Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Informationsintegration

ETL & Datenherkunft (*Lineage*)

28.6.2007

Felix Naumann

Einschub von Peter Bunemann

- Vortrag: The two cultures of digital curation – why databases matter!
- Berlin 2005

Provenance – an old problem

3

W.V. Quine. "Words enough…" New York Review of Books XIII(10):3-4, 1969

A well-known encyclopedia, following tradition, incorrectly describes Monaco as having an area of "8 square miles."

A new edition using adds "...the length being 2 1/4 miles and the width varying from 165 to 1100 yards."

An editor, spotting the inconsistency, removes the *correct* information from the subsequent edition.

Quelle: Peter Bunemann

The area of Monaco today

4

Most sources	1.95 sq km
www.atlapedia.com	1.94 sq km
military.countrywatch.com	2 sq km

Most sources	1.95 sq km	0.75 sq m
www.state.gov	1.95 sq km	0.8 sq m
www.atlapedia.com	1.94 sq km	1 sq m

(1.95 sq km = 0.753 sq m)

Quelle: Peter Bunemann

Atlapedia online

Countries A to Z
World Maps
Search
Class Resources

OFFICIAL NAME: Principality of Monaco
CAPITAL: Monaco-Ville
SYSTEM OF GOVERNMENT: Constitutional Monarchy
AREA: 1.94 Sq Km (1 Sq Mi)
ESTIMATED 2000 POPULATION: 32,500

LOCATION & GEOGRAPHY: Monaco is located on the Mediterranean Riviera, close to the Italian border with France and is surrounded by the French department of Alpes-Maritimes. Monaco is the second smallest country in the world and the principality has four distinct divisions. (1.) La Condamine, the business district. (2.) The Casino or Monte Carlo. (3.) Monaco-Ville which is on a rocky promontory and (4.) Fontvieille.

CLIMATE: Monaco has a Mediterranean climate which is characterized by warm summers and mild winters. Rainfall is light with an average annual precipitation of about 730 mm (29 inches) while sunshine usually lasts 300 days a year without rainfall. Average temperature ranges are from 8 to 12 degrees Celsius (46 to 54 degrees Fahrenheit) in January to 22 to 26 degrees Celsius (72 to 79 degrees Fahrenheit) in August.

PEOPLE: Around 18% of the population are Monegasque of Rhaetian descent. The remainder are foreigners which include the French who account for 38% of the population while 17% are Italian.

RELIGIONS: Mostly Christians with 95% of the population Roman Catholic while 5% are Protestant.

LANGUAGES: The official language is French, although Monegasque, a mixture of French and Italian is also spoken. English is also widely spoken.

MODERN HISTORY - WWII TO 1993: In 1949 Prince Rainier III succeeded his grandfather Prince Louis. In Apr. 1956 he married Grace Kelly, a famous American movie star, who took the title of Princess Grace and they had two daughters as well as a son, Prince Albert the heir apparent. In 1959 Prince Rainier dissolved the National Council and in 1962 under pressure from France, the Prince restored the National Council. In the same year, Prince Rainier III granted Monaco a new constitution that gave women the right to vote and abolished the death penalty. In 1963 again under pressure form France, Monaco signed a convention which placed certain Monaco based companies under French fiscal law. Since the early 1980's Monaco has reclaimed land from the sea for the development of new beaches. In Sept. 1982 Princess Grace died in an automobile accident and in Oct. 1990 Stefano Casiraghi, the husband of Princess Caroline's second marriage was killed in a speedboat accident. On March 19, 1991 the Conseil Communal elected Anne-Marie Campora mayor of Monaco-Ville to replace Jean-Louis Medecin who had held that position since 1971. In Oct. 1991 Monaco was elected to the committee of the Mediterranean Action Plan which consists of 18 nations working together for the protection of the Mediterranean

Fertig

The population of Monaco today?

HPI Hasso Plattner Institut

6

2004 35,000		http://www.cybevasion.fr/tourisme/monaco.html
2004 33,300		http://www.internetworldstats.com/europa2.htm
2004 32,270	(July 2004 est.)	http://www.cia.gov/cia/publications/factbook/geos/mn.html
2004 32,000		http://www.studentsoftheworld.info/pageinfo_pays.php3?Pays=MCO
2004 29,972		http://worldatlas.com/webimage/countrys/europe/mc.htm
2003 32,130	(July 2003 est.)	http://www.greenfacts.org/studies/climate_change/index.htm
2003 32,130	(mid 2003)	http://www.infoplease.com/ipa/A0004379.html
2003 32,000	(July 2003 estimate)	http://www.gesource.ac.uk/worldguide/html/962_people.html
2003 30,000		http://www.tlfq.ulaval.ca/axl/europe/monaco.htm
2002 31,987	(July 2002 est.)	http://www.greekorthodoxchurch.org/wfb2002/monaco/monaco_people.html
2001 31,842	(July 2001 est.)	http://wonderclub.com/Atlas/mccia.htm
2001 31,842	(July 2001 est.)	http://www.worldfactsandfigures.com/countries/monaco.php
2001 31,842	(July 2001 est.)	http://www.workmall.com/wfb2001/monaco/monaco_people.html
2000 32,500	(est 2000)	http://www.atlapedia.com/online/countries/monaco.htm
2000 32,020	(C 2000-05-03)	http://www.citypopulation.de/Monaco.html
2000 32,020	(2000).	http://www.worldtravelguide.net/data/mco/mco.asp
2000 32,020	(2000 census[!])	http://www.state.gov/r/pa/ei/bgn/3397.htm
2000 31,693	(July 2000 est.)	http://geography.about.com/library/cia/blcmonaco.htm
2000 31,693	(July 2000 est.)	http://www.abacci.com/atlas/demography.asp?countryID=269
2000 31,693	(July 2000 est.)	http://www.mapquest.com/atlas/?region=monaco
2000 31,842		http://www.fact-index.com/m/mo/monaco.html
2000 31,842		http://en.wikipedia.org/wiki/Monaco
2000 31,700	(e2000m)	http://www.library.uu.nl/wesp/populstat/Europe/monacoc.htm
1999 32,149	(July 1999 est.)	http://www.photius.com/wfb1999/monaco/monaco_people.html
1999 32,000		http://geography.about.com/library/weekly/aa012599.htm
1990 29,972	(1990 census)	http://www.monte-carlo.mc/us/presentation/keyfigur/

Quelle: Peter Bunemann

The area and population of Monaco today

7

- The 2004 figures vary by 17% !!
- Most sites copy from the CIA world fact-book
  - The US state department does not – and contradicts itself!
- Only two sites give attribution.
- No evidence is given for how the estimates were made.
- The last census appears to have been taken in 1990!

