Gathering Real OLAP Analysis Sessions: A Feedback

Julien Aligon*

*Laboratoire d'Informatique – Université François Rabelais Tours, France
julien.aligon@univ-tours.fr

Abstract. The use of OLAP sessions, conducted by professional analysts, seems to be the best way to assess the relevance of OLAP solutions based on former queries (in particular with user-centric approaches, like recommendation or personalization of queries). However, for scholar research teams, obtaining such logs is often difficult. Moreover, the complexity of the queries produced in these logs can lead to an important treatment of them, denaturing the performed analysis. In this paper, we propose a feedback from real OLAP sessions performed by graduate students in Business Intelligence. This feedback reports the design of questionnaires and the use of an original user interface to easily conduct real OLAP sessions.

1 Introduction

In the context of relational databases ((Chaudhuri et al., 2003), (Golfarelli, 2003), (Khoussainova et al., 2011), (Akbarnejad et al., 2010)) or multidimensional databases ((Aligon et al., 2011), (Aligon et al., 2013b)), the use of logs is essential for assessing the relevance of solutions based on former queries. It is obvious that real logs (and more generally real data) are most relevant to assess the user-centric approaches, like recommendation or personalization of queries. Unfortunately, it is often difficult for scholar research teams to obtain real logs from professional analysts, in particular by the fact they can contain sensitive data. Even when the case exists, the complexity of the queries can lead to an important treatment of them. Indeed, the user-centric approaches are generally based on more basic query definition than those implemented in the tools used by professional analysts. Consequently, the number of queries can be strongly reduced, or too simplified. Finally the analysis performed in these types of logs can be denatured.

In the context of multidimensional databases, we propose in this paper a feedback reporting the gathering of real logs, according to a pre-defined query model. Precisely, this feedback is based on tests conducted with graduate students in Business Intelligence. Indeed, it has been assumed in (Runeson, 2003) that graduate students could perform analysis sessions as good as professional analysts. In order to control the type of queries we have to generate, we propose a new user interface for designing OLAP sessions. Note that the aim of this paper is not to propose a benchmark of OLAP sessions, but is a first approach for this long-term perspective by describing a feedback from sessions designed by students.

The paper is organized as follows. A related work is given Section 2. Section 3 describes the database as the OLAP cube model used during the tests with the students. Section 4 refers to the design of the questionnaires and the original GUI allowing to easily conduct OLAP sessions, according to a pre-defined query model. Section 5 gives statistical results about the collected logs. Section 6 concludes the paper and discusses perspectives.

2 Related work

The problem of lack of real logs (and its consequences) is well summarized in (Gerhard Weikum, 2013), which reads in part: "Academic research on Big Data is excessively based on boring data and nearly trivial workloads. On the other hand, Big Data research aims to obtain insights from interesting data and cope with demanding workloads. This is a striking mismatch."

Frequently, the tests for assessing the solutions based on former queries in the database context (multidimensional or not) are regularly performed with synthetic logs, as in (Chaudhuri et al., 2003), (Golfarelli, 2003), (Akbarnejad et al., 2010), (Aligon et al., 2013b) or (Aligon et al., 2011). (Akbarnejad et al., 2010) uses the SkyServer[1] query log for recommending queries in a relational context. But as depicted in (Singh et al., 2007), an important filtering has been given to the SkyServer query log. For example, the division of the log in different sessions has been arbitrarily conducted by setting a period of time between them. Thus, the SkyServer query log cannot ensure exact sessions. (Chaudhuri et al., 2003) and (Golfarelli, 2003) especially use the TPC-H benchmarch to form their logs. Even if the queries of this benchmark are closed to those performed by the industry, they cannot represent real analysis sessions.

In order to get as close as possible to real sessions, several solutions propose synthetic generators which try to simulate the behaviors of analysts. For instance in (Aligon et al., 2011) and (Aligon et al., 2013b), a session is generated by producing a sequence of queries which the result of the last one has a significant aggregate (supposed interesting for an analyst). In (Aligon et al., 2013a), different patterns between sessions are proposed for testing the relevance of OLAP similarity measures. But none of these works demonstrate that the synthetic sessions would be the same as those performed with analysts. Indeed, these synthetic generators always use objective quality criteria such as heterogeneity and closeness. They are unable to produce sessions with more subjective criteria (for instance by designing sessions with different difficulty degrees). Our paper is precisely focused on the subjectivity of the designed analysis.

