High Performance Computing and High Performance Data Analytics—What is the Missing Link?

Bastian Koller, Michael Gienger and Michael M. Resch

Abstract Within this book chapter, technologies for data mining, data processing and data interpreting are introduced, evaluated and compared. Especially, traditional High Performance Computing, and the newly emerging fields High Performance Data Analytics and Cognitive Computing are put into context in order to understand their strengths and weaknesses. However, the technologies have not been evaluated solely, but also the missing links between them have been identified and described.

1 Introduction

At this point of time, there are various technologies in the market that target data analysis, data processing, data interpreting and data mining. So far, it has not been clear if all of those technologies are direct competitors or can be seen in a complementary fashion. This book chapter therefore analyses the technologies carefully and introduces as well as compares their direct angles. Being more concrete, traditional High Performance Computing, the newly emerging field High Performance Data Analytics as well as Cognitive Computing are evaluated. In particular, the interactions between those technological fields are visualized in addition.

The book chapter is organized as follows: Section 2 is providing the High Performance Computing context, Sect. 3 is introducing High Performance Data Analytics whereas Sect. 4 compares the approaches and describes the missing links. Finally, Sect. 5 concludes this book chapter.

B. Koller (✉) · M. Gienger · M. Resch
High Performance Computing Center Stuttgart, Nobelstrasse 19,
70569 Stuttgart, Germany
e-mail: koller@hlrs.de

M. Gienger
e-mail: gienger@hlrs.de

M. Resch
e-mail: resch@hlrs.de

M.M. Resch et al. (eds.), *Sustained Simulation Performance 2016*,
DOI 10.1007/978-3-319-46735-1_2

2 The Evolution of High Performance Computing

Within this section of the book chapter, a generic view on High Performance Computing (HPC) and its evolution over time is given. Although the purpose of such HPC systems is in principle the same, the available performance, the customer base as well as the computational and applications models changed in the last decade. In summary, various application areas such as computational fluid dynamics, climate or physics simulations are considered HPC relevant at the moment, which are executed on innovative systems that may be equipped by vector central processing units, by commonly used x86 processors or even accelerators.

2.1 Traditional High Performance Computing

High Performance Computing has been traditionally designed to solve problems that are too large and complex for common desktop computers or even workstations. Those systems enable a maximum of performance for memory, compute, storage or input/output (I/O) intensive applications and operations. However, with respect to their special design and the corresponding drastic costs, they clearly lack on the flexibility to combine all requirements into a unique general-purpose system.

Although there are self-appointed general-purpose systems in the worldwide HPC market, there is always a key application that drives the selection of such systems. Applications that require solely a high computational demand will result in a system architecture that is based on accelerators, whereas applications that require thousands of memory operations per second will rather tend to the vector or x86 architecture. Thus, due to the main area of applications and the corresponding costs, a HPC system is always tailored to its common applications so that "real" general-purpose systems cannot be seen in the markets.

2.2 Evolution Over Time

Within the last decade, there was a huge evolution with regards to the HPC systems. Reaching from vector machines to the widely adopted x86 architecture and modern accelerators, especially hardware evolved quickly. In the meantime, HPC systems with more than 1.000.000 cores are not an utopia any more[1] so that besides the efficiency of the systems, also the models and applications can benefit from the huge amount of provided computational performance.

But not just the hardware evolved, also the customer basis changes: industrial applications from the automotive world, academic applications dealing with, for instance, climate simulation as well as applications from small and medium sized

[1] Top500: http://www.top500.org

enterprises from various kinds of areas are targeting the High Performance Computing systems. However, with the evolved systems and the immense performance, also the execution models get more complicated. On the one hand, there are still traditional applications that require a huge amount of resources for a single run and on the other, parametric studies with less constant performance requirements but generating a huge amount of results are common in state-of-the-art HPC systems.

Nevertheless, HPC driving applications are still usual in the High Performance Computing area, but due to the changing application and executions models, general-purpose systems are becoming more evident as large computational intensive applications typically produce a huge amount of results. So there is currently a trade-off between providing generic systems that are flexible enough to cope with different kinds of workloads and such systems that are solely made to provide one single key performance type.

