**Issae-Cnam liban.**

**Département :Génie informatique.**

**Professeur :Fares Pascal.**

**Etudiant :Haidar Mohammad.**

**Matière :Construction rigoureuse 1.**

**Sujet :Big Data.**

2013/2014

**Indexe :**

1-Définition…………………………………………………………………………

2-Dimensions des big data………………………………………………….

3-Informatique décisionnelle………………………………………………

4-Stockage………………………………………………………………………….

5-Applications des big data…………………………………………………

6-Evolutions……………………………………………………………………….

7-Big Data et sécurité…………………………………………………………

Définition

**Définition :**

Chaque jour, nous générons 2,5 trillions d’octets de données.Les big data, littéralement les « grosses données », ou volumes massifs des données, désignent des ensembles de données qui deviennent tellement volumineux comme les messages sur les sites de médias sociaux, d'images numériques et de vidéos publiées en ligne, d'enregistrements transactionnels d'achats en ligne et de signaux GPS de téléphones mobiles, qu'ils en deviennent difficiles à travailler avec des outils classiques de gestion de base de données ou de gestion de l'information.

Dimensions des big data

Le Big Data s'accompagne du développement d'applications à visée analytique, qui traitent les données pour en tirer du sens. Ces analyses sont appelées Big Analytics ou «broyage de données». Voici un rapport de recherche en 2001 du META Group (devenu Gartner) définit les enjeux inhérents à la croissance des données comme étant tridimensionnels : les analyses complexes répondent en effet à la règle dite «des 3V» (volume, vélocité et variété). Ce modèle est encore largement utilisé aujourd'hui pour décrire ce phénomène.En plus,on peut ajouter une autre dimension aux «3V»,c’est la véracité. Voici une couverture générale de Big Data par quatre dimensions : volume, vélocité, variété et véracité.

**Volume :**

Le volume des données stockées aujourd’hui est en pleine expansion. les entreprises sont submergées de volumes de données croissants de tous types, qui se comptent en téraoctets, voire en pétaoctets. Selon une étude IDC sponsorisée par EMC Gartner, les données numériques créées dans le monde seraient passées de 1,2 zettaoctets par an en 2010 à 1,8 zettaoctets en 2011, puis 2,8 zettaoctets en 2012 et s'élèveront à 40 zettaoctets en 2020. À titre d'exemple, Twitter générait en janvier 2013, 7 téraoctets de données chaque jour et Facebook 10 téraoctets.

**Variété :**

Le volume des Big Data met les data centers devant un réel définition: la variété des données. Il ne s'agit pas de données relationnelles traditionnelles, ces données sont brutes, semi-structurées voire non structurées (cependant, les données non-structurées devront, pour utilisation, être structurées). Ce sont des données complexes provenant du web (Web Mining), au format texte (Text Mining) et images (Image Mining). Elles peuvent être publiques (Open Data, Web des données), géo-démographiques par îlot (adresses IP).Ce qui les rend difficilement utilisables avec les outils traditionnels. De nouvelles connaissances sont issues de l’analyse collective de ces données.

 Utiliser les centaines de flux vidéo des caméras de surveillance pour contrôler les points d'intérêt.

 Tirer parti de la croissance de 80 % du volume de données image, vidéo et documentaires pour améliorer la satisfaction client.

**Vélocité :**

La vélocité représente à la fois la fréquence à laquelle les données sont générées, capturées et partagées.Des flux croissants de données doivent être analysés en temps réel(Data Stream Mining) pour répondre aux besoins des processus chrono-sensibles. Pour les processus chronosensibles tels que la détection de fraudes, le Big Data doit être utilisé au fil de l'eau, à mesure que les données sont collectées par votre entreprise afin d'en tirer le maximum de valeur.

* Scruter 5 millions d'événements commerciaux par jour afin d'identifier les fraudes potentielles.
* Analyser en temps réel 500 millions d’enregistrements détaillés d’appels quotidiens.

**Véracité :**

La Véracité traduit la mise en exergue du besoin de disposer de données de qualité dans un système Big Data. Cette nécessaire qualité sert tous les métiers de l’entreprise, à commencer par le marketing qui l’utilise pour mieux cibler les campagnes, pour mieux adresser les clients, pour être plus pertinent.

