REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOHAMED BOUDIAF - M'SILA

FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE

DEPARTEMENT D'INFORMATIQUE
MEMOIRE de fin d'étude

Présenté pour l'obtention du diplôme de MASTER
Domaine : Mathématiques et Informatique
Filière : Informatique
Spécialité : informatique décisionnelle et d'optomisation

Par**: messeguem abdel djalil**

SUJET

Analyse des données avec apache spark

Soutenu publiquement le : / /2019 devant le jury composé de :

Nom et prénom Enseignant		
……	**Université de M'sila**	**président**
Saoudi laila	**Université de M'sila**	**Rapporteur**
Bentrissa rahima……	**Université de M'sila**	**Examinateur**
……	**Université de M'sila**	**Examinateur**

Promotion : 2018 /2019

Remerciments :

Je remercie allah de m'avoir donner le courage et la volonté ainsi que la conscience et la patience d'avoir pu terminer la mémoire de master

Je tiens à éxprimer mes vifs remerciments à mon encadreur Mme. **chalabi baya** pour m'avoir donner l'opportunité de réaliser se sujet sous sa direction ,la confiance faite ainsi que ses conseils et ses motivations,et son temps consacré tout au long du travail.

Je tiens à remercie **saoudi laila** d'avoir accepté d'etre notre président de jury, ainsi que l'ensemble des examinateurs ma damme **bentrissa rahima** pour leurs remarques, leurs perspectives et leurs suggestions.

Je remercie mon amie **lounnas abderrahim** qui me donne des conseil et m’a aidé de comprendre le projet

Dédicaces

Je dédie ce travail :
À mes parents
À tout ma famille
À mes très chères amis
À tous mes collègues
À tous ceux qui m'ont encouragé et aidé

Table des matières

Chapitre I : Généralités sur le cloud computing

CHAPITRE 2 : LES BIG DATA

CHAPITRE III : Mise en œuvre, Test et Evaluation

Liste des figures :

Introduction générale

Nous sommes confrontés actuellement à une explosion de données structurées ou non structurées produites massivement par les différentes sources de données numériques.

D'une part les applications qui génèrent des données issues des logs, des réseaux de capteurs, des traces de GPS, etc., et d'autre part, les utilisateurs produisent beaucoup de données telles que des photographies, des vidéos et des musiques. Selon IBM, chaque heure 2.5 trillions d'octets de données sont générées. Selon les prévisions faites, d'ici 2020 cette croissance sera supérieure à 40 Zettaoctets, alors qu'un Zettaoctet de données numériques, seulement, ont été générées de 1940(début de l'informatique) à 2010.

Beaucoup de concepts «inséparables» dominent actuellement le marché de l'IT : «Cloud Computing», «Big Data», «NoSQL» ou «MapReduce».

La problématique :

De rudes contraintes opposent les différents chercheurs dans le domaine, quant au stockage et à l'analyse de ces masses de données. Les prévisions de taux de croissance des volumes de données traitées dépassent les limites des technologies traditionnelles à savoir les bases de données relationnelles ou les Datawarehouses. On parle de pétaoctet (billiard d'octets 10^{15}), voir d'exaoctet (trilliard d'octets 10^{18}) et encore le Zettaoctet(10^{21}) ou le Yottaoctet(10^{24}).

Notre problématique est comment prendre en charge l'accroissement rapide des volumes de données géographiquement éloignées et comment gérer cette puissante montée en charge. Quel sont les technologies et les modèles de programmation proposées pour pallier ces différents problèmes engendrés par ce déluge de données ?

La solution de la problématique :

Cette révolution scientifique qui envahisse le monde de l'information et l'Internet a imposé aux différents chercheurs depuis quelques années, de nouveaux défis et les a poussé à concevoir de nouvelles technologies pour contenir, traiter ces volumes énormes de données. Plusieurs modèles de programmation parallèle et systèmes de gestion de fichiers, principalement, l'Hadoop se place comme la solution la plus répandue dans le marché informatique. et un système d'analyse et traitement de données basé le modèle de programmation MapReduce pour réaliser des traitements parallèles et distribués sur des gros volumes de données.

Notre projet de fin d'étude a pour but d'étudier les méthodes et les technologies du Big Data, Nous nous intéresserons particulièrement aux technologies Hadoop avec ses composantes HDFS et MapReduce.

Chapitre1
cloud computing

1) Introduction

Lorsque vous stockez vos photos en ligne plutôt que sur votre ordinateur à la maison, ou de l'utilisation webmail ou d'un site de réseau social, vous utilisez un service de "cloud computing". Si vous êtes une organisation, et que vous voulez utiliser, par exemple, un service de facturation en ligne est un service de «cloud computing ».

L'avancement rapide des technologies de l'information et de la communication a permis le développement de nouveaux paradigmes informatiques, où les techniques de traitement, de stockage, de communication, de partage et de diffusion de l'information ont radicalement changés. Les individus et les organisations sont de plus en plus recours à des serveurs externes pour le stockage et la diffusion efficace et fiable d'informations.

Cloud computing fait référence à la livraison de ressources informatiques plus l'Internet. Au lieu de garder des données sur votre propre disque dur ou la mise à jour applications pour vos besoins, vous utilisez un service sur Internet, à un autre emplacement, pour stocker vos informations ou utiliser ses applications.

2) Historique :

Le terme « Cloud », ou bien « nuage », a été utilisé historiquement comme une métaphore de l'Internet. Cet usage a été à l'origine dérivé de sa représentation commune dans les diagrammes de réseau comme l'ébauche d'un nuage, utilisé pour représenter le transport des données entre les réseaux fédérateurs porteurs (qui détenait le nuage) à un emplacement de point final sur l'autre côté du nuage. Ce concept remonte dès 1961, lorsque le professeur John McCarthy a suggéré que la technologie d'ordinateur en temps partagé « time-sharing » pourrait conduire à un avenir où la puissance de calcul et même les applications spécifiques pourraient être vendus comme un service public [1,2].Cette idée est devenue très populaire dans les années 1960, mais au milieu des années 1970 l'idée s'évanouit quand il est devenu clair que les technologies liées aux IT de la journée étaient incapables de soutenir un tel modèle informatique futuriste. Cependant, depuis le tournant du millénaire, le concept a été revitalisé. C'est au cours de cette période de relance que le terme Cloud Computing a commencé à émerger dans les milieux technologiques [1]. Le Cloud a entraîné un changement perturbateur dans la technologie. Ce changement, a été et, va continuer à révolutionner la façon dont les entreprises acquit et fournit des services IT.

Le Cloud Computing en version exploitable, est le fruit des investigations effectuées par Amazon Web Services (IaaS) en 2002. Cette société leader du e-business, satisfaisait

régulièrement les grosses commandes ponctuelles sur son site, lors des fêtes de Noël. Elle a investi dans un parc gigantesque de machines. Et ces dernières ne sont pas exploitées

correctement le reste de l'année. La diminution de la puissance du parc, ne pouvait pas résoudre le problème. En effet, il subsistait toujours des pointes d'appels, lors des fêtes. Et l'indisponibilité de leur site serait cruciale pour leurs affaires, car elle représentait la majorité de son chiffre d'affaire. Ce sera un impact négatif difficile à rattraper.L'idée est alors venue chez Amazon, de louer ces ressources à des entreprises, durant les périodes hors fêtes, et à la demande. Le résultat ne s'est pas fait attendre, puisque les avantages de ce concept sont nombreux pour les entreprises. Elles n'ont pas à se soucier de l'investissement en grosses

machines, ou de la gestion de machines et d'hommes, alors que ses services sont effectués dans les normes et au moindre coût.

Ses clients augmentent continuellement, et Amazon effectue des extensions de ses parcs et de ses prestations pour satisfaire les demandes. D'autres sociétés de service IT comme Google et Microsoft, ont suivi le courant. Elles se sont mises dernièrement à fournir des services identiques. Il y a également FlexiScale, RackSpace et GoGrid. On les classe comme des fournisseurs d'environnement Cloud. Selon l'Institut de consulting Gartner, une forte référence pour le domaine, le Cloud Computing arrivera bientôt au même niveau d'affaire que celui du E-business en son temps. 2013 était l'année de son adoption massive par les entreprises [3].

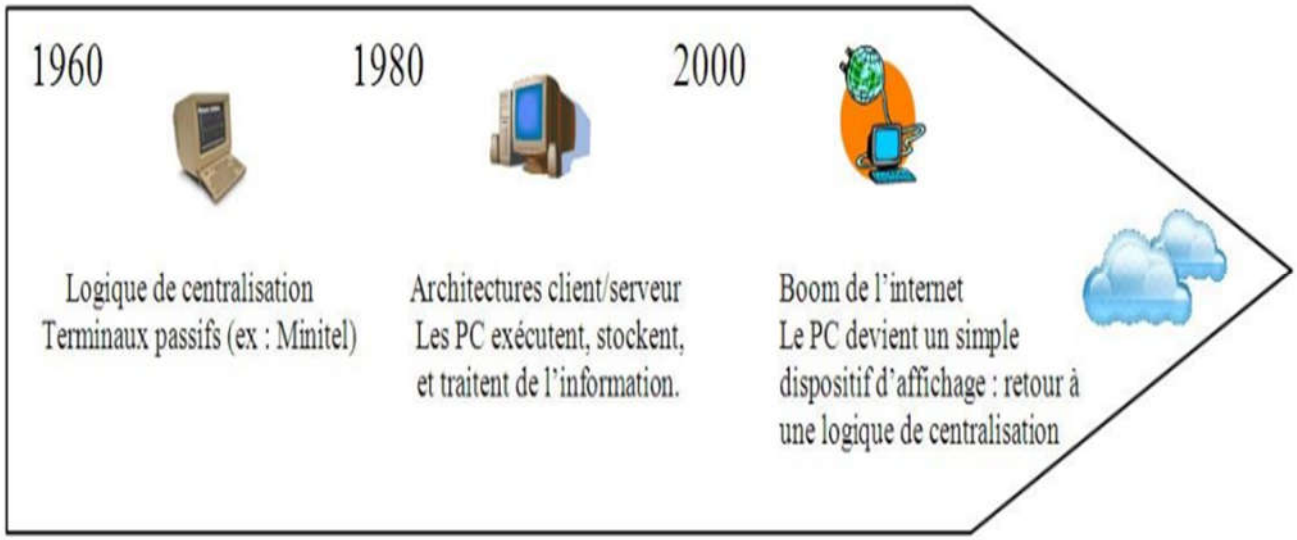


Figure I.1 : Evolution de l'informatique depuis le Minitel jusqu'au Cloud Computing [4]

3) Définitions de cloud computing :

Le Cloud signifie « nuage » et Computing « informatique », le Cloud Computing est donc l'informatique en nuage pour une traduction littérale anglais français

* la définition du National Institute of Standards and Technology (NIST), le cloud computing est l'accès via un réseau de télécommunications, à la demande et en libre-service, à des ressources informatiques partagées configurables (réseaux, serveurs, stockage, applications et services), qui peuvent être provisionnées rapidement et libérées avec un effort de gestion minimale ou prestataire de services interaction. Ce modèle de nuage est composé de cinq caractéristiques essentielles, trois modèles de services et quatre modèles de déploiement [5].

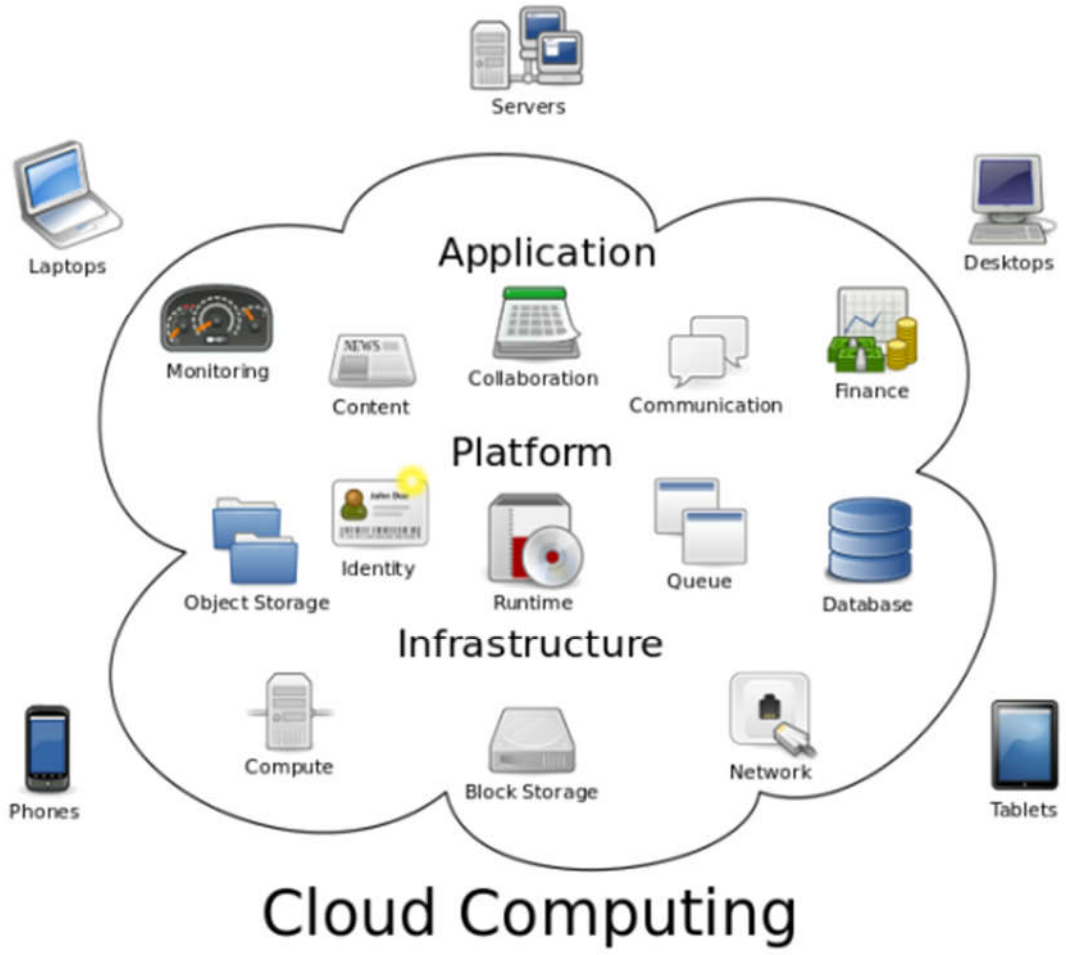


Figure I.2 : cloud computing

4) Les caractéristiques du cloud computing:

Les cinq caractéristiques essentielles de l'informatique en nuage définies par le NIST [5]sont :

– **Service à la demande**: un client peut unilatéralement allouer des ressources,telles que du temps de calcul ou de l'espace de stockage. Cette opération est faite automatiquement, sans nécessiter d'interactions humaines avec les fournisseurs.

– **Accès réseau**: les services sont disponibles sur le réseau et accessibles par des mécanismes standard permettant l'accès par des clients lourds ou légers et hétérogènes.

– **Partage des ressources**: les ressources informatiques du fournisseur sont mises en commun afin de répondre aux besoins de plusieurs clients en suivant un modèle multi-utilisateurs. Les ressources physiques et virtuelles sont allouées et ré-allouées

dynamiquement en fonction des demandes des clients. Le client ne peut générale-ment ni contrôler, ni connaître la localisation exacte des ressources fournies, mais a parfois la possibilité de la spécifier à un niveau d'abstraction différent (pays, centre de données, etc.). Les ressources fournies sont par exemple de l'espace de stockage,de la puissance de calcul, de la mémoire ou de la bande passante.

