Enhancing the interactive visualisation of a data preparation tool from in-memory fitting to Big Data sets

Gorka Epelde[1,2], Roberto Álvarez[1,2], Andoni Beristain[1,2], Mónica Arrúe[1,2], Itsasne Arangoa[1,2] and Debbie Rankin[3]

[1] Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia-San Sebastián, Spain
[2] Biodonostia Health Research Institute, eHealth Group, 20014 San Sebastián, Spain

{gepelde,ralvarez, aberistain, marrue, iarangoa}@vicomtech.org

[3] School of Computing, Engineering and Intelligent Systems, Ulster University, Derry~Londonderry, Northern Ireland, UK

d.rankin1@ulster.ac.uk

Abstract. In order to derive reliable insights or make evidence-based decisions, the starting point is to assess and meet a minimum quality of data, either by those that publish the data (preferably) or alternatively by those that prepare data for analysis and develop specific analytics. Much of the (open) data shared by governments and different institutions, or crowdsourced, is in tabular format, and the amount and size of it is increasing rapidly. This paper presents the challenges faced and the solutions adopted while evolving the web-based graphical user interface (GUI) of a tabular data preparation tool from in-memory fitting to Big Data sets. Traditional standalone processing and rendering solutions are no longer usable in a Big Data context. We report on the approach adopted to asynchronously pre-compute the visualisations required for the tool, in addition to the applied visualisation aggregation strategies. The implementation of this approach has allowed us to overcome web-browsers' client-side data handling limitations and to avoid information overload when using granular information charts from our existing in-memory data preparation tool with Big Data sets. The developed solution provides the user with an acceptable GUI interaction time.

Keywords: Big data visualisation, data preparation, data quality, exploratory data analysis, visual information cluttering, data reduction, asynchronous pre-processing

1 Introduction

With the advent of mobile technology and the Internet of Things, together with the trend to share, either publicly or under request, different datasets for research and analysis, has led to data sets that are too large and complex for traditional data processing and data management applications.

The Big Data era has brought massive datasets that are noisy and heterogeneous, requiring new processing and visualisation approaches, given that traditional databases and architectures are not able to efficiently store and process them. The heterogeneous data sources have to be accessed using different protocols, transmission rates, with different data quality levels and schema representations.

In the field of Big Data information visualisation and data management, research has classified the wishful characteristics of a Big Data Visualisation tool [1]. The first desirable characteristic is defined as scalable data management to handle and enable real-time interaction over datasets with a huge number of objects. Coupled with data management, scalable and efficient visualisation of large and dynamic sets of volatile raw data is an advisable feature to have. Regarding the consumer of such tools, the other two recommended attributes are: visual scalability to avoid problems related to visual information cluttering, and customisation capabilities of visualisations to meet the expectations and analysis needs of different user types.

Moreover, increasing data democratization, i.e. societal and technological evolution making data accessible to everyone, is leading to the availability of very diverse and large datasets for analysis, to people that might lack data analysis expertise (e.g. as research scientists, policy makers, or individuals).

Visualisation techniques, used in data visualisation tools, provide users with intuitive means to interactively explore the content of the data, identify interesting patterns, infer correlations, and support sense-making activities.

Currently, the challenge is to implement the best combination of underlying data-management technologies and visualisation techniques to enable end-users to gain value and insights out of the data quickly, minimizing the role of IT-experts in the loop. This is especially critical in the Big Data context and for data preparation and Quality of Data (QoD) improvement tools, where users are not limited to exploration and analyse of data and therefore need to be able to transform the dataset to meet their goals.

The contribution of this paper is the description of detected problems and implemented solutions for the visualisations of a Data Preparation Tool for Big Data sets.

In this paper, we first present related work in the visualisation and data preparation domains (section 2). Then, we introduce our original in-memory data preparation tool in section 3. In section 4, we describe the challenges faced and solutions adopted when evolving its visualisations from in-memory fitting to Big Data sets. Finally, we present our conclusions and discuss potential directions of future work in section 5.

2 Related Work

Traditional data visualisation tools are usually restricted to small datasets, processed offline and limited to accessing and visualising pre-processed sets of static data.

In an attempt to handle the characteristics of the Big Data era, the research community has proposed different visualisation approaches [1].

The most common techniques are those of data reduction, which aim to summarise the dataset by using different approximations. The approaches followed for data summarisation include sampling (i.e. visualising a representative subset or filtering non-

contributing sets of data) [2, 3] and aggregation (i.e. visualising an aggregated or abstracted version of the dataset by using binning or clustering techniques) [4, 5].

