

Département de Mathématiques Filière d'Ingénieurs Génie Mathématique et Informatique

Rapport de Projet de Fin d'Études

Présenté par :

HANI Moad

Data Science et aide à la décision : Analyse des sentiments et création de nouvelles connaissances pour fidéliser la clientèle de L'Union IT Services

Soutenu : le 25 juin 2018

Sous la direction de :

S. Amine	Professeure à la Faculté des Sciences et Techniques Mohammedia.
N. Boukouchi	Directeur pédagogique à L'UITS et expert en virtualisation, stockage et réseaux.
K. Katkout	Directeur général de L'UITS et expert en sécurité et administration réseaux.
D. Moumida	Professeur à la Faculté des Sciences et Techniques Mohammedia.
.	

Devant le jury composé de :

S. Amine	Professeure à la Faculté des Sciences et Techniques Mohammedia.
S. Chaira	Professeure à la Faculté des Sciences et Techniques Mohammedia.
K. Katkout	Directeur général de L'UITS et expert en sécurité et administration réseaux.
D. Moumida	Professeur à la Faculté des Sciences et Techniques Mohammedia.
S. Sajid	Professeur à la Faculté des Sciences et Techniques Mohammedia.

2017-2018

**Faculté des Sciences et techniques Mohammedia – BP 146, Mohammedia 20650-
Maroc. Tel : 0523314705 -Fax : 0523315353 - Site web : www.fstm.ac.ma**



Je dédie ce travail à
mes Chers Parents, ma Soeur
Aînée et mes Professeur(e)s
de "vie" qui m'ont encou-
ragé et soutenu durant mes
études. Sans eux, tout
cela n'aurait pas
été possible.
♥

Remerciements

Tout d'abord, je tiens à remercier tout particulièrement Monsieur Chakib ABCHIR notre Chef de département mathématique qui planifie et coordonne nos stages. Ma reconnaissance va aussi à tout le corps Professoral de la filière "Génie Mathématique et Informatique", notamment mon Professeur Ahmed TAIK pour m'avoir encouragé à compléter mon projet de fin d'année sur la virtualisation par son corollaire Data Science.

Je tiens également à exprimer toute ma gratitude à mes Professeurs Madame Saida AMINE et Monsieur Driss MOUMIDA pour m'avoir encadré et orienté par leurs précieux remarques et conseils, ce qui a rendu la tâche moins ardue et plus agréable.

Mes sincères remerciements vont aussi aux membres du jury Mesdames Saida AMINE et Soumia CHAIRA et Messieurs Driss MOUMIDA et Said SAJID qui ont bien voulu examiner ce travail et m'honorer par leur présence.

J'adresse mes chaleureux remerciements à la Direction de l'Union IT Services pour sa bienveillante ouverture sur l'université et en particulier Monsieur Khalid KAT-KOUT pour m'avoir accordé ce stage. Je le remercie pour son aide précieuse et sa sympathie. Je n'oublierai pas de remercier vivement mes tuteurs d'entreprise, Monsieur Nabil BOUKOUCHI et Madame Siham CHOUIKH pour leur disponibilité à me fournir toutes les explications nécessaires qui m'ont aidé à l'élaboration de ce travail avec aisance.

Je n'oublierai pas d'exprimer mes sincères remerciements à tous ceux qui ont pu contribuer de près ou de loin à la réalisation de ce travail.

Table des matières

Dédicace

Remerciements

Table des matières

Table des figures

Glossaire

Présentation et contexte du projet – Introduction

- 0.1 Introduction
- 0.2 Présentation de l'entreprise
- 0.3 La problématique
- 0.4 Valeur ajoutée
- 0.5 Technologies et outils utilisés

Chapitre 1 : Synthèse bibliographique..... 11

1.1 Big Data et Machine learning.....	11
1. Big Data – Introduction.....	11
2. Définition et avantages du Big Data.....	11
3. Caractéristiques du Big Data.....	13
4. Ecosystème d'Hadoop.....	13
5. Les principales distributions Hadoop.....	19
5.1. Cloudera.....	19
5.2. Hortonworks.....	20
5.3. MapR.....	20
6. Base de données NoSQL.....	20
7. L'apprentissage automatique ou Machine Learning.....	21
8. Les différents types de Machine Learning.....	22
9. Les principaux algorithmes.....	24
9.1. La régression linéaire.....	24
9.2. La classification naïve bayésienne.....	24
9.3. La régression logistique.....	25
9.4. L'algorithme des k-moyennes	25
9.5. Les arbres de décision	26
9.6. Les machines à vecteurs de support.....	27
9.7. Gradient Descent.....	27
10. Choisir son type d'apprentissage et son algorithme.....	28
1.2 L'analyse des sentiments	29
1. Opinion.....	30
1.1. Opinion Régulière et opinion Comparative.....	31
1.2. Opinion Explicite et opinion Implicit.....	31
2. Sentiment.....	31
3. Approches de catégorisation de sentiments.....	32
4. Catégorisation basée sur le lexique	32
o Principe.....	33
o Technique.....	33

5. Catégorisation basée sur l'apprentissage automatique.....	34
○ Principe.....	34
○ Technique.....	34
6. Catégorisation hybride.....	35
7. Niveaux d'analyse des sentiments.....	35
○ Niveau document.....	36
○ Niveau phrase.....	36
○ Niveau aspect	36
8. Tâches de l'analyse de sentiments.....	37
8.1. Analyse de la subjectivité et détection de l'opinion.....	37
8.2 Catégorisation de sentiments.....	37
8.3 Identification du sujet et du porteur d'opinion.....	37
8.4 Résumé de l'opinion.....	38
8.5 Détection de l'ironie et du sarcasme.....	38
8.6 Détection des spams.....	38
9. Préparation des données.....	39
○ Définition	39
○ Corpus d'entraînement et corpus de test.....	39
10. Prétraitement.....	39
○ Définition	39
○ Techniques.....	40
11. Extraction des caractéristiques.....	40
12. Représentation.....	42
13. Évaluation.....	43
13.1. Précision.....	45
13.2. Rappel.....	45
13.3. F-Mesure.....	45
 Chapitre 2 : Contribution.....	47
2.1 Mise en œuvre et applications.....	47
1. Création et comparaison d'architectures Hadoop et Spark.47	
2. Analyse des sentiments – Les exemples d'exploitation de Tweets et de traitement d'images.....	89
3. Applications en R.....	94
4. Applications en Python	103
5. Machine learning et résolution de cas probables.....	127
6. Chatbot et marketing, la combinaison gagnante.....	154
○ Premier chatbot : Assister à distance les clients effectifs.....	155
○ Second chatbot : Marketier les formations et services de l'UTS auprès des clients potentiels.....	157
 Conclusion et perspectives.....	167
Annexes.....	171
Bibliographie.....	183

Table des figures

Diagramme illustrant l'opération de lecture de fichier dans Hadoop.....	15
Diagramme illustrant l'opération d'écriture de fichier dans Hadoop.....	16
Diagramme illustrant les composants de Cloudera.....	19
Diagramme illustrant les composants de Hortonworks	20
Diagramme illustrant les composants de MapR	21
Courbe de la régression linéaire.....	24
Courbe de la régression logistique.....	25
Classification selon l'algorithme des K-means.....	26
Illustration du principe des arbres de décision.....	26
Principe de l'algorithme de classification binaire	27
Gradient Descent.....	28
Proximité entre documents.....	42
Matrice de confusion.....	43
Précision et rappel (« recall »).....	44
Réseau Twitter et exploitation des données.....	91
Twitter en chiffres.....	92
Exemple de Tweet (2010 – l'API a beaucoup changé depuis).....	93

Glossaire

- SSH :** *Secure Shell (SSH) est à la fois un programme informatique et un protocole de communication sécurisé. Le protocole de connexion impose un échange de clés de chiffrement en début de connexion. Par la suite, tous les segments TCP sont authentifiés et chiffrés.*
- DNS :** *Domain Name Server (DNS) est un système essentiel au fonctionnement d'Internet. C'est entre autres, le service qui permet d'établir la correspondance entre le nom de domaine et son adresse IP.*
- VM :** *Une machine virtuelle, ou VM (Virtual Machine), est un environnement d'application ou de système d'exploitation (OS, Operating System) installé sur un logiciel qui imite un matériel dédié. Un logiciel spécialisé, appelé hyperviseur, émule intégralement les différentes ressources matérielles d'un serveur ou d'un PC client, telles que l'unité centrale, la mémoire, le disque dur ou le réseau, et permet à des machines virtuelles de les partager.*
- IP :** *Protocole informatique de connexion qui gère la transmission des données par Internet.*
- API :** *Une interface de programmation applicative (souvent désignée par le terme API pour application programming interface) est un ensemble normalisé de classes, de méthodes ou de fonctions qui sert de façade par laquelle un logiciel offre des services à d'autres logiciels. Elle est offerte par une bibliothèque logicielle ou un service web, le plus souvent accompagnée d'une description qui spécifie comment des programmes consommateurs peuvent se servir des fonctionnalités du programme fournisseur.*
- Maître-Esclave :** *Dans cette Architecture, un serveur (appelé noeud maître) est considéré comme maître. Il est responsable de la répartition du travail entre les différents postes clients ainsi que de leur synchronisation. Il récupère également les erreurs et les résultats des calculs. On lui soumet les différents travaux qu'il gère à l'aide d'une file d'attente. Les postes client sont appelés noeud. Ils sont connectés au réseau et capables d'exécuter du code informatique sous l'ordre du serveur. On peut leur transmettre non seulement les données mais aussi le code à appliquer à celles-ci.*
- Mononiveau :** *Dans cette Architecture, tous les noeuds ont le même rôle. Ils réalisent eux même la répartition à l'aide d'échange de messages.*
- BI :** *L'informatique décisionnelle ou BI (Business Intelligence) est un processus technologique qui analyse des données pour présenter des informations*

exploitables par les dirigeants, les cadres commerciaux et les autres utilisateurs, afin de leur permettre de prendre des décisions plus avisées.

Data Science : *La science des données est l'extraction de connaissance d'ensembles de données^{1,2}. Elle emploie des techniques et des théories tirées de plusieurs autres domaines plus larges des mathématiques, la statistique principalement, la théorie de l'information et la technologie de l'information, notamment le traitement de signal, des modèles probabilistes, l'apprentissage automatique, l'apprentissage statistique, la programmation informatique, l'ingénierie de données, la reconnaissance de formes et l'apprentissage, la visualisation, l'analytique prophétique, la modélisation d'incertitude, le stockage de données, la compression de données et le calcul à haute performance. Les méthodes qui s'adaptent aux données de masse sont particulièrement intéressantes dans la science des données, bien que la discipline ne soit généralement pas considérée comme limitée à ces données.*

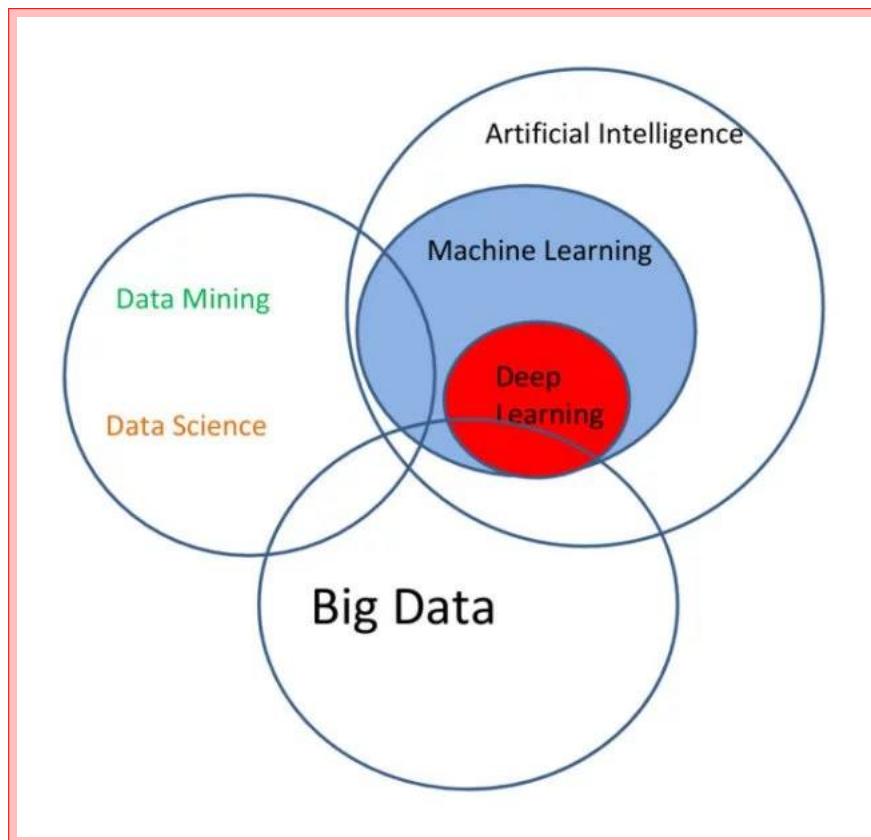
Big Data : *La Big Data ou mégadonnées désignent l'ensemble des données numériques produites par l'utilisation des nouvelles technologies à des fins personnelles ou professionnelles. Cela recoupe les données d'entreprise (courriels, documents, bases de données, historiques de processeurs métiers...) aussi bien que des données issues de capteurs, des contenus publiés sur le web (images, vidéos, sons, textes), des transactions de commerce électronique, des échanges sur les réseaux sociaux, des données transmises par les objets connectés (étiquettes électroniques, compteurs intelligents, smartphones...), des données géolocalisées, etc.*

ML : *(en anglais machine learning, littéralement « l'apprentissage machine ») ou apprentissage statistique, champ d'étude de l'intelligence artificielle, concerne la conception, l'analyse, le développement et l'implémentation de méthodes permettant à une machine (au sens large) d'évoluer par un processus systématique, et ainsi de remplir des tâches difficiles ou problématiques par des moyens algorithmiques plus classiques.*

AI : *Ensemble des théories et des techniques développant des programmes informatiques complexes capables de simuler certains traits de l'intelligence humaine (raisonnement, apprentissage).*

DL : *Le deep learning ou apprentissage profond est un type d'intelligence artificielle dérivé du machine learning (apprentissage automatique) où la machine est capable d'apprendre par elle-même, contrairement à la programmation où elle se contente d'exécuter à la lettre des règles prédéterminées.*

Glossaire



Relations entre Big Data, Data Science, Machine Learning, Intelligence Artificielle et Deep Learning

Corpus : *Un corpus est un ensemble de documents, artistiques ou non (textes, images, vidéos, etc.), regroupés dans une optique précise. On peut utiliser des corpus dans plusieurs domaines : études littéraires, linguistiques, scientifiques, philosophie1, etc.*

Dataset : *Un jeu de données (en anglais dataset ou data set) est un ensemble de valeurs (ou données) où chaque valeur est associée à une variable (ou attribut) et à une observation.*

Cluster : *Groupe de serveurs constituée de deux serveurs au minimum (appelé aussi nuds) et partageant une baie de disques commune, pour assurer une continuité de service.*

Datacenter : *Installation utilisée pour remplir une mission critique relative à l'informatique et à la télématique. Il comprend en général un contrôle sur l'environnement (climatisation, système de prévention contre l'incendie, etc.), une alimentation d'urgence et redondante, ainsi qu'une sécurité physique élevée.*

Virtualisation : *La virtualisation est un ensemble de techniques permettant d'héberger plusieurs systèmes d'exploitation sur une même couche physique.*

Présentation et contexte du projet - Introduction

0.1 Introduction

Chaque jour, 2,5 quintillions de bytes de données sont générés. Ce nombre impressionnant reflète bien l'explosion du volume de données. D'ici 2020, les entreprises qui se servent du Big Data seront largement avantagées et gagneront au total 1,2 billion de dollars par an de plus que les entreprises qui ne s'en servent pas. Telle est la prédition des analystes de Forrester. Les réseaux sociaux sont des espaces qui permettent à tous publics de créer des réseaux d'amis et de relations professionnelle, en générant chaque jour des millions de postes, de likes et de tweets et des photos envoyés sur Facebook, Twitter, Instagram et autres soit un volume mondial de données qui double tous les trois ans, selon McKinsey Global Institute. Ces nouvelles pratiques ont constraint les entreprises et les marques à trouver une nouvelle manière d'analyser et d'utiliser ce flux de données appelé «Big Data», dont les techniques d'exploration et d'analyse sont adaptées à ces nouveaux ordres de grandeurs.

C'est en 2009 qu'apparut le terme « Social media analytics » chez Google pour analyser, convertir et transformer en connaissances signifiants des données volumineuses et non structurées grâce à des outils et des solutions adaptées. L'analyse des conversations sur les réseaux représente un enjeu identitaire pour les marques afin de déterminer leur influence, les sentiments des utilisateurs et même prédire des sujets émergents. Mais ces données représentent également un enjeu financier pour les collecter et les mettre à profit. Ce rapport a pour vocation de répondre à la problématique : « comment préparer une décision stratégique dans un contexte de données (concurrenrielles, évolutives, désordonnées et encombrantes) ? » et donner à mon entreprise d'accueil, L'Union IT Services, l'opportunité de mieux cibler et de mieux s'adresser à leur publics sur les réseaux sociaux ayant pour toile de fond la prise de décision stratégique pour assurer la démarcation de la concurrence, alors que l'utilisation et l'exploitation de ces données s'avère être très complexe et pose de nombreux problèmes et ceux grâce à des outils concrets tels l'administration de l'une des architectures Hadoop ou Spark, les algorithmes d'apprentissage automatique, les librairies de la science de données propres à R et python ainsi que les outils de visualisation permettant d'appréhender les 3V de la donnée : Volume, Vitesse et Variété par des stratégies d'aide à la décision et de Business Intelligence efficientes et prédictives.

Nous commencerons notre étude avec une synthèse bibliographique assez com-

plète qui se fera sur deux axes : En premier lieu, la Data Science et le Machine Learning ensuite l'analyse des sentiments. Dans le second chapitre nous allons créer et comparer les architectures Big Data les plus utilisées dans les entreprises. L'architecture Big Data étant conçue pour gérer le stockage, le traitement et l'analyse de données trop volumineuses ou complexes pour les systèmes de base de données traditionnels. Nous avons choisi de faire étape par étape l'installation et la configuration de trois architectures les plus utilisés par les experts Data Scientists en citant à chaque fois les points forts de chacune d'elles.

Le chapitre suivant abordera l'analyse des sentiments dans deux objets complexes :

- Données non structurées textuelles en « Text Mining » : les Tweets
- Données non structurées et non textuelles en « Image Processing » : les images.

Dans les activités de clientèle, les informations que renferment les données non structurées sont analysées pour améliorer le marketing relationnel et la gestion des relations clientèle. La capacité à extraire des informations à partir des données sociales étant une pratique largement adoptée par les organisations à travers le monde, nous allons, en effet, faire une extraction, analyse et traitement basiques d'opinions avec R issues de Twitter et d'un fichier texte que nous avons créé pour aboutir au nettoyage et à l'analyse de 1.6 million de tweets avec Python et la création d'un script pour l'opinion mining pour un sujet et un nombre de tweets précisés par l'utilisateur. Nous verrons ensuite comment il est possible de détecter les émotions d'un client par une analyse faciale à un instant donné par les techniques de l'apprentissage automatique.

Nous allons mener une étude statistique assez complète sur le Churn Analysis en implémentant un modèle de régression logistique pour la segmentation géodémographique afin de prédire le taux de départ des clients dans une entreprise.

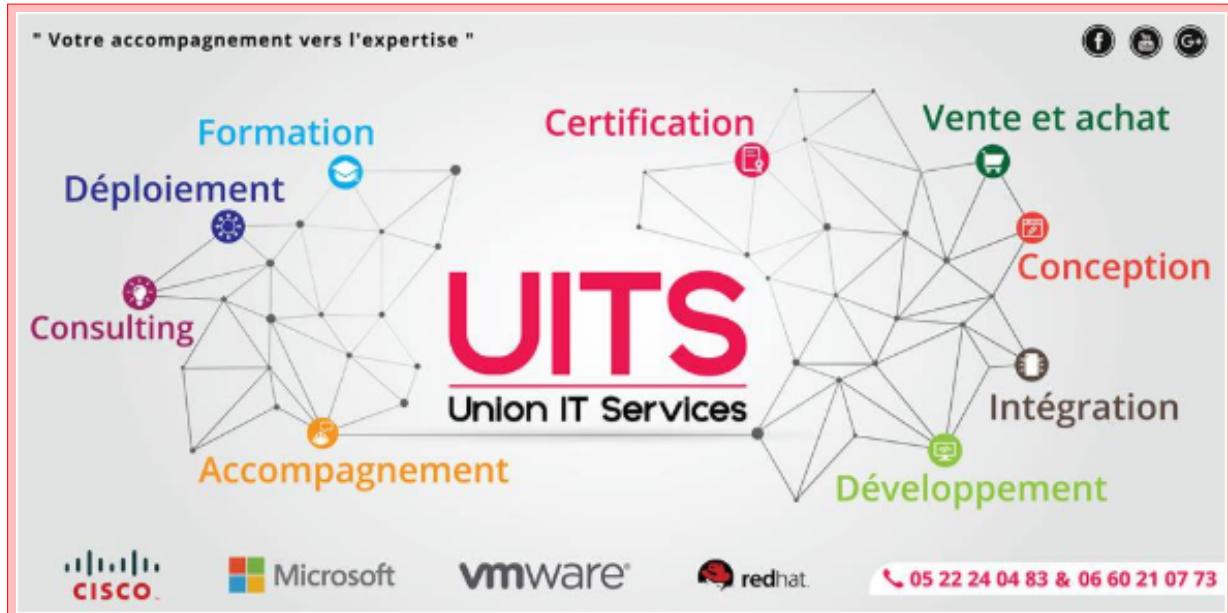
Par la suite, nous nous intéresserons à la création de Chatbots. Avec l'accord de l'entreprise, nous avons pris l'initiative de créer deux sortes de Chatbots avec Rivescript et DialogFlow (Api.ai). La première sera intégrée dans le site de l'Union It Services pour l'assistance à distance des clients effectifs et la seconde dans la Page Facebook pour marquer les formations et services de l'entreprise visant les clients potentiels.

Enfin, nous allons conclure et donner des perspectives à chacune de nos réalisations.

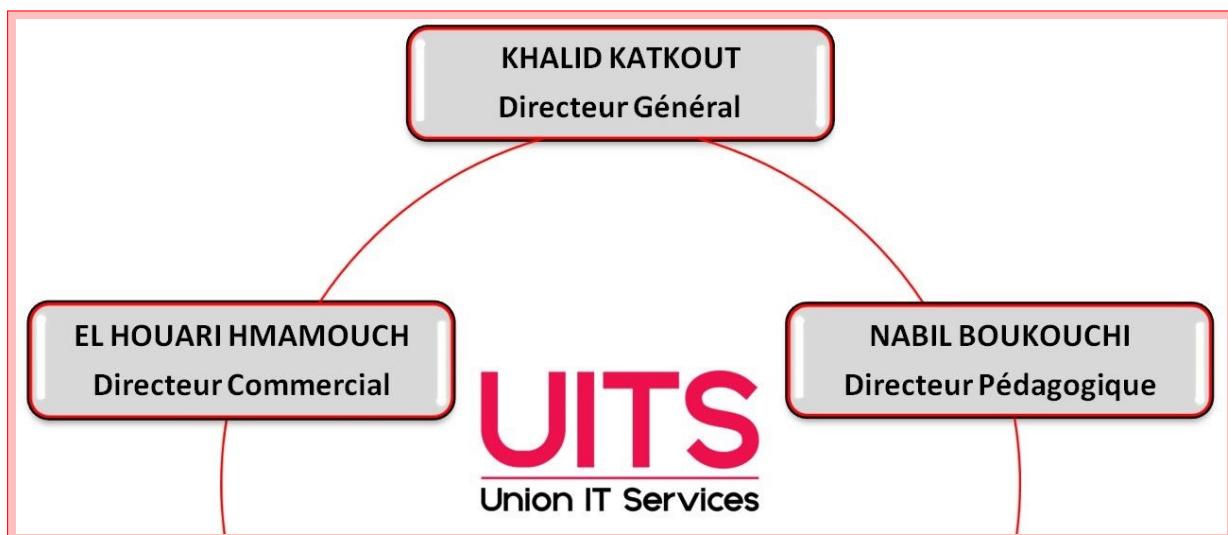
0.2. Présentation de l'entreprise

0.2 Présentation de l'entreprise

L'entreprise Union It Services est spécialisée dans tous les Services IT. Elle fournit des prestations d'audit et accompagne les entreprises dans l'étude de leurs projets, solutions IT Software et Hardware, développement applicatif et production Multimédia. C'est une start-up créée en 2017, à Casablanca. Son staff est formé de cofondateurs opérant dans les domaines suivants : Formateurs en IT et Datacenter et Administrateurs réseaux, Support technique et systèmes informatiques.



Organigramme :



Contact :



Le Data Center de L'UITs sur lequel je me suis entraîné pour créer les architectures Hadoop et Spark



0.3 La problématique

Elle s'articule autour de comment préparer une décision stratégique dans un contexte de données concurrentielles, évolutives, non structurées et gigantesques ?

0.4 Valeur ajoutée

L'intérêt principal de mon travail est d'aider l'Union It Services à prendre des décisions valables au sujet de la gestion interne, des formations dispensées, des voies possibles d'amélioration etc, en faisant entendre les avis de leur clientèle actuelle ou potentielle.

0.5 Technologies et outils utilisés

Technologies, logiciels et langages	Définitions
VMware vSphere Client 5.5	Un logiciel d'infrastructure de Cloud Computing de l'éditeur VMware, c'est un hyperviseur de type 1 (Bare Metal), basé sur l'architecture VMware ESXi.
VMware vCenter Server 6.0	Un logiciel qui offre aux administrateurs informatiques un contrôle simple et automatisé sur l'environnement virtuel pour le déploiement d'une infrastructure VMware sécurisée.
Google Cloud Plateforme	Une plateforme de cloud computing fournie par Google, proposant un hébergement sur la même infrastructure que celle que Google utilise en interne pour des produits tels que son moteur de recherche. La Google Cloud Platform est composée d'une famille de produits, chacun comportant une interface web, un outil de lignes de commande, et une interface de programmation applicative REST.
Putty	Un émulateur de terminal doublé d'un client pour les protocoles SSH, Telnet, rlogin, et TCP brut. Il permet également d'établir des connexions directes par liaison série RS-232.
Tableau	Le logiciel Tableau Desktop permet d'interroger et d'analyser rapidement vos données sous tous les angles via de simples glisser-déposer.
Gretl	Gretl (acronyme de Gnu Regression, Econometrics and Time-series Library) appartient à la famille des logiciels d'économétrie libres.
Dialogflow	C'est avant tout une interface qui va me permettre d'utiliser l'intelligence de Google. DialogFlow contient l'API Cloud Natural Language qui permet de reconnaître des phrases envoyées par l'utilisateur. Avec les phrases récupérées et un peu de machine learning, Google reconnaît la phrase, et lance en adéquation une action proposée par ma configuration. Le nom de la marque a changé récemment et s'appelle désormais "api.ai". Une fourchette de prix avec le support en bonus pour les entreprises, et gratuit avec limitation de volume pour les utilisateurs non commerciaux.

0.5. Technologies et outils utilisés

Technologies, logiciels et langages	Définitions
Firebase	Un ensemble de services d'hébergement pour n'importe quel type d'application (Android, iOS, Javascript, Node.js, Java, Unity, PHP, C++ ...). Il propose d'héberger en NoSQL et en temps réel des bases de données, du contenu, de l'authentification sociale (Google, Facebook, Twitter et Github), et des notifications, ou encore des services, tel que par exemple un serveur de communication temps réel. Lancé en 2011 sous le nom d'Envolve, par Andrew Lee et par James Templin, le service est racheté par Google en octobre 2014. Il appartient aujourd'hui à la maison mère de Google : Alphabet.
RiveScript	Un langage de script simple pour la création de chatbots avec une syntaxe conviviale et facile à apprendre.
R(logiciel : RStudio)	Un langage de programmation et un logiciel libre dédié aux statistiques et à la science des données soutenu par la R Foundation for Statistical Computing. Le langage R est largement utilisé par les statisticiens, les data miner, data scientist pour le développement de logiciels statistiques et l'analyse des données.
Python(Environnement : Canopy et Jupyter kernel 3)	Un langage de programmation objet, multi-paradigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet. Il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire par ramasse-miettes et d'un système de gestion d'exceptions ; il est ainsi similaire à Perl, Ruby, Scheme, Smalltalk et Tcl.

Présentation et contexte du projet - Introduction

OpenCV (pour Open Computer Vision)	Une bibliothèque graphique libre, initialement développée par Intel, spécialisée dans le traitement d'images en temps réel. La société de robotique Willow Garage et la société ItSeez se sont succédé au support de cette bibliothèque.
Slack	C'est une plate-forme de communication collaborative propriétaire ainsi qu'un logiciel de gestion de projets créé par Stewart Butterfield en août 2013 et officiellement lancée en février 2014.

Technologies, logiciels et langages de programmation utilisés

Chapitre 1

Synthèse bibliographique

1.1 Big Data et Machine Learning

1.1.1 Big Data - Introduction

"LES DATA SONT LE PÉTROLE DU XXI E SIÈCLE" GILLES BABINET (de L'Ère numérique, un nouvel âge de l'Humanité, JANVIER 2014)

Big Data est un terme global pour les stratégies non-traditionnelles et les technologies qui ont besoin de cueillir, organiser et traiter de grands datasets. Pendant que le problème de travailler avec les données qui excède le pouvoir informatique ou le stockage d'un ordinateur simple n'est pas nouveau, l'échelle et la valeur de ce type d'informatique s'est beaucoup développée ces dernières années.

1.1.2 Définition et avantages du Big Data

Pour comprendre la 'Big Data', nous avons besoin de savoir d'abord ce qui est "la data". Le dictionnaire d'Oxford définit 'Les données' comme : "Les quantités, les caractères ou les symboles sur lesquels les opérations sont exécutées par un ordinateur, qui peuvent être conservés et transmises sous la forme de signaux électriques et enregistrés sur les supports d'enregistrement magnétiques, optiques, ou mécaniques."

Big data se trouve sous trois formes :

— Structuré

Toutes les données qui peuvent être stockées, consultées et traitées sous la forme d'un format fixe sont appelées données «structurées». Au fil du temps, les talents en informatique ont réussi à développer des techniques pour travailler avec ce genre de données (où le format est bien connu à l'avance) et en tirer également des avantages. Cependant, de nos jours, nous prévoyons des problèmes lorsque la taille de ces données augmente considérablement, les tailles typiques sont de plusieurs zettabyte.

Exemple : Une table "Employés" dans une base de données est un exemple de données structurées

— Non structuré

Les données non structurées sont une désignation générique qui décrit toute donnée extérieure à un type de structure. Les données non structurées textuelles sont générées par les courriels, les présentations PowerPoint, les documents Word, ou encore les logiciels de collaboration ou de messagerie instantanée. Les données non structurées non textuelles, quant à elles, sont générées via des supports tels que les images JPEG, les fichiers audio MP3, ou encore les fichiers vidéo Flash. En plus de leur taille énorme, les données non structurées posent de nombreux défis en termes de traitement pour en tirer de la valeur. Aujourd’hui, les organisations disposent d’une mine de données, mais malheureusement, elles ne savent pas les mettre en valeur. Ces données sont dans leur forme brute ou format non structuré.

Exemples de données non structurées : Sortie renvoyée par "Recherche Google"

— Semi-structuré

Les données semi-structurées constituent une forme intermédiaire. Elles ne sont pas organisées selon une méthode complexe rendant possible un accès et une analyse sophistiqués ; cependant, certaines informations peuvent leur être associées, telles que des balises de métadonnées, qui permettent l’adressage des éléments qu’elles renferment.

Exemple :

Un document Word est généralement considéré comme un ensemble de données non structurées. Cependant, vous pouvez lui ajouter des métadonnées sous la forme de mots-clés qui représentent le contenu du document et qui permettent de le retrouver plus facilement lorsqu’une recherche est effectuée sur ces termes. Les données sont alors semi-structurées.

Avantages du traitement de Big Data

La capacité à traiter les «Big Data» apporte de multiples avantages, tels que :

- Les entreprises peuvent utiliser l’intelligence artificielle pour prendre des décisions.
- L’accès aux données sociales à partir des moteurs de recherche et des sites comme Facebook, Twitter permettent aux organisations d’affiner leurs stratégies d’affaires.
- Service à la clientèle amélioré : Les systèmes traditionnels de retour des clients sont remplacés par de nouveaux systèmes conçus avec les technologies «Big Data». Dans ces nouveaux systèmes, les technologies du Big Data et du traitement du langage naturel sont utilisées pour lire et évaluer les réactions des consommateurs :
- Identification précoce des risques pour le produit / les services, le cas échéant.
- Meilleure efficacité opérationnelle : Les technologies «Big Data» peuvent être utilisées pour créer une zone de transit ou une zone d’atterrissement pour de nouvelles données avant d’identifier quelles données doivent être déplacées vers l’entre�ôt de données. En outre, une telle intégration des technologies «Big Data» et de l’entre�ôt de données aide l’organisation à décharger des données rarement consultées.

1.1.3 Caractéristiques du Big Data

- Volume - Le nom "Big Data" lui-même est lié à une taille qui est énorme. La taille des données joue un rôle crucial pour déterminer la valeur des données. En outre, si une donnée particulière peut réellement être considérée comme un Big Data ou non, dépend du volume de données. Par conséquent, le «volume» est une caractéristique qui doit être prise en compte lorsqu'il s'agit de «Big Data».
- Variété - L'aspect suivant de «Big Data» est sa variété. La variété fait référence à des sources hétérogènes et à la nature des données, à la fois structurées et non structurées. Auparavant, les feuilles de calcul et les bases de données étaient les seules sources de données prises en compte par la plupart des applications. De nos jours, les données sous la forme d'e-mails, de photos, de vidéos, d'appareils de surveillance, de fichiers PDF, d'audio, etc. sont également prises en compte dans les applications d'analyse. Cette variété de données non structurées pose certains problèmes pour le stockage, l'extraction et l'analyse des données.
- Vitesse - Le terme «vitesse» fait référence à la vitesse de génération des données. La vitesse à laquelle les données sont générées et traitées pour répondre aux demandes détermine le potentiel réel des données.
Divers individus et organisations ont suggéré d'élargir les trois V originaux, bien que ces propositions aient eu tendance à décrire les défis plutôt que les qualités des mégadonnées. Quelques ajouts communs sont :
- Vérité : La variété des sources et la complexité du traitement peuvent entraîner des défis dans l'évaluation de la qualité des données (et par conséquent, la qualité de l'analyse qui en résulte)
- Variabilité : La variation des données entraîne une grande variation de qualité. Des ressources supplémentaires peuvent être nécessaires pour identifier, traiter ou filtrer les données de faible qualité afin de les rendre plus utiles.
- Valeur : Le défi ultime du big data est la création de valeur. Parfois, les systèmes et les processus en place sont suffisamment complexes pour que l'utilisation des données et l'extraction de la valeur réelle deviennent difficiles.

1.1.4 Ecosystème d'Hadoop

Hadoop Distributed File System

Hadoop est fourni avec un système de fichiers distribué appelé HDFS (HADOOP Distributed File Systems). Les applications basées sur HADOOP utilisent HDFS. HDFS est conçu pour stocker de très gros fichiers de données, fonctionnant sur des clusters de matériel de base. Il est tolérant aux pannes, évolutif et extrêmement simple à développer.

Le cluster HDFS est principalement constitué d'un NameNode qui gère les métadonnées du système de fichiers et d'un DataNodes qui stocke les données réelles.

NameNode : NameNode peut être considéré comme un maître du système. Il gère l'arborescence du système de fichiers et les métadonnées pour tous les fichiers et répertoires présents dans le système. Deux fichiers 'Namespace image' et le 'edit log' sont

utilisés pour stocker les informations de métadonnées. Namenode a connaissance de tous les datanodes contenant des blocs de données pour un fichier donné, mais ne stocke pas les emplacements de bloc de manière persistante. Cette information est reconstruite à chaque fois à partir des datanodes au démarrage du système. DataNode : DataNodes sont des esclaves qui résident sur chaque machine dans un cluster et fournissent le stockage réel. Il est responsable de servir, lire et écrire des demandes pour les clients.

Les opérations de lecture / écriture dans HDFS fonctionnent à un niveau de bloc. Les fichiers de données dans HDFS sont divisés en blocs de 64 Mo, qui sont stockés en tant qu'unités indépendantes.

HDFS fonctionne sur un concept de réPLICATION de données dans lequel plusieurs répliques de blocs de données sont créées et distribués sur des noeuds dans un cluster pour permettre une haute disponibilité des données en cas de panne de noeud.

Lire dans HDFS :

La demande de lecture de données est possible par HDFS, NameNode et DataNode. Appelons le lecteur en tant que «client». Le diagramme ci-dessous illustre l'opération de lecture de fichier dans Hadoop.

1- Le client lance la demande de lecture en appelant la méthode 'open ()' de l'objet FileSystem ; c'est un objet de type DistributedFileSystem.

2- Cet objet se connecte à namenode en utilisant RPC et obtient des informations de métadonnées telles que les emplacements des blocs du fichier.

3- En réponse à cette demande de métadonnées, les adresses des DataNodes ayant une copie de ce bloc sont renvoyées.

4- Une fois les adresses des DataNodes reçues, un objet de type FSDataInputStream est renvoyé au client. FSDataInputStream contient DFSInputStream qui prend en charge les interactions avec DataNode et NameNode. À l'étape 4 illustrée dans le diagramme ci-dessus, le client appelle la méthode 'read()' qui permet à DFSInputStream d'établir une connexion avec le premier DataNode (avec le premier bloc de fichier).

5- Les données sont lues sous la forme de flux dans lesquels le client appelle la méthode 'read()' à plusieurs reprises. Ce processus d'opération read() continue jusqu'à ce qu'il atteigne la fin du bloc.

6- Une fois la fin du bloc atteinte, DFSInputStream ferme la connexion et passe à la recherche du prochain DataNode pour le bloc suivant.

7- Une fois que le client a terminé la lecture, il appelle la méthode close().

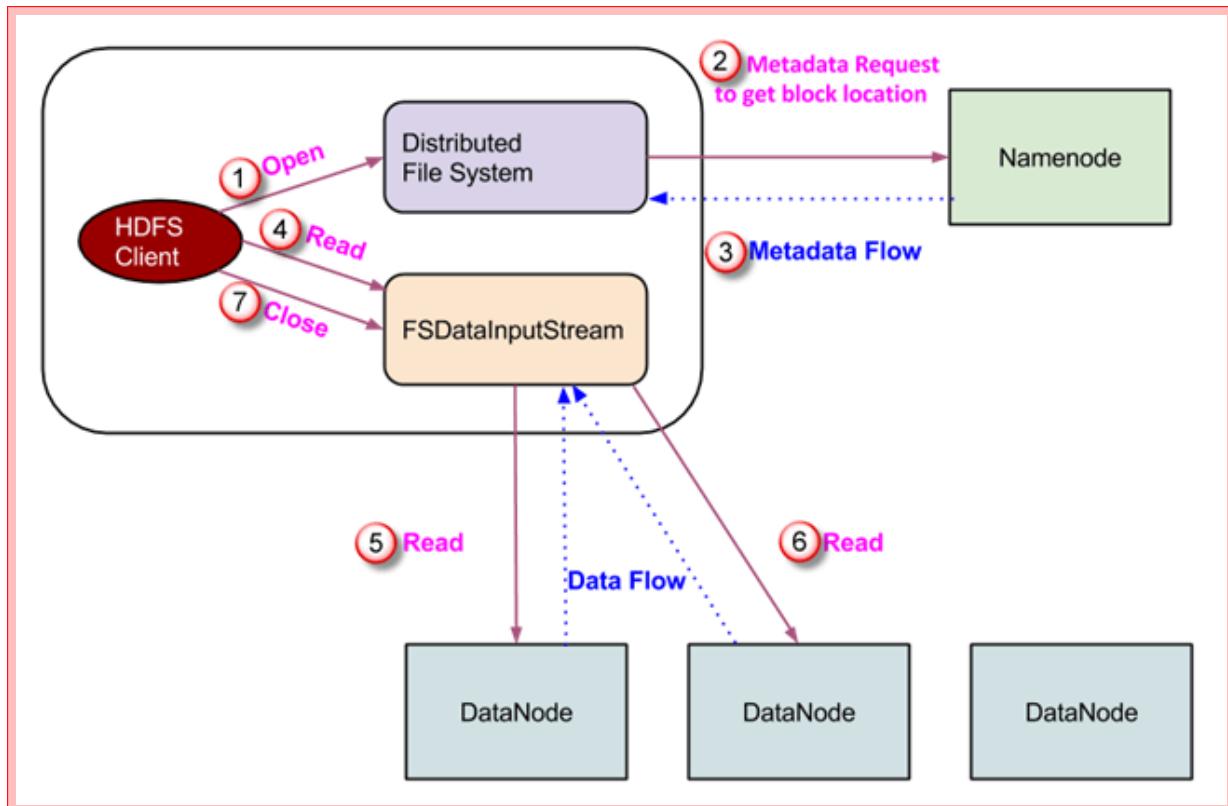


Diagramme illustrant l'opération de lecture de fichier dans Hadoop

Opération d'écriture dans HDFS :

Dans ce chapitre, nous allons comprendre comment les données sont écrites dans HDFS via des fichiers.

