Infrastructure technique intégrée
Équipements et Interconnexion des sites

GAIA Data

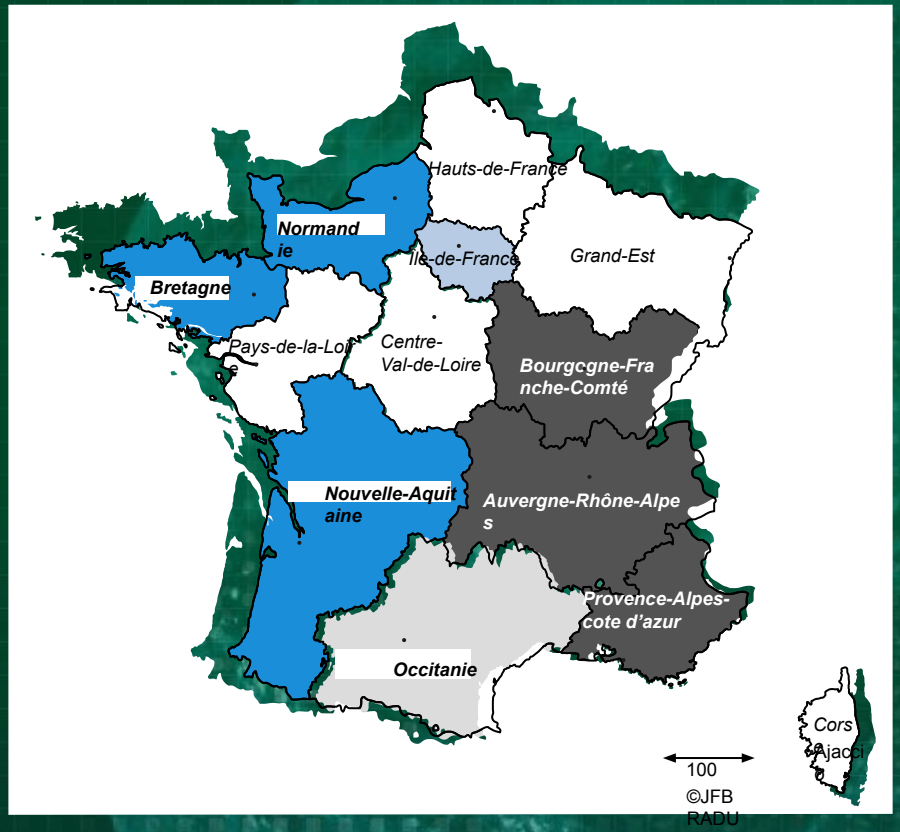
KICK-OFF
12 AVRIL 2022

Karim Ramage (CNRS/IPSL)
Coordinateur technique adjoint du projet GAIA DATA

SOMMAIRE

01

Contexte et Enjeux



Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE
DE L'ENSEIGNEMENT SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'INNOVATION

DATA TERRA

CLIMERI-France
Infrastructure de recherche nationale de modélisation du climat

PNDB
Pôle National de Données de Biodiversité

Contexte numérique national ESRI : InfraNum

Rationaliser les infrastructures informatiques de l'ESR

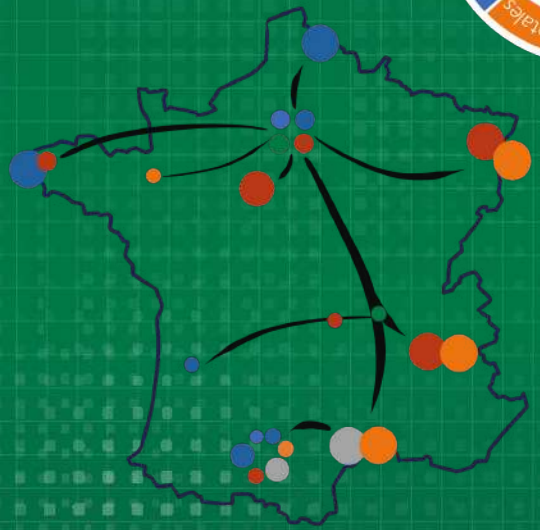
- réduire l'impact écologique des salles informatiques (datacentre)
- regrouper les RH nécessaires pour la gestion des datacentres
- offrir des hébergements à l'état de l'art, sécurisés, résilients et performants énergétiquement
- au-delà de l'hébergement physique : tendre vers une offre de cloud pour l'ESRI
- en s'appuyant sur les grandes infrastructures de calcul et données de l'ESRI : GENCI, Mésocentres régionaux, France-Grilles

