http://numerique.anap.fr/publication/1505-big-data-en-sante-donnees-concernees-usages-entrepot-bio-heterogenes-et-outils-d-exploitation

Avis d'experts

Big Data en santé : données concernées, usages, entrepôt bio-hétérogènes et outils d'exploitation

Cet avis d'expert a été rédigé par Hélène SOL.

Périmètre traité dans ce propos

Les aspects juridiques

Ce propos ne vise pas à traiter les aspects juridiques liés à l'exploitation des données de santé « sensibles », vaste sujet, très réglementé. Il vise plutôt à décrire une méthode, une organisation projet et lister les outils pouvant permettre de démarrer l'exploitation de big data en santé. Cependant, une réponse « pratique » peut être apportée à la question : « *Quel est le support légal pour les utiliser ?* » : pour des raisons « éthiques », l'établissement de santé se doit d'héberger les données qu'il gère et qu'il génère, surtout s'il s'agit de données issues de la recherche et de protocoles thérapeutiques. Les utilisateurs médecins rencontrés sur le sujet n'acceptent de partager leurs données qu'avec la certitude qu'elles seront gardées confidentielles, à savoir hébergées au sein de l'ES, et en sécurité (accès très restreints). Il n'est pas envisageable à ce jour pour eux d'accepter d'être hébergé en mode externalisé chez un Hébergeur de Données de Santé certifié. Ils exigent que les données restent souveraines.

Avant-propos

Le propos dans ce texte sera orienté Big Data. Le terme de « **donnée calculée** » sera vu comme la manière dont les NTIC peuvent exploiter et donner du sens à la donnée brute.

« *Les Big Data sont avant tout un dispositif technologique dont l'objectif est de transformer une donnée brute en connaissance directement exploitable, jusqu'au fondement de la méthode scientifique, par une structure* » (Guillaud, 2011).

Notre propos évoquera le rassemblement de grands volumes de données dans les Big Data et leur analyse à l'aide des nouvelles techniques de data mining.

Les données exponentielles :

Avec la croissance d'Internet, de l'usage des réseaux sociaux, de la téléphonie mobile, du Cloud Computing[1], des objets connectés et communicants, l'information est aujourd'hui plus abondante que jamais et sa croissance est chaque jour plus rapide. L'entreprise Ericsson prédit qu'il y aura 50 milliards d'objets connectés dans le monde d'ici à 2020. L'Organisation des Nations unies (ONU) a estimé que plus de données ont été créées en 2011 que dans toute l'histoire de l'humanité. La masse de données numériques est passée de 480 milliards de gigaoctets en 2008 à 2,72 zettaoctets en 2012. Jusqu'en 2020 cette masse va continuer à progresser à une vitesse exponentielle pour atteindre les 40 zettaoctets. D'après le rapport d'étude de Global Investor (Crédit Suisse), le monde digital aura grandi en 2020 de 300 fois sa taille de 2005, grâce notamment à l'émergence des objets connectés. Chaque jour, l'humanité génère 2,5 trillions d'octets de données. **Le monde crée en 2 jours autant de données que toute l'humanité en a créées pendant 2000 ans. 90 % des data dans le monde ont été créées durant les deux dernières années**. Ainsi, le volume de données produites chaque année dans le monde devrait être multiplié par 44, d'ici 2020. La quantité d'information archivée croît 4 fois plus vite que l'économie mondiale, pendant que la puissance de traitement informatique croît 9 fois plus vite.

Le volume total des données **d'e-santé** dans le monde double tous les 73 jours. Cette importante volumétrie de données ouvre dès lors le champ aux systèmes experts.

