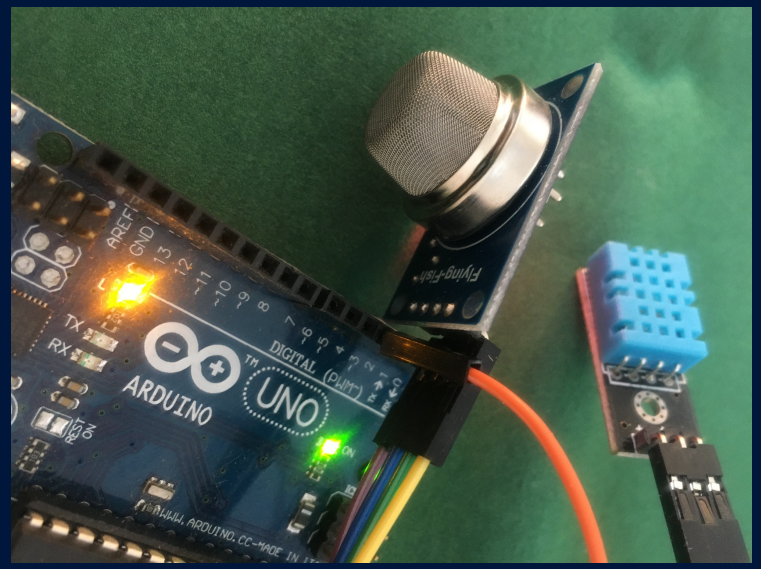
cphbusiness

COPENHAGEN BUSINESS ACADEMY





DATA ENGINEERING



FLOW 3 – Foreløbig plan

Uge 10	07.11.2022	intro til dataforespørgsler	Intro - API, Mongo, SQL og webscraping
	08.11.2022	Webscraping	Webscrapping: Case EDC, Bilbasen
	09.11.2022		
	10.11.2022	Webscraping / MongoDB	Mongodb
	11.11.2022	Webscraping	Præsentation af OLA
Uge 11	14.11.2022	SQL	MySQL og R
	15.11.2022	SQL	MySQL: Case Northwind
	16.11.2022		
	17.11.2022	SQL	MySQL: Case Northwind
	18.11.2022	SQL	Arbejde med OLA
Uge 12	21.11.2022	Cloud Computing	AWS - server og services
	22.11.2022	Cloud Computing	API og Mongo: Casse smart city Aarhus
	23.11.2022		
	24.11.2022	Cloud Computing	Case: PR Flights, R & Mongo på AWS
	25.11.2022	Cloud Computing	ML på AWS
Uge 13	28.11.2022	IOT	Internet of Things
	29.11.2022	IOT	Case: Afstands-sensor
	30.11.2022		
	01.12.2022	IOT	Case:Afstands-sensor
	02.12.2022	OLA	
Uge 14	05.12.2022	Webscraping & NLP	Intro til NLP
	06.12.2022	Webscraping & NLP	Sentiment på boligannoncer
	07.12.2022		
	08.12.2022		
	09.12.2022	Opsamling	Præsentation af OLA, eksamensforberedelse

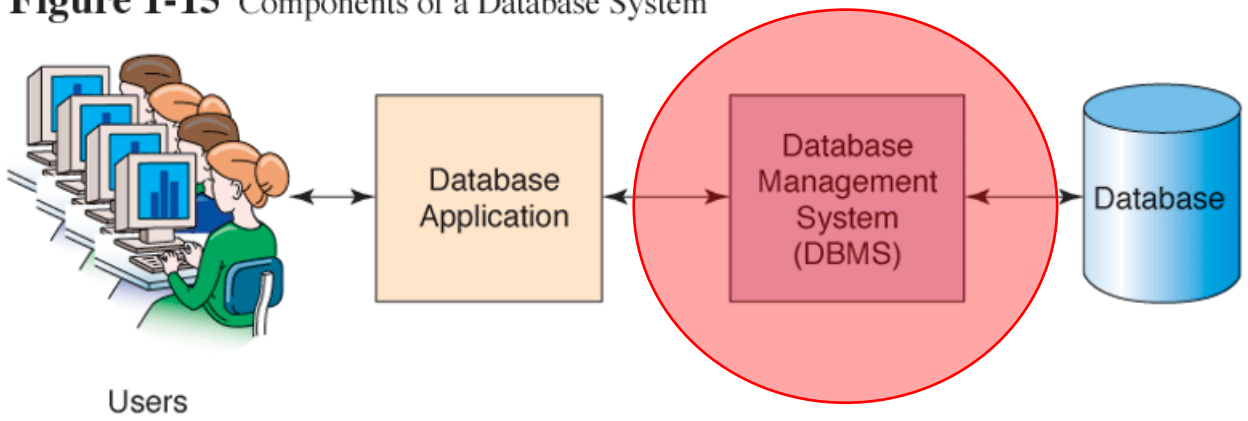
Agenda - CRUD

- WEBSCRAPE
  - Afslutning med EDC
- SQL
  - Intro til WorkBench
  - SELECT (conditions,join,aggregation)
    - WORLD-databasen
    - Øvelser i MySQL
  - UPDATE og INSERT (Northwind)
    - Demo
    - Øvelser
  - CREATE (Cars)
    - Tilføj pris og forhandler
- SQL i R
  - SQL-queries fra R

Database Management System (DBMS)

- Et **software system** som giver brugere mulighed for at definere, oprette og vedligeholde en database samt kontrolleret adgang til denne.

Figure 1-15 Components of a Database System



Relationel Database

Den mest udbredte DBMS type.

- En database har et navn
- En database har en eller flere tabeller

- Hver tabel har et navn
- Hver tabel har en eller flere kolonner
- Hver kolonne har navn og datatype

Eksempel:
Database hedder **test**
indeholder 9 tabeller

Tabel eksempel
Medarbejdere (emp)

Kolonner – har navn og simpel datatype

Rækker – indeholder relaterede værdier

empno	ename	job	mgr	hiredate	sal	deptno
7369	SMITH	CLERK	7902	12/17/1980	800	20
7499	ALLEN	SALESMAN	7698	02/20/1981	1600	30
7521	WARD	SALESMAN	7698	02/22/1981	1250	30
7566	JONES	MANAGER	7839	04-02-1981	2975	20
7654	MARTIN	SALESMAN	7698	09/28/1981	1250	30
7698	BLAKE	MANAGER	7839	05-01-1981	2850	30
7782	CLARK	MANAGER	7839	06-09-1981	2450	10
7788	SCOTT	ANALYST	7566	04/19/1987	3000	20
7839	KING	PRESIDENT		11/17/1981	5000	10
7844	TURNER	SALESMAN	7698	09-08-1981	1500	30
7876	ADAMS	CLERK	7788	05/23/1987	1100	20
7900	JAMES	CLERK	7698	12-03-1981	950	30
7902	FORD	ANALYST	7566	12-03-1981	3000	20
7934	MILLER	CLERK	7782	01/23/1982	1300	10

Tabel eksempel 2
Medarbejdere (emp)
Afdelinger (dept)

emp

empno	ename	job	mgr	hiredate	sal	deptno
7369	SMITH	CLERK	7902	12/17/1980	800	20
7499	ALLEN	SALESMAN	7698	02/20/1981	1600	30
7521	WARD	SALESMAN	7698	02/22/1981	1250	30
7566	JONES	MANAGER	7839	04-02-1981	2975	20
7654	MARTIN	SALESMAN	7698	09/28/1981	1250	30
7698	BLAKE	MANAGER	7839	05-01-1981	2850	30
7782	CLARK	MANAGER	7839	06-09-1981	2450	10
7788	SCOTT	ANALYST	7566	04/19/1987	3000	20
7839	KING	PRESIDENT		11/17/1981	5000	10
7844	TURNER	SALESMAN	7698	09-08-1981	1500	30
7876	ADAMS	CLERK	7788	05/23/1987	1100	20
7900	JAMES	CLERK	7698	12-03-1981	950	30
7902	FORD	ANALYST	7566	12-03-1981	3000	20
7934	MILLER	CLERK	7782	01/23/1982	1300	10

dept

deptno	dname	loc
10	ACCOUNTING	NEW YORK
20	RESEARCH	DALLAS
30	SALES	CHICAGO
40	OPERATIONS	BOSTON

Er tabellerne logisk forbundne?

SQL - flere formål

DML (**D**ata **M**anipulation **L**anguage)

- Kommandoer som ændrer data i databasen

DDL (**D**ata **D**efinition **L**anguage)

- Kommandoer som definerer databasen

Database forespørgsler har formatet:

```
select …
from …
where …
```

Eksempel:

```
select ename
from emp
where mgr = 7698
```

SQL

Data Definition (DDL)

- **CREATE**
- ALTER
- DROP

Data Manipulation (DML)