– **Élasticité rapide**: les ressources fournies au client peuvent être ajustées automatiquement (en allouant ou libérant des ressources) afin de s'adapter rapidement à la demande. Du point de vue de l'utilisateur, les ressources disponibles apparaissent le

plus souvent comme illimitées et pouvant être allouées à tout moment et en toute quantité.

– **Service mesuré**: l'utilisation des ressources est automatiquement contrôlée et op-timisée en utilisant une métrique adaptée au type de service. Cette utilisation est

communiquée au client et au fournisseur de manière transparente. Ces cinq caractéristiques permettent de définir si un service fourni est, ou non, un service d'informatique en nuage. Les différents types d'informatique en nuage sont alors

classifiés selon leur modèle de service et leur modèle de déploiement.

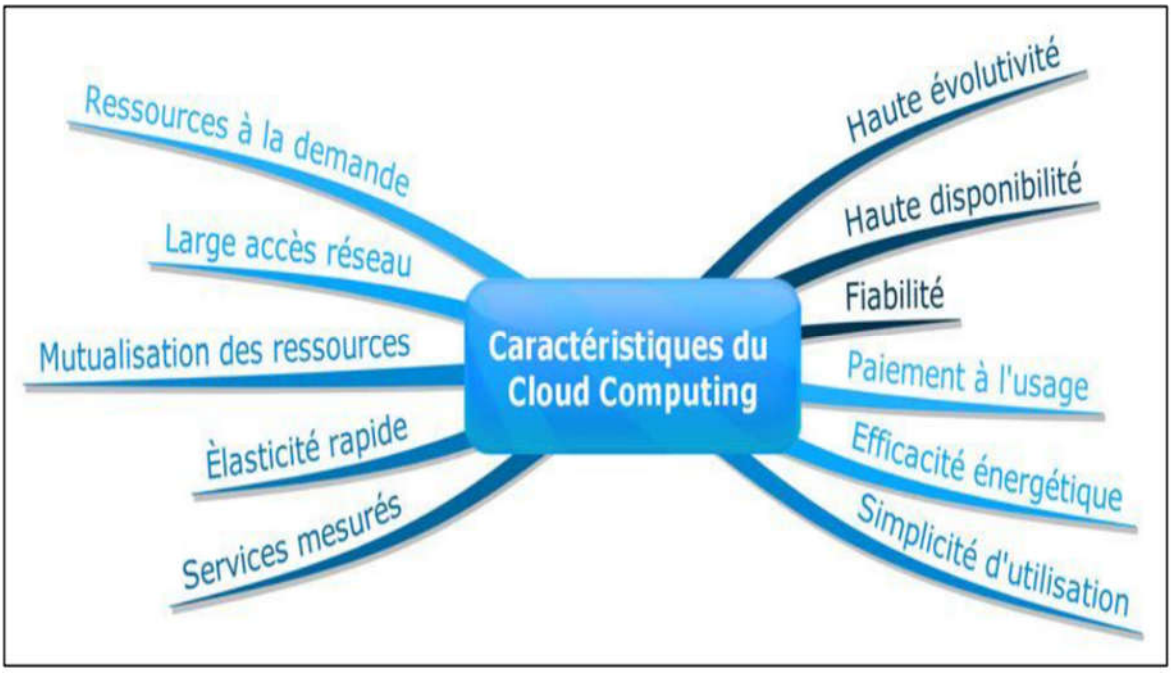


Figure I.3: Caractéristiques du Cloud Computing [5]

5) Bénéfices du cloud Computing :

Les retombées des principes du cloud sont bénéfiques à la fois pour son fournisseur, les entreprises délocalisant leurs infrastructures. Généralement, ils assurent aux deux premiers une meilleure rentabilité. De plus, ils permettent à l'entreprise de se concentrer sur les taches de production autres que la maintenance de systèmes informatiques.

5.1) Pour le fournisseur :

Les bénéfices du fournisseur sont uniquement dus au fait de la mutualisation des ressources. En effet, après son investissement dans la mise en place des infrastructures pour le cloud, il fait payer aux entreprises la marge nécessaire pour sa rentabilisation. Comme pour une entreprise disposant d'une plateforme interne, il paie pour les frais d'administration de

l'ensemble. Cette dépense peut être amortie par facturation aux entreprises. En plus de cette marge, il bénéficie des couts de réutilisation des ressources. En effet, compte tenu de la non

appartenance des ressources aux entreprises, elles (les

ressources) leurs sont facturées à chaque usage. La même ressource peut ainsi faire l'objet de plusieurs facturations.

5.2) Pour l'entreprise :

Ce qui rend le Cloud particulièrement attirant pour les entreprises particulières c'est les coûts de développement très bas pour la création et l'hébergement d'applications ou de sites web sur le Cloud, en sachant que l'utilisation des produit de ces entreprise, ne nécessitent presque aucun payement, ce qui augmente le nombre de clients.Une entreprise particulière peut bénéficier directement des services de publicités pour mettre en ligne des annonces de ces produit, ou d'utiliser des services de vente telle que ebay, ou même avoir des contrats de financement avec des banque électronique qui se trouve sur le Cloud. C'est elle la première gagnante de cette technologie. Elle réalise des bénéfices en argent [6].

6) Les trois modèles de services de cloud computing :

Trois modèles de services peuvent être offerts sur le cloud : Software as a Service (SaaS), Platform as a Service (PaaS) et Infrastructure as a Service (IaaS). Ces trois modèles de service doivent être déployés sur des infrastructures qui possèdent les cinq caractéristiques essentielles citées plus haut pour être considérées comme du cloud computing.

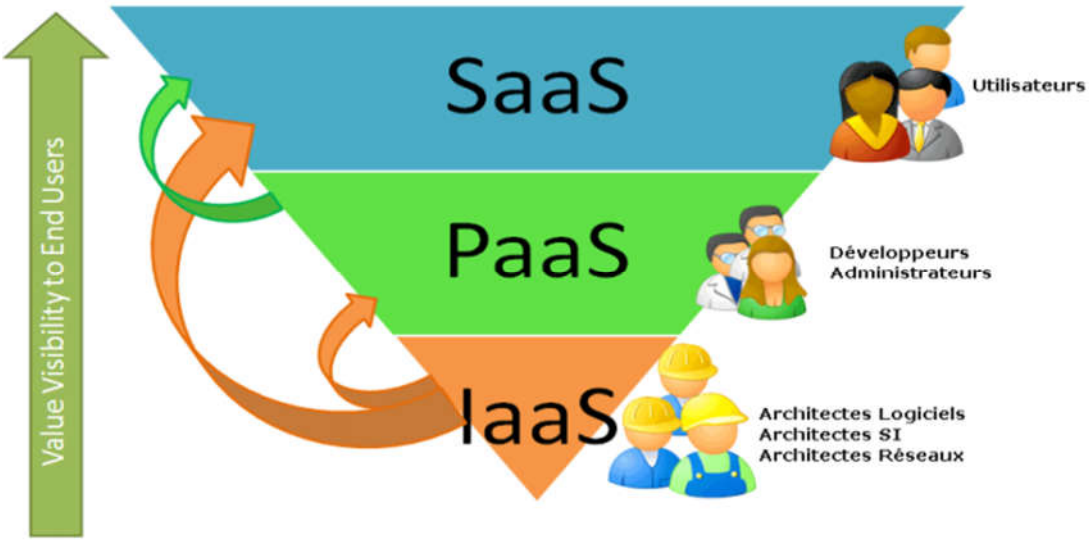


Figure I.4 : Définitions et contours du Cloud Computing [7]

6.1) Software as a Service :

L'acronyme « SAAS » est le plus connu dans le monde du Cloud Computing. Sa signification est « Software as a Service », autrement dit, application en tant que service, c'est un modèle de déploiement d'application dans lequel un fournisseur loue une application clé en main à ses clients en tant que service à la demande au lieu de leur facturer des licences.

De cette façon, l'utilisateur final n'a plus besoin d'installer tous les logiciels existants sur sa machine de travail. Cela réduit également la maintenance en supprimant le besoin de mettre à jour les applications. Ce type de modèle transforme les budgets logiciels en dépenses variables et non plus fixes et il n'est plus nécessaire d'acquérir une version du logiciel pour chaque personne au sein de l'entreprise [8].

6.2) Plateform as a Service :

Le PAAS qui signifie « Platform as a Service » est une architecture composée de tous les éléments nécessaires pour soutenir la construction, la livraison, le déploiement et le cycle de vie complet des applications et des services exclusivement disponibles à partir d'internet. Elle est également connue sous le nom de « CloudWare » [8].

Le PAAS offre des facilités à gérer le déroulement des opérations lors de la conception, du développement, du test, du déploiement et de l'hébergement d'applications web à travers des outils et des services tels que [8]:

- Le travail collaboratif (« team collaboration »).
- L'intégration des services web et bases de données.

Ces services sont fournis au travers une solution complète destinée aux développeurs et disponible immédiatement via l'internet.

6.3) Infrastructure as a Service :

L'IAAS (Infrastructure as a Service) est un modèle qui permet de fournir des infrastructures informatiques en tant que service. Ce terme était originellement connu sous le nom de (Hardware as a Service).Ces infrastructures virtuelles composent un des domaines du « As a Service » en empruntant la même philosophie de fonctionnement et de tarification que la plupart des services du Cloud Computing [8].

Plutôt que d'acheter des serveurs, des logiciels, et l'espace dans un centre de traitement de données et/ou de l'équipement réseau, les clients n'ont plus qu'à louer les ressources auprès des prestataires de service. Le service est alors typiquement tarifé en fonction de l'utilisation et de la quantité des ressources consommées.

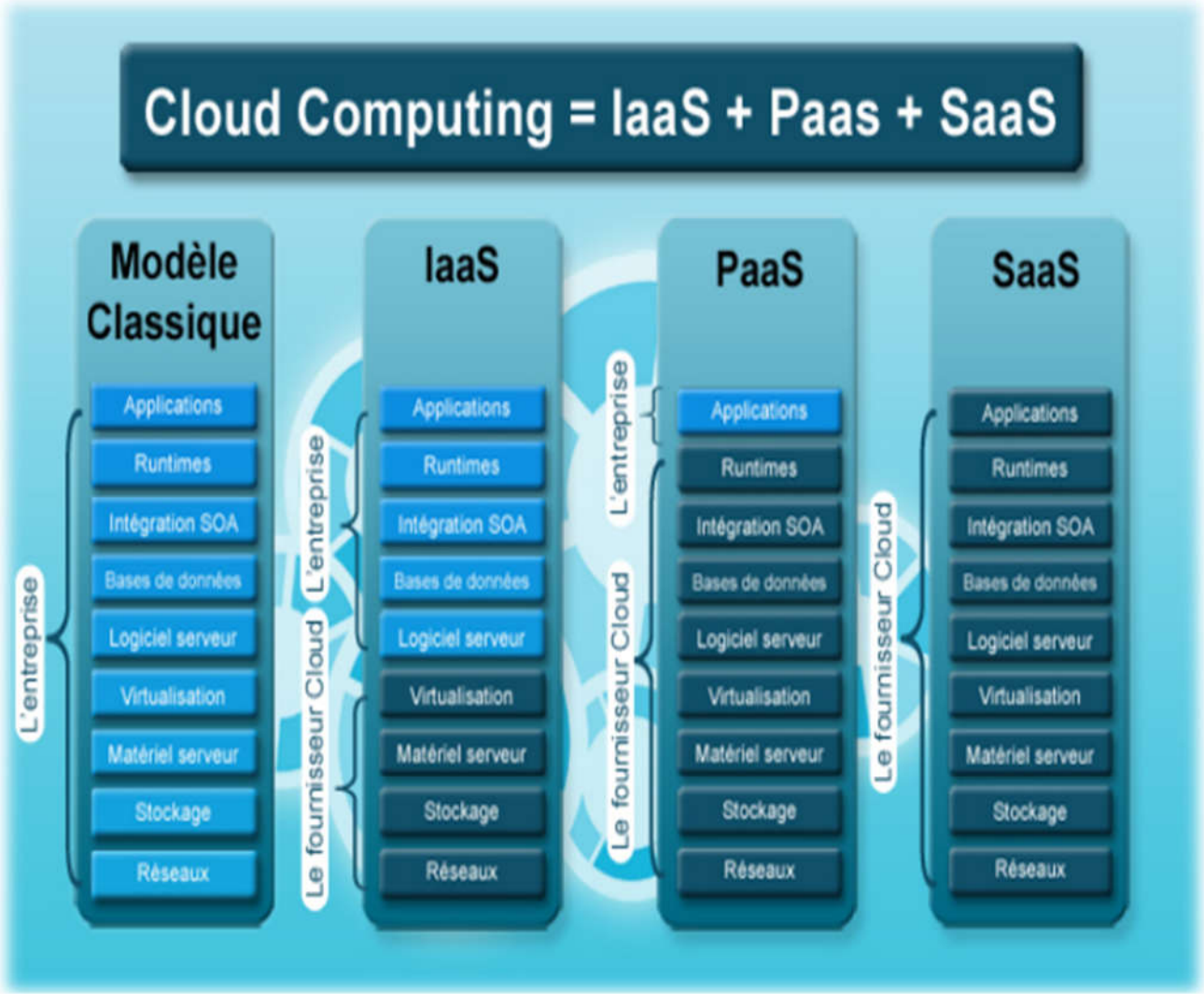


Figure I.5 :les modèles de service dans le cloud computing [9]

6.4) Application des trois modèles de service du le cloud computing :

SaaS messagerie web, applications internet, bureaux virtuels, jeux, etc... Exemples: Amazon EC2, Rackspace Cloud, Joyent

PaaS environnement d'exécution, serveurs web, bases de données, etc... Exemples: Google App Engine, Cloud Foundry, Force.com

IaaS machines virtuelles, serveurs, systèmes de stockage, réseaux, etc... Exemples: Google Docs, Microsoft Office 365, Dropbox

7) Les principaux acteurs dans le cloud computing :

les acteurs d´efinis dans l'architecture de référence de NIST. Chaque acteur est une entité (une personne ou une organisation) qui participe à une transaction ou à un processus et/ou effectue des tâches dans le cloud computing.

consommateur cloud : Une personne ou une organisation qui entretient une relation d'affaires et/ou utilise le service de fournisseur cloud.

fournisseur cloud : Une personne, une organisation ou une entit´e responsable de services à la disposition des parties
intéressées.

vérificateur cloud :Une partie qui peut procéder à une évaluation indépendante des services de cloud computing, au fonctionnement du système d'information, à la performance et à la sécurité de la mise en œuvre du cloud.

courtier cloud : Une entit´e qui g`ere l'utilisation, la performance et la presta tion des services de cloud computing et n´egocie les relations entre les fournisseurs clouds et les consommateurs clouds.

transporteur cloud : Un interm´ediaire qui fournit la connectivit´e et le transport des services de cloud computing depuis le fournisseur cloud jusqu'au consommateur cloud.

8) Modèles de déploiement dans le cloud computing :

Le NIST a défini quatre modèles de déploiement du cloud computing : le cloud privé, le cloud communautaire, le cloud public et le cloud hybride[5].

8.1) Le cloud privé :
l'infrastructure en nuage est utilisable par une seule organisation comprenant plusieurs utilisateurs. Elle peut appartenir ou être gérée par l'organisation, par une tierce partie ou par une combinaison de ces entités. Elle peut être localisée sur ou hors-site.

8.2) Le cloud communautaire :
l'infrastructure est utilisable par une communauté d'utilisateurs, appartenant à des organisations partageant des intérêts communs (objectifs, besoins de sécurité, etc.). Elle peut appartenir ou être gérée par une ou plusieurs organisations de la communauté, par une tierce partie ou par une combinaison de ces entités, et peut être localisée sur ou hors-site.

