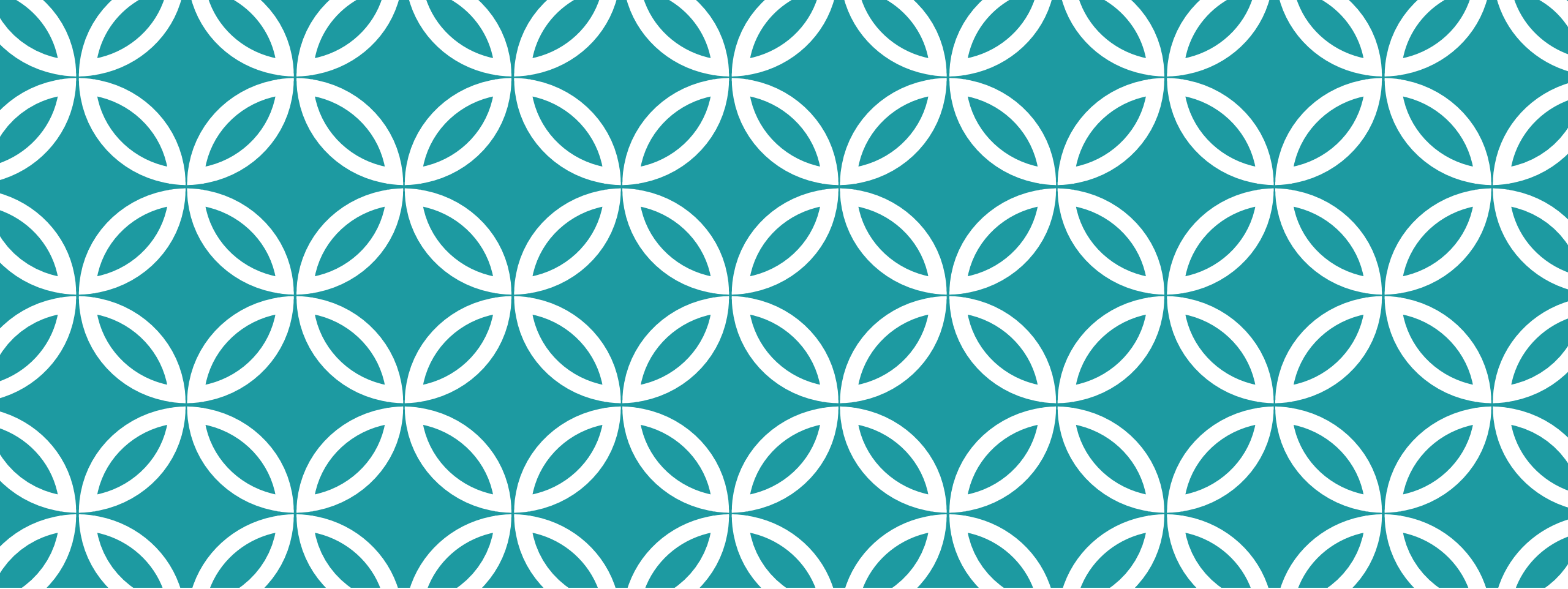


Traitement de données massives Ecosystème Hadoop

David Auber
Alexandre Perrot



Big Data ?

Révolution des données
Révolution technologique
Révolution des usages

Révolution des données



La taille des données numérisées générées ne cesse d'augmenter.

- Texte
- Image/Son
- Vidéo
- Fréquence cardiaque
- Position géographique
- Rythme de sommeil
- etc...



Révolution des données



Quizz des données générées tous les jours.

- **SMS par jours**
- Tweets par jours
- Recherche Google
- Emails échangés



Révolution des données



Quizz des données générées tous les jours.

- **SMS par jours**
- Tweets par jours
- Recherche Google
- Emails échangés

540 Millions
540Go/j => 200To/a

Révolution des données



Quizz des données générées tous les jours.

- SMS par jours
- **Tweets par jours**
- Recherche Google
- Emails échangés



Révolution des données



Quizz des données générées tous les jours.

- SMS par jours
- **Tweets par jours**
- Recherche Google
- Emails échangés

500 Millions

Révolution des données



Quizz des données générées tous les jours.

- SMS par jours
- Tweets par jours
- **Recherche Google**
- Emails échangés



Révolution des données



Quizz des données générées tous les jours.

- SMS par jours
- Tweets par jours
- Recherche Google
- Emails échangés

4,5 Milliards

Révolution des données



Quizz des données générées tous les jours.

- SMS par jours
- Tweets par jours
- Recherche Google
- **Emails échangés**



Révolution des données



Quizz des données générées tous les jours.

- SMS par jours
- Tweets par jours
- Recherche Google
- **Emails échangés**

145 Milliards !!!

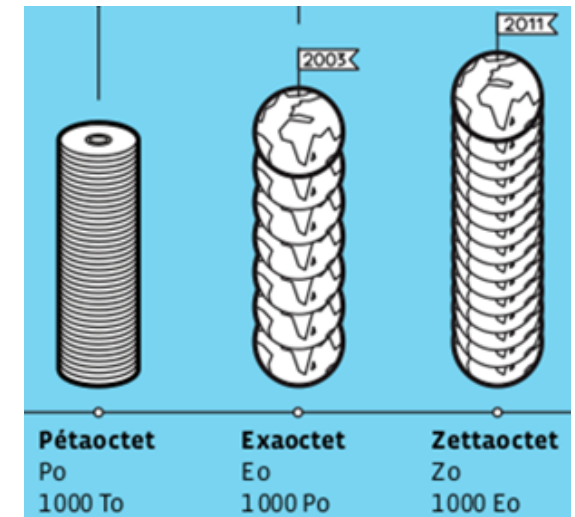
Révolution des données



Ordre de grandeur

Entre 2003 et 2011 on a enregistrées 1000 fois plus de données que depuis le début de l'humanité.

- 1 Po = La hauteur de la tour Montparnasse en DVD
- Exaoctets = Donnée numérisées depuis 2003
- Zettaoctets = donnée stockées en 2011

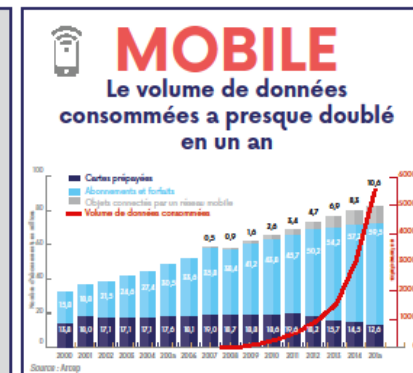
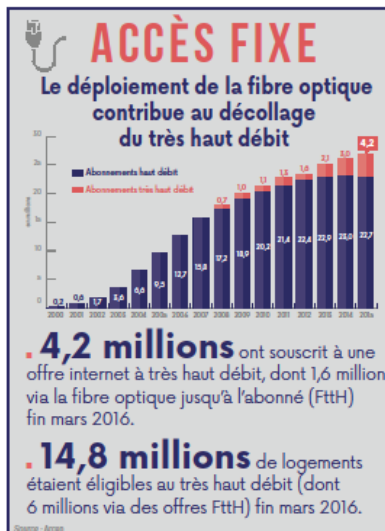


Révolution des données

RÉGULER POUR CONNECTER



CHIFFRES CLÉS 2015/2016



7,8 MILLIARDS D'EUROS
C'EST LE MONTANT DES INVESTISSEMENTS DES OPÉRATEURS EN 2015, EN HAUSSE DE 10% EN UN AN
EN INCLUANT LES ACHATS DE FRÉQUENCES, CE MONTANT S'ÉLÈVE À
10,6 MILLIARDS D'EUROS

