REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mohamed Khider – BISKRA
Faculté des Sciences Exactes, des Sciences de la Nature et de la Vie
Département d'informatique

N° d'ordre : SIOD 09 /M2/2018

Mémoire

Présenté pour obtenir le diplôme de master académique en

Informatique

Parcours : System Informatique Optimisation et Décisionnelle

Réalisation d'une architecture pour l'assurance de la disponibilité des Big Data.

Par :

DJERBOUA HAMZA

Soutenu le 24 Juin 2018, devant le jury composé de :

Bendahmane Toufik	M A A	Président
SAOULI Hamza	M C B	Rapporteur
Bouguetitiche Amina	M A A	Examinateur

Remerciement

Grand merci à Allah, le tout puissant qui m'a donné la force et le courage d'arriver jusque-là.

Je tiens à remercier profondément mon encadreur : Monsieur SAOULI Hamza qui m'a encouragé à faire le maximum d'efforts dans ce travail ainsi que pour sa disponibilité.

Je remercie par la même occasion les membres de jury qui ont bien voulu accepter d'examiner et évaluer mon travail et participer à ma soutenance.

Je remercie aussi ma famille et tous les enseignants du département d'informatique pour leurs efforts afin de nous assurer la meilleure formation possible durant notre cycle d'étude.

Enfin, je remercie tous ceux qui ont contribué de près ou de loin à l'aboutissement de ce travail.

Dédicace

Louange a Allah de m'avoir donné la patience d'aller jusqu'au bout de mes rêves et le bonheur de lever mes mains vers le ciel et de dire

"Al Hamdoulillah"

Je dédie ce modeste travail à ceux qui mon aider dans ma vie, aux personnes généreuses, vertueuses qui m'ont été d'un grand secours, chacun a sa manière, qui ont sacrifiés leur temps pour mon bonheur et pour ma réussite, depuis l'école de mon enfance, et grâce a le éducation, mes parents, qui mon permis d'arriver à la fin de mon cycle d'étude

A l'esprit de mon père, et ainsi que ma mère qu'ils seront fière de moi.

Qu Allah les garde et les protèges ainsi que toute ma famille

Mes frères et mes sœurs,

A mes oncles, mes tantes, et leurs familles

A mes amis de la promotion

A tous ceux qui me sont chers.

Je dédie ce travail également a toutes les personnes chers a mon cœurs .Qu'elles trouvent ici l'expression de toute ma gratitude et mon amour.

Résumé

Big data est le résultat de la révolution de la technologie de l'information. Il a une grande influence sur le présent et l'avenir grâce à son large spectre d'application dans le domaine de la technologie de l'information et de la communication, la recherche scientifique et l'économie.

Il faut permettre l'accessibilité à l'information en continu sans interruption ni dégradation et avec un temps de réponse acceptable. Si le nom d'un fichier est supprimé du répertoire des fichiers qui le contient, les informations contenues dans les fichiers ne seront plus accessibles à l'ordinateur, donc inexploitables. Il est à signaler que les cinq éléments de sécurité restant (confidentialité, intégrité, utilité, authentification et possession) ne pouvaient pas faire face à ce problème. Par conséquent la préservation de la disponibilité doit être considérée comme un objectif vital. Toute perte d'information, qu'elle soit temporelle ou définitive constitue un problème de disponibilité bien que les dégâts causés ne soient pas de même proportions.

Mots clés : Big data , Disponibilité, Cloud, Hadoop , Sécurité, Travaux connexes.

Table matière

Remerciement……..I

Dédicace……II

Résumé……..III

Table des matières……IV

Liste des figures……..…XI I

Introduction générale …….....1

Chapitre I : Sécurité des Big Data

I.1. Introduction…….3

I.2. BigData……3

I.2.1. Définition…….3

I.2.2. Modèle 5V……...…4

I.2.2.1. Volume……..…..…4

I.2.2.2. Vitesse……..4

I.2.2.3. Variété……..4

I.2.2.4. Valeur ……..4

I.2.2.5. Véracité……....4

I.2.3. Méthode de traitement des BigData ……..5

I.2.3.1. Data Mining ……..5

I.2.3.2. Réseaux de neurones ……..6

I.2.3.3. Machine Learning……6

I.2.3.4. Traitement de signal ……..6

I.2.3.5. Méthode de visualisation…….6

I.2.4. Qualité des BigData…….6

I.2.4.1. Cohérence contextuelle……6

I.2.4.2. Cohérence temporelle……..7

I.2.4.3. Cohérence opérationnelle……7

I.2.5. Framework d'implémentation…….7

I.2.5.1. Capture Data……7

I.2.5.2. Organiser Data……7

I.2.5.3. Analyzer Data……..7
I.2.5.4. Les valeurs et les décisions d'affaires……..8
I.2.6. Domaine d'application de Big data……8
I.2.6.1. Agriculture……..……8
I.2.6.2. Assurance……8
I.2.6.3. Marketing……8
I.2.6.4. Au-delà du marketing……..9
I.2.6.5. Achat programmatique……9
I.2.6.6. Compétitivité et Innovation de produit……..10
I.2.6.7. Gestion de catastrophes naturelles……10
I.2.6.8. Contrôle d'épidémies……10
I.2.6.9. Prévention d'attaques cybernétiques……10
I.2.7 Défis et Enjeux……..11
I.3. Sécurité et vie privé……...11
I.3.1. Sécurité : survole général……..11
I.3.2. Les six éléments de sécurité……..12
I.3.3. Disponibilité……..12
I.3.4. Utilitaire……12
I.3.5. Intégrité……..12
I.3.6. Authenticité……...12
I.3.7. La Confidentialité……..13
I.3.8. La Possession……13
I.3.9. Sécurisation : SI Via BigData……13
I.3.10. Types de vie privée……13
I.3.10.1. Sécurité physique……...13
I.3.10.2. Sécurité politique……...13
I.3.10.3. Sécurité socio-économique……14
I.3.10.4. Sécurité culturelle……..14
I.3.10.5. Sécurité environnementale……14
I.3.10.6. Sécurité d'incertitude……..14

I.3.10.7. Sécurité de l'information……14
I.3.11. Défis de protection et sécurisation……14
I.3.12. Vie privé : survole général……15
I.3.13. Types de vie privée……15
I.3.13.1. La vie privée de la personne :……..…… …15
I.3.13.2. Comportement de la vie privée et travail……...……15
I.3.13.3. la vie privée communicatio……..……..16
I.3.13.4. Vie privée et image des données….…….……..16
I.3.13.5. la vie privée des pensées et des sentiments……16
I.3.13.6. la vie privée de l'emplacement et de l'espace…….16
I.3.13.7. la vie privée de l'association……..16
I.3.14. Défis et enjeux de la vie privée…….16
I.3.15. Sécurité via la vie privée……..17
I.3.16. Cryptage de données…….18
I.3.16.1. Cryptage recherché……18
I.3.16.2. Commander-Préserver le cryptage…….18
I.3.16.3. Cryptage structuré……..18
I.3.16.4. Cryptage homomorphie…….18
I.3.17. Gestion de la confiance…….19
I.3.18. Définition et Types de vulnérabilité…….19
I.3.18.1. Vulnérabilités de logiciels…….19
I.3.18.2. Vulnérabilités du personnel……...20
I.3.18.3. Vulnérabilités de planification de récupération après sinistre...20
I.3.18.4. Vulnérabilités du protocole réseau…….20
I.3.18.5. Infrastructure critique et BigData……..20
I.3.19. Contrôle de sécurité et protection d'infrastructure……21
I.3.19.1. Contrôles préventifs……...21
I.3.19.2. Contrôle des détectives……..21
I.3.20. Comment sécuriser les BigData……21
I.3.20.1. Gestion des données……..21

I.3.20.2. Gestion de l'identité et de l'accès……...22
I.3.20.3. Protection des données et confidentialité……...22
I.3.20.4. Sécurité réseau……...22
I.3.20.5. Sécurité et intégrité de l'infrastructure……...22
I.4. Conclusion……22

Chapitre II: Approches

II.1. Introduction……...……23
II.2. Problème est recommandation lies à la disponibilité……...……23
II.2.1. Limites d'utilisation des données……...……23
II.2.2. Données non existantes ou manquantes……..…..23
II.2.3. Calendrier et synchronisation des données……...24
II.2.4. Données de fréquence……24
II.2.5. Formats de données, normes et spécifications……..24
II.2.6. Incertitude et fiabilité des données……24
II.2.7. Accès, stockage et traitement des données……24
II.2.8. Temps de préparation des données et autres ressources……25
II.3. Disponibilité dans les BigData……..25
II.3.1. Base de données NoSQL……...25
II.3.1.1. Les inconvénients……..26
II.3.2. Plateforme géographique……..26
II.3.2.1. Le rôle le cloud computing dans la plateforme géographiq …26
II.3.2.2. Base de données distribuée……26
II.3.2.3. Système existant CityServer3D…….27
II.3.2.4. Disponibilité du système……27
II.3.3. Gestion de la tolérance aux fautes…….27
II.3.3.1. Méthodes gestion de la tolérance aux fautes……..27
II.3.3.2. Inconvénients 28
II.3.4. Assurer la disponibilité des données.....28
II.3.4.1. Performance dans la disponibilité de Big data.....29
II.3.4.2. Télécharger les données sur l'appareil.....29

II.3.4.3. Les données sont situées dans plusieurs endroits.....29
II.3.4.4. Les méthodes d'accès aux données.....29
II.3.5. Oracle au service de la disponibilité.....30
II.3.5.1. Architecture oracle Big data.....30
II.3.6. La disponibilité dans le Cloud.....31
II.3.7. Approche multi-Cloud.....31
II.3.7.1. Système de partage secret.....31
II.3.7.2. Environnement supposé pour l'utilisation de la technique proposée.....31
II.3.7.3. Environnement supposé de la technique proposée.....31
II.3.7.4. L'utilisateur demande des statistiques et un SLA Cloud.....31
II.3.7.5. Inconvénients.....32
II.3.8. Approche basé prix/réplication de données.....32
II.3.8.1. Hadoop Distributed File System.....32
II.3.8.2. Modèles d'avantages financiers.....32
II.3.8.3. Modèles de coûts d'exploitation.....32
II.3.8.4. Formulation du problème.....33
II.3.9. Gestion de désastre.....33
II.3.9.1. Phase configuration du compte.....33
II.3.9.2. Phase chargement des données.....33
II.3.9.3. phase de téléchargement des données.....34
II.3.9.4. Les inconvénients.....34
II.3.10. Intégrité et disponibilité basé réplication.....34
II.3.10.1. Structure de cloud hiérarchique.....34
II.3.10.2. Construction de MRVR.....34
II.4. Disponibilité hors BigData.....35
II.4.1. Approche évolutive.....35
II.4.1.1. Puissance opérationnelle.....35
II.4.1.2. Détermination de la réplication.....36
II.4.2. Modèle comparaison 36
II.5. Conclusion.....37

Chapitre III: Conception et modélisation

III.1. Introduction.....38
III.2. Conception générale du système proposé.....38
III.2.1. Architecture globale.....38
III.2.2. Architecture détaillée 39
III.2.2.1. Composant gestion base de données 39
2.2.1.1. L'architecture composant gestion base de données.....39
2.2.1.2. les rôles composant gestion base de données.....39
III.2.2.2. Composant réplication des données.....40
2.2.2.1. l'architecture composant réplication.....40
2.2.2.2. les rôles composant réplication.....40
III.2.2.3. Composant Backup.....41
2.2.3.1. L'architecture composant Backup.....41
2.2.3.2. Les rôles composant Backup.....41
III.2.2.4. Composant cloud.....42
2.2.4.1. L'architecture composant cloud.....42
2.2.4.2. Les rôle composant cloud.....42
III.2.2.5. Composant supervision.....42
2.2.5.1. L'architecture composant supervision.....42
2.2.5.2. les rôles composant supervisé.....43
III.3. Projection sur Hadoop.....44
III.3.1. NameNode.....44
III.3.2. Secondary Name Node.....45
III.3.3. DataNode.....45
III.3.4. JobTracker.....45
III.3.5. TaskTracker.....45
III.4. Conception et Modélisation détaillée avec UML.....46
III.4.1. Diagramme de séquence général.....46
III.4.2. Diagramme d'activité des composants.....47
III.4.2.1. Diagramme d'activité de composant gestion base de données.....47

III.4.2.2. Diagramme d'activité de composant réplication les donnée.....47
III.4.2.3. Diagramme d'activité de composant Backup.....48
III.4.2.4. Diagramme d'activité de composant 49
III.4.2.5. Diagramme d'activité de composant supervision.....49
III.5. Conclusion.....50

Chapitre IV: Implémentation

IV.1. Introduction.....51
IV.2. Outils et langages de programmation utilisés.....51
IV.2.1. Langages de programmation 51
2.1.1. Java 51
IV.2.2. Outils de développement.....52
2.2.1. Netbeanse.....52
2.2.2. Hadoop.....53
2.2.3. XAMPP.....54
2.2.4. MySQL.....54
2.2.5. phpMyAdmin 55
IV.3. Description des Interfaces Graphiques.....55
IV.3.1. interface accès a system.....55
IV.3.2. Interface d'authentification.....56
IV.3.3. Interface User 57
IV.3.4. Interface Ajouter des informations patientes.....57
IV.3.5. Interface modifier information patient.....58
IV.3.6. Interface supprimé patient.....58
IV.3.7. Interface de l'administrateur.....59
IV.3.8. Interface Backup.....59
IV.3.9. Interface réplication les données.....60
IV.3.10. Interface supervision 61
IV.3.11. Interface cloud service.....62
IV.3.12. L'interface de la base de données.....62
3.12.1. Tableau d'information patient.....62

3.12.2. Tableau fournisseurs.....63
IV.4. Les principaux codes source.....64
IV.4.1. Connection avec base de données.....64
IV.4.2. Fonction de rechercher.....64
IV.4.3. Code exporte dans fichier.....65
IV.4.4. Code copier fichier.....66
IV.4.5. Code Import fichier.....66
IV.5. Conclusion.....67
Conclusion générale 68
Bibliographie…….69

Liste des figures

Chapitre I : Sécurité des Big Data

Figure I.1 : Modèle 5V……...……5

Chapitre III: Conception et modélisation

Figure III.1 : L'architecture globale du système proposé……38

Figure III.2 : L'architecture composant gestion base de données……...39

Figure III.3 : L'architecture composant réplication de données.……40

Figure III.4 : L'architecture composant Backup…….41

Figure III.5 : L'architecture composant cloud…….42

Figure III.6 : L'architecture composant supervisé……..43

Figure III.7 : Rôles de serveur Hadoop……...44

Figure III.8 : Diagramme de séquence général……...46

Figure III.9 : Diagramme d'activité de composant gestion base de données……...47

Figure III.10 : Diagramme d'activité de composant réplication des données……..48

Figure III.10 : Diagramme d'activité de composant Backup……..48

Figure III.10 : Diagramme d'activité de composant cloud…….49

Figure III.10 : Diagramme d'activité de composant cloud…….50

Chapitre IV: Conception et modélisation

Figure IV.1 : Logo de java……...52

Figure IV. 2 : Logo de netbeanse…….53

Figure IV.3 : Logo d'hadoop……...53

Figure IV.4 : Logo de XAMPP……54

Figure IV.5 : Logo de mysql……54

Figure IV.6 : Logo de Phpmyadmin……55

Figure. VI.7 Interface accès a system……...55

Figure. VI.8: Interface d'authentification……56

Figure. VI.11 Interface User……57

Figure. VI.12 Interface ajouté patient……..57

Figure. VI.13 Interface modifié patient……...58

Figure. VI.13 Interface supprimé patient……58

Figure. VI.14 Interface de l'administrateur……59

Figure. VI.15 Interface Backup……60

Figure. VI.16 Interface Backup……61

Figure. VI.17 Interface supervision……61

Figure. VI.18 Interface supervision……62

Figure. VI.19 Tableau d'information patient……...63

Figure. VI.20 Tableau fournisseurs……..63

Introduction générale

Introduction générale

Ces mégadonnées sont maintenant au centre des préoccupations des acteurs de tous les domaines d'activité. Ainsi le taux de croissance annuel moyen mondial du marché de la technologie et des services autour du Big Data sur la période 2011-2016 est estimé à plus de 30 %. D'après une étude IDC de 2013, ce marché devrait ainsi atteindre 23,8 milliards de dollars en 2016. Sur le plan européen, l'activité autour des mégadonnées devrait représenter autour de 8 % du PIB européen en 2020 (AFDEL février 2013). D'après le cabinet Markess International, le marché français des solutions et services en analytique, big data et gestion des données aurait atteint 1,9 milliard d'euros en 2015. Son taux de croissance annuel moyen d'ici 2018 est attendu à plus de 12 % (d'après Le monde informatique du 15 mars 2016).

La définition initiale de big data s'orientait d'abord vers la question technologique, avec la célèbre règle des 3V : un grand Volume de données, une importante Variété de ces mêmes données et une Vitesse de traitement s'apparentant parfois au temps réel. Ces technologies étaient censées répondre à l'explosion des données dans le paysage numérique (le « data deluge »). Puis, ces qualificatifs ont évolué, avec une vision davantage économique portée par le 4ème V de la définition, celui de Valeur, et une notion qualitative véhiculée par le 5e V, celui de Véracité des données (disposer de données fiables pour le traitement).

L'augmentation rapide des données, il est plus difficile de traiter des données, ce qui crée plusieurs défis pour faire l'accord avec les données facilement. et il peut être divisé en ces défis en trois branches: données traitement, défis est les traitements et défis de gestion.

Tout en traitant de grandes quantités d'informations face aux défis tels que 3V (Volume, Vitesse, Variété). la multiplicité des différentes données résulte de la problématique de sa gestion.

La question de la disponibilité des données est un défi majeur et comporte de nombreuses complexités. Cela inclut la complexité, le type, l'emplacement et la fréquence de la source de données. La difficulté est lorsque les sources ne sont pas synchronisées ou utilisent des formats de sortie différents.

L'augmentation significative du nombre d'utilisateurs de sites Web modernes et de grandes entreprises fait face à de nouveaux défis. La technique le système de gestion de base de données relationnelle (SGBDR) ne peut pas atteindre les nouvelles exigences, le défi est passé à la règle relationnelle.

La base de données NoSQL(données non relationnelles) a été créée pour traiter les entrepôt de ou données les Big data . C'est une classe très large dans les systèmes de gestion de base de données, qui ne suit pas les données relationnelles

Malgré l'évolution rapide et continue des outils et plateformes big data, les mesures de sécurité adéquates nécessitent un effort supplémentaire car ils ne répondent pas aux conditions exigées. Actuellement, la plateforme Hadoop est largement utilisée par l'industrie et la recherche scientifique toutefois il convient de signaler qu'Hadoop est une technologie open source et la sécurité ne constitue pas une priorité. La sécurité du système de gestion des bases de données relationnelles traditionnelles (SGBDR) a été améliorée à plusieurs reprises au cours de plusieurs années. Par contre Hadoop fait ses premiers pas en matière de sécurité, il reste beaucoup à faire pour combler les défaillances sécuritaires.

Ce projet est structuré de la façon suivante :

Dans le chapitre I on va présenter des notions générales et d'autres informations ayant rapport à la sécurité en général, au big Data en tant que nouvelle technologie et enfin aux exigences sécuritaires imposées par cette dernière.

Dans le chapitre II on va étudier des travaux de recherche qui ont proposé des solutions pour assurer l'intégrité dans le big Data. On va résumer les méthodes proposées par différents chercheurs en la matière et on va tenter de mettre en lumière les inconvénients afin de présenter plus tard des solutions susceptibles de combler les défaillances observées. Les solutions qui nous aident à construire notre architecture pour réalisation d'une architecture pour l'assurance de la disponibilité des Big Data.

Dans le chapitre III on va présenter la conception détaillée, dans laquelle nous avons fixé la structure globale de l'application. On va proposer un ensemble de composants qui représente disponibilité système Bigdata .

Le dernier chapitre de notre projet est la partie réalisation qui a été consacrée à la présentation des outils du travail et les interfaces les plus significatives de notre application.

Chapitre I

Sécurité des Big data

I.1. Introduction

Les données sont devenues partie intégrante de l'histoire, de la politique, de la science, de l'économie et des structures commerciales, et maintenant même de la vie sociale. Cette tendance est clairement visible dans les réseaux sociaux tels que Facebook, Twitter et Instagram où les utilisateurs produisent quotidiennement un énorme flux de différents types d'informations (musique, images, texte, etc.).Les technologies BigData et Data Intensive deviennent une nouvelle tendance technologique dans la science, l'industrie et les affaires.

Les BigData sont en train de devenir liés avec tous les aspects de l'activité humaine, depuis l'enregistrement des événements jusqu'à la recherche, la conception, la production et les services numériques ou la livraison des produits, jusqu'à le consommateur final.

L'utilisation des BigData et leur disponibilité dans des nombreuses tâches sensibles liées à la confidentialité, cette dernière fait partie de la sécurité des données et de la confidentialité une exigence de plus en plus critique. Le fait que les données doivent être partagées et rendues disponibles, éventuellement en temps réel, à une grande variété d'utilisateurs et d'applications compliqué davantage le problème de la protection des données.

I.2. BigData

I.2.1. Définition

Chaque jour, nous gérons 2,5 trillions d'octets des données, A tel point que 90% des données dans le monde ont été créées au cours des deux dernières années seulement.

Les BigData ou méga données désignent l'ensemble des données numériques produites par l'utilisation des nouvelles technologies à des fins personnelles ou professionnelles. Ces données proviennent de partout (de capteurs utilisés pour collecter les informations climatiques, de messages sur les sites de médias sociaux, d'images numériques et de vidéos publiées en ligne, d'enregistrements transactionnels d'achats en ligne et de signaux GPS de téléphones mobiles, pour ne citer que quelques sources). Ces données sont appelées BigData ou volumes massifs de données

I.2.2. Modèle 5V

Pour comprendre le phénomène qu'est la BigData est souvent décrit par les 5V.à prendre en compte et à optimiser dans le cadre d'une démarche d'optimisation de la gestion la BigData.

Il faut qu'il réponde à trois principaux critères 3V: Volume, Vitesse, et Variété, Mais il est courant, d'ajouter deux autres critères pour compléter la typologie 3V, à savoir : Valeur et Véracité.

I.2.2.1. Volume

De données crées et gérées par les utilisateurs sont en constante augmentation. Grande taille devient de plus en plus chaque jours, et il augmente autour de 40% chaque année.

