

Design of a system for visualizing trends and behaviors based on customer data

Design av ett system för visualisering av trender och beteenden baserat på kunddata.

Oskar Andersson

Supervisor: Erik Berglund

Examiner: Anders Fröberg

Upphovsrätt

Detta dokument hålls tillgängligt på Internet – eller dess framtida ersättare – under 25 år från publiceringsdatum under förutsättning att inga extraordinära omständigheter uppstår.

Tillgång till dokumentet innebär tillstånd för var och en att läsa, ladda ner, skriva ut enstaka kopior för enskilt bruk och att använda det oförändrat för ickekommersiell forskning och för undervisning. Överföring av upphovsrätten vid en senare tidpunkt kan inte upphäva detta tillstånd. All annan användning av dokumentet kräver upphovsmannens medgivande. För att garantera äktheten, säkerheten och tillgängligheten finns lösningar av teknisk och administrativ art.

Upphovsmannens ideella rätt innefattar rätt att bli nämnd som upphovsman i den omfattning som god sed kräver vid användning av dokumentet på ovan beskrivna sätt samt skydd mot att dokumentet ändras eller presenteras i sådan form eller i sådant sammanhang som är kränkande för upphovsmannens litterära eller konstnärliga anseende eller egenart.

För ytterligare information om Linköping University Electronic Press se förlagets hemsida <https://ep.liu.se/>.

Copyright

The publishers will keep this document online on the Internet – or its possible replacement – for a period of 25 years starting from the date of publication barring exceptional circumstances.

The online availability of the document implies permanent permission for anyone to read, to download, or to print out single copies for his/hers own use and to use it unchanged for non-commercial research and educational purpose. Subsequent transfers of copyright cannot revoke this permission. All other uses of the document are conditional upon the consent of the copyright owner. The publisher has taken technical and administrative measures to assure authenticity, security and accessibility.

According to intellectual property law the author has the right to be mentioned when his/her work is accessed as described above and to be protected against infringement.

For additional information about the Linköping University Electronic Press and its procedures for publication and for assurance of document integrity, please refer to its www home page: <https://ep.liu.se/>.

Abstract

Big amounts of data are produced every day in companies. By analyzing and visualizing the data a lot of insights can be gained. The company Solution Xperts wanted to create a system that could import and visualize Big Data. In this work a system was created and evaluated. The report shows that it can be difficult to visualize Big Data, but when a system is created it can easily be adapted to data coming from different companies and provide a lot of value to companies and organizations.

Acknowledgement

I would like to thank my supervisor Erik Berglund and my examiner Anders Fröberg for their support and guidance during this work. I also want to thank the team at Solution Xperts for being supportive from the very start. All support I have received from both Linköping University and Solution Xperts has been very valuable, especially since the work was performed remotely due to the Corona Pandemic.

Table of Contents

Design of a system for visualizing trends and behaviors based on customer data	0
Abstract	3
Acknowledgement	4
Figures.....	7
1. Introduction.....	8
1.1 Motivation.....	8
1.2 Research Question	8
1.3 Figures use dummy data	8
2. Theory.....	8
2.1 Big Data	8
2.1.1 Introduction and definition.....	8
2.1.2 The “5 Vs”	9
2.1.3 The value of Big Data	9
2.2 Analyzing Big Data.....	9
2.2.1 Visual analysis tools	9
2.2.2 Sampling and clustering.....	10
2.3 Usages of Business Intelligence	11
2.4 Power BI	11
2.4.1 Visualization.....	12
2.4.2 Reports and Dashboards.....	12
2.5 Visualizing Big Data	12
2.6 Interview questions.....	12
3. Method	13
3.1 Visual analysis tool	13
3.2 The work process	13
3.3 Demonstration and interview	14
4. Results.....	15
4.1 The work process	15
4.1.1 Creating views based on figure 1 and 2.....	15
4.1.2 Creating a view based on figure 3.....	16
4.1.3 Writing SQL queries	16
4.1.4 Measures	16
4.1.5 Features of Power BI.....	17

4.1.6 Creating visualizations	20
4.2 The demonstration and the interview	22
5. Discussion	23
5.1 Results	23
5.1.1 The work process	24
5.1.2 The demonstration and the interview	24
5.2 Trends and behaviors	24
5.3 Replicability	25
5.4 Future work	25
6. Conclusion	26
References	27

Figures

Figure 1: Old Flow Tracking view.

Figure 2: Old Interchange view.

Figure 3: The third old view.

Figure 4: First Artifact measure.

Figure 5: Last Artifact measure.

Figure 6: SQL query.

Figure 7: Invoices per day measure.

Figure 8: Number of invoices measure.

Figure 9: Duration measure.

Figure 10: The Power Query Editor.

Figure 11: Report view options.

Figure 12: Model view.

Figure 13: Order view.

Figure 14: Order view buttons activated.

Figure 15: Middle Step view.

Figure 16: New Flow Tracking view.

1. Introduction

1.1 Motivation

In today's world almost everything revolves around information. Information is flowing constantly from all kinds of sources and a lot of it is recorded and stored. The company Solution Xperts is a consulting firm that has developed their own solution for storing data from integrations. Their solution is now deployed at some bigger companies and organizations. For some years there has been a wish from some of the companies to make use of the data to better understand their organizations. If the data was visualized in a way that it became easy to understand and if there was a possibility to find trends and behaviors in the data, it could bring a lot of value to these companies and organizations. Today Solution Xperts offer their customers a couple of views that provide some insights based on the customers data, but the goal is to create more comprehensive visualizations that can be interacted with. Depending on the choice of architecture of the system it could become a general visualization product that could easily be implemented for other customers.

1.2 Research Question

How should a system for visualizing customer data be constructed to make it easy to spot trends and discover behaviors from the data?

The research question is delimited to creating a system that is good enough to fulfil the needs of one of Solution Xperts customers. The wishes of the customer are to have views of the data in table form where the customer can drill through to see more details of the data, the possibility to select different parts of the data and have the different views adapt to that selection and the possibility to create new views and calculated values based on their data. To be able to create new views and calculated values the customer will still need some help from consultants at Solution Xperts since it will require some coding at this point.

1.3 Figures use dummy data

All figures showing data in this report contain dummy data.

2. Theory

2.1 Big data

2.1.1 Introduction and definition

In our digitalized world there is a lot of data created and stored every day. The data can come from production processes and different communication channels. This data is often unstructured and consists of different kinds of data [1] [2]. Big Data is the term that is often used when referring to databases containing this kind of data. Defining Big Data is not the easiest thing, but one definition is the following [3]:

“The term Big Data describes a data environment in which scalable architectures support the requirements of analytical and other applications which process, with high velocity, high volume data which may have a variety of data formats and which may include high velocity data acquisition.”

2.1.2 The “5 Vs”

When talking about Big Data it is usually described by the “5 Vs” [1] [4]:

- *Variety* stands for the variation of different kinds of data often found in Big Data. A database can store data such as images, text, videos, and sensor data.
- *Velocity* is about the speed of which data is created, sent, and handled. The importance of velocity varies between different applications. Some applications use stored data while other applications get a constant flow of new information.
- *Volume* means the volume of the data.
- *Value* is about the value you can get from the data. By understanding the data, you can get value from it in many ways. Seeing trends or patterns in for example HR data, employee data, customer data, or order data can help reduce costs and be beneficial to management when taking decisions.
- *Veracity* is about the correctness of the data. Big Data often consists of lots of data coming from different sources. This means there might be data that is incomplete or missing. The data is supposed to be analyzed and therefore it is important to assess its correctness and try to improve its quality.

