
Big Data : une introduction

DRIO 5101A : Big Data Analytics avec Spark
Thibaud Vienne - ESIEE Paris

Fonctionnement du cours

- 11 heures de cours :
 - Big Data : Une introduction (2h)
 - Panorama de l'écosystème Hadoop (3h)
 - Le framework Spark (3h)
 - Machine learning avec Spark (3h)

- 15 heures de TP :
 - TP 1 : Tutorial + Comptage de mots avec Spark.
 - TP 2 : Analyse de logs Apache.
 - TP 3 : Prédiction de dates de sorties de chansons.
 - TP 4 : Classification de clics internet.

- Evaluation finale

Prérequis

- Programmation Python :
 - Fonctionnalités basiques python.
 - Manipulation de listes et dictionnaires.
 - “List Compréhension” et fonctions “lambda”.
- Machine Learning :
 - Régressions linéaires, logistiques, généralisées, régularisation.
 - Techniques d’arbre (arbre de décision, forêts aléatoires, boosting).
 - Processus de validation croisée.

Sommaire

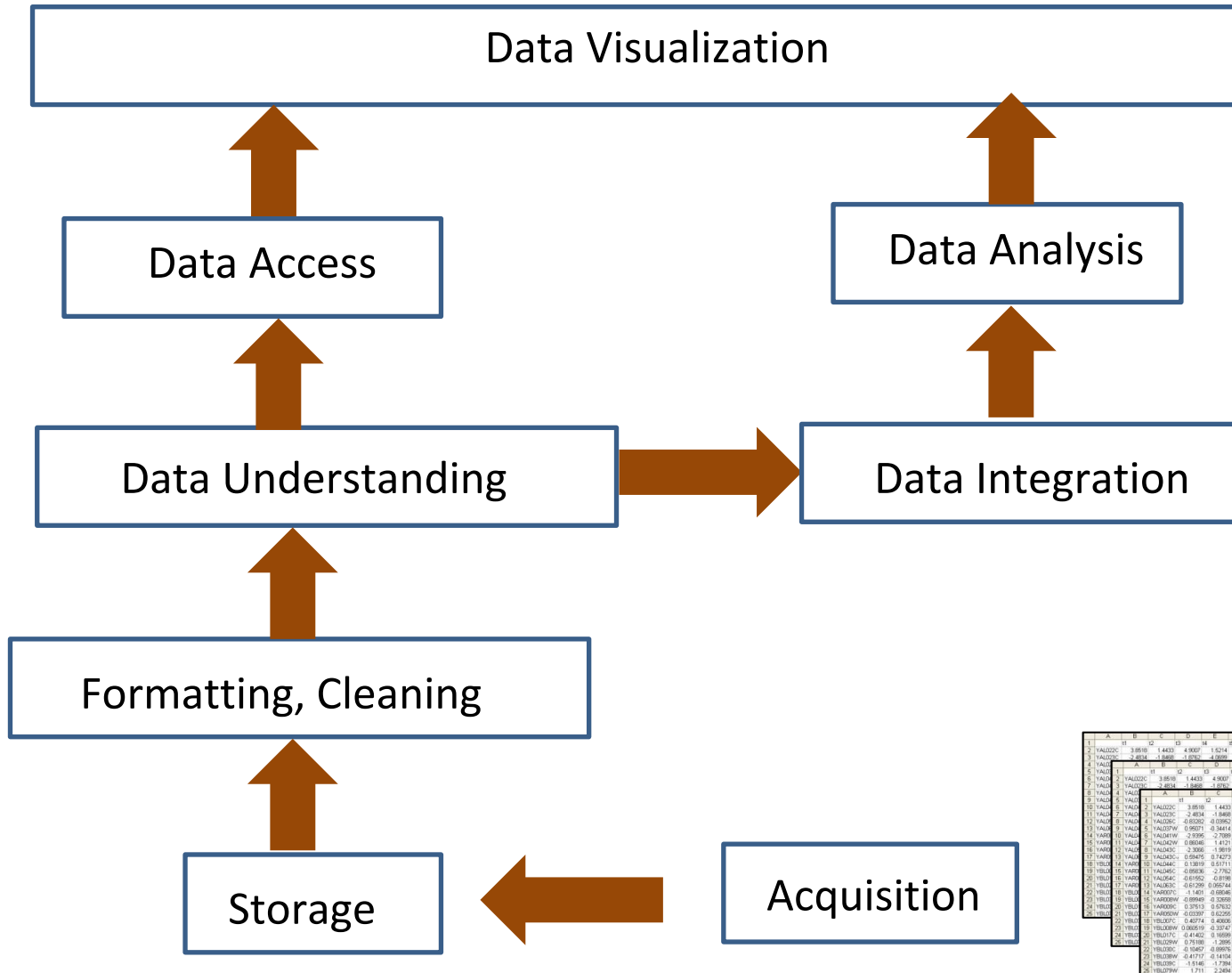
1. Big Data : première approche
2. Les solutions Big Data: définition et principes
3. Un exemple de projet Big Data
4. Les métiers de la Data

Big Data : première approche

Big Data : c'est quoi ?

- Terme assez vague. Pas de définition particulièrement homogène.
- **Un Big Data** (littéralement « grandes données ») est une collection de données si grande et si complexe qu'il devient très difficile de pouvoir traiter ces données avec des outils classiques de gestion de bases de données.
- **Une solution Big Data** est un ou un ensemble d'outils informatiques étant capable de gérer et de traiter des données Big Data.
- Le challenge Big Data comprend l'acquisition, le stockage, le nettoyage, le transfert, l'intégration, l'analyse et la visualisation de ces données.

Big Data : Un ensemble de processus

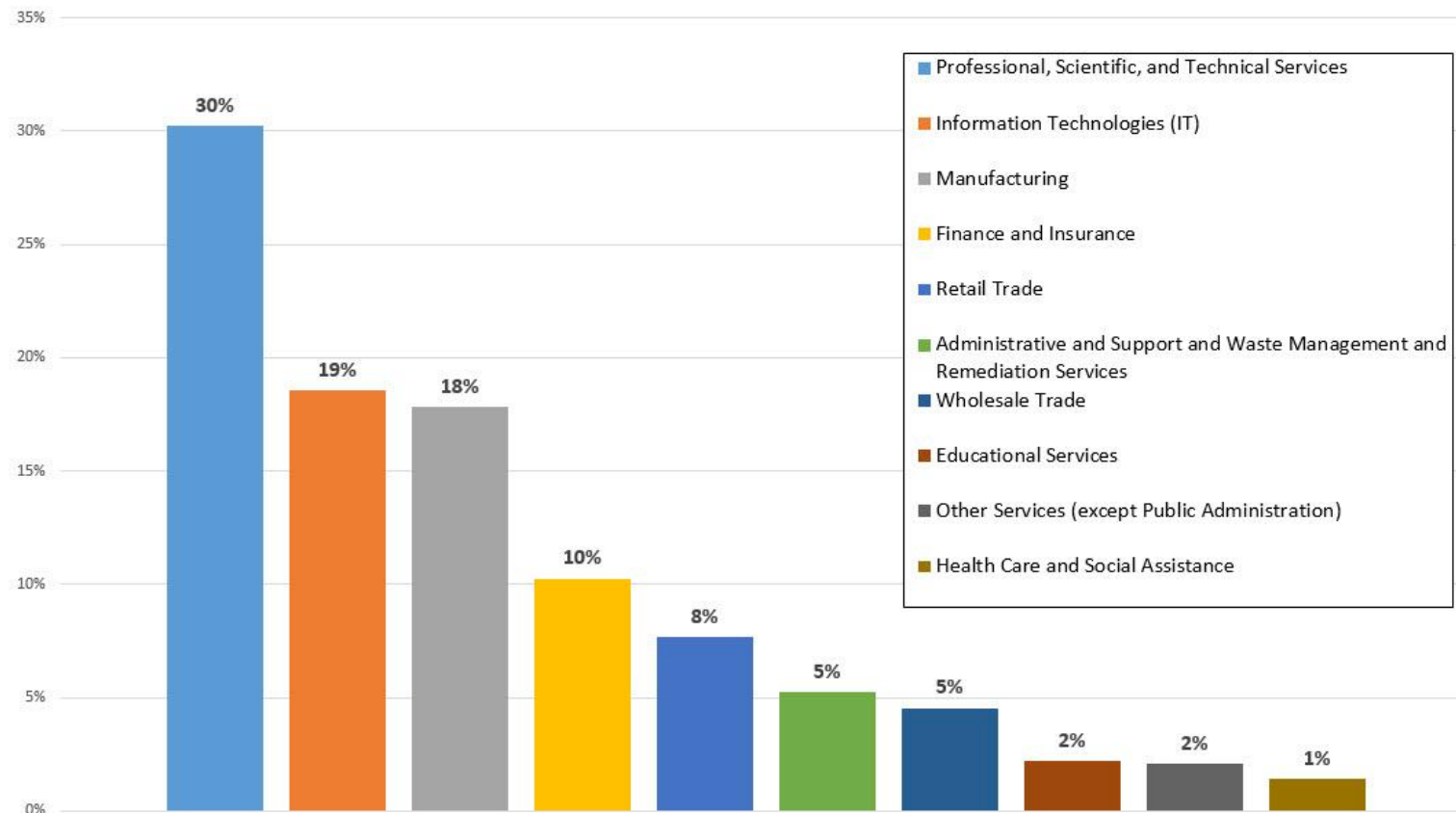


Big Data : Pourquoi tout le monde en parle ?

