



Universität Karlsruhe
Fakultät für Informatik
Institut für Programmstrukturen
und Datenorganisation (IPD)



Konzeption und Umsetzung eines Data Warehouse zur Erstellung von Ganglinien und zur Bewertung von Prognosen im Verkehr

STUDIENARBEIT

von

Ling Kong

im Oktober 2004

Verantwortlicher Betreuer:
Betreuender Mitarbeiter:

Prof. Dr.-Ing. Dr. h.c. Peter C. Lockemann
Dipl.-Inform. Heiko Schepperle

Dieser Arbeit wurde von mir selbständig angefertigt. Alle verwendeten Literaturstellen sind im Literaturverzeichnis aufgeführt; eine Verwendung anderer Hilfsmittel erfolgte nicht.

Ich versichere dies ausdrücklich mit nachstehender Unterschrift:

Ling Kong

Karlsruhe, 30.10.2004

Kurzfassung

Diese Arbeit beschreibt ein Data Warehouse Schema zur Erstellung von Ganglinien und zur Bewertung von Prognosen im Verkehr. Mit diesem Schema können die Daten im Verkehr wie z.B. Simulationsdaten, Prognosewerten, erwartete Daten in Form von Ganglinien erstellt und gespeichert werden. Durch Gangliniendarstellungen für verschiedene Arten von Daten im Verkehr können diese Daten intuitiv mit einander verglichen und dann die Bewertung von Verkehrsprognosen unterstützt werden. Das Data Warehouse Schema wird auf die konzeptuelle Ebene mit durch ein mE/R-Modell dargestellt und erläutert. Auf die physische Modellierungsebene wird ein physisches Data Warehouse Schema in Form vom Star Schema entworfen und mit relationaler Datenbank implementiert. Basierend auf das physische Schema werden Ganglinien-Abfragen und Darstellungen mit OLAP-System Cognos Series 7 umgesetzt.

Inhalt

1. Einleitung.....	1
1.1 Projekt OVID	1
1.2 Zielsetzung	1
1.3 Aufbau der Arbeit.....	2
2. Analyse von Daten im Verkehr.....	3
2.1 Simulationsdaten	3
2.2 Floating Car Daten	4
2.3 Güterverkehrsdaten	4
2.4 Zusammenfassung.....	5
3. Verkehrsprognose.....	7
3.1 Allgemeine Grundlagen für Prognose.....	7
3.2 Ein generelles Prognosemodell	8
3.3 Kurzzeitprognose im Verkehr	9
3.3.1 Kurzzeitprognoseart	10
3.3.2 Prognosemodell am Querschnitt	10
3.4 Zusammenfassung.....	11
4. Data Warehouse und OLAP	13
4.1 Data Warehouse	13
4.1.1 Multidimensionale Datenmodelle	14
4.1.2 Snowflake Schema und Star Schema.....	14
4.2 OLAP	15
4.2.1 ROLAP, MOLAP, DOLAP und HOLAP	17
4.2.2 Standardfunktionen von OLAP	18
4.3 Zusammenfassung.....	18
5. Konzeptuelle Modellierung.....	19
5.1 Ganglinien im Verkehr.....	19
5.2 Konzeptuelles Datenmodell	20
5.2.1 Dimension Ort.....	21
5.2.2 Dimension Simulationslauf.....	21
5.2.3 Dimension Zeitreihe	22
5.2.4 Dimension Verfahren	22
5.3 Zusammenfassung.....	22
6. Umsetzung des konzeptuellen Datenmodells.....	23

6.1	Physische Data Warehouse Schema.....	23
6.2	Dimensionen.....	24
6.3	Faktentabelle	25
6.4	Datenquelle für Simulationsdaten	25
6.5	Zeitreihendarstellung.....	26
6.6	Daten Integration.....	28
6.7	Gangliniendarstellung	28
6.8	Zusammenfassung	30
7.	OLAP Einsatz.....	31
7.1	Einführung Cognos Series 7.....	31
7.1.1	Cognos Architect.....	32
7.1.2	Cognos Transformer Edition	33
7.1.3	Cognos PowerPlay	33
7.2	Implementierung mit Cognos PowerPlay	34
7.2.1	Erstellung Cognos Architect Modell.....	34
7.2.2	Erstellung PowerCube mit Cognos Transformer Edition	34
7.2.3	Ganglinien-Abfragen mit Cognos PowerPlay.....	35
7.3	Zusammenfassung	35
8.	Zusammenfassung	37

1. Einleitung

In diesem Kapitel wird zuerst die Motivation des grundlegenden Projekts OVID kurz erläutert. Die Zielsetzung dieser Arbeit wird anschließend vorgestellt. Eine Übersicht über den Aufbau der Arbeit wird zum Schluss gegeben.

1.1 Projekt OVID

Belastungen der Straßen werden ein immer größeres Problem in vielen Ländern. Alltäglicher Stau im Verkehr ist ein Zeichen des Problems. Dies ist auch als ökologische und ökonomische Konsequenzen bemerkbar. Eine Lösung zu diesem Problem ist der Einsatz intelligent und vernetzt arbeitender Routenplanungssysteme. Mit einem solchen System kann allumfassende Verfügbarkeit von Information für den Verkehrsteilnehmer sowohl im Personen- als auch im Güterverkehr gewährleistet werden. Diese Tendenz gegenüber ist das Projekt OVID ins Leben gerufen worden. Projekt OVID ist ein vom Bundesministerium für Bildung und Forschung gefördertes Projekt. OVID steht für selbst Organisationsfähigkeit im Verkehr durch I+K-gestützte Dienste. Ein Teilprojekt des Projekts OVID ist Teilprojekt B1 „Verlässliche Datenbanken für die Informationsbereitstellung im Verkehr“. Im Zusammenhang mit diesem Teilprojekt entsteht die vorliegende Arbeit.

Das Ziel des Projekts OVID ist der Aufbau und die Nutzung einer Plattform zur Modellierung und Bewertung von verkehrsinfrastrukturellen, verkehrstelematischen und logistischen Maßnahmen im Verkehrs- und sozi-ökonomischen System (vgl.[OVID]). An diesem Projekt sind neben dem Institut für Programmstrukturen und Datenorganisation (IPD) noch das Institut für Verkehrswesen (IFV), das Institut für Fördertechnik und Logistiksysteme (IFL) und das Institut für Wirtschaftspolitik und Wirtschaftsforschung (IWW) der Universität Karlsruhe sowie das Institut für Informations- und Datenverarbeitung der Fraunhofergesellschaft (IITB) beteiligt. Als Industriepartner sind die PTV AG und die LOCOM Consulting GmbH beteiligt.

1.2 Zielsetzung

Im Rahmen des Projekts OVID wird die Wirkung von Diensten auf den Verkehr simuliert. Diese Dienste erstellen unter anderem eine Prognose über zukünftige Verkehrszustände. Vor jedem Simulationslauf müssen die Daten bestimmt werden, auf die sich die Prognose während der Simulation stützen kann. Diese „erwarteten Daten“ werden aus den in vorigen Läufen gemessenen Daten erzeugt. Während der Simulation werden die „erwarteten“ Daten laufend angepasst und somit „prognostiziert“ Daten erzeugt, mit deren Hilfe die zukünftige Verkehrssituation vorhergesagt werden kann. Die Qualität der Prognose kann anschließend durch Vergleich der prognostizierten Daten mit den tatsächlich gemessenen Daten bewertet werden. Für die Erzeugung und Datenhaltung von sich zeitlich verändernden Daten, die als Ganglinien bezeichnet sind, wird ein Datenmodell auf Basis von Data Warehouse

Technologie konzipiert. Die Bewertung der durchgeführten Prognose soll dadurch ebenfalls unterstützt werden.

Anhand dieser Zielsetzung besteht diese Arbeit im wesentlichen aus den folgenden Aufgaben:

- Analyse von Daten im Verkehr
- Analyse von Verkehrsprognosen
- Analyse der Grundlage für Data Warehouse und OLAP (online Analytical Processing)
- Konzeption eines Data Warehouse Schemas (auf konzeptuelle Ebene und physische Modellierungsebene)
- Umsetzung der Konzeption mit Oracle und OLAP-Werkzeuge Cognos Serie 7

1.3 Aufbau der Arbeit

Diese Arbeit ist in 8 Kapiteln gegliedert. Das vorliegende Kapitel dient als eine kurze Einleitung für die gesamte Arbeit.

In Kapitel 2 werden die Grundlagen für Verkehrsdaten erläutert.

Um die Prognosen im Verkehr und deren Bewertung gut zu verstehen, werden die Grundlagen für Verkehrsprognose untersucht und in Kapitel 3 vorgestellt.

Da Data Warehouse Technologie als die technische Basis dieser Arbeit verwendet, wird Kapitel 4 auf die Analyse von Data Warehouse Technologie und Untersuchung von OLAP-Systemen eingehen.

In Kapitel 5 wird zuerst das Konzept Ganglinien im Verkehr mit einem Beispiel skizziert. Dann wird das konzeptuelle Datenmodell für die Darstellung von Ganglinien in mE/R-Modell vorgestellt.

Auf Basis des konzeptuellen Datenmodells wird ein physisches Data Warehouse Schema für Erstellung von Ganglinien entwickelt. In Kapitel 6 wird dieses Schema und dessen Umsetzung erläutert.

Basierend auf das physische Schema werden Ganglinien-Abfragen und Darstellungen mit OLAP-Systemen erstellt und damit die Bewertung der in Ganglinien dargestellten Prognosen unterstützt. In Kapitel 7 wird der Einsatz von OLAP-System Cognos Series 7 erklärt.

Kapitel 8 schließt die Arbeit mit einer Zusammenfassung und einer Evaluation der Ergebnisse dieser Arbeit.

2. Analyse von Daten im Verkehr

In diesem Kapitel wird untersucht, wie die Verkehrsdaten aussehen und welche Besonderheiten die Verkehrsdaten besitzen. Die Daten im Verkehrsumfeld können in folgende Arten untergeteilt werden (siehe [Koop04]).

- Messdaten: Diese Daten sind unter anderem durch Sensor erhoben oder durch statistische Verfahren aus den Vergangenheitsdaten erfasst. Messdaten beschreiben die Verkehrssituation und werden benutzt, um weitere Verkehrsinformationen abzuleiten.
- Infrastrukturdaten: Diese Daten sind die Basisinformation über die Verkehrsnetze, wie z.B. Straßenlänge, Straßenkategorie usw.
- Ereignisdaten und Störungsdaten: Diese Daten sind von menschlichen beobachtet und aufgenommen. Solche Daten beschreiben verschiedene Arten von Ereignissen, von Wetter bis zu Staumeldungen.

Infrastrukturdaten enthalten die Fundamentaldaten zum Verkehrsnetz, auf dem Straßenverkehr geplant werden kann. Ereignisdaten und Störung beinhalten viele so genannte Imperfektdaten, die unsicher, unscharfe und ungenau sind. Bei einem Stau kann z.B. von einen als sehr schlimm aber von anderen als nicht so schlimm angemeldet. In dieser Arbeit werden hauptsächlich die Messdaten verwendet. Die anderen Arten von Verkehrsdaten werden in dieser Arbeit wenig berücksichtigt.

Mit verschiedenen Messverfahren kann man die Messdaten erfassen. In der Literatur werden demgemäß die Messdaten nach Messverfahren untergliedert. Im folgenden werden drei Arten von Messdaten und zugehörigen Messverfahren vorgestellt, d.h. Simulationsdaten, Floating Car Daten und Güterverkehrsdaten.

2.1 Simulationsdaten

Simulationsdaten beziehen sich in dieser Arbeit auf die Daten, die vom Programm VISSIM erzeugt werden. VISSIM ist ein Produkt der Firma PTV AG. VISSIM ist ein mikroskopisches, verhaltensbasiertes und universell einsetzbares Simulationstool. Im Verkehrsbereich eignet sich VISSIM für die Modellierung von inner- und außerstädtischen Gebieten. Neue Version von VISSIM berücksichtigt sich auf zunehmende Anwendung in der Verkehrsplanung. In einem Editor können einzelne Fahrspuren, Ampeln, Messstation mit bestimmten Eigenschaften zu einem komplexen Netz zusammengefügt, und verschiedene Fahrzeugtypen mit bestimmten Eigenschaften wie Länge, Wunschbeschleunigungen, befahrbare Spuren usw. definiert werden. Die Simulationsdaten von VISSIM lassen sich an bestimmte Messstation einer Fahrspur anlegen und erfassen von den vorüberfahrenden Fahrzeugen.

Die simulierten Daten umfassen alle Verkehrsinformationen im Detail (vgl. [Merk04]):

- Die eindeutige Fahrzeugnummer des simulierten Fahrzeugs

- Der Typ des Fahrzeuges
- Die Zeitpunkte des Einfahren in die und Ausfahren aus der Messstation
- Die Geschwindigkeit und die Beschleunigung des Fahrzeuges
- Die Anzahl der Personen, die sich im Fahrzeug befinden
- Die Zeit, die das Fahrzeug während der Simulation schon gestanden hat
- Die Länge des Fahrzeugs

Diese Simulationsdaten können mit dem Datenbankschema, das im [Sand04] fertiggestellt ist, gespeichert werden. In dieser Arbeit werden diese Daten als Datenquelle verwendet.

2.2 Floating Car Daten

Im Vergleich zu Simulationsdaten entstehen Floating Car Daten (FCD) aus den Messungen von wirklich existierenden Verkehrsgeschehen. Durch Sensoren im FCD-Fahrzeuge werden Daten gesammelt und an einem Dienstzentrum übermittelt. Anschließend werden diese Daten als aufbereitete Verkehrslageinformation an beliebige Verkehrsteilnehmer zurückgibt. Sie können dann als Grundlage für Navigation und Routenplanung dienen (siehe [Merk04]).

