100 Puzzles to Learn Data Warehousing

Cristian Scutaru

Table of Contents

Introduction

Learning how to design a data warehouse may be difficult. Ralph Kimball has some great legacy books on the dimensional modeling techniques, but they are verbose, with complicated examples. Our illustrated examples here are kept simple on purpose, to help you better understand complicated concepts like periodic snapshot fact tables, degenerate dimensions, late arriving facts or dimensions.

We focus our puzzles on Ralph Kimball's dimensional modeling techniques, but we also introduce you to Extract-Transform-Load (ETL) basics, OLAP fundamentals and some other important things you must know about data warehouses in general. We dive deep into slowly changing dimensions (SCD), with other illustrated examples to help you get the ideas in no time.

You need just some basic prior knowledge about Data Warehouses in general. The explanations and external references from the answers to our questions will help you learn the rest. We also assume you already have some basic background in data modeling for relational databases, and SQL.

These puzzles are for Software Developers and Engineers, Database Engineers and Architects, or Data Analysts. Difficulty level is from beginner to advanced.

We've split the 100 questions into 5 quizzes with 20 single and multi-choice questions each. Try solving each quiz separately, writing down on a piece of paper the answer to each question. Then go to the Answers and Explanations section, and learn more from our solutions to the puzzles. Follow the links to external references for a deep dive on the subject.

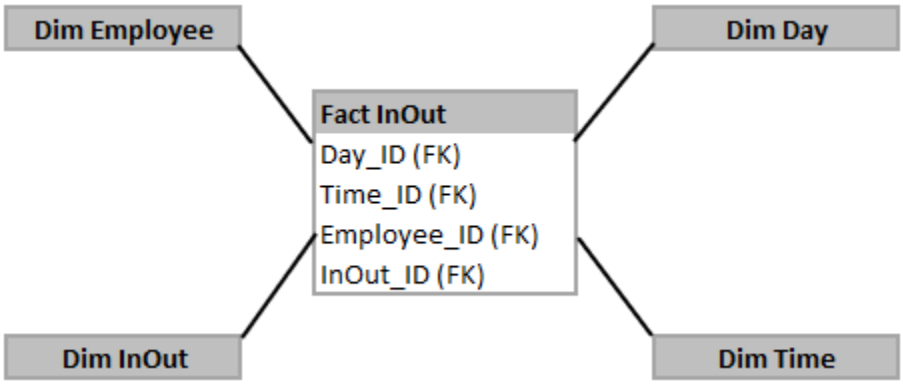
An interactive version of this book has been implemented on Udemy as **100 Puzzles to Learn Data Warehousing**.

Quiz 1

Question 1:

The Fact InOut table records the date and time when employees are hired and then leave the company.

What kind of fact table is this? (select one)



A) non-additive fact table
B) factless fact table
C) degenerate fact table
D) star table

Question 2:

What is this called? (select one)

	Date	Prospect	Customer	Account	Product	Household	Branch
New Business Solicitation	X	X			X		
Lead Tracking	X	X			X		X
Account Application Pipeline	X	X	X	X	X		X
Account Initiation	X	X	X	X	X	X	X
Account Transactions	X		X	X	X	X	X
Account Monthly Snapshot	X		X	X	X	X	X
Account Servicing Activities	X		X	X	X	X	X

A) intersection table
B) dimensional choice table
C) business matrix
D) bus matrix

Question 3:

Which is NOT a data warehouse commercial product? (select one)

A) Amazon Redshift
B) Google's BigQuery
C) Apache HBase
D) Apache Hive
E) Snowflake

Question 4:

Is this a valid periodic snapshot fact table, with daily account balances? (select one)

Account	Date	Balance
12345	3/1/2017	$1,000
12345	3/2/2017	$1,000
12345	3/5/2017	$1,400
12345	3/6/2017	$1,325
12345	3/7/2017	$1,325

A) Yes, because it keeps track of the daily balance as it should.
B) No, because there should be more account numbers in such a table.
C) No, because there are some days with no entries.
D) No, because periodic snapshots should record only new transactions.

Question 5:

What does ETL stand for? (select one)

A) Extract, Transfer, Load
B) Export, Transfer, Load
C) Extract, Transform, Load

Question 6:

What data flow was recommended by William Inmon? (select one)

A) OLTP to ETL to staging to DW to ETL to data marts to OLAP cubes
B) OLTP to data marts to ETL to staging to DW to OLAP cubes
C) OLTP to ELT to DW to staging to OLAP cubes

Question 7:

Which is NOT a typical pair of ETL data transformation? (select one)

A) splitting and joining
B) mapping and reducing
C) transposing and pivoting
D) filtering and sorting
E) cleaning and validating

Question 8:

Which is NOT a type of data mart? (select one)

A) Dependent
B) Independent
C) Combined
D) Hybrid

Question 9:

Which are NOT correct SCD types? (check all that apply)

A) Type 0: Retain original
B) Type 1: Overwrite
C) Type 2: Add new attribute
D) Type 3: Add new row
E) Type 4: Add mini-dimension

Question 10:

What is a conformed fact? (select one)

A) A same measurement that appears in separate fact tables, with the same consistent meaning.
B) A measurement referenced by a conformed dimension.
C) A measurement that appears with the same name in separate fact tables.

