Centrale
CASABLANCA

DATA ENGINEERING

CONCEVOIR DES ENTREPOTS DE DONNÉES « RESPONSABLES »

PLAN DE L'INTERVENTION

Présentation générale

Atelier
Data Engineering

Conclusions

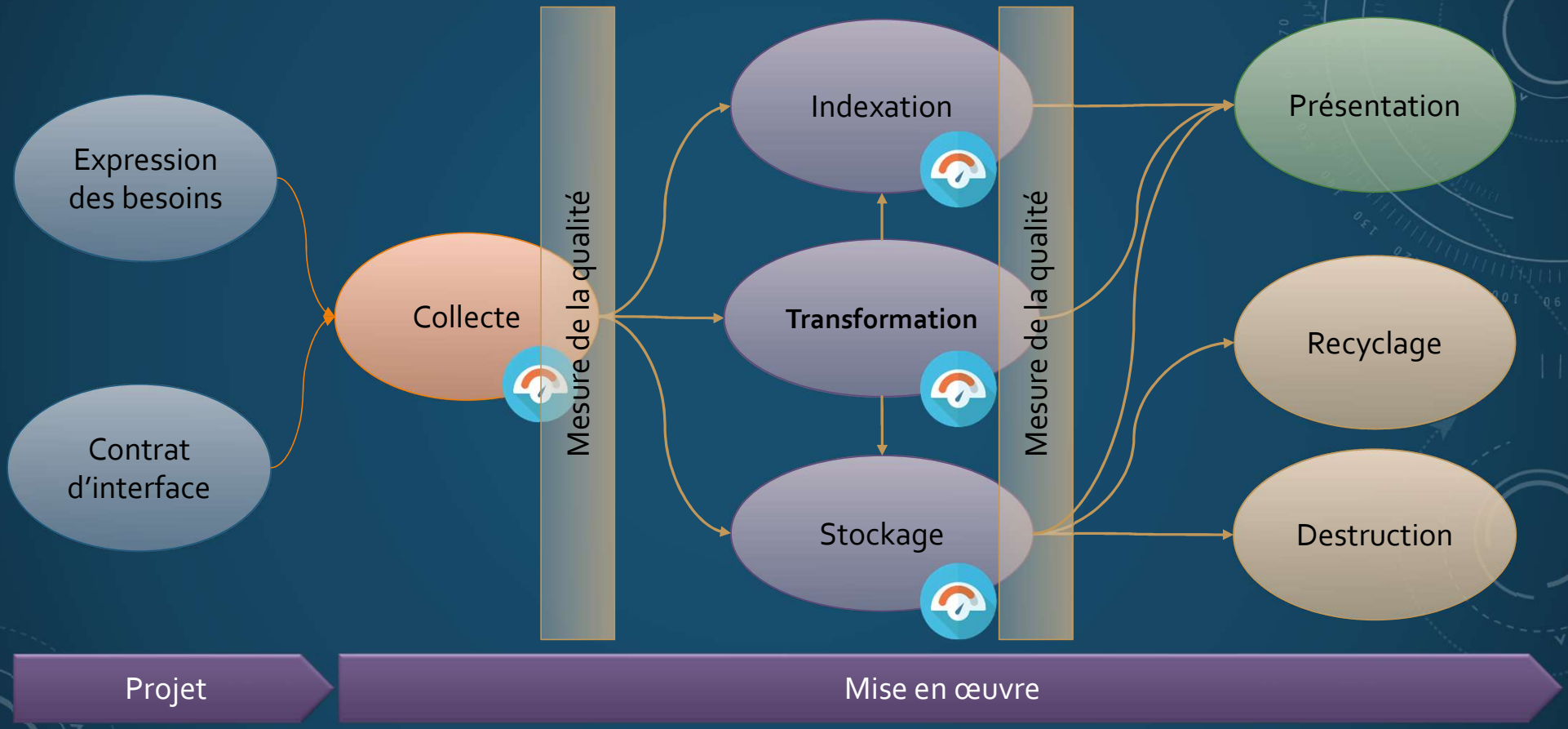
Data et gouvernance

Expositions

Collecte et stockage

Traitement et raffinage

LE PARCOURS DES DONNÉES EN ENTREPRISE



PARCOURS DES DONNÉES : PHASE PROJET

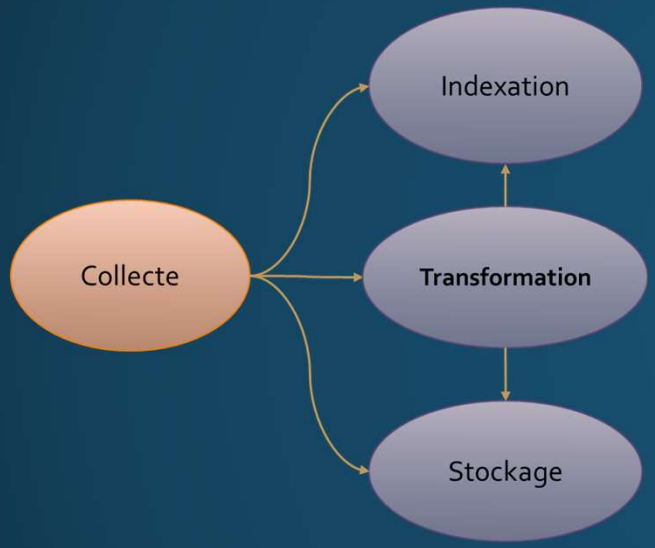
Expression des besoins

- En quoi mes cas d'usages répondent ils à la stratégie de l'entreprise ?
- Quelle problématique d'entreprise ces cas d'usages mettent ils en évidence ?
- Quelles sont les données utiles dont mes cas d'usage ont besoin ?
- Les sources de données identifiées sont elles fiables et pérennes ?
- Quelles sont les transformations ?

Contrat d'interface

- Description du contenant des données (fichier, flux, base de données, …)
- Description du format des données (typologie, liste de valeurs, …)
- Description du protocole de transmission (flux, FTP, HTTPs, …)
- Description de la fréquence d'acquisition des données
- Description des mesures de sécurité à prendre lors de l'acquisition des données

PARCOURS DES DONNÉES : PRÉPARATION



La **collecte** des données met en œuvre le contrat d'interface et permet techniquement de faire le lien entre la source de données et le DataLake de l'entreprise

