

Gestion des données

Cours 1 – Introduction

Olivier Schwander

`<olivier.schwander@sorbonne-universite.fr>`

2021-2022

Page web

http://www-connex.lip6.fr/~schwander/enseignement/m2stat_gd/

Contenu du cours

Contenu

- ▶ Bases de données
- ▶ Business intelligence
- ▶ Extraction de données (services web, pages web, emails)

Évaluation

- ▶ Devoir maison
- ▶ Examen final

Travaux pratiques

Sur vos machines

Machine virtuelle

- ▶ Fichier fourni sur le site du cours
- ▶ Machine Linux avec tout installé dessus

Éventuellement

- ▶ Sur votre machine directement
- ▶ Mais à éviter sauf nécessité
- ▶ **Support technique très limité**

Pour un statisticien

Chaîne de traitement complète

- ▶ *Acquérir les données*
- ▶ *Stocker les données*
- ▶ *Mettre à jour les données*
- ▶ Traiter les données
- ▶ Visualiser, faire des rapports

Objectifs

- ▶ Maîtriser la totalité de la chaîne
- ▶ Comprendre le rôle des données en entreprise
- ▶ Connaître les outils utilisés en entreprise

Pour une entreprise

Métiers

- ▶ Rarement *traiter des données*
- ▶ Vendre des choses: produits, services, publicités

Objectif

- ▶ Comprendre la clientèle, les utilisateurs, l'entreprise
- ▶ Sur le long terme
- ▶ Sur des secteurs et des domaines variés
- ▶ Intégrer l'analyse des données dans la vie de l'entreprise

Bases de données

Stockage structuré des données

Logiciels pour le stockage

- ▶ Un serveur accessible à travers un réseau (souvent)
- ▶ Permettant de lire et d'écrire
- ▶ Garantissant des propriétés intéressantes

Langage de requête

- ▶ Exprimer des demandes
- ▶ Filtrer pour n'obtenir que les résultats intéressants
- ▶ Efficacement
- ▶ Standardisé (SQL) ou pas (noSQL)

Extraction des données

Comment récupérer les données

- ▶ Depuis une base de données
- ▶ Depuis une interface documentée
- ▶ Depuis des fichiers dans un format structuré
- ▶ Depuis des fichiers dans un format non-structuré
- ▶ Depuis une interface non-documentée
- ▶ Depuis une interface non-coopérative, non-documentée, qui peut changer

Quelques bibliothèques, techniques et outils

- ▶ Pour des services web
- ▶ Pour des pages web
- ▶ Pour des emails

Business intelligence

Définition

L'Informatique Décisionnelle (ID), en anglais Business Intelligence (BI), est l'informatique à l'usage des décideurs et des dirigeants des entreprises. Les systèmes de ID/BI sont utilisés par les décideurs pour obtenir une connaissance approfondie de l'entreprise et de définir et de soutenir leurs stratégies d'affaires, par exemple : d'acquérir un avantage concurrentiel, d'améliorer la performance de l'entreprise, de répondre plus rapidement aux changements, d'augmenter la rentabilité, et d'une façon générale la création de valeur ajoutée de l'entreprise.

Wikipédia: article *Informatique décisionnelle*

Business intelligence

ETL: Extract - Transform - Load

- ▶ Récupérer les données là où elles sont
- ▶ Transformer les données si besoin
- ▶ Stocker les données de façon exploitable

Data Warehouse

- ▶ Stockage des données
- ▶ Toute l'histoire de l'entreprise
- ▶ Stable dans le temps

OLAP: Online Analytical Processing

- ▶ Données en grande dimension
- ▶ Visualisation, structuration
- ▶ Pas forcément de traitement statistique compliqué