- **SELECT**
- **INSERT**
- **UPDATE**
- DELETE

SQL SELECT eksempler

```
SELECT *
FROM emp
```

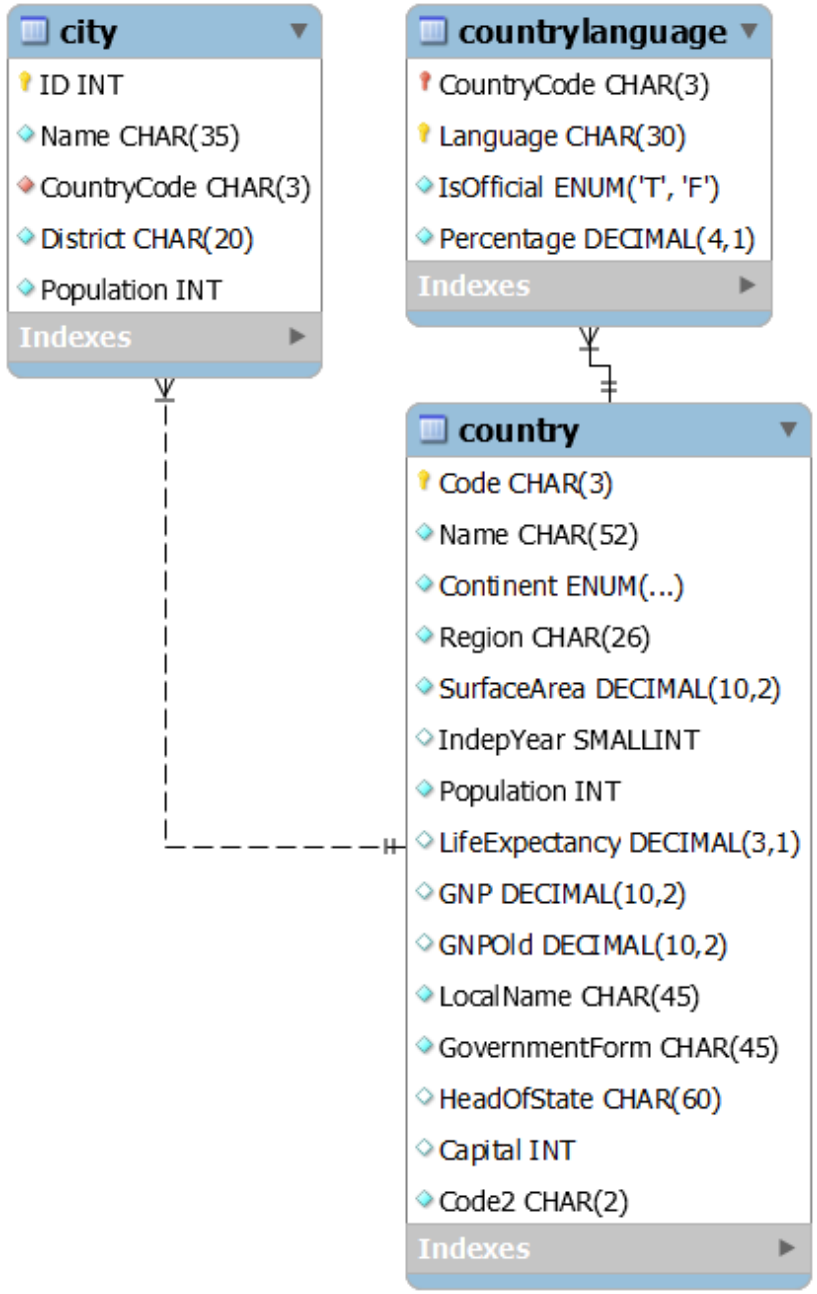
```
SELECT avg(sal)
FROM emp
```

```
SELECT ename, hiredate, sal
FROM emp
WHERE sal > 1000
```

```
SELECT empno
FROM emp
WHERE ename  = 'Smith'
```

empno	ename	hiredate	sal	deptno
7369	SMITH	12/17/1980	800	20
7876	ADAMS	05/23/1987	1100	20
7900	JAMES	12-03-1981	950	30
7934	MILLER	01/23/1982	1300	10

World DB ER-diagram



World DB select

1) I hvilket distrikt ligger byen 'Stanley’?

2) Er færøsk et officielt sprog på Færøerne?

3) Hvad er `CountryCode` for 'Sri Lanka'

4) Hvilket land har det mindste areal?

5) Hvor mange amerikanske byer er med i DB'en?

6) I hvilket land taler mere end halvdelen af befolkningen 'Pashto'?

7) Hvad er den samlede befolkning i de danske byer der er med i DB'en?

8) Hvilke sprog tales i byen 'Nassau'?

9) Hvilket land har den højeste `LifeExpectancy`?

10) Hvilke lande har flere indbyggere end Russland?

World DB - Joins

1) I hvilket distrikt ligger byen 'Stanley'?

2) Er færøsk et officielt sprog på Færøerne?

```
#sp 2
select language, name from Country
join CountryLanguage on country.code=CountryLanguage.CountryCode
where name like "%oe%"
order by 2;
```

'Pashto'?

7) Hvad er den samlede befolkning i de danske byer der er med i DB'en?

8) Hvilke sprog tales i byen 'Nassau'?

9) Hvilket land har den højeste `LifeExpectancy`?

10) Hvilke lande har flere indbyggere end Russland?

World DB - subqueries

1) I hvilket distrikt ligger byen 'Stanley'?

2) Er færøsk et officielt sprog på Færøerne?

3) Hvad er `CountryCode` for 'Sri Lanka'

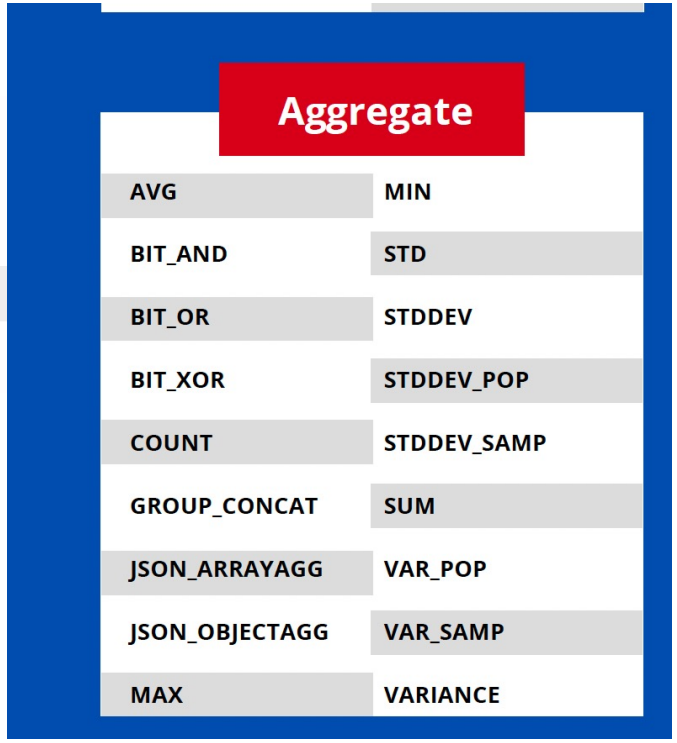
4) Hvilket land har det mindste areal?

```
15
16 •  select name,surfacearea from Country
17    where SurfaceArea in (select min(surfacearea) from country);
```

World DB - Aggregation

5) Hvor mange amerikanske byer er med i DB'en?

```
20    #sp 5
21 •  select count(*) as "cities", co.name from city ci, country co
22    where ci.countrycode=co.code
23    group by co.name
24    order by 1 desc;
```



World DB join

6) I hvilket land taler mere end halvdelen af befolkningen 'Pashto'?

"list all land og deres sprog"

```
26    #sp 6
27 •  select cl.language, cl.percentage,co.name, co.population from CountryLanguage cl, country co
28    where cl.CountryCode=co.code
29    and cl.language like "Pash%"
30    order by 2 desc;
```

World DB Aggregation

7) Hvad er den samlede befolkning i de danske byer der er med i DB'en?

```
32    # sp 7
33 •  select co.name,sum(ci.population) as "sum pop"from city ci, country co
34    where ci.countrycode=co.code
35    #and co.name like "Den%"
36    group by co.name
37    order by 1 desc;
```

World DB joins

8) Hvilke sprog tales i byen 'Nassau'?

```
#sp 8 Sprog i nassau?
select cl.language,ci.name,co.name from countrylanguage cl, city ci, Country Co
where ci.countrycode=co.Code
and co.code=cl.countrycode
and ci.name like "Nassa%"
order by 2;
```

World DB subquery

9) Hvilket land har den højeste `LifeExpectancy`?

```
45
46  # sp 9 Højest life-expect
47  Select lifeexpectancy, name from Country
48  where LifeExpectancy in (select max(lifeexpectancy) from country)
49
```

World DB select

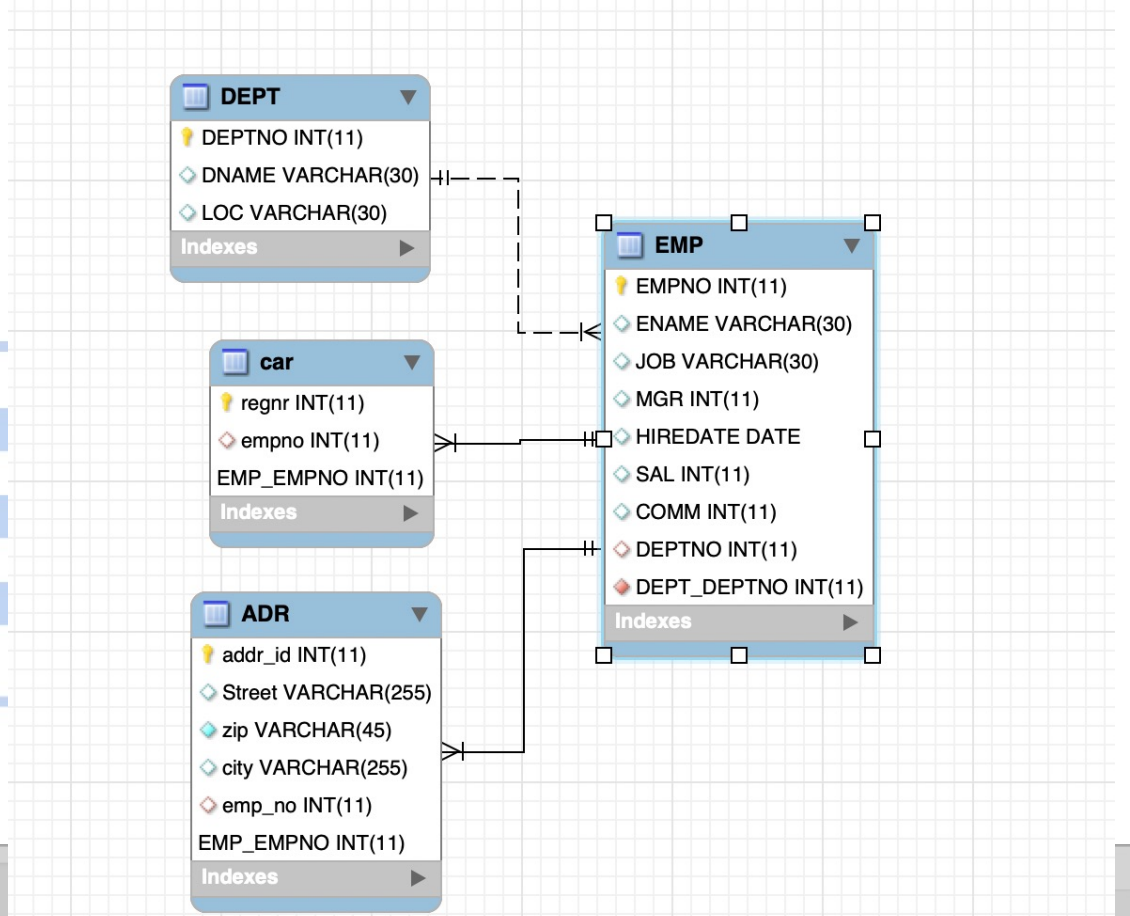
10) Hvilke lande har flere indbyggere end Rusland?

```
50  # sp 10 Flere indb end Rusland?
51  select name, population from country
52  where population > (select population from country
53                      where name like "Rus%")
54  order by 2;
55
```