Contexte des 3 Infrastructures de Recherche

- 30 Centres de données et Services regroupant les experts :
  - ingénierie de la données
  - ingénierie logicielle
  - ingénierie systèmes et réseau
- 50 Po (2020) ; 100 Po (2023) ; 150 Po (2025)
- 50 000 cœurs de calcul cumulés sur les centres de traitement

Stratégie Gaia Data pour les Equipements

Concentrer les investissements en **équipement de Gaia Data** sur les Centres de Calcul et Données des pôles de **Data-Terra**, de **ClimERI** et du **PNDB** compatibles avec la politique InfraNum, tout en **gardant un ancrage régional** essentiel au fonctionnement et au financement des activités des trois IRs.



8 sites principaux
30 sites existants

INFRASTRUCTURE GAIA DATA

8 Sites regroupant :

- Centres de Calcul Nationaux (CINES, IDRIS)
- Centres de Calcul et données d'organismes (CNES, Ifremer, BRGM)
- Mésocentres Régionaux (GRICAD, UniStra, Univ Lille, Meso@LR)
- Mésocentres Thématiques (ICARE, ESPRI, IPGP-Dante)

en développant **les liens avec les infrastructures nationales** et en tenant compte des **évolutions du paysage numérique national et européen** au cours du projet : Projets PIA3 FITS et MesoNet, développement du **Cloud et du calcul national** (FG-Cloud, Projet Clusster, PEPR Cloud, GENCI), et **européen** (EOSC, DIAS, Gaia-X, EUROPHPC)

Projets Equipex+ ou PIA4 infra

- FITS
- MesoNet
- Clusster

PEPR thématiques

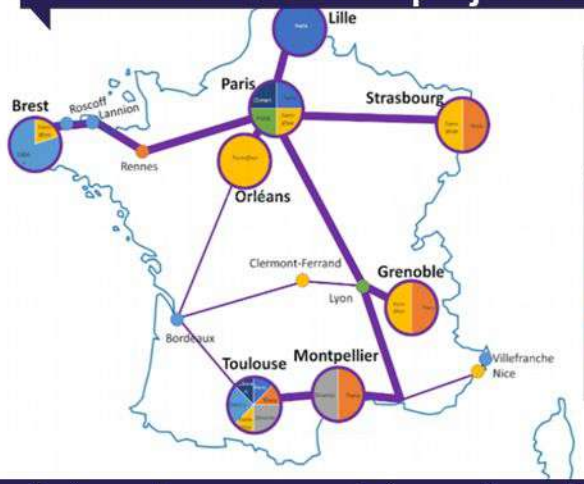
- PEPR Cloud

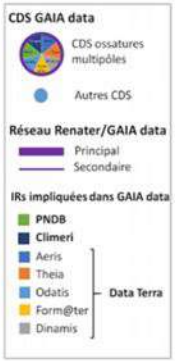
Projets H2020 – Horizon Europe

- IS-ENES
- PHIDIAS
- EOSC-Pillar
- FAIR EASE

Projets CPER en région

relation avec des projets connexes





Intégré dans le paysage international / Européen

02

Objectifs et Moyens

Renforcer les moyens dans les Centres de Calcul et Données pour assurer les missions des pôles et IR

- Stockage des données de référence
- Production des données à valeur ajoutée
- Distribution

Développer les infrastructures pour permettre l'interopérabilité des accès aux données

- Interconnexion réseau
- Grille de données
- Datalakes

Renforcer les équipements et développer les systèmes pour assurer l'interopérabilité des services entre les Centres de Calcul et Données

- Containerisation, cloud et plateforme Iaas/PaaS/IaaC
- Outils de déploiements logiciels
- Authentication and Authorization Infrastructure / Single Sign On

Renforcer les architectures spécialisées pour le service de la données

- Nœuds pour la visualisation, traitements à la demande (Virtual Research Environment, Earth Analytic Labs)
- IA / Machine Learning

