Spark et Hadoop pour l'analyse des Big Data

Module 1 Glossaire : Qu'est-ce que le Big Data ?

Bienvenue ! Ce glossaire alphabétique contient de nombreux termes utilisés dans ce cours. Ce glossaire complet comprend également des termes supplémentaires reconnus par l'industrie qui ne sont pas utilisés dans les vidéos de cours. Ces termes sont essentiels pour que vous puissiez les reconnaître lorsque vous travaillez dans l'industrie, participez à des groupes d'utilisateurs et à d'autres programmes de certification professionnelle.

Temps de lecture estimé : 12 minutes

Terme	Définition
Apache Spark	Un framework d'application open source en mémoire utilisé pour le traitement de données distribuées et l'analyse itérative de grands ensembles de données.
Apache HBase	Un magasin de données NoSQL robuste qui gère efficacement les ressources de stockage et de calcul indépendamment de l'écosystème Hadoop.
Intelligence d'affaires (BI)	Comprend divers outils et méthodologies conçus pour convertir efficacement les données en informations exploitables.
Big Data	Ensembles de données dont le volume, la vitesse ou la variété dépassent la capacité des bases de données relationnelles classiques à gérer, capturer et traiter efficacement avec une latence minimale. Les principales caractéristiques du Big Data sont un volume substantiel, une vitesse élevée et une grande variété.
Analyse de Big Data	Utilise des techniques d'analyse avancées sur des ensembles de données volumineux et diversifiés, comprenant des données structurées, semi-structurées et non structurées provenant de différentes sources et tailles, allant des téraoctets aux zettaoctets. Il aide les entreprises à tirer des enseignements des données collectées par les appareils IoT.
Outils de programmation Big Data	Les outils de programmation sont le dernier composant des outils commerciaux de Big Data. Ces outils de programmation effectuent des tâches analytiques à grande échelle et opérationnalisent le Big Data. Ils fournissent également toutes les fonctions nécessaires à la collecte, au nettoyage, à l'exploration, à la modélisation et à la visualisation des données. Parmi les outils populaires que vous pouvez utiliser pour la programmation, citons R, Python, SQL, Scala et Julia.
Commitant	La plupart des projets open source disposent de processus formels de contribution au code et incluent différents niveaux d'influence et d'obligation envers le projet : contributeur, contributeur, utilisateur et groupe d'utilisateurs. En règle générale, les contributeurs peuvent modifier le code directement.
Cloud computing	Permet aux clients d'accéder à l'infrastructure et aux applications via Internet sans nécessiter d'installation et de maintenance sur site. En exploitant le cloud computing, les entreprises peuvent utiliser la capacité du serveur à la demande et évoluer rapidement pour gérer les exigences informatiques étendues du traitement de grands ensembles de données et de l'exécution de modèles mathématiques complexes.
Fournisseurs de services cloud	Proposer une infrastructure et un support essentiels, en fournissant des ressources informatiques partagées englobant la puissance de calcul, le stockage, la mise en réseau et les logiciels d'analyse. Ces fournisseurs proposent également un modèle de logiciel en tant que service comprenant des solutions spécifiques, permettant aux entreprises de collecter, de traiter et de visualiser efficacement les données. AWS, IBM, GCP et Oracle sont des exemples marquants de fournisseurs de services cloud.
Processus d'extraction, de transformation et de chargement (ETL)	Une approche systématique qui consiste à extraire des données de diverses sources, à les transformer pour répondre à des exigences spécifiques et à les charger dans un entrepôt de données ou un autre référentiel de données centralisé.
Hadoop	Un framework logiciel open source qui fournit un traitement distribué fiable pour de grands ensembles de données grâce à l'utilisation de modèles de programmation simplifiés.
Système de fichiers distribué Hadoop (HDFS)	Système de fichiers distribué sur plusieurs serveurs de fichiers, permettant aux programmeurs d'accéder aux fichiers ou de les stocker à partir de n'importe quel réseau ou ordinateur. Il s'agit de la couche de stockage de Hadoop. Il fonctionne en divisant les fichiers en blocs, en créant des répliques des blocs et en les stockant sur différentes machines. Il est conçu pour accéder aux données en streaming de manière transparente. Il utilise une interface de ligne de commande pour interagir avec Hadoop.
Ruche	Infrastructure d'entrepôt de données utilisée pour l'interrogation et l'analyse de données, dotée d'une interface de type SQL. Elle facilite la génération de rapports et utilise un langage de programmation déclaratif, permettant aux utilisateurs de spécifier les données qu'ils souhaitent récupérer.
Internet des objets (IoT)	Un système d'objets physiques connectés via Internet. Un objet ou un appareil peut inclure un appareil intelligent dans nos maisons ou un appareil de communication personnel tel qu'un smartphone ou un ordinateur. Ceux-ci collectent et transfèrent des quantités massives de données sur Internet sans intervention manuelle en utilisant des technologies intégrées.
Données de la machine	Désigne les informations générées par diverses sources, notamment les capteurs de l'Internet des objets (IoT) intégrés aux équipements industriels, ainsi que les blogs qui capturent le comportement et les interactions des utilisateurs.
Carte	MapReduce convertit un ensemble de données en un autre ensemble de données, et les éléments sont fragmentés en tuples (paires clé ou valeur).
MapReduce	Modèle de programme et technique de traitement utilisés dans le calcul distribué basé sur Java. Il divise les données en unités plus petites et traite les données volumineuses. Il s'agit de la première méthode utilisée pour interroger les données stockées dans HDFS. Il permet une évolutivité massive sur des centaines ou des milliers de serveurs dans un cluster Hadoop.
Bases de données NoSQL	Les bases de données NoSQL sont conçues dès le départ pour stocker et traiter de grandes quantités de données à grande échelle et prendre en charge un nombre croissant d'entreprises modernes. Les bases de données NoSQL stockent les données dans des documents plutôt que dans des tables relationnelles. Les types de bases de données NoSQL incluent les bases de données de documents pures, les magasins de valeurs clés, les bases de données à colonnes larges et les bases de données graphiques telles que MongoDB, CouchDB, Cassandra et Redis.
Logiciels open source	Non seulement la version exécutable du code est gratuite, mais le code source est également entièrement ouvert, ce qui signifie que chaque ligne de code est disponible pour que les utilisateurs puissent la visualiser, l'utiliser et la réutiliser selon leurs besoins.

