

DATALAKE

Jacques FOURNIER

Directeur Général des Statistiques

Renaud LACROIX

Directeur de l'ingénierie et de la méthodologie (DGS)

1^{er} décembre 2017

Contexte et positionnement stratégique

☐ **Le Datalake de la Banque de France est un projet stratégique**

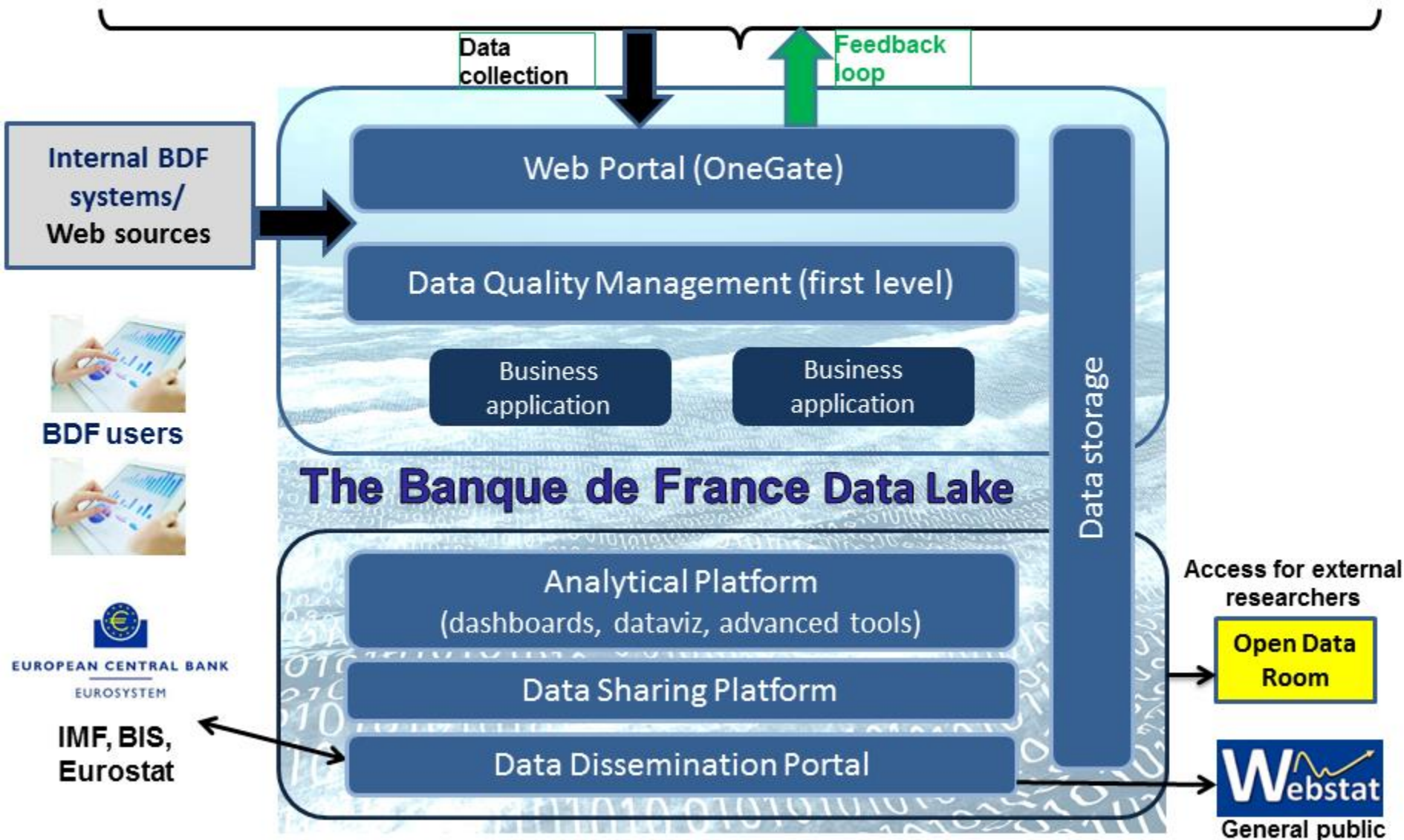
☐ **5 objectifs majeurs :**

1. **Face à la perspective de données granulaires massives, adapter sa technologie**
 - ✓ dès Octobre 2018, Anacredit impliquera la collecte mensuelle de 94 attributs pour 21 millions de lignes de crédit. Idem pour les titres (règlement Securities Holding Statistics)
2. **Assurer le meilleur service public au moindre coût, en devenant un acteur du Big Data et en réalisant des synergies internes**
3. **Accroître et moderniser les capacités d'analyse**
4. **Offrir aux remettants (institutions financières et non financières) un outil moderne, vecteur pour eux d'économies et d'efficacité**
5. **Maintenir la qualité de l'information économique et financière : la mauvaise donnée ne doit pas chasser la bonne.**