The next set of proposed techniques target the hierarchical exploration of a dataset, allowing the visual exploration of large datasets at different levels of detail [6, 7]. These are computed by a hierarchical aggregation of the dataset, which allows the user to get a synopsis of the dataset and retrieve details of the data at different levels.

These two types of strategies (i.e. data reduction and hierarchical techniques) aim to contribute to the visual scalability characteristic discussed in the introduction. Other research has targeted the real-time interaction (with the dataset) feature by working on the progressive result delivery and different caching and prefetching strategies. Regarding progressive techniques, these tend to combine both user interaction-based dynamic result calculations [8] and incremental computation and delivery of results [9].

Moreover, visualisation approaches have been developed to tackle the dynamic nature of datasets by implementing incremental and adaptive strategies that allow for an on-the-fly exploration of large and dynamic datasets [6, 10].

Finally, regarding visualisation techniques, another area of research has focussed on assisting the user by recommending visualisations that are more appropriate for the specific characteristics of the data (or identified trends) [11], or the user behaviour and preferences [12].

Regarding commercial tools, a body of research work has analysed some popular visualisation tools (i.e. Tableau, PowerBI, Plotly, Gephi and Excel) and techniques, and how well they fit into the size, heterogeneity and dynamism properties of Big Data [13]. These tools are more focussed on visualising data prepared for analysis.

A recent market analysis report has analysed data preparation tools [14], studying the integration and exploration features, data manipulation features and user experience and user interface features among others, as part of the technical assessment of these tools. As the studied features prove, Big Data management and visualisation techniques are key to these data preparation and QoD improvement tools, whereas commercial tools are focussed on data preparation following the extract, transform, load (ETL) procedure, and not in the data exploration task for QoD assessment and improvement.

In this state-of-the-art context, we report on our initially developed in-memory data preparation and QoD improvement TAQIH tool, and on the experience of enhancing this tool from in-memory dataset visualisation and preparation to Big Data sets.

3 TAQIH – In-Memory Data Preparation Tool

TAQIH [15] is a data preparation tool developed to support non-technical users on 1) the exploratory data analysis (EDA) process of tabular health data, and 2) the assessment and improvement of its quality. A web-based tool was implemented with a simple yet powerful visual interface.

First, it provides interfaces to understand the dataset, to gain an understanding of the content, structure and distribution. Then, it provides data visualisation and data quality improvement utilities for the dimensions of completeness, accuracy, redundancy and readability [16].

TAQIH was designed and developed with in-memory data preparation technologies, so visualisation, data management and data transformations are limited to a single computer's memory and web-browser capabilities. Experimentally we have been able to manage datasets under 200MB using a desktop machine with 8 GB of RAM, but for providing the end-user with an acceptable interaction time (10 seconds for keeping the user's attention [17]), this is reduced to few tens of MBs. Data transformations are synchronously applied as they are requested, and visualisations updated accordingly.

TAQIH contains a main navigation bar at the top of the GUI, where items are placed from left to right following the usual iterative pipeline in EDA. First, 'General Stats' and 'Features' menu items provide global and detailed views of the data to gain insights about content, distribution and quality. Then, the 'Missing Values' section deals with the completeness dimension of data quality. After that, the 'Correlations' section presents the statistical relationship among variables, to help the identification of possible redundancies among variables or incoherent data, related to the redundancy and accuracy dimensions of data quality. Next, the 'Outliers' section identifies observations that differ significantly from others in the features and instances axes which is also related to accuracy, redundancy, readability and trust dimensions in data quality. All views include a small sample of the dataset (following data reduction by sampling) to help interpreting the dataset and identifying the transformation actions needed.

TAQIH is composed of both pre-processed property visualisation and summary visualisation (histograms or density plots), but also includes binary heatmaps (for missing values) or boxplots (for outliers) representing instance level data, which can be cumbersome when moving to very large datasets.

4 Enhancing Data Preparation Tool Visualisations for Big Data

Volumes considered in the Big Data context (i.e. large datasets not fitting in a computer's memory and expected to be increasing) impose new challenges over traditional datasets which could be totally managed in a computer's memory. When it comes to data preparation and QoD assessment, traditional Python-based or R-based methods do not directly handle datasets that do not fit into a computer's memory.

Additionally, many traditional general statistics or quality assessment algorithms need to keep global variables for their computation, for example, cardinality calculations might require expansion as large as the data source size. This makes existing data quality algorithms unsuitable for distributed parallel computing.

We have also identified two more issues when moving QoD assessment to large datasets: the visualisations used to allow the users to explore the data to evaluate its quality and that data preparation tasks cannot be run synchronously anymore.