Bei FCD wurde Verkehrsnetz zuerst in Linken geteilt, jede Link besitzt einen Start- und Endknoten. Eine Link stellt daher eine Verbindungsstrecke zwischen die Start- und Endknoten dar. Ein Fahrzeug beginnt seine Messung bei Erreichen des Startknotens und übermittelt die gemessenen Daten bei Erreichen des Endknotens. Pro Link werden folgende Daten erfasst und übermittelt (vgl. [Merk04]):

- Der Typ des Fahrzeugs
- Die Einfahrt- und Ausfahrtzeitpunkte
- Die Straßenkategorie
- Die Linknummer
- Die durchschnittliche Geschwindigkeit
- Die Standardabweichung von diesem Durchschnitt
- Der prozentuale Stillstandsanteil
- Die minimale und die Maximale Geschwindigkeit
- Die durchschnittliche positive und negative Beschleunigung

2.3 Güterverkehrsdaten

Güterverkehrsdaten stellen die Daten im Güterverkehrsbereich dar. Im Rahmen des Projekts OVID werden die Güterverkehrsdaten über alle Güterverkehrstransporte verwendet, die in den Berichtsjahren 1999, 2001 und 2002 von in Deutschland gemeldeten Fahrzeugen durchgeführt wurden ([KBB01]). Das verwendete Messverfahren ist Stichprobe. Die gesammelten Ergebnisse wurden aggregiert und anschließend mit Hochrechnungsfaktoren multipliziert, damit die Aussage für gesamte Transportaufkommen treffen können. Die

Verkehrssituation kann durch unten genannte Aggregation und Gruppierung von gesammelter Information erhoben ([Merk04]).

Zur Gruppierung wurden verwendet:

- Das betrachtete Jahr
- Die Art des Fahrzeuges und des Anhängers
- Der Be- und Entladenort
- Der genutzte Rauminhalt
- Die Art der transportierten Güter
- Die Form der Ladung
- Die Größe der Container
- Die Art der Fahrt
- Das Verkehrsmittel, auf das die Fracht anschließend verladen wurde

Zur Aggregation wurden verwendet:

- Die zurückgelegte Entfernung
- Die Anzahl der zurückgelegten Einzelstrecken
- Das transportierte Gesamtgewicht
- Das zulässige Gesamtgewicht
- Die Nutzlast

2.4 Zusammenfassung

Verkehrsdaten mit verschiedenen Eigenschaften werden in drei Arten aufgeteilt: Messdaten, Infrastrukturdaten und Ereignisdaten. Die drei Arten von Daten können zusammen Verkehrssituation beschreiben und umfassen. Auf die Messdaten wurde näher eingegangen. In diesem Kapitel wurden drei Arten von Messdaten vorgestellt. Die eine ist die Simulationsdaten, die durch das Programm VISSIM erzeugt werden. Diese Daten enthalten die Verkehrsinformation von simulierter Messstation. Eine andere ist Floating Car Daten, die durch Sensor in Fahrzeugen gesammelt werden. Floating Car Daten beschreiben die wirklich existierende Verkehrssituation von Straßenlinken. Güterverkehrsdaten handeln sich um bereits aggregierte und hochgerechnete Werte. Dieser Arbeit berücksichtigt sich vorwiegend auf die Simulationsdaten.

3. Verkehrsprognose

Die Grundlagen für Verkehrsprognose werden in diesem Kapitel untersucht und vorgestellt. Zuerst wird auf die allgemeinen Grundlagen für Prognose eingegangen. Danach wird ein generelles Prognosemodell auf Basis von Ganglinien skizziert. Kurzzeitprognose am Querschnitt ist das zu diskutierende Verfahren im Rahmen des Projekts OVID und wird deshalb anschließend kurz vorgestellt.

3.1 Allgemeine Grundlagen für Prognose

Prognosen spielen eine wichtige Rolle in vielen Bereichen. Um die Zukunft zu antizipieren, sich auf eintreffende Situationen vorzubereiten oder prognostizierte Ereignisse zu beeinflussen, werden Prognosen auf Basis der vergangenen Informationen durchgeführt. Informationen der Vergangenheit können sowohl in Form von persönlichen Erfahrungen, also implizit verschlüsselt, aber als auch explizit in Form von Vergangenheitsdaten vorhanden sein. Bei Prognosen werden bestimmte empirische, mathematische oder statistische Methode bzw. Algorithmen benötigt. Diese werden zusammen als Prognoseverfahren bezeichnet. Prognoseverfahren werden normalerweise in eine qualitative und eine quantitative Gruppe klassifiziert. Qualitative Verfahren werden häufig in den Bereichen verwendet, in denen die kausalen Modelle häufig fehlen oder zu ungenau sind. Bei den quantitativen Verfahren werden häufig die kausalen Modelle und reinen Zeitreihenmodellen gegliedert. Kausale Modelle stellen Zusammenhänge zwischen der zu prognostizierenden Größe und anderen unabhängigen Variablen her. Demgegenüber basiert ein reines Zeitreihenmodell ausschließlich auf historischen Werteverläufen der zu prognostizierten Größe. Für Prognose im Verkehr finden reine Zeitreihenmodelle häufig Verwendung. Zur Einstellung der unterschiedlichen Prognosemodelle werden viele unterschiedliche Analysedaten bzw. unterschiedlich lange Analysezeiträume benötigt. In der Praxis wird ein quantitatives Prognoseverfahren häufig nach dem folgenden Prozess entwickelt (vgl. [Wild96]):

- Zuerst wird ein Prognosemodell ausgewählt auf eine Analyse der vorliegenden historischen Daten eines Analysezeitraums
- Im zweiten Schritt wird das Prognosemodell mit den historischen Daten des Analysezeitraums verwendet
- Die Prognoseergebnisse werden dann analysiert und ausgewertet, um die Parameter des Prognosemodells optimal einzustellen.

In dieser Arbeit wird ein Datenmodell auf Basis von Data Warehouse Technologie entwickelt, damit der obengenannte Prozess in Verkehrsbereichen unterstützt werden können.

3.2 Ein generelles Prognosemodell

In diesem Abschnitt wird ein allgemeines Prognosemodell auf Basis von Ganglinien vorgestellt. Die Verkehrsstärken eines Verkehrsquerschnitts (häufig durch Fahrzeugmenge dargestellt) werden nach jedem Messintervall gemessen und gespeichert. So erhält man eine Zeitreihe, deren graphische Darstellung als Ganglinie bezeichnet. In der Literatur wurden bereits viele Diskussionen über Verkehrsprognosen auf Basis von Ganglinien gegeben. Eine Einführung kann man bei [Wild96] finden. Hier wird als Beispiel ein generelles Prognosemodell auf Basis von Ganglinien vorgestellt. Abbildung 3.1 stellt dieses generelle Prognosemodell dar.

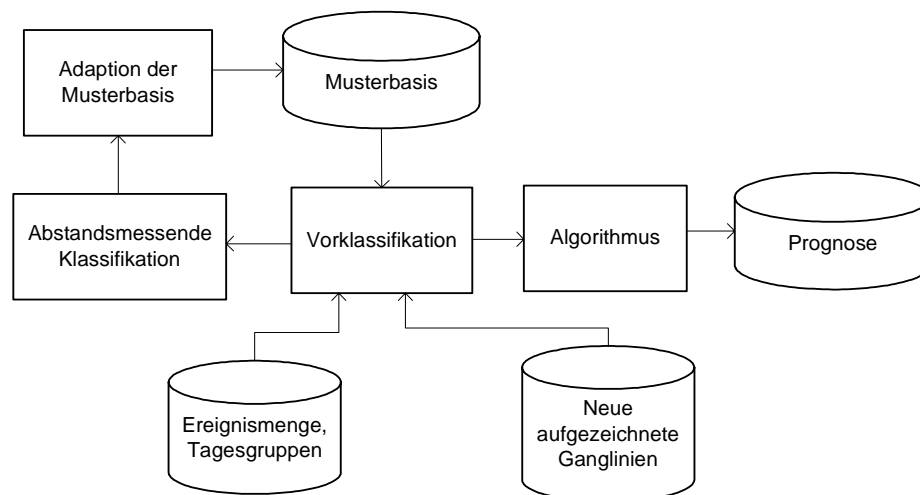


Abbildung 3.1 Ein generelles Prognosemodell (vgl. [Wild96])

Das generelle Prognosemodell besteht aus den folgenden Bestandteilen (siehe [Wild96]):

- Neue aufgezeichnete Ganglinien stellt die aktuelle aufgenommene Verkehrsinformation dar. Diese ist die Datenbasis für Verkehrsprognosen.
- Ereignismenge und Tagesgruppen: Verkehrsstärke hängt stark von Wochentagen ab. Da immer wiederkehrenden Ereignisse im ihrem Umfeld den Verkehr wesentlich beeinflussen, sollen bei Verkehrsprognose verschiedene Ereignisse und Tagesgruppe berücksichtigt werden. Die Daten über Ereignisse und Tagsgruppe werden zuerst erfasst und gespeichert. Beispiele für vorhersehbare Verkehrbeeinflussende Ereignisse sind:
 - Sportveranstaltungen in Stadien
 - Kulturelle Veranstaltungen
 - Wochenmärkte
 - Messen und Ausstellungen
 - Feiertag
 - Verkaufsoffener Sonntag
- Vorklassifikation stellt ein Prozess zur Bearbeitung der Datenbasis für Verkehrsprognosen dar. In diesem Prozess werden Ereignisse und Tagesgruppe aufgrund von Erfahrungen einen vergleichbaren Verkehrsablauf aufweisen und dadurch zu ähnlichen Ganglinien führen. Ganglinien, die sowohl in der Tagesgruppe als auch in allen

Ereignissen übereinstimmen, gelangen in dieselbe Vorklasse. Die erste Ganglinie einer neuen Vorklasse kann mit keiner anderen historischen Ganglinie verglichen werden, und führt deshalb sofort zur Bildung einer eigenen neuen Ganglinienklasse (im folgenden wird auch als Musterkurve verwendet). Jede weitere neu aufgezeichnete Ganglinie in derselben Vorklasse wird mit allen Musterkurven der in dieser Vorklasse bereits vorhandenen Ganglinienklassen verglichen.

- Abstandmessende Klassifikation: Nach der Vorklassifikation wird eine abstandsmessende Klassifikation durchgeführt, sofern in der ermittelten Vorklasse bereits ein oder mehrere Ganglinien vorliegen. Falls der Abstand kleiner als ein Kriterium ist, wird die Ganglinie in die am besten passende Klasse integriert, sonst wird die neue Ganglinie und damit der ihr zugrund liegenden Verkehr als neue zusätzliche Musterkurve in die Musterbasis aufgenommen.
- Die Musterbasis: Die Musterbasis enthält die Musterkurven aller Ganglinienklassen (zur vergleichener historischen Daten). Die ist auch nach Ereignisse und Tagesgruppe klassifiziert.
- Adaption der Musterbasis: Die Musterbasis wird fortlaufend an die sich verändernden Verkehrsabläufe angepasst.
- Algorithmen: Mit den erstellten Algorithmen können aus der Vergleichskurve und aktuelle Information Prognosewerte berechnet werden. Der Algorithmus in [Wild96] berechnet die Prognosewerte $\bar{q}(t_p)$ wie folgt:

$$\bar{q}(t_p) = q_v(t_p) + k \cdot \Delta q(t_0)$$

$$\text{wobei } \Delta q(t_0) = q_a(t_0) - q_v(t_0)$$

q_v = Verkehrsstärke der Vergleichsganglinie

q_a = aktuelle Verkehrsstärke

t_0 = Zeitpunkt der Prognoseerstellung

t_p = Prognosezeitpunkt

Der Faktor k soll durch Testreihen bestimmt werden. Als Beispiel kann Faktor k wie folgt berechnet werden:

$$k = -0,52 t_h + 0,842, \text{ falls } 0 < t_h < 1,6$$

$$k = 0, \text{ falls } t_h > 1,6$$

t_h = Prognosehorizont in Stunden

- Prognose: die Prognose soll die Verkehrsstärke in Zukunft vorhersagen.

3.3 Kurzzeitprognose im Verkehr

Im Verkehr werden Prognosen mit unterschiedlichen Zielsetzungen und unterschiedlich langen Zeithorizonten verwendet. Auf Basis unterschiedlich langer Zeithorizonten werden Prognosen häufig als Langzeitprognose, Mittelfristigprognose und Kurzzeitprognose untergegliedert. Prognosen mit langen Prognosehorizonten von einigen Monaten bis zu Jahrzehnten werden als Langzeitprognose bezeichnet und oftmals für Planungsaufgabe wie z.B. Verkehrswegeplanung, Planung des öffentlichen Verkehrs benötigt. Mittelfristig-

prognosen haben die Prognosehorizonten von einigen Stunden bis zu einigen Monaten und werden für Einsatzplanung wie z.B. Vorbereitung von Veranstaltungen benutzt. Im Vergleich zu Langzeitprognosen und Mittelfristigprognosen besitzen Kurzzeitprognosen den Prognosehorizonten von wenigen Minuten bis zu mehreren Stunden. Kurzzeitprognosen basieren auf den ständig ändernden Verkehrsparametern und unterstützen die Steuerungs- und Regelungsaufgaben in verkehrbeeinflussenden Systemen, d.h. sie werden oftmals zur Kontrolle der gegenwärtigen Verkehrssituationen gebraucht. Im Rahmen des Projekts OVID sind Kurzzeitprognosen im Verkehr von großer Bedeutung. Im folgenden wird deshalb auf die Kurzzeitprognosen eingegangen.

3.3.1 Kurzzeitprognoseart

Kurzzeitprognosen lassen sich in drei Teilbereichen unterteilen ([ZaHe80]):

- Prognose am Querschnitt
- Prognose über die Strecke (zeitlich-räumlich)
- Prognose von Strömen im Netz.

Im Rahmen des Projekts OVID werden ausschließlich Prognosen am Querschnitt betrachtet.

3.3.2 Prognosemodell am Querschnitt

Die Prognose am Querschnitt bezieht sich auf eine einzelne Messstelle, ohne benachbarte Bereiche einzubeziehen. Sie ist eine rein zeitliche Prognose ohne räumlichen Bezug. Die am Querschnitt erhobenen Messwerte bilden die Grundlage für die Prognose. Der Analysezeitraum wird in Intervalle untergliedert, die so groß sein sollen, dass Veränderungen des Verkehrszustandes, die möglicherweise einen Steuerungseingriff erfordern, erkannt werden können.

