République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université A. MIRA - Bejaïa
Faculté des Sciences Exactes
Département d'Informatique

جامعة بجاية
Tasdawit n Bgayet
Université de Béjaïa

Mémoire de fin de cycle

En vue de l'obtention du diplôme de Master en informatique

Spécialité : Administration et Sécurité des Réseaux Informatique

Thème :

Authentification et Autorisation dans un environnement BIG DATA Sous Hadoop

Réalisé par :

- YAKOUBEN GHILAS
- KERNAF HICHAM

Devant le jury composé de :

Examinatrice :	Dr EL BOUHISSI Houda	M.C.A – Université de Bejaïa.
Examinatrice :	Dr CHIBANI Samia	M.C.A – Université de Bejaïa.
Encadrante :	Dr BATTAT Nadia	M.C.B – Université de Bejaïa.

Promotion 2021-2022

Remerciement

En premier lieu, on remercie le bon Dieu de nous avoir donné la volonté, le courage, la persistance et la patience de réaliser et finaliser ce modeste travail.

On adresse un vif remerciement à Madame **BATTAT.N** d'avoir accepté de nous encadrer, aussi pour ses orientations et ses conseils au long de ce travail. Nos remerciements s'adressent également aux membres du jury d'avoir accepté d'évaluer ce travail.

On tient également à remercier :

Tous ceux qui ont participé de près ou de loin,

de façon directe ou indirecte, à la réalisation de ce projet, nos enseignants, nos amis,

nos collègues et toute la promotion de 2022.

Enfin, nous remercions l'ensemble du personnel du département d'informatique et tout le corps professoral de l'Université Abderrahmane Mira de Bejaia pour la qualité de leur enseignement et pour les valeurs qu'ils nous ont inculquées et tous ceux qui ont contribué de près ou de loin à la concrétisation de ce travail.

DÉDICACE

Je dédie ce modeste travail à toutes personnes qu'on aime:

Essentiellement à mes ***CHERS PARENTS*** *qui sont la source de la tendresse, de la patience, de la générosité*

et ceux qui m'ont appris le secret de la réussite, pour leurs sacrifices et encouragements pendant ma formation et que dieu les protègent et les gardes en bonne santé.

Mes chers frères et sœurs pour leurs amours et leurs conseils et soutient.

A mon cher binôme ***HICHAM*** *avec qui j'ai partagé cette période en beauté.*

Mes chers collègues et amis ***YANIS***, ***FARINAS***, ***GHANI***.

Qui m'ont vraiment soutenue et aidé durant cette période

En fin, à toutes les personnes qui comptent pour moi, qui ont intervenu dans ma vie et qui m'ont accompagné et soutenu soit de près ou de loin, vous êtes une source de force pour moi

Y.GHILAS

DÉDICACE

A l'aide de DIEU, le tout puissant

Ce travail est achevé, je le dédie à toutes

Personnes qu'on aime

À ***MON PÈRE*** *et à* ***MA MÈRE****.*

L'honneur de ce travail revient à mes très chers parents pour leur affection, leur sacrifices et encouragements pendant ma formation et que dieu les protège et les garde en bonne santé.

A mes frères ***TOUFIK, BRAHIM ET SALIM****.*

A ma sœur ***SAMIA***

A mon meilleur ami et collègue ***GHILAS*** *qui m'a soutenu et aidé pour la réalisation de ce modeste travail.*

A tous mes amis : ***RAHIM, KACI, AHMED ET OKBA****.*

A tous ceux qui n'ont aidé et contribué à ma formation.

A toutes les personnes qui m'ont vraiment soutenue et aidé même si de loin ; vous êtes une source de force pour moi.

K.HICHAM

Table des matières

Liste des Figures

Liste des Tableaux :

Résumé :

Apache Hadoop est l'une des plates-formes les plus populaires pour le traitement des Big Data en utilisant du matériel de base pour l'analyse et le traitement des données. Les organisations ont créé leurs propres clusters pour gérer leurs données. La sécurité de cette plateforme est la plus grande préoccupation qui tourne autour de la protection des données et des services. Les données augmentent de jour en jour, tout comme les risques associés aux Big Data. Il devient important de fournir des mesures de sécurité efficaces pour les données. Ce mémoire fait une synthèse des différents aspects et protocole d'authentification et d'autorisation. Dans la partie applicative de ce mémoire, nous avons mis en place le protocole d'authentification et d'autorisation ranger. Ensuite, nous avons présenté les étapes nécessaires à son utilisation.

Mots clés: Big Data, Hadoop, HDFS, MapReduce, authentification, autorisation, Kerberos, ranger.

Abstract:

Apache Hadoop is one of the most popular platforms for Big Data processing using commodity hardware for data analysis and processing. Organizations have created their own clusters to manage their data. The security of this platform is the biggest concern that revolves around the protection of data and services. As the data increases day by day, so do the risks associated with Big Data. It becomes important to provide effective security measures for the data. This dissertation makes an overview of the different aspects and protocol of authentication and authorization. In the application part of this dissertation, we have setted up the ranger authentication and authorization protocol. Then, we presented the necessary steps to use it.

Key words: Big Data, Hadoop, HDFS, MapReduce, authentication, authorisation, Kerberos, ranger.

Liste des abréviations:

ABAC	Attribute-based Access Control
AD	Active Directory
ADN	Acide DésoxyriboNucléique
AMPLab	Algorithms, Machines and People Lab
API	Application Programming Interface
AS	Authentication Server
CA	*Certificate Authority*
CERN	European Council for Nuclear Research
cm	cloudera manager
CPU	Central Processing Unit
DBMS	data base management system
DAC	Discretionary Access Control
DB	Data Base
ECDSA	Elliptic Curve Digital Signature Algorithm
GCP	Google Cloud Platform
GFS	Google File System
GPS	Global Positioning System
HA	High Available
HBase	Hadoop Database
HCs	Hadoop Clusters
HDFS	Hadoop Distributed File System
HDP	Hadoop
HFT	High-frequency trading
HTTP	Hypertext Transfer Protocol
IP	Internet Protocol
IBM	International Business Machines
IoT	Internet of Thing
JDK	Java Development Kit
JSON	JavaScript Object Notation
JVM	Java Virtual Machine
KDC	Key Distribution Center
LDAP	Lightweight Directory Access Protocol

MAC	Mandatory Access Control
MD5	Message Digest algorithm 5
MIT	Massachusetts Institute of Technology
Mo	Mégaoctet
MR	MapReduce
No-SQL	No-Structured Query Language
NSA	National Security Agency
PIN	Personal Identification Number
RBAC	Role-Based Access Control
REST	representational state transfer
RDBMS	Relational data base management system
RH	Human resources
RHEL	Red Hat Enterprise Linux
RSA	Rivest Shamir Adleman
SAS	Statistical Analysis System
SEC	Securities and Exchange Commission
SHA	Secure Hashing Algorithm
SSH	Secure Socket Shell
SSL	secure sockets layer
SSO	Single Sign On
SPSS	Statistical Package for the Social Sciences
SQL	Structured Query Language
TB	TeraByte
TGS	Ticket Granting Server
TGT	Ticket Granting Ticket
TLS	Transport Layer Security
TS	Service Ticket
URI	Universal Resource Identifier
URL	Uniform Resource Locators
U-SQL	Unified-Structured Query Language
XML	Extensible Markup Language
SGBDR	Système de Gestion de Base de Données Relationnelle

Introduction générale :

Le monde actuel est confronté à une explosion importante de données. Selon les statistiques faites par le site planetoscope.com en 2022, on trouve que, chaque seconde, près de 65.000 recherches qui sont faites sur le moteur de recherche Google par les internautes, près de 43 000 vidéos qui sont visionnées sur le site de vidéos YouTube, 29.000 Giga-octets (Go) d'informations sont publiés dans le monde, environ 2320 Tweets qui sont postés puis expédiés sur Tweeter, 29.000 Giga-octets (Go) d'informations sont publiés dans le monde.[94]

Le développement et l'accès à ces données a conduit à l'apparition du terme Big Data qui possède ses origines dans le Data Science et le Cloud Computing. Ce phénomène impacte en particulier les entreprises qui sont amenées à manipuler des Téraoctets voir des Pétaoctets de données nécessitant une infrastructure spécifique pour leur création, leur stockage, leur traitement, leur analyse et leur récupération. En d'autres termes, il s'agit du développement en temps réel d'une masse de données volumineuse qui dépasse la capacité des outils de traitement et d'analyse traditionnels (bases de données relationnelles, … etc.) [95]. Ce qui nécessite l'utilisation des plateformes et outils dédiés à la gestion de ces données parmi lesquels la plateforme Hadoop qui intègre le stockage et le traitement des données, la gestion du système et un outil d'entreposage de données.[66]

Hadoop est un Framework open-source pour le stockage et le traitement des Big data dans un environnement distribué. *Il* dispose d'une architecture maître-esclave pour le stockage des données et le traitement distribué des données à l'aide l'utilisation des MapReduce et HDFS. MapReduce est un Framework logiciel dérivé de Java, pour analyser les données à grande échelle. Il utilise un modèle de traitement des données distribuées. HDFS est un autre composant d'Hadoop, qui stocke de grands volumes de données.[96]

Alors que l'adoption d'Apache Hadoop s'accélère, les capacités d'authentification et d'autorisation sont une préoccupation majeure pour la sécurité de l'accès aux données.[97]

Problématique :

Dans ce contexte, de nombreux travaux et recherches portent sur des solutions basées sur le protocole Kerberos pour l'authentification et le contrôle d'accès. D'autres travaux sont basés sur Kerberos pour assurer l'authentification combiné avec des outils de sécurité (ranger, sentry, etc.…) pour assurer l'autorisation.

Aujourd'hui, nous avons le choix entre plusieurs protocoles différents. Il est facile de voir comment quelqu'un pourrait être confus et même frustré en essayant de choisir celui à adopter. Chacun de ces protocoles offre plusieurs avantages et atouts, mais aussi il n'existe pas un protocole

parfait, chaque protocole a des inconvénients également. Il est essentiel de comparer ces protocoles en considérant un ensemble de critères (comme, temps de traitement) afin de fournir des indicateurs et des mesures aux personnes intéressées (chef d'une entreprise, etc.). Ces indicateurs et des mesures peuvent aider ces personnes à prendre de décisions sur le protocole d'authentification et d'autorisation à mettre en place.

Contribution

L'objectif initial de ce projet était de comparer au moins trois protocoles d'authentification et d'autorisation. Cependant, nous avons pu seulement mettre en place un seul protocole qui est ranger.

Structure de ce mémoire

Ce mémoire est composé de (03) trois chapitres :

- Le premier chapitre donne une vision globale sur l'état de l'art de Big Data, qui portera sur quelques définitions, ainsi que les domaines d'applications et l'architecture, puis on présentera sa technologie Hadoop.
- Dans le deuxième chapitre nous allons présenter les techniques et méthodes d'authentification et d'autorisation. Ensuite nous présenterons certains travaux de recherche concernant l'authentification et l'autorisation dans Hadoop.
- Dans le troisième chapitre, nous détaillerons les étapes de la mise en place du protocole ranger ensuite nous présenterons les étapes nécessaires à son utilisation.
- Nous clôturons ce mémoire par une conclusion générale et des perspectives.

CHAPITRE 1

1.1 INTRODUCTION:

Lorsque les systèmes de bases de données relationnelles traditionnels n'ont pas pu traiter les données non structurées (blogs, vidéos, photographies, mises à jour sociales et activité humaine) crées par les organisations, les médias sociaux ou toute autre source générant des données, le big data est né. Le terme "big data" désigne des données dont le volume est énorme, dont la variété est diverse et qui se déplacent à un rythme élevé. Il ne s'agit pas d'une chose, mais d'un concept ou d'un paradigme qui définit la collecte et l'utilisation croissantes de différents ensembles de données.

Dans ce chapitre, nous présentons le big data et ses concepts, ainsi que les domaines d'application, les approches d'analyse des données et les avantages et problèmes liés au big data.

1.2 Historique :

Le terme "big data" a été inventé avant le développement de la technologie des bases de données, en raison de la nécessité de trouver des solutions pour faire face à l'afflux massif d'ensembles de données et, par conséquent, à la rareté de l'espace de stockage.

Comme le montre la **figure 1.1**, le concept de Big data a évolué au fil des ans, chaque décennie étant définie en termes d'espace disque informatique, allant du mégaoctet (Mo) dans les années 1970 à YottaByte (YBs) en 2022. [1,2,3 ,4]

A partir de l'an 2010, le volume de données augmente de façon exponentielle dicté par la variété et de nombreuses sources de données numérisées.

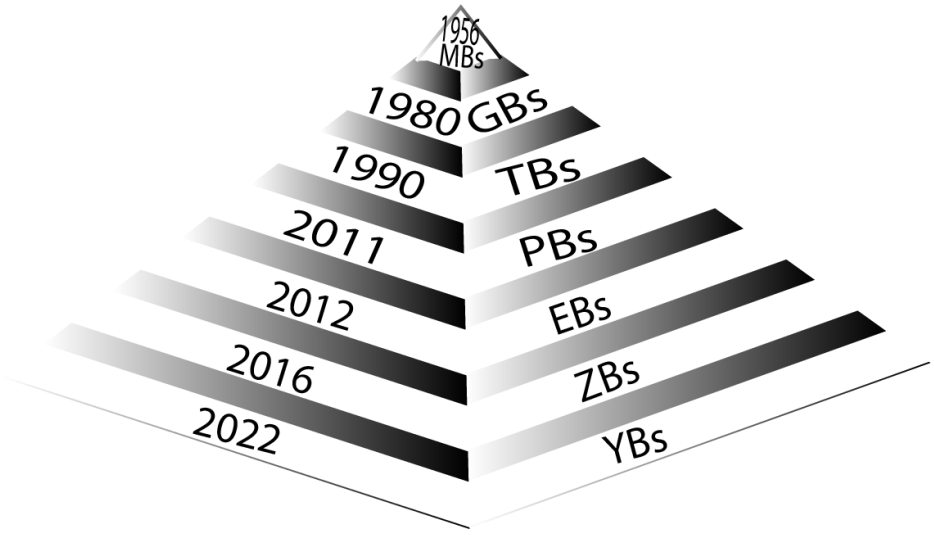


Figure 1. 1 Historique du big data [1]

Le cadre du Big Data tente de diviser l'évolution du Big Data en trois phases distinctes :[5]

Phase 1.0 : Le big data était principalement défini par le stockage et l'analyse des données, et il était considéré comme une extension des systèmes de gestion de bases de données existants et de la technologie d'entreposage des données. [5]

Phase 2.0 : Suite à l'essor du Web 2.0 et à la diffusion de contenus semi-structurés et non structurés, le concept de Big data s'est développé pour englober des solutions techniques avancées permettant d'extraire des informations utiles de types de données disparates et hétérogènes.[5]

Phase 3.0 : avec l'introduction des smartphones et des appareils mobiles, des données de capteurs, de l'Internet des objets (IoT), des dispositifs portables et d'une foule d'autres générateurs de données, le Big Data est entré dans une nouvelle ère et a ouvert un tout nouvel éventail de possibilités. [5]

Le tableau 1.1[5] présente un résumé des trois phases proposées pour le Big Data.

Phase 01	Phase 02	Phase 03
Période 1970-2000	Période 2000-2010	Période 2010-présent
Contenu structure, basé sur DBMS (data base management system) : -RDBMS (Relational DBMS) & warehousing. -Charge de transfert d'extraits. -Traitement analytique en ligne. -Tableau de bord et analyse statistique.	Contenu non structuré, basé sur le web : -Recherche et extraction d'informations. -Exploitation des opinions. -Réponse aux questions. -Analyse et intelligence web. -Analyse des media sociaux. -Analyse des réseaux sociaux. -L'analyse spatio-temporelle.	Contenu mobile et basé sur des captures : -Analyse de localisation. -Analyse centré sur la personne. -Analyse contextuelle. -Visualisation mobile. -Interaction homme-machine.

Tableau 1. 1 Les trois phases du Big data [5]

1.3 DEFINITIONS :

Il n'existe pas de définition unique du big data.

- Le big data est un domaine qui traite des méthodes d'analyse, d'extraction méthodique d'informations ou de traitement des volumes de données qui sont trop importants ou trop compliqués pour les logiciels d'application de traitement de données typiques. Les données

comportant de nombreux champs (colonnes) ont une plus grande puissance statistique, mais les données comportant de nombreux attributs ou colonnes ont un taux de fausse découverte plus élevé [6].

- " Le big data est un terme populaire utilisé pour caractériser le développement exponentiel, la disponibilité et l'utilisation de l'information, à la fois structurée et non structurée ", selon SAS. [7]
- "Les données, qui arrivent de partout ; les capteurs utilisés pour recueillir des informations climatiques, les messages sur les sites de réseaux sociaux, les images et vidéos numériques, les enregistrements de transactions d'achat et le signal GPS des téléphones portables, pour n'en citer que quelques-uns", voilà comment IBM décrit le Big Data. [8]

1.4 Architecture du Big Data :

L'architecture du big data est la disposition qui permet d'ingérer, de traiter et d'analyser les données de manière optimale. En d'autres termes, l'architecture du big data est le pivot de l'analyse des données et permet aux outils d'analyse du big data d'extraire des informations vitales de données autrement obscures et de prendre des décisions commerciales significatives et stratégiques.[9]

Voici un bref aperçu de certains des composants les plus courants de l'architecture big data [10] :

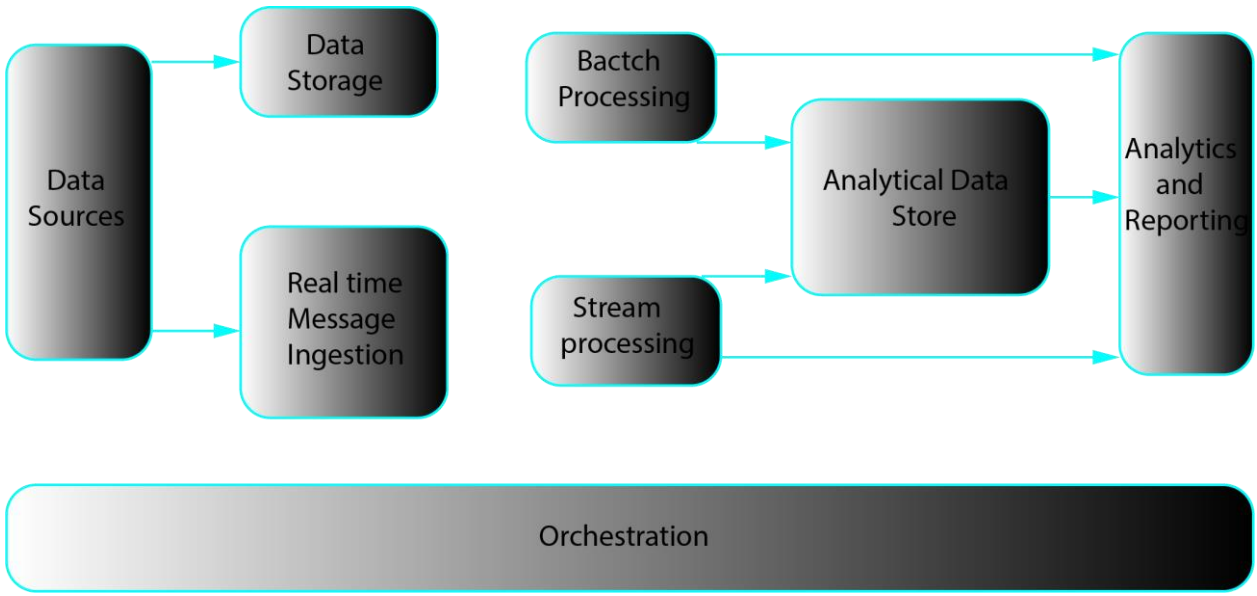


Figure 1. 2 Composants les plus courants de l'architecture big data[10]

1. Sources de données(Data Source) :

Les sources de données comprennent toutes les sources en or à partir desquelles le pipeline d'extraction de données est construit. On peut donc dire que c'est le point de départ du pipeline de big data.

2. Stockage des données(Data Storage) :

Les données qui sont gérées pour les opérations construites par lots sont stockées dans les magasins de fichiers qui sont distribués par nature et sont également capables de contenir de grands volumes de fichiers volumineux de formats différents. On l'appelle le lac de données.

3. Traitement par lots(Batch Processing) :

Toutes les données sont séparées en différentes catégories ou morceaux qui font appel à des travaux de longue haleine utilisés pour filtrer et agréger et également préparer les données à l'état traité pour l'analyse. Ces tâches utilisent généralement des sources, les traitent et fournissent la sortie des fichiers traités aux nouveaux fichiers. Le traitement par lots est effectué de diverses manières, en utilisant des tâches Hive ou des tâches basées sur U-SQL, ou en utilisant Sqoop ou Pig avec les tâches de réducteur de carte personnalisées qui sont généralement écrites dans l'un des langages suivants : Java, Scala ou tout autre langage tel que Python.

4. Ingestion de messages en temps réel(Real Time-Based Message Ingestion):

Il s'agit souvent d'un simple entrepôt ou magasin de données responsable de tous les messages entrants qui sont déposés dans le dossier nécessairement utilisé pour le traitement des données. Il existe cependant une majorité de solutions qui requièrent un magasin d'ingestion basé sur les messages qui agit comme un tampon de messages et prend également en charge le traitement à l'échelle, fournit une livraison relativement fiable ainsi que d'autres sémantiques de mise en file d'attente des messages. Parmi les options possibles, citons Apache Kafka, Apache Flume, les concentrateurs d'événements d'Azure, etc.

5. Traitement en flux(Stream Processing):

Il existe une légère différence entre l'ingestion de messages en temps réel et le traitement en continu. Le premier prend en considération les données ingérées qui sont d'abord collectées et ensuite utilisées comme un outil de type publish-subscribe. Le traitement en continu, quant à lui, est utilisé pour traiter toutes les données en continu qui se produisent dans des fenêtres ou des flux, puis écrit les données sur le puits de sortie. Cela inclut Apache Spark, Apache Flink, Storm, etc.

6. Stockage de données à des fins analytiques(Analytics-Based Datastore):

Il s'agit du magasin de données utilisé à des fins analytiques. Les données déjà traitées sont donc interrogées et analysées à l'aide d'outils analytiques qui peuvent correspondre aux solutions de Big

Data. Les données peuvent également être présentées à l'aide d'une technologie d'entrepôt de données NoSQL comme HBase ou toute utilisation interactive de la base de données Hive qui peut fournir l'abstraction de métadonnées dans le magasin de données. Les outils comprennent Hive, Spark SQL, Hbase, etc.

7. Rapports et analyses(Reporting and Analysis):

Les informations doivent être générées à partir des données traitées, ce qui est fait efficacement par les outils de rapport et d'analyse qui utilisent leur technologie et leur solution intégrées pour générer des graphiques, des analyses et des informations utiles aux entreprises. Ces outils comprennent Cognos, Hyperion, etc.

8. Orchestration :

Les solutions basées sur le Big Data consistent en des opérations liées aux données qui sont répétitives par nature et sont également encapsulées dans les flux de travail qui peuvent transformer les données sources et également déplacer les données à travers les sources ainsi que les puits et charger dans les magasins et pousser dans les unités analytiques. Les exemples incluent Sqoop, oozie, data factory, etc.

1.5 Caractéristiques :

Les diverses représentations du problème des Big data ont naturellement conduit à fournir une pléthore de perceptions techniques sur le paradigme des Big data. Cela a ouvert les portes à plusieurs descriptions du Big data dans des contextes différents afin de mieux comprendre le paradigme du Big data, ses défis et ses avantages, et donc d'en obtenir la véritable valeur. Pour le différencier des systèmes traditionnels de traitement des données, le Big data a été amplement caractérisé par les fameux 3V (Volume, Vélocité et Variété). [11]

Cependant, les entreprises ont trouvé ces trois dimensions moins adéquates pour traiter correctement les Big data et donc insuffisantes pour fournir les informations précieuses espérées.
Les informations précieuses espérées. C'est pourquoi sept caractéristiques supplémentaires ont été proposées, ce qui a permis d'obtenir une représentation consolidée du problème des Big Data. Ces caractéristiques sont principalement la véracité, la variabilité, la valeur, la validité, la vulnérabilité, la volatilité et la visualisation. Voici une brève discussion sur les 10 V du Big Data.

Volume : fait allusion à l'augmentation massive de la croissance des données. [12]

Vélocité : fait référence à la collecte rapide de données provenant de diverses sources de données en temps quasi réel et en temps réel. La vélocité englobe à la fois des caractéristiques temporelles et de retard. [13]

La figure 1.3 illustre la quantité de données générées par minute en 2019 à partir de diverses sources.

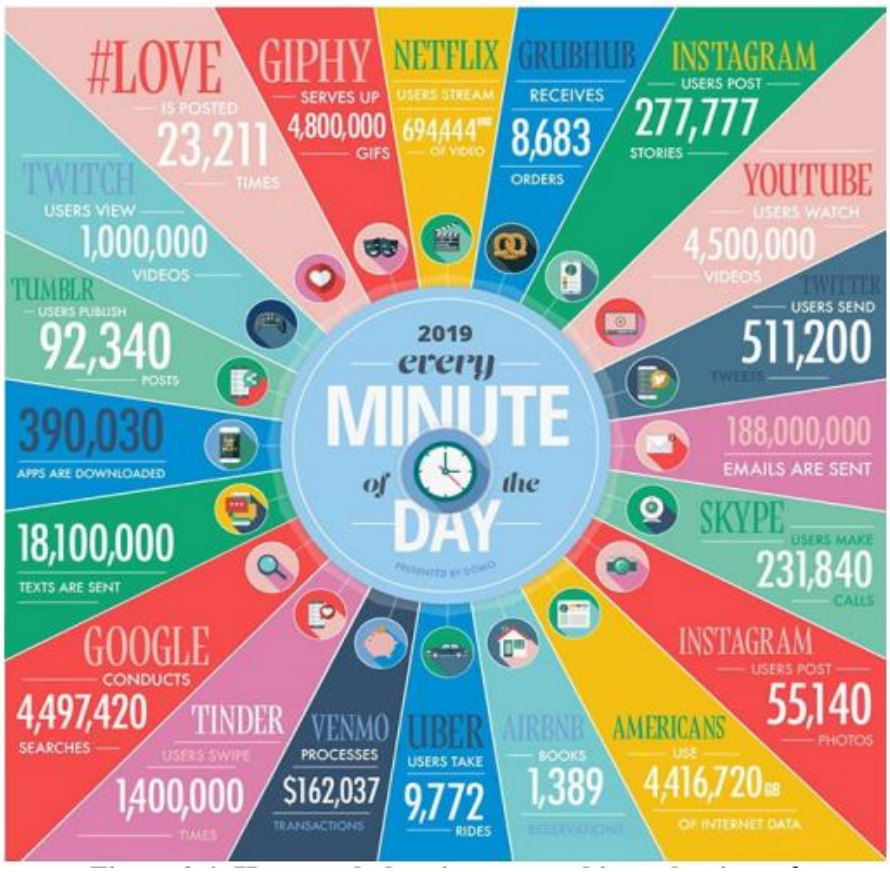


Figure 1. 3 Combien de données sont générées chaque minute[14]

Variété : il faut recueillir des informations à partir de diverses sources et dans divers formats [15]. Cela inclut l'importation de données dans une variété de formats, notamment structurés (tableaux dans des bases de données relationnelles - SGBDR, par exemple), semi-structurés (courrier électronique, XML, JSON et autres langages de balisage, par exemple) et non structurés (données dans une variété de formats comme :texte, images, fichiers audio, vidéo, données de capteurs, etc.) L'importation de jeux de données à partir d'autres référentiels fait également partie de la variété.

La véracité : fait référence à la provenance, l'exactitude et la justesse des données notamment l'objectivité vs la subjectivité, la véracité vs la tromperie et la crédibilité vs l'invraisemblance [16] [17]. Ces caractéristiques incluent, mais ne sont pas limitées à : (i) la fiabilité de l'origine des données ; (ii) la fiabilité et la sécurité du stockage des données ; et (iii) l'accessibilité des données.

Variabilité : fait référence aux différences de sens, à la quantité d'incohérences, au nombre de dimensions des données et à la vitesse à laquelle les données sont reçues. [18-21]

Validité : elle est définie comme "des faits dont il a été démontré (ou dont on sait) qu'ils constituent une indication précise de l'affirmation faite" [22]. La validité diffère de la véracité en ce qu'elle "signifie l'exactitude et la précision des données par rapport à leur utilisation prévue" [23]. En d'autres termes, on peut faire confiance aux données, ce qui satisfait à l'exigence de véracité. Cependant, une interprétation erronée des données pourrait entraîner une application non intentionnelle. En outre, les mêmes données véridiques peuvent être valables pour une application mais non valables pour une autre. [24]

Vulnérabilité : Ce terme fait référence à la sécurité des ensembles de données qui seront collectés et analysés ultérieurement [25]. Il fait également référence aux faiblesses du système qui permettent d'effectuer des actes nuisibles sur les ensembles de données obtenus. Par conséquent, l'acquisition de jeux de données doit garantir la disponibilité de systèmes immunitaires capables de protéger les données recueillies contre les intrusions. [25]

Volatilité : fait référence à la durée pendant laquelle les données peuvent être conservées et utilisées avant qu'elles ne deviennent obsolètes ou non pertinentes [26]. C'est un facteur important car le coût du stockage et de l'entretien augmente en fonction de la durée de conservation des Big data. [27]

Visualisation : désigne la capacité de fournir des Big data dans un contexte visuel, tel que des diagrammes, des graphiques, des cartes, etc. afin de mieux comprendre et interpréter les données [28]. Elle aide également les individus et les organisations à identifier les modèles, les corrélations, les tendances, les liens et les interdépendances. Les décideurs peuvent utiliser la visualisation des big data pour accéder à d'énormes données, les analyser, les comprendre en temps réel et agir en conséquence. [29]

Valeur : Ce terme fait référence au résultat final d'une analyse de big data (c'est-à-dire de nouveaux aperçus).[30]

L'importance du Big Data va au-delà des données liées aux entreprises pour inclure les données générées par les entreprises industrielles, les établissements d'enseignement, les données politiques et gouvernementales, les données sur les soins de santé, et une variété d'autres industries.[31]

1.6 Classification des données:

Les données sont classées dans des catégories appropriées afin d'être utilisées ou utilisées plus efficacement. La catégorisation des données permet à l'utilisateur de trouver facilement ce qu'il cherche. Lorsqu'il s'agit de la sécurité et de la conformité des données, ainsi que de la réalisation de divers types d'objectifs d'entreprise ou personnels, la classification des données est essentielle. C'est également un élément important car les données doivent être accessibles dans un certain délai.[32]

1.6.1 Types de classification des données :

Les données peuvent être divisées en trois catégories [32]:

1.6.1.1 Les données structurées :

Les données structurées sont des données stockées dans une structure qui définit leur format. Elles peuvent être stockées sous forme de tableaux et sont préparées à l'aide d'un schéma prédéterminé. Un exemple typique peut être considéré comme un système de gestion de base de données relationnelle (SGBDR), des données de transaction, des fichiers de données tels que des feuilles de calcul. Un autre exemple, Microsoft Excel est un excellent outil relativement simple pour travailler avec des données structurées.[32]

1.6.1.2 Données non structurées :

Elles sont décrites comme des données qui n'adhèrent pas à une norme prédéterminée ou qui ne suivent aucun format ordonné. Ce type de données ne convient pas non plus à une base de données relationnelle puisque les données de cette dernière sont structurées de manière prédéfinie. Les données non structurées sont également très importantes dans le secteur du big data, et il existe plusieurs systèmes pour les gérer et les stocker, comme les bases de données No-SQL.
Word, PDF, texte, journaux de médias, etc. en sont quelques exemples.[32]

1.6.1.3 Les données semi-structurées :

Les données semi-structurées sont des données qui ne sont pas stockées dans une base de données relationnelle mais qui possèdent certaines qualités organisationnelles qui facilitent leur examen. Certains processus peuvent être stockés dans une base de données relationnelle, mais les données semi-structurées sont difficiles à stocker, bien que les données semi-structurées existent pour économiser de l'espace.

