



## ONYXIA : UNE PLATEFORME OPEN SOURCE ORIENTÉE DATASCIENCE

---

Cédric Couralet   Alexis Guyot   Olivier Levitt

# Modernisation de l'offre scientifique du GENES

La plateforme Onyxia

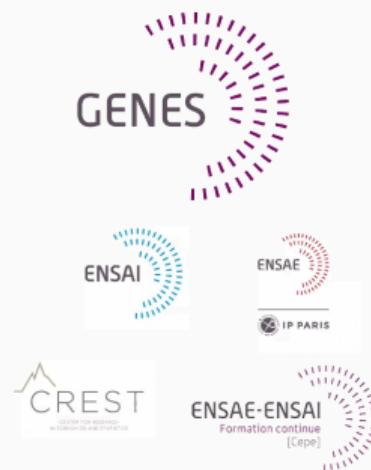
L'instance SSPCLOUD

Installation et utilisation d'Onyxia au GENES

## **MODERNISATION DE L'OFFRE SCIENTIFIQUE DU GENES**

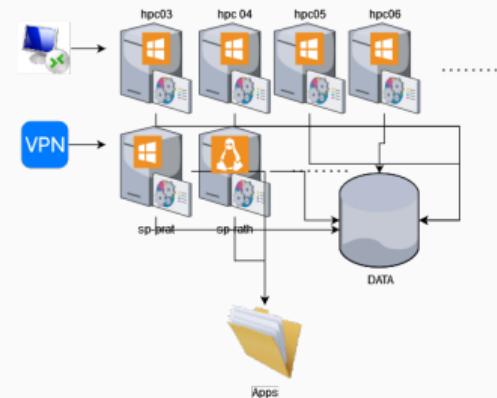
---

- Groupe des écoles nationales d'économie et statistique
- 2 écoles d'ingénieurs, un centre de formation continue, un centre de recherche (UMR)
- Tutelle Technique : Insee.
- Domaines : Statistique, économie, sociologie...
- **Fort penchant vers la science de la données**



# OFFRE SCIENTIFIQUE HISTORIQUE

- Serveurs publics/privés
- Baie de stockage
- Logiciels déployés sur partage réseau



- 😡 Allocation/partage des ressources
- 😡 Gestion des packages (r, python...)
- 😡 Gestion des mise à jour (serveur/application)
- 😑 Performances
- 😑 ...« je veux faire du LLM »
- 😑 Reproductibilité / Science Ouverte

## NOUVELLE SOLUTION

- Moins couteuse pour l'exploitation
- Proposant une meilleure allocation des ressources et leur partage
- Favorisant la reproductibilité
- Favorisant les bonnes pratiques de séparation code/data/traitement



 **kubernetes**

 **git**



## Comment embarquer les utilisateurs?



**kubernetes**



**git**



## **LA PLATEFORME ONYXIA**

---



**Un cloud opensource pour la datascience (By Insee)**

- Pas d'enfermement dans une solution
- Cloud Native
- 100% Open Source (MIT)
- Déploiement facile

# ONYXIA, C'EST QUOI?

- Une application web permettant le déploiement de service sur cluster kubernetes
- Un catalogue de services spécialement construit pour l'intégration de services externes (Stockage Objet, Gestion de secrets, Git...)
- Un catalogue de formation

The screenshot shows the Onyxia web interface with the following details:

- Header:** Onyxia - SSP Cloud Databab, Hélène personal project.
- Top Bar:** Tutorials, AIML4OS, Déconnexion.
- Left Sidebar:** Réduire, Accueil, Mon compte, Paramètres du projet, Catalogue de services, Mes services (selected), Mes secrets, Mes fichiers, Explorateur de Données.
- Middle Content:**
  - Mes services:** A callout box with instructions: "Lancer, visualiser et gérer rapidement vos différents services en cours d'exécution. Il est recommandé de supprimer vos services après chaque session de travail." Buttons: Rafraîchir, Nouveau service, Supprimer tous, Events.
  - Services en cours:** A list of running services:
    - Vscode-torch-fastText (Status: Hier, Last check: 1 minute ago, Open button)
    - Miflow (Status: Hier, Last check: 2 days ago, Open button)
  - Enregistrés:** A list of registered services:
    - vscode-torch-fastText (Status: Hier, Last check: 2 days ago, Lancer button)
    - vscode-pytorch (Status: Hier, Last check: 2 days ago, Lancer button)
- Bottom Footer:** 2017 - 2024 Onyxia, Contribuer au projet, Français, Conditions d'utilisation, v0.1.2.