Quelle: Peter Bunemann

Swissprot: a curated database

8

```
ID   11SB_CUCMA     STANDARD;      PRT;   480 AA.
AC   P13744;
DT   01-JAN-1990 (REL. 13, CREATED)
DT   01-JAN-1990 (REL. 13, LAST SEQUENCE UPDATE)
DT   01-NOV-1990 (REL. 16, LAST ANNOTATION UPDATE)
DE   11S GLOBULIN BETA SUBUNIT PRECURSOR.
OS   CUCURBITA MAXIMA (PUMPKIN) (WINTER SQUASH).
OC   EUKARYOTA; PLANTA; EMBRYOPHYTA; ANGIOSPERMAE; DICOTYLEDONEAE;
OC   VIOLALES; CUCURBITACEAE.
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CV. KUROKAWA AMAKURI NANKIN;
RX   MEDLINE; 88166744.
RA   HAYASHI M., MORI H., NISHIMURA M., AKAZAWA T., HARA-NISHIMURA I.;
RL   EUR. J. BIOCHEM. 172:627-632(1988).
RN   [2]
RP   SEQUENCE OF 22-30 AND 297-302.
RA   OHMIYA M., HARA I., MASTUBARA H.;
RL   PLANT CELL PHYSIOL. 21:157-167(1980).
CC   -!- FUNCTION: THIS IS A SEED STORAGE PROTEIN.
CC   -!- SUBUNIT: HEXAMER; EACH SUBUNIT IS COMPOSED OF AN ACIDIC AND A
CC       BASIC CHAIN DERIVED FROM A SINGLE PRECURSOR AND LINKED BY A
CC       DISULFIDE BOND.
CC   -!- SIMILARITY: TO OTHER 11S SEED STORAGE PROTEINS (GLOBULINS).
DR   EMBL; M36407; G167492; -.
DR   PIR; S00366; FWPU1B.
DR   PROSITE; PS00305; 11S_SEED_STORAGE; 1.
KW   SEED STORAGE PROTEIN; SIGNAL.
FT   SIGNAL        1     21
FT   CHAIN        22    480       11S GLOBULIN BETA SUBUNIT.
FT   CHAIN        22    296       GAMMA CHAIN (ACIDIC).
FT   CHAIN       297    480       DELTA CHAIN (BASIC).
FT   MOD_RES      22     22       PYRROLIDONE CARBOXYLIC ACID.
FT   DISULFID    124    303       INTERCHAIN (GAMMA-DELTA) (POTENTIAL).
FT   CONFLICT     27     27       S -> E (IN REF. 2).
FT   CONFLICT     30     30       E -> S (IN REF. 2).
SQ   SEQUENCE   480 AA;  54625 MW;  D515DD6E CRC32;
     MARSSLFTFL CLAVFINGCL SQIEQQSPWE FQGSEVWQQH RYQSPRACRL ENLRAQDPVR
     RAEAEAIFTE VWDQDNDEFQ CAGVNMIRHT IRPKGLLLPG FSNAPKLIFV AQGFGIRGIA
     IPGCAETYQT DLRRSQSAGS AFKDQHQKIR PFREGDLLVV PAGVSHWMYN RGQSDLVLIV
     FADTRNVANQ IDPYLRKFYL AGRPEQVERG VEEWERSSRK GSSGEKSGNI FSGFADEFLE
     EAFQIDGGLV RKLKGEDDER DRIVQVDEDF EVLLPEKDEE ERSRGRYIES ESESENGLEE
     TICTLRLKQN IGRSVRADVF NPRGGRISTA NYHTLPILRQ VRLSAERGVL YSNAMVAPHY
     TVNSHSVMYA TRGNARVQVV DNFGQSVFDG EVREGQVLMI PQNFVVIKRA SDRGFEWIAF
     KTNDNAITNL LAGRVSQMRM LPLGVLSNMY RISREEAQRL KYGQQEMRVL SPGRSQGRRE
//
```


