RC25281 (AUS1204-004) April 23, 2012
Computer Science

IBM Research Report

Understanding System and Architecture for Big Data

Anne E. Gattiker, Fadi H. Gebara, Ahmed Gheith, H. Peter Hofstee, Damir A. Jamsek, Jian Li, Evan Speight
IBM Research Division
Austin Research Laboratory
11501 Burnet Road
Austin, TX 78758
USA

Ju Wei Shi, Guan Cheng Chen
IBM Research Division
China Research Laboratory
Building 19, Zhouguancun Software Park
8 Dongbeiwang West Road, Haidian District
Beijing, 100193
P.R.China

Peter W. Wong
IBM Linux Technology Center
11501 Burnet Road
Austin TX 78758
USA

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Understanding System and Architecture for Big Data

Anne E. Gattiker, Fadi H. Gebara, Ahmed Gheith, H. Peter Hofstee, Damir A. Jamsek, Jian Li, Evan Speight
IBM Research - Austin
{gattiker,fhgebara,ahmedg,hofstee, jamsek,jianli,speight}@us.ibm.com

Ju Wei Shi, Guan Cheng Chen
IBM Research - China
{jwshi,chengc}@cn.ibm.com

Peter W. Wong
IBM Linux Technology Center
wpeter@us.ibm.com

ABSTRACT

The use of Big Data underpins critical activities in all sectors of our society. Achieving the full transformative potential of Big Data in this increasingly digital world requires both new data analysis algorithms and a new class of systems to handle the dramatic data growth, the demand to integrate structured and unstructured data analytics, and the increasing computing needs of massive-scale analytics. In this paper, we discuss several Big Data research activities at IBM Research: (1) Big Data benchmarking and methodology; (2) workload optimized systems for Big Data; (3) case study of Big Data workloads on IBM Power systems. In (3), we show that preliminary infrastructure tuning results in sorting 1TB data in 14 minutes[1] on 10 Power 730 machines running IBM InfoSphere BigInsights. Further improvement is expected, among other factors, on the new IBM PowerLinux™ 7R2 systems.

1. INTRODUCTION

The term "Big Data" refers to the continuing massive expansion in the data *volume* and *variety* as well as the *velocity* and *veracity* of data processing [9]. Volume refers to the scale of the data and their processing need. Whether for data at rest or in motion, i.e., being a repository of information or a stream, the desire for high speed is ever-present. Variety indicates that Big Data comes in many forms, and techniques are needed to understand and process such variation. Veracity is the truth in the information. Yet data uncertainty can come from many different places, from the data itself, sensor precision, model approximation to inherent process uncertainty. Emerging social media also brings in new types of data uncertainty such as rumors, lies, falsehoods and wishful thinking. We believe the *4 Vs* capture the main characteristics of Big Data nowadays.

Undoubtedly, the use of Big Data underpins critical activities in all sectors of our society. Achieving the full transformative potential of Big Data in this increasingly digital world requires both new data analysis algorithms and a new class of systems to handle the dramatic data growth, the demand to integrate structured and unstructured data analytics, and the increasing computing needs of massive-scale analytics. As a result, research ideas are spawned in all critical aspects of emerging analytics systems for Big Data. A short list of these include but are not limited to: processor, memory, and system architectures for data analysis; benchmarks, metrics, and workload characterization for analytics and data-intensive computing; debugging and performance analysis tools for analytics and data-intensive computing; accelerators for analytics and data-intensive computing; implications of data analytics to mobile and embedded systems; energy efficiency and energy-efficient designs for analytics; availability, fault tolerance and recovery issues; scalable system and network designs for high concurrency or high bandwidth data streaming; data management and analytics for vast amounts of unstructured data; evaluation tools, methodologies and workload synthesis; OS, distributed systems and system management support; MapReduce and other processing paradigms and algorithms for analytics.

This short paper briefly discuss three topics that a few IBM researchers and colleagues from other IBM divisions have been working on:

- Big Data benchmarking and methodology
- Workload optimized systems for Big Data
- Case study of Big Data workloads on IBM Power systems

We highlight the research directions that we are pursuing in this paper. More technical details and progress updates will be covered in several papers in an upcoming issue of the IBM Journal of Research and Development.

[1] All performance data contained in this publication was obtained in the specific operating environment and under the conditions described below and is presented as an illustration. Performance obtained in other operating environments may vary and customers should conduct their own testing.

