Data Warehouse Project

Advanced Database Management Technologies 2017-18

4.jan 2018

The Independent Media Corporation

By:

Farzad Mushfiq

Frida Strand Kristoffersen

Table of contents

Advanced Database Management Technologies 2017-18

The Independent Media Corporation is one of Scandinavia's Media conglomerates. The company is a mother company of different printed newspaper companies (e.g. Nordlys, Aftenposten, etc.). They also offer a common online newspaper where every printed newspaper can publish their news, so it does not belong to a particular website. The papers, both the printed newspapers, but also the online one, are sold all over Scandinavia.

Instead of having a database containing all on the data about each subscriber and each sale, a data ware (DW) is needed to be able to analyse and make better business decisions. With a data warehouse the company can calculate the profits to run the business smoothly to achieve its long term vision.

The data warehouse would help to analyse business needs, trends, answer business management questions as the corporation wants to increase the sales amount both regarding the newspapers and the online subscriptions. By implementing a data warehouse, The Independent Media Corporation can easily get a more complete picture of their business in order to maximize their profits.

A data warehouse makes the analyse process of the business easy. It will be an important equipment now that the organization is going through a tough time because the trend all over the world shows that the sales numbers regarding printed newspapers are declining dramatically[1]. Through a data warehouse these important analyzes can be done within seconds and maybe the DW could be exactly what makes especially regarding the printed newspapers make the right decisions at the right time in order to survive.

[1] https://www.theatlantic.com/business/archive/2016/11/the-print-apocalypse-and-how-to-survive-it/506429/

Business Process Modelling

Due to that the Independent Media Corporation is a big corporation company and therefore has a lot of business processes, we have decided to analyze their online subscription offer and their printed newspaper sales. We have collected the information which is relevant to be able to support the most important decisions within each business process. Below we have described information about the specific processes.

Online Subscription

Every online subscription includes information about the subscriber's information (e.g. name, birthday, address), subscription bundle (student or normal), subscription price, subscription period etc.

Printed newspaper sales

Every newspaper house is connected to different stores which sells their papers to the actual customer. When we built this data warehouse, we defined a "printed newspaper sales" as something that is happening at one specific day, with one specific printed newspaper house (daughter company), in potentially more than one store, with specific production costs, shipping costs and a specific price per paper where the amount of papers sold in total is specified.

With this data warehouse, we are able to find out e.g. the sales numbers for a specific printed newspaper house on a daily, weekly and monthly basis. Regarding the online subscription, we are able to do analyzed based on e.g. subscription bundle and information about the subscriber.

Query examples

To be able to analyze the business of The Independent Media Corporation, we need to get an overview of which data we need to extract from our data warehouse. Therefore we have listed some queries to make it easier to visualize what type of service the data warehouse can offer.

Online subscription

- How many have been a subscriber for >5 years?
- How many subscribers is registered for the online newspaper?
- What year did recruited most new subscribers?
- In what countries do the online subscribers live (compared to the paper subscribers)?
- What is the most popular subscription bundle?
- What is the average age of the online subscribers (compared to a paper subscriber)?
- At what weekday gets normally most new subscribers?
- What is the total revenue for online subscriptions last year?

Printed newspaper sales

- How many printed papers were sold 19.08.2017?
- How much was the shipping costs for 2017?
- How many papers from VG were printed in a specific city (in a printing house located in this city)?
- How much was the total revenue for June 2017?
- How many different stores sell printed newspapers?
- What quarter had the highest revenue in 2017?
- What region has most stores?
- What was the profit for printed paper sales in 2017? (profit = total revenue - total costs)

Conceptual design

To be able to build a data warehouse, we need to understand what is the most relevant information to the two processes. Each dimension below contains attributes and meaningful information which makes us able to look at the facts from different perspectives.

