Data Science, Data Engineering, Data Management... Data Art?

Big Data in 30 hours

Lecture 3: Data warehousing

Cracow University of Technology (Politechnika Krakowska)
Faculty of Physics, Mathematics and Computer Science
Original course title:
Methods and tools for big data analysis, code WFMiI I oIIS D2 17/18
Graduate Master's degree studies, 2nd year
2018/2019
Venue: Room F017, ul. Podchorążych 1, Kraków (budynek wydz. Fiz Mat i Inf)
Time: Every Monday, 12:45, starting Mon Oct 8th 2018 through end Jan 2019

Let's start with previous lecture recap

Lecture 2 was about: Relational Databases

Lecture 2 recap: sqlite ?

Task	command
To launch	?
To exit / interrupt	?
Data exploration	?
Ingest csv	?
Measure time	?
Write to file	?
Explain plan	?

Lecture 2 recap, contd

```
/*what will this return?*/

select c.lastname, count(c.lastname) from invoices as i
inner join customers as c where i.customerid =
c.customerid order by count(c.lastname);

/*what will this return?*/
select i.invoiceid, c.customerid from invoices as i
inner join customers as c where c.lastname =
'Nielsen';

/*what will this return?*/
select i.invoiceid, c.customerid from invoices as i
inner join customers as c;
```

Lecture 2 recap, contd

- /*what will this return?*/

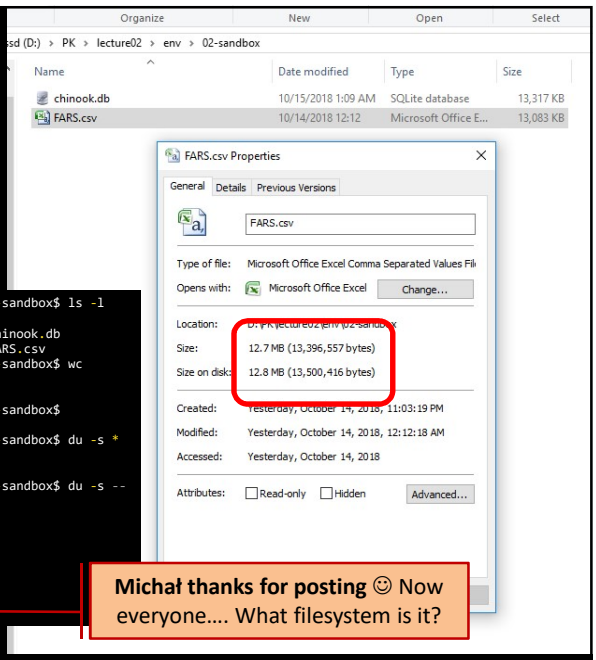
```
mean=`sed 's/,/./' scores.txt | awk '{
total += $0; count++ } END { print
total/count }'`
```


Storage considerations:
1. How big is this file?

2. How big will the database grow?

```
pawel@DESKTOP-NO0DIQE:/mnt/d/PK/lecture02/env/02-sandbox$ ls -l
total 26624
-rwxrwxrwx 1 pawel pawel 13636608 Oct 15 01:09 chinook.db
-rwxrwxrwx 1 pawel pawel 13396557 Oct 14 00:12 FARS.csv
pawel@DESKTOP-NO0DIQE:/mnt/d/PK/lecture02/env/02-sandbox$ wc
FARS.csv
  151159   151159 13396557 FARS.csv
pawel@DESKTOP-NO0DIQE:/mnt/d/PK/lecture02/env/02-sandbox$

pawel@DESKTOP-NO0DIQE:/mnt/d/PK/lecture02/env/02-sandbox$ du -s *
13440   chinook.db
13184   FARS.csv
pawel@DESKTOP-NO0DIQE:/mnt/d/PK/lecture02/env/02-sandbox$ du -s --
apparent-size *
13317   chinook.db
13083   FARS.csv
```



Michał thanks for posting ☺ Now everyone…. What filesystem is it?

Lecture 2 recap, contd.

```
pawel@DESKTOP-NO0DIQE:/mnt/d/PK/lecture02/env/02-sandbox$ ls -l
total 26624
-rwxrwxrwx 1 pawel pawel   884736 Nov 29  2015 chinook.db
-rwxrwxrwx 1 pawel pawel 13396557 Oct 14 00:12 FARS.csv
pawel@DESKTOP-NO0DIQE:/mnt/d/PK/lecture02/env/02-sandbox$ wc FARS.csv
  151159   151159 13396557 FARS.csv
pawel@DESKTOP-NO0DIQE:/mnt/d/PK/lecture02/env/02-sandbox$

sqlite> .mode csv
sqlite> .import FARS.csv fars
sqlite> select count(*) from fars;
151158

pawel@DESKTOP-NO0DIQE:/mnt/d/PK/lecture02/env/02-sand
total 26624
-rwxrwxrwx 1 pawel pawel 13636608 Oct 15 01:09 chinoo
-rwxrwxrwx 1 pawel pawel 13396557 Oct 14 00:12 FARS.c
```

Homework 2:
Explain why that much? Then make an experiment: delete the table, and check the size again. Re-import the table again, and check the size 3rd time.Conclusions?
Dominik thanks for posting the answer

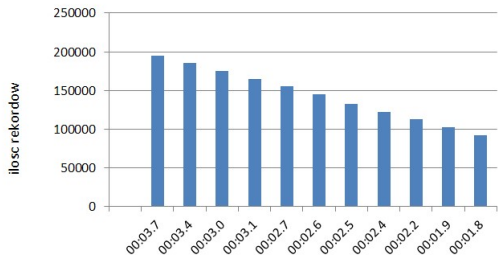
Lecture 2 recap, contd.