MySQL Skema

Browsing

```
SHOW DATABASES
SHOW TABLES
SHOW FIELDS FROM table / DESCRIBE table
SHOW CREATE TABLE table
SHOW PROCESSLIST
KILL process_number
```



Info | Columns | Indexes

Column	Type	Default...	Nullable	Charac			ement
addr_id	int(11)		NO				
Street	varchar(255)		YES	utf8mb4	utf8mb4_090...	select,insert,update,references	
zip	varchar(45)		NO	utf8mb4	utf8mb4_090...	select,insert,update,references	
city	varchar(255)		YES	utf8mb4	utf8mb4_090...	select,insert,update,references	
emp_no	int(11)		YES			select,insert,update,references	

Mysql – Filtre

Select

```
SELECT * FROM table
SELECT * FROM table1, table2, ...
SELECT field1, field2, ... FROM table1, table2, ...
SELECT ... FROM ... WHERE condition
SELECT ... FROM ... WHERE condition GROUPBY field
SELECT ... FROM ... WHERE condition GROUPBY field HAVING condition2
SELECT ... FROM ... WHERE condition ORDER BY field1, field2
SELECT ... FROM ... WHERE condition ORDER BY field1, field2 DESC
SELECT ... FROM ... WHERE condition LIMIT 10
SELECT DISTINCT field1 FROM ...
SELECT DISTINCT field1, field2 FROM ...
```

Conditions

```
field1 = value1
field1 <> value1
field1 LIKE 'value _ %'
field1 IS NULL
field1 IS NOT NULL
field1 IS IN (value1, value2)
field1 IS NOT IN (value1, value2)
condition1 AND condition2
condition1 OR condition2
```

Mysql – Filtre

What is Inner Join?

An Inner Join returns only the rows that have matching values in both the tables (we are considering here the join is done between the two tables).

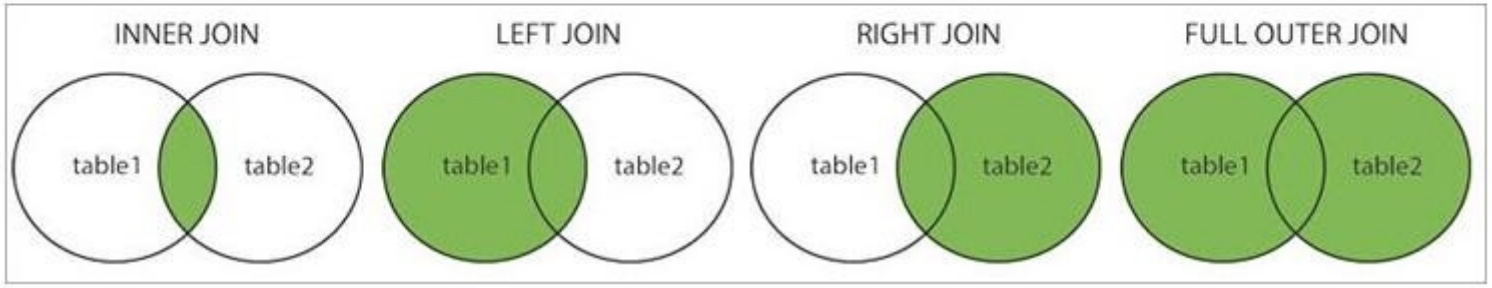
What is Outer Join?

The Outer Join includes the matching rows as well as some of the non-matching rows between the two tables. An Outer join basically differs from the Inner join in how it handles the false match condition.

There are 3 types of Outer Join:

- **Left Outer Join**: Returns all the rows from the LEFT table and matching records between both the tables.
- **Right Outer Join**: Returns all the rows from the RIGHT table and matching records between both the tables.
- **Full Outer Join**: It combines the result of the Left Outer Join and Right Outer Join.

Difference between Inner and Outer Join



https://www.softwaretestinghelp.com/inner-join-vs-outer-join/

avgsal	dname
1567	SALES
2175	RESEARCH
2917	ACCOUNTING

avg sal	dname
NULL	OPERATIONS
1567	SALES
2175	RESEARCH
2917	ACCOUNTING

WORLD Databasen

Spørgsmål til "world" databasen

1) I hvilket distrikt ligger byen 'Stanley'?

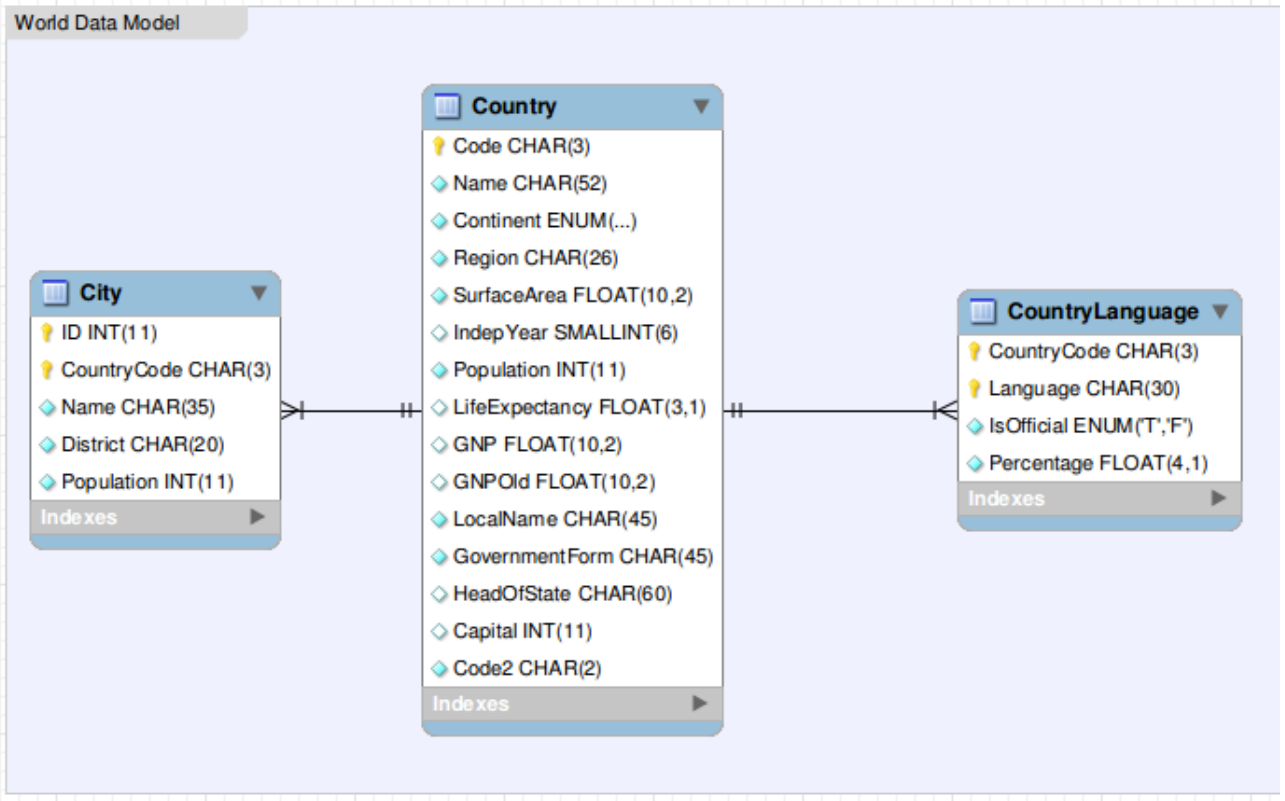
2) Er færøsk et officielt sprog på Færøerne?

3) Hvad er `CountryCode` for 'Sri Lanka'

4) Hvilket land har det mindste areal?

5) Hvor mange amerikanske byer er med i DB'en?

6) I hvilket land taler mere end halvdelen af befolkningen 'Pashto'?



