



Welcome!!



UE: BD de nouvelle génération

Niveau 5 Génie Logiciel

TP: Présentation de MapReduce

Membres:

MASSO YETNA FERDINAND HERVE (18A0516P)

Tuteur:

Guidedi KALADZAVI, PhD

plan

01

Généralités sur les bigs data

02

Généralités sur Hadoop

03

Representation de Mapreduce

04

Conclusion



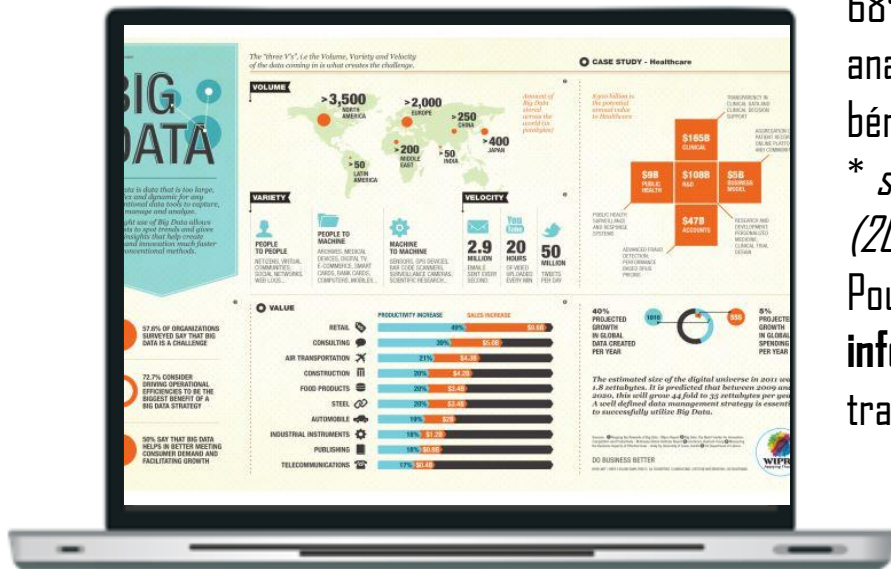
US GOV'T DATA EXPLOSION

- 7.5 PETABYTES** EUSIS ARCHIVE
- 15 PETABYTES / YEAR** HIGH ENERGY PHYSICS
- 300 TERABYTES / DAY** LIGHT SOURCE'S
- 9.6 BILLION DEVICES** GLOBAL RESEARCH PERSONAL DEVICES
- 409 MILLION WEARABLES**
- 7.1 BILLION DEVICES** PETASCALE COMPUTING
- 58 MILLION WEARABLES**
- 43 TERABYTES / DAY / DRONE** 30,000+ DRONES IN USE
- 800 TERABYTES / DAY**
- 100 EXABYTES STORED**
- 28.1 BILLION INSTALLED**
- 13.7 BILLION INSTALLED**
- 140 BILLION DNA BASES SEQUENCED**
- TERABYTES / HOUR**
- 10 PETABYTES STORED** CANCER GENOMICS ATLAS
- 2.5 PETABYTES STORED** CHINESE DIPLOMACY ATLAS
- 84 EXABYTES / MONTH**
- 225 EXABYTES / MONTH** GLOBAL IP TRAFFIC
- 2050 INDEX, M.E. & C.**

- ❑ L'extraction de connaissances à partir de grands volumes de données
- ❑ l'apprentissage statistique
- ❑ l'agrégation de données hétérogènes, la visualisation et la navigation dans de grands espaces de données et de connaissances

permettent d'observer des phénomènes, de valider des hypothèses, d'élaborer de nouveaux modèles ou de prendre des décisions en situation critique

