Social-Media-Auswertungen im Data Warehouse

Martin Frisch, OPITZ CONSULTING Deutschland GmbH

Big-Data-Ansätze ermöglichen es nicht nur, immer größere Datenmengen zu verarbeiten, sondern auch, neue Arten von Datenquellen für Auswertungen und Analysen heranzuziehen. Insbesondere Web-2.0-Inhalte, die Nutzer sozialer Netzwerke erstellen, bieten vielfältige Potenziale.

Dieser Artikel beschreibt ein Fallbeispiel aus einer Master-Thesis, in dem Kommentare aus Twitter zur Fernsehserie „Tatort" für die Erweiterung eines Data Warehouse (DWH) verwendet wurden. Das Beispiel zeigt, wie Big-Data-Komponenten ein auf Oracle basierendes DWH ergänzen können, um Kommentare aus Twitter zu extrahieren, mit einer Sentiment-Analyse zu bewerten, im DWH bereitzustellen – und schließlich in Berichten aufzubereiten. Für die Ausgangssituation wurde zunächst ein Business-Intelligence-System mit einem klassischen DWH für die Tatort-Analyse aufgebaut. Dieses umfasst die Bereiche Datenquelle, Datensicherung, Verarbeitung und Auswertung *(siehe Abbildung 1)*.

Als Datenquelle liegen Stammdaten zu den Ermittlern, den Teams und den Episoden des Tatorts sowie die absoluten und prozentualen Einschaltquoten in CSV-Dateien vor. Der Bereich „Datensicherung" entspricht einem DWH mit den drei Schichten „Stage", „Core" und „Mart". Diese befinden sich in einer Oracle-12*c*-Datenbank. Die ETL-Verarbeitung geschieht mit dem Oracle Data Integrator (ODI), der die Daten aus den CSV-Dateien in die Stage-Schicht extrahiert. Von dort integrieren die Ladeprozesse die Daten in den Core, der nah an der dritten Normalform modelliert ist. Der Mart entspricht dagegen einem Starschema. Auf diesem setzt ein Dashboard Reporting mit Oracle Answers auf, einem Werkzeug der Oracle Business Intelligence Suite (Version 11*g*). In den Berichten können Anwender die Einschaltquoten nach folgenden Kriterien auswerten:

- Datum
- Episode
- Spielort
- Ermittler

Damit der Erfolg der Episoden nicht nur an den Einschaltquoten messbar ist, wurde das DWH um Kennzahlen erweitert, die die Stimmung der Twitter-Gemeinde zu den Tatort-Episoden oder bestimmten Komponenten und Eigenschaften wie Ermittlern, Teams, Handlung, Musik etc. widerspiegeln. Dabei war zu berücksichtigen, dass Daten aus sozialen Netzwerken im Vergleich zu Daten aus klassischen Quellen wie zum Beispiel einem operativen System einige Besonderheiten aufweisen. So sind Social-Media-Daten weitaus weniger strukturiert. Zudem ist eine auswertbare Größe zur Stimmung eines Tweets in diesem nicht direkt enthalten, sondern muss mit Methoden des Text-Mining aus dem jeweiligen Text heraus er-