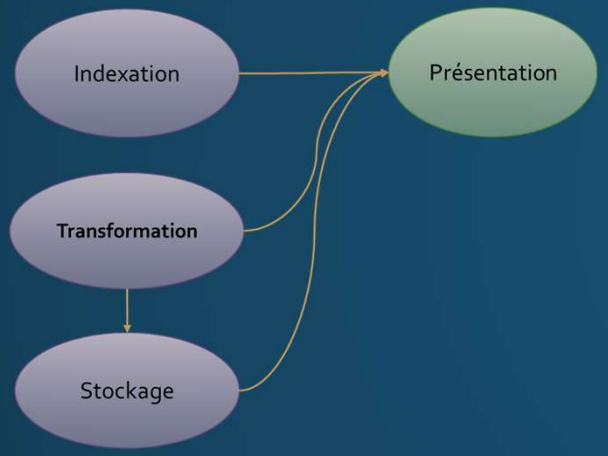
L'indexation permet de connaître précisément la localisation de chaque donnée ou sous ensemble de données. Cette indexation permettra aux utilisateurs de « fouiller » de manière globale et intelligente

La **transformation** permet une modification, des calculs, des agrégations sur les données en cours d'intégration

Le **stockage** permet de conserver de manière pérenne les informations brutes aussi bien que les données transformées

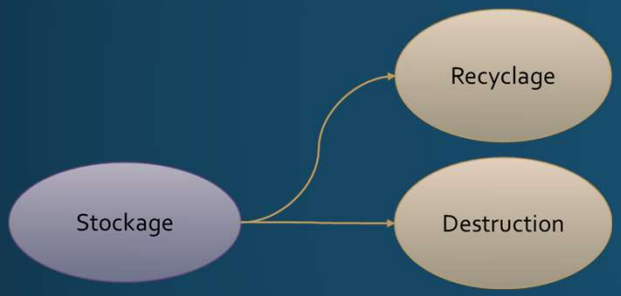
En permanence , toutes ces activités de préparation sont sous « monitoring » et font l'objet d'audit et de mesures de la qualité

PARCOURS DES DONNÉES : PRÉSENTATION



- L'indexation est un des outils de la présentation des données, car elle précède toute utilisation.
- Les données sont préparées en amont ou à la demande, lors de la phase de **transformation** pour répondre aux exigences fonctionnelles et techniques des utilisateurs finaux
- La **présentation** met en œuvre les moyens techniques permettant aux « utilisateurs » d'accéder aux données (API, HTTP, …)
- Les « utilisateurs » possèdent un périmètre limité de vision des données défini au travers d'une **politique de sécurité**

