BIG data opportunités & stratégies

Sommaire

- Introduction : généralités
- Nouveaux paradigmes & nouveaux métiers
- Stats / IA / machine learning
- Utilisations pratiques
- Question de droit : sécurité & vie privée
- Stratégies

BIG comment?

• 1995 : 30 Gb

● 1997: 2 Tb (x60)

2014

	Bande passante / jour	Stockage
Facebook	600 Tb	300 Pb
Google	100.000 Tb	15,000 Pb
Twitter	100 Tb	?
NSA	29.000 Tb	10.000 Pb

```
1 Tb (Terabytes) = 1000 Gb
1 Pb (Petabytes) = 1000 Tb
```

Evolution des usages

- Réseau sociaux : Crowd sourcing
- Smartphones / IoT : senseurs automatiques
- Massification d'internet :
 - **3.4 milliards** (40 % vs 1% en 1995)

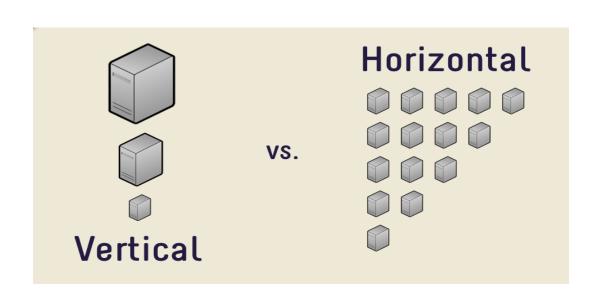
Evolution des technologies

- Prix du stockage de données
- Centralisation d'internet (GAFAM)
- Virtualisation des systèmes :
 Cloud // laaS (infrastructure as a service)
- Technologies de stockage / calcul distribué
- APIs (Wikipedia, Google, Facebook)
- Horizontal scaling

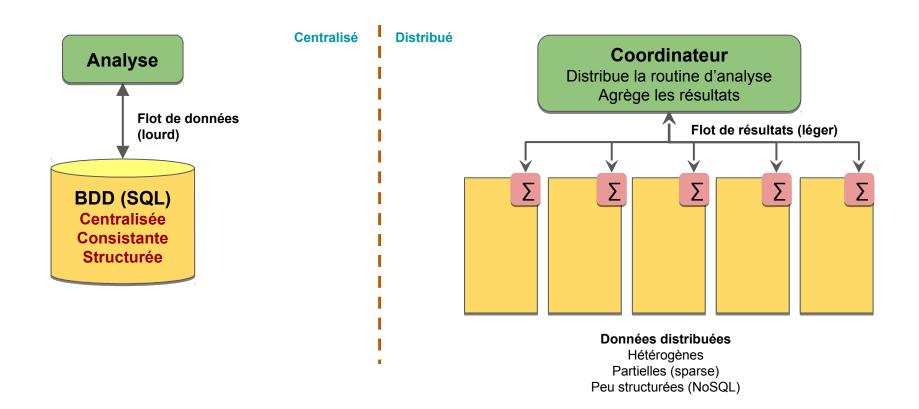
Nouveaux paradigmes: horizontal scaling

Scaling horizontal

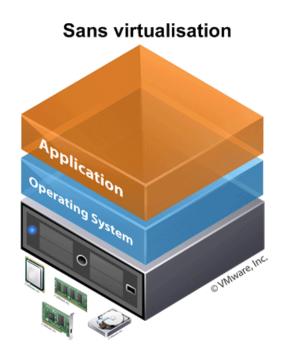
- économique
- sans coupure
- sans limite de taille



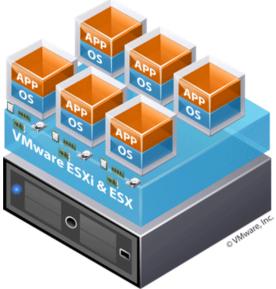
Nouveaux paradigmes : BDD



Nouveaux paradigmes : Virtualisation







Nouveaux paradigmes : Virtualisation

- Isolation des VMs ⇒ Sécurité
- Décorrélation physique / logique
- Adaptation à la charge (paiement ressource * temps)
- Provision rapide (qq jours ⇒ qq minutes)
- Sauvegarde, clonage, migration facile
- Automatisation de l'administration : laaS