I.2.2.2. Vitesse

Une des grandes forces du BigData, c'est de pouvoir utiliser les données à mesure qu'elles sont collectées. La collecte et l'analyse des données doivent être effectuées rapidement.

I.2.2.3. Variété

Indique les différents types des données, les données structurées ou non structurées (voix, données faciales, données transactionnelles, web analytiques, textes, images, etc.)

I.2.2.4. Valeur

Lorsque la collecte et l'analyse des données, il devient de précieuses données, et peuvent être exploitées Pour aider à la prise de décision, en outre il peut être vendu sociétés commerciales.

I.2.2.5. Véracité

Sont difficiles à traiter avec de grandes données et la vérification de l'exactitude des données en éliminant le bruit. Par des moyens d'organiser les données, Ceci est d'assurer la qualité des données d'analyse et de décision dans l'étude. [5]

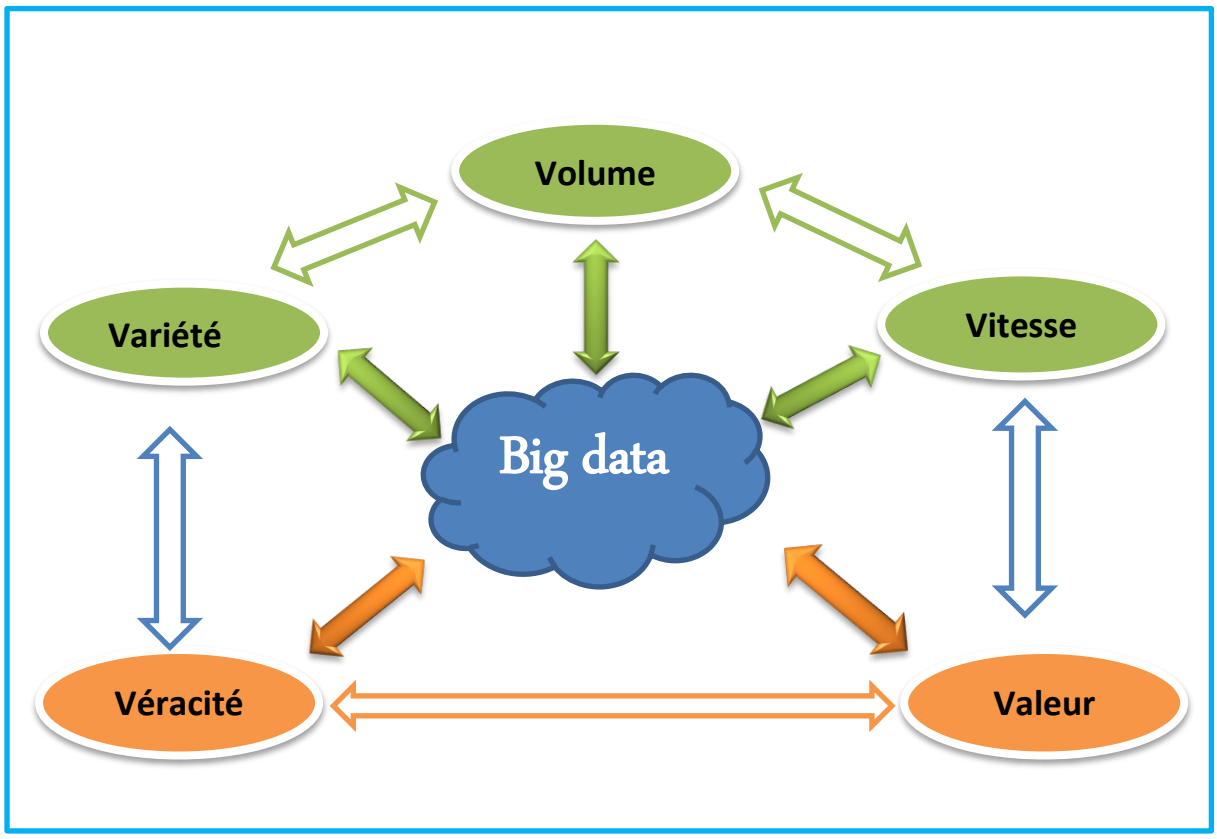

Figure I.1 : Modèle 5V

I.2.3. Méthode de traitement des BigData

Pour l'analyse des BigData , il existe plusieurs techniques différentes pour l'analyse des données. utilisée dans les statistiques et de l'informatique. les techniques les plus avancées pour l'analyse. tels que: réseaux de neurones artificiels , méthodes d'analyse prédictive , statistiques , Traitement du langage naturel…ect.

Les méthode de traitement des BigData diverses disciplines Les mathématiques appliquées, les statistiques, l'informatique et l'économie. Ce sont les bases de techniques d'analyse de données telles que: Data Mining , les Réseaux de neurones , Machine Learning , Traitement de signal et Méthodes de visualisation.

I.2.3.1. Data Mining

Cette méthode pour identifier et extraire des informations utiles à partir des données ou des ensembles de données à grande échelle l'utilisation : classification, régression et association Réglez les techniques d'apprentissage.

I.2.3.2. Réseaux de neurones

est un ensemble d'algorithmes dont la conception , et optimisés par des méthodes d'apprentissage de type probabiliste .

I.2.3.3. Machine Learning

Fait partie de l'une des approches de l'intelligence artificielle, et est donc une discipline scientifique centrée sur le développement, l'analyse et l'implémentation de méthodes automatisables, qui offrent la possibilité à une machine d'évoluer grâce a un processus d'apprentissage.

I.2.3.4. Traitement de signal

Cette méthode est l'analyse des signaux discrets et continus. Et étudie les techniques de traitement, d'analyse et d'interprétation des signaux. En d'autres termes, il permet la représentation analogique des grandeurs physiques (par exemple, des signaux radio ou des sons, etc.).

I.2.3.5. Méthode de visualisation

Est un ensemble de méthodes est une technique d'exploration et d'analyse des données numériques à l'aide de graphiques. sont à même de lire et de comprendre sans ambiguïté le sens porté par le graphique .Ces processus sont nécessaires pour diminuer la taille des données avant le rendu réel [1].

I.2.4. Qualité des BigData

Un modèle de qualité d'utilisation pour les données volumineuses les principales caractéristiques de qualité de BigData pour évaluer le niveau de qualité de l'utilisation des ensembles de données hétérogènes pour les projets de BigData sont la cohérence.

I.2.4.1. Cohérence contextuelle

Re réfère à la capacité des ensembles de données à utiliser dans le même domaine d'intérêt du problème indépendamment de tout format de toute taille, ou venant à des vitesses différentes.

I.2.4.2. Cohérence temporelle

Fait référence au fait que l'ensemble de données est produit par un générateur de données, utilisé pour effectuer une analyse, et compris dans un intervalle de temps cohérent.

I.2.4.3. Cohérence opérationnelle

se réfère à la mesure dans laquelle l'ensemble de données peut être inclus dans la même analyse, d'un point de vue technologique[3].

I.2.5. Framework d'implémentation

Il a été étudié diverses approches de solutions Big Data mises en place jusqu'ici Pour diverses organisations. afin de mettre Framework global pour Solutions Big Data ce qui est utile pour tous les utilisateurs Big data. et il comprend en quatre étapes pour d'un traitement de BigData[3].

I.2.5.1. Capture Data

Cette phase capture les données Semi-structuré ou non structuré à savoir Créés ou recueillis. Selon une variété de données divers outils comme relationnel, data base Management System , NoSQL et Hadoop.

I.2.5.2. Organiser Data

Dans la phase Organiser data , et divisé en bases de données relationnelles ou Datawarehouse , pour éliminer les redondances . Il existe différents outils ou logiciels pouvant être utilisés dans cette phase Comme Hive , Cloudera Distribution avec Hadoop Apache , Big Data connectors et Intégrateurs.

I.2.5.3. Analyzer Data

L'objectif principal de cette phase est d'utiliser divers outils d'analyse et d'intelligence d'affaires (BI) pour extraire la valeur des données organisées et raffinées , Pour une analyse plus approfondie, à la recherche et les problèmes de performance et d'améliorer les valeurs de l'entreprise.

I.2.5.4. Les valeurs et les décisions d'affaires

Il peut prédire les résultats, basée sur l'analyse des données et les décisions commerciales peuvent être prises [20].

I.2.6. Domaine d'application de Big data

Dans cette section nous présentons quelques principaux domaines d'applications du Big data :

I.2.6.1. Agriculture

D'ici 2050 on prévoie le dépassement de 9 milliards d'êtres humains sur le globe, ce qui rend l'agriculture un domaine prioritaire pour gérer les besoins alimentaires de la population mondiale. Le Big data représente un atout considérable pour l'organisation de l'agriculture à travers le monde, notamment pour la gestion de l'irrigation (l'eau potable étant une ressources de plus en plus rare), où nous avons besoin de gérer de gigantesque masses de données qui concernent les prédictions météorologique et la sécheresse du sol.

I.2.6.2. Assurance

L'assurance représente l'un des domaines directe d'application de Big data, vue qu'on est amené à effectuer des statistiques et des analyses sur les risques liés au comportement de millions d'individus.

La possibilité de récolter de gigantesques masses d'informations qui concernent la vie des individus permet de concevoir un modèle de vie pour chacun d'eux : hygiène de vie, conduite de voiture, amende, consommation électrique, relation professionnelle ….Etc. Ces modèles de vie permettent aux agences d'assurances d'améliorer leurs offres, d'optimiser leurs méthodes, et même de mener des enquêtes plus précises.

I.2.6.3. Marketing

Avec le marketing on est amené à gérer de gigantesques masses d'informations qui proviennent de divers sites et réseaux sociaux que des clients potentiels peuvent visiter. Mais ce qui révolutionne vraiment le marketing de nos jours c'est l'omniprésence de capteurs publics sur les centres commerciaux, métros, aéroports et universités, et qui sont destinés à capter le comportement des consommateurs, ce qu'il achète, ce à quoi ils s'intéressent, et les produits

qu'ils ne trouvent pas aux marchés, ce qui permet d'analyser et d'étudier leurs besoins en temps réel afin de produire des solutions et des méthodes de marketing plus efficaces.

L'utilisation des capteurs permet de capter des données de diverses formes : images de visages pour analyse émotionnelle, vidéos pour description comportementale, données textuelles pour décrire la nature des produits achetés, données numériques et statistiques. Cette diversité qui nécessite un traitement en temps réel ne peut être résolue qu'avec des méthodes de stockage et de traitement d'informations issues de Big data.

I.2.6.4. Au-delà du marketing

Le Big data a permis de refaçonner le monde du marketing en offrant les techniques et les stratégies nécessaires pour bénéficier des données que publient les consommateurs et fournisseurs en utilisant les : Réseau sociaux, applications mobiles, magasins, TV, catalogues, blog, presse, radios, etc. Sans la techniques Big data il sera tout simplement impossible de traiter les gigantesques masses d'informations que produisent ces moyens de publication. L'émergence de Big data a permis l'apparition de nouvelles notions telle que le pré-marketing et re-marketing qui représentant une nouvelle vision d'atteindre et de convaincre les consommateurs finaux.

I.2.6.5. Achat programmatique

L'achat programmatique est devenu la technique la plus utilisée pour l'achat/vente sur Internet, vue que cette technique permet d'utiliser un logiciel ou une plateforme intermédiaire entre les clients et les fournisseurs pour effectuer des opérations de : publicité, choix du meilleur prix, et payement électronique. L'achat programmatique permet d'alléger les tâches qui correspondent au processus d'achat/vente en s'occupant automatiquement du processus de négociation entre client et fournisseur ainsi que de toute opération manuelle traditionnellement demandée par le fournisseur. Cependant, l'achat programmatique impose la manipulation en temps réel de gigantesques masses d'informations qui sont échangées entre clients et fournisseurs en compétition pour trouver et acheter les meilleurs espaces publicitaires sur le Net. Les techniques de gestion des données issues du domaine Big data représentent un atout considérables et une alternative prometteuse pour la gestion des plateformes d'achat/vente intermédiaire.

I.2.6.6. Compétitivité et Innovation de produit

La possibilité de traiter de gigantesques masses d'informations en temps réel permet aux entreprises d'analyser les besoins de leurs clients afin de pouvoir optimiser et améliorer leurs propres produits et augmenter leur compétitivité sur le marché. C'est ainsi, que les services qu'offrent les fournisseurs de téléphonie mobile permettent aux touristiques de localiser, en temps réel, leurs clients habituels afin de leurs envoyer des offres d'excursions, les lieux et la nature des évènements touristiques, et les réductions hôtelières et les billets d'avion par exemple. Les techniques d'analyse en temps réel de gigantesques masses d'informations, issues de Big data, permettent également aux entreprises de contrôler et d'être à jours par rapport aux produits des entreprises concurrentes ce qui garantit l'innovation et la compétitivité des produits.

I.2.6.7. Gestion de catastrophes naturelles

L'une des applications les plus intéressantes de Big data, est la possibilité d'analyser des données météorologiques en temps réel, ce traitement permet de suivre et de visualiser le déplacement des ouragans et de prédire les endroits géographiques où ces derniers vont frapper. C'est ainsi que les gouvernements locaux et les organisations internationales d'assistance humanitaire peuvent préparer les ressources nécessaires (couverture, alimentations, médicaments) ainsi que les moyens de transport et d'intervention rapide pour aider la population en détresse.

I.2.6.8. Contrôle d'épidémies

Le Big data peut contribuer à contrôler la propagation d'épidémies à travers le monde en surveillant par exemple la migration des insectes porteurs de maladies à travers le globe. Le big data est également utilisé pour traquer la population des rats dans les grandes villes telles que New-York ou Chicago où la police locale utilise un système Big data pour la surveillance visuelle et l'analyse des itinéraires des rats, afin de contrôler leurs croissances.

I.2.6.9. Prévention d'attaques cybernétiques

De nos jours, les techniques d'analyse de données qu'offre le Big data sont devenues incontournables pour pouvoir détecter les intrusions, les failles sécuritaires ainsi que les attaques cybernétiques, vue que le volume de données transportées sur le Net est devenu gigantesque, diversifier, et nécessitant un traitement en temps réel. Avec les techniques de

traitement de données Big data on arrive à tracer le schéma relationnel entre les données et effectuer des calculs statistiques qui permettent de surveiller et d'intervenir, en temps réel, sur les menaces et les attaques cybernétiques à l'échelle mondiale. [19]

I.2.6.10. Défis et Enjeux

L'augmentation rapide des données, il est plus difficile de traiter des données, ce qui crée plusieurs défis pour faire l'accord avec les données facilement. et il peut être divisé en ces défis en trois branches: données traitement, défis est les traitements et défis de gestion.

Tout en traitant de grandes quantités d'informations face aux défis tels que 3V (Volume, Vitesse, Variété). la multiplicité des différentes données résulte de la problématique de sa gestion.

- Le volume fait référence à la grande quantité de données.

- La vitesse se réfère à la vitesse de la nouvelle génération de données et de la distribution.

- La variété se réfère à la complexité des données qui peut conduire à un manque de qualité et de précision.

La deuxième branche des défis est les traitements, il comprend la collecte de données, la résolution des similitudes trouvées dans différentes sources, les données de modification à un type acceptable pour l'analyse.

La dernière branche des défis est le défi de la gestion, cette classification permet la gestion des données, le contenu de stockage de données, traitées et collectées. Les sujets d'étude les plus importants sont la vie privée des données, la sécurité .[4]

I.3. Sécurité et vie privé

I.3.1. Sécurité : survole général

Travailler beaucoup de plates-formes et outils de grandes quantités de données sont en train d'émerger, la gestion de ces des Big data, et pour disponibilité les données, doivent être des mesures de sécurité appropriées ou la vie privée. Il peut également être une solution de sécurité des données pour répondre à trois exigences:

1) la vie privée se réfère à la protection des données contre la divulgation non autorisée.

2) l'intégrité se réfère à la prévention de la modification non autorisée et incorrecte des données.

3) la disponibilité désigne la prévention et la récupération des erreurs matérielles et logicielles. [8]

I.3.2. Les six éléments de sécurité

La sécurité de les informations un point important dans le maintien de l'intégrité de l'information, et il y a six éléments de sécurité nécessaires à la sécurité des informations.

I.3.3. Disponibilité

Ont utilisé plusieurs contrôles pour préserver ou la disponibilité des fichiers de données, Ainsi la préservation de la disponibilité doit être en tant que but de la sécurité de l'information.

I.3.4. Utilitaire

Dans ce cas, a chiffré la seule copie d'informations précieuses, pour préserver l'utilité de l'information dans ce cas, la gestion doit nécessiter des copies de sauvegarde obligatoires de toutes les informations critiques et devrait contrôler l'utilisation de mécanismes de protection puissants tels que la cryptographie.

I.3.5. Intégrité

L'intégrité est un engagement à des règles d'éthique, La plupart des contrôles peuvent être appliqués pour prévenir la perte de l'intégrité de l'information, en utilisant et en vérifiant les nombres de séquence, les sommes de contrôle et les totaux de hash pour assurer l'intégralité et la totalité d'une série d'éléments.

I.3.6. Authenticité

Pour assurer l'exactitude de l'information peut être appliquée à un certain nombre de contrôles, il s'agit notamment de confirmer les transactions, les noms, les livraisons et les adresses, La validation des produits, Et en utilisant des signatures numériques et des filigranes pour authentifier des documents.

I.3.7. La Confidentialité

Les contrôles pour maintenir la confidentialité comprennent l'utilisation de la cryptographie, de résister aux attaques trompeuses d'ingénierie sociale visant à obtenir leurs connaissances techniques et à contrôler l'utilisation d'ordinateurs.

I.3.8. La Possession

Devrait inclure un modèle de sécurité pour protéger la possession d'informations afin d'éviter le vol, les renseignements sont confidentiels ou non, et des contrôles qui peuvent protéger la possession de l'information, y compris l'utilisation des lois sur le droit d'auteur et la mise en œuvre des limites d'utilisation physique et logique, maintenir et analyser vos journaux d'audit informatique. [16]

I.3.9. Sécurisation : SI Via BigData

Les domaines de sécurité et administratifs sont les concepts clés, autour desquels les services de sécurité et les protocoles sont construits. Un domaine fournit un contexte pour établir un contexte de sécurité et une relation de confiance. Pour le maintien du contexte de sécurité support de source de données et les technologies de virtualisation améliorées peuvent potentiellement constituer une base pour développer des solutions appropriées, et peuvent fournir un environnement sécurisé.

I.3.10. Types de vie privée

Afin de maintenir la confidentialité Nous avons identifié sept types généraux de contextes de sécurité et les mesures d'accompagnement pour protéger ces contextes et protection :

I.3.10.1. Sécurité physique

Il vise à protéger les caractéristiques physiques comme les propriétés des systèmes et objets.

I.3.10.2. Sécurité politique

Assure la protection des droits acquis pour des institutions et politique reconnu.

I.3.10.3. Sécurité socio-économique

Traite des mesures économiques visant à protéger le système économique et le développement.

I.3.10.4. Sécurité culturelle

Maintenir la continuité des régimes traditionnels de la langue, la culture et les associations ... etc.

I.3.10.5. Sécurité environnementale

Basée sur la fourniture de la sécurité les risques environnementaux et réduction les problèmes environnementaux.

I.3.10.6. Sécurité d'incertitude

Menaces exceptionnelles et rares, il continue de menacer sérieusement la détérioration de la qualité de vie.

I.3.10.7. Sécurité de l'information

Pour protéger l'information accès non autorisé l'utilisation : divulgation, modification, lecture, inspection, enregistrement ou destruction. [11]

I.3.11. Défis de protection et sécurisation

L'augmentation des données est un défi fondamental pour les principes de protection des données , le principal objectif de la sécurité des Big data est de protéger la vie privée des utilisateurs.

Nous risquons de perdre les avantages qui peuvent être tirés de Big data si nous essayons les anciennes méthodes de protection Big data, et besoin d'un nouveau modèle de protection .les données peuvent être protégées par : Garder l'intégrité de plusieurs copies de données, Défis pour le maintien de l'anonymat des données, Confidentialité de l'analyse des données, Identifient duplicate data sets.

Celui qui vise à contrôler la collecte et la conservation de données personnelles peut ne plus être suffisant pour protéger la vie privée et la nécessité de sanctions légales contre l'utilisation abusive. [10]

I.3.12. Vie privé : survole général

Avec l'utilisation croissante des services en ligne et l'échange d'informations personnelles pour collecter, analyser et stocker des données personnelles, cela soulève des inquiétudes concernant la perte de données ou l'utilisation abusive.

Il y a eu beaucoup de recherches et d'études afin de préserver la vie privée. Le problème de la confidentialité des données à l'ère des données importantes est difficile. Et des tendances de la recherche (efficacité, sécurité et confidentialité, propriété des données, contrôle d'accès et confidentialité statistique).

Malgré les réalisations importantes de la recherche sur la protection de la vie privée ,la vie privée demeure l'un des principaux défis des big data. [8]

I.3.13. Types de vie privée

Pour déterminer la vie privée et essayer d'organiser d'éventuelles violations visant à protéger la vie privée, il a été suggéré que sept types différents de la vie privée qui doivent être protégés et qui sont d'un intérêt et d'évaluer différentes

I.3.13.1. La vie privée de la personne :

Cet aspect de la vie privée comprend le maintien des fonctions et des caractéristiques du corps

I.3.13.2. Comportement de la vie privée et travail

Y compris les questions délicates telles que les douanes, les activités politiques et les pratiques religieuses. Le concept de vie privée du comportement personnel qui se produit dans le traitement de l'environnement externe

I.3.13.3. la vie privée communications

Maintient les communications et évite l'interception des communications, telles que l'interception du courrier, e-mail, du téléphone ou des télécommunications.

I.3.13.4. Vie privée et image des données

Inclure conservateur des données personnelles Pour être disponibles automatiquement ou peuvent être accessibles par d'autres personnes et organisations.

I.3.13.5. la vie privée des pensées et des sentiments

Il est l'un de la vie privée personnelles est un droit de ne pas partager des idées ou des sentiments ou ces pensées ou sentiments peuvent ne pas être disponibles.

I.3.13.6. la vie privée de l'emplacement et de l'espace

Les personnes ont le droit de voyager dans tous les lieux publics ou privés sans les identifier, le suivre ou les surveiller.

