



Guru Charan Gaddam

 Gcharan1810@gmail.com ||  +1 475-319-3980

Professional Overview:

Results-driven and visionary Data Engineer with extensive 11+ years of experience in managing high-performing cross-functional teams and driving transformative data strategies. Skilled in cloud engineering, data pipelines, data migration, strategic planning, and ETL processes, with expertise across AWS, Azure, GCP, Snowflake, and Databricks. Innovative with extensive experience in AI/ML, improvising data pipeline architectures, data integration, data modeling, and implementing AI-driven projects to deliver impactful results.

Committed to leveraging deep expertise to advance AI and cloud engineering, while fostering continuous improvement and delivering scalable, high-impact solutions in Big Data environments. Passionate about leveraging my expertise to contribute to the success of data-driven projects, enhance operational efficiencies, and drive technological innovation across organizations.

Cloud Technologies & AI/ML Solutions:

- Experienced in designing and implementing scalable cloud data architectures and AI/ML solutions across AWS, Azure, and GCP, utilizing technologies Snowflake, BigQuery, and IBM services.
- Skilled in big data processing, data migration, and building reliable, distributed systems with a focus on performance optimization and data modeling.
- Expertise in driving data-driven decision making through machine learning integration, while leveraging platforms SAP BusinessObjects, Teradata, and Oracle Data Visualization for actionable insights and improved system reliability.

Cloud Data Architecture & Data Migration:

- Proficient in designing and implementing robust cloud data architectures using Azure Synapse Analytics, AWS Redshift, Snowflake, and BigQuery to support scalable data lakes and data warehousing solutions.
- Expertise in data migration, data modeling, and dimensional modeling, with a strong focus on data partitioning, data sharding, and data syncing to ensure seamless data flow across platforms.
- Skilled in leveraging IBM Cloud and applying feature engineering to optimize performance and enhance data processing efficiency, driving improved data management and analytics capabilities.

Data Preparation & ETL Process:

- Led the design and automation of ETL (Extract, Transform, Load) processes using tools Azure Data Factory, AWS Glue, Informatica, and IBM DataStage to process and transform data from various sources, expertising in data extraction, data integration, data transformation, and data aggregation to prepare data for analysis.
- Enhanced data quality through data cleansing, data validation, and data orchestration, ensuring data is both actionable and accurate for decision-making.

Real-Time Data Processing & Streaming:

- Delivered real-time data processing and data streaming solutions using Apache Kafka, AWS Kinesis, Azure Event Hubs, and Google Pub/Sub, enabling scalable, high-performance dataflow architectures to process data from IoT devices, transactional systems, and external APIs. Managed data pipelines and implemented batch processing for large-scale data applications.
- Enhanced data pipeline monitoring and applied best practices for data performance optimization and data query optimization.

Big Data & Advanced Analytics:

- Led initiatives in big data environments using Apache Spark, Databricks, and GCP BigQuery to process and analyze vast amounts of structured and unstructured data by applying advanced analytics techniques data clustering, data aggregation, and regression analysis.
- Enhanced data performance optimization and data query optimization strategies to improve processing speed and reduce costs, leveraging SQL, Python, and Java for analytics and data transformation tasks.

Machine Learning & AI Solutions:

- Integrated machine learning models using platforms Azure Machine Learning, AWS SageMaker, IBM Watson Analytics, and Databricks, developing predictive models and deployed AI-driven solutions for data automation and decision-making.
- Worked closely with data scientists to implement artificial intelligence models, including computer vision and natural language processing (NLP) algorithms, driving impactful business results.

Cloud Data Security & Compliance:

- Ensured data security by implementing robust encryption protocols, data backup and recovery mechanisms, and strong access control policies.
- Utilized AWS KMS, Azure Key Vault, and IBM Guardium to ensure data is protected and compliant with industry standards GDPR and HIPAA, implemented data governance frameworks that track data lineage, metadata management, and ensure data fault tolerance.

Data Integration & Automation:

- Implemented data integration using data APIs and streamlined data orchestration processes with Apache Cassandra, SQL, and Teradata to ensure seamless data flow across various platforms.
- Emphasized data containerization for improved scalability and streamlined deployment of data solutions, enabling quicker time-to-market, automated efforts to integrate and orchestrate data across distributed environments, improving overall system efficiency.

Data Visualization & Reporting:

- Developed comprehensive data visualization solutions using tools Power BI, Tableau, Google Analytics, Oracle Data Visualization, and SAP BusinessObjects to build interactive dashboards that allowed stakeholders to visualize key business metrics, track performance, and improve data-driven decision-making and applied data performance optimization to ensure fast and accurate reporting.
- Successfully empowered teams in implementing cloud data engineering best practices and enhanced systems architecture.
- Utilized tools JIRA, PowerBI, and Tableau for project tracking, reporting, and data visualization.