Terme	Définition
Analyse des prix	Aide à comprendre la segmentation du marché, à identifier les meilleurs prix pour une gamme de produits et à effectuer une analyse des marges pour une rentabilité maximale.
Bases de données relationnelles	Les données sont structurées sous forme de tableaux, avec des lignes et des colonnes, formant collectivement une base de données relationnelle. Ces tableaux sont interconnectés à l'aide de clés primaires et étrangères pour établir des relations au sein de l'ensemble de données.
Analyse des sentiments	Utilise les conversations sur les réseaux sociaux pour obtenir des informations sur les opinions des consommateurs sur un produit. Il est utilisé pour développer des stratégies marketing efficaces et établir des liens avec les clients en fonction de leurs sentiments et de leurs préférences.
Données sociales	Elles proviennent des mentions « J'aime », des tweets et des retweets, des commentaires, des téléchargements de vidéos et des médias généraux qui sont téléchargés et partagés via les plateformes de médias sociaux les plus populaires au monde. Les données générées par les machines et les données générées par les entreprises sont des données que les organisations génèrent dans le cadre de leurs propres opérations.
Données transactionnelles	Généré à partir de toutes les transactions quotidiennes qui ont lieu en ligne et hors ligne, telles que les factures, les ordres de paiement, les enregistrements de stockage et les reçus de livraison.
Vitesse	La vitesse à laquelle les données arrivent. La vélocité est l'un des quatre principaux éléments utilisés pour décrire les dimensions du Big Data.
Volume	L'augmentation de la quantité de données stockées au fil du temps. Le volume est l'un des quatre principaux éléments utilisés pour décrire les dimensions du Big Data.
Variété	La diversité des données ou les différentes formes de données qui doivent être stockées. La diversité est l'un des quatre principaux éléments utilisés pour décrire les dimensions du big data.
Véracité	La certitude des données, comme celle des grandes quantités de données disponibles, rend difficile de déterminer si les données collectées sont exactes. La véracité est l'un des quatre principaux éléments utilisés pour décrire les dimensions du big data.
Encore un autre négociateur de ressources (YARN)	Il fait office de gestionnaire de ressources fourni avec Hadoop et est généralement le gestionnaire de ressources par défaut pour de nombreuses applications Big Data, telles que HIVE et Spark. Bien qu'il reste un gestionnaire de ressources robuste, il est important de noter que des gestionnaires de ressources basés sur des conteneurs plus contemporains, tels que Kubernetes, émergent progressivement comme les nouvelles pratiques standard dans le domaine.

Auteur(s)

- Rashi Kapoor