Institut
Mines-Télécom

Une brève introduction aux Données Massives - Challenges et perspectives

Romain Picot-Clémente
Cécile Bothorel
Philippe Lenca

Plan

1 Big Data

2 4Vs

3 Hadoop et son écosystème

4 Nouveaux challenges, nouvelles formations

5 Conclusion

Plan

1 Big Data

2 4Vs

3 Hadoop et son écosystème

4 Nouveaux challenges, nouvelles formations

5 Conclusion

Big Data – Données Massives

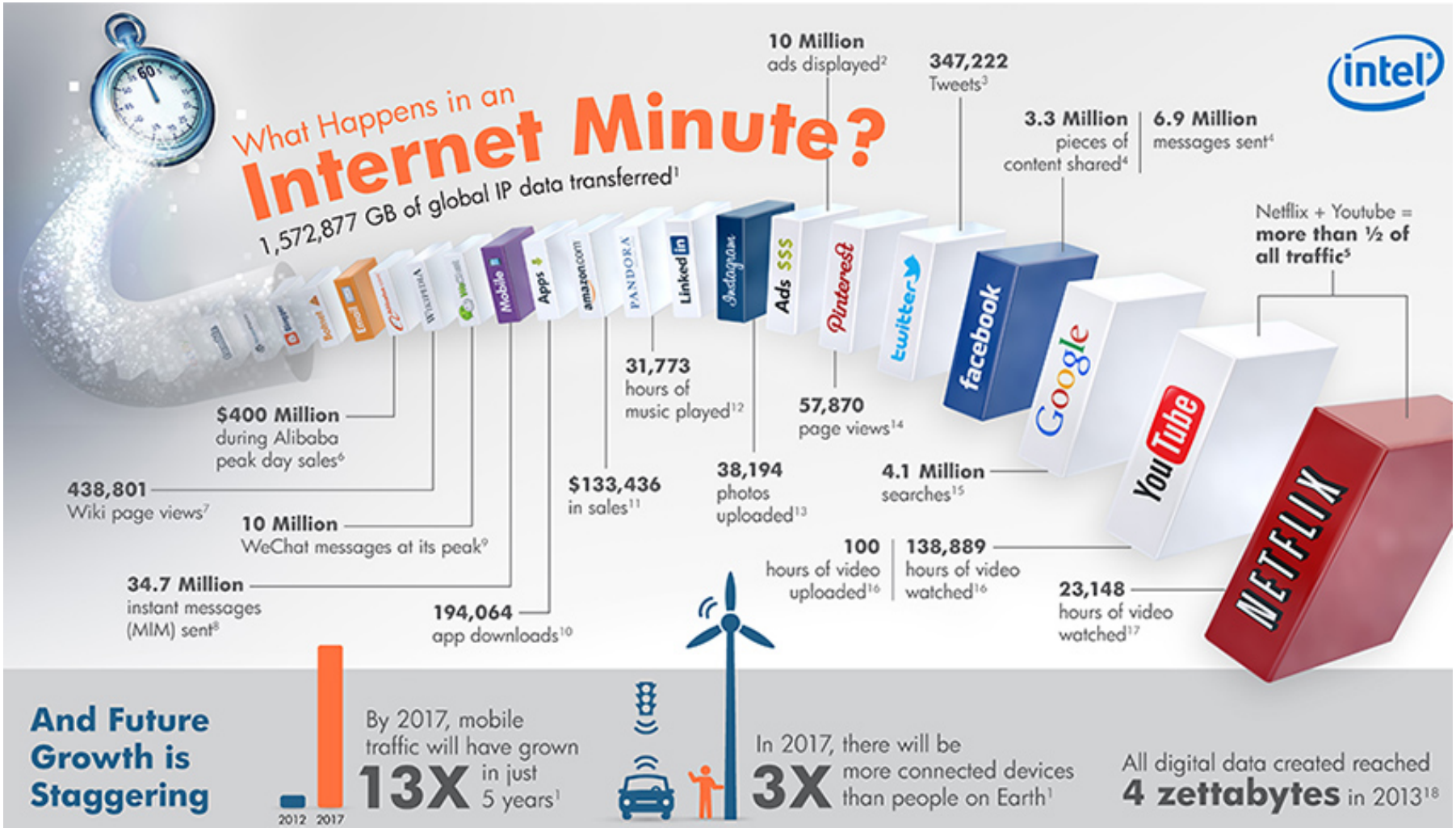
Tout le monde en parle mais qu'est-ce que c'est ?

Les big data, (grosses données, ou mégadonnées, données massives), désignent des ensembles de données qui deviennent tellement **volumineux** qu'ils en deviennent **difficiles** à travailler avec des **outils classiques** de gestion de base de données ou de gestion de l'information. [fr.wikipedia.org/wiki/Big_data]

"Big data is like teenage sex : everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it." **[Dan Ariely]**

▶ **Changement de pratique (outils, méthodologie).**

Sources de ces données massives (1/4)



[www.intel.com]

TELECOM Bretagne

Sources de ces données massives (2/4)

Mais encore...

- génomique
- téléphonie
- objets connectés, capteurs
- open data
- astrophysique, météo
- observation de la terre (climat, catastrophes)
- ...

▶ **Quatre grands challenges (cyber sécurité, ville intelligente, transport intelligent et le médical).**

Sources de ces données massives (3/4)

Bref, elles sont partout, c'est le déluge **[The economist, 2010]**

▶ **Et on y a de plus en plus accès (capacité de stockage et de traitement en forte hausse –liée à une forte baisse des prix–).**

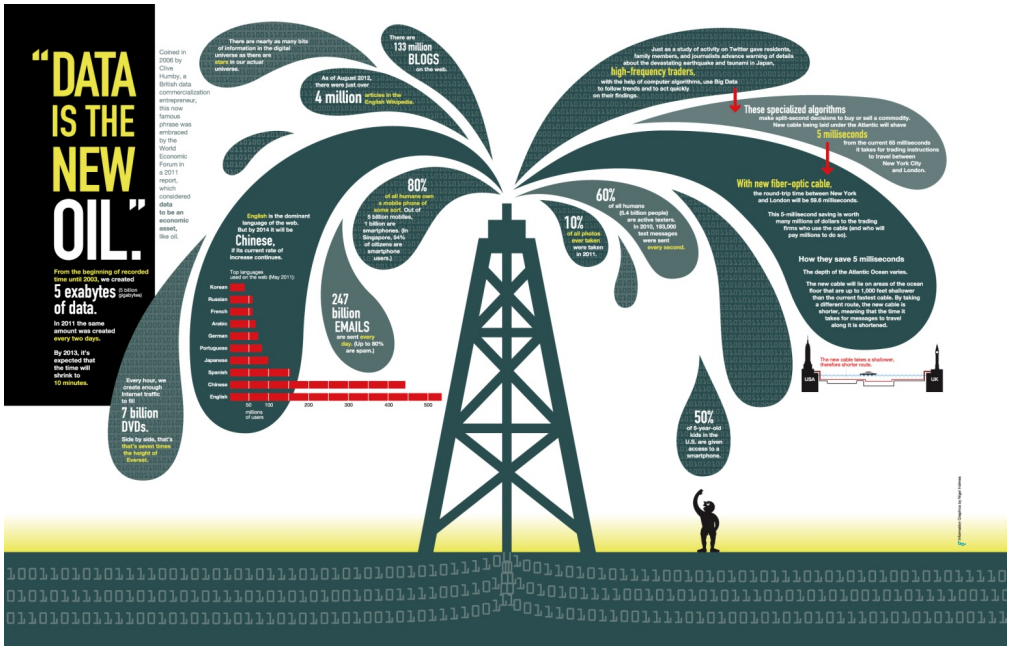
Sources de ces données massives (4/4)

Bref, elles sont partout, c'est le déluge

- qui n'a pas de téléphone portable ?
- qui n'a pas de compte sur un réseau social ?
- qui n'a jamais réalisé un achat sur internet ?
- qui n'a pas un objet connecté ?
- qui n'a pas d'assurance ?
- qui n'a pas de compte bancaire ?
- ...