Facts, dimensions and measures

Online subscription:

- **Time dimension**
  - id: The key for the time dimension.
  - hours: Represents the hour for when the subscription was registered.
  - minutes: Represents the minutes for when the subscription was registered.
- **Subscriber dimension**
  - id: Represent the unique id number of subscriber.
  - firstname: Represents subscriber's first name.
  - surname: Represents subscriber's last name.
  - birthday: Represents the date of when the subscriber was born.
  - addressid: Represents the key to the address dimension where you can find the subscriber's address.
- **Subscription type**
  - id: The key for the subscription type dimension
  - bundle: Decribes type of subscription: Student (discount) or normal
- **Measures**
  - discount: It is possible to get a family discount for online subscription if you are a student. The size of the discount depends whether it is a campaign or not, which is not further taken into account regarding this data warehouse.
  - price: The original price that the subscriber paid for the subscription.
  - lengthofsubscription: Represents the length of the subscribers' subscription in months

Printed newspaper sales

- **Shipping dimension**
  - id: The key for the shipping dimension
  - costs : represents the cost of the shipping of newspapers to a specific location. The costs depends on the shipping method, and does also include packaging.
  - Method: represents the methods of transportation of newspapers via plane, boat, train, car.

- **Printing dimension**
  - id: The key for the printing dimension
  - printinghouse: represents which printing (child-)company sold it's newspapers.
  - costs: represents the cost of the newspapers printed in printing house.
  - addressid: The key to the address of the printing house in the address dimension which is also used by store to save their addresses as well.

- **Store dimension**

- id: The key for the store dimension
- name: Name of the store
- addressid: The key to the address of the store in the address dimension which also printing house uses to save their addresses as well.

- **Production dimension**
  - id: The key for the production dimension
  - costs: Represents the overall costs regarding producing the papers. This includes, marketing costs, administration costs, etc.
- **Measures:**
  - quantity: Represents the number of papers the specific printing house sold on a specific day when using a specific shipping method.
  - priceperpaper: Represents the price for one paper only. In order to calculate the total revenue for this printing house for this day, is by multiplying priceperpaper with quantity, for each shipping method for this day, and then add them all together.

Dimensions in common

Both of the business processes use the date dimension in their fact table.

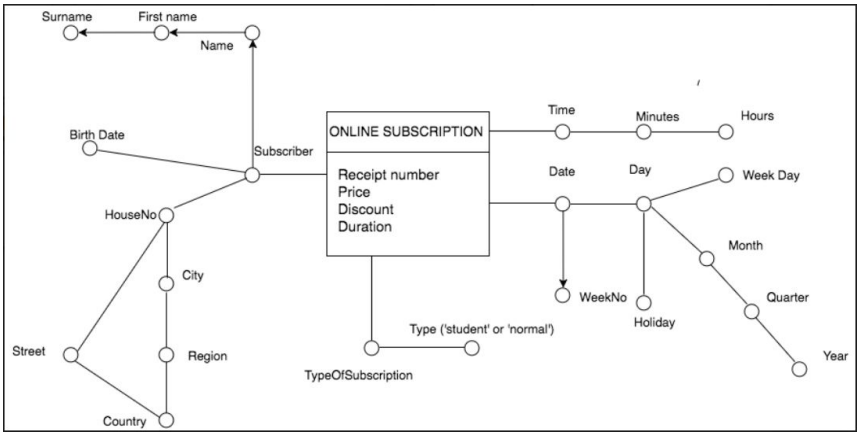
- **Date dimension**
  - id: The key for the date dimension.
  - date: The whole date for the subscription registration (YYYY-MM-DD).
  - day: The day in the month for the subscription registration/papers sold.
  - month: The number of month in the year of the registration/papers sold.
  - quarter: Represents which quarter of the year the registration was registered/papers sold.
  - year: The year the registration was registered/papers sold.
  - weekday : The day of the week as a string for the subscription registration/papers sold (e.g. ‘monday’).
  - holiday: Boolean value e.g. if the registration/paper was made/sold on a Sunday.
  - weeknumber: Represents the number of week in the year the subscription was registered/papers sold.
- **Address dimension**
  - This dimension is shared between the subscriber, store and printing dimension.
  - id: The key for the address dimension.
  - houseno: Represents the number on the house.
  - street: Represents the street name.
  - region: Represents the region.
  - city: Represents the city.