Für Kurzzeitprognosen wurde eine Vielzahl von Modellen in der Literatur diskutiert. Hier werden zwei Ansätze als Beispiele vorgestellt, dabei stellt q_p den Prognosewert dar und die beiden Ansätze gehen von Messwerten aus 5 Minuten Intervallen aus ([ZaHe80]):

- (1) Mittelwert aus 10 min, d.h. aus den zweijüngsten Messwerten q_{-1} und q_{-2} :

$$q_p = 1/2 (q_{-1} + q_{-2})$$

- (2) Gleitender Mittelwert, überlagert mit einem gleitenden Trend, der bis in die ersten 5 Minuten der Prognosezeit übertragen wird:

$$\overline{q_0} = \alpha \cdot q_{-1} + (1 - \alpha) \cdot \overline{q_{alt}}$$

$$\overline{\Delta q_0} = \alpha \cdot (q_{-1} - q_{alt}) + (1 - \alpha) \cdot \overline{\Delta q_{alt}}$$

$$q_p = \overline{q_0} + \overline{\Delta q_0}$$

α Glättungsfaktor

$\overline{q_0}$ gleitender Mittelwert im aktuellen Intervall

$\overline{q_{alt}}$ geleiteter Mittelwert im vorhergegangenen Intervall

$\overline{\Delta q_0}$ geleiteter Zuwachs im aktuellen Intervall

$\overline{\Delta q_{alt}}$ geleiteter Zuwachs im vorhergehenden Intervall.

3.4 Zusammenfassung

In diesem Kapitel wurde zuerst allgemeines Prinzip für Prognosen vorgestellt. Danach wurde auf ein generelles Prognosemodell eingegangen. Da im Rahmen des Projekts OVID Kurzzeitprognosen, insbesondere Kurzzeitprognose am Querschnitt, von großer Bedeutung sind, wurde die Grundlagen für Kurzzeitprognosen am Querschnitt ebenfalls erläutert.

4. Data Warehouse und OLAP

In dieser Arbeit wird die zu entwickelnde Konzeption mit Data Warehouse Technologie umgesetzt und die Anfragenprozess mit Online Analytical Processing (OLAP) implementiert. Deshalb werden die Grundlagen für Data Warehouse und OLAP-Systeme kurz erläutert.

4.1 Data Warehouse

Der Begriff Data Warehouse wurde von W. H. Inmon geprägt: „*A Data Warehouse is a subject-oriented, time-variant, and nonvolatile collection of Data in support of managements Decision support process.*“ Diese Definition für Data Warehouse impliziert mehrere Annahmen. Erstens, ein Data Warehouse wird immer auch physikalisch von den operativen Quelldatensystemen getrennt. Zweitens enthält ein Data Warehouse sowohl aktuelle als auch historische Detaildaten, zusätzlich auch eine Menge von leicht (moderat) bzw. hoch aggregierten (konsolidierten) Daten. Drittens wird die Struktur des Data Warehouse und der Daten in einem zentralen Metadaten-Repository hinterlegt. Die Information über die Daten (bzw. zumeist mit „Daten über Daten“ beschrieben) bilden das Fundament jedes Data Warehouse (vgl. [Kurz99]).

Ein Data Warehouse besitzt die folgenden Eigenschaften:

- **Subjekt-orientiert:** Die Daten eines Data Warehouse werden nach dem Anwendungsumfeld organisiert.
- **Integriert:** Ein Data Warehouse wird aus einer Vielzahl interner sowie externer Datenquelle gefüllt. Dabei spielt die Datenqualität eine wesentliche Rolle.
- **Zeit-Variant:** Die Daten eines Data Warehouse werden langfristig gespeichert. Ein Data Warehouse System soll „Zeit“ Dimension beinhalten. Damit werden die Zeitreihenanalysen unterstützt.
- **Schreibgeschützt:** Die Daten werden permanent gespeichert. Auf ein Data Warehouse wird nur lesend zugegriffen. Die Daten werden durch die Analysen nicht manipuliert.
- **Entscheidungsgestützt:** Der Bedarf des Entscheides an analytischer bzw. strategischer Information muss abgedeckt werden. Der Prozess der Entscheidungsunterstützung muss sinnvoll durch qualitative Analysen unterstützt wird.

Ein Data Warehouse wird häufig als ein spezielles Datenbanksystem implementiert. Im Vergleich zum Datenbanksystem soll allerdings ein Data Warehouse sich weniger an den Funktionen operativen Anwendungssystemen, sondern vielmehr an Analysethemen und Analysezwecke operativer Daten orientieren. Außerdem soll ein Data Warehouse ermöglichen, unterschiedliche Sichte auf die Datenobjekte abzuleiten. Für die Entwicklung eines Data Warehouse sollen, wie bei der Entwicklung von Datenbanksystemen, zuerst ein Datenmodell erstellt werden. Dazu wird auf die konzeptuelle Ebene das sogenannte multidimensionale Datenmodell verwendet. In der physikalischen Modellierungsebene werden das Star Schema oder das Snowflake Schema abgeleitet. Im folgenden werden die beiden Themen behandelt.

4.1.1 Multidimensionale Datenmodelle

Im multidimensionalen Datenmodell unterscheiden sich die Daten in qualifizierende und quantifizierende Daten. Die Quantifizierenden Daten sind die eigentlichen zu analysierenden Daten. Demgegenüber dienen die qualifizierenden Daten hauptsächlich zur Beschreibung der quantifizierenden Daten ([BaGü01]).

Multidimensionale Datenmodelle bestehen wesentlich aus zwei Elemente: Fakt und Dimension. Synonym zum Fakt sind Kennzahl und Kenngröße, die quantifizierende Daten darstellen und meistens numerische Daten sind. Der Fakt hat zentrale Bedeutung in jedem multidimensionalen Datenmodell. In Normalfällen muss Fakt mit mindestens einer Dimension assoziiert.

Eine Dimension stellt innerhalb des multidimensionalen Datenmodells eine ausgewählte Entität, mit der eine Auswertungssicht eines Anwendungsbereichs definiert wird. Dimension dient der eindeutigen orthogonalen Strukturierung des Datenraums in einem Data Warehouse System. Durch die Einführung von mehreren Dimensionen können dann die Daten für unterschiedliche Kombinationen aus diesen Dimensionen gespeichert, analysiert und ausgewertet werden. Eine wichtige Aufgabe eines Data Warehouse ist Speicherung, Analyse und Auswertung aggregierte Daten. Deshalb kann eine Dimension hierarchisch-zusammenhängende Merkmale enthalten, d.h. damit die Analyse bzw. Auswertung von Faktdaten auf verschiedene Granularitäten durchgeführt werden kann, werden Dimensionen hierarchisch organisiert. Die Daten auf der niedrigsten Stufe bestimmen die Datengranularität eines multidimensionalen Datenmodells und daher die zu speichernde Datenmenge. Die Hierarchie einer Dimension bildet mittels einer Baumstruktur eine Abstraktionshierarchie über die Elemente einer Dimension. Die Elemente einer Dimension bilden die Blätter eines Baums, die auch als basisgranulare Klassifikationsknoten bezeichnet. Die mittels einer Baumstruktur dargestellte Hierarchie wird wiederum als Klassifikationshierarchie bezeichnet. Sie beschreiben damit die verschiedenen Verdichtungsstufen einer Dimension. In einem Managementunterstützungssystem, ist es sinnvoll, das Zustandekommen von hochverdichteten Zahlen nachvollziehen oder auch eventuell noch eine weitere darunter liegende Verdichtungsebene sehen zu können. Hierbei findet ausgehend von Fakt mit Dimensionshierarchie eine Informationsverdichtung statt. Das Herunterbewegen in einer Dimensionshierarchie zu Elementen mit niedrigerem Verdichtungs niveau wird als *Drill-Down* und umgekehrte Richtung als *Roll-Up* bezeichnet.

4.1.2 Snowflake Schema und Star Schema

Nachdem ein Data Warehouse auf die konzeptuelle Ebene mit einem multidimensionalen Datenmodell dargestellt wurde, ist ein physisches Schema für das Data Warehouse zu erstellen. Der Grund hierfür liegt daran, dass Data Warehouse vorwiegend mit Datenbanktechnik umgesetzt werden. Die physische Modellierung entspricht der logischen Modellierung eines Datenbanksystems. Bei der Umsetzung mit Datenbanktechnik bestehen die Möglichkeit, Fakt und Dimensionen in einer relationalen Datenbank umzusetzen bzw. eine eigene Tabelle für Fakt und jede Dimension anzulegen.

Faktentabelle liegt im Zentrum einer physischen Modellierung. Alle Kenngröße werden innerhalb der Faktentabelle verwaltet. Zu Verbindungen zu alle Dimensionen muss die Faktentabelle noch Fremdeschlüsselbeziehungen zu den jeweils niedrigsten Klassifikationsstufen der verschiedenen Dimensionen beinhalten. Die Fremdschlüssel entsprechen den Zellkoordinationen in der multidimensionalen Datensicht. Sie bilden daher den

zusammengesetzten Primärschlüssel für die Faktentabelle. Mit dieser Struktur kommt der Name Snowflake, die wie eine Schneeflocke aussieht ([BaGü01]). Der Vorteil von Snowflake Schema besteht in der optimalen Unterstützung von Aggregaten, der besseren Unterstützung von N:M Beziehungen zwischen Hierarchieobjekten und der Redundanzfreiheit. Die Probleme des Snowflake Schema bestehen durch ein komplexeres Modell, welches für den Benutzer das Verständnis erschwert, und auch bei der Anfrage kompliziert erscheint, da viele JOIN-Operation bei Anfrage ausgeführt werden müssen, um viele Tabellen miteinander zu verbinden.

Da die Verbindungen zwischen Dimensionsstufen ziemlich zeitaufwendig sind, wird häufig das Star Schema in Data Warehouse System benutzt. Im Star Schema wurden alle Dimensionsstufen, die zu einer Dimension gehören, in einer Dimensionstabelle eingesetzt ([BaGü01]). Dieses führt zur Denormalisierung der Dimensionstabelle. An dieser Stelle weicht man von Prinzip der Normalisierung ab, um schnellere Anfragebearbeitung zu erreichen. Eine Faktentabelle wird mit alle Dimensionstabellen assoziiert. Im Modell ergibt sich dann ein Stern: die Faktentabelle bildet den Kern des Sterns, die dazugehörigen Dimension stellen die Strahlen des Sternes dar. Daher kommt der Name Star Schema.

Zusammenfassend besitzt das Star Schema also folgende Eigenschaften, die es für Anwendungen geeignet erscheinen:

- Einfache Struktur: Das Star Schema hat eine einfache Strukturierung. Bei dieser Struktur werden Anfragen einfacher und leichter formuliert und erstellt.
- Einfache und flexible Darstellung von Klassifikationshierarchien: Klassifikationshierarchien werden nicht mehr wie in Snowflake Schema durch viele zusammenhängende Tabelle dargestellt, sondern einfach innerhalb von Dimensionstabellen als Spalten angegeben.
- Effiziente Anfrageverarbeitung innerhalb der Dimension: Durch die Denormalisierung der Dimensionstabellen sind bei Selektionsprädikaten, die höhere Dimensionsstufen zur Einschränkung verwenden, keine Verbundoperationen zwischen verschiedenen Tabellen nötig, um die Menge von Tuple zu bestimmen, die mit der Faktentabellen verbunden werden müssen.

Im Vergleich mit dem Snowflake Schema hat das Star Schema geringer Speicherplatzbedarf und bietet bessere Änderungsmöglichkeit. Welches Entwurfsschema für ein Data Warehouse System eingesetzt werden soll, hängt von den konkreten Daten- und Anfragecharakteristiken. Eine Mischung von Star Schema und Snowflake Schema kann in der Praxis auch sinnvoll sein.

4.2 OLAP

Ein Data Warehouse dient meistens nur als eine Datenhaltung bzw. Datenbasis einem speziellen Analysezzweck. Die wirklichen Aufgaben der Analyse werden mit sogenannte OLAP-Systemen erledigt. Als ein analytisches und strategisches Informationssystem ist ein geeignetes Analyse-System für ein Data Warehouse sehr wichtig. OLAP ist die Abkürzung für Online Analytical Processing. Dieser Begriff wurde 1993 von Edgar F. Codd in seinem Werk „Providing OLAP to User Analysten: An IT Mandate“ wie folgende geprägt[Kurz99]:

„OLAP is a numerous, speculative >>what-if<< and/or >>why<< data model scenarios executed within the context of some specific historical basis and perspective. Dynamic enterprise analysis is required to create, manipulate,

animate and synthesize information from Enterprise Data Models. This includes the ability to discern new or unanticipated relationships between variables, the ability to identify the parameters necessary to handle large amounts of data, to create an unlimited number of dimensions (consolidation paths) and to specify cross-dimensional conditions and expressions”

Eine neue OLAP Definition wurden 1995 von Pendese und Greeth im OLAP Report eingeführt, die als FASMI (*Fast Analysis Shared Multidimensional Information*) bezeichnet. Die fünf Schlüsselwörter werden wie folgt erläutert (vgl. [Kurz99]):

- **Fast:** Ein OLAP-System sollte vernünftige Antwortzeit garantieren, das heißt, bei einfachen Abfragen innerhalb von 5 Sekunden, bei komplexere nicht mehr als 20 Sekunden.
- **Analysis:** Ein OLAP-System sollte Analysefunktion für Anwender einfach stellen ohne zusätzlichen Programmieraufwand auszuführen.
- **Shared:** Ein OLAP-System sollte die Daten im gemeinsamen Zugriff für mehren Benutzen zur Verfügung stellen.
- **Multidimensional:** Die gesamte Analysedatenbank muss semantisch und multidimensional modelliert sein. Diese bildet die Grundlage jeder multidimensionalen Analyse.
- **Information:** OLAP Datenbanken besetzten die Fähigkeit, aus Daten die für Analysen notwendige Information zu erzeugen.