Source : Arcep

58 MIN SONT CONSACRÉES PAR JOUR À SURFER SUR INTERNET VIA UN TÉLÉPHONE MOBILE EN FRANCE

LE TRAFIC MONDIAL DE DONNÉES MOBILES EN 2020 REPRÉSENTERA **4 FOIS** LE TRAFIC INTERNET EN 2005

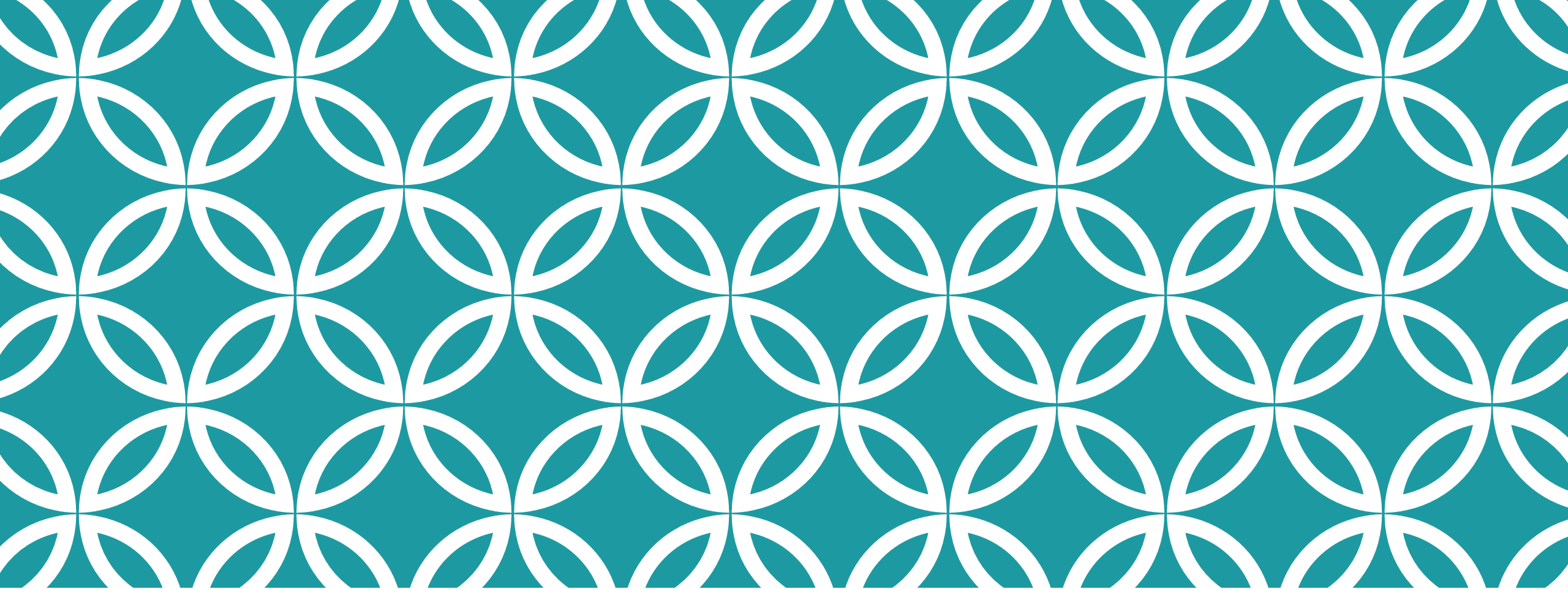
Source : Visual networking index - Cisco

DES PME CONNECTÉES
10,1 MILLIARDS D'EUROS HT
C'EST LE CHIFFRE D'AFFAIRES DES VENTES DES OPÉRATEURS AUX ENTREPRISES (30% DU MARCHÉ DE DÉTAIL)

Source : Arcep (2014)

9,1% DES 15-24 ANS se connectent sur INTERNET uniquement via UN SMARTPHONE

Source : Médiamétrie - Audience Internet global en France - Janvier 2016



Big Data ?

Révolution des données
Révolution technologique
Révolution des usages

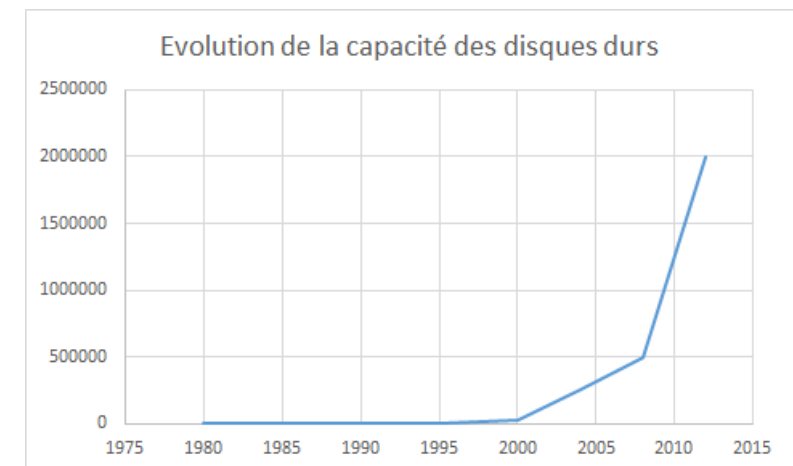
Révolution technologique



Les avancées technologiques sont à l'origine du phénomène Big Data.

- **Progression des disques**
- Progression des processeur
- Progression des Mémoire
- Progression des réseaux

1980	26Mo
1995	3Go
2000	30Go
2004	250Go
2012	2To



Révolution technologique



Les avancées technologiques sont à l'origine du phénomène Big Data.

- Progression des disques
- Progression des processeur
- Progression des mémoires
- **Progression des réseaux**

1983	10MB/s
1994	100MB/s
1996	1GB/S
2002	10GB/s
2010	100GB/s

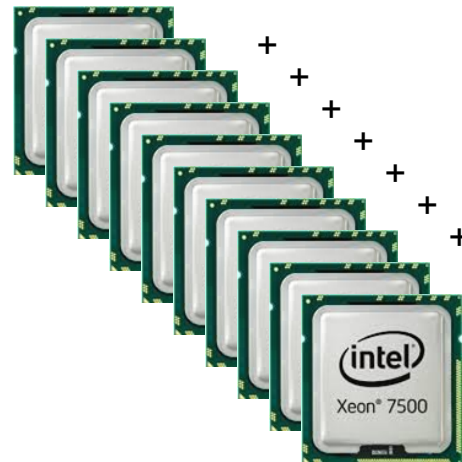


Révolution technologique



L'interconnexion des machines permet l'avènement de la scalabilité horizontale.

- Interconnexion des disques durs
- Interconnexion des mémoires
- Interconnexion des processeurs



= Machine de puissance et stockage illimitée.

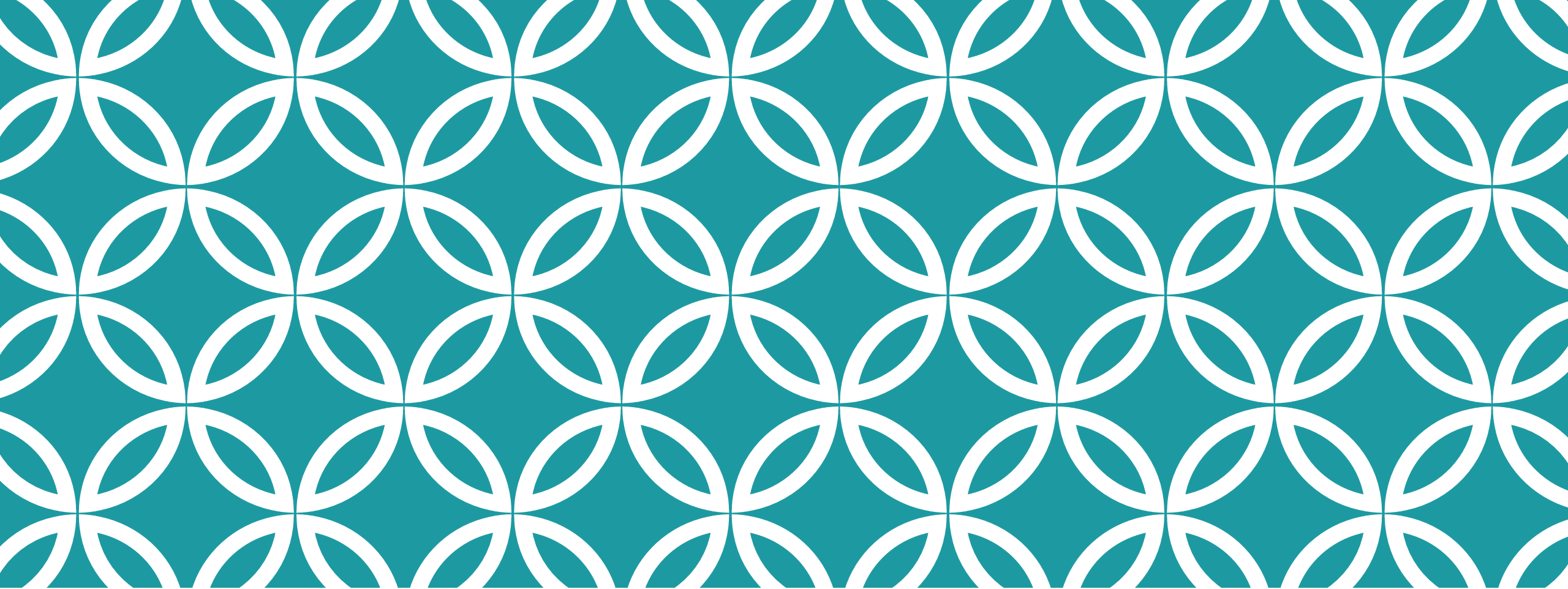
Révolution technologique



La scalabilité horizontale est d'ores et déjà utilisée pour créer des "machines" d'une capacité inimaginable il y a quelques années.

- NSA Data Center Utah 2014
- 100.000 m² surface
- 10.000 m² serveurs
- 2 Millards\$ de matériel
- 10.000 Racks de serveurs
- 12 Exabyte
- 1 an un de communication US = 272 PetaByte
 - 20 ans d'enregistrement possible





Big Data ?

