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

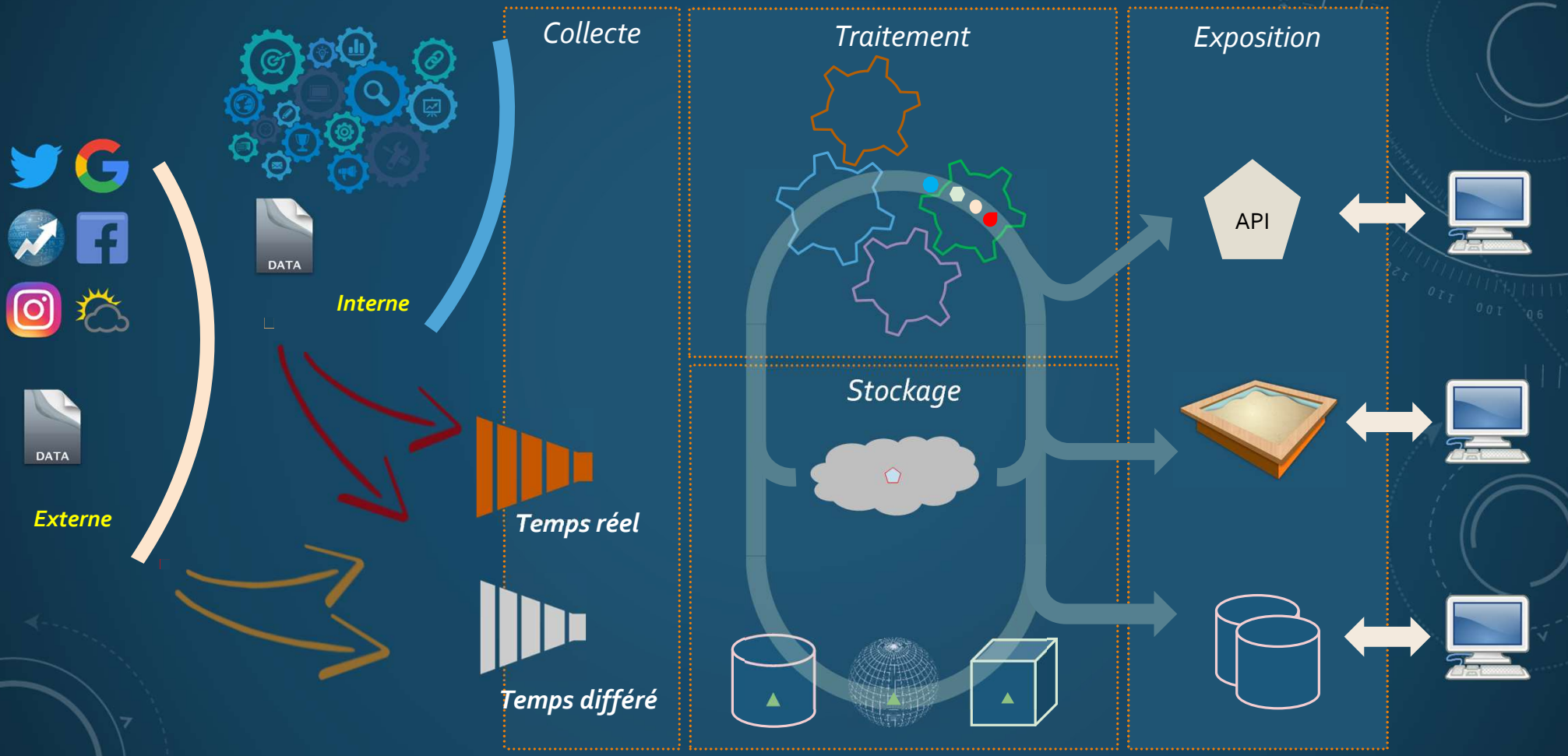
BIG DATA : DES POSSIBILITÉS INFINIES

- L'entreprise doit faire face à une masse de données qu'elle « doit » prendre en compte
- Les infrastructures pour les exploiter peuvent être internes ou externes
- Les couches matérielles (i.e. les ordinateurs) sont « encore » sous contrôle des constructeurs
- Les couches logicielles peuvent être « open source » ou « propriétaire »

