

# Exposé: Implementing an imperative language to do machine learning on web-scale data within parallel data flow systems.

Master's thesis

Till Rohrmann  
Technical University of Berlin, Germany  
till.rohrmann@mailbox.tu-berlin.de

## 1. INTRODUCTION

Since the begin of the information age, our environment becomes more and more infused with digital technology gathering, processing, storing and broadcasting information of various kind. Nowadays, computing devices are almost ubiquitous in our daily lives, which can hardly be imagined without them anymore. We can find information technology in our cars making our ride safer, in our washing machines controlling the proper cleansing and in our mobile phones acting as our personal assistant just to name a few. Apart from the personal use, digital devices found their way into and propel almost all important branches of science and industry such as finance, engineering, health, genetics, meteorology etc.

The information technology's key to success and the quick adoption is its unprecedented amplification in computing, storage and telecommunication capacity over the last couple of decades. The growth has been mainly spurred by the technological progress on the fields of integrated circuits, broadcasting technology and algorithms. Since digital devices are strongly linked to the semiconductor technology, the aforementioned capacities exhibit a similar exponential growth rate [6] as Moore [8] predicted for the number of components on integrated circuits. Nowadays, it is estimated that 2.5 ZB are produced daily and this number is supposed to rise up to 40 ZB by the year of 2040 [3].

With the increase of data, its usefulness became obvious and thus the demand for analyzing tools harnessing the gathered data increased as well. Currently, insights gained from collected data already helps to improve a variety of processes. For example, retail stores analyze their sales, customer, pricing and weather data in order to decide which products to offer or when to do a discount sale. Police departments try to detect probable crime locations by extracting patterns from previously recorded criminal acts and then reinforce the policemen in this region [7]. Hospitals analyze

their patients' records and scientific studies in order to find the cancer treatment best complying with the specifics of the patient and thus having the highest chance of success [2]. These examples underline the importance and utility of data analysis tools. The majority of these tools are based on statistic techniques. Consequently, they can leverage the massive amounts of data collected every day and improve their descriptive and predictive results. However, this comes at the price of an increased computing time which requires fast computer systems to make the computation feasible.

In recent years, we have seen that clock rates of CPUs stagnated. Before, there was a simple receipt to increase the computing power of micro-controllers; increase the clock rate, which demands more power, and shrink the technology to mitigate for the increased power consumption. However, the shrinkage induce the problem of leakage, which increases the power demand again. The consumed power is limited by the amount of energy you can dissipate and thus there is a technological limit for the increase of clock rate. When it became clear that the micro-controller would hit this so-called power wall, one duplicated the micro-controller's functionality to support simultaneous execution of multiple applications and to harness the inherent parallelism of programs.

The emerging multi-core and distributed systems pose new challenges for programmers, since now they have to know about locking, deadlocks, race-conditions and inter-process communication in order to make most of the available hardware. Due to this, parallel program development became cumbersome and error-prone. Therefore, new programming models are conceived which relieve the programmer from tedious low-level tasks related to parallelization, such as load-balancing, scheduling of parallel tasks or fault recovery. Instead, the programmer can concentrate on the actual algorithm and the goal he wants to achieve.

These are the reasons why Google's MapReduce [5] framework and its open source reimplementation Hadoop [1] became so popular among scientists as well as engineers. MapReduce is a programming framework for concurrent computations on vast amounts of data running on a large cluster. Originally, it is inspired by the often in the context of functional programming occurring functions *map* and *reduce*. Each MapReduce program is separated into a map-phase and a reduce-phase for which the user has to provide a user-defined function each. The input of the program is split up into key-value pairs and each pair is fed to the map operation which applies the user-defined map-function to it. This

operation can be run completely in parallel since each call to the map-function is independent of the other calls. Each map-function produces a list of new key-value pairs which are then grouped according to their key value. Pairs with the same key are then given to the reduce-function which then produces a new list of key-value pairs. Again, each call of the reduce function can be run in parallel since there is no dependency.

However, MapReduce exhibits also some deficiencies in realizing common sub-tasks such as joining two datasets. These deficiencies result from the strict programming model and cause some serious performance losses. There are some other distributed programming models such as Stratosphere [4] and Spark [9] which try to tackle these problems. They both support iterative algorithms which is a necessary requirement for many data analytical algorithms. Furthermore, they both offer a more flexible pipeline not requiring to store intermediate results always to hard disks.

