Week 6 - Data visualization

The objective of the sixth week is to create the data visualization for the project.

As the data mart layer is completed, the project is ready to jump into the visualization part through Power BI. The proper presentation of the data is essential as the main objective is to create actions from the stakeholders through insights gained from the dashboards. Last part of the project is to create visualizations as a foundation for these insights.

Learning materials

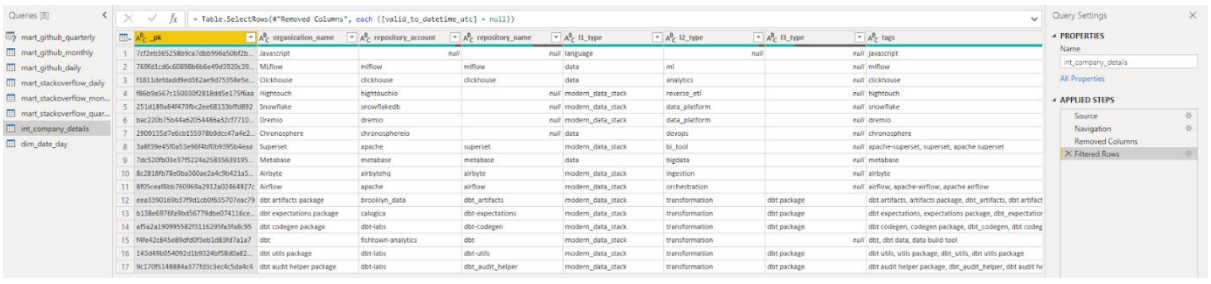
- Power BI Tutorial
- Power BI Beginner Tutorial
- Power BI Map

Prerequisites

- Completed data mart layer
- Installed Power BI Desktop through this link (not from Microsoft Store)

List of actions in Power BI

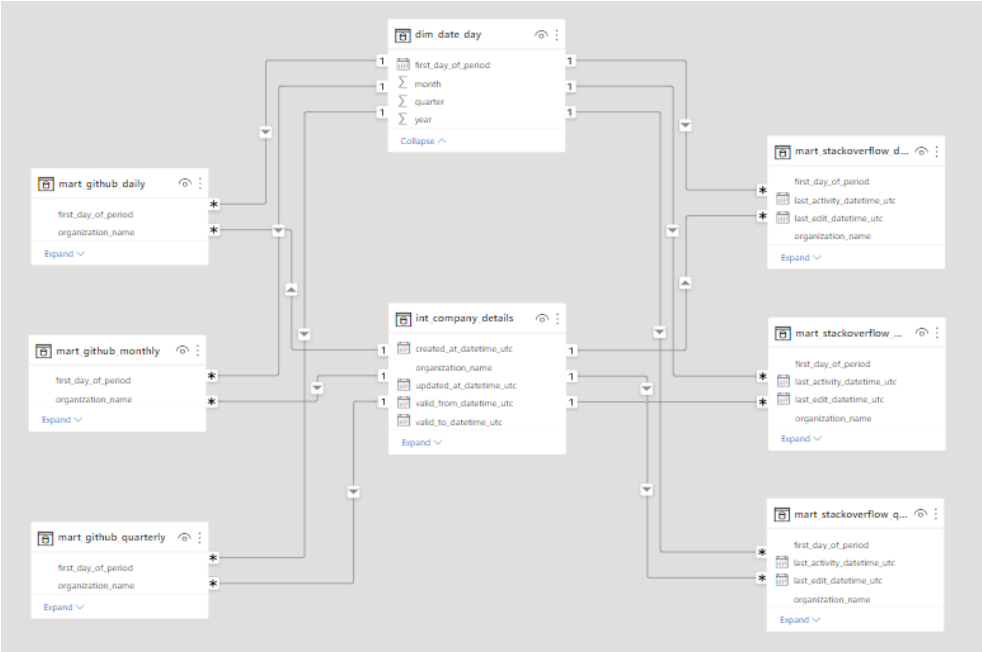
- Add the Big Query data sources. Get data button from the panel.
- Select your project and add all the mart tables and the intermediate Company Details only.
- Edit and sort the Company Details table in the Power Query Editor. The Company details should only contain the actual version of the sheet.



- Create a dim_date_day table from scratch. One of the daily data mart tables can be duplicated, sorted, and renamed for this purpose. The final table should be the following with unique values.