2.1.3 The value of Big Data

Businesses have realized there is a possibility to get value from their Big Data [5]. To be able to get this value it is important that data collection, data storage, data analysis and data visualization are well thought through. All these parts play an important role in getting value out of Big Data. It is also important to think of the security aspect of using visualization tools on Big Data since there can be sensitive information connected to the visualizations. Big Data analysis can for example help understand when orders are arriving and if some orders should be prioritized for some reason [6] [7]. It can also help optimizing logistics and Human Resources.

2.2 Analyzing Big Data

2.2.1 Visual analysis tools

It is difficult to analyze Big Data in its raw form due to its nature of being unstructured [8]. To address this issue, we can use visual analysis tools. By using visual analysis tools, we gain the ability to easily manipulate and work with the data. This means we can analyze and understand the data by creating visualizations of it. Visualizations could for example be different types of diagrams where related data is placed on different axes. It could also be lists where related data is placed side by side

in different columns. Powerful visual analysis tools can be of help in the entire process of working with Big Data. Cleaning data from different sources that is incomplete or erroneous is a tedious work. This can be done more efficiently by using a visual analysis tool. For a visual analysis tool to be useful it must be flexible and fast, which means it should have a quick response time and few confusing options or widgets that need to be interacted with [9].

Visual analysis tools need to provide the right type of controls for the analyst to be able to specify what data he wants to see [9]. By using the controls, the analyst can perform visualization, filtering, sorting, and derivation. Filtering is important since there are usually only parts of the data set that are interesting to look at. It can be performed in many ways, for example using radio buttons, selecting a set of data to include or exclude, or by using more advanced techniques such as entering queries using a query language.

View manipulation is the next important part where an analyst should be able to select different parts of the data, highlight patterns and drill down to get a more detailed view. Coordinated multiple views are useful when analyzing data. An example can be that an analyst is working with customer data and creates a textual list with the customers most recent purchase records, a map containing data of where the customers live and a pie chart showing how many products are in stock at the central warehouse. If one or a few items are selected in one of the views, the other views will adapt to that selection by for example highlighting the selected items. This is powerful since it makes it easier to see connections between different groups of data and finding patterns in the data. Coordinated multiple views could also be used to filter some data.

A tiled window layout is helpful to organize all views. This means all views and selectors can be viewed at once. When adding new views there are a couple of good alternatives for how the window organization could change. The new view could appear next to the existing ones with minor adjustments. A second alternative is to have the option of adding tabs where you can switch between tabs to see different views.

Visual analysis tools can be quite smart and provide suggestions to analysts of steps to take to make their work easier. The tools can have options to publish visualizations to the web where those with access can interact with it. Visualization tools provide analysts the ability to learn new things about data and share that knowledge with others through sharing visualizations. This means the social aspect of visualization tools is also very important. Not only data scientists need to work with data visualizations [10]. It is important for users with no background in computer science to also be able to interact with data visualizations to gain insights.

2.2.2 Sampling and clustering

Analyzing Big Data can be difficult due to its huge size [2]. By sampling data, a smaller amount of data can be used to get the same analytical results as if all the data would have been used. It is a good idea to use sampling when analyzing Big Data. Clustering is widely used when working with Big Data [11]. The idea of clustering is to put data that shares common traits together in groups. Performing clustering on Big Data can be difficult.

2.3 Usages of Business Intelligence

Business Intelligence (BI) is used to gain insights from data. Looking at raw data coming from a database is not that interesting but if the data for example is visualized in an interactive graph it can provide insights of great value [7]. Using a three-tier architecture (Database layer, Application layer, Presentation layer) makes it hard to provide a Business Intelligence based application supposed to run in real-time. This is due to the difficulty of knowing what the execution times will be in the database layer.

The next generation of BI consist of three kinds of BI:

Operational (Real-time) BI:

The idea is to have a continuous update of data, making it possible to find trends and behaviors in the data that was just created. The hard part of realizing Operational BI is to get an acceptable delay in the update of the data.

Situational BI:

Situational BI is about getting valuable information from external parts. The information could bring new point of views and come from different sources such as the internet. A difficulty of implementing Situational BI is that information found in external sources is often unstructured and need to be added to the already existing database to make it possible to perform an analysis of all the data.

Self-service BI:

Self-service BI means that users not very skilled in IT should be able to create their own reports. The positive parts of Self-service BI are that there are no need to have an IT-expert create reports or work with the data that is to be analyzed since non-technical users can perform it themselves.

2.4 Power BI

Power BI is a visual analysis tool [12]. It consists of a desktop version, a mobile version, and a web version. In Power BI you can import many types of data. It can be Excel documents, data from webpages, SQL-databases, and data from web-services such as Azure. When data is imported, you can choose to do operations on the data before it is loaded into the program in a separate window called the Query editor. You can for example choose to filter data or enter an SQL query that determines how the data is imported.

There is a tab in Power BI called model that contains a visual overview of all relations of the data. When a new project is started relations are automatically created by Power BI. Sometimes Power BI sets up relations that should not be there but in that case it is easy to remove them. Relations can also be added manually. New columns can be created using the language DAX. Sometimes you do not want to create a new column of stored data. It could be that you want to make calculations on the data or that you do not want to increase the file size of your report. In that case you use something called a measure instead.

2.4.1 Visualization

There are many different visual components that can be chosen and customized to show your data. Some of the more common ones are tables and column charts. When a visualization component is chosen you can choose what data should be visualized. Data from different columns and tables can be combined into one visual component. Filters can be used to choose more precisely what should be shown. Many visual components can be placed side by side on the same report and visualize the same or part of the same data in different kinds of visualization components. If a row or a data point is selected in one of the components the rest will adapt, just like coordinated multiple views that were mentioned in part 2.2.

2.4.2 Report and Dashboard

A report is the visual page you create using the visual components. Along with diagrams, buttons can be added that perform an action of your choice. Reports usually have many pages that visualize various things or act as differently detailed layers. You can for example create one layer that acts as an overview containing a selected number of data records. From this overview an action called drill through lets the user select one record and “drill through” to another page of the report that displays more details about that record. When the report is created the next step is to upload it to the web version of Power BI. Parts from uploaded reports can be chosen to be part of a dashboard. When the dashboard is finished it can be shared. Users can interact with dashboards by for example selecting a record and make the visualization components show data regarding the selected record.

2.5 Visualizing Big Data

Visualizing data in graphs makes it easy to spot trends and behaviors in Big Data [13]. Without visualization it is very hard understand the data. Choosing the right kind of graph for different kinds of data is important since different kinds of data should be presented in different ways. It is also important to think of ways to reduce or distribute the amount of processing needed to visualize Big Data. One approach to visually look at data is to use the hierarchical approach [14]. The hierarchical approach offers an overview where you can use actions such as drill-down or zoom to look at more details of the data. This is very similar to how Power BI is used [12]. To be able to visualize big sets of data not only filters and zoom are required [15]. If a set of data contains billions of records it is impossible to show them all on the screen, therefore aggregate markers showing for example thousands in one point become necessary. Modern visual analysis tools handle this well.

2.6 Interview questions

The interview questions will be open-ended since the purpose of the interview is to get the opinions of the test user [16]. By using open-ended questions, the test user is not limited when thinking about what kind of functionality he wishes the system to have. This could lead to wishes for functionality that are not possible to implement in the created system, but that could be solved by for example changing the design of the system.

