**Ram**

**Senior Data Engineer**

**PROFESSIONAL SUMMARY**

* 9+ years of experience designing and implementing data solutions for complex business problems.
* Expertise in Database Management Systems (DBMS), large-scale data warehousing, reporting solutions, data streaming, and real-time analytics.
* Proven experience in the Big Data ecosystem, including data acquisition, ingestion, modeling, storage analysis, integration, processing, and management.
* Experience across the entire Software Development Life Cycle (SDLC) using Agile and Waterfall methodologies.
* Skilled in cloud environments (AWS, Azure) with experience automating, configuring, and deploying instances.
* Proficient in Azure and AWS technologies, data warehousing solutions, ETL tools (ADF, SSIS, Glue), data ingestion, cleansing, staging, loading, and transformation.
* Expertise in data ingestion from various sources including relational databases (SQL Server, Oracle), unstructured data (CSV, JSON), and real-time data streams (Kafka, Kinesis).
* Proficient in data transformation using Apache Spark (Azure Databricks, AWS EMR) with PySpark and SQL.
* Experience in data warehousing solutions with Azure Synapse Analytics (Azure) and AWS Redshift (AWS).
* Experience in building and automating data workflows using cloud-based tools (Azure Data Factory, AWS Glue) and triggers (scheduled, event-based, tumbling window).
* Successfully migrated on-premises data to cloud data lakes (Azure Data Lake Store & AWS S3) using respective migration tools (ADF, DMS).
* Experience in managing and optimizing data lake resources (ADLS, S3) for cost-efficiency.
* Designed and implemented data pipelines for batch and real-time data processing using cloud services (Azure Stream Analytics, Event Hubs, Service Bus Queues) and Apache Airflow.
* Utilized DevOps practices (Azure DevOps & Jenkins) for CI/CD deployments.
* Strong understanding of cloud concepts (PaaS, IaaS) and numerous services offered by both Azure and AWS (Storage, Functions, Logic Apps, Monitoring, Key Vault, EC2, S3, Lambda).
* Experience in designing data models for data warehouses supporting multi-source data integration.
* Experience with data pre-processing and cleaning using Python libraries (NumPy, Pandas) for feature engineering and data imputation.
* Developed data analysis scripts and notebooks using Python and Spark for data manipulation.
* Experience in designing and developing dashboards for internal metrics using tools like Power BI and Azure Synapse Analytics.
* Working knowledge of data processing tools like SQL Server Integration Services (SSIS), Sqoop, and Apache Spark Streaming.
* Experience with NoSQL databases like Cassandra for real-time data loading, manipulation, and querying using CQL.
* Experience with Hadoop cluster management on AWS EC2 using Amazon EMR and S3 for backup storage.
* Experience with Elastic Load Balancing for auto-scaling in EC2 instances.
* Experience with workflow orchestration tools like Apache Airflow and Oozie for managing and scheduling Hadoop and Spark jobs.
* Utilized Python for pattern matching and formatting warnings/errors in build logs.
* Experience in using analytical applications like Python, R, and Java to identify trends and translate analytical findings into actionable insights.
* Developed data analysis scripts using Python libraries to filter and structure unstructured data.
* Built Spark applications in Databricks for data transformation and aggregation before loading into Azure Synapse Analytics.
* Result-oriented professional with experience in creating data mapping documents, writing functional specifications and queries, normalizing data from 1NF to 3NF/4NF.
* Good understanding of DataModelling (Dimensional and Relational) concepts like Star-Schema Modeling, Snowflake Schema Modeling, Fact and Dimension Tables.
* Managed ingestion of data from different data sources into HDFS using Sqoop, and Flume and performed transformations using Hive, and Map Reduce and then loaded data into HDFS.
* In-depth knowledge of Hadoop architecture and its components like YARN, HDFS, Name Node, Data Node, Job Tracker, Application Master, Resource Manager, Task Tracker, and Map Reduce programming paradigm.
* Created notebooks in Azure Databricks for data extraction, transformation, and cleansing using spark SQL and PySpark

