Case Study

Aufbau eines Modern Data Warehouse als Ausgangspunkt für künftige Data-Science-Initiativen

"Wir verfolgen echtes datengetriebenes Wirtschaften in Form der präskriptiven Datenanalyse. Gleichzeitig erweitern wir mit Advanced Analytics nicht nur die emotionale Kundenbindung, sondern steigern zudem die positive Customer Experience."

Sara Burdenski
Teamleitung Kundenbindung & Analytisches Marketing
evm

Aufgabe

Aufbau eines neuen Modern Data Warehouse in Azure für kundenzentriertes Data Management als Basis für künftige Data-Science-Initiativen.

Über evm

Die Energieversorgung Mittelrhein (evm) ist das größte kommunale Energie- und Dienstleistungsunternehmen aus Rheinland-Pfalz. Das Einzugsgebiet reicht im nördlichen Rheinland-Pfalz vom Westerwald über den Hunsrück und die Eifel bis hin zur Landesgrenze Nordrhein-Westfalens. Rund 1.000 Mitarbeiter versorgen hier Kunden mit Ökostrom, Erdgas, Wärme, Trinkwasser, Telekommunikation und kompetentem Service. Die evm gehört zu den wichtigsten Arbeitgebern der Region und ist sich ihrer Verantwortung bewusst. Deshalb setzt sie sich aktiv und aus Überzeugung für transparentes, umweltschonendes und ressourcenorientiertes Handeln sowie soziales Engagement ein.

Projekt im Überblick

Kunde

Energieversorgung Mittelrhein AG

Branche

Energiedienstleister

Lösung

> Modern Data Warehouse

Leistungen

> Data Architecture

> Cloud Data Warehouse

> Data Pipelines

> Data Science Consulting

Technologien

> Microsoft Azure Data Factory

> Spark

> Databricks

> Microsoft Power BI

Power BI

databricks

Ausgangslage

Die bislang eingesetzte Data-Warehouse-Lösung der Energieversorgung Mittelrhein sorgt für Unzufriedenheit. Gängige deskriptive Fragestellungen können über Datenauswertungen zwar beantwortet und statische Scores entwickelt werden, doch die Qualität der Daten ist nicht zufriedenstellend. Eine vollumfängliche und progressive Datenanalyse ist mit der vorhandenen Datenbasis nicht möglich. Zudem ist der Wartungsaufwand des vorhandenen Technologie-Stack zu hoch. Der Marketingbereich Kundenbindung plant daher parallel zum bestehenden veralteten Data Warehouse den Neubau eines Modern Data Warehouse. Mit der bereinigten Datenbasis sollen geplante Data-Science-Initiativen deutlich vorangetrieben werden.

Ziel all dieser Bemühungen ist die Aktivierung einer validen 360° Kundensicht. Unter anderem werden dafür ein Prognosemodell für den Grundversorgerstatus angestrebt, eine Kündigeranalyse, die Ermittlung relevanter dynamischer Scores wie Vertragswert und Kundenwert sowie weitere zielgruppenspezifische Analysen. Die evm ist bereits tief in die Datenanalyse eingestiegen, stößt jedoch technologisch begründet an ihre Grenzen. Es fehlt vor allem die Kundensicht in den Daten, insbesondere bei Massendaten, die aus dem Abrechnungssystem stammen.

Zudem weisen unterschiedliche Datenquellen aufgrund fehlender ETL-Strecken keine Verbindung untereinander oder zu einem zentralen Speicherort auf. Die Anbindung von Quellsystem und Datenfeldern leidet unter fehlender Flexibilität und Rohdaten werden teilweise nicht korrekt verarbeitet. Für das Anforderungs- und Fehlermanagement sollen strukturierte Prozesse eingeführt werden. Außerdem wird IT-seitige Unterstützung bei der ETL-Programmierung sowie der Implementierung automatisierter Prüfmechanismen benötigt.

Ein Kunde, zu viele Datensätze

Aufgrund unterschiedlicher Datenquellen werden die vorhandenen Datensätze bislang nur aus Vertragssicht behandelt und sind durch unterschiedliche Transformationswege zudem fehleranfällig. So kann es vorkommen, dass einem Haushalt mehrere Attribute über unterschiedliche Datensätze zugesprochen werden.
Beispielsweise schließt ein Ehepaar einen Vertrag über Gas ab. Offiziell ist die Ehefrau der Vertragspartner. Nimmt ihr Mann später eine vertragliche Änderung vor,

indem er beispielsweise einen zusätzlichen Stromvertrag abschließt, existieren bereits zwei unterschiedliche Datensätze für einen Haushalt. Diese werden jedoch nicht automatisch miteinander in Verbindung gebracht, denn die Datensätze werden im alten System singulär betrachtet. Es kann nun sein, dass der Haushalt bei einer Datenabfrage zwar als Gaskunde, nicht aber als Stromkunde geführt wird.

