eBook Gratuit

APPRENEZ bigdata

eBook gratuit non affilié créé à partir des **contributeurs de Stack Overflow.**

#bigdata

Table des matières

À propos

You can share this PDF with anyone you feel could benefit from it, downloaded the latest version from: bigdata

It is an unofficial and free bigdata ebook created for educational purposes. All the content is extracted from Stack Overflow Documentation, which is written by many hardworking individuals at Stack Overflow. It is neither affiliated with Stack Overflow nor official bigdata.

Use the content presented in this book at your own risk; it is not guaranteed to be correct nor accurate, please send your feedback and corrections to info@zzzprojects.com

Chapitre 1: Démarrer avec bigdata

Remarques

Cette section fournit une vue d'ensemble de ce qu'est bigdata et pourquoi un développeur peut vouloir l'utiliser.

Les données massives sont les données caractérisées par les 4 V. Ce sont le volume, la vélocité, la variété et la véracité.

1. Volume - Lorsque la quantité de données est en volume énorme, comme des téraoctets ou des pétaoctets. Comme l'indique un rapport, nous avons généré 90% de données mondiales au cours des 2 ou 3 dernières années.
2. Vitesse - La vitesse à laquelle les données circulent dans le système. Par exemple, des millions d'utilisateurs qui téléchargent simultanément leur contenu sur des sites de réseaux sociaux génèrent des données aussi élevées que dans la plage des téraoctets / s.
3. Variété - Différents types de données basées sur sa nature. Il peut être structuré (dont la plupart des anciens SGBDR traitent), semi-structuré (courrier électronique, XML, etc.) et non structuré (vidéos, audios, données de capteurs, etc.).
4. Veracity - C'est le moyen par lequel nous obtenons un aperçu significatif de nos données disponibles. Cela peut être considéré comme l'aspect le plus important des données, car la plupart des décisions commerciales dépendent de l'utilité des données.

Hadoop Framework est la plate-forme la plus utilisée pour stocker et traiter les données volumineuses **.** Il se compose de 2 choses:

1. Système de fichiers distribués Hadoop (HDFS) - Les données sont stockées sur le système de fichiers distribués (HDFS) Hadoop, qui est en réalité un cluster de matériel de base différent de la méthode primitive de stockage sur les serveurs. Les données résident sur HDFS et cadres.
2. MapReduce (MR) - Il s'agit de la structure de traitement par défaut pour Hadoop. MapReduce (fait partie d'Apache Hadoop)

Avec l'avancement de Hadoop, de nouveaux outils de traitement ont vu le jour dans la communauté Hadoop. Quelques outils / frameworks parmi les plus populaires:

1. Apache Spark
2. Apache Storm
3. Apache Flink

   Et beaucoup plus..

Peu de mécanismes de stockage autres que le simple HDFS:

1. Ruche

2. HBase
3. Cassandra

Et beaucoup plus..

Un développeur peut être intéressé par les capacités de traitement des données massives afin que cela puisse se révéler être une différence majeure dans la façon dont nous examinons nos données. Dans un univers parallèle, nous pouvons également appeler des données volumineuses en tant que données riches non contrôlées. Nous devons maîtriser ces énormes données. Avec le big data, nous pourrions peut-être traiter le potentiel caché de données déjà existantes.

Un bon exemple peut être cité dans le comportement des clics des clients sur les sites Web d'achat où leurs vues, clics et le temps passé sur ce site Web indiquent au revendeur en ligne de se procurer des produits et d'envoyer des recommandations basées sur le comportement des utilisateurs.

Examples

Qu'est-ce que le Big Data?

Le Big Data, dans sa forme la plus élémentaire, peut être décrit comme le terme générique métalisé par différents aspects des données. Ces différents aspects sont

> Volume (quantité énorme de données), vitesse (débits de données plus importants), variété (données structurées, non structurées et semi-structurées) et véracité (prendre les bonnes décisions en fonction des données).

Ces bases de données étaient difficiles à gérer par les bases de données relationnelles du troisième âge. Un besoin de nouveau système est apparu et le traitement Big Data est venu à la rescousse. Alors que beaucoup de personnes ont une compréhension différente de ce qu'est le Big Data, voici quelques définitions des Big Data données par les leaders du secteur Data:

Définitions:

- «Les données massives dépassent la portée des environnements matériels et des outils logiciels couramment utilisés pour les capturer, les gérer et les traiter dans un délai raisonnable pour les utilisateurs.» (Article de Teradata Magazine, 2011)
- «Le Big Data fait référence à des ensembles de données dont la taille dépasse la capacité des outils logiciels de base de données classiques à capturer, stocker, gérer et analyser.»

(The McKinsey Global Institute, 2012)

- «Le Big Data est un ensemble d'ensembles de données si vaste et complexe qu'il devient difficile de traiter en utilisant des outils de gestion de base de données.» (Wikipedia, 2014)
- "Les Big Data sont des ressources d'information à grand volume, à grande vitesse et / ou très variées qui nécessitent de nouvelles formes de traitement pour permettre une prise de décision améliorée, la récupération des informations et l'optimisation des processus" (Gartner, 2012)

Quand les données deviennent "Big"?


```
IOPS:Input/Output Operations Per Second
```

Exemple de Big Data

Les données massives désignent des ensembles de données si volumineux ou complexes que les applications de traitement de données traditionnelles ne sont pas adaptées pour y faire face. Les défis incluent l'analyse, la capture, la conservation des données, la recherche, le partage, le stockage, le transfert, la visualisation, l'interrogation, la mise à jour et la confidentialité des informations.