Question 11:

How would you qualify the Balance measure from the following periodic snapshot fact table? (check all that apply)

Account	Date	Balance
12345	3/1/2017	$1,000
12345	3/2/2017	$1,000
12345	3/3/2017	$1,400
12345	3/4/2017	$1,325
12345	3/5/2017	$1,325

A) additive
B) semi-additive
C) running sum
D) moving average

Question 12:

Which is NOT a known type of OLAP? (select one)

A) WOLAP
B) ROLAP
C) MOLAP
D) GOLAP
E) YOLAP

Question 13:

A query listing insurance policies with cancelation date greater than today is not able to return the active policies.

What could be the cause? (select one)

A) cancelation date value is NULL
B) cancelation date is recorded as a string

C) should use the BETWEEN operator

Question 14:

What is a staging area? (select one)

A) A special table in a data warehouse.
B) A special file used during the ETL process.
C) An intermediate storage area used for data processing during the ETL process.

Question 15:

You populate a PostgreSQL Countries table with population per state, in millions of people:

country	state	pop
USA	California	40
USA	Texas	29
Canada	Ontario	15

You issue the following SQL query:

```
SELECT country, state, SUM(pop) AS totals
FROM Countries
GROUP BY *** (country, state)
ORDER BY country, state;
```

What should * be to get the result below? (select one)**

country	state	*totals*
Canada	Ontario	15
Canada		15
USA	California	40
USA	Texas	29
USA		69
		84

A) ROLLUP
B) CUBE
C) GROUPING SETS

Question 16:

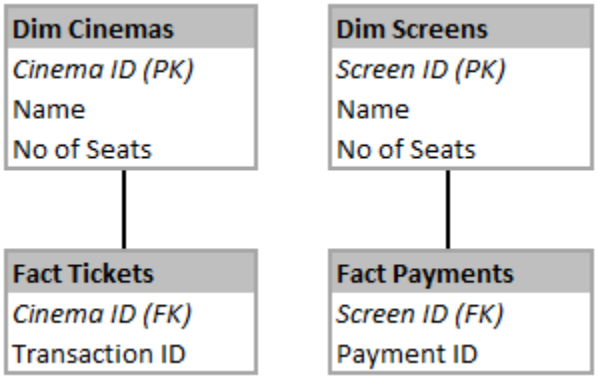
What values can go into a calendar date dimension? (check all that apply)

A) Week number
B) Time-of-day
C) Fiscal period
D) Holiday indicator
E) Numeric primary key

Question 17:

Different teams import similar data, but create tables with different names: Cinemas-Tickets and Screens-Payments are similar systems, processing cinema sales.

What do you call the two similar dimension tables? (select one)



A) similar dimensions
B) degenerate dimensions
C) conformed dimensions
D) shrunken dimensions

Question 18:

What is the main difference between an ETL and an ELT? (select one)

A) ELT loads data directly into the data warehouse, which also serves as staging area.
B) ETL loads data directly into the data warehouse, which also serves as staging area.
C) There is no such term as an ELT.

Question 19:

The following fact table has an entry with a NULL value for a numeric measure column (Discount), and another NULL value for a dimension foreign key value (Store ID).

How would you improve the overall design? (select one)

Date	Amount	Discount	Store ID
3/1/2017	$1,000	$200	3456
3/7/2018	$4,300	NULL	NULL

A) Replace all NULLs in Discount with either 0 or -1.
B) Replace all NULLs in Store ID column with some reserved value.
C) Add a reserved value in the related Stores dimension table for NULL entries, used also as FK.
D) Replace all NULLs in both Discount and Store ID as described before.
E) Do nothing, as best practice is to use NULLs as imported.

Question 20:

What is a late arriving fact? (select one)

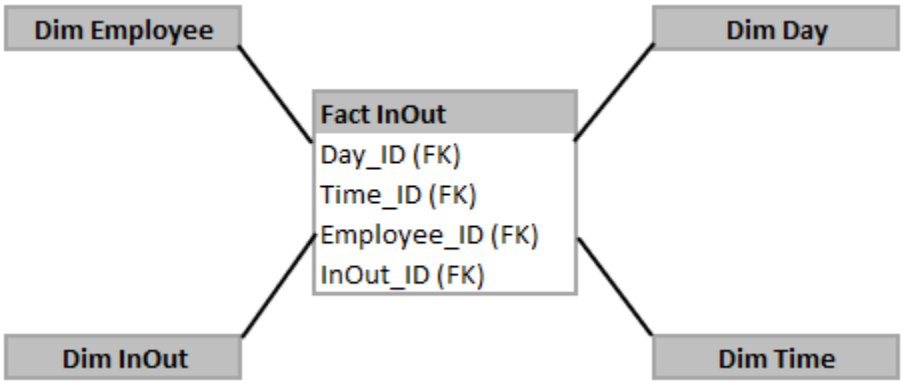
A) A fact received in a different order than it was recorded in the OLTP system.
B) A fact to be recorded, when one of its dimension key values already changed.
C) A fact received before one of its related dimensions.

