|  |
| --- |
| **SHAYZA**  Senior Data Engineer |

[sandyjones@coherenttechinc.com |](mailto:sandyjones@coherenttechinc.com | ) (407) 205 2187 | <https://www.linkedin.com/in/shayza-shavi-857051239/>

**Overview**

Dedicated and versatile Data Engineer with a strong proficiency in cloud platforms, including AWS and Azure, alongside advanced skills in Python and SQL. Proven track record in leveraging cloud technologies to drive data-driven decision-making and enhance data pipelines for organizations.

**Professional Summary**

* Overall **9+ years** of experience in project development, implementation, deployment, and maintenance.
* Extensive experience in designing, implementing, and maintaining data solutions on both **Azure and AWS cloud platforms.**
* Proficient in developing robust **Extract, Transform, Load (ETL)** processes using tools like Apache Spark, PySpark, and Azure Data Factory.
* Deployed and configured **data infrastructure** in both Azure and AWS environments.
* Experience in designing and implementing **data ingestion pipelines** using Azure Data Factory and AWS Glue to collect data from various sources, including databases, APIs, and file systems.
* Proficiency in utilizing Azure Event Hubs and AWS Kinesis for **real-time data streaming.**
* Created **data transformation** workflows with Azure Databricks and AWS EMR to process and cleanse raw data.
* Managed **data storage** solutions in Azure, including Azure Blob Storage and Azure Data Lake Storage, and in AWS using Amazon S3.
* Strong experience in designing and implementing **data warehousing** solutions in Azure SQL Data Warehouse and AWS Redshift for scalable and high-performance analytics.
* Orchestrated **data workflows and dependencies** using Azure Logic Apps and AWS Step Functions to ensure data processing accuracy and reliability.
* Experience in managing Azure Key Vault and AWS Key Management Service (KMS) for **key and secret management.**
* Established **monitoring and alerting systems** using Azure Monitor and AWS CloudWatch to proactively identify and address performance issues.
* Implemented Infrastructure as Code (IaC) using Azure DevOps and AWS CodePipeline to **automate deployment** and updates of data infrastructure.
* Implemented **disaster recovery strategies** and backup solutions to ensure data resilience in Azure and AWS environments.
* Extensive experience with a variety of **big data tools** and frameworks such as Hadoop, Spark, Kafka, and Hive.
* Exceptional understanding of **Hadoop architecture** and various components such as HDFS, Job Tracker, Task Tracker, Name Node and Data Node.
* Utilized **Python** for data extraction, transformation, and loading (ETL) tasks, enabling efficient data processing.
* Developed **custom Python scripts** for data cleansing, data validation, and automation of ETL processes.
* Conducted data transformation and cleansing using **SQL queries**, Python and PySpark.
* Designed and implemented end-to-end data pipelines for batch and real-time data processing, ensuring data integrity and reliability.
* Designed and implemented **real-time data processing** solutions using technologies like Apache Kafka, Azure Stream Analytics, and AWS Kinesis.
* Experience in working with Databases like **Oracle 10g, DB2, SQL Server 2008 and MySQL** and proficiency in writing complex SQL queries.
* Successfully executed data migration projects, transferring large datasets from on-premises environments to Azure and AWS cloud using **MySQL** as the database engine.
* Conducted performance tuning and optimization of **MySQL databases**, resulting in improved query performance and reduced response times for critical applications.
* Proficient in integrating **PL/SQL** code with third-party applications using APIs and web services, enabling seamless data exchange between systems.
* Designed and implemented **NoSQL database** solutions on both Azure and AWS platforms, including Azure Cosmos DB, Amazon DynamoDB, and MongoDB.
* Proficient in optimizing **NoSQL queries** for improved performance by employing indexing, partitioning, and sharding strategies.
* Proficient in database schema design, indexing, query optimization, and data replication strategies.
* Hands-on experience with **MongoDB, HBase, Cassandra, and Cosmos DB**, including data modeling, performance tuning, and scalability.
* Skilled in designing data-cleansing, configuration-driven ETL frameworks and pipelines in Python.
* Utilized **Python libraries** such as **Pandas and NumPy** to clean and transform raw data into structured formats suitable for analysis.
* Managed and maintained **relational databases (e.g., PostgreSQL, MySQL) using Python** for tasks like schema design, indexing, and query optimization.
* Developed **Python scripts** for data migration and database backup processes.
* Identified and resolved performance bottlenecks in data pipelines by optimizing **Python** code, SQL queries, and resource utilization.
* Proficient in optimizing **Spark** jobs for performance and resource utilization.
* Implemented ETL (Extract, Transform, Load) pipelines using **PySpark** for data ingestion and transformation.
* Proficient in working with various **data serialization formats, including Parquet, Avro, ORC, and JSON,** for efficient storage and retrieval.
* Ensured compliance with **data privacy regulations (e.g., GDPR, HIPAA)** in data handling.
* Proficient in using **Git for version control**, ensuring efficient tracking and management of code changes throughout data engineering projects.
* Conducted code reviews on **GitHub**, identifying and addressing issues related to code quality, performance, and adherence to coding standards.
* Designed and developed **Tableau dashboards** and reports to visualize complex data sets, providing actionable insights for business stakeholders.
* Conducted regular quality assurance checks on **Tableau** reports to identify and rectify data inconsistencies and visualization errors.
* Leveraged **Informatica Intelligent Cloud Services (IICS)** to orchestrate complex data integration tasks, integrating on-premises and cloud data sources seamlessly.
* Designed and implemented data warehousing solutions using Apache Hive, enabling efficient querying and analysis of large-scale structured and semi-structured data.
* Knowledge of **GCP** services such as BigQuery, Cloud DataProc, Cloud Storage and Cloud Composer.
* Conducted performance tuning and capacity planning exercises for **Kubernetes** clusters to handle increasing data processing demands effectively.
* Expertise in using various **Airflow Operators** including PythonOperator, BashOperator, SQLOperator, BigQueryOperator, S3FileTransformOperator and DataProcSparkOperator.
* Good experience with various IDEs like Eclipse, IntelliJ.
* Created Git repos and Jenkins jobs for the build and deployment of code to different environments and also to the Docker Container.
* Strong experience in setting up **CI/CD workflows** using AWS CodePipeline, which automates the build, test, and deployment processes.
* Proficient in integrating AWS CodeCommit, CodeBuild, and CodeDeploy to achieve end-to-end automation in the CI/CD pipeline.
* Creating and managing project & issue tracking in JIRA for Agile methodology to track bugs, stories, epics, and other tasks.
* Involved in daily SCRUM meetings to discuss the development/progress and was active in making scrum meetings more productive.
* Good analytical and communication skills and ability to work independently with minimal supervision and perform as part of a team.
* Proactive mindset, a strong dedication towards work responsibilities and a quick learning ability.