I.3.13.7. la vie privée de l'association

L'association a le droit de pratiquer des activités sans suivre et surveiller leurs relations et activer l'association. [18]

I.3.14. Défis et enjeux de la vie privée

Afin de traiter la vie privée des BigData , cela crée de plusieurs enjeux et défis. Il existe certainement beaucoup de discussions sur la vie privée .Nous trouvons des défis de sécurité et de protection de la vie privée à MapReduce (Le style de traitement parallèle massif) , Ce qui est très différent du calcul classique, Parmi les défis à MapReduce:

- *Taille des données d'entrée et son stockage* :Les entrées dans de Big data sont décrites comme 4V (volume, vitesse, variété et véracité), Un défi dans la vie privée des comptes MapReduce ,Les données sont réparties en divisions ,petite taille copiée et distribuée sur plusieurs nœuds. Chaque partition est transférée de manière sûre et spéciale.

- *La nature distribuée des calculs MapReduce* : MapReduce contient de nombreux nœuds et des données réparties en parallèle, la probabilité d'attaques par contre au système central

-*Flux de données* :Les calculs MapReduce nécessitent un flux de données complexe entre différents nœuds de stockage, différents nœuds informatiques et nuages généraux, comme suit:

-Entre le stockage de données et les nœuds informatiques: les calculs de MapReduce sont effectués près de l'emplacement des données pour minimiser le flux de données et cela entraîne un flux de données du stockage vers les nœuds informatiques.

-Nuages (clouds en anglais) généraux: le compte MapReduce peut être implémenté dans plus d'un nuage différent et est vulnérable aux attaques

-*Nuage hybride* : Les nuage hybride a une gestion efficace et économique des ressources et a un traitement efficace des données. Malheureusement, MapReduce est conçu pour fonctionner dans un nuage et cela représente des défis dans le nuage hybride.

- *Les défis économiques* : Le nuage général est définie pour trois facteurs économiques: le stockage de données, les coûts de communication et le temps de compte. La vie privée économique doit être intégrée aux calculs de MapReduce.

Les défis susmentionnés reflètent les efforts déployés par MapReduce dans le tirage au sort en ce qui concerne les nouvelles exigences en matière de sécurité et de confidentialité. [19]

I.3.15. Sécurité via la vie privée

L'une des raisons les plus importantes est la nature personnelle, qui vise à réduire les risques de sécurité qui affectent la vie privée de l'individu. Parmi les méthodes de traitement ce problème, isoler et masquer les données.

Lorsqu'il s'agit de big data , il est difficile de fournir pour déterminer la crédibilité des données lors de l'utilisation de données anonymes .Afin d'assurer la vie privée, nous aurons beaucoup de ces ensembles de données qui peuvent être liés. Le défi auquel nous sommes confrontés est l'augmentation des ensembles de données, la vie privée peut être perdue Lorsque l'anonymat, devenant également douteuse avec la sécurité. [13]

I.3.16. Cryptage de données

Afin de protéger les informations stockées dans de big data , l'utilisation de la technologie de cryptage est nécessaire, et le traitement des big data en raison de sa grande volume et de sa diversité entraîne des défis pour le cryptage des données. Contrairement aux méthodes de cryptage des données précédentes petite et moyenne taille, dans cette section, Les avancées récentes dans la méthodologie de cryptage, y compris le cryptage consultable, le cryptage conservateur d'ordre, le cryptage structuré et le cryptage Homomorphique.

I.3.16.1. Cryptage recherché

Au cours des dernières années, le cryptage consultable est apparu dans le cloud computing et les Big data. Il s'agit d'un cryptage primitif qui permet à l'utilisateur de rechercher en toute sécurité des données cryptées, par des mots-clés sans décryptage préalable. Ceux-ci sont cryptés de manière à ce que leur contenu soit caché sur le serveur, à moins que les symboles appropriés ne soient donnés.

I.3.16.2. Commander-Préserver le cryptage

Est une façon de préserver les textes chiffrés, étudiés à l'origine en théorie dans la base de données. Exécuter sur les opérations du système du texte chiffré, de la même manière que dans un chiffrage sur les plaintexts, et cette propriété fonctionne bien, comme l'une des stratégies cryptographiques clés, qui permet d'exécuter des requêtes de base de données efficaces.

I.3.16.3. Cryptage structuré

En cryptage structuré élimine le stockage de grandes portions de données sous forme de texte simple, mais possède des structures de données. Le cryptage structuré a été proposé, afin d'accéder à ce type de données chiffrées. Le chiffrement structuré peut être considéré comme une généralisation du cryptage basé sur l'index.

I.3.16.4. Cryptage homomorphie

Le premier codage homomorphique a été proposé en 2009. Afin d'assurer le décryptage, il a fallu la construction d'un diagramme qui nécessite la gestion de la partie aléatoire restante appelée bruit, pour un fonctionnement homomorphique, développé par le «bootstrapping», qui

permet de passer d'un certain texte chiffré à un nouveau texte chiffré qui crypte le même, mais a moins de bruit, Malheureusement le bootstrapping est potentiellement lourd. [12]

I.3.17. Gestion de la confiance

été étudié confiance dans la plusieurs disciplines, y compris la sociologie, la psychologie, l'économie et l'informatique, largement liées à la sécurité et à la vie privée, et la confiance dépend de deux facteurs importants: le risque et l'interdépendance. La source de risque est l'incertitude quant à l'intention de l'autre partie, et l'interdépendance du fait que les intérêts des parties sont liés et ne peuvent être atteints sans compter l'un sur l'autre.

L'évolution que vous connaissez la Big data en croissance, cela signifie la nécessité de fournir une confiance totale pour fournir les meilleurs services. Il existe plusieurs problèmes concernant la fiabilité des Big data dans le cycle de vie des big data, Devrait recevoir plus d'attention et doit être largement étudié. [12]

I.3.18. Définition et Types de vulnérabilité

Les points des faiblesses apparaissent souvent lorsque les systèmes de sécurité sont exposés à les cyber-attaques, Cela entraîne non disponible d'informations qui ne sont pas disponibles pour les utilisateurs autorisés. Nous trouvons souvent des abus dans l'accès à l'information préjudiciable en volant des informations confidentielles, en sabotant des informations et en bloquant l'accès à l'information. Les agents des menaces réussissent souvent à endommagé ou à abuser d'actifs car ils exploitent les vulnérabilités présentes dans l'infrastructure d'information critique.

Afin de travailler au renforcement de l'infrastructure de l'information, il faut des mesures contre les attaques électroniques, Nécessite des mesures pour identifier et renforcer les faiblesses.

I.3.18.1. Vulnérabilités de logiciels

La vulnérabilité des logiciels permet aux logiciels malveillants d'exploiter la vulnérabilité des logiciels pour accéder à un système. Cela peut conduire à l'accès des informations stockées dans la base de données.

I.3.18.2. Vulnérabilités du personnel

L'exploitation inverse de la discipline par le personnel est la faiblesse des employés mécontents qui ont la capacité de nuire à l'information de leurs organisations. Comme la plupart des employés ont accès à des informations confidentielles dans leur organisation. Les faiblesses des employés incluent non seulement les employés mécontents. Les attaques électroniques peuvent viser le sujet et utiliser ces méthodes dans le domaine de l'ingénierie sociale.

I.3.18.3. Vulnérabilités de planification de récupération après sinistre

Lors de la planification de faire face aux cybera-ttaques, il y a plusieurs fallacieuse et aboutissent à de faibles de planification. En raison de ces vulnérabilités, l'organisation peut ne pas être en mesure de récupérer des informations perdues en raison de cyber-attaques.

I.3.18.4. Vulnérabilités du protocole réseau

Les attaques sur les protocoles réseau sont exploitées pour perturber ou résoudre des sites. Les attaquants exploitent ce protocole pour sa vulnérabilité, permettant d'utiliser des informations confidentielles, il est important que les organisations prennent des mesures pour empêcher l'exploitation de ces protocoles. [7]

I.3.18.5. Infrastructure critique et BigData

L'infrastructure d'information est importante dans de nombreux pays, et beaucoup d'efforts sont faits pour protéger infrastructure d'information critique. Les big data ont des implications importantes comme l'on cherche à protéger infrastructure d'information critique. En matière d'infrastructure, le plus grand défi n'est pas la confidentialité, mais la disponibilité est importante.

L'«infrastructure d'information» était protégée contre l'accès non-Internet et la protection a été compromise par sa connexion à Internet. Et a entraîné une augmentation de la taille des données en raison de la connexion de nombreux objets intelligents, par exemple: à l'aide du réfrigérateur intelligent, tout le réfrigérateur d'information devient un grand entrepôt de données. La sécurisation des big data est un aspect clé du défi auquel nous sommes confrontés, des défis à surmonter. [13]

I.3.19. Contrôle de sécurité et protection d'infrastructure

Afin de sécuriser et protéger l'infrastructure contre les menaces et les cyber-attaques, Devrait être discuté Trois contrôles de sécurité: Contrôles préventifs, correctifs et détective.

I.3.19.1. Contrôles préventifs

Pour mettre en œuvre des contrôles préventifs, suivre des stratégies visant à prévenir et prévenir les vulnérabilités : politiques de prévention, pare-feu, antivirus, tests d'intrusion.

I.3.19.2. Contrôle des détectives

Lorsque les contrôles de prévention ne permettent pas d'éviter les attaques électroniques, la stratégie des contrôles de détection est utilisée. La stratégie vise à réduire l'impact de l'exploitation de la vulnérabilité et à atténuer les dégâts causés par les cyber-attaques, car elle permet de détecter rapidement les attaques. Un certain nombre de contrôles de détective seront vérifiés ci-dessous. [7]

I.3.20. Comment sécuriser les BigData

L'un des problèmes les plus importants traités par de big data de sécurité et la vie privée, la taille, la variété et la rapidité des big data, il est difficile de traiter de des big data grâce à la sécurité. Nous discuterons Comment sécuriser les big data

Dans cette section, les principaux éléments proposée pour la sécurité les big data : Gestion des données, Gestion de l'identité et de l'accès, Protection des données et confidentialité, Sécurité réseau et Sécurité et intégrité de l'infrastructure.

I.3.20.1. Gestion des données

Afin de bien gérer les données, nous devons examiner trois branches de principales. La classification des données qui facilite le suivi de la sécurité des données. La découverte de données est à risque et fonctionne sur la protection. Décrivez les données en comprenant les flux de données de bout en bout dans un environnement les big data.

I.3.20.2. Gestion de l'identité et de l'accès

Afin d'assurer la sécurité les big data, il faut identifier les utilisateurs et comment se connecter aux données et accéder aux utilisateurs aux données via la gestion des politiques d'accès centralement. Accès aux données personnalisées par rôle et non par l'utilisateur.

I.3.20.3. Protection des données et confidentialité

Afin de fournir une plus grande protection et une plus grande vie privée big data , nécessite le travail de cryptage des données, peut fournir des capacités de sécurité et de protection précises. Il peut également protéger les fichiers chiffrés de l'accès et protéger le cryptage des données sensibles vers le fondateur.

I.3.20.4. Sécurité réseau

La sécurité réseau est importante pour assurer la sécurité les big data. Pour la sécurité des réseaux, la protection des données en transit. Utilisation du protocole HTTPS pour empêcher la divulgation d'informations et assurer la confidentialité des communications. Et dans le zonage de la sécurité du réseau pour le contrôle d'accès dans le réseau.

I.3.20.5. Sécurité et intégrité de l'infrastructure

La sécurité de l'infrastructure est une analyse de tous les changements dans le système. Travailler sur la sécurité de la sécurité des infrastructures est très important dans la cohésion les big data. Linux Security Optimiser a été créé, ce qui permet d'offrir plus de protection et d'intégrité à l'infrastructure. [15]

I.4. Conclusion

Le Big Data joue un rôle important dans plusieurs domaines et cette tendance devrait augmenter

il y a un certain beaucoup de défis de sécurité et de confidentialité qui doivent être abordés pour nous permettre d'exploiter tout le potentiel des big data pour disponibilité, telles que des algorithmes de cryptage et de décryptage efficaces, des mécanismes de préservation de la vie privée, la fiabilité et la vérification de l'intégrité des mégadonnées.

Chapitre II

Approches et travaux connexes

II.1. Introduction

La disponibilité dans la BigData permettant de maintenir le bon fonctionnement du système d'information, L'objectif de la disponibilité est de garantir l'accès à un service ou à des ressources.

Pour réaliser un system permettant l'assurance de la disponibilité des BigData nécessitant les approches de travaux connexes.Resumons les méthodes proposées par différentes recherche dans ce domaine Pour trouver des points plus importants afin de travailler sur ce projet.

II.2. Problème est recommandation lies à la disponibilité

La question de la disponibilité des données est un défi majeur et comporte de nombreuses complexités. Cela inclut la complexité, le type, l'emplacement et la fréquence de la source de données. La difficulté est lorsque les sources ne sont pas synchronisées ou utilisent des formats de sortie différents.

Cette section décrit les problèmes et les recommandations concernant la disponibilité des données.

II.2.1.Limites d'utilisation des données

Il existe des limites dans l'utilisation de certaines données, et cela a un impact dans l'environnement de test expérimental, mais les restrictions sont créées de différentes façons.

- *Propriété des données:* il existe plusieurs quantités de données et il est difficile pour les chercheurs d'y accéder en raison des limitations dans l'utilisation des données.

- *Sécurité des données:* pour la sécurité des données et la vie privée, il faut développer des restrictions qui protègent les données.

II.2.2.Données non existantes ou manquantes

Le manque de données pour diverses raisons, selon les régions ou les données, coûte cher à capturer ou à dissimuler des données importantes. Dans les temps, il peut être impossible de capturer des données. Aller vers non existantes ou manquantes les données.

II.2.3. Calendrier et synchronisation des données

La synchronisation des données est un problème majeur, et les entreprises possède plusieurs producteurs d'informations. Pour utiliser plusieurs sources, vous devez comprendre quand cette information a été créé et elle devra être développé en horaires connus, y compris une mesure ou une estimation de la différence de temps.

II.2.4. Données de fréquence

Le traitement des algorithmes peu performants, ce qui entraîne de petites augmentations de la taille des données, elle peuvent entraîner des modèles de détection supplémentaires. Les problèmes de fréquence sont une préoccupation majeure à cause de l'impact sur la précision des données. Par exemple, pour certains algorithmes d'apprentissage automatisés, augmenter la taille du package de formation peut produire des données supplémentaires.

II.2.5. Formats de données, normes et spécifications

Les données différenciées sont stockées, finalement dans des fichiers formatés selon une variété de critères. Des normes et des spécifications ont été établies pour aider à certaines questions et l'utilisation de normes bien reconnues à chaque niveau de coordination des données et de description des formats de données et de la signification.

II.2.6. Incertitude et fiabilité des données

Parmi d'autres choses qui traitent d'incertitude des Big data, la fiabilité et l'exactitude des données. Certaines de ces caractéristiques devraient être traitées en fournissant des métadonnées appropriées, comme dans le cas précédent, et il est proposé que plusieurs ensembles de données correspondant aux données originales et modifiées soient disponibles.

II.2.7. Accès, stockage et traitement des données

Afin de lier aux données nécessitant plusieurs analyses de données et de disponibilité. Les données sont en train d'être restructurées par les chercheurs pour stocker, et accéder à des données non séquentielles. Étant donné que la taille de l'ensemble de données est très grande, les chercheurs n'ont besoin que d'un accès relativement restreint. Le processus de traitement

des données a lieu dans l'environnement des propriétés contrôlées de sorte que les résultats soient renvoyés sans exposer les données brutes.

II.2.8. Temps de préparation des données et autres ressources

Le traitement des bases de données en coordination et en mobilisation nécessite un temps limité pour les ingénieurs et le fabricant doit prendre le temps de décider de la propriété et de la manière de le séparer. Pour les accords de non-divulgation qui peuvent différer entre les consommateurs et les ensembles de données. Il oblige les entreprises à distribuer des données avant de lancer des projets, afin d'assurer la diffusion précise et utile des données.[34]

II.3. Disponibilité dans les BigData

II.3.1. Base de données NoSQL

L'augmentation significative du nombre d'utilisateurs de sites Web modernes et de grandes entreprises fait face à de nouveaux défis. La technique le système de gestion de base de données relationnelle (SGBDR) ne peut pas atteindre les nouvelles exigences,si le défi passe à la règle relationnelle.

La base des données NoSQL(données non relationnelles) ont été créée pour traiter les entrepôt ou les Big data . C'est une classe très large dans les systèmes de gestion de base de données, qui ne suit pas les données relationnelles. Il est caractérisé par NoSQL:

- Possibilité de répliquer et de diviser les données sur de plusieurs serveurs

- Ajouter dynamiquement de plusieurs attributs aux enregistrements de données.

- Le modèle de synchronisation est plus faible que le SGBDR

Le modèle de données de base du magasin de colonnes est que les enregistrements peuvent être divisés verticalement et horizontalement, et chaque ligne et chaque colonne sont divisées en plusieurs nœuds, les bases de données de colonnes qui stockent les informations dans plusieurs emplacements sur le disque.

Le modèle de données de base du magasin de colonnes est que les enregistrements peuvent être divisés verticalement et horizontalement, et chaque ligne et chaque colonne sont

divisées en plusieurs nœuds, les bases de données de colonnes qui stockent les informations dans plusieurs emplacements sur le disque. [21]

II.3.1.1. Les inconvénients

- La plupart des systèmes NoSQL sont des projets open source, qui peuvent perdre leur crédibilité.

- Soutien du projet par la plupart des producteurs de bases de données relationnelles.

- Manque de recherche combinant NoSQL avec la technologie de base de données relationnelle afin de créer des bases de données hybrides.

- Manque de traitement avec le cloud computing.

II.3.2. Plateforme géographique

Les travaux de développement afin d'élargir le champ d'application d'un système de serveur géographique nécessite le développement de la croissance les données spatiales. La disponibilité des ressources joue un rôle de plus en plus important dans la disponibilité des services et la fourniture d'un temps de réponse acceptable. Et l'ajout de services et leur évolutivité.

II.3.2.1. Le rôle le cloud computing dans la plateforme géographique

Le cloud computing est destiné à externaliser les services informatiques. Le cloud computing est un modèle de développement de diffusion et distribution de l'information. Avec le cloud computing les ressources peuvent être utilisées plus efficacement. Le cloud computing est une bonne alternative à un client système standard. De nombreuses entreprises font confiance à leurs propres applications et services de conseil.

II.3.2.2. Base de données distribuée

Le travail des bases de données distribuées vise à assurer la transparence dans la distribution. La base de données est distribuée à de plusieurs ordinateurs dans le système, parce que l'échec de l'ordinateur peut continuer à accéder à la base de données. Dans les

systèmes de base de données distribués, les performances augmentent car plusieurs serveurs gèrent les données disponibles afin de raccourcir le temps de réponse.

II.3.2.3. Système existant CityServer3D

La technologie CityServer3D consiste en une base de données géographique, CityServer3D crée automatiquement des modèles tridimensionnels et réalise ainsi des simulations dans le monde 3D. Le programme peut gérer des données bidimensionnelles, tridimensionnelles et inter graphiques ensemble.

II.3.2.4. Disponibilité du système

La disponibilité d'un standard de qualité, et il a été démontré que deux des serveurs MongoDB ont un impact négatif sur la disponibilité. Un nouveau serveur doit être élu par trois serveurs. La haute disponibilité est obtenue en utilisant au moins deux serveurs CityServer3D et trois serveurs MongoDB dans le système. [22]

II.3.3. Gestion de la tolérance aux fautes

La tolérance aux pannes des lignes de traitement de données est l'une des préoccupations les plus importantes dans le développement de la planification de l'emploi pour le traitement de Big data, mais la plupart des tables n'assurent qu'une tolérance aux pannes pour les nœuds dépendants Le facteur est souvent considéré comme un nœud principal très complexe. Afin de faire face à la probabilité accrue d'erreur humaine, ils séparent le nœud principal du reste de la masse.

II.3.3.1. Méthodes gestion de la tolérance aux fautes

II.3.3.1.1. Formulaire de compte

Un modèle parallèle synchrone est utilisé afin de déduire le modèle de tolérance de panne qui convient aux applications de Big data. Ce modèle est qu'un programme parallèle consiste en plusieurs étapes parallèles consécutives en interne. Toutes les étapes de la séquence sont placées là où c'est possible.

II.3.3.1.2. Modèle Failover

La capacité du système de conversion d'un élément défaillant à sauvegarder. Est la possibilité de passer d'un nœud principal défaillant à un nœud de sauvegarde avec une restauration complète de l'état d'exécution. Ce sont les capacités de base qui composent la capacité du système répartir sur l'échec.

II.3.3.1.3. Modèle de programmation

Pour construire une hiérarchie composée d'un ensemble de nœuds de cluster. Un algorithme qui rend le groupe plus gênant a été développé en entrant plusieurs nœuds principal. Si un nœuds principal échoue, un autre nœud du même niveau ou supérieur est remplacé dans la hiérarchie.

II.3.3.1.4. Gestion des échecs des nœuds principaux

Cette méthode nécessite la synchronisation de tous les nœuds qui effectuent des subordonnés à partir du premier noyau. Cette méthode nécessite la synchronisation de tous les nœuds qui effectuent des subordonnés à partir du premier noyau. Cependant le nœud de sauvegarde est également déconnecté au milieu de la mise en œuvre de certains noyaux secondaires. [23]

II.3.3.2. Inconvénients

- En cas d'algorithmes, d'autres erreurs peuvent se produire si elles n'affectent pas les nœuds principaux.

- Il n'y a aucun moyen de rendre disponible l'argument existant sans réécrire les instructions.

- Difficulté à gérer les rôles serveur / client.

II.3.4.Assurer la disponibilité des données.

Traiter les applications BigData pour l'accès aux données nécessite des mécanismes pour assurer la disponibilité des données. Nous discutons de quelques techniques communes pour s'assurer que l'application de BigData

II.3.4.1. Performance dans la disponibilité de Big data

Les performances concernent l'utilisation des ressources de données pour exécuter rapidement et efficacement les données, que toutes les données requises soient accessibles en temps opportun en raison de la taille le données.

II.3.4.2. Télécharger les données sur l'appareil

La plupart des solutions lors du transfert les Big data garantissent l'utilisation de tables vers de chargement pour les grandes applications. Suivi par un processus qui charge de nouvelles données dans l'appareil. Un délai peut apparaître au moment où le périphérique est chargé dans des données volumineuses.

