GUIDED REFLECTIVE EXPLORATORY ANALYSIS

D13

IDEALVis Consortium

http://idealvis.inspirecenter.org/

Executive Summary

The overall objective of Work Package 5 (WP5) is to design and develop an intelligent data analytics component that facilitates the enhancement of a data analyst's efficiency and effectiveness with respect to data exploration processes by: (i) quantifying the user's experience through recorded interactions; and (ii) identifying effective analysis practices through machine learning techniques.

This deliverable presents the set of mechanisms developed for recording the user's interaction with the IDEALVis platform. The first mechanism is a tracker that enables the platform to record the navigation of the user when performing a specific data analysis task. The second mechanism is a tracker that records the time spent by a user actively gazing (i.e., viewing/interpreting) a specific analysis report produced by the platform. The last mechanism is a dashboard that analyses/aggregates the collected data and provides a personalized assessment of the user's efficiency with respect to his/her overall data analysis process (e.g., the time consumed on specific tasks, any repetitive execution of same tools, etc). Through this dashboard the users can review their analytical practices by comparing against the performance of other users, which allows them to understand their relative strengths and weaknesses. The final section of this deliverable provides an analysis of all users' tracked interaction records, which have been captured during the pilot user study.

Table of Contents

List of Figures

1 Introduction

One of the goals of IDEALVis is to provide data analysts with the necessary tools for quantifying and comparing the efficiency and effectiveness of their analysis workflows with that of other analysts. To this end, the *Analysis Tracker Dashboard* is powered by two tracking mechanisms (*Analysis Construction Tracker* or ACT, and *Analysis Viewing Tracker* or AVT), which monitor and capture the data analyst's interactions, as well as several procedures that analyze and aggregate the tracking data to produce meaningful quantitative reports and relevant statistics. Indicatively, the tracking mechanisms collect information about (i) the user; (ii); the data analysis construction process (e.g., duration and frequency of each step of the Analysis Wizard) including the task's query specification parameters and timestamps; and (iii) the time spent by a user actively gazing (i.e., viewing/interpreting) a specific analysis report produced by the platform. Using this information, the dashboard allows for quickly identifying personal bottlenecks and/or inefficiencies related to the methodology adopted by analysts, as well as comparing their performance on specific tasks with that of other users.

1.1 Deliverable Structure

The remainder of this deliverable is organized as follows:

- Section 2 presents the IDEALVis platform's Realtime Activity Tracking mechanisms, their operations, the collected data, and their persistence.

- Section 3 presents the *Analysis Tracker Dashboard*, the insights it provides to users, and an overview of the procedures that generate those insights.

- Section 4 presents the analysis of all the users' tracked interaction records, which were captured during the pilot user study.

- Section 5 concludes this deliverable.

2 Realtime Activity Tracking

In IDEALVis, the user's interaction is captured seamlessly and in real time, requiring no additional input from the corresponding user. This is achieved via the embedded Realtime Activity Tracker module, which uses two mechanisms, the Analysis Construction Tracker, and the Analysis Viewing Tracker, both of which are described in the following sections.

2.1 Analysis Construction Tracker

The Analysis Construction Tracker (ACT) runs on the client-side (i.e., it executes on the user's browser) and captures information regarding the user's interaction with the Analysis Wizard, which is used for constructing analysis reports to address a particular analysis task. When the user initiates a new analysis session through the wizard, the ACT starts collecting data as the user progresses through the wizard's steps. Steps tracked by ACT include the (i) Select Dataset step; (ii) Select Analysis step; (iii) Select Attributes step; and (iv) Select Output step (Figure 1). The data collected includes the time spent by a user on each of the above steps (measured in milliseconds). Since the wizard allows a user to navigate forward (i.e., to the next step) and backwards (i.e., to the previous step) when constructing an analysis, the tracker continuously aggregates the total time a particular step is visited by the user.

Figure 1 - Analysis Wizard Steps

At the wizard's final step (i.e., Final Result), users can view the report produced by their analysis. At this step, the time is not recorded by the ACT. Instead, the ACT uses this step, which marks the end of the analysis construction process, to persist all recorded information for to the database. This is achieved using an asynchronous HTTP POST request that runs in the client's browser background.