Révolution des données
Révolution technologique
Révolution des usages

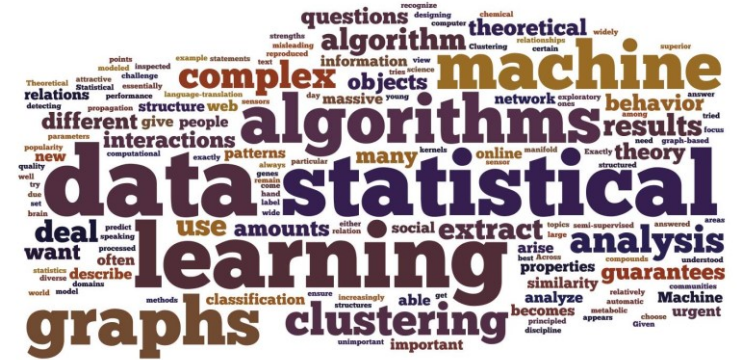
La puissance de calcul et les capacités de stockage permettent maintenant de traiter les problèmes dans leur globalité.

Figure 1 consists of two plots. The left plot is a histogram of variable x (black bars) with a red normal distribution curve overlaid. The right plot is a scatter plot of y versus x . It shows a green shaded elliptical region representing the joint distribution. A red line represents the regression of y on x , and a blue line represents the regression of x on y . The angle between these two lines is labeled ϕ . The variance of x is labeled $\text{Var}(x)$ and the variance of y is labeled $\text{Var}(y)$. The covariance is labeled $\text{Cov}(x; y)$. The regression equations are given as:

$$y = a_y + b_y x = \bar{y} + b_y (x - \bar{x})$$

$$x = a_x + b_x y = \bar{x} + b_x (y - \bar{y})$$

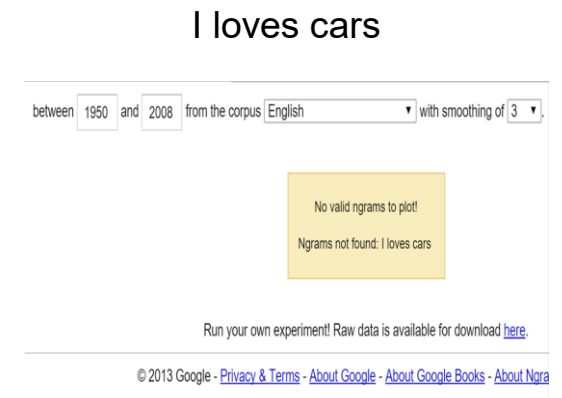
Révolution des usages



Des données pour apprendre aux machines

Les données permettent aux machines de résoudre des problèmes que nous ne pouvons pas résoudre jusque là.

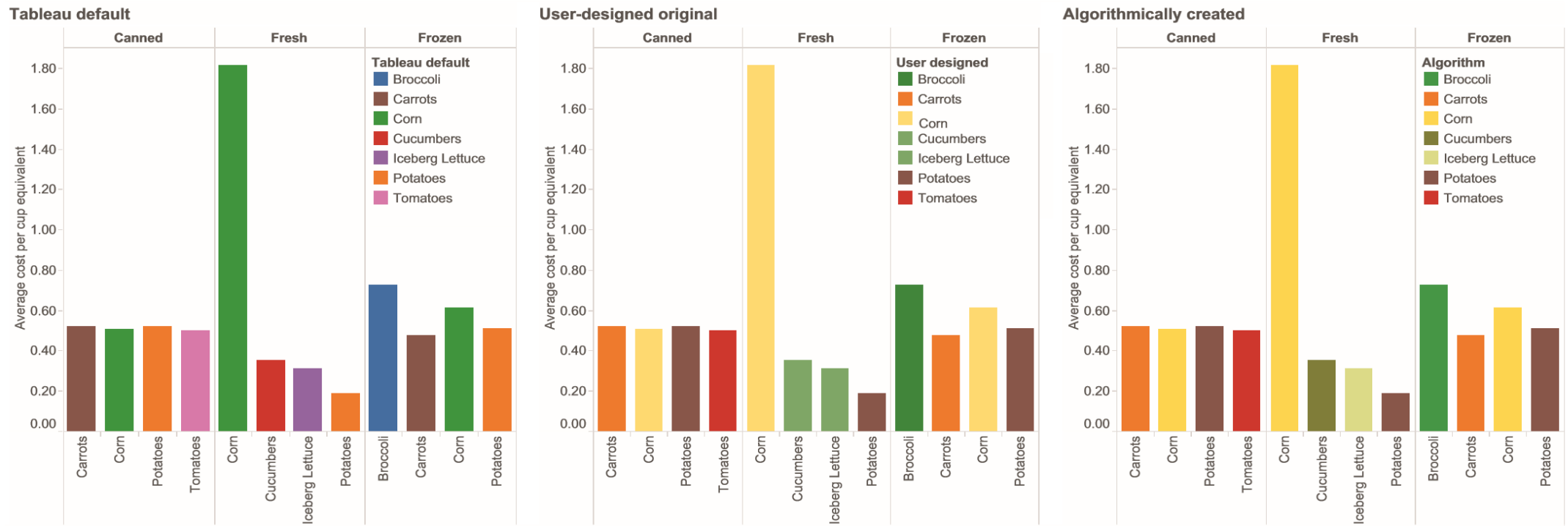
Par exemple: Les N-GRAMs, permettent de compter combien de fois des mots apparaissent ensemble. En utilisant ces statistiques la machine est capable de détecter les erreurs.



<https://books.google.com/ngrams/>

A word cloud visualization featuring various terms associated with machine learning and data science. The words are arranged in a dense, overlapping manner, with colors ranging from dark blue/purple at the top to light yellow/green at the bottom. The most prominent words include "machine", "algorithms", "data", "statistical", "learning", "graphs", "complex", "questions", "theoretical", "behavior", "results", "theory", "analysis", "guarantees", "social", "extract", "use", "amounts", "classification", "clustering", "objects", "information", "view", "reference", "massive", "network", "sites", "need", "graph-based", "exactly", "structural", "patterns", "particular", "always", "great", "best", "hand", "label", "size", "popularity", "new", "quality", "well", "try", "data", "set", "series", "deal", "product", "speaking", "want", "often", "describe", "domains", "world", "model", "methods", "relationships", "certain", "superior", "answer", "good", "focus", "need", "theory", "understood", "improved", "answered", "areas", "best", "properties", "similarity", "analyze", "becomes", "prioritized", "discipline", "relatively", "automatic", "multiple", "appears", "chosen", "Given", "urgent", "important", "unimportant", "structures", "increasingly", "able", "get", "communication", "communities", "Machine". Other visible words include "points", "impacted", "example", "statements", "strengths", "misleading", "reproduced", "computer", "designing", "chemical", "recognize", "detecting", "attractive", "challenge", "essential", "performance", "propagation", "structure", "web", "parameters", "interactions", "computational", "exactly", "patterns", "particular", "always", "great", "best", "hand", "label", "size", "popularity", "new", "quality", "well", "try", "data", "set", "series", "deal", "product", "speaking", "want", "often", "describe", "domains", "world", "model", "methods", "relationships", "certain", "superior", "answer", "good", "focus", "need", "theory", "understood", "improved", "answered", "areas", "best", "properties", "similarity", "analyze", "becomes", "prioritized", "discipline", "relatively", "automatic", "multiple", "appears", "chosen", "Given", "urgent", "important", "unimportant", "structures", "increasingly", "able", "get", "communication", "communities", "Machine".