3 Towards High Performance Data Analytics

In contrast to Sect. 2, this chapter focuses High Performance Data Analytics (HPDA), a new emerging field for the High Performance Computing sector. High Performance Data Analytics target the efficient analytics of various kinds of data, reaching from structured up to unstructured as well as streaming data, which cannot be analysed anymore on standard workstations or Clouds due to their volume, their variety or their velocity.

3.1 Where Is It Needed?

As already highlighted in the introduction of this section, High Performance Data Analytics target the analysis of available (e.g. stored) or real-time streaming data. In contrast to HPC applications, HPDA requires typically not an extraordinary huge amount of compute performance, but rather a very broad I/O backend that is able to transfer data quickly enough to the actual processing engines.

The applications that typically cause such data intensive workloads are settled in the sensor technologies area, such as the evolving Internet of Things, the aligned Industry 4.0 and the cyber physical systems area. The physical sensors produce a huge amount of data that has to be analysed in time, sometimes even real-time to provide the corresponding actions. However, not just those industrial areas require the implicitly described system architecture, but also modern Internet stores with their designed customer marketing require as much as knowledge possible about their customers. This fact results in strong correlations of data that have to be analysed on huge-scale systems, since Clouds are not performing enough. Finally, not only the described applications require HPDA functionality, fine-grained models and their

corresponding applications produce Terabytes of data in the meanwhile that cannot be analysed on a state-of-the-art HPC system anymore.

3.2 HPDA Concepts and Technologies

As already highlighted in the sections before, HPC and HPDA approaches in terms of hardware and software require different technologies. Therefore, these requirements will be discussed and addressed in particular in this sub-section to bridge the gap between both technologies.

In terms of hardware, data intensive workloads require different key performance indicators than standard HPC applications. The differences between both approaches are highlighted below:

- **Processors**
  In traditional HPC systems, fast processors with fast memory pipelines are focused. For HPDA systems, the amount of Floating Point Operations Per Second is still important, however the performance of the system is determined by the storage system.
- **Memory**
  The more memory available for data analytics, the better for the overall application execution since most of the data and results can be kept in memory instead of check pointing them to the storage backend. For HPC systems, the same statement holds, although much smaller memory systems are targeted than in the HPDA area.
- **Networks**
  Whenever data needs to be transferred, fast interconnects come into play. So both, HPC and HPDA systems require fast memory and latency-oriented networks in order to transfer the data efficiently.
- **Storage**
  Typical HPC systems provide a central system storage from which all the required data gets read and written. An approach like this is not possible for HPDA since the data accessibility is the key performance indicator for the whole applications. Therefore, data analytics systems provide fast local disks that can be used to provide and cache the data in order to optimize the application execution.

As can be seen, the main differences between HPC and HPDA systems are located in the area of processors and storages, since fast number-crunching processors are required for HPC only. In contrast, very fast input/output systems with large capacity are mandatory for efficient data processing.

The software requirements come along with the hardware requirements. In contrast to traditional HPC applications, which require programming models and paradigms such as message passing or shared memory parallelism, data analytics applications rely on in-memory processing and programming languages such as Java, Python or Scala. So the most important applications for data analytics are currently

the Apache tools Spark[2], Hadoop[3], Storm[4] and Flink[5] as well as some smaller projects such as Disco Project[6], DataTorrent[7] or BashReduce[8].

Most of those applications build on the MapReduce algorithm, which has been introduced by the global player Google[9]. The MapReduce algorithm consists of three phases—map, shuffle and reduce, whereas the map and the reduce parts are directly specified by the user in order to allow parallel processing of data on manifold machines. Using his concept enables processing different kinds of data, reaching from structured data including files and databases up to unstructured and real-time data such as online data composed of several data structures.