Cloud Data Solutions & Automation:

- Managed cross-cloud data solutions, enabling seamless integration and workflow between multi-cloud environments.
- Implemented data mesh and data federation strategies to improve consistency and reliability in cloud-based data solutions, ensuring scalability, reliability, and performance optimization by applying data load balancing, data partitioning, and data indexing.

DevOps for Data Engineering:

- Developed and maintained CI/CD pipelines for data pipeline automation and deployment, improving data pipeline monitoring and reducing manual intervention.
- Implemented observability practices, ensuring real-time monitoring, logging, and data validation, Focusing on data fault tolerance, ensuring high availability and continuous integration of complex data workflows.

Cross-Functional & Strategic Collaboration:

- Collaborated with technical teams and business leaders to implement data-driven solutions that optimized operational workflows, financial attribution, and supply chain processes.
- Utilized data visualization, analytics, and advanced machine learning algorithms to drive strategic insights, ensuring that all data processes complied with internal data compliance regulations and best practices for data governance.

Tech Stack:

ETL Tools/OS	: Informatica Power center, DataStage, SSIS, SSRS, Ab Initio, Airflow, MuleWindows: 9x/NT/2000/XP/Vista/7/8/10, UNIX, Mulesoft, ELK stack, Splunk monitoring.
Database	: Oracle 19c/12c, MS Access 2016, SQL Server 2019, SSIS, Sybase and DB2, Teradata r15, Hive 2.3, Impala, Cassandra 3.11, Amazon Aurora, Google Cloud Spanner, Azure Cosmos DB, PostgreSQL, MariaDB.
Programming	: Python (NumPy, Pandas, matplotlib), SQL, T-SQL, PLSQL, Shell Scripting, R, C#, C++, Java, HTML5, Power shell, ASP, Visual Basic, XML, Angular.
visualization tools/DataScience	: Tableau, Tableau server, Tableau Reader, SAP Business Objects, SSIS, SSRS, Crystal Reports, Power BI Business Objects, Advanced Excel, Machine Learning, Deep Learning Models, PCAData Science Pipeline, TensorFlow, PyTorch, Scikit-learn, Keras.
Machine Learning/AI	: Scikit-learn, TensorFlow, Keras, PyTorch, XGBoost, LightGBM, CatBoost, Hugging Face Transformers, OpenAI GPT models, AutoML, MLFlow, Google AI Platform, Azure ML Studio, AWS SageMaker, Deep Learning Models, Natural Language Processing (NLP), Computer Vision, Reinforcement Learning.
Data Modeling Tools	: Erwin 9.7, ER/Studio, Star-Schema Modeling, Snowflake Schema Modeling, FACT and dimension tables, ETL Informatica, Pivot Tables.
Clouds / Services	: Data Lakes, Data Factory, SQL Data warehouse, Data Lake Analytics, Databricks, Synapse, Blob storage, Azure Monitor, Azure Application Insights, Virtual Machines, and other Azure services, Salesforce CRM, Dockers, Kubernetes, AWS (S3, EMR, EC2, Glue, Athena) IAM, EMR, Kinesis, VPC, Dynamo DB, Redshift, Amazon RDS, Lambda, DMS, Quick Sight, Amazon Elastic Load Balancing, Auto Scaling, CloudWatch, SNS, SQS, AWS SageMaker, AWS Lambda, AWS Kinesis, AWS CloudTrail, Azure Machine Learning, Azure Cognitive Services, Azure Kubernetes Service (AKS), Azure Synapse Analytics, Azure Functions, GCP, Google BigQuery, Google Cloud AI Platform, Google Cloud Dataproc, Google Pub/Sub, Google Dataflow
Applications	: Toad for Oracle, Oracle SQL Developer, MS Word 2017, MS Excel 2016, MS PowerPoint 2017, Teradata r15.
Big Data	: Hadoop 3.0, Spark 2.3, MongoDB 3.6, MapReduce, Sqoop, Kafka, Snowflake, Apache Flink, Apache Beam, Apache Kafka Streams

Professional Experience:

SENIOR DATA ENGINEER| CLOUD & AI | WALMART, HOBOKEN, NJ

FEB 2025 – Current

The project focused on modernizing retail and enterprise data platforms by migrating legacy on-premise data warehouses to AWS, implementing real-time data streaming, and integrating advanced AI and machine learning models. This transformation aimed to enhance operational efficiency, enable personalized customer experiences, optimize inventory management, and ensure data security and regulatory compliance across the retail ecosystem