By proposing in this paper a test conducted with graduate students, we can intuitively think that the performed analysis sessions are necessarily of lower quality than sessions provided by professional analysts. In the context of relational databases, a test in (Khoussainova et al., 2011) has already been conducted with students to demonstrate that browsing through past SQL query sessions helped speed up query composition. Moreover, (Runeson, 2003) supposes that graduate students can perform analysis sessions as good as experimented analysts, even if more investigation is needed. On the contrary, freshmen students are not recommended for conducting analysis sessions.

[1] Skyserver. http://www.skyserver.org.

3 Preliminaries for the test

In this section, we present the context of the test conducted with graduate students and then we describe the multidimensional schema, the query model and the data used.

3.1 Context of the Test

The aim of the test is to gather analysis sessions devised by graduate students in Business Intelligence (from the University Francois Rabelais of Tours (France) and the University of Bologna (Italy)). The students have to play the role of analysts and devise sessions to answer pre-defined analysis needs (detailed in Section 4.1) The different sessions of all students populate an OLAP query log. Note that this test is not intended to provide a benchmark of OLAP sessions but is a first approach to complete this project in a long term perspective.

3.2 Cube and query models

As discussed in Section 1, we want to limit the possibilities for querying an OLAP cube. Indeed, the devised sessions will use to assess future user-centric works. In our case, we want to consider the following cube and query definitions:

Definition 3.1 (Multidimensional Schema) *A* multidimensional schema *(briefly, a* schema*) is a triple $\langle L, H, M \rangle$ where:*

- *$L = \{l_1, \dots l_p\}$ is a finite set of* levels*, i.e., categorical attributes;*
- *$H = \{h_1, \dots, h_n\}$ is a finite set of* hierarchies*, each characterized by (1) a subset $Lev(h_i) \subseteq L$ of levels and (2) a* roll-up *tree-structured partial order $\succeq_{h_i}$ of $Lev(h_i)$;*
- *$M = \{m_1, \dots, m_l\}$ is a finite set of* measures*, i.e., numerical attributes.*

Definition 3.2 (Group-by Set) *Given schema $\langle L, H, M \rangle$, let $Dom(H) = Lev(h_1) \times \dots \times Lev(h_n)$; each $G \in Dom(H)$ is called a* group-by set *of the schema.*

Definition 3.3 (OLAP Query) *A* query *over schema $\langle L, H, Meas \rangle$ is a triple $q = \langle G, P, M \rangle$ where:*

1. *$G \in Dom(H)$ is the query group-by set;*
2. *$P = \{p_1 = L_1 \in X_1, \dots, p_n = L_n \in X_n\}$ is a set of predicates, one by dimension, whose conjunction is of the form $L_1 \in X_1 \wedge \dots \wedge L_n \in X_n$, where L_j is a dimension and X_j is a set of constants in that dimension.*
3. *$M \subseteq Meas$ is the measure set whose values are returned by q.*

This query model is the one used in the user interface for querying the OLAP cube, as depicted in Section 4.2.

3.3 Multidimensional Database and Schema

Because we need real logs with real analysis sessions, we need real data to analyze. The multidimensional database comes from real census data for social and economic research, called IPUMS [2].

[2] Integrated Public Use Microdata Series, Minnesota Population Center. http://www.ipums.org, 2008.

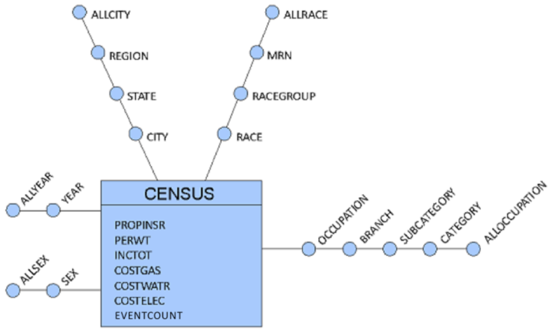

FIG. 1 – *Multidimensional Schema used for the test with the students*

The multidimensional schema is given Figure 1. The cube analyzed by the students is composed with 5 hierarchies and 25 measures. Here is a more detailed description of the different measures, aggregated either sum, maximum, minimum or average.

