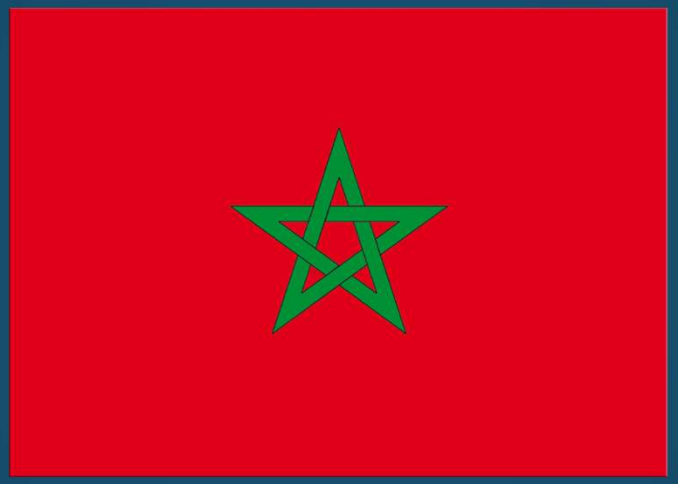
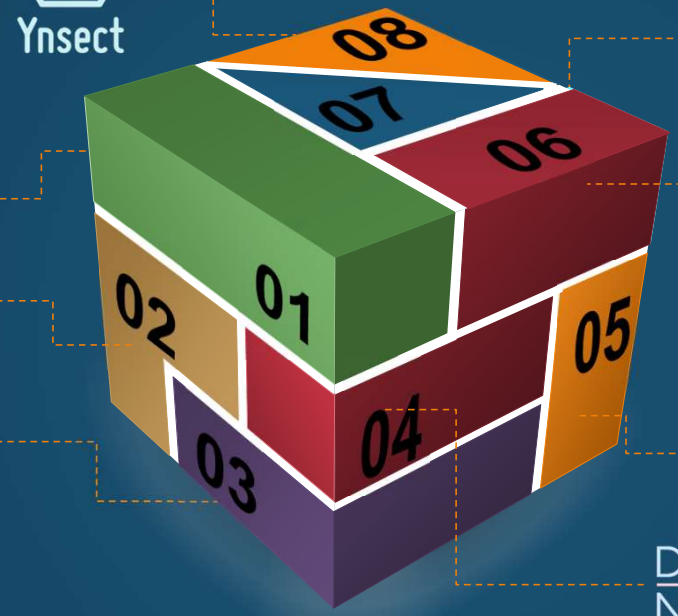
Centrale
CASABLANCA

DATA ENGINEERING

CONCEVOIR DES ENTREPOTS DE DONNÉES « RESPONSABLES »



LES PRÉSENTATIONS …



Centrale CASABLANCA

UNIVERSITÉ PARIS 1 PANTHÉON SORBONNE
IAE PARIS
SORBONNE BUSINESS SCHOOL

UniLaSalle
Terre & Sciences

École des Ponts
ParisTech

SUPINFO International University

UNIVERSITY OF LONDON

Stanford University

PLAN DE L'INTERVENTION

Présentation générale

Atelier
Data Engineering

Conclusions

Data et gouvernance

Expositions

Collecte et stockage

Traitement et raffinage

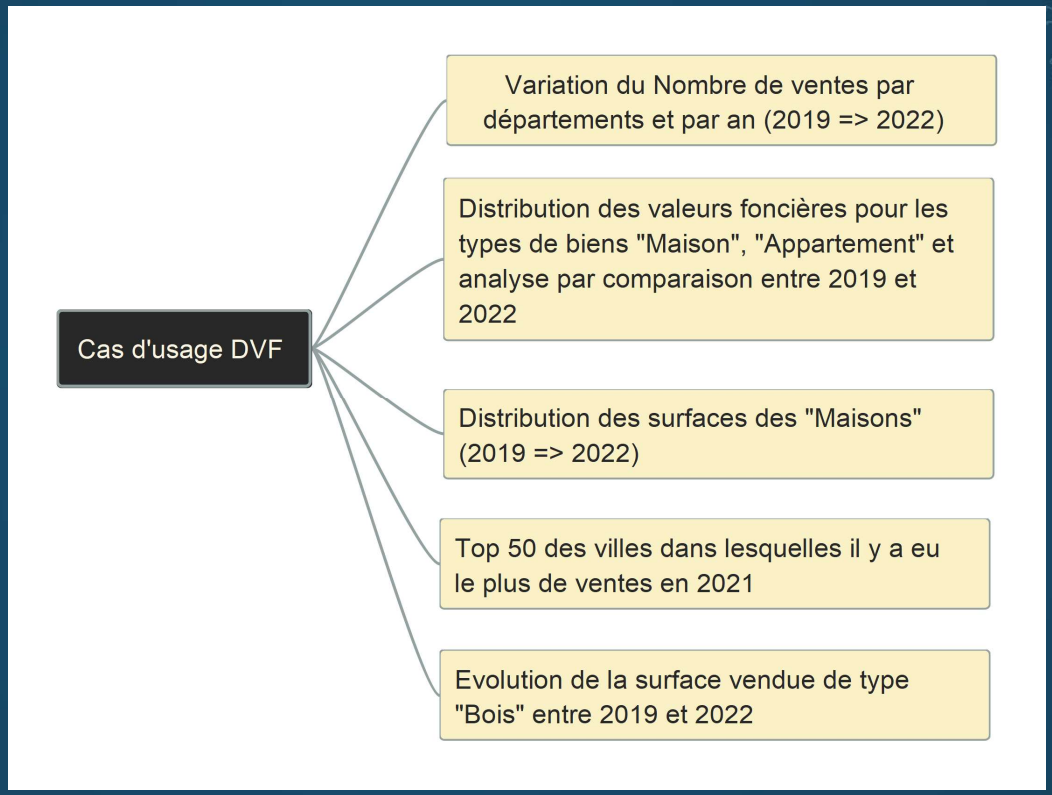
EN GUISE D'INTRODUCTION ….