Exemple : Affecter automatiquement des couleurs sur un diagramme.
Setlur & Stone IEEE Trans. Vis. Comp. Graphics 2105



Révolution des usages

Test de tous les NGRAMS

mots – couleurs

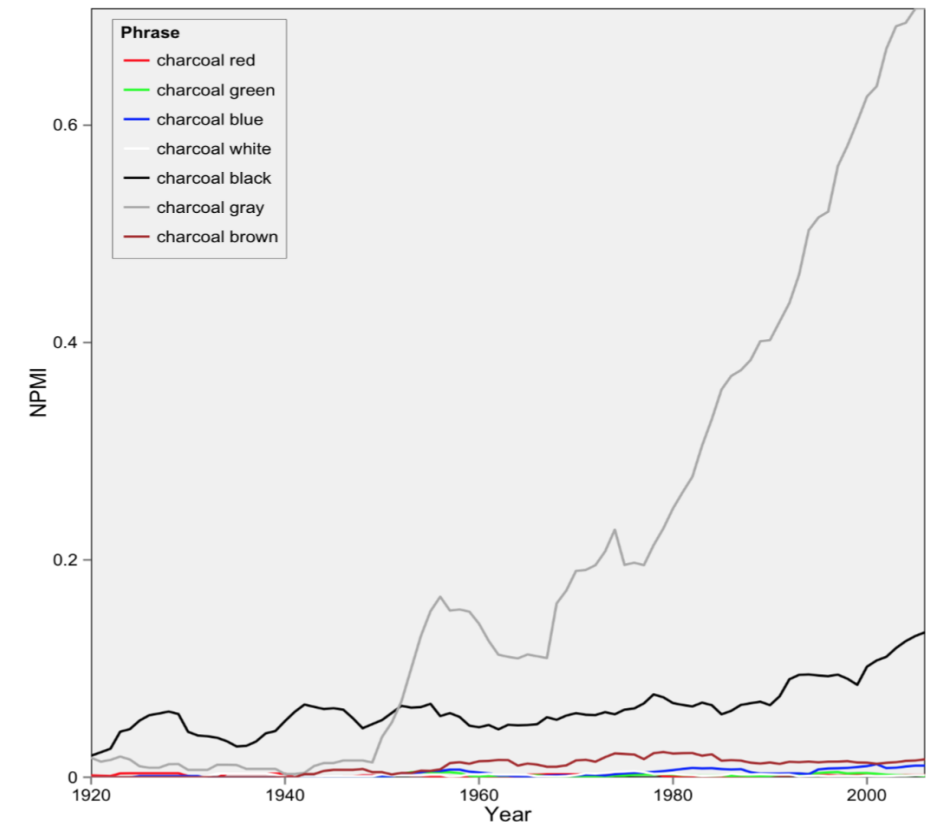
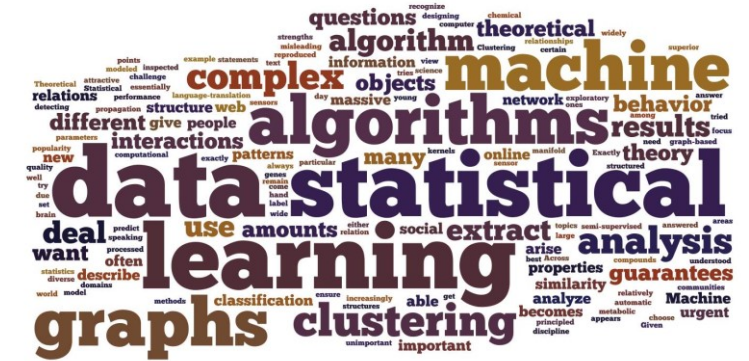
Impossible de recalculer les NGRAMS à la volée.

Besoin de stocker tous les NGRAMS

2 GRAMS N * N,

3 GRAMS N * N * N

10 000 mots -> 64 To (sans index)

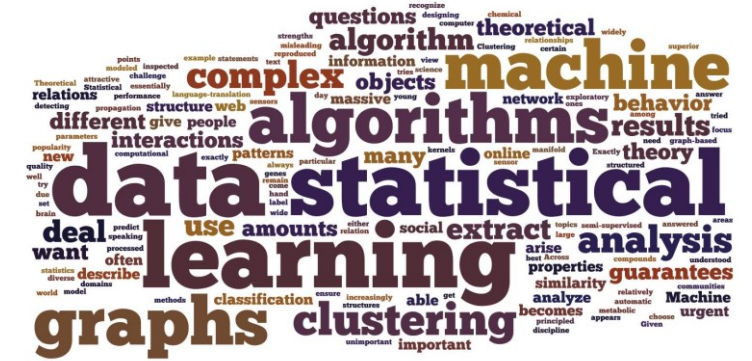








Révolution des usages

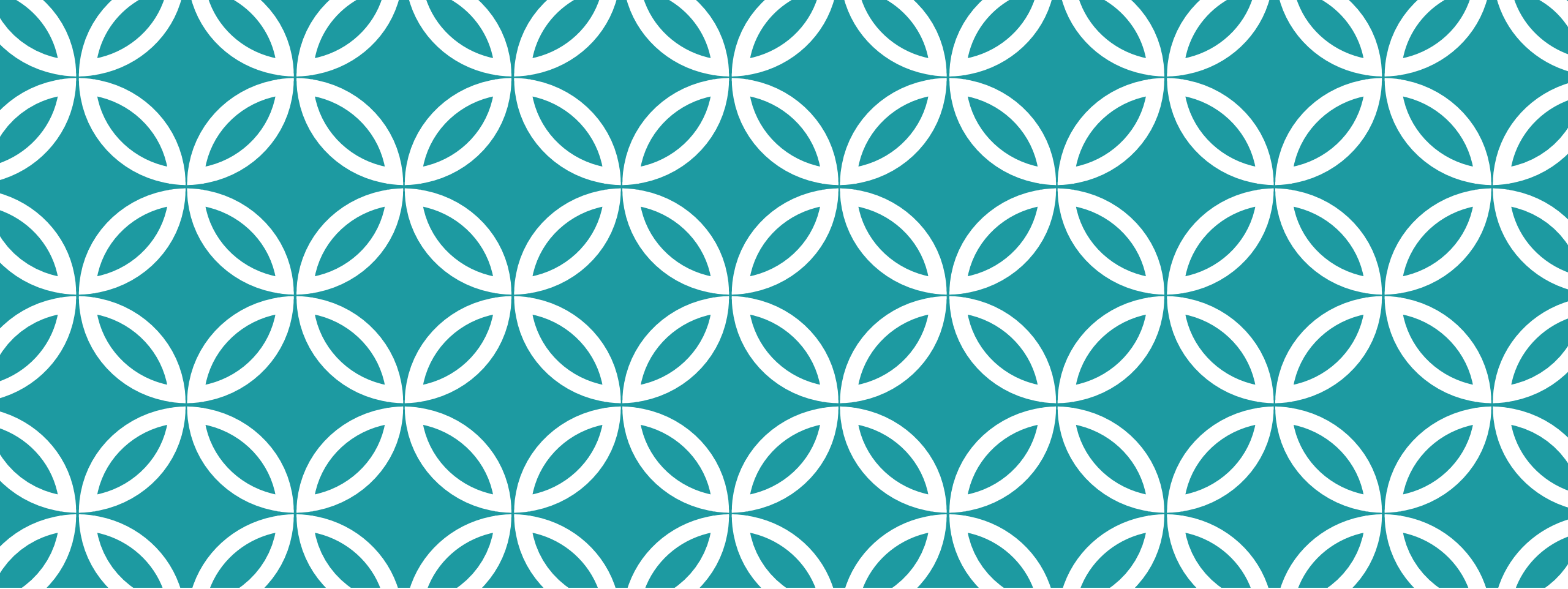
Récupération de toutes les images indexées avec la couleur retenu.

Détermination de la couleur exact...

- Base association sémantique
 - nom-couleur



Input term	Top clustered images	Canonical color
taxi		
lizard		
saffron		



Hadoop

Pourquoi
Histoire
Concepts

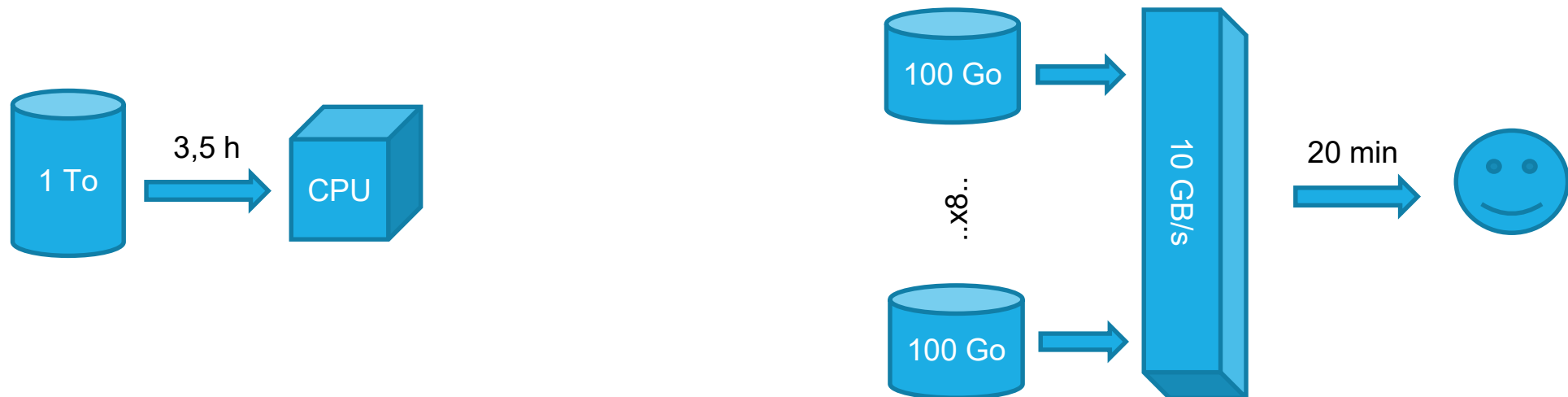
Problématique de stockage

Coût et vitesse d'écriture des disques d'un Téra en 2016

- SSD 250€ 1 To / 500 Mo/s
- HDD 50€ 1 To / 80 Mo/s

Coût du Péta:

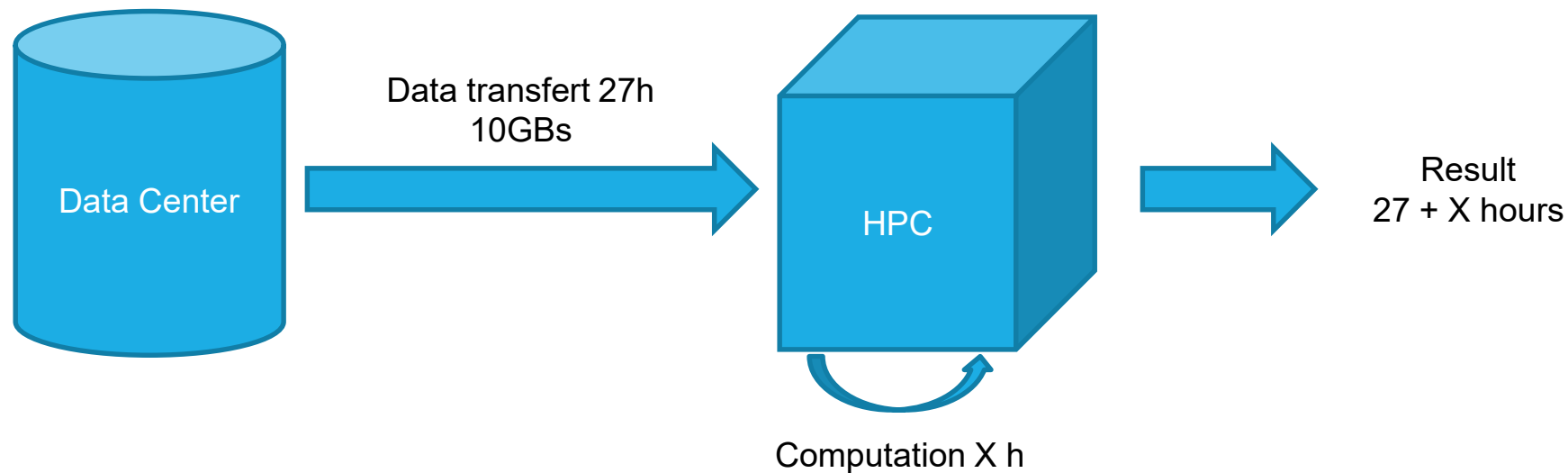
- SSD 250.000 € -> 2.000.000 secondes => 23,1 jours (sur 1 disque)
- HDD 50.000 € -> 12.500.000 secondes => 144 jours => 4,8 mois (sur 1 disque)



Problématique de calcul

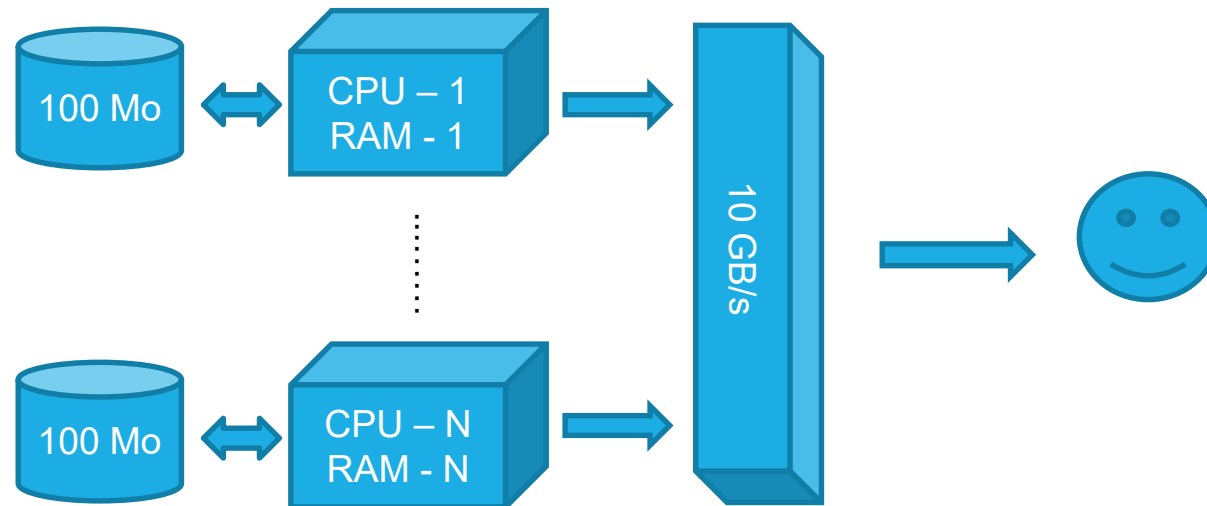
Temps de transfert d'un Péta octets vers un processeur:

- Débit réseaux 10Gb/s
- transfert d'un Péta (si disque à 10Gb/s) => 100.000 s -> 27h



Unification du stockage et du calcul

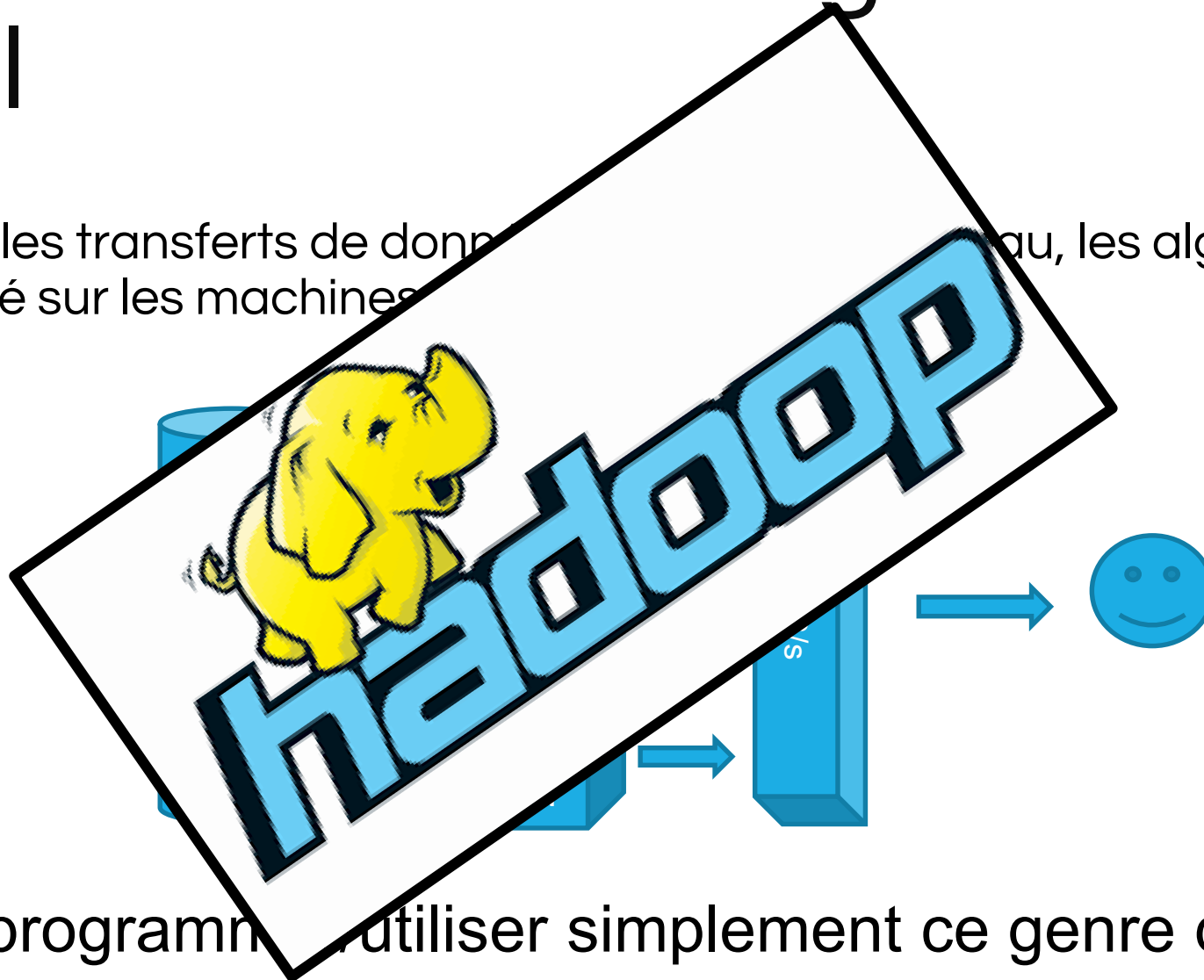
Pour limiter les transferts de données inutiles sur le réseau, les algorithmes sont exécutés sur les machines qui stockent les données.