- Led the design and deployment of scalable cloud data platforms using Snowflake Data Warehouse alongside Azure Data Lake Storage, Azure SQL Database, and Azure Synapse Analytics, integrating ETL workflows with Azure Databricks (PySpark), and Snowpipe for real-time inventory visibility and accelerating retail analytics across IBM-based systems using ANSI SQL, Snowflake SQL (SnowSQL), and TSQL.
- Migrated large-scale transactional and customer datasets from IBM Netezza and legacy on-premises SQL-based warehouses to Snowflake and Azure, ensuring GDPR and PCI-DSS compliance through secure Java-, Python-, and Snowflake Scripting-driven ETL workflows orchestrated with Azure Data Factory and transformed using PLSQL and Snowflake Tasks.
- Built advanced customer segmentation and recommendation engines using Python (scikit-learn), PySpark MLlib in Azure Databricks, and Snowpark for Java/Python, integrating outputs with Java-based microservices and leveraging Snowflake Variant columns to feed data into Azure Cognitive Search and Azure Dashboards (Power BI).
- Implemented real-time streaming architectures with Azure Event Hubs, Azure Stream Analytics, Snowflake Streams & Tasks, and Kafka connectors, combining Java, PySpark, and C++ modules for high-speed data ingestion and routing ETL output to Azure Cognitive Search and Kibana via Azure Monitor logs.
- Collaborated with marketing, finance, and supply chain teams to build ML models in Azure Machine Learning using TensorFlow, PyTorch, Python, and Snowpark, integrating predictive analytics into Java-based reporting systems and enforcing access control using Snowflake Secure Views and Row Access Policies.
- Enhanced data governance by automating metadata management with Azure Purview and Snowflake's Information Schema, integrating PLSQL-based catalog queries from IBM databases with Azure Cognitive Search indexing through Java APIs and Cerebro for end-to-end lineage tracking.
- Deployed AI-based anomaly detection using Azure Functions and Isolation Forest models built in Python and C++, streaming results to Azure Monitor Logs and visualizing them with Kibana while orchestrating end-to-end pipelines using Azure Data Factory and Snowflake Tasks integrated with IBM logs.
- Created dynamic Power BI dashboards powered by Azure Cognitive Services – Text Analytics API, integrating Python APIs and SQL queries including Snowflake SQL with Azure Monitor log insights to enhance retail manager visibility, and supporting backend logic through Java-based ETL scripts and Snowflake Secure Data Sharing.
- Integrated multi-source customer behavior data into Azure Data Lake Storage and Snowflake external stages via Java-based ingestion pipelines, blending data with PySpark in Azure Databricks and Snowpark, storing unified customer views in IBM SQL, and querying via ANSI SQL, Snowflake SQL, and TSQL.
- Streamlined data migration using reusable dbt models written in PLSQL and Snowflake SQL, automated testing through Azure DevOps Pipelines, and orchestrated workflows via Azure Data Factory and PySpark-based ETL pipelines on Azure Databricks to ensure minimal downtime and integration continuity with Java-based systems.
- Led Agile data engineering sprints utilizing Azure Data Factory and Snowflake Tasks to manage PySpark and Java-based ETL pipelines, delivering real-time insights through SQL, Azure Monitor, Kibana dashboards, and Snowflake Resource Monitors aligned with evolving business objectives.
- Developed robust real-time order processing ETL pipelines using PySpark, Snowpark, and Java on Azure Databricks, integrating SQL and C++ microservices for peak-season transaction processing and monitoring workflows through Azure Monitor, Cerebro, and Snowflake usage dashboards.
- Automated customer segmentation workflows using Python (LightGBM, CatBoost), orchestrated with Azure Data Factory and Snowflake Tasks, trained in Azure Databricks with Java logic wrappers, and ingested data through PLSQL, SQL, and Snowflake Connectors integrated with IBM data lakes.
- Implemented data security best practices using Azure Key Vault, Snowflake RBAC, dynamic data masking, tokenization, and Azure Role-Based Access Control (RBAC) integrated with IBM Guardium, enabling encryption via Java and C++, and streaming access audits to Azure Monitor, Azure Sentinel, and Kibana for monitoring.
- Created predictive maintenance models using Python, Azure Databricks, PySpark, and Snowpark, with C++-based feature generation modules and dashboards integrated into real-time monitoring pipelines using Azure Monitor, Kibana, Azure Sentinel, and Snowflake Billing Dashboards.