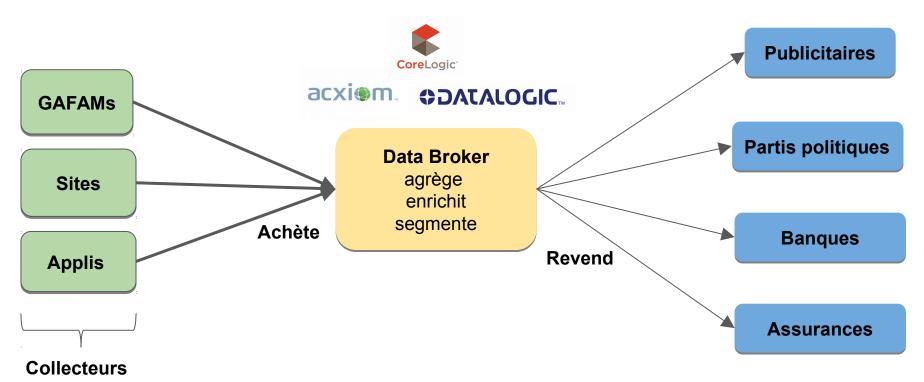
Big Data Landscape 2016 (Version 2.0) Infrastructure Analytics Applications Data Science Hadoop Hadoop in Analyst Analytics Sales & Marketing Customer Service Human Visualization Spark Cluster Services Legal On-Premise the Cloud Platforms Platforms Platforms Capital RADIUS Gainsight MEDALLIA amazon ++ableau ccontext relevant cloudera and the amazen Microsoft A databricks Q Palantir Microsoft Sbloomreach Zeta ğıld RAVEL CONTINUUM & DataRobot ATTENJITY -(b) kuberseti W EVERSTRING livefyre JUDICATA gunvus MAPR Pivotal AYASDI Alpine III. Qlik @ looker CLARABRIDG GridGain 4 docker blueyonder Lattice M CLICKFOX IBM InfoSphere Connectifi Quid enigma Datameer Roambi MODE plotly ⊕kahuna √infer SAILTHRU Everlav STELLAService MESOSPHERE AZENA TREAS #dataiku @tonian textic IBM InfoSphere SECHOLO TO THE PERSON TACHYON Core OS popperdata Digital Reasoning NG@DATA' 🚅 Presc Bottlenose. persado AVISO ÓSENSE C DOMINO Sense ②Brevia entelo i bluedata jethro 🙆 altiscale 🛄 bole Stack IQ CRRITAL INSIGHT inter ana ўhat ______ ∧ысовтнены S QUANTIFIND ACTIONIC CHARTIO hi📮 PRIMW MITION fuse(machines #NGAGIO appuri Wisele NewSQL Databases BI Platforms Statistical Log Analytics Social NoSQL Databases Ad Optimization Security Vertical AI Computing Analytics Omamobe Google Cloud Platfor Clustrix Pivotal Power BI wamazon splunk> Hootsuite Applications # CYLANCE AppNexus **S**sas **ÖRACLE** criteol. MedaMan # paradigm4 sumologic CounterTack cybereason Microsoft Azure MarkLogic NETBASE memsal Threat Metrix. kıbana **■ DATASIFT** DATASTAX ₯ birst SPSS OpenX procketfuel ■mongoDB 🚰 GoodDets Sentine/One tracx bitly 10 Integral the TradeDesk Clara **≪EROSPIKE** Couchbase Recorded Future platfora ▲ MATLAB **synthesio** Sequoia DB redislabs @ Influxdata deepdb Trafedien Cockreach Lans loggly KASIST@ (V) atscale **IFORT**SCALE **⁵**siftscience dstillery DataXU Oppier TAPAD ⊕ Kuybass feedzai ♥SICNIFYD MPP Data Graph Speech & NLP Horizontal Al Cloud EDW Real-Time Machine Learning Transformation Integration Databases Databases NarrativeScience 🕖 Watson amazon Finance Publisher Govt / Regulation informatica amazon ARRIA ERADATA sentient altervx ⊿ffirm III Lendina Club @neo4j Tools METAMARKETS H₂O. C Socrata Put potential to work VERTICA. √ vicarious Microsoft Azun talend Outbrain OnDeck> "Kreditech Striim MuleSoft Dato > NINETEZZA Pivotal noro 🕝 👭 Numenta OPENGOV TRIFACTA oo 👛 apilal Tab90la 🚌 finance Lendüp 💔 Kabbage snap**Logic** confluen G cortical in (Oction) tamr 🐅 🤈 quantcast FN FiscalNote tidemark (2000) INSIKT Bedrock Data Claepsonge to VISENZ 🌺 🏭 clarifai EMINTORRENT Alle Mindfield OrientDB kognitio StreamSets Chartbeat EXASOL Ordremic dataArtisans IDIBON (8) PREDPOL" Z UOra : Dataminr : 100 Lenddo InfiniteGraph Alation A Infoworks xplenty jieldbot کے mark43 **iSENTIUM** Management Security Storage Crowd-Search Data For Business Web / Mobile App Dev "1# OpenDataSoft Quantopian sentient Yieldmo M Asterony ORACLE Services / Monitoring **TANIUM** sourcing Analysts / Commerce amazon apigee OPERA Google Analytics New Relic. 25 EXALERO Nigami..ogic Life Sciences Industries Goegle Do mages mechanicalturi Education/ © Counsyl CODE42 mixpanel Microsoft Azu Lucidworks amazon octifio OP@WER @Harmony* ClearStory Learning DataGravity panasas/ DATASCIONO RJMetrics @ BLUECOR 🏰 elastic 🖸 ThoughtSpot Numerify KNEWTON RetailNext splunk> X Recombine **S** Typesafe A AMPLITUDE (1) granify CrowdFlowe CIRRO **S**VECTR∧ STITCH FIX . M∧∧N∧ Ø swiftype Clever KYRUUS FLATIRON соно SUMAII Airtable o⊕s⊛zymergen HealthTap® DIMESSIVES Ýrocona ^{sene} Qumulo import (i) retention custora TACHYUS Seeq FarmLogs @eclara METABIOTA ZEPHYR Cross-Infrastructure/Analytics PANORAMA Gingerio * transcriptic Glow #HowGood celect @#22Time @enlitic @AiCure 🗘 🗅 know 🖼 statmuse B⊕XEVER Tamazon Google !! Microsoft TRIM SAP SSAS III (m) Arrives VERTICA VITAVORACI | TERADATA ORACLE II Netapp Open Source Framework Query / Data Flow Data Access Coordination Real-Time Stat Tools Machine Learning 16% Search Security accumulo HBASE mongoDB Apache SINGA MALlib. A STORM Spark R Apache Ranger talend DOGGE Aerosolve cassandra YARN 🍌 🃸 MESOS Scala Apache Zookeeper Caffe SciDB & kafka CNTK Visualization Flink SLAMDATA DELL Solr Spark TEZ VELES WEKA 🈻 TACHYON 🔵 druid CouchDB ::rick :: CouchDB Apacho Ambori Flink @CDAP DIMSUM Lucene Data Sources & APIs Incubators & Schools Health IOT Financial & Economic Data Air / Space / Sea Location / People / Entities Other DataCamp Bloomberg D | DOW JONES acxiem. Experian JAWBONE GARMIN △ spire ThingWerx "INSIGHT, YEDLEE PREMISE CAPITAL IQ GARMIN CONSTITUTE SInsideView @esri * WINDWARD 🙀 practice fusion : 🖷 fit bit 🕖 helium 🔯 samsara A DataElite STREETLINE CARTECO Factual. Place panjiva Withings VALIDIC netatm quandl xignite CBINSOHTS Crimson Hexagon D placemeter BASIS TySense VOQ.ATA.GOV The Data Incubator kinsa (/) Human API mottermark @estimize @ PLAND Airware ToponeOoploy Last Updated 2/12/2016 Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap) FIRSTMARK