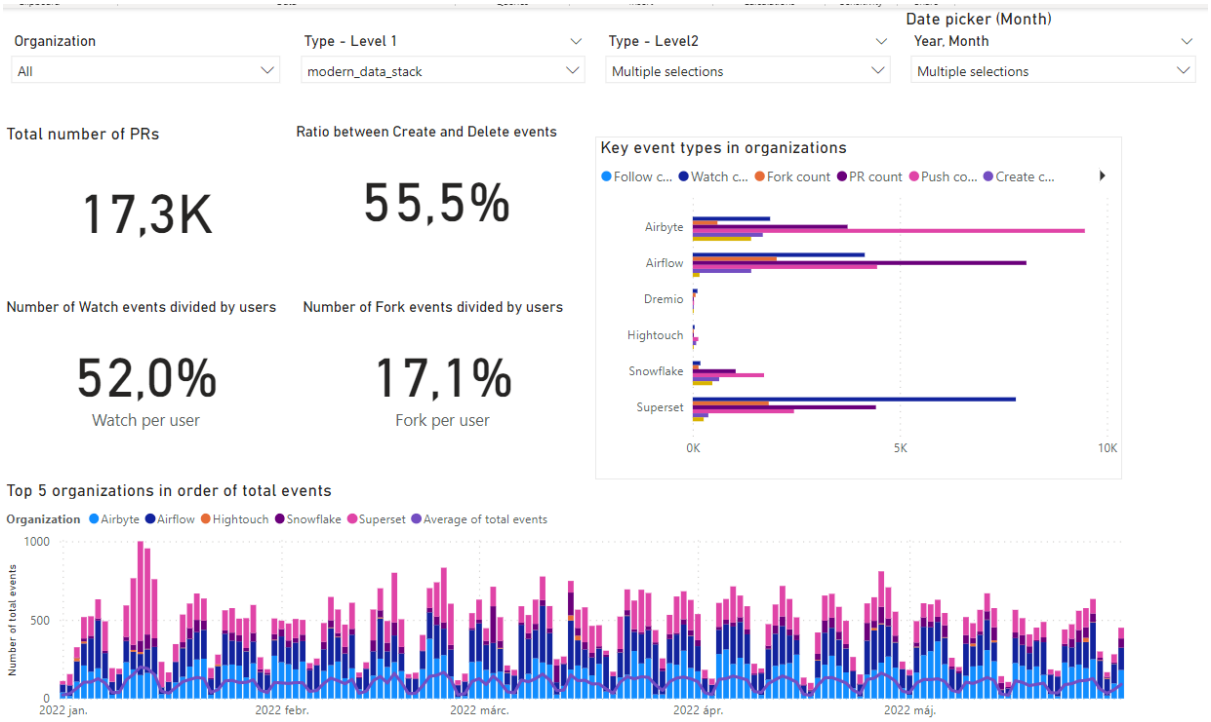
	first_day_of_period	month	quarter	year
1	2022. 01. 01. 0:00:00	1	1	2022
2	2022. 01. 02. 0:00:00	1	1	2022
3	2022. 01. 03. 0:00:00	1	1	2022
4	2022. 01. 04. 0:00:00	1	1	2022
5	2022. 01. 05. 0:00:00	1	1	2022
6	2022. 01. 06. 0:00:00	1	1	2022
7	2022. 01. 07. 0:00:00	1	1	2022
8	2022. 01. 08. 0:00:00	1	1	2022
9	2022. 01. 09. 0:00:00	1	1	2022
10	2022. 01. 10. 0:00:00	1	1	2022

- Create the relational model. This is a possible output, but all the six data mart tables should be mapped to the company details table through organization and the dim_date_day table through the first_day_of_period column.



First dashboards

Now everything is set to start the creation of the visualizations. At this stage, the task is to start to explore Power BI. An MVP is added here, it is just a few filters and charts to start the visualization. It is up to your imagination, but try to use various charts, filters, diagrams and create some custom KPIs using DAX.



Dashboard implementation

The purpose of the dashboards is to answer questions that could be relevant, easily interpretable and logically correct for the business objective.

In this case, it is to have a better understanding of trends, correlations, and insights on emerging and already existing data companies and between data categories.