Comment programmer/utiliser simplement ce genre de machines ?

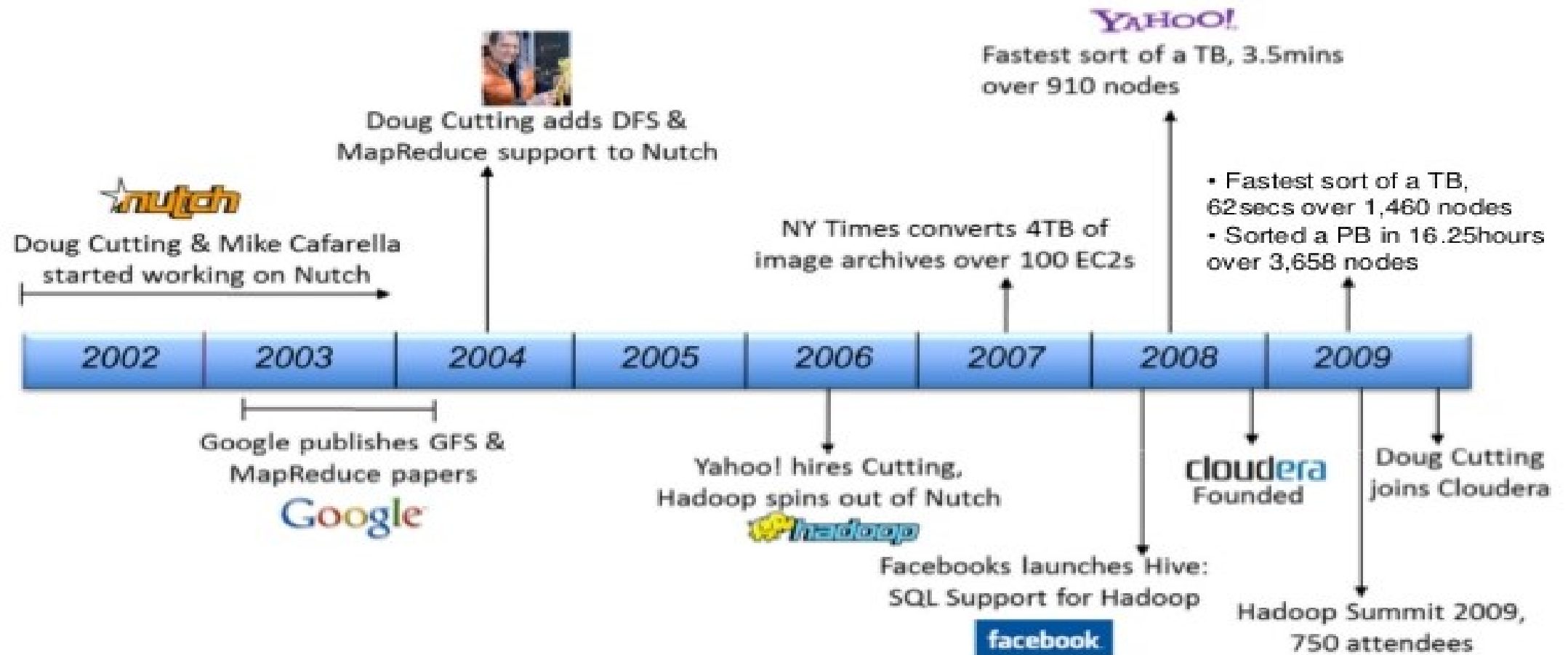
Unification du stockage et du calcul

Pour limiter les transferts de données, les algorithmes sont exécutés sur les machines.

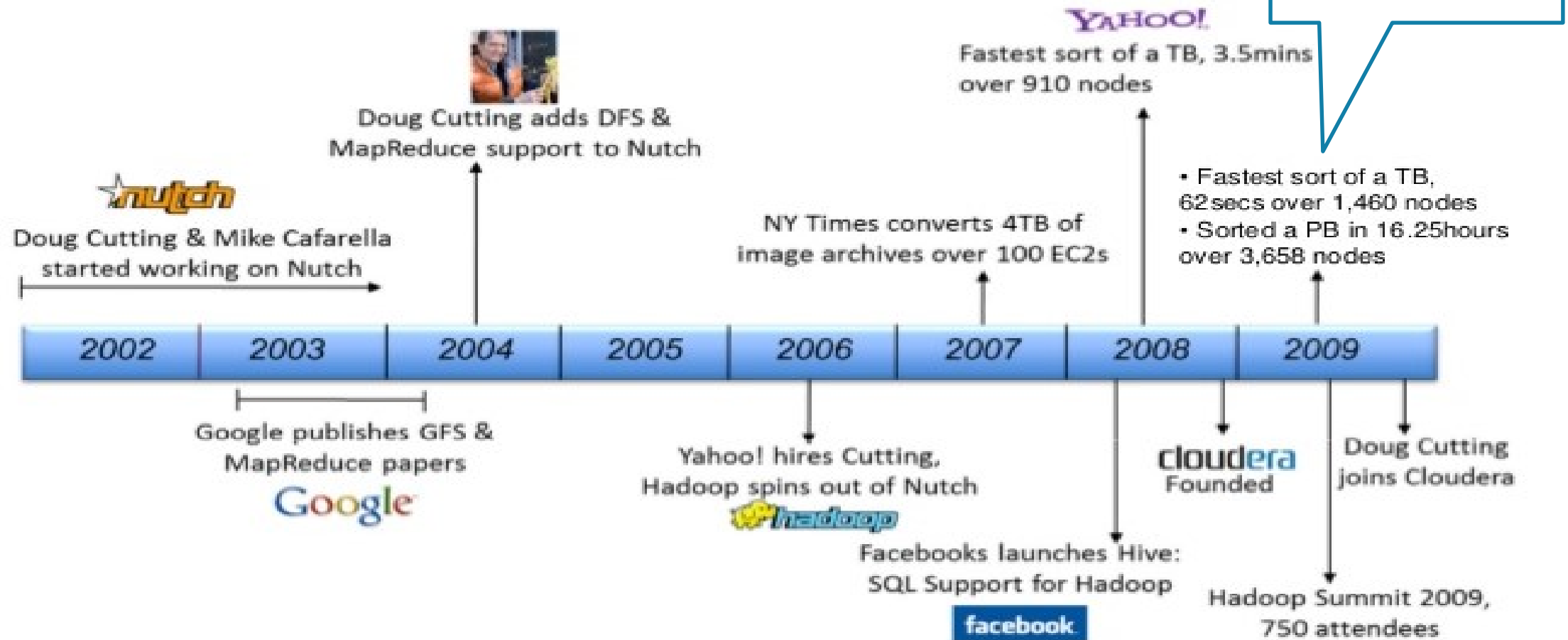


Comment programmer et utiliser simplement ce genre de machines ?