3.3 A Practical Application Making Use of HPC and HPDA

In order to proof the statements of the last sections and sub-sections, the information shall be complemented with a practical example from the Global Systems Science community, which represents an emerging field in the HPC sector. Within the EC-funded CoeGSS project[10], a set of applications is focused that require particular workflows to retrieve the results. In particular, the workflow foresees HPDA, huge-scale HPC, small-scale HPC and visualization to generate synthetic populations, execute the resulting agent-based models and finally, visualize the results [1]. For clarification, the workflow and its targeted technologies is depicted in Fig. 1.

Thus, those kinds of applications demonstrate that there is a new need to support other methods and techniques than the classical HPC applications demand. As a consequence, being competitive in terms of hardware and software reaches a new level of complexity.

4 The Missing Link

Summarizing the previously mentioned evolution scenarios for High Performance Computing and the raise of High Performance Data Analytics, this seems as a promising and valuable way to go. However the deeper one dives into the implications of the use of these technologies and the potential they provide, it becomes obvious that

[2]Apache Spark: http://spark.apache.org

[3]Apache Hadoop: http://hadoop.apache.org

[4]Apache Storm: http://storm.apache.org

[5]Apache Flink: http://flink.apache.org

[6]DiscoProject: http://www.discoproject.org

[7]DataTorrent RTS: https://www.datatorrent.com

[8]BashReduce: https://github.com/erikfrey/bashreduce

[9]Google Inc.: http://www.google.com

[10]Centre of excellence for Global Systems Science: http://www.coegss.eu

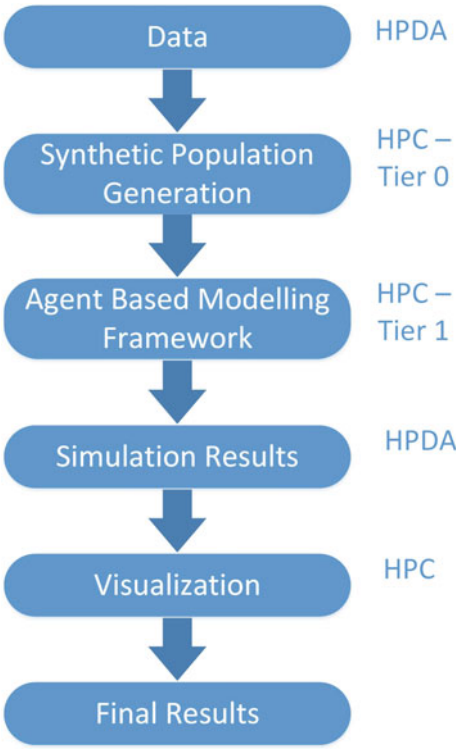

Fig. 1 CoeGSS application workflow

the resulting outputs, especially in terms of data variety and data size get hard to handle for a human in the loop.

We see a tendency in so-called "business-ready solutions" to stress the support of the human in the loop by application of technological fields such as machine learning, artificial intelligence or cognitive computing. For the remainder of this paper we will stick to the term cognitive computing as a placeholder for the above mentioned disciplines, which can be described as the variety of scientific disciplines of Artificial Intelligence and Signal Processing[11]. A similar view has been presented by James Kobielus, Big Data Evangelist, 2013, in a blog entry on *Cognitive Computing: Relevant at all Speeds, Scales and Scopes of Thought*, where he defines cognitive computing as

> the ability of automated systems to handle the conscious, critical, logical, attentive, reasoning mode of thought that humans engage in when they, say, play Jeopardy or try to master some academic discipline.

[11] Wikipedia Definition of Cognitive Computing: https://en.wikipedia.org/wiki/Cognitive_computing

4.1 Cognitive Computing

The principles of cognitive computing are not new, and nearly everyone who is in the Information Technology business has at a certain point in time heard of this topic. Thus is it also not surprising, that it's base assumptions and ideas were even reported already at the end of the 19th century, when Boole proposed its book on "The Laws of Thoughts" [2]. Even though this was just a conceptual approach, and the first programmable computer by Zuse needed As already mentioned before, during the evolution of these principles, the domain of cognitive methodologies and artificial intelligence went either side by side or showing clear overlaps. A variety of theories and implementation approaches were taken, the probably most prominent ones being so far IBM's Watson [3] and the recently presented AlphaGo [4].