**Top of Form**

**Bottom of Form**

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Azure services** | Azure Data Factory (ADF v2), Azure Data Lake, Blob Storage, Databricks, ADLS Gen 2, Azure Cosmos DB, Azure Event Hub, Azure Machine Learning, Azure DevOps Pipelines, Polybase, Azure Monitor, Azure Stream Analytics, Azure Synapse Analytics, Azure Event Hubs, Azure Service Bus Queues, Azure Logic Apps, Apache Air flow, Azure SQL Database, Azure Functions Apps. |
| **Aws Services** | Amazon S3, Amazon EC2, AWS Glue, Amazon Athena, Amazon Redshift, AWS EMR, Amazon SNS, Amazon SQS, AWS DMS, Amazon Kinesis, AWS Lambda, Amazon CloudWatch, AWS Glue Data Brew Recipes. |
| **Hadoop Components/Big Data Technologies** | HDFS, MapReduce, Hive, airHBase, Oozie, Sqoop, Zookeeper, Kafka, Yarn, Cloudera Manager, Spark, PySpark, Airflow, Kafka, Cloudera. |
| **Languages** | Scala, Python, SQL, Hive QL, Java, UNIX Shell Scripting, CQL, T-SQL |
| **IDE Tools** | Eclipse, PyCharm, SQL PowerShell |
|  |  |
| **Data Analysis Libraries** | Pandas, NumPy Seaborn, SciPy, Scikit-learn, Matplotlib |
| **Containerization Tools** | Docker, Kubernetes |
| **CI/CD Tools** | Jenkins, GitLab |
| **Version Control Tools** | GitHub, Azure DevOps, |
| **Development Tools** | Eclipse, PyCharm, IntelliJ, SSMS, Microsoft Office Suite |
| **Databases** | Oracle, SQL Server, MySQL, Cassandra, HBase and Azure Cosmos DB |
| **Reporting Tools/ETL Tools** | Power BI, Tableau, SQL Server Integration Services (SSIS), SQL Server Reporting services (SSRS) |
| **Software Methodologies** | Agile, Scrum, Waterfall |

# EXPERIENCE

**Client**: Deutsche bank, New York, NY May 2022 - Present **Role:** Sr. Azure Data Engineer **Responsibilities:**

* Implemented Azure Data Factory (ADF) extensively for ingesting data from different source systems like relational and unstructured data to meet business functional requirements.
* Design and developed Batch processing and real-time processing solutions using ADF, Databricks clusters and stream Analytics.
* Created numerous pipelines in Azure using Azure Data Factory v2 to get the data from disparate source systems by using different Azure Activities like Move &Transform, Copy, filter. Provided support for optimal pipelines, data flows and complex data transformations and manipulations using ADF and PySpark with Databricks.
* Automated jobs using different triggers like Events, Schedules and Tumbling in ADF.
* Created, provisioned different Databricks clusters, notebooks, jobs and autoscaling.
* Performed data flow transformation using the data flow activity. Used Polybase to load tables in Azure synapse.
* Implemented Azure, self-hosted integration runtime in ADF. Implemented batch processing solutions using PySpark and Spark SQL to manage large volumes of data efficiently.
* Improved performance by optimizing computing time to process the streaming data by optimizing the cluster run time.
* Ingested the data from azure synapse analytics to snowflakes data warehouse.
* Scheduled and automated business processes and workflows using Apache Airflow for orchestrating data pipelines.
* Designed and developed a new solution to process the NRT data by using Azure stream analytics, Azure Event Hub, and Service Bus Queue.
* Used Azure Devops & Jenkins pipelines to build and deploy different resources (Code and Infrastructure) in Azure.
* Ensure the developed solutions are formally documented and signed off by business.
* Collaborated with team members to resolve any technical issue, Troubleshooting, Project Risk & Issue identification, and management.
* Worked on the cost estimation, billing, and implementation of services on the cloud.
* Experience managing Azure Data Lakes (ADLS) and Data Lake Analytics and an understanding of how to integrate with other Azure Services.
* Migration of on-premises data (Oracle) to Azure Data Lake Storage (ADLS) using Azure Data Factory (ADF).
* Work closely across teams (Support, Solution Architecture) and peers to establish and follow best practices while solving customer problems.
* Created infrastructure for optimal extraction, transformation, and loading of data from a wide variety of data sources.
* Created pipelines in Azure using ADF to get the data from different source systems and transform the data by using many activities.
* Created Linked service to land the data from various sources to Azure Data Factory. Created diverse types of triggers to automate the pipeline in ADF.
* Created, provisioned different Databricks clusters needed for batch and continuous streaming data processing and installed the required libraries for the clusters.
* Used PySpark and Spark SQL for data validation and quality checks, ensuring data integrity and consistency throughout the pipeline.