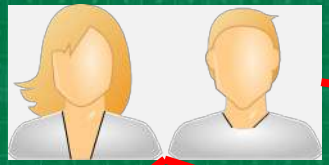
Budget WP2 Infrastructure Gaia Data

- **Equipement : 8 777 445 €**
  - Renforcement/Acquisition des équipements réseaux et sécurité…….... 758 503 €
  - Renforcement des espaces "capacitifs" pour les données de référence……..... 2 361 950 €
  - Renforcement/Acquisition des espaces de stockage d'échange ……..…….…… 1 530 959 €
  - Renforcement des espaces de disques "rapide" pour le traitement de données……..……..…. 1 384 892 €
  - Renforcement des processeurs pour les plateformes d'analyses……..…. 1 069 191 €
  - Renforcement des processeurs graphiques pour la visualisation……..…….. 509 190 €
  - Renforcement des plateformes de virtualisation/conteneurisation pour l'hébergement des services… 1 162 760 €
- **Prestations : 610 768 €**
  - Location Stockage/Calcul/Locaux dans les mésocentres régionaux …..……..…… 281 728 €
  - Installation / Déploiement des équipements……... 329 040 €
- **Réduction par rapport au budget initial :** 22 % en moyenne pour l'ensemble des sites
  - Changement de la repartition entre phase de développement et phase d'exploitation de 60/40 à 70/30, permettant de faire porter la reduction essentiellement sur la phase d'exploitation
- Recherche de financements complémentaires au niveau national (CPER, PEPR) et Européens (EOSC)
- Optimisation des marchés en lien avec les autres projets PIA3 (MesoNet) et GENCI (CINES, IDRIS)
- **Architecture logicielle** : Contribution du WP3 pour le développement et le déploiement