II.3.4.3. Les données sont situées dans plusieurs endroits

La présence de données dans plusieurs endroits est une augmentation de la disponibilité des données dans la distribution de l'accès aux données. La plupart de leurs le Big data résident à au moins deux endroits: tables et périphériques. En fait, il peut être situé ailleurs aussi

II.3.4.4. Les méthodes d'accès aux données

II.3.4.4.1. indexation:

lors de la conception les Big data, est créé un index pour chaque colonne ou ensemble de colonnes. Pour faciliter le processus de recherche et être équilibré.

II.3.4.4.2. Fractionnement horizontal

il est courant d'affecter des lignes de données à des supports physiques distincts. Cette séparation peut avoir de nombreux avantages.

II.3.4.4.3. Le partitionnement par plage de dates est courant

Cela permet aux applications d'accéder à des sections séparées les unes des autres. C'est une purge qui s'exécute sur les données dans la section la plus ancienne tandis que les nouvelles lignes de table sont insérées dans une autre

II.3.4.4.4. Division verticale

De cette manière, l'administrateur de base de données analyse les données de colonne dans une table pour déterminer si certaines colonnes ont été mises à jour ou interrogées plus ou moins fréquemment. [24]

II.3.5. Oracle au service de la disponibilité

Oracle travaille pour réaliser les meilleures études en disponibilité des Big data. C'est un système composé de nombreux appareils et logiciels. Utilisation de technologies de données Oracle modernes, Oracle Big Data Appliance et Oracle Exadata Data base Machine. Vérifier la haute disponibilité et éliminer les problèmes de temps d'arrêt.

Le projet Oracle comprend deux phases:

Phase 1: Scénarios de haute disponibilité et de répartition dans un site

Phase 2: Scénarios de récupération après sinistre sur plusieurs sites.

Oracle offre une grande flexibilité et des plates-formes sécurisées et prend également en charge l'analyse sur de big data. Les données peuvent être traitées et intégrées en fonction de leurs structures géographiques

II.3.5.1. Architecture oracle Big data

Oracle Big Data Architecture comprend les technologies suivantes :

- L'entrepôt de données est utilisé comme une nouvelle ressource pour de nombreuses sources de données structurée et non structurée. Les appareils sont connectés les uns aux autres afin de traiter les données en termes de stockage, de performance et de croissance.

- Exadata système de secours est utilisé qui est une copie pour maintenir les bases de données synchronisés

- Réplication de données appareils big data Assure la haute disponibilité et la cohérence des données.

- peut être utiliser système de gestion de big data .Que ce soit dans le Cloud ou service Cloud le Big data Oracle.

Est considéré Oracle big data disponibilité complète et intégrée. Et peut être déployé dans le cloud avec service de cloud de big data d'Oracle. Fournit la technologie Oracle Big Data SQL Technologie haute performance Pour accéder aux données.Force est activée le format complet d'Oracle, SQL pour disponibilité une vue unifiée dans la base de données Oracle,Hadoop, et sources Nosql. [33]

II.3.6.La disponibilité dans le Cloud

II.3.7.Approche multi-Cloud

Il a été proposé une nouvelle approche, Très confidentiel et Très disponibilité à un coût raisonnable. en utilisant la technologie de schéma de partage secret.

II.3.7.1. Système de partage secret

Un système de partage de secret est la méthode qui génère ce que l'on appelle des "shares" à partir des données d'origine. Les données d'origine peuvent être récupérées si le nombre d'actions est supérieur ou égal à un seuil

II.3.7.2. Environnement supposé pour l'utilisation de la technique proposée

Chaque nuage gère (SLA) son propre, Et envoie le cloud (sla) à l'utilisateur, en comparant l'utilisateur entre la demande de cloud et (sla) et détermine la séparation des services de cloud par l'utilisateur.

II.3.7.3. Environnement supposé de la technique proposée

Quant à cette technique proposée, Il est supposé que l'utilisateur utilise les services de cloud pour stocker des données, l'utilisateur configure les données avec la disponibilité de la commande, la confidentialité, le coût et la performance.

II.3.7.4. L'utilisateur demande des statistiques et un SLA Cloud

Les statistiques de demande utilisateur sont composées de disponibilité, de confidentialité, de coût et de performance. pour la disponibilité est le taux d'utilisation de l'utilisation de plusieurs services de cloud

. pour le coût est le coût total pour l'utilisation de plusieurs services de cloud. pour la performance est le temps de transfert de données. pour la confidentialité est le degré de risque.[25]

II.3.7.5. Inconvénients

-l'efficacité en utilisant hétéro-nuage n'est pas présentée

-Manque d'efficacité qui correspond à la mesure du rendement

-il est nécessaire d'effectuer une étude sur la conception générale du système en utilisant la méthode proposée

II.3.8. Approche basé prix/réplication de données

II.3.8.1. Hadoop Distributed File System

L'architecture HDFS se compose principalement de deux types de composants: un NameNode et multi-DataNodes. Le NameNode gère l'espace de noms et le fichier du système de fichiers contrôle d'accès des clients. les DataNodes gèrent les périphériques de stockage attachés aux nœuds sur lesquels ils s'exécutent en interne. chaque fichier à stocker dans DataNodes est divisé en plusieurs blocs, et le NameNode décide de l'endroit où chaque bloc sera stocké parmi les DataNodes disponibles.

II.3.8.2. Modèles d'avantages financiers

supposons qu'un CSP exploite un HDFS avec un total DataNodes homogènes possédant une capacité de stockage totale. Avec un nombre total d'utilisateurs sur la plate-forme, sans perdre le généralité.

II.3.8.3. Modèles de coûts d'exploitation

Pour satisfaire les demandes des utilisateurs de stockage et de réplication, CSP doit payer pour les coûts d'exploitation, qui sont évidemment inférieurs aux revenus obtenus des utilisateurs et le coût d'opération pour le service de stockage

II.3.8.4. Formulation du problème

. Comme toutes les répliques, est stocké Répliques dans des serveurs géographiquement dispersés, réplication de plusieurs blocs de données, et augmentation disponibilité des données. [21]

II.3.8.5. Les inconvénients

- Difficulté à organiser demandé les utilisateurs ont simultanément l'espace de stockage des fournisseurs de services cloud Avec une capacité de stockage limitée.

-CSP font face au défi de sélectionner les utilisateurs qui seront servis pour maximiser le bénéfice final.

II.3.9. Gestion de désastre

Le système de fournisseur de services travaille pour résoudre le problème de désastre et discuter d'une solution directe. notre modèle de système de fournisseur de services se compose d'utilisateurs, de serveur de stockage et de serveur de sauvegarde.

Le serveur de stockage fournit un service de stockage et le serveur de sauvegarde fournit une sauvegarde, une restauration et une réponse à services de désastre, travailler indépendamment.

Notre système de service de sauvegarde comprend trois phases: (Configuration du compte, chargement des données et téléchargement de données)

II.3.9.1. Phase configuration du compte

D'utilisateur s'enregistre dans fournisseur de services de cloud pour accéder aux fonctionnalités fournies. Créer trois répertoires pour remplir le but de la sécurité et de la restauration pendant pour déclarer désastre.

II.3.9.2. Phase chargement des données.

Lors du chargement des données par l'utilisateur étant crypté et envoyé au serveur de stockage. ensuite, sauvegardez sur le serveur de sauvegarde où il est d'abord déchiffré et stocké dans un répertoire pour gérer les problèmes de désastre .

II.3.9.3. phase de téléchargement des données

Lorsque l'utilisateur veut les données qu'il demande au serveur. Il y a deux cas

premier cas les données sont disponibles est faite la demande à l'utilisateur. Un autre cas quand une désastre se produit sont fait opération restauration Utilisation de données de sauvegarde. [30]

II.3.9.4. Les inconvénients

-Système de service de sauvegarde nécessite un coût élevé.

II.3.10.Intégrité et disponibilité basé réplication

II.3.10.1. Structure de cloud hiérarchique

L'architecture hiérarchique du cloud est donnée en fonction des différents types de services fournis.

Où il se compose de trois couches. chaque couche organise les nœuds dans une séquence appropriée et construit la structure BAT pour aider les protocoles de localisation et de récupération.

-*Couche de service*

HCSP organisateur de multi-cloud. Lors de la vérification, il agrège les preuves de tous les coulds et répond à la preuve finale aux auditeurs désignés dans la procédure de récupération.

Couche de calcul.

Cette couche repose sur la capacité de calcul élevée du multi-cloud. Nous pouvons diviser cette couche en niveaux K

Couche de stockage

Cette couche est composée de nœuds de stockage physiques, qui sont des nœuds feuille sous la structure de BAT.

II.3.10.2. Construction de MRVR

a trois phases concrètes et chaque phase contient plusieurs fonctions.

II.3.10.2.1.Phase de préparation

DataFrag (): être un composant de fichier privé qui a sa signification physique

KegGen (): une clé symétrique est générée pour chiffrer le composant du fichier. Ensuite, l'algorithme de génération de clé au hasard

RepGen (): Pour réaliser plusieurs réplicas en multi-cloud, MRVR demande aux propriétaires de définir le nombre de réplicas

TagGen (): L'algorithme de génération de tags prend la clé secrète, le paramètre, l'ensemble de blocs de données et leurs répliques, et la liste de cloud comme entrées

II.3.10.2.2.Phase de vérification

Challenge (): L'algorithme de challenge prend l'information abstraite Finfo et le paramètre de vérification en entrée.

Prove (): Le HCSP reçoit le challenge Chal de l'auditeur, et il décompose Chal en sous-challenges sur chaque cloud

Verify (): L'algorithme de vérification prend en entrée le challenge Chal, la preuve, la clé publique et la table d'index, et calcule la valeur de hachage des blocs de données contestés

II.3.10.2.3.Phase de récupération des données

pour construire un protocole efficace pour localiser et récupérer les blocs corrompus en utilisant BAT

Localiser corruptions (): Pour localiser les blocs corrompus est une procédure de déplacement BAT.

Récupérer corruptions (): propose une stratégie sans téléchargement pour récupérer les blocs corrompus. [21]

II.4. Disponibilité hors BigData

II.4.1.Approche évolutive

Décrire notre modèle de disponibilité proposé pour la gestion des nœuds de service évolutifs et distribués, étant donné que la haute disponibilité est une exigence importante pour le service de gestion de confiance.

II.4.1.1. Puissance opérationnelle

Dans notre approche, nous proposons de répartir les noeuds de service de gestion de confiance sur divers nuages et de diriger de manière dynamique les demandes vers le nœud de

service de gestion de confiance approprié. afin que son niveau de disponibilité souhaité puisse toujours être maintenu.

Il est crucial de développer un mécanisme qui aide à déterminer le nombre optimal de nœuds de services de gestion de confiance. proposer que chaque nœud hébergeant une instance de service de gestion de confiance signale sa puissance opérationnelle. La puissance opérationnelle d'un nœud de service de gestion de confiance particulier est calculée comme la moyenne de la distance euclidienne

II.4.1.2. Détermination de la réplication

Dans notre cadre de gestion de confiance, propose d'exploiter les techniques de réplication afin de minimiser la possibilité qu'un nœud héberge une instance de service de gestion de confiance

L'instance de service de gestion de confiance peut tomber en panne pour plusieurs raisons telles que surcharge, réparation, mise à jour de service, etc., ce qui empêche les consommateurs de donner des avis de confiance ou de demander une évaluation de confiance pour les services cloud.

La réplication permettra à l'instance de service de gestion de confiance de récupérer toutes les données perdues pendant la période d'indisponibilité de sa réplique. [39]

II.4.2.Modèle comparaison

Méthode proposition	Performance	Évolutivité	Flexibilité	Fonctionnalité	Confidentialité	Cout
Base de données NoSQL	✓	✓	✓	✓	-	-
Plateforme Géographique	✓	×	-	×	✓	×
Gestion de la tolérance aux fautes	✓	-	-	-	✓	-

Assurer la disponibilité des données**.**	✓	×	×	✓	✓	×
Oracle au service de la disponibilité	✓	-	✓	✓	✓	×
Approche multi-Cloud	✓	-	✓	✓	✓	-
Approche basé prix/réplication de données	-	-	×	✓	-	✓
Gestion de désastre	✓	×	×	✓	×	×
Intégrité et disponibilité basé réplication	✓	-	-	✓	✓	×
Approche évolutive	-	✓	-	✓	-	×

II.5. Conclusion

La majeure partie de l'étude et de la recherche à l'heure actuelle, le lien vers les meilleurs services dans la Big data, à l'amélioration des projets ont été voir les approches et travaux connexes.

Dans ce chapitre, nous avons vu plusieurs travaux liés au projet sur lequel nous travaillons pour réaliser un system permettant l'assurance de la disponibilité des BigData

Chapitre III

Conception et Modélisation

III.1. Introduction

L'objectif du projet est la fin de l'étude comment faire réalisation d'une architecture pour l'assurance de la disponibilité des BigData.

Au travers du chapitre Approches et travaux connexes. Nous avons étudié des travaux de recherche pour tirer parti de leur savoir-faire. Dans ce chapitre, nous présenterons une proposition de conception détaillée, dans laquelle nous avons réformé la structure globale de l'application.

III.2. Conception générale du système proposé

III.2.1. Architecture globale

Dans cette section, un ensemble des composants qui représentent notre système. et aident à présenter l'architecture globale du système proposé.

Suppose que le système actuel pour l'assurance de la disponibilité des BigData

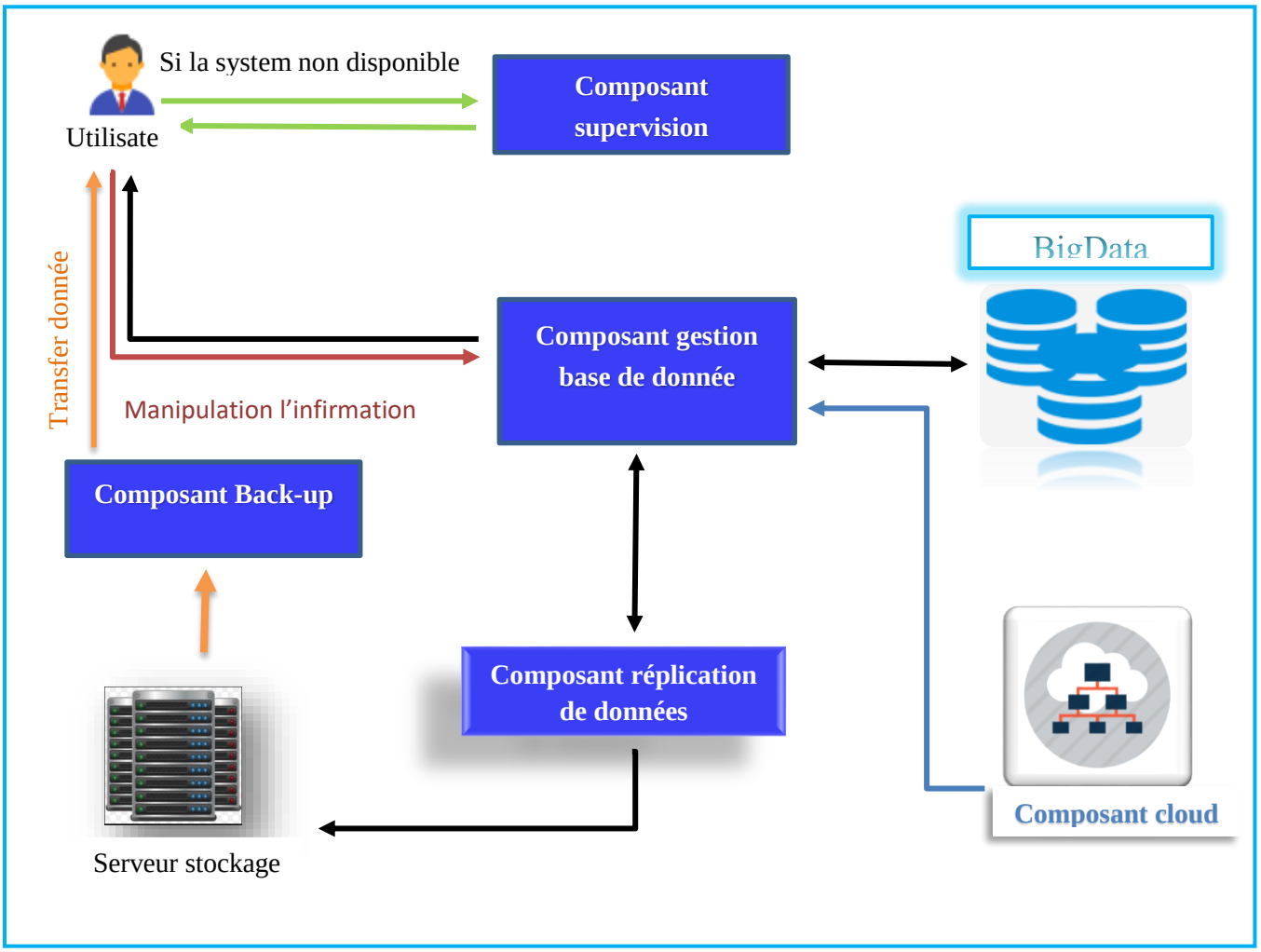

Figure III.1 : L'architecture globale du système proposé.

III.2.2. Architecture détaillée

Dans cette section, Expliquer chaque élément avec l'architecture de chaque composant et on détail le fonctionnement et les rôles de chaque composant.

III.2.2.1. Composant gestion base de données

2.2.1.1. L'architecture composant gestion base de données

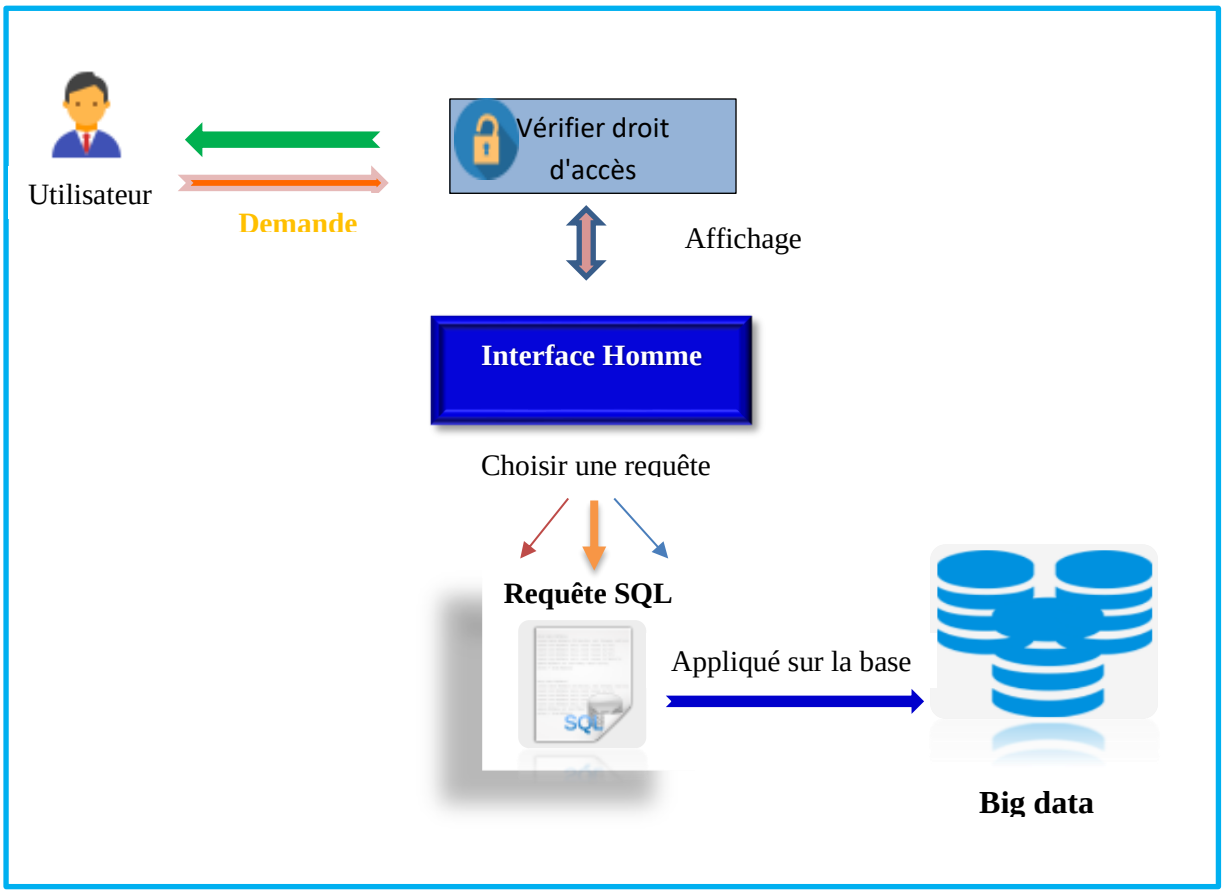

Figure III.2 : L'architecture composant gestion base de

2.2.1.2. les rôles composant gestion base de données

- ✓ Accéder facilement aux informations
- ✓ Communication utilisateur avec le système
- ✓ Les opérations de recherche et de manipulation des données
- ✓ Organisation et classification Informations

III.2.2.2. Composant réplication des données

2.2.2.1. l'architecture composant réplication

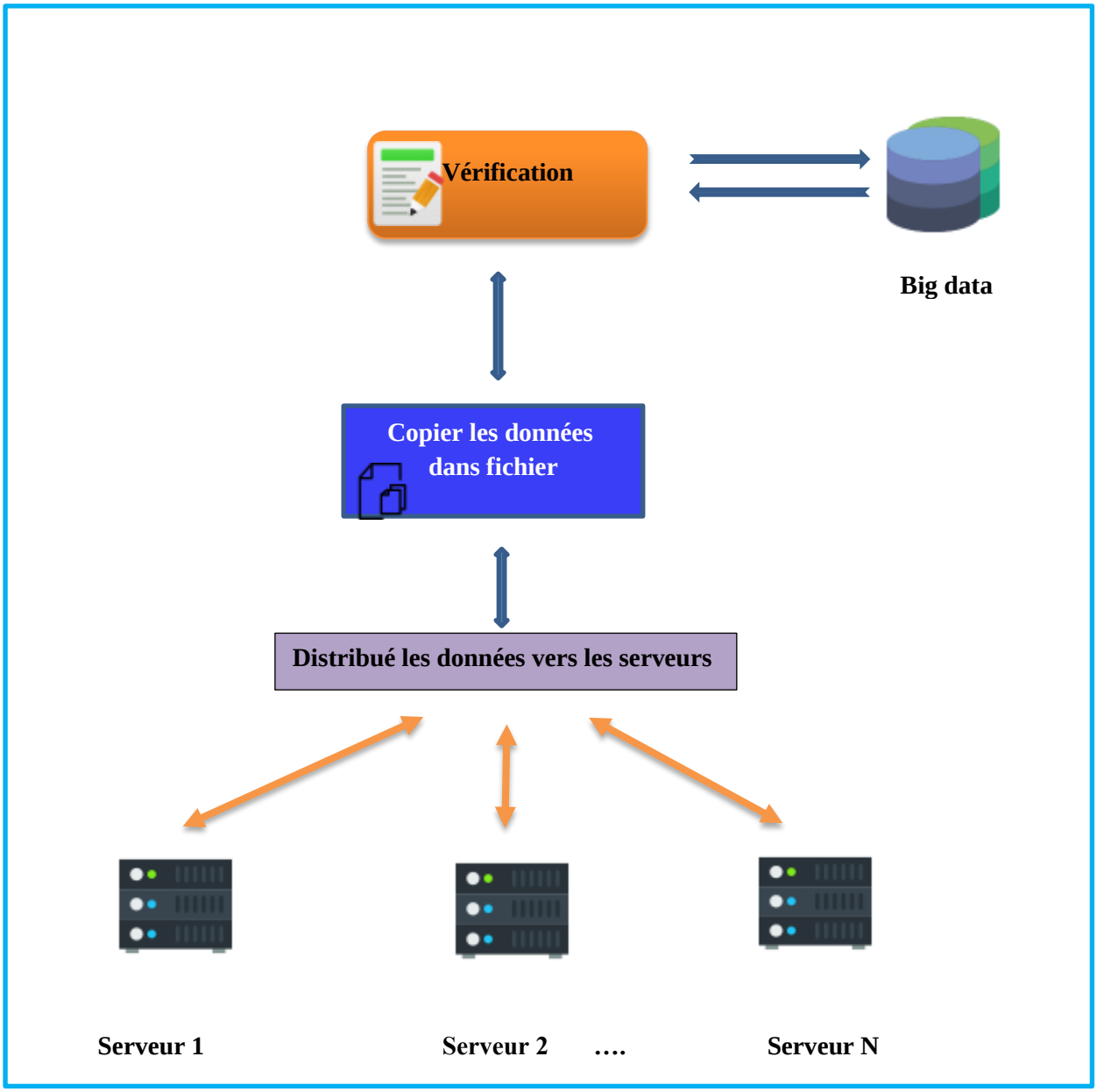

Figure III.3 : L'architecture composant réplication de données.