Assuming that the user has already constructed a particular analysis, and has explored the resulting analysis report, at any point in time he/she may opt to navigate backwards to a previous wizard step. In such cases, the ACT spawns a new tracking instance (i.e., the user's time on each step is reset to 0).

2.1.1 ACT: TRACKING DATA AND PERSISTENCE

During the final step of the Analysis Wizard, the user's analysis/query specification and unique identifier are submitted for processing. The former is annotated with the ACT tracking data. Prior to persisting the tracking records to the database, the system executes the user's analysis to produce the corresponding analysis report's data and metadata (i.e., the Visualization Engine instructions generated from the system's Adaptation Engine, which guide how the resulting data is to be visualized). The reason the analysis is run prior to persisting the ACT tracking records is because along with the tracking records, the system requires to also persist the generated metadata, which carries the adaptation engine's visual modifications and visualization rendering definitions.

Figure 2 illustrates the database table used to store all ACT tracking records and related metadata. Below a short explanation is provided for each data property included in this table.

- **Id:** A unique record identifier, representing a specific wizard usage/construction instance.

- **Configuration:** The analysis/query specification representing all of a user's selections in the specific wizard instance.

- **UserId:** The unique identifier of the user related to this tracking record.

- **AnalysisReportMetadata:** The analysis report metadata that resulted from the user's executed analysis/query specification.

- **AnalysisTaskId:** The analysis task the user was addressing during the specific wizard instance.

- **SelectDatasetDuration:** The total time spent selecting a dataset for the analysis.

- **SelectAnalysisDuration:** The total time spent selecting an analysis procedure.

- **SelectAttributesDuration:** The total time spent selecting data attributes for the analysis.

- **SelectOutputDuration:** The total time spent selecting an output method for the analysis.
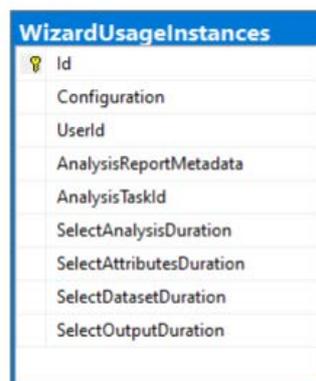
Figure 2 - Wizard Usage Instances Database Table for ACT

When the ACT tracking record is successfully persisted, the database automatically assigns a unique identifier to the persisted wizard usage/construction tracking record (i.e., **Id**). Prior to returning the resulting analysis report, the report is annotated with the specific tracking record's unique identifier. Therefore, every analysis report in the system has a unique wizard usage/construction instance associated with it. This association is later used by the AVT.

2.2   Analysis Viewing Tracking

Like ACT, the Analysis Viewing Tracker (AVT) runs on the client-side. This tracker can determine whether the user is actively gazing (i.e., viewing /interpreting) an analysis report found in the IDEALVis platform. Analysis reports are available in the system in two locations: (i) at the final step of the Analysis Wizard; or (ii) on the user's dashboard. The AVT records the time a user spends on viewing a particular analysis report. The mechanism infers that a user is actively focusing on a specific analysis report when three rules are met: (i) the user's browser is not minimized and has active focus; (ii) the selected/active tab in the user's browser is the one corresponding to the

platform; and (iii) a specific analysis report, i.e., data visualization, is visible in the available browser window, denoting that the report is almost at the center of the screen. The AVT initiates the tracking process only once the user has maintained at least 1.5 seconds of active focus (i.e., initial view time) on the analysis report. This threshold indicates that the user has paused navigating, and thus the analysis report found on screen is the report of interest. AVT persists tracking data to the database using an asynchronous HTTP POST request that runs in the client's browser background. The persistence operation is initiated only once the user decides to switch focus from the current analysis report to something else (i.e., every time one of the above focus rules is invalidated).

2.2.1   AVT: TRACKING DATA AND PERSISTENCE

As described in Section 2.1.1, every analysis report found in the system is always annotated with a specific Analysis Wizard usage/construction instance unique identifier. This unique identifier enables the system to locate the source, or the process carried out by the user to produce a specific analysis report. Additionally, the identifier is required by the AVT when persisting a tracking view time record for a particular user. Figure 3 (left) showcases the database table used by the AVT to store all user's viewing instances for every analysis report. The tracking records of the AVT are linked to the corresponding tracking records of the ACT, shown in Figure 3 (right), using the WizardUsageInstanceId foreign key. This relationship denotes that a user can have multiple viewing instances related with a single analysis report, which is represented in the system by an ACT tracking record.

