Expérimentation d'une architecture de calcul distribué « Code over Data » au sein d'une infrastructure de recherche

Sujet pour le Tremplin Recherche d'ESIEE Paris, 2025–2026

30 septembre 2025

Laboratoire et équipe

Le projet se déroulera au sein de la plateforme Cortext – l'une des plateformes scientifiques de l'UGE – spécialisée dans l'analyse de données pour les sciences sociales, en particulier l'analyse de documents à travers de réseaux socio-sémantiques, cartographies géo-spatiales et autres approches algorithmiques. Cortext est hébergée par le Laboratoire Interdisciplinaire Sciences Innovations Sociétés (LISIS), une unité mixte de recherche entre UGE, CNRS et INRAE.

Tuteur

Le tuteur du projet sera le Dr. Alexandre Hannud Abdo, Ingénieur de recherche au CNRS et coordinateur scientifique de la plateforme Cortext.

Adresse électronique : alexandre.hannud-abdo@univ-eiffel.fr

Filière visée

Informatique ; Datascience et intelligence artificielle ; Artificial intelligence and cybersecurity.

Ouvert à des étudiants en E4 (nov-avr) avec possibilité de poursuite par un stage (mai-aoû), ou en E5 (nov-jan) avec un stage de fin d'études dans le laboratoire (fév-jul). Le sujet sera adapté selon le niveau.

Présentation générale

Les analyses de gros ensembles de documents et traces numériques assistées par méthodes computationnelles sont une démarche de plus en plus présente dans les recherches en sciences sociales. Davantage qu'un outil, il s'agit d'une nouvelle façon de mobiliser les concepts clés de ses disciplines pour mieux transiter entre les échelles micro et macro qui caractérisent les systèmes sociaux. La plateforme Cortext est l'une des rares infrastructures dédiées à mettre des instruments computationnels et statistiques de pointe à portée de main et au service des compétences et connaissances de chercheurs dont le métier est ancré, loin des méthodes quantitatives, dans des liens profonds avec leurs terrains à travers d'observations, ethnographies, études de cas, et participation dans la vie sociale.

Objectif du projet

Il s'agit de participer à la conception et expérimentation d'une nouvelle infrastructure de gestion de ressources computationnels pour l'application web d'analyse de données Cortext Manager[1]. L'actuelle infrastructure a exécuté, depuis ses débuts, un demi-million d'analyses lancées par environ dix mil utilisateurs dans une centaine de pays, mais atteint des limites conceptuels et matériels face aux demandes de nouveaux méthodes d'analyse et données plus exigeantes.

Afin de livrer une proposition d'architecture ou même un prototype pour l'avenir des services fournis par la plateforme Cortext, l'étudiant travaillera en réalisant des nouveaux développements et des tests, sur la base d'une première expérimentation déjà conduite en employant une approche de systèmes distribuées du type « Compute over Data » avec le logiciel Bacalhau[2].

Les buts de cette architecture incluent :

— La distribution de tâches à travers un réseau de multiples nœuds, dynamiques et même distants, à des fins de répartition de charge (*load balancing*) comme également pour permettre l'ajout *ad hoc* de capacité de calcul y compris par les utilisateurs mêmes – de leurs propres

ordinateurs ou via renseignement d'une clé d'accès pour des infrastructures tierces de *cloud computing*). L'orchestrateur de Bacalhau semble compatible avec nos demandes, mais une précédent de référence est le distributeur du projet BOINC[3].

— L'isolement de tâches dans des conteneurs Docker[4] permettant leur exécution dans des ordinateurs quelconques où l'on fait tourner le *daemon* Bacalhau. Par rapport au système de *runtime* du projet BOINC[5], l'architecture de Bacalhau en se basant sur Docker offre un socle d'exécution plus générique ;

— L'intégration sécurisé du réseau de nœuds avec le stockage de données de l'application Cortext Manager et les concepts de « projet » et d'« analyse » pour le contrôle d'accès ;

— La traçabilité des tâches et de la recette des images de conteneurs afin d'améliorer leur reproductibilité[6] ;