# DÉMO

Onyxia - SSP Cloud Datalab h4njlg personal project ▾

Tutoriels AIML4OS Déconnexion

Réduire Accueil Mon compte Paramètres du projet Catalogue de services Mes services Mes secrets Mes fichiers Explorateur de Données

## Mes services

Lancer, visualiser et gérer rapidement vos différents services en cours d'exécution. Il est recommandé de supprimer vos services après chaque session de travail.

Rafraîchir + Nouveau service Supprimer tous Events

### Services en cours

```
$ helm list --namespace user-h4njlg
```

Service	Démarré :	Action
Vscode-torch-fastText	hier	Ouvrir
Mlflow	il y a 2 jours	Ouvrir

### Enregistrés

Développer (2)

Service	Démarré :	Action
vscode-torch-fastText		Lancer ⋮
vscode-pytorch		Lancer ⋮

## FONCTIONNEMENT : CONFIGURATION DES SERVICES

- Utilisation de helm pour le lancement des services
- Utilisation du values.schema.json pour l'affichage du formulaire

```
"accessKeyId": {  
    "description": "AWS Access Key",  
    "type": "string",  
    "x-form": {  
        "value": "{{s3.AWS_ACCESS_KEY_ID}}"  
    },  
    "hidden": {  
        "value": false,  
        "path": "s3/enabled"  
    }  
}
```

## FONCTIONNEMENT : MÉCANISME DE DÉCOUVERTE

Utilisation de la fonction lookup de Helm

```
 {{ range $index, $secret := (lookup "v1" "Secret" $namespace "").items }}  
 {{- if (index $secret "metadata" "annotations") }}  
 {{- if and (index $secret "metadata" "annotations" "onyxia/discovery")  
 "annotations" "onyxia/discovery" | toString) }}  
 {{- end }}  
 {{- end }}
```

## AUTRES FONCTIONNALITÉS

---

- Lien partageable pour lancement de services
- Gestion de projet/groupe
- Possibilité d'avoir son propre catalogue (*simple* dépôt de chart Helm)
- Restriction sur le lancement de service

roles :

```
- roleName: vip
  files:
    - relativePath: nodeSelector-gpu.json
      content: |
        {
          "$schema": "http://json-schema.org/draft-07/schema#",
          [...]
          "default": "NVIDIA-A2",
          "enum": ["NVIDIA-A2", "Tesla-T4", "NVIDIA-H100-PCIe"]
          [...]
        }
```

## **L'INSTANCE SSPCLOUD**

---

## L'INSTANCE SSPCLOUD

<https://datalab.sspcloud.fr>

- Instance d'onyxia maintenue par l'Insee
- Expérimentation sur données non sensible, partage de connaissances (formations)
- Ouverte d'abord au data scientist du service statistique public, étendue à l'ensemble des agents de l'état et aux étudiants

- Bac à sable ML/IA pour la statistique publique européenne et canadienne
- 1000 utilisateurs quotidiens (7000 inscrits)

# CATALOGUE DE FORMATION

Plateforme SSP Cloud



Français ▾



## Formations et tutoriels

python |

Damien Dotta et 2 autres

### Polars

Des tutoriels R et Python pour prendre en main Polars, une librairie Rust qui offre des performances exceptionnelles sur les DataFrames.

[Consolider](#) [Découvrir](#) [Apprendre](#)

Ouvrir

Lucas Malherbe

### Appariement de données individuelles

Des tutoriels en Python et en R pour s'initier à l'appariement de données individuelles.

[Découvrir](#) [Apprendre](#)

Ouvrir

Inseefrlab

### Funathon 2024

Des tutoriels pour découvrir et pratiquer la data science autour du thème 'Décollage imminent pour la data science'

[Consolider](#) [Découvrir](#)

Ouvrir

inseefrlab

### Initiation à Python

Cours introductif à Python : fondamentaux du langage et premières manipulations de données

[Découvrir](#) [Apprendre](#)

Ouvrir

36h10 Lino Gallana

### Python pour la data science

Approfondissement de Python pour la data science : manipulation de données, visualisation, modélisation, traitement du langage naturel

[Consolider](#) [Apprendre](#)

