and stored them in file and the database separately. We performed 10,000 times of query and plot the total response time. To be precise, in each test, the results stored in file are all read into memory and the query is performed in memory. On the other hand, the in-memory database is started from the first test and keeps running as a distributed server. We can observe that using in-memory database, the query response is constant and is only proportional to the number of query items. While the traditional 'offline file search' keeps increasing the response time as more results accumulated. Our evaluation shows that the distributed in-memory database is promising to help reducing the data movement for big data analytic problems.

VI. Related Work

Compared to caching raw data, caching result is a relatively new research area. In cloud computing, caching results is a technique to achieve fault tolerance (RDD [20])cite linkedin and incremental computing [7], [4]. The idea of RDD is to keep the partitioned operation and recompute the data using lineage for fast fault tolerance. In contrast, Our in-advance system is designed to reuse the analysis results by detecting the computation and I/O overlapping. Knowledge discovery is another area that is related to our work. Knowledge discovery is 'the nontrivial extraction of implicit, previously unknown, and potentially useful information from data' [8]. It focuses on using various machine learning or statistical methods to explore the data for unknown knowledge. Our work is similar in the way of predicting the useful results, but the difference is that we design a lightweight system that observes the user's analysis habit and tries to make a recommendation for scientists.

VII. conclusion

In this study, we have introduced a new *in-advance data analytics* method for reducing data movement for big data analysis and big data applications. The proposed in-advance data analytics leverages a prediction method that uses minimal computing resources to generate useful analysis results in advance. As data movement dominates the run time of big data analysis, and computing is virtually free for big data problem, the in-advance data analytics can be a promising solution that fully leverages data locality and reduces the data movement and the time to solution.

VIII. Acknowledgments

This research is sponsored in part by the National Science Foundation under grant CNS-1338078 and CNS-1162488. We also acknowledge the High Performance Computing Center (HPCC) at Texas Tech University for providing resources.

References

[1] Climate data operator. https://code.zmaw.de/projects/cdo.

[2] General circulation model. http://en.wikipedia.org/wiki/Community_Climate_System_Model.

[3] H. Abbasi, G. Eisenhauer, M. Wolf, K. Schwan, and S. Klasky. Just in time: adding value to the io pipelines of high performance applications with jitstaging. In *HPDC*, pages 27–36, 2011.

[4] P. Bhatotia, A. Wieder, R. Rodrigues, U. A. Acar, and R. Pasquin. Incoop: Mapreduce for incremental computations. In *Proceedings of the 2Nd ACM Symposium on Cloud Computing*, SOCC '11, pages 7:1–7:14, New York, NY, USA, 2011. ACM.

[5] K. M. Curewitz, P. Krishnan, and J. S. Vitter. Practical prefetching via data compression. In P. Buneman and S. Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 257–266, Washington, D.C., 26–28 May 1993.

[6] T. L. Delworth. Gfdl's cm2 global coupled climate models. *J. Climate*, 19:643674, 2006.

[7] J. Ekanayake, H. Li, B. Zhang, T. Gunarathne, S. hee Bae, J. Qiu, and G. Fox. Twister: A runtime for iterative mapreduce. In *In The First International Workshop on MapReduce and its Applications*, 2010.

[8] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus. Knowledge discovery in databases: An overview. *AI Mag.*, 13(3):57–70, Sept. 1992.

[9] M. Gardner, W. chun Feng, J. Archuleta, H. Lin, and X. Ma. Parallel genomic sequence-searching on an ad-hoc grid: Experiences, lessons learned, and implications. In *SC'2006 Conference*, Tampa, FL, Nov. 2006.

[10] J. Gray, D. T. Liu, M. A. Nieto-Santisteban, A. S. Szalay, G. Heber, and D. DeWitt. Scientific data management in the coming decade. Technical Report MSR-TR-2005-10, Microsoft Research (MSR), Jan. 2005.

[11] F. Kaspar, U. Schulzweida, and R. Muller. Climate data operators as a user-friendly processing tool for cmsaf's satellite-derived climate monitoring products. In *the Proc. of the EUMETSAT Meteorological Satellite Conference*, 2010.

[12] S. Lakshminarasimhan, J. Jenkins, I. Arkatkar, Z. Gong, H. Kolla, S.-H. Ku, S. Ethier, J. Chen, C.-S. Chang, S. Klasky, R. Latham, R. B. Ross, and N. F. Samatova. ISABELA-QA: query-driven analytics with ISABELA-compressed extreme-scale scientific data. In *SC 2011, Seattle, WA, USA, November 12-18, 2011*, page 31, 2011.

[13] J. Liu, S. Byna, and Y. Chen. Segmented analysis for reducing data movement. In *the Proc. of the IEEE International Conference on Big Data,(Bigdata'13)*, 2013.

[14] X. Ma and A. L. N. Reddy. MVSS: An active storage architecture. *IEEE Transactions on Parallel and Distributed Systems (TPDS)*, PDS-14(10):993–1005, Oct. 2003.

[15] A. Parameswaran, N. Polyzotis, and H. Garcia-Molina. Seedb: Visualizing database queries efficiently. Technical report, Stanford University, 2013.

[16] J. Piernas, J. Nieplocha, and E. J. Felix. Evaluation of active storage strategies for the lustre parallel file system. In *SC'07*. ACM/IEEE, Reno, NV, Nov. 2007.

[17] D. Wang, C. Zender, and S. Jenks. Efficient clustered server-side data analysis workflows using swamp. *Earth Science Informatics*, 2(3):141–155, 2009.

[18] R. Wickremesinghe, J. S. Chase, and J. S. Vitter. Distributed computing with load-managed active, storage. In *Proc. 11th IEEE International Symposium on High Performance Distributed Computing (11th HPDC'02)*, pages 13–23, July 2002.

[19] H. Yu, C. Wang, R. W. Grout, J. H. Chen, and K.-L. Ma. In situ visualization for large-scale combustion simulations. *IEEE Computer Graphics and Applications*, 30(3):45–57, May/June 2010.

[20] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. Franklin, S. Shenker, and I. Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. Technical Report UCB/EECS-2011-82, EECS Department, University of California, Berkeley, Jul 2011.