**Advait  Mahendra Narvekar**

**Lead Azure Data Engineer**

**mahendrahere1@gmail.com**

**607 713-8155**

# SUMMARY

As a seasoned Senior Data & Cloud Engineer with over 13 years of IT experience in Data Engineering, Analytics, Data Modeling, and Azure Cloud, I have successfully navigated the full lifecycle of implementing robust data solutions, including data marts, data warehouses, and lake houses. My expertise lies in utilizing Azure Cloud Services, such as Azure Synapse Analytics, Azure Data Lake Storage, Azure SQL Database, Dedicated SQL Pool, Spark Pools, Azure DevOps, Azure Databricks, Azure Key Vault, Azure Logic Apps, and Cosmos DB, alongside technologies like Delta Lake, Hadoop, Spark, Hive, Sqoop, PySpark, Python, SQL Server, and NoSQL databases.

I have a proven track record of executing complex Azure cloud migrations and hybrid architectures that deliver comprehensive data platforms. My hands-on experience spans key Azure services—including Azure Data Factory (ADF), Synapse Analytics, Azure Functions, Azure Active Directory, and Azure Monitor—coupled with strong competencies in Terraform, CI/CD pipelines, Python, SQL, and Apache Spark. I also possess a deep understanding of both Windows and Linux systems administration, networking, and cloud security protocols, allowing me to architect secure and efficient solutions.

Recognized for my pioneering work in data warehousing and ETL frameworks, I have led numerous projects from the initial design phases to deployment, significantly enhancing system reliability, data accuracy, and cloud cost-efficiency. I am particularly passionate about mentoring teams, documenting architectural designs, and fostering stakeholder confidence through effective knowledge transfer and agile collaboration, including pre-sales consulting.

**Key Highlight’s:**

* Excellent communication skills, technical skills, managing skills and extensive client facing background in Finance, Healthcare, Banking and data analytics and database management domains.
* Experience in design, development, and implementation of database systems using MS-SQL Server for both OLTP & Data Warehousing Systems applications.
* Experience in building data pipelines using Azure Data Factory, Wintel / Linx, Azure Databricks, and loading data to Azure Data Lake, Azure SQL Database, Azure SQL Datawarehouse (Synapse Analytics), snowflakes.
* Experience with Implementing Databricks Delta Lake Architecture (Bronze, Silver and Gold layers) and the Delta Live Tables (DLT) and creating pipelines for DLT and using Auto Loader to process new data files incrementally and efficiently.
* Experience with SSIS ,Power BI Desktop, Power BI Services, M Language ) Interactions, DAX.
* In-depth understanding of Spark Architecture including Spark Core, Spark SQL, Spark, Data.
* Experienced in developing scripts using Python, PySpark, Scala, Teradata, Shell Scripting to do Extract, Load, and Transform data working knowledge of Azure Data bricks, Mops, snowflakes.
* Strong experience in writing applications using Python using different libraries like Pandas, NumPy, SciPy etc.
* Designed and configured data access policies within Unity Catalog, facilitating fine-grained access controls and maintaining data security for sensitive datasets while promoting a self-service analytics environment.
* Experience in AI, NLP, Deep learning LLM application development
* Demonstrated ability to build and manage end-to-end data pipelines using Apache Airflow and DBT, facilitating seamless data integration from various sources, including Oracle and MongoDB
* Expertise in using advanced statistical methods and machine learning algorithms using SQL, R and Python.
* Experience in developing Spark applications using SQL and PySpark in Azure Databricks for data extraction, transformation, and aggregation from multiple file formats for Analyzing & transforming the data to uncover insights into the customer usage patterns.
* Responsible for estimating the cluster size, monitoring, and troubleshooting of the Spark Databricks cluster.
* Hands on experience on Unified Data Analytics with Databricks, Databricks Workspace User Interface, Managing Databricks Notebooks, Delta Lake with Python, Delta Lake with Spark SQL, Scala, snowflakes.
* Strong experience with Azure services like Storage Account (ADLS Gen2), Key Vault, Data Factory (ADF V2), Logic App, Azure Databricks (ADB), Active Directory (AAD), Storage Explorer and Data Studio.
* Created Pipelines in ADF using Linked Services, Datasets and Dataflows to Extract, Transform and load data from different sources into Azure SQL, Blob storage, Azure SQL Data warehouse.
* Experience in creating various datasets in ADF using linked services to connect to different source and target systems like SQL Server, Oracle, Azure Blob Storage, Azure Data Lake Storage and Azure Synapse Analytics and Azure SQL DB.
* Experience on ETL methodology for supporting Data Extraction, transformations and loading processing using Hadoop.
* Worked on data visualization tools like Tableau and also integrated the data using ETL tool Talend.
* Experienced in Migrating SQL database to Azure Data Lake, Azure data lake Analytics, Azure SQL Database, Databricks, and Azure SQL Data warehouse and controlling and granting database access and Migrating On-premises databases to Data Lake store using Azure Data factory, Wintel / Linx, snowflakes, Teradata.
* Designed Star and Snowflake Data Models for Enterprise Data Warehouse using ER Studio.
* Extensively used regular expressions and core features in Python using lambda, map, reduce etc and effectively implemented logging feature using python logging library and profiling using cProfile
* Working knowledge with IAAS (Infrastructure as a Service) tools such as Terraform and Pulumi.
* Hands - on experience in Azure Cloud Services (PaaS & IaaS), Azure Synapse Analytics, SQL Azure, Data Factory, Azure Analysis services, Application Insights, Azure Monitoring, Key Vault, Azure Data Lake .
* Experience in creating Tables, Stored Procedures, Functions, Views, Indexes and Constraints.