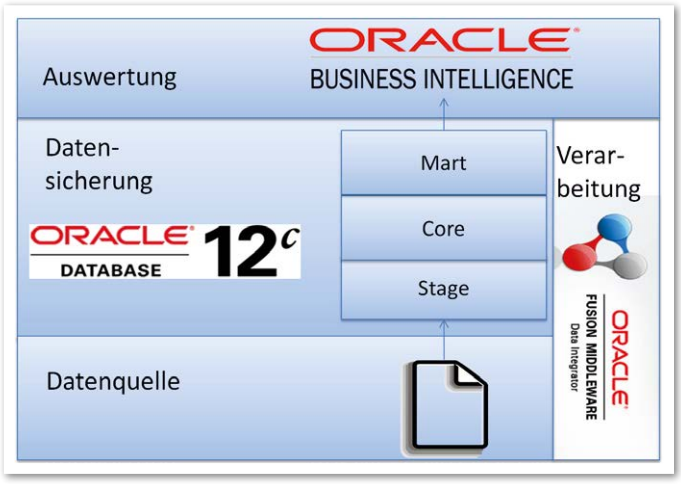

Abbildung 1: Klassisches Data Warehouse

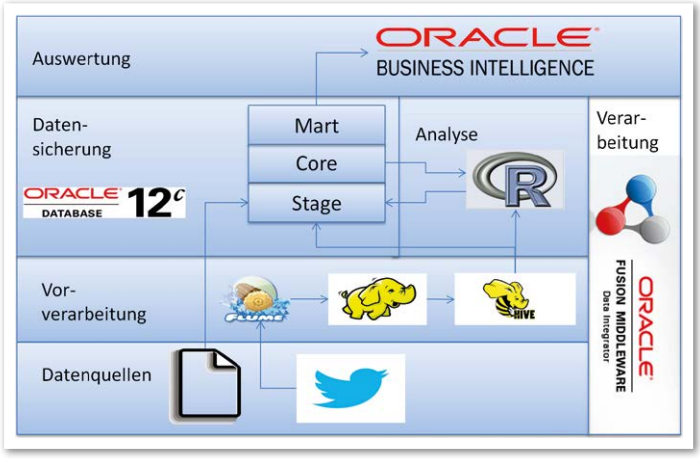

Abbildung 2: Erweitertes Data Warehouse

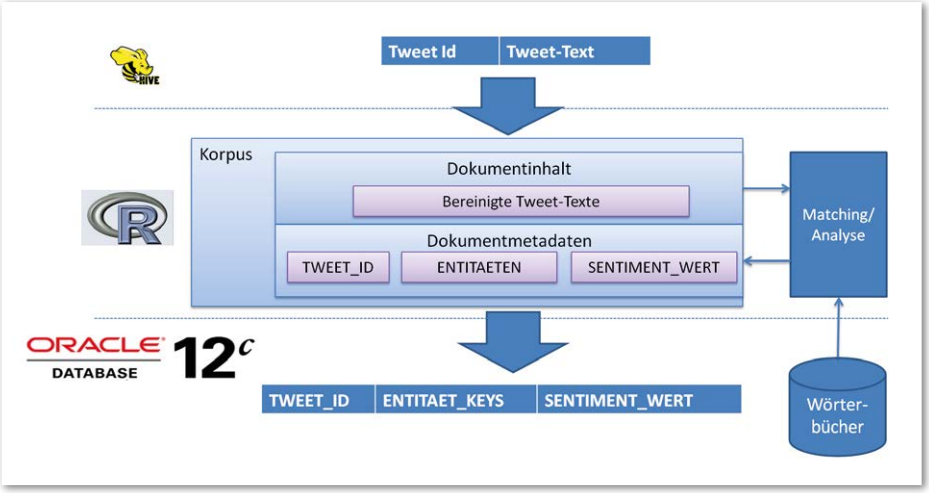

Abbildung 3: Analyse in R

mittelt werden. Für diese Aufgaben sind spezielle Werkzeuge erforderlich.

Erweiterungen für die Social-Media-Analyse

Die bisherige Architektur wurde um einen Vorverarbeitungs- und einen Analyse-Bereich erweitert *(siehe Abbildung 2)*. Zur Vorverarbeitung kommen Werkzeuge des Hadoop-Rahmenwerks zum Einsatz. Dessen Cloudera-Distribution steht auch auf der Oracle Big Data Appliance zur Verfügung. Mit diesen Tools können die relevanten Twitter-Daten extrahiert und in eine relationale Form überführt werden. Die Tweets werden mithilfe des Twitter-Streaming-API während der Episoden-Ausstrahlung mitgeschnitten.