Homework 3

- We would like to see the relation between joined table sizes and execution time. What kind of correlation should we expect?
- Calculate this for 10 different table sizes, and graph on a bar graph.
- One possibility: jupyter %%timeit magic function + matplotlib. Another possibility: shell script
- **Post results on the linkedin group**

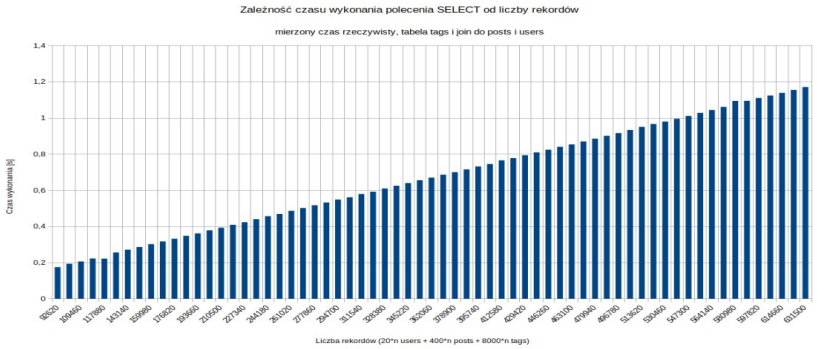
Homework 4

- Perform the same, for triple join (three tables joined)
- What do we expect?
- **Post results on the linkedin group**



From Adrian and Wojtek – thanks ☺ but…

Why are we seing this?



Lecture 2 recap, contd.

Material from Wojtek

N+1 query in Star Wars World

One table, 400 users:
- fast, single query
- clean application log

```
Started GET "/" for 127.0.0.1 at 2018-10-18 00:33:02 +0200
Processing by UsersController#index as HTML
  Rendering users/index.html.erb within layouts/application
  User Load (2.1ms)  SELECT "users".* FROM "users"
  Rendered users/index.html.erb within layouts/application (85.9ms)
Completed 200 OK in 110ms (Views: 106.5ms | ActiveRecord: 2.1ms)
```

Users

Name			
0. Dion Treutel	Show	Edit	Destroy
1. Brooks Harris	Show	Edit	Destroy
2. Laurence Vandervort I	Show	Edit	Destroy
3. Isaiah Graham	Show	Edit	Destroy
…			
396. Farrah O'Hara	Show	Edit	Destroy
397. James Gottlieb	Show	Edit	Destroy
398. Sharron Langworth	Show	Edit	Destroy
399. Meghan Boyle	Show	Edit	Destroy

Name: Dion Treutel

Han, ma bookie, keel-ee catleya ku kah. (Han, my boy, you disappoint me.)
It's against my programming to impersonate a deity.
Careful, Ren, that your personal interests not interfere with orders from Leader Snoke.
Your presence is like a violent storm in this quiet world
I have a bad feeling about...
You're tenacious, Jedi.
You're smarter than a tree, aren't you?
I assume you need something. Desperately.
You need a teacher. I can show you the ways of the Force.
We have no choice, General Calrissian! Our cruisers can't repel firepower of that magnitude!
Twice the pride, double the fall.
An object can not make you good, or evil. The temptation of power, forbidden knowledge, even the desire to do good can lead some down that path. But only you can change yourself.
I will start my operations here, and pull the rebels apart piece by piece. They'll be the architects of their own destruction.
Now, witness the power of this fully operational battle station.
I was not elected to watch my people suffer and die while you discuss this invasion in a committee!
A Jedi uses the Force for knowledge and defense, never for attack.
Anakin you're breaking my heart, you're going down a path I can't follow.
Let's just say my master will always do what needs to be done. I'm not even sure how peacetime will agree with him.
Eek bon chee ko pa na green. (I'll put a price on your head so big...)
The senate will decide your fate.
I have a bad feeling about...
The Republic doesn't exist out here. We must survive on our own.
Remember your focus determines your reality.
The Jedi are extinct. Their fire has gone out of the universe. You, my friend, are all that's left of their religion.
Chosky nowy u chusis. (I'm lucky I found you first.)

```
def index
  @users = User.all
end
```

Lecture 2 recap, contd.

Material from Wojtek

N+1 query in Star Wars World

1. Users with their posts on index page - 400+1 queries. Much longer query time, dirty app log.

2. After optimization: 2 complicated queries, twice better speed.

```
Started GET "/users" for 127.0.0.1 at 2018-10-18 00:48:12 +0200
Processing by UsersController#index as HTML
  Rendering users/index.html.erb within layouts/application
  User Load (2.0ms)  SELECT "users".* FROM "users"
  Post Load (39.6ms)  SELECT "posts".* FROM "posts" WHERE "posts"."user_id" IN (1, 2, 3, 4, 5, 6, 7, 8, 9, 10
  Rendered users/index.html.erb within layouts/application (709.0ms)
Completed 200 OK in 756ms (Views: 709.0ms | ActiveRecord: 42.9ms)
```

```
Post Load (0.2ms)  SELECT "posts".* FROM "posts" WHERE "posts"."user_id" = ?  [["user_id", 377]]
Post Load (0.2ms)  SELECT "posts".* FROM "posts" WHERE "posts"."user_id" = ?  [["user_id", 378]]
Post Load (0.2ms)  SELECT "posts".* FROM "posts" WHERE "posts"."user_id" = ?  [["user_id", 379]]
Post Load (0.2ms)  SELECT "posts".* FROM "posts" WHERE "posts"."user_id" = ?  [["user_id", 380]]
Post Load (0.2ms)  SELECT "posts".* FROM "posts" WHERE "posts"."user_id" = ?  [["user_id", 381]]
Post Load (0.2ms)  SELECT "posts".* FROM "posts" WHERE "posts"
Post Load (0.1ms)  SELECT "posts".* FROM "posts" WHERE "posts"
Post Load (0.2ms)  SELECT "posts".* FROM "posts" WHERE "posts"
Post Load (0.2ms)  SELECT "posts".* FROM "posts" WHERE "posts"
Post Load (0.2ms)  SELECT "posts".* FROM "posts" WHERE "posts"
Rendered users/index.html.erb within layouts/application (1033.1ms)
Completed 200 OK in 1060ms (Views: 968.5ms | ActiveRecord: 90.2ms)
```

- N+1 query for each item
- Lazy loading
- Joins - just one query
- Eager loading - +1 query for each resource (not item!)
- What if db server has different localization?

