1. **Data Warehouse Characteristics** (6 points)
According to a widely accepted definition by Inmon (1996), *"a data warehouse is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management's decision-making process"*. Explain the each of these four defining characteristics.

subject-oriented: It is a subject who ~~triggers~~ manages/uses the datawarehouse, not automatized.

integrated: There are specific platforms / integrated software products for managing the datawarehouse. The database is normalized.

time-variant: The database always has a relation to the time, and the has date/time attribute(s), telling when something happene

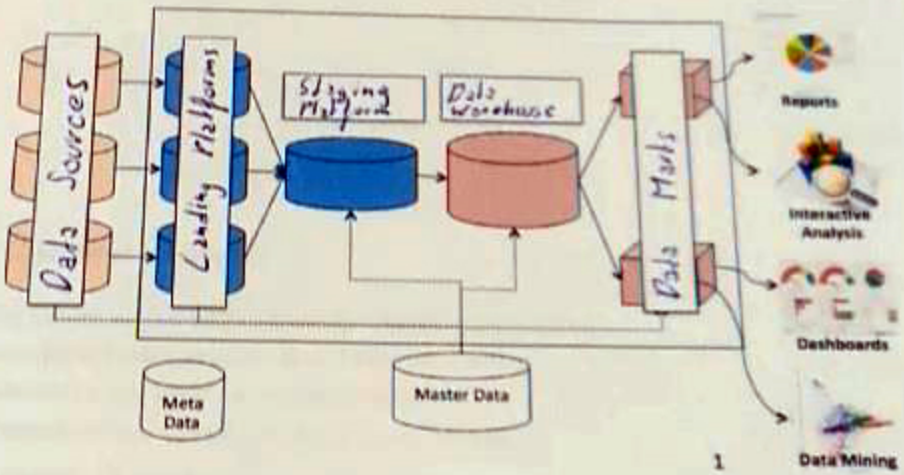
nonvolatile: rows in the database should not be deleted/updated d but insert a new row with the new data instead.

2. **Information Integration Approaches** (6 points)
Discuss the three basic architectural approaches for information integration and explain them succinctly. Also discuss the most significant issues associated with each approach.

3. **DWH use cases** (3 points)
List three typical use cases for data warehousing

3. Decision Support Systems (DSS)
3. Customer Management Systems
3. Operational Systems e.g. Rechnungslegung, Lieferscheine, Reservierungen, etc

4. **DWH Architecture Reference Model** (5 points)
Complete the following figure that illustrates the Data Warehouse Reference Architecture covered in the lecture by filling in the basic architectural elements in the respective fields



Briefly describe the role of each of these five building blocks below (1-2 sentences each).

DataSource: raw data, Tables in database

Landing Platform: pre-selected, filtered data, subsets of Data Source.

Staging Platform: transformed, aggregated data from Landing Platforms

Data Warehouse: Integrated data, operative Setup

Data Marts: used for analytic purposes, use-cases per user

5. **OLTP vs OLAP** (4 points)

- ☒ OLAP systems are optimized for complex queries
- ☒ OLTP systems tend to use normalized schemas
- ☐ OLTP accepts controlled redundancy for the sake of maximized query performance
- ☒ OLAP systems store historized data

6. **Normalization** (3 points)
Complete the following verdict that summarizes the formal requirements for a 3 NF relation:

"A nonkey attribute must . . .

. . . (so help me Codd)."

