---
test 1 ausarbeitung
---

ignoriere was am test steht - ist mit sicherheit grš§teils falsch

1) DATA WAREHOUSE CHARACTERISTICS (6 points)
subject-oriented:
- In the DW, data is not stored by operational applications, but by business subjects
- Data is grouped around subjects
- its structure is designed to make querying the data simple.
integrated:
- DW contains consolidated data from several (operational) applications/databases
- Usually the data in the DW is not updated or deleted
- Data inconsistencies are removed
time-variant and nonvolatile:
- The data stored in operational systems contains the current values
- DW: When new current data becomes available, the "old" data is not overwritten
- A data warehouse has to contain historical data to
    - Allow for analysis of the past
    - Relate information to the present
    - Enable forecasts for the future

2) INFORMATION INTEGRATION APPROACHES (6 points)
- Federation: Everybody talks directly to everyone else.
    - issue: n applications / data stores => up to n^2 connections
- Warehouse: Sources are translated from their local schema to a global schema and copied to a central DB.
    - issue: usually only one-directional data flows supported
- Mediator: Virtual warehouse – turns a user query into a sequence of source queries and assembles the results of this queries into an “aggregate” result.
    - issue: complex architecture, potentially slow, difficult to maintain

3) DWH USE CASES (3 points)
   List three typical use cases for data warehousing
- Manufacturing: Logistics management
- Resource management
- Finance: Risk management
- Airlines: Route profitability

4) DWH ARCHITECTURE REFERENCE MODEL (5 points)
picture from left to right:
Source DBs
Landing Area
Staging Area
Data Warehouse
Data Marts

Source DB: DB of an application which supports types of business transactions.
Landing Area: DB that is able to store a single data extract of a subset of one Source DB. Schema 1:1 with source DB scheme.
Staging Area: DB that is able to store matching data extracts from various Landing Areas in an integrated format, waiting for the upload to the DW once data from all Landing Areas are available. Schema 1:1 with DW scheme.
Data Warehouse: DB containing the history of all complete Staging Areas. Schema: 3NF.
Data Mart: DB – on disk or in main memory – containing data describing the (present and past) performance of one or more types of business transactions, taken form the DW. Schema: denormalized star.

5) OLTP VS OLAP (4 points)
X
X
O (mit etwas bauchweh: fkeys, indexes -> redundant?)
X

6) NORMALIZATION (3 points)
A nonkey attribute must provide information about the key, the whole key, and nothing but the key (so help me Codd).
(Anspielung auf 3NF)

7) FACTS VS DIMENSIONS(4 points)
X
X
O
X
(juhu)

8) SNOWFLAKE SCHEMA (4 points)
X
X
O
O

9) OLAP OPERATIONS (4 points)
3
1
2
3

10) R-OLAP VS M-OLAP VS H-OLAP (4 points)
X
X
X
0

11) HORIZONTAL PARTITIONING (4 points)
X
0
X
0

12) VERTICAL PARTITIONING AND COLUMN STORES (4 points)
0
0
X
X

13) BITMAP INDICES (13 points)

a)

	index_rating_avg_1	index_rating_avg_2	index_rating_avg_3	index_rating_avg_4
101	0	0	1	0
102	0	1	0	0
103	1	0	0	0
104	0	0	0	1

b)

