

REPUBLICQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

وزارة التعليم العالي و البحث العلمي

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE
SCIENTIFIQUE



RAPPORT DU PROJET DE FIN D'ÉTUDES

Pour l'obtention du diplôme de

Poste Graduation Spécialisée en Big Data et Calcul Intensif

**Une solution Big Data pour l'analyse temps réel
des fichiers logs**

Réalisé par :

- CHAHMANA Samir
- MAZOUZ Imed

Devant le Jury :

Président :

Mr SAADALAH Madjid

Examineurs :

Mr BOUCENNA Fateh

Mr DABAH Adel

L'Encadreur

Mr Yahiaoui Saïd

Co-encadreur

Mr. Bouchama Nadir

Année Universitaire : 2017 /2018

ملخص

تعتبر الاتصالات الهاتفية المتنقلة هي واحدة من أكثر التقنيات المستخدمة في العالم الحديث.

إنها ليست أداة اتصال فحسب ، بل هي أيضاً خزان معلومات في عدة قطاعات كقطاع الإدارة ، المالية ، التعليم، النقل، الزراعة والصحة.

حيث أن لها دور مهم بشكل خاص في مجال إدارة الطوارئ والكوارث من خلال تقديم تقديرات دقيقة لأعداد السكان وحركاتهم قبل أو أثناء هذه الأحداث.

الغرض من هذا العمل هو إقتراح و تطوير هندسة البيانات الضخمة، لإستغلال سجل بيانات المكالمات CDR في الوقت الأنّي و ذلك من أجل تقرير كثافة حركة المرور على الطرق في المدينة الذكية.

الكلمات المفتاحية : البيانات الضخمة، المدينة الذكية، سجل بيانات المكالمات،

هندسة البيانات الضخمة، الاتصالات الهاتفية المتنقلة، خوارزم k-means.

Abstract

Mobile telephony is one of the most used technologies in the modern world.

It is not only a communication tool, it is also a vector of information in sectors such as administration, finance, education, transport, agriculture and health.

It has a particularly important role in the field of emergency and disaster management by providing critical estimates of population numbers and movements before or during such events.

The purpose of this work is to propose and develop a Big Data architecture to exploit CDR (Call Detail Record) data in real time in order to estimate the density of road traffic in a smart city.

Key Words: Big Data, Call Detail Record, Hadoop, Big Data architecture, mobile telephony, K-means.

Résumé

La téléphonie mobile est l'une des technologies les plus utilisées dans le monde moderne. Elle n'est pas seulement un outil de communication, elle est aussi un vecteur d'informations dans des secteurs tels que l'administration, la finance, l'éducation, le transport, l'agriculture et la santé.

Elle joue un rôle particulièrement important dans le domaine de la gestion des situations d'urgence et des catastrophes en fournissant des estimations cruciales sur le nombre et les mouvements de population avant ou pendant de tels événements.

Ce travail a pour but de proposer et de développer une architecture Big Data pour exploiter en temps réel les données CDR (Call Detail Record) afin d'estimer la densité du trafic routier dans une ville intelligente.

Mots clés: Big Data, ville intelligente, Call Detail Record, Hadoop, architecture Big Data, téléphonie mobile, K-means.

Sommaire

Introduction générale.....	1
Chapitre I : <i>Bigdata</i>.....	4
I.1 Introduction.....	4
I.2 Définition du <i>Big Data</i>.....	4
I.3 Technologies du <i>Big Data</i>.....	5
I.4 Caractéristiques de <i>Big Data</i>.....	5
I.5 Les enjeux du <i>Big Data</i>.....	6
I.6 Architecture <i>Big Data</i>.....	7
I.7 Ecosystème <i>Hadoop</i>.....	8
I.8 <i>Apache Spark</i>.....	10
I.9 Comparaison entre <i>Mapreduce</i> et <i>Spark</i>.....	12
I.10 <i>Apache Storm</i>.....	13
I.11 Conclusion.....	15
Chapitre II : Analyse des Fichiers Logs.....	17
II.1 Introduction.....	17
II.2 Les fichiers <i>logs</i>.....	17
II.3 Traitement des fichiers <i>logs Big Data</i>.....	19
II.4 Les différents types de données.....	20
II.5 Enjeux sur le cycle de vie de la donnée.....	22
II.6 Source de données.....	23
II.7 Les défis de l'analyse des fichiers <i>log</i>.....	24
II.8 Données ouvertes (<i>OPENDATA</i>).....	26
II.9 Typologie des méthodes d'analyse de données massives.....	28
II.10 Analyse des fichiers <i>logs</i>.....	29
II.11 Outils d'analyse de fichiers logs dans le bigdata en temps réel.....	31
II.12 Comparaison entre <i>ELK/Splunk/Graylog</i>.....	33
II.13 Architectures de Traitement des Fichiers <i>Logs</i> dans <i>Big Data</i>.....	35
a. Architecture <i>Lambda</i>.....	35
b. Architecture <i>KAPPA</i>.....	37
c. Architecture <i>SMACK</i>.....	39
II.14 Conclusion.....	41
Chapitre III : Analyse de Relevés détaillés des communications <i>CDR</i>.....	43
III.1 Introduction.....	43
III.2 Technologie de téléphonie mobile.....	44
III.3 Relevés détaillés des communications (<i>CDR</i>).....	45

III.4	Estimation la densité de population d'une région donnée	46
III.5	L'intérêt d'analyser les données mobile.....	47
III.6	Analyse des données mobiles « CDR »	50
III.8	Choix d'algorithme (DBSCAN ,KMEANS)	51
III.9	Conclusion	55
Chapitre IV : Conception.....		57
IV.1	Introduction.....	57
IV.2	Objectif de l'étude	57
IV.3	Architecture générale	58
IV.4	Jeu de données	59
IV.5	Analyse de données.....	60
5.1	Nombre de <i>clusters</i> (Méthode <i>Elbow Curve</i>)	60
5.2	L'algorithme <i>k-means</i>	62
IV.6	Conclusion	64
Chapitre V : Implémentation.....		66
V.1	Introduction.....	66
V.2	Déploiement de l'architecture	67
V.3	<i>Dataset</i> d'analyse	71
V.4	Plateforme d'analyse	72
V.5	Conclusion	81
Conclusion générale		82
Référence		83
Annexe...		84