BIG DATA

Analyse et valorisation de masses de données

SMILE

I.T IS OPEN

I - PREAMBULE

I.1 SMILE

Smile est une société d'ingénieurs experts dans la mise en œuvre de solutions open source et l'intégration de systèmes appuyés sur l'open source. Smile est membre du CNLL, le Conseil National du Logiciel Libre, association d'associations pour la promotion et la défense du logiciel libre.

Smile compte 1200 collaborateurs dans le monde, ce qui en fait la première société en France et en Europe spécialisée dans l'open source.

Depuis 2000, Smile mène une action active de veille technologique qui lui permet de découvrir les produits les plus prometteurs de l'open source, de les qualifier, de les évaluer, puis de les déployer, de manière à proposer à ses clients les produits les plus aboutis, les plus robustes et les plus pérennes.

Cette démarche a donné lieu à toute une gamme de livres blancs couvrant différents domaines d'application. La gestion de contenus, les portails, le décisionnel, les frameworks PHP, la virtualisation, la Gestion Electronique de Documents, les ERP, le big data …

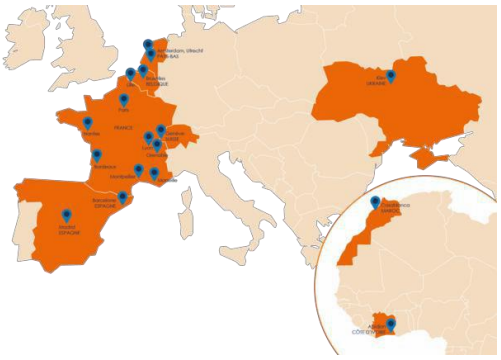
Chacun de ces ouvrages présente une sélection des meilleures solutions open source dans le domaine considéré, leurs qualités respectives, ainsi que des retours d'expérience opérationnels.

Au fur et à mesure que des solutions open source solides gagnent de nouveaux domaines, Smile est et sera présent pour proposer à ses clients d'en bénéficier sans risque.

Smile apparaît dans le paysage informatique français et européen comme le prestataire intégrateur de choix pour accompagner les plus grandes entreprises dans l'adoption des meilleures solutions open source.

Ces dernières années, Smile a également étendu la gamme des services proposés. Depuis 2005, un département consulting accompagne nos clients, tant dans les phases d'avant-projet, en recherche de solutions, qu'en accompagnement de projet. Depuis 2000, Smile dispose d'une Agence Interactive, proposant outre la création graphique, une expertise e-marketing, éditoriale, et interfaces riches. Smile dispose aussi d'une agence spécialisée dans la Tierce Maintenance Applicative, l'infogérance et l'exploitation des applications.

Enfin, Smile est implanté à Paris, Lyon, Nantes, Bordeaux, Lille, Marseille et Montpellier. Et présent également en Suisse, en Ukraine, aux Pays-Bas, au Maroc ainsi qu'en Côte d'Ivoire. Quelques références de Smile



I.2 OPEN SOURCE

En parallèle à ces publications, pour bien comprendre la révolution en marche de l'open source, Smile a publié plusieurs livres blancs expliquant les spécificités, les modèles économiques, les sous-jacents ainsi que les atouts de l'open source.

I.3 QUELQUES REFERENCES DE SMILE

Smile est fier d'avoir contribué, au fil des années, aux plus grandes réalisations Web françaises et européennes ainsi qu'à l'implémentation de systèmes d'information pour des sociétés prestigieuses. Vous trouvez ci-dessous quelques clients nous ayant adressé leur confiance.

Web

EMI Music, Salon de l'Agriculture, Mazars, Areva, Société Générale, Gîtes de France, Groupama, Eco-Emballage, CFnews, CEA, Prisma Pub, Véolia, JCDecaux, 01 Informatique, Spie, PSA, Boiron, Larousse, Dassault-Systèmes, Action Contre la Faim, BNP Paribas, Air Pays de Loire, Forum des Images, IFP, BHV, Gallimard, Cheval Mag, Afssaps, Bénéteau, Carrefour, AG2R La Mondiale, Groupe Bayard, Association de la Prévention Routière, Secours Catholique, Canson, Veolia, Bouygues Telecom, CNIL, Danone, Total, Crédit Agricole…

E-Commerce

Krys, La Halle, The North Face, Kipling, Vans, Pepe Jeans, Hackett, Minelli, Un Jour Ailleurs, Decitre, ANWB, Solaris, Gibert Joseph, De Dietrich, Macif, Figaroclassifieds, Furet du Nord, Gîtes de France, Camif Collectivité, GPdis, Projectif, ETS, Yves Rocher, Bouygues Immobilier, Nestlé, Stanhome, AVF Périmédical, CCI, Snowleader, Darjeeling, Cultura, Belambra …

Collaboratif

HEC, Bouygues Telecom, Prisma, Veolia, Arjowiggins, INA, Primagaz, Croix Rouge, Eurosport, Invivo, Faceo, Château de Versailles, Eurosport, Ipsos, VSC Technologies, Sanef, Explorimmo, Bureau Veritas, Région Centre, Dassault Systèmes, Fondation d'Auteuil, Gaz Electricité de Grenoble, Ville de Niort, Ministère de la Culture, PagesJaunes Annonces, Primagaz, UCFF, Apave, Géoservices, Renault F1 Team, INRIA, CIDJ, SNCD, CS informatique, Serimax, Véolia Propreté, Netasq, Corep, Packetis, Alstom Power Services, Mazars, COFRAC, Assemblée Nationale, DGAC, HEC …

Systèmes d'Information

Veolia Transport, Solucom, Casden Banque Populaire, La Poste, Christian Louboutin, PubAudit, Effia Transport, France 24, Publicis, Nouvelles Frontières, Jus de Fruits de Mooréa, Espace Loggia, Bureau Veritas, Skyrock, Lafarge, Cadremploi, Groupe Vinci, IEDOM, Carrefour, Corsair, Le Bon Coin, Jardiland, Trésorerie Générale du Maroc, Ville de Genève, ESCP, Faiveley Transport, INRA, Yves Rocher, ETS, Perouse Médical, Auchan ecommerce, Viapresse, Danone …

Infrastructure

Agence Nationale pour les Chèques Vacances, Pierre Audoin Consultants, Rexel, Motor Presse, OSEO, Sport24, Eco-Emballage, Institut Mutualiste Montsouris, Ionis, Osmoz, SIDEL, Atel Hotels, Cadremploi, Institut Français du Pétrole, Mutualité Française, Bouygues Telecom, Total, Ministère de l'écologie, Orange, Carrefour, Jardiland, Kantar, Coyote, France Televisions, Radio France, …

Consultez nos références, en ligne, à l'adresse : http://www.smile.fr/clients.

II - SOMMAIRE

I - PREAMBULE 2
I.1 SMILE 2
I.2 OPEN SOURCE 3
I.3 QUELQUES REFERENCES DE SMILE 4
II - SOMMAIRE 6
III - EN RESUME 8
III.1 LE BIG DATA GENERATEUR D'OPPORTUNITES POUR LES ENTREPRISES ET COLLECTIVITES 8
III.1.a Une croissance des masses de données 8
III.1.b Des gisements de données qui se multiplient 8
III.1.c La transformation de la matière première data en valeur pour l'entreprise et ses clients/usagers 8
III.2 UNE TENDANCE DE FOND POUR L'ANALYSE DE DONNEES MASSIVES 9
III.3 CHECKLIST D'UN PROJET BIG DATA 10
III.3.a Cadrer les opportunités métier 10
III.3.b Cadrer l'architecture 10
IV - CE LIVRE BLANC 11
IV.1 VERSION 2015 11
IV.2 APPROCHE 11
IV.3 SUJETS TRAITES 11
V - CONCEPTS ET DEFINITIONS 12
V.1 BIG DATA 12
V.2 ENTREPOT DE DONNEES OU DATAWAREHOUSE 12
V.3 STOCKAGE DISTRIBUE - NOSQL 12
V.3.a Limites des SGBDR dans les architectures distribuées 13
V.3.b Principes de distribution et de réplication des données 13
V.3.c Structures des bases et organisation des données NoSQL 14
V.3.d Fédération de données NoSQL dans des bases relationnelles 16
V.4 INTEGRATION ET TRAITEMENT (DISTRIBUE) DE DONNEES MASSIVES 16
V.4.a ETL 16
V.4.b Frameworks de traitements distribués - Map-Reduce 17
V.5 L'ANALYSE MULTIDIMENSIONNELLE OU OLAP 17
V.6 REQUETAGE AD-HOC EN LANGAGE NATUREL 17
V.7 DATA MINING 17
VI - CAS D'USAGES 18
VI.1 USAGES COUVERTS PAR LES SOLUTIONS BIG DATA POUR L'ANALYSE ET LA VALORISATION 18
VI.2 MARKETING 18
VI.2.a Vue à 360° des clients et analyse des comportements de consommation 18
VI.2.b Ressenti sur les services, produits et concepts 18
VI.2.c Rétention de clients 19

Edition octobre 2015

VI.2.d Implantation de points de vente 19
VI.3 E-COMMERCE 20
VI.3.a Le Big Data accélérateur des ventes 20
VI.3.b Le NoSQL pour gérer facilement des catalogues produits riches 20
VI.4 LOGISTIQUE ET CHAINE D'APPROVISIONNEMENT 20
VI.4.a Le Big Data au service de la traçabilité 20
VI.4.b Le Big Data facteur d'optimisation de la chaîne d'approvisionnement 20
VI.5 OBJETS CONNECTES 21
VI.6 TELECOMS 21
VII - PANORAMA DES SOLUTIONS BIG DATA 22
VII.1 OBSERVATION SUR LE POSITIONNEMENT ACTUEL DES COMPOSANTS 22
VII.2 SYNTHESE DES SOLUTIONS BIG DATA 22
VII.3 HADOOP 25
VII.4 SPARK 29
VII.5 MONGODB 30
VII.6 ETL TALEND FOR BIG DATA 31
VII.7 SUITE PENTAHO 36
VII.8 ELASTICSEARCH 41
VII.9 JASPERSOFT 43
VII.10 APACHE ZEPPELIN 45
VII.11 SPAGOBI 47

III - EN RESUME

III.1 LE BIG DATA GENERATEUR D'OPPORTUNITES POUR LES ENTREPRISES ET COLLECTIVITES

III.1.a Une croissance des masses de données

Chaque jour, la quantité de données créées et manipulées ne cesse d'augmenter, et ce quel que soit le secteur d'activité concerné.

