



République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

**Université des Sciences et de la Technologie Houari Boumediene**

**Faculté d'Electronique et d'Informatique  
Département Informatique**

Mémoire de Master

**Filière : Informatique**

**Spécialité :  
Mathématiques et Informatique Décisionnelle**

---

**Thème**

**Conception et mise en œuvre d'un Data Lake**

---

**Sujet Proposé par :**

**Mme. BESSAI F.**

**Mme. ALIANE H.**

**Présenté par :**

**SADAOUI Abdelwahab**

**ZIOUCHI Nassim**

**Soutenu le : 11/09/2019**

**Devant le jury composé de :**

**Mr. KHEMISSA H. Président**

**Mme. GUEBAILI R. Membre**

Binôme N° : 044 / 2019

# *Remerciements*

*Nous tenons à remercier en premier lieu ALLAH, le tout puissant de nous avoir donné la patience et le courage durant cette année, ainsi que l'audace pour dépasser toutes les difficultés.*

*Nous tenons à exprimer notre profonde gratitude à nos encadrants Mme **Bessai** et Mme **Aliane** de nous avoir guidé et orienté tout au long de la réalisation de notre projet, de nous faire bénéficier de leurs qualités scientifiques et pour leurs précieux conseils qu'elles nous ont prodigué.*

*Nous adressons aussi nos vifs remerciements à tous les membres du jury pour l'honneur qu'ils nous ont fait en acceptant de juger notre travail.*

*Nos remerciements s'étendent également à tous nos enseignants durant les années des études qui ont collaboré à notre bonne formation.*

*Finalement nous remercions tous ceux qui nous ont encouragé et aidé et que nous n'avions pas pu citer leurs noms, qu'Allah les protège.*

# Table des matières

|                                                              |           |
|--------------------------------------------------------------|-----------|
| Introduction générale                                        | 1         |
| <b>I ÉTAT DE L'ART</b>                                       |           |
| <b>1 Big data</b>                                            | <b>3</b>  |
| 1.1 Introduction . . . . .                                   | 3         |
| 1.2 Définitions . . . . .                                    | 3         |
| 1.3 Caractéristiques de Big Data . . . . .                   | 4         |
| 1.4 Types de données . . . . .                               | 5         |
| 1.5 Types de traitements de données Big Data . . . . .       | 6         |
| 1.6 Architecture Big Data . . . . .                          | 7         |
| 1.6.1 Architecture Lambda . . . . .                          | 8         |
| 1.7 Solutions Big Data . . . . .                             | 10        |
| 1.7.1 Hadoop . . . . .                                       | 10        |
| 1.7.2 Spark Apache . . . . .                                 | 10        |
| 1.8 Conclusion . . . . .                                     | 10        |
| <b>2 Data Lake</b>                                           | <b>11</b> |
| 2.1 Introduction . . . . .                                   | 11        |
| 2.2 Histoire de Data Lake . . . . .                          | 11        |
| 2.3 Définitions . . . . .                                    | 12        |
| 2.4 Caractéristiques des Data Lake . . . . .                 | 12        |
| 2.5 Architecture du Data Lake . . . . .                      | 14        |
| 2.6 Catégories de Data Lake . . . . .                        | 15        |
| 2.7 Utilisateurs de Data Lake . . . . .                      | 15        |
| 2.8 Avantages et inconvénients de Data Lake . . . . .        | 16        |
| 2.9 Les usages du data lake . . . . .                        | 16        |
| 2.10 Différences entre Data Lake et Data Warehouse . . . . . | 17        |
| 2.11 Conclusion . . . . .                                    | 18        |
| <b>3 Écosystème Apache Hadoop</b>                            | <b>19</b> |
| 3.1 Introduction . . . . .                                   | 19        |

|         |                                                              |    |
|---------|--------------------------------------------------------------|----|
| 3.2     | Caractéristiques de Hadoop . . . . .                         | 19 |
| 3.3     | Fonctionnement de Hadoop . . . . .                           | 20 |
| 3.4     | Systèmes où Hadoop peut être utilisé . . . . .               | 20 |
| 3.5     | Composants Hadoop . . . . .                                  | 21 |
| 3.5.1   | Système de fichiers distribués Hadoop (HDFS) . . . . .       | 21 |
| 3.5.1.1 | Le nœud principal (Namenode) . . . . .                       | 22 |
| 3.5.1.2 | Le nœud secondaire (Secondary Namenode) . . . . .            | 22 |
| 3.5.1.3 | Le nœud de données (Datanode) . . . . .                      | 22 |
| 3.5.1.4 | Stockage de données dans HDFS . . . . .                      | 23 |
| 3.5.1.5 | Le flux opérationnel de lecture/écriture dans HDFS . . . . . | 23 |
| 3.5.2   | MapReduce . . . . .                                          | 25 |
| 3.5.2.1 | Exemple MapReduce . . . . .                                  | 25 |
| 3.5.3   | YARN . . . . .                                               | 28 |
| 3.6     | Écosystème Hadoop . . . . .                                  | 29 |
| 3.7     | Apache Spark . . . . .                                       | 32 |
| 3.8     | Conclusion . . . . .                                         | 33 |