- La gestion des données au sens « large »
- Implications en entreprise
- Environnements techniques
- Responsabilités sociales et environnementales

METHODE DE TRAVAIL

Date et Heure	Cours et objectifs
07/11/2023 [8h45 => 10h30]	Présentation de la méthode et du projet Introduction du cours, présentation des objectifs et présentation générale des enjeux autour des données
07/11/2023 [10h45 => 12h15]	Poursuite du cours sur les données et sur les nécessités d'une gouvernance. Présentation du framework Spark, utilisations et méthodes
07/11/2023 [14h => 16h]	Atelier pratique, installation de l'environnement et premiers cas pratiques
10/11/2023 [8h45 => 10h30]	Processus de collecte et de stockage de données
10/11/2023 [10h45 => 12h15]	Poursuite du cours sur les données, notions de traitements et de raffinages et d'expositions, conclusion du cours
10/11/2023 [14h => 16h]	Suite et fin de l'atelier de « Data Engineering »

MINI PROJET


Cas d'usage DVF
Variation du Nombre de ventes par départements et par an (2019 => 2022)
Distribution des valeurs foncières pour les types de biens "Maison", "Appartement" et analyse par comparaison entre 2019 et 2022
Distribution des surfaces des "Maisons" (2019 => 2022)
Top 50 des villes dans lesquelles il y a eu le plus de ventes en 2021
Evolution de la surface vendue de type "Bois" entre 2019 et 2022

PLAN DE L'INTERVENTION

Présentation générale

Atelier
Data Engineering

Conclusions

Data et gouvernance

Expositions

Collecte et stockage

Traitement et raffinage

DONNEES / INFORMATIONS / CONNAISSANCES

Donnée

Représentation d'une information dont on va se servir dans un cadre numérique

Information

Représentation quantitative d'une donnée. Donnée mise dans un contexte

Connaissance

Mise en relation de plusieures informations pour parvenir à la compréhension d'un phénomène

DONNEES / INFORMATIONS / CONNAISSANCES

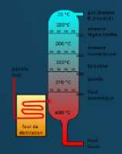
Donnée	Information	Connaissance
1000	1000°C [Température]	Température d'un four à céramique
	1000 feet [Altitude]	Température dans un réacteur d'avion
	1000 kg [Masse]	Altitude en tour de piste
	1000 km/h [Vitesse]	1 m3 d'eau
	1000 personnes [Dénombrement]	Vitesse prévue pour l'hyperloop
		Echantillon d'un sondage

LES FLUX ET LES STOCKS

- **Les données ont toujours existé**
  - *C'est notre **perception du monde** qui nous fait prendre la mesure de ce que nous pouvons en extraire*

- **Internet, IoT, Téléphone Intelligent**
  - *La démocratisation des moyens mobiles de communication et la multiplication des senseurs génèrent des données sur tout type d'activités (humaine ou non)*

- **Collecter, Stocker, Traiter**
  - *Les données générées n'ont pas de **modèle a priori**. On les collecte, puis on les stocke et ensuite on regardera (??) si elles peuvent servir à un ou plusieurs cas d'usage*



- **Raffiner, analyser et prévoir**
  - *La finalité est de pouvoir extraire de la valeur (analyse, corrélation, tendances, prédiction, apprentissage)*

PRENDRE LA MESURE

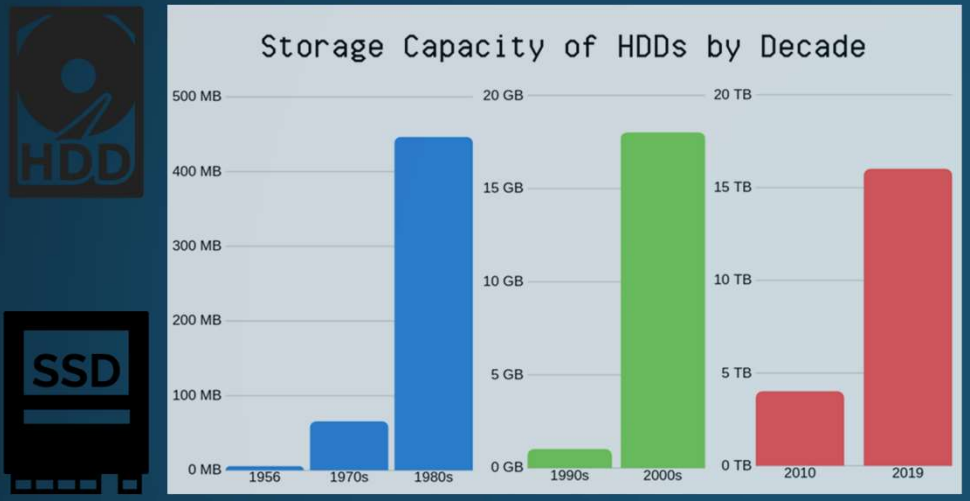
A Un octet

`0 100 0001`

1002 octets

1024 x 1024 octets = 1 Mo

LES FLUX ET LES STOCKS



https://www.securedatarecovery.com/blog/wp-content/uploads/2019/02/stats1.png

$1\ \text{Zio} = 1024^3\ \text{Tio}$
$1\ \text{Eio} = 1024^2\ \text{Tio}$
$1\ \text{Pio} = 1024^1\ \text{Tio}$

