meRoo : Un écosystème logiciel pour l'apprentissage des sciences des données installé sur un cluster de Raspberry Pi.

Frédéric BLANCHARD 1* Guillaume DOLLÉ 2† Philippe REGNAULT 3‡

Résumé

Les projets d'ingénierie reposant sur une volumétrie importante de données nécessitent une collaboration maîtrisée entre les différents acteurs impliqués. Cette collaboration est caractérisée par des flux importants d'informations (données, code, documents, etc) produites et échangées par divers outils, logiciels et langages. La maîtrise de la gestion de ces flux, communément désignée par « *workflow* » est un enjeu essentiel dans la formation des ingénieurs en sciences des données.

La mise en place de tels workflows nécessite une infrastructure matérielle et logicielle qui peut ne pas être facile à déployer dans le cadre de formations académiques. C'est une difficulté à laquelle nous avons été confrontés dans le cadre de nos enseignements au sein du master *Statistique et Évaluation pour la Prévision* de l'Université de Reims Champagne-Ardenne (URCA). Nous avons pris le parti d'y répondre en privilégiant une approche économique, frugale, accessible et... fun : déployer un ensemble de ressources libres sur un cluster de nano-ordinateurs Raspberry Pi, mises en musique par RStudio Server !

Notre « super-nano-calculateur » **meRoo**, anagramme de *Romeo*, nom du super-calculateur de l'URCA, est utilisé pour les enseignements depuis quelques mois. Requêtage de bases SQL, noSQL, développement et hébergements d'applications Shiny, calcul parallèle avec R ou python, calcul distribué avec Hadoop/Spark, intégration continue avec GitLab, les possibilités sont étonnantes. Retours sur cette expérience pédagogique enthousiasmante !

Mots-clefs : Enseignement, data science, workflow, cluster, Raspberry Pi.

Développement

Le cluster **meRoo** a vocation à illustrer le fonctionnement d'un *workflow* de sciences de données, en proposant un ecosystème de ressources logicielles libres :

- des systèmes de gestion de base de données (SGBD) relationnel (PostgreSQL) ou noSQL (MongoDB) ;
- un *framework* de stockage et calcul distribués (*Hadoop*/*Spark*) ;
- des langages de programmation haut-niveau avec des bibliothèque dédiées aux sciences des données (R (essentiellement) et Python (un peu)) ;
- un système de gestion de versions (Git) et un service centralisé associé (GitLab) ;
- un éditeur de supports de communication puissant (R Markdown/Quarto) ;
- un serveur Shiny ;
- un environnement de développement intégré (EDI) puissant et interfaçable avec les ressources précédentes (RStudio Server).

Cet écosystème logiciel permet de montrer l'ensemble des étapes de la création d'un dashboard dynamique servi par Shiny Server, du requêtage des données nécessaires à l'intégration continue de son code source, en passant par le développement collaboratif de ce code et la parallélisation de certaines procédures.

*CReSTIC, Université de Reims Champagne-Ardenne, frederic.blanchard@univ-reims.fr

†Laboratoire de Mathématiques de Reims UMR CNRS 9008, Université de Reims Champagne-Ardenne, guillaume.dolle@univ-reims.fr

‡Laboratoire de Mathématiques de Reims UMR CNRS 9008, Université de Reims Champagne-Ardenne, philippe.regnault@univ-reims.fr

Inspirés par de nombreux tutoriels mettant en avant l'utilisation de nano-ordinateurs Raspberry Pi (RPi) dans des projets de data science ou d'intelligence artificielle (voir par exemple Giger, Srikugan, and Persaud (2020), Castro Socolich (2021), Evans (2020)), nous avons pris le parti de déployer cet écosystème sur un cluster de 4 RPis 4B (le modèle le plus évolué). Cette architecture offre des performances certes limitées ; mais aussi et surtout la possibilité d'illustrer des mécanismes complexes pour des coûts financier et énergétique dérisoires (environ 500 €, puissance voisine de 30 W). Elle offre également un avantage pédagogique à ne pas sous-estimer : la matérialité du système d'information utilisé : ça clignote, ça ventile et ... ça capte l'attention et fait sourire ! :-)

Références

Castro Socolich, Andrés. 2021. "Automaticcaly Installing Shiny and RStudio Server on Raspberry Pi OS with Ansible. Andres' Blog." January 13, 2021. https://andresrcs.rbind.io/2021/01/13/raspberry_pi_server/.

Evans, PJ. 2020. "Build a Raspberry Pi Cluster Computer." *MagPi*. https://magpi.raspberrypi.com/articles/build-a-raspberry-pi-cluster-computer.

Giger, Peter, Sajan Srikugan, and Badrie Persaud. 2020. "A Raspberry Pi Cluster for Teaching Big-Data Analytics." Master Project. Universistät Zürich. https://www.ifi.uzh.ch/dam/jcr:9c6065e2-10aa-442b-915c-57246020c23c/ReportMScProjektGigerSrikuganPersaud.pdf.