CPER 2021-2027

Acronyme du projet /

IDEAL

Titre du projet en français	de l'**I**ntelligence **D**istribué**E A L**'Apprentissage automatique : approches innovantes pour la gestion des données massives de la recherche
Responsable du projet	Nom, Prénom : Vincent Breton – Farouk Toumani Fonctions : chercheur CNRS – professeur UCA Etablissement : CNRS - UCA
Coordonnées	Vincent.breton@clermont.in2p3.fr (0686325751) Farouk.Toumani@uca.fr (0607570529)
Partenaires	**Organismes : CNRS – INRAE – UCA** **Laboratoires : IP, LIMOS, LMBP, LPC, Mésocentre Clermont-Auvergne, TSCF**
Objet du projet	Innovations technologiques et logicielles pour la collecte, la gestion et l'analyse des données massives de la recherche
Coût total du projet	940 k€ d'équipement dont 100k€ de co-financement par le CNRS IN2P3
Calendrier du projet/Phasage	5 actions annuelles programmées de 2022 à 2027

Etablissement gestionnaire

Nom de l'établissement	Statut
CNRS ou UCA	EPSCP

RESUME

Mots-clefs : données massives, intelligence artificielle, apprentissage automatique, internet des objets, intelligence distribuée, réseaux de capteurs, services mutualisés, performance énergétique

1. INTRODUCTION – CONTEXTE

Les données massives sont désormais présentes dans pratiquement tous les secteurs de la société et de l'économie. Leur gestion pose des défis scientifiques et techniques majeurs. Dans la plupart des projets actuels qui ont pour objectif l'analyse ou l'extraction de connaissances à partir des données, entre **70 à 80% du temps** est consacré à la phase de collecte, de nettoyage, d'intégration et d'organisation des données.

Au cœur de l'explosion des données collectées, ce sont aujourd'hui plus de 20 milliards d'objets qui sont connectés à Internet (source : Strategy Analytics) et une estimation prévoit un parc mondial de 40 milliards d'appareils d'ici 2025[1].

Si le domaine des objets connectés, au cœur de cette remontée massive de données, permet de proposer des solutions dans de nombreux domaines de recherche, il est important que ces travaux prennent en compte leur propre coût énergétique. En effet, actuellement internet et ses infrastructures représentent déjà une consommation électrique équivalente à celle d'un état qui se situerait au sixième rang des pays consommateurs d'électricité. L'avènement annoncé et débuté des objets connectés devraient continuer à augmenter ce coût énergétique. Il est donc indispensable de réfléchir dès la conception de ces objets connectés à leur autonomie énergétique afin de ne pas alourdir trop fortement ce bilan. En cohérence avec ces enjeux et problématiques nationaux et internationaux, les travaux de recherche s'orientent ainsi vers l'amélioration énergétique des dispositifs de collecte des données. Dans ce contexte les cœurs de microprocesseur, dont la consommation devrait encore chuter avec les nouvelles technologies à faible pertes (FDSOI et FINFET), permettront une fois embarqués dans des systèmes dédiés, de réaliser des objets communicants à très faible consommation nécessaires à l'essor de l'internet des objets.

Dans le même temps, la complexité des traitements liée à la quantité et la diversité des données alors remontées nécessitent de revisiter de manière profonde les fondamentaux en matière de gestion de données. L'intelligence artificielle, notamment au cœur du réseau de capteurs, apparaît aujourd'hui comme la solution la plus prometteuse à la fois pour les traitements de ces données mais également dans la gestion intelligente de la contrainte énergétique de ces réseaux d'objets connectés.

Le site clermontois a développé une expertise en conception de détecteurs ainsi que dans le domaine de la gestion et de l'analyse de données massives, en s'intéressant à la fois à des questions fondamentales mais aussi appliquées, en particulier dans le cadre de grands projets dans les domaines de l'environnement et de l'agriculture, de la santé et celui de la cosmologie/physique des particules.
Il est ainsi un des premiers sites universitaires en France sur lequel une plate-forme instrumentale de réseau de capteurs sans fil est déployée pour l'ensemble des acteurs. Développée notamment pour les besoins de surveillance des agroécosystèmes, elle propose une chaîne opérationnelle grâce à laquelle des nœuds communicants transmettent à l'aide d'un protocole de communication ouvert et sécurisé (LoRa) des données de capteurs de tous types jusqu'à un cloud hébergé au Mésocentre Clermont-Auvergne.