Les données au format XML en sont un exemple.[32]

1.6.2 Caractéristiques de la classification des données :

L'objectif principal de l'organisation des données est de structurer les données de manière à ce qu'elles soient accessibles aux utilisateurs. Par conséquent, elle présente les caractéristiques de base suivantes.[32]

-Homogénéité : Les éléments de données de chaque groupe doivent être similaires les uns aux autres.

-Clarté : Il ne doit y avoir aucune ambiguïté dans le placement d'un élément de données dans un groupe.

-Stabilité : La collecte des données doit être stable, ce qui signifie que toute étude ne doit pas avoir d'impact sur le même ensemble de classifications.

-Elasticité : La base de la catégorisation doit pouvoir être modifiée lorsque l'objectif de la classification change.

1.7 DOMAINE D'APPLICATION :

Au cours des dernières années, le Big Data a changé la donne dans la plupart, sinon tous les types d'industries modernes. Alors que le Big Data devient de plus en plus omniprésent dans notre vie quotidienne, l'attention s'est détournée du battage médiatique pour se concentrer sur la découverte d'une véritable valeur dans son application.

Les principaux domaines d'application du Big Data sont:

- **Banque et valeurs mobilières :**

Le Big Data est utilisé par la Securities and Exchange Commission (SEC) pour surveiller les activités des marchés financiers. Elle détecte actuellement les activités commerciales illicites sur les marchés financiers par le biais de l'analyse de réseau et des processeurs de langage naturel.

Le Big Data est aussi utilisé par les traders de détail, les grandes banques, les fonds spéculatifs et d'autres marchés financiers pour l'analyse des transactions, l'analyse d'aide à la décision avant les transactions et la surveillance des sentiments, etc. [33]

- **Santé :**

Le secteur de santé a accès à de vastes quantités de données, mais il a du mal à les utiliser pour contrôler l'augmentation des coûts, ainsi que les systèmes inefficaces qui empêchent des prestations

de santé plus rapides et de meilleure qualité dans tous les domaines, principalement parce que les données électroniques sont soit indisponibles, soit insuffisantes, soit inadaptées. En outre, les bases de données sur la santé qui enregistrent les données relatives à la santé ont rendu impossible l'établissement de liens entre les données susceptibles de révéler des modèles utiles pour le corps médical. Parmi les autres défis du Big Data, citons l'exclusion des patients de la prise de décision et l'utilisation de données provenant de divers capteurs facilement disponibles.[33]

- **Assurance :**

L'un des domaines d'application directe du Big Data est l'assurance, où des statistiques et l'analyse du comportement à risque de millions de personnes sont nécessaires. La capacité de récolter des quantités massives de données sur la vie des gens peut être utilisée pour créer un modèle de vie pour chaque individu : style de vie, conduite de voiture, amendes, utilisation de l'électricité, relations professionnelles, etc. Ces modèles permettent aux compagnies d'assurance-vie d'améliorer leurs produits et leurs opérations, voire de réaliser des enquêtes plus approfondies.[34]

- **Gestion des catastrophes naturelles :**

L'une des applications les plus fascinantes du Big Data est la capacité d'analyser les données météorologiques en temps réel ; cette technique permet de suivre et de visualiser les mouvements des ouragans, ainsi que de prédire où ils vont frapper. Par conséquent, les gouvernements locaux et les groupes humanitaires internationaux peuvent préparer les ressources nécessaires (couverture, fournitures et médicaments), ainsi que les modes de transport et les interventions rapides, pour aider les personnes dans le besoin.[34]

- **Prévenir les cyber-attaques :**

Aujourd'hui, les approches Big Data pour l'analyse des données sont devenues essentielles pour détecter les intrusions, les failles de sécurité et les cyber-attaques, car le volume de données envoyées sur Internet a énormément augmenté, s'est diversifié et nécessite un traitement en temps réel [35].

- **Contrôle d'épidémies :**

Le big data peut aider à contrôler la propagation des épidémies dans le monde entier en surveillant par exemple la migration des insectes porteurs de maladies à travers le monde. Les big data sont également utilisées pour la chasse aux rats dans les grandes villes comme New York ou Chicago, où la police locale utilise des systèmes de big data pour le suivi visuel et l'analyse des parcours de rats afin de contrôler leur croissance.[34]

1.8 Techniques d'analyse de données :

Certaines approches d'analyse des données sont énumérées ci-dessous. [36]

Analyse statistique : Il s'agit de l'une des méthodes les plus populaires puisqu'elle peut être appliquée à des ensembles de données de la bibliothèque, qu'ils soient petits ou grands. Il s'agit d'un jeu de chiffres, mais les chercheurs la définissent comme une science qui consiste à acquérir, analyser et démontrer d'énormes volumes de données afin d'identifier les tendances et les modèles fondamentaux. Cette stratégie peut être utilisée dans n'importe quelle situation. Un rédacteur qui rédige un document d'étude, par exemple, peut utiliser l'analyse statistique pour évaluer les données reçues de ses clients et améliorer son travail.[37]

Analyse diagnostique : L'objectif de cette méthode est d'obtenir un examen approfondi de la situation et de connaître les sources d'un certain événement, concept ou occurrence. Les entreprises peuvent utiliser les diagnostics pour identifier les relations entre les mesures de l'entreprise. Par exemple, il est courant de voir des entreprises de rédaction de contenu s'appuyer sur des recherches de diagnostic pour comprendre pourquoi les meilleures critiques de résumé sont meilleures ou pires que prévu.[37]

Analyse descriptive : Il s'agit d'une expression pour l'analyse des données qui aide à décrire, montrer ou résumer les données d'une manière compréhensible afin que des modèles puissent émerger. Cette stratégie est fréquemment utilisée pour suivre les indicateurs clés de performance. L'objectif est simple dans ce cas, ce qui implique que la personne dispose d'un ensemble clair d'indicateurs clés de performance et de statistiques descriptives qui indiquent les résultats basés sur les données réelles de l'entreprise.[37]

Analyse prédictive : est un type d'analyse de données qui est principalement utilisé pour trouver des tendances et prévoir les résultats futurs de l'entreprise. Au lieu de dépendre de son intuition pour prendre des décisions, la personne peut utiliser cette analyse pour prendre des décisions fondées sur les données. Il s'agit d'un type de modélisation statistique qui permet d'élaborer des conclusions pertinentes sur un sujet donné. Cette stratégie peut être utilisée pour l'évaluation des risques et la prévision des ventes.[37]

Analyse narrative : Si une personne préfère effectuer une analyse commerciale en utilisant des mots plutôt que des chiffres, cette méthode peut être idéale pour l'entreprise. L'objectif majeur de cette enquête est d'examiner les idées, les opinions, les attitudes et les histoires. En conséquence, l'organisation peut découvrir des préférences critiques parmi les employés et leur logiciel RH, ainsi que reconsidérer la culture entière de l'entreprise.[37]

Analyse prescriptive : L'objectif fondamental de l'intelligence économique est de trouver de nouvelles et meilleures techniques pour faire de meilleurs jugements. Pour cette raison, l'analyse prospective est incroyablement utile, mais elle est aussi difficile. L'analyse prédictive et l'analyse descriptive sont combinées dans cette méthode. Elle peut être définie comme la branche de l'analyse d'entreprise qui s'occupe de déterminer le plan d'action optimal dans une situation donnée. L'inconvénient logistique de ce système est qu'il nécessite une équipe importante et un budget conséquent.[37]

Analyse de texte : processus assisté par ordinateur permettant de décoder de grandes quantités de données non structurées afin de découvrir des idées, des modèles et des tendances dans des données quantitatives. L'exploration de texte est un autre nom pour cette technologie. Cette stratégie est fréquemment utilisée en conjonction avec des supports de visualisation de données, ce qui permet d'obtenir une représentation plus précise de la procédure souhaitée. Étant donné qu'Internet regorge de divers types de contenu textuel, une entreprise peut utiliser cette stratégie pour interpréter et évaluer le contenu lié à la marque. Cela permet à de nombreuses entreprises de maintenir leur réputation en ligne malgré l'interaction avec un grand nombre de clients du monde entier [38].

Analyse du contenu : Si une personne est impliquée dans l'analyse de données qualitatives, alors elle doit essayer cette méthode. Elle est utilisée pour déterminer l'existence de certains mots, concepts, thèmes au sein de certaines données qualitatives fournies. Cette tactique est généralement

utilisée dans l'analyse textuelle. Elle est donc bien connue des clients des services d'aide à la rédaction de documents et de devoirs. Cette méthode est excellente pour les entreprises qui souhaitent comprendre leurs consommateurs, car elle aide les décideurs à connaître la signification réelle des enquêtes, des critiques et d'autres types de commentaires. [39]

1.9 Technologie du Big Data :

La technologie Big Data est définie comme une plateforme et un utilitaire logiciel conçus pour l'analyse, le traitement et l'extraction d'informations à partir d'un grand nombre de structures extrêmement complexes et de grands ensembles de données, ce qui est très difficile à gérer pour les systèmes traditionnels. La technologie du big data est utilisée pour traiter les données en temps réel et les données par lots. L'apprentissage automatique est devenu un élément essentiel de la vie quotidienne et de tous les secteurs d'activité. Par conséquent, la gestion des données par le biais du Big Data devient très importante.[40]

Voici quelques plateformes les plus connues [41]:

Apache Storm :

Est un Framework de calcul et de traitement de flux distribué écrit principalement dans le langage de programmation Clojure. Initialement créé par Nathan Marz et l'équipe de Back Type, le projet a été ouvert après avoir été acquis par Twitter. Il utilise des «becs» et des «boulons» créés sur mesure pour définir les sources d'information et les manipulations afin de permettre le traitement par lot et distribué de données en continu. La première publication a eu lieu le 17 septembre 2011.[41]

Teradata :

Est une société informatique américaine qui vend des plateformes de données analytiques, les applications et les services connexes. Ses produits sont destinés à consolider les données provenant de différentes sources et de rendre les données disponibles pour l'analyse. Les services proposés par Teradata pour le Big Data sont: Concentrer et Unifier les données. [41]

SPSS (Statistical Package for the Social Sciences):

Également connu sous le nom de IBM SPSS Statistics, est un logiciel utilisé pour l'analyse de données statistiques, il a été créé en 1968 par SPSS Inc. et a été acquis par IBM en 2009. Bien que le nom de SPSS reflète son utilisation originale dans le domaine des sciences sociales, son utilisation s'est depuis étendue à d'autres marchés de données. SPSS est couramment utilisé dans les domaines de la santé, du marketing et de l'éducation.

SPSS fournit des analyses de données pour les statistiques descriptives et bivariées, les prédictions de résultats numériques et les prédictions d'identification de groupes. Le logiciel offre également des fonctions de transformation des données, de création de graphiques et de marketing direct.[42]

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. Développé à l'université de Californie à Berkeley par AMPLab, 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.[41]

Hadoop :

Crée par Doug CUTTINGen2006, Hadoop est un Framework open source écrit en java qui permet le traitement distribué de grands ensembles de données sur des grappes d'ordinateurs en utilisant des modèles de programmation simples. L'application Hadoop fonctionne dans un environnement qui fournit un stockage et un calcul distribué sur des grappes d'ordinateurs. Hadoop est conçu pour passer d'un seul serveur à des milliers de machines, chacune offrant des possibilités de calcul et de stockage local.[43]

Cette plateforme est :

— Convient pour l'analyse de données volumineuses : comme les Big Data ont tendance à être distribuées et non structurées par nature, les clusters Hadoop sont les mieux adaptés à l'analyse des Big Data. Comme c'est la logique de traitement (et non les données réelles) qui circule vers les nœuds de calcul, la consommation de bande passante du réseau est moindre. Ce concept est

appelé "concept de localité des données", ce qui permet d'accroître l'efficacité des applications basées sur Hadoop. [43]

— Évolutive : les clusters Hadoop peuvent être facilement mis à l'échelle en ajoutant des nœuds de cluster supplémentaires et permettent ainsi la croissance des Big Data. De plus, la mise à l'échelle ne nécessite pas de modifications de la logique d'application. [43]

— une plateforme tolérant les fautes : l'écosystème Hadoop a une disposition permettant de reproduire les données d'entrée sur d'autres nœuds du cluster. Ainsi, en cas de défaillance d'un nœud de grappe, le traitement des données peut toujours se poursuivre en utilisant les données stockées sur un autre nœud de grappe. [43]

- **Composants de Hadoop :**

Hadoop dispose d'une architecture maître-esclave pour le stockage des données et le traitement distribué des données à l'aide l'utilisation des MapReduce et HDFS. [43]

A. Hadoop Distributed File System (HDFS) : est un système de fichiers distribués, extensible et portable pour Hadoop inspiré par le Google File System (GFS). Il a été conçu pour stocker de très gros volumes de données dans le cluster. Il se diffère du reste des systèmes de gestion de fichier traditionnel par les principales caractéristiques suivantes : [43]

- Portabilité : le système de fichiers HDFS est indépendant du noyau du système d'exploitation.
- Distributivité : HDFS est un système distribué où chaque nœud d'un cluster correspond à un sous-ensemble du volume global de données du cluster.
- HDFS utilise des tailles de blocs largement supérieures à ceux des systèmes classiques.
- HDFS fournit un système de réplication des blocs dont le nombre de réplications est configurable.

Le système de fichiers HDFS contient une architecture maître/esclave. Cette architecture consiste souvent en un seul NameNode qui joue le rôle de maître, et plusieurs DataNodes qui jouent le rôle d'esclave.[43]

— NameNode (le maître) : Il s'agit d'un nœud maître unique qui existe dans le cluster et qui s'occupe de gérer l'état du HDFS et l'espace de noms du système de fichiers en effectuant une opération telle que l'ouverture, le renommage et la fermeture de fichiers. [43]

— SecondaryNameNode : est aussi un nœud maître. il est chargé de la maintenance des nœuds du cluster, où il garde sous contrôle l'espace disque utilisé, limite la charge processeur du NameNode et il permet la continuité de fonctionnement du cluster Hadoop en cas de panne. [43]

— DataNode (l'esclave) : Le cluster HDFS contient plusieurs DataNodes (nœuds esclaves), chaque DataNode contenant plusieurs blocs. Ces blocs sont utilisés pour stocker des données. Il est de la responsabilité du DataNode de lire et d'écrire les requêtes des clients du système de fichiers. Il effectue la création, la suppression et la réplication des blocs sur instruction du NameNode.[43]

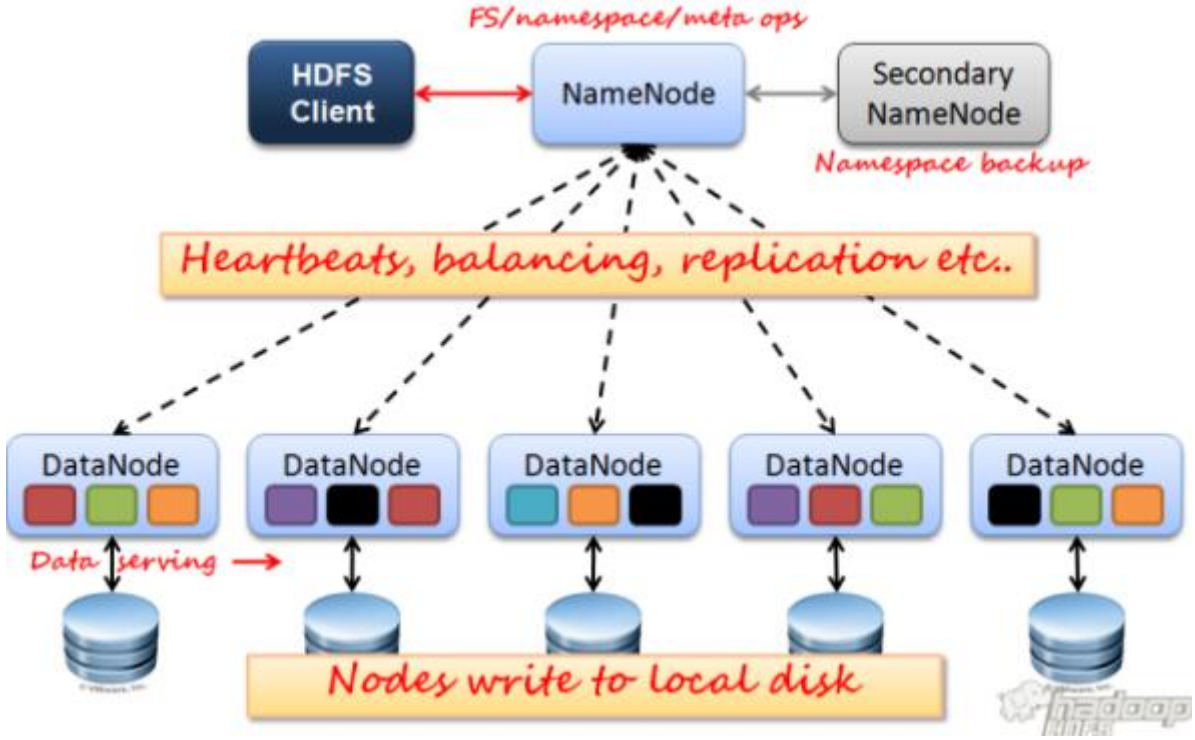


Figure 1. 4 Architecture possible de HDFS [44]

B. MapReduce :

Est un cadre logiciel et un modèle de programmation utilisé pour le traitement d'énormes quantités de données du HDFS. Le programme MapReduce fonctionne en deux phases, à savoir Map et Reduce [43] :

— **Map :** prend une paire clef/valeur du lecteur d'entrée, effectue un calcul sur celle-ci, puis produit le résultat sous forme de paire clef/valeur également. Les résultats des tâches de Map sont d'abord transmis à la mémoire tampon de la mémoire principale, puis lorsqu'elle est presque pleine, ils sont transférés sur le disque. La fonction Map s'écrit de la manière suivante : Map (clé1, valeur1) → List (clé2, valeur2). [43]

— **Reduce** : est définie par l'utilisateur est invoquée une fois pour chaque clé distincte et est appliquée sur l'ensemble des valeurs associées à cette clef ; c'est-à-dire que les paires ayant la

même clé seront traitées comme un seul groupe. La fonction Reduce s'écrit de la manière suivante : Reduce (clé2, List (valeur2)) → List (valeur2). [43]

Le MapReduce possède une architecture maître-esclave [43]:

— Le maître JobTracker qui :

— Ordonnance les différentes tâches des jobs soumis ;

— Assigne les tâches aux TaskTrackers ;

— Gère l'ensemble des ressources du système ;

— Reçoit les jobs des clients.

— L'esclave TaskTracker permet de

— Exécuter des tâches dans une autre JVM (Child) ;

— Heartbeat avec le JobTracker ;

— Exécuter les tâches données par le Jobtracker ;

1.10 Les avantages et les limites :

1.10.1 Les avantages :

Le Big Data présente d'importants avantages sociétaux, scientifiques et technologiques. Il s'agit de la manière dont il est appliqué aux êtres humains. Voici quelques-uns de ces avantages [45] :

Comprendre les clients : Le big data est utilisé pour mieux comprendre les actions et les préférences des clients. Pour acquérir une vision plus complète de leurs clients, les entreprises sont désireuses de compléter leurs ensembles de données standard avec des données de médias sociaux, des journaux de navigation, des analyses de texte et des données de capteurs. Dans de nombreuses circonstances, l'objectif principal est de développer des modèles de prédiction.

Comprendre et optimiser les processus d'entreprise : Les big data sont également de plus en plus utilisées pour améliorer les processus d'entreprise. Les détaillants peuvent utiliser des projections dérivées des données des médias sociaux, des tendances de recherche sur le Web et des prévisions météorologiques pour maximiser leurs stocks. L'analyse des big data est également utilisée pour améliorer les processus opérationnels des ressources humaines (RH). Il s'agit notamment d'employer des outils de big data pour optimiser l'acquisition de personnel ainsi que pour surveiller la culture d'entreprise et l'engagement des employés.

Améliorer la science et la recherche : Les nouvelles possibilités offertes par le big data transforment actuellement la science et la recherche. Par exemple, le CERN, le laboratoire suisse de physique nucléaire, possède le plus grand et le plus puissant accélérateur de particules du monde, le Grand collisionneur de hadrons. Les expériences visant à découvrir les secrets de notre univers – comment il a commencé et fonctionne – produisent des volumes massifs de données. Pour analyser ses 30 pétaoctets de données, le centre de données du CERN dispose de 65 000 unités centrales. Mais pour évaluer ces données, il utilise la capacité de calcul de milliers d'ordinateurs répartis dans 150 centres de données à travers le monde. Une telle capacité de calcul est susceptible de modifier un large éventail de domaines scientifiques et de recherche.

Améliorer le milieu de la santé et la santé publique : La puissance de calcul et l'analyse des big data nous permet de décoder des chaînes entières d'ADN en quelques minutes et nous permettra de trouver de nouveaux remèdes et de mieux comprendre et prévoir les schémas pathologiques. Les essais cliniques du futur ne seront pas limités par des échantillons de petite taille, mais pourront potentiellement inclure tout le monde.

Optimisation des performances des machines et des appareils : les machines et les appareils deviennent plus intelligents et plus autonomes grâce à l'analyse des big data. Les outils de big data, par exemple, sont utilisés pour faire fonctionner la voiture à conduite autonome de Google. La Toyota Prius est équipée de caméras, d'un GPS, d'ordinateurs puissants et de capteurs qui lui permettent de rouler en toute sécurité sur la route sans nécessiter d'interaction humaine. Les données des compteurs intelligents sont utilisées pour optimiser les systèmes énergétiques grâce aux techniques de big data. Les outils de big data peuvent même être utilisés pour améliorer les performances des ordinateurs et des entrepôts de données.

Trading financier : le big data est actuellement largement utilisé dans le trading à haute fréquence (HFT). Dans ce cas, les décisions de trading sont prises à l'aide d'analyses de big data. La majorité des opérations sur actions se déroulent aujourd'hui par le biais d'algorithmes de données, qui utilisent de plus en plus les signaux des réseaux de médias sociaux et des sites d'information pour prendre des décisions d'achat et de vente en une fraction de seconde.

Améliorer la sécurité et l'application de la loi : le big data est largement utilisé pour améliorer la sécurité et faciliter l'application de la loi. Selon les allégations, l'Agence nationale de sécurité

américaine (NSA) utilise l'analyse des big data pour déjouer les plans des terroristes (et peut-être nous espionner). D'autres utilisent l'analyse des big data pour détecter et éviter les cyber-attaques. Les sociétés de cartes de crédit utilisent le big data pour détecter les transactions frauduleuses, tandis que les services de police utilisent les techniques du big data pour appréhender les criminels et même prédire leurs activités.

1.10.2 Les limites :

Les big data présentent une variété de problèmes. Plusieurs spécialistes travaillant dans le domaine du big data ont reconnu un certain nombre de problèmes qui se posent lors de l'analyse de grandes quantités de données. De nombreuses entreprises se retrouvent piégées dans leurs projets Big Data dès les premières étapes, car elles ne sont pas conscientes des problèmes et ne disposent pas des ressources nécessaires pour les résoudre.

Voici quelques-unes des limites du Big Data[37] :

Manque de compréhension adéquate : Parce qu'elles ne disposent pas d'une compréhension suffisante, les entreprises sont incapables de réussir leurs initiatives en matière de Big Data. Les employés n'ont aucune idée de ce que sont les données, de la manière de les stocker, de les distribuer ou de leur pertinence, et ils n'ont aucune idée de leur provenance. Même si les spécialistes des données sont conscients de ce qui se passe, d'autres personnes ne le sont pas forcément. Par exemple, si les recrues ne sont pas conscientes de la nécessité de stocker des données, elles peuvent omettre de sauvegarder des informations cruciales. Ils peuvent ne pas faire un usage approprié de la base de données. Par conséquent, lorsque ces informations critiques sont requises, elles ne peuvent pas être récupérées rapidement.[37]

Confusion lors d'une sélection d'outil de Big Data : La plupart du temps, les entreprises sont perplexes lorsqu'il s'agit de sélectionner le bon outil d'analyse et de stockage des Big Data. Parce qu'il y a tant d'options, les entreprises deviennent excessivement confuses, font de mauvaises sélections et choisissent une technologie qui n'est pas appropriée. En conséquence, du temps, des heures de travail et de l'argent sont gaspillés.[37]

Problèmes liés à la croissance des données : Un autre problème important du big data est de stocker correctement ces volumes massifs de données. Il a été remarqué que la quantité de données conservées dans les bases de données et les centres de données augmente rapidement. Plus ces ensembles de données augmentent en taille, plus il devient difficile d'en assurer le suivi. La

majorité des données sont non structurées et proviennent de sources diverses, notamment de vidéos, de fichiers, d'audios, de documents et d'autres médias. En clair, cela signifie qu'elles ne sont pas consultables dans des bases de données. [46]

Sécurisation des données : L'un des aspects les plus terrifiants du Big Data est la sécurisation des grandes collections de données. On rapporte souvent que les entreprises sont tellement préoccupées par la compréhension, le stockage et l'étude de leurs ensembles de données que la sécurité des données est repoussée à des étapes ultérieures. Cependant, ce n'est pas une décision judicieuse, car les installations de stockage de données non protégées peuvent servir de terreau à des pirates informatiques malveillants.[47]

Manque de professionnels des données : Les entreprises ont besoin de spécialistes des données formés pour exploiter les outils et la dernière technologie Big Data. Il s'agira notamment d'analystes de données, d'ingénieurs de données et de scientifiques des données ayant une connaissance approfondie des outils de big data et la capacité de donner un sens à des ensembles de données massifs [48]. Les organisations sont confrontées à une pénurie de spécialistes du Big Data en raison de l'évolution rapide des techniques de traitement des données, mais ce n'est pas le cas des professionnels des données. Il est essentiel de prendre des mesures concrètes pour combler cet écart.

Coûteux : L'adoption de projets Big Data implique un investissement financier important. Si l'entreprise opte pour une solution sur site, elle doit prendre en compte le coût du matériel, de l'électricité et de l'embauche de nouveaux employés. Malgré le fait que les cadres essentiels soient open source, l'entreprise sera responsable de l'installation, de l'expansion, de la configuration et de la maintenance du nouveau logiciel. [47]

Problèmes de mise à l'échelle : La conception de la solution peut être prise en compte, et des ajustements de mise à l'échelle peuvent être effectués sans exiger davantage de travail. Cependant, le véritable problème ne réside pas dans le lancement de capacités de thésaurisation supplémentaires, mais dans la complexité de la mise à l'échelle afin que le système de performance ne se détériore pas et reste dans les limites du budget. [47]

1.11 CONCLUSION :

Dans ce chapitre, nous avons passé en revue les différentes technologies permettant de classifier les données et les techniques permettant de les analysées. Nous avons également cité les caractéristiques du big data (volume, variété, vélocité, valeur, véracité), ses domaines d'applications et les plateformes les plus connus ainsi que ses avantages et limites.

CHAPITRE 2

2.1 Introduction :

L'authentification et l'autorisation sont deux mécanismes de sécurité de l'information essentiels que les administrateurs utilisent pour protéger les systèmes et les informations. Bien que ces deux termes se ressemblent, ils jouent des rôles distincts mais tout aussi essentiels dans la sécurisation des applications et des données.