Die Herausforderung besteht im Datenabgleich und der kontextbezogenen Bereinigung und Zusammenführung der unterschiedlichen Datensätze.

Ziele

Es soll eine einheitliche Kundendatenbank als zentrale Informationsquelle aufgebaut werden, um die Customer Journey nachzuvollziehen und Kunden wie Kundinnen zielgerichtet ansprechen zu können. Wichtig ist in diesem Zusammenhang die kundenzentrierte Speicherung der Daten, wie beispielsweise Bestandsdaten, Vertragsdaten, Verhaltensdaten oder mikrographische Daten, um sowohl kundenspezifische Automatisierungen als auch eine konsistente Kundenansprache gewährleisten zu können. Weiterhin sollen Prognosen und Trends entwickelt werden.

Im Fokus steht zudem die Umsetzung geplanter Data-Science-Initiativen. Das vorhandene Datenverständnis innerhalb der evm soll durch den technologischen Ausbau massiv unterstützt werden. Aus Sicht der Datenanalyse müssen hierfür unterschiedliche Anforderungen an Systeme und Datenbanken erfüllt sein:

> Relevantes Programmierwerkzeug (R, SQL, Python, etc.)

> Cluster für unabhängige Datenverarbeitung vom Produktivsystem

> Kundenzentriertes Datenmodell

> Datenbereinigung

> Möglichkeit der schnellen Datenerweiterung (z. B. Anbindung externer Quellen über ETL-Strecken)

> Bidirektionaler Zugriff (relevante Kundendaten für Analysen nutzen und gewonnene Erkenntnisse zurückspielen, wie Kennzahlen, Kennzeichen, Regelwerke)

> Veränderungen von Rohdaten in Echtzeit betrachten, um relevante Daten abzugreifen

Das neue Modern Data Warehouse soll somit ein Ort der zentralen Datensammlung für Analysezwecke und wirtschaftliche Entscheidungshilfe werden. Um dies zu erreichen, müssen die Daten vollständig und wahrheitsgemäß erfasst werden. Alle relevanten Quellsysteme werden hierfür an den Data Lake, der dem neuen Warehouse zugrunde liegt, angebunden. Relevante Kundendaten müssen auf Rohdatenebene und idealerweise in Echtzeit zugänglich sein. Idealerweise wird das neue Modern Data Warehouse von den Nutzern ohne Umweg über CSV als direkte Datenquelle für Auswertungen genutzt.

Jetzt Beratung anfragen!

"Wichtig für die Weiterentwicklung des Datenmodells im neuen Modern Data Warehouse ist, dass wir weg von der reinen Vertragssicht hin- ein in die komplexere Kundensicht wechseln."

Sara Burdenski
Marktmanagement und Innovation
Energieversorgung Mittelrhein

Projektverlauf

In einem Data Thinking Workshop sowie einem darauffolgenden Data Discovery Workshop erfassen die Energieversorgung Mittelrhein und taod die Ist-Situation. Die Identifikation erster Use Cases sorgt für eine detaillierte Anforderungsanalyse, die schließlich zu einem MVP-Ansatz führt. Auf Basis der vorhandenen Datenstrategie erfolgt die Sprint-Planung, die permanente agile Erweiterungen zulässt.

Aus Sicht der evm ist die agile Datenmodellierung grundlegend für den Erfolg ihrer Vision. Mit dem neuen Modern Data Warehouse soll die Basis für eine kundenorientierte Prozessgestaltung und Ansprache geschaffen werden. Das kundenzentrierte Datenmodell ist somit die Basis für wertschöpfende Datenanalyse und aufkeimende Data-Science-Initiativen. Für den Modellaufbau müssen folgende Faktoren besonders berücksichtigt werden:

> Auswertungslogik muss bedacht werden

> Eindeutigkeit eines Datensatzes muss gewährleistet sein

> Performante Datenstruktur sicherstellen

> Historisierung von Bestandsbewegungen ermöglichen

Record-Linkage-Verfahren zur Kundensichtgenerierung

Bislang wurden Datenabfragen und Datenabgleiche manuell vorgenommen. Das führte zu erheblichen Problemen. Beispielsweise wurden für einen Haushalt mehrere Kundennummern geführt, wenn innerhalb des Haushalts unterschiedliche Personen diverse Kontaktpunkte der evm berührten. Gewünscht ist daher die Entwicklung einer eindeutigen Kundensicht-ID, um Dopplungen und Missverständnisse zu vermeiden, Datensätze zusammenzuführen und Daten zu bereinigen.