Below a short explanation is provided for each data property included in the AVT's records table.

- **Id:** The unique identifier of a viewing instance.

- **WizardUsageInstanceId:** The unique identifier of the analysis report originating from the ACT tracking records.

- **ViewTime:** The time in milliseconds for the specific viewing instance.
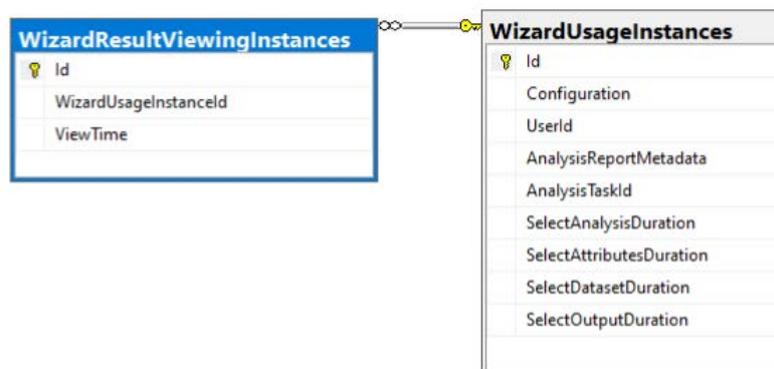
Figure 3 - Wizard Result Viewing Instances Database Table for AVT

3 Analysis Tracker Dashboard

This section presents the *Analysis Tracker Dashboard* and discusses the insights it offers to a business data analyst. Moreover, an overview of the procedure employed to generate this information is provided.

All data collected by both the ACT and AVT mechanisms are compiled into detailed analytics revealing: (i) analysis construction patterns and tactics adopted by each data analyst; and (ii) insights regarding the user's data exploration practices. These analytics are accessible in the Analysis Tracker Dashboard, which is available to all analyst users, thus enabling a better understanding of their relative strengths and weakness across all / or specific tasks. The dashboard is broken down in two sections: (i) the user's overall analysis statistics; and (ii) the user's analysis statistics for specific tasks performed using the platform (e.g., the analysis tasks provided during the pilot study). A description of both sections is provided next.

3.1 Overall Analysis Statistics

The overall analysis statistics section is illustrated in Figure 4. This section of the dashboard aims at providing the user with a set of key interaction and performance indicators, presented as widgets on the top of the dashboard. Selected indicators are also plotted in a column chart against the average indicator values across all other system users. These statistics consider all user analysis interactions, i.e., interactions related to the construction, or the exploration of all analysis reports produced within the system, and are not tied to a specific analysis task.
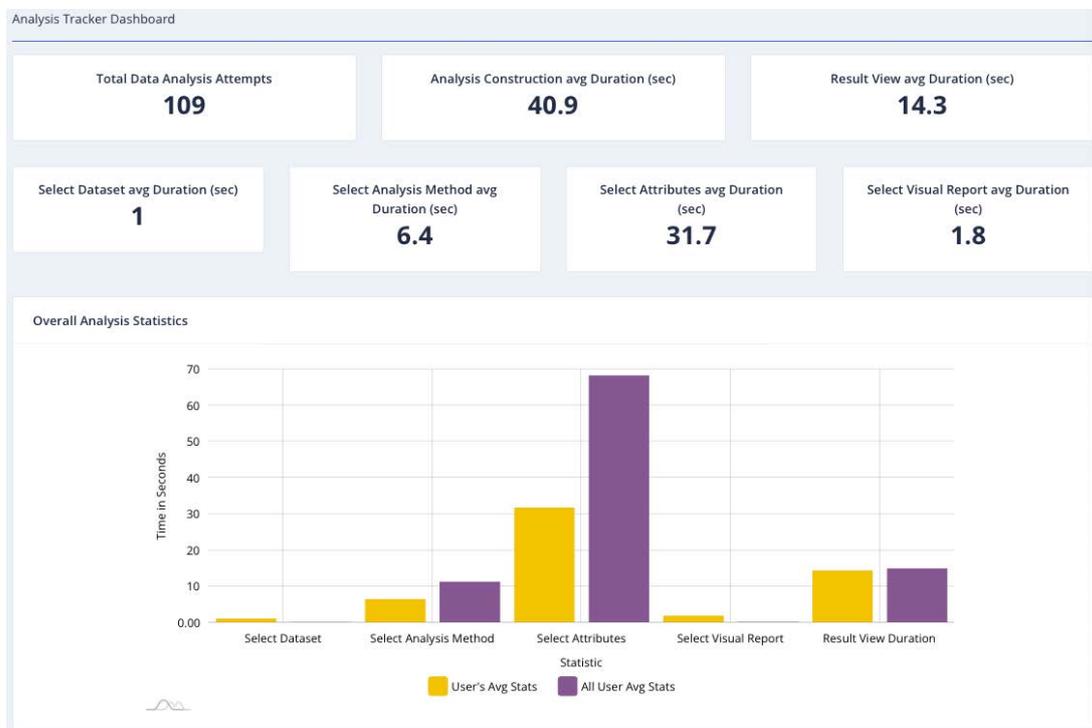
Figure 4 - Overall Analysis Statistics (Analysis Tracker Dashboard)