Quelle: Peter Bunemann

9

```
ID   11SB_CUCMA     STANDARD;      PRT;   480 AA.
AC   P13744;
DT   01-JAN-1990 (REL. 13, CREATED)
DT   01-JAN-1990 (REL. 13, LAST SEQUENCE UPDATE)
DE   11S GLOBULIN BETA SUBUNIT PRECURSOR.
OS   CUCURBITA MAXIMA (PUMPKIN) (WINTER SQUASH).
OC   EUKARYOTA; PLANTA; EMBRYOPHYTA; ANGIOSPERMAE; DICOTYLEDONEAE;
OC   VIOLALES; CUCURBITACEAE.
RC   STRAIN=CV. KUROKAWA AMAKURI NANKIN;
RX   MEDLINE; 88166744.
RA   HAYASHI M., MORI H., NISHIMURA M., AKAZAWA T., HARA-NISHIMURA I.;
CC   -!- FUNCTION: THIS IS A SEED STORAGE PROTEIN.
CC   -!- SUBUNIT: HEXAMER; EACH SUBUNIT IS COMPOSED OF AN ACIDIC AND A
CC       BASIC CHAIN DERIVED FROM A SINGLE PRECURSOR AND LINKED BY A
CC       DISULFIDE BOND.
CC   -!- SIMILARITY: TO OTHER 11S SEED STORAGE PROTEINS (GLOBULINS).
DR   EMBL; M36407; G167492; -.
FT   CHAIN        22    480       11S GLOBULIN BETA SUBUNIT.
FT   CHAIN        22    296       GAMMA CHAIN (ACIDIC).
FT   CHAIN       297    480       DELTA CHAIN (BASIC).
FT   MOD_RES      22     22       PYRROLIDONE CARBOXYLIC ACID.
FT   DISULFID    124    303       INTERCHAIN (GAMMA-DELTA) (POTENTIAL).
FT   DISULFID    124    303       INTERCHAIN (GAMMA-DELTA) (POTENTIAL).
FT   CONFLICT     27     27       S -> E (IN REF. 2).
FT   CONFLICT     30     30       E -> S (IN REF. 2).
SQ   SEQUENCE   480 AA;  54625 MW;  D515DD6E CRC32;
     MARSSLFTFL CLAVFINGCL SQIEQQSPWE FQGSEVWQQH RYQSPRACRL ENLRAQDPVR
     RAEAEAIFTE VWDQDNDEFQ CAGVNMIRHT IRPKGLLLPG FSNAPKLIFV AQGFGIRGIA
     IPGCAETYQT DLRRSQSAGS AFKDQHQKIR PFREGDLLVV PAGVSHWMYN RGQSDLVLIV
     FADTRNVANQ IDPYLRKFYL AGRPEQVERG VEEWERSSRK GSSGEKSGNI FSGFADEFLE
     EAFQIDGGLV RKLKGEDDER DRIVQVDEDF EVLLPEKDEE ERSRGRYIES ESESENGLEE
     TICTLRLKQN IGRSVRADVF NPRGGRISTA NYHTLPILRQ VRLSAERGVL YSNAMVAPHY
     TVNSHSVMYA TRGNARVQVV DNFGQSVFDG EVREGQVLMI PQNFVVIKRA SDRGFEWIAF
     KTNDNAITNL LAGRVSQMRM LPLGVLSNMY RISREEAQRL KYGQQEMRVL SPGRSQGRRE
//
```

Where does this information come from? Which editor? Or was it the cited papers?

Quelle: Peter Bunemann

10

```
ID   11SB_CUCMA     STANDARD;      PRT;   480 AA.
AC   P13744;
DT   01-JAN-1990 (REL. 13, CREATED)
DT   01-JAN-1990 (REL. 13, LAST SEQUENCE UPDATE)
DT   01-NOV-1990 (REL. 16, LAST ANNOTATION UPDATE)
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CV. KUROKAWA AMAKURI NANKIN;
RX   MEDLINE; 88166744.
RA   HAYASHI M., MORI H., NISHIMURA M., AKAZAWA T., HARA-NISHIMURA I.;
RL   EUR. J. BIOCHEM. 172:627-632(1988).
RN   [2]
RP   SEQUENCE OF 22-30 AND 297-302.
RA   OHMIYA M., HARA I., MASTUBARA H.;
RL   PLANT CELL PHYSIOL. 21:157-167(1980).
CC   -!- FUNCTION: THIS IS A SEED STORAGE PROTEIN.
CC   -!- SUBUNIT: HEXAMER; EACH SUBUNIT IS COMPOSED OF AN ACIDIC AND A
CC       BASIC CHAIN DERIVED FROM A SINGLE PRECURSOR AND LINKED BY A
CC       DISULFIDE BOND.
CC   -!- SIMILARITY: TO OTHER 11S SEED STORAGE PROTEINS (GLOBULINS).
DR   EMBL; M36407; G167492; -.
DR   PIR; S00366; FWPU1B.
DR   PROSITE; PS00305; 11S_SEED_STORAGE; 1.
KW   SEED STORAGE PROTEIN; SIGNAL.
FT   SIGNAL        1     21
FT   CHAIN        22    480       11S GLOBULIN BETA SUBUNIT.
FT   CHAIN        22    296       GAMMA CHAIN (ACIDIC).
FT   CHAIN       297    480       DELTA CHAIN (BASIC).
FT   MOD_RES      22     22       PYRROLIDONE CARBOXYLIC ACID.
FT   DISULFID    124    303       INTERCHAIN (GAMMA-DELTA) (POTENTIAL).
FT   CONFLICT     27     27       S -> E (IN REF. 2).
FT   CONFLICT     30     30       E -> S (IN REF. 2).
SQ   SEQUENCE   480 AA;  54625 MW;  D515DD6E CRC32;
     MARSSLFTFL CLAVFINGCL SQIEQQSPWE FQGSEVWQQH RYQSPRACRL ENLRAQDPVR
     RAEAEAIFTE VWDQDNDEFQ CAGVNMIRHT IRPKGLLLPG FSNAPKLIFV AQGFGIRGIA
     IPGCAETYQT DLRRSQSAGS AFKDQHQKIR PFREGDLLVV PAGVSHWMYN RGQSDLVLIV
     FADTRNVANQ IDPYLRKFYL AGRPEQVERG VEEWERSSRK GSSGEKSGNI FSGFADEFLE
     EAFQIDGGLV RKLKGEDDER DRIVQVDEDF EVLLPEKDEE ERSRGRYIES ESESENGLEE
     TICTLRLKQN IGRSVRADVF NPRGGRISTA NYHTLPILRQ VRLSAERGVL YSNAMVAPHY
     TVNSHSVMYA TRGNARVQVV DNFGQSVFDG EVREGQVLMI PQNFVVIKRA SDRGFEWIAF
     KTNDNAITNL LAGRVSQMRM LPLGVLSNMY RISREEAQRL KYGQQEMRVL SPGRSQGRRE
//
```

History of the entry is incomplete (only version numbers of last updates are kept.)

