Towards Streamlined Big Data Analytics

by András A. Benczúr, Róbert Pálovics (MTA SZTAKI) , Márton Balassi (Cloudera) , Volker Markl, Tilmann Rabl, Juan Soto (DFKI) ,Björn Hovstadius, Jim Dowling and Seif Haridi (SICS)

Big data analytics promise to deliver valuable business insights. However, this will be difficult to realise using today's state-of-the-art technologies, given the flood of data generated from various sources. The European STREAMLINE project develops scalable, fast reacting, and high accuracy machine learning techniques for the needs of European online media companies.

Big data analytics promise to deliver valuable business insights. However, this will be difficult to realise using today's state-of-the-art technologies, given the flood of data generated from various sources. The European STREAMLINE project [L1] develops scalable, fast reacting, and high accuracy machine learning techniques for the needs of European online media companies.

A few years ago, the term "fast data" arose to capture the idea that streams of data are generated at very high rates, and that these need to be analysed quickly in order to arrive at actionable intelligence.

To this end, the EU Horizon 2020 funded STREAMLINE project aims to address the aforementioned technical and business challenges. The STREAMLINE consortium's three research partners, MTA SZTAKI (Hungary), SICS (Sweden), and DFKI (Germany) and four industry partners, Rovio (Finland), Portugal Telecom, NMusic (Portugal), and Internet Memory Research (France) will jointly tackle problems routinely faced by online media companies, including customer retention, personalised recommendation, and web-scale data extraction. Collectively, these four industrial partners serve over 100 million users, offer services that produce billions of events yielding over 10 TB of data daily, and possess over a PB of data-at-rest.

Today's big data analytics systems face two types of latency bottlenecks, namely, system latency and human latency. System latency issues arise due to the absence of appropriate (data) stream-oriented analytics tools and more importantly the added complexity, cost, and burden associated with simultaneously supporting analytics for both data-at-rest and data-in-motion. Human latency results from the heterogeneity of existing tools and the low-level programming languages required for product or service development that rely on an inordinate number of boilerplate codes are system specific (e.g., Hadoop, SolR, Esper, Storm, and relational database management systems) and demand a plethora of scripts to glue systems together.

Developing analytics that that are well-suited for both data-at-rest and data-in-motion is non-trivial. Prior to our development, even the simplest case had no integral solution when we train a predictor on historic data and use the streaming API to give predictions with the trained model. In our solution using Flink's unified runtime environment [L2], the model's input for the fitting is drawn from the batch environment and the unlabelled data for the predictor is drawn from the streaming environment without ever needing to explicitly store models or switch between different runtime environments.

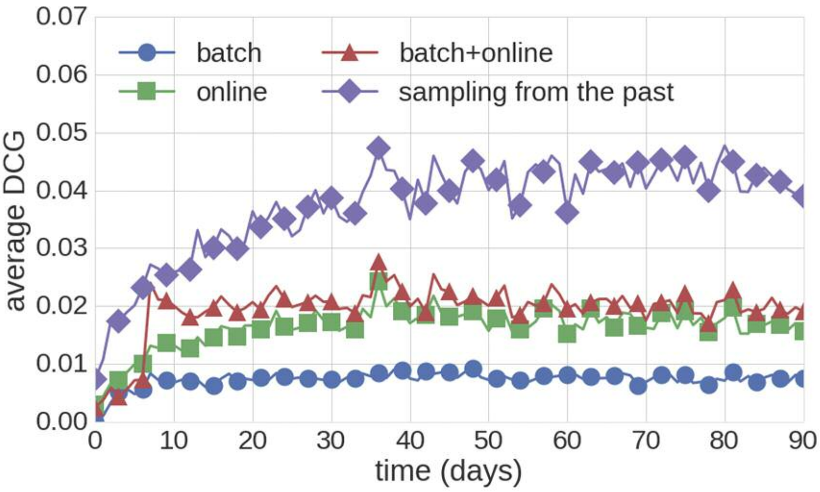

Figure 1: Combined batch and online method system performance.

As another machine learning example, let us consider a typical recommendation problem of our partners. Their tasks are implicit as users give no ratings: we only have information on their interaction with the items (e.g., clicks, listening, view). In this top-k recommendation task, we have to provide a list of the best k items for a given user. In this task, we contrasted batch and online learning methods. Batch machine learning repeatedly reads all training data multiple times, e.g., via stochastic gradient descent, which uses records multiple times in random order, or via elaborate optimisation procedures, e.g., involving SVMs. The common belief is that these methods are more accurate and easier to implement than online machine learning.