Traditional visualisations (e.g. missing values or outliers) mainly work by plotting all the instances of the dataset, which requires pulling all instances from the dataset, and having the user's client applications manage all the data to visualise and respond to users' interactions. This is no longer feasible and it is unrealistic to expect the user to wait until a data cleansing task, over a large dataset that might require hours, is complete.

4.1 Migration to an Asynchronous Distributed Architecture

To overcome the data volume challenges identified, we have opted to use algorithms that provide approximations to evolve the TAQIH tool into an asynchronous processing framework. Big Data computing infrastructures have been used, for those algorithms which have distributable or parallelized versions, whilst for those requiring adaptations, state-of-the-art proposals have been implemented following Big Data computing approaches where possible, and per-chunk processing where more fine-grain control of shared global variables is required.

Figure 1 illustrates the previous TAQIH architecture contrasted to the architecture redesign for the GYDRA architecture.

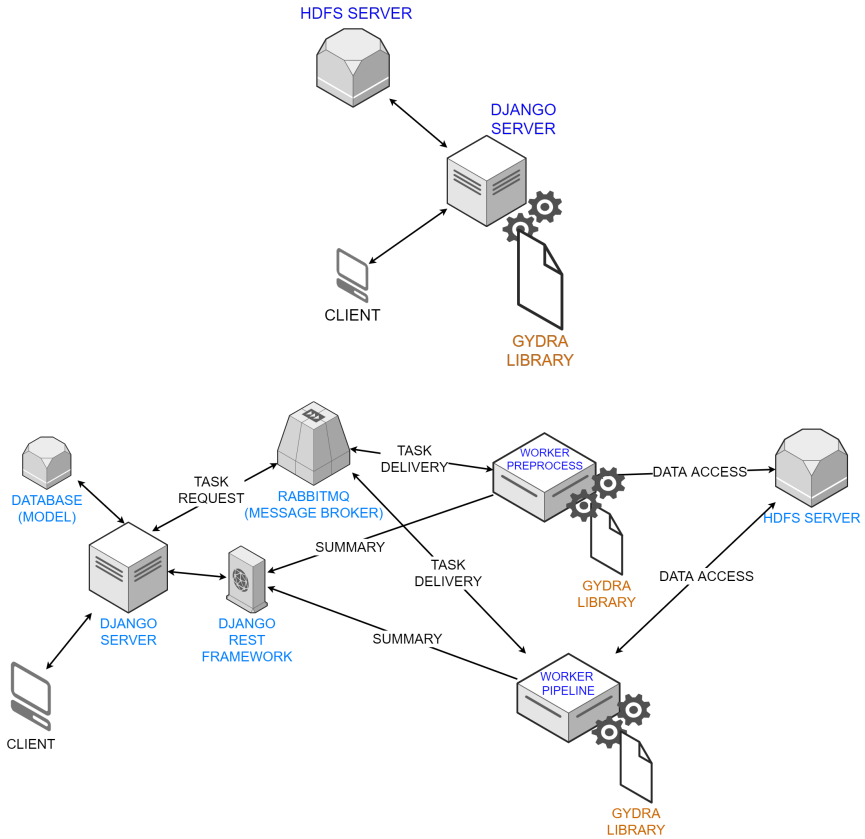

Fig. 1. TAQIH solution single machine focussed architecture (top) and GYDRA solution distributed architecture (bottom)

The GYDRA tool's user interface has been developed using the Django framework, HTML5, Asynchronous JavaScript (AJAX) and Bootstrap responsive web library. Distributed asynchronous tasking is managed through the Celery Distributed Task Queue tool with RabbitMQ as a message broker to implement the real task queue. The Celery

worker (depicted as Worker Preprocess and Worker Pipeline in Figure 1) can either run data pre-processing or transformation tasks by using the GYDRA Python library reading and handling HDFS stored datasets per chunks or by submitting applications to an Apache Spark cluster.

4.2 QoD Assessment Visualisation Updates

For the Big Data QoD indicators visualisation issues, approximations requiring a limited and controlled but representative amount of data to be displayed have been implemented. The computation and generation of the visualisation is performed by the asynchronous workers to reduce processing load and smooth the user experience on the client side. Subsequently, the different visualisation components of the GYDRA preparation tool are analysed individually.

Tab-Based Navigation Approach and General Stats
GYDRA has retained the main tab-based navigation approach and sample of the dataset described for TAQIH, while a new data transformation pipeline section has been added across the different views, to better understand the dataset and dynamically add transformations during EDA (since transformations have to be processed offline now).