1- Le client lance l'opération d'écriture en appelant la méthode 'create ()' de l'objet `DistributedFileSystem` qui crée un nouveau fichier - Étape no.1 dans le diagramme ci-dessous.

2- L'objet `DistributedFileSystem` se connecte au `NameNode` à l'aide de l'appel RPC et lance la création d'un nouveau fichier. Cependant, cette opération de création de fichier n'associe aucun bloc au fichier. Il est de la responsabilité de `NameNode` de vérifier que le fichier (en cours de création) n'existe pas déjà et que le client dispose des autorisations nécessaires pour créer un nouveau fichier. Si le fichier existe déjà ou si le client n'a pas les droits suffisants pour créer un nouveau fichier, alors `IOException` est envoyé au client. Sinon, l'opération réussit et un nouvel enregistrement pour le fichier est créé par `NameNode`.

3- Une fois qu'un nouvel enregistrement est créé dans `NameNode`, un objet de type `FSDataOutputStream` est renvoyé au client. Le client l'utilise pour écrire des données dans le HDFS. La méthode d'écriture de données est invoquée (étape 3 du diagramme).

4- `FSDataOutputStream` contient l'objet `DFSOutputStream` qui s'occupe de la communication avec `DataNodes` et `NameNode`. Pendant que le client continue d'écrire des données, `DFSOutputStream` continue de créer des paquets avec ces données. Ces paquets sont mis en file d'attente appelée `DataQueue`.

5- Il y a un autre composant appelé `DataStreamer` qui consomme cette `DataQueue`.

DataStreamer demande également à NameNode d'attribuer de nouveaux blocs, choisissant ainsi les DataNodes souhaitables à utiliser pour la réPLICATION.

6- Maintenant, le processus de réPLICATION commence en créant un pipeline à l'aide de DataNodes. Dans notre cas, nous avons choisi le niveau de réPLICATION de 3 et il y a donc 3 DataNodes dans le pipeline.

7- DataStreamer déVERSE des paquets dans le premier DataNode du pipeline.

8- Chaque DataNode dans un pipeline stocke le paquet reçu par celui-ci et le transmet au second DataNode dans le pipeline.

9- Une autre file d'attente, 'Ack Queue' est maintenue par DFSOutputStream pour stocker les paquets qui attendent un accusé de réCEPTION de DataNodes.

10- Une fois que l'accusé de réCEPTION d'un paquet dans la file d'attente est reçU de tous les DataNodes dans le pipeline, il est supprimé de la 'file d'attente d'accusé de réCEPTION'. En cas de défaillance de DataNode, les paquets de cette file d'attente sont utilisés pour réINITIALISER l'opération.

11- Une fois que le client a terminé avec les données d'écRITURE, il appelle la méthode close () (étape 9 dans le diagramme). Appel à close (), entraîne le vidage des paquets de données restants dans le pipeline suivi de l'attente d'accusé de réCEPTION.

12- Une fois l'accusé de réCEPTION final reçU, NameNode est contacté pour lui indiquer que l'opération d'écRITURE de fichier est terminée.

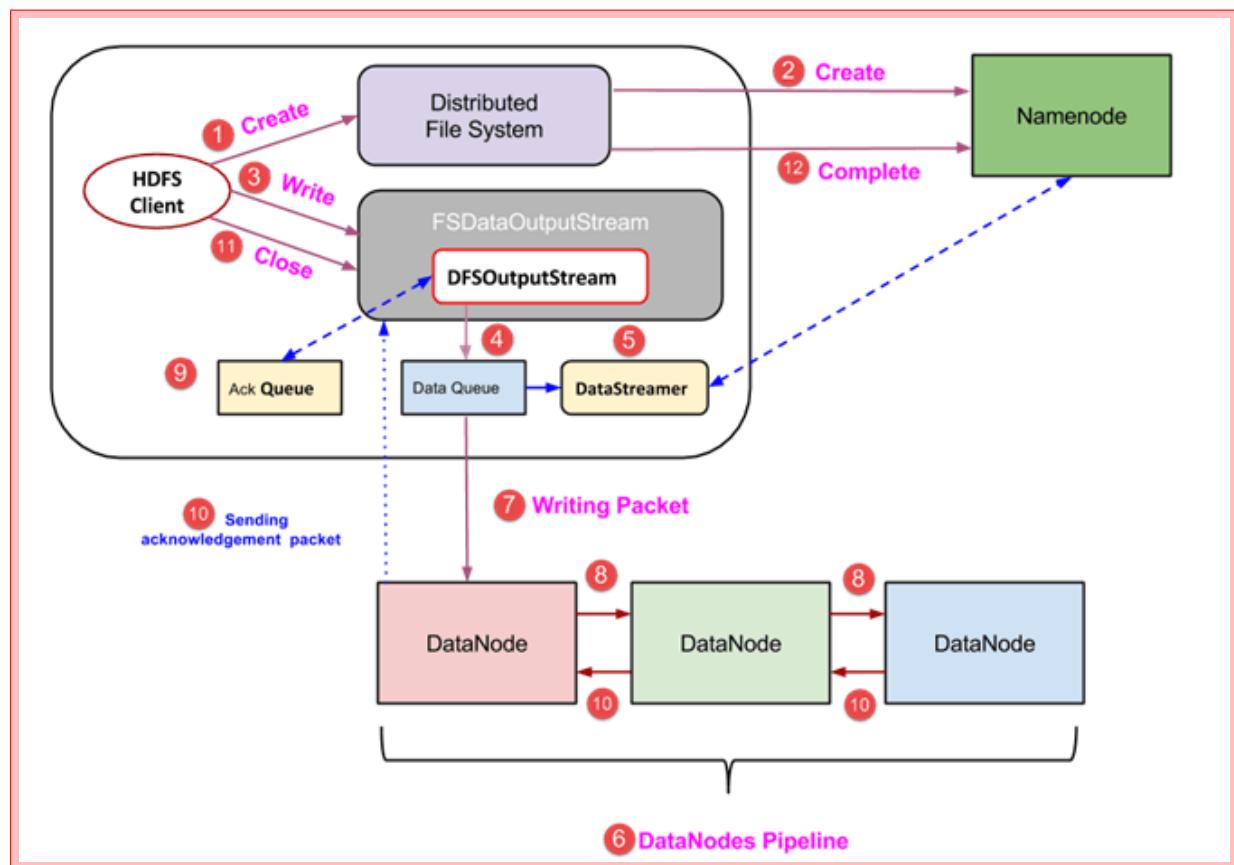


Diagramme illustrant l'opération d'écriture de fichier dans Hadoop

MapReduce et YARN

MapReduce est un modèle de programmation adapté au traitement de données volumineuses. Hadoop est capable d'exécuter des programmes MapReduce écrits en plusieurs langages : Java, Ruby, Python et C++. Les programmes MapReduce sont de nature parallèle et sont donc très utiles pour effectuer une analyse de données à grande échelle en utilisant plusieurs machines du cluster.

Apache Hadoop YARN (Yet Another Resource Negotiator) est une technologie de gestion de clusters. Elle rend l'environnement Hadoop mieux adapté aux applications opérationnelles qui ne peuvent pas attendre la fin des traitements par lots. YARN combine d'une part un gestionnaire centralisé des ressources qui harmonise l'exploitation des ressources système Hadoop par les applications et d'autre part des agents du gestionnaire de noeuds responsables de surveiller les opérations de traitement effectuées sur chaque noeud d'un cluster.

HBase

HBase est un sous-projet d'Hadoop. C'est un système - écrit en Java - de gestion de base de données non relationnelles et distribuées. Il dispose d'un stockage structuré pour les grandes tables. HBase est inspirée des publications de Google sur BigTable. Comme BigTable, c'est une base de données orientée colonnes. HBase est souvent utilisé conjointement au système de fichiers HDFS, ce dernier facilitant la distribution des données de HBase sur plusieurs noeuds. Contrairement à HDFS, HBase permet de gérer les accès aléatoires read/write pour des applications de type temps réel.

ZooKeeper

ZooKeeper est un service de coordination des services d'un cluster Hadoop, en particulier, le rôle de ZooKeeper est de fournir aux composants Hadoop les fonctionnalités de distribution. Pour cela il centralise les éléments de configuration du cluster Hadoop, propose des services de clusters et gère la synchronisation des différents éléments (événements). ZooKeeper est un élément indispensable au bon fonctionnement de Hbase.

Pig

Dans le cadre de Map Reduce, les programmes doivent être traduits en une série d'étapes Map et Reduce. Cependant, ce n'est pas un modèle de programmation avec lequel les analystes de données sont familiers. Donc, pour combler cette lacune, une abstraction appelée Pig a été construite sur Hadoop. Pig est un langage de programmation de haut niveau utile pour l'analyse de grands ensembles de données. Pig était le résultat d'un effort de développement chez Yahoo! Il permet aux utilisateurs de se concentrer davantage sur l'analyse des ensembles de données en masse et de passer moins de temps à écrire des programmes Map-Reduce.

Hive

Hive est à l'origine un projet Facebook qui permet de faire le lien entre le monde SQL et Hadoop. Il permet l'exécution de requêtes SQL sur un cluster Hadoop en vue d'analyser et d'agréger les données. Le langage SQL est nommé HiveQL. C'est un langage de visualisation uniquement, C'est pourquoi seules les instructions de type Select sont supportées pour la manipulation des données. Dans certains cas, les développeurs doivent faire le Mapping (la mise en cohérence entre deux types d'informations distincts) entre les structures de données et Hive.

Oozie

Apache Oozie est un planificateur de workflow pour Hadoop. C'est un système qui exécute le workflow des travaux dépendants. Ici, les utilisateurs sont autorisés à créer des graphes (acycliques et dirigés) de flux de travail, qui peuvent être exécutés en parallèle et séquentiellement dans Hadoop. Il se compose de deux parties :

- Le moteur de workflow dont la responsabilité d'un moteur de workflow est de stocker et d'exécuter des workflows composés de travaux Hadoop, par exemple, MapReduce, Pig, Hive.
- Le moteur de coordination qui exécute des tâches de flux de travail basées sur des calendriers prédéfinis et la disponibilité des données.

Oozie est évolutif et peut gérer l'exécution en temps opportun de milliers de workflows (chacun consistant en des dizaines de tâches) dans un cluster Hadoop. Oozie est très flexible aussi. On peut facilement démarrer, arrêter, suspendre et réexécuter des tâches. Oozie rend très facile la réexécution des workflows ayant échoué. On peut facilement comprendre à quel point il peut être difficile de rattraper les tâches manquées ou ratées en raison de temps d'arrêt ou d'échec. Il est même possible d'ignorer un noeud défaillant spécifique.

Flume

Apache Flume est un système utilisé pour déplacer des quantités massives de données en continu dans HDFS. La collecte des données de journal présentes dans les fichiers journaux des serveurs Web et leur agrégation dans HDFS pour analyse, est un exemple d'utilisation courant de Flume. Flume prend en charge plusieurs sources comme :

1. 'tail' (qui canalise les données du fichier local et écrit dans HDFS via Flume, similaire à la commande Unix 'tail')
2. Journaux système
3. Apache log4j (permet aux applications Java d'écrire des événements dans HDFS via Flume).

Sqoop

Apache Sqoop (SQL vers Hadoop) est conçu pour prendre en charge l'importation en masse de données dans HDFS à partir de magasins de données structurés tels que des bases de données relationnelles, des entrepôts de données d'entreprise et des systèmes NoSQL. Sqoop est basé sur une architecture de connecteur qui supporte les plugins pour fournir une connectivité aux nouveaux systèmes externes.

1.1.5 Les principales distributions Hadoop

Dans une distribution Hadoop on va retrouver les éléments suivants (ou leur équivalence) HDFS, MapReduce, ZooKeeper, HBase, Hive, HCatalog, Oozie, Pig, Sqoop,etc. Ces solutions sont des projets Apache et donc disponibles mais l'intérêt d'un package complet est évident : compatibilité entre les composants, simplicité d'installation, support, etc. Dans ce rapport nous évoquerons les trois distributions majeures que sont Cloudera, HortonWorks et MapR, toutes les trois se basant sur Apache Hadoop.

Cloudera

Cloudera est une start-up de la Silicon Valley, basée à Burlingame, qui se consacre au développement de solutions de type Big Data basées sur le framework Hadoop. La distribution Hadoop de Cloudera embarque plusieurs composants Open Source. Elle est déclinée en plusieurs éditions, chacune intégrant des outils d'administration et de déploiement différents.

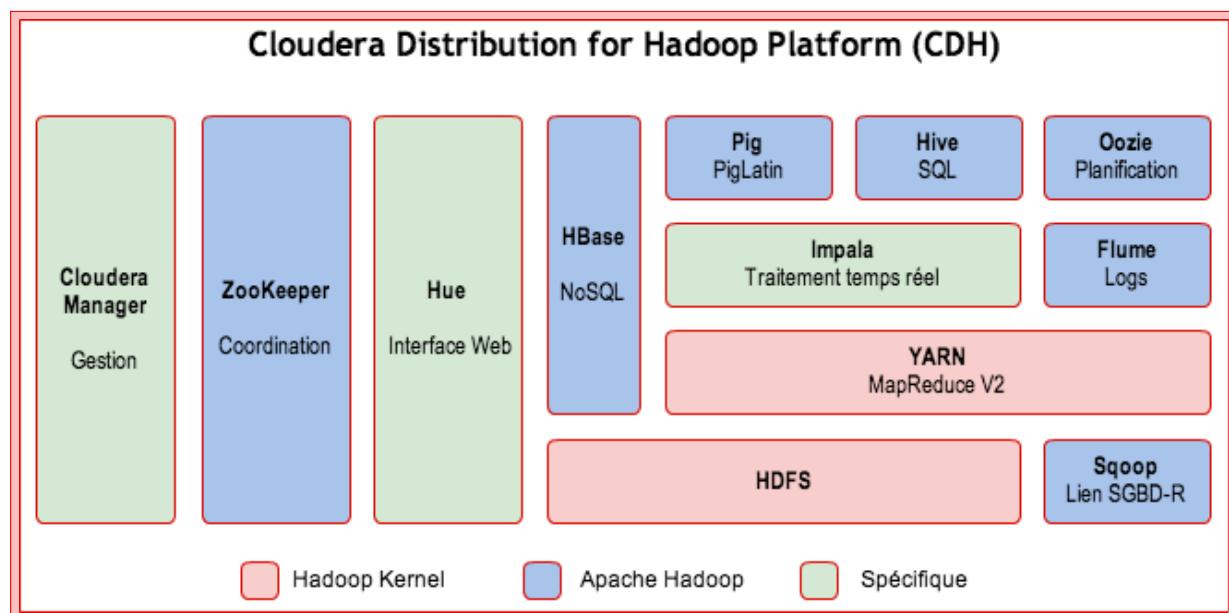


Diagramme illustrant les composants de Cloudera

Hortonworks

Hortonworks est une société de logiciels informatique basée à Santa Clara, en Californie. La société se concentre sur le développement et le soutien de Hadoop, un framework Java qui permet le traitement distribué de grands volumes de données par des grappes de serveurs. Hortonworks est une société de logiciels informatique basée à Santa Clara, en Californie. La société se concentre sur le développement et le soutien de Hadoop, un framework Java qui permet le traitement distribué de grands volumes de données par des grappes de serveurs.

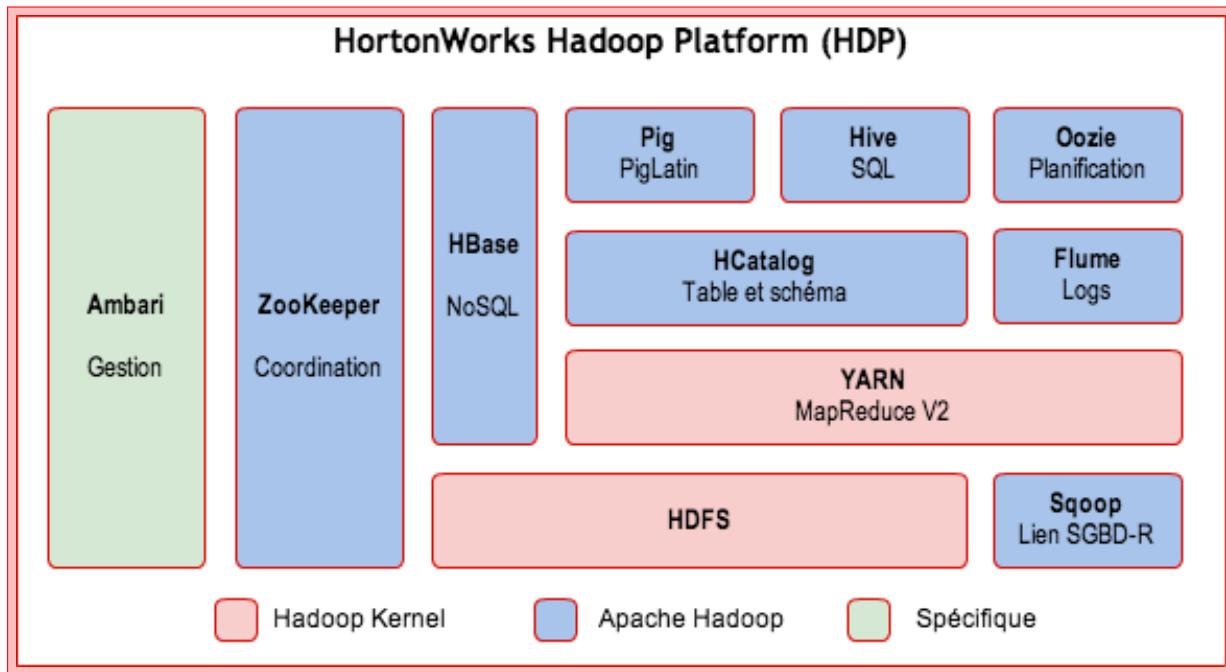


Diagramme illustrant les composants de Hortonworks

MapR

MapR est une entreprise de création de logiciels, fondée en 2009, et située à San Jose, en Californie. Elle est à l'origine de plusieurs des principaux projets open source Hadoop, dont Apache HBase, Apache Hive, Apache Zookeeper ou encore Apache Pig. Cette entreprise vend ses propres projets Hadoop à des clients en provenance de nombreuses et diverses industries telles que la vente au détail, les services financiers, les médias, la santé, la manufacture, les télécommunications et le secteur public.

Elle propose trois versions de son produit Apache Hadoop. Ces trois versions sont nommées M3, M5 et M7. M3 est une version gratuite, M5 est payante et propose davantage de fonctionnalités, tandis que M7 ajoute une version modifiée d'HBase implémentant l'API HBase directement dans le système de fichiers.

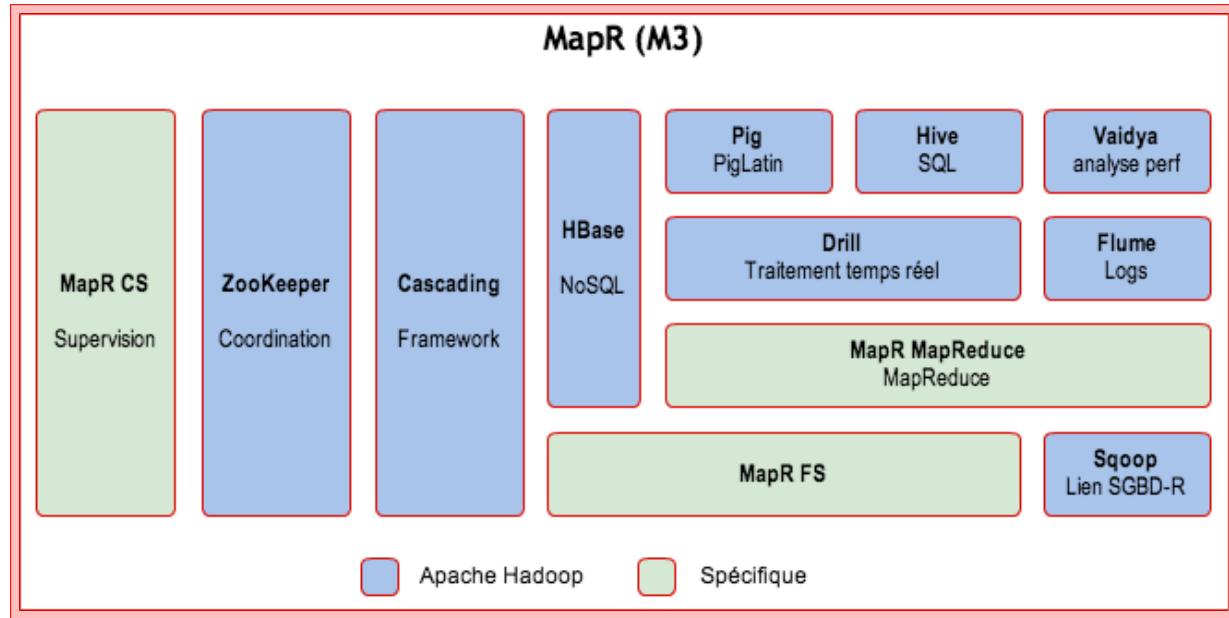


Diagramme illustrant les composants de MapR

1.1.6 Base de données NoSQL

Les bases de données NoSQL (No-SQL ou Not Only SQL) sont un sujet très à la mode en ce moment. Le terme NoSQL désigne une catégorie de systèmes de gestion de base de données destinés à manipuler des bases de données volumineuses pour des sites de grande audience. Les bases données NoSQL sont scalables, elles permettent de traiter les données d'une façon distribuée. Parmi les avantages du NoSQL on trouve :

- Leurs performances ne s'écroulent jamais quel que soit le volume traité. Leur temps de réponse est proportionnel au volume ;
- Elles se migrent facilement. En effet, contrairement aux SGBDR classiques, il n'est pas nécessaire de procéder à une interruption de service pour effectuer le déploiement d'une fonctionnalité impactant les modèles des données
- Elles sont facilement scalables. A titre d'exemple, le plus gros cluster de NoSQL fait 400 To, tandis qu'Oracle sait traiter jusqu'à une vingtaine de Téraoctet (pour des temps de réponse raisonnables).

1.1.7 L'Apprentissage Automatique ou Machine Learning

Définition 1. (Arthur Samuel) L'apprentissage automatique peut-être vu comme l'ensemble des techniques permettant à une machine d'apprendre à réaliser une tâche sans avoir à la programmer explicitement pour cela.

Définition 2. (Wikipédia) L'apprentissage automatique ou apprentissage statistique est un domaine de l'intelligence artificielle qui concerne la conception, l'analyse, le développement et l'implémentation de méthodes permettant à un ordinateur d'évoluer

par un processus systématique, et ainsi de remplir des tâches difficiles ou problématiques à remplir par des moyens algorithmiques plus classiques.

Les possibilités sont très grandes et l'on retrouve l'apprentissage automatique dans de nombreux domaines de la vie réelle comme :

- La vision par ordinateur
- La conduite automobile sans chauffeur
- La prédition de pannes
- L'identification des personnes indécises susceptibles de se laisser convaincre par un candidat à l'élection présidentielle américaine
- L'optimisation de consommation électrique chez Google
- La surveillance de la pêche illégale ou de la déforestation
- La traduction linguistique

1.1.8 Les différents types de Machine Learning

Parmi les types d'apprentissage automatique les plus répandus, on citera :

1. Supervisé,
2. Non supervisé,
3. Par renforcement,

Yann Le Cun qui est considéré comme l'un des inventeurs du Deep Learning résume ainsi ces différentes classes :

La majorité des types d'apprentissages chez l'homme et l'animal sont non supervisés. Si l'apprentissage automatique était un gâteau, l'apprentissage non supervisé serait ce gâteau, l'apprentissage supervisé serait le nappage du gâteau et l'apprentissage par "renforcement" serait la cerise sur le gâteau. Nous savons faire le nappage et la cerise, mais nous ne savons pas faire le gâteau.

Chaque type d'apprentissage peut s'appuyer sur différents algorithmes :

- **Linear Regression** (régression linéaire),
- **Logistic Regression** (régression logistique),
- **Decision Tree** (arbre de décision),
- **Random forest** (forêts d'arbres/arbes aléatoires),
- **SVM** (machines à vecteur de support),
- **Naive Bayes** (classification naïve bayésienne),
- **KNN** (Plus proches voisins),
- **Gradient Boost et Adaboost**,
- **Dimensionality reduction**,
- **Q-Learning**,
- **Réseaux de neurones ...**

Apprentissage supervisé

L'apprentissage supervisé permet de répondre à des problématiques de **classification** et de **régression**. L'idée consiste à associer un label à des données sur lesquelles vous possédez des mesures.

- Si les labels sont discrets (des libellés ou valeurs finies) on parlera de classification.
- Si au contraire les labels sont continus (comme l'ensemble des nombres réels), on parlera de régression.

1. Classification

La classification consiste à donner des étiquettes à ses données :

I- Vous disposez d'un ensemble de données connues que vous avez déjà classé (photos, plantes, individus...) Vous souhaitez, à partir de cette première classification, dite connaissance, classer de nouveaux éléments.

Certains logiciels d'albums photos utilisent ce type d'apprentissage pour classer vos images. Ils vous permettent de désigner un ensemble de photos contenant votre enfant et d'indiquer où se trouve ce dernier dans ces images. **C'est la phase d'apprentissage.**

Puis vous lui dites, voici ma collection de 15000 photos, retrouve toutes celles qui contiennent mon enfant. Le logiciel analyse alors votre collection et tente de retrouver celles qui présentent une similitude avec le jeu de données que vous lui avez enseigné. Certains logiciels vous indiquent même où se situe votre enfant dans l'image. **C'est la phase de prédiction.**

II- Un autre exemple très parlant est la détection automatique de Spams :

Vous disposez d'un grand nombre d'emails déjà classés avec une étiquette Spam/- Valide Vous souhaitez classer les nouveaux emails entrant sur la connaissance des emails déjà classés Les résultats sont généralement très bons.

Ce type de classification permet de répondre à de nombreux problèmes d'identification : reconnaissance de plantes, de personnes, de produits, reconstitution de valeurs manquantes (en remplacement d'une interpolation), etc...

Il peut utiliser différents types d'algorithmes, comme les plus proches voisins, les machines à vecteurs de supports, les arbres décisionnels, ...

2. Régression

Imaginez un ensemble de points sur une image en 2 ou 3 dimensions, ayant une intensité lumineuse différente. Par exemple une image satellite des étoiles lointaines ou encore une photo de nuit des feux des véhicules en circulation. Cette intensité lumineuse est mesurable et peut être considérée comme un label, mais il est différent pour chaque étoile. C'est un label continu qui peut prendre toute valeur au dessus de 0.

Les régressions permettent de s'approcher d'une équation idéale permettant de déterminer la luminosité de chacune de nos étoiles en fonction de leur position ou

inversement la distance des véhicules en fonction de la luminosité de leurs phares.

En astronomie ce procédé est utilisé pour identifier la distance des galaxies à partir de multiples observations comme les mesures de l'intensité lumineuse de chacune d'elles dans les différentes longueurs d'ondes. Cela s'appelle le décalage vers le rouge photométrique.

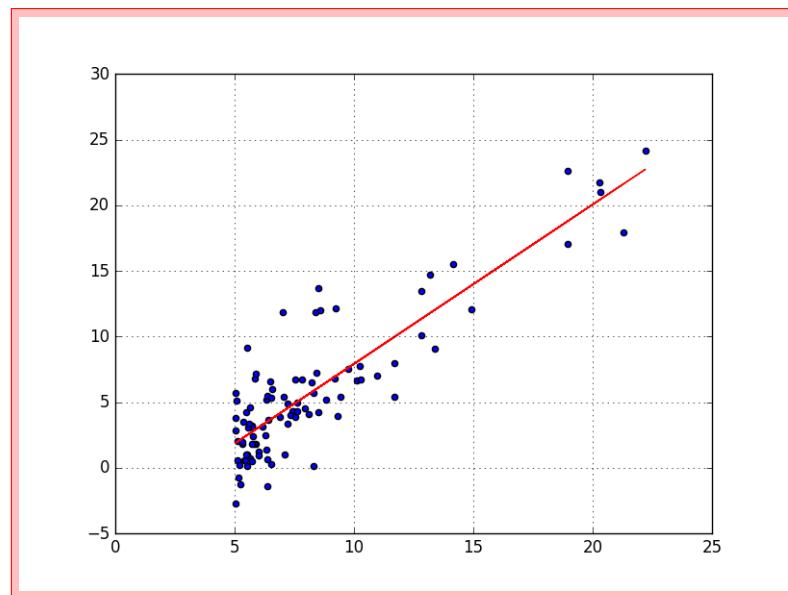
Les algorithmes tels les régressions linéaires, machines à vecteur de support ou encore les forêts d'arbres de décision sont taillés pour ce type de traitement.

1.1.9 Les principaux algorithmes

L domaine du Machine Learning regorge d'algorithmes pour répondre à différents besoins. Chacun a ses spécificités mathématiques et algorithmiques.

La régression linéaire

Les algorithmes de régression linéaire modélisent la relation entre des variables prédictives et une variable cible. La relation est modélisée par une fonction mathématique de prédiction. Le cas le plus simple est la régression linéaire univariée. Elle va trouver une fonction sous forme de droite pour estimer la relation. La régression linéaire multivariée intervient quand plusieurs variables explicatives interviennent dans la fonction de prédiction. Et finalement, la régression polynomiale permet de modéliser des relations complexes qui ne sont pas forcément linéaires.



Courbe de la régression linéaire

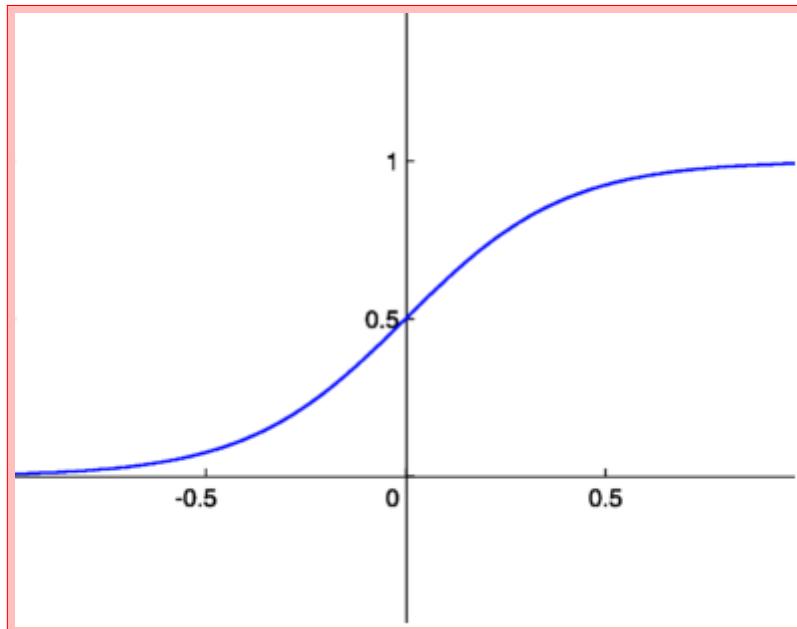
La classification naïve bayésienne

Naïve Bayes est un classifieur assez intuitif à comprendre. Il se base sur le théorème de Bayes des probabilités conditionnelles.

Naïve Bayes assume une hypothèse forte (naïve). En effet, il suppose que les variables sont indépendantes entre elles. Cela permet de simplifier le calcul des probabilités. Généralement, le Naïve Bayes est utilisé pour les classifications de texte (en se basant sur le nombre d'occurrences de mots).

La régression logistique

La régression logistique est une méthode statistique pour effectuer des classifications binaires. Elle prend en entrée des variables prédictives qualitatives et/ou ordinaires et mesure la probabilité de la valeur de sortie en utilisant la fonction sigmoïde (représentée dans la photo).



Courbe de régression logistique

On peut effectuer la classification multi-classes (par exemple classifier une photo en trois possibilités comme moto, voiture, tramway). En utilisant la régression logistique et la méthode un-contre-tous (One-Versus-All classification).

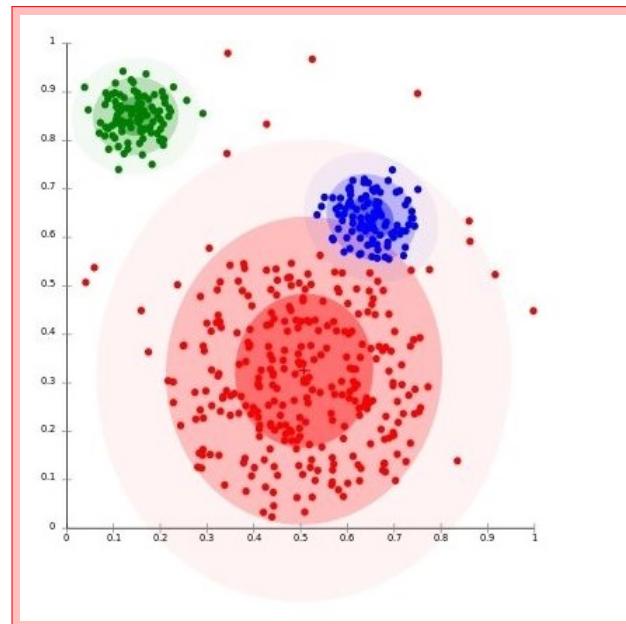
La régression logistique permettra de répondre à des problèmes comme :

Est-ce que le client est solvable pour lui accorder un crédit ?

Est-ce que la tumeur diagnostiquée est bénigne ou maligne ?

L'algorithme des k-moyennes

Imaginez que vous souhaitez lancer une campagne publicitaire et que vous voulez envoyer un message publicitaire différent en fonction du public visé. Vous devez dans un premier lieu regrouper la population ciblée sous forme de groupes. Les individus de chaque groupe auront un degré de similarité (age, salaire etc). C'est ce que fera L'algorithme K-Means !



Classification selon l'algorithme de K-means

K-Means est un algorithme de clustering en Unsupervised Learning. On lui donne un ensemble d'éléments (des données), et un nombre de groupes K. K-means va segmenter en K groupes les éléments. Le regroupement s'effectue en minimisant la distance euclidienne entre le centre du cluster et un élément donné.

Les arbres de décision

L'arbre de décision est un algorithme qui se base sur un modèle de graphe (les arbres) pour définir la décision finale. Chaque noeud comporte une condition, et les branchements sont en fonction de cette condition (Vrai ou Faux). Plus on descend dans l'arbre, plus on cumule les conditions. L'image ci-dessous illustre ce fonctionnement.

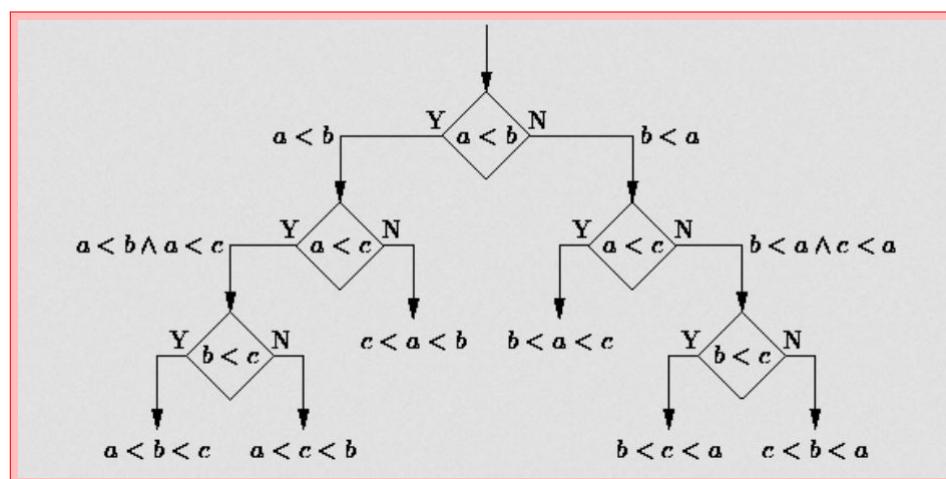
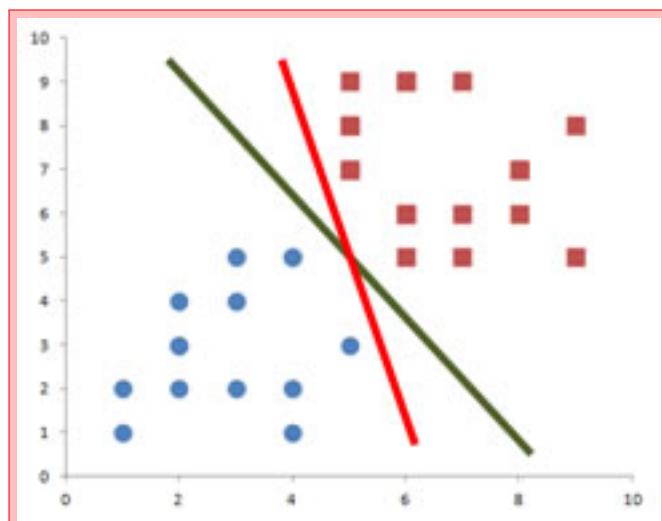


Illustration du principe des arbres de décision

Les machines à vecteurs de support

Machine à Vecteurs de Support (SVM) est lui aussi un algorithme de classification binaire. Tout comme la régression logistique. Si on prend L'image ci-dessous, nous avons deux classes (Imaginons qu'il S'agit de e-mails, et que les mails Spam sont en rouge et les non spam sont en bleu). La régression Logistique pourra séparer ces deux classes en définissant le trait en rouge. le SVM va opter à séparer les deux classes par le trait vert.

Le SVM choisira la séparation la plus nette possible entre les deux classes (comme le trait vert). C'est pour cela qu'on le nomme aussi Large Margins classifier (classifieur aux marges larges).



Principe de l'algorithme de classification binaire

Gradient Descent

Vu son importance, j'inclus L'algorithme Gradient Descent dans cette liste bien qu'il ne soit pas vraiment un algorithme de machine Learning. En effet, Gradient Descent est un algorithme itératif de minimisation de fonction de coût. cette minimisation servira à produire des modèles prédictifs comme la régression logistique et la régression linéaire.

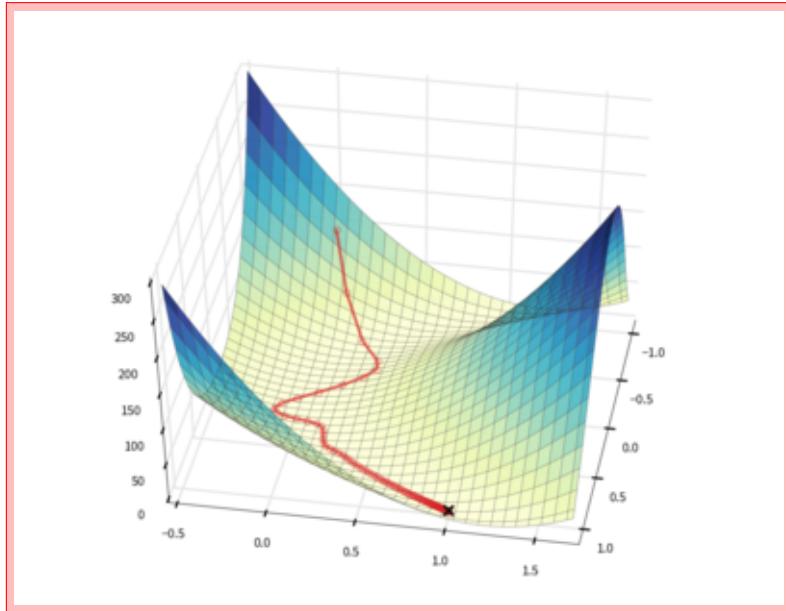
La fonction de coût d'erreur (Cost Function) :

On définit l'erreur unitaire entre une valeur observée y_i et une valeur prédictée $h(x_i)$, comme suit : $(h(x_i) - y_i)^2$

Trouver le meilleur couple (θ_0, θ_1) revient à minimiser le coût global des erreurs unitaires qui se définit comme suit : $\sum_{i=0}^m (h(x_i) - y_i)^2$; m est la taille du training set.

