Hadoopizer : a cloud environment for bio-informatics data analysis

Anthony Bretaudeau (1), Olivier Sallou (2), Olivier Collin (3)
(1) anthony.bretaudeau@irisa.fr, INRIA/Irisa, Campus de Beaulieu, 35042, RENNES, Cedex, France
(2) olivier.sallou@irisa.fr, INRIA/Irisa, Campus de Beaulieu, 35042, RENNES, Cedex, France
(3) olivier.collin@irisa.fr, INRIA/Irisa, Campus de Beaulieu, 35042, RENNES, Cedex, France

Overview:

Biology is evolving into a big data science, particularly with the new sequencing technologies which have emerged during the last years. Cloud computing appears as one of the answers to face the rapidly increasing volume of bioinformatics data.
Here we present a private cloud environment deployed on the GenOuest bioinformatics platform. After an overview of the software publicly available for bioinformatics treatments in the cloud, we present a new framework (Hadoopizer) which is a generic tool for the parallelisation of bioinformatics analysis in the cloud using the MapReduce paradigm.
These developments are available online at this address: http://genocloud.genouest.org

Contexte :

L'émergence des nouvelles technologies de séquençage a transformé de manière radicale la Biologie qui est devenue une science générant des masses de données. Le changement d'échelle induit par cette génération massive de données nécessite de trouver de nouveaux modes de calcul, que ce soit pour les ressources mais également pour la façon de distribuer ces calculs. La démocratisation des appareils de séquençage dans les laboratoires complique la situation car tous les laboratoires ne disposent pas forcément d'une infrastructure adaptée pour le calcul. L'utilisation d'une ressource de type cloud apparaît donc comme une alternative intéressante pour accompagner l'évolution des laboratoires [1].

Les plates-formes bio-informatiques sont concernées au premier chef par l'évolution des techniques en Biologie et sont impliquées dans des réflexions sur ces nouveaux modes de calcul. La plate-forme bio-informatique GenOuest a lancé plusieurs projets afin d'obtenir un retour d'expérience sur l'apport des technologies du cloud et de l'utilisation de Map-Reduce pour l'analyse des données de séquençage. Le premier projet consiste en la mise en place d'un cloud privé basé sur OpenNebula. Le second projet repose sur l'exploration de Map-Reduce pour l'exploitation des séquences. A cet effet, un environnement de déploiement automatique Hadoop a été développé.

GenoCloud :

GenOuest a déployé un cloud de 120 cœurs, 470 Go de mémoire et 8 To de stockage pour proposer un environnement de démonstration et d'expérimentation pour la communauté. Cet environnement est disponible à l'adresse http://genocloud.genouest.org/.

Basé sur OpenNebula, ce cloud propose des images pré-configurées (réseau, accès par clé SSH, …), un proxy pour l'accès à partir de l'extérieur vers les applications web, ainsi que des outils de déploiement automatique d'environnement de calculs. Cela inclut pour le moment la possibilité de déployer un cluster Grid Engine ou un cluster Hadoop dynamiquement en sélectionnant le nombre de nœuds voulus et la possibilité de l'étendre, le tout s'effectuant via une interface graphique pour en faciliter l'usage. L'interface offre également un monitoring des instances ainsi créées.

Ces images ont un but pédagogique, afin que les utilisateurs puissent tester simplement et efficacement leurs outils sans s'occuper de l'implémentation des couches techniques.

Un outil de workflow extensible développé par GenOuest (Manband) sera bientôt implémenté de la même façon dans ces images.

Un outil de messaging (RabbitMQ) permet aux applications de communiquer de façon robuste entre elles sans connaître la configuration de leurs voisines.

Ce cloud propose également aux utilisateurs un accès disque partagé entre toutes les VM, allant au delà des solutions type EBS ou S3, car plus souple pour l'utilisateur habitué à des accès type NFS entre serveurs. Ceci facilite la transition ou les tests d'applications existantes. Chaque instance possède un montage automatique sur ce disque vers un répertoire personnel.

Enfin, il supporte l'interface EC2 et propose un stockage compatible S3 pour permettre un passage simple vers un prestataire privé.

Hadoopizer :

Un outil exploitant le framework Hadoop est en cours de développement afin d'évaluer l'apport de l'utilisation de MapReduce pour le traitement de données bio-informatiques. Cet outil, baptisé Hadoopizer, constitue un environnement générique pour l'utilisation de Hadoop en bio-informatique. L'outil est capable de prendre une ligne de commande traditionnelle et de lancer son exécution de manière distribuée sur des quantités importantes de données, en utilisant un cluster Hadoop déployé de façon automatique sur l'environnement de cloud de la plate-forme GenOuest.

