**TRISHA**

**Email:Sam@uipathtek.com**

**PH:  980-983-2003**

**Sr. Data Engineer**

**PROFESSIONAL SUMMARY:**

* Around **8+** years of experience into Big - data related technologies on various domains like Insurance & Finance.
* Experience as Azure Cloud Data Engineer in Microsoft Azure Cloud technologies including Azure Data Factory (ADF), Azure Data Lake Storage (ADLS), Azure Synapse Analytics (SQL Data warehouse), Azure SQL Database, Azure Analytical services, Polybase, Azure Cosmos NoSQL DB, Azure Key vaults, Azure DevOps, Azure HDInsight Big Data Technologies like Hadoop, Apache Spark and Azure Data bricks.
* Big Data - Hadoop (MapReduce & Hive), Spark (SQL, Streaming), Azure Cosmos DB, SQL Datawarehouse, Azure DMS, Azure Data Factory, AWS Redshift, Athena, Lambda, Step Function and SQL.
* Strong knowledge in Spark ecosystems such as Spark core, Spark SQL, Spark Streaming libraries.
* Very Good experience working in Azure Cloud, Azure DevOps, Azure Data Factory, Azure Data Lake Storage, Azure Synapse Analytics, Azure Analytical services, Azure Cosmos NO SQL DB, Azure HD Insight Big Data Technologies (Hadoop and Apache Spark) and Data bricks.
* Experience in designing Azure Cloud Architecture and Implementation plans for hosting complex application workloads on MS Azure.
* Experience working in reading Continuous json data from different source system using Kafka into Databricks Delta and processing the files using Apache Structured streaming, PySpark and creating the files in parquet format.
* Created manual Test Cases to check that each of the deliverables meet user's requirement.
* Good knowledge in Apache Hadoop ecosystem components Spark, Cassandra, HDFS, Hive, SQOOP, Airflow.
* Experienced in working with different data formats CSV, JSON and Parquet.
* Strong in Data Warehousing concepts, Star schema and Snowflake schema methodologies, understanding Business process/requirements.
* Expert in building hierarchical and Analytical sql queries that helps in reporting.
* Expert in implementing Business Rules by creating re-usable transformations like mapplets and mappings.
* Expert in using debugger in Informatica designer tool to test and fix errors in the mappings. Supported ad-hoc reporting and analytics request with an eye for creating scalable self-service or automated solutions
* Developed and worked on Machine Learning algorithms for predictive modelling
* Architected complete scalable data pipelines, data warehouse for optimized data ingestion.
* Collaborated with data scientists and architects on several projects to create data mart as per requirement.
* Conducted complex data analysis and report on results.
* Constructed data staging layers and fast real-time systems to feed BI applications and machine learning algorithms
* Understanding of AWS, Azure webservices and at least hands on experience working in projects. Knowledge of the software development life cycle, agile methodologies, and test-driven development.
* Develop scalable and reliable data solutions to move data across systems from multiple sources in real time (Kafka) as well as batch modes (Sqoop)
* Extensive experience in ETL methodology for performing Data Profiling, Data Migration, Extraction, Transformation and Loading using Talend and designed data conversions from wide variety of source systems including Netezza, Oracle, DB2, SQL server, Teradata, Hive, Hana and non - relational sources like flat files, XML and Mainframe files.
* Built Enterprise ingestion Spark framework to ingest data from different sources (s3, Salesforce, Excel, SFTP, FTP and JDBC Databases) which is 100% metadata driven and 100% code reuse which lets Junior developers to concentrate on core business logic rather spark/Scala coding.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Big Data Technologies:** | HDFS, Hive, Spark, MapReduce, YARN, Spark-Core, Spark-Sql. |
| **Programming Languages:** | .NET, C/C++, HTML, SQL, PL/SQL, and Scala. |
| **Scripting Languages:** | Shell Scripting, Bash, Powershell, Python. |
| **Operating Systems:** | UNIX, Windows, LINUX. |
| **Web technologies:** | ASP.NET, MVC Framework |
| **Cloud Technologies:** | AWS EC2, ELB, S3, Azure. |
| **Azure Stack:** | Azure Data lake, Data factory, Azure Databricks, Azure SQL database, Azure SQL Data warehouse. |
| **Databases:** | Oracle, SQL-Server, MySQL server, MS SQL, IBM DB2, MongoDB. |
| **Build Tools:** | ANT, Maven, Gradle, Docker and Jenkins |
| **IDE / Tools:** | Eclipse, IntelliJ, Spring Tool Suite (STS) |
| **Testing/Test Management / Defect Management tools:** | Selenium Web Driver/RC/IDE/Grid, HP Quick Test Pro(QTP), Load Runner, JIRA, Quality Center, ALM, Clear Quest, SOAP UI |
| **Version Control:** | Tortoise SVN, CVS and GIT |
| **Platforms:** | Windows, Mac, Linux and Unix. |
| **Methodologies:** | Agile, Waterfall, Test Driven Development |