La fonction de coût est définie comme suit : $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=0}^m (h(x_i) - y_i)^2$ En remplaçant le terme $h(x)$ par sa valeur on obtient : $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=0}^m (\theta_0 + \theta_1 x_i - y_i)^2$

Cette formule représente la fonction de coût (cost function / Error function) pour la régression linéaire univariée.



Gradient Descent

L'algorithme peut sembler compliqué à comprendre, mais L'intuition derrière est assez simple : Imaginez que vous soyez dans une colline, et que vous souhaitez la descendre. A chaque nouveau pas (analogie à L'itération), vous regardez autour de vous pour trouver la meilleure pente pour avancer vers le bas. Une fois la pente trouvée, vous avancez d'un pas d'une grandeur α .

Dans la définition de L'algorithme on remarque ces deux termes :

le terme α S'appelle le Learning Rate : il fixe la grandeur du pas de chaque itération du Gradient Descent. $\frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$: Ce terme est la dérivée partielle pour chacun des termes θ_0 et θ_1 . Les dérivées partielles de θ_0, θ_1 sont : $\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=0}^m (h(x_i) - y_i)$

$$\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=0}^m (h(x_i) - y_i) x_i$$

A chaque itération, L'algorithme avancera d'un pas et trouvera un nouveau couple de θ_0 et θ_1 . Et à chaque itération, le coût d'erreur global se réduira.

1.1.10 Choisir son type d'apprentissage et son algorithme

Il est nécessaire de réunir plusieurs ingrédients qui doivent être savamment mélangés pour réussir une IA :

- La qualité des données
- La puissance de calcul (le matériel)
- Les algorithmes
- Et le talent

Quand on demande à un expert en apprentissage automatique quel algorithme est le mieux pour tel problème, il vous répond en général : "ça dépend, il faut en essayer plusieurs et voir celui qui fonctionne le mieux sur ce cas". En effet cela dépend :

- De la qualité des données dont vous disposez pour l'apprentissage et pour la classification

- Des paramètres que vous utilisez avec vos données
- De la quantité des données sources et à classer
- Du temps d'exécution requis
- Des paramètres disponibles pour influencer le comportement de l'algorithme
- ...

Machine Learning et Python

Python a su s'installer/s'imposer dans l'univers scientifique et industriel. Le domaine du machine learning n'est pas resté à l'écart, bien au contraire... Les formidables possibilités de calcul du langage ont permis de percer ce secteur et de multiples librairies ont vu le jour.

- Annoy : librairie extrêmement rapide implémentant la recherche des plus proches voisins
- Caffe : Deep learning framework
- Chainer : Framework intuitif pour les réseaux de neurones
- Neon : Deep Learning framework extrêmement performant
- NuPIC : Plateforme d'IA implémentant les algorithmes d'apprentissage HTM
- Shogun : Large Scale Machine Learning Toolbox
- TensorFlow : Réseau de neurones disposant d'une API de haut niveau
- Theano : Librairie d'apprentissage automatique destinée à évaluer et optimiser des expressions mathématiques
- Torch : Framework d'algorithmes d'apprentissage très performant disposant de binding Python
- Theanets : deep learning
- ...

Le plus épanté est qu'elles sont toutes généralement d'une grande qualité et utilisées dans des environnements professionnels. Toutefois, Scikit-Learn est probablement la plus populaire des librairies disponibles pour ce langage. Elle possède un grand nombre de fonctionnalités spécialisées dans l'analyse de données et le data Mining qui en font un outil de choix pour les chercheurs et développeurs.

1.2 L'analyse des sentiments

Dans un contexte de Big Data, où les données sont collectées dans des volumes importants et à une vitesse impressionnante, où les données sont de nature variées (données structurées, données non structurées comme les textes, les images, les sons), le traitement et l'analyse de cette matière première devient une étape incontournable.

Pour le traitement des données textuelles, l'analyse des sentiments est un traitement automatique des langues très utile pour automatiser la synthèse des multiples

avis pour obtenir efficacement une vue d'ensemble des opinions sur un sujet donné. En effet, les sources de données textuelles porteuses d'opinion disponibles sur le web se multiplient : avis d'internautes, forums, réseaux sociaux L'intérêt de ces données est considérable, pour les annonceurs qui souhaitent obtenir un retour client sur leurs produits ou leur image de marque, leur e-réputation, mais également pour les personnes souhaitant se renseigner pour un achat, une sortie, ou un voyage.

Une expression d'opinion possède une polarité, qui peut être soit positive, soit négative, soit neutre. La valeur neutre correspond à une opinion de polarité ambiguë, qui sera éventuellement désambiguisee par le contexte.

Les termes analyse de sentiments et fouille d'opinions sont très souvent utilisés de manière interchangeable. Selon [Pang and Lee, 2008], le terme "analyse de sentiments" est plus utilisé dans le domaine du traitement automatique du langage naturel tandis que celui de "fouille d'opinion" est adopté dans le domaine de la recherche d'information. Dans notre travail, nous emploierons les termes analyse de sentiments et fouille d'opinion de manière interchangeable.

Nous entamons ce chapitre par la présentation de quelques définitions nécessaires pour la compréhension du domaine. Nous passerons ensuite aux différents types d'opinions selon la classification trouvée dans la littérature. Nous aborderons en second lieu le domaine de l'analyse de sentiments en présentant ses différents niveaux et les tâches qui le composent. Pour finir nous nous attarderons sur les techniques utilisées dans deux principales tâches de l'analyse de sentiments, la construction du lexique d'opinion et la catégorisation de sentiments.

Les termes sentiment ou opinion sont les concepts de base du domaine en question, c'est pourquoi nous allons tenter de présenter quelques définitions afin de mettre en avant les différences entre ces derniers. Plusieurs chercheurs tentent de donner des définitions aux opinions et aux sentiments tandis que d'autres considèrent que ces derniers représentent un seul concept et accordent plus d'importance à leurs propriétés à savoir : la polarité, l'intensité etc.

1.2.1 Opinion

Une opinion est donc un jugement ou une information subjective, que porte un individu envers un sujet contrairement à un fait qui est une information objective supposée vraie à tout moment. La représentation la plus répandue de l'opinion dans la littérature est un tuple contenant ses différentes propriétés. [Kim and Hovy, 2004] utilisent un quadruplet (porteur, sujet, revendication, sentiment). [Kobayashi et al., 2007] rejoignent [Kim and Hovy, 2004] dans leur représentation en quadruplet de l'opinion, cependant les éléments qui la composent sont légèrement différents. En effet, le quadruplet en question est de la forme (porteur, objet, aspect, évaluation) où :

- le porteur représente celui qui émet l'opinion,
- l'objet est une entité représentant la cible de l'opinion,
- l'aspect est une partie spécifique de l'objet évalué,
- l'évaluation est la qualité de l'aspect qui forme l'évaluation.

[Liu, 2012] reprend à son tour le modèle de [Kim and Hovy, 2004] et celui de [Kobayashi et al., 2007] en y apportant une nouvelle dimension qui est celle du temps. De ce fait, l'opinion est représentée sous forme de quintuple à savoir (porteur, objet, aspect, sentiment, temps). Quand une opinion touche l'entité toute entière, l'aspect particulier GENERAL est utilisé.

L'opinion peut être classée selon sa nature : opinion régulière ou opinion comparative. Elle peut également être classée en se basant sur la manière avec laquelle elle est exprimée en opinion explicite et opinion implicite.

Opinion Régulière et opinion Comparative :

1. Opinion Régulière

Une opinion régulière, ou simplement opinion, se divise à son tour en deux sous-catégories d'opinion [Liu, 2012].

- Opinion Directe : est une opinion exprimée directement sur une entité ou un aspect d'entité, par exemple la qualité d'image est superbe.
- Opinion Indirecte : est une opinion exprimée indirectement sur une entité ou un aspect d'entité en se basant sur son effet sur les autres entités. Par exemple : Après l'injection du médicament, l'état de mes articulations a empiré.

2. Opinion Comparative

L'opinion comparative serait plutôt de la forme La qualité d'image de la caméra X est meilleure que celle de la caméra Y. Elle utilise donc des structures de langue différentes de celle de l'opinion régulière et à tendance à exprimer une comparaison entre deux entités ou plus à l'égard de leurs caractéristiques ou attributs communs, par exemple, la qualité de l'image [Ganapathibhotla and Liu, 2008].

Opinion Explicite et opinion Implicite

1. Opinion Explicite

Une opinion explicite est une déclaration subjective qui donne lieu à une opinion régulière ou comparative. Elle se distingue par l'utilisation des mots qui expriment ouvertement une opinion ou un sentiment.

2. Opinion Implicite

Une opinion implicite est une déclaration objective qui, elle aussi, implique une opinion régulière ou comparative. Cette déclaration est porteuse de sentiments sous-entendus positifs ou négatifs, e.g. :J'ai acheté le matelas la semaine passée, et il a déjà formé des creux.

1.2.2 Sentiment

L'analyse de sentiments ou fouille d'opinion est le domaine d'étude qui analyse les opinions, les sentiments, les évaluations, les émotions des gens à partir du langage écrit.

C'est l'un des domaines de recherche les plus actifs dans le traitement automatique du langage naturel, de la fouille de données, la fouille du web ainsi que la fouille de texte [Liu, 2012].

1.2.3 Approches de catégorisation de sentiments

Les techniques de catégorisation de sentiments peuvent être réparties en trois grandes familles , à savoir les approches basées sur le lexique, les approches basées sur l'apprentissage automatique ainsi que les approches hybrides. L'approche basée sur le lexique repose principalement sur un lexique de mots d'opinion prédéfinis ainsi que sur des règles syntaxiques et linguistiques. L'approche basée sur l'apprentissage automatique est implémentée en construisant un classifieur tandis que l'approche hybride tire le meilleur de la combinaison des deux précédentes pour atteindre une précision plus élevée. [Bahrainian and Dengel, 2013, Liu, 2012, Medhat et al., 2014a] Il faut noter que le nom "approche basée sur le lexique" donné dans la littérature à la première approche n'exclut pas l'utilisation d'un lexique dans l'approche basée sur l'apprentissage automatique. La différence est dans le fait que la première approche utilise un lexique où l'orientation de chaque mot est connue au préalable. L'approche basée sur l'apprentissage automatique à son tour représente les documents en vecteur en utilisant le lexique afin de construire un modèle de prédiction en se basant sur des exemples.

Les premiers pas de la recherche dans le domaine de l'analyse de sentiments remontent à l'an 2001 qui marque la prise de conscience des chercheurs à propos des problèmes et des opportunités que peut soulever ce domaine [Pang and Lee, 2008].

Ceci dit, la recherche connaît un essor considérable que les auteurs attribuent à plusieurs facteurs, parmi eux :

- le développement des techniques d'apprentissage automatique dans le domaine du traitement automatique du langage naturel et la recherche d'information,
- la disponibilité des données pour l'entraînement des algorithmes d'apprentissage automatique et ce grâce à l'épanouissement du web et plus spécialement, le développement des sites web offrant une agrégation des critiques,
- la réalisation de challenges intellectuels fascinants ainsi que les applications commerciales et Business Intelligence qu'offre le domaine.

1.2.4 Catégorisation basée sur le lexique

On considère que les mots et les expressions de sentiments sont les principaux indicateurs utilisés pour exprimer les sentiments, et que la polarité collective d'un document ou d'une phrase est la somme des polarités des mots ou des expressions individuelles qui le constitue [Turney, 2002].

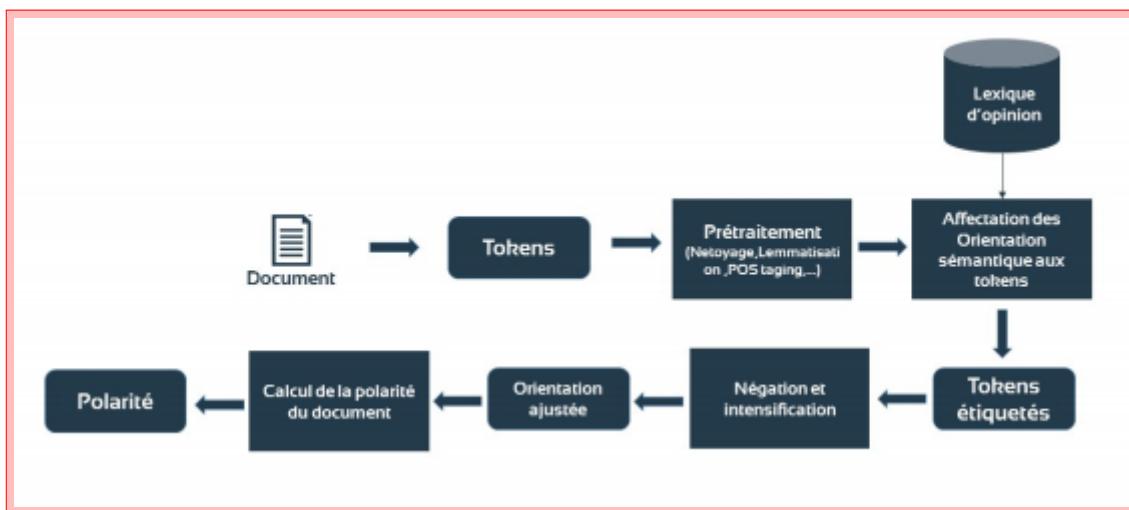
La première approche de l'analyse de l'opinion a été construite sur ces hypothèses, elle est dite une approche basée sur le lexique car elle repose essentiellement sur un lexique de mots d'opinion prédéfinis dont la polarité est connue à priori.

Principe

L'approche basée sur le lexique consiste à évaluer la polarité d'un texte à l'aide de deux groupes d'indicateurs : ceux qui expriment un sentiment positif et ceux qui expriment un sentiment négatif. Le système extrait du texte t tous les mots w positifs et négatifs et les met dans le groupe correspondant. La somme des mots $g(t)$ représente une évaluation du sentiment global dans le texte ; si la quantité des mots positifs l'emporte sur celle des mots négatifs, le système tend à dire que le texte exprime une opinion positive, dans le cas inverse, l'opinion est considérée comme négative [Turney,2002]. La formule ci-dessous explique la méthode de calcul :

$$Polarité(t) = \begin{cases} Positive & \text{si } g(t) > 0 \\ Neutre & \text{si } g(t) = 0 \\ Negative & \text{si } g(t) < 0 \end{cases} \quad \text{Ou} \quad g(t) = \sum_{w \in t} Orientation(w)$$

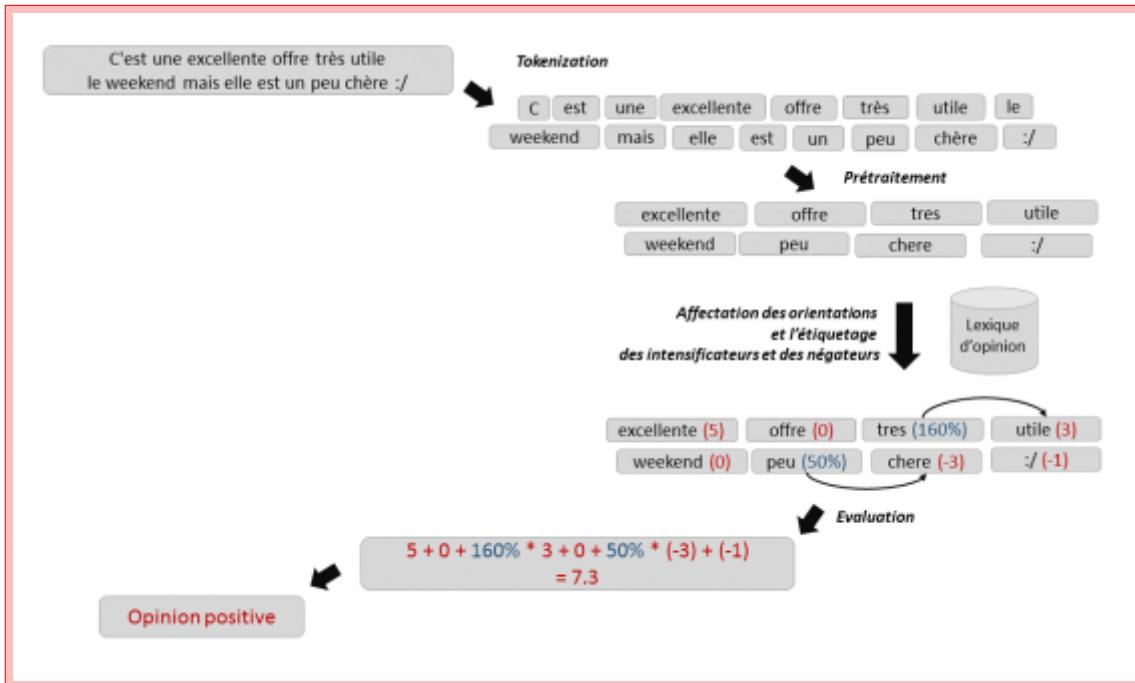
Technique



Présente les différentes tâches effectuées par le système. Une phase de tokenization du texte est nécessaire afin de représenter le texte en un ensemble d'unités linguistiques. Cette phase est suivie par le prétraitement qui consiste à normaliser les diverses manières d'écrire un même mot, à corriger les fautes d'orthographe évidentes ou les incohérences et à expliciter certaines informations lexicales comme les abréviations. Les unités sont annotées par leur orientation sémantique en utilisant un lexique d'opinion. Une phase d'ajustement d'orientation est primordiale pour traiter les négations et les intensifications. Enfin, le système calcule une évaluation de la polarité globale du texte à partir de l'orientation sémantique des unités lexicales annotées.

Nous présentons sur la figure un exemple de catégorisation basée sur le lexique. Le texte "c'est une excellente offre très utile le weekend mais elle est un peu chère :/" a un score de polarité de 7,3 suite à l'évaluation obtenue en passant par les différentes

étapes de traitement. Le score est supérieur à zéro, la polarité globale du texte est donc considérée comme étant positive.



1.2.5 Catégorisation basée sur l'apprentissage automatique

Principe

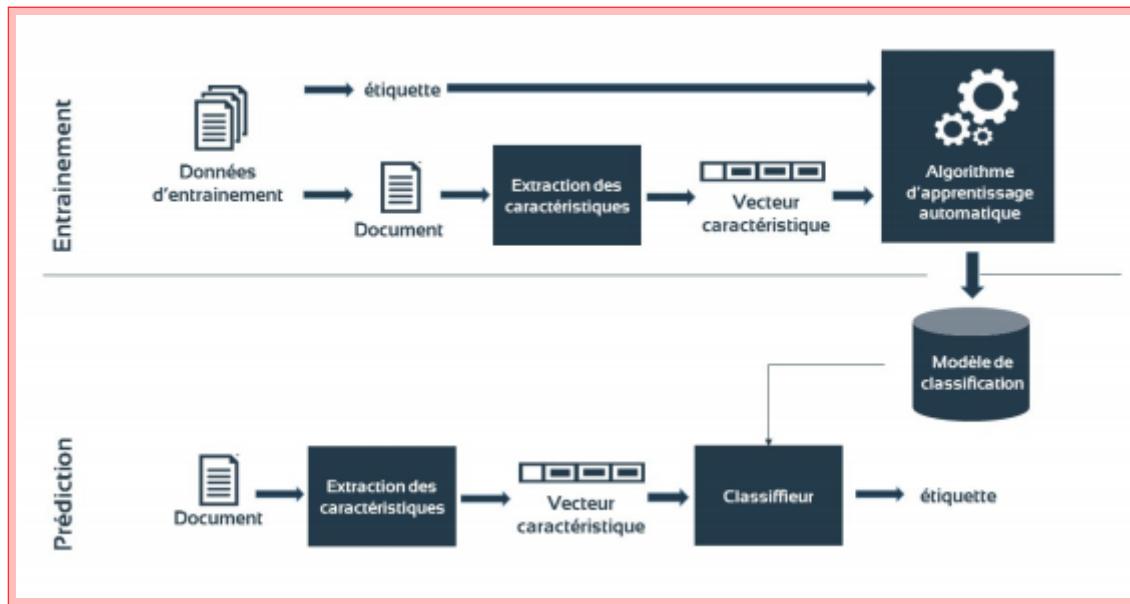
Cette approche considère l'analyse de sentiments à partir du texte comme étant un problème de catégorisation de texte [Medhat et al., 2014a]. En effet, le problème consiste en une catégorisation à un seul label sur deux classes ; positive et négative, ou bien sur trois classes ; positive, négative et neutre.

Technique

Tout système de classification basé sur l'apprentissage automatique repose sur trois points très importants : le corpus de documents étiquetés, l'extraction des caractéristiques et l'algorithme d'apprentissage automatique. La Figure indique clairement les étapes de la catégorisation à base d'apprentissage automatique en mettant en évidence la phase d'entraînement et la phase de prédiction.

Lors de la phase d'entraînement, une masse de documents étiquetés avec la classe correspondante sont représentés selon un vecteur de caractéristiques construit à partir du lexique composé des mots du corpus, ces caractéristiques sont les critères sur lesquels se base le système pour prendre une décision. Les vecteurs sont ensuite envoyés en entrée à un algorithme qui se charge d'ajuster ses paramètres de calcul pour rapprocher au plus sa prédiction des classes assignées au préalable. L'ajustement des paramètres peut passer par plusieurs itérations. A l'issue de cette phase, un modèle de

prédition est mis en place et sera utilisé lors de la phase de prédition pour assigner une classe à de nouveaux textes.



1.2.6 Catégorisation hybride

Cette approche adopte une combinaison de l'approche lexicale et l'approche basée sur l'apprentissage automatique [Medhat et al., 2014a]. Elle tend à pallier les inconvénients des deux précédentes approches en réduisant d'une part le temps d'entraînement ainsi que la quantité de données utilisées et d'autre part, la dépendance du domaine des données. Nous citons, à titre d'exemple, [Khan et al., 2015] qui ont utilisé une méthode basée sur le lexique pour étiqueter les textes et construire automatiquement un corpus d'entraînement qui servira ensuite pour entraîner un classifieur. Nous citons également [Mudinas et al., 2012] qui proposent d'utiliser un vecteur de caractéristiques construit à partir de lexique composé des mots du corpus et d'introduire une nouvelle caractéristique dans ce vecteur appelée "score" calculée par la méthode lexicale.

1.2.7 Niveaux d'analyse de sentiments

[Liu, 2012] distingue trois différents niveaux d'analyse de sentiments en se basant sur la granularité de l'unité de texte considérée par la méthode en question. Ces niveaux sont cités en partant du plus général au plus fin : le niveau document, le niveau phrase et pour finir le niveau aspect.

Nous allons, dans ce qui suit, aborder plus en détails chacun des trois niveaux en mettant en évidence leur hypothèse ainsi que leur démarche.

Niveau document

L'analyse de sentiments au niveau document part de l'hypothèse que le document exprime une seule opinion envers une seule entité provenant d'une même source. La tâche principale est donc la détermination de l'orientation générale du sentiment du document selon les classes qui peuvent être positives, négatives ou neutres. En effet, soit un document d'évaluant une entité e, le but de l'analyse est donc de déterminer le sentiment s du porteur de l'opinion p à propos de l'entité e. Le sentiment s concerne l'aspect GENERAL dans la représentation en quintuple de [Liu, 2012]. Dans la pratique, cette représentation affiche plusieurs limites. A vrai dire, un document peut évaluer plusieurs entités avec des avis différents envers ces dernières, comme il peut aborder une seule entité mais avoir des avis mitigés envers ses différents aspects. Malgré ces limites, cette représentation s'avère utile lorsque nous traitons des documents courts où l'hypothèse de départ est souvent vérifiée.

Niveau phrase

L'analyse de sentiments au niveau document est jugée trop brute pour une grande partie des applications c'est pourquoi, la recherche descend à un niveau de détail plus fin qui est le niveau phrase. La classification des sentiments au niveau phrase considère chaque phrase composant le document d comme étant une unité de base de l'analyse et part à son tour de l'hypothèse que la phrase exprime une seule opinion envers une seule entité.

L'hypothèse émise est valable quand il s'agit de phrases simples mais reste limitée quand il est question de phrases composées où une phrase peut exprimer plus d'un sentiment.

L'analyse de sentiments au niveau phrase consiste en deux tâches qui sont la catégorisation de la subjectivité et la catégorisation du sentiment. Ces deux tâches sont définies comme suit :

1. Catégorisation de la subjectivité : Cette étape classifie les phrases en deux catégories : subjective et objective. Une phrase objective exprime une information factuelle tandis qu'une phrase subjective exprime un point de vue personnel et une opinion qui peut faire référence à un sentiment positif ou négatif.
2. Catégorisation du sentiment : A l'issue de l'étape de classification de la subjectivité, si une phrase est jugée subjective, nous devons déterminer si cette dernière exprime un sentiment positif ou négatif.

Niveau aspect

Le terme aspect fait référence à un attribut ou une fonction de l'entité évaluée. Pour une analyse plus complète, il faut détecter les aspects d'un sujet et déterminer les sentiments relatifs à ces derniers [Liu, 2012]. L'objectif est de découvrir tous les quintuples (porteur, objet, aspect, sentiment, temps) dans un document d donné. Par exemple dans la phrase "La qualité d'image de la caméra est géniale mais elle est très

chère", l'analyse de sentiments au niveau aspect doit détecter un sentiment positif envers l'aspect "qualité d'image" ainsi qu'un sentiment négatif envers l'aspect "prix".

1.2.8 Tâches de l'analyse de sentiments

Nous allons, dans ce chapitre, aborder les différentes tâches qui composent un système d'analyse de sentiments. Ce plan se réfère principalement au modèle de [Liu,2012] et fournit une définition de chaque tâche. Nous nous focaliserons par la suite sur la tâche de catégorisation de sentiments.

Analyse de la subjectivité et détection de l'opinion

L'analyse de la subjectivité ou détection de l'opinion consiste à déterminer si un texte donné contient une opinion ou non. Ce problème a été abordé dans un premier temps indépendamment de l'analyse de sentiments avant de devenir une tâche de base, mais n'en reste pas moins l'une des plus difficiles.

La recherche dans la détection automatique de l'opinion à partir du texte a été initiée par [Wiebe et al., 1999] avec des travaux où ils proposent des méthodes discriminatives entre du texte objectif et du texte subjectif au niveau document, phrase et expression en utilisant un classifieur Naïve Bayes. Ce classifieur utilise un ensemble de caractéristiques à savoir la présence ou l'absence de classes syntaxiques particulières, la ponctuation et la position des phrases. Ces caractéristiques sont jugées indicatrices de subjectivité.

Par la suite, [Hatzivassiloglou and Wiebe, 2000] démontrent que les adjectifs *gradables*¹ automatiquement détectés sont une caractéristique utile pour la classification de la subjectivité. Plus récemment, [Wilson et al., 2005] ont effectué un travail pour la classification de la subjectivité au niveau document en utilisant l'algorithme des k plus proches voisins basé sur le nombre total de mots et expressions de subjectivité dans chaque document.

Catégorisation de sentiments

La tâche de catégorisation de sentiments consiste à déterminer si un texte exprime une opinion positive ou négative de son auteur vis-à-vis d'un sujet du texte. Cette tâche utilise les techniques de traitement du langage naturel et d'apprentissage automatique qui seront détaillées par la suite.

Identification du sujet et du porteur d'opinion

Une autre tâche de base de l'analyse de sentiments est la détection du porteur d'opinion et l'identification du sujet. L'avantage de cette tâche est de pouvoir filtrer

1. Des adjectifs qui peuvent être employés avec des intensificateurs tels que très ou peu. Par exemple l'adjectif grand est gradable tandis que solaire ne l'est pas.

les opinions selon un sujet particulier ou alors de regrouper les opinions d'une personne particulière pour des fins de personnalisation en sélectionnant les sujets que ce dernier préfère.

Résumé de l'opinion

Les applications de l'analyse de sentiments requièrent l'étude des opinions de beaucoup de personnes car un seul avis ne suffit pas, de ce fait, une certaine forme de résumé s'impose [Liu, 2012]. La récapitulation d'opinions consiste finalement à générer un résumé concis et digeste d'un grand nombre d'opinions. [Hu and Liu, 2004] sont les premiers à proposer des résumés basés sur les aspects à partir de critiques de clients vis-à-vis des produits vendus en ligne. Ils résument leur travail en trois étapes :

1. Identification des aspects du produit que les clients ont mentionnés dans leurs opinions,
2. Identification des phrases qui contiennent une opinion positive ou négative pour chaque aspect,
3. Production d'un résumé en utilisant les informations découvertes.

Détection de l'ironie et du sarcasme

Le *sarcasme*² et l'*ironie*³ sont considérés dans l'analyse de sentiments comme des modificateurs de la polarité, de la même manière que la négation [Liu, 2012]. La détection de ces derniers est importante pour identifier correctement les opinions présentes dans les textes. La compréhension des phrases sarcastiques n'est pas toujours facile, même pour les humains, ainsi une solution informatique est une tâche intéressante et difficile. L'approche générale pour la détection du sarcasme est basée sur l'apprentissage automatique en utilisant des traits lexicaux simples en complément de dictionnaires [Davidov et al., 2010, González-Ibáñez et al., 2011, Rilo et al., 2013].

Détection des spams

Aujourd'hui, à travers les réseaux sociaux, les blogs et les micro-blogs, il est très facile pour les gens d'exprimer leurs opinions d'une façon anonyme. Malgré ses avantages, l'anonymat a produit de nouvelles difficultés pour l'analyse de l'opinion. Il permet aux gens avec des intentions malveillantes de fausser les résultats des systèmes en postant de faux avis afin de promouvoir ou de discréditer des produits cibles, des services, des organisations ou des individus sans divulguer leurs véritables intentions.

La tâche de la détection des spams vise essentiellement à repérer ces gens (les spammeurs d'opinion) afin d'assurer la fiabilité des sources. Contrairement à l'extraction d'opinions, la détection de spams n'est pas seulement un problème de traitement du langage naturel car elle est considérée aussi comme étant un problème d'extraction de données.

2. Désigne le fait de dire le contraire de ce que l'on pense sans laisser de signes indicatifs.
 3. Consiste à dire le contraire de ce que l'on pense en le faisant comprendre par des signes.

1.2.9 Préparation des données

Un des points les plus importants dans l'apprentissage automatique supervisé est sans aucun doute la disponibilité des données labellisées au préalable. En effet, la phase d'apprentissage requiert une grande quantité d'exemples dont la classe est déjà connue. Ceci permettra à la fonction de prédire la classe de nouveaux documents en fonction des exemples déjà rencontrés lors de la phase d'apprentissage.

De plus, les données labellisées doivent être représentatives des données qui devront être catégorisées par la suite grâce au modèle construit. A titre d'exemple, si nous souhaitons prédire les sentiments des commentaires issus de réseaux sociaux, le corpus doit contenir le même type de documents et couvrir toutes les classes.

Définition

[Sinclair, 1996] définit un corpus comme étant une collection de textes sous forme électronique, sélectionnés selon des critères externes dans le but de représenter, autant que possible, un langage ou une variété et faire office de source de données pour des recherches en linguistique. Les critères externes font référence à des notions comme le type de texte, le domaine ou encore l'opinion. [Biber, 1993] définit la représentativité d'un échantillon par sa capacité à couvrir la gamme de variabilité de toute la population.

Corpus d'entraînement et corpus de test

1. Corpus d'entraînement ou corpus d'apprentissage :

Constitue la majeure partie de l'ensemble initial allant de 70 à 80 %. Les documents du corpus d'apprentissage sont utilisés comme exemples pour construire le modèle de prédiction.

2. Corpus de test

Cet ensemble de documents, composé du reste des documents, sert à évaluer la qualité de l'apprentissage effectué sur le corpus d'apprentissage et mesurer la précision de la prédiction sur de nouveaux documents.

1.2.10 Prétraitement

Les textes contenus dans le corpus jusqu'à présent sont brutes, ils ont donc besoins de quelques traitements additionnels avant de passer à la phase d'analyse, ces traitements rentrent dans le cadre de la phase de prétraitement.

Définition

[Haddi et al., 2013] dénissent le prétraitement des données comme étant le processus de nettoyage et de préparation du texte pour la classification. Ils affirment que les

textes disponibles en ligne contiennent beaucoup de bruit et de parties non informatives comme les balises HTML ou les scripts. De plus, beaucoup de mots trouvés dans les textes n'ont aucun impact sur l'orientation générale de ce dernier.

C'est une des étapes parmi les plus importantes mais aussi les plus gourmandes en termes de temps dans le processus global. En effet, tout processus traitant du texte se doit de réduire au maximum les données incomplètes, bruitées et incompatibles et de ne garder que les mots susceptibles de porter une information utile à l'analyse. Cette phase permet de transformer les données textuelles brutes en une structure adaptée à la fouille. La phase fait appel à des techniques comme la suppression des mots vides ou encore la lemmatisation, des techniques que nous détaillerons dans la suite du rapport.

La qualité de l'analyse est étroitement liée à l'efficacité du prétraitement appliqué. En effet, le fait de garder tous les mots sous une forme brute augmente la dimensionnalité du problème et rend la distinction entre les documents de différentes classes plus difficile.

Techniques

La phase de prétraitement fait appel à des techniques qui peuvent modifier la forme d'un mot ou l'éliminer complètement. Les techniques en question sont, à titre d'exemple, la suppression des caractères spéciaux, les marques de ponctuation, la normalisation de la casse ou encore la suppression des mots vides. On peut également appliquer des fonctions plus élaborées comme la lemmatisation ou la racinisation.

Le choix des techniques et l'ordre de leur application est déterminé en fonction des besoins, du contexte et du type de données. En effet, certaines techniques sont pertinentes pour un certain type de textes tandis que d'autres ne le sont pas. Nous allons dans ce qui suit présenter les méthodes de prétraitement les plus utilisées dans la fouille de textes.

1.2.11 Extraction des caractéristiques

La tokenisation : L'opération qui permet de partitionner une chaîne de caractères en ses composants, c'est une étape fondamentale pour toute tâche de traitement automatique du langage naturel. Il n'y a pas de méthode formelle et juste pour faire la tokenisation, la technique dépend étroitement de son application.

Cette étape est d'autant plus importante pour l'analyse de sentiments car l'information utile pour déterminer la polarité du texte est souvent représentée de manière particulière à l'image des émoticônes, exemple ":-)". Dans le cas présent, la tokenisation doit absolument capturer les émoticônes en plus des autres termes.

Suppression des mots vides : Les mots qui sont les plus fréquemment utilisés dans le langage sont souvent inutiles et sont vides de tout sens pour l'analyse de sentiments. Ces mots sont appelés "mots vides" et ne sont que des mots outils tels que les pronoms ou les prépositions de conjonction. Il faut donc supprimer ces mots vides, une étape qui a bien fait ses preuves dans les rapports de [Xue and Zhou, 2009] qui ont démontré l'apport de cette suppression sur la précision du système. Cette étape permet de

réduire le nombre de mots à considérer dans la représentation des documents en éliminant les mots les moins signifiants pour la classification.

Racinement : La stemmatisation ou racinement est le nom donné au procédé qui vise à transformer les *flexions*⁴ en leur radical ou stemme en ôtant les suffixes et les préfixes grâce à des algorithmes bien connus. Cette étape permet de considérer les mots qui partagent la même racine comme une seule entité dans le vocabulaire. Beaucoup d'algorithmes de racinement ont été proposés depuis les années 1960, nous citons celui de Porter qui est de loin le plus utilisé [Jivani et al., 2011].

La façon de procéder est commune à tous les stemmatiseurs, ils se basent sur certaines règles qui dépendent de la langue pour supprimer de manière grossière les suffixes dans l'espoir de retrouver la forme correcte du radical, ceci n'étant pas toujours le cas. La racinement ne prend pas en compte le contexte ni la nature grammaticale du mot (nom, verbe ou autre), de plus, elle ne puise pas dans des dictionnaires pour trouver le lemme ce qui explique certains cas où le lemme n'est pas un mot connu et correctement écrit [Stanford, 2008]. Pour illustrer le processus, la racinement permet de réduire les mots "connecter" et "connection" à une seule racine qui est "connect".

Lemmatisation : La lemmatization désigne l'analyse lexicale du contenu d'un texte qui permet de regrouper les mots d'une même famille. Chacun des mots se trouve ainsi réduit en une entité appelée lemme (forme canonique). La lemmatization regroupe les différentes formes que peut prendre un mot, soit : le nom, le pluriel, le verbe à l'infinitif ou encore sa conjugaison [Stanford, 2008].

Contrairement à la racinement qui fonctionne uniquement avec une base de connaissance des règles syntaxiques et grammaticales de la langue, la lemmatization repose sur une base de connaissances de toutes les formes dérivées de la langue. On associe à chaque forme un seul lemme possible. En reprenant l'exemple présenté plus haut, le lemmatiseur classerait les mots "connection", "connecter", "connectivité" ou encore "connecterait" dans la même famille.

Autres techniques : Le prétraitement peut inclure d'autres étapes comme la **normalisation de la casse** ou encore la **suppression des accents**. Le contexte des réseaux sociaux implique, lui aussi, un certain nombre de traitements particuliers, nous citions :

1. **La suppression des liens hypertexte :** En général les URLs n'apportent aucune information dans le cas d'analyse de sentiments dans du texte, bien au contraire ils peuvent fausser la prédiction de la subjectivité et la polarité d'un texte donné. Prenons pour exemple www.excellent.dz, ce texte aurait eu une classe positive alors qu'il est totalement neutre, cette mauvaise prédiction serait due à la présence d'un mot d'opinion dans l'URL.
2. **La suppression des lettres répétées :** Les internautes ont tendance à utiliser une suite de lettres répétées dans un même mot pour exprimer l'intensité du sens que porte ce dernier. Ceci dit, cette répétition des lettres génèrent des mots incorrects et absents des dictionnaires c'est pourquoi il faut éliminer les lettres obsolètes.
3. **Substitution des hashtags :** Le hashtag est un marqueur qui prend la forme d'un ou plusieurs mots accolés précédés par un dièse. Cette manière d'écrire est couramment utilisée dans les réseaux sociaux afin de marquer un contenu avec des
4. les formes différentes que peut prendre un mot lorsqu'il est accordé, décliné ou conjugué

mots clés dans le but de mettre l'accent sur eux.

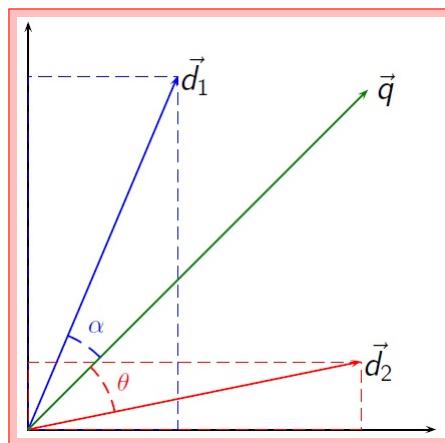
1.2.12 Représentation

A l'issue de la phase de prétraitement, il faut avoir une représentation formelle et efficace pour les documents en fonction des caractéristiques choisies avant d'effectuer une analyse. Pour ce faire, plusieurs modèles ont été proposés, le plus connu étant le modèle vectoriel proposé par [Salton et al., 1975].

Le modèle vectoriel est une représentation mathématique du contenu d'un document, selon une approche algébrique. L'ensemble de représentation des documents est un vocabulaire comprenant des termes d'indexation. Ceux-ci sont typiquement les mots les plus significatifs du corpus considéré : noms communs, noms propres, adjetifs... Ils peuvent éventuellement être des constructions plus élaborées comme des expressions ou des entités sémantiques. À chaque élément du vocabulaire est associé un index unique arbitraire.

Chaque contenu est ainsi représenté par un vecteur v , dont la dimension correspond à la taille du vocabulaire. Chaque élément v_i du vecteur v consiste en un poids associé au terme d'indice i et à l'échantillon de texte. Un exemple simple est d'identifier v_i au nombre d'occurrences du terme i dans l'échantillon de texte. La composante du vecteur représente donc le poids du mot i dans le document.