L'information immatérielle et multiforme

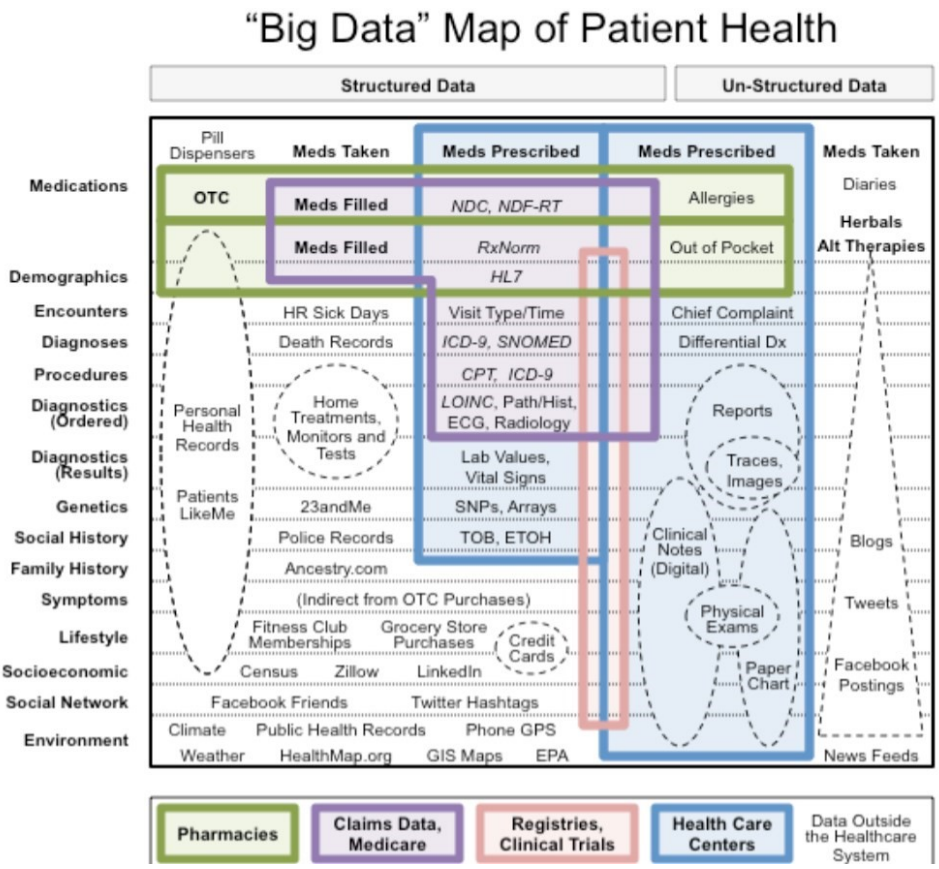
La communication et l'échange de données ont toujours motivé l'homme. Depuis les codes morse en passant par des fils télégraphiques, la société est désormais propulsée dans l'ère du numérique où jpeg et mp3 se côtoient dans les réseaux d'Internet où l'information est devenue plus que jamais immatérielle. L'homme ne cesse de d'innover en nouvelles technologies qui favorisent la circulation d'une information multiforme : celle-ci nécessite moins de place et devient plus facile à déplacer, à partager, à analyser... Les nouvelles technologies de l'information et de la communication (NTIC) jouent un rôle majeur et central notamment dans le domaine de la santé.

Contexte : la mine d'or que constituent les SIH

Les Systèmes d'Information Hospitalier représentent une mine d'or : les établissements de santé stockent et traitent de très gros volumes de données qui constituent un réservoir extrêmement riche, en particulier de données cliniques : résultats patients, données des essais cliniques, données génétiques, données biocliniques, données pathologiques, prescriptions de pharmacie, résultats de laboratoires, données médicales de toute sorte... Les suivis des patients à domicile, par exemple dans le cadre de l'HAD, permettront d'enrichir ce réservoir de nouvelles données recueillies par les appareils connectés aux patients à domicile.

