**3.1 Einführung in Data Warehousing und Business Intelligence (1/3)**

Alle Inhalte dieser Lerneinheit unterliegen dem Copyright des Autors © Westenberger / TH Köln  
Weitergabe der Inhalte oder Veröffentlichung in jedweder Form ist untersagt.

Das Lernziel dieses Lernraumes haben Sie erreicht, wenn Sie in der Lage sind, die Informationsbedarfe von Stakeholdern einer Diskurswelt zu analysieren und zu strukturieren, ihre relevanten Anforderungen an die Bereitstellung von Informations-, bzw Analysediensten zu identifizieren und eine geeignete Infrastruktur zu entwerfen, die diese Anforderungen erfüllt.  
Dies wird erreicht, indem Sie  
- verstehen, wie Unternehmensziele operationalisiert werden, eine Klassifizierung von tyischen Berichtsgegenständen kennen und anwenden lernen,   
- Datenstrukturen entwerfen können, die für Berichtszwecke und Analysezwecke besonders geeignet sind  
- die wichtigsten Aspekte der Transformation und Aufbereitung von operativen Daten verstehen und  
- typische Referenzarchitekturen für Data Warehousing und Business Intelligence kennen

Am Ende des Lernraumes 2 „Data Warehousing und Business Intelligence“ sollten folgende Kompetenzen erworben worden sein:  
1. Sie können Ereignisse, Zustände und Prozesse in einer Fachdomäne identifizieren und verstehen deren Bedeutung für die Ableitung von Berichtsanforderungen.  
2. Sie können zwischen Messwerten und Messkontext unterscheiden, dieses Wissen auf Fachdomänen anwenden und passende Dimensionsstrukturen entwerfen.  
3. Sie kennen die Faktentabellentypen und können diese passend für Szenarien auswählen.  
4. Sie können STAR-Schemata passend zu den Faktentabellentyen entwerfen und den Entwurf begründen.  
5. Sie kennen die klassische Referenzarchitektur einer BI-Infrastruktur und können die Unterschiede der Modelle nach Kimball und Inmon erläutern.  
6. Sie können den ETL-Prozess erläutern und passende Profilingkennzahlen für Quelldaten benennen und exemplarisch deren Berechnung beschreiben.

# 3.1 Einführung in Data Warehousing und Business Intelligence (2/3)

Der Begriff Intelligence im zusammengesetzten Term "Business Intelligence (BI)" sollte im Sinne von Informationsbeschaffung übersetzt werden. In der Literatur findet sich eine Vielzahl von unterschiedlichen BI-Definitionen [[1]](https://ilu.th-koeln.de/ilias.php?baseClass=illmpresentationgui&cmd=layout&ref_id=467010&obj_id=52439#fn1). Einige davon sehen BI als eine Menge von Techniken und Technologien, um Daten für die Entscheidungsunterstützung für das Management nutzbar zu machen. Andere betrachten BI als einen Prozess zu demselben Zweck. Wir werden in diesem Lernmodul BI als eine Disziplin verstehen, die sich mit der Nutzbarmachung von Daten zur Entscheidungsunterstützung des Managements in Unternehmen, bzw Organisationen sowie der Informationsversorgung weiterer Stakeholder auseinandersetzt und dazu einsetzbare Technologien, Prozesse, Werkzeuge, Infrastrukturen sowie Organisationsstrukturen betrachtet.  
Eine für BI-Zwecke spezialisierte relationale Datenhaltung wird im Allgemeinen als Data Warehouse bezeichnet. Unter Data Warehousing werden die Datenbereitstellungsprozesse für ein Data Warehouse verstanden. Eine für Analyse-Zwecke spezialisierte Bereitstellung von Rohdaten, die auch nicht-relationale Strukturen einschließt, fasst man häufig unter dem Begriff Data Lake zusammen.

Die klassische und häufigste BI-Aufgabe besteht in der Bereitstellung von Berichten über bestimmte Aspekte eines Systems. Da jede Informationsbereitstellung einen Aufwand bedeutet, wird damit ein Zweck verfolgt, zB einer behördlichen Auflage entsprechend oder wirtschaftlichen Gesichtspunkten dienend. Der Berichtsgegenstand kann sowohl reale als auch planerische Aspekte eine Systems betreffen. Solche Aspekte sind beispielsweise  
  