The list below summarizes each of the dashboard indicators in Figure 4 and provides an overview of the procedure that generates this information.

Total Data Analysis Attempts: This indicator specifies the total number of times a user utilised the Analysis Wizard to produce an analysis report. This value is retrieved by counting the total number of tracking records persisted by the ACT for a particular user. This is shown only in the dashboard's widgets.

Select Dataset Duration: This indicator presents the average time (in seconds) spent by a user on selecting a dataset when constructing an analysis report via the Analysis Wizard. This value is retrieved by performing an average aggregation on the **SelectDatasetDuration** property of ACT tracking records for a particular user. This indicator is shown both in the dashboard's widgets and the overall analysis statistics column chart. The column chart presents the value of this indicator for the user against the average indicator value for all system users.

Select Analysis Method Duration: This indicator presents the average time (in seconds) spent by a user on selecting an analysis method when constructing an analysis report via the Analysis Wizard. This value is retrieved by performing an average aggregation on the **SelectAnalysisDuration** property of ACT tracking records for a particular user. This indicator is shown in both the dashboard's widgets and the overall analysis statistics column chart. The column chart presents the value of this indicator for the user against the average indicator value for all system users.

Select Attributes Duration: This indicator presents the average time (in seconds) spent by a user on selecting the analysis data attributes when constructing an analysis report via the Analysis Wizard. This value is retrieved by performing an average aggregation on the **SelectAttributesDuration** property of ACT tracking records for a particular user. This indicator is shown in both the dashboard's widgets and the overall analysis statistics column chart. The column chart presents the value of this indicator for the user against the average indicator value for all system users.

Select Visual Report Duration: This indicator presents the average time (in seconds) spent by a user on selecting the type of data visualization to be used in their analysis report (one of the steps performed when constructing an analysis report via the Analysis Wizard). This value is retrieved by performing an average aggregation on the **SelectOutputDuration** property of ACT tracking records for a particular user. This indicator is shown both in the dashboard's widgets and the overall analysis statistics column chart. The column chart presents the value of this indicator for the user against the average indicator value for all system users.

Analysis Construction Duration: This indicator presents the total time (in seconds) spent on average by a user on constructing an analysis report using the Analysis Wizard. This value is retrieved by summing the values of the following user's indicators: Select Dataset Duration, Select Analysis Method Duration, Select Attributes Duration, and Select Visual Report Duration.

Result View Duration: This indicator presents the average time (in seconds) spent by a user on viewing an analysis report. This value is retrieved by performing an average aggregation on the **ViewTime** property of AVT tracking records for a particular user. This indicator is shown in both the dashboard's widgets and the overall analysis statistics column chart. The column chart presents the value of this indicator for the user against the average indicator value for all system users.

3.2 Specific Task Analysis Statistics

The Analysis Tracker Dashboard's section that presents analysis statistics across specific analysis tasks is illustrated in Figure 5. This section of the dashboard summarizes statistics that are specific to the user's exploration interactions performed for a specific analysis task.