Da die Daten im Abrechnungssystem eine mangelnde Qualität aufweisen, nicht über mehrere Attribute hinweg geprüft werden können und auch Ähnlichkeitsvergleiche nicht möglich sind, soll das Record-Linkage-Verfahren zum Einsatz kommen. Hierbei wird jedes Attribut (zum Beispiel Name oder Anschrift) zwischen zwei Datensätzen verglichen, eine Ähnlichkeitswahrscheinlichkeit bestimmt und kategorisiert in gleich, ähnlich oder ungleich. Auf Basis der Ähnlichkeitseinschätzungen der ausgewählten Attribute lernt das Record-Linkage-Modell automatisiert wie wahrscheinlich es ist, dass zwei Datensätze zusammengehören. Da diese Berechnungen komplexe statistische Verfahren (wie zum Beispiel Expectation Maximization) erfordern, ist der Aufbau einer adäquaten Cloud-Infrastruktur unumgänglich.

Herausforderungen

> Sehr aufwändige Datenquellanbindung

> Umfangreiche Entwicklung der Software-Umgebung

> Datenaufbereitung nicht automatisiert

> Viele Fehlerquellen (eigene Software-Pakete)

> Fehlende Transparenz und Dokumentation

> Hohe Fehleranfälligkeit

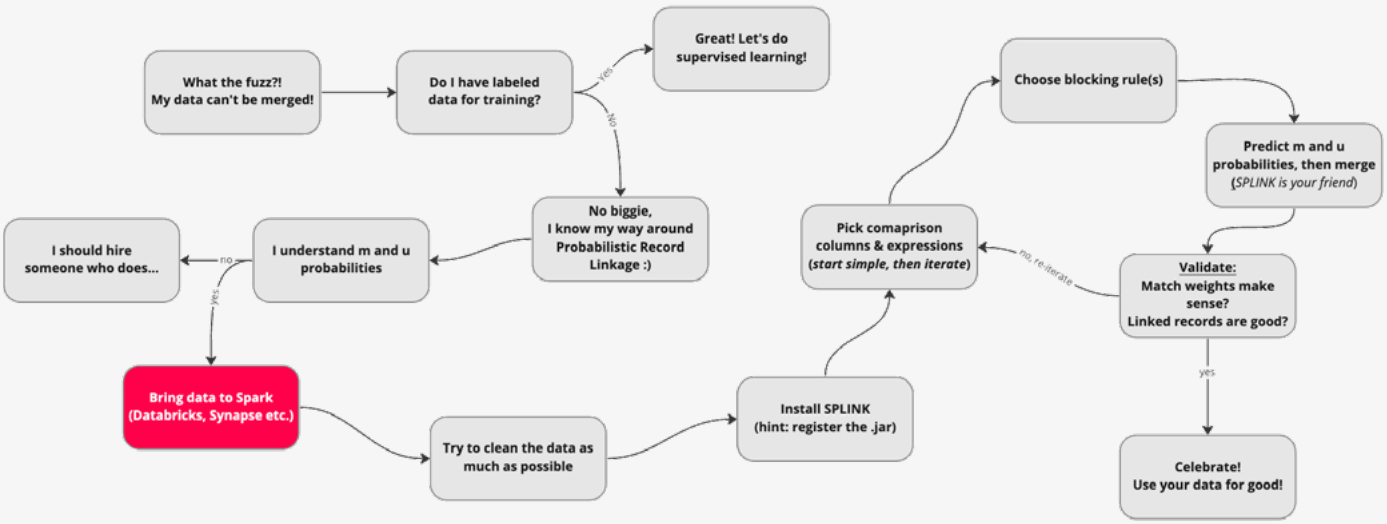
> Mangelnde Skalierbarkeit

Record Linkage for beginners

A Hitchhikers Guide by taod

Werden umfangreiche Daten und Datensätze aus unterschiedlichen Datenquellen verarbeitet, kann es immer wieder zu Problemen mit Duplikaten kommen. Durch Eingabefehler oder Probleme bei der Übertragung von Daten, unterschiedlicher Schreibweisen oder Abkürzungen, aber auch aufgrund unterschiedlicher Datenschemata können aus einem einzigen Datensatz mehrere Dateneinträge entstehen, die alle dieselben Informationen beinhalten. Diese Duplikate können entweder identisch oder nichtidentisch sein. Identische Duplikate sind relativ leicht zu erkennen und zusammenzuführen. Nichtidentische Duplikate hingegen unterscheiden sich in ein bis mehreren Werten. Ihre Bereinigung wird somit um ein Vielfaches komplexer. Denn die überflüssigen Duplikate können nicht einfach gelöscht werden, sondern müssen konsolidiert und miteinander verglichen werden.

