Vast and Fast Data in the era of the large astrophysics and particle physics experiments

Gazagnes, Simon

DOI:
10.33612/diss.179743481

RÉSUMÉ

Dans ce chapitre, je présente un résumé, en français, des travaux réalisés dans le cadre de cette thèse. Ce résumé détaille les principaux résultats obtenus de manière simplifié afin qu'ils soient accessibles à un public sans expertise scientifique (et qui comprend la langue de Molière).

Depuis le début du troisième millénaire, nous sommes entrés dans l'ère de l'Information, une ère dominée par les vastes volumes de données et d'informations générés par nos sociétés modernes. Cette nouvelle ère a des conséquences importantes pour la recherche, dont les progrès reposent désormais essentiellement sur des expériences scientifiques a la pointe de la technologie. Les données collectées par ces expériences sont complexes à traiter pour plusieurs raisons. Dans cette thèse, nous avons principalement fait face a des problemes de *vast* et *fast data*, c'est-à-dire des défis liés a des vastes volumes de données, ou à des systemes d'acquisition extremement rapide qui nécessite des systemes de traitement capable d'analyser les données collectées en temps réel. L'extraction et l'analyse du contenu informationnel de ces données est l'un des principaux défis que nous devons relever pour pouvoir ouvrir la voie vers de nouvelles découvertes scientifiques.

Pour faire face a ces défis, la communauté scientifique travaillant sur ces aspects s'est développée à un rythme incroyable au cours des dernières années afin de fournir des solutions adaptées pour surmonter les obstacles auxquels seront confrontées les futures expériences scientifiques de pointe. Les projets interdisciplinaires, regroupant des scientifiques avec une expertise dans divers domaines comme les mathématiques, l'informatique, la physique et la biologie, constituent des approches modernes pour relever les nouveaux défis liés a ces nouvelles experiences scientifiques. Cette thèse en est un exemple: le projet [VF]ast data, financé par une subvention du *Centre for Data Science and Systems Complexity* (DSSC) de l'Université de Groningue, a été proposé afin de bénéficier des derniers développements

en morphologie mathématique, une branche en pleine expansion dans le domaine du traitement des images et des signaux, dans le but de fournir de nouveaux outils nous permettant de traiter et d'analyser de manière optimale les données collectées par plusieurs projets d'experiences scientifiques en astrophysique et en physique des particules.

Ce projet de doctorat interdisciplinaire a donné lieu à cinq sous-projets scientifiques portant sur différents aspects. Dans les **Chapitres 2** et **3**, j'ai présenté deux articles qui détaillent un nouvel outil de traitement d'images et de signaux, DISCCOFAN (DIStributed Connected COmponent Filtering and ANalysis), basé sur les développements récents dans le domaine de la morphologie mathématique. Cet outil fournit une méthode efficace pour analyser les structures observées dans des images à deux ou trois dimensions. Plusieurs domaines tels que l'imagerie biomédicale et l'astronomie sont basées sur l'analyse des propriétés d'objets (par exemple un vaisseau sanguin ou une galaxie). La plupart des techniques disponibles dans la litterature ne sont pas optimales pour traiter ces applications. Recemment, une classe spécifique de techniques de morphologie mathématique, appelée "component trees", a été développée. Les *component trees* sont des structures hiérarchiques (sous la forme d'un arbre) qui représentent les relations imbriquées des régions connectées (c'est-à-dire les structures) dans l'image. Ces structures nous permettent d'analyser bien plus efficacement les objets dans les images. Lorsque j'ai commencé ce doctorat, l'application de ces techniques était limité à des images de taille relativement modeste, jusqu'à quelques gigapixels (quelques milliards de points). Cependant, elles ne pouvaient pas traiter des images avec des centaines ou des milliers de milliards de pixels (des images gigantesques), car la manipulation de ces données nécessite une puissance de calcul gigantesque sur une seule machine et aucune machine dans le monde ne dispose de telles capacités. Profitant des progrès du calcul parallèle, nous avons développé DISCCOFAN, un nouvel outil de calcul qui permet d'utiliser les *component trees* pour traiter d'énormes images a deux et trois dimensions. DISCCOFAN utilise un systeme de parallelisation qui distribue les operations sur plusieurs "noeudsïndépendants qui accomplissent les tâches simultanément. Il s'agit donc d'un outil prometteur pour traiter les images et les volumes de données tres volumineux qui seront collectés par les prochaines générations d'expériences scientifiques dans divers domaines.