Cette croissance, exponentielle, est liée à :

- l'évolution du nombre d'utilisateurs des solutions IT
- la génération de données par des machines et capteurs
- l'évolution des périmètres couverts et des usages (mobile,...)
- la finesse de l'information tracée
- la croissance des volumes opérationnels
- l'évolution de l'historique de données disponible.

III.1.b Des gisements de données qui se multiplient

Ces données sont issues de sources multiples :
RFID, compteurs d'énergie, opérations commerciales en volumes, transactions financières, blogs, réseaux de capteurs industriels, réseaux sociaux, téléphonie, indexation Internet, parcours de navigation GPS, détails d'appels en call center, e-commerce, dossiers médicaux, informatique embarquée, Internet des objets, données biologiques, données de jeux massivement en ligne, textes de tickets ou mails, sondages, logs,...

Ces masses de données apportent des opportunités d'analyses plus larges et plus fines ainsi que de nouveaux usages de l'information, qu'elle soit pleinement ou partiellement structurée à la source.

III.1.c La transformation de la matière première data en valeur pour l'entreprise et ses clients/usagers

La question n'est plus "Le Big Data peut-il devenir un avantage concurrentiel pertinent ?" mais "**Comment pouvons-nous exploiter les possibilités offertes par ces solutions pour optimiser nos processus d'analyse et de prise de décision ?**.

En effet, les masses de données constituent un matériau brut. Au delà de leur exploitabilité (pertinence, disponibilité et qualité), c'est la capacité à les transformer en analyse et en service qui apporte une valeur maximale.

Le Big Data transforme progressivement les organisations autour de la valorisation de l'information. Avec la finesse d'information sur les opérations passées et de plus en plus d'informations prospectives, le Big Data va permettre l'éclosion de modèles prédictifs plus pertinents.

III.2 UNE TENDANCE DE FOND POUR L'ANALYSE DE DONNEES MASSIVES

Les systèmes de base de données relationnelles et les outils d'aide à la décision n'ont initialement pas été créés afin de manipuler une telle quantité et richesse de données, et il peut vite devenir compliqué et improductif pour les entreprises d'accéder à ces masses de données avec les outils classiques.

Cette nouvelle problématique a donné naissance aux systèmes de gestion de base de données appelés « NoSQL » (Not Only SQL), qui ont fait le choix d'abandonner certaines fonctionnalités des SGBD classiques au profit de la simplicité, la performance et de la capacité à monter en charge.
Des frameworks comme Hadoop ont également été créés et permettent le requêtage, l'analyse et la manipulation de ces données en masse.

Nous relevons que les principales solutions de Big Data sont open source. Ce contexte favorise leur vitesse de développement et de diffusion au sein des entreprises et collectivités.
Et ce à moindre coût par rapport à des solutions dont l'évolution de la capacité est verticale : coût des ressources matérielles, licences,...

Il est possible de mettre en place une solution Big Data complète uniquement basée sur des solutions open source sans coût de licence. Toutefois, des versions commerciales basées sur de l'open source apportent des facilités qui vont dans le sens de la productivité de mise en oeuvre et de l'exploitabilité des solutions avec des outils d'administration complémentaires notamment.

Beaucoup d'entreprises et de collectivités publiques utilisent déjà des solutions Big Data, souvent hébergées dans le cloud (ex : Google Analytics, réseaux sociaux, Salesforce,...).
Les solutions Big Data ont fait leurs preuves et sont mûres pour un déploiement en production.

Les fonctionnalités de visualisation graphique (DataViz), pour illustrer des analyses portant sur des masses de données, et de datamining prennent avec le Big Data toute leur importance.
Ce mouvement va de pair avec le développement des librairies JavaScript de visualisation graphique avancées (d3.js,...) et des frameworks Javascript d'interactivité avec les données.

Nous relevons aussi les possibilités de consolidation de données (massives) et hétérogènes ɛà la voléeɛ en complément de l'entrepôt de données : la fédération de données.

III.3 CHECKLIST D'UN PROJET BIG DATA

Au delà des principes et bonnes pratiques de mises en oeuvre de solutions IT, une vigilance sur les points suivants peut éviter des écueils lors du cadrage d'un projet Big Data:

III.3.a Cadrer les opportunités métier

- identifier des leviers de gain d'exploitation de masses de données sur les activités de l'entreprise
- identifier le périmètre (légal, technique, historique) d'information disponible : SI interne, données fournies par des partenaires, OpenData, ...
- identifier le ou les cas d'utilisation résultant de l'adéquation entre les leviers de gain et le périmètre d'information disponible

III.3.b Cadrer l'architecture

- définir une architecture flexible adaptée au(x) cas d'utilisation; il n'existe pas un modèle d'architecture Big Data idéal adapté à tous les usages
- valider la disponibilité et l'exploitabilité des données sources
- valider l'architecture (matérielle, réseau, applicative) par un test de montée en charge.

IV - CE LIVRE BLANC

IV.1 VERSION 2015

Cette nouvelle version du livre blanc (la première datant de février 2014) nous permet de compléter les usages et de prendre en compte les derniers apports de l'éco-système Big Data qui voit des évolutions rapides, notamment autour d'Hadoop et de Spark, ainsi que des nouvelles versions de solutions open source.

IV.2 APPROCHE

Comme les autres livres blancs publiés par Smile, cet ouvrage s'efforce de réunir :

- une approche générale de la thématique, ici : l'analyse et la valorisation de masses de données, ses concepts, ses champs d'application, ses besoins spécifiques.
- un recensement des meilleures solutions open source dans ce domaine.
- une présentation assez complète de ces solutions, de leurs forces, de leurs limites, de leur maturité et de leur aptitude à satisfaire des besoins opérationnels.

Cette étude, réalisée par notre équipe de consultants, a été fondée sur plusieurs années de travail de recherche et de premiers déploiements effectifs de solutions Big Data.
Cet ouvrage vient compléter livres blancs Smile Décisionnel et NoSQL.
Les marques et logos présents dans ce livre blanc sont la propriété des entreprises concernées.

IV.3 SUJETS TRAITES

Ce livre blanc est concentré sur les solutions applicatives de collecte et de valorisation de masses de données.
D'autres aspects de l'exploitation des masses de données sont importants mais non décrits ici :

- **Qualité des données :** prendre en compte la qualité et le nettoyage des données, ainsi que la gestion du cycle de vie des données référentielles dans le scope du projet évite d'aboutir à une masse de données inexploitable. Des solutions de traitement, qualification et nettoyage automatique des données existent : fonctionnalités intégrées aux flux de données ETL, briques complémentaires telles DataQuality de Talend.
- **Infrastructures techniques :** les solutions Big Data nécessitent une architecture répartie. La composante système et réseaux est un facteur clé de performance et d'exploitabilité d'une solution Big Data.
- **Sécurité de l'information :** les aspects de sécurisation des accès et de gestion de l'intégrité des données sont importants pour la mise en oeuvre d'une solution pérenne.

- **Respect de la vie privée :** les solutions Big Data peuvent apporter une puissance informative importante. Cette puissance doit respecter les libertés individuelles.
- **Solutions :** l'écosystème des solutions Big Data est riche et évolutif. Il nous serait difficile de détailler toutes les solutions. Nous nous sommes concentrés sur les solutions les plus pertinentes à l'heure actuelle.

V - CONCEPTS ET DEFINITIONS

V.1 BIG DATA

Le Big Data consiste en un ensemble de données plus ou moins structurées qui deviennent tellement volumineuses qu'elles sont difficiles à travailler avec des outils classiques de gestion de base de données.

En 2012, Gartner a posé les bases de la définition du Big Data, basée sur les 3V :

- Volume
- Vitesse
- Variété des données.

→ "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."

Sans seuil ni repère, beaucoup de bases de données classiques peuvent prétendre à répondre à ces trois critères.
Dans le présent livre blanc, pour les usages d'analyse, nous compléterons pragmatiquement la combinaison des 3V avec une considération de volumétries en dizaines de millions d'enregistrements minimum.

V.2 ENTREPOT DE DONNEES OU DATAWAREHOUSE

L'entrepôt de données est une base de données qui concentre de l'information issue de différents systèmes d'information de l'entreprise, à des fins d'analyse et de reporting des activités et marchés.

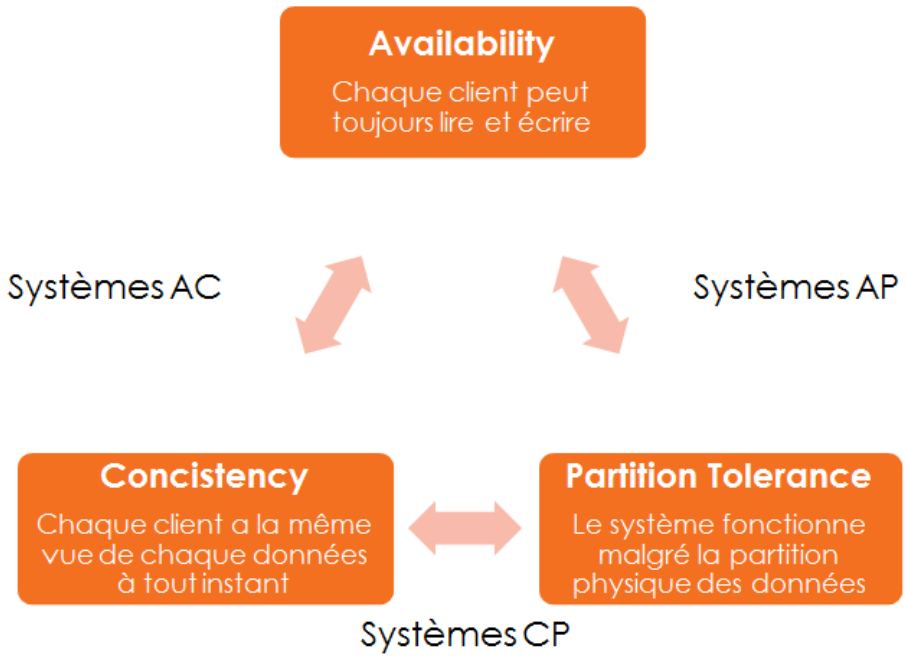
V.3 STOCKAGE DISTRIBUE - NOSQL

NoSQL, pour Not Only SQL désigne les systèmes de gestion de base de données qui ne s'appuient plus, du fait des volumétries et de la variété des données contenues, sur une architecture relationnelle et transactionnelle.
Ces systèmes privilégient la simplicité et l'évolutivité de la capacité via des architectures distribuées.

V.3.a Limites des SGBDR dans les architectures distribuées

Outre leur modèle relationnel, la plupart des moteurs de SGBDR (Système de Gestion de Bases de Données Relationnelles) sont transactionnels ce qui leur impose le respect des contraintes Atomicity Consistency Isolation Durability, communément appelé par son acronyme ACID.