Ereignisse - Ereignisse finden zu einem Zeitpunkt statt und haben im Prinzip keine Dauer (zumindest aus makroskopischer Sicht). Da der Fokus des klassischen Berichtswesens nicht auf einzelnen Ereignissen liegt, werden Ereignisse in der Regel für Berichtszwecke über einen Zeitraum aggregiert. Ereignisse ändern in der Regel Systemzustände. zB ist ein Fahrraddiebstahl ein Ereignis, das den Zustand eines Fahrrades von "nutzbar" in "gestohlen" verändert. Die Warenentnahme eines Materials aus einem Lager ist ein Ereignis, das den Systemzustand des Lagers "Bestand an Material x" reduziert. In mikroskopischer Betrachtung stellt sich die Warenentnahme zwar als Aktivität dar, die eine bestimmte Dauer hat, aber wenn diese Granularität für die Berichtszwecke nicht relevant und damit vernachlässigbar ist, würde man das Ereignis "Beginn" oder das Ereignis "Ende der Warenentnahme" als das repräsentative Ereignis für die Warenentnahme ansehen und den entsprechenden Zeitstempel nehmen. Messbar an einem solchen Ereignis wären die Menge und der Wert der entnommenen Ware. Zusätzlich wären nach messbar die Zustände vor und nach dem Ereignis, hier also die Bestände vor uns nach der Entnahme (als Menge und als Wert). Zustandsmessungen widmen wir eine eigene Messklassifikation (s. nachfolgenden Topic).  
  
Systemzustände - Der Zustand eines Systems wird häufig durch Messgrößen charakterisiert. Zustandsgrößen wie Temperatur oder eine Lagerbestand werden in der Regel für einen bestimmten Zeitpunkt gemessen. Will man eine Zustandsgröße für einen Berichtszeitraum angeben, wird in der Regel ein Referenzzeitpunkt für eine Messung in diesem Intervall angegeben, oder -wenn mehrere Messungen in dem Zeitraum gemacht wurden- Funktionen wie Minimum, Maximum, Mittelwert auf die Messwertreihe angewendet, oder es wird der letzte Wert genommen. Häufig bleibt der Zustand eines Systems über einen bestimmten Zeitraum konstant, bevor er durch den Eintritt eines Ereignisses verändert wird. Dann ist auch die Zustandsdauer etwas, was man messen kann. So bleibt zum Beispiel der Bestand eines Lagers solange konstant, bis er durch einen Lagerzugang oder -abgang verändert wird. Der Zustand eines ausgeliehenen Buches der Bibliothek ändert sich durch dessen Rückgabe. Wie lange ein Lagerbestand einen bestimmten Wert angenommen hat, ist selten von Interesse, wie lange Bücher ausgeliehen werden, allerdings schon eher.  
Neben zahlenmäßig messbaren System- oder Objektattribute werden auch Strukturen, Zusammenhänge unter Zuständen subsummiert. Die Messung besteht hier in der Feststellung der Existenz einer Beziehung "Mitarbeiter Müller gehört zur Abteilung Produktion". Hier wird in der Regel ebenfalls ein Bezugszeitpunkt gewählt.  
Auch ein Vorgang wie die Aktivität eines Akteurs in einem System kann wie ein Zustand betrachtet werden.  
Hinweis: Eine Zustandsmessung zu einem Zeitpunkt, an dem sich gerade der Zustand ändert, ist problematisch. Hier kann man allenfalls den Zustände vor, bzw nach Eintreten eines Ereignisses messen (s. oben Warenentnahme).  
  
(Diskrete) Prozesse: Menge von Ereignissen und Aktivitäten, die einen gerichteten, zusammenhängenden Graph mit einem Startknoten und Endknoten bilden. Ein Prozess dient in der Regel zur Erreichung eines Ergebnisses. Seine Ereignisse und Aktivitäten ändern Zustände in dem System, in dem der Prozess abläuft. Auch einem Prozess selber können Zustände zugeordnet werden.   
o Dauer von Aktivitäten und Zuständen, bzw Prozessen (Zeitbezug mit Referenzpunkt, zb dem Start- oder Endzeitpunkt eines Prozesses oder eines Zustandes)  
o Da Prozesse als Container angesehen werden können, die Ereignisse und Aktivitäten, bzw Zustände umfassen, können die Messzahlen aus den Prozesselementen ggf auf der Prozessebene aggregiert werden, man siehe zB Prozesskosten oder Prozessdauern.  
  