Gemäß der obengenannten Definitionen ergeben sich für OLAP-Systemen die folgenden Anforderungen:

- **Multidimensionale konzeptionelle Sicht:** Die gespeicherten Datenbestände können auf vielfältige Art und Weise und von verschiedenen Sichten betrachtet werden und entsprechend in multidimensionalen Strukturen oder Modellen abbilden lassen.
- **Funktionale Transparenz für den Anwender:** Die für den Anwender sichtbaren und abfragbaren Datenbestände können aus mehreren Quellen stammen. Die zugrunde liegende Basistechnologie kann sehr kompliziert sein. Ein OLAP-System soll eine klare Trennung zwischen einer intuitiven Benutzerschnittstelle und der komplizierten Basistechnik gewährleisten.
- **Variable Zugriffsmöglichkeit:** In einem OLAP-System muss der Integrationsprozess einfach gestaltet werden, damit die Anwender in der Lage sind, unbeschränkte Zugriffe auf die Datenbestände möglich zu machen.
- **Konsistente Berichtsgenerierung:** Weder zunehmende Datenbankgröße noch wachsende Zahl von Dimensionen oder Anwendern dürfen zum Leistungsabfall bei der Berichtsgenerierung führen. Ein konstantes Antwortzeitverhalten ist schon im Hinblick auf dem Akzeptanten eines OLAP-Systems ein kritischer Erfolgsfaktor.
- **Client/Server Architektur:** Eine Client/Server Architektur soll je nach Anforderungen skalierbar und leicht erweiterbar sein. Eine OLAP Server Architektur zur gleichmäßigen Verteilung der Last ist von Vorteil.
- **Gleichgestellte Dimension:** Die Möglichkeiten hinsichtlich der Strukturieren und Berechnen sollen bei allen Dimensionen gleichermaßen möglich sein.
- **Automatische Anpassung der physikalischen Speicherform:** Die OLAP Datenbanken sollen in der Lage sein, das optimale Verhältnis zwischen zuvor kalkulierten und den

zur Laufzeit berechneten Werten selbst zu finden und umzusetzen, d.h. die optimale Ausnutzung der Ressourcen sicherzustellen.

- Unterstützung von mehreren Anwendern: Ein OLAP-System soll konkurrierendes und gleichzeitiges Lesen und Schreiben unterstützen.
- Unbeschränkte dimensionsübergreifende Operationen: Berechnung und andere Aktivitäten zwischen und über den Dimensionen dürfen nicht den Eingriff des Anwenders erfordern; das muss von OLAP automatisch leisten.
- Intuitive Datenmanipulation: Die Navigation von OLAP soll die Anwender leicht machen, die Daten bis zu Zellen steuern zu können.
- Flexible Berichterstellung: Die Anordnung der Daten sollte ohne Beschränkungen möglich sein.

4.2.1 ROLAP, MOLAP, DOLAP und HOLAP

Ein OLAP-System wird als analytisches Werkzeug bzw. System auf Basis eines Data Warehouse verwendet. Mit unterschiedlichen Architekturkonzepten können OLAP-Systeme in folgenden Typen verdeutlicht werden.

- ROLAP – Relationales OLAP

Der Relationale Ansatz ist am weitesten verbreitet und wird am häufigsten bei den großen Data Warehouse Projekten sehr erfolgreich eingesetzt. Das multidimensionale Datenmodell wird in zweidimensionalen Tabellen gespeichert. Dazu wird das Star Schema oder das Snowflake Schema angewendet. SQL wird zur Datentransformation, Verwaltung des Data Warehouse und für die OLAP Abfragen verwendet. Der Einsatz von relationalen Datenbanktechnologien als eine Lösung von Data Warehouse hat Nachteile. Erstens ist Standard SQL nur bedingt ausreichend. Zweitens müssen Sicherheitsmaßnahmen zusätzlich noch durch Metadaten gewährleistet werden.

- MOLAP - Multidimensionales OLAP

Bei MOLAP wird Data Warehouse mit der effizienten multidimensionalen Datenbank implementiert. Kennzahlen werden in Form eines großen Speicherarrays, der Zugriff auf einzelnen Datenzellen über die assoziierten Dimensionslisten wird, welche das Element der einzelnen Hierarchieobjekte enthalten, berechnet. Die operativen Daten aus den Vorsystemen werden in einer Art von Initialisierungsprozess ausgelesen und in die multidimensionale Struktur übernommen. Meisten MOLAP Systemen haben eigene Multidimensionale Abfragesprache. Durch diese effizienten Multidimensionalen Speicherstrukturen erfolgen gute Antwortzeiten. Powervolle Abfragesprache liefern umfangreiche Analysefunktionalität. Einige Nachteile sind z.B. aufwendiger Ladeprozess, eingeschränktes Datenvolumen und leere Zellen bei leicht verdichtete Daten.

- HOLAP – Hybrid OLAP

HOLAP ist gekennzeichnet durch die Vereinigung der Vorteilen Relationaler Datenbanktechnologie und multidimensionaler Speicherstrukturen. Dabei wird das relationale Datenbanksystem zur Speicherung der leicht verdichteten historischen Detaildaten verwendet, und zu effizienten speichern höher verdichteten Datenwürfeln wird ein multidimensionales Datenbanksystem aufgebaut. Die bestehenden Nachteile sind komplexes Architekturkonzept und uneinheitliche OLAP Abfragesprache.

- DOLAP – Desktop OLAP

DOLAP wird auch als Client OLAP bezeichnet. Diese benötigen kein Server Backend, um OLAP Abfragen stellen zu können. Stattdessen werden nur benötigten Teildaten aus den zur Verfügung stehenden Datenquellen auf den Client geladen und dort multidimensional aufbereitet. DOLAP ist sehr geeignet für kleine und klar abgegrenzte Anwendungsgebiete. Der Vorgang des Ladens von Daten zu Client verursacht aber einen hohen Netzwerkverkehr.

4.2.2 Standardfunktionen von OLAP

OLAP-Standardfunktionen lassen sich im folgenden erläutern:

- *Drill Down*: Der Detaillierungsgrad der Faktendaten wird erhöht, indem die aktuellen Hierarchieobjekte entlang Dimensionspfad in verdichtete Richtungen wechseln.
- *Roll Up*: Roll Up ist die umgekehrte Operation zum Drill Down, um Detaillierungsgrad der Faktendaten zu verringern. Dabei wechseln die aktuellen Hierarchieobjekte entlang Dimensionspfad in gegen Richtung.
- *Pivot*: Die Berichtesdaten werden aus unterschiedlichen Perspektiven zusammengestellt. Dabei wird die Reihenfolge der aktuell dargestellten Dimensionen vertauscht, der Datenwürfel wird virtuell gedreht.
- *Slice*: Die ursprüngliche OLAP Abfrage wird um eine Dimension verringert, entspricht dem Abschneiden einer Scheibe des Datenwürfels.
- *Dice*: Ein Datenfilter wird ausgeführt, entspricht dem Ausschneiden eines Teiles aus dem bestehenden Datenwürfel. Der ursprüngliche Abgefragte Dimensionalität bleibt erhalten, nur die dargestellten Elemente der Hierarchieobjekte verändern.
- *Drill Across*: In einem OLAP Bericht können die Daten aus mehreren Faktentabellen, welche eindeutig über gemeinsame Dimensionstabellen zusammenhängen, dargestellt werden.

Es gibt noch weitere Funktionen wie z.B. *Drill Aside*, *Drill Through*. Diese hängen von jeweiligen Produkte ab.

4.3 Zusammenfassung

In dieser Arbeit wird Data Warehouse als Lösungsansatz eingeführt. Deshalb wurden in diesem Kapitel die Grundlagen für Data Warehouse und OLAP vorgestellt. Der Begriff multidimensionales Datenmodell mit Fakten und Dimensionen wurde erklärt. Zwei wichtige logische Modelle für Data Warehouse, Star und Snowflake Schema, wurden vorgestellt. OLAP mit unterschiedlichen Architekturkonzepten und ihre Standardfunktionen wurden in diesem Kapitel ebenfalls erläutert.

5. Konzeptuelle Modellierung

In der Arbeit ist ein Data Warehouse Schema für Erstellung von Ganglinien und Bewertung von Prognosen im Verkehr zu entwickeln. Mit diesem Schema können die Daten im Verkehr wie z.B. Simulationsdaten, Prognosewerten, erwartete Daten in Form von Ganglinien erstellt, gespeichert und ausgewertet werden. Bei der Entwicklung des Data Warehouse Schemas ist zuerst ein konzeptuelles Datenmodell zu entwickeln, damit die grundlegenden Ideen auf die konzeptuelle Ebene dargestellt und erläutert werden können. Im Zentrum des konzeptuellen Datenmodells steht das Konzept Ganglinie, d.h. das zu entwickelnde Datenmodell dient überwiegend zu Darstellung, Speicherung und Auswertung von Ganglinien. Deshalb wird zuerst das Konzept Ganglinie im Verkehr mit Beispiele skizziert. Dann wird auf das konzeptuelle Datenmodell eingegangen.

5.1 Ganglinien im Verkehr

Ganglinie wird allgemein als eine grafische Darstellung von quantitativen Daten in einer Zeitreihe dargestellt. Die quantitativen Daten werden zuerst einer bestimmten Zeitreihe zugeordnet und dann in verschiedenen graphischen Formen dargestellt. Im Verkehrsbereich können die Verkehrsstärken, wie z.B. die Anzahl von Fahrzeugen, die innerhalb eines Zeitintervalls durch einen Verkehrsquerschnitt fahren, nach jedem Zeitintervall gemessen und gespeichert werden. So erhält man eine Reihe von Daten über Verkehrsstärken in einer Zeitreihe. Eine graphische Darstellung dieser Datenreihe wird dann als Ganglinie bezeichnet.

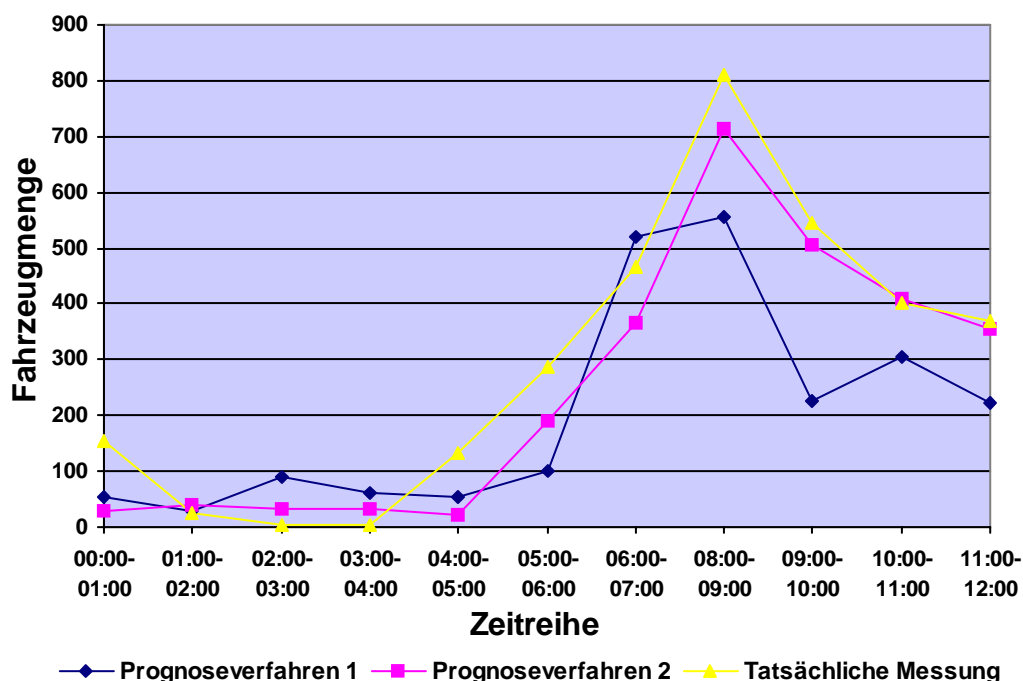


Abbildung 5.1 Beispiele für Ganglinien im Verkehr

Die im Rahmen des Projekts OVID verwendeten Daten wie z.B. gemessene Daten, Prognosedaten, Simulationsdaten und erwartete Daten werden als Ganglinien dargestellt und gespeichert. In Form von Ganglinien können die Verkehrsdaten anschaulich dargestellt und intuitiv miteinander verglichen werden. Insbesondere können verschiedene Prognoseergebnisse verglichen und bewertet werden.

Abbildung 5.1 stellt ein Beispiel für Ganglinien dar. Auf dieser Abbildung stehen drei Ganglinien. Eine Ganglinie für tatsächlich gemessene Daten und zwei Ganglinien für die Daten aus unterschiedlichen Prognoseverfahren. Während die x-Achse des Diagramms die Zeitreihe mit Zeitintervall von 1 Stunde darstellt, deutet die y-Achse die Fahrzeugmengen an. Aus der Abbildung ist deutlich zu sehen, dass die Abweichung zwischen der Ganglinie für Prognoseverfahren 2 und der Ganglinie für tatsächlich gemessene Daten meistens kleiner als dies zwischen der Ganglinie für Prognoseverfahren 1 und der Ganglinie für tatsächlich gemessene Daten ist. Es ist deshalb erkennbar, dass in diesem Fall Prognoseverfahren 2 besser als Prognoseverfahren 1 ist. Durch derartige Vergleiche kann die Qualität von Prognosen intuitiv bewertet werden.