Théorème de CAP



Il est actuellement impossible d'obtenir ces trois propriétés en même temps dans un système distribué. Sur de nombreux SGBDR classiques, la réplication devient plus complexe avec de fortes volumétries et une forte vélocité des données.

V.3.b Principes de distribution et de réplication des données

Les capacités de montée en charge des bases NoSQL reposent, au delà de leur simplicité, sur la distribution (sharding) et la réplication des données sur différents noeuds (cluster de quelques serveurs à plusieurs DataCenter).
Pour simplifier, une analogie peut être faite entre les mécanismes de partitionnements verticaux (sur plusieurs tables physiques de la même instance) de certains moteurs de bases de données relationnelles et la distribution horizontale (sur plusieurs serveurs) des données en NoSQL.
Les données peuvent également être répliquées, sur un principe analogue aux mécanismes de stockage en RAID, afin de garantir un haut niveau de service, même en cas de problème ou de maintenance d'un nœud du cluster.

V.3.c Structures des bases et organisation des données NoSQL

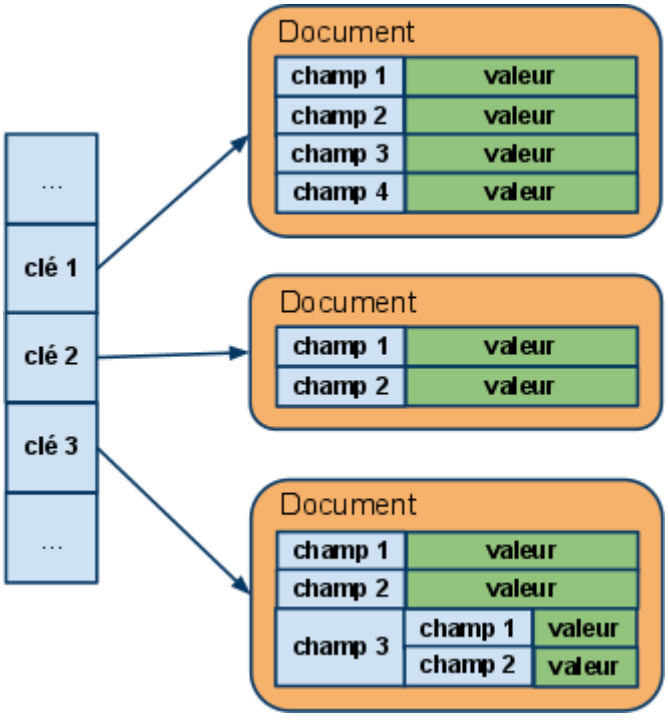
Il existe plusieurs paradigmes au niveau des systèmes de stockage NoSQL :

Type documentaire

Les bases de données documentaires sont constituées de collections de documents. Les collections sont généralement assimilées à des tables d'un modèle relationnel.

Bien que les documents soient structurés, ces bases sont sans schéma de données prédéfini. Il n'est donc pas nécessaire de définir au préalable l'ensemble des champs utilisés dans un document. Les documents peuvent donc avoir une structure hétérogène au sein de la base.

Un document est composé de champs et de valeurs associées, ces dernières pouvant être
requêtées. Les valeurs peuvent être, soit d'un type simple (entier, chaîne de caractère,
date, ...), soit composées de plusieurs couples clé/valeur (imbrications nested sets).
Les structures de données sont donc très souples.



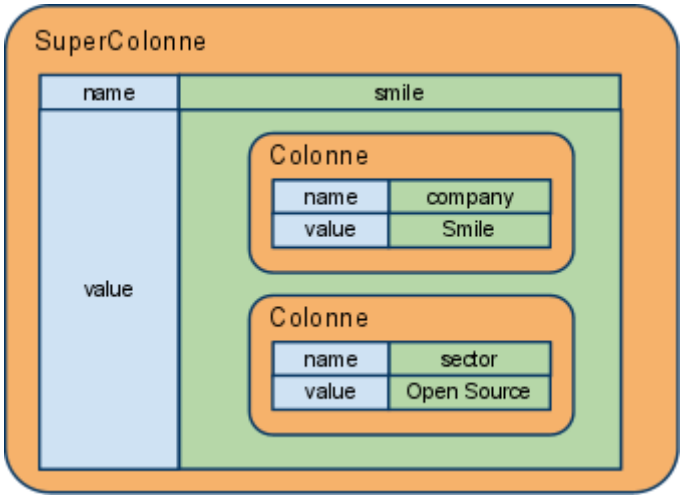
La souplesse du modèle de données, les performances et les capacités de requêtage orientent l'usage des bases documentaires vers du stockage opérationnel de masse (ODS) dans un système décisionnel.

Type graphe

Au delà du moteur de stockage sous la forme d'une base documentaire, ce type de base propose également des relations entre objets. Ces derniers sont orientés et peuvent porter des propriétés.

Type orienté colonnes

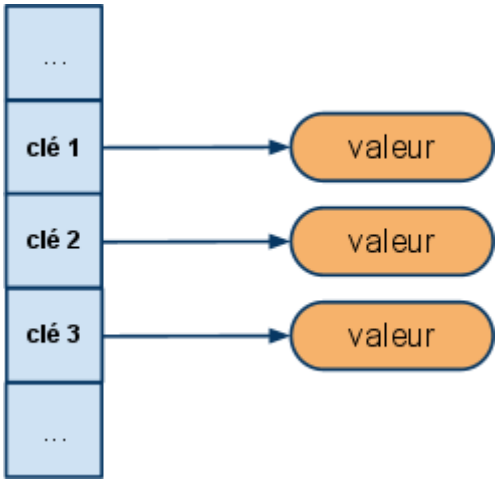
La colonne représente l'entité de base de la structure de données. Chaque colonne d'un objet est défini par un couple clé / valeur. Une colonne contenant d'autres colonnes est nommée super- colonne.



Ces types de bases sont adaptés au stockage opérationnel de masse (ODS) et de source d'analyses massives dans un système décisionnel.

Type clé/valeur

Dans ce modèle, chaque objet/enregistrement est identifié par une clé unique.
La structure de l'objet est libre.



Dans ce modèle on ne dispose généralement que des quatre opérations Create, Read, Update, Delete (CRUD) en utilisant la clé de l'enregistrement à manipuler.

Du fait des limites fonctionnelles d'accès aux données de ces types de base, nous ne leur voyons pas d'application décisionnelle.

V.3.d Fédération de données NoSQL dans des bases relationnelles

Plusieurs moteurs de bases de données relationnelles permettent de fédérer des lacs de données massives NoSQL externes au sein de bases de données classiques.
Le modèle est ici d'utiliser un moteur de stockage NoSQL (réparti et qui reste accessible de manière autonome) au sein d'une base de données relationnelle pour son exploitation.
Citons par exemple le mécanisme de Foreign Data Wrapper de PostGreSQL ou le connecteur Cassandra de MariaDB.

Ces mécanismes offrent l'avantage d'intégrer facilement des données de bases NoSQL au sein d'un ODS ou un entrepôt de données de type base de données relationnelle et ainsi d'y accéder avec un langage SQL classique.
Par exemple, cela peut être une source de fait MongoDB à très forte volumétrie et vélocité intégrée de manière ξtransparenteξ à un ODS PostGreSQL.

Par contre, il faut bien garder à l'esprit les limites de ce modèle :

- limites de performances techniques du moteur de la base de données fédératrice par rapport à un système de traitement réparti (agrégation de masses de données notamment)
- perte de performance due à l'intégration d'un système tiers
- mapping ξrigideξ des champs entre la base NoSQL et les tables virtuelles de la base de données fédératrice.

V.4 INTEGRATION ET TRAITEMENT (DISTRIBUE) DE DONNEES MASSIVES

V.4.a ETL

Afin d'alimenter un datawarehouse à partir des différentes sources de données ou de synchroniser en batch des données entre systèmes, on utilise une gamme d'outils appelés ETL, pour « Extract, Transform, Load ».
Comme le nom l'indique, ces outils permettent d'extraire des données à partir de différentes sources, de les transformer (rapprochement, format, dénomination, calculs), et de les charger dans une cible, comme un datawarehouse dans le cas d'un projet décisionnel.

L'ETL permet de masquer, grâce à une modélisation visuelle, la complexité de manipulations (réparties) des données (hétérogènes) au sein des traitements et ainsi d'en réduire fortement les coûts de développement, de maintenance et d'exploitation.
Un ETL est généralement composé d'un studio de modélisation des traitements ainsi que d'un ou plusieurs environnements d'exécution et des outils d'administration voire de visualisation de données suivant les versions.

V.4.b Frameworks de traitements distribués - Map-Reduce

Modèle d'architecture portant sur la distribution et la répartition des traitements de données sur plusieurs noeuds d'une grappe de serveurs (cluster).
Dans l'étape Map, les données à traiter et traitements à effectuer sont répartis sur les noeuds de traitement.
Dans l'étape Reduce, les nœuds de traitements remontent leur résultat pour agrégation (il peut y avoir plusieurs niveaux de traitement).

V.5 L'ANALYSE MULTIDIMENSIONNELLE OU OLAP

L'analyse multidimensionnelle permet l'analyse de mesures suivant différents aspects métiers appelés dimensions ou axes d'analyse et ce, à plusieurs niveaux de regroupement.
Par exemple, la mesure de Montant HT d'une ligne de facture peut être agrégée par :

- jour → mois → trimestre → année
- produit → catégorie de produits → ligne de produits
- client → segment de client.

V.6 REQUETAGE AD-HOC EN LANGAGE NATUREL

Le requêtage ad-hoc permet à des non informaticiens de construire visuellement des requêtes, en s'appuyant sur un dictionnaire d'informations en langage naturel (métadonnées) faisant abstraction du langage technique d'accès aux bases de données (SQL, JSON).

V.7 DATA MINING

Le data mining consiste à rechercher des informations statistiques utiles cachées dans un grand volume de données.
L'utilisateur est à la recherche d'une information statistique qu'il n'identifie pas encore.

VI - CAS D'USAGES

VI.1 USAGES COUVERTS PAR LES SOLUTIONS BIG DATA POUR L'ANALYSE ET LA VALORISATION

Il existe de nombreux cas d'usage des solutions de valorisation et d'analyse massive de données.
Nous en avons détaillé quelques uns ci-dessous mais nous pouvons aussi citer l'analyse fine de processus, la recherche scientifique, les analyses politiques et sociales, l'analyse de données de capteurs sur les chaînes industrielles...