**Environment:** Azure Data Factory (ADF v2), Azure SQL Database, Azure functions Apps, Azure Data Lake, BLOB Storage, SQL server, snowflake, AZURE PowerShell, Data bricks, Python, ADLS Gen 2, Azure Cosmos DB, Azure Event Hub, Azure Machine Learning., Azure DevOps pipelines, Jenkins pipelines, Polybase, Azure Monitor, Cost Management, Azure Stream Analytics, Azure Synapse Analytics, Azure Event Hubs, Azure Service Bus Queues, Azure Logic Apps

**Client**: State Farm Insurance, Richardson, TX July 2020 – April 2022 **Role:** Data Engineer **Responsibilities:**

* As an Azure data engineer, I oversaw implementing the ETL process for loading data from various sources into Databricks tables and Azure Synapse Tables.
* Designed & customized data models for the Data warehouse supporting data from multiple sources in real-time.
* Involved in building the ETL architecture & Source to Target mapping to load data into the Data warehouse.
* Design and implement ETL and data movement solutions using Azure Data Factory.
* Worked on On-premises data migration to Azure Data Lake Storage via Azure Data Factory
* Experience with data pre-processing and cleaning to perform feature engineering and data imputation techniques for missing values in a dataset using Python.
* Worked on Data analysis by using NumPy and Pandas libraries in python to filter unstructured data to structured format.
* Created Notebooks that extract raw data from a Data Bricks database, transform it, and then insert it into a cleansed Data Bricks database.
* Created Linked services to connect the external resources to ADF.
* Designed and developed the data warehouse models by using Snowflake schema.
* Created PowerShell scripts to automate ADF pipeline triggers using the Control-M scheduler.
* Developed stored procedure, lookup, execute pipeline, data flow, copy data, and azure function features in ADF.
* Designed and implemented multiple dashboards for internal metrics using Azure Synapse - PowerPivot & Power Query tools.
* Configured Spark Streaming to receive real time data from Kafka. Used Backpressure to control message queuing in the topic.
* Used Data Factory to develop pipelines and performed batch processing using Azure Batch processing.
* Designed and developed pipelines to move the data from Azure blob storage to AWS S3 then moved the data from AWS S3 to data warehouse AWS Redshift which served as backup for the project.
* Migrated data from an on-premises Cassandra database to Cassandra on Azure Cosmos DB
* Developed Spark applications in Databricks using PySpark and Spark SQL to perform transformations and aggregations on source data before loading it into Azure Synapse Analytics for reporting.
* Managed Docker orchestration and Docker containerization using Kubernetes.
* Used Kubernetes to orchestrate the deployment, scaling and management of Docker Containers.
* Utilized PySpark SQL for querying structured data and performing data manipulations before loading into Amazon Redshift or other data warehouses.
* Optimized PySpark jobs for performance by tuning Spark configurations, partitioning strategies, and leveraging in-memory processing capabilities.
* Integrated PySpark with third-party libraries and frameworks such as Pandas and NumPy for advanced data analytics and machine learning tasks.
* Developed and tested PySpark scripts and notebooks in Databricks for interactive data exploration and prototyping of ETL pipelines.
* Implemented data validation and quality checks within PySpark jobs to ensure data integrity and consistency throughout the ETL process.
* Designed and implemented custom PySpark functions (UDFs) for complex data transformations and business logic implementation.
* Optimized workflows by building DAGs in Apache Airflow to schedule the ETL jobs and implemented additional components in Apache Airflow like Pool, Executors, and multi-node functionality.