Généralités sur les bigs data

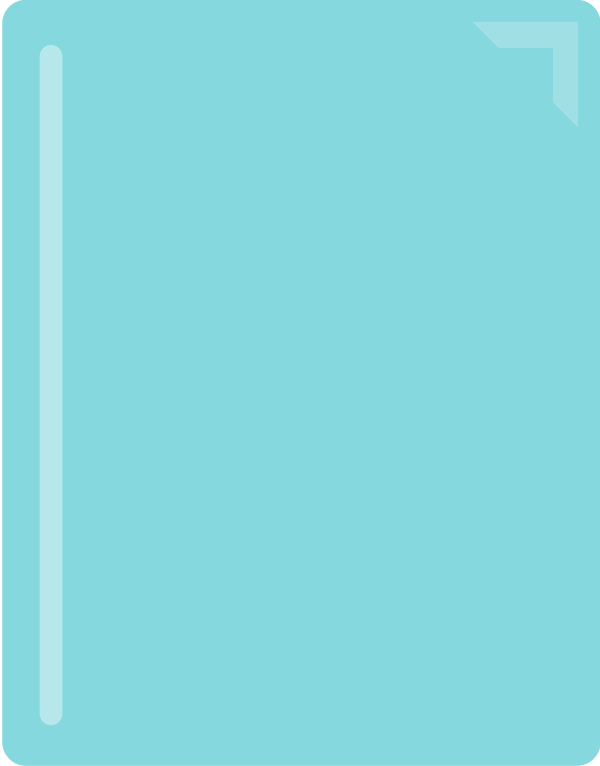


68% des entreprises qui ont systématiquement recours à une analyse de données dans leurs prises de décision voient leurs bénéfices augmenter

* *selon une étude menée par the Economist Intelligence Unit (2014)*

Pour qui réussit à optimiser son usage, la donnée devient **information**, puis, bien partagée au sein de l'entreprise, elle se transforme en **connaissance** et constitue son **savoir**

Généralités sur les bigs data



en 2012, le postulat selon lequel, de 2005 à 2020, le volume de données croîtrait d'un facteur 300, de 130 exaoctets à 40 000 exaoctets, soit 40 trillions de gigaoctets, ce qui représente plus de 5 200 gigaoctets créés pour chaque homme, femme et enfant en 2020. Actuellement, le volume des données en circulation connaît une démultiplication permanente : 5 exaoctets de données sont désormais produits tous les deux jours, soit le même volume que l'ensemble des données produites de l'aube de la civilisation jusqu'à 2003. En 2014, 90 % de toutes les données jamais générées par l'homme l'ont été au cours des deux dernières années. Cisco renchérit sur ce constat lorsqu'il estime le trafic réseau global annuel à 1,3 zettaoctets en 2016

Généralités sur les bigs data

Contexte

Très grand Volume de Données

Google = 20 milliards de pages Web = 400 TeraOctets,
30-35 MB/sec (lecture sur disque) / 4 mois de lecture.

Données Distribuées

Comment effectuer des calculs sur ces grandes masses de données?

Pour répondre à ces défis, l'idée de Google est de développer une approche conceptuelle consistant, d'une part, à distribuer le stockage des données et, d'autre part, à paralléliser le traitement de ces données sur plusieurs nœuds d'une grappe de calcul (un cluster d'ordinateurs).

Généralités sur les bigs data

❑ Indexation

- Moteurs de Recherche
- Indexation d'images

❑ Reporting

- Caractériser les utilisateurs d'un réseau social
- Détecter la fraude à la CB
- Statistiques sur des logs de connexions (publicité)

❑ Analyse

- Détection de Communautés
- Analyse du Churn

HADOOP



Généralités sur HADOOP

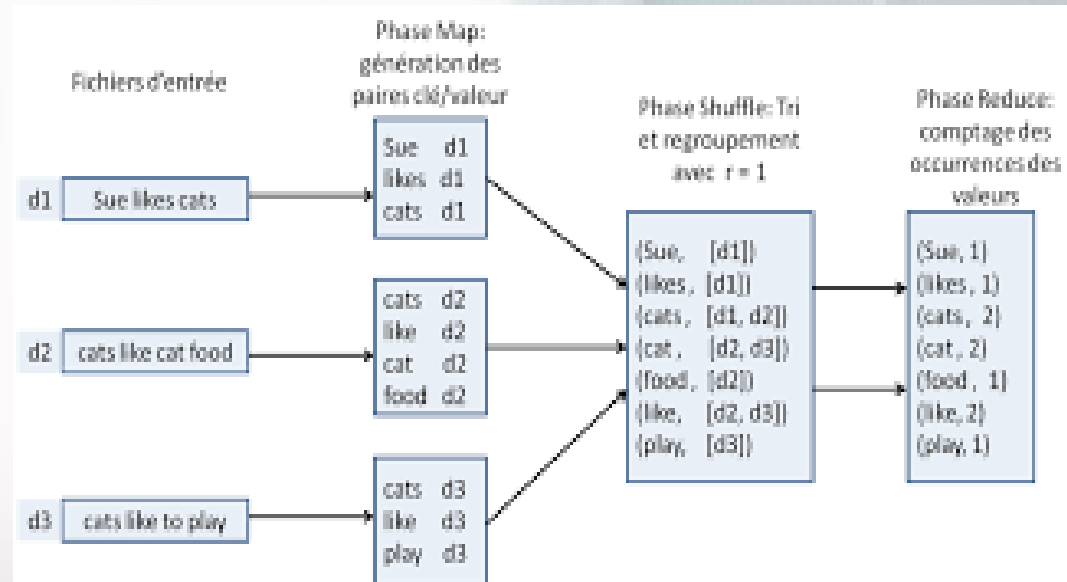
- ☐ Trop de données (GB,TB,PB)
- ☐ Améliorer des résultats existants
- ☐ Obtenir de nouveaux résultats
- ☐ Combiner des données hétérogènes
- ☐ Croissance rapide (et constante) des données
- ☐ Temps de traitement lent (minutes, heures)
- ☐ Budgets limités
- ☐ Plusieurs ordinateurs déjà disponibles