General

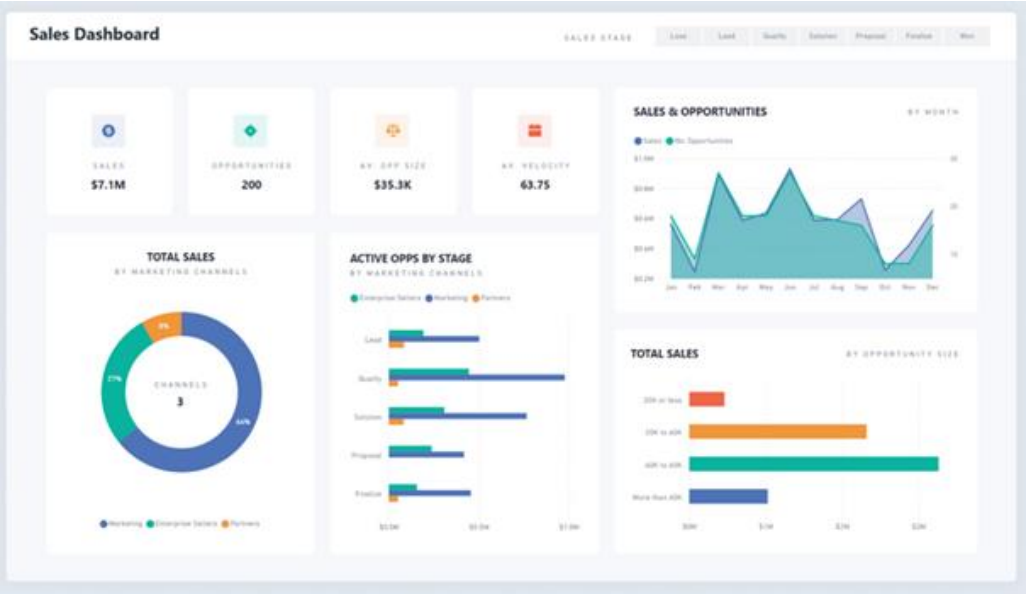
In BI projects, it is common to have a design with the expected output before the actual implementation. It is called a mock-up; mock-ups are models or replicas of a final output that can be used to illustrate the look and feel of a dashboard before any actual development work begins. There are several tools in the market in this category, one of them is called Figma.

It is also important to note that the dashboards should handle all the possible outcomes regarding filtering and drilling down in subjects. It is a common problem in visualization and using BI tools to create dashboards that can show misleading or incorrect data. It means it should not be possible for an end user that with various clicks, creates a state where it shows false or logically incorrect data. On the other hand, there are cases when it is unavoidable as a report or page focuses only on presenting one business case, but in that case, it should be communicated clearly to the end users or lock those routes which can end up in false results.

One of the key aspects of creating visualizations is the right balance in the depth and in the number of visualizations. More graphic visualizations per page makes it harder to interpret. Filters usually applied in one location, usually the upper part of the page or the left side.

Make sure that you have a title for the page itself and for every visualization as well. Labels can be used if there are not easily interpretable metrics or measures.

The overall design and the placement of the visualizations is part of the task. It is up to the creator's aesthetics and imagination to create the look and design of each dashboard.



GitHub dashboard

In GitHub, there will be two main pages. One is to analyze one company in depth and the other is to compare companies from the same category type.

1. page – Company analysis

This page focuses only on analyzing one company.

The objective of this page is to have an understanding on the performance of the organization and the progress from previous periods.

Filters	Description
Company	No multiple choice, only one company. Category of the company should be indicated as a label.
Event types	It should only affect the related line chart, which is about event totals
Date	Monthly aggregation. The monthly data mart table should be used
Open source	Y/N filter
Visualizations	**Description**
Line chart	All the events should be presented here from the daily data mart table. It can be filtered with the event type slicer. Additional task is to show the previous month's value. It can be visualized with any type of line chart.

Table	It should show all the event types without filter applied. Columns: Event types || Total counts events || Change in percentage of the event from the previous month. It should be ordered from the change percentage from the previous month. Conditional formatting can be applied (higher than the previous month in green, lower in red)
KPIs	**Description**
User count	Total and the change from the previous month
Event count	Total and the change from the previous month
Custom standardized KPIs	The totals without context can be misleading, it should be normalized as comparable metrics. For example the number of commits or forks depends on the number of active users that contributed to the project. 2-3 KPIs would be enough here and it should be decided by the creator.

2. page – Category analysis

This page focuses on the companies within the same data category. There will be a company filter because it is important to examine a company against the direct competitors within the category, but it can give an average look about the company's overall performance by categories. This is the main objective of this page, and the visualizations tend to satisfy this pursuit.