2.2.2.2. les rôles composant réplication

- ✓ partage d'informations pour assurer la cohérence de données entre plusieurs sources de données redondantes.
- ✓ les mêmes données sont dupliquées sur plusieurs périphériques.

- ✓ créer plusieurs sauvegardes
- ✓ vérification l'information plus consulté
- ✓ Définir des lieux pour créer des sauvegardes
- ✓ Compenser la base de données si elle n'est pas accessible

III.2.2.3. Composant Backup

2.2.3.1. L'architecture composant Backup

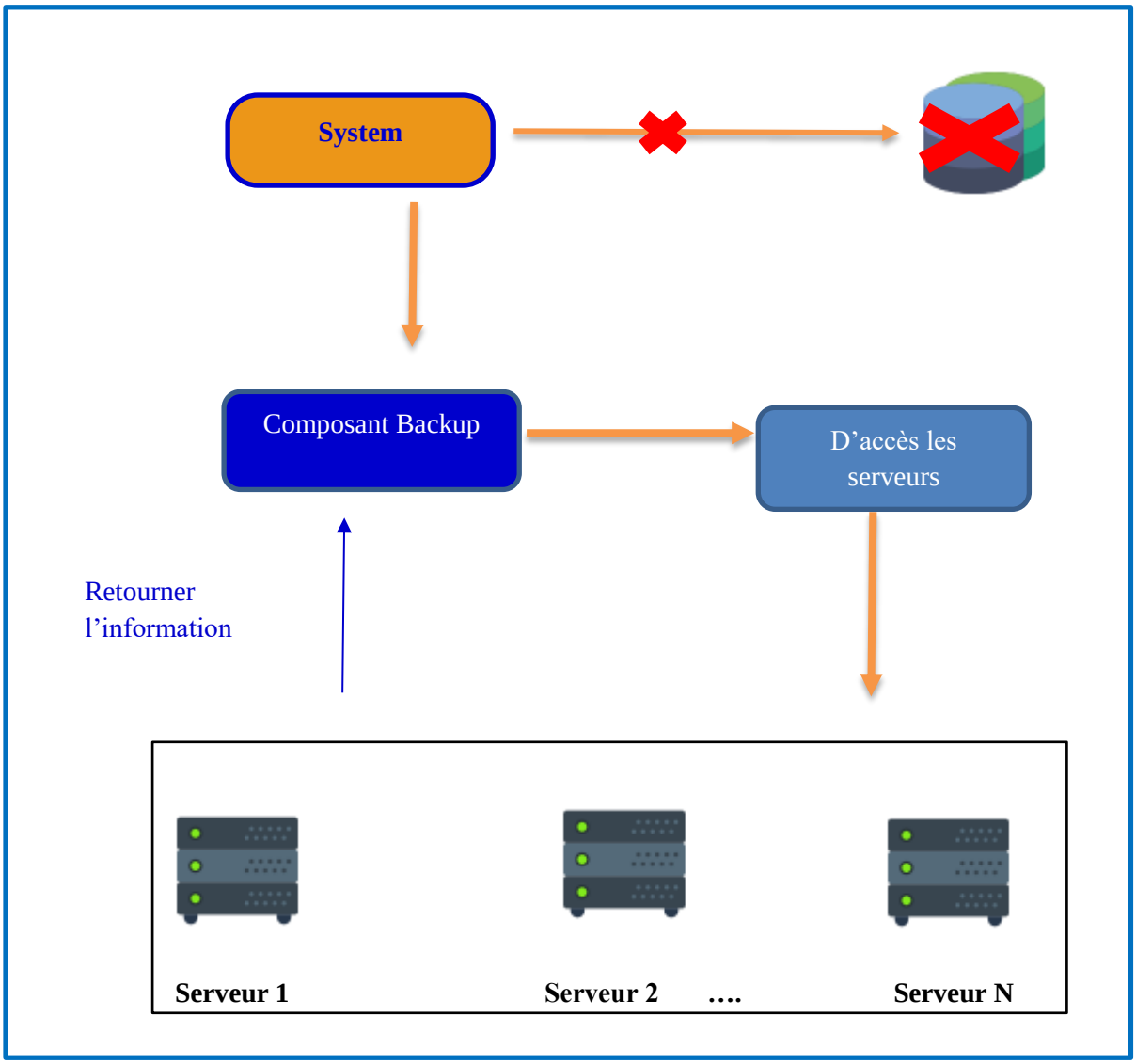

Figure III.4 : L'architecture composant Backup.

2.2.3.2. Les rôles composant Backup

- Permet d'accéder aux données stockées dans la sauvegarde
- Transférer les informations stockées dans la sauvegarde à l'utilisateur

- Mettre à jour les données dans la sauvegarde

III.2.2.4. Composant cloud

2.2.4.1. L'architecture composant cloud

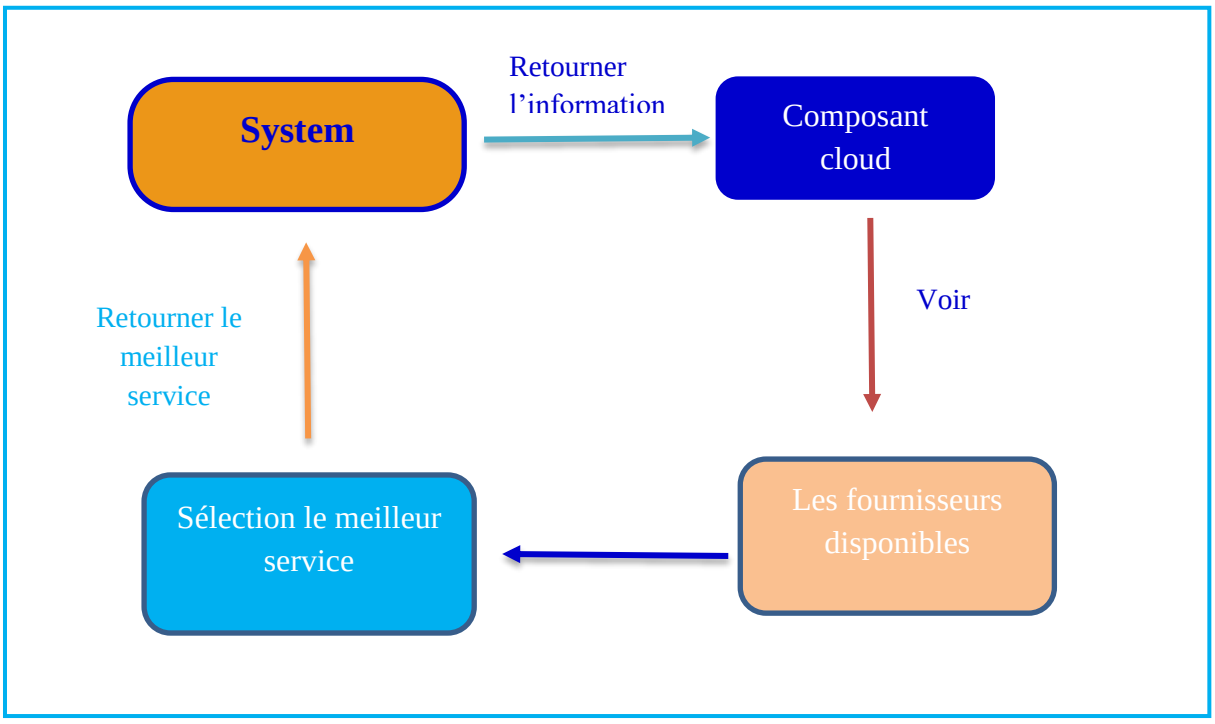

Figure III.5 : L'architecture composant cloud.

2.2.4.2. Les rôle composant cloud

- Fourniture les besoins de stockage sont bien plus importants.
- Voir les fournisseurs (service cloud) disponible
- sélection meilleur service cloud pour stocker les données

III.2.2.5. Composant supervision

2.2.5.1. L'architecture composant supervision

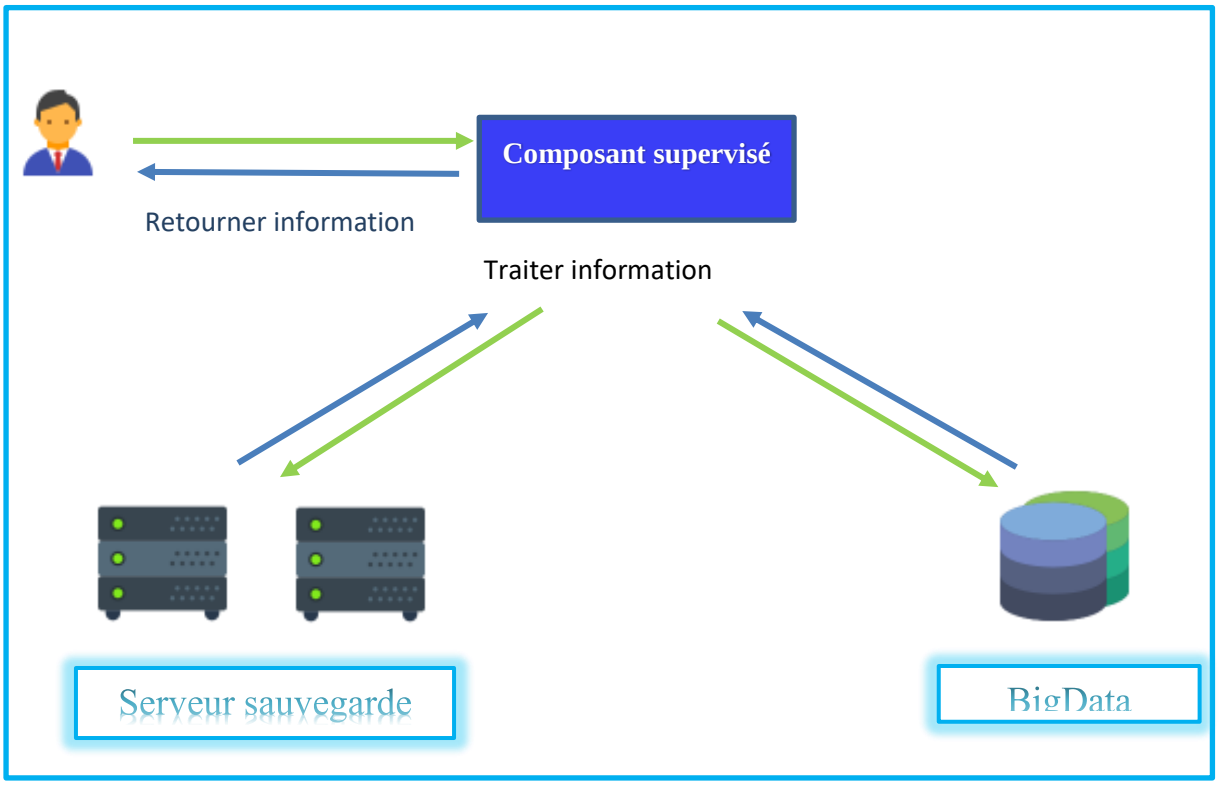

Figure III.6 : L'architecture composant supervisé.

2.2.5.2. les rôles composant supervisé

- ✓ Séparer le composant sur le système
- ✓ Connectez l'utilisateur avec la base de données et la sauvegarde
- ✓ Travaille pour disponibilité des données continues

III.3. Projection sur Hadoop

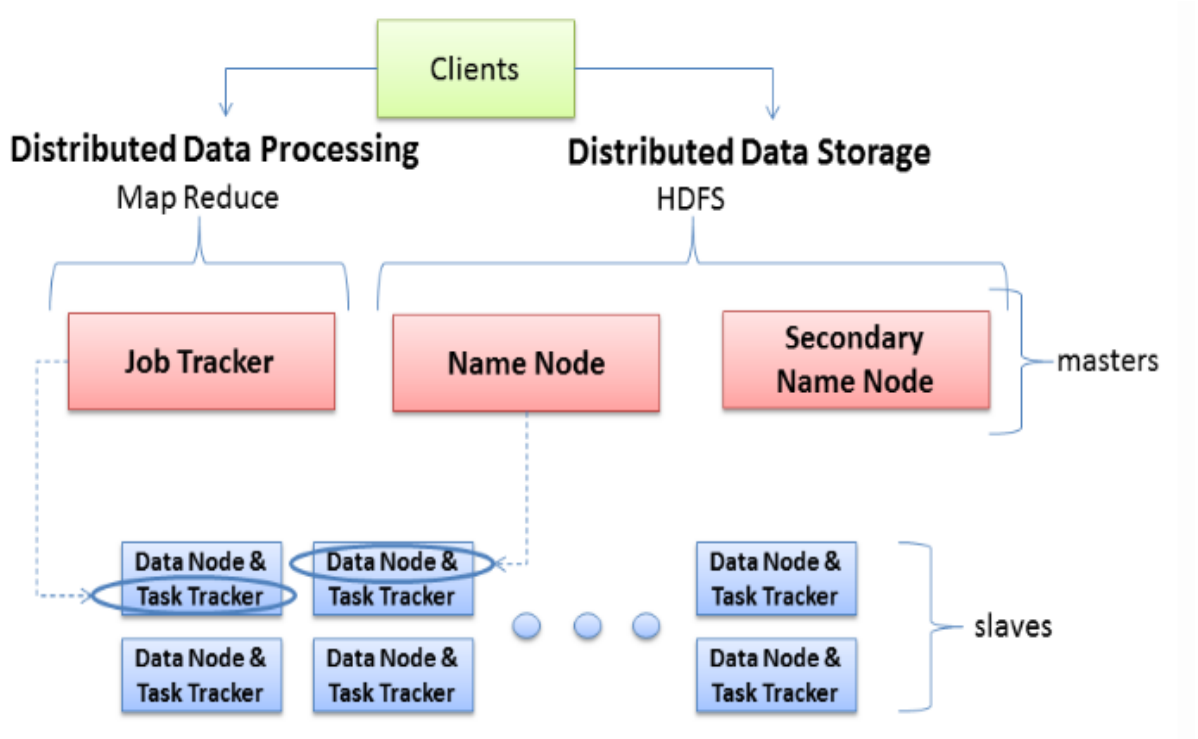

Figure III.7 : Rôles de serveur Hadoop

Les trois principales catégories de rôles de la machine dans un déploiement Hadoop sont les machines client, les nœuds *Maîtres* «Masters» et les nœuds esclaves « Slave». Les noeuds *Maîtres* supervisent les deux pièces fonctionnelles clés qui composent Hadoop : stockant beaucoup de données (HDFS) et exécutant des calculs parallèles sur toutes ces données (Map Reduce). Le Name Node supervise et coordonne la fonction de stockage de données (HDFS), tandis que Job Tracker supervise et coordonne le traitement parallèle des données à l'aide de Map Reduce. Les nœuds esclaves constituent la grande majorité des machines et fait le stockage des données et d'exécuter les calculs. Chaque slave tourne à la fois un Data Node et Task Tracker qui communiquer et de recevoir des instructions de leurs nœuds Maître.

III.3.1. NameNode

Le NameNode dans Hadoop est le noeud où Hadoop stocke toutes les informations de localisation des fichiers dans HDFS.

III.3.2. Secondary Name Node

Le secondary name node est responsable de l'exécution des fonctions d'entretien périodiques pour le NameNode. Il ne crée que des points de contrôle du système de fichiers présents dans le NameNode.

III.3.3. DataNode

Le DataNode est chargé de stocker les fichiers dans HDFS. Il gère les blocs de fichiers dans le noeud. Il envoie des informations au NameNode sur les fichiers et les blocs stockés dans ce nœud et répond à la NameNode pour toutes les opérations du système de fichiers.

III.3.4. JobTracker

JobTracker est chargé de prendre des demandes d'un client et l'attribution des Task Trackers avec les Task à effectuer. Le JobTracker tente d'assigner des tâches à Task Tracker sur le Data Node où les données sont présentes localement (Data Localité). Si cela est impossible, il va au moins essayer d'assigner des Task à TaskTrackers dans le même rack. Si, pour une raison quelconque, le node échoue au Job Tracker affecte la tâche à l'autre Task Tracker où la réplique des données existe depuis les blocs de données sont reproduits à travers les DataNodes. Cela garantit que le travail ne manque pas même si un nœud échoue au sein du cluster.

III.3.5. TaskTracker

Task Tracker accepte Task (Map,Reduce and Shuffle) de la JobTracker. Le Task Tracker continue à envoyer un message de heart beat à un Job Tracker de notifier qu'elle est vivante. Avec le rythme cardiaque il envoie aussi les emplacements libres disponibles à l'intérieur pour traiter des tâches. TaskTracker démarre et surveille le Map & Reduce Tasks et envoie progrès / informations d'état vers le Job Tracker. Le système externe travaille avec le HDFS (Name Node, Data Node), et le système interne : l’agent scanner travaille avec Secondary Name Node et Acces level agent travaille avec MapReduce (JobTracker, TaskTracker).

III.4. Conception et Modélisation détaillée avec UML

III.4.1. Diagramme de séquence général

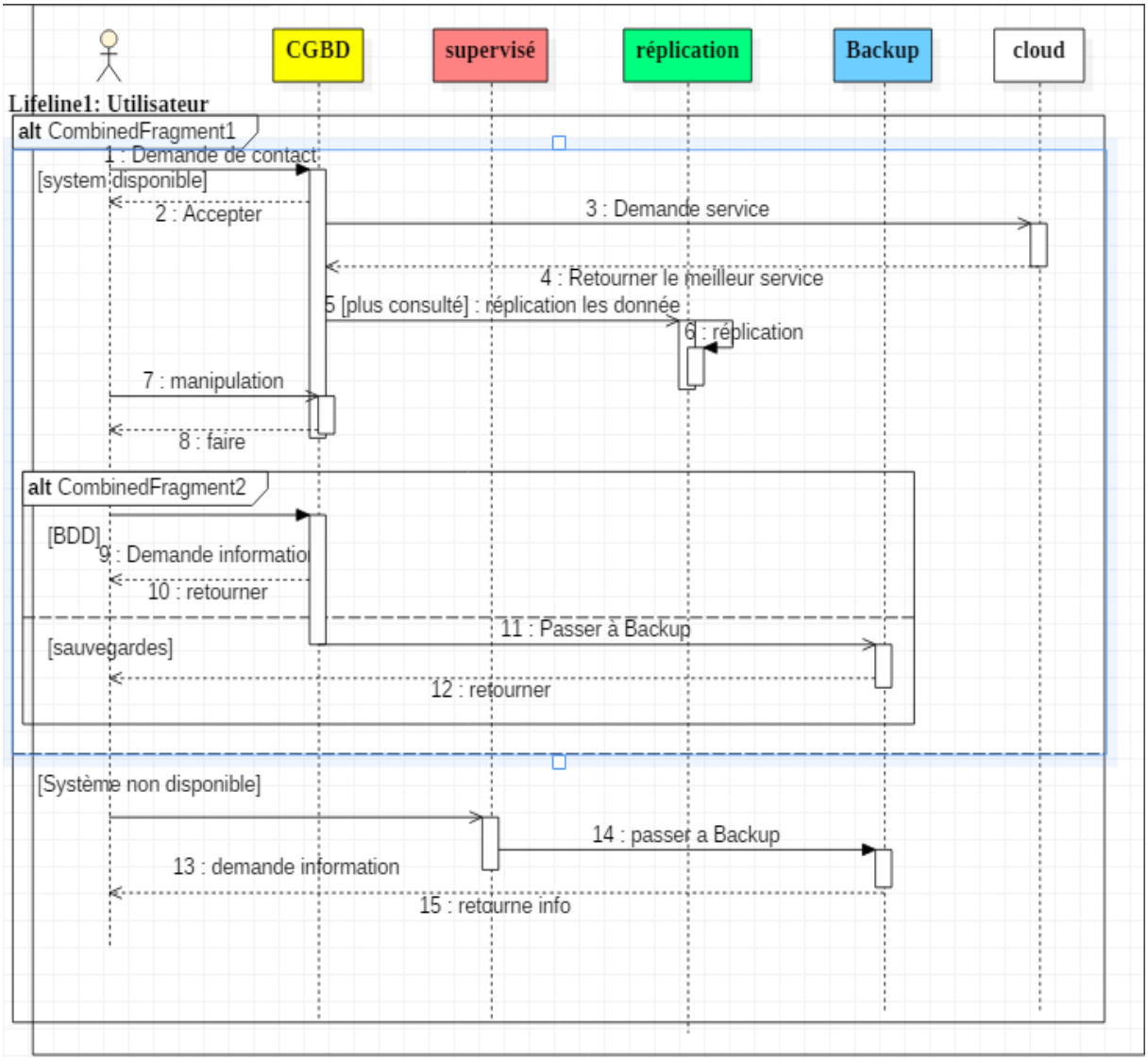

Figure III.8 : Diagramme de séquence général.