- Extraction de valeur supplémentaire rendue possible grâce aux solutions Big Data!

Top 10 Industries Hiring Big Data Expertise - Positions Advertised For In 2015

Source: WANTED Analytics, a CEB Company, 2015



Big Data : Effet de mode ou valeur ajoutée ?



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.

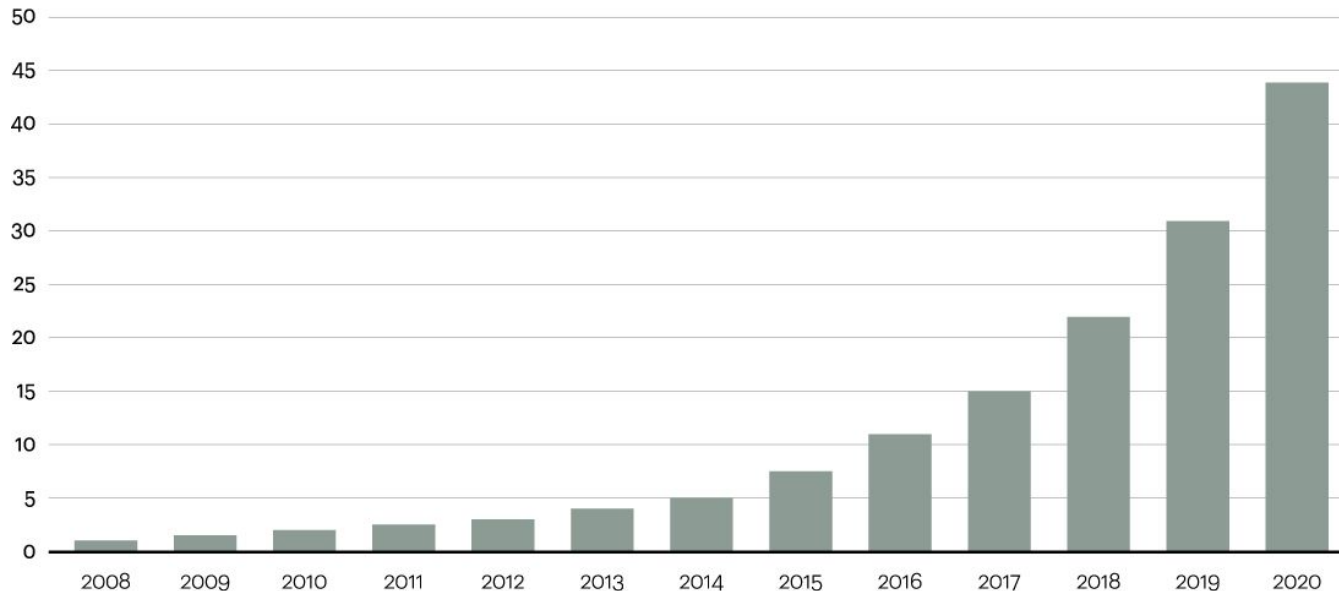
— *Dan Ariely* —

AZ QUOTES

Big Data : Des volumes toujours croissants

Data is growing at a 40 percent compound annual rate, reaching nearly 45 ZB by 2020

Data in zettabytes (ZB)



Source: Oracle, 2012

Système international (SI)		
Unité	Notation	Valeur
bit	bit	1 bit
kilobit	kbit	10^3 bits
mégabit	Mbit	10^6 bits
gigabit	Gbit	10^9 bits
térabit	Tbit	10^{12} bits
pétabit	Pbit	10^{15} bits
exabit	Ebit	10^{18} bits
zettabit	Zbit	10^{21} bits
yottabit	Ybit	10^{24} bits

Big Data : D'où viennent les données ?

➤ Les données générées par l'utilisateur:

- Réseaux sociaux
- Applications de partages de contenus



➤ Les données de supervisions issues de serveurs informatiques:

- Apache logs
- Machines syslog



➤ Les données scientifiques, financières et médicales:

- Données météorologiques, données spatiales
- encodage d'un génome
- ...



➤ Les données issues de capteurs:

- objets connectés
- tags RFID
- GPS



Big Data : Des leviers multiples



Innovations technologiques:

- Objets connectés + Internet des objets
- Cloud Computing
- Baisse des coûts de stockage de données



Changements sociaux:

- Internet accessible à tous
- Emergence des réseaux sociaux
- Partage d'informations devenu courant.



Nouvelles opportunités Business:

- Fidélisations clients supermarchés
- Extraction de valeur de données (assurance, marketing...)
- Business Models orientés « numériques »

Les solutions Big Data

Définition des 5 Vs

De projets à une solution Big Data

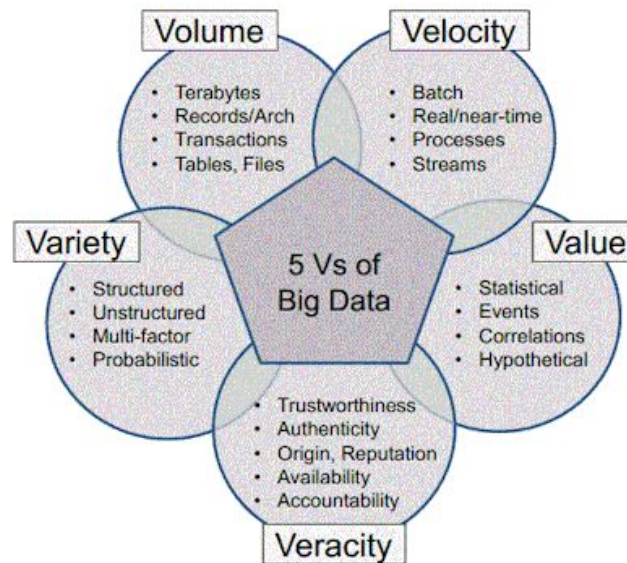
La solution Hadoop

Définition des 5V

- Les approches traditionnelles de gestion de bases de données et de l'exploitation de la donnée ne sont plus adaptées aux applications métiers actuelles tant en terme de:
- **Volume:** énormes (bien souvent supérieur à la centaine de TO) qui prennent trop de place pour pouvoir être stockées et exploitées sur une seule machine.
- **Vélocité:** la plupart des applications (surtout en présence de grands volumes) sont bien trop lentes aux vu des exigences métiers actuelles.
- **Variété:** Les données sont de structures différentes et parfois difficilement exploitables (csv, table SQL, vidéo, email...)
- **Véracité:** Les données sont-elles erronées? Sont-elles correctes?
- **Valeur :** la valeur contenue dans les données.

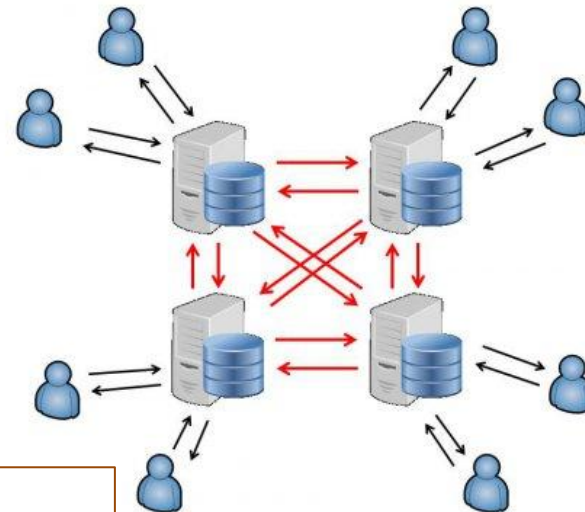
Définition des 5V

- Une solution Big Data est un ou un ensemble de technologies capable de gérer un très grand Volume de données à variétés multiples, avec une Vélocité suffisante.
- On retient de la définition d'une solution de Big data **la définition dites des 5Vs.**



Définition des 5V : Volume

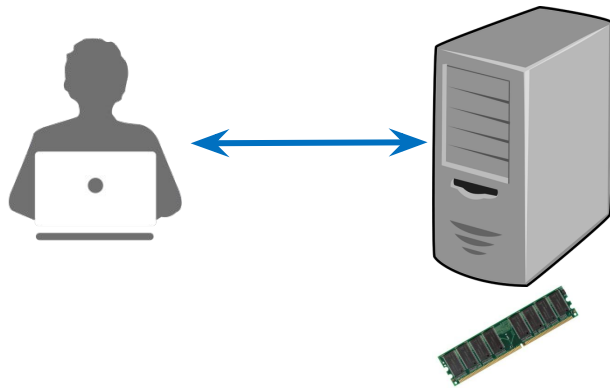
- Le volume réfère à la capacité d'une solution Big data à pouvoir gérer de gros volumes de données.
- Au vu de la quantité de données, il est impossible de stocker et traiter ces données sur une seule machine. On utilise alors des technologies utilisant des « **systèmes distribués** ».
- Processus de systèmes distribués:
 - Processus de stockage distribué
 - Processus de calcul distribué



Systeme
distribué

Définition des 5V : Volume

id	Prénom	Note
1	Nicolas	12
2	Mickael	16
3	Daphnée	11
4	Thomas	9
5	Marie	14
6	Yassine	13