5.2 Konzeptuelles Datenmodell

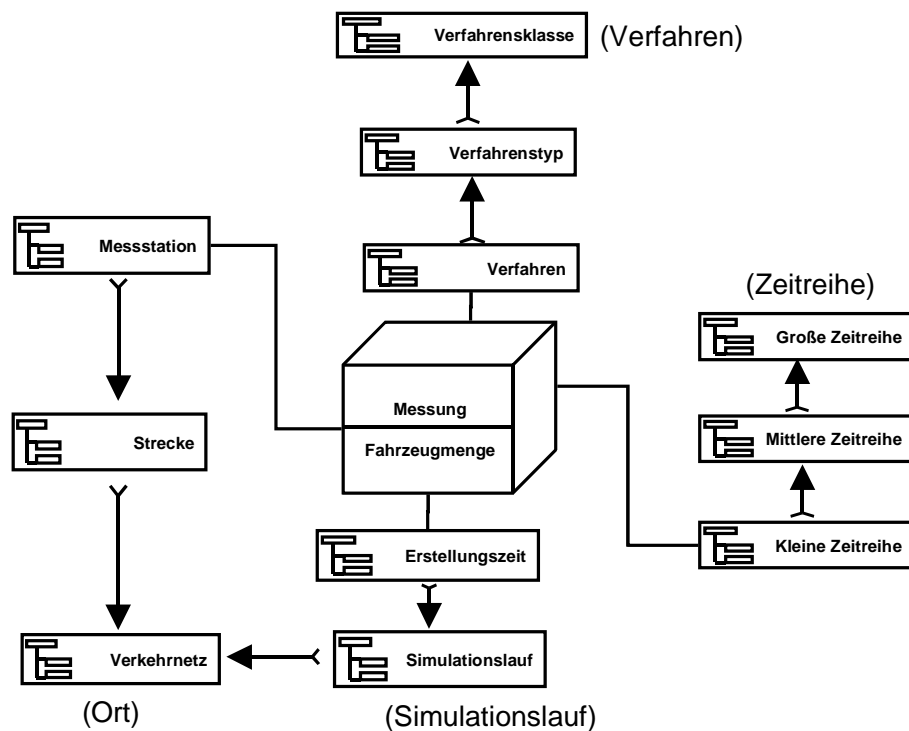


Abbildung 5.2 Konzeptuelles Datenmodell (mE/R-Model)

Um eine Ganglinie vollständig darzustellen, müssen außer Zeitreihe noch andere qualifizierenden Merkmale für die Ganglinie erfasst werden. Ganglinien können mit verschiedenen Simulationen, Prognosemethoden und Algorithmen erzeugt werden. Deshalb spielen Prognoseverfahren eine wichtige Rolle bei der Erstellung von Ganglinien. Eine Ganglinie ist außerdem einer bestimmten Messstation bzw. Straßenstrecke zugeordnet. Ortinformation ist deswegen auch ein wichtiges Merkmal für Ganglinien. Darüber hinaus wird eine Ganglinie oft durch die Erstellungszeit beschrieben. Um Ganglinie mit

verschiedenen qualifizierenden Merkmalen auf konzeptuelle Ebene darzustellen, eignet sich das multidimensionale Datenmodell.

Die konventionellen Datenmodelle wie z.B. E/R-Modell reichen nicht mehr aus für multidimensionales Datenmodell. Zu einen unterstützt das E/R-Modell keine semantische Unterscheidung zwischen qualifizierenden und quantifizierenden Daten. Zu anderen gibt es Schwierigkeiten bei der Abbildung von Dimensionshierarchien im E/R-Modell. Demzufolge wurde das mE/R-Modell (multidimensional E/R-Model) als spezielle Modellierungstechnik für multidimensionales Datenmodell entwickelt. Sie stellt eine Erweiterung des bekannten E/R-Modells für relationale Schema dar. Details über das mE/R-Modell finden Sie bei [SBHD99]. In dieser Arbeit wird das mE/R-Modell für die Darstellung des konzeptuellen Datenmodells verwendet.

Abbildung 5.2 stellt das in dieser Arbeit entwickelte konzeptuelle Datenmodell im mE/R-Modell dar. Im Zentrum des Modells steht der Fakt, der durch die Entität Messung dargestellt ist. Die zu analysierende Fahrzeugmenge ist die einzige Kennzahl und gehört zu Entität Messung. Mit Entität Messung stehen vier Dimensionen in Verbindung, das heißt Dimension Ort, Dimension Simulationslauf, Dimension Verfahren und Dimension Zeitreihe. Jede Kennzahl Fahrzeugmenge wird durch die vier Dimensionen festgestellt und charakterisiert. Im folgenden werden die vier Dimensionen detailliert vorgestellt.

5.2.1 Dimension Ort

Eine Messung wird für eine bestimmte Messstation bzw. eine Straßenstrecke durchgeführt. Dimension Ort umfasst Information über den Ort bzw. über die Messstation. Diese Dimension besitzt drei Klassifikationsstufen, die in eine hierarchischer Struktur zusammengeführt werden. Die niedrigste Stufe ist Messstation, die durch verschiedene Information wie z.B. Name, ID usw. bezeichnet. Auf der Mittelestufe steht Strecke, da eine Strecke mehrere Messstationen enthält. Strecke werden durch Name, Nummer/ID sowie Informationen über Strecke wie z.B. Länge, Kategorie (Autobahn, Bundesstraße, Landstraße usw.) beschrieben. Ein bestimmtes Verkehrsnetz umfasst dann viele Strecken. Daher steht Verkehrsnetz auf die oberste Stufe.

5.2.2 Dimension Simulationslauf

Durch Dimension Simulationslauf werden Erstellungsdatum, Erstellungsuhrzeit und gekennzeichnete Simulationslauf für eine Messung erfasst. In einer Verkehrssimulation gehört jeder Messwert, der durch Berechnung von erwarteten Werten, durch Prognose oder durch Messung während der Simulation erzeugt wird, zu einem Simulationslauf. Diese Dimension verfügt über zwei Klassifikationsstufen. Die niedrige Stufe ist Simulationszeit, die durch Erstellungszeit einer Messung dargestellt ist. Da mehrere Simulationen können für dasselbe Verkehrsnetz gleichzeitig durchgeführt werden, steht auf die Oberstufe deshalb Simulationslauf. Simulationslauf wird durch Simulationslaufnummer gekennzeichnet. Außerdem wird ein Simulationslauf immer für ein bestimmtes Verkehrsnetz durchgeführt. Dies wird in diesem konzeptuellen Datenmodell durch die Beziehung zwischen Verkehrsnetz und Simulationslauf dargestellt.

5.2.3 Dimension Zeitreihe

Dimension Zeitreihe ist grundlegend für Ganglinie-Darstellungen. Mit Dimension Zeitreihe werden die Tagszeit in gewünschte Intervallgröße geteilt und als eine Reihe dargestellt. Diese Dimension wird ebenso in hierarchische Stufe strukturiert. Die kleine Zeitreihe kann z.B. die Intervallgröße eine Viertelstunde haben. Eine mittlere Zeitreihe hat die Intervallgröße eine halbe Stunde. Für große Zeitreihe wird z.B. die Intervallgröße eine Stunde eingesetzt.

Es gibt eigentlich zwei Möglichkeiten für Modellierung von Zeitreihen. Darstellung von Zeitreihen ist einerseits eine Standardfunktionalität herkömmlicher OLAP-Systemen. Deshalb können Zeitreihen mit OLAP-Systemen direkt umgesetzt werden. Andererseits lassen sich Zeitreihen auch über Datenmodell zu modellieren. Zeitreihe in OLAP-Systemen ist unflexibel, da eine Zeitreihe in OLAP-System oft nur mit gleich langen Intervallgrößen dargestellt werden kann. In dieser Arbeit werden Zeitreihen mit Datenmodell modelliert, damit eine Ganglinie einer Zeitreihe mit unterschiedlich langem Intervallgröße zugeordnet werden kann. Für stark zeitabhängige Verkehrsdaten können unterschiedlich lange Intervallgrößen gewählt werden. Die Straßen sind von 6:00 Uhr bis 9:00 Uhr und von 16:00 Uhr bis 18:00 Uhr meistens stark belastet, da viele Leute in diesem Zeitraum zur Arbeit bzw. zurück nach Hause fahren. Eine viertelstündliche Intervallgröße ist deshalb für diesen Zeitraum geeignet. Von 22:00 Uhr bis 04:00 Uhr sind wenige Fahrzeuge unterwegs. Für diesen Zeitraum ist zweistündliche Intervallgröße angemessen. Demgegenüber macht stündliche Intervallgröße für anderen Zeitraum sinn.

5.2.4 Dimension Verfahren

Mit Dimension Verfahren wird eine Messung dadurch charakterisiert, aus welchem Verfahren bzw. Algorithmus die Kennzahl entstanden wurde. Diese Dimension besitzt drei Klassifikationsstufen. Die niedrigste Stufe enthält Information über einzelnes Verfahren. In zweite Stufe werden einzelne Verfahren in Verfahrenstypen gruppiert. Für Prognoseverfahren gibt es z.B. den Typ Langzeitprognose und den Typ Kurzzeitprognose. In oberster Stufe stehen die generelle Verfahrensklasse wie z.B. gemessene Daten, Prognosedaten und erwartete Daten.

5.3 Zusammenfassung

In diesem Kapitel wurde zuerst das Konzept Ganglinien im Verkehr mit einem Beispiel skizziert. Dann wurde das konzeptuelle Datenmodell für die Darstellung von Ganglinien mit mE/R-Modell vorgestellt. Im Zentrum des konzeptuellen Modells steht die Entität Messung, die Kennzahl Fahrzeugmenge enthält. Kennzahl Fahrzeugmenge wird durch vier Dimensionen festgestellt und charakterisiert.

6. Umsetzung des konzeptuellen Datenmodells

6.1 Physische Data Warehouse Schema

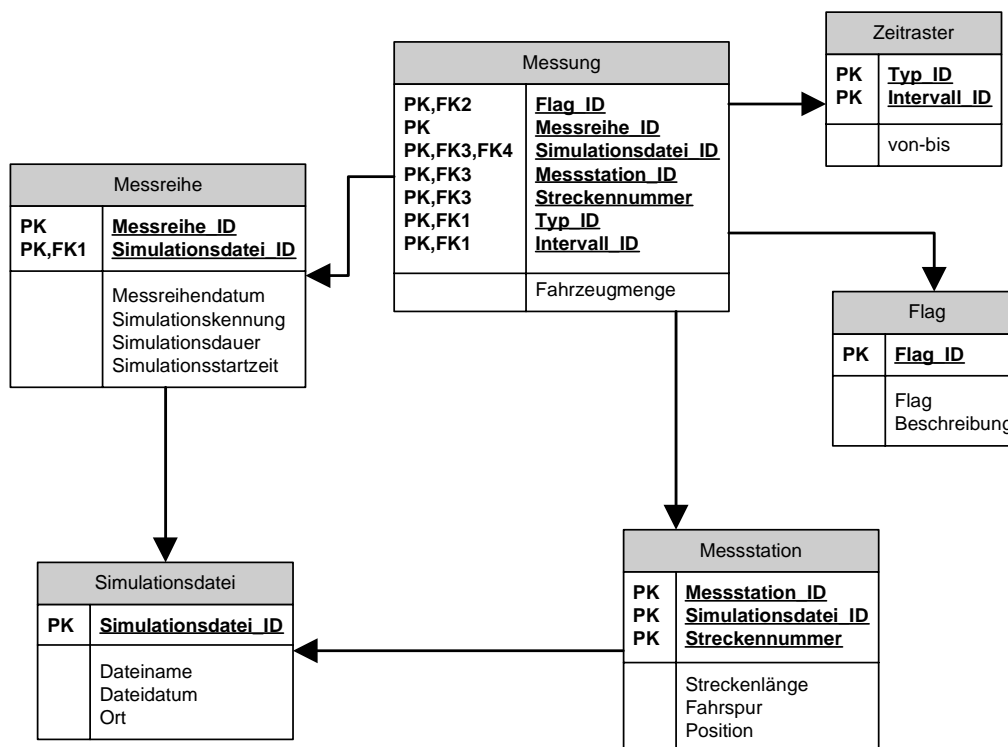


Abbildung 6.1 Das physische Data Warehouse Schema

Auf Basis des konzeptuellen Datenmodells ist ein physischer Entwurf für ein Data Warehouse zu entwickeln. Dabei werden die Einschränkungen und Besonderheiten der eingesetzten Techniken und Systemen berücksichtigt. In dieser Arbeit wird relationale Datenbank für die Implementierung eingesetzt. Der physische Entwurf besteht in diesem Fall aus der Transformation von Entities und Beziehungen in die entsprechenden Relationen bzw. Tabellen. Das konzeptuelle Datenmodell kann bei physischem Entwurf in einem Star Schema oder Snowflake Schema abgebildet. In der Arbeit wurde das Star Schema verwendet. Abbildung 6.1 stellt das physische Data Warehouse Schema dar.

Dieses Schema ist ein 4-dimensionales Schema. Tabelle Messung steht als Faktentabelle im Zentrum des Schemas und bezieht sich auf 4 Dimensionen: Messreihe, Messstation, Flag und Zeitraster. Die Fahrzeugmenge ist die einzige Kennzahl der Faktentabelle und wird durch die 4 Dimensionen charakterisiert. Der Fakt und die Dimensionen des konzeptuellen Datenmodells werden wie folgt in den Tabellen des physischen Schemas abgebildet:

- Tabelle Messung implementiert den Fakt Messung

- Tabelle Messreihe entspricht Dimension Simulationslauf
- Mit Tabelle Messstation und Simulationsdatei wird Dimension Ort implementiert.
- Tabelle Flag realisiert Dimension Verfahren.
- Tabelle Zeitraster bildet Dimension Zeitreihe ab.

Das Schema verfolgt zwar das Star Schema, aber nicht das reine Star Schema, d.h. nicht wie ein gängiges Star Schema, deren Dimensionen voneinander unabhängig sind. Die Unabhängigkeit von Dimensionen wird in der Literatur als Orthogonalität von Dimensionen bezeichnet (siehe [BaGü01]). In diesem Schema kann die Orthogonalität von Dimensionen nicht vollständig erreichen, da die Simulationsdatei etwas besondere behandelt werden sollen. Da Dimension Messreihe und Dimension Messstation sich auf Tabelle Simulationsdatei beziehen, sind die beiden Dimensionen deshalb voneinander abhängig. Die Orthogonalität von Dimensionen im Schema ist daher verletzt. Durch das redundante Vorhanden der Simulationsdatei in den beiden Dimensionen können Dimension Messreihe und Messstation von einander physisch getrennt werden und ein reines Star Schema kann dann erreichen. Dies ist zwar gegen Datenbank-Normalisierung, aber von Data Warehouse zugelassen. Allerdings könnte Tabelle Simulationsdatei sehr vielen Daten enthalten und demzufolge sehr hoher Speicherplatzaufwand entstehen. Um die Redundanz und hoher Speicherplatzaufwand zu vermeiden, wird in diesem Schema nur eine Tabelle für die beiden Dimensionen abgebildet.