Les objectifs du Datalake

- ☐ Le projet vise à bâtir un espace de données granulaires pluridisciplinaires (Datalake), offrant un service flexible et innovant aux utilisateurs internes, qui bénéficiera également aux entreprises soumises à des obligations de reporting à la Banque de France
- ☐ La plateforme offrira les services clés permettant l'exploitation efficace et « intelligente » des données tout au long du cycle de production (output)
- ☐ La plateforme est au service de tous les métiers de la Banque de France, SURFI (supervision) incluse
- ☐ La plate-forme sera complète : stockage et contrôles statistiques automatisés, plate-forme analytique (édition de tableaux de bord, machine learning, ...)
- ☐ Un travail en plateau, gestion mode « start-up », suivi rapproché du management

**REPORTING AGENTS : FINANCIAL INSTITUTIONS, NON
FINANCIAL CORPORATIONS**



☐ **Mise en œuvre de la réglementation ANACREDIT**

- Plus de 2000 contrôles à mettre en œuvre à la demande de la BCE sur les données granulaires

☐ **Mise en œuvre de la collecte DATAGAPS (G20, IMF/BRI):**

- Données d'exposition sur l'actif et le passif des groupes bancaires systémiques

☐ **Rénovation du traitement des remises des banques et des assurances pour les besoins prudentiels et statistiques :**

- Collecte et premiers contrôles qualité des remises prudentielles

☐ **Réingénierie des activités de calcul et de contrôle dans plusieurs métiers de la Banque**

- Automatiser et centraliser les contrôles simples, de premier niveau, actuellement répartis
- Privilégier les logiciels libres massivement utilisés par la communauté scientifique et disponibles sur Datalake

☐ **Réingénierie de systèmes décisionnels**

- Plateforme décisionnelle de la Direction des services bancaires (clients externes de la Banque)