Nouveaux acteurs

- Infrastructure / Cloud : Google, Amazon, Microsoft, OVH
- BDD distribuées, NoSQL : CouchDB, MongoDB, Hive
- Framework : Hadoop, Apache Spark
- Ecosystème fourni d'outils et d'applications
- Intégration complexe
- Domaine évoluant très vite

⇒ **Distributions** : Cloudera, Mapr

Nouveaux acteurs : Les Data Brokers

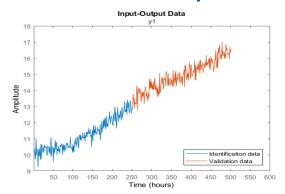


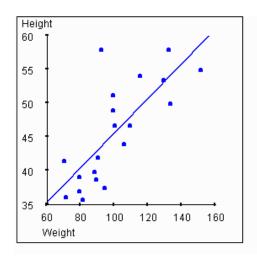
Nouveaux métiers

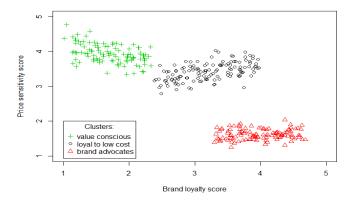
- Nouveaux langages : Scala (calcul parallèle, fonctionnel)
- Spécialisations en informatique
 - O Data analyst: maths, stats
 - Architecte de données
 - IA / Machine learning :
 "psychologues" de réseaux de neurones.