Le processus par lequel un système informatique vérifie l'identité d'une entité (personne, ordinateur, etc.) est appelé authentification. Il va comparer les informations des utilisateurs autorisés stockées dans une base de données (sur un serveur d'authentification local ou distant) aux informations fournies. Par défaut Hadoop ne prend pas en charge l'authentification des utilisateurs ou des services Hadoop. Un utilisateur s'authentifie uniquement avec les systèmes d'exploitation pendant le processus d'ouverture de session. Par conséquent, Hadoop est soumis à plusieurs risques de sécurité c'est pour ça qu'il est recommandé à activer l'authentification pour le cluster afin de protéger les données contre les menaces internes et externes au réseau.[66]

L'autorisation est le processus de définition des droits/privilèges d'accès aux ressources, qui est lié à la sécurité de l'information en général et à la sécurité informatique en particulier, ainsi qu'au contrôle d'accès. L'autorisation est une fonction de la phase de définition de la politique, qui précède la phase d'application de la politique, dans laquelle les demandes d'accès sont acceptées ou refusées en fonction des autorisations qui ont été définies précédemment [83]. L'autorisation est généralement associée à l'authentification afin que le serveur ait une idée de l'identité du client qui demande l'accès.[84]

L'autorisation de niveau de service définit les permissions des utilisateurs pour les différents objets du cluster. Ces autorisations permettent de contrôler les différentes actions qu'un utilisateur peut effectuer, par exemple, soumettre un travail MapReduce, accéder à un fichier sur HDFS, etc..[85]

2.2 L’Autorisation :

2.2.1 Définition :

L'autorisation est un mécanisme de sécurité permettant de déterminer les niveaux d'accès ou les privilèges des utilisateurs/clients liés aux ressources du système, notamment les fichiers, les services, les programmes informatiques, les données et les fonctionnalités des applications. Il s'agit du processus d'octroi ou de refus d'accès à une ressource du réseau qui permet à l'utilisateur d'accéder à diverses ressources en fonction de son identité. [49]

2.2.2 Permissions utilisées dans l'autorisation :

L'autorisation est basée sur les "permissions", qui définissent ce qu'un utilisateur authentifié peut et ne peut pas faire dans un système informatique. [50]

- **Permissions basées sur les rôles**: accorde des autorisations en fonction d'un groupe d'utilisateurs ayant un rôle professionnel partagé. Les autorisations basées sur les rôles spécifient les ressources auxquelles ce groupe est autorisé à accéder. Ce modèle d'autorisations prend en charge le principe d'accès au moindre privilège, selon lequel un système doit accorder à chaque utilisateur les ressources minimales dont il a besoin pour remplir son rôle professionnel. [50]
- **Autorisations du dispositif**: accorde des autorisations en fonction du périphérique qui accède à la ressource. Ce modèle d'autorisation peut accorder des autorisations différentes pour des dispositifs fiables, tels qu'un ordinateur portable d'entreprise, ou des dispositifs non fiables, tels qu'un dispositif mobile personnel. Les systèmes d'autorisation doivent ajuster les autorisations des dispositifs en fonction d'une évaluation de la sécurité de chaque dispositif. [50]
- **Permissions d'emplacement** : accorde des autorisations en fonction de l'emplacement de l'utilisateur ou de l'entité. Les systèmes d'autorisation utilisent ce type de permission pour limiter l'accès aux ressources sensibles pour les utilisateurs se connectant depuis leur domicile ou pour d'autres entités se connectant à distance. [50]

2.2.3 Méthodes d'autorisation :

- **Contrôle d'accès discrétionnaire (DAC)** : DAC détermine les privilèges en fonction de l'utilisateur spécifique et de ses groupes d'accès. Un modèle DAC permet à chaque objet d'un système d'être accessible par un groupe ou une identité particulière. Les personnes chargées d'accorder les autorisations peuvent fournir une autorisation d'administration à d'autres utilisateurs. [50]

Figure 2. 1 Contrôle d'accès discrétionnaire(DAC) [51]

- **Contrôle d'accès obligatoire (MAC)** : MAC détermine l'autorisation des entités au niveau du système d'exploitation. Le MAC régit généralement les autorisations pour les threads et les processus, en définissant les fichiers et les objets mémoire auxquels ils peuvent accéder. [50]

Figure 2. 2 Contrôle d'accès obligatoire (MAC) [51]

- **Contrôle d'accès basé sur les rôles (RBAC)** : RBAC est utilisé pour appliquer les contrôles d'accès définis dans le modèle DAC ou MAC. Le RBAC s'appuie sur des rôles et des privilèges prédéfinis, affecte les utilisateurs aux rôles et configure un système de sorte que seuls des rôles spécifiques puissent accéder à chaque objet. [50]

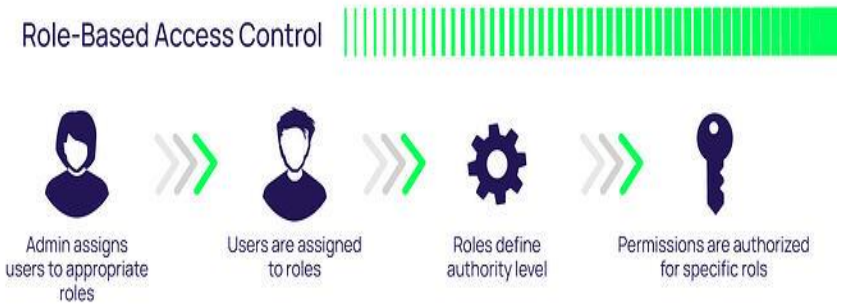


Figure 2. 3 Contrôle d'accès basé sur les rôles (RBAC) [51]

- **Contrôle d'accès par attributs (ABAC)** : ABAC est utilisé pour mettre en œuvre des contrôles d'accès en fonction d'une politique. Elle utilise des attributs, qui peuvent être attachés à un utilisateur, une ressource, un objet ou un environnement entier. Une entité est autorisée si le système d'authentification constate que tous les attributs définis dans la politique sont vrais. [50]

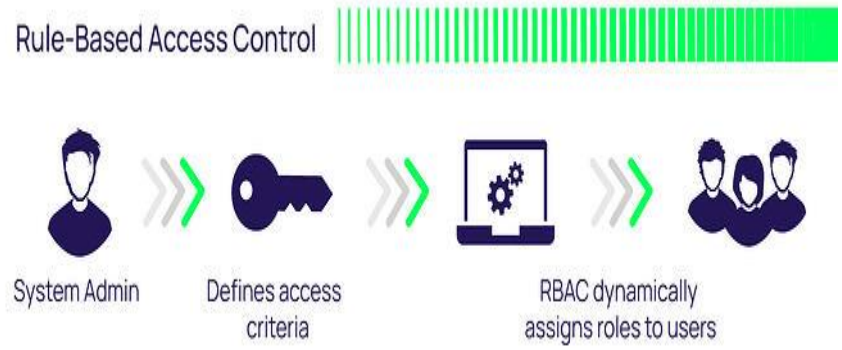


Figure 2. 4 Contrôle d'accès par attributs (ABAC) [51]

2.2.4 Objectifs de l'Autorisation:

- Une fois qu'un sujet est authentifié, l'autorisation est le processus qui consiste à déterminer si l'identité donnée (par exemple, un utilisateur) est autorisée à accéder à la ressource demandée et, si oui, quelles actions il est autorisé à entreprendre. L'objectif est de donner aux utilisateurs authentifiés l'accès aux ressources (telles que les réseaux, les applications ou

les données) dont ils ont besoin pour faire leur travail et rien de plus (également connu sous le nom de principe du moindre privilège) et de refuser tout autre accès.[80]

- L'autorisation est une petite partie de l'équation du contrôle d'accès, les organisations mettent en place les mesures d'authentification pour gérer efficacement l'accès aux données sensibles. L'autorisation consiste à donner à un utilisateur authentifié la permission d'effectuer une action donnée sur des ressources spécifiques, elles sont toutes deux nécessaires pour traiter des données sensibles. Sans l'une ou l'autre, les données restent vulnérables aux violations de données et aux accès non autorisés. [81]

Dans les environnements sécurisés, l'autorisation doit toujours suivre l'authentification. Les utilisateurs doivent d'abord prouver que leur identité est authentique avant que les administrateurs d'une organisation ne leur accordent l'accès aux ressources demandées.[82]

2.3 L'Authentification :

2.3.1 Définition :

L'authentification est le processus de reconnaissance de l'identité d'un dispositif ou d'un utilisateur. Il s'agit du mécanisme qui associe une demande entrante à un ensemble d'informations d'identification. Les informations d'identification fournies sont comparées à celles d'un fichier dans une base de données contenant les informations de l'utilisateur autorisé sur un système d'exploitation local ou dans un serveur d'authentification. [52]

2.3.2 Facteurs d'Authentification :

Un facteur d'authentification est une catégorie spéciale de justificatif de sécurité qui est utilisée pour vérifier l'identité et l'autorisation d'un utilisateur qui tente d'accéder, d'envoyer des communications ou de demander des données à partir d'un réseau, d'un système ou d'une application sécurisée. [53]
On dénombre cinq catégories [53]:

2.3.2.1 Facteurs de connaissance :

Les facteurs de connaissance exigent que l'utilisateur fournisse certaines données ou informations avant de pouvoir accéder à un système sécurisé. Un mot de passe ou un numéro d'identification personnel (PIN) est le type le plus courant de facteur d'authentification basé sur la connaissance utilisée pour restreindre l'accès à un système. La plupart des connexions génériques à des

applications ou à des réseaux nécessitent un nom d'utilisateur ou une adresse électronique et un mot de passe ou un code PIN correspondant pour y accéder. Le nom d'utilisateur ou l'adresse électronique en soi n'est pas considéré comme un facteur d'authentification - c'est la façon dont l'utilisateur revendique son identité auprès du système. Un mot de passe ou un code PIN est utilisé pour authentifier que le nom d'utilisateur ou l'adresse électronique est fourni par la bonne personne.[53]

2.3.2.2 Facteurs de possession

Les facteurs de possession exigent que l'utilisateur possède un élément d'information ou un dispositif spécifique avant de pouvoir accéder au système. Les facteurs de possession sont généralement contrôlés par un dispositif dont on sait qu'il appartient au bon utilisateur, comme la carte de paiement.[53]

2.3.2.3 Facteurs d'inhérence

Les facteurs d'inhérence permettent d'authentifier les justificatifs d'accès sur la base de facteurs qui sont uniques à l'utilisateur. Il s'agit notamment des empreintes digitales, des empreintes de pouce et des empreintes de paume ou de main. La reconnaissance vocale et faciale et les scans de la rétine ou de l'iris sont également des types de facteurs d'authentification inhérents.[53]

2.3.2.4 Facteurs de localisation

Les administrateurs réseau peuvent mettre en œuvre des services qui utilisent des contrôles de sécurité de géolocalisation pour vérifier l'emplacement d'un utilisateur avant de lui accorder l'accès à une application, un réseau ou un système.[53]

2.3.2.5 Facteurs comportementaux

Un facteur d'authentification basé sur le comportement est basé sur les actions entreprises par l'utilisateur pour accéder au système. Les systèmes qui prennent en charge les facteurs d'authentification basés sur le comportement peuvent permettre aux utilisateurs de préconfigurer un mot de passe en exécutant des comportements dans une interface définie et en les répétant ultérieurement comme méthode de vérification de l'identité.[53]

2.3.3 Type d'authentification :

Avec les progrès de l'Internet, diverses méthodes d'authentification de réseau ont vu le jour. Les mots de passe, l'authentification à deux facteurs, les jetons, la biométrie, la reconnaissance des ordinateurs, les CAPTCHA et l'authentification unique SSO **(**Single Sign On) sont autant

d'exemples d'approches d'authentification générales. Nous allons maintenant examiner les méthodes d'authentification les plus populaires [54] :

2.3.3.1 Authentification par mot de passe:

Ce type d'authentification exige que le fournisseur se rappelle ce qu'il sait. Il y a deux parties dans cette méthode. Premièrement, le fournisseur entre le nom d'utilisateur et, deuxièmement, le mot de passe. Le mot de passe est la combinaison secrète de mots et de chiffres que le fournisseur connaît. [55]

2.3.3.2 Authentification basée sur la cryptographie :

Appeler aussi l'authentification par clé est le processus qui consiste à utiliser des clés cryptographiques dans une poignée de main de type défi-réponse pour prouver l'identité d'une personne. Elle entre dans la catégorie "quelque chose que vous avez". [56]

2.3.3.3 Fonctions de hachage cryptographiques :

Une fonction de hachage est un algorithme qui transforme des données de taille arbitraire en une sortie de taille fixe. La sortie est un texte chiffré appelé valeur de hachage ou résumé. L'objectif principal d'une fonction de hachage cryptographique est de vérifier l'intégrité des données.

Les fonctions de hachage sont souvent utilisées dans les mots de passe. Les mots de passe de toute base de données sécurisée sont stockés sous la forme de valeurs de hachage ou de résumés. Il n'est pas sûr de stocker des mots de passe sous forme de texte brut dans une base de données. Chaque fois que vous vous connectez, votre mot de passe est haché dans un résumé et comparé à celui stocké dans une base de données.[57]

Les fonctions de hachage sont aussi utilisées pour générer les signatures numériques qui peuvent être utilisé pour assurer la non-répudiation.

2.3.3.4 Authentification par clé symétrique :

Dans l'authentification par clé symétrique, l'utilisateur partage une seule, clé secrète avec un serveur d'authentification (normalement la clé est intégrée à un jeton) [59]. L'utilisateur peut être s'authentifié en envoyant au serveur d'authentification son nom d'utilisateur ainsi qu'un message de défi aléatoire qui est chiffré par la clé secrète. Ainsi, l'utilisateur est considéré comme un utilisateur authentifié si le serveur peut faire correspondre le message crypté reçu en utilisant sa clé secrète commune. [60]

2.3.3.5 Authentification par clé asymétrique :

L'authentification basée sur la cryptographie asymétrique repose sur deux clés : une clé privée et une clé publique. La clé privée n'est connue que par le dispositif à authentifier, tandis que la clé publique peut être divulguée à toute entité désireuse d'authentifier le dispositif. L'hôte envoie un défi au dispositif. Ce dernier calcule une signature sur la base du défi et de la clé privée et la renvoie à l'hôte. Mais ici, l'hôte utilisera la clé publique pour vérifier la signature. Il est également essentiel que la fonction utilisée pour calculer la signature possède certaines propriétés mathématiques.[61]
L'échange de clés publiques est vulnérable aux attaques de l'homme du milieu, pour éviter cette attaque, on utilise un certificat signé pour échanger une clé publique, ce certificat signé contient des informations d'identification comme un email, la clé publique associée et la signature du certificat.[62]

Un certificat est un document électronique émis par une tierce partie de confiance qui permet de garantir l'authenticité d'une clé publique.
Les fonctions les plus couramment utilisées pour les schémas asymétriques sont RSA et ECDSA. Ici aussi, le dispositif prouve qu'il a connaissance d'un secret, la clé privée, sans le divulguer.[61]

2.3.3.6 Authentification Biométrique :

L'authentification biométrique est une méthode d'identification et/ou de vérification de l'identité d'un utilisateur basée sur la mesure de ses caractéristiques physiologiques ou comportementales uniques. L'empreinte numérique, la reconnaissance faciale et la géométrie de la main sont autant d'exemples de biométrie physiologique. La reconnaissance verbale, l'action, et le balayage distinctif constituent la biométrie comportementale.[64]

2.3.3.7 Authentification unique (SSO) :

L'authentification unique ou SSO (Single Sign On) est une méthode permettant d'accéder à plusieurs systèmes logiciels indépendants de manière à ce que lorsqu'un utilisateur se connecte à un système, sans être amené à se reconnecter dans chaque application, obtient l'accès à tout le système. Ce processus aide les utilisateurs à accéder à de nombreux services et réduit la menace pour les administrateurs de diriger les utilisateurs de manière pratique. En empêchant l'utilisateur de se souvenir de nombreux mots de passe, il contribue à améliorer l'efficacité de l'utilisateur et à réduire le temps nécessaire à l'utilisateur pour saisir plusieurs mots de passe. [65]

2.4 Travaux relatifs :

Depuis quelques années, l'usage de l'informatique a permis une production en très grande quantité de données à travers plusieurs secteurs différents. Il existe quelques technologies permettant de traiter et manipuler ces énormes volumes de données. Hadoop est l'une de ces

technologies qui stocke les données qui peuvent inclure explicitement des informations sensibles (financières, personnelles et commerciales) dans un cluster.[66]

Les clusters Hadoop (HCs) permettent de stocker et d'analyser d'énormes quantités de données dans un environnement de traitement parallèle et distribué en supportant du matériel de base. Les HCs sont :

- Hautement évolutifs, en permettant de booster les applications d'analyse de données.
- Flexibles pour ajouter de la puissance de traitement lorsque la génération de données augmente en ajoutant un nœud supplémentaire dans le cluster.

-Très résistants aux défaillances des données.[67]

Cependant, bien qu'il s'agisse d'un excellent outil de traitement des données volumineuses, L'absence de mécanisme de vérification de la sécurité dans Hadoop peut permettre aux utilisateurs illégaux d'envahir facilement le cluster. Les utilisateurs illégaux peuvent accéder de manière malveillante aux composants d'un cluster et soumettre des travaux malveillants, déguisés en d'autres utilisateurs, pour altérer les permissions et intercepter ou altérer les données sur HDFS. [68]

Pour combler le manque de sécurité dans les clusters Hadoop, plusieurs protocoles d'autorisation ont été proposés. Ces protocoles peuvent être classés comme suit :

1- **Protocoles basés sur l'utilisation du protocole Kerberos :** Kerberos est un protocole d'authentification qui a été construit pour un système qui fournit des services de sécurité à l'échelle du réseau. Il s'agit d'un système d'authentification distribuée qui permet à un client de prouver son identité à un serveur sans envoyer de données à travers le réseau qui pourraient permettre à un attaquant de se faire passer ensuite pour ce client. Kerberos peut résoudre un grand nombre des problèmes de sécurité de réseaux hétérogènes de grande taille, notamment l'authentification mutuelle entre clients et serveurs. L'idée de base de Kerberos est la suivante est qu'une tierce partie de confiance (le serveur de sécurité Kerberos) fournit un moyen par lequel les constituants du réseau peuvent se faire confiance entre eux[69],il repose sur trois serveurs pour assurer l'authentification : [69]

- Un serveur d'authentification (AS : Authentication Server) qui prend en charge toute la partie authentification pure du client. C'est lui seul qui peut permettre au client de communiquer au TGS (grâce à un ticket d'accès). [70]
- Le serveur de distribution de tickets –TGS : (TicketGranting Server) prend en charge les demandes d'accès aux services des clients déjà authentifiés. L'ensemble des infrastructures serveur de Kerberos AS et TGS est appelé le centre de distribution de clés (KDC : KeyDistribution Center). Ils sont généralement regroupés sur le même serveur.[70]
- Serveur d'application : C'est un serveur qui exécute un service particulier.[71]

2- Protocoles combinant Kerberos avec un autre outil de sécurité :

L'écosystème Hadoop contient plusieurs outils comme soutien à la sécurité Hadoop, et chacun d'entre eux présente des caractéristiques différentes et une efficacité différente dans un contexte différent.[86]

2.4.1 Protocoles basés sur l'utilisation du protocole Kerberos :

Dans cet article [72], ils ont proposé une méthode pour incorporer le protocole Kerberos entre le client et le cluster HDFS est proposée pour sécuriser le système. Kerberos est un protocole d'authentification de réseau qui assure une communication sécurisée entre le client et le serveur sur un réseau non sécurisé. De plus, un agent a été incorporé qui est autorisé à accéder au tampon du client, à en extraire des données et à les charger dans le cluster HDFS. Afin de fournir l'authenticité au client, le système a ajouté un protocole d'authentification qui est KERBEROS entre le client et le cluster HDFS, de sorte que lorsque le client demande au Name Node du cluster HDFS de stocker ses données, le Name Node redirige le client vers Kerberos qui authentifie le client et autorise l'agent. L'agent, après avoir obtenu l'autorisation, prend les données du client dans le tampon du client et les place dans le nœud de données. De cette façon, l'ensemble du système de stockage des données volumineuses dans le cluster HDFS devient plus sûr et le client n'a plus besoin de charger lui-même ses données dans le nœud Data.

Dans l'article [73], ils ont étudié le mécanisme d'authentification du protocole Kerberos sous HDFS, et souligne les problèmes que le mécanisme d'authentification d'identité du protocole Kerberos dans un environnement de cluster HDFS : synchronisation temporelle, sécurité du KDC, les attaques par dictionnaire et le mécanisme de déni. En vue de résoudre ces problèmes de sécurité, ce document présente tout d'abord un aperçu du processus d'authentification de l'actuel système Kerberos dans un environnement de cluster HDFS, deuxièmement, il modifie le protocole Kerberos en utilisant le chiffrement à clé publique et le mécanisme de signature des données. Enfin, il fournit le processus d'authentification du protocole Kerberos amélioré dans l'environnement HDFS et il vérifie la faisabilité du protocole amélioré par une analyse spécifique.

Dans l'article [74], les auteurs ont proposé une nouvelle direction pour Kerberos est la cryptographie à clé publique. La cryptographie à clé publique facilite grandement la distribution des clés. En utilisant uniquement la cryptographie symétrique, le KDC et le client doivent partager une

clé ; en utilisant la cryptographie asymétrique, le client peut présenter la clé publique, qui peut être utilisée pour chiffrer les messages qui lui sont destinés. Le grand avantage de cette version de Kerberos est que le KDC ne doit plus enregistrer les clés des clients dans sa base. Pour obtenir un ticket d'octroi, le client doit présenter sa clé publique. Le KDC utilise cette clé pour chiffrer le ticket et la clé de session. Comme tout le monde est capable de créer une paire de clés pour la cryptographie à clé publique, une infrastructure supplémentaire est nécessaire. Une autorité de certification (CA) de confiance doit signer chaque clé publique valide. Le client peut présenter sa clé qui est signée par l'autorité de confiance. L'intégration dans Kerberos est facile car seule l'interaction avec le service d'authentification doit être modifiée pour utiliser la cryptographie asymétrique ; tout le reste peut rester tel quel. Si le client présente sa clé publique, le service d'authentification vérifie s'il possède une signature valide d'une autorité de confiance et renvoie ensuite une clé de session. Le client déchiffre la clé de session avec la clé privée de sa paire de clés. La communication suivante est gérée comme dans Kerberos sans support de cryptographie à clé publique.[74]

2.4.2 Protocoles basés sur Kerberos combinés avec un autre outil de sécurité:

Dans cet article [75], les auteurs ont proposé d'utiliser Kerberos et Ranger pour assurer l'authentification et l'autorisation des utilisateurs et des services dans la plateforme distribuée Hadoop (HDP). Ils ont utilisé MIT Kerberos pour l'authentification de manière automatisée, ce qui implique de créer des principaux et de les donner à Ambari avec des privilèges d'administrateur ca veut dire on peut choisir de faire en sorte qu'Ambari se connecte au KDC et crée automatiquement les services nécessaires et les principaux Ambari, générer et distribuer les keytabs ("Configuration automatisée de Kerberos "). Ambari fournit également une option avancée pour configurer manuellement Kerberos, si onchoisit cette option, on doit créer les principaux, générer et distribuer les keytabs. Ambari ne le fera pas automatiquement ("Configuration manuelle de Kerberos ").
Apache Ranger qui est une application web centralisée, comprend des fonctions d'autorisation sur tous les composants d'Apache et le système de fichiers, d'administration de politiques, d'audit et de rapports. Les utilisateurs autorisés peuvent accéder à la console web Apache Ranger pour gérer les politiques de sécurité. Ces politiques de sécurité sont déployées comme des processus légers sur Namenode.

Dans [76], les auteurs ont proposé d'utiliser Apache sentry pour protéger les données sensibles dans le HDFS et assurer l'authentification. Apache sentry a des acteurs pour définir les politiques

d'autorisation. Les acteurs dans Apache Sentry sont : l'utilisateur, le groupe d'utilisateurs, les ressources, les privilèges et le rôle. L'acteur utilisateur sert à authentifier l'utilisateur dont l'identité peut être obtenue à partir du contexte de la session. L'authentification est fournie par quelques techniques comme Kerberos, LDAP Active Directory, LDAP, AD intégré avec Kerberos, établissant un point de vérité unique et une signature unique. L'acteur groupe d'utilisateurs est défini au-delà de la politique sentry qui est obtenue à partir du répertoire des utilisateurs (LDAP, AD, HDFS) et peut également être disponible à partir du contexte de la session. Acteur-Ressources sont utilisés pour protéger les données dans les fichiers, le répertoire sur HDFS, etc. L'acteur privilège est l'action ou l'opération associée à une ressource. Les rôles d'acteur sont une collection de privilèges définis dans la politique de sentry.

Dans le travail cité dans [77], le protocole d'authentification Kerberos est associé au protocole TLS (Transport Layer Security) pour protéger les données contre les attaques et le rejeu. Le système proposé permet aux clients HDFS d'être authentifiés par le nœud de données à l'aide de jetons d'accès aux blocs.
TLS est utilisé pour permettre le cryptage des données transférées. Il assure l'intégrité et la confidentialité des données entre deux entités de communication [77].

Dans [78], les auteurs ont proposé une nouvelle méthode pour sécuriser les données au sein de HDFS. Elle est mise en œuvre en utilisant la Wire Line Security (appelée Open Standard for Authorization). Ils ont utilisé un protocole d'authentification ouvert qui aide à surmonter les problèmes du modèle d'authentification client-serveur conventionnel. Dans le modèle client-serveur conventionnel, le client demande à accéder à une ressource protégée sur le serveur en s'authentifiant à l'aide du passeport du propriétaire de la ressource. Afin de permettre à des applications tierces d'accéder à des ressources restreintes, le propriétaire de la ressource vérifie son autorisation auprès de la tierce partie.

Dans le système proposé, la sécurité filaire est utilisée pour authentifier l'utilisateur et elle renvoie également un jeton unique pour chaque utilisateur qui tente de se connecter avec succès. Le jeton renvoyé par le serveur d'authentification est utilisé dans la méthode de cryptage, ce qui assure la confidentialité et l'intégrité des données de l'utilisateur. Les fichiers sont cryptés avant d'être chargés sur HDFS et décryptés lorsque l'exécution du travail est en cours. L'algorithme de cryptage en temps réel utilise le jeton d'authentification comme clé et crypte les données par XoRing avec cette clé. [78]

Cette méthode utilise Knox qui est un Framework de proxy inverse sans état. Un cluster Hadoop entièrement sécurisé nécessite Kerberos. Kerberos nécessite une bibliothèque côté client et une configuration complexe côté client. En encapsulant Kerberos, Knox élimine le besoin de logiciel client ou de configuration client et simplifie ainsi le modèle d'accès.

2.5 Conclusion :

Dans ce chapitre, nous avons cité quelques méthodes de l'autorisation ainsi que les différents type et facteurs de l'authentification. Ces mécanismes sont les deux considérations les plus importantes dont chaque architecture de sécurité d'un système Hadoop a toujours besoin. Par la suite nous avons décrit quelques protocoles d'authentification et d'autorisation existants.

A l'heure actuelle, le monde de la sécurité offre plusieurs protocoles d'authentification et d'autorisation pour sécuriser les environnements Hadoop. Pour choisir le meilleur outil de sécurité à mettre en place, il est nécessaire d'évaluer les performances de chaque protocole.

CHAPITRE 3

3.1 Introduction

Ce chapitre est consacré à la mise en place et l'utilisation d'un protocole d'authentification et d'autorisation qui est ranger. Nous présenterons en premier lieu l'environnement de travail et l'installation des composants nécessaire à l'exécution. Nous présentons par la suite quelques tests.

3.2 Environnement de travail:

3.2.1 Cloudera

Les clusters Hadoop seront intégrés dans la plateforme de données Cloudera. Cloudera est une société de logiciels américaine cofondée en 2008 par le mathématicien Jeff Hammer Bach, un ancien de Facebook. Les autres cofondateurs sont Christophe Bisciglia, ex-employé de Google, Amr Awadallah, ex-employé de Yahoo, Mike Olson, PDG de Cloudera. La firme Cloudera se consacre au développement de logiciels fondés sur Apache Hadoop, permettant l'exploitation de Big Data, à savoir des bases de données accumulant plusieurs pétaoctets.[87]

3.2.2 Windows Server 2012 R2 Datacenter Workstation:

Windows Server est une collection de systèmes d'exploitation Microsoft qui offrent une administration, un stockage de données, des applications et des communications de niveau entreprise. Windows Server 2012R2 apporte des améliorations à la virtualisation, à l'administration, au stockage, à la mise en réseau, à l'infrastructure des bureaux virtuels, à la protection des accès, à la sécurité des informations, aux services Web et à l'infrastructure de la plate-forme d'applications.[88]

3.2.3 Linux CentOS7:

Community Enterprise Operating System (CentOS) est un système d'exploitation gratuit et open-source. CentOS Linux est une plateforme stable, prévisible, gérée et reproductible, développée à partir du code source de Red Hat Enterprise Linux(RHEL). CentOS a été créé par "GregoryKurtzeren"2004 et a été la première édition de RHEL à intégrer "systemd" comme une fonctionnalité standard**.** CentOS 7 a été lancé en 2014 et comprenait diverses améliorations qui ont aidé la communauté.[89][90]

3.3 Mise en place du protocole:

3.3.1 Installation de Hadoop version 6.1

Hadoop peut fonctionner en trois modes qui sont [91] :

1. Mode local (local mode) : Hadoop fonctionne sur un seul poste de travail, et chacun de ses cinq démons (NameNode, SecondaryNameNode, DataNode, JobTracker et TaskTracker) tourne dans la même JVM (Java Virtual Machine). Par conséquent, la portée des variables diffère considérablement de ce qu'elle est dans les modes pseudo-distribué ou complètement distribué.

2. Mode pseudo-distribué (pseudo-distributed mode) :Hadoop fonctionne sur un seul poste de travail en mode pseudo-distribué, bien que chacun des cinq démons s'exécute dans sa propre JVM.Les développeurs Hadoop utilisent fréquemment le mode pseudo-distribué car il leur permet de construire et de tester des applications dans un environnement qui ressemble beaucoup à un véritable cluster Hadoop.

3. Mode totalement distribué (fully-distributed mode) : Le mode entièrement distribué simule le comportement d'un véritable cluster Hadoop, avec plusieurs postes de travail reliés entre eux par un réseau. Chaque démon Hadoop fonctionne sur sa propre JVM, généralement sur un ordinateur séparé.

Pour réaliser notre travail, nous allons utiliser trois machines qui seront connectées sur le même réseau et un serveur Windows Server 2012 R2 Datacenter pour gérer les trois machines.

Machines	Nom de la machine	Type de nœud	Système d'exploitation
Machine01	Cloudera	Hadoop master	CentOS,7.
Machine02	Node1	Hadoop slave	CentOS,7.
Machine03	Node2	Hadoop slave	CentOS,7.

Tableau 3. 1 Les machines utilisées

Remarque : Avant de mettre en place les deux protocoles à comparer, nous devons configurer notre réseau. Les étapes de la configuration sont bien détaillées dans l'annexe A.

Création des machines du cluster :

La création des machines du cluster est la même que pour le serveur mais avec les différences suivantes:

Série et type de la machine master:

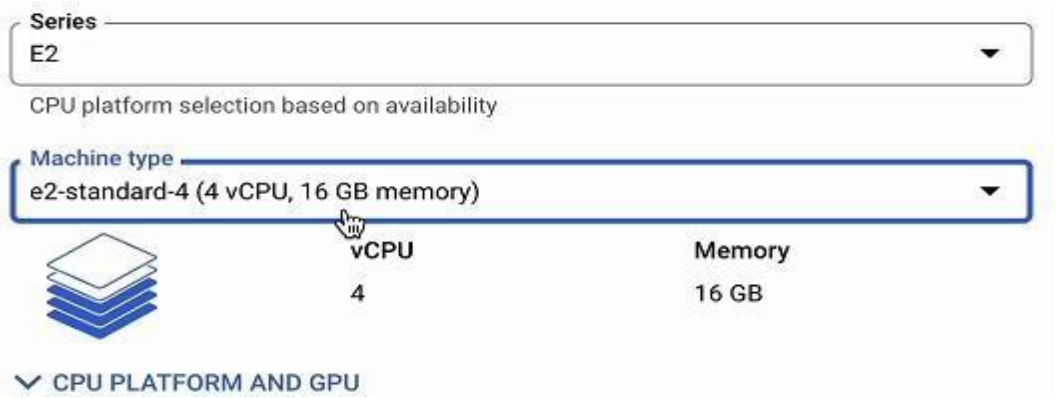


Figure 3. 1 capacité de la machine master

Série et type des machines slaves:

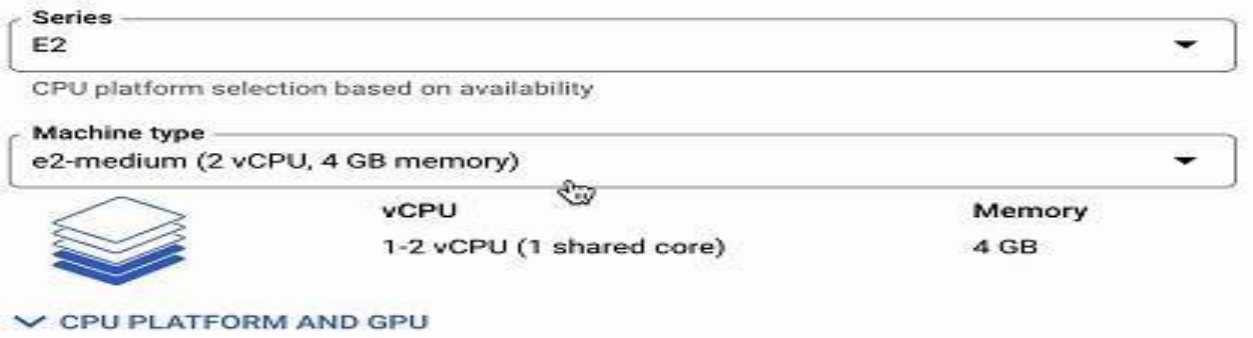


Figure 3. 2 Capacité de la machine esclave

Les machines du cluster obtiendront leur adresse IP (adresse IP interne et adresse IP externe)à partir de notre réseau préconfiguré "vpc-netw" automatiquement, dans la plage10.128.0.0/20.