**Technical Skills**

|  |  |
| --- | --- |
| Big Data Technologies | HDFS, Hive, Flume, Oozie, Avro, Hadoop Streaming, Zookeeper, Kafka, Impala, Apache Spark, hue, Ambari.  Apache ignite. |
| Hadoop Distributions | Cloudera (CDH4/CDH5), Horton Works |
| Cloud | S3, Glue, Lambda, Athena, DynamoDB, EMR, EC2, Cloud Watch, Blob Storage, ADF, Databricks, HDInsight, Azure Functions, Synapse |
| Languages | Java, Python, SQL, PL/SQL, HQL |
| IDE Tools | Eclipse, IntelliJ, Net Beans, Anaconda |
| Framework | Hibernate, Spring, Struts, Junit |
| Operating Systems | Windows (XP,7,8), UNIX, LINUX, Mac OS-X, Ubuntu, CentOS |
| Application Servers | J Boss, Tomcat, Web Logic, Web Sphere, Servlets |
| Reporting Tools/ETL Tools | Tableau, Informatica, Teradata |

**Certifications**

* Microsoft Certified: Azure Data Engineer Associate <https://learn.microsoft.com/api/credentials/share/en-us/ShaikShayzaLnu-1984/2003E95D72A8F6B4?sharingId=98D62121A6EE90BB>