03

Evolutions des Centres de Gaia Data

Architecture Gaia Data



IAM
SSO

Accès aux données :
S3, OpenDAP, GridFTP

Accès aux Services :
OGC, notebooks, …

Centre Gaia Data

IAM
SSO

Stockage
Datalake / Data
Repositories

Services et
Calcul extensible

HPC

Sécurité

Sécurité

iRods / S3

Docker
Singularity

Centre Gaia Data

IAM
SSO

Stockage
Datalake / Data
Repositories

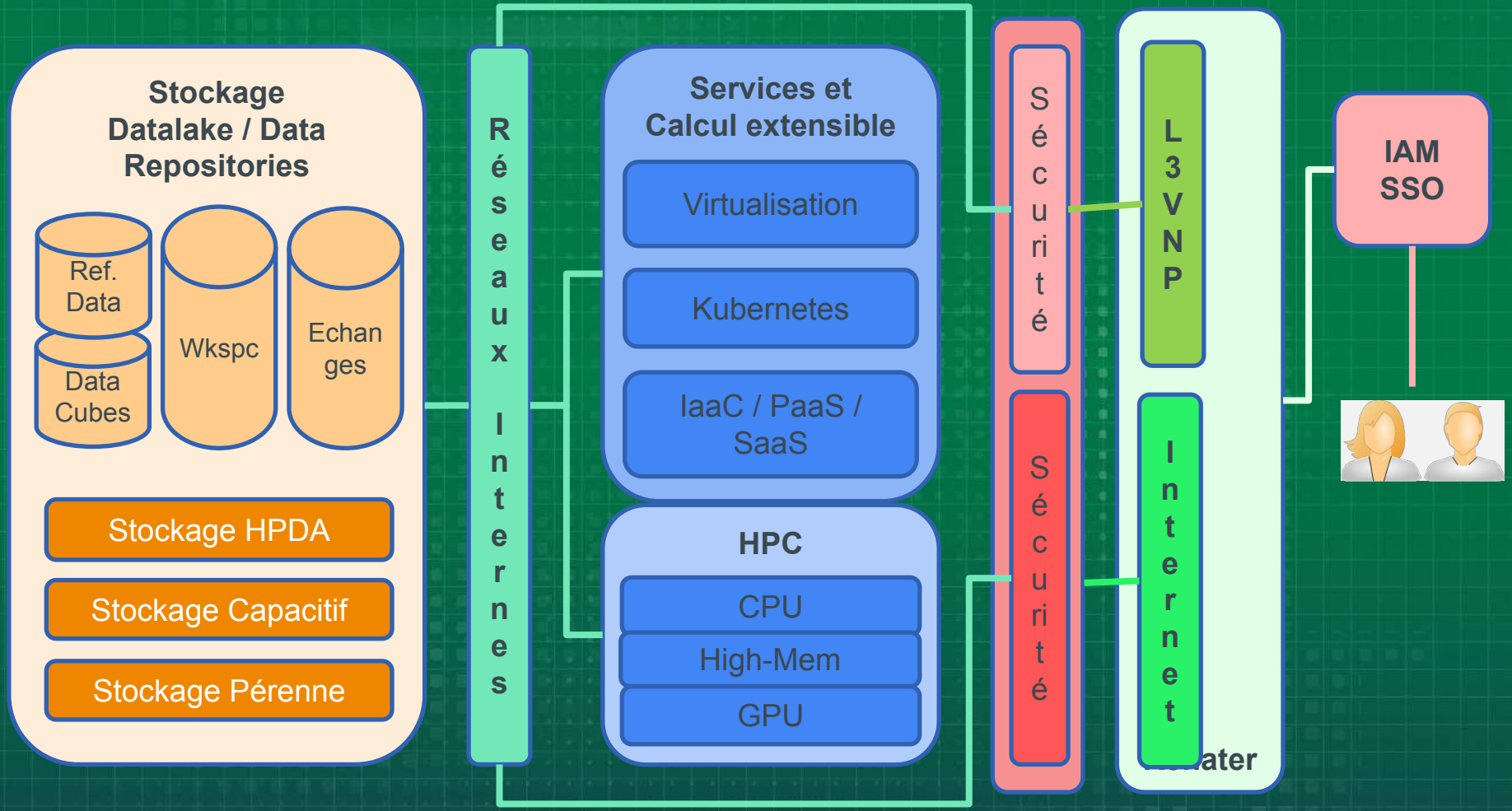
Services et
Calcul extensible

HPC

Sécurité

Sécurité

Architecture Cible des Centres de Gaia Data



Extension des systèmes de stockage capacitif des 8 centres pour les données de références

- **Garantir la capacité des pôles et des IR** à assurer leurs missions pour l'acquisition, le traitement, **l'hébergement** et la distribution des données du système Terre et de la biodiversité

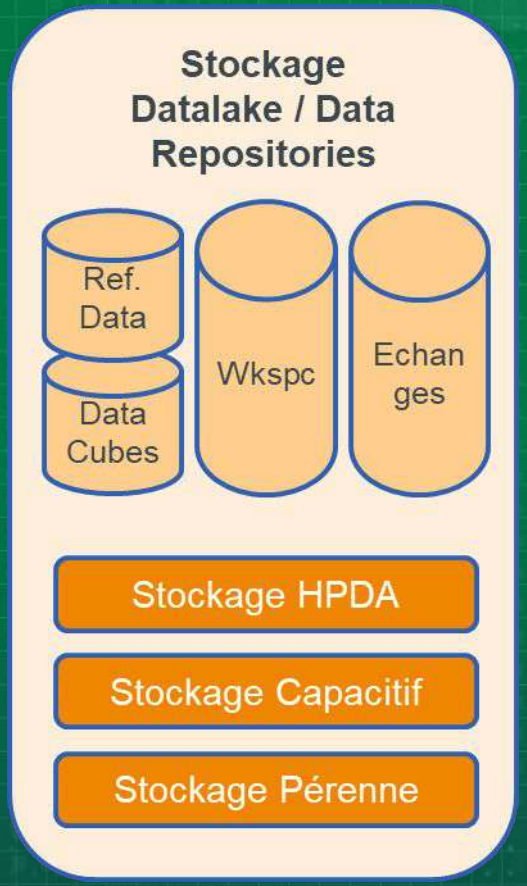
Renforcement des espaces de disques rapides :

- **Stockage au plus près des ressources de calcul** : espaces tampons pour héberger des copies des données du format d'origine, proche de l'observation ou de la simulation, vers des formats de type « **DataCubes** » pour l'analyse (algorithmes parallèles de type IA, par exemple) ou pour la visualisation : images tuilées, facettes 3D, …

Acquisition d'espaces de stockage « tampon » pour les échanges inter-centres

- Hébergement de **données de références communes** et utiles à plusieurs centres
- **Transfert de données** d'un hébergeur vers un centre de calcul (GENCI par exemple) pour des (re-)traitements massifs
- **Regroupement de jeux de données** multi-centres pour les traitements à la demande de Earth Analytics Labs