Le projet reçoit un accueil favorable

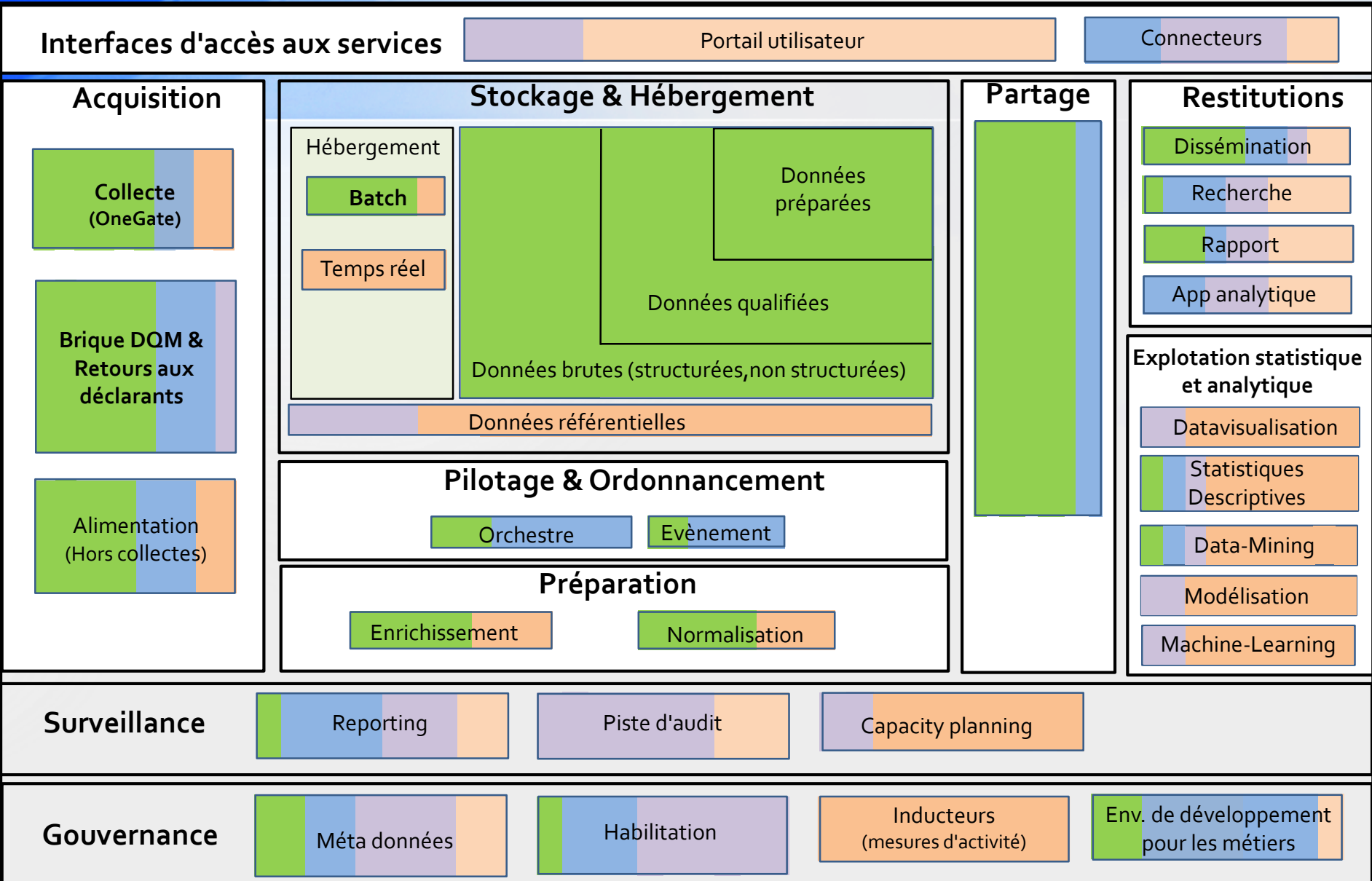
❑ En interne :

- Projet transversal stratégique, en cours de réalisation.

❑ En externe :

- Plusieurs pays européens et extra-européens ont manifesté un net intérêt pour la démarche et souhaitent s'en inspirer
- Les points forts du projet vus de nos partenaires :
 - La capacité du Datalake à **dépasser les silos** pour développer le **partage des données** et des outils statistiques
 - **Un savoir-faire avéré** sur des technologies statistiques innovantes (acquis avec la plate forme de 'data sharing' interne, l'ouverture des données à la recherche au travers de l'Open Data Room, la plateforme européenne MMSR - Money Market Statistical Reporting)

Le socle de services du Datalake



Avancement des travaux

Zoom sur la brique de contrôle du Datalake

1/2

❑ Le module DQM est alimenté avec :

- ✓ Les données au format tableau
- ✓ Un fichier « Méta » qui décrit la structure des tableaux
- ✓ Un fichier « Matrice » qui contient les contrôles paramétrés par les utilisateurs

