Du phénotypage au Big Data

Alexandre Journaux[1], Jonathan Mineau[2], Vincent Negre[2]

Résumé. Avec l'émergence des nouvelles technologies d'acquisition de données, les solutions actuelles de stockage et d'interrogation vont très vite atteindre leurs limites. Les Catis (centres automatisés de traitement de l'information) Sicpa (systèmes d'informations et calcul pour le phénotypage animal) et Codex (connaissances et données expérimentales) s'intéressent donc aux différentes technologies autour du Big Data.

Mots clés : phénotypage, haut débit, Big Data, NoSQL

Introduction

Les missions des Catis (centres automatisés de traitement de l'information) Sicpa (systèmes d'informations et calcul pour le phénotypage animal) et Codex (connaissances et données expérimentales) sont de mettre en place des outils pour l'intégration de données de phénotypage (de l'acquisition au traitement statistique) et d'en assurer leur valorisation. Ces outils permettent la mesure de différents caractères phénotypiques, que ce soit dans le domaine animal (ex : comportement alimentaire, social, sexuel, réaction au stress) ou végétal (ex : mise en place d'organes, évolution de la biomasse, impact du stress). Les axes de développements sont toujours orientés pour favoriser l'ergonomie de travail et l'amélioration de la qualité/traçabilité des données. Pour des activités dites de terrains ou manuelles, ces outils sont une aide pour les utilisateurs et facilitent la **saisie** (application déployée sur boîtiers PDA (personal digital assistant), automate de pesées, tablettes, etc.), l'**enregistrement** (lecture de puce RFID (radio frequency identification), code-barre, QrCode, etc.), la **fiabilisation** (récupération de poids par automate de pesée, etc.), le **contrôle** (vérification de dates, de seuil, etc.), la **valorisation** (calcul de gain de poids ou biomasse moyen quotidien, etc.).

Depuis quelques années, les Catis doivent faire face à de nouvelles demandes utilisateurs. Les avancées technologiques dans des domaines aussi variés que l'informatique, l'électronique ou la physique appliquée impactent directement nos dispositifs expérimentaux. Nous voyons l'émergence de nano-ordinateurs nomades ou de capteurs embarqués, miniaturisés, autonomes, géo localisés et dont l'enregistrement et la sauvegarde des données produites massivement sont assurés par des capacités de stockages de plus en plus grandes. Désormais nous pouvons assurer l'enregistrement automatique de mesures standards, fiables, répétées, fréquentes et à haut-débit. Dans ce contexte, l'opportunité s'offre aux équipes de développements de créer de nouveaux outils de collectes pour répondre aux nouveaux sujets de recherche.

Vers un phénotypage haut débit

Au niveau du Cati Sicpa, nous avons mis en place plusieurs protocoles :

- ✓ étude du comportement alimentaire du canard avec la mise en place de distributeurs automatisés de concentrés (DAC). Ces appareils enregistrent, de manière continue, le poids de la mangeoire et le poids de l'animal identifié à l'entrée du DAC ;
- ✓ étude du comportement de la truie par rapport à ses porcelets avec la mise en place d'un accéléromètre sur la mère. Ces appareils enregistrent, de manière continue, la position de la mère ;
- ✓ étude de l'activité des déplacements de la reine dans la ruche avec la mise en place d'une puce RFID sur la reine et d'antennes sur les cadres de la ruche. Ces appareils enregistrent, de manière continue, la localisation de la reine dans la ruche ;

1 *UMR GenPhySE, Inra, 31326 Castanet-Tolosan Cedex, France*
2 *UMR LEPSE, Inra, Campus Supagro Montpellier, 2 place Viala, 34060 Montpellier Cedex 2*
alexandre.journaux@inra.fr

- ✓ étude du comportement social de la vache dans l'étable avec la mise en place de capteurs et de radars dans les bâtiments d'élevage. Ces appareils enregistrent, de manière continue, la localisation de chaque vache dans l'étable,
- ✓ etc.

Au niveau du Cati Codex le projet Silex (système d'information pour l'expérimentation) est un projet collaboratif dont l'objectif est de fournir des briques logicielles facilitant la mise en place de systèmes de gestion des données ou des connaissances. Il aborde plusieurs aspects : mesures en-ligne, mesures hors-ligne, architecture, modèles de données et de connaissances, accès web et traitements statistiques.

Tous ces outils collectent des masses très importantes de données à des fréquences très rapides. Ils permettent donc de faire du phénotypage à haut débit. Cependant, nos outils de stockage et d'interrogation des données restent, à ce jour, les mêmes que pour des données classiques : bases de données relationnelles (MySQL ou Oracle), requêtages par des scripts SQL, SAS ou R. Avec ces solutions nous rencontrons des problèmes de performances pour l'interrogation et, à terme, nous aurons aussi des problèmes d'optimisation du stockage.

Afin d'anticiper et de résoudre ces problèmes, les Catis Sicpa et Codex s'intéressent, désormais, aux problématiques du Big Data et à ses solutions autour des technologies NoSQL (not only structured query language).

Le Big Data

Définition

On parle de Big Data quand des ensembles de données deviennent si grands et si complexes qu'ils deviennent difficiles voire impossibles à traiter en utilisant les méthodes et les outils traditionnels.