S'appuyant sur ces solides fondements, l'objectif du projet IDEAL est de maîtriser la chaîne complète de collecte et de traitement de données, du capteur à l'apprentissage automatique.
Le projet identifie quatre champs d'investigation :

[1] https://news.strategyanalytics.com/press-release/iot-ecosystem/strategy-analytics-internet-things-now-numbers-22-billion-devices-where

- L'intelligence distribuée. L'objectif est d'améliorer la fiabilité et l'efficacité des applications en déportant le traitement des données vers les objets connectés au sein d'infrastructures auto-adaptables. L'enjeu est d'analyser la donnée au plus proche de sa source de production et de développer des approches d'apprentissage automatique pour des systèmes autonomes ou embarqués. Les verrous associés portent sur le développement d'architectures dynamiques et reconfigurables, basées sur des capteurs autonomes à faible consommation.

- L'autonomie énergétique des objets connectés. L'objectif est de réaliser les dispositifs connectés à faibles consommation pouvant devenir autonomes grâce aux développements récents dans le domaine de la récupération d'énergie ambiante. Les solutions matérielles issues de ces recherches permettront un développement à large échelle et à coût énergétique réduit des outils de remonter des données en proposant les premiers nœuds intelligents communicants autoalimentés.

- La gestion et l'analyse des données massives à grande échelle. De nouveaux paradigmes émergent autour de la distribution des données ou de la parallélisation massive dans le cadre de nouveaux modèles de calcul tandis que les enjeux de qualité de donnée et d'incertitude de mesure restent centraux. D'autres voies à explorer sont le développement de méthodes d'extraction de connaissances et de fouilles de données centrées sur l'utilisateur, l'exploitation de techniques de représentation des connaissances et du raisonnement pour développer de nouvelles approches sémantiques et l'interaction entre la gestion des données et l'apprentissage automatique. L'ensemble de ces enjeux doit intégrer l'optimisation énergétique des infrastructures et des logiciels.
- L'apprentissage à partir de données de grande dimension. L'enjeu est d'étudier l'architecture de réseaux pour l'analyse de données capteurs denses et hétérogènes, pour le traitement d'images de très haute résolution, de séries temporelles, et pour l'apprentissage multimodal.

Ces développements seront mutualisés pour leur utilisation par les autres défis du site clermontois sur les infrastructures du mésocentre Clermont-Auvergne. Ils seront aussi intégrés à l'offre de services vers les autres acteurs de la région dans le cadre du projet CINAURA.

2. LES RETOMBEES SCIENTIFIQUES, TECHNIQUES ET ECONOMIQUES

Le projet a vocation à développer des solutions innovantes qui feront l'objet de publications scientifiques dans les revues d'Electronique, d'Informatique et d'Intelligence Artificielle. Son succès permettra d'irriguer tous les domaines applicatifs du site qui bénéficieront à termes des avancées du projet à travers des capteurs intelligents, des stratégies nouvelles de gestion ou d'analyse de données ou des solutions d'apprentissage de lots de données massives. Des innovations technologiques sont particulièrement attendues dans les domaines de l'e-santé, de l'étude de la biodiversité, de la surveillance des volcans actifs et de l'énergétique.

Sur le plan technique, il rassemble les laboratoires de recherche en informatique, en mathématique, en physique, en sciences de l'ingénieur, en agroenvironnement, en santé et Science de la Terre, laboratoires concepteurs de solutions innovantes en big data et/ou producteurs, consommateurs de données massives. Les échanges de compétences et de savoir-faire entre les personnels feront du site clermontois un site d'excellence sur les technologies numériques en Auvergne dans les domaines de l'internet des objets, l'intelligence artificielle et la science des données.
Sur ces quatre volets extrêmement porteurs sur le plan économique, le projet va permettre de développer puissamment les partenariats déjà actifs avec les start-up et les PMI/PME de la région et apporter de nouvelles opportunités, notamment dans le domaine de l'énergétique.
Les travaux précédents autour de la plate-forme instrumentale de réseau de capteurs sans fil développée sur le site clermontois ont démontré que la thématique était riche en possibilité de

valorisation. Actuellement ces travaux ont en effet conduit au dépôt d'une licence logicielle et d'une déclaration d'invention concernant la partie réseau du système. Un projet de prématuration sur le nœud capteur sans fil est également en cours de dépôt.

