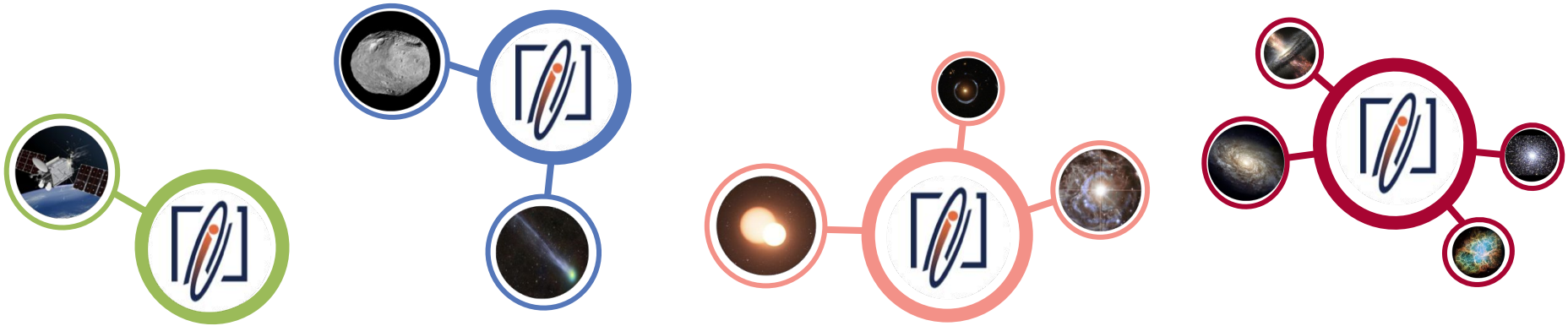
Fink: Analyse de millions d'événements astrophysiques en temps réel

Julien Peloton, IJCLab

18/10/2022

Contexte scientifique



Astronomie du ciel transitoire: étude de la variabilité des objets et phénomènes transitoires de notre système solaire, de notre galaxie, et au-delà.

L'observatoire Rubin sera le principal pourvoyeur d'alertes de la décennie (2024+)

- 10 millions d'alertes par nuit, jusqu'à une magnitude 25!
- Notre approche: **Analyse systémique** du flux de données d'alertes.

Livre blanc: https://dx.doi.org/10.1093/mnras/staa3602

Challenges techniques

Survey à grande échelle: Avec l'Observatoire Rubin (2024-2034), l'astronomie entre dans le domaine du big data (>20 Téraoctets de données par nuit).

- Connection haut-débit & sans interruption
- Analyse des alertes avec une basse latence
- Filtrage efficace des alertes en sortie
- Archivage & mise à disposition de l'intégralité des données sur 10 ans, et en temps-réel (interactivité).
- Ré-analyses en continu.

Astronomie multi-messager: La combinaison des données en temps réel de plusieurs observatoires devient nécessaire.

- Interopérabilité des outils, protocoles de communication standardisés, …

Fink en quelques dates

2019: lancement du projet (à partir d'une R&D informatique à IJCLab)

2020: MoU avec le télescope ZTF

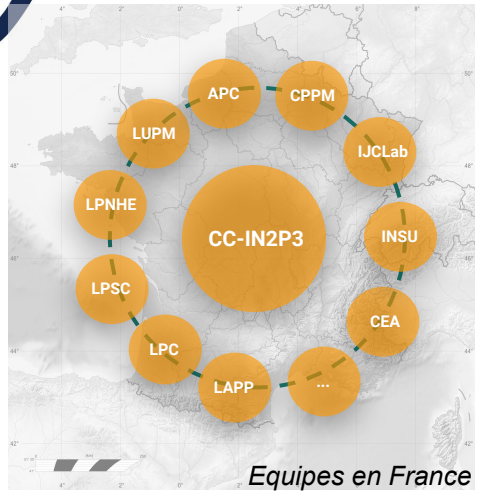
- Pathfinder pour Rubin
- 200 000 alertes/nuit