- Big Data signifie à la fois grand volume de stockage et grande capacité de calcul
- Dans leur ADN, les logiciels de l'écosystème Big Data permettent une distribution (stockage,calcul)

- Les technologies Big Data sont par nature « open source » (<> GRATUIT)
- Mais l'écosystème Big Data est complexe et en perpétuel changement
- Les relations entre les différents logiciels peuvent être complexes
- Il y a peu de société qui gèrent la pérennité des logiciels
- La communauté « open source » s'occupe principalement de la gestion des améliorations

BIG DATA : UNE ARCHITECTURE À PENSER



BIG DATA : DISTRIBUTION OBLIGATOIRE ?

Un système distribué est un groupe d'ordinateurs qui travaillent ensemble et apparaissant comme un seul ordinateur pour l'utilisateur final

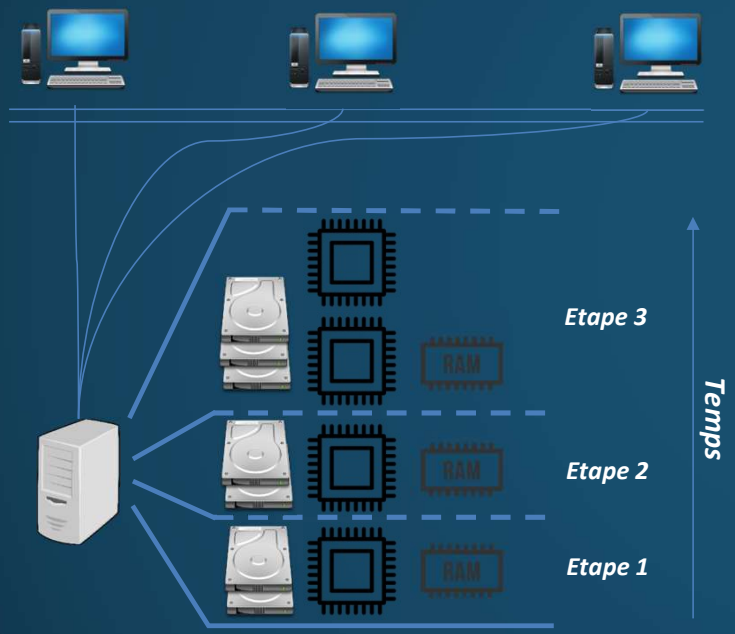
- **Les avantages :**
  - Partage des ressources
  - Mise à l'échelle
  - Tolérance aux pannes / aux fautes
  - Performance et vitesse d'exécution
  - Coût

- **Les inconvénients :**
  - Eléments hétérogènes
  - Latence inter éléments
  - Mémoire distante vs Mémoire locale
  - Synchronisation
  - Panne partielle

- **Les « challenges » :**
  - La sécurité
  - Mise à l'échelle
  - Transparence

BIG DATA : MISE A L'ECHELLE

Scalabilité : « Mise à l'échelle », Adaptation à un accroissement



Scalabilité verticale

Scalabilité horizontale

BIG DATA : COLLECTE DES DONNÉES

La collecte « externe » de données est possible grâce :

- à l'utilisation d'API disponible sur le WEB
- au web scraping (aspiration de site Web)
- à la récupération de fichiers via FTP

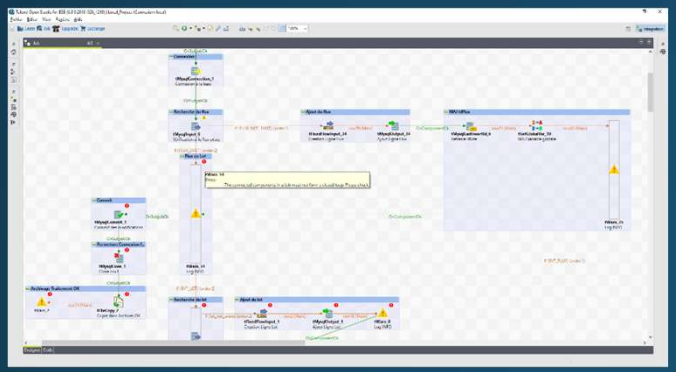
En interne les données s'échangent grâce aux mêmes technologies:

- Par utilisation d'API disponible dans les applications « legacy »
- Via la récupération de fichiers via FTP

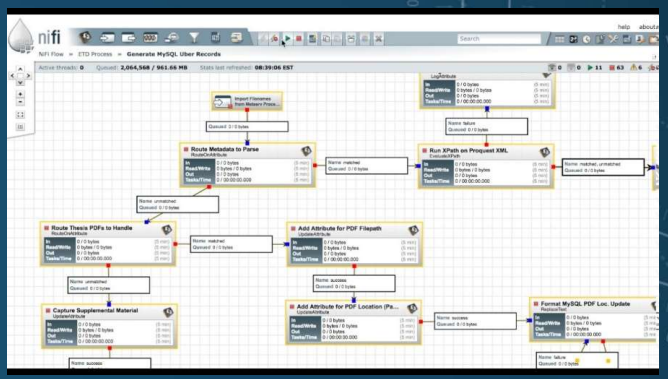
Deux types d'outils peuvent être utilisés :

- **EAI** (Enterprise Application Integration)
  - Permet l'urbanisation Inter Application
  - Normalise les formats d'échanges et les protocoles
- **ETL** (Extracting Transforming and Loading)
  - Permet une connexion multi-protocole aux sources de données
  - Autorise le pré-traitement des données avant stockage

BIG DATA : TRAVAILLER AVEC DES FLUX DE DONNÉES



Talend



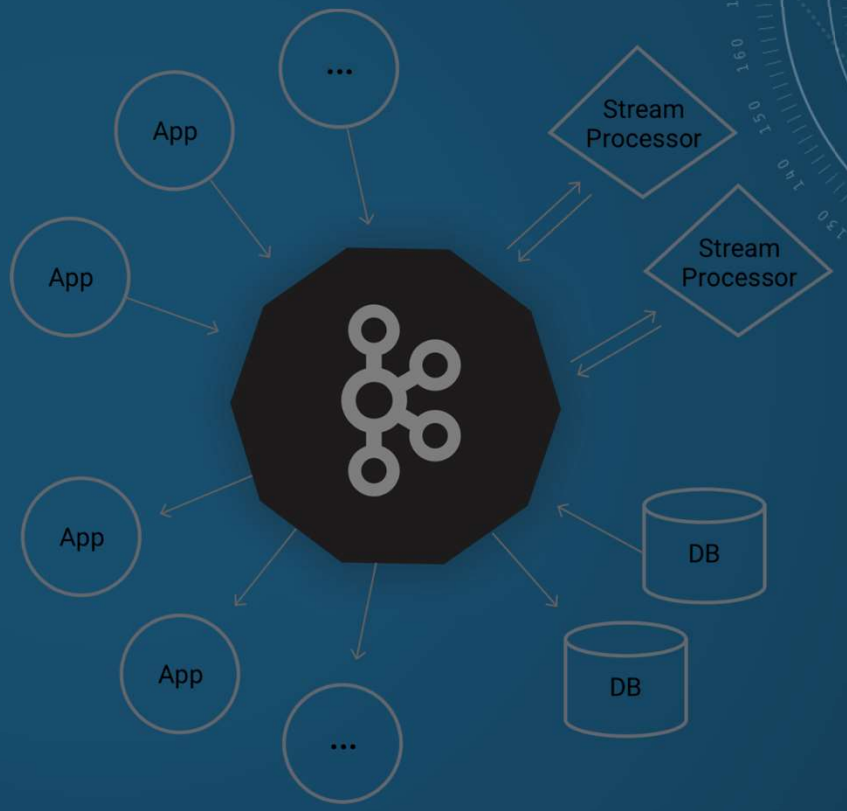
Nifi

Streamsets

BIG DATA : UN FLUX PERMANENT À GÉRER

Kafka est une plateforme :

- De traitement des données en temps réel
- Hautement tolérante aux pannes
- Possibilité de mise à l'échelle
- Possédant de multiples connecteurs
- Interfaçable avec de nombreux langages



BIG DATA : STOCKAGE DES DONNÉES

Le stockage des données dépend de plusieurs critères :