For instance, during the pilot study, all participants were asked to address a set of 38 analysis tasks using the Analysis Wizard. As part of this process, each user had to choose an analysis task before being able to utilise the wizard. All tracking records are classified in the system as belonging to a specific analysis task. Using this classification, the dashboard can produce statistics per analysis task a user attempts to solve. Figure 5 presents only a part of the dashboard summarising statistics for two of the analysis tasks a particular user has addressed.

The statistics presented for each analysis task use the same indicators as the ones described in the previous section. However, only the tracking data from the analysis instances related to a specific analysis task is used.
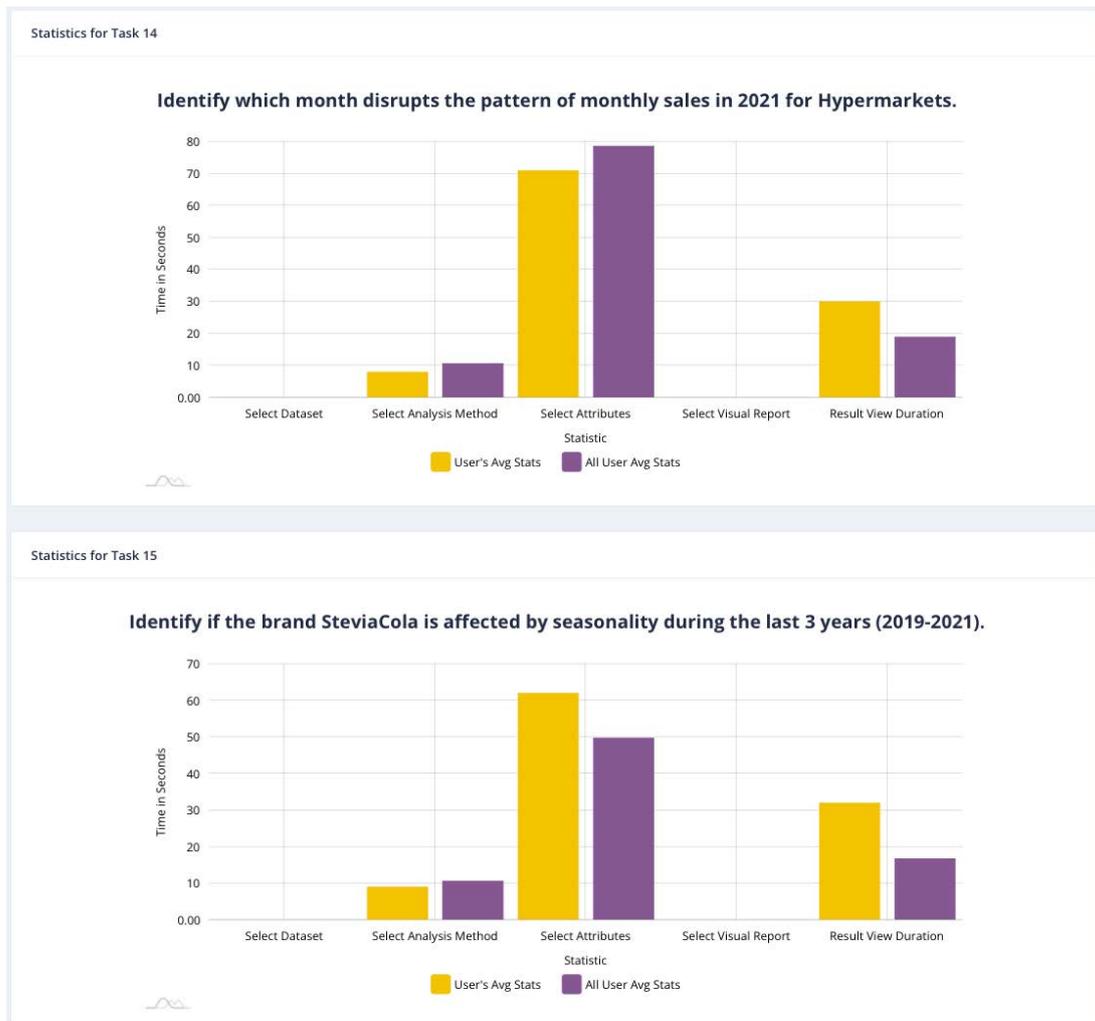
Figure 5 - Specific Task Analysis Statistics (Analysis Tracker Dashboard)

The indicators **Select Dataset** and **Select Visual Report** for both analysis tasks presented in Figure 5 have a value of 0. This is due to the fact that the tasks were used during the pilot study, where the Select Dataset and Select Output steps of the Analysis Wizard were disabled.