WORLD Databasen

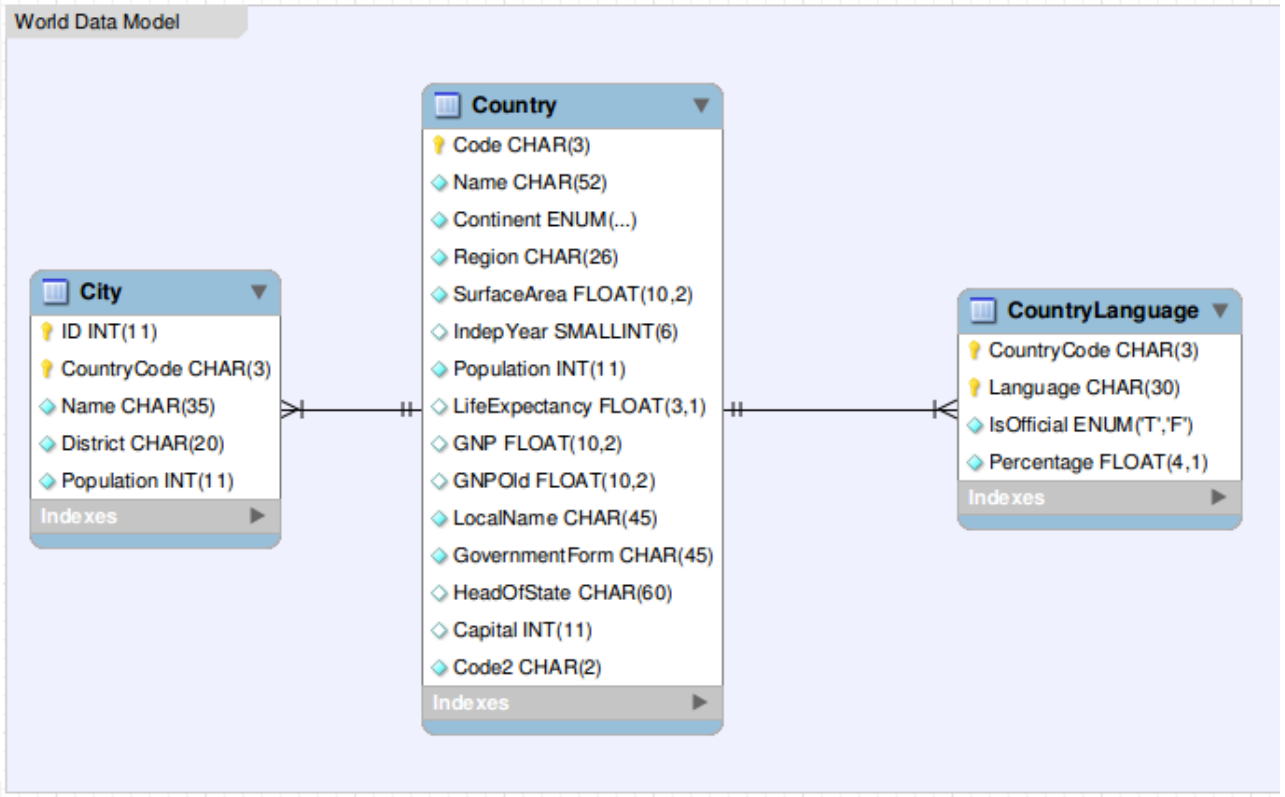
Spørgsmål til "world" databasen

7) Hvad er den samlede befolkning i de danske byer der er med i DB'en?

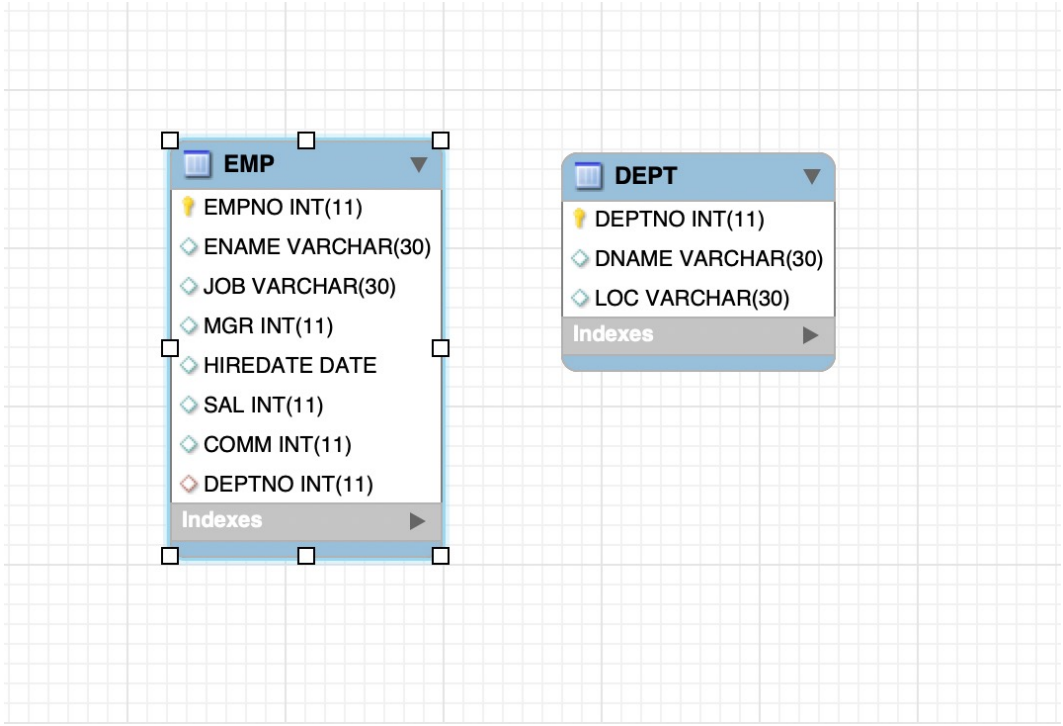
8) Hvilke sprog tales i byen 'Nassau'?

9) Hvilket land har den højeste `LifeExpectancy`?

10) Hvilke lande har flere indbyggere end Rusland?



EMPLOYEE Databasen



Øvelse:

Find max-værdien af DEPTNO

Indsæt en ny afdeling, DATASCIENCE (Seattle) med passende DEPTNO

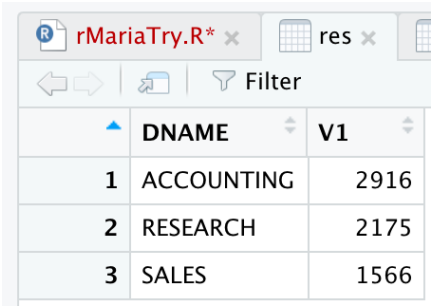
Tilføj dig selv som medarbejder med passende data (brug transaktion)

Prøv at tilføje en medarbejder til en ikke-eksisterende dept

Kan du lave en query så du får flg:

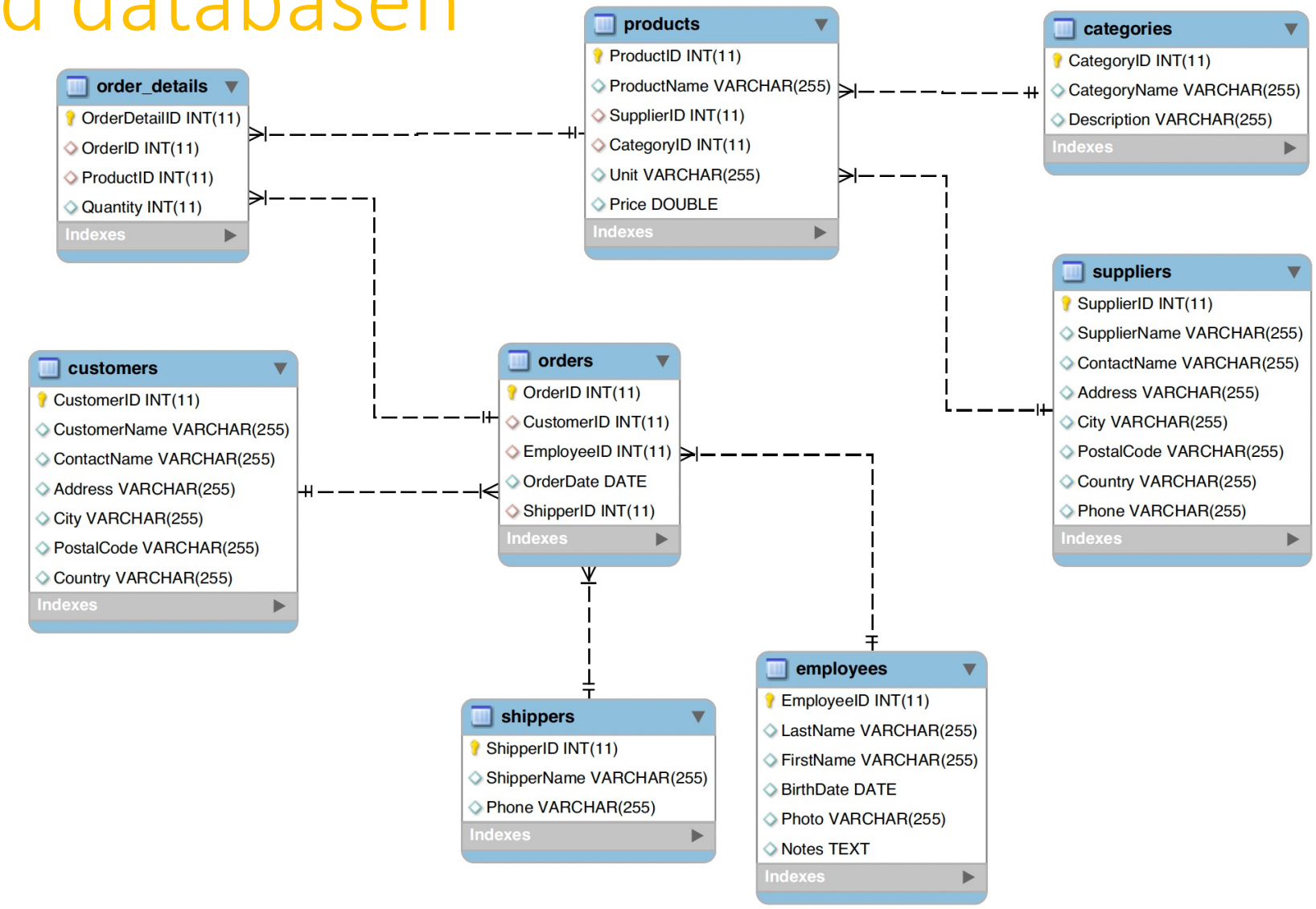
gennemsnit	afdeling
1566	SALES
2175	RESEARCH
2916	ACCOUNTING