DFM diagrams

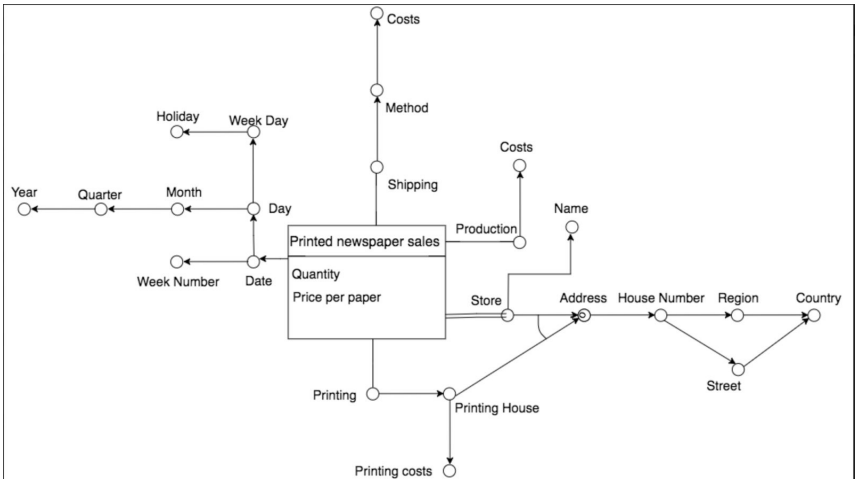
In this section, we will show you the DFM diagrams for online subscription and the printed newspapers to give a better visualisation and to provide an effective support to our next phase of conceptual design.

- Since we are interested in the address of both the printing houses and stores, we make the address a shared hierarchy since the address hierarchy is by two dimensions.
- We make use of multiple arcs for the newspaper sales' stores because one sale is defined to be one delivery from one "package" of newspapers from the printing house. And a printing house will often distribute through different stores.
- Region: Represents the region area where the subscriber is located. Regions will normally include more than one city.
- Street: A street can connect different cities and in this data warehouse we assume that a street cannot cross countries. Thus, street and region creates a convergence since both of them are connected to country.

Online Subscription



Printed newspaper sales



Logical Design

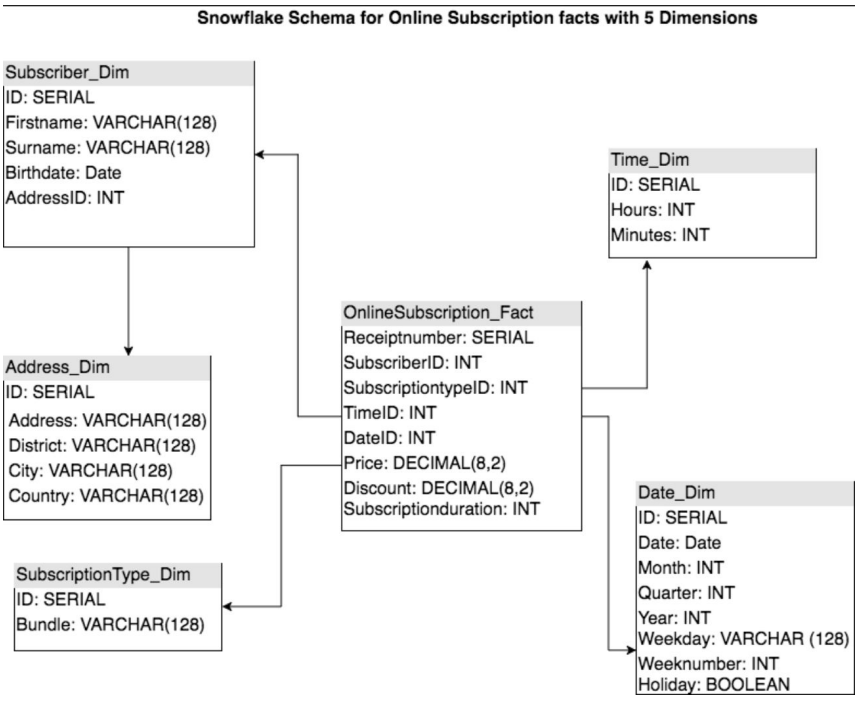
In this section, we are going to show how we implemented our DW. For the logical design of our database, we have chosen the snowflake schema because the address dimension is used three times (two in printed newspaper sales and one time in online subscriptions).