Les plates-formes d'expérimentation du capteur au cloud et celle dédiée à la performance énergétique d'objets connectés seront mise à disposition pour les formations et ouvertes aux entreprises dans le cadre de prestations et de projets de recherche dans une optique de transfert de compétences et d'expertise vers l'écosystème des entreprises régionales. Un tel transfert a déjà été entrepris dans le secteur de l'agriculture dans le contexte du projet PIA3 AgriLITech porté par le groupe Limagrain.
Le mésocentre, hébergé dans le datacenter de l'Université Clermont Auvergne, mutualise pour l'ensemble de la communauté scientifique du site auvergnat des ressources humaines, des équipements de calcul et de stockage, et des logiciels et applications scientifiques. Il participe à la démarche de labellisation des datacenters avec les sites de Lyon et de Grenoble (projet CINAuRA).

3. L'INSCRIPTION DU PROJET DANS UN AXE OU LES AXES PRIORITAIRES RETENUS PAR L'ETAT ET/OU LA REGION

Le projet s'inscrit au cœur des enjeux du numérique. Des retombées sont également attendues dans les autres domaines que sont la santé, l'agriculture et l'environnement.
Il s'inscrit dans la priorité nationale du numérique (et tout particulièrement l'intelligence artificielle) avec des retombées attendues sur certains enjeux de la transition énergétique (autonomie des objets connectés) et de la transition écologique (surveillance de la biodiversité).

4. LA COHERENCE AVEC LES ORIENTATIONS STRATEGIQUES REGIONALES DE RECHERCHE ET D'INNOVATION ET AVEC LA STRATEGIE DE SITE

Le projet s'appuie sur l'axe transverse « Instrumentation » et le domaine en structuration « Intelligence artificielle » de la politique de site. Il est co-construit avec le programme de développement instrumental du projet I-Site CAP2025 (https://cap2025.fr), devenu programme DATA, qui a grandement contribué à poser les fondations de la plate-forme de réseau de capteurs opérationnelle aujourd'hui avec comme objectif de collecter, transférer, intégrer et exploiter des ensembles massifs de données agro-environnementales.
Plus globalement, ce programme, adresse des problématiques de recherche fondamentale et appliquée liées aux objets connectés et au Big-Data dans les domaines de l'e-agriculture, de l'e-santé, de la mobilité et de la surveillance des volcans, au cœur des 4 challenges scientifiques du projet I-site CAP2025.

5. LES MODALITES DE PARTENARIATS ET LES ASPECTS ORGANISATIONNELS DU PROJET.

Le projet IDEAL sera structuré autour de quatre actions thématiques et d'une action transversale, en forte interaction :

- A1 : Intelligence distribuée.
- A2 : Autonomie énergétique des objets connectés.
- A3 : Gestion et analyse de données massives.
- A4 : Apprentissage à partir de données de grande dimension.
- A5 : Optimisation énergétique de l'infrastructure de calcul et des applications informatiques (action transversale)

Les actions sont en charge de la définition et de la mise en œuvre du programme scientifique dans leur périmètre thématique.
Gouvernance : Le comité de pilotage est composé des porteurs du projet, de l'animateur du programme DATA du projet I-Site CAP2025, des responsables des actions et des directeurs des laboratoires partenaires (ou leurs représentants). Pour conserver un ancrage transversal en lien avec les différents domaines d'application du site, une version élargie du comité de pilotage intégrera les représentants des autres projets du CPER 2021-2026 ainsi que ceux des challenges du projet I-Site CAP2025.

Ce projet, ou un projet proche, a-t-il été soumis pour PIA, au CPER 2015-2020, à un financement national (ANR, ADEME, autres…), aux Fonds européens ?	**Non**
Ce projet est-il la suite, pour tout ou partie, d'un ou plusieurs projets soumis à PIA, au CPER 2015-2020, à un financement national (ANR, ADEME, autres…), aux Fonds européens ?	OUI Acronymes des projets : Coordinateurs AUDACE : Vincent Breton PIA ISITE/CAP2025 : programme transverse. Instrumentation et Big Data : Dominique Pallin

PLAN DE FINANCEMENT PREVISIONNEL

Les moyens envisagés ont été pensés de manière globale de façon à accompagner les développements scientifiques et technologiques du projet IDEAL. Des interactions fortes apparaissent avec les 4 Centres Internationaux de Recherche (CIR) du projet I-Site CAP2025 :

- CIR 1 : projet Fenomenes (en lien avec les capteurs et outils numériques pour les agroécosystèmes) ;
- CIR 2 : projet MODE (en lien avec les thèmes « Robotique mobile et usages associés » et « interaction, simulation et aide à la décision »),
- CIR 3 : projet BIOTIC (en lien avec la bio-informatique et la santé connectée),
- CIR 4 : projet 3R (en lien avec les questions liées à l'évaluation quantitative des risques naturels et à leur gestion).

