 

**Senior Data Engineer**

Data engineer with over 8+ years of experience specializing in data migration projects with technologies like **Databricks,** **Azure, PySpark**, and **Azure Analytics** for seamless data integration. Proficient in crafting and deploying ETL solutions using PySpark, Azure Data Factory, and Azure Databricks. Strong technical skills in driving data migration and data engineering initiatives. A professional with agile expertise, adept at working seamlessly in multicultural environments. Demonstrated ability to contribute effectively to team dynamics and excel in individual tasks as per project requirements.

|  |  |
| --- | --- |
| * Enterprise Data Management | * ETL Solutions/ELT |
| * Data Analysis & Data Discovery | * Data Validation |
| * Data Security * Data Quality * Data Warehousing | * Cloud Migration Strategy * Data Governance * Data Modelling |

**Tools & Technologies**

|  |  |
| --- | --- |
| **Databases** | MS SQL Server |
| **Programming** | **Python,** **PySpark, Spark**, Hive, Azure DevOps, SQL U-SQL Oozie, NoSQL (Cosmos DB) |
| **Data Integration & ETL** | **Azure Data Factory** |
| **Issue Tracking & Project Management** | JIRA |
| **Visualization Tools** | Power BI |
| **Cloud Data** | **Databricks, Azure Data Lake**, Azure Synapse |

|  |
| --- |
| **Functional knowledge –** Exploratory Data Analysis, Data Insights Generation, Problem Solving, Requirements Gathering  **Job Roles –** Data Engineer, Data Management Analyst, ETL Developer, Data Quality Engineer  **Development Methodologies –** Agile, Waterfall  **Organizations -** BrewDat CSA, JLL, Cognizant Technology Solutions |

**PROFESSIONAL EXPERIENCE**

**Key Projects:**

**Project - Platform Engineering**

Cloud Scale Analytics (CSA) is a next-gen data platform architecture and implementation that enables an effective construction and operationalization of a data platform with Databricks and Azure, at scale. The design is modular and natively integrates with the Enterprise scale design pattern

**Role: Senior Data Engineer (Aug 2021-Present)**

* Optimized Databricks workloads using deletion vectors, partition pruning, spot VMs, and cluster optimization, achieving overall cost savings of 50%.
* Implemented the migration from Hive Metastore to Unity Catalog, enhancing data management and governance.
* Created an ACL framework in Unity Catalog to enable granular access control at the table, row, and column levels.
* Designed and implemented metadata-driven pipelines for extraction and orchestration in data factory which is leveraged by users to build data products.
* Developed PySpark notebooks as standardized templates for project users, facilitating the streamlined loading of data into bronze, silver, and gold tiers through autoloader and delta streaming processes.
* Implemented comprehensive access management for users in Databricks workspaces using PowerShell, configuring essential objects such as clusters, SQL warehouses, global init scripts, and more to optimize the overall onboarding of projects and users into workspace.

• Developed and implemented data quality metrics, focusing on completeness, uniqueness, accuracy, integrity, validity, and timeliness to enhance overall data quality within the organization.

• Developed a versatile library containing generic methods tailored for various load types, including APPEND, UPSERT, and SCD2, enhancing efficiency and flexibility in data loading processes.

• Constructed PySpark transformation reusable methods to streamline data processing tasks, including flattening structures, deduplication, and standardizing column names etc. for building data products faster.

• Implemented narrow quality data checks, including type and null checks, enabling users to identify and quarantine or reject flawed data before its integration into the silver tier.

• Established a comprehensive framework empowering user to implement data masking seamlessly. The framework supports various masking techniques such as TEXT\_SUBSTITUTION, NUMERIC\_SUBSTITUTION, KEY\_MASKING, INTEGER\_RANDOMIZATION, and DECIMAL\_RANDOMIZATION, ensuring flexibility and security in handling sensitive information.

**Project – Enterprise Data Repository for Global Property Solution**

Establish an enterprise data repository with enterprise services in UDP (unified data platform), real time Property User Interfaces (UI) and CRM integration that serves complete information related to Properties, Availabilities, Listings, Leases, Sales transactions and Capital stack. Address data governance through integration with enterprise MDM & common Data Quality framework.

**Role: Senior Data Engineer (Oct 2020-Aug2021)**

* Building ingestion frameworks for various sources using python.
* Developed generic framework for building snapshot of all the data ingested in lake using Pyspark.
* Built required reusable spark code for ETL transformations.
* Develop data pipelines using Pyspark based on the mappings.
* Develop data quality and enrichment frameworks to be used in jobs to clean the data.
* Write unit tests in python to validate different modules developed.
* Provide fixes for bugs raised and optimize the performance of the spark jobs.