Lecture 2 recap, contd.

Material from Wojtek

N+1 query in Star Wars World

Users

Name				
0. Dion Treutel	Han, ma bookie, keel-ee calleya ku kah. (Han, my boy, you disappoint me.) | Shimrit Ilmwe It's against my programming to impersonate a deity. | Anakin Skywalker Careful, Ren, that your personal interests not interfere with orders from Leader Snoke. | Padme Amidala Your presence is like a violent storm in this quiet world | Anakin Skywalker I have a bad feeling about... | Greedo You're tenacious, Jedi. | Panaka	Show	Edit	Destroy
	Now, witness the power of this fully operational battle stat...			
	It's a trap! | Darth Vader Twice the pride, double the fall. | Jan...			

```
Tag Load (0.1ms)  SELECT "tags".* FROM "tags" WHERE "tags"."post_id" = ?
Tag Load (0.1ms)  SELECT "tags".* FROM "tags" WHERE "tags"."post_id" = ?
Tag Load (0.1ms)  SELECT "tags".* FROM "tags" WHERE "tags"."post_id" = ?
Rendered users/index.html.erb within layouts/application (35961.2ms)
Completed 200 OK in 35997ms (Views: 35004.8ms | ActiveRecord: 990.5ms)
```

10x faster

```
def index
  @users = User.all
end
```

```
def index
  @users = User.includes(posts: [:tags]).all
end
```

```
b within layouts/application (2045.8ms)
ews: 1990.6ms | ActiveRecord: 92.3ms)
```

Today's special

Data warehousing intro

but why – isn't OLAP dead?

- Dimensions, measures, hierarchies, drill-down are not dead
- The need for BI will also remain
- The underlying technology might change
- It'll be long time for the legacy warehouses to disappear
- So… good idea to understand what DWH is about!

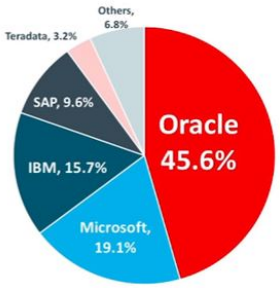
Created by Freepik

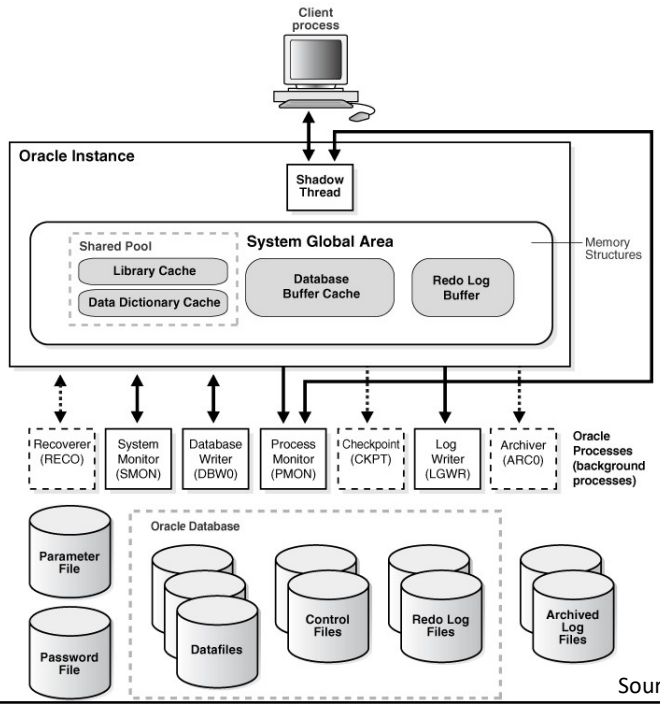
Oracle Database

Oracle Database server = database + instance (or more). The database is a set of files. The instance consists of memory segments (SGA, PGA) and background processes.

The term Oracle Database is often used to refer to both instance and database.

Gartner RDBMS market research (2016)





Source: oracle.com

First steps in SQL*Plus

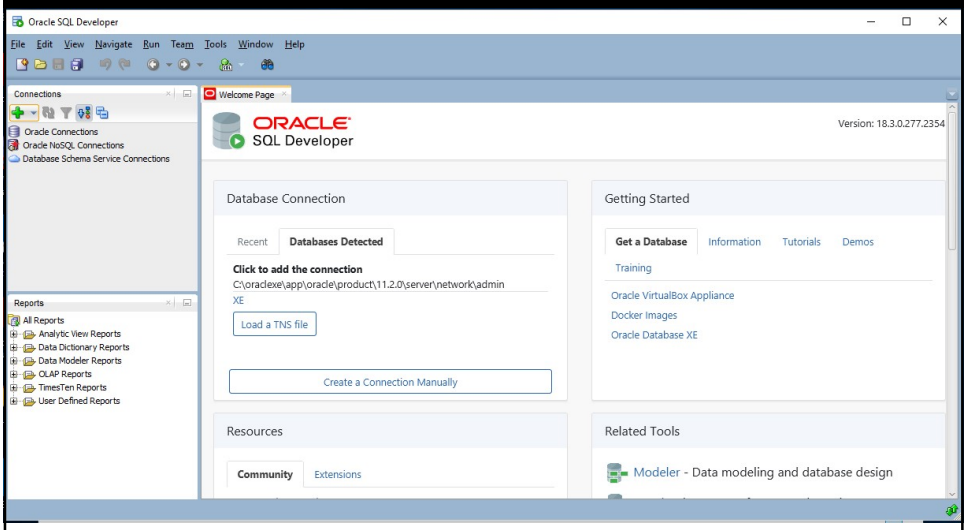
101

- sqlplus / as sysdba
- select * from dba_users;
- whoami: select user from dual;
- To capture output to file: spool output.txt
- To change one's password: ALTER USER name IDENTIFIED BY "new-password-here"
- select table_name from user_tables