Filters	**Description**
Company	Can be multiple companies.
Category	It is the level 2 type category. Once a company is picked, this filter has to be jumped to the company's category only.
Date	Daily aggregation in hierarchy, drill down from year to day. The daily data mart table should be used.
Open source	Y/N filter
Visualizations	**Description**

Treemap	Event and user volume by companies. Event is mandatory, but it can be added additionally to change between events and users within the same visualization.
Line chart 1	Line chart with the total users and events volume by time. An average should be added as well which is the average of the category.
Line chart 2	The three most meaningful event types for companies. It should be picked individually. An additional task is the average line with the average of the top three performers from each event type within the category.
Table	List of companies from the category with the following columns: Organization || Average monthly active users || Average monthly PR actions || Average monthly watch events || Commits per user || Total events per user || Watch events per user. Monthly data mart table can be used here

Stack Overflow

The same two pages will be used in Stack Overflow, but with different metrics. There will be an additional page to answer other types of questions that do not depend on categories and companies itself.

Most of the metrics and measures are previously calculated in the database layer. It was part of the project, but in other cases the visualization tool is capable of performing these easy calculations. A possible additional task is to create some of the already calculated metrics with DAX. For example, the average on accepted answers, which is the accepted answers divided by the number of posts.

1. page – Company analysis

This page focuses only on analyzing one company.

The objective of this page is to have an understanding about the company regarding Stack Overflow trends.

Filters	Description
Company	No multiple choice, only one company. Category of the company should be indicated as a label.
Date	Monthly aggregation. The monthly mart table should be used.

Open source	Y/N filter
Visualizations	**Description**
Line chart	Various trends on metrics over time: Number of posts || Number of answers || Number of comments || Number of favorites
Table	There should be three columns. One is the current month value, the previous month's value and the difference in percentage. The rows are the metrics. Every metric can be added here.
KPIs	**Description**
Unanswered questions	Number of questions without any answers. And it is divided by the total number of questions.
Score	Individually created logically correct KPI based on the creator's score metric.

2. page – Category analysis

This page focuses on the companies within the same data categories. It will follow the same logic as in GitHub.

Filters	**Description**
Company	No multiple choice, only one company. Category of the company should be indicated as a label.
Category	It is the level 2 type category. Once a company is picked, this filter has to be jumped to the company's category only.
Date	Daily aggregation. The daily mart table should be used.
Open source	Y/N filter
Visualizations	**Description**

Line chart	Line chart on metrics over time which can be chosen by metrics from a list. Metrics: View count, average view count, answer count, average answer count, favorite count, average favorite count, average no answer count, score. The line chart should show only one of the metrics at a time and it can be chosen by the end user.
Stacked bar chart	All companies within a category where the stacks are the following: View count || Number of questions || Number of answers
KPIs	**Description**
KPIs next to each other	Two cards next to each other for the same measure. Left is the chosen company, the right is the average of the companies within the category. 2-3 KPIs should be added using this approach, which can be picked individually. Additional task is the conditional formatting as the higher is green and the other is red.

3. page – Top/low and other non-category related questions

In this page, highest and lowest performers are shown and interesting questions that do not really match any of the previous reports.

Filters	**Description**
Category	It is the level 2 type category.
Date	Daily aggregation. The daily monthly mart table should be used.
Open source	Y/N filter
Visualizations	**Description**
Table 1	Companies with the most of the unanswered questions in order with the following columns: Organization || Unanswered questions || Unanswered question per total questions

Table 2	List of the companies where the latest activity is not later than February. Companies with zero activity should be excluded.
Table 3	List of companies without any questions. List of categories with most companies without any questions. It should be two separate tables.
Table 4	Top ten most answered questions. Columns: Organization name || Number of answers || Link to the question as an URL (https://stackoverflow.com/questions/ concat with question_id
KPIs	**Description**
Most commented category	Category with most comments
Most viewed category	Category with most views
Month with less activity	The month with less activity

Additional task

Third dashboard where both data sources are analyzed. It is a completely individual task. The objective is to create a scale between 1 and 10 which is derived from various weighted measures. It can be anything and the goal is to create a top-down list of the companies derived from both GitHub and Stack Overflow activity. It is a common process where different metrics can be put together on the same scale. The goal is to have a weighted score which can be coming from any measures, because at the end, it is the same for every company. It can be the total view counts divided by the questions, added the favorites per questions and the score overall, it should be distributed between a 1-10 range and the same can apply to GitHub as well. At the end, there will be a distinct comparable list for companies with a weighted scale.

The whole dashboard should be built around this metric. It should be the center of the dashboard and other KPIs and visualizations should only highlight the source data that this metric is derived from.