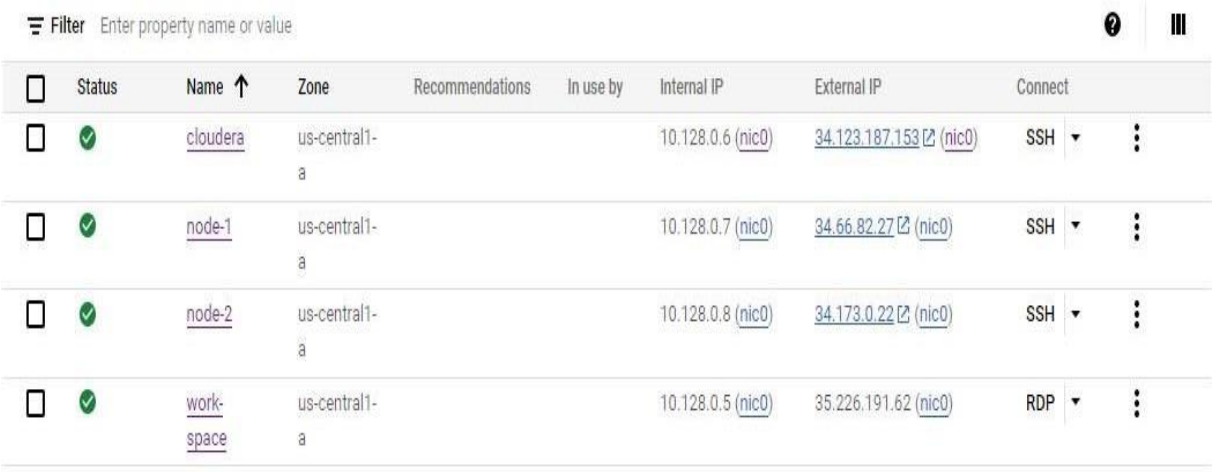


Figure 3. 3 Espace de travail et les machines du cluster

Remarque:

Pour se connecter au serveur(le type de serveur est Windows Server, 2012 R2 Datacenter) on tape l'adresse IP externe du serveur sur le programme de connexion au bureau à distance ainsi que le nom d'utilisateur et le mot de passe qui a été préconfiguré dans Google Cloud Platform**.** La manipulation de cluster via le serveur est bien expliquée dans l'annexe C**.**

3.3.2 Installation de Cloudera :

Nous allons installer Cloudera manager version7.7.4surlenodemastereton va créer notre cluster: À partir de https://archive.cloudera.com/cm7/7.4.4/, nous copions l'adresse du lien de l'installateur de Cloudera manager.

Sur le terminal de la machine master nous téléchargeons le paquet Cloudera manager:

```
[user@cloudera ~]$ sudo wget https://archive.cloudera.com/cm7/7.4.4/cloudera-manager-installer.bin
--2022-09-18 01:42:42--  https://archive.cloudera.com/cm7/7.4.4/cloudera-manager-installer.bin
Resolving archive.cloudera.com (archive.cloudera.com)... 151.101.0.167, 151.101.64.167, 151.101.128.167, ...
Connecting to archive.cloudera.com (archive.cloudera.com)|151.101.0.167|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 757794 (740K) [application/octet-stream]
Saving to: 'cloudera-manager-installer.bin'

100%[=======================================================================================>] 757,794     --.-K/s   in 0.05s

2022-09-18 01:42:42 (15.0 MB/s) - 'cloudera-manager-installer.bin' saved [757794/757794]
```

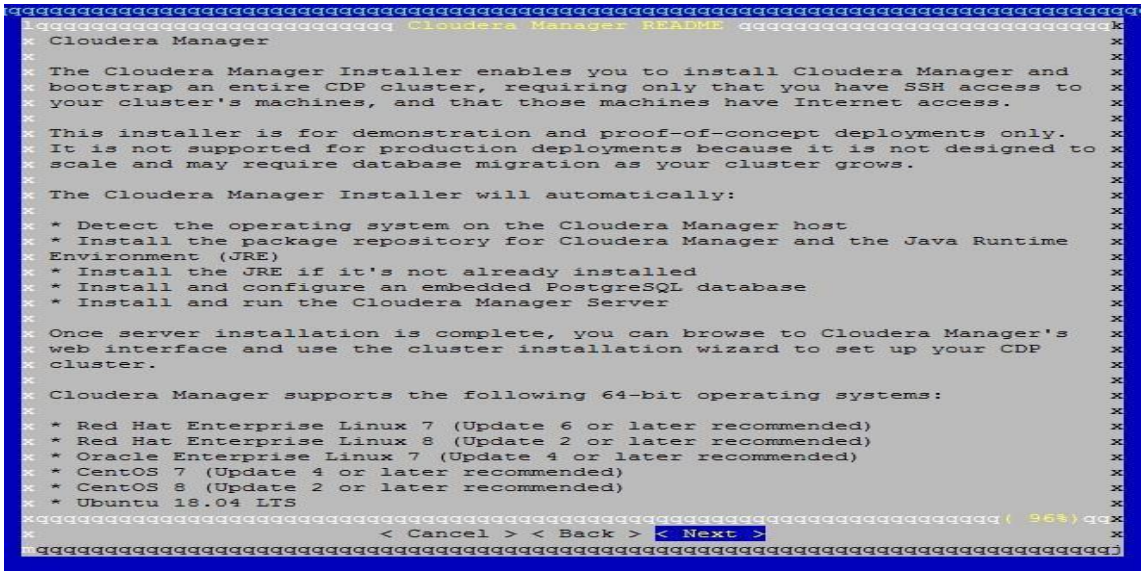
Pour convertir le fichier d'installation en fichier exécutable, nous exécutons la commande suivante :

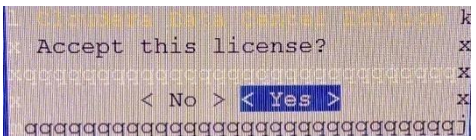
```
[user@cloudera ~]$ sudo  chmod u+x cloudera-manager-installer.bin
```

Ensuite on installe Cloudera manager:

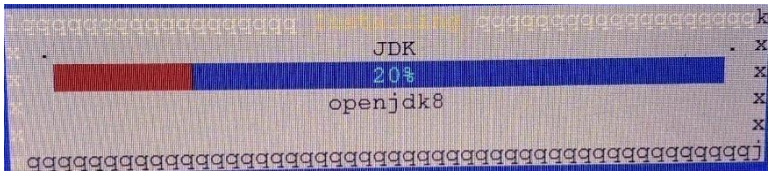
```
[user@cloudera ~]$ sudo ./cloudera-manager-installer.bin
```

En cliquant plusieurs fois sur le bouton suivant:

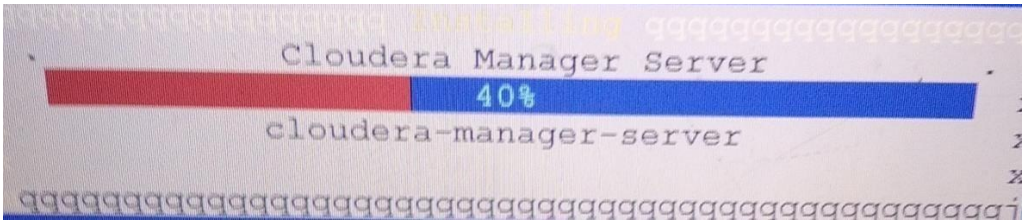


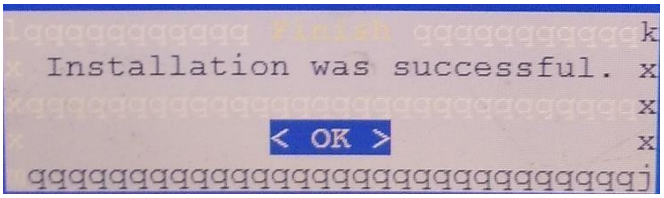


Installation des paquets JDK:

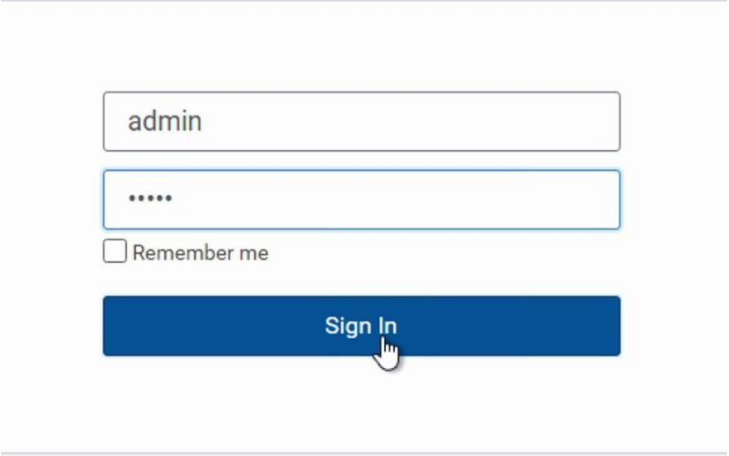


Installation de Cloudera manager server:





Pour accéder à Cloudera manager allez à:http://cloudera:7180/:



Après s'être connecté, nous commençons à créer notre cluster. Nous fournissons un nom et le type de cluster.

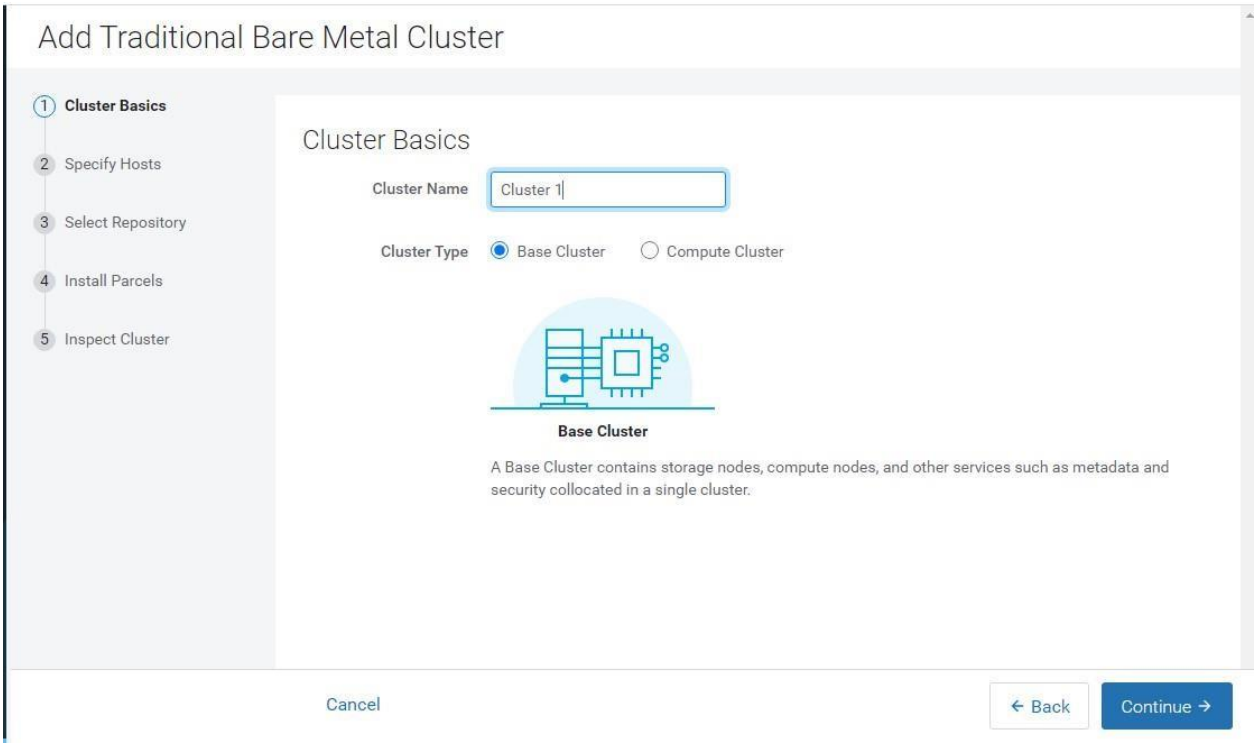


Figure 3. 4 Définition de nom et type cluster

Nous sélectionnons les adresses IP internes(disponibles sur GCP) de nos hôtes:

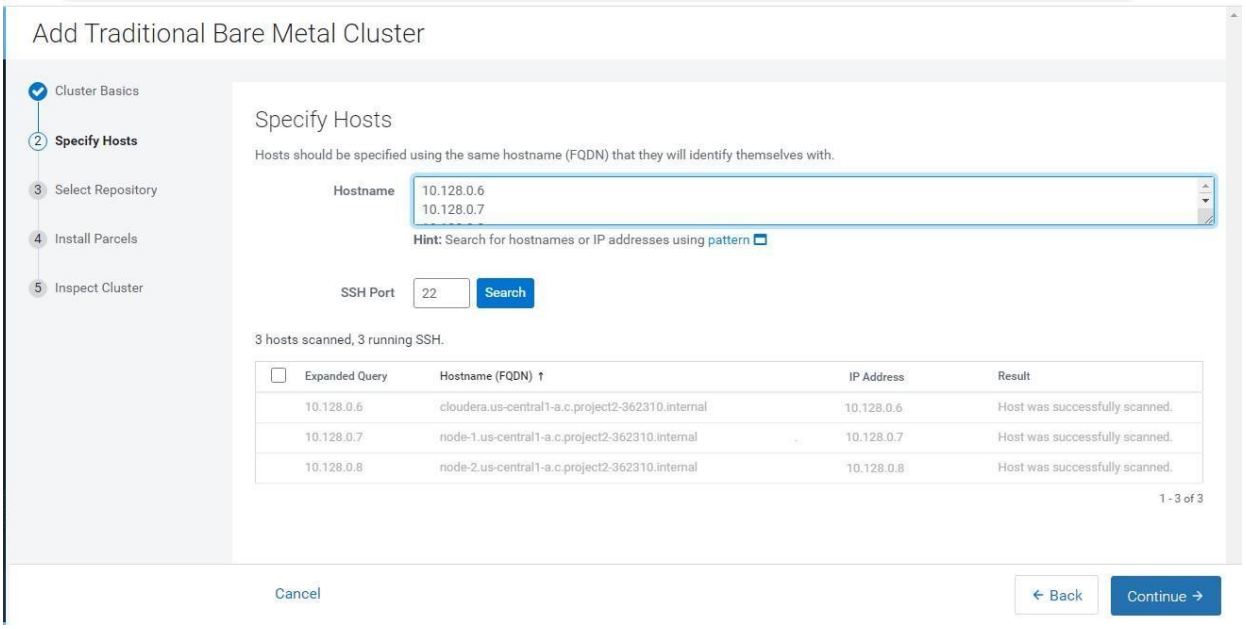


Figure 3. 5 Spécification des hôtes du cluster

Nous Choisissons l'emplacement du dépôt Cloudera à installer sur nos hôtes.

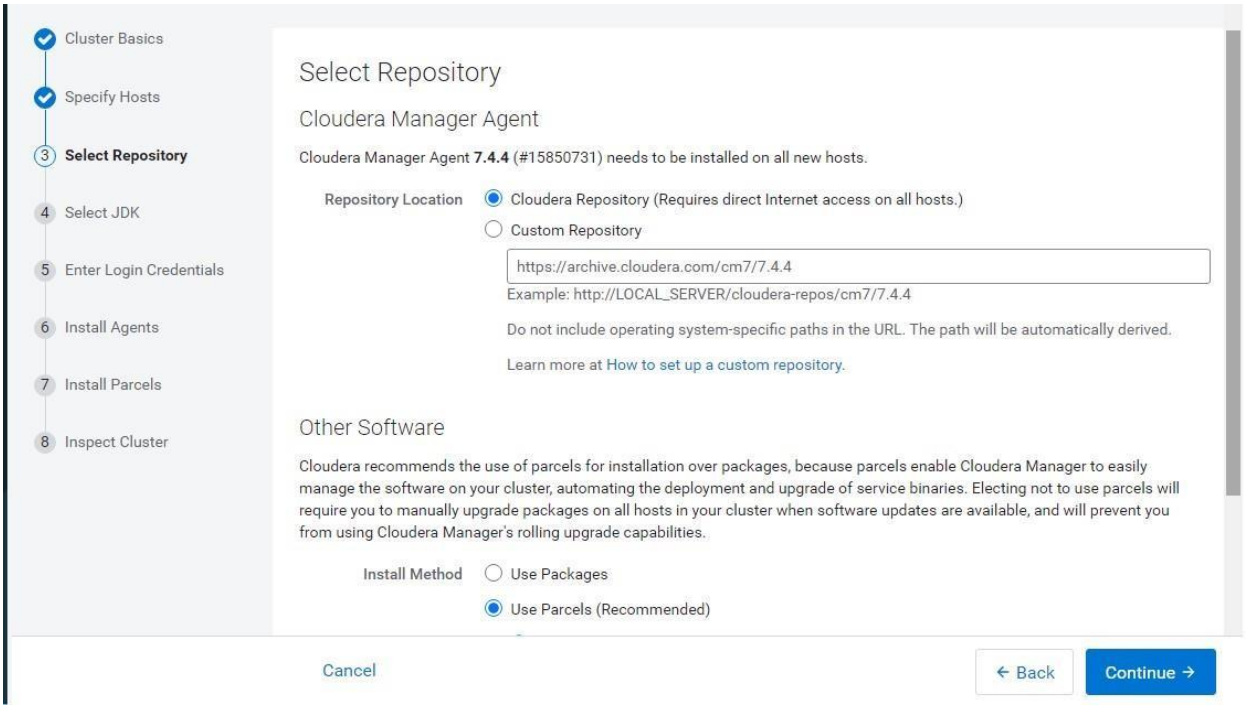


Figure 3. 6 Spécification de la localisation de l'agent répertoire de Cloudera manager

Nous installons la version 8 d'openjdk supportée et fournie par le gestionnaire Cloudera.

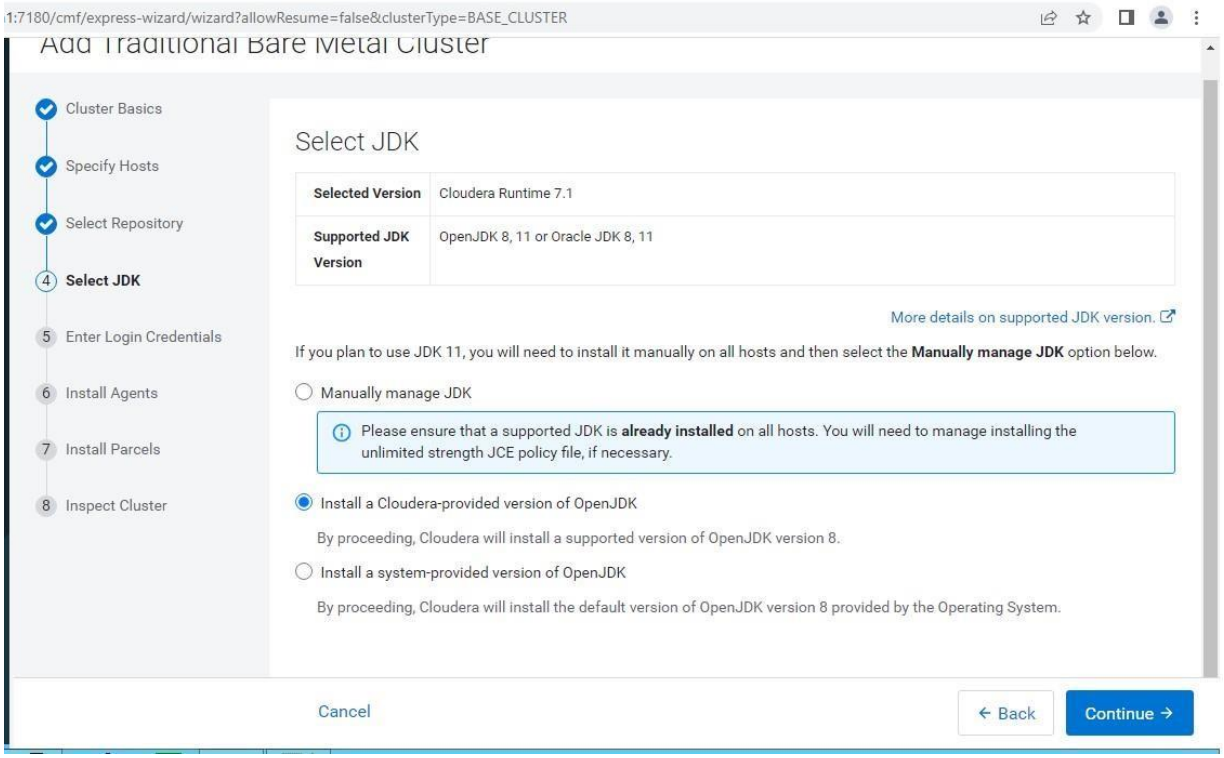


Figure 3. 7 Installation de JDK sur Cloudera manager

Nous accédons aux hôtes(user) en introduisant ses informations d'identification.

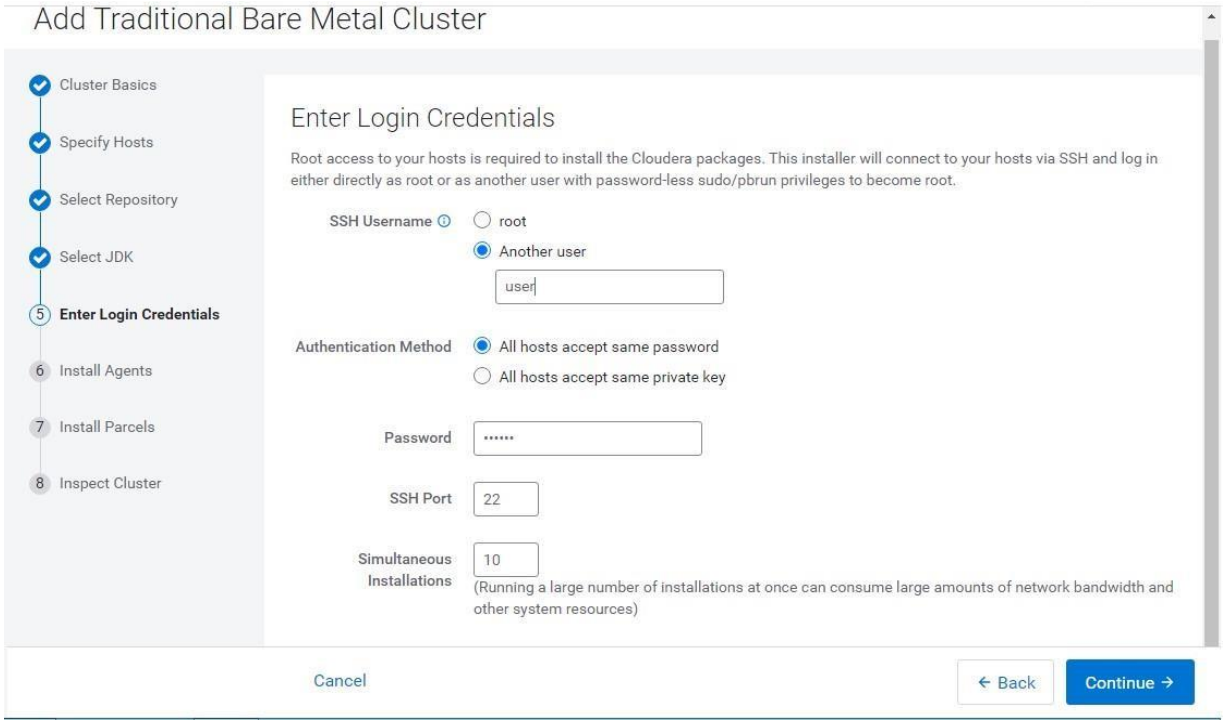


Figure 3. 8 Interface d'authentification

Nous installons l'agent Cloudera Manager qu'est responsable du démarrage et de l'arrêt des processus, du déballage des configurations, du déclenchement des installations et de la surveillance de tous les hôtes d'un cluster.

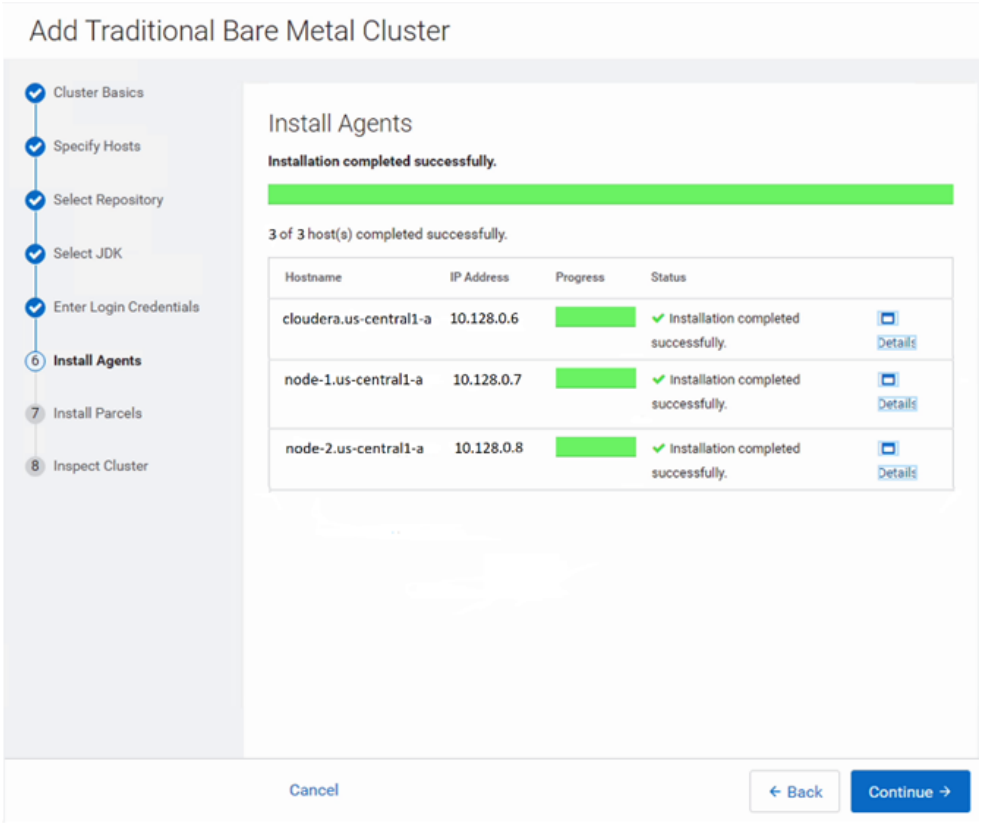


Figure 3. 9 Installation des agents sur les hôtes du cluster

Nous installons les Parcelles. Parcel est un format de distribution binaire qui nous permet d'installer, de mettre à jour ou même de supprimer facilement un ensemble de fichiers d'une manière simple, uniforme, version née, cohérente et distribuée dans un environnement Cloudera.

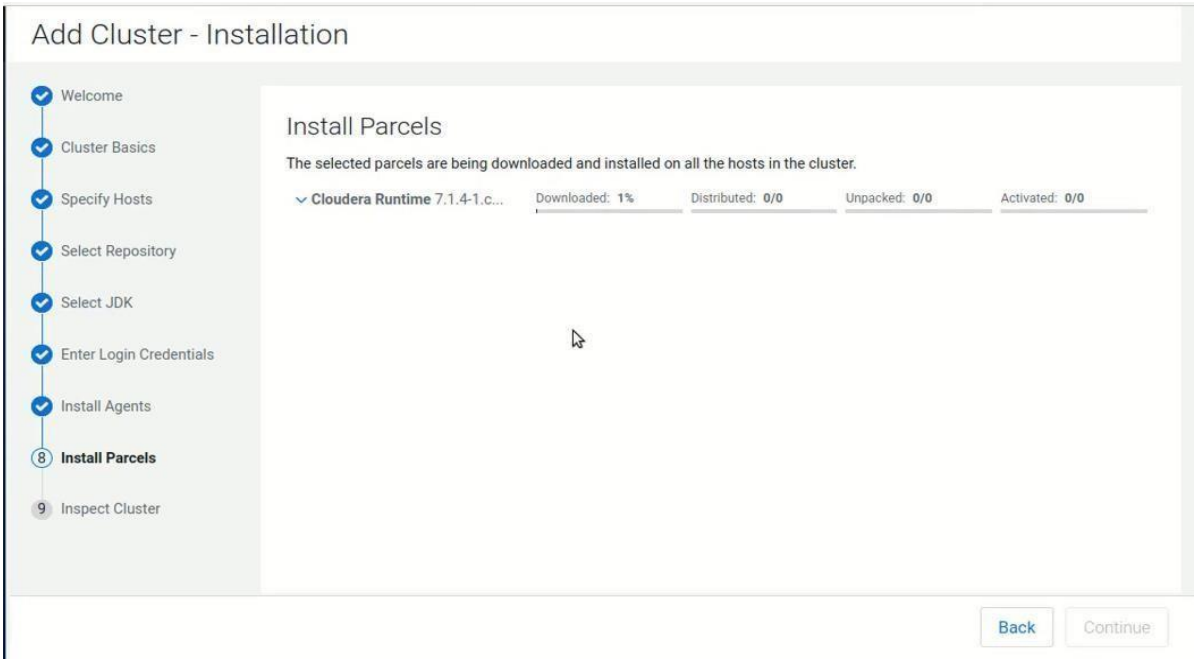


Figure 3. 10 Installation des parcelles

Nous Inspectons le cluster pour détecter les éventuels problèmes.

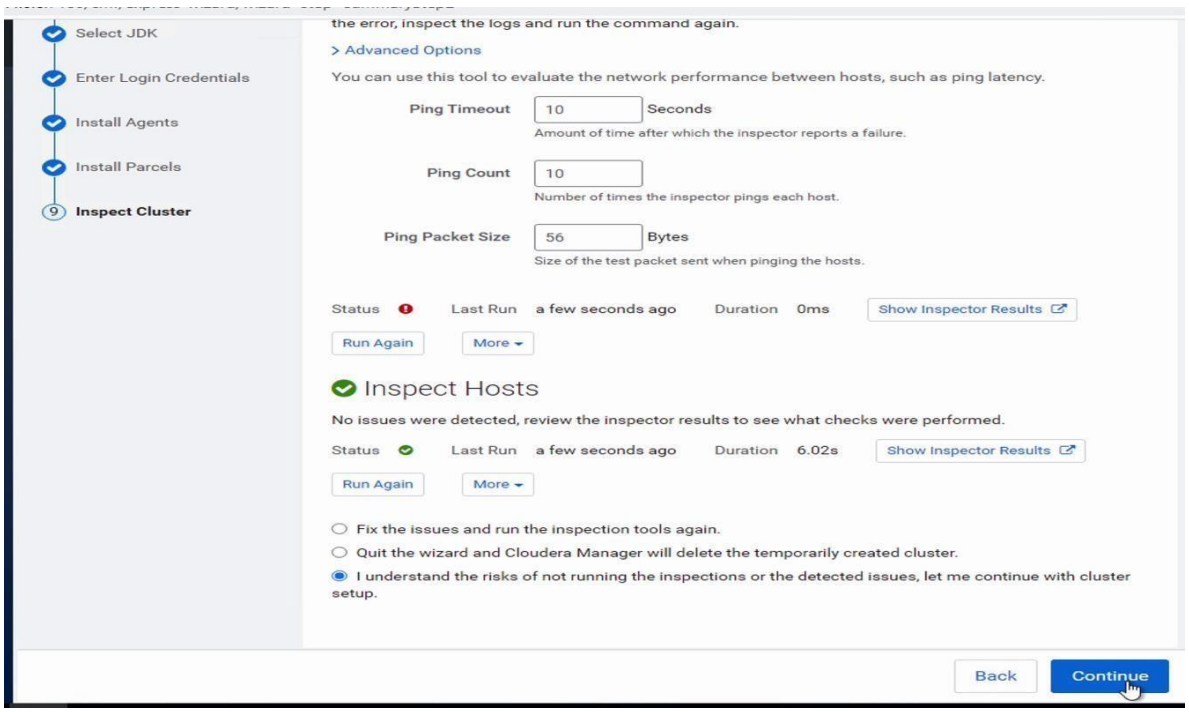


Figure 3. 11 Etat d'inspection du cluster

Configuration du cluster consiste à:

Pour configurer le cluster, il faut suivre les étapes suivantes:

1. Sélectionner les services.
2. Attribuer des règles aux hôtes du cluster.
3. Configurer la base de données
4. Introduire les paramètres requis

Ces étapes sont bien expliquées dans l'annexe D.