Og samme resultat i R

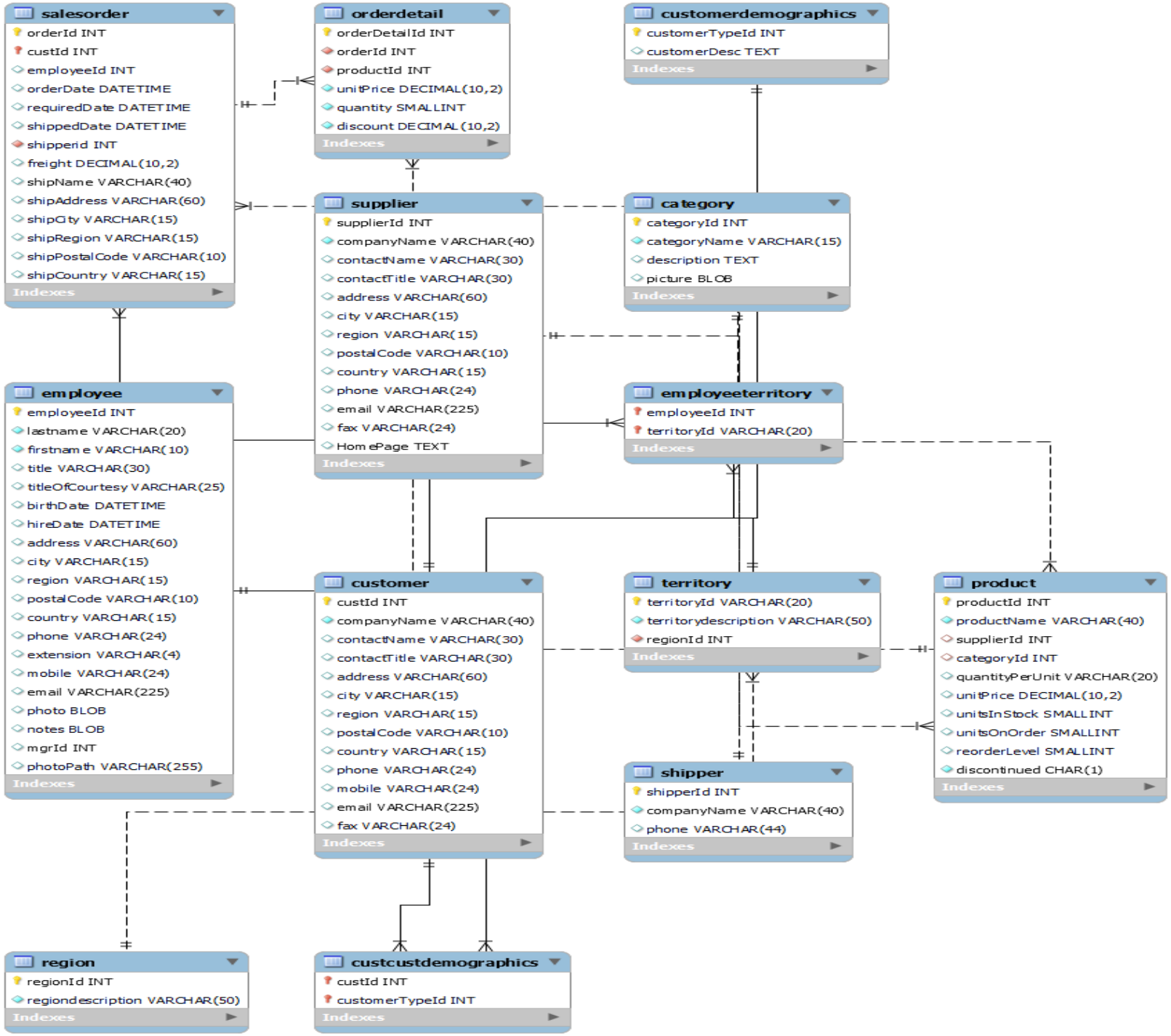


	DNAME	V1
1	ACCOUNTING	2916
2	RESEARCH	2175
3	SALES	1566

Northwind databasen

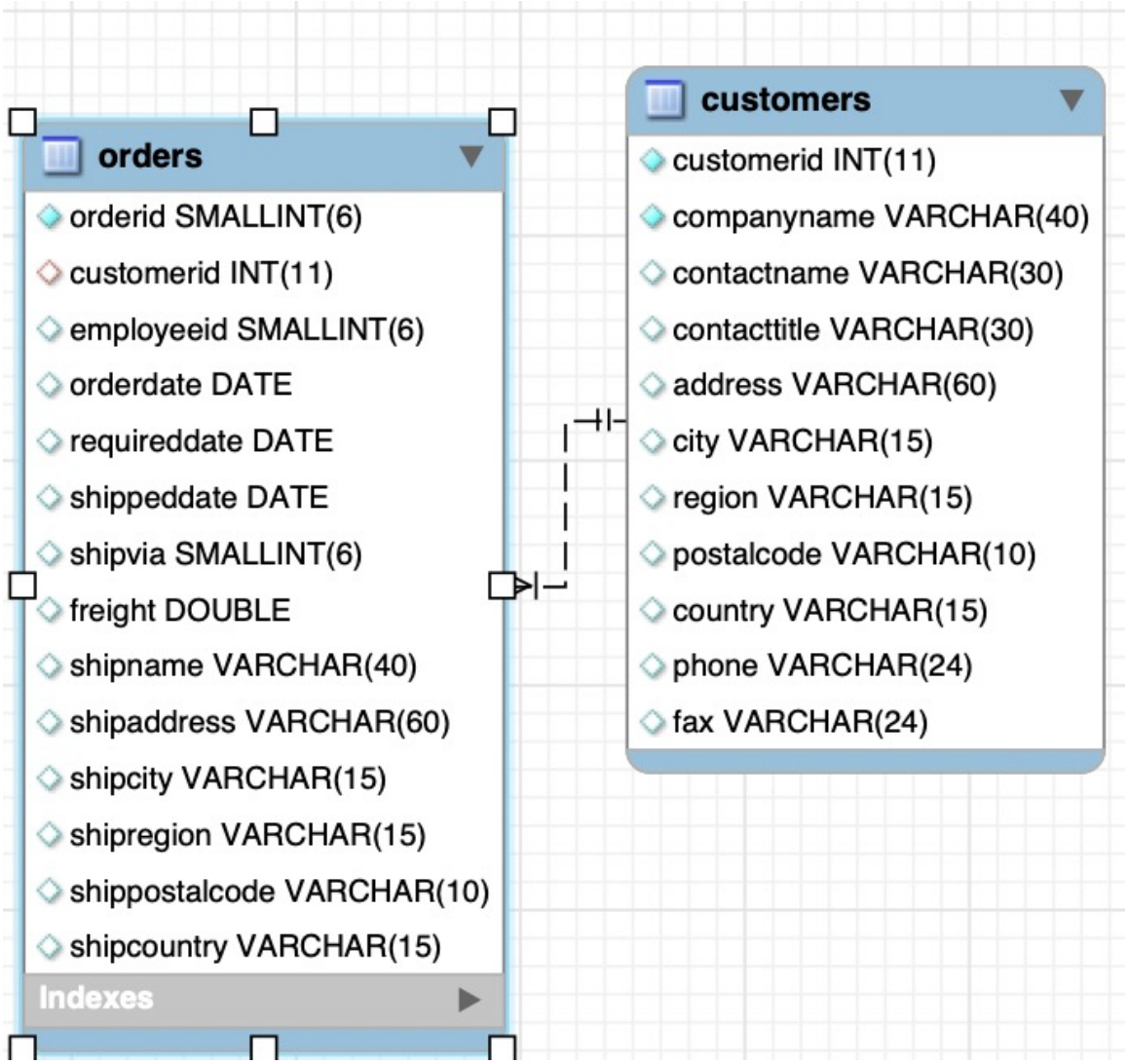


Northwind databasen



Northwind databasen - videodude

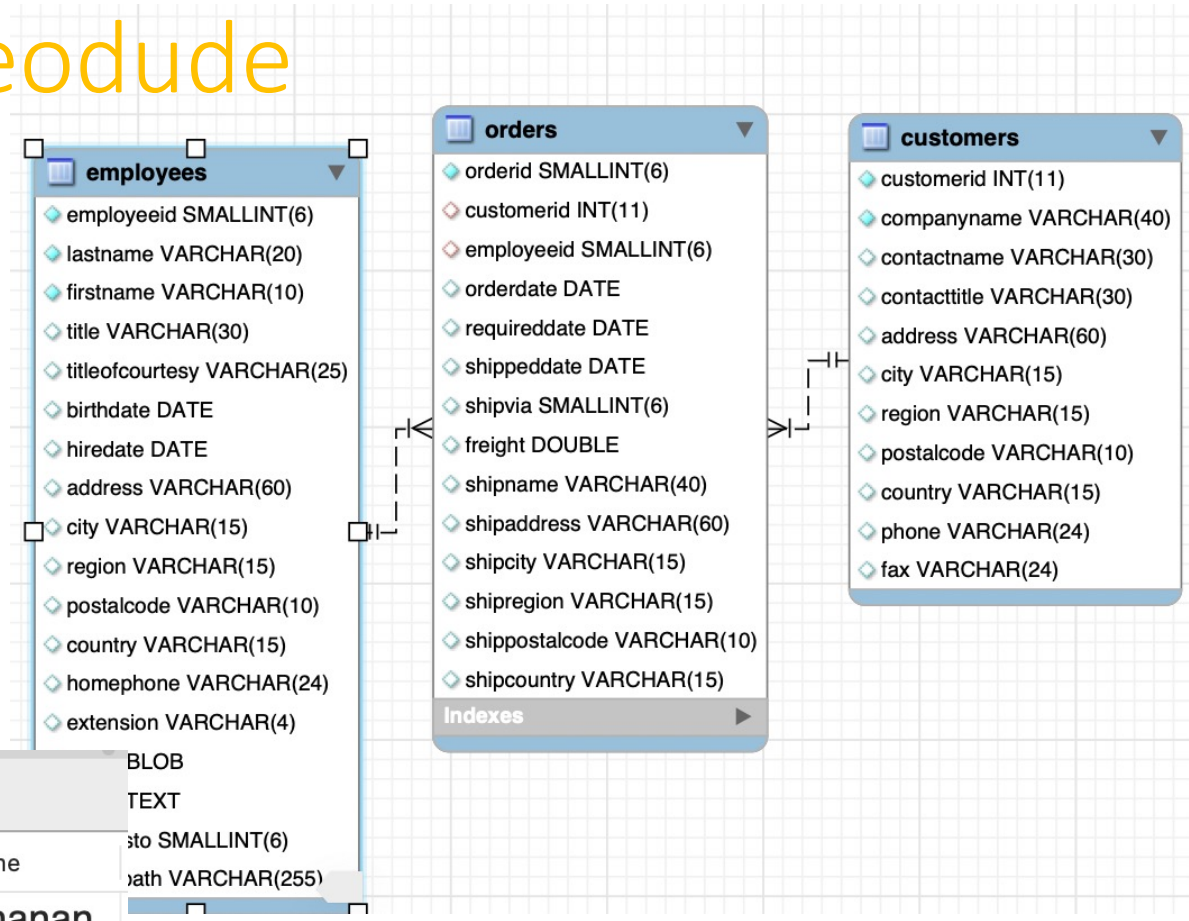
Liste over alle kunders ordrer og de medarbejdere som hjalp dem?



