Google Summer of Code '23

Add more summarization functions

Project: Add more summarization functions

Organization: Mathesar

Name: Aritra Majumder

Email ID: aritramajumder8438@gmail.com

Synopsis:

The Mathesar Data Explorer enables an action called "Summarize" that let a user view a summary of their data, which is in fact an aggregation of some column(s), grouped by some other column(s). Currently, the only possible aggregation functions are counting or listing. At the end of the GSoC period, more aggregation functions like sum, min, max and mean and median should be added for better user experience

Feature Description:

- ❖ We want to add more summarization (aggregation) functions to the Mathesar Data Explorer. The functions should either come from the PostgreSQL aggregate functions, or the implementer could create their own. Functions to prioritize are:
  - ➢ Summing numeric columns
  - ➢ Joining array (list) columns into a single array (list)

➢ Merging JSON Object columns

➢ Statistical aggregations (Mean, Median, Max, Min)

Architectural Problems:

The only real UX issue to solve here is how to present the different options in a way that is understandable to the user. It may be that the current drop-down list needs to be enriched somehow. It's possible that the implementer could want to do an aggregation that needs some kind of presentation of the output, but that's doubtful.

Project Implementation:

Data is essentially the plain facts and statistics collected during the operations of a project/business. They can be used to record a wide range of activities and to analyze them to get an overview of the same. **Aggregation functions are one of those functions which give a broad overview on considering all the information (data that carries logical meaning).**

Some of the aggregation functions are:

- **Mean (or average):** The arithmetic Mean is the most popular measure of central tendency. It is defined as the sum of a collection of numbers divided by the count of numbers in the collection.

- **Median:** The median is the middle number in a sorted, ascending or descending list of numbers and can be more

descriptive of that data set than the average. It is the point above and below which half (50%) of the observed data falls, and so represents the midpoint of the data. When there are outliers in the dataset, the median is considered to be the best measure of central tendency.

- **Sum:** Sum is also one of the most popular measures of aggregation. It is extensively used in accounting.
- **Min:** Min (or minimum) is the smallest number or quantity among a set of the same. Generally, it is used to get an idea about the range and the outliers of a dataset.
- **Max:** Max (or maximum) is the largest number or quantity among a set of the same. It is also used to get an idea about the range and the outliers of a dataset.

Postgres Implementation:

- **Mean (or average):** The `AVG()` function is one of the most commonly used aggregate functions in PostgreSQL. The `AVG()` function allows us to calculate the average value of a set.

```
SELECT AVG("data") FROM public.data_table;
```

- **Median:** The `PERCENTILE_CONT(0.5)` function allows us to calculate the median value of a set.

```
SELECT PERCENTILE_CONT(0.5) within group(order by "data")
FROM public.data_table
```

- **Min:** The `MIN()` function allows us to calculate the min value of a set.

```
SELECT MIN("data") FROM public.data_table;
```

- **Max:** The `MAX()` function allows us to calculate the max value of a set.

```
SELECT MIN("data") FROM public.data_table;
```

- **Sum:** The `SUM()` function allows us to calculate the max value of a set.

```
SELECT MIN("data") FROM public.data_table;
```