But still, these frameworks force the user to express the program in a certain way, which is often not natural or intuitive for a user coming from a different domain. This implies that one has to overcome a particular entry-barrier to use the system and this might already be too high for some users. Furthermore, the actual program might become lengthy and complicated expressed within the programming model. This makes developing and debugging the program difficult as well as time-consuming, thus expensive. Especially in the field of data analytics and machine learning programs are usually expressed in a mathematical form. Therefore, systems such as Matlab and R are widely used and recognized for their fast prototyping capabilities and their extensive mathematical libraries. However, these linear algebra systems lack the support for automatic parallelization and thus restricting the user to a single workstation. Therefore, the amount of processable data is limited to the size of the main memory, which constitutes a serious drawback for real-world applications. Moreover, machine learning experts are usually no experts on the field of distributed computing and neither vice versa. The group of experts on both fields is negligible small. Consequently, it is very difficult, time-consuming and not least expensive to implement machine learning algorithms on distributed systems.

This problem would be mitigated by having a distributed sparse linear algebra system supporting a Matlab- and R-like language. Assuming that such a system is realizable, then one could run existing Matlab- or R-code directly or at most with minor adjustments in a distributed fashion. Furthermore, new distributed algorithms could be quickly implemented benefiting from the expressiveness of linear algebra. This would drastically speed up the application of machine learning and data analysis algorithms on web-scale data.

## 2. THESIS APPROACH

Within my master's thesis I want to answer the following questions:

- Which are the language primitives of a distributed sparse linear algebra system in order to support a multitude of current machine learning algorithms?
- How can the supported operations be realized in a distributed system?

- What kind of automatic optimization can be applied to the program in order to speed it up?

## 2.1 Language primitives

The set of language primitives has to include besides the operational primitives of linear algebra also an iteration abstraction in order to realize algorithms based on convergence. Furthermore, it is of particular interest to keep this set as small as possible, because this would alleviate a possible optimization step prior to execution. Another objective, though a smaller one, is to design the set of language primitives so general that they can be realized within different distributed systems.

## 2.2 Runtime

I want to research how the most common linear algebra operations can be efficiently represented within a MapReduce-like system. One key aspect is going to be the matrix matrix multiplication because it is one of the central operations of many machine learning algorithms and is one of the most time consuming operations as well. Therefore, it is crucial to efficiently implement this operation. This requires a thorough investigation and evaluation of all possible implementations. Since this will strongly depend on the actual distributed system, I will implement and evaluate these operations within Stratosphere. The Stratosphere implementation acts at the same time as a proof of concept of the whole distributed sparse linear algebra system. I have chosen Stratosphere over other systems, such as Spark, because of its expressiveness and its powerful parallel data-flow optimizer. Stratosphere extends the MapReduce framework with additional  $2^{nd}$  level functions, such as *match*, *cross* and *coGroup*. Last but not least, Stratosphere is developed at the chair of Prof. Markl, where I write my master's thesis.

## 2.3 Optimization

If I have still some time left, then I will research possible optimization strategies for linear algebra programs. One obviously important choice of execution would be the order of subsequent matrix matrix multiplications. Different execution plans might emit different intermediate results which vary in size and density and thus making one of the plans favorable over the other. I intend to use a similar cost-based plan enumeration approach as it is used for database query optimizations.

## 2.4 Implementation

## 2.5 Evaluation

## 3. RELATED WORK

- Spark and Stratosphere
- Pegasus
- SystemML

## 4. REFERENCES

- [1] Hadoop. <http://hadoop.apache.org/>, 2008. Cited January 8th 2014.

- [2] Ibm watson hard at work: New breakthroughs transform quality care for patients. <http://www.mskcc.org/pressroom/press/ibm-watson-hard-work-new-breakthroughs-transform-quality-care-patients>, 2013. Cited January 8th 2014.
- [3] Big data within ibm. <http://www.ibm.com/big-data/>, 2014. Cited January 8th 2014.
- [4] D. Battré, S. Ewen, F. Hueske, O. Kao, V. Markl, and D. Warneke. Nephele/pacts: a programming model and execution framework for web-scale analytical processing. In *Proceedings of the 1st ACM symposium on Cloud computing*, pages 119–130. ACM, 2010.
- [5] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.
- [6] M. Hilbert and P. López. The worlds technological capacity to store, communicate, and compute information. *Science*, 332(6025):60–65, 2011.
- [7] S. Lohr. The age of big data. *New York Times*, 11, 2012.
- [8] G. E. Moore et al. Cramming more components onto integrated circuits, 1965.
- [9] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: cluster computing with working sets. In *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*, pages 10–10, 2010.