INTRODUCTION DATA INTELLIGENCE & DATA SCIENCE

MODULE 3 – INTRODUCTION AU BIG DATA

Informatique – orientation IA – 1DA/IA

PLAN

- Introduction
- Bases de données No SQL
- ETL / ELT
- Architectures Big Data
- Hadoop
- Lien entre IA et Big Data

INTRODUCTION & DÉFINITIONS

INTRODUCTION

BIG DATA - DÉFINITION

« Domaine technologique dédié à l'analyse de très grands volumes de données informatiques (petaoctets), issus d'une grande variété de sources, tels les moteurs de recherche et les réseaux sociaux ; ces grands volumes de données.»

(larousse.fr)

« Le terme Big Data décrit des ensembles de très gros volumes de données – à la fois structurées, semi-structurées ou non structurées – qui peuvent être traitées et exploitées dans le but d'en tirer des informations intelligibles et pertinentes.»

(https://www.lemagit.fr/definition/Big-Data)

« Le big data, les mégadonnées, ou les données massives, désigne les ressources d'informations dont les caractéristiques en termes de volume, de vélocité et de variété imposent l'utilisation de technologies et de méthodes analytiques particulières pour créer de la valeur, et qui dépassent en général les capacités d'une seule et unique machine et nécessitent des traitements parallélisés.»

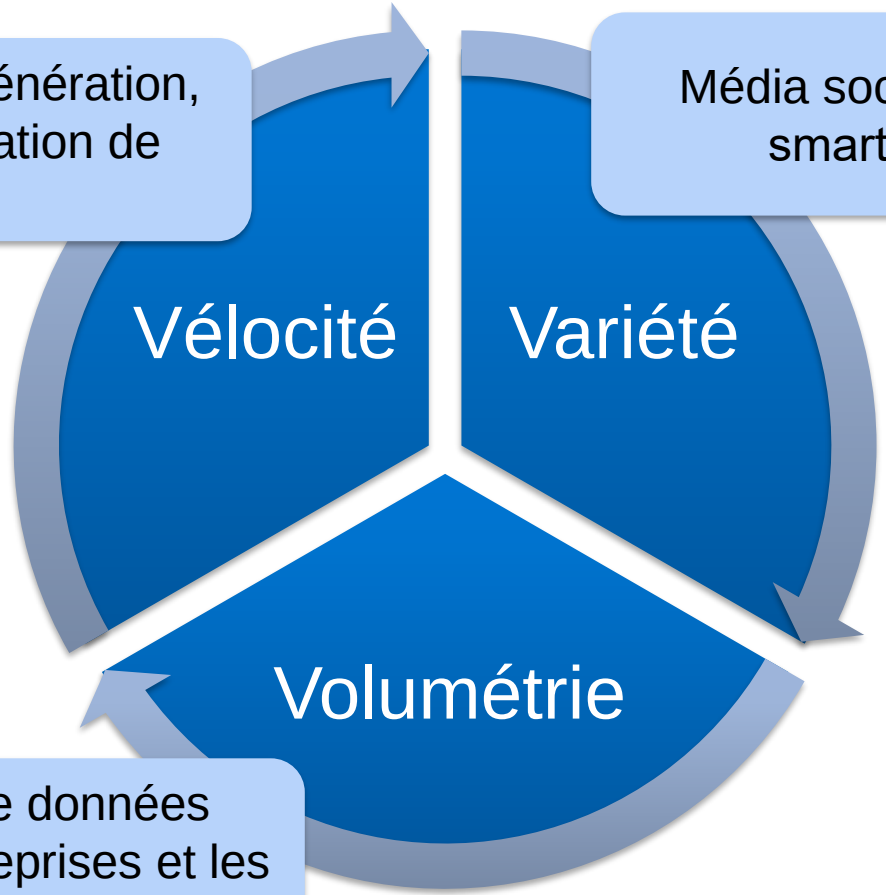
(wikipedia.org)

INTRODUCTION

BIG DATA - DÉFINITION (3V)

Fréquence élevée de génération, récupération, visualisation de données

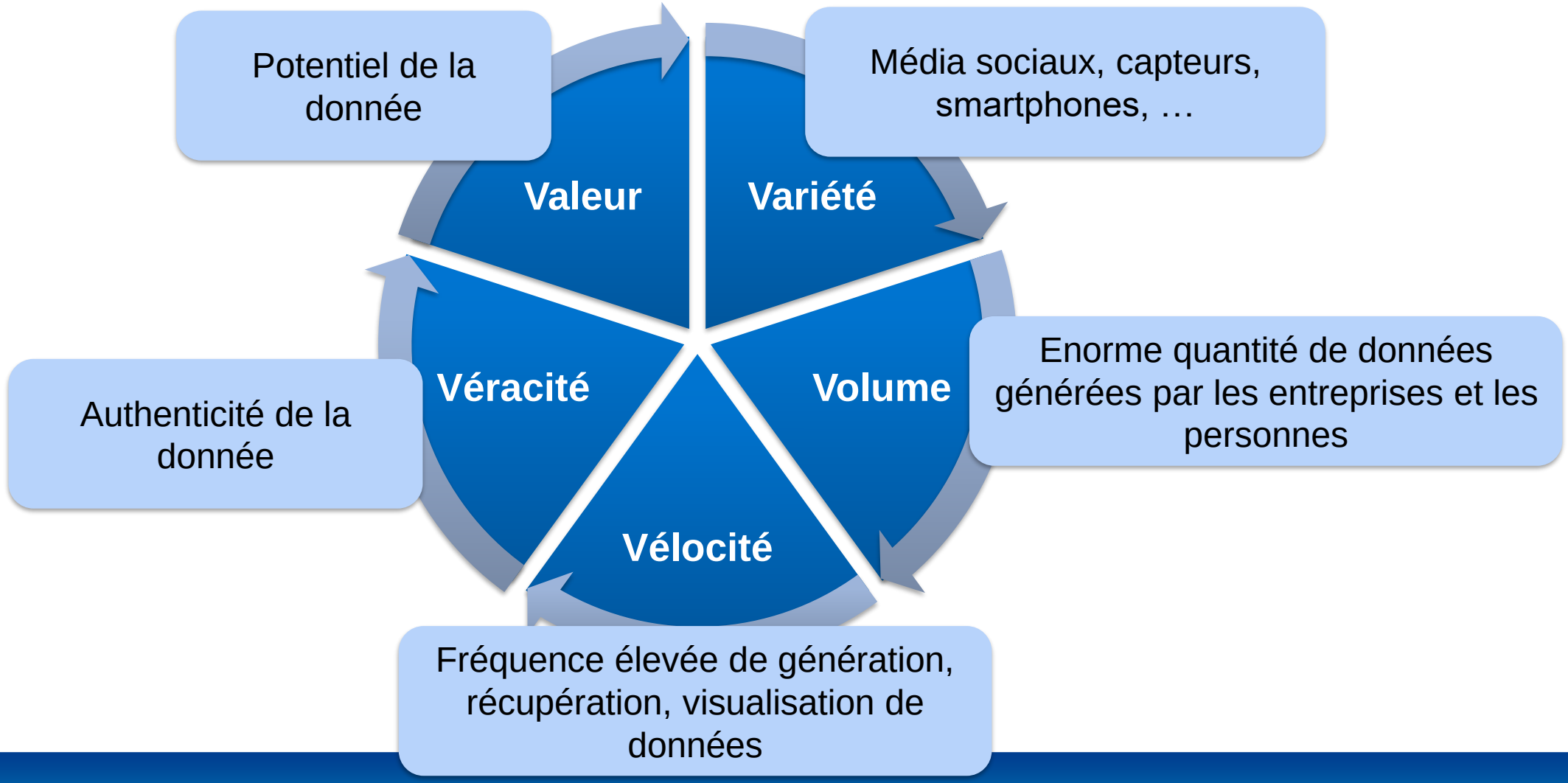
Média sociaux, capteurs, smartphones, …



Enorme quantité de données générées par les entreprises et les personnes

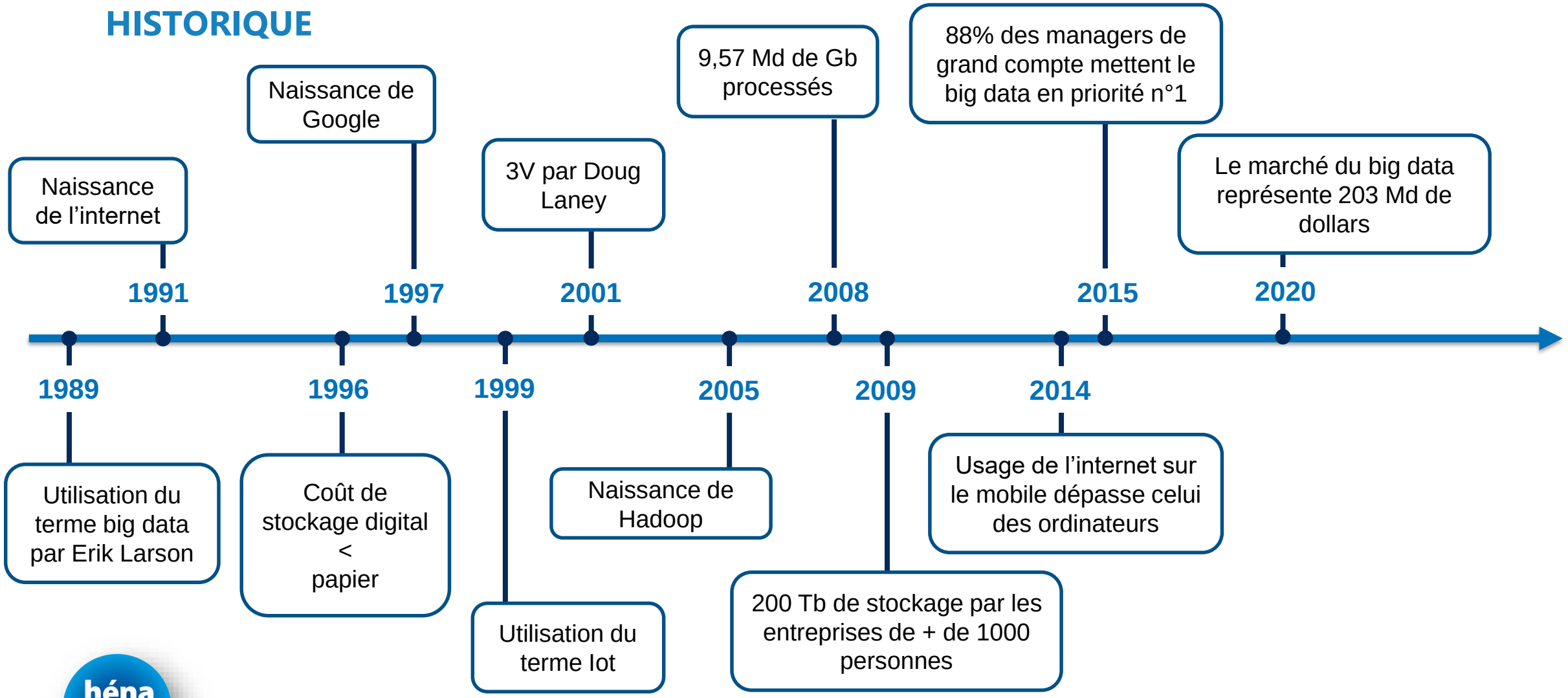
INTRODUCTION

BIG DATA - DÉFINITION (5V)