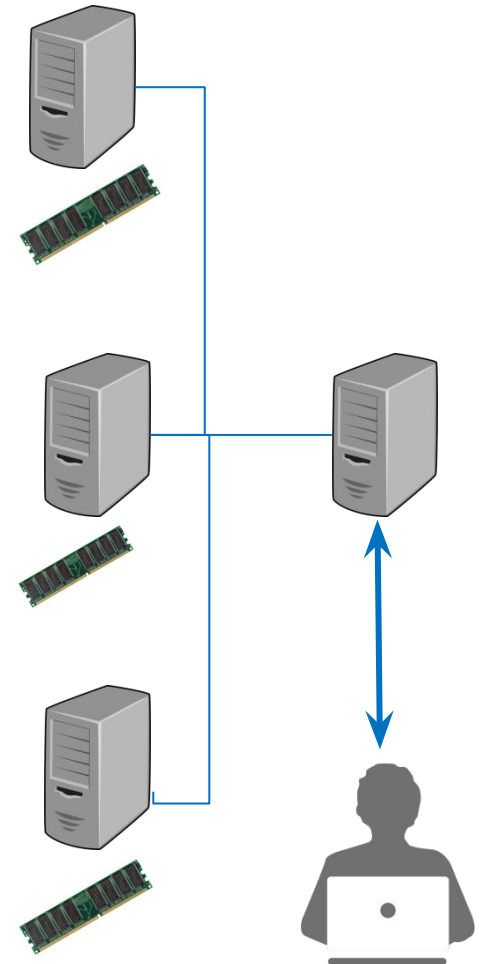


Système centralisé

id	Prénom	Note
1	Nicolas	12
2	Mickael	16

id	Prénom	Note
3	Daphnée	11
4	Thomas	9

id	Prénom	Note
5	Marie	14
6	Yassine	13



Système distribué

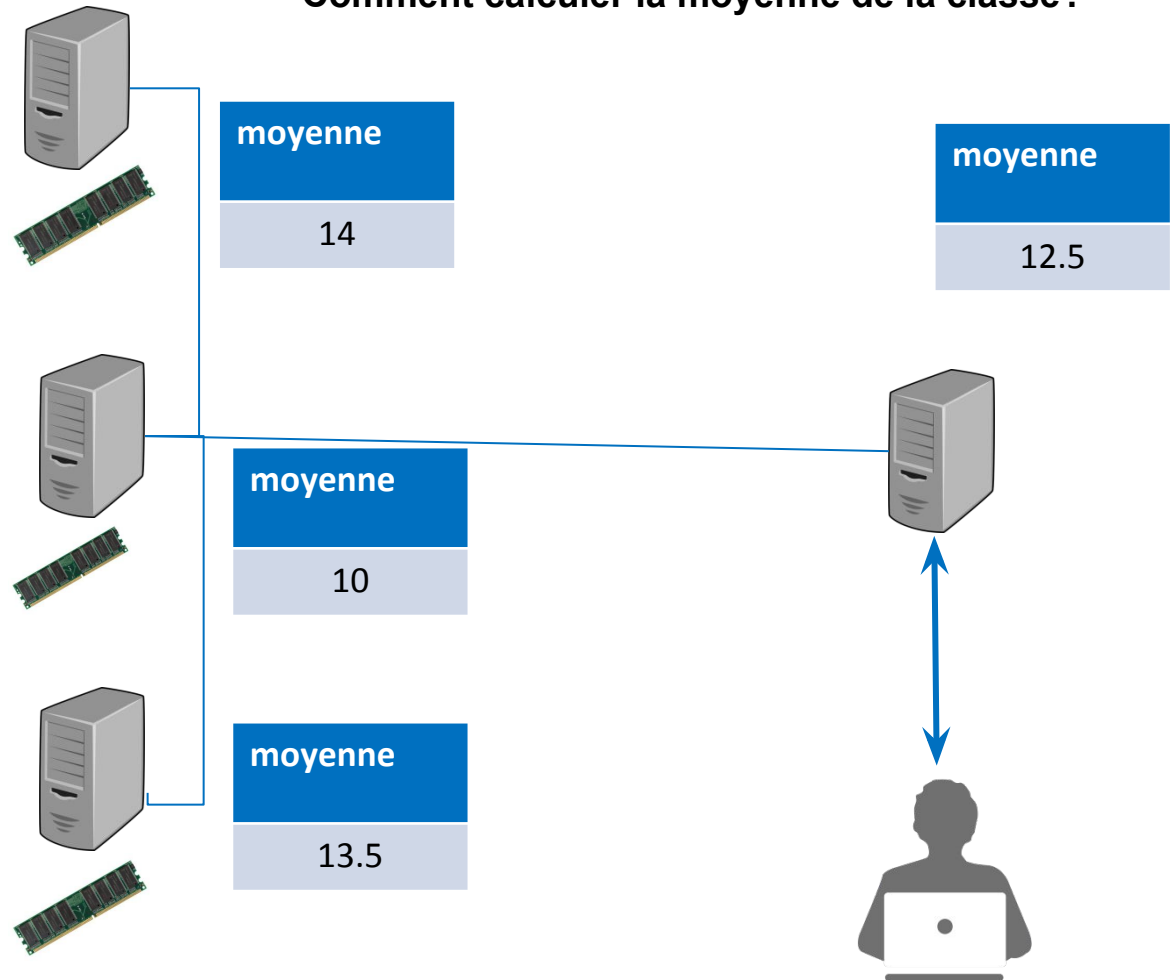
Définition des 5V : Volume

Comment calculer la moyenne de la classe?

id	Prénom	Note
1	Nicolas	12
2	Mickael	16

id	Prénom	Note
3	Daphnée	11
4	Thomas	9

id	Prénom	Note
5	Marie	14
6	Yassine	13



Systeme distribué

Définition des 5V : Volume

- Transparence pour l'utilisateur.
- Interopérabilité du système.
- Mise à l'échelle (fonctionne efficacement dans différentes échelles).
- Tolérance aux pannes.
- Sécurité (authentification, intégrité des données, disponibilité).

Définition des 5V : Volume

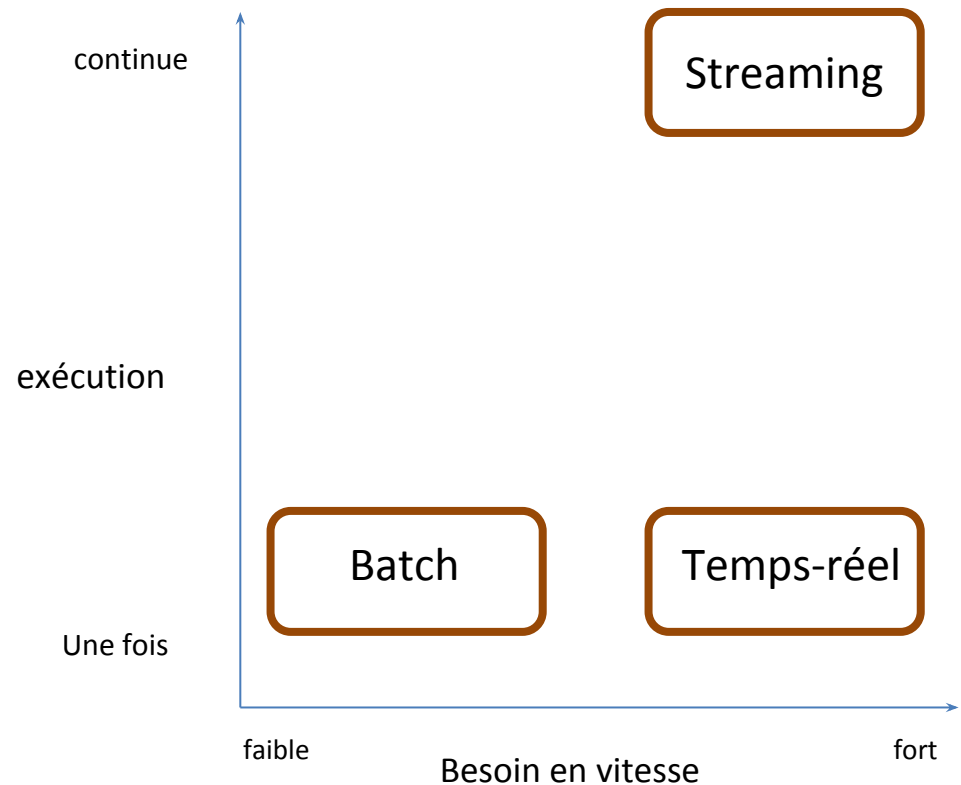
Avantages	Inconvénients
<ul style="list-style-type: none">• Bon rapport performance/prix.• Utilisation d'un maximum de ressources de calcul.• Système encore opérationnel même en cas de panne d'une machine.	<ul style="list-style-type: none">• Logiciels de gestion complexes.• Nécessite de bonnes performances en termes de télécommunications.• Nécessite des mécanismes de synchronisation et de sécurité.

