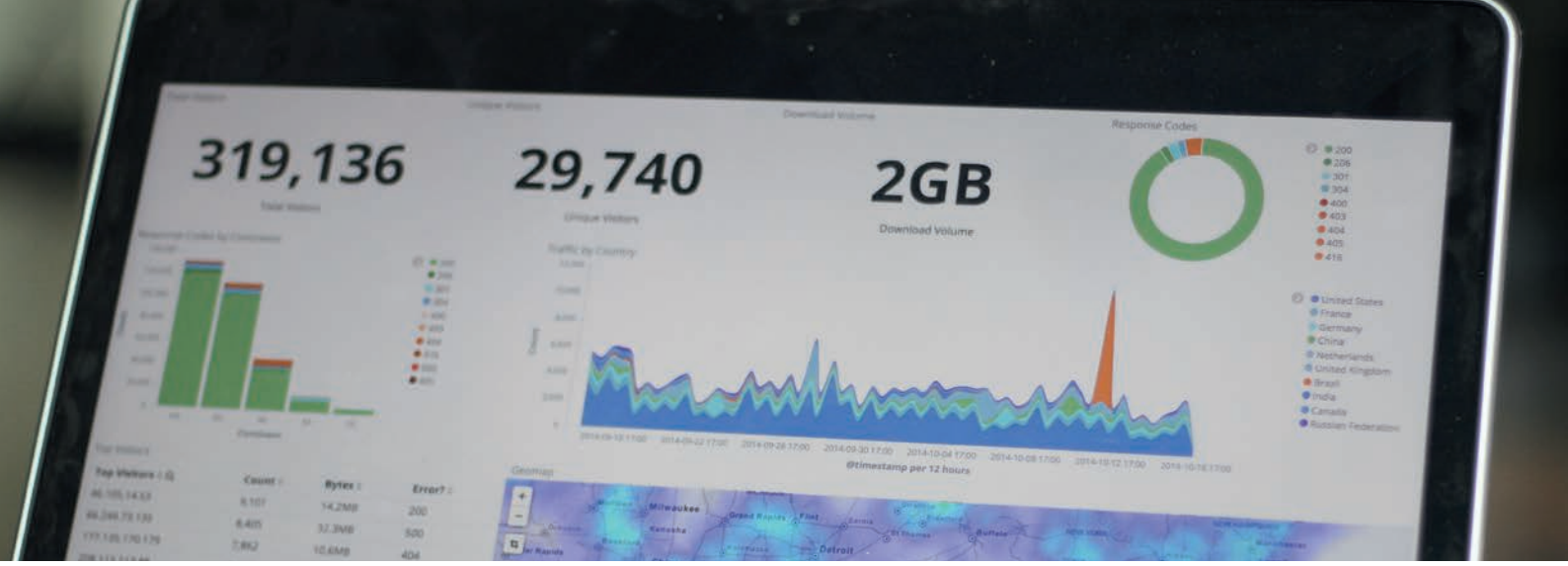


Case Study

Best of Both Worlds: Aufbau einer hybriden BI-Architektur mit Big-Data-Komponenten für dmTECH

inovex

Die IT-Tochter von dm, Deutschlands umsatzstärkstem Drogeriemarkt, fokussiert sich mit mehr als 800 Mitarbeitern auf die Digitalisierung des Handels. Im Mittelpunkt steht die Entwicklung von innovativen Lösungen: für den Online-Shop, Kunden- und Mitarbeiter-Apps sowie für die IT in den dm-Märkten, den Verteilzentren und der Zentrale. Bei vielen Innovationen spielen der effiziente Umgang mit und die Analyse von großen Datenmengen eine zentrale Rolle, um Mehrwerte aus der Kombination von Datenquellen zu generieren (Big Data). Vor diesem Hintergrund arbeitet dmTECH gemeinsam mit inovex an verschiedenen Projekten, die jeweils neue Big-Data-Möglichkeiten implementieren. Das Ziel des Projekts „BI Analytics“ bestand darin, die Möglichkeiten einer bestehenden Data-Warehouse-Lösung mit Big-Data-Technologien zu erweitern. In der neuen hybriden Architektur sollen neuartige Use Cases umgesetzt werden, beispielsweise eine skalierbare Lösung zur Einzelbelegauswertung.