Entsprechende heuristische Analysen in Verbindung mit Algorithmen unterstützen die Korrektur der Daten. Für die Erkennung und Bereinigung von Duplikaten wird häufig auf das sogenannte Record Linkage zurückgegriffen. Hierbei handelt es sich um ein automatisches Verfahren, das Fehler bei der Zusammenführung mehrere Datenquellen oder bei der Datenbereinigung identifiziert und behebt. Befinden sich Unternehmen an dem Punkt, an dem die Zusammenführung ihrer Daten notwendig ist, stehen sie vor unterschiedlichen Herausforderungen. Die Qualität der Daten ist genauso ausschlaggebend für eine erfolgreiche Konsolidierung der Datensätze wie der Einsatz korrekter Algorithmen innerhalb eines leistungsfähigen technologischen Settings. Wie ein solcher Entscheidungsbaum aussehen kann illustriert der taod Hitchhikers Guide.



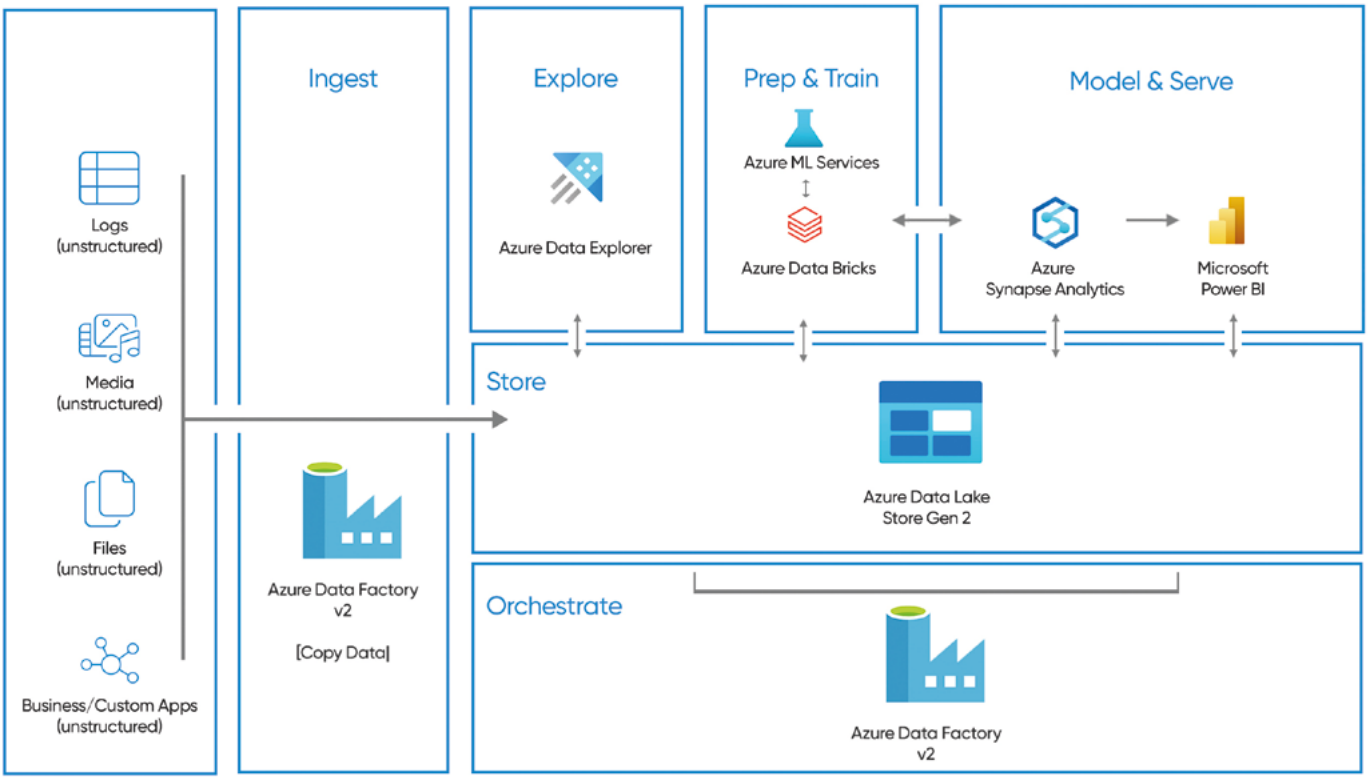
Daten intelligent zusammenführen: Der Hitchhikers Guide von taod

