWS services** like **EC2, S3, EMR, Sage Maker, RDS(Aurora), Redshift, DynamoDB, Elastic Cache (Memcached & Redis) & Quick Sight**, **Databricks**, **Athena, Glue** and other services of the AWS family
* Experienced in Designing, Developing, Documenting, and **Testing** **ETL** jobs and mappings in Server and Parallel jobs using Data Stage to populate tables in **Data Warehouse** and **Data marts.**
* Designed and built data pipelines using **Azure Data Factory (ADF)** to extract, transform, and load data from diverse sources like **Azure SQL**, **Blob storage**, and Azure SQL Data Warehouse.
* Hands-on experience in using **Hadoop ecosystem** components like **Hadoop, Hive, Pig, Sqoop, HBase, Cassandra, Spark Streaming, Spark SQL, Zookeeper, Kafka, Flume, MapReduce framework, Yarn, Scala, and Hue**
* Experience with cloud technologies like **AWS Step Functions, AWS Lambda,** **Quick Sight**, **Cloud Watch, Glue**, **Athena, Redshift, IAM, EMR, IAM, SNS.**
* This professional possesses in-depth expertise in the Hadoop ecosystem, including components like **HDFS**, **MapReduce**, **Hive**, **Pig, HBase**, and **Spark**.
* With a hands-on approach, they have successfully managed the installation, configuration, and maintenance of **Hadoop clusters**, showcasing a strong command over cluster management.
* Their programming proficiency extends to key languages such as **Java**, **Python**, and **Scala**, enabling them to excel in writing and optimizing MapReduce jobs tailored for large-scale data processing.
* In summary, their comprehensive skill set positions them as a valuable asset in leveraging **Hadoop** for effective and efficient **data analytics** and processing.
* Demonstrated proficiency in **Java programming** specifically tailored for data engineering tasks.
* Specialized focus on developing solutions for **data processing** and analytics applications.
* Extensive hands-on experience in crafting and fine-tuning **MapReduce** jobs.
* Proven ability to optimize these jobs for efficient large-scale data processing within the **Hadoop** ecosystem.
* Experience **in ETL/ELT ingesting** data from various sources into **Data Warehouses** using business transformations and **star/snowflake data modeling.**
* Experience in data processing like collecting, aggregating, and moving data from various sources using **AWS Glue** creating **Glue crawler jobs, Glue Catalogues, and Python scripts.**
* Experience in developing **CI/CD** (continuous integration and continuous deployment) and automation using **Git and Kubernetes**
* Expertise in **Data Migration, Data Profiling, Data Cleansing**, **Transformation, Integration, Data Import,** and **Data Export** using multiple **ETL** **tools** such as **Informatica** **Power Centre, SSIS**
* Extensively used **Python pytest, pyodbc, NumPy, Pandas, MySQL dB, sqlite3, snowflake-python-connector,** and other packages
* Developed various **shell scripts** and **Python scripts** to automate **Spark jobs** and **Hive jobs.**
* Strong Knowledge of the Architecture of Spark distributed and parallel processing using **Spark SQL**, **Spark Data frames APIs**, and Spark execution framework.
* Experience in **OLAP, OLTP**, Business Intelligence, and **Data Warehousing** with emphasis on ETL and Business Reporting needs
* Worked on Key Performance Indicators **KPIs**, design of **star schema**, and **snowflake schema** in Analysis Services (**SSAS)**
* Designed and implemented data warehouses and data marts using conformed **Facts & Dimensions, Slowly Changing Dimensions (SCD)**, change data capture **(CDC)**, Surrogate Keys, Star, and Snowflake Schema
* Experience in **Performance Tuning** and **Debugging** of existing ETL processes using **Informatica** tools.
* Experience in creating Views, **clustered and non-clustered Indexes**, Stored Procedures, Triggers, **Window functions**, aggregate functions, and Table Value Functions using DDL, DML and PL/T-SQL
* Experience in converting **SQL** queries into Spark Transformations using **Spark RDDs, Spark SQL, Spark Data frames**, and developing **python** objects in the **OOPS programming paradigm.**
* Experience in working with **CSV, text, Excel, parquet, Jason** formats of data, and data from **RESTful API**
* Experience with **Requests**, **NumPy, SciPy, Matplotlib, HTTPLib2, Urllib2, Beautiful Soup, Data Frame**, and **pandas python** **libraries** during the development lifecycle
* Employed **AWS Lambda** and **Python scripts** for event-driven processing and achieved a 30% reduction in testing time through enhanced data validation with **Query Surge**.
* Expertly managed **Hadoop clusters** expanded Hive functionality with custom **UDFs**, and leveraged **Sqoop** for **data extraction.**
* Proficient in **data modeling** and **statistical analysis techniques**, including machine learning on **Spark** and **Hadoop clusters**.
* Demonstrated expertise in real-time data stream processing with **Apache Kafka** and integrated it seamlessly with **Hadoop components.**
* Collaborated across cross-functional teams to develop **analytics-driven** solutions and conducted **data migration** projects to improve **scalability** and **performance**.
* Constructed robust **ETL processes** for **Druid** using **Apache Nifi** and **AWS Glue**, resulting in enhanced data processing efficiency.
* Developed and maintained data governance practices, ensuring **data quality, lineage**, and **metadata management**.
* Successfully crafted microservices architectures using Docker, Kubernetes, and Apache Mesos to manage and scale big data clusters.
* Proficient in **SQL, PL/SQL,** and database management for optimizing query performance.
* Skilled in creating technical documentation and providing mentorship to fellow **ETL developers**.
* Skilled in correlation and multivariate analysis, **causal analytics, and advanced Regression Modelling** using statistical modeling concepts.
* Experience in analyzing the data using MS Excel (**VBA, V - lookup, H – lookup, pivot tables, what-if analysis**, scenarios, link cells, charts) and MS Access
* Translated intricate business challenges into scalable **Big Data solutions**, **devising strategies**, and **crafting roadmaps** for data initiatives.
* Managed multiple tasks and worked under **tight deadlines and in a fast-paced environment.**
* Team player, quick learner, organized and self-motivated.