Hadoopizer propose un parallélisme à gros grain, concernant des traitements qualifiés d'*embarrassingly parallel*, c'est-à-dire où chaque sous-tâche est exécutée de façon indépendante des autres sous-tâches.

D'autres outils bio-informatiques utilisant Hadoop sont disponibles [2,3,4,5,6,7] (tableau 1). Cependant, chacun de ces outils ne permet d'utiliser qu'un algorithme précis (assemblage, mapping). Implémenter d'autres algorithmes dans ce cadre impose des développements importants.

Le logiciel Eoulsan [8] est un peu plus polyvalent puisqu'il propose de base plusieurs outils (tableau 1). Il offre aussi la possibilité d'intégrer d'autres outils grâce à une API Java, mais ceci nécessite un temps de développement non négligeable.

Outil	Application	Évolutivité	Programmes interfacés
CloudAligner [3]	Mapping	+	Algorithme spécifique
CloudBurst [4]	Mapping	+	RMAP
Contrail [5]	Assemblage	+	Algorithme spécifique
Crossbow [6]	Détection de SNPs	+	Bowtie + SOAPsnp
Myrna [7]	Expression différentielle	+	Bowtie + R
Eoulsan [8]	Expression différentielle	++	Bowtie, BWA, SOAP2, Gsnap, Gmap, R

Tableau 1 : quelques outils bio-informatiques utilisant le cloud computing

Il nous a semblé important de disposer d'un outil beaucoup plus polyvalent. La caractéristique principale de Hadoopizer est d'être générique en permettant l'exécution de n'importe quelle ligne de commande shell sur un cluster Hadoop. Il propose un framework qui permet d'encapsuler tout le programme, en prenant en charge la parallélisation des exécutions. Ceci apporte une grande souplesse d'utilisation, l'exploitation d'un nouvel algorithme ne nécessitant aucun développement spécifique. Ceci constitue un avantage important dans le domaine de la bio-informatique qui est caractérisé par une durée de vie relativement courte des logiciels d'analyse.

Hadoopizer prend en charge différents formats de données couramment utilisés : fichiers de séquence fasta ou fastq ou encore alignements au format sam ou bam. Ces formats peuvent être utilisés comme données de sortie ou d'entrée ; dans ce cas, elles sont alors découpées et envoyées aux différents nœuds de calculs. Il est aussi possible d'utiliser des données statiques, qui doivent être entièrement accessibles en lecture sur chaque nœud (comme un génome de référence par exemple) : Hadoopizer prend alors en charge les transferts nécessaires pour que ces données soient accessibles sur tous les nœuds de calcul.

Les transferts des données d'entrée et de sortie sont gérés par Hadoopizer à travers plusieurs protocoles (local, HDFS, S3). La compression de ces données est aussi supportée dans différents formats (gzip, bzip2).

Il est possible de déployer automatiquement les binaires nécessaires à l'exécution sur tous les nœuds de calculs. Les binaires doivent être fournis sous forme d'une archive qui est envoyée puis décompressée automatiquement sur chaque nœud de calcul au début de l'exécution.

Enfin dans le cas d'un enchaînement de plusieurs outils, les données de sortie peuvent être enregistrées dans un fichier au format optimisé et placé sur le système de fichier HDFS du cluster Hadoop. Ceci permet de maximiser les performances de lecture pour une réutilisation comme données d'entrée d'un nouveau calcul.

Résultats scientifiques :

La première application sur laquelle a été testé cet environnement correspond au mapping de séquences sur un génome de référence. Ce type de traitement est très utilisé pour l'analyse de données de séquençage a haut débit. Dans ce cas

d'utilisation, on dispose d'un nombre important (quelques dizaines de Go) de séquences de faible taille (appelés 'reads', typiquement 100 caractères). On souhaite alors positionner chacun de ces reads sur un génome de référence disponible sous la forme d'une très longue séquence (de quelques Mo à quelques Go). Cet exemple est facilement parallélisable puisque chaque read peut être positionné indépendamment des autres reads. De nombreux algorithmes sont disponibles pour effectuer cette étape de mapping, chacun ayant des spécificités (tolérance aux erreurs dans les reads, support des gaps).