## II CONCEPTION ET MISE EN ŒUVRE

|          |                                                                     |           |
|----------|---------------------------------------------------------------------|-----------|
| <b>4</b> | <b>Conception de la solution de stockage Data Lake</b>              | <b>34</b> |
| 4.1      | Introduction . . . . .                                              | 34        |
| 4.2      | Architecture générale . . . . .                                     | 34        |
| 4.2.1    | Ingestion des données . . . . .                                     | 35        |
| 4.2.2    | Interrogation des données . . . . .                                 | 36        |
| 4.2.3    | Visualisation des données . . . . .                                 | 36        |
| 4.3      | Description de l'architecture . . . . .                             | 36        |
| 4.3.1    | Ingestion des données . . . . .                                     | 36        |
| 4.3.1.1  | Ingestion des données structurées (Bases de données) . . . . .      | 37        |
| 4.3.1.2  | Ingestion des données semi-structurées et non-structurées . . . . . | 38        |
| 4.3.2    | Stockage des données . . . . .                                      | 39        |
| 4.3.3    | Transformation et exploitation des données . . . . .                | 40        |
| 4.3.3.1  | Avantages de Hive . . . . .                                         | 40        |
| 4.3.3.2  | Avantages de Pig . . . . .                                          | 41        |
| 4.3.4    | Visualisation des données et résultats . . . . .                    | 41        |
| 4.4      | Conception UML de la solution Data Lake . . . . .                   | 41        |
| 4.4.1    | Identification des acteurs . . . . .                                | 42        |
| 4.4.2    | Diagramme de cas d'utilisation . . . . .                            | 42        |
| 4.4.2.1  | Diagramme de cas d'utilisation général . . . . .                    | 42        |
| 4.4.3    | Diagramme de séquence . . . . .                                     | 43        |
| 4.4.3.1  | Diagramme de séquence «Authentification» . . . . .                  | 44        |

|          |                                                                     |           |
|----------|---------------------------------------------------------------------|-----------|
| 4.4.3.2  | Diagramme de séquence «Alimentation du Data Lake» . . .             | 44        |
| 4.4.3.3  | Diagramme de séquence «Consulter les données» . . . . .             | 46        |
| 4.4.4    | Diagramme de classe . . . . .                                       | 47        |
| 4.5      | Conclusion . . . . .                                                | 48        |
| <b>5</b> | <b>Implémentation et mise en œuvre</b>                              | <b>49</b> |
| 5.1      | Introduction . . . . .                                              | 49        |
| 5.2      | Présentation de l'environnement de travail . . . . .                | 49        |
| 5.2.1    | Environnement matériel . . . . .                                    | 49        |
| 5.2.2    | Environnement logiciel . . . . .                                    | 51        |
| 5.3      | Configuration de l'environnement de travail . . . . .               | 52        |
| 5.3.1    | Configuration commune des machines . . . . .                        | 53        |
| 5.3.2    | Configuration personnalisée des machines . . . . .                  | 53        |
| 5.3.2.1  | Configuration du cluster Hadoop (Data Lake) . . . . .               | 54        |
| 5.3.2.2  | Configuration de la machine d'utilisateur . . . . .                 | 59        |
| 5.4      | Présentation de l'application . . . . .                             | 64        |
| 5.4.1    | Lancement du cluster Hadoop . . . . .                               | 65        |
| 5.4.2    | Présentation de l'interface Hue . . . . .                           | 66        |
| 5.4.3    | Alimentation du Data Lake . . . . .                                 | 67        |
| 5.4.3.1  | Base de données du CERIST . . . . .                                 | 67        |
| 5.4.3.2  | Données semi-structurées et non-structurées . . . . .               | 69        |
| 5.4.4    | Exemples de traitements et d'analyses sur les données du Data Lake  | 71        |
| 5.4.4.1  | Cas des données structurées . . . . .                               | 71        |
| 5.4.4.2  | Cas des données semi-structurées . . . . .                          | 73        |
| 5.4.4.3  | Cas des données non-structurées . . . . .                           | 77        |
| 5.4.5    | Gestion des droits d'accès et privilèges des utilisateurs . . . . . | 80        |
| 5.5      | Conclusion . . . . .                                                | 82        |
|          | <b>Conclusion générale et perspectives</b>                          | <b>83</b> |
|          | <b>Bibliographie</b>                                                | <b>85</b> |