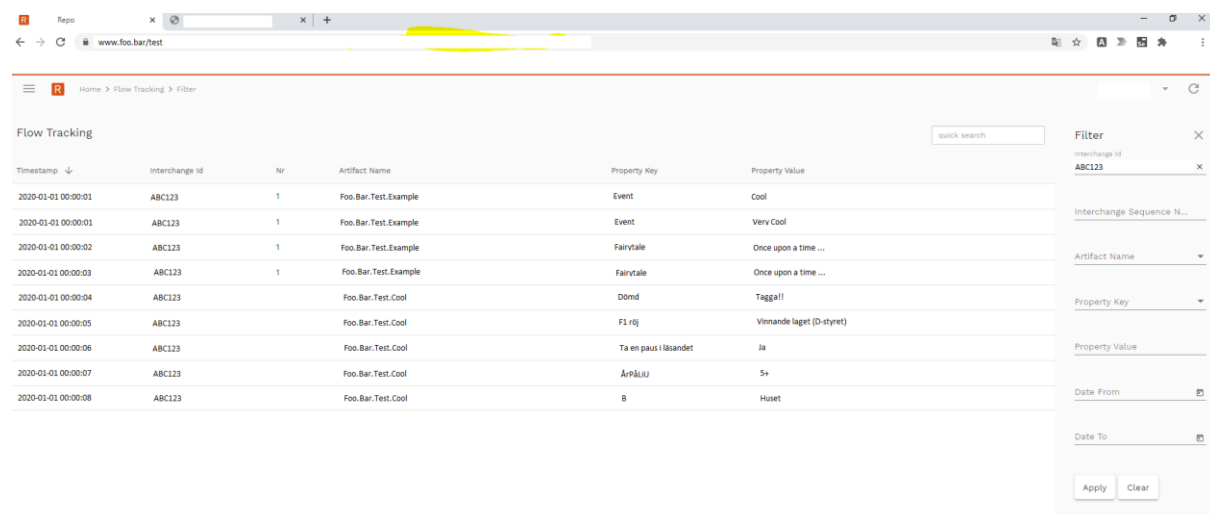
3. Method

3.1 Visual analysis tool

Solution Xperts have experience working with the visual analysis tool Power BI. This meant that they could provide a good support throughout the development of the system if Power BI was chosen. Power BI seemed to have the required properties to develop the system, so it became the tool of choice for this work. The focus at the start of the work was to get to know Power BI. By recommendation of experienced Power BI developers at Solution Xperts, the learning started by watching tutorials of Power BI on YouTube, followed by reading articles and creating a simple prototype. Other tools that were helpful to use when creating SQL queries and DAX expressions were SSMS (Microsoft SQL Server Management Studio) and DAX Studio. After getting to know Power BI the work on the system began.

3.2 The work process

The work was based on three figures (figure 1,2, and 3) showing the visual appearance of the old system accompanied by written instructions of its functionality and the connections between different views. The work process was performed in an iterative manner where one or a couple of smaller parts of the system were created and tested at a time. When one part was finished, it was shown to the supervisor at Solution Xperts for feedback. This way of working always ensured the work to be performed in the right direction. When the new system was finished, meaning the visual appearance was right, all functionality that was asked for was in place and all views were connected, the system was demonstrated to a person at Solution Xperts that had experience in using the old system.



Timestamp	Interchange Id	Nr	Artifact Name	Property Key	Property Value
2020-01-01 00:00:01	ABC123	1	Foo.Bar.Test.Example	Event	Cool
2020-01-01 00:00:01	ABC123	1	Foo.Bar.Test.Example	Event	Very Cool
2020-01-01 00:00:02	ABC123	1	Foo.Bar.Test.Example	Fairytale	Once upon a time ...
2020-01-01 00:00:03	ABC123	1	Foo.Bar.Test.Example	Fairytale	Once upon a time ...
2020-01-01 00:00:04	ABC123		Foo.Bar.Test.Cool	Domd	Tagga!!
2020-01-01 00:00:05	ABC123		Foo.Bar.Test.Cool	F1 r0j	Vinnande laget (D-styret)
2020-01-01 00:00:06	ABC123		Foo.Bar.Test.Cool	Ta en paus i l0sandet	Ja
2020-01-01 00:00:07	ABC123		Foo.Bar.Test.Cool	Är plöj	5+
2020-01-01 00:00:08	ABC123		Foo.Bar.Test.Cool	B	Huset

Figure 1: Old Flow Tracking view.

Timestamp	Interchange Id	Type	Partner Id	First Artifact	Last Artifact
2020-01-01 00:00:01	abc123	Three		One.Two.Three.Four	You.Have.Three.Cereals
2020-01-01 00:00:01	abc1234	Three		One.Two.Three.Four	You.Have.Three.Cereals
2020-01-01 00:00:02	abc12345	Three		One.Two.Three.Four	You.Have.Three.Cereals
2020-01-01 00:00:03	abc123456	Example		Bar.Foo.Example.Test	Bar.Foo.Example.Cool
2020-01-01 00:00:04	abcd223	Example		Bar.Foo.Example.Test	Bar.Foo.Example.Cool
2020-01-01 00:00:05	abc323	Example		Bar.Foo.Example.Test	Bar.Foo.Example.Cool
2020-01-01 00:00:06	abc554	Test		Foo.Bar.Test.Example	Foo.Bar.Test.Cool
2020-01-01 00:00:07	abcde123456	Test		Foo.Bar.Test.Example	Foo.Bar.Test.Cool
2020-01-01 00:00:08	abc6767	Test		Foo.Bar.Test.Example	Foo.Bar.Test.Cool
2020-01-01 00:00:09	abc65432	Test		Foo.Bar.Test.Example	Foo.Bar.Test.Cool

Figure 2: Old Interchange view.

Timestamp	Partner Id	Customer Number	Customer Order Number	Order Number	Order Response	Invoice Number	Delivery Number	Temporary Order Number
2020-10-02 13:37:37	Foo Co	123456	1234	123456	N/A	2342523		234234
2020-10-03 13:37:37	Bar Co	234567	12345	10000001	OrderConfirmation			233333
2020-10-04 13:37:39	Test Co	345678	123456	100000012	OrderConfirmation	123456	123123	455545
2020-10-05 13:38:38	Test Co	45678	1234567	10000002	OrderConfirmation	123456	123124	54321
2020-10-05 13:38:40	Test Co	567890	123123	2000004	OrderConfirmation	123456	123125	234123
2020-10-10 13:37:37	Test Co	01234	123412	2354234	OrderConfirmation	123456	123126	789987

Figure 3: The third old view.

3.3 Demonstration and interview

The demonstration was combined with an interview. Due to the Corona pandemic the demonstration and the interview were conducted through Microsoft Teams. One person familiar with the old system was selected for the demonstration and the interview. This person has an extensive knowledge about the old system and a good idea of what functionality would be useful in the new system. During the demonstration, the person was guided through the system. The demonstration started by going through each tab and showing the functionality of its added filters, buttons, and diagrams. It was also showed that if one specific order was selected, all diagrams and tables adapted to that selection and showed information about that order. The next step was to drill through to the Middle Step view (figure 15). This view showed all ID:s associated with the selected order. One ID was selected in the Middle Step view and a drill through was performed to the Flow Tracking view (figure 16) where details about the selected ID were presented.

The test person was encouraged to ask questions along the way and some of the interview questions got discussed and answered before we reached the final interview step. When the guiding of the system was done, I asked the following open-ended questions to get the opinions of the test person:

1. What do you think about this new system?
2. Is there any specific functionality you would like this system to have?
3. Do you have any other thoughts or questions?