- PROPINSR: annual property insurance cost
- PERWT: person weight
- INCTOT: total personal income
- COSTGAS: annual gas cost
- COSTWATR: annual water cost
- COSTELEC: annual electricity cost
- EVENTCOUNT: number of facts (default measure, only aggregated by sum)

Example 3.1 *A possible query that a student could conduct over this cube would be: "I want the yearly evolution of the AVGINCTOT measure for Female", formally defined by:*
$q_1 = \langle\{Sex, Year, AllCity, AllRace, AllOccupation\}, \{Sex = "Female"\},$
$\{AVGINCTOT\}\rangle$

4 Test with graduate students

In this section we describe the questionnaires and the user interface used by the students for devising OLAP sessions, based on the OLAP cube defined in Section 3.3.

4.1 Design of the questionnaires

We describe here the design of the questionnaires which the students have to answer, for producing several analysis sessions. Note that all questionnaires are available in (Aligon et al., 2013).

The analysis sessions devised by the students will be used in future works, to especially assess the relevance of user-centric approaches. For this purpose, the log of sessions has to be as complete as possible in terms of diversity and complexity of analysis. Indeed, the diversity of analysis allows to cluster the analysis sessions according to the same expressed needs. Thus different context of analysis will be available to test the user-centric approaches. Regarding the different levels of complexities, we suppose that the log has sessions whose expressed needs are more or less detailed. Indeed, if a need is well expressed we expect that all sessions devised by different students will be close. On the other hand, if the need is less detailed we can suppose that the sessions will not look alike. Thus it can help, for instance, to test if a user-centric approach proposes recommendations or personalizations in the case where a current query is specific (i.e not frequent in the log) or common. By taking into account the requirements of diversity and complexity of analysis, our questionnaires are organized as follows.

Three type of needs have been identified over the cube depicted in Section 3.3:

- the $individual\ profile$ analysis. A profile is defined as the combination of sex and $race$ dimensions.
- the $occupation$ analysis.
- the $mixed$ analysis, i.e. an analysis is not specifically related to an $individual\ profile$ or $occupation.$

For each analysis described previously, two sub-types of analysis focused on particular measures are considered:

- the INCTOT measure (measuring the personal income)
- the energy measures (i.e COSTGAS, COSTWATR and COSTELEC)

Other analysis could be possible but the number of students doing the test is not large enough (see Section 5 for statistical details).

For each questionnaire, three levels of difficulties have been chosen:

- the basic needs. For this level, the needs are explicitly given. A possible question could be: *Is there a trend in the evolution of the average cost of gas for some profiles?*
- the intermediate needs. For this level, the needs are less explicit than the basic needs but not too complex. A possible question could be: *Compare the evolution of the minimum of energy costs, for the highest income, with the evolution of the maximum energy costs for the lowest incomes.*
- the advanced needs. For this level, the needs are deliberately fuzzy. A possible question could be: *Where is it better to live in terms of incomes, for an occupation?*

Consequently, 6 questionnaires have been designed for the tests with the students.

4.2 GUI for OLAP sessions

We now present the user interface querying an OLAP cube. As indicated in Section 1, a new user interface is required for especially taking into account the query model used in future user-centric approaches. We can note that few works have been focused on the combination between the HCI and structured data (see (Li and Jagadish, 2012) and (Nandi and Jagadish, 2011)). The GUI is depicted Figure 2.

Because the students are not so familiar with a particular language (like the MDX language) for designing queries, we chose to abstract this by implementing a user interface allowing to graphically design OLAP queries. This functionality can be seen in the part 1 of Figure 2. The interface is inspired by the Dimension Fact Model (DFM, developed in (Golfarelli and Rizzi, 2009)). It allows to design a query respecting the formal model defined in Section 3. A group-by set is created by linking a level between each dimension. A selection predicate can be added by selecting a level and the wished values (part 2 of Figure 2). The same principle can be done for adding the measures. When a query is designed, the user has to execute it (part