Des analyses de mapping sur des jeux de données réels ont été menées en utilisant Hadoopizer sur le cloud privé de GenOuest. Deux algorithmes de mapping ont été testés : bowtie [9] et bwa [10]. Les données d'entrées correspondaient à 11Go de reads à aligner sur un génome de référence de 400Mo. Le lancement de ces analyses avec Hadoopizer comporte plusieurs étapes : (1) Hadoopizer lit un fichier xml contenant la ligne de commande à exécuter et la localisation des données (figure 1) ; (2) les binaires et le génome de référence sont copiés sur chacun des nœuds de calcul ; (3) les reads sont répartis sur les nœuds de calcul par paquets (343 paquets pour 11Go de reads) ; (4) chaque nœud exécute la ligne de commande sur chaque paquet de read reçu (étape de mapping) ; (5) les fichiers de résultats générés pour chaque paquet de reads sur chaque nœud sont regroupés dans un même fichier final de résultat (étape de reducing).

Ces premiers tests ont montrés que l'approche proposée permet bien de mettre en œuvre une parallélisation à gros grains sur ce type de données, en se focalisant sur le comportement des programmes vis à vis des données. Par ailleurs, Hadoopizer étant un outil générique, il a été possible de tester différents algorithmes de mapping dans un temps réduit.

```
<?xml version="1.0" encoding="utf-8"?>
<job>
    <command>
        bowtie -m 1 --best --strata -S ${genome} ${reads} > ${mapped}
    </command>

    <input id="reads" split="true">
        <url splitter="fastq">/home/example/reads.fastq</url>
    </input>

    <input id="genome">
        <url autocomplete="true">/home/example/indexed_genome</url>
    </input>

    <outputs>
        <url>/home/example/output_mapping/</url>
        <output id="mapped" reducer="sam" />
    </outputs>
</job>

Executed with the following command line:

hadoop jar hadoopizer.jar -b bowtie_bin.tar.gz -c job_config.xml -w hdfs://192.168.2.20/bowtie_tmp/
```

Figure 1 : Exemple de fichier xml pour un mapping utilisant le logiciel bowtie.
Un fichier fastq est automatiquement découpé, le génome de référence est rendu accessible automatiquement sur chaque nœud de calcul, le logiciel a exécuter (bowtie) est déployé par Hadoopizer et les données de sortie sont enregistrées au format SAM.

Perspectives :

Plusieurs évolutions sont envisagées concernant l'outil Hadoopizer. En particulier, le développement de modules permettant la lecture et l'écriture d'autres formats de données couramment utilisés en bio-informatique (bed, wiggle, gff, blast, ...) sera nécessaire. Des modules plus spécialisés pourront aussi être développés pour prendre en charge des modes de parallélisme spécifiques : fenêtres glissantes et chevauchantes sur des séquences ou encore support de données de séquençage pairées.

Il est aussi envisagé d'intégrer Hadoopizer à un outil de workflow (slicee [11]). Dans ce cadre, l'utilisateur bio-informaticien pourra ainsi concevoir des workflows pouvant exploiter de façon transparente plusieurs environnements de parallélisme selon les ressources de calcul disponibles : cluster SGE ou cluster Hadoop.

En complément à ces développements, ce projet a pour but d'évaluer l'intérêt d'une approche basée sur le cloud en bio-informatique. Des benchmarks seront ainsi réalisés afin de pouvoir comparer de façon objective les performances offertes par cet environnement.

Références :

[1] Fox. Computer science. Cloud computing--what's in it for me as a scientist?. Science (2011) vol. 331 (6016) pp. 406-7

[2] Taylor. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC Bioinformatics (2010) vol. 11 Suppl 12 pp. S1

[3] Nguyen et al. CloudAligner: A fast and full-featured MapReduce based tool for sequence mapping. BMC Res Notes (2011) vol. 4 pp. 171

[4] Schatz. CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics (2009) vol. 25 (11) pp. 1363-9

[5] Contrail http://sourceforge.net/apps/mediawiki/contrail-bio/index.php?title=Contrail

[6] Langmead et al. Searching for SNPs with cloud computing. Genome Biol (2009) vol. 10 (11) pp. R134

[7] Langmead et al. Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biology (2010) vol. 11 (8) pp. R83

[8] Jourdren et al. Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses. Bioinformatics (2012) vol. 28 (11) pp. 1542-3

[9] Langmead et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology (2009) vol. 10 (3) pp. R25

[10] Li et Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (2009) vol. 25 (14) pp. 1754-60

[11] Piat et al. SLICEE: A Service oriented middleware for intensive scientific computation. 7th IEEE 2011 World Congress on Services (SERVICES 2011), Washington DC, USA