

Survey name (something about improving Apache Spark)

Mario Becerra Contreras

November 4, 2017

Abstract

We got a pretty nice abstract regarding the subject that is being studied and analyzed. This is a superb abstract, like, really really good. It's like the best abstract ever, for real. This is the best abstract you'll see in quite a long time.

1 Summary

[illegible][illegible]

2 Comments

[illegible]

la la la la la l. Lala ala la ala la la al al al laal alal la la la la la la la la l. Lala ala la ala la la al al al laal alal
la la la la la la la la la l. Lala ala la ala la la al al al laal alal la la la la la la la la la l.

References

- [1] Nagraj Alur, ed. *DB2 UDB's high function business intelligence in e-business*. 1st ed. IBM redbooks. San Jose, Calif: IBM Corp, 2002. ISBN: 978-0-7384-2460-6.
- [2] Michael Armbrust et al. "Spark SQL: Relational Data Processing in Spark". en. In: ACM Press, 2015, pp. 1383–1394. ISBN: 978-1-4503-2758-9. DOI: 10.1145/2723372.2742797. URL: <http://dl.acm.org/citation.cfm?doid=2723372.2742797> (visited on 10/13/2017).
- [3] Reza Bosagh Zadeh et al. "Matrix Computations and Optimization in Apache Spark". en. In: ACM Press, 2016, pp. 31–38. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939675. URL: <http://dl.acm.org/citation.cfm?doid=2939672.2939675> (visited on 10/13/2017).
- [4] Aaron Davidson and Andrew Or. "Optimizing shuffle performance in spark". In: *University of California, Berkeley-Department of Electrical Engineering and Computer Sciences, Tech. Rep* (2013).
- [5] Nusrat Sharmin Islam et al. "Performance characterization and acceleration of in-memory file systems for Hadoop and Spark applications on HPC clusters". In: *Big Data (Big Data), 2015 IEEE International Conference on*. IEEE, 2015, pp. 243–252. URL: <http://ieeexplore.ieee.org/abstract/document/7363761/> (visited on 10/13/2017).
- [6] Xiaoyi Lu et al. "Accelerating Spark with RDMA for Big Data Processing: Early Experiences". In: IEEE, Aug. 2014, pp. 9–16. ISBN: 978-1-4799-5860-3. DOI: 10.1109/HOTI.2014.15. URL: <http://ieeexplore.ieee.org/document/6925713/> (visited on 10/13/2017).
- [7] Xiaoyi Lu et al. "High-Performance Design of Hadoop RPC with RDMA over InfiniBand". In: IEEE, Oct. 2013, pp. 641–650. ISBN: 978-0-7695-5117-3. DOI: 10.1109/ICPP.2013.78. URL: <http://ieeexplore.ieee.org/document/6687402/> (visited on 10/13/2017).
- [8] Andre Luckow, Pradeep Mantha, and Shantenu Jha. "Pilot-Abstraction: A Valid Abstraction for Data-Intensive Applications on HPC, Hadoop and Cloud Infrastructures?" In: *arXiv preprint arXiv:1501.05041* (2015). URL: <https://arxiv.org/abs/1501.05041> (visited on 10/13/2017).
- [9] Andre Luckow et al. "Hadoop on HPC: integrating Hadoop and pilot-based dynamic resource management". In: *Parallel and Distributed Processing Symposium Workshops, 2016 IEEE International*. IEEE, 2016, pp. 1607–1616. URL: <http://ieeexplore.ieee.org/abstract/document/7530058/> (visited on 10/13/2017).
- [10] Peilong Li et al. "HeteroSpark: A heterogeneous CPU/GPU Spark platform for machine learning algorithms". In: IEEE, Aug. 2015, pp. 347–348. ISBN: 978-1-4673-7891-8. DOI: 10.1109/NAS.2015.7255222. URL: <http://ieeexplore.ieee.org/document/7255222/> (visited on 10/13/2017).
- [11] Konstantin Shvachko et al. "The hadoop distributed file system". In: *Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on*. IEEE, 2010, pp. 1–10. URL: <http://ieeexplore.ieee.org/abstract/document/5496972/> (visited on 10/13/2017).
- [12] Li Wang et al. "Understanding the Behavior of Spark Workloads from Linux Kernel Parameters Perspective". en. In: ACM Press, 2016, pp. 1–2. ISBN: 978-1-4503-4666-5. DOI: 10.1145/3007592.3007593. URL: <http://dl.acm.org/citation.cfm?doid=3007592.3007593> (visited on 10/13/2017).
- [13] Matei Zaharia et al. "Apache Spark: a unified engine for big data processing". en. In: *Communications of the ACM* 59.11 (Oct. 2016), pp. 56–65. ISSN: 00010782. DOI: 10.1145/2934664. URL: <http://dl.acm.org/citation.cfm?doid=3013530.2934664> (visited on 10/13/2017).
- [14] Matei Zaharia et al. "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing". In: *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. USENIX Association, 2012, pp. 2–2. URL: <http://dl.acm.org/citation.cfm?id=2228301> (visited on 10/13/2017).

- [15] Matei Zaharia et al. “Spark: Cluster computing with working sets.” In: *HotCloud* 10.10-10 (2010), p. 95. URL: http://static.usenix.org/legacy/events/hotcloud10/tech/full_papers/Zaharia.pdf (visited on 10/13/2017).