Moreover, to implement Joining array (list) columns into a single array (list) and Merging JSON Object columns we have to create custom PostgreSQL functions.
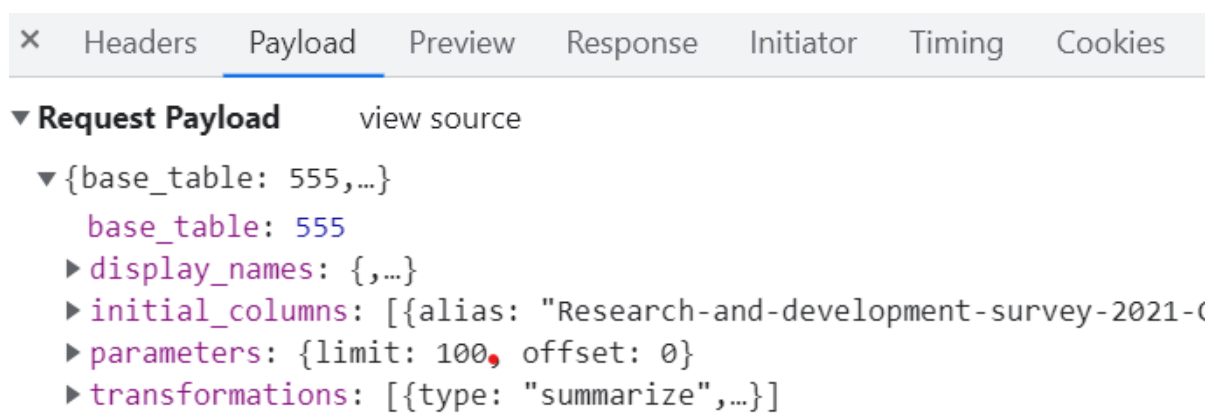
UX Design Problems:
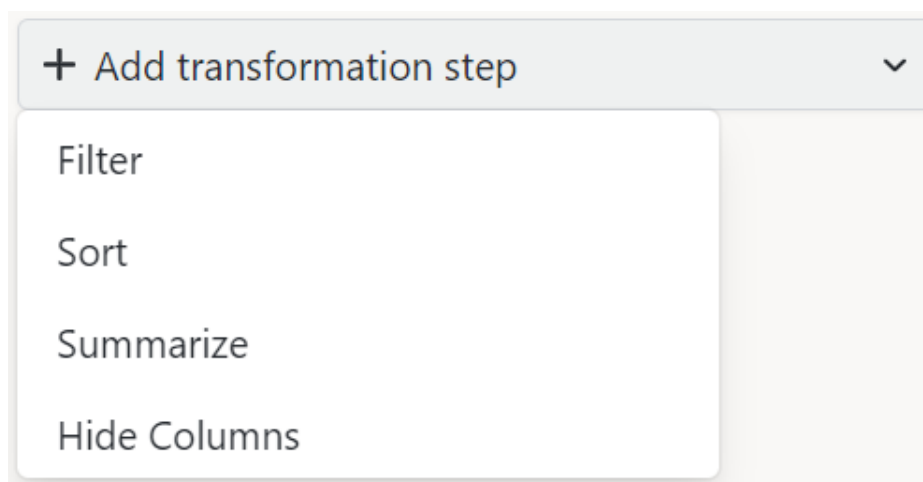
We can add aggregation functions (Mean, Median, Sum, Min, Max) to this dropdown list and therefore show the result to the topmost highlighted cell of the next free column.

API Spec:

Currently, whenever summarize function is called,

The request is sent through the post method to the corresponding URL with the following payload.

We have to modify the transformation.type to the corresponding aggregation functions and subsequently call

the database query based on the type received at the backend.

Timeline and Deliverables:

May 4	Proposal accepted or rejected.	• Community Bonding Period - Discussing the project with mentors and finalizing the project implementation.
May 29	Pre-work complete	**Coding Officially begins!**
Jun 6	Milestone #1	• Determine which aggregation functions might be necessary after discussing with mentors. • Go through the PostgreSQL aggregate functions documentation to determine which of them might be useful
June 20	Milestone #2	• Write PostgreSQL queries to implement the aggregation functions. • Write custom functions to join array (list) columns into a single array (list) and to merge JSON Object columns.
July 6	Milestone #3	• Add the API endpoints to enable users to perform aggregation functions to transform data.
July 20	Milestone #4	• Enrich the current dropdown list by adding all the aggregation functions. • Integrate the frontend and the backend.
August 4	Milestone #5	• Thoroughly test the APIs to assess the functionality, reliability, performance and failure-handling capability.

August 18	Milestone #6	- During this period, I need to submit a report to my mentors on the performance of the API after integrating Excel support. - Based on the feedback received, I might work on some additional features.
September 1	Milestone #7	- By now, Mathesar UI should be able to allow users to transform data using aggregation functions, - Finalize everything done till now, add (or update) the documentation, add tests for the new features.
September 15	Final week	- Submit my final work and wait for the evaluation. - Write a blog about my experience as a contributor to Mathesar under GSoC.

This timeline is tentative and adjustable based on the mentors' feedback, advice and current progress.

Questionnaire:

1. Why are you interested in working on Mathesar?

   I had chosen Mathesar because I liked the idea behind Mathesar and it's tech-stack perfectly aligns with my expertise. I stayed with the organization because of the wonderful community it has.

2. Why are you interested in working on this project idea?

   This project needs interaction with the database, backend as well as frontend. I will have a lot of opportunities to learn from the mentors as well as to implement what I have already learnt.

3. What about your skills and experience makes you well-suited to take on this project?

I have previous experience working with PostgreSQL and Django ORM. I have also worked with Javascript and VueJS in many of my projects.

4. Do you have any other commitments during the program period? Provide dates, such as holidays, when you will not be available.

After my final semester ends (May 15), I don't have any other commitments.

5. If your native language is not English, are you comfortable working closely with a mentor in English?

Yes, absolutely.