Die Extraktion des Twitter-Streams geschieht mit Flume, einem Werkzeug, mit dem sich große Datenmengen bewegen lassen und das Streaming-Daten in Echtzeit verarbeiten kann. Dazu benötigt Flume die Konfiguration eines sogenannten „Agenten", der aus Quelle („Source"), Zwischenspeicher und Ziel („Sink") besteht. Als Quelle dient ein Java-Programm aus dem Cloudera GitHub *[1]*, um das Streaming-API von Twitter als benutzerdefinierte Quelle verwenden und dafür die öffentliche Java-Bibliothek „Twitter4J" nutzen zu können *[2]*.

Zur Identifikation der relevanten Tweets wird das Hashtag „#tatort" als Such-Schlüssel mitgegeben. Als Zwischenspeicher reicht der interne Speicher von Flume aus, da dort keine Transformationen stattfinden. Der Agent verwendet das Hadoop-Filesystem (HDFS) als Ziel. Dort wird zunächst das JSON-Format beibehalten, in dem Twitter neben den Tweet-Texten noch zahlreiche weitere Attribute wie Zeitpunkt, Ort und Autor der Veröffentlichung liefert. Der Speicherpfad ist dynamisch, sodass Flume für jede Episoden-Ausstrahlung einen eigenen Ordner anlegt.

Hive kann in Kombination mit einem Serializer/Deserializer („SerDe") unterschiedlichste (semi-strukturierte) Formate wie JSON in eine relationale Struktur überführen und mit der SQL-ähnlichen Abfragesprache HiveQL selektieren. Ein fertiger „SerDe" für das JSON-Format findet sich beispielsweise auch im Cloudera GitHub. Mit diesem wurde eine nach Episoden partitionierte „External Table" angelegt, die eine Tabellen-Struktur über die JSON-Daten im HDFS legt. Jede Partition zeigt auf genau einen durch Flume angelegten Ordner. Im ODI sind Befehle zum Anfügen und Löschen von Partitionen in Prozeduren gekapselt.

Da immer eine Episode in einem Batch verarbeitet wird, eignet sich die Partitionierung hier auch zum dynamischen Filtern der Daten. Die Ladeprozesse bestimmen zur Laufzeit das Datum der zu verarbeitenden Episode und legen zu Beginn der Verarbeitung eine neue Partition an. Nach erfolgreicher Verarbeitung löschen die Ladeprozesse diese Partition wieder, wobei die Daten physisch im HDFS erhalten bleiben. Somit verweist die Tabelle während der Verarbeitung nur auf Daten zur aktuellen Episode und muss nicht mehr gefiltert werden. Vor und nach der Verarbeitung ist die Tabelle leer.

Zu diesem Zeitpunkt liegen die Twitter-Daten in einer Tabellen-Struktur vor, die aber noch stark denormalisiert ist. Sie enthält Redundanzen und nicht-atomare Attribute (wie Arrays oder zusammengesetzte Datentypen). Um diese aufzulösen, finden bereits in Hive erste Aufbereitungen mithilfe von vier Views statt, die die Informationen nach den folgenden Entitäten gliedern:

- Tweet(-Text)
- Twitter-Nutzer
- Ort
- Anzahl der Retweets

Informationen, die später im DWH und in den Berichten nur als Dimension oder Detail-Information Verwendung finden, werden direkt in den Stage-Bereich des DWH geladen. Dazu bietet der ODI einen entsprechenden Hive-Konnektor. Im Mart wird mithilfe dieser Daten eine (faktenlose) Fakten-Tabelle zu den Tweets und jeweils eine eigene Dimension für Twitter-Nutzer und Twitter-Orte bereitgestellt.

