Fiche de lecture – The Seattle Report on Database Research[1]

Voici cinq défis en lien avec le big data, présents dans le rapport et listés par ordre d'importance. On notera que, bien souvent, les défis sont interconnectés entre eux et ne représentent qu'une partie des défis présentés dans le rapport. D'une certaine manière, tous les challenges cités dans le rapport sont tout autant importants les uns que les autres.

Le défi de la gestion des systèmes de données à grande échelle et des systèmes distribués complexes

En raison du volume et de la vélocité croissants des données, de leur type pouvant être structuré, semi-structuré ou non structuré, les données deviennent difficiles à prévoir et à stocker. Pour surmonter ce défi, les systèmes ont besoin de technologies efficaces, fiables et sécurisées, telles que les entrepôts de données, les bases de données distribuées, les *data lakes* et les bases de données NoSQL. Le défi consiste à concevoir, gérer et exploiter ces systèmes de manière efficace, ce qui implique souvent l'utilisation de plusieurs centres de données, du *cloud computing* et d'autres technologies avancées. *Pour approfondir* : Abadi, D. et al. (2016). The Beckman Report on Database Research. *Commun. ACM, 59*(2), 92–99. doi:10.1145/2845915

Le défi de l'intégration du big data à l'écosystème de la science des données

Ce défi consiste à rendre les bases de données plus accessibles et plus faciles à intégrer aux flux de travail de la science des données. Cela nécessite le développement de solutions de gestion des données évolutives et flexibles, capables de gérer le volume, la vitesse et la variété des données générées par les applications de science des données. En outre, une gestion efficace du cycle de vie des modèles de science des données est nécessaire, y compris le *versioning*, l'audit et le déploiement. Pour relever ce défi, il faut une collaboration étroite entre les chercheurs en bases de données et les scientifiques des données afin d'élaborer des solutions transparentes. *Pour approfondir* : Parameswaran, A. G. (2019). Enabling Data Science for the Majority. *Proc. VLDB Endow.*, *12*(12), 2309–2322. doi:10.14778/3352063.3352148

Le défi du développement de modèles d'apprentissage automatique pour le big data

Ce défi consiste à créer des algorithmes capables de traiter efficacement de grandes quantités de données et de fournir des résultats en temps réel. Cela nécessite une expertise technique et une bonne compréhension des besoins de l'entreprise. Les algorithmes doivent être capables de traiter une variété de types de données et de gérer les valeurs manquantes, la haute dimensionnalité et l'apprentissage en temps réel. L'objectif est de construire des modèles qui sont précis, évolutifs, robustes et efficaces. À terme, il faut voir les modèles comme des données à part entière que l'on pourrait requêter directement, qui en ferait un outil puissant pour l'analyse de données. *Pour approfondir* : Kara, K., Eguro, K., Zhang, C., & Alonso, G. (2018). ColumnML: Column-Store Machine Learning with On-The-Fly Data Transformation. *Proc. VLDB Endow., 12*(4), 348–361. doi:10.14778/3297753.3297756

Le défi de l'auto-tuning

L'*auto-tuning* est une technique utilisée pour optimiser les performances des systèmes de bases de données. Elle vise à ajuster les paramètres de configuration afin d'optimiser les performances pour des charges de travail spécifiques. Le processus d'*auto-tuning* est automatisé, ce qui élimine le besoin d'intervention manuelle, et peut être réalisé par des algorithmes de *machine learning*. L'*auto-tuning* permet d'améliorer les performances et l'évolutivité des systèmes de bases de données, ce qui les rend plus accessibles à un plus grand nombre d'utilisateurs et garantit qu'ils sont optimisés pour des charges de travail spécifiques, ce qui est essentiel pour fournir des services de haute qualité et répondre aux exigences des applications axées sur les données.

Le défi de la confidentialité et de la sécurité des données

La quantité d'informations sensibles collectées et analysées ne cesse d'augmenter. Il est crucial de protéger ces informations contre les violations, les accès non autorisés et les pertes, tout en respectant les réglementations en matière de protection des données. Pour relever ce défi, il faut mettre en œuvre de solides mesures de confidentialité et de sécurité, telles que des bases de données cryptées et des méthodes de transmission de données sécurisées. En outre, un aspect important de la confidentialité et de la sécurité des données consiste à s'assurer que les données sont collectées et utilisées de manière éthique, en veillant à ce que les individus aient le contrôle de leurs informations.

[1] Abadi, D. et al. (2022). The Seattle Report on Database Research. *Commun. ACM, 65*(8), 72–79. doi:10.1145/3524284