**En plus:**

Selon IBM Global Entrepreneur France,elle visualise la regle **‘‘**les 3 V sont morts, vive les 5 V !’’.Elle ajoute la “cinquieme V” aux “4V”precedents,c’est la visibilite.

La Visibilité offerte par une plate-forme Big Data est la faculté de comprendre ce que l’on consulte et de trouver les bonnes données en naviguant au sein d’une plate-forme Big Data fédérée.

Différence avec l'informatique décisionnelle

Si la définition du Gartner en 3V est encore largement reprise, la maturation du sujet fait apparaitre un autre critère plus fondamental de différence d’avec l'informatique décisionnelle et concernant les données et leur utilisation :

**Informatique décisionnelle** : utilisation de statistique descriptive, sur des données à forte densité en information afin de mesurer des phénomènes, détecter des tendances… ;

**Big Data**:utilisation de statistique inférentielle, sur des données à faible densité en information dont le grand volume permet d’inférer des lois (régressions….) donnant dès lors (avec les limites de l’inférence) au big data des capacités prédictives.

**Stockage**

Pour répondre aux problématiques Big Data l’architecture de stockage des systèmes doit être repensée et les modèles de stockage se multiplient en conséquence. Les solutions de stockage Big data **commencent à être adoptées par les entreprises, qui les utilisent sur leurs périmètres stratégiques** pour en tirer **des gains de performance** là où les solutions standards étaient limitantes. **L’intégration de ces solutions reste l’étape clé** : l’étude des données est nécessaire pour choisir la bonne implémentation Big data, faciliter sa configuration et ainsi profiter pleinement des  bénéfices announces.

Voici ces 3 modèles :

le cloud computing : Les applications et les données ne se trouvent donc plus sur les réseaux internes, mais dans le nuage, et l’accès aux données, aux services et à l’infrastructure se fait à distance.

Ce mode de traitement des données implique *grosso modo*que les logiciels, la mémoire ou les capacités de calcul soient utilisées en réseau selon les besoins, par ex. sur Internet ou au sein d'un réseau privé virtuel. En d'autres termes, ces capacités sont **louées**: l'environnement informatique (centre de calculs, espaces de stockage, logiciels de messagerie et de collaboration, environnements de développement et logiciels spéciaux tels que la gestion des relations avec la clientèle) n'est plus la propriété de l'entreprise ou de l'autorité et n'est plus géré par celles-ci, mais est loué à un ou plusieurs prestataires de services.

les super calculateurs hybrides : Les HPC pour High Performance Computing, qu’on retrouve en France dans les centres nationaux de calculs universitaire tels quel’IDRIS, le CINES, mais aussi au CEA ou encore le HPC-LR.

Les systèmes de fichiers distribuées (DFS - Distributed files system): les données ne sont plus stockées sur une seule machine car la quantité à stocker est beaucoup trop importante. Les données, les fichiers sont "découpés" en morceaux d'une taille définie et chaque morceau est envoyé sur une machine bien précise utilisant du stockage local. Le stockage local est préféré au stockage SAN/NAS pour des raisons de goulots d'étranglement au niveau du réseau et des interfaces réseaux des SAN. De plus, utiliser un stockage de type SAN coute bien plus cher pour des performances bien moindres. Dans les systèmes de stockage distribué pour le Big Data on introduit le principe de "Data locality". Les données sont sauvegardées là où elles peuvent être traitées.

Applications des big data

Les Big Data trouvent une application dans de nombreux domaines : de grands programmes scientifiques (CERN28Mastodons), de grandes entreprises (IBM29,Amazon Web Services, BigQuery, SAP HANA) des entreprises spécialisées (Teradata, Jaspersoft30, Pentaho31...) de l'Open Source (Apache Hadoop, Infobright32, Talend33...) et des Start-up (aleph-networks,Bionatics, Hariba Médical, SafetyLine, KwypeSoft, Vigicolis, PredicSis):

**Recherche scientifique :**

Les expériences du Large Hadron Collider représentent environ 150 millions de capteurs délivrant des données 40 millions de fois par seconde. Il y a autour de 600 millions de collisions par seconde, et après filtrage, il reste 100 collisions d’intérêt par seconde. En conséquence, il y a 25 Po de données à stocker chaque année, et 200 Po après réplication.