3.3.3 Installation de Kerberos version5:

Cloudera Manager fournit un assistant pour intégrer l'instance Kerberos au cluster afin de fournir des services d'authentification.

Kerberos doit déjà être déployé dans notre nœud maître et le centre de distribution de clés Kerberos (KDC)doit être prêt à être utilisé, avec un royaume établi.

3.3.3.1 Client Kerberos.

Nous devons installer les paquets clients sur tous les nœuds de notre cluster. Nous allons exécuter la commande suivante:

```
[root@node-1 ~]# yum install krb5-workstation -y
```

```
Installed:
  krb5-workstation.x86_64 0:1.15.1-54.el7_9

Complete!
```

Ensuite, nous configurons le fichier /etc/krb5.conf qui contient les informations de configuration de Kerberos notamment:

```
[root@node-1 ~]#   vim /etc/krb5.conf
```

-Le nom du royaume (realm) Kerberos par défaut, le realm doit être en majuscule. Dans notre cas nous avons choisi comme realm (HADOOPSECURITY.COM).

-Le nom du serveur Kerberos (la machine sur laquelle le serveur Kerberos été installée). Dans notre cas, nous allons installer le serveur Kerberos sur la Machine Cloudera, la même machine sur laquelle nous avons installé Hadoop en mode maître.

-Le nom du serveur d'administration aussi c'est Cloudera.

```
[libdefaults]
 default_realm = HADOOPSECURITY.COM
 dns_lookup_realm = false
 dns_lookup_kdc = false
 ticket_lifetime = 24h
 renew_lifetime = 7d
 forwardable = true

 default_tgs_enctypes = aes256-cts-hmac-sha1-96 aes128-cts-hmac-sha1-96 arcfour-hmac-md5
 default_tkt_enctypes = aes256-cts-hmac-sha1-96 aes128-cts-hmac-sha1-96 arcfour-hmac-md5
 permitted_enctypes = aes256-cts-hmac-sha1-96 aes128-cts-hmac-sha1-96 arcfour-hmac-md5

[realms]
 HADOOPSECURITY.COM = {
  kdc = cloudera.us-central1-a.c.project2-362310.internal
  admin_server = cloudera.us-central1-a.c.project2-362310.internal
  max_renewable_life = 7d
 }
```

3.3.3.2 Serveur Kerberos:

Sur le terminal de la machine maître :nous allons exécuter la commande suivante Pour installer les paquets pour un serveur Kerberos.

```
[root@cloudera ~]# sudo yum install -y krb5-server
```

Nous configurons le fichier/etc/krb5.conf:

```
[root@cloudera ~]# vim /etc/krb5.conf
```

```
[libdefaults]
 default_realm = HADOOPSECURITY.COM
 dns_lookup_realm = false
 dns_lookup_kdc = false
 ticket_lifetime = 24h
 renew_lifetime = 7d
 forwardable = true

 default_tgs_enctypes = aes256-cts-hmac-sha1-96 aes128-cts-hmac-sha1-96 arcfour-hmac-md5
 default_tkt_enctypes = aes256-cts-hmac-sha1-96 aes128-cts-hmac-sha1-96 arcfour-hmac-md5
 permitted_enctypes = aes256-cts-hmac-sha1-96 aes128-cts-hmac-sha1-96 arcfour-hmac-md5

[realms]
 HADOOPSECURITY.COM = {
  kdc = cloudera.us-central1-a.c.project2-362310.internal
  admin_server = cloudera.us-central1-a.c.project2-362310.internal
  max_renewable_life = 7d
 }
```

Nous pouvons modifier le fichier/var/krb5kdc/kdc.conf pour ajouter quelque type de cryptage.

```
[root@cloudera ~]# vim /var/kerberos/krb5kdc/kdc.conf
```

```
[kdcdefaults]
 kdc_ports = 88
 kdc_tcp_ports = 88

[realms]
 HADOOPSECURITY.COM = {
  #master_key_type = aes256-cts
  acl_file = /var/kerberos/krb5kdc/kadm5.acl
  dict_file = /usr/share/dict/words
  admin_keytab = /var/kerberos/krb5kdc/kadm5.keytab
  supported_enctypes = aes256-cts:normal aes128-cts:normal des3-hmac-sha1:normal arcfour-hmac:normal des-hmac-sha1:normal des-cbc-md5:nor
mal des-cbc-crc:normal
max_renewable_life = 7d
 }
```

Un type de chiffrement Kerberos**:** est une combinaison spécifique d'un algorithme de chiffrement et

d’un algorithme d’intégrité pour assurer la confidentialité et l’intégrité des données et améliore la sécurité globale du service Kerberos.

Dans la section suivante[realms]du fichier kdc.conf on trouve:

acl file: emplacement du fichier de liste de contrôle d’accès.

Admin_keytab: emplacement du fichier kadm5.keytab.

Kdc_ports: Liste des ports sur lesquels le serveur Kerberos doit écouter les requêtes UDP Par défaut, les ports 88 et 750 sont utilisés pour le KDC, et le port 749 est utilisé pour le démon d'administration du KDC.

Master_key_type : Le type de clé de la clé principal. Il est utilisé pour déterminer le type de cryptage qui chiffre les entrées de principal dans la base de données.

Max_renewable_life: spécifie la période maximale pendant laquelle un ticket valide peut être renouvelé dans ce domaine.

Supported_enctypes: spécifie la clé par défaut des principaux pour ce realm.

Kerberos utilise un fichier de liste de contrôle d’accès (ACL)pour gérer les droits d’accès à la base de données Kerberos.

L’emplacement par défaut du fichier Kerberos (ACL) c’est:”/var/krb5kdc/kadm5.acl”.

Le fichier kadm5.acl vous permet d’autoriser ou d’interdire les privilèges pour les principaux individuels.

```
[root@cloudera ~]# vim /var/kerberos/krb5kdc/kadm5.acl
```

L’entrée suivante dans le fichier kadm5.acl donne à tout principal dans le domaine HADOOPSECURITY.COM avec l’instance admin tous les privilèges sur la base de données Kerberos.

```
*/admin@HADOOPSECURITY.COM      *
```

Ensuite, nous créons une base de données KDCaveckdb5_util.

L’utilitaire kdb5_util permet à un administrateur de créer, vider, charger et détruire la base de données KerberosV5.

```
[root@cloudera ~]# kdb5_util create
Loading random data
Initializing database '/var/kerberos/krb5kdc/principal' for realm 'HADOOPSECURITY.COM',
master key name 'K/M@HADOOPSECURITY.COM'
You will be prompted for the database Master Password.
It is important that you NOT FORGET this password.
Enter KDC database master key:
Re-enter KDC database master key to verify:
```

3.3.3.3 Kerberos Principal:

Un principal :est le nom unique d'un utilisateur ou d'un service autorisé à s'authentifier à la de Kerberos.

Pour accéder directement à la base de données KDC ont tape la commande suivante:

```
[root@cloudera ~]# kadmin.local
Authenticating as principal root/admin@HADOOPSECURITY.COM with password.
kadmin.local:
```

Kadmin. Local : est une interface de ligne de commande du système d'administration Kerberos. Pour ajouter un nouveau principal, on tape la commande**« addprinc »,** en demandant deux fois un mot de passe.

Pour chaque instance du service Hadoop, on doit créer un principal Kerberos correspondant à ce service.

```
kadmin.local:  addprinc cm/admin
WARNING: no policy specified for cm/admin@HADOOPSECURITY.COM; defaulting to no policy
Enter password for principal "cm/admin@HADOOPSECURITY.COM":
Re-enter password for principal "cm/admin@HADOOPSECURITY.COM":
Principal "cm/admin@HADOOPSECURITY.COM" created.
kadmin.local:
```

Une fois que les nouveaux principaux sont ajoutés, on tape « list_principals » dans kadmin.local pour les afficher:

```
kadmin.local:  list_principals
HTTP/cloudera.us-central1-a.c.project2-362310.internal@HADOOPSECURITY.COM
HTTP/node-1.us-central1-a.c.project2-362310.internal@HADOOPSECURITY.COM
K/M@HADOOPSECURITY.COM
abc@HADOOPSECURITY.COM
cm/admin@HADOOPSECURITY.COM
hdfs/cloudera.us-central1-a.c.project2-362310.internal@HADOOPSECURITY.COM
hdfs/node-1.us-central1-a.c.project2-362310.internal@HADOOPSECURITY.COM
hdfs@HADOOPSECURITY.COM
hive/cloudera.us-central1-a.c.project2-362310.internal@HADOOPSECURITY.COM
hue/cloudera.us-central1-a.c.project2-362310.internal@HADOOPSECURITY.COM
kadmin/admin@HADOOPSECURITY.COM
kadmin/changepw@HADOOPSECURITY.COM
kadmin/cloudera.us-central1-a.c.project2-362310.internal@HADOOPSECURITY.COM
kiprop/cloudera.us-central1-a.c.project2-362310.internal@HADOOPSECURITY.COM
krbtgt/HADOOPSECURITY.COM@HADOOPSECURITY.COM
mapred/cloudera.us-central1-a.c.project2-362310.internal@HADOOPSECURITY.COM
user2@HADOOPSECURITY.COM
yarn/cloudera.us-central1-a.c.project2-362310.internal@HADOOPSECURITY.COM
yarn/node-1.us-central1-a.c.project2-362310.internal@HADOOPSECURITY.COM
zookeeper/cloudera.us-central1-a.c.project2-362310.internal@HADOOPSECURITY.COM
kadmin.local:
```

Maintenant que tout est configurée nous pouvons redémarre les services krb5-kdc et krb5 admin-server pour que les modifications seront prises en compte

```
[root@cloudera ~]# systemctl restart krb5kdc.service
[root@cloudera ~]# systemctl restart kadmin.service
[root@cloudera ~]#
```

➢ Obtenir des Tickets:

Si on exécute la commande **klist** pour afficher les tickets on reçoit le message suivant :

```
[root@cloudera ~]# klist
klist: No credentials cache found (filename: /tmp/krb5cc_0)
```

Cela veut dire que nous avons aucun utilisateur (principal)authentifiée, il n'ya pas de ticket. Pour obtenir des tickets ont tape simplement kinit suivi du nom d'un principal, puis on tape le mot de passe. kinit obtient et échange un ticket initial d'attribution de ticket TGT pour le principal :

```
[root@cloudera ~]# kinit cm/admin
Password for cm/admin@HADOOPSECURITY.COM:
[root@cloudera ~]#
```

Affichage des tickets : on utilise la commande klist pour afficher les informations sur le ticket.

```
[root@cloudera ~]# klist
Ticket cache: FILE:/tmp/krb5cc_0
Default principal: cm/admin@HADOOPSECURITY.COM

Valid starting       Expires              Service principal
09/17/2022 11:25:58  09/18/2022 11:25:58  krbtgt/HADOOPSECURITY.COM@HADOOPSECURITY.COM
        renew until 09/24/2022 11:25:58
```

Ticket cache : est l'emplacement du fichier de ticket. Dans l'exemple ci-dessus, ce fichier s'appelle /tmp/krb5cc_0.Le principal par défaut est notre principal Kerberos 'cm/admin@HADOOPSECURITY.COM'.

Les champs **Valid starting** et **Expires** décrivent la période de validité du ticket. Le **Service principal** décrit chaque ticket. Le tgt a un premier composant krbtgt et un second composant qui est le nom de domaine.

-Détruire un ticket(**kdestroy**):cet utilitaire sert à détruire les tickets d'autorisation Kerberos actifs de l'utilisateur en supprimant le cache des informations d'identification qui les contient

```
[root@cloudera ~]# kinit cm/admin
Password for cm/admin@HADOOPSECURITY.COM:
[root@cloudera ~]# klist
Ticket cache: FILE:/tmp/krb5cc_0
Default principal: cm/admin@HADOOPSECURITY.COM

Valid starting       Expires              Service principal
09/17/2022 11:25:58  09/18/2022 11:25:58  krbtgt/HADOOPSECURITY.COM@HADOOPSECURITY.COM
        renew until 09/24/2022 11:25:58
[root@cloudera ~]# kdestroy
[root@cloudera ~]# klist
klist: No credentials cache found (filename: /tmp/krb5cc_0)
[root@cloudera ~]#
```

3.3.4 Configuration de Hadoop avec Kerberos:

Pour lancer Kerberos, on clique sur Enable Kerberos.

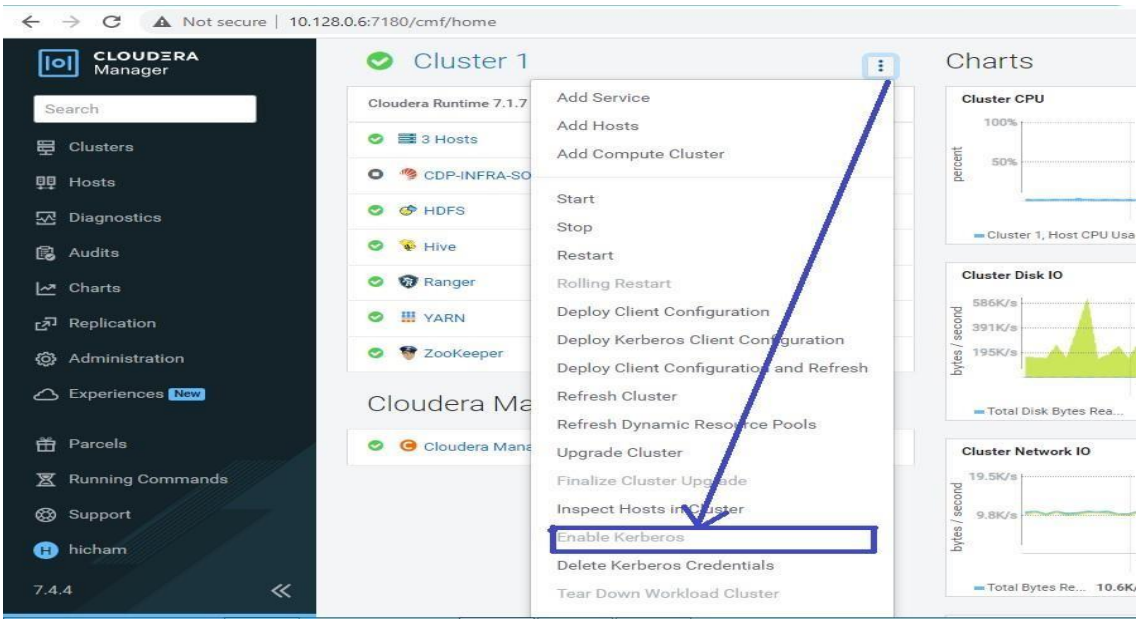


Figure 3. 12 Activation de Kerberos

Ensuite, une page de démarrage apparaît. On Sélection le type de KDC applicable pour afficher les étapes de configuration pour votre type de KDC. Après la configuration, on coche la case : J'ai terminé toutes les étapes ci-dessus. Puis on clique sur Continuer

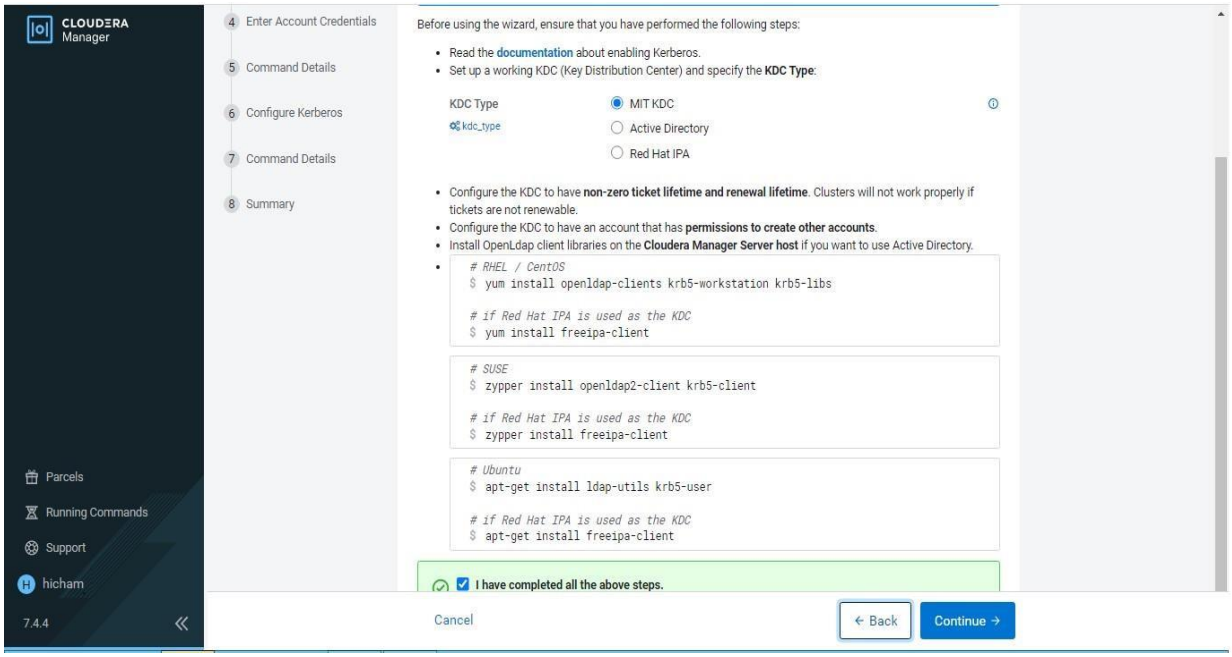


Figure 3. 13 Spécification de type de kdc

Nous allons saisir des valeurs pour les types de cryptage Kerberos, le domaine de sécurité Kerberos, le serveur KDC.

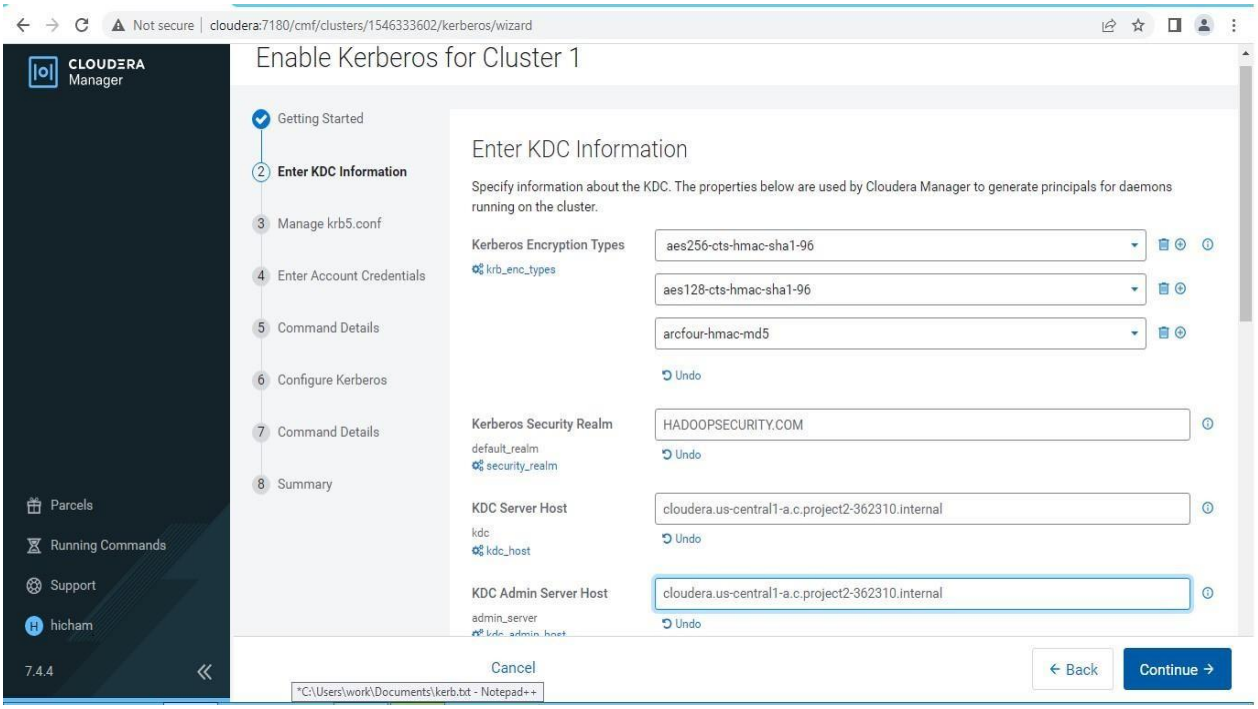


Figure 3. 14 Spécification des informations relatives au KDC

Gérer **krb5.conf** : Cette étape permet de spécifier si Cloudera Manager déploie et gère ou non le fichier **krb5.conf** sur votre cluster. Dans notre cas nous avons préconfiguré le fichier **krb5.conf**.

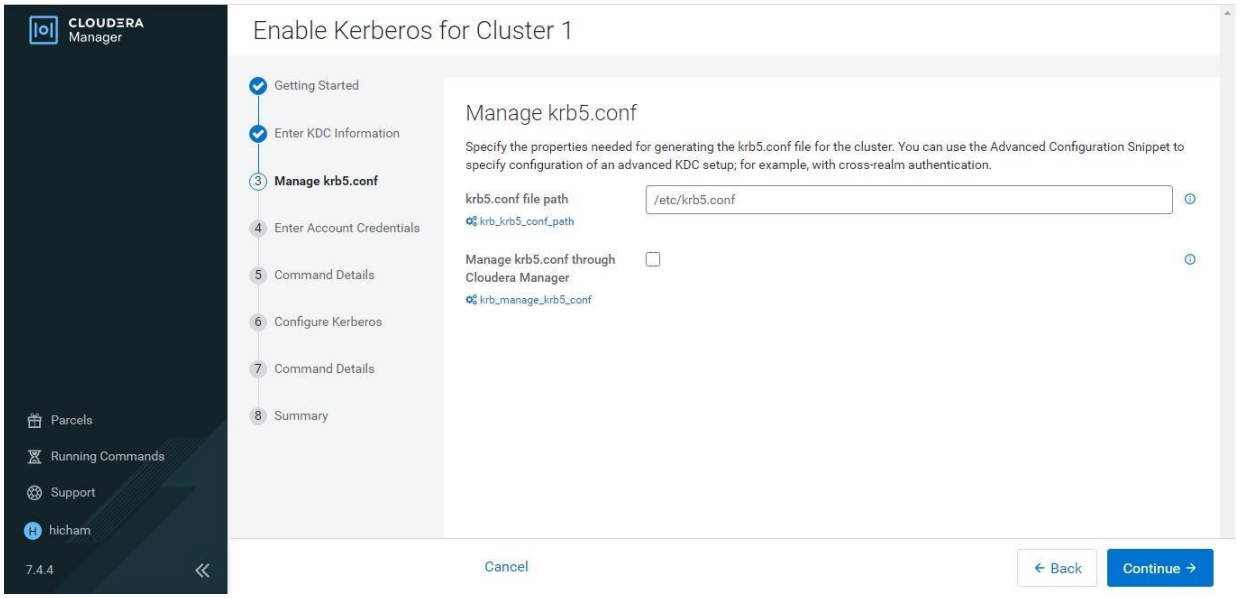


Figure 3. 15 Définition du chemin du fichier préconfiguré krb5.conf

Nous allons saisir les informations d'identification du compte : saisissons le nom d'utilisateur et le mot de passe de l'utilisateur.

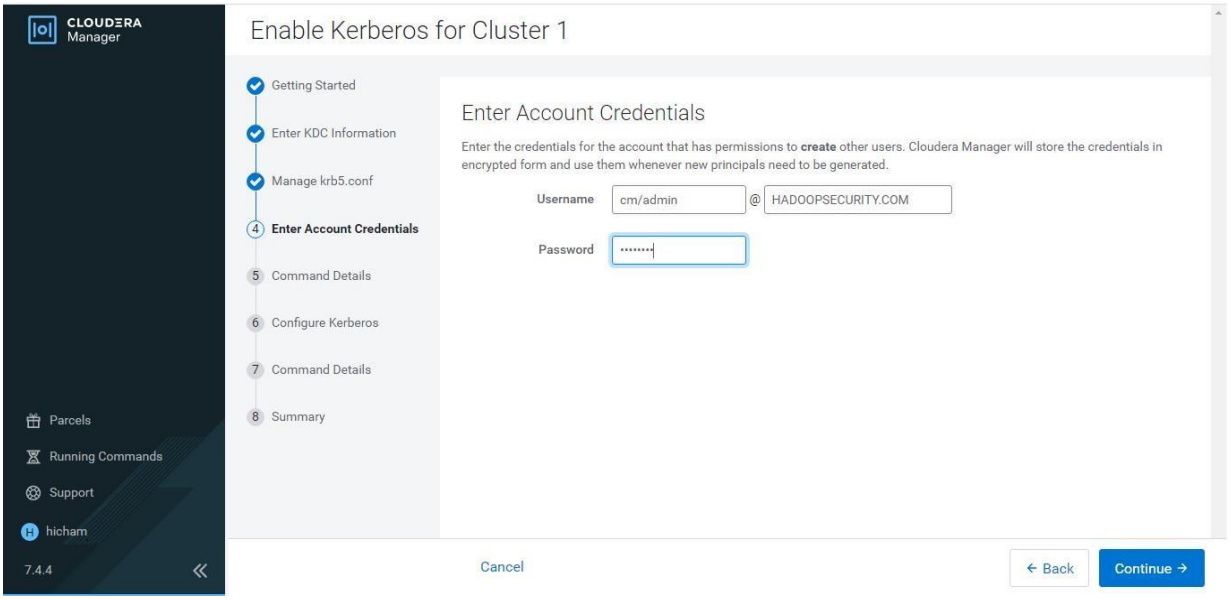


Figure 3. 16 Définition de nom utilisateur et mot de passe du l'administrateur

L'assistant définit automatiquement les ports privilégiés nécessaires au protocole de l'émetteur-

récepteur DataNode et à l'interface Web HTTP. Web UI, mais nous l'avons défini sur Default Kerberos Principals

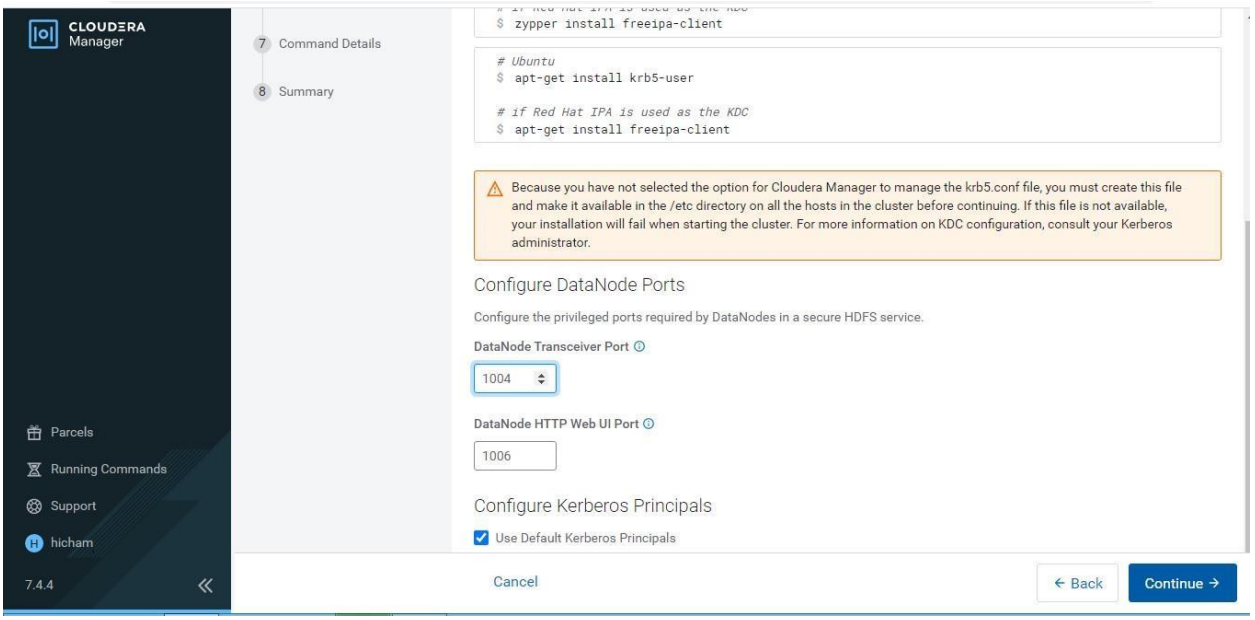


Figure 3. 17 Configuration des ports de DataNode

La page Détails de la commande affiche le résultat de la commande Enable Kerberos.

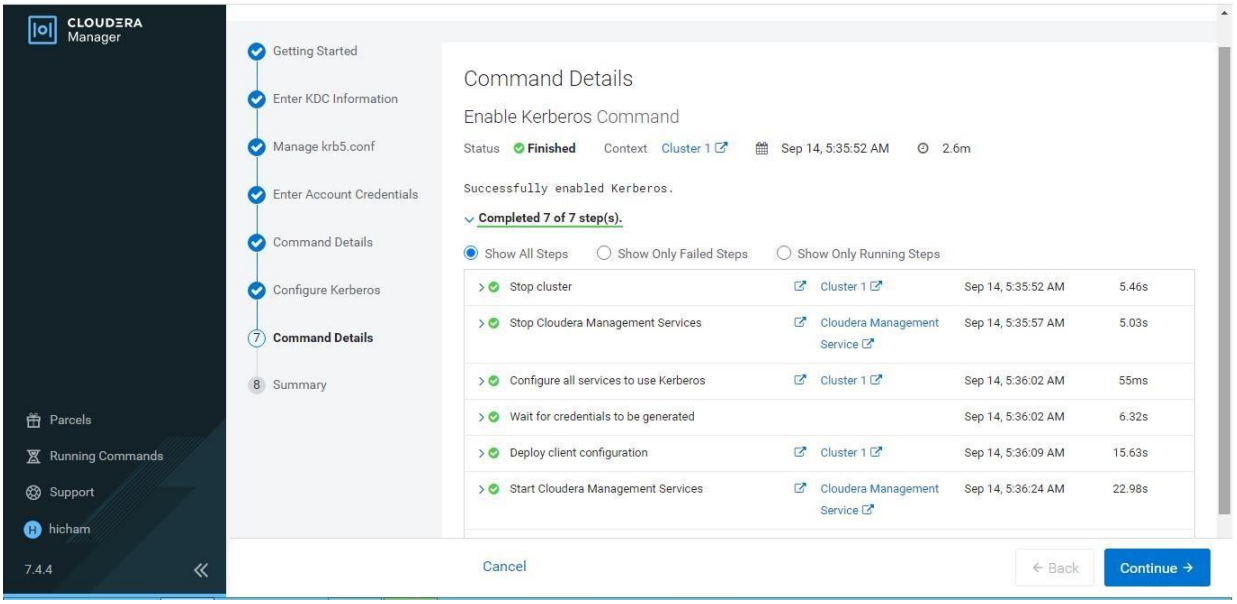


Figure 3. 18 résultats des commandes d'activation kerberos

La dernière étape de la liste de l'assistant montre que Kerberos a été activé avec succès.