Northwind databasen - videodude

Hvilke medarbejdere hjalp med handlerne?

```
select companyname, orderdate, shipcountry, firstname, lastname
from orders
join customers on orders.customerid=customers.customerid
join employees on orders.employeeid=employees.employeeid
order by 2;
```

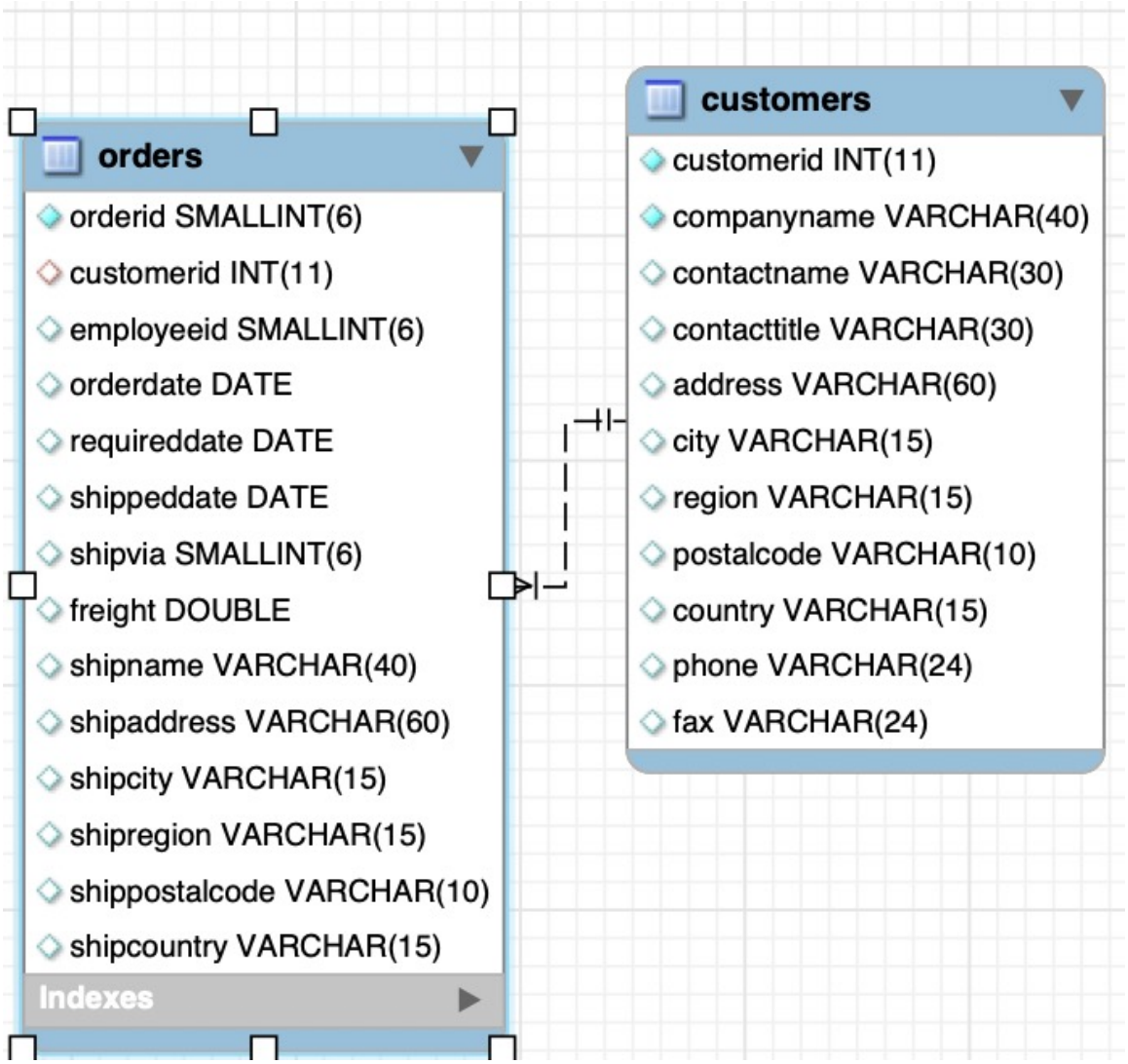


Result Grid Filter Rows: Search Export:

companyname	orderdate	shipcountry	firstname	lastname
Vins et alcools Chevalier	1996-07-04	France	Steven	Buchanan
Toms Spezialitäten	1996-07-05	Germany	Michael	Suyama
Victuailles en stock	1996-07-08	France	Janet	Leverling
Hanari Carnes	1996-07-08	Brazil	Margaret	Peacock

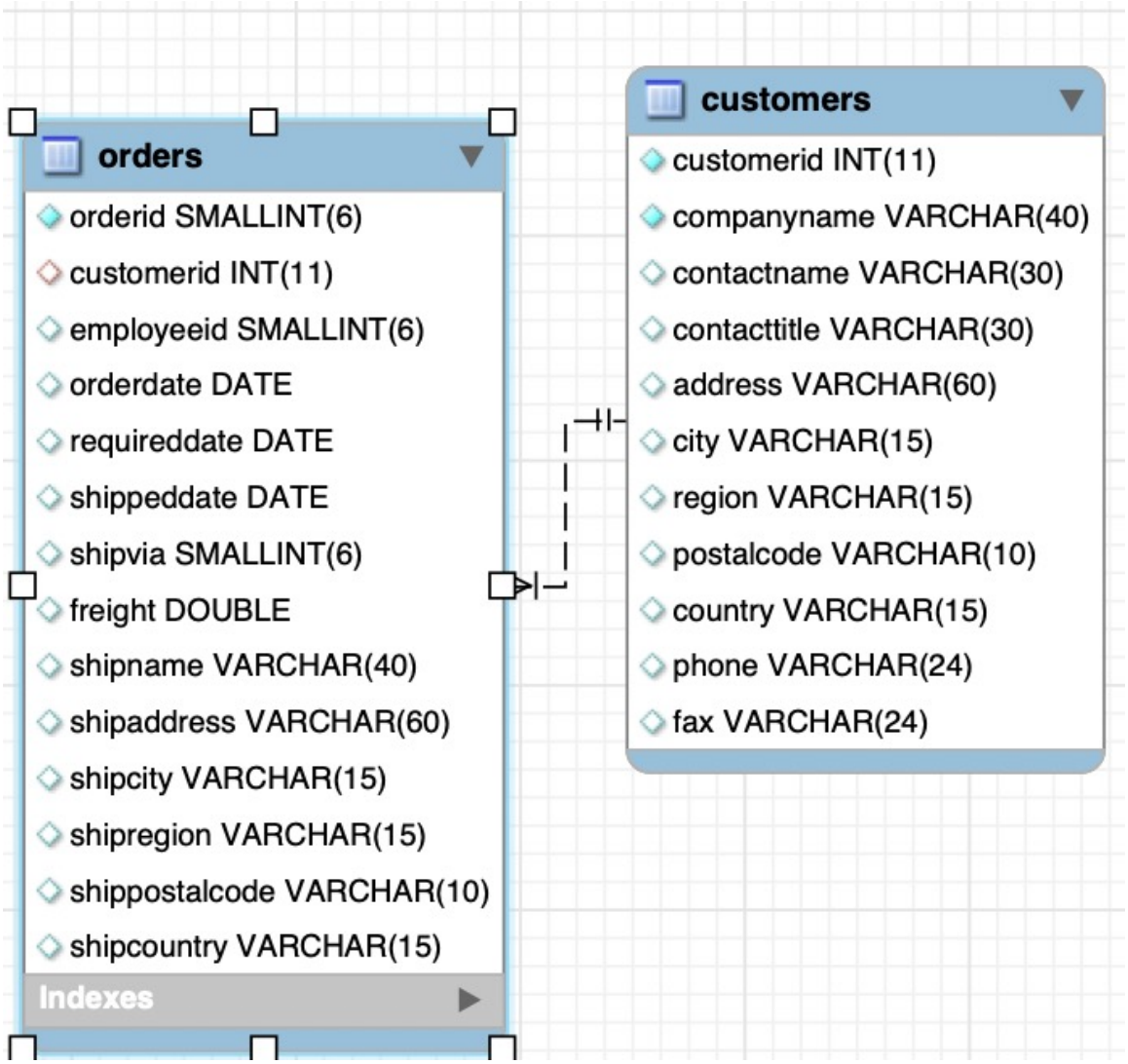
Northwind databasen

Liste over alle kunders ordrer og de medarbejdere som hjalp dem?