Définition des 5V : Vitesse

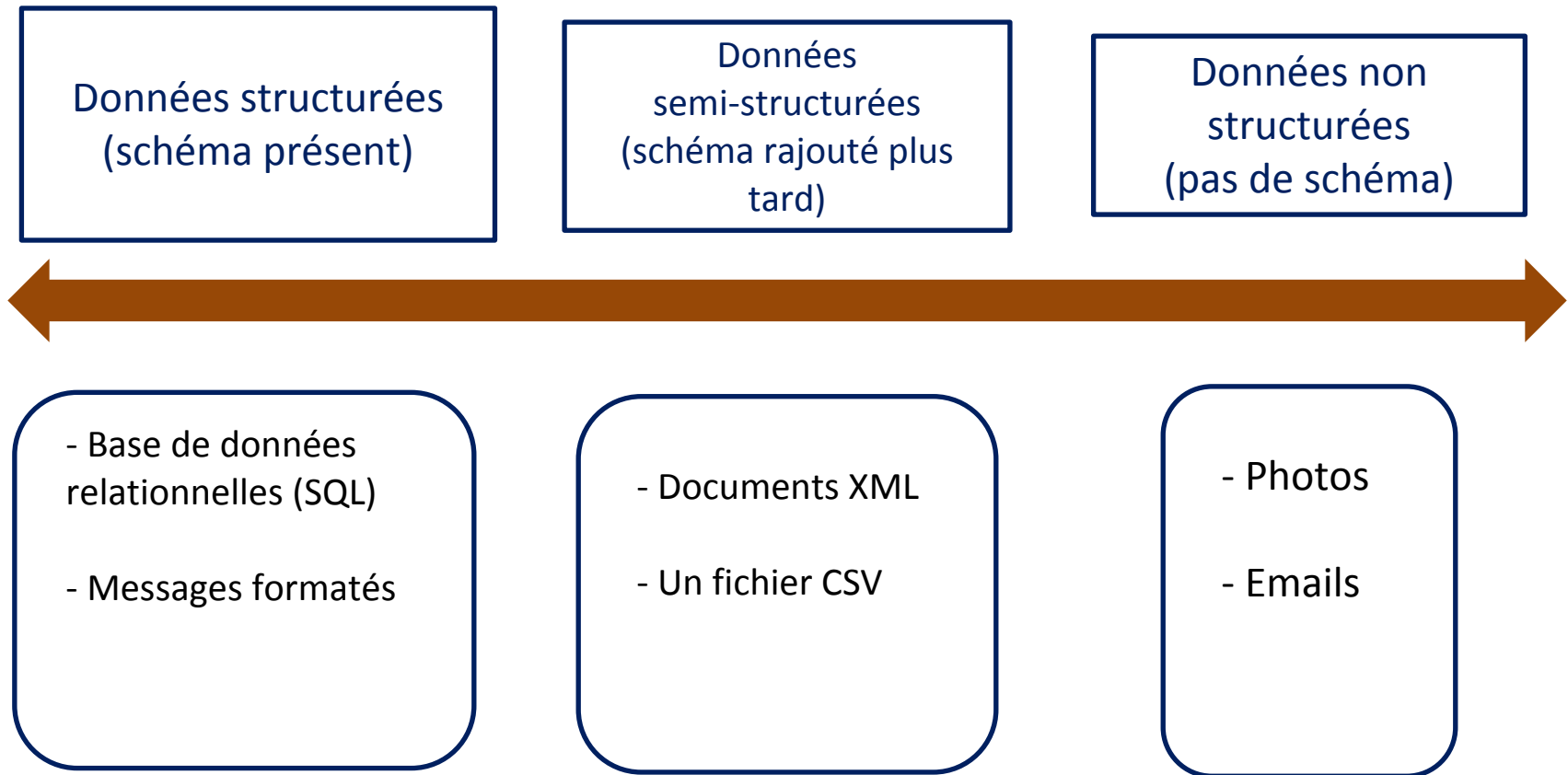
- La vitesse réfère à la capacité d'un système à pouvoir interagir suffisamment vite avec les données qu'il contient. Que ce soit:
 - Les données entrantes à traiter et à stocker.
 - Les requêtes clientes à traiter et à transmettre.
- Exemples:
 - Gérer l'ensemble de tweets qui arrivent chaque seconde sur les serveurs, et ce, en temps réel.
 - Voitures automatiques.
 - Construction d'un fil d'actualité temps réel sur facebook.

Définition des 5V : Vitesse

- **Batch:** L'ensemble des données est envoyé d'un seul coup au programme. La vitesse du programme importe peu.
- **Real Time / Interactive:** L'ensemble des données est envoyé d'un seul coup au programme. Cependant, la vitesse est primordiale pour pouvoir répondre au besoin.
- **Streaming:** Les données sont traitées dès leur arrivée dans le système. La vitesse est également primordiale pour le bon fonctionnement du programme.



Définition des 5V : Variété



Définition des 5V : Variété

- Les données sont structurées quand on connaît par avance le schéma.
- Chaque colonne possède un nom, un type et éventuellement des valeurs manquantes.

Nom	Code	Type de données	Longueur	Précision
cli activité	CLI_ACTIVITE	Caractère variable (256)	256	
cli adr1	CLI_ADR1	Caractère variable (25)	25	
cli adr2	CLI_ADR2	Caractère variable (25)	25	
cli ca	CLI_CA	Décimal (8,2)	8	2
cli comment	CLI_COMMENT	Caractère variable (100)	100	
cli cp	CLI_CP	Caractère (5)	5	
cli effectif	CLI_EFFECTIF	Entier		
cli fax	CLI_FAX	Caractère (20)	20	
cli id	CLI_ID	Entier		
cli lib court	CLI_LIB_COURT	Caractère variable (10)	10	
cli lib long	CLI_LIB_LONG	Caractère variable (50)	50	
cli rais soc	CLI_RAIS_SOC	Caractère variable (100)	100	
cli tel	CLI_TEL	Caractère (10)	10	
cli type	CLI_TYPE	Caractère (1)	1	
cli type id	CLI_TYPE_ID	Entier		
cli ville	CLI_VILLE	Caractère variable (10)	10	
col adr1	COL_ADR1	Caractère variable (25)	25	
col adr2	COL_ADR2	Caractère variable (25)	25	
col cp	COL_CP	Caractère (5)	5	
col deb contrat	COL_DEB_CONTRAT	Date		

Définition des 5V : Variété

- Quelque soit le format, les données semi-structurées peuvent être représentées sous forme tabulaire (lignes + colonnes).
- Chaque colonne possède un nom, un type et éventuellement des valeurs manquantes.
- Cependant, le schéma sera appliqué après chargement des données.

المحطة ونسبة موانق ونسبة غير موانق وعدد الدخايعين والأصوات الصريحة والأصوات الباطلة ونسبة الميثارة وموانق وغير موانق
070,341", "055,512", 9.32, "224,15", "125,853", "808,639,2", 0.40, 0.60, "الخليويونية"
675,497", "417,995", 6.34, "105,24", "092,493,1", "701,383,4", 3.33, 7.66, "الجزيرة"
327,280,1", "371,974", 8.34, "342,36", "698,254,2", "478,580,6", 8.56, 2.43, "القاهرة"
670,56", "839,307", 8.22, "743,6", "509,364", "713,629,1", 5.15, 5.84, "قنا"
566,113", "378,205", 2.37, "354,4", "944,318", "773,868", 6.35, 4.64, "دمياط"
825,382", "503,737", 0.32, "143,21", "328,120,1", "351,565,3", 2.34, 8.65, "الشرقية"
689,83", "248,466", 7.38, "054,13", "937,549", "278,454,1", 2.15, 8.84, "بني سويف"
841,40", "779,133", 0.26, "512,3", "620,174", "009,685", 4.23, 6.76, "الأقصر"
316,26", "116,44", 7.30, 841, "432,70", "388,232", 4.37, 6.62, "البحر الأحمر"
911,518", "219,631", 5.31, "013,21", "130,150,1", "758,719,3", 1.45, 9.54, "الدقهلية"
866,6", "157,12", 6.29, 328, "023,19", "407,65", 1.36, 9.63, "جندب سيناء"
393,6", "237,70", 5.36, 863, "630,76", "495,212", 3.8, 7.91, "مرسى مطروح"
975,76", "235,179", 4.36, "435,3", "210,256", "963,713", 0.30, 0.70, "الإسماعيلية"
999,380", "374,364", 0.34, "951,14", "373,745", "898,236,2", 1.51, 9.48, "المنوفية"
201,139", "506,442", 0.28, "176,14", "707,581", "688,127,2", 9.23, 1.76, "أسوط"
842,43", "061,104", 7.38, "880,1", "903,147", "522,387", 6.29, 4.70, "السيديس"
219,57", "890,486", 2.35, "441,12", "109,544", "694,579,1", 5.10, 5.89, "الفيوم"
390,155", "704,760", 5.34, "165,23", "094,916", "947,718,2", 0.17, 0.83, "المنيا"
939,5", "687,40", 9.32, 665, "626,46", "584,143", 7.12, 3.87, "الوادى الجديد"
578,81", "353,85", 0.38, "298,2", "931,166", "322,445", 9.48, 1.51, "بورسعيد"
517,126", "029,467", 4.25, "320,13", "546,593", "672,393,2", 3.21, 7.78, "سوهاج"
238,14", "726,50", 6.30, 949, "964,64", "618,215", 9.21, 1.78, "شمال سيناء"
009,512", "488,468", 9.33, "596,18", "497,980", "656,948,2", 2.52, 8.47, "الغربية"
716,529", "975,663", 2.36, "883,16", "691,193,1", "770,347,3", 4.44, 6.55, "الأسكندرية"
687,265", "755,818", 7.33, "505,18", "442,084,1", "930,276,3", 5.24, 5.75, "البحيرة"
560,187", "994,360", 6.29, "992,8", "554,548", "212,886,1", 2.34, 8.65, "نفس الشينخ"
396,45", "020,149", 7.22, "691,3", "416,194", "740,872", 3.23, 7.76, "أسيوط"
686,83", "795,160", 0.42, "926,1", "481,244", "491,586", 2.34, 8.65, "المصينون بالخير"
101,061,6", "911,693,10", 9.32, "395,303", "012,755,16", "866,918,51", 2.36, 8.63, "البحر الأحمر"

Définition des 5V : Variété

- Les données non structurées ne peuvent pas basiquement contenir de schéma.
- Ces données ne contiennent généralement « qu'une seule colonne » et il est alors nécessaire d'en extraire l'information pertinente.