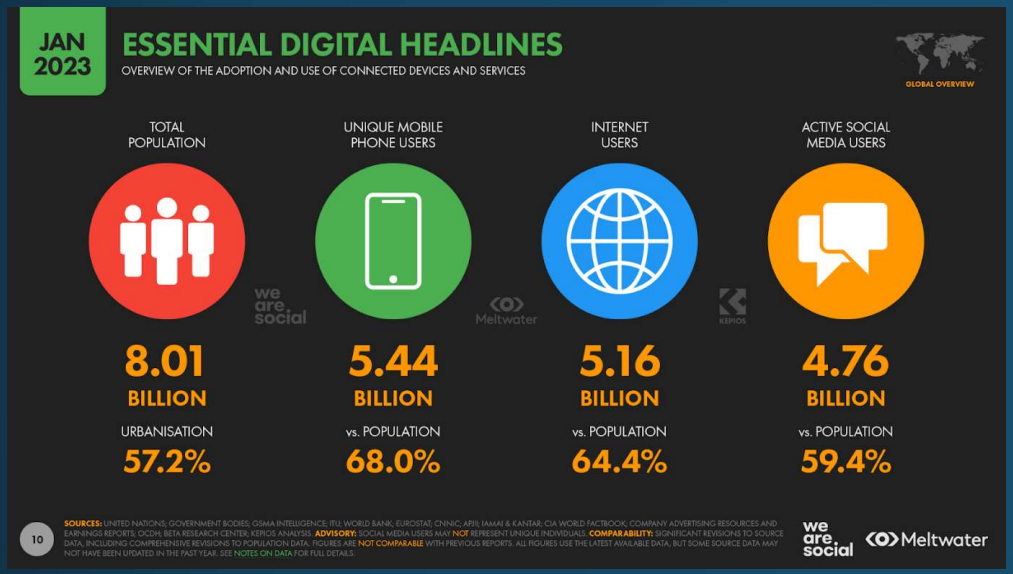
$1024^3 = 1\,073\,741\,824$

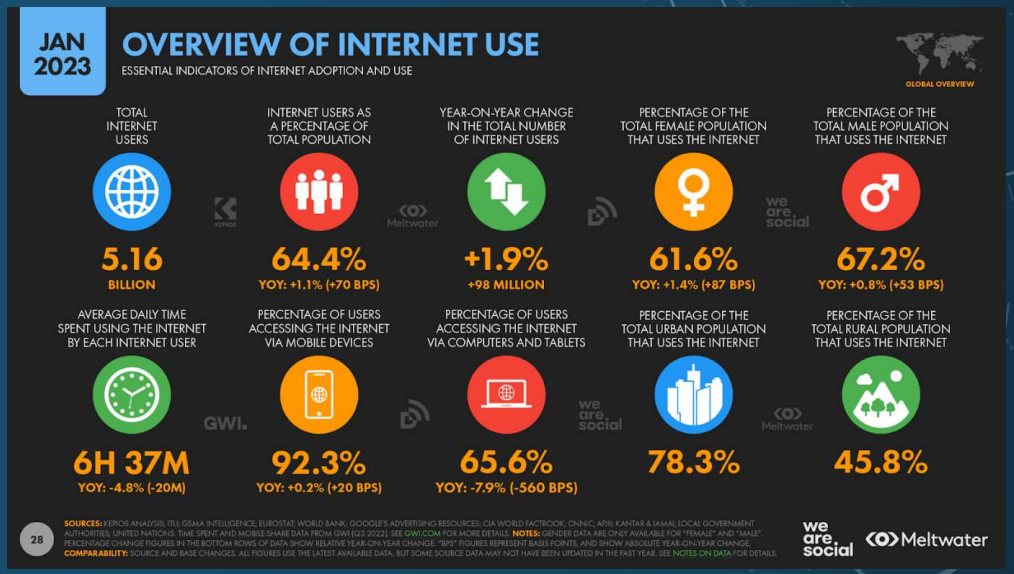
262 000 photos de 4 Mio

170 heures de vidéo
[1920x1080]

210 000 livres
[5 Mio par livres]

UTILISATION DIGITALE





LE VOLUME D'INTERNET (2023)

120 ***Zebioctet(Zio)*** =
$33x1024^3$ *Tébioctets*

GPS, GSM,Réseau sociaux,web marchant,..

Données marchandes, comptables, produits, finances, personnel ...

Réseau sociaux,web marchant, ...

Interaction homme <-> Objet
Interaction Objet <-> Objet

LE PARADOXE

2/3 des entreprises ne savent pas ou ne peuvent pas traiter les données qu'elles collectent

A elle seule , les ROT Data représenteront en 2020 un coût de près de 900 Milliards de $ pour les entreprises

La plupart des entreprises ne savent plus faire face au déluge des données qu'elles ont devant elles !

- **Absence de stratégie data ?**
- **Le stockage dans le cloud est il « gratuit » ?**
- **Les ressources informatiques sont elles infinies et gratuites en entreprise ?**