Big data & Statistiques

- Analyse, métriques
- Prédictions
- Corrélations, régression ⇒ Fonction
- Segmentation
- Périodicité, prévision temporelle

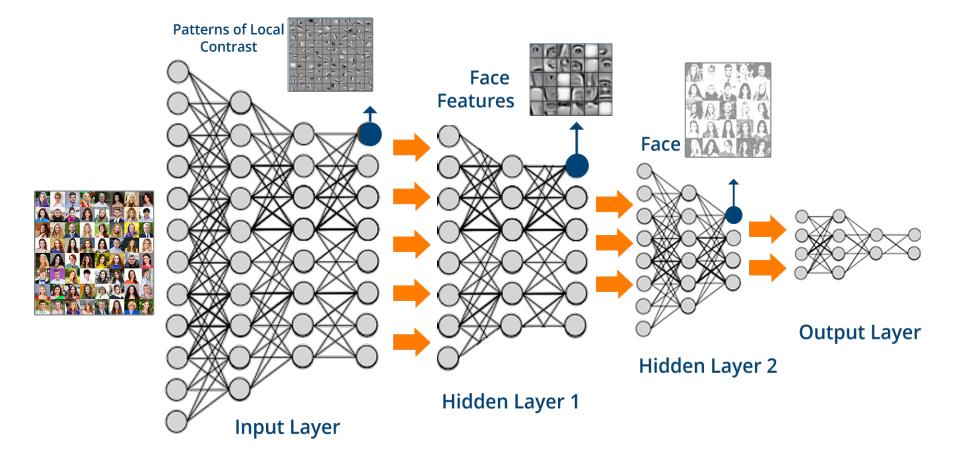




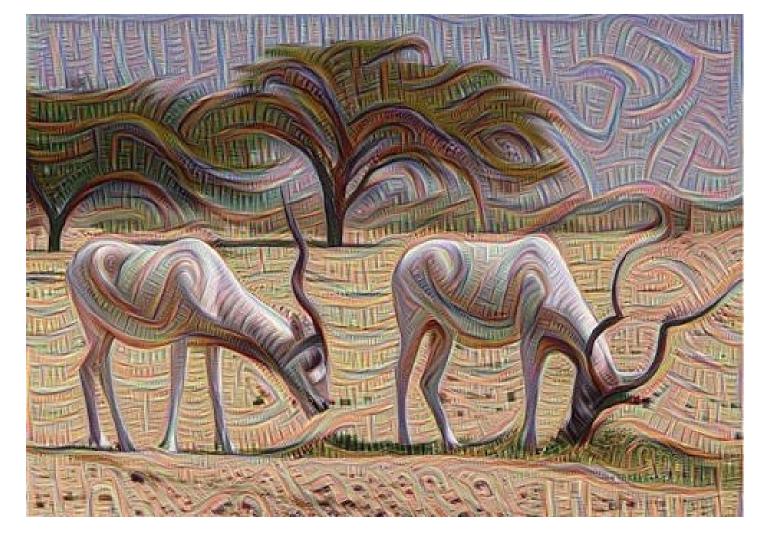


Big data: Deep learning

- Machine learning :
 apprentissage sans programmation explicite
- Le programme s'améliore avec l'expérience
- Deep learning = imitation du cerveau
- Réseau de neurones artificiels
- Essai / erreur ⇒ renforcement







Big data: Deep learning

- Apprentissage : nécessite énormément de données
- BIG DATA ⇒ progrès fulgurant des IA :
 - Classification d'images
 - Reconnaissance vocale
 - Reconnaissance faciale
 - Conduite automatique
 - Traduction automatique