**PROFESSIONAL EXPERIENCE:**

**PPL Electric utilities and Corporation June 2021 to Present**

**Role: Sr. Data Engineer**

**Responsibilities:**

* Used Azure Data Factory extensively for ingesting data from disparate source systems.
* Used Azure Data Factory as an orchestration tool for integrating data from upstream to downstream systems.
* Extracted data from different sources (text, Parquet, csv files, Azure and On-Premises Databases) to ADL, blob and used PySpark for transformation of data on Azure Databricks.
* Creating Batch Pipelines in Azure Data Factory (ADF) by configuring Linked Services/Integration Runtime to Extract, Transform and load data from different sources into Azure Data Lake.
* Developed spark applications using Pyspark and spark-SQL for data extraction, transformation, and aggregation from multiple file formats for analyzing and transforming the data into customer usage patterns.
* Used Cosmos DB for storing catalog data and for event sourcing in order processing pipelines.
* Designed and developed user defined functions, stored procedures, triggers for Cosmos DB
* Analyzed the data flow from different sources to target to provide the corresponding design Architecture in Azure environment.
* Used Pandas, OpenCV, NumPy, Seaborn, TensorFlow, Matplotlib, Sci-kit-learn, NLTK in Python for developing data pipelines and various machine learning algorithms.
* Programmed in PySpark, Spark SQL and Python to streamline the incoming data and build the data pipelines to get useful insights, and orchestrated pipelines using Azure services.
* Imported data from different data sources into ADLS and created internal, external tables in Databricks which can be accessed by downstream consumers.
* Involved in spark performance optimizations like partitioning, bucketing and perform several types of joins on spark tables and persisted the data for the faster performance.
* Perform Data Cleaning, features scaling, features engineering using pandas and NumPy packages in python.
* Created High level technical design documents and Application design documents as per the requirements and delivered clear, well-communicated and complete design documents.
* Created Build definition and Release definition for Continuous Integration and Continuous Deployment.
* Created Application Interface Document for the downstream to create new interface to transfer and receive the files through Azure Data Share.
* Planning and Developing roadmaps and deliverables to advance the migration of existing solutions on-premise systems/applications to Azure cloud and proposing architectures considering cost/spend in Azure and develop recommendations to right-size data infrastructure.
* Designed and developed azure devops pipelines to manage resource across the multiple subscriptions in Azure.
* 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.
* Created, provisioned different Databricks clusters needed for batch and continuous streaming data processing and installed the required libraries for the clusters.
* Integrated Azure Active Directory authentication to every Cosmos DB request sent and demoed feature to Stakeholders
* Improved performance by optimizing computing time to process the streaming data and saved cost to company by optimizing the cluster run time.
* Developed email notification alerts framework using Azure Logic Apps to send the emails alerts to business teams and also orchestrated azure services using Azure Logic Apps.
* Involved in analyzing the different cluster types like Interactive Cluster, Job Cluster and Pool clusters and decreased the cluster usage by setting up the appropriate cluster type to the scheduled jobs.
* Designed and developed a new solution to process the NRT data by using Azure stream analytics, Azure Event Hub and Service Bus Queue.
* Involved in designing the data warehouses and data lakes on regular (Oracle, SQL Server) high performance (Netezza and Teradata) and big data (Hadoop - MongoDB, Hive, Cassandra and HBase) databases.
* Created Linked service to land the data from SFTP location to Azure Data Lake.
* Created several Databricks Spark jobs with PySpark to perform several tables to table operations.
* Extensively used SQL Server Import and Export Data tool.
* Created database users, logins and permissions to setup.
* Take initiative and ownership to provide business solutions on time.
* Working with complex SQL, Stored Procedures, Triggers, and packages in large databases from various servers.
* Helping team member to resolve any technical issue, Troubleshooting, Project Risk & Issue identification and management.
* Addressing resource issues, Monthly one on one, Weekly meeting.

**Environment**: Hadoop, Azure Cloud, Azure Data Factory (ADF v2), Azure functions Apps, Azure Data Lake, BLOB Storage, SQL server, Teradata Utilities, Windows remote desktop, UNIX Shell Scripting, AZURE PowerShell, Data bricks, Python, pandas, Erwin Data Modelling Tool, Azure Cosmos DB, Azure Stream Analytics, Azure Event Hub, Azure Machine Learning.