A dedicated Data Engineer with a strong commitment to driving data-driven insights and delivering impactful solutions. Excited to contribute expertise to the success of a dynamic data engineering team.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Big Data Technologies** | Kafka, Cassandra, Apache Spark, Spark Streaming, HBase, Impala, HDFS, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Zookeeper, Flink, Presto, Druid, Delta Lake, Apache Beam, Apache Nifi, Apache NiFi Registry, Apache Airflow, Apache Druid, Apache Pulsar, AWS Kinesis, Google Cloud Dataflow, Google BigQuery, Quantexa, Databricks Delta, Snowflake, Azure Data Lake, Azure Stream Analytics. |
| **Hadoop Distribution** | Cloudera CDP, Hortonworks HDF, Apache Hadoop, Amazon EMR, Microsoft HDInsight, Google Cloud Dataproc, Hadoop |
| **Programming Languages** | SQL, PL/SQL, Python, R, PYSpark, Pig, Hive QL, Scala, Shell Scripting, Regular Expressions, Java, Kotlin, Go, Rust |
| **Spark components** | RDD, Spark SQL (Data Frames and Dataset), Spark Streaming, Spark MLlib, Spark GraphX, SparkR, PySpark |
| **Cloud Infrastructure** | AWS, Azure, GCP, Alibaba Cloud, IBM Cloud, Oracle Cloud |
| **Databases** | Oracle, Teradata, MySQL, SQL Server, NoSQL Databases (HBase, MongoDB, Cassandra, Couchbase, DynamoDB) |
| **Scripting & Query Languages** | Shell scripting, SQL, GraphQL, Apache Drill |
| **Version Control** | CVS, SVN, Clear Case, GIT, Bitbucket, GitHub, GitLab |
| **Build Tools** | Maven, SBT, Gradle |
| **Reporting Tools** | Tableau Desktop, Tableau Server, Power BI |
| Data Quality & Monitoring | Nagios; Ganglia; Query Surge; Datagaps ETL Validator |