Our job

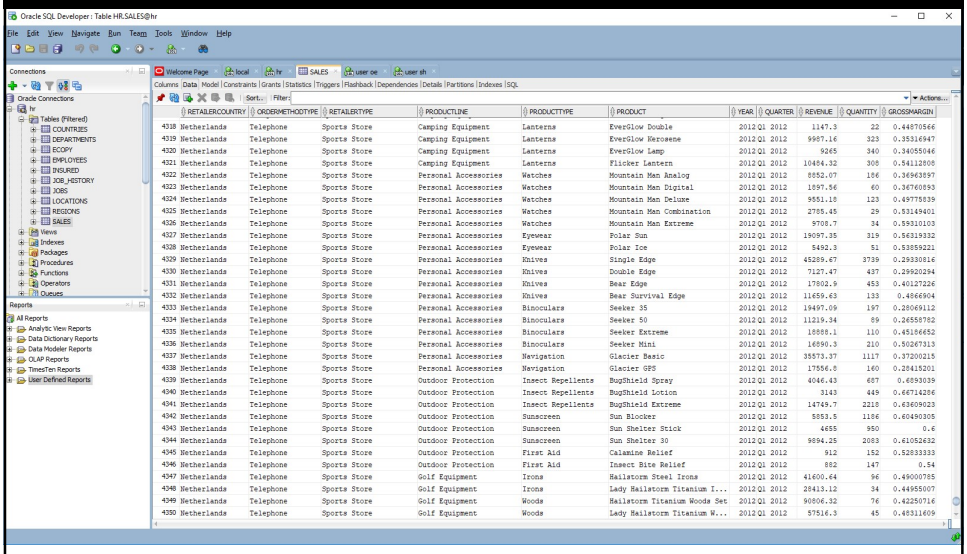
- Install the sample HR schema
- https://docs.oracle.com/database/121/COMSC/installation.htm#COMSC001
- ALTER USER hr ACCOUNT UNLOCK IDENTIFIED BY Password;
- SQL> @?/demo/schema/human_resources/hr_main.sql
- Newpassword/users / temp / yourpassword / *$ORACLE_HOME*/demo/schema/log/
- You are now HR!

We will import a csv in SQL Developer



https://community.watsonanalytics.com/wp-content/uploads/2015/08/WA_Sales_Products_2012-14.csv

The result



To compare with SQLite, Oracle is…

similar

- SQL*Plus – like sqlite3
- SQL Developer – like sqlite studio
- PL/SQL: most basic syntax is the same
- But PL/SQL is procedure-driven and has more advanced concepts – procedures, triggers, constraints and more

different

- Server
- Memory management, SGA, PGA
- On-disk management: blocks, extents, segments, tablespaces, Control files, log files
- Users, roles / privileges
- Big family of tools
- High availability
- Disaster Recovery
- Manageability
- Data movement
- Backup/recovery
- …

Oracle features: performance-wise

Indexes

- B-tree
- Function-based
- Reverse key
- Bitmap
- Compressed
- Partitioned
- Cluster
- domain
- …

More…

- Materialized views
- In-memory
- Cache
- Columnar store architecture
- Temporary tables
- Database replay
- Flashback
- Partitioning
- Tablespace-level storage integration (RAID)
- Exadata

Let's try some PL/SQL!

PL/SQL triggers

```
/* triggers are cool */

CREATE [OR REPLACE ] TRIGGER trigger_name
{BEFORE | AFTER | INSTEAD OF }
{INSERT [OR] | UPDATE [OR] | DELETE}
[OF col_name]
ON table_name
[REFERENCING OLD AS o NEW AS n]
[FOR EACH ROW]
WHEN (condition)
DECLARE
   Declaration-statements
BEGIN
   Executable-statements
EXCEPTION
   Exception-handling-statements
END;
```

My first dummy trigger

```
SQL> create table sales3 as select * from sales
where rownum <10;

Table created.


SQL> CREATE OR REPLACE TRIGGER mytrig
  2  AFTER INSERT OR UPDATE ON sales3
  3  FOR EACH ROW
  4
  5  BEGIN
  6    dbms_output.put_line("my trigger works,
yay!")
  7  END mytrig;
  8  /
```


My first dummy trigger

```
SQL> CREATE OR REPLACE TRIGGER mytrig
  2  AFTER INSERT OR UPDATE ON sales3
  3  FOR EACH ROW
  4
  5  BEGIN
  6    dbms_output.put_line("my trigger works, yay!")
  7  END mytrig;
  8  /

Warning: Trigger created with compilation errors.


SQL> show errors
Errors for TRIGGER MYTRIG:

LINE/COL ERROR
-------- -----------------------------------------------------------
----
3/1      PLS-00103: Encountered the symbol "END" when expecting one of
the
         following:
         := . ( % ;
         The symbol ";" was substituted for "END" to continue.
```

My first dummy trigger

```
SQL> CREATE OR REPLACE TRIGGER mytrig
  2  AFTER INSERT OR UPDATE ON sales3
  3  FOR EACH ROW
  4
  5  BEGIN
  6    dbms_output.put_line('my trigger works, yay!');
  7  END mytrig;
  8  /

Trigger created.


SQL> describe sales3;
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 RETAILERCOUNTRY                                    VARCHAR2(26)
 ORDERMETHODTYPE                                    VARCHAR2(26)
 RETAILERTYPE                                       VARCHAR2(128)
 PRODUCTLINE                                        NUMBER(38)
 PRODUCTTYPE                                        VARCHAR2(26)
 PRODUCT                                            NUMBER(38,2)
 YEAR                                               NUMBER(38)
 QUARTER                                            NUMBER(38,8)
 REVENUE
 QUANTITY
 GROSSMARGIN


SQL> insert into sales3(product, year, quantity, revenue) values('tent', '2018', '10',
'450');

1 row created.
```

☹

My first dummy trigger

```
SQL> set serveroutput on
SQL> insert into sales3(product, year,
quantity, revenue) values('tent', '2018',
'10', '450');
my trigger works, yay!

1 row created.
```

☺

Give yourself kudos
You will need this. More triggers later

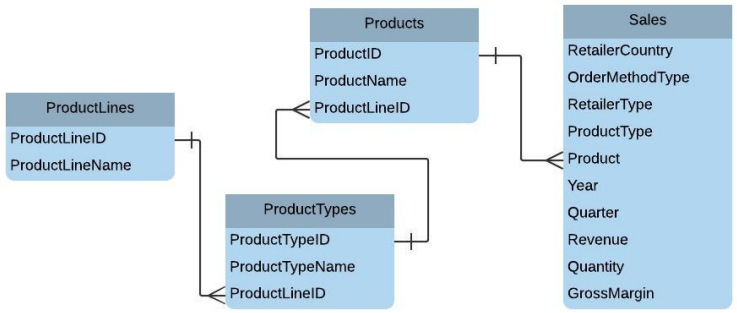
The fun is over

Theory now.