LE COUT ENVIRONNEMENTAL

1 Gb (1024 Mb) => 0.0042 kg de CO_2
1 Zb = 1024^3 Tb = 1024^4 Gb => 1024^4 x 0,0042 = 4 617 948 tonnes de CO_2

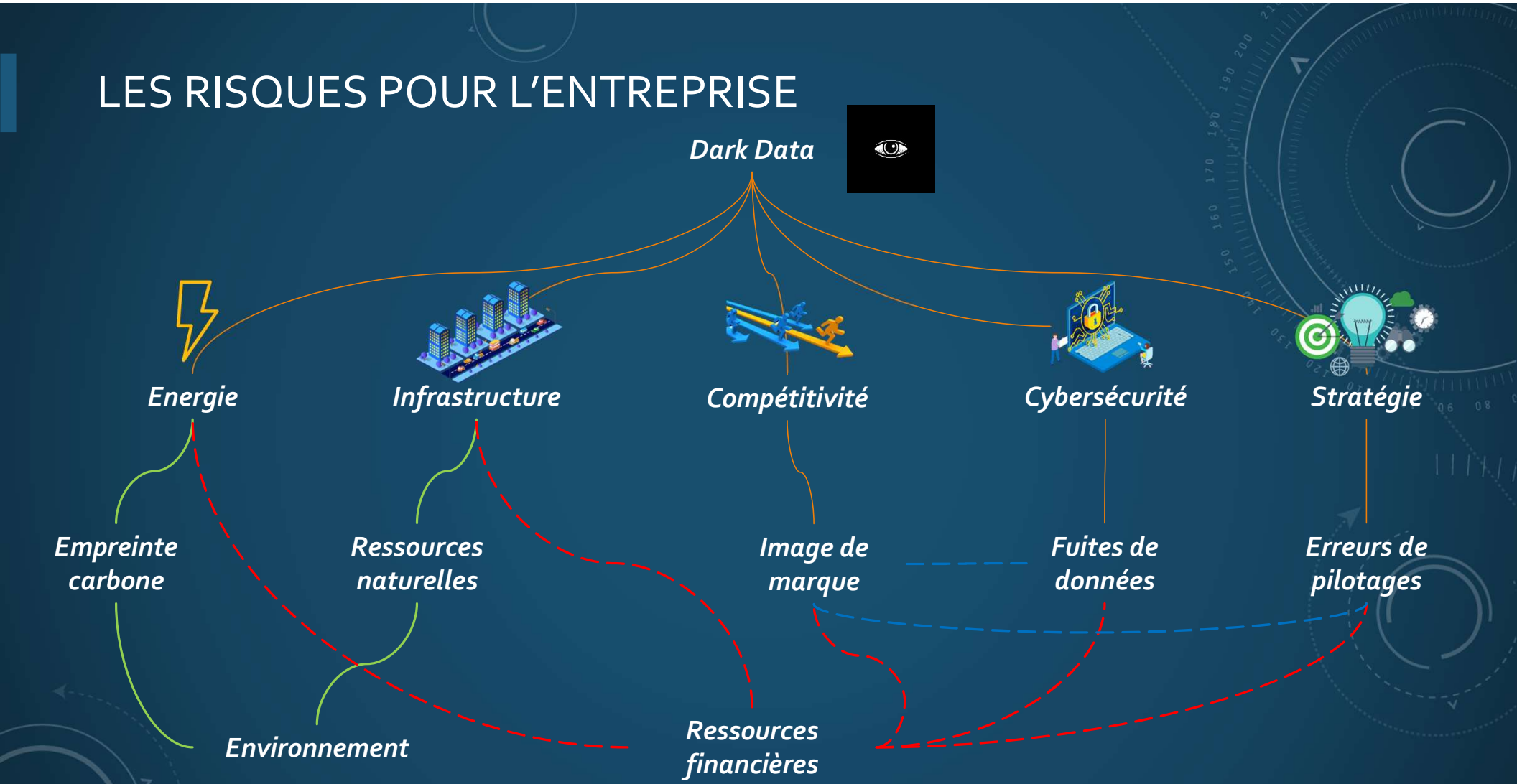
Le volume global d'internet (120 Zb) coute donc par an :

554 153 860 tonnes de CO_2

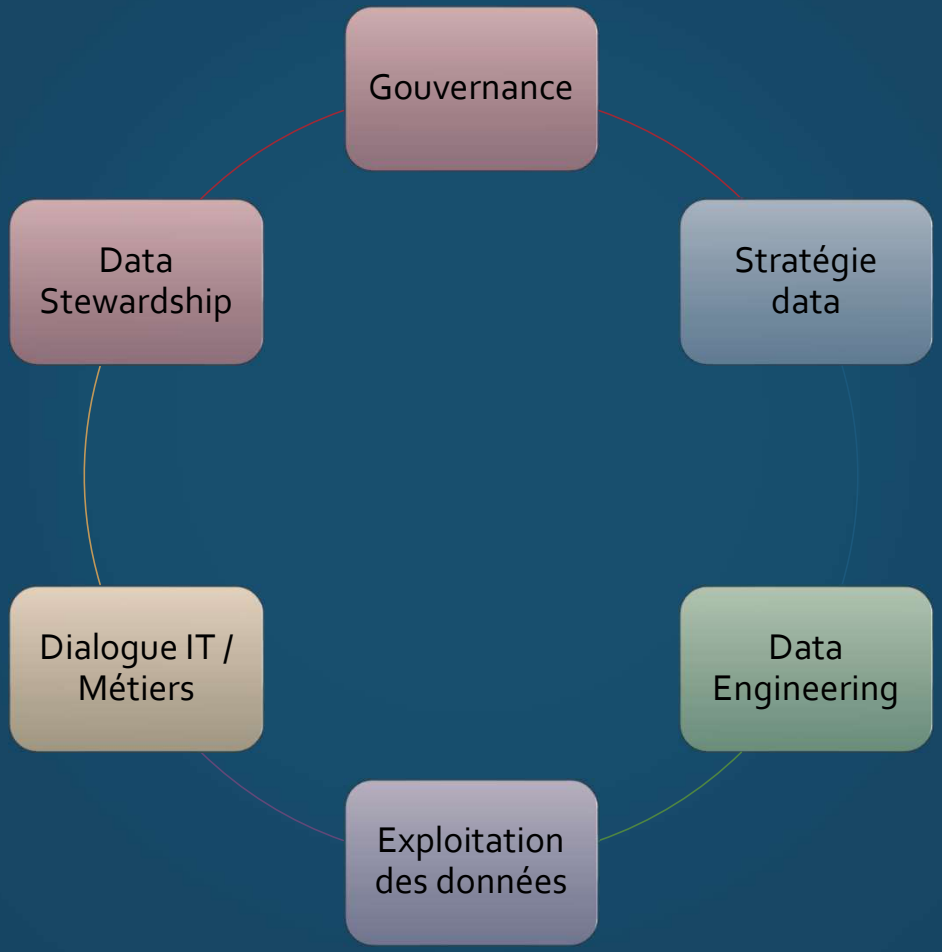
Equivalent de l'émission en CO_2 de 121 millions de véhicules par an

Equivalent au volume de 221 661 piscines olympiques (2500 m^3)

LES RISQUES POUR L'ENTREPRISE



CE QU'IL MANQUE



UN MÉTIER D'INGÉNIEUR

Mesure de débit

Mesure de pression

Usine de filtration

Compteur de consommation

Usine de séparation

Usine de valorisation

L'eau mérite un traitement particulièrement complexe pour permettre sa consommation.

