

Business Intelligence – ein Überblick

Der Autor Markus Ehl berät und unterstützt seit 10 Jahren bei der Entwicklung von Business Intelligence Lösungen und ist Managing Senior Consultant des Beratungs- und IT-Dienstleisters committance AG.

Vorwort und Einführung in BI

Die Verwendung des Begriffs Business Intelligence geht auf den Analysten Howard Dresner der Gartner Group zurück und wurde von ihm Anfang der 90er geprägt.

Business Intelligence bezeichnet dabei weniger ein konkretes Verfahren oder eine bestimmte Methodologie, sondern beschreibt vielmehr ein Füllhorn von Anwendungen und Technologien mit dem Zweck der datenbasierten geschäftlichen Entscheidungsfindung.

Hieraus resultiert die teilweise schwammige Verwendung des Begriffs Business Intelligence in der Praxis. Business Intelligence muss gemäß dieser Verwendung nicht zwingend „intelligente“ Verfahren oder Methoden im Sinne von Methoden der künstlichen Intelligenz bezeichnen. Business Intelligence kann sich demnach je nach Hersteller und Dienstleister genauso gut auf die reine Datenhaltung mit einfachen Auswertungen und Reportings beziehen.

Ein großer Mehrwert aus Unternehmensdaten liegt aber genau in der operativen Verwendung von Analyseergebnissen aus Data Mining Verfahren wie bspw. Neuronale Netze oder Entscheidungsbäume.

Für solche tiefgehenden und weiterführenden Analysen von Daten wird in jüngerer Zeit auch der Begriff Business Analytics verwen-

det, um sich von der allgemeineren Verwendung des Begriffs Business Intelligence abzuheben. Diese Unterscheidung wirkt aber künstlich.

Wir verwenden den Begriff Business Intelligence daher immer im Sinne der kombinierten Anwendung von Data Warehousing zur Datenhaltung, Monitoring und Reporting und Data Mining Verfahren und Techniken zur tiefgehenden Analyse.

Im Folgenden geben wir eine kurze Beschreibung dieser Begriffe.

Data Warehousing

Da die IT-Strukturen innerhalb der Unternehmen historisch gewachsen sind, gibt es oft eine Vielzahl von Systemen, die sich zumeist auf heterogenen Plattformen befinden. Die Datenmenge innerhalb dieser Systeme wächst stetig und wird immer unüberschaubarer. Dies ist umso problematischer, sofern die Daten innerhalb der Systeme völlig getrennt betrachtet werden. Aufgrund der mangelnden Integration kann nicht der optimale Nutzen aus den Daten gezogen werden, welcher sich als Wettbewerbsvorteil am jeweiligen Markt darstellen würde. Um die Konkurrenzfähigkeit sicherzustellen, gehen vor allem größere Unternehmen dazu über, eine übergreifende Sicht auf Daten über möglichst viele Systeme hinweg aufzubauen.

Ein einzelnes System, welches eine themenorientierte, integrierte und zeitbezogene sowie dauerhafte Ansammlung von Unternehmensinformationen gewährleistet, wird *Data Warehouse* genannt.

Data Warehouses haben in der Regel einen sehr großen Speicherbedarf, da sich zumeist bereits in den einzelnen Systemen große Datenmengen befinden und sich die Datenmengen beim Zusammenfassen – zumindest bei semantisch unterschiedlichen Daten – summieren. Das Data Warehouse kann zusätzlich auch mit externen Daten angereichert werden, welche nicht aus operativen Systemen stammen. Dies können beispielsweise Listen mit zusätzlichen Informationen wie etwa Geocodes sein. Des Weiteren erfordert der Aufbau eines Data Warehouses spezielle Prozesse und Methoden, die sich mit der Datenextraktion aus den heterogenen Systemen, der Datenzusammenführung, der Datenhaltung und der Datenauswertung beschäftigen. Die transaktionsorientierte Sicht auf Daten innerhalb operativer Systeme ist zumeist nicht für Analysen geeignet. Aus diesem Grund werden die Daten aller Systeme in einem übergreifenden Modell zusammengefasst, welchem eine spezielle Modellierungsmethode fürs Data Warehousing, die *Dimensionale Modellierung*, zugrunde gelegt werden kann. Werden Geschäftslogiken auf die Daten innerhalb des Data Warehouses angewendet, so kann zur Steigerung des Nutzens auch eine Rückkopplung zu operativen Systemen stattfinden.