Quelle: Peter Bunemann

Two kinds of provenance

11

name	born	period
J.S. Bach	1685	baroque
G.F. Handel	1685	baroque
W.A. Mozart	1756	classical

```
SELECT name, born
FROM composer
```

```
SELECT name, born
FROM composer
WHERE born < SELECT AVERAGE born FROM composer
```

name	born
J.S. Bach	1685
. . .	…

Why is this element in the output?

Where does this element come from?

Quelle: Peter Bunemann

Example (stolen from Wang-Chiew Tan)

12

Serves fine French Cuisine in elegant setting. Jackets required.

Extensive wine list!

Yummy chicken curry!!

NYRestaurants (Source Table)

Restaurant	Cost	Type	Zip
Peacock Alley	$$$	French	10022
Bull & Bear	$$$	Seafood	10022
Pacifica	$	Chinese	10013
Soho Kitchen & Bar	$	American	10022

All Restaurants (View 1)

Restaurant	Cost	Type
Peacock Alley	$$$	French
Bull & Bear	$$$	Seafood
Pacifica	$	Chinese
Soho Kitchen & Bar	$	American

Cheap Restaurants (View 2)

Restaurant	Cost	Type
Pacifica	$	Chinese
Soho Kitchen & Bar	$	American

Quelle: Peter Bunemann

Überblick

13

- Motivation und Beispiel
- Datentransformationen (jeweils Definition und Tracing Prozedur)
  - Dispatcher
  - Aggregatoren
  - Black Boxes & MISOs
  - [Inverse Transformation]
- Transformationssequenzen
- Data Lineage nach [CW03]

Data Lineage – Motivation

14

- Data Lineage
  - Data Lineage ist das Problem, zu Objekten im integrierten System diejenigen Objekte in den Quellen zu bestimmen, aus denen das integrierte Objekt abgeleitet wurde.
  - Auch: Data Provenance
  - Auch: Data Pedigree
- Data Warehouses
  - Datenanalyse
  - Decision Support
  - Data Mining
  - Aggregation

Hilfe durch Data Lineage

Data Lineage – Motivation

15

- Schwierigkeit des Data Lineage hängt von Transformationen ab
  - SQL: Leichter aber unrealistisch
    - Data Lineage durch SQL Sichten
    - Data Lineage durch Operatoren der relationalen Algebra
  - Allgemeine Transformationen: Schwierig aber wichtig
    - Data Lineage durch komplexe, nutzerdefinierte Transformationen
    - Data Lineage durch ETL Prozesse
    - Data Lineage durch Ketten von 60+ Transformationen
- Data Lineage geschieht auf Datenebene.
  - Metadata Lineage
    - Schema Mapping
    - Schemaintegration

Data Lineage – Motivation

16

A	B
a	-1
a	2
b	0

A	B
a	2
b	0

Herkunft des Tupels (a, 2)?

- T = $\sigma_{B \geq 0}$
  - $\Rightarrow$ lin(a,2) = {(a,2)}
- T = Gruppierung nach A und Aggregation: 2x SUM(B)
  - $\Rightarrow$ lin(a,2) = {(a,-1); (a,2)}
- T = Gruppierung nach A und Aggregation: MAX(B)
- …

Data Lineage – Motivation

- Zusätzliche Schwierigkeiten
  - Runtime overhead
    - ETL
    - Bei virtueller Integration
  - Speicherbedarf
    - Metadaten
  - Transformationen
    - Einzeln
    - In Ketten
    - In (azyklischen) Graphen
- Trade-off zwischen Nutzen und Kosten

Data Lineage – Beispiel

Produkt(PID, Name, Kategorie, Preis, Gültig)

Bestellung(BID, KundenID, Datum, Produkte)

prod-id	prod-name	category	price	valid
111	Apple IMAC	computer	1200	10/1/1998–
222	Sony VAIO	computer	3280	9/1/1998–11/30/1998
222	Sony VAIO	computer	2250	12/1/1998–9/30/1999
222	Sony VAIO	computer	1950	10/1/1999–
333	Canon A5	electronics	400	4/2/1999–
444	Sony VAIO	computer	2750	12/1/1998–

Figure 1: Source data set for `Product`

order-id	cust-id	date	prod-list
0101	AAA	2/1/1999	333(10), 222(10)
0102	BBB	2/8/1999	111(10)
0379	CCC	4/9/1999	222(5), 333(5)
0524	DDD	6/9/1999	111(20), 333(20)
0761	EEE	8/21/1999	111(10)
0952	CCC	11/8/1999	111(5)
1028	DDD	11/24/1999	222(10)
1250	BBB	12/15/1999	222(10), 333(10)

Figure 2: Source data set for `Order`

Quelle: [CW03]

Data Lineage – Beispiel

19

Ziel: Tabelle „Verkaufssprung“

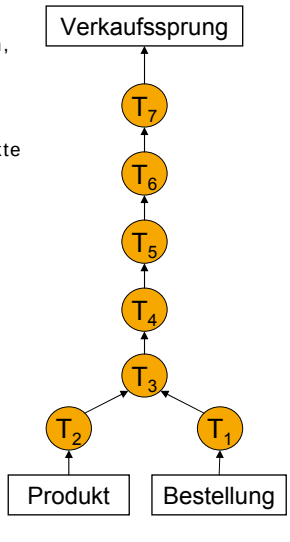
- Computer-Produkte, die sich im letzten Quartal mehr als doppelt so viel verkauften wie im Durchschnitt der drei vorigen Quartale

