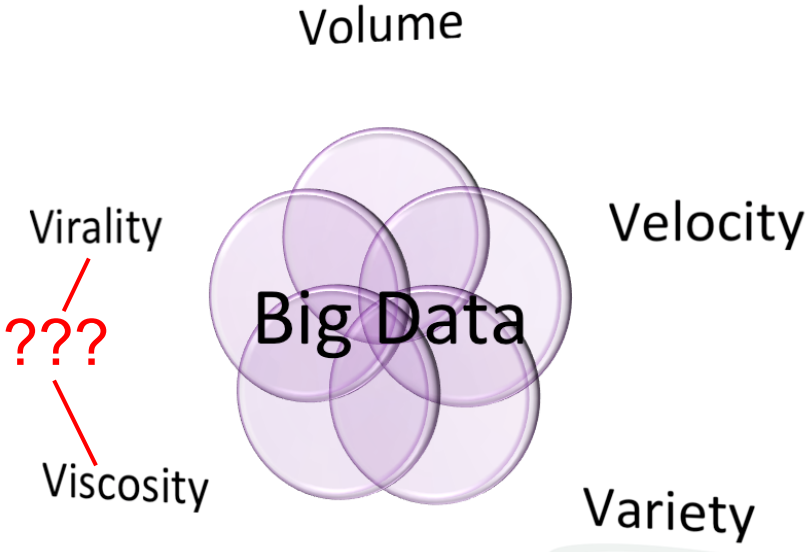
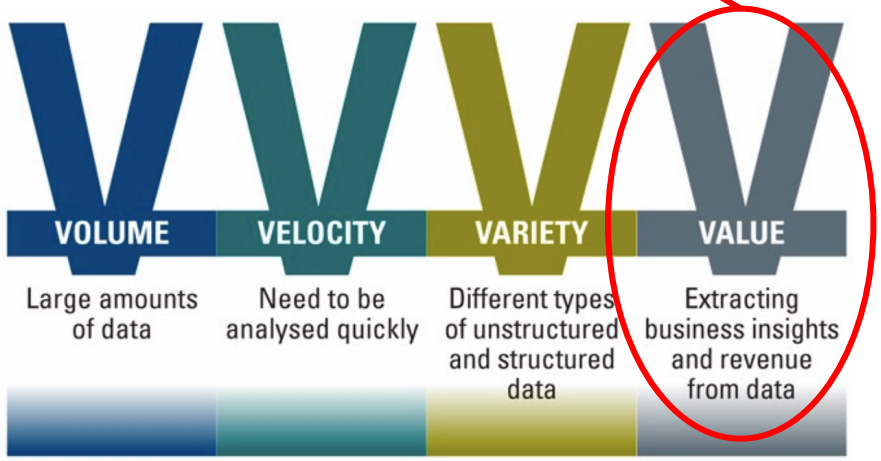
Benoit Da Mota

Dans le pire des cas, vous pouvez toujours dire que

“vos données ont de la valeur”





Quelques grands concepts

MapReduce est un patron d'architecture de développement informatique, inventé par Google, dans lequel sont effectués des calculs parallèles, et souvent distribués, de données potentiellement très volumineuses (...).

Les termes « *map* » et « *reduce* », et les concepts sous-jacents, sont empruntés aux langages de programmation fonctionnelle utilisés pour leur construction (map et réduction de la programmation fonctionnelle et des langages de programmation tableau).

NoSQL (*Not only SQL* en anglais) désigne une catégorie de systèmes de gestion de base de données (SGBD) qui n'est plus fondée sur l'architecture classique des bases relationnelles. L'unité logique n'y est plus la table, et les données ne sont en général pas manipulées avec SQL. (...) Il renonce aux fonctionnalités classiques des SGBD relationnels au profit de la simplicité. (...) Un modèle typique en NoSQL est le système clé-valeur, avec une base de données pouvant se résumer topologiquement à un simple tableau associatif unidimensionnel avec des millions — voire des milliards — d'entrées.

Un **système de fichiers distribué** est un système de fichiers qui permet le partage de fichiers à plusieurs clients au travers du réseau informatique. Contrairement à un système de fichiers local, le client n'a pas accès au système de stockage sous-jacent, et interagit avec le système de fichiers via un protocole adéquat. Les informations sont stockées sur plusieurs noeuds et souvent de façon redondante (ex. LUSTRE, HDFS, etc.).

(*Source : Wikipédia*)

Quelques grands concepts

MapReduce est un patron d'architecture de développement informatique, inventé par Google, dans lequel sont effectués des calculs parallèles, et souvent distribués, de données potentiellement très volumineuses (...).

Les termes « *map* » et « *reduce* », et les concepts sous-jacents, sont empruntés aux langages de programmation fonctionnelle utilisés pour leur construction (map et réduction de la programmation fonctionnelle et des langages de programmation tableau).