Aufbau einer Cloud-Architektur in Azure Synapse

Azure Synapse bietet für den Aufbau der Cloud-Umgebung für die Energieversorgung Mittelrhein nicht nur die nötige Datenanbindung und Datenaufbereitung durch ETL-Pipelines, sondern auch eine nahtlose Integration in Cluster-Compute-Instanzen in Apache Spark. Als Analyse-Engine verarbeitet Spark große Datenmengen und nutzt integrierte Module für SQL, Streaming, maschinelles Lernen und die Verarbeitung von Graphen.

In Azure Synapse ist die jeweils beste Technologie aus unterschiedlichen Bereichen vereint: SQL-Technologie für Data Warehousing in Unternehmen, Spark-Technologie für Big-Data-Zwecke, Data Explorer für die Analyse von Protokollen und Zeitreihen, Pipelines für die Datenintegration und ETL/ELT sowie eine tiefe Integration in andere Azure-Dienste, wie zum Beispiel Power BI, Cosmos DB und Azure ML.

Mit diesem Tech-Stack ist es gelungen, bestehende Ansätze zur eindeutigen Kundenidentifizierung zu automatisieren, die Ergebnisse flexibel in das neue Modern Data Warehouse zu integrieren und mit einem adäquaten Monitoring in die Geschäftsprozesse einzubinden. Die aufgebaute Struktur spart dabei zusätzlich unnötige Compute-Kosten, da für neue Kundendaten in der Regel kein neuer Modelldurchlauf trainiert werden muss. Es kann so lange auf bestehende Parameterschätzungen zurückgegriffen werden, wie das Modell im Monitoring eine eigens dafür definierte Qualitätskenngröße nicht unterschreitet.



Aufbau eines Modern Data Warehouse in Azure

“Die wichtigsten Säulen im Record-Linkage-Verfahren sind konsequente Datenbereinigung und intelligente Wahl der Vergleichsattribute. Module wie SPLINK bieten dabei das nötige Werkzeug, damit sich Data Scientisten auf diese beiden Fragen konzentrieren können, bei denen Erfahrung ebenso wie technisches und statistisches Verständnis eine entscheidende Rolle spielen.”

Christopher König
Managing Consultant
taod Consulting

Ergebnis: Technisches Enablement für geplante Data-Science-Initiativen

Gemeinsam mit taod ist der Energieversorgung Mittelrhein die komplexe und aufwändige Bereinigung ihrer Datenbasis unter Einsatz des Rekord-Linkage-Verfahrens gelungen. Der Aufbau eines neuen cloudbasierten Modern Data Warehouse in Azure bildet zudem eine zuverlässige und performante Grundlage für alle zukünftigen Data-Science-Initiativen.

Jetzt Beratung anfragen!

taod Services

> Cloud Data Platform Consulting

Wie können wir dich beraten?

Kontaktiere uns gerne unverbindlich, wenn du dich für das Thema Modern Data Stack interessierst und dich von uns beraten lassen möchtest.

Andreas Huppert
Managing Consultant
+49 151 53429328
andreas.huppert@taod.de

Jetzt Beratung anfragen!

taod

Raise the potential in every byte

Hinweis: Zur besseren Lesbarkeit wird in dieser Case Study, neben Doppelformen und Partizipialformen, das generische Maskulinum verwendet. Die verwendeten Personenbezeichnungen beziehen sich – sofern nicht anders kenntlich gemacht – auf alle Gender.

Verwendete Bilder
taod Consulting GmbH
istockphoto.com/topae
istockphoto.com/firina

Datenschutzbeauftragter
Frank Gundlach
GTB – Genossenschafts-Treuhand Bayern GmbH Wirtschaftsprüfungs-gesellschaft

Türkenstrasse 22 - 24
80333 München

+49 170 9416034
fgundlach@gv-bayern.de

Kontakt
taod Consulting GmbH
Oskar-Jäger-Str. 173, K4
50825 Köln

+49 221 975 849 70
info@taod.de

Vertreten durch
Simon Biela, Matthias Steinforth, Benedikt Stienen

Amtsgericht Köln HRB 95089