# PROJECT REPORT:
# Data Warehousing
# SuperX - Procurement

submitted by

Anna Anisienia

Stella Valcheva

Philipp Riedel

created as part of the Master Study Program
“Business Intelligence and Process Management”
at the Berlin School of Economics and Law

Winter semester 2017/18

Supervisor: Prof. Dr. Roland Müller

# Table of content

# Business Requirements of Interest

| Question | Business Requirement | Importance | High level entities | Measures |
|---|---|---|---|---|
| Q1 | What is the average quantity bought for each material and material type per year and month? [measured in units of products] | Medium | Material, Material type, Month, Year | Quantity |
| Q2 | What is the total quantity bought for each material and material type over time? [purchase volume measured in units of products] | High | Material, Material type, Month, Quarter, Year | Quantity |
| Q3 | What is the highest, lowest and average price per material and material type per month and year? [measured in €] | Medium | Material, Material type, Month, Year | Price |
| Q4 | What is the total order volume [measured in units of products] and value [measured in €] per month, quarter and year? | High | Month, Quarter, Year | Quantity, Total costs |
| Q5 | What is the total amount of materials delivered by specific suppliers and supplier categories filtered by specific materials? [measured in units of products] | High | Supplier, Supplier category, Material type, Material | Quantity |
| Q6 | What are the total order value [measured in €] and order volume [measured in units of products] per employee per year? | Low | Employee, Year | Total costs, Quantity |
| Q7 | Who is the cheapest and most expensive supplier for each material per month and year? [measured in price of products in €] | High | Supplier, Supplier category, Month, Year | Price [Min, Max] |
| Q8 | What is the total order value per supplier and supplier category per month and year? [measured in €] | High | Supplier, Supplier category, Month, Year | Total costs |
| Q9 | What is the total order volume per supplier and supplier category per month and year? [measured in units of products] | Medium | Supplier, Supplier category, Month, Year | Quantity |
| Q10 | What are the total order value [measured in €] and total order volume [measured in units of products] per supplier and supplier country over time? | Medium | Supplier, Supplier country, Month, Year | Total costs, Quantity |
| Q11 | What is the total quantity ordered per country? | Low | Country | Quantity |
| Q12 | What is the number of suppliers per material and supplier category per month and year? | High | Supplier, Supplier category, Material, Month, Year | Number of suppliers |

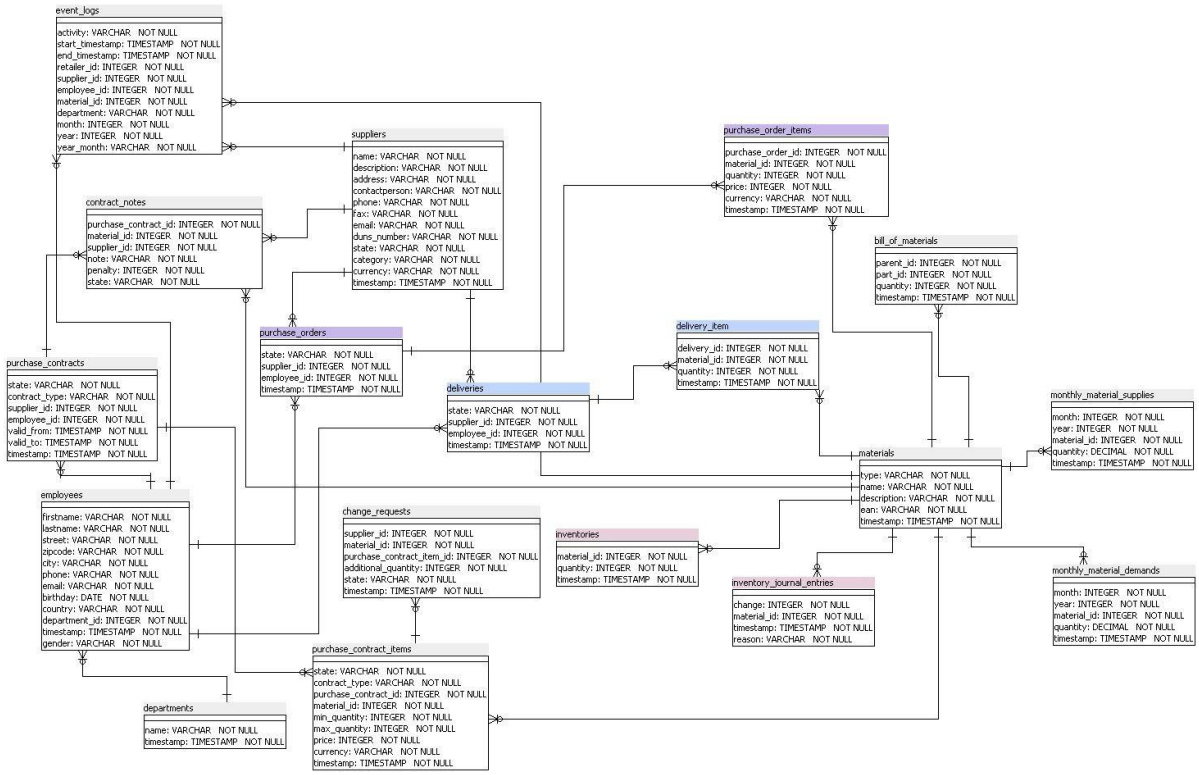
# Intention of KPIs

1. **What is the average quantity bought for each material and material type per month and year?**
   Find out which materials are bought in small, medium or large quantities in order to classify the different materials in A, B and C goods.
2. **What is the total quantity bought for each material and material type over time?**
   Find out which materials are bought in small, medium or large quantities in order to classify the different materials in A, B and C goods.
3. **What is the highest, lowest and average price per material and material type per month and year?**
   Find out the best time to buy a product. By observing the price fluctuations of each material over time, it is possible to reveal trends and patterns.
4. **What is the total order volume and value per month, quarter and year?**
   Find out the fluctuation of the order volume over time to get a general overview or trend of the procurement department.
5. **What is the total amount of materials delivered by specific suppliers and supplier categories filtered by specific materials?**
   Find out which supplier is the most important one for a particular material.
6. **What is the total order value and order volume per employee per year?**
   Find out which employee is ordering the most to create an internal ranking.
7. **Who is the cheapest and most expensive supplier for each material per month and year?**
   Find out who is the most critical supplier of a specific material in order to classify them into A, B and C suppliers.
8. **What is the total order value per supplier and supplier category per month and year?**
   Find out who is the most crucial supplier to classify them into A, B and C suppliers.
9. **What is the total order volume per supplier and supplier category per month and year?**
   Find out who is the most significant supplier, in general, to classify them into A, B and C suppliers.
10. **What are the total order value and total order volume per supplier country over time?**
    Find out on which countries the procurement department strongly depends on.
11. **What is the total quantity ordered per country?**
    Find out on which countries the procurement department strongly depends on.
12. **What is the number of suppliers per material and supplier category per month and year?**
    Find out on which supplier the procurement department strongly depends on.

# Analysis of data sources

## Significant tables

Since the goal of this project is the creation of a data mart, the team decided to start by investigating the data source and identifying tables that are relevant for Procurement department. This way, the first draft of OLTP schema has been created to facilitate further analysis of the dataset in the Data Profiling process. The following figure presents the OLTP schema for Procurement department.

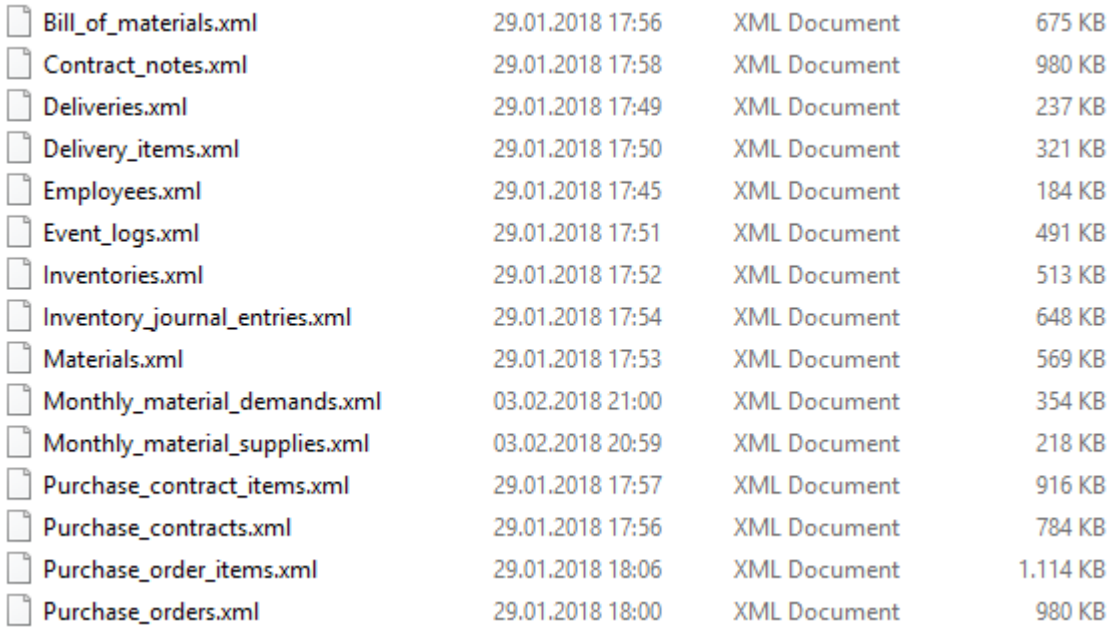



It is important to note that many Sales and Production related tables have been excluded from further analysis in order to focus solely on the Procurement department:

- **Sales:** forecasts, orders, order_items, retailers, shippings, shipping_items
- **Production:** machines, machine_types, monthly_employee_hours, production_job_definitions, production_orders.

## Data profiling

In the next step, the team performed data profiling by using SQL Server Integration Services. After establishing of Data Profiling Task in an Integration Services project and definition of profile requests, the team created several XML files including a number of data profiles that will be discussed in this section.



| | | | |
|---|---|---|---|
| Bill_of_materials.xml | 29.01.2018 17:56 | XML Document | 675 KB |
| Contract_notes.xml | 29.01.2018 17:58 | XML Document | 980 KB |
| Deliveries.xml | 29.01.2018 17:49 | XML Document | 237 KB |
| Delivery_items.xml | 29.01.2018 17:50 | XML Document | 321 KB |
| Employees.xml | 29.01.2018 17:45 | XML Document | 184 KB |
| Event_logs.xml | 29.01.2018 17:51 | XML Document | 491 KB |
| Inventories.xml | 29.01.2018 17:52 | XML Document | 513 KB |
| Inventory_journal_entries.xml | 29.01.2018 17:54 | XML Document | 648 KB |
| Materials.xml | 29.01.2018 17:53 | XML Document | 569 KB |
| Monthly_material_demands.xml | 03.02.2018 21:00 | XML Document | 354 KB |
| Monthly_material_supplies.xml | 03.02.2018 20:59 | XML Document | 218 KB |
| Purchase_contract_items.xml | 29.01.2018 17:57 | XML Document | 916 KB |
| Purchase_contracts.xml | 29.01.2018 17:56 | XML Document | 784 KB |
| Purchase_order_items.xml | 29.01.2018 18:06 | XML Document | 1.114 KB |
| Purchase_orders.xml | 29.01.2018 18:00 | XML Document | 980 KB |

The files mentioned above contain the following types of data profiles:

- Candidate Key Profiles: expresses the percentage of unique values in each attribute, thereby identifying candidates for the primary key.
- Column Length Distribution Profiles: shows the length distribution of string attributes.
- Column Null Ratio Profiles: indicates the percentage of NULL values in each column.
- Column Pattern Profiles: presents a pattern distribution in string attributes, which is based on regular expressions. This way, it enables to see invalid data entries, blank characters as well as values that deviate from a typical data format, such as EAN code or ISBN.
- Column Statistics Profiles: outlines the summary statistics for numerical values (i.e., minimum, maximum, mean and standard deviation). Furthermore, it shows the minimal and maximal value for Date attributes.
- Column Value Distribution Profiles: presents all unique attribute values in the distribution and its frequency. It could be used to create a list of valid domain values.
- Functional Dependency Profiles: shows the ratio of attribute dependencies.

In following we present our findings for each table:

1. **Bill of materials**
    a. no NULL values identified
    b. Column Statistics Profiles (min to max values):
        i. id: 1 to 35
        ii. parent_id: 16 to 31
        iii. part_id: 1 to 26
        iv. quantity: 1 to 14
        v. timestamp: all values created on 31.12.2009 and never changed
    c. Column Value Distribution Profile:
        i. most common parent_id: 19, 20, 21
        ii. most common part_id: 21, 19, 1, 2
        iii. quantity: 24 times value 1, 6 times value 4
2. **Deliveries**
    a. no NULL values identified
    b. Column Statistics Profiles (min to max values):
        i. id: 1 to 2572
        ii. employee_id: 6 to 116
        iii. supplier_id: 1 to 158
        iv. timestamp: 13.01.2010 to 17.12.2013
    c. state for all deliveries is “OK”
3. **Delivery items**
    a. no NULL values identified
    b. Column Statistics Profiles (min to max values):
        i. id: 1 to 10056
        ii. delivery_id: 1 to 2572
        iii. material_id: 1 to 15
        iv. quantity: 420 to 219678 with a mean = 15305 and sd = 38409
        v. timestamp: 13.01.2010 to 17.12.2013
    c. Column Value Distribution Profile:
        i. most commonly delivered material_id: 6 and 10
        ii. most common quantity: 507

4. **Employees**
    a. no NULL values identified
    b. Candidate Key Profile: first and last names are not unique
    c. Column Value Distribution Profile:

i. all employees have different birth dates
ii. all employees are from Berlin
iii. 23% is from Sales (#1), 22% from Procurement (#3), 20% from Production (#4), 18% from Production Planning (#2) and last 16% from Management (#5)
iv. no potential duplicates identified by looking at email addresses, first and last names, phones, street names and even zip codes
v. **all employees have gender value of "M"** which is obviously incorrect since there are a lot of female first names in the data.

d. Column Length Distribution Profile:
i. one last name with length 2 is **suspicious**:

| email | firstname | gender | id | lastname |
|---|---|---|---|---|
| jamal@marahrens.net | Jamal | M | 2 | Ne |

ii. all **zip codes** have the length of 5, which is fine, as they should come from Germany. However, there values are **not compatible with the City** they belong to! All rows are indicated as from Berlin, however zip codes in Berlin ranges from 10115 to 14199. Zip codes from the data: 99159, 98063, 96827, 96437, 96263, 95876, 94578, 94018, 93822, 93728, 92614, 92263, 90929, 90551, 89707, 89596, 88996, 88134, 87419, 86916, 85943, 85781, 85014, 83797, 83379, 82673, 81899, 80933, 79568, 76995, 76169, 74909, 73992, 71812, 71330, 70443, 69725, 69542, 68155, 67681, 66145, 65089, 64571, 64217, 63245, 61468, 61455, 61232, 61180, 61003, 60573, 60452, 60144, 60024, 59435, 59426, 58835, 57222, 56647, 55508, 54381, 53254, 52711, 52121, 51437, 49805, 49704, 49437, 48985, 48792, 48550, 44030, 41808, 41794, 40908, 40527, 37748, 36866, 36718, 36662, 36401, 35312, 35292, 34523, 34367, 33940, 33034, 31029, 30845, 30530, 28617, 25921, 24517, 24426, 24378, 23557, 23490, 22926, 22076, 21881, 21603, 21206, 20871, 19751, 19564, 18779, 18405, 16200, 15563, 15283, 14925, 14398, 14137, 14042, 13820, 13448, 12841, 12411, 10518, 10485
iii. email address:
1. one email address with a length of 34 identified → domain value "schellenbeckgummelt.name" which has not been found in the Internet → **suspicious!**
2. another email with length of 32 not found in the Internet → domain "beutelspacherkarsten.org" → **suspicious!**

3. another email domain with length of 30: “schnballbichler.name” → **suspicious!**
4. another email domain with length of 30: “stckertmotzenbbcker.net” → **suspicious!**
5. it turns out that **all email addresses are suspicious:**

| email | firstname | gender | id | lastname |
|---|---|---|---|---|
| frederike@hildenbrandgaba.de | Frederike | M | 36 | Wolf |
| ludwig@schemberaschirmer.com | Ludwig | M | 53 | Stang |
| dorian@siewertkalinowski.com | Dorian | M | 112 | Niklaus |

| email | firstname | gender | id | lastname |
|---|---|---|---|---|
| jonathan@paeslersiewert.net | Jonathan | M | 17 | Knorr |
| domenic@achillesanggreny.ch | Domenic | M | 47 | Weller |
| evelin@schneltinglorenz.org | Evelin | M | 104 | Ruch |

| email | firstname | gender | id | lastname |
|---|---|---|---|---|
| shawn@beushausendolch.info | Shawn | M | 11 | Lauckner |
| rocco@hermeckehillard.info | Rocco | M | 40 | Jungton |
| madeleine@bremerspahn.name | Madeleine | M | 92 | Scharf |
| melisa@siewertschoberg.com | Melisa | M | 110 | Heuck |

| email | firstname | gender | id | lastname |
|---|---|---|---|---|
| fine@kraftboruschewski.ch | Fine | M | 32 | Moritz |
| josefine@gollingkhnert.de | Josefine | M | 75 | Moormann |
| edwin@muckenthalervogt.de | Edwin | M | 77 | Hansen |
| svenja@schirmeister.name | Svenja | M | 79 | Letzelter |
| sophie@hillardtrampeli.ch | Sophie | M | 91 | Gamper |
| tamara@schfersteinecke.ch | Tamara | M | 116 | Illing |

e. Column Pattern Profile:

i. there is **no common corporate email address domain**, it rather looks like either all email addresses are fake, or those are employees private email addresses:

| Pattern | Pattern Percentage |
|---|---|
| \p{L}+@\p{L}+\.DE | 19 |
| \p{L}+@\p{L}+\.NET | 16 |
| \p{L}+@\p{L}+\.CH | 16 |
| \p{L}+@\p{L}+\.NAME | 15 |
| \p{L}+@\p{L}+\.INFO | 13 |
| \p{L}+@\p{L}+\.ORG | 13 |
| \p{L}+@\p{L}+\.COM | 8 |

ii. phone values are **not standardized**:

Pattern Distribution - phone

| No. | Pattern | Pattern Percentage |
|---|---|---|
| 1 | \(\d\d\d\d\d\) \d+ | 30 |
| 2 | \+49-\d\d\d\d-\d+ | 24 |
| 4 | \(\d\d\d\d\) \d+ | 24 |
| 3 | \+49-\d\d\d-\d+ | 22 |

f. Column Statistics Profiles (min to max values):

i. id: 1 to 120

ii. birthday: 3.02.1963 to 13.10.1997

iii. timestamp: all values created on 31.12.2009 and never changed → **suspicious due to natural fluctuation and changes in master data** such as change of address because someone moved, etc.

## 5. Event logs

a. there are some NULL values for supplier_id, retailer_id, material_id and employee_id, but it is probably fine, as not all event logs are related to them.

b. Column Statistics Profiles (min to max values):

Column Statistics Profiles - [dbo].[event_logs]

| Column | Minimum | Maximum | Mean | Standard Deviation |
|---|---|---|---|---|
| employee_id | 5 | 119 | 63.675010940919 | 30.3213703200436 |
| end_timestamp | 01.01.2010 10:46:27 | 01.01.2014 18:00:00 | | |
| id | 1 | 27868 | 13934.5 | 8044.79864570892 |
| material_id | 16 | 31 | 21.3946127511143 | 4.06815134329389 |
| month | 1 | 12 | 6.78943591215731 | 3.46343036038292 |
| retailer_id | 1 | 105 | 51.9854624750136 | 29.9961892212831 |
| start_timestamp | 01.01.2010 09:00:00 | 01.01.2014 14:29:15 | | |
| year | 2010 | 2014 | 2011.56642026697 | 1.12288677001162 |

c. most of the event logs are related to Production and Sales:

| Value | Count | Percentage |
|---|---|---|
| Production | 15481 | 55.5512 % |
| Sales | 6437 | 23.0982 % |
| Logistics | 3746 | 13.4419 % |
| External | 1272 | 4.5644 % |
| Procurement | 617 | 2.2140 % |
| Production Planning | 315 | 1.1303 % |

d. the most common material id: 21, 19, 20

e. majority of event logs created in December and the year 2013:

Frequent Value Distribution (0.1000 %) - month — Encrypted Connection — 1000 Rows

| Value | Count | Percentage |
|---|---|---|
| 12 | 3091 | 11.0916 % |
| 5 | 2821 | 10.1227 % |
| 7 | 2626 | 9.4230 % |
| 6 | 2577 | 9.2472 % |
| 11 | 2564 | 9.2005 % |

Frequent Value Distribution (0.1000 %) - year — Encrypted Connection

| Value | Count | Percentage |
|---|---|---|
| 2013 | 7501 | 26.9162 % |
| 2012 | 7081 | 25.4091 % |
| 2011 | 6756 | 24.2429 % |
| 2010 | 6472 | 23.2238 % |
| 2014 | 58 | 0.2081 % |

Frequent Value Distribution (0.1000 %) - year_month — Encrypted Connection

| Value | Count | Percentage |
|---|---|---|
| 2013/12 | 931 | 3.3407 % |
| 2013/7 | 783 | 2.8097 % |
| 2013/5 | 745 | 2.6733 % |
| 2012/12 | 744 | 2.6697 % |
| 2011/5 | 735 | 2.6374 % |

## 6. Inventories

a. no NULL values

b. Column Statistics Profiles:

Column Statistics Profiles - [dbo].[inventories]

| Column | Minimum | Maximum | Mean | Standard Deviation |
|---|---|---|---|---|
| material_id | 1 | 31 | 16 | 8.94427190999916 |
| quantity | 0 | 95271701 | 4808530.41935484 | 16779946.6683297 |
| timestamp | 25.12.2013 09:00:00 | 01.01.2014 09:00:00 | | |

c. Column Value Distribution Profile:

Frequent Value Distribution (0.1000 %) - timestamp — Encrypted Connection

| Value | Count | Percentage |
|---|---|---|
| 01.01.2014 09:00:00 | 10 | 32.2581 % |
| 25.12.2013 09:00:00 | 9 | 29.0323 % |
| 31.12.2013 09:00:00 | 8 | 25.8065 % |
| 27.12.2013 09:00:00 | 3 | 9.6774 % |
| 26.12.2013 09:00:00 | 1 | 3.2258 % |

7. **Inventory journal entries**
   a. no NULL values
   b. Column Statistics Profiles:

Column Statistics Profiles - [dbo].[inventory_journal_entries]

| Column | Minimum | Maximum | Mean | Standard Deviation |
|---|---|---|---|---|
| change | -504 | 219678 | 1355.30379321004 | 12428.9949823731 |
| id | 1 | 109986 | 54993.5 | 31750.2233522328 |
| material_id | 1 | 31 | 13.6773862127907 | 8.76662972876727 |
| timestamp | 13.01.2010 09:00:00 | 01.01.2014 09:00:00 | | |

   c. Column Value Distribution Profile:

Frequent Value Distribution (0.1000 %) - reason — Encrypted Connection

| Value | Count | Percentage |
|---|---|---|
| Production | 58029 | 52.7604 % |
| Shipping | 41901 | 38.0967 % |
| Delivery | 10056 | 9.1430 % |

Frequent Value Distribution (0.1000 %) - material_id — Encrypted Connection

| Value | Count | Percentage |
|---|---|---|
| 2 | 8108 | 7.3718 % |
| 8 | 5709 | 5.1907 % |
| 9 | 5650 | 5.1370 % |
| 4 | 5593 | 5.0852 % |
| 21 | 5500 | 5.0006 % |
| 6 | 5497 | 4.9979 % |

Frequent Value Distribution (0.1000 %) - change — Encrypted Connection

| Value | Count | Percentage |
|---|---|---|
| -36 | 16305 | 14.8246 % |
| -2 | 11408 | 10.3722 % |
| 36 | 8146 | 7.4064 % |
| -3 | 7673 | 6.9763 % |
| -72 | 4974 | 4.5224 % |
| -4 | 4665 | 4.2414 % |

## 8. Materials

a. no NULL values

b. all values created on 31.12.2009 and never changed.

c. Column Value Distribution Profile:

Frequent Value Distribution (0.1000 %) - type — Encrypted Connection

| Value | Count | Percentage |
|---|---|---|
| OemProduct | 12 | 38.7097 % |
| SemiFinishedProduct | 11 | 35.4839 % |
| FinishedProduct | 5 | 16.1290 % |
| RawMaterial | 3 | 9.6774 % |

Frequent Value Distribution (0.1000 %) - description — Encrypted Connection — 1000 Rows

| Value | Count | Percentage |
|---|---|---|
| Autobody with Sticker | 5 | 16.1290 % |
| Logo Stickers | 5 | 16.1290 % |
| Remote Receiver | 2 | 6.4516 % |
| Useable Controller with removable joysticks... | 1 | 3.2258 % |

d. Column Statistics Profiles: id between 1 and 31

e. Column Pattern Profile:

Pattern Distribution - description — Encrypted

| No. | Pattern | Pattern Percentage |
|---|---|---|
| 4 | (\p{L}+ )+ | 55 |
| 1 | AUTOBODY WITH STICKER | 16 |
| 2 | LOGO STICKERS | 16 |
| 3 | REMOTE RECEIVER | 6 |
| 5 | .* | 6 |

Pattern Distribution - name — Encrypted

| No. | Pattern | Pattern Percentage |
|---|---|---|
| 1 | (\p{L}+ )+LOGO STICKERS | 16 |
| 2 | (\p{L}+ )+AUTOBODY | 16 |
| 4 | SUPER-X (\p{L}+ )+ | 16 |
| 3 | \p{L}+ | 10 |
| 10 | .* | 10 |

## 9. Monthly material supplies

a. the **timestamp** attribute has **not** been **used**:

Column Null Ratio Profiles - [dbo].[monthly_material_supplies]

| Column | Null Count | Null Percentage |
|---|---|---|
| id | 0 | 0.0000 % |
| material_id | 0 | 0.0000 % |
| month | 0 | 0.0000 % |
| quantity | 0 | 0.0000 % |
| timestampe | 720 | 100.0000 % |
| year | 0 | 0.0000 % |

b. Column Statistics Profile:

Column Statistics Profiles - [dbo].[monthly_material_supplies]

| Column | Minimum | Maximum | Mean | Standard Deviation |
|---|---|---|---|---|
| id | 1 | 720 | 360.5 | 207.845896439325 |
| material_id | 1 | 15 | 8 | 4.32049379893857 |
| month | 1 | 12 | 6.5 | 3.45205252953466 |
| quantity | 6012 | 94308914 | 4983582.82638889 | 13613902.3054384 |
| year | 2010 | 2013 | 2011.5 | 1.11803398874989 |

**c.** all attributes in the data are uniquely distributed, i.e., for each material_id there is exactly one row per month within four years *2010-2013* (1*12*4 = 48):

Frequent Value Distribution (0.1000 %) - material_id

| Value | Count | Percentage |
|---|---|---|
| 1 | 48 | 6.6667 % |
| 2 | 48 | 6.6667 % |
| 3 | 48 | 6.6667 % |
| 4 | 48 | 6.6667 % |
| 5 | 48 | 6.6667 % |
| 6 | 48 | 6.6667 % |
| 7 | 48 | 6.6667 % |
| 8 | 48 | 6.6667 % |
| 9 | 48 | 6.6667 % |
| 10 | 48 | 6.6667 % |
| 11 | 48 | 6.6667 % |
| 12 | 48 | 6.6667 % |
| 13 | 48 | 6.6667 % |
| 14 | 48 | 6.6667 % |
| 15 | 48 | 6.6667 % |

**d.** the same is true for month attribute, 15 material_ids * 4 years = 60 rows for each month:

Frequent Value Distribution (0.1000 %) - month

| Value | Count | Percentage |
|---|---|---|
| 1 | 60 | 8.3333 % |
| 2 | 60 | 8.3333 % |
| 3 | 60 | 8.3333 % |
| 4 | 60 | 8.3333 % |
| 5 | 60 | 8.3333 % |
| 6 | 60 | 8.3333 % |
| 7 | 60 | 8.3333 % |
| 8 | 60 | 8.3333 % |
| 9 | 60 | 8.3333 % |
| 10 | 60 | 8.3333 % |
| 11 | 60 | 8.3333 % |
| 12 | 60 | 8.3333 % |

e. and for a year as well, i.e., 15 material_ids * 12 months = 180 rows per year:

Frequent Value Distribution (0.1000 %) - year

| Value | Count | Percentage |
|---|---|---|
| 2010 | 180 | 25.0000 % |
| 2011 | 180 | 25.0000 % |
| 2012 | 180 | 25.0000 % |
| 2013 | 180 | 25.0000 % |

However, there are **31 material_ids, not only fifteen**. Why only around half of the materials are accounted for in monthly_material_supplies table? Are only those supplied or is it just a data quality issue? The SuperX employees need to elaborate on this.

## 10. Monthly material demands

a. also here, the **timestamp** attribute has **not** been **used**:

Column Null Ratio Profiles - [dbo].[monthly_material_demands]

| Column | Null Count | Null Percentage |
|---|---|---|
| id | 0 | 0.0000 % |
| material_id | 0 | 0.0000 % |
| month | 0 | 0.0000 % |
| quantity | 0 | 0.0000 % |
| timestampe | 1488 | 100.0000 % |
| year | 0 | 0.0000 % |

b. Column summary statistics:

Column Statistics Profiles - [dbo].[monthly_material_demands]

| Column | Minimum | Maximum | Mean | Standard Deviation |
|---|---|---|---|---|
| id | 1 | 1488 | 744.5 | 429.548503276017 |
| material_id | 1 | 31 | 16 | 8.94427190999916 |
| month | 1 | 12 | 6.5 | 3.45205252953466 |
| quantity | 131 | 121572 | 3838.07190860215 | 12496.3637725469 |
| year | 2010 | 2013 | 2011.5 | 1.11803398874989 |

This time we can see that demand exists for all material_ids. Why is **only half** of them regularly **supplied**? Is it only a data quality issue or do SuperX have problems with reconciliation between supply and demand? Again, SuperX employees need to dig deeper on this matter.

c. the most common quantity values:

Frequent Value Distribution (0.1000 %) - quantity

| Value | Count | Percentage |
|---|---|---|
| 159 | 11 | 0.7392 % |
| 300 | 9 | 0.6048 % |
| 237 | 7 | 0.4704 % |
| 2146 | 7 | 0.4704 % |
| 242 | 7 | 0.4704 % |

d. Common value distribution looks very similar to monthly_material_supplies.

## 11. Suppliers

a. in this column we can see that **contact person** was **missing** 9 times:

Column Null Ratio Profiles - [dbo].[suppliers]

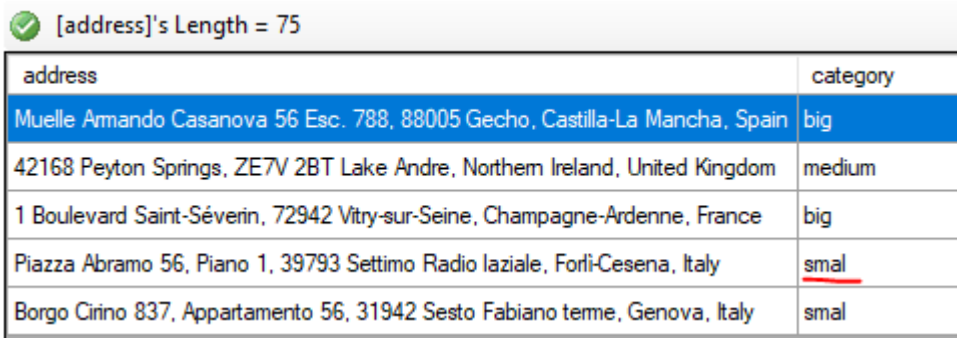
| Column | Null Count | Null Percentage |
|---|---|---|
| address | 0 | 0.0000 % |
| category | 0 | 0.0000 % |
| contactperson | 9 | 5.5901 % |
| currency | 0 | 0.0000 % |
| description | 0 | 0.0000 % |
| duns_number | 0 | 0.0000 % |
| email | 0 | 0.0000 % |
| fax | 0 | 0.0000 % |
| id | 0 | 0.0000 % |
| name | 0 | 0.0000 % |
| phone | 0 | 0.0000 % |
| state | 0 | 0.0000 % |
| timestamp | 0 | 0.0000 % |

Column Null Ratio Profiles

| address | category | contactperson | currency | description | duns_number | email | fax | id |
|---|---|---|---|---|---|---|---|---|
| Timstraat 685 I, 4... | small | | EUR | Function-based ... | 26-064-8717 | broek.jr.isa@koni... | 0614469077 | 25 |
| Borgo Pacifico 6... | big | | EUR | Conoscenza bas... | 56-968-2563 | furio.greco@rec... | +50 921 94399381 | 29 |
| Kieselstr. 32c, 40... | small | | EUR | Front-line human-... | 77-017-5642 | nolte.stefan@jan... | +49-725-2897372 | 37 |
| 34851 Abbigail B... | small | | USD | Cloned multi-taski... | 29-057-5732 | v.dietrich.mervin... | 502.719.0547 x0... | 70 |
| Heinrich-Böll-Str. ... | medium | | EUR | Enhanced eco-c... | 27-092-3492 | lenn_ba er@solln... | +49-2588-27979... | 76 |
| Regensburger Str... | medium | | EUR | User-friendly cont... | 48-600-3484 | hessler.philine@ri... | +49-1767-75211... | 132 |
| Sandenstraat 149... | medium | | EUR | Synchronised sta... | 71-688-2542 | finn.beek.den@j... | (9799) 842508 | 134 |
| In den Dehlen 22... | small | | EUR | Innovative interm... | 09-974-5155 | hasan.reusse@br... | (999) 999-999 | 137 |
| Subida Gabriel, 7... | medium | | EUR | éxito nueva gene... | 13-952-6789 | paz_eva_pagan... | 944-504-152 | 146 |

Also, we can see that some **fax numbers** are incorrect, probably because some employees were forced to enter something into NOT NULL fax field, even though this supplier does not have a fax.

**b.** In general, this table has many data quality problems, for instance, the address column is a combined field for street, city, zip-code, and country and **need to be parsed** in order to be analyzed later on in a data warehouse. It looks like the field can be split into separate parts after the comma.



[address]'s Length = 75

| address | category |
|---|---|
| Muelle Armando Casanova 56 Esc. 788, 88005 Gecho, Castilla-La Mancha, Spain | big |
| 42168 Peyton Springs, ZE7V 2BT Lake Andre, Northern Ireland, United Kingdom | medium |
| 1 Boulevard Saint-Séverin, 72942 Vitry-sur-Seine, Champagne-Ardenne, France | big |
| Piazza Abramo 56, Piano 1, 39793 Settimo Radio laziale, Forlì-Cesena, Italy | smal |
| Borgo Cirino 837, Appartamento 56, 31942 Sesto Fabiano terme, Genova, Italy | smal |

At the same time, we can see some typing errors, such as “smal” → “small”.