**Client: CBRE, Dallas June 2019 to May 2021**

**Role: Data Engineer**

**Responsibilities:**

* Implemented a proof of concept deploying this product in Amazon Web Services AWS.
* Developed solutions for import/export of data from Teradata, Oracle to HDFS, S3 and S3 to Snowflake.
* Worked on setting up AWS DMS and SNS for data transfer and replication and used SQL on the new AWS Databases like RedShift and Relation Data Services.
* Predominantly using Python and AWS (Amazon web services), and MySQL along with NoSQL (mongodb) databases for meeting end requirements and building scalable real time system.
* Used AWS Data Pipeline to schedule an Amazon EMR cluster to clean and process web server logs stored in Amazon S3 bucket.
* Implemented Real-time streaming of AWS CloudWatch Logs to Splunk using Kinesis Firehose
* Designed and developed user defined functions, stored procedures, triggers for Cosmos DB
* Parsed Semi Structured JSON data and converted to Parquet using Data Frames in PySpark and Created Hive DDL on Parquet and Avro data files residing in both HDFS and S3 bucket.
* Involved in designing the data warehouses and data lakes on regular (Oracle, SQL Server) high performance (Netezza and Teradata) and big data (Hadoop - MongoDB, Hive, Cassandra and HBase) databases.
* Analyzed the data flow from different sources to target to provide the corresponding design Architecture in Azure environment.
* Responsible for creating on-demand tables on S3 files using Lambda Functions and AWS Glue using Python and PySpark.
* Expertise in working with multiple databases like Oracle, SQL Server, Netezza, Mongo DB, Salesforce.
* Involved in designing or developing the data pipelines where we get a lot of data from multiple data sources which ingest the data from Oracle RDBMS
* ETL/SSIS data processing experience with various source systems DB2, Salesforce, Dynamics CRM, Dynamics Finance, Informix, XML, Json, Parquet
* Developed AWS CloudWatch Dashboards for monitoring API Performance
* Used custom developed PySpark scripts to pre-process, transform data and map to tables inside the CIF (Non-corporate Information Factory) data warehouse
* Developed JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the Sql Activity.
* Developed shell scripts of Sqoop jobs for loading periodic incremental imports of structured data from various RDMS to S3 and used Kafka to ingest real-time website traffic data to HDFS.
* As part of reverse engineering discussed issues/complex code to be resolved and translated them into Informatica logic and prepared ETL design documents.
* Experienced working with team, lead developers, interfaced with business analysts, coordinated with management and understand the end user experience.
* Used Informatica Designer to create complex mappings using different transformations to move data to a Data Warehouse.
* Developed mappings in Informatica to load the data from various sources into the Data Warehouse using different transformations like Source Qualifier, Expression, Lookup, aggregate, Update Strategy and Joiner.
* Optimized the performance of the mappings by various tests on sources, targets and transformations.
* Scheduling the sessions to extract, transform and load data in to warehouse database on Business requirements using scheduling tool.
* Extracted (Flat files, mainframe files), Transformed and Loaded data into the landing area and then into staging area followed by integration and sematic layer of Data Warehouse (Teradata) using Informatica mappings and complex transformations (Aggregator, Joiner, Lookup, Update Strategy, Source Qualifier, Filter, Router and Expression Optimized the existing ETL pipelines by tuning SQL queries and data partition techniques
* Created independent data marts from existing data warehouse as per the application requirement and updated them on bi-weekly basis.
* Demonstrated expert level technical capabilities in areas of Azure Batch and Interactive solutions, Azure Machine learning solutions and operationalizing end to end Azure Cloud Analytics solutions.
* Decreased the Azure billing by pivoting from using Redshift storage to Hive tables for unpaid services and implemented various techniques like Partitioning and Bucketing over hive tables to improve the query performance
* Used Presto distributed query engine over hive tables for its high performance and low cost
* Automated and validated data pipelines using Apache Airflow

**Environment**: AWS, S3 bucket, Sqoop, Informatica, Amazon EMR/Redshift, Presto, Apache Airflow, Hive, Azure datafactory.