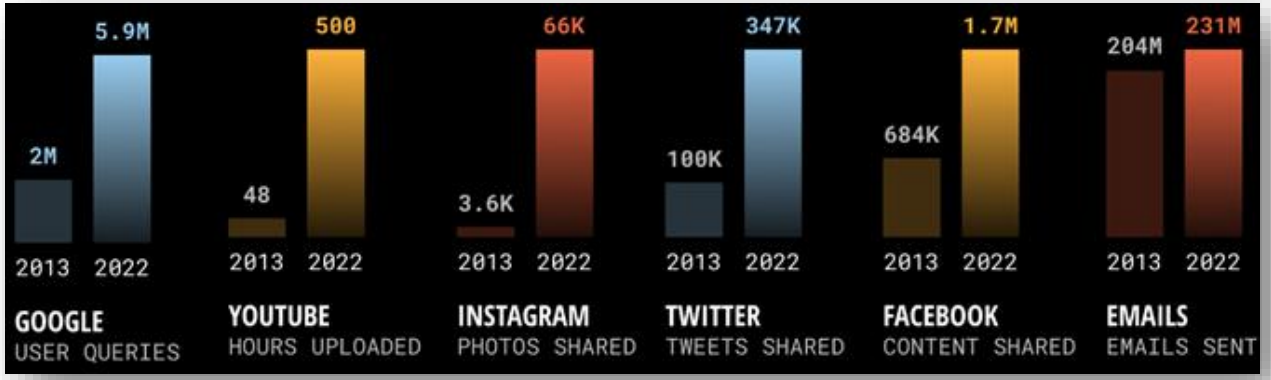
INTRODUCTION

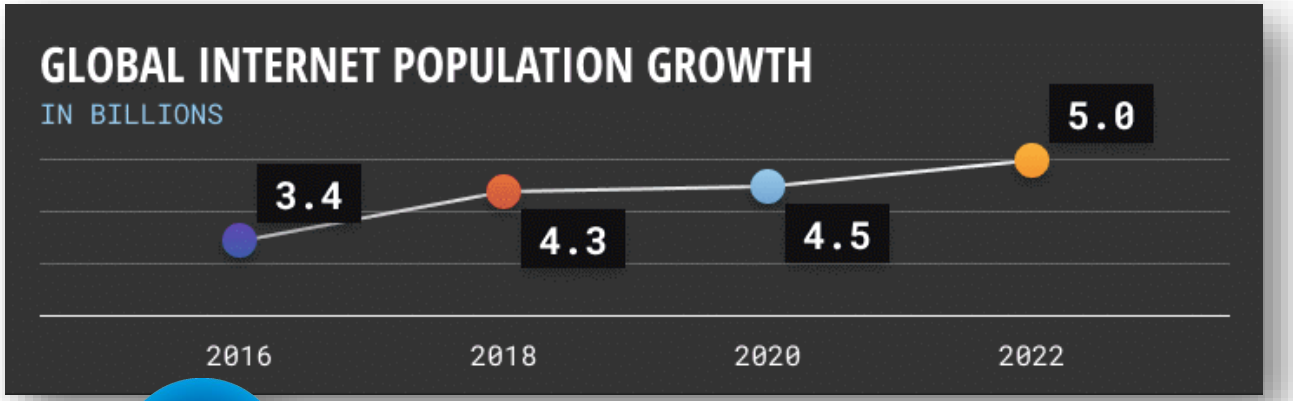
HISTORIQUE

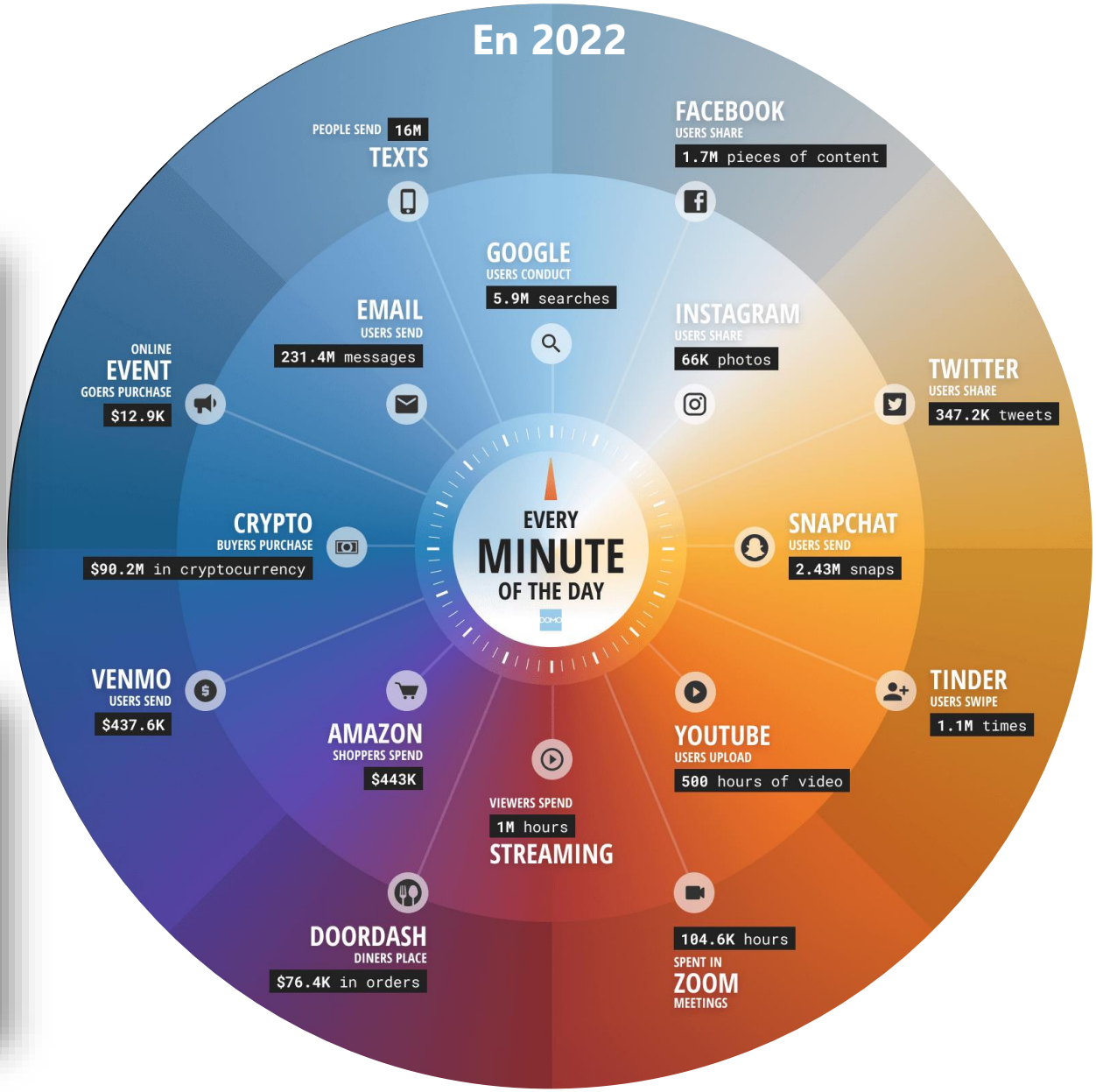


INTRODUCTION

QUELQUES CHIFFRES (DOMO.COM)