Utilisations

- Trafic temps réel / bouchon (Waze)
 Smart cities ⇒ optimisation des flux :
 Circulation: lignes de bus.
- Smart GRIDS : réseau électrique:
 - Prédiction / pilotage de la consommation (Linky)
 - Prédiction de la production (ENRs)
 - ⇒ Pilotage de stockage & vente sur le marché D+1

Utilisations

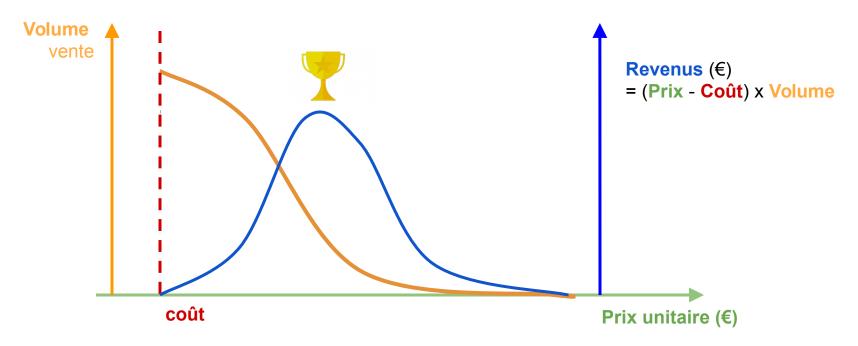
- Détection épidémies (google trends) / catastrophes
 (Twitter)
- Gestion des communs : Réseau hydraulique (fuites, suivi perso, quotas, pénurie ...)
- Médecine : aide au diagnostique
- Droit : aide aux juristes, recherche de jurisprudence

Utilisations: marketing

- Vers une personnalisation des offres
- Fidéliser le client : promos ciblées, rappels
- Provoquer l'achat : "Vous aimerez aussi"
- Prédiction de la demande :
 Gestion des stocks, demande périodique (WE,
 vacances), dépendant de la météo (Cinema, Restaurant)
- Etudes de tendances du marché :
 Ouvertures de lignes aériennes, création de produits

Utilisations: A/B testing

La courbe de prix : Le Graal du commercial



Utilisations: A/B testing

- Variation aléatoire du prix (ou % promo, pub, design, ..)
 sur un échantillon de visiteurs (qq %)
- Impact négligeable sur les ventes pendant la campagne
- Puis statistiques ⇒ courbe de prix
- Bonus : segmentarisée !
- ⇒ Action : changement du prix, voire variable par segment
- ⇒ Maximisation des revenus

Question de droit : sécurité & vie privée

- Les données critiques :
 - Données client ⇒ Réputation en cas de vol
 - Secret industriel :
 Innovation ? △ NSA / GAFAM
 - Secret défense / légal / admin FR ⇒ Sol Français
- Droit, espionnage ⇒ préférer UE / France (OVH)

Question de droit : données personnelles

- Données personnelles : définition (FR / UE)
 - O Identifiants uniques (nom, tel, sécurité sociales, ..)
 - ou identification par recoupement
- Obligations :
 - FR: Déclaration CNIL + accès, modif, suppression
 - + UE [RGPD] Mai 2018 :
 - Collecte limitée, explicite, légitime ⇔ finalités
 Cf droits applis sous android
 - Expiration

Stratégies : Open Source

- Open Source : Logiciel avec accès au code source + droit de modification et redistribution
- Big DATA: Beaucoup d'outils open source :
 - Suite Apache Spark
 - Apache Impala
 - Elasticsearch
- Gratuits, puissants, maintenus + communauté
- Open Source != bénévole :
 - support, formation, consulting

Stratégies : résumé

	Outils ⇒ Données	Données ⇒ Cloud
Vie privée	Déclaration CNIL	 Déclaration CNIL + Anonymiser si possible + Préférer Cloud privé (+ cher)
Sécurité		Souvent bonne politique sécurité
Outils d'analyse	Open source Installation spécifique Adaptation au business	 Open source ou logiciel "maison" ⇒ Pas de communauté
Coûts / travail	Consultingou compétences internes	 Licence / abonnement Cloud Préparation des données Transfert Config

À emporter



- Secteur porteur : marchés & métiers
- Technologies jeunes / mouvantes
- Données = mines d'or
- Analyse = extraction de valeur
- Gardez-en le contrôle !

Questions?



Des artisans développeurs, au service du bien commun.

http://atelier-des-communs.fr/
raphael@atelier-des-communs.fr