# TECHNICAL SKILLS:

|  |  |
| --- | --- |
| Hadoop Ecosystem | Hadoop, HDFS, MapReduce, YARN, Pig, Hive, Spark, Sqoop, Kafka |
| DevOps/Infra | Git, Jenkins, Maven, SBT, Terraform, Kubernetes, Cloud Formation etc. |
| Programming/Scripting | SQL, Java, Scala, Python, CLI, shell scripting |
| Tools | Git, BitBucket, WinSCP, Putty, mRemoteNG, IntelliJ, VS Code |
| Databases | AWS RDS, Redshift, PostgreSQL, Oracle 10g/11g, MySQL and Teradata,NoSQL Database (Hbase, Cassandra, Mongo DB) |
| ETL & Visualization | Tableau, Spotfire, Cognos, SAP BO/BI and Informatica |
| AWS Services | Glue, Databricks, Redshift, Lambda, CloudFormation, EMR, s3, DynamoDB, Athena. |
| Azure Cloud | Azure Databricks PySpark, Azure SQL Database, ADLS Gen2, Azure Data Factory (ADF), Key Vault, Logic App |
| Others | Postgres SQL, SQL Server, MySQL, NoSQL, Git, Jenkins, CI/CD, Kafka, Rabbit M |
| Hadoop Ecosystem | Hadoop, HDFS, MapReduce, YARN, Pig, Hive, Spark, Sqoop, Kafka |

**CERTIFICATIONS:**

Advance Python Certified   
Databricks Certified   
Azure Data Engineer Certified

**EDUCATIONAL DETAILS:**

K. J. Somaiya College of Engineering | India | Bachelor of Computer Eng. (B.E.) |2012

**PROFESSIONAL EXPERIENCE:**

**Prudential Financial,, New York, USA (Remote) Nov 2023 – Present**

**Lead Azure Data Engineer**

# Responsibilities

# Worked as Data Engineer to collaborate with other Product Engineering team members to develop, test, and support data-related initiatives.