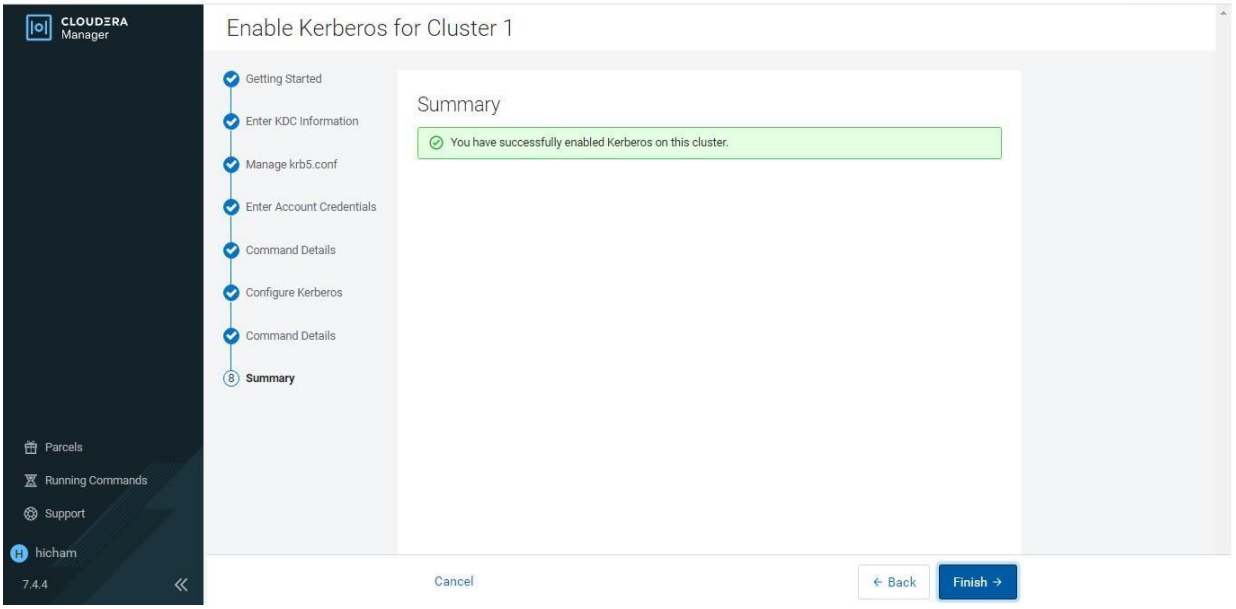


Figure 3. 19 Résumé des étapes d'installation de Kerberos dans le cluster

3.3.5 Activation d'Apache Ranger :

Après l'installation de Cloudera Manager et l'ajout d'un cluster, des étapes supplémentaires sont nécessaires pour terminer l'installation d'Apache Ranger.

3.3.5.1 Activer les plugins:

Les plugins Ranger pour HDFS peuvent ne pas être activés par défaut. Les plugins Ranger permettent aux composants de la pile Cloudera Manager - tels que HDFS et Solr - de se connecter à Ranger et d'accéder à ses services d'autorisation et d'audit.

Pour activer le plugin HDFS, il faut suivre les étapes suivantes :

1-Accéder à la page d'état du service HDFS et Cliquer sur l'onglet Configuration.

2-Rechercher la propriété de configuration Enable Ranger Authorization.

3-Si la propriété Enable Ranger Authorization n'est pas sélectionnée, sélectionner-la et enregistrer les modifications.

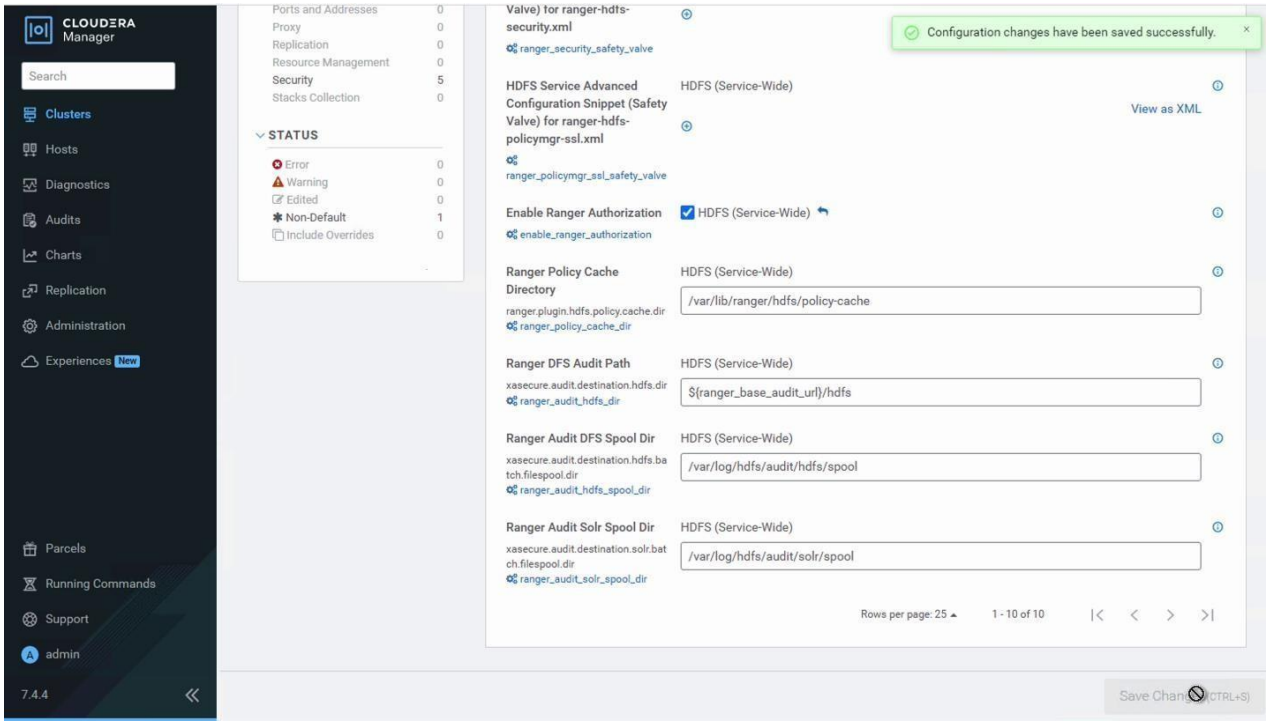


Figure 3. 20 Activation de l'option Ranger autorisation

Ensuite, il faut redémarrer le service HDFS :

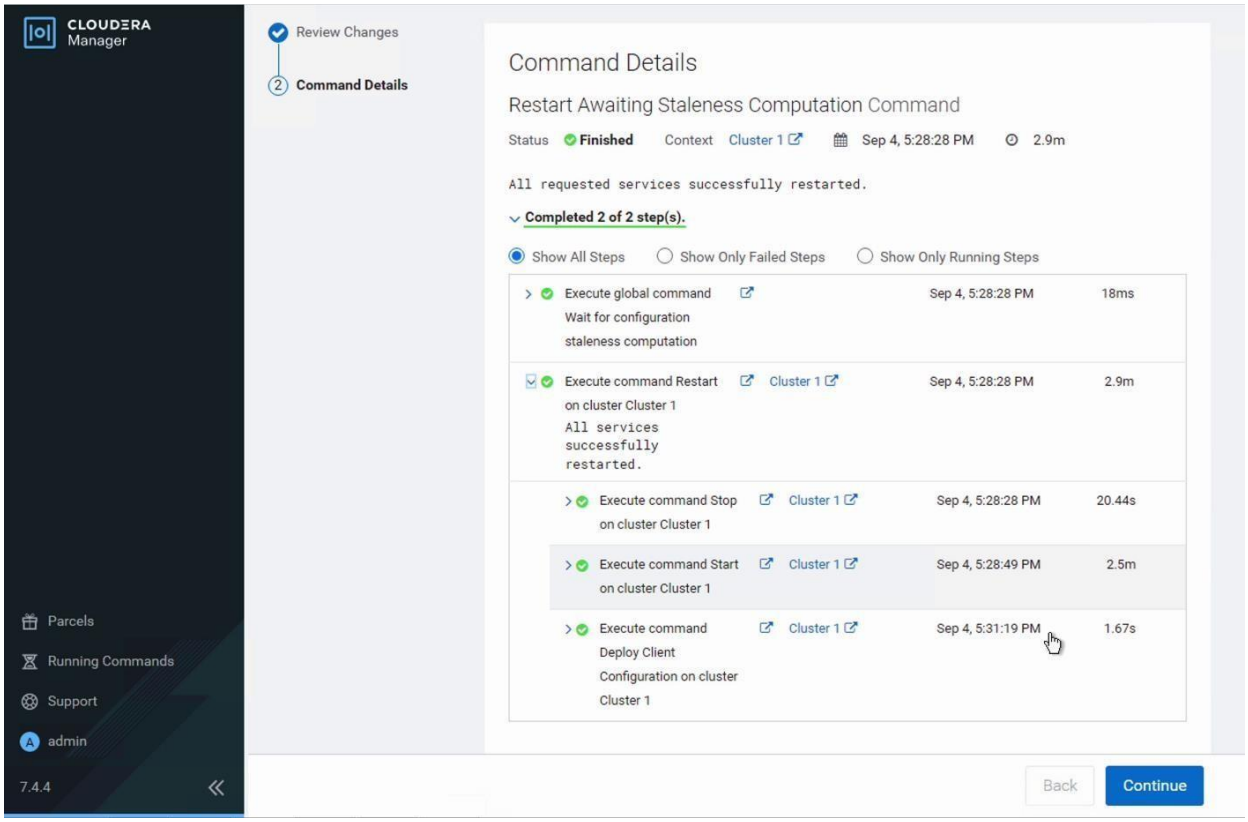


Figure 3. 21 Redémarrage de service HDFS

Nous allons sur la page d'état du service Ranger et nous cliquons sur Actions > Configurer le service plugin Ranger.

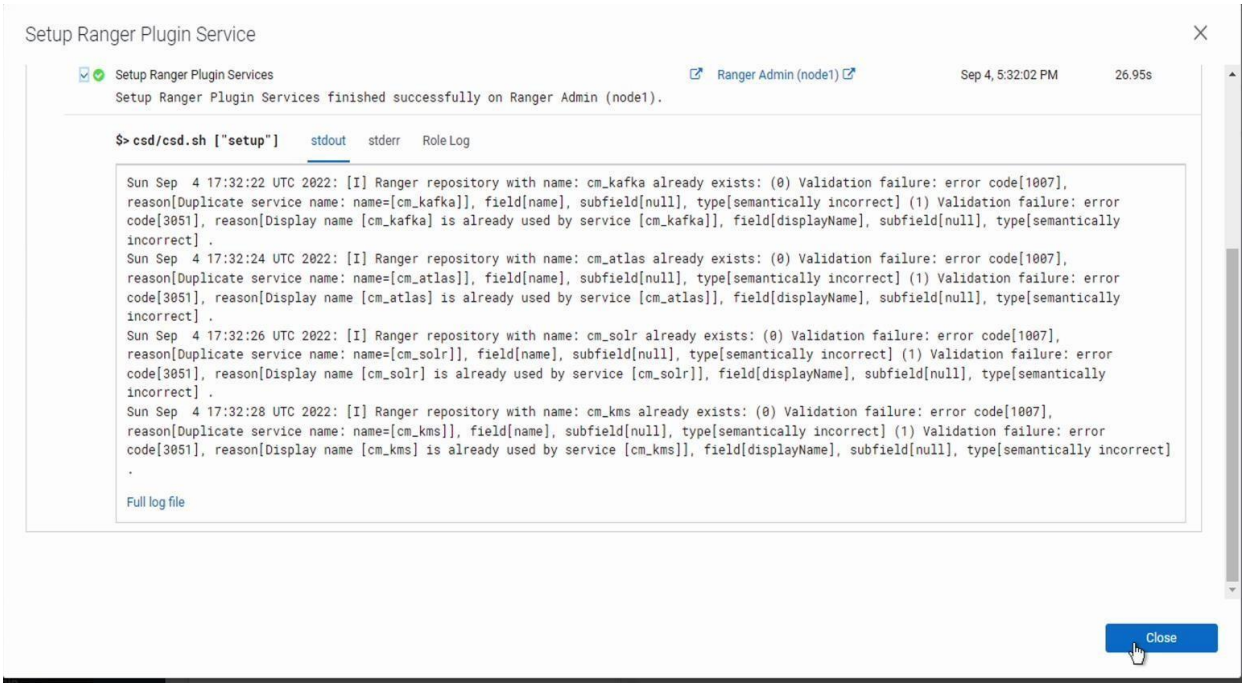


Figure 3. 22 Configuration de ranger plugin services

3.3.6 Vérification du ranger:

Pour vérifier l'utilité de Ranger on va simuler deux scénario :

- **Scénario 1:** nous devons créer un répertoire dans le HDFS en passant par le processus d'authentification Kerberos pour obtenir un ticket.

Pour créer un nouveau groupe d'utilisateurs, on exécute la commande suivante**.**

```
[centos@node1 ~]$ sudo groupadd mlops
```

Pour créer un nouvel utilisateur dans le groupe créé (mlops), on exécute la commande suivante**.**

```
[centos@node1 ~]$ sudo useradd hicham -g mlops
```

Nous devons créer un principal "hicham" pour que le nouvel utilisateur "hicham" puisse obtenir un ticket Kerberos.

Pour que l'utilisateur " hicham " soit autorisé, nous devons créer une politique pour cet utilisateur dans Ranger en suivant les étapes suivantes:

Aller sur la page d'état du service Ranger et cliquer sur Interface web d'administration de ranger:

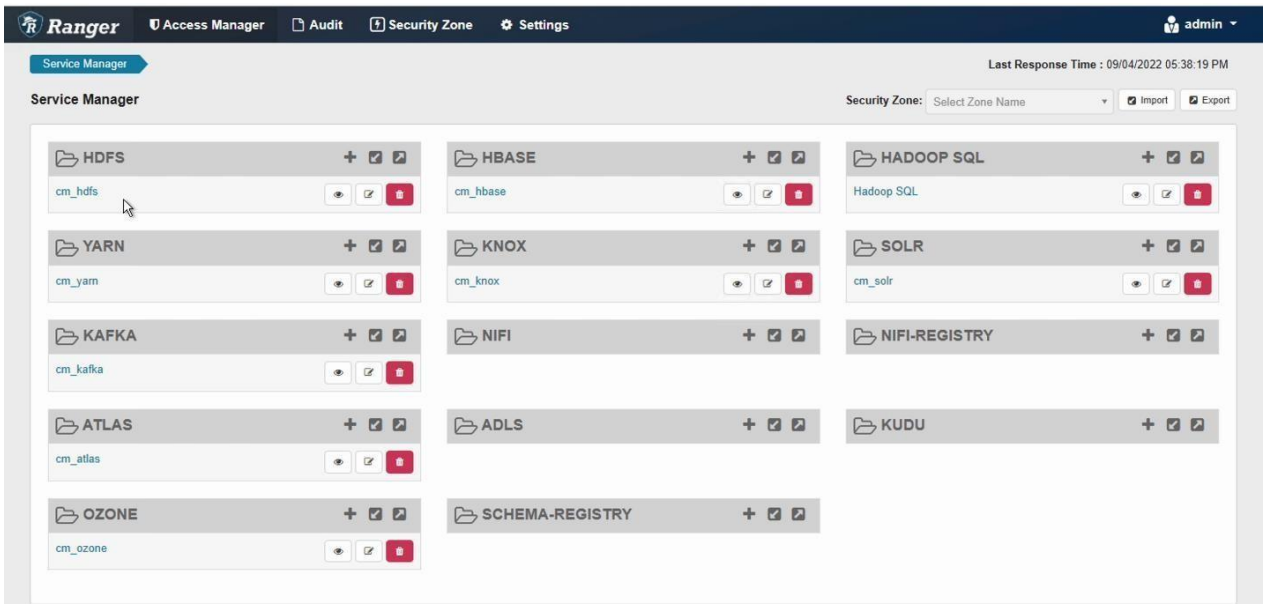


Figure 3. 23 Tableau des services manipulés par ranger

Nous ajoutons une politique en cliquant sur HDFS puis nous définirons le nom de la politique, le chemin de ressource et les permissions accordée, le nom d'utilisateur et le groupe auquel il appartient.

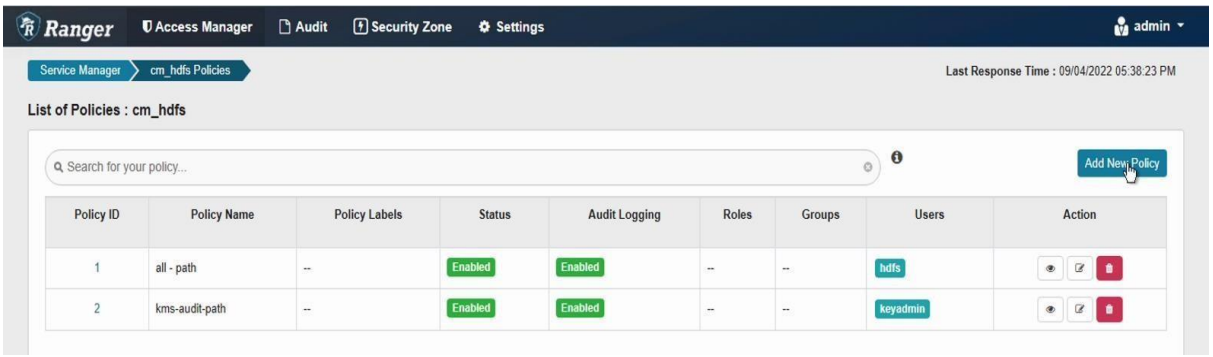


Figure 3. 24 Liste des politiques de cm_hdfs

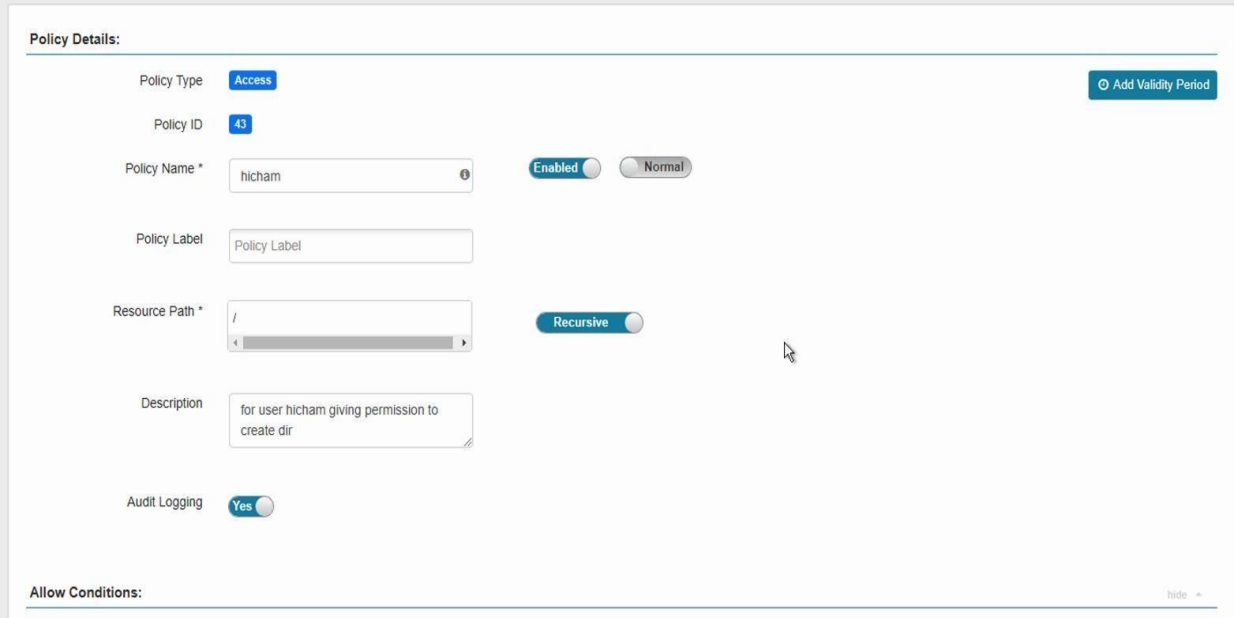


Figure 3. 25 Ajout d'une politique-1

Nous sélectionnerons les conditions:

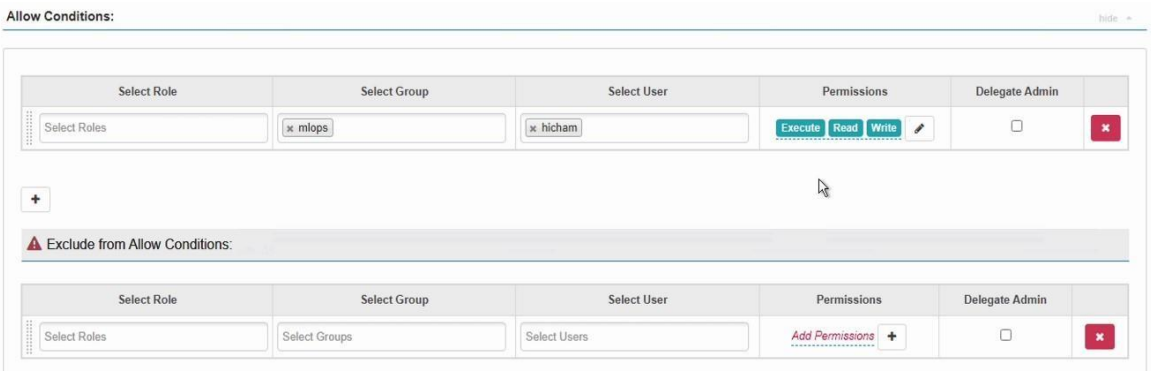
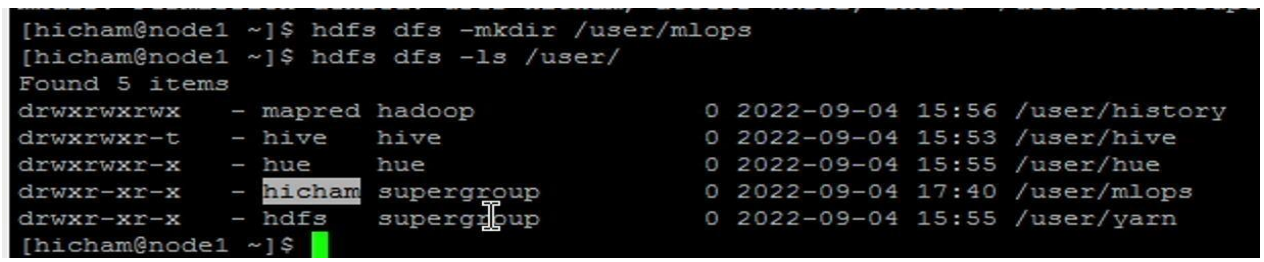


Figure 3. 26 Ajout d'une politique-2

Ensuite, nous créons le répertoire /user/mlops:



- **Scénario 2: Exécutions des Job :**

Dans ce scénario, nous avons exécuté un job dans l'HDFS en suivant les étapes suivantes :

1) Dans le cluster, nous gérons l'accès par le biais de Ranger, donc nous nous assurons d'abord que l'utilisateur a la permission appropriée.

2) Nous avons utilisé les jars fournies par Cloudera pour cette activité.

3) Exécuter les commandes respectives pour obtenir la sortie et la confirmer sur HDFS/Local.

3.3.6.1 TeraSort benchmark :

Le benchmark TeraSort a pour but de trier 1TB de données (ou toute autre quantité de données que vous voulez) aussi vite que possible. Il s'agit d'un benchmark qui combine le test des couches HDFS et MapReduce d'un cluster Hadoop.[99]

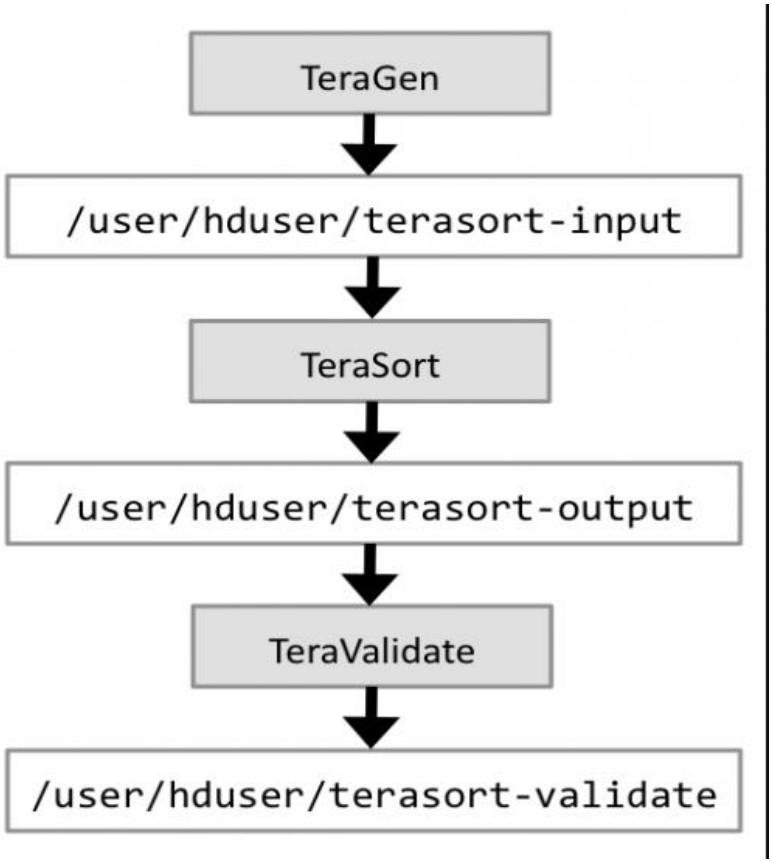


Figure 3. 27 montre le flux de données de base [99]

- **TeraGen:** génère des données aléatoires qui peuvent être facilement utilisées comme données d'entrée pour une exécution ultérieure de TeraSort.[99]

La syntaxe pour exécuter TeraGen est la suivante :

$ hadoop jar hadoop-*examples*.jar teragen <number of 100-byte rows><output dir>

- pour notre cas la commande a passé dans notre cluster est :

```
[hdfs@cm jars]$ hadoop jar hadoop-mapreduce-examples-3.1.1.7.1.7.0-551.jar teragen 1000 /tmp/terasort
```

- **TeraSort:** est implémenté comme une tâche de tri MapReduce avec un partitionneur personnalisé qui utilise une liste triée de n-1 clés échantillonnées qui définissent la plage de clés pour chaque réduction.[99]

La syntaxe pour exécuter le benchmark TeraSort est la suivante :

$ hadoop jar hadoop-*examples*.jar terasort <input dir><output dir>

- pour notre cas la commande a passé dans notre cluster est :

```
[hdfs@cm jars]$ hadoop jar hadoop-mapreduce-examples-3.1.1.7.1.7.0-551.jar terasort /tmp/terasort /tmp/terasort-output
```

- **TeraValidate:** TeraValidate garantit que les données de sortie de TeraSort sont globalement triées.[99]

La syntaxe pour exécuter le test TeraValidate est la suivante :

$ hadoop jar hadoop-*****examples*****.jar teravalidate <terasort output dir**(=** input data**)**><teravalidate output dir>

- pour notre cas la commande a passé dans notre cluster est :

```
[hdfs@cm jars]$ hadoop jar hadoop-mapreduce-examples-3.1.1.7.1.7.0-551.jar teravalidate /tmp/terasort-output /tmp/terasort-validate
```

3.3.6.2 MapReduce benchmark (MRbench) :

MRBench vérifie si les petites tâches sont réactives et fonctionnent efficacement sur votre cluster. Il se concentre sur la couche MapReduce, car son impact sur la couche HDFS est très limité.[99]

La commande pour exécuter une boucle de 50 petits travaux de test est la suivante :

```
$ hadoop jar hadoop-*test*.jar mrbench -numRuns 50
```

- pour notre cas la commande a passé dans notre cluster est :

3.4 Conclusion:

Dans ce chapitre, nous avons décrit les différentes étapes de la mise en place de l'outil ranger. Ensuite nous avons présenté les étapes nécessaires à l'utilisation de cet outil.

Conclusion Générale et perspectives :

Dans le cadre du processus Hadoop, plusieurs types de données sont combinés et stockés dans un lac de données Hadoop, puis les données stockées sont traitées en conséquence.[98]
Les données peuvent être sensibles (les détails de la carte de crédit de l'utilisateur, les détails bancaires, les mots de passe). Pour les sécuriser, il est possible de mettre en place diverses stratégies telles que l'exclusion des utilisateurs non autorisés et des intrusions à l'aide de pare-feu, la fiabilité de l'authentification des utilisateurs, la formation des utilisateurs finaux, etc. [98]

Les administrateurs utilisent l'authentification et l'autorisation comme deux processus de sécurité de l'information essentiels pour protéger les environnements Hadoop. L'identité d'un utilisateur ou d'un service est confirmée par l'authentification, et ses privilèges d'accès sont établis par l'autorisation. Plusieurs techniques et protocoles d'authentification et d'autorisation sont actuellement disponibles pour sécuriser les systèmes Hadoop.

Dans ce projet, nous avons pu mettre en place le protocole d'authentification et d'autorisation pour Hadoop qui est ranger. Ce protocole a été utilisé pour assurer les deux mécanismes de sécurité (authentification et autorisation).

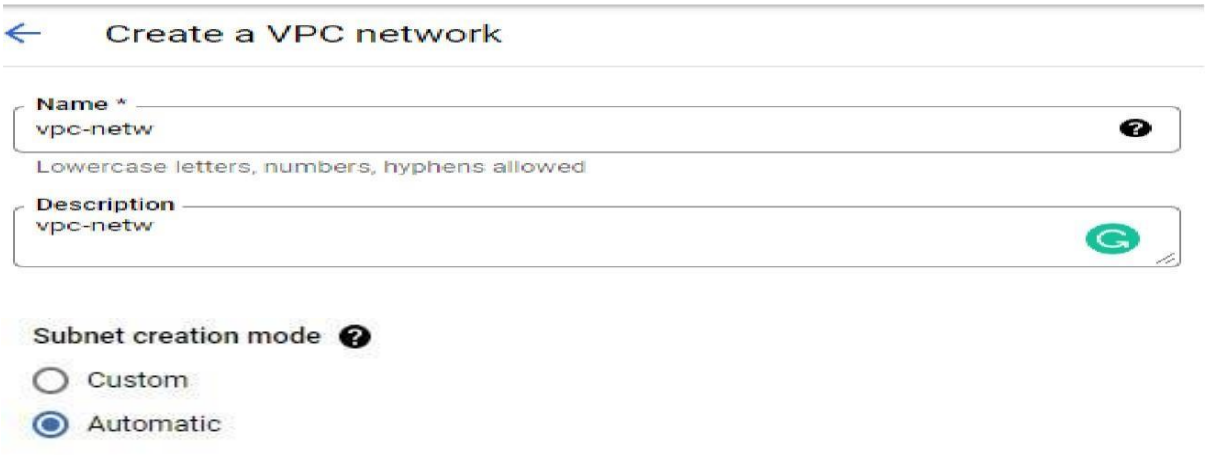
En guise de perspective pour les futurs travaux, nous envisagerons de réaliser l’objectif initial de ce projet qui consiste à faire une étude comparative des protocoles d'authentification et d'autorisation afin d’évaluer leurs performances.