Un exemple général de big data:

Données collectées par le site de réseautage social facebook. Facebook collecte des centaines de téraoctets (To) de données chaque jour. Les données collectées peuvent être des images, des vidéos, des publications, des mises à jour, etc. Les données varient de structurées à non structurées. Un like, share ou reaction peut structurer des données car nous en connaissons clairement la structure. Alors que les mises à jour ou les publications sont des données non structurées qui ne suivent pas exactement une structure. Toutes ces données forment ensemble BigData!

Qu'est-ce qui se passe sous Big Data?

Le Big Data implique les données produites par différents appareils et applications. Voici quelques-uns des domaines qui relèvent du Big Data.

- Données de la boîte noire: il s'agit d'une composante de l'hélicoptère, des avions et des avions à réaction, etc.
- Données sur les médias sociaux: les médias sociaux tels que Facebook et Twitter contiennent des informations et des opinions publiées par des millions de personnes à travers le monde.
- Données boursières: Les données boursières contiennent des informations sur les décisions d'achat et de vente prises sur une part des différentes entreprises réalisées par les clients.
- Données du réseau électrique: les données du réseau électrique contiennent les informations consommées par un nœud particulier par rapport à une station de base.
- Données de transport: Les données de transport incluent le modèle, la capacité, la distance et la disponibilité d'un véhicule.
- Données du moteur de recherche: Les moteurs de recherche récupèrent un grand nombre de données provenant de différentes bases de données.
- Données de capteur: données provenant de différents appareils fonctionnant sur des capteurs, par exemple: données météorologiques (météorologiques et climatiques), données sismiques (séismes), données océaniques (marées, tsunami, etc.).

Ainsi, le Big Data inclut un volume important, une grande vitesse et une variété de données extensible. Les données qu'il contient seront de trois types.

```
1. Structured data : Mostly data from Relational Databases.

2. Semi Structured data : XML data, email data.

3. Unstructured data : Word, PDF, Text, Media Logs.
```

Lire Démarrer avec bigdata en ligne: https://riptutorial.com/fr/bigdata/topic/6890/demarrer-avec-bigdata

Chapitre 2: Démarrer avec Big Data / Hadoop Security

Introduction

Nous pouvons sécuriser les données dans Hadoop en utilisant différentes méthodes. Chaque méthode a ses propres avantages. Nous pouvons également combiner plusieurs méthodes pour un meilleur résultat. Cette rubrique couvre les avantages et les limites de chaque méthode

Remarques

1. Kerberos est un protocole d'authentification réseau:

une. Avantage: authentifier les utilisateurs au niveau d'entrée.

b. Limitation: Kerberos empêche les utilisateurs non autorisés d'accéder à l'environnement. Mais après la connexion, il ne fournira pas d'authentifications de niveau détaillées telles que table, colonne, dossier, niveau de fichier, etc.

2. Apache Sentry est un système d'implémentation du rôle fin

une. Avantage: authentifications au niveau de l'application telles que Hive, Impala, Solr, etc. Il peut contrôler l'accès au niveau de la base de données, de la table et de la colonne pour un utilisateur / groupe donné.

b. Limitation: Il ne peut pas contrôler les dossiers HDFS soulignés par des applications comme Hive, Impala, etc. Ex: Hive table prod.table1 stockée dans /user/hive/warehouse/prod.db/table1. La configuration du rôle de sentinelle dans Hue ne peut contrôler que l'accès aux tables / colonnes dans Hue, mais il est possible que l'utilisateur puisse accéder aux dossiers directement dans HDFS.

c. Limitation: les dossiers HDFS qui ne sont pas liés à Hive, Impala, etc. ne seront pas contrôlés

3. Une liste de contrôle d'accès (ACL) est une liste d'entrées de contrôle d'accès (ACE). Chaque entrée de contrôle d'accès d'une liste de contrôle d'accès identifie un mandataire et spécifie les droits d'accès autorisés, refusés ou audités pour ce dépositaire.

une. Avantage: L' accès au niveau des dossiers est possible par les utilisateurs utilisant

4. HDFS Encryption implémente un cryptage transparent de bout en bout des données lues et écrites sur HDFS

une. Avantage: Crypter les données fournira un niveau de sécurité supplémentaire. En général, le chiffrement des données est requis par un certain nombre d'entités gouvernementales, financières et réglementaires différentes.

Examples

Exemple pour ACL

```
hadoop fs -setfacl
```

Lire Démarrer avec Big Data / Hadoop Security en ligne:
https://riptutorial.com/fr/bigdata/topic/9869/demarrer-avec-big-data---hadoop-security

Crédits

S. No	Chapitres	Contributeurs
1	Démarrer avec bigdata	Ani Menon, Community, Mr. P, NeoWelkin, Sayali Sonawane
2	Démarrer avec Big Data / Hadoop Security	saranvisa