FELDM

Datenintegration und -architektur

Aufbau eines globalen Online- und Offline-Datenarchivs in der Azure Cloud

Ein internationaler Einzelhändler mit hoher Markenbekanntheit und riesigem Online- und Offline-Kundenstamm beauftragte FELD M mit der Konzeption und Entwicklung einer IT-Architektur für die Speicherung und Handhabung aller Online-Rohdaten aus über 40 Märkten. Die Lösung sollte täglich mehr als 500 GB Daten verarbeiten, anreichern und zur weiteren Analyse an eine hochperformante Datenbank übermitteln können.

Wir hatten eine Reihe von Herausforderungen zu meistern: So war das Erstellen globaler Reports bisher zeitaufwendig und erforderte hohen manuellen Aufwand. Es war nicht möglich, die verfügbaren Online-Daten im Detail zu analysieren. Data Scientists und Analysten konnten auf die Daten nur über das Adobe-Interface zugreifen, was ihre Analysemöglichkeiten stark begrenzte – speziell im Hinblick auf Langzeitstudien. Das vorhandene firmeneigene Data Warehouse (DWH) war für die riesige Menge an Rohdaten ungeeignet. Zudem gab es aufgrund verschiedener Parallel-Architekturen keine zentrale und verbindliche Datenquelle (Single Source of Truth).

Die Kernanforderungen an das Projekt:

- Ausgelegt auf die Verarbeitung von täglich mehr als 500 GB an Daten
- ETL-Strecke in Python und Spark ohne Lock-in-Effekte auf der Plattform
- Integration von Adobe Analytics-Rohdaten aus mehr als 40 Märkten

Das gemeinsame Projekt lief über einen Zeitraum von 18 Monaten. Unsere Aufgaben reichten von der Konzeption der geeigneten IT-Architektur für die Datenverarbeitung und-aufbereitung über die Übertragung der Web Analytics-Rohdaten bis hin zur Dashboard-Implementierung. Hauptziele waren die Umsetzung einer kosteneffizienten Lösung zur Verarbeitung des enormen Datenvolumens und die Erstellung automatisierter Performance-Dashboards sowie ein Warn- und Monitoring-System zur Sicherung der Datenqualität.

Eine skalierbare ETL-Strecke, um das Potenzial der Adobe Clickstream-Rohdaten zu heben

Um einen kompletten Lock-in-Effekt auf der Cloud-Plattform zu vermeiden, wurde die ETL-Strecke in Python und Spark implementiert. Die Aufgaben der ETL-Strecke bestanden darin, die historischen Daten aus den vergangenen Jahren zu archivieren und zu verarbeiten, sie mit weiteren Datenquellen (z.B. Produktdaten) zu verknüpfen und die benötigten Aggregationen in einer Datenbank (SQL Server) zu speichern. Diese Daten bilden die Grundlage der PowerBI-Dashboards für marktübergreifende Analysen (Performance-Dashboards).
Im Ergebnis ist eine kostengünstige und hochperformante Lösung entstanden, die täglich mehr als 500 GB Daten speichern, verarbeiten und aggregieren kann. Der Kunde verfügt nun über eine einheitliche Datenbasis inklusive verschiedenster Aggregationslevel, die Auswertungen durch die Data Science- und Analytics-Abteilungen ermöglichen.

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