Les technologies modernes, smartphones, puces génomiques, senseurs GPS pour mesurer les déplacements ou l'activité des patients, sont la source d'une énorme quantité d'informations utilisables en épidémiologie, informations qu'il aurait été impossible de recueillir par des approches traditionnelles. Les données disponibles couvrent un champ de plus en plus vaste : régime alimentaire, pollution sous toutes ses formes, mode de vie, hygiène de vie, déplacements, infections, traitements médicamenteux, stress, etc...

Schéma 1 : La carte des données de santé d'un patient



Malheureusement, ces données sont encore souvent peu exploitées.

Les Big Data...

Les nouvelles technologies informatiques favorisent le stockage de ces données multiformes dans un entrepôt dédié, que l'on peut désormais qualifier de Big Data[2]. Le terme de Big Data désigne une nouvelle discipline qui se situe au croisement de plusieurs secteurs tels que les technologies, les statistiques, les bases de données et les métiers en santé. Cette approche est une réponse à l'explosion des données non structurées voire parfois multi-structurées observée dans l'univers numérique (Internet, RFID, mobile).

Cette activité permet de capter les données numériques, de les traiter à très grande vitesse et de les rendre ainsi exploitables quelle que soit la nature de cette donnée. De plus, conformément à la loi de Moore (1985), la technologie permet désormais de pouvoir traiter ces millions d'informations en quelques instants ; des outils informatiques existent (voir plus bas) : pour traiter une énorme quantité de données, même non structurée, en des temps record.

… Dans la santé…

Aujourd'hui, la médecine moderne est devenue presque inconcevable sans l'utilisation des données personnelles numérisées. L'émergence de la e-santé, la télémédecine, la m-health, les NBIC (nanotechnologies, biotechnologies, informatique, et sciences cognitives), et des Big Data modifient la prestation de santé, la relation médecin patient, et la compréhension scientifique du corps humain et des maladies. Le temps est maintenant venu de favoriser l'accès à ces données massives en Santé et l'inter-opérabilité des systèmes d'informations afin de mettre en place des « centres de données cliniques » (appelés plus bas « Entrepôt bio-hétérogène ») , et de permettre les croisements de données de santé et de recherche qui permettront des analyses multiparamétriques corrélant les données épidémiologiques, médico-technique, cliniques, issues des capteurs, de recherche clinique, voire de recherche fondamentale.

… Pour la médecine de demain …

La prise en compte de l'ensemble de ces données dans les études épidémiologiques suscite des attentes et des espoirs en termes de compréhension des causes et des mécanismes des maladies comme pour la personnalisation du suivi médical. De par la prédictivité offerte par nouveaux outils mise en place sur les Big Data, une pratique proactive de la médecine est amenée à se développer, intégrant l'analyse complexe sur les multiples données disponibles : biologiques, pathologiques, leur évolution, les données « environnementale » (climat, pollution…), les données relevées par des applications dites « de santé » connectées aux patients… Le croisement de toutes ces informations et les calculs poussés d'indicateurs, permettront d'orienter la médecine vers des axes thérapeutiques novateurs. Ces axes préfigurent la médecine de demain : exploiter ces Big Data de manière souveraine rejoignent pleinement les concepts de la Médecine 4P, médecine systémique qui est intrinsèquement : **P**rédictive et **P**réventive -- agissant par anticipation avant l'apparition de symptômes – et qui est également **P**ersonnalisée et **P**articipative -- adaptant interventions et traitements aux caractéristiques et réactions individuelles --.

Quels usages de l'exploitation des données stockées dans le SIH ?