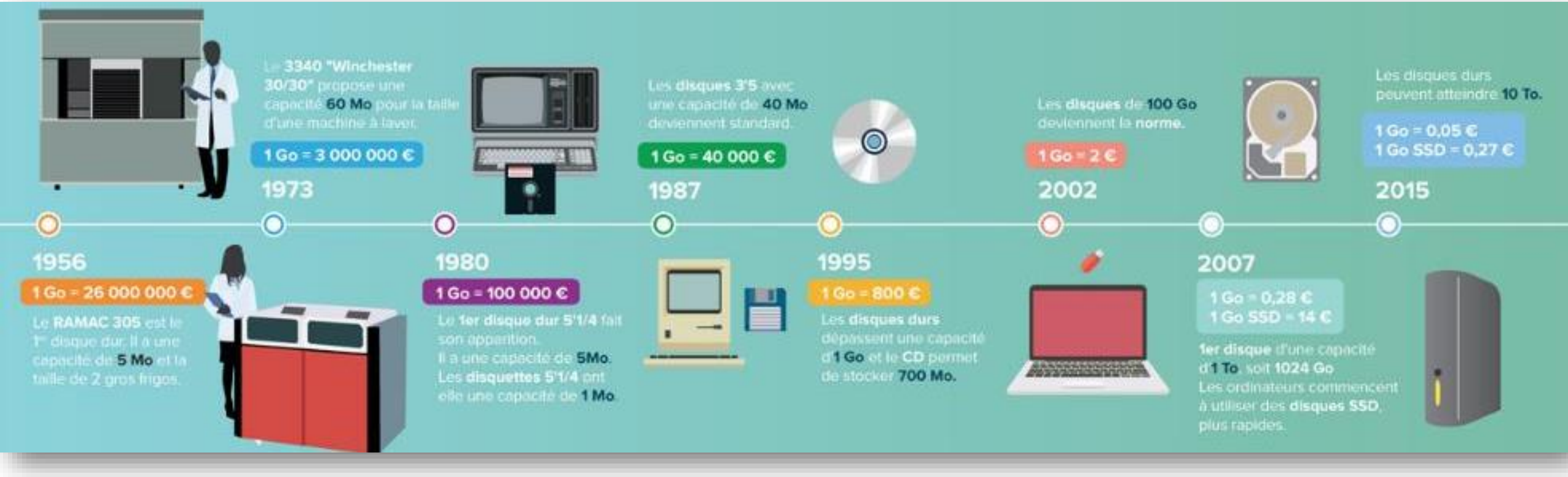






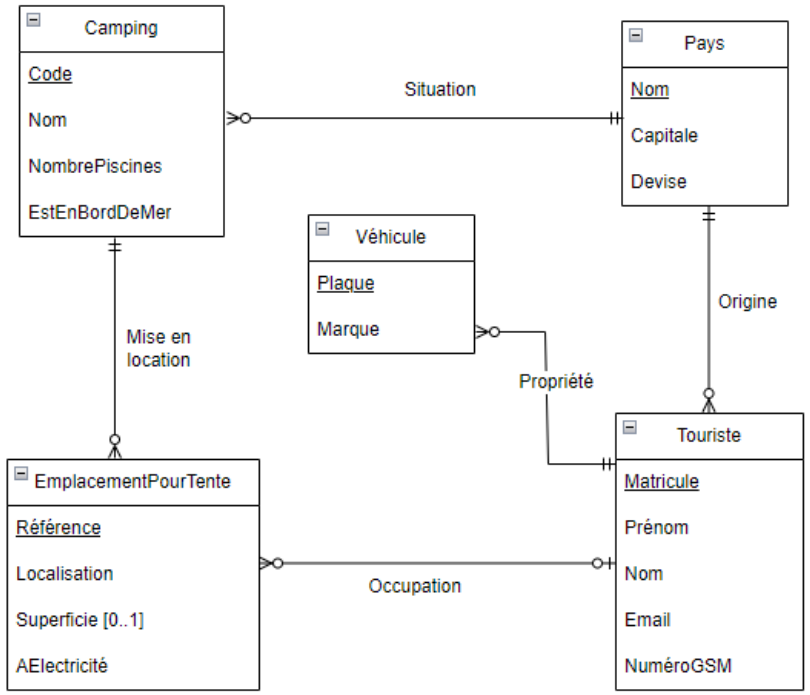
INTRODUCTION

QUELQUES CHIFFRES – COÛT DE STOCKAGE

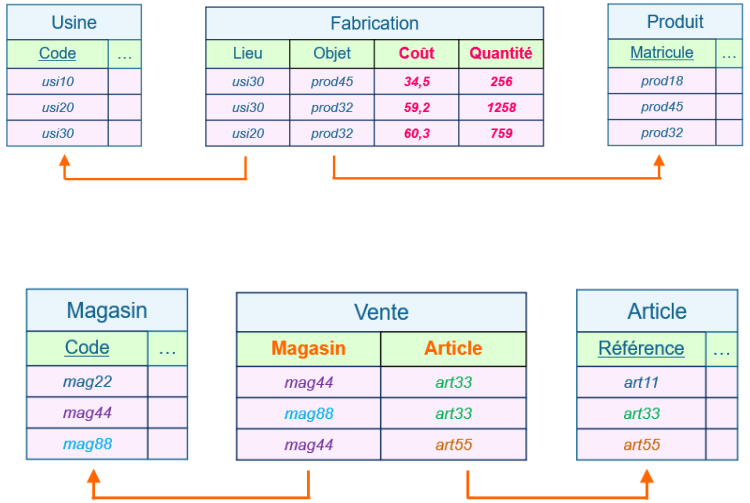


INTRODUCTION

SYSTÈME DE STOCKAGE TRADITIONNEL - SGBD RELATIONNEL



Modèle conceptuel



Modèle logique

Exemples de SGDBr

(Source : cours d'OSD)

INTRODUCTION

SYSTÈME DE STOCKAGE TRADITIONNEL – PROPRIÉTÉS ACID

Atomicité : Les mises à jours des données doivent être « atomiques ». (Totalement réalisées ou pas du tout)

Cohérence : Les modifications apportées à la base doivent être valides. (respecter le modèle implémenté dans la base de données)

Isolation : Les transactions lancées au même moment ne doivent jamais interférer entre elles.

Durabilité : Toutes les transactions sont lancées de manière définitive.

INTRODUCTION

SYSTÈME DE STOCKAGE NO SQL – PROPRIÉTÉS BASE

Basically **A**vailable : le système garantit un taux de disponibilité de la donnée

Soft-State : La base NoSQL n'a pas à être cohérente à tout instant

Eventually Consistent : À terme, la base atteindra un état cohérent

INTRODUCTION

POURQUOI LE BIG DATA ? LIMITES DES SYSTÈMES TRADITIONNELS

Système traditionnel (SGBDR)		Big Data
Données structurées		Données non structurées
Temps de traitement		Flots continus (temps réel)
Pas d'historique		Données méta-taguées
Source unique		Sources très disparates

INTRODUCTION

SOURCES DE DONNÉES

- Données structurées

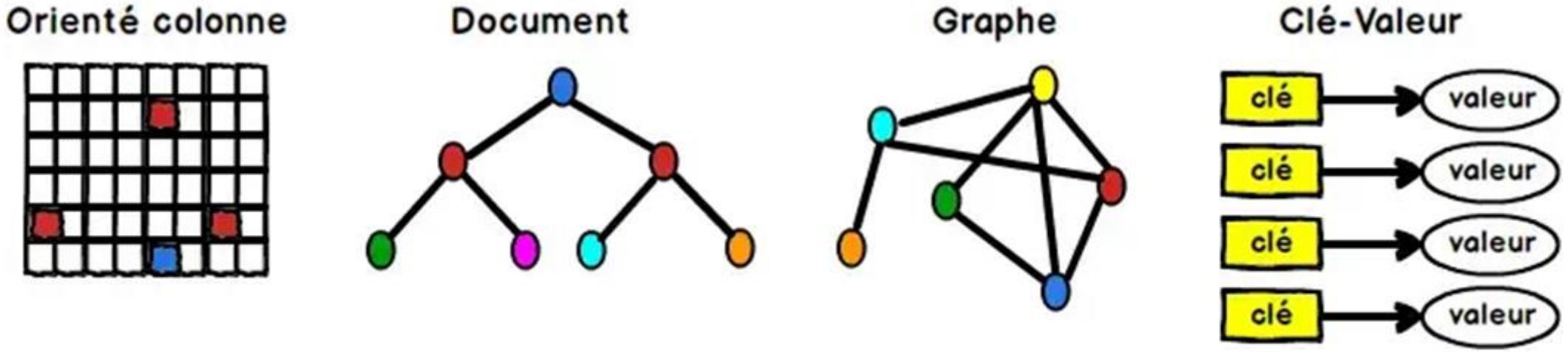
- Données semi-structurées

- Données complexes

Sources internes

Open Data
Brussels open datastore
cirb-cibg

Sources externes



BASES DE DONNÉES NO SQL

BASES DE DONNÉES NO SQL

DÉFINITION

> **Définition :** Bases not only SQL : Ce terme désigne l'ensemble des bases de données qui s'opposent à la notion relationnelle des SGBDR.

- Principes
  - Elles privilégient la disponibilité et le partitionnement à la cohérence
  - Langages permettant d'aller plus loin dans la programmation que le SQL (Java, PHP)
  - Pas de schéma pour les données (Flexibilité)
  - Données de structures complexes ou imbriquées
  - Scalabilité Horizontale et linéaire(Partitionnement des données et partage des calculs)
  - Mode d'utilisation : peu d'écritures, beaucoup de lectures
  - Données distribuées : on a souvent la possibilité d'utiliser des algorithmes MapReduce.
- Quatre types :

BASES DE DONNÉES NO SQL

BASE DE DONNÉES ORIENTÉE « CLÉ – VALEUR »

Définition :

- Paradigme « clé-valeur »
- Valeur = string ou objet plus complexe

Forces :

- Simplicité, scalabilité, disponibilité

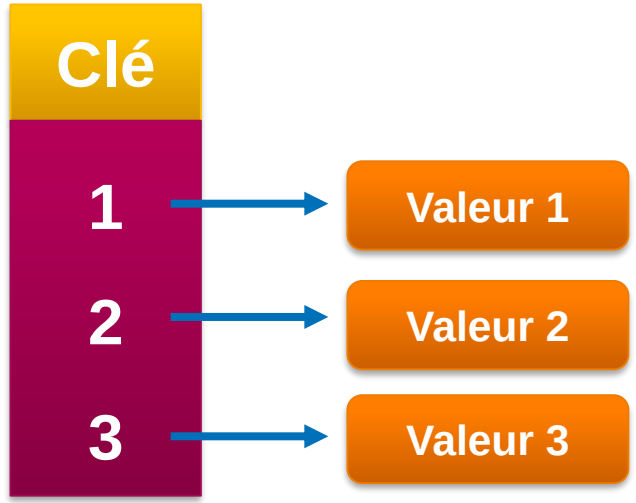
Faiblesses :

- pas de requête sur le contenu des objets
- pas adaptée pour les modèles complexes

Opérations possibles :

- Création : créer un nouveau couple (k, v)
- Lecture, modification et suppression : en connaissant la clé

Définition



Représentation

Principales solutions :

redis, memcached, (Voldemort), riak, pickleDB, apache Ignite, ORACLE BERKELEY DB

Exemples d'utilisation :

- Gestion de panier d'achat
- Collecte d'événements (jeu en ligne)

Exemples

BASES DE DONNÉES NO SQL

BASE DE DONNÉES ORIENTÉE « DOCUMENT »

Définition

Définition

- Paradigme « clé-valeur » ;
- La valeur = un document (JSON ou XML) ;

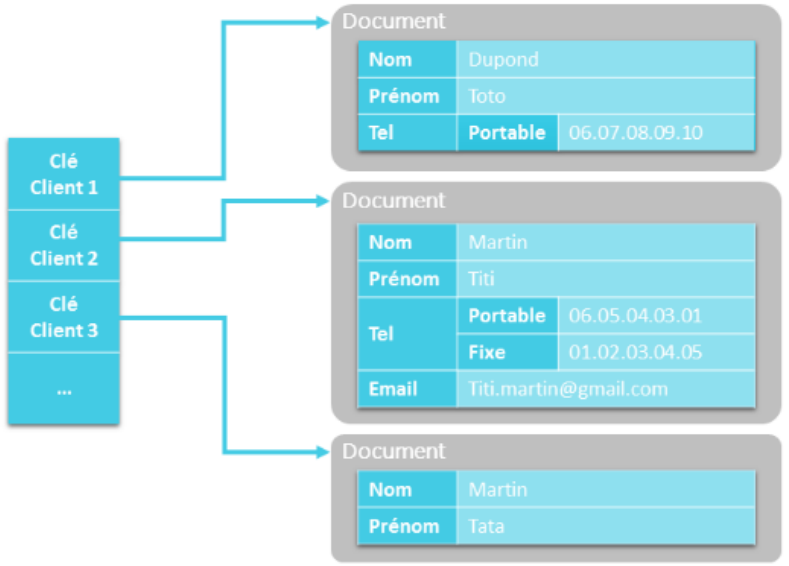
Forces :