❑ Typologies de contrôles couverts

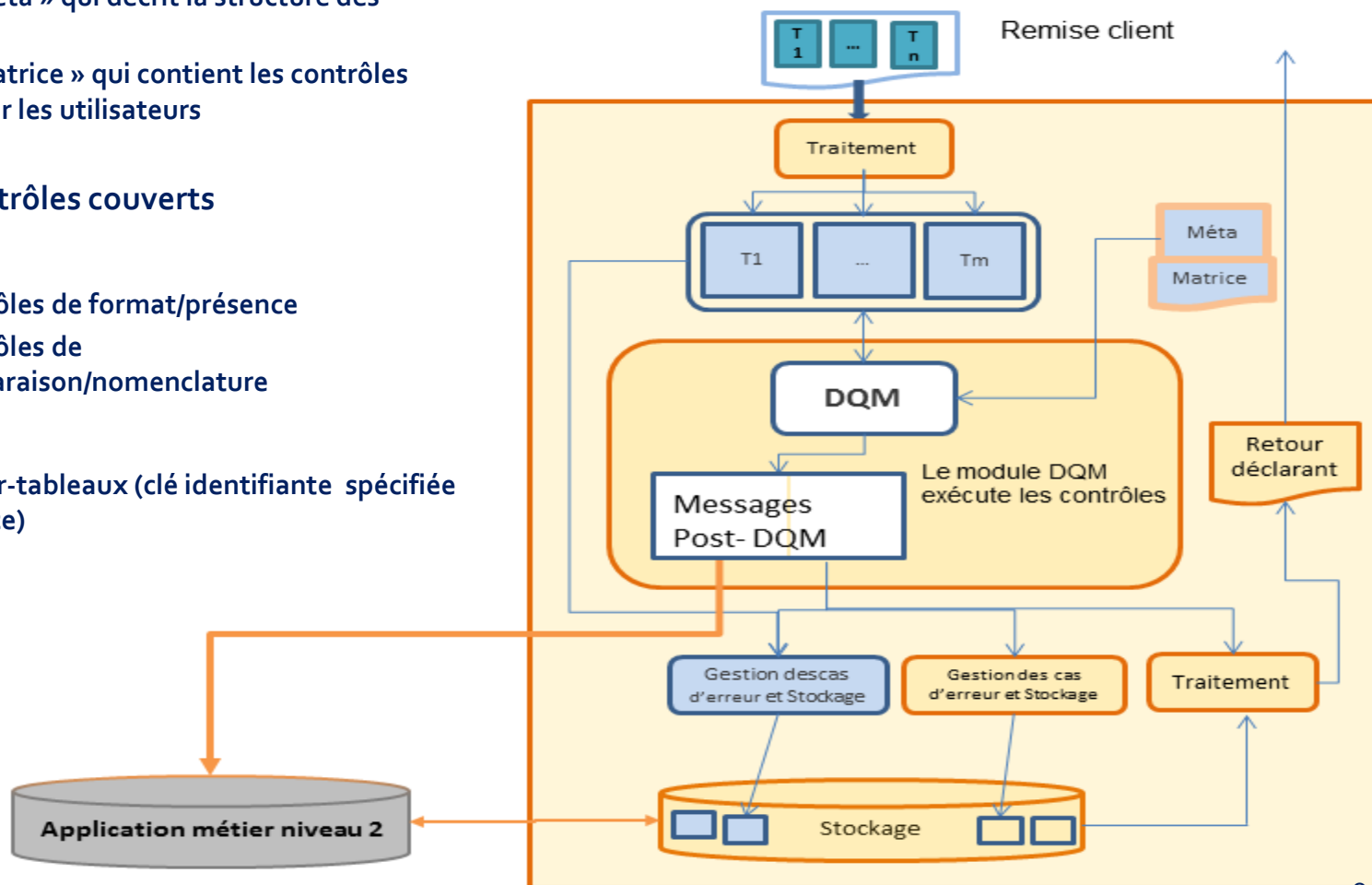
1. Intra tableau

- ✓ Tous les contrôles de format/présence
- ✓ Tous les contrôles de somme/comparaison/nomenclature

2. Inter tableaux

- ✓ Contrôles inter-tableaux (clé identifiante spécifiée dans la matrice)