4. Results

4.1 The work process

4.1.1 Creating views based on figure 1 and 2

The views were created using Power BI. At the start the Flow Tracking and Interchange views were created (see figure 1 and 2). The data was retrieved from an SQL database where some values already were separated into their own columns while other values needed to be extracted or calculated to be visualized. In the Flow Tracking view all values that were supposed to be visualized already had their own column in the database, meaning they could easily be selected and visualized in Power BI. The new Flow Tracking view can be seen in figure 16. The Interchange view (figure 2) was a bit more complicated to create. In this view we wanted to present the first and the last value of a certain kind where the first value was the one connected to the first event of an Interchange and the last value was connected to the last event of the same Interchange. To achieve this, a measure (a column with calculated data in Power BI) was created using the language DAX, see figure 4 and 5. In figure 4 the first artifact was chosen by filtering artifact names by the minimum timestamp which means the earliest timestamp. Due to how measures are constructed a table column cannot be the expression to be calculated in a calculate function. Fortunately, there are ways around this. Using a max- or a min- function (it does not matter which one is chosen in this case) on the table column (row 4 in figure 4) fixes the problem. The same idea was applied for getting the last artifact in figure 5 by getting the artifact name with the maximum value on timestamp.

```
1 FirstArtifact =  
2 var mindate = MIN('FlowTracking'[Timestamp])  
3 return  
4 CALCULATE(MAX('FlowTracking'[ArtifactName]),FILTER(FlowTracking, FlowTracking[Timestamp]=mindate))
```

Figure 4: First Artifact measure.

```
1 LastArtifact =  
2 var last_date = MAX('FlowTracking'[TimeStamp])  
3 return  
4 CALCULATE(MAX('FlowTracking'[ArtifactName]),FILTER(FlowTracking, FlowTracking[Timestamp]=last_date))
```

Figure 5: Last Artifact measure.

4.1.2 Creating a view based on figure 3

To create the functionality of figure 3 some more operations needed to be performed on the data. To be able to visualize the information in the columns named Customer, Customer Order, Response, and Invoice in figure 3 an operation called pivot needed to be performed. The pivot function groups information of the same kind together from a column and creates new columns for the different groups of information. An example could be a column that contains the names of different fruits such as banana, orange, and apple. If you want to get a better overview of all rows containing the value apple in the fruit column you may want to extract all apple values out of the fruit column and create an apple column. That operation is a pivot operation, and it can be done for a lot of different values coming from the same column. Power BI has a built-in function for pivoting and unpivoting rows in columns but for some reason it did not work in this case. Luckily, you can enter SQL queries in Power BI, so the pivot function was created in SQL. The SQL query that was used for pivoting the rows is seen in figure 6 below. The query is restricted by not returning null values of TemporaryOrderNumber and by only returning data that is later than 2020-07-01. This is mostly due to what data is interesting to look at, but it also reduces the amount of data being loaded into Power BI.

```
;with cte_ as (
Select distinct PropertyValue as TemporaryOrderNumber, InterchangeID from FlowTracking
where PropertyKey = 'TemporaryOrderNumber' and [Timestamp] > '2020-07-01' and isnull(PropertyValue ,'' )<>' '
)
select  TemporaryOrderNumber,
        TenantId,
        Min(Timestamp) as 'Timestamp',
        Min(CreatedDate) as CreatedDate,
        Min(Case PropertyKey When 'CustomerNumber' Then PropertyValue End) CustomerNumber,
        Min(Case PropertyKey When 'SenderPartyName' Then PropertyValue End) SenderPartyName,
        Min(Case PropertyKey When 'CustomerOrderNumber' Then PropertyValue End) CustomerOrderNumber,
        Min(Case PropertyKey When 'OrderNumber' Then PropertyValue End) OrderNumber,
        Min(Case PropertyKey When 'OrderResponseType' Then PropertyValue End) OrderResponseType,
        Min(Case PropertyKey When 'InvoiceNumber' Then PropertyValue End) InvoiceNumber,
        Min(Case PropertyKey When 'DeliveryNumber' Then PropertyValue End) DeliveryNumber
from FlowTracking f
inner join cte_ c on c.InterchangeID = f.InterchangeID
group by TemporaryOrderNumber, TenantId
order by 1 desc
```

Figure 6: SQL query.

4.1.3 Writing SQL queries

It was not very efficient to develop and evaluate SQL queries directly in Power BI since it had no good interface for writing queries. Every change to the queries also required a refresh of the data, which is a heavy and slow operation. A better solution was to use SSMS which is a good application for writing SQL queries. In SSMS you can run your queries on the top 1000 rows of a table in the database which makes testing your queries an easy and quick operation. Once an SQL query was ready to be used it was simply copied into Power BI.

4.1.4 Measures

Measures share some common traits with functions found in programming languages. A measure can for example be used inside another measure, just like a function calls another function. This is a smart feature making it possible to reduce the need of code duplication and making it possible to

arrange the measures conveniently. An example of a measure used inside another measure is seen in figure 7 where “Antal fakturor” is the name of the measure seen in figure 8.

```
1 Antal fakturor per dag = ceiling([Antal fakturor]/DATEDIFF(min(Interchanges[Timestamp]),max(FlowTracking[Timestamp]),DAY), 1)
```

Figure 7: Invoices per day measure.

```
1 Antal fakturor = DISTINCTCOUNT(Interchanges[InvoiceNumber])
```

Figure 8: Number of invoices measure.

There was a wish to see how long interchanges were running. To create that functionality a measure was needed to be created. The resulting measure called Duration can be seen in figure 9. To get the time difference between the first and the last event of an interchange a built-in function called DATEDIFF was used. The time difference was calculated as seconds and then run through an algorithm that split the seconds into days, hours, minutes, and seconds. At row seven the result was created and saved in a variable called res1. An if-statement seen at row eight checked the result and returned an empty string if there were no time difference at all and otherwise it returned the duration. In the report all rows that had a duration bigger than zero showed their duration while the rows that had a duration of zero had empty columns.

```
1 Duration =
2 var seconds = DATEDIFF(min(Interchanges[Timestamp]),max(FlowTracking[Timestamp]),SECOND)
3 var Day_ = INT(seconds/(24*60*60))
4 var Hours_ = MOD(INT(seconds/(60*60)),24)
5 var Min_ = MOD(INT(seconds/60),60)
6 var Sec_ = MOD(seconds,60)
7 var res1 = Day_ & "d " & right("0"&Hours_, 2) & ":" & right("0"&Min_, 2) & ":" & right("0"&Sec_, 2)
8 var res2 = if(res1 == "d 0:0:0", " ", res1)
9 return res2
```

Figure 9: Duration measure.

4.1.5 Features of Power BI

The Power Query Editor is the part of Power Bi where data is imported and prepared (figure 10). It had some features that were helpful. Data could be prepared by for example unselecting and thereby removing unwanted values. In the right part of figure 10 there was a box called Applied Steps where all changes were logged. Selecting an earlier step made it possible to see what the data looked like at that step, making it easy to go back through filtering, removals, name changes and so on. This was helpful for reminding yourself of how to perform certain steps and it also functioned as a check list of already performed steps.