# Designed and developed data warehouse and business intelligence architectures, creating ETL processes to ingest data from various sources into Hadoop/HDFS for analysis and further processing.

# Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL, Scala, and Azure Databricks, snowflakes.

# Worked on Spark using Python as well as PySpark and Spark SQL for faster testing and processing of data.

# Worked with various formats of files like delimited text files, Wintel / Linx, click stream log files, Apache log files, Avro files, JSON files, XML Files. Mastered in using different columnar file formats like Parquet formats.

# Worked on Azure Databricks cloud to organizing the data into notebooks and making it easy to visualize data using dashboards.

# Followed the Medallion Architecture principles for data processing within data bricks from Bronze to Silver to Gold.

# Knowledge on Configuration files in Splunk props. conf, Transforms.confg, Output.confg.

# Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Databricks.

# Involved in building the ETL architecture & Source to Target mapping to load data into the Data warehouse.

# Architect and implement ETL and data movement solutions using Azure Data Factory, SSIS

# Worked with Spark Session Object on Spark SQL and Data-Frames for faster execution of Hive queries.

# Used Broadcast Join in Spark for making smaller datasets to large datasets without shuffling data across nodes.

# Validated and transformed data for end-to-end ETL workflows, loading data from XMLs into enterprise data warehouses.

# Developed data pulls from Cosmos using Scope Scripts.

# Created and leveraged existing KQL queries to pull real-time data for analysis.

# Conducted in-depth technical investigations and root cause analyses using KQL and SQL Profiler, swiftly resolving complex customer issues.

# Designed and developed data flows (streaming sources) using Azure Databricks features.

# Built application platforms in the cloud by leveraging Azure Databricks.

# Responsible for estimating the Cluster size, Monitoring, and troubleshooting of the Spark Data Bricks cluster.

# Created Pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform, and load data from different sources, Scheduled Triggers, Mapping data flows using Azure Data factory and using key Vaults to Store Credentials.

# Designed and configured data access policies within Unity Catalog, facilitating fine-grained access controls and maintaining data security for sensitive datasets while promoting a self-service analytics environment.

# Created Dax Queries to generated computed columns in Power BI.

# Developed analysis reports and visualization using DAX functions like table function, aggregation function.

# Created Data Sets, Linked Services, Control Flows and Azure Logic Apps for sending Emails and Alerts.

# Worked with Azure BLOB and Data Lake storage and loading data into Azure SQL Synapse analytics.

# Created Parameterized datasets and pipelines for reusability and avoid duplicates code.

# Implemented Watermark table concept to only pull incremental data from source systems.

# Mentor and share knowledge with customers as well as provide architecture reviews, discussions, and prototypes.

# Used Microsoft SQL Server Management Studio and SSIS package - Analyzing Technical Design documents, Data Analysis and analyzing complex logics in SQL server, Lift and Shift to Teradata from SQL Server.

# Implemented MLOps guide for the projects from business requirement gathering till Model Operationalization.

# Created SQL scripts to migrate data from SQL server to Cosmos DB.

# Use various Teradata load utilities for data load and UNIX shell scripting for file validation process.

# Worked on Ingesting high volumes of tuning events generated by client Set-top boxes from elastic search in batch mode and from Amazon Kinesis Streaming in real-time via Kafka brokers to Enterprise Data Lake using Python and NiFi.

# Ensure overall system stability, reliability, and manageability of Azure Synapse Analytics deployments.

# Transferred data outside the synapse workspace using the PySpark connector

# Worked on migration of data from On-prem SQL server to Cloud databases(Azure Synapse Analytics (DW) & Azure SQL DB).

# Experience developing Splunk Queries and Dashboards targeted at understanding application performance and capacity analysis.

# Worked on Real-time ML and experienced in handling streaming services for Machine Learning.

# Designed and developed Business intelligence dashboards, Analytical reports and Data visualizations using power BI by creating multiple measures using DAX expressions for user groups like sales, operations, and finance team teams.