▶ **Un individu-jour génère aujourd'hui plus de données qu'un Néandertalien (☺) pendant toute sa vie. On sait tout de nous.**

Analyse des Données Massives



[Clive Huby, 2006 ; Infographie : Nigel Holmes]

Ensemble des processus et technologies permettant de stocker, traiter et analyser des données massives.

▶ **Données - Connaissances - Valeur**

Plan

1 Big Data

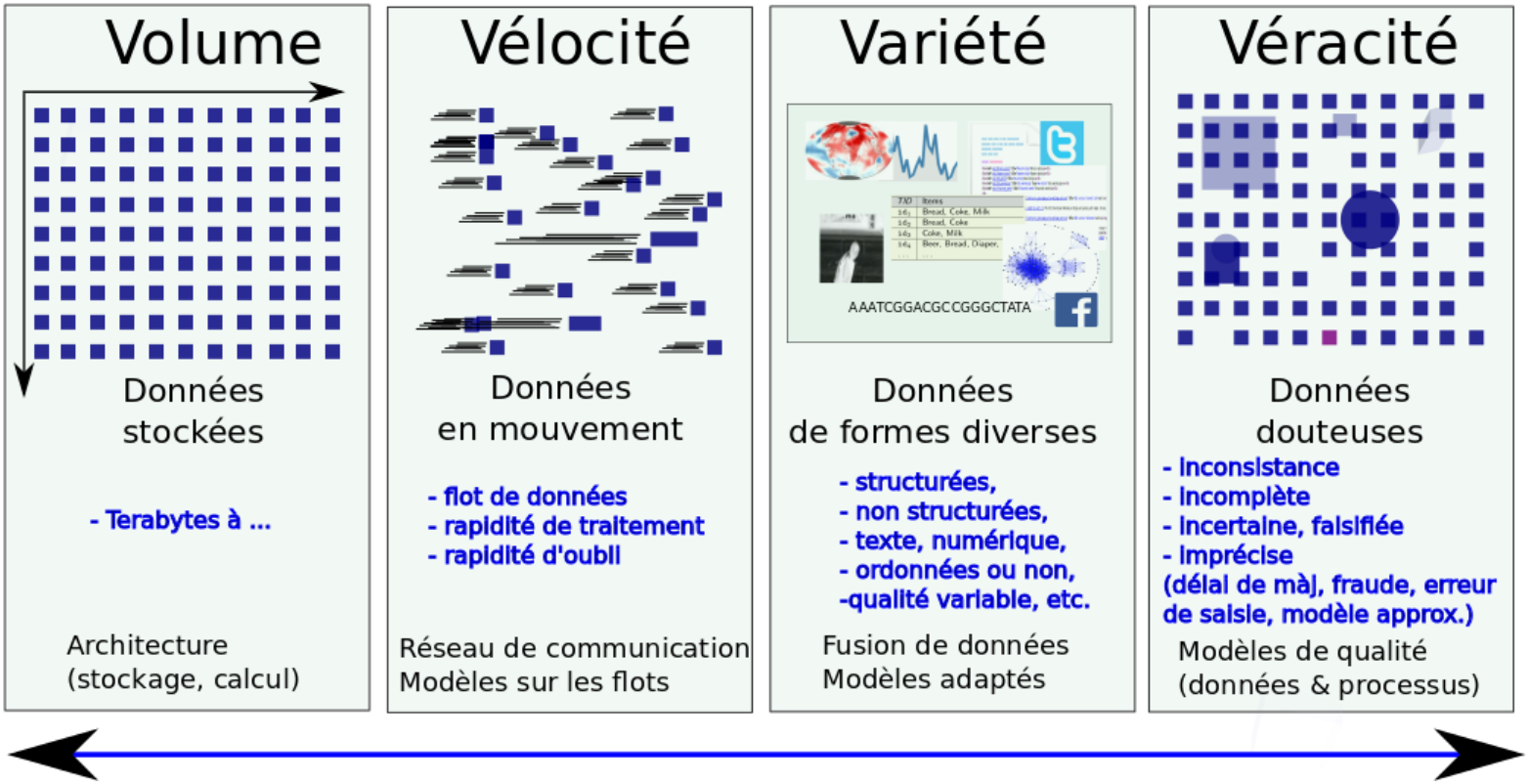
2 4Vs

3 Hadoop et son écosystème

4 Nouveaux challenges, nouvelles formations

5 Conclusion

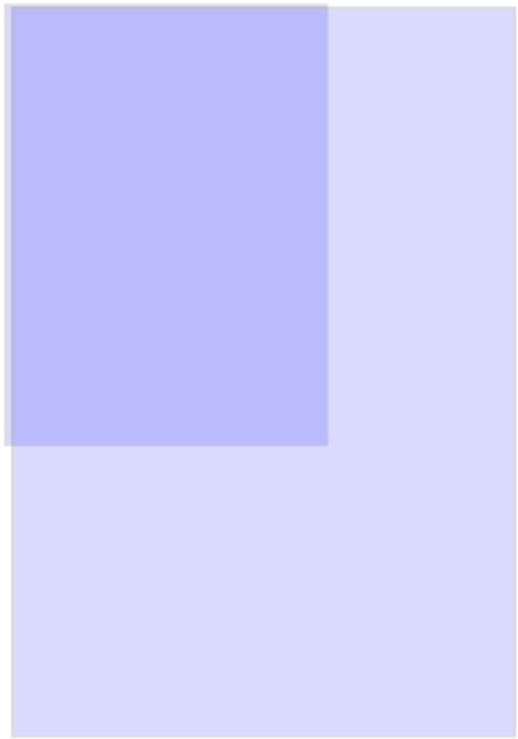
Données massives



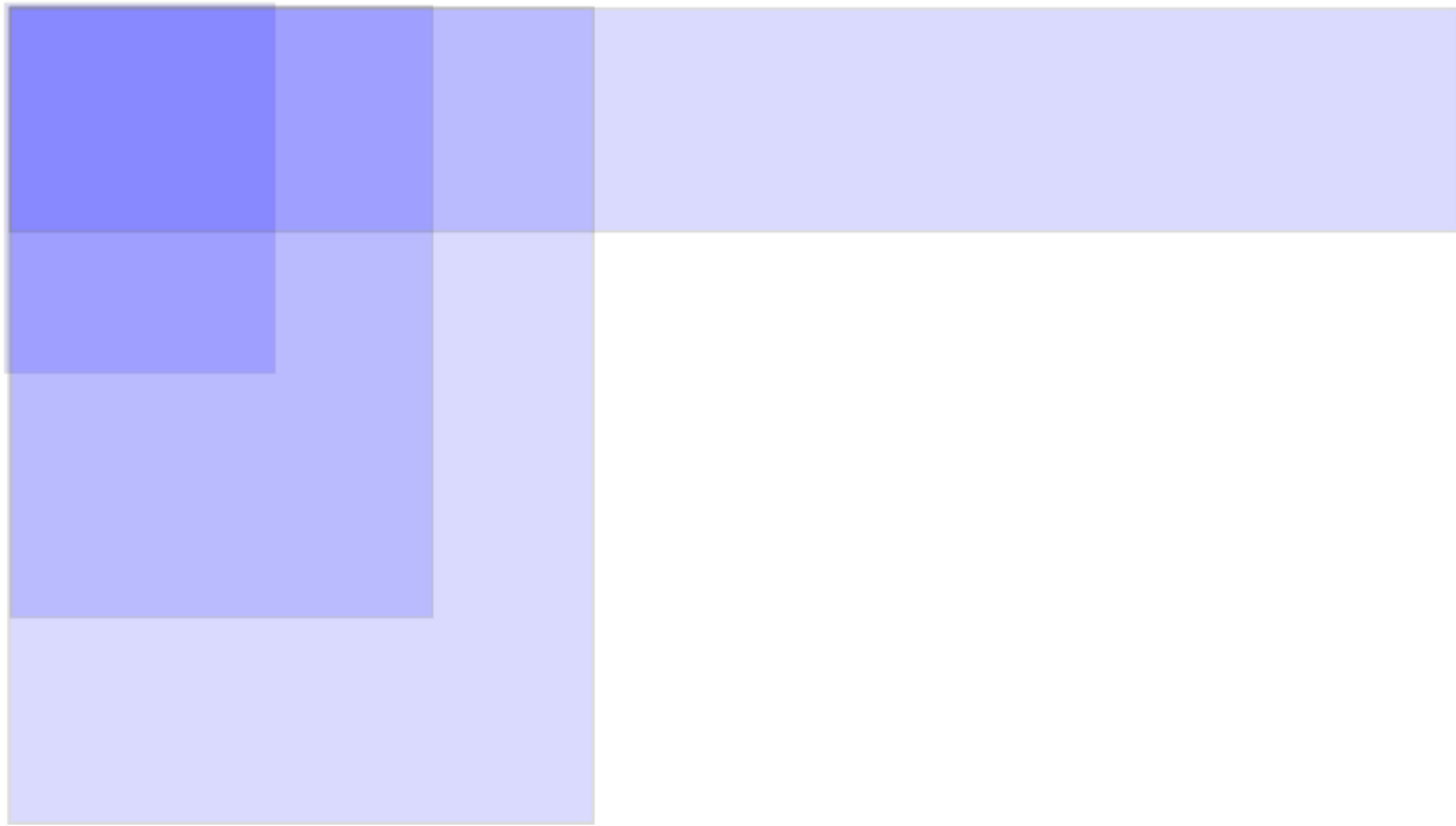
- **Valeur, Visualisation, . . .**

Données massives

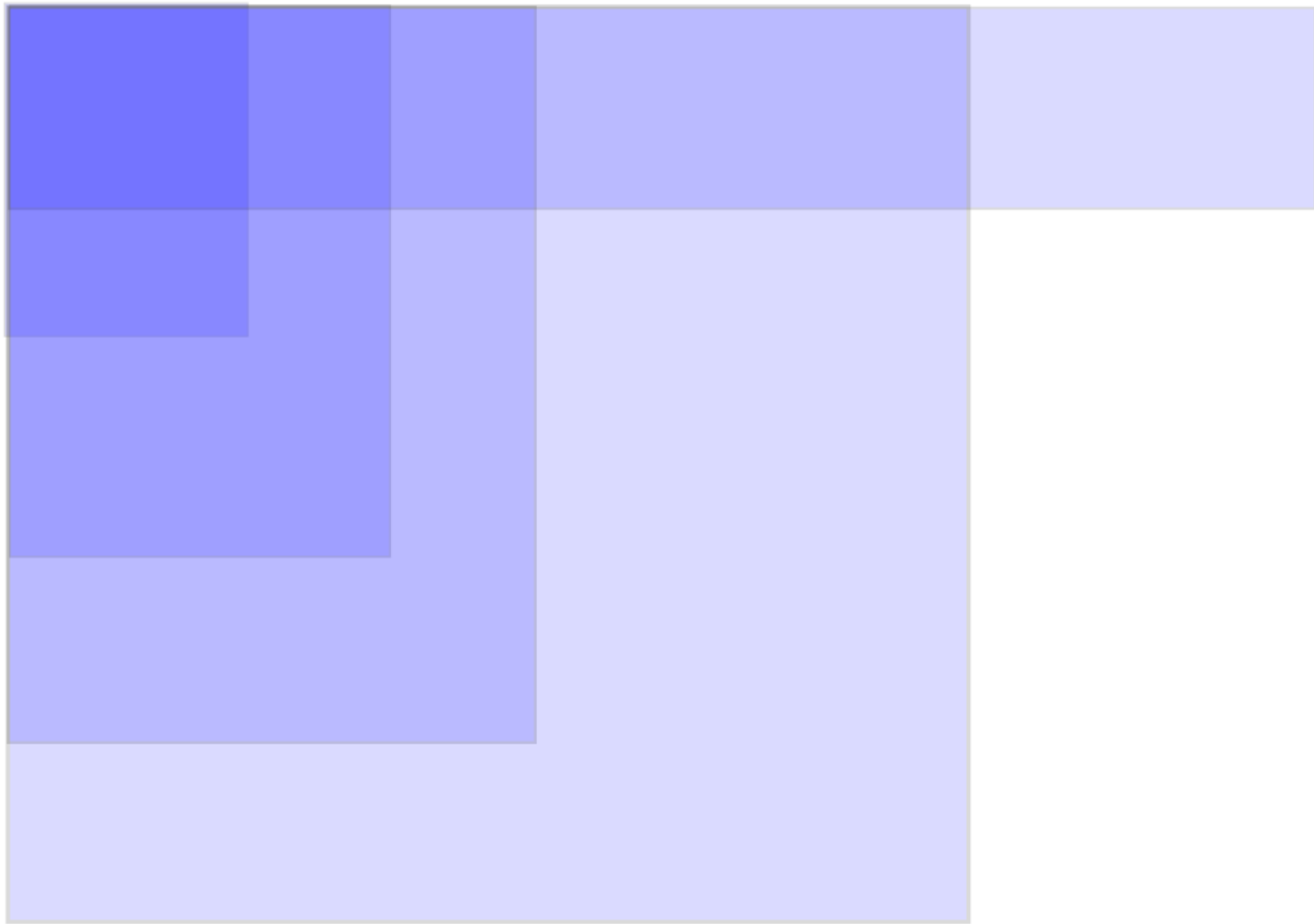
Données massives



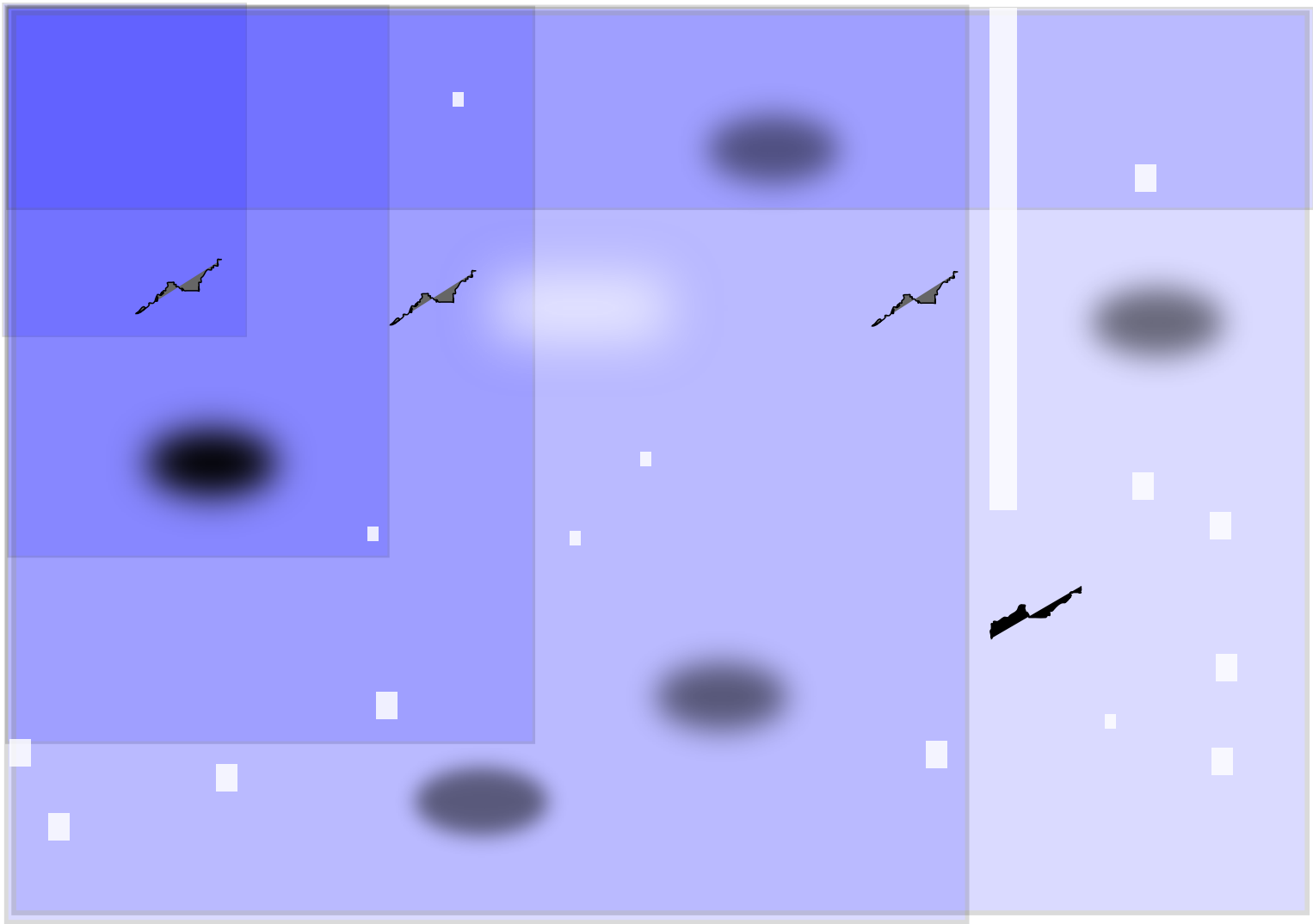