6.2 Dimensionen

In diesem Abschnitt werden die Dimensionen des Schemas ausführlich erklärt.

- **Dimension Messreihe:** eine Messreihe bezeichnet einen Simulationslauf. Messreihe_ID und Simulationsdatei_ID zusammen bilden die Primärschlüssel dieser Tabelle. Mit Dimension Messreihe wird Kennzahl Fahrzeugmenge mit entsprechender Information über Simulation beschrieben. Eine Simulation wird durch Simulationskennung, Simulationsdatum (am Datum wurde eine Simulation ausgeführt bzw. eine Messreihe erzeugt), Simulationsstartzeit und Simulationsdauer charakterisiert. Eine Simulation bezieht sich noch auf eine Simulationsdatei. Tabelle Simulationsdatei enthält die Information (wie z.B. Dateiname, Dateidatum und Ort), die bei der Simulation mit VISSIM benötigt sind. An einem bestimmten Ort kann eine Simulation mehrmals durchgeführt werden.
- **Dimension Messstation:** eine Messstation wird eindeutig durch Messstation_ID, Simulationsdatei_ID und Streckennummer dargestellt. Durch Dimension Messstation wird Kennzahl Fahrzeugmenge mit Ortinformation beschrieben. Eine Streckennummer bezeichnet eine Strassstrecke, auf die entsprechende Messstationen legen. Dimension Messstation enthält noch zusätzliche Information über Strecke, wie Strecklänge, Fahrspur und Position. In einer Simulationsdatei werden mehrere Strecken definiert und jede Strecke kann mehrere Messstation enthalten.
- **Dimension Zeitraster:** Diese Dimension entspricht Dimension Zeitreihe im konzeptuellen Datenmodell und enthält die Information über Zeitintervalle. Ein Zeitintervall wird durch Intervalltyp, Intervall_ID und Start-/Endzeitpunkt dargestellt. Intervalltyp kann gemäß konkreter Anforderung definiert werden. Wir können z.B. Zeitintervall von 15, 30 oder 60 Minuten als unterschiedliche Intervalltype definieren. Intervall_ID ist die Reihenummer für Intervalle eines Zeitrasters. Datenfeld „VON-BIS“ ist in Form „Startzeitpunkt-Endzeitpunkt“ darzustellen. Mit dieser Dimension kann die

Kennzahl Fahrzeugmenge einem bestimmten Zeitraster zugeordnet und dann als Ganglinien dargestellt werden.

- **Dimension Flag:** Diese Dimension dient zu Unterscheidung der verschiedenen Arten von Messdaten. Durch diese Dimension wird festgelegt, ob eine Kennzahl einem Simulationswert, Prognosewert oder erwarteten Wert wie z.B. Durchschnittwert, Gewichtwert usw. entspricht.

6.3 Faktentabelle

Im Zentrum des Schemas steht die Faktentabelle, welche die zu analysierende und prognostizierende quantifizierende Daten enthält. In diesem Schema enthält die Faktentabelle nur eine einzige Kennzahl, die Fahrzeugmenge. Die folgenden Fremdschlüssel zu Dimensionen bilden zusammen die Primärschlüssel der Faktentabelle:

- Flag_ID
- Messreihe_ID
- SimulationsDatei_ID
- Messstation_ID
- Streckennummer
- TYP_ID
- Intervall_ID

Informationen in Dimensionen dienen dazu, diese Kennzahl zu qualifizieren und zu charakterisieren. Mit den qualifizierenden Daten aus Dimension Flag, Messreihe, Messstation und Zeitraster kann Kennzahl Fahrzeugmenge vollständig definiert werden.

6.4 Datenquelle für Simulationsdaten

Mit dem Datenbankschema in [Sand04] wird jedes durchquertes Fahrzeug mit Zeitpunkt (sekundengenau), Messstation, Simulationsdatei, Simulationsparameter und Fahrzeugtyp bereits zusammengestellt. Basiert auf dieses Datenbankschema wird eine Datenbanksicht aufgebaut, damit die Daten durch das Datenbankschema geladen und als Simulationsdaten erfassen werden. Der SQL-Befehlen (in Oracle 9i) zur Erstellung der Simulationsdaten sieht wie folgt aus:

```
SELECT  MessReihe_id AS MessReihe,
        MessStation_id AS MessStation,
        to_char(DateiDatum+(round(tausfahrt/60)/1440, DD.MM.YYYYHH24:MI')
              AS Simulationszeit,
        round(tausfahrt/60) AS Simulationsminuten,
        count(*) AS Verkehrsmenge
FROM    OVIDSA01.Simulationsdatei
WHERE    MessReihe.sim_datei_id=OVIDSA01.MessReihe.generatede-key AND
```

OVIDSA01.MessReihe.sim_datei_id=OVIDSA01.Simulationsdatei.sim_id

GROUP BY **MessReihe_id,**
 Messstation_id,
 DateiDatum+(round(tausfahrt/60/1440),
 Round(tausfahrt/60))

Durch die Funktion DateiDatum+round(tausfahrt/60)/1440 wird die Eintrittszeit eines Fahrzeugs in Messstation in der Format DD.MM.YYYYHH24:MI umgerechnet. Mit GROUP BY werde die Anzahl der Fahrzeuge, die eine Messstation um die Eintrittszeit durchquerten, aggregiert.

6.5 Zeitreihendarstellung

Um die Verkehrsstärke als Ganglinie darzustellen, wird Tabelle Zeitraster eingeführt. In dieser Tabelle bilden Intervalltyp und Intervall_ID die Primärschlüssel. Datenfeld Von-Bis repräsentiert Anfangen- und Endzeitpunkt einer Zeitreihe.

Wie bereits im Abschnitt 5.2.3 erwähnt wurde, ist eine Zeitreihe mit unterschiedlich langer Intervallgröße bei Verkehrsprognose von Vorteil. Tabelle Zeitraster soll unterschiedlich lange Intervallgröße speichern können. Die Tagszeit (24 Stunden) wird deshalb zuerst in gewünschte Intervalltype aufgeteilt. Als Beispiel definieren wir Intervalltyp 1 für eine viertelstündige Intervallgröße, Intervalltyp 2 für eine stündliche Intervallgröße und Intervalltyp 3 für eine zweistündliche Intervallgröße. Für verschiedene Tagszeit werden verschiedenen Intervalltype wie folgt verwendet:

- Von 6:00 Uhr bis 9:00 Uhr: Intervalltyp 1, viertelstündliche Intervallgröße
- Von 16:00 Uhr bis 18:00 Uhr: Intervalltyp 1, viertelstündliche Intervallgröße
- Von 22:00 Uhr bis 04:00 Uhr: Intervalltyp 3, zweistündliche Intervallgröße
- Sonstige Tagszeit: Intervalltyp 2, stündliche Intervallgröße

Als ein Beispiel stellt Tabelle 6.1 eine Zeitreihe mit unterschiedlich langer Intervallgröße dar.

Intervalltyp	Intervall_ID	Von-Bis
3	0	00:00-02:00
3	1	02:00-04:00
2	2	04:00-05:00
2	3	05:00-06:00
1	4	06:00-06:15
1	5	06:15-06:30
1	6	06:30-06:45
1	7	06:45-07:00
1	8	07:00-07:15
1	9	07:15-07:30

1	10	07:30-07:45
1	11	07:45-08:00
1	12	08:00-08:15
1	13	08:15-08:30
1	14	08:30-08:45
1	15	08:45-09:00
2	16	09:00-10:00
2	17	10:00-11:00
2	18	11:00-12:00
2	19	12:00-13:00
2	20	13:00-14:00
2	21	14:00-15:00
2	22	15:00-16:00
1	23	16:00-16:15
1	24	16:15-16:30
1	25	16:30-16:45
1	26	16:45-17:00
1	27	17:00-17:15
1	28	17:15-17:30
1	29	17:30-17:45
1	30	17:45-18:00
2	31	18:00-19:00
2	32	19:00-20:00
2	33	20:00-21:00
2	34	21:00-22:00
3	35	22:00-00:00

Tabelle 6.1 Beispiel für Tabelle Zeitraster

Um Zeitreihendaten in Tabelle Zeitraster einzutragen, wird für jeden Intervalltyp jeweils eine Prozedur gebraucht. Ein Beispiel für Intervalltyp 4 (halbstündlich) sieht wie folgend aus:

AS

TypNUMBER:=4;

Intervallminuten NUMBER:=30;

i NUMBER:=1;

BEGIN

FOR i IN 1..24

LOOP

```

INSERT INTO ZeitRaster VALUES (TypNumber, i*1-1, to_char((i-1), to_char(i-1)||':00-'|| to_char(i-1)||':30'));

INSERT INTO ZeitRaster VALUES (TypNumber, i*2-1, to_char(i-1)||'30-'|| to_char(i)||'00');

END LOOP;

END;

```

6.6 Daten Integration

Ein Ziel dieses Data Warehouse Schemas ist die Unterstützung der Bewertung von Prognosen. Um das Ziel zu erreichen, sollen Simulationsdaten und Prognosewerte in der Faktentabelle gleichzeitig gespeichert werden. Zusätzlich sollen die erwarteten Daten, wie z.B. Durchschnittswerte von ausgewählten Messreihen oder Gewichtswerte mit gegebenen Gewichtsfaktoren auch in der Faktentabelle gehalten werden. Um verschiedene Arten von Messdaten im Data Warehouse Schema gleichzeitig zu behandeln, wird Dimension Flag eingeführt. Mit einem Flag kann identifiziert werden, ob eine Kennzahl Fahrzeugmenge in Tabelle Messung ein Simulationswert, Prognosewert oder ein erwarteter Wert ist. Erzeugte Flags und entsprechende Beschreibungen werden in Tabelle Flag gespeichert. Für die Erzeugung von Flag ist beispielweise die Regel zu verfolgen, dass ein Flag für Simulationswerte mit 1, für erwartete Daten mit 2 und für Prognosewerte mit 3 anfangen. Tabelle 6.2 stellt ein Beispiel für Tabelle Flag dar.

Flag_ID	Flag	Beschreibung
1	100	Simulationswerte mit Zeitintervall 1h
2	210	Durchschnittswerte aus Messreihe 4, 21, 41, 66
3	220	Gewichtswerte nach Gewichtsfaktoren 4(0.1), 21(0.2), 41(0.3), 66(0.4)
4	300	Prognosewerte
...

Tabelle 6.2 Beispiel für Tabelle Flag

6.7 Gangliniendarstellung

Um Messdaten (Simulationsdaten, Prognosedaten und erwartete Daten) als Ganglinie mit dem Data Warehouse Schema darzustellen, werden die Daten in der Faktentabelle Messung mit Prozedur erfasst. Als ein Beispiel stellt die folgende Prozedur für das Erzeugen von Ganglinien (Intervalltyp 4, Intervall 30 Minuten) dar.

```

AS
Flag1 NUMBER:=2;
Typ2 NUMBER:=4;
Intervall NUMBER:=30;

```

```

BEGIN
DELETE FROM MESSMENGE;
INSERT INTO MESSMENGE
SELECT Flag1 AS Flag
      MessReihe,
      MessStation,
      substr(Simulationszeit,1,10) AS SimulationsDatum,
      Typ2 AS Typ,
      To_number(substr(Simulationszeit,12,2)*2+floor(to_number(substr(Simulation
szeit,15,2)/Intervall) AS Intervall_id,
      SUM(Verkehrsmenge) AS Verkehrsmenge
FROM      MESSDATEN
GROUP BY  MessReihe,
          MessStation,
          substr(Simulationszeit,1,10)
          Substr(Simulationszeit,12,2),
          floor(to_number(substr(Simulationszeit (15,2))/Intervall);
END;

```

Zum Sparen von Speicherplatz wird in Tabelle Messung nur die Messmenge in aktuellem gewünschtem Intervall gespeichert. Deshalb steht am Anfang der Prozeduren die Anweisung DELETE. Mit dem Zusatz GROUP BY wird die Anzahl der Fahrzeuge, die bei derselben Messreihe im gleichen Intervall und in der gleichen Messstation eintraten sind, addiert. Nach Ausführung dieser Prozedur kann Tabelle Messung mit Gangliniendaten gefüllt werden. Tabelle 6.3 stellt ein Beispiel für Tabelle Messung dar.

Flag_ID	MESSREIHE	MESSSTATION	SIMULATIONSdatum	TYP	INTERVALL	MENGE
1	41	64	04-Mai-2004	1	4	485
1	41	64	04-Mai-2004	1	5	495
...
2	0	24	03-Mai-2004	1	20	26
2	0	24	03-Mai-2004	1	21	38
...
3	0	58	04-Mai-2004	1	2	54
3	0	58	04-Mai-2004	1	3	45

Tabelle 6.3 Beispiel für Tabelle Messung

In Tabelle Messung bezeichnet Messreihe einen Simulationslauf mit dem Simulationsdatum. Messstation wird durch die Nummer der Messstation dargestellt. Typ und Intervall stellen die Zeitreihe dar. Verkehrsmenge fasst die Anzahl von Fahrzeugen zusammen, die bei demjenigen Simulationslauf innerhalb einer bestimmten Zeitreihe durch eine bestimmte Messstation gefahren sind. Messreihe, Messstation, Simulationsdatum, Typ und Intervall bilden zusammen die Primärschlüssel der Tabelle. TYP und Intervall zusammen stellen die Fremdschlüssel für Tabelle Zeitraster dar. Flag ist die Fremdschlüssel von Tabelle Flag. Die

Verkehrsmenge mit Flag_ID 2 bezeichnet die Durchschnittswerte aus Messreihen 4, 21, 41 und 66. Flag_ID 3 stellt die Gewichtswerte nach Gewichtfaktoren. Hier im Beispiel gilt die Regel, jüngste Messreihe hat der größte Faktor. Nämlich Messreihe 4 hat Faktor 0.1, Messreihe 21 hat Faktor 0.2, Messreihe 41 hat Faktor 0.3 und Messreihe 66 hat Faktor 0.4. Für die Durchschnittswerte und Gewichtswerte werden die Messreihen mit dem Wert 0 eingefügt. Die Rechnungsprozesse wurden durch Prozeduren (in ORACLE 9i) realisiert.