VI.2 MARKETING

Le Big Data transforme en profondeur les métiers du marketing, avec les facilités suivantes :

VI.2.a Vue à 360° des clients et analyse des comportements de consommation

Une vue complète de chaque client nécessite la consolidation de larges ensembles de données:

- informations sur le client stockées dans le SI : ERP, CRM, bases opérationnelles...
- segmentation
- parcours omnicanal / historique de la relation depuis la prospection
- comportements d'achat : détail des commandes, fréquence, canaux
- enquêtes de satisfaction, réseaux sociaux publiques
- niveau d'engagement; parrainage d'autres clients
- expérience d'utilisation d'objets connectés
- utilisation des services après-vente.

La collecte et la consolidation de toutes ces données représente une tâche fastidieuse, rarement faite ou uniquement sur un petit panel de clients. Les solutions Big Data peuvent permettre d'automatiser cela et apporter les gains suivants :

- détecter des besoins rendus visibles après corrélation de données
- optimiser l'adéquation des produits et services proposés
- affiner les ciblages et optimiser les communications avec chaque client : canal, message,...

VI.2.b Ressenti sur les services, produits et concepts

Corréler les données provenant :

- des activités et échanges de support après-vente
- des avis, enquêtes de satisfaction et de l'analyse de mots postés sur les réseaux sociaux publics.

VI.2.c Rétention de clients

Le Big Data permet de détecter des signes de désengagements de clients en utilisant la consolidation d'informations de plusieurs capteurs digitaux de la relation/de l'interaction client tels les appels ou tickets Back Office, la lecture de procédures de retrait sur le site web, des messages de réseaux sociaux publics,...

VI.2.d Implantation de points de vente

La technologie Big Data offre la possibilité de corréler des données de différentes natures et de différentes sources pour déterminer le meilleur emplacement pour un point de vente :

- OpenData mises à disposition par les collectivités
- données géographiques
- données socio-économiques
- informations disponibles sur le marché et la concurrence.

VI.3 E-COMMERCE

L'e-commerce est par nature une activité où la relation client est digitale et donc consommatrice et génératrice de données utiles au processus marketing et de vente. Les principales solutions d'analyse d'audience web (pages visitées, recherches,...) du marché utilisent des solutions Big Data.

VI.3.a Le Big Data accélérateur des ventes

Le Big Data peut apporter des solutions pour :

- analyser les tunnels de vente dans un contexte omnicanal --> les leviers et freins de transformation à partir de plusieurs sources de données / canaux
- analyser des comportements d'achat des clients afin d'optimiser leur expérience
- corréler les ventes avec les retours, livraisons et données financières
- analyser finement l'usage et les interactions des utilisateurs avec le site e-commerce :
    - prédire à chaud la prochaine étape dans le processus de vente
    - Real User Monitoring
- analyser un positionnement tarifaire par rapport au marché et aider à l'optimisation des prix dans des objectifs de volumes et de rentabilité
- détecter des fraudes.

VI.3.b Le NoSQL pour gérer facilement des catalogues produits riches

Par ailleurs, les bases NoSQL documentaires sont particulièrement adaptées à l'entreposage et l'analyse de données souples et complexes, telles les caractéristiques de produits.

VI.4 LOGISTIQUE ET CHAINE D'APPROVISIONNEMENT

VI.4.a Le Big Data au service de la traçabilité

Les solutions Big Data permettent une pleine traçabilité des opérations logistiques :

- mouvements de stock - RFID
- produits frais ou sensibles
- suivi de flotte ou de colis, y compris lors de transport intermodal.

Ces solutions facilitent les opérations de suivi des voyages dans le temps : geo corridoring, analyse des voyages et taux de rotation

VI.4.b Le Big Data facteur d'optimisation de la chaîne d'approvisionnement

La masse de données disponible sur tous les mouvements permet d'analyser et de piloter plus finement les processus logistiques et d'approvisionnement.

La richesse d'information permet de combiner les différents facteurs de qualité (délais, défauts, qualité de service,...) et économiques (prix d'achat, coût de possession et de stockage,...) dans les analyses.

Le Big Data permet d'intégrer plus facilement les données logistiques dans les informations du cycle de vie des objets (commande, logistique, exploitation, recyclage,...) et permet ainsi une vision à 360° autour de la fonction d'approvisionnement.

VI.5 OBJETS CONNECTES

Ces dernières années ont vu le développement et la diffusion de masse d'objets connectés grand public, notamment autour des thèmes de la santé et du sport ainsi que de la maison connectée.
Une illustration de l'apport des technologies Big Data est l'utilisation de thermostats connectés qui permettent de réaliser des économies d'énergies grâce à l'application de machine learning sur les données issues des sondes consolidées avec des prévisions météorologiques et paramètres de l'utilisateur.

VI.6 TELECOMS

Les télécoms génèrent des masses de données sur les flux transités. Le Big Data est une solution utile pour :

- l'analyse de capacité
- la segmentation des usagers et des comportements d'usage des réseaux
- la corrélation avec les processus de vente et de support
- la qualité de service de réseaux complexes, la corrélation avec les appels aux call center.

VII - PANORAMA DES SOLUTIONS BIG DATA

VII.1 OBSERVATION SUR LE POSITIONNEMENT ACTUEL DES COMPOSANTS

L'écosysteme big data est riche, et une solution unique ne répond pas à tous les besoins, imposant une interopérabilité forte entre les solutions. Aussi, au delà du choix de telle ou telle solution, il sera important de savoir associer les solutions entre elles pour en tirer le meilleur, dans votre contexte.

Par exemple :

- l'intégration de briques de traitement et requêtage Hadoop avec du stockage MongoDB ou Cassandra.
- plusieurs ETL peuvent s'appuyer sur les frameworks de traitement distribué Hadoop.

VII.1.a Projets de la fondation Apache

Les principales technologies big data ont été initiées par les acteurs du web comme Google, Facebook, Twitter ou Yahoo, puis reversées en open source, sous licence libre. Ceci leur assure un développement communautaire et une diffusion plus large, avec l'ambition de constituer un standard sinon un socle réutilisable.

La grande majorité des projets reversés est placé sous la gouvernance de la fondation Apache, ce qui en fait le leader actuel en termes de big data.

VII.2 SYNTHESE DES SOLUTIONS BIG DATA

VII.2.a Composants d'intégration et de traitement de données

Type	Solution	Site web de la solution
Intégration de logs	Apache Flume	http://flume.apache.org
Intégration de flux	Apache NIFI	https://nifi.apache.org
Interface SQL	Apache Hive	http://hive.apache.org
Interface SQL	Apache Drill	https://drill.apache.org
Interface SQL	Presto	https://prestodb.io
Framework de	Apache Pig	https://pig.apache.org

requêtage et traitement		
Interface SQL	Cloudera Impala	http://www.cloudera.com/content/cloudera/en/products-and-services/cdh/impala.html
ETL	Talend for Big Data	http://fr.talend.com/products/big-data
ETL	Pentaho Data Integration	http://www.pentaho.fr/explore/pentaho-data-integration
Framework de traitement	Hadoop YARN & MapReduce	https://hadoop.apache.org
Outil et langage haut niveau de développement	Pig	https://pig.apache.org
Framework de traitement orienté temps réel	Storm	http://storm-project.net
Framework de traitement	Spark	http://spark.apache.org
Framework de traitement	Tez	https://tez.apache.org
Intégration de données en SGBDRs	Apache Sqoop	http://sqoop.apache.org
Système de messages distribué	Apache Kafka	http://kafka.apache.org

Composants de stockage de données

Type	Solution	Site web de la solution
NoSQL Colonne	Apache Cassandra Base de données répartie en Peer to Peer	http://cassandra.apache.org
NoSQL Colonne	Apache HBase Base de données du framework Hadoop Voir Hadoop pour sa	http://hbase.apache.org

	description	
NoSQL Colonne	Kudu	http://getkudu.io
NoSQL Document	MongoDB	http://www.mongodb.org
NoSQL Document	ElasticSearch	http://www.elasticsearch.org
NoSQL Graph	Neo4j	http://www.neo4j.org
Système de fichiers distribué	Hadoop HDFS	https://hadoop.apache.org

VII.2.b Composants d'analyse et de restitution

Type	Solution	Site web de la solution
Portail décisionnel complet	Pentaho Business Analytics	http://www.pentaho.fr
Portail décisionnel complet	JasperSoft BI Suite	http://www.jaspersoft.com/fr
Portail décisionnel complet	Spago BI	http://www.spagobi.org
Portail de tableaux de bord web	ElasticSearch Kibana	http://www.elasticsearch.org/overview/kibana
Portail décisionnel complet	Vanilla Platform	http://bpm-conseil.com
Portail d'analyse et de visualisation de données	Apache Zeppelin	https://zeppelin.incubator.apache.org/
Moteur OLAP Big Data	Apache Kylin	http://kylin.incubator.apache.org
Framework web pour R	Shiny	http://shiny.rstudio.com
Framework web de data-visualisation	D3.js	http://d3js.org
Portail de reporting	JSReport	http://jsreport.net

VII.3 HADOOP

Editeur : Fondation Apache Licence : Apache Licence V2 et commerciales (suivant la distribution et la version) Version actuelle : 2 (pour le cœur)	

VII.3.a

Hadoop est un ensemble de projets et d'outils open source de la fondation Apache permettant de stocker et traiter massivement des données. Hadoop a été développé à l'origine par Facebook et Yahoo.

VII.3.b Principes

- Répartir le stockage et les traitements
- traiter au plus proche du stockage, afin de limiter les échanges de données massives entre noeuds du cluster

VII.3.c Distributions Hadoop

De manière analogue à Linux, et s'il est possible de compiler, paramétrer et intégrer εmanuellementε les différents composants, il existe **plusieurs distributions de Hadoop, simplifiant le déploiement et l'administration**, telles que **Hortonworks**, **Cloudera** et **MapR**.

Chaque distribution apporte une valeur ajoutée différente, et il n'existe pas une solution unique qui correspond à tous les usages.

VII.3.d Principaux composants Hadoop

Framework de traitements parallélisés Map-Reduce

Hadoop Map-Reduce est un puissant framework Java de traitement de données massives.
A noter que dans le cas de l'utilisation conjointe avec HDFS et HBase et suivant la configuration du cluster Hadoop, une partie des traitements sont effectués au niveau des noeuds de stockage.

HDFS : Hadoop Distributed File System

HDFS est un système de fichiers distribué sur des noeuds d'un cluster Hadoop.
HDFS est adapté au stockage et la réplication de fichiers de grande taille (>256MB).
A noter qu'il existe plusieurs formats de stockage des données dans HDFS dont certains en colonne comme ORC, Parquet,...

