ERP Data Pipeline Automation using PySpark for a Manufacturing Company

Client Overview

- The client is a leading steel and aluminum manufacturer, serving a huge group of customers across diverse geographic locations, while manufacturing more than 10,000 products with a widely spread network.

Project Overview

- ETL process was established using PySpark to migrate the Sales data stored on MySQL on-perm database, which was in-turn fetched from multiple ERP systems, to Redshift on AWS cloud. Data processing and analytics was performed using Amazon EMR. The data underwent several data quality checks and validation by using Google API and Python packages.

Business Requirement

- To migrate the data from on-perm database to cloud with minimal migration time.
- Make the migrated data analytics ready for downstream advanced analytics.
- Process the data migration in a secure way, with several data quality checks.

Our Solutions

- Indium successfully migrated the data from on-perm MySQL database to Redshift in AWS cloud using PySpark.
- Amazon EMR enabled tuning and monitoring of the cluster constantly with EC2 instance.
- The data for migration to be extracted from S3 bucket, which contains daily incremental data fetched from multiple ERP systems.
- Perform several data quality checks and validations. Pre-process of the data to handle null values, along with data type checks. Validate addresses using Google API and other python packages, such as pyusps, pycountry, uszipcode, geopy etc.

Business Impact

- Technology choices lead to an 80% reduction in data migration time from on-perm database to cloud.
- Feasibility and ease of data migration was increased by 80% by leveraging PySpark.
- Security protocols were established on Amazon EMR by configuring EC2 firewall setting to control the access to the instances.
- The usage of cloud database minimized the maintenance and management overheads.

Tools used:

- PySpark, Python, Python Packages (Pyusps, Pycountry, GeoPy), Redshift, Amazon EMR, S3 Bucket, MySQL

INDIA
Chennai | Bengaluru | Mumbai
Toll-free: 1800-123-1191

USA
Cupertino | Princeton
Toll-free: +1-888-207-5969

SINGAPORE
+65 9630 7959

UK
London

MALAYSIA
Kuala Lumpur
+60 (3) 2298 8465