- Service natif DataLake
- Spécifique solution client
- Hors DL



Avancement des travaux

Zoom sur la brique de contrôle du Datalake

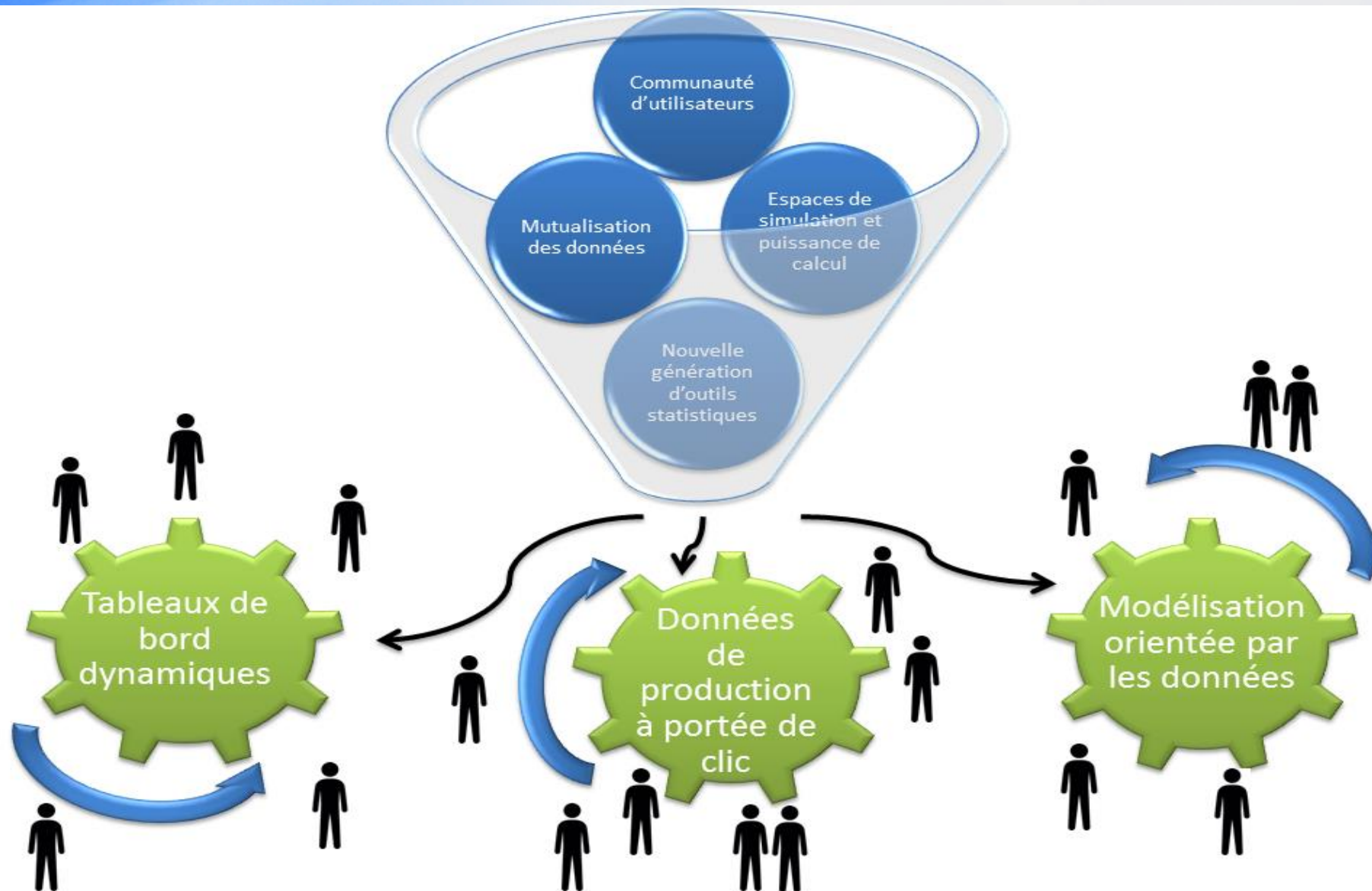
2/2

- ☐ Les retours vers les déclarants répondront à plusieurs objectifs :
 - Transmettre aux établissements déclarants des retours immédiats (contrôles techniques) et différés (contrôles fonctionnels) dans un délai n'excédant pas 24H
 - Faciliter l'intégration des fichiers d'anomalies dans le système d'information des déclarants en vue de leur analyse par des outils automatisés
 - Faciliter le dialogue entre la BDF et les établissements déclarants autour de la qualité des données transmises

