Eoulsan : analyse du séquençage à haut débit dans le cloud et sur la grille

Sartirana Andréa(1), Jourdren Laurent(2), Mora de Freitas Paulo (3), Chamont David(4), Busson Philippe (5), Le Crom Stéphane (6)

(1) sartiran@llr.in2p3.fr, Laboratoire Leprince-Ringuet, CNRS UMR 7638, École Polytechnique, 91128 Palaiseau Cedex, France.
(2) jourdren@biologie.ens.fr, École normale supérieure, Institut de Biologie de l'ENS, IBENS, Plateforme Génomique, Paris, F-75005 France; Inserm, U1024, Paris, F-75005 France; CNRS, UMR 8197, Paris, F-75005 France.
(3) paulo.moradefreitas@dr2.cnrs.fr, Délégation Régionale Paris B, CNRS MOY200, 75005 Paris
(4) chamont@llr.in2p3.fr, Laboratoire Leprince-Ringuet, CNRS UMR 7638, École Polytechnique, 91128 Palaiseau Cedex, France.
(5) busson@llr.in2p3.fr, Laboratoire Leprince-Ringuet, CNRS UMR 7638, École Polytechnique, 91128 Palaiseau Cedex, France.
(6) lecrom@biologie.ens.fr, École normale supérieure, Institut de Biologie de l'ENS, IBENS, Plateforme Génomique, Paris, F-75005 France; Inserm, U1024, Paris, F-75005 France; CNRS, UMR 8197, Paris, F-75005 France. Université Pierre et Marie Curie - Paris 6, UMR7622 Paris, F-75005 France; Centre National de la Recherche Scientifique, UMR7622, Laboratoire de Biologie du Développement, Equipe analyse des données à haut débit en génomique fonctionnelle, Paris, F-75005 France

Overview:

New generations of high throughput sequencing machines will spread over the scientific community an exponentially increasing amount of data. As an answer to the analysis bottleneck, the cloud computing approach is an economic and scalable solution. We developed Eoulsan, a versatile framework based on the Hadoop implementation of the MapReduce algorithm, dedicated to high throughput sequencing data analysis on distributed computers. However cloud-computing solutions can be limited in term of data security and integrity. Moreover, some research groups may not have a cloud-computing resource available for their computation or may not be able to effort the cost of it. What we seek is to have a tool which is able to exploit a number of different computing resources, with minimal changes in the user interface, so that each research group may adopt the solution which best fits its needs and possibilities. The purpose of this work is to set up a Hadoop system on the EGI grid infrastructure in order to offer to the user a computer resource alternative for their analyses. Our results demonstrate that we were able to use Eoulsan and Hadoop over the grid infrastructure. We perform several tests in order to assess the performances of such a solution.

Enjeux scientifiques :

Ces dernières années, les techniques de séquençage à haut débit ont révolutionné la façon dont les chercheurs appréhendent les études de génomique et les applications possibles semblent sans limites. Ces méthodes ont évolué pour atteindre une meilleure précision, une plus grande facilité d'utilisation et des débits plus importants (1). Dans le même temps, la compétition entre les fournisseurs a permis une baisse des coûts significative. Aujourd'hui, la dernière génération d'appareil, de plus petite dimension, a pour but la mise à disposition des technologies de séquençage sur la paillasse des chercheurs. Toutes les conditions sont réunies pour que ces outils se disséminent largement à travers la communauté scientifique.
Cependant, il reste une limitation majeure qui ralentit ce processus d'adoption : l'étape d'analyse des données. Avec l'augmentation exponentielle de la quantité des données générées, la demande en ressources informatiques doit être capable de suivre ces évolutions (2). Les efforts pour remplir ces attentes sont très importants et visibles à travers le nombre important des solutions logicielles disponibles (3-4). Malgré tout, l'analyse des données nécessite des infrastructures informatiques de grande taille associées avec des coûts importants en terme de maintenance et de ressources humaines. Avec le nombre limité des personnels supports présents dans les laboratoires, il y a un besoin fort pour des ressources bioinformatiques en génomique mutualisées.
Avec les algorithmes de parallélisation, comme MapReduce, l'analyse des données, distribuée sur plusieurs machines fonctionnant en parallèle, peut être effectuée rapidement (5). Néanmoins, les logiciels qui utilisent la parallèlisation basée sur MapReduce ne sont pas très nombreux pour traiter du séquençage à haut débit (6-8). C'est pour cette raison que nous avons développé Eoulsan (9), un logiciel open-source pour analyser les données de séquençage grâce au calcul distribué, qui permet d'automatiser l'analyse d'un grand nombre d'échantillons à la fois, de simplifier la configuration d'une infrastructure de calcul distribué et de travailler avec de nombreuses solutions d'analyse disponibles dans la littérature. Cet outil fonctionne actuellement sur une architecture de cloud-computing au travers des solutions Amazon Web Services (AWS). Cependant ces solutions, si elles sont flexibles et