Généralités sur HADOOP

Hadoop est un framework libre et open source destiné à faciliter la création d'applications distribuées et échelonnables permettant aux applications de travailler avec des milliers de nœuds et des pétaoctets de données.

- ❑ La phase **Map** consiste à assigner à chaque nœud du cluster la tâche d'attribuer à chaque mot de la page web un indice correspondant à la page dans laquelle il est.
- ❑ La phase **Shuffle** consiste, pour chaque nœud, à trier par ordre alphabétique les mots auxquels il a affecté un index.
- ❑ La phase **Reduce** consiste, pour chaque mot contenu dans l'ensemble des nœuds du cluster, à regrouper tous ces indices



Utilisation de l'approche MapReduce pour créer un index inversé

Généralités sur HADOOP

Terminologie d'Hadoop

Avant de parler de l'exécution des *jobs* MapReduce dans Hadoop, nous allons en présenter la terminologie.

Un ***job MapReduce*** est une unité de travail que le client veut exécuter.

Il consiste en trois choses :

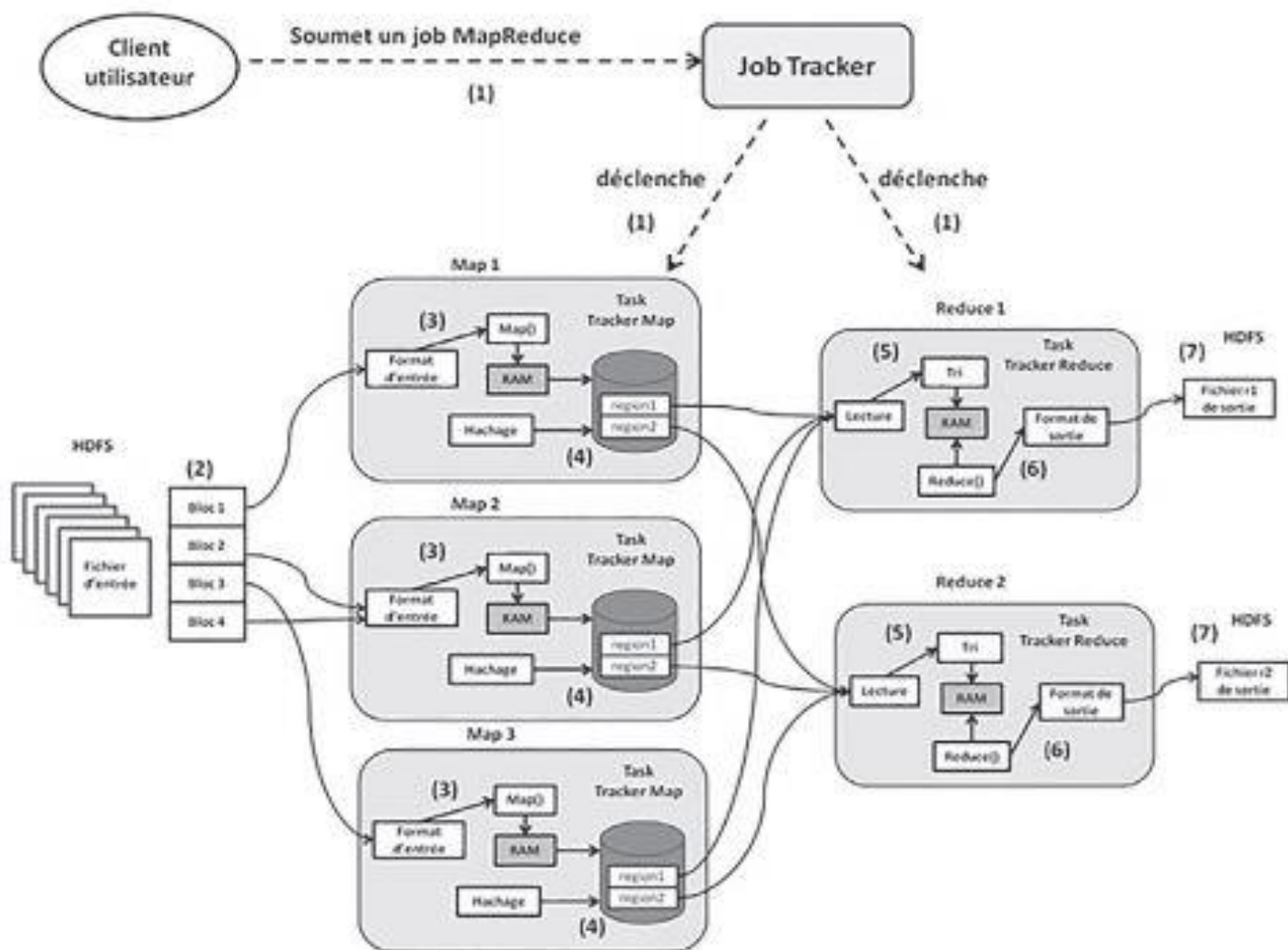
- le fichier des données à traiter (*input file*) ;
- le programme MapReduce ;
- les informations de configuration (métadonnées).

