Une image contenant texte, carte de visite

Description générée automatiquement

Table des matières

[I. Etat de l’art 5](#_Toc89343257)

**Aucune entrée de table d'illustration n'a été trouvée.**

Introduction

Dans le cadre du cours de Big Data, il nous est demandé d’effectuer un travail de recherche afin d’en apprendre plus sur le stockage dans le Big Data et d’évaluer notre capacité à faire des recherches sur ce sujet.

Le stockage étant un sujet vaste, ce rapport fera l’objet de documentation sur l’ETL (Data Warehouse) et l’ELT (Data lake) dans un premier temps. Par la suite, le sujet se recentrera sur le data Lake et plus précisément sur la différence entre un mode distribué et un mode non-distribué.

# Etat de l’art

Afin de bien établir le sujet de ce travail, il est important de définir ce qu’est le stockage de manière générale.  
  
Le stockage c’est un ensemble de méthodologie permettant la conservation des produits sous leurs formes initiale ou en passant par une transformation dans un espace donné. En informatique, cette définition s’applique et d’autant plus quand on parle de big data.

A l’heure actuelle, dans une ère ou tout est sujet à prétexte pour récolter des données, le point du stockage est très important. Lorsque qu’on parle de stockage, on parle forcément de comment le stocker et dans qu’elle mesure. Ce qui nous amène à la question, comment stocker un nombre conséquent et croissant de données ?  
  
En Big Data, il existe deux grands types de stockage ; le Data Lake et le Data Warehouse.

## Le Data Warehouse (1 page)

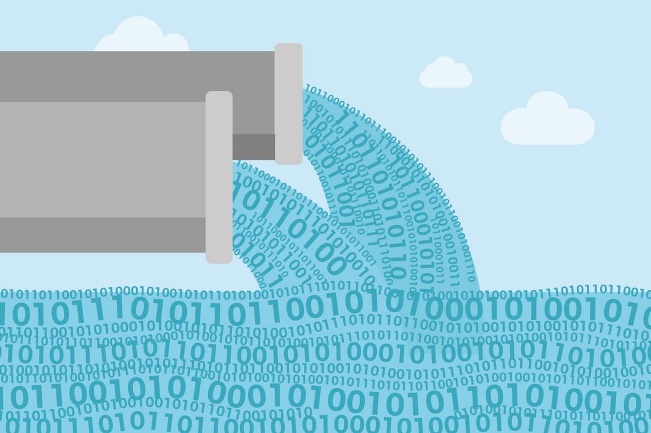
Le Data Warehouse est le premier type de stockage beaucoup utilisé en big data.

* ETL
  + Données stockées transformer
* Axé ???
* Modification a un moment précis, les données mise dans le DW ne peuvent pas être alterées
* Garde l’intégrité des données
* Utilisé par des business analyst

## Le Data Lake (1 page)

Le deuxième grand type de stockage en big data est le data lake. Le Data Lake ou DL est une technologie qui possède de nombreux avantages dans l’ère de la collecte de données.

Le DL utilise le concept nommé ELT qui signifie Extraction, Insertion, Transformation. Ce nom est simplement la suite d’étape utilisé durant la mise en œuvre.

Effectivement, le lac de données (Data lake) accorde la possibilité de stocker d’innombrable données dans n’importe quelle forme. Nous n’avons donc pas besoin de se soucier du format de nos données lors de la récolte et ensuite du stockage de celle-ci. Ce n’est que l’hors de la récupération des données dans le data lake que la transformation des données sera faite afin de les adaptées à l’utilisation que l’on souhaite en faire.

Le machine learning étant une méthodologie d’apprentissage pour ordinateur consistant a l’ordinateur d’apprendre sans que l’humain n’intervienne dans ses lignes de code. Cet apprentissage se fait par un biais d’étude mathématiques et statistiques en analysant les données. Plus le nombre de données est important, plus l’apprentissage est efficace dans toutes les situations. Il est indéniable que le stockage brut de données constitue un avantage considérable. De plus le stockage dans un Data lake est volatile ce qui signifie que les données peuvent être alimenté en temps réel, cela pourrait permettre à un algorithme d’apprentissage de se mettre à jour constamment sur des données collecter continuellement.

Figure 1 - Data lake métaphore

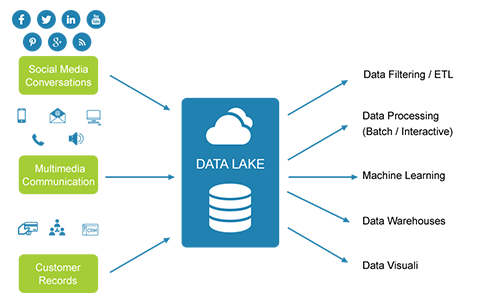
 Cependant, le lac de données possède aussi ses défauts. Une technologie aussi avantageuse pour le machine learning et le stockage de masse ne peut pas être que positive. La data lake est un cout important et constitue une perte d’intégrité des données (LIEN DEFINITION BAS DE PAGE).  
  
En effet, un data lake peut couter plusieurs centaines de milliers d’euros pour une très importante utilisation. Il est donc important pour choisir ce type de technologie, d’en avoir le besoin. De plus, pour une utilisation autre que le machine learning, la perte d’intégrité des données peut poser de gros problèmes, la cohérence des données étant en causes. Le stockage de données brut et la perte d’intégrité des données n’étant pas un problème pour le machine learning, le data lake est principalement utilisé par des data scientist et des data analyst.

Figure 2 - Data lake

* ELT
  + Données stockées brutes sans alteration
  + Transformation seulement quand on souhaite l’utiliser
* Axé machine learning
* Stockage de masse
  + Constitue un cout plus important
  + Nécessaire si on possède énormément de données (lien machine learning)
* Modification en temps réel
* Constitue une perte d’intégrité des données (Intégrité def bas de page)
* Utilisé par des data scientist | data analyst

## Hadoop ( ½ page)

* Définition
* A quoi ça sert

## Docker (1 page)

* Définition
* Utilisation
* Docker Compose

# Problématique

# Proposition de solution

# Aspect pratique

# Conclusion

# Références bibliographiques