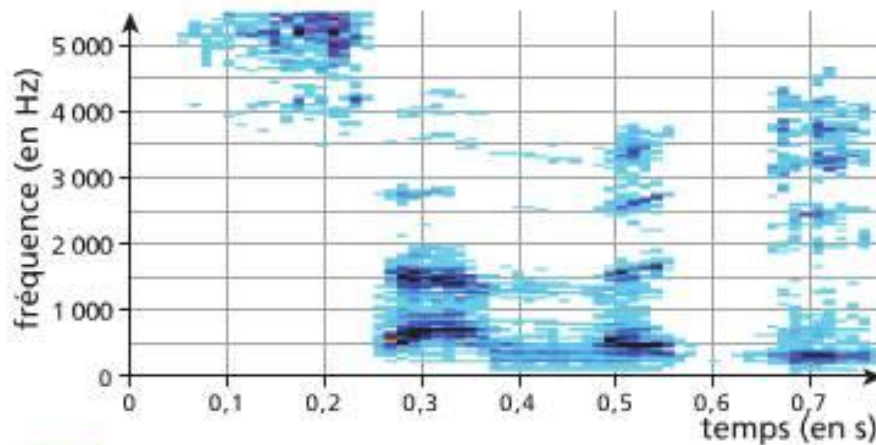
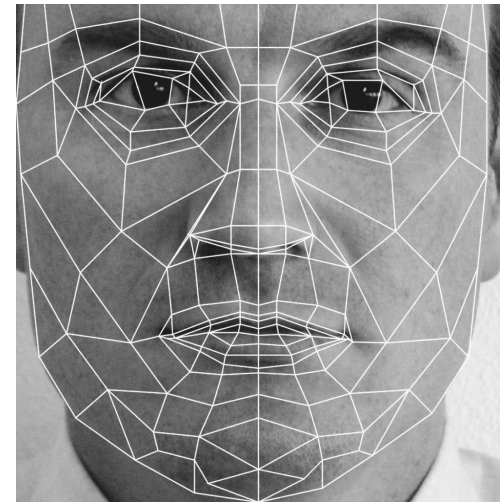


Fig. 1 Spectrogramme d'un mot.

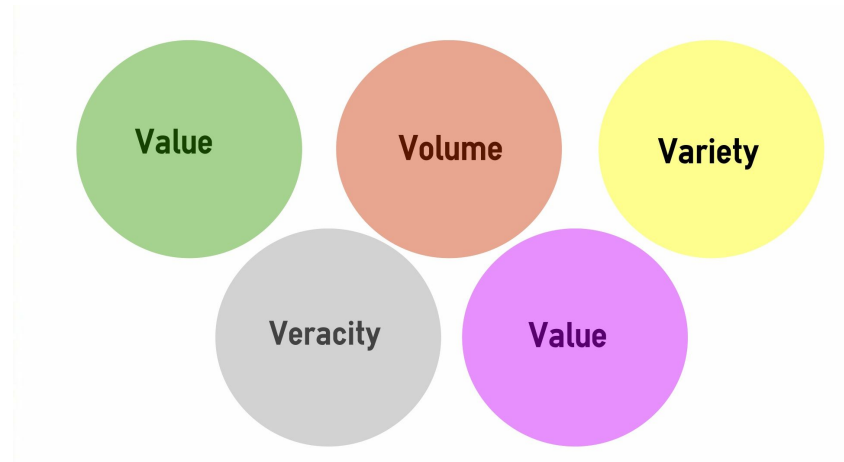


Définition des 5V : Véracité

- La véracité est un « V » qui a été rajouté récemment dans la définition d'une solution Big Data. En effet certaines données peuvent être:
 - trop anciennes pour s'appliquer à l'application métier.
 - Sans valeur supplémentaire pour l'application métier.
 - Données peu adaptées.
 - Données issues de formulaires.
 - Données imprécises (données météo, données type textes...)
 - Données externes peu sûres.

Définition des 5V : la Valeur

- On parle parfois d'un cinquième « V ». Cependant, le nom et la définition de celui-ci sont très différents d'une entreprise à l'autre.
- Une caractéristique pouvant être mise en avant est la Valeur. Celle-ci désigne la valeur contenue dans un jeu de données.



Définition des 5V

- Il est commun de définir un Big Data à partir des caractéristiques « V »
- Contenant originellement 3 composantes, les caractéristiques « V » voient parfois apparaître selon les sources et les années de nouveaux V.
- Les 3 V originaux restent cependant:
 - Le Volume
 - La vitesse
 - La variété







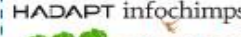
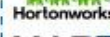





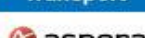
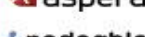





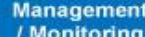

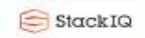









De projets à une solution Big Data

Open Source Project

Framework	Query / Data Flow	Data Access	Coordination / Workflow	Real-Time	Statistical Tools	Machine Learning	Cloud Deployment	Messaging Framework
 HDFS Apache Hadoop YARN Spark	 HIVE	 Cassandra  CouchDB  SciDB  mongoDB	 ZooKeeper  talend  Oozie	Storm	 SciPy  R	 Mahout  Spark	 Hadoop	 kafka

Infrastructure

NoSQL Databases	Hadoop Related
 DATASTAX  Neo4j  HYPERTABLE  Couchbase  Cloudant  basho	 infochimps  cloudera  Zettaset  IBM
NewSQL Databases	Collection / Transport
 VoltDB  memsql  <small>DRAWNSMSCALE</small>	 aspera  nodeable
MPP Databases	Management / Monitoring
 VERTICA  kognitio  PARACCEL  GREENPLUM <small>A DIVISION OF EMC</small>	 oceanSync  StackIQ  bunday  DATADOG
Storage	Crowdsourcing
 Cleversafe  panasas  nimble storage	 CROWDCOMPUTING SYSTEMS CrowdFlower  amazon
Security	
	 Stormpath  IMPERA codefortytwo software DATAGUISE

Analytics

Analytics Solutions	Data Visualization
 Palantir platforma  PERVASIVE Datameer  PRECOG DIGITAL REASONING	 Quid ACTUATE  Kitenga visual.ly  centrifuge metaLayer
Statistical Computing	Social Media
 sas REVOLUTION ANALYTICS  SKYTREE pik	 bit.ly tracx  bluefin Dataminr
Sentiment Analysis	Analytics Services
 GENERAL SENTIMENT crimson hexagon	 THINK BIG accenture McKinsey&Company
Big Data Search	IT Analytics
 elasticsearch Autonomy	 splunk  sumologic
Location / People / Events	Real-Time
 RapLeaf FlipTop  Recorded Future PlaceIQ	 CONTINUITY  feedzai
Crowdsourced Analytics	SMB Analytics
 DataKind kaggle	 sumAll custora

Applications

Ad Optimization			
 DataXu Data. Insight. Action.	 aggregate knowledge	 m6d	 ai Match ad intelligence
 thetradedesk	 bluekai	 rocketfuel	
 TURN	 33 across	 MediaMath	
Publisher Tools		Application Service Provider	
 VISUAL REVENUE	 Yieldex REVENUE MANAGER	 collective iT	
 yieldbot			
Industry Application		Marketing	
 NEXT BIG SOUND	 KNEWTON	 LATTICE ENGINES	
 BILLS GUARD	 Bloomberg	 Sailthru	
 numberFire	 Climate Solutions	 bloomreach ad revenue	
 zest cash		 CLICKFOX	