Dans un premier temps utilisateur demande connectez-vous au système, pour le manipulation de l'information. Il y a deux cas si le système est disponible ou si le système n'est pas disponible.

Si le système est disponible, l'utilisateur sauvegarde les informations ou voir les données, Se connecter à un composant gestion base de donnée afin de traiter les données, Ainsi que le composant réplication de données contrôle l'information trouvé dans les Big data. vérifier l'information la plus consulté dans Big data, En cas vérifiez la condition est fait Sauvegarde sur plusieurs disques.

Lorsque vous demandez des informations à un composant gestion base de données, il y a deux cas. si Big data disponible retourner directe si Big data non disponible passer a composant Backup pour retourner l'information de la sauvegarde.

Dans le cas où le système non disponible, L'utilisateur se connecte avec un composant supervisé qui manipule les données l'utilisateur.

III.4.2. Diagramme d'activité des composants

III.4.2.1. Diagramme d'activité de composant gestion base de données

Dans cette composant, l'utilisateur manipuler et rechercher les données et les demande au besoin.

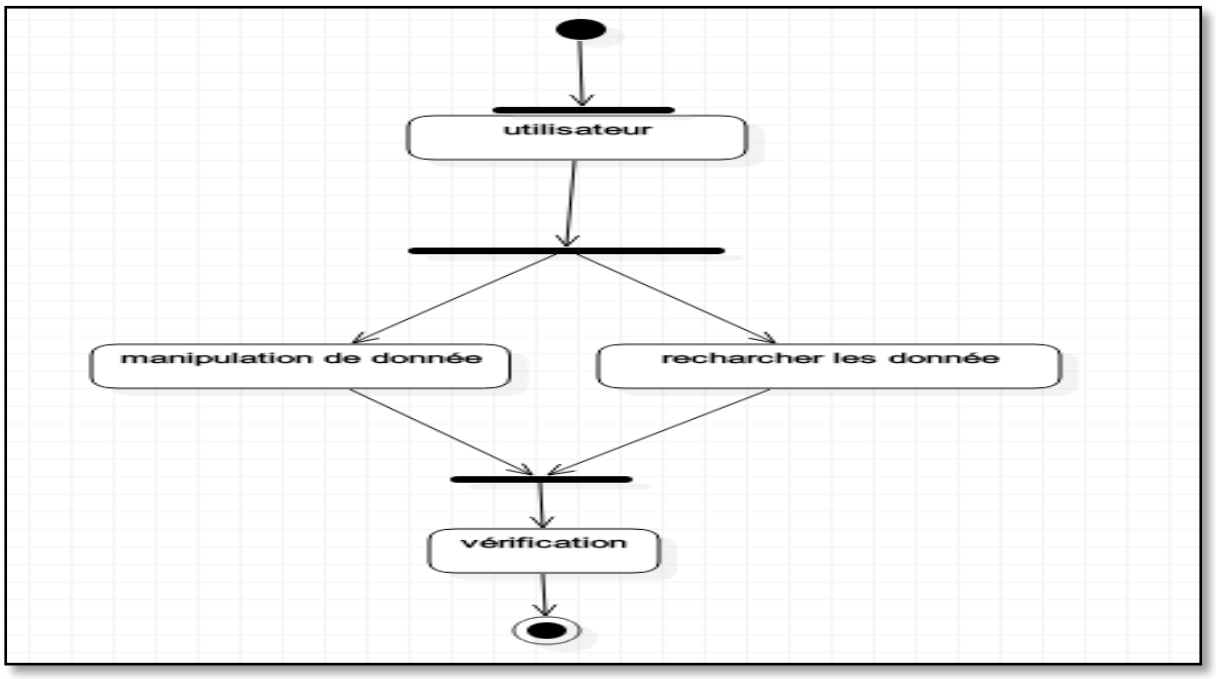

Figure III.9 : Diagramme d'activité de composant gestion base de données

III.4.2.2. Diagramme d'activité de composant réplication les donnée

Ç'est fait component contrôler les donnée la plus consulté par les utilisateurs afin de les stocker dans des sauvegardes.

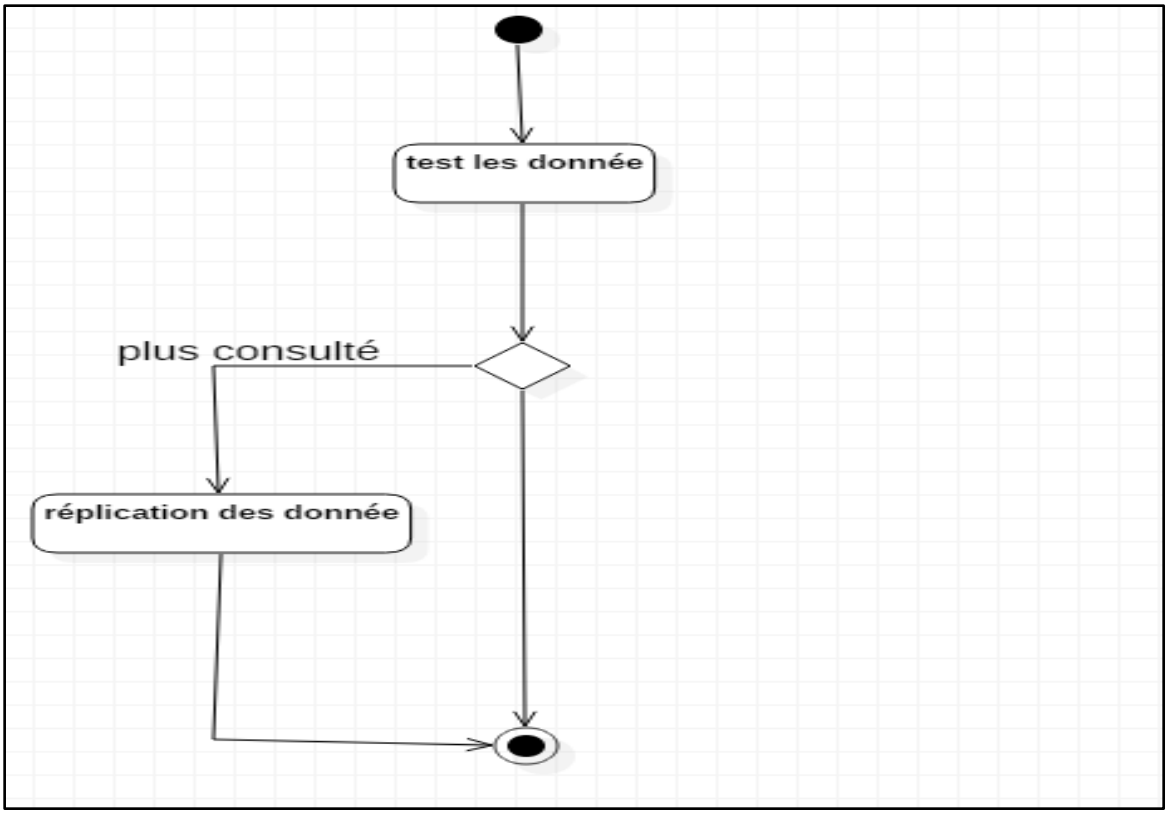

Figure III.10 : Diagramme d'activité de composant réplication des données

III.4.2.3. Diagramme d'activité de composant Backup

Dans cette composant l'utilisateur est connecté à la sauvegarde si BigData non disponible.

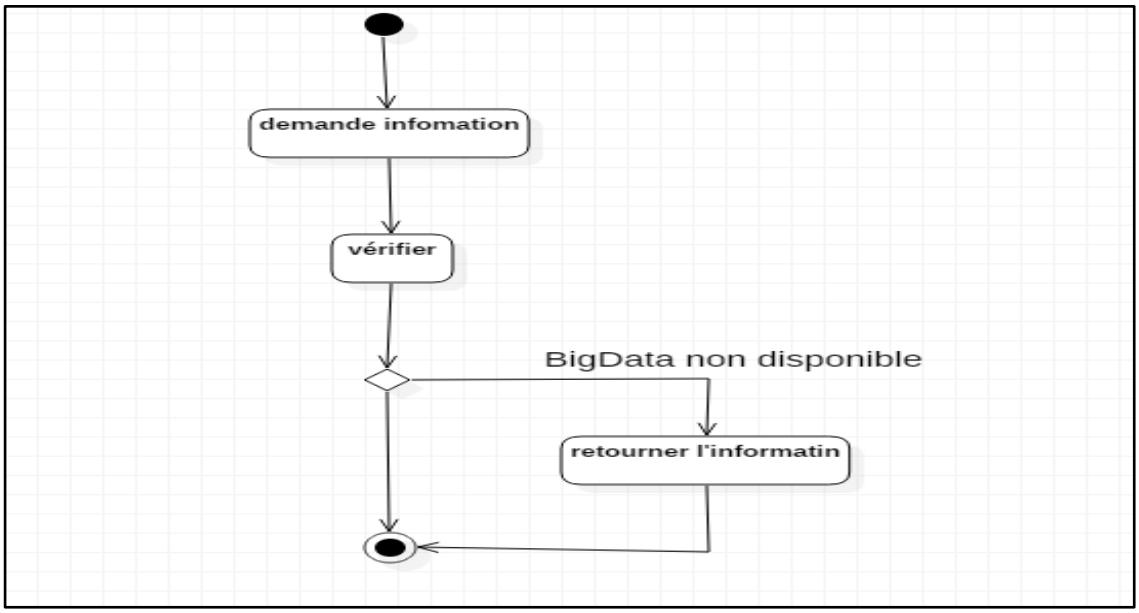

Figure III.10 : Diagramme d'activité de composant Backup

III.4.2.4. Diagramme d'activité de composant

Dans ce diagramme, il explique les étapes Choisissez le meilleur serveur cloud et voir les fournisseurs disponible pour dans cette service.

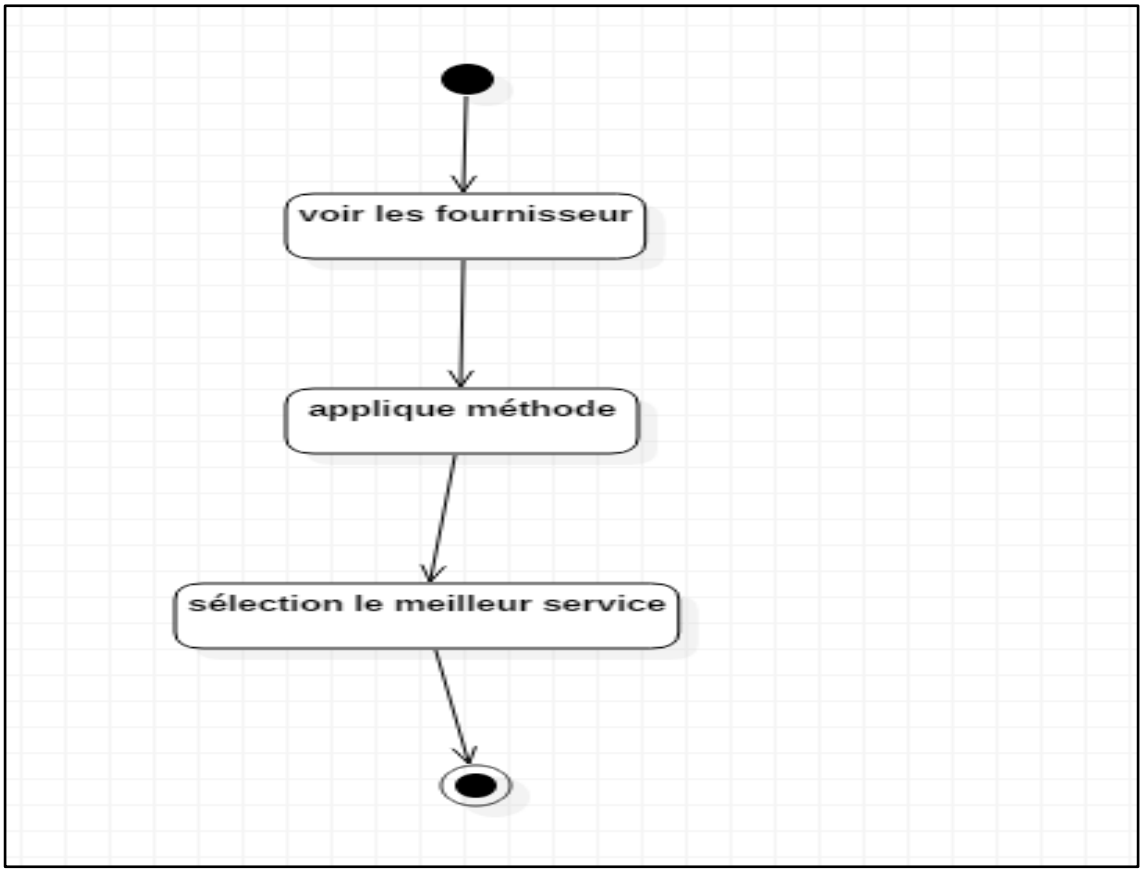

Figure III.10 : Diagramme d'activité de composant cloud

III.4.2.5. Diagramme d'activité de composant supervision

dans ce composant lance une copie du système en cas de panne du système

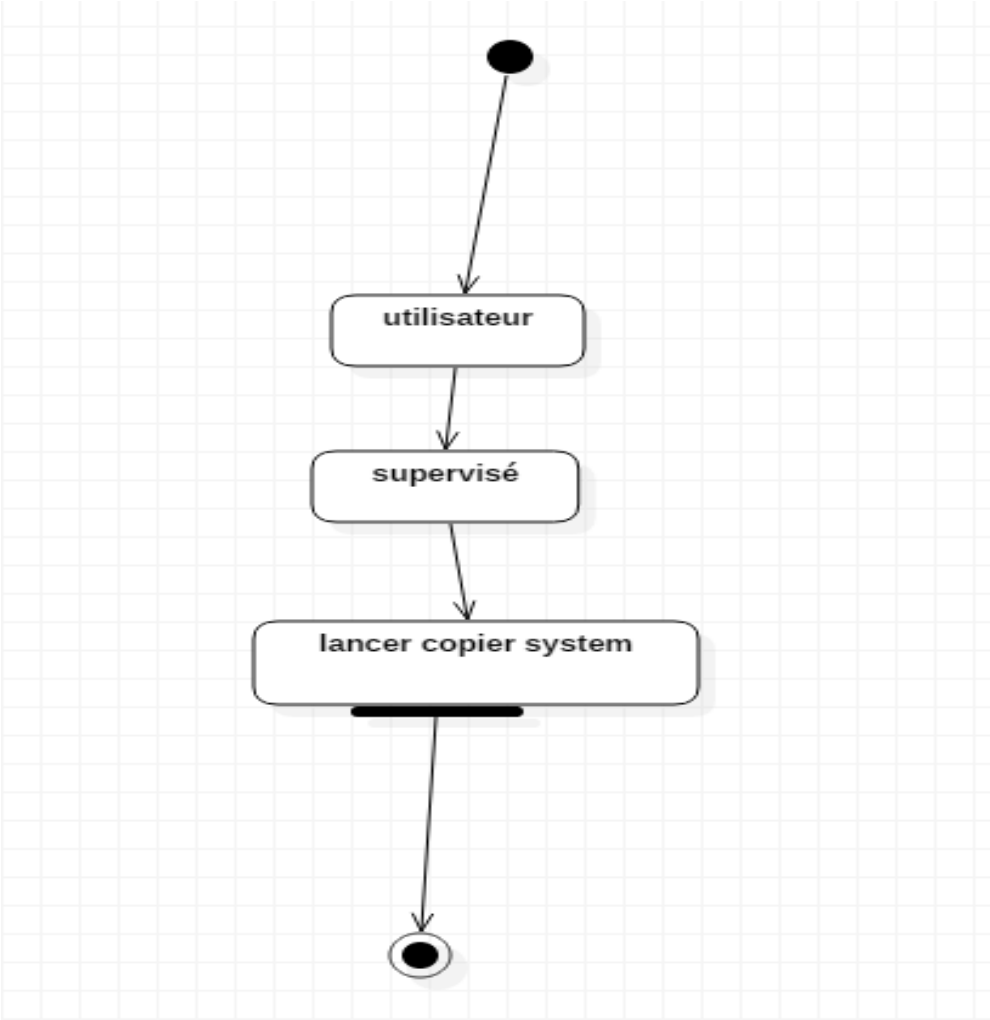

Figure III.10 : Diagramme d'activité de composant cloud

III.5. Conclusion

Dans ce chapitre nous avons présenté notre système de pour l'assurance de la disponibilité des BigData. Notre architecture est composée d'un ensemble de composants. Aussi bien architecture détaillée et les role chaque composant.

A été créé diagramme séquence de système et les diagrammes d'activité choque composants.

Chapitre IV
Implémentation

IV.1. Introduction

Le chapitre suivant est consacré à la description des détails d'implémentation d'architecture de l'intégrité dans le big data. Nous commençons par la présentation des langages de programmation et les outils de développement utilisés pour la mise en œuvre du système conçu dans le chapitre précédent. Nous donnons par la suite une description textuelle et graphique de quelques interfaces du système réalisé puis l'architecture de système, la description des interfaces graphiques et enfin les principaux codes source.

IV.2. Outils et langages de programmation utilisés

Pour la résiliation du système d'intégrité dans le big data, nous avons utilisé un langage de programmation, et quelques environnements de développement. Nous les décrivons brièvement dans les sous sections suivantes.

IV.2.1. Langages de programmation

2.1.1. Java

Java est un langage de programmation informatique orienté objet créé par James Gosling et Patrick Naughton, employés de Sun Microsystems, avec le soutien de Bill Joy (cofondateur de Sun Microsystems en 1982), présenté officiellement le 23 mai 1995 au SunWorld.

La société Sun a été ensuite rachetée en 2009 par la société Oracle qui détient et maintient désormais Java.

La particularité et l'objectif central de Java est que les logiciels écrits dans ce langage doivent très facilement portables sur plusieurs systèmes d'exploitation tels que UNIX, Windows, Mac OS ou GNU/Linux, avec peu ou pas de modifications. Pour cela, divers plateformes et frameworks associés visent à guider, sinon garantir, cette portabilité des applications développées en Java. Outre son orientation objet, le langage Java a l'avantage d'être modulaire (on peut écrire des portions de code génériques, c'est-à-dire utilisables par plusieurs applications), rigoureux (la plupart des erreurs se produisent à la compilation et non à l'exécution) et portable (un même programme compilé peut s'exécuter sur différents environnements).

Java est un langage interprété, ce qui signifie qu'un programme compilé n'est pas directement exécutable par le système d'exploitation mais il doit être interprété par un autre programme, qu'on appelle interpréteur.

Un programmeur Java écrit son code source, sous la forme de classes, dans des fichiers dont l'extension est .java . Ce code source est alors compilé par le compilateur javac en un langage appelé bytecode et enregistre le résultat dans un fichier dont l'extension est .class. Le byecode ainsi obtenu n'est pas directement utilisable. Il doit être interprété par la machine virtuelle de Java qui transforme alors le code compilé en code machine compréhensible par le système d'exploitation. C'est la raison pour laquelle Java est un langage portable : le bytecode reste le même quelque soit l'environnement d'exécution.

Figure IV.1 : Logo de java

IV.2.2. Outils de développement

2.2.1. Netbeanse

Netbeanse IDE : est un environnement de développement intégré (EDI), placé en open source par Sun en juin 2000 sous licence CDDL (Common Development and Distribution License) et GPLv2. En plus de Java, NetBeans permet également de supporter différents autres langages, comme C, C++, JavaScript, XML, Groovy, PHP et HTML de façon native ainsi que bien d'autres (comme Python ou Ruby) par l'ajout de greffons. Il comprend toutes les caractéristiques d'un IDE moderne (éditeur en couleur, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages Web).

Figure IV. 2 : Logo de netbeanse

2.2.2. Hadoop

Hadoop est un framework libre et open source écrit en Java destiné à faciliter la création d'applications distribuées (au niveau du stockage des données et de leur traitement) et échelonnables (scalables) permettant aux applications de travailler avec des milliers de nœuds et des pétaoctets de données. Ainsi chaque nœud est constitué de machines standard regroupées en grappe. Tous les modules de Hadoop sont conçus dans l'idée fondamentale que les pannes matérielles sont fréquentes et qu'en conséquence elles doivent être gérées automatiquement par le framework.

Hadoop a été inspiré par la publication de MapReduce, GoogleFS et BigTable de Google. Hadoop a été créé par Doug Cutting et fait partie des projets de la fondation logicielle Apache depuis 2009.

Le noyau d'Hadoop est constitué d'une partie de stockage: HDFS (Hadoop Distributed File System), et d'une partie de traitement appelée MapReduce. Hadoop fractionne les fichiers en gros blocs et les distribue à travers les nœuds du cluster.

Figure IV.3 : Logo d'hadoop

2.2.3. XAMPP

XAMPP est un ensemble de logiciels permettant de mettre en place facilement un serveur Web local, un serveur FTP et un serveur de messagerie électronique. Il s'agit d'une distribution de logiciels libres (X (cross) Apache MariaDB Perl PHP) offrant une bonne souplesse d'utilisation, réputée pour son installation simple et rapide. Ainsi, il est à la portée d'un grand nombre de personnes puisqu'il ne requiert pas de connaissances particulières et fonctionne, de plus, sur les systèmes d'exploitation les plus répandus.

Figure IV.4 : Logo de XAMPP.

2.2.4. MySQL

MySQL est un système de gestion de bases de données relationnelles (SGBDR). Il fait partie des logiciels de gestion de base de données les plus utilisés au monde. MySQL fait référence au Structured Query Language, le langage de requête utilisé.

Figure IV.5 : Logo de mysql.

2.2.5. phpMyAdmin

PhpMyAdmin est une interface d'administration pour le SGBD MySQL. Il est écrit en langage PHP et s'appuie sur le serveur HTTP Apache.