Evolution des infrastructures Ressources de Stockage



- **Renforcement des ressources HPC pour la production régulière**
- **Acquisition de nœuds à large mémoire pour:**
  - les traitements rapides à la demande
  - sélections immédiates des données d'intérêt parmi des données extrêmement nombreuses (moteurs d'indexation, TripleStore et bases NoSQL)
- **Nœuds pour la visualisation des données**
  - Processeurs graphiques pour générer des images à la volée
- **Développement des plateformes de virtualisation / containerisation pour l'hébergement des services**
- **Interopérabilité des traitements entre les centres de Gaia Data**
  - environnements logiciels reproductibles, standardisés et paramétrables
  - Packaging des applications et environnements (Guix, Spack, …)
  - Conteneurisation des codes et applications (docker, singularity)
  - Orchestrateurs d'infrastructure (Terraform, Openstack, K8S) pour faciliter les déploiements de services
- **Développements en lien avec France-Grilles, Clusster, EOSC-Association TF, ESA Cloud**

Evolution des infrastructures Ressources de Calcul

Déploiement d'interconnexions réseaux pour la grille de données et les services de Gaia Data

- **Réseau privé dédié (L3VPN)** entre les centres partenaires de Gaia Data pour sécuriser les services distribués (ex. : vlan dédiés pour la redondance des services d'authentification)
- **Très haut-débit dédié** pour le transfert et l'accès aux données entre les centres de Gaia Data, puis avec les centres européens partenaires (Projets EOSC, IS-ENES, …)
- **Haut-Débit vers Internet** pour la distribution de données – 10Gbps

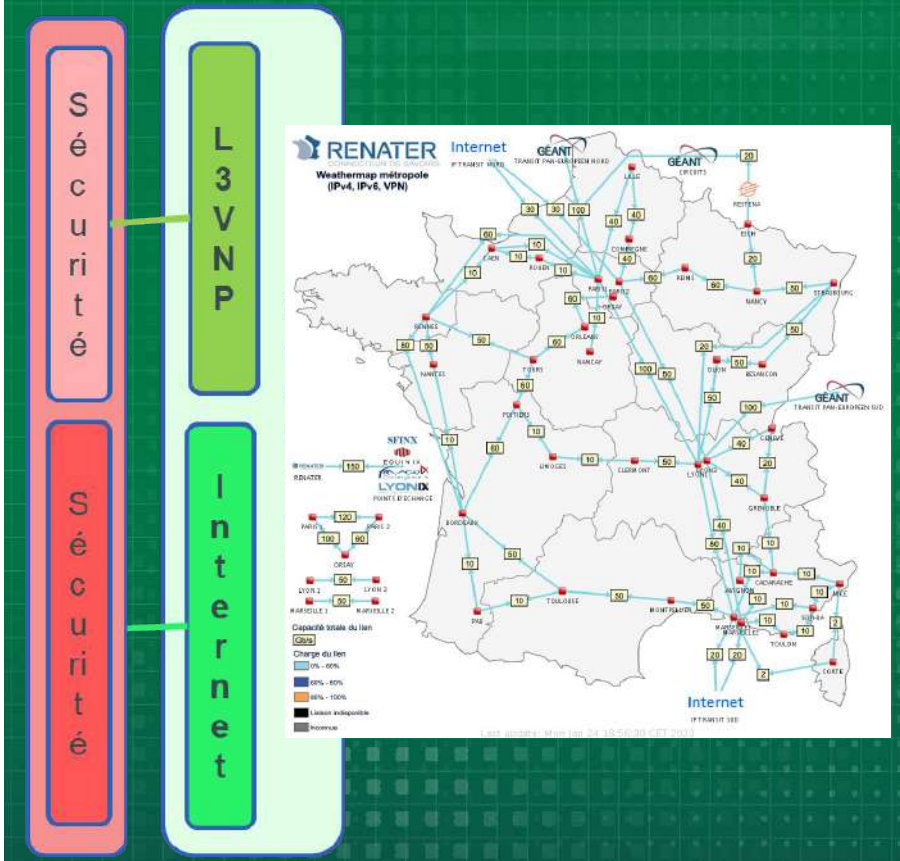
- **Mise à niveau des équipements réseaux** sur chaque site en partenariat avec les gestionnaires des réseaux d'accès (boucles locales)
- **Discussions au niveau Gaia Data avec Renater** pour la mise en œuvre des interconnexions

Mise à niveau des équipements de sécurité :