I)
```
SELECT a.name
    FROM article AS a
    WHERE a.rating_avg = 4;
```

Option 2:
```
SELECT name
    FROM article
    WHERE index_rating_avg_4 = 1;
```

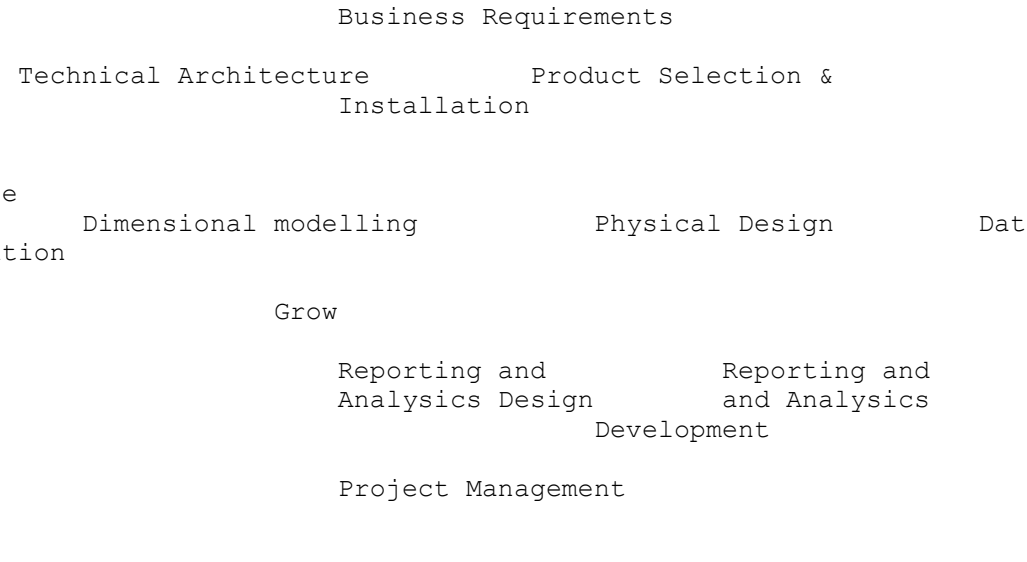
II)
```
SELECT count(article_id)
    FROM article AS a
    WHERE a.lactoseFree = 'TRUE' AND ( a.rating_avg = 3 OR a.rating_avg = 4 );
```

Option 2:
```
SELECT count(article_id)
    FROM article
    WHERE (index_rating_avg_4 = 1 OR index_rating_avg_3 = 1) AND actoseFree = 1;
```

14) ETL (4 points)
X

X
X
0

15) KIMBALLS DATA WAREHOUSE DEVELOPMENT LIFECYCLE (4 points)



16) AGILE BI (6 points)
Relevance:
There are lot of companies with very dynamic business intelligence applications. This means that there are many on going changes in the business intelligence applications due to changing requirements form business.
These requirements are regarding reports, KPIs.

Classical waterfall model has issues with handling a very dynamic environment like mentioned above..
There is a huge overhead when using a waterfall model which is not efficient for a environment where the requirements change nearly every day.
Nevertheless there are companies which do not face such a huge dynamic in their business intelligence applications, so in this case a waterfall model can be prefered.

Problems: Compliance or Regulation issues in areas like finance can stop you from using Agile BI
Risks: Keep the history after changes

17) SCHEMA-ON-WRITE VS SCHEMA-ON-RED (6 points)
Comparision:
Schema-on-write
(1) Schema must be created before any data can be loaded
(2) Explicit load operation transforms data to DB internal structure
(3) New columns must be added explicitly before new data for such columns can be loaded into DB

Schema-on-read
(1) Data is simply copied to the file store, no transformation
(2) Late binding: Serializer/Deserializer extracts required columns at read time
(3) New data can start flowing any time and will appear retroactively once the Ser/De is updated to parse it

Advantages:
Schema-on-write
+ Read is fast
+ Standards/Governance

Schema-on-read
+ Load is fast
+ Flexibility/Agility

18) HORIZONTAL VS. VERTICAL SCALING ISSUES (unfortunately unclear)
- Vertical (i.e., bigger machines): unfavorable economics, fundamental limits
- Horizontal (i.e., partitioning): Data Sharding for RDBMS developed as an add-on,

inherently difficult
- Price vs. performance:
“The cost structures of data warehouses are anywhere from 10 to 100 times higher than what they can drive per terabyte on a Hadoop cluster. Now they can store multiple years of data, not just a month or two.”

Storage (Kryder’s law) grows much faster than processing power (Moore’s law) and network speed (Nielsen’s law).

Solution: Distributed data stores
BUT: ACID (Atomicity, Consistency, Isolation, Durability) difficult in a distributed environment

19) CAP THEOREM (unfortunately unclear)
– Consistency: all nodes see the same data at the same time
– Availability: every request receives a response about whether it was successful or failed
– Partition tolerance: system continues to operate despite arbitrary message loss or failure of part of the system
Impossible to achieve all three simultaneously!

20) MC Frage: HDFS is
X a specific file system that runs natively on a local machine ("Sits on top of the native filesystem")
0 conceptually based on Google's Big Table
X conceptually based on Google's GFS
0 stored in HDFS can be changed ("Files are “write once” - can be replaced, but not changed")

21) MC Frage: HBase
0 distributed relational database
X is based on a column-family oriented data model
X is a key-value data store
0 uses typed columns

22 zusatzfrage) mehr daten sind immer besser als weniger daten - stimmen sie dem zu? (10)
siehe reading "The Economist: A different game"