- Documents structurés
- Pas de définition de structure nécessaire
- Récupération possible de données structurées via la clé uniquement

Faiblesse :

- Pas adaptée pour les données non structurées ou interconnectées

Représentation



Exemples

Principales solutions :

mongoDB, MarkLogic, OrientDB, Amazon DynamoDB, CouchDB relax, IBM Informix database software, CRATE.IO, RethinkDB, RAVENDB Safe by Default, Optimized for Efficiency, CLUSTERPOINT

Exemples d'utilisation :

- Les données clients
- La gestion catalogue de produits

BASES DE DONNÉES NO SQL

STRUCTURE D'UN DOCUMENT (JSON)

- JSON : Javascript Object Notation
- Format de données permettant un transfert standardisé
- Simple à lire et écrire
- Se compose de différents types de données :
  - **String :** `{ "name" : "Jones" }`
  - **Nombre :**
```
{ "number_1" : 210, "number_2" : 215, "number_3" : 21.05, "number_4" : 10.05 }
```
  - **Objet :**
```
{ "Influencer" : { "name" : "Jaxon" , "age" : "42" , "city" : "New York" } }
```
  - **Tableau :**
```
{ "Influencers" : [ { "name" : "Jaxon", "age" : 42, "Works At" : "Tech News" },{ "name" : "Miller", "age" : 35, "Works At" : "IT Day" } ] }
```
  - **Booléen :** `{ "AllowPartialShipment" : false }`
  - **Valeur nulle :** `{ "Special Instructions" : null }`

BASES DE DONNÉES NO SQL

STRUCTURE D'UN DOCUMENT (JSON) - EXEMPLES

```
{
 "fruits": [
   { "kiwis": 3,
     "mangues": 4,
     "pommes": null
   },
   { "panier": true },
 ],
 "legumes":
   { "patates": "amandine",
     "figues": "de barbarie",
     "poireaux": false
   }
}
```

Exemple de panier

```
{
  "menu": "Fichier",
  "commandes": [
      {
          "titre": "Nouveau",
          "action":"CreateDoc"
      },
      {
          "titre": "Ouvrir",
          "action": "OpenDoc"
      },
      {
          "titre": "Fermer",
          "action": "CloseDoc"
      }
   ]
}
```

Exemple de menu contextuel

BASES DE DONNÉES NO SQL

BASE DE DONNÉES ORIENTÉE « COLONNE »

Définition

- Tables constituées de lignes et colonne
- « Ressemble » aux SGBDr
- Colonnes dynamiques

Forces :

- Flexibilité
- Temps de traitement
- Non stockage des valeurs « null »
- Historisation à la valeur

Faiblesses :

- Pas adaptée pour les données non structurées ou interconnectées

Définition

Clé	Identité		
	Suffixe	Prénom	Nom
1		Mick	Jameson
2	Dr	Jack	Mickeal
3	PhD	Min	Lee

Clé	Identité		
	Email	Téléphone	Adresse
1	mick.jam@gmail.com	+33XXXXXX	
2	jack.mich@outlook.com	+33XXXXXX	75000 Paris
3	min.lee@orange.com	+33XXXXXX	

Représentation

Principales solutions :

Exemples d'utilisation :

- Suivi de colis
- Analyse de données issues de capteurs (Iot)

Exemples

BASES DE DONNÉES NO SQL

BASE DE DONNÉES ORIENTÉE « GRAPHE »

Définition

- Basée sur la théorie des graphes
- Nœuds, relations, propriétés
- Représente bien le monde réel

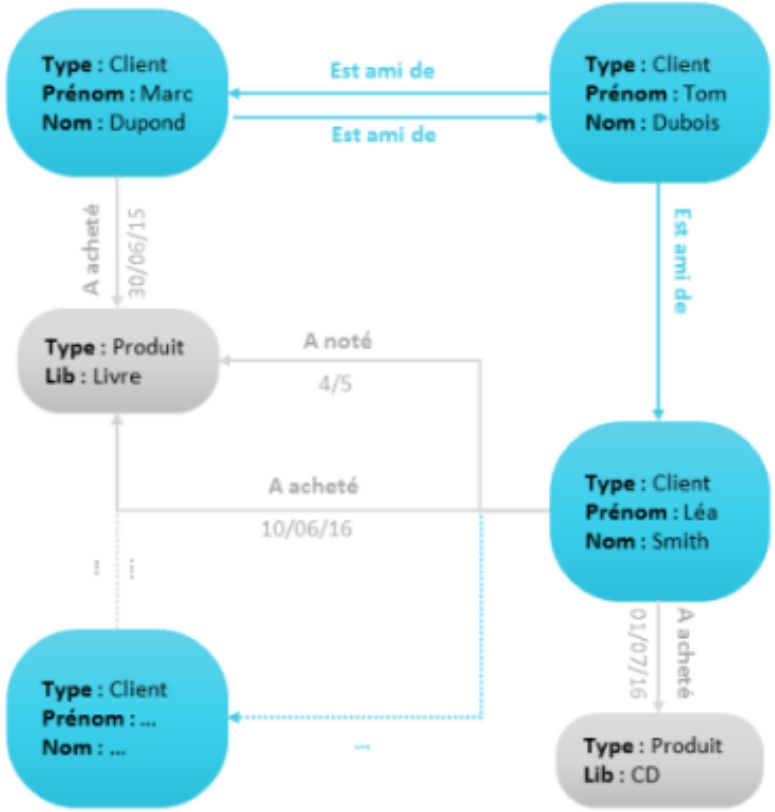
Forces :

- Adaptées aux objets organisés en réseaux
- Application des algorithmes de théorie des graphes
- Visualisation native en graphe

Faiblesses :

- Pas adaptée pour les données non fortement connectées

Définition



Représentation

Principales solutions :

neo4j, OrientDB, Titan, weaver alpha, InfiniteGraph, INFO GRID, *Sparksee, Cayley, ArangoDB, HyperGraphDB

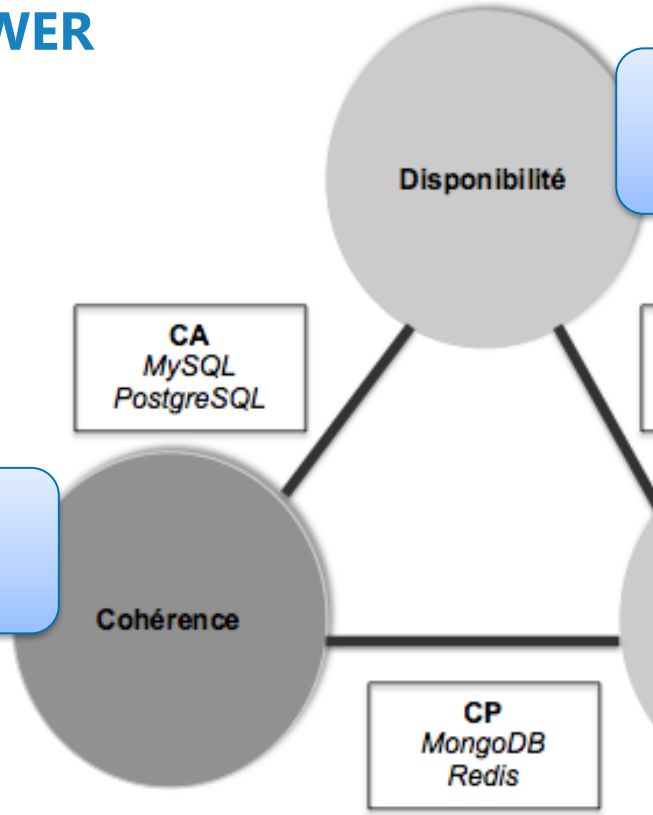
Exemples d'utilisation :

- Moteur de recommandation
- Détection de la fraude
- Données géo spatiales

Exemples

BASES DE DONNÉES NO SQL

THÉORÈME DE BREWER



Tant que le système tourne (distribué ou non), la donnée doit être disponible

Une donnée n'a qu'un seul état visible quel que soit le nombre de réplicas

Quel que soit le nombre de serveurs, toute requête doit fournir un résultat correct

Dans toute base de données, vous ne pouvez respecter au plus que 2 propriétés parmi la cohérence, la disponibilité et la distribution.

Eric A. Brewer


ETL
Extract
Transformation
Load
VS
ELT
Extract/Load
Transformation

ETL & ELT

ETL & ELT

DÉFINITION

- ETL signifie **Extract, Transform, Load**
- Le processus ETL comprend 3 étapes :
  - Extraction des données d'un système source (SI opérationnel)
  - Transformation de ces données (filtres, calculs, rejets…)
  - Loading (chargement) des données transformées dans un système cible (DataWarehouse)

- ELT signifie **Extract, Load, Transform**
- Le processus ETL comprend 3 étapes :
  - Extraction des données d'un système source (SI opérationnel)
  - Loading (chargement) des données non-transformées dans un système cible
  - Transformation de ces données via le moteur de stockage (filtres, calculs, rejets…)

ETL & ELT

L'EXTRACTION

- **Extraction de données provenant de diverses sources**
- **Types d'extraction :**
  - **Notification de mise à jour** : le système source vous avertit lorsqu'un enregistrement de données change.
  - **Extraction progressive :** le système recherche les modifications à intervalles réguliers, par exemple une fois par semaine, une fois par mois ou à la fin d'une campagne.
  - **Extraction complète :** rechargement de l'ensemble des données.
- **Solutions d'extraction de données :**