Données massives



Données massives



Données massives



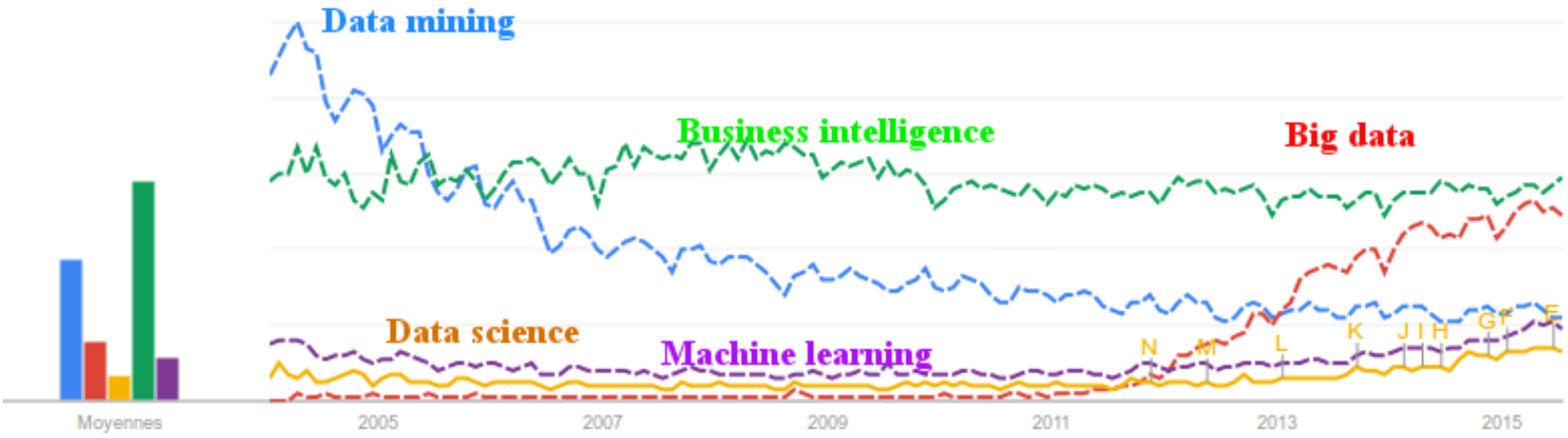
Changements

Quelques problèmes

- fausses corrélation
- difficultés à évaluer les modèles
- estimation et tests
- pas de contrôle sur la production des données
- temps d'analyse (qualité des données)
- outils classiques ne savent pas traiter les grands Volumes
- récence, représentativité des données

▶ Massive n'est pas Meilleure, et les Algorithmes dans tout cela ? – Big, Rich and Right Data –

Nouvelle science ?



Nouvelle science & nouveaux enjeux ?

Et demain ?

- 8/9 solutions industrielles françaises (ville durable, mobilité écologique, transports de demain, médecine du futur, économie des données, objets intelligents, confiance numérique, alimentation intelligente)
  [Industrie du futur, 18 mai 2015]
- développements informatiques et nouvelles approches d'analyse
- changements de comportements
- éthique : vie privée et risque de Big Brother, qui possède la donnée, droit à l'oubli, etc.

▶ **Enjeux scientifiques, économiques et sociétaux.**

Plan

1 Big Data

2 4Vs

3 Hadoop et son écosystème

4 Nouveaux challenges, nouvelles formations

5 Conclusion

Hadoop

Généralités

- Une approche proposée dans la suite des travaux de Google
- ≪ Un framework Java libre destiné à **faciliter la création d'applications distribuées et échelonnables** (scalables), permettant aux applications de travailler avec des milliers de nœuds et des péta-octets de données ≫ (wikipédia)