**PROFESSIONAL EXPERIENCE:**

**Sr . Data Engineer**

**Peapod Digital Labs - Chicago, Illinois July 2021 - Till Date**

**Responsibilities:**

* Led the installation, configuration, and maintenance of advanced data pipelines, ensuring optimal performance and reliability.
* Comprehensive understanding of ETL (Extract, Transform, Load) and ELT (Extract, Load, Transform) concepts.
* Proficient in utilizing popular ETL tools such as Apache NiFi, Talend, and Informatica.
* Skilled in designing and implementing efficient data extraction processes from various sources.
* Translated intricate business challenges into scalable Big Data solutions, crafting meticulous data strategies and roadmaps.
* Extracted data from Hadoop and securely transferred it to S3 storage at regular intervals.
* Designed and executed end-to-end data solutions on Google Cloud Platform (GCP), including data ingestion, ETL processes, data modeling, and data warehousing through BigQuery and Cloud Storage.
* Developed custom Python (PySpark) scripts for User-Defined Functions (UDFs) to manage data transformations, cleansing, and aggregation.
* Utilized Pig scripts to generate MapReduce jobs for Hadoop Distributed File System (HDFS) data processing.
* Identified process enhancement opportunities and implemented ETL solutions using Informatica, Python, and scheduling tools.
* Expertise in transforming and cleansing data to ensure accuracy and consistency.
* Experience in loading data into data warehouses or data lakes using tools like Amazon Redshift or Apache Hadoop.
* Strong knowledge of data integration techniques and best practices.
* Analyzed root causes of production issues and resolved data anomalies, ensuring data reliability.
* Fine-tuned code for optimization, performance, and efficient data processing, enhancing overall data workflows.
* Actively participated in requirement gathering, analysis, and comprehensive documentation of business needs.
* Built diverse ETL solutions from various sources, including SQL scripting, ETL tools, Python, and scheduling tools.
* Proficient in data profiling and wrangling of XML, web feeds, and files using Python, Unix, and SQL.
* Transformed and aggregated data from multiple sources into a data warehouse using Python.
* Expertise in automating OpenStack and AWS deployments through Cloud Formation, Ansible, Chef, and Terraform.
* Familiarity with data modeling and schema design for optimal ETL/ELT processes.
* Ability to troubleshoot and optimize ETL workflows for performance improvements.
* Hands-on experience with scripting languages (e.g., Python, SQL) for data manipulation and transformation.
* Knowledge of parallel processing and distributed computing in the context of ETL/ELT operations.
* Played a key role in implementing Datagaps ETL Validator for end-to-end validation of complex ETL workflows, improving ETL efficiency.
* Stay updated on emerging trends and technologies in the ETL/ELT space for continuous improvement.
* Integrated data sources into a unified GCP data platform with Pub/Sub and Cloud Dataflow.
* Designed and deployed Sqoop jobs for incremental data extraction from DB2 and loading into Hive tables for interactive reporting.
* Utilized Sqoop to transfer data from diverse sources to HDFS and RDBMS.
* Developed PySpark and Spark SQL applications for data extraction, transformation, and aggregation from various file formats.
* Implemented data pipelines with Apache Druid for real-time and batch data storage and querying.
* Built Talend data pipelines for event-driven architectures, real-time data streaming, and API workflows to ensure data integrity.
* Implemented data quality checks and monitoring processes with QuerySurge, reducing data discrepancies and enhancing reliability.
* Proficient in working with the Hadoop ecosystem components such as HDFS, MapReduce, Hive, Pig, HBase, Spark, etc.
* Experience with the installation, configuration, and maintenance of Hadoop clusters.
* Strong proficiency in programming languages commonly used in the Hadoop ecosystem, such as Java, Python, and Scala.
* Writing and optimizing MapReduce jobs for large-scale data processing.
* Optimized Druid clusters for performance and cost-efficiency.
* Orchestrated data pipelines using Kubernetes and Docker.
* Proficient in dimensional modeling, including Star and Snowflake Schemas.
* Familiarity with the BI Stack, including SSRS, Tableau, and Power BI.
* Proficient in Java programming for data engineering tasks, with a focus on data processing and analytics applications.