économiques peuvent poser des problèmes en terme de sécurité et de temps de transfert des données. C'est la raison pour laquelle nous avons travaillé à rendre Eoulsan utilisable sur les grilles de calcul EGI. La grille EGI est un réseau de ressources de calcul distribué sur des centaines de sites dans le monde, qui sont tous accessibles de façon uniforme et sécurisée par un inter-logiciel. Cette ressource a été déployée à l'origine pour soutenir l'activité informatique des expériences du Large Hadron Collider au CERN et elle est utilisée aujourd'hui par plusieurs communautés scientifiques.

Développements et utilisation des infrastructures :

Le premier objectif de ce projet était de définir et mettre au point un système basé sur Hadoop et MapReduce sur les grilles de calcul. Une fois cette infrastructure mise en place nous voulions dans un second temps faire fonctionner l'outil Eoulsan sur cette infrastructure. Les caractéristiques du système sont les suivantes : un nombre de machine nécessaire pour effectuer les calculs relativement faible (moins de 10), une disponibilité rapide des ressources (inférieure à une journée) et un temps de calcul limité (inférieur à une journée). La solution mise en place et présentée ici, est pour le moment limité à faire tourner chaque tache sur un seul site de la grille de calcul. Cependant les analyses effectuées par Eoulsan sur la grille sont exactement les mêmes que celles réalisées sur le Cloud d'Amazon.

Outils et difficultés rencontrées :

Parmi les différents problèmes rencontrés pour rendre Hadoop compatible avec les grilles de calcul, les deux difficultés principales sont la localisation/répartition du maître et des esclaves sur les différents nœuds de la grille, et la mise à disposition du système dédié de stockage de données HDFS pour Hadoop. Dans un premier temps, le système de démonstration étant localisé sur un site unique, nous avons choisi de lancer aussi bien la machine maître que les esclaves sur la grille. Cette façon de procéder est plus simple à implémenter et ne nécessite pas de ressources matérielles dédiées. La première tâche exécutée sur la grille lance le processus maître puis au fur et à mesure de l'exécution des tâches les machines esclaves sont recrutées et incluses dans le cluster Hadoop. La limitation principale de cette solution est qu'elle rendra plus difficile la distribution des tâches sur différents sites.
Concernant le stockage des données, plusieurs solutions sont disponibles pour rendre le HDFS accessible pour le cluster Hadoop : utiliser un cluster HDFS dédié, passer par un stockage POSIX/NFS pour lancer le cluster HDFS sur chaque nœud ou lancer un cluster HDFS en utilisant les disques locaux des nœuds. La première solution semble la plus performante et la plus flexible mais n'est pas standard sur les infrastructures grilles actuelles. La seconde offre le plus de compromis entre flexibilité et performance mais nécessite que tous les esclaves soient lancés avant de démarrer l'analyse. Enfin, si la dernière solution ne demande aucun stockage permanent particulier, elle impose également que tous les esclaves soient lancés avant de démarrer l'analyse et surtout elle peut poser des problèmes de performance et de limitation en espace disque.
Nous avons donc cherché à évaluer à l'aide de différents tests, les performances et le fonctionnement des différentes options possibles sur plusieurs jeux de données. Nous avons également comparé les résultats obtenus avec la grille de calcul avec ceux obtenus sur le cloud.