Il en va de même pour la data qui est l'actif circulant de valeur dans une entreprise.

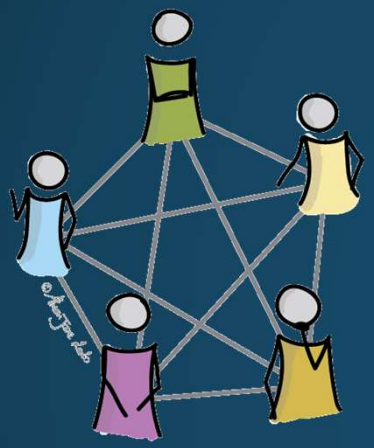
L'INGENIERIE DES DONNÉES

- Appliquer la stratégie « data » de l'entreprise
- Travailler en étroite collaboration avec l'architecte data et la DSI
- Prendre soin des données (traitement, stockage, exposition)
- Délivrer des données de qualité aux utilisateurs métiers
- Exploiter de manière industrielle les sources et stockage de données
- Construire des « pipelines » robustes, exploitables et maintenables

STRATÉGIE DATA DE L'ENTREPRISE

- Les usages métiers doivent « tirer » les travaux du data engineering.
- Les chantiers sont identifiés, jalonnés et sont suivis au niveau financier.
- Les technologies utilisées sont celles de l'entreprise.
- Les ressources de l'entreprises sont utilisées de manière rationnelles.
- Une documentation est mise à jour sur les travaux effectués et en cours.
- Le personnel est formé, entraîné pour utiliser les outils mis en place.

UN TRAVAIL D'ÉQUIPE



- L 'équipe data est composée de data engineer , d'architecte(s) data et de Lead Data Engineers.
- L'architecte (maitrise d'ouvrage) et le Lead Data Engineer (maîtrise d'œuvre) travaillent en étroite collaboration dans les chantiers.
- La DSI est le partenaire privilégié de l'équipe data car elle fournit les infrastructures et leur exploitation.
- Les équipes métiers et l'équipe data travaillent ensemble dans les projets identifiés comme stratégiques pour l'entreprise
- Les data scientists contribuent aux analyses des données en cours de traitement.

TRAITEMENT DES DONNÉES

- Les sources de données font l'objet d'un audit et d'une cartographie.
- Le data steward assure l'interface entre les besoins et les sources de données.
- Seules sont captées les données en rapport avec les besoins exprimés.
- Les technologies utilisées sont celles à l'état de l'art, pérennes et documentées.
- Les traitements réalisés sont équipés de sondes logicielles (journalisation).
- Un ordonnanceur de tâches est mis en place pour enchainer les traitements.

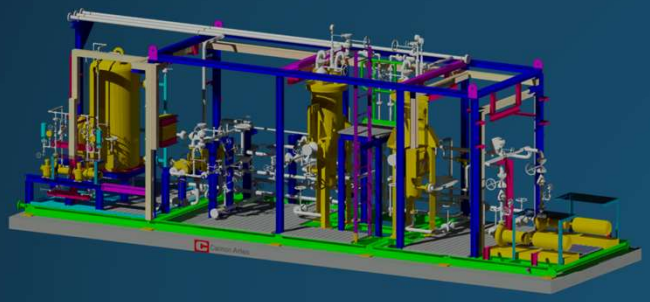
QUALITE DES DONNÉES

- La qualité des données est indispensable dans toute société « data driven ».
- Durant tous les processus de traitement la qualité des données est mesurée.
- Prélèvement d'échantillon (échantillothèque).
- Mise en place des seuils « qualité » avec les experts métiers.
- Décider les actions à mettre en place en cas de données polluées.
- Communiquer régulièrement sur les métriques « qualité » data.