**Client: Credit Suisse, New York, NY November 2017 to May2019**

**Role: Data Engineer**

**Responsibilities:**

* Designed and developed the real-time matching solution for customer data ingestion
* Worked on converting the multiple SQL Server and Oracle stored procedures into Hadoop using Spark SQL, Hive, Scala, and Java.
* Created production Data-lake that can handle transactional processing operations using Hadoop Eco-System.
* Developed PySpark and SparkSQL code to process the data in Apache Spark on Amazon EMR to perform the necessary transformations.
* Involved in validating and cleansing the data using Pig statements and hands-on experience in developing Pig MACROS.
* Analyzed dataset of 14M record count and reduced it to 1.3M by filtering out rows with duplicate customer IDs and removed outliers using boxplots and univariate algorithms.
* Worked with Hadoop Big Data Integration with ETL on performing data extract, loading, and transformation process for ERP data.
* Performed extensive exploratory data analysis using Teradata to improve the quality of the dataset and created Data Visualizations using Tableau.
* Experienced in various Python libraries like Pandas, One dimensional NumPy, and Two dimensional NumPy.
* Experienced in using PyTorch library and implementing natural language processing.
* Developed data visualizations in Tableau to display day to day accuracy of the model with newly incoming Data.
* Worked with R for statistical modeling like Bayesian and hypothesis test with dplyr and BAS packages, and visualized testing results in R to delivery business insight
* Model validation by Confusion Matrix, ROC, AUC, and developed diagnostic tables and graphs that demonstrated how a model can be used to improve the efficiency of the selection process
* Presented and reported business insights by SSRS and Tableau dashboard combined with different diagrams
* Utilized Jira as project management methodology and Git for version control to build the program
* Reported and displayed the analysis result in the web browser with HTML and JavaScript
* Involved constructively with project teams, supported the project's goal through principle and delivered the insights for team and client

**Environment**: Hadoop, Spark SQL, Hive, Scala, Java, MS Access, SQL Server, Pig, PySpark, Tableau, Excel, Pyspark

**Client: Careator Technologies Pvt Ltd Hyderabad, India January 2016 to July 2017**

**Role: Data Analyst**

**Responsibilities:**

* Worked with Data Analyst for requirements gathering business analysis and project coordination.
* Performed migration of Reports (Crystal Reports, and Excel) from one domain to another domain using Import/Export Wizard.
* Wrote complex SQL, PL/SQL, Procedures, Functions, and Packages to validate data and testing process.
* Used advanced Excel formulas and functions like Pivot Tables, Lookup, If with and/index, match for data cleaning.
* Redesigned some of the previous models by adding some new entities and attributes as per the business requirements.
* Reviewed Stored Procedures for reports and wrote test queries against the source system (SQL Server) to match the results with the actual report against the Data mart (Oracle).
* Involved with data profiling for multiple sources and answered complex business questions by providing data to business users.
* Performed SQL validation to verify the data extracts integrity and record counts in the database tables
* Created Schema objects like Indexes, Views, and Sequences, triggers, grants, roles, Snapshots.
* Effectively used data blending feature in Tableau to connect different databases like Oracle, MS SQL Server.
* Transferred data with SAS/Access from the databases MS Access, Oracle into SAS data sets on Windows and UNIX.
* Provided guidance and insight on data visualization and dashboard design best practices in Tableau
* Performed Verification, Validation and Transformations on the Input data (Text files) before loading into target database.
* Executed data extraction programs/data profiling and analyzing data for accuracy and quality.
* Wrote complex SQL queries for validating the data against different kinds of reports generated by Business Objects.
* Documented designs and Transformation Rules engine for use of all the designers across the project.
* Designed and implemented basic SQL queries for testing and report/data validation.
* Used ad hoc queries for querying and analyzing the data.
* Performed Gap Analysis to check the compatibility of the existing system infrastructure with the new business requirements.

**Environment**: SQL, PL/SQL, Oracle9i, SAS, Business Objects, Tableau, Crystal Reports, T-SQL, SAS, UNIX, MS Access 2010

**Client: Yana Software Private Limited Hyderabad, India June 2014 to December 2015**

**Role: QA Analyst**

**Responsibilities:**

* Created Test Plans and Test Scripts by analyzing the Business requirements and System Requirements of the Application.
* Worked closely with a project team for gathering the business requirements and interacted with business analysts to translate business requirements into technical specifications
* Involved in physical and logical design of the applications.
* Worked closely with Application Architects, Business Analysts, Project Managers to discuss and collect business requirements
* Developed detailed Testing Strategy for the entire application and developed various test cases.
* Functional testing the new developments / enhancements.
* Analyzed and tested various reports produced by the application.
* Developed Manual scripts using Quality Center to perform functional and regression testing.
* Conducted Functionality testing during various phases of the application
* Performed UAT Testing
* Checked the data flow through the front end to backend and used SQL queries, to extract the data from the database.
* Created test data for testing specific billing functionalities.
* Work involves testing and reporting bugs using bug-tracking system Mainframe Based System and verification of bug reviews with Development team.
* Used Parameterization to fetch data to test the application using QTP
* Inserted Check Points to Check for the broken Links, Text, and standard properties of an object using QTP
* Working extensively towards testing the performance of the whole application

**Environment**: Quality Center 10.0, Team Tracker, Quick test professional, MSOffice, XML, Agile, SQL, Internet Explorer 9.0, Unix and Windows 2000.