**Experience**

**Sr. Data Engineer**

Microsoft- Redmond, WA January 2022 – Present

**Responsibilities:**

* Designed, developed, and maintained data pipelines using **Azure Data Factory**, enabling efficient data extraction, transformation, and loading (ETL) processes.
* Managed and optimized **Azure SQL databases** for high-performance data storage.
* Developed data processing and analytics workflows using **Azure Databricks** and Apache Spark.
* Designed scalable data warehousing solutions on **Azure Synapse Analytics**.
* Managed data lakes and storage accounts on **Azure Blob Storage** and Azure Data Lake Storage Gen2.
* Developed real-time data processing solutions using **Azure Stream Analytics**.
* Set up and managed Hadoop and Spark clusters using **Azure HDInsight**.
* Designed globally distributed, highly available NoSQL databases using **Azure Cosmos DB**.
* Built and optimized data processing jobs using U-SQL on **Azure Data Lake Analytics**.
* Implemented continuous integration and continuous deployment (CI/CD) pipelines using **Azure DevOps.**
* Implemented Azure Monitor and Azure Security Center for proactive monitoring and security compliance.
* Used **Azure Data Catalog** to create a centralized metadata repository and data lineage documentation.
* Data ingestion into one or more Azure Services such as Azure Data Lake, Azure SQL, Azure Storage, Azure DW and processing in Azure Databricks.
* Integrated Azure Databricks with Azure Data Factory for big data processing, optimizing data workflows and improving data processing speed.
* Implemented complex data transformation logic using ADF Data Flows, including data mapping, data type conversions, aggregations, filtering, and custom data manipulations.
* Utilized **Delta Lake** features such as schema evolution, time travel, and optimization techniques for query performance.
* Designed and implemented data governance workflows and processes, ensuring compliance with industry regulations such as **GDPR, CCPA, and HIPAA.**
* Integrated the data catalog with data governance frameworks, ensuring compliance with regulatory requirements and facilitating data stewardship activities.
* Designed and built data pipelines on Azure Databricks using Spark-based transformations, including data cleansing, aggregation, data type conversions, and advanced analytics.
* Integrated Databricks with other Azure services like Azure Blob Storage, Azure Data Lake Storage, Azure SQL Database, etc.
* Automated ETL workflows by developing reusable mappings and workflows in Informatica, reducing manual intervention and ensuring timely data processing.
* Successfully migrated on-premises Informatica ETL processes to Azure Data Factory, optimizing performance, and reducing operational costs.
* Integrated Azure Functions with other Azure services like Azure Storage, Azure Event Hubs, Azure Cosmos DB, etc., to build robust and scalable serverless applications.
* Wrote Databricks notebooks (Python) for handling large volumes of data, transformations, and computations to operate with several types of file formats.
* Developed and optimized Spark applications using Spark Core, Spark SQL, Spark Streaming, or Spark MLlib.
* Utilized Databricks built-in libraries and tools, such as Delta Lake, MLflow, and Databricks Runtime, for advanced data engineering and machine learning tasks.
* Designed and implemented data warehousing solutions using **Snowflake**, including schema design, data loading, and optimizing query performance.
* Implemented medium to large scale BI solutions on Azure using Azure Data Platform services.
* Build de-bugging and production monitoring DAGs to help track the production workflow.
* Implemented **CI/CD pipelines** in Azure for automated build, test, and deployment of applications and infrastructure.
* Configured source code repositories, such as Azure Repos or GitHub, and integrated them with CI/CD pipelines.
* Developed and run UNIX shell scripts and implemented auto deployment process.
* Utilized machine learning algorithms such as linear regression, multivariate regression, PCA, K-means, & KNN for data analysis.
* Expertise with both version control platforms Git, and Agile methodologies and supporting tools Jira and Jira Service Desk.
* Involved in various phases of development analyzed and developed the system going through Agile Scrum methodology.
* Responsible for production support, on-call responsibilities, debugging, and rapid problem resolution.