Projekt-Setting

Die BI-Experten bei dmTECH betreiben seit vielen Jahren eine Data-Warehouse-Infrastruktur, in der über kontrollierte ETL-Prozesse Daten aus vielen verschiedenen Bereichen des Unternehmens gesammelt, aufbereitet und zentral abrufbar gemacht werden. Dieser Ansatz eignet sich für strukturierte Daten, die über standardisierte ETL-Entwicklungsbausteine für unternehmensweit harmonisierte Auswertungen bereitgestellt werden.

Neue Analyseanforderungen werden häufig durch zwei verschiedene Zielsetzungen motiviert: Einerseits sollen größere Mengen an möglicherweise weniger stark strukturierten Daten oder Rohdaten effizient verarbeitet werden können; andererseits sollen über einen Streaming-Ansatz Ergebnisse nahezu in Echtzeit zur Verfügung gestellt werden können.

Für die Adressierung beider Aspekte eignen sich Big-Data-Technologien wie beispielsweise Apache Hadoop, die den Funktionsumfang der Data-Warehouse-Lösung ergänzen.

Als konkretes Projekt für die Implementierung des hybriden Ansatzes soll im Rahmen dieser Case Study die Plattform zur Einzelbelegauswertung vorgestellt werden. Aus regulatorischen und steuerrechtlichen Gründen muss dmTECH in der Lage sein, die historischen Kassenbelege, etwa im Zuge einer steuerlichen Prüfung, auswertbar zur Verfügung zu stellen. Hierbei entstehen durchgängig sehr große Mengen an Kassenbelegen, die kontinuierlich als Datenstrom von allen dm-Märkten in die IT-Systeme eingespeist werden.

Ziel der Einzelbelegauswertung ist es, die großen Datenmengen innerhalb dieser Datenströme aufzubereiten und damit Realtime Reporting zu ermöglichen.

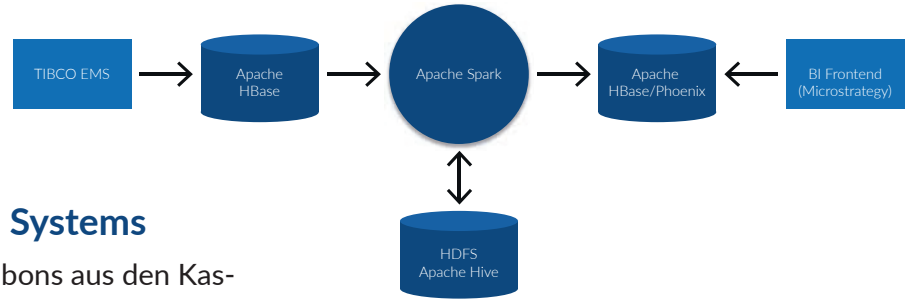
Konkret ist es von Interesse, punktuelle Einsichten und Informationen von einzelnen Kassenbelegen zu erhalten. Damit kann einerseits der Abruf eines bestimmten Belegs gemeint sein („Ich benötige den Beleg vom 24.12.2018 aus der Filiale 123 um 11:33 Uhr an Kasse 2.“) oder eine aggregierte Sicht auf Teile der Daten („Ich benötige den summierten Umsatz aller Kassenbelege aus KW 13 im Jahr 2018, auf denen die Warengruppe 12 enthalten ist.“).

Hadoop als Enabler

Das Ökosystem von Apache Hadoop bietet einige Komponenten, mit denen sich die Anforderungen sehr gut abbilden lassen – insbesondere die skalierbare Ablage und Verarbeitung von großen Datenmengen.