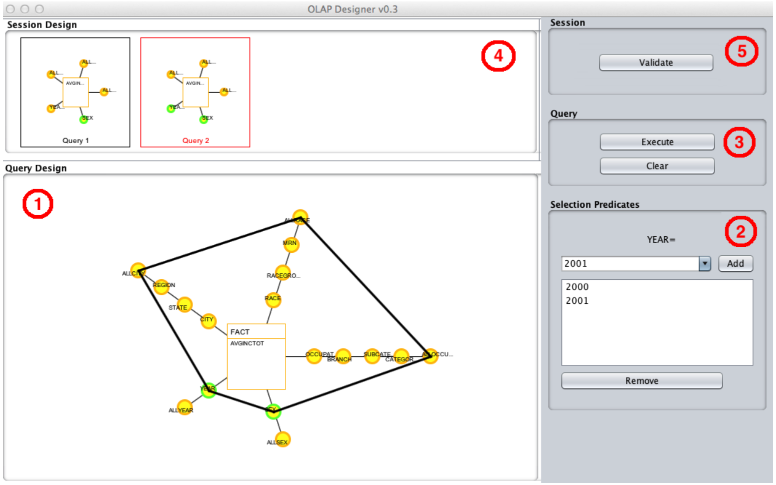

FIG. 2 – *User Interface for Designing OLAP Sessions*

3 of Figure 2). The query result is displayed to the user and the query is added to the current session (part 4 of Figure 2). Once the user considers he answered to his need , he validates his session (part 5 of Figure 2) which is automatically added in the log. Note that we assume three queries are needed to form a session. The GUI takes into account this constraint.

5 Statistics over the logs

In this section, we discuss the statistical results obtained from the tests conducted by the graduate students in Business Intelligence from the University of Francois Rabelais of Tours and the University of Bologna. 40 students participated to the tests (18 from France and 22 from Italy) by answering to one of six questionnaires given. We develop the statistics about the obtained sessions but also about the components of the queries (i.e. an element of the group-by set, measure set or selection set as defined in 3.3), shortly named $fragment$. Note that all the results and more can be consulted in (Aligon et al., 2013).

5.1 Statistical results over the sessions

The log is composed of 810 queries, distributed among 182 sessions (85 from France and 97 from Italy). Each questionnaire has been done 4 or 5 times. Figure 3 shows the number of sessions of the log for each complexity of question. We can note that the number of sessions for the advanced questions is half as large as the basic or intermediate questions. This is simply because the number of advances questions in the questionnaires is less important than the others.

Figure 4 reports the average number of queries for each complexity of question. For each level of difficulty, the average number seems low but constant between them. This result can seem strange when, intuitively, we could think that the more difficult the question is, the more important the number of queries should be. A first answer could be that the advanced questions were too difficult for the students or the questionnaires were too long (tiring the students).

Figure 5 indicates the average time for designing the sessions for each complexity of question. We can see that the time for designing the sessions related to the basic questions is the highest. This is due to the fact that the basic questions were the first conducted by the students. Consequently, a period of adaptation needs to be taken into account. The low period of time for devising the answers to the advanced questions can confirm the previous comment about Figure 4.

Figure 6 shows the number of sessions for each questionnaire. We can note that questionnaires 3 and 6 have less sessions than others. This is because more students did not answer all the questions of their questionnaires.

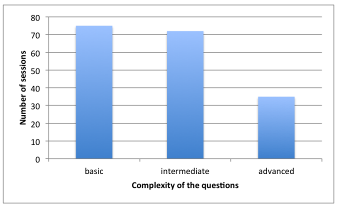

FIG. 3 – *Number of sessions per complexity of questions*

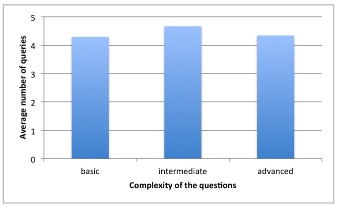

FIG. 4 – *Average number of queries per complexity of questions*

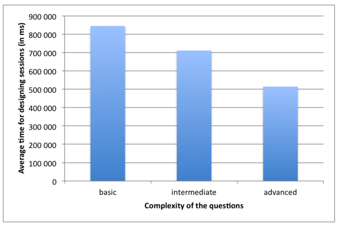

FIG. 5 – *Average time per complexity of questions*

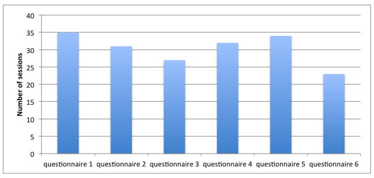

FIG. 6 – *Number of sessions per questionnaire*

5.2 Statistical results over the query fragments

We describe here more detailed statistics about query fragments (the elements included in the group-by set, measure set or selection set).

