Labex UCN@Sophia PhD Proposal , January 2015

Main advisor

Francoise Baude (PR), Scale team CNRS I3S
Mail: Francoise.baude@unice.fr
Web page:http://www-sop.inria.fr/members/francoise.baude/

Co-advisor

Didier Parigot (HDR), Zenith team, INRIA CRISAM
Mail: Didier.Parigot@inria.fr
Web page: http://www-sop.inria.fr/members/Didier.Parigot/

Predictive Big Data Analytics: continuous and dynamically adaptable data request and processing

Context:
As the popularity of Big Data explodes, more and more use cases are implemented using this kind of technologies. But there are some use cases that are not properly tackled by classic Big Data models and platforms like Apache Hadoop MapReduce because of these models intrinsic batch nature. These cases are those where online processing of new data is required as soon as they enter the system, in order to aggregate to the current analysis results the newest information extracted from these incoming data. Such on-line and continuous processing pertains to what is known as continuous query and trigger in the more focused context of databases [1][2], or also as complex event processing in publish-subscribe systems [3]. More generally, processing the incoming data is known as Data Stream processing, and in the big data area is known as real-time data analytics.

Social networks are nowadays the medium of choice for data delivery among end-users, be it in public circles or in more private spheres as professional dedicated networks. Analysing, understanding, recommending, rating all the vast amount of data of various kinds (text, images, videos, etc) is a feature more and more required to be supported by the underlying systems of those social networks. In particular, the Beepers startup associated with the Zenith team, is working towards an alerting tool to be part of their offered toolbox for building a social network. The goal of this joint PhD is partly motivated by this perspective as deploying an alert pertains to a persistent -possibly sophisticated- query and thus a data streaming program. Thanks to the partnership established through this common PhD research, we want to offer Beepers as an added-value, the capability for the query to dynamically evolve. The motivation for this evolution is to better stick to the discovery of a trend, a sentiment, an opinion, etc. that is emerging as extracted by the continuous data flow analytics.

Indeed, some situations have the need of what could be named *anticipatory* analytics: given gathered data originating from various sources and combined to get meaningful information out of them, the goal is to adapt the current analytics in such a way that it can

match to the anticipated coming situation, somehow ahead of time. For instance, doing short-term weather forecast for local places : if suddenly the speed of the wind increases and changes direction while intense rain falls, there is a need for (1) updating the short-term weather predictions, but also for (2) deploying appropriate supervision of the now-in-danger zone, so that, in case of a flooding risk of the new targeted zone, assuming the new zone is an inhabited one, the system gets able to trigger alerts towards the right actors: if now a flooding can reach the hospital, evacuation effectively should start, whereas it was not necessary few minutes before when wind was still soft and given its direction the risky zone was a field full of corn; now, supervising also electricity delivery is required to make sure any blackout during evacuation risk can be anticipated. Another example of anticipated analytics is in multimedia stream processing: for example, a data journalist that is extracting data from its preferred social network and runs analytics on videos to learn about happenings in a specific city area prior to writing its press article, suspects from the broadcasted content to recognize someone possibly involved in a crime scene shown in a concomitant published video; and, accordingly, decides to adapt its current analytics to combine the two information sources around some common data which is the relatives whose accommodation lies in the specific city area, and continues its search focusing now on these relatives acts while still monitoring the initially recognized suspect.
From these exemplary scenarios, we clearly foresee that one trend in big data technologies is in predictive analytics [16][17][18] which by nature requires a strong capability of dynamic adaptation given partial already gained results. In this trend, our claim is that the analytics require to adapt (even better self-adapt) to what is happening, given of course some previously user-defined rules dictating which adaptation it could be relevant to decide to trigger.

Work:
Several big data platforms geared at real-time analytics have emerged recently: Spark Streaming [4], Twitter's Storm, S4[5]. These platforms allow one to define a program as taking eventually, after a compilation process, the form of a DAG (directed acyclic graph) but to our knowledge, none allows to adapt the program at runtime with respect to its functional/business nature. This is because these languages such as StreamSQL, CQL, StreamIt, and IBM's SPL [7] are generally declarative. As a result, developers focus on expressing data processing logic, but not orchestrating runtime adaptations (i.e. in which situation functional adaptation should happen, how to monitor the adaptation needs, how to effectively modify the program without redeploying a new one from scratch) . Some data workflow engines start to appear in the community of large-scale stream processing but without yet an adequate behavioural/functional adaptation capability, as the so far only focus was to allow adaptation to face requirements for non-functional changes dictated by scalability, fault-tolerance, performance needs through node replication or elasticity [6]. From our literature survey done so far, only [8] sketches a solution for functional adaptivity in stream processing languages of big data platforms, even if some anterior works on stream platforms or active databases are a good starting point [9].