- Le type de données à stocker
  - Relationnel, graphe, document, fichiers, …
- La criticité du stockage
  - Stockage mono-serveur
  - Stockage distribué
- L'écosystème déjà en place
- La formation du personnel

BIG DATA : HADOOP OUI MAIS …

- Hadoop est un ensemble d'outils qui permet le stockage et le traitement d'un important volume de données
- Hadoop désigne aussi , plus généralement , un ensemble d'outils de traitement de données

- Hadoop recouvre deux composants principaux :
  - HDFS (Hadoop Distributed File System)
  - Le framework MapReduce

- HDFS est un système de fichier distribué dans un cluster avec une tolérance aux pannes
- Mais :
  - HDFS ne convient pas pour les applications « Fast Analytics ».
  - Met en place le concept de « Write Once – Read Only »

BIG DATA : TRAITEMENT DES DONNÉES

Le traitement des données couvre les points suivants :

- Le traitement des métadonnées et l'indexation
- Le nettoyage , les corrections de données
- Les interactions avec les stocks de données
- La maintenance des données
- Les audits sur les données
- L'ordonnancement des traitements
- Les traitements demandés par les utilisateurs

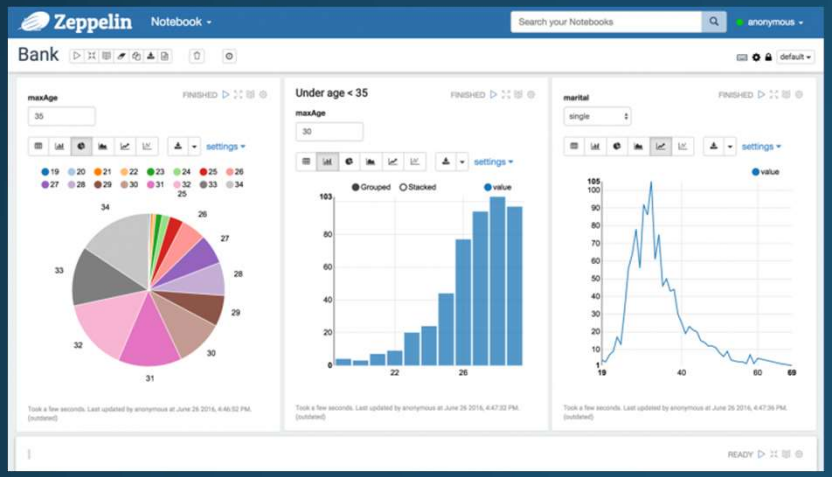
Airflow

BIG DATA : EXPOSITION DES DONNÉES

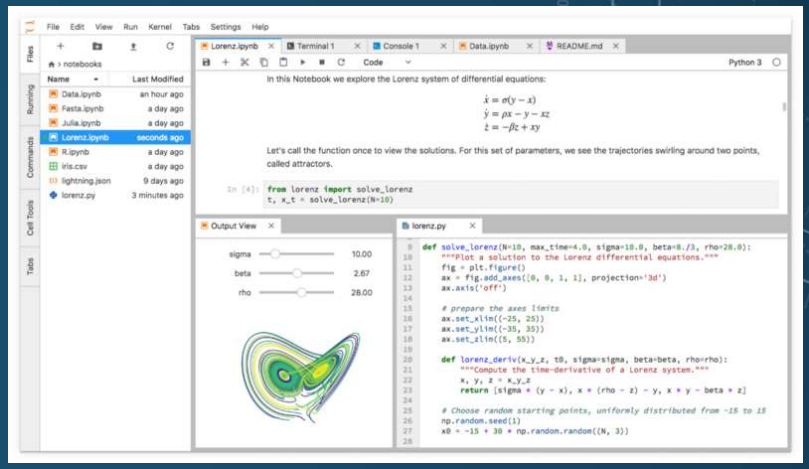
Les données stockées sont présentées aux utilisateurs :

- Par l'intermédiaire d'API, utilisable à la demande et permettant la récupération de données complexes ou d'objets métiers
- Sous la forme d'un « bac à sable », espace voué à la recherche en data science. Cet espace permet l'expérimentation sans débordement sur l'espace de production
- Dans une base de données « classique », permettant ainsi de profiter d'un ensemble de données issu du « stockage » massif, mais dans un container facilement utilisable
- En utilisant un « server » adapté aux besoins de l'analyste