Tools and Environments: BigQuery | Cloud Storage | Cloud SQL | Cloud Data Fusion | Dataflow | Pub/Sub | Vertex AI | Dataproc | Cloud Functions | Secret Manager | Cloud IAM | Cloud Natural Language API | Cloud Vision API | Dialogflow | Data Catalog | dbt | Python (scikit-learn, XGBoost, LightGBM, CatBoost, TensorFlow, PyTorch) | Spark MLlib | Facebook Prophet | SHAP & LIME | Agile | Cloud Build | Looker

SENIOR DATA ENGINEER | MACHINE LEARNING | BCBSA, WALLINGFORD, CT

JUNE 2023 – Jan 2025

The project aimed to enhance the scalability, performance, and security of healthcare data systems by migrating on-premise data warehouses, implementing real-time data streaming, and enabling advanced analytics to improve operational efficiency and compliance within the healthcare organization.

- Designed and deployed cloud-based solutions leveraging Snowflake Data Warehouse, Azure Synapse Analytics, Azure Data Lake, Azure SQL Database, Power BI, GCP BigQuery, Cloud Storage, Looker, and Java-based microservices to ensure scalability, availability, and performance for healthcare data systems, utilizing Python, Pyspark, and Snowpark scripts within Databricks for advanced data processing and orchestration with Airflow and Snowflake Tasks.
- Migrated large healthcare datasets from on-premise systems to Snowflake, Azure, and GCP environments using ETL pipelines built with Python, Java, and SQL (including PLSQL, TSQL, and Snowflake SQL), applying Snowpipe for continuous data ingestion while ensuring data integrity and security throughout cloud adoption.

- Built, tested, and maintained complex data pipelines using Azure Data Factory, Azure Synapse Analytics, Azure SQL Database, Snowflake Tasks, GCP Dataflow, and Cloud Composer, leveraging Python, Databricks, Pyspark, and Snowflake Streams to support seamless ETL, migration, and CDC capabilities.
- Designed and implemented data models for large-scale healthcare datasets with optimized transformation and storage strategies using Snowflake Secure Views, Azure Synapse, GCP BigQuery, and Elastic Search for fast retrieval, applying advanced SQL, PLSQL, TSQL, and Snowflake Scripting to support analytics and reporting.
- Enhanced data migration processes with automated workflows using Azure Data Factory, Airflow, GCP Cloud Composer, and Snowflake Tasks, orchestrating ETL with Python, Databricks, Snowpipe, and Pyspark to ensure minimal downtime and improved migration reliability.
- Collaborated with stakeholders to gather business needs and implemented robust validation methods across Snowflake, Azure, and GCP platforms using SQL, Python, and Java, leveraging Elastic Search and Snowflake's Row Access Policies for auditing and quality assurance.
- Developed and implemented transformation and ETL strategies using dbt, Azure Data Factory, Snowflake Scripting, GCP Dataflow, and Airflow with Python, Java, and Pyspark to ensure seamless integration across healthcare systems while enforcing governance policies.
- Designed, tested, and implemented dashboards in Power BI and Looker to deliver clinical insights, using SQL, Python, and Snowflake data sharing capabilities for aggregation and Elastic Search integration to enhance dashboard responsiveness and interactivity.
- Performed data analysis and profiling on raw healthcare datasets using Azure tools, GCP tools (Dataprep, Data Catalog), and Snowflake Variant/OBJECT columns with Python and Pyspark in Databricks to resolve inconsistencies and prepare clean data for analytics.
- Conducted advanced analytics and modeling using Python, Spark, Databricks, Azure Databricks, GCP Vertex AI, and Snowflake Snowpark, building prescriptive and predictive pipelines with Pyspark, SQL, and Java to derive actionable insights at scale.
- Integrated Snowflake with Azure Synapse Analytics, Azure Data Lake, Power BI, GCP BigQuery, and Looker to build end-to-end processing pipelines, automating workflows with Airflow and orchestrating ETL with Python, Databricks, Snowpipe, and Snowflake Tasks.
- Enhanced data reliability and quality with Snowflake's Time Travel and Fail-safe features, while implementing validation techniques using Azure Purview and GCP DLP, and securing pipelines through Azure Key Vault, GCP Secret Manager, and Snowflake RBAC with Java modules and IAM policies.
- Developed and optimized dbt models and testing frameworks in SQL, Python, Java, and Snowflake SQL to ensure data consistency and quality during transformation and migration efforts across Azure, GCP, and Snowflake platforms.
- Conducted data segmentation and modeling using Azure ML Studio, GCP AI Platform, BigQuery ML, and Snowflake Snowpark, building ML models with Python, Spark, Pyspark, and Java to deliver operational, financial, and clinical insights.
- Built and deployed real-time data ingestion pipelines using Azure Data Factory, Azure Stream Analytics, GCP Pub/Sub, Dataflow, and Snowpipe, developing integration layers using Python, Java, and Airflow to ensure accurate and timely healthcare reporting.