Contact Author: Jian Li; Email: jianli@us.ibm.com
IBM Research Technical Report April 24th, 2012

2. BENCHMARKING METHODOLOGY

Massive Scale Analytics is representative of a new class of workloads that justifies a re-thinking of how computing systems should be optimized. We start by tackling the problem of the absence of a set of benchmarks that system hardware designers can use to measure the quality of their designs and that customers can use to evaluate competing hardware offerings in this fast-growing and still rapidly-changing market. Existing benchmarks, such as HiBench [10], fall short in terms of both scale and relevance. We conceive a methodology for peta-scale data-size benchmark creation that includes representative Big Data workloads and can be used as a driver of total system performance, with demands balanced across storage, network and computation. Creating such a benchmark requires meeting unique challenges associated with the data size and its unstructured nature. To be useful, the benchmark also needs to be generic enough to be accepted by the community at large. We also observe unique challenges associated with massive scale analytics benchmark creation, along with a specific benchmark we have created to meet them.

A first consequence of the massive scale of the data, is that the benchmark must be descriptive, rather than prescriptive. In other words, our proposed benchmark is provided as instructions for acquiring the required data and processing it, rather than providing benchmark code to run on supplied data. We propose the use of existing large datasets, such as the 25TB ClueWeb09 dataset [7] and the over 200TB Stanford WebBase repository [8]. Interestingly, we also observe the challenges of using such real-world large datasets, which include physical data delivery, e.g., via shipped disk drives, and data formatting/"cleaning" of the data to allow robust processing.

We propose compute- and data-intensive processing tasks that are representative of key massive-scale analytics workloads to be applied to this unstructured data. These tasks include major Big Data application areas, text analytics, graph analytics and machine-learning. Specifically our benchmark efforts focused on document categorization based on dictionary-matching, document and page ranking, and topic determination via non-negative matrix factorization. The first of the three, in particular, required innovation in benchmark creation, as there is no "golden reference" to establish correct document categorization. Existing datasets typically used as references for text-categorization assessments, such as the Enron corpus [2], are orders of magnitude smaller than what we required. Our approach for overcoming this challenge includes utilizing publicly-accessible documents coded by subject, such as US Patent Office patents and applications, to create subject-specific dictionaries against which to match documents. Unique challenges of ensuring "real-world" relevance includes covering non-word terms of importance, such as band names that include regular expression characters, and a "wisdom of crowds" approach that helps us meet those challenges.

We plan to make our benchmark public.

3. WORKLOAD OPTIMIZED SYSTEMS

While industry has made substantial investments in extending its software capabilities in Analytics and Big Data, thus far these new workloads are being executed on systems that were designed and configured in response to more traditional workload demands.

In this paper we present two key improvements to traditional system design. The first is the addition of reconfigurable acceleration. While reconfigurable acceleration has been used successfully in commercial systems and appliances before (e.g. DataPower [5] and Netezza [4]), we have prototyped and demonstrated that such technology can benefit the processing of unstructured data.

The second innovation we discuss is a new modular and dense design that also leverages acceleration. Achieving the computational and storage densities that this design provides requires an increase in processing efficiency that is achieved by a combination of power-efficient processing cores and offloading of performance-intensive functions to the reconfigurable logic.

With an eye towards how analytics workloads are likely to evolve, and towards executing such workloads efficiently, we conceive the system that leverages the accelerated dense scale-out design in combination with powerful global server nodes that orchestrate and manage the computation and that can also be used to perform deeper in-memory analytics on selected data.

4. BIG DATA ON POWER SYSTEMS

To date, Apache Hadoop [1] has been widely deployed on clusters of relatively large numbers of moderately sized, commodity servers. However, it has not been widely used on large, multi-core, heavily threaded machines even though smaller systems have increasingly large core and hardware thread counts. We describe an initial performance evaluation and tuning of Hadoop on a large multi-core cluster with only a few machines.

The evaluation environment comprises IBM InfoSphere BigInsights [3] on a 10-machine cluster of IBM Power 730 Express servers[2]. IBM InfoSphere BigInsights brings the power of Apache Hadoop to the enterprise. BigInsights enhances the Hadoop technology to withstand the demands of the enterprise, adding administrative, workflow, provisioning, and security features, along with best-in-class analytical capabilities from IBM Research. This version of BigInsights, version 1.3, uses Hadoop 1.0 and its built-in HDFS file system.

Each Power 730 machine includes 12 P7 cores @ 3.7GHz, up to 48 hardware threads, 64 GB of memory, and $24 \times 600GB$ SAS drives in an external Direct Attached Storage (DAS) drawer. The machines are connected by 10Gb Ethernet network. We used SUSE Linux (SLES11 SP1).