▶ **Un socle pour un écosystème riche**

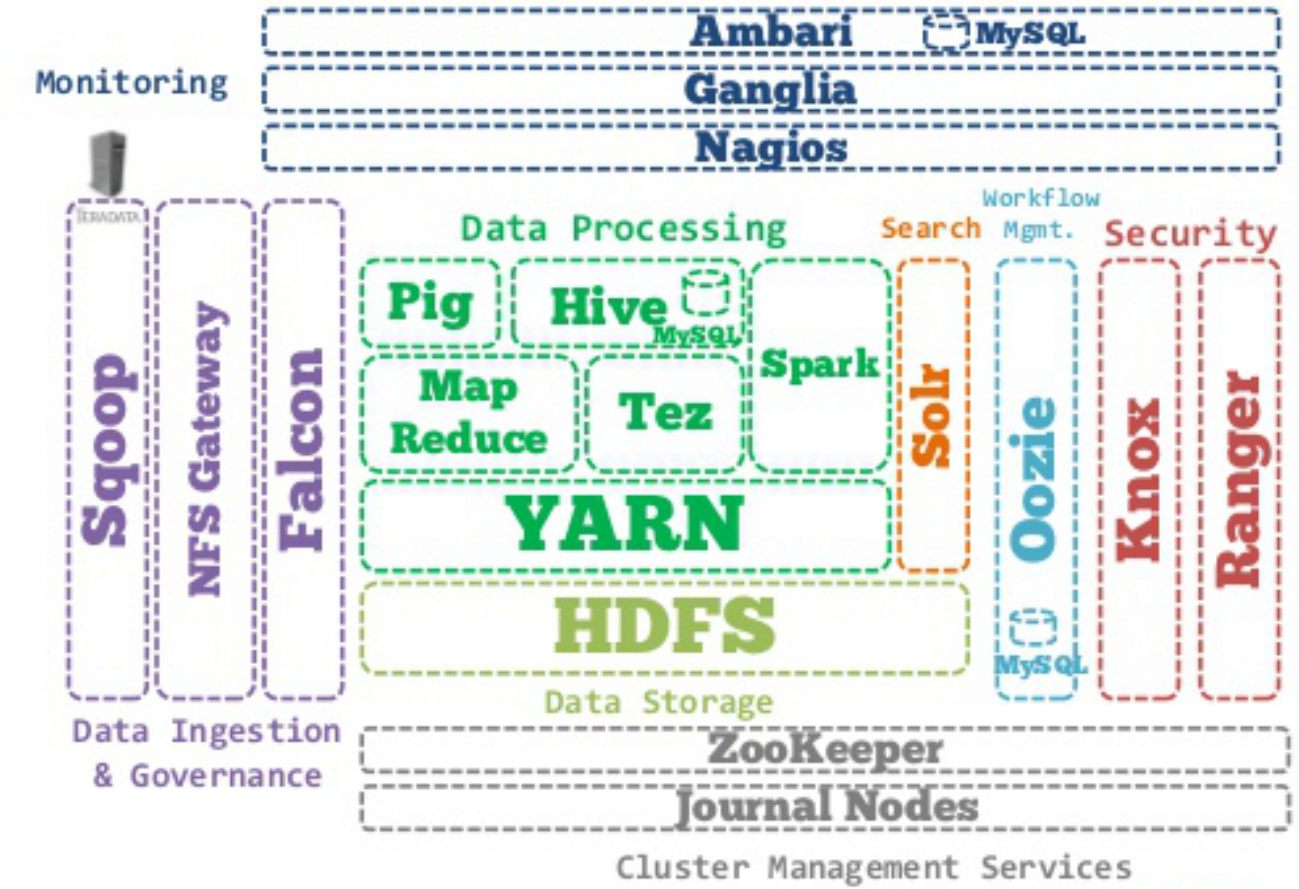
Hadoop

Cahier des charges lors de la conception d'Hadoop

- Un cluster Hadoop doit pouvoir **regrouper plusieurs dizaines, centaines ou milliers de nœuds** : chaque nœud permet d'offrir du stockage et une puissance de calcul
- Un cluster Hadoop doit pouvoir **stocker et traiter des gros volumes de données** dans des délais et coûts acceptables
- Si un nœud tombe, cela ne doit pas entrainer l'arrêt du calcul ou la perte de données (**résistance à la panne**)
- Une machine peut être rajoutée dans le cluster pour améliorer les performances (**évolutivité**)

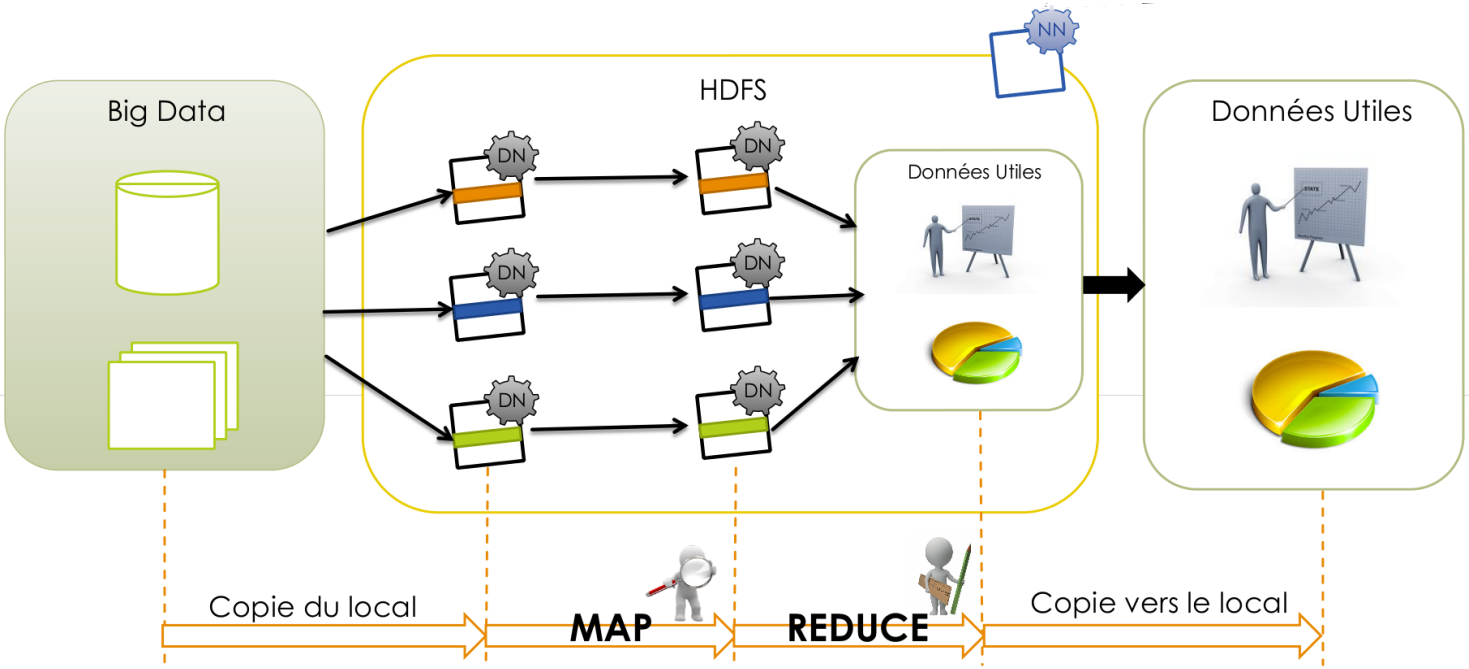
Un écosystème en expansion

Apache Hadoop Ecosystem



Hadoop à l'origine : HDFS et Map-Reduce

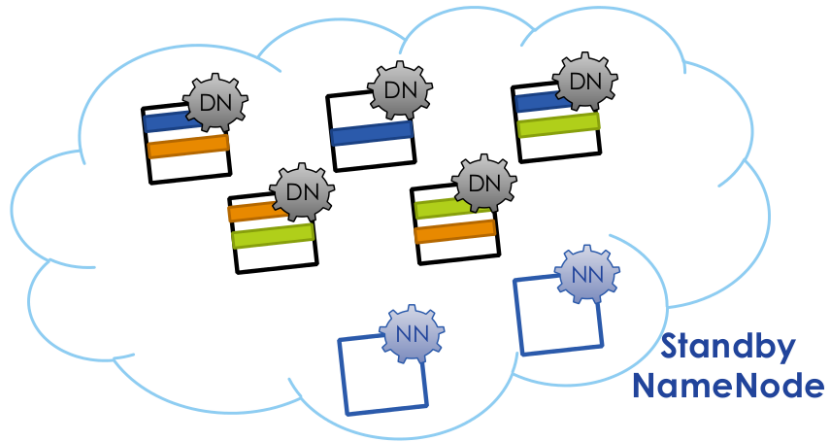
- Le projet Hadoop constitué originellement de deux parties :
  - Stockage des données : **HDFS** (**H**adoop **D**istributed **F**ile **S**ystem)
  - Traitement des données : **Map-Reduce**
- Principe :
  - Diviser les données
  - Les sauvegarder sur un ensemble de machines, appelé cluster
  - Traiter les données directement là où elles sont stockées

