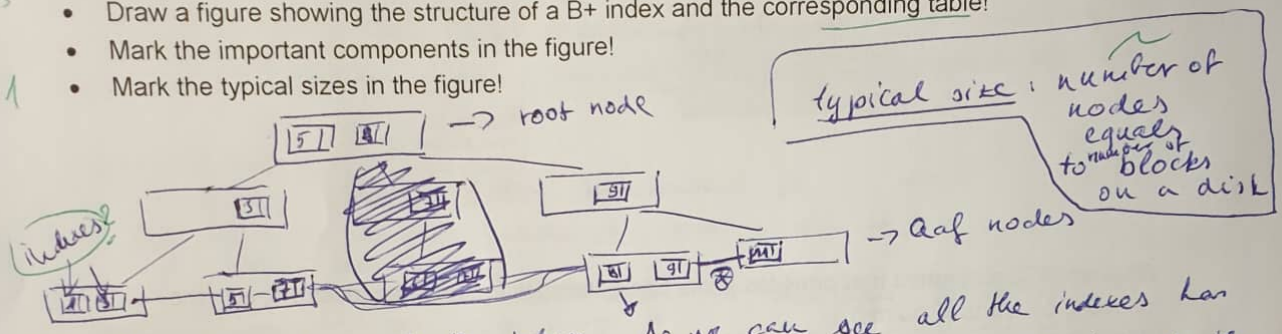
Exam
Database, advanced, 07.01.2020.

Name: Grohmann T. Levente 1
Neptun-code: F93V21

(10+5+5+10+3+3+3+3+6= 48points)

6 A. Indexes (10 points)

2.5
- Draw a figure showing the structure of a B+ index and the corresponding table!
- Mark the important components in the figure!
1
- Mark the typical sizes in the figure!



The leafnodes contains all the indexes. As we can see all the indexes has a pointer that shows us the next item (next index). ⊛ it points towards the next leaf node too.

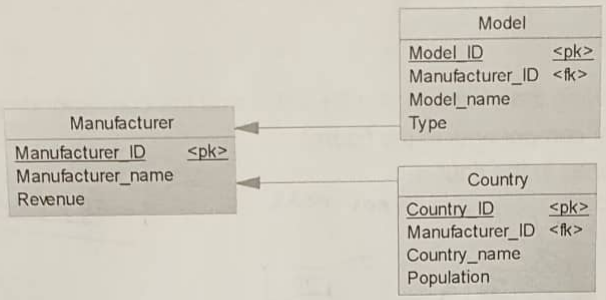
1 Which size of a B+ tree determines the cost of a query (let us consider a simple case, when the query is very selective, returns only a few records)?

The (height) height of the B+ tree determines the cost of a query, that is how fast the tree can return an index for a query.

1.5 What can we say about index usage in case of a query with bad selectivity (high ratio of rows returned)? Please explain your answer!

In case of bad selectivity index usage (~~doesn't~~) doesn't make the query cost less. However if the index is composed for the selected query it can return the records much faster.

B. The following relational schema diagram is given, describing bike manufacturers, their bike models and the countries in which they have representatives.



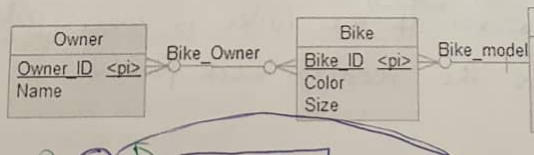
Write an SQL statement that provides some basic statistics for the manufacturers, with the following columns (5 points):

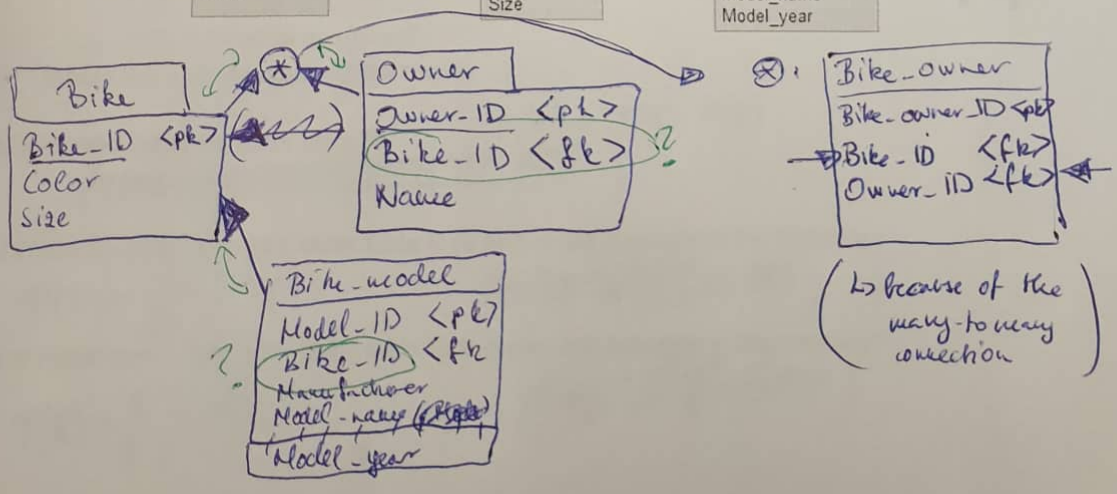
Manufacturer_name	**Sum of population** (sum of population of countries in which the manufacturer has a representative)	**Number of models** (number of models the manufacturer produces)

Joining the tables:
```
FROM (Manufacturer LEFT JOIN Model ON (~~Manufacturer_ID~~)
Manufacturer.Manufacturer_ID = Model.Manufacturer_ID ) AS (~~Table~~) T1
LEFT JOIN Country ON T1.Manufacturer_ID = Country.Manufacturer_ID
```

The query:
```
SELECT Manufacturer_name, SUM(Population), Count(Model_ID)
```

C. Create a relational diagram for the following ER diagram. Mark primary and foreign keys clearly! Use the relational schema diagram notation of task B! (5 points)





Exam
Database, advanced, 07.01.2020.

