PARConnector : sans détour du Langage R au Cloud Big Data Application à l'analyse Metagenomique

I. Alshabani[b], J.-M. Batto[a], M. Berland[a], D. Caromel[b], N. A. Gaye[a], E. Le Chatelier[a], L. Pellegrino[b], N. Pons[a], E. Prifti[a] , F. Viale[b]

[a] MetaGenoPolis, US1367 INRA
Domaine de Vilvert, 78352 Jouy-en -Josas
CEDEX, France
jean-michel.batto@jouy.inra.fr

[b] ActiveEon
Les Algorithmes - Pythagore B
06560 Sophia Antipolis
denis.caromel@activeeon.com

Mots clefs : HPC, Big Data, ProActive, R, Quantitative Analysis, Metagénomique Quantitative

L'analyse quantitative du microbiote humain a été développée conjointement par l'équipe INRA-MetaGenoPolis (www.mgps.eu) dans le cadre du projet Européen MetaHIT (www.metahit.eu) qui impliquait 13 partenaires académiques et industriels. Cette approche explore l'information génomique des bactéries vivant dans le tractus digestif de manière globale et quantifiée [1]. L'analyse repose sur 'R' qui présente l'avantage d'avoir une gratuité et un vaste choix en modules et traitements.
Dans un environnement de traitement HPC, l'orchestration permet de présenter à l'utilisateur une puissance de traitement sans en exposer la complexité sous-jacente. Nous présentons ici notre approche et les outils utilisés pour effectuer du traitement Big Data en lien avec la metagénomique quantitative à travers l'utilisation de l'orchestrateur open source ProActive [2].
L'intérêt de cette approche portée par l'éditeur ActiveEon est que la complexité de l'invocation de la chaîne HPC est masquée. Ainsi dans une invocation de type variation paramétrique, l'utilisateur peut se concentrer sur son script 'R' sans devenir spécialiste du métier du HPC tels les protocoles réseaux ou les couches de liaisons (i.e. MPI).

PARConnector: L'API ProActive pour 'R'
L'orchestrateur ProActive permet de piloter un ensemble de noeuds de calculs. Ce pilotage est réalisé soit en ligne de commande soit avec une interface graphique qui permet de superviser l'ensemble des moyens de calcul. Chaque traitement est un job qui va être exécuté sur un ou plusieurs noeuds de calculs.

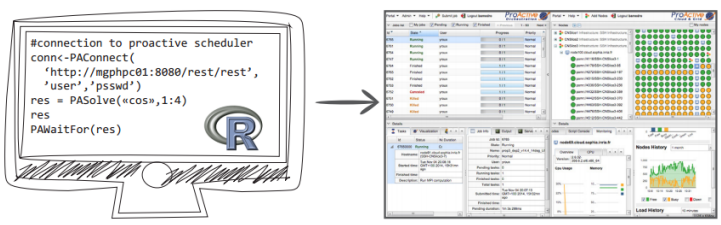

Figure 1 : Usage interactif du PARConnector.

Depuis 2014, "PARConnector" est un module 'R' qui apporte une solution simple de partitionnement et de parallélisation des calculs scientifiques à travers l'orchestrateur ProActive [3]. La soumission des jobs à exécuter se fait par l'intermédiaire d'une fonction nommée `PASolve` du

package 'R' PARConnector. Cette fonction, qui prend comme arguments une fonction R et un jeu de paramètres à évaluer, se charge de la distribution des tâches associées aux jobs via l'intermédiaire de l'orchestrateur ProActive qui a pour objectif de tirer avantage de toutes les ressources de calcul disponibles quelle que soit l'infrastructure cible (ordinateurs de bureau, les serveurs, clusters ou encore Clouds privés et publics) ou le système d'exploitation utilisé (Linux, Mac OS X, Windows). En plus de pouvoir orchestrer l'exécution des jobs, la console de supervision permet le monitoring en temps réel des jobs et des ressources disponibles (cf. Figure 1) tout en offrant des mécanismes de tolérance aux pannes et d'auto-scaling.

Étude de cas: Metagénomique Quantitative

La metagénomique quantitative se concentre sur la collecte du génome des espèces qui composent un écosystème donné. Le microbiote intestinal constitue un écosystème d'intérêt majeur pour le domaine biomédical. La metagénomique quantitative analyse l'ADN bactérien diversifiée et aide à établir l'abondance relative de ses composantes (espèces bactériennes). Cela produit une mesure de l'abondance de l'ADN sur une matrice dense avec des milliers de colonnes et des dizaines de millions de lignes ou chaque valeur est une valeur à virgule flottante.

Travailler sur de telles structures de données avec des méthodes classiques relève du défi en raison du nombre très important de dimensions à supporter. Les bioanalystes utilisent habituellement la plate-forme R pour son excellence dans l'analyse et de traitement des données. Toutefois, dans les cas de données massives comme c'est le cas ici, une approche de type traitement data-parallèle doit être envisagée. En segmentant l'information et en ayant des noeuds de traitements 'R' pilotés par l'orchestrateur ProActive, il est possible de faire du traitement de données massive par l'intermédiaire de PASolve. En résumé, PARConnector offre un modèle à la Map/Reduce qui s'intègre facilement avec les scripts R des utilisateurs.

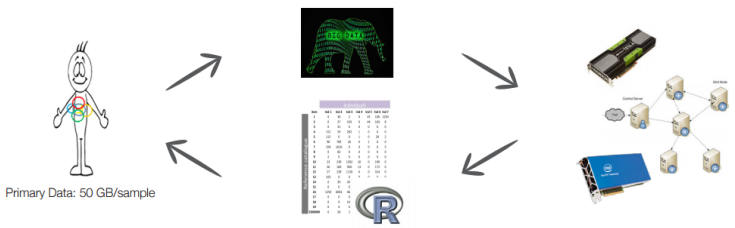

Figure 2 : Interactions dans notre cas metagénomique.

De part l'excellente gestion des infrastructures sous-jacentes (GPU, XeonPHI, Cloud, ...) dans l'orchestrateur ProActive, l'invocation d'un script via PARConnector tirera bénéfice des meilleures architectures matérielles sans que l'utilisateur ait à développer une expertise métier pour ces architectures. Le rôle d'ActiveEon est d'amener la puissance HPC dans sa globalité vers l'utilisateur 'R' en lui permettant de se concentrer sur son métier d'analyse statistique.

La versatilité de l'orchestrateur permet également de mixer des scripts et des traitements à partir d'autres langages. De ce fait, l'utilisateur 'R' peut également construire un workflow mixant des traitements 'R' et d'autres traitements, comme illustré par la Figure 2.

Nous proposons donc pour cette rencontre une présentation et des démonstrations « live » du package PARConnector, appliqués à la metagénomique quantitative.

Références

[1] Le Chatelier, Emmanuelle, et al. "Richness of human gut microbiome correlates with metabolic markers." Nature 500.7464 (2013): 541-546

[2] Caromel, D. "ProActive Parallel Suite: Multi-cores to Clouds to autonomicity." Intelligent Computer Communication and Processing, 2009. ICCP 2009. IEEE 5th International Conference

[3] ProActive Live Tour, Online: https://try.activeeon.com/tutorials/r/r.html