7. **Facts vs. Dimensions** (4 points)

- ☒ Fact tables are usually relatively "long" (many rows)
- ☒ Dimension tables provide the "business context"
- ☐ Fact tables are used for categorization
- ☒ Dimensions can usually thought of as "nouns"

8. **Snowflake Schema** (4 points)

- ☒ The Snowflake schema normalizes selected dimension tables of a star schema
- ☒ Snowflake schemas are somewhat more complex and less intuitive than Star schemas
- ☐ Denormalized structures are easier to update and maintain
- ☐ The performance of queries on a snowflake schema is typically better than on a star schem

9. **OLAP Operations** (4 points)
Going from a coarser level of aggregation to a finer (more detailed) level is called. . .
○ Slicing ● Dicing ☒ Drill-down ○ Roll-up

Adding filter conditions in one (or more) dimension(s) is called. . .
☒ Slicing ○ Dicing ○ Drill-down ○ Roll-up

In a Data Warehousing context, what is typically called *"pivoting"* or *"cross tabulation"* in s
sheet software is known as. . .
○ Slicing ☒ Dicing ● Drill-down ○ Roll-up

Which operation can be implemented in SQL by adding a `group by` clause along a dimensi
archy?
○ Slicing ○ Dicing ○ Drill-down ☒ Roll-up

10. **R-OLAP vs M-OLAP vs. H-OLAP** (4 points)
Which of the following statements regarding R-OLAP/M-OLAP/H-OLAP/D-OLAP are correct?
   - ☒ M-OLAP stores data in DBMSs optimized for storage of multidimensional data
   - ☒ Compared to M-OLAP, R-OLAP systems tend to be more scalable for very large data volumes
   - ☒ H-OLAP is a hybrid of R-OLAP and M-OLAP technologies that aims to combine the greater scalability of R-OLAP with the faster computation of M-OLAP.
   - ☐ M-OLAP stores multidimensional data in relational DBMSs

11. **Horizontal Partitioning** (4 points)
Which of the following statements regarding Horizontal Partitioning are correct?
   - ☒ Horizontal partitioning is based on the idea of splitting a table into disjoint parts with the same schema
   - ☐ Horizontal partitioning is based on the idea of pre-computing aggregated query results that can efficiently answer other queries over a star schema
   - ☒ The union of all horizontal partitions equals the original table
   - ☐ Horizontal partitioning involves a space-time trade-off

12. **Vertical Partitioning and Column Stores** (4 points)
Which of the following statements regarding vertical partitioning are correct?
   - ☒ Vertical partitioning is based on the idea of pre-computing aggregated query results that can efficiently answer other queries over a star schema.
   - ☐ Full-tuple access is always faster when the tuples are stored in column stores organized based on a decomposed storage model.
   - ☐ Vertical partitioning is based on the idea of dividing a table into multiple tables that contain fewer columns
   - ☐ The unique key column is replicated in all tables for joining

13. **Bitmap Indices** ()

(a) Suppose the relation in Table 1 is used in the DWH of a pastry producer. The field `rating_avg` contains rounded integer average ratings from 1 to 4. What does a bitmap on `rating_avg` for the table instance given above look like? Add the index columns and respective row values to the table. [5]

article_id	name	lactoseFree	price	rating_avg	index_rating_avg
101	Brownie	TRUE	3.0	3	3
102	Cheesecake	FALSE	3.5	2	2
103	Cookie	TRUE	2.5	1	1
104	Sachertorte	FALSE	5.0	4	4

Table 1: Relation `Article` - add a bitmap index column for `rating_avg`

(b) Assume there are separate bitmap indices on column 'lactoseFree' and column 'rating'. How could you use these bitmap indices to answer the following queries efficiently? Provide the SQL code for the following sample queries:

i. What are the names of the highest-rating articles (i.e., those with the highest value in rating_avg)? [3]

SELECT name, max(rating_avg) FROM Article GROUP BY name;

