SUCCESS STORY

ERP Data Pipeline Automation Using Pyspark For A Manufacturing Company

BUSINESS

Data Engineering

DOMAIN

Manufacturing

TOOLS

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

KEY HIGHLIGHTS

- 80% reduction in data migration time from on-perm database to cloud
- PySpark increased the feasibility and ease of data migration by 80%
- The usage of cloud database minimized the maintenance and management overheads

ERP Data Pipeline Automation Using Pyspark For A Manufacturing Company

CLIENT

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. In dealing with a vast number of customers, there was significantly difficulty in maintaining and managing all the data, stored on their on-perm database. The bulk of the data was fetched from multiple ERP systems and connected with the on-perm MySQL database.

PROJECT OVERVIEW

For a large Manufacturing company, an 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 REQUIREMENTS

The entire process is carried out with the focus to fulfill the following 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.

SOLUTIONS

To perform a secured data migration and minimise the time consumption, Indium proposed the following solution:

- The data migration from the 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.

About Indium

Indium is a Digital Engineering Services leader and Full Spectrum Integrator that helps customers embrace and navigate the Cloud-native world with Certainty. With deep expertise across Applications, Data & Analytics, AI, DevOps, Security and Digital Assurance we “Make technology work” and accelerate business value, while adding scale and velocity to customer’s digital journey on AWS.

indium

Make Technology Work

USA	INDIA	UK	SINGAPORE
Cupertino | Princeton	Chennai | Bengaluru | Mumbai	London	Singapore
Toll-free: +1-888-207-5969	Toll-free: 1800-123-1191	Ph: +44 1420 300014	Ph: +65 6812 7888

www.indiumsoftware.com

For Sales Inquiries
sales@indiumsoftware.com

For General Inquiries
info@indiumsoftware.com