# Responsible for the management of Power BI assets, including reports, dashboards, workspaces, and the underlying datasets that are used in the reports.

# Used Power BI and Power Pivot to develop data analysis prototype and used Power view and Power map to visualize reports.

# Tools and Technologies Used: Azure Data factory, Azure Data Lake, Unity catalog, Wintel / Linx, Azure Data Bricks, Azure SQL Database, Azure Synapse Analytics, Active Directory, Application Insights, Azure Monitoring, Azure Search, snowflakes, Data factory, Key Vault, Azure Analysis services, Spark, Power BI, Python Scripting, Data Migration Assistant, Azure Database Migration services

**Kaiser Permanent, Oakland, CA Jun 2020 – Oct 2023**

**SR.Azure Data engineer**

# Responsibilities

# Designed and Developed extensive Azure Data Factory (ADF) pipelines for ingesting data from different source systems like relational/non-relational to meet business functional requirements.

# Implemented Databricks Delta Lake Architecture (Bronze, Silver and Gold layers) and the Delta Live Tables (DLT) and creating pipelines for DLT and using Auto Loader to incrementally and efficiently process new data files

# Built HIPAA-compliant data pipelines using Azure Data Factory (ADF) to ingest and harmonize EHR/ECG data, lab results, and EPIC cardiovascular extracts from disparate sources (Google Cloud, email attachments, ADLS), improving data accessibility for clinical analytics.

# Designed and developed Batch processing and real-time processing solutions using ADF, Wintel / Linx, Databricks.

# Python health checks for OpsGenie. Moved MySQL scripts to health checks in Icinga2.

# Created pipelines, data flows and complex data transformations and manipulations using ADF and Pyspark with Databricks.

# Used Spark-Streaming APIs to perform necessary transformations and actions on the fly for building the common learner data model which gets the data from Kafka in near real time and Persists into Cassandra.

# Created Power BI tabular model, DAX measures, reports, and Apps for operational reporting and KPI dashboards.

# Designed and configured data access policies within Unity Catalog, facilitating fine-grained access controls and maintaining data security for sensitive datasets while promoting a self-service analytics environment.

# Performed BI Developer role for: tabular model including custom measures using DAX.

# Developed Azure Databricks notebooks to apply the business transformations and perform data cleansing operations.

# Secured PHI (Protected Health Information) via Azure Key Vault integration, automating credential rotation for lab data pipelines and ensuring compliance with HIPAA/HITECH regulations.

# Optimized clinical data warehousing in Azure Synapse Analytics, enabling faster reporting on patient outcomes, lab trends, and EHR anomalies.

# Designed scalable ETL pipelines for EPIC cardiovascular data integration, ensuring seamless synchronization with Azure Data Lake Storage (ADLS Gen2) to support predictive modeling for cardiac care teams.

# Expertise in building and managing data warehouses in Azure Synapse.

# Secured PHI (Protected Health Information) via Azure Key Vault integration, automating credential rotation for lab data pipelines and ensuring compliance with HIPAA/HITECH regulations.

# Handled the data enrichment on the raw data using Databricks.

# Developed the pipelines in Azure Data factory for various scenarios to meet business requirement using blob storages and ingesting the data into azure synapse analytics.

# Expertise in writing complex SQL queries, Joins, Stored procedures using, SQL Server, and Oracle.

# Created Master Job Sequences for integration, (ETL Control) logic to capture job success, failure, error and audit, information for reporting

# Created, provisioned different Databricks clusters, notebooks, jobs and autoscaling.

# Extensively worked on Spark Context, Spark-SQL, Scala, RDD's Transformation, Actions and Data Frames.

# Developed custom ETL solutions, batch processing and real-time data ingestion pipeline to move data in and out of Hadoop using Pyspark and Spark SQL.