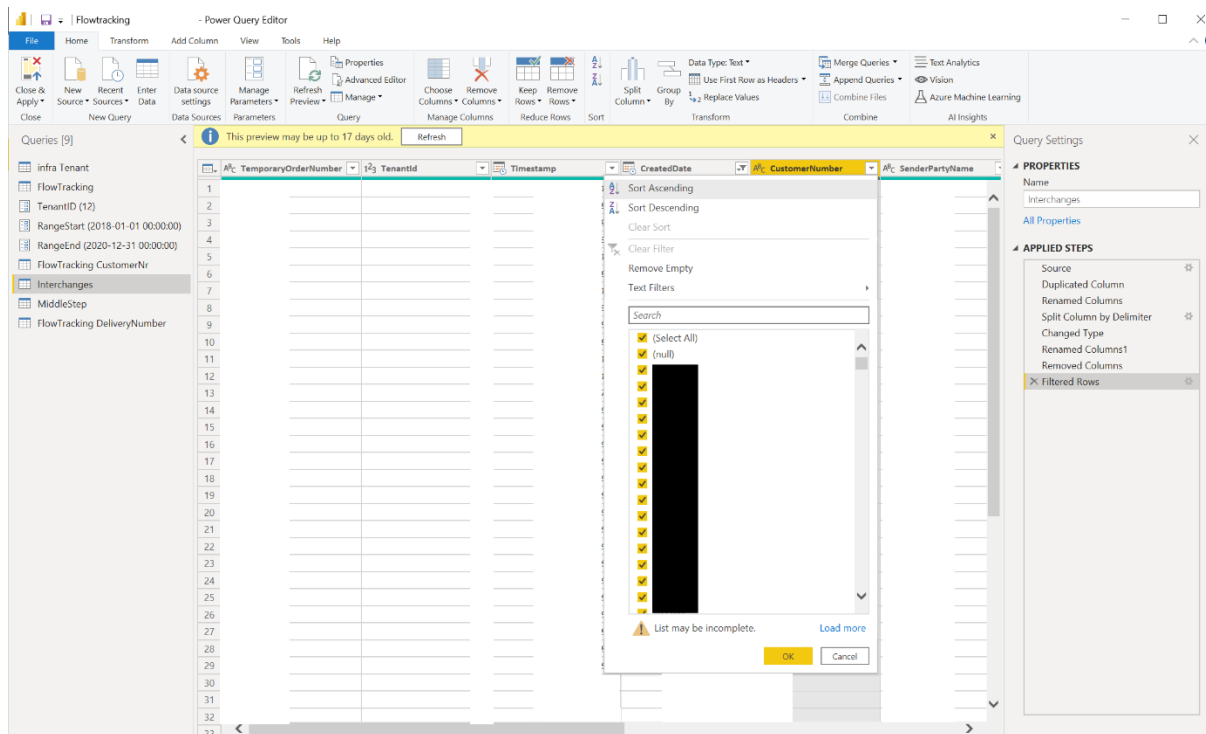


Figure 10: The Power Query Editor.

In the right part of the report view there are a lot of available options (see figure 11). Filters (left part of figure 11) can be applied to a specified visual, to the page, or to all pages. SenderPartyName was added as a filter for the page in figure 11. SenderPartyName was also selected as a filter for all pages, meaning the whole report. In this case the filter for SenderPartyName was added twice which was unnecessary but shows the concept of filters on different levels. If we take another look at the bottom of the filter panel, we see some of the options that can be selected for a filter, for example show everything except blank values. Filters have proved to be very useful during this work due to the many possibilities to filter on dates, values and so on.

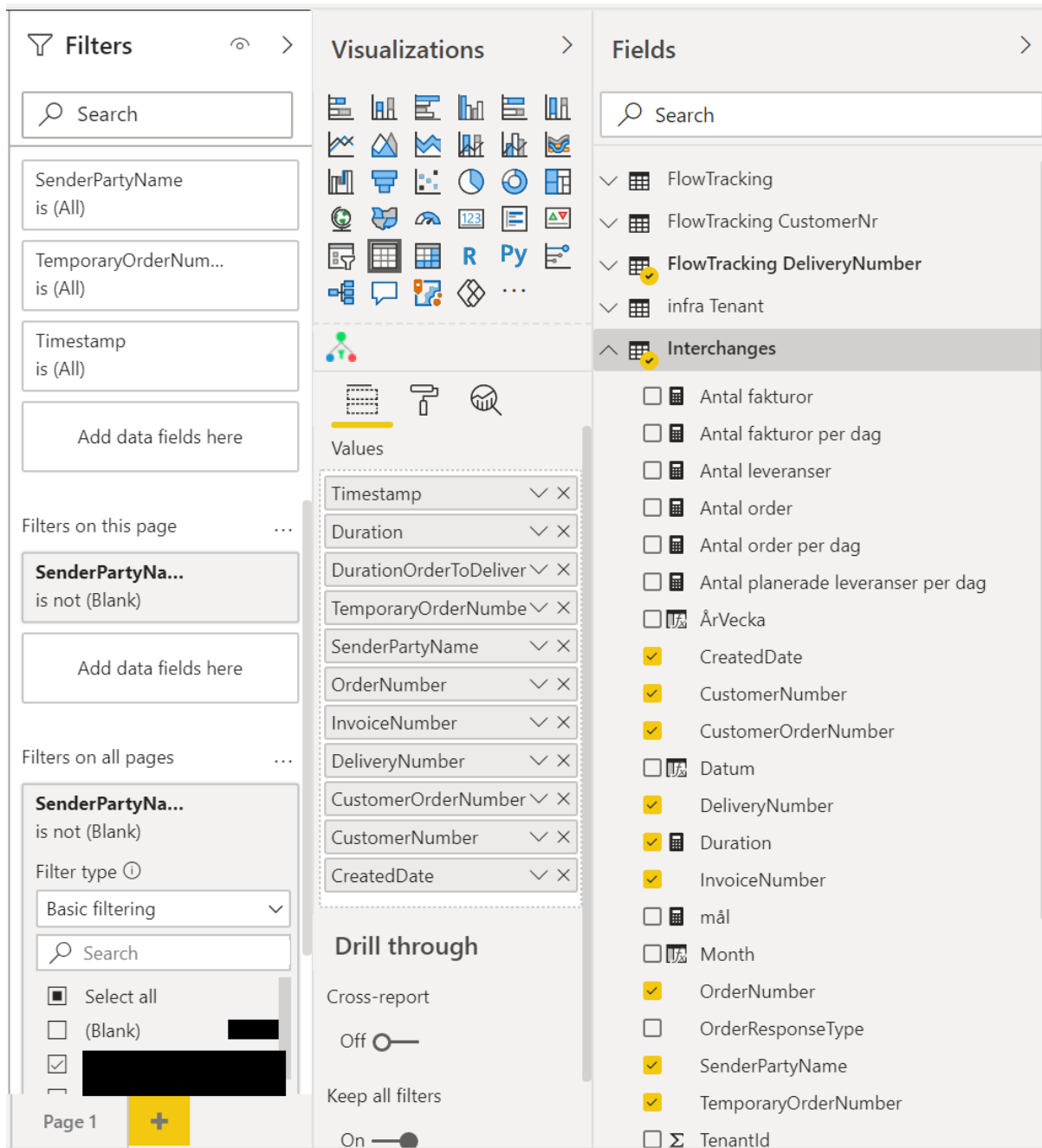


Figure 11: Report view options.

The model view was very useful for getting an overview of all relations. For someone having knowledge of database technology it was also very convenient to set up and manage relations in the model view (figure 12).