Tools and Environments: Azure (Azure Synapse Analytics, Azure Data Lake, Azure SQL Database, Azure Blob Storage, Azure Data Factory, Azure Stream Analytics, Azure Functions, Azure Key Vault, Azure Databricks) | GCP (Cloud Storage, BigQuery, Dataflow, AI Platform, Cloud Monitoring, Cloud Key Management, Cloud Pub/Sub, Dataproc, Cloud Build, Cloud Data Catalog, Secret Manager, Compute Engine) | Snowflake | SQL | dbt | Python | Spark | PySpark | Apache Kafka | Power BI | SSRS | Predictive Analytics and Data Mining Algorithms | HIPAA | Parquet | JSON | HL7 FHIR | Jenkins | PowerShell | Apache Iceberg |

DATA ENGINEER | BANK OF AMERICA, JERSEY CITY, NJ

MARCH 2021 – MAY 2023

The project focused on optimizing and securing the Bank's data infrastructure on AWS, improving real-time analytics capabilities while ensuring compliance with financial regulations. It involved migrating legacy data systems, streamlining data pipelines, improving performance, and reducing operational costs through advanced AWS services and automation.

- Led the migration of on-premise data warehouses to Amazon Redshift, Snowflake, and Databricks, enhancing scalability, performance, and security, achieving a reduction in operational costs by leveraging optimized ETL frameworks using SQL, Snowflake SQL, T-SQL, and PL/SQL, while applying Snowflake-based data modeling and PySpark logic to improve transformation and query performance across Redshift, Snowflake, and Databricks SQL environments.
- Designed and orchestrated end-to-end ETL data pipelines using AWS Glue, Airflow, Snowflake Tasks, and Databricks, ensuring efficient ingestion, transformation, and loading of financial data using PySpark, Snowpipe, SQL, and PL/SQL, while embedding Python logic for custom reporting and enabling high-performance ETL with Snowflake Scripting and SQL-driven transformations.
- Managed and optimized large-scale datasets using Amazon S3, HDFS, Snowflake External Tables, and Databricks Delta Lake, processing petabyte-scale financial data with PySpark, Snowflake Streams, and T-SQL to ensure low-latency, high-throughput data access and CDC support.
- Developed and tuned SQL, PL/SQL, Snowflake SQL, and Python scripts to improve ETL efficiency and scalability in Snowflake, while integrating Elasticsearch and Snowflake Secure Views for enhanced searchable data indexing and fine-grained access control.
- Developed, tested, and implemented custom ETL solutions using Python, PySpark, Scala, and Snowpark within AWS Glue, Databricks, and Snowflake. Integrated ML models using Python libraries and deployed predictive financial analytics using SQL, Snowflake SQL, T-SQL, and PL/SQL across structured and semi-structured datasets stored in Snowflake's Variant and Object columns.
- Enhanced Amazon S3 and Snowflake Lakehouse storage using lifecycle policies and tiered retention strategies via Time Travel and Fail-safe. Automated archival workflows with Python, PySpark, and Snowflake Tasks to meet financial regulations, while integrating Elasticsearch for audit log indexing and compliance retrieval.
- Integrated Apache Airflow, Databricks, and Snowflake orchestration to automate complex ETL pipelines across financial domains. Scheduled and monitored PySpark, PL/SQL, T-SQL, Snowflake Streams & Tasks, and Python transformations with Airflow sensors and Elasticsearch log diagnostics for real-time visibility.

