Projet Scientifique de l'équipe PAMDA
LIFO
Université d'Orléans - INSA CVL

Axes de recherche

Les axes de recherche de notre équipe sont **les bases de données** et le **le parallélisme**. L'axe **bases de données** se focalise sur les données intelligentes avec deux buts majeurs

1. générer des bases de données graphes à partir des données textuelles,
2. interroger et manipuler ces bases de données graphes de manière intelligente, offrant à l'utilisateur un système d'interrogation qui intègre l'analyse de données et la prédiction.

L'axe **parallélisme** mène des recherches autour du parallélisme implicite et du traitement de grandes masses de données en utilisant différents modèles (MapReduce, patterns parallèles spécialisés). Les travaux de cet axe se focalisent sur les outils d'abstraction du parallélisme dans le but

1. d'offrir à des utilisateurs non spécialistes des interfaces de programmation séquentielles pour décrire des traitements de gros volumes de données,
2. de paralléliser le traitement de ces grandes masses de données par la mise en place de nouvelles approches garantissant un traitement efficace en terme de temps d'exécution, de consommation d'énergie, d'équilibrage de charges et de coûts de communication sur des systèmes à grande échelle.

L'originalité de l'équipe peut se résumer par son positionnement sur un parallélisme efficace et accessible aux non spécialistes et l'assurance d'utiliser des données de qualité.

Dans le cadre de l'axe bases de données l'équipe participe à l'action DOING qui met en place des collaborations entre chercheurs en bases de données, traitement automatique des langues et intelligence artificielle ayant la santé comme domaine d'application. DOING aborde une interaction directe entre la recherche en TAL et les bases de données, c'est-à-dire, une abstraction par étapes, du bas vers le haut, de l'instance (des données textuelles) vers le schéma d'une base. L'action s'intéresse aussi aux questions des requêtes prédictives sur graphe, pour établir une intégration entre requêtes et méthodes d'apprentissage automatique.

Projet Scientifique

Notre projet scientifique s'inscrit dans une extension de l'action DOING pour répondre aux enjeux du HPDA (*High Performance data analytics*). En effet, lorsque les graphes deviennent grands et même trop grands, les algorithmes utilisés pour les traiter, les explorer et les analyser deviennent coûteux en termes de temps d'exécution et de coût de communications. En général, le traitement distribué de grands graphes est une tâche difficile en raison de la taille et de la structure irrégulière inhérente aux calculs sur graphes.

L'action DOING via son extension **DOINGbyHPDA** s'intéresse aux conditions d'exécution des **requêtes data science sur des grands graphes**. Seule l'exécution de ces requêtes sur des **architectures haute performance et *exascale*** peut permettre de traiter l'extraction et l'analyse de connaissances sur des grands graphes (par exemple dans l'étude de cas cliniques

liés à des maladies ou l'analyse de phénomènes environnementaux, etc.). Dans ce cadre il est nécessaire de lever les verrous suivants :

1. Concevoir un système d'interrogation *déclaratif* sur deux aspects principaux :
   — la demande d'analyse d'un utilisateur qui serait exprimée de manière implicite dans une requête, pour ensuite être traduite automatiquement en un workflow ;
   — la stratégie de parallélisation des analyses demandées par ces requêtes qui serait exprimée via des outils offrant un minimum d'effort de programmation. Ces outils dans le cadre du parallélisme implicite s'appuieraient sur des patterns à définir et dédiés aux traitements de gros graphes de données.

   L'Exascale permet aux scientifiques d'accéder à des moyens (puissance de calculs, capacité de stockage) capables de traiter leurs données dans leur ensemble sans faire d'approximations. Cependant, la conception et le déploiement de ces applications sur une architecture qui permet l'exascale n'est pas une tâche facile surtout pour un scientifique non informaticien. **Il est essentiel de fournir à ce scientifique les outils qui lui permettent d'interroger ces données de manière déclarative.**
2. Permettre le passage à l'échelle de l'exécution des jointures (corrélations par similarité) sur des données massives par la conception de stratégies d'équilibrage de charges et par la mise en place de techniques de redistribution des données permettant de réduire les coûts de traitement et de communication.
3. Rendre possible l'exécution de workflows complexes qui combinent à la fois des requêtes, des algorithmes hétérogènes, la gestion et le traitement de divers graphes.
4. Rendre possible le traitement des graphes dynamiques qui prennent en compte des mises à jour (insertions, modifications, suppressions), de sorte que les états actuels et précédents peuvent être interrogés de manière transparente.