Als konkretes Werkzeug erlaubt beispielsweise Apache HBase eine skalierbare Ablage von (Roh-)Kassenbelegen mit zeitgleichem effizienten Zugriff auf einzelne Bons. In der weiteren Verarbeitung können die Daten zur Massenverarbeitung im Hadoop Distributed Filesystem (HDFS) abgelegt und mit Apache Spark performant analysiert und aufbereitet werden. Für den Zugriff über eine SQL-Schnittstelle eignen sich Apache Phoenix oder Apache Hive, wodurch die

Integration in bestehende BI-Frontends und damit interaktives Reporting möglich wird.



Technische Details des Systems

Technisch werden die Kassenbons aus den Kassen aller Filialen über einen Filial-Server an die bestehende Middleware-Lösung (TIBCO EMS) bei dmTECH geschickt. In dieser Middleware wurde ein Konnektor eingerichtet, der jeden einzelnen Kassenbon direkt in das NoSQL-System Apache HBase schreibt. Somit steht jeder einzelne Bon kurz nach seiner Entstehung als Rohdatum in der Auswertungslösung zur Verfügung. Innerhalb von HBase können die Bons über einen eindeutigen Schlüssel identifiziert werden – dies erlaubt beispielsweise eine direkte Deduplizierung der Daten im Falle von Mehrfachlieferungen.

In der so abgelegten rohen Form sind die Kassenbelege allerdings nicht performant entlang aller relevanter Dimensionen auswertbar. Um dies zu ermöglichen, werden die Rohdaten mit Apache Spark gelesen, transformiert und über verschiedene Zwischenschritte in die finalen auswertbaren Tabellen geschrieben. Diese sind wiederum in HBase gespeichert und werden über die SQL-Schicht Apache Phoenix zugänglich gemacht. Zwischenergebnisse werden im Hadoop Distributed Filesystem (HDFS) in Form von Hive-Tabellen vorgehalten, um eine effiziente Berechnung zu gewährleisten.

Alle Komponenten entlang dieser Aufbereitung sind verteilte Systeme, die über Parallelverarbeitung gute horizontale Skalierungseigenschaften aufweisen. Somit kann beispielsweise bei wachsender Datenmenge der Hadoop Cluster um Rechenknoten erweitert werden, um die Verarbeitung weiter in einem akzeptablen Zeitrahmen durchführen zu können.

Die folgende Skizze veranschaulicht das Zusammenspiel der Systemkomponenten.

Auswertung

Apache Phoenix verwendet als Ablagespeicher die performante NoSQL-Datenbank HBase. Durch die Auswertungsschicht lassen sich Daten schnell analysieren und entsprechende Kennzahlen generieren. Ein wichtiger Aspekt beider Technologien ist, dass Tabellen in HBase/Phoenix jeweils für bestimmte Auswertungsmuster im Voraus optimiert werden können, vergleichbar mit dem Erstellen von Indizes in relationalen Datenbanken. Durch die Möglichkeit, in Spark performant aus den Rohdaten oder den Zwischenschichten neue Auswertungstabellen zu generieren, können auch zukünftige Reporting-Anforderungen gut abgedeckt werden.

Kern der Auswertung

Als Schnittstelle zur Benutzung der Einzelbelegauswertung wird das vorhandene BI-Werkzeug Microstrategy verwendet. Dieses wird auch in vielen anderen Use Cases bei dmTECH genutzt und ist dementsprechend gut etabliert. Die Integration von Microstrategy mit den Daten in der Auswertungslösung erfolgt über Apache Phoenix. Diese Integration mit Hadoop ist vorteilhaft, weil Phoenix ein SQL Interface bietet, das standardisiert über JDBC angesprochen werden kann. Dies beinhaltet eine Abstrahierung der komplizierten Zugriffsmechanismen der darunter liegenden HBase-Datenbank. Durch diese Standards (SQL und JDBC) ist der Zugriff auf Hadoop für außenstehende Systeme einfach umzusetzen.