# Created DataFrames and RDD’s using spark by reading data from Hive Tables, parquet files in ADLS Gen2.

# Developed ETL’s using PySpark. Used both DataFrame API and Spark SQL API, Scala, MLops, snowflakes.

# Engineered automated data ingestion workflows for lab test results stored in Google Cloud, reducing manual processing by 25% and minimizing errors in diagnostic data reporting.

# Leveraged PySpark for distributed processing of large-scale ECG datasets, enhancing anomaly detection accuracy by 30% for cardiology research teams.

# Visualized data using different visualization tools R, Azure ML and Power BI

# Used Spark, Scala, performed various transformations and actions and the result data is saved back to HDFS.

# Collaborated with clinical teams to standardize EPIC cardiovascular data extracts, reducing data reconciliation efforts and improving cross-departmental reporting.

# Used Broadcast variables and joins fact tables, dimensional tables efficiently. Used various optimizations in Spark to run the job effectively.

# Worked on migration of data from On-prem SQL server to Cloud databases(Azure Synapse Analytics (DW) & Azure SQL DB).

# Experience in Tuning SQL queries using utilities like EXPLAIN PLAN, SQL Trace and TKPROF to improve the performance.

# Used Microsoft SQL Server Management Studio and SSIS package - Analyzing Technical Design documents, Data Analysis and analyzing complex logics in SQL server, Lift and Shift to Teradata from SQL Server.

# Use various Teradata load utilities for data load and UNIX shell scripting for file validation process.

# Ingested data in mini-batches and performs RDD transformations on those mini-batches of data by using Spark Streaming to perform streaming analytics in Data bricks, Teradata.

# Used Azure Databricks, created Spark clusters and configured high concurrency clusters to speed up the preparation of high-quality data.

# Designed and developed user defined functions, stored procedures, triggers for Cosmos DB

# Worked with various formats of files like delimited text files, click stream log files, Apache log files, Avro files, JSON files, XML Files. Mastered in using different columnar file formats like Parquet formats.

# Developed Spark applications using PySpark and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns

# Extensive knowledge in migrating applications from internal data storage to Azure. Experience in building streaming applications in Azure Notebooks using Kafka and Spark.

# Worked on designing and developing the Real - Time Tax Computation Engine using Oracle, Stream Sets, Kafka, Spark Structured Streaming.

# Created Dax Queries to generated computed columns in Power BI.

# Developed analysis reports and visualization using DAX functions like table function, aggregation function.

# Created, provisioned different Databricks clusters needed for batch and continuous streaming data processing and installed the required libraries for the clusters.

# Developed Databricks Python notebooks to Join, filter, pre-aggregate, and process the files stored in Azure Data Lake storage.

# Used Cosmos DB for storing catalog data and for event sourcing in order processing pipelines.

# Created modular DBT models for legacy code migration, improving transformation accuracy.

# Created numerous pipelines in Azure using Azure Data Factory v2 to get the data from disparate source systems by using different Azure Activities like Copy Data, filter, for each, Get Metadata etc.

# Automated jobs using different triggers like Events, Schedules and Tumbling in ADF.

# Integrated Azure Active Directory authentication to every Cosmos DB request sent and demoed feature to Stakeholders

# Experience with making API (Application programming Interface) calls from ADF using Web Activity as POST call to get the access tokens and retrieve data into ADLS Gen2 and Azure SQL Server, snowflakes.

# Well-versed with Azure authentication mechanisms such as Service principal, Managed Identity, Key vaults.

# Experience with creating Secrets/Keys/Certificates in Azure Key Vault to store sensitive information.

# Experience in working on Agile/SCRUM methods in a fast-paced manner.

# Assisted AI/ML team to implement data profiling, gap analysis, anomaly detection out of teh data mart.

# Tools and Technologies Used: Azure SQL Server, Azure Data Factory, Azure Data Lakes Gen1/Gen2, Visual Studio Code, Azure Databricks, Apache Spark, Azure Synapse, Teradata, Azure Dev-ops, Wintel / Linx, Python, Power BI, Azure Logic Apps and Azure Cloud Services, Azure functions Apps, unity catalog, Azure Monitoring, Azure Search, Key Vault, Power BI, snowflakes, Python, MLops, Data Migration Assistant

**Erie Insurance Erie, PA Mar 2017 – May 2020**

**Azure Data engineer**

**Responsibilities**

# Integrated RESTful APIs to ingest, transform, and store high-volume datasets in Delta Lake, improving data accessibility for downstream analytics.