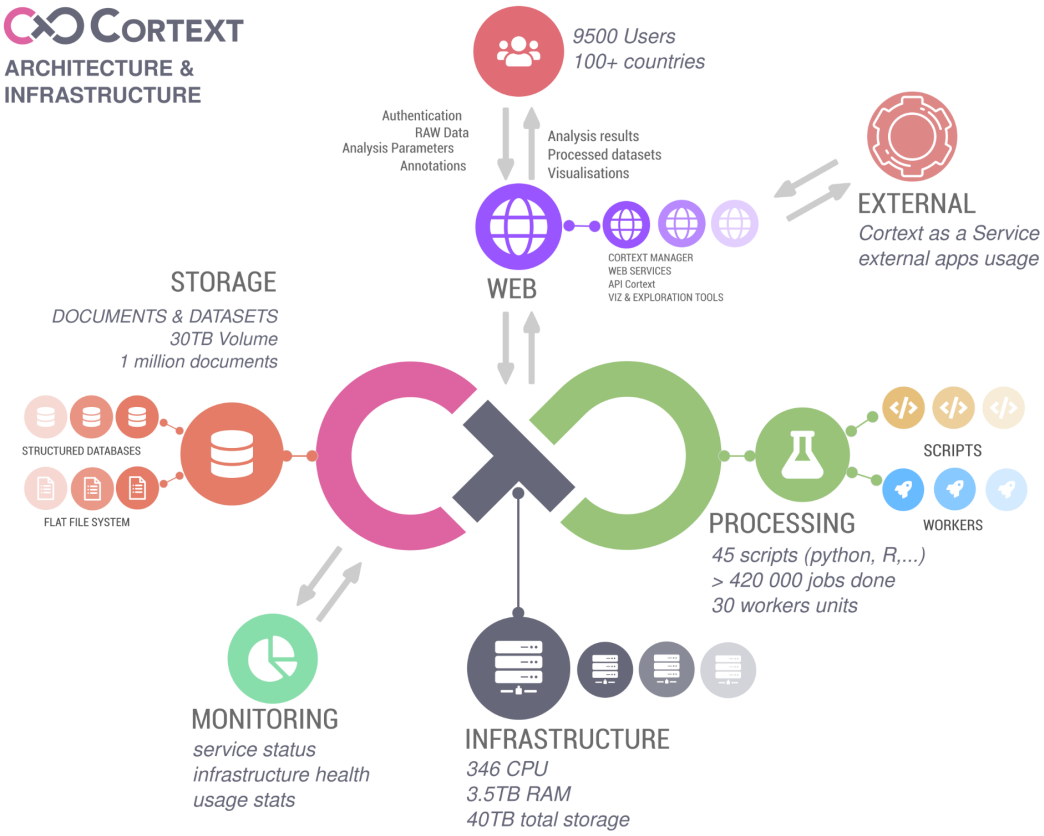

FIGURE 1 – Diagramme du fonctionnement des services autour de l'application Cortext Manager.

A travers son travail de recherche l'étudiant développera également une compréhension des systèmes distribuées, d'enjeux de sécurité et reproductibilité, et des compétences dans l'utilisation de systèmes de contrôle de version et de conteneurs Linux, dans la mise en œuvre d'un réseau de conteneurs pour le prototypage, et dans la spécification et documentation d'une architecture et d'un plan de transition d'une infrastructure existante ; son travail contribuera à l'avenir d'une plateforme de l'université utilisée à l'échelle internationale.

Bibliographie

[1] P. BREUCKER et al., *Cortext Manager*, M. BARBIER et L. VILLARD, éd., version v2, 28 oct. 2016. adresse : https://docs.cortext.net.

[2] *Distributed Compute Over Data | Bacalhau*, en. visité le 30 sept. 2025. adresse : https://bacalhau.org/.

[3] D. Anderson, E. Korpela et R. Walton, « High-Performance Task Distribution for Volunteer Computing, » in *First International Conference on e-Science and Grid Computing (e-Science'05)*, Pittsburg, PA, USA : IEEE, 2005, p. 196-203, ISBN : 9780769524481. DOI : 10.1109/E-SCIENCE.2005.51. visité le 30 sept. 2025. adresse : http://ieeexplore.ieee.org/document/1572226/.

[4] C. Boettiger, « An introduction to Docker for reproducible research, » en, *ACM SIGOPS Operating Systems Review*, t. 49, n° 1, p. 71-79, jan. 2015, ISSN : 0163-5980. DOI : 10.1145/2723872.2723882. visité le 30 sept. 2025. adresse : https://dl.acm.org/doi/10.1145/2723872.2723882.

[5] D. P. Anderson, C. Christensen et B. Allen, « Grid resource management—Designing a runtime system for volunteer computing, » en, in *Proceedings of the 2006 ACM/IEEE conference on Supercomputing - SC '06*, Tampa, Florida : ACM Press, 2006, p. 126, ISBN : 9780769527000. DOI : 10.1145/1188455.1188586. visité le 30 sept. 2025. adresse : http://portal.acm.org/citation.cfm?doid=1188455.1188586.

[6] N. Vallet, D. Michonneau et S. Tournier, « Toward practical transparent verifiable and long-term reproducible research using Guix, » en, *Scientific Data*, t. 9, n° 1, p. 597, oct. 2022, ISSN : 2052-4463. DOI : 10.1038/s41597-022-01720-9. visité le 30 sept. 2025. adresse : https://www.nature.com/articles/s41597-022-01720-9.