Innerhalb eines Data Warehouses kann man mindestens zwischen drei logischen Ebenen unterscheiden: der *ETL-Ebene* (von Extraktion, Transformation und Laden), der *Data Warehouse-Ebene* und der *Data Mart-Ebene*.

Jede dieser Ebenen adressiert einen anderen Gesichtspunkt eines Data Warehouses:

- Die **ETL-Ebene** sammelt alle verfügbaren Daten aus den operativen Systemen.
- Diese Daten werden dann aufbereitet in der **Data Warehouse-Ebene** vorgehalten.
- Für Auswertungen werden in der **Data Mart Ebene** zumeist Daten gemäß bestimmter Geschäftsproblematiken in Data Marts bereitgestellt.

Nach William H. Inmon, einem der Urväter des Data Warehousing, der u.a. auch den Begriff des Data Warehousing geprägt hat, ist ein Data Warehouse wie folgt definiert:

Definition: Data Warehouse (nach William H. Inmon)

„A data warehouse is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management's decision making process.“

In diesem Zusammenhang bedeutet:

„**subject-oriented**“, dass Daten thematisch, unabhängig von der Datenquelle, abgelegt werden, so dass sie einfach für Reportings benutzt werden können. Dies bedeutet wiederum, dass in den meisten Fällen die transaktionsorientierte Sicht der operativen Systeme aufgegeben wird und eine Umstrukturierung der Daten erfolgen muss. „**integrated**“, dass Daten aus verschiedenen Systemen integriert werden. Dies folgt auch aus dem ersten Punkt, da thematisch zusammengehörige Daten auch aus verschiedenen Systemen stammen können.

„**time-variant**“, dass die Daten aus den operationalen Systemen historisiert werden, so dass Analysen über Zeiträume hinweg möglich werden.

„**non-volatile**“, dass Daten innerhalb des Data Warehouses weder von den Endbenutzern verändert noch eingefügt, sondern lediglich abgefragt werden.

Der Sinn und Zweck eines Data Warehouses ist der Datenzugriff auf unternehmensweite Daten mit einer gesicherten Datenbasis. Das Ziel ist idealerweise ein einziges großes Data Warehouse mit allen verfügbaren Daten aus allen Datenquellen.

Ein Data Warehouse dient, wie auch in der Definition von Inmon erkennbar, zur Unterstützung von Analysen und Entscheidungen in dem entsprechenden Geschäftsumfeld des Unternehmens. Somit ergibt sich für die Gruppe der Endbenutzer eines Data Warehouses, dass es sich dabei hauptsächlich um Entscheidungsträger und Business Analysten handelt. Dies ist eine relativ kleine Benutzergruppe, wenn man den unternehmensweiten und geschäftskritischen Aspekt eines Data Warehouses in Betracht zieht. Die Endbenutzer greifen nur lesend auf die Daten zu - allerdings erfordert so gut wie jede Anfrage die Analyse einer großen Datenmenge, die oft über mehrere Tabellen bzw. Joins verteilt ist. Neue Daten werden durch periodische Inserts bzw. Updates per Batch-Lauf in das Data Warehouse übernommen. Dadurch können Daten hinreichend genug historisiert werden, auch wenn Datenänderungen zwischen den Datenübernahmen verloren gehen können. Dies ist der Fall, wenn in den Quellsystemen mehrfache

Änderungen an Daten innerhalb einer Übernahmeperiode vorgenommen werden.



Data Marts

Ein *Data Mart* kann als ein kleines Data Warehouse aufgefasst werden, welches gegenüber einem unternehmensweiten Data Warehouse einen auf die Daten sinnvoll eingeschränkte Zweck hat.

Häufig wird ein Data Mart auf einem Data Warehouse eben mit der eingeschränkten Sicht aufgebaut, bspw. als Marketing Data Mart, kann aber auch eigenständig oder nebenläufig aufgebaut werden. Innerhalb eines Data Marts können Daten für einen be-

stimmten Verwendungszweck nochmals aggregiert vorgehalten werden.

