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

**Reviews For Paper**

| | |
|---|---|
| **Track** | Industrial |
| **Paper ID** | 412 |
| **Title** | Rethinking Data-Intensive Science Using Scalable Analytics Systems |

---

**Masked Reviewer ID:** Assigned_Reviewer_1

**Review:**

| Question | |
|---|---|
| Overall rating | Accept |
| Summary of review and rating rationale | This work presents a data analysis framework for eScience, with a focus on genomics and astronomy. The authors propose a layered model (i.e., like computer networks) for addressing problems in large-scale science. On one hand, what the paper promulgates doesn't sound very new: put data into a columnar format, use a scalable file system, and process it with your favorite parallel processing tool. On the other hand, separating on disk representation from data schema (for instance) is a near revolutionary idea in these scientific disciplines. The downside is that the sound and fury — excessive performance flag waving and sound engineering instead of novel research — detract from a subtle but important message. |
| 3 strong points | + Hard important problem: eScience does still use text files<br>+ High-level architecture will be impactful<br>+ Solid engineering lessons for a couple of application areas |
| 3 weak points | - Exacerbates performance gains, e.g., fixates on GATK when many large sequencing service centers use faster analytics<br>- Focus on performance gains from good engineering (pipeline through memory, algorithmic enhancements)<br>- Grab-bag of algorithmic advancements for particular field — could have focused more on "narrow waist." |
| | There is much to like and discuss in this paper. To outsiders of these particular fields this may feel obvious. But smart people didn't apply (or know about) them for a long time. And following this "stack" is non-trivial (as is knowing how to apply the end-to-end principle (which seems to be a missing part of the network stack analogy)). Finding bugs in existing analytics due to "stack smashing" is nice evidence.<br><br>On the other hand, I find that the authors occasionally protest too much. In the area of genomics, they continue to assert that alignment and calling are bottlenecks, taking 120 hours on a single box. While it's true GATK remains the standard bearer, some big whole-genome sequencing services (I.e., Illumina) don't use it, and instead use software that takes < 5-10 hours depending on machine and parameters. Similarly, many deliver 30-40x germline genomes, not 60x. And I'm pretty sure the authors know this, as they refer to other modern (and emerging) proposed aligners and callers that run 10x faster and produce decent results. In particular, SNAP (from Pandya's ASHG 2014 slides online) can align in 1.2 hours and sort, index, and mark duplicates in 2 hours, on a single 16 core machine with an SSD. So should we compare ADAM to something everyone knows is bust that takes 79 hours on an ill-configured box? Why the comparisons to HDDs on AWS? Especially now |

| | |
|---|---|
| Detailed comments (please refer to strong and weak points in your review) | that SSD is the default allocation? The times in Table 1 seem to be constructed to favor ADAM. Even with GATK, I've heard that re-alignment and BQSR can each be up to 2x shorter than the numbers reported here when using a large memory machine with SSDs.<br><br>So the authors should adjust the timings and costs of the prior approach. And it would also be helpful to put throughput (with a fixed budget) as a column in Table 1 as well. Even better, throughput vs machine count not just for ADAM but for running existing tools with separate genomes per box. ADAM still provides a benefit, right? It might also be instructive to comment on whether systems like Spark are also good for alignment/calling.<br><br>Along the same lines there is considerable discipline-centered detail and individual algorithmic enhancements. It is understood that getting the imperfect audience to pay attention results in content bias. So the paper is heavy into using existing techniques to solve eScience problems: columnar formats, in memory processing (for genomic pre-processing), range queries, and reading/writing from S3. It's ok that this isn't highly novel (but at least one of the many papers from the SpatialHadoop project should be cited), and it is well done.<br><br>But digging into that is missing the larger point. With regard to genomics, these seemingly simple changes could be fundamental. Genomics is plagued with crappy flat file formats. And with tools that are inconsistent. Everyone study is a one-off, ad-hoc process that courts disaster with incompatible tools, and poorly formatted files. These may not be the right layers (the 7 layer OSI model is often a simpler stack in practice), but it's going in the right direction. |