Proximité entre documents Étant donnée une représentation vectorielle d'un corpus de documents, on peut introduire une notion d'espace vectoriel sur l'espace des documents en langage naturel. On en arrive à la notion mathématique de proximité entre documents.



Proximité entre documents

En introduisant des mesures de similarité adaptées, on peut quantifier la proximité sémantique entre différents documents. Les mesures de similarité sont choisies en fonction de l'application. Une mesure très utilisée est la similarité cosinus, qui consiste à quantifier la similarité entre deux documents en calculant le cosinus entre leurs vecteurs. La proximité d'une requête q à un document d_1 sera ainsi donnée par :

$$\cos \alpha = \frac{\mathbf{d}_1 \cdot \mathbf{q}}{\|\mathbf{d}_1\| \|\mathbf{q}\|}$$

En conservant le cosinus, nous exprimons bien une similarité. En particulier, une valeur nulle indique que la requête est strictement orthogonale au document. Physiquement, cela traduit l'absence de mots en commun entre \mathbf{q} et \mathbf{d}_1 . Parmi les applications existantes, on peut citer :

La catégorisation : regrouper automatiquement des documents dans des catégories pré-définies.

La classification : étant donné un ensemble de documents, déterminer automatiquement les catégories qui permettront de séparer les documents de la meilleure façon possible (catégorisation non supervisée).

La recherche documentaire : trouver les documents qui répondent le mieux à une requête (ce que fait un moteur de recherche) ; la requête de l'utilisateur est considérée comme un document, traduite en vecteur, et comparée aux vecteurs contenus dans le corpus des documents indexés.

Le filtrage : classer à la volée des documents dans des catégories pré-définies (par exemple, identifier un spam sur la base d'un nombre suspect d'occurrence d'un mot suspect dans un courriel et l'envoyer automatiquement à la corbeille).

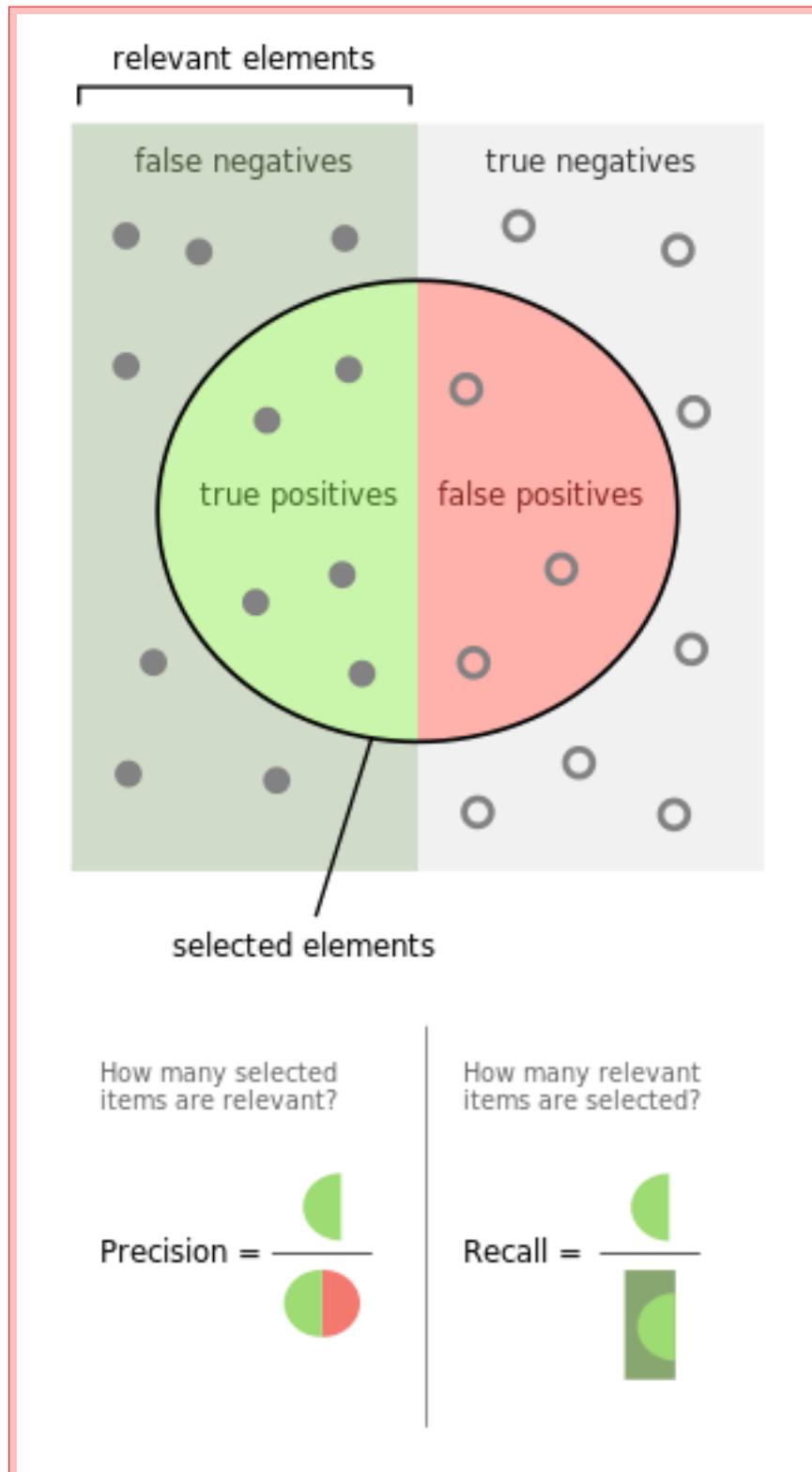
1.2.13 Évaluation

Les chercheurs utilisent les méthodes d'évaluation appliquées à la recherche d'information pour mesurer les performances des systèmes d'analyse de sentiments. Nous citons le rappel, la précision et la F-Mesure. Les faux positifs et les faux négatifs sont deux concepts importants dans la mesure de performances de la classification. On parle de faux négatif quand le résultat du test indique que le fait étudié est négatif alors que dans la réalité, celui-ci est positif. Par analogie, on dit d'un test que le fait est faux positif s'il indique qu'un cas est positif alors que ce dernier est négatif. Nous introduisons donc la matrice de confusion représentée ci-dessous qu'on peut générer pour chacune des classes.

		Prédiction	
		Positive	Négative
Réalité	Positive	TP	FN
	Négative	FP	TN

Matrice de confusion

La précision compte la proportion d'items pertinents parmi les items sélectionnés alors que le rappel compte la proportion d'items pertinents sélectionnés parmi tous les items pertinents sélectionnables.



Précision et rappel (« recall »)

Précision

La précision est le nombre de documents pertinents retrouvés rapporté au nombre de documents total proposé par le moteur de recherche pour une requête donnée.

Le principe est le suivant : quand un utilisateur interroge une base de données, il souhaite que les documents proposées en réponse à son interrogation correspondent à son attente. Tous les documents retournés superflus ou non pertinents constituent du bruit. La précision s'oppose à ce bruit documentaire. Si elle est élevée, cela signifie que peu de documents inutiles sont proposés par le système et que ce dernier peut être considéré comme "précis". On calcule la précision avec la formule suivante :

$$\text{précision}_i = \frac{\text{nb de documents correctement attribués à la classe } i}{\text{nb de documents attribués à la classe } i}$$

Pour simplifier les formules, nous utilisons les données de la table de confusion. La précision se calcule donc comme suit :

$$\boxed{\text{Précision} = \frac{\text{TP}}{\text{TP} + \text{FP}}}$$

Rappel

Le rappel est défini par le nombre de documents pertinents retrouvés au regard du nombre de documents pertinents que possède la base de données. Cela signifie que lorsque l'utilisateur interroge la base il souhaite voir apparaître tous les documents qui pourraient répondre à son besoin d'information. Si cette adéquation entre le questionnement de l'utilisateur et le nombre de documents présentés est importante alors le taux de rappel est élevé. À l'inverse si le système possède de nombreux documents intéressants mais que ceux-ci n'apparaissent pas dans la liste des réponses, on parle de silence. Le silence s'oppose au rappel.

$$\text{rappel}_i = \frac{\text{nb de documents correctement attribués à la classe } i}{\text{nb de documents appartenant à la classe } i}$$

En statistique, le rappel est appelé sensibilité. De même que pour la précision, nous simplions la formule du rappel comme suit :

$$\boxed{\text{Rappel} = \frac{\text{TP}}{\text{TP} + \text{FN}}}$$

Dans le cadre multi-classes (où n est supérieur à 1), les moyennes globales de la précision et du rappel sur l'ensemble des classes i peuvent être évaluées par la macro-moyenne qui calcule d'abord la précision et le rappel sur chaque classe i suivie d'un calcul de la moyenne des précisions et des rappels sur les n classes :

$$\text{précision} = \frac{\sum_{i=1}^n \text{précision}_i}{n}$$

$$\text{rappel} = \frac{\sum_{i=1}^n \text{rappel}_i}{n}$$

F-Mesure

Une mesure populaire qui combine la précision et le rappel est leur moyenne harmonique, nommée F-mesure (soit F-measure en anglais) ou F-score :

$$F = 2 \cdot \frac{(\text{précision} \cdot \text{rappel})}{(\text{précision} + \text{rappel})} \xrightarrow{\text{Le Cas Général}} F_\beta = \frac{(1 + \beta^2) \cdot (\text{précision} \cdot \text{rappel})}{(\beta^2) \cdot \text{précision} + \text{rappel}}$$

Chapitre 2

Contribution

2.1 Mise en oeuvre et applications :

2.1.1 Cr ation et comparaison d'architectures "Hadoop" et "Spark"

Selon le cabinet, entre 60 % et 73 % des donn es accessibles aux entreprises sont inutilis es pour la Business Intelligence et les analyses. Allant du fait qu'il n'y a pas de projet Big Data sans architecture Big Data capable de traiter les donn es vari es (structur , semi-structur  et non structur ), volumineuses et qui arrivent   des vitesses diff rentes, les analystes Big Data de Forrester Research, consid rent qu'adopter Hadoop est indispensable pour toute organisation souhaitant entreprendre une strat gie d'analyse de donn es. Cependant bien que Hadoop domine le march , Apache Spark demeure la solution   la mode dans le monde complexe du Big Data. En effet, au del  du buzz, Spark dispose de vrais atouts. Dans cette partie, nous avons cr er et d ployer trois topologies modernes les plus utilis es par les entreprises dans le monde,   savoir : Cloudera et Hortonworks qui sont propres   Hadoop Vs Apache Spark. Ces r alisations se feront de mani res totalement diff rentes, tout en les comparant entre-elles pour voir les atouts et limites de chaque solution.

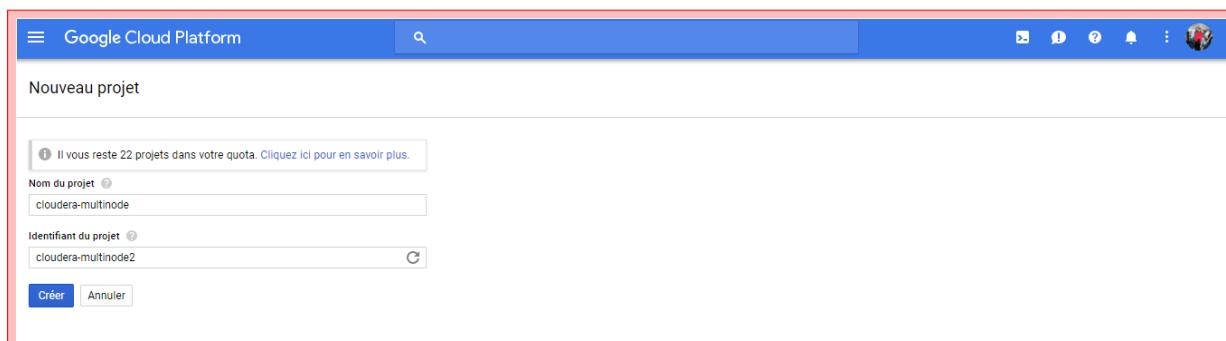
Cloudera

Cloudera est le leader, ce qui lui donne une l gitimit  avec un nombre de clients sup rieur   celui de ses concurrents. Le fait de disposer du cr ateur du framework Hadoop dans ces rangs est un grand avantage. La distribution Cloudera a  t  utilis e dans le rapport pour plusieurs raisons. Tout d'abord le fait que Cloudera propose une version open source qui utilise les principaux composants de Hadoop. Ensuite, la distribution de Cloudera est la plus mature sur le march  avec d j  la cinqui me version nomm e CDH5. Mais surtout, la distribution de Cloudera est la plus utilis e en entreprise. En effet, selon le livre blanc « O  en est l'adoption du Big Data ? » publi  par Talend en 2013, "12 %" des personnes ont

répondu qu'elles utilisaient déjà ou comptaient utiliser la distribution de Cloudera, contre "4 %" pour la distribution de MapR et "3 %" pour la distribution d'Hortonworks. Le reste des réponses concernant d'autres solutions. (Talend, 2013). Cloudera existe en trois versions : Free Edition, Standard et Enterprise. Nous avons décidé d'utiliser la version Enterprise afin d'explorer les fonctionnalités qu'elle offre vu que celle-ci est adaptée pour un contexte d'entreprise. Cloudera propose un outil pour superviser et automatiser le déploiement des clusters Hadoop nommé Cloudera Manager. C'est ce composant que nous avons utilisé pour installer le cluster Hadoop. Les fonctionnalités clés de Cloudera sont les suivantes :

- Gestion du cluster : elle permet de déployer, configurer et exploiter facilement des clusters de façon centralisée, avec une administration intuitive pour tous les services, les hôtes et les workflows.
- Monitoring du cluster : elle permet de maintenir une vue centralisée de toutes les activités de la grappe (noeuds du cluster), ses contrôles proactifs et des alertes.
- Diagnostique du cluster : cette fonctionnalité permet de diagnostiquer et résoudre facilement les problèmes avec l'aide des rapports opérationnels et des tableaux de bord, des événements, de l'affichage des journaux, des pistes d'audit.
- Intégration : cette fonctionnalité permet d'intégrer les outils de surveillance existants (SNMP, SMTP) avec Cloudera Manager.

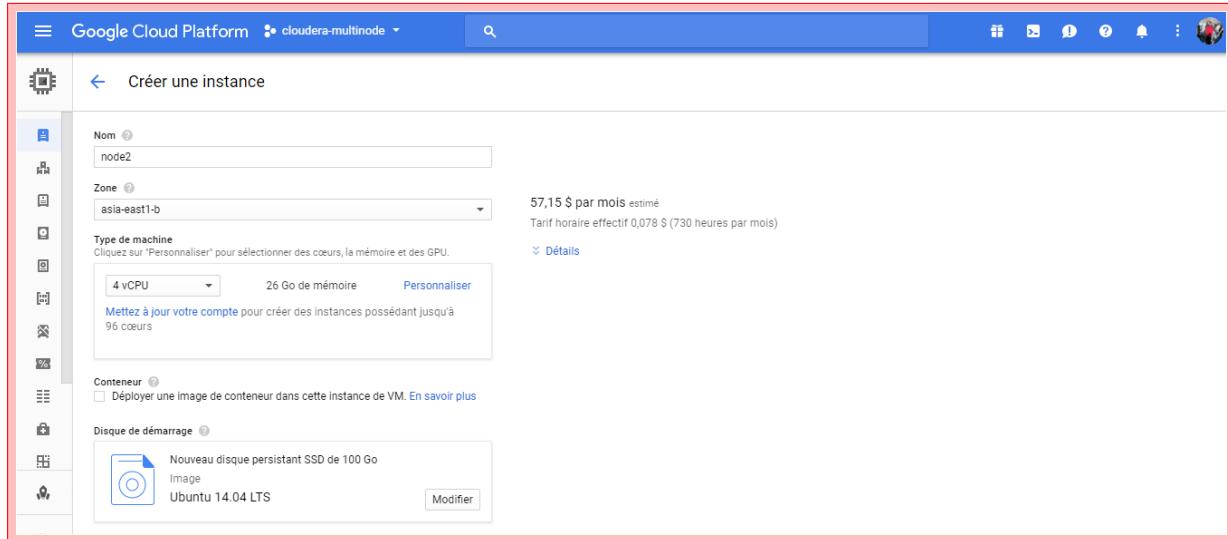
Nous allons voir comment créer Hadoop Cluster en temps réel à l'aide du gestionnaire Cloudera (CDH-5.6) sur Ubuntu -14.04 dans Google Cloud. Tout d'abord, on commence par l'attribution de nom à notre cluster. Ici, nous l'appelons "cloudera-multinode" avec l'identifiant "cloudera-multinode2" et nous cliquons sur "créer" pour commencer les configurations.



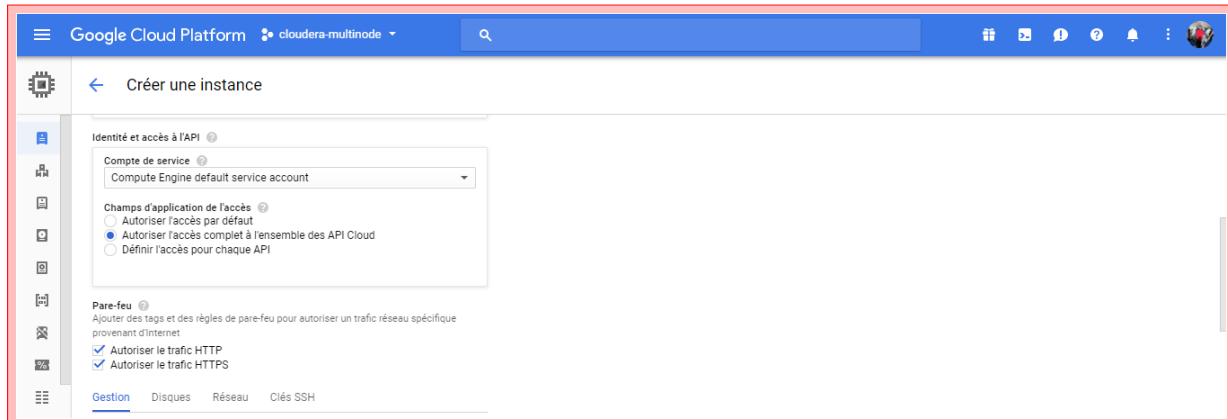
Nous rappelons que pendant la version d'essai on ne peut créer des machines virtuelles dans la même zone étant donné qu'on est limité par le quota et le nombre d'heures de fonctionnement. C'est pourquoi nous avons décidé d'attribuer à chaque noeud une zone géographique. Ce qui ne pose aucun problème puisque les serveurs Google connectent le monde ! Sachant que la configuration des noeuds est similaire, prenons le cas du second noeud :

- Nom : node2

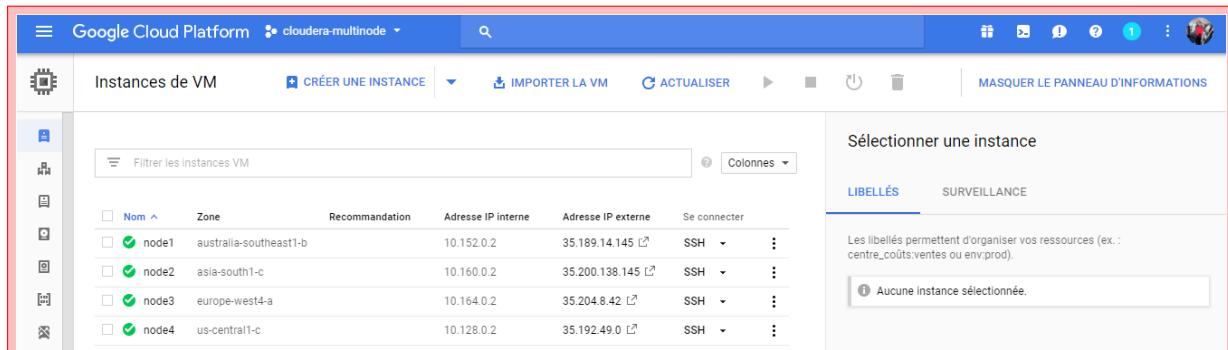
- Zone : Asia-east1-b
- Type de Machine : 4 vCPU et 26 Go de mémoire.
- Disque de démaragement : Ubuntu 14.04 LTS



On doit cependant autoriser l'accès à toutes les API CLOUD et autoriser le trafic Http et Https.



On se retrouve alors avec le cluster Multi-noeuds suivant :



Il sera question maintenant de créer une clé de service. Dans la rubrique "identifiants" nous cliquons sur "créer des identifiants" après avoir spécifié qu'il s'agit dans ce cas de "Clé de compte de service"

The screenshot shows the Google Cloud Platform API Identifiers page. In the center, there is a dropdown menu titled 'Créer des identifiants'. The options listed are:

- Clé API
- ID client OAuth
- Clé de compte de service
- Aidez-moi à choisir

At the bottom of the dropdown menu is a blue button labeled 'Créer des identifiants'.

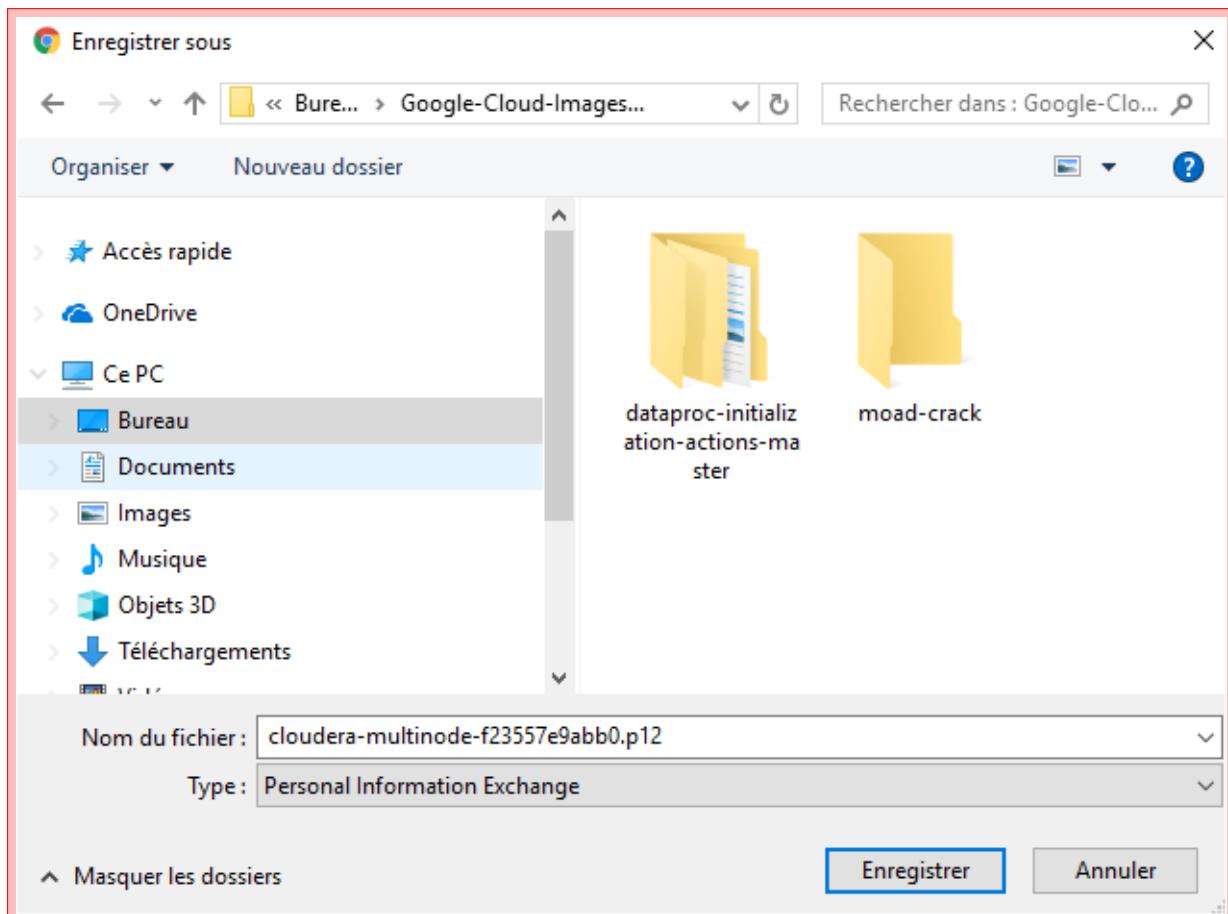
En cliquant sur le nouveau service, nous remplirons la case du nom et nous avons précisé P12 comme type de clé pour une raison de rétrocompatibilité avec le code au format P12.

The screenshot shows the 'Créer une clé de compte de service' dialog box. The fields filled in are:

- Compte de service: Nouveau compte de service
- Nom de compte de service: Moad
- Rôle: Rôle
- ID de compte de service: moad-387 @cloudera-multinode2.iam.gserviceaccount.com
- Type de clé: P12 (selected)

At the bottom of the dialog box are two buttons: 'Créer' and 'Annuler'.

Nous téléchargeons la clé Publique/Privée et nous gardons le mot de passe.



Voilà comment nous avons configuré les paramètres d'authentification.

The screenshot shows the 'Identifiers' section of the Google Cloud Platform API page. It displays a single service account entry:

ID	Date de création	Compte de service
f23557e9abb0439ec4b0f846aad5536ad760ab67	1 avr. 2018	Moad

Nous nous connectons en SSH pour apporter les premiers changements sur le système.

Remarque 1 : Grâce à Secure Shell (SSH) nous pouvons établir une connexion sécurisée entre notre ordinateur et le serveur sur lequel se trouve notre site web. Les web-masters peuvent ainsi accéder par Shell au serveur web et y exécuter des commandes. Le transfert des données est chiffré.

```
moadhaniii@node1: ~ - Google Chrome
Secure | https://ssh.cloud.google.com/projects/cloudera-multinode2/zones/australia-southeast1-b/instances/node1?authuser=0&hl=fr&p...
Connected, host fingerprint: ssh-rsa 2048 4C:F3:21:FB:F4:FC:36:CB:1C:81:C4:10:17:18:FE:45:BE:35:6B:CC
Welcome to Ubuntu 14.04.5 LTS (GNU/Linux 4.4.0-116-generic x86_64)

 * Documentation:  https://help.ubuntu.com/
System information as of Sun Apr 1 18:03:46 UTC 2018

System load: 0.0          Processes:           103
Usage of /: 1.2% of 98.40GB   Users logged in:  0
Memory usage: 0%            IP address for eth0: 10.152.0.2
Swap usage: 0%

Graph this data and manage this system at:
https://landscape.canonical.com/

Get cloud support with Ubuntu Advantage Cloud Guest:
http://www.ubuntu.com/business/services/cloud

0 packages can be updated.
0 updates are security updates.

New release '16.04.4 LTS' available.
Run 'do-release-upgrade' to upgrade to it.

Your Hardware Enablement Stack (HWE) is supported until April 2019.

Last login: Sun Apr 1 18:03:46 2018 from 35.205.132.171
moadhaniii@node1:~$
```

Nous commençons par modifier le mot de passe d'entrée.

```
Last login: Sun Apr 1 18:03:46 2018 from 35.205.132.171
moadhaniii@node1:~$ sudo -i
root@node1:~# passwd
Enter new UNIX password:
Retype new UNIX password:
passwd: password updated successfully
root@node1:~#
```

Nous essayons de nous connecter comme "sudo" via la commande "su" en Linux.

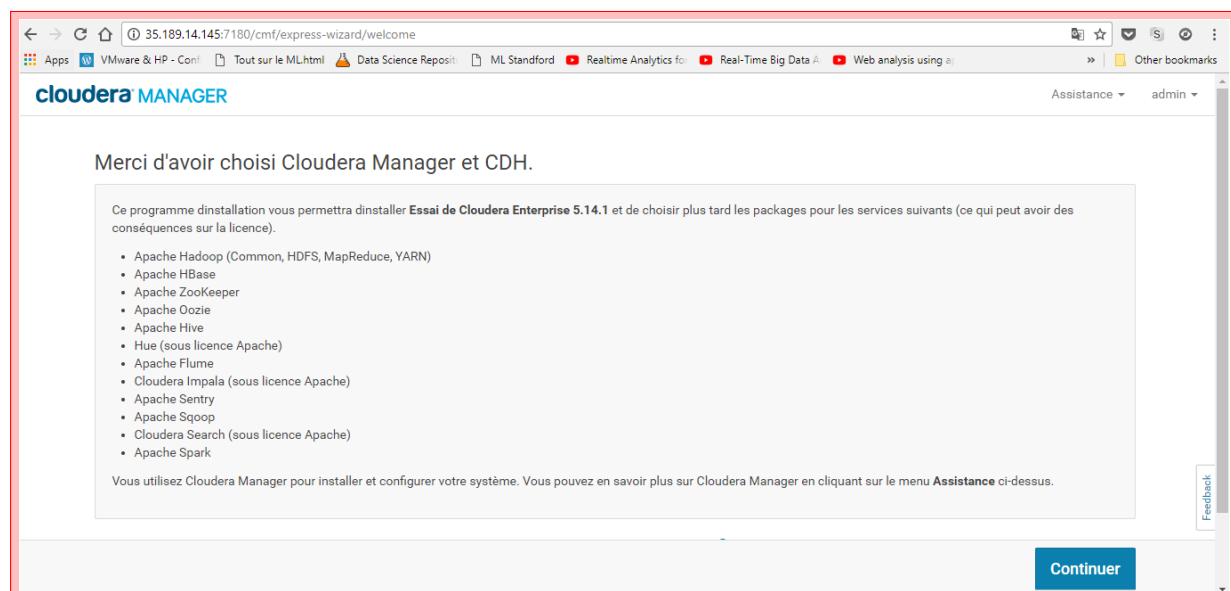
Remarque 2 : La commande "su" (Switch User) permet d'ouvrir une session avec l'ID (IDentifiant) d'un autre utilisateur, ou de démarrer un nouveau shell de connexion.

```
Last login: Sun Apr 1 18:27:45 2018 from 173.194.92.36
moadhaniii@node1:~$ su
Password:
root@node1:/home/moadhaniii# nano /etc/ssh/sshd_config
```

Cloudera Manager est une application de gestion pour les clusters Hadoop. Il permet de déployer des services basés sur Hadoop via une interface web graphique. Il a été intégré à CloudMan en tant que méthode efficace de déploiement et de gestion des services Hadoop. Dans ce contexte, CloudMan est utilisé pour provisionner les ressources de cloud nécessaires (par exemple, les instances, les disques) tandis que Cloudera Manager est utilisé pour déployer et gérer les services Hadoop.

Lors de l'exécution de Cloudera Manager, il est nécessaire d'utiliser un type d'instance avec au moins 16 Go de RAM. Une fois qu'une instance est lancée, nous démarrons l'application Cloudera Manager depuis la page d'administration CloudMan en cliquant sur le bouton Démarrer en regard du nom du service. Pour nous connecter, nous utilisons les mêmes informations d'identification que celles utilisées pour l'authentification avec CloudMan.

Pour le moment, il est nécessaire de créer et de configurer manuellement un cluster Hadoop en suivant l'assistant dans l'application Cloudera Manager après la connexion. Nous nous connectons avec le mot de passe que nous avons spécifié sur le formulaire Cloud Launch lorsque nous avons démarré le cluster.



Nous spécifions les hôtes pour notre installation du cluster Cloudera.

Spécifiez les hôtes pour votre installation de cluster CDH.

Les hôtes doivent être spécifiés à l'aide du même nom d'hôte (FQDN) qui leur permet de s'identifier.

Cloudera recommande d'inclure l'hôte de Cloudera Manager Server. Cela permet également d'activer la surveillance de l'état d'intégrité pour cet hôte.

Astuce: Recherchez les noms dhôte et les adresses IP à l'aide de modèles.

4 hôtes analysés, 4 exécutant SSH.

Requête étendue	Nom d'hôte (FQDN)	Adresse IP	Actuellement géré	Résultat
<input checked="" type="checkbox"/>	node1	node1.c.cloudera-multinode2.internal	10.152.0.2	Non ✓ Hôte prêt : temps de réponse de 0 ms.
<input checked="" type="checkbox"/>	node2	node2.c.cloudera-multinode2.internal	10.160.0.2	Non ✓ Hôte prêt : temps de réponse de 228 ms.
<input checked="" type="checkbox"/>	node3	node3.c.cloudera-multinode2.internal	10.164.0.2	Non ✓ Hôte prêt : temps de réponse de 280 ms.
<input checked="" type="checkbox"/>	node4	node4.c.cloudera-multinode2.internal	10.128.0.2	Non ✓ Hôte prêt : temps de réponse de 172 ms.

Précédent Continuer

Les installations des quatre noeuds dans le cluster.

Installation de cluster

Install Agents

Installation terminée avec succès.

4 hôte(s) sur 4 terminé(s) avec succès.

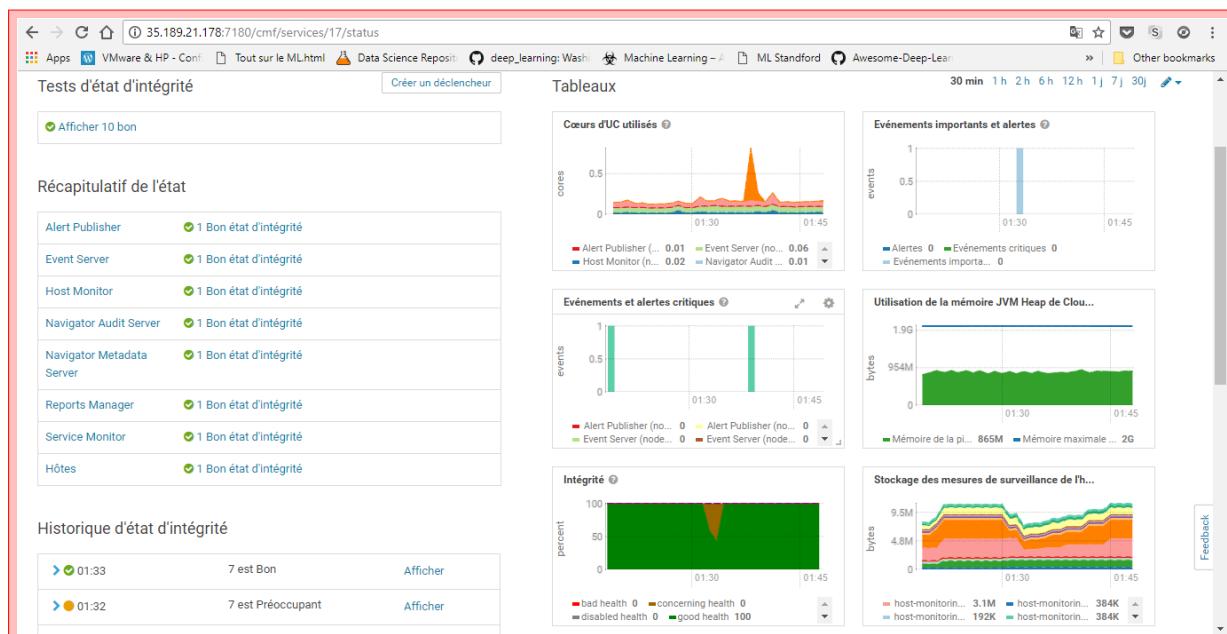
Nom d'hôte	Adresse IP	Progression	Etat	Détails
node1.c.cloudera-multinode2.internal	10.152.0.2	<div style="width: 100%; background-color: #2e7131;"></div>	✓ Installation terminée avec succès.	Détails
node2.c.cloudera-multinode2.internal	10.160.0.2	<div style="width: 100%; background-color: #2e7131;"></div>	✓ Installation terminée avec succès.	Détails
node3.c.cloudera-multinode2.internal	10.164.0.2	<div style="width: 100%; background-color: #2e7131;"></div>	✓ Installation terminée avec succès.	Détails
node4.c.cloudera-multinode2.internal	10.128.0.2	<div style="width: 100%; background-color: #2e7131;"></div>	✓ Installation terminée avec succès.	Détails

Précédent 1 2 3 4 5 6 7 Continuer

Dernière configuration du cluster.

The screenshot shows the 'Configuration du cluster' (Cluster Configuration) screen in Cloudera Manager. It is specifically focused on 'Configuration de la base de données' (Database Configuration). The interface includes a note about configuring and testing database connections, with options for using personalized databases or embedded databases. For Hive, a PostgreSQL database is selected with 'hive' as the name, user, and password. For Hue, another PostgreSQL database is selected with 'hue' as the name, user, and password. A navigation bar at the bottom shows steps 1 through 6, with step 2 highlighted. Buttons for 'Précédent' (Previous) and 'Continuer' (Next) are visible.

Nous pouvons alors visualiser les tableaux récapitulatifs qui illustrent l'état du cluster.



Résultat final du cluster Cloudera multi-noeuds doté de 11 services.

The screenshot shows the Cloudera Manager interface for Cluster 1 (CDH 5.14.0, Parcels). The left sidebar lists various services: Hôtes 4, HBase, HDFS, Hive, Impala, Key-Value St..., Oozie, Solr, Spark, YARN (MR2 I...), and ZooKeeper. The main area displays four performance charts:

- UC du cluster**: Utilisation d'UC de l'hôte dans Hôtes. Current value: 3.8%.
- E/S disque du cluster**: bytes / second. Current values: Total Octets de disque... 0, Total Octets de... 164K/s.
- E/S réseau du cluster**: bytes / second. Current values: Total Octets re... 10.5K/s, Total Octets tr... 10.8K/s.
- E/S HDFS**: bytes / second. Current values: Total Octets lus d... 1b/s, Total Octets écr... 2.8b/s.

At the bottom, there is a search bar and a taskbar with icons for various applications.

Hortonworks

Le principal concurrent de Cloudera dans ce comparatif Hadoop n'est autre que Hortonworks, second sur le marché en terme de présence. Ce vendeur compte dans le top 100 Red Herring. Ce pure player propose une distribution open source à 100 % de la plateforme de traitement de données. La firme cherche à proposer ses innovations à travers la plateforme open data et à bâtir un écosystème de partenaires pour accélérer le processus d'adoption d'Hadoop en entreprise. Grâce à ce choix, les clients bénéficient d'une flexibilité complète durant l'utilisation du logiciel s'ils souhaitent changer de distribution.

Selon Mike Gualtieri, principal analyste de Forrester, quand la communauté open source n'avance pas assez vite dans certains domaines, Hortonworks lance de nouveaux projets pour l'aider. Par exemple, Apache Ambari est une console de gestion de cluster développée par Hortonworks pour la provision, la gestion et la surveillance des clusters Hadoop. Malheureusement, cette approche open source se fait au détriment de certaines fonctionnalités.

Dans ce comparatif Hadoop, l'entreprise fait montrer de ses pouvoirs de séduction. Chaque trimestre, Hortonworks attire à peu près 60 nouveaux clients dont certains géants comme Samsung, Spotify, Bloomberg ou eBay. La firme est également partenaire de RedHat, Microsoft, SAP ou Teradata.

