Université Paris 8 Vincennes – Saint-Denis UFR MITSIC

Réalisation d'une application Big Data pour traiter les données Covid19

Dirigé par :
Madame Rakia JAZIRI

Auteurs :
Amassin NACERDDINE
Larbi BENMEDJBOUB

19 janvier 2021

Table des matières

Table des figures

Introduction générale

Depuis déjà quelques années, les évolutions technologiques en informatique se succèdent à une vitesse impressionnante pour répondre principalement au besoin toujours croissant des utilisateurs particulièrement aux données généré par ces derniers.

En effet, non seulement les machines actuelles tel que les serveurs ont des capacité de traitement de plus en plus croissante,permettant d'effectuer de nombreuses opération très complexe mais aussi les évolutions en télécommunication qui rendent l'information de plus en plus disponible et ceci quasiment en temps réel.

Nous allons donc tirer patrie de cette augmentation de puissance ainsi que l'existence de bibliothèques et procédures de haut niveau dans le but de traiter et de présenter les données de la pandémie de covid-19.

Pour ce faire nous avons développé une application sous Python afin de présenter sous forme de graphe la grosse quantité de données récupérées .

Ce rapport est organisé comme suit :

— Chapitre 1 : nous aborderons l'aspect technique et le fonctionnement du système.
— Chapitre 2 : nous présenterons la conception et les scénarios envisagés ainsi que la méthode Scrum suivie.
— Chapitre 3 : nous présenterons l'environnement matériel et logiciel du développement et certaines spécifications de notre application.

Quant à la conclusion, elle dressera les perspectives du projet.

Chapitre 1

Étude théorique et fonctionnement du système

1.1 Introduction

Dans cette section nous allons identifier et analyser les différents problèmes et besoins auxquels font face.

1.2 État de l'art

L'idée de réaliser une application afin de traiter et de présenter les données covid-19 nous est venu car c'est une événement d'actualité et les données de la pandémie sont disponible en open source.
Les outils proposés par des plate formes comme Google et les applications gouvernementales sont très performantes,néanmoins il nous a paru intéressant de faire des statistiques propre a nous selon des paramètres et des critères bien précis.

1.3 Fonctionnalités

Nous présenterons dans cette section les différentes fonctionnalités du système

1.3.1 Trois Fonctionnalités majeurs

Notre application propose trois fonctionnalités majeurs

— la première permet a l'utilisateur de visualiser une MAP qui va contenir un code couleur selon si le pays est gravement ou moins touché,on va prendre comme critères différents points et les passer a une fonction floue qui va nous retourner un Score(entre 0 et 100) pour chaque pays.

— la seconde lui permettra de voir les graphes du nombre de cas et du nombre de décès par pays en moyenne par mois.

— la troisième fonctionnalité va lui permettre de voir les différents Tweets liés a l'actualité corvid-19 .

Chapitre 2

Méthodologie d'analyse et de conception

2.1 Introduction

Dans cette section nous présenterons la méthode de conception adoptée pour la réalisation de notre projet.

2.2 Méthode SCRUM

Pour la réalisation de notre projet nous avons adopté la méthode agile SCRUM qui est parfaitement adapté pour un développement rapide flexible et efficace de logiciels.

Cette méthode tire son nom de la mêlée du rugby.Elle sous entend donc un grand travail d'équipe[4] (.
L'approche SCRUM suit les principes de la méthodologie Agile, c'est-à-dire l'implication et la participation active du client tout au long du projet.
Ainsi notre équipe a du se réunit quotidiennement lors d'une réunion de synchronisation, appelée mêlée quotidienne, afin de suivre l'avancement du projet et la répartition des taches quotidienne.[1]

2.3 Diviser pour régner

2.3.1 Diviser notre problème

Notre problème étant complexe il nous a donc fallu le diviser en plusieurs sous problèmes qui étaient plus faciles a appréhender.

Les IHM

Les interfaces homme machine étant très importantes car elles représentent le premier contact avec le clients on a dù les optimiser pour une meilleure ergonomie..

Les données

Les données étant la partie la plus importante de notre application nous ne devions en aucun cas négliger cette aspect la.
Par ailleurs la méthode pour la sauvegarde de données que nous avons choisi et le un serveur distant sous AWS.

Nous nous sommes souvent référé a la documentation très riche de ce dernier.[**?**]

Les APIs et les Frameworks

Les APIs et les Frameworks étant nombreuses nous avions l'obligation en apprendre le plus possible grâce a la documentation et en maîtriser un maximum.Pour pouvoir passer au codage de l'application.

2.3.2 Diviser le temps

Sprint 1

durant notre premier Sprint nous nous somme mis d'accord sur le fonctionnement du système et avons émis les différents cas d'utilisations.

Sprint 2

Durant notre second sprint nous avons schématisé les interfaces de notre applications.

Sprint 3

Durant le 3ème Sprint nous avons validé les technologies et platformes a utiliser.Et avons synchronisé notre travail dans un service web d'hébergement et de gestion de développement de logiciels.

2.3.3 Diviser l'équipe

Notre temps étant très réduit et notre équipe très peu fournis on se devait de bien se répartir les taches entre nous et ceci en adéquation avec les compétences et points forts de chacun d'entre nous.

Chapitre 3

Visualisation des données

3.1 Introduction

Dans cette section nous présenterons notre application qui permet la présentation des données aux utilisateurs.