Figure IV.6 : Logo de Phpmyadmin.

IV.3. Description des Interfaces Graphiques

IV.3.1. interface accès a system

Cette interface est pour que l'utilisateur décide s'il veut entrer en mode utilisateur ou en mode admin

Figure. VI.7 Interface accès a system

IV.3.2. Interface d'authentification

Tout utilisateur doit être doté d'un compte. Le composant de control d'accès vérifie la requête de l'utilisateur dans la base de données des comptes avant toute autorisation d'accès.

L'utilisateur reçoit un message d'erreur si le mot de passe ou le nom d'utilisateur incorrecte, et le administrateur ne nécessite pas d'être dans la base de données parce qu'il y a un administrateur pour le système.

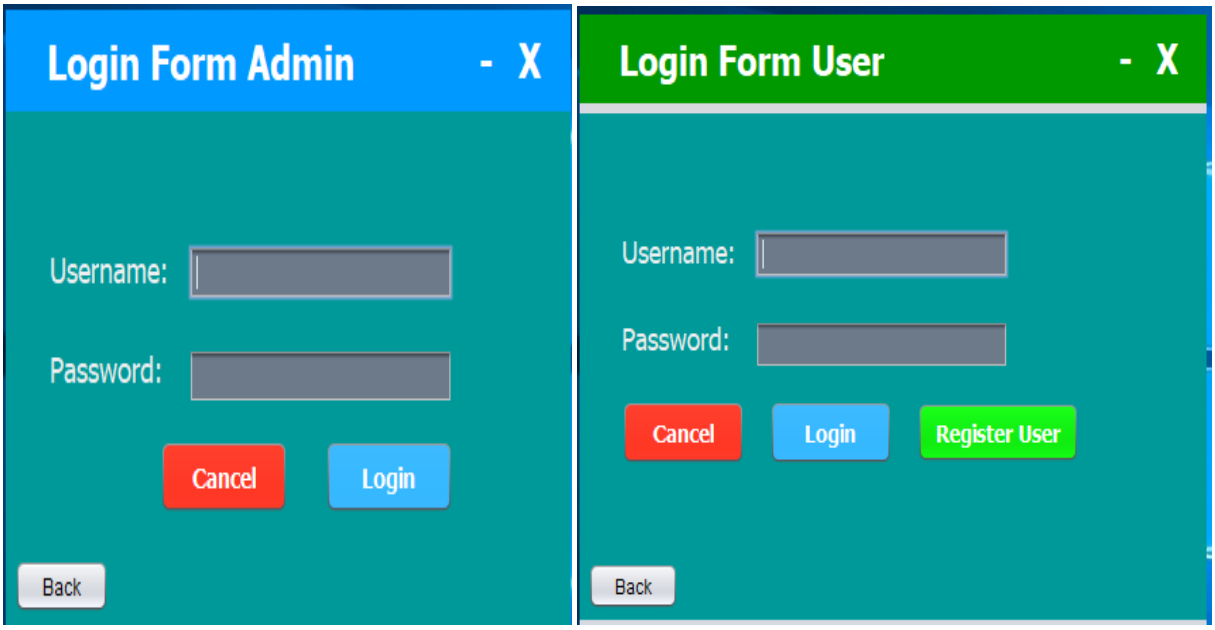

Figure. VI.8: Interface d'authentification

Ainsi que l'utilisateur peut inscription dans le système

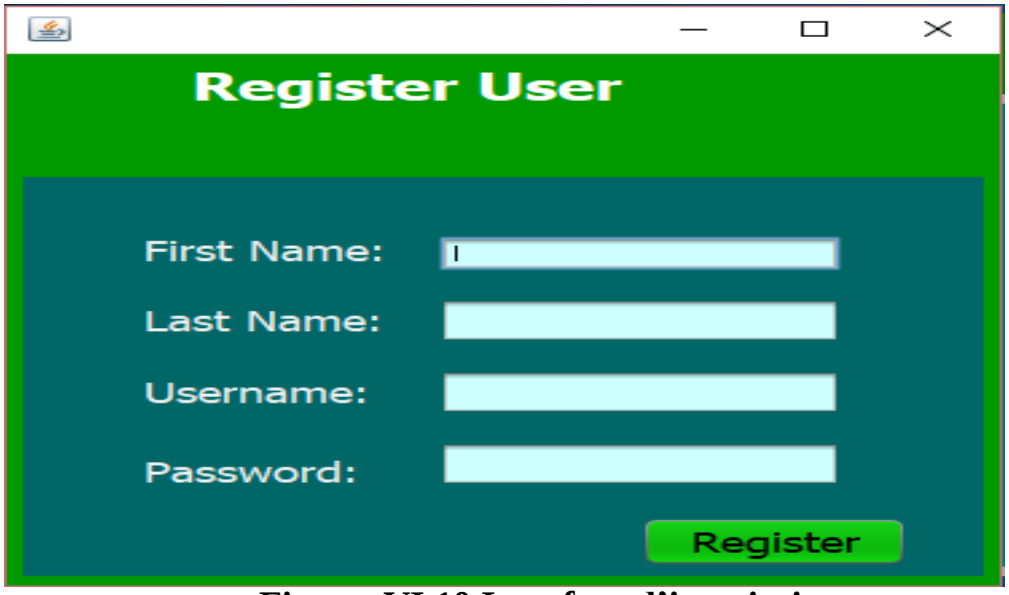

Figure. VI.10 Interface d'inscription

IV.3.3. Interface User

Cette interface est pour que user ajouter, supprimer, modifier ou visualiser les informations patient

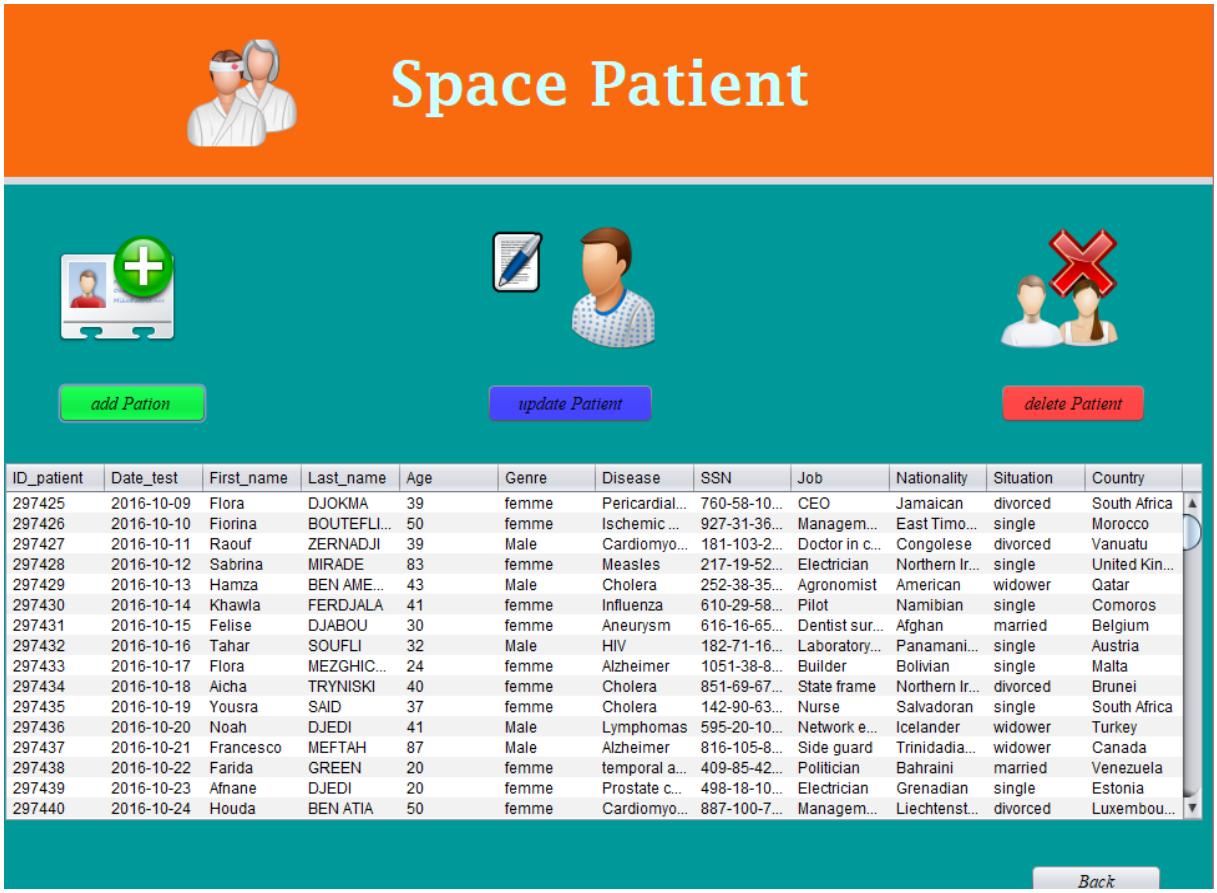

Figure. VI.11 Interface User

IV.3.4. Interface Ajouter des informations patientes

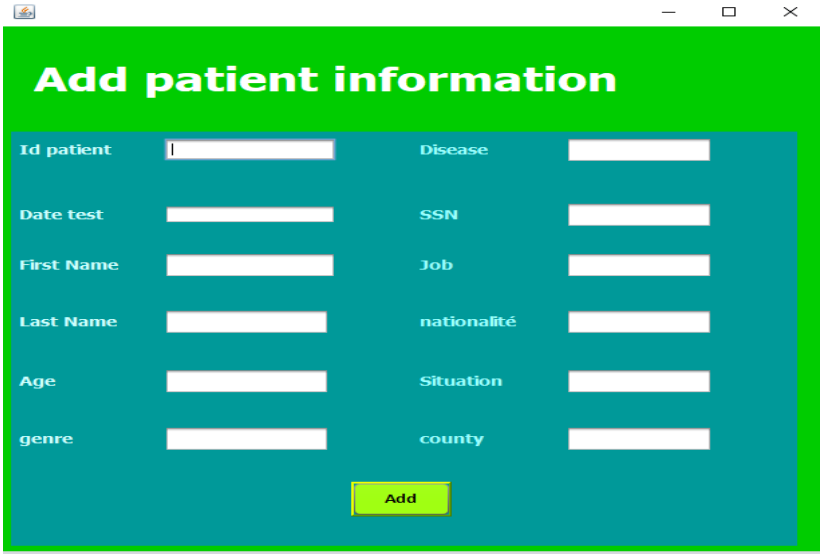

Figure. VI.12 Interface ajouté patient

IV.3.5. Interface modifier information patient

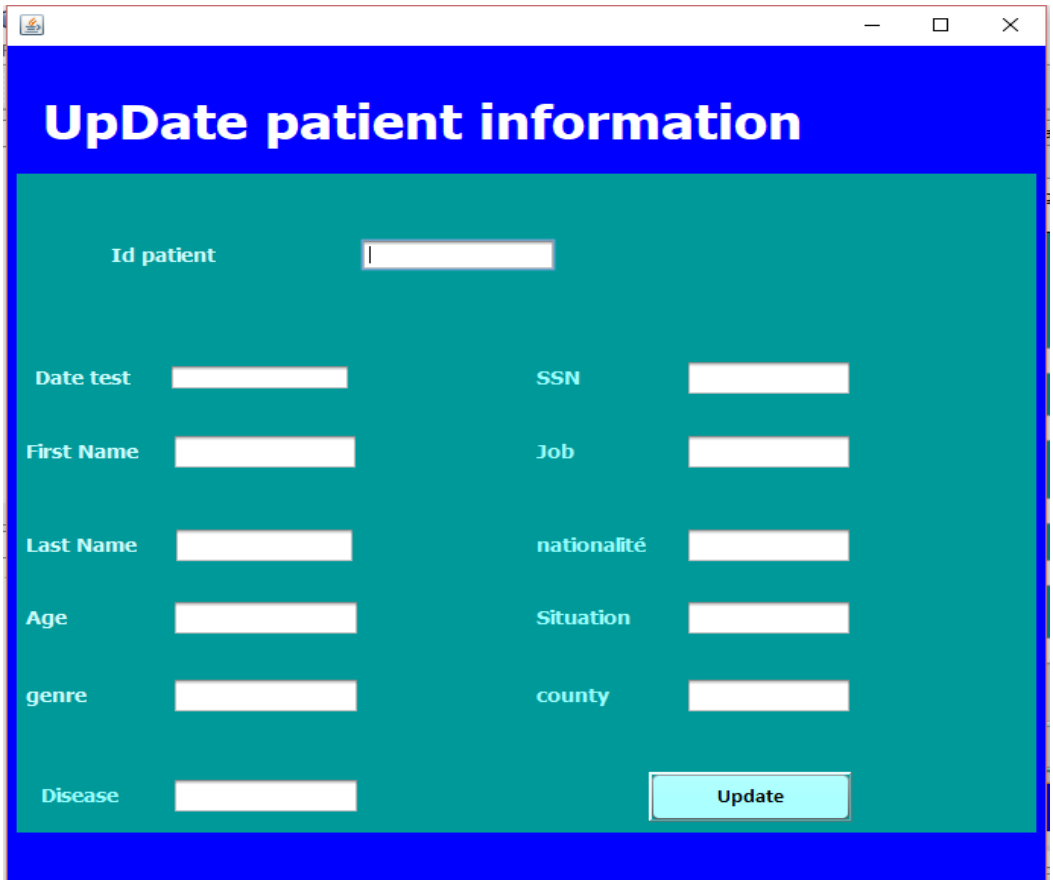

Figure. VI.13 Interface modifié patient

IV.3.6. Interface supprimé patient

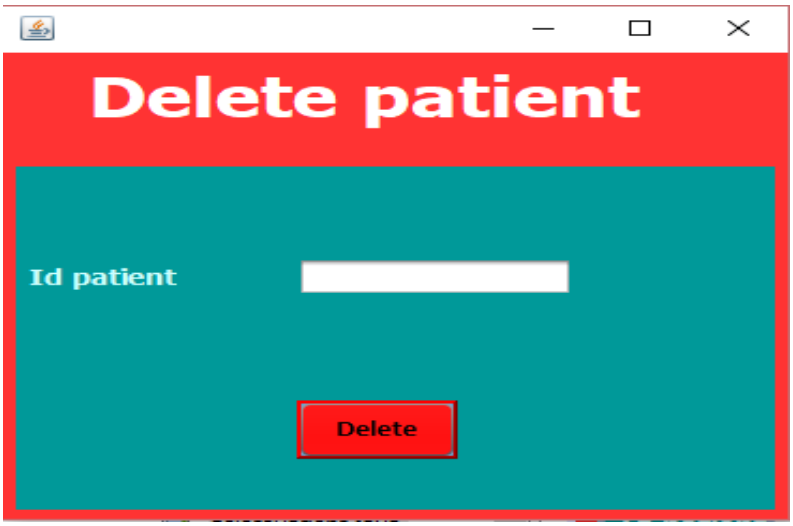

Figure. VI.13 Interface supprimé patient

IV.3.7. Interface de l'administrateur

Dans cette interface, l'administrateur peut se connecter à tous les tache système espace Users , espace patient ,Backup , réplication les données et les fournisseurs.

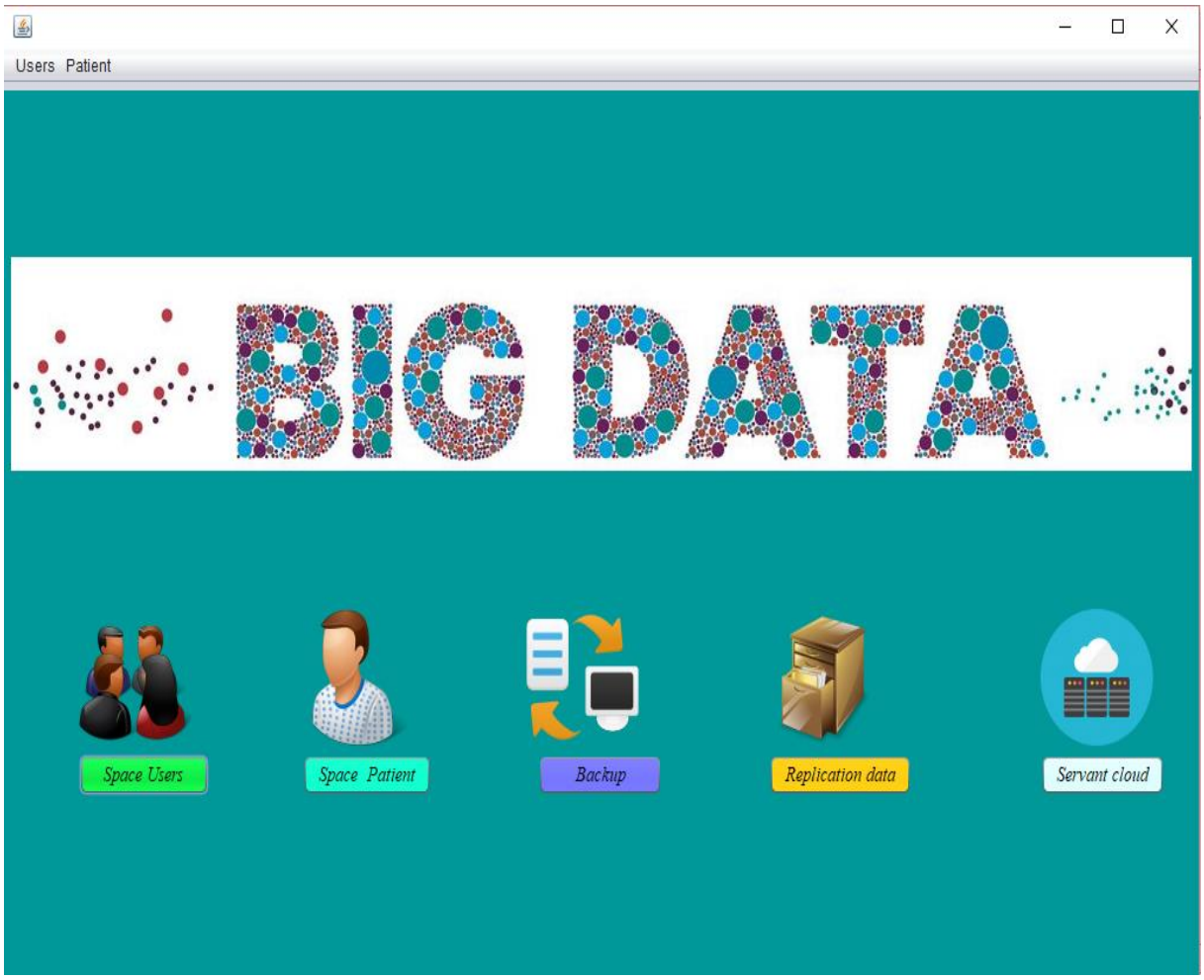

Figure. VI.14 Interface de l'administrateur

IV.3.8. Interface Backup

Dans cette interface, l'administrateur peut connecter tous les fichiers stockés dans la sauvegarde afin de les visualiser

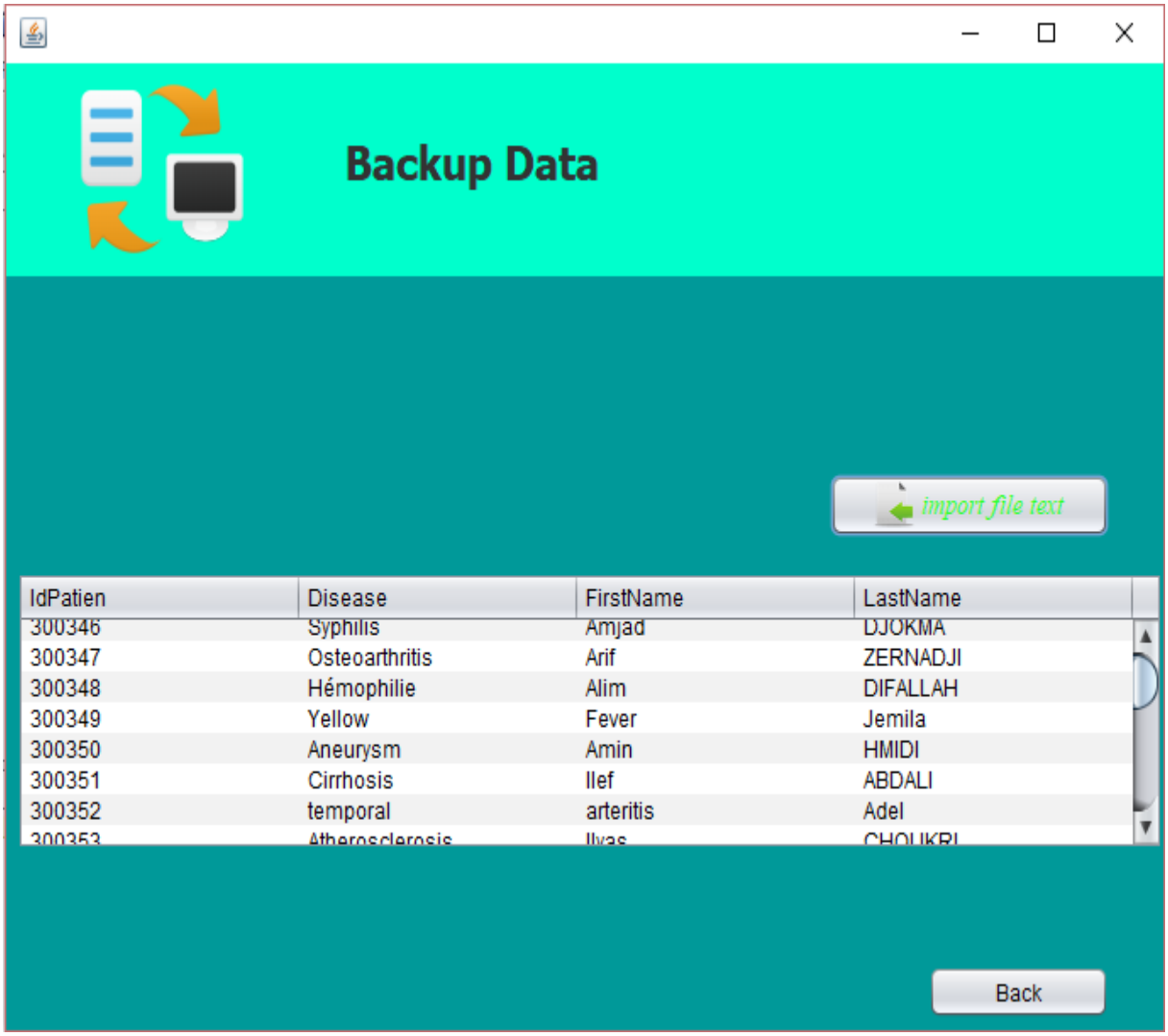

Figure. VI.15 Interface Backup

IV.3.9. Interface réplication les données

Dans cette interface, vous pouvez rechercher des informations patientes et choisir où enregistrer vos résultats de recherche.