6.8 Zusammenfassung

In diesem Kapitel wurde das physische Data Warehouse Schema für Erstellung von Ganglinien und Bewertung von Prognose im Verkehr vorgestellt. Dieses Schema verfolgt das Star Schema und besitzt eine Faktentabelle mit 4 Dimensionen. Als Datenquell wird eine Datenbanksicht auf Basis des Datenbankschemas in [Sand04] aufgebaut. Mit der Datenbanksicht wurden die Simulationsdaten erzeugt. Damit die Daten im Verkehr als Ganglinien dargestellt werden, wird Dimension Zeitraster eingeführt. In Dimension Zeitraster werden die Tagszeit in gewünschten Intervalltype aufgeteilt. Um die verschiedenen Arten von Messdaten in diesem Schema integriert zu behandeln, wird die Dimension Flag eingesetzt. Die Messdaten werden durch Prozeduren in ORACLE 9i erzeugt und in Faktentabelle Messung gefüllt. Mit diesem Schema können Verkehrsdaten wie z.B. Simulationsdaten, Prognosewerten, erwartete Daten in Form von Ganglinien erstellt, gespeichert und für weitere Bewertung zur Verfügung gestellt. Alle obengenannte Prozesse wurden mit verschiedenen Prozeduren in ORACLE 9i realisiert. Damit ist die Realisierbarkeit des Data Warehouse Schemas bewiesen.

7. OLAP Einsatz

In diesem Kapitel wird das in dieser Arbeit verwendete OLAP-System Cognos Series 7 vorgestellt. Dabei wird erläutert, wie mit OLAP-Systemen auf die Daten des implementierten Data Warehouse zu zugreifen und wie Ganglinien-Abfrage zu erstellen.

7.1 Einführung Cognos Series 7

Cognos Series 7 besitzt umfangreiche OLAP-Funktionalitäten und besteht aus vielen Softwarekomponente, wie z.B. Cognos PowerPlay, Cognos PowerPlay Web, Cognos Query, Cognos Architect, Cognos Transformer Edition usw.. Zur Ganglinie-Abfrage werden in dieser Arbeit Cognos PowerPlay, Cognos Architect and Cognos Transformer verwendet. In diesem Abschnitt werden die drei Komponenten sowie deren Beziehungen skizziert.

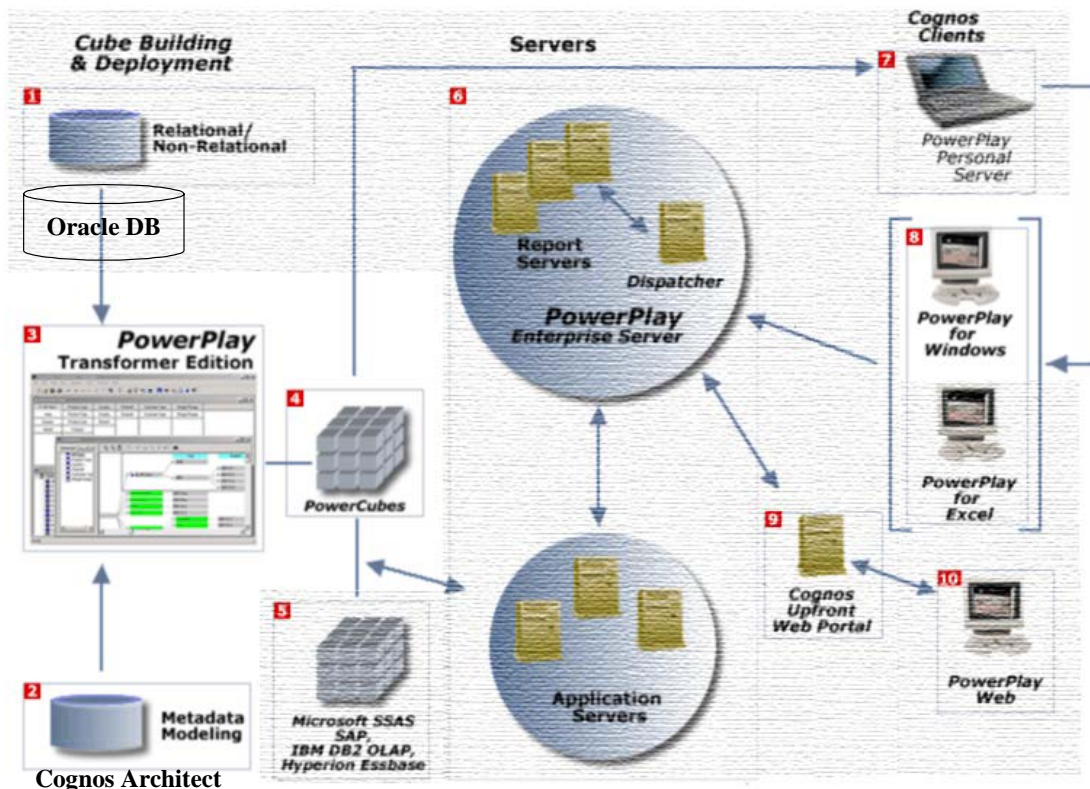


Abbildung 7.1 Überblick über Cognos Series 7 (adaptiert aus Cognos Webseite)

Abbildung 7.1 stellt einen Überblick über Cognos Series 7 dar. Die Komponente, auf die sich diese Arbeit nicht direkt bezieht, werden in Abbildung 7.1 verdunkelt dargestellt. In dieser Arbeit werden die folgenden Komponenten von Cognos Series 7 verwendet:

- **Oracle DB als Data Mart:** Data Mart stellt die Datenquellen für ein OLAP-System dar. In dieser Komponente werden verschiedenen Softwarekomponenten für Datenextraktion,

Transformation und Datenladen zur Verfügung gestellt, damit die zu analysierenden Daten sowohl aus virtuellen als auch aus physischen Systemen geladen werden können. Anschließend werden die Daten in Form multidimensionaler Datenmodelle dargestellt und für die weitere Analyse bereitgestellt. In dieser Arbeit werden die in Kapitel 6 bereitgestellten Datenbanktabellen (in Oracle 9i) als Data Mart eingesetzt.

- **Cognos Architect:** Mit Cognos Architect werden die Metadaten für die zu analysierenden Daten an einer zentralen Stelle definiert und verwaltet. Die Metadaten werden nachher bei der Erzeugung von PowerCube sowie bei der Darstellung von Ganglinien verwendet.
- **Cognos PowerCube:** Cognos PowerCube dient als die Datenhaltung für Analyse und unterstützt unter anderem das Star Schema. Kontinuierlich werden Daten aus Datenquellen wie z.B. relationalen Datenbanken hochgeladen, um sie zu einem oder mehreren PowerCubes hinzuführen. PowerCubes ermöglichen dank ihres multidimensionalen Datenmodells, gewünschte Datenabfrage schnell und bequem zusammenzustellen und zu vielfältigen Ausgaben aufzubereiten.
- **Cognos Transformer Edition:** Mit Cognos Transformer Edition werden PowerCubes modelliert und erzeugt.
- **Cognos PowerPlay (for Windows):** Über Cognos PowerPlay for Windows können die Ergebnisse von Datenabfragen visuell dargestellt oder an anderen Anwendungen weitergeleitet werden. Dabei kann Analyse auf jede beliebige Detailebene durchgeführt werden und viele verschiedene Arten von OLAP-Berichten können anschließend erzeugt werden. Mit Cognos PowerPlay for Excel kann man auf multidimensionale Daten über Microsoft Excel zugreifen. Cognos Serie 7 liefert außerdem eine einfache und leistungsfähige Benutzeroberfläche, die auf Internet Explorer basiert und die neueste Browser-Technologie nutzt.

7.1.1 Cognos Architect

Cognos Architect ist ein Tool für Metadaten Management (siehe [CognosArch]). Als Basis für Cognos Series 7 unterstützt Cognos Architect alle Report- und Analyseprodukte. Mit Cognos Architect werden die nötigen Metadaten für ein Analysethema bzw. Analysezwecke zusammen in Form eines sogenannten Cognos Architect Modell dargestellt. Cognos Architect Modell liefert eine businessorientierte zentrale Lokation für Metadaten von Data Warehouse sowie eine wiederbenutzbare Umgebung für alle Analyseverfahren. Cognos Architect Modell werden von vielen anderen Komponenten verwendet. Beispielweise verwendet Impromptu ein Cognos Architect Modell, um Katalogen zu erzeugen. PowerPlay Transformer Edition nutzt Cognos Architect Modell bei Modellierung von PowerCubes.

Durch eine dreistufige Architektur isoliert Cognos Architect die Geschäftsregeln von den Datenquellen und den Endanwendungen. Von einer zentralen Stelle aus können Metadaten und Geschäftsregeln entwickelt und verwaltet werden. Die dreistufige Architektur sieht wie folgenden aus:

- **Data Access Layer:** In Data Access Layer werden alle wichtigen Metadaten über die Datenerfassung aus verschiedenen Systeme angelegt. Tabelle, Spalten und andere Datenobjekte werden in Data Access Layer definiert. Außerdem kann man SQL-Anweisungen sowie benutzerdefinierte Funktionsmodule auch in dieser Layer anlegen. Über Data Access Layer können Daten aus verschiedenen Quellen wie z.B. Tabelle von

relationalen Datenbank, eingebettete Prozeduren eines Datenbanksystems oder SQL-Anweisungen geladen werden.

- **Business Layer:** In Business Layer werden Datensichten für Berichte und Query definiert. Dabei wird insbesondere das sogenannte Business Modell erstellt. Ein Business Modell enthält verschiedene Objekte wie z.B. Entity und Attribute. Die Objekte und deren Beziehungen können ebenfalls in Business Modell definiert werden. Business Modell wird ähnlich wie E/R-Diagramm dargestellt werden. Außerdem kann man die Geschäftslogik in Form von Kalkulation und Filter darstellen. Für die Darstellung der Daten in verschiedenen Sichten können sogenannte Anzeigeregeln angelegt werden. Die Objekte in Business Layer werden anhand von den Objekten in Data Access Layer erzeugt. Durch die Erstellung der Business Layer können die Quellen und Darstellung von Daten getrennt werden, damit kann mehrere Flexibilität für Berichten und Query gewährleistet werden.
- **Package Layer:** In der Package Layer werden eine Menge von Objekten in Business Layer als eine Package zusammen gepackt. Eine Package bezieht sich normalerweise auf ein spezielles Cognos Produkt wie z.B. Cognos PowerPlay, Cognos Transformer Edition, Cognos Impromptu oder Cognos Query. Für unterschiedliche Benutzer oder Zwecke werden unterschiedliche Package erzeugt und gespeichert.

7.1.2 Cognos Transformer Edition

Cognos Transformer Edition ist ein Modellierungswerkzeug für die Erstellung von Cognos PowerCubes (siehe [CognosTran]). Um PowerCubes zu erstellen, sollen Quelldaten zuerst ausgewertet werden. Jedes in Cognos PowerPlay verwendete PowerCube basiert auf ein Modell, das mit Cognos Transformer Edition definiert ist. Die Erstellung verfolgt mit Angabe von Datenquellen und Definition von Dimensionen und Kennzahlen.

7.1.3 Cognos PowerPlay

Cognos PowerPlay enthält umfassende Funktionen für Analyse, die im folgenden zusammengefasst werden (siehe [CognosPow]):

- Visualisieren von Leistungsindikatoren: Leistungsindikatoren sind die Kennzahlen, die für Bewertungen wichtig sind.
- Untersuchen der Auswirkungen: Mit Cognos PowerPlay können die Geschäftsinformationen aus multidimensionalem Sicht detaillierten analysiert werden, wodurch die Entscheidung für Geschäfts beeinflusst
- Analysieren und zusammenfassen die Daten in Berichten: Mit Cognos PowerPlay können für die zu analysierenden Daten viele verschiedene Operationen, wie z.B. Zerlegen, Filtern, Ändern von Darstellungen, Bestimmen des prozentualen Anteils, Rangordnung, Hervorheben von Ausnahmen, Berechnen usw., durchgeführt werden.
- Darstellung der Analyseergebnisse: Analyseergebnisse können in vielen Formen wie z.B. Kreuztabellen, Kreisdiagramm, Balkendiagramm und Liniendiagramm angezeigt werden.

7.2 Implementierung mit Cognos PowerPlay

In dieser Arbeit werden überwiegend Cognos PowerPlay, Cognos Architect and Cognos Transformer Edition eingesetzt, um die Ganglinie-Abfrage über OLAP-Werkzeuge zu realisieren.

7.2.1 Erstellung Cognos Architect Modell

Wie in Abschnitt 7.1.1 vorgestellt wurde, dient Cognos Architect als ein Metadaten Management Werkzeug. Ein Architect Modell ist die zentrale Lokation für Metadaten, die wiederbenutzbar für weitere Funktionen ist. Für die Darstellung und Abfragen von Ganglinien wird ein Cognos Architect Modell in drei Layers wie folgt definiert:

- **Data Access Layer:** In dieser Layer werden Datenbanktabelle MESSUNG, FLAG und ZEITRASTER aufgenommen.
- **Business Layer:** für jede Tabelle im Date Access Layer wird eine Entity im Business Layer erzeugt. Jede Spalte in der Tabelle wird als ein Attribut betrachtet. Die Schlüssel in Data Access Layer bleiben in Business Layer erhalten. Die Beziehung zwischen die aufgenommenen Tabellen bzw. Entities wird durch Abbildung 7.2 dargestellt.

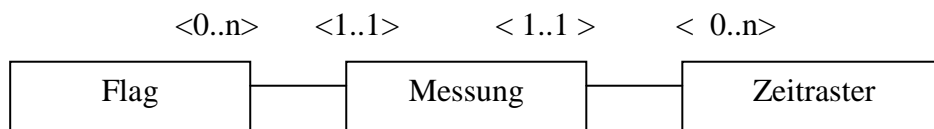


Abbildung 7.2 Beziehung zwischen Objekten in Business Layer

Die Kardinalität <1..1> zwischen Messung und Zeitraster besagt, dass jede Tulpe von Messung ist genau mit einer Tulpe von Zeitraster verknüpft. Umgekehrt besagt die Kardinalität <0..n> zwischen Zeitraster und Messung, das sich jede Tulpe von Zeitraster auf keine oder mehre Tulpen in Messung beziehen. Die Beziehung zwischen Messung und Flag wird analog aufgebaut.