Information summarisation is provided through a navigation approach by providing a hierarchical view, starting from 'General Stats' and moving to the 'Features' section, and by including a raw data sample across all sections. Next, the 'Missing Values', 'Correlations' and 'Outliers' sections are placed to assist the user through common EDA tasks. Figure 2 and Figure 3 depict the navigation approach and 'General Stats' sections of TAQIH and GYDRA tools respectively.

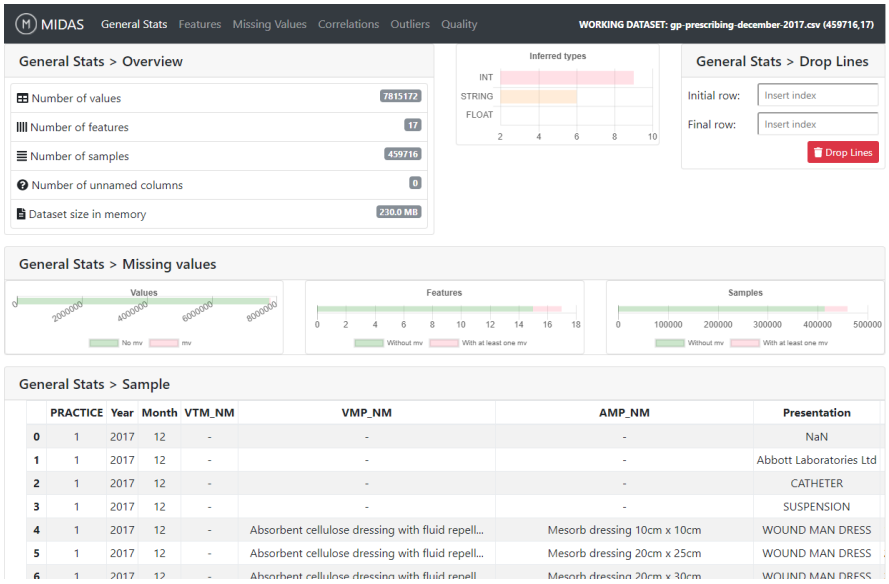

Fig. 2. TAQIH navigation and the 'General Stats' section. This figure is reproduced from [16]