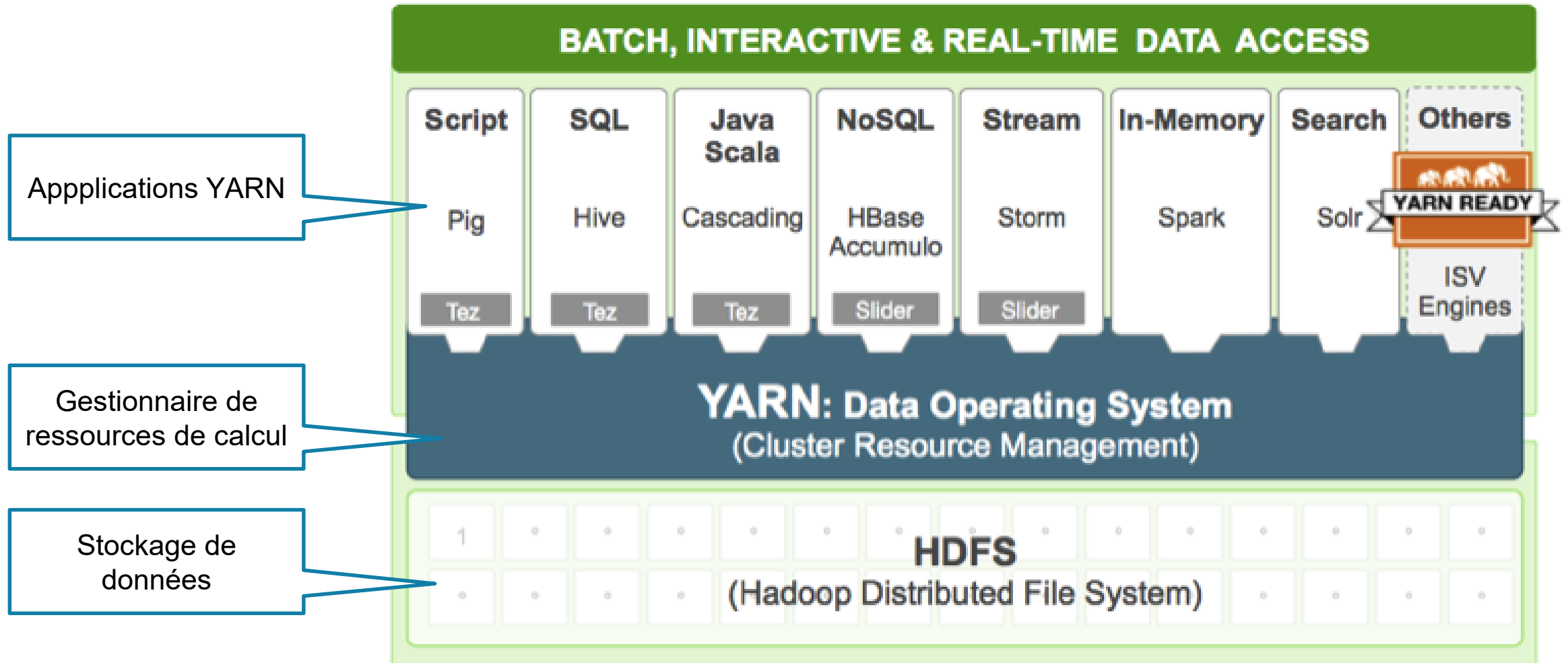
Histoire



Histoire

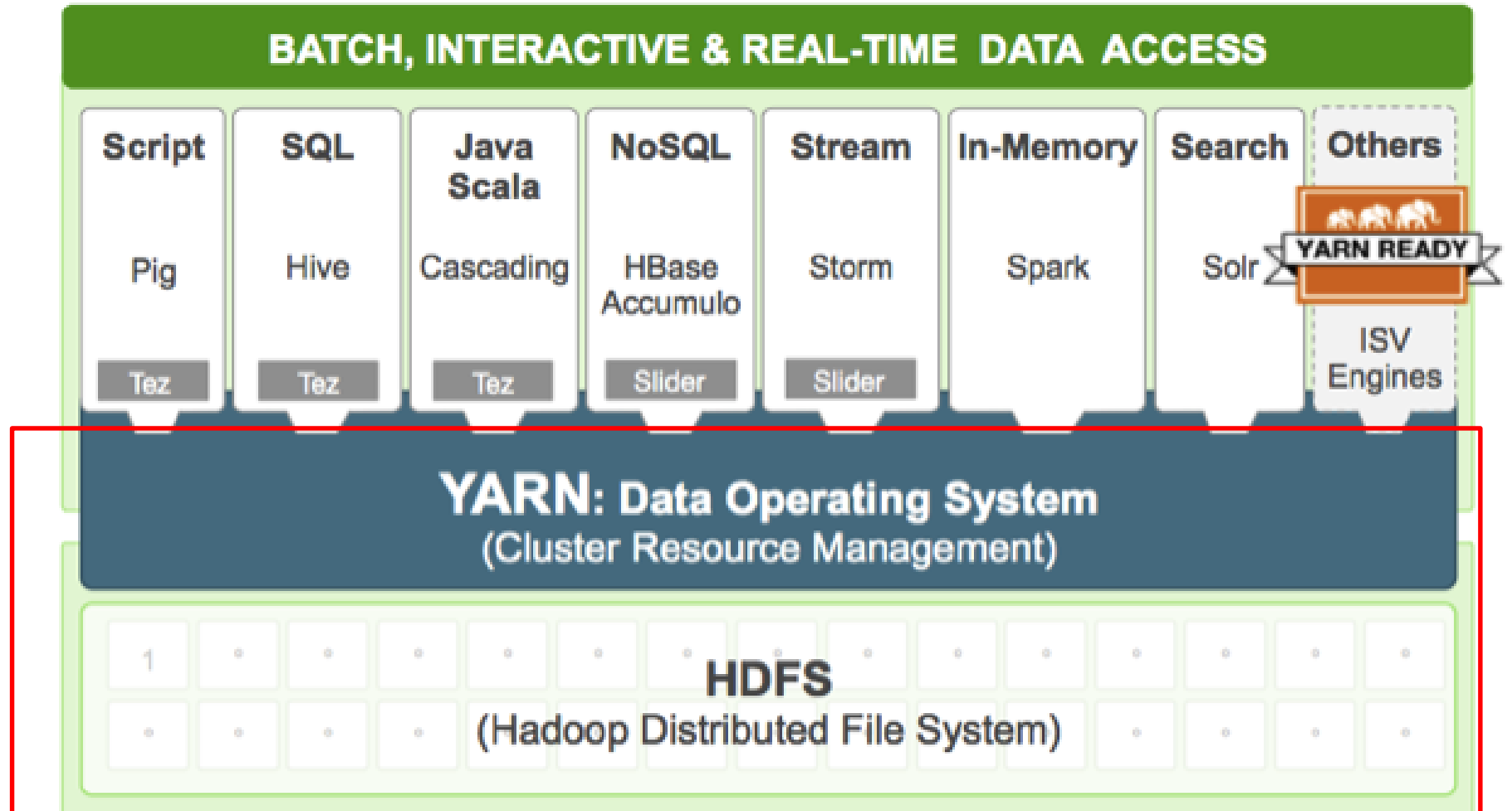


Ecosystème Hadoop

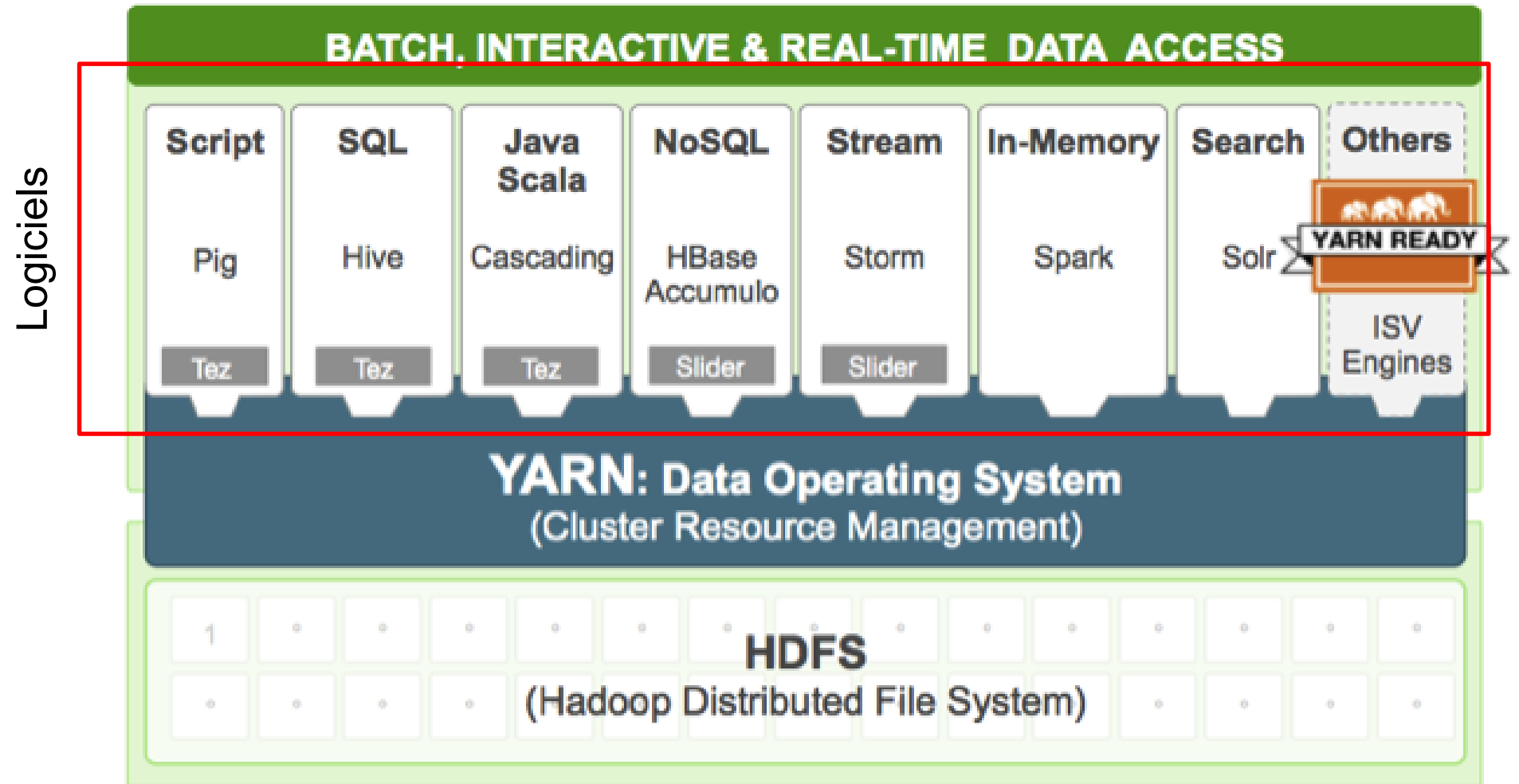


Ecosystème Hadoop

Système d'exploitation

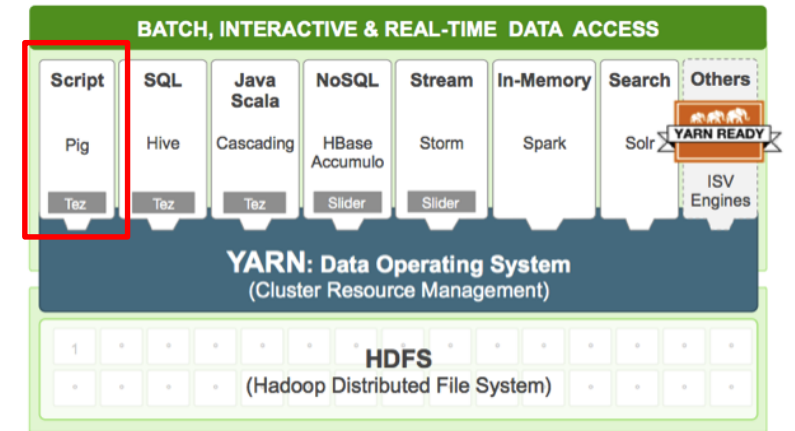


Ecosystème Hadoop



Ecosystème Hadoop

Pig



Apache Pig est un moteur d'exécution de script de haut niveau pour écrire des programmes d'analyse de données. Couplé avec Hadoop MapReduce et maintenant Tez + Yarn + HDFS, il permet l'évaluation de ces programmes de manière distribuée et transparente. L'utilisation de Pig permet ainsi le traitement de données très massives en toute simplicité.

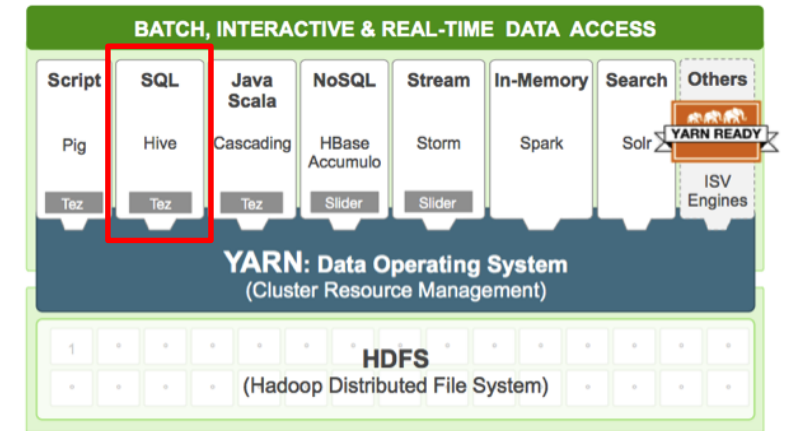
Le langage Pig Latin a les propriétés suivantes:

- Facile
- Optimisé
- Extensible

```
A = LOAD 'data' USING PigStorage() AS (f1:int, f2:int, f3:int);  
B = GROUP A BY f1;  
C = FOREACH B GENERATE COUNT ($0);  
DUMP C;
```

Ecosystème Hadoop

Hive



Apache Hive TM est une base de donnée qui facilite creation de requête et la manipulation des jeux de donnée stockés de manière distribuée.

La particularité de Hive est de permettre l'utilisation d'un langage de type SQL (HiveQL) pour effectuer l'administration et les requêtes sur les données. Hive permet simultanément la création de code Map Reduce pour créer des opération plus efficaces.

Propriété:

- Langage SQL
- Support de MapReduce