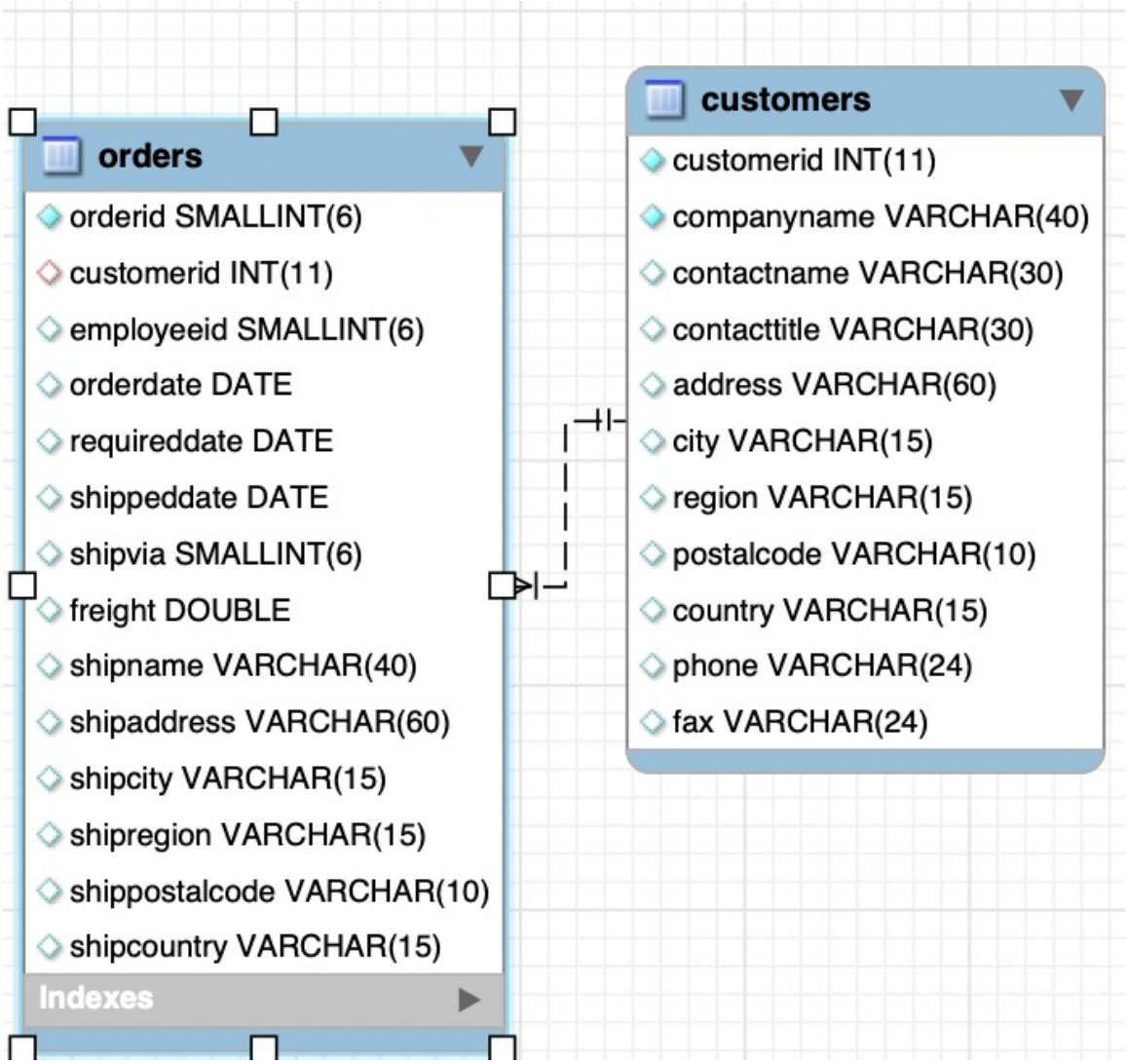
Northwind databasen

Hvilke tre kunder placerede flest ordrer?



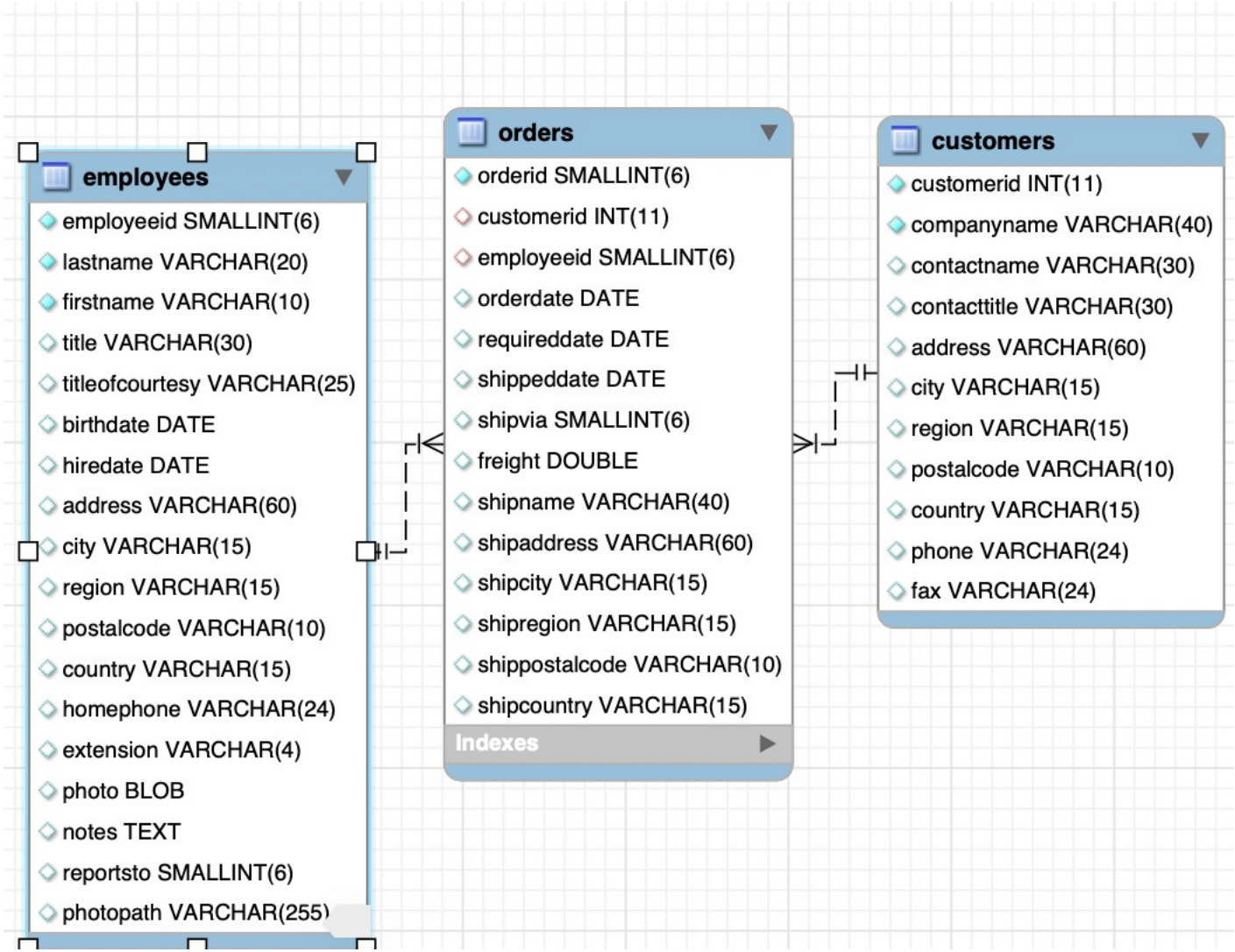
Northwind databasen

Hvilke kunder placerede ingen ordre?



Northwind databasen

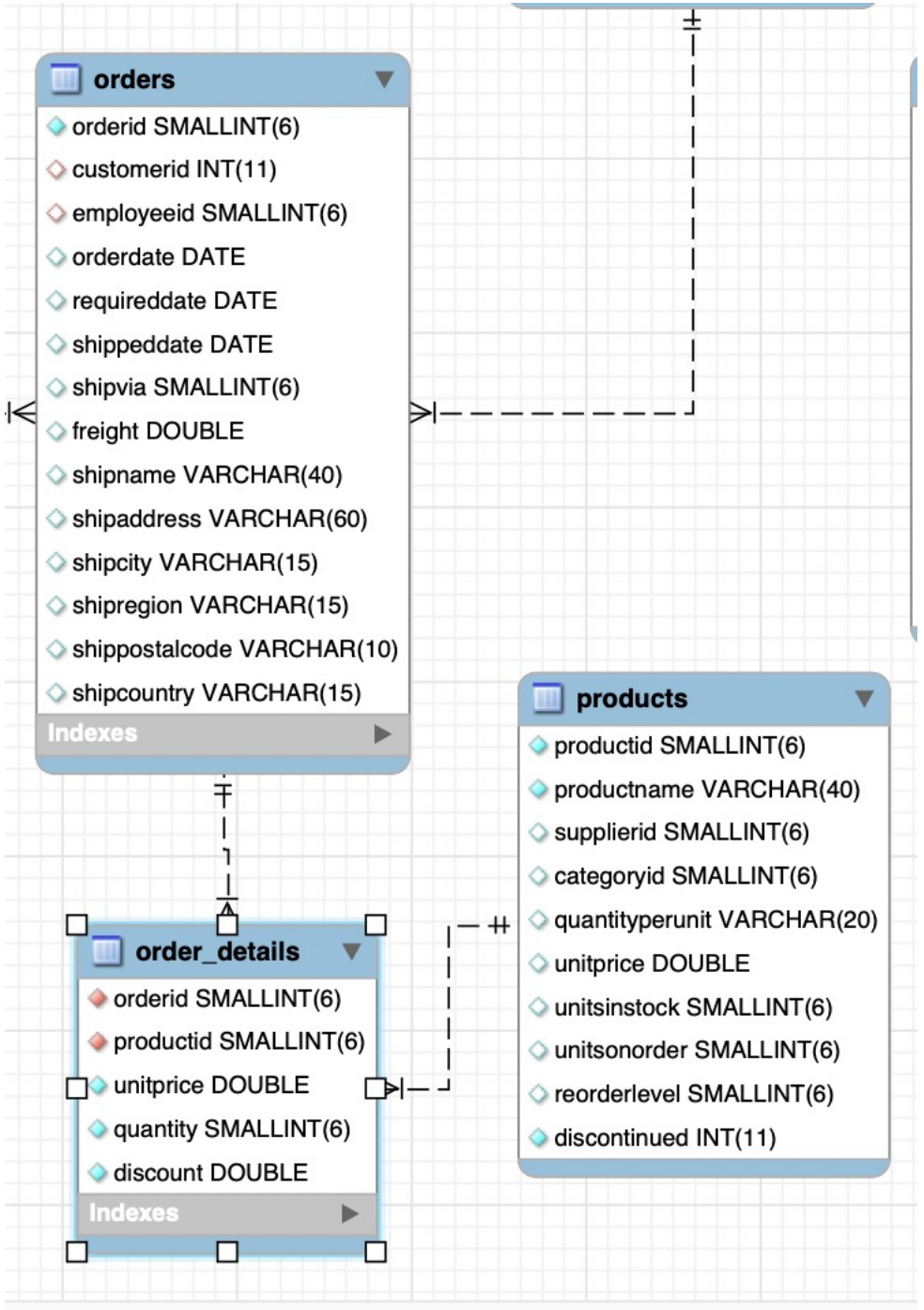
Hvilke medarbejdere hjalp ”Ernst Handel”?