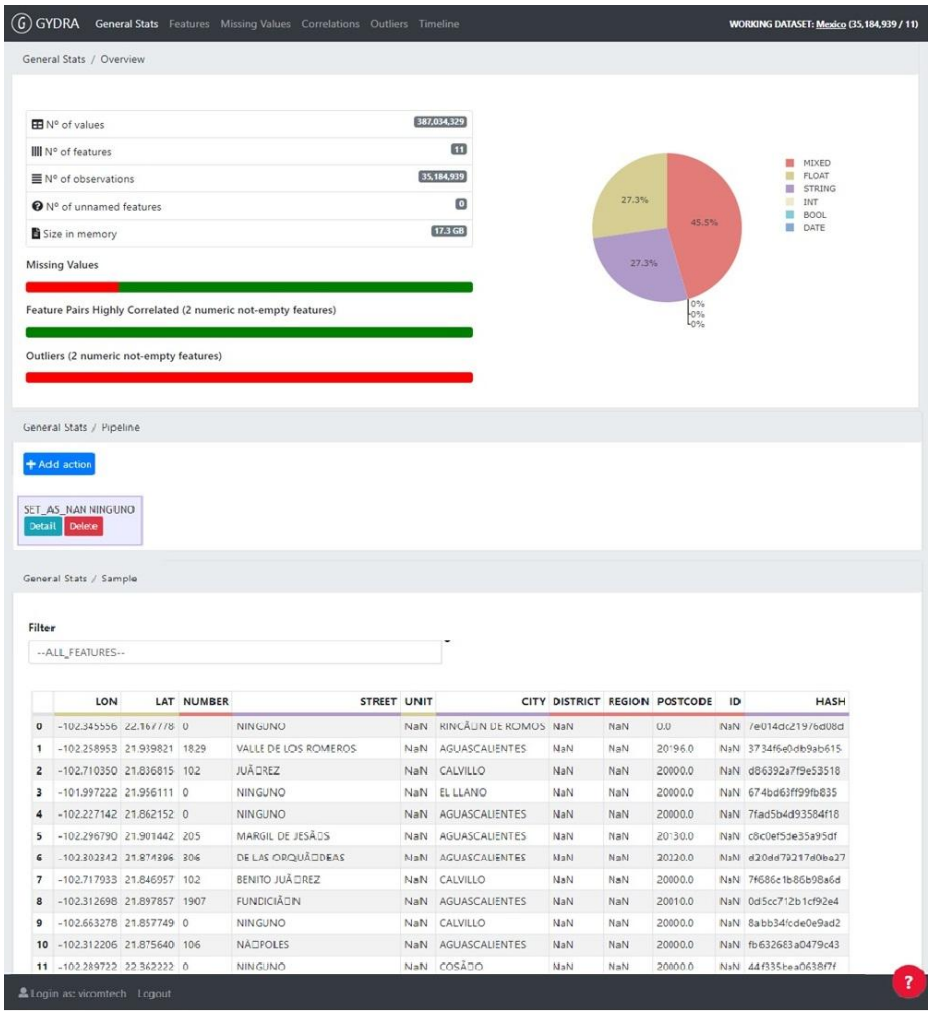

Fig. 3. GYDRA main application navigation and 'General Stats' section including transformations pipeline and a sample of the dataset

Features

Concerning the 'Features' section, visualisation remains quite similar, moving feature-related transformations from a feature specific section to the common transformation specification pipeline section. Additionally, cardinality and the number of appearances per each feature variable has been dropped, considering the scalability limitations when moving to big data sets. We have replaced this feature with the visualisation of the Top 10 values. Figure 4 and Figure 5 depict the feature analysis section of the TAQIH and GYDRA tools respectively.

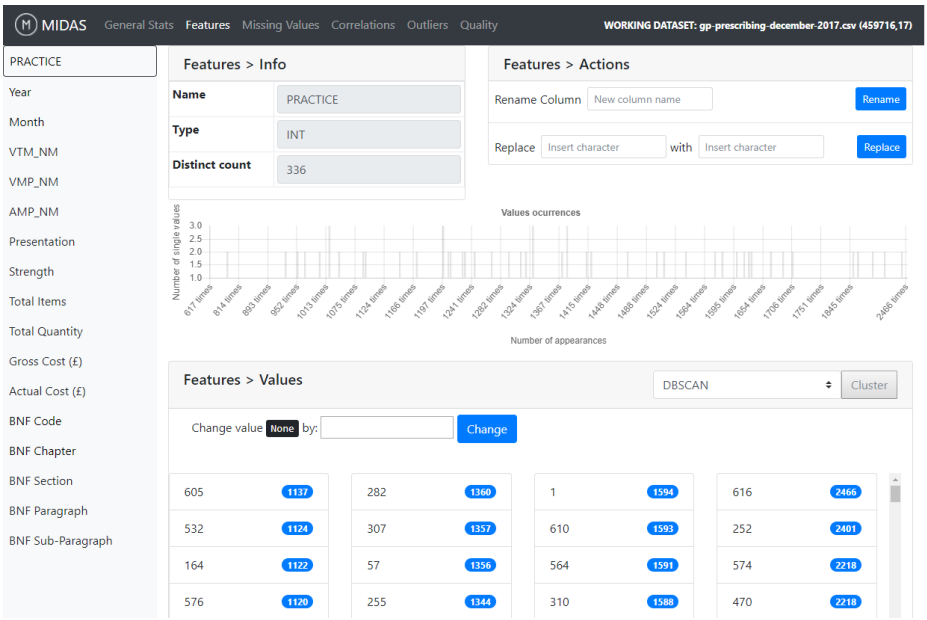

Fig. 4. TAQIH per feature analysis section. This figure is reproduced from [16]