Fazit und Wiederverwendbarkeit

Die beschriebene Erweiterung der bestehenden BI-Architektur bei dmTECH mit Komponenten aus dem Hadoop-Ökosystem erlaubt es, Kennzahlen performant aus einem extrem umfangreichen Datenbestand abzuleiten. Dieser Ansatz erweitert somit die Möglichkeiten der bestehenden Plattform und erlaubt durch die horizontale Skalierbarkeit eine kostengünstige Anpassung an zukünftig steigende Datenmengen und komplexere Auswertungen. Über die Standardschnittstellen SQL und JDBC werden die neuen Systeme nahtlos in die bestehende Landschaft integriert. Dies erlaubt die individuelle Entscheidung bei weiteren Use Cases, welche Technologie sich für die jeweiligen Anforderungen am besten eignet – Hadoop oder die bestehende Infrastruktur. Dieser hybride Ansatz kann also – *das beste aus beiden Welten* – zusammenbringen und damit die Möglichkeiten in Summe erweitern.

Die Ausrichtung auf Hadoop wird sich auch in zukünftigen Projekten und bei weiteren Use Cases auszahlen. Großes Interesse besteht etwa an der Auswertung der Kennzahlen in Echtzeit. Ein entsprechendes Realtime Dashboard lässt sich mit den bereits beschriebenen Hadoop-Frameworks unter Hinzunahme beispielsweise des Streaming Frameworks Apache Nifi umsetzen. Damit ist es möglich, Umsätze anhand der Kassenbelege in Near-Realtime zu untersuchen.

Technologien

› Apache Hadoop
› Apache Spark
› Apache Hive
› Apache HBase
› Apache Phoenix
› TIBCO EMS
› Microstrategy

„Dieses Projekt war für uns der Türöffner zur Anwendung des Data Lake-Konzeptes auf Basis der Hadoop Technologien. Der hybride Ansatz ermöglicht uns das effiziente Speichern großer Datenmengen, um diese in Realtime zu analysieren und Einsichten gewinnen zu können. Mit inovex an unserer Seite haben wir erfolgreich die bestehende DWH/BI-Infrastruktur um eine State-of-the-Art Data-Lake-Landschaft erweitert und damit die Basis geschaffen, um anstehende Analytics-Anforderungen umsetzen zu können."

— Thomas Herwerth
(Produktverantwortlicher DWH/BI, dmTECH)

Über inovex

Über 350 IT-Expert:innen unterstützen Unternehmen umfassend bei der Digitalisierung und Agilisierung ihres Kerngeschäfts und bei der Realisierung von neuen digitalen Use Cases.

Unser Lösungsangebot umfasst Application Development (Web Platforms, Mobile Apps, Smart Devices und Robotics – vom UI/UX Design bis zu den Backend Services), Data Management & Analytics (Business Intelligence, Big Data, Search, Data Science & Deep Learning, Machine Perception und Artificial Intelligence) und die Entwicklung von skalierbaren IT Infrastructures (IT Engineering, Cloud Services), auf denen die digitalen Lösungen im DevOps-Modus betrieben werden. Wir modernisieren vorhandene Lösungen (Replatforming), härten Systeme gegen Angriffe von außen (Security) und vermitteln unser Wissen durch Trainings und Coachings (inovex Academy).

inovex ist in Karlsruhe, Pforzheim, Stuttgart, München, Köln und Hamburg ansässig und bundesweit in Projekte involviert.

Nehmen Sie Kontakt auf

› Haben Sie Fragen zum Thema Business Intelligence?
› Suchen Sie einen Partner, der Sie bei der Optimierung Ihrer Big-Data-Lösung unterstützt?
› Möchten Sie mehr über inovex und unser Portfolio für die digitale Transformation erfahren?

Ihr Ansprechpartner

Dr. Dominik Benz
Head of ML Engineering
dominik.benz@inovex.de
+49 173 3181 94
inovex.de/business-intelligence