EXPLOITATION INDUSTRIELLE DES DONNÉES

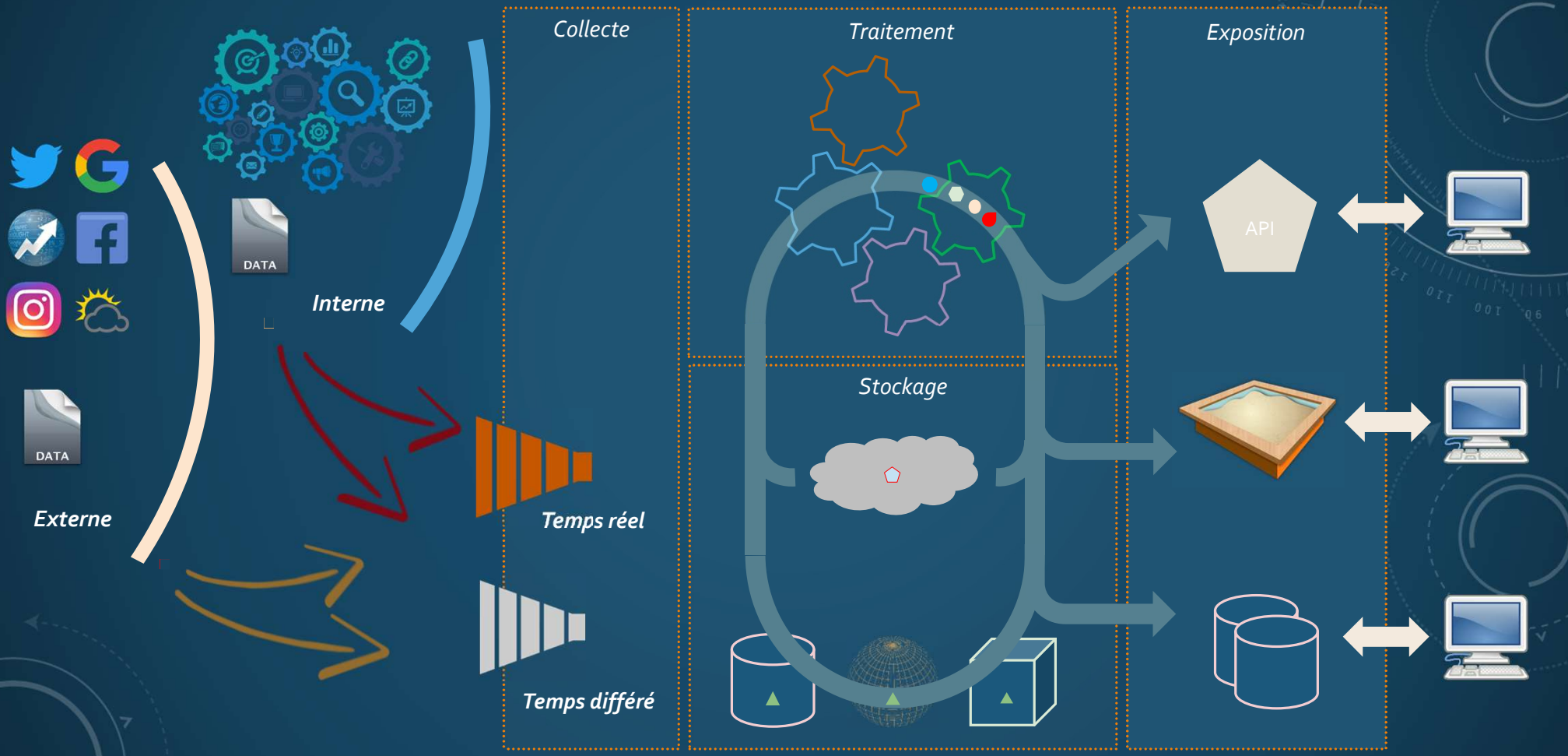
- Les traitements sur les données sont identifiés et cartographiés.
- Chacun des algorithmes est connecté à un système centralisé de « log ».
- Quand c'est nécessaire le calcul distribué est utilisé.
- Les traitements s'enchaînent de manière conditionnelle (succès/échec).
- Les stocks de données utilisent des technologies appropriées et maîtrisées.
- Limiter le nombre de langages de programmation utilisé.

EXPLOITABILITÉ



- Tous les traitements sont associés à une documentation d'exploitation.
- Les erreurs font l'objet d'une nomenclature et de procédures de reprises.
- L'exploitabilité de la solution est prise en charge par un « exploitant ».
- Les mises à jours sont réalisées par l'exploitant et sous sa responsabilité.
- Un monitoring des opération est réalisé en permanence.
- Des indicateurs d'exploitation sont publiés à fréquence régulière.
- La consommation des ressources est mesurée en permance (disque, cpu, réseau,ram, ...).
- Organiser, concevoir et mettre en place un PRA et un PCA

MISE EN OEUVRE



EN PRATIQUE …

Les sources de données sont d'un grand nombre de types :

- Systèmes d'entreprise (ERP, CRM)
- Systèmes industriels (automates)
- Systèmes temps réels (production, aéronautique,…)
- Système de mesures (pression , température, …)
- Mise à disposition de données (open data, …)
- … Et bien d'autres encore

Tous ces systèmes utilisent pour la plupart des bases de données relationnelles pour y stocker l'ensemble des informations. On trouve également des stockages de fichiers.

Peut-on , doit-on connecter une Plateforme Data sur les systèmes opérationnels ?

COLLECTE DES DONNÉES

Les tâches à réaliser avant toute collecte de données :

- Identifier les protocoles de communication.
- Mettre en place la sécurité d'accès aux sources de données.
- Choisir la stratégie de connexion (synchrone / asynchrone).
- Identifier la pérennité de la source.
- Identifier la fréquence de collecte.
- Identifier les données à collecter.
- Identifier les espaces de stockages cibles.

La collecte des données est indissociable de deux étapes :

- Le contrôle qualité
- L'indexation

META DONNÉES

- Avant tout stockage on fera un audit qualité des données qui sont collectées. Ces métriques feront partie de l'ensemble des méta data utilisées pour l'indexation.
- Toutes les données « brutes » qui proviennent des systèmes opérationnels doivent être stockées dans un « sanctuaire » sans transformation et être indexées en utilisant les méta data extraites.
- Peut alors commencer la phase de transformation des données :
  - Filtrage
  - Ajout / suppression de données
  - Agrégations
  - Stockage des éléments calculés

INDEXATION DES DONNÉES

Chaque container de données (structuré, non structuré) doit être indexé en prenant en compte :