**Environment:** Azure Data Factory (ADF), Azure SQL Database, Azure Databricks, Azure Synapse Analytics, Azure Blob Storage and Azure Data Lake Storage, Azure Stream Analytics, Azure HDInsight, Azure Cosmos DB, Azure Data Lake Analytics, Azure DevOps, CI/CD, Azure Monitoring and Security, Azure Data Catalog, Spark SQL, Delta Lake, Azure Functions, Python, Azure Synapse, Snowflake, Azure App Service, Azure Kubernetes Service (AKS), Azure Container Registry, NoSQL, Data Mining, Machine learning algorithms, Git, Jira.

**AWS Data Engineer**

Workday- Pleasanton, CA July 2020 - December 2021

**Responsibilities:**

* Worked on AWS services like **S3** for data storage, big data processing and analysis using EMR.
* Implemented data extraction, transformation, and loading (ETL) processes using **AWS Glue** for large-scale data pipelines, resulting in improved data quality and reduced processing time.
* Optimized **EMR** cluster configurations, including instance types, cluster size, and auto-scaling policies, resulting in significant cost savings and improved job performance.
* Developed and deployed serverless functions using **AWS Lambda** to process real-time data streams, handle data transformations, and trigger downstream data workflows.
* Leveraged **AWS CDK** constructs to define and deploy AWS Step Functions, orchestrating complex data workflows and ETL (Extract, Transform, Load) pipelines, resulting in optimized data processing and analysis.
* Utilized **AWS Step Functions** state management and error handling capabilities to handle exceptions, retries, and branching logic within the data processing workflows.
* Designed, implemented, and maintained real-time data pipelines using **AWS EventBridge** to capture and route events from various sources to target AWS services and custom applications.
* Implemented **AVRO schemas** to ensure data compatibility and efficient serialization in data processing workflows, reducing storage costs and improving data retrieval times.
* Demonstrated proficiency in working with **Parquet**, a columnar storage file format, to improve query performance and reduce data storage costs on Amazon S3.
* Implemented robust **IAM policies**, roles, and permissions to ensure the principle of least privilege, granting appropriate access to AWS resources for various team members and applications.
* Designed and implemented **AWS Lake Formation** as part of the data platform architecture, ensuring secure data storage and access controls for the organization's data lake.
* Implemented security measures within **Impala**, including role-based access controls and encryption, to safeguard sensitive data and comply with data privacy regulations.
* Utilized **SNS and SQS** for notification mechanisms and error handling, enabling efficient and fault-tolerant data processing and system.
* Integrated CodePipeline with source code repositories like GitHub or AWS CodeCommit, enabling seamless code versioning and continuous integration.
* Importing data from various sources like **MySQL database, MongoDB, SFTP** folder for converting raw data in structured format and analyzing.
* Utilized **AWS Airflow operators** (S3KeySensor, S3PrefixSensor, S3ToRedshiftTransfer, RedshiftToS3Transfer, EMRCreateJobFlowOperator, EMRStepOperator, LambdaFunctionOperator, ECSOperator, AWSBatchOperator, AWSAthenaOperator, GlueContext, EMRAddStepsOperator, DataPipelineOperator, EMRTerminateJobFlowOperator) to enhance the automation, scalability, and reliability of data processing workflows, ensuring timely and accurate data delivery to downstream systems.
* Created reusable **Terraform** modules to abstract common infrastructure patterns and configurations.
* Identified and resolved issues related to **Terraform** configurations and infrastructure deployments.
* Used Jira for bug tracking and Bit Bucket to check-in and checkout code changes.
* Wrote Shell scripts for several day-to-day processes and worked on its automation.
* Generating the necessary reports in Tableau by creating the workflow model with data lake in Hadoop.