Notre projet vise des applications dans le domaine des sciences environnementales ou de la santé. Pour ces domaines le traitement des requêtes data science sur de grands graphes permettra de révéler l'évolution caractéristique des modèles de réponse. Par exemple sur des données liées à la santé, cela permettra de confronter des politiques de traitement des maladies en comparant des cas cliniques et des données sociales concernant l'adoption de certains traitements par des groupes de malades des milieux culturels et socio-économiques variés dans des pays ou des régions différentes.

Notre projet scientifique a pour **objectif sur 5 ans** de concevoir et développer notre framework **DoingHPDA** de requêtes data-science sur de grands graphes permettant l'exécution des traitements sur une architecture exascale. Pour les deux premières années, il s'appuie sur la poursuite des travaux de recherche menés dans les 2 axes pour concevoir et développer les briques nécessaires au framework

— Pour l'axe parallélisme il s'agit de définir les patterns de traitement adaptés à la manipulation et aux traitements des données sous forme de grands graphes, de proposer des parallélisations efficaces qui répondent aux contraintes de l'exascale mais qui tiennent compte de la nécessiter de réduire la consommation d'énergie. Il s'agit également de définir les outils d'aide à la conception du *workflow* afin d'être capable à partir de la requête initiale de proposer automatiquement les patterns pertinents aux différentes étapes de ce *workflow*.
— Pour l'axe bases de données il s'agit de travailler dans la mise en place d'un système d'interrogation et manipulation des bases de données graphes, dont le langage de requêtes (déclaratif) engloberait une analyse prédictive - terme qui combine la gestion des données, l'apprentissage automatique et l'optimisation. Il existe une demande croissante de tels outils pour le traitement des problèmes de sciences des données. Ce travail comprend des aspects théoriques et pratiques, aussi bien sur les couches hautes du framework, pour concevoir et optimiser des *workflows*, vus comme les plans d'exécution des requêtes, que

sur les couches basses en interaction avec l'axe parallélisme, pour la mise en place de ces *workflows*.

À partir de ces briques il s'agira de concevoir et de développer notre framework permettant aux scientifiques d'interroger leurs données sans se préoccuper de la parallélisation des traitements et du déploiement des données sur la machine. Pour cette partie de notre projet sur 3 ans au delà du développement à réaliser, des verrous restent à lever sur la gestion des données et la performance des traitements

- définir une structure de données permettant une répartition intelligente
- définir les méthodes de répartition des données en fonction du workflow décrit par l'utilisateur et pour optimiser les différents traitements.
- intégrer des critères d'optimisation en vue d'obtenir un workflow économe en terme de consommation d'énergie.

Un objectif majeur sur la dernière année de notre projet est de construire une collaboration afin de valider notre framework sur une application réelle impliquant de gros volumes de données. Une collaboration avec des utilisateurs permettrait de faire des tests de performances sur des données réelles et également d'ajuster nos interfaces en fonction des besoins et des retours des utilisateurs.

Les nouvelles thématiques scientifiques. Notre projet scientifique a des liens forts avec des thématiques à développer dans notre équipe pour que notre framework réponde aux enjeux sociétaux actuels. Une forte interdisciplinarité est nécessaire pour que notre framework permette de concevoir des workflows adaptés aux données étudiées. En particulier les techniques issues de l'intelligence artificielle sont des éléments essentiels d'un workflow répondant à une requête d'analyse de gros volumes de données dans de nombreux domaines scientifiques.