Nutzen eines Data-Mart-Konzeptes

- Anpassungen und Eingrenzungen von Daten für einen speziellen Nutzerkreis z.B. mit Hinblick auf Datenschutz
- Beschleunigen der Geschwindigkeit von allgemeinen Abfragen durch Verringerung des Datenvolumens und der Komplexität
- Unabhängigkeit von Aktualisierungszyklen im Data Warehouse

Analyseverfahren

Für die Auswertung von Daten innerhalb eines Data Warehouses bieten sich neben der generellen Möglichkeit von Anfragen durch die Standard Datenbanksprache SQL (Structured Query Language) oder durch vordefinierte statische Reports weitere spezielle Analyseverfahren an. Diese Verfahren setzen normalerweise auf speziell für diesen Zweck bereitgestellte Data Marts auf. Im Folgenden werden die Verfahren *Data Mining* und *OLAP* (On-Line Analytic Processing) näher beschrieben.

Für beide Verfahren wird eine unterschiedliche Datenorganisation benötigt, so dass beide Verfahren üblicherweise auf unterschiedlichen Data Marts aufsetzen. Beide Verfahren unterscheiden sich generell in ihrer Anwendung:

- Beim **Data Mining** werden Methoden der Statistik sowie der künstlichen Intelligenz (Regression, Neuronales Netz, Entscheidungsbaum, Fuzzy Logic, Support Vector Machines, Self-organizing Maps) verwendet, wobei Zusammenhänge und Muster in den Daten über die jeweilige Methode gefunden werden.

Zum erfolgreichen Einsatz von Data Mining wird ein sehr gutes Verständnis der Statistik und der jeweilig angewendeten Methode beim Benutzer vorausgesetzt insbesondere bei der Aufbereitung der Daten. Des Weiteren findet die Interpretation der Ergebnisse hinsichtlich Qualität und Anwendbarkeit durch den Benutzer nach der eigentlichen Analyse statt.

- Beim **OLAP** werden die zu analysierenden Attribute und die (meist grafische) Präsentationsform vom Benutzer unter dem Gesichtspunkt eigener Annahmen ausgewählt.

Diese Annahmen entsprechen der beim Data Mining nachgelagerten Interpretation der Ergebnisse und werden vom Benutzer normalerweise a priori eingebracht. Die OLAP Analyse bestätigt oder verwirft diese Annahmen und die Ergebnisse können für weiterführende Analysen verwendet werden. Die Verwendung der OLAP Clients ist typischerweise sehr einfach und intuitiv über eine Benutzeroberfläche möglich. Der Benutzer selbst schaut sich vom ihm ausgewählte Datenattribute in einer von ihm ausgewählten Relation zueinander beispielsweise als Plot-Chart oder in der

einfachsten Form als Histogramm an. Die Anwendung von OLAP setzt bei dem Benutzer daher insbesondere ein gutes Verständnis der geschäftlichen Sachverhalte voraus.

Data Mining

Das Ziel des *Data Minings* ist es, anhand verschiedener Analyseverfahren neue, zuvor unbekannte Zusammenhänge und Muster innerhalb großer Datenmengen zu erkennen.

Häufig geht der Einsatz von Data Mining deshalb mit Data Warehouses einher. Beim Data Mining kommen computer-gestützte statistische und künstliche Intelligenz Verfahren wie Regression, Neuronale Netze, Entscheidungsbäume oder Clusteringverfahren wie Self-organizing Maps zum Einsatz. Erst durch den Einsatz dieser Methoden können Computer selbstständig Muster erkennen und Klassifikationen durchführen. Auch wird der Umgang mit großen, schnell wachsenden Datenmengen mit solchen computergestützten Methoden erst ermöglicht.

Data Mining wird häufig als vorhersagende oder prädiktive Analyse verwendet, bei der mit Hilfe historischer Daten Vorhersagen von Ereignissen oder Verhalten in der Zukunft gemacht werden.

Hierunter fallen *Affinitätsvorhersagen* und auch *Assoziationsanalysen* bei denen Zusammenhänge zwischen den untersuchten Datensätzen einer Datenmenge bestimmt werden. Es werden Muster innerhalb der Eingabeattribute und der historisierten Werte

erkannt. Für die Vorhersagen wird vorausgesetzt, dass diese Muster auch in der näheren Zukunft den Sachverhalten beschreiben. Häufig werden Ergebnisse durch Regeln ausgedrückt und beim Database Marketing zur Responseoptimierung oder für Cross Selling eingesetzt. Beispielsweise kann, wie bei Online-Buchhändlern üblich, ein Zusammenhang zwischen Kunden, die Bücher gleichen Typs gekauft haben und deren Interessen für weitere Bücher zu einem hohen Prozentsatz festgestellt werden.

