ROHAN KATARIA

SENIOR AZURE DATA ENGINEER

Professional Summary

- Skilled Azure Databricks and Azure Datafactory Developer
- Design and Devlopment using Software Engineering best practices (eg. 12factor app)
- Experienced working in Azure/AWS/Linux/Cloud/DevOps
- 7 years in Data warehouse Development, Support, and Administration
- Proficient in creating Automated ADF Pipelines and leveraging Databricks to create Data Migration Framework
- Experienced in creating Data products for a successful Data Engineering Project
- Data Migration experience from on-premise Datawarehouse to Cloud Datawarehouse.
- Extensive usage of AZURE for connecting to data sources on cloud and on-prem
- Experience in Designing and deploying AZURE solutions using ADF, AZF, ADLS, AZURE BUS SERVICES
- Data Modeling using PowerBI Tabular Data
- Expertise in creating Databricks Spark applications using PySpark Dataframes and SarkSQL
- Experienced in writing Python and Bash Scripts.
- Deep analytics and understanding of Big Data and algorithms using Hadoop, MapReduce, NoSQL and distributed computing tools.
- Used Kubernetes to orchestrate the deployment, scaling and management of Docker Containers.
- Experience in code deployment, Orchestration and Scheduling using tools such as Kubernetes, Docker Swarm
- Debugging production issues, root cause analysis and fixing.
- Expertise in Requirements gathering/analysis, Data Mining, Design, Development, Testing and Production rollover of Reporting and Business Intelligence Analysis projects.
- Experience in gathering and writing detailed business requirements and translating them into technical specifications and design.

Technical Experience

- BI Tools: Tableau (Desktop/Server), Power BI, Qlikview, BIRST, SSRS, Spotfire
- ETL : Talend, Informatica, Databricks, ADF, Oracle Warehouse Builder and SSIS
- Databases: Oracle, MS SQL Server, DB2, No SQL, Teradata, Mongo DB, Cassandra, Synapse SQL Pool, CosmosDB
- Languages: SQL, PL/SQL, R, PySpark, Shell, Python, SnowSQL
- Big Data: Spark, Hive, Kafka, MapReduce, DataBricks

Professional Experience

Senior Azure Data Engineer, Publicis Sapient, Remote **June 2021 – March 2022**
(Client-Marcel.ai)

- Primarily involved in Replatforming, Stabilizing and Maturing of Data Science Databicks Platform
- Architected and replatformed the current ELT/ETL architecture using Azure Data factory, Databricks and Azure Synapse SQL Pool for efficient data storage and cloud warehouse capabilities.
- Configured Azure KeyVault to Encode/Hide exposed keys inside Databricks Notebooks/ADF Linked services and refactored all available resources as a measure to increase security around sensitive data.
- Designed and implemented end-to-end data solutions (integration, processing, testing, storage) in Azure
- Create linked services to connect to REST-API, Azure Storage, Azure Synapse, Sharepoint, Google Analytics & Adobe Analytics.
- Used MLFlow within Databricks to Track, Log and Deploy Machine Learning Models to Production and serve Front End.
- Architected and Implemented a Transformation Feature Store to feed the data to ML Models.
- Enabled DataLakehouse Architecture on top of ADLS to persist ingested parquet/json/csv/avro files to LANDING, STAGING and CONSUMPTION Layers of cloud warehouse (ELT Process)

- Utilized SQL End Points functionality of Databricks to power business intelligence reports leveraging custom/parameterised Views and Tables as source.
- Configured Azure EventHub to Ingest realtime “User Reactions” data from Front-End to ADLS and Transformed through Databricks.
- Created Data Validation framework using Databricks and GreatExpectations Python Library on top of Staging and Consumption Data to compare with Quantitative and Qualitative Expected output.
- Identified common patterns in the ELT/ETL Processes and Created a Common Parameterised Databricks Notebook to accept parameters like JSON Files to be Flattened, Business Transformations, Table Names to be created inside DeltaLake, Schema Changes, Sink of Processed CSV files.
- Configured LogAnalytics to track and monitor diagnostic data of ADF Pipelines and Databricks Jobs, and increased the Overall Stability of Pipelines/Jobs to 80% by Analysing the Root Causes of Failures and Longer Running Pipelines.
- Created Semantics and Data Models in Synapse SQL Pool to store Consumption level data from Databricks DeltaLake
- Increased Query Performancy and Query Processing Time for Spark Dataframes from over 30mins to a few miliseconds by Caching Lazily Evaluating query-operations inside memory.
- Created Proof of Concept to read, edit and drop entities and tables inside Azure Tables using Databricks
- Identified Common Jobs and Reference Files, and persisted them to DeltaLake of Databricks which impacted in faster performance of Databricks Jobs/ADF Runs
- Worked on MERGE/UPDATE/DELETE SparkSQL queries to efficiently work with User Data on the basis of user behavior
- Identified and Assigned DataTypes to Ingested and Transformed data using Databricks Notebooks.
- Developed PowerBI applications to Monitor ADF Pipelines and Databricks Jobs, and analyze the Overall Performance
- Efficiently worked with CI/CD framework inside BitBucket and Jenkins to Deploy ADF Pipelines to Production from DEV, TEST and UAT Environment.
- Administored Databricks Clusters and Instances, and successfully upgraded from Databricks Version 7.3 to 9.1 LTS
- Created custom Email Alerts inside Azure Databricks, Azure Datafactory and PowerBI to notify Pipeline Failures, Over-Average-Processing-Time and Over-Average-Memory-Consumption.