Les usages :

Du point de vue médecin, il y a des domaines très différents :

- L'aide au diagnostic.
- La médecine personnalisée.
- La recherche épidémiologique à l'échelle de la société.
- …

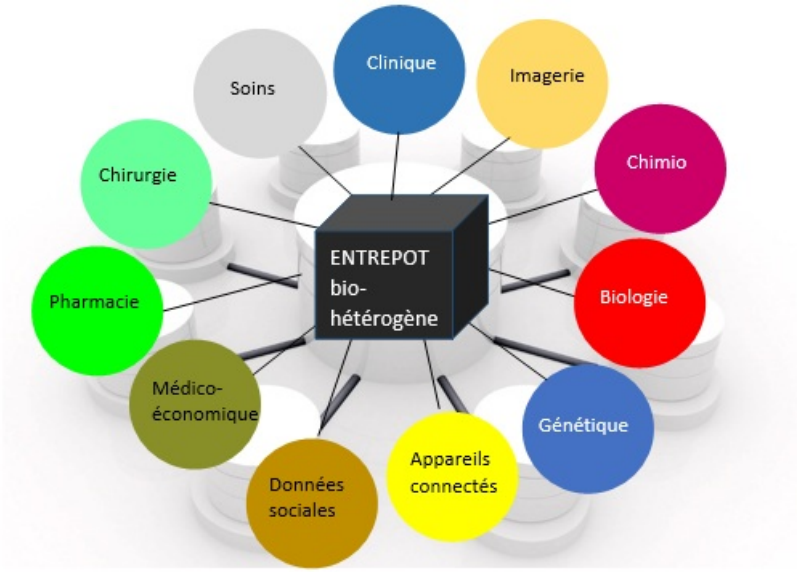
Les populations :

Les usages de l'exploitation de ces données multiparamétriques concernent plusieurs populations :

- Les professionnels de la santé, avec la possibilité d'un diagnostic plus rapide et précis, voire d'un traitement totalement personnalisé pour les patients.
- Le patient, afin de lui appliquer une médecine dorénavant ciblée et personnalisée rejoignant les principes de la médecine 4P.
- Les collectivités afin de remplir leur rôle de vigilances et de veille sanitaire.
- La recherche :
  - La recherche épidémiologique pour la détection de cohortes.
  - La recherche clinique pour les études de faisabilité et la recherche de patients éligibles.
  - L'exploitation de ces données massives peut contribuer au développement de nouveaux produits de santé, détection des signaux faibles lors d'épidémies ou d'effets indésirables graves…
- Les services médico-administratif afin de proposer l'organisation raisonnée des soins, le pilotage des activités, l'analyse des trajectoires de santé.
- L'enseignement : aux professionnels de santé, l'information de citoyens, l'éducation thérapeutique

Les solutions : comment intégrer et exploiter des données patient hétérogènes ?

Schéma 2 : créer un entrepôt de données bio-hétérogène



L'entrepôt intègre les données cliniques hétérogènes depuis leur source pour permettre des analyses trans-domaines.

Les outils : comment, qui peut et avec quoi traiter ces données ?

Les Big Data permettent de connecter de nombreux flux de données. Ils promettent de réconcilier, d'harmoniser, d'unifier, d'interconnecter, et de fluidifier d'importants volumes de données digitales, dans un monde où tout est désormais numérisé. **Différents outils technologiques permettent de traiter tous types de données, dans de considérables quantités, et en un temps limité.**

Comment ?

Afin d'exploiter ces immenses volumes de données brutes et hétérogènes, nous devons nous tourner vers les nouvelles sciences des données. La valorisation des Big Data passe par la mise en place **d'analyses sophistiquées** de type **vectorielles** et la mise en œuvre **d'algorithmes mathématiques**, déclinés par les outils cités ci-dessous. Les outils sont les mêmes quels que soient les usages.

Qui ?

Le profil idéal pour traiter ces données est constitué d'une triple compétence :

1. « Ingénieur en informatique » : capable de cibler et d'extraire les données pour les placer dans l'entrepôt.
2. « Médecin » : pour identifier si les corrélations établies par les outils informatique de datamining sont légitimes et exploitables.
3. « Mathématicien » et « Statisticien » : pour orienter les outils de datamining vers de nouveaux modèles d'analyses et valider les corrélations établies.