6. Have you worked on a project remotely and/or with people in other time zones before? If you have, please provide details.

No, I did work on some projects remotely but not with people in other time zones.

7. Are you interested in contributing to Mathesar after the program is complete?

Yes. I started open-source contribution recently and am enjoying it very much. After this project is completed, some issues and bugs might be raised by the users and contributors. I will love to solve those issues and other issues as well.

About Myself:

I am a senior pursuing a B.E. degree in electrical engineering at Jadavpur University, Kolkata, India. I have been into web development and competitive programming since 2021. Apart from them, I enjoy playing football and athletics and watching soccer games, anime etc.

- Name: Aritra Majumder
- IRC: @arima_kun:matrix.org

- Mail Id: aritramajumder8438@gmail.com
- Phone No: +918240484924
- GitHub: https://github.com/Aritra8438
- LinkedIn:
  https://www.linkedin.com/in/aritra-majumder-juee23/
- Country: India
- Primary language: English, Bengali.
- Typical work hours: 10:00 AM - 8:00 PM IST (UTC + 05:30)
- Previous GSoC participation: This is the first time I am participating in GSoC.

Education:

- Institution: Jadavpur University
- Degree: Bachelor of Electrical Engineering
- Major: Electrical Engineering
- Graduation Year: 2023
- Courses Taken:
  - Web Development
  - Data Structures and Algorithms
  - Object-oriented Programming
  - Database Management System

Skills:

Skill name	Proficiency	Where I've used this skill
Python	4	Internship at PwC
Django	4	During Internship at Shiksha-sopan and in projects under Mercor
Javascript	3	In my various projects.
DBMS	3	Internship at PwC
SQL	4	Internship at PwC
Teamwork	5	Throughout my life

Experience:

Brief Description	Relevant links	Where you've used this skill
Worked on developing an online assessment platform operated under H.C.Verma	[NAEST website](#) This project is confidential.	Worked on the backend development written using **Django**.
Worked on developing "Twilio Medicine Reminder" under Mercor.	[Live demo](#) The actual implementation is private under Mercor	Worked on the backend (**Django**) and frontend (**VueJS**) and the database (**PostgreSQL**) of the site and integrated Twilio Twilio messaging API.
Developed a demo E-commerce site from scratch	[Code](#)	In my various projects.
Project Lighthouse (under PwC)	This project is confidential.	An industry-oriented project where I have extensively learnt and used Python, SQL and packages like OpenPyXL,

		XlWings and pandas.

Why I might be a Good Contributor:

While I have a comprehensive electrical background, my emphasis is on software development. I have a considerable background in PostgreSQL (and MySQL), Python and Javascript and the frameworks built on them like Django and VueJS (I have mentioned VueJS because Svelte and VueJS have surprisingly similar structures). I have worked on Django and PostgreSQL to a great extent in my projects and internships.

Moreover, I have a good understanding of Django ORM (object-relational mapper) which might be relevant to this project. I have never used SQLAlchemy in my projects before but as one of the Python ORMs, I have an understanding of how Python ORMS works. I also have a good knowledge of SQL as I had to work on it during my internship period.

I have a strong knowledge of **Data Structures** and **Algorithms** which might be helpful when it comes to optimizing or improving the time complexity of an algorithm making the user experience. As I have a good

understanding of the prerequisites, I can focus more on the project goals and implement many relevant features.

Other Commitments:

I have my final semester starting on May 1 and ending on May 15. After that, I'll be able to devote a minimum of 4 hours daily during May, June and July.

And in August, I have to choose a career option, but irrespective of what I choose, I will be able to devote 2 hours daily and 4 hours on weekends.

Contributions to Mathesar:

I have started contributing to Mathesar very recently and my pull requests have not been merged yet.

Here is the list of bugs and feature requests I have reported:

1. [Support Importing Semicolon Separated Values file #2753](#)
2. [Exploration Inconsistency when Corresponding Table is Deleted #2747](#)
3. [Database Page not reflecting updated Table Count unless reloaded #2736](#)

Here are the pull requests:

1. [#2720](#) - This might optimize [this issue](#).
2. [#2759](#) - This might fix [this issue](#).
3. [#2736](#) - This might fix [this issue](#).

Thank You, Mathesar:

I hardly had any open-source contributions until I recently found Mathesar (to be specific, on the 20th of this month). But thanks to all the mentors and the maintainers of the community, their constant guidance and quick responses made open-source contribution easier and much more interesting for me in such a small time. Thank you for this beautiful community.

It's my earnest request to you to grant me a little time so that I can contribute towards the community. I am looking forward to being a part of Mathesar. Thank you, again.