Data Warehouse Bus Matrix

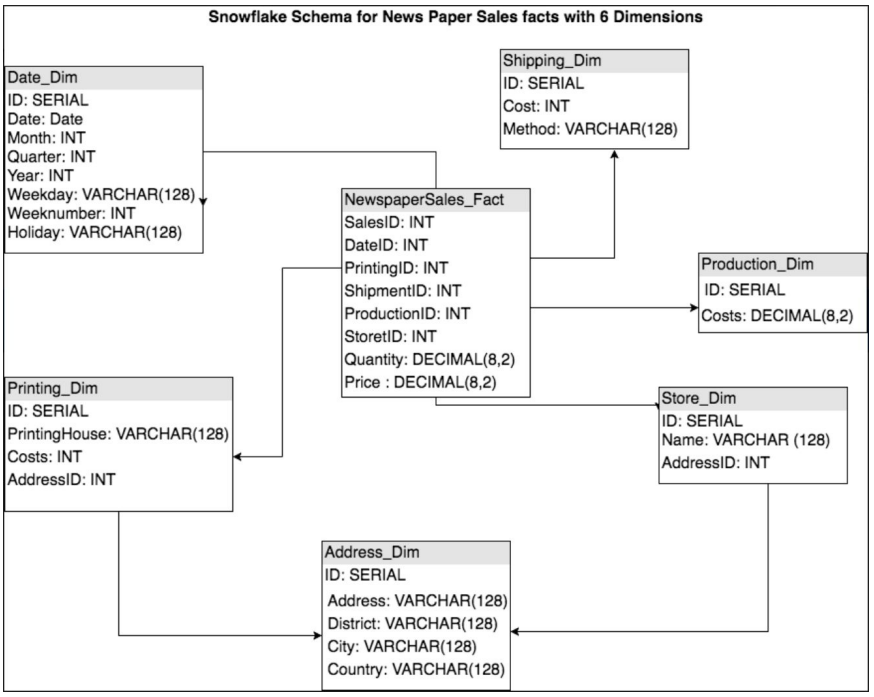
Before we implemented the snowflake schema we wanted to get an overview of the different dimensions by creating a DW bus matrix. With this matrix, our intention is to give an overview of the different dimensions and which one the two business processes share.

Business process	Date	Address	Time	Subscriber	Printing	Store	Shipping	Production
Online subscriptions	x	x	x	x				
Printed newspaper sales	x	x			x	x	x	x

Online subscription:



Newspaper sales:



Data cleaning

When data is transferred into the data warehouse, it might occur some errors regarding inconsistency. Under the data cleaning part, we have had our main focus on:

City and countries can be different from different languages

- We have chosen to have every place names in English. For every input, we have used a translator to check what language the name corresponds to, and if it is not the English version of it, we translate it.

Customer names, printing house names and store names can be misspelled

- It is possible to check this by looking at their address. If two stores (or customers or printing houses) has the same address, but not the same name, we will check if it is the same store by using mapping where we work on the similarities. This can also be done e.g. if an address is assumed to be inconsistent and a name occurs twice.

Currency

- Everything is translated into Euro.

The tables

Address table

id [PK] integer	houseno integer	city character varying (128)	region character varying (128)	country character varying (128)	street character varying (128)
1	99	trondheim	elgeseter	sweden	kongens gate
2	5	stavanger	storo	norway	storogaten
3	5	lillehammer	kringsja	norway	jernvinna
4	1	tromso	prestvannet	norway	prestvannvegen
5	8	helsinki	helles	finland	aqqavegen
6	91	tromso	sentrum	norway	torggata
7	28	bergen	sentrum	norway	skippergata
9	45	stockholm	sentrum	sweden	jomfrugata
12	5	copenhagen	copen	denmark	kartoffelgate
16	10	tromso	sentrum	norway	stortorget
21	110	Tromso	Workinn	Norway	Workinnmarka
22	6	Sandefjord	Hunsrod	Norway	Hunsrodgrenda
23	10	Tromso	Hamna	Norway	Fiskeribølgen
24	6	Tromso	Troms	Norway	Myrengvegen
27	109	Tromso	Hapet	Norway	Prestvannvegen