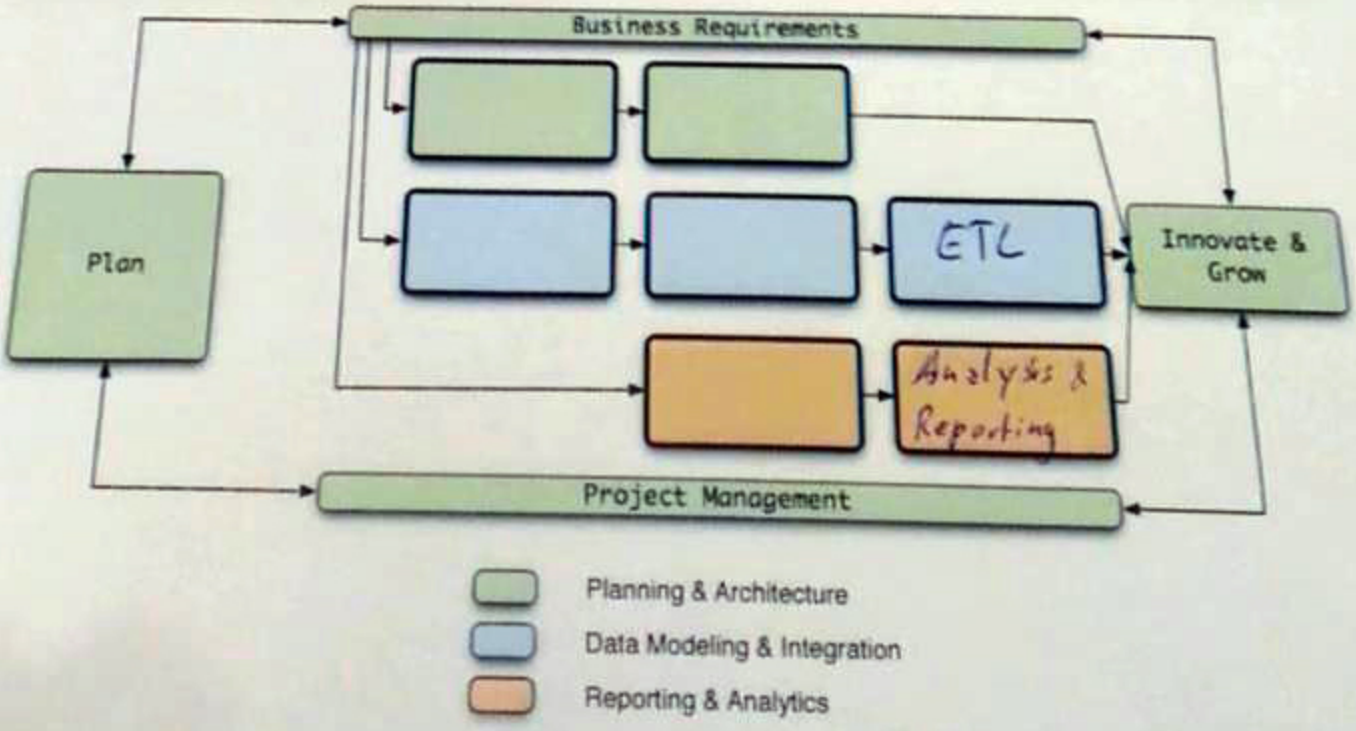
ii. How many articles with an avg. rating higher or equal than 3 are lactose-free? [5]

SELECT COUNT(article_id) FROM Article WHERE lactose_free = 'TRUE';

14. **ETL** (4 points)
Which of the following statements regarding ETL are correct?

- ☒ ETL stands for "Extraction, Transformation, Loading"
- ☒ Data scrubbing uses domain knowledge to detect "dirty" data
- ☐ Business harmonization ensures the use of a common set of business indicators (local units, currencies, periodization etc.) in the DWH.
- ☐ Only virtual DWHs need an ETL process

15. **Kimball's Data Warehouse Development Lifecycle** (4 points)
Complete the following figure that illustrates Kimball's Data Warehouse Development Lifecycle by filling in the activities in each of the three concurrent tracks.



16. **Agile BI** (6 points)
Outline the relevance of agile methods in Business Intelligence. Discuss motivation and objectives and contrast Agile BI methods with more traditional waterfall-style development approaches. Highlight potential advantages as well as problems/risks.

17. **Schema-on-Write vs. Schema-on-Read** (6 points)
Compare and contrast schema-on-write (used in traditional RDBMSs) and schema-on-read (NoSQL) from a Business Intelligence perspective. Explain the differences between these approaches and discuss their respective advantages/disadvantages.