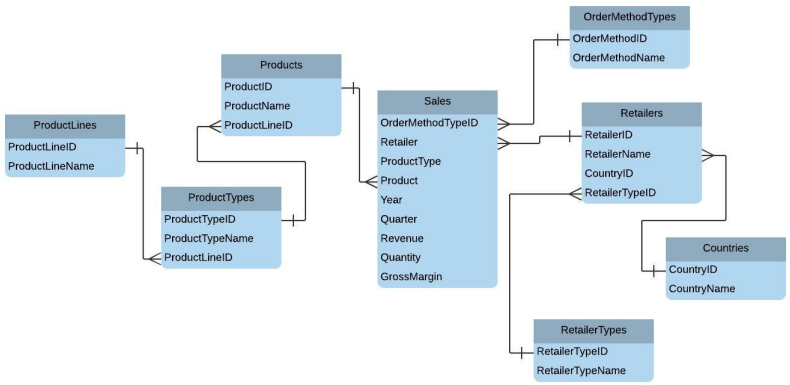
How would we normalize the sales table?

Sales
RetailerCountry
OrderMethodType
RetailerType
ProductLine
ProductType
Product
Year
Quarter
Revenue
Quantity
GrossMargin

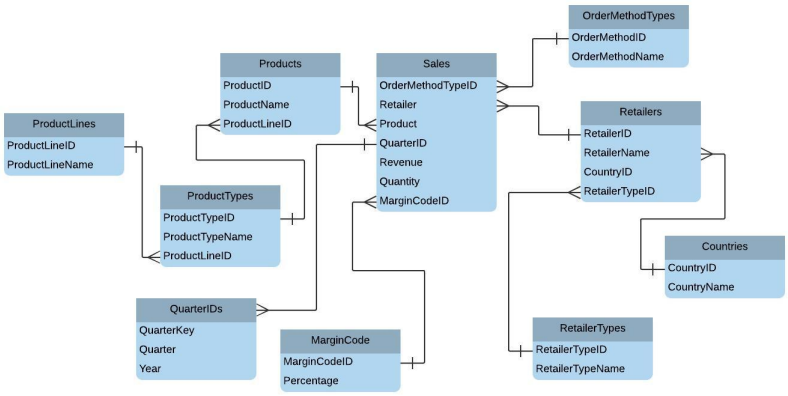
Step 1


Products
ProductID
ProductName
ProductLineID
Sales
RetailerCountry
OrderMethodType
RetailerType
ProductType
Product
Year
Quarter
Revenue
Quantity
GrossMargin
ProductLines
ProductLineID
ProductLineName
ProductTypes
ProductTypeID
ProductTypeName
ProductLineID

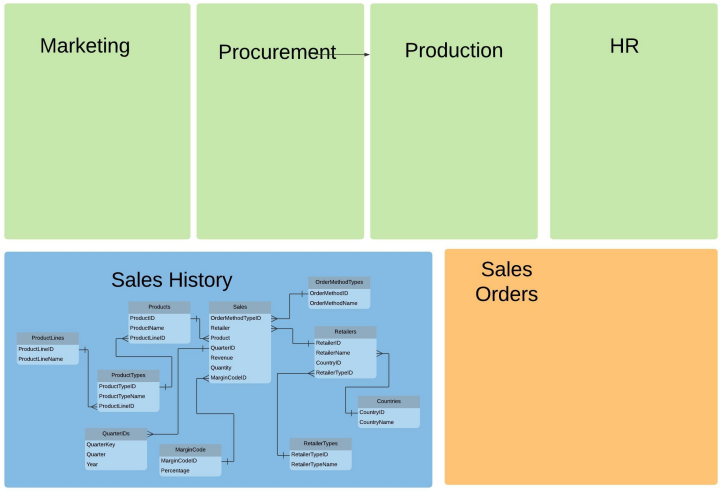
Step 2


OrderMethodTypes
OrderMethodiD
OrderMethodName
Products
ProductID
ProductName
ProductLineID
Sales
OrderMethodTypeID
Retailer
ProductType
Product
Year
Quarter
Revenue
Quantity
GrossMargin
Retailers
RetailerID
RetailerName
CountryID
RetailerTypeID
ProductLines
ProductLineID
ProductLineName
ProductTypes
ProductTypeID
ProductTypeName
ProductLineID
Countries
CountryID
CountryName
RetailerTypes
RetailerTypeID
RetailerTypeName

Step 3



1000s tables in typical production OLTP



Why?

Why 3NF?

Some thoughts

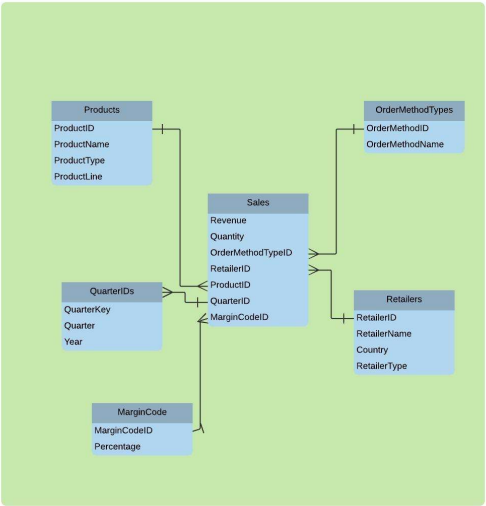
Remember indexing is good for SELECT, but bad for UPDATE?