**Enterprise Data Lake Analytics**

Develop unified layer of data from different sources including oracle, salesforce, sql server etc to enable users and analysts provide insights in the organization business and build required reports**.**

**Role: Cloud Data Engineer (Nov 2019-Sep 2020)**

* Developed ingestion pipelines using Azure Data Factory from various sources into azure data lake.
* Developed reusable classes for ETL transformations, scd2 and scd1
* Flatten and load complex data types including array and struct.
* Develop Databricks jobs using Pyspark to transform the raw data into curated data as per TDD.
* Created Databricks jobs to load data from lake to Azure SQL database using bulk load.
* Developed stored procedures to load data from staging to base table and necessary indexes to improve the performance of query.
* Written automated unit test cases to ensure good quality codes and zero defects

**Project: Foresight and Intelligence**

The purpose of the project was to migrate legacy data warehouse on oracle Azure. The objective of the program is to provide scalable, accurate and global reporting incorporating multiple dimensions and facts in cohesive data warehouse for reporting environment.

**Role: Associate (Nov 2018 – Oct 2019)**

* Developed ADF pipelines using numerous activities and logged the pipeline run status and other information in Azure SQL DB.
* Created integration framework for moving source systems to Azure Data lake store.
* Worked on PySpark in data bricks for moving the files to different layers in ADLS by transforming the data.
* Developed pipeline for moving the incremental load to Azure SQL DW through polybase.
* Developed Azure function app for moving data from cloud to target system.
* Provide root cause analysis and fixes on production bugs.
* Implemented validation mechanism to maintain the integrity of data within source systems and cloud.
* Develop hql scripts for loading data into hadoop for different load strategies such as insert, overwrite and incremental load.
* Worked on continuous integration and delivery (CI/CD) of Azure data factory pipelines in Azure DevOps

**Project: Data as a Service**

The purpose of this project is to build a new Data as a service platform on the Microsoft Azure could

environment. The data is loaded from a wide variety of sources into a data lake environment. Data marts and Curated Data Models (CDMs) were created for self-service reporting, Customer Insights and analysis capabilities. The Curated Data Models is built based on specific requirements from business needs and projects.

**Role: Programmer Analyst (Jan 2018 – Oct 2018)**

* Created Azure Data Factory pipelines, Connections, Datasets and triggers extensively to meet the functional requirement.
* Written U-SQL scripts to do the necessary transformations and loading the data in Azure data lake Store and USQL catalog through Azure data lake analytics.
* Created logic apps for the purpose of job failure/success notifications.
* Written stored procedures for incremental loads and logging of jobs execution information.
* Written hive scripts to load/extract the data and create dynamic partitions in hdinsight Hadoop.
* Written U-SQL table valued functions in U-SQL catalog to parse the xml data into CSV.
* Developed U-SQL scripts with various programming constructs including Extractors, Dynamic Partition Outputter, Exceptions, File Sets, Virtual Columns and File Wild Cards etc.

**Project: Data Lake for Supply Chain Management**

The purpose of this project was to implement the data lake categorized into three storage areas-Raw Data Source, Staging Data Model and Reporting Data Model. The data in RDS is in raw and original format. For business users to create reports and apply analytics, the data in RDM is cleaned and transformed leveraging HDFS, hive, pig, Sqoop, HBase and Oozie. The data lake can store the enterprise-wide information in organized, easy to identify and have the ability to share with different user groups. The data stored can be easily transformed into different layouts to process with big data frameworks, tools available.

**Role: Programmer Analyst Trainee (Dec 2016 – Dec 2017)**

* Worked on importing and exporting data from different databases like Oracle, Mysql into HDFS and Hive using Sqoop.
* Involved in creating Hive tables, loading with data and writing hive queries.
* Execution of Unit test cases and reporting bugs.
* Implemented the partitioning and bucketing to improve the performance of the hive queries.
* Developed Hive UDFs to extract the data in desired format from staging tables.
* Designed Oozie Workflow utilizing hive, pig, shell etc actions and scheduled through coordinator job with data dependency.
* Involved in fixing the failures of oozie job and reprocess them.
* Involved in the Support and implementation of the identified change request.
* Involved in optimizing the performance of Hive by adjusting numeric configuration parameters.

**CERTIFICATIONS**

* Databricks Certified Data Engineer Associate

**EDUCATION**

|  |  |  |
| --- | --- | --- |
| Year | Degree | Institution |
|  |  |  |