Points positifs :

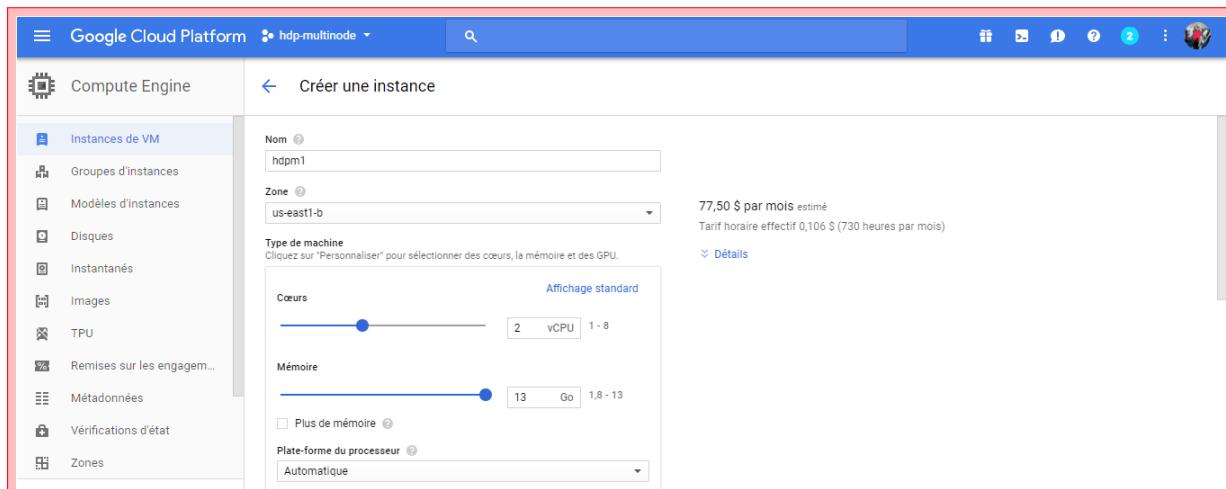
- La seule distribution du comparatif Hadoop à supporter Windows
- N'enferme pas ses utilisateurs dans un silo distributif
- Système de partenariats et de certifications

Points négatifs :

- Manque de certaines fonctionnalités
- Interface basique

D'abord nous créons 3 machines virtuelles dans des zones différentes de :

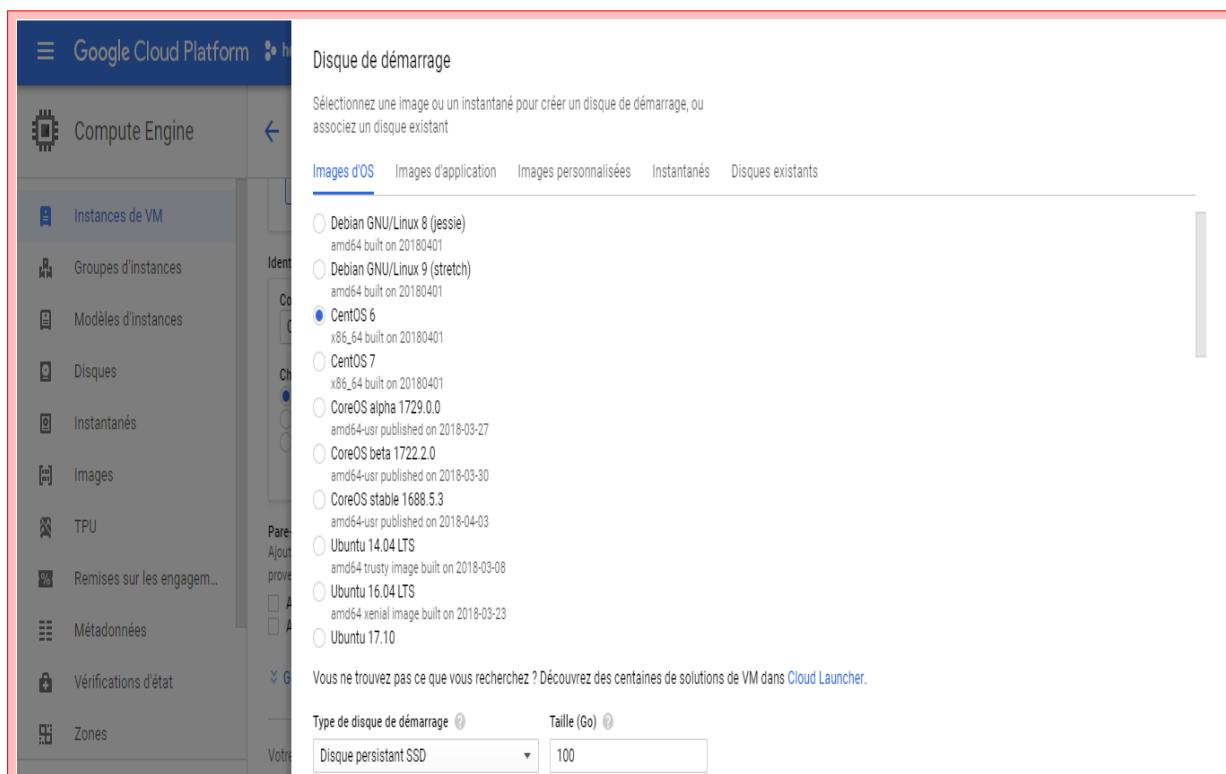
- 2 vCPU
- 13 Go de mémoire



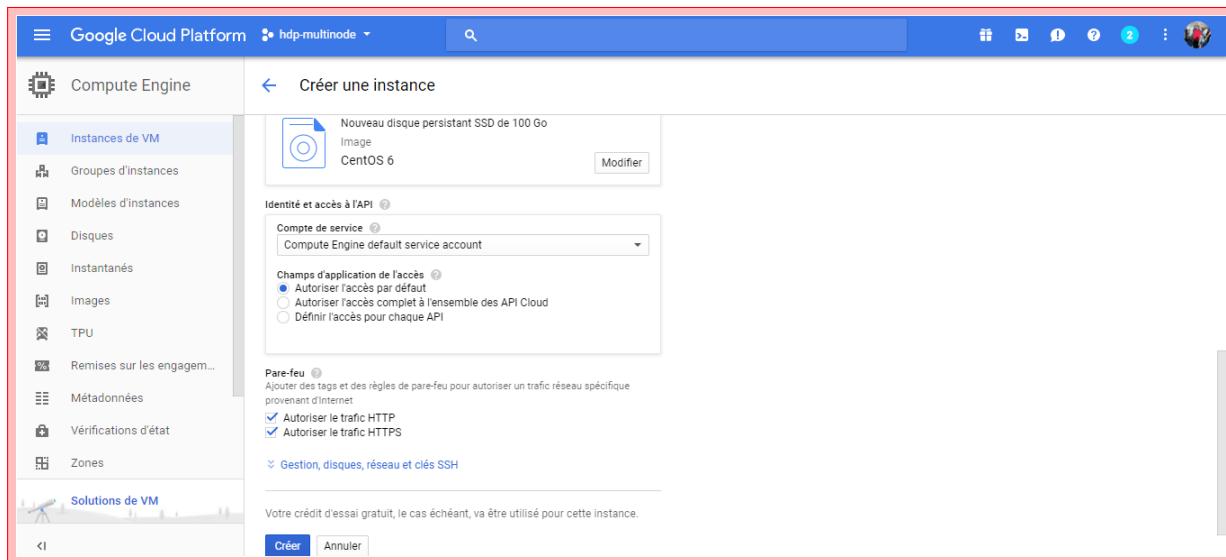
Nous avons opté pour CentOS car cette distribution est basé sur une solution commerciale de qualité supérieure d'autant plus qu'elle soit très stable et compatible, ce qui est très souhaitable dans les environnements professionnels.

Nous choisissons un disque persistant de 100 Go. Les disques persistants sont des périphériques de stockage réseau durables auxquels nos instances peuvent accéder comme des disques physiques sur un ordinateur de bureau ou un serveur. Les données sur chaque disque persistant sont réparties sur plusieurs disques physiques. Compute Engine gère les disques physiques et la distribution des données pour assurer la redondance et optimiser les performances pour nous. Les disques persistants standard sont protégés par des disques durs standard (HDD). Les disques persistants SSD sont protégés par des disques SSD.

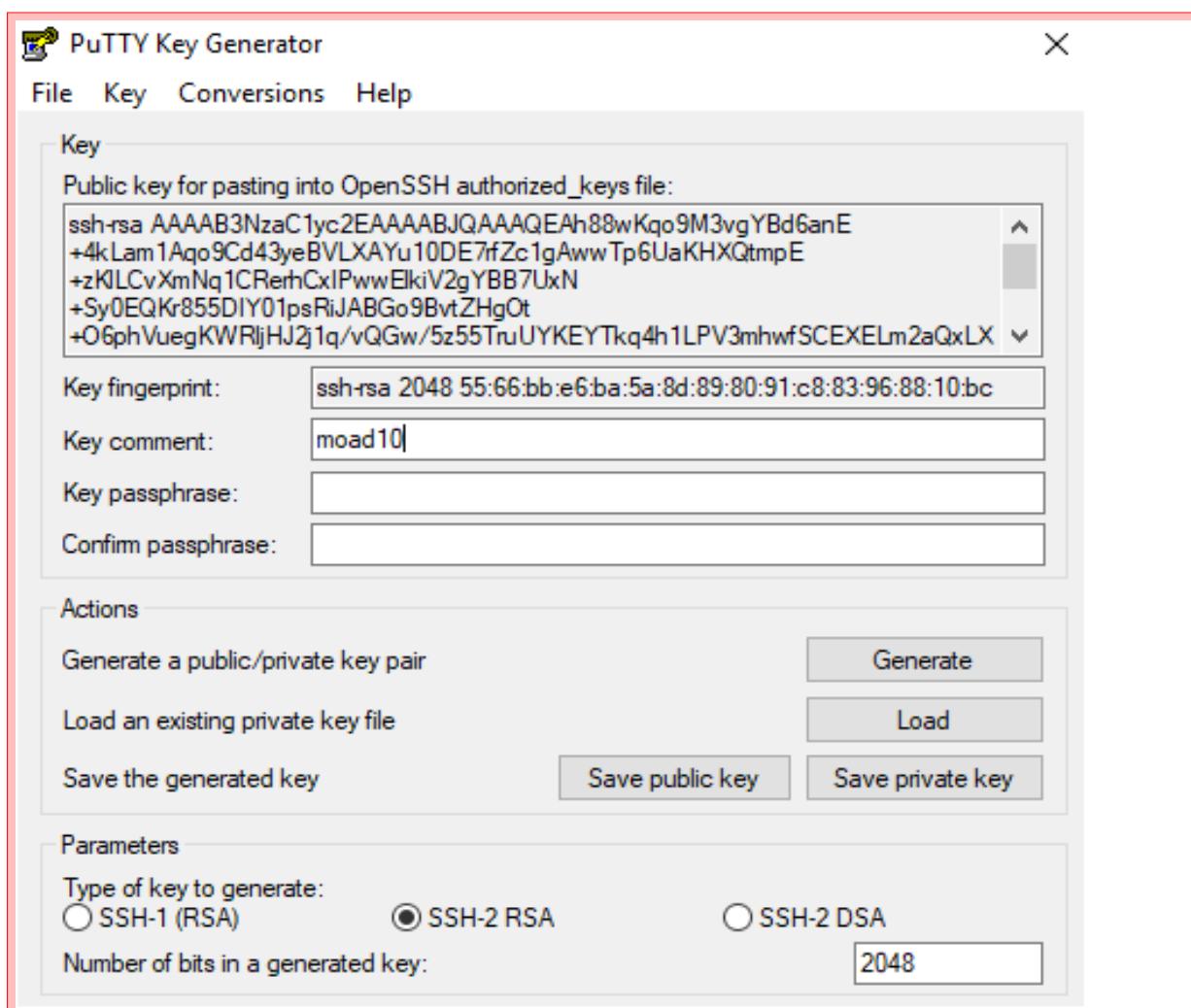
Les disques persistants sont situés indépendamment des instances de notre machine virtuelle. Nous pouvons donc détacher ou déplacer des disques persistants pour conserver nos données même après la suppression de nos instances. Les performances des disques persistants évoluent automatiquement en fonction de la taille, ce qui nous permet de redimensionner nos disques persistants existants ou d'ajouter des disques persistants à une instance pour répondre à nos besoins en termes de performances et d'espace de stockage.



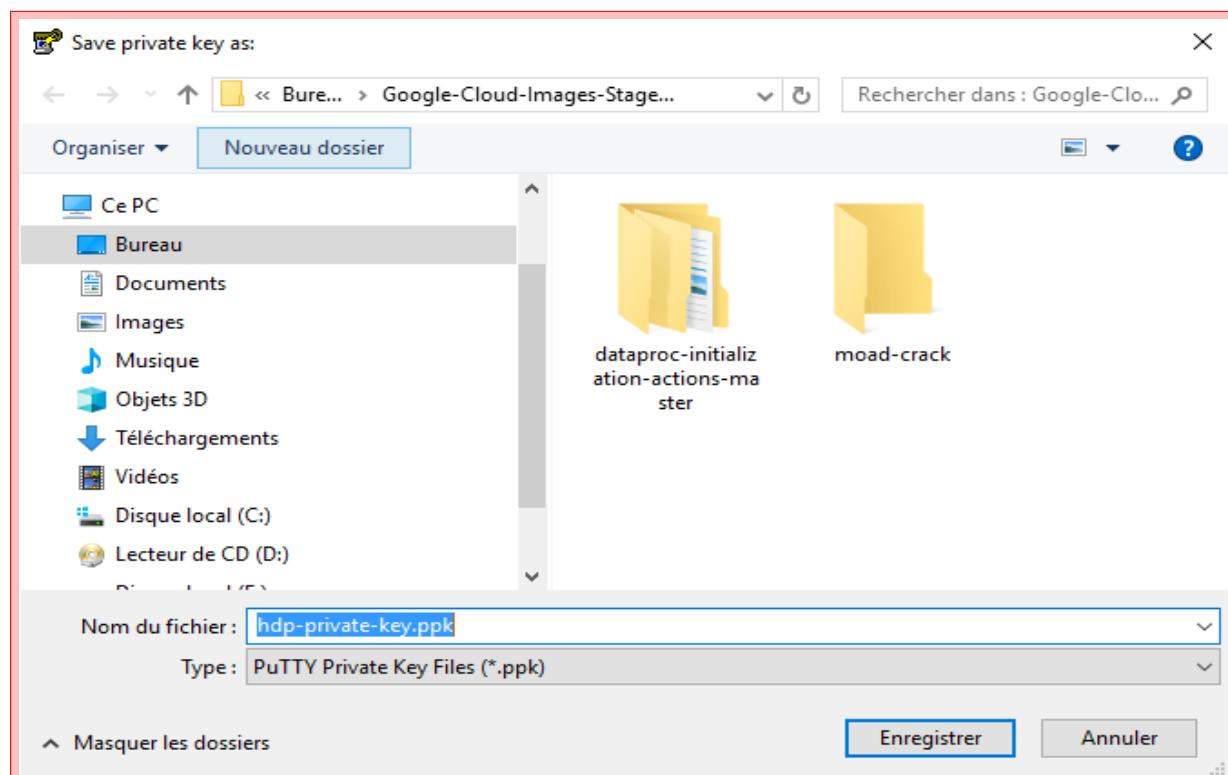
Nous autorisons l'accès internet via le navigateur WEB.



Après la création de nos noeuds (Machines virtuelles) nous voulons les connecter entre eux, c'est la raison pour laquelle nous allons utiliser Putty Key Generator.



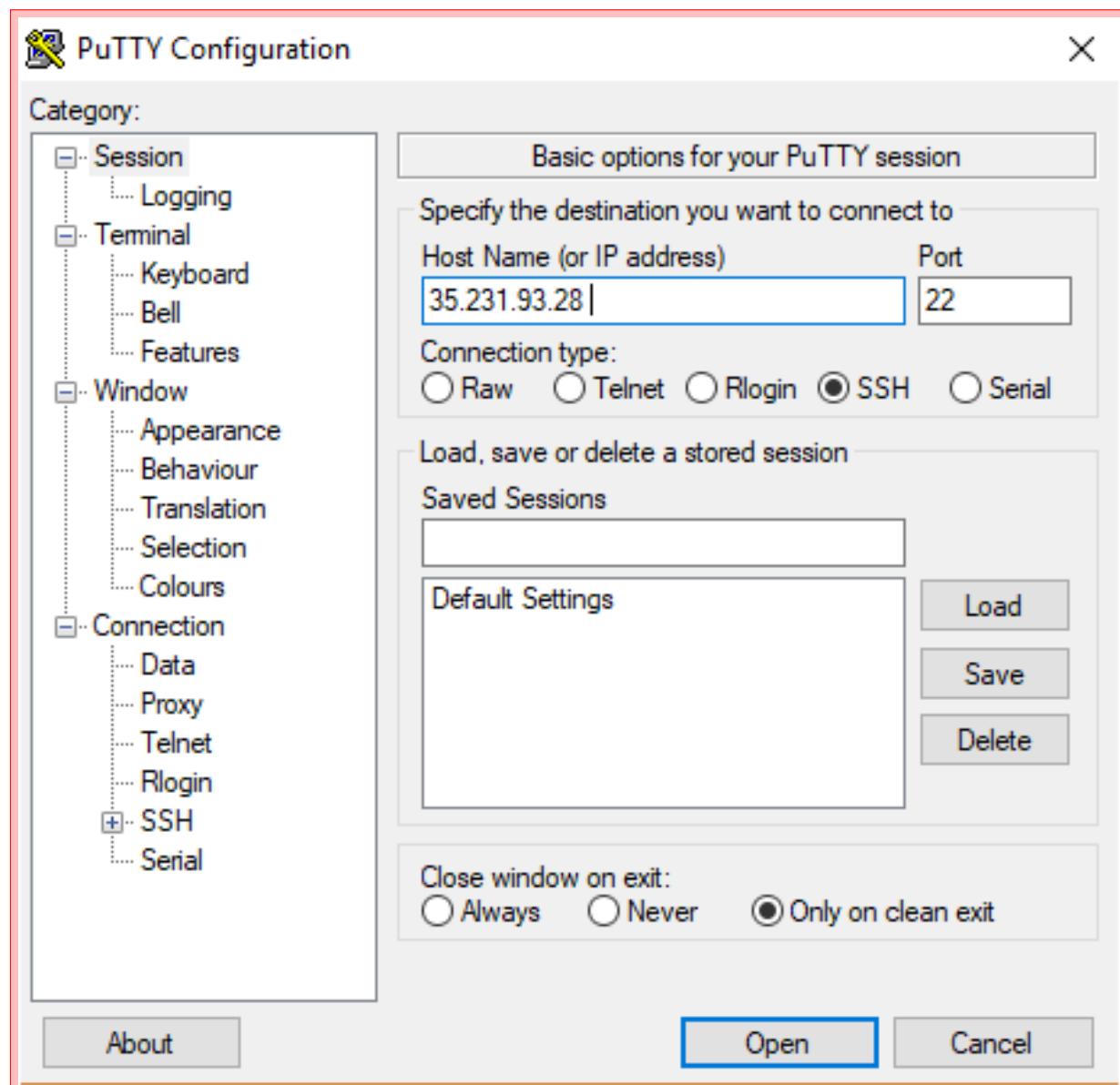
Après l'avoir créée, nous l'enregistrons dans un fichier.



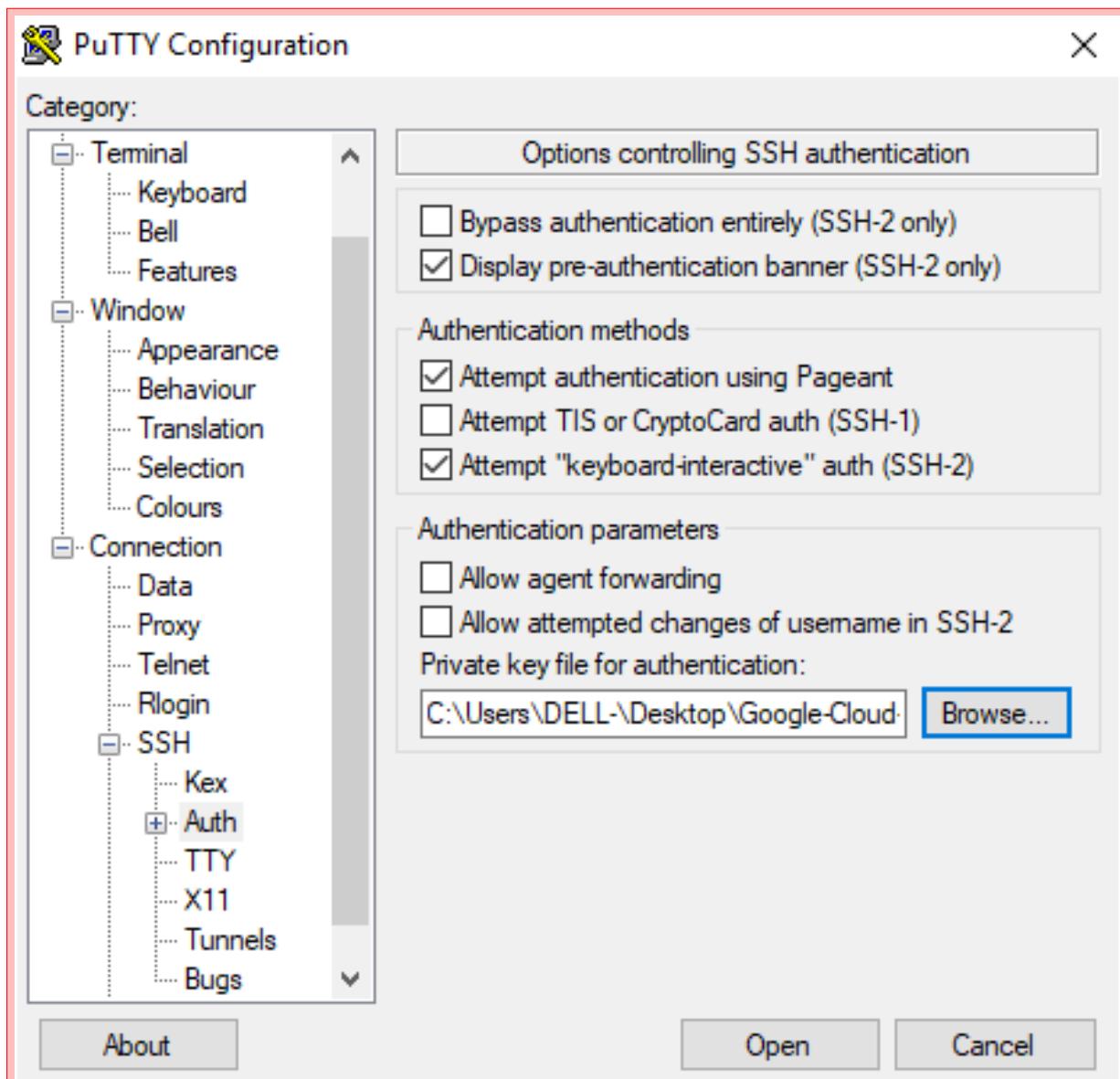
La clef SSH du serveur a été copié sur cette instance pour permettre une connexion sans demande de mot de passe.

The screenshot shows the 'Clés SSH' (SSH Keys) section of the Google Cloud Platform interface. At the top, there are tabs: 'Gestion', 'Disques', 'Réseau', and 'Clés SSH', with 'Clés SSH' being the active tab. Below the tabs, a message states: 'Ces clés n'autorisent l'accès qu'à cette instance, contrairement aux clés SSH à l'échelle du projet.' followed by a link 'En savoir plus'. There is a checkbox labeled 'Bloquer les clés SSH à l'échelle du projet' with a descriptive text below it: 'Si vous cochez cette case, vous bloquez les clés SSH au niveau du projet pour cette instance.' followed by another link 'En savoir plus'. On the left, the instance name 'moad10' is listed. In the center, a modal window displays a long string of characters representing the SSH key content: 'PwwE1kiV2gYBB7UxN+Sy0EQKr855DIY01psRiJABGo9BvtZHg0t+O6phVuegKWR1jHJ2j1q/vQGw/5z55TrUYKEYTkq4h1LPV3mhwfSCEXELm2aQxLXPfLfFe153/y7kZS8SaygdRs5z/MDn+clue5SxAYMTty9KnKXJXgvWcSkhQRGyPy+19+nyF6RT5qL5u2/UeXoqR7RKZDLU9fPm/e/+G5IMcVYPrFf6ykPVLUmH+sjPXVwA6i//iFQ== moad10'. At the bottom, there is a blue button with a plus sign and the text '+ Ajouter un élément'.

Il y a beaucoup de pages d'options, comme nous pouvons le voir au niveau de la section « Category » sur le côté. Pour le moment : nous avons juste besoin de remplir le champ en haut « Host Name (or IP address) ». Nous y entrons l'adresse IP de notre première machine virtuelle (notre Master node).

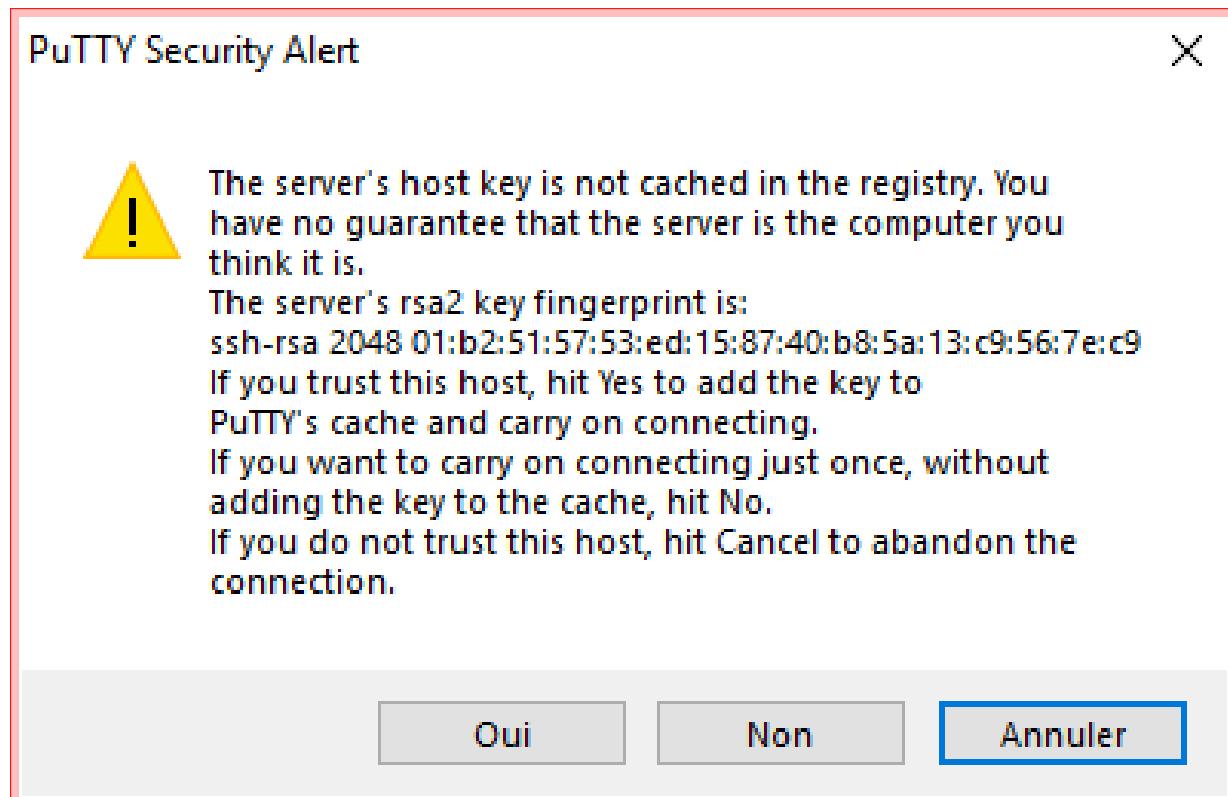


Maintenant, nous nous rendons dans Connection \Rightarrow SSH \Rightarrow Auth. Nous cliquons sur le petit bouton « Browse » pour sélectionner notre clé privée.



Que se passe-t-il? On nous dit que le fingerprint (empreinte) du serveur est 01 : b2 : 51 : 57 : 53 : ed : 15 : 87 : 40 : b8 : 5a : 13 : c9 : 56 : 7e : c9. C'est un numéro unique qui nous permet d'identifier le serveur. Si demain quelqu'un essaie de se faire passer pour le serveur, le fingerprint changera forcément et nous saurons qu'il se passe alors quelque chose d'anormal. En effet, SSH nous avertira de manière très claire si cela arrive.

Nous tapons « oui » pour confirmer que c'est bien le serveur auquel nous voulons nous connecter. Le serveur et le client vont alors s'échanger une clé de chiffrement.



Normalement, le serveur devrait nous demander au bout de quelques secondes notre mot de passe.

A screenshot of a terminal window. The title bar shows the user is logged in as "moad10@hdpm1:~". The window content shows the following text:

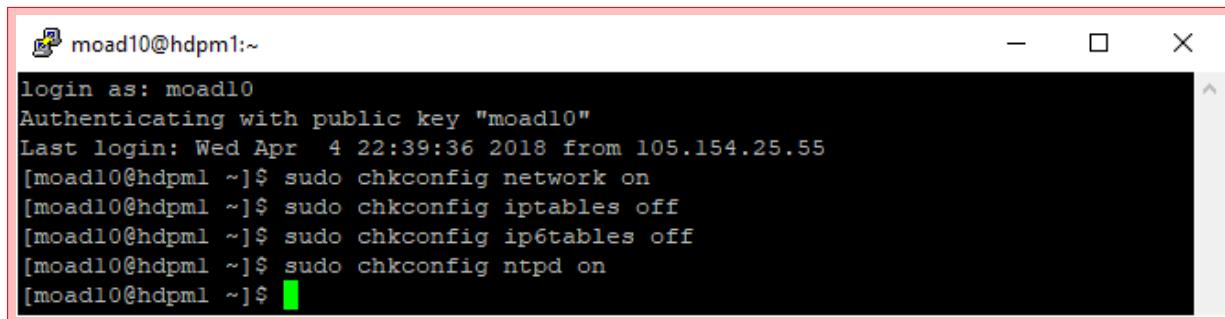
```
login as: moad10
Authenticating with public key "moad10"
[moad10@hdpm1 ~]$
```

Nous devons désactiver le pare-feu sous Linux à des fins de test. Nous utilisons CentOS. Comment désactiver le pare-feu sous Linux ?

Un pare-feu Linux est un pare-feu logiciel qui fournit une protection entre notre serveur (station de travail) et un contenu nuisible sur Internet ou sur le réseau. Il va essayer de protéger nos machines contre les utilisateurs malveillants et les logiciels tels que les virus / vers.

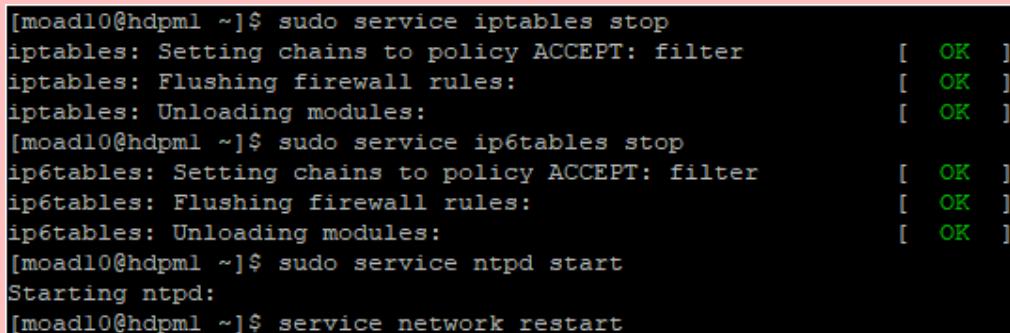
- iptables est un logiciel libre de l'espace utilisateur Linux grâce auquel l'administrateur système peut configurer les chaînes et règles dans le pare-feu en espace noyau
- Ip6tables est utilisé pour configurer, maintenir et inspecter les tables de règles de filtrage de paquets IPv6 dans le noyau Linux. Plusieurs tables différentes peuvent être définies. Chaque table contient un certain nombre de chaînes intégrées et peut également contenir des chaînes définies par l'utilisateur. Chaque chaîne est une liste de règles pouvant correspondre à un ensemble de paquets. Chaque règle spécifie que faire avec un paquet

correspondant. C'est ce qu'on appelle une «cible», qui peut être un saut à une chaîne définie par l'utilisateur dans la même table.



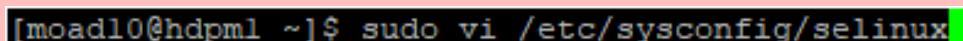
```
moad10@hdpm1:~$ login as: moad10
Authenticating with public key "moad10"
Last login: Wed Apr  4 22:39:36 2018 from 105.154.25.55
[moad10@hdpm1 ~]$ sudo chkconfig network on
[moad10@hdpm1 ~]$ sudo chkconfig iptables off
[moad10@hdpm1 ~]$ sudo chkconfig ip6tables off
[moad10@hdpm1 ~]$ sudo chkconfig ntpd on
[moad10@hdpm1 ~]$
```

Nous stoppons les services iptables et ip6tables puis nous démarrons le service ntpd. "ntpd" qui est un sigle de Network Time Protocol Daemon, est un daemon qui définit et maintient l'heure sur un système d'exploitation par synchronisation avec les serveurs dédiés à donner cette information.



```
[moad10@hdpm1 ~]$ sudo service iptables stop
iptables: Setting chains to policy ACCEPT: filter [ OK ]
iptables: Flushing firewall rules: [ OK ]
iptables: Unloading modules: [ OK ]
[moad10@hdpm1 ~]$ sudo service ip6tables stop
ip6tables: Setting chains to policy ACCEPT: filter [ OK ]
ip6tables: Flushing firewall rules: [ OK ]
ip6tables: Unloading modules: [ OK ]
[moad10@hdpm1 ~]$ sudo service ntpd start
Starting ntpd: [ OK ]
[moad10@hdpm1 ~]$ service network restart
```

SELinux (Security Enhanced Linux) est un système de contrôle d'accès obligatoire (Mandatory Access Control) qui s'appuie sur l'interface Linux Security Modules fournie par le noyau Linux. Concrètement, le noyau interroge SELinux avant chaque appel système pour savoir si le processus est autorisé à effectuer l'opération concernée. SELinux s'appuie sur un ensemble de règles (policy) pour autoriser ou interdire une opération. La commande "vi" permet d'éditer ce dernier.



```
[moad10@hdpm1 ~]$ sudo vi /etc/sysconfig/selinux
```

Maintenant nous allons voir le SWAPPINESS.

En fait le système est simple puisque on dit à Linux quand il doit commencer à écrire sur le disque dur dans l'espace d'échange appelé SWAP afin de délester la Ram. Il peut paraître évident qu'avec des temps d'accès en millisecondes pour les HD et en nanoseconde pour la mémoire vive, il est préférable d'écrire dans cette dernière.

Et on ne parle même pas des vitesses d'écriture qui sont immensément plus rapides que celles des HD

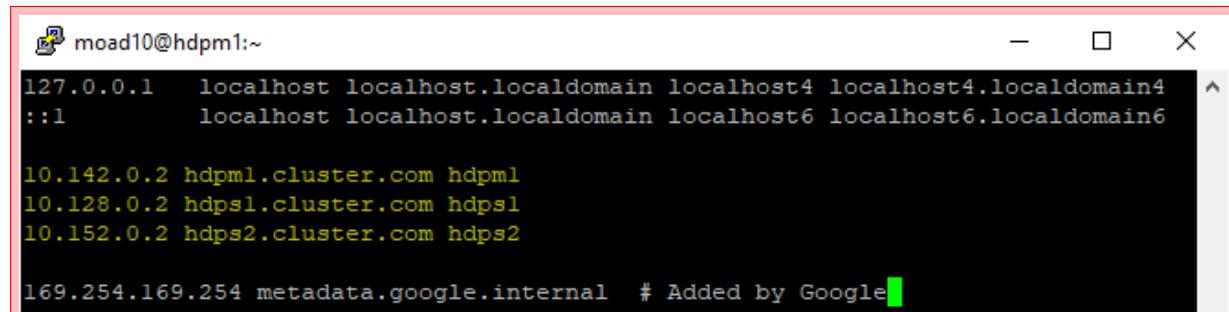
En clair :

- vm.swappiness = 0 Linux utilisera le HD en dernière limite pour éviter un manque de RAM.
- vm.swappiness = 60 Valeur par défaut de Linux : à partir de 40% d'occupation de Ram, le noyau écrit sur le disque.
- vm.swappiness = 100 Tous les accès se font en écriture dans la SWAP.

Pour s'informer des réglages de swappiness on utilise la commande : cat /proc/-sys/vm/swappiness Pour notre machine virtuelle (16G de Ram), nous estimons qu'un swappiness à 90 est suffisant puisque la swap sera utilisée lorsqu'il ne restera plus que 1,6G). A nouveau, il faut éditer le fichier suivant /etc/sysctl.conf en root et ajouter cette ligne : vm.swappiness = 10

```
[moad10@hdpm1 ~]$ sudo sysctl vm.swappiness=10
vm.swappiness = 10
```

Nous éditons le fichier "rc.local", puis le fichier "hosts". Nous modifions les 3 fichiers des "hosts" puisque nous disposons de 3 noeuds et à chaque fois nous ajoutons les 3 lignes en jaune pour que la machine se connaît et connaît les deux autres machines afin de créer une inter-connexion entre les noeuds.

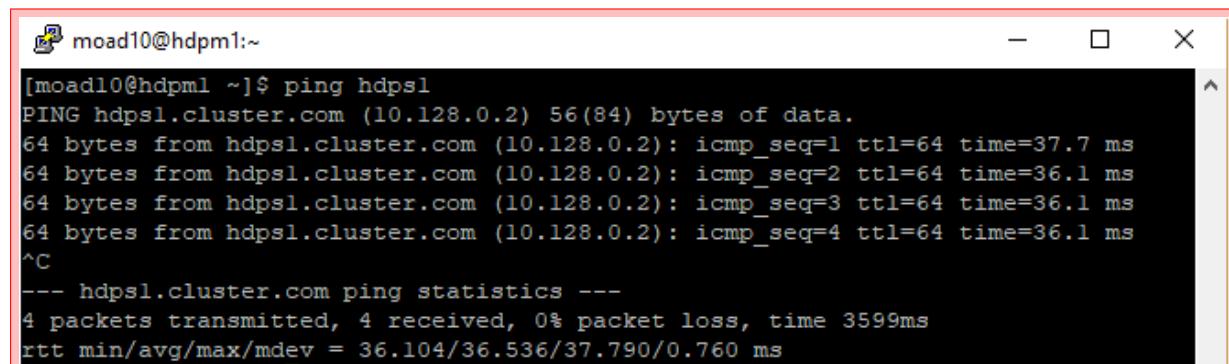


```
moad10@hdpm1:~
127.0.0.1 localhost localhost.localdomain localhost4 localhost4.localdomain4
::1 localhost localhost.localdomain localhost6 localhost6.localdomain6

10.142.0.2 hdpm1.cluster.com hdpm1
10.128.0.2 hdps1.cluster.com hdps1
10.152.0.2 hdps2.cluster.com hdps2

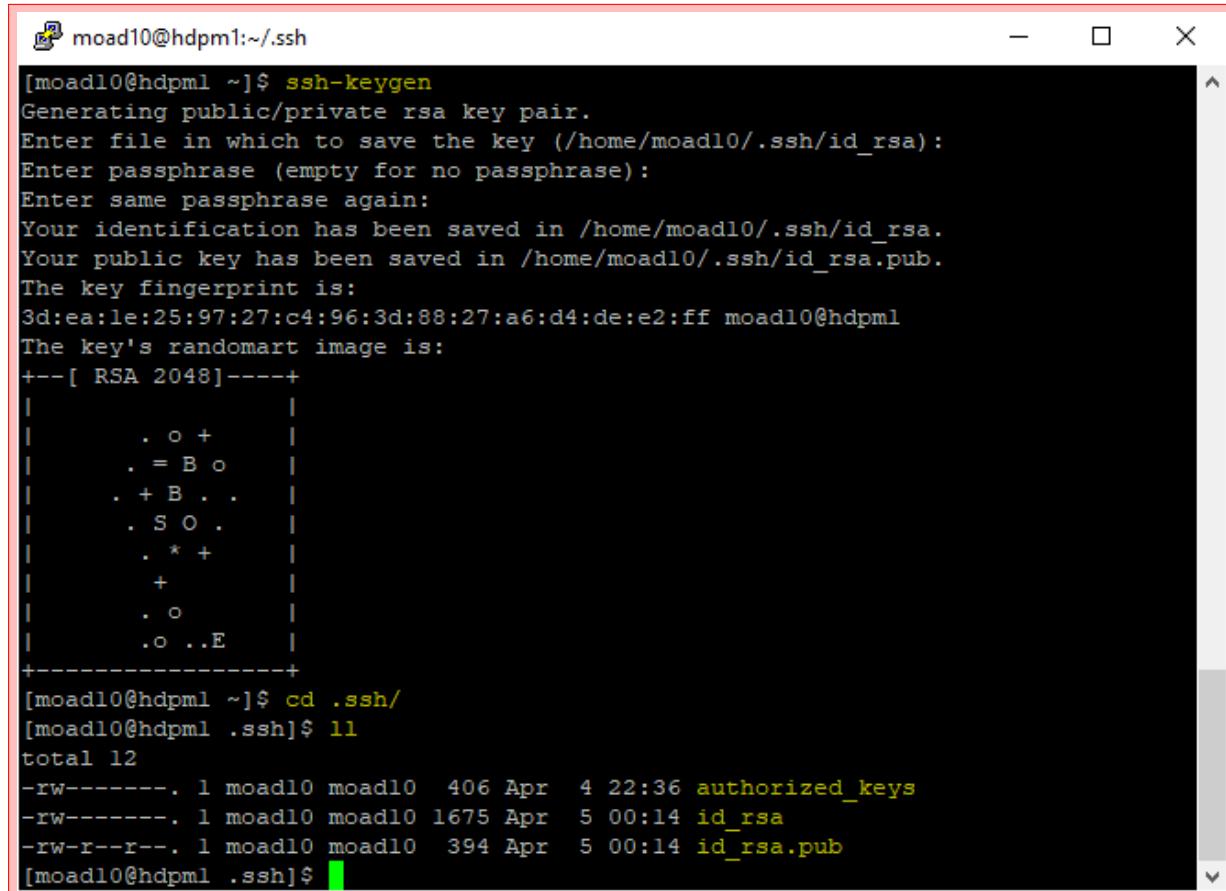
169.254.169.254 metadata.google.internal # Added by Google
```

Nous testons le ping et cela marche très bien.



```
[moad10@hdpm1 ~]$ ping hdps1
PING hdps1.cluster.com (10.128.0.2) 56(84) bytes of data.
64 bytes from hdps1.cluster.com (10.128.0.2): icmp_seq=1 ttl=64 time=37.7 ms
64 bytes from hdps1.cluster.com (10.128.0.2): icmp_seq=2 ttl=64 time=36.1 ms
64 bytes from hdps1.cluster.com (10.128.0.2): icmp_seq=3 ttl=64 time=36.1 ms
64 bytes from hdps1.cluster.com (10.128.0.2): icmp_seq=4 ttl=64 time=36.1 ms
^C
--- hdps1.cluster.com ping statistics ---
4 packets transmitted, 4 received, 0% packet loss, time 3599ms
rtt min/avg/max/mdev = 36.104/36.536/37.790/0.760 ms
```

Nous voulons utiliser Linux et OpenSSH pour automatiser nos tâches. Par conséquent, nous avons besoin d'une connexion automatique d'un hôte A à un hôte B. Nous ne voulons plus entrer de mot de passe, car nous voulons appeler ssh depuis un script shell.



```
moadl0@hdpml:~/ssh
[moadl0@hdpml ~]$ ssh-keygen
Generating public/private rsa key pair.
Enter file in which to save the key (/home/moadl0/.ssh/id_rsa):
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/moadl0/.ssh/id_rsa.
Your public key has been saved in /home/moadl0/.ssh/id_rsa.pub.
The key fingerprint is:
3d:ea:le:25:97:27:c4:96:3d:88:27:a6:d4:de:e2:ff moadl0@hdpml
The key's randomart image is:
+--[ RSA 2048]--+
| . o +
| . = B o
| . + B . .
| . S O .
| . * +
| +
| . o
| .o ..E
+-----+
[moadl0@hdpml ~]$ cd .ssh/
[moadl0@hdpml .ssh]$ ll
total 12
-rw-----. 1 moadl0 moadl0 406 Apr  4 22:36 authorized_keys
-rw-----. 1 moadl0 moadl0 1675 Apr  5 00:14 id_rsa
-rw-r--r--. 1 moadl0 moadl0 394 Apr  5 00:14 id_rsa.pub
[moadl0@hdpml .ssh]$
```

Dans un premier temps, le client génère une paire de clés (« Generating public/-private rsa key pair »). Il doit ensuite sauvegarder ces clés dans des fichiers (un pour la clé publique, un pour la clé privée). On nous propose une valeur par défaut.