* Extensive experience in developing and optimizing MapReduce jobs for large-scale data processing in Hadoop ecosystem.
* Strong understanding of Java libraries and frameworks commonly used in data engineering, such as Apache Hadoop and Apache Spark.
* Expertise in designing and implementing scalable and efficient data pipelines using Java technologies.
* Extensive experience in Amazon Web Services (AWS), including EC2, S3, DynamoDB, Redshift, CloudFormation, CloudTrail, Kinesis, and more.
* Implemented customized UDFs in Python.
* Utilized ETL methodology for Data Migration, Data Profiling, Extraction, Transformation, and Loading with Talend.
* Scalability Microservicesenable scalable data engineering solutions by breaking down complex systems into smaller, independent services.
* Modularity Data engineering microservices are designed as modular components, allowing for easier development, testing, and maintenance.
* Data Processing Specialized microservices can handle distinct data processing tasks, such asingestion, transformation, and storage, promotinga more streamlined workflow.
* Flexibility Microservices architecture provides flexibility in choosing the right technology stack for each specific data processing task, optimizing performance.
* Fault Isolation Isolating different data processing tasks within microservices helps contain failures, preventing cascading issues throughout the entire system.
* APIsand Communication Microservices communicate through well-defined APIs, enabling seamless integration with other services and facilitating data flow between components.
* Monitoring and Logging Each microservice can have its own monitoring andlogging, aiding in identifying and resolving issues efficiently.
* Parallel Development Teams can work on different microservices concurrently, promoting parallel development and faster time-to-market for data engineering solutions.
* Resource Efficiency Microservices allow for resource-efficient scaling, as resources can be allocated based on the specific needs of each service.
* Distributed Architecture Data engineering microservices often operate in a distributed architecture, enhancing resilience and reducing single points of failure.
* Microservices support CI/CD practices, enabling rapid and reliable deployment of changes to the data engineering infrastructure.
* Design and implement data infrastructure to support AI applications.
* DevelopETL (Extract, Transform, Load) processes for ingesting and processing large datasets.
* Collaborate with data scientists to integrate machine learning models into data pipelines.
* Optimize databases and storage systems for efficientAI model training and inference.
* Implement data quality and validation processes to ensure accuracy in AI-driven insights.
* Utilize cloud platforms and distributed computing for scalable AI data processing.
* Design and maintain data warehouses to store structured and unstructured data forAI applications.
* Ensure data security and compliance with privacy regulations in AI data handling.
* Collaborate with cross-functional teams to understand AI requirements and deliver effective data solutions.
* In-depth knowledge of life science concepts, terminology, and data requirements Data Proficient in integrating diverse data sources within the life science domain, including clinical trials, genomics, and healthcare data .
* Expertise in designing and implementing data models tailored to life science data, ensuring effective storage and retrieval Skilled in developing Extract, Transform, Load (ETL) processes to cleanse, transform, and move data between systems, ensuring data quality and consistency.
* Experience in designing and managing data warehouses specific to life science needs for efficient data storage and retrieval Familiarity with big data tools and technologies for processing and analyzing large-scale life science datasets.
* Adherence to regulatory requirements and implementing robust security measures to protect sensitive life science data Implementation of data governance practices to ensure data quality, privacy, and compliance with industry standards.
* Effective collaboration with cross-functional teams, including scientists, researchers, and IT professionals, to understand data requirements and deliver solutions.
* Keeping abreast of evolving technologies and methodologies in both data engineering and life science to enhance skills and stay relevant in the field.