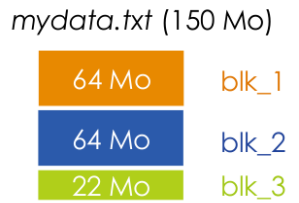


HDFS : Hadoop Distributed File System

Système de fichiers distribué, extensible et portable

- Permet de **stocker de très gros volumes de données** sur un grand nombre de machines (noeuds) équipées de disques durs usuels
- Quand un fichier mydata.txt est enregistré dans HDFS, il est décomposé en grands blocs (64Mo par défaut)
- Chaque bloc est enregistré sur un nœud et **répliqué** sur deux autres nœuds

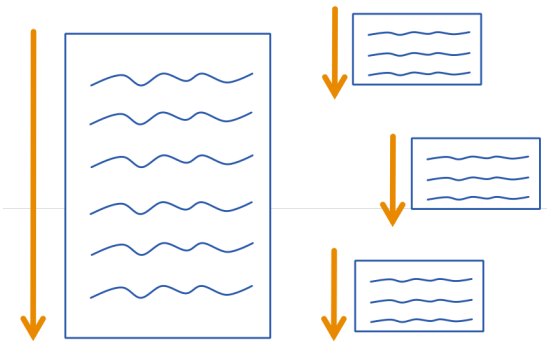




Map-Reduce

Patron d'architecture de développement pour du calcul haute performance sur plusieurs milliers de machines

- **Réduit le déplacement des données** entre machines, qui est la source de la complexité en distribué
- Fournit un **haut niveau de transparences** aux utilisateurs
  - Masque la parallélisation des traitements
  - Prend en charge la tolérance aux pannes
  - Gère l'équilibrage des charges et la coordination
- Fournit un modèle ≪ relativement ≫ **simple à comprendre et à programmer**



Discussion

Atouts de Hadoop

- Forte **tolérance aux pannes** (sécurité et persistance des données)
- Capacité de **montée en charge maximale**
- Capacité d'**analyse** et de **traitement** des données **à grande échelle**
- **Coût matériel très faible**

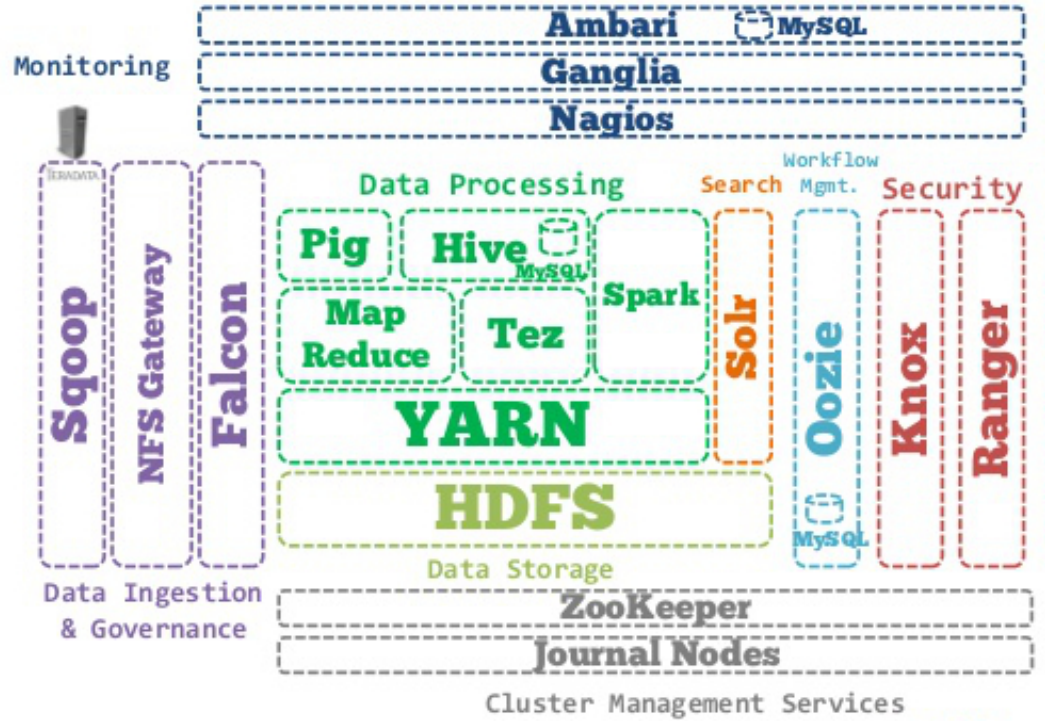
Discussion

Inconvénients de Hadoop

- **Administration complexe** : besoin de développer une expertise spécifique Hadoop ou faire appel à des prestataires extérieurs
- Map-Reduce conçu pour du “batch processing”, n’est **pas adapté à des traitements temps réel**
- Map-Reduce **difficilement utilisable pour des algorithmes de machine learning** car il n’y a pas de moyen aisé pour communiquer un état partagé d’une itération à une autre

Plus loin avec l'écosystème Hadoop

- Des **alternatives à Map-Reduce** qui utilisent la RAM partagée, particulièrement **adaptées au machine learning, à l'analyse de graphe** : Tez, Spark (très grande communauté)
- Des **langages de haut niveau** pour le traitement des données (étapes Map-Reduce transparentes) : Hive, Pig
- De **nombreux outils** pour la gestion des données, la gestion du cluster, la sécurité, ...



Comment se lancer ?

Se familiariser avec Hadoop

- Apprendre avec une machine virtuelle de type Cloudera :
  - Utilisation de HDFS, premiers programmes avec map-reduce, initiation aux outils de plus haut niveau, etc.
- Utiliser ensuite les différents modes de fonctionnement d'Hadoop
  - Modes local, pseudo-distribué, totalement distribué ou virtualisé
  - Bâtir progressivement son cluster à partir de matériels recyclés puis achetés spécifiquement

Pour une entreprise

Un choix du DSI

- Un cluster dédié : installé physiquement dans l'entreprise ou chez un hébergeur
  - \+ confidentialité des données, maitrise du cluster, totale liberté
  - \- investissement, administration du cluster et ≪ tuning ≫
- Un cluster dans le cloud
  - \+ tarif compétitif, fiabilité et disponibilité élevée
  - \- limitations techniques, configurations imposées, confidentialité des données