Column Length Distribution Profiles - [dbo].[suppliers]

| Column | Minimum Length | Maximum Length | Ignore Leading Spaces |
|---|---|---|---|
| address | 45 | 85 | ☐ |
| category | 3 | 6 | ☐ |
| contactperson | 8 | 28 | ☐ |
| currency | 3 | 3 | ☐ |
| description | 19 | 51 | ☐ |
| duns_number | 11 | 11 | ☐ |
| email | 18 | 46 | ☐ |
| fax | 9 | 21 | ☐ |
| name | 8 | 36 | ☐ |
| phone | 9 | 21 | ☐ |
| state | 6 | 6 | ☐ |

Length Distribution - address

| Length | Count | Percentage |
|---|---|---|
| 85 | 1 | 0.6211 % |
| 80 | 1 | 0.6211 % |
| 79 | 1 | 0.6211 % |
| 77 | 2 | 1.2422 % |
| 75 | 5 | 3.1056 % |
| 74 | 2 | 1.2422 % |
| 73 | 2 | 1.2422 % |
| 72 | 3 | 1.8634 % |
| 71 | 2 | 1.2422 % |
| 70 | 5 | 3.1056 % |
| 69 | 5 | 3.1056 % |
| 68 | 6 | 3.7267 % |

[address]'s Length = 85

| address | category | contactperson |
|---|---|---|
| Partida Cristian Lozada 99 Esc. 231, 41241 Alcalá de Henares, Región de Murcia, Spain | smal | Francisca Sauce... |

In this column we can also see some **potential duplicates:**

Column Value Distribution Profiles - [dbo].[suppliers]

| Column | Number Of Distinct Values |
|---|---|
| address | 161 |
| category | 4 |
| contactperson | 152 |
| currency | 5 |
| description | 161 |
| duns_number | 161 |
| email | 161 |
| fax | 126 |
| id | 161 |
| name | 161 |
| phone | 141 |
| state | 1 |
| timestamp | 48 |

Frequent Value Distribution (0.1000 %) - address

| Value | Count | Percentage |
|---|---|---|
| 1 Boulevard du Bac, 75450 Vitry-sur-Seine, Champagne-Ardenne, France | 1 | 0.6211 % |
| 1 Boulevard Saint-Séverin, 72942 Vitry-sur-Seine, Champagne-Ardenne, France | 1 | 0.6211 % |

**c.** The email addresses contain **spaces in between**:

Column Value Distribution Profiles - [dbo].[suppliers]

| Column | Number Of Distinct Values |
|---|---|
| address | 161 |
| category | 4 |
| contactperson | 152 |
| currency | 5 |
| description | 161 |
| duns_number | 161 |
| email | 161 |
| fax | 126 |

Frequent Value Distribution (0.1000 %) - email

| Value | Count |
|---|---|
| ada.buczek@biernack rzepka.pl | |
| adame_cortez_eva@zarate.com | |
| ak.leon@paszkowskigrzekowiak.com.pl | |
| alexandre.mme.martin@roux.eu | |
| angelis.de.mari o@cattaneo.it | |
| anouk_prof_linden@ruiter.net | |
| anton.wanner@priemer.info | |

Column Value Distribution Profiles

| contactperson | description | email | fax |
|---|---|---|---|
| Marino De Angelis | Struttura polarizz... | angelis.de.mari o@cattaneo.it | (999) 999-999 |

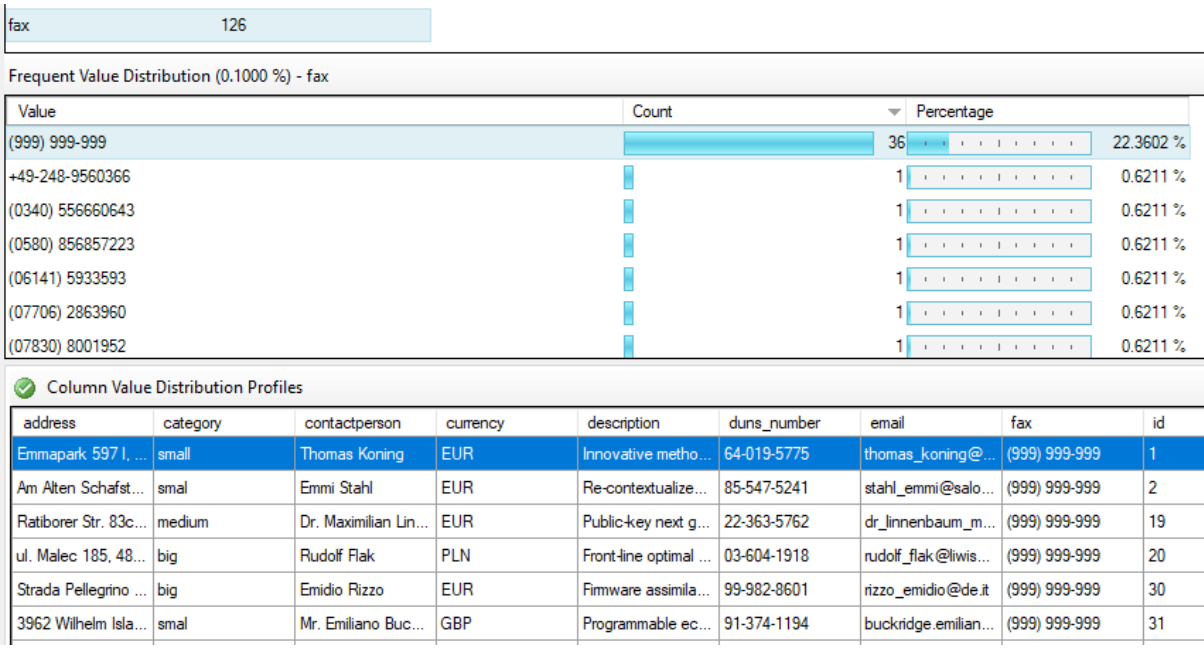
**d.** there are also some typing errors in duns_number:

| duns_number | 161 |
|---|---|

Frequent Value Distribution (0.1000 %) - duns_number

| Value | Count | Percentage |
|---|---|---|
| 07-486-2739 | 1 | 0.6211 % |
| -0-76453798 | 1 | 0.6211 % |

**e.** for the fax, the problem with this **fictive number** is existent among 22% of all cases:



fax 126

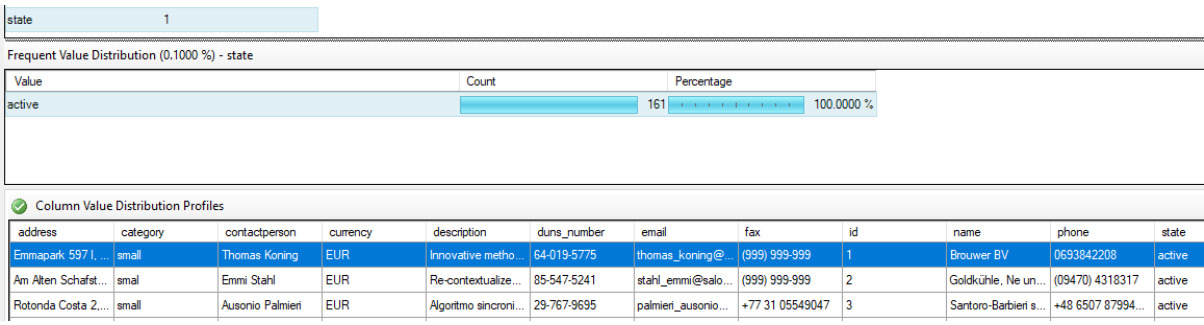
Frequent Value Distribution (0.1000 %) - fax

| Value | Count | Percentage |
|---|---|---|
| (999) 999-999 | 36 | 22.3602 % |
| +49-248-9560366 | 1 | 0.6211 % |
| (0340) 556660643 | 1 | 0.6211 % |
| (0580) 856857223 | 1 | 0.6211 % |
| (06141) 5933593 | 1 | 0.6211 % |
| (07706) 2863960 | 1 | 0.6211 % |
| (07830) 8001952 | 1 | 0.6211 % |

Column Value Distribution Profiles

| address | category | contactperson | currency | description | duns_number | email | fax | id |
|---|---|---|---|---|---|---|---|---|
| Emmapark 597 I, ... | small | Thomas Koning | EUR | Innovative metho... | 64-019-5775 | thomas_koning@... | (999) 999-999 | 1 |
| Am Alten Schafst... | smal | Emmi Stahl | EUR | Re-contextualize... | 85-547-5241 | stahl_emmi@salo... | (999) 999-999 | 2 |
| Ratiborer Str. 83c... | medium | Dr. Maximilian Lin... | EUR | Public-key next g... | 22-363-5762 | dr_linnenbaum_m... | (999) 999-999 | 19 |
| ul. Malec 185, 48... | big | Rudolf Flak | PLN | Front-line optimal ... | 03-604-1918 | rudolf_flak@liwis... | (999) 999-999 | 20 |
| Strada Pellegrino ... | big | Emidio Rizzo | EUR | Firmware assimila... | 99-982-8601 | rizzo_emidio@de.it | (999) 999-999 | 30 |
| 3962 Wilhelm Isla... | smal | Mr. Emiliano Buc... | GBP | Programmable ec... | 91-374-1194 | buckridge.emilian... | (999) 999-999 | 31 |

**f.** Column **state** seems to be **no longer maintained**, as **all** suppliers existing in the database are marked as **"active"**.



state 1

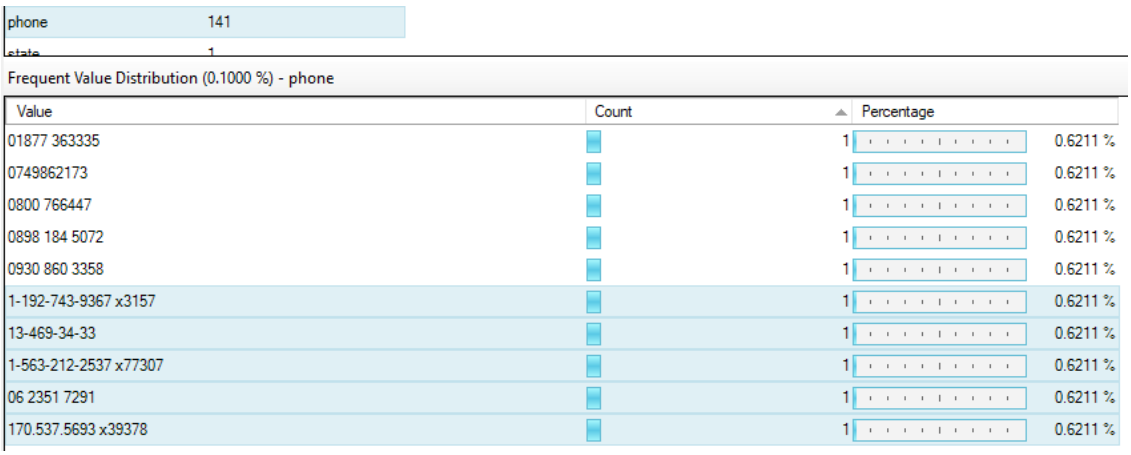
Frequent Value Distribution (0.1000 %) - state

| Value | Count | Percentage |
|---|---|---|
| active | 161 | 100.0000 % |

Column Value Distribution Profiles

| address | category | contactperson | currency | description | duns_number | email | fax | id | name | phone | state |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Emmapark 597 I, ... | small | Thomas Koning | EUR | Innovative metho... | 64-019-5775 | thomas_koning@... | (999) 999-999 | 1 | Brouwer BV | 0693842208 | active |
| Am Alten Schafst... | smal | Emmi Stahl | EUR | Re-contextualize... | 85-547-5241 | stahl_emmi@salo... | (999) 999-999 | 2 | Goldkühle, Ne un... | (09470) 4318317 | active |
| Rotonda Costa 2,... | small | Ausonio Palmieri | EUR | Algoritmo sincroni... | 29-767-9695 | palmieri_ausonio... | +77 31 05549047 | 3 | Santoro-Barbieri s... | +48 6507 87994... | active |

**g.** the **phone** number has an **inconsistent format**; sometimes the field seems to be misused, i.e., used for a different purpose than storing a phone number (some resemble more a product key rather than a phone number):



phone 141

state 1

Frequent Value Distribution (0.1000 %) - phone

| Value | Count | Percentage |
|---|---|---|
| 01877 363335 | 1 | 0.6211 % |
| 0749862173 | 1 | 0.6211 % |
| 0800 766447 | 1 | 0.6211 % |
| 0898 184 5072 | 1 | 0.6211 % |
| 0930 860 3358 | 1 | 0.6211 % |
| 1-192-743-9367 x3157 | 1 | 0.6211 % |
| 13-469-34-33 | 1 | 0.6211 % |
| 1-563-212-2537 x77307 | 1 | 0.6211 % |
| 06 2351 7291 | 1 | 0.6211 % |
| 170.537.5693 x39378 | 1 | 0.6211 % |

**h.** some supplier **names** contain **multi-value columns**, which denotes not normalized data structure:

Column Value Distribution Profiles - [dbo].[suppliers]

| Column | Number Of Distinct Values |
|---|---|
| duns_number | 161 |
| email | 161 |
| fax | 126 |
| id | 161 |
| name | 161 |
| phone | 141 |
| state | 1 |
| timestamp | 48 |

Frequent Value Distribution (0.1000 %) - name

| Value | Count | Percentage |
|---|---|---|
| Clement et Bertrand | 1 | 0.6211 % |
| Collet et Denis | 1 | 0.6211 % |
| Crona, Huels and Koelpin | 1 | 0.6211 % |
| Cummerata, Prosacco and Wisozk | 1 | 0.6211 % |
| Dam-Boer | 1 | 0.6211 % |
| Daugherty-Kassulke | 1 | 0.6211 % |
| De Santis e figli | 1 | 0.6211 % |
| De Santis SPA | 1 | 0.6211 % |
| Dekker BV | 1 | 0.6211 % |
| Douglas, Cummings and Upton | 1 | 0.6211 % |
| Doyle and Sons | 1 | 0.6211 % |
| Drabik, Wrona and Pawlikowski | 1 | 0.6211 % |
| Duval et Colin | 1 | 0.6211 % |
| Duval SA | 1 | 0.6211 % |
| Engel, Bogenrieder und Dieckmann | 1 | 0.6211 % |
| Enríquez Soto e Hijos | 1 | 0.6211 % |

Column Value Distribution Profiles

| address | category | contactperson | currency | description | duns_number | email | fax | id | name |
|---|---|---|---|---|---|---|---|---|---|
| al. Filipek 45698, ... | medium | Pani Reginald Kli... | PLN | Cloned explicit pr... | 97-530-8737 | reginald_pani_kli... | 12-197-06-57 | 113 | Drabik, Wrona and Pawlikowski |

**i.** also, the format of the fax number is lacking consistency:

| fax | 126 |
|---|---|

Frequent Value Distribution (0.1000 %) - fax

| Value | Count | Percentage |
|---|---|---|
| (08577) 5795670 | 1 | 0.6211 % |
| (1968) 199430 | 1 | 0.6211 % |
| (209) 318-4330 x0195 | 1 | 0.6211 % |
| (2157) 064120 | 1 | 0.6211 % |
| (303) 260-3072 x8456 | 1 | 0.6211 % |
| (577)158-3405 x08531 | 1 | 0.6211 % |
| (6217) 577290 | 1 | 0.6211 % |
| (734) 706-3187 | 1 | 0.6211 % |
| (9799) 842508 | 1 | 0.6211 % |
| (999) 999-999 | 36 | 22.3602 % |
| +10 70 40043449 | 1 | 0.6211 % |
| +33 158936798 | 1 | 0.6211 % |
| +33 159272905 | 1 | 0.6211 % |
| +33 167808689 | 1 | 0.6211 % |

## 12. Change_requests

**a.** since change request has no rows, data profiling cannot be performed.

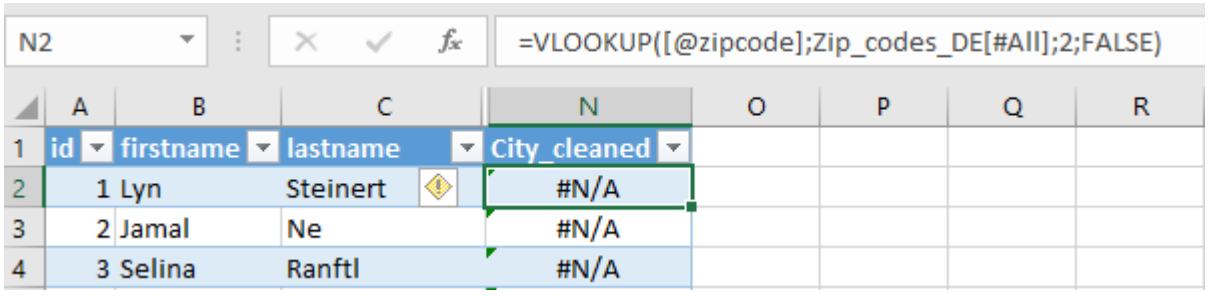
## Data cleansing - employees table

In the following section, we will present the process of cleaning the data, especially with regard to the aforementioned data quality issues.

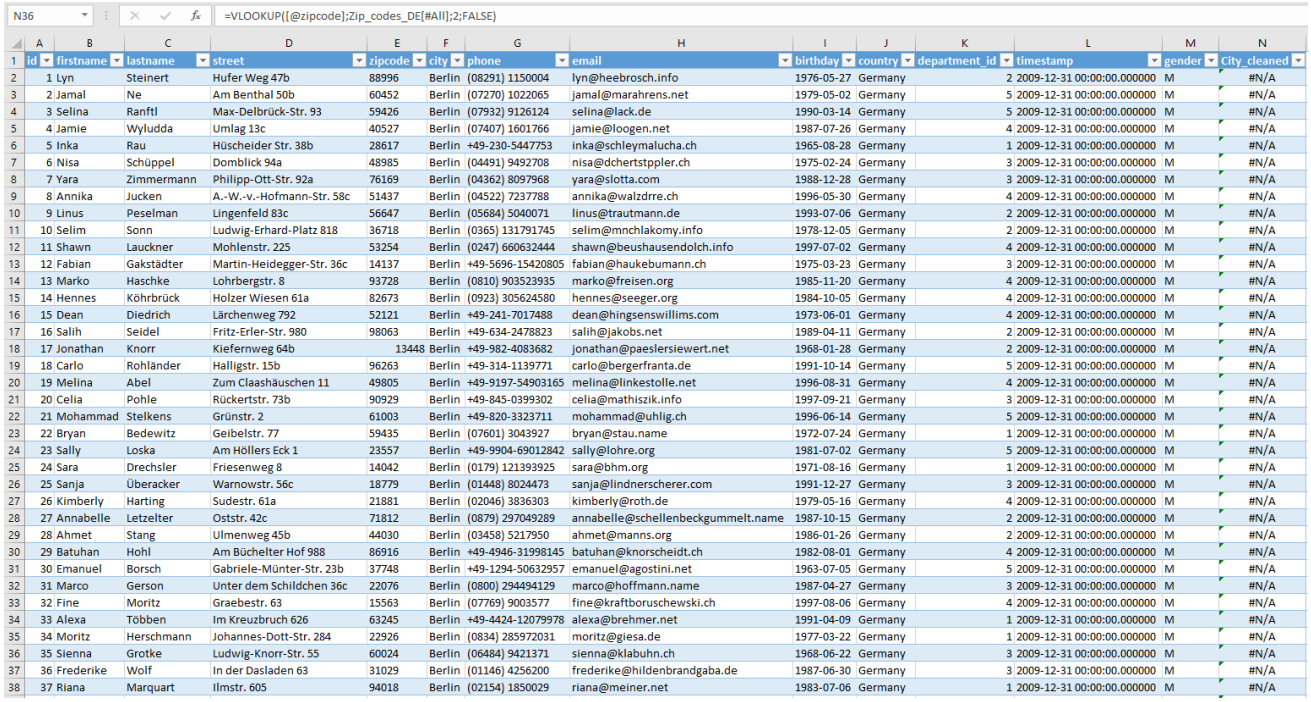
**Wrong addresses**

In the first step, we mapped the zip codes from the data to a lookup zip codes list found online[1] in order to find the right city names corresponding to the zip codes. As it turned out, the zip codes existing in the SuperX Employees database table do not exist, what we can see by looking at NAs resulting from a simple lookup:

| Plz | Ort | Vorwahl | Bundesland |
|---|---|---|---|
| 54298 | Aach | 651 | Rheinland-Pfalz |
| 78267 | Aach | 7774 | Baden-Wuerttemberg |
| 52062 | Aachen | 241 | Nordrhein-Westfalen |
| 52064 | Aachen | 241 | Nordrhein-Westfalen |
| 52066 | Aachen | 241 | Nordrhein-Westfalen |
| 52068 | Aachen | 241 | Nordrhein-Westfalen |
| 52070 | Aachen | 241 | Nordrhein-Westfalen |
| 52072 | Aachen | 241 | Nordrhein-Westfalen |
| 52074 | Aachen | 241 | Nordrhein-Westfalen |
| 52076 | Aachen | 241 | Nordrhein-Westfalen |
| 52078 | Aachen | 241 | Nordrhein-Westfalen |
| 52080 | Aachen | 241 | Nordrhein-Westfalen |
| 73430 | Aalen | 7361 | Baden-Wuerttemberg |
| 73431 | Aalen | 7361 | Baden-Wuerttemberg |
| 73432 | Aalen | 7361 | Baden-Wuerttemberg |
| 73433 | Aalen | 7361 | Baden-Wuerttemberg |
| 73434 | Aalen | 7361 | Baden-Wuerttemberg |
| 65326 | Aarbergen | 6120 | Hessen |
| 25560 | Aasbuettel | 4892 | Schlewig-Holstein |
| 29416 | Abbau Ader | | Sachsen-Anhalt |




[1] List of German zip codes, retrieved on 3.02.2018 from https://gist.github.com/jbspeakr/4565964#file-german-zip-codes-csv



N36 =VLOOKUP([@zipcode];Zip_codes_DE[#All];2;FALSE)

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | id | firstname | lastname | street | zipcode | city | phone | email | birthday | country | department_id | timestamp | gender | City_cleaned |
| 2 | 1 | Lyn | Steinert | Hufer Weg 47b | 88996 | Berlin | (08291) 1150004 | lyn@heebrosch.info | 1976-05-27 | Germany | 2 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 3 | 2 | Jamal | Ne | Am Benthal 50b | 60452 | Berlin | (07270) 1022065 | jamal@marahrens.net | 1979-05-02 | Germany | 5 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 4 | 3 | Selina | Ranftl | Max-Delbrück-Str. 93 | 59426 | Berlin | (07932) 9126124 | selina@lack.de | 1990-03-14 | Germany | 5 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 5 | 4 | Jamie | Wyludda | Umlag 13c | 40527 | Berlin | (07407) 1601766 | jamie@loogen.net | 1987-07-26 | Germany | 4 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 6 | 5 | Inka | Rau | Hüscheider Str. 38b | 28617 | Berlin | +49-230-5447753 | inka@schleymalucha.ch | 1965-08-28 | Germany | 1 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 7 | 6 | Nisa | Schüppel | Domblick 94a | 48985 | Berlin | (04491) 9492708 | nisa@dchertstppler.ch | 1975-02-24 | Germany | 3 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 8 | 7 | Yara | Zimmermann | Philipp-Ott-Str. 92a | 76169 | Berlin | (04362) 8097968 | yara@slotta.com | 1988-12-28 | Germany | 3 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 9 | 8 | Annika | Jucken | A.-W.-v.-Hofmann-Str. 58c | 51437 | Berlin | (04522) 7237788 | annika@walzdrre.ch | 1996-05-30 | Germany | 4 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 10 | 9 | Linus | Peselman | Lingenfeld 83c | 56647 | Berlin | (05684) 5040071 | linus@trautmann.de | 1993-07-06 | Germany | 2 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 11 | 10 | Selim | Sonn | Ludwig-Erhard-Platz 818 | 36718 | Berlin | (0365) 131791745 | selim@mnchlakomy.info | 1978-12-05 | Germany | 2 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 12 | 11 | Shawn | Lauckner | Mohlenstr. 225 | 53254 | Berlin | (0247) 660632444 | shawn@beushausendolch.info | 1997-07-02 | Germany | 4 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 13 | 12 | Fabian | Gakstädter | Martin-Heidegger-Str. 36c | 14137 | Berlin | +49-5696-15420805 | fabian@haukebumann.ch | 1975-03-23 | Germany | 3 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 14 | 13 | Marko | Haschke | Lohrbergstr. 8 | 93728 | Berlin | (0810) 903523935 | marko@freisen.org | 1985-11-20 | Germany | 4 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 15 | 14 | Hennes | Köhrbrück | Holzer Wiesen 61a | 82673 | Berlin | (0923) 305624580 | hennes@seeger.org | 1984-10-05 | Germany | 4 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 16 | 15 | Dean | Diedrich | Lärchenweg 792 | 52121 | Berlin | +49-241-7017488 | dean@hingsenswillims.com | 1973-06-01 | Germany | 4 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 17 | 16 | Salih | Seidel | Fritz-Erler-Str. 980 | 98063 | Berlin | +49-634-2478823 | salih@jakobs.net | 1989-04-11 | Germany | 2 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 18 | 17 | Jonathan | Knorr | Kiefernweg 64b | 13448 | Berlin | +49-982-4083682 | jonathan@paeslersiewert.net | 1968-01-28 | Germany | 2 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 19 | 18 | Carlo | Rohländer | Halligstr. 15b | 96263 | Berlin | +49-314-1139771 | carlo@bergerfranta.de | 1991-10-14 | Germany | 5 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 20 | 19 | Melina | Abel | Zum Claashäuschen 11 | 49805 | Berlin | +49-9197-54903165 | melina@linkestolle.net | 1996-08-31 | Germany | 4 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 21 | 20 | Celia | Pohle | Rückertstr. 73b | 90929 | Berlin | +49-845-0399302 | celia@mathiszik.info | 1997-09-21 | Germany | 3 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 22 | 21 | Mohammad | Stelkens | Grünstr. 2 | 61003 | Berlin | +49-820-3323711 | mohammad@uhlig.ch | 1996-06-14 | Germany | 5 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 23 | 22 | Bryan | Bedewitz | Geibelstr. 77 | 59435 | Berlin | (07601) 3043927 | bryan@stau.name | 1972-07-24 | Germany | 1 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 24 | 23 | Sally | Loska | Am Höllers Eck 1 | 23557 | Berlin | +49-9904-69012842 | sally@lohre.org | 1981-07-02 | Germany | 5 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 25 | 24 | Sara | Drechsler | Friesenweg 8 | 14042 | Berlin | (0179) 121393925 | sara@bhm.org | 1971-08-16 | Germany | 1 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 26 | 25 | Sanja | Überacker | Warnowstr. 56c | 18779 | Berlin | (01448) 8024473 | sanja@lindnerscherer.com | 1991-12-27 | Germany | 3 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 27 | 26 | Kimberly | Harting | Sudestr. 61a | 21881 | Berlin | (02046) 3836303 | kimberly@roth.de | 1979-05-16 | Germany | 4 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 28 | 27 | Annabelle | Letzelter | Oststr. 42c | 71812 | Berlin | (0879) 297049289 | annabelle@schellenbeckgummelt.name | 1987-10-15 | Germany | 2 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 29 | 28 | Ahmet | Stang | Ulmenweg 45b | 44030 | Berlin | (03458) 5217950 | ahmet@manns.org | 1986-01-26 | Germany | 2 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 30 | 29 | Batuhan | Hohl | Am Büchelter Hof 988 | 86916 | Berlin | +49-4946-31998145 | batuhan@knorscheidt.ch | 1982-08-01 | Germany | 4 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 31 | 30 | Emanuel | Borsch | Gabriele-Münter-Str. 23b | 37748 | Berlin | +49-1294-50632957 | emanuel@agostini.net | 1963-07-05 | Germany | 5 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 32 | 31 | Marco | Gerson | Unter dem Schildchen 36c | 22076 | Berlin | (0800) 294494129 | marco@hoffmann.name | 1987-04-27 | Germany | 3 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 33 | 32 | Fine | Moritz | Graebestr. 63 | 15563 | Berlin | (07769) 9003577 | fine@kraftboruschewski.ch | 1997-08-06 | Germany | 4 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 34 | 33 | Alexa | Többen | Im Kreuzbruch 626 | 63245 | Berlin | +49-4424-12079978 | alexa@brehmer.net | 1991-04-09 | Germany | 1 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 35 | 34 | Moritz | Herschmann | Johannes-Dott-Str. 284 | 22926 | Berlin | (0834) 285972031 | moritz@giesa.de | 1977-03-22 | Germany | 1 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 36 | 35 | Sienna | Grotke | Ludwig-Knorr-Str. 55 | 60024 | Berlin | (06484) 9421371 | sienna@klabuhn.ch | 1968-06-22 | Germany | 3 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 37 | 36 | Frederike | Wolf | In der Dasladen 63 | 31029 | Berlin | (01146) 4256200 | frederike@hildenbrandgaba.de | 1987-06-30 | Germany | 3 | 2009-12-31 00:00:00.000000 | M | #N/A |
| 38 | 37 | Riana | Marquart | Ilmstr. 605 | 94018 | Berlin | (02154) 1850029 | riana@meiner.net | 1983-07-06 | Germany | 1 | 2009-12-31 00:00:00.000000 | M | #N/A |

Therefore, to solve the problem of wrong addresses, we looked up several addresses from the table online. A simple Google search proved that the addresses come from Leverkusen, not from Berlin. As we assume, the employees must be working in the same city, since SuperX is a small company with just 120 employees. To improve the data quality, we therefore simply set the “city” to Leverkusen.

```
UPDATE employees
SET city = 'Leverkusen'
WHERE city = 'Berlin';
```

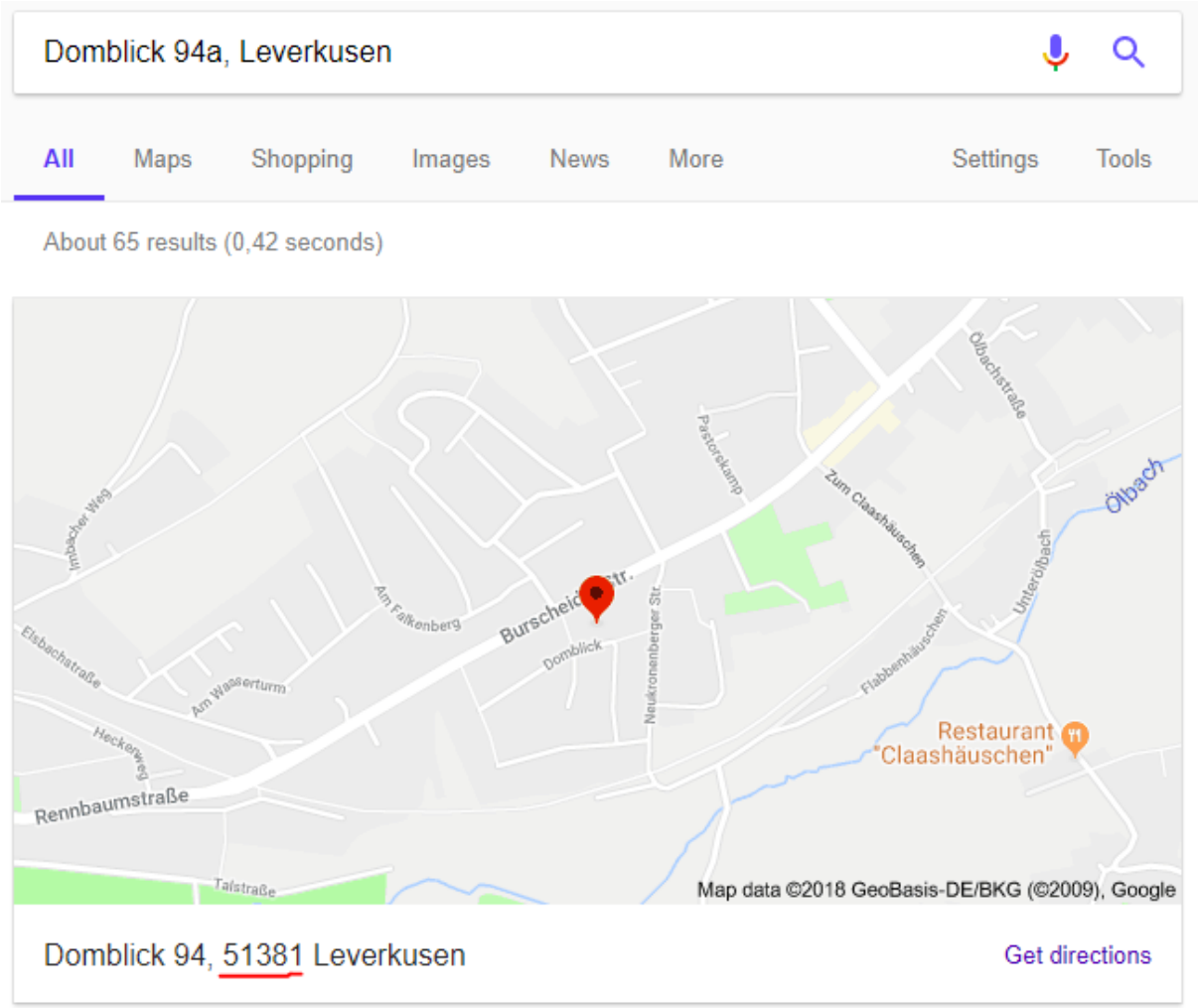
100 %

Messages

```
(120 rows affected)
```

**To find the right postal codes, we could either:**

- set all to the most common zip code,
- we could also buy German address list from Deutsche Post[2]
- make a compromise and just delete the column zip_code
- manually look up every single address in google, such as here:




We decided on the first option, because:

- This value is not relevant for our Data Mart - we will not aggregate Procurement related data such as purchase orders by the employees' place of residence!
- It still enables to improve the data quality later on
- by using the most common zip code the addresses still can be right for some of the records.
- Leverkusen has, according to our data, just six zip codes. To save time, we naïvely assume, that all employees live in the most populated district 51381:

[2] German address list, retrieved on 3.02.2018 from https://www.deutschepost.de/de/a/adressleistungen.html

| Plz | Ort | Vorwahl | Bundesland |
|---|---|---|---|
| 51371 | Leverkusen | 214 | Nordrhein-Westfalen |
| 51373 | Leverkusen | 214 | Nordrhein-Westfalen |
| 51375 | Leverkusen | 214 | Nordrhein-Westfalen |
| 51377 | Leverkusen | 214 | Nordrhein-Westfalen |
| 51379 | Leverkusen | 214 | Nordrhein-Westfalen |
| 51381 | Leverkusen | 214 | Nordrhein-Westfalen |

```
UPDATE employees
SET zipcode = '51381';
```

100 %

Messages

```
(120 rows affected)
```

**Wrong genders**

| name | sex |
|---|---|
| Aaden | boy |
| Aaliyah | girl |
| Aarav | boy |
| Aaron | boy |
| Aaron | girl |
| Ab | boy |
| Abagail | girl |
| Abb | boy |
| Abbey | girl |
| Abbie | boy |
| Abbie | girl |
| Abbigail | girl |
| Abbott | boy |
| Abby | girl |

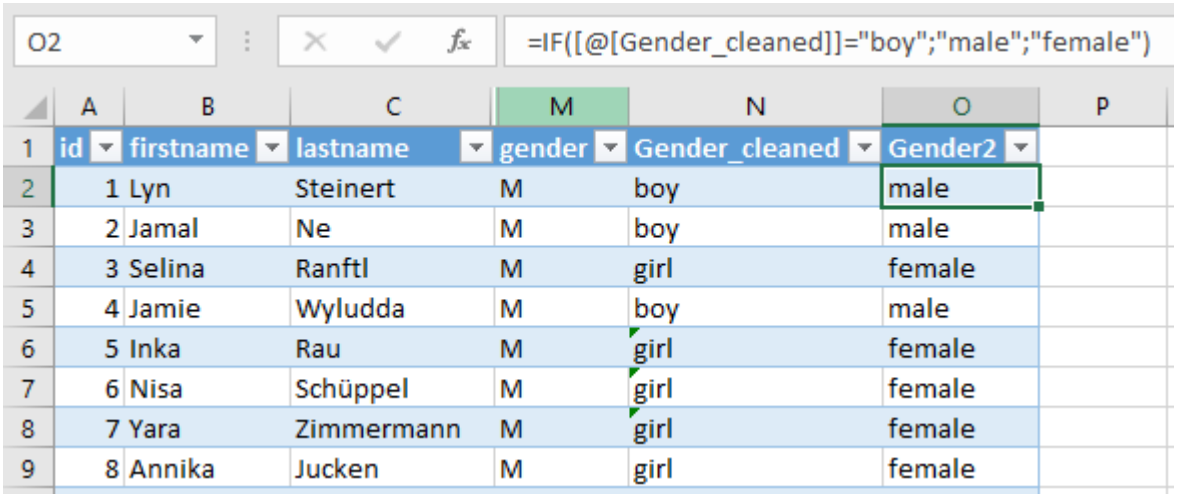
In the next step, we addressed the issue of incorrect genders. Since we saw that all genders are set to “M”, even though some of the employee names are female, we downloaded from the Internet a list containing first names, associated with each gender[3]. This way, we want to assign the right gender to each employee record automatically.

