Ecole Doctorale STIC

Sciences & Technologies de l'Information et de la Communication

ED STIC - Proposition de Sujets de Thèse

pour la campagne d'Allocation de thèses 2017

Axe Sophi@Stic :	SystèmesRéseaux|
Titre du sujet :	Predictive Big Data Analytics: safe dynamically adaptable data streaming analytics
Mention de thèse :	Informatique
HDR Directeur de thèse inscrit à l'ED STIC :	Françoise Baude

Co-encadrant de thèse éventuel :

Nom :	HUET
Prénom :	Fabrice
Email :	fabrice.huet@unice.fr
Téléphone :	0492942691

Email de contact pour ce sujet :	fabrice.huet@unice.fr
Laboratoire d'accueil :	I3S

Description du sujet :

As the popularity of Big Data explodes, more and more use cases are implemented using this kind of technologies, and batch processing is not anymore sufficient.
Processing the incoming data is known as Data Stream processing, and in the big data area is known as real-time data analytics.

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. 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.
Several big data platforms geared at real-time analytics have emerged recently: Spark Streaming, Twitter's Storm, S4. 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.
Also, an open question remains about the way data streams have to be managed (stopped, paused,...) during the reconfiguration process. It depends about what properties about the data are seek (should all tuples be handled, or can the application afford to lose some of them, etc). Ensuring the needed guarantees requires relying upon a sound data stream analytics programming model and support.

As a result from some years of research in the Scale group, Grid Component Model (GCM) [1] is a component model for applications to be run on distributed infrastructures, that extends the Fractal component model. Autonomic capabilities can be expressed in the membranes of GCM components, that can drive their reconfiguration at functional level, in a totally autonomous manner.
If the DAG of a streaming application translates into a component oriented program, then it can naturally benefit from its intrinsic reconfiguration properties, and we can work on providing reconfiguration support at both functional and non functional level (aka elasticity for e.g. better performance and lower energy consumption [2]) . 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. We have started to define a streaming platform, based upon GCM. The goal of this thesis is to pursue this preliminary work [3], and as such innovate even more in the context of self-adaptation of big data stream analytics at both application and platform support sides.

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

[2]Tiziano De Matteis, Gabriele Mencagli, Proactive elasticity and energy awareness in data stream processing. The journal of systems and Software, 127 (2017) 302--319, Elsevier

[3] F. Baude, L. El Bèze, M. Oliva Towards a flexible data stream analytics platform based on the GCM autonomous software component technology In HPCS'2016, workshop on Autonomic HPC. IEEE, July 2016.

English version:

As the popularity of Big Data explodes, more and more use cases are implemented using this kind of technologies, and batch processing is not anymore sufficient.
Processing the incoming data is known as Data Stream processing, and in the big data area is known as real-time data analytics.

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. 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.
Several big data platforms geared at real-time analytics have emerged recently: Spark Streaming, Twitter's Storm, S4. 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.
Also, an open question remains about the way data streams have to be managed (stopped, paused,...) during the reconfiguration process. It depends about what properties about the data are seek (should all tuples be handled, or can the application afford to lose some of them, etc). Ensuring the needed guarantees requires relying upon a sound data stream analytics programming model and support.

As a result from some years of research in the Scale group, Grid Component Model (GCM) [1] is a component model for applications to be run on distributed infrastructures, that extends the Fractal component model. Autonomic capabilities can be expressed in the membranes of GCM components, that can drive their reconfiguration at functional level, in a totally autonomous manner.
If the DAG of a streaming application translates into a component oriented program, then it can naturally benefit from its intrinsic reconfiguration properties, and we can work on providing reconfiguration support at both functional and non functional level (aka elasticity for e.g. better performance and lower energy consumption [2]) . 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. We have started to define a streaming platform, based upon GCM. The goal of this thesis is to pursue this preliminary work [3], and as such innovate even more in the context of self-adaptation of big data stream analytics at both application and platform support sides.

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

[2]Tiziano De Matteis, Gabriele Mencagli, Proactive elasticity and energy awareness in data stream processing. The journal of systems and Software, 127 (2017) 302--319, Elsevier

[3] F. Baude, L. El Bèze, M. Oliva, Towards a flexible data stream analytics platform based on the GCM autonomous software component technology In HPCS'2016, workshop on Autonomic HPC. IEEE, July 2016.