# Designed and built scalable ETL pipelines to process and integrate data from APIs, databases, and external sources.

# Automated data ingestion and transformation workflows, reducing manual effort by 40% and enhancing data freshness.

# Contributed to the design and deployment of a hybrid data lake architecture utilizing Azure Blob Storage and on-premises Hadoop clusters to support banking analytics and reporting use cases.

# Participated in enterprise data lake design using Azure Blob Storage and Delta Lake, enabling scalable ingestion and analytics of high-volume financial data.

# Developed Python-based scripts to automate data validation, anomaly detection, and reconciliation processes.

# Built analytical dashboards using Power BI and Tableau to provide real-time business insights.

# Built and deployed secure APIs on Azure Functions and App Services, integrating them with monitoring via Azure Application Insights and custom logging.

# Participated in proof-of-concept (PoC) initiatives to evaluate and test early-stage Azure Data Factory v1, assessing its suitability for replacing legacy ETL workflows in a hybrid cloud environment.

# Optimized SQL queries and Snowflake warehouse configurations, leading to a 20% reduction in query execution time.

# Developed Spark code using Scala and Spark-SQL/Streaming for faster testing and processing of data.

# Analyzed the SQL scripts and designed the solution to implement using Scala.

# Ensured compliance with GDPR and CCPA regulations by implementing secure identity and access controls through Azure AD B2C and automated token validation mechanisms.

# Used Spark-SQL to Load JSON data and create Schema R DD and loaded it into Hive Tables and handled structured data using Spark SQL.

# Collaborated with infrastructure teams to integrate Azure Active Directory for secure identity federation with internal systems during early adoption of Microsoft cloud-based services.

# Implemented Spark Scripts using Scala, Spark SQL to access hive tables into Spark for faster processing of data.

# Developed Tableau dashboards for internal team review, enabling comprehensive data analysis and customization using dynamic filters and actions.

# Designed resilient ETL pipelines with Azure SQL and Logic Apps, supporting near-real-time transformations and reducing latency for fraud detection analytics.

# Set design patterns and standards to load daily deltas and ensured over 2 million records loaded from Amazon S3 to Snowflake within seconds and optimized the initial load from hours to within minutes.

# Developed scalable solutions for real-time data processing using Spark Streaming, Kafka, and NoSQL databases, facilitating agile analytics and insights generation.

# Tools and Technologies Used: Azure, Tableau, Cloud Computing, ETL, Data Visualization, Query Tuning, Online Transaction Processing (OLTP), Kinesis, Talend.

**Nike, Beaverton, OR**   **Oct 2015 – Feb 2017**

**ETL/ Data engineer**

**Responsibilities**

# Working as a MSSQL BI developer to generate the reports as per the business requirements.

# Created/Updated Views, Stored Procedure, Triggers, User Defined Functions (UDF) and Scripts using T-SQL Code.

# Debugging the codes which is developed by other developers in the team.

# Involved in requirement gathering from the business.

# Experience in conversion of Oracle codes to T-SQL and conversion of T-SQL codes to Oracle.

# Experience in analyzing the base requirement and build the initial draft reports for the demo to the business

# Extract Transform Load (ETL) development using SQL Server 2008, SQL 2012 Integration Services (SSIS).

# Experience in moving the data from heterogeneous data sources like oracle, xml etc. to SQL using SSIS.

# Worked on various possible performance-related issues including Database Tuning with the sole intention of achieving faster response times and better throughput

# Developed JSON Scripts for deploying the pipelines in Azure Data Factory (ADF) that process the data using SQL Activity.