**Environment:** AWS, Azure, Google Cloud, SQL Server, T-SQL, SQL Server Integration Services (SSIS),

SQL Server Reporting Services (SSRS), Databricks, SQL Server Analysis Services (SSAS), Management Studio

(SSMS), Advance Excel (creating formulas, pivot tables, Hlookup, Vlookup, Macros), Terraform, Spark,

Kafka, Impala, Python, Power BI, Tableau, Presto,Hive/Hadoop, Snowflakes, TensorFlow,Keras,

Scikit-Learn, Kubernetes, CDP, Docker, Jenkins, Trifacta, Talend, Informatica, Snowflakes.

**Sr . Big Data Engineer**

**Johnson & Johnson - Edison, New Jersey April 2019 - June 2021**

**Responsibilities:**

* Developed and maintained data pipelines on Google Cloud Platform (GCP), ensuring efficient data flow and processing.
* Leveraged Apache Nifi to streamline data ingestion from various sources into the cloud environment.
* Creating pipelines in Azure Data Factory (ADF) using Linked Services, Datasets, and Pipelines to extract, transform, and load data from various sources such as Azure SQL, Blob storage, Azure SQL Data Warehouse, and write-back tools.
* Comprehensive understanding of ETL (Extract, Transform, Load) and ELT (Extract, Load, Transform) concepts.
* Proficient in utilizing popular ETL tools such as Apache NiFi, Talend, and Informatica.
* Skilled in designing and implementing efficient data extraction processes from various sources.
* Extracting files from Hadoop and transferring them to S3 on a daily or hourly basis. Working with data governance and data quality to design models and processes.
* Creating a comprehensive data dictionary and mapping from sources to the target in MDM (Master Data Management) data models, covering all steps and scope of the project.
* Managing Azure Data Lakes (ADLS) and Data Lake Analytics, integrating with other Azure services. Familiarity with USQL.
* Working with cross-functional teams to develop analytics-based solutions targeting customer subscribers.
* Utilizing AWS Lambda to run Python scripts and Java for event-driven processing. Configuring Lambda jobs and roles using AWS CLI.
* Expertise in transforming and cleansing data to ensure accuracy and consistency.
* Experience in loading data into data warehouses or data lakes using tools like Amazon Redshift or Apache Hadoop.
* Strong knowledge of data integration techniques and best practices.
* Utilized QuerySurge to validate data completeness, accuracy, and consistency across various data sources, ensuring compliance with regulatory standards.
* Performing data ingestion using Sqoop and HDFS commands, accumulating partitioned data in various storage formats like text, JSON, Parquet, etc. Loading data from Linux file systems to HDFS.
* Storing data files in Google Cloud Storage (GCS) buckets on a daily basis. Developing and maintaining GCP cloud-based solutions using DataProc and BigQuery.
* Assisted in data migration projects, transferring legacy data systems to GCP for improved scalability and performance.
* Utilizing AWS for storage and handling terabytes of data for customer BI reporting tools.
* Building and managing Hadoop clusters, including adding and decommissioning nodes for maintenance.
* Monitoring cluster health by setting up alerts using Nagios and Ganglia.
* Built Terra grunt project to manage Terraform conffiguration file DRY while working with multiple terraforms modules and worked with Terraform templates to automate the Azure Iaas virtual machines using terraform modules and deployed virtual machine scalle seets in prodjuctiion environment.
* Managing user and group permissions as per client requests.
* Developed and maintained data governance practices, including metadata management, data lineage, and data quality checks.
* Built robust ETL processes to extract, transform, and load data from diverse sources into Druid using technologies such as Apache Nifi and AWS Glue.
* Collaborated with AWS Lambda for serverless data processing tasks, optimizing resource utilization and cost.
* Administered data quality checks using Datagaps ETL Validator, reducing data anomalies and improving data reliability.
* Utilized Terraform for infrastructure as code, automating cloud resource provisioning.
* Managed version control and code repositories on GitLab, enabling a collaborative and organized development environment.
* Worked with Docker and Kubernetes for containerization and orchestration of data processing components.
* Designed and implemented real-time data streaming solutions using Apache Kafka and Apache Druid, enabling real-time analytics.
* Integrated Tableau Server for secure and interactive data visualization accessible to stakeholders.
* Using Apache Spark, Spark SQL, Spark MLlib, Scala, and ML libraries extensively for data processing, analysis, and POC (Proof of Concept) development.
* Led the adoption of best practices for data governance and data security in a multi-cloud environment.
* Implemented a data catalog and metadata management system to facilitate data discovery and lineage tracking.
* Familiarity with data modeling and schema design for optimal ETL/ELT processes.
* Ability to troubleshoot and optimize ETL workflows for performance improvements.
* Hands-on experience with scripting languages (e.g., Python, SQL) for data manipulation and transformation.
* Collaborated with data scientists to optimize machine learning model deployments on GCP.
* Implemented data partitioning and compaction strategies to improve query performance in big data storage.
* Managed AWS S3 and Azure Blob Storage for data storage and archival, reducing storage costs.
* Designed and automated data migration processes for seamless transition to cloud-based solutions.
* Implemented data retention policies and archiving strategies to manage data lifecycle.
* Collaborated with the DevOps team to create Dockerized data processing containers, simplifying deployment.
* Led efforts to implement continuous monitoring and anomaly detection in the data pipelines.
* Assisted in designing and optimizing the company's data governance framework and compliance policies.
* Conducted regular training sessions for the team on cloud-native data engineering best practices and technologies.
* Scalability Microservicesenable scalable data engineering solutions by breaking down complex systems into smaller, independent services.
* Modularity Data engineering microservices are designed as modular components, allowing for easier development, testing, and maintenance.
* Knowledge of parallel processing and distributed computing in the context of ETL/ELT operations.
* Stay updated on emerging trends and technologies in the ETL/ELT space for continuous improvement.
* Data Processing Specialized microservices can handle distinct data processing tasks, such asingestion, transformation, and storage, promotinga more streamlined workflow.
* Flexibility Microservices architecture provides flexibility in choosing the right technology stack for each specific data processing task, optimizing performance.
* Fault Isolation Isolating different data processing tasks within microservices helps contain failures, preventing cascading issues throughout the entire system.
* APIsand Communication Microservices communicate through well-defined APIs, enabling seamless integration with other services and facilitating data flow between components.
* Monitoring and Logging Each microservice can have its own monitoring andlogging, aiding in identifying and resolving issues efficiently.
* Parallel Development Teams can work on differentmicroservices concurrently, promoting parallel development and faster time-to-market for data engineering solutions.
* Resource Efficiency Microservices allow for resource-efficient scaling, as resources can be allocated based on the specific needs of each service.
* Distributed Architecture Data engineering microservices often operate in a distributed architecture, enhancing resilience and reducing single points of failure.
* Microservices support CI/CD practices, enabling rapid and reliable deployment of changes to the data engineering infrastructure.
* Design and implement data infrastructure to support AI applications.
* DevelopETL (Extract, Transform, Load) processes for ingesting and processing large datasets.
* Collaborate with data scientists to integrate machine learning models into data pipelines.
* Optimize databases and storage systems for efficientAI model training and inference.
* Implement data quality and validation processes to ensure accuracy in AI-driven insights.
* Utilize cloud platforms and distributed computing for scalable AI data processing.
* Design and maintain data warehouses to store structured and unstructured data forAI applications.
* Ensure data security and compliance with privacy regulations in AI data handling.
* Collaborate with cross-functional teams to understand AI requirements and deliver effective data solutions.