**Environment:** AWS Services (S3, EMR, Glue, Lambda, Impala, Step Functions, DynamoDB, RDS, Redshift, SNS, SQS, CodePipeline, CloudWatch), MySQL, MongoDB, Hive, Spark, Apache Hudi, Kafka, Apache Airflow, Git, Apache Tomcat, BitBucket, Jira, Tableau, Shell Scripting, Terraform

**BIG Data Engineer**

Change healthcare- Nashville, TN May 2019 - June 2020

**Responsibilities:**

* Implemented and maintained Hadoop clusters, including **HDFS, YARN, and MapReduce**, to process and store large-scale data efficiently.
* Optimized the Hive tables using optimization techniques like partitions and bucketing to provide better performance with HiveQL queries.
* Involved in converting **Hive/SQL queries** into Spark transformations using Spark RDDs and Scala.
* Performed tuning of Spark applications to set batch interval time and correct level of parallelism and memory tuning.
* Implemented Spark Kafka integration to ingest real-time data streams from **Apache Kafka** into Spark for processing and analysis.
* Configured Spark Streaming's integration with Kafka to manage offsets and ensure fault-tolerant processing of Kafka data.
* Developed and managed data pipelines for extracting, transforming, and loading (ETL) data from various sources into **Teradata**, ensuring data integrity and accuracy.
* Developed complex **PL/SQL scripts** to extract, transform, and load (ETL) data from various sources into the data warehouse, ensuring data accuracy and consistency.
* Utilized Spark Kafka connectors and libraries to establish robust and scalable connections between Spark and Kafka.
* Configured and managed **Apache Tomcat** servers for hosting data-intensive web applications and APIs.
* Worked on **Apache Hudi** datasets on insert, upsert and bulk insert.
* Maintained the data in Data Lake (ETL), which originated from the Teradata.
* Implemented **Oracle RAC** (Real Application Clusters) to achieve high availability and load balancing, ensuring seamless data access even during hardware failures.
* Experienced in building automation using Jenkins, Maven, ANT.
* Leveraged **Apache Hive's** integration with **Apache Spark** and Hadoop ecosystems to perform distributed data processing and handle large volumes of data.
* Implemented data quality checks and validation rules in Java, ensuring data accuracy and completeness throughout the data pipeline.
* Integrated Docker with Kubernetes for orchestration and management of data engineering applications, leading to improved scalability and fault tolerance.
* Involved in Agile – Sprint methodologies to do requirements gathering, analysis and planning.
* Designed, developed, and implemented complex SSIS packages, asynchronous ETL processing, Ad hoc reporting, and SSRS report server, and data mining in SSAS.
* Prepared dashboards using calculations, parameters in Tableau and Developing Tableau reports that provide clear visualizations of various industry specific KPIs.

**Environment:** Python, Scala, Java, Apache Spark, Apache Kafka, AWS S3, Amazon EMR, Amazon Redshift, AWS Glue, Apache Hadoop, Teradata, PySpark, AWS Lambda, AWS Glue, AWS Data Pipeline, AWS EMR, Apache Hive, Amazon Redshift, DynamoDB, Git, Jenkins, Maven, ANT, Agile, Scrum, Tableau, SSIS, SSRS, SSAS, Oracle, Amazon Web Services (AWS) (S3, EMR, Redshift, DynamoDB, Lambda, Glue, CloudWatch), Tableau.