3.2 Logiciels et outils utilisé

3.2.1 Python3

Le langage de programmation interprété, multi-paradigme et multiplateformes python a était utilisé et choisi pour ça productivité ainsi que pour ses outils de haut niveau et une syntaxe simple à utiliser..[?]

FIGURE 3.1 – Logo Python

3.2.2 Flask 1.1.2

Le micro framework open-source de développement web en Python a était choisi car il est car il est très léger très facile a déployer.[?]

FIGURE 3.2 – Logo Flask

3.2.3 Pycharm

L'environnement de développement intégré utilisé pour programmer en Python a était utilisé pour réaliser ce projet.

FIGURE 3.3 – Logo Pycharm

3.2.4 Chart.js

. La bibliothèque JavaScript open source a était utilisé pour l'affichage des graphes

FIGURE 3.4 – LogoCart.js

3.2.5 JavaScript Object Notation

. JavaScript Object Notation est un format de données textuelles dérivé de la notation des objets du langage JavaScript. Il permet de représenter de l'information structurée comme le permet XML par exemple.

FIGURE 3.5 – JSON

3.2.6 AWS EC2

Amazon Elastic Compute Cloud ou EC2 est un service proposé par Amazon permettant à des tiers de louer des serveurs sur lesquels exécuter leurs propres applications web. Un cluster avec un Master et deux Slaves a été utilisé dans notre cas

FIGURE 3.6 – Logo AWS

3.2.7 AWS Elastic Beanstalk

Le service web proposé par Amazon Web Services pour le déploiement d'applications a était utilisé pour le déploiement de notre application.

FIGURE 3.7 – Logo AWS Elastic Beanstalk

3.2.8 Hadoop

Le framework libre et open source a été installé ainsi que ses différente brique sur notre cluster dans le but de traiter les données que nous disposons et ceci d'une manière distribuées .

FIGURE 3.8 – Logo Hadoop

Hive

Apache Hive est une infrastructure d'entrepôt de données intégrée sur Hadoop permettant l'analyse, le requêtage via un langage proche syntaxiquement de SQL ainsi que la synthèse de données[2]

FIGURE 3.9 – Logo Hive

Flume

Apache Flume est un logiciel de la fondation Apache destiné à la collecte et à l'analyse de fichiers de log. L'outil est conçu pour fonctionner au sein d'une architecture informatique distribuée et ainsi supporter les pics de charge.[3] Il nous a été utile pour la collécte de données a partir de Twitter

FIGURE 3.10 – Logo flume

3.2.9 Git & GitHub

Le service d'hébergement basé sur le Web GitHub ainsi que le logiciel de gestion de versions Git on aussi étaient très utiles pour le traitement des différentes versions de l'application.

FIGURE 3.11 – Logo Git & GitHub

Conclusion générale

Ce travail nous a permit de souligner la difficulté de la présentation de données afin de ne garder uniquement celles pertinentes. Nous avons réussi a réaliser la majorité des fonctionnalités théorique de base que l'on s'était fixé .Il reste néanmoins des améliorations à faire afin d'aboutir a un travail finale. On peut citer

— Améliorer les IHM pour une meilleure visualisation.
— Optimiser les requêtes afin d'avoir le temps d'attente le plus bas possible.

Le développement de cette application nous a permit d'enrichir nos connaissances dans les différentes API utilisé.

En ce qui concerne l'aspect humain, ce travail nous a donné un aperçu sur la vie professionnelle, à mieux nous organiser dans notre travail et ce malgré les circonstances actuel qui sont plus que particulières , afin d'accomplir les tâches qui nous sont confiées dans les meilleures conditions et dans les plus brefs délais.

Bibliographie

[1] S. S. Apoorva Srivastava, Sukriti Bhardwaj. *SCRUM model for agile methodology*, page 1.2.3, 6 May 2017.

[2] H. L. Manual. *https ://cwiki.apache.org/confluence/display/Hive/LanguageManual*.

[3] M. M. e. J.-L. R. Pirmin Lemberger, Marc Batty. *Big Data et machine learning : Manuel du data scientist, Dunod*, page 240p, 2015.

[4] K. Schwabe. *SCRUM Development Process*, 2018.

Résumé —

La pandémie de Covid-19 est une pandémie d'une maladie infectieuse émergente, appelée la maladie à coronavirus 2019 ou Covid-19, provoquée par le coronavirus SARS-CoV-2, apparue à Wuhan le 17 novembre 2019, dans la province de Hubei (en Chine centrale), avant de se propager dans le monde.
L'écosystème très complet proposé par Hadop et AWS nous permet de faire des traitements de plus en plus complexe. Notre imagination reste donc la seul limite a tout cela.

Mots clés : Hadoop . AWS . Covid19 . Big data . Cloud . Loique Floue . Data visualisation

Abstract— The Covid-19 pandemic is a pandemic of an emerging infectious disease, called the 2019 coronavirus disease or Covid-19, caused by the SARS-CoV-2 coronavirus, which appeared in Wuhan on November 17, 2019, in the province of Hubei (in vspace 0.2cm Central China), before spreading around the world.
The very complete ecosystem offered by Hadop and AWS allows us to perform increasingly complex treatments. Our imagination therefore remains the only limit to all this.

Keywords : Hadoop . AWS . Covid19 . Big data . Cloud . Fuzzy logic . Data visualisation