Ensuite, on nous demande une passphrase. C'est une phrase de passe qui va servir à chiffrer la clé privée pour une meilleure sécurité. Là, nous avons deux choix :

- soit de taper Entrée directement sans rien écrire, et la clé ne sera pas chiffrée sur notre machine,
- soit de taper un mot de passe de notre choix, et la clé sera chiffrée.

Tout le monde ne met pas une phrase de passe. En fait, cela dépend du risque que quelqu'un d'autre utilise la machine du client et puisse lire le fichier contenant la clé privée qui doit rester confidentielle. Si nous avons notre propre machine virtuelle et que personne d'autre ne l'utilise, il y a assez peu de risques (à moins d'avoir un virus, un spyware). Si c'est en revanche une machine public, nous

recommandons vivement de mettre une passphrase pour chiffrer la clé qui sera enregistrée.

Si l'on hésite entre les deux méthodes, nous recommandons de rentrer une passphrase : c'est quand même la méthode la plus sûre.

Il faut maintenant envoyer au serveur notre clé publique pour qu'il puisse nous chiffrer des messages.

Notre clé publique devrait se trouver dans *e/.ssh/id_rsa.pub* (pub comme public). correspond à notre home (*/home/moad10/* dans notre cas). Notons que .ssh est un dossier caché. Notre clé privée se trouve dans *e/.ssh/id_rsa*. Elle est normalement chiffrée puisque nous avons entré une passphrase, ce qui constitue une sécurité de plus.

Nous pouvons déjà nous rendre dans le dossier .ssh, pour commencer. Les trois fichiers sont :

id_rsa : notre clé privée, qui doit rester secrète. Elle est chiffrée puisque nous avons rentré une passphrase;

id_rsa.pub : la clé publique que nous pouvons partager, et que nous devons envoyer au serveur ;

L'opération consiste à envoyer la clé publique (*id_rsa.pub*) au serveur et à l'ajouter à son fichier *authorized_keys* (clés autorisées). Le serveur y garde une liste des clés qu'il autorise à se connecter.

```
[moad10@hdpm1 .ssh]$ cat id_rsa.pub
ssh-rsa AAAAB3NzaC1yc2EAAAABIwAAQEA0bh5Z2ORUFH+SN/h5kXDky47GMaqx1V+k8Vlilu4YIJ+
My/jLtUVAoRrNQEEVkbBZEfAFM8XOsRbN5nq0+AoWcAe+w2b/FwIwLCXZ4F5T60vKSImOgZ+OqYzG1w0
C/DwX1AXMVzEjnR/5carQ1OiFn8X4Ir0p4fYdHwdgu/ouwDeW2X5rts6rcApr0/OKuSxo2vtsd/1A+84
jHDRghogH1IQkIB9AF9Up4Ijab71cz042Ztk3Qr9ceWojmP7wkX4iVUYDngUOCtB1IcTGmjvL1XmXW1C
2g2kLyrWZ5Sv+bpCh9AppmZUz7kXCDeQohAZcN56snvfOVvSxjmJAefn9Q== moad10@hdpm1
[moad10@hdpm1 .ssh]$ ssh-rsa AAAAB3NzaC1yc2EAAAABIwAAQEA0bh5Z2ORUFH+SN/h5kXDky4
7GMaqx1V+k8Vlilu4YIJ+My/jLtUVAoRrNQEEVkbBZEfAFM8XOsRbN5nq0+AoWcAe+w2b/FwIwLCXZ4F
5T60vKSImOgZ+OqYzG1w0C/DwX1AXMVzEjnR/5carQ1OiFn8X4Ir0p4fYdHwdgu/ouwDeW2X5rts6rcA
pr0/OKuSxo2vtsd/1A+84jHDRghogH1IQkIB9AF9Up4Ijab71cz042Ztk3Qr9ceWojmP7wkX4iVUYDng
UOCtB1IcTGmjvL1XmXW1C2g2kLyrWZ5Sv+bpCh9AppmZUz7kXCDeQohAZcN56snvfOVvSxjmJAefn9Q=
= moad10@hdpm1
```

Les systèmes de gestion de configuration sont conçus pour faciliter le contrôle d'un grand nombre de serveurs pour les administrateurs et les équipes d'exploitation. Ils nous permettent de contrôler de nombreux systèmes différents d'une manière automatisée à partir d'un emplacement central. Bien qu'il existe de nombreux systèmes de gestion de configuration populaires disponibles pour les systèmes Linux, tels que Chef et Puppet, ils sont souvent plus complexes que ce que beaucoup de gens veulent ou ont besoin. Ansible est une excellente alternative à ces options, car il a un coût beaucoup plus petit.

Ansible fonctionne en configurant des machines client à partir d'un ordinateur avec des composants Ansible installés et configurés. Il communique sur les canaux SSH normaux afin de récupérer des informations à partir de machines distantes, d'émettre des commandes et de copier des fichiers. Pour cette raison, un

système Ansible ne nécessite aucun logiciel supplémentaire pour être installé sur les ordinateurs clients. C'est une façon qu'Ansible simplifie l'administration des serveurs. Tout serveur ayant un port SSH exposé peut être placé sous le parapluie de configuration d'Ansible, quelle que soit l'étape de son cycle de vie.

Ansible adopte une approche modulaire, ce qui facilite l'extension pour utiliser les fonctionnalités du système principal afin de traiter des scénarios spécifiques. Les modules peuvent être écrits dans n'importe quelle langue et communiquer en langage JSON standard. Les fichiers de configuration sont principalement écrits dans le format de sérialisation des données YAML en raison de leur nature expressive et de leur similarité avec les langages de balisage populaires. Ansible peut interagir avec les clients via les outils de ligne de commande ou via ses scripts de configuration appelés Playbooks.

Dans ce guide, nous allons installer Ansible sur le noeud maître et apprendre quelques notions de base sur l'utilisation du logiciel.

Pour commencer à explorer Ansible, comme un moyen de gérer nos différents serveurs, nous devons installer le logiciel Ansible sur au moins une machine. (Ici, le noeud maître).

```
[moad10@hdpm1 .ssh]$ sudo yum install ansible
```

Nous changeons de répertoire, pour nous situer dans le dossier ansible. Ensuite, la commande ll=ls+l sous entend l'utilisation du caractère -l (petite lettre L) qui affichera une longue liste du contenu du répertoire courant. Sur tous les exemples, nous combinons le paramètre -l (principalement) pour obtenir un meilleur résultat.

```
moad10@hdpm1:/etc/ansible
[moad10@hdpm1 .ssh]$ cd /etc/ansible
[moad10@hdpm1 ansible]$ ll
total 28
-rw-r--r--. 1 root root 19155 Nov 29 22:39 ansible.cfg
-rw-r--r--. 1 root root 1016 Nov 29 22:39 hosts
drwxr-xr-x. 2 root root 4096 Nov 29 22:39 roles
```

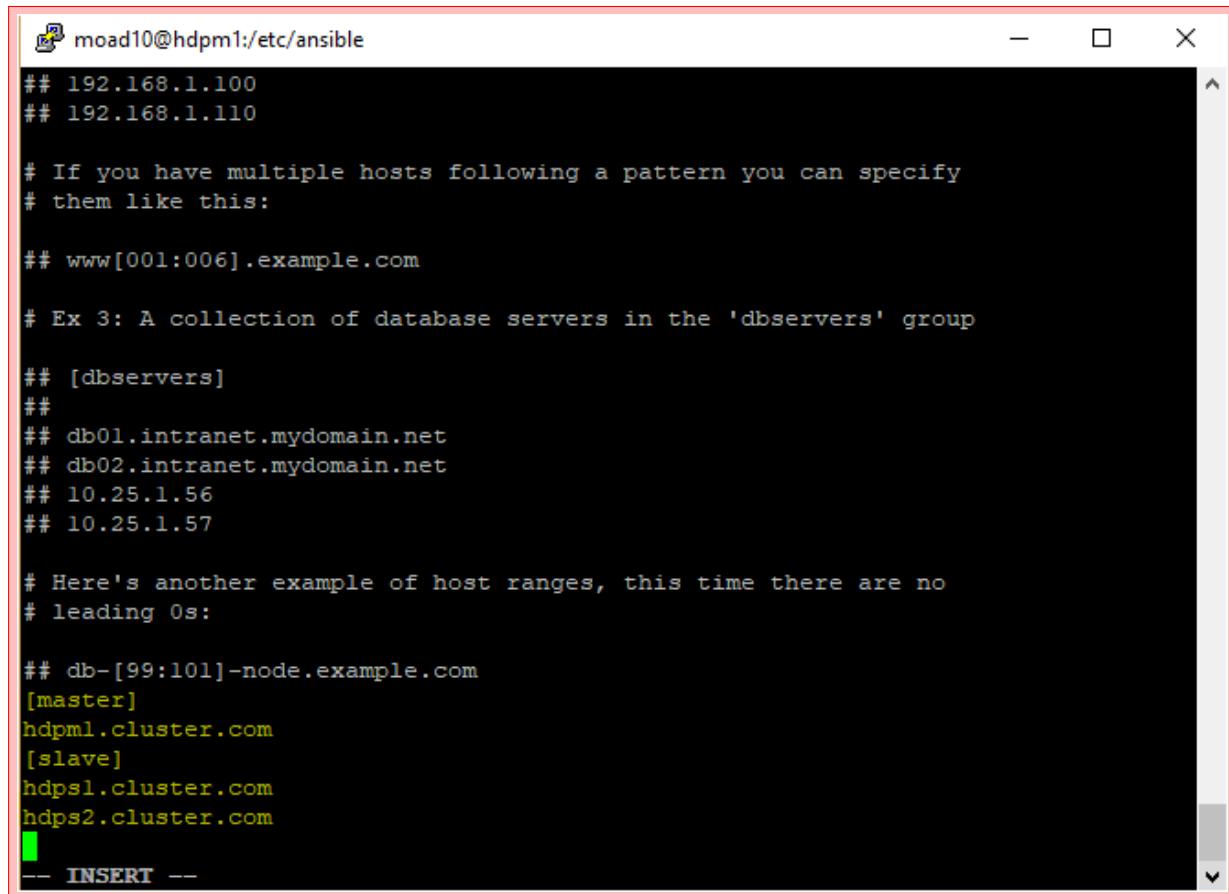
Ansible garde la trace de tous les serveurs qu'il connaît à travers un fichier "hosts". Nous devons d'abord configurer ce fichier avant de pouvoir commencer à communiquer avec nos autres ordinateurs.

Nous ouvrons le fichier avec les privilèges root comme ceci :

```
[moad10@hdpm1 ansible]$ sudo vi hosts
```

Nous spécifions nos 3 clients :

- Un noeud maître : hdpm1,
- Deux noeuds secondaires : hdps1 et hdps2.



```
moad10@hdpm1:/etc/ansible
## 192.168.1.100
## 192.168.1.110

# If you have multiple hosts following a pattern you can specify
# them like this:

## www[001:006].example.com

# Ex 3: A collection of database servers in the 'dbservers' group

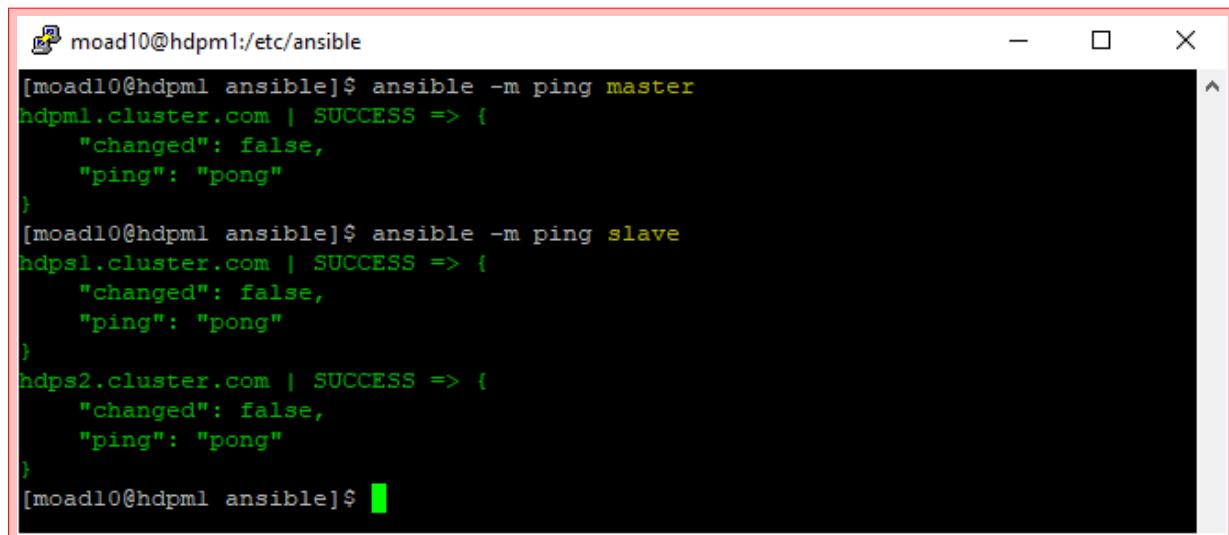
## [dbservers]
##
## db01.intranet.mydomain.net
## db02.intranet.mydomain.net
## 10.25.1.56
## 10.25.1.57

# Here's another example of host ranges, this time there are no
# leading 0s:

## db-[99:101]-node.example.com
[master]
hdpm1.cluster.com
[slave]
hdps1.cluster.com
hdps2.cluster.com

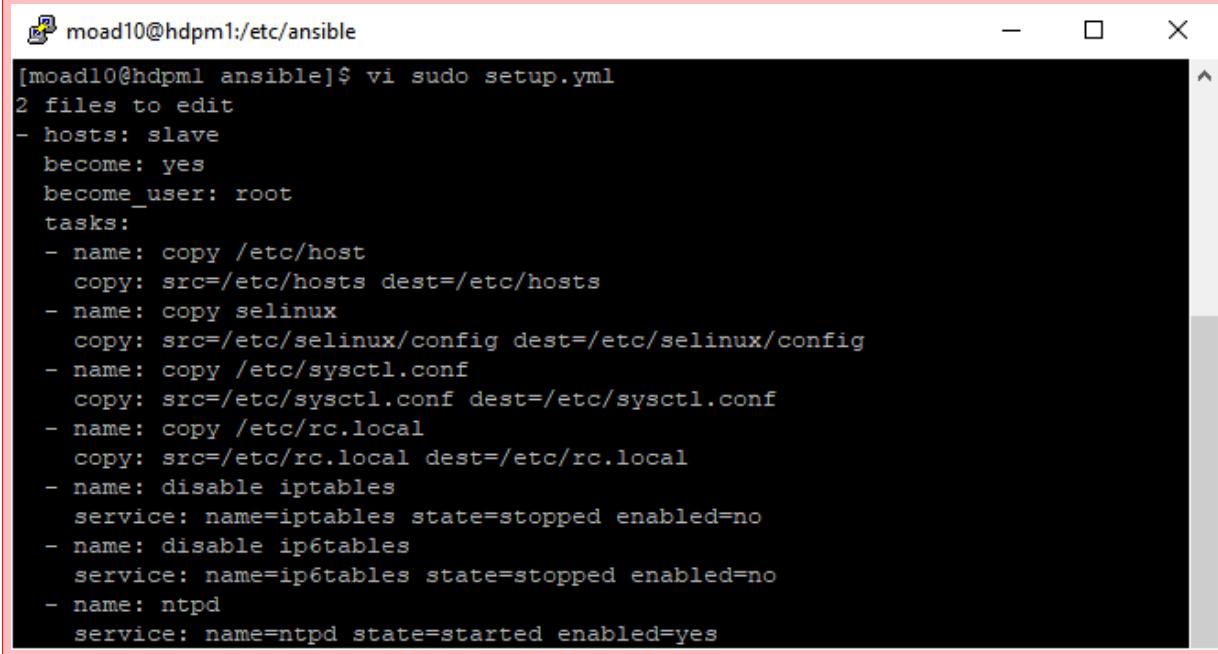
-- INSERT --
```

Ici, nous testons la connexion et l'échange "inter-noeuds", c'est-à-dire que nous pouvons maintenant joindre n'importe quelle machine virtuelle des deux autres et donner des odres aux deux machines esclaves moyennant le noeud maître.



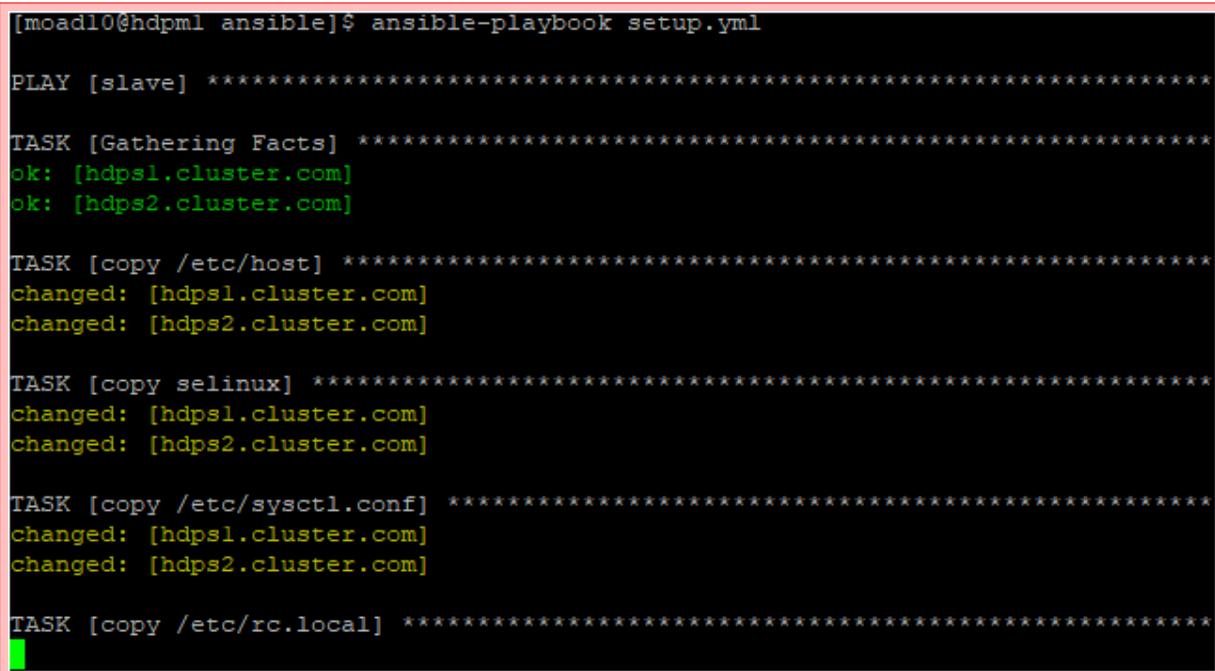
```
[moad10@hdpm1 ansible]$ ansible -m ping master
hdpm1.cluster.com | SUCCESS => {
    "changed": false,
    "ping": "pong"
}
[moad10@hdpm1 ansible]$ ansible -m ping slave
hdps1.cluster.com | SUCCESS => {
    "changed": false,
    "ping": "pong"
}
hdps2.cluster.com | SUCCESS => {
    "changed": false,
    "ping": "pong"
}
[moad10@hdpm1 ansible]$
```

Nous créons un fichier "setup.yml" pour automatiser les tâches en donnant des ordres aux deux machines esclaves à travers le noeud maître.



```
moad10@hdpm1:/etc/ansible
[moad10@hdpm1 ansible]$ vi sudo setup.yml
2 files to edit
- hosts: slave
become: yes
become_user: root
tasks:
- name: copy /etc/host
copy: src=/etc/hosts dest=/etc/hosts
- name: copy selinux
copy: src=/etc/selinux/config dest=/etc/selinux/config
- name: copy /etc/sysctl.conf
copy: src=/etc/sysctl.conf dest=/etc/sysctl.conf
- name: copy /etc/rc.local
copy: src=/etc/rc.local dest=/etc/rc.local
- name: disable iptables
service: name=iptables state=stopped enabled=no
- name: disable ip6tables
service: name=ip6tables state=stopped enabled=no
- name: ntpd
service: name=ntpd state=started enabled=yes
```

Le playbook est composé de modules. Il s'exécute lorsque l'administrateur exécute la commande ansible-playbook sur les machines cibles. L'administrateur doit utiliser un fichier d'inventaire pour spécifier les hôtes sous la gestion du playbook. Le fichier d'inventaire contient une liste de tous les hôtes gérés par Ansible et offre une option pour regrouper les hôtes en fonction de leurs fonctionnalités. Par exemple, un administrateur peut appliquer une lecture à un groupe de serveurs Web dans le playbook, et une lecture différente à un groupe de serveurs de base de données.



```
[moad10@hdpm1 ansible]$ ansible-playbook setup.yml
PLAY [slave] ****
TASK [Gathering Facts] ****
ok: [hdps1.cluster.com]
ok: [hdps2.cluster.com]

TASK [copy /etc/host] ****
changed: [hdps1.cluster.com]
changed: [hdps2.cluster.com]

TASK [copy selinux] ****
changed: [hdps1.cluster.com]
changed: [hdps2.cluster.com]

TASK [copy /etc/sysctl.conf] ****
changed: [hdps1.cluster.com]
changed: [hdps2.cluster.com]

TASK [copy /etc/rc.local] ****
```

```
moad10@hdpm1:/etc/ansible
changed: [hdps1.cluster.com]
changed: [hdps2.cluster.com]

TASK [copy /etc/rc.local] *****
changed: [hdps1.cluster.com]
changed: [hdps2.cluster.com]

TASK [disable iptables] *****
changed: [hdps1.cluster.com]
changed: [hdps2.cluster.com]

TASK [disable ip6tables] *****
changed: [hdps1.cluster.com]
changed: [hdps2.cluster.com]

TASK [ntpd] *****
ok: [hdps1.cluster.com]
ok: [hdps2.cluster.com]

PLAY RECAP *****
hdps1.cluster.com      : ok=8    changed=6    unreachable=0    failed=0
hdps2.cluster.com      : ok=8    changed=6    unreachable=0    failed=0
```

Nous créons deux sessions d'échanges entre le noeud maître et les deux esclaves (hdps1 et hdps2). "ssh hdps1" ouvre un session tandis que la commande "exit;" la ferme.

```
[moad10@hdpm1 ansible]$ ssh hdps1
Last login: Sun Apr  8 00:41:43 2018 from hdpm1.c.hdp-multinode.internal
[moad10@hdps1 ~]$ cat /etc/hosts
127.0.0.1  localhost localhost.localdomain localhost4 localhost4.localdomain4
::1        localhost localhost.localdomain localhost6 localhost6.localdomain6

10.142.0.2 hdpm1.cluster.com hdpm1
10.128.0.2 hdps1.cluster.com hdps1
10.152.0.2 hdps2.cluster.com hdps2

10.142.0.2 hdpm1.c.hdp-multinode.internal hdpm1 # Added by Google
169.254.169.254 metadata.google.internal # Added by Google
[moad10@hdps1 ~]$ exit;
logout
Connection to hdps1 closed.
[moad10@hdpm1 ansible]$ ssh hdps2
Last login: Sun Apr  8 00:41:47 2018 from hdpm1.c.hdp-multinode.internal
[moad10@hdps2 ~]$ cat /etc/hosts
127.0.0.1  localhost localhost.localdomain localhost4 localhost4.localdomain4
::1        localhost localhost.localdomain localhost6 localhost6.localdomain6

10.142.0.2 hdpm1.cluster.com hdpm1
10.128.0.2 hdps1.cluster.com hdps1
10.152.0.2 hdps2.cluster.com hdps2

10.142.0.2 hdpm1.c.hdp-multinode.internal hdpm1 # Added by Google
169.254.169.254 metadata.google.internal # Added by Google
[moad10@hdps2 ~]$
```

Le projet Apache Ambari vise à simplifier la gestion de Hadoop en développant des logiciels pour provisionner, gérer et surveiller les clusters Apache Hadoop. Ambari fournit une interface utilisateur de gestion Hadoop intuitive et facile à utiliser, soutenue par ses API RESTful.

Ambari permet aux administrateurs système de :

- **Provisionner un cluster Hadoop** Ambari fournit un assistant étape par étape pour installer les services Hadoop sur un nombre illimité d'hôtes. Ambari gère la configuration des services Hadoop pour le cluster.
- **Gérer un cluster Hadoop** Ambari fournit une gestion centralisée pour démarrer, arrêter et reconfigurer les services Hadoop sur l'ensemble du cluster.
- **Surveiller un cluster Hadoop** Ambari fournit un tableau de bord pour surveiller l'intégrité et l'état du cluster Hadoop. Ambari exploite Ambari Metrics System pour la collecte des métriques. Ambari exploite Ambari Alert Framework pour l'alerte du système et nous avertit lorsque notre attention est nécessaire (par exemple, un nud tombe en panne, l'espace disque restant est faible, etc.).

Ambari permet aux développeurs d'applications et aux intégrateurs de systèmes d'intégrer facilement les fonctionnalités de provisioning, de gestion et de surveillance Hadoop à leurs propres applications avec les API REST Ambari. Après avoir lancé une commande pour installer le répertoire initial de Ambari, nous installons un serveur ambari qui est à la base de cette démarche.

```
[moadl0@hdpml ~]$ sudo service ambari-server start
Using python /usr/bin/python
Starting ambari-server
Ambari Server running with administrator privileges.
Organizing resource files at /var/lib/ambari-server/resources...
Ambari database consistency check started...
No errors were found.
Ambari database consistency check finished
Server PID at: /var/run/ambari-server/ambari-server.pid
Server out at: /var/log/ambari-server/ambari-server.out
Server log at: /var/log/ambari-server/ambari-server.log
Waiting for server start.....
Ambari Server 'start' completed successfully.
```

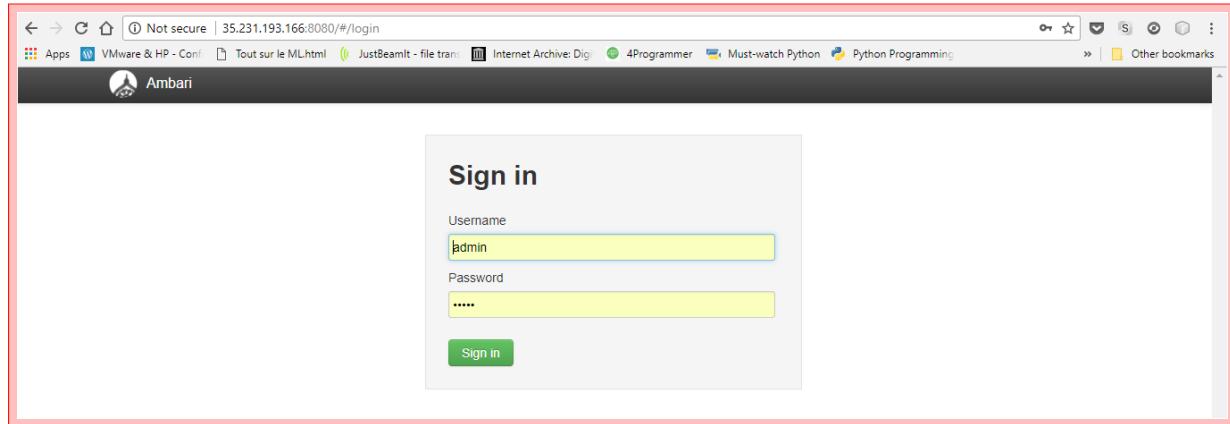
Maintenant, nous pouvons accéder à l'interface web Ambari (hébergée sur le port 8080).

```
[moadl0@hdpml ~]$ sudo netstat -nltp | grep 8080
tcp        0      0 ::::8080          ::::*                      LISTEN
EN        4885/java
```

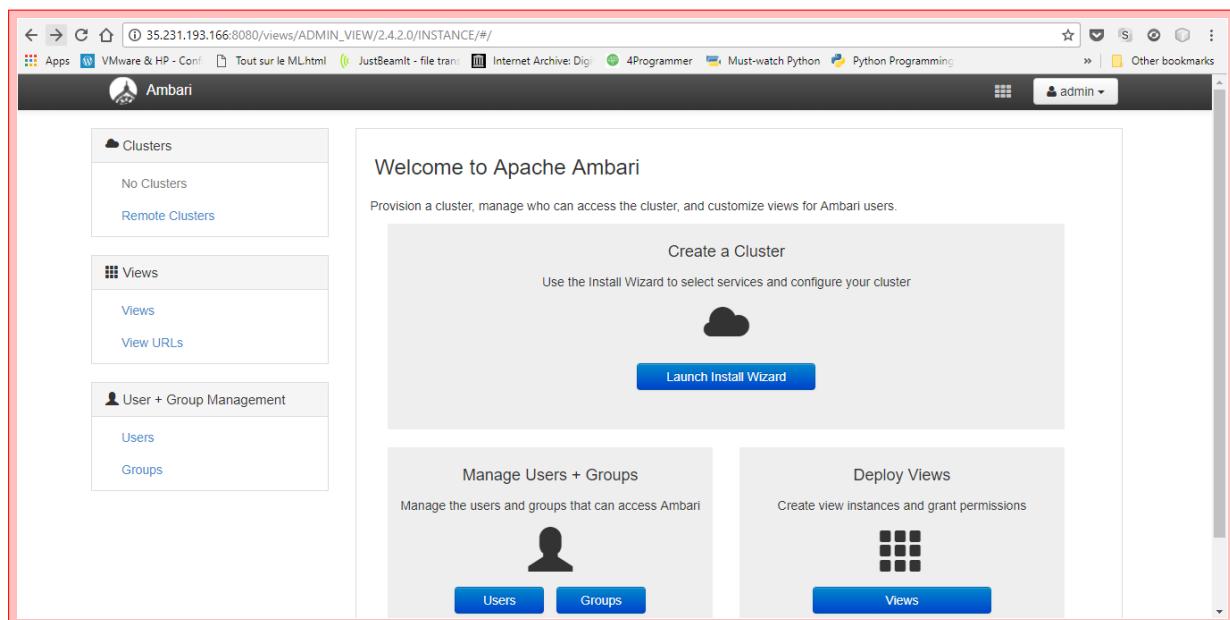
Configuration du cluster Hortonworks Hadoop

1. page de démarrage :

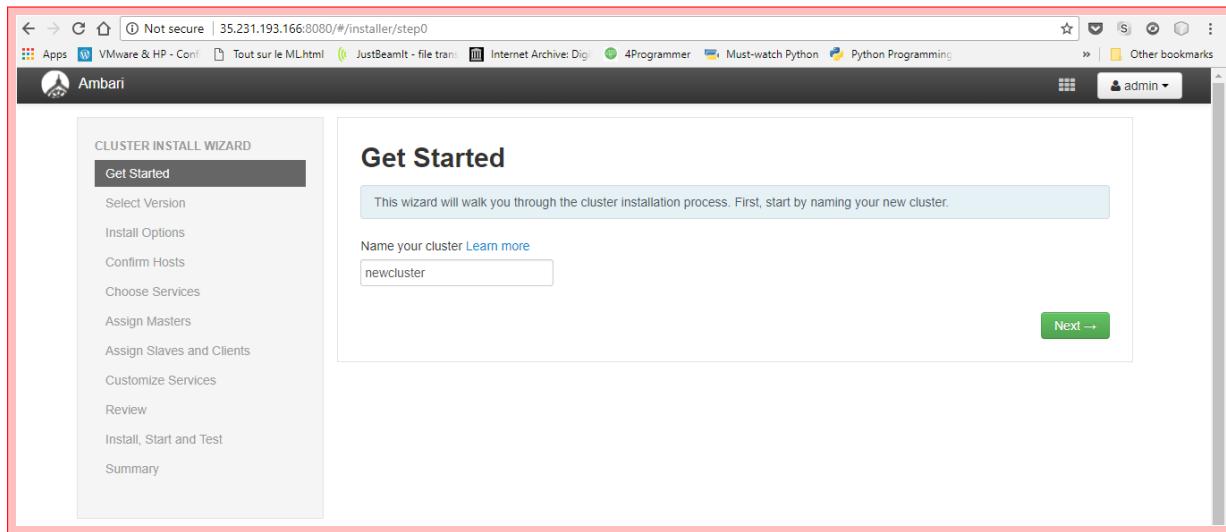
Connectons-nous à Ambari avec le nom d'utilisateur par défaut "admin" et le mot de passe par défaut "admin".



Nous cliquons sur "Lancer l'assistant d'installation" pour démarrer la configuration du cluster.



2. Nom du cluster : Nous donnons un nom à notre cluster.

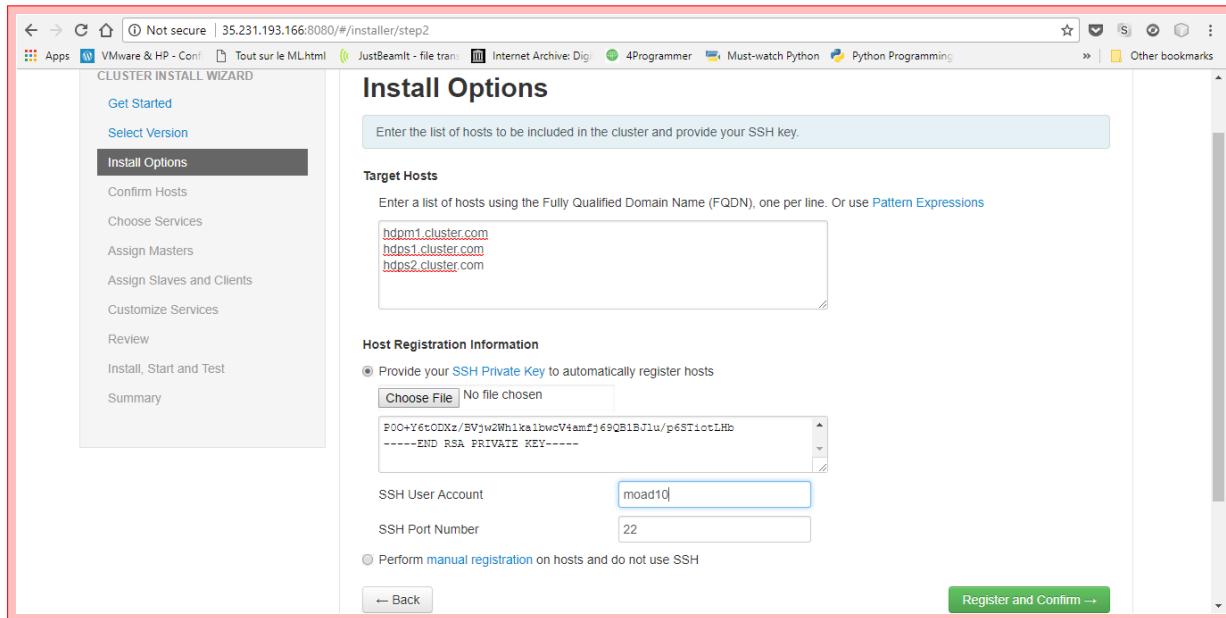


3. Sélection de la pile :

Cette page répertorie les piles disponibles pour l'installation. Chaque pile est pré-emballée avec le composant d'écosystème Hadoop. Ces piles proviennent de Hortonworks.

4. Entrée d'hôte et entrée de clé SSH

Avant d'aller plus loin dans cette étape, nous devrions avoir un mot de passe moins de configuration SSH pour tous les nuds participants.



Nous ajoutons les noms d'hôtes des noeuds, une seule entrée sur chaque ligne. [Ajouter FQDN qui peut être obtenu par la commande hostf -f]. Nous sélectionnons la clé privée utilisée lors de la configuration du mot de passe moins SSH et du nom d'utilisateur à l'aide de la clé privée créée.

5. Hôte le statut d'enregistrement :

CLUSTER INSTALL WIZARD

- [Get Started](#)
- [Select Version](#)
- [Install Options](#)
- Confirm Hosts**
- [Choose Services](#)
- [Assign Masters](#)
- [Assign Slaves and Clients](#)
- [Customize Services](#)
- [Review](#)
- [Install, Start and Test](#)
- [Summary](#)

Confirm Hosts

Registering your hosts.
Please confirm the host list and remove any hosts that you do not want to include in the cluster.

<input type="checkbox"/> Remove Selected		Show: All (3) Installing (0) Registering (0) Success (3) Fail (0)	
Host	Progress	Status	Action
hdpm1.cluster.com	<div style="width: 100%; background-color: #2e7131;"></div>	Success	<input type="button" value="Remove"/>
hdps1.cluster.com	<div style="width: 100%; background-color: #2e7131;"></div>	Success	<input type="button" value="Remove"/>
hdps2.cluster.com	<div style="width: 100%; background-color: #2e7131;"></div>	Success	<input type="button" value="Remove"/>

Show: 25 ▾ 1 - 3 of 3 ⏪ ⏴ ⏵ ⏩ ⏹

Some warnings were encountered while performing checks against the 3 registered hosts above [Click here to see the warnings.](#)

[← Back](#) [Next →](#)

Nous pouvons voir certaines opérations en cours, ces opérations incluent la configuration de Ambari-agent sur chaque nud, créant des configurations de base sur chaque nud. Une fois que nous voyons tout vert, nous sommes prêts à passer à autre chose. Parfois, cela peut prendre du temps car il installe quelques paquets.