Annexes:

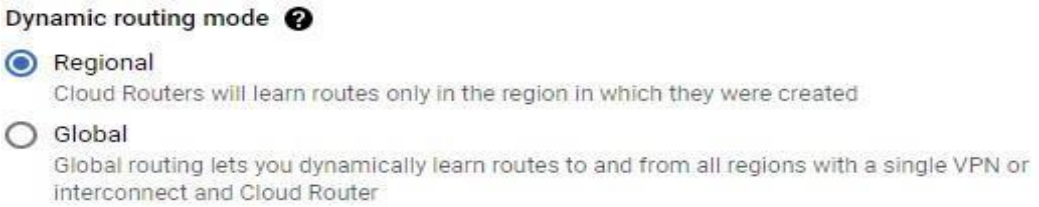
Annexe A:

- Configuration du réseau:

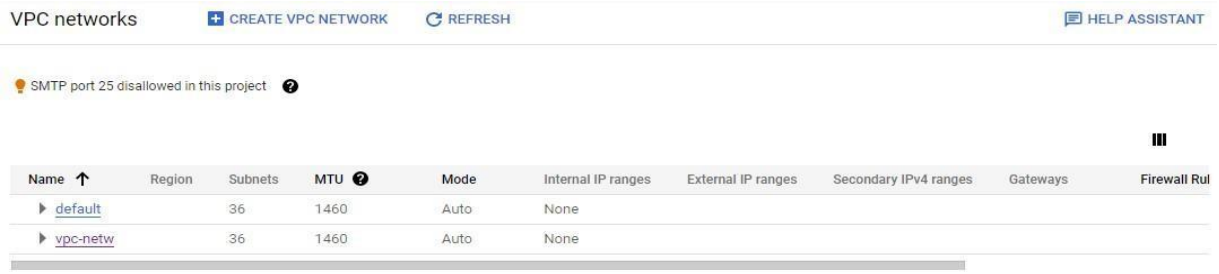
A partir du panneau gcp cliquez sur VPC network>create vpc network : Créer notre réseau en spécifiant le nom, le mode de routage dynamique et le mode de création de sous-réseau, le reste est par défaut.



Le mode de création automatique prend en charge l'IPv4. Ces plages d'adresses IP seront attribuées à chaque région de notre réseau VPC. Lorsqu'une instance est créée pour notre réseau VPC, une IP lui sera attribuée à partir de la plage d'adresses de la région appropriée.



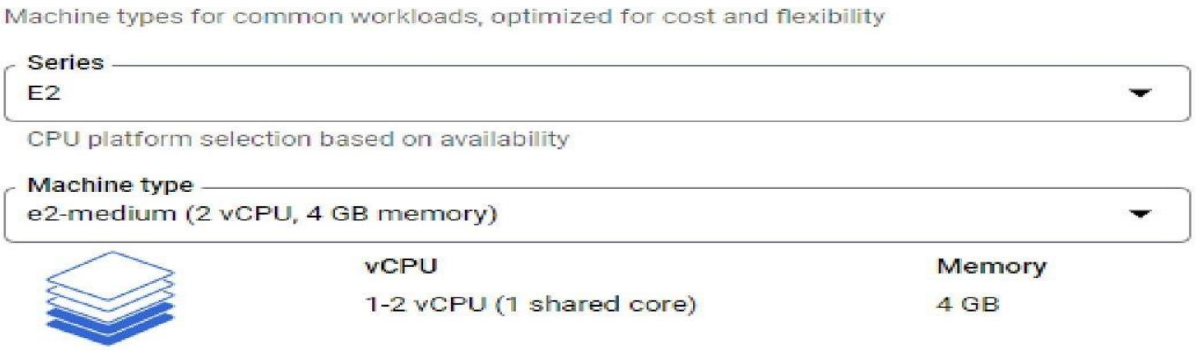
Mode de routage dynamique régional : Chaque routeur cloud du réseau VPC annonce les plages de sous-réseau IPv4 primaires et secondaires dans la même région que le routeur cloud.



Ensuite, nous créons notre environnement de travail qui est le Windows Server, 2012 R2Datacenter.

A partir du panneau gcp cliquez sur Compute Engine>create Instance. Nom et région du serveur:

La série et le type de serveur :



Boot Disk image pour notre serveur

Le serveur va fonctionner et accéder au trafic de réseau "vpc-netw" que nous avons créé.

Basic information

Name	workspace
Instance Id	5625969117228508607
Description	None
Type	Instance
Status	Running
Creation time	Sep 4, 2022, 2:34:25 PM UTC+02:00
Zone	us-central1-a
Instance template	None
In use by	None
Reservations	Automatically choose
Labels	None
Deletion protection	Disabled
Confidential VM service	Disabled
Preserved state size	0 GB

- Serveur créé avec succès.

Annexe B:

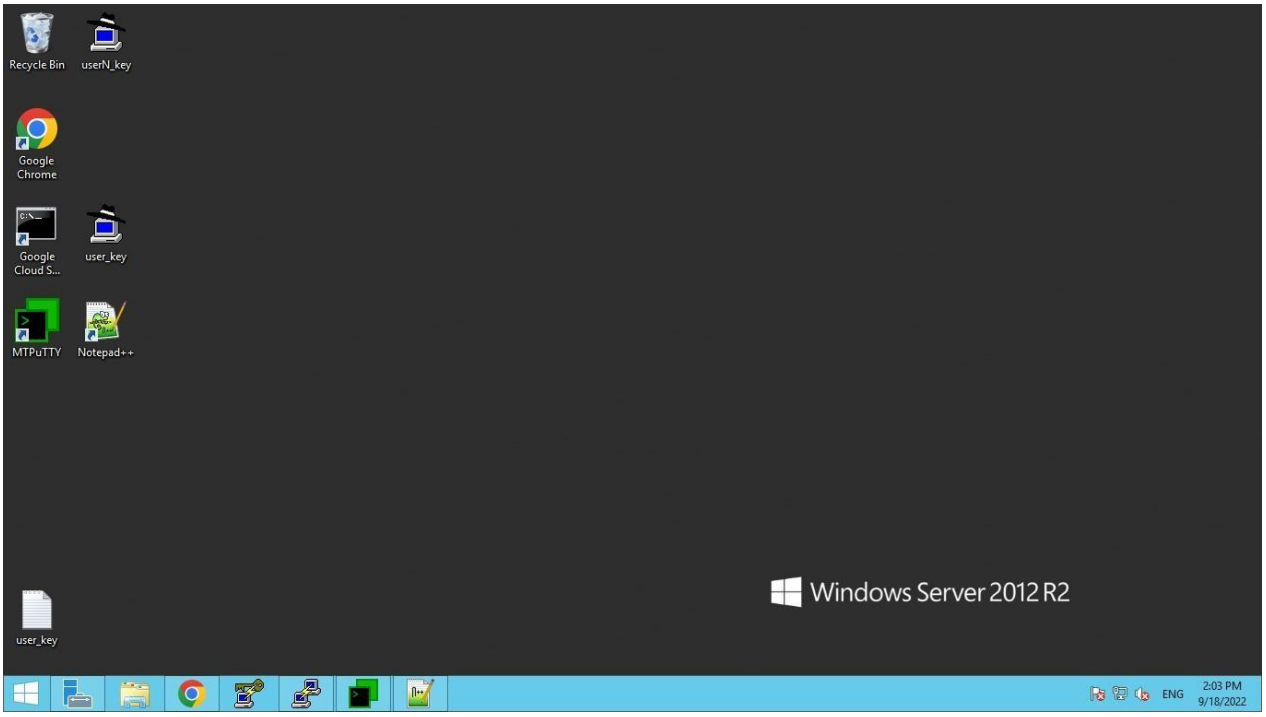
Boot disk image pour les machines master et esclaves:

Storage

Boot disk

Name	cloudera
Image	centos-7-v20220822
Size	50 GB
Interface type	SCSI
Type	SSD persistent disk
Encryption type	Google-managed
Mode	Boot, read/write
Snapshot schedule	None

Annexe C:



- **Manipulation de cluster via le serveur:**

Afin d'accéder et de manipuler notre cluster depuis le serveur, nous aurons besoin d'une connexion ssh établie par Putty entre le serveur et les machines du cluster.

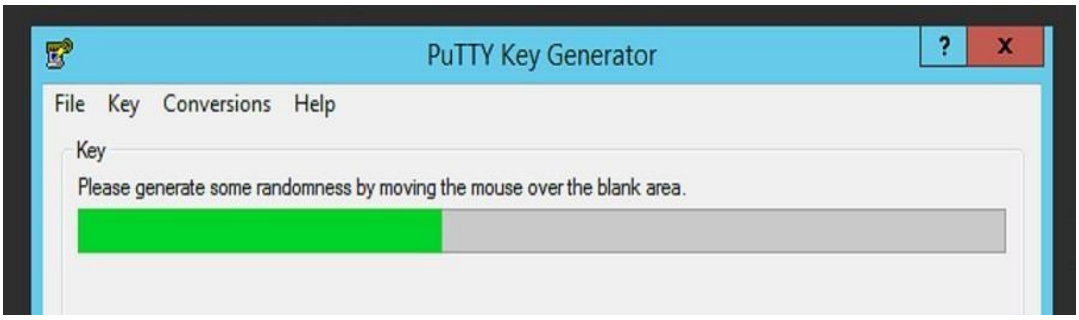
PuTTY est un logiciel de terminal flexible pour Windows. C'est le client SSH libre le plus utilisé sur la planète. Il prend en charge les connexions SSH, telnet et raw socket, ainsi qu'une bonne émulation de terminal. Il prend en charge l'authentification par clé publique et l'authentification unique Kerberos. Il contient également des implémentations SFTP et SCP en ligne de commande. PuTTY utilise tous les protocoles énumérés ci-dessus pour permettre une session à distance sur une machine via un réseau. Il s'agit d'un outil de communication textuel populaire, ainsi que d'un programme permettant de connecter des serveurs Linux à des postes de travail basés sur le système d'exploitation Microsoft.[92][93]

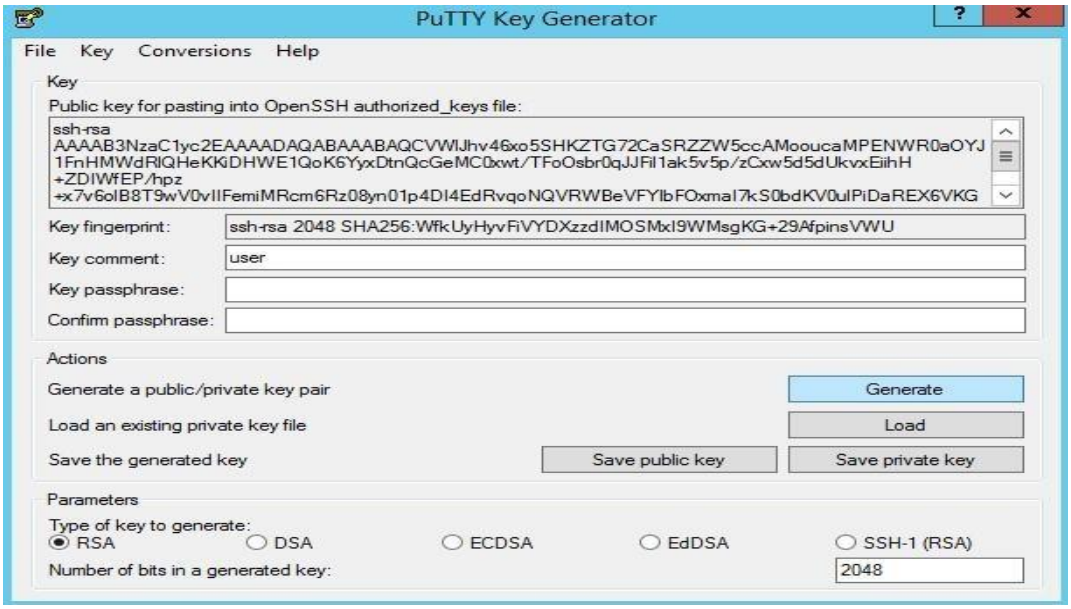
Nous utiliserons MTPuTTY car il offre la possibilité d'ouvrir une seule fenêtre et d'avoir plusieurs onglets ouverts pour exécuter toutes nos connexions multiples en même temps.

Pour authentifier le serveur avec les machines du cluster, nous aurons besoin de clés privées partagées entre eux.

Après l'installation de PuTTY et de MTPuTTY, nous générons les clés privées en utilisant PuTTY key Generator. :

Générateur de clé PuTTY>générer.

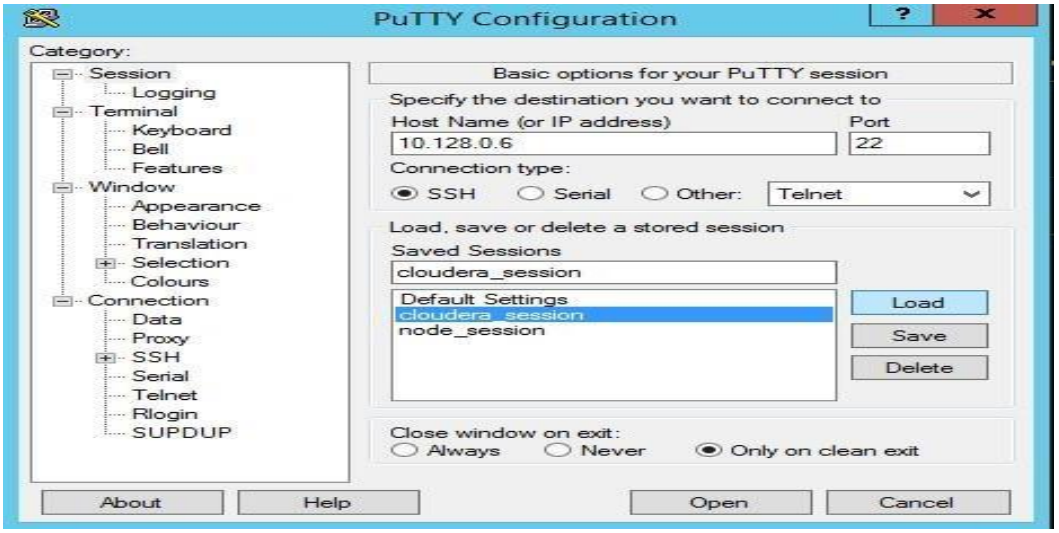




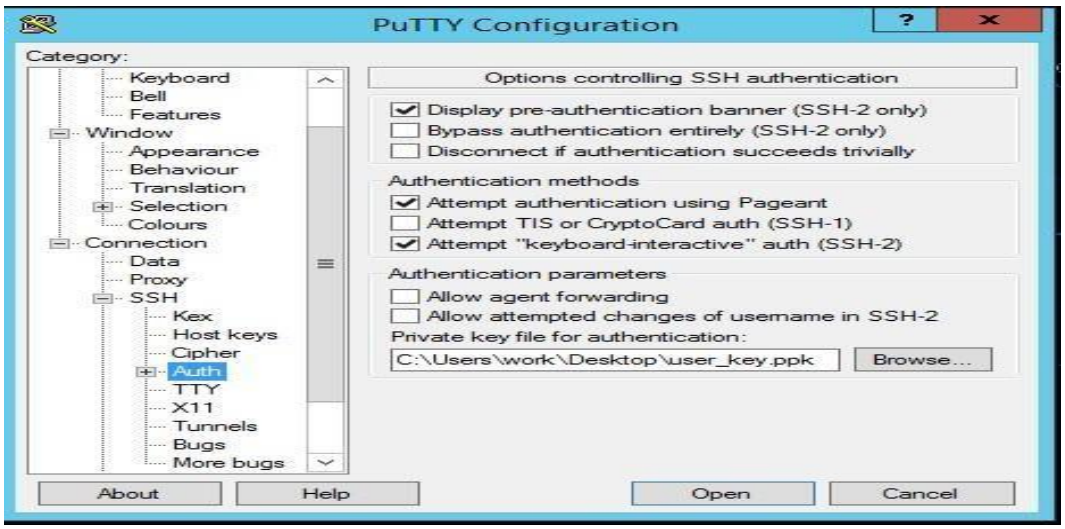
✓Clé privée générée avec succès.

Nous créons une session dans PuTTY afin d'accéder à la machine depuis le serveur:

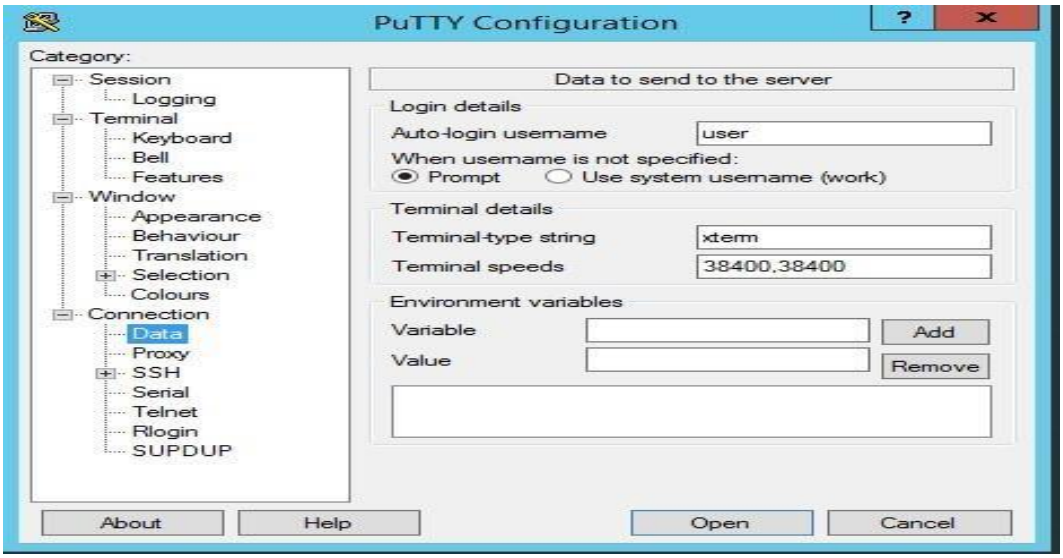
1-Nous soumettons l'HostName(adresse IP interne)de la machine ciblée.



2- Nous soumettons la clé privée que nous avons générée (user_key.ppk) à partir du générateur de clé PuTTY.



Sur Windows>data>nous définissons les détails de connexion du nom d'utilisateurDans notre cas "user".



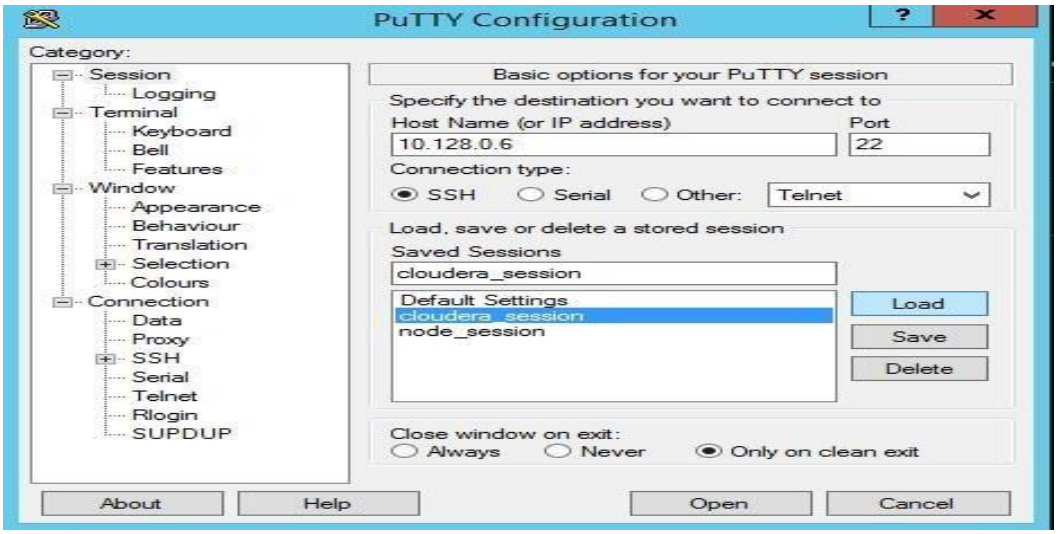
Ensuite, nous sauvegardons la session.

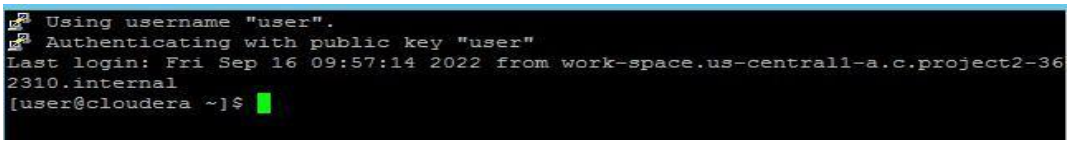
Nous soumettons la même clé dans les machines du cluster (node master et les nodes esclaves) sur la plateforme Google cloud afin qu'elle puisse établir la connexion ssh correctement.

Compute engine>cliquez sur la machine ciblée>edit>naviguez vers la section ssh>ajoutez la clé privée.

✓Connexion à la machine maître (même chose pour les machines esclaves)

Nous sélectionnons la session ciblée "cloudera session">load>open.

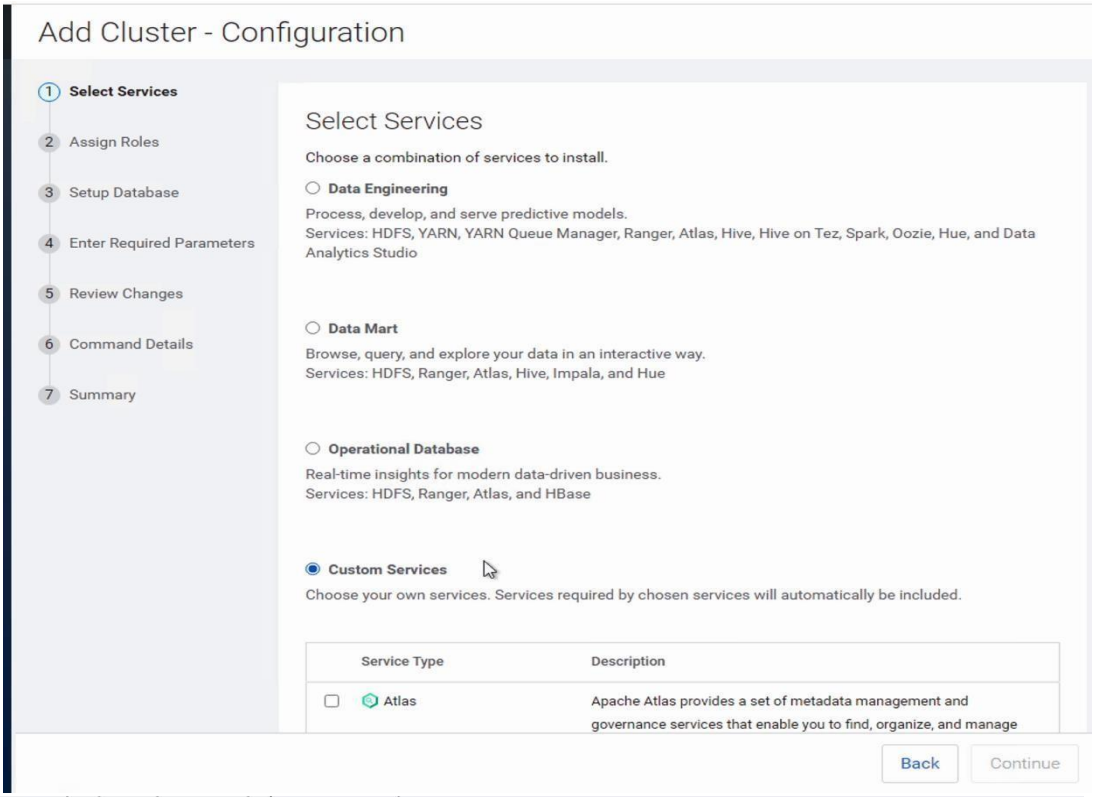


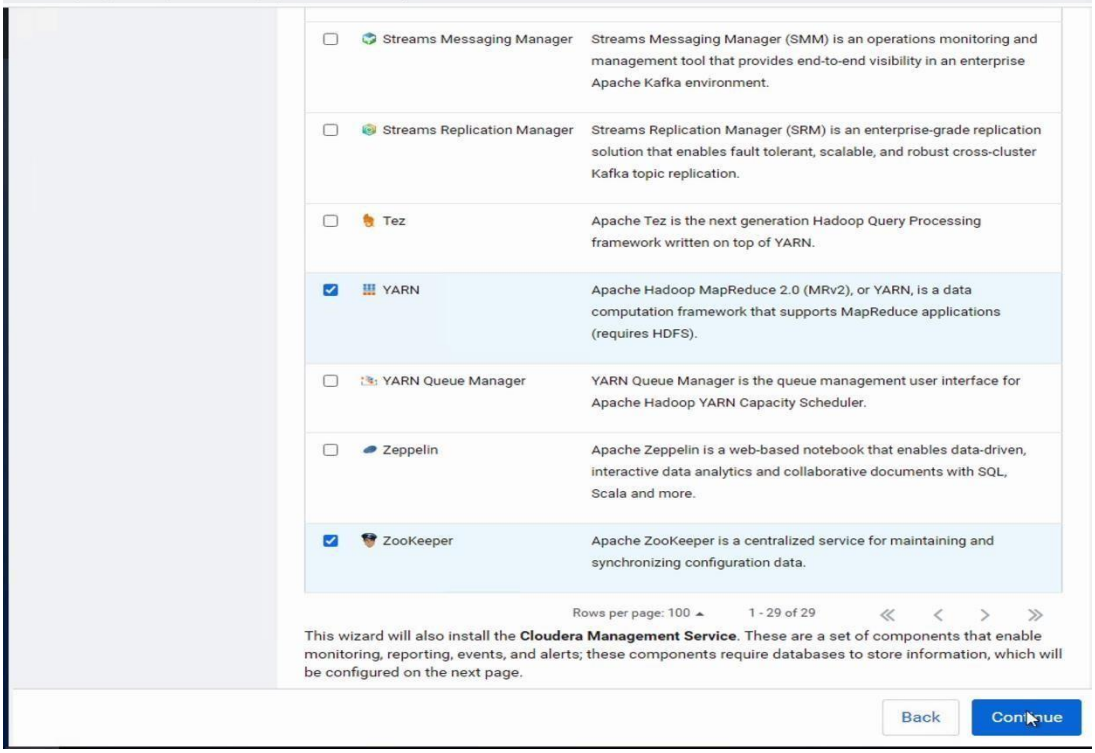


Annexe D

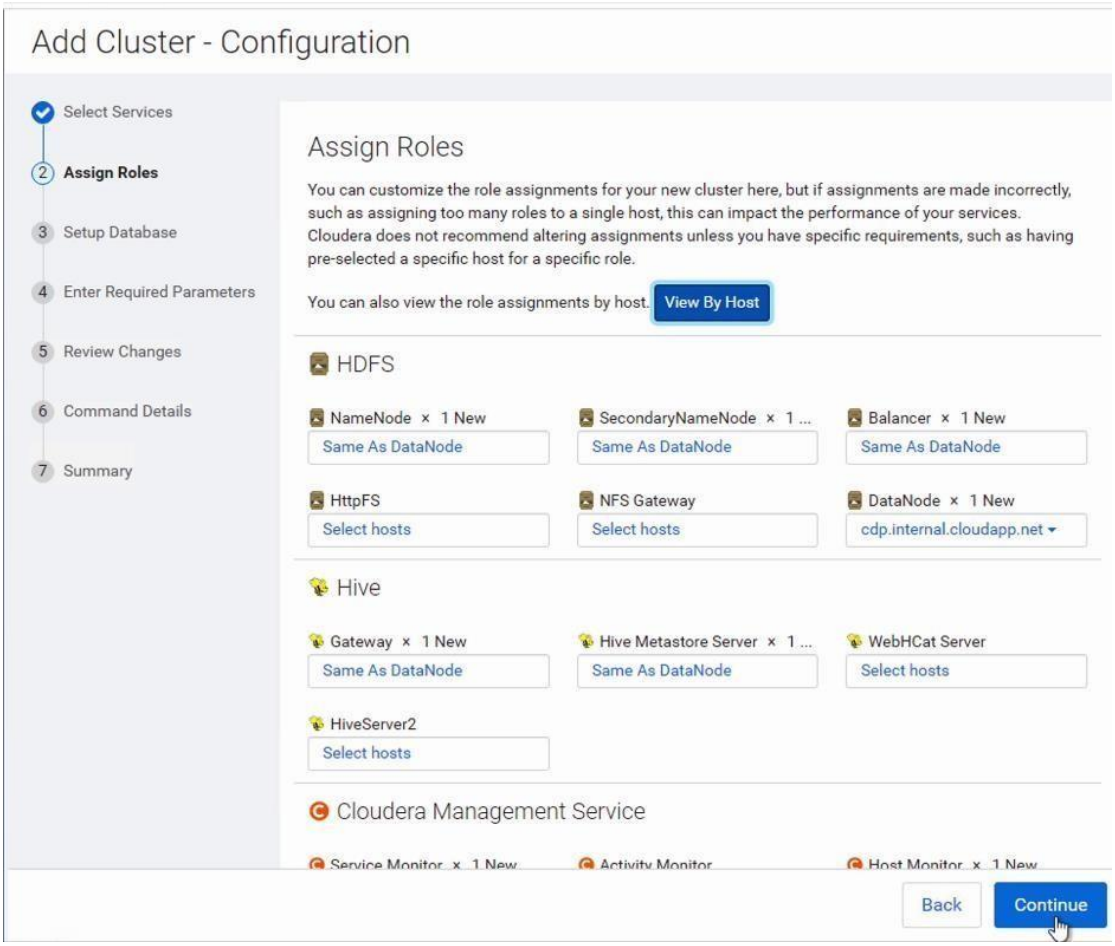
1-Sélectionner les services:

Pour le cluster, nous allons installer ces service : Hdfs, Hive, Ranger, Yarn et Zookeeper.

