SIGMOD2015
SIGMOD2015
May 31- June 4, 2015, Melbourne, Australia

Reviews For Paper

Track	Research 2nd Submission
Paper ID	412
Title	Rethinking Data-Intensive Science Using Scalable Analytics Systems

Masked Reviewer ID: Assigned_Reviewer_1

Review:

Question	
Overall Rating	Revise and Resubmit
Summary of the paper (what is specifically being proposed, and in what context) and a brief justification for your overall recommendation; one paragraph	In this paper, the authors report case studies of scientific computing using open source large scale analytic systems including Avro, Parquent and Spark. The reference layering is proposed, and experimental results show the cost and performance of the proposed approaches in comparison with traditional MPI-based approach. This paper addresses a very important topic to extend the application of Apache Hadoop ecosystem to scientific computing. But I feel that this paper needs to be revised to highlight the technical contributions and novelty if we want to accept it as a research paper.
Three (or more!) strong points about the paper. Please be precise and explicit; clearly explain the value and nature of the contribution.	S1. The paper addresses an important topic that may lead to a lots of discussions and follow-up research works S2. The attempt to propose the layering approach and have schema as the "narrow waist" of the stack. S3. The performance and cost results in the evaluation section is impressive.
Three (or more!) weaknesses of the paper. Please indicate clearly whether the paper has any mistakes, missing related work, or results that cannot be considered a contribution. Write it so the authors understand what are seen as the negatives.	W1. The workload characteristic of scientific computation need further clarification. W2. The novelty in the two case studies need to be highlighted in section 4 and 5. W3. The experimental evaluations need to be extended for further analysis where time goes compared to MPI-based implementation.
Is the paper relevant for SIGMOD?	Yes

Significance	The paper will start a new line of research and/or products
Technical depth and quality of content	Syntactically complete but with limited contribution
Validation - experiments and proofs	OK, but do not cover all of the claims
Presentation	Reasonable: improvements needed
Discussion of related work - recall that the new page limits allow for material outside the 12pp including references, hence we expect good coverage of related work	Some description of related work, but could provide better perspective
Detailed Evaluation (contribution, pros/cons, errors); please number each point	D1 (W1). I would like to see more detailed analysis of the workload characteristic of scientific computing from different layers with a focus on evidence and materialized layer. For the evidence access layer, does Spark/MapReduce perfectly fit for the scientific computing workloads? I would like to see the analysis from different dimensions including data, computing logic and resource usage. Data: Take Spark for example, it fits for the scenario that the raw data of the workflow is very large on HDFS, but the working set of intermediate jobs (iterative or not) can fit into memory. What is the characteristics (e.g. selectivity of Map output, etc.) of scientific computing workloads? Job: Both MapReduce and Spark are based on the functional programming model, which means that we have to design algorithms following BSP. Compared to the MPI, BSP may lead to higher computational cost for some algorithms (e.g. pagerank, see the time complexity difference of Pagerank between GraphLab and GraphX), and it would be nice if the authors can add such analysis on scientific computing workloads. Resource usage: The jobs are CPU, memory or network bounded? Will the shuffle phase become the bottleneck of a job? For the materialized layer, the authors may highlight the requirements on optimizing I/O using data management techniques such as column store. Then, could the authors conclude: (a) which framework portfolio (e.g. Spark + Column store, MapReduce + Column store, MPI + Column store, etc) fits scientific computing workloads? (b) do we need to extend the framework (e.g. add more transformation operators API in Spark) to better support scientific computing workloads? These findings would be valuable for data management guys to extend existing systems like Spark to fit for scientific computing. D2. (W2) In section 4, is it possible to drive the presentation using example algorithms? I would like to see more details on designing and optimizing the algorithms from the two case studies, and understand which part of the optimizations is nontrivial. Based on current presentation, it seems that it is

	hard to find a single technique that has not been explored. D3. (W3) In section 6.1 and 6.2, the authors may add more analysis to further dissect why Spark is better than MPI in the two cases. In section 6.1, the authors do not explain why Spark is better than MPI. In section 6.2, the authors explain it at very high level in the third paragraph. But I would like to see numerical results to evaluate the impact on the execution time and resource usage with and without each of the optimizations. For example: (a) Spark with column store vs. Spark w/o column store to show both execution time and I/O saved on the column store. (b) Spark with caching vs. Spark w/o caching to show the impact of caching intermediate results, etc. Then we can better understand how each of the proposed techniques help for scientist computing workloads. D4. I am still curious about why Spark outperforms MPI. How about we use MPI+Avro+ Parquent? Hope that I can understand it after the authors address D1-D3. D5. Some typos. For example, in the abstract, it claims that "we demonstrate an example genomics pipeline that leverages open-source MapReduce and …", but it should be "Spark" if I understand correctly.
If revision is required, please list specific revisions you seek from the authors	Please address D1-D5.