**Environment**: Hadoop, Kafka, Spark, Sqoop, Spark SQL, Spark-Streaming, Hive, Terraform,Impala, Scala, pig,

NoSQL, Oozie, Hbase, Data Lake, Python , QuerySurge, Azure, Databricks, AWS(Glue, Lambda, StepFunctions,

SQS, Code Build, Code Pipeline, EventBridge, Athena), Unix/Linux Shell Scripting, CDP, Informatica PowerCenter.

**Data Engineer/AWS**

**JP Morgan Chase - Fairfax, Virginia Aug 2017 - Mar 2019**

**Responsibilities:**

* Proficiently transformed complex business problems into robust Big Data solutions.
* Created feature, scenario, and step definitions for Behavior-Driven Development (BDD) and Test-Driven Development (TDD) using Cucumber, Gherkin, and Ruby.
* Designed data collection approaches based on project scope and Software Development Life Cycle (SDLC) methodologies.
* Constructed pipelines in Azure Data Factory for data extraction, transformation, and loading from diverse sources.
* Extracted files from Hadoop and transferred them to S3 storage on a regular basis.
* Created comprehensive data dictionaries and mappings for Master Data Management (MDM) data models.
* Managed Azure Data Lakes and Data Lake Analytics.
* Collaborated with QA and BA teams to expedite defect resolution and requirements clarification.
* Integrated AWS Lambda for Python and Java-based event-driven processing.
* Utilized QuerySurge for data validation and testing.
* Handled data ingestion via Sqoop and HDFS commands.
* Stored data files in Google Cloud Storage (GCS) with data processing in DataProc and BigQuery.
* Led data migration projects from legacy systems to Google Cloud Platform (GCP).
* Leveraged AWS for storage and data handling.
* Managed Hadoop clusters, health monitoring, and alert setup using Nagios and Ganglia.
* Used Terraform for managing infrastructure.
* Designed and implemented scalable and efficient data architectures on AWS.
* Developed data models for optimized storage, retrieval, and analytical processing.
* Implemented end-to-end ETL processes using AWS Glue, Apache Spark, and custom scripts.
* Managed data pipelines for ingestion, processing, and transformation of large datasets.
* Set up and maintained data warehouses on Amazon Redshift, optimizing for performance and query efficiency.
* Managed data distribution, indexing, and partitioning strategies.
* Implemented real-time data processing using AWS Kinesis for timely insights.
* Developed solutions for real-time analytics and monitoring.
* Integrated diverse data sources into a unified data platform, ensuring consistency and integrity.
* Collaborated with cross-functional teams to understand and meet data integration requirements.
* Managed user and group permissions.
* Implemented Datagaps ETL Validator for ETL workflow validation.
* Achieved data integration with Pub/Sub and Cloud Dataflow.
* Designed and deployed Sqoop jobs for data extraction and loading into Hive tables.
* Developed Spark applications using PySpark and Spark SQL.
* Created data pipelines using Apache Druid.
* Constructed Talend data pipelines for various use cases.
* Implemented data quality checks and monitoring with QuerySurge.
* Optimized Druid clusters.
* Utilized Kubernetes and Docker for cluster services.
* Created automated multi-dimensional cubes using SQL Server Integration Services (SSIS).
* Achieved real-time data processing with Spark Streaming and NoSQL databases.
* Accomplished data integration with Apache NiFi.
* Collaborated with Anti-Money Laundering (AML) compliance teams for Actimize implementation.