Les **Chapitres 4** et **5** présentent deux projets astrophysiques axés sur l'étude des processus physiques qui ont régi l'evolution de notre

Univers pendant “l’époque de la réionisation”, une transition importante qui s’est produite au cours du premier milliard d’années et au cours de laquelle les atomes d’hydrogene dans l’Universe ont été ionisés par le rayonnement emis par les premières étoiles et galaxies. Cette époque a des implications majeures pour comprendre l’Univers tel que nous le connaissons aujourd’hui, notamment parce que la première génération d’objets astronomiques s’est formée pendant cette période. Comme elle s’est produite il y a des milliards d’années, il est extrêmement difficile d’obtenir des observations sur ce qu’il s’est passé a cette epoque. Par conséquent, bien qu’il s’agisse d’une étape critique dans l’histoire de l’Univers, nous n’avons encore que trop peu d’informations sur les processus physiques qui ont régi cette époque. Cependant, plusieurs télescopes, actuellement en construction, nous permettront d’observer les premières étoiles et galaxies qui ont existé pendant cette période. Ces observations nous aideront à répondre à des questions fondamentales liées à la formation des premiers objets astronomiques et à leur impact sur l’évolution de notre Univers.

Dans le **Chapitre 4**, j’ai analysé les propriétés de galaxies, proches de nous, qui ont des propriétès similaires aux premières galaxies qui ont été formées dans l’Univers. Le but de cette analyse étant d’établir un modèle théorique cohérent qui pourra être utilisé pour interpréter efficacement les futures observations des premières galaxies qui existaient pendant l’époque de la réionisation. Cette analyse a fourni des informations précieuses pour comprendre l’impact de ces objets sur l’evolution de l’Universe à cette époque. Le prochain *James Webb Space Telescope* et les très grands télescopes terrestres a venir nous fourniront des observations révolutionnaires pour évaluer les résultats de cette étude.

Le **Chapitre 5** s’appuie sur une approche differente, appelée “observations à 21 cm”, qui permet d’étudier de manière similaire les processus physiques qui ont régi l’époque de la réionisation. Plutôt que d’observer directement des objets astronomiques tels que des étoiles ou des galaxies, les observations à 21-cm offrent un moyen unique d’explorer l’Univers au cours du premier milliard d’années en se basant sur l’évolution du gaz entre les étoiles et les galaxies. La formation des premiers objets astronomiques a eu un impact crucial sur l’évolution des charactéristiques morphologiques et topologiques des régions de gaz dans l’Univers, de sorte que nous pouvons utiliser les observations à 21 cm pour analyser les propriétés de ces sources. En utilisant la méthode présentée dans les deux premiers chapitres, j’ai étudié la morphologie du gaz à ces époques à l’aide de simulations. J’ai montré que cette approche devrait fournir des informations précieuses sur les phenomènes astrophysiques qui ont pris place à cette époque. Ce travail est particulièrement pertinent dans le contexte du *Square Kilometer*

Array, un radiotélescope à venir qui recueillera des observations à 21-cm révolutionnaires.

Enfin, le **Chapitre 6** détaille le dernier projet réalisé dans le cadre de cette thèse. Ce projet s'est concentré sur la conception d'un algorithme efficace pour identifier les trajets de particules fondamentales dans des experiences scientiques basées sur des accélérateurs de particules. Ces derniers utilisent des collisions de particules fondamentales (par exemple, protons, neutrons, électrons) pour sonder les propriétés des éléments fondamentaux de la matière et la physique régissant les très petites échelles. Afin de détecter et d'analyser des particules très rares, les expériences d'accélérateurs doivent fonctionner à des taux d'interaction extrêmement élevés, avec des milliards de milliards de collisions de particules par seconde. La quantité de données produites par ces expériences étant trop importante pour être stockée, elle nécessite des systèmes de traitement des données en temps réel qui permettent de sélectionner les interactions les plus pertinentes. Ces systèmes utilisent un algorithme pour identifier les trajectoire de particules en temps réel à partir des données collectées par les detecteurs embarqués dans les accélérateurs. La reconstruction des trajectoires des particulers permet d'extraire leurs propriétés clés. Dans ce dernier chapitre, nous avons conçu un algorithme de reconstruction des trajectoires de particules pour une experiences scientifique à venir qui fonctionnera à des taux d'interaction très élevés (donc, collectant des données de maniere extrêmement rapide). Nous avons montré que cette méthode est particulièrement rapide, et donc prometteuse pour reconstruire les trajectoires des particules en temps reel.

Dans l'ensemble, cette thèse interdisciplinaire a fourni de nouvelles méthodes d'analyse de données pour relever les défis liés aux futures expériences scientifiques de pointe en astrophysique et en physique des particules. Plus important encore, elle ouvre la voie à de futures travaux, étendant ces recherches dans diverses directions. Enfin, cette thèse souligne que les progrès actuels et à venir liés aux techniques d'analyse de données permettront des découvertes clés, fournissant des informations cruciales pour répondre à des questions fondamentales dans divers domaines scientifiques. Par conséquent, bien que cette période revolutionnaire aura des défis critiques à surmonter, l'avenir est prometteur pour les decouvertes scientifiques qui émergeront de ces futures expériences scientifiques.