Maybe some other things are good for UPDATE / DELETE / INSERT, but bad for SELECT?

For a moment, imagine that our database only serves SELECTs

How would we redesign the database?

Rethink


Products
ProductID
ProductName
ProductType
ProductLine
OrderMethodTypes
OrderMethodID
OrderMethodName
Sales
Revenue
Quantity
OrderMethodTypeID
RetailerID
ProductID
QuarterID
MarginCodeID
QuarterIDs
QuarterKey
Quarter
Year
Retailers
RetailerID
RetailerName
Country
RetailerType
MarginCode
MarginCodeID
Percentage


Snowflakes collection
designed by freepik.com
Created by Freepik

Snowflake schema

Summary

- Single joints in all directions
- Simple SELECTs
- Facts table: business facts (measures) + foreign keys
- Dimensions tables
- Measures
- Fact attributes

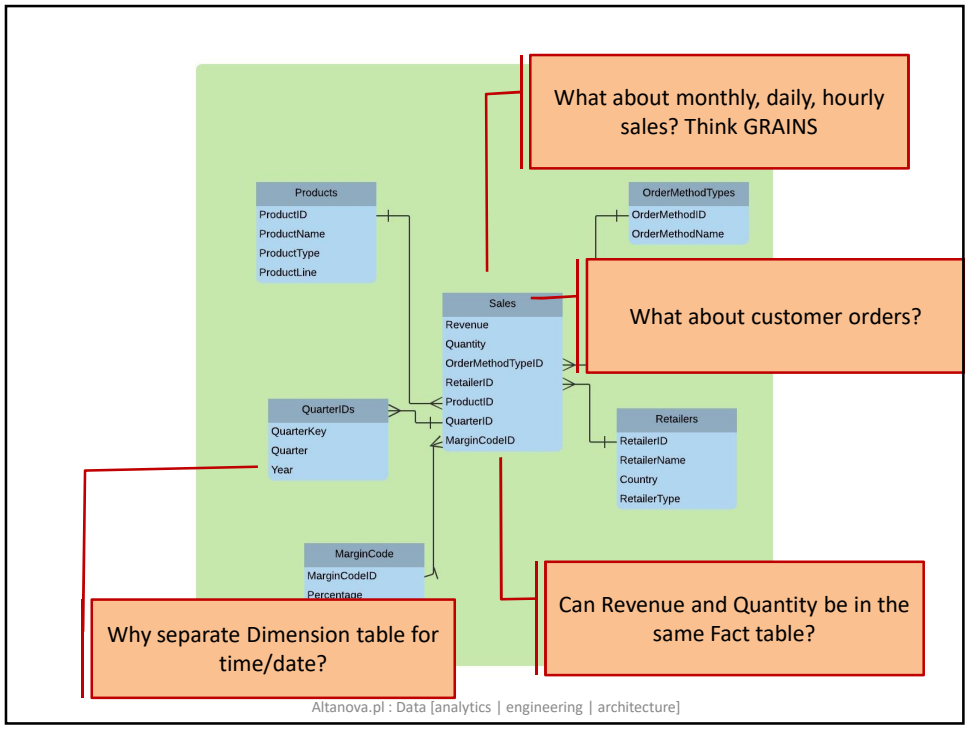
Decisions to make

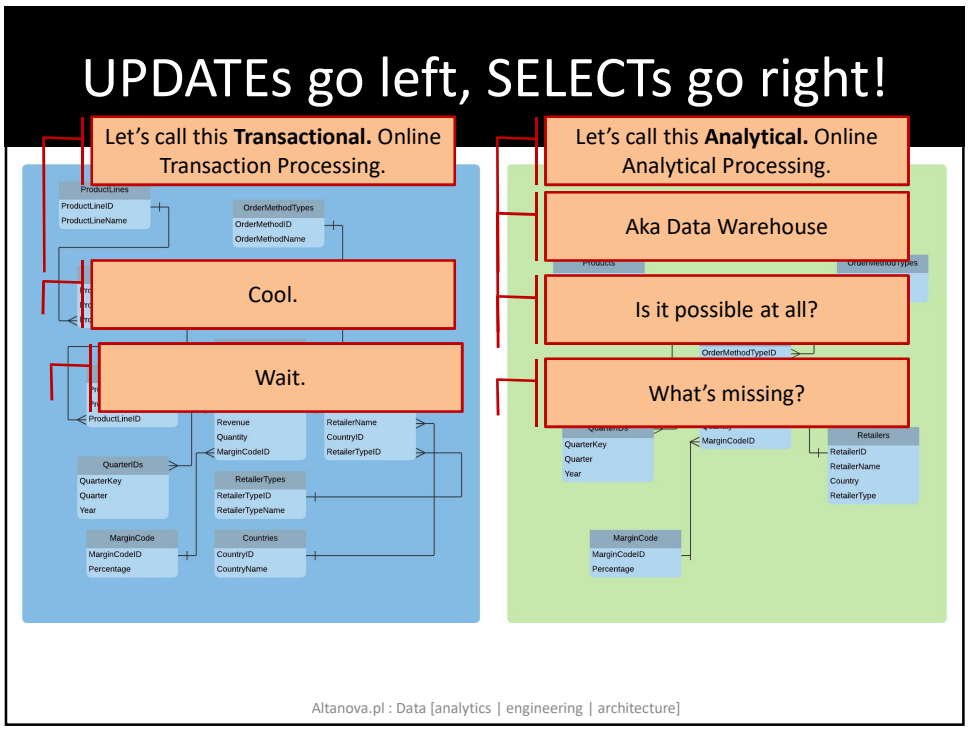
- What are the facts
- What is the grain
- Star or snowflake or constellation
- ROLAP / MOLAP?