ETL & ELT

LA TRANSFORMATION

- Différents types de transformation existent :
  - **Nettoyage** : supprime les erreurs et mappe les données source avec le format des données cibles.
  - **Révision du format :** convertit les données, telles que les jeux de caractères, les unités de mesure et les valeurs date/heure en un format cohérent. (exemple : les différentes unités)
  - **Déduplication :** suppression des doublons
  - **Dérivation** : génération de nouvelles valeurs à partir d'autres (Total à partir de Prix et quantité)
  - **Division** : diviser un attribut de colonne ou de données en de multiples colonnes dans le système cible
  - **Résumé :** créer de nouvelles métriques (ex. CA / Client à partir des factures)
  - **Chiffrement** : Chiffrer les données avant leur transfert vers la base de données cible.
  - …

ETL & ELT

LE CHARGEMENT

Bases de données relationnelles

Bases de données No SQL

Bases de données

Définition

- architecture de stockage de données
- permettant de stocker des données brutes provenant de sources diverses

Solutions existantes :

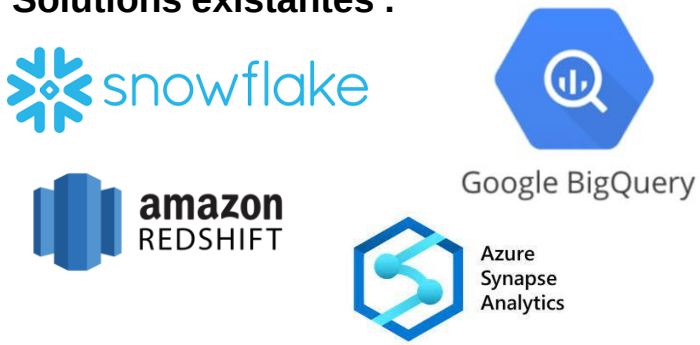
Datalake

Définition

- L'entrepôt de données est un modèle pour soutenir le flux de données des systèmes opérationnels vers les systèmes décisionnels.
- Données traitées et structurées

➔ Business Intelligence

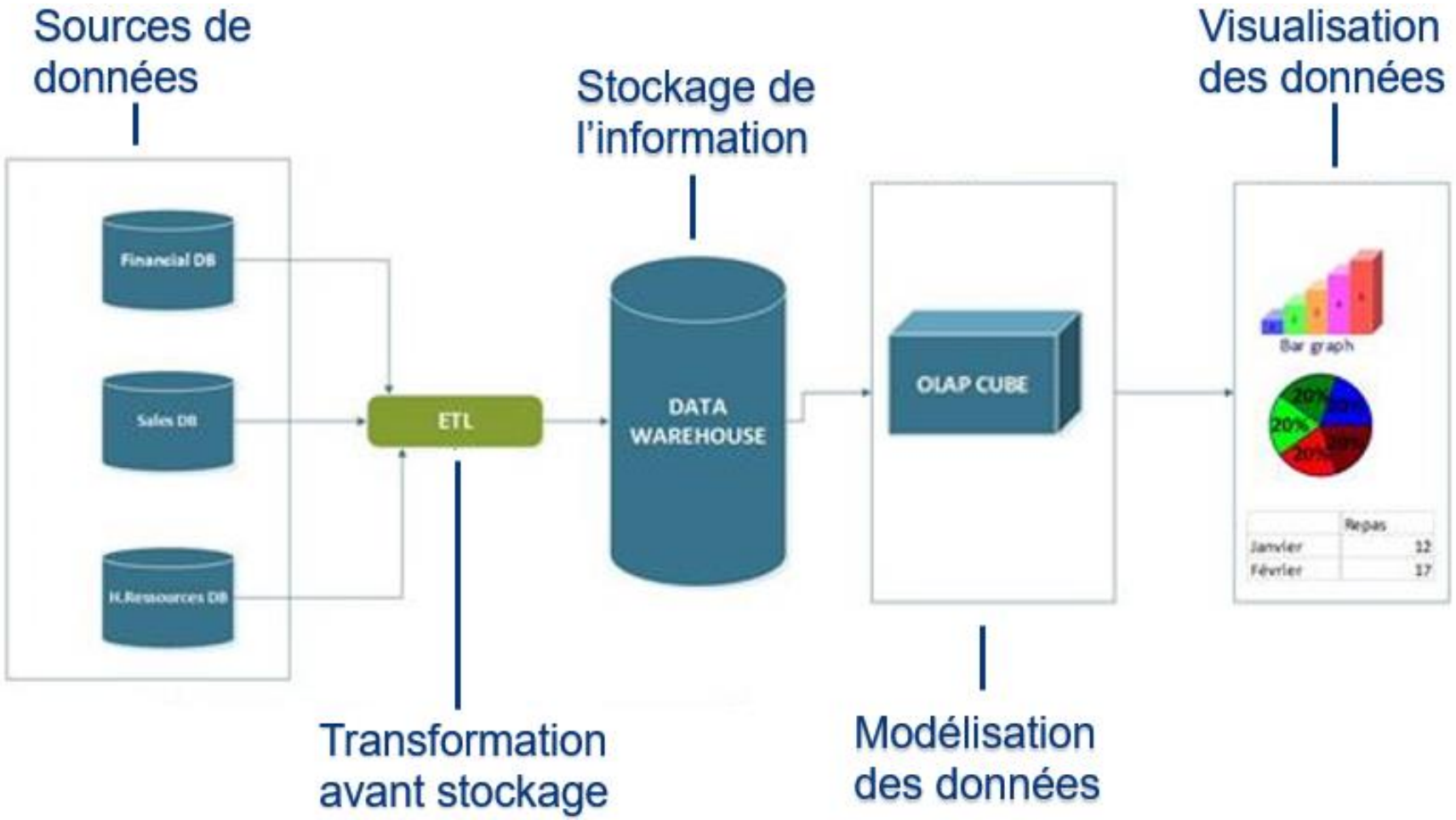
Solutions existantes :