Quand le Sloan Digital Sky Survey (SDSS) a commencé à collecter des données astronomiques en 2000, il a amassé plus de données en quelques semaines que toutes les données collectées dans l’histoire de l’astronomie. Il continue à un rythme de 200 Go par nuit, et a aujourd’hui stocké plus de 140 téraoctets d’information. Des prévisions annoncent que le Large Synoptic Survey Telescope, dont la mise en route est prévue en 2015, amassera ce même montant tous les cinq jours.

Décoder le génome humain a originellement pris 10 ans, cela peut désormais être fait en moins d'une semaine : les séquenceurs d'ADN ont progressé d'un facteur 10 000 les dix dernières années, soit 100 fois la loi de Moore (100 environ sur 10 ans). En biologie, les approches massives basées sur une logique d’exploration des données et de recherche d’induction sont légitimes et complémentaires des approches classiques basées sur l'hypothèse initiale formulées.

Le NASA Center for Climate Simulation (NCCS) stocke 32 Po de données d’observations et de simulations climatiques.

**Politique :**

L’analyse de Big Data a joué un rôle important dans la campagne de ré-élection de Barack Obama, notamment pour analyser les opinions politiques de la population.

Depuis l'année 2012, le Département de la défense américain investit annuellement sur les projets de Big Data plus de 250 millions de dollars.

Le gouvernement américain possède six des dix plus puissants supercalculateurs de la planète.

La National Security Agency est actuellement en train de construire le Utah Data Center. Une fois terminé, ce data center pourra supporter des yottaoctets d’information collectés par la NSA sur internet.

En 2013, le Big Data faisait partie des 7 ambitions stratégiques de la France déterminées par la Commission innovation 2030.

**Secteur privé :**

Walmart traite plus d’un million de transactions client par heure, celles-ci sont importées dans des bases de données dont on estime qu’elles contiennent plus de 2,5 Po d’information.

Facebook traite 50 milliards de photos.

D’une manière générale l'exploration de données de Big Data permet l’élaboration de profils clients dont on ne supposait pas l’existence.

Evolutions

L'un des principaux enjeux de productivité du Big Data dans son évolution va porter sur la logistique de l'information, c'est-à-dire sur comment garantir que l'information pertinente arrive au bon endroit au bon moment. Il s'agit d'une approche micro-économique. Son efficacité dépendra ainsi de celle de la combinaison entre les approches micro- et macro-économique d'un problème.

Selon une étude IDC, les données numériques créées dans le monde atteindraient 40 zettaoctets d'ici 2020. A titre de comparaison, Facebook générait environ 10 téraoctets de données par jour au début 2013. Le développement de l'hébergement massif de données semble avoir été accéléré par plusieurs phénomènes simultanément: la pénurie de disques durs due aux inondations en Thaïlande en 2011, l'explosion du marché des supports mobiles (smartphones et tablettes notamment), etc. Ajouté à cela, la démocratisation du cloud-computing de plus en plus proche, grâce à des outils comme Dropbox, amène le big data au centre de la logistique de l'information.

Afin de pouvoir exploiter au maximum le Big Data, de nombreuses avancées doivent être faites, et ce en suivant trois axes :

**Modélisation de données :**

Les méthodes actuelles de modélisation de données ainsi que les systèmes de gestion de base de données ont été conçus pour une utilisation à des fins commerciales de l’information. La fouille de données a des caractéristiques fondamentalement différentes et les technologies actuelles ne permettent pas de les exploiter. Dans le futur il faudra des modélisations de données et des langages de requêtes permettant :

une représentation des données en accord avec les besoins de plusieurs disciplines scientifiques ;

de décrire des aspects spécifiques à une discipline (modèles de métadonnées) ;

de représenter la provenance des données ;

de représenter des informations contextuelles sur la donnée ;

de représenter et supporter l’incertitude ;

de représenter la qualité de la donnée.

De très nombreux autres thèmes de recherche sont liés à ce thème, citons notamment: la réduction de modèle pour les EDP, le compressive sensing en imagerie, l'étude de méthodes numériques d'ordre élevé... Probabilités, statistiques, analyse numérique, Equations aux Dérivées Partielles déterministes et stochastiques, approximation, calcul haute performance, algorithmique.... Une grande partie de la communauté scientifique, notamment en mathématiques appliquées et en informatique, est concernée par ce thème porteur.