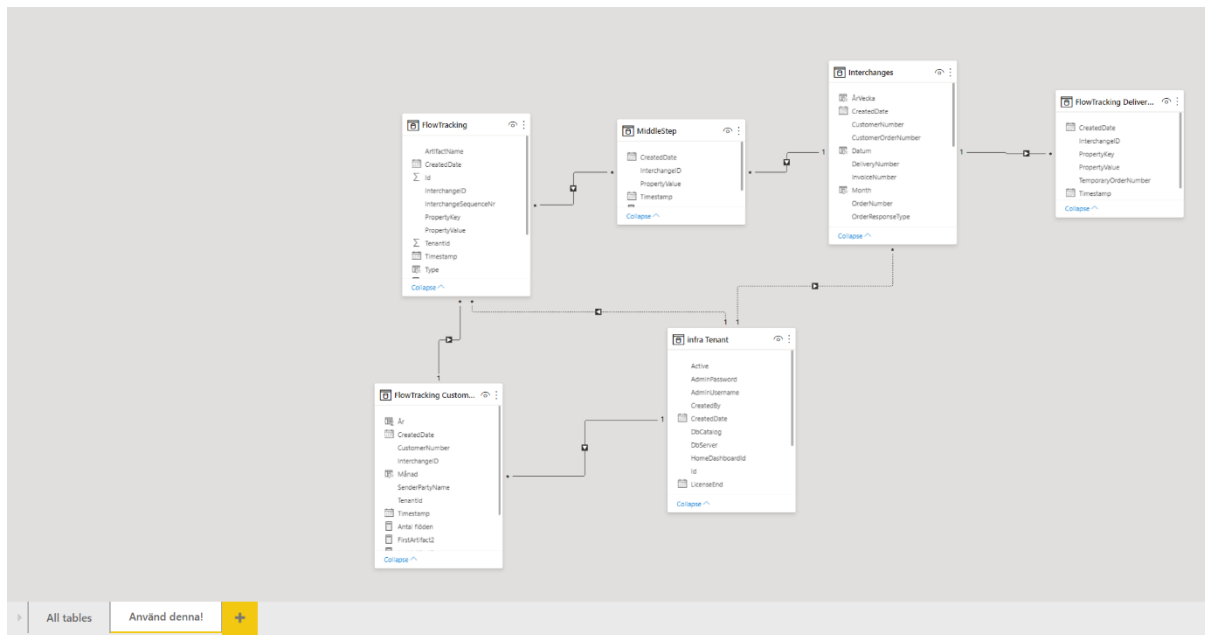


Figure 12: Model view.

4.1.6 Creating visualizations

When all views and functionalities of the old system were implemented the work process of visualizing trends and insights from the data started. Figure 13 shows an example of some of the insights that could be derived and displayed from the data. Here it can be seen that different graphical visualizations can enhance the visual experience of the data. Calculations such as orders per day could easily be created and visualized. A powerful feature was that the visualizations adapted to selections in the data set. A common selection of data was done by adding or removing filters. The view in figure 13 was connected to some other views. Switching between views could be performed by pushing buttons with the names of the different views. To get more detailed information about one row of data the row could be selected. When selecting a row, the button named “Se del-flöden” became active (figure 14). After pushing the “Se del-flöden” button the view changed to a view called Middle Step, see figure 15. In the Middle Step view all data flows of the selection were showed. A calculated value that was created and added to this view was RunningTime, which showed the duration of each flow. All events of a flow could be displayed by selecting a flow and then clicking the button named “Se flöde”, changing the view to the Flow Tracking view (figure 16). Changing views by clicking buttons was done by using the feature drill through.

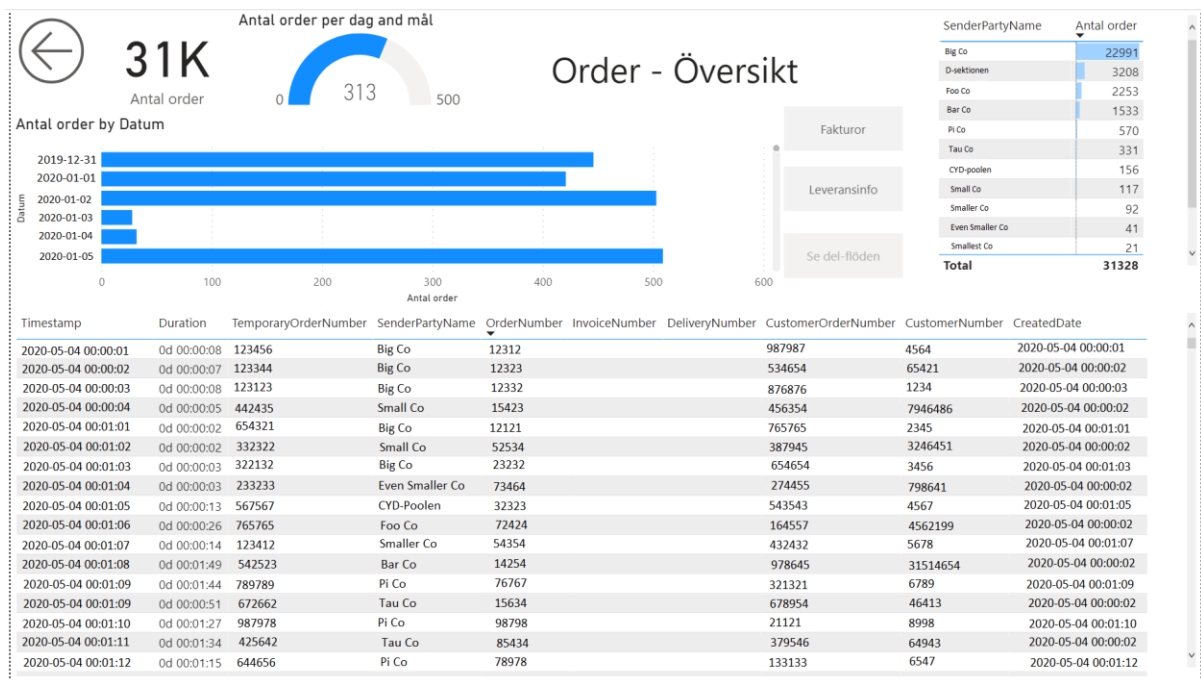


Figure 13: Order view.

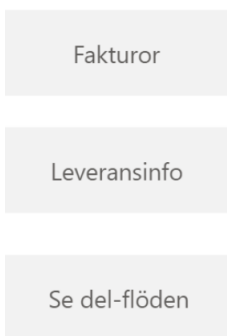


Figure 14: Order view buttons activated.

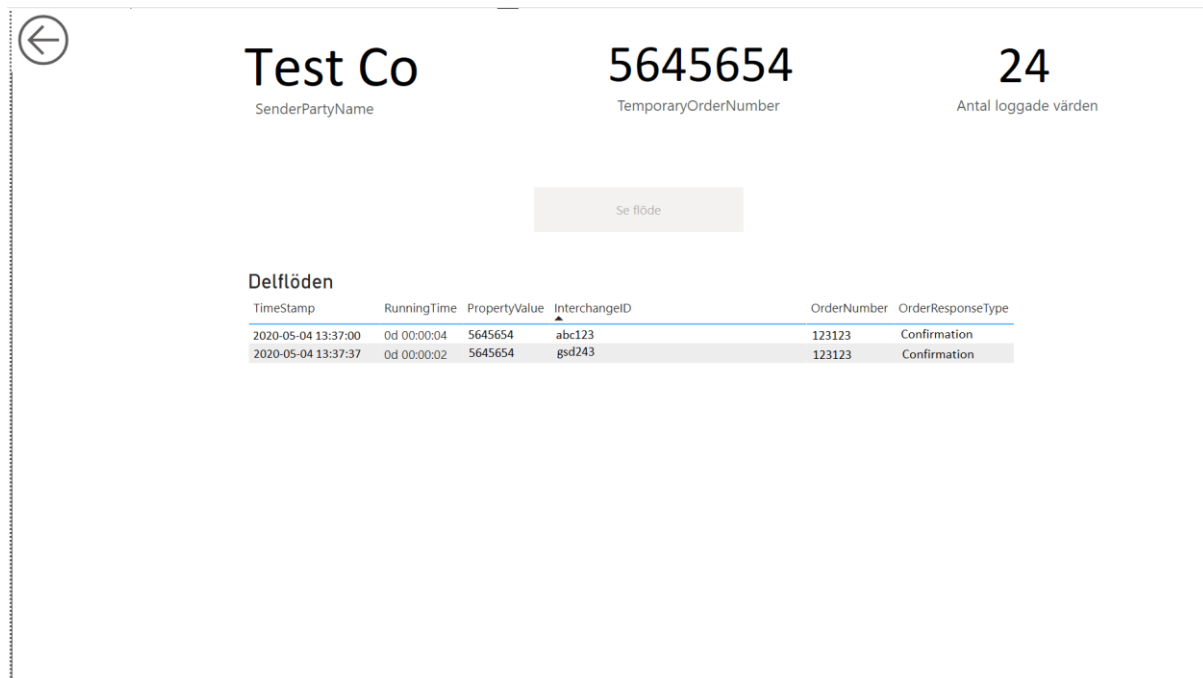


Figure 15: Middle Step view.