# Experienced in developing audit, balance and control framework using SQL DB audit tables to control the ingestion, transformation, and load process in Azure.

# Extensively used SSIS Import/Export Wizard, for performing the ETL operations.

# Create and optimize complex T-SQL queries and stored procedures to encapsulate reporting logic

# Created SQL Server Reports using SSRS 2012.

# Scheduled the Reports to run automatically in tidal.

# Parsed high-level design specifications to simple ETL coding & mapping standards.

# Experience with creating and using CI/CD (Continuous Integration and Continuous Delivery) tools such as Jenkins, Octopus and Azure Devops pipelines.

# Used Spark-Streaming APIs to perform necessary transformations and actions on the data got from Kafka.

# Consumed the data from Kafka using Apache spark.

# Generated parameterized reports, sub reports, tabular reports using SSRS 2008R2.

# Designed, Developed and Deployed reports in MS SQL Server environment using SSRS-2008R2.

# Generated Sub-Reports, Cross-tab, Conditional, Drill down reports, Drill through reports and parameterized reports using SSRS 2012.

# Developed Analysis Services Project. Developed, deployed and monitored SSIS Packages including upgrading DTS to SSIS. Responsible for identifying and defining Data Source and Data Source views.

# Designed different type of Star schemas using ERWIN with various Dimensions like time, services, customers and Fact tables.

# Used Performance Point Server to create interactive reports with an ability to drill down into details.

# Wrote T-SQL procedures to generate DML scripts that modified database objects dynamically based on user inputs.

# Designed and Implemented migration strategies for traditional systems on Azure Lift and Shift, Data Migration Assistant, Azure Database Migration services and other third-party tools.

# Designed, developed, and delivered large-scale data ingestion, data processing, and data transformation projects on Azure.

# Performed Documentation for all kinds of reports and DTS and SSIS packages.

# Designed and implemented stored procedures and triggers for automating tasks.

# Maintained the table performance by following the tuning tips like normalization, creating indexes and collect statistics.

# Tested applications for performance, data integrity and validation issues.

# Tools and Technologies Used: SQL Server 2014/2012, SSIS, SSRS, XML, MS PowerPoint, MS SharePoint, MS Project, MS Access 2007/2003, Agile, Shell Scripting, Oracle, Crystal Reports, SVN Tortoise, Tidal, DART Tool

**Mastek Ltd, Addison, TX Apr 2013 – Oct 2015**

**Big Data Engineer**

**Responsibilities**

# Assisted in building foundational data pipelines to move customer and sales data from on-prem databases to AWS S3 using Sqoop and custom shell scripts.

# Configured AWS S3 directory structures for staging and processed data zones, ensuring consistency across jobs.

# Wrote Hive queries to transform raw data into partitioned tables supporting business intelligence dashboards.

# Ran data validation checks to ensure schema integrity and data accuracy between source databases and processed outputs.

# Implemented EC2-based data processing jobs using Python and Bash for ingestion and transformation tasks.

# Configured IAM roles for secure access to S3 and Hive metastore, adhering to least-privilege principles.

# Built Redshift tables for analytics, loading data from S3 with COPY commands while optimizing file formats and compression for performance.

# Developed error triage processes using CloudWatch logs and basic alert scripts to notify when jobs failed.

# Participated in orchestration workflow development using cron jobs on EC2 for sequencing data pulls and transformations.

# Documented pipeline behavior, wrote UDFs in Python for Hive, and followed handover processes to onboard into workflows seamlessly.

# Tools and Technologies Used: AWS (S3, EC2, EMR, Redshift, IAM, CloudWatch), Hive, Sqoop, Python, Shell Scripting, MySQL, Jenkins, Linux