Le Big Data se définit souvent à partir de trois notions appelées les 3 V :

- ✓ volume : le volume de données déjà important augmente en permanence. Ce qui peut générer des difficultés de stockage et d'analyse ;
- ✓ variété : les données proviennent de différentes sources, disciplines, formats… Ce qui peut générer des difficultés de compréhension et d'intégration ;
- ✓ vitesse : des volumes importants de données sont collectés de manière très fréquente (cela peut être en temps réel). Ce qui peut générer des difficultés de traitement de ces données.

Un ensemble de données caractérisé par ces notions doit donc être géré à partir d'outils dédiés au Big Data.

Les outils pour gérer du Big Data

Plusieurs outils existent désormais pour le traitement et l'analyse des données Big Data. Le Cati Sicpa s'intéresse essentiellement à deux technologies : les architectures MapReduce et les bases de données NoSQL.

MapReduce est un patron d'architecture qui permet de manipuler des grandes quantités de données en les distribuant sur plusieurs machines.

Le terme de NoSQL regroupe des systèmes de bases de données qui ne sont pas relationnels. Ces bases de données proposent une alternative aux bases de données classiques pour le traitement de gros volumes de données. Elles ne présentent pas les propriétés d'une base de données relationnelle (propriétés ACID (atomicité, cohérence, isolation et durabilité)) et ne sont pas, en général, interrogeables avec le langage SQL. Elles ont été conçues pour gérer des volumes de données importants (elles possèdent pour cela des

propriétés de « scalabilité horizontale » car les données peuvent être réparties sur plusieurs bases) et des données non structurées (leur modèle de données est plus souple que celui des SGBDRs (systèmes de gestion de bases de données relationnels)). Les bases de données NoSQL peuvent être classés en quatre catégories : les bases de données de type clé-valeur (Dynamo, Reddis, Voldemort…), les bases de données de type documents (MongoDB, CouchDb…), les bases de données de type colonnes (Hbase, Cassandra, BigTable…) et les bases de données de type graphes (Neo4J, InfoGrid…).

Les collaborations et les solutions expérimentées

Grâce aux crédits incitatifs du Département Génétique Animale, nous avons pu acheter un serveur qui nous permet, dans un premier temps, d'expérimenter les différentes solutions techniques puis qui nous permettra de mettre en production la solution retenue.

En collaboration avec l'Irstea, nous avons testé Cassandra (NoSQL) et Hadoop (MapReduce). Les premiers résultats prometteurs nous conduisent à poursuivre les recherches en participant au dépôt d'un projet FUI (fonds unique interministériel). Le but de ce projet est d'automatiser le passage d'un modèle relationnel de données vers un modèle optimisé pour le Big Data.

En collaboration avec le Cati Codex, nous avons testé MongoDB (NoSQL).

Témoignage et retour d'expérience de MongoDB au Cati Codex

Nous avons opté pour MongoDB, seule technologie qui semblait en 2012 être assez mûre avec une communauté relativement développée. Ce choix s'est finalement avéré judicieux au regard de l'essor de MongoDB dans la communauté informatique internationale. Nous avons confronté en test final les performances de MongoDB au SGBDR PostgreSQL avec son module NoSql. Nous avons généré un jeu de données équivalant à 5 années de stockage. Nous avons ensuite injecté un ensemble de requêtes complexes correspondant aux requêtes utilisateurs les plus courantes. Les résultats obtenus ont consolidé et entériné notre choix sur l'utilisation de MongoDB.

Dans l'approche NoSQL la conception de la base de données doit suivre une approche déstructurée (c'est à dire qu'on ne respecte pas les formes normales des bases de données relationnelles en acceptant de la redondance d'informations dans les documents). La puissance de MongoDB venant en parti de ses moteurs, nous avons pu vérifier que la bonne utilisation des index et des bons moteurs de stockage décuple grandement les performances (les moteurs sont interchangeables dans MongoDB, mais le dernier - « WiredTiger » - se révèle très performant. Pour que cette redondance ne soit pas au final limitante et que les index soient pertinents, une connaissance parfaite du schéma relationnel d'origine et des requêtes utilisateurs est nécessaire.

Enfin, il est à noter que ce qui fait aussi le succès de MongoDB c'est l'architecture des serveurs stockant les documents. Cette configuration garantit la résistance aux pannes (grâce au mécanisme de replicat-set) et optimise la répartition de la charge (grâce au mécanisme de sharding). Cependant, elle est complexe car les données sont stockées sur plusieurs serveurs pouvant être répartis sur différents sites. Cette complexité demande non seulement des compétences humaines pour le déploiement et la maintenance mais aussi des infrastructures adaptées (datacenters, une bonne liaison réseau entre les datacenters) et un financement à inclure avant d'utiliser cette technologie dans un projet.

Conclusion

Grâce aux nouvelles technologies proposées autour du Big Data, les Catis Sicpa et Codex vont donc pouvoir développer des outils pour la gestion de très gros volumes de données. Ces outils permettront un stockage optimisé des données collectées dans les Unités et Installations Expérimentales mais aussi, pour les scientifiques, des solutions d'interrogations fiables et rapides.