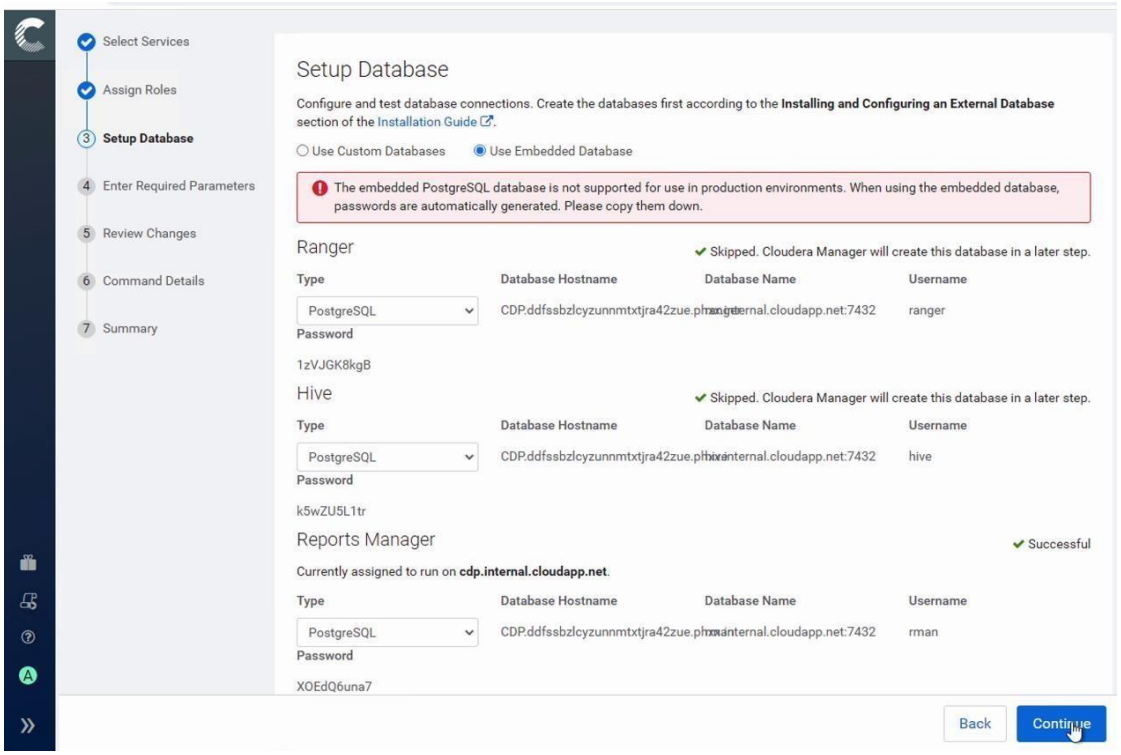




2- Attribuer des règles aux hôtes du cluster.

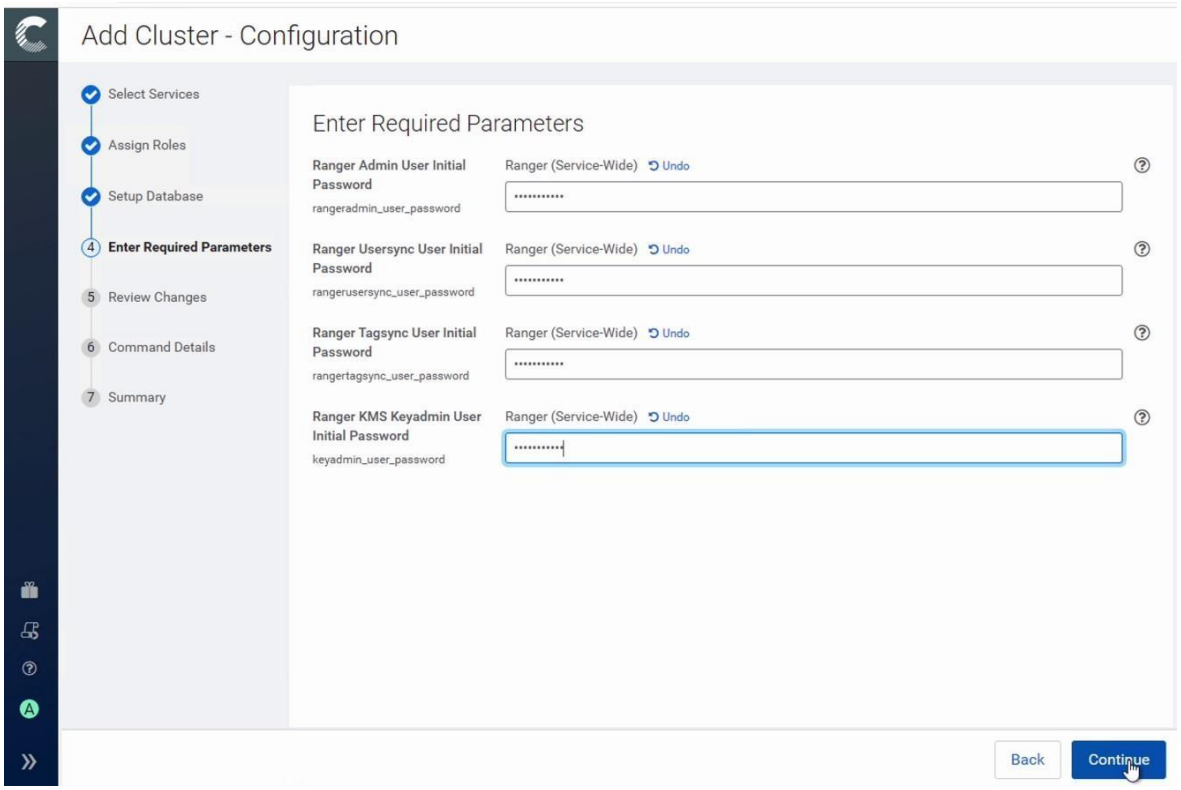


3- Configurer la base de données :

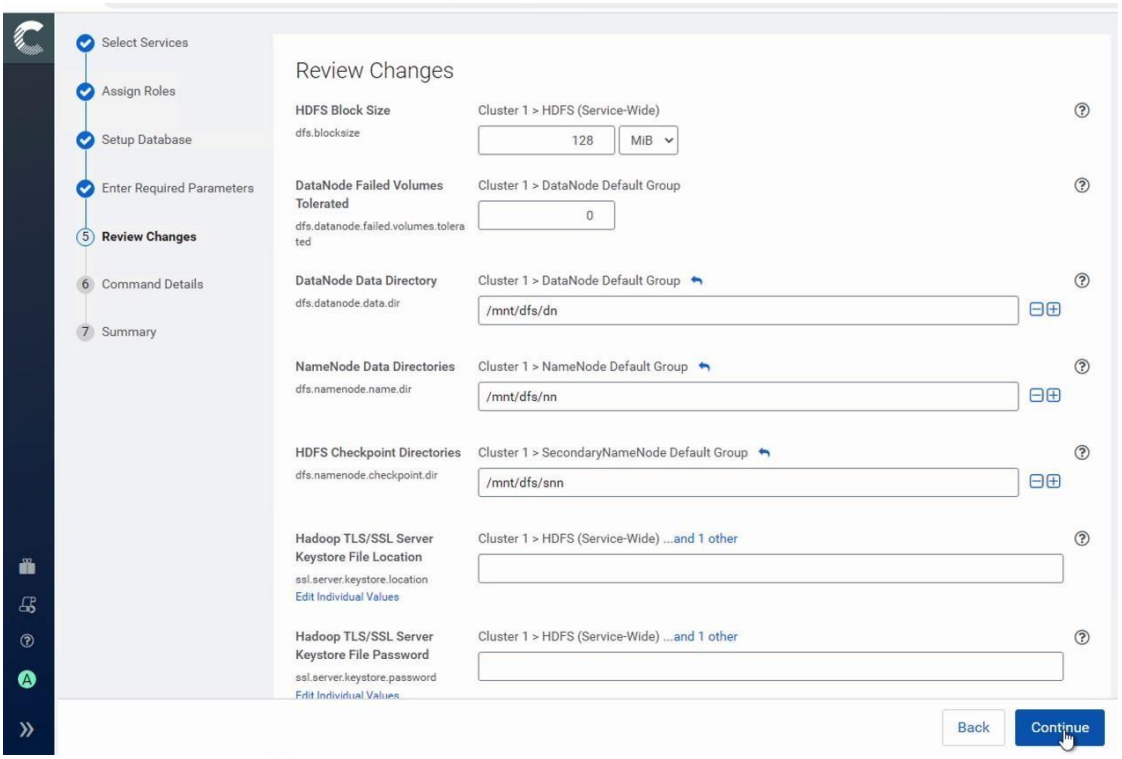


✓ La base de données a été crée et testée avec succès.

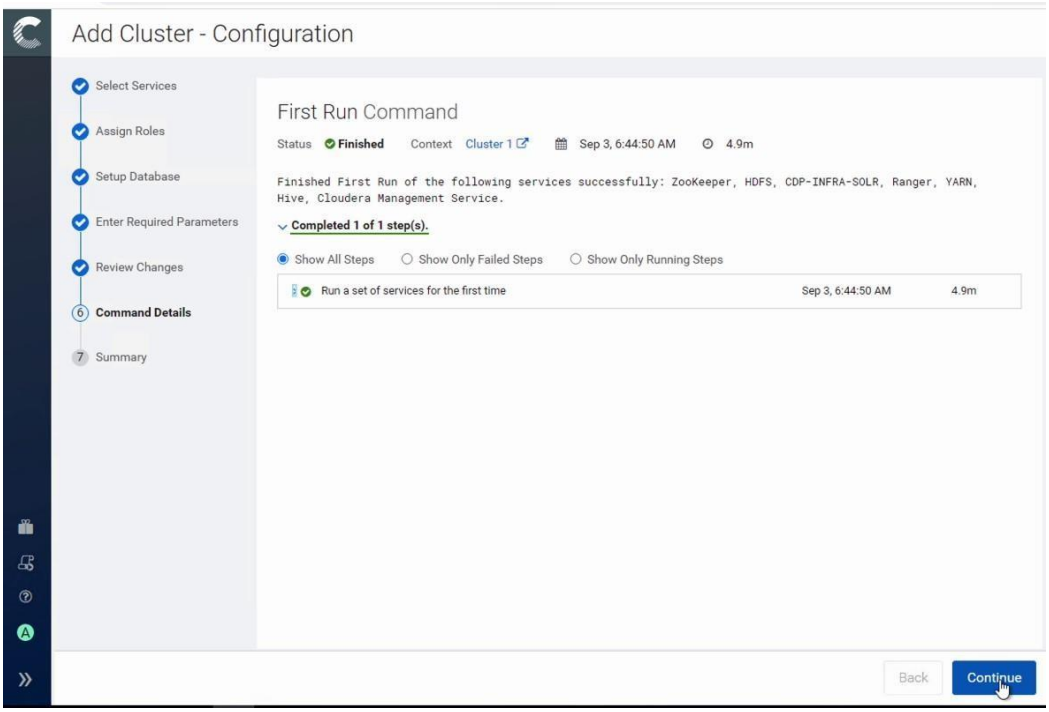
4-Introduire les paramètres requis:



✓ Revoir les changements : tout est bon :

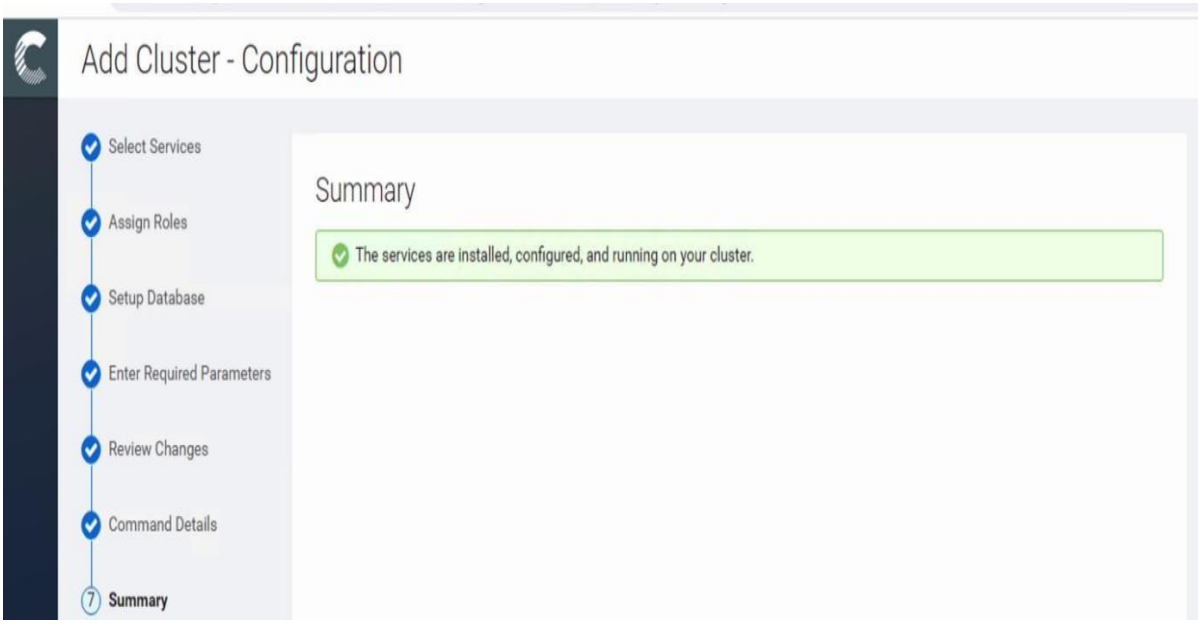


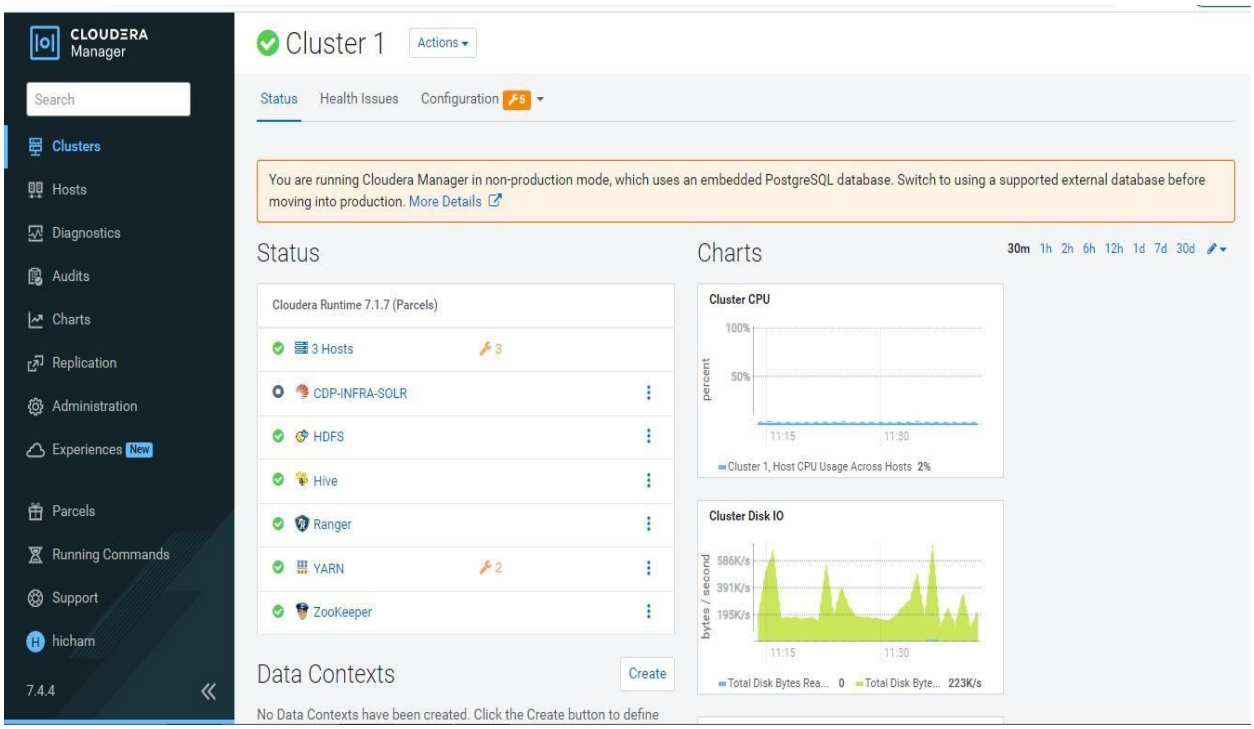
Détails de la commande: Exécutez les services pour la première fois:



✓ Les services ont été exécutés avec succès.

Sommaire:





✓ Notre cluster est mis en place avec succès.

REFERENCES :

Webographie :

[1].https://www.ionos.com/digitalguide/websites/web-development/what-is-a-megabyte/Consulté le 30 Juin 2022.

[2].https://www.zmescience.com/science/how-big-data-can-get/Consulté le 30 Juin 2022.

[3].https://blogs.cisco.com/sp/the-zettabyte-era-officially-begins-how-much-is-thatConsulté le 30 Juin 2022.

[4].https://now.northropgrumman.com/zipping-past-the-zettabyte-era-whats-next-for-the-internet/Consulté le 30 Juin 2022.

[5]https://www.bigdataframework.org/short-history-of-big-data/ Consulté le 27 Juin 2022.

[6].https://www.datakwery.com/techniques/big-data/ Consulté le 28 Juin2022

[7].https://www.sas.com/en_us/insights/big-data/what-is-big-data.html Consulté le 27 Juin 2022.

[8].http://www-01.ibm.com/software/data/bigdata/Consulté le 28 Juin 2022

[9].https://www.upgrad.com/blog/what-is-big-data-architecture-definition-layers-process-best-practices/ Consulté le 30 Juin 2022.

[10].https://www.educba.com/big-data-architecture/Consulté le 30 Juin 2022.

[14].https://www.domo.com/learn/infographic/data-never-sleeps-7 Consulté le 28 Juin 2022.

[32].https://www.geeksforgeeks.org/classification-of-data/Consulté le 28 Juin 2022.

[33].https://www.simplilearn.com/tutorials/big-data-tutorial/big-data applications?fbclid=IwAR1LrPlnb-DSdoEG H5iHsb3PPaYcKLFHMqDkfygTw2yLvgvycwhnZlmPG8Consulté le 28 Juin 2022.

[40].https://www.educba.com/big-data-technologies/ Consulté le 30 Juin 2022.

[42].https://www.techtarget.com/whatis/definition/SPSS-Statistical-Package-for-the-Social-Sciences#:~:text=SPSS%20(Statistical%20Package%20for%20the%20Social%20Sciences)%2C%20also%20known,expanded%20into%20other%20data%20markets.Consulté le 30 Juin 2022.

[44].https://yoyoclouds.wordpress.com/2011/12/15/hdfsarchitecture/ Consulté le 30 Juin 2022.

[49].https://economictimes.indiatimes.com/definition/authorization consulté le 30-08-2022

[50].https://frontegg.com/blog/authentication-vs-authorization consulté le 31-08-2022

[51].https://delinea.com/blog/access-control-models-methods consulté le 25-08-2022

[52].https://economictimes.indiatimes.com/definition/authenticationConsulté le 30 Juin 2022.

[53].https://www.sumologic.com/glossary/authentication-factor/Consulté le 30 Juin 2022.

[54].https://www.n-able.com/blog/network-authentication-methodsConsulté le 30 Juin 2022

[56].https://garantir.io/key-based-authentication/Consulté le 29 Juin 2022

[57].https://www.section.io/engineering-education/understand-hashing-in-cryptography/Consulté le 30 Juin 2022.

[58]. http://igm.univ-mlv.fr/~dr/XPOSE2006/depail/fonctionnement.html consulté le 21-08-2022

[59].https://www.gov.hk/en/residents/communication/infosec/digitalcert.htmConsulté le 30 Juin 2022.

[61].A1372_Maxim Integrated_Authentication, Implementing Secure Authentication Without Being a Cryptography Expert De: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiRgsC8scv4AhUZg_0HHVFTAMM4HhAWegQIJhAB&url=http%3A%2F%2Fdatasheet.octopar

t.com%2FMAXQ1061-KIT%2523-Maxim-Integrated-datasheet-82969687.pdf&usg=AOvVaw0rxP8AjgsTmkwuifK-oYxgConsulté le 30 Juin 2022.

[62].https://openclassrooms.com/fr/courses/1757741-securisez-vos-donnees-avec-la-cryptographie/6031874-creez-des-certificats-numeriques consulté le 26-08-2022

[80]. https://www.f5.com/labs/articles/education/what-is-access-control consulté le 09-09-2022

[81].https://www.infoguardsecurity.com/why-you-need-both-authorization-and-authentication/ consulté le 09-09-2022

[82].https://www.okta.com/identity-101/authentication-vs-authorization/consulté le 10-09-2022

[83].https://www.educba.com/authorization-types/ consulté le 15-09-2022

[84].https://similaranswer.fr/quest-ce-que-lauthentification-et-lautorisation-des-utilisateurs/consulté le 16-09-2022

[85].https://www.oreilly.com/library/view/cloudera-administration handbook/9781783558964/ch06s06.html consulté le 15-09-2022

[86].https://towardsdatascience.com/apache-hadoop-a-review-on-security-issues-and-solutions-for-hdfs-5ba06861b7cdconsulté le 20-09-2022

[88].https://learn.microsoft.com/en-us/windows/win32/srvnodes/windows-serverconsulté le 22-09-2022

[89].https://www.centos.org/about/ consulté le22-09-2022.

[90].https://linuxhint.com/everything-you-want-to-know-about-centos-as-linux-distribution/ consulté le 22-09-2022

[92].https://www.ssh.com/academy/ssh/putty consulté le 22-09-2022.

[93].https://www.techopedia.com/definition/4335/putty consulté le 22-09-2022.

[94].https://www.planetoscope.com/developpement-durable/Internet-consulté le 24-09-2022

[98].https://www.xenonstack.com/blog/big-data-security consulté le 24-09-2022

[99].https://www.michael-noll.com/blog/2011/04/09/benchmarking-and-stress-testing-an-hadoop-cluster-with-terasort-testdfsio-nnbench-mrbench/ consulté le 15-09-2022

Bibliographie :

[11]. Sagiroglu, S. and D. Sinanc. Big data: A review. In 2013 international conference on collaboration technologies and systems (CTS). 2013. IEEE

[12].Manyika, J., et al., Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, 2011

[13]. Furht, B., Villanustre, F. (2016). Introduction to Big Data. In: Big Data Technologies and Applications. Springer, Cham. https://doi.org/10.1007/978-3-319-44550-2_1

[15]. Hofmann, E., Big data and supply chain decisions: the impact of volume,variety and velocity properties on the bullwhip effect. International Journal of Production Research, 2017

[16]. Rubin, V. and T. Lukoianova, Veracity roadmap: Is big data objective, truthful and credible? Advances in Classification Research Online, 2013.

[17]. Demchenko, Y., et al. Addressing big data issues in scientific data infrastructure. In Collaboration Technologies and Systems (CTS), 2013 International Conference on. 2013. IEEE

[18]. Gandomi, A. and M. Haider, Beyond the hype: big data concepts, methods, and analytics. International journal of information management, 2015.

[19]. Fan, W. and A. Bifet, Mining big data. ACM SIGKDD Explorations Newsletter, 2013. **14**(2): p. 1.

[20]. Jukić, N., et al., Augmenting data warehouses with big data. Information Systems Management, 2015. **32**(3): p. 200-209.

[21]. Kacfah Emani, C., N. Cullot, and C. Nicolle, Understandable Big Data: A survey. Computer Science Review, 2015.

[22]. Wasser, T., et al., Using ‘big data’to validate claims made in the pharmaceutical approval process. Journal of medical economics, 2015. 18(12): p. 1013-1019.

[23]. Uddin, M.F. and N. Gupta. Seven V’s of Big Data understanding Big Data to extract value. In Proceedings of the 2014 zone 1 conference of the American Society for Engineering Education. 2014. IEEE

[24]. Uddin, M.F. and N. Gupta. Seven V’s of Big Data understanding Big Data to extract value. In Proceedings of the 2014 zone 1 conference of the American Society for Engineering Education. 2014. IEEE

[25]. Hackenberger, B.K., Data by data, Big Data. Croatian medical journal, 2019. **60**(3): p. 290

[26].Asokan, G. and V. Asokan, Leveraging “big data” to enhance the effectiveness of “one health” in an era of health informatics. Journal of epidemiology and global health, 2015. **5**(4): p. 311-314.

[27]. Uddin, M.F. and N. Gupta. Seven Vs of Big Data understanding Big Data to extract value. In Proceedings of the 2014 zone 1 conference of the American Society for Engineering Education. 2014. IEEE.

[28]. Sun, G., F. Li, and W. Jiang, Brief Talk About Big Data Graph Analysis and Visualization. 2019.

[29]. Elgendy, N.and A.Elragal. Big data analytics: a literature review paper. in Industrial Conference on Data Mining. 2014. Springer.

[30]. Demchenko, Y., et al. Addressing big data issues in scientific data infrastructure. In Collaboration Technologies and Systems (CTS), 2013 International Conference on. 2013. IEEE

[31]. Bilal Abu-Salih, Pornpit Wongthongtham,Dengya Zhu , Kit Yan Chan , Amit Rudra,”Chapter 2 Introduction to Big data Technology”,1-46.

[34]. kassimi dounya, Une approche de sécurité Big Data dans le Cloud Computing, thèse de doctorat, Université Mohamed Khider Biskra, 2019-2020

[35]. H. Saouli, O. Kazar, D. Kassimi : Applications et enjeux des Big Data dans le contexte des défis mondiaux. Proceedings 10th of Les Avancées des Systèmes Décisionnels (ASD), Annaba, Algérie (2016) 14-16 May

[36]. Lee, I. "Big data: Dimensions, evolution, impacts, and challenges." Business Horizons, 60(3): 293-303, (2017).

[37]. Gurram Bhaskar & Motati Dinesh Reddy, Analysis of Big Data Challenges and Different Analytical Methods, International Journal of Engineering Research and Advanced Technology (IJERAT), Volume 7, Issue 3, March 2021

[38]. Khan, Z. and Vorley, T., 2017. "Big data text analytics: an enabler of knowledge management" Journal of Knowledge Management

[39]. Krippendorff, K., "Content analysis: An introduction to its methodology." Sage publication, (2017).

[41]. Medfouni Hayet, Validation de clustering des données dans un contexte Big Data, Mémoire de Master, Université Larbi Ben Mhidi, 2017-2018.

[43]. Merri Nassim, Bouhaoui Walid, Mise en place d'un cluster hadoop automatisé sur Docker, Mémoire de Master, Université de Abderrahmane Mira Bejaia, 2019-2020.

[45]. Katal, A., Wazid, M., & Goudar, R. H. (2013). Big Data: Issues, Challenges, Tools and Good Practices. IEEE, 404-409.

[46]. Sharma, 2020, "Augmenting Data Warehouses with Big Data", Information Systems Management Volume 32, Issue 3: Business Intelligence.

[47]. Alex Bekker, "The Scary Seven big data challenges", Science soft, 2017.

[48]. L'heureux, A., Grolinger, K., Elyamany, H. F., and Capretz, M. A. (2017), "Machine learning with big data: Challenges and approaches". IEEE Access, 5 : 7776-7797.

[55]. Chabani Rabah, Implémentation d'un Protocole d'Élection d'un Serveur d'Authentification dans l'Internet des Objets. Mémoire de master, Université Mohamed SeddikBenyahiaJijel. 2021.

[60]. Mohmmad A. Alia, AbdelfatahArefTamimi, and Omaima N. A. AL-Allaf, <<Cryptography Based Authentication Methods>>, Proceedings of the World Congress on Engineering and Computer Science, Volume 1, WCECS 2014, 22-24 October, 2014.

[64]. Chabani Rabah, Implémentation d'un Protocole d'Élection d'un Serveur d'Authentification dans l'Internet des Objets. Mémoire de master, Université Mohamed SeddikBenyahiaJijel. 2021.

[65]. HelalSonya, Authentification Anonyme et Contrôle d'Accès dans un Environnement Cloud : Application au Domaine e-santé. Mémoire de Master, Université Saad Dahlab Blida. 2019.

[66]. Medjri Kahina, HaddoucheSoumia, Mise en place d'un système d'authentification pour Hadoop. Mémoire de Master, UniversitéAkliMohandOulhadj-Bouira. 2019.

[67]. Balaraju.J., PVRD. Prasada Rao, <<Designing Authentication for Hadoop Cluster using DNA Algorithm.>>, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-3, September 2019

[68]. PengShen, Xiaoming Ding, WenjunRen, <<Research on Kerberos Technology Based on Hadoop Cluster Security >>, Advances in Engineering Research, volume 155, 228-233, 2018.

[69]. Sufyan T. Faraj Al-Janabi and Mayada Abdul-salam Rasheed, <<Public-Key Cryptography Enabled Kerberos Authentication>>, 2011 Developments in E-Systems Engineering.

[70]. R. AOUDJIT, Mise en place d'un système de sécurité basé sur l'authentification dans un réseau IP, Mémoire de Master, UNIVERSITE MOULOUD MAMMERI DE TIZI-OUZOU, 2015.

[71]. Chandni Grover, Manpreet Kaur Aulakh, <<Big Data Authentication and Authorization in HDP (Hadoop Distributed platform) using Kerberos and Ranger>>, IIMT College of Engineering, Greater Noida, ISBN: 978-93-86171-50-4, 2017

[72]. Jolly Khurana, Dr. Kamlesh Sharma, <<A Secured Method for accessing HDFS in Hadoop>>,International Journal for Research in Applied Science & Engineering Technology (IJRASET**),** Volume 8 Issue VII July 2020.

[73]. Daming Hu, Deyun Chen, Yuanxu Zhang and Shujun Pei, <<Research on Hadoop Identity Authentication Based on Improved Kerberos Protocol>>, International Journal of Security and Its Applications, Vol.9, No.11 (2015), pp.429-438

[74]. Pushkar Bhadle1, Sonal Gugale2, Sakshi Trar3, Harjot Kaur4, Shital Salve, <<Kerberos Authentication System using Public key Encryption >>, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1930-1933

[75]. Chandni Grover, Manpreet Kaur Aulakh, <<Big Data Authentication and Authorization in HDP (Hadoop Distributed platform) using Kerberos and Ranger>>, IIMT College of Engineering, Greater Noida, ISBN: 978-93-86171-50-4, 2017

[76].N.Sirisha, K.V.D.Kiran,<<Authorization of Data In Hadoop Using Apache Sentry>>,International Journal of Engineering & Technology, 7 (3.6) (2018) 234-236.

[77]. Ravi Raju, Yallapragada and Donavalli, Haritha (2020). Data Authorization in Hadoop using Kerberos Authentication System and Transport Layer Security. International Journal on Emerging Technologies, 11(1): 403–408.

[78]. Prof. S.A.Darade, KiranKamble,GauraviKhanapure,SnehaChavan ,KomalKumbhar, <<Network Level Security in Hadoop using wire encryption>>, International Journal of

Advanced Research in Science Management and Technology, Volume1, Issue 6, November 2015

[87]. Benaoudasid Ahmed Amine,Implantation Du modèle mapreduce Dans L'environnementDistribué Hadoop : Distribution Cloudera, mémoire de Master, Université de Abou Bakr Belkaïd-Tlemcen,2015.

[91]. Ahmad, S., Yasin, A., & Shafi, Q. (2018). DDoS attacks analysis in bigdata (Hadoop)environment. 2018 15th International Bhurban Conference on Applied Sciences andTechnology(IBCAST).doi:10.1109/ibcast.2018.8312270

[95]. BOUMRAOU KAHINA, KEDJAR HAKIM, Mise en place d'un Cluster Hadoop de dix (10) postes avec interface d'exécution de Jobs MapReduce à l'EcoleNationale Supérieure en Science et Technologie de l'Informatique (ENSTI), Université de Bejaïa, Mémoire de Master, 2020

[96]. Saminath.V, Sangeetha.M.S,<<Internals of Hadoop Application Framework andDistributed File System>>, International Journal of Scientific and Research Publications, Volume 5, Issue 7, July 2015

[97].K. Zheng and W. Jiang, "A token authentication solution for Hadoop based on Kerberos pre-authentication," 2014 International Conference on Data Science and Advanced Analytics (DSAA), 2014, pp. 354-360, DOI: 10.1109/DSAA.2014.7058096.