PARCOURS DES DONNÉES : RECYCLAGE ET DESTRUCTION



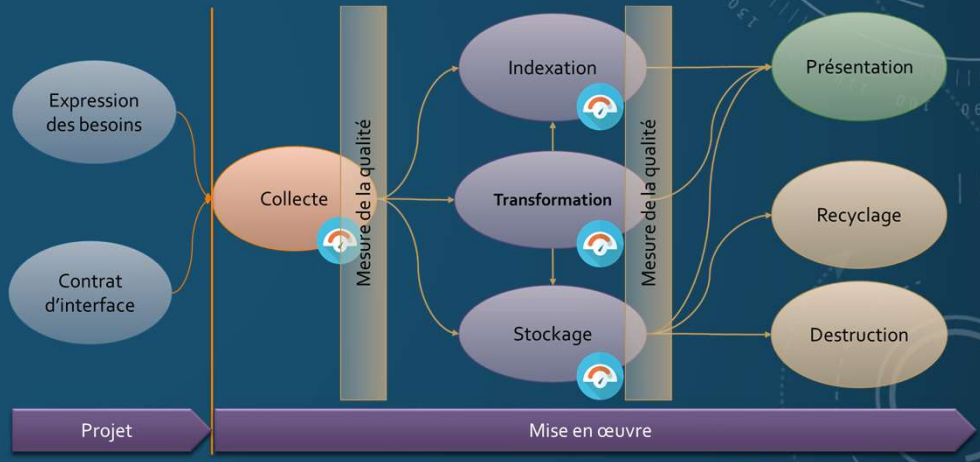
- Les données ont une date limite d'utilisation. Cette DLU permet d'enrichir les Méta Données disponibles dans l'indexation

- A partir de la DLU, les données doivent être :
  - soit recyclées et mises à disposition d'autres ensembles de données
  - soit conservées dans des « congélateurs » à données, soit définitivement effacées.

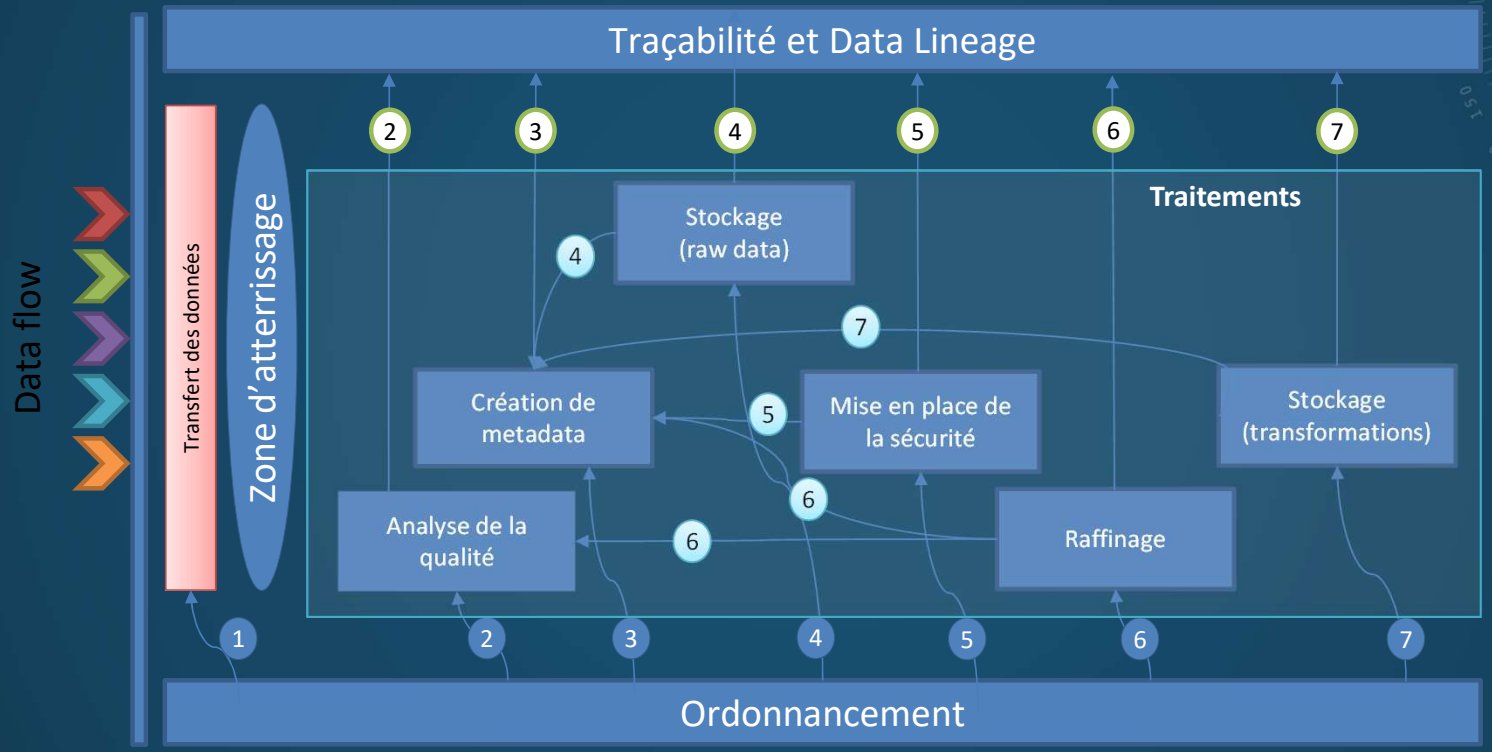
PARCOURS DES DONNÉES : UN TRAVAIL D'ÉQUIPE