Date table

id [PK] integer	date date	day integer	month integer	year integer	holiday boolean	weeknumber integer	quarter integer	weekday character varying
1	2016-11-01	1	11	2016	false	44	4	tuesday
2	2016-10-01	1	10	2016	false	39	4	saturday
3	2016-06-15	15	6	2016	false	24	2	wednesday
5	2017-06-08	8	6	2017	false	23	2	thursday
6	2016-01-08	8	1	2016	false	1	1	friday
7	2016-12-08	8	12	2016	false	49	4	thursday
8	2016-10-10	10	10	2016	false	41	4	monday
9	2017-10-04	4	10	2017	false	40	4	wednesday
14	2017-10-18	18	10	2017	false	42	4	wednesday
18	2017-10-11	11	10	2017	false	41	4	wednesday
20	2017-10-03	3	10	2017	false	40	4	tuesday
22	2017-10-21	21	10	2017	false	42	4	saturday
23	2017-10-07	7	10	2017	false	40	4	saturday
24	2017-10-14	14	10	2017	false	41	4	saturday
26	2017-12-25	25	12	2017	false	52	4	monday
27	2017-12-26	26	12	2017	false	52	4	tuesday
28	2017-12-27	27	12	2017	false	52	4	wednesday
30	2017-11-24	24	11	2017	false	47	4	friday
31	2017-11-26	26	11	2017	true	47	4	sunday
33	2017-11-19	19	11	2017	true	46	4	sunday
34	2017-11-14	14	11	2017	false	46	4	tuesday

Online subscription fact table

id [PK] integer	time integer	subscriber integer	typeofsubscription integer	date integer	discount integer	duration integer	price integer
1	7	7	7	7	200	6	2500
2	8	8	7	8	500	6	2500
3	6	6	6	6	0	6	2500
4	1	1	7	1	500	12	2500
5	3	2	7	2	500	12	2500
6	3	3	6	3	0	12	2500
7	5	4	7	5	300	6	2500
8	5	5	6	5	0	6	2500

Store table

id [PK] integer	name character varying (128)	addressid integer
1	delideluca	6
9	narvesen	4
12	spar	3
14	europris	2
21	rema	1
25	kmarket	5

Newspaper sales fact table

id integer	date integer	printing integer	productioncost integer	store integer	shipping integer	quantity integer	priceperpaper integer
73962	23	15	22	14	19	2000	3
73963	24	15	16	14	16	2500	3
73964	27	15	22	14	19	2000	3
73965	28	15	22	14	19	2000	3
73966	22	17	18	14	18	22000	3
73967	26	17	18	14	18	22000	3
73968	26	17	18	14	18	22000	3
73969	28	23	22	21	19	2000	3
73970	30	23	22	21	22	2000	3
73971	31	23	22	21	25	2000	3
73972	34	27	22	25	27	1000	3
73973	31	25	22	25	25	2000	3
73974	33	25	22	25	25	2000	3
73975	34	26	26	25	26	5000	3
73976	14	7	6	9	6	1500	3
73977	18	7	6	9	6	200	3
73978	14	3	3	1	3	4000	3
73979	14	4	5	1	4	4500	3
73980	14	4	5	9	4	4500	3
73981	9	1	1	1	1	5000	3
73982	18	2	2	1	2	7000	3
73983	18	9	9	9	6	1500	3
73984	18	9	9	9	9	1500	3
73985	18	9	9	12	10	1500	3
73986	18	11	22	12	19	2200	3
73987	20	11	22	12	19	2200	3
73988	22	11	22	12	19	2200	3

Time table

timeid [PK] integer	hours integer	minutes integer
1	12	4
3	20	35
5	20	48
6	8	48
7	1	21
8	5	21

Subscription type table

bundle character varying (128)	subscriptionid [PK] integer
normal	6
student	7

Production table

cost double precision	id [PK] integer
1000	1
1200	2
800	3
900	5
500	6
400	9
350	16
3000	18
300	22
600	26

Printing table

printinghouse character varying (128)	costs integer	id [PK] integer	addressid integer
nordlys	300	1	16
nordlys	400	2	16
nordlys	200	3	16
nordlys	280	4	16
nordlys	180	7	16
nordlys	480	9	16
nordlys	600	11	16
aftenposten	800	15	7
aftenposten	8000	17	7
bergenstidene	800	23	9
frittland	800	25	12
frittland	900	26	12
frittland	300	27	12