**Environment:** AWS, Azure, Google Cloud, SQL Server, T-SQL, SQL Server Integration Services (SSIS), SQL Server Reporting Services (SSRS), Databricks, SQL Server Analysis Services (SSAS), Management Studio (SSMS), Advance Excel (creating formulas, pivot tables, Hlookup, Vlookup, Macros), Terraform, Spark, Kafka, Impala, Python,

Power BI, Tableau, Presto, Hive/Hadoop, Snowflakes, TensorFlow, Keras, Scikit-Learn, Kubernetes, CDP, Docker, Jenkins, Trifacta, Talend, Informatica, Snowflakes.

**Big Data Engineer**

**KPMG - Orlando, Florida Feb 2016 - Aug 2017**

**Responsibilities:**

* Designed and constructed end-to-end ETL pipelines, leveraging GCP data services like Cloud Dataflow, Cloud Storage, and BigQuery.
* Employed Query Surge to enhance data validation performance, resulting in a 30% reduction in testing time and improved ETL processing speed.
* Proficiently utilized AWS components like EC2, S3, and implemented Continuous Delivery pipelines through Docker and GitHub.
* Created serverless data ingestion pipelines using GCP Cloud Functions and Python to load real-time CSV data into BigQuery from GCS buckets.
* Employed Apache Kafka for seamless data streaming and seamlessly integrated it with Hadoop ecosystem components for real-time data processing.
* Established data quality and profiling frameworks using Impala, Apache Atlas, and Apache NiFi to maintain data accuracy, completeness, and consistency.
* Worked extensively within GCP and AWS cloud environments, developing and deploying scalable data pipelines and machine learning back-end pipelines, with the use of Apache Airflow, Kubernetes, and Docker.
* Demonstrated proficiency in data processing, analysis, and visualization by utilizing Pandas, NumPy, and Matplotlib libraries.
* Possessed expertise in Hadoop ecosystem components such as HDFS, Apache Hive, and Apache Spark for efficient storage, processing, and analysis of big data.
* Applied experience in fact dimensional modeling, transactional modeling, and Slowly Changing Dimensions (SCD) within data warehousing.
* Demonstrated proficiency in statistical modeling and machine learning techniques, including Decision Trees and Linear/Logistic Regressors, and implemented machine learning models on Spark and Hadoop clusters.
* Developed PL/SQL Stored Procedures, Functions, Triggers, Views, and Packages for optimizing database management systems and query performance.
* Leveraged Apache NiFi, Apache Flink, and Apache Beam for real-time data stream processing, enabling insights and alerts generation.
* Proficiently crafted microservices architectures using containerization and orchestration technologies like Docker, Kubernetes, and Apache Mesos, to effectively manage and scale big data clusters.
* In-depth involvement in ETL pipeline creation using GCP data services.
* Utilized QuerySurge for data validation and testing.
* Managed AWS components such as EC2 and S3.
* Implemented serverless data ingestion with GCP Cloud Functions.
* Worked with Apache Kafka for data streaming.
* Conducted data quality and profiling using Impala, Apache Atlas, and Apache NiFi.
* Operated in AWS and GCP cloud environments for scalable data pipelines.
* Utilized Pandas, NumPy, and Matplotlib for data processing and analysis.
* Proficient in Hadoop ecosystem components, including HDFS, Hive, and Spark.
* Applied fact dimensional modeling and transactional modeling techniques.
* Implemented machine learning on Spark and Hadoop clusters.
* Developed PL/SQL solutions for database management.
* Accomplished real-time data stream processing with Apache Flink and Apache Beam.
* Managed containerization and orchestration using Docker, Kubernetes, and Apache Mesos.

**Environment**: Hadoop, Kafka, Spark, Sqoop, Spark SQL, Spark-Streaming, Hive, Terraform,Impala, Scala, pig, NoSQL, Oozie, Hbase, Data Lake, Python , QuerySurge, Azure, Databricks, AWS(Glue, Lambda, StepFunctions, SQS, Code Build, Code Pipeline, EventBridge, Athena), Unix/Linux Shell Scripting, CDP, Informatica PowerCenter.