Plan

1 Big Data

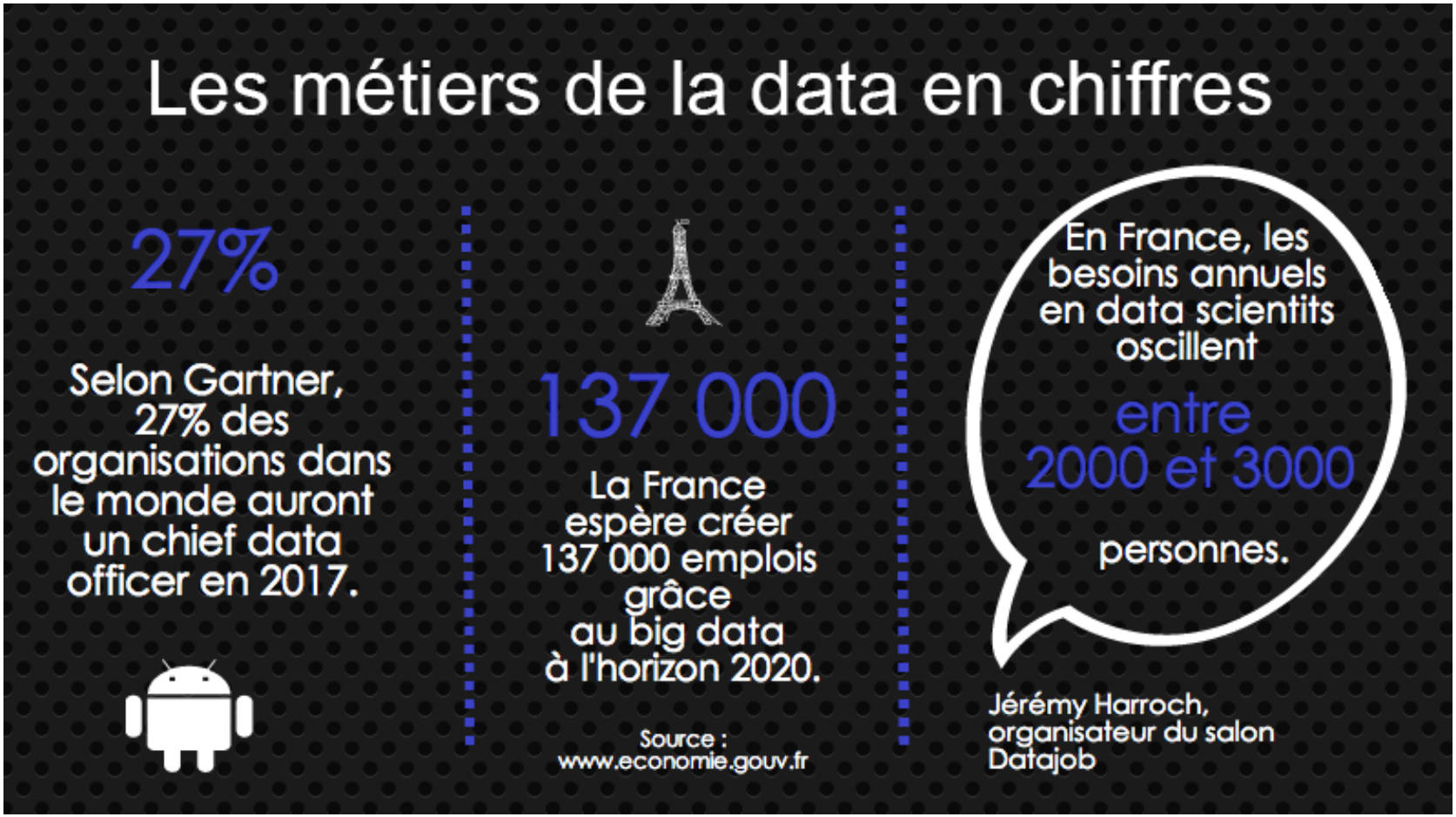
2 4Vs

3 Hadoop et son écosystème

4 Nouveaux challenges, nouvelles formations

5 Conclusion

Nouveaux challenges, nouveaux besoins



Métiers en émergence 1/2

Données métiers

- Volume et Vélocité ne sont pas une réelle nouveauté pour les actuaires
- Volume important pour modéliser les risques
- Vélocité très élevée en finance de marché (trading haute fréquence)
- Mais toute la richesse n'a pas été exploitée

3 profils

- Le **Data Miner**, explorateur des données
- Le **Data Scientist**, informaticien spécialisé dans l'analyse des données
- Le **Data Analyst**, responsable des opérations de bases de données et appui analytique à l'exploration de données

Source : Apec, BIG DATA : Les actuaires en première ligne (Institut des Actuaires)

Métiers en émergence 2/2

Données non structurées

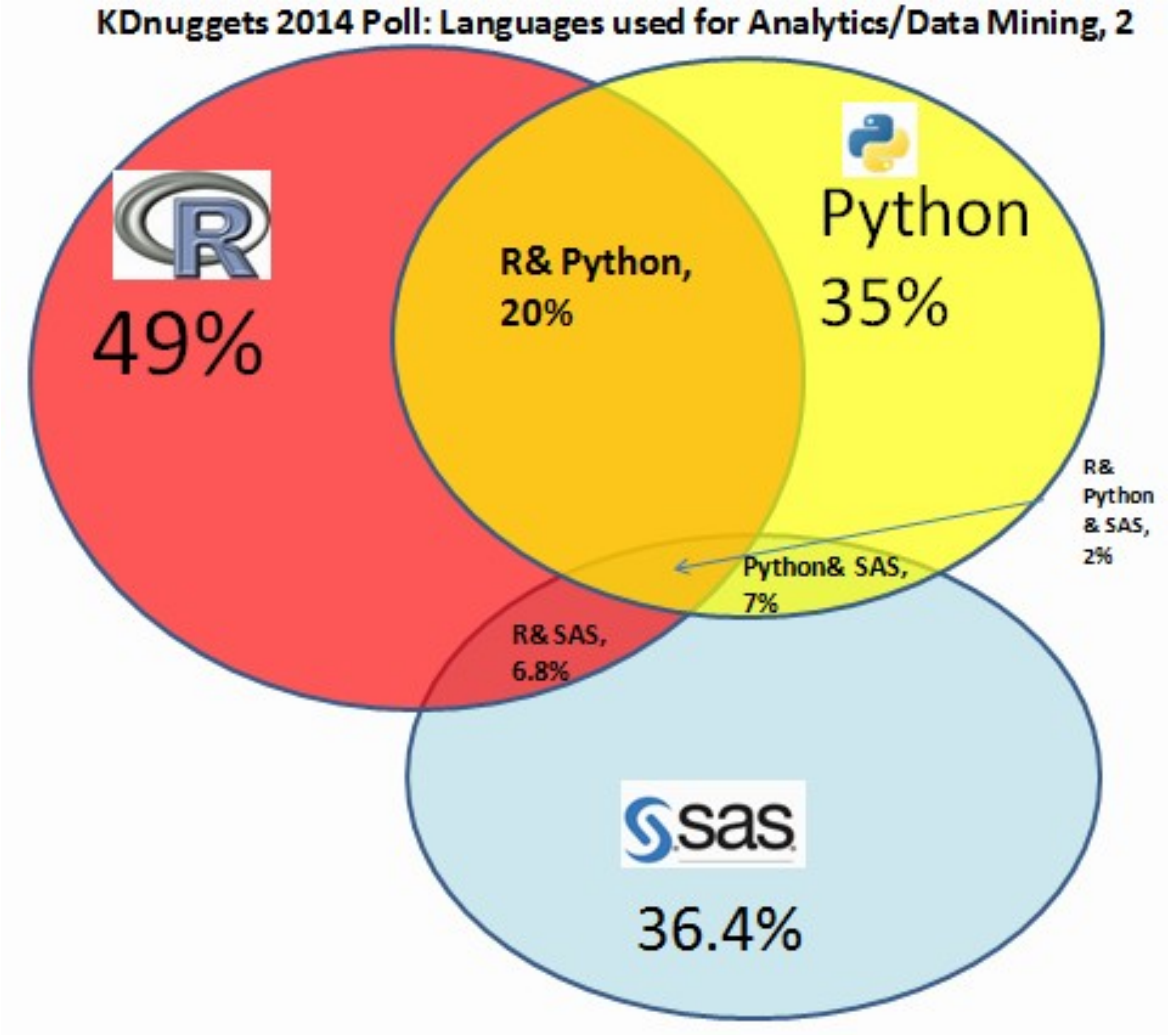
- Le troisième V **Variété** est encore très peu exploité, potentiel de **différentiation stratégique**
- Données non structurées : mails, photos, tweets, etc.
- Open Data : fichiers de cartes grises, statistiques de vente de médicaments