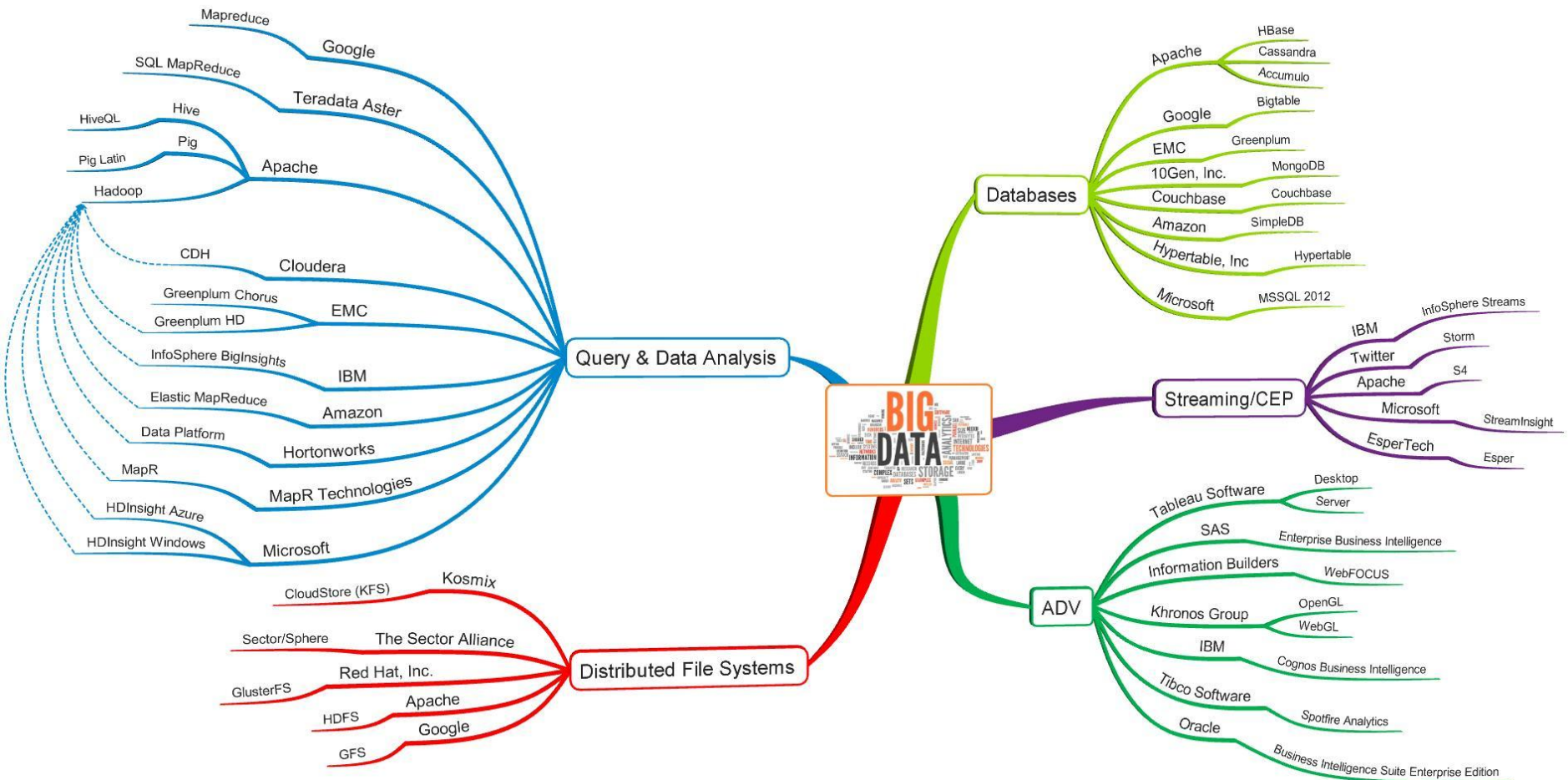
Data Sources

Data Marketplaces	Data Sources
 DataMarket  Windows Azure  factual	 DATA SIFT  Gnip  knoema  infochimps
Ad Optimization	
 Withings  RunKeeper JAWBONE 	

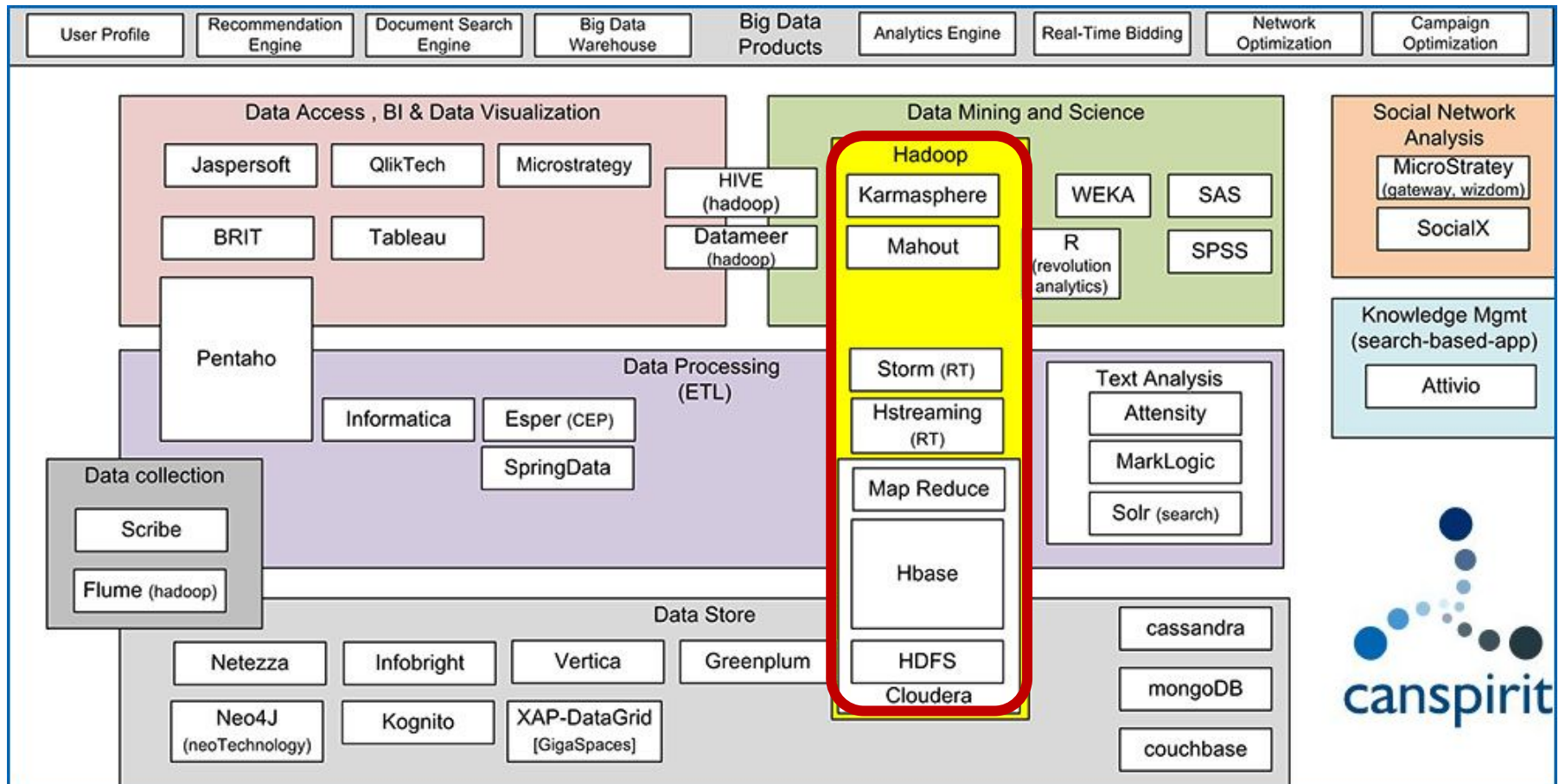
De projets à une solution Big Data

- Systèmes de fichiers distribués.
- Des outils de traitements distribués (Data Processing).
- Des outils de gestion de base de données.
- Des outils temps réel.
- Des outils d'exploitation de données (data Mining, machine learning).
- Des outils de visualisation et de restitution.
- Des outils d'ingestion de données.

De projets à une solution Big Data

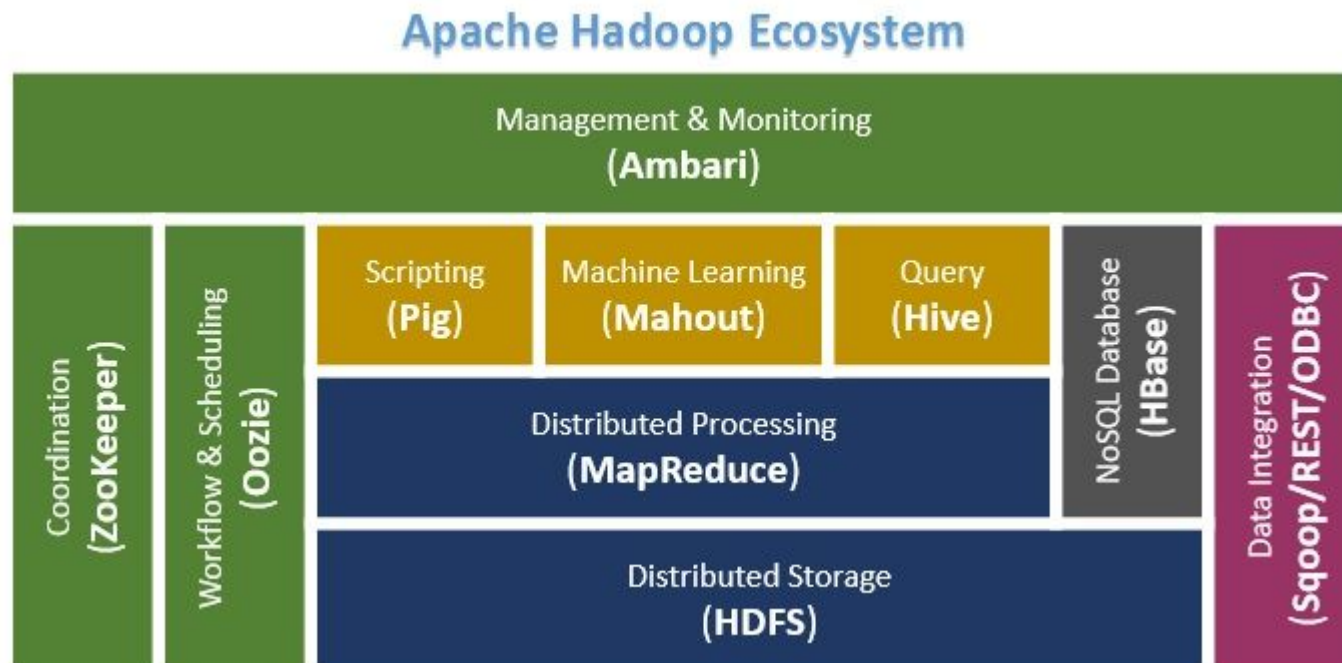


La solution Hadoop



La solution Hadoop

- Hadoop est un framework (un ensemble d'outils et de logiciels) spécialement désigné pour être une solution Big Data.
- Il contient un ensemble de briques logiciels, qui, assemblées les unes avec les autres permettent de répondre à des problématiques Big Data.