8.3) Le cloud public :
l'infrastructure est ouverte au public. Elle peut appartenir ou être gérée par une entreprise, une université, une organisation gouvernementale ou une combinaison de ces entités. Elle est localisée sur le site du fournisseur.

8.4) Le cloud hybride :

l'infrastructure est une combinaison d'au moins deux infrastructures en nuage distinctes (privée, communautaire, public). Elles demeurent des entités uniques mais interagissent par des technologies standard ou propriétaires, ce qui permet la portabilité des applications et des données.

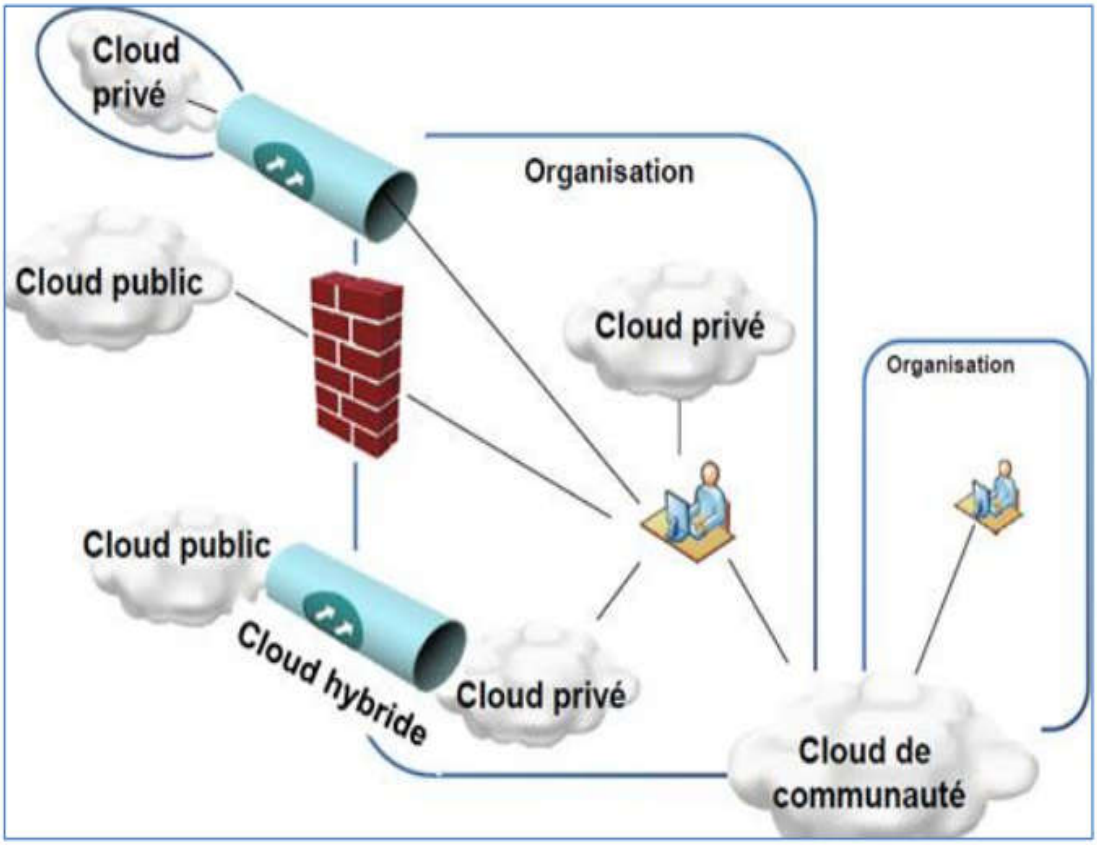


Figure I.6: Modèles de déploiement Cloud Computing

9) Avantages et inconvénients du cloud computing :

9.1) Avantages :

Nous citons quelques avantages du cloud computing:

- Réduction des couts de gestion: avec le Cloud les entreprises non plus à se soucier de la gestion des ressources ou du personnel nécessaire à la supervision de leurs plateformes… il n'ont qu'à former le corps principal du personnel à utiliser les
- applications Cloud visés par l'entreprise [10].

- Réduction des couts d'utilisation: le modèle économique du Cloud permet aux clients de réduire et de Controller leurs dépenses, puisque ils ne payent que ce qu'ils utilisent comme ressources Cloud [10].

- Système anti désastre: la récupération des données et des applications après un désastre (séisme par exemple) est gérée par un backend qui stocke et relance le système à nouveau pour assurer une disponibilité permanente des services Cloud [11].
- Un démarrage rapide : Le cloud computing permet de tester le business plan rapidement, à coûts réduits et avec facilité.

- Mobilité et accès facile: en stockant nos données et en déployant nos applicati ons sur le Cloud l'accès à ces derniers devient seulement une question de connexion Internet.

9.2) Inconvénients :

Le cloud computing présente également des inconvénients, parmi lesquels nous pouvons citer:

- Connexion Internet obligatoire : sans celle-ci, nous ne pouvons pas accéder aux ressources stockées dans le cloud computing
- Les performances des applications peuvent être amoindries : un cloud public n'améliorera définitivement pas les performances des applications par rapport à un cloud privé [12].
- La fiabilité du Cloud : Un grand risque lorsqu'on met une application qui donne des avantages compétitifs ou qui con
- tient des informations clients dans le Cloud [12].
- Conformité réglementaire: lors du transfert des données du client vers le fournisseur du service Cloud, le client est seul responsable de l'intégrité et de la sécurité des données [10].
- Dépendance des services: un client n'a pas la possibilité de changer le type de services à consommer chez un fournisseur donné [10].

10) Conclusion :

Dans ce chapitre nous avons abordé une description globale de la technologie de Cloud Computing telle que le Le Cloud Computing est une nouvelle technologie d'utilisation des services informatiques, nous pouvons être beaucoup plus flexibles et productif dans l'utilisation des ressources allouées dynamiquement. Le Cloud Computing va continuer à évoluer comme le fondement de l'Internet du futur, où nous seront interconnectés dans un réseau de contenus et des services.

Chapitre 2
les big data

1) Introduction :

De grandes quantités d'informations sont mises en ligne sur le web par des milliers d'entreprises, d'organisations et d'individus, la charge ainsi que le volume de données à gérer ont crû de façon exponentielle, pour cela les sociétés ont recouru au Datawarehouse ou entrepôt de données pour l'analyse et le stockage de données.

Généralement le Datawarehouse est centralisé dans un serveur connecté à une baie de stockage, cette solution est difficilement scalable (ajout de puissance à la demande) en plus du fait qu'elle ne gère que les données structurées dans des SGBD.

Pour faire face à l'explosion du volume des données, on parle actuellement de pétaoctet (billiard d'octets) voir de zettaoctet (trilliard d'octets) et aussi face à la grande variété des données (image, texte, web, etc.) un nouveau domaine technologique a vu le jour : le Big Data inventé par les géants du web, au premier rang comme Yahoo, Google et Facebook, qui ont été les tous premiers à déployer ce type de technologie.

Ce concept apporte une architecture distribuée et scalable pour le traitement et le stockage de données. Ce nouveau paradigme a pour principal objectif l'amélioration des performances et l'augmentation de la vitesse d'exécution des requêtes et des traitements.

2) Définition de Big Data :

Le terme de Big Data a été évoquée la première fois par le cabinet d'études Gartner en 2008 mais la naissance de ce terme effective remonte à 2001 et a été évoquée par le cabinet Meta Group.

Il fait référence à l'explosion du volume des données (de par leur nombre, la vitesse à laquelle elles sont produites et leur variété) et aux nouvelles solutions proposées pour gérer cette volumétrie tant par la capacité à stocker et explorer, et récemment par la capacité à analyser et exploiter ces données dans une approche temps réel [13].

Big data, littérairement les grosses données, est une expression anglophone utilisée pour désigner des ensembles de données qui deviennent tellement volumineux qu'ils en deviennent difficiles à travailler avec des outils classiques de gestion de base de données. Il s'agit donc d'un ensemble de technologies, d'architecture, d'outils et de procédures permettant à une organisation très rapidement de capter, traiter et analyser de larges quantités et contenus hétérogènes et changeants, et d'en extraire les informations pertinentes à un coût accessible[14].

3) Caractéristiques du Big Data :

Le Big Data (en français "Grandes données") regroupe une famille d'outils qui répondent à une triple problématiques : C'est la règle dite des 3V [15].

Figure II.1:Le Big Data, les 3 V [13]

3.1 Volume :

Le Big Data est associé à un volume de données vertigineux, se situant actuellement entre quelques dizaines de téraoctets et plusieurs péta-octets en un seul jeu de données. Les entreprises et tous les secteurs d'activités confondus, devront trouver des moyens pour gérer le volume de données en constante augmentation qui est créé quotidiennement. Les catalogues de plus de 10 millions de produits sont devenus la règle plutôt que l'exception.

Voici quelques chiffres pour illustrer ce phénomène :

- 90% des données actuelles ont été créées dans les deux dernières années seulement ;
- Twitter comme exemple, génère 7 To de données chaque jour.

3.2 Vitesse (vélocité) :

La vitesse décrit la fréquence à laquelle les données sont générées, capturées et partagées. Les entreprises doivent appréhender la vitesse non seulement en termes de création de données, mais aussi sur le plan de leur traitement, de leur analyse et de leur restitution à l'utilisateur en respectant les exigences des applications en temps réel.

3.3 Variété :

La croissance de la variété des données est la conséquence des nouvelles données multi structurelles et de l'expansion des types de données provenant de différentes sources hétérogènes. Aujourd'hui, on trouve des capteurs d'informations aussi bien dans les appareils électroménagers, les trains, les automobiles ou les avions, qui produisent des informations très variées.

Ces nouvelles données dites non-structurées sont variées :

- Des photos ;
- Des mails (avec l'analyse sémantique de leur contenu) ;
- Les données issues des réseaux sociaux (commentaires et avis des internautes sur Facebook ou Twitter par exemple) ;

Ces trois caractéristiques illustrées par les trois « V », sont les principes définissant le Big Data. Avant tout, il s'agit d'un changement d'orientation sur l'utilisation de la donnée. En somme, le point clé du Big Data est de donner un sens à ces grosses données et pour cela, il faut les analyser.

4) Processus de chargement et de collecte de données dans Big Data :

La couche responsable du chargement de données dans Big Data, devrait être capable de gérer d'énorme volume de données, avec une haute vitesse, et une grande variété de données.

Cette couche devrait avoir la capacité de valider, nettoyer, transformer, réduire (compression), et d'intégrer les données dans la grande pile de données en vue de son traitement. La Figure illustre le processus et les composants qui doivent être présent dans la couche de chargement de données [16][17].

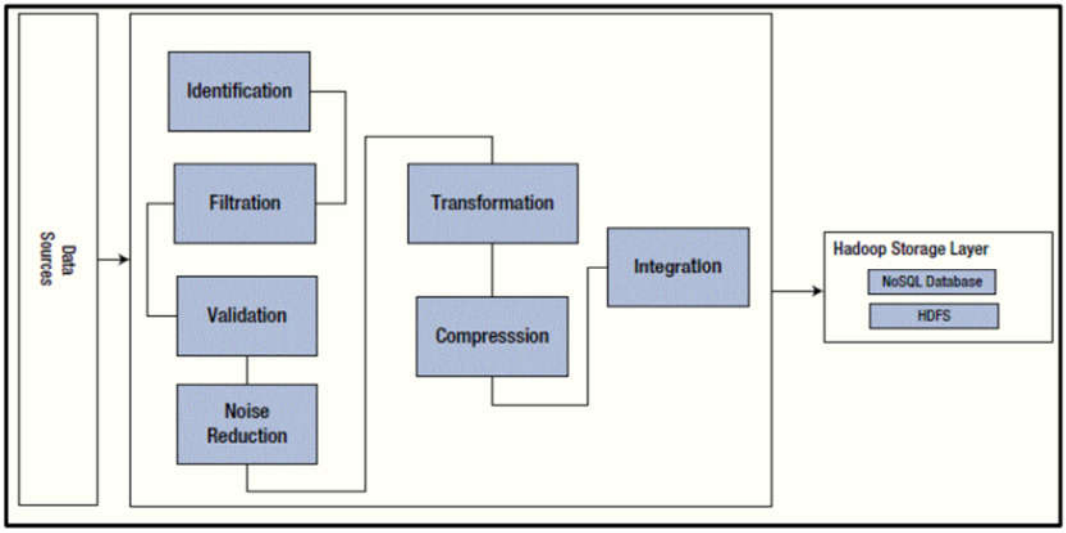


Figure II.2: Couche de chargement des données dans le Big Data[16]

La couche de chargement de données de Big Data collecte les informations pertinentes finales, sans bruit, et les charge dans la couche de stockage de Big Data (HDFS ou NoSQL base).Elle doit inclure les composants suivants :

- Identification des différents formats de données connues, par défaut Big Data cible les données non structurées ;
- Filtration et sélection de l'information entrante pertinente pour l'entreprise ;
- Validation et analyse des données en permanence ;
- Réduction de bruit implique le nettoyage des données en supprimant le bruit ;
- La transformation peut entrainer le découpage, la convergence, la normalisation ou la synthèse des données ;
- Compression consiste à réduire la taille des données, mais sans perdre de la pertinence des données ;
- Intégration consiste à intégrer l'ensemble des données dans le stockage de données de Big Data (HDFS ou NoSQL base).

5) Différence entre BI (Business intelligence) et Big Data :

La méthodologie BI traditionnel fonctionne sur le principe de regrouper toutes les données de l'entreprise dans un serveur central (Datawarehouse ou entrepôt de données). Les données sont généralement analyse en mode déconnecté.

Les données sont généralement structurées en SGBDR avec très peu de données non structurées [16][17].

Une solution Big Data, est différente d'une BI traditionnel dans les aspects suivants :

- Les données sont conservées dans un système de fichiers distribué et scalable plutôt que sur un serveur central ;
- Les données sont de formats différents, à la fois structurées ainsi que non structurées;
- Les données sont analysées en temps réel ;
- La technologie Big Data s'appuie sur un traitement massivement parallèle (concept MPP).

6) Architecture Big Data :

On distingue principalement les couches suivantes :

- Couche matériel (infrastructure Layer) :peut-être des serveurs virtuels VMware, ou des serveurs lame blade ;
- Couche stockage (Storage layer) : les données seront stockées soit dans une base NoSQL, ou bien directement dans le système de fichier distribué ou les Datawarehouse;
- Couche management et traitement : on trouve dans cette couche les outils de traitement et analyse des données comme MapReduce ou Pig [16] .
- Couche visualisation : pour la visualisation du résultat du traitement.