- Intégration facilitée

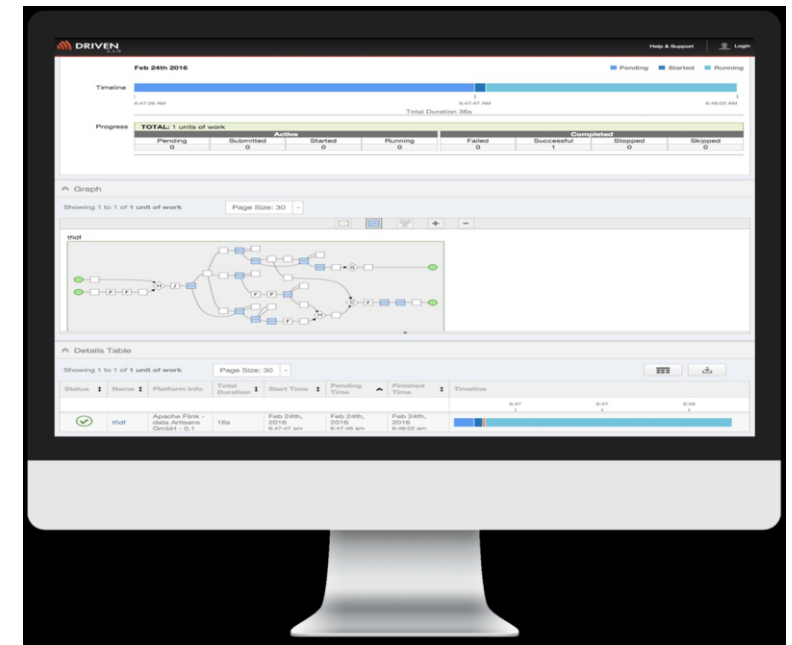
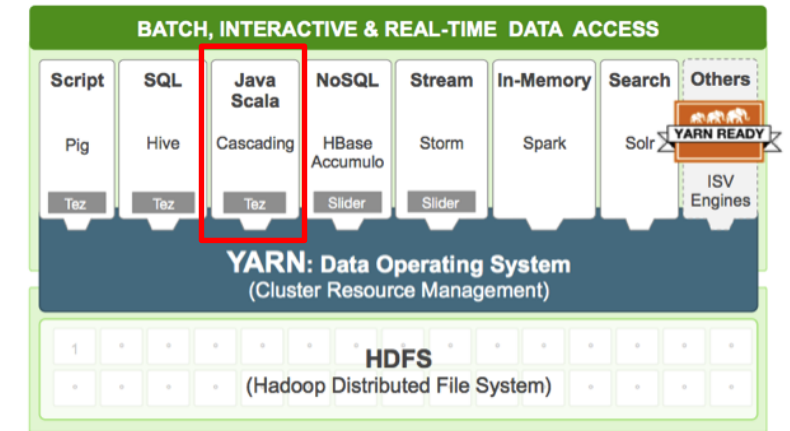
```
SELECT COUNT(*) FROM u_data;
```


Ecosystème Hadoop Cascading

Cascading, est un logiciel qui permet de modéliser des flux de transformation de données. Il permet de chainer des processus de traitement (notamment des Job Map-Reduce) les un avec autres.

Cascading est sous licence Apache.

Cascading permet de créer des flux de complexes et garanti que toutes les opérations s'exécutent bien correctement. Gestion des dépendances de Job.

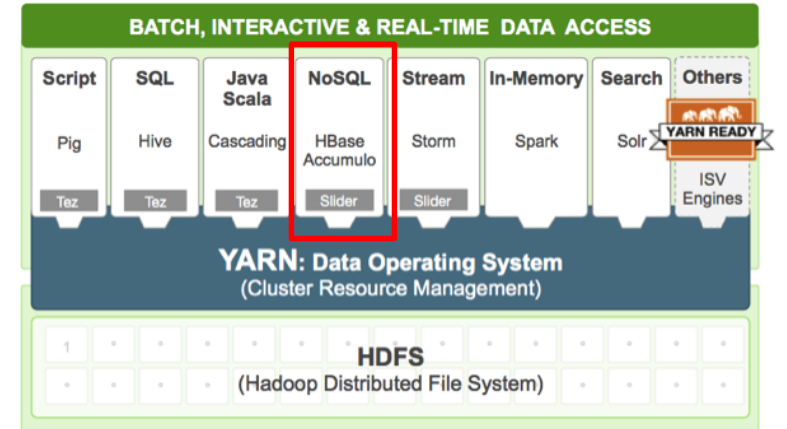


Ecosystème Hadoop HBase

Apache HBase est la base de donnée de Hadoop. Elle est distribuée et elle passe l'échelle horizontalement.

C'est une base de données de type « NoSQL ». Cela signifie que HBase n'offre pas toutes les fonctionnalités des « Base de données relationnelle ».

Par exemple, le langage SQL n'est pas supporté, les colonnes ne sont pas typées ... HBase s'inspire très fortement de la base de données de Google appelé BigTable.



Ecosystème Hadoop Storm

Apache Storm permet de faire du calcul distribué à la volée. Là où Map Reduce travaille en batch.

Spark est une implémentation du paradigme map reduce en utilisant uniquement un stockage en mémoire. Il permet d'être plus efficace que MapReduce (x10).

Solr, permet de faire de l'indexation de texte basée sur Lucène.