3 profils s'ajoutent

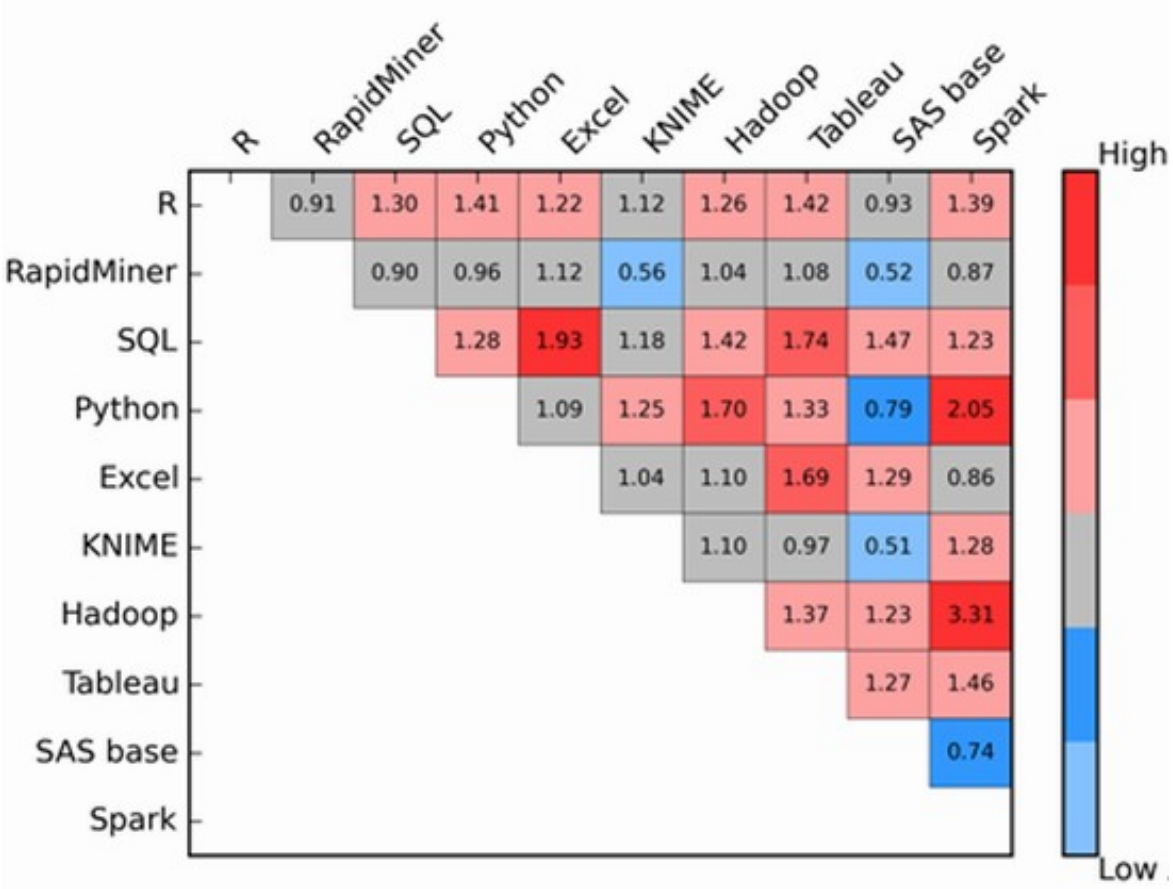
- Le **Chief Data Officer**, responsable de la collecte des données et garant de l'éthique
- Le **Responsable de la relation client** et le **Chef de projet e-CRM**, pour une expérience utilisateur au coeur de la stratégie

Sources : Apec, chronique "Les clés de demain" (Le Monde), BIG DATA : Les actuaires en première ligne (Institut des Actuaires)

Outils



Outils



- Logiciels d'informatique décisionnelle s'intègrent petit à petit aux environnements Big Data (Data Loader for Hadoop de SAS)
- Librairies Python, R gratuites disponibles... et utilisables sous Hadoop et Spark

FIGURE : KDnuggets' Association Matrix Heat Map for top 10 most popular data mining tools

Formation à Telecom Bretagne : un parcours Data Scientist

`http://www.telecom-bretagne.eu/formations/`

Cycle ingénieur : parcours Systèmes d'information décisionnels
Master of Science : Informatique et systèmes de décision
Mastère spécialisé : Informatique appliquée à la décision bancaire et actuarielle

- Enseignements des fondamentaux
  - Probabilités, Statistique, TPs en Python et R
  - Fouille de données, méthodologie CRISP-DM
  - Recherche opérationnelle, aide à la décision
  - Base de données, reporting & Business Intelligence, ETL Talend Open Studio
- Architectures Big Data
- Maîtrise du décisionnel avec SAS
- Big Data Analytics
  - Sous une machine virtuelle de type Cloudera
  - Utilisation de HDFS, premiers programmes Map/Reduce en Python
  - Initiation aux outils de plus haut niveau comme la librairie Mahout
  - Bientôt Spark, langage Python et Scala

▶ **Fouille de données et Big Data Analytics dispensés à l'Euria**

Formation pour l'actuariat

Formation Data Science de l'Institut des Actuaires

Spécifiquement destinée aux actuaires, vise à compléter les formations en actuariat (initiales et continues) par une formation opérationnelle en extraction, gestion et analyse des données massives et hétérogènes.

La profession développe ce domaine

- AXA par exemple investit 800M€ dans le Big Data sur 3 ans, des équipes se créent chez Allianz, Bnp Paribas Cardif, dans les cabinets de conseil de type Optimind Winter.
- Des actuaires rédigent des livres sur le sujet (ex : **[M. Dupuis & E. Berthelé, Big Data dans l'assurance, 2014]**.
- L'Institut des Actuaires encourage la recherche au sens large, et ce sujet en particulier. Si des papiers intéressants voient le jour, l'Institut les mettra en valeur (revues actuarielles, conférences, etc).

Plan

1 Big Data

2 4Vs

3 Hadoop et son écosystème

4 Nouveaux challenges, nouvelles formations

5 Conclusion

Conclusion

Quelques challenges

- challenges techniques, massif $\neq$ meilleur
- challenges éthiques, droit à l'oubli, privacy
- challenges politiques, de la mutualisation à l'hyper-segmentation

Quelques opportunités nouvelles

- de mettre le client au centre de la stratégie
- de marché avec l'ouverture des données

Une seul exemple : évaluation des risques

A priori (tranche d'âge, zone géographique, CSP, etc.) vs. in situ fondée sur les comportements réels (et donc individualisés).

▶ **Les actuaires ont un rôle clé à jouer.**