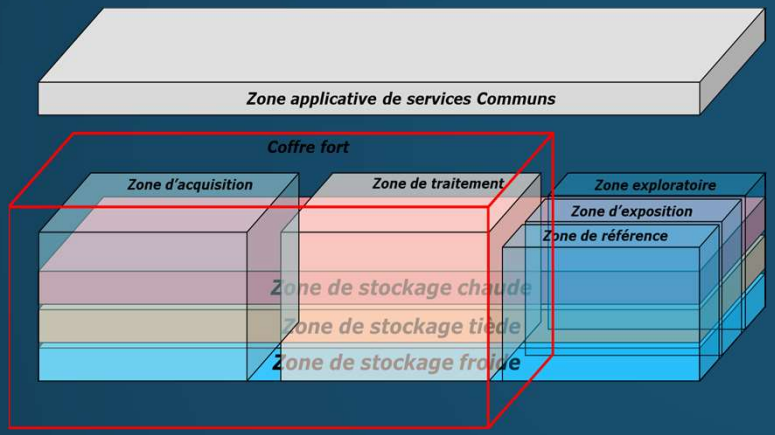
- Les informations de base (nom, date de création, signature [sha256, sha1]
- Les informations sur le contenu (structuré, non structuré, nb de colonnes, types des colonnes, nb d'enregistrements, … toutes autres méta données)
- L'origine des données (producteur)
- Les métriques qualités
- Les cas d'usages qui utilisent ces données
- La date limite d'utilisation

STOCKAGE DES DONNÉES

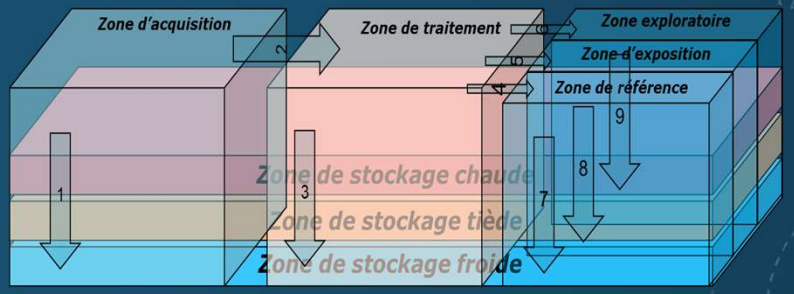
- Stocker les données c'est garder de manière permanente des informations de bases, les informations transformées pour les utiliser ultérieurement.
- Les données doivent être accessibles facilement aux équipes qui en auront besoin dans des structures « utiles » et correspondantes à leurs métiers.
- Tout stockage doit être obligatoirement accompagné d'une indexation.
- Se poser la question de la volumétrie à stocker a court terme, moyen terme, long terme.

LAC DE DONNÉES (DATALAKE)

Une vision « multidimensionnelle » des zones de stockages au sein d'un DataLake



Les différentes zones de stockage



La dynamique inter zone

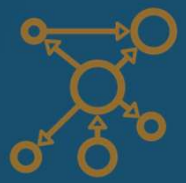
Chaque zone de stockage possède des propriétés particulières liées au cycle de vie des données

AUTRES TYPES DE STOCKAGE

Plusieurs types de stockages de données existent. Ces stockages seront choisis en fonction des licences, des couts, de leurs types.