It turned out that 45 out of 120 names were not accounted for in the list. Therefore we manually adjusted those values. Consequently, we obtained the following result:

[3] Baby names, retrieved on 03.02.2018 from https://raw.githubusercontent.com/hadley/data-baby-names/master/baby-names.csv

| id | firstname | lastname | email | birthday | country | depa | timestamp | gender | Gender_cleaned |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Lyn | Steinert | lyn@heebrosch.info | 1976-05-27 | Germany | 2 | 2009-12-31 00:00:00.000000 | M | boy |
| 2 | Jamal | Ne | jamal@marahrens.net | 1979-05-02 | Germany | 5 | 2009-12-31 00:00:00.000000 | M | boy |
| 3 | Selina | Ranftl | selina@lack.de | 1990-03-14 | Germany | 5 | 2009-12-31 00:00:00.000000 | M | girl |
| 4 | Jamie | Wyludda | jamie@loogen.net | 1987-07-26 | Germany | 4 | 2009-12-31 00:00:00.000000 | M | boy |
| 5 | Inka | Rau | inka@schleymalucha.ch | 1965-08-28 | Germany | 1 | 2009-12-31 00:00:00.000000 | M | girl |
| 6 | Nisa | Schüppel | nisa@dchertstppler.ch | 1975-02-24 | Germany | 3 | 2009-12-31 00:00:00.000000 | M | girl |
| 7 | Yara | Zimmermann | yara@slotta.com | 1988-12-28 | Germany | 3 | 2009-12-31 00:00:00.000000 | M | girl |
| 8 | Annika | Jucken | annika@walzdrre.ch | 1996-05-30 | Germany | 4 | 2009-12-31 00:00:00.000000 | M | girl |
| 9 | Linus | Peselman | linus@trautmann.de | 1993-07-06 | Germany | 2 | 2009-12-31 00:00:00.000000 | M | boy |
| 10 | Selim | Sonn | selim@mnchlakomy.info | 1978-12-05 | Germany | 2 | 2009-12-31 00:00:00.000000 | M | boy |
| 11 | Shawn | Lauckner | shawn@beushausendolch.inf | 1997-07-02 | Germany | 4 | 2009-12-31 00:00:00.000000 | M | boy |
| 12 | Fabian | Gakstädter | fabian@haukebumann.ch | 1975-03-23 | Germany | 3 | 2009-12-31 00:00:00.000000 | M | boy |
| 13 | Marko | Haschke | marko@freisen.org | 1985-11-20 | Germany | 4 | 2009-12-31 00:00:00.000000 | M | boy |
| 14 | Hennes | Köhrbrück | hennes@seeger.org | 1984-10-05 | Germany | 4 | 2009-12-31 00:00:00.000000 | M | boy |
| 15 | Dean | Diedrich | dean@hingsenswillims.com | 1973-06-01 | Germany | 4 | 2009-12-31 00:00:00.000000 | M | boy |
| 16 | Salih | Seidel | salih@jakobs.net | 1989-04-11 | Germany | 2 | 2009-12-31 00:00:00.000000 | M | boy |
| 17 | Jonathan | Knorr | jonathan@paeslersiewert.net | 1968-01-28 | Germany | 2 | 2009-12-31 00:00:00.000000 | M | boy |
| 18 | Carlo | Rohländer | carlo@bergerfranta.de | 1991-10-14 | Germany | 5 | 2009-12-31 00:00:00.000000 | M | boy |
| 19 | Melina | Abel | melina@linkestolle.net | 1996-08-31 | Germany | 4 | 2009-12-31 00:00:00.000000 | M | girl |
| 20 | Celia | Pohle | celia@mathiszik.info | 1997-09-21 | Germany | 3 | 2009-12-31 00:00:00.000000 | M | girl |

Now we will only exchange the labels “boy” → “male” and “girl” → “female”.




| | A | B | C | M | N | O | P |
|---|---|---|---|---|---|---|---|
| 1 | id | firstname | lastname | gender | Gender_cleaned | Gender2 | |
| 2 | 1 | Lyn | Steinert | M | boy | male | |
| 3 | 2 | Jamal | Ne | M | boy | male | |
| 4 | 3 | Selina | Ranftl | M | girl | female | |
| 5 | 4 | Jamie | Wyludda | M | boy | male | |
| 6 | 5 | Inka | Rau | M | girl | female | |
| 7 | 6 | Nisa | Schüppel | M | girl | female | |
| 8 | 7 | Yara | Zimmermann | M | girl | female | |
| 9 | 8 | Annika | Jucken | M | girl | female | |

**Inconsistent format of phone numbers**

Even though the phone numbers are not in a consistent format, we leave it unchanged, since this process requires reconciliation with SuperX employees, as we cannot know for sure, which of them are landline and mobile phone numbers.

**Wrong email addresses**

Before we can upload the cleaned data into SQL Server, we need to deal with the problematic email domains. At the moment, the email addresses from employees appear to be fake. Even if those are correct private email addresses from employees, we want to make sure that SuperX uses correct data and this can only be achieved by setting certain communication standards. We, therefore, investigated free email domains and recommend reserving a domain "**superx.web**" which is currently available for free[4].

To demonstrate how this change could look like, and at the same time to improve data quality, we replaced the current email addresses with newly created email addresses in the form: "**firstname.lastname@superx.web**". The following figure demonstrates the final version that has been uploaded to SQL Server:

```
SELECT TOP (1000) [id]
      ,[firstname]
      ,[lastname]
      ,[street]
      ,[zipcode]
      ,[city]
      ,[phone]
      ,[birthday]
      ,[country]
      ,[department_id]
      ,[timestamp]
      ,[gender]
      ,[email]
  FROM [SuperX].[dbo].[cleansed_employees]
```

Results | Messages

| id | firstname | lastname | street | zipcode | city | phone | birthday | country | department_id | timestamp | gender | email |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Lyn | Steinert | Hufer Weg 47b | 51381 | Leverkusen | (08291) 1150004 | 1976-05-27 00:00:00.0000000 | Germany | 2 | 2009-12-31 00:00:00.0000000 | male | lyn.steinert@superx.web |
| 2 | Jamal | Ne | Am Benthal 50b | 51381 | Leverkusen | (07270) 1022065 | 1979-05-02 00:00:00.0000000 | Germany | 5 | 2009-12-31 00:00:00.0000000 | male | jamal.ne@superx.web |
| 3 | Selina | Ranftl | Max-Delbrück-Str. 93 | 51381 | Leverkusen | (07932) 9126124 | 1990-03-14 00:00:00.0000000 | Germany | 5 | 2009-12-31 00:00:00.0000000 | female | selina.ranftl@superx.web |
| 4 | Jamie | Wyludda | Umlag 13c | 51381 | Leverkusen | (07407) 1601766 | 1987-07-26 00:00:00.0000000 | Germany | 4 | 2009-12-31 00:00:00.0000000 | male | jamie.wyludda@superx.web |
| 5 | Inka | Rau | Hüscheider Str. 38b | 51381 | Leverkusen | +49-230-5447753 | 1965-08-28 00:00:00.0000000 | Germany | 1 | 2009-12-31 00:00:00.0000000 | female | inka.rau@superx.web |
| 6 | Nisa | Schüppel | Domblick 94a | 51381 | Leverkusen | (04491) 9492708 | 1975-02-24 00:00:00.0000000 | Germany | 3 | 2009-12-31 00:00:00.0000000 | female | nisa.schüppel@superx.web |
| 7 | Yara | Zimmermann | Philipp-Ott-Str. 92a | 51381 | Leverkusen | (04362) 8097968 | 1988-12-28 00:00:00.0000000 | Germany | 3 | 2009-12-31 00:00:00.0000000 | female | yara.zimmermann@superx.web |
| 8 | Annika | Jucken | A.-W.-v.-Hofmann-Str. 58c | 51381 | Leverkusen | (04522) 7237788 | 1996-05-30 00:00:00.0000000 | Germany | 4 | 2009-12-31 00:00:00.0000000 | female | annika.jucken@superx.web |
| 9 | Linus | Peselman | Lingenfeld 83c | 51381 | Leverkusen | (05684) 5040071 | 1993-07-06 00:00:00.0000000 | Germany | 2 | 2009-12-31 00:00:00.0000000 | male | linus.peselman@superx.web |
| 10 | Selim | Sonn | Ludwig-Erhard-Platz 818 | 51381 | Leverkusen | (0365) 131791745 | 1978-12-05 00:00:00.0000000 | Germany | 2 | 2009-12-31 00:00:00.0000000 | male | selim.sonn@superx.web |
| 11 | Shawn | Lauckner | Mohlenstr. 225 | 51381 | Leverkusen | (0247) 660632444 | 1997-07-02 00:00:00.0000000 | Germany | 4 | 2009-12-31 00:00:00.0000000 | male | shawn.lauckner@superx.web |
| 12 | Fabian | Gakstädter | Martin-Heidegger-Str. 36c | 51381 | Leverkusen | +49-5696-15420805 | 1975-03-23 00:00:00.0000000 | Germany | 3 | 2009-12-31 00:00:00.0000000 | male | fabian.gakstädter@superx.web |
| 13 | Marko | Haschke | Lohrbergstr. 8 | 51381 | Leverkusen | (0810) 903523935 | 1985-11-20 00:00:00.0000000 | Germany | 4 | 2009-12-31 00:00:00.0000000 | male | marko.haschke@superx.web |
| 14 | Hennes | Köhrbrück | Holzer Wiesen 61a | 51381 | Leverkusen | (0923) 305624580 | 1984-10-05 00:00:00.0000000 | Germany | 4 | 2009-12-31 00:00:00.0000000 | male | hennes.köhrbrück@superx.web |
| 15 | Dean | Diedrich | Lärchenweg 792 | 51381 | Leverkusen | +49-241-7017488 | 1973-06-01 00:00:00.0000000 | Germany | 4 | 2009-12-31 00:00:00.0000000 | male | dean.diedrich@superx.web |
| 16 | Salih | Seidel | Fritz-Erler-Str. 980 | 51381 | Leverkusen | +49-634-2478823 | 1989-04-11 00:00:00.0000000 | Germany | 2 | 2009-12-31 00:00:00.0000000 | male | salih.seidel@superx.web |
| 17 | Jonathan | Knorr | Kiefernweg 64b | 51381 | Leverkusen | +49-982-4083682 | 1968-01-28 00:00:00.0000000 | Germany | 2 | 2009-12-31 00:00:00.0000000 | male | jonathan.knorr@superx.web |
| 18 | Carlo | Rohländer | Halligstr. 15b | 51381 | Leverkusen | +49-314-1139771 | 1991-10-14 00:00:00.0000000 | Germany | 5 | 2009-12-31 00:00:00.0000000 | male | carlo.rohländer@superx.web |
| 19 | Melina | Abel | Zum Claashäuschen 11 | 51381 | Leverkusen | +49-9197-54903165 | 1996-08-31 00:00:00.0000000 | Germany | 4 | 2009-12-31 00:00:00.0000000 | female | melina.abel@superx.web |
| 20 | Celia | Pohle | Rückertstr. 73b | 51381 | Leverkusen | +49-845-0399302 | 1997-09-21 00:00:00.0000000 | Germany | 3 | 2009-12-31 00:00:00.0000000 | female | celia.pohle@superx.web |

[4] State: 04.02.2018: https://www.united-domains.de/domain-registrieren/

## Data cleansing - supplier table

In following, we cleanse the data in order to use it in the Data Mart. For this purpose, we will now address all issues identified during Data Profiling such as the problem of inconsistent phone or fax number format, as we will not use those columns in Data Mart.

**Parse addresses**

In the first step we parsed the combined address column into five separate fields:

- street
- zipcode
- city
- region
- country.

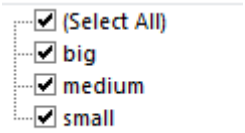
We also adjusted some typing errors (ex. " Breme%" → "Bremen") and inconsistent naming such as "U.S.A" → "USA".

For parsing we used just some simple Excel functions such as:

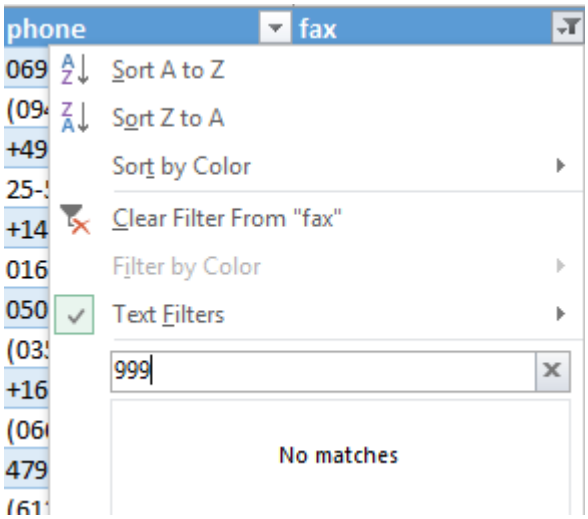
- =LEFT(F2;(FIND(" ";F2;1)-1))
- =MID(F2;FIND(" ";F2)+1;256)
- =TRIM()

| address | street | zipcode | city | region | country |
|---|---|---|---|---|---|
| Emmapark 597 I, 5326 NH Maas aan de IJssel, Limburg, Netherlands | Emmapark 597 I | 5326 | NH Maas aan de Ijssel | Limburg | Netherlands |
| Am Alten Schafstall 85b, 70119 Ost Fabianland, Hamburg, Deutschland | Am Alten Schafstall 85b | 70119 | Ost Fabianland | Hamburg | Deutschland |
| Rotonda Costa 2, 41981 Sesto Elga, Varese, Italy | Rotonda Costa 2 | 41981 | Sesto Elga | Varese | Italy |
| Meckhofer Feld 18, 14490 Jarosburg, Bremen, Deutschland | Meckhofer Feld 18 | 14490 | Jarosburg | Bremen | Deutschland |
| 498 Rutherford Row, M1O 5I8 Tremblaymouth, Manitoba, Canada | 498 Rutherford Row | M1O-5I8 | Tremblaymouth | Manitoba | Canada |
| 191 Leannon Ville, U3R3R6 East Savanna, Ontario, Canada | 191 Leannon Ville | U3R3R6 | East Savanna | Ontario | Canada |
| Glorieta Gabriela Cardenas, 85, 99566 Almería, Región de Murcia, Spain | Glorieta Gabriela Cardenas 85 | 99566 | Almería | Región de Murcia | Spain |
| Vriesplantsoen 888, 6760 QK Oud Annesluus, Limburg, Netherlands | Vriesplantsoen 888 | 6760-QK | Oud Annesluus | Limburg | Netherlands |
| 990 Patsy View, 89563-9582 Lake Samson, North Dakota, USA | 990 Patsy View | 89563-9582 | Lake Samson | North Dakota | USA |
| 97 Quai de la Harpe, 94423 Neuilly-sur-Seine, Basse-Normandie, France | 97 Quai de la Harpe | 94423 | Neuilly-sur-Seine | Basse-Normandie | France |
| Ronda Sergio Mondragón, 74 Esc. 719, 26010 Huelva, Aragón, Spain | Ronda Sergio Mondragón 74 Esc. 719 | 26010 | Huelva | Aragón | Spain |
| 7 Rue de Presbourg, 88873 Montreuil, Corse, France | 7 Rue de Presbourg | 88873 | Montreuil | Corse | France |
| Rotonda Eufemia 9, 63803 Ferrara laziale, Firenze, Italy | Rotonda Eufemia 9 | 63803 | Ferrara laziale | Firenze | Italy |
| 37707 Douglas Haven, 23066 West Reannaborough, South Carolina, USA | 37707 Douglas Haven | 23066 | West Reannaborough | South Carolina | USA |
| Solinger Str. 35a, 28176 Tuchscheid, Sachsen, Deutschland | Solinger Str. 35a | 28176 | Tuchscheid | Sachsen | Deutschland |
| 819 Noble Island, O7D 9X5 Hammesport, Northwest Territories, Canada | 819 Noble Island | O7D-9X5 | Hammesport | Northwest Territories | Canada |
| Masía Carolina, 19, 86678 Orense, Galicia, Spain | Masía Carolina 19 | 86678 | Orense | Galicia | Spain |
| Kaiserplatz 4, 59239 Ianburg, Hessen, Germany | Kaiserplatz 4 | 59239 | Ianburg | Hessen | Germany |
| Ratiborer Str. 83c, 66728 Alt Levi, Berlin, Deutschland | Ratiborer Str. 83c | 66728 | Alt Levi | Berlin | Deutschland |
| ul. Malec 185, 48-181 Pruchnik, Łódzkie, Poland | ul. Malec 185 | 48-181 | Pruchnik | Łódzkie | Poland |
| Am Alten Sc$afstall %5b, 70119 Ost Fab*anland, Hamburg, Deutschland | Am Alten Schafstall 85b | 70119 | Ost Fabianland | Hamburg | Deutschland |
| R%nda Ser*io Mondragón, 74 Esc. 719, 26010 Huelva, Aragón, Spain | Ronda Sergio Mondragón 74 Esc. 719 | 26010 | Huelva | Aragón | Spain |
| Meckhofer*Feld 18, 1449$ Jarosburg, Breme%, Deutschland | Meckhofer Feld 18 | 1449 | Jarosburg | Bremen | Deutschland |
| Am Alten Schafst$%I 85b, 70119 O*t Fabianland, Hamburg, Deutschland | Am Alten Schafstal 85b | 70119 | Ost Fabianland | Hamburg | Deutschland |
| Timstraat 685 I, 4691 IF Noord Rickdorp, Limburg, Netherlands | Timstraat 685 I | 4691 | IF Noord Rickdorp | Limburg | Netherlands |
| 6 Quai La Boétie, 75090 Mérignac, Rhône-Alpes, France | 6 Quai La Boétie | 75090 | Mérignac | Rhône-Alpes | France |
| 95 Impasse de Montmorency, 78190 Tourcoing, Nord-Pas-de-Calais, France | 95 Impasse de Montmorency | 78190 | Tourcoing | Nord-Pas-de-Calais | France |
| 8981 Anabel Causeway, 92756-1799 Mozellefurt, Pennsylvania, U.S.A. | 8981 Anabel Causeway | 92756-1799 | Mozellefurt | Pennsylvania | USA |
| Borgo Pacifico 67, 95947 Borgo Jarno, Massa-Carrara, Italy | Borgo Pacifico 67 | 95947 | Borgo Jarno | Massa-Carrara | Italy |
| Strada Pellegrino 9, Piano 3, 86800 Settimo Sasha veneto, Taranto, Italy | Strada Pellegrino 9, Piano 3 | 86800 | Settimo Sasha veneto | Taranto | Italy |
| al. Sienkiewicz 93362, 22-712 Giżycko, Zachodniopomorskie, Poland | al. Sienkiewicz 93362 | 22-712 | Giżycko | Zachodniopomorskie | Poland |
| 3962 Wilhelm Islands, W7 7RJ West Reyesstad, Northern Ireland, United Kingdom | 3962 Wilhelm Islands | W7-7RJ | West Reyesstad | Northern Ireland | United Kingdom |
| 79094 Macejkovic Keys, 67887 Carrollmouth, Texas, USA | 79094 Macejkovic Keys | 67887 | Carrollmouth | Texas | USA |
| Speestr. 20, 82319 Gürbigscheid, Schleswig-Holstein, Germany | Speestr. 20 | 82319 | Gürbigscheid | Schleswig-Holstein | Germany |
| Heidberg 84b, 90469 West Tizian, Hessen, Germany | Heidberg 84b | 90469 | West Tizian | Hessen | Germany |
| 5519 Ewald Station, E3 5UY West Rosalynbury, England, United Kingdom | 5519 Ewald Station | E3-5UY | West Rosalynbury | England | United Kingdom |

Even though column state and timestamp are not important for Data Mart, we leave them for now. Then we adjusted the category “smal” into “small”:




Then we set all fax numbers equal to (999) 999-999 to NULL.




Since we could not talk to neither SuperX employees nor to their suppliers, we could not find out the right contact person for missing values in the column “contact_person”.

We did not elaborate too detailed on the duplicates, except for the one “usual suspect” which we discovered during Data Profiling:

| id | name | description | address |
|---|---|---|---|
| 93 | Roche et Michel | Re-contextualized holistic interface | 1 Boulevard Saint-Séverin, 72942 Vitry-sur-Seine, Champagne-Ardenne, France |
| 106 | Laine SEM | Business-focused eco-centric challenge | 1 Boulevard du Bac, 75450 Vitry-sur-Seine, Champagne-Ardenne, France |

| street | zipcode | city | region | country | contactperson | phone | fax | email | duns_number |
|---|---|---|---|---|---|---|---|---|---|
| 1 Boulevard Saint-Séverin | 72942 | Vitry-sur-Seine | Champagne-Ardenne | France | Roux Paul | +33 226480482 | | paul_roux@caron.com | 80-048-0728 |
| 1 Boulevard du Bac | 75450 | Vitry-sur-Seine | Champagne-Ardenne | France | Prof Noa Bernard | 0749862173 | +33 456956472 | prof_noa_berna d@carpentier.fr | 99-678-1154 |

We did not find enough evidence for duplicate between those two rows. However, they are definitely similar, especially the address.

Also, we corrected two records with the wrong pattern, presumably typing error:

| email | duns_number |
|---|---|
| kujawa.hieronim@turekkrupa.ne | -0-76453798 |
| karpi_j_zef_ski@kruszewski.org | -5-54785241 |

Next, we removed all spaces within email addresses:

=SUBSTITUTE([@email];" ";"")

The final version that has been loaded into SQL Server:

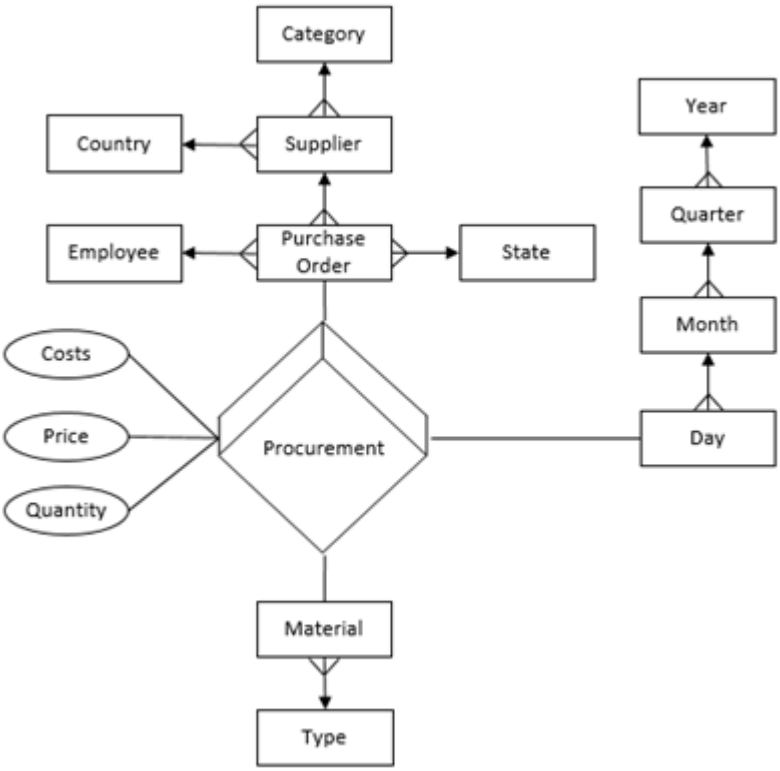
| id | name | description | street | zipcode | city | region | country | contactperson |
|---|---|---|---|---|---|---|---|---|
| 1 | Brouwer BV | Innovative methodical core | Emmapark 597 I | 5326 | NH Maas aan de I | Limburg | Netherlan | Thomas Koning |
| 2 | Goldkühle, Ne un | Re-contextualized holistic conglomeratic | Am Alten Schafstall 85b | 70119 | Ost Fabianland | Hamburg | Deutschla | Emmi Stahl |
| 3 | Santoro-Barbieri s | Algoritmo sincronizzata sensibile al cont | Rotonda Costa 2 | 41981 | Sesto Elga | Varese | Italy | Ausonio Palmieri |
| 4 | Henkel-Ullrich | Enterprise-wide radical capability | Meckhofer Feld 18 | 14490 | Jarosburg | Bremen | Deutschla | Brian Schmidt |
| 5 | Crona, Huels and I | Balanced logistical installation | 498 Rutherford Row | M1O-5I8 | Tremblaymouth | Manitoba | Canada | Vernie Breitenberg |
| 6 | Bins, Kuvalis and I | Balanced grid-enabled system engine | 191 Leannon Ville | U3R3R6 | East Savanna | Ontario | Canada | Deonte Conn |
| 7 | Gamez, Viera y Ra | instalación multimedia Extendido | Glorieta Gabriela Cardenas | 99566 | Almería | Región de Murcia | Spain | Sra. Gabriela Cuella |
| 8 | Leeuwen V.O.F. | Fundamental disintermediate product | Vriesplantsoen 888 | 6760-QK | Oud Annesluus | Limburg | Netherlan | Msc Eva Meijer |
| 9 | Wisozk-Tremblay | Mandatory national algorithm | 990 Patsy View | 89563-9582 | Lake Samson | North Dakota | USA | Abbigail Gutmann |
| 10 | Riviere EURL | Stand-alone system-worthy instruction s | 97 Quai de la Harpe | 94423 | Neuilly-sur-Seine | Basse-Normandie | France | M Clara Renault |
| 11 | Enríquez Soto e Hi | groupware cliente servidor Universal | Ronda Sergio Mondragón | 26010 | Huelva | Aragón | Spain | María Cristina Ocho |
| 12 | Baron EURL | Team-oriented zero defect core | 7 Rue de Presbourg | 88873 | Montreuil | Corse | France | Dr Mael Brunet |
| 13 | Grasso Group | Sistema aperto migliorata non-volatile | Rotonda Eufemia 9 | 63803 | Ferrara laziale | Firenze | Italy | Clea De Santis |
| 14 | Smith-Vandervort | Self-enabling asynchronous structure | 37707 Douglas Haven | 23066 | West Reannabor | South Carolina | USA | Colton Conn |
| 15 | Eplinius, Klapper | Reactive neutral instruction set | Solinger Str. 35a | 28176 | Tuchscheid | Sachsen | Deutschla | Fr. Erik Kock |
| 16 | Gorczany Inc | Ergonomic maximized challenge | 819 Noble Island | O7D-9X5 | Hammesport | Northwest Territories | Canada | Eunice Luettgen I |
| 17 | Malave Pabón Her | aprovechar no-volátil Reducido | Masía Carolina 19 | 86678 | Orense | Galicia | Spain | Sr. Marta Iglesias Zú |
| 18 | Seeger, Fuchs und | Optimized reciprocal hierarchy | Kaiserplatz 4 | 59239 | Ianburg | Hessen | Germany | Tamino Ahrenberg |
| 19 | Figura-Gehrig | Public-key next generation emulation | Ratiborer Str. 83c | 66728 | Alt Levi | Berlin | Deutschla | Dr. Maximilian Linn |
| 20 | Leśniak-Panek | Front-line optimal forecast | ul. Malec 185 | 48-181 | Pruchnik | Łódzkie | Poland | Rudolf Flak |
| 21 | Gdlokühle, Ne un | Advanced motivating contingency | Am Alten Schafstall 85b | 70119 | Ost Fabianland | Hamburg | Deutschla | Pani Apollo Fijałkow |
| 22 | Eoríquez Soto e Hi | Profound methodical conglomeration | Ronda Sergio Mondragón | 26010 | Huelva | Aragón | Spain | Hieronim Kujawa |
| 23 | HInkel-Uelrich | Open-source web-enabled adapter | Meckhofer Feld 18 | 1449 | Jarosburg | Bremen | Deutschla | Serafina Skowron |
| 24 | Goldkühle, Ne un | Enhanced non-volatile synergy | Am Alten Schafstal 85b | 70119 | Ost Fabianland | Hamburg | Deutschla | Józef Karpiński |
| 25 | Koning-Bosch | Function-based multimedia moderator | Timstraat 685 I | 4691 | IF Noord Rickdor | Limburg | Netherlands | |
| 26 | Clement et Bertra | Organized well-modulated orchestration | 6 Quai La Boétie | 75090 | Mérignac | Rhône-Alpes | France | Mlle Noémie Meyer |
| 27 | Lucas et Bourgeoi | Function-based incremental moderator | 95 Impasse de Montmoren | 78190 | Tourcoing | Nord-Pas-de-Calais | France | Benoit Quentin |
| 28 | Legros Inc | Streamlined coherent encryption | 8981 Anabel Causeway | 92756-1799 | Mozellefurt | Pennsylvania | USA | Amaya Ernser PhD |
| 29 | Valentini-Longo S | Conoscenza base virtuale metodica | Borgo Pacifico 67 | 95947 | Borgo Jarno | Massa-Carrara | Italy | |
| 30 | Marini-Basile SPA | Firmware assimilata radicale | Strada Pellegrino 9, Piano | 86800 | Settimo Sasha ve | Taranto | Italy | Emidio Rizzo |

| phone | fax | email | duns_number | state | category | currency | timestamp |
|---|---|---|---|---|---|---|---|
| 0693842208 | | thomas_koning@koning.com | 64-019-5775 | active | small | EUR | 2009-12-31 |
| (09470) 4318317 | | stahl_emmi@salowrittweg.net | 85-547-5241 | active | small | EUR | 2009-12-31 |
| +48 6507 879942 | +77 31 05549047 | palmieri_ausonio@barone.t | 29-767-9695 | active | small | EUR | 2009-12-31 |
| (04192) 1055307 | +49-205-6742714 | brian.shmidt@schnherr.de | 70-738-2722 | active | medium | EUR | 2009-12-31 |
| (318)684-0822 x0 | 1-119-625-6463 x5 | bretenberg_vernie@braunmcdermott.ca | 00-168-3179 | active | medium | CAD | 2009-12-31 |
| 1-563-212-2537 x | 649.545.8145 x40 | deonte.conn@kaulke.biz | 63-220-2865 | active | small | CAD | 2009-12-31 |
| (999) 999-999 | 917 612 160 | gabriela_valle_sra_cuellar@grego.info | 46-519-2738 | active | big | EUR | 2009-12-31 |
| 06 4505 0322 | (1968) 199430 | meijer.msc.eva@brink.com | 89-373-0638 | active | small | EUR | 2009-12-31 |
| (956) 682-2777 x5 | 425.515.7426 | gutmann_abbigail@okeefe.org | 42-800-3061 | active | medium | USD | 2009-12-31 |
| 0474268594 | +33 412671988 | m_renault_clara@le.inf | 09-566-4621 | active | small | EUR | 2009-12-31 |
| 957.617.303 | 930955855 | miranda.a.ochoa.crstina.mar@escobedo.info | 50-764-3798 | active | big | EUR | 2009-12-31 |
| 0553691753 | +33 580738595 | dr.brunet.mael@renard.org | 11-063-3566 | active | small | EUR | 2009-12-31 |
| +39 398 978 036 | +39 014 094 135 | de_santis_clea@bendettiroetti.it | 47-703-8105 | active | small | EUR | 2009-12-31 |
| 915-970-2973 | 270-985-3116 x51 | conn_colton@zieme.net | 04-799-8401 | active | small | USD | 2009-12-31 |
| +49-706-8067469 | +49-3798-840872 | fr.erik.kock@vahlensieckheinemann.info | 87-260-9381 | active | big | EUR | 2009-12-31 |
| 498-943-9299 x7 | 188-092-1817 | eunice.i.lettgen@langosh.net | 76-112-4722 | active | medium | CAD | 2009-12-31 |
| 954-712-373 | 955161954 | iga.sr.iglesias.marta.z@hinojosa.info | 55-093-3416 | active | big | EUR | 2009-12-31 |
| (01858) 0159354 | +49-590-0550446 | tamino.ahrenerg@brner.de | 32-382-4361 | active | big | EUR | 2009-12-31 |
| +49-8849-58951328 | | dr_linnenbaum_maximilian@lenksteinert.com | 22-363-5762 | active | medium | EUR | 2009-12-31 |
| 25-527-21-08 | | rudolf_flak@liwiskimasowski.com.pl | 03-604-1918 | active | big | PLN | 2009-12-31 |
| 94-571-63-67 | 86-672-25-40 | fija_pani_kowski_apollo@bednarczyk.net | 85-527-5441 | active | small | EUR | 2009-12-31 |
| 34-425-89-70 | 15-148-70-53 | kujawa.hieronim@turekkrupa.net | 70-76453798 | active | big | EUR | 2009-12-31 |
| 13-469-34-33 | 25-729-34-32 | skowron_serafina@baraski.net | 70-238-7722 | active | medium | EUR | 2009-12-31 |
| 83-522-49-08 | 91-785-41-11 | karpi_j_zef_ski@kruszewski.org | 50-54785241 | active | small | EUR | 2009-12-31 |
| (2819) 189851 | 0614469077 | broek.jr.isa@konig.com | 26-064-8717 | active | small | EUR | 2010-01-01 |
| 0721565778 | +33 159272905 | meyer_miemlle_no@meyer.org | 52-637-3689 | active | small | EUR | 2010-01-01 |
| +33 167704521 | 0693329641 | benoit_quentin@faure.fr | 01-149-5137 | active | medium | EUR | 2010-01-01 |
| 320-815-3951 | 667-615-5330 x57 | ernser.phd.amaya@carrollhintzname | 62-932-1721 | active | small | USD | 2010-01-01 |
| +67 5164 204559 | +50 921 94399381 | furio.greco@reco.org | 56-968-2563 | active | big | EUR | 2010-03-01 |

# Multidimensional design (conceptual design)

## ME/R diagram

We started the planning of the future data mart by creating a Multidimensional Entity Relationship Diagram which served as a blueprint for further logical design in the form of a star schema. This model shows that a cube should be created, centered around Procurement.