As a result from some years of research in the Scale team, Grid Component Model (GCM) [10] is a component model for applications to be run on distributed infrastructures, that

extends the Fractal component model. Also, Zenith' research have resulted in strong competences in component oriented platforms featuring high dynamicity like SON [11] (and that the co-supervised PhD student could also take advantage from). Fractal defines a component model where components can be hierarchically organized, reconfigured, and controlled offering functional server interfaces and requiring client interfaces. GCM extends that model providing to the components the possibility to be remotely located, distributed, parallel, and deployed in a large-scale computing environment (cluster, grid, cloud, etc), and adding collective communications (multicast and gathercast interfaces). Autonomic capabilities can be expressed in the membranes of GCM components, that can drive their reconfiguration at functional level, in a totally autonomous manner [10]. If the DAG of a streaming application translates into a component oriented program, then it can naturally benefit from its intrinsic reconfiguration properties. This is one of the expected research question to be addressed in the scope of this PhD: how to benefit from autonomic, high-expressivity, clear functional versus non-functional separation of concerns features of the GCM component-oriented approach in order to support dynamic adaptation of the analytics the streaming application corresponds to.

Work:

Towards the global goal of designing a working solution for anticipated big data analytics, the development of the PhD could be organized along the following guidelines :

- Investigate on the stream processing languages (mainly Domain Specific Languages) that can be used atop of GCM composition framework, and extended to express adaptability functionalities. An approach developed in [15] relying upon data flow analysis of the DSL code could be reused to infer the GCM-based graph.
- Extend the GCM model and runtime to be stream processing aware
- Implement the framework to be able to write adaptable stream processing workflows relying in fine on GCM and on SON(to handle the needed dynamic code deployment features).
- Study existing and emerging stream processing platforms (that are mainly open source and supported by Apache) and select or extend the most appropriate one to plug to it the GCM-based dynamic reconfiguration of analytics feature and new DSL; alternatively but more ambitious, if existing platforms extensibility is not easy, develop a complete GCM-based big data real-time analytics solution, and make sure it can interact with existing big data providers (famous public social networks, as twitter, facebook, or dedicated social networks as built with the Beepers technology)
- Benchmark and test on relevant big data analytics use-cases, extracting relevant information from social networks contents (with semantic annotation) like images, videos, text; applying to these contents for instance sentimental analysis to predict future situations [12], and accordingly adjust recommendations provided to the users of the social network [13].

Complementarity and perspectives

This research proposal is at the crossing of middleware for large-scale platforms working on large data volume, languages (including DSLs), data mining, social networking. Françoise Baude from SCALE team is an expert in runtime and middleware for distributed languages, and a recent EU project, *PLAY*, allowed her to gain experience in situational-awareness through complex event processing and supporting publish/subscribe platforms for web-semantic described events [14]. Didier Parigot from Zenith is addressing DSLs, data flow models, and also middleware and databases [15] to support social networking, through the *iLab* collaboration with Beepers. Overall, the two teams share a common background, while having complementary assets applicable to the emerging technologies for big data in general. The two teams have here a first opportunity to collaborate. They also have a strong willingness proved by past and current efforts, to transfer research results towards the industry, which has a rising interest in data analytics supporting solutions. This common PhD research stands up as a nice catalyst and opportunity to demonstrate usability and relevance of past developed platforms (ProActive/GCM, and SON respectively, that are going to be used in complementary aspects), on this exciting emerging area of adaptable big data analytics.

Innovation potential:
Due to the applied nature of the research there are obvious perspectives of valorization as a "product" handled by existing or to come SMEs of the Sophia-Antipolis ecosystem. Consequently, the candidate may apply for an ICT Labs Doctoral Training Center additional funding. Indeed, the Labex@UCN has been labelled as the foundation for the newly funded Doctoral Training Center in Sophia-Antipolis from 2015. In this context, six additional months of PhD funding, plus an Innovation & Entrepreneurship curriculum, and the opportunities to evaluate the project in other ICT Labs ecosystems are allocated. There are at least two natural ICT Labs collaborating ecosystems that are strongly engaged in Big Data analytics efforts, which consequently deserve our attention. These are:

- Technische Universität Berlin (TUB), which coordinates the Berlin Big Data Center (BBDC), a national center on Big Data recently established by the German Federal Ministry of Education and Research (BMBF). TUB established the Data Analytics Laboratory (DAL) in 2011 to serve as a focal point for innovative research. TUB is the birthplace of one of the leading open-source big data analytics platform called Stratosphere, now Apache Flink, with an active worldwide user community. This is one of the possible systems that could be extended by the thesis work, to provide adaptive analytics.
- The University of Trento (UNITN), an associated partner of EIT ICT labs Trento CLC, as the Politechnico de Milano. Both also offer an ecosystem active in big data analytics solutions. Polimi and its ecosystem, benefit from the expertise of the researchers working in the Collaborative Innovation Center on Big Data (CIC) jointly developed with IBM.