**Gestion de données :**

Le besoin de gérer des données extrêmement volumineuses est flagrant et les technologies d’aujourd’hui ne permettent pas de le faire. Il faut repenser des concepts de base de la gestion de données qui ont été déterminés dans le passé. Pour la recherche scientifique, par exemple, il sera indispensable de reconsidérer le principe qui veut qu’une requête sur un SGBD fournisse une réponse complète et correcte sans tenir compte du temps ou des ressources nécessaires. En effet la dimension exploratoire de la fouille de données fait que les scientifiques ne savent pas nécessairement ce qu’ils cherchent. Il serait judicieux que le SGBD puisse donner des réponses rapides et peu coûteuses qui ne seraient qu’une approximation, mais qui permettraient de guider le scientifique dans sa recherche.

Dans le domaine des données clients, il existe également de réels besoins d'exploitation de ces données, en raison notamment de la forte augmentation de leur volume des dernières années. Le big data et les technologies associées permettent de répondre à différents enjeux tels que l'accélération des temps d’analyse des données clients, la capacité à analyser l’ensemble des données clients et non seulement un échantillon de celles-ci ou la récupération et la centralisation de nouvelles sources de données clients à analyser afin d’identifier des sources de valeur pour l’entreprise.

**Outils de gestion des données :**

Les outils utilisés à l’heure actuelle ne sont pas en adéquation avec les volumes de données engendrés dans l’exploration de Big Data. Il est nécessaire de concevoir des instruments permettant de mieux visualiser, analyser, et cataloguer les ensembles de données afin de permettre une optique de recherche guidée par la donnée. La recherche en Big Data ne fait que commencer. La quantité de data évolue beaucoup plus rapidement que nos connaissances sur ce domaine. Le site the Gov Lab prévoit qu'il n y aura pas suffisamment de scientifiques du Data. En 2018, les États-Unis auraient besoin de 140 000 à 190 000 scientifiques spécialisés en Big Data.

Big Data et sécurité

## Dans un tel environnement, les activités sécurité à prendre en compte sont celles déroulées pour la sécurisation d’un système d’information : la gouvernance, la protection et enfin la supervision.

## La Gouvernance:

## L’introduction d’une solution d’analyse de données impose à la fois une révision des politiques de sécurité existantes pour intégrer les nouveaux usages des données métier, ainsi qu’une extension de celles-ci pour intégrer les problématiques propres aux nouvelles données obtenues de différentes sources telles que les réseaux sociaux et l’analyse elle-même produisant de nouvelles données métier à très forte valeur ajoutée.

## La protection:

## Selon la sensibilité des données analysées, la confidentialité est respectée en :

## Assurant un contrôle fin de l’accès aux données via une gestion des accès et des identités.

## Implémentant, si cela est nécessaire d’un point de vue règlementaire, une solution de chiffrement des données les plus sensibles.

## La valorisation des traces d’activité, essentielles pour la sécurité, est rendue possible par deux facteurs technologiques :

## La disponibilité de volumes de stockage considérables permettant la production massive des traces de l’activité des systèmes et utilisateurs.

## Le développement d’outils d’analyse conférant une nouvelle valeur métier à ces traces tout en les rendant exploitables grâce à une supervision de sécurité plus intelligente.

## La supervision:

## Avec l’émergence de menaces avancées persistantes (ou Advanced Persistent Threats - APT), dont le but premier est d’exfiltrer des données valorisables par l’attaquant, l’intégration de mesures de sécurité périmétriques est un impératif. Cependant, du fait des nouvelles menaces qui les contournent, aujourd’hui, leur seule supervision relève uniquement de l’hygiène.

## Ainsi, pour faire face à ces menaces de manière efficace et plus proactive, il est dorénavant nécessaire d’intégrer un Security Operations Center de deuxième génération (SOC V2) permettant de superviser l’ensemble des traces produites par le système, de les corréler et de les analyser dans l’optique d’identifier les signaux faibles témoignant d’une attaque évoluée active.

## Dans ce domaine, les technologies d’analyse apportées par le big data fournissent des moyens de compléter l’arsenal sécuritaire en rendant la supervision plintelligente et de ce fait plus efficace contre des menaces concrètes.