Figure 7 refers to the number of fragments for each complexity of question. We can notice that the number of fragments for the advanced questions is very low compared to the others. Related to the previous results of the Figures 4 and 5, we can conclude that the advanced questions were less well conducted than others.

Figure 8 shows the number of fragment type. We notice that there are twice as many projection fragments as measures. In the same way, there are twice as many measure fragments as selections. These results seem fully regular. Indeed each OLAP query must include a group-by set (in our case, composed with 5 levels) and at least one measure (if the user does not specify a measure, the default measure is used) whereas a selection set can be empty.

Figure 9 details the fragments related to the levels of the selection set. We can see that the frequency between each level of selection is very different. For instance, the $Year$, $Category$, $Branch$ and Sex levels are widely used but it reflects the type of asked questions.

Figure 10 indicates the number of fragments per questionnaires. We can note the same result than the explanation of Figure 6 for questionnaires 3 and 6, but we can also see that questionnaire 2 has few fragments whereas his number of sessions seems correct. It can mean that the sessions from this questionnaire are lower quality than others.

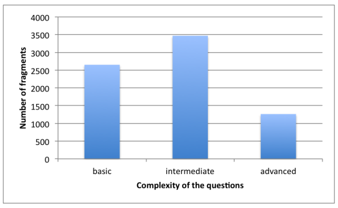

FIG. 7 – *Number of fragments per complexity of questions*

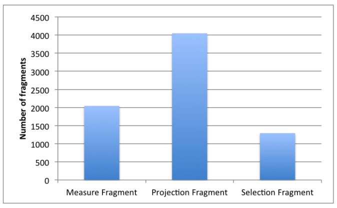

FIG. 8 – *Number of fragment type*

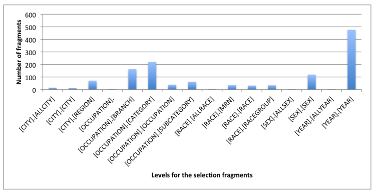

FIG. 9 – *Number of fragments per level of selections*

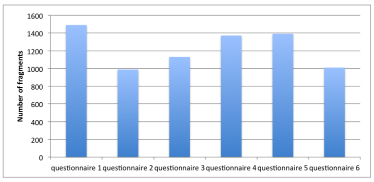

FIG. 10 – *Number of fragments per questionnaires*

5.3 Log filtering

According to the preliminary results of Sections 5.1 and 5.2 the conducted sessions can potentially not meet the need expressed in the questionnaires. Therefore, it is interesting to identify these types of sessions and put aside them if needed. We just present here a preliminary work about the log filtering:

- Identifying the identical successive queries for a same session. If we delete them (it seems relevant since these queries do not provide new information to the user), 25 sessions (about 14% of all sessions) have less than 3 queries.

- Identifying the queries having each member of a given level l in the selection set. If we delete this set, the queries are simplified without loosing information. 128 queries have been identified (about 15% of all queries).
- Identifying the sessions having high variations of OLAP operations between their successive queries. A solution is to compute the minimal atomic OLAP operations that transform a given query into its next query. The atomic operations considered can be: change level along one hierarchy in the group-by set, add or remove a clause from the selection predicate, change the constant appearing in a selection clause, and add or remove a measure.

6 Conclusion & Discussion

We have reported a feedback from real OLAP sessions devised by students from needs defined in different types of questionnaire. The statistical results show that the gathered sessions are workable if we pay attention to the quality of those coming from advanced questions. Regarding the log filtering, it seems easy to identify sessions having strange behaviors, like successive identical queries, or high variations in terms of OLAP operations. However, the definition of a relevant session remains difficult. A solution could be to identify a pattern of session for each question. For instance by analyzing and understanding the different sessions answering to a same question, we could identify a general schema of analysis. In any case, the definition of the relevance of an OLAP session is the key problem and has to be achieved, precisely in order to propose, in a long term perspective, a benchmark for OLAP sessions. With the same aim, specific metrics for measuring the quality of OLAP sessions from existing sessions in a corpus have to be developed. We will also continue to gather sessions with graduate students in order to freely propose a corpus to the OLAP community.