6. Choisir les services que nous souhaitons installer :

Selon les piles sélectionnées à l'étape 3, nous avons le nombre de services que nous pouvons installer dans le cluster. Nous pouvons en choisir un que nous voulons. Ambari sélectionne intelligemment les services dépendants si nous ne l'avons pas sélectionné.

CLUSTER INSTALL WIZARD

- [Get Started](#)
- [Select Version](#)
- [Install Options](#)
- [Confirm Hosts](#)
- Choose Services**
- [Assign Masters](#)
- [Assign Slaves and Clients](#)
- [Customize Services](#)
- [Review](#)
- [Install, Start and Test](#)
- [Summary](#)

Choose Services

Choose which services you want to install on your cluster.

Service	Version	Description
<input checked="" type="checkbox"/> HDFS	2.7.1.2.5	Apache Hadoop Distributed File System
<input checked="" type="checkbox"/> YARN + MapReduce2	2.7.1.2.5	Apache Hadoop NextGen MapReduce (YARN)
<input type="checkbox"/> Tez	0.7.0.2.5	Tez is the next generation Hadoop Query Processing framework written on top of YARN.
<input type="checkbox"/> Hive	1.2.1.2.5	Data warehouse system for ad-hoc queries & analysis of large datasets and table & storage management service
<input type="checkbox"/> HBase	1.1.2.2.5	A Non-relational distributed database, plus Phoenix, a high performance SQL layer for low latency applications.
<input type="checkbox"/> Pig	0.16.0.2.5	Scripting platform for analyzing large datasets
<input type="checkbox"/> Sqoop	1.4.6.2.5	Tool for transferring bulk data between Apache Hadoop and structured data stores such as relational databases
<input type="checkbox"/> Oozie	4.2.0.2.5	System for workflow coordination and execution of Apache Hadoop jobs. This also includes the installation of the optional Oozie Web Console which relies on and will install the ExtJS Library.

[← Back](#) [Next →](#)

7. Mappage des services maîtres avec les nuds

Installation de Hadoop avec Ambari Comme nous le savons, l'écosystème Hadoop dispose d'outils basés sur l'architecture maître-esclave. Dans cette étape, nous allons associer les processus maîtres au noeud. Nous nous assurons d'équilibrer correctement notre cluster. Gardons également à l'esprit que les services primaires et secondaires tels que Namenode et Namenode secondaire ne sont pas sur la même machine.

8. Mappage des esclaves avec des nuds

Installation de Hadoop avec Ambari Similaire aux maîtres, nous mappons les services esclaves sur les nuds. En général, tous les nuds auront un processus esclave en cours d'exécution au moins pour les Datanodes et les Nodemangers.

Host	all none	all none	all none	all none
hdpm1.cluster.com*	<input checked="" type="checkbox"/> DataNode	<input type="checkbox"/> NFSGateway	<input checked="" type="checkbox"/> NodeManager	<input checked="" type="checkbox"/> Client
hdps1.cluster.com*	<input checked="" type="checkbox"/> DataNode	<input type="checkbox"/> NFSGateway	<input checked="" type="checkbox"/> NodeManager	<input type="checkbox"/> Client
hdps2.cluster.com	<input checked="" type="checkbox"/> DataNode	<input type="checkbox"/> NFSGateway	<input checked="" type="checkbox"/> NodeManager	<input type="checkbox"/> Client

9. Personnaliser les services

Installation de Hadoop avec Ambari. C'est une page très importante pour les administrateurs. Ici nous pouvons configurer les propriétés de notre cluster pour le rendre le plus adapté à nos cas d'utilisation. En

outre, il va y avoir certaines propriétés requises comme le mot de passe métastore Hive, etc. Ceux-ci seront pointés avec des erreurs rouges comme des symboles.

10. Vérifier et démarrer le provisionnement

Nous nous assurons de revoir la configuration du cluster avant le lancement, car cela nous évitera d'installer des configurations incorrectes.

11. Lancer et rester en arrière jusqu'à ce que le statut devienne VERT.

CLUSTER INSTALL WIZARD

- Get Started
- Select Version
- Install Options
- Confirm Hosts
- Choose Services
- Assign Masters
- Assign Slaves and Clients
- Customize Services
- Review
- Install, Start and Test**
- Summary

Install, Start and Test

Please wait while the selected services are installed and started.

100 % overall

Show: All (3) In Progress (0) Warning (0) Success (3) Fail (0)		
Host	Status	Message
hdpm1.cluster.com	100%	Success
hdps1.cluster.com	100%	Success
hdps2.cluster.com	100%	Success

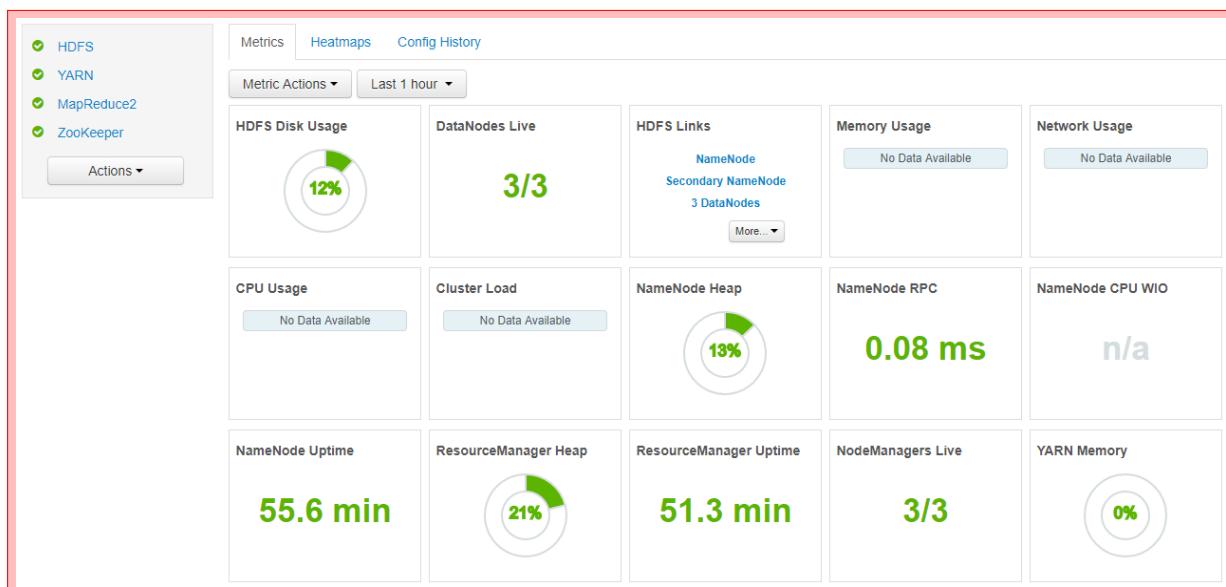
3 of 3 hosts showing - Show All

Show: 25 ▾ 1 - 3 of 3 ⏪ ⏴ ⏵ ⏩ ⏷ ⏸ ⏹

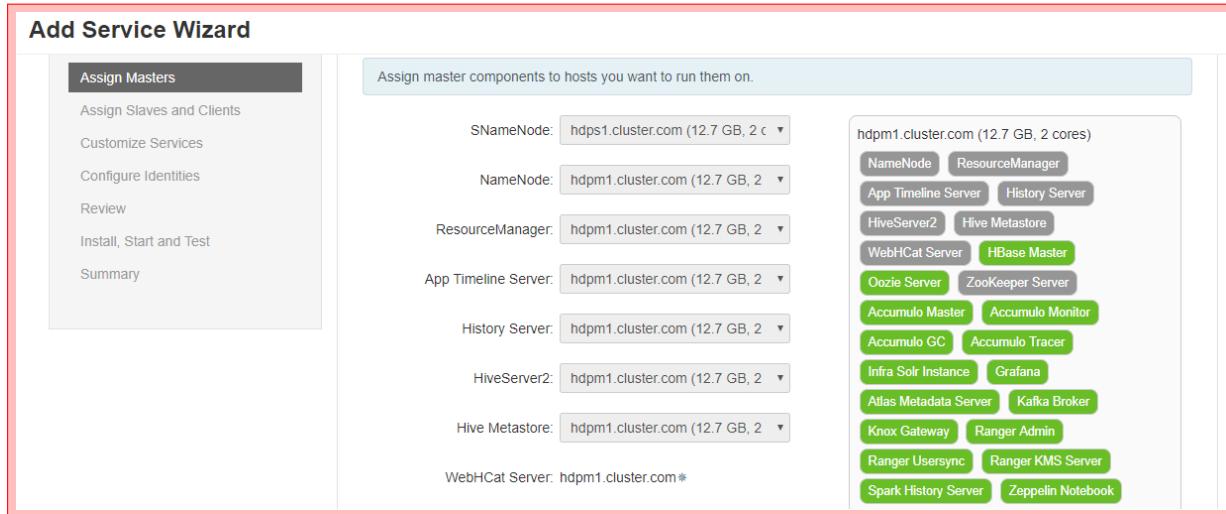
Successfully installed and started the services.

Next →

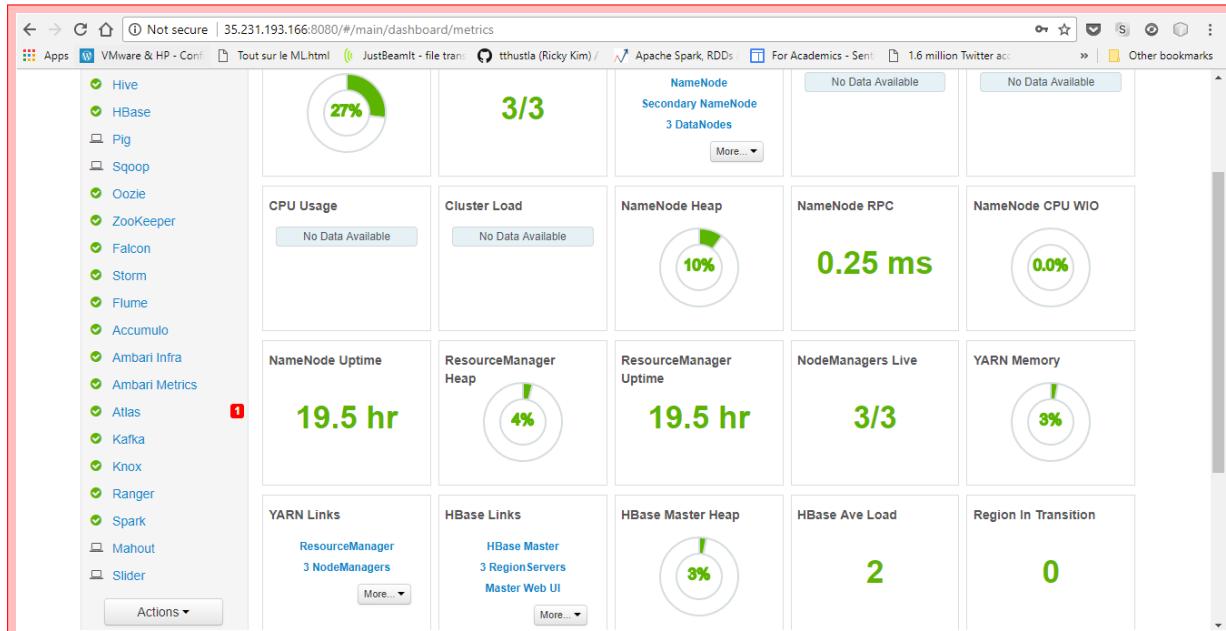
Ici, nous avons installé les 4 principaux packages HDFS (Le système de fichier distribué), YARN (Technologie de gestion de clusters), MapReduce2 (conçu pour lire, traiter et écrire des volumes massifs de données.) et ZooKeeper (Un logiciel de gestion de configuration pour systèmes distribués).



Nous ajoutons plusieurs autres services pour créer un cluster plus au moins complet.



Enfin, nous pouvons visualiser le résultat de plus de 84 configurations : Un cluster Hortonworks multi-noeuds doté de tous les services essentiels aux procédés Big Data.



Spark

Qu'est-ce que Spark? La programmation Spark n'est rien d'autre qu'une plateforme de calcul groupée polyvalente et rapide comme l'éclair. En d'autres termes, il s'agit d'un moteur de traitement de données open source. Cela révèle des API de développement, qui qualifient également les travailleurs de données pour accomplir le streaming, l'apprentissage automatique ou les charges de

travail SQL qui exigent un accès répété aux ensembles de données. Cependant, Spark peut effectuer un traitement par lots et un traitement de flux.

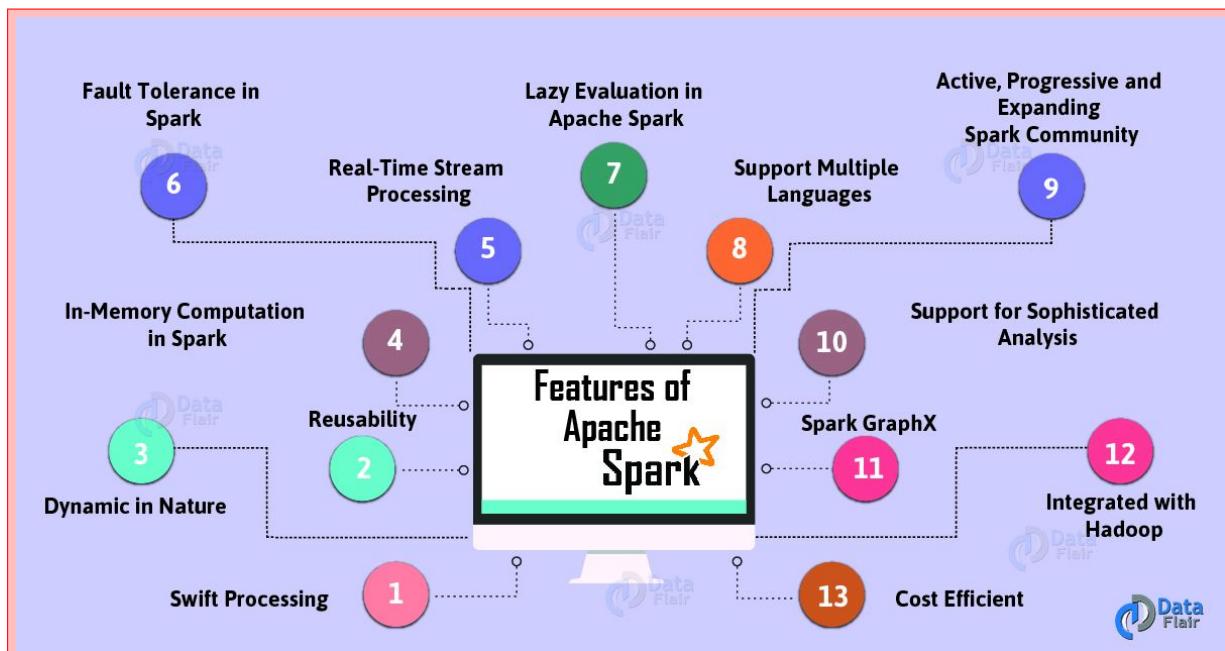
1. Le traitement par lots : est un mode de traitement des données suivant lequel les programmes à exécuter ou les données à traiter sont groupés en lots. Équivalent étranger : batch processing.
2. Le traitement de flux : Un flux de données est une séquence infinie d'éléments générés de façon continue à un rythme rapide. Le terme "rapide" signifie ici que la vitesse d'arrivée des nouveaux éléments est grande devant les capacités de traitement et de stockage disponibles. Les applications produisant des flux de données ou une sortie pouvant être modélisée comme telles ont connus une véritable explosion ces dernières années. On peut par exemple citer les logs de sites web, les tickets de communications fixes ou mobiles, ou bien encore les données de capteurs, de trafic routier par exemple mais aussi d'indices boursiers ou de météorologie. En réponse à ce nouveau besoin, des algorithmes pour traiter ces flux de données ont donc été développés. Ces derniers se trouvant au carrefour de trois champs disciplinaires, l'algorithmique, les bases de données, et bien sûr, les statistiques.

De plus, Spark est conçu de telle sorte qu'il s'intègre à tous les outils Big Data. Comme spark peut accéder à n'importe quelle source de données Hadoop, Spark peut également fonctionner sur les clusters Hadoop. En outre, Apache Spark étend Hadoop MapReduce au niveau suivant. Cela inclut également les requêtes itératives et le traitement de flux. Une croyance plus commune à propos de Spark est que c'est une extension de Hadoop. Bien que ce ne soit pas vrai. Cependant, Spark est indépendant de Hadoop car il possède son propre système de gestion de cluster. Fondamentalement, il utilise Hadoop à des fins de stockage uniquement. Fondamentalement, Apache Spark offre des API de haut niveau aux utilisateurs, tels que Java, Scala, Python et R. Bien que Spark soit écrit en Scala, il propose toujours des API riches en Scala, Java, Python, ainsi que R. Plus important encore, en comparant Spark avec Hadoop, il est 100 fois plus rapide que Big Data Hadoop et 10 fois plus rapide que l'accès aux données à partir du disque.

Resilient Distributed Dataset - RDD L'abstraction clé de Spark est RDD. RDD est un acronyme de Resilient Distributed Dataset. C'est l'unité de données fondamentale dans Spark. Il s'agit d'une collection distribuée d'éléments à travers les noeuds de cluster qui effectue des opérations parallèles. Il existe 3 façons de créer des RDD Spark :

- Collections parallèles : En invoquant la méthode parallelize dans le programme du pilote, nous pouvons créer des collections parallélisées.
- Jeux de données externes : On peut créer des RDD Spark en appelant une méthode `textFile`. Par conséquent, cette méthode prend l'URL du fichier et le lit comme une collection de lignes.
- RDD existants : De plus, nous pouvons créer un nouveau RDD dans spark, en appliquant une opération de transformation sur les RDD existants

Caractéristiques de Apache Spark :



1. Traitement rapide : Apache Spark offre une vitesse de traitement des données élevée. C'est environ 100x plus rapide en mémoire et 10x plus rapide sur le disque. Cependant, cela n'est possible qu'en réduisant le nombre de lecture-écriture sur le disque.
2. Dynamique dans la nature : Fondamentalement, il est possible de développer une application parallèle dans Spark. Comme il y a 80 opérateurs de haut niveau disponibles dans Apache Spark.
3. Calcul en mémoire à Spark : L'augmentation de la vitesse de traitement est possible grâce au traitement en mémoire. Cela améliore la vitesse de traitement.
4. Réutilisation : Nous pouvons facilement réutiliser le code spark pour le traitement par lots ou joindre le flux aux données historiques. Également pour exécuter des requêtes ad-hoc sur l'état du flux.
5. Tolérance aux fautes : Spark offre une tolérance aux pannes. C'est possible grâce à l'abstraction de base de Spark - RDD. Fondamentalement, pour gérer l'échec de tout noeud de travail dans le cluster, les RDD Spark sont conçus. Par conséquent, la perte de données est réduite à zéro.
6. Traitement de flux en temps réel : Nous pouvons faire du traitement de flux en temps réel dans Spark. Fondamentalement, Hadoop ne supporte pas le traitement en temps réel. Il ne peut traiter que les données déjà présentes. Par conséquent, avec Spark Streaming, nous pouvons résoudre ce problème.
7. Évaluation paresseuse dans Spark : Toutes les transformations que nous faisons dans Spark RDD sont paresseuses dans la nature, c'est-à-dire qu'elles

ne donnent pas le résultat tout de suite plutôt qu'un nouveau RDD est formé à partir de celui existant. Ainsi, cela augmente l'efficacité du système.

8. Soutenir plusieurs langues : Spark prend en charge plusieurs langues. Tels que Java, R, Scala, Python. Par conséquent, il montre la dynamicité. En outre, il surmonte également les limitations de Hadoop car il ne peut que construire des applications en Java.
9. Prise en charge d'une analyse sophistiquée : Il y a des outils dédiés dans Apache Spark. Comme pour le streaming de données interactives / requêtes déclaratives, l'apprentissage de la machine qui ajoutent pour mapper et réduire.
10. Intégré avec Hadoop : Comme nous le savons, Spark est flexible. Il peut fonctionner indépendamment et également sur Hadoop YARN Cluster Manager.
11. Spark GraphX : Dans Spark, un composant pour le calcul de graphes et de graphes parallèles, nous avons GraphX. Fondamentalement, il simplifie les tâches d'analyse graphique par la collecte d'algorithmes de graphes et de constructeurs.
12. Rentable : Pour les problèmes de Big data comme dans Hadoop, une grande quantité de stockage et le grand centre de données sont requis lors de la réPLICATION. Par conséquent, la programmation Spark s'avère être une solution rentable.

Mise en place d'une architecture Spark à 6 noeuds

Dans cette partie nous expliciterons les techniques récentes pour mettre en place une topologie Apache Spark avec un master et 5 esclaves. Cette dernière sera dotée d'un notebook appelé Jupyter qui a un rôle majeur dans le prétraitement et la fouille de données au milieu des Data Scientists. Tout d'abord nous nous dirigeons vers "Dataproc" et nous cliquerons sur "cluster" en vue de créer notre cluster Apache Spark.

The screenshot shows the Google Cloud Platform dashboard. On the left, there's a sidebar with various services: Accueil, Endpoints, BIG DATA (Google BigQuery, Pub/Sub, Dataproc, Dataflow, ML Engine, IoT Core, Genomics, Dataprep), and a pinned note 'Les épingles apparaissent ici'. The main area has tabs for 'TABLEAU DE BORD' and 'ACTIVITÉ'. Under 'TABLEAU DE BORD', there are sections for 'Informations sur le projet' (Nom du projet: My First Project, ID du projet: festive-courier-198619, Numéro du projet: 467095716681), 'API' (RQL API, showing requests per second from 0 to 1 over time), 'État de Google Cloud Platform' (Fonctionnement normal de tous les services), 'Facturation' (Frais estimés: 0,00 USD \$), and 'Error Reporting' (Aucun signe d'erreur). A red box highlights the 'Clusters' section under the 'Dataproc' category in the sidebar.

Une fenêtre apparaît, indiquant le chargement du cluster.

The screenshot shows the 'Clusters' page for the 'Dataproc' service. It displays a single cluster entry: 'Cloud Dataproc Clusters'. The description states: 'Google Cloud Dataproc vous permet de configurer des clusters Apache Hadoop et de vous connecter à des datastores analytiques sous-jacents.' Below this is a button labeled 'Activation...'. A red box highlights this cluster entry.

A cette étape, nous avons attribué un nom au cluster "spark" créé dans la région "us-east1" dans la zone "c". Il y a cependant plusieurs régions couvrant le globe, chacune d'elle à un prix de location donné.

Zones et régions :

Certaines ressources de Compute Engine résident dans des régions ou des zones. Une région est un lieu géographique spécifique où nous pouvons gérer nos ressources. Chaque région a une ou plusieurs zones. Par exemple, la région us-central1 désigne une région du centre des États-Unis qui comporte des zones us-central1-a, us-central1-b, us-central1-c et us-central1-f.

Les ressources qui résident dans une zone, telles que les instances ou les disques persistants, sont appelées ressources zonales. Les autres ressources, telles que les adresses IP externes statiques, sont régionales. Les ressources régionales peuvent être utilisées par toutes les ressources de cette région, indépendamment de la zone, tandis que les ressources zonales ne peuvent être utilisées que par d'autres

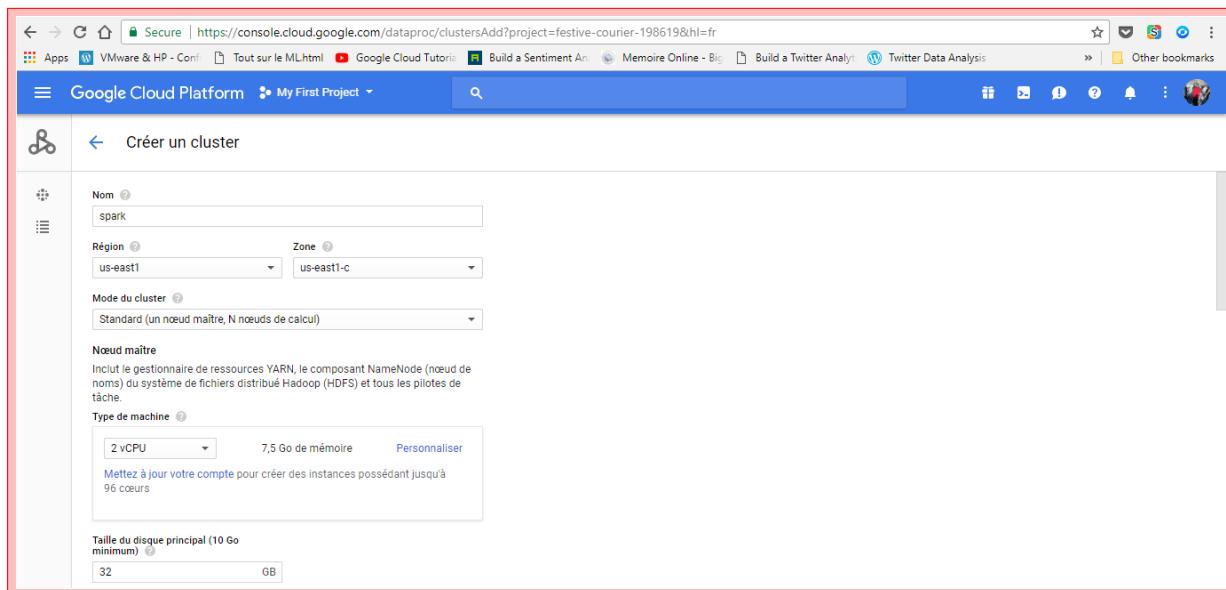
ressources de la même zone. Par exemple, les disques et les instances sont les deux ressources zonales. Pour attacher un disque à une instance, les deux ressources doivent être dans la même zone. De même, si l'on souhaite attribuer une adresse IP statique à une instance, celle-ci doit se trouver dans la même région que l'adresse IP statique.

1. Noeud principal

Type de machine : Nous changeons la valeur par défaut : (2 vCPU, 7,5GO) Cluster Mode : Il y a 3 modes ici. Mode simple (1 maître, 0 travailleur), Mode standard (1 maître, travailleur N) et Mode élevé (3 maîtres, N travailleurs). Taille de disque primaire : Pour mes tests, 32 GB.

2. Noeud de travail

Configuration similaire à celle du noeud maître. Nous utilisons 5 noeuds de travail et la taille du disque est de 32 Go. Vous remarquerez qu'il existe une option pour utiliser le stockage SSD local. Vous pouvez connecter jusqu'à 8 périphériques SSD locaux à l'instance de machine virtuelle. Chaque disque a une taille de 375 Go. Les disques SSD locaux sont physiquement attachés au serveur hôte et offrent des performances supérieures et un stockage de latence inférieur à celui du stockage sur disque persistant de Google. Les disques SSD locaux sont utilisés pour les données temporaires, telles que le brassage des données dans MapReduce. Les données sur le stockage SSD local ne sont pas persistantes. Pour plus d'informations, consultez la page <https://cloud.google.com/compute/docs/disks/local-ssd>.



Nous cliquons sur "Créer" lorsque tout est terminé. Après quelques minutes, le cluster à 6 noeuds est créé.

The screenshot shows the Google Cloud Platform Compute Engine Instances page. On the left, there's a sidebar with options like Instances de VM, Groupes d'instances, Modèles d'instances, Disques, Instantanés, Images, TPUs, Remises sur les engagements, Métadonnées, Vérifications d'état, and Solutions de VM. The main area displays a table of instances:

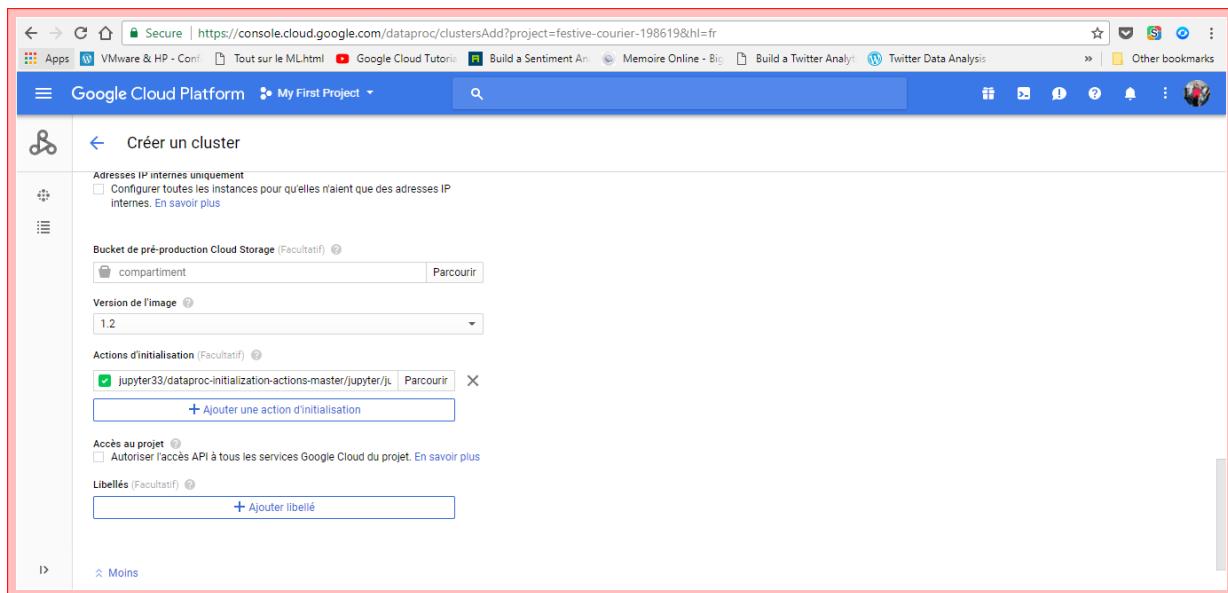
Nom	Zone	Recommandation	Adresse IP interne	Adresse IP externe	Se connecter
spark-m	us-east1-c		10.142.0.3	104.196.215.242	SSH
spark-w-0	us-east1-c		10.142.0.4	35.231.127.175	SSH
spark-w-1	us-east1-c		10.142.0.5	35.196.115.183	SSH
spark-w-2	us-east1-c		10.142.0.2	35.231.79.168	SSH
spark-w-3	us-east1-c		10.142.0.6	35.227.88.21	SSH
spark-w-4	us-east1-c		10.142.0.7	35.185.19.121	SSH

Ensuite, nous devons configurer Jupyter. Sachant que gsutil nous permet d'accéder à la ligne de commande pour gérer des compartiments et des objets Cloud Storage. La commande : "gsutil ls gs://dataproc-initialization-actions" nous montre qu'on peut ajouter Hive, Drill ou encore Jupyter et plein d'autres outils stratégiques à notre topologie via des scripts pré-configurés par l'équipe Apache Spark.

The screenshot shows a Command Prompt window with the following text:

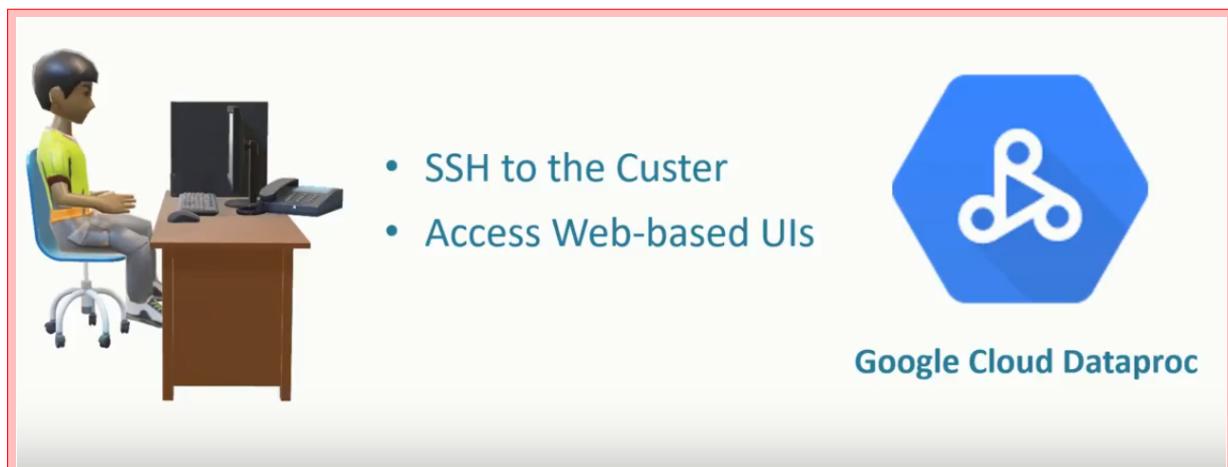
```
$ gcloud components update
gs://dataproc-initialization-actions/CONTRIBUTING.md
gs://dataproc-initialization-actions/LICENSE
gs://dataproc-initialization-actions/README.md
gs://dataproc-initialization-actions/favicon.ico
gs://dataproc-initialization-actions/apache-zppelin/
gs://dataproc-initialization-actions/cloud-sql-proxy/
gs://dataproc-initialization-actions/conda/
gs://dataproc-initialization-actions/databab/
gs://dataproc-initialization-actions/drill/
gs://dataproc-initialization-actions/flink/
gs://dataproc-initialization-actions/ganglia/
gs://dataproc-initialization-actions/hive-hcatalog/
gs://dataproc-initialization-actions/hue/
gs://dataproc-initialization-actions/ipython-notebook/
gs://dataproc-initialization-actions/jupyter/ ←
gs://dataproc-initialization-actions/kafka/
gs://dataproc-initialization-actions/list-consistency-cache/
gs://dataproc-initialization-actions/oozie/
gs://dataproc-initialization-actions/post-init/
gs://dataproc-initialization-actions/presto/
gs://dataproc-initialization-actions/stackdriver/
gs://dataproc-initialization-actions/tez/
gs://dataproc-initialization-actions/user-environment/
gs://dataproc-initialization-actions/util/
gs://dataproc-initialization-actions/zppelin/
gs://dataproc-initialization-actions/zookeeper/
```

Comme le montre l'image ci-dessous, nous ajoutons Jupyter aux actions l'initialisation.



The screenshot shows the 'Créer un cluster' (Create a cluster) page in the Google Cloud Platform. It includes fields for internal IP addresses, a pre-production Cloud Storage bucket, image version (1.2), initialization actions (Jupyter notebook), project access (API permissions), and labels. A red border highlights the entire form area.

Comment accéder à Dataproc cluster ? En effet il existe au moins deux manières : (SSH et accès Web)



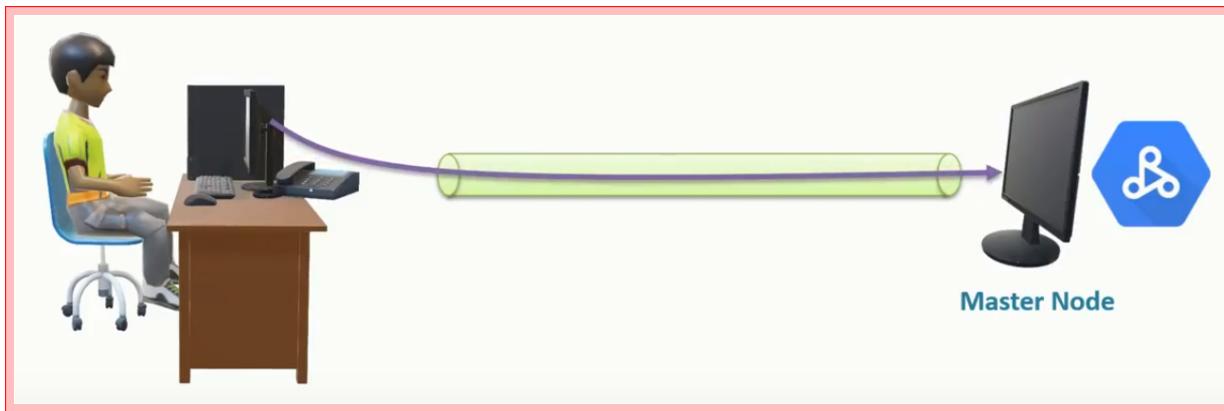
The diagram illustrates two ways to access a Google Cloud Dataproc cluster: 'SSH to the Cluster' and 'Access Web-based UIs'. It features a cartoon character sitting at a desk with a computer, a blue hexagonal logo for Google Cloud Dataproc, and the text 'Google Cloud Dataproc'.

Voyons à présent la première méthode : La connexion sécurisée à distance avec SSH. Secure Shell (SSH) est à la fois un programme informatique et un protocole de communication sécurisé. Le protocole de connexion impose un échange de clés de chiffrement en début de connexion. Par la suite, tous les segments TCP sont authentifiés et chiffrés.