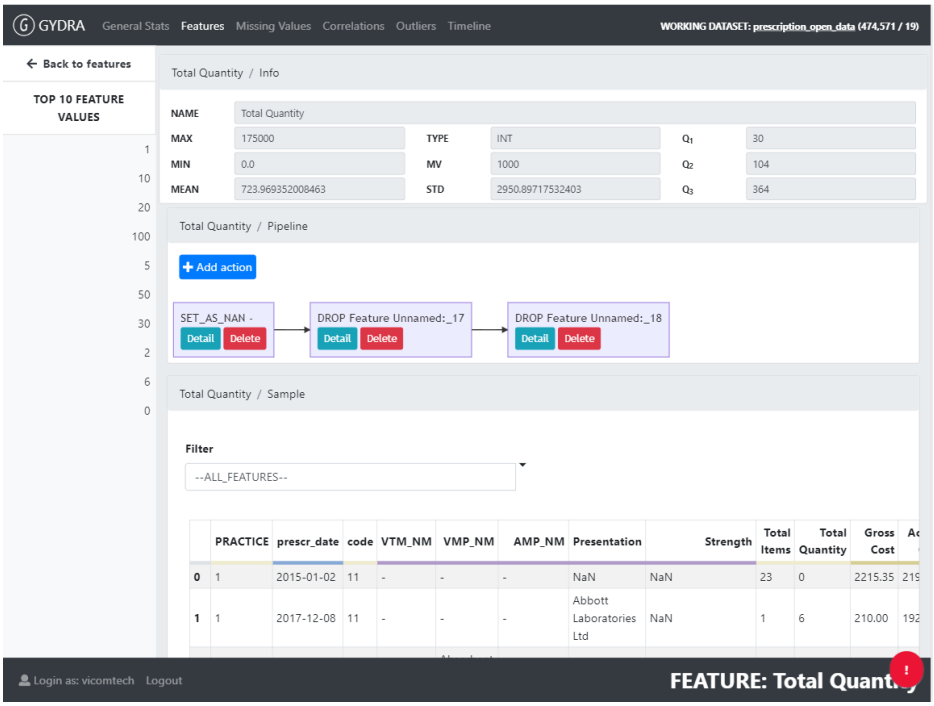

Fig. 5. GYDRA per feature analysis section.

Missing Values

For the Missing Values section, previously a binary heatmap with individual data was displayed, while in the GYDRA tool, percentages of missing values have been computed and visualised. TAQIH had a specific Missing Value imputation section, while in GYDRA, this has been moved to the common transformation pipeline. Figure 6 depicts the TAQIH tool's Missing values section, while Figure 7 shows the GYDRA tool's implementation.

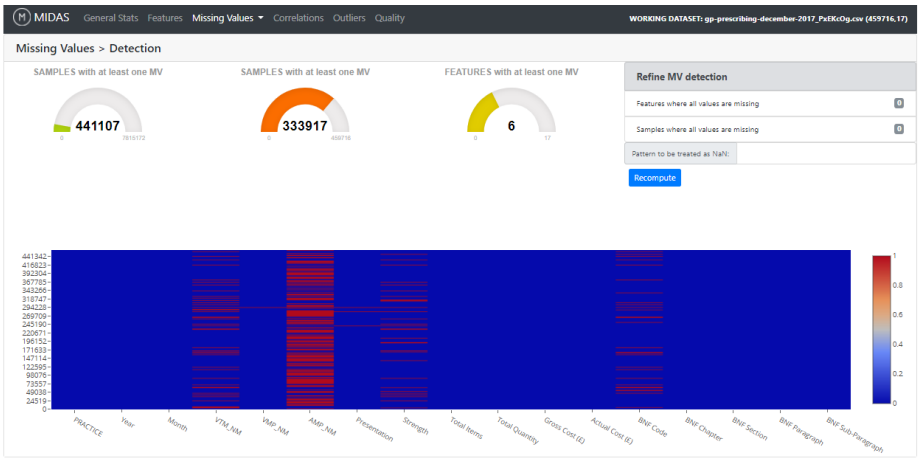

Fig. 6. TAQIH tool's Missing values section. This figure is reproduced from [16]

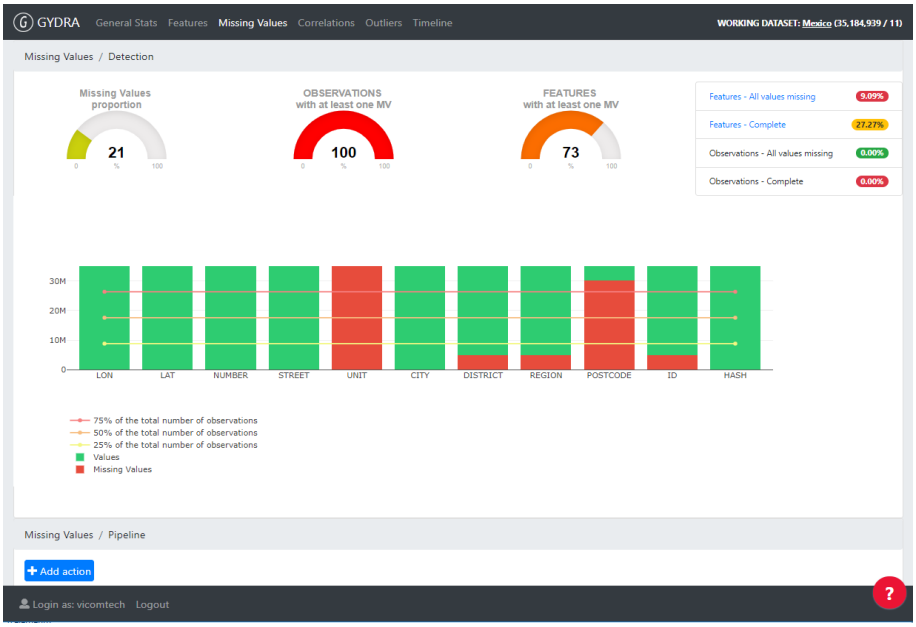

Fig. 7. GYDRA tool's Missing values section

Correlations

The Correlations section remains similar considering its most important visualisation is a correlation matrix, displaying feature association values (see Figure 8). The density plot used in TAQIH to compare two features among their value set was dropped.