In summary, the statistics for analysis tasks 14 (Figure 5 top) and 15 (Figure 5 bottom) can be interpreted as follows:

- When exploring data for both analysis tasks, the user was on average faster than the rest of the users in selecting the appropriate analysis method.

- When exploring data for analysis task 14, the user was on average faster than the rest of the users in selecting the appropriate analysis data attributes. This statement though, is not true for analysis task 15 since the user was on average slower than the rest of the users in selecting the appropriate analysis data attributes.

- Finally, compared to others, the user spent more time on average on exploring (i.e., viewing/interpreting) the analysis reports of tasks 14 and 15.

4 Analyzing Users' Interaction Records

During the IDEALVis pilot study, the ACT and AVT mechanisms were constantly tracking the users' analysis interaction and exploration. This section provides an analysis of the tracked interaction records captured from 45 users. This analysis captures a wholistic view on the practices adopted by all users regardless of their characteristics. The metrics explored are the following:

- **Data Analysis Attempts**
- **Select Analysis Method Duration**
- **Select Attributes Duration**
- **Result View Duration**

As explained above, the reason why only four metrics are presented is because during the pilot study, some of the steps of the Analysis Wizard (i.e., Select Dataset and Select Output) were unavailable/disabled for the users' data exploration; thus, records for those analysis steps were not captured.

4.1 Identified Interaction Patterns

Data Analysis Attempts: During the pilot study, all users had to complete a total of 38 analysis tasks. For addressing each of the analysis tasks, a user had to perform at least a single data analysis attempt (i.e., construct a single analysis report using the Analysis Wizard). Users, however, were given the option to perform unlimited analysis attempts. Analysis results revealed that on average, a user performed 50.6 ± 6.8 analysis attempts to address all 38 analysis tasks. The user with the minimum number of data analysis attempts throughout the study performed 38 attempts, while the user with the maximum number of attempts performed 66 attempts.

Select Analysis Method Duration: On average, users during the pilot study spent 11.0 ± 6.4 seconds on selecting the appropriate analysis method/type for constructing their analysis report. The fastest user in terms of selecting an analysis method achieved an average time of 5 seconds, while the slowest user had an average time of 34 seconds for this metric.

Select Attributes Duration: On average, users during the pilot study spent 69.7 ± 21.7 seconds on selecting the required data attributes for constructing their analysis report. The fastest user in terms of selecting data attributes achieved an average time of 36 seconds, while the slowest one had an average time of 151 seconds for this metric. This is the analysis step for which all users spent most of their time, which is explained by the fact that it entails performing most of the analysis configuration.

Result View Duration: On average, users during the pilot study spent 16.0 ± 5.4 seconds on viewing/interpreting analysis reports. The fastest user achieved an average time of 6.7 seconds, while the slowest one had an average time of 26.6 seconds for this metric.

5 Conclusions

This deliverable provided an overview of the Analysis Tracker Dashboard, which enables data analysts: (i) to quickly identify their bottlenecks or inefficiencies related to the adopted analysis methodologies used when addressing analysis tasks; (ii) to compare their efficiency and effectiveness with that of other users across specific tasks; (iii) to explore their overall performance across all analysis tasks they perform; and (iv) to better allocate or reallocate their time to different types of analysis tasks based on their performance statistics.

The different sections of the dashboard were described by discussing the key indicators that enable a user to better understand their relative performance with regards to analysis construction and exploration across different tasks. The dashboard's underlying, real-time activity tracking mechanisms were described, which include the Analysis Construction Tracker and Analysis Viewing Tracker mechanisms. Finally, the deliverable presented an analysis of all users' interaction records during the project's corresponding pilot study, in an attempt to characterize the user practices adopted.