Northwind databasen

Hvilke tre produkter solgte mest

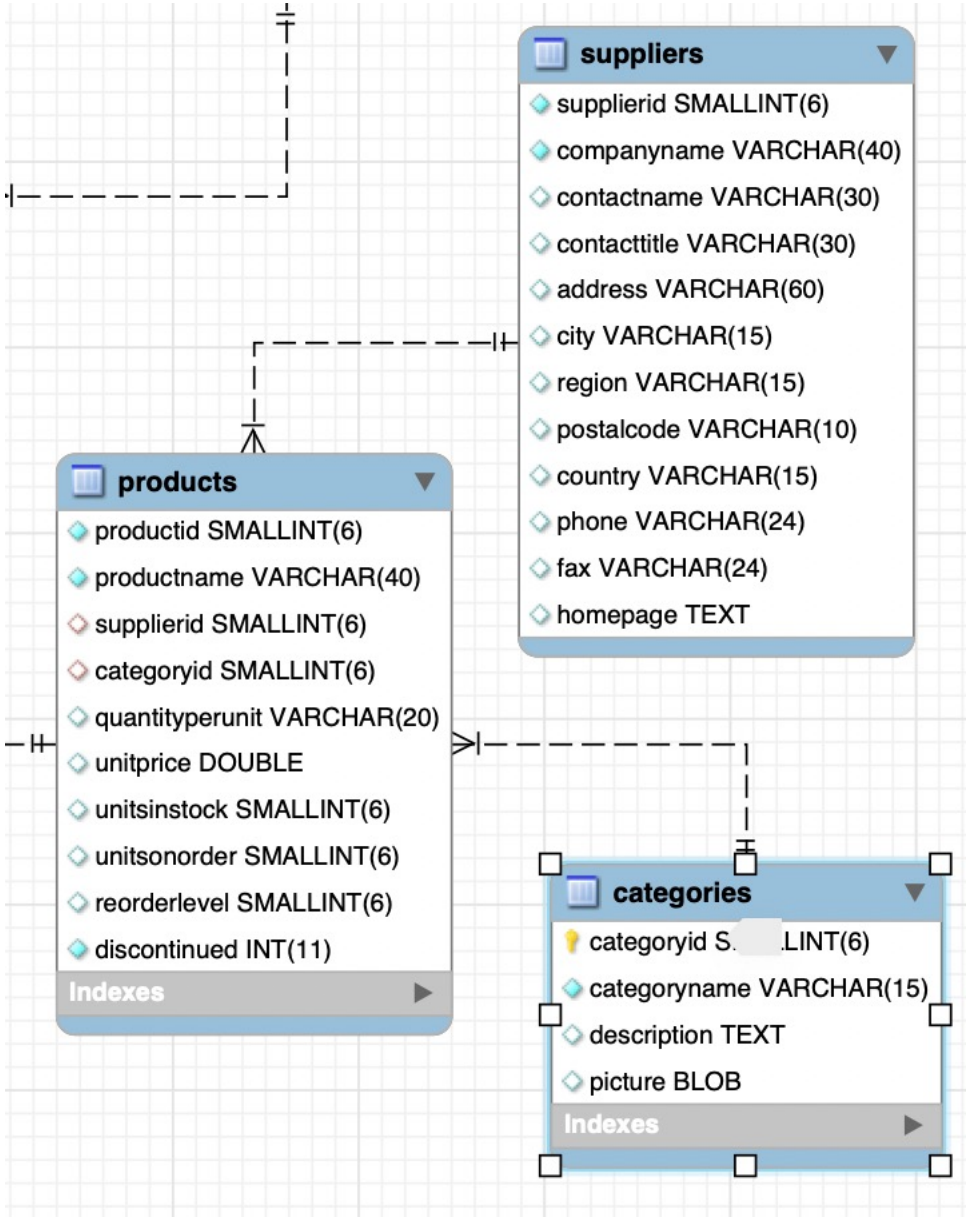
- mængde
- kroner



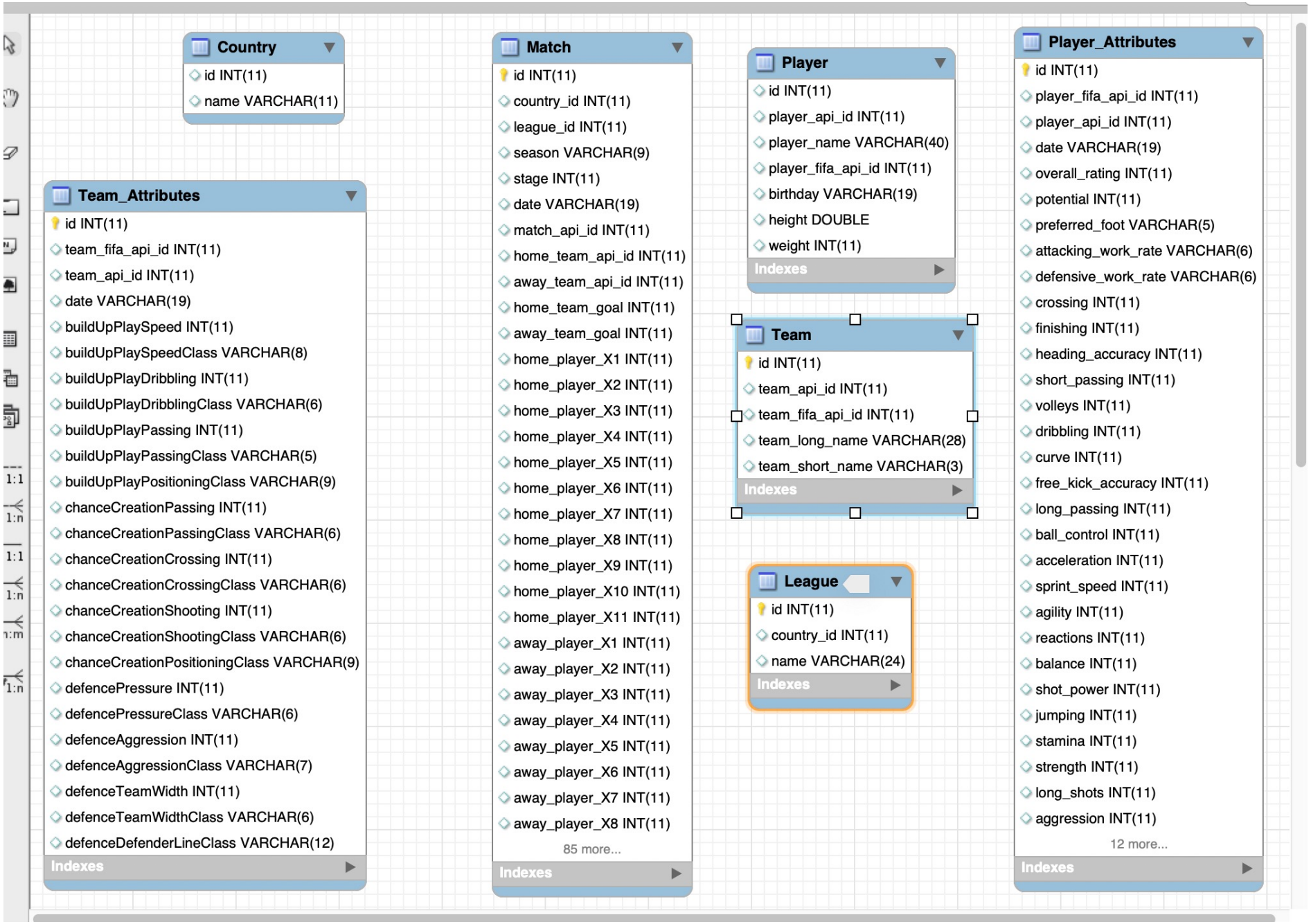
Northwind databasen

Hvilke tre produkter solgte mest

- mængde
- kroner



SOCCER databasen



MySQL from R

- Connecting and disconnecting
  - Connecting to and disconnecting from databases
    - dbConnect(MariaDB(), ..)
- Tables
  - Reading and writing entire tables
    - dbWriteTable(con, "mycarstable”,mycarsdf)
    - mycardf <-dbReadTable(con, "mycarstable”)
- Results
  - More control for sending queries and executing statements
    - dbGetQuery(con, "SELECT * FROM city limit 3")
    - dbExecute (con,”INSERT INTO city (Name,Population) VALUES (‘Lviv’,123123)”)

MySQL from R

rMariaTry.R* | res | dfEmpDB | suppl | sorted | cat | prod | cat_prod_join | nrs

Filter

	DEPTNO	EMPNO	ENAME	JOB	MGR	HIREDATE	SAL	COMM	DNAME	LOC
1	10	7782	CLARK	MANAGER	7839	1981-06-09	2450	NA	ACCOUNTING	NEW YORK
2	10	7839	KING	PRESIDENT	NA	1981-11-17	5000	NA	ACCOUNTING	NEW YORK
3	10	7934	MILLER	CLERK	7782	1982-01-23	1300	NA	ACCOUNTING	NEW YORK
4	20	7369	SMITH	CLERK	7902	1980-12-17	800	NA	RESEARCH	DALLAS
5	20	7876	ADAMS	CLERK	7788	1987-05-23	1100	NA	RESEARCH	DALLAS
6	20	7566	JONES	MANAGER	7839	1981-04-02	2975	NA	RESEARCH	DALLAS
7	20	7902	FORD	ANALYST	7566	1981-12-03	3000	NA	RESEARCH	DALLAS
8	20	7788	SCOTT	ANALYST	7566	1987-04-19	3000	NA	RESEARCH	DALLAS
9	30	7499	ALLEN	SALESMAN	7698	1981-02-20	1600	300	SALES	CHICAGO
10	30	7698	BLAKE	MANAGER	7839	1981-05-01	2850	NA	SALES	CHICAGO
11	30	7521	WARD	SALESMAN	7698	1981-02-22	1250	500	SALES	CHICAGO
12	30	7654	MARTIN	SALESMAN	7698	1981-09-28	1250	1400	SALES	CHICAGO
13	30	7844	TURNER	SALESMAN	7698	1981-09-08	1500	0	SALES	CHICAGO
14	30	7900	JAMES	CLERK	7698	1981-12-03	950	NA	SALES	CHICAGO

Indlæse tabellerne hver for sig.
Udfør operationer i R (merge,aggregate)

rMariaTry.R* | res

Filter

	DNAME	V1
1	ACCOUNTING	2916
2	RESEARCH	2175
3	SALES	1566