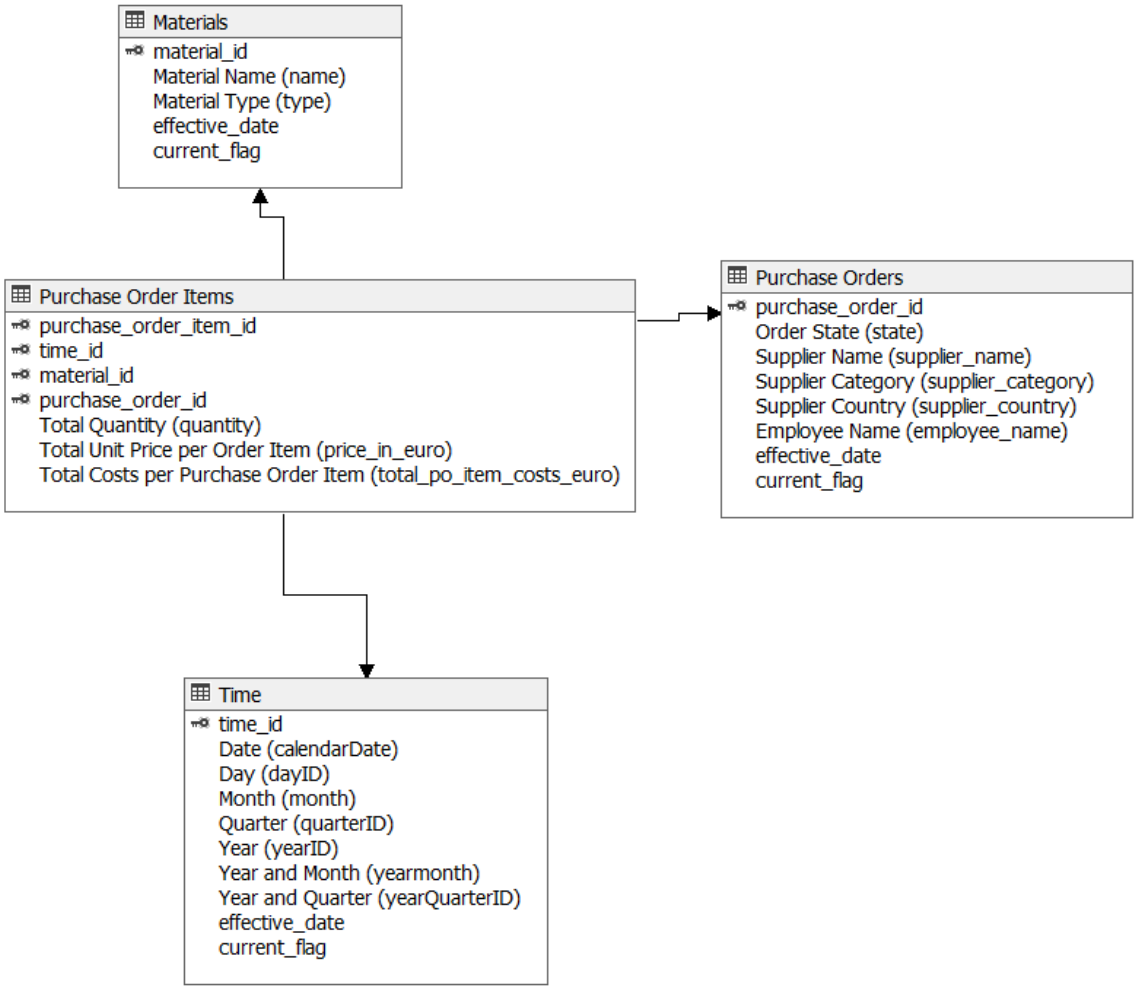



The planned data mart should have the following characteristics:

- **subject-related:** the focus of the analysis are the business processes in the Procurement department
- **integrated:** the data mart should integrate attributes from different tables and store them in a single data mart
- **non-volatile:** the data mart implements Slowly Changing Dimension strategy two so that changes to the dimension tables can be later imported into dimension tables as new rows
- **time-variant:** the data mart stores mainly historical data. New data can be regularly loaded into data mart by using a combination of the snapshot and queryable refresh strategy, which can be implemented as a stored procedure in SQL Server.

## Star Schema

By using the following logical design of the star schema, the team defined the *Purchase Order Items* as the fact table, and *Materials*, *Time* as well as *Purchase Orders* as dimensions. At this point, the Slowly Changing Dimension 2 strategy has been implemented to account for the later changes in the dimension tables. These changes should be imported as new rows, and the binary attribute *current_flag* should be then adjusted accordingly. The expressions in brackets indicate the original database column name since those were given a friendly name so that the end user can intuitively understand their meaning.



# Proof-of-concept implementation - Microsoft Technologies

## ETL process

Even though we already cleaned the relevant data, there have still been some preprocessing steps that needed to be tackled before creating the Data Mart. Within the scope of ETL process, the data has been extracted from SuperX OLTP system, then all required transformations such as joining related tables, changing column labels and leaving out unnecessary information such as ex. fax or phone number have been performed, and finally, we were able to load the transformed data into reporting database *DataMart_NewSuperX* that serves as a data mart, i.e., a small-scale data warehouse.

```
/* data imported into SQL SERVER*/
USE DataMart_NewSuperX;
GO

-- Dim Time
CREATE TABLE DimTime (
time_id int NOT NULL CONSTRAINT [pkDimTime] PRIMARY KEY,
calendarDate Date NOT NULL,
dayID int NOT NULL,
month nvarchar(50) NOT NULL,
quarterID int NOT NULL,
yearID int NOT NULL,
yearmonth nvarchar(7) NOT NULL,
yearQuarterID nvarchar(7) NOT NULL,
effective_date Date NOT NULL,
current_flag bit NOT NULL
);
go

-- Dim Material
CREATE TABLE DimMaterial (
material_id int NOT NULL CONSTRAINT [pkDimMaterial] PRIMARY KEY,
name nvarchar(50) NOT NULL,
type nvarchar(50) NOT NULL,
effective_date Date NOT NULL,
current_flag bit NOT NULL
);
go
```

```
-- Dim Purchase Order
CREATE TABLE DimPurchaseOrder (
purchase_order_id int NOT NULL CONSTRAINT [pkDimPurchaseOrder] PRIMARY KEY,
state nvarchar(50) NOT NULL,
supplier_name nvarchar(50) NOT NULL,
supplier_category nvarchar(50) NOT NULL,
supplier_country nvarchar(50) NOT NULL,
employee_name nvarchar(50) NOT NULL,
effective_date Date NOT NULL,
current_flag bit NOT NULL
);
go

-- Fact table Purchase Order Items
CREATE TABLE FactPurchaseOrderItems (
    purchase_order_item_id int NOT NULL, --PK
    time_id int, -- PK, FK1
    material_id int, -- PK, FK2
    purchase_order_id int, -- PK, FK3
    quantity int,
    price_in_euro money,
    total_po_item_costs_euro money
    CONSTRAINT [pkFactPurchaseOrderItems] PRIMARY KEY
    (purchase_order_item_id, time_id, material_id, purchase_order_id)
);
go

-- FK constraints
ALTER TABLE dbo.FactPurchaseOrderItems
ADD CONSTRAINT fkFactToDimTime
FOREIGN KEY (time_id) REFERENCES dbo.DimTime (time_id);
go

ALTER TABLE dbo.FactPurchaseOrderItems
ADD CONSTRAINT fkFactToDimMaterial
FOREIGN KEY (material_id) REFERENCES dbo.DimMaterial (material_id);
go

ALTER TABLE dbo.FactPurchaseOrderItems
ADD CONSTRAINT fkFactToDimPurchaseOrder
FOREIGN KEY (purchase_order_id) REFERENCES dbo.DimPurchaseOrder (purchase_order_id);
go
```

```
-- INSERT DATA
INSERT INTO [DataMart_NewSuperX].[dbo].[DimTime]
SELECT distinct [time_id] = concat(YEAR(timestamp), MONTH(timestamp), DAY(timestamp)),
 [calendarDate] = cast(timestamp as date),
 [dayID] = DAY(timestamp),
 [month] = DATENAME(month,timestamp),
[quarterID] = DATEPART(quarter, timestamp),
[yearID] = YEAR(timestamp),
[yearmonth] = concat(YEAR(timestamp),'-', MONTH(timestamp)),
[yearQuarterID] = concat(YEAR(timestamp),'-', DATEPART(quarter, timestamp)),
[effective_date] = cast(timestamp as date),
[current_flag] = 1
from NewSuperX.dbo.purchase_order_items;
go
```

```
INSERT INTO [DataMart_NewSuperX].[dbo].[DimMaterial]
SELECT DISTINCT material_id, name, type,
[effective_date] = cast(NewSuperX.dbo.materials.timestamp as date),
[current_flag] = 1
  FROM NewSuperX.dbo.purchase_order_items
  join NewSuperX.dbo.materials ON purchase_order_items.material_id = materials.id;
go

INSERT INTO [DataMart_NewSuperX].[dbo].[DimPurchaseOrder]
SELECT DISTINCT purchase_order_id, [state] = purchase_orders.state,
[supplier_name] = suppliers.name,
[supplier_category] = CASE WHEN category='smal' THEN 'small' ELSE category END,
[supplier_country] = CASE WHEN right(address, CHARINDEX(' ', REVERSE(address))-1)='U.S.A.' THEN 'USA'
WHEN right(address, CHARINDEX(' ', REVERSE(address))-1)='Deutschland' THEN 'Germany'
WHEN right(address, CHARINDEX(' ', REVERSE(address))-1)='Kingdom' THEN 'United Kingdom'
ELSE right(address, CHARINDEX(' ', REVERSE(address))-1) END,
[employee_name] = concat(employees.firstname, ' ', employees.lastname),
[effective_date] = cast(NewSuperX.dbo.purchase_orders.timestamp as date),
[current_flag] = 1
  FROM NewSuperX.dbo.purchase_order_items
  join NewSuperX.dbo.purchase_orders ON purchase_order_items.purchase_order_id = purchase_orders.id
  join NewSuperX.dbo.employees ON purchase_orders.employee_id = employees.id
  join NewSuperX.dbo.suppliers ON purchase_orders.supplier_id = suppliers.id;
go
```

```
CREATE OR ALTER VIEW cleaned_po_items
as with notnullcurrencies (purchase_order_id, nncurrency)
   as (SELECT distinct purchase_order_id, currency as nncurrency
   from NewSuperX.dbo.purchase_order_items where currency is not null)
   SELECT [purchase_order_item_id] = id,
  [time_id] = concat(YEAR(timestamp), MONTH(timestamp), DAY(timestamp)),
  material_id, [purchase_order_id] = purchase_order_items.purchase_order_id, quantity,
   case when purchase_order_items.currency is null then nncurrency
   else purchase_order_items.currency end as currency
     FROM NewSuperX.dbo.purchase_order_items
     join notnullcurrencies on notnullcurrencies.purchase_order_id = purchase_order_items.purchase_order_id;
go

INSERT INTO [DataMart_NewSuperX].[dbo].[FactPurchaseOrderItems]
SELECT c.purchase_order_item_id, c.time_id, c.material_id, c.purchase_order_id, c.quantity,
[price_in_euro] =
  CASE WHEN c.currency='CAD' THEN price*0.64
  WHEN c.currency='USD' THEN price*0.81
  WHEN c.currency='PLN' THEN price*0.24
  WHEN c.currency='GBP' THEN price*1.12
  ELSE price END,
  [total_po_item_costs_euro] = c.quantity*CASE WHEN c.currency='CAD' THEN price*0.64
  WHEN c.currency='USD' THEN price*0.81
  WHEN c.currency='PLN' THEN price*0.24
  WHEN c.currency='GBP' THEN price*1.12
  ELSE price END
FROM cleaned_po_items c
join [NewSuperX].[dbo].[purchase_order_items] on purchase_order_items.id = c.purchase_order_item_id;
```

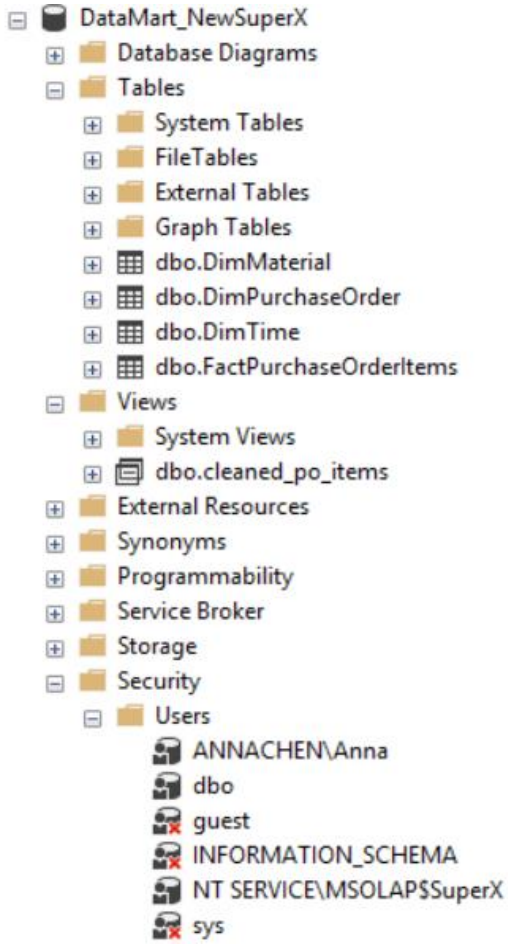
In the script above, we performed the following transformations:

- **data definition:** the creation of the data structure for the fact and dimension tables, including adding columns for Slowly Changing Dimensions and referential integrity constraints to indicate relationships between tables,
- **surrogate keys:** ex. *time_id* comprised of year, month and date value
- **table joins:** the end user does not want to see only key columns such as *material_id*, therefore it is necessary to perform a join to see ex. *material name* instead.
- **data type casts:** conversion from ex. TIMESTAMP to DATE column
- **corrections & data cleansing:** “smal” → “small”, “Deutschland” → “Germany”, “Kingdom” → “United Kingdom”
- **standardization:** “U.S.A.” → “USA”, different currencies → recalculated to a single EURO currency to enable correct summarization in the cube later on,

- **handling of missing values:** the column currency had many missing values in the single purchase order items, but we were able to extrapolate them from other purchase order items, belonging to the same purchase order, by using a view that served as a lookup table.

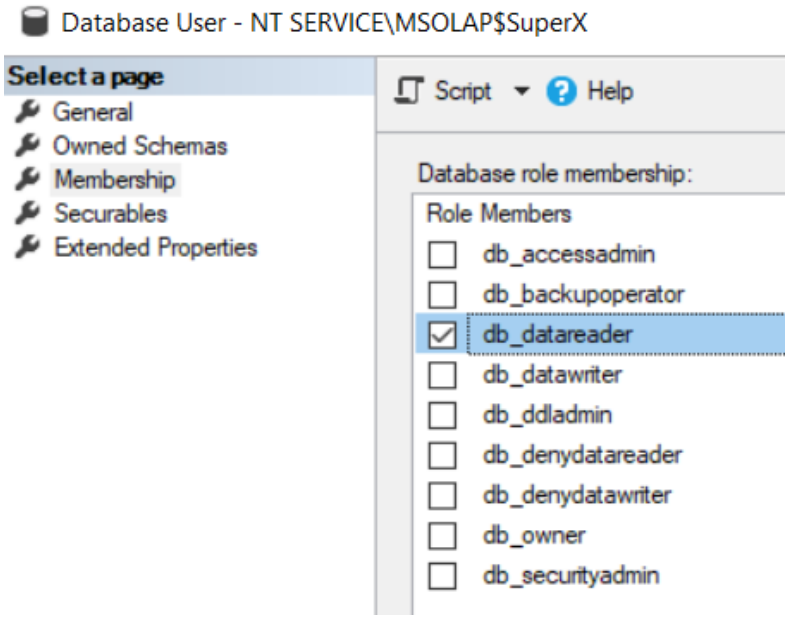
## Final Cube Design

After all necessary transformations and loading the selected data into database *DataMart_NewSuperX*, the multi-dimensional implementation has been initiated.




## Challenges

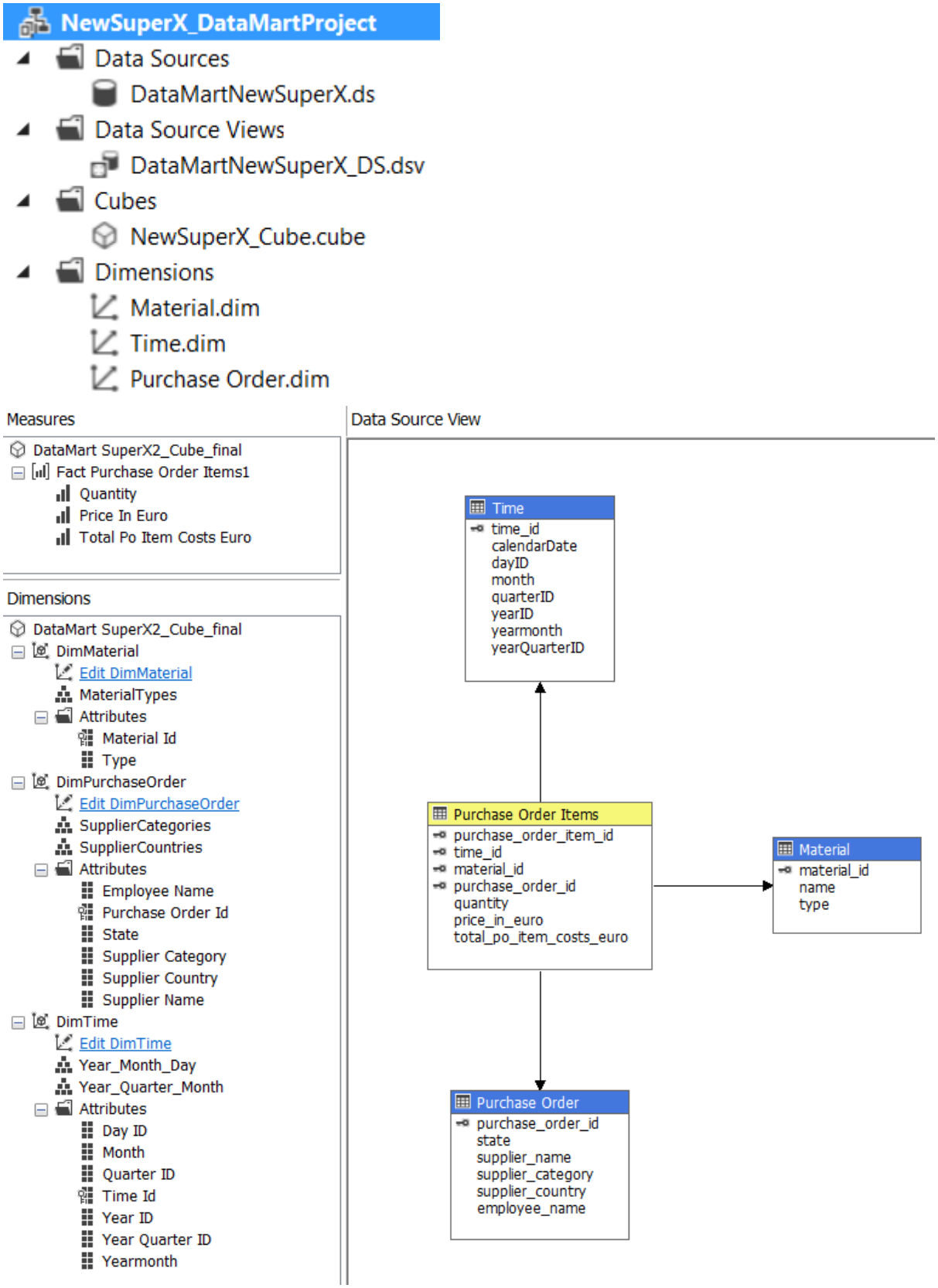
The cube was planned to be implemented with Microsoft SQL Server Analysis Services and a particularly important step for the further process was the creation of a special user **NT SERVICE\MSOLAP$SuperX**, which should have only reading permission to the multidimensional database. This turned out to be a crucial step for later deployment and caused the team much investigation to figure out the solution to the problem in order to process the cube.




## Multidimensional implementation

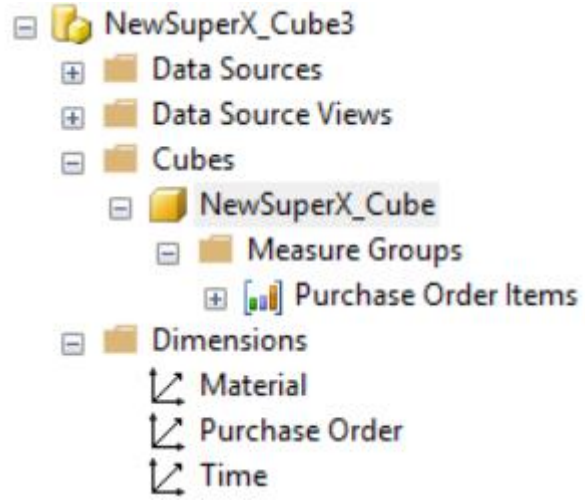
Next, a multidimensional project has been created in Analysis Services. The whole process was comprised of the following steps:

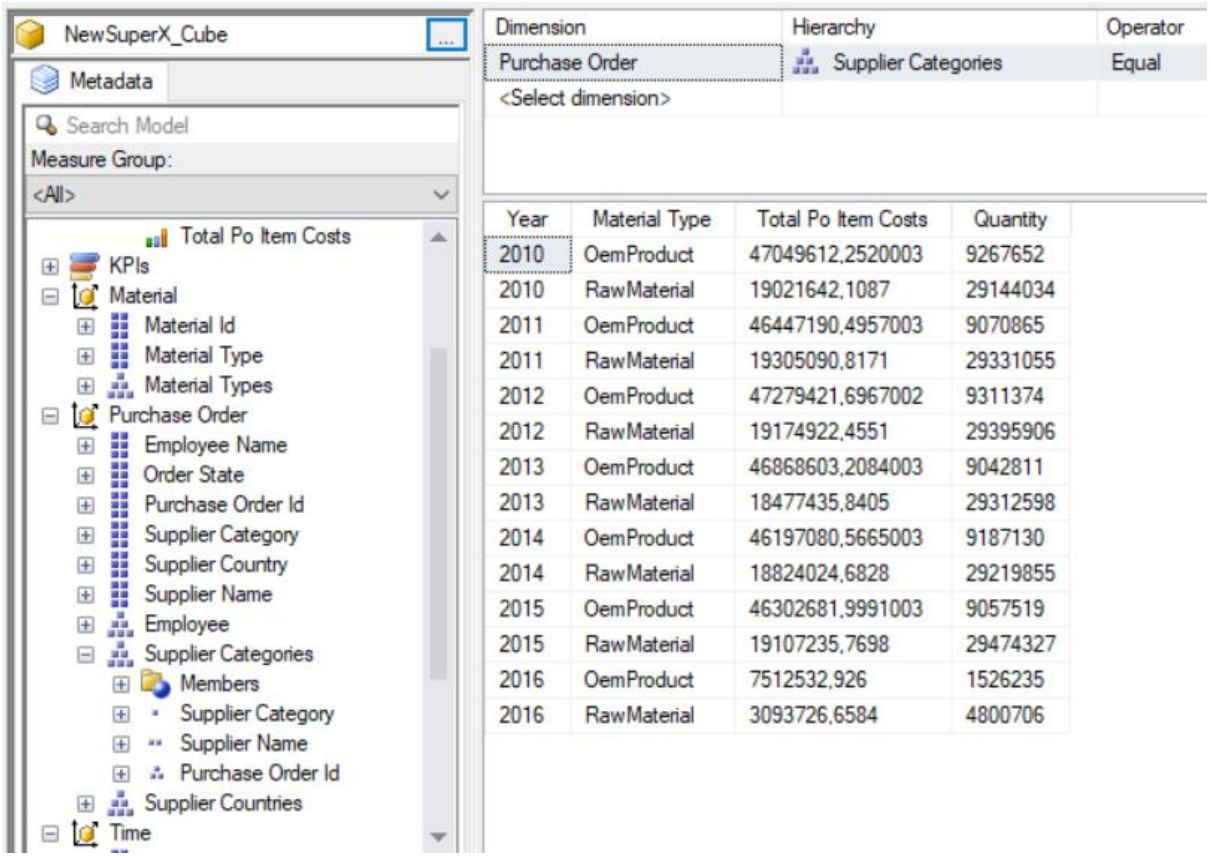
- **Data Source:** creating a connection to the data source, i.e., to the Data Mart
- **Data Source View:** determining the fact and dimension tables, as well as specifying friendly names for the users and adjusting the format of measures, ex. to display them in currency units.
- **Cube:** specifying dimensional attributes as well as their hierarchies, adjusting the measures ex. to display either SUM, COUNT or AVERAGE. At this point the team noticed a drawback of the Microsoft solution, as it required determining all aggregate functions needed for the analysis in advance. It speeds up the eventual cube's calculations performance, but at the same time it is rather inflexible, because users could change their mind and wish instead of SUM to see, for instance, the AVERAGE value and this would require anew processing of the fact table.


NewSuperX_DataMartProject
Data Sources
DataMartNewSuperX.ds
Data Source Views
DataMartNewSuperX_DS.dsv
Cubes
NewSuperX_Cube.cube
Dimensions
Material.dim
Time.dim
Purchase Order.dim
Measures
DataMart SuperX2_Cube_final
Fact Purchase Order Items1
Quantity
Price In Euro
Total Po Item Costs Euro
Dimensions
DataMart SuperX2_Cube_final
DimMaterial
Edit DimMaterial
MaterialTypes
Attributes
Material Id
Type
DimPurchaseOrder
Edit DimPurchaseOrder
SupplierCategories
SupplierCountries
Attributes
Employee Name
Purchase Order Id
State
Supplier Category
Supplier Country
Supplier Name
DimTime
Edit DimTime
Year_Month_Day
Year_Quarter_Month
Attributes
Day ID
Month
Quarter ID
Time Id
Year ID
Year Quarter ID
Yearmonth
Data Source View
Time
time_id
calendarDate
dayID
month
quarterID
yearID
yearmonth
yearQuarterID
Purchase Order Items
purchase_order_item_id
time_id
material_id
purchase_order_id
quantity
price_in_euro
total_po_item_costs_euro
Material
material_id
name
type
Purchase Order
purchase_order_id
state
supplier_name
supplier_category
supplier_country
employee_name

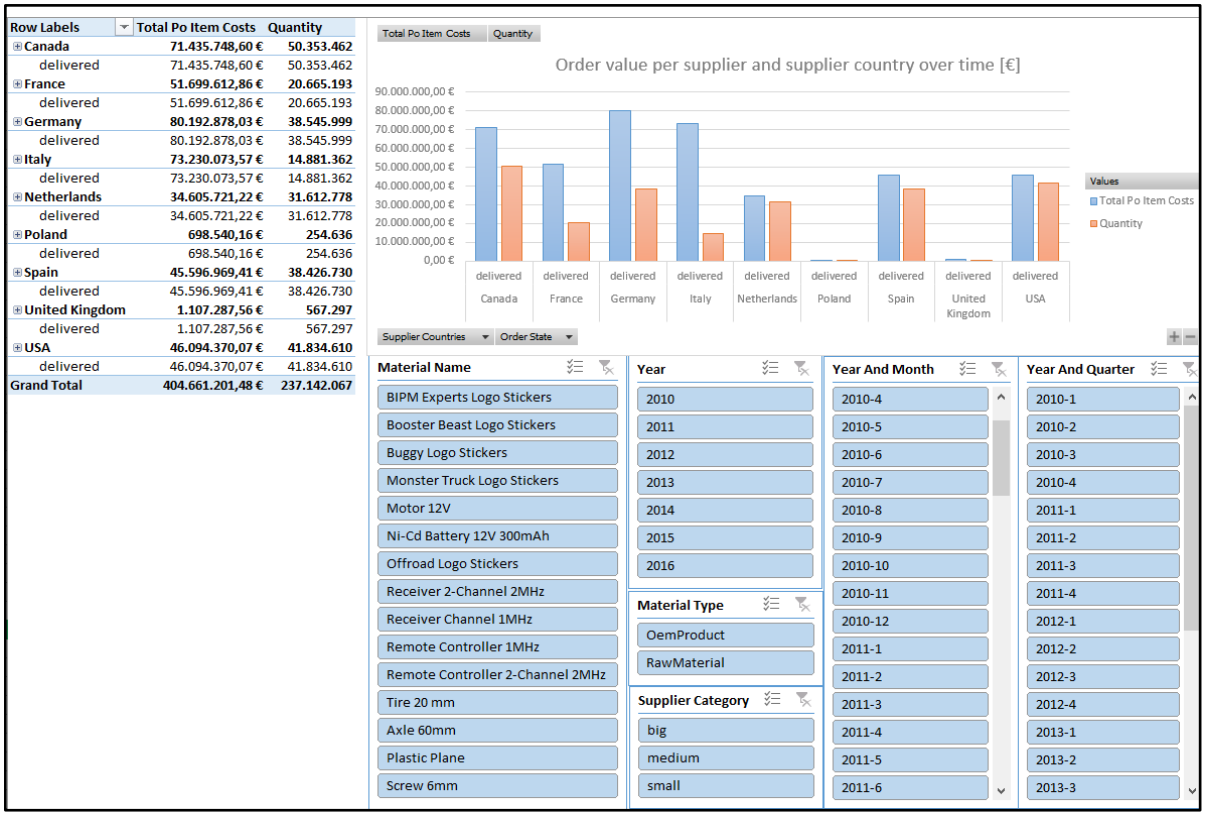
## Dashboard implementation

The team first created the cube by using Microsoft SQL Server Analysis Services, as described in the previous section. The resultant cube, after deployment and processing, looks as follows:



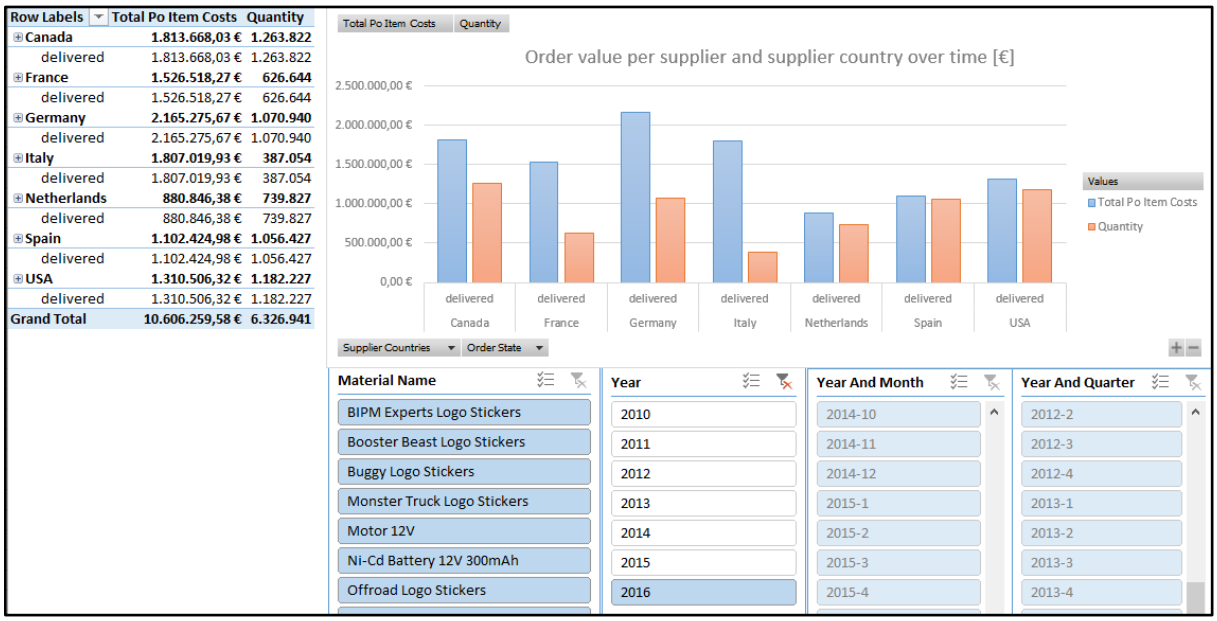




Based on the implemented cube, the team created a dashboard by using Microsoft Excel connected to the cube in Analysis Services.




In the dashboard above, there is a summary of supplier statistics with regard to the purchase order volume [in units of products] and value [in €], as well as delivery reliability, indicated by the column Order State (ex. *delivered*). The advantage of this view is the ability to first look at the highly aggregated data, for instance showing only the total purchase order item costs and quantity, and then drill down, slice and dice the aggregated data to find specific information that someone is interested in, and roll up again if needed. The slicer presented below the bar chart, enable to filter the data by a specific condition quickly, ex. to see only orders for a specific product from some individual supplier at a desired point of time. To illustrate this, we want to see only data for 2016 in order to compare the transactions made with German vendors to suppliers from other countries, and later to drill down into individual transactions undertaken only with German suppliers:

- **aggregated** **overview:**




- **detailed overview** (in which every row represents a single purchase order item, indicating the fact table grain, i.e., the lowest level of granularity, obtained by double-clicking the *Total Po Item Costs* corresponding to the German suppliers):

Data returned for Total Po Item Costs, Germany, All - All - 2016 - All (First 1000 rows).

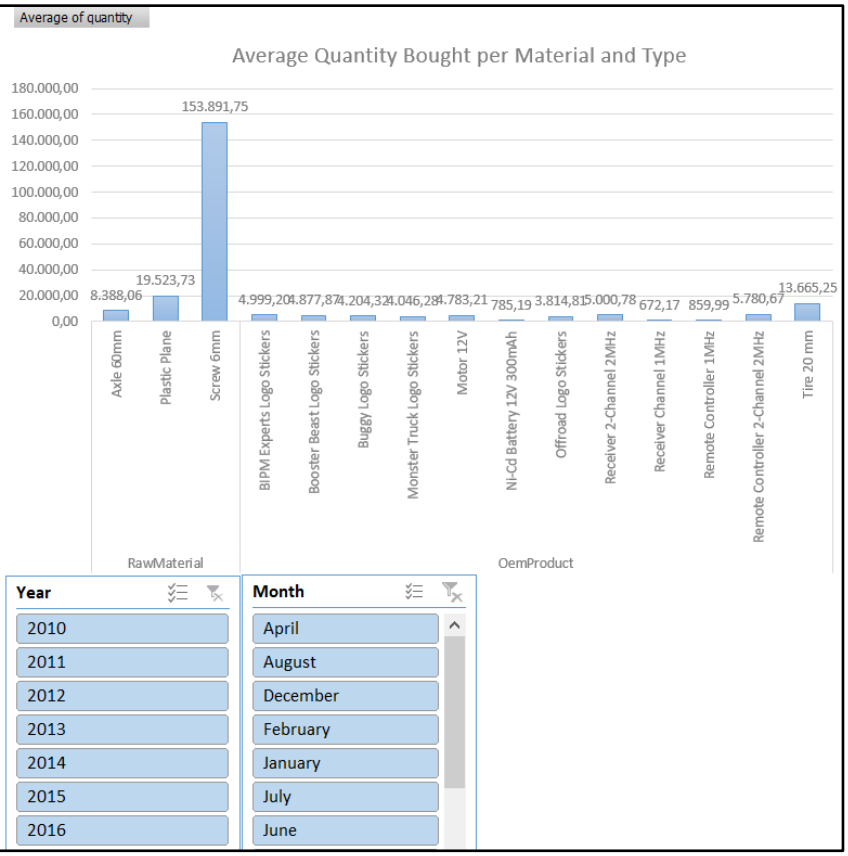
| [Purchase Order Items].[$Time.Time Id] | [Purchase Order Items].[$Purchase Order.Purchase Order Id] | [Purchase Order Items].[$Material.Material Id] | [Purchase Order Items].[Quantity] | [Purchase Order Items].[Price] |
|---|---|---|---|---|
| 2016-01-11 | 3884 | Screw 6mm | 168751 | 0,12 |
| 2016-01-11 | 3884 | Axle 60mm | 4973 | 2,12 |
| 2016-01-11 | 3884 | Remote Controller 2-Channel 2MHz | 4095 | 22,72 |
| 2016-01-11 | 3884 | Remote Controller 1MHz | 1089 | 11,34 |
| 2016-01-11 | 3884 | Motor 12V | 6807 | 8,32 |
| 2016-01-11 | 3884 | Receiver Channel 1MHz | 743 | 8,2 |
| 2016-01-11 | 3884 | Buggy Logo Stickers | 2630 | 0,48 |
| 2016-01-11 | 3884 | Booster Beast Logo Stickers | 4680 | 0,46 |
| 2016-01-11 | 3884 | BIPM Experts Logo Stickers | 5708 | 0,89 |
| 2016-01-11 | 3886 | Plastic Plane | 18771 | 4,27 |
| 2016-01-11 | 3886 | Axle 60mm | 5205 | 2,23 |
| 2016-01-11 | 3886 | Remote Controller 2-Channel 2MHz | 5517 | 23,05 |
| 2016-01-11 | 3886 | Remote Controller 1MHz | 943 | 10,96 |
| 2016-01-11 | 3886 | Motor 12V | 3149 | 7,78 |
| 2016-01-11 | 3886 | Receiver 2-Channel 2MHz | 5388 | 15,06 |
| 2016-01-11 | 3886 | Receiver Channel 1MHz | 420 | 8,06 |
| 2016-01-11 | 3886 | Monster Truck Logo Stickers | 3511 | 0,5 |
| 2016-01-11 | 3897 | Tire 20 mm | 19460 | 1,18 |
| 2016-01-11 | 3897 | Offroad Logo Stickers | 4045 | 0,49 |
| 2016-01-11 | 3900 | Ni-Cd Battery 12V 300mAh | 486 | 2,3 |
| 2016-01-11 | 3908 | Buggy Logo Stickers | 6101 | 0,46 |
| 2016-01-12 | 3911 | Screw 6mm | 168751 | 0,12 |
| 2016-01-12 | 3911 | Axle 60mm | 4973 | 2,12 |
| 2016-01-12 | 3911 | Remote Controller 2-Channel 2MHz | 4095 | 22,72 |
| 2016-01-12 | 3911 | Remote Controller 1MHz | 1089 | 11,34 |
| 2016-01-12 | 3911 | Motor 12V | 6807 | 8,32 |
| 2016-01-12 | 3911 | Buggy Logo Stickers | 2630 | 0,48 |

The dashboards were created to satisfy the business requirements, defined at the beginning of this report. In order to address them, the following section provides a short explanation of how the designed cube can answer those questions.