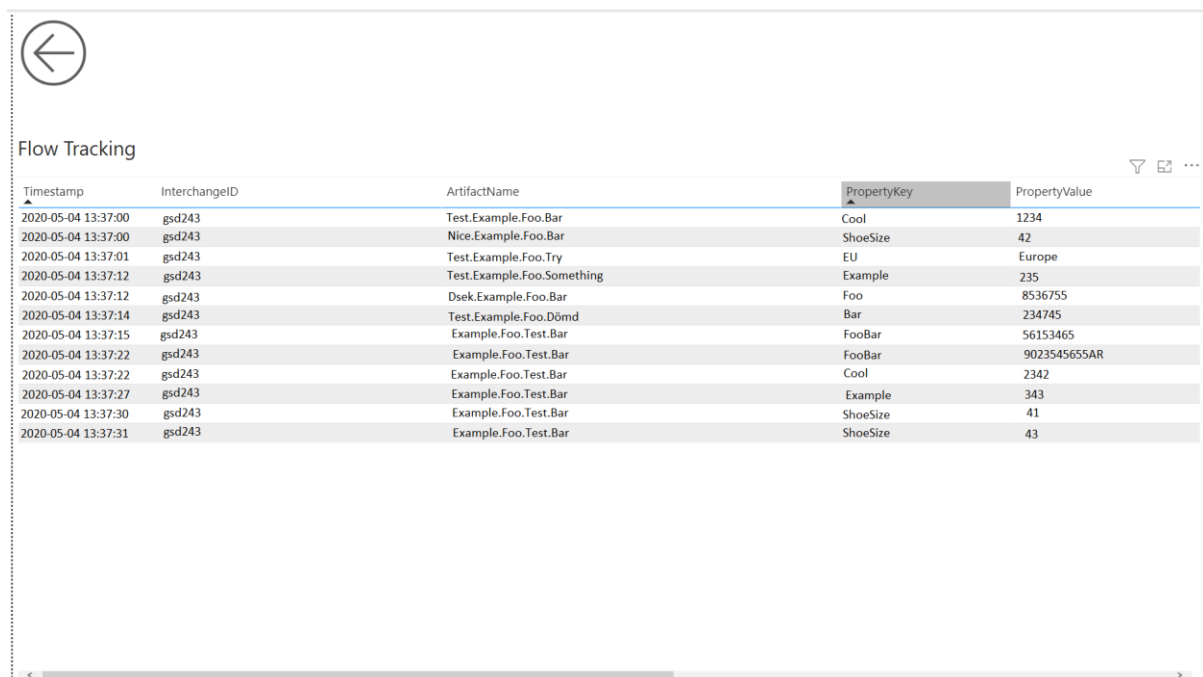


Figure 16: New Flow Tracking view.

4.2 The demonstration and the interview

The system was demonstrated to the test person. After the demonstration, the following questions were asked. Some of the questions were discussed during the demonstration, mostly due to

questions asked by the test user about functionality that was about to be demonstrated but had not yet been demonstrated.

1. What do you think about this new system?

The test person was happy to see that the old system was successfully replicated using the new tool Power BI. During the demonstration, the test person asked if the different graphs and diagrams would adapt if a selection of data was made. The answer to the question was yes and it was also demonstrated visually.

2. Is there any specific functionality you would like this system to have?

While demonstrating the drill through connection between the Order view (figure 13) and the Middle Step view (figure 15) a question was asked if more details could be shown based on a row selection in the Middle Step view. The answer was yes and a drill through was made to the Flow Tracking view (figure 16) by simply pushing a button. The test person was satisfied with that solution.

The test person wondered if it was possible to add functionality that would enable a non-technical user of the system to add columns and data in the tables. It was explained that SQL queries were required to create new columns. By pivoting data different columns had been created. To create new columns additional pivot operations would be required. The reason this solution was used had to do with the design of the database. One way of creating this functionality would have been to create columns for all kinds of data that could potentially be of interest and then create functionality for the user to choose what columns should be visualized.

3. Do you have any other thoughts or questions?

The test user thought it would be interesting to add some additional functionality to the new system compared to the old system. We discussed what functionality could be added that would be usable and at the same time technically possible due to limitations and possibilities in the system. This resulted in the creation of the "Duration" column in the Order view (figure 13) and the "RunningTime" column in the Middle Step view (figure 15).

5. Discussion

5.1 Results

The system worked as intended. It made it possible to visualize customer data and spot trends and behaviors in the data. This connects to the theory chapter, which described visualization as an important technique to understand Big Data. Self-service BI was described in the theory chapter as something that will probably be more used in the future. The question, if users could do their own modifications to the data and change the data, was raised during the interview, showing there was interest in implementing Self-service BI. The system showed that Self-service BI could be implemented, but some changes would be needed to be done to the database. The difficulties of working with Big Data mentioned in the theory chapter were also evident. Due to the big size of the data, there will be difficulties creating new reports in the future due to the file size limit of Power BI that will be exceeded in the future if no changes are made to the system.

5.1.1 The work process

Power BI showed to be an efficient tool to use. All steps from importing data to creating the visualizations worked well and could be performed inside Power BI. The most obvious part that required some workarounds was to create new tables and columns through SQL queries and DAX code. This would not have been necessary if the database would have been prepared for being used by a visualization software. Even though there were some parts that needed some more work, the work process was efficient for someone experienced in writing code. The old system could be replicated in the new system and visualizations showing trends or behaviors could easily be added to the new system.

One part of the work that appeared to be very important was the design of the database. Power BI imports data and visualizes it, it does not create or add data to the database. This means databases that are supposed to be used by Power BI should be designed with Power BI in mind, making the work in Power BI much easier. At one point during the development of the system a connection between two different tables could not be done in Power BI since they shared no common data that could be used to create a key. If a discussion would have been held between the developers of the database and people knowledgeable of Power BI while designing the database these kinds of issues could more easily have been prevented.

5.1.2 The demonstration and the interview

The test person had a good idea of what trends and behaviors Solution Xperts customers would like to see in their data. When the test person saw what could be done using Power BI he wanted to know more about the possibilities of what functionality could be added. After discussing what functionality could be interesting to add in the future it could be concluded that almost all of the requested functionalities could be added where some parts would be quite easy to add while other parts would be trickier to add.

It would have been a good idea to interview more than one person but all persons using the old system work close together so probably they would have had similar thoughts about the new system. Due to the extensive knowledge of the person that was interviewed the demonstration and the interview seem to have provided a lot of value, even though more interviews would have been good for confirmation of the thoughts and ideas that were discussed.