BIG DATA : EXPOSITION DES DONNÉES



Apache Zeppelin est un « serveur » de « notebook », permettant à l'utilisateur d'interagir de différentes manières avec les entrepôts de données



Jupyter Notebook est une plateforme permettant de s'interfacer avec des entrepôts de données avec un grand nombre de langages et des « widgets » interactifs

BIG DATA : VOIR ET ANALYSER LES DONNÉES

Les utilisateurs des données en entreprise peuvent :

- Utiliser les services d'un serveur distant (avec des applications mutualisées)
- Utiliser une application cliente « lourde » installée sur leur ordinateur
- Construire une application « clé en main »

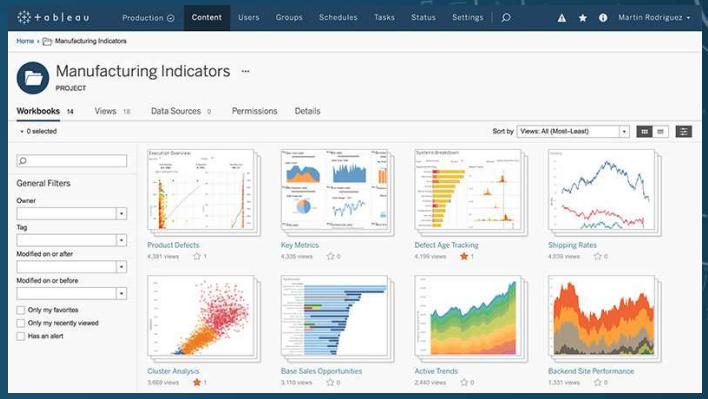
Les applications permettent :

- De visualiser des données de manière « intelligible »
- De concevoir, exécuter des traitements sur des ensembles de données

BIG DATA : VOIR ET ANALYSER LES DONNÉES

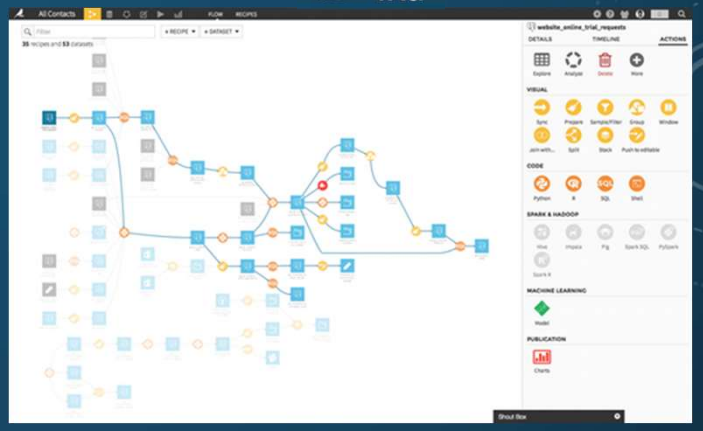
Qlik

+ a b | e a u



Power BI

data iku



BIG DATA : PAAS [PLATEFORME AS A SERVICE]

- Réaliser une architecture Big Data est une « aventure » risquée pour une entreprise
- Des éditeurs ont mis en place des solutions « clé en main » pour les entreprises
- Le cloud est une autre manière de construire de manière maîtrisée une architecture « data »

CONCLUSIONS

QU'AVEZ-VOUS RETENU / QUELLES SONT VOS QUESTIONS ?

BIBLIOGRAPHIE

- *https://www.talend.com/*
- *https://streamsets.com/*
- *https://nifi.apache.org/*
- *https://kafka.apache.org/*
- *http://couchdb.apache.org/*
- *https://flink.apache.org/*
- *https://www.g2.com/categories/graph-databases*
- *https://www.arangodb.com/*
- *https://www.mongodb.com/*
- *https://ignite.apache.org/*
- *https://zeppelin.apache.org/*
- *https://www.dremio.com/*
- *https://www.saagie.com/fr/*
- *https://fr.wikipedia.org/wiki/Hadoop*