- pour les centres exposant les services Gaia Data vers le public
- pour les services inter-centres Gaia Data

- Surveillance et maintenance sécurité

Evolution des infrastructures Réseau et sécurité



Déploiement d'une Infrastructure d'Authentification et Autorisation interopérable pour l'accès aux données et services de Gaia Data

- Solution basée sur les travaux menés par AERIS : **Keycloack**

Création d'un annuaire des acteurs et des utilisateurs

- Annuaire centralisé des acteurs et des utilisateurs
- Mise en place d'un mécanisme d'authentification unique
- Proposer un service de gestion des autorisations centralisé

Gestion des niveaux de confiance

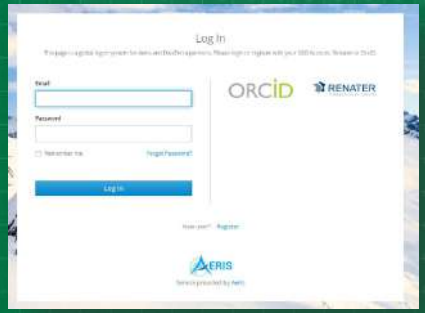
- Non authentifié
- Authentification déléguée (Fédération Renater, ORCid, réseaux sociaux, …)
- Utilisateur authentifié d'un organisme (annuaires existants)

Mise en place de workflow de validation de compte

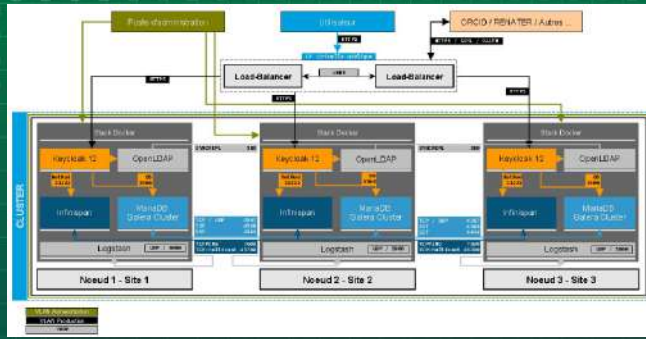
- Sécurisation fine des jeux de données des catalogues
- Statistique et traçabilité des téléchargements

Possibilité de synchroniser une fédération d'identité déjà en place dans les pôles et IR

Evolution des infrastructures Authentification / Autorisations - SSO



IAM
SSO



Evolution des infrastructures Supervision et Métriques

Développement d'un système de supervision et d'accounting mutualisé pour:

- répondre à un niveau de service (SLA) défini,
- faciliter l'exploitation de l'infrastructure,
- piloter les déploiements de ressources,
- mesurer l'empreinte carbone des systèmes

 Supervision des infrastructures systèmes et des Services

 Système de détection et d'alerte d'indisponibilité des services, de surcharge des ressources

 Métriques sur l'utilisation des ressources calcul, stockage

 Métriques sur l'utilisation de la donnée

Dashboards communs Gaia Data

04

Articulation avec les Projets PIA3/PIA4

Articulation avec les Projets PIA3/PIA4 : MESONET

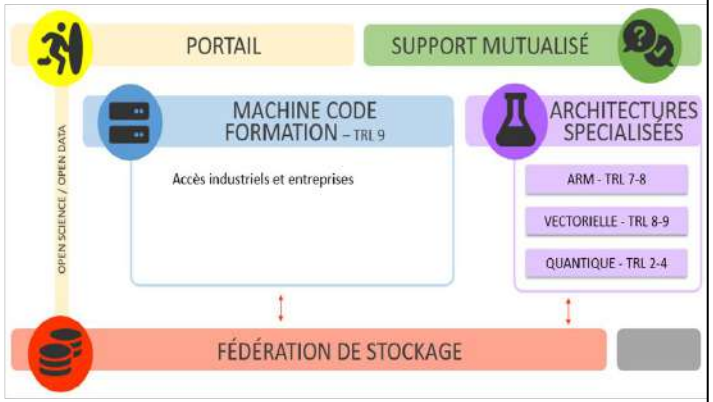
MESONET
Le mésocentre des mésocentres

Infrastructure nationale distribuée de type *mésocentre*

- Renforcer la structuration de l'offre régionale
- Disposer d'infrastructures calcul / IA au meilleur niveau technologique
- Intégrer les nouvelles communautés
- Encourager les échanges Tiers1-Tiers2
- Fournir une Infrastructure agile pour le développement des codes et la formation
- S'intégrer à la vision nationale et européenne