5.2 Trends and behaviors

During the interview it became evident that the drill through function was highly appreciated by the interviewee. This means the test user was satisfied all in all, but it does not say much about the system's ability to make it easy to spot trends and behaviors, which was asked in the research question. If you think of drill through you can drill into more detailed views, but you can also take a step back when you wish to see the overview again. In the system the ability to spot trends and discover behaviors is present in the overview. One example that connects to the interview is the measure "duration" that was created. Duration found the timestamp where an event started and the timestamp where it ended. The time difference was calculated, and a duration was created that could be presented as a column in the table of the overview. Even though every event often consisted of many rows shown in the detailed view only the first row of every event was presented

in the table in the overview. This single row presentation of each event was accompanied by the measure duration, giving the user an overview of all events and their duration. The overview table could be filtered by a user by for example showing which events had the longest duration. By doing such a simple filtering it might be possible to see the behavior that for example orders from a particular customer are the ones that has the longest duration.

The map visualization can be very useful for understanding data. As an example, a company can have warehouses spread over a country. If the company management want to see how many products of a certain kind are sold from every warehouse in a year, using this system would be very helpful. A map visualization could be used to show every warehouse as a dot on a map using the longitude and latitude of every warehouse. Depending on the annual sales of the warehouses the dots can vary in size, showing bigger dots representing warehouses selling a lot in a year, and the other way around for warehouses selling less. In addition to varying sizes of the dots on the map the dots could also have different colors. Warehouses selling more than one hundred of a certain product could have the color green, warehouses selling fifty to one hundred of the product could be yellow and the ones selling less than fifty of the product could be red. This gives the management a quick and powerful overview of the performance of every warehouse and how they compare to each other. By combining sizes and colors of the dots it is easy to find for example big warehouses that does not perform very well selling this product.

To really get a trend out of the map visualization a slider could be added where the user can slide or type different years to see how the selling of this product has changed over years at every location. More visuals are usually added as well. It can be visuals showing numbers such as total sales in a year or average sales in the last five years. Graphs are also usually added to show for example trend lines. The most powerful part here is that the visuals and values can be interacted with by the user. If a particular year is chosen all visuals on the page adapt to that selection showing data for that particular year, if a range of years are chosen the visuals will adapt to that selection as well. This means the user can find trends and behaviors himself without the need of a developer once all visuals are created and connected to the data.

5.3 Replicability

It should be possible to reproduce the system and the results presented in this report. The system created in this work was adapted to a specific customers data and therefore another system probably would look a bit different. It should also be no problem to use another Business Intelligence software than Power BI for creating visualizations.

5.4 Future work

The system can visualize trends and behaviors in customer data, but it is not necessarily easy for a user to do all kinds of operations. To make it easy for a user to perform any type of operation he might be interested in, some changes would be necessary. The system seems to have a good construction overall, but the database would need to be redesigned with Power BI in mind. By redesigning the database, the pivot functions would not be necessary, and it would be easier to create a powerful Self-service BI functionality.

A problem with Power BI is that it has a limit of how big the reports can be. When the database is constantly growing, a point will be reached in the coming years where the report becomes too big. One solution to this problem which has been performance tested by a team at Solution Xperts seems promising. The solution is to use Microsoft Synapse which makes it possible to work with enormous amounts of data at a high speed and at a reasonable cost [17]. Currently Microsoft Synapse is not mature enough to be used (the team at Solution Xperts found around 5 bugs that was reported to Microsoft when trying it) but it seems very promising long term.

In the future it would be a good idea to give users the ability to create new views and eventually even create calculated values. To achieve this quite a lot of code writing and architectural work needs to be done. Luckily, Power BI is continuously developed meaning we could see more of the functionality wished for soon.

6. Conclusion

The aim of this work was to be able to answer the following question:

How should a system for visualizing customer data be constructed to make it easy to spot trends and discover behaviors from the data?

The research question was delimited to creating a system that was good enough to fulfil the needs of one of Solution Xperts customers. The research question was investigated through the development of a system that was supposed to have the ability to visualize trends and behaviors in customer data. The system was created as a new version of an old system that lacked the ability to show trends and behaviors in the data. When the new system had all features of the old system it was demonstrated to a person who had been using the old system and had a good idea of what new features would be good to add in the new system. New features were added, and it became evident that the system could be used to visualize trends and behaviors in customer data. The new system is usable by Solution Xperts customers and due to its construction, it can easily be configured with data from different customers, making it possible for many customers to enjoy the benefits of being able to visualize trends and behaviors in their data.

For the future it might be necessary to both create a new SQL database with a different design and use Synapse to be able to handle the big amount of data that will be stored due to the continuous addition of data to the database.

References

- [1] J. Magenheimer and C. Schulte, "Data Science Education," in *Encyclopedia of Education and Information Technologies*, Cham, Springer, 2020, pp. 493-514.
- [2] W. Albattah, "The Role of Sampling in Big Data Analysis," in *BDAW '16: Proceedings of the International Conference on Big Data and Advanced Wireless Technologies*, Blagoevgrad, 2016.
- [3] I. Emmanuel and C. Stanier, "Defining Big Data," in *BDAW '16: Proceedings of the International Conference on Big Data and Advanced Wireless Technologies*, Blagoevgrad, 2016.
- [4] X. L. Dong and D. Srivastava, "Big data integration," *Proceedings of the VLDB Endowment*, vol. 6, no. 11, pp. 1188-1189, 2013.
- [5] R. Venkatraman and S. Venkatraman, "Big Data Infrastructure, Data Visualization and Challenges," in *BDIOT 2019: Proceedings of the 3rd International Conference on Big Data and Internet of Things*, Melbourn, 2019.
- [6] C.-P. Lopez, M. Segura and M. Santórum, "Data Analytics and BI Framework based on Collective Intelligence and the Industry 4.0," in *ICISS 2019: Proceedings of the 2019 2nd International Conference on Information Science and Systems*, Tokyo, 2019.
- [7] Q. V. Duy, J. Thomas, S. Cho, P. De and B. j. Choi, "Next Generation Business Intelligence and Analytics," in *ICBIM '18: Proceedings of the 2nd International Conference on Business and Information Management*, Barcelona, 2018.
- [8] J. Heer and S. Kandel, "Interactive Analysis of Big Data," *The ACM Magazine for Students Volume 19*, pp. 50-54, 2012.
- [9] J. Heer and B. Shneiderman, "Interactive Dynamics for Visual Analysis: A taxonomy of tools that support the fluent and flexible use of visualizations," *Queue*, pp. 45-54, February 2012.
- [10] K. Morton, M. Balazinska, D. Grossman and J. Mackinlay, "Support the data enthusiast: challenges for next-generation data-analysis systems," *Proceedings of the VLDB Endowment*, vol. 7, no. 6, pp. 453-456, 2014.
- [11] L. Chen, Z. Pan and Y. Lina, "Study on Clustering Computing Methods of Big Data," in *ICITEE-2019: Proceedings of the 2nd International Conference on Information Technologies and Electrical Engineering*, Zhuzhou Hunan, 2019.
- [12] Microsoft, "Power BI documentation," 2020. [Online]. Available: <https://docs.microsoft.com/en-us/power-bi/>. [Accessed 05 11 2020].
- [13] M. Mani and S. Fei, "Effective Big Data Visualization," in *IDEAS 2017: Proceedings of the 21st International Database Engineering & Applications Symposium*, Bristol, 2017.
- [14] N. Bikakis, "Big Data Visualization Tools," in *Encyclopedia of Big Data Technologies*, Cham, Springer, 2018.

- [15] B. Shneiderman, "Extreme visualization: squeezing a billion records into a million pixels," in *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, Vancouver Canada, 2008.
- [16] M. Arvola, Interaktionsdesign och UX - om att skapa en god användarupplevelse, Lund: Studentlitteratur AB, 2014.
- [17] Microsoft, "Azure Synapse Analytics," 2020. [Online]. Available: <https://azure.microsoft.com/sv-se/pricing/details/synapse-analytics/>. [Accessed 01 12 2020].