Hive

Hadoop Hive permet de fournir une interface SQL à Hadoop, de manière analogue à une base de données classique. La présence de connecteurs JDBC et ODBC permet une connexion facile depuis des portails BI, tableurs, applicatifs métier,…
Hive permet de définir des tables appuyées sur des données du cluster Hadoop et externes.
Hive s'appuie sur les moteurs de traitement MapReduce, Spark et Tez (le choix du moteur est paramétrable) pour opérer les requêtes. Depuis la version 0.14, il est possible de réaliser des opérations de type INSERT, UPDATE et DELETE.
De part l'initiative Stinger (et Stinger.next), les performances de Hive ont été grandement améliorées, permettant de passer à un usage batch/forte latence à un usage interactif.

Hbase

HBase est une base de données NoSQL répartie en colonnes, inspirée de Google BigTable.
La mise en oeuvre de HBase repose généralement sur un système de fichiers répartis HDFS.
HBase peut être exploité en SQL avec une connectivité JDBC au travers d'Apache Phoenix ou de Hive.

Pig

Pig est un outil de développement haut-niveau de flux Big Data pour manipuler des ensembles de données. Dans la pratique, Pig est surtout utilisé pour du raffinage de données.
Pig permet l'intégration de fonctions et librairies externes afin d'étendre ses capacités de traitement. L'exécution peut exploiter les moteurs Spark et Tez au delà de MapReduce.

Tez

Tez est un moteur de traitement apportant la capacité d'effectuer les traitements répartis et successifs sans stockage intermédiaire (directed-acyclic-graph), améliorant ainsi les performances/réduisant la latence par rapport à MapReduce.

Kafka

Kafka permet l'intégration de messages applicatifs (broker) à forte volumétrie.

Flume

Flume permet l'intégration distribuée de logs et de données issues de réseaux sociaux.

Sqoop

Sqoop intègre des données à partir et vers des bases de données relationnelles.

SolR

SolR est un puissant moteur de recherche, basé sur Apache Lucene, intégré à Hadoop.

Oozie

Oozie est un moteur d'ordonnancement, de workflow et de coordination de tâches Hadoop (Mapr-Reduce, Pig,...).

Zookeper

Zookeper est un module de gestion de configuration pour les systèmes distribués.

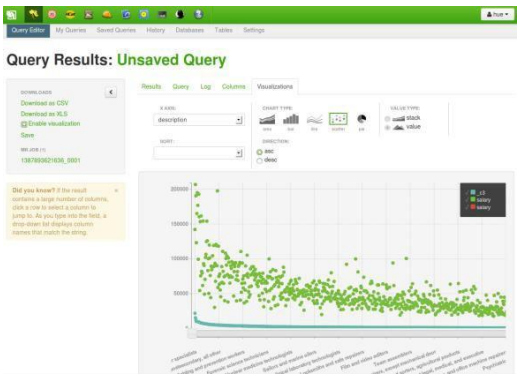
Mahout

Mahout est une librairie Java qui permet d'implémenter différents algorithmes de data mining sur un cluster Hadoop.
Ces algorithmes sont développés à partir de MapReduce. Cependant, ils ne se limitent pas uniquement à Hadoop et certains fonctionnent sur d'autres environnements, dont non distribués.

Hue

Hue est un portail web d'exploitation de clusters Hadoop qui permet de:

- réaliser des requêtes Hive (Beeswax) :



- éditer, gérer et exécuter des traitements (jobs MapReduce, scripts Pig et Spark avec coloration syntaxique)
- construire des tableaux de bords interactifs avec un filtrage basé sur la recherche.

VII.3.e Usages et possibilités pour le décisionnel

Big Data

L'ensemble Hadoop fournit un éco-système permettant de traiter de nombreux cas d'usages pour le décisionnel Big Data :

- l'entreposage de données opérationnelles (ODS HDFS ou Hbase) ou en entrepôt de données (Hbase et Hive).
- l'intégration et le traitement parallélisé de données (YARN, Map-Reduce, Pig, Spark)
- le requêtage et l'analyse de masses de données (Hive+YARN, Map-Reduce, Pig, Spark)
- le datamining (Mahout, Spark MLlib, RHadoop).

Notons que les principaux portails décisionnels open source intègrent directement un connecteur Hive pour une exploitation des données traitées dans un cluster Hadoop.

VII.4 SPARK

Editeur : Fondation Apache Licence : Apache V2 Version actuelle : 1.5	

Spark est un moteur de traitement de données distribué orienté mémoire. Il permet ainsi de traiter massivement des données avec une faible latence.
Spark peut être utilisé seul et s'intègre avec HADOOP, Cassandra, MongoDB, ElasticSearch, des bases de données avec connecteur JDBC,...

C'est une brique de traitement de plus en plus utilisée, et un éco-système s'est développé autour de Spark :

- SparkSQL
- Spark Streaming
- MLlib
- GraphX
- Spark-jobserver

VII.4.a Intégration avec HADOOP

S'il peut fonctionner de manière autonome, Spark, intégré à plusieurs distributions HADOOP, permet d'exploiter les données stockées dans HDFS (et Hbase).
Il peut également servir de moteur de traitement à Pig et Hive.

VII.5 MONGODB

Type NoSQL : document Editeur : MongoDB Licences : GNU AGPL v3.0 et commerciale (suivant la version) Version actuelle : 3	

VII.5.a

MongoDB est une base de données NoSQL de type document, **la définition des données est très souple et chaque enregistrement a sa propre structure**, dont les objets sont stockés au format JSON binaire (BSON).

MongoDB permet de gérer la réplication et la répartition de données sur un ensemble de serveurs (cluster), ce qui assure un **service en très haute disponibilité**.

VII.5.b Connectivité, requêtage et traitement

L'avantage du format JSON est son utilisation native dans de nombreux langages de programmation, notamment le Javascript; la console MongoDB est d'ailleurs un interpréteur Javascript.
MongoDB fournit également des fonctions JavaScript de traitement réparti : MongoDB Map-reduce.

VII.5.c Usages Big Data BI

MongoDB peut servir d'Operating Data Store.
Avec ses connecteurs disponibles au sein de la plupart des solutions BI, open source ou non, MongoDB peut aussi servir d'entrepôt de données de masse à des fins de requêtage et de reporting.

L'analyse multidimensionnelle (OLAP) avec MongoDB nécessite actuellement l'emploi combiné d'un composant supplémentaire, tel :

- Hadoop Hive+Map-Reduce
- une fédération de données JDBC :
  - l'ETL Pentaho Data Integration avec son connecteur JDBC et du moteur Map-Reduce de MongoDB
  - Foreign Data Wrapper de PostGreSQL.

VII.5.d Conclusion

A l'heure où nous écrivons ces lignes, **MongoDB est la base NoSQL la plus populaire** d'après le site db-engines.com, bénéficiant d'une relative facilité de mise en oeuvre ainsi que d'un scope fonctionnel utile à l'entreposage opérationnel de masse de données.

VII.6 ETL TALEND FOR BIG DATA

Editeur : Talend Licences : Apache V2 et commerciale (suivant la version) Version actuelle : 6 (TOS); 5.6 (EE)	talend

Éditeur et solutions

Talend est un éditeur basé en France (Talend SA) et en Californie (Talend Inc.). La société Talend, fondée en 2005, est soutenue dans son développement par des investisseurs tels Idinvest Partners (AGF Private Equity), Silver Lake Sumeru, Balderton Capital, Bpifrance et Iris Capital. Talend a réussi une levée de fonds de 40 millions de dollars fin 2013.

Talend offre un large éventail de solutions middleware répondant aux besoins de gestion de données et d'intégration d'applications, à travers une plateforme unifiée et flexible :

- l'intégration de données (ETL)
- la qualité de données (DQ)
- les architectures orientées services (ESB)
- la gestion de référentiels de données (MDM).

Talend obtient une reconnaissance forte de la part des observateurs tel le Gartner (Magic Quadrants).
Les solutions sont disponibles en version communautaire (Talend Open Studio for Data Integration / Big Data) et en version commerciale avec des fonctionnalités supplémentaires et un support éditeur.
Les fonctionnalités ETL classiques de Talend sont présentées plus en détail dans le livre blanc Décisionnel de Smile (http://www.smile.fr/Livres-blancs/Erp-et-decisionnel/Le-decisionnel-open-source).

Talend et le Big Data

Talend propose depuis début 2012 une gamme de solutions Big Data, allant de la version Open Studio à la plateforme d'intégration massive de données (Talend Platform for Big Data).
Talend a établi des partenariats avec des acteurs majeurs du Big Data, notamment : Cloudera, EMC Greenplum, Google, HortonWorks, MapR.

Plus d'informations :

- http://fr.talend.com/solutions/etl-analytics
- http://www.talend.com/solutions/big-data
- http://fr.talend.com/products/platform-for-big-data

VII.6.a Fonctionnalités

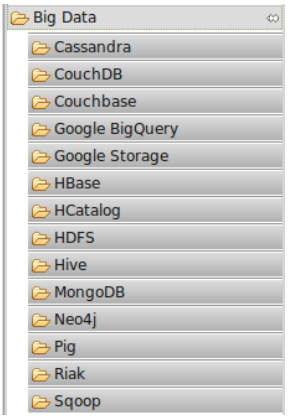
ETL Talend Open Studio for Big Data

Talend est un ETL de type « générateur de code », c'est-à-dire qu'il offre la capacité de créer graphiquement des processus (répartis) de manipulation et de transformation de données puis de générer l'exécutable correspondant sous forme de programme Java (et scripts Pig).

Ce programme peut ensuite être déployé sur un ou plusieurs serveur(s) d'exécution.

La modélisation des traitements se fait dans le Studio Talend, qui permet d'utiliser des connexions prédéfinies et les tâches de transformations pour collecter, transformer et charger les données par simple **glisser-déposer** dans l'espace de modélisation.

⊕ Palette de connecteurs Big Data



L'ETL Talend fournit nativement une large palette de connecteurs permettant de s'interfacer à la plupart des systèmes existants : bases de données, fichiers locaux ou distants, web services, annuaires,...

Si l'ETL classique Talend peut se connecter aux principales bases NoSQL via des connecteurs communautaires ou APIs, la version Talend Open Studio for Big Data fournit nativement toute la flexibilité et les connecteurs d'intégration de masses de données, dont :

- les bases NoSQL : MongoDB, Apache Hadoop/Hive, Cassandra, Google BigQuery, Neo4j
- HDFS, HCatalog
- le chargement massif de bases NoSQL MongoDB et Cassandra ainsi qu'Apache Sqoop.

⊕ Composants de transformation

Les composants de transformation permettent entre autres :

- les multiplexages et jointures
- les filtrages (lignes, colonnes), le dédoublonage
- l'exécution d'opérations sur des évènements en base ou sur des fichiers
- les manipulations de fichiers locaux ou distants...