NoSQL (*Not only SQL* en anglais) désigne une catégorie de systèmes de gestion de base de données (SGBD) qui n'est plus fondée sur l'architecture classique des bases relationnelles. L'unité logique n'y est plus la table, et les données ne sont en général pas manipulées avec SQL. (...) Il renonce aux fonctionnalités classiques des SGBD relationnels au profit de la simplicité. (...) Un modèle typique en NoSQL est le système clé-valeur, avec une base de données pouvant se résumer topologiquement à un simple tableau associatif unidimensionnel avec des millions — voire des milliards — d'entrées.

Un **système de fichiers distribué** est un système de fichiers qui permet le partage de fichiers à plusieurs clients au travers du réseau informatique. Contrairement à un système de fichiers local, le client n'a pas accès au système de stockage sous-jacent, et interagit avec le système de fichiers via un protocole adéquat. Les informations sont stockées sur plusieurs noeuds et souvent de façon redondante (ex. LUSTRE, HDFS, etc.).

(*Source : Wikipédia*)

Quelques grands concepts

MapReduce est un patron d'architecture de développement informatique, inventé par Google, dans lequel sont effectués des calculs parallèles, et souvent distribués, de données potentiellement très volumineuses (...).

Les termes « *map* » et « *reduce* », et les concepts sous-jacents, sont empruntés aux langages de programmation fonctionnelle utilisés pour leur construction (map et réduction de la programmation fonctionnelle et des langages de programmation tableau).

NoSQL (*Not only SQL* en anglais) désigne une catégorie de systèmes de gestion de base de données (SGBD) qui n'est plus fondée sur l'architecture classique des bases relationnelles. L'unité logique n'y est plus la table, et les données ne sont en général pas manipulées avec SQL. (...) Il renonce aux fonctionnalités classiques des SGBD relationnels au profit de la simplicité. (...) Un modèle typique en NoSQL est le système clé-valeur, avec une base de données pouvant se résumer topologiquement à un simple tableau associatif unidimensionnel avec des millions — voire des milliards — d'entrées.

Un **système de fichiers distribué** est un système de fichiers qui permet le partage de fichiers à plusieurs clients au travers du réseau informatique. Contrairement à un système de fichiers local, le client n'a pas accès au système de stockage sous-jacent, et interagit avec le système de fichiers via un protocole adéquat. Les informations sont stockées sur plusieurs noeuds et souvent de façon redondante (ex. LUSTRE, HDFS, etc.).

(*Source : Wikipédia*)

Le “~~Big Data~~” ça fonctionne ?

hadoop

Hortonworks · hadapt · cloudera · EXASOL (Excellent data expert) · INFOBRIGHT · Pivotal · calpont · kognitio · MAPR TECHNOLOGIES · Actian · VERTICA · NETEZZA an IBM Company · HSTREAMING · PRECOG · Google bigquery · DataGravity

Gros utilisateurs: Yahoo, Facebook, Google, Amazon, Microsoft ...