## Answers to business questions

**Q1: What is the average quantity bought for each *Material* and *Material Type* per Year and Month?** [measured in units of products]

| Row Labels | Average of quantity |
|---|---|
| RawMaterial | 58.245,80 |
| Axle 60mm | 8.388,06 |
| Plastic Plane | 19.523,73 |
| Screw 6mm | 153.891,75 |
| OemProduct | 4.565,67 |
| BIPM Experts Logo Stickers | 4.999,20 |
| Booster Beast Logo Stickers | 4.877,87 |
| Buggy Logo Stickers | 4.204,32 |
| Monster Truck Logo Stickers | 4.046,28 |
| Motor 12V | 4.783,21 |
| Ni-Cd Battery 12V 300mAh | 785,19 |
| Offroad Logo Stickers | 3.814,81 |
| Receiver 2-Channel 2MHz | 5.000,78 |
| Receiver Channel 1MHz | 672,17 |
| Remote Controller 1MHz | 859,99 |
| Remote Controller 2-Channel 2MH | 5.780,67 |
| Tire 20 mm | 13.665,25 |
| Grand Total | 15.330,15 |

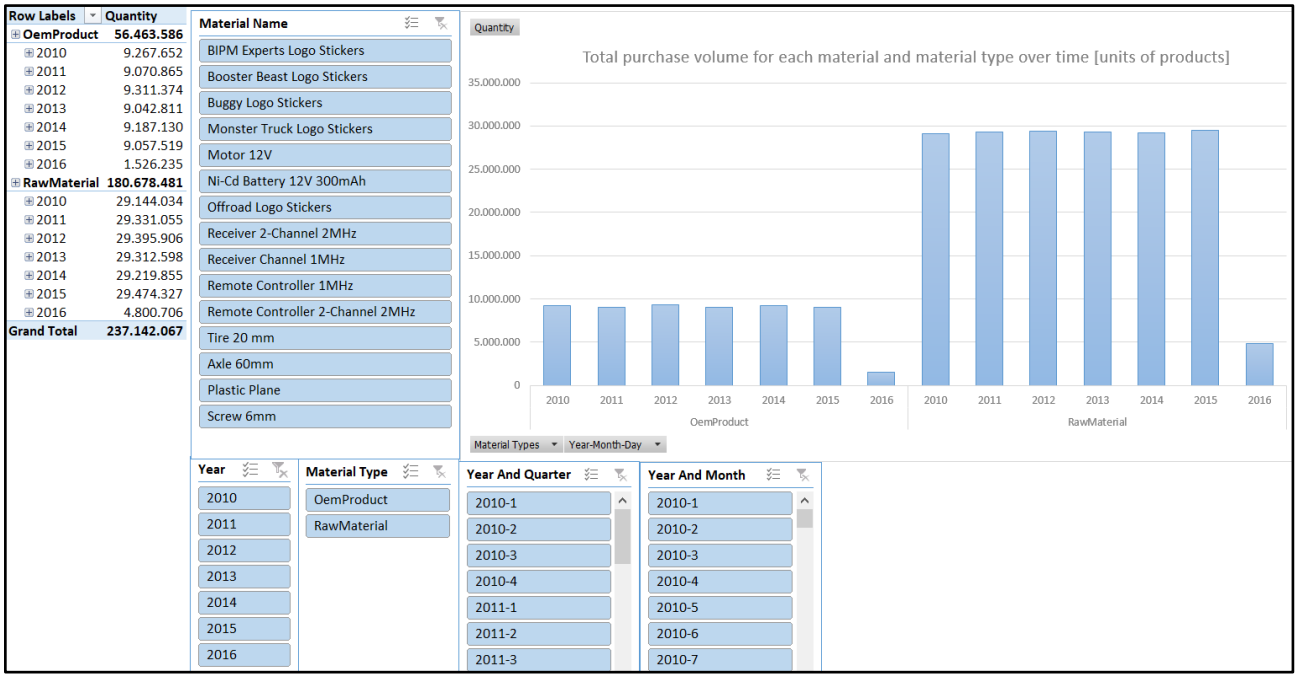



**Findings:**

- *Screw 6mm* is the most commonly bought material with an average quantity of 154 thousand units purchased each year.
- If one filters this data for each year, the distribution is nearly identical, implying a steady purchase volume per each material and material type every year.

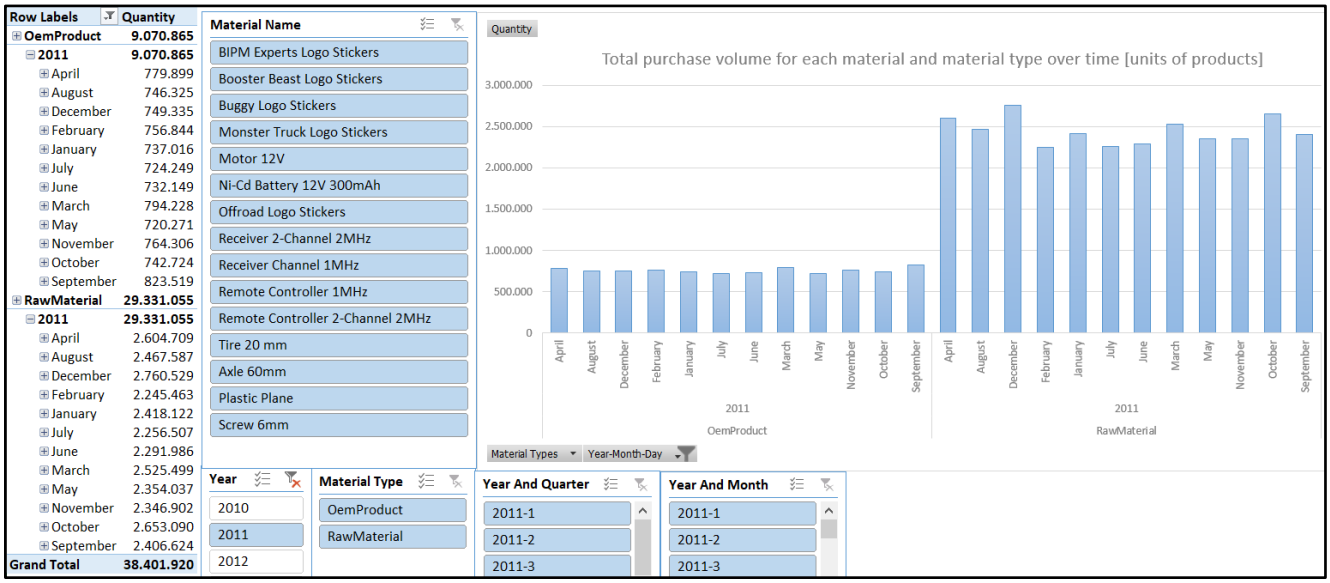
**Q2: What is the total quantity bought for each *Material* and *Material Type* over time?**

[purchase volume measured in units of products]



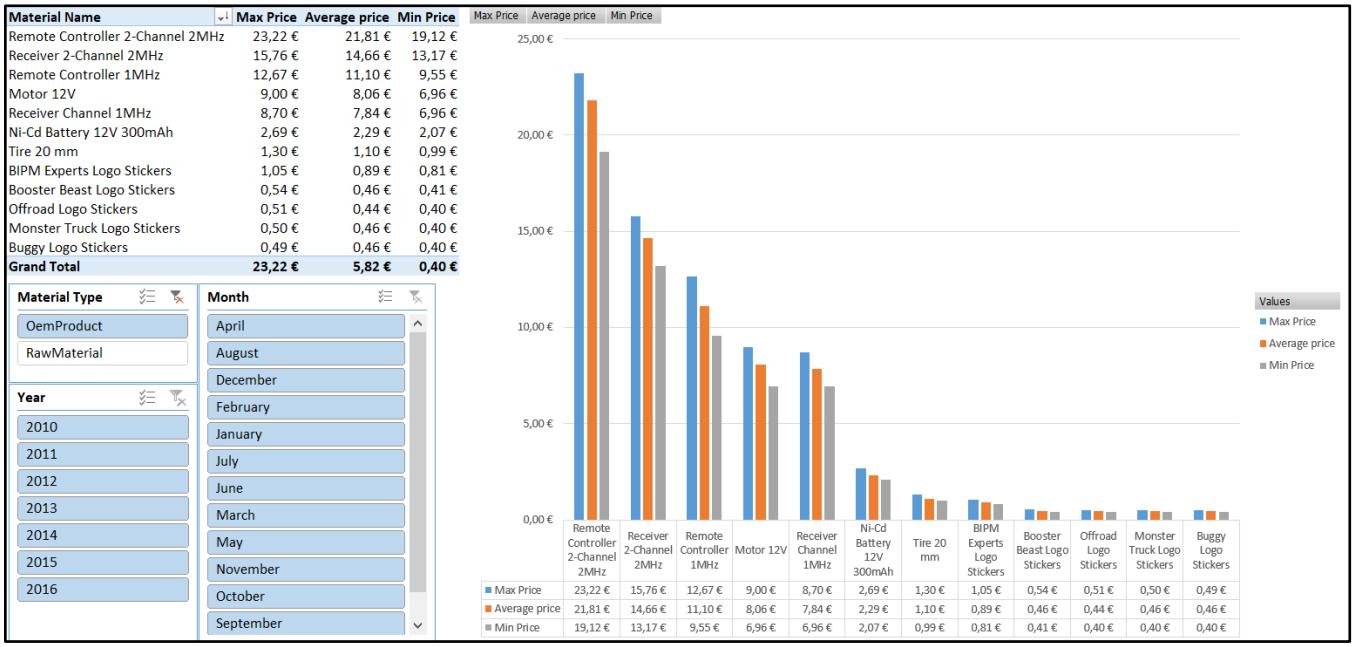

**Findings:**

- Raw materials are ordered in higher quantities than OEM products
- The distribution for years 2010-2015 is nearly identical → around 9 million units of OEM products bought every year and around 29 million units of raw materials, implying that the company is not growing, but maintains a steady level of items ordered every year.
- However, there are some slight seasonal differences in specific months of the year:



**Q3: What is the highest, lowest and average price per *Material* and *Material Type* per Month and Year?** [measured in €]

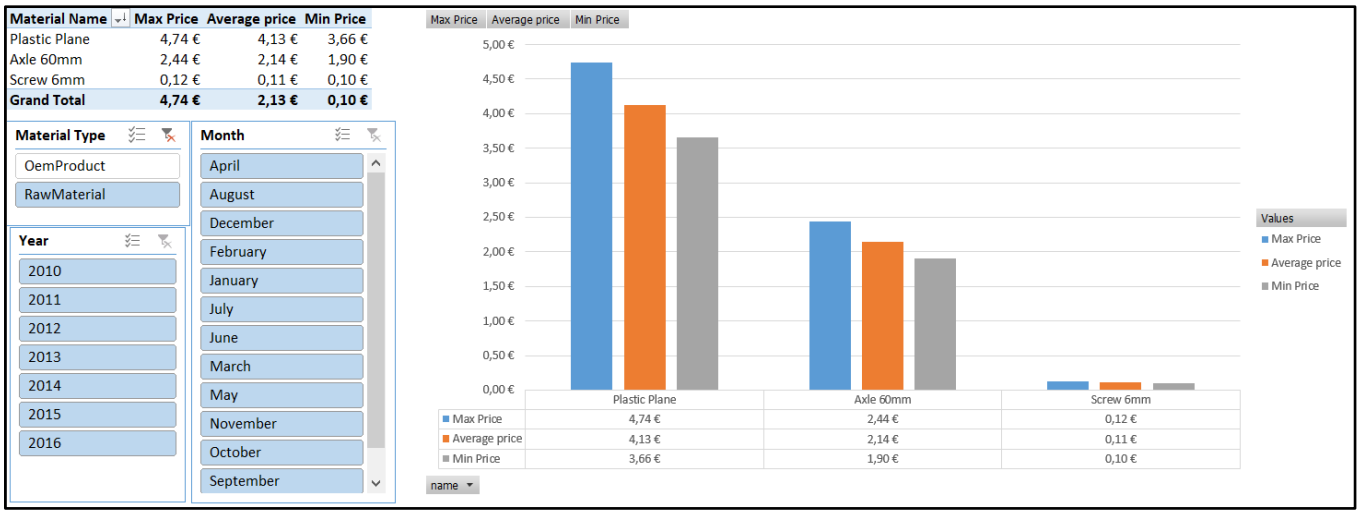
## 1. Material Type → OEM Products



| Material Name | Max Price | Average price | Min Price |
|---|---|---|---|
| Remote Controller 2-Channel 2MHz | 23,22 € | 21,81 € | 19,12 € |
| Receiver 2-Channel 2MHz | 15,76 € | 14,66 € | 13,17 € |
| Remote Controller 1MHz | 12,67 € | 11,10 € | 9,55 € |
| Motor 12V | 9,00 € | 8,06 € | 6,96 € |
| Receiver Channel 1MHz | 8,70 € | 7,84 € | 6,96 € |
| Ni-Cd Battery 12V 300mAh | 2,69 € | 2,29 € | 2,07 € |
| Tire 20 mm | 1,30 € | 1,10 € | 0,99 € |
| BIPM Experts Logo Stickers | 1,05 € | 0,89 € | 0,81 € |
| Booster Beast Logo Stickers | 0,54 € | 0,46 € | 0,41 € |
| Offroad Logo Stickers | 0,51 € | 0,44 € | 0,40 € |
| Monster Truck Logo Stickers | 0,50 € | 0,46 € | 0,40 € |
| Buggy Logo Stickers | 0,49 € | 0,46 € | 0,40 € |
| **Grand Total** | **23,22 €** | **5,82 €** | **0,40 €** |



## 2. Material Type → Raw Materials



| Material Name | Max Price | Average price | Min Price |
|---|---|---|---|
| Plastic Plane | 4,74 € | 4,13 € | 3,66 € |
| Axle 60mm | 2,44 € | 2,14 € | 1,90 € |
| Screw 6mm | 0,12 € | 0,11 € | 0,10 € |
| **Grand Total** | **4,74 €** | **2,13 €** | **0,10 €** |



**Findings:**

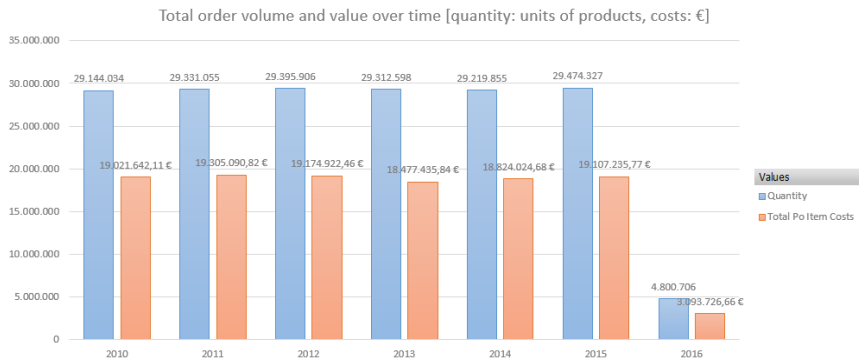
- "Remote Controller 2-Channel 2MHz" turned out to be the most expensive OEM Product
- "Plastic Plane" is the most expensive Raw Material
- Even though there are some significant deviations between MIN, MAX and AVERAGE price, the prices are stable over time, which can be observed by filtering over time.

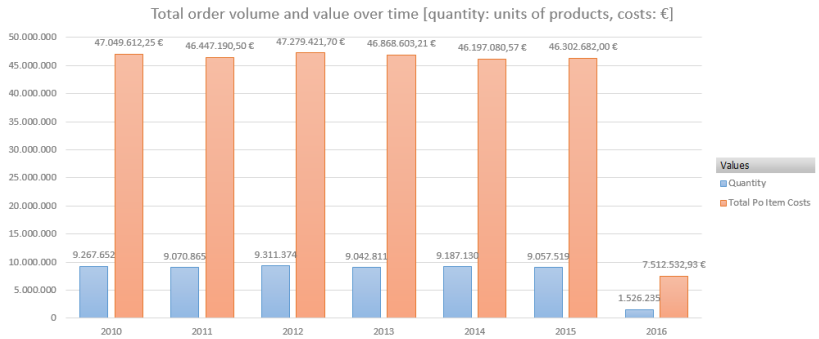
**Q4: What is the total order volume** [measured in units of products] **and value** [measured in €] **per *Month, Quarter* and *Year*?**


**Findings:**

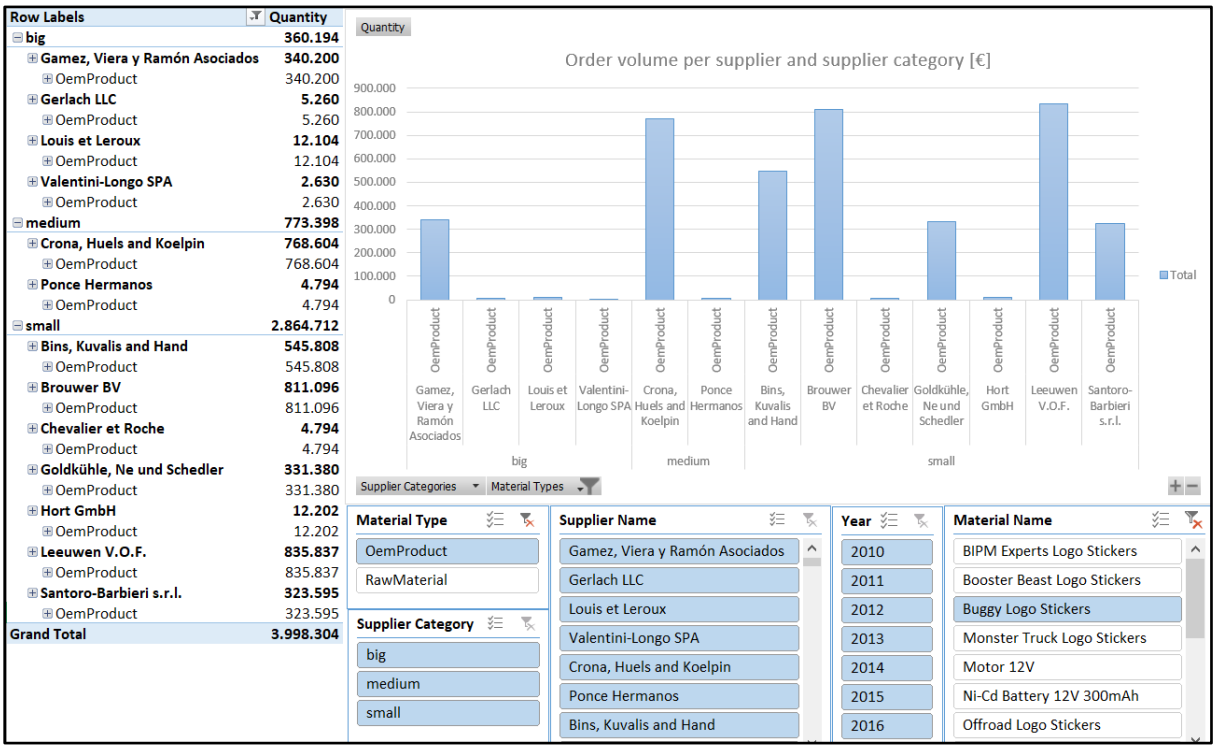
- The distribution of quantity and total costs of bought products is steady over time.
- While raw materials are purchased in larger quantities than OEM products, OEM products are more expensive, resulting in higher overall purchase order costs.

**Raw Materials (upper figure) vs. OEM products (lower figure):**



**Q5: What is the total amount of materials delivered by specific Suppliers and Supplier categories filtered by specific Materials?** [measured in units of products]

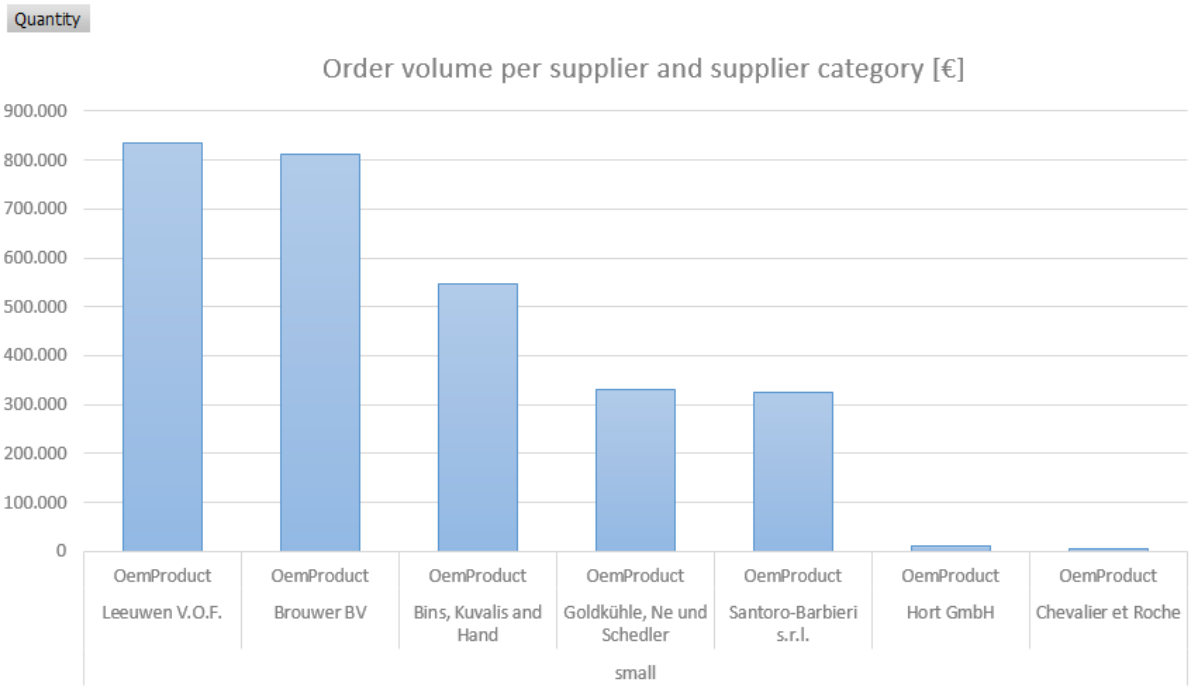


| Row Labels | Quantity |
|---|---|
| big | 360.194 |
| Gamez, Viera y Ramón Asociados | 340.200 |
| OemProduct | 340.200 |
| Gerlach LLC | 5.260 |
| OemProduct | 5.260 |
| Louis et Leroux | 12.104 |
| OemProduct | 12.104 |
| Valentini-Longo SPA | 2.630 |
| OemProduct | 2.630 |
| medium | 773.398 |
| Crona, Huels and Koelpin | 768.604 |
| OemProduct | 768.604 |
| Ponce Hermanos | 4.794 |
| OemProduct | 4.794 |
| small | 2.864.712 |
| Bins, Kuvalis and Hand | 545.808 |
| OemProduct | 545.808 |
| Brouwer BV | 811.096 |
| OemProduct | 811.096 |
| Chevalier et Roche | 4.794 |
| OemProduct | 4.794 |
| Goldkühle, Ne und Schedler | 331.380 |
| OemProduct | 331.380 |
| Hort GmbH | 12.202 |
| OemProduct | 12.202 |
| Leeuwen V.O.F. | 835.837 |
| OemProduct | 835.837 |
| Santoro-Barbieri s.r.l. | 323.595 |
| OemProduct | 323.595 |
| Grand Total | 3.998.304 |



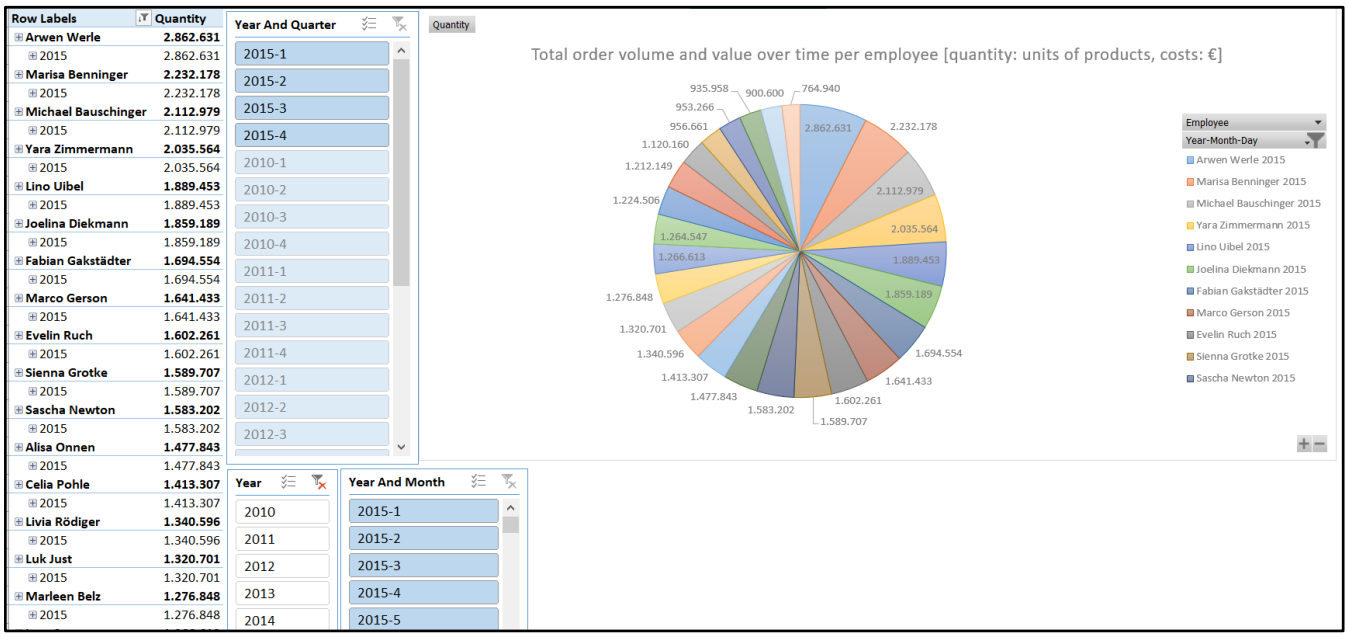
**Findings:**

- There is some dependency between individual suppliers and materials that they deliver, for instance, the Material “Buggy Logo Stickers” is mainly delivered by suppliers from Category “small” such as *Leeuwen V.O.F.* or *Brouwer BV*:

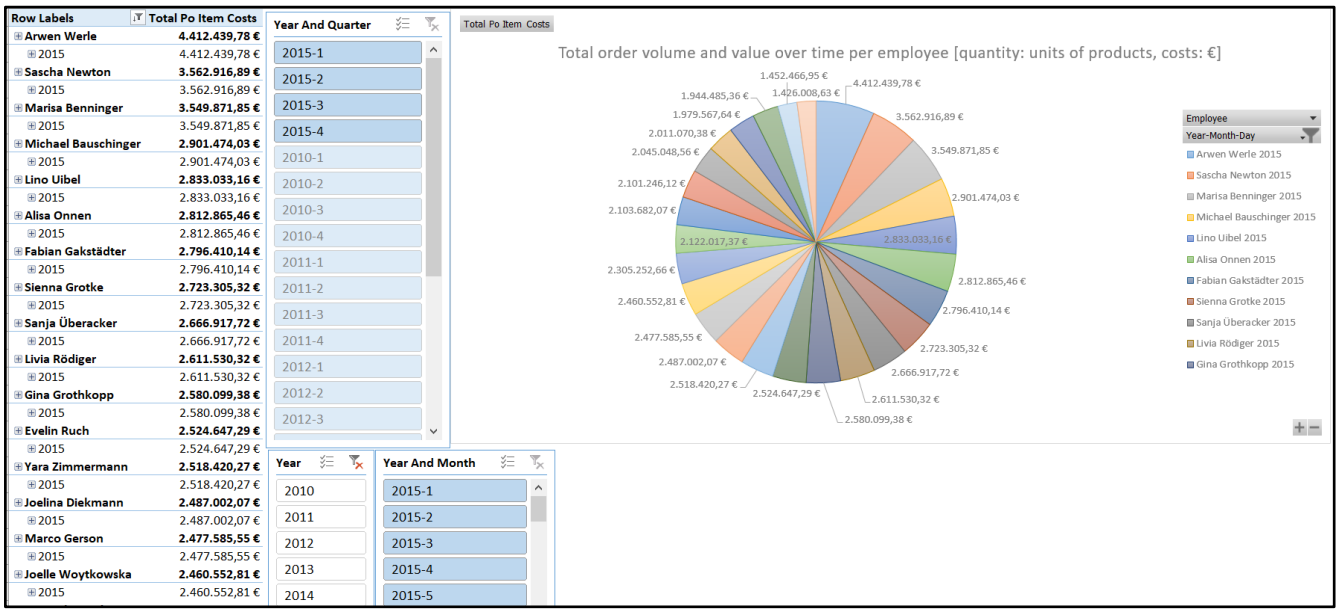


**Q6: What is the total order value** [measured in €] **and order volume** [measured in units of products] **per Employee per Year?**

Results for the year 2015 concerning **order volume** [measured in units of products]:



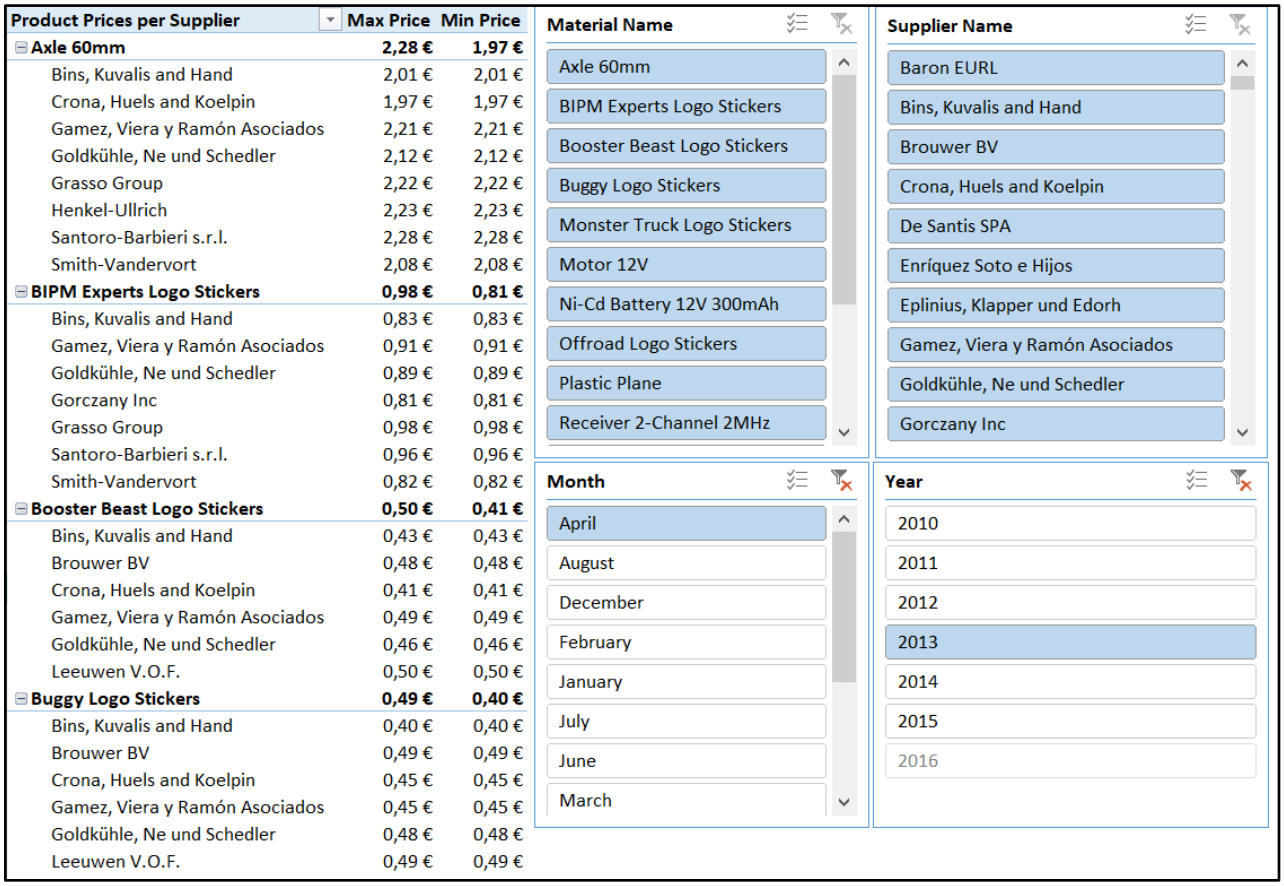

Results for the year 2015 concerning **order value** [measured in €]:




**Findings:**

- the results slightly vary, if one considers either order quantity or order value, for instance, four employees responsible for the most orders initiated in 2015:
    - **first pie chart:** Arwen Werle, Marisa Benninger, Michael Bauschinger, Yara Zimmerman
    - **second pie chart:** Arwen Werle, Sascha Newton, Marisa Benninger, Michael Bauschinger

**Q7: Who is the cheapest and most expensive Supplier for each Material per Month and Year?** [measured in Price of products in €]



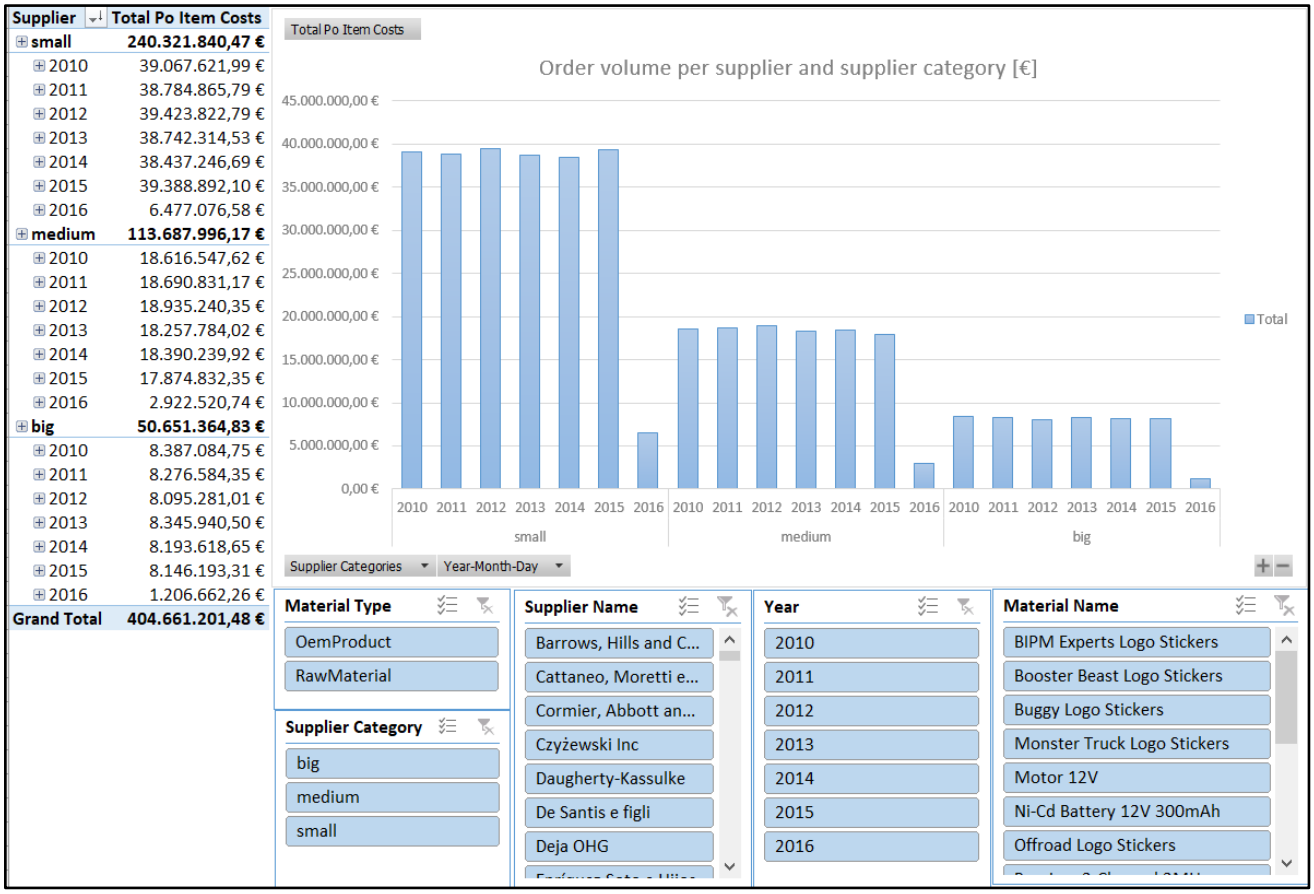
| Product Prices per Supplier | Max Price | Min Price |
|---|---|---|
| **Axle 60mm** | **2,28 €** | **1,97 €** |
| Bins, Kuvalis and Hand | 2,01 € | 2,01 € |
| Crona, Huels and Koelpin | 1,97 € | 1,97 € |
| Gamez, Viera y Ramón Asociados | 2,21 € | 2,21 € |
| Goldkühle, Ne und Schedler | 2,12 € | 2,12 € |
| Grasso Group | 2,22 € | 2,22 € |
| Henkel-Ullrich | 2,23 € | 2,23 € |
| Santoro-Barbieri s.r.l. | 2,28 € | 2,28 € |
| Smith-Vandervort | 2,08 € | 2,08 € |
| **BIPM Experts Logo Stickers** | **0,98 €** | **0,81 €** |
| Bins, Kuvalis and Hand | 0,83 € | 0,83 € |
| Gamez, Viera y Ramón Asociados | 0,91 € | 0,91 € |
| Goldkühle, Ne und Schedler | 0,89 € | 0,89 € |
| Gorczany Inc | 0,81 € | 0,81 € |
| Grasso Group | 0,98 € | 0,98 € |
| Santoro-Barbieri s.r.l. | 0,96 € | 0,96 € |
| Smith-Vandervort | 0,82 € | 0,82 € |
| **Booster Beast Logo Stickers** | **0,50 €** | **0,41 €** |
| Bins, Kuvalis and Hand | 0,43 € | 0,43 € |
| Brouwer BV | 0,48 € | 0,48 € |
| Crona, Huels and Koelpin | 0,41 € | 0,41 € |
| Gamez, Viera y Ramón Asociados | 0,49 € | 0,49 € |
| Goldkühle, Ne und Schedler | 0,46 € | 0,46 € |
| Leeuwen V.O.F. | 0,50 € | 0,50 € |
| **Buggy Logo Stickers** | **0,49 €** | **0,40 €** |
| Bins, Kuvalis and Hand | 0,40 € | 0,40 € |
| Brouwer BV | 0,49 € | 0,49 € |
| Crona, Huels and Koelpin | 0,45 € | 0,45 € |
| Gamez, Viera y Ramón Asociados | 0,45 € | 0,45 € |
| Goldkühle, Ne und Schedler | 0,48 € | 0,48 € |
| Leeuwen V.O.F. | 0,49 € | 0,49 € |



**Findings:**

- For the Material "*Axle 60mm*", for April 2013:
  - **the cheapest Supplier:** Crona, Huels and Koelpin
  - **the most expensive Supplier:** Santoro-Barbieri s.r.l.
- For the Material "*BIPM Experts Logo Stickers*", for April 2013:
  - **the cheapest Supplier:** Gorczany Inc
  - **the most expensive Supplier:** Grasso Group
- For the Material "*Booster Beast Logo Stickers*", for April 2013:
  - **the cheapest Supplier:** Crona, Huels and Koelpin
  - **the most expensive Supplier:** Leeuwen V.O.F.