La liste des composants Talend est disponible à l'adresse suivante :
http://www.talendforge.org/components/index.php
Les capacités de traitement des données peuvent être étendues avec :

- les composants communautaires disponibles sur Talend Exchange
- l'intégration de librairies externes
- des fonctions de traitement spécifiques Java ou Pig.

⊕ Gestion des différents environnements d'exécution des traitements

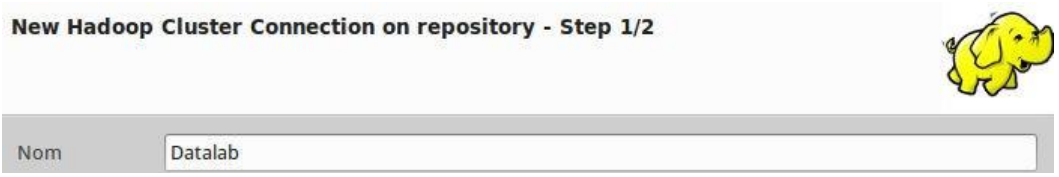
L'ETL Talend gère des contextes d'exécution permettant d'externaliser l'ensemble des paramètres d'accès et variables d'exécution utilisés dans les composants / jobs.

Les utilisateurs peuvent ainsi configurer les paramètres à la volée lors de l'exécution ou utiliser des paramètres différents pour chaque contexte d'exécution : le développement, la recette et la production.

⊕ Intégration avec Hadoop

Paramétrage de cluster

L'ETL Talend for Big Data permet de paramétrer un cluster de manière analogue à une connexion classique à une base de données au travers d'un assistant :



Une fois le cluster paramétré, l'interface propose une **découverte automatique des services Hadoop déployés et accessibles du cluster**, afin d'en faciliter l'utilisation au sein des traitements ETL:

Découverte automatique et sélection des services d'un cluster Hadoop

Génération de traitements répartis Pig

Talend for BigData propose de **produire visuellement des traitements (répartis) Hadoop.**

En effet, à partir d'une modélisation de flux avec des composants graphiques prédéfinis disponibles dans la palette, **Talend for Big Data génère le code Pig**, permettant de bénéficier de la puissance de traitement du cluster Hadoop sans avoir à saisir du code.

Intégration et requêtage SQL avec Hive

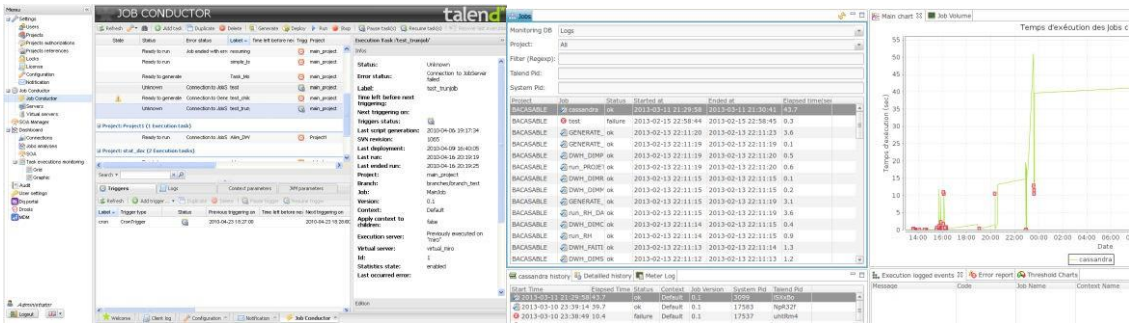
Il est également possible d'utiliser le mode ELT (Extract, Load and Transform) avec Hive pour répartir des requetes et traitements sur un cluster Hadoop.

Talend Enterprise for Big Data

De manière analogue à Talend Enterprise for Data Integration pour l'ETL, cette version commerciale apporte notamment :

⊕ La console Talend Administration Center

- gestion des référentiels des projets d'intégration, utilisateurs et droits associés
- ordonnancement des traitements (Job Conductor)
- console de monitoring AMC (Activity Monitoring Console) web
- gestion des reprises de traitements sur erreur d'exécution
- gestion des environnements d'exécution des traitements.



Job Conductor Talend - Activity Monitoring Console Talend

⊕ Autres fonctionnalités de productivités et d'exploitabilité

Cette version apporte également :

- le versionning des traitements
- la capacité de définir des points de reprise des traitements en cas d'erreur d'exécution
- un moteur de règles (Drools)
- joblets : morceaux de jobs réutilisables pour la factorisation des développements
- design de jobs à partir de templates
- visualisateur de données en sortie des composants
- change data capture

Jobs MapReduce et Spark

Cette version offre la possibilité de développer visuellement des traitements purement MapReduce ou Spark, dont l'exécution peut se faire sur un cluster Hadoop. L'exécution de jobs MapReduce depuis le studio offre un suivi d'avancement visuel de chaque étape map et reduce.

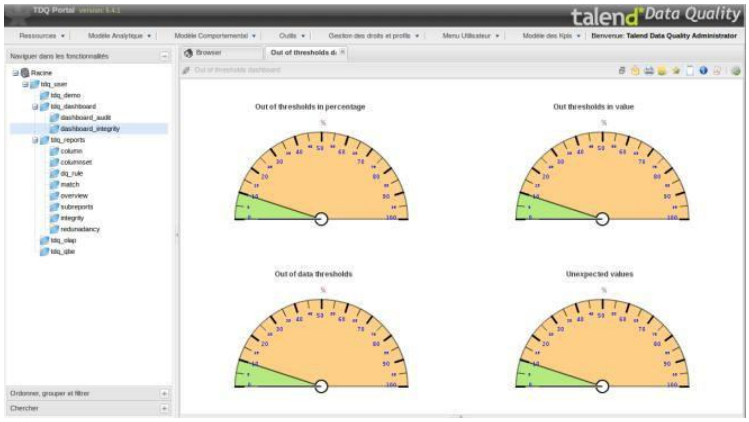
Talend Platform for Big Data

Cette version apporte notamment des fonctionnalités complémentaires et intégrées de qualité de données et de gestion de mapping complexes (XML, EDI) :

⊕ Profilage des données

Les analyses modélisées depuis le studio unifié, se font sur des sources, dont la définition peut être partagée avec les métadonnées définies au niveau de l'intégration. L'outil produit des métriques sur le taux d'unicité, de remplissage, la conformité à un format, la diversité des formats…
Des rapports, tableaux de bords et données requêtables peuvent être produits et publiés sur un portail décisionnel intégré (basé sur SpagoBI, présenté plus loin dans le

document) à partir des analyses de données afin de piloter le processus d'amélioration de la qualité des données :



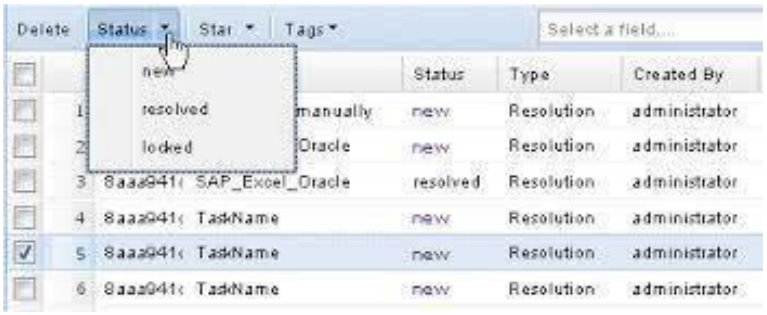
⊕ Composants de correction et enrichissement des données

Le studio de modélisation est enrichi de composants de traitement et correction supplémentaires de qualité des données :

- correction/enrichissement d'adresses postales via des services tiers QAS, Google
- rapprochements complexes en utilisant des technologies de logique floue
- création de tâches de correction manuelle des données.

⊕ Workflow web de correction des données

La solution intègre la console web Data Stewardship avec la définition de workflows de correction et validation de données :



Liste des tâches de correction/validation de données

Détail d'une tâche de résolution de données

VII.7 SUITE PENTAHO

Editeur : Pentaho Licence : Apache V2 et commerciale (suivant la version) Version actuelle : 5.4	

VII.7.a Présentation

Editeur et solutions

Pentaho est un éditeur basé en Floride et en Californie, avec des bureaux en France. L'éditeur est un acteur impliqué de l'open source, qui a rallié dès le début des produits open source comme Kettle ou Mondrian et qui anime sa communauté.

Au delà de la solution d'intégration de données, Pentaho fournit aussi une solution complète d'analyse et d'exploitation décisionnelle des données : Pentaho Business Analytics, présentés plus loin dans le document.

Pentaho et le Big Data

Pentaho a établi des partenariats avec des acteurs majeurs du Big Data, notamment : MongoDB, HortonWorks, Cloudera, MapR et DataStax.
L'éditeur publie également un portail web dédié aux problématiques Big Data : http://www.pentahobigdata.com

VII.7.b Fonctionnalités de l'ETL Pentaho Data Integration

Pentaho Data Integration (PDI) est un ETL qui permet de concevoir et exécuter des opérations de manipulation et de transformation de données.
Grâce à un modèle graphique à base d'étapes, il est possible de créer dans le studio de modélisation (Spoon), sans programmation, des processus composés d'imports et d'exports de données, et de différentes opérations de transformation (conversions, jointures, application de filtres, ou même exécution de fonctions Javascript si besoin).

Les fonctionnalités ETL classiques de Pentaho Data Integration sont présentées plus en détail dans le livre blanc Décisionnel.

PDI Community Edition

L'ETL Pentaho Data Integration propose des connecteurs aux principales Bases NoSQL/Big Data telles Hadoop (HDFS, HBase, Hive et MapReduce), Cassandra, CouchDb, MongoDB, ElasticSearch ainsi qu'aux bases de données Amazon S3 et aux réseaux sociaux Twitter et Facebook.