- Implemented comprehensive monitoring using Amazon CloudWatch, AWS X-Ray, Airflow, and Snowflake Query History and Usage dashboards, capturing detailed ETL and SQL performance metrics, identifying Snowflake bottlenecks with diagnostic Python scripts and profiling Databricks clusters with PySpark.
- Collaborated with data science teams to integrate Amazon SageMaker, Snowflake, Databricks, and Airflow into unified pipelines, feeding ML workflows with feature-engineered datasets stored in Snowflake and Redshift, and exposing results through SQL Views and Elasticsearch dashboards.
- Developed and implemented data retention strategies using Amazon S3, Snowflake Zero-Copy Cloning, Delta Tables, and Glacier, building automated retention logic using Airflow, AWS Glue, and Snowflake Tasks, and leveraging PySpark, SQL, and Snowflake SQL for auditing and archiving metadata indexed into Elasticsearch.
- Enhanced data models in Amazon RDS, Redshift, and Snowflake by applying advanced normalization, indexing, and partitioning techniques to optimize SQL, T-SQL, and Snowflake SQL performance, while building ETL jobs using Python, PySpark, PL/SQL, and Snowflake Scripting to support financial KPIs and reporting.
- Managed the integration of external financial data using AWS Lambda, API Gateway, Snowflake Connector for Kafka, and Airflow, automating ingestion from APIs into Snowflake and Databricks. Applied Python and SQL logic in AWS Glue and Snowflake for processing, with Elasticsearch used for indexed search-based analytics.
- Implemented robust SSRS report security using Snowflake Role-Based Access Control (RBAC) and dynamic data masking, enforced by ETL pipelines developed with Python, Airflow, and Snowflake Secure Views. Applied granular security at SQL, T-SQL, and PL/SQL layers, logging access trails in Elasticsearch for compliance.

Tools and Environments: AWS (Redshift, Glue, S3, Lambda, API Gateway, CodePipeline, CodeBuild, CloudWatch, X-Ray, CloudTrail, Elastic Disaster Recovery, RDS, Trusted Advisor, Cost Explorer, Glacier, SageMaker, Elastic Disaster Recovery, Cross-Region Replication for S3 | SQL | Python | Scala | Apache Airflow | Scikit-learn | NumPy | Pandas | Matplotlib | Seaborn | StatsModels | HDFS | SSRS | CI/CD (YAML pipelines) | Machine Learning | SQL Query Optimization |

CLOUD DATA ENGINEER | CENTURYLINK, MIDDLEFIELD, CT

JAN 2020 – OCT 2020

The implementation aimed to enhance customer data analytics, improve data governance, and deliver real-time insights for marketing and customer service within the telecom industry. The project focused on optimizing data management practices to ensure compliance with GDPR and CCPA regulations, while also supporting data migration strategies to modernize legacy systems. This initiative enabled more accurate customer insights, improved decision-making, and streamlined operations, all while maintaining regulatory compliance and enhancing the overall customer experience.

- Contributed to migrating huge customer data from on-premises SQL Server databases to a hybrid cloud environment, leveraging AWS S3 for the data lake and GCP Cloud Storage for machine learning model storage. Integrated ETL pipelines built using Java, PySpark, and SQL (including T-SQL and ANSI SQL) to ensure high availability, data consistency, and secure storage across platforms.
- Developed, tested, and managed ETL pipelines using AWS Glue and Airflow to ingest, transform, and load data from diverse sources into AWS Redshift. Employed GCP Dataflow, Databricks, and PySpark to process real-time data streams, while using SQL and PL/SQL for data transformations and validations.
- Leveraged AWS SageMaker for building predictive models and GCP AI Platform for deployment, utilizing PySpark, Databricks, and ETL pipelines for preprocessing data. Employed SQL, ANSI SQL, and Java to create dynamic customer segments for targeted marketing and used Airflow for automated retraining workflows.
- Tuned performance for large-scale data processing using Apache Spark on Dataproc, while optimizing query execution through partitioning and indexing strategies in Redshift and BigQuery integrating PySpark, SQL, and Databricks for high-efficiency data aggregation jobs and job orchestration with Airflow.
- Designed and implemented batch and streaming data ingestion pipelines using AWS Glue, AWS Kinesis, GCP Pub/Sub, and GCP Dataflow. Streamlined data flows using Airflow, built resilient ETL jobs in PySpark and Databricks, and enriched data using SQL, PL/SQL, and Java.
- Developed a multi-tiered data lake architecture on GCP Cloud Storage, leveraging ETL pipelines written in Java, PySpark, and orchestrated via Airflow. Used SQL, PL/SQL, and T-SQL to manage raw, cleansed, and curated layers and employed C++ for parallel processing where applicable.
- Implemented CI/CD pipelines using Jenkins, and Google Cloud Build to deploy ETL workflows developed in PySpark, Java, and Databricks. Automated versioning and testing for SQL, ANSI SQL, and PL/SQL components while monitoring deployment health via Airflow.
- Integrated AWS Glue and GCP Cloud Data Catalog to classify and catalog data assets, improving discoverability, lineage, and governance. Used ETL tools with metadata tagging in SQL, and analytics monitoring with Elasticsearch, Kibana, and Cerebro dashboards built in Java and PySpark.
- Configured role-based access control (RBAC) and managed identities for AWS and GCP using Airflow, SQL, and Java to enforce fine-grained access. Ensured ETL scripts and PySpark jobs followed security best practices, and tracked access audits using Kibana, Elasticsearch, and Cerebro.
- Implemented a disaster recovery plan using GCP Backup and Restore, and geo-redundant storage. Maintained backup pipelines in Java and Airflow, ensuring recoverability of data processed through ETL, PySpark, and SQL workflows.
- Developed complex data transformation workflows using AWS Glue and GCP Dataproc, leveraging PySpark, Databricks, Java, and C++ to cleanse and enrich large datasets, transforming structured and semi-structured data using SQL, T-SQL, ANSI SQL, and PL/SQL to feed analytics dashboards.
- Used AWS CloudWatch and GCP Cloud Monitoring to set up automated alerts for pipeline failures and integrated logging with Elasticsearch, visualization in Kibana, and cluster inspection using Cerebro. Triggered recovery steps via Airflow and analyzed root causes using Java, PySpark, and SQL diagnostics.