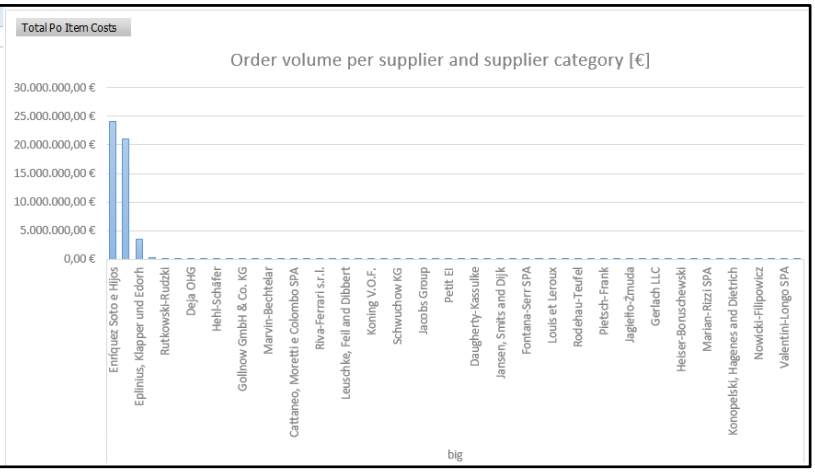
**Q8: What is the total order value per Supplier and Supplier Category per Month and Year?** [measured in €]



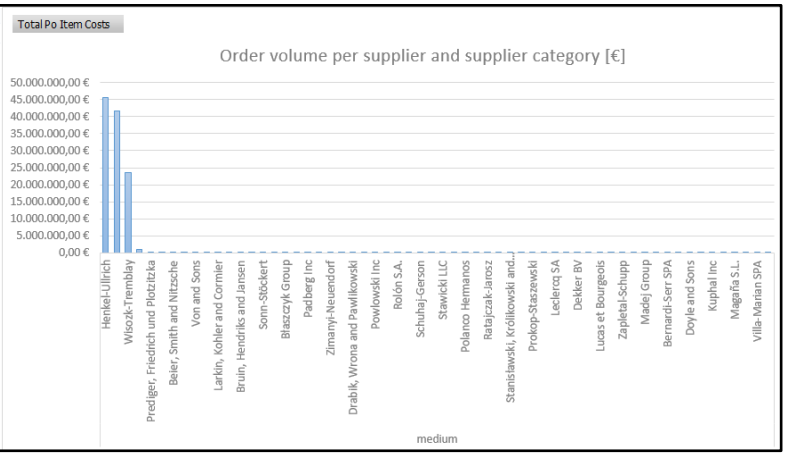

**Findings:**

- Again we can see that the distribution over time is very steady, implying that SuperX is able to maintain continuous expenditures on purchased materials and products, however the company is not growing.
- Purchase orders undertaken with Suppliers from category “small” prevail, reaching the total purchase order value of over 240 million € over the 6 years (2010-2016).
- In contrast, the value of transactions made with “big” Suppliers reaches in the same period only around 50 million €.
- After removing the Time dimension from the rows, we can find out the three top Suppliers with the highest order value over the entire period for each category:
    - **Big:** Enríquez Soto e Hijos; Gamez, Viera y Ramón Asociados; Eplinius, Klapper und Edorh
    - **Medium:** Henkel-Ullrich; Crona, Huels and Koelpin; Wisozk-Tremblay
    - **Small:** Baron EURL; Grasso Group; Santoro-Barbieri s.r.l.

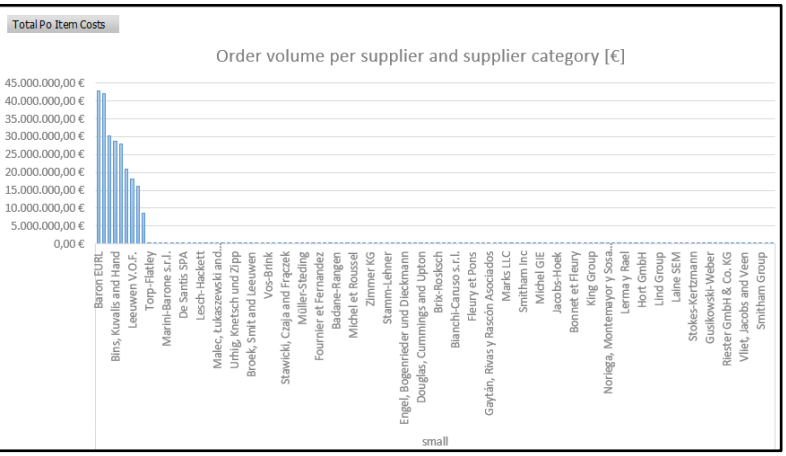
| Supplier | Total Po Item Costs |
|---|---|
| **big** | **50.651.364,83 €** |
| Enríquez Soto e Hijos | 24.141.931,47 € |
| Gamez, Viera y Ramón Asociados | 21.024.692,04 € |
| Eplinius, Klapper und Edorh | 3.465.352,10 € |
| Sören, Ihly und Patzwahl | 261.475,24 € |
| Rutkowski-Rudzki | 178.039,18 € |
| Knorr-Krol | 174.849,40 € |
| Deja OHG | 161.430,60 € |
| Suthe, Roecker und Lott | 157.709,16 € |
| Hehl-Schäfer | 154.378,40 € |
| Seeger, Fuchs und Sussmann | 135.253,80 € |
| Gollnow GmbH & Co. KG | 112.209,24 € |
| De Santis e figli | 98.974,42 € |
| Marvin-Bechtelar | 52.716,49 € |
| Marini-Basile SPA | 49.943,14 € |
| Cattaneo, Moretti e Colombo SPA | 48.845,76 € |
| Fadel-Ledner | 44.930,59 € |
| Riva-Ferrari s.r.l. | 36.812,16 € |
| Lakin-Durgan | 34.020,20 € |
| Leuschke, Feil and Dibbert | 33.352,98 € |
| Czyżewski Inc | 29.551,05 € |
| Koning V.O.F. | 24.763,86 € |



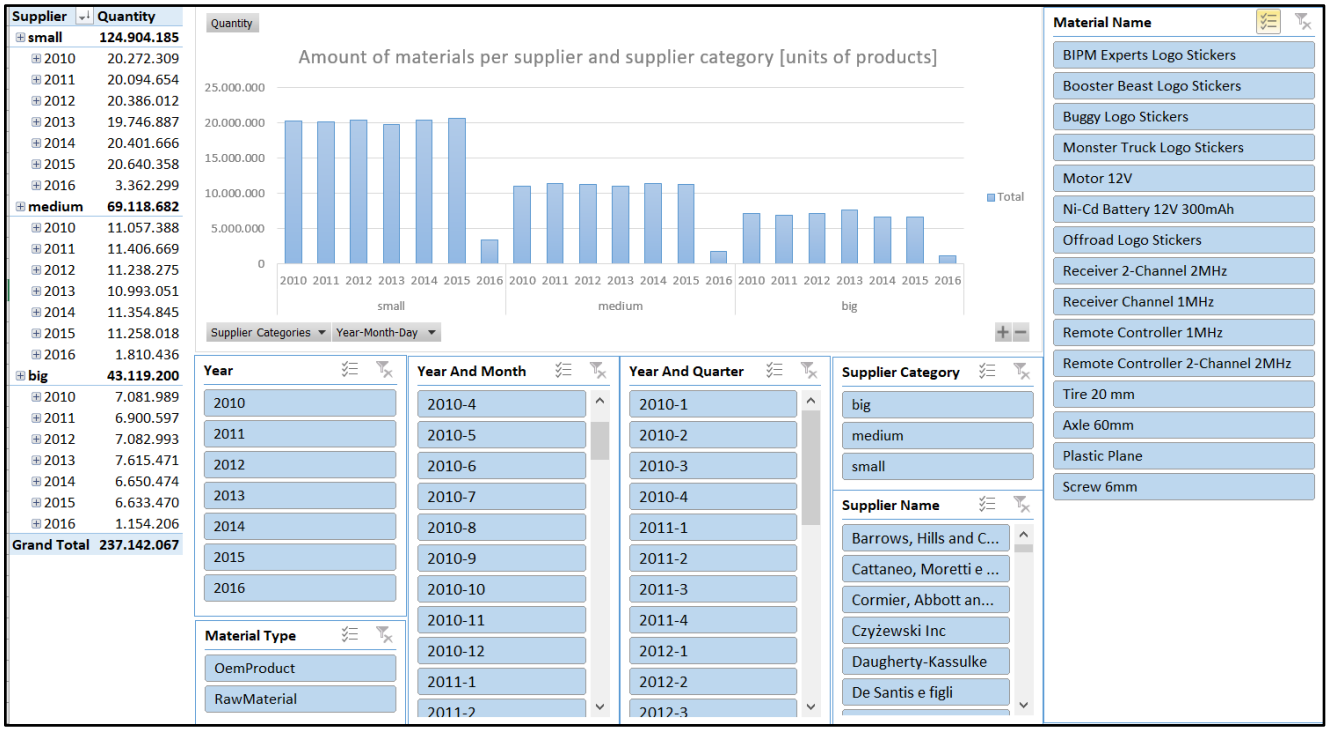

| Supplier | Total Po Item Costs |
|---|---|
| **medium** | **113.687.996,17 €** |
| Henkel-Ullrich | 45.718.721,39 € |
| Crona, Huels and Koelpin | 41.668.261,46 € |
| Wisozk-Tremblay | 23.481.852,38 € |
| Gorczany Inc | 982.628,24 € |
| Prediger, Friedrich und Plotzitzka | 261.836,56 € |
| MacGyver-Ruecker | 169.721,89 € |
| Beier, Smith and Nitzsche | 151.969,84 € |
| Hartmann-Runte | 144.746,94 € |
| Von and Sons | 143.855,87 € |
| Turcotte-Greenholt | 143.355,34 € |
| Larkin, Kohler and Cormier | 96.412,98 € |
| Günz, Dörre und Abel | 54.275,40 € |
| Bruin, Hendriks and Jansen | 52.722,72 € |
| Douglas-Rippin | 47.045,84 € |
| Sonn-Stöckert | 46.314,80 € |
| Cont, Bianchi e Coppola s.r.l. | 46.314,80 € |
| Błaszczyk Group | 45.156,40 € |
| Heidenreich Group | 33.137,22 € |
| Padberg Inc | 30.176,31 € |
| Duval SA | 28.069,72 € |
| Zimanyi-Neuendorf | 26.890,32 € |




| Supplier | Total Po Item Costs |
|---|---|
| **small** | **240.321.840,47 €** |
| Baron EURL | 42.780.210,06 € |
| Grasso Group | 42.203.079,08 € |
| Santoro-Barbieri s.r.l. | 30.194.576,81 € |
| Bins, Kuvalis and Hand | 28.615.023,24 € |
| Goldkühle, Ne und Schedler | 28.066.128,76 € |
| Smith-Vandervort | 21.042.160,17 € |
| Leeuwen V.O.F. | 18.085.327,41 € |
| Brouwer BV | 16.060.467,45 € |
| Riviere EURL | 8.612.316,88 € |
| Torp-Flatley | 276.005,95 € |
| Gibson Inc | 230.330,66 € |
| Riepl-Oberem | 212.771,92 € |
| Marini-Barone s.r.l. | 189.532,20 € |
| Scherbarth OHG | 175.920,24 € |
| Ureña, Ochoa y Tapia Asociados | 169.290,96 € |
| De Santis SPA | 163.329,48 € |
| Keebler-Douglas | 154.040,36 € |
| Legros Inc | 148.550,27 € |
| Lesch-Hackett | 143.521,03 € |
| Cummerata, Prosacco and Wisozk | 135.100,32 € |
| Wawrzyniak, Wierzbicki and Sosnowski | 119.767,68 € |



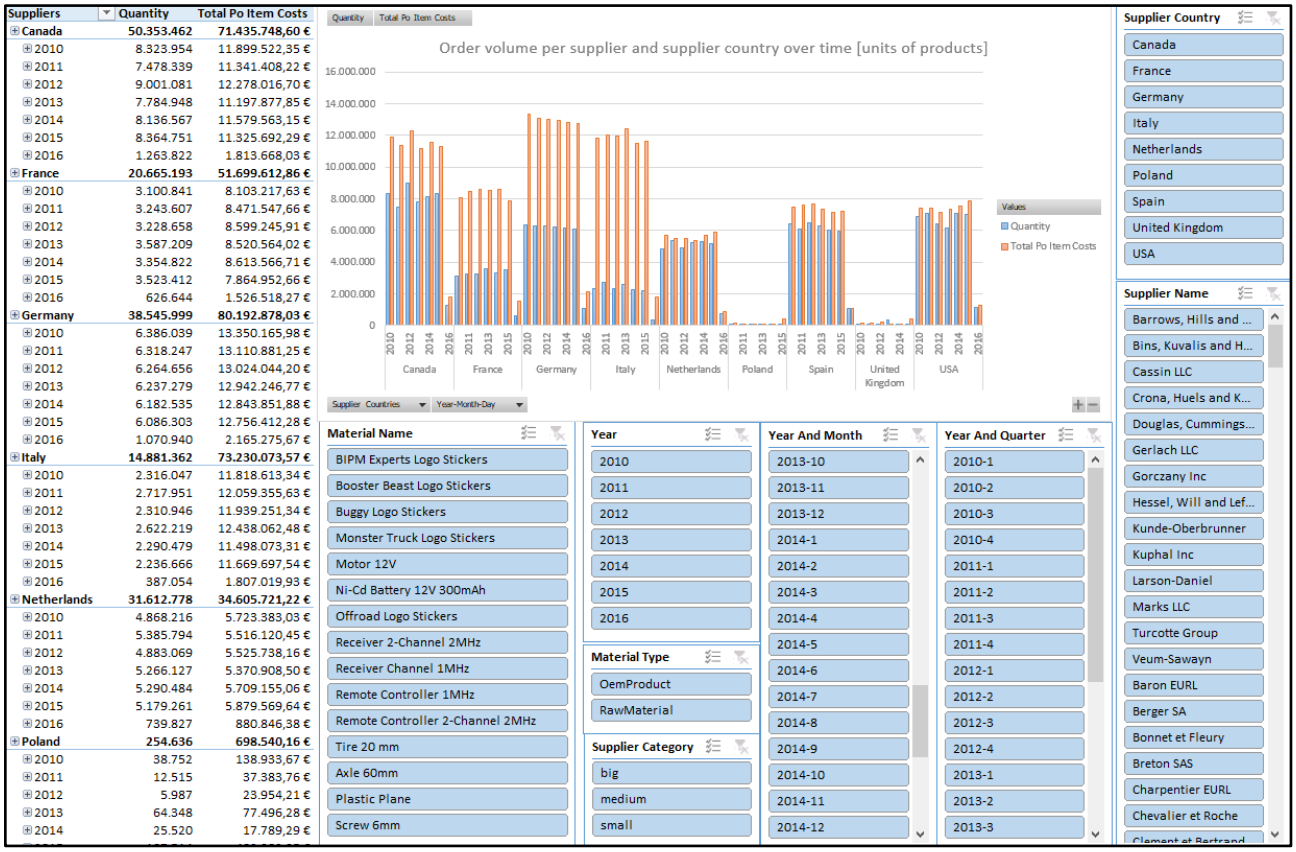
**Q9: What is the total order volume per Supplier and Supplier Category per Month and Year?** [measured in units of products]




**Findings:**

- Also here, the distribution over time is very steady, implying that SuperX can maintain the continuous amount of purchased materials and products, nevertheless the company is not growing.
- Purchase orders undertaken with Suppliers from category “small” prevail, reaching the total purchase order volume of over 124 million units of bought products over the 6 years period (2010-2016).
- In contrast, the value of transactions made with “big” Suppliers reaches in the same period only around 43 million units.
- After removing the Time dimension from the rows, we can find out the three top Suppliers with the highest order volume over the entire period for each category:
    - **Big:** Enríquez Soto e Hijos; Gamez, Viera y Ramón Asociados; Eplinius, Klapper und Edorh
    - **Medium:** Wisozk-Tremblay; Crona, Huels and Koelpin; Henkel-Ullrich
    - **Small:** Leeuwen V.O.F.; Goldkühle, Ne und Schedler; Bins, Kuvalis and Hand
- This way, we can observe that the results differ slightly, if one considers the order quantity rather than total costs.

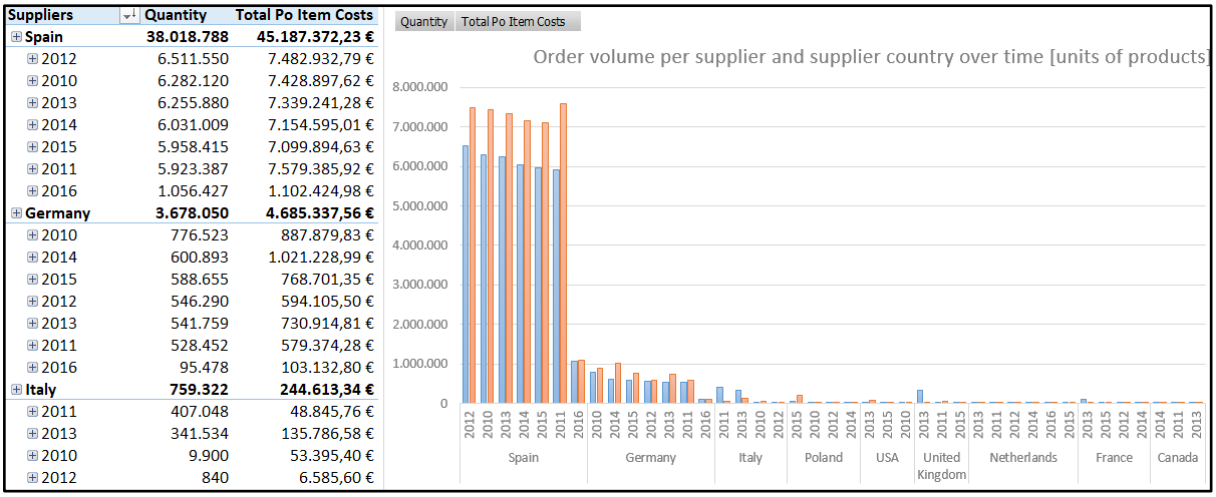
**Q10: What is the total order value** [measured in €] **and total order volume** [measured in units of products] **per Supplier Country over time?**




| Suppliers | Quantity | Total Po Item Costs |
|---|---|---|
| Canada | 50.353.462 | 71.435.748,60 € |
| 2010 | 8.323.954 | 11.899.522,35 € |
| 2011 | 7.478.339 | 11.341.408,22 € |
| 2012 | 9.001.081 | 12.278.016,70 € |
| 2013 | 7.784.948 | 11.197.877,85 € |
| 2014 | 8.136.567 | 11.579.563,15 € |
| 2015 | 8.364.751 | 11.325.692,29 € |
| 2016 | 1.263.822 | 1.813.668,03 € |
| France | 20.665.193 | 51.699.612,86 € |
| 2010 | 3.100.841 | 8.103.217,63 € |
| 2011 | 3.243.607 | 8.471.547,66 € |
| 2012 | 3.228.658 | 8.599.245,91 € |
| 2013 | 3.587.209 | 8.520.564,02 € |
| 2014 | 3.354.822 | 8.613.566,71 € |
| 2015 | 3.523.412 | 7.864.952,66 € |
| 2016 | 626.644 | 1.526.518,27 € |
| Germany | 38.545.999 | 80.192.878,03 € |
| 2010 | 6.386.039 | 13.350.165,98 € |
| 2011 | 6.318.247 | 13.110.881,25 € |
| 2012 | 6.264.656 | 13.024.044,20 € |
| 2013 | 6.237.279 | 12.942.246,77 € |
| 2014 | 6.182.535 | 12.843.851,88 € |
| 2015 | 6.086.303 | 12.756.412,28 € |
| 2016 | 1.070.940 | 2.165.275,67 € |
| Italy | 14.881.362 | 73.230.073,57 € |
| 2010 | 2.316.047 | 11.818.613,34 € |
| 2011 | 2.717.951 | 12.059.355,63 € |
| 2012 | 2.310.946 | 11.939.251,34 € |
| 2013 | 2.622.219 | 12.438.062,48 € |
| 2014 | 2.290.479 | 11.498.073,31 € |
| 2015 | 2.236.666 | 11.669.697,54 € |
| 2016 | 387.054 | 1.807.019,93 € |
| Netherlands | 31.612.778 | 34.605.721,22 € |
| 2010 | 4.868.216 | 5.723.383,03 € |
| 2011 | 5.385.794 | 5.516.120,45 € |
| 2012 | 4.883.069 | 5.525.738,16 € |
| 2013 | 5.266.127 | 5.370.908,50 € |
| 2014 | 5.290.484 | 5.709.155,06 € |
| 2015 | 5.179.261 | 5.879.569,64 € |
| 2016 | 739.827 | 880.846,38 € |
| Poland | 254.636 | 698.540,16 € |
| 2010 | 38.752 | 138.933,67 € |
| 2011 | 12.515 | 37.383,76 € |
| 2012 | 5.987 | 23.954,21 € |
| 2013 | 64.348 | 77.496,28 € |
| 2014 | 25.520 | 17.789,29 € |

**Findings:**

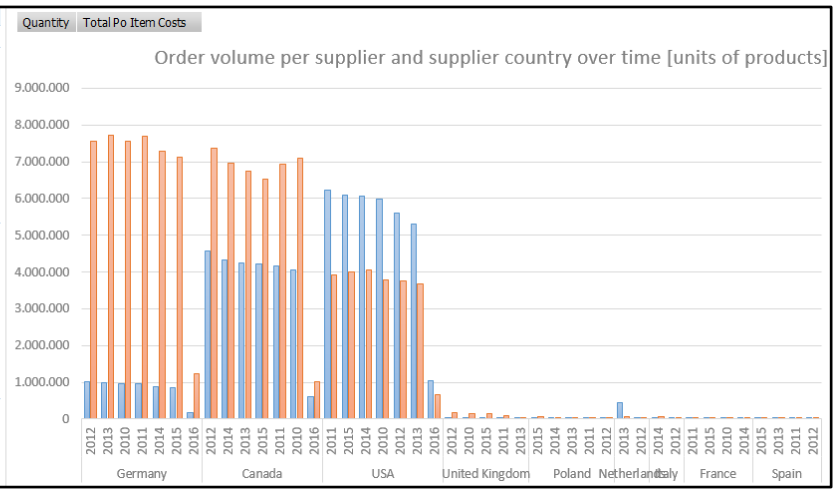
- Country with highest total order value is Germany, with 80.192.878,03 €
- Country with smallest total order value is Poland, with 698.540,16 €
- The same situation can be observed for purchase order value, i.e., quantity.
- Suppliers from category “big” are predominantly from Spain:




| Suppliers | Quantity | Total Po Item Costs |
|---|---|---|
| Spain | 38.018.788 | 45.187.372,23 € |
| 2012 | 6.511.550 | 7.482.932,79 € |
| 2010 | 6.282.120 | 7.428.897,62 € |
| 2013 | 6.255.880 | 7.339.241,28 € |
| 2014 | 6.031.009 | 7.154.595,01 € |
| 2015 | 5.958.415 | 7.099.894,63 € |
| 2011 | 5.923.387 | 7.579.385,92 € |
| 2016 | 1.056.427 | 1.102.424,98 € |
| Germany | 3.678.050 | 4.685.337,56 € |
| 2010 | 776.523 | 887.879,83 € |
| 2014 | 600.893 | 1.021.228,99 € |
| 2015 | 588.655 | 768.701,35 € |
| 2012 | 546.290 | 594.105,50 € |
| 2013 | 541.759 | 730.914,81 € |
| 2011 | 528.452 | 579.374,28 € |
| 2016 | 95.478 | 103.132,80 € |
| Italy | 759.322 | 244.613,34 € |
| 2011 | 407.048 | 48.845,76 € |
| 2013 | 341.534 | 135.786,58 € |
| 2010 | 9.900 | 53.395,40 € |
| 2012 | 840 | 6.585,60 € |

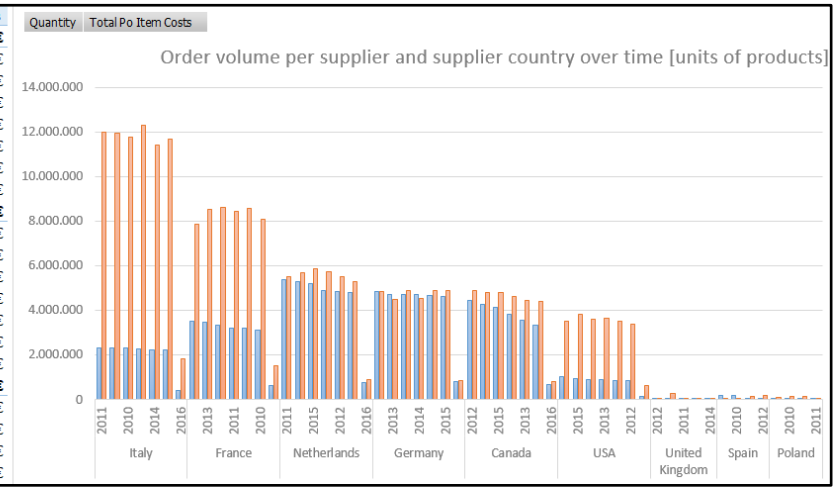
- In the category “medium”, suppliers are mainly from Germany, Canada and USA:

| Suppliers | Quantity | Total Po Item Costs |
|---|---|---|
| **Germany** | **5.837.848** | **46.154.547,74 €** |
| 2012 | 1.016.282 | 7.547.751,66 € |
| 2013 | 981.960 | 7.732.336,98 € |
| 2010 | 963.079 | 7.564.519,67 € |
| 2011 | 958.350 | 7.687.309,41 € |
| 2014 | 890.697 | 7.280.353,51 € |
| 2015 | 861.381 | 7.109.264,76 € |
| 2016 | 166.099 | 1.233.011,75 € |
| **Canada** | **26.184.519** | **42.654.134,19 €** |
| 2012 | 4.571.194 | 7.373.546,59 € |
| 2014 | 4.320.828 | 6.961.302,04 € |
| 2013 | 4.241.634 | 6.754.974,87 € |
| 2015 | 4.232.356 | 6.524.288,79 € |
| 2011 | 4.155.807 | 6.921.799,70 € |
| 2010 | 4.056.972 | 7.096.284,04 € |
| 2016 | 605.728 | 1.021.938,16 € |
| **USA** | **36.311.455** | **23.879.518,74 €** |
| 2011 | 6.238.641 | 3.932.808,00 € |
| 2015 | 6.083.841 | 3.993.245,28 € |
| 2014 | 6.067.199 | 4.058.622,63 € |
| 2010 | 5.988.699 | 3.796.473,55 € |



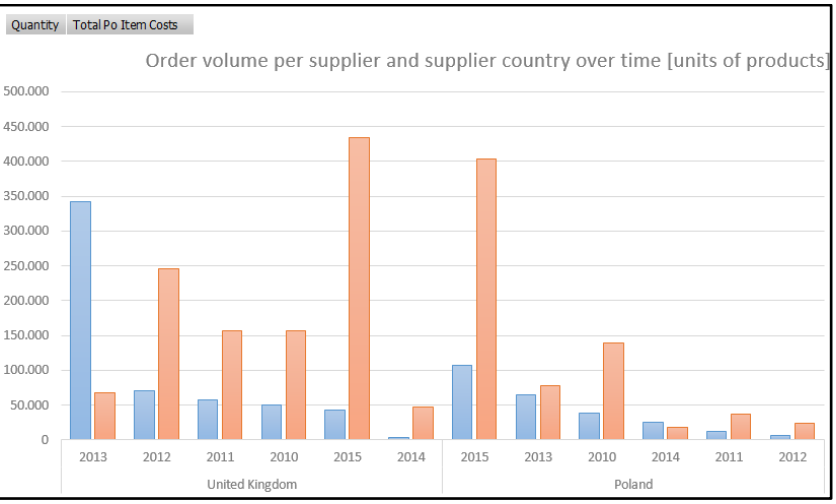

- In the category “small”, most of the suppliers are from Europe:

| Suppliers | Quantity | Total Po Item Costs |
|---|---|---|
| **Italy** | **14.071.689** | **72.913.722,37 €** |
| 2011 | 2.310.903 | 12.010.509,87 € |
| 2012 | 2.309.134 | 11.930.303,78 € |
| 2010 | 2.306.147 | 11.765.217,94 € |
| 2013 | 2.280.685 | 12.302.275,90 € |
| 2014 | 2.241.100 | 11.428.697,41 € |
| 2015 | 2.236.666 | 11.669.697,54 € |
| 2016 | 387.054 | 1.807.019,93 € |
| **France** | **20.480.380** | **51.627.389,07 €** |
| 2015 | 3.500.144 | 7.854.007,30 € |
| 2013 | 3.473.676 | 8.508.075,39 € |
| 2014 | 3.351.602 | 8.605.911,95 € |
| 2011 | 3.220.019 | 8.443.477,94 € |
| 2012 | 3.216.648 | 8.593.481,11 € |
| 2010 | 3.091.647 | 8.095.917,11 € |
| 2016 | 626.644 | 1.526.518,27 € |
| **Netherlands** | **31.117.195** | **34.459.682,66 €** |
| 2011 | 5.372.942 | 5.494.429,13 € |
| 2014 | 5.285.960 | 5.706.938,30 € |
| 2015 | 5.177.083 | 5.854.805,78 € |
| 2010 | 4.868.216 | 5.723.383,03 € |




- The distribution for each separate country is steady over time, except for Poland and United Kingdom:

| Suppliers | Quantity | Total Po Item Costs |
|---|---|---|
| **United Kingdom** | **567.297** | **1.107.287,56 €** |
| 2013 | 341.853 | 67.157,42 € |
| 2012 | 70.880 | 246.209,42 € |
| 2011 | 57.985 | 156.332,14 € |
| 2010 | 50.174 | 156.536,53 € |
| 2015 | 42.966 | 434.523,76 € |
| 2014 | 3.439 | 46.528,29 € |
| **Poland** | **254.636** | **698.540,16 €** |
| 2015 | 107.514 | 402.982,95 € |
| 2013 | 64.348 | 77.496,28 € |
| 2010 | 38.752 | 138.933,67 € |
| 2014 | 25.520 | 17.789,29 € |
| 2011 | 12.515 | 37.383,76 € |
| 2012 | 5.987 | 23.954,21 € |
| **Grand Total** | **821.933** | **1.805.827,72 €** |



**Q11: What is the total quantity ordered per country?** [measured in units of products]


**Findings:**

- It turned out that the highest purchase order volume is observed in Canada in the USA, and the smallest in the United Kingdom and Poland.
- The distribution varies slightly for different years, for instance in 2013 Spain turned out to outstrip USA and Germany with regard to the highest amount of ordered products:

# Proof-of-concept implementation - Postgres, Pentaho, Tableau

This section contains a step by step description of how the business intelligence architecture is implemented using Postgres as a database server, DBeaver as database management system, Spoon from Pentaho for the ETL process and Tableau for cube implementation and dashboards.