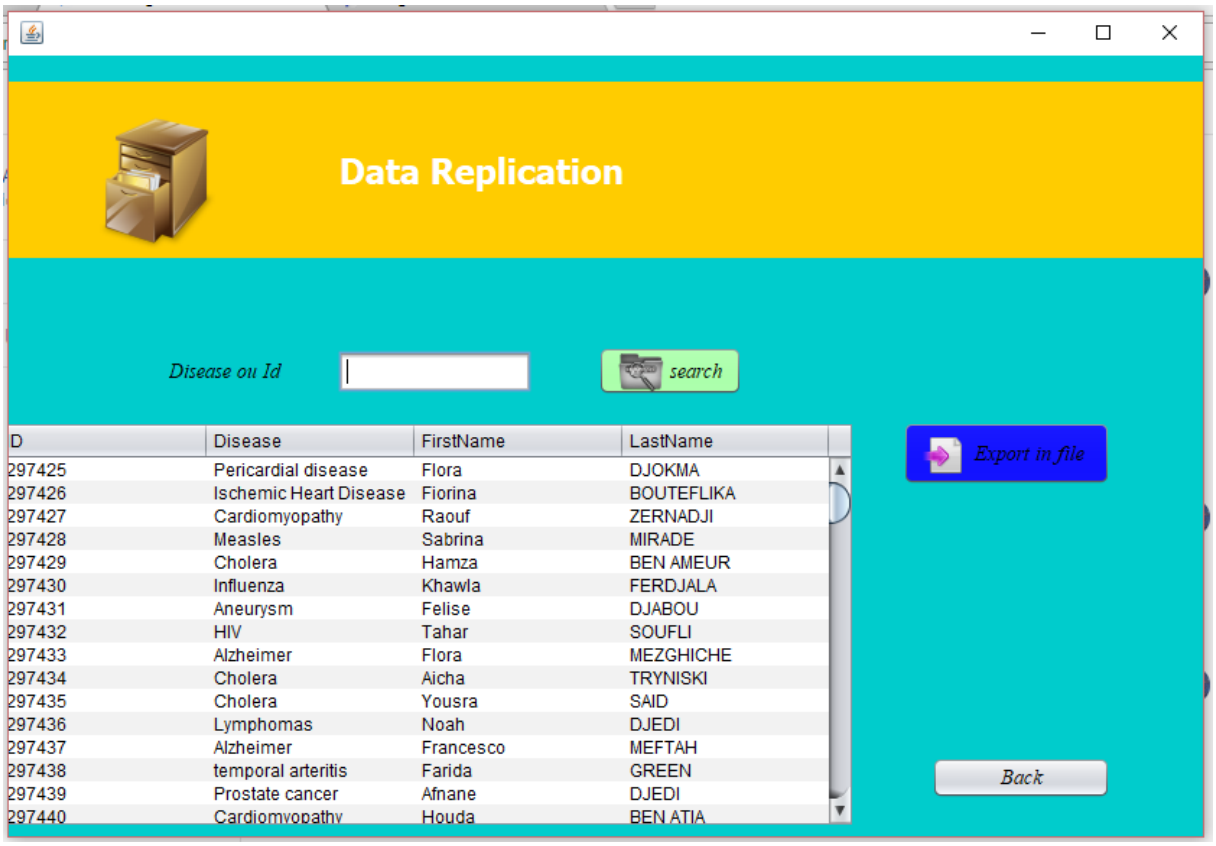

Figure. VI.16 Interface Backup

IV.3.10. Interface supervision

Cette interface explique comment L'administrateur utilise le composant de supervision pour redémarrer le système.

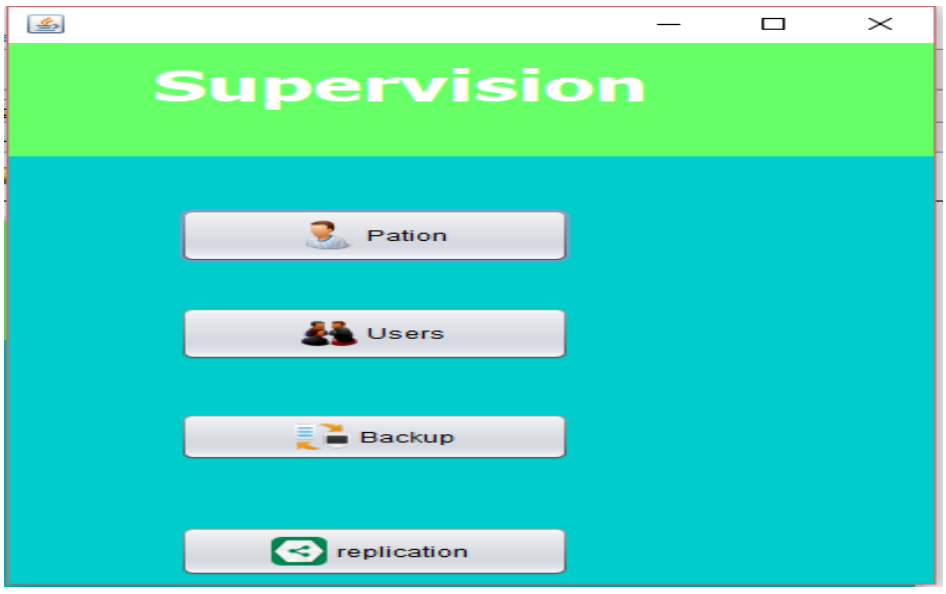

Figure. VI.17 Interface supervision

IV.3.11. Interface cloud service

Cette interface voir les fournisseurs disponible et sélection la meilleur service cloud.

Figure. VI.18 Interface supervision

IV.3.12. L'interface de la base de données

3.12.1. Tableau d'information patient

Ce tableau contient 12 colonne (ip patient , date test, firstname , lastname, age, genre, disease, SSN, job, nationality, situation et county) et sur plus de 150 000 lignes. C'est ce qui fait cette base Big Data.

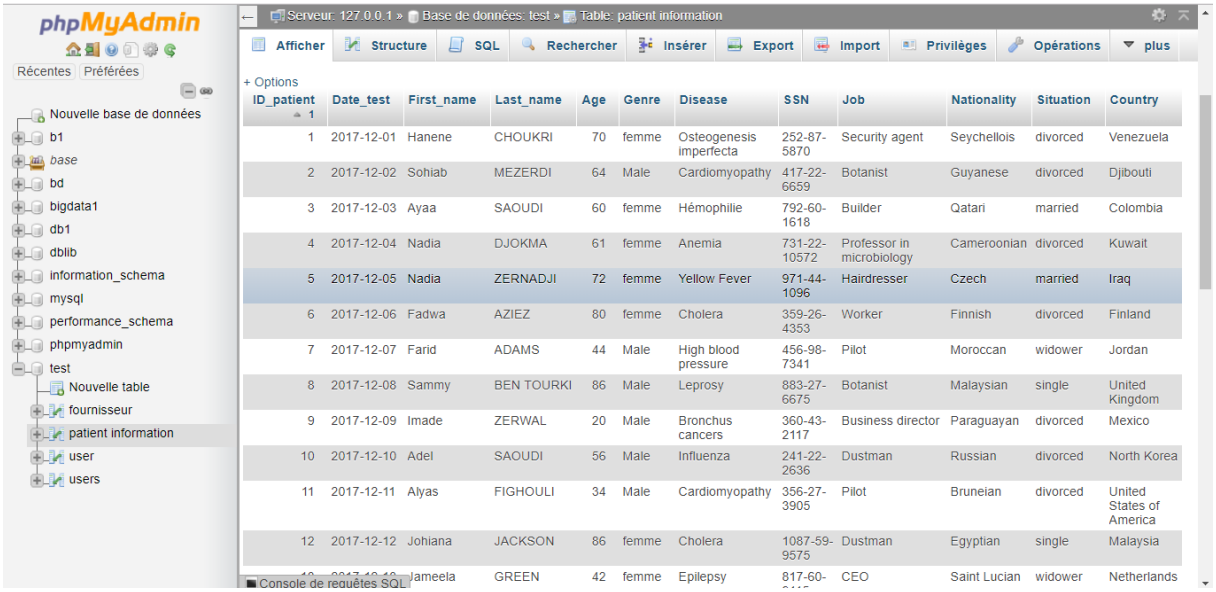

Figure. VI.19 Tableau d'information patient

3.12.2. Tableau fournisseurs

tableau les fournisseurs pour l'aide de stockage les données, cette table contient 4 colonnes (id fournisseur , nome fournisseur , CUP fournisseur et prix) et 10 fournisseurs.

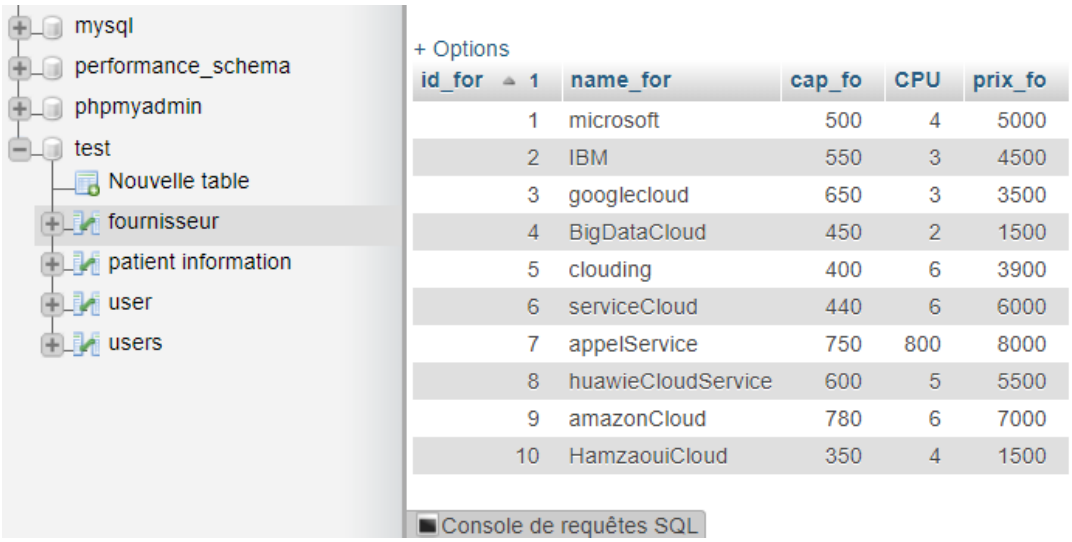

Figure. VI.20 Tableau fournisseurs

IV.4. Les principaux codes source

IV.4.1. Connection avec base de données

```
*/
public class MyConnection {


    // create a function to connect with mysql database

    public static Connection getConnection(){

        Connection con = null;
        try {
            Class.forName("com.mysql.jdbc.Driver");
            con = DriverManager.getConnection("jdbc:mysql://localhost/test","root","");
        } catch (Exception ex) {
            System.out.println(ex.getMessage());
        }

        return con;
    }

}
```

IV.4.2. Fonction de rechercher

```
public ArrayList<InfoPatient> ListUsers(String ValToSearch)
 {
     ArrayList<InfoPatient> usersList = new ArrayList<InfoPatient>();

     Statement st;
     ResultSet rs;

     try{
         Connection con = getConnection();
         st = con.createStatement();
         String searchQuery = "SELECT * FROM `patient information` "
                 + "WHERE CONCAT(`ID_patient`, `Disease`) LIKE '%"+ValToSearch+"%'";
         rs = st.executeQuery(searchQuery);

         InfoPatient infoPatient;

         while(rs.next())
         {
             infoPatient = new InfoPatient(
                                 rs.getInt("ID_patient"),
                                 rs.getString("Disease"),
                                 rs.getString("First_name"),
                                 rs.getString("Last_name")
                               );
             usersList.add(infoPatient);
         }
     }catch(Exception ex){
         System.out.println(ex.getMessage());
```

```
public void findUsers()
 {
     ArrayList<InfoPatient> users = ListUsers(jText_Search.getText());
     DefaultTableModel model = new DefaultTableModel();
     model.setColumnIdentifiers(new Object[]{"ID","Disease","FirstName","LastName"});
     Object[] row = new Object[4];

     for(int i = 0; i < users.size(); i++)
     {
         row[0] = users.get(i).getId();
         row[1] = users.get(i).getFname();
         row[2] = users.get(i).getFnam();
         row[3] = users.get(i).getlname();

         model.addRow(row);
     }
    jTable1.setModel(model);

 }
```

IV.4.3. Code exporte dans fichier

```
    private void jButton3ActionPerformed(java.awt.event.ActionEvent evt) {

         JFileChooser chooser = new JFileChooser();
chooser.setAcceptAllFileFilterUsed(false);
FileNameExtensionFilter filter = new FileNameExtensionFilter("Text Files","txt","text");
chooser.setFileFilter(filter);
int returnVal = chooser.showOpenDialog(this.view);
if (returnVal == JFileChooser.APPROVE_OPTION)
filePath = chooser.getSelectedFile().getAbsolutePath();
        File file = new File(filePath);
        try {
            FileWriter fw = new FileWriter(file);
            BufferedWriter bw = new BufferedWriter(fw);

            for(int i = 0; i < jTable1.getRowCount(); i++){//rows
                for(int j = 0; j < jTable1.getColumnCount(); j++){//columns
                    bw.write(jTable1.getValueAt(i, j).toString()+" ");
                }
                bw.newLine();
            }

            bw.close();
            fw.close();

        } catch (IOException ex) {
            Logger.getLogger(recharcher.class.getName()).log(Level.SEVERE, null, ex);
        }
    }
```

IV.4.4. Code copier fichier

```
protected static void ReadWrite(File file1,File file2) throws Exception{
     FileWriter fileWrite = new FileWriter(file2 );
      FileReader fileReader = new FileReader(file1);

      int i=0;
 while((i=fileReader.read())!= -1){
     fileWrite.write(i);
 }
 fileWrite.flush();
 fileWrite.close();
 fileReader.close();

 }
```

IV.4.5. Code Import fichier

```
JFileChooser chooser = new JFileChooser();
chooser.setAcceptAllFileFilterUsed(false);
FileNameExtensionFilter filter = new FileNameExtensionFilter("Text Files","txt","text");
chooser.setFileFilter(filter);
int returnVal = chooser.showOpenDialog(this.view);
if (returnVal == JFileChooser.APPROVE_OPTION)
filePath = chooser.getSelectedFile().getAbsolutePath();
  File file = new File(filePath);

       try {
           FileReader fr = new FileReader(file);
           BufferedReader br = new BufferedReader(fr);

           DefaultTableModel model = (DefaultTableModel)jTable1.getModel();
           Object[] lines = br.lines().toArray();

           for(int i = 0; i < lines.length; i++){
               String[] row = lines[i].toString().split(" ");
               model.addRow(row);
           }

       } catch (FileNotFoundException ex) {
           Logger.getLogger(BackupPage.class.getName()).log(Level.SEVERE, null, ex);
       }
```

IV.5. Conclusion

Dans ce chapitre nous avons présenté les étapes de la mise en œuvre de notre projet avec tous les outils, les langages et les plateformes utilisés ainsi que la présentation avec l'explication du rôle de chaque outil. Nous avons illustré les interfaces graphiques avec une description textuelle.

Conclusion générale

L'ère numérique est caractérisée par une croissance exponentielle de la création de données . Alors que l'on parlait il y a peu de gigaoctets, on parle plutôt de téraoctets, de pétaoctets, d'exaoctets et même de zettaoctets des données.

En plus de l'aspect « Volume », les big Data se définissent par quatre autres « grands V » : Vélocité, Variété, Véracité, Valeur. La vélocité désigne tout d'abord la vitesse à laquelle il est possible d'actualiser les analyses de données numériques. Par ailleurs, on ne traite plus que des données préalablement échantillonnées et structurées, mais on peut traiter tous les types de données, structurées et non structurées« Variété ». La valeur du big Data est aujourd'hui reconnue par les industriels et les gouvernements. L'exploitation efficace et efficiente du big Data permet d'assurer un avantage concurrentiel et apporte de la valeur pour plusieurs secteurs économiques, scientifiques et sociaux.

La majorité des technologies traditionnelles ne sont plus adéquates pour prendre en charge ces big Data car souvent ils manquent de performance, de flexibilité et d'évolutivité. En effet, big Data requièrent de nouvelles technologies plus flexibles et plus performantes, ainsi que de nouvelles méthodes d'analyses fiables et robustes pour stocker, traiter, analyser, sécuriser et visualiser des billions de données en un temps record. Les big Data nécessitent également le développement de nouvelles compétences, méthodes et modèles pour assurer de la disponibilité des Big Data.

La majeure partie de l'étude et de la recherche à l'heure actuelle, le lien vers les meilleurs services dans la Big data, à l'amélioration des projets ont été voir les approches et travaux connexes. Nous avons vu plusieurs travaux liés au projet sur lequel nous travaillons pour réaliser un system permettant l'assurance de la disponibilité des Big Data

Nous avons présenté notre système de pour l'assurance de la disponibilité des BigData. Notre architecture est composée d'un ensemble de composants.

Nous avons présenté les étapes de la mise en œuvre de notre projet avec tous les outils, les langages et les plateformes utilisés ainsi que la présentation avec l'explication du rôle de chaque outil. Nous avons illustré les interfaces graphiques avec une description textuelle.

Bibliographie

[1]. Lisbeth Rodríguez-Mazahua1 · Cristian-Aarón Rodríguez-Enríquez1 · José Luis Sánchez-Cervantes1 · Jair Cervantes2 · Jorge LuisGarcía-Alcaraz3 · Giner Alor-Hernández1- A general perspective of Big Data: applications, tools, challenges and trends

[2]. Ekaterina Olshannikova, Aleksandr Ometov, Yevgeni Koucheryavy and Thomas Olsson.

[3]. Ismael Caballero, Manuel Serrano, and Mario Piattini Paseo de la Universidad 4, 13071, Ciudad Real, Spain- A Data Quality in Use Model for Big Data.

[4]. Manas Kumar Sanyal, Sajal Kanti Bhadra and Sudhangsu Das- A Conceptual Framework for Big Data Implementation to Handle Large Volume of Complex Data

[5]. Hamza Saouli*, Kazar Okba*Dounya Kassimi*- Applications et enjeux des Big Data dans le contexte des défis mondiaux

[6]. Min Chen · Shiwen Mao · Yunhao Liu- Big Data: A Survey

[14]. Shuyu Li and Jerry Gao- Security and Privacy for Big Data

[8]. Seymour Bosworth, M.E. Kabay, Eric Whyne -COMPUTER SECURITY HANDBOOK

[10]. Big data and data protection.

[9]. Yuri Demchenko1(&), Canh Ngo1, Cees de Laat1, Peter Membrey2,

and Daniil Gordijenko3-Big Security for Big Data: Addressing Security

Challenges for the Big Data Infrastructure

[12]. Jan Camenisch, Simone Fischer-Hübner, Marit Hansen - Privacy andIdentity

Management for the Future Internet in the Age of Globalisation

[13]. Sithu D. Sudarsan, Raoul P. Jetley and Srini Ramaswamy -Security and Privacy of Big Data

[12]. Leslie P. Francis- Introduction: Technology and New Challenges for Privacy

[15]. Tamir Tsegaye, Stephen Flowerday-Controls for Protecting Critical Information Infrastrcture from Cyberattacks.

[17]. Philip Derbeko, Shlomi Dolev, Ehud Gudes, Shantanu Sharma- Security and privacy aspects in MapReduce on clouds: A survey.

[16]. Ajit Gaddam- Securing Your Big Data Environment.

[34]. Don Libes, Seungjun Shin, Jungyub Woo- Considerations and Recommendations for Data Availability for Data Analytics for Manufacturing.

[21]. Yu Huang and Tiejian Luo- NoSQL Database: A Scalable, Availability, High Performance Storage for Big Data.

[22]. Tim F¨ orster1, Simon Thum2, and Arjan Kuijper- High Availability of Big-Geo-Data as a Platform as a Service

[23]. Ivan Gankevich(B), Yuri Tipikin, Vladimir Korkhov, Vladimir Gaiduchok, Alexander Degtyarev, and Alexander Bogdanov- Factory: Master Node High-Availability for Big Data Applications and Beyond.

[24]. http://it.toolbox.com/blogs/database-administration/big-data-ensuring-data-availability-75688

[33]. Oracle Big Data Appliance - Maximum Availability Architecture

[25]. Atsushi Kanai, Yuuki Kajiura- A File-distribution Approach to Achieve High Availability and Confidentiality for Data Storage on Multi-cloud

[26]. Seungmin Kang∗, Bharadwaj Veeravalli∗, Khin Mi Mi Aung*y*, Chao Jin*y* -An Efficient Scheme to Ensure Data Availability for a Cloud Service Provider

[27]. Rajesh Vargheese, Yannis Viniotis- Influencing Data Availability in loT Enabled Cloud based e-Health in a 30 day Readmission Context

[28]. R.K. Banyal, V.K. Jain, Pragya Jain -Data Management System to Improve Security and Availability in Cloud Storage

[29]. Tsozen Yeh, Huichen Lee- Enhancing Availability and Reliability of Cloud Data through Syncopy

[30]. Aditi Tripathi, Mayank Deep Khare, Dr. Pradeep kumar Singh- Efcient and Secured Approach for Faster Data Availability and Restoration in Disaster Cloud Data Management

[31]. Qutaibah Althebyan, Rami Mohawesh, QussaiYaseen, YaserJararweh -Mitigating Insider Threats in a Cloud Using a Knowledgebase Approach while Maintaining Data Availability

[32]. Xin Pei1, Jiuchuan Lin1, Bo Jin1 and Yongjian Wang- Bo Jin1 and Yongjian Wang- Ensuring Replication-based Data Integrity and Availability in Multicloud Storage

[35]. Yuuki Ueda, Hideharu Kojima and Tatsuhiro Tsuchiya- On the Availability of Replicated Data Managed by Hierarchical Voting

Erratum

On remercie d'avance toute personne qui nous signale, les erreurs qu'il pourrait déceler. À l'adresse suivante : hamzadjerboua7@gmail.com