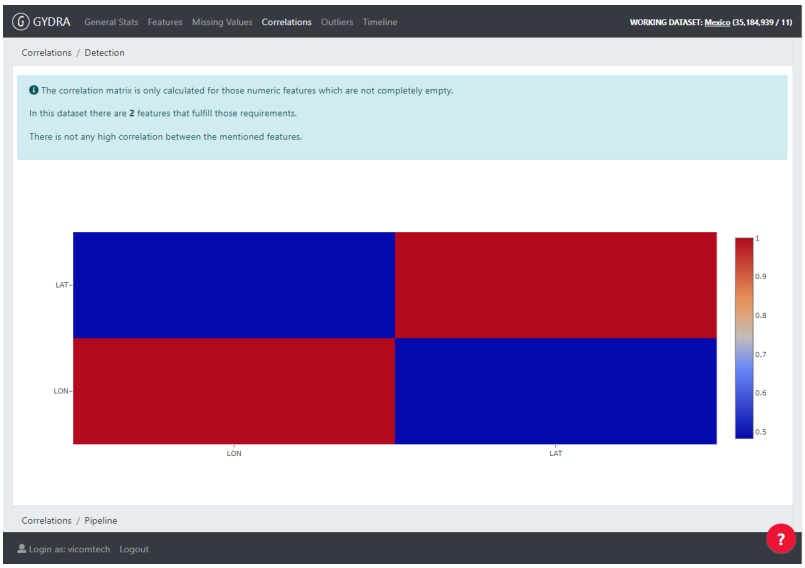

Fig. 8. GYDRA tool's Correlations section

Outliers

Regarding the Outliers section, in the TAQIH tool both a traditional box plot and a histogram with each different value occurrence (classified as in or out layer) were used. For GYDRA, a novel box plot visualisation has been proposed (see Figure 9). The outlier distribution is plotted as two histograms, one for low values and the other one for high values (suspected outliers and outliers according to Tukey algorithm). Multivariate outlier detection was dropped from TAQIH while an alternative for Big Data was implemented.

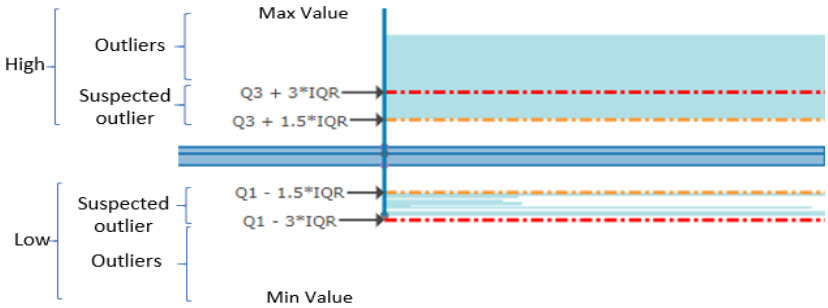

Fig. 9. GYDRA's Outlier diagram with outliers binning

5 Conclusions And Future Work

In this paper, we report on the enhancements implemented to a tabular data preparation and QoD improvement tool, focussed on its visualisation features, when moving from in-memory datasets to Big Data sets. Architectural and visualisation solutions adopted have been described accordingly. It was necessary to move from an in-memory synchronous architecture to an asynchronous distributed processing architecture to be able to operate the datasets, as well as the remote creation of visualisations. Asynchronous pre-processing of visualisations was adopted instead of on-the-fly processing, given the need to compute general statistics and per feature indicators (e.g. top n values or correlations). For Big Data sets it is not possible to calculate these indicators and provide a timely response to the user otherwise. Next, we have adopted different data reduction approaches to ensure visual scalability and meet local memory limitations for visualisation, introducing a novel box plot for Big Data.

Future work is planned on researching and implementing features lost in the transition from TAQIH to GYDRA. Next, the focus will be on the integration of streaming data and researching exploration techniques to help the user in understanding the dynamic of the sources to better define their ingestion pipelines. In line with this, enabling users to navigate through the dataset to find missing values should be enhanced to help users identify and understand underlying trends. Related to user assistance, machine learning techniques are being researched to support the user by suggesting appropriate preparation tasks.