2021: projet IN2P3

- Au sein du Master Projet LSST

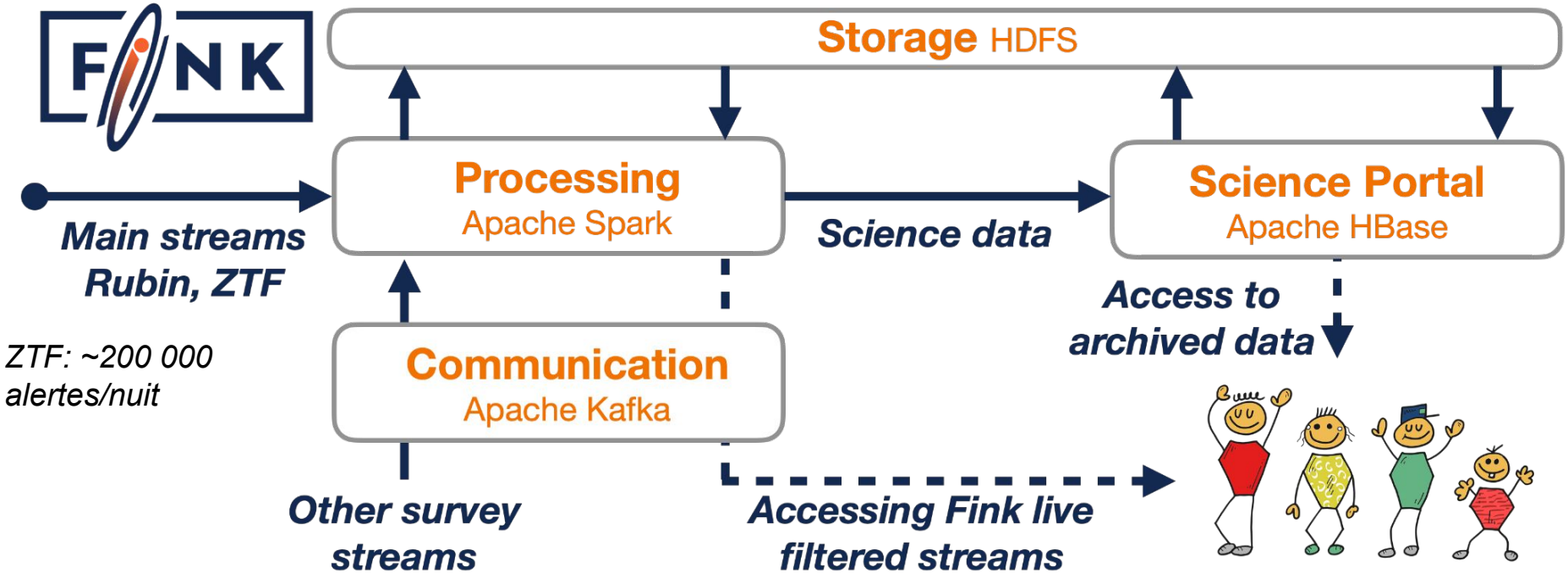
2022: Déploiement au CC-IN2P3

2023/2024: Commissioning & opérations de Rubin



Equipes en France

Prototype @ VirtualData

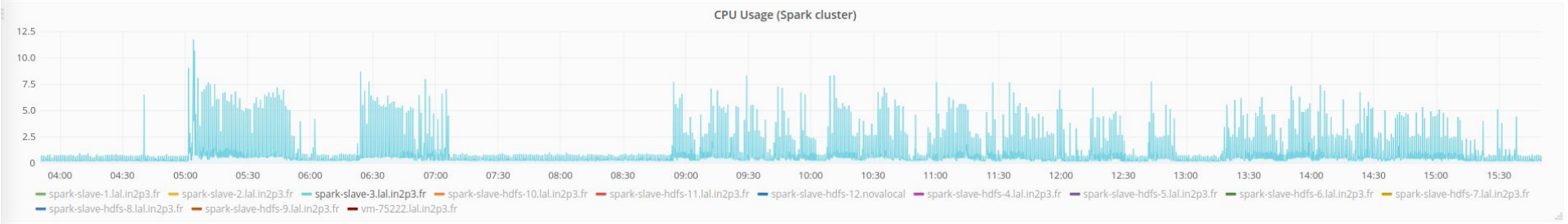


Apache Spark

R&D lancée en 2014 pour le traitement de grosses données en interactif.