Shipping table

costs integer	method character varying	id [PK] integer
100	car	1
300	car	2
200	car	3
220	car	4
110	car	6
50	train	9
50	plane	10
150	boat	16
1000	boat	18
100	boat	19
100	train	22
100	plane	25
200	plane	26
10	car	27

Subscriber table

firstname character varying	surname character varying (128)	birthday date	subscriberid [PK] integer	addressid integer
Frida Strand	Kristofferen	1993-06-15	1	24
Marie Voll	Eek	1993-07-01	2	21
Jonas Modin	Rismyhr	1988-09-06	3	22
Erica Josefine	Olsen	1994-05-01	4	23
Kjersti	Strand	1960-04-15	5	24
Trond	Strand	1955-08-19	6	24
Brage Strand	Kristoffersen	1991-12-13	7	24
Ane Skrove	Nossum	1991-12-13	8	27

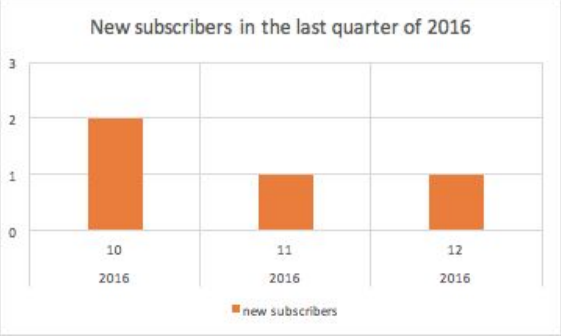
QUERIES

QUERY 1: How many new subscribers the online newspaper achieved in the last quarter of 2016?

- **The output is showing explicitly the number for each month and can conclude that in October the online newspaper gained most new subscribers.**

```
select  date.year,
        date.month,
        count(*) AS "new subscribers"
FROM    date, fact_onlinesub
WHERE   fact_onlinesub.date = date.id and
        date.year = 2016 and
        date.month IN (10,11,12)
GROUP BY date.year, date.month
ORDER BY date.year, date.month
```

year integer	month integer	new subscribers bigint
2016	10	2
2016	11	1
2016	12	1



QUERY 2: What is the total costs, revenue and then the final surplus per printinghouse in 2017?

- **We order the output by descending surplus. Aftenposten is clearly the printing house that had the highest surplus in 2017.**

```
select  printing.printinghouse,
        SUM(fact_newspapersales.quantity*fact_newspapersales.priceperpaper) AS "Revenue",
        SUM(shipping.costs) AS "Shipping costs",
        SUM(printing.costs) AS "Printing costs",
        SUM(production.costs) AS "Production costs",
        SUM(shipping.costs+printing.costs+production.costs) AS "Total_costs",
        SUM(fact_newspapersales.quantity*fact_newspapersales.priceperpaper
           -shipping.costs
           -printing.costs
           -production.costs) AS "Surplus in 2017"
FROM    printing, fact_newspapersales, date, production, shipping
WHERE   fact_newspapersales.printing = printing.id and
        fact_newspapersales.production = production.id and
        fact_newspapersales.shipping = shipping.id and
        fact_newspapersales.date=date.id and
        date.year=2017
GROUP BY printing.printinghouse
ORDER BY "Surplus in 2017" DESC
```

printinghouse character varying (128)	Revenue bigint	Shipping costs bigint	Printing costs bigint	Production costs double precision	Total_costs double precision	Surplus in 2017 double precision
aftenposten	223500	3450	27200	10250	40900	182600
nordlys	113400	1770	5060	7900	14730	98670
frittland	30000	410	2800	1500	4710	25290
bergenstidene	18000	300	2400	900	3600	14400

QUERY 3: Get an overview over the total shipping costs and total quantity shipped per day per shipping method in October 2017 where the store is located in Norway, and what was the total shipping costs and total quantity shipped per printing and store location in October 2017 where the store is located in Norway.