Hinweise:

* Manchmal bedarf es bzgl der Abgrenzung von Ereignissen und Zuständen, bzw Beständen auf der Basis natürlicher Sprache einer Klärung: "Die Anzahl gestohlener Fahrräder" könnte als die "Anzahl der Fahrraddiebstähle in einem Zeitraum", also als Ereigniskennzahl gemeint sein. Aber auch eine Interpretation als Zustandskennzahl "Bestand an Fahrrädern an einem Stichzeitpunkt mit dem derzeitigen Zustand gestohlen" ist zulässig. In solchen Fällen ist eine sprachliche Präzisierung im Zuge der Anforderungsermittlung nötig.
* Jede Buchung im Rahmen der Finanzbuchhaltung, zB für Aufwände oder Erlöse kann als ein Ereignis angesehen werden, das einen Sachkontostand verändert. Wenn mehrere solcher Buchungen in einem sächlichen Zusammenhang stehen, zB im Rahmen eines Prozesses erfolgen, kann es sinnvoll sein, diese Messungen auf der Prozessebene zu verdichten (s.o. Prozesskosten).
* Wenn man über die Dauern von Zuständen oder Prozessen eine Kennzahl wie zB "mittlere Dauer der Auftragsabwicklung" in einem Berichtszeit ermitteln will, ergibt sich das Problem, dass Zustände möglicherweise über die Intervallgrenzen hinweg andauern. Je nach Kontext kann es sinnvoll sein, Dauern dann dem Anfangs- oder Endzeitpunkt des Zustandes oder Prozesses zuzuordnen (s.o. Prozesse) oder auch den Zustand an den Intervallgrenzen zu splitten. Beispiele:
  + Ist nach der mittleren Ausleihdauer von Büchern in einem speziellen Jahr gefragt, ist die Zuordnung der Dauer einer Ausleihe zum Beginn oder Ende einer Ausleihe sicher vernünftig
  + Ist nach der Gesamtausleihzeit von Büchern in einem speziellen Jahr gefragt, zB um den Prozentsatz zu ermitteln, zu dem ein Buch oder eine Buchtitel ausgeliehen war, ist eine Aufsplittung der Ausleihen an Jahresgrenzen sinnvoll.

Aus den Messungen werden Maßzahlen, in dem die Messungen auf einem Ensemble von Ereignissen oder Prozessen durchführt und aus den Messwerten für das Ensemble aussagekräftige Kennzahlen ableitet. Dies kann im Zählen von Ereignissen, dem Aufsummieren von Messgrößen, wie zB Umsatzbeträge sein. Auch der Mittelwert ist eine solche charakterisierende Kennzahl.  Zählen, Summen-, Differenzen- sowie Mittelwertbildung wie auch die Anwendung von Min- oder Max-Funktionen auf Messungen an Ereignissen, Zuständen oder Prozessen führen auf Absolutkennzahlen. Beispiele sind die Anzahl von Studienabschlüssen in einem Zeitraum, die Anzahl von Lehrenden an einem Zeitpunkt und die durchschnittliche Studiendauer. Werden zwei oder mehre Absolutkennzahlen zueinander in Relation gesetzt, entsteht eine Relativkennzahl oder Verhältniskennzahl. Das können zwei unabhängige Absolutkennzahlen sein, die zueinander in Beziehung gesetzt werden, zB die Zahl der Absolventen mit Zahl der Lehrenden. Daraus ergibt sich die Verhältniszahl "Zahl der Absolventen pro Lehrenden", oder Absolutkennzahlen, die in einer Abhängigkeit zueinander stehen, zB die Anzahl Frauen unter den Studierenden. Absolutkennzahlen besitzen im Gegensatz zu Relativkennzahlen dieselbe Maßeinheit wie die Messwerte, auf denen sie basieren.  Um diese Kennzahlen aussagekräftig zu machen, muss die Grundgesamtheit, auf der gemessen wird, klar definiert werden. Die Anwendung dieses Kennzahlenkonstruktes wäre beispielsweise die Maßzahl "Anzahl der Studienabschlüsse im Studiengang x im Studienjahr y". Die Metrik, also was gemessen wird, ist hier die Zählung von Studienabschlüssen. Das Ensemble, also die Grundgesamtheit, auf der gemessen wird, wird durch die Angaben  "... im Studiengang x im Studienjahr y" definiert.  Entsprechende kommt bei der Relativzahl  "Die Anzahl der Studienabschlüsse im Studiengang x im Studienjahr y dividiert durch die Anzahl der Lehrenden in diesem Studiengang am Stichtag d" eine weitere Absolutzahl ins Spiel, die sich aus der Metrik "Zählung der Lehrenden" auf der Grundgesamtheit "im Studiengang x zum Zeitpunkt d" festgelegt ist. Die zu bemessende Grundgesamtheit wird häufig partitioniert, sodass die Kennzahlen als Aufriss dargestellt werden können.  Ein Beispiel wäre die Partitionierung eines Zeitraumes in Jahre, zB  "Die Anzahl der jährlichen Studienabschlüsse im Studiengang x" wird im Aufriss der Studienjahre 2010 bis 2020 dargestellt. Damit wird ausgedrückt, dass die Grundgesamtheit "Studiengang x und Studienjahre 2010 bis 2020 " in jährliche Partitionen unterteilt wird, in denen jeweils separat die Maßzahl ermittelt wird.