**Environment:** Azure Data Storage, Azure Services, Azure SQL server, Azure Data Factory Azure data warehouse, MySQL, ETL, PowerBI, SQL Database, U-SQL, Azure Data Lake, Kafka, Azure Databricks, T-SQL, SQL PowerPivot, Power Query, PowerShell, Control-M Scheduler. ETL (Extract, Transform, Load), data warehouse, data modeling, source-to-target mapping, data migration, data pre-processing, data cleaning, feature engineering, data imputation, real-time data processing, Apache Kafka, batch processing, Spark Streaming, backpressure, Azure Blob storage, Azure Function.

**Client**: Verizon, Irving, TX January 2018 – May 2020 Role**:** AWS Data Engineer **Responsibilities:**

* Designed and implemented Lambda architecture for comprehensive data processing, encompassing both batch and real-time streams, ensuring fault tolerance and low-latency access.
* Orchestrated data ingestion using AWS Glue Crawlers and AWS DMS, extracting structured data from MySQL, Salesforce, and JSON files stored in Amazon S3.
* Developed AWS Glue ETL Jobs to transform diverse data formats (e.g., CSV to Parquet, JSON parsing) and loaded processed data into Amazon Redshift for efficient querying.
* Implemented real-time data capture using Amazon Kinesis Data Streams to manage streaming data from multiple sources, enhancing data availability and responsiveness.
* Utilized AWS Lambda functions for real-time data processing, ensuring immediate data transformation and integration into the data warehouse.
* Managed containerized applications with Kubernetes, optimizing resource allocation and ensuring scalability and reliability of data processing workflows.
* Leveraged Amazon Athena for ad-hoc querying and Amazon Quick Sight for interactive dashboard creation, facilitating data-driven decision-making and business insights.
* Established robust monitoring and logging mechanisms using AWS CloudWatch, ensuring initiative-taking management of ETL jobs, Lambda functions, and streaming services.
* Conducted thorough end-to-end testing and performance assessments of data pipelines, ensuring data integrity, reliability, and adherence to project timelines.
* Collaborated closely with cross-functional teams to analyze business requirements, design scalable solutions, and provide technical guidance and support throughout the project lifecycle.
* Implemented CI/CD pipelines using Jenkins and Git for automated testing, deployment, and version control of Python/PySpark codebase, ensuring continuous integration and delivery.
* Demonstrated proficiency in using PyCharm IDE for Python/PySpark development, Git for version control, and Docker for containerization, streamlining development and deployment processes.
* Optimized workflows by designing DAGs in Apache Airflow, scheduling ETL jobs, and enhancing orchestration capabilities with features like pooling and multi-node execution.
* Successfully migrated heterogeneous database tables from on-premises to AWS Cloud using AWS DMS, ensuring seamless data transfer and minimal downtime.
* Utilized Spark SQL capabilities within AWS Glue to normalize data structures across disparate sources, adhering to standardized schemas and formats.
* Provided technical leadership and support in resolving design and implementation challenges, ensuring alignment with project objectives and client requirements.
* Documented architectural designs, workflows, and implementation details, facilitating knowledge sharing and ensuring comprehensive project documentation.
* Received positive feedback from stakeholders for delivering a scalable, resilient, and cost-effective data migration and analytics solution on Amazon Redshift.
* Continuously updated skills and knowledge in AWS services and data engineering best practices, contributing to the team's expertise and innovation in cloud-based solutions.
* Recognized for exceptional problem-solving abilities, initiative-taking communication, and commitment to delivering high-quality solutions that meet business needs and exceed expectations.
* Utilized PySpark within AWS Glue ETL jobs to write custom data transformations and business logic in Python.
* Developed PySpark scripts to manipulate, filter, and aggregate data from diverse sources, ensuring flexibility and customization in data processing workflows.