Data Warehouse

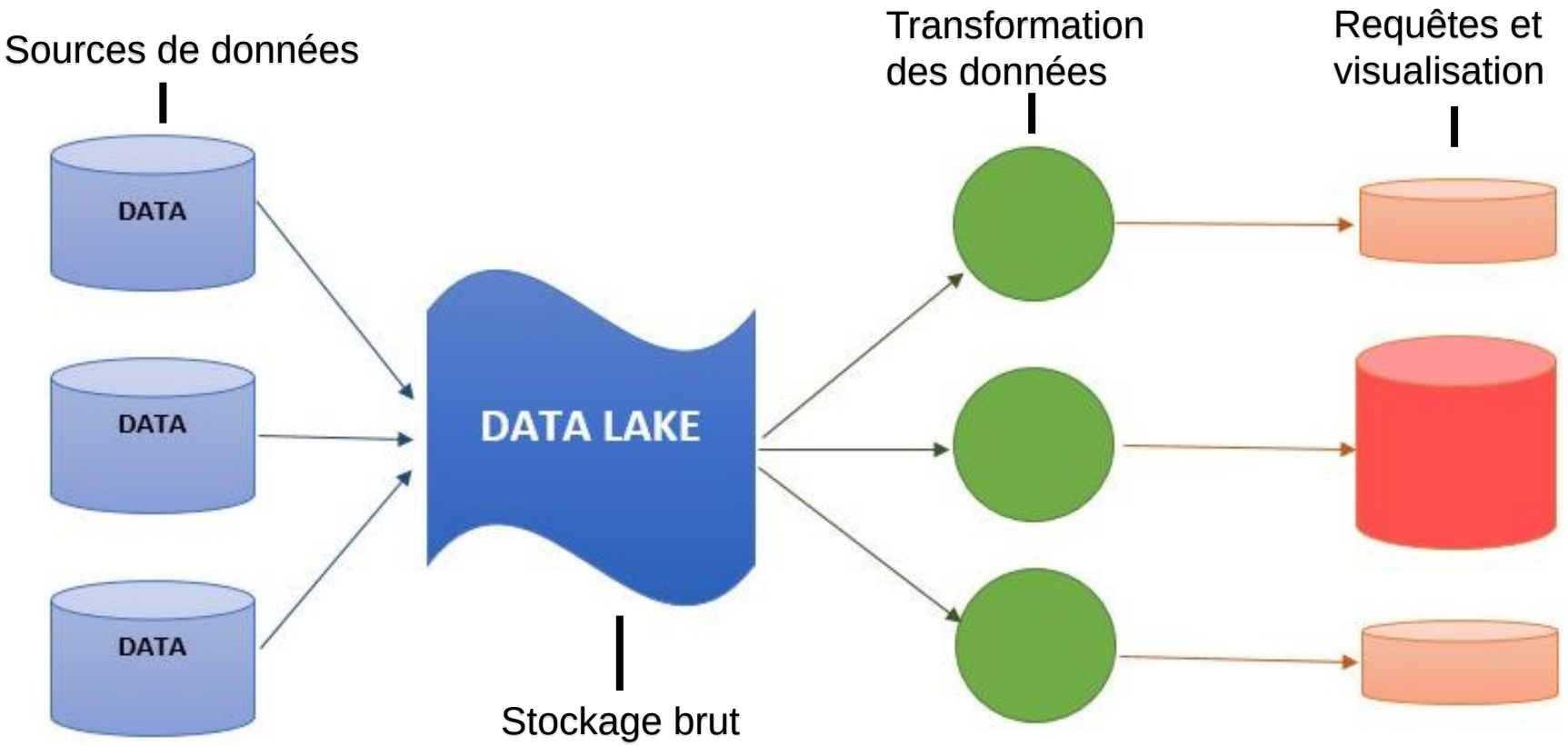
ETL & ELT

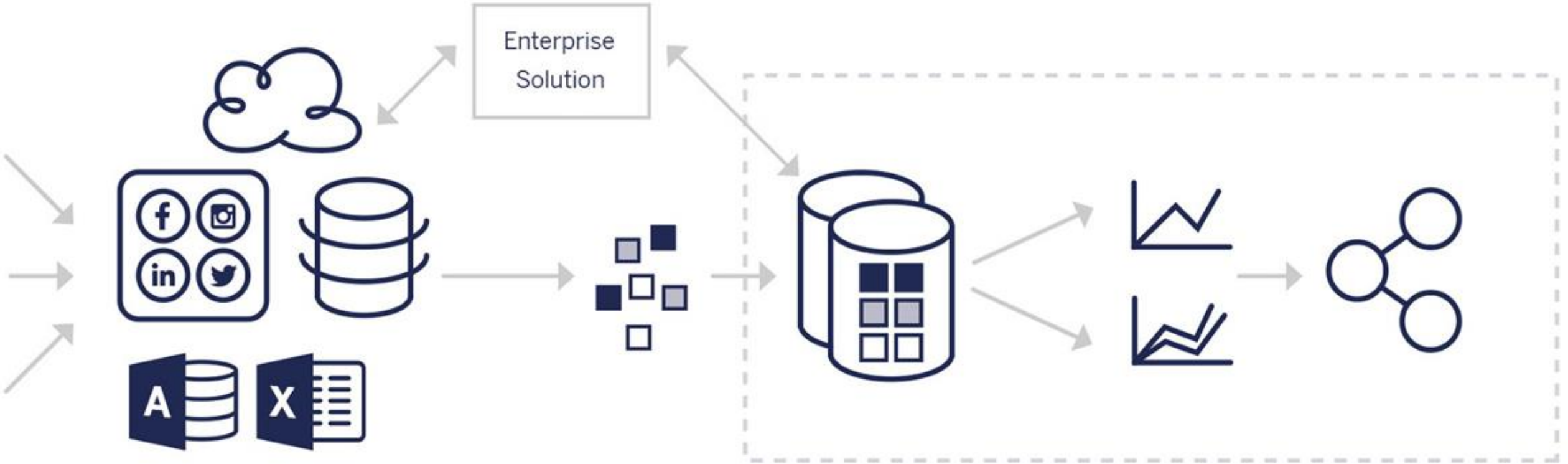
PROCESSUS ETL



ETL & ELT

PROCESSUS ELT





ARCHITECTURES BIG DATA

ARCHITECTURES BIG DATA

LE DATA LAKE

« Le data lake est une méthode de stockage rapide qui regroupe un ensemble de données dans un même espace. Ce système permet aux entreprises de stocker des données brutes, issues de très nombreuses sources, dans l'attente de les organiser et de les analyser. »

(https://blog.hubspot.fr/marketing/data-lakes)

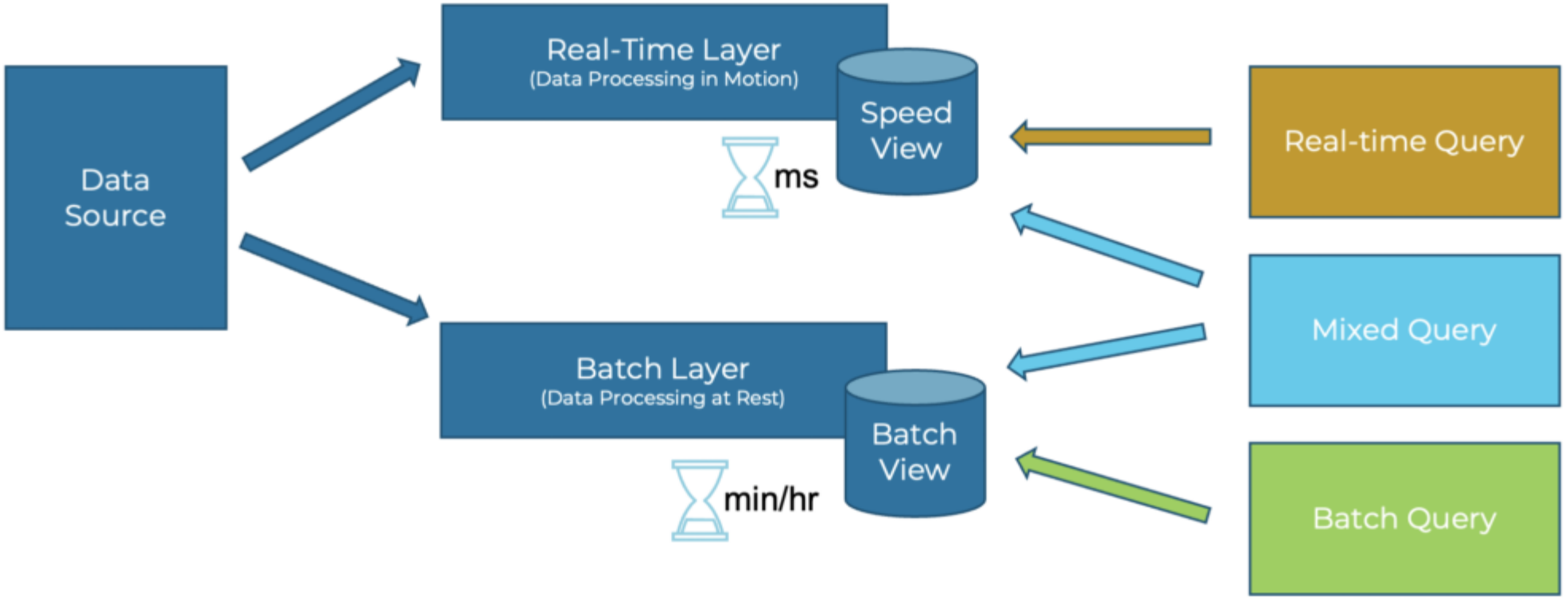
« Un lac de données (en anglais datalake) est une méthode de stockage de données massives utilisée par le big data. Ces données sont gardées dans leurs formats originaux ou sont très peu transformées. Le lac de données donne la priorité au stockage rapide et volumineux de données hétérogènes en adoptant une architecture en cluster. Il n'est pas optimisé pour les requêtes SQL comme les SGBD relationnels classiques, et s'écarte des Propriétés ACID traditionnelles. On parle depuis 2010 de SGBD NoSQL. »

(wikipedia.org)

- ➔ Système de stockage
- ➔ Formats originaux de données
- ➔ Données hétérogènes
- ➔ Architecture en cluster
- ➔ ~~Propriétés ACID~~
- ➔ SGBD NoSQL

ARCHITECTURES BIG DATA

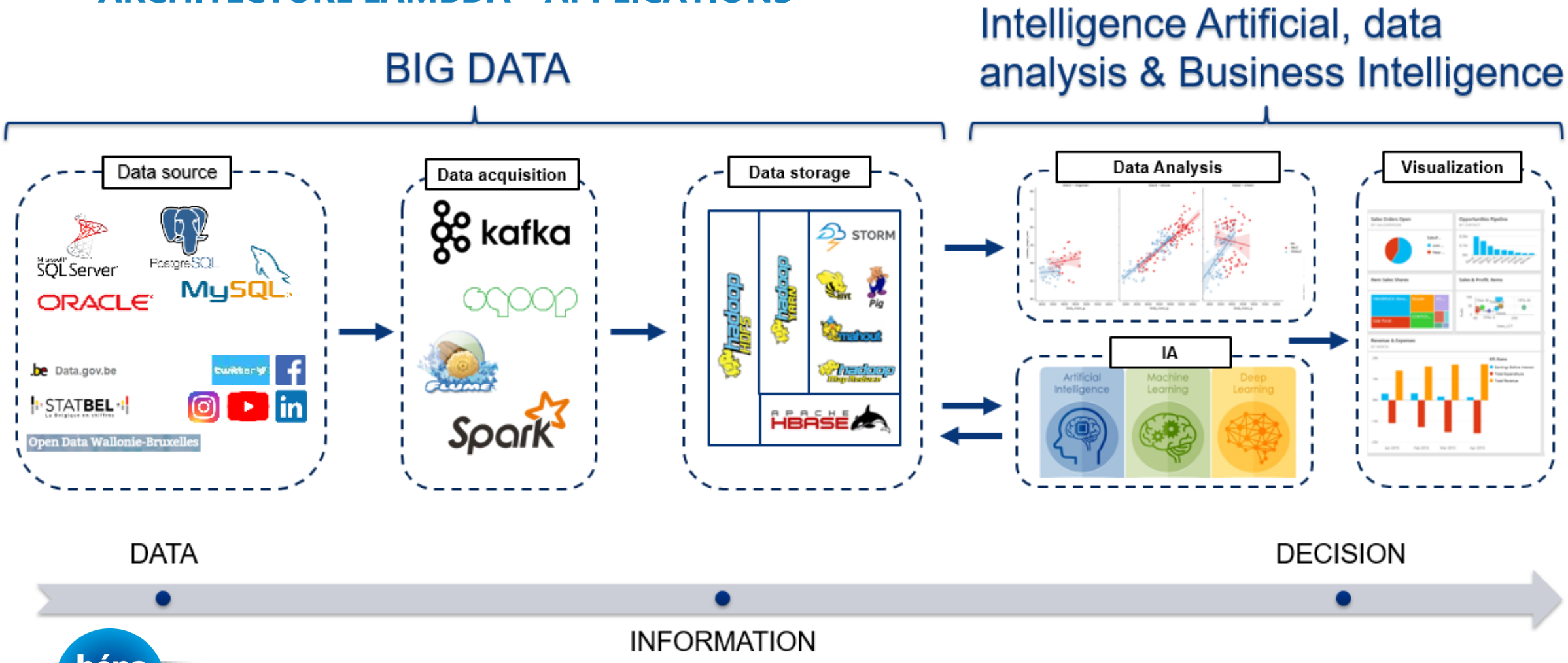
ARCHITECTURE LAMBDA



Source : kai-waehner.de/blog/2021/09/23/real-time-kappa-architecture-mainstream-replacing-batch-lambda/