➢ **Créer une Infrastructure de Recherche**

14,2 M€ financés sur un budget total de 30,4 M€
début du projet au 01/10/2021 pour une durée de 6 ans



- **Mésocentres régionaux participant aux deux projets**
- **Problématiques communes :**
  - Fédération du stockage : 20 Po distribués iRods (workdir / staging)
  - Interconnexions réseau
  - Authentification fédérée
  - Sécurité (12 sites démarche d'audit sécurité)
- Participation croisée aux GT des différents projets
- Co-financements pour certains équipements

Articulation avec les Projets PIA3/PIA4 : FITS

Ambition

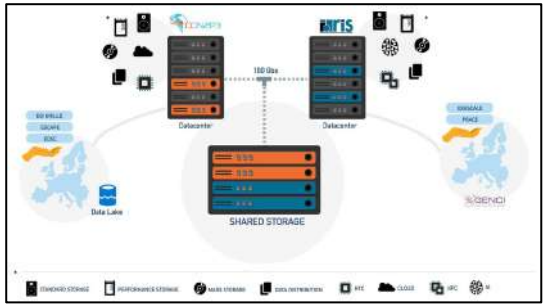
Fédérer les services et savoir-faire de l'IDRIS et du CC-IN2P3, dans le respect de leurs missions spécifiques, à travers la mise en oeuvre d'une infrastructure répartie de stockage, de traitement et de mise à disposition, diffusion et valorisation des données, hébergées dans des conditions environnementales à faible empreinte carbone.

Objectifs

- Permettre aux IR, disposant d'un modèle de calcul établi et nécessitant l'accès à des technologies variées d'y accéder de manière simple et facile
- Leur proposer des moyens d'accès et de distributions à leurs données ouvertes
- Permettre aux IR disposant de leur infrastructure de les héberger et opérer dans un environnement à l'état de l'art

Acteurs

- Acteurs du calcul au CNRS
- Communautés de recherche



Convergences possibles entre les projets :

- Authentification fédérée
- Portail d'accès aux ressources
- Débordement vers les mésocentres et cloud via France-Grille
- Hébergement des infrastructures informatiques de Gaia Data

Articulation avec les Projets PIA3/PIA4 : CLUSSTER

CLUSSTER

Cloud Unifié Souverain de Services, de TEchnologies et d'infrastructuRes

Réponse à AMI de Bpifrance relatif à la stratégie d'accélération cloud : Développement et renforcement de la filière française et européenne du Cloud

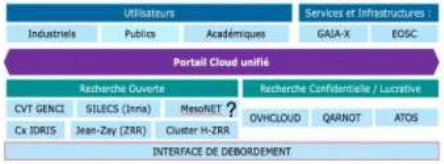
Contexte

- AI For humanity (2018)
- Recherche ouverte: Jean Zay (GENCI/IDRIS)
- Recherche confidentielle : OVH, Atos, Activeeon, Qarnot, etc.
- Des expertises académiques
- Des expertises industrielles métiers
- *Aucune offre unique adressant Recherche ouverte et confidentielle et offrant des services d'expertise*

Un portail unifié et souverain

- Fédérer l'ensemble des infrastructures existantes des acteurs privés et publics et des offres de services à valeur ajoutée
- Recherche ouverte, confidentielle et activités lucratives
- Secteur Académique, Industrielle, public
- Intégration écosystème européen; GAIA-X et EOSC

- Faciliter la lisibilité de l'écosystème pour les utilisateurs et l'usage de toutes les ressources/services existantes en France
- Accompagner: Formation, Veille techno, expertise métier



Perspective

- Evolution des services : support à l'adhésion pour couvrir de nouveaux verticaux métiers
- Evolution des domaines: Extension au quantique et à la simulation numérique

Convergence possibles entre les projets :

- Gaia Data : Use Case de Clusster
- Authentification fédérée
- Portail d'accès unifié
- Interopérabilité des traitements
- Capacité de débordement
- Hébergement des services Gaia Data pour les utilisateurs du secteur aval
- Portail vers les infrastructures européennes (EOSC, Gaia-X)

Ce travail a bénéficié d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du programme Investissements d'Avenir Equipex+.

contact@gaia-data.org

www.gaia-data.org