Un exemple de projet Big Data

La géolocalisation SFR

La géolocalisation SFR

1 Milliard
d'événements par
jours en France
métropolitaine

SFR



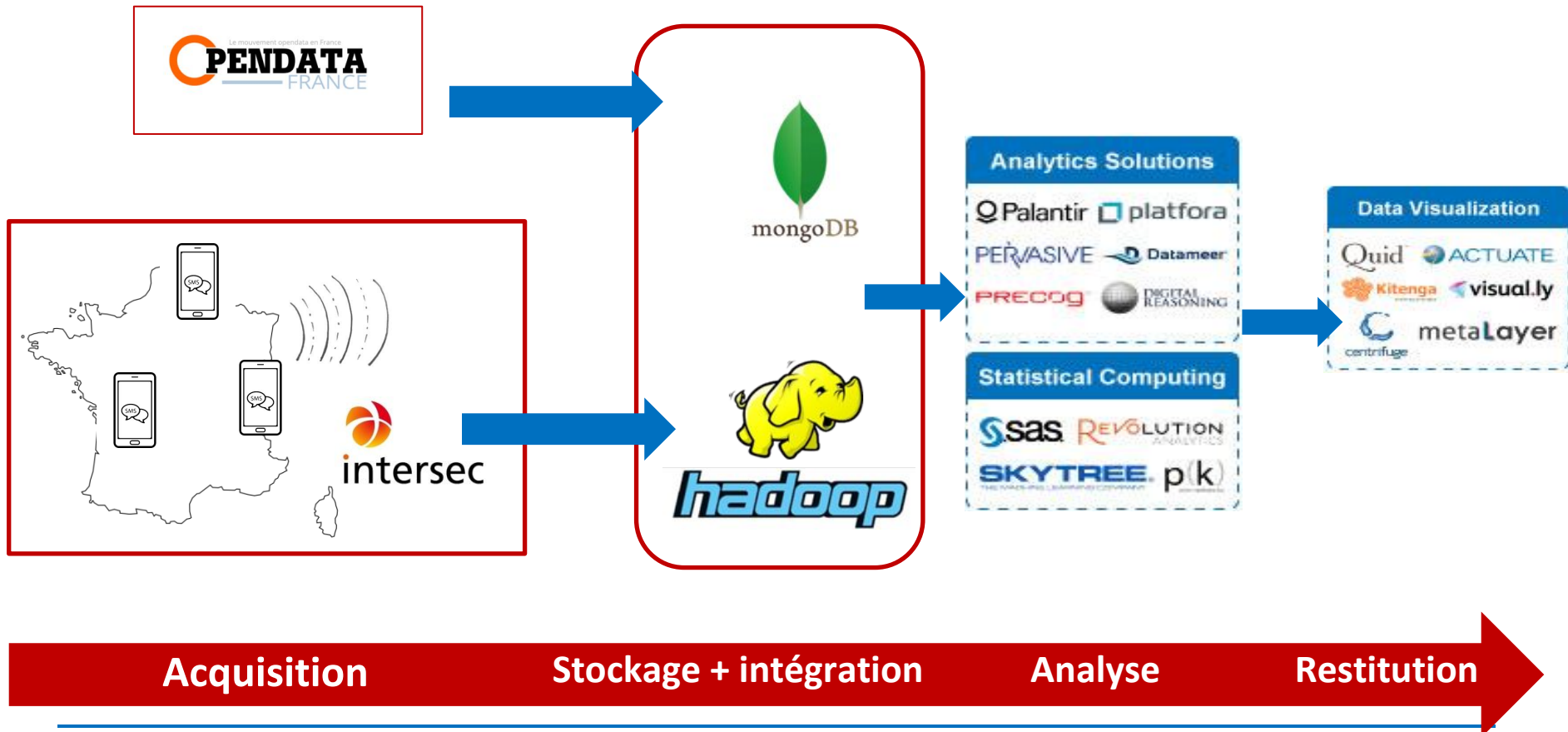
2011 - 2013



But :

- Meilleure collecte des informations de géolocalisation des utilisateurs.
- Analyser la fréquentation de lieux publics et flux de population.
- En déduire des indicateurs marketing en temps réel.

Infrastructure de la solution



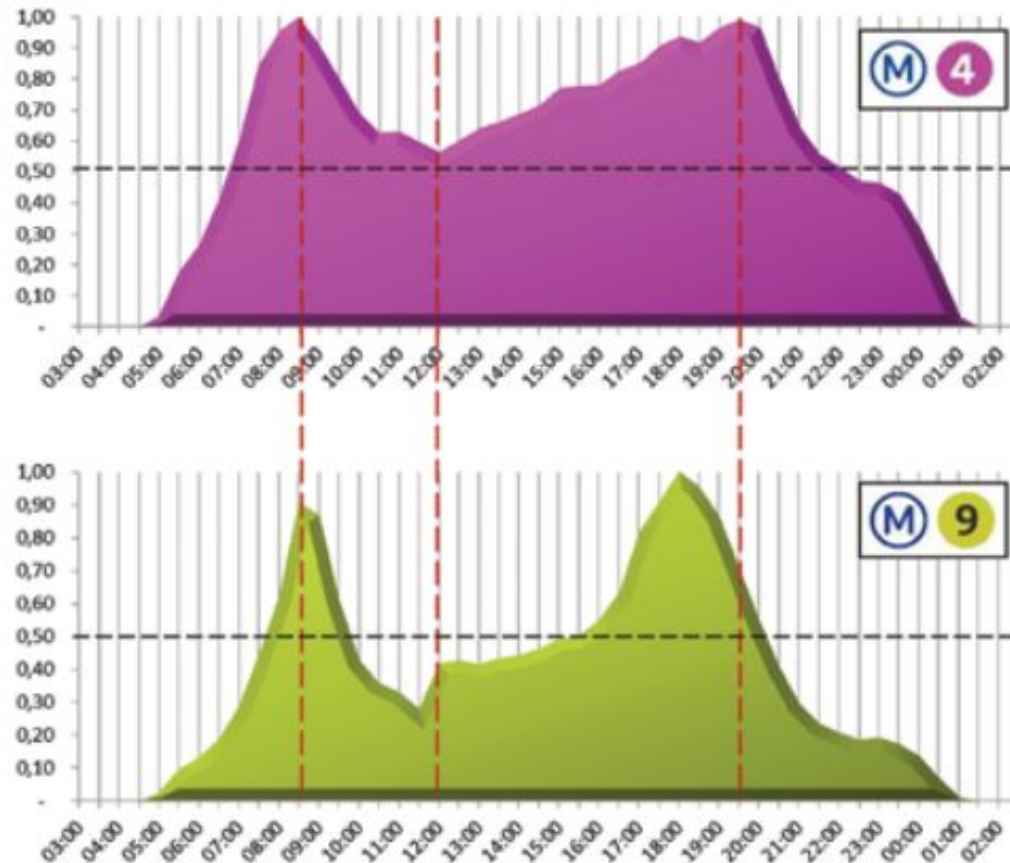
Résultats



Résultats

- Comportements presque identiques jusqu'au premier pic de **8h30**
- La fréquentation de la ligne 4 reste au-dessus de l'indice 0,5 entre **7h** et **23h**
→ Pics beaucoup plus marqués sur la ligne 9
- Pic du soir à **18h** (ligne 9) contre **19h30** (ligne 4)
- Sur la ligne 4, la fréquentation est encore importante jusqu'à **23h**

Comparaison des indices de présence (lundi-jeudi)



Les métiers de la Data

Les métiers de la Data
Process Data chez Natixis

Les métiers de la Data

- Il existe plusieurs plusieurs métiers autour de la data. A l'heure actuelle, la différenciation entre ces métiers et la nomenclature associée varie encore d'une entreprise à l'autre mais tend vers une convergence.
- Les principaux métiers autour de l'univers de la data sont :
 - Le Data Analyst / Expert Data Viz'.
 - Le Data Scientist.
 - Le Data Engineer.
 - Le Data Architect.
 - Le Chief Data Officer (CDO).
- Une bonne organisation de ces métiers et la connaissance de cette nomenclature peut être un gage de qualité sur les projets et en tant que process d'entreprise.