Data Mining wird aber auch zur Segmentierung (Clustering) verwendet. Hierbei werden die Daten in Teilmengen oder Cluster basierend auf einer bestimmten Menge von Attributen und Attributwerten eingeteilt.

Die Cluster können dabei durch statistische Methoden wie z.B. der klassischen ABC-Analyse oder auch k-means gebildet werden. Des Weiteren können Cluster durch Methoden der Künstlichen Intelligenz ermittelt werden wie z.B. durch die auf Neuronalen Netzen basierenden Self-Organizing Maps (SOMs). SOMs bilden die inhärente topologische Struktur der Daten auf Karten ab. Durch die visuelle Darstellung der Karten können Cluster in Anzahl und Größe durch den Benutzer nach der eigentlichen Analyse bestimmt werden, was gegenüber anderen Verfahren große Vorteile bietet.

Mit dem visuellen Verfahren der *selbstorganisierenden Merkmalskarten* bekommt der Anwender einen intuitiven Zugang zur Datenstruktur. Eine SOM visualisiert dabei statistische Eigenschaften einer Datenmenge als intuitiv interpretierbare Berg- und Tal-Landschaft.

Die Nutzung von SOMs setzt kein mathematisches Vorwissen voraus. Der User richtet seine Aufmerksamkeit ausschließlich auf die leicht erkennbaren visuellen Muster. Die visuellen Methoden der Datenanalyse nutzen dabei die hoch spezialisierte menschliche Fähigkeit zur visuellen Strukturerkennung und -differenzierung.

Zur Anwendung der meisten Data Mining Verfahren müssen die zugrunde liegenden Daten in eine flache, denormalisierte Form gebracht werden. Alle Attribute müssen dabei in Abhängigkeit des zu analysierenden Attributs angeordnet sein. Soll beispielsweise das Kundenverhalten analysiert werden, so werden alle Attribute, wie etwa Stammdaten und Rechnungsdaten der letzten sechs Monate, der Kundennummer des jeweiligen Kunden zugeordnet.

OLAP (On-Line Analytic Processing)

Unter *OLAP* versteht man die allgemeine Aktivität des Abfragens und Präsentierens von Daten unter Berücksichtigung bestimmter Geschäftsdimensionen. Die Präsentation der Daten erfolgt dabei zumeist in graphischer Form von Diagrammen (z.B. Balken-, Torten-, Linien-, Punktdiagramme) aber auch in tabellarischer Form (z.B. Pivot-Tabellen).

Normalerweise werden für OLAP sogenannte OLAP-Würfel oder Data Marts bereitgestellt, die dimensionaler Natur sind und sich auf einen speziellen Geschäftsaspekt beziehen. Dadurch kann die Datenmenge eingeschränkt und somit die Abfragegeschwindigkeit erhöht werden. Dies ist ein wichtiger Punkt, da bei den Anfragen jeweils

große Datenmengen verarbeitet werden müssen und Anfragen entsprechend lange laufen können. Für die Akzeptanz von OLAP bei den Endbenutzern spielt die Verarbeitungsgeschwindigkeit eine wichtige Rolle. Wartezeiten über mehreren Minuten für eine Anfrage werden vor allem bei der Benutzergruppe der Entscheidungsträger nicht akzeptiert, zumal OLAP Analysen normalerweise iterativ durchgeführt werden.

Im Allgemeinen werden beim OLAP spezielle OLAP-Clients verwendet, die auf den Umgang mit dimensionalen Modellen ausgelegt sind. Manche Clients bieten neben der zusammengefassten tabellarischen oder graphischen Präsentation von Daten auch die Möglichkeit, auf die dahinter liegenden Daten durchzugreifen – dies wird als *reach-through* oder *pass-through* bezeichnet. Dadurch wird ermöglicht, dass man sich die einzelnen Datensätze, die auf die entsprechende Abfrage zutreffen, detailliert anschauen kann.

committance AG

Ein auf Enterprise Information Management spezialisiertes Beratungsunternehmen und IT-Dienstleister mit Kernkompetenzen in analytischem CRM und Business Intelligence Systemen wie Data Warehousing, Data Mining und Kampagnenmanagement sowie Portaltechnologie. Darüber hinaus bietet die committance AG operative Unterstützung und technischen Support an.

Weitere Informationen finden Sie unter www.committance.com.