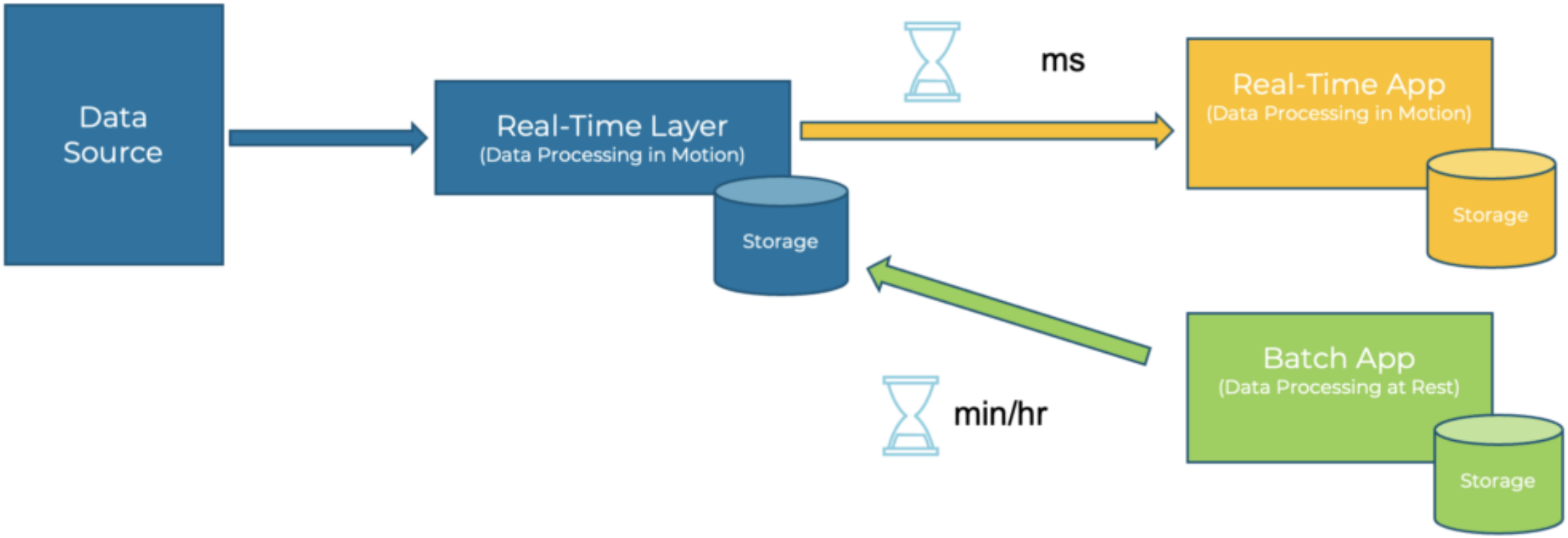
ARCHITECTURES BIG DATA

ARCHITECTURE LAMBDA - APPLICATIONS



ARCHITECTURES BIG DATA

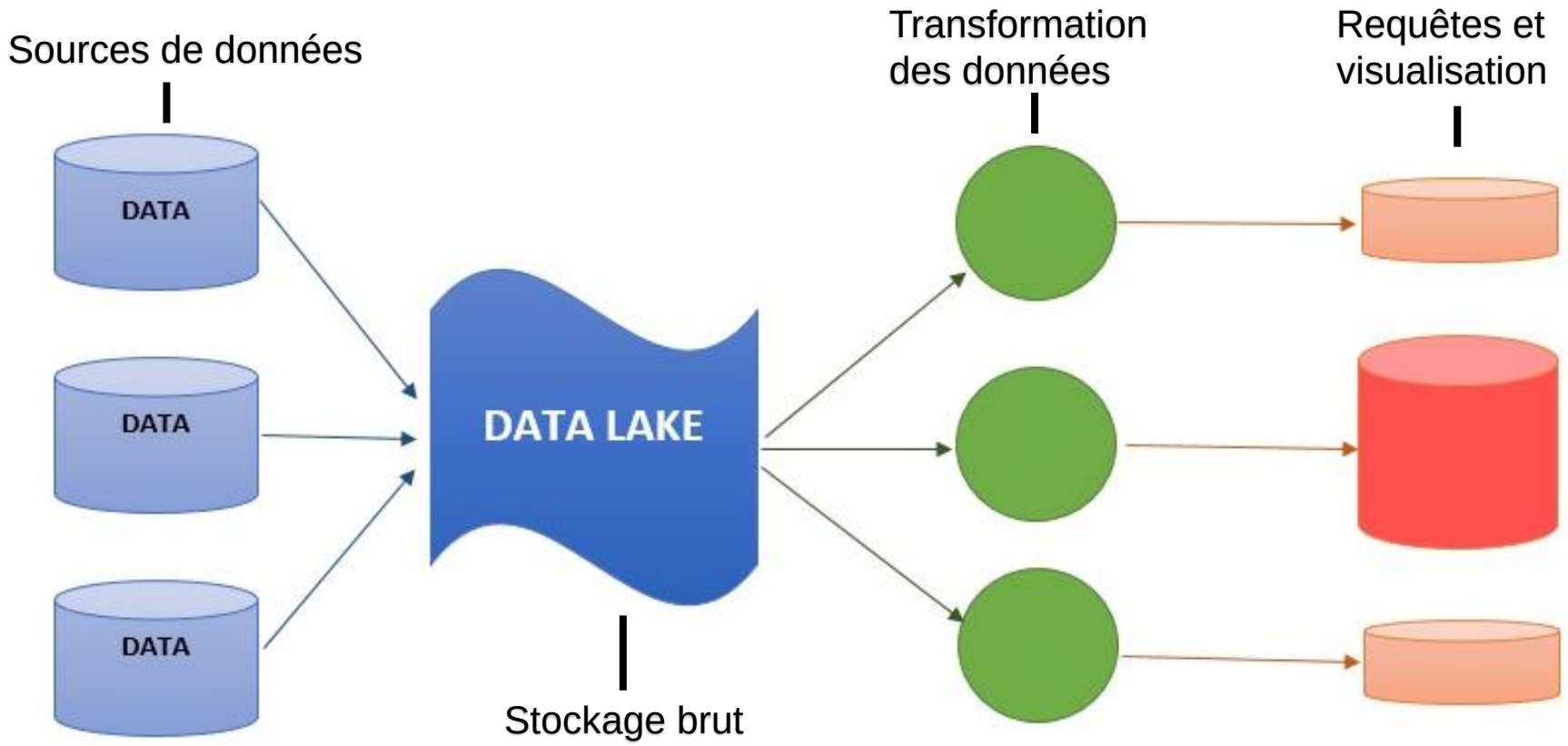
ARCHITECTURE KAPPA

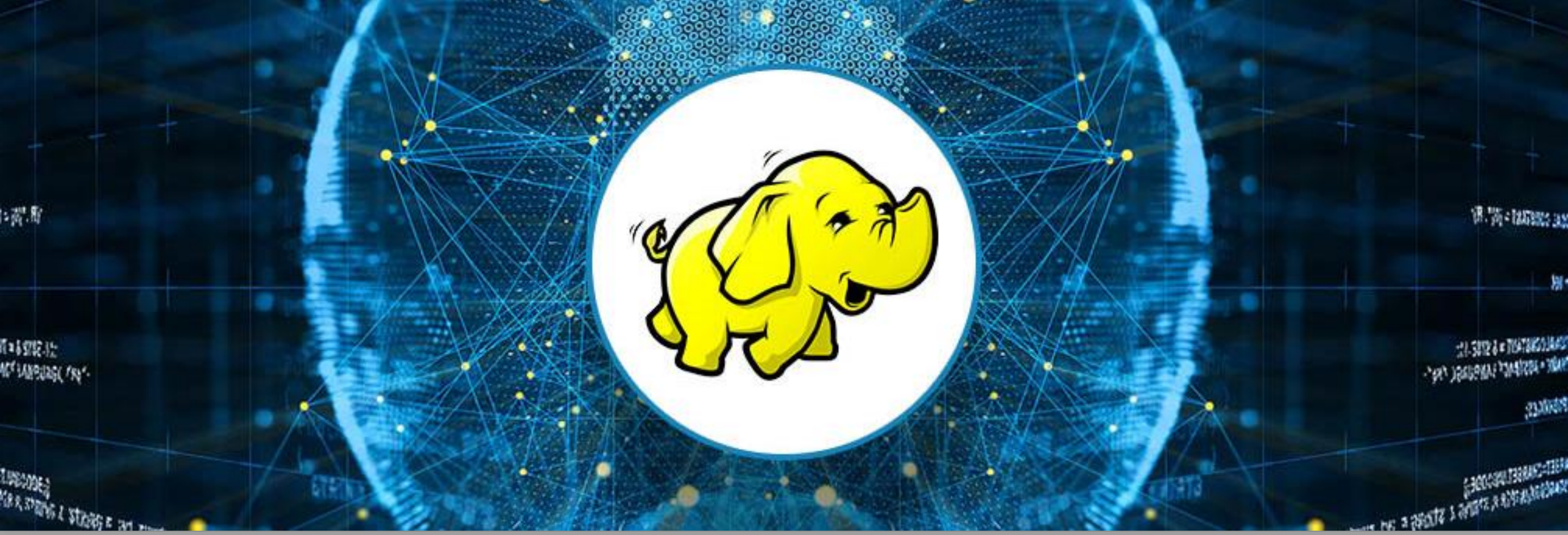


Source : kai-waehner.de/blog/2021/09/23/real-time-kappa-architecture-mainstream-replacing-batch-lambda/

ARCHITECTURES BIG DATA

ARCHITECTURE DATALAKE





ECOSYSTÈME HADOOP

ECOSYSTÈME HADOOP

DÉFINITION

- Définition : Hadoop est un framework open source java géré par la fondation Apache conçu pour réaliser des traitements sur des grosses volumétries de données.

- Composition :

  - Bibliothèque de classe java et d'outils (Hadoop Common)
  - Système HDFS
  - Une implémentation Map Reduce

- Historique :

  - Google crée Map/Reduce en 2004 (Jeffrey Dean et Sanjay Ghemahat) et GoogleFS
  - Doug Cutting et Michael J. Cafarella créent Hadoop en 2005 dans le cadre du projet Nutch (Moteur de recherche Open source Yahoo)