Équipement : plan de financement prévisionnel :

- 2023 : plate-forme de développement de la performance énergétique d'objets connectés (150 k€)
- 2023 : dispositif de mesure des performance énergétiques de l'infrastructure de calcul et des applications informatiques (50k€)
- 2023 : plate-forme de capteurs intelligents et infrastructure informatique pour les services numériques adossés - phase 1 (100 k€)
- 2024 : plate-forme distribuée pour l'acquisition et la gestion de données massives - phase 1 (100k€).
- 2025 : plate-forme de capteurs intelligents et infrastructure informatique pour les services numériques adossés - phase 2 (280 k€).
- 2026 : plate-forme distribuée pour l'acquisition et la gestion de données massives - phase 2 (100k€).
- 2027 : ferme de calcul pour l'apprentissage profond à grande capacité mémoire (160 k€).

Les équipements envisagés concernent 4 catégories de besoins :

- En lien avec le champ d'investigation 1 (intelligence distribuée), il est prévu la réalisation d'une plateforme sur le déploiement de capteurs intelligents dont la conception et le monitoring feront l'objet de travaux de recherche dans le cadre de ce projet.
- En support des activités de recherche liées au champ d'investigation 2 (autonomie énergétique des objets connectés) Des équipements mutualisés de caractérisation énergétiques à très faible niveau des systèmes IOTs sont prévus. Ils permettront ainsi la conception de nouveaux objets

connectés plus performant et soutenant ainsi les nombreuses thématiques de recherches en lien.

- En support des activités de recherche liées au champ d'investigation 3 (gestion et l'analyse des données massives à grande échelle), il est prévu l'acquisition d'une plate-forme distribuée pour la gestion de données massives. La plateforme constitue un objet de recherche en soit et se doit donc d'être distribuée, modulable et reconfigurable à souhait. Par ailleurs, un ensemble de capteurs sera acquis pour analyser les performances optimisation énergétiques de l'infrastructure de calcul du mésocentre et des applications informatiques.
- En support des travaux de recherche du champ 4 (apprentissage à partir de données de grande dimension), il est prévu l'acquisition d'une ferme de calcul pour l'apprentissage profond nécessitant des serveurs à grande capacité mémoire et équipés de processeurs à calcul accéléré.

Les ressources humaines nécessaires à l'acquisition, la mise en service et l'opération des équipements acquis dans le cadre du projet IDEAL seront fournies par le personnel permanent du mésocentre et des laboratoires partenaires (LIMOS, LPC). Ces équipements serviront de support à des formations.

Année du début de l'investissement : 2022

Nature de Dépenses*	Montants	Financement sollicité auprès de la région	Co-financement	Année d'acquistion de l'équipement
Optimisation énergétique des infrastructures de calcul et des logiciels	50k€	50k€		2023
Plateforme gestion de données	200k€	200k€		2024 - Phase 1 : 100k€ 2026 - Phase 2 : 100k€
Plate-forme de mesure performance énergétique d'objets connectés	200k€	150	50k€ (CNRS IN2P3)	2023
Capteurs intelligents et infrastructure adossée	330k€	280k€	50k€ (CNRS IN2P3)	2023- Phase 1 : 100k€ 2025 - Phase 2 : 280k€
Apprentissage automatique	160k€	160k€		2027
Total	**940k€**	**840k€**	**100k€**	

Répartition des financements par année

Année	**Montant total**
2023	300k€
2024	100k€
2025	280k€
2026	100k€
2027	160k€
Total	**940K€**

Synthèse des financements demandés à la région

Financeurs	Financement global sollicité	Nature de Dépenses*	Montants
Etat /Région	840k€	Optimisation énergétique de l'infrastructure de calcul et des applications informatiques	50k€
		Plateforme gestion de données	200k€
		Plate-forme de mesure performance énergétique d'objets connectés	150k€
		Capteurs intelligents et infrastructure adossée	280k€
		Apprentissage automatique	160k€
Total	**840k€**		**840k€**