1. Tabelle anlegen
2. Transformationen als Graph definieren
3. Transformationen ausführen

Data Lineage – Beispiel

20

- T1: Bestellungen (Produktlisten) aufspalten
  - Neues Schema: Bestellung(BID, KundenID, Datum, PID, Menge)
- T2: Kategorie selektieren
  - Filter für Computer Kategorie
- T3: Join (und Projektion) über Bestellungen und Produkte
  - Neues Schema: (BID, Datum, PID, Menge, Name, Preis, Gültig)
- T4: Aggregation und Pivotisierung
  - Verkaufsmenge pro Quartal und Produkt
  - Neues Schema: (Name, Q1, Q2, Q3, Q4)
- T5: Durchschnittsberechnung
  - Neues Schema: (Name, Q1, Q2, Q3, AVG3, Q4)
- T6: Selektion für Verkaufsprünge
- T7: Projektion
  - Neues Schema: Verkaufssprung(Name, AVG3, Q4)



Produkt(PID, Name, Kategorie, Preis, Gültig)
Bestellung(BID, KundenID, Datum, Produkte)

Data Lineage – Beispiel

prod-id	prod-name	category	price	valid
111	Apple IMAC	computer	1200	10/1/1998–
222	Sony VAIO	computer	3280	9/1/1998–11/30/1998
222	Sony VAIO	computer	2250	12/1/1998–9/30/1999
222	Sony VAIO	computer	1950	10/1/1999–
333	Canon A5	electronics	400	4/2/1999–
444	Sony VAIO	computer	2750	12/1/1998–

order-id	cust-id	date	prod-list
0101	AAA	2/1/1999	333(10), 222(10)
0102	BBB	2/8/1999	111(10)
0379	CCC	4/9/1999	222(5), 333(5)
0524	DDD	6/9/1999	111(20), 333(20)
0761	EEE	8/21/1999	111(10)
0952	CCC	11/8/1999	111(5)
1028	DDD	11/24/1999	222(10)
1250	BBB	12/15/1999	222(10), 333(10)

Data Lineage für Verkaufssprung Tupel (Sony VAIO, 11250, 39600):

Order

order-id	cust-id	date	prod-list
0101	AAA	2/1/1999	333(10), 222(10)
0379	CCC	4/9/1999	222(5), 333(5)
1028	DDD	11/24/1999	222(10)
1250	BBB	12/15/1999	222(10), 333(10)

Product

prod-id	prod-name	category	price	valid
222	Sony VAIO	computer	2250	12/1/1998–9/30/1999
222	Sony VAIO	computer	1950	10/1/1999–

Quelle: [CW03]

Überblick

- Motivation und Beispiel
- Datentransformationen (jeweils Definition und Tracing Prozedur)
  - Dispatcher
  - Aggregatoren
  - Black Boxes & MISOs
  - [Inverse Transformation]
- Transformationssequenzen
- Data Lineage nach [CW03]

Transformationen

23

- Datenmenge
  - Menge aus beliebigen Daten
    - Tupel, Werte, Objekte
    - Hier: i.d.R. Tupel
- Transformation
  - Beliebige Prozedur, mit einer Datenmenge als Input und einer Datenmenge als Output.
  - $T(I) = O$
- Komposition von Transformationen
  - $T = T_1 \circ T_2$: $T(I) = T_2(T_1(I))$
  - [Assoziativ: $(T_1 \circ T_2) \circ T_3 = T_1 \circ (T_2 \circ T_3)$]

Transformationen – Eigenschaften

24

- Stabil: Kein erfundener Output
  - Also: $T(\varnothing) = \varnothing$
  - Gegenbeispiel?
    - Transformationen, die jedem Tupel einen festen Wert anhängen
- Deterministisch: Immer gleicher Output bei gleichem Input
  - Gegenbeispiel?
    - Transformationen, die einen zufälligen Sample produziert.

Data Lineage – Definition

25

- Allgemein gilt: Transformationen können für jeden Outputwert alle Inputwerte betrachten.
  - I.d.R. ist das nicht so.
- Sei T(I) = O und $o \in O$
- $I^* \subseteq I$ ist die Menge der Inputwerte, die zum Output o beitragen.
- $I^* = T^*(o,I)$
- Sei $O^* \subseteq O$, dann $T^*(O^*,I) = \cup_{o \in O^*} T^*(o,I)$
  - O^* ist der „interessante“ Output

Data Lineage – Motivation

26

A	B
a	-1
a	2
b	0

Transformation T

A	B
a	2
b	0

Herkunft des Tupels (a, 2)?

- $T = \sigma_{B \geq 0}$
  - $\Rightarrow T^*((a,2),I) = \{(a,2)\}$
- T gruppiert nach A und aggregiert 2x SUM(B)
  - $\Rightarrow T^*((a,2),I) = \{(a,-1); (a,2)\}$

Transformationen

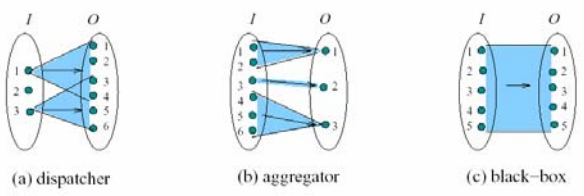
27

- Zwei Extreme
  - Relationale Operatoren oder Sichten:
    - Exakte Data Lineage kann bestimmt werden.
  - Völlig unbekannte Transformation
    - Der gesamte Input ist Data Lineage.
- Realität liegt dazwischen.
- Drei Transformationsklassen
- Hinzu kommen
  - Schema Mappings (nicht hier)
  - [Inverse Transformation]en

Transformationen - Klassifikation

28

- Dispatcher (wörtlich: „Abfertiger“)
  - Jeder Input produziert null oder mehr Outputs
- Aggregatoren
  - Gruppen von Inputs produzieren einen Output
- Black-Boxes
  - Alles andere



Quelle: [CW03]

Überblick

29

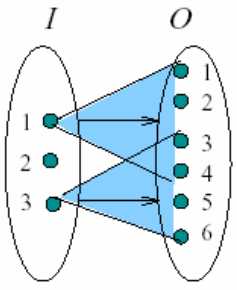
- Motivation und Beispiel
- Datentransformationen (jeweils Definition und Tracing Prozedur)
  - Dispatcher
  - Aggregatoren
  - Black Boxes & MISOs
  - [Inverse Transformation]
- Transformationssequenzen