Acknowledgement

Thanks to the graduate students in Business Intelligence, from the University François Rabelais of Tours (France) and the University of Bologna (Italy), for their participation to the tests . Thanks to those who assisted in the successful achievement of the tests: Dr. Matteo Golfarelli, Dr. Patrick Marcel, Dr. Veronika Peralta, Pr. Stefano Rizzi and Dr. Elisa Turricchia.

References

Akbarnejad, J., G. Chatzopoulou, M. Eirinaki, S. Koshy, S. Mittal, D. On, N. Polyzotis, and J. S. V. Varman (2010). Sql querie recommendations. *PVLDB 3*(2), 1597–1600.

Aligon, J., M. Golfarelli, P. Marcel, V. Peralta, and S. Rizzi (2013). Real OLAP logs. http://www.julien.aligon.fr/index.php/real-olap-logs/.

Aligon, J., M. Golfarelli, P. Marcel, S. Rizzi, and E. Turricchia (2011). Mining Preferences from OLAP Query Logs for Proactive Personalization. In *Advances in Databases and Information Systems - 15th International Conference, ADBIS 2011, Vienna, Austria, September 20-23, 2011. Proceedings*, pp. 84–97.

Aligon, J., M. Golfarelli, P. Marcel, S. Rizzi, and E. Turricchia (2013a). Similarity measures for olap sessions. *KAIS 34*(3).

Aligon, J., P. Marcel, and E. Negre (2013b). Summarizing and querying logs of olap queries. In *Advances in Knowledge Discovery and Management*, pp. 99–124.

Chaudhuri, S., P. Ganesan, and V. R. Narasayya (2003). Primitives for Workload Summarization and Implications for SQL. In *VLDB*, pp. 730–741.

Gerhard Weikum (2013). Where's the data in the big data wave? ACM SIGMOD Blog, http://wp.sigmod.org.

Golfarelli, M. (2003). Handling Large Workloads by Profiling and Clustering. In *Data Warehousing and Knowledge Discovery, 5th International Conference, DaWaK 2003, Prague, Czech Republic, September 3-5,2003, Proceedings*, pp. 212–223.

Golfarelli, M. and S. Rizzi (2009). *Data Warehouse Design: Modern Principles and Methodologies* (1 ed.). New York, NY, USA: McGraw-Hill, Inc.

Khoussainova, N., Y. Kwon, W.-T. Liao, M. Balazinska, W. Gatterbauer, and D. Suciu (2011). Session-Based Browsing for More Effective Query Reuse. In *Scientific and Statistical Database Management - 23rd International Conference, SSDBM 2011, Portland, OR, USA, July 20-22, 2011. Proceedings*, pp. 583–585.

Li, F. and H. V. Jagadish (2012). Usability, databases, and hci. *IEEE Data Eng. Bull. 35*(3), 37–45.

Nandi, A. and H. V. Jagadish (2011). Guided interaction: Rethinking the query-result paradigm. *PVLDB 4*(12), 1466–1469.

Runeson, P. (2003). Using students as experiment subjects - an analysis on graduate and freshmen psp student data. In *Proceedings of 7th International Conference on Empirical Assessment & Evaluation in Software Engineering*, pp. 95–02.

Singh, V., J. Gray, A. Thakar, A. S. Szalay, J. Raddick, B. Boroski, S. Lebedeva, and B. Yanny (2007). Skyserver traffic report - the first five years. *CoRR abs/cs/0701173*.

Résumé

L'utilisation de sessions OLAP, effectuées par des analystes professionnels, semble être la meilleure manière pour vérifier la pertinence de solutions OLAP basées sur des requêtes passées (notamment avec les approches centrées utilisateurs, comme la recommandation ou la personalisation de requêtes). Cependant, pour les équipes de recherche universitaire, obtenir de tels logs est souvent difficile. De plus, la complexité des requêtes produites dans ces logs peut mener à un important traitement, dénaturant les analyses conduites. Dans ce papier, nous proposons un retour d'expérience à partir de vrais logs OLAP effectués par des étudiants de Master en Aide à la Décision. Ce retour d'expérience rapporte la mise en place de questionnaires et l'utilisation d'une interface utilisateur originale pour réaliser facilement de vraies sessions OLAP.