(Client-Chevron)

- Primarily involved in Data Migration using Teradata, SQL Server, Azure Storage, Azure Data Factory, Databricks and SSIS.
- Designed and implemented end-to-end data solutions (storage, integration, processing, visualization) in Azure
- Create linked services to connect to Azure Storage and on-premises SQL Server and Teradata.
- Used Databricks and leveraged Delta Lake to store data in Raw, Refined and Produce layers (ELT Process)
- Created Data Validation framework using Databricks on top of Staging and Produce data to compare with Source
- Created Data Products to support the migration of Data from Teradata to Azure Synapse
- Built majority of Interfaces for a successful data migration from Source SQL Warehouse to Target Synapse Server to migrate data periodically
- Transformed Sportfire Reports to PowerBI reports with stored procedures in the backend pulling data from Synapse
- Translated Semantics and Data Models from Teradata to Synapse
- Upgraded project from on-premise data to the cloud using Azure Delta Lake, Azure Data Factory, Databricks, Azure Data lake Storage.
- Designed and implemented database solutions in Azure Synapse Warehouse.
- Succesful documentation of all the data products created and data migrations performed as a deliverable
- Experienced in managing Azure Data Lake Storage (ADLS), Databricks Delta Lake and an understanding of how to integrate with other Azure Services.
- Experienced in creating a Data Lake Storage Gen2 storage account and a file system.
- Used Copy Activity in Azure Data Factory to copy data among data stores located on-premises and in the cloud.
- Implemented Copy activity, Custom Azure Data Factory Pipeline Activities for On-cloud ELT processing
- Created Pipelines in Azure Data Factory using Linked Services, Datasets, Pipelines to Extract, Transform and load data from different sources like Azure SQL, SSIS, Blob storage and Azure SQL Data warehouse.
- Created queries in SparkSQL and PySpark to reduce processing time by 50%
- Writing PySpark script to Validate the data from Teradata to cloud Environment.
- Experienced in creating Event-based triggers in Azure Data Factory, and in setting up Self-hosted Integration runtime in Azure Data Factory to connect On-Premise SQL Server.
- Created data integration and technical solutions for Azure Data Lake Analytics, Azure Data Lake Storage, Azure Data Factory, SSIS, Azure SQL databases and Azure SQL Data Warehouse for providing synapse analytics and reports for improving marketing strategies.
- Involved in converting SQL queries into Spark transformations using Spark data frames, Scala and Python.
- Worked on Azure Logic apps to schedule orchestration pipelines

Azure Data Engineer, Fannie Mae, Reston- VA **Apr 2018- June 2021**

- Designed and deployed Azure Solutions using Azure Data Factory, Azure Functions, Databricks and Data Lake Storage
- Created scheduled pipelines using Azure Data Factory to move data from Data Lake Storage to Snowflake Warehouse.
- Designed database and warehouse in Snowflake, promoted the code into Test and Prod environments using CI/CD.
- Migration of enterprise data warehouse from on-premise SQL Server to Snowflake.
- Created Snowpipes, Stored Procedures and Tasks for data ETL using Cloud Snowflake.
- Designed Data Models using PowerBI for Data inside Snowflake and UAT on data models.
- Utilized Power BI to replicate SSAS cubes and Analyze data using Excel
- Created ETL DASHBOARD using PowerBI to monitor data migration activity
- Conducted meeting with business stakeholders, clients and director level members for requirement gatherings and presented various solutions using best practices.
- Developed Unit Testing framework before pushing solutions to Stage and Production Environment.
- Experience in building data pipelines using Azure Data Factory, Azure Databricks, and loading data to Azure Data Lake.
- Enabled Snowflake platform for business users and monitored the warehouse usage, database usages.
- Designed, developed, maintained R Shiny Applications based on the requirements of the user.
- Developed Loss Allowance Forecasting using quantitative analysis on state and zip-code level to estimate the total losses on mortgage loans quantitative data.
- Worked on installation of Docker using Docker toolbox. Created custom Docker container images, tagging and pushing the images.
- Dockerized applications by creating Docker images from Docker file.
- Hands on with Git / GitHub for code check-ins/checkouts and branching etc.
- Provided production support and involved with root cause analysis, bug fixing and promptly updating the business users on day-to-day production issues.
- Used Spark Data frames, Spark-SQL, Spark MLLib extensively and developing and designing POC's using Spark SQL, and MLLib libraries.
- Extensive experienced in developing Stored Procedures, Functions, Views and Triggers, Complex SQL queries using SQL Server, TSQL and Oracle PL/SQL.
- Proficient in developing Entity- relationship diagrams, star/snowflake schema designs and expert in modelling the transactional databases and data warehouse.