## ETL Process

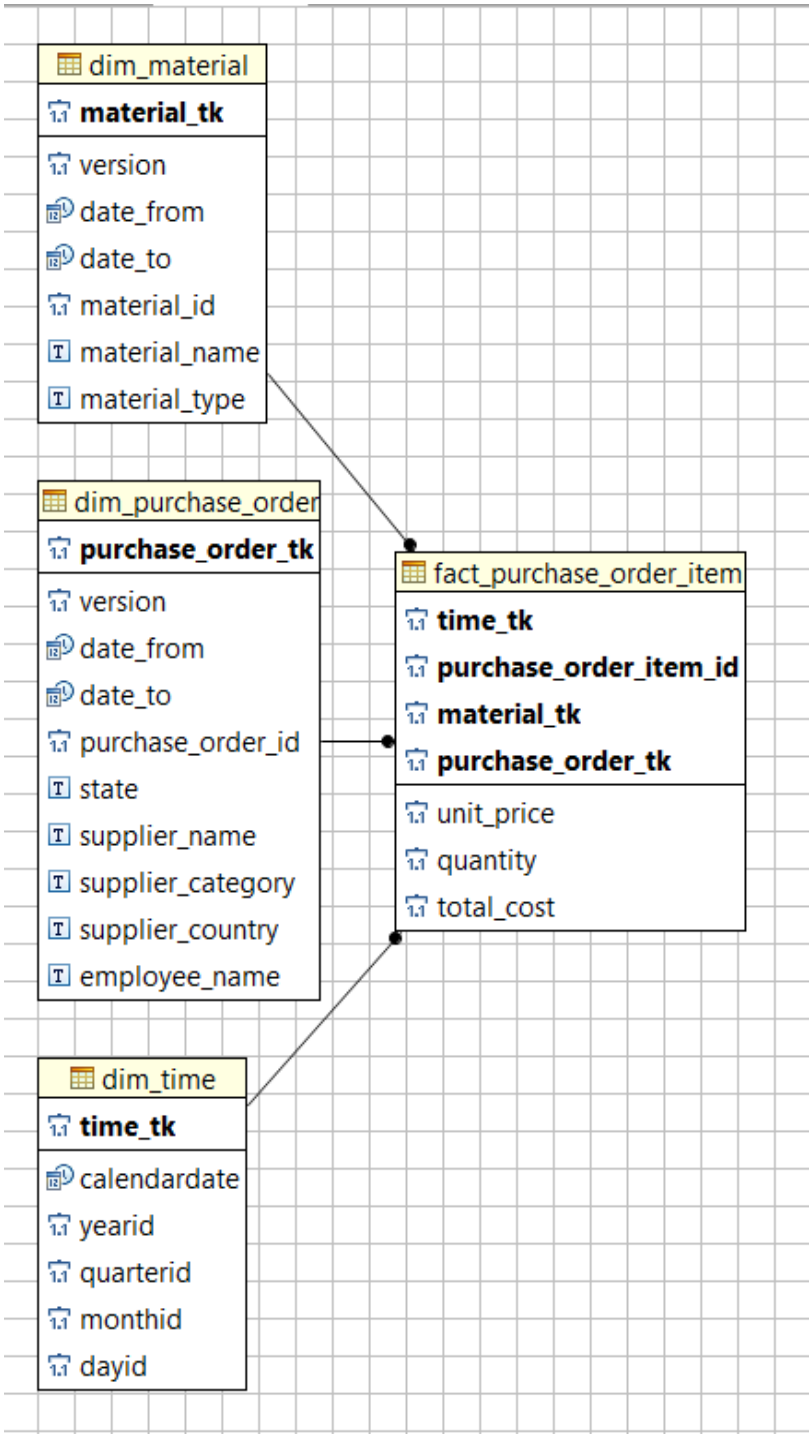
The ETL process begins after the data has been cleansed. The purpose is to create a star schema, then to initially upload the data to the OLAP database and to handle the slowly changing dimensions (SCD).

First of all, an empty star schema is created in Postgres using the MER diagram. After that the tables are created with SQL script. For this all fields and hierarchies from the MER diagram were identified and created as tables and columns in the star schema: three dimension tables: material, purchase order and time and one fact table: purchase order items, containing three measures: quantity, price and cost. The cost is derived by multiplying quantity and price. All measures are additive, and the fact table type is transactional.

For the procurement data mart, technical keys are used as the primary keys for the dimensions, and the combination of all those keys referred by the fact table will be its composite primary key. The time dimension will be pre-populated with the dates and their corresponding months and years. The materials and purchase order dimensions of SCD type 2, therefore their tables contain columns for version, date_from and date_to.

```sql
CREATE  TABLE IF NOT EXISTS dim_purchase_order (
  purchase_order_tk BIGINT PRIMARY KEY ,
  version INT NULL DEFAULT NULL ,
  date_from TIMESTAMP NULL DEFAULT NULL ,
  date_to TIMESTAMP NULL DEFAULT NULL ,
  purchase_order_id BIGINT NULL DEFAULT NULL ,
  state VARCHAR(100) NULL DEFAULT NULL ,
  supplier_name VARCHAR(255) NULL DEFAULT NULL ,
  supplier_category VARCHAR(100) NULL DEFAULT null,
  supplier_country VARCHAR(255) NULL DEFAULT null,
  employee_name VARCHAR(255) NULL DEFAULT null);

CREATE INDEX idx_dim_purchase_order_tk ON dim_purchase_order (purchase_order_tk);

CREATE TABLE IF NOT EXISTS dim_time (
  time_tk BIGINT PRIMARY KEY,
  calendarDate DATE NULL DEFAULT NULL ,
  yearID INT NULL DEFAULT NULL ,
  quarterID INT NULL DEFAULT NULL ,
  monthID INT NULL DEFAULT NULL ,
  dayID INT NULL DEFAULT null);

CREATE  INDEX idx_dim_time_tk ON dim_time (time_tk);

CREATE TABLE IF NOT EXISTS dim_material (
  material_tk SERIAL PRIMARY KEY ,
  version INT NULL DEFAULT NULL ,
  date_from TIMESTAMP NULL DEFAULT NULL ,
  date_to TIMESTAMP NULL DEFAULT NULL ,
  material_id BIGINT NULL DEFAULT NULL ,
  material_name VARCHAR(255) NULL DEFAULT NULL ,
  material_type VARCHAR(100) NULL DEFAULT null);

  CREATE  INDEX idx_dim_material_lookup ON dim_material (material_id);

  CREATE  TABLE IF NOT EXISTS fact_purchase_order_item (
  time_tk BIGINT NULL DEFAULT NULL REFERENCES dim_time (time_tk),
  purchase_order_item_id BIGINT NULL ,
  material_tk BIGINT NULL REFERENCES dim_material (material_tk),
  purchase_order_tk BIGINT NULL REFERENCES dim_purchase_order (purchase_order_tk),
  unit_price DOUBLE PRECISION NULL DEFAULT NULL ,
  quantity INT NULL DEFAULT NULL ,
  total_cost DOUBLE PRECISION NULL DEFAULT NULL ,
  CONSTRAINT fact_purchase_order_item_pkey
  PRIMARY KEY (time_tk, purchase_order_item_id, material_tk, purchase_order_tk) );

CREATE  INDEX idx_purchase_order_tk ON fact_purchase_order_item (purchase_order_tk);
CREATE  INDEX idx_material_tk ON fact_purchase_order_item (material_tk);
CREATE  INDEX idx time tk ON fact purchase order item (time tk);
```

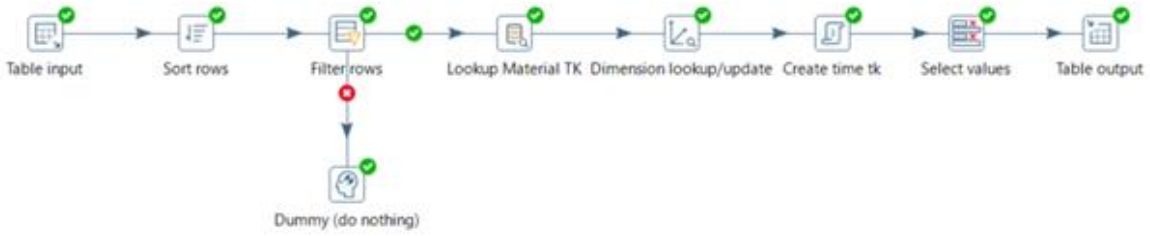
**Star Schema:**


dim_material
material_tk
version
date_from
date_to
material_id
material_name
material_type
dim_purchase_order
purchase_order_tk
version
date_from
date_to
purchase_order_id
state
supplier_name
supplier_category
supplier_country
employee_name
fact_purchase_order_item
time_tk
purchase_order_item_id
material_tk
purchase_order_tk
unit_price
quantity
total_cost
dim_time
time_tk
calendardate
yearid
quarterid
monthid
dayid

Using Spoon as ETL tool gives the advantage of the intuitive user interface and a minimal amount of coding. On the one hand this prevents from mistakes, on the other, it is also suitable for users who do not feel comfortable in writing code. Their inputs, outputs and different rules were specified for each table in the star schema. For the dimension tables *material* and *purchase order*, the transactions include input table, where the querying of the OLTP database is specified, and dimension lookup, where the logic for the SCD of type 2 is implemented. It uses the timestamp from the OLTP DB and compares to the date_to column in the star schema to select the correct version and thus always shows the last update of the fields.

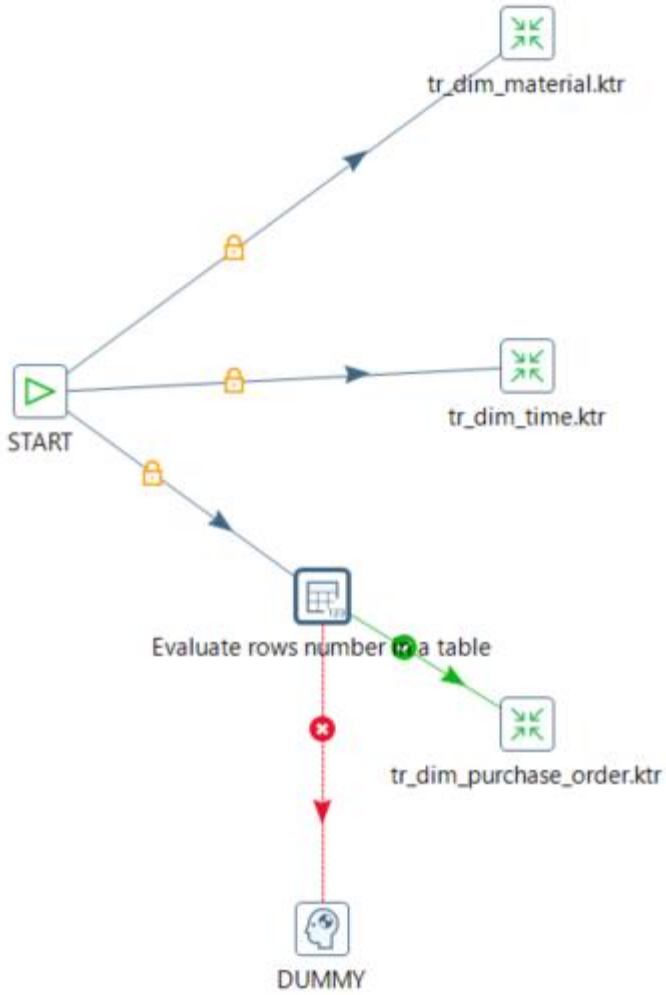

In the ETL process, the time dimension is created using sequence and calculators to generate the technical keys and days, months and years for 5000 dates, beginning from the first date we have in the OLTP database.


The last table created and populated is the fact table. It is required that it is last, because it contains all relations to the other tables and without the dimensions, the fact table cannot be populated. For this purpose. first the technical keys for the material and purchase order dimensions are mapped. Then the technical key for the time dimension is created based on the timestamp from the purchase order item table and the  date in the time dimension.




After all, transactions have been created, they are combined into jobs, that even allow for parallel execution. The table *time* must be created only once in the beginning, a condition that checks the row numbers is added. The first job creates the dimension tables in parallel, and the last job specifies the order and populates the fact table.

Job for dimensions:




Final job:


Now the star schema is populated with data and ready to be used for the creation of the cube, analysis and visualisations.

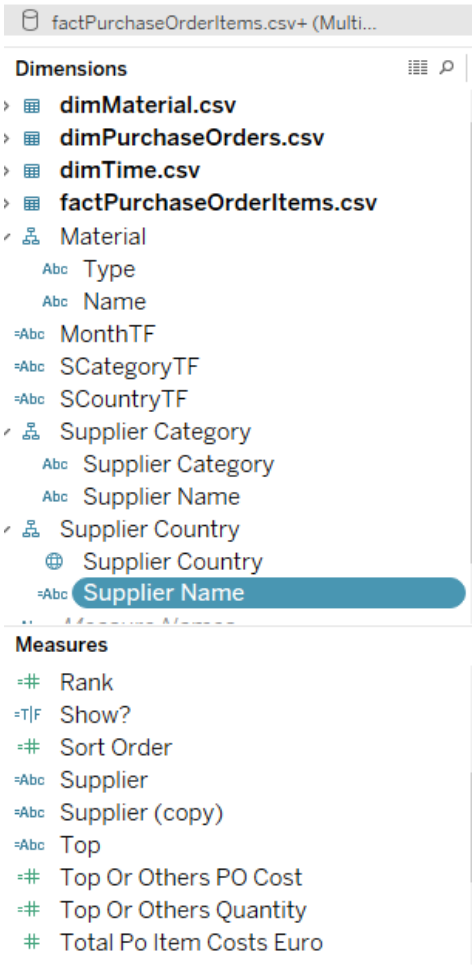
## Multidimensional Analysis and Visualizations

An OLAP database deployed on Postgres server allows for great flexibility and a vast choice of tools for analysis and visualizations as any tool that can be connected to a database can be used. For this implementation Tableau is used as analysis tool. Its main advantages are:

- Intuitive user interface
- Vivid and easy to interpret graphics
- Great computational power - in-memory database
- Large choice of possible data sources

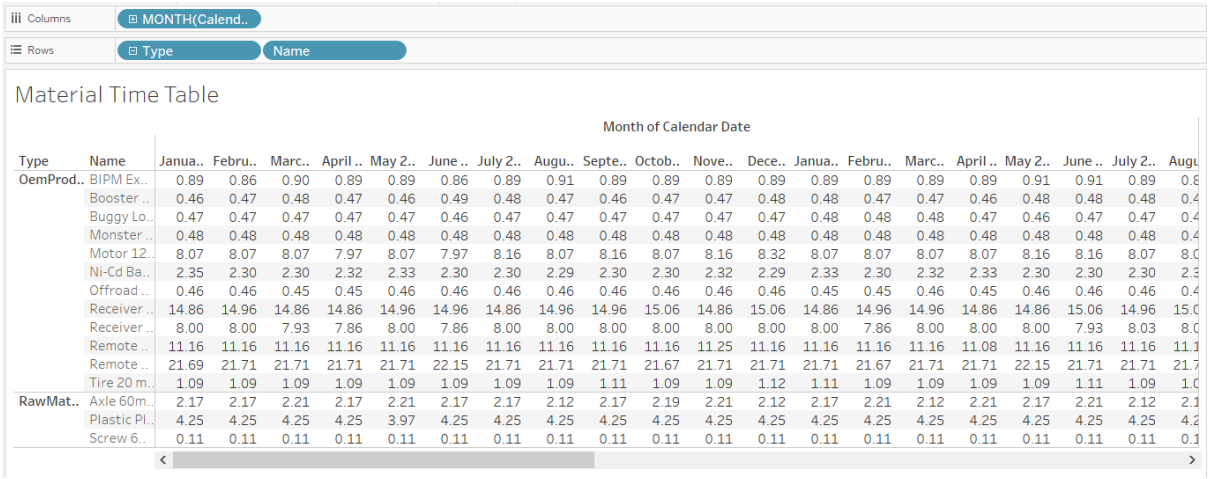
For this project, the data from the Microsoft implementation was first used for the Tableau dashboards. Afterwards it was connected to the Postgres' star schema for testing and comparison purposes.

The first task after loading the data is to create the hierarchies. In Tableau this is a very easy process - simple assignment of the dimension's members to their hierarchies.




**Example of roll-up - drill-down tables in Tableau:**

Most common unit price per material or material type over time - year, quarter, month.




Material Time Table

| | | Month of Calendar Date | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Type | Name | Janua.. | Febru.. | Marc.. | April .. | May 2.. | June .. | July 2.. | Augu.. | Septe.. | Octob.. | Nove.. | Dece.. | Janua.. | Febru.. | Marc.. | April .. | May 2.. | June .. | July 2.. | Augu |
| OemProd.. | BIPM Ex.. | 0.89 | 0.86 | 0.90 | 0.89 | 0.89 | 0.86 | 0.89 | 0.91 | 0.89 | 0.89 | 0.89 | 0.89 | 0.89 | 0.89 | 0.89 | 0.89 | 0.91 | 0.91 | 0.89 | 0.8 |
| | Booster .. | 0.46 | 0.47 | 0.48 | 0.47 | 0.46 | 0.49 | 0.48 | 0.47 | 0.46 | 0.47 | 0.47 | 0.48 | 0.48 | 0.47 | 0.47 | 0.46 | 0.48 | 0.48 | 0.48 | 0.4 |
| | Buggy Lo.. | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.46 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.48 | 0.48 | 0.48 | 0.47 | 0.46 | 0.47 | 0.47 | 0.4 |
| | Monster .. | 0.48 | 0.48 | 0.48 | 0.48 | 0.48 | 0.48 | 0.48 | 0.48 | 0.48 | 0.48 | 0.48 | 0.48 | 0.48 | 0.48 | 0.48 | 0.48 | 0.48 | 0.48 | 0.48 | 0.4 |
| | Motor 12.. | 8.07 | 8.07 | 8.07 | 7.97 | 8.07 | 7.97 | 8.16 | 8.07 | 8.16 | 8.07 | 8.16 | 8.32 | 8.07 | 8.07 | 8.07 | 8.07 | 8.16 | 8.16 | 8.07 | 8.0 |
| | Ni-Cd Ba.. | 2.35 | 2.30 | 2.30 | 2.32 | 2.33 | 2.30 | 2.30 | 2.29 | 2.30 | 2.30 | 2.32 | 2.29 | 2.33 | 2.30 | 2.32 | 2.33 | 2.30 | 2.30 | 2.30 | 2.3 |
| | Offroad .. | 0.46 | 0.46 | 0.45 | 0.45 | 0.46 | 0.46 | 0.46 | 0.46 | 0.46 | 0.46 | 0.46 | 0.46 | 0.45 | 0.45 | 0.46 | 0.45 | 0.46 | 0.46 | 0.46 | 0.4 |
| | Receiver .. | 14.86 | 14.96 | 14.86 | 14.86 | 14.96 | 14.96 | 14.86 | 14.96 | 14.96 | 15.06 | 14.86 | 15.06 | 14.86 | 14.96 | 14.96 | 14.86 | 14.86 | 15.06 | 14.96 | 15.0 |
| | Receiver .. | 8.00 | 8.00 | 7.93 | 7.86 | 8.00 | 7.86 | 8.00 | 8.00 | 8.00 | 8.00 | 8.00 | 8.00 | 8.00 | 7.86 | 8.00 | 8.00 | 8.00 | 7.93 | 8.03 | 8.0 |
| | Remote .. | 11.16 | 11.16 | 11.16 | 11.16 | 11.16 | 11.16 | 11.16 | 11.16 | 11.16 | 11.16 | 11.25 | 11.16 | 11.16 | 11.16 | 11.16 | 11.08 | 11.16 | 11.16 | 11.16 | 11.1 |
| | Remote .. | 21.69 | 21.71 | 21.71 | 21.71 | 21.71 | 22.15 | 21.71 | 21.71 | 21.71 | 21.67 | 21.71 | 21.71 | 21.71 | 21.67 | 21.71 | 21.71 | 22.15 | 21.71 | 21.71 | 21.7 |
| | Tire 20 m.. | 1.09 | 1.09 | 1.09 | 1.09 | 1.09 | 1.09 | 1.09 | 1.09 | 1.11 | 1.09 | 1.09 | 1.12 | 1.11 | 1.09 | 1.09 | 1.09 | 1.09 | 1.11 | 1.09 | 1.0 |
| RawMat.. | Axle 60m.. | 2.17 | 2.17 | 2.21 | 2.17 | 2.21 | 2.17 | 2.17 | 2.12 | 2.17 | 2.19 | 2.21 | 2.12 | 2.17 | 2.21 | 2.12 | 2.21 | 2.17 | 2.21 | 2.12 | 2.1 |
| | Plastic Pl.. | 4.25 | 4.25 | 4.25 | 4.25 | 3.97 | 4.25 | 4.25 | 4.25 | 4.25 | 4.25 | 4.25 | 4.25 | 4.25 | 4.25 | 4.25 | 4.25 | 4.25 | 4.25 | 4.25 | 4.2 |
| | Screw 6.. | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.1 |

Total quantity and cost per material or material type per supplier category and supplier.

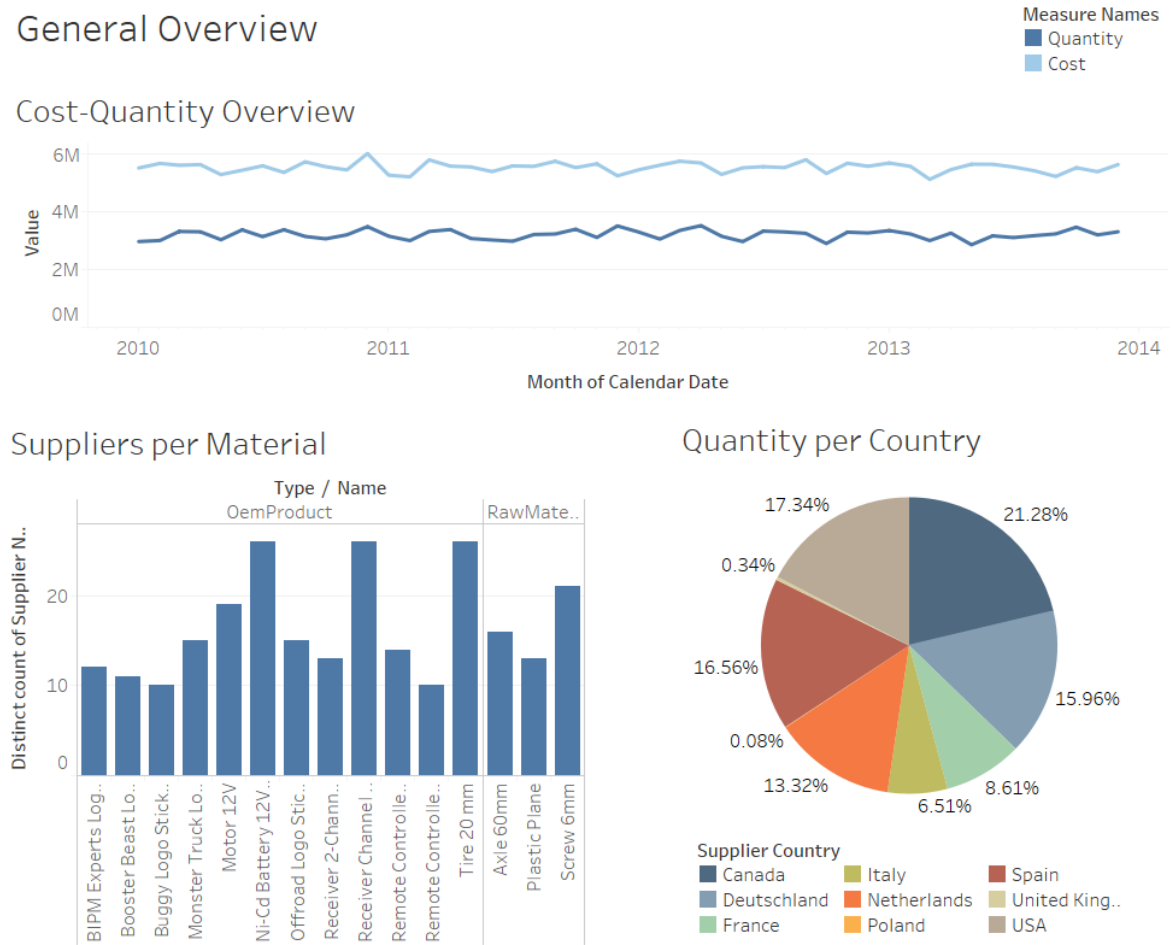

Material Supplier Table

| | | Supplier Category | | |
|---|---|---|---|---|
| Type | Name | big | medium | small |
| OemProduct | BIPM Experts Logo Stickers | 118,726<br>108,041 | 122,360<br>100,213 | 511,246<br>510,989 |
| | Booster Beast Logo Stickers | 88,159<br>43,198 | 122,720<br>68,218 | 405,109<br>191,315 |
| | Buggy Logo Stickers | 55,440<br>24,948 | 127,092<br>57,827 | 455,310<br>214,027 |
| | Monster Truck Logo Stickers | 139,233<br>66,832 | 266,517<br>120,609 | 266,962<br>123,999 |
| | Motor 12V | 62,085<br>521,625 | 222,329<br>1,701,876 | 553,984<br>4,486,777 |
| | Ni-Cd Battery 12V 300mAh | 51,075<br>117,049 | 39,227<br>94,510 | 37,682<br>110,340 |
| | Offroad Logo Stickers | 93,035<br>45,668 | 226,250<br>95,682 | 318,440<br>145,300 |
| | Receiver 2-Channel 2MHz | 221,834<br>3,357,286 | 180,536<br>2,613,493 | 468,957<br>6,870,152 |
| | Receiver Channel 1MHz | 35,382<br>282,639 | 42,213<br>311,846 | 55,877<br>448,255 |
| | Remote Controller 1MHz | 17,365<br>191,015 | 22,632<br>248,047 | 94,560<br>1,048,801 |
| | Remote Controller 2-Channel 2MHz | | 286,254<br>6,146,001 | 620,661<br>13,613,331 |
| | Tire 20 mm | 651,710<br>750,200 | 451,173<br>468,170 | 1,621,105<br>1,989,959 |
| RawMaterial | Axle 60mm | 182,472<br>403,263 | 247,743<br>527,935 | 1,065,411<br>2,294,037 |
| | Plastic Plane | 432,212<br>1,836,901 | 1,266,996<br>5,117,279 | 1,327,102<br>5,655,212 |
| | Screw 6mm | 5,466,743<br>594,592 | 7,379,438<br>769,362 | 11,944,481<br>1,375,549 |

For this project three Tableau dashboards were created. They are mostly for managerial use and offer an overview and more detailed analysis for the suppliers and materials in procurement department. As the current processes are monthly, the update frequency would also be monthly.

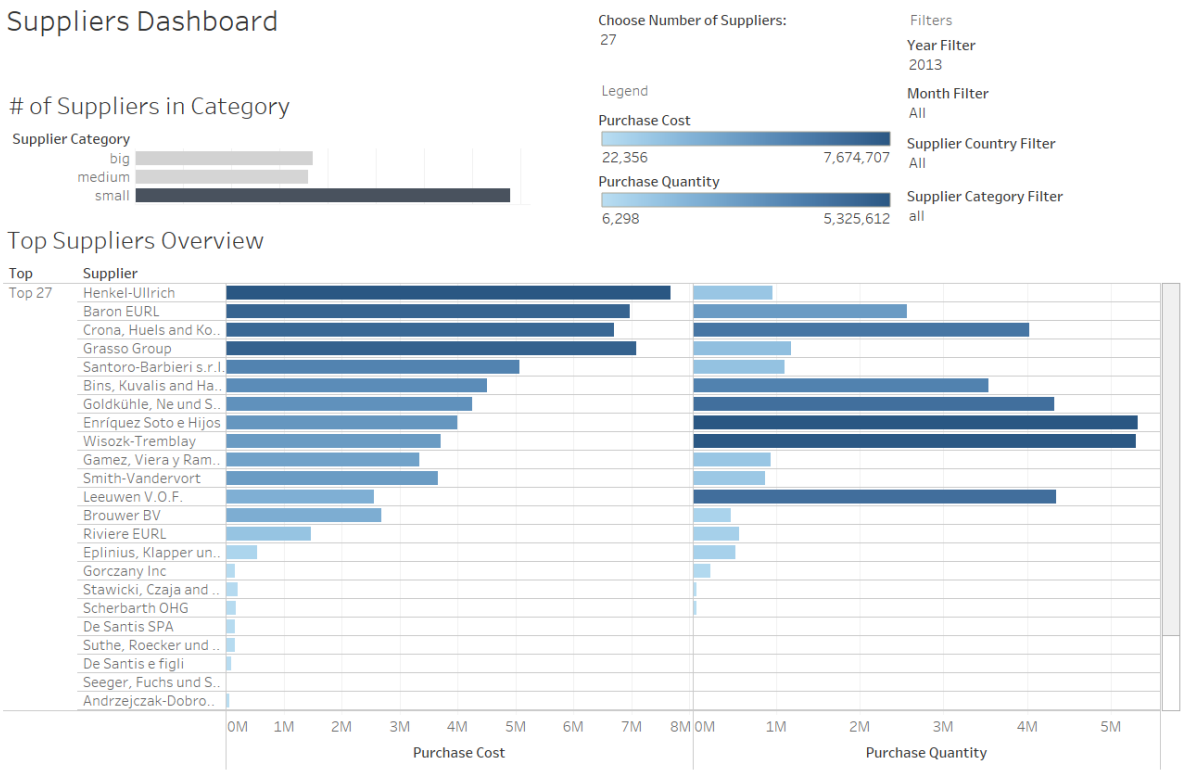
**General Overview**

On this dashboard, there are three graphs. The Cost-Quantity Overview gives a representation of how the cost and quantity changed over time. It could be used to track trends or detect increase or decrease in cost of the materials. The next graph - Suppliers per Material shows from how many Suppliers the materials can be delivered. It is meant to prevent the case when only one supplier can deliver particular material and the thus the entire production process is dependent on this supplier. For procurement, this would be a red light to find an alternative supplier. The pie chart shows where the materials are ordered from. It is crucial to know that more than 30% of the materials are delivered from North America. It indicates that for example international trade laws should be monitored closely.

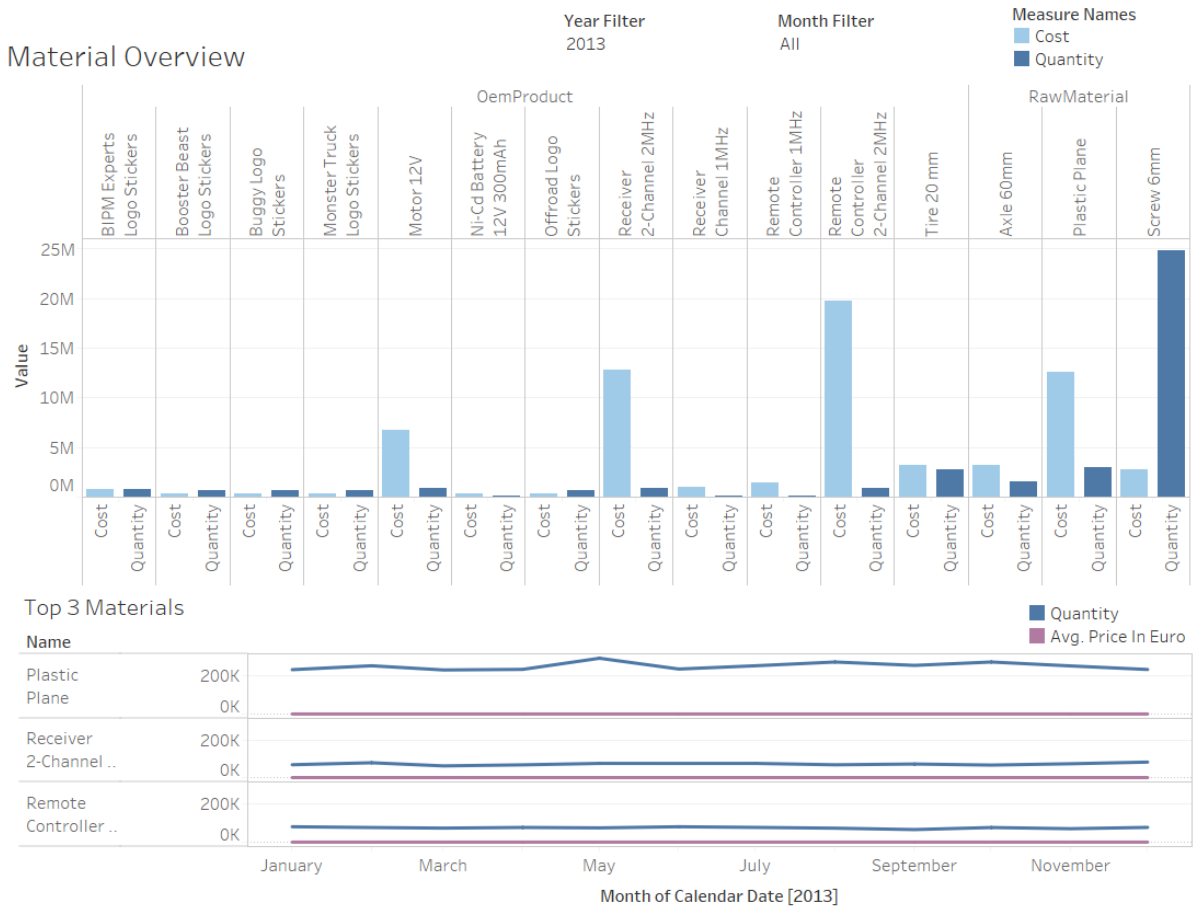



The Suppliers dashboard gives all the information relevant to Suppliers. The small bar chart shows the distribution of suppliers per category. It is interesting to notice that the category small has double the number of suppliers. Additionally, the user can choose how many of the top suppliers or all suppliers they would like to analyze and then filter by year, month, supplier category or country. The two bars show the difference between the total price paid to the

supplier and the total quantity received. Thus it can be inferred that the most expensive products come from Helkel-Ullrich.




The last dashboard gives an overview of the materials prices, total cost and quantity. It allows for filtering of month and year and easy detection of the most essential products in terms of cost and quantity. The second graph tracks the median unit price vs. the total quantity of the most expensive products. It could be an indicator to change the supplier of those products.


Material Overview
Year Filter
2013
Month Filter
All
Measure Names
Cost
Quantity
OemProduct
RawMaterial
BIPM Experts Logo Stickers
Booster Beast Logo Stickers
Buggy Logo Stickers
Monster Truck Logo Stickers
Motor 12V
Ni-Cd Battery 12V 300mAh
Offroad Logo Stickers
Receiver 2-Channel 2MHz
Receiver Channel 1MHz
Remote Controller 1MHz
Remote Controller 2-Channel 2MHz
Tire 20 mm
Axle 60mm
Plastic Plane
Screw 6mm
Value
25M
20M
15M
10M
5M
0M
Cost
Quantity
Top 3 Materials
Name
Quantity
Avg. Price In Euro
Plastic Plane
Receiver 2-Channel ..
Remote Controller ..
200K
0K
January
March
May
July
September
November
Month of Calendar Date [2013]

## Comparison of technologies

| | **Advantages** | **Disadvantages** |
|---|---|---|
| **Microsoft SQL Server:**<br>- SQL Server Management Studio as Database Management System,<br>- SQL Server Integration Services for ETL Process & Data Profiling,<br>- SQL Server Analysis Services for multidimensional implementation,<br>- SQL Server Reporting Services or Microsoft Excel as a base for reports and visualization) | - Microsoft SQL Server covers the entire process for creation of Data Warehouse, from OLTP data management, over ETL process, Data Profiling, to creation of dashboards<br>- The full functionality provided by a single software tool→ no need for third party tools<br>- User-friendly Excel Dashboards as merit, as anyone can use it<br>- Possibility of drilling down into single observations, which is highly appreciated among business users<br>- Easy refresh strategy, by clicking on “Refresh all”<br>- Well established product with a long history on the market (since 1989), extensive documentation and support provided by the vendor as well as by communities online. | - Rigid structure: every aggregate function for measures should be pre-calculated<br>- Requires substantial amount of resources (expensive)<br>- Not easy to use in the beginning<br>- Cumbersome security restrictions, for instance, deployment of cubes requires creation of special users with only read permissions (and those users are not predefined) |
| **Postgres and Tableau**<br>- Postgres server,<br>- DBeaver as Database Management System,<br>- Pentaho Spoon Data | - Open source technology, which may not require many resources<br>- High level of flexibility in the choice of tools | - Different providers<br>- more effort to contact each (ex: maintenance)<br>- The developers need to learn how to use very |

| | | |
|---|---|---|
| Integration,<br>- Tableau for visualizations | -Independent products that can be substituted<br>- In-memory database for the analysis in Tableau<br>- User-friendly interface<br>- Very high levels of flexibility | different programs |

# Process Intelligence

Human mining: Business Process Model and Notation

Process mining: Automatic process discovery

Extract the relevant event log info from the DB:

```
select id, activity, start_timestamp, end_timestamp, employee_id, department, "month", "year", year_month
from event_logs
where department = 'Procurement'
```

Load the data to disco:

The Case ID for this events log is the year/month combination as the processes in the procurement department are monthly. Activities, employees as resource and start and end timestamps are also included.