How a little open source project came to dominate big data

by Katherine Noyes @noyesk JUNE 30, 2014, 3:49 PM EDT

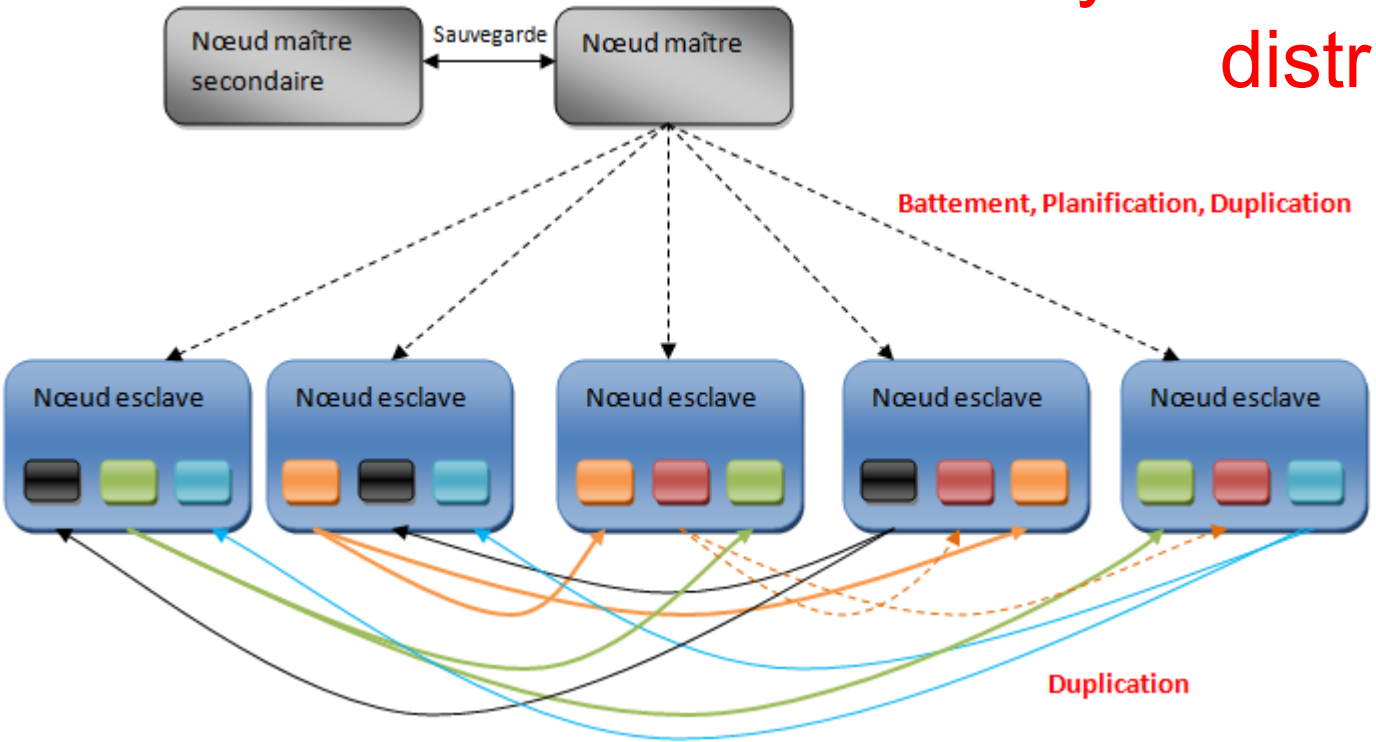
It began as a nagging technical problem that needed solving. Now, it's driving a market that's expected to be worth $50.2 billion by 2020.

Hadoop allows some of the world's largest companies to store and process datasets on clusters of commodity hardware.

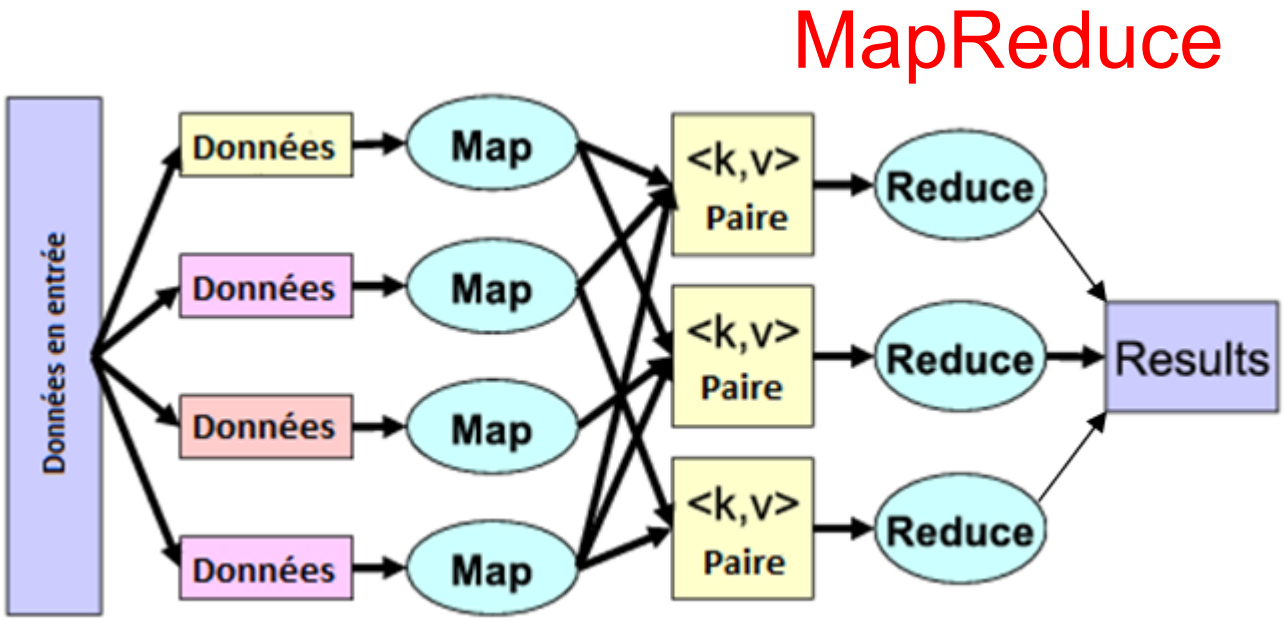
Jetta Productions—Getty Images

ça fonctionne comment? (1/2)

