Contribution Title

Romain Avouac and
Thomas Faria

Abstract Abstract here

1 Context

1.1 Objectif

- Traitement de la donnée au sens large : innovation en statistique publique → ML, big data, confidentialité ...
- ESS Net Big Data I et II

1.2 Freins à l'innovation

- Thème général : donner de l'autonomie
- Limites du poste de travail : littérature sur scaling horizontal / vertical
- Observation commune aux différents INS :
  - Insee / SSM : homogénéité des parcours, pourtant grande diversité d'infra, de moyens DSI → difficulté à partager des environnements, des formations → idée de fournir une "sandbox", un commun technologique (2020) [NB : dans la continuité, sandbox à l'échelle européenne via le one-stop-shop (2024)]
  - Visions/incitations différentes DSI/statisticien → sécurité avant le fonctionnel

Romain Avouac
Insee, 88 avenue François Verdier, Montrouge, e-mail: romain.avouac@insee.fr

Thomas Faria
Insee, 88 avenue François Verdier, Montrouge, e-mail: thomas.faria@insee.fr

- Inspirations : DevOps, DataOps

1.3 Innovation technologique

Observation : convergence d'éco-sytèmes.
Axe : big data is dead → architecture découplage.

- Transition éco big data → éco découplage : co-localisation plus très justifiée
- Stockage objet
- Infra BD tradi très spécialisées (calcul distribué). Aujourd'hui avec ML etc cas d'usages bcp plus diversifiés → outils d'automatisation, MLOPS, GPUs
- Insee : déjà culture fichier SAS + volumétries limitées → sauté l'étape BDD (cf. big data is dead)

Axe : conteneurisation comme moyen d'autonomisation.

- Conteneurisation = light virtualization vs. VM
- Tendance DevOps → DataOps, MLOps
- Reproductibilité des traitements

2 Implementation

2.1 Onyxia

Axe : mise à dispo des technos cloud → favoriser l'autonomie.

- Convergence des choix d'archi. Mais suffisant pour garantir l'autonomie : non → les outils de l'éco-système s'adressent plutôt à des informaticiens (ex : difficulté de configurer Spark sur du stockage objet en mode kube)
- Eco système découplé, mais exigeant → compétences diverses.
- Enjeu : faciliter l'acces aux ressources cloud pour les statisticiens (qui doit déjà s'acculturer à la reproductibilité → convergence avec les outils des développeurs) → double décalage qui demande une assistance
- IHM Onyxia comme liant technique

Axe : principes

- production-ready : outils d'automatisation (-> autonomie)
- no vendor-lockin (enfermement de la structure → coût (licences) et des pratiques → fige les compétences)
- cloud-native : onyxia n'est pas le choix fondamental, le parti pris est sur le choix sous-jacent : conteneurisation + stockage objet

2.2 SSP Cloud

- Orientation plateforme : instance vivante d'Onyxia, ouverte, collaborative, sandbox (cf. ref papier SSP Cloud sur l'aspect plateforme)
- Innovation ouverte → littérature
- Open-data
- Instance de partage : formations reproductibles + utilisation dans les écoles de stats + hackathons (organisation annuelle du funathon cf. one-stop-shop)
- A catalog of services which covers the entire lifecycle of a data science project
- Acculturation aux bonnes pratiques par l'usage

2.3 Application : MLOps APE

- Illustration de la diversité des taches nécessaires dans un projet de ML et modularité indispensable de l'infra utilisée (reprendre infra BD trop spécifique et onyxia cool)
- Possible car équipe au pied du mur → Innovation possible mais pas voulu
- Notebook avec méthode de data science classique
- Avantages données ouvertes → utilisation ssp cloud possible
- Parler des problèmes liés à ce type de développement → MLOps pour les résoudre (model versionning, logging parameters)
- Logiciel qui permet de suivre cette approche = MLflow et c'est dispo sur ssp cloud
- Distribution des entrainements : scaling horizontal (argo workflow)
- Déploiement API avec fastAPI → conteneurisation (liberté vis à vis de l'informatique)
- Automatisation des processus avec argoCD pour déploiement API et dashboard de monitoring
- Environnement dev et production très proche → passage en prod facilité
  - Transmission d'une image
  - Transmission d'une API
- Monitoring indispensable

