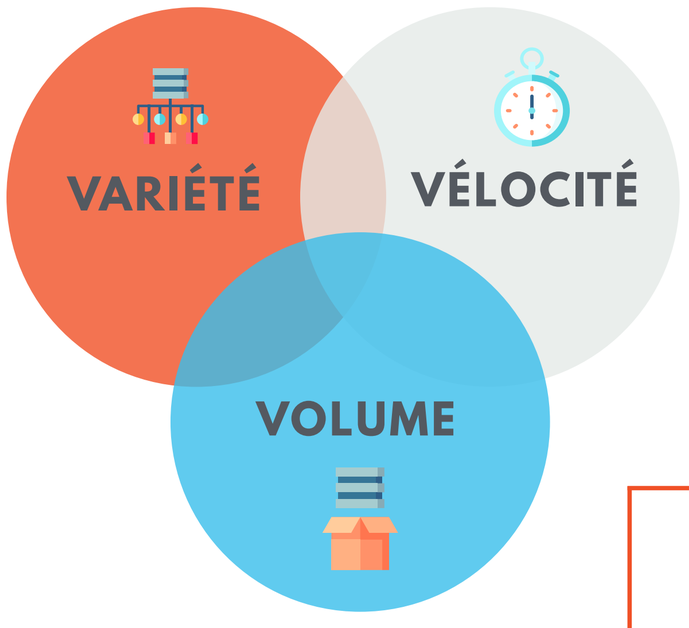
BIG DATA

Algorithmes PySpark pour traitement de données

Nous créons environ 2,5 trillions d'octets de données par jour. Ce volume de données se caractérise aussi par sa variété et sa vélocité. Ces données proviennent de différentes sources : réseaux sociaux, envoi de mails et autres applications informatique sur internet auxquels on se connecte à l'aide des ordinateurs, des smartphones et d'autres objet connectés. Il devient de plus en plus difficile de traiter ces données par les technologies classiques : le big data a donné naissance à de nouvelles technologies. L'importance de ces masses de données vient de la valeur qu'on peut en extraire : le business de la data est valorisé à plusieurs milliards de dollars.



Data Science

La Data-Science consiste à collecter, nettoyer et analyser des données hétérogènes pour en extraire de la valeur. Elle aide à la prise de décision grâce à une plus grand lisibilité de ces dernières. L'analyse de données est réalisée à l'aide d'outils mathématiques et informatiques.

Hadoop

Hadoop est l'une des premières technologies big data. Il consiste à traiter les données en parallèle (plusieurs machines formant un réseau qui collaborent à la résolution du même problème). Hadoop est aussi un système de stockage de données.

Qu'est ce que Apache Spark

Framework de traitements Big Data open source réalisé pour effectuer des analyses sophistiquées et conçu pour la rapidité et la facilité d'utilisation. Spark apporte également des améliorations à MapReduce grâce à des étapes de shuffle moins coûteuses.

Framework

Spark propose un framework complet et unifié pour subvenir aux besoins de traitements Big Data pour divers jeux de données, divers par leur nature aussi bien que par le type de source.

MILib

Spark offre également une librairie de machine learning contenant tous les algorithmes et utilitaires d'apprentissage classiques, comme la classification, la régression, le clustering , le filtrage collaboratif, la réduction de dimensions, en plus des primitives d'optimisation sous-jacentes.

PySpark

Pour réaliser ce projet nous allons utiliser l'API de Spark Python appelée PySpark. Cet API nous permet d'utiliser toute la puissance de Spark avec le langage de programmation Python.

Applications

Il existe de nombreux algorithmes spécifiques à l'analyse de données. Nous en avons étudié quelques-uns tout au long du projet.

Régression Linéaire

En statistiques la régression dite linéaire est un modèle de regression qui cherche à établir une relation linéaire entre plusieurs variables.
Cela nous permet par exemple d'observer et trouver une relation entre le poids et la taille d'une population. Ce modèle est très souvent utilisé pour faire de la prédiction ou encore expliquer un phénomène.

K-means

K-means (ou K-Moyennes) est une méthode de partitionnement de données. Étant donné un ensemble de données et un entier *k*, l'objectif est de répartir les données en *k* groupes, appelés clusters, de façon à minimiser une certaine fonction. K-means doit minimiser la distance entre les points à l'intérieur de chaque cluster. Cet algorithme est utilisé par certains logiciels pour diviser un groupe hétérogène de données en sous-groupes plus homogènes.

Text Minning

Application des méthodes de data mining aux données textuelles. Une application importante est la classification des documents (articles de presses, tweets…) en fonction de leur contenu.

TF.IDF

Le TF-IDF est une méthode de pondération souvent utilisée en recherche d'information. Son rôle est d'évaluer l'importance d'un terme dans un document par le rapport entre la fréquence du terme dans le document et le logarithme de l'inverse de la proportion de documents qui contiennent le terme.

Classification de notes

Étant donné un ensemble de notes d'informatique, nous devions écrire un algorithme qui récupérait ces notes d'un fichier pour ensuite appliquer différents algorithmes d'analyse de données sur cet ensemble. En appliquant K-means avec $k = 2$, nous pouvons extraire deux groupes d'élèves, les meilleurs et les moins bons. Nous avons donc appliqué plusieurs algorithmes pour essayer d'en tirer des conclusions quant à leur significations.

Conclusion

Tout au long de ce projet nous avons été amenés à étudier le principe de Big Data et les enjeux liés à la maîtrise de celui-ci. Le traitement de données notamment grâce à des algorithmes permet d'en extraire une certaine valeur marchande. Nous avons donc été amenés à étudier certains de ces algorithmes tels que K-Means ou encore TF.IDF pour comprendre les différents types de traitement de données. Ces données sont en perpétuelle évolution. C'est pour cela qu'elles sont si importantes aujourd'hui et qu'il est décisif de savoir les traiter.

EISTI

David RIGAUX
Mehdi DALAA
Mandry MBUNDU
Maxime LUNDQUIST