- Data Lineage nach [CW03]

Transformationen – Dispatcher

30

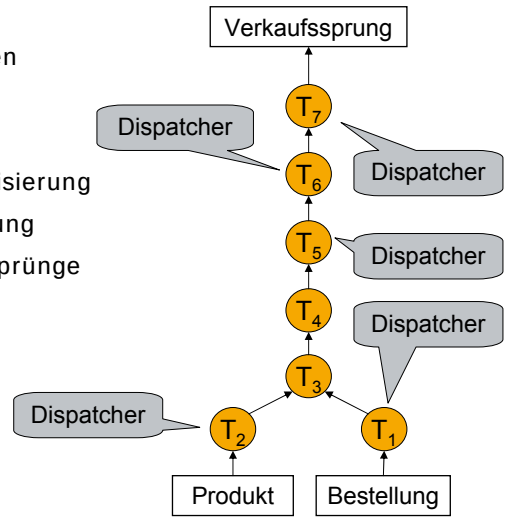
- Jeder Input produziert unabhängig null oder mehr Outputs.
- Formal:
  - $\forall I,\ T(I) = \cup_{i \in I} T(\{i\})$
- Lineage:
  - $T^*(o,I) = \{i \in I \mid o \in T(\{i\})\}$



Data Lineage – Beispiel

31

- T1: Bestellungen aufspalten
- T2: Kategorie selektieren
- T3: Join und Projektion
- T4: Aggregation und Pivotisierung
- T5: Durchschnittsberechnung
- T6: Selektion für Verkaufsprünge
- T7: Projektion



Transformationen – Dispatcher

32

Tracing Prozedur

- Definiert für Outputmengen, deshalb geeignet für Kompositionen

procedure TraceDS($\mathcal{T}, O^*, I$)
  $I^* \leftarrow \varnothing$;
  for each $i \in I$ **do**
    if $\mathcal{T}(\{i\}) \cap O^* \neq \varnothing$ **then** $I^* \leftarrow I^* \uplus \{i\}$;
return I^*;

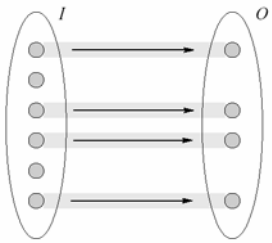
Aufwand:

- Vollständiger Scan des Input
- Transformationsaufruf für jeden Inputwert

Transformationen – Dispatcher

33

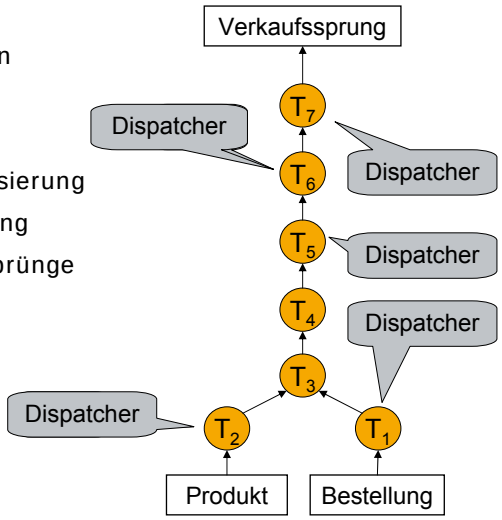
- Dispatcher-Spezialfall: Filter
- Filter:
  - $\forall i \in I$, $T(\{i\}) = \{i\}$ oder $T(\{i\}) = \varnothing$
- Data Lineage:
  - $T^*(o) = \{o\}$
  - Bzw. $T^*(O) = O$
- Tracing Prozedur trivial



Data Lineage – Beispiel

34

- T1: Bestellungen aufspalten
- T2: Kategorie selektieren
- T3: Join und Projektion
- T4: Aggregation und Pivotisierung
- T5: Durchschnittsberechnung
- T6: Selektion für Verkaufsprünge
- T7: Projektion



Überblick

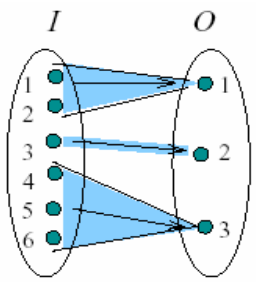
35

- Motivation und Beispiel
- Datentransformationen (jeweils Definition und Tracing Prozedur)
  - Dispatcher
  - Aggregatoren
  - Black Boxes & MISOs
  - [Inverse Transformation]
- Transformationssequenzen
- Data Lineage nach [CW03]

Transformationen – Aggregatoren

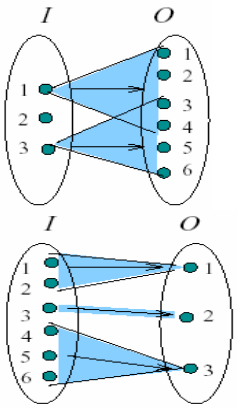
36

- Zwei Bedingungen müssen gelten
  1. Partitionierung: Inputs können partitioniert werden, so dass jede Partition für genau ein Output verantwortlich ist.
     - Sei $T(I) = \{o_1,\ldots, o_n\}$.
     - $\forall I$ existiert eine eindeutige, disjunkte Partitionierung $I_1, \ldots, I_n$ so dass $T(I_k) = o_k$ für alle k.
  2. Vollständigkeit: Jeder Input ist an mindestens einem Output beteiligt.
     - $\forall I \neq \emptyset$, $T(I) \neq \emptyset$
- Lineage: $T^*(o_k,I) = I_k$