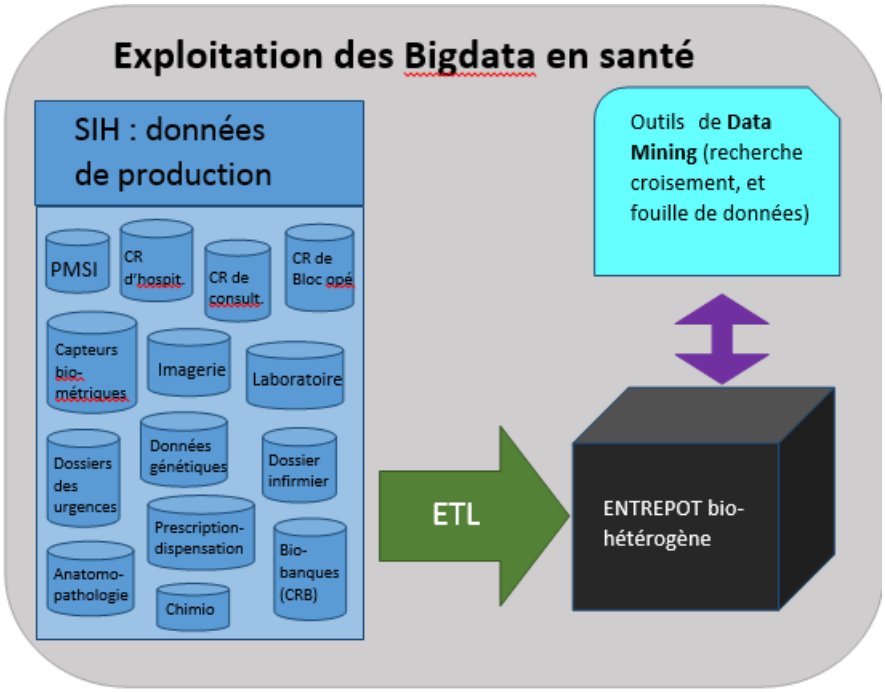
Avec ces profils et à partir d'outils ad hoc, il sera possible désormais de détecter et d'optimiser, de tracer et de cibler, voire de prédire et de prévoir des informations précises. **Le pilotage opérationnel**, voir la **décision médicale**, grâce à l'approche métier du médecin, seront également grandement améliorés : l'afflux et le croisement de data multiparamétriques en temps réel permettront une compréhension plus fine de l'environnement. La prise de décision et les diagnostics du médecin seront étayés (éclairés), et les actions des services seront pilotés plus efficacement à partir d'éléments tangibles. De plus, la granularité et le large spectre de sources des données étudiées autoriseront la vision rassemblée mais aussi le forage de données à un niveau très fin selon le cheminement en axes d'analyse de l'esprit humain.

Les profils d'intervenants sont également les mêmes sont les mêmes quels que soient les usages, le respect de la pluridisciplinarité est le facteur de réussite.

Avec quoi (quels outils ?)

Les nouveaux logiciels des NTIC ont la capacité de détecter l'information intéressante pour obtenir un traitement optimum des données : c'est le concept de **Data Mining**. Cette approche utilise une **méthode inductive** et non plus déductive, en cherchant à **établir des corrélations** entre plusieurs informations disparates **sans hypothèses prédéfinies**. Les techniques d'« **analytics avancés** »[3], s'appuient sur ces larges volumes de données pour chercher des « signaux faibles »[4], au sein d'une arborescence de catégories identifiées.

Schéma 3 : l'exploitation des Big Data en santé



Il s'agit de copier, via des outils d'ETL, les données de production en améliorant la qualité des données : sélectionner la donnée « propre » la plus pure.