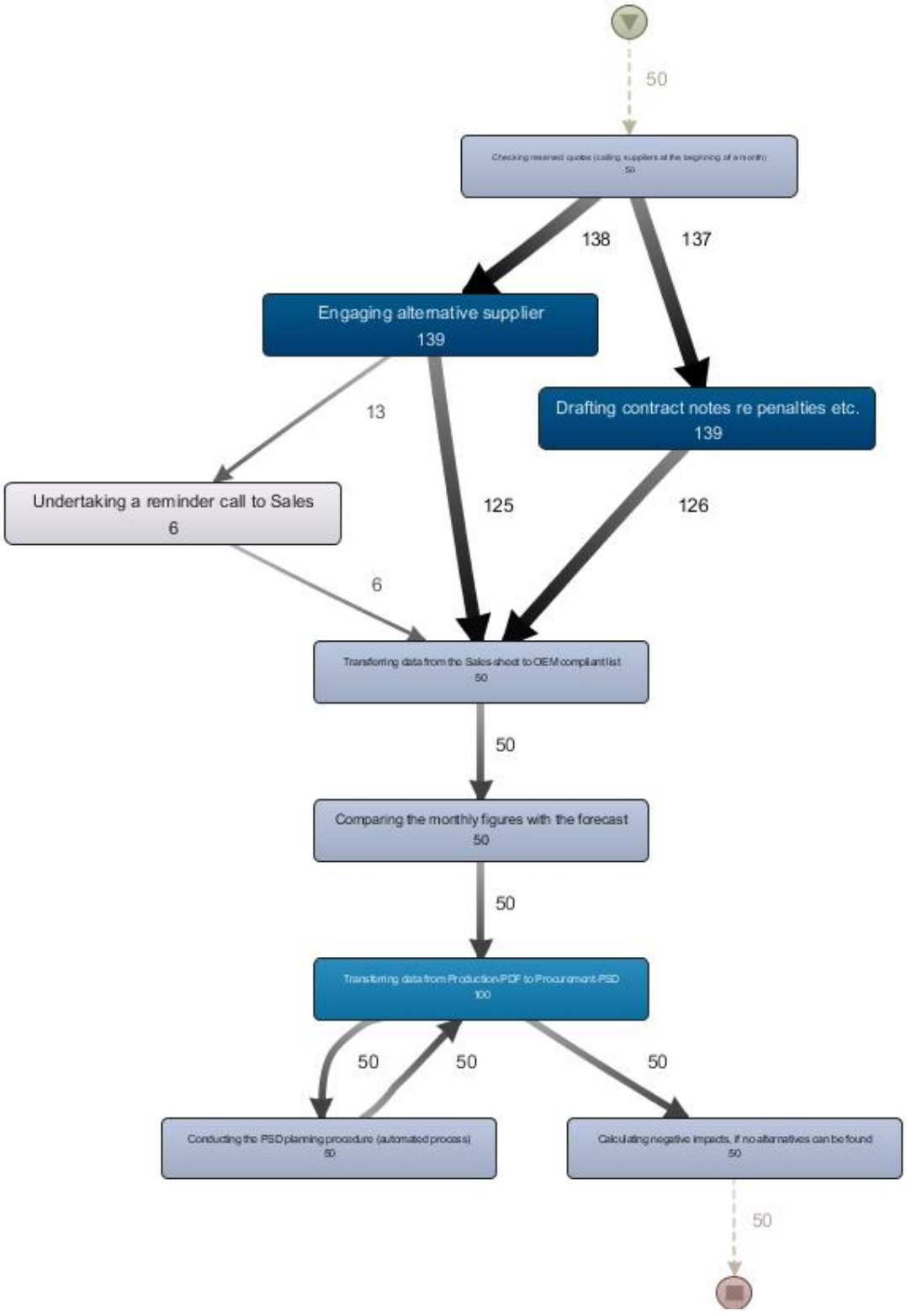
Overview:

In total, the event log provides data from January 2010 to February 2014. Those are 50 months and therefore 50 cases. Over that time span, 11 different Variants of execution of the process were conducted. The total duration per case varies between 22 days and 25 days. The average waiting time varies between 4 days and 7 hours and 6 days and 3 hours. The minimum number of events per case is 7, and the maximum is 19.

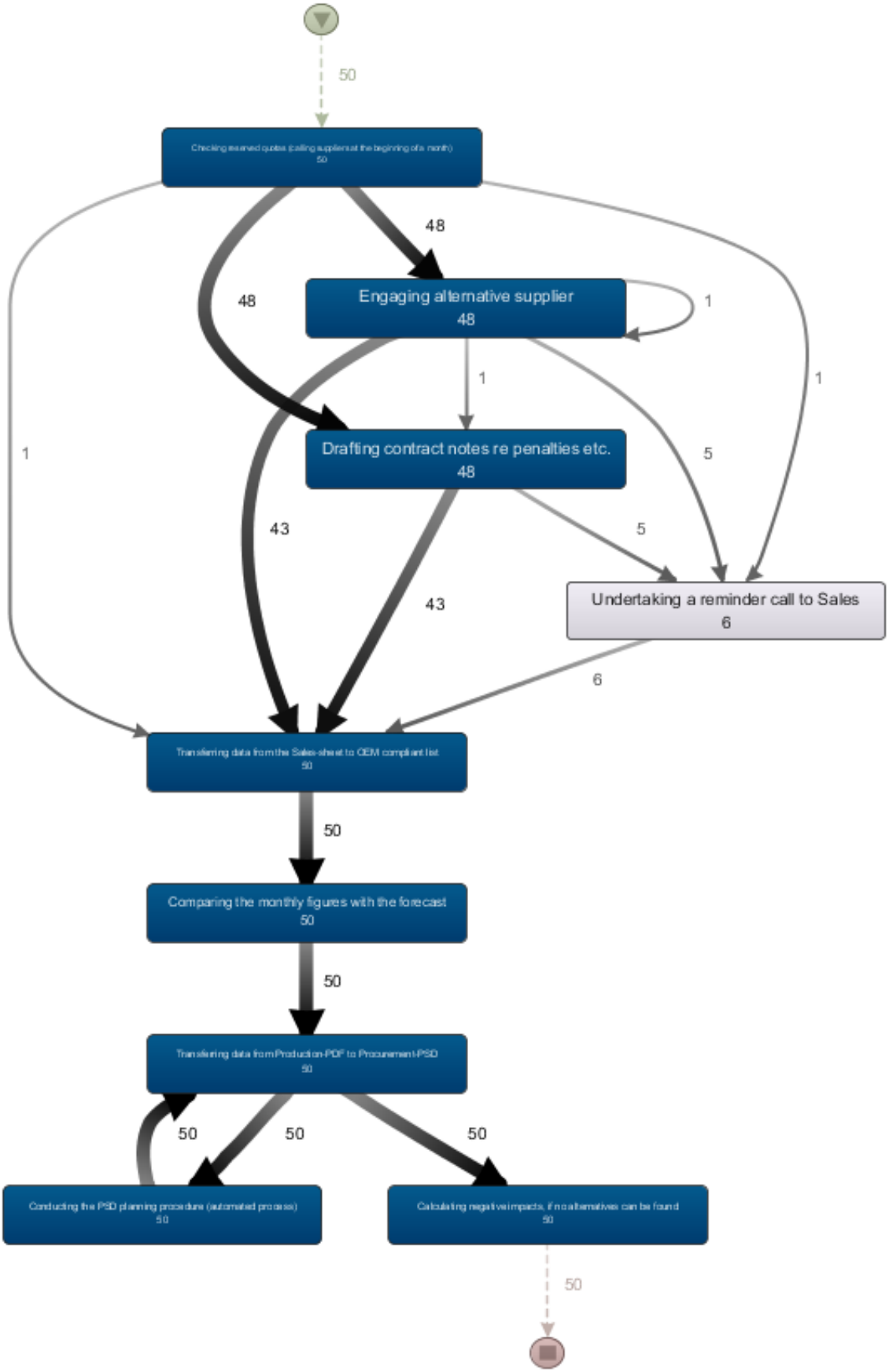
Results:

The first graph represents the overview of the process execution and the absolute frequencies per activity and path throughout the whole period.

It is easy to conclude that the two most frequently executed activities are *"Engaging alternative supplier"* and *"Drafting contract notes"* with 139 occurrences for 50 cases. This indicates that they are repeated almost three times per process execution. The other frequently executed activity - 100 times is *"Transferring data from Sales-sheet to OEM compliant sheet*" The outlier on the lower boundary is the activity *"Undertaking a reminder call to Sales"*, which occurs only six times.




Furthermore, the case frequencies (in how many cases a specific activity or path appears) show well-distributed load and low variance of process paths. In the second map, all paths were included. Almost all activities are present in all cases.

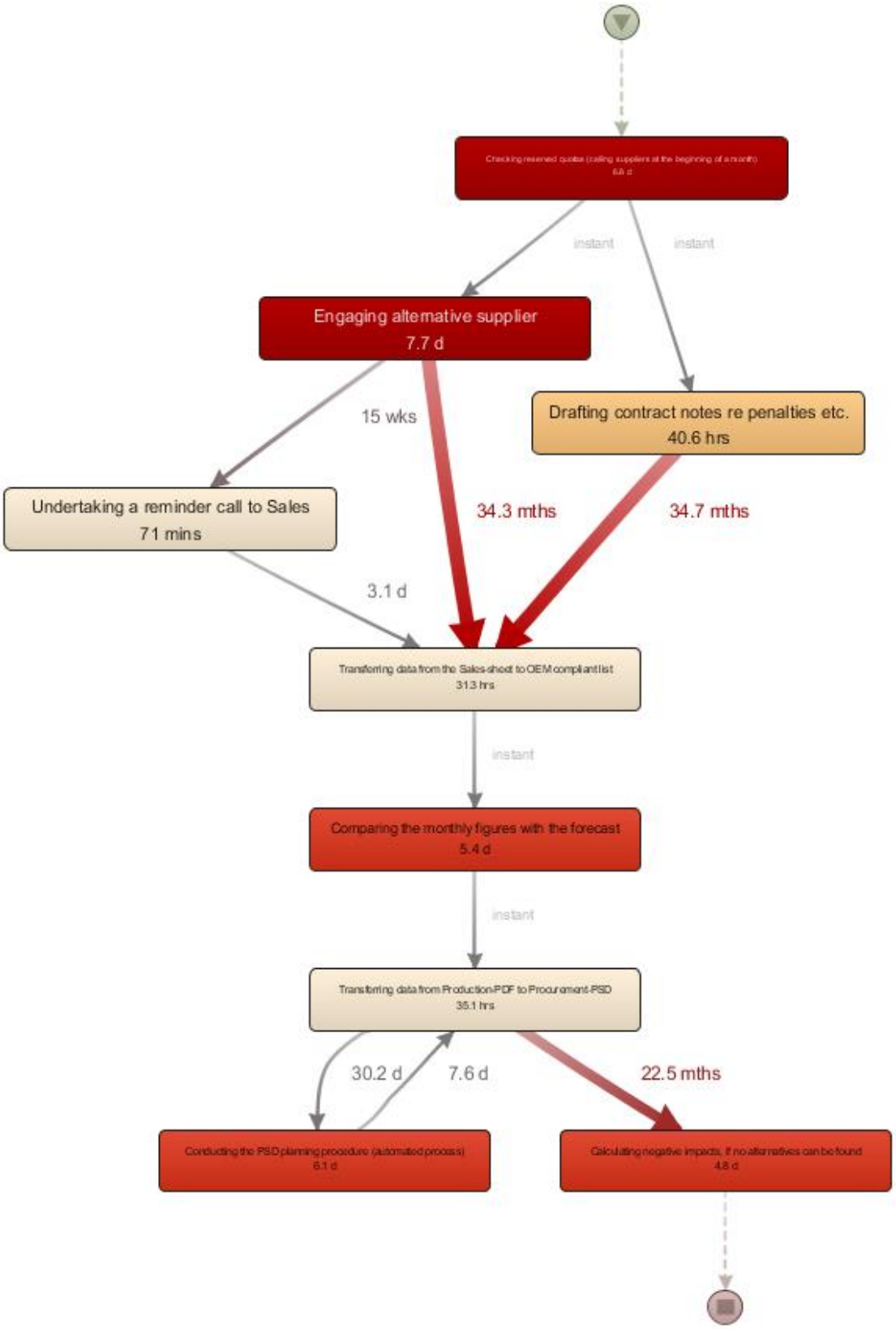

50
Checking received quotes (calling suppliers at the beginning of a month)
50
48
48
Engaging alternative supplier
48
1
1
1
1
Drafting contract notes re penalties etc.
48
5
43
5
43
Undertaking a reminder call to Sales
6
6
Transferring data from the Sales-sheet to OEM compliant list
50
50
Comparing the monthly figures with the forecast
50
50
Transferring data from Production-PDF to Procurement-PSD
50
50
50
50
Conducting the PSD planning procedure (automated process)
50
Calculating negative impacts, if no alternatives can be found
50
50

The two frequency graphs show very clearly the difference between the expected process sequence and the actual one. In the process descriptions, it is identified that the activities related to the engagement of new suppliers should happen very rarely. However, the process diagrams indicate that they occur a couple of times per process instance. Therefore it can be concluded that this part of the process needs to be improved.

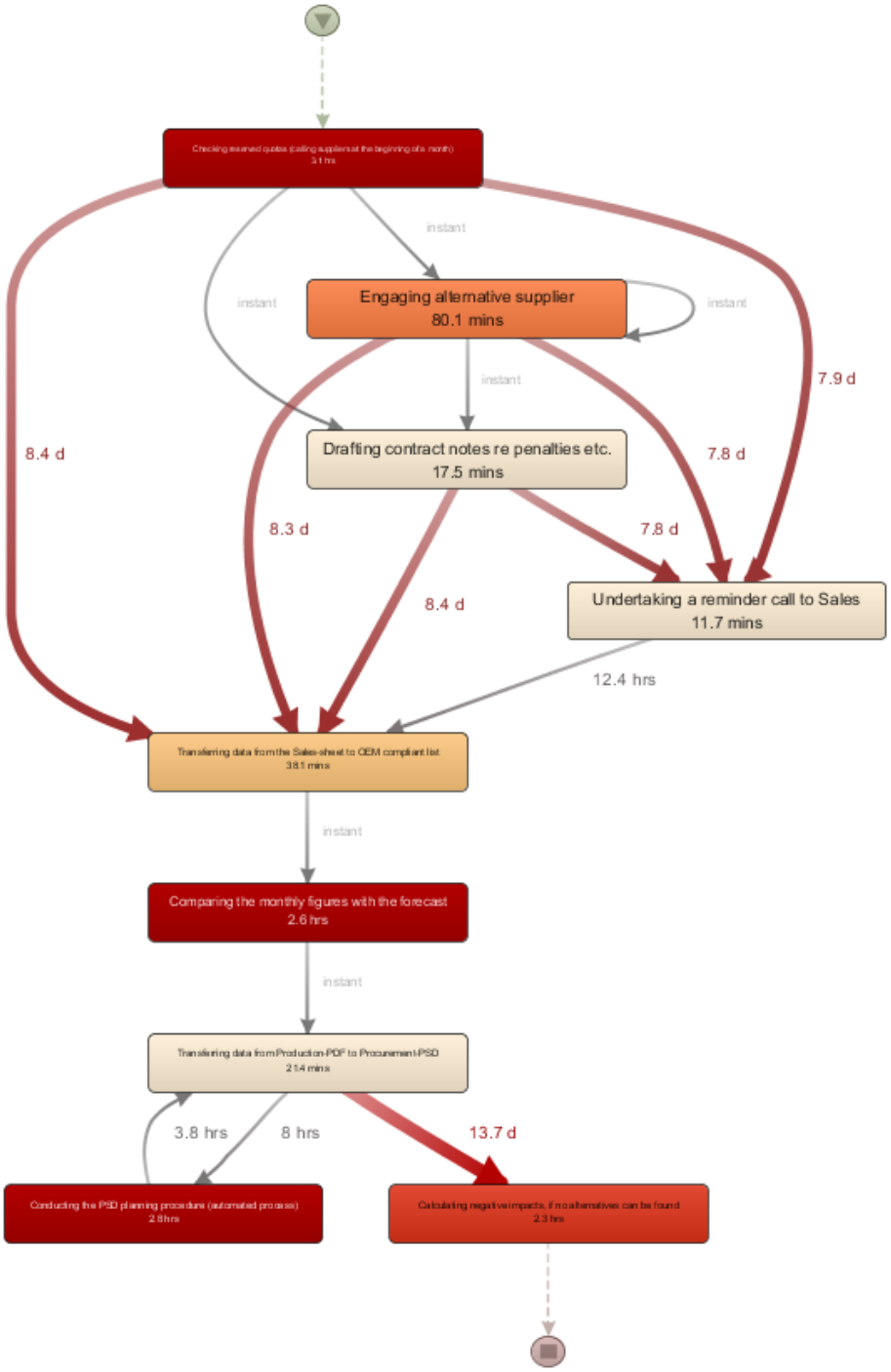
Graph 3 shows the total duration per activity and the waiting times. It is easy to spot the most time-consuming activities in the process, namely:

- *Checking reserved quotas (calling suppliers at the beginning of the month)*
- *Engaging alternative suppliers*
- *Comparing monthly figures with the forecast*
- *Checking negative impacts, if no alternatives can be found*

Throughout the whole period, each took over five days. This could be explained either by the duration of the activity itself or by the fact that it has been repeated many times.
Furthermore, the highest waiting times, represented by the thick red lines, are identified. In total, each of them takes more than 22 months in the period 2010 - 2014. The two highest ones are connected to the activities of engaging alternative suppliers as well.


Checking reserved quotas (calling suppliers at the beginning of a month)
6.6 d
instant
instant
Engaging alternative supplier
7.7 d
15 wks
34.3 mths
34.7 mths
Drafting contract notes re penalties etc.
40.6 hrs
Undertaking a reminder call to Sales
71 mins
3.1 d
Transferring data from the Sales-sheet to OEM compliant list
31.3 hrs
instant
Comparing the monthly figures with the forecast
5.4 d
instant
Transferring data from Production-PDF to Procurement-PSD
35.1 hrs
30.2 d
7.6 d
22.5 mths
Conducting the PSD planning procedure (automated process)
6.1 d
Calculating negative impacts, if no alternatives can be found
4.8 d

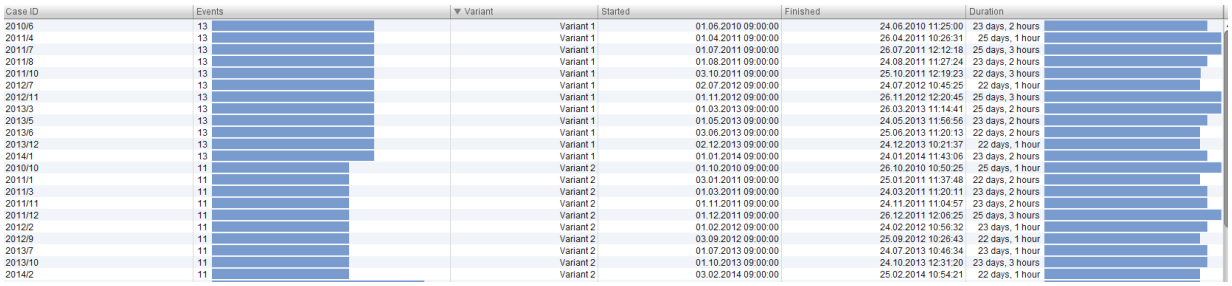
The diagram showing the median times provides an explanation for the repetition of the activities. For example, the activity “Engaging alternative suppliers” usually takes about 80 minutes, compared to the activities with the longest duration - over 2.3 hours, it is relatively low, but the fact that it has a very high number of repetitions in total increases its total duration. This graph also shows the most frequent waiting times. The highest among the process instances are about eight days.



All in all the process graphs are a very useful tool to identify the pain points in the process. For the procurement process, the are definitely related to the activities for engaging alternative suppliers. There are the most repetitions, most time consumption and the longest waiting times.

Another useful insight is that there are eleven versions of the process that contain between 7 and 19 events. 44% of the cases are represented by variant 1 and 2, containing respectively, 13 and 11 events. Therefore it can be concluded that in the most of the cases an average number of events take place.

Statistics of the most frequent variants: 1 & 2; 44% of the times:



| Case ID | Events | Variant | Started | Finished | Duration |
|---|---|---|---|---|---|
| 2010/6 | 13 | Variant 1 | 01.06.2010 09:00:00 | 24.06.2010 11:25:00 | 23 days, 2 hours |
| 2011/4 | 13 | Variant 1 | 01.04.2011 09:00:00 | 26.04.2011 10:26:31 | 25 days, 1 hour |
| 2011/7 | 13 | Variant 1 | 01.07.2011 09:00:00 | 26.07.2011 12:12:18 | 25 days, 3 hours |
| 2011/8 | 13 | Variant 1 | 01.08.2011 09:00:00 | 24.08.2011 11:27:24 | 23 days, 2 hours |
| 2011/10 | 13 | Variant 1 | 03.10.2011 09:00:00 | 25.10.2011 12:19:23 | 22 days, 3 hours |
| 2012/7 | 13 | Variant 1 | 02.07.2012 09:00:00 | 24.07.2012 10:45:25 | 22 days, 1 hour |
| 2012/11 | 13 | Variant 1 | 01.11.2012 09:00:00 | 26.11.2012 12:20:45 | 25 days, 3 hours |
| 2013/3 | 13 | Variant 1 | 01.03.2013 09:00:00 | 26.03.2013 11:14:41 | 25 days, 2 hours |
| 2013/5 | 13 | Variant 1 | 01.05.2013 09:00:00 | 24.05.2013 11:56:56 | 23 days, 2 hours |
| 2013/6 | 13 | Variant 1 | 03.06.2013 09:00:00 | 25.06.2013 11:20:13 | 22 days, 2 hours |
| 2013/12 | 13 | Variant 1 | 02.12.2013 09:00:00 | 24.12.2013 10:21:37 | 22 days, 1 hour |
| 2014/1 | 13 | Variant 1 | 01.01.2014 09:00:00 | 24.01.2014 11:43:06 | 23 days, 2 hours |
| 2010/10 | 11 | Variant 2 | 01.10.2010 09:00:00 | 26.10.2010 10:50:25 | 25 days, 1 hour |
| 2011/1 | 11 | Variant 2 | 03.01.2011 09:00:00 | 25.01.2011 11:37:48 | 22 days, 2 hours |
| 2011/3 | 11 | Variant 2 | 01.03.2011 09:00:00 | 24.03.2011 11:20:11 | 23 days, 2 hours |
| 2011/11 | 11 | Variant 2 | 01.11.2011 09:00:00 | 24.11.2011 11:04:57 | 23 days, 2 hours |
| 2011/12 | 11 | Variant 2 | 01.12.2011 09:00:00 | 26.12.2011 12:06:25 | 25 days, 3 hours |
| 2012/2 | 11 | Variant 2 | 01.02.2012 09:00:00 | 24.02.2012 10:56:32 | 23 days, 1 hour |
| 2012/9 | 11 | Variant 2 | 03.09.2012 09:00:00 | 25.09.2012 10:26:43 | 22 days, 1 hour |
| 2013/7 | 11 | Variant 2 | 01.07.2013 09:00:00 | 24.07.2013 10:46:34 | 23 days, 1 hour |
| 2013/10 | 11 | Variant 2 | 01.10.2013 09:00:00 | 24.10.2013 12:31:20 | 23 days, 3 hours |
| 2014/2 | 11 | Variant 2 | 03.02.2014 09:00:00 | 25.02.2014 10:54:21 | 22 days, 1 hour |

All activities in a sequential way in Variant 1: 12 cases (24%)

| | Activity | Resource | Date | Time | Duration |
|---|---|---|---|---|---|
| 1 | Checking reserved quotas (calling suppliers at the beginning of a month) | 73 | 01.06.2010 | 09:00:00 | 3 hours, 29 mins |
| 2 | Engaging alternative supplier | 73 | 01.06.2010 | 12:29:02 | 1 hour, 1 min |
| 3 | Drafting contract notes re penalties etc. | 96 | 01.06.2010 | 12:29:02 | 16 mins, 26 secs |
| 4 | Engaging alternative supplier | 88 | 01.06.2010 | 12:29:02 | 1 hour, 35 mins |
| 5 | Drafting contract notes re penalties etc. | 77 | 01.06.2010 | 12:29:02 | 22 mins, 51 secs |
| 6 | Engaging alternative supplier | 34 | 01.06.2010 | 12:29:02 | 1 hour, 19 mins |
| 7 | Drafting contract notes re penalties etc. | 67 | 01.06.2010 | 12:29:02 | 25 mins, 6 secs |
| 8 | Transferring data from the Sales-sheet to OEM compliant list | 5 | 09.06.2010 | 21:36:27 | 37 mins, 27 secs |
| 9 | Comparing the monthly figures with the forecast | 96 | 09.06.2010 | 22:13:54 | 2 hours, 2 mins |
| 10 | Transferring data from Production-PDF to Procurement-PSD | 34 | 10.06.2010 | 00:16:41 | 31 mins, 32 secs |
| 11 | Conducting the PSD planning procedure (automated process) | 110 | 10.06.2010 | 09:00:00 | 2 hours, 25 mins |
| 12 | Transferring data from Production-PDF to Procurement-PSD | 88 | 10.06.2010 | 14:41:43 | 27 mins, 34 secs |
| 13 | Calculating negative impacts, if no alternatives can be found | 5 | 24.06.2010 | 09:00:00 | 2 hours, 25 mins |

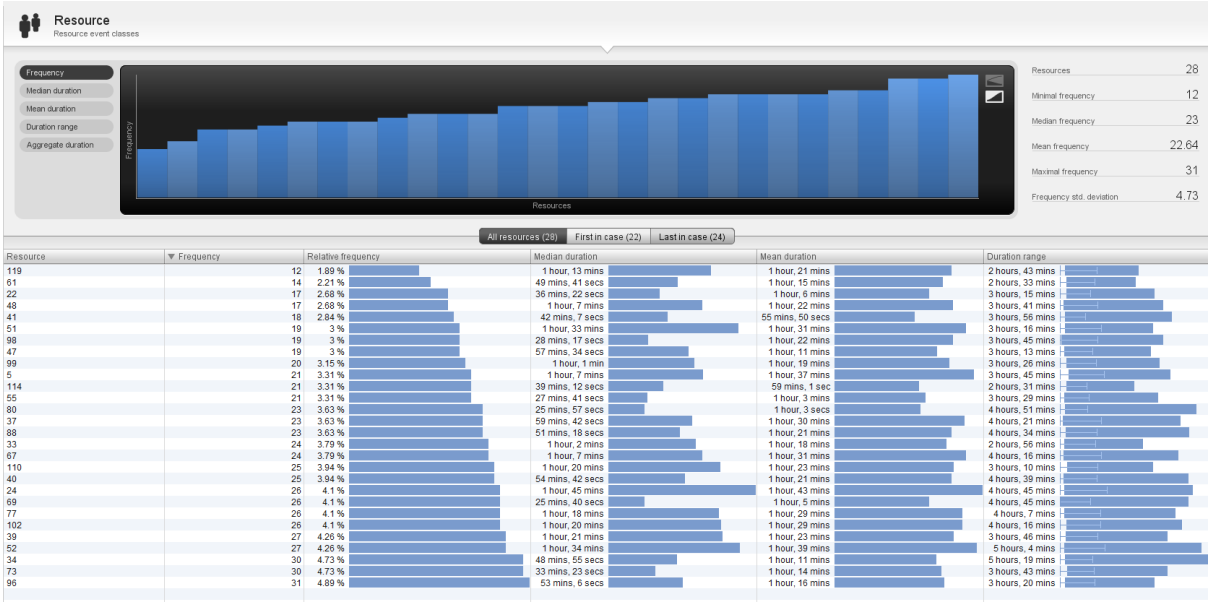
All activities in a sequential way in Variant 2: 10 cases (20%)

| | Activity | Resource | Date | Time | Duration |
|---|---|---|---|---|---|
| 1 | Checking reserved quotas (calling suppliers at the beginning of a month) | 37 | 01.10.2010 | 09:00:00 | 3 hours, 33 mins |
| 2 | Engaging alternative supplier | 34 | 01.10.2010 | 12:33:43 | 1 hour, 28 mins |
| 3 | Drafting contract notes re penalties etc. | 98 | 01.10.2010 | 12:33:43 | 3 mins, 3 secs |
| 4 | Engaging alternative supplier | 47 | 01.10.2010 | 12:33:43 | 1 hour, 11 mins |
| 5 | Drafting contract notes re penalties etc. | 69 | 01.10.2010 | 12:33:43 | 19 mins, 37 secs |
| 6 | Transferring data from the Sales-sheet to OEM compliant list | 34 | 11.10.2010 | 21:17:50 | 56 mins, 19 secs |
| 7 | Comparing the monthly figures with the forecast | 61 | 11.10.2010 | 22:14:09 | 2 hours, 5 mins |
| 8 | Transferring data from Production-PDF to Procurement-PSD | 34 | 12.10.2010 | 00:19:52 | 17 mins, 26 secs |
| 9 | Conducting the PSD planning procedure (automated process) | 80 | 12.10.2010 | 09:00:00 | 5 hours, 3 mins |
| 10 | Transferring data from Production-PDF to Procurement-PSD | 24 | 12.10.2010 | 16:03:08 | 23 mins, 45 secs |
| 11 | Calculating negative impacts, if no alternatives can be found | 102 | 26.10.2010 | 09:00:00 | 1 hour, 50 mins |

In both variants, the activities “Engaging alternative supplier” and “Drafting contract notes” are repeated - 4 and 2 times respectively. This further confirms that this part of the process should be improved.

Moreover, statistics per resource are also available. In the procurement process, the employees are marked as a resource. The table below shows the frequency and duration per resource/employee. The employees with the highest frequency are 96; 73; 34. On the other hand, the ones with lowest are 61 and 119.

Frequency and duration per employee:




| Resource | Frequency | Relative frequency | Median duration | Mean duration | Duration range |
|---|---|---|---|---|---|
| 119 | 12 | 1.89 % | 1 hour, 13 mins | 1 hour, 21 mins | 2 hours, 43 mins |
| 61 | 14 | 2.21 % | 49 mins, 41 secs | 1 hour, 15 mins | 2 hours, 33 mins |
| 22 | 17 | 2.68 % | 36 mins, 22 secs | 1 hour, 6 mins | 3 hours, 15 mins |
| 48 | 17 | 2.68 % | 1 hour, 7 mins | 1 hour, 22 mins | 3 hours, 41 mins |
| 41 | 18 | 2.84 % | 42 mins, 7 secs | 55 mins, 50 secs | 3 hours, 56 mins |
| 51 | 19 | 3 % | 1 hour, 33 mins | 1 hour, 31 mins | 3 hours, 16 mins |
| 98 | 19 | 3 % | 28 mins, 17 secs | 1 hour, 22 mins | 3 hours, 45 mins |
| 47 | 19 | 3 % | 57 mins, 34 secs | 1 hour, 11 mins | 3 hours, 13 mins |
| 99 | 20 | 3.15 % | 1 hour, 1 min | 1 hour, 19 mins | 3 hours, 26 mins |
| 5 | 21 | 3.31 % | 1 hour, 7 mins | 1 hour, 37 mins | 3 hours, 45 mins |
| 114 | 21 | 3.31 % | 39 mins, 12 secs | 59 mins, 1 sec | 2 hours, 31 mins |
| 55 | 21 | 3.31 % | 27 mins, 41 secs | 1 hour, 3 mins | 3 hours, 29 mins |
| 80 | 23 | 3.63 % | 25 mins, 57 secs | 1 hour, 3 secs | 4 hours, 51 mins |
| 37 | 23 | 3.63 % | 59 mins, 42 secs | 1 hour, 30 mins | 4 hours, 21 mins |
| 88 | 23 | 3.63 % | 51 mins, 18 secs | 1 hour, 21 mins | 4 hours, 34 mins |
| 33 | 24 | 3.79 % | 1 hour, 2 mins | 1 hour, 18 mins | 2 hours, 56 mins |
| 67 | 24 | 3.79 % | 1 hour, 7 mins | 1 hour, 31 mins | 4 hours, 16 mins |
| 110 | 25 | 3.94 % | 1 hour, 20 mins | 1 hour, 23 mins | 3 hours, 10 mins |
| 40 | 25 | 3.94 % | 54 mins, 42 secs | 1 hour, 21 mins | 4 hours, 39 mins |
| 24 | 26 | 4.1 % | 1 hour, 45 mins | 1 hour, 43 mins | 4 hours, 45 mins |
| 69 | 26 | 4.1 % | 25 mins, 40 secs | 1 hour, 5 mins | 4 hours, 45 mins |
| 77 | 26 | 4.1 % | 1 hour, 18 mins | 1 hour, 29 mins | 4 hours, 7 mins |
| 102 | 26 | 4.1 % | 1 hour, 20 mins | 1 hour, 29 mins | 4 hours, 16 mins |
| 39 | 27 | 4.26 % | 1 hour, 21 mins | 1 hour, 23 mins | 3 hours, 46 mins |
| 52 | 27 | 4.26 % | 1 hour, 34 mins | 1 hour, 39 mins | 5 hours, 4 mins |
| 34 | 30 | 4.73 % | 48 mins, 55 secs | 1 hour, 11 mins | 5 hours, 19 mins |
| 73 | 30 | 4.73 % | 33 mins, 23 secs | 1 hour, 14 mins | 3 hours, 43 mins |
| 96 | 31 | 4.89 % | 53 mins, 6 secs | 1 hour, 16 mins | 3 hours, 20 mins |

# Business recommendations

After creating the proof of concept and solving the problems, we would like to give the procurement department some business recommendations for the future.

First, it is suggested to introduce some OLTP adjustments such as the use of integrity rules and constraints in the database in order to have a consistent basis for the Data Mart. This way, one could avoid incorrect entries and standardize the data, allowing for accurate aggregation and summarization in the later multidimensional analysis. Furthermore, it is imperative to improve the current database design, for example, by parsing the single column “address” in the *Supplier* table into five separate fields (street, zip code, city, region, country). Additionally, it is recommended to improve the data administration, especially when it comes to order and delivery date, as we found out that these dates often have been identical. By introducing those changes, it would be possible to derive some interesting KPIs such as “Delivery time” or “Delivery reliability” to get a better overview of all supplier’s issues.

What is more, we would like to encourage the use of presented analytical insights for the daily business, which allows for comparison of material prices per supplier over time which helps to decide when to buy specific materials. To make even better decisions, one could go a step further by using predictive analytics in order to predict prices for the near future.

Finally, we would recommend improving the efficiency of the purchasing process as we identified some bottlenecks and long waiting times at individual process steps. It is suggested to dig deeper to understand the reasons for this and find out a way to avoid this in the future.