Eine besondere Herausforderung stellte die Harmonisierung der Ortsinformationen dar. Häufig sind diese leer oder enthalten unterschiedliche Granularitä-

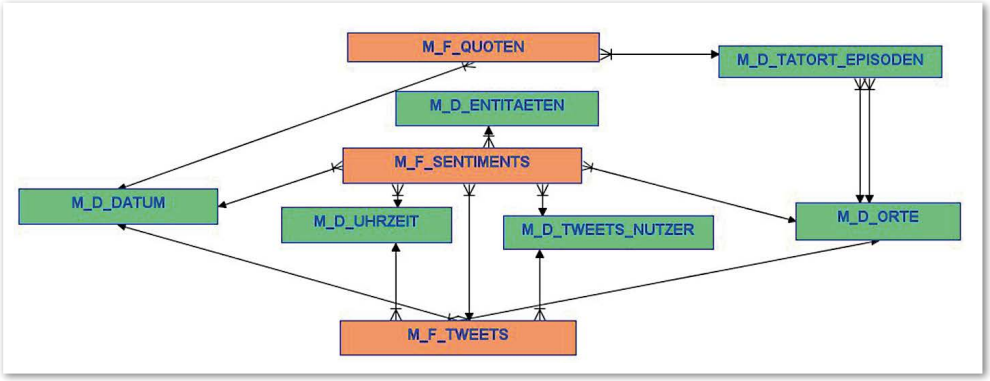

Abbildung 4: Das Mart-Datenmodell

ten, Sprachen und Schreibweisen für den gleichen Ort. Da die Ortskoordinaten vom Twitter-API mitgeliefert werden, konnte dieses Problem mit Referenzdaten gelöst werden, die ebenfalls über Ortskoordinaten verfügen.

Analyse in R

Die Ermittlung der eigentlichen Stimmung geschieht im Analyse-Bereich mit der Oracle-Distribution des Statistik-Werkzeugs R (Version 3.01). Diese Distribution verfügt über Konnektoren, um auf Hive zuzugreifen und Daten mit der Oracle-Datenbank auszutauschen. R importiert, bereinigt und analysiert die Tweet-Texte aus Hive und sichert die Ergebnisse in der Oracle Datenbank *(siehe Abbildung 3)*.

Neben den reinen Tweet-Texten importiert R auch die zugehörigen Tweet-IDs aus Hive. Die Texte werden als Dokumente in einem sogenannten „Korpus" vorgehalten. Diese speziell für Dokumente ausgelegte Speicherstruktur erlaubt es, den eigentlichen Inhalten auch Metadaten zuzuordnen. Der Korpus für die Tatort-Tweets speichert die IDs und die Ergebnisse der Analyse als solche Metadaten. Die Analyse selbst erfolgt in drei Schritten:

1. Bereinigung der Tweet-Texte. Dabei werden insbesondere überflüssige und störende Textteile wie Stoppwörter, Sonderzeichen, Zahlen oder Leerzeichen entfernt. Zudem werden alle Buchstaben kleingeschrieben, Umlaute aufgelöst und Wortstämme gebildet.
2. Für jeden Tweet wird mit einem lexikonbasierten Ansatz der Sentiment-Analyse ein Stimmungswert ermittelt.
3. Jedem Tweet werden Entitäten zugeordnet, auf die sich der Tweet bezieht, etwa ein bestimmter Tatort-Ermittler.

Sowohl der lexikonbasierte Ansatz der Sentiment-Analyse als auch die Zuordnung der Entitäten verwenden Wörterbücher. Da R selbst über keinen persistenten Speicher verfügt, verwaltet die Oracle-Datenbank diese Wörterbücher und stellt sie in Schnittstellen-Views für R zur Verfügung.

Das Wörterbuch für die Sentiment-Analyse enthält Phrasen, die eine bestimmte Stimmung ausdrücken. Diese Stimmung wird im Wörterbuch durch eine Zahl zwischen „–1" (sehr negativ) und „1" (sehr positiv) ausgedrückt. Während der Analyse werden diese Werte für die in einem Tweet-Text gefundenen Phrasen zusammengerechnet.

Im Wörterbuch für die Entitäten-Zuordnung befinden sich hingegen Kontext-Informationen und Eigenschaften, die im Wesentlichen auf den Stammdaten aus dem DWH basieren. Das sind in diesem Fall zum Beispiel Namen der Ermittler, der Spielorte oder der Episoden. Darüber hinaus besteht die Möglichkeit, zusätzliche Begriffe manuell zu konfigurieren.