**Environment:** Amazon S3, Amazon EC2, AWS Glue, Amazon Athena, Amazon Redshift, AWS EMR, Amazon SNS, Amazon SQS, AWS DMS, Amazon Kinesis, AWS Lambda, Amazon CloudWatch, Apache Airflow, AWS Glue crawlers, AWS Glue ETL jobs, AWS Glue Data Brew recipes, AWS Glue Catalog, PySpark, Jenkins, Git Bucket, PyCharm IDE, Parquet, Text files, JSON, XML data lake, ETL (Extract, Transform, Load), data warehousing, business intelligence, streaming data, CI/CD (Continuous Integration/Continuous Delivery), version control, Docker, Kubernetes.

**Client**: Accenture, Hyderabad, India May 2013 – June 2016 **Role:** Junior Big Data Engineer **Responsibilities:**

* Involved in analyzing business requirements and prepared detailed specifications that follow project guidelines required for project development.
* Used Sqoop to import data from Relational Databases like MySQL, Oracle.
* Involved in importing structured and unstructured data into HDFS.
* Fetched real-time data using Kafka and processed using Spark and Scala.
* Worked on Kafka to import real-time weblogs and ingested the data to Spark Streaming.
* Developed business logic using Kafka Direct Stream in Spark Streaming and implemented business transformations.
* Worked on Building and implementing real-time streaming ETL pipeline using Kafka Streams API.
* Worked on Hive to implement Web Interfacing and stored the data in Hive tables.
* Migrated Map Reduce programs into Spark transformations using Spark and Scala.
* Experienced with Spark Context, Spark-SQL, Spark YARN.
* Implemented Sqoop jobs for large data exchanges between RDBMS and Hive clusters.
* Extensively used Zookeeper as a backup server and job scheduled for Spark Jobs.
* Developed Spark scripts using Scala shell commands as per the business requirement.
* Worked on Cloudera distribution and deployed on AWS EC2 Instances.
* Experienced in loading real-time data to a NoSQL database like Cassandra.
* Well versed in using Data Manipulations, Compactions, in Cassandra.
* Experience in retrieving the data present in Cassandra cluster by running queries in CQL (Cassandra Query Language).
* Implemented usage of Amazon EMR for processing Big Data across a Hadoop Cluster of virtual servers on Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3).
* Deployed the project on Amazon EMR with S3 connectivity for setting a backup storage.
* Well versed in using Elastic Load Balancer for Auto scaling in EC2 servers.
* Configured step functions workflows that involve Hadoop actions using Oozie.
* Used Python for pattern matching in build logs to format warnings and errors.
* Integrated GitLab with Continuous Integration/Continuous Deployment (CI/CD) pipelines to automate build, test, and deployment processes.

**Environment:** Hadoop, Map Reduce, Hive, Spark, Oracle, GitHub, Tableau, UNIX, Cloudera, Kafka, Sqoop, Scala, HBase, Amazon EC2, S3. Python, Zookeeper, Oozie, RDBMS, CQL (Cassandra Query Language), HDFS, MY-SQL, ORACLE

**Top of Form**

**Bottom of Form**

**EDUCATION:**

**Masters:** Master of science computer science, University of North Texas, TX, USA | Graduation Year: Dec 2017

**Bachelors:** GITAM University, India |Graduation Year: May 2013