3 Future perspective and discussion

3.1 Future

- Onyxia, un bien commun opensource largement réutilisé (Insee, SSB) → faciliter les contributions pour la postérité du projet open-source, qui dépasse l'Insee

- One-stop-shop : SSP Cloud comme plateforme de référence pour les projets de ML → croissance de l'offre de formation (+ traduction)
- Accompagner les réinstanciations (datafid, POCs dans le secteur privé)
- Multiplication des projets qui passent en prod (applications de dataviz, modèles de ML avec MLOps, webscraping : Jocas/WINs)

3.2 Discussion

- Cout d'entrée important pour l'organisation : stockage objet, cluster kube/conteneurisation
  - Choix fondamental d'archi → limite à la diffusion d'onyxia
  - Assumer le choix : compétences, organisation ...
  - Mais globalement : tendance favorable car beaucoup d'orga et INS font ce choix
- Cout d'entrée important pour le statisticien :
  - Non-persistence de l'environnement → git + stockage objet
  - Travail dans un conteneur → perte de repères sur l'environnement
  - Mais formation : bonnes pratiques + écoles de formation Insee + accompagnements
- SSP Cloud :
  - Instance ouverte → absence de données sensibles → grosse limitation des cas d'usage réalisables + frustrations → en résumé, difficile de maximiser à la fois innovation et sécurité (pb sur-contraint)
  - → résolution via le choix de l'innovation max car sujet des échanges inter-administration de données complexe + le SSP Cloud a pavé la voie à des instances internes, plus fermées → stratégie assumée "platform-as-a-package" : projet open-source packagé → facilité ++ de réinstanciation
  - Pas une plateforme de diffusion de données → pas de stratégie globale de gouvernance → le sujet de la méta-donnée n'est pas abordé.
- Gouvernance :
  - Quelle organisation ? Equipe DS centralisée qui vient en appui ou data scientists dans les orgas métiers ? Collaboration avec les équipes infos ? (cf. graphique orga/compétences de Romain)

Appendix

References

1. Broy, M.: Software engineering — from auxiliary to key technologies. In: Broy, M., Dener, E. (eds.) Software Pioneers, pp. 10-13. Springer, Heidelberg (2002)
2. Dod, J.: Effective substances. In: The Dictionary of Substances and Their Effects. Royal Society of Chemistry (1999) Available via DIALOG. http://www.rsc.org/dose/title of subordinate document. Cited 15 Jan 1999
3. Geddes, K.O., Czapor, S.R., Labahn, G.: Algorithms for Computer Algebra. Kluwer, Boston (1992)
4. Hamburger, C.: Quasimonotonicity, regularity and duality for nonlinear systems of partial differential equations. Ann. Mat. Pura. Appl. **169**, 321–354 (1995)
5. Slifka, M.K., Whitton, J.L.: Clinical implications of dysregulated cytokine production. J. Mol. Med. (2000) doi: 10.1007/s001090000086
6. J. Dod, in *The Dictionary of Substances and Their Effects*, Royal Society of Chemistry. (Available via DIALOG, 1999), http://www.rsc.org/dose/title of subordinate document. Cited 15 Jan 1999
7. H. Ibach, H. Lüth, *Solid-State Physics*, 2nd edn. (Springer, New York, 1996), pp. 45-56
8. S. Preuss, A. Demchuk Jr., M. Stuke, Appl. Phys. A **61**
9. M.K. Slifka, J.L. Whitton, J. Mol. Med., doi: 10.1007/s001090000086
10. S.E. Smith, in *Neuromuscular Junction*, ed. by E. Zaimis. Handbook of Experimental Pharmacology, vol 42 (Springer, Heidelberg, 1976), p. 593
11. Brown B, Aaron M (2001) The politics of nature. In: Smith J (ed) The rise of modern genomics, 3rd edn. Wiley, New York
12. Dod J (1999) Effective Substances. In: The dictionary of substances and their effects. Royal Society of Chemistry. Available via DIALOG. http://www.rsc.org/dose/title of subordinate document. Cited 15 Jan 1999
13. Slifka MK, Whitton JL (2000) Clinical implications of dysregulated cytokine production. J Mol Med, doi: 10.1007/s001090000086
14. Smith J, Jones M Jr, Houghton L et al (1999) Future of health insurance. N Engl J Med 965:325–329
15. South J, Blass B (2001) The future of modern genomics. Blackwell, London