Le cluster exécute le job en divisant le programme MapReduce en deux : les tâches Map d'un côté et les tâches Reduce de l'autre. Dans le cluster Hadoop, il y a deux types de processus qui contrôlent l'exécution du job : le ***JobTracker*** et un ensemble de ***TaskTrackers***.



Détails d'exécution d'un modèle de calcul dans Hadoop

- ❑ Au départ, l'utilisateur configure le job : il écrit la fonction Map, la fonction Reduce, spécifie le nombre de tâches Reduce, le format de lecture des fichiers d'entrée, le format des fichiers de sortie, éventuellement la taille des blocs du fichier d'entrée et le Facteur de réplication.
- ❑ Le HDFS découpe le fichier d'entrée en M blocs de taille fixe, généralement 64 Mo.
- ❑ Par défaut, le JobTracker déclenche M TaskTrackers sur les M nœuds de données dans lesquels ont été répartis les M blocs du fichier d'entrée, pour exécuter les tâches Map, soit un TaskTracker Map pour chaque bloc de fichiers.
- ❑ Périodiquement, dans chaque nœud, les paires clé/valeur sont sérialisées dans un fichier sur le disque dur local du nœud.
- ❑ Lorsque les r TaskTrackers Reduce reçoivent les informations de localisation, ils utilisent des appels de procédures distantes (protocole RPC) pour lire – depuis le disque dur des nœuds sur lesquels les tâches Map se sont exécutées – les régions des fichiers Map leur correspondant.
- ❑ Les TaskTrackers Reduce itèrent à travers toutes les données triées, puis ils passent chaque clé unique rencontrée, avec sa valeur, à la fonction Reduce écrite par l'utilisateur.
- ❑ Le job s'achève là ; à ce stade, les r fichiers Reduce sont disponibles et Hadoop applique, selon la demande de l'utilisateur, soit un Print Ecran, soit leur chargement dans un SGBD, soit encore leur passage comme fichiers d'entrée à un autre job MapReduce.



Détails d'exécution d'un modèle de calcul dans Hadoop

❑ Au départ, l'utilisateur configure le job : il écrit la fonction Map, la fonction Reduce, spécifie le nombre de tâches Reduce, le format de lecture du fichiers d'entrée, le format des fichiers de sortie, éventuellement la taille des blocs du fichier d'entrée et le facteur de réplication.





Thank you