We have some early experiences with measuring and tuning standard Hadoop programs, including some of the ones used in the HiBench [10] benchmark, and some from real-world customer engagements. In this paper, we use Terasort, a widely used sort program in Hadoop distributions, as a case study. While Terasort in Hadoop is flexible to be configured to sort a variety of amount of data, we sort 1TB input data by literally following the term "Tera Sort". In addition we compress neither input nor output data.

Our initial trial is done with the default Hadoop map-reduce configuration, e.g. limited map and reduce task slots, which does not utilize the 12 processor cores in a single P730

[2]The new PowerLinux™ 7R2 systems have been measured to perform at least as well as similarly configured IBM Power 730 systems.

system. As expected, the test takes hours to finish. After initial adjustment of the number of mappers and reducers to fit to the parallel processing power of 12 cores in a Power 730 system, the execution time drastically decreases to 47 minutes. Of course, this is just the beginning of the exercise.

We then apply the following public-domain tuning methods to gradually improve the sort performance:

- Publicly available LZO compression [6] for intermediate data compression[3];
- Aggressive read ahead setting, deadline disk IO scheduling and large block size to improve storage performance;
- Four-way Simultaneous Multithreading (SMT4) to further increase computation parallelism;
- Preliminary intermediate control of map and reduce stages to better utilize available memory capacity;
- Reconfiguration of storage subsystem to remove failover support of storage adapters for effective bandwidth improvement;
- JVM, Jitting and GC tuning that fit to Power architecture.

As of this writing, we were able to achieve less than 14 minutes execution time of sorting 1TB input data from disk and writing back 1TB output data to disk storage. As more comprehensive tuning are underway, we expect that further improvement is possible.

Note that we have only applied infrastructure tuning in this stage of the study. In the next stage, we plan to incorporate the performance enhancement features in IBM InfoSphere BigInsights for further improvement. We are also working on patches for better intermediate control in MapReduce. Furthermore, we plan to apply reconfigurable acceleration technology as indicated in Section 3. In the meantime, scalability study is also important to understand the optimal approach to (A) *strong-scaling* of the BigInsights cluster with constant input and (B) *weak-scaling* of the BigInsights cluster that scales up with input. While we have not observed that the network is a bottleneck in our 10-node cluster, this may change when the system scales up and the workload changes.

5. CONCLUSIONS

In this paper, we have presented our initial study on Big Data benchmarking and methodology as well as workload optimized systems for Big Data. We have also discussed our initial experience of sorting 1TB data on a 10-node Power 730 cluster. As of this writing, it takes less than 14 minutes to complete the sort, and we expect additional improvements in the near future.

6. ACKNOWLEDGMENTS

We would like to thank the IBM InfoSphere BigInsights team for their support. Particularly, we owe a debt of gratitude to Berni Schiefer, Stewart Tate, Hui Liao and Mei-Mei Fu for their guidance and insights. We also would like to thank Steve Pratt for his expertise in file system and IO performance tuning and Bill Buros for his guidance in various Power Linux performance matters. Susan Proietti Conti, Gina King, Angela S Perez, Raguraman Tumaati-Krishnan, Demetrice Browder and Chuck Bryan have been supporting us to streamline the process all along the way. Lastly but not least importantly, we cannot reach this stage without the generous support and advice from, Dan P. Dumarot, Richard W. Bishop, Pat Buckland, Clark Anderson, Ken Blake, Geoffrey Cagle and John Williams, to name a few.

[3]Among others, we expect IBM CMX compression library to have better performance. Note that LZO is not part of IBM InfoSphere BigInsights.

7. REFERENCES

[1] Apache hadoop. http://hadoop.apache.hadoop.
[2] Enron Email Dataset. http://www.cs.cmu.edu/ enron/.
[3] IBM InfoSphere BigInsights. http://www-01.ibm.com/software/data/infosphere/biginsights/.
[4] IBM Netezza Data Warehouse Appliances. http://www-01.ibm.com/software/data/netezza/.
[5] IBM WebSphere DataPower SOA Appliances. http://www-01.ibm.com/software/integration/datapower/.
[6] LZO: A Portable Lossless Data Compression Library. http://www.oberhumer.com/opensource/lzo/.
[7] The ClueWeb09 Dataset. http://lemurproject.org/clueweb09/.
[8] The Stanford WebBase Project. http://diglib.stanford.edu:8091/ testbed/doc2/WebBase/.
[9] IBM Global Technology Outlook. 2012.
[10] S. Huang, J. Huang, J. Dai, T. Xie, and B. Huang. The HiBench benchmark suite: Characterization of the MapReduce-based data analysis. In *IEEE International Conference on Data Engineering Workshops (ICDEW)*, Long Beach, CA, USA, 2010.