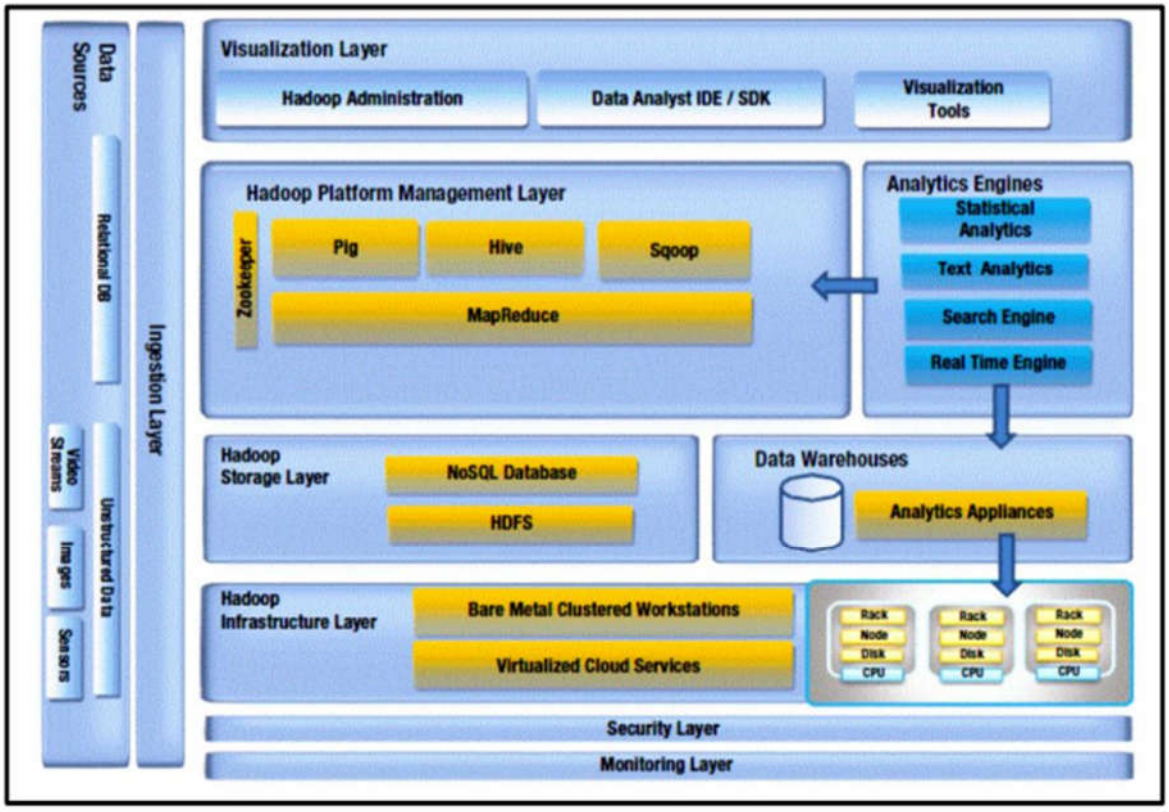


Figure II.3: Architecture de Big Data [16]

6.1) Avantages de l'architecture Big Data :

Plusieurs avantages peuvent être associés à une architecture Big Data, nous pouvons citer par exemple :

- **Evolutivité** (scalabilité) : Quelle est la taille que devra avoir votre infrastructure ? Combien d'espace disque est nécessaire aujourd'hui et à l'avenir ? le concept Big Data nous permet de s'affranchir de ces questions, car il apporte une architecture scalable.
- **Performance** : Grâce au traitement parallèle des données et à son système de fichiers distribué, le concept Big Data est hautement performant en diminuant la latence des requêtes.
- **Coût faible** : Le principal outil Big Data à savoir Hadoop est en Open Source, en plus on n'aura plus besoin de centraliser les données dans des baies de stockage souvent excessivement chère, avec le Big Data et grâce au système de fichiers distribués les disques internes des serveurs suffiront.
- **Disponibilité** : On a plus besoin des RAID disques, souvent coûteux. L'architecture Big Data apporte ses propres mécanismes de haute disponibilité.

7) Sources et types de données :

Les données collectées, stockées et traitées dans les Big Data peuvent être issues de différents domaines et crées par plusieurs sources de données hétérogènes, ce qui génère une masse de données de types différents structurés et non structurés [18].

7.1) Sources de données structurées :

Le terme « données structurées » désigne généralement des données dont la longueur et le format sont définis. Les exemples de données structurées comprennent des nombres, des dates et des chaînes (par exemple, le nom d'un employé, son poste de travail, etc.). On rappelle que la plupart des études déclarent que ce type de données représente environ de 10 à 20 pour cent des données globales. Les données structurées sont les données qu'on a l'habitude de traiter, généralement stockées dans une base de données relationnelle ayant un schéma, interrogées à l'aide du langage de requête structuré (SQL). Elles sont recueillies à partir de sources traditionnelles.

Dans le monde des Big Data, les données structurées prennent un nouveau rôle avec l'évolution de la technologie qui offre de nouvelles sources de données structurées produites souvent en temps réel et en grands volumes.

7.2) Sources de données non structurées :

Les données non structurées sont des données qui n'ont pas un format spécifié. Elles représentent 80 à 90 pour cent de l'ensemble des données disponibles. Jusqu'à quelques années auparavant, les outils disponibles n'offraient pas des traitements particuliers à ce type de données, mise à part le stockage ou l'analyse manuelle. Des données non structurées sont un peu partout : pdf, doc, email ou post dans un réseau social. En fait, un grand nombre d'individus et organisations mènent leurs activités, leurs revenus autour de ces types de données.

Tout comme pour les données structurées, les données non structurées sont générées soit par la machine ou par l'humain. Voici quelques exemples de données non structurées générées par une machine :

- Images satellites : Incluent les données météorologiques ou les données images de surveillance par satellite capturées par les services gouvernementaux. Un exemple typique de ces systèmes est Google Earth.
- Données scientifiques : Cela comprend l'imagerie sismique, les données atmosphériques,les données astronomiques, l'environnement, la génomique, la physique subatomique et la physique des hautes énergies.
- Photographies et vidéo : Concerne la sécurité, la surveillance et la vidéo de trafic.
- Voici d'autres exemples de données non structurées générées par l'homme :
- Textes et courrier interne d'une entreprise : Inclue tous les textes contenus dans les documents, les journaux, les rapports, les procés verbaux, les bilans, les résultats d'enquêtes, les sondages et les courriers. L'information d'entreprise représente en fait un grand pourcentage des données textuelles dans le monde d'aujourd'hui.
- Données des médias sociaux : Ces données sont générées à partir des plateformes des réseaux sociaux tels que YouTube, Facebook, Twitter, LinkedIn et Flickr.
- Données mobiles : Cela inclut des données telles que les messages texte et les informations de localisation.
- Contenu du site Web : Ceci provient de n'importe quel site offrant des contenus non structurés, tels que YouTube, Flickr ou Instagram [18,19].

Notons à la fin de cette section qu'il y a une autre catégorie de données qualifiée comme semi-structurée qui se place entre les deux catégories structurée et non structurée. Les données semi-structurées ne sont pas nécessairement conformes à une structure fixe prédéfinie mais peuvent être auto-descriptives et définies par des simples couples marque/valeur. Par exemple, ces couples peuvent inclure :<Famille> = Matallah, <Père> = Abdelkader, et <fils> = Oussama.

Des exemples de données semi-structurées incluent XML (Extensible Markup Language), CSV file (Comma-Separated Values), EDI (Electronic Data Interchange) et SWIFT (Langage de script implicitement parallèle).

.

8) Quelques domaines d'utilisation du Big Data :

Avant de conclure, citons rapidement quelques domaines d'utilisation du Big Data. Le Big Data trouve sa place dans de nombreux domaines :

Dans la première catégorie, on retrouve des secteurs qui manipulent quotidiennement des volumes de données très important, avec des problématiques de vitesse associées. On y trouve:

- Les Banques : la sanctuarisation de données anciennes dues à des contraintes réglementaires ;
- La Télécommunication : l'analyse de l'état du réseau en temps réel ;
- Les Médias Numériques : le ciblage publicitaire et l'analyse de sites web ;
- Les Marchés Financier : l'analyse des transactions pour la gestion des risques
- et la gestion des fraudes, ainsi que pour l'analyse des clients.

- La deuxième catégorie de secteur est plus hétérogène, les besoins, mais aussi l'utilisation qui est faite du Big Data, peuvent être très différents. On y trouve :
- Les Services Publics : l'analyse des compteurs (gaz, électricité, etc.) et la gestion des équipements ;
- Le Marketing : le ciblage publicitaire et l'analyse de tendance ;
- La Santé : l'analyse des dossiers médicaux et l'analyse génomique [15].

9) Big Data et Datawarehouse :

Les entrepôts de données sont traditionnellement des supports des données structurées et ont été étroitement lié aux systèmes opérationnels et transactionnels de l'entreprise (SGBDR). Ces systèmes qui sont soigneusement construits sont maintenant au milieu d'importants changements après l'émergence de Big Data [20][17].

Datawarehouse est une base de données (données structurées) regroupant une partie ou l'ensemble des données fonctionnelles d'une entreprise. Il entre dans le cadre de l'informatique décisionnelle ; son but est de fournir un ensemble de données servant de référence unique, utilisées pour la prise de décisions dans l'entreprise par les baies de statistiques et de rapports réalisés via des outils de reporting. D'un point de vue technique, il sert surtout à 'délester' les bases de données opérationnelles des requêtes pouvant nuire à leurs performances [21].

Les organisations continueront inévitablement à utiliser des entrepôts de données pour gérer le type de données structurées et opérationnelles qui caractérise les systèmes relationnels(SGBDR). Ces entrepôts seront toujours fournis aux analystes avec la capacité pour analyser les données clés, les tendances, et ainsi de suite.Cependant, avec l'avènement du Big Data, le défi pour les entrepôts de données est de réfléchir à une approche complémentaire avec le Big Data, on pourrait concevoir un modèle hybride. Dans ce modèle les restes de données optimisées opérationnelles très structurées seront stockées et analysées dans l'entrepôt de données, tandis que les données qui sont fortement distribuées et non structurées seront contrôlées par Big Data (Hadoop ou NoSQL) [20][17].

La tendance serait de stocker la grande masse de données non structurées dans une vaste gamme de serveurs Big Data (Hadoop/MapReduce) pour tirer profit de la scalabilité et la rapidité d'analyse de Big Data, ensuite à l'aide d'outils, ces données seront déplacées dans le modèle relationnel de sorte qu'elles peuvent être interrogées avec le langage SQL traditionnel (SGBDR et Datawarehouse).

On peut donc interfacer Big Data avec le Datawarehouse(DW), effectivement les données non structurées provenant de différentes sources peuvent être regroupées dans un HDFS avant d'être transformées et chargées à l'aide d'outils spécifiques dans le Datawarehouse et les outils traditionnels de BI [16][17].

Comme les DW traditionnels ne gèrent pas les données non structurées, Big Data peut servir comme un moyen de stockage et d'analyse des données non structurées qui seront chargées dans les DW.

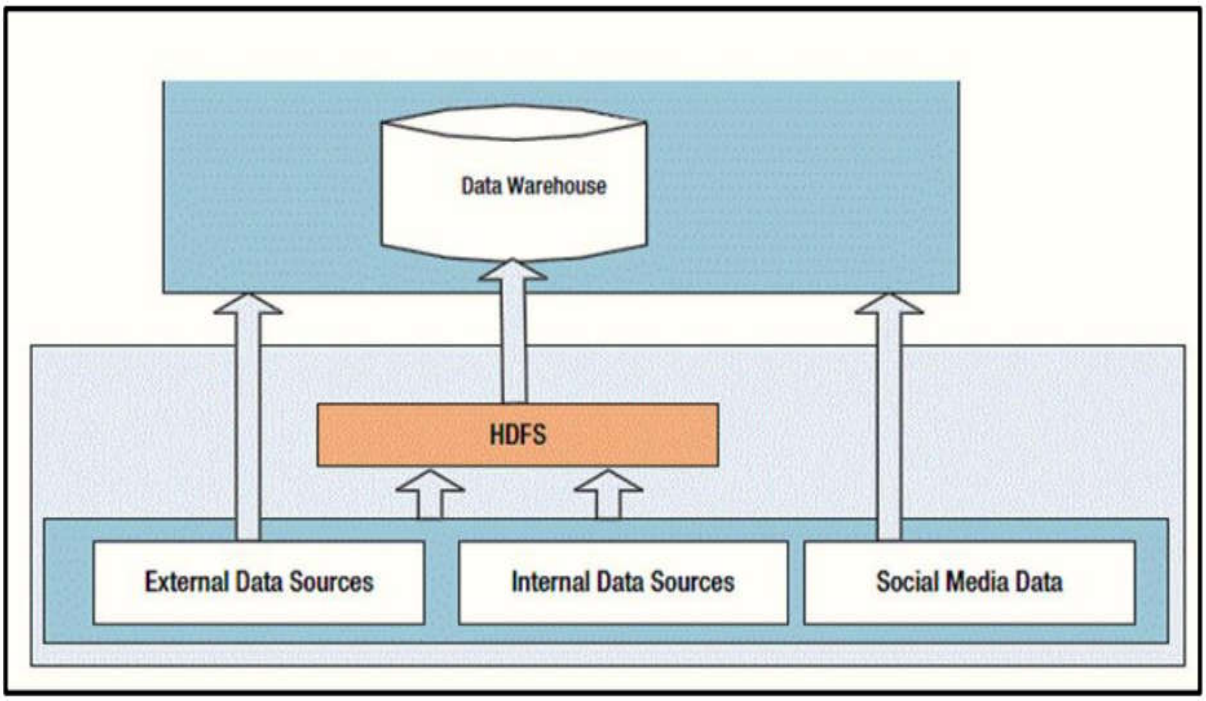


Figure II.4 : Lien entre Big Data et DW [16]

10) Big Data et les ETL (extraction, transformation et chargement) :

Certains outils ETL traditionnels comme Talent commence à s'adapter avec le monde Big Data. Les outils ETL sont utilisés pour transformer les données dans le format requis par l'entrepôt de données(Datawarehouse). La transformation est effectivement faite dans un endroit intermédiaire avant que les données ne seront chargées dans l'entrepôt de données.

Pour le Big Data des outils ETL comme Informatica ont été utilisés pour permettre une solution d'ingestion rapide et flexible des données non structurées (supérieure à 150 Go/jour)[16].

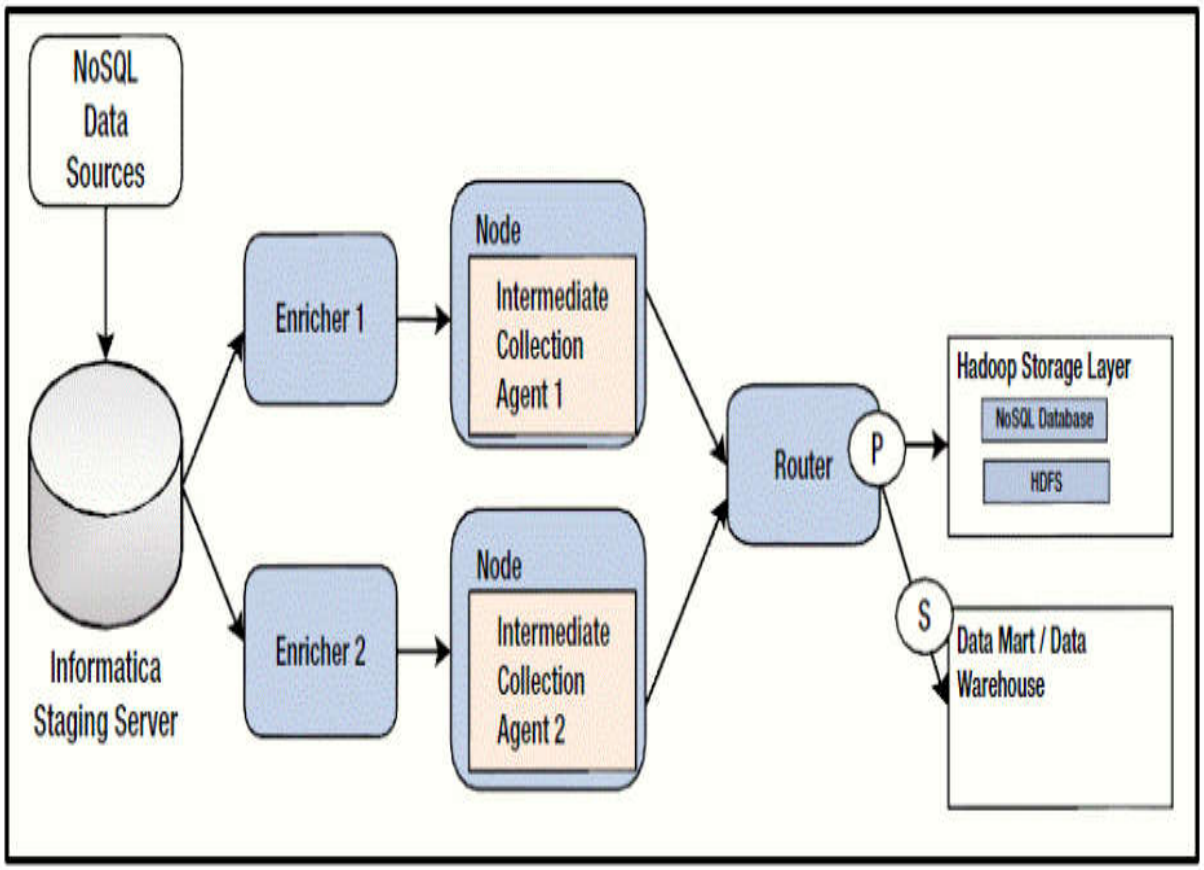


Figure II.5 : Utilisation d'ETL Informatica pour Big Data [16]

11) Les principales technologies de Big Data :

Elles sont nombreuses. Pour optimiser les temps de traitement sur des bases de données géantes, plusieurs solutions peuvent entrer en jeu :

***Des bases de données NoSQL** (comme MongoDB, Cassandra ou Redis) qui implémentent des systèmes de stockage considérés comme plus performants que le traditionnel SQL pour l'analyse de données en masse (orienté clé/valeur, document, colonne ou graphe).