Masked Reviewer ID: Assigned_Reviewer_2

Review:

Question	
Overall Rating	Reject
Summary of the paper (what is specifically being proposed, and in what context) and a brief justification for your overall recommendation; one paragraph	The paper presents a layered framework for supporting scientific applications. With a schema as the middle layer, data independence can be achieved. Two applications, genomic and astronomy image processing, are used to demonstrate the generality of such an approach. The backend is built using existing tools, and the experimental results show that the systems are more efficient than existing systems.
Three (or more!) strong points about the paper. Please be precise and explicit; clearly explain the value and nature of the contribution.	1. A reasonable layered design for building analytic systems. 2. Summarized the characteristics of scientific applications well. 3. Demonstrated the generality of the architecture through 2 case studies. 4. Paper is well structured, and easy to read.
Three (or more!) weaknesses of the paper. Please	

indicate clearly whether the paper has any mistakes, missing related work, or results that cannot be considered a contribution. Write it so the authors understand what are seen as the negatives.	1. Technical contribution is weak. 2. Performance gain comes essentially from the existing systems used. 3.
Is the paper relevant for SIGMOD?	Yes
Significance	The paper improves on existing work
Technical depth and quality of content	Syntactically complete but with limited contribution
Validation - experiments and proofs	Very nicely support the claims made in the paper
Presentation	Excellent: careful, logical, elegant, understandable
Discussion of related work - recall that the new page limits allow for material outside the 12pp including references, hence we expect good coverage of related work	Some description of related work, but could provide better perspective
Detailed Evaluation (contribution, pros/cons, errors); please number each point	1. This paper is a nice paper but may not be technically strong for sigmod. 2. There are also many works that have attempted to build mapreduce based database systems. These systems essentially guarantee data independence too. Not exactly clear what advantage the proposed approach has over these systems. 3. The performance gain is essentially derived from the features supported in existing systems, be it compression, selection, projection, etc. As long as one uses these features in the same way in the same system, one can derives the benefit.
If revision is required, please list specific revisions you seek from the authors	none.

Masked Reviewer ID: Assigned_Reviewer_4

Review:

Question	
Overall Rating	Reject
Summary of the paper (what is specifically being proposed, and in what context) and a brief justification for your overall recommendation; one paragraph	This paper describes a system that uses several database and system technologies, MapReduce and Columnar, for Astronomy and Genomics data processing pipelines, as an example of scientific analysis. I enjoyed reading the paper, however, there is not much real technical depth or material at the level that is suitable for SIGMOD. I strongly recommend the authors resubmit to the industrial track since many of the techniques adopted, lessons learned, and performance compared can be interesting even if the technical depth is not enough.
Three (or more!) strong points about the paper. Please be precise and explicit; clearly explain the value and nature of the contribution.	1. well written; 2. important topic; 3. nice case studies.
Three (or more!) weaknesses of the paper. Please indicate clearly whether the paper has any mistakes, missing related work, or results that cannot be considered a contribution. Write it so the authors understand what are seen as the negatives.	Technical depth is the weakness here. The paper is about adopting existing techniques and applying to two case studies. Quite suitable as a CIDR paper.
Is the paper relevant for SIGMOD?	Yes
Significance	The paper improves on existing work
Technical depth and quality of content	Insignificant contribution
Validation - experiments and proofs	Very nicely support the claims made in the paper
Presentation	Excellent: careful, logical, elegant, understandable

Discussion of related work - recall that the new page limits allow for material outside the 12pp including references, hence we expect good coverage of related work	Some description of related work, but could provide better perspective
Detailed Evaluation (contribution, pros/cons, errors); please number each point	REPLACE THIS WITH YOUR ANSWER
If revision is required, please list specific revisions you seek from the authors	REPLACE THIS WITH YOUR ANSWER