Name: Graham F. Leverto 3
Neptun-code: F83V21

5.5

D. Explain major concepts of data warehousing! (10 points)

Motivation, definition: Data warehousing is used for collect data from relational databases and from outher data sources to ensure and structure the data for enterprise usage, decision making, statistics etc.

2

Architecture, major elements:

0.5

- Data Source: Relational DBs, Data centrums, etc ← E
- Ensure the data, clean it, and reorder it to a well-structured for. ← T
- End-users: who are making queris from the Ware house datasystem and can even change it.

ETL:
Extract
Transition
Load

Challenges (some):

0

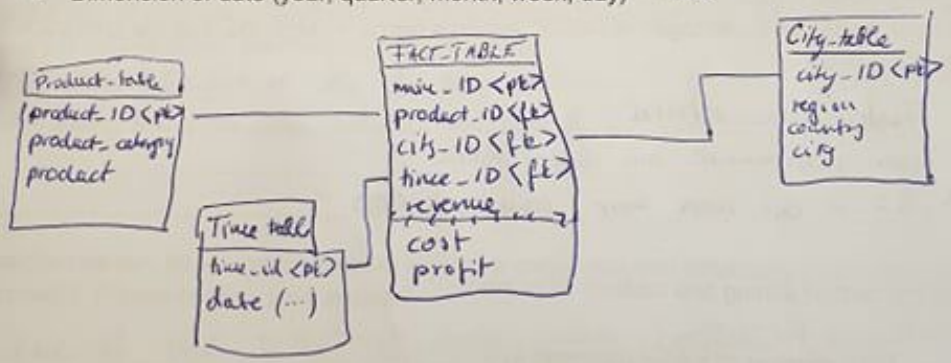
Eventual consistency: (all the infos are the same for all the users after times after every query and change.)

CAP theorem: Capacity, Availability, Partical Nleave ⇒ only 2 can be guaranteed.

Draw a star schema for the following case! (Use the relational schema diagram notation in task ...!

3

- Measures: revenue, cost, profit
- Dimension 1: region, country, city Place
- Dimension 2: product category, product Product
- Dimension 3: date (year, quarter, month, week, day) Time



0

Possibilities to improve the performance/optimize a data warehouse (in a wide sense)?

→ decentralize the system
→ using indexes (for querys too)
→ user friendly interface to communicate with the DW
→ „Clean" the data from unteresting/not useful information
↳ (Normalize the infos)
→ optimize the structure after time, for example reorder or restructure after many changes.

0 E. What advantages/disadvantages has Hibernate over JDBC? (Mark advantages with "+" sign and the disadvantages with "-" sign!) (3 points)

⊖ (Less) JDBC has (~~better~~) an Object Oriented background with inheritance

0 F. Compare object-relational mappers and objectrelational databases! What are the typical scenarios for their usage? (3 points)

Object-relational mappers are (~~structured for~~ ...) used for mapping huge data with many inconsistent data and info.
Objectrelational database is used for containing huge amount of data from many well-structured tables.

3 G. What advantages/disadvantages has a relational database over the Python Pandas package? (Mark the advantages with "+" sign and the disadvantages with "-" sign!) (3 points)

⊖ Rel. Dat. doesn't have a toolkit for displaying the data and queries like Pandas
⊕ (~~...~~) Performs better with queries (with high selectivity)
⊖ Python Pandas can be used for machine learning, deep learning, and to make Neural Networks

2 H. Compare NoSQL databases (key-value/document stores, MongoDB), to relational ("SQL") databases! Mark the advantages of NoSQL systems with a "+" sign, the disadvantages with a "-" sign. (3 points)

⊕ Faster data retrieval
⊕ Better performance on bigger data
⊖ with SQL we can make more complex queries

I. What is the cost of joining two tables? (6 points)

- *Table R*
  - *T(R): number of tuples (records) in R*
  - P(R) – number of pages in R
- Table S
  - T(S) – number of tuples in S
  - P(S) number of pages in S
- Ignore the cost of outputting the result!

Cost of nested loop join, outer table R: $P(R) + T(R) \cdot P(S)$

2 (~~P(R) + P(S) · [T(S) · T(R)]~~ ·)

Cost of block nested loop join, outer table R (B pages are available in the memory):

2 $P(R) + \frac{P(R)}{B+1} \cdot P(S)$ (~~...~~)

Cost of block nested loop join, outer table S (B pages are available in the memory):

2 $P(S) + \frac{P(S)}{B-1} P(R)$ (~~...~~)