- **We use GROUPING SETS to be able to get two different "overviews" within the same table.**

```
SELECT  date.day AS "Day in October 2017",
        shipping.method AS "Shipping method",
        address.city AS "Printing House Location City",
        address2.city AS "Store Location City",
        SUM(fact_newspapersales.quantity) AS "Total quantity shipped",
        sum(shipping.costs) as "Shipping costs"
FROM    fact_newspapersales, date, shipping, address, store, printing, address AS address2
WHERE   fact_newspapersales.shipping=shipping.id and
        fact_newspapersales.date=date.id and
        fact_newspapersales.store=store.id and
        fact_newspapersales.printing=printing.id and
        store.addressid=address.id and
        printing.addressid=address2.id and
        date.year=2017 and date.month=10 and
        address.country = 'norway'
GROUP BY shipping.method,
        GROUPING SETS ((date.day), (address.city, address2.city))
ORDER BY date.day
```

Day in October 2017 integer	Shipping method character varying	Printing House Location City character varying (128)	Store Location City character varying (128)	Total quantity shipped bigint	Shipping costs bigint
3	boat	[null]	[null]	2200	100
4	car	[null]	[null]	5000	100
7	boat	[null]	[null]	2000	100
11	boat	[null]	[null]	2200	100
11	train	[null]	[null]	1500	50
11	plane	[null]	[null]	1500	50
11	car	[null]	[null]	8700	520
14	boat	[null]	[null]	2500	150
18	car	[null]	[null]	14500	750
21	boat	[null]	[null]	24200	1100
[null]	boat	stavanger	bergen	26500	1250
[null]	plane	lillehammer	tromso	1500	50
[null]	boat	lillehammer	tromso	6600	300
[null]	train	tromso	tromso	1500	50
[null]	car	tromso	tromso	28200	1370

QUERY 4: Ranking the monthly surplus per printing house in 2017?

- **Here we use RANK OVER () to rank the different printing houses by their monthly surplus. Each printing house will therefore appear several times since the same printing house will appear one time for each time it sold printed papers in a month.**

```
SELECT  date.month AS "Month in 2017",
        printing.printinghouse AS "Printing house",
        SUM(fact_newspapersales.quantity*fact_newspapersales.priceperpaper) AS "Revenue",
        SUM(shipping.costs+production.costs+printing.costs) AS "Total costs",
        SUM(fact_newspapersales.quantity*fact_newspapersales.priceperpaper
                                -(shipping.costs+production.costs+printing.costs)) AS "Surplus",
        RANK() OVER (ORDER BY SUM(fact_newspapersales.quantity*fact_newspapersales.priceperpaper
                                -(shipping.costs+production.costs+printing.costs)) DESC)
FROM    fact_newspapersales, date, shipping, printing, production
WHERE   fact_newspapersales.shipping=shipping.id and
        fact_newspapersales.date=date.id and
        fact_newspapersales.production=production.id and
        fact_newspapersales.printing=printing.id and
        date.year=2017
GROUP BY printing.printinghouse, date.month
```

Month in 2017 integer	Printing house character varying (128)	Revenue bigint	Total costs double precision	Surplus double precision	rank bigint
12	aftenposten	144000	26400	117600	1
10	nordlys	113400	14730	98670	2
10	aftenposten	79500	14500	65000	3
11	frittland	30000	4710	25290	4
11	bergenstidene	12000	2400	9600	5
12	bergenstidene	6000	1200	4800	6

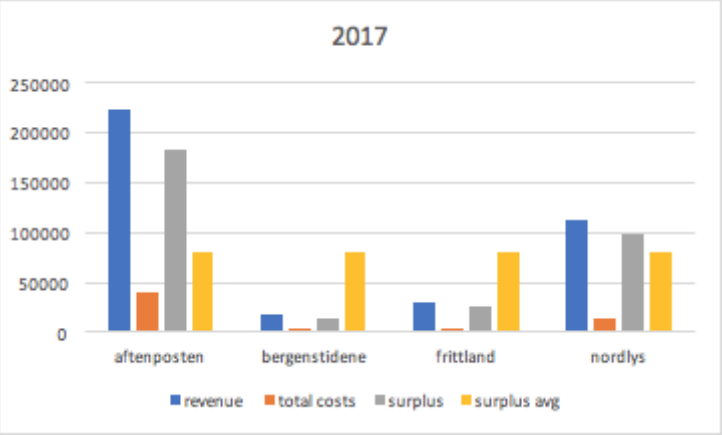
QUERY 5: How was the printing papers surplus compared to the average surplus for all the printed newspapers in 2017?