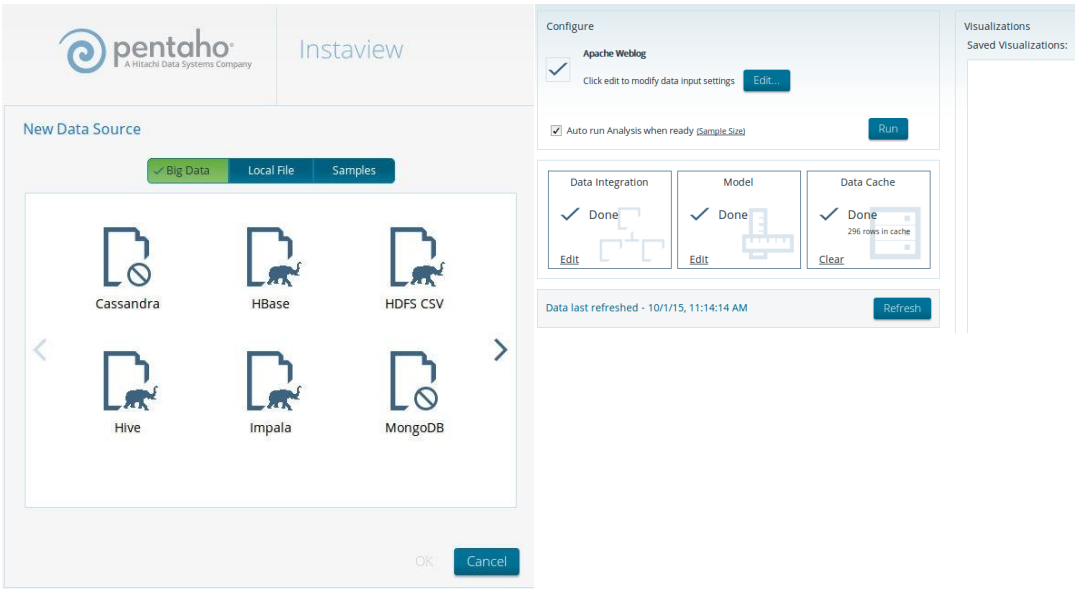
Pour les traitements en masse, la connectivité avec Hadoop Map-Reduce et le moteur MongoDB Map-reduce sont intéressants, tout comme les capacités de répartition de charge des traitements ETL dans une configuration cluster de PDI.

En sus des composants et techniques dédiés aux technologies Big Data, il y a d'autres options de PDI qui permettent une meilleure gestion de grosses volumétries de données :

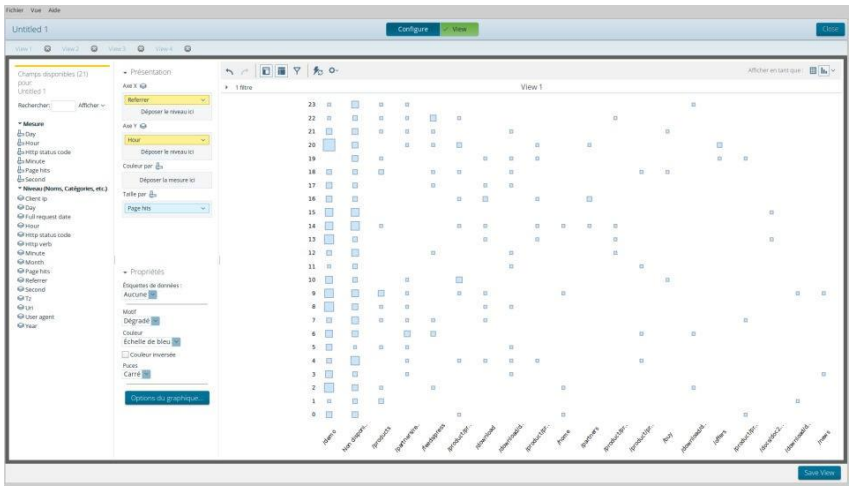
- lecture en parallèle de fichiers plats de grande taille tels que des fichiers de logs
- exécution concurrente de plusieurs copies d'une même étape d'une transformation avec distribution aléatoire en entrée des données en conséquence
- partitionnement, même option que la précédente avec une distribution plus intelligente des données à l'aide d'algorithmes proposés ou possibilité de développer des algorithmes de répartition spécifiques
- pour un environnement distribué, possibilité depuis la version 5.0 de faire du load balancing pour la distribution des données entre deux étapes d'une transformation.

PDI Enterprise Edition

La version Enterprise apporte plusieurs outils pour plus de productivité dans la manipulation de données massives.
Les **possibilités de visualisation et d'analyse intégrées Instaview** sont utiles aux Data Scientists[1] pour développer rapidement des applications analytiques Big Data, en limitant les allers-retours entre outils :



1 http://blog.smile.fr/Pentaho-4-8-l-analyse-instantanee-et-interactive-des-donnees-mobiles-et-big-data



Perspective Instaview de Pentaho Data Integration Enterprise Edition

En effet, dans le cadre de la méthodologie AgileBI, cette perspective intégrée au studio de modélisation des traitements ETL permet d'analyser avec l'outil Analyzer Pentaho des données, Big Data ou non, issues des transformations et mises en cache dans une base MongoDB.

⊕ Fédération de données

La version Enterprise propose également des possibilités de fédération de données au travers d'un connecteur JDBC. Ce dernier permet de projeter une transformation PDI comme source de données JDBC : cela ouvre des perspectives intéressantes de connectivité et de restitutions en quasi temps réel sur des processus métiers distribués au niveau applicatif.
Cela permet également de faire une interface entre des technologies Big Data, NoSQL et certains outils de restitutions plutôt orientés SQL (workbench/Mondrian). Et ainsi, permet d'éviter dans certains cas une structure de stockage hybride (NoSQL / SQL).

⊕ Connectivité Hadoop

Pentaho Data Integration propose une interface de paramétrage de cluster Hadoop :

⊕ Pentaho MapReduce

Pentaho MapReduce permet le développement de traitements MapReduce (mettant en œuvre une transformation pour l'étape map et une transformation pour l'étape reduce) depuis le studio de modélisation des traitements ETL.
Ils sont ensuite exécutables sur un cluster Hadoop.

⊕ Pentaho Predictive Analytics

En plus des méthodes d'analyse classiques (analyse d'événements passés et/ou présents), un des enjeux du Big Data notamment dans le domaine scientifique est de faire parler ces gros volumes de données pour de la prévision.

Weka est un projet data mining open source dont Pentaho est un acteur majeur, dans ce contexte de nombreux plugins sont disponibles par défaut ou non pour l'utilisation de certaines briques de Weka (Scoring, Knowledge Flow, ...) via Pentaho Data Integration.
Pour plus de précision sur les possibilités en termes de Data Mining via Pentaho, rendez-vous sur :
http://wiki.pentaho.com/display/DATAMINING/Pentaho+Data+Mining+Community+Documentation.

VII.7.c Fonctionnalités du portail Pentaho Business Analytics

Pentaho Business Analytics est un portail décisionnel qui permet la distribution d'outils d'analyse et requêtage en langage naturel ainsi que des documents décisionnels à un grand nombre de personnes par l'intermédiaire d'une interface web :

Page d'accueil de Pentaho Business Analytics

Pentaho est proposé en version communautaire et en version entreprise soumise à souscription annuelle, avec des modules supplémentaires (Pentaho Analyzer) ainsi qu'un support produit.

La communauté enrichit le portail en version communautaire sous forme de modules disponibles depuis le Pentaho MarketPlace, parmi lesquels l'interface d'analyse Saiku et les CTools qui ont le vent en poupe.

Pentaho fournit un portail décisionnel complet, permettant aux utilisateurs finaux :

- l'analyse multidimensionnelle : Pentaho Analyzer, Saiku Analytics
- le requêtage ad-hoc : Interactive Report, Saiku Reporting (une nouvelle version de Saiku Reporting compatible avec les nouvelles versions du portail Pentaho est annoncée par Meterit.bi), WAQR
- l'exploitation de tableaux de bords dynamiques (CTools).

Les capacités de répartition de charge (load balancing) entre plusieurs instances Pentaho Business Analytics sont intéressantes dans le cadre d'analyses en masses.

Connectivité NoSQL et exploitation de données massives

Pentaho fournit nativement des connecteurs Big Data au niveau des connections du portail pour les sources NoSQL offrant une connectivité JDBC :

- Hive
- Impala
- connecteur JDBC générique.

A noter qu'il est également possible d'accéder à d'autres sources de données NoSQL au sein du portail en passant par de la fédération de données, en utilisant l'ETL PDI ou un mécanisme de stockage externe d'une base relationnelle.

L'outil Pentaho Report Designer permet de plus d'élaborer et de publier des rapports à partir d'une source MongoDB.

VII.8 ELASTICSEARCH

Type NoSQL : document Editeur : Elastic Licence : Apache V2 Version actuelle : 1.7.2 (moteur ElasticSearch); 4 (Kibana)	elastic

L'éditeur Elastic (ex ElasticSearch) a publié une pile applicative avec :

- un moteur de recherche : ElasticSearch, propulsé par Apache Lucene et une base de données NoSQL documentaire.
- un module de chargement de données dans ElasticSearch à partir de logs (et autres sources avec le paramétrage de modules complémentaires) : Logstash
- un module de dashboard : Kibana, qui permet d'associer la puissance du moteur de recherche d'ElasticSearch (des recherches complexes peuvent être faites pour filtrer les données pertinentes à l'analyse) aux modules de reporting classiques.
- un connecteur Hadoop.

L'éditeur Elastic propose un service de support en production pour ces composants.

La notoriété et l'utilisation d'Elastic prennent de plus en plus d'ampleur, y compris en France. Deux exemples parmi tant d'autres, dans des contextes spécifiques :

- ɛLe moteurɛ, lemoteur.orange.fr, moteur d'indexation du web, développé par Orange, et refondu autour de la technologie Elastic[2], supportant plus de 150 millions de documents, avec des temps de réponse adaptés au web.
- Le plugin développé par Smile pour Magento et les sites e-commerce, remplaçant la fonctionnalité de recherche native avec une performance et surtout un niveau de pertinence optimisés et paramétrables en fonction de critères de comportement ou de contexte[3].

VII.8.a Moteur de recherche et base NoSQL ElasticSearch

Persistence

ElasticSearch permet la mise en cluster pour la réplication et la répartition de données
A noter que les indexes (de recherche/requêtage) générés sont de type colonne.

Connectivité, requêtage et traitement

L'accès et la manipulation de données se fait simplement via l'API REST et le format JSON.
Le moteur de requêtage propose des capacités d'agrégation et d'analyse, utile pour du requêtage décisionnel.

Usages Big Data BI

[2] https://www.elastic.co/blog/how-elasticsearch-helped-orange-to-build-out-their-website-search
[3] https://github.com/Smile-SA/smile-magento-elasticsearch

ElasticSearch peut servir d'Operating Data Store et à la mise en oeuvre de datamarts combinés avec des outils de restitution compatibles.

VII.8.b Portail Kibana

Exemple de tableau de bord Kibana

L'usage unique de Kibana est la publication de **tableaux de bords visuels, souples, hautement paramétrables par l'utilisateur final**, grâce aux fonctionnalités de recherche et de filtrage offertes par ElasticSearch.
L'outil propose un rafraîchissement automatique, adapté à des problématiques de monitoring de processus en temps quasi réel.