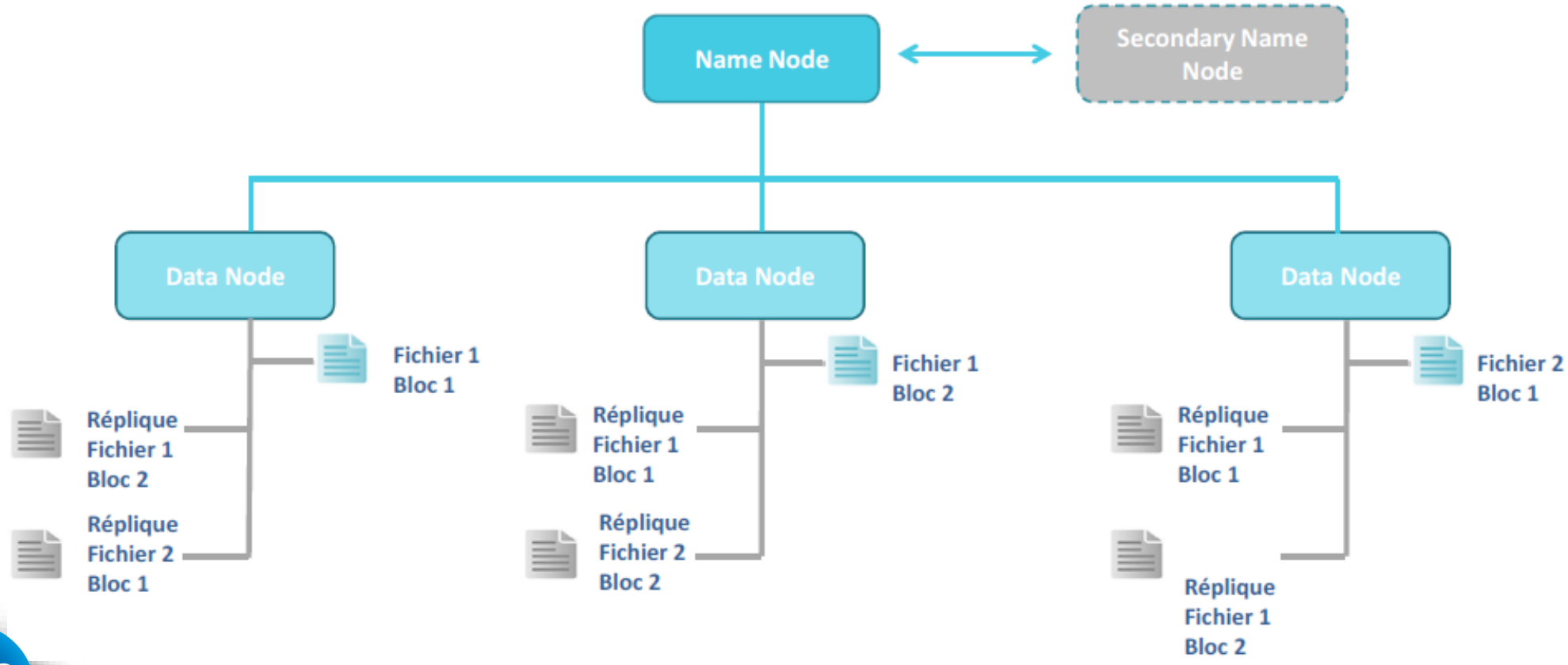
ECOSYSTÈME HADOOP

COMPOSITION

ECOSYSTÈME HADOOP

HADOOP HDFS

- Hadoop Distributed Files System (HDFS) : est un système de stockage de fichiers distribué, extensible et portable développé par Hadoop en java à partir du GoogleFS, basé sur le shared nothing.



ECOSYSTÈME HADOOP

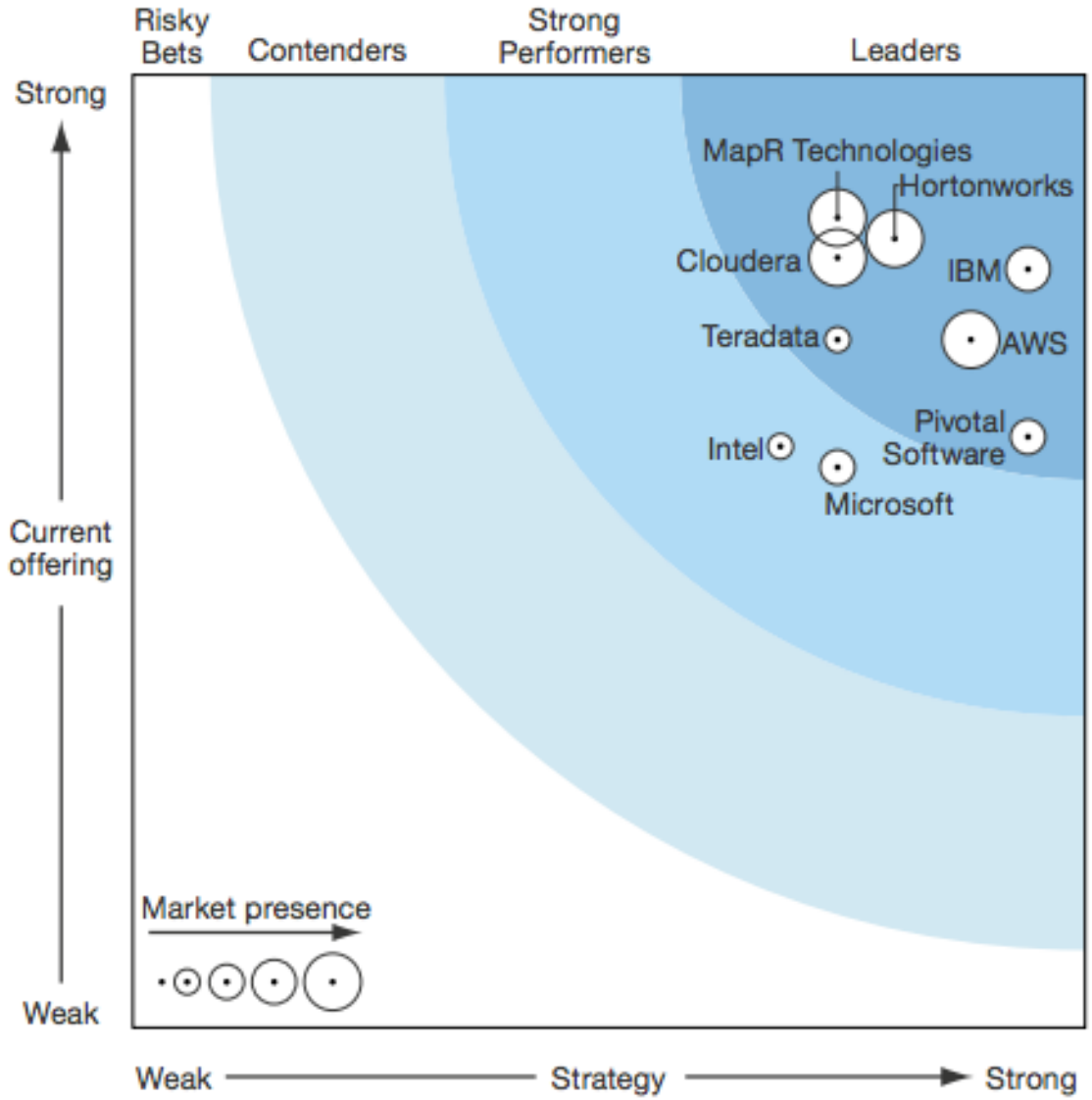
DISTRIBUTIONS

- **Définition**
  - Solutions basées sur Hadoop
  - propres caractéristiques, avantages et intégrations spécifiques,
  - facilitent la mise en place et la gestion de clusters Hadoop
- **Principales solutions :**

cloudera Ask Bigger Questions

Hortonworks

MAPR TECHNOLOGIES



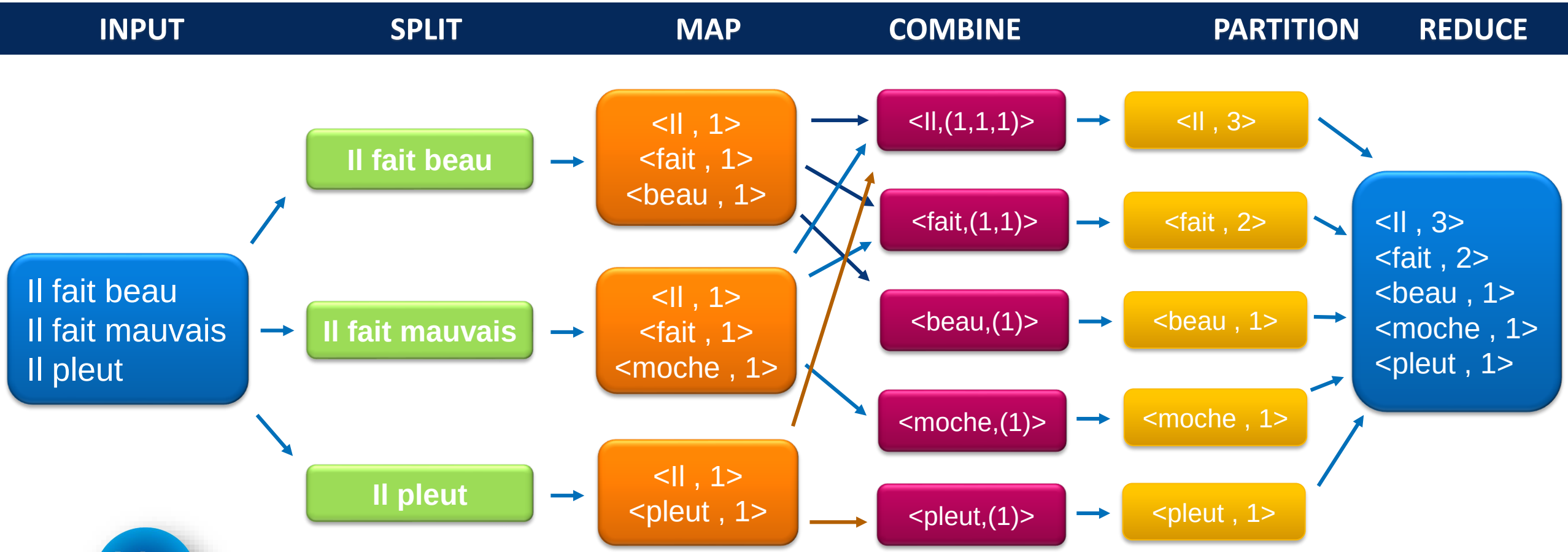
ECOSYSTÈME HADOOP

MAP REDUCE - DÉFINITION

- Modèle de programmation créé par Google
- Traitement sur des clusters
- Calculs distribués (C, C++, Java, Ruby, Pearl, Python)
- Tolérance aux erreurs
- Framework Apache Hadoop
- Fonctionnalités principales : MAP et REDUCE

ECOSYSTÈME HADOOP

MAP REDUCE – PRINCIPE DE FONCTIONNEMENT



LIENS ENTRE IA & BIG DATA

LIENS ENTRE L'IA ET LE BIG DATA

EXEMPLES

- **Alimentation des modèles d'IA :** Plus les données fournies au modèle sont nombreuses et variées, meilleure est la capacité de l'IA à apprendre et à prendre des décisions.
- **Prétraitement des données : L**e nettoyage, la transformation et l'agrégation de données sont des tâches où les techniques d'IA peuvent être appliquées pour faciliter la préparation des données.
- **Analyse de données avancée :** L'IA permet d'identifier des modèles, des tendances et des relations qui seraient difficiles à détecter par des méthodes traditionnelles dans des grands volumes de données.
- **Personnalisation et recommandation :** L'IA permet de créer des systèmes de recommandation personnalisés en analysant les données sur le comportement de l'utilisateur.
- **Traitement du langage naturel :** Les techniques d'IA, telles que le traitement automatique du langage naturel (NLP), sont utilisées pour extraire des informations à partir de textes non structurés, comme des articles de presse, des médias sociaux et des rapports.