Résultats scientifiques :

Les premiers tests ont été effectué sur un jeu de données provenant d'expérience de séquençage de transcriptome entier (RNA-Seq) obtenu chez la souris et correspondant à 8 échantillons pour lesquels un total de 188 millions de lectures en Single Read de 76 bases de long, soit près de 14 Gb, a été obtenu. L'infrastructure de calcul grille a été mise à disposition par le site EGI GRIF_LLR. En particulier le site héberge un cluster dédié aux activités R&D qui a été utilisé pour les tests. Les données ont été d'abord téléchargées de l'IBENS vers le stockage du site GRIF_LLR, simplement par scp. Pendant les phases successives, upload de données et exécution de la tache, les transferts ont eu lieu exclusivement à l'intérieur du réseau local du site GRIF_LLR. Plusieurs nœuds identiques ont été réservés sur la grille pour effectuer les calculs, chaque machine possédant un processeur E5520 à 2,27 Ghz avec l'équivalent de 16 cœurs virtuels, 48 Gb de RAM, un disque dur de 250 Gb, une connexion Ethernet à 1 Gb/s et la version 6.3 64 bit de l'OS Linux Scientific. La version 1.0.4-1 du système Hadoop a été installée sur chaque nœud et en ce qui concerne le stockage nous avons choisi de comparer les solutions utilisant soit un système de stockage HDFS dédié soit celui basé sur POSIX/NFS. Les résultats sont présentés Figure 1.
Ces premiers résultats démontrent que le système Hadoop peut fonctionner sur l'infrastructure des grilles de calcul. Le choix du système de stockage n'a pas d'influence sur les résultats obtenus ce qui laisse plus de liberté en terme de sélection de la solution la plus flexible. En outre il est clair qu'héberger le processus maître sur le même nœud qu'un esclave ne ralenti pas les performances

donc cette solution est préférable comme elle permet d'économiser les machines mises à dispositions. Cette possibilité n'est pas offerte lorsque l'on utilise le système Hadoop sur l'infrastructure AWS. Pour finir l'analyse des résultats comparés à ceux obtenus lors des tests effectués chez Amazon (9) montrent de meilleures performances sur la grille. Cependant ces résultats sont à prendre avec précaution car les infrastructures ne sont pas totalement comparables.

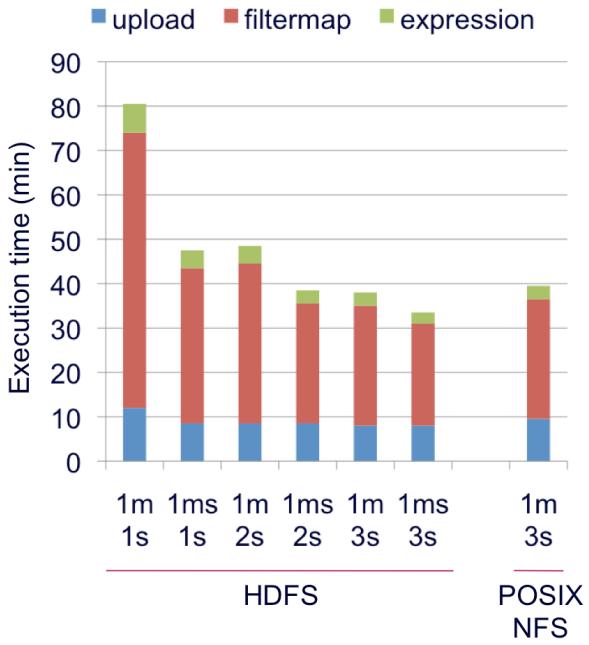

Figure 1 : Temps d'exécution d'Eoulsan en fonction du nombre de nœuds maître (m), esclaves (s) et maître/esclave (ms) réservé sur la grille. Le temps est présenté pour les 3 étapes de l'analyse à savoir le téléchargement des données sur le cluster (upload) depuis le stockage locale du site GRIF_LLR, l'alignement et le filtrage des lectures (filtermap) et la mesure d'abondance des transcrits (expression). La comparaison entre l'accès au stockage directement sur un système HDFS dédié ou au travers d'un montage POSIX/NFS est également indiquée.
Eoulsan execution time according to the number of master (m), slave (s) and master/slave (ms) nodes scheduled on the grid. Running time is shown for the three analysis steps: data upload from the GRIF_LLR local storage to the HDFS cluster, read mapping and filtering (filtermap) and transcript abundance estimation (expression). Comparison between direct accesses to dedicated HDFS storage or through the POSIX/NFS mounting process is also indicated.