Quiz 1 - Answers and Explanations

Question 1:

The Fact InOut table records the date and time when employees are hired and then leave the company.

What kind of fact table is this? (select one)



A) non-additive fact table
B) factless fact table
C) degenerate fact table
D) star table

Answer: B

Domain: Modeling

Explanation:

Although most measurement events capture numerical results, it is possible that the event merely records a set of dimensional entities coming together at a moment in time. For example, an event of a student attending a class on a given day may not have a recorded numeric fact, but a fact row with foreign keys for calendar day, student, teacher, location, and class is well-defined. Likewise, customer communications are events, but there may be no associated metrics.

Factless fact tables can also be used to analyze what didn't happen. These queries always have two parts: a factless coverage table that contains all the possibilities of events that might happen and an activity table that contains the events that did happen. When the activity is subtracted from the coverage, the result is the set of events that did not happen.

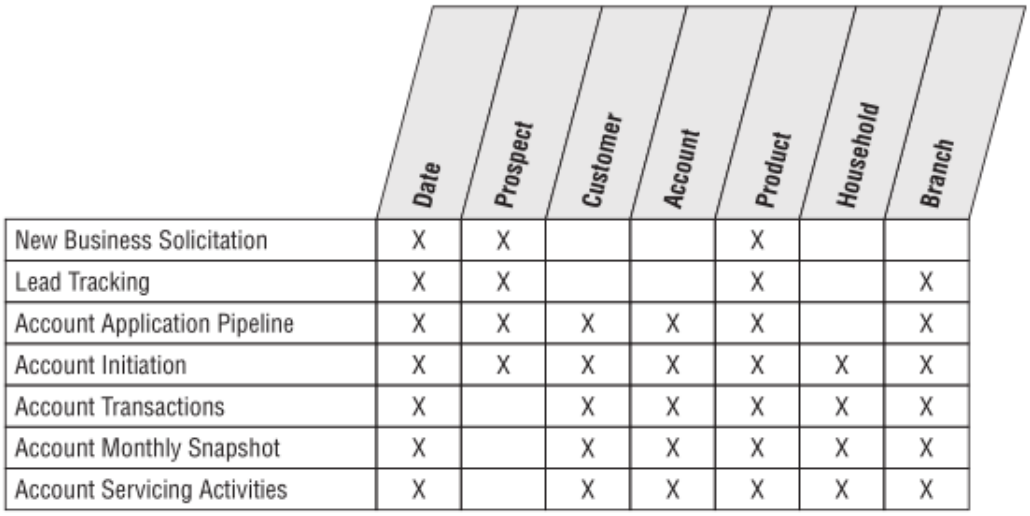
References:

https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/factless-fact-table/

https://panoply.io/data-warehouse-guide/#conceptual-logical-and-physical-data-models

Question 2:

What is this called? (select one)



	Date	Prospect	Customer	Account	Product	Household	Branch
New Business Solicitation	X	X			X		
Lead Tracking	X	X			X		X
Account Application Pipeline	X	X	X	X	X		X
Account Initiation	X	X	X	X	X	X	X
Account Transactions	X		X	X	X	X	X
Account Monthly Snapshot	X		X	X	X	X	X
Account Servicing Activities	X		X	X	X	X	X

A) intersection table
B) dimensional choice table
C) business matrix
D) bus matrix

Answer: D

Domain: Modeling

Explanation:

The **enterprise data warehouse bus matrix** is the essential tool for designing and communicating the enterprise data warehouse bus architecture. The rows of the matrix are business processes and the columns are dimensions. The shaded cells of the matrix indicate whether a dimension is associated with a given business process. The design team scans each row to test whether a candidate dimension is well-defined for the business process and also scans each column to

see where a dimension should be conformed across multiple business processes. Besides the technical design considerations, the bus matrix is used as input to prioritize DW/BI projects with business management as teams should implement one row of the matrix at a time.

The **detailed implementation bus matrix** is a more granular bus matrix where each business process row has been expanded to show specific fact tables or OLAP cubes. At this level of detail, the precise grain statement and list of facts can be documented.

References:

https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/enterprise-data-warehouse-bus-matrix/

Question 3:

Which is NOT a data warehouse commercial product? (select one)

A) Amazon Redshift
B) Google's BigQuery
C) Apache HBase
D) Apache Hive
E) Snowflake

Answer: C

Domain: General

Explanation:

Apache HBase is a NoSQL database, providing Bigtable-like capabilities for Hadoop. Apache Hive is the data warehouse built on top of Apache Hadoop.

This may be an easy question, but:

(a) It is very easy to mix-up HBase with Hive, when talking about data warehouses on Hadoop.
(b) Redshift, BigQuery and Snowflake are the most popular cloud data warehouses today, and this is worth remembering.

References:

https://en.wikipedia.org/wiki/Apache_HBase