Un hôpital peut simplement faire naître son entrepôt de données s'il en a le souhait, la motivation et les compétences. Un éditeur ne peut pas s'opposer à ce que l'Etablissement de Santé (ES) s'introduise, recopient et forent les données internes à son logiciel, partant du principe que les données produites sur les patients de l'ES appartiennent à cet ES ; il en a besoin pour soigner, elles lui sont présentées sous forme d'interface Homme Machine, mais il en a aussi besoin pour sa facturation (transmises sous forme d'interfaces), pour les échanges avec ses autres logiciels (générées sous forme de comptes-rendus électroniques), mais également pour son pilotage et son évolution (son progrès). Dans la plupart du cas, aucun outil d'extraction n'est fourni par l'éditeur. Un éditeur, au mieux, fournira, dans les livrables, le dictionnaire des données de sa base de données (à demander dans les cahiers des charges d'acquisition logiciel), afin de faciliter cet accès, et au pire, rien du tout.

La récupération technique des données se fait donc, quasiment dans tous les cas, grâce à l'ETL, par des informaticiens « futés » et très bons techniciens, jusqu'au-boutiste, capables de trouver les astuces pour accéder aux bases de données des différents logiciels et des différents éditeurs. Croyez-moi, cela est tout à fait réalisable. Toutes les données internes sont extractibles par un bon informaticien. La principale difficulté qu'il rencontrera est le non respect des règles de l'art –de la part de l'éditeur– dans l'architecture de la base de données, ne respectant pas les formes dites « normales », en sus avec des noms de colonnes au nom peu évocateur (exemple : la colonne « Nom_patient » qui porte le nom « Col1 ») et des tables peu parlantes avec des données répétées en différents endroits.

En effet, comme dans le bâtiment, un architecte en base de données doit être affecté lors de la conception et l'évolution d'un logiciel (ainsi qu'un concepteur de normes et procédures de développement), ce qui est rarement le cas. Les bases de données mal conçues sont fréquentes, en outre, dans les solutions logicielles développées par croissance externe, et /ou sur différentes technologies.

Le nom des outils

Les outils à l'œuvre pour l'exploitation des Big Data constituent la réelle innovation de ces dernières

années ; citons :

- **Hadoop** : « *framework libre et open source écrit en Java destiné à faciliter la création d'applications distribuées (au niveau du stockage des données et de leur traitement) et échelonnables (scalables) permettant aux applications de travailler avec des milliers de nœuds et des pétaoctets de données* ». https://fr.wikipedia.org/wiki/Hadoop

  Hadoop est devenue la référence en matière de parallélisation des Big Data, customisée par les Big Four américains GAFA (Google, Apple, Facebook, et Amazon), les trois géants chinois BAT (Baidu, Alibaba, et Tencent), et les quatre grandes entreprises emblématiques de la disruption numérique à savoir NATU (Netflix, Airbnb, Tesla, et Uber) pour mettre en place des technologies de promotion et de ciblage de leurs utilisateurs.

- **MapReduce**[5]: « *patron d'architecture de développement informatique, inventé par Google, dans lequel sont effectués des calculs parallèles, et souvent distribués, de données potentiellement très volumineuses, typiquement supérieures en taille à 1 téraoctet* ». https://fr.wikipedia.org/wiki/MapReduce

Une des autres principales caractéristiques de l'émergence des Big Data est constituée des systèmes de gestion de bases de données non relationnelles distribuées, employant des modalités de requête **NoSQL**[6] (Not Only SQL), qui dépassent donc les codifications du langage SQL (Structured Query Language). Ces bases de données NoSQL sont classées selon la manière dont elles stockent les données. On retrouve ainsi des catégories comme les bases orientées clé-valeur, document, colonne ou encore les bases structurant les données en se fondant sur **la théorie des graphes et la théorie vectorielle**. On peut donner comme exemples des solutions telles que : SenseiDB ; Voldemort (LinkedIn) – Cassandra ; Hive ; HBase (Facebook) – Dynomo ; S3 (Amazon) – CouchDB (Ubuntu One) – MongoDB (SourceForge.net) – MapReduce ; BigTable (Google) – No Database SQL (Oracle) – Storm ; FlockDB (Twitter) ; S4 (Yahoo).