Impliquer l'ensemble des forces de l'entreprise pour réussir la stratégie data

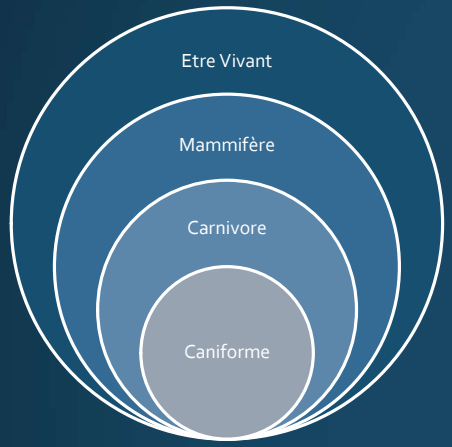
- Marketing
- Finance
- SI
- Commerce
- Production
- Service Client



PARCOURS DES DONNÉES : UNE VISION ORIENTÉE SI



PARCOURS DES DONNÉES : IMPORTANCE DES MÉTADATA



Peau avec des poils
Quatre pattes
Domestique

Meilleur « ami » de l'homme

Les métadata permettent de décrire la donnée et de pouvoir la retrouver grâce à des concepts communs et connus de tous

En entreprise, toutes les données ou ensembles de données doivent être décrits avec un vocabulaire commun à l'entreprise afin de pouvoir les retrouver, les croiser, les transformer

Un DataLake sans indexation des données (i.e. thésaurus) et sans métadata est amené à ne plus être utilisé et à devenir dangereux pour la santé de l'entreprise

LE DATA MANAGEMENT PLAN

Le Data Management (ou gestion des données) consiste à valoriser, à la fois, les données internes et externes dès lors qu'elles enrichissent le patrimoine de données d'une entreprise.

Dans la gestion des données on trouve :

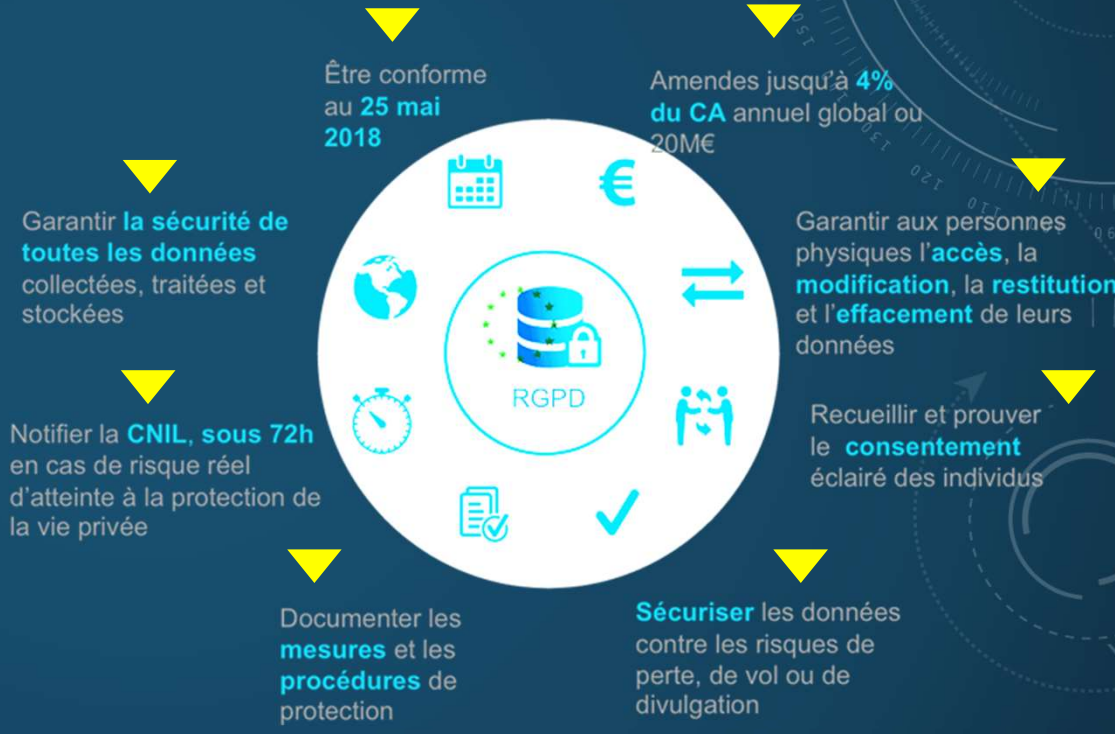
- La collecte
- Le stockage
- L'accessibilité
- L'exploitation
- La sécurisation