4.2 Benefits

Figure 2 shows how High Performance Computing, High Performance Data Analytics and Cognitive Techniques can complement each other. High Performance Computing (HPC) delivers the needed processing power for those kind of applications, requiring massive parallel execution. At the same time, these kind of applications produce partially enormous amounts of data, which may be too big to be manually analysed, even having current support tools at hand. Thus the discipline of High Performance Data Analytics can be used to analyse and handle these (and other sources' data sets) in a sufficient way. Cognitive techniques can provide support to both disciplines, to help to interpret and present the results in a best possible way.

In a general way, the expected benefits from applying these concepts, are manifold. In general support for those fields where big amounts of data are collected, handled and interpreted is improved, examples are:

- Enhanced analysis of business potentials of new offerings/new activities. This can reach from the virtual testing of new opportunities, e.g. in drug design or on combined virtual and real world simulations such as finding new geographic locations for drilling
- Support of staff (e.g. engineers) in decision processes by providing them a selection of potential paths to follow
- Improving Operations by understanding of performed operations and their parameters, so that either in real time or after longer-duration analysis processes can be optimized

Taken this complementarity into account, the workflow as described in Fig. 1 can be extended to the one presented in Fig. 3.

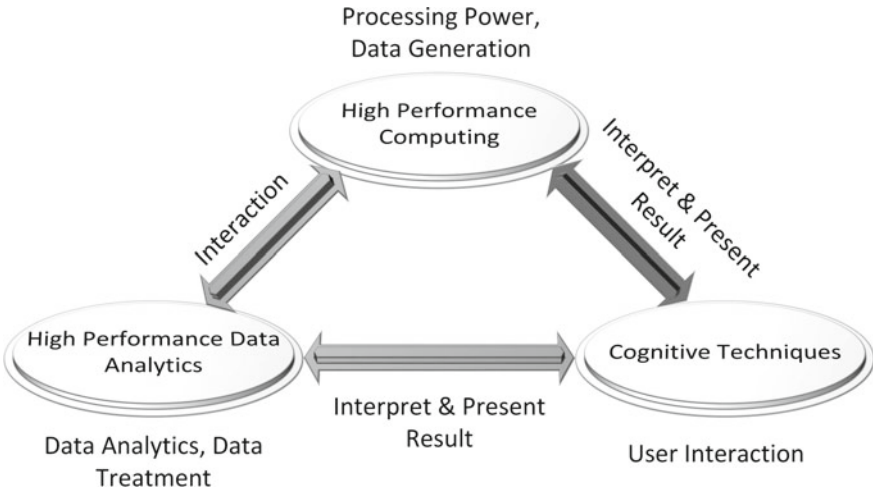

Fig. 2 Cognitive Techniques complementing the global picture of HPC and HPDA

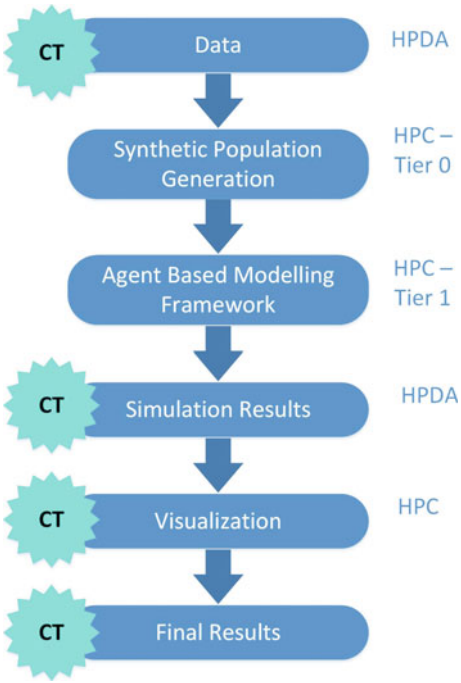

Fig. 3 Extending the GSS workflow with cognitive techniques support

4.3 Available Technologies

Within this document, we also want to have a short look at those technologies, which may act as baseline to realize an integration of cognitive concepts into a traditional HPC/HPDA based workflow (e.g. the one presented in Fig. 3.