Tools and Environments: AWS (S3, Glue, Redshift, Kinesis, CloudWatch, SageMaker, Secrets Manager, KMS, EMR, CodePipeline, EC2, Backup) | GCP (Cloud Storage, BigQuery, Dataflow, AI Platform, Cloud Monitoring, Cloud Key Management, Cloud Pub/Sub, Dataproc,

DATA ENGINEER | ADP, HYDERABAD, IND

MAY 2016 – JUN 2019

The project aimed to optimize the existing data integration and analytics platform by automating ETL processes to enhance the management of workforce data, including payroll, employee records, and HR analytics. The focus was on improving data quality, performance, and ensuring compliance with regulations GDPR. Additionally, the project supported data migration efforts to streamline the transfer of legacy data, enabling more efficient real-time data processing. This enhanced the overall decision-making and operational efficiency in Human Capital Management (HCM), driving better insights and compliance across the organization.

- Led the migration of on-premises databases to Azure cloud solutions, significantly improving system uptime, scalability, and reducing maintenance costs.
- Built a scalable and efficient data integration and analytics platform that collected, processed, and analyzed customer data from various sources, enabling data migration strategies and enhancing decision-making processes to improve customer engagement.
- Designed and developed scalable ETL pipelines using Azure services Azure Data Factory, Azure Functions, and Python-based scripts to automate data extraction, transformation, and loading from multiple sources, databases, APIs, log files, supporting seamless data migration across systems.
- Integrated data from various internal and external sources into a central data lake on Azure Data Lake Storage, ensuring efficient data storage, retrieval, and optimized partitioning strategies while facilitating smooth migration of legacy data.
- Developed and maintained data models in Azure Synapse Analytics, supporting efficient querying, reporting, and analytics for business stakeholders, ensuring consistency in data migration efforts across platforms.
- Used Azure Data Factory and Apache Spark to perform large-scale data processing and transformation tasks, ensuring high data quality, consistency, and accuracy throughout the data migration process.
- Implemented data validation and cleansing routines, ensuring the consistency of data across different sources and maintaining data accuracy for downstream processes, including data migration to new platforms.
- Automated routine data tasks using Azure Functions and Python scripts, reducing manual intervention and improving operational efficiency, especially for data migration and ongoing data management tasks.
- Enhanced data pipelines for performance and cost efficiency, leveraging services Azure Virtual Machines, Azure Synapse Analytics, and Azure Blob Storage, ensuring smooth and cost-effective data migration.
- Ensured security best practices by encrypting data at rest and in transit using Azure Key Vault, implementing Azure Active Directory (AAD) policies to secure access, and maintaining network security through Azure Virtual Networks for data in transit, including during migration.
- Ensured compliance with GDPR by implementing strict access controls, data retention policies, and audit logging through Azure Monitor and Azure Purview, facilitating secure and compliant data migration processes.
- Set up monitoring and alerting using Azure Monitor to track data pipeline performance, implementing real-time failure recovery mechanisms via Azure Logic Apps, minimizing disruption during data migration.
- Automated scaling of Azure Virtual Machines using Virtual Machine Scale Sets, reducing infrastructure costs, while ensuring scalability and availability for large-scale data migration projects.
- Architected ETL workflows with Azure Data Factory, leveraging dynamic data flows and PySpark for semi-structured data, including handling inputs from Azure SQL Database, Cosmos DB, and on-premises databases, supporting complex data migration scenarios.
- Integrated with external APIs using Azure API Management and Azure Functions, streamlining real-time data collection and ingestion processes, and enhancing data migration from external platforms.
- Automated CI/CD pipelines for ETL workflows, optimizing code deployments and infrastructure updates using Terraform, reducing manual effort and improving the reliability of data migration pipelines.