Cluster en place depuis 2017, basé sur un cloud OpenStack:

- **Calcul:** 200 vCPU total, 2GB RAM/vCPU
- **Orchestrateur:** Apache Mesos */ transition vers Kubernetes*
- **Stockage:** 35 TB sur HDFS */ transition vers Ceph natif*
- **Distribution logicielle:** CernVM FS
- **Publications techniques:** Peloton et al 2018, Plaszczynski et al 2019, Plaszczynski et al 2020, Möller et al 2021



L’utilisateur au centre

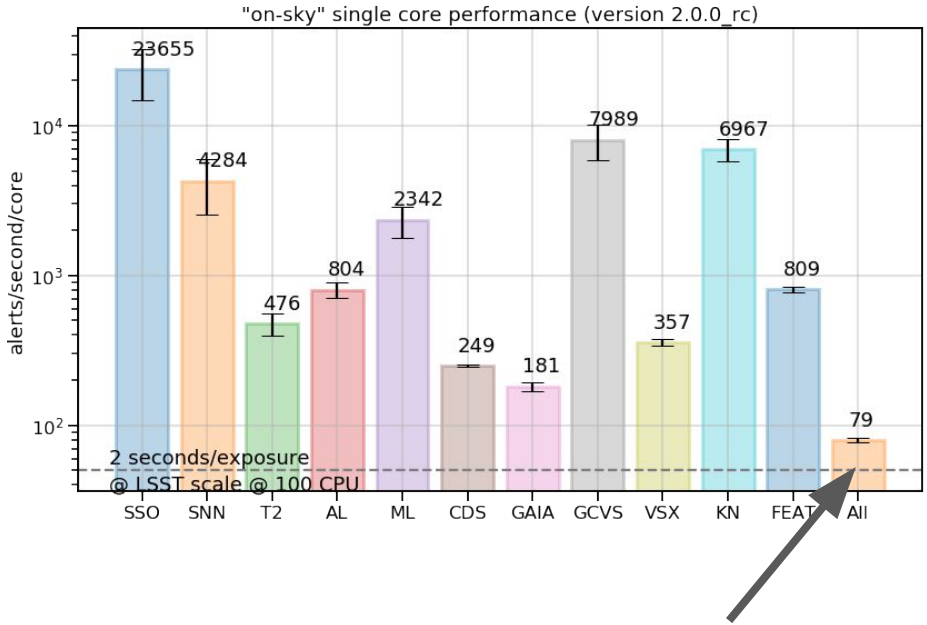
Fink est une plateforme*, où les scientifiques intègrent leur logique d’extraction.

Challenge: comment mettre l’utilisateur au centre, sans lui demander d’être expert en big data ?

- Essentiellement Machine Learning/Deep learning

La containerisation est essentielle! Le *templating* aussi…

Exemple de performance module utilisateur (plus c’est haut, mieux c’est!)