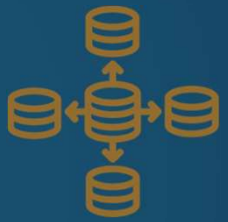
SGBD/R

Base Objet



Base graphe

Base séries temporelles



Base distribuée

Base NRT

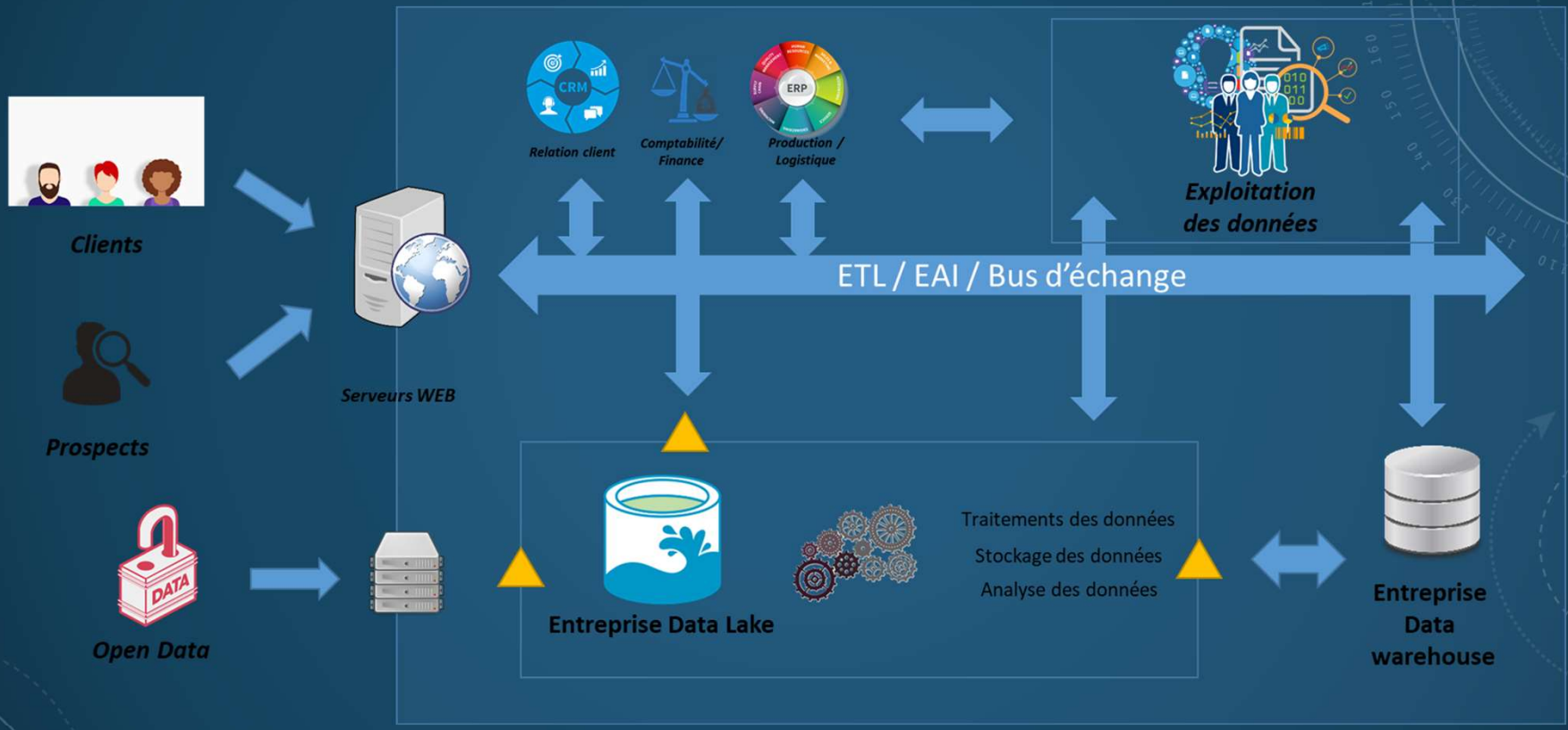
Système de fichiers

Système de fichiers distribué

ANALYSE ET EXPOSITION DES DONNÉES

- L'analyse des données peut dans certains cas être traitée par les équipes de Data Engineer quand aucune compétence n'est présente dans les équipes métiers.
- L'analyse permet la mise en place de :
    - *KPI*
    - *Dashboards*
    - *Rapports*
    - *Workflows de décisions*
- Les technologies utilisées sont principalement les outils de programmation, qui permettent de faire des calculs complexes ainsi que la la transmission d'informations

UNE ENTREPRISE PILOTÉE PAR LES DONNÉES



UN DERNIER POINT

Après l'exposé de toutes ces notions, que valent vraiment vos données ?

ESTIMATION DES INVESTISSEMENTS

Cout IT

Poste budgétaire	Montant
Réseau : transport des données	
Stockage (raw data)	
Stockage (raffinage)	
Exposition des données	
Sauvegarde des données	
Sécurité des données / accès des données	

Cout Data Engineering

Poste budgétaire	Montant
Spécifications / Conception / Réalisations	
Tests / MEP / Orchestration / MOE	
Transformation / Indexation / Valorisation	
Réalisation d'objets métiers	

Cout Data Science

Poste budgétaire	Montant
Analyse des besoins clients	
Création des modèles / Tests des modèles	
Mise à l'échelle	
Analyse des écarts et adaptation des algorithmes	

Cout Pilotage

Poste budgétaire	Montant
Définition des plans projets	
Elaboration des modèles de coûts	
Suivi et animation des projets	

Cout Personnel

Poste budgétaire	Montant
Recrutement / Formation permanente	
Salaires + charge	

ESTIMATION DE LA VALEUR / ESTIMATION DES PERTES…

Valeurs pour l'entreprise

Poste budgétaire	Montant
Avantage concurrentiel	
Attractivité des produits et des services	
Suivi et recrutement de nouveaux clients	
Adhérence aux besoins des clients	
Réduction des pertes clients	
Innovation	

Pertes pour l'entreprise

Poste budgétaire	Montant
Destruction des données (volontaire ou non)	
Vol de données (fuite de données)	
Piratage des sites web / transactions	
Altération de l'image de marque	
Perte de clients	
Perte d'attractivité / compétitivité	

ESTIMATION DE LA VALEUR / ESTIMATION DES PERTES…

Investissement	Profit	Perte
Coûts IT		
Coûts Data Engineering		
Coûts Data Science		
Cout Pilotage		
Cout Personnel		
	Valeur entreprise	
		Pertes entreprise

La valeur des données de l'entreprise est très importante (matériel et immatériel) et son succès et sa survie en dépendent très fortement.

Savoir « valoriser » les données c'est les considérer comme un actif et non plus comme une réalité « virtuelle »

Le cours de vos données sur le Dark Web…

- Données de carte de crédit et banque en ligne (de **1$ à 30$**)
- Identifiants de compte crypto "vide" (**de 300 $ et 600 $**)
- Identifiants e-commerce (**de 10$ à 250$**)
- Identifiants de réseaux sociaux (**de 40$ et 50$**)
- Identifiant de compte de streaming et gaming (**de 5$ à 10$**)
- Identifiants d'email (**de 50$ à 70$**)
- Identifiants pour outil de paiement (**de 25$ et 1000$**)
- Documents d'identification (**de 10$ et 4000$**)
- ***50 000 lignes d'informations clients*** : (**de 2000$ à 3000$**)

BIBLIOGRAPHIE

- https://wearesocial.com/fr/blog/2023/01/digital-report-levolution-du-numerique-en-2023/
- https://dataintell.io/2023/03/hidden-costs-rot/
- https://www.veritas.com/fr/fr/resources/dark-data
- https://www.tandfonline.com/doi/full/10.1080/14778238.2023.2192580
- https://digitaldecarb.org/the-figures/
- https://www.supplychainbrain.com/articles/37198-data-corporate-asset-or-waste-by-product-of-business-processes
- https://co2.myclimate.org
- https://www.tessi.eu/fr/emmanuelle-ertel-sur-b-smart/
- https://explodingtopics.com/blog/data-generated-per-day
- https://hai.stanford.edu/news/quantifying-value-data
- https://www.hpe.com/us/en/insights/articles/how-do-you-measure-the-value-of-data-2203.html
- https://en.wikipedia.org/wiki/Data_valuation
- https://www.talend.com/resources/data-value/
- https://positivr-fr.cdn.ampproject.org/c/s/positivr.fr/classement-8-donnees-personnelles-vendues-dark-web/?amp=1