References :

[1] Shivnath Babu and Jennifer Widom. 2001. Continuous queries over data streams. *SIGMOD Rec.* 30, 3 (September 2001), 109-120

[2] Scheuermann, Peter, and Goce Trajcevski. "Active Database Systems." *Wiley Encyclopedia of Computer Science and Engineering* (2008).

[3] Eugene Wu, Yanlei Diao, and Shariq Rizvi. 2006. High-performance complex event processing over streams. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data* (SIGMOD '06)

[4]Matei Zaharia, Tathagata Das, Haoyuan Li, Timothy Hunter, Scott Shenker, and Ion Stoica. 2013. Discretized streams: fault-tolerant streaming computation at scale. In *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles* (SOSP '13)

[5] Neumeyer, Leonardo, et al. "S4: Distributed stream computing platform." *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*. IEEE, 2010.

[6] Raphaël Barazzutti, Thomas Heinze, Andre Martin, Emanuel Onica, Pascal Felber, Christof Fetzer, Zbigniew Jerzak, Marcelo Pasin, Etienne Riviere:
Elastic Scaling of a High-Throughput Content-Based Publish/Subscribe Engine. ICDCS 2014: 567-576

[7] Martin Hirzel, Henrique Andrade, Bugra Gedik, Vibhore Kumar, Giuliano Losa, Mark Mendell, Howard Nasgaard, Rboert Soulé, Kun-Lung Wu - *SPL Stream Processing Language Specification* - IBM, 2009

[8] Gabriela Jacques-Silva, Bugra Gedik, Rohit Wagle, Kun-Lung Wu, Vibhore Kumar - *Building User-defined Runtime Adaptation Routines for Stream Processing Applications* - The Very Large Data Base Endowment Journal (VLDB), 2012

[9] Trajcevski, Goce, et al. "Evolving triggers for dynamic environments." *Advances in Database Technology-EDBT 2006*. Springer Berlin Heidelberg, 2006. 1039-1048.

[10] F. Baude, L. Henrio, C. Ruz *Programming Distributed and Adaptable Autonomous Components - the GCM/ProActive Framework* Software: Practice and Experience, Wiley, In Press, 2015

[11] Ayoub Ait Lahcen, Didier Parigot, "A Lightweight Middleware for Developing P2P Applications with Component and Service-Based Principles", *CSE*, 2012, 2012 IEEE 15th International Conference on Computational Science and Engineering (CSE 2012),

[12] Asur, Sitaram, and Bernardo A. Huberman. "Predicting the future with social media." *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*. Vol. 1. IEEE, 2010.

[13] Fady Draidi, Esther Pacitti, Didier Parigot, and Guillaume Verger. 2011. P2Prec: a social-based P2P recommendation system. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (CIKM '11)

[14] N. Stojanovic, R. Stühmer, F. Baude, P. Gibert. Tutorial: *Where Event Processing Grand Challenge meets Real-time Web: PLAY Event Marketplace* DEBS'12, the 6th ACM International conference on Distributed Event-based system, ACM, 2012, p. 341-352 July 2012.

[15]Ayoub Ait Lahcen, Developing Component-Based Applications with a Data-Centric Approach and within a Service-Oriented P2P Architecture: Specification, Analysis and Middleware, PhD thesis co-supervised by D. Parigot, Dec 2012

[16] Boulos, Maged N. Kamel, et al. "Social Web mining and exploitation for serious applications: Technosocial Predictive Analytics and related technologies for public health, environmental and national security surveillance." *Computer Methods and Programs in Biomedicine* 100.1 (2010): 16-23.

[17]Lozada, Brian A. "The Emerging Technology of Predictive Analytics: Implications for Homeland Security." *Information Security Journal: A Global Perspective* 23.3 (2014): 118-122.

[18]Doyle, Andy, et al. "The EMBERS architecture for streaming predictive analytics." *Big Data (Big Data), 2014 IEEE International Conference on*. IEEE, 2014.