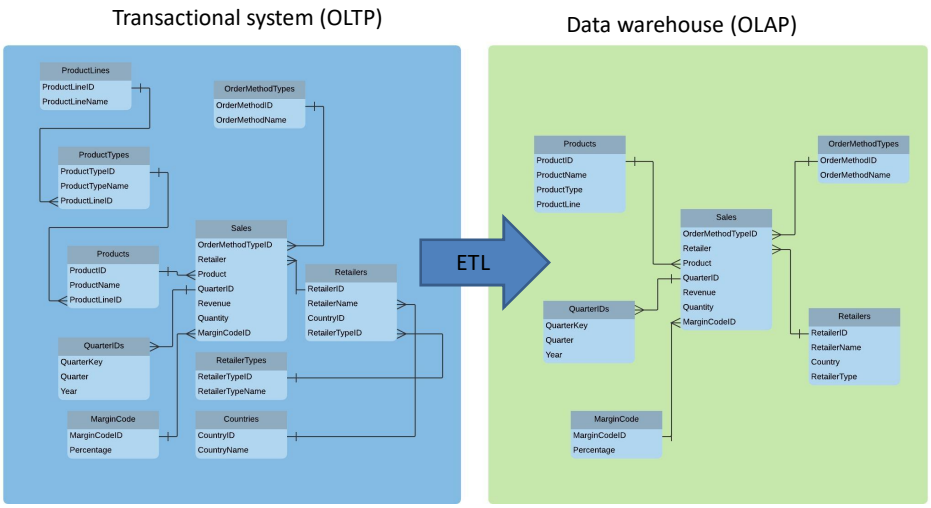
Over to you

	3NF	Snowflake
The good		
The bad		
Optimal for what processing?		
Terrible for…?		
Is it readable for end-users?		
Queries: batch or real-time?		
Business-critical?		
Database size		
Database design		
High-availability requirements		
Disaster Recovery requirements		

Homework 1 ☺
No, I did not forget


What about monthly, daily, hourly sales? Think GRAINS
Products
ProductID
ProductName
ProductType
ProductLine
OrderMethodTypes
OrderMethodID
OrderMethodName
Sales
Revenue
Quantity
OrderMethodTypeID
RetailerID
ProductID
QuarterID
MarginCodeID
What about customer orders?
QuarterIDs
QuarterKey
Quarter
Year
Retailers
RetailerID
RetailerName
Country
RetailerType
MarginCode
MarginCodeID
Percentage
Can Revenue and Quantity be in the same Fact table?
Why separate Dimension table for time/date?
Altanova.pl : Data [analytics | engineering | architecture]


UPDATEs go left, SELECTs go right!
Let's call this **Transactional.** Online Transaction Processing.
Let's call this **Analytical.** Online Analytical Processing.
Aka Data Warehouse
Cool.
Is it possible at all?
Wait.
What's missing?
ProductLines
ProductLineID
ProductLineName
OrderMethodTypes
OrderMethodID
OrderMethodName
Revenue
Quantity
MarginCodeID
RetailerName
CountryID
RetailerTypeID
QuarterIDs
QuarterKey
Quarter
Year
RetailerTypes
RetailerTypeID
RetailerTypeName
MarginCode
MarginCodeID
Percentage
Countries
CountryID
CountryName
Retailers
RetailerID
RetailerName
Country
RetailerType
Altanova.pl : Data [analytics | engineering | architecture]



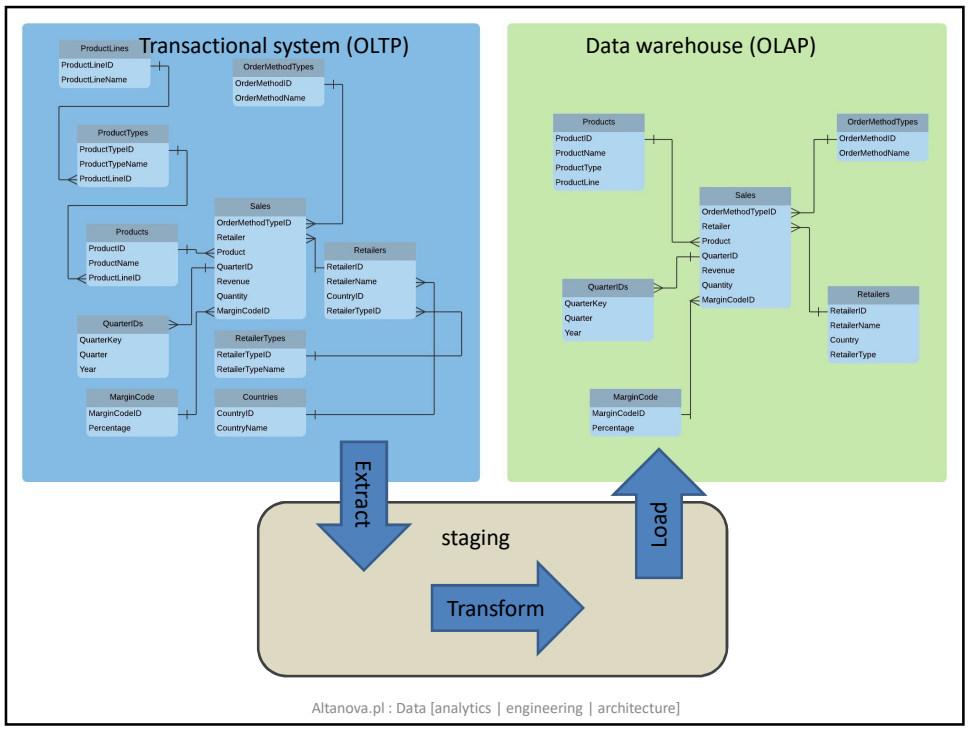
Transporting data from transactional do DWH