Une seconde série de test a été effectuée avec un jeu de données correspondant à une plus grande quantité de résultats. Nous avons pour cela travaillé chez l'homme mais avec 6 échantillons séquencés en Paired-End de 100 bases de long. Les données contiennent 888 millions de lecture soit un peu plus de 88 Gb. Nous avons lancé les analyses sur la grille de calcul et en parallèle sur AWS en utilisant les instances m1.xlarge qui se rapprochent le plus de celles utilisées sur la grille. Pour ces tests nous avons utilisés 4 nœuds/instances. Les résultats sont présentés dans le tableau ci-dessous (Tableau 1).

	Upload	filtermap	expression	Total
Standalone	154	1,146	4	1,304
Grid	53	388	2.5	467
AWS	80	810	64	1,120

Tableau 1 : Comparaison des temps d'exécution (en minutes) d'Eoulsan sur la grille ou sur le cloud d'Amazon (AWS) pour chaque étape de l'analyse. Dans les 2 cas, 4 nœuds/instances sont réservées, un pour le maître et les 3 autres pour les esclaves. Pour la grille c'est le stockage HDFS direct qui est utilisé. Pour le test « standalone » un nœud de la grille du GRIF_LLR a été utilisé.
Comparison of Eoulsan running times (in minutes) between grid and Amazon cloud (AWS) for each analysis step. In both cases, 4 nodes/instances were scheduled, one for the master and 3 for the slaves. On the grid, we use direct HDFS storage. For the "standalone" test, one node from the GRIF_LLR grid cluster was used.

Ces résultats montrent clairement que les calculs sont effectués plus rapidement sur la grille que sur l'infrastructure de cloud-computing d'Amazon. Encore une fois, ces résultats sont à prendre avec précaution car la comparaison directe des configurations entre ces deux systèmes est difficile, de nombreux paramètres n'étant pas connus et maîtrisés chez AWS.

Perspectives :

Ces premiers résultats sont très encourageants et ouvrent de nouvelles perspectives dans la gestion des ressources de calcul dans le domaine de l'analyse des données de séquençage à haut débit. En effet le but final devrait être de permettre à l'utilisateur de ne pas se préoccuper de quel type de ressource de calcul il va utiliser pour effectuer son analyse. Mais en fonction de sa localisation, de ses partenariats ou de la disponibilité des ressources, de pouvoir lancer son calcul sur les grilles, dans le cloud ou sur n'importe quel serveur dédié localement ou de façon déportée. Cela passe aussi sans doute par la création de machines virtuelles génériques et universelles permettant de lancer Hadoop n'importe où.

Références :

1. Mardis, E.R. (2011) A decade's perspective on DNA sequencing technology. Nature, 470, 198-203.
2. Pennisi, E. (2011) Human genome 10th anniversary. Will computers crash genomics? Science, 331, 666-668.
3. Pepke, S., Wold, B. and Mortazavi, A. (2009) Computation for ChIP-seq and RNA-seq studies. Nat Methods, 6, S22-32.
4. Trapnell, C. and Salzberg, S.L. (2009) How to map billions of short reads onto genomes. Nat Biotechnol, 27, 455-457.
5. Schatz, M.C., Langmead, B. and Salzberg, S.L. (2010) Cloud computing and the DNA data race. Nat Biotechnol, 28, 691-693.
6. A. McKenna, M. Hanna, E. Banks et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20 (9), 1297 (2010).
7. B. Langmead, K. D. Hansen, and J. T. Leek (2010). Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biol 11 (8), R83 (2010).
8. B. Langmead, M. C. Schatz, J. Lin et al. (2009). Searching for SNPs with cloud computing. Genome Biol 10 (11), R134 (2009).
9. Jourdren, L., Bernard, M., Dillies, M.A. and Le Crom, S. (2012) Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses. Bioinformatics, 28, 1542-1543.