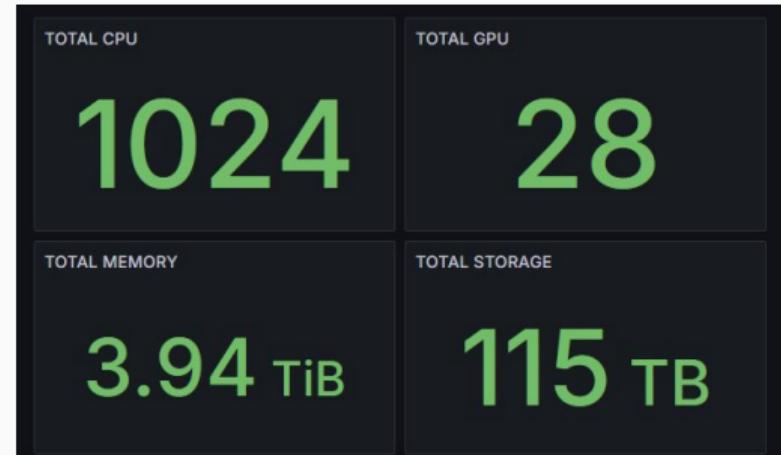
Ouvrir

## **INSTALLATION ET UTILISATION D'ONYXIA AU GENES**

---

- Service de gestion du code → **gitea**
- Service de stockage objet → **minio**
- Accessible depuis l'ensemble des utilisateurs du GENES → **OIDC/Keycloak**
- Plateforme d'exécution des traitements → **Kubernetes**
- Point d'entrée → **Onyxia**

- Installation avec KubeSpray (ansible)
- Control Plane : 3 VM
- Ingress : 2 VM
- 4 serveurs physiques :
  - 200 CPU
  - 1 To RAM
  - 1 GPU A100 40Gbps (7 slices)
- longhorn pour les volumes
- argo-cd pour le déploiement de services
- Publication http/https avec un Ingress Nginx
- Monitoring avec prometheus/grafana



## DÉPLOIEMENT D'ONYXIA

---

- Keycloak pour l'authentification (utilisation de l'annuaire **GENES**)
- Minio pour stockage S3. Policy pour 1 bucket par utilisateur
- Configuration de l'api Kubernetes pour 1 namespace par utilisateur

- Onxia déployé à partir du chart helm fourni
- Utilisation des catalogues par défaut
- Personnalisation minimale

# DÉPLOIEMENT D'ONYXIA

Datalab ccouralet-ensae personal project ▾

Rejoignez nous sur Teams Documentation Déconnexion

< Réduire

Accueil

Mon compte

Paramètres du projet

Catalogue de services

Mes services

Mes secrets

Mes fichiers

Coquille SQL OLAP

Explorateur de Données

Bienvenue Cedric!

Travaillez avec Python ou R et disposez de la puissance dont vous avez besoin !

Nouvel utilisateur du datalab ?



**Un environnement ergonomique et des services à la demande**

Analysez les données, faites du calcul distribué et profitez d'un large catalogue de services. Réservez la puissance de calcul dont vous avez besoin.

Consulter le catalogue

**Une communauté active et enthousiaste à votre écoute**

Profitez et partagez des ressources mises à votre disposition : tutoriels, formations et canaux d'échanges.

Rejoindre la communauté

**Un espace de stockage de données rapide, flexible et en ligne**

Pour accéder facilement à vos données et à celles mises à votre disposition depuis vos programmes - Implémentation API S3

Consulter des données

# UTILISATION DU DATALAB

## Déploiement d'application oTree pour le laboratoire d'expérimentations sociales du CREST :

- Utilisation de gitea, drone-ci et argocd pour développement collaboratif et déploiement en continue sur kubernetes sous forme de chart helm

## Démarrage d'environnement de formation préconfiguré

- Lien autosuffisant vers un service Onyxia ((python + extensions + notebooks) donné aux stagiaire

## Business Data Challenge de l'ENSAE

- Données sur Minio en lecture seule
- Chaque équipe a son espace de travail collaboratif sur Onyxia

## Ajout de nouveaux services au catalogue

- vscode-c++
- limesurvey, wordpress

# CHALLENGES ET PERSPECTIVES

---

## Challenges

- Compétences Kubernetes à acquérir
- Accompagnement des utilisateurs (*Comment j'accède en ssh ?*)
- Expliquer les bonnes pratiques
- Maîtriser l'ensemble

## Perspectives

- Réservation de ressources pour projets (GPU)
- Ajout de nouveaux services au catalogues
- Reflexion sur logiciels non web/non libres (matlab/sas)
- Ajout de formations spécifiques

## CONCLUSION

---

**Pour le GENES, Onyxia a permis une modernisation franche de son offre scientifique.**