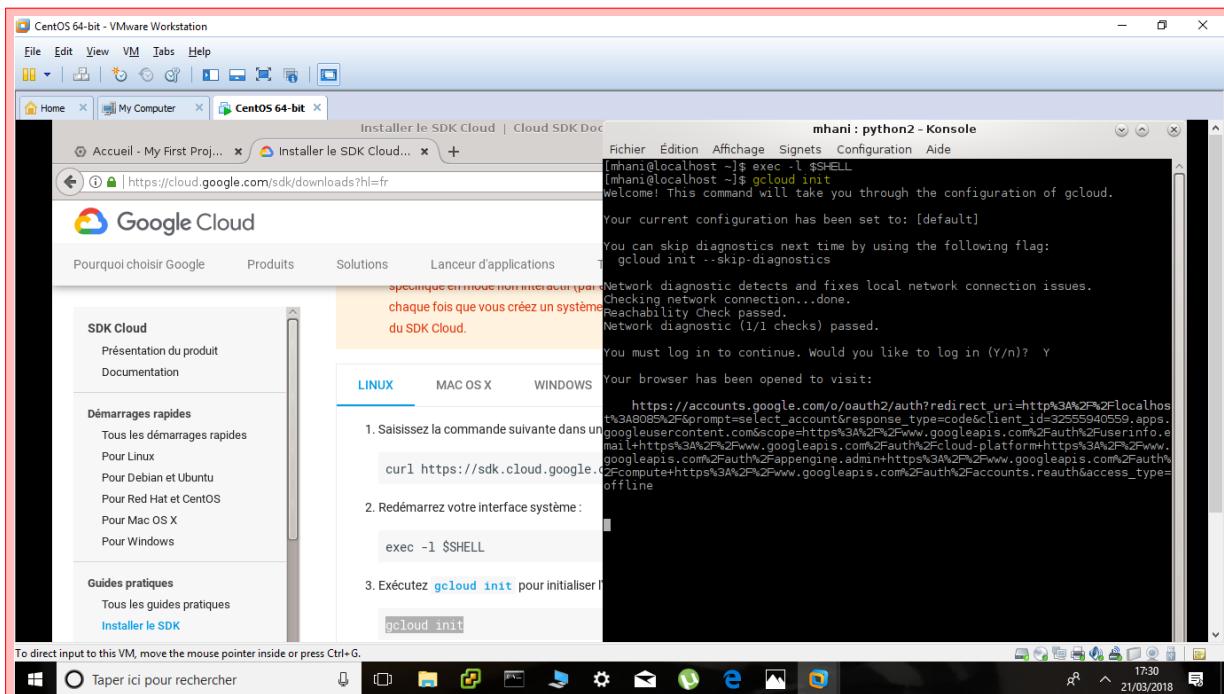
```
moadhaniii@spark-m: ~ - Google Chrome
Secure | https://ssh.cloud.google.com/projects/festive-courier-198619/zones/us-east1-c/instances/spark-m?authuser=0&hl=fr&projectNu...
Connected, host fingerprint: ssh-rsa 2048 58:B2:F8:5E:74:4C:13:A9:59:C8:D0:7B:69:EB:89:94:76:0A:48:94
The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*copyright.

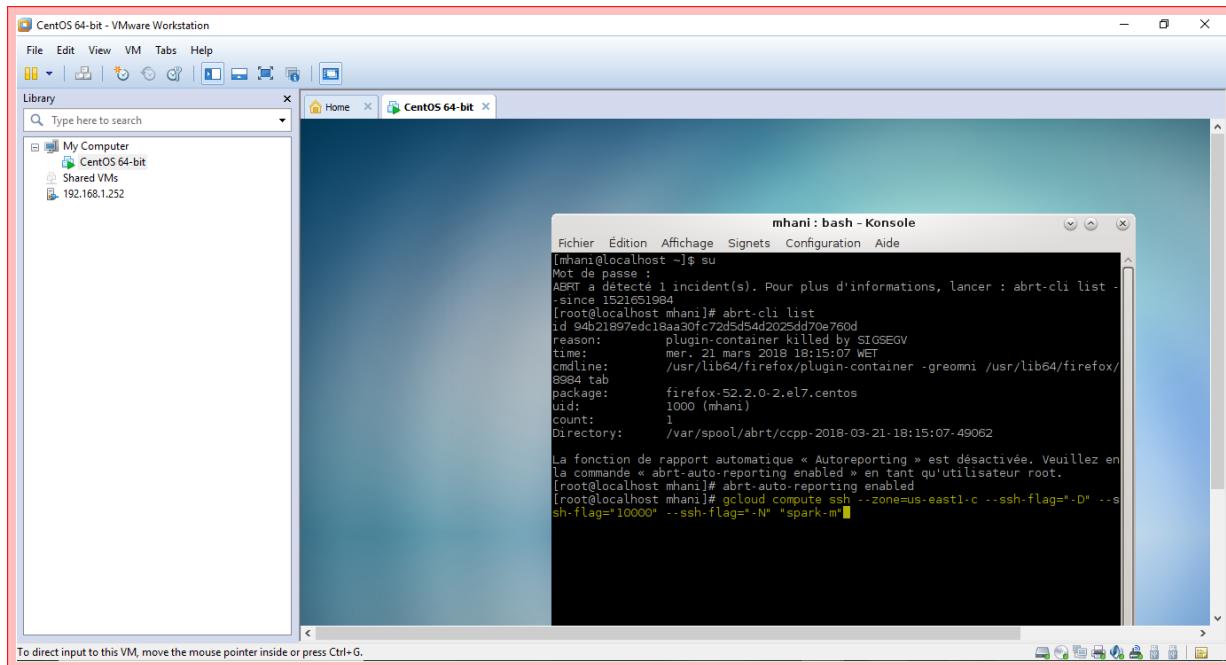
Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
moadhaniii@spark-m:~$ hadoop fs -ls /
18/03/21 02:37:30 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.3-hadoop2
Found 2 items
drwxrwxrwt - mapred hadoop 0 2018-03-21 02:15 /tmp
drwxrwxrwt - hdfs hadoop 0 2018-03-21 02:15 /user
moadhaniii@spark-m:~$ hadoop fs -ls /user
18/03/21 02:39:27 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.3-hadoop2
Found 8 items
drwxrwxrwt - hdfs hadoop 0 2018-03-21 02:15 /user/hbase
drwxrwxrwt - hdfs hadoop 0 2018-03-21 02:15 /user/hdfs
drwxrwxrwt - hdfs hadoop 0 2018-03-21 02:15 /user/hive
drwxrwxrwt - hdfs hadoop 0 2018-03-21 02:15 /user/mapred
drwxrwxrwt - hdfs hadoop 0 2018-03-21 02:15 /user/pig
drwxrwxrwt - hdfs hadoop 0 2018-03-21 02:15 /user/spark
drwxrwxrwt - hdfs hadoop 0 2018-03-21 02:15 /user/yarn
drwxrwxrwt - hdfs hadoop 0 2018-03-21 02:15 /user/zookeeper
moadhaniii@spark-m:~$
```

Quant à la seconde méthode : l'accès Web est ouvert via des ports qu'on peut voir par le biais de la commande "netstat -a | grep LISTEN | grep tcp". On constatera que jupyter est accessible via le port : 8123. Bien qu'on a l'option de configurer une "Firewall Rule" qui va nous permettre d'ouvrir Jupyter dans un navigateur cela s'avère très risqué du fait que ces services ne sont pas sécurisés. Nous avons pensé à une alternative : Tunnel SSH vers le noeud principal.



Pour nous connecter au serveur Web local sur le noeud maître, vous créerez un tunnel SSH entre notre ordinateur et le noeud maître. Ceci est également connu sous le nom de redirection de port. Si nous créons notre tunnel SSH à l'aide du transfert de port dynamique, tout le trafic acheminé vers un port local inutilisé spécifié est transmis au serveur Web local sur le noeud maître. Cela créera un proxy SOCKS. Nous pouvons ensuite configurer notre navigateur Internet pour utiliser un add-on tel que FoxyProxy ou SwitchySharp pour gérer nos paramètres de proxy SOCKS. L'utilisation d'un module complémentaire de gestion de proxy nous permet de filtrer automatiquement les URL en fonction des modèles de texte et de limiter les paramètres de proxy aux domaines qui correspondent à la forme du nom DNS public du noeud maître. Le module complémentaire du navigateur gère automatiquement l'activation et la désactivation du proxy lorsque nous basculons entre l'affichage des sites Web hébergés sur le noeud maître et ceux sur Internet. Note : SOCKS est un protocole réseau qui permet à des applications client-serveur d'employer d'une manière transparente les services d'un pare-feu.





La commande "gcloud compute ssh" à pour rôle va ouvrir un tunnel depuis le port 10.000 de notre machine locale vers la zone GCP qui est "us-east1-c" et le noeud principal. La commande ' –ssh-flag="-D" ' est pour permettre la redirection de port dynamique tandis que la seconde : ' –ssh-flag="-N" ' est pour empêcher gcloud d'ouvrir "remote shell".

The screenshot shows the Hadoop Cluster Metrics page with a table of cluster metrics. A specific Spark application entry is highlighted with a red box.

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Contain
application_150988677773_0001	prashant	Spark shell	SPARK	default	0	Sun Nov 5 19:49:03 +0500	N/A	RUNNING	UNDEFINED	2

The screenshot shows a Jupyter Notebook interface with a single cell labeled "In []:". The status bar at the bottom indicates "Kernel starting, please wait..." and "Trusted".

2.1.2 Analyse des sentiments - Les exemples d'exploration de Tweets et de traitement d'images

Comment une machine peut-elle comprendre les émotions, les sentiments et les opinions d'un humain ? Il est possible grâce à l'analyse du sentiment, qui utilisent le langage naturel de traitement connu sous le nom de PNL ou d'apprentissage automatique qui aide à découvrir le contexte derrière le contenu ! Prenons l'exemple - "J'ai raté le vol ! Impressionnant". Bien sur, personne n'aime perdre son vol. Mais une analyse générale du texte choisira le mot "Impressionnant" et le présentera comme un commentaire positif même s'il s'agit d'un "commentaire sarcastique". La complexité d'un langage humain est beaucoup plus élevée car c'est un jeu complexe de mots et d'émotions. Par conséquent, l'analyse des sentiments ne nous aide pas seulement à dire «ce qui est mauvais», mais nous permet également de savoir «pourquoi c'est mauvais».

La généralisation des accès internet haut-débit et l'usage massif des Smartphones et des réseaux sociaux ont démocratisé la création, la production et la diffusion des images : elles envahissent aujourd'hui la toile ! Dans cet océan d'images, comment les entreprises peuvent-elles surveiller leurs visuels produits, leurs campagnes publicitaires, leurs logos et toutes les autres images faisant référence positivement ou négativement à leurs marques ? Une image vaut mille mots. Et bien que les écrits n'appartiennent pas encore tout à fait au passé, le contenu visuel fait partie intégrante de la manière dont les messages sont partagés sur le Web. Nous pouvons analyser les contenus visuels 60 000 fois plus vite qu'un texte, et nous sommes de plus en plus impatients.

AVANTAGES DE CETTE ETUDE POUR L'UITS (ET POUR TOUTE AUTRE ENTREPRISE QUI MISE SUR L'ANALYSE DES SENTIMENTS) :

1. Améliorer l'expérience client - L'analyse des sentiments aide à surveiller les avis en ligne affichés par les clients sur différents canaux d'un même panel, aidant ainsi à identifier les problèmes rencontrés par les clients et à les résoudre le plus tôt possible.
2. Construire la réputation en ligne - L'analyse du sentiment vous aide à évaluer l'opinion de vos produits ou de votre marque. Il vous indique également si votre marque / produit est discuté et ce qui en est dit, en particulier dans le cas des sites de réseaux sociaux.
3. Identifier les opportunités et améliorer les caractéristiques du produit - Les commentaires mis à jour vous permettent d'identifier les failles dans votre produit ou service et ainsi vous donner l'occasion de vous améliorer et de vous tenir au-dessus de la foule.

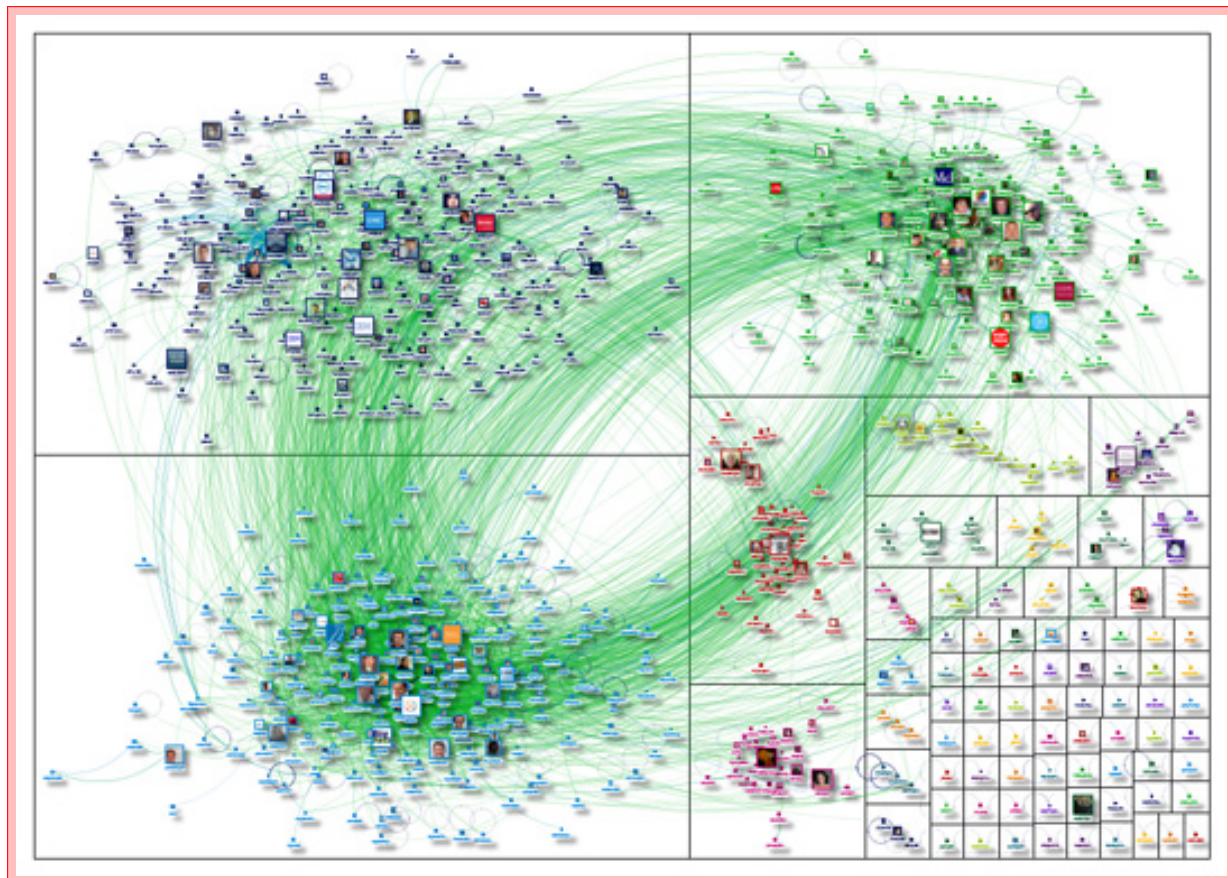
Twitter représente un instrument fondamentalement nouveau pour faire des mesures sociales. Il regorge d'informations utiles pour le marketing et le ciblage de la clientèle. Des millions de personnes expriment volontairement des opinions sur n'importe quel sujet imaginable. Cette source de données est incroyablement

précieuse pour la recherche et les affaires. Par exemple, les chercheurs ont montré que «l'humeur» de la communication sur Twitter reflète les rythmes biologiques et peut même servir à prédire le marché boursier. Un tweet est un objet complexe dans la science des données car, en plus d'être non structuré, il a plusieurs attributs d'où la nécessité de la phase de nettoyage du texte. Plus on donne de l'importance à cette dernière, plus les résultats seront précis. Nous souhaitons dans cette section appliquer les méthodes de data mining aux tweets. Il faut pour cela que nous les transformions en matrice de données numériques. En effet, les algorithmes ne savent pas apprêhender en l'état les documents textuels.

Qu'est qu'un tweet? Un tweet est un post sur Twitter. L'acte d'écrire un tweet s'appelle "Tweeting" ou "Twittering". Les tweets peuvent contenir jusqu'à 140 caractères, espaces compris, et peuvent inclure des URL et des hashtags. Comme tout autre site de médias sociaux, Twitter connaît beaucoup de choses sur nous, grâce aux "métadonnées". En effet, pour un message de 140 caractères, nous obtiendrons BEAUCOUP de métadonnées, soit plus de 20 fois la taille du contenu initial que nous avons saisi! La quasi-totalité de ces métadonnées est accessible via l'API Twitter. Voici quelques exemples qui pourraient être exploités par n'importe qui pour «prendre des empreintes digitales» et suivre quelqu'un :

- Fuseau horaire et langue définis pour l'interface Twitter
- Langues détectées dans les tweets
- Sources utilisées (application mobile, navigateur web, ...)
- Géolocation
- Les hashtags les plus utilisés, la plupart des utilisateurs retweetés, etc.
- Activité quotidienne / hebdomadaire

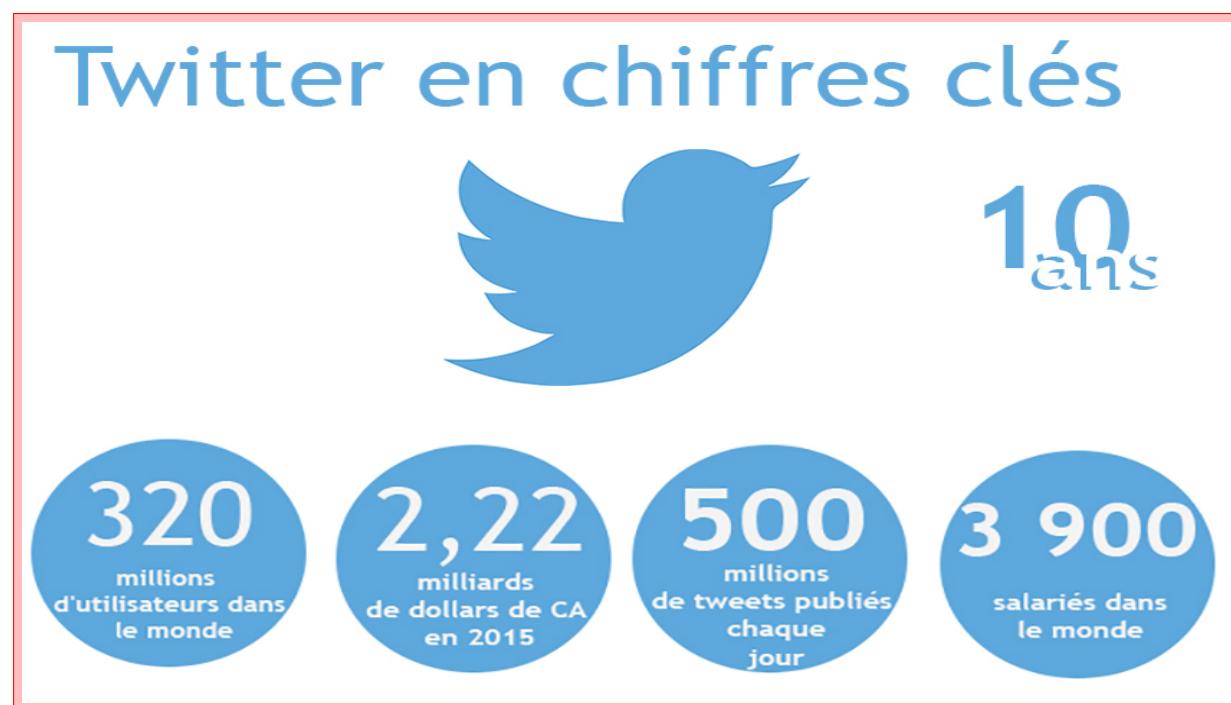
Twitter et sécurité : Tout le monde connaît le danger des fuites de géolocalisation et comment cela peut affecter la vie privée. Mais peu réalisent que juste tweeting régulièrement peut en dire beaucoup sur leurs habitudes. La mise à part d'un seul tweet peut révéler des métadonnées intéressantes.



Réseau Twitter et exploitation des données

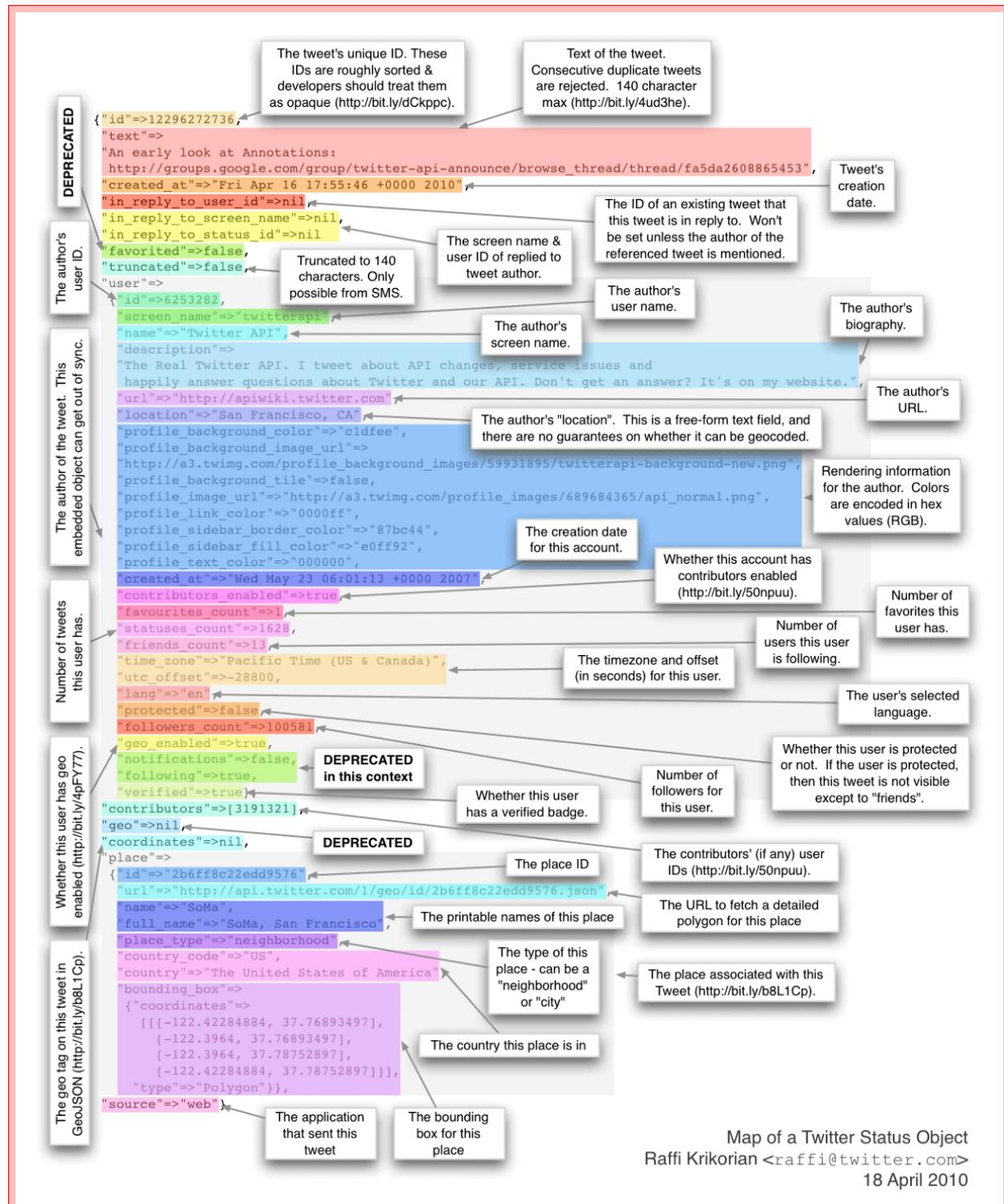
Twitter, un outil marketing plus puissant que Facebook Twitter est un réseau social majeur pour la recherche, l'engagement et la mise en avant de la marque. En effet, les utilisateurs de Twitter sont trois fois plus susceptibles de suivre une marque que les utilisateurs de Facebook. De plus, une étude de Twitter a révélé que 69 % des personnes interrogées avait déjà acheté auprès d'une PME après l'avoir suivie sur Twitter. L'étude a également indiqué que 79 % des personnes qui suivent une PME retweetent du contenu de cette entreprise. Une autre étude révèle que 63,5 % des responsables des médias sociaux ont listé Twitter comme l'une des principales plateformes en matière de ROI.

- **Générer des leads** en apprenant à connaître les personnes qui interagissent avec votre marque, à savoir pourquoi elles partagent votre contenu et avec qui.
- **Trouver les influenceurs** de votre secteur avec qui communiquer.
- **Analyser la concurrence** pour trouver des informations détaillées sur leurs tweets, mentions, hashtags, abonnés et bien plus encore.
- **Trouver les sujets tendance** par contenu, hashtag, terme de recherche, source, et plus encore.



Twitter en chiffres

En regardant la structure d'un tweet, on constate que les objets Tweet sont complexes ; ils contiennent même des sous-objets. Par exemple, nous pouvons extraire l'objet utilisateur du tweet pour obtenir des informations sur l'utilisateur.



Exemple de Tweet (2010 - l'API a beaucoup changé depuis)

2.1.3 Applications en R

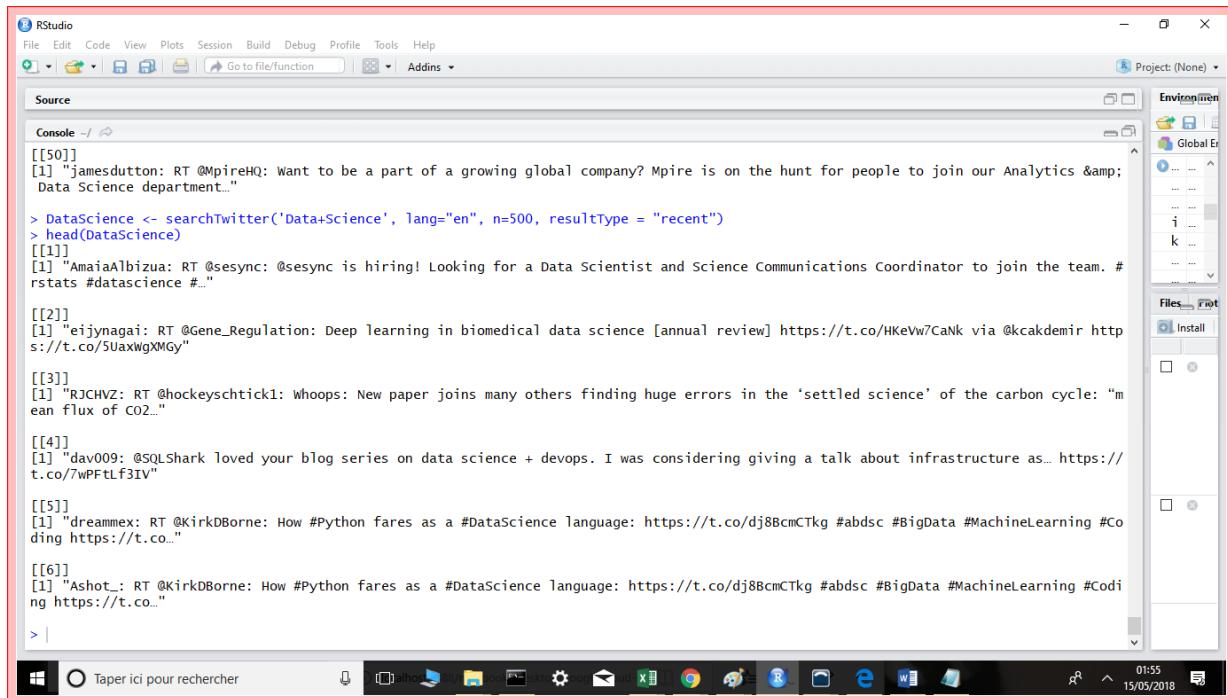
Nous devons d'abord créer une application Twitter en se rendant sur le site : <https://apps.twitter.com/> Nous présentons dans ce chapitre notre solution pour assurer une collecte efficace de toutes les données textuelles qui se rapportent à l'activité des internautes sur Twitter en plus des dispositions prises pour leur sauvegarde. Nous introduirons également les étapes de nettoyage, de structuration et de visualisation qui interviennent à la suite de la collecte. Nous avons pu établir une première connexion pour extraire la data ciblée sur Twitter. Voici un extrait du code utilisé pour nous connecter à notre compte twitter et faire des premières manipulations :

Procédure d'authentification à Twitter via R

```
require(twitteR)  
  
consumer_key <- ''  
consumer_secret <- ''  
  
access_token <- ''  
access_secret <- ''  
  
setup_twitter_oauth()
```

Nous voudrons chercher et analyser les derniers tweets sur le net à propos de "la Data Science" afin de faire de l'analyse d'opinions et aboutir à réaliser un genre de graphes très utilisés par les Data Scientists nommé : "Word Cloud" ou "nuage de mots-clés" qui est une représentation visuelle des mots-clefs (tags) les plus utilisés sur twitter. Généralement, les mots s'affichent dans des tailles et des couleurs de caractères d'autant plus visibles qu'ils sont utilisés ou populaires.

Nous commençons par faire une recherche des 500 derniers tweets en anglais contenant les deux mots "Data" et (+) "Science". On enregistre le résultat dans la variable DataScience et on visualise les 6 premiers à l'aide de la fonction head() qui, à l'inverse de tail(), renvoie la première partie d'un vecteur, d'une matrice, d'une table, d'un bloc de données ou d'une fonction.



The screenshot shows the RStudio interface with a red border around the main window. The console tab is active, displaying a list of 500 tweets. The tweets are printed in blue, indicating they are URLs. The RStudio environment pane on the right shows various global variables like i, k, etc. The taskbar at the bottom shows several open applications.

```
[50]
[1] "jamesdutton: RT @MpireHQ: Want to be a part of a growing global company? Mpire is on the hunt for people to join our Analytics & Data Science department...">
> DataScience <- searchTwitter('Data+Science', lang="en", n=500, resultType = "recent")
> head(DataScience)
[1]
[1] "AmaiaAlbizua: RT @sesync: @sesync is hiring! Looking for a Data Scientist and Science Communications Coordinator to join the team. #rstats #datascience #...">
[2]
[1] "eijynagai: RT @Gene_Regulation: Deep learning in biomedical data science [annual review] https://t.co/HKeVw7CaNk via @kcakdemir http://t.co/5UaxWgXMGY">
[3]
[1] "RJCHVZ: RT @hockeyschtick1: Whoops: New paper joins many others finding huge errors in the 'settled science' of the carbon cycle: "mean flux of CO2...">
[4]
[1] "dav009: @SQLShark loved your blog series on data science + devops. I was considering giving a talk about infrastructure as... https://t.co/7wPFtLf3IV">
[5]
[1] "dreammex: RT @KirkDBorne: How #Python fares as a #DataScience language: https://t.co/dj8BcmCTkg #abdsc #BigData #MachineLearning #Coding https://t.co...">
[6]
[1] "Ashot_: RT @KirkDBorne: How #Python fares as a #DataScience language: https://t.co/dj8BcmCTkg #abdsc #BigData #MachineLearning #Coding https://t.co...">
```

Notre variable est une liste, ce que nous voudrons faire est de la convertir en un vecteur caractère afin de pouvoir créer un corpus (Ensemble fini de textes choisi comme base de notre étude). La dernière ligne montre qu'après l'extraction de texte on a obtenu un vecteur de 500 éléments.

```
> class(DataScience)
[1] "list"
> DataScience_text <- sapply(DataScience, function(x) x$getText())
> class(DataScience_text)
[1] "character"
> str(DataScience_text)
chr [1:500] "RT @sesync: @sesync is hiring! Looking for a Data Scientist and Science Communications Coordinator to join the " | __truncated__ ...
```

Ici, nous sommes en mesure de pouvoir créer notre corpus (ligne 1), de vérifier le type de la variable (ligne 2) et d'inspecter le contenu d'une ligne donnée, en précisant son numéro à l'intérieur de la fonction inspect (ligne 3).

```
> DataScience_corpus <- Corpus(VectorSource(DataScience_text))
> DataScience_corpus
<<SimpleCorpus>>
Metadata: corpus specific: 1, document level (indexed): 0
Content: documents: 500
> inspect(DataScience_corpus[500])
<<SimpleCorpus>>
Metadata: corpus specific: 1, document level (indexed): 0
Content: documents: 1
[1] RT @JOMadeke: Having a background in tech (coding, computer science, data analytics) should not determine whether or not a female founder c...
```

A l'étape passionnante de nettoyage des données, nous commençons par enlever toute sorte de ponctuation existante, ensuite nous enlevons les espaces et URL et écrivons tous les caractères du texte en minuscules et puis nous supprimons les mots vides (non significatifs), les nombres et espaces vides.

```
> DataScience_clean <- tm_map(DataScience_corpus, removePunctuation)
> toSpace <- content_transformer(function(x, pattern) {return (gsub(pattern, " ",x))})
> DataScience_clean<- tm_map(DataScience_clean,toSpace,"[^[:graph:]]") 
> DataScience_clean <- tm_map(DataScience_clean, content_transformer(tolower))
> DataScience_clean <- tm_map(DataScience_clean, removeWords, stopwords("english"))
> DataScience_clean <- tm_map(DataScience_clean, removeNumbers)
> DataScience_clean <- tm_map(DataScience_clean, stripWhitespace)
```

Nous faisons appel à la fonction wordcloud et y passons quelques paramètres (nombre maximal de mots et couleurs...)



Nous allons passer au "Text mining", en prenant comme support de text, l'article qui aborde le même thème de ce projet dont le titre : " Comment la Data Science conduit à une meilleure prise de décision" publié dans ce site : <https://towardsdatascience.com/how-data-science-leads-to-better-decision-making-2de2116e586e>. Nous créons notre fichier 'pl.txt' qui contient l'ensemble de l'article.

```

pl.txt - Bloc-notes
Fichier Edition Format Affichage ?
How Data Science leads to better decision-making

This week's news in the US, in Europe, and in the Middle East is a vivid reminder of the fallacy of rational decision-making. Whether we are reading about business, ec
We live in a time and space in which the which data is constantly mistaken for facts. Taking better decisions, rather than crunching the data, is the ultimate benchmark
What does improving decision-making entail? In decision science, we learn that the major challenges to effective management are the perceptions of the complexity, a
What is a better decision? In line with David Snowden's work on sense-making[v], we believe there is a clear distinction between good, better, and great decisions. Go
Although machine learning is currently marketed to management as a mystical elixir, it's nothing more than a technological tool used to explore the nature of the proble
How can the study of data science help us become better decision-makers? Business analytics is a four-step process designed to help people make better decisions
The practice of data science is the heart and soul of the Business Analytics Institute. The 2018 BAI Summer Program will explore the mission-critical skills in leveragin

```

La fonction `readLines()` considère chaque ligne comme étant séparée du reste (après lecture d'une ligne, elle l'affiche et si elle trouve une ligne vide elle retourne " " dans une ligne et passe à la suivante). Pour voir la structure de ce qu'elle retourne dans sa totalité , nous nous servons de `str()` qui nous dit que c'est un vecteur de caractères de 15 éléments et chaque élément est une ligne séparée dans le texte original.

```

> str(readLines('pl.txt'))
chr [1:15] "How Data Science Leads to better decision-making" "" ...

```

Cependant, nous voulons que le texte soit considéré comme une seule entité, chose possible grâce à la fonction `paste` et son attribut "collapse" qui regroupe toutes les lignes et met un espace entre elles.

```

> text <- paste(readLines('pl.txt'), collapse = " ")
> text
[1] "How Data Science leads to better decision-making. This week's news in the US, in Europe, and in the Middle East is a vivid reminder of the fallacy of rational decision-making. Whether we are reading about business, economics or society, each day seems to bring its load of conspicuously poor decision-making. Jack Zenger and Joseph Folkman have outlined several reasons why decision-makers fail including negligence, lack of anticipation, indecisiveness, and isolation. [i]. Are fake news, faked facts, and manipulated opinions the cause or the result of poor decisions? [ii] Most importantly, what can be done to improve our decision-making skills for our organizations, our customers, and our careers? We live in a time and space in which the which data is constantly mistaken for facts. Taking better decisions, rather than crunching the data, is the ultimate benchmark for improving management. We are currently producing roughly 2.5 quintillion bytes of data each day? more data in th... <truncated>
> gsub(pattern = "\\w", replace= " ", text)
[1] "How Data Science leads to better decision making This week s news in the US in Europe and in the Middle East is a vivid reminder of the fallacy of rational decision making Whether we are reading about business economics or society each day seems to bring its load of conspicuously poor decision making Jack Zenger and Joseph Folkman have outlined several reasons why decision makers fail including negligence lack of anticipation indecisiveness and isolation i Are fake news faked facts and manipulated opinions the cause or the result of poor decisions ii Most importantly what can be done to improve our decision making skills for our organizations our customers and our careers we live in a time and space in which the which data is constantly mistaken for facts taking better decisions rather than crunching the data is the ultimate benchmark for improving management we are currently producing roughly 2 5 quintillion bytes of data each day more data in th... <truncated>

```

La fonction `gsub` avec a pour rôle de supprimer la ponctuation. En effet, dans la syntaxe des expressions régulières en R, lorsqu'on écrit "`w`" c'est pour considérer un mot (word), mais "`W`" c'est pour tout ce qui diffère du "mot" dans le texte (Not word). Encore une fois, nous changeons tous les caractères pour qu'ils deviennent en minuscules.

```

> tolower(text2)
[1] "how data science leads to better decision making this week s news in the us in europe and in the middle east is a vivid reminder of the fallacy of rational decision making whether we are reading about business economics or society each day seems to bring its load of conspicuously poor decision making jack zenger and joseph folkman have outlined several reasons why decision makers fail including negligence lack of anticipation indecisiveness and isolation i are fake news faked facts and manipulated opinions the cause or the result of poor decisions ii most importantly what can be done to improve our decision making skills for our organizations our customers and our careers we live in a time and space in which the which data is constantly mistaken for facts taking better decisions rather than crunching the data is the ultimate benchmark for improving management we are currently producing roughly 2 5 quintillion bytes of data each day more data in th... <truncated>
> text3 <- tolower(text2)

```

Nous supprimons tous les mots de liaison, de possession etc du texte car ils ne sont pas porteur de sentiments. En effet, en recherche d'information, un mot vide (ou stop word, en anglais) est un mot qui est tellement commun qu'il est inutile de l'indexer ou de l'utiliser dans une recherche. Un mot dont la distribution est uniforme sur les textes de la collection est dit « vide ». En d'autres termes, un mot qui apparaît avec une fréquence semblable dans chacun des textes de la collection n'est pas discriminant, ne permet pas de distinguer les textes les uns par rapport aux autres.

```
> stopwords()
[1] "i"      "me"     "my"     "myself"  "we"      "our"    "ours"   "ourselves" "you"
[10] "your"   "yours"   "yourself" "yourselves" "he"      "him"    "his"    "himself"   "she"
[19] "her"     "hers"    "herself"  "itself"    "it"      "its"    "itself"  "they"     "them"
[28] "theirs"  "themselves" "what"    "which"    "who"     "whom"   "this"    "that"     "these"
[37] "those"   "am"      "is"      "are"      "was"     "were"   "be"     "been"     "being"
[46] "have"    "has"     "had"     "having"   "do"      "does"   "did"    "doing"    "would"
[55] "should"  "could"   "ought"   "i'm"      "you're"  "he's"   "she's"   "it's"     "we're"
[64] "they're" "i've"    "you've"  "we've"    "they've" "i'd"    "you'd"   "he'd"     "she'd"
[73] "we'd"    "they'd"  "i'll"    "you'll"   "he'll"   "she'll"  "we'll"   "they'll"  "isn't"
[82] "aren't"  "wasn't"  "weren't" "hasn't"   "haven't" "hadn't" "doesn't" "don't"    "didn't"
[91] "won't"   "wouldn't" "shan't"  "shouldn't" "can't"   "cannot" "couldn't" "mustn't"  "let's"
[100] "that's"  "who's"   "what's"  "here's"   "there's" "when's" "where's" "why's"   "how's"
[109] "a"       "an"      "the"     "and"     "but"    "if"     "or"     "because"  "as"
[118] "until"   "while"   "of"      "at"      "by"     "for"    "with"   "about"    "against"
[127] "between" "into"   "through" "during"   "before"  "after"   "above"   "below"    "to"
[136] "from"    "up"     "down"    "in"      "out"    "on"     "off"    "over"    "under"
[145] "again"   "further" "then"    "once"    "here"   "there"   "when"   "where"   "why"
[154] "how"     "all"    "any"    "both"    "each"   "few"    "more"   "most"    "other"
[163] "some"   "such"   "no"     "nor"    "not"    "only"   "own"    "same"   "so"
[172] "than"   "too"    "very"   ""        ""       ""       ""       ""       ""
> removeWords(text3, stopwords())
[1] " data science leads better decision making week s news us europe middle east vivid reminder fallacy rational decision making whether reading business economics society day seems bring load conspicuously poor decision making jack zenger jose ph folkman outlined several reasons decision makers fail including negligence lack anticipation indecisiveness isolation fake news faked facts manipulated opinions cause result poor decisions ii importantly can done improve decision making skills organizations customers careers live time space data constantly mistaken facts taking better decisions rather crunc hing data ultimate benchmark improving management currently producing roughly 2 5 quintillion bytes data day data last two years previous history mankind iii klaus schwab suggests entered fourth industrial revolution value defined ability cap ture analyze vast amount data ... <truncated>
```

Nous remarquons l'existence de lettres seules comme 's', donc nous executons la commande qui demande de supprimer les mots commençons par une lettre de l'alphabet et se terminant par une lettre [a-Z] mais qui sont de longueur égale à UN.

```
> removeWords(text3, stopwords())
[1] " data science leads better decision making week s news us europe middle east vivid
reminder fallacy rational decision making whether reading business economics society day se
ems bring load conspicuously poor decision making jack zenger joseph folkman outlined several r
easons decision makers fail including negligence lack anticipation indecisiveness isolation
fake news faked facts manipulated opinions cause result poor decisions ii importantly c
an done improve decision making skills organizations customers careers live time space
data constantly mistaken facts taking better decisions rather crunching data ultimate b
enchmark improving management currently producing roughly 2 5 quintillion bytes data day dat
a last two years previous history mankind iii klaus schwab suggests entered fourth indus
trial revolution value defined ability capture analyze vast amount data ... <truncated>
> text3 <- removeWords(text3, stopwords())
> gsub(pattern="\b[A-z]\b{1}", replace= " ", text3)
[1] " data science leads better decision making week news us europe middle east vivid
reminder fallacy rational decision making whether reading business economics society day se
ems bring load conspicuously poor decision making jack zenger joseph folkman outlined several r
easons decision makers fail including negligence lack anticipation indecisiveness isolation
fake news faked facts manipulated opinions cause result poor decisions ii importantly c
an done improve decision making skills organizations customers careers live time space
data constantly mistaken facts taking better decisions rather crunching data ultimate b
enchmark improving management currently producing roughly 2 5 quintillion bytes data day dat
a last two years previous history mankind iii klaus schwab suggests entered fourth indus
trial revolution value defined ability capture analyze vast amount data ... <truncated>
```

Ainsi, nous pouvons comparer entre le texte original et celui obtenu après la phase du nettoyage.

```
> text3 <- gsub(pattern="\b[A-z]\b{1}", replace= " ", text3)
> text3 <- stripWhitespace(text