FELDM

Data Integration and Architecture

Development of a global online and offline data archive in the Azure Cloud

An international retailer with high brand awareness and huge online and offline customer bases asked for support from FELD M in designing and developing an IT architecture for storing and handling all online raw data across more than 40 markets. The solution should process, enrich, and deliver more than 500 GB of data daily in a high-performance database for further analyses.

FELD M was tasked with overcoming a number of problems: Global reports were time-consuming and needed a lot of manual effort. It was impossible to drill down into the details of the available online data. Data scientists and analysts were able to access the data via the Adobe interface only, which limited their capabilities, especially for long-term analyses. The traditional in-house data warehouse (DWH) was unable to deal with massive volume of raw data. Due to the existence of several parallel architectures, there was no single source of truth.

The data integration and architecture project focused on the following key points:

- Processing of more than 500 GB of data daily
- ETL pipeline in Python and Spark without lock-in effects on the platform
- Integration of Adobe Analytics raw data from more than 40 markets

Over a period of 18 months, FELD M accompanied the project. The tasks ranged from the conception and implementation of the appropriate IT architecture for data processing and enrichment to the translation of the Web Analytics raw data and the dashboard implementation. The ultimate goal was to realize a cost-effective solution to process the massive volumes of data via the ETL pipeline and visualize this data in automated performance dashboards, including an alerting and monitoring system.

A scalable ETL pipeline to unlock the potential of Adobe Clickstream raw data:

To avoid a complete lock-in effect on the cloud platform, we set up an ETL pipeline in Python and Spark. The tasks of the ETL pipeline were to archive and process the historical data of past years, to link them with additional data sources (e.g. product data) and to store the required aggregations in a database (SQL Server). The PowerBI dashboards for cross-market analyses were based on these data.

The resulting cost-effective, high-performance solution can store, process, and aggregate more than 500 GB of data daily. The client now has a uniform data basis including various levels of pre-aggregations for analyses by the Data Science and Analytics departments.

Contact us: FELD M GmbH Sandstraße 33 80335 Munich
Phone +49 (0)89 / 5529 756.0 hello@feld-m.de www.feld-m.de