Donner du sens aux données c'est les valoriser dans le patrimoine de l'entreprise

GESTION DES DONNÉES ET RGPD

Le Data Management constitue un moyen pour les entreprises de se conformer à la RGPD

Règlement **G**énéral sur la **P**rotection des **D**onnées

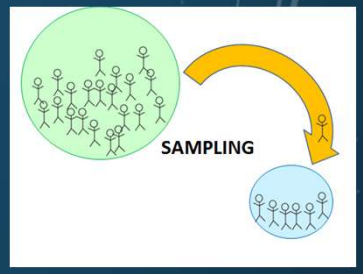


QUALITÉ DES DONNÉES

Quelles sont les dimensions qui décrivent la qualité des données ?

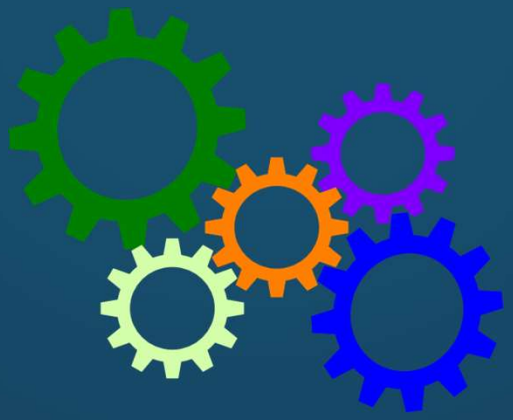
LES BONNES PRATIQUES DE LA QUALITÉ DES DONNÉES

- Le processus de collecte de données doit mettre en place des mesures relatives à la qualité de « bout en bout »
- La qualité des données doit faire l'objet de métriques précis, obtenus par échantillonnages ou par examen global
- La méthode dite « par échantillon » est issue des bonnes pratiques de mesure de la qualité dans le domaine de l'industrie
- Les mesures doivent être horodatées et stockées pour analyser la variation de qualité des données.



QUALITÉ ET SANTÉ DES DONNÉES

- La qualité des données s'exprime à partir de règles et de seuils
- Les mesures de qualité des données sont faites à intervalles réguliers
- L'emploi des sondes logicielles dans les applicatifs est nécessaire
- Les résultats des mesures sont stockés puis analysés
- Un tableau de bord global doit être produit au Data Manager à intervalle régulier
- Un plan de « remédiation » doit être mis en place en cas de problèmes critiques
- L'amélioration continue doit être une habitude et pas une contrainte



BIBLIOGRAPHIE

- *http://www.cil.cnrs.fr/CIL/spip.php?article2949*
- *https://ec.europa.eu/info/law/law-topic/data-protection_fr*
- *https://www.lebigdata.fr/data-2020-2310*
- *http://www.mc2i.fr/Accompagnement-dans-la-mise-en-conformite-RGPD*
- *https://fr.wikipedia.org/wiki/Ontologie_(informatique)*
- *https://protege.stanford.edu/products.php#web-protege*
- *https://www.economie.gouv.fr/entreprises/reglement-general-sur-protection-des-donnees-rgpd*
- *https://www.riskinsight-wavestone.com/2018/05/rgpd-que-va-t-il-se-passer-apres-le-25-mai/*
- *https://www.enssib.fr/bibliotheque-numerique/documents/66017-big-data-et-bibliotheques-traitement-et-analyse-informatiques-des-collections-numeriques.pdf*