Avancement des travaux

Zoom sur la plateforme analytique

1/2



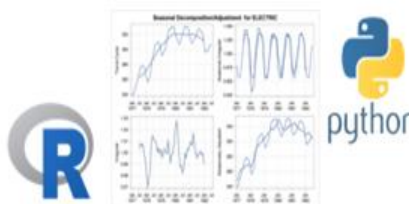
Avancement des travaux

Zoom sur la plateforme analytique

2/2



Requêtage interactif



Bibliothèques de modules statistiques,
outils de développement



Data science : machine learning,
deep learning, réseaux de neurones



Du prototypage...
...à la production de masse

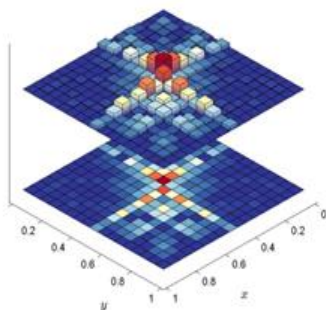


- Cahier des charges

- Echantillons

- Production de masse

- Produit fini



Produit final développé
par le métier

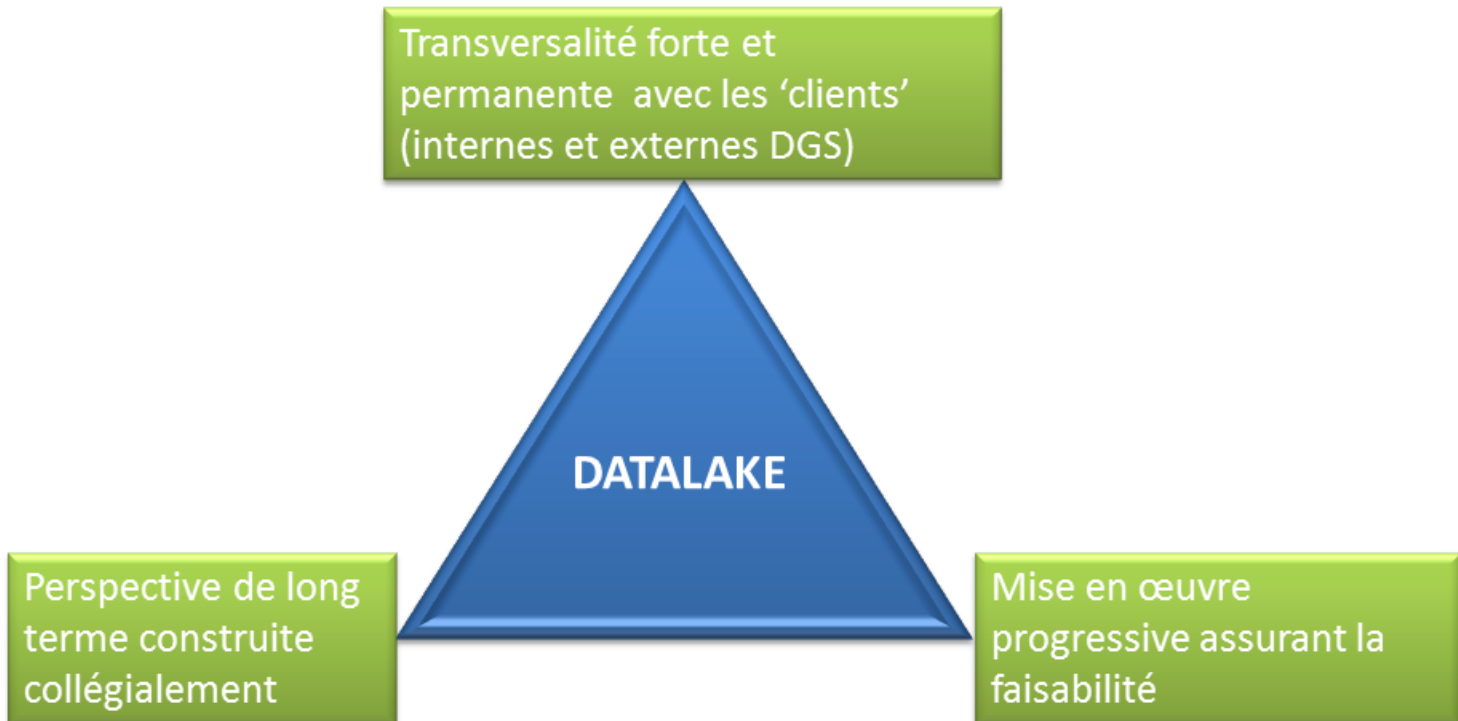


Produit final « industrialisé »
avec le concours des équipes
projet



Mise en œuvre

Principes



Mise en œuvre

Architecture technique : CLOUD privé BDF

- 4 Serveurs physiques pour héberger exclusivement les serveurs virtuels de données DataLake « DataNodes » (~16 Machines Virtuelles) ~210To brut (espace de stockage)
- 6 nouveaux serveurs pour héberger les Machines Virtuelles applicatives du Datalake (~40 VM).
- Qualification, installation et déploiement physique terminés fin Septembre
- Mise en œuvre de l'infrastructure Cloud et livraison des environnements pour les besoins du Datalake
- Attention particulière portée à la gestion de versions différente selon les environnements (production / intégration / recette / développement)