Comme déjà mentionné dans le rapport d'auto-évaluation, l'équipe a déjà des liens avec ces thématiques comme la thèse sur la parallélisation implicite de très larges réseaux de neurones ou encore de récentes collaborations multidisciplinaires avec des chercheurs en TAL (traitement automatique des langues) et IA (intelligence artificielle) sur la transformation de données en information puis en connaissances. Ces thèmes de recherche doivent être étendues au sein de notre équipe dans l'objectif de spécialiser notre framework DoingHPDA aux applications actuelles. Dans ce contexte il est important que l'équipe s'agrandisse pour l'extension de ses travaux vers ces enjeux du HPDA soit en amont du framework avec la gestion et la préparation des données soit sur la conception et la génération automatique du workflow avec les briques nécessaire à la mise en place de méthodes d'analyse issues de l'IA.

Forces

Pour développer son projet, PAMDA compte déjà avec certains moyens issus des actions en cours ou des collaborations nationales et internationales.

- *Action parallélisme :* HITL du RTR-DIAMS
- *Action DOING :* GT du RTR-DIAMS, action du GTR MADICS, projet APR-IA
- *Collaborations nationales :* LLL ; LIFAT ; LIRIS ; LISN-Paris Saclay ; BRGM ; Ennov et Huawei (entreprises)
- *À l'international :* workshop DOING associé à la conférence ADBIS depuis 2020, collaborations avec le Brésil (UFPR ; UFRN ; 1ers contacts avec HU-UFJF et USP)
- *Lien enseignement/recherche :* participation au nœud français du projet EUMaster4HPC

De plus, l'équipe peut également s'appuyer sur ses expériences dans les domaines des géosciences (anciens projets et 3 thèses) et de la santé.

Pour progresser, PAMDA a comme objectif de continuer et d'élargir ses collaborations nationales et internationales. PAMDA est déjà collaborateur dans deux grands projets institutionnels brésiliens, le PrInt (programme d'internationalisation des universités brésiliennes) qui correspond à un vaste programme de la Capes (Coordination pour l'amélioration du personnel de

l'enseignement supérieur), au Brésil. Dans ce cadre, auprès de l'UFRN [1](*Universidade Federal do Rio Grande do Norte*), PAMDA va poursuivre ses collaborations avec des chercheurs de cette université, avec, comme dans le passé récent, des co-directions de thèses. Auprès de l'UFPR [2]. (*Universidade Federal do Paraná*), PAMDA compte avec la visite de chercheurs brésiliens pour des durées allant de 1 à 3 mois. En effet, l'équipe a l'intention de s'investir davantage dans les collaborations liées à l'action DOING mais les élargir pour prendre en compte d'autres aspects de son projet scientifique, DOINGbyHPDA.

L'équipe appuie et s'investit dans les projets institutionnels de l'Université d'Orléans. Elle reconnaît le besoin d'avancer encore plus dans cette direction. PAMDA est sensible à la politique de l'établissement de renforcer les liens entre les formations et la recherche. Dans ce contexte, des membres de l'équipe participent régulièrement aux enseignements du Master IMIS du pôle informatique. Soit dans les modules dédiés à l'initiation à la recherche soit dans les modules liés au parallélisme. Nous avons également très souvent encadrés des étudiants pour leur stage de M2. En outre, depuis la mise en place de GSON (DU Data Sciences) à laquelle nous avons fortement participé, nous contribuons toujours avec entre autres deux modules sur le parallélisme et le BigData. Dans le cadre du projet MINERVE (PIA4 ExcellencES) les thématiques de recherche de l'équipe se retrouvent complètement dans le pôle thématique "sciences et technologies augmentées par les data sciences" que l'établissement souhaite créer. C'est une opportunité importante pour notre équipe de s'investir dans de nouveaux enseignements en lien avec nos axes de recherche en proposant aux étudiants soit des projets de recherche en lien avec notre framework soit de les intégrer à notre équipe pour la formation par la recherche.

1. `https://print.ufrn.br/`
2. `http://www.prppg.ufpr.br/site/print/pb/print-ufpr/`