Le design des tableaux de bord se fait via l'insertion de panels (graphiques, listes, tendances, cartographies,...) dans une structure de type tableau. Un tableau de bord peut ainsi être bâti en quelques minutes. Les panels communiquent entre eux : recherche, zoom,...
Les tableaux de bord peuvent être enregistrés dans une base ElasticSearch afin d'être ré-exécutés et partagés.

Techniquement, le portail Kibana est maintenant motorisé par Node.js, avec une interface utilisateur écrite en javascript.
L'intégration avec le module Shield permet d'apporter une sécurité des accès à Kibana.

VII.9 JASPERSOFT

Editeur : JasperSoft Licences : GPL et commerciale (suivant la version) Version actuelle : 6.1	TIBCO Jaspersoft

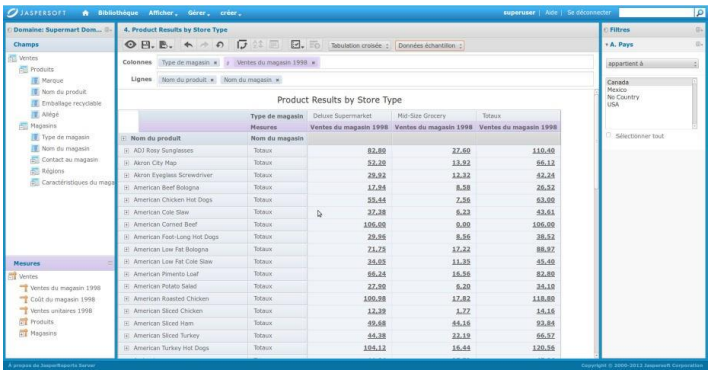
VII.9.a

JasperSoft BI Suite est la plateforme décisionnelle de TIBCO JasperSoft, société qui développe également le générateur d'états JasperReports, disponible depuis 2001. La plateforme propose des fonctionnalités de reporting et d'analyse et est disponible sous deux licences : GPL et commerciale.

VII.9.b Fonctionnalités

JasperServer, dans ses versions Professionnelle et Entreprise, offre des fonctionnalités supplémentaires par rapport à la version open source, limitée à la publication et la diffusion de rapports :

- outil de création de rapports ad-hoc en ligne (listes, graphiques ou tableaux croisés), accessible à tout utilisateur
- outil de composition de tableaux de bord.



Module de requêtage ad-hoc de JasperServer

Connectivité NoSQL et exploitation de données massives

JasperSoft BI fournit nativement, en versions commerciales Professionnal et Entreprise, un outil de requêtage et d'analyse ad-hoc qui permet une exploitation directe de sources de données NoSQL :

- MongoDB
- Hadoop via Hive

Un système de cache de données est présent, pour optimiser le temps de réponse des requêtes.

JasperSoft Studio fournit également une large palette de connecteurs au delà du JDBC classique pour le reporting et les tableaux de bord :

- MongoDB

- Hadoop via Hive
- Cassandra
- JSON.

Il existe aussi des connecteurs communautaires pour d'autres bases NoSQL, comme Google BigQuery ou Neo4j.

VII.10 APACHE ZEPPELIN

Editeur : NFLabs Licence : Apache 2.0 Version actuelle : 0.5	Apache Zeppelin

VII.10.a

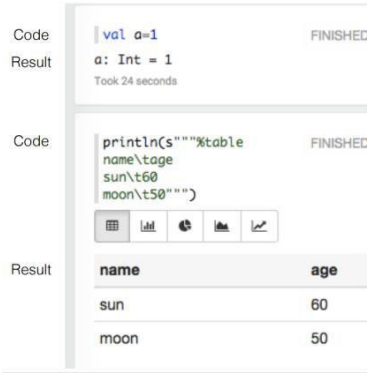
Zeppelin est une application permettant de représenter les données sous forme graphique et fonctionnant comme un carnet de notes. Zeppelin supporte plusieurs langages comme Scala (avec SparkContext). Il implémente Spark et d'autres implémentations sont possibles comme Hive, D3 ou Markdown.
Notons que Zeppelin ne s'adresse pas aux utilisateurs finaux car il nécessite une connaissance de certains langages, mais plutôt à des data scientists/analysts ou à des développeurs.

Techniquement, Apache Zeppelin est basé sur une architecture web solide avec d3.js, grunt, bower et AngularJS. La communication client/serveur se fait via Http REST/Websocket. La gestion des dépendances est réalisée avec Maven.
Apache Zeppelin s'intègre avec Apache Spark et bien d'autres interpréteurs dont:

- PySpark
- Hive
- Mysql (JDBC)
- Markdown
- Shell
- SparkSQL.

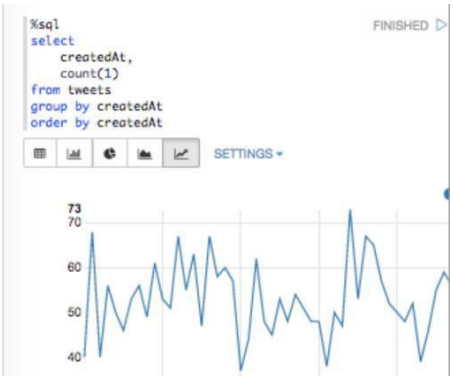
VII.10.b Fonctionnalités

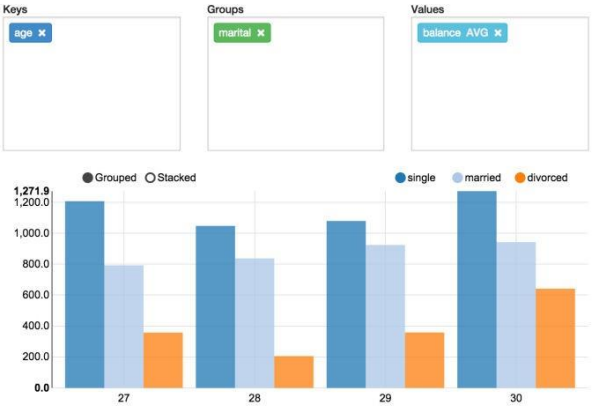
Carnet de note



Visualisation de données et pivot

Zeppelin permet de transformer directement le résultat de requêtes en graphiques, ici avec une requête SQL :





Formulaires dynamiques

Zeppelin permet de créer des paramètres, utilisables directement dans les tableaux de bord :

VII.11 SPAGOBI

Editeur : Engineering Group / OW2 Consortium Licence : Mozilla Public License V2 Version actuelle : 5.1	spagobi

VII.11.a

SpagoBI est une suite décisionnelle uniquement distribuée sous licence open source, développée par la société italienne Engineering Ingegneria Informatica au sein du consortium OW2.

VII.11.b Fonctionnalités

Afin de couvrir les différents besoins fonctionnels propres à la valorisation et l'analyse de données, SpagoBI propose une vingtaine de modules (ou « moteurs ») complémentaires, offrant des fonctionnalités de reporting/dashboarding, requêtage et analyse OLAP ad-hoc, geoBI, KPI et datamining :

Exemples de restitutions SpagoBI

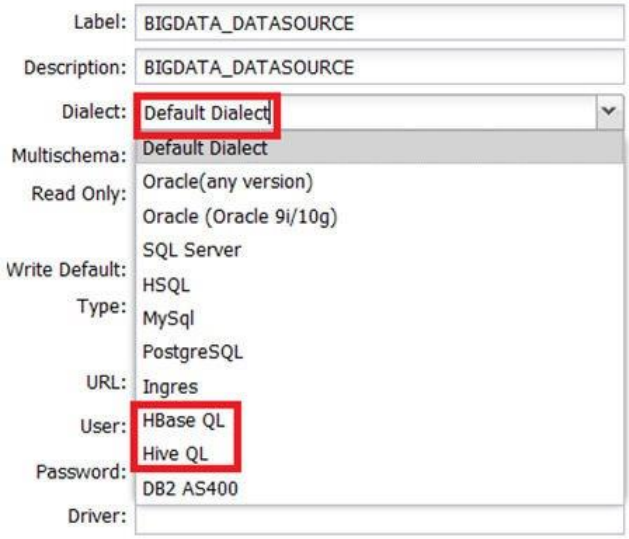
Ces modules s'appuient sur un ensemble de projets open source phares, offrant ainsi une grande richesse de modules fonctionnels : l'ETL Talend, le moteur OLAP Mondrian, les moteurs de reporting BIRT et Jasper, R et weka datamining.

Modules de SpagoBI

⊕ Connectivité NoSQL et exploitation de données massives

Afin de répondre à la problématique du Big Data, SpagoBI a développé de nouveaux connecteurs permettant le requêtage de bases de données NoSQL via des datasets :

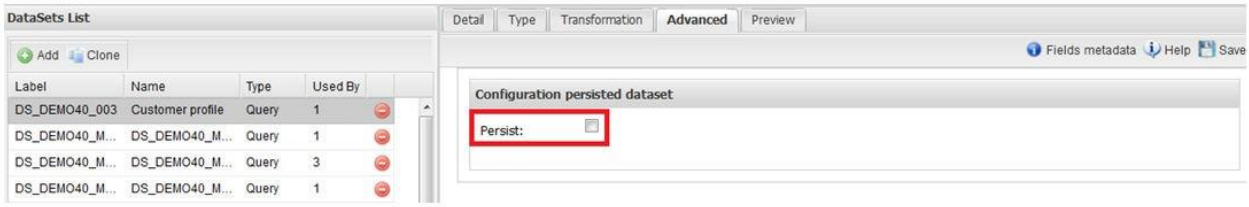
- HBase: développement de requête HBQL, langage de requête Hbase, intégré nativement dans SpagoBI
- Hive: développement de requête HQL, langage de requête Hive très proche du SQL, intégré nativement dans SpagoBI
- Impala: connecteur Cloudera Impala JDBC, rendu disponible par Cloudera
- Cassandra: développement de requêtes CQL, langage de requête Cassandra.



Sélection du langage d'un connecteur

Dans la version 4 de SpagoBI, la définition de dataset a évolué afin de permettre des temps de réponses plus courts sur les larges volumes de données :

- possibilité de planifier l'alimentation des datasets pour une restitution différée
- possibilité de définir des datasets persistants où les données sont stockées en cache.



Définition d'un dataset persistant

SpagoBI travaille actuellement à introduire les problématiques d'accès en temps réel ainsi que la mise en place d'une couche sémantique sur les données Big Data.

N'hésitez pas à nous transmettre vos avis et évaluations sur ce livre blanc.
Une seule adresse : contact@smile.fr

Edition octobre 2015