Sr. Data Engineer, Cuna Mutual, Madison- WI **Oct 2017–Mar 2018**

- Designed, developed and implemented performant ETL pipelines using python API (PySpark) of Apache Spark using Databricks.
- Performance tuning of PySpark scripts.
- Create PySpark frame to bring data from DB2 to Amazon S3.
- Used Toad Data Point to extract data out of the enterprise data warehouse using structured query language or SQL.
- Developed multiple Data Models using Toad Data Modeler for Proof of Concepts.
- Developed POC for Optimization Toolchain in Investment Portfolio and Product Mix to assist clients.
- Developed deployment of Forecasting Models with Quantitative Sales Data on company's products for weekly, monthly, quarterly and yearly sales.
- Performed application server builds in EC2 environment and monitoring them using CloudWatch.
- Led technical implementation of advanced analytics projects and defined the time-series approaches.
- Spun up clusters and used Hadoop ecosystem tools like Spark and databricks
- Develop new and effective analytics algorithms and wrote the key pieces of mission-critical source code.
- Writing PySpark script to Validate the data from Teradata to cloud Environment.
- Conducted meeting with business stakeholders, clients and director level members for requirement gatherings and presented various solutions using best practices.
- Created many calculated columns and measure using DAX in Power BI based on report requirements.
- Installed on premise data gateway and schedule daily data refresh in Power BI.
- Worked on Data Analysis Expressions (DAX) for accessing data directly from tabular SSAS database.
- Import data into power BI using different ETL methods, Direct Query, and Restful API.
- Geneated relational model within Power BI generate marketing Campaigns metrics related to landing pages, Email Analysis and visitors.

Data Engineer, Numero Data, Herndon- VA **Jul 2016 – Sep 2017**

- Strong knowledge in designing and developing Power BI visualization according to business requirement documents and plans for creating interactive dashboards
- Migrated existing Tableau reports to Power BI reports for different teams to analyze log information/supply chain data
- Managed access of reports and data for individual's users utilizing roles by embedding Power BI reports
- Involved in providing maintenance and development of bug fixes for the existing and new Power BI reports
- Designed dashboards utilizing custom filters and DAX expressions with Power BI
- Presented data through dashboards and scorecards with different visuals such as funnel charts, donut charts, scatter plots in Power BI
- Developed the transformation/business logic to load data into data warehouse
- Designed Complex mappings for Slowly Changing Dimensions using Lookup (connected and unconnected), Update strategy and filter transformations for retaining consistent historical data
- Performed source data analysis and captured metadata, reviewed results with business.
- Corrected data anomalies as per business recommendation
- Instigated duplicates removal mechanism using Spark data frame and executed 'Upsert' & 'MERGE' logic using SparkSQL
- Implemented the entire flow of data processing in Spark using Python API and optimized the processing time.
- Worked on automated data pipelines
- Designed and built ETL pipelines to automate ingestion of structures and unstructured data.

Data Engineer, Cleveland Clinic, Cleveland- OH **Jul 2014- Jun 2016**

- Developed or enhanced several strategic models to assist in long term planning of the health
- Developed report automation processes using Tableau to improve turnaround time and reporting capabilities
- Provided advice and support in the usage of healthcare data within the product portfolio, and provided expert subject-matter support to product design and development initiatives
- Designed, coded, and implemented R code for analytic studies to fulfill internal and external client requests, maintained, evaluated, and reported on recurring analytic studies. Performed data hygiene and created data validation programs
- Performing the analyses of health care data, including medical and pharmacy claims, membership files and health advisory/coaching
- Responsible for analyzing health data and producing, verifying and interpreting client reports with very little oversight
- Consolidated data in a cross functional environment from multiple data sources (i.e. membership, claims/encounters, pharmacy, lab, radiology) to create ad hoc reports to support large group proposals
- Worked on editing data connections and changing data sources in Tableau
- Specialized in analyzing and editing metadata (dimensions and measures) and saving data sources.
- Created worksheets by creating join/data blending, custom SQL and data source filters.
- Created Actions in Worksheets and Dashboards for interactivity and to compare data against different views.
- Provided Production support to Tableau users and Wrote Custom SQL to support business requirements
- Created ETL packages using Heterogeneous data sources (DB2, ORACLE and Flat Files etc.) and then loaded the data into destination tables by performing different kinds of transformations using SSIS
- Created different types of reports such as tabular, matrix, drill down and sub reports

Education & Certifications

- **B.S Automation Engineering**, *GGSIPU 2009*