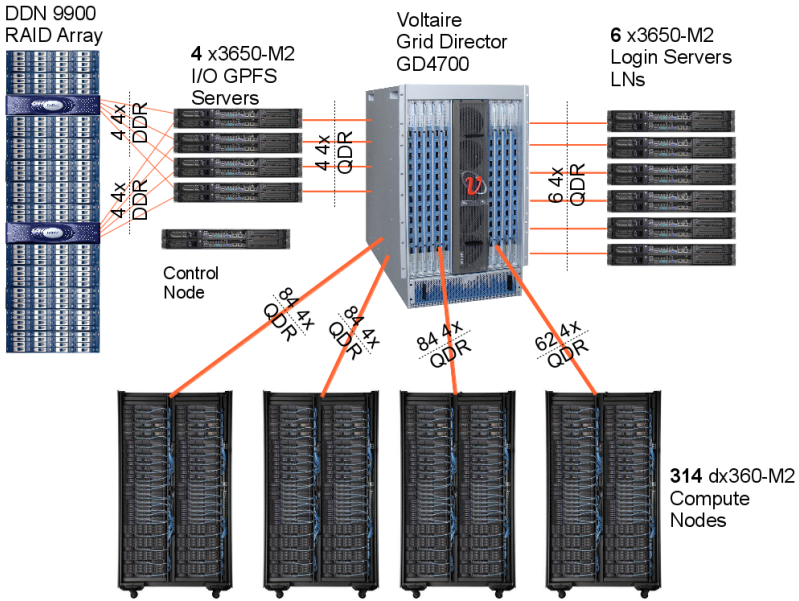
Système de fichiers distribué



ça fonctionne comment? (2/2)



Et matériellement ?



?

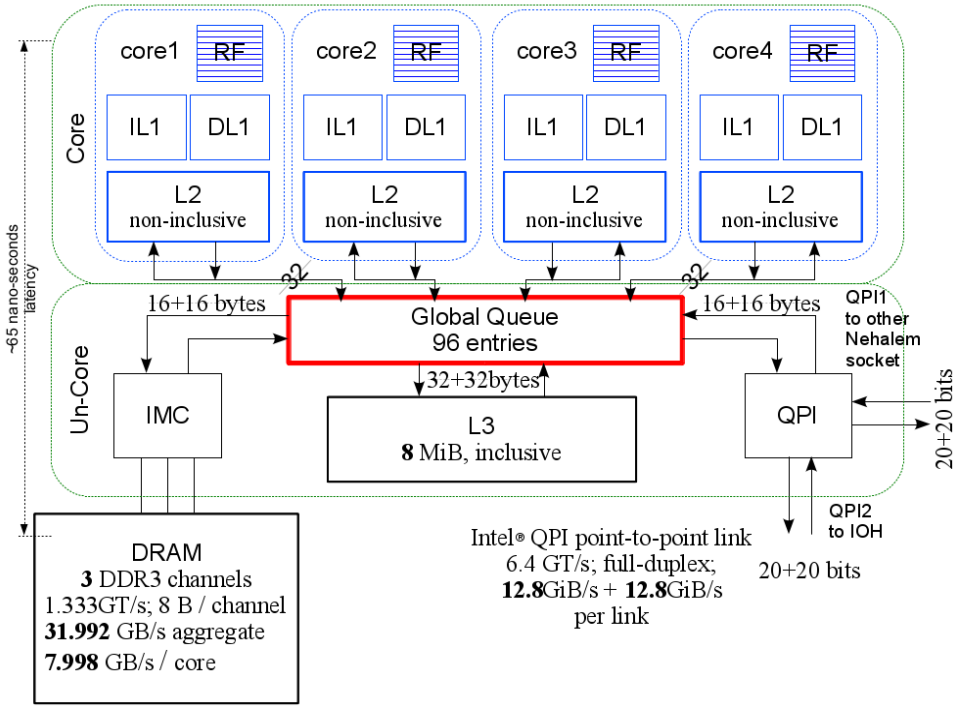
les vrais problèmes “informatiques”

- Compute bounded
  limité par la puissance de calcul
- Memory bounded
  limité par la bande passante mémoire
- I/O bounded
  limité en accès aux données

=> Selon le ou les problèmes à traiter il faut bâtir une solution hardware, middleware, software et méthodologique* adaptée.
=> On parle de co-design.

* statistique/ML

version simplifiée d’un processeur moderne :



Est-ce que cela a un sens ?

- Doit-on uniquement parler de technologies ?
  - Hardware : quelles machines, stockages et réseaux
  - Middleware : Cloud, MapReduce, Hadoop, etc.

- Peut-on poser n'importe quelles questions à des grandes données ?
  - De quelles grandes données parlons nous ? Grand nombre de descripteurs (p>>n) ou grand nombre d'échantillons (n>>p) ?
  - Mon modèle est-il adapté à des mises à jour (inline) ?
  - Doit-on structurer ou valoriser les données/méta-données (SGBD)?

- Que veulent entendre les financeurs ?