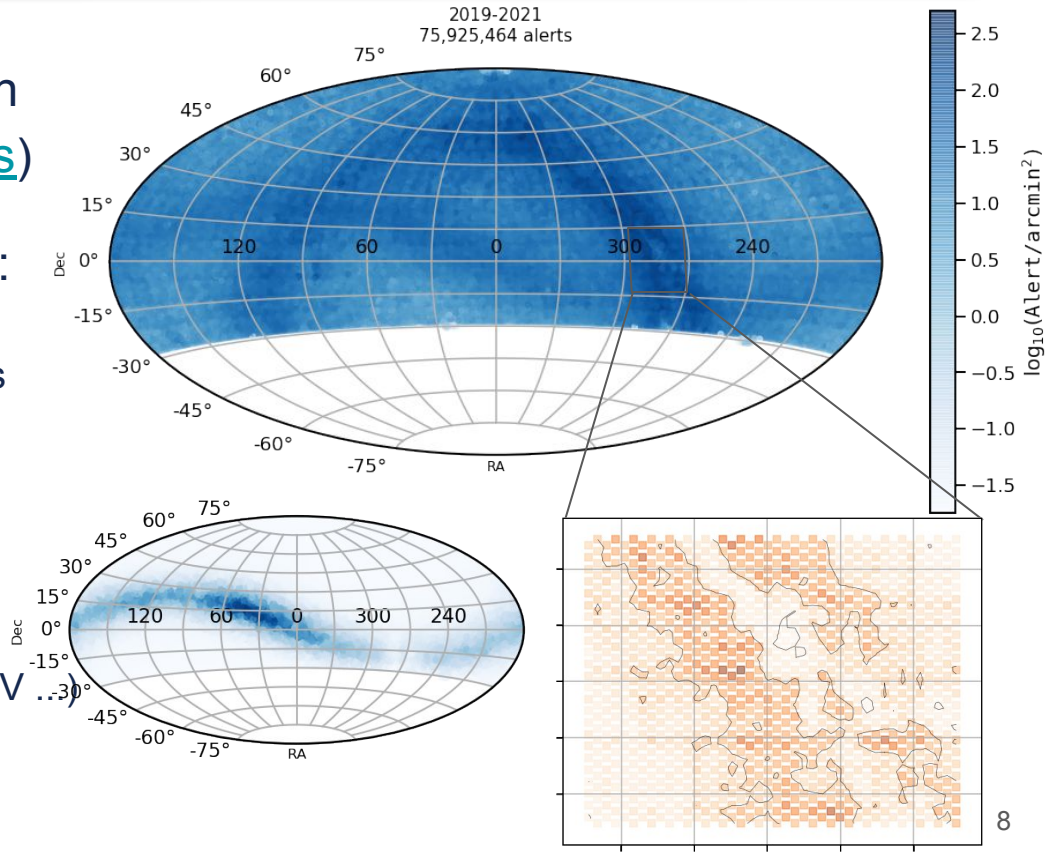
**entre le Software-as-a-Service (SaaS) et le Platform-as-a-Service (PaaS)*

Statistiques ZTF/Fink

150 million alertes reçues, 105 million analysées (https://fink-portal.org/stats)

Taux typiques (pour 200,000 alertes):

- ~75,000 étoiles variables connues
- ~25,000 objet du Système Solaire connus
- ~100 nouveaux candidats SSO
- ~100 nouveaux candidats supernovae & core-collapse
- ~10 satellites (non)identifiés
- ~5 nouveaux candidats SN Ia
- ~1 candidat transient rapide (KN, GRB, CV …)
- ~1 nouveau candidat microlentille



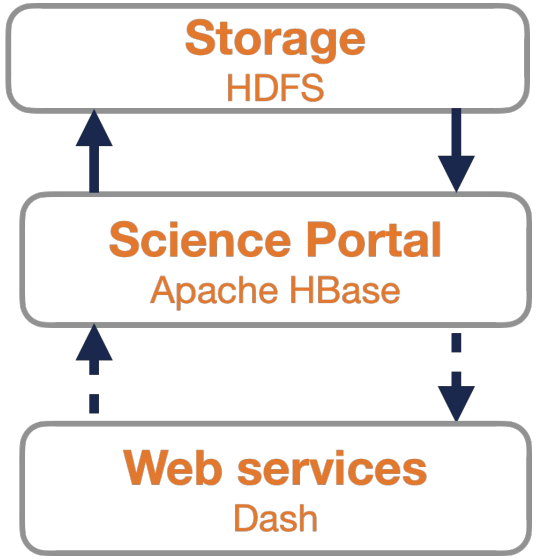
Base de données

Base non-relationnelle (Apache HBase)