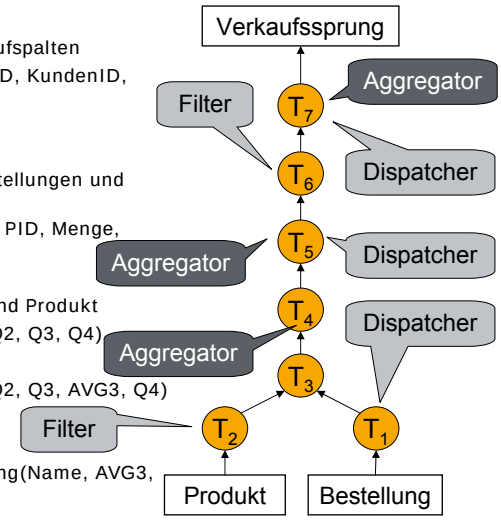
Vergleich: Dispatcher vs. Aggregator

- Dispatcher
  - Jeder Input produziert unabhängig null oder mehr Outputs.
- Aggregator
  - Jeder Input ist an mindestens einem Output beteiligt.
  - Inputs können partitioniert werden, so dass jede Partition für genau ein Output verantwortlich ist.
- Transformationen, die zugleich Dispatcher und Aggregator sind?
  - Identität
  - Projektion (ohne Duplikateliminierung)



Data Lineage – Beispiel

- T1: Bestellungen (Produktlisten) aufspalten
  - Neues Schema: Bestellung(BID, KundenID, Datum, PID, Menge)
- T2: Kategorie selektieren
  - Filter für Computer Kategorie
- T3: Join (und Projektion) über Bestellungen und Produkte
  - Neues Schema: (BID, Datum, PID, Menge, Name, Preis, Gültig)
- T4: Aggregation und Pivotisierung
  - Verkaufsmenge pro Quartal und Produkt
  - Neues Schema: (Name, Q1, Q2, Q3, Q4)
- T5: Durchschnittsberechnung
  - Neues Schema: (Name, Q1, Q2, Q3, AVG3, Q4)
- T6: Selektion für Verkaufsprünge
- T7: Projektion
  - Neues Schema: Verkaufssprung(Name, AVG3, Q4)



Transformationen – Aggregatoren

- Tracing Prozedur

procedure TraceAG($\mathcal{T}, O^*, I$)
  $L \leftarrow$ all subsets of I sorted by size;
  for each $I^* \in L$ in increasing order **do**
    if $\mathcal{T}(I^*) = O^*$ **then**
      if $\mathcal{T}(I - I^*) = O - O^*$ **then break**;
      else $L =$ all supersets of I^* sorted by size;
return I^*;

Potenzmenge

Mindestens I* ist Lineage

I* ist vollständiges Lineage

Eingrenzung: I* ist größer

- Aufwand: 2|I| Aufrufe von T
  - Zu viel!
  - Deshalb zwei Unterklassen
    - Kontextfreie Aggregatoren
    - Schlüsselerhaltende Aggregatoren

Transformationen – Aggregatoren

- Kontextfreie Aggregatoren
  - Input gehört zu einer Partition, unabhängig von den Werten andere Inputs in der Partition.
  - Alle bisherigen Aggregatoren sind kontextfrei.
  - Gegenbeispiel: Clustering und Durchschnittsbildung über Cluster
    - Mitgliedschaft in einem Cluster ist von anderen Werten abhängig.
- Tracing Prozedur einfacher
  - Intuition: Bildung der Partitionen I ist linear. Danach nur noch Zugehörigkeit prüfen.

Transformationen – Aggregatoren

41

```
procedure TraceCF(T, O*, I)
  I* ← ∅;
  pnum ← 0;
  for each i ∈ I do
    if pnum = 0 then I_1 ← {i}; pnum ← 1; continue;
    for (k ← 1; k ≤ pnum; k++) do
      if |T(I_k ∪ {i})| = 1 then I_k ← I_k ∪ {i}; break;
    if k > pnum then pnum ← pnum + 1; I_pnum ← {i};
  for k ← 1..pnum do
    if T(I_k) ⊆ O* then I* ← I* ∪ I_k;
return I*;
```

Finde für jedes *i* eine Partition

Initialisierung der ersten Partition

Prüfe ob *i* in eine vorhandene Partition passt

Sonst erzeuge neue Partition

|I²| Transformationen

Suche Partitionen, die O* erzeugen.

|I| Transformationen

Transformationen – Aggregatoren

42

- Schlüsselerhaltende Aggregatoren
  - Sei I partitioniert I_1, ..., I_n, so dass T(I) = $\{o_1,...o_n\}$.
  - ∀k, ∀I‘⊆I_k : T(I‘) = $\{o‘_k\}$ und $o‘_k$.key = o_k.key
  - Beispiel: „Normale“ Gruppierung und Aggregation
  - Gegenbeispiel: Gruppierung, die Gruppierungsattribut nicht erhält.
- Tracing Prozedur
  - Aufwand: |I|
  - Intuition: Schlüssel im Transformationsergebnis wird verwendet, um Zugehörigkeit zu prüfen.

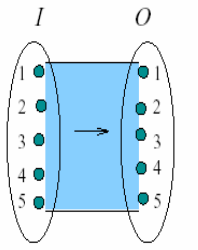
```
procedure TraceKP(T, O*, I)
  I* ← ∅;
  for each i ∈ I do
    if π_key(T({i})) ⊆ π_key(O*)
    then I* ← I* ⊎ {i};
return I*;
```

Überblick

- Motivation und Beispiel
- Datentransformationen (jeweils Definition und Tracing Prozedur)
  - Dispatcher
  - Aggregatoren
  - Black Boxes & MISOs
  - [Inverse Transformation]
- Transformationssequenzen
- Data Lineage nach [CW03]