**Data Engineer/GCP**

**First Republic Bank - Raleigh, North Carolina Aug 2014 - Feb 2016**

**Responsibilities:**

* Assisting in designing the overall ETL solutions, including analyzing data, preparing high-level and detailed design documents, test and data validation plans, and deployment strategies.
* Preparing technical mapping specifications, process flows, and error handling documents.
* Developing both simple and complex mappings, implementing business logic using various transformation logic such as unconnected and connected Lookups, Router, Filter, Expression, Aggregator, Joiner, Update Strategy, Unconnected and Connected Stored Procedures, Normalizer, and more.
* Creating tasks such as Pre/Post Session Commands, Timer, Event Wait, Event Raise, Email, and Command tasks.
* Experience in writing real-time processing using Spark Streaming with Kafka.
* Using HiveQL to analyze partitioned and bucketed data and compute metrics for reporting.
* Querying data using SparkSQL on top of the Spark engine.
* Managing and monitoring Hadoop clusters using Cloudera Manager.
* Using Python and Shell scripting to build data pipelines.
* Developing data pipelines using Sqoop, HQL, Spark, and Kafka to ingest enterprise message delivery data into HDFS.
* Developing workflows in Oozie and Airflow to automate tasks for loading data into HDFS and pre-processing with Pig and Hive.
* Assisting in creating and maintaining technical documentation for launching Hadoop clusters and executing Hive queries and Pig Scripts.
* Writing Autosys Jil files to run workflow components and deploy files.
* Extensive work with database components such as SQL, PL/SQL, Stored Procedures, Stored Functions, Packages, and Triggers.
* Performing code review and troubleshooting of existing Informatica mappings and deploying code from development to test to production environments.
* Supporting other ETL developers by providing mentoring, technical assistance, troubleshooting, and alternative development solutions.
* Playing a key role in designing comprehensive ETL solutions, encompassing data analysis, design documentation, test plans, and deployment strategies.
* Creating technical mapping specifications, process flow diagrams, and error handling protocols.
* Developing both straightforward and complex data transformations, implementing various business logic using tools like Lookups, Routers, Filters, Expressions, Aggregators, Joiners, and more.
* Configuring tasks including Pre/Post Session Commands, Timers, Event Waits, Event Raises, Emails, and Command tasks.
* Proficient in real-time data processing using Spark Streaming with Kafka.
* Utilizing HiveQL to analyze partitioned and bucketed data, extracting metrics for reporting purposes.
* Querying data with SparkSQL on top of the Spark engine.
* Managing and monitoring Hadoop clusters via Cloudera Manager.
* Employing Python and Shell scripting to construct data pipelines.
* Developing data pipelines with Sqoop, HQL, Spark, and Kafka to ingest enterprise message delivery data into HDFS.
* Design and implement scalable and reliable data architectures on GCP.
* Create and maintain data models for efficient storage and retrieval.
* Develop and implement data pipelines for ingesting, processing, and transforming data.
* Use tools like Apache Beam or Cloud Dataflow for scalable ETL (Extract, Transform, Load) processes.
* Utilize GCP's Big Data services, such as Bigtable, Dataproc, and Datastore, for large-scale data processing.
* Optimize and tune performance for big data applications.
* Document data engineering processes, pipelines, and architectures.
* Adhere to best practices in data engineering and GCP services usage.
* Designing workflows using Oozie and Airflow for automating tasks related to data loading into HDFS and preprocessing with Pig and Hive.
* Collaborating on the creation and maintenance of technical documentation for launching Hadoop clusters and executing Hive queries and Pig Scripts.
* Writing Autosys Jil files to orchestrate workflow components and file deployments.
* Demonstrating proficiency in database technologies such as SQL, PL/SQL, Stored Procedures, Functions, Packages, and Triggers.
* Conducting code reviews, troubleshooting existing ETL components, and managing code deployment across development, testing, and production environments.
* Providing mentoring, technical assistance, troubleshooting support, and alternative development solutions to fellow ETL developers.

**Environments:** Hadoop, HDFS, Spark, Hive, Pig, Sqoop, Oozie, DB2, Java, Python, Oracle, Sql, Splunk, UNIX, Shell Scripting, Apache Kafka, Apache NiFi, Apache Flume, Amazon EMR, Amazon Redshift, Google BigQuery,

TensorFlow, PyTorch, Apache MXNet, Apache Ranger.

**EDUCATION QUALIFICATION**:

Bachelors - Computer Science St. Mary’s College, Hyderabad, India Sep 2009 - May 2013

Masters - Computer Science Campbellsville University, Kentucky, USA June 2013 - July 2014