In the case of Watson, a variety of APIs is available for selected developers and business users, as well as the Watson Analytics Solution[12]. Furthermore there is a variety of Open Source alternative available, which shall be discussed on a high level in the following overview:

DARPA DeepDive
DeepDive [5, 6] is a free version of a Watson like system. It was developed within the frame of the US Defense Advanced Research Projects Agency (DARPA) and in opposite to Watson has the aim to extract structured data from unstructured data sources. DeepDive uses machine learning technologies to train itself and targets especially those users with moderate to no machine learning expertise.

UIMA
Apache Unstructured Information Management (UIMA)[13] is supporting the analysis of large sets of unstructured information. Its an implementation of the Oasis Unstructured Information Management standard[14] **OpenCog**

OpenCog [7] is a project targeting artificial intelligence and delivering an open source framework. One output of OpenCog is the cognitive architecture OpenCog Prime [8] for robot and virtual embodied cognition.

5 Conclusions

The previous sections have pointed out that High Performance Computing and High Performance Data Analytics can be seen as rather complementary approaches, then as direct competitors. Even though there are activities to provide a common software stack, which may run on both, HPC and HPDA specific hardware, there is only a subset of concrete problems in the problem space which can be addressed efficiently in such a manner. Mainly, this is a result of the partially quite different hardware set up of the respective technological environment.

Now, assuming that HPC and HPDA work with a high performance, we also have to face the fact that the size and amount of data sets proceeded and again resulting from this processing enter a dimension, which makes a satisfactory manual processing by a human in the loop (e.g. an engineer) nearly impossible. Thus we see that even if there is an issue (e.g. data analytics) solved with those appliances, another issue pops up which is the understanding and respectively handling of information.

[12] http://www.predictiveanalyticstoday.com/ibm-watson-analytics-beta-open-business/
[13] http://uima.apache.org/
[14] https://www.oasis-open.org/committees/download.php/28492/uima-spec-wd-05.pdf

For that purpose we have introduced cognitive technologies, which can act as some sort of "helper" technology to simplify the life of the end user and enable for improved use of simulation results. This technology, even if it appears to be still in its infancy, can support the (human) end user and provide decision baselines allowing improved processing of information. We have shown that a variety of implementations already exist, next steps need to see in how far they can cover the requirements of selected use cases.

References

1. Wolf, S., Paolotti, D., Tizzoni, M., Edwards, M., Fuerst, S., Geiges, A., Ireland, A., Schuetze, F., Steudle, G.: D4.1 - First report on pilot requirements. http://coegss.eu/wp-content/uploads/2016/03/CoeGSS_D4_1.pdf
2. Boole, G.: Investigation of the Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities (1853)
3. Ferrucci, D.A., Brown, E.W., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J.M., Schlaefer, N., Welty, C.A.: Building Watson: an overview of the DeepQA project. AI Mag. **31**(3), 59–79 (2010)
4. Silver, D., Hassabis, D.: AlphaGo: mastering the ancient game of Go with Machine Learning, Blogpost. https://research.googleblog.com/2016/01/alphago-mastering-ancient-game-of-go.html (2016)
5. Niu, F., Zhang, C., Re, C., Shavlik, J.W.: DeepDive: web-scale knowledge-base construction using statistical learning and inference. 884. In: VLDS: CEUR-WS.org. (CEUR Workshop Proceedings), pp. 25–28 (2012)
6. Zhang, C.: DeepDive: a data management system for automatic knowledge base construction, Ph.D. Dissertation, University of Wisconsin-Madison (2015)
7. Hart, D., Goertzel, B.: OpenCog: a software framework for integrative artificial general intelligence. In: Wang, P., Goertzel, B., Franklin, S. (eds.) 'AGI', pp. 468–472. IOS Press (2008)
8. Goertzel, B.: OpenCog Prime: a cognitive synerfy based architecture for artificial general intelligence
9. Hurwitz, J.S., Kaufman, M., Bowles, A.: Cognitive Computing and Big Data Analytics. Wiley, Indianapolis (2015)