Transformationen – Black Boxes

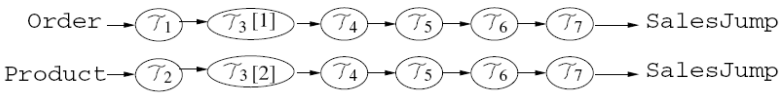
- Transformationen, die weder Dispatcher noch Aggregatoren sind, noch eine explizite Tracing Prozedur aufweisen.
- Beispiel:
  - Sortierung und Einfügen der Ordnungszahl.
  - Kein Dispatcher, weil Output nicht unabhängig
  - Kein Aggregator, weil ein Output nur mittels aller Inputs erzeugt werden kann.
- Lineage:
  - T*(o,I) = I
- Tracing Prozedur:
  - Trivial,
  - aber nutzlos



MISOs – Multiple Input Single Output Transformationen

45

- Exklusive MISOs
  - Unabhängige Transformation jeder Inputmenge
  - Beispiel: UNION
  - Lineage:
    - Teilen der Transformation in unabhängige Teile
    - Bestimmung der Eigenschaften der Teile
    - Lineage gemäß der Eigenschaften
- Inklusive MISOs
  - Lineage
    - Teilen der Transformation in Einzelteile
      - » Jeweils anderer Input als Konstante



Überblick

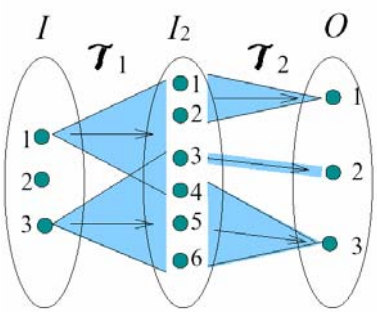
46

- Motivation und Beispiel
- Datentransformationen (jeweils Definition und Tracing Prozedur)
  - Dispatcher
  - Aggregatoren
  - Black Boxes & MISOs
  - [Inverse Transformation]
- Transformationssequenzen
- Data Lineage nach [CW03]

Transformationssequenzen

- Bisher: Lineage und Tracing für einzelne Transformationen
- Nun: Sequenzen von Transformationen
- Sei $I_2^* = T_2^*(o, I_2)$
- Sei $I^* = T_1^*(I_2^*, I)$
- Dann gilt $I^* = (T_1 \circ T_2)^*(o, I)$

- Beispiel:
  - $(T_1 \circ T_2)^*(3, I) = \{1,3\}$



Transformationssequenzen

Naive Tracing Prozedur für Sequenzen $T_1 \circ \ldots \circ T_n$:

- Speicherung aller Zwischenergebnisse I_k
- Tracing Prozedur rückwärts für jeden Transformationsschritt.
- Nicht effizient:
  - Hoher Speicherbedarf
  - Viele Transformationsschritte
- Besser: Explizite Kombination von Transformationen

Transformationssequenzen

49

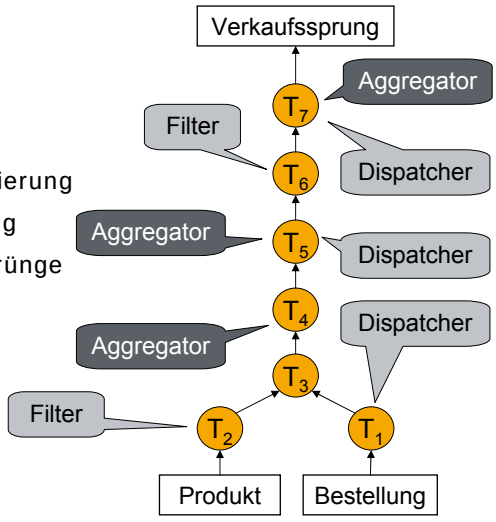
Gegeben eine Transformationssequenz

1. Normalisiere Sequenz durch geeignete Kombinationen
2. Bestimme für Tracing benötigte Zwischenergebnisse
3. Bei Transformation, speichere diese Zwischenergebnisse
4. Iteratives Tracing durch normalisierte Sequenz

Rückblick

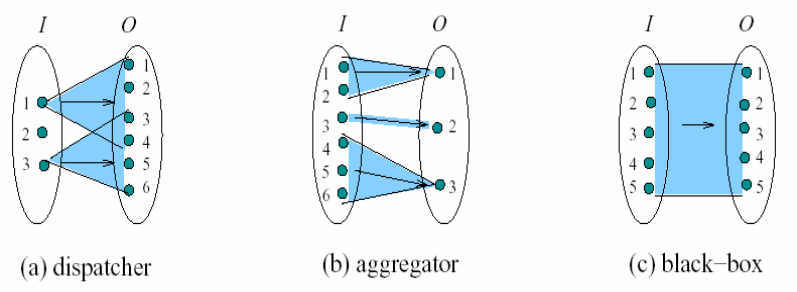
50

- T_1: Bestellungen aufspalten
- T_2: Kategorie selektieren
- T_3: Join und Projektion
- T_4: Aggregation und Pivotisierung
- T_5: Durchschnittsberechnung
- T_6: Selektion für Verkaufsprünge
- T_7: Projektion



Rückblick

51

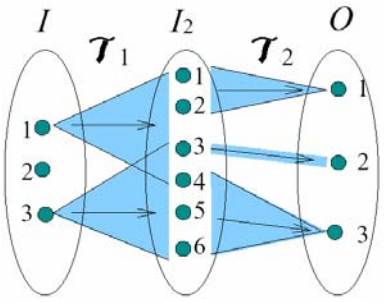


Definition | Beispiel | Tracing Prozedur | Tracing Aufwand

Rückblick

52

- Transformationssequenzen
- Normalisierungsalgorithmus



Literatur

[CW03] Yingwei Cui, Jennifer Widom: Lineage tracing for general data warehouse transformations. VLDB J. 12(1): 41-58 (2003)

Ergänzend:

- [BKT01] Peter Buneman, Sanjeev Khanna, Wang Chiew Tan: Why and Where: A Characterization of Data Provenance. ICDT 2001: 316-330