Acknowledgments

This work was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 727721 (MIDAS).

This work was supported by the Gipuzkoan Science, Technology and Innovation Network Programme funding of the HIDRA project.

6 References

1. Bikakis, N.: Big Data Visualization Tools. In: Sakr, S. and Zomaya, A.Y. (eds.) Encyclopedia of Big Data Technologies. pp. 336–340. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-319-77525-8_109.
2. Battle, L., Stonebraker, M., Chang, R.: Dynamic reduction of query result sets for interactive visualizaton. In: 2013 IEEE International Conference on Big Data. pp. 1–8 (2013). https://doi.org/10.1109/BigData.2013.6691708.
3. Park, Y., Cafarella, M., Mozafari, B.: Visualization-aware sampling for very large databases. In: 2016 IEEE 32nd International Conference on Data Engineering (ICDE). pp. 755–766 (2016). https://doi.org/10.1109/ICDE.2016.7498287.

4. Jugel, U., Jerzak, Z., Hackenbroich, G., Markl, V.: VDDA: automatic visualization-driven data aggregation in relational databases. VLDB J. 25, 53–77 (2016). https://doi.org/10.1007/s00778-015-0396-z.
5. Lins, L., Klosowski, J.T., Scheidegger, C.: Nanocubes for Real-Time Exploration of Spatiotemporal Datasets. IEEE Trans. Vis. Comput. Graph. 19, 2456–2465 (2013). https://doi.org/10.1109/TVCG.2013.179.
6. Bikakis, N., Papastefanatos, G., Skourla, M., Sellis, T.: A hierarchical aggregation framework for efficient multilevel visual exploration and analysis. Semantic Web. 8, 139–179 (2017). https://doi.org/10.3233/SW-160226.
7. Elmqvist, N., Fekete, J.-D.: Hierarchical Aggregation for Information Visualization: Overview, Techniques, and Design Guidelines. IEEE Trans. Vis. Comput. Graph. 16, 439–454 (2010). https://doi.org/10.1109/TVCG.2009.84.
8. Stolper, C.D., Perer, A., Gotz, D.: Progressive Visual Analytics: User-Driven Visual Exploration of In-Progress Analytics. IEEE Trans. Vis. Comput. Graph. 20, 1653–1662 (2014). https://doi.org/10.1109/TVCG.2014.2346574.
9. Im, J.-F., Villegas, F.G., McGuffin, M.J.: VisReduce: Fast and responsive incremental information visualization of large datasets. In: 2013 IEEE International Conference on Big Data. pp. 25–32 (2013). https://doi.org/10.1109/BigData.2013.6691710.
10. Zoumpatianos, K., Idreos, S., Palpanas, T.: Indexing for interactive exploration of big data series. In: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. pp. 1555–1566. Association for Computing Machinery, Snowbird, Utah, USA (2014). https://doi.org/10.1145/2588555.2610498.
11. Mackinlay, J., Hanrahan, P., Stolte, C.: Show Me: Automatic Presentation for Visual Analysis. IEEE Trans. Vis. Comput. Graph. 13, 1137–1144 (2007). https://doi.org/10.1109/TVCG.2007.70594.
12. Gotz, D., Wen, Z.: Behavior-driven visualization recommendation. In: Proceedings of the 14th international conference on Intelligent user interfaces. pp. 315–324. Association for Computing Machinery, Sanibel Island, Florida, USA (2009). https://doi.org/10.1145/1502650.1502695.
13. S. M. Ali, N. Gupta, G. K. Nayak, R. K. Lenka: Big data visualization: Tools and challenges. In: 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I). pp. 656–660 (2016). https://doi.org/10.1109/IC3I.2016.7918044.
14. Ovum: Ovum Decision Matrix: Selecting a Self-Service Data Prep Solution, 2018–19. (2018).
15. Álvarez Sánchez, R., Beristain Iraola, A., Epelde Unanue, G., Carlin, P.: TAQIH, a tool for tabular data quality assessment and improvement in the context of health data. Comput. Methods Programs Biomed. SI Data Qual. Assess. 181, 104824 (2019). https://doi.org/10.1016/j.cmpb.2018.12.029.
16. The Dama UK Working Group: The Six Primary Dimensions For Data Quality assessment, https://www.dqglobal.com/wp-content/uploads/2013/11/DAMA-UK-DQ-Dimensions-White-Paper-R37.pdf, last accessed 2018/03/08.
17. Nielsen, J.: Usability Engineering. Morgan Kaufmann, Amsterdam (1993).