- **Cassandra** : « *Apache Cassandra est un système de gestion de base de données (SGBD) de type NoSQL conçu pour gérer des quantités massives de données sur un grand nombre de serveurs, assurant une haute disponibilité en éliminant les points individuels de défaillance* ». https://fr.wikipedia.org/wiki/Cassandra_(base_de_donn%C3%A9es)

Quelques concepts

- Du fait de la sensibilité des données exploitées, le Big Data doit rester souverain et être hébergé dans le SIH de l'établissement
- Dé-identification des données
- Traçabilité des accès
- Conservation du contexte : dualité Document / Données atomiques
- Traitement automatique des données textuelles
- Extraction des concepts
- Détection de la négation, de l'incertitude
- Expansion sémantique

Quelques caractéristiques techniques

- Bases de données NoSQL déjà utilisées dans le monde hospitalier : No Database SQL (Oracle) et MongoDB (SourceForge.net)
- Intégration basée sur les standards d'interopérabilité et sur les profils IHE : HL7, PN13, HPRIM
- Nomenclatures médicales intégrées : CIM10, ADICAP, SNOMED, LOINC
- Exemples d'ETL industriels : ENOVACOM ou TALEND
- Module d'analyse et de fouille de données : R
- Outils intégrés : I2B2 / SHRINE : https://www.i2b2.org/ https://catalyst.harvard.edu/services/shrine/

Retrouvez le catalogue des productions des experts HN.

[1]Le Cloud Computing, ou « informatique des nuages » en français, désigne la représentation métaphorique de l'ensemble des services, proposant, à travers une connexion Internet, l'entreposage de fichiers ou la gestion de contenu en passant par l'exécution d'applications en ligne.

[2]Le terme Big Data est un nom anglais traduit par « grandes données » ou « méga données » ; le mot data est le pluriel latin de datum. Les bigdata, en sciences de l'information, décrivent le recueil et la gestion de bases de données se distinguant par un volume important, une variété des types de données de sources hétérogènes et une grande vitesse de génération.

[3]Le terme « analytics avancés » comprend les techniques et les méthodes suivantes : statistiques non-paramétriques, réduction de dimension, règles d'association, analyse de données réticulaires (network analysis), classification non supervisée (cluster analysis), algorithmes génétiques, etc...

[4]Le signal n'est pas faible par la nature de la source d'information (formelle ou non) mais par le rattachement entre cette source et une entité en mesure de prendre une décision après avoir mis en relation le signal et un scénario. Un signal faible est difficilement interprétable, informel, improbable mais généralement annonciateur d'un évènement à venir.

[5]Le MapReduce est une technique qui segmente le traitement d'une opération (appelée « job » chez Hadoop) en plusieurs étapes, dont deux élémentaires, afin de faciliter la parallélisation des traitements sur les données : le « Mapping » et le « Reducing »

[6]Le terme « NoSQL » désigne une catégorie de systèmes de gestion de base de données (SGBD) destinés à manipuler des bases de données volumineuses pour des sites de grande audience. Ces SGBD ne sont plus basés sur l'architecture classique de bases relationnelles.

Ressources associées

DÉMARCHE
D2 - Dossier patient informatisé interopérable

DÉMARCHE
D3 - Prescription électronique alimentant le plan de soins

DÉMARCHE
D1 - Résultats d'imagerie, de laboratoire et d'anapath

DÉMARCHE
D4 - Programmation des ressources

DÉMARCHE
Pilotage médico-économique

PERSONNE RESSOURCE
Elise MORICHON

PERSONNE RESSOURCE
Katia LE NEDIC

PERSONNE RESSOURCE
Jérôme CHAUVET

Glossaire

datamining
ES
ETL
HL7
HAD
IHE
interface
loi
pharmacie
pilotage
PN13
SIH

Date de parution : 07/01/2016