- Backend pour les services web (+API REST)

Pour: Simple à déployer, passe facilement à l'échelle, simultanéité, pas de schémas à maintenir…

Contre: noSQL (un cauchemar pour la communauté astro!), index unique, se déploie sous HDFS, un peu vieillissant (moins maintenu), …

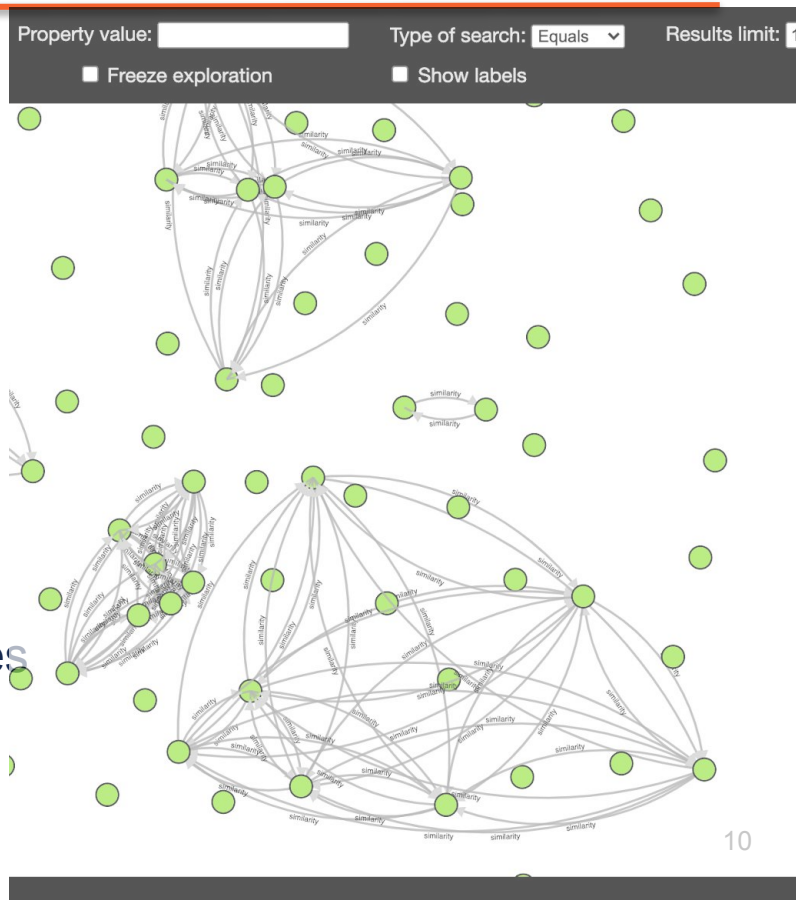


Travaux en cours

Objectif: architecture d'accès aux données de manière transparente en utilisant plusieurs types de bases de données:

- base de données SQL
- base de données NoSQL (hors graphes)
- base de données orientée graphe

Chaque base de données est utilisée pour le domaine où elle est la plus appropriée et un accès inter-technologie transparent est offert aux utilisateurs (J. Hrivnac).



Conclusion

2019-2022: Première phase de Fink déployée et réussie (2019-2022)

- Preuve de concept validée, premiers papiers scientifiques publiés.

2022: Déploiement au CC-IN2P3 de la plateforme de production pour Fink.

- R&D → Production! Passage toujours délicat…
- Le volume de données va être décuplé dans les 2 ans à venir.
- De nouvelles R&D sont en cours pour faire face à de nouveaux défis (orchestrateur, système de stockage, base de données)

La formation est cruciale!

- Pour les utilisateurs.
- Mais aussi en interne, e.g. thèse informatique/physique à IJCLab: un profil pertinent pour les projets à forte composante informatique.