- **Here we use a nested aggregation to compare the average overall surplus to each of the printing house's surplus by adding a seperate column called "avg". The result can be presented in a diagram where you can easily see that Aftenposten is the printed paper with the highest surplus of 2017 and it is also simple to compare each surplus with the**

average.

```
SELECT  printing.printinghouse AS "Printing house",
        SUM(fact_newspapersales.quantity*fact_newspapersales.priceperpaper) AS "Revenue",
        SUM(shipping.costs+production.costs+printing.costs) AS "Total costs",
        SUM(fact_newspapersales.quantity*fact_newspapersales.priceperpaper
                                -(shipping.costs+production.costs+printing.costs)) AS "Surplus",
        AVG(SUM(fact_newspapersales.quantity*fact_newspapersales.priceperpaper
                                -(shipping.costs+production.costs+printing.costs)))
                                OVER ()
FROM    fact_newspapersales, date, shipping, printing, production
WHERE   fact_newspapersales.shipping=shipping.id and
        fact_newspapersales.date=date.id and
        fact_newspapersales.production=production.id and
        fact_newspapersales.printing=printing.id and
        date.year=2017
GROUP BY printing.printinghouse
```

Printing house character varying (128)	Revenue bigint	Total costs double precision	Surplus double precision	avg double precision
aftenposten	223500	40900	182600	80240
bergenstidene	18000	3600	14400	80240
frittland	30000	4710	25290	80240
nordlys	113400	14730	98670	80240



QUERY 6: How many new online subscribers did the company gain compared to the last term that gained >0 online subscribers?

- **We can use the LAG function to "drag down" the number of registered subscribers from the last quarter with more than zero registered subscribers. This will appear in its own column called "Last active quarter". This makes us able to compare period-to-period.**

```
SELECT date.year, date.quarter,
    count(*) AS "New Subscribers",
    LAG(count(*),1) OVER (ORDER BY date.year, date.quarter) as "Last active quarter"
FROM date, fact_onlinesub
WHERE fact_onlinesub.date = date.id
GROUP BY date.year, date.quarter
ORDER BY date.year, date.quarter
```

year integer	quarter integer	New Subscribers bigint	Last active quarter bigint
2016	1	1	[null]
2016	2	1	1
2016	4	4	1
2017	2	2	4

QUERY 7: Get an overview of what day of the week each specific subscriber chose to subscribe and what bundle each subscriber chose

- **Now we can use GROUPING which allows us to recognize if a row was grouped by a specific column.**

```
SELECT  subscriber.firstname, subscriber.surname,
        CASE WHEN GROUPING (bundle)=1 THEN 'ALL' ELSE bundle END AS "Bundle",
        CASE WHEN GROUPING (date.weekday)=1 THEN 'ALL' ELSE date.weekday END AS "Week day"
FROM    fact_onlinesub, subscriber, subscriptiontype, date
WHERE   fact_onlinesub.subscriber=subscriber.id AND
        fact_onlinesub.typeofsubscription=subscriptiontype.id AND
        fact_onlinesub.date=date.id
GROUP BY subscriber.firstname, subscriber.surname, ROLLUP (bundle, date.weekday)
```

firstname character varying	surname character varying (128)	Bundle character varying	Week day character varying
Ane Skrove	Nossum	student	monday
Brage Strand	Kristoffersen	student	thursday
Kjersti	Strand	normal	thursday
Marie Voll	Eek	student	saturday
Erica Josefine	Olsen	student	thursday
Trond	Strand	normal	friday
Jonas Modin	Rismyhr	normal	wednesday
Frida Strand	Kristofferen	student	tuesday
Ane Skrove	Nossum	student	ALL
Erica Josefine	Olsen	student	ALL
Frida Strand	Kristofferen	student	ALL
Kjersti	Strand	normal	ALL
Jonas Modin	Rismyhr	normal	ALL
Trond	Strand	normal	ALL
Marie Voll	Eek	student	ALL
Brage Strand	Kristoffersen	student	ALL
Frida Strand	Kristofferen	ALL	ALL