**Data Engineer**

**The Hartford- Hartford, CT** October 2017 - April 2019

**Responsibilities:**

* Perform data analysis to evaluate the data quality and resolve the data related issues.
* Worked on components of the Hadoop ecosystem such as Cloudera Hive, Apache Spark, and Apache Oozie for big data processing and analytics.
* Developed Hive queries and optimized them for better performance by using techniques like partitioning, bucketing, and indexing.
* Designed and maintained Hive tables, views, and functions to support data analysis and reporting requirements.
* Implemented Spark jobs using Spark Core, Spark SQL, and Spark Streaming to process and analyze large-scale datasets.
* Implemented error handling and exception mechanisms in PL/SQL code, leading to a more robust and resilient data processing system.
* Designed and implemented scalable Kubernetes clusters for container orchestration, managing critical data engineering applications and microservices.
* Optimized Spark applications by tuning parameters like batch interval, parallelism, and memory allocation.
* Monitored and managed Oozie workflows using the Oozie web console and command-line interface for troubleshooting and performance optimization.
* Implemented error handling and recovery mechanisms in IICS workflows, reducing data integration failures by 20% and enhancing data reliability.
* Implemented Oozie actions to execute tasks such as Hive queries, Spark jobs, and file operations.
* Implemented continuous integration and delivery pipelines using AWS CodePipeline.
* Configured and monitored system metrics, logs, and alarms using AWS CloudWatch.
* Implemented data migration strategies to move data from traditional RDBMS (Relational Database Management Systems) to Big Data platforms.
* Utilized data integration tools like Apache Kafka, Apache NiFi, or custom scripts to stream or batch process the migrated data.
* Developed logical and physical data models using Teradata's data modeling tools, ensuring data consistency and accuracy.
* Wrote Shell scripts to monitor load on database and Perl scripts to format data extracted from data warehouse based on user requirements.
* Developed Oozie workflows and scheduled them in Control-M as daily jobs to load incremental updates from the RDBMS source systems.
* Involved in gathering, understanding and documenting the data mapping requirements from the user through daily calls and continuous follow-ups.
* Monitored and optimized the performance of NoSQL databases, identifying and resolving bottlenecks to enhance overall system efficiency.
* Error validation of the data moving from flat file to the Oracle database.
* Used Bulk Collections for better performance and easy retrieval of data, by reducing context switching between SQL and PL/SQL engines.
* Debugged data issues between source systems and EDW.
* Conduct code review to ensure the work delivered by the team is of high-quality standards.

**Environment:** Cloudera Hive, Apache Spark, Apache Oozie, AWS services, Apache Kafka, Apache NiFi, Control-M, Oracle database, Prometheus monitoring toolkit, Splunk, Bulk Collections (in PL/SQL), EDW, Git, Jira, CI/CD tools (Jenkins), Tableau, SQL and HiveQL.

**Hadoop DEVeloper**

DITCOS- Hyderabad, India August 2014 - July 2017

**Responsibilities:**

* Used Sqoop to import and export data between Hadoop and relational databases.
* Utilized Sqoop's built-in support for Kerberos authentication to securely transfer data between Hadoop and external systems.
* Developed and optimized Hive queries using techniques like partitioning, bucketing, and indexing.
* Created and managed Hive tables, views, and functions to organize and manipulate data efficiently.
* Developed Spark applications using Spark Core, Spark SQL, and Spark Streaming.
* Worked with Spark SQL for querying structured data in Hadoop.
* Monitored Oozie workflows for troubleshooting and optimization.
* Designed and developed automated data ingestion pipelines using Sqoop and Apache Oozie, allowing seamless data movement from various data sources to Hadoop clusters on a scheduled basis, enhancing data availability for analytics teams.
* Integrated Oozie with other Hadoop ecosystem tools, such as Sqoop and Hive, to create end-to-end data processing pipelines.
* Created reports, dashboards, and visualizations based on SQL query results using tools like Tableau, Power BI, or Excel.
* Conducted data validation and integrity checks using SQL constraints, data profiling, and data cleansing techniques to ensure data accuracy and consistency.
* Collaborated with cross-functional teams to define key performance indicators (KPIs) and metrics, and developed SQL queries to measure and monitor performance.
* Conducted ad hoc data analysis using SQL to answer specific business questions and provide data-driven recommendations.
* Configured Flume to stream data into HDFS and Hive using HDFS Sinks and Hive sinks.
* Designed and implemented MapReduce-based large-scale parallel relation-learning system.
* Developed an application for filtering out data from consumer database and creating a deduped dataframe to store the latest and greatest data in the Hive tenants.
* Created several Databricks Spark jobs with PySpark to perform several tables to table operations.
* Developed automatic job flows and ran through Oozie daily and when needed which runs MapReduce jobs internally.
* Designed interfaces using Boom where quality and performance were of utmost importance.

**Environment:** Apache Sqoop, Apache Hive, Apache Spark, Apache Oozie, Apache HDFS, Hadoop, SQL, Tableau, Power BI, Microsoft Excel, Jenkins, Git, DevOps Pipelines, Flume, Data Ingestion Pipelines, MapReduce, PySpark, Databricks, Boom, Kerberos Authentication, Data Lake, KPIs, Metrics, Ad hoc Data Analysis, Data Validation, Data Migration.

**Education**

* Bachelor of Technology from JNTUH.