JIDM - Comments to Reviewers

Ref: Paper #176 (Real time data loading and OLAP queries: Living together in next generation BI environments)

Dear JIDM editors

Please find below our comments in response to the reviewers that gently conducted a careful revision on our article (#176 – Real time data loading and OLAP queries: Living together in next generation BI environments), submitted to the JIDM.

We sincerely consider that our modifications address all the reviewers suggestions, and would like to thank them for having contributed to improving the paper quality.

Best regards,
Fernanda Baião
(on behalf of the authors)

Reviewer A:

The paper has been extended from its original version.:
Yes.

The paper has been translated from its original version.:
Yes, but the translation needs improvement (which requires minor/major review).

Detailed comments (please number them C1, C2, ...). This may include suggestions for minor corrections and major reviews.:
The paper proposes a new approach for loading data in real time DW that is based on data fragmentation. The idea is very interesting and challenging. This paper is recommended for acceptance but the following improvements should be carried out by the authors.

(C1) Aside from [Inmon et al. 2008], [Kimball and Caserta 2004] and the related work given in section 3, there are several other real time DW architectures including approaches to avoid refreshment anomalies (see reference below). The authors should allude to them briefly and some details about these and how they might relate to their work would be helpful.
Thomas Jorg and Stefan Dessloch. Near Real-Time Data Warehousing Using State-of-the-Art ETL Tools. In:Third International Workshop on Enabling Real-Time for Business Intelligence (BIRTE 2009).

Response from authors: we analyzed this work and referred it in the paper, both in Section 2 and in the beginning of Section 4. In particular, the formalization of the anomalies that may occur when traditional ETL tools are used to address real time data is very interesting.

(C2) One of the main problems with the paper is the low quality of the English text (e.g.” THO and TJOA [2004] and DOKA et al.[2010] proposes…”; " or another data ranges.",….) .

Response from authors: the text was thoroughly reviewed with regard to the proper use of the English language.

(C3) The paper is difficult to follow because it contains a mixture of

basic concepts used in the paper and the authors´ experiments and their proposal (e.g. section 4 describes the work done by the authors together with fragmentation and distribution techgniques).

Response from authors: Section 4 is now focused only on describing the proposal of the paper.

(C4) In section 4, the paper would benefit from the use of a formalism to present the proposed distributed architecture (e.g. an UML components diagram).

Response from authors: We decided to keep the existing notation, since it is adopted by important authors in the area (e.g., the book "Principles of Distributed Systems from Ozsu and Valduriez) to illustrate the elements of a distributed architecture

(C5) A workload for performing evaluations over real time DW is given in Section 5.1 by extending the SSB queries. However, the query selectivity indicator is not given for these new queries.

Response from authors: we recognized the importance of this information, and added a comment at the end of Section 5.1.

(C6) Authors should improve the readability of their figures.

Reviewer B:

The paper has been extended from its original version.:
Yes, but not enough (which requires minor/major review).

The paper has been translated from its original version.:
No, the original was already in English.

Detailed comments (please number them C1, C2, ...). This may include suggestions for minor corrections and major reviews.:
OBJECTIVE OF THE PAPER: the paper proposes a physical architecture that applies data fragmentation techniques for loading operational data in real time in warehouses

---

COMMENTS (MINOR CORRECTIONS)

C1. SOME IMPROVEMENTS
- The authors should define the meaning of DW in the first time it appears.
- The authors should define the meaning of OLAP in the first time it appears.
- The authors should define the meaning of OLTP in the first time it appears.
- The authors should define the meaning of DDBM in the first time it appears.

C2. INTRODUCTION, PARAGRAPH 3: CHANGE FROM: "… improve decision-making processes in Organizations, especially …" TO " … improve decision-making processes in organizations, especially …"

C3. INTRODUCTION, PARAGRAPH 5: CHANGE FROM: "related works" TO "related work"

C4. SECTION 2, ITEM 2: The authors should improve the writing of the sentence: “Processing data considers data transformations tools are built
focusing on ETL traditional”

C5. SECTION 3, TITLE: CHANGE FROM “related works” TO “related work”

C6. SECTION 3, PARAGRAPH 1: The authors should improve the writing of paragraph 1 so that this paragraph should encompass more than 1 sentence.

C7. The authors should use expressions such as “proposals” or “approaches” instead of “works” in several places of the paper.

C8. SECTION 3: Some proposals are described using the “past tense”, while some proposals are described using the “present tense”. The authors should use the “past tense” or the “present tense”.

C9. What happens when node 1 (architecture of Fig.1) fails? The authors should improve the description of their architecture by explaining this situation.

Response from authors: This is a limitation of our current approach. We made it clear in the future work.

C10. SECTION 4, PARAGRAPH 5: CHANGE FROM: … this column point to the current date. TO: … this column pointS to the current date.

C11. CHANGE FROM: … an balanced distribution TO: a balanced distribution

C12. SECTION 4, PARAGRAPH 7: The authors should improve the writing of the
sentence: “Fig. 1 illustrates distributed architecture proposed.” Also,
the authors should always use node 1 (e.g., instead of the first node)

C13. SECTION 5, PARAGRAPH 2: CHANGE FROM: in an efficient way TO: efficiently, SO THAT: to efficiently process queries ….