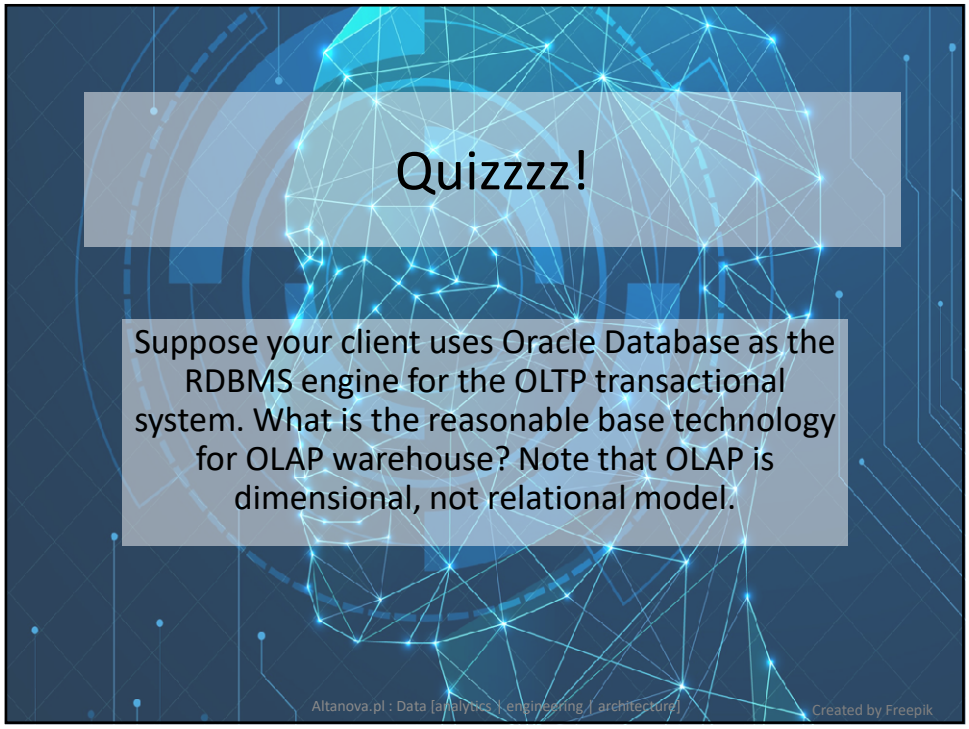
Significant Processing needed on the fly

- Cleansing
- Stripping unncesessary columns
- Formatting
- Removing dummy/test records
- Denormalizing

Again… why?

- Aha. Performance.
- Really? ☺
- How about… quick-and-dirty extract?


Transactional system (OLTP)
ProductLines
ProductLineID
ProductLineName
OrderMethodTypes
OrderMethodID
OrderMethodName
ProductTypes
ProductTypeID
ProductTypeName
ProductLineID
Sales
OrderMethodTypeID
Retailer
Product
QuarterID
Revenue
Quantity
MarginCodeID
Products
ProductID
ProductName
ProductLineID
Retailers
RetailerID
RetailerName
CountryID
RetailerTypeID
QuarterIDs
QuarterKey
Quarter
Year
RetailerTypes
RetailerTypeID
RetailerTypeName
MarginCode
MarginCodeID
Percentage
Countries
CountryID
CountryName
Data warehouse (OLAP)
Products
ProductID
ProductName
ProductType
ProductLine
OrderMethodTypes
OrderMethodID
OrderMethodName
Sales
OrderMethodTypeID
Retailer
Product
QuarterID
Revenue
Quantity
MarginCodeID
QuarterIDs
QuarterKey
Quarter
Year
Retailers
RetailerID
RetailerName
Country
RetailerType
MarginCode
MarginCodeID
Percentage
Extract
staging
Load
Transform
Altanova.pl : Data [analytics | engineering | architecture]


Quizzzz!
Suppose your client uses Oracle Database as the RDBMS engine for the OLTP transactional system. What is the reasonable base technology for OLAP warehouse? Note that OLAP is dimensional, not relational model.
Altanova.pl : Data [analytics | engineering | architecture]
Created by Freepik

Over to you

	OLTP	OLAP
Business purpose		
Used by what corporate roles?		
Typical selects. Are they frequent? Are they intensive?		
Updates, inserts – how often they happen, how long they take?		
Are queries mostly complex?		
Queries are batch-type or real-time?		
Business-critical?		
Database size		
Database design		
High-availability requirements		
Disaster Recovery requirements		

Homework 2☺

Homework 3

Easy [1 hour]

- Install Tableau
- Connect to Oracle
- View SALES
- Explore and have fun

Homework 4 & 5

3. Medium [2 hours]

- Build a simple automatic ETL
- From SALES table copy new records directly to a new table SALES2. Both OLTP and OLAP in one database. No separate staging
- We have no operations in Netherlands. Netherlands records is dummy data for testing. Netherlands data

Hint: Triggers!
This might be useful:
WHERE tableB.columnX = :NEW.columnX;

4. Medium hard [4 hours]

- Install Oracle on 2 laptops
- Connect 2 Oracle Databases between them. One is OLTP, the other is OLAP
- Implement the same ETL between 2 machines!

Hint1: Disable firewalls… Hint 2:
INSERT INTO test_tbl@remote_db

Homework 6 - for die-hards. work in team

Hard [1 day]

- Create a snowflake!
- Use the sales data or any other data with 10,000 records and at least 3 dimensions. More dimensions better.
- What is your fact?
- Grain?
- Measures?
- Dimensions?
- Hint: import csv to 1 table, then use SQL Developer or SQL*Plus to transfer data to newly created dimension tables

Data Science, Data Engineering, Data Management... Data Art?

Big Data in 30 hours

Thank you!
Next lecture: easy. OLAP + Tableau maybe?

Cracow University of Technology (Politechnika Krakowska)
Faculty of Physics, Mathematics and Computer Science
Original course title:
Methods and tools for big data analysis, code WFMiI I oIIS D2 17/18
Graduate Master's degree studies, 2nd year
2018/2019
Venue: Room F017, ul. Podchorążych 1, Kraków (budynek wydz. Fiz Mat i Inf)
Time: Every Monday, 12:345, starting Mon Oct 8th 2018 through end Jan 2019

Altanova.pl : Data [analytics | engineering | architecture]