Heterogenen Datenquellen zum Trotz – Möglichkeiten der Vernetzung einer Community mit dem Big Data-Ansatz

Patrick Pongratz

European IT Consultancy EITCO GmbH
Am Bonner Bogen 6
53227 Bonn
PPongratz@eitco.de

Derzeit gibt es im Agro-Food-Sektor keine nennenswerte nationale Vernetzung von Datenquellen. Wir verfügen in Deutschland über massive Datenvolumina, die jedoch nicht ausgewertet werden, da ein zentraler Knotenpunkt als Anlaufstelle für die diversen Stakeholder fehlt. Das zu hebende Potential liegt im Cloud-Ansatz in Kombination mit föderalistischer Datenhoheit und Datensicherheit. Insbesondere behördliche Einrichtungen tun sich in dieser Kombination schwer. Hier können technologisch mit dem Smart-Data / BigData Ansatz vollkommen neuartige Möglichkeiten aufgezeigt und in Kombination mit der Rechenleistung aus der Cloud eine deutlich verbesserte Reaktionsgeschwindigkeit eingebracht werden.

Mit einer Community-getragenen Knowledge Base im Agro-Food-Sektor kann ein technologischer Standard definiert werden, wie die Daten im Hinblick eines zentralen BigData-Hubs geliefert werden müssen. Die Herausforderung besteht darin, aus einer Vielzahl von Quellen und Datenhoheiten Informationen zu sammeln, sie nutzerspezifisch in Echtzeit aufzubereiten und über diverse Kanäle via automatisierte Prozessunterstützung innerhalb einer Community-getragenen Knowledge Base im Agro-Food-Sektor zur Verfügung zu stellen. Hierfür eignet sich beispielsweise ein sogenannter „Semantic Service Bus“, der nach festen Regeln und Prioritätsstufen zuerst Informationen im Rahmen der Community zusammenträgt, mit Hilfe eines BigData-Hubs / Netzwerk-Knotens aufbereitet und im Anschluss an dezidierte Empfänger versendet, sowie diese in der Knowledge Base zur Verfügung stellt. Diesbezüglich könnten dem, innerhalb der Stakeholder verantwortlichen Personenkreis, vier Kategorien von Informationen zur Verfügung gestellt werden:

- Beschreibende zur Beurteilung (Visualisierung)
- Vergleichende zur Auswahl von Handlungsalternativen
- Vorhersagende aus entsprechenden Simulationsprogrammen zur Abschätzung der Entwicklung
- Vorschreibende zur schnellen Umsetzung von Maßnahmen.

Diesbezüglich bedienen wir uns eines sogenannten ETL-Prozesses (Extract, Transform, Load), bei dem Daten aus mehreren ggf. unterschiedlich strukturierten Datenquellen in einer Zieldatenbank vereinigt werden. Für die Speicherung der Daten eignen sich NoSQL oder NotOnlySQL Datenbanken die häufig Key-Value Stores implementieren. Die benötigte Software-Infrastruktur kann aus einer BigData-Plattform wie Hadoop, Konnektoren zu den relevanten Datenquellen sowie Analyse-Tools wie Hive für Data-Warehousing, Mahout für Machine Learning oder Pig als interaktive Shell bestehen.

Eine Datenanreicherung mit unbekannten Datenquellen ist innerhalb der Knowledge-Base jederzeit möglich, da neue Datensets jederzeit neu gemischt oder zugeordnet werden können. So können neben dauerhaft beteiligten Partnern innerhalb der Community weitere interessierte Institutionen und Unternehmen auch im Falle einer weitreichenderen Simulation zu wertvollen Analysen beitragen.

Es ist innerhalb der Community möglich, die dauerhaft zur Verfügung stehenden Datenquellen in einem zu definierenden Zyklus anzusteuern, diese innerhalb zu definierender Algorithmen und Warnschwellen zu analysieren, daraus Anomalitäten oder versteckte Zusammenhänge automatisiert zu visualisieren und vordefinierte Stakeholder einzubinden bzw. zu alarmieren (Machine Learning / Data Mining).

Gerade die zu erwartende Menge an Daten verlangt nach einer Abstraktionsschicht oder einer Qualifizierung der Daten um über visuelles Data Mining bestimmte Muster aufzudecken. Nur so können Ursache und Wirkung miteinander verknüpft, Gruppen geclustert und klassifiziert, Ausreißer erkannt und die geographische Herkunft bestimmt werden. Eine blitzschnelle Datenverarbeitung und eine „Instant Advanced Analytics“ Auswertung kann vorab mit direktem Laden in eine spaltenorientierte in-Memory-Datenbank realisiert und die Ergebnisse mit einer integrierten Galerie mit anderen Anwendern geteilt werden. Sollte sich aus den Anomalien ein ernstzunehmender Handlungsbedarf abzeichnen, werden ein aktuelles Bild der zu erwartenden Auswirkung visualisiert, die Quelle nachvollzogen sowie Prognosen abgeleitet. Eine Metadatenmodellierung erfolgt erst dann, wenn sich erkennen lässt, was modelliert und operationalisiert werden muss. Hierfür wird unmittelbar und kurzfristig eine enorme Rechenleistung benötigt, die uns die Cloud-Technologie gewährleisten kann. Nur so lassen sich die bestehenden strukturierten Datenbanken in Kombination mit unstrukturierten Ad-hoc-Daten aus Smartphones, Social-Media-Kanälen (Consumer / Customer) und mobilen wie stationären Messdaten aus Sensorik (Handel, Weiterverarbeiter, Labore, Lebensmittelkontrolleure, etc.) nahezu in Echtzeit aufbereiten.

Technologisch unterscheidet sich der Ansatz deutlich von herkömmlichen transaktionalen Datenbank-Anwendungen oder Data-Warehouse-Lösungen, da der Schwerpunkt auf der vollständigen Verarbeitung innerhalb einer heterogenen Datenspeicherung (SQL, NoSQL) und der Verarbeitung der Daten im Cloud-Hauptspeicher („in memory“ / „on the fly“) liegen wird.