Les métiers de la Data : Data Analyst



Le rôle du Data Analyst est d'explorer et d'exploiter des données, et ce, afin de les valoriser au travers de plusieurs KPI métiers. Il est également chargé de la restitution des résultats auprès du métier, notamment par l'intermédiaire d'applications de Data Visualisation.

Les tâches réalisées par un Data Analyst sont souvent :

- Exploration, nettoyage, analyse et exploitation de données.
- Création, valorisation et restitution de KPIs métier.
- Création d'applications de visualisation (reporting, dashboarding...).

Compétences :

- Python, R, SAS.
- Data Visualisation (Qlikview, Qlik Sense, Tableau...).
- Connaissances métiers.

Les métiers de la Data : Data Scientist

Le rôle du Data Scientist est de valoriser la donnée de façon avancée grâce à des outils mathématiques, statistiques, algorithmiques et d'analyse prédictive afin de répondre à un besoin métier.



Les tâches réalisées par un Data Scientist sont souvent :

- la réalisation d'algorithmes prédictifs en vue d'appréhender et/ou de prédire un phénomène déterministe.
- Segmentation et clustering de populations diverses (clients, produits, observations...).
- Traitement et analyse de données non structurées (Texte, Images, Vidéos, bande-sons...)

Compétences :

- Python, R.
- Mathématiques, statistiques.
- Analyse descriptive et prédictive.
- Machine learning, Deep learning, intelligence artificielle.

Les métiers de la Data : Data Engineer



Le rôle du Data Engineer est de préparer, en amont, la donnée afin de la rendre accessible et utilisable par les Data analysts et Data Scientists. Lorsque l'on parle de métiers dit "Big Data", on réfère bien souvent, en premier lieu à un travail de Data Engineer.

Les tâches réalisées par un Data Engineer sont souvent :

- la réalisation de pipelines informatiques afin d'acheminer des données brutes jusqu'à un espace de données centralisé (solution Big Data, base de données...).
- Intégration, agrégation et structuration des données.
- Rendre disponible la donnée aux autres métiers Data aux travers de différents clients et API.

Compétences :

- Java, Scala, C++.
- Maîtrise des outils de la stack Hadoop.
- Base de données, algèbre relationnel.

Les métiers de la Data : Data Architect



Le rôle du Data Architect est de concevoir et administrer une ou des solutions Big Data afin de pouvoir rendre possible la collecte, l'intégration, l'exploitation et la restitution des données aux utilisateurs. Les solutions qu'ils déploient peuvent être de différents type, allant d'une solution Hadoop native à des solutions dans le cloud comme AWS service ou Microsoft Azure.

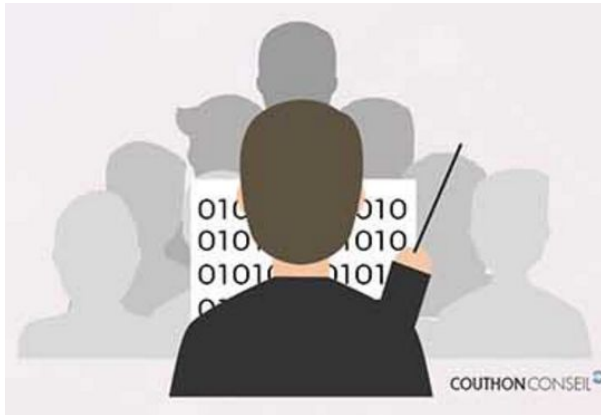
Les tâches réalisées par un Data Architect sont souvent :

- le déploiement et la maintenance d'infrastructures Big Data (clusters Hadoop...).
- Support auprès des utilisateurs.
- Le management des données et la mise en place de stratégies relatives à l'ingestion, le stockage, l'exploitation et la restitution des données.

Compétences :

- Administration de systèmes d'exploitation.
- Maîtrise des outils de la stack Hadoop, frameworks Big Data.
- Base de données, architectures de données.

Les métiers de la Data : Chief Data Officer



Le rôle du Chief Data Officer est d'assurer la gouvernance des données et leur valorisation pour répondre aux enjeux décisionnels de l'entreprise.

Les tâches réalisées par un Chief Data Officer sont souvent :

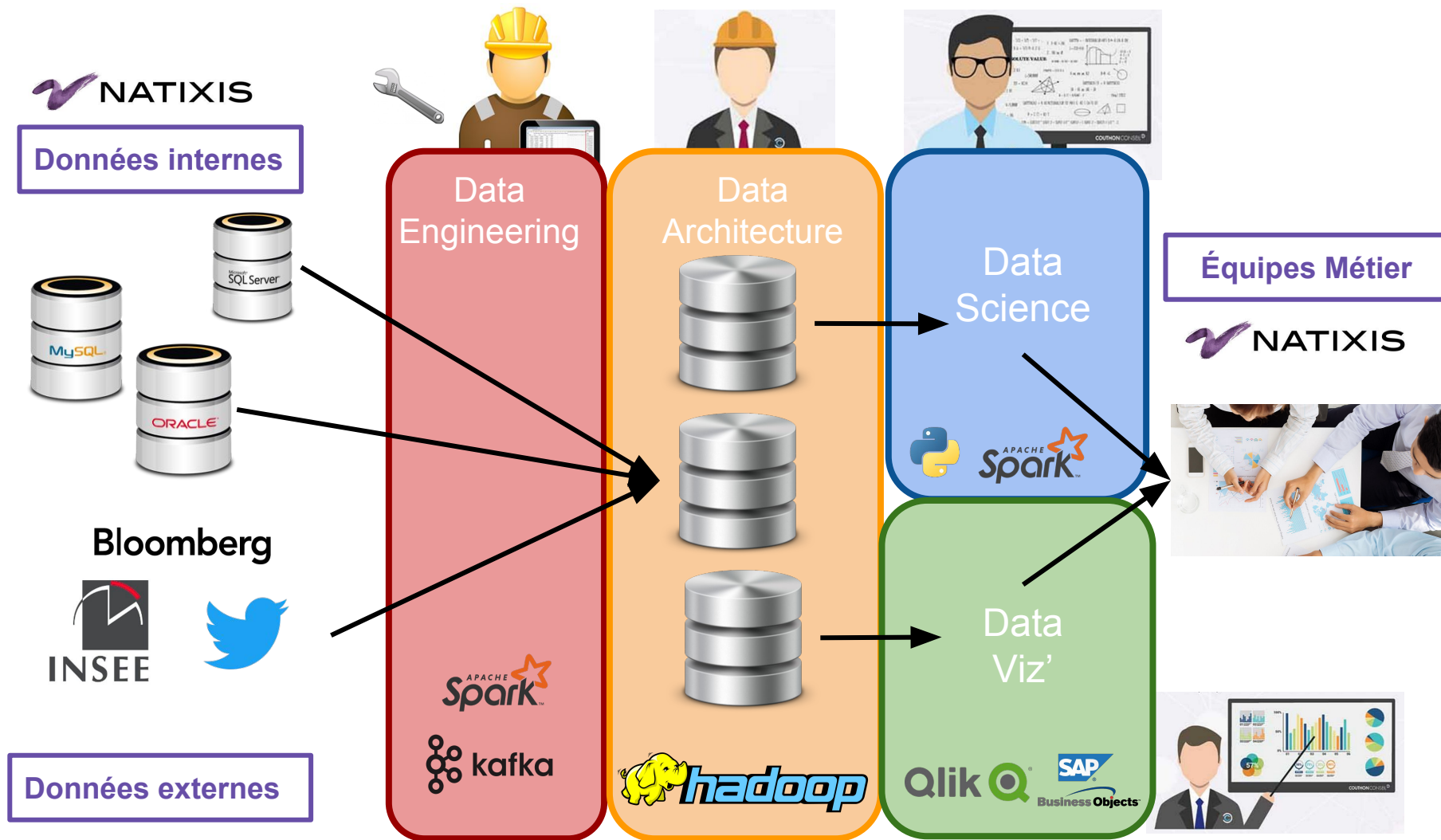
- l'acquisition centralisée de données auprès de partenaires, fournisseurs ou clients de l'entreprise au regard de la stratégie de l'entreprise.
- Définir et assurer la gouvernance de la donnée.

Compétences :

- Management, Leadership.
- Connaissances métier, stratégie d'entreprise.

Note : Le cabinet Gartner, spécialisé en connaissance IT, estime qu'en 2019, 90% des entreprises auront embauché un Chief Data Officer.

Processus Data chez Natixis



Conclusion

- Le Big data est un ensemble de processus. Il consiste en l'acquisition, le stockage, le nettoyage, l'intégration, l'analyse et la visualisation de grands volumes de données.
- Le Big Data est favorisé par la croissance exponentielle des données, des innovations technologiques, comportementales et business.
- Un Big Data se définit à travers les caractéristiques des « 5V », Volume, Vitesse, Variété, Véracité et Valeur.
- Une solution Big Data est un ensemble de projets qui, utilisés ensemble, permettent de traiter une problématique Big Data au regard des caractéristiques des 5V.
- Hadoop est un framework open-source spécialement conçu pour traiter des problématiques Big Data.