Online learning recommenders perform updates immediately, in the data-streaming model by reading events only once. These models adapt fast to the actual mood of the user, as seen in our experiments over the 30M music listening dataset, crawled by the CrowdRec [L3] team. As shown in Figure 1, online learning outperforms batch trained models in terms of Discounted Cumulative Gain (DCG). We may only gain slight improvements by a loose integration of a linear combination of batch and streaming predictions. The strongest method, on the other hand, requires tight integration by injecting past samples into a data stream matrix factorisation method.

Although Apache Spark is currently a clear leader in the next generation open source big data platform scene both in terms of market penetration and community support, Flink is gaining solid momentum to rival it. We aim to determine whether Apache Flink as a data processing platform has the potential to become a leading player of the open source Big Data market. For example, in Figure 2 we have repeated the benchmarks of Data Artisans [L4] with the latest version of Flink and Spark to depict Alternating Least Squares performance. Another important benchmark that emphasises Flink's low latency capabilities is performed by Yahoo [L5] .

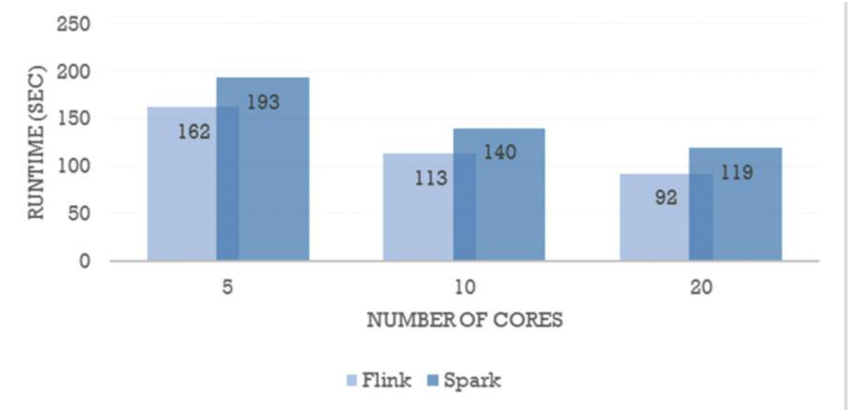

Figure 2: Matrix Factorization, 1 billion entries.

In addition to evaluating system latency in comparison with existing alternatives (Spark, Storm), we also aim to evaluate human latency, defined as the effort and time spent on setup, preparation, and development. To do that we will conduct some test subjects to implement data analysis problems, including (i) Bachelor and Master students, as well as data scientists from our industrial partners and participants at Flink Hackathons.

The goal of our evaluation under various evaluation criteria and use cases is to determine whether Flink is technologically suited to eventually replace Spark or is destined to remain a niche solution for streaming analytics.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 688191.

Links:
[L1] http://streamline.sics.se/
[L2] http://flink.apache.org/
[L3] http://crowdrec.eu/
[L4] http://data-artisans.com/computing-recommendations-at-extreme-scale-with-apache-flink/#more-81
[L5] http://yahooeng.tumblr.com/post/135321837876/benchmarking-streaming-computation-engines-at

Please contact:
Björn Hovstadius, STREAMLINE Project Coordinator, SICS, Sweden
bjornh@sics.se

Autonomous Machine Learning

by Frederic Alexandre (Inria)

Inspiration from human learning sets the focus on one essential but poorly studied characteristic of learning: Autonomy.

One remarkable characteristic of human learning is that, although we may not excel in any specific domain, we are quite good in most of them, and able to adapt when a new problem appears. We are versatile and adaptable, which are critical properties for autonomous learning: we can learn in a changing and uncertain world. With neither explicit labels, nor data preprocessing or segmentation, we are able to pay attention to important information and neglect noise. We define by ourselves our goals and the means to reach them, self-evaluate our performances and apply previously learned knowledge and strategies in different contexts. In contrast, recent advances in machine learning exhibit impressive results, with powerful algorithms surpassing human performance in some very specific domains of expertise, but these models still have very poor autonomy.

Our Mnemosyne Inria project-team is working in the Bordeaux Neurocampus with medical and neuroscientist teams to develop systemic models in computational neuroscience, focusing on these original characteristics of human learning. Our primary goal is to develop models of the different kinds of memory in the brain and of their interactions, with the objective to exploit them to study neurodegenerative diseases, and another important outcome of our work is to propose original models in machine learning, integrating some of these important characteristics.

We believe that important steps toward autonomous learning can be made along the following lines of research:

Developing an interacting system of memories

Specific circuits in the brain are mobilised to learn explicit knowledge and others to learn procedures. In addition to modelling these circuits, studying their interactions is crucial to understanding how one system can supervise another, resulting in a more autonomous way of learning. In the domain of perceptual learning in the medial temporal lobe, we model episodic memories storing important events in one trial, and forming later, by consolidation in other circuits, new semantic categories. In the domain of decision-making in the loops between