Exemplarisch wurden einige Begriffe, die Charakteristika einer Krimiserie wie Handlung, Schauspieler oder anderes beschreiben, eingepflegt. Findet die Analyse einen Begriff aus diesem Wörterbuch in einem Tweet-Text, speichert sie in den Korpus-Metadaten für dieses Dokument einen eindeutigen Entitäten-Schlüssel zu diesem Begriff, der ebenfalls aus dem Wörterbuch stammt.

Nach der Analyse schreibt R die Ergebnisse in den Stage-Bereich des Tatort-DWH. Die gesamte R-Verarbeitung ist in kaskadierenden Skripten umgesetzt, die über einen einzigen Betriebssystembefehl vom ODI aufgerufen werden. Oracle R Enterprise kann die Skripte in der Oracle-Datenbank verwalten und über entsprechende SQL-Befehle starten, was in einem produktiven System mit vielen Skripten eine sinnvolle Alternative sein kann.

Nachdem sich die Ergebnisse der Analyse im DWH befinden, werden sie über die Tweet-ID den Twitter-Daten zugeordnet, die direkt von Hive in das DWH geladen wurden. Gleiches geschieht mit den Entitäten und den Stammdaten anhand der Entitäten-Schlüssel. Da die Stimmung der Tweets im Sentiment-Wert quantifiziert ist, kann sie im DWH wie klassische Quelldaten verarbeitet und aufbereitet werden.

Im Mart werden zusätzlich verschiedene Kennzahlen auf Basis des Sentiment-Wertes aus der Analyse berechnet. Beispielsweise wird der Sentiment-Wert auf einer Skala von „–1" bis „1" normalisiert. Durch eine Aggregation im Frontend kann daraus dann im Anschluss ein Durchschnittswert berechnet werden, der die Gesamtstimmung repräsentiert. Zudem werden die Tweets mit Flags versehen, die zeigen, ob der Tweet aufgrund des Stimmungswertes als positiv, neutral oder negativ zu bewerten ist.

Dank dieser Kennzeichnung kann in Berichten leicht die Komplettanzahl oder auch der Anteil positiver oder negativer Tweets ermittelt werden. *Abbildung 4* zeigt das komplette Datenmodell der Mart-Schicht, das aus drei Faktentabellen (rot) und sechs Dimensionstabellen (grün) besteht.

Fazit

Die Tatort-Analyse zeigt, wie durch das Anfügen eines Vorverarbeitungs- und Analyse-Bereichs auch Daten aus sozialen Netzen in einem DWH ausgewertet werden können. Oracle-Datenbank, ODI, Hadoop und R harmonieren dabei sehr gut miteinander. So profitiert beispielsweise die Analyse vom DWH als persistenter Speicherkomponente und Informationslieferant. Von dort können relevante Kontext-Informationen zur Domäne gefunden werden, was im Beispiel die Entitäten sind. Zudem können hier Konfigurationsdaten für die Analyse (etwa Wörterbücher) verwaltet werden.

Für die Umsetzung muss das DWH nicht umgebaut werden, sondern es reicht aus, es um einige Komponenten zu erweitern, die Spezialaufgaben beim Umgang mit Social-Media-Daten übernehmen. Der ODI kann weiterhin zur Steuerung und Überwachung der Gesamtverarbeitung eingesetzt werden. Gleiches gilt für die Oracle BI Suite als Frontend.

Durch das Beibehalten dieser Werkzeuge profitieren die Mitarbeiter eines Unternehmens weiterhin von ihren Kenntnissen in den bereits für sie bekannten Technologien. Dadurch haben sie weniger Berührungsängste mit den neuen Datenquellen und hinzugekommenen Technologien.

Quellen

[1] *https://github.com/cloudera/cdh-twitter-example*
[2] *http://twitter4j.org/en/index.html*

Martin Frisch
martin.frisch@opitz-consulting.com