- **Package Layer:** Nachdem alle Objekte in Business Layer erzeugt bzw. definiert sind, werden sie in einer Package eingepackt. Bevor eine Package erzeugt ist, führt das System eine Überprüfung durch. Die Package wird dann in Cognos Transformer Edition bei Definition von PowerCube benutzt.

7.2.2 Erstellung PowerCube mit Cognos Transformer Edition

Cognos Transformer Edition ist eine OLAP Design Werkzeuge, mit dem ein Transformer Modell erzeugt werden kann. In einem Transformer Modell wird ein PowerCube definiert. Bei Definition von PowerCube in Cognos Transformer wird die in Cognos Architect definierte Package als Datenquelle benutzt. Die folgenden Objekte werden in Cognos Transformer Edition definiert:

- **Kennzahl:** Fahrzeugmenge aus Relation Messung wird hier als Kennzahl definiert.
- **Dimensionen:** Die verwendeten Dimensionen sind Flag, Messreihe, Messstation und Zeitraster.

- PowerCube: Mit definierter Kennzahl Fahrzeugmenge und die vier Dimensionen wurde eine vier dimensionale PowerCube erzeugt.

7.2.3 Ganglinien-Abfragen mit Cognos PowerPlay

Mit Cognos PowerPlay for Windows können die Daten in PowerCube ausgelesen und die Analyse durchgeführt werden. Cognos PowerPlay ist ein powervolles Analysewerkzeug, mit dem schwierige Abfragen direkt gebildet und schnell in vielen Formen wie z.B. Kreuztabellen, Kreisdiagramm, Balkendiagramm und Liniendiagramm angezeigt werden können.

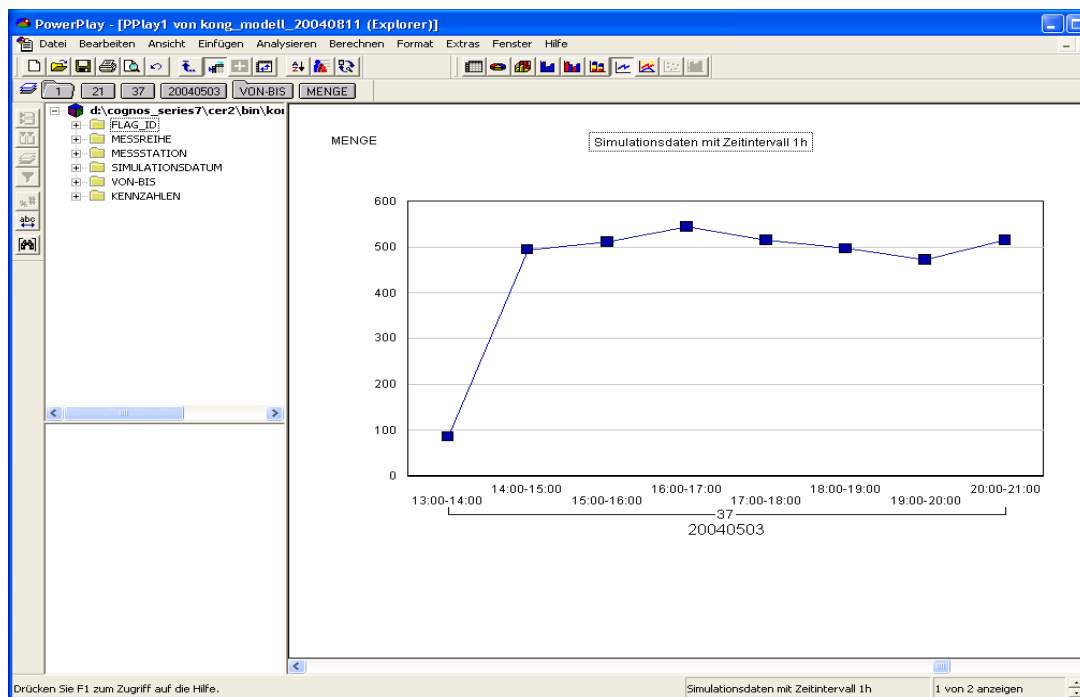


Abbildung 7.3 Erstes Beispiel für Ganglinien-Abfragen

Anwendung von Cognos PowerPlay ist nahezu intuitiv. Mit einigen Mausklicken werden die zugrundeliegenden Daten aus PowerCube untersucht und analysiert. Abbildung 7.3 stellt ein Beispiel für Ganglinien-Abfrage dar. Das Beispiel zeigt eine Ganglinie durch ein Liniendiagramm an. Das Diagramm bezeichnet die Verkehrssituation bei Messstation 37 aus Messreihe 21 am 03.05.2004. Während die x-Achse eine Zeitreihe darstellt, zeichnet die y-Achse die Fahrzeugmenge auf. Aus dem Diagramm ist die Verkehrstendenz leicht zu erfassen.

Zweites Beispiel stellt Abbildung 7.4 die Fahrzeugmenge aus allen Messstationen mit 1 Stunden Zeitintervall dar. Diese Fahrzeugmenge sind erwartete Werte von Messreihe 4, 21, 41, 66 mit jeweils Gewichtungsfaktoren 0,1, 0,2, 0,3, 0,4. Abbildung 7.4 stellt einen Ausschnitt von einer Cognos Kreuztabelle dar. Die zweite Zeile stellt ganze Zeitreihe von einem Tag dar und umfasst die zweite Spalte alle Messstation.

7.3 Zusammenfassung

Dieser Kapitel beschäftigt sich mit dem Einsatz von OLAP-System Cognos Series 7 für die Ganglinien-Abfragen. Cognos Series 7 ist ein umfassendes Softwareprodukt und besteht aus

vielen Funktionen. Anhand von Aufgaben dieser Arbeit werden hauptsächlich Cognos Architect, Cognos Transformer Edition und Cognos PowerPlay verwendet. Mit Cognos Architect wurden die Metadaten erstellt. Ein PowerCube wurde mit Cognos Transformer Edition definiert. Über Cognos PowerPlay wurden Ganglinien-Abfragen erstellt.

The screenshot shows the Cognos PowerPlay interface with a pivot table. The table displays 'Messdaten nach Gewichtsfaktor' (Measurement data by weight factor) for '20040503'. The rows represent weight factors from 21 to 42, and the columns represent time intervals from 13:00-14:00 to 18:00-19:00. The values represent the quantity of measurements in each category.

		20040503					
		13:00-14:00	14:00-15:00	15:00-16:00	16:00-17:00	17:00-18:00	18:00-19:00
Messdaten nach	21	72	488	515	475	468	464
Gewichtsfaktor	22	14	99	115	108	99	104
4(0.1),	23	42	203	178	202	212	221
21(0.2), 41(0.3),	24	12	36	36	40	39	45
66(0.4)	25	78	518	536	506	484	488
	26	9	68	90	81	78	79
	27	50	239	212	239	249	263
	28	0	4	2	0	3	3
	29	111	732	729	758	734	705
	30	13	117	86	102	122	103
	31	41	236	281	238	261	245
	32	90	522	527	486	528	500
	33	70	426	423	446	429	423
	34	21	135	135	159	153	117
	35	25	135	167	132	156	140
	36	84	493	517	473	501	486
	37	77	445	459	490	464	447
	38	24	140	126	141	142	122
	39	16	67	87	74	66	63
	40	115	695	742	686	717	711
	41	66	396	392	410	371	358
	42	99	522	481	504	497	508

Abbildung 7.4 Zweites Beispiel für Ganglinien-Abfrage

8. Zusammenfassung

In der Arbeit wurde unter anderem ein Data Warehouse Schema für Erstellung von Ganglinien und Bewertung von Prognose im Verkehr entwickelt. Mit diesem Schema können die Daten im Verkehr wie z.B. Simulationsdaten, Prognosewerten, erwartete Daten in Form von Ganglinien erstellt, gespeichert und weiterhin bewertet.

In der Entwurfsphase wurde zuerst ein multidimensionales Datenmodell entwickelt, das Datenmodell stellt die grundlegenden Ideen auf die konzeptuelle Ebene dar. Das Datenmodell wurde gemäß des mE/R-Diagramms dargestellt. Im Mittelpunkt des Datenmodells steht die Entität Messung, die durch eine Kennzahl Fahrzeugmenge gekennzeichnet wird. Vier Dimensionen Zeit, Ort, Messverfahren und Simulationslauf wurden definiert, um die Kennzahl Fahrzeugmenge umfassend zu charakterisieren.

Ein physisches Data Warehouse Schema wurde dann auf Basis des konzeptuellen Datenmodells entwickelt. Die Komponenten des Schemas sind durch eine Transformation von mE/R Datenmodell zum Star Schema erzeugt worden. Die vier wichtige Dimensionen, welche eine Messung charakterisieren, sind jeweils durch eine Tabelle im physischen Schema dargestellt. In der Faktentabelle befindet sich die Kennzahl Fahrzeugmengen, welche durch die vier Dimensionen eindeutig beschrieben ist. Statt einer Tabelle für Dimension Verfahren wurde Tabelle Flag eingefügt, um mehrere Arten von Daten zu integrieren. Durch die Integration können Simulationsdaten, Prognosedaten und erwartete Daten gleichzeitig als Ganglinien dargestellt und verglichen werden. Durch die in ORACLE 9i implementierten Prozeduren, werden die Faktentabelle sowie Dimensionstabellen erzeugt und mit Daten eingetragen.

Um die Ganglinien-Abfrage zu erstellen, wurde OLAP-Produkt Cognos Series 7 eingesetzt. In dieser Arbeit wurden hauptsächlich Cognos Architect, Cognos Transformer Edition und Cognos PowerPlay verwendet. Mit Cognos Architect wurden die nötigen Metadaten in Data Access Layer, Business Layer und Package Layer erstellt. Ein PowerCube wurde dann mit Cognos Transformer Edition definiert. Über Cognos PowerPlay wurden verschiedene Ganglinien-Abfragen erstellt und die Ganglinien in verschieden Form angezeigt.

Dieser Arbeit konzentriert sich auf Erzeugung von Gangliniendarstellung im Verkehr, damit Verkehrsprognosen unterstützt and bewertet werden können. Im Verkehr sind viele Arten von Daten vorhanden. Einer der Vorteile des entwickelten Data Warehouse Schemas besteht darin, dass dieses Schema ermöglicht, alle solchen Daten in einem System zu speichert, zu bearbeiten.

Für Gangliniendarstellung können die gesammelten Fahrzeugmengen (in Form von Simulationsdaten, erwarteten Daten und Prognosewerten) in dem Data Warehouse Schema mit kontinuierlicher Zeitreihe zusammen erfassen. In der Arbeit wurden mit drei Intervallen untersucht, nämlich 1 Stunde, 30 Minuten und 15 Minuten. Mit anderen Intervallen können dann ähnliche Untersuchungen durchgeführt werden.

In dieser Arbeit wurden Durchschnittswerte und Gewichtswerte mit gegebenen Faktoren als erwarteten Daten berechnet und gespeichert. Weitere Arten von erwarteten Daten können mit diesem Data Warehouse Schema analog bearbeitet werden.

Mit diesem Schema können die Prognosewerte aus anders Schema auch in diesem System integriert werden. Diese Integration verschiedener Arten von Daten ermöglicht, Verkehrsprognosen leicht zu bewerten.

Literatur

- [BaGü01] Bauer, A.; Günzel, H.: Data Warehouse Systeme: Architektur, Entwicklung, Anwendung; 1. Auflage; 2001.
- [CognosArch] Cognos Architect, Model with Architect, 2003 Cognos Inc.
- [CognosTran] Cognos PowerPlay, Entdecken Sie Transformer, 2003 Cognos Inc.
- [CognosPow] Cognos PowerPlay for Windows, Entdecken Sie PowerPlay, 2003 Cognos Inc.
- [KBB01] Statistische Mitteilungen: Güterkraftverkehr deutscher Lastkraftfahrzeuge, Kraftfahrt-Bundesamt und Bundesamt für Güterverkehr, 2001.
- [Koop04] Erik Alexander Koop, Datenbankunterstützung für imperfekte Daten im Verkehrsumfeld, Diplomarbeit IPD, Fakultät für Informatik, Universität Karlsruhe 2004
- [Kurz99] Andreas Kurz: Data Warehousing: Enabling Technology, 1. Auflage 1999.
- [Merk04] Andreas Merkel, Konzeption und Umsetzung einer Schemaerweiterung durch Kontexte unter Berücksichtigung von Aggregierungsanfragen, Studienarbeit IPD, Fakultät für Informatik, Universität Karlsruhe, Mai 2004.
- [Sand04] Jan Sandberger, Konzeption und Evaluation eines Data Warehouse zur Analyse von Daten im Verkehrsbereich, Studienarbeit IPD, Fakultät für Informatik, Universität Karlsruhe, 2004.
- [StWi03] Frank Steinert und Ralf Willenbrock, Dokumentation zum Projekt "Bereitstellung und Consulting Floating-Car-Daten", gedas Deutschland GmbH, Version 1.0, 2003.
- [SBHD98] Sapia, C., Blaschka, M., Höfling, G., Dinter, B.: extending the E/R Model for the Multidimensional Paradigm. In Kambayashi, Y. et. Al.(Hrsg.), Advances in Database Technologies, LNCS Vol. 1552, Springer, 1999.
- [Wild96] Dieter Wild, Die Prognose von Verkehrsstärken anhand klassifizierter Ganglinien, (Bericht aus der Informatik, Zugl.: Karlsruhe, Univ., Diss., 1996), Aachen: Shaker, 1996.
- [ZaHe80] H. Zackor, S. Herkt, Kurzzeitprognose von Verkehrsströmen auf der Grundlage aktueller Querschnittsmessungen, in Forschung Straßenbau und Straßenverkehrstechnik, herausgegeben vom Bundesminister für Verkehr, Abteilung Straßenbau, Bonn Bad Godesberg, Heft 313, 1980.