* **Des infrastructures de serveurs pour distribuer les traitements** sur des dizaines, centaines, voire milliers de nœuds. C'est ce qu'on appelle le traitement massivement parallèle. Le Framework Hadoop est sans doute le plus connu d'entre eux. Il combine le système de fichiers distribué HDFS, la base NoSQL HBase et l'algorithme MapReduce.

* **Le stockage des données en mémoire** : On parle de traitement in-memory pour évoquer les traitements qui sont effectués dans la mémoire vive de l'équipement informatique, plutôt que sur des serveurs externes.

L'avantage du traitement in-memory est celui de la vitesse puisque les données sont immédiatement accessibles. En revanche, ces données ne sont pas stockées sur le long terme, ce qui peut poser des problèmes d'historisation [22].

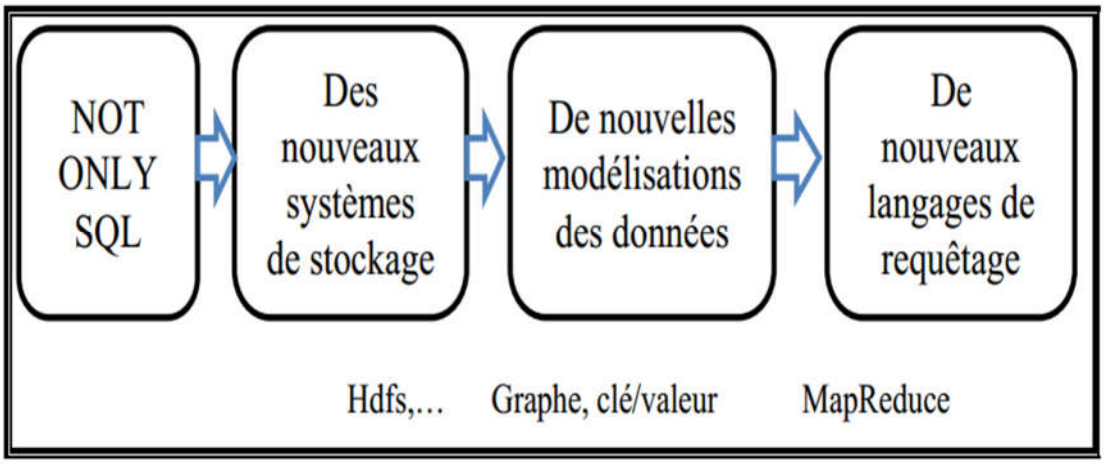


Figure II.6: Les solutions de stockage

12) Bases de données NoSQL :

Les bases de données NoSQL (No-SQL ou Not Only SQL) sont un sujet très à la mode en ce moment. Le terme NoSQL désigne une catégorie de systèmes de gestion de base de données destinés à manipuler des bases de données volumineuses pour des sites de grande audience. Les bases données NoSQL sont scalables, elles permettent de traiter les données d'une façon distribuée. Parmi les avantages du NoSQL on trouve :

* Leurs performances ne s'écroulent jamais quel que soit le volume traité. Leur temps de réponse est proportionnel au volume ;
* Elles se migrent facilement. En effet, contrairement aux SGBDR classiques, il n'est pas nécessaire de procéder à une interruption de service pour effectuer le déploiement d'une fonctionnalité impactant les modèles des données ;
* Elles sont facilement scalable. A titre d'exemple, le plus gros cluster de NoSQL fait 400 To, tandis qu'Oracle sait traiter jusqu'à une vingtaine de Téraoctet (pour des temps de réponse raisonnable).

12.1Caractéristiques NoSQL :

- Gros volume de données ;
- Réplication scalable et distribution ;
- Des centaines de machines voire des milliers;
- Distribuées partout dans le monde.
- Des requêtes qui exigent une réponse rapide;
- Asynchronicité des insertions et updates ;
- Acidité non respectée dans la plupart du temps ;
- Des performances en lectures/écritures ;
- Centaines de milliers de lectures/secondes ;
- Centaines de milliers d'écritures/secondes ;

12.2 Les types des bases NoSQL :

Il en existe 4 types distincts qui s'utilisent différemment et qui se prêtent mieux selon le type données que l'on souhaite y stocker [23] .

*** Clé-Valeur**

Les BD NoSQL fonctionnant sur le principe Clé-Valeur sont les plus basiques que l'on peut trouver.

- Elles fonctionnent comme un grand tableau associatif et retourne une valeur dont elle ne connaît pas la structure ;
- Leur modèle peut être assimilé à une table de hachage (hashmap) distribuée ;
- Les données sont simplement représentées par un couple clé/valeur ;
- La valeur peut être une simple chaîne de caractères, ou un objet sérialisé.

*** Document**

Elles sont basées sur le modèle « clé-valeur » mais la valeur est un document en format semi-structuré hiérarchique de type JSON ou XML (possible aussi de stocker n'importe quel objet, via une sérialisation). Elles stockent une collection de "documents"

*** Colonnes**

Les données sont stockées par colonne, non par ligne, on peut facilement ajouter des colonnes aux tables, par contre l'insertion d'une ligne est plus coûteuse quand les données d'une colonne se ressemblent, on peut facilement compresser la colonne.

C'est un modèle proche d'une table dans un SGBDR mais ici le nombre de colonnes:

- est **dynamique** ;
- peut **varier d'un enregistrement à un autre**, ce qui évite de retrouver des colonnes ayant des valeurs NULL.

*** Graphe**

Elles permettent la modélisation, le stockage et la manipulation de données complexes liées par des relations non-triviales ou variables

- modèle de représentation des données basé sur la théorie des graphes
- s'appuie sur les notions de nœuds, de relations et de propriétés qui leur sont rattachées.

12.3 Les principales bases de données NoSQL :

- MongoDB

La plus populaire des bases NoSQL documentaires est écrite en C et n'utilise pas de machine virtuelle JAVA.. Elle possède également une documentation de premier ordre [24].

- Cassandra

Cassandra est le projet open source qui découlent de la technologie de stockage Facebook. Cassandra est une base de données en colonnes écrite en JAVA [24].

- HBase

Hbase est inspirée des publications de Google sur BigTable. Comme BigTable, elle est une base de données oriente colonne. Basée sur une architecture maitre/esclave, les bases de données HBase sont capables de gérer d'énormes quantités d'informations (plusieurs milliards de lignes par table) [17].

Conclusion :

Dans Dans ce chapitre, nous avons présenté les principes des Big Data, ces caractéristiques, son fonctionnement ainsi que les différents domaines dans lesquels elles sont utilisées.

On a aussi recensé les différents modèles de bases de données NoSQL qui existent actuellement dans le marché en accentuant sur les solutions les plus populaires.

chapitre 03
Mise en œuvre, Test et Evaluation

1) Présentation d'Hadoop :

Hadoop est un Framework Java open source d'Apache pour réaliser des traitements sur des volumes de données massifs, de l'ordre de plusieurs pétaoctets (soit plusieurs milliers de To).

Hadoop a été conçu par Doug Cutting en 2004, également à l'origine du moteur Open Source Nutch. Doug Cutting cherchait une solution pour accroître la taille de l'index de son moteur. Il eut l'idée de créer un Framework de gestion de fichiers distribués. Yahoo! en est devenu ensuite le principal contributeur, le portail utilisait notamment l'infrastructure pour supporter son moteur de recherche historique. Comptant plus de 10 000 clusters Linux en 2008, il s'agissait d'une des premières architectures Hadoop digne de ce nom. Crée spécialement pour les gros volumes. Facebook pour l'analyse des logs, Google pour l'analyse des requêtes, etc...

Il est caractérisé par :

- **Robuste** : si un nœud de calcul tombe, ses tâches sont automatiquement réparties sur d'autres nœuds. Les blocs de données sont également répliqués;
- **Coût** : il optimise les coûts via une meilleure utilisation des ressources présentées;
- **Souple** : car il répond à la caractéristique de variété des données en étant capable de traiter différents types de données;
- **Virtualisation** : ne plus se reposer directement sur l'infrastructure physique (baie de stockage coûteuse), mais choisir la virtualisation de ses clusters Hadoop.

Trois principales distributions Hadoop sont aujourd'hui disponibles : Cloudera, Hortonworks, MapR.

Nous allons présenter deux concepts fondamentaux d'Hadoop : Sa propre version de l'algorithme MapReduce à savoir Hadoop MapReduce et son système de fichiers distribué HDFS.

2) Le système de fichier distribué d'Hadoop HDFS :

Hadoop utilise un système de fichiers virtuel qui lui est propre : le HDFS (Hadoop Distributed File System). HDFS est un système de fichier distribué, extensible et portable inspiré par le Google File System (GFS).

Il a été conçu pour stocker de très gros volumes de données sur un grand nombre de machine équipées de disques durs banalisés, il permet de l'abstraction de l'architecture physique de stockage, afin de manipuler un système de fichier distribué comme s'il s'agissait d'un disque dur unique. [25]

3) MapReduce

MapReduce est un paradigme (modèle) de programmation parallèle proposé par Google. Il est principalement utilisé pour le traitement distribué sur de gros volumes de données aux seins d'un cluster de nœuds. Il est conçu pour la scalabilité et la tolérance aux pannes.

Le modèle de programmation fournit un cadre à un développeur afin d'écrire une fonction Map et une fonction Reduce. Tout l'intérêt de ce modèle de programmation est

de simplifier la vie du développeur. Ainsi, ce développeur n'a pas à se soucier du travail de parallélisation et de distribution du travail. MapReduce permet au développeur de ne s'intéresser qu'à la partie algorithmique [26].

Un programme MapReduce peut se résumer à deux fonctions Map () et Reduce ()

* La première, **MAP**, va transformer les données d'entrée en une série de couples clef /valeur. Elle va regrouper les données en les associant à des clefs, choisies de telle sorte que les couples clef/valeur aient un sens par rapport au problème à résoudre. Par ailleurs, cette opération doit être parallélisable: on doit pouvoir découper les données d'entrée en plusieurs fragments, et faire exécuter l'opération MAP à chaque machine du cluster sur un fragment distinct. La fonction Map s'écrit de la manière suivante :

Map (clé1, valeur1) → List (clé2, valeur2).

* La seconde, **REDUCE**, va appliquer un traitement à toutes les valeurs de chacune des clefs distinctes produite par l'opération MAP. Au terme de l'opération REDUCE, on aura un résultat pour chacune des clefs distinctes. Ici, on attribuera à chacune des machines du cluster une des clefs uniques produites par MAP, en lui donnant la liste des valeurs associées à la clef. Chacune des machines effectuera alors l'opération REDUCE pour cette clef. La fonction Reduce s'écrit de la manière suivante : Reduce (clé2, List (valeur2)) → List (valeur2).

L'exemple classique est celui du **WordCount** qui permet de compter le nombre d'occurrences d'un mot dans un fichier. En entrée l'algorithme reçoit un fichier texte qui contient les mots suivants :

voiture la le elle de elle la se la maison voiture

Dans notre exemple, la clé d'entrée correspond au numéro de ligne dans le fichier et tous les mots sont comptabilisés à l'exception du mot « se ».

Le résultat de la fonction Map est donné ci-dessous.

(voiture, 1) /(la,1) /(le,1) / (elle,1) / (de,1) / (elle,1) / (la,1) / (la,1) / (maison,1) / (voiture,1)

Avant de présenter la fonction Reduce, deux opérations intermédiaires doivent être exécutées pour préparer la valeur de son paramètre d'entrée. La première opération appelée shuffle permet de grouper les valeurs dont la clé est commune. La seconde opération appelée sort permet de trier par clé. A la différence des fonctions Map et Reduce, shuffle et sort sont des fonctions fournies par le Framework Hadoop, donc, il n'a pas à les implémenter.

Ainsi, après l'exécution des fonctions shuffle et sort le résultat de l'exemple est le suivant :

(de, [1]) / (elle, [1,1]) / (la, [1, 1,1]) / (le, [1]) / (maison, [1]) / (voiture, [1,1]

Suite à l'appel de la fonction Reduce, le résultat de l'exemple est le suivant :

(de, 1) / (elle, 2) / (la, 3) / (le, 1) / (maison, 1) / (voiture, 2)

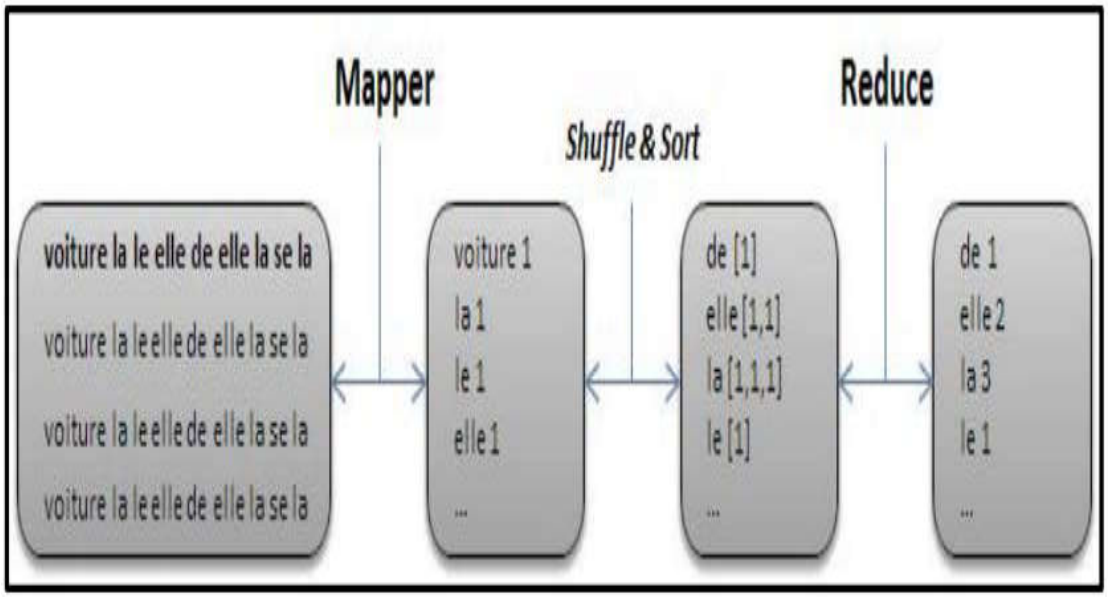


Figure III.1: Exemple d'un programme MapReduce (WordCount) [26]

4) Présentation de spark :

Spark (ou Apache Spark) est un framework open source de calcul distribué. Il s'agit d'un ensemble d'outils et de composants logiciels structurés selon une architecture définie [27]. Développé à l'université de Californie à Berkeley par AMPLab [28] , Spark est aujourd'hui un projet de la fondation Apache. Ce produit est un cadre applicatif de traitements big data pour effectuer des analyses complexes à grande échelle.