Tools and Environments: Azure (Data Factory, Functions, Data Lake Storage, Synapse Analytics, Blob Storage, Key Vault, AAD, Virtual Networks, Monitor, Logic Apps, Virtual Machines, SQL Database, Cosmos DB, API Management, Purview, Virtual Machine Scale Sets) | SQL | Python | Apache Spark | PySpark | Terraform | GDPR | ETL | Machine Learning | CI/CD | Terraform | Azure DevOps | Data Migration | Data Integration | Azure Resource Management | GDPR | Data Architecture |

DATABASE ENGINEER | HSBC, HYDERABAD, IND

SEP 2013 – APR 2016

The project focused on designing and implementing scalable, secure, and high-performance database solutions for financial applications, with an emphasis on high availability, regulatory compliance, and seamless integration with cloud platforms. It also involved the migration of critical data to modern cloud environments, ensuring smooth transitions and enabling real-time data processing and analytics for improved decision-making and operational efficiency.

- Designed and implemented relational databases using Oracle, Microsoft SQL Server, and IBM DB2 for high-volume applications, optimizing database schemas for financial reporting and fast data retrieval in Investment Banking, Corporate Finance, Asset Management, and Securities Trading.
- Led the migration of on-premises databases to AWS cloud solutions, significantly improving system uptime, scalability, and reducing maintenance costs.
- Implemented AWS DMS for smooth database migration to the cloud, ensuring data consistency and high availability using AWS services RDS Multi-AZ, Aurora Global Databases, AWS IAM, and AWS Glue, critical for Data Integration.
- Leveraged programming languages Python, SQL, Java, Scala, and R for building and optimizing Trading Algorithms, Quantitative Analysis, and Financial Risk Modeling solutions.
- Integrated advanced database encryption techniques to protect sensitive financial data, ensuring compliance with PCI-DSS, Basel III, and RBI guidelines in line with Banking Regulation and Financial Compliance.

- Enhanced database schemas for banking applications, ensuring scalability, performance, and data integrity in high-transaction environments, supporting Derivatives, Capital Markets, Quantitative Finance, and Equity Research.
- Implemented event-driven architecture using Apache Kafka for real-time data streaming, enhancing fraud detection capabilities and reducing fraud response time in Risk Analytics and Investment Strategies.
- Led ETL processes using SSIS and Informatica, migrating large datasets into Teradata for improved reporting, and facilitating cloud migration to AWS for enhanced scalability and real-time processing.
- Applied big data technologies Hadoop, Spark, and Kafka for fraud detection and customer behavior analysis, improving Risk Management and Quantitative Analysis in Hedge Funds.
- Enhanced performance using AWS EC2, Elastic Load Balancing, and integrated robust security measures in line with industry regulations, ensuring encryption, access control, and threat protection for Trading Algorithms and Market Data.
- Designed high-availability architectures using Oracle RAC and SQL Server Always On to ensure continuous access to financial data during system failures, critical for Stress Testing and Operational Risk management.
- Executed disaster recovery strategies with Veritas NetBackup and RMAN to ensure fast data restoration during system failures, safeguarding Financial Compliance and Capital Adequacy.
- Utilized tools Oracle Enterprise Manager, SQL Diagnostic Manager, and Tableau to monitor and improve database performance, resolving bottlenecks in high-volume environments, crucial for Financial Data Management.
- Developed complex PL/SQL and T-SQL stored procedures, functions, and triggers for enhanced performance and rapid data retrieval in Financial Risk Modeling, Portfolio Management, and Financial Analysis.
- Applied Business Intelligence (BI) tools SAP-BI, Cognos, and OBIEE to create Compliance Reporting and manage Financial Data Warehousing for Capital Markets and Financial Modeling applications.

Tools and Environments: Java, J2EE | Spring | Hibernate | Oracle | SQL Server | MySQL | MongoDB | Kafka | AWS (RDS, Lambda, S3, Redshift, Aurora) | Oracle RAC | Veritas NetBackup | Teradata | Hadoop | Spark | SQL | PL/SQL | T-SQL | AWS DMS | IAM | KMS | AWS Shield.

EDUCATION DETAILS:

Bachelors in Computer Science, KLU university, TamilNadu, 2013
Masters in Computer Science, University of Bridgeport, CT, 2020