C14. SECTION 5.1, PARAGRAPH 1:
- customer; supplier; part, and date. Please, use ; or , .
- CHANGE FROM: thirteen TO: 30
- INCLUDE: into four categories, NAMED Q1, Q2, Q3 AND Q4.
- The authors should join paragraphs 2 and 3. Also, include “, as follows” after “this feature”, so that: We extended SSB to include this feature, as follows. Original DBGEN ….
- CHANGE FROM: It was changed to include ‘01/01/1999’ day representing TO: We changed it
- INCLUDE: Besides, a new set of queries, NAMED Q5 (Fig. 3),
- INCLUDE: … from each of the four SSB query groups, GENERATING QUERIES
Q5.1 TO Q5.4,

C15. The authors should include a comma after “Also,”, “Thus,” throughout the paper.

Response from authors: All adjustments were performed as requested in C1 to C14.

COMMENTS (MAJOR CORRECTIONS)

C16. SECTION 3: The last sentence seems to be very confusing. The authors should explain why this paragraph is placed here and highlight the mining of this paragraph with regard to the proposed work. Also, the authors should precisely define what it means an "acceptable performance."

Response from authors: This sentence was moved to the end of the paragraph that describes the work of Inmon, and refined to make it more clear.

C17. The authors should improve Section 3 by specifying the differentials of their work with regard to each related work, i.e. the authors should include one or more sentences at the end of each paragraph to highlight the differentials of their work.

Response from authors: We pointed the major drawbacks of each related work, that are not present in our proposal.

C18. The authors should clarify the differences between this current article and Pereira et al. [2011]. Please, detail each improvement introduced in this paper. Also, this comparison should be placed in Section 1.

Response from authors: We listed the major improvements of the paper, as requested. In addition, the paper passed through an overall text revision.

C19. The definition of shared nothing distributed architecture (paragraph 6) should be placed in Section 2. Section 4 only should specify the relationship between the shared nothing distributed architecture and the proposed work.

Response from authors: We kept this definition in Section 4, since Section 2 is focused on real time dw concepts only. However, the definition was moved to a more adequate place, where we present the proposal architecture.

C20. Regarding the sentence "but this technique could also be used to create distinct date ranges to the historical fragments implementing data life cycle concept from Inmon et al. [2008].", the authors should explain this sentence better.

Response from authors: We clarified that a data range partition on the historical fragments may be considered an implementation of the data life cycle proposed by Inmon.

C21. SECTION 4, PARAGRAPH 9: The authors should improve the writing of paragraph 9. It is out of the scope of this section. Are the authors comparing their work with the work of Santos and Bernardino? If so, this paragraph should be placed in Section 3. Are the authors explaining some concepts related to their proposal? If so, only these concepts should remain in this section together with a better explanation.

Response from authors: we kept this paragraph, and clarified that it explains our approach for defining the right moment to redistribute real time data among the historical fragments, based on the several possible criteria definied by Santos and Bernardino.

C22. The authors detail several related work in Section 3, but compare their work with Furtado (2004) in the performance tests. Why? Why the work of Furtado is not described in Section 3? Also, why the proposals of Section 3 are not considered in this first test?

Response from authors: The work from Furtado does not handle real-time loading specifically (and therefore may not be considered a related work), but does apply distribution and parallel techniques, which surely impact on performance results.

C23. SECTION 5.3, PARAGRAPH 1: The authors should explain why in the proposed scenario the performance of the queries remains almost unchanged, i.e. they should detail the characteristics of the proposal that contribute to this conclusion. Also, the authors should explain why the execution of Query Q3.4 was more costly. Furthermore, the authors should include the percentage of difference between their proposal and the traditional proposal when arguing that "the

performance of traditional fragmentation was higher than the proposed fragmentation when loads are not concurrently executed with queries".

Response from authors: All requested information and explanations were added to the text.

C24. SECTION 5, LAST PARAGRAPH: According to the paper, "It's clear that query complexity and cost increases significantly when employing Santos and Bernardino approach". The authors should explain why the cost increases significantly when compared with their proposal. In fact, there are no performance tests in the paper that supports this conclusion.

Response from authors: for this claim, re relied on the cost of the query plans calculated by PostgreSQL (all statistics were up to date when those queries were explained). Those costs are shown in the query plan

C25. The authors should explain better that they are applying horizontal fragmentation. They should also explain the differences between vertical and horizontal fragmentation and detail why they applying horizontal fragmentation instead of vertical fragmentation or both. Also, the authors should improve the description of their architecture (Fig.1) to also include details about applying horizontal fragmentation.

Response from authors: we made it clear at some points along the text, and explained (Section 4, 4th paragraph) that horizontal fragmentation was preferred due to distribution transparency.

C26. Regarding performance tests, what happens when the number of nodes is not 5? The authors should also analyze the use of 2, 3, 4, 5, 6, etc. nodes.

Response from authors: Since at this moment our goal was not on evaluating the impact of adding/removing nodes in the distributed environment, we did not focus on this analysis. We do recognize it as an important future work, and mentioned it accordingly.

C27. The paper proposes a physical DW architecture that applies data fragmentation techniques on the fact table to provide real time data loading. What happens with the dimension tables? Are they fragmented? Are they replicated? This should be explained in the paper, perhaps in the proposed architecture.

Response from authors: we clarified that the dimension tables (except from the time one) are replicated on all nodes.

C28. The references are not uniform.

Response from authors: we carefully revised them along the text.