Apache Spark est un moteur de traitement de données rapide dédié au Big Data. Il permet d'effectuer un traitement de larges volumes de données de manière distribuée (cluster computing). Très en vogue depuis maintenant quelques années, ce Framework est en passe de remplacer Hadoop. Ses principaux avantages sont sa vitesse, sa simplicité d'usage, et sa polyvalence.

Son principal avantage est sa vitesse, puisqu'il permet de lancer des programmes 100 fois plus rapidement que Hadoop MapReduce in-memory, et 10 fois plus vite sur disque. Son moteur d'exécution DAG avancé supporte le flux de données acyclique et le computing in-memory. Il est également facile à utiliser, et permet de développer des applications en Java, Scala, Python et R. Son modèle de programmation est plus simple que celui d' Hadoop. Grâce à plus de 80 opérateurs de haut niveau, le logiciel permet de développer facilement des applications parallèles.

5) L'agorithme de youtube :

Analyse de données You Tube :

- Ce blog explique comment effectuer une analyse de données You Tube dans hadoop, mapreduce.

- Ces données You Tube sont disponibles publiquement et l'ensemble de données You Tube est décrit ci-dessous sous l'en-tête Description de l'ensemble de données.
- À l'aide de cet ensemble de données, nous effectuerons des analyses et dégagerons des idées telles que les 10 vidéos les mieux notées sur You Tube, qui ont téléchargé le plus grand nombre de vidéos.
- En lisant ce blog, vous comprendrez comment gérer des ensembles de données qui ne sont pas correctement structurés et comment trier la sortie du réducteur.

- **DESCRIPTION DE L'ENSEMBLE DE DONNÉES :**

Colonne 1**:** identifiant vidéo de 11 caractères.
Colonne 2**:** Téléchargeur de la vidéo
Colonne 3**:** Intervalle entre le jour d'établissement de You tube et la date de téléchargement de la vidéo.
Colonne 4**:** Catégorie de la vidéo.
Colonne 5**:** Longueur de la vidéo.
Colonne 6**:** Nombre de vues pour la vidéo.
Colonne 7**:** Note sur la vidéo.
Colonne 8**:** Nombre de notes attribuées à la vidéo
Colonne 9**:** Nombre de commentaires formulés sur les vidéos.
Colonne 10**:** Identifiants vidéo associés à la vidéo téléchargée.

- **DÉCLARATION DE PROBLÈME :**

Ici, nous allons découvrir quelles sont les 5 premières catégories avec le nombre maximum de vidéos téléchargées.

- **CODE SOURCE :**

Maintenant, à partir du mappeur, nous voulons obtenir la catégorie vidéo en tant que clé et la valeur int finale '1' en tant que valeurs qui seront transmises à la phase de lecture aléatoire et de tri et sont ensuite envoyées à la phase de réduction, où l'agrégation des valeurs est effectuée.

- **COMMENT EXECUTER**

hadoop jar top5.jar

Ici 'hadoop' spécifie que nous exécutons une commande Hadoop et jar spécifie le type d'application que nous exécutons. Top5.jar est le fichier jar créé avec le code source

Dans notre cas, le chemin du fichier d'entrée est le répertoire racine de hdfs désigné par /youtubedata.txt et l'emplacement du fichier de sortie dans lequel stocker la sortie a été défini par top5_out.

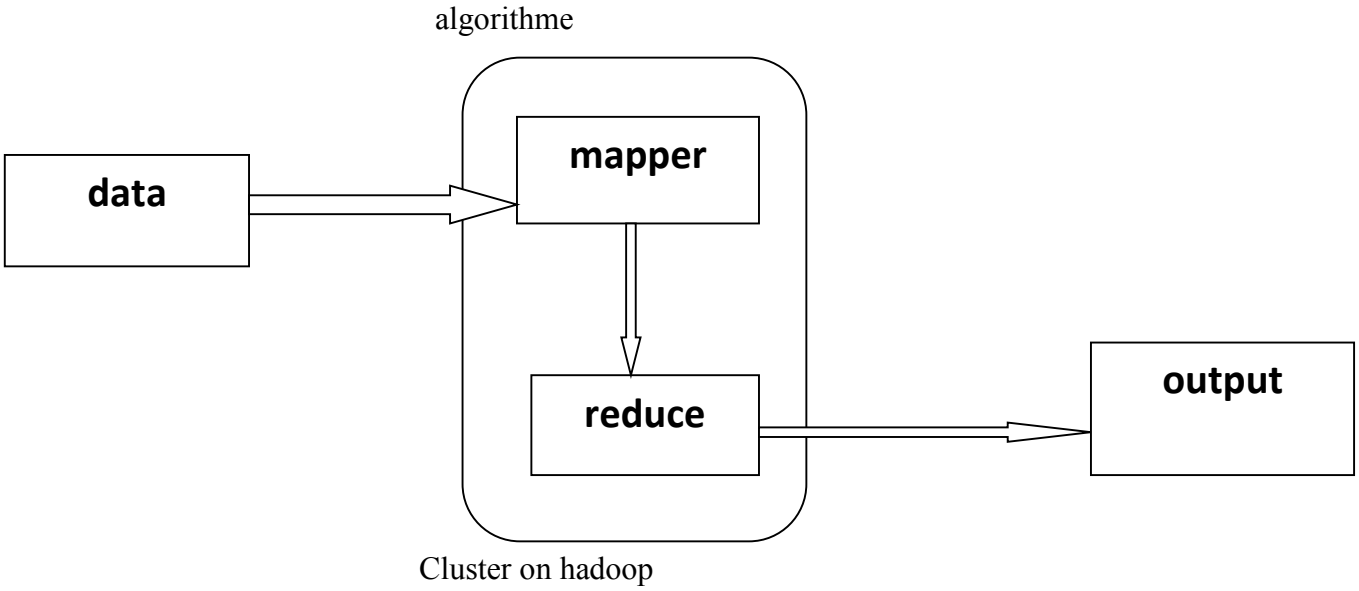
- **Comment voir la sortie**

hadoop fs -cat / top5_out

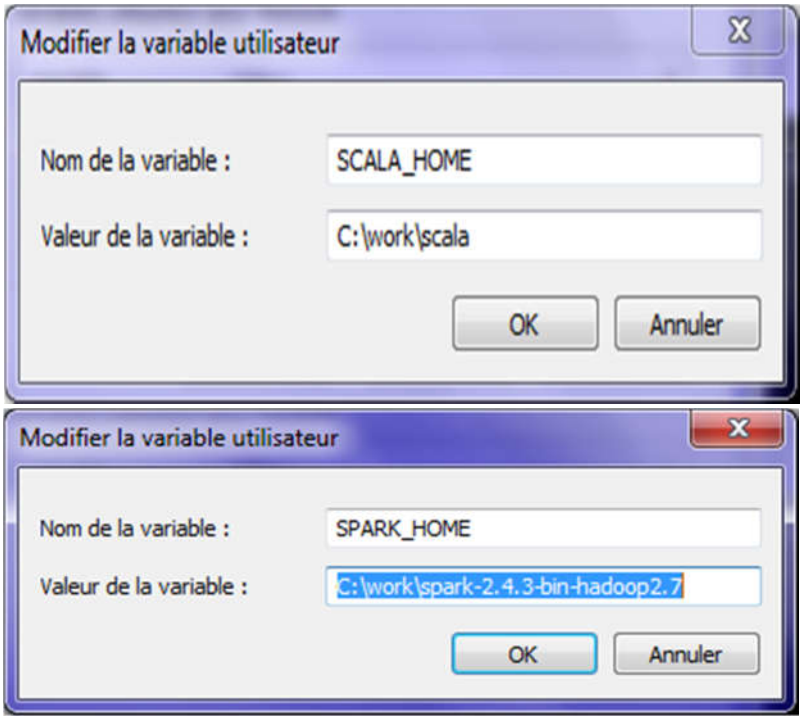
ici 'hadoop' spécifie que nous exécutons une commande Hadoop et dfs spécifie que nous effectuons une opération liée au système de fichiers distribué Hadoop et que '- cat' permet

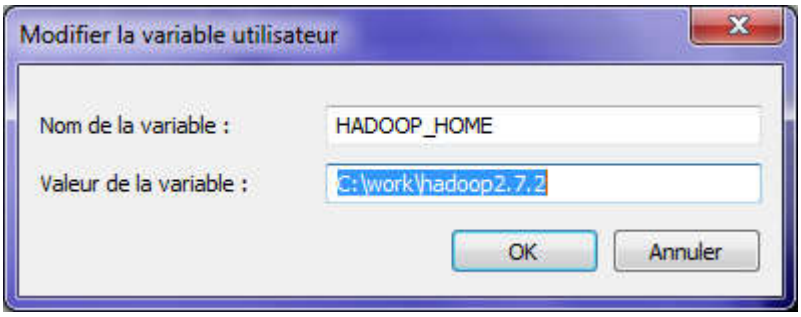
d'afficher le contenu d'un fichier et de top5_out est le fichier où la sortie est stockée. .
Ici, présente les 5 premières catégories avec le nombre maximal de vidéos téléchargées.
et on va voir le résulta final

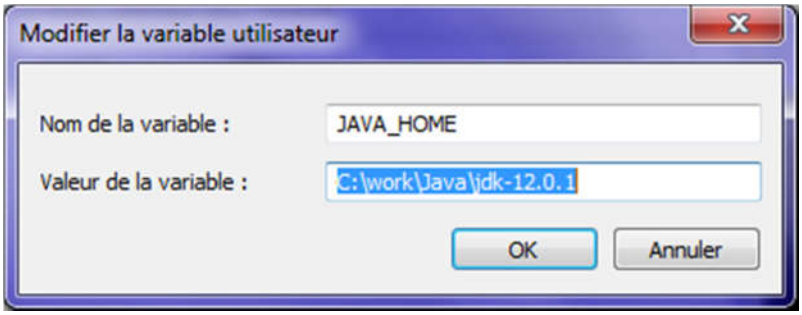
- **le chéma d'algorithme :**



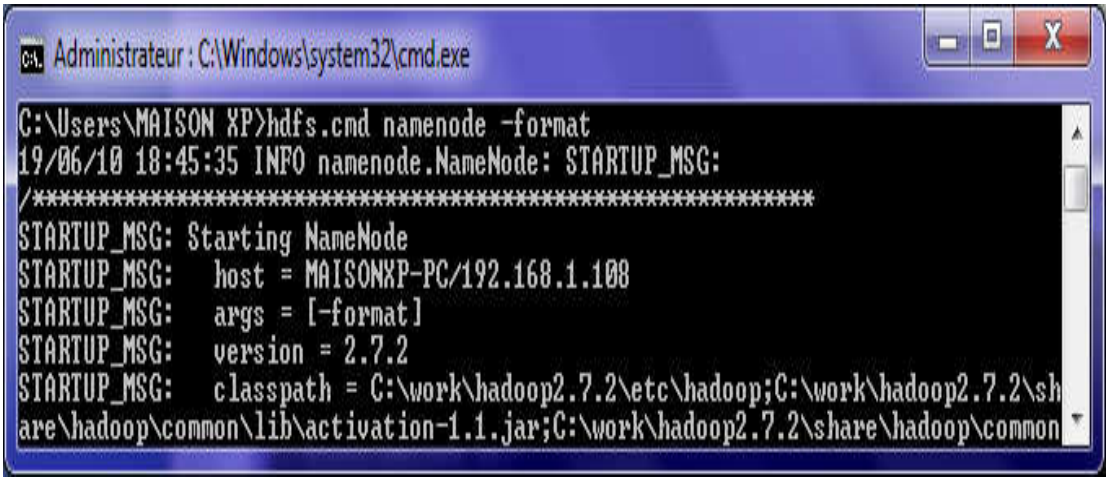
Comment installer apache spark :

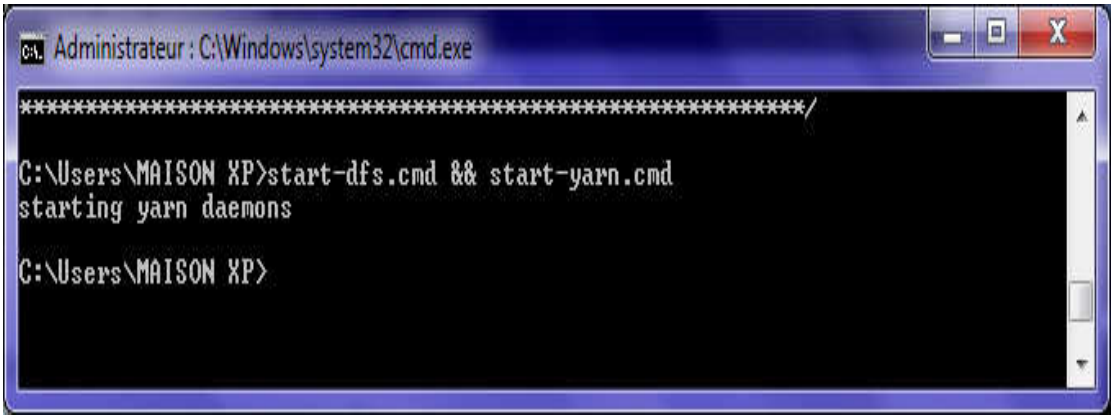


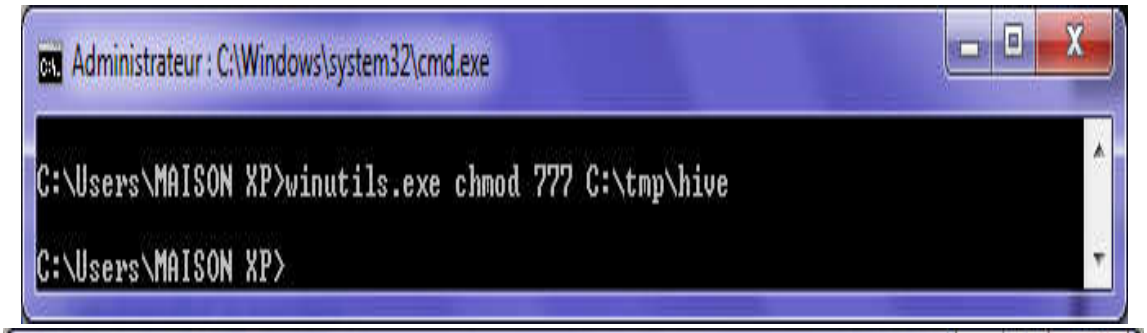


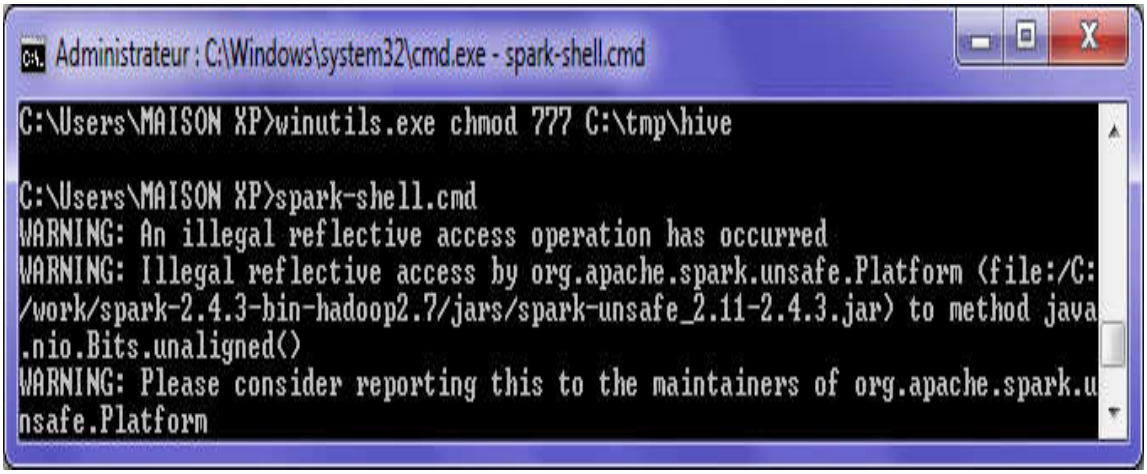


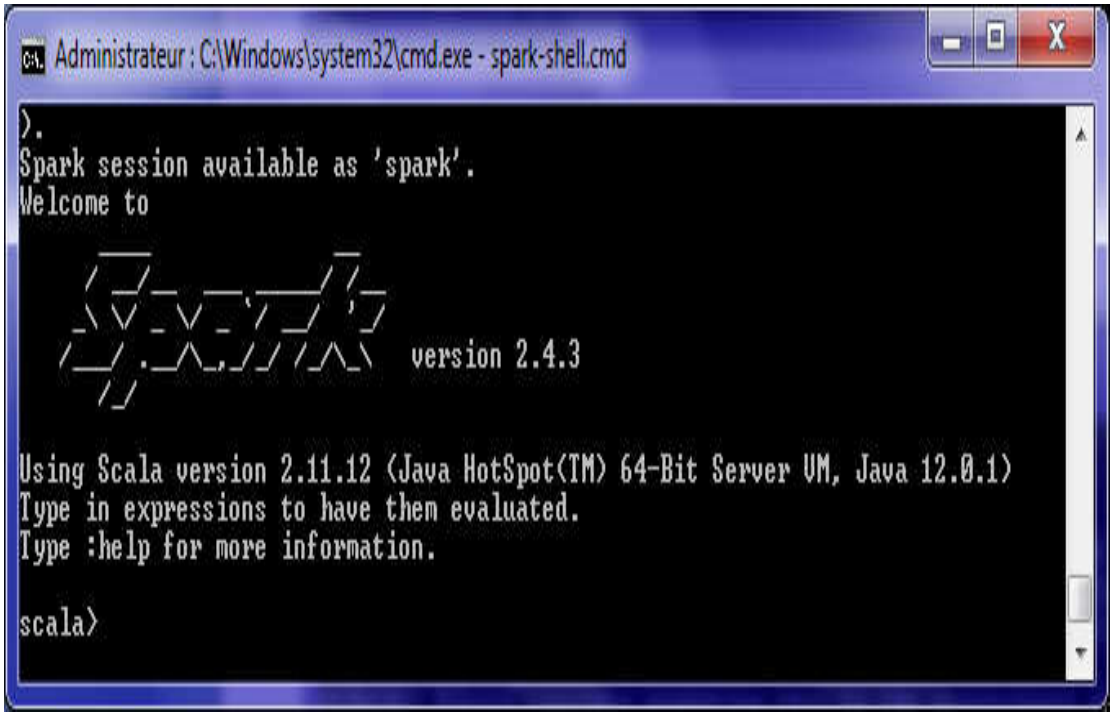
hdfs.cmd namenode -format






Administrateur : C:\Windows\system32\cmd.exe
C:\Users\MAISON XP>winutils.exe chmod 777 C:\tmp\hive
C:\Users\MAISON XP>

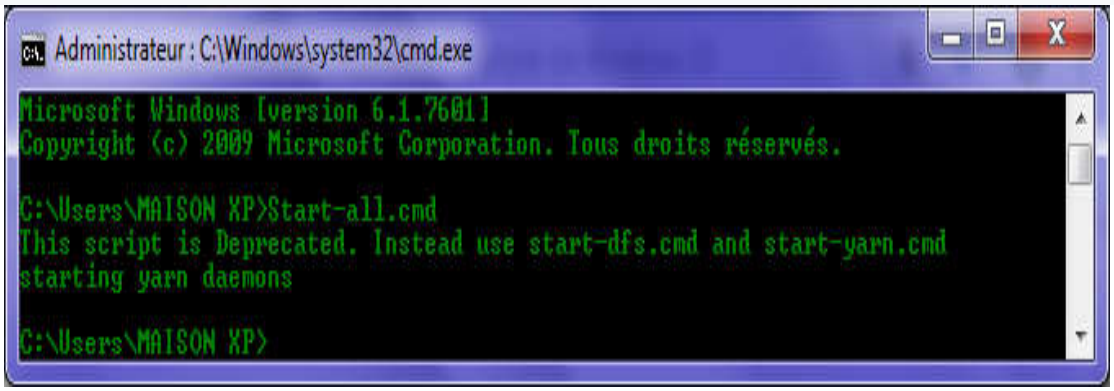

Administrateur : C:\Windows\system32\cmd.exe - spark-shell.cmd
C:\Users\MAISON XP>winutils.exe chmod 777 C:\tmp\hive
C:\Users\MAISON XP>spark-shell.cmd
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/C:/work/spark-2.4.3-bin-hadoop2.7/jars/spark-unsafe_2.11-2.4.3.jar) to method java.nio.Bits.unaligned()
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform


Administrateur : C:\Windows\system32\cmd.exe - spark-shell.cmd
).
Spark session available as 'spark'.
Welcome to
version 2.4.3
Using Scala version 2.11.12 (Java HotSpot(TM) 64-Bit Server VM, Java 12.0.1)
Type in expressions to have them evaluated.
Type :help for more information.
scala>

6) L'éxécution de l'algorithme :

1. Ouvrez cmd en mode administratif et démarrez le cluster

```
Start-all.cmd
```



2. Créez un répertoire d'entrée dans HDFS.

```
hadoop fs -mkdir /input_dir
```

3. Copiez le fichier texte d'entrée nommé fichier_entrée.txt dans le répertoire d'entrée (rép_entrée) de HDFS.

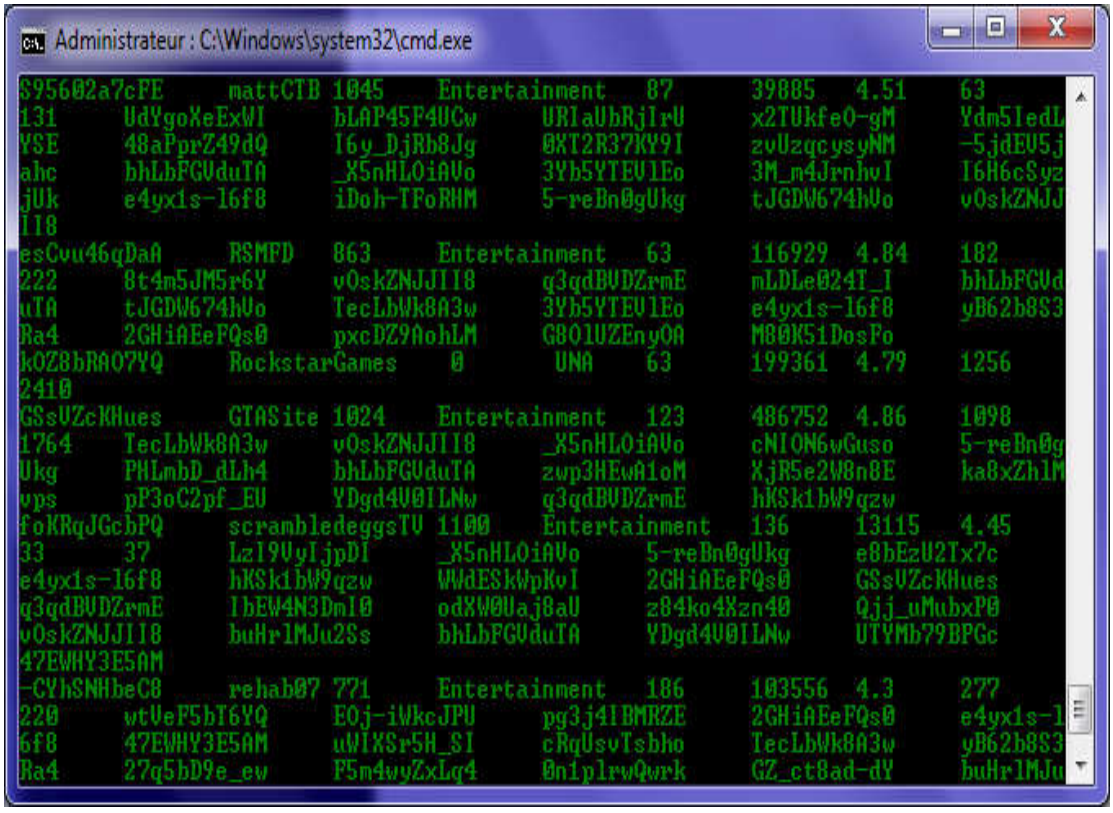
```
hadoop fs -put C:/input_file.txt /input_dir
```

4. Vérifiez input_file.txt disponible dans le répertoire d'entrée HDFS (input_dir).

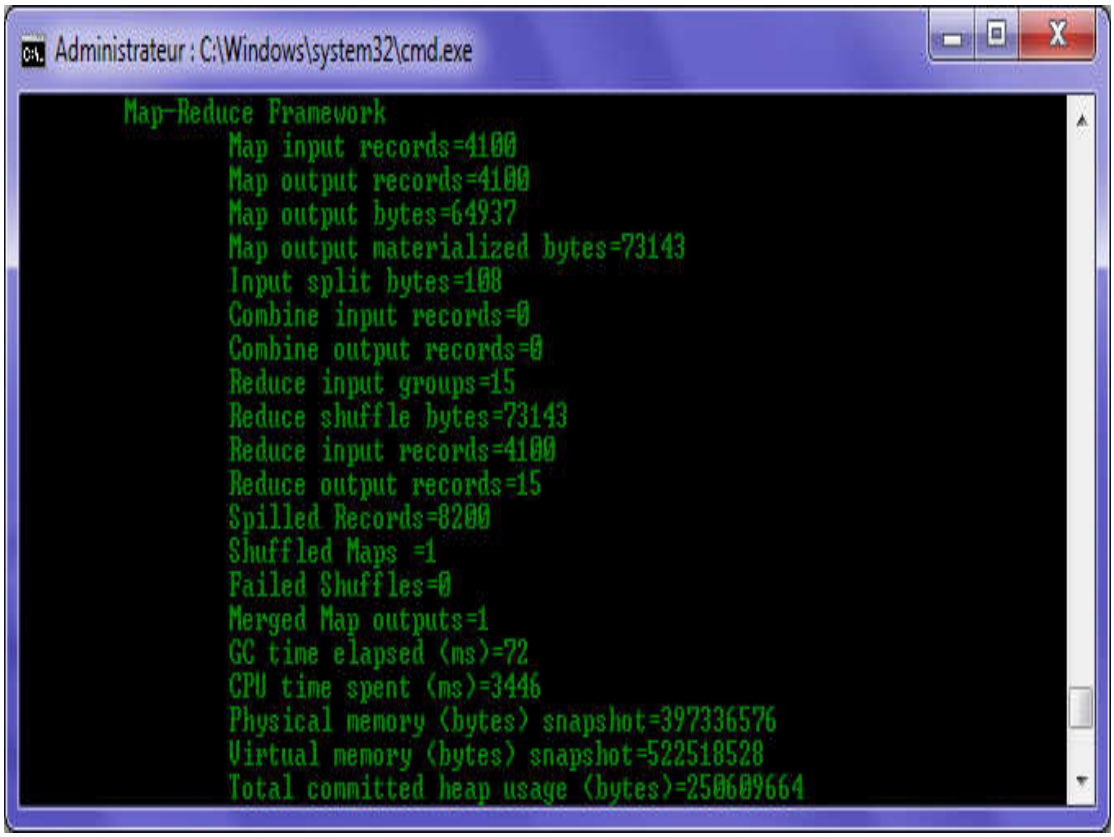
hadoop fs -ls /input_dir/

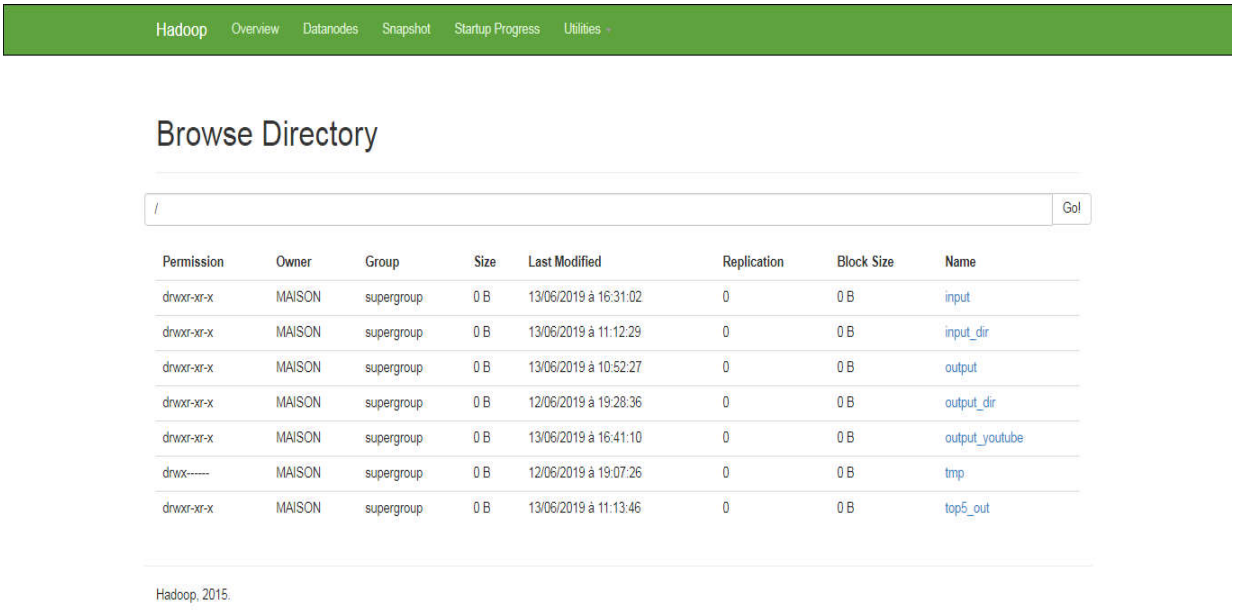
/input Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	MAISON	supergroup	946.67 KB	13/06/2019 à 16:31:01	1	128 MB	youtubedata.txt

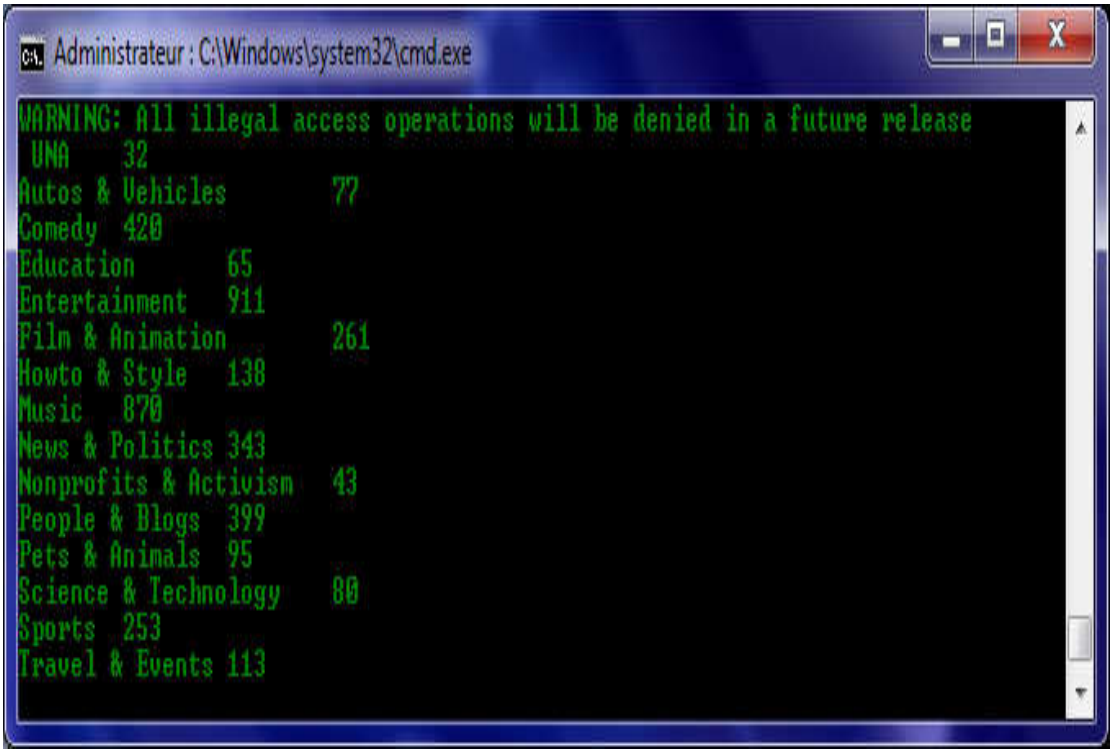


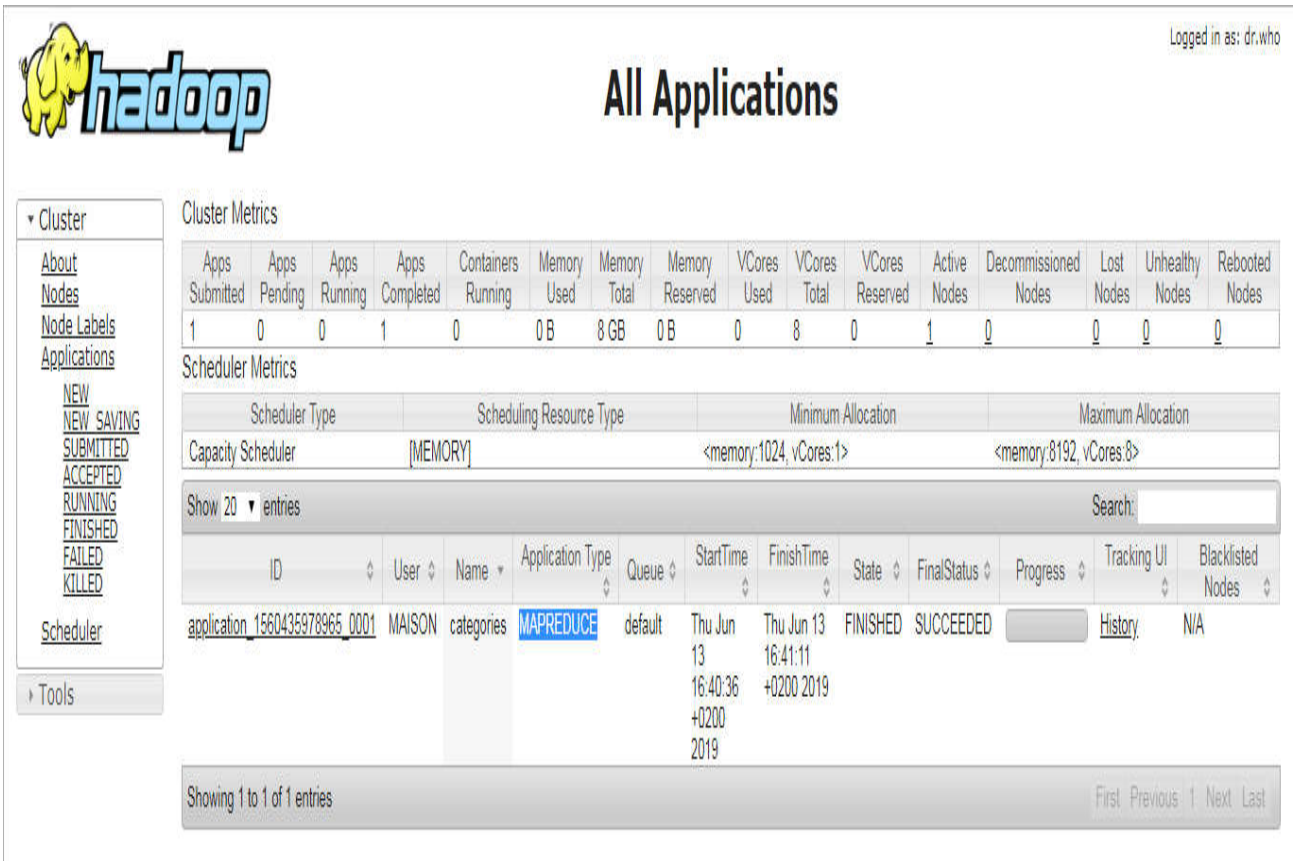
```
hadoop jar C:/MapReduceClient.jar wordcount /input_dir /output_dir
```





```
hadoop dfs -cat /output_dir/*
```




Logged in as: dr.who
All Applications
Cluster
About
Nodes
Node Labels
Applications
NEW
NEW_SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
Scheduler
Tools
Cluster Metrics
Apps Submitted | Apps Pending | Apps Running | Apps Completed | Containers Running | Memory Used | Memory Total | Memory Reserved | VCores Used | VCores Total | VCores Reserved | Active Nodes | Decommissioned Nodes | Lost Nodes | Unhealthy Nodes | Rebooted Nodes
1 | 0 | 0 | 1 | 0 | 0 B | 8 GB | 0 B | 0 | 8 | 0 | 1 | 0 | 0 | 0 | 0
Scheduler Metrics
Scheduler Type | Scheduling Resource Type | Minimum Allocation | Maximum Allocation
Capacity Scheduler | [MEMORY] | <memory:1024, vCores:1> | <memory:8192, vCores:8>
Show 20 entries
Search:
ID | User | Name | Application Type | Queue | StartTime | FinishTime | State | FinalStatus | Progress | Tracking UI | Blacklisted Nodes
application_1560435978965_0001 | MAISON | categories | MAPREDUCE | default | Thu Jun 13 16:40:36 +0200 2019 | Thu Jun 13 16:41:11 +0200 2019 | FINISHED | SUCCEEDED | | History | N/A
Showing 1 to 1 of 1 entries
First Previous 1 Next Last

Conclusion Général :

Notre projet a fait l'objet d'une étude des nouvelles technologies du cloud computing et big data, Le premier chapitre vis à donner une description globale de la technologie de Cloud Computing, Historique**,**Les caractéristiques du cloud computing et Modèles de déploiement dans le cloud computing, Ainsi son développement remarquable ces dernières années qui suscite de plus en plus l'intérêt des différents utilisateurs de l'internet et de l'informatique,

dans autre coté Nous avons abordé les principes des Big Data, ces caractéristiques, son fonctionnement ainsi que les différents domaines dans lesquels elles sont utilisées. On a aussi recensé les différents modèles de bases de données NoSQL qui existent actuellement.

La partie applicative consistait à un déploiement d'Hadoop avec ses composants MapReduce et HDFS dans un environnement distribué ,La mise en œuvre d'un cluster Hadoop avec ses composants et l'exécution d'un algorithme qui on a choisi à une seule machine.

Références Bibliographiques

[1]: John W. Rittinghouse, James F. Ransome. Cloud Computing Implementation ,Management , and Security. CRC Press 2010;

[2] : Maurice Audin. «Etat de l'art du Cloud Computing et adaptation au Logiciel Libre».2009;

[3] http://www.claranet.fr/cloud.html

[4] www.idf.direccte.gouv.fr (le cloud Computing une nouvelle filière fortement structurante)

[the NiST 2011]

[5] Mell, grance. (2011) 'The NIST Definition of Cloud Computing' Timothy Grance Special Publication 800-145, National Institute of Standards and Technology , Gaithersburg.

[6] Mather, T. Kumaraswamy, S. and Latif, S. (2009) 'Cloud Security and Privacy: An Enterprise Perspective on Risks and Compliance', O'Reilly Media, ISBN-13: 978 0596802769.

[7] http://www.yeswecloud.fr

[8] N.Grevet. Le cloud computing : évolution ou révolution ? Pourquoi, quand, comment et surtout faut-il prendre le risque ?, Août 2009.

[9] claud kesako.cloud –serveur http//www.cloud-serveur.fr/fr/le cloud/cloud-kesako,(consulté mars 2016).(cité en page viii et 17)

[10] Mathur, P. (2010) 'Cloud Computing: New challenge to the entire computer industry', Proceediongs of the 1St International Conference on Parallel, Distributed and Grid Computing, Solan, India, pp. 223-228.

[11] Jian J.Z., Chengdu, C. and Nan, Zhang. (2011) 'Cloud Computing-based Data Storage and Disaster Recovery', Proceediongs of the International Conference on Future computer science and education, Xi'an, China, pp. 629-632.

[12] Mémoire "Cloud Computing" IUT Nancy Charlemagne

[13] http://www.redsen-consulting.com/2013/06/big-data/, consulté le 9/06/ 2015.

[14]http://webcache.googleusercontent.com/search?q=cache:2DfXDe1Z_mEJ:www.aubay.com/fileadmin/user_upload/Publications_FR/Regard_Aubay__Big_Data_Web.pdf+&cd=1&hl=ar&ct=clnk&gl=dz, consulté le2/03/2015.

[15]M.CORINUS, T.Derey, J.Marguerie, W.Techer, N.Vic, Rapport d'étude sur le Big Data, SRS Day, 54p, 2012.

[16] Big data application and archetecteure; de himanshu et soumendra mohanty.

[17] Mekideche Mounir, Conception et implémentation d'un moteur de recherche à base d'une architecture Hadoop (Big Data), Avril 2015.

[18]. Hurwitz, J, Nugent, A, Halper, F, and Kaufman, M (2013). Big data for dummies? Ebook. https://eecs.wsu.edu/~yinghui/mat/courses/fall%202015/resources/Big%20data%20for%20dummies.pdf.

[19]. Fermigier, S (2012). Big Data and Open Source: Une Convergence Inévitable. Livre Blanc (http://fermigier.com/blog/2012/03/new-whitepaper-big-data-open-source/).

[20] Big Data for Dummies, par Wiley Brand.

[21] http://fr.wikipedia.org/wiki/Entrep%C3%B4t_de_donn%C3%A9es,consulté le 9/04/2015.

[22] http://www.journaldunet.com/solutions/analytics/big-data, consulté le1/03/ 2015.
[23]Bernard ESPINASSE, Introduction aux systèmes NoSQL (Not Only SQL), Ecole Polytechnique Universitaire de Marseille, 19p, Avril 2013.
[24] http://www.technologies-ebusiness.com/langages/les-bases-no-sql, consulté le.6/06/2015
[25] http://fr.wikipedia.org/wiki/Hadoop, consulté le 10/04/2015.
[26] http://mbaron.developpez.com/tutoriels/bigdata/hadoop/introduction-hdfs-map-reduce/, consulté le 27/04/ 2015.
[27]https://databricks.com/spark/about
[28]https://amplab.cs.berkeley.edu/

1. Ouvrez cmd en mode administratif et déplacez-vous vers "C: /Hadoop-2.8.0/sbin" et démarrez le cluster

```
Start-all.cmd
```

2. Créez un répertoire d'entrée dans HDFS.

```
hadoop fs -mkdir /input_dir
```

3. Copiez le fichier texte d'entrée nommé fichier_entrée.txt dans le répertoire d'entrée (rép_entrée) de HDFS.

```
hadoop fs -put C:/input_file.txt /input_dir
```

4. Vérifiez input_file.txt disponible dans le répertoire d'entrée HDFS (input_dir).

```
hadoop fs -ls /input_dir/
```

5. Vérifier le contenu du fichier copié.

```
hadoop dfs -cat /input_dir/input_file.txt
```

6. Exécutez MapReduceClient.jar et fournissez également des répertoires d'entrée et de sortie.

```
hadoop jar C:/MapReduceClient.jar wordcount /input_dir /output_dir
```

7. Vérifier le contenu du fichier de sortie généré.

```
hadoop dfs -cat /output_dir/*
```

Résumé :

La révolution technologique intégrant de multiples sources d'informations, la vulgarisation de l'informatique dans les différents secteurs et domaines ont amené à l'explosion de la volumétrie des données, qui reflète le changement d'échelle des volumes, du nombre et de types. Ces accroissements massifs ont poussé à l'évolution des manières de gestion, de stockage, de localisation et d'accès aux données. Les dernières étapes de cette évolution informatique ont émergé de nouvelles technologies : Cloud Computing et Big Data.
Cloud computing peut être définie comme l'utilisation des ressources informatiques (matériels et logiciels) via Internet (virtuellement), où il est prévu sur la forme des services.dans autre coté Le Big Data est un ensemble de technologies basées sur les bases de données NoSQL « Not Only SQL » permettant le passage à grande échelle en volumes, en nombres et en types de données
dans la partie applicative on a propose un algorithme de youtube pour analysée les données avec l'hadoop et on a utlisée mapreduce

mot clé : Cloud Computing, BigData, NoSQL , Hadoop, MapReduce

abstract :

The technological revolution integrating multiple information sources and extension of computer science in different sectors led to the explosion of the data quantities, which reflects the scaling of volumes, numbers and types. These massive increases have resulted in the development of new locations techniques and access to data. The final steps in this evolution have emerged new technologies : Cloud Computing and Big Data.
Cloud computing can be defined as the use of computing ressources (hardware and software)via the internt (virtualy),where it is planned on the form of services.in other side
Big Data is a set of technologies based on NoSQL databases allowing scalability of volumes, numbers and types of data.
In the applicatif part we have offers the algorithme of youtube for for analyze big data with hadoop and map reduce.

Keywords: Cloud Computing, BigData, NoSQL , Hadoop, MapReduce

ملخص :

أدت الثورة التكنولوجية التي أدمجت المصادر المتعددة للمعلومات وكدا تعميم تكنولوجيا المعلومات في مختلف القطاعات و المجالات الى انفجار حجم البيانات مما يعكس تغير السلم على مستوى الأحجام ،الأعداد و الأنواع و قد أدت هده الزيادة الهائلة الى تغييرات في طرق التسيير ,تخزين و التوصل الى موقع البيانات,المراحل الأخيرة من تطور المعلوماتية أفرز تقنيات جديدة :الحوسبة السحابية و البيانات الضخمة
و يمكن أن نعرف الحوسبة السحابية على أنها استخدام المصادر الحوسبية (العتاد و البرمجيات) عن طريق الأنترنت (افتراضية) حيث أنها مقدمة على شكل خدمات, من جهة أخرى البيانات الضخمة هي مجموعة من القنيات المبنية على قواعد بيانات no sql ليس فقط sql مما يسمح بقياس واسع النطاق في الأحجام و الأرقام و أنواع البيانات
في الجزء التطبيقي لدينا خوارزمية اليوتيوب بحيث نقوم بتحليل بياناته الضخمة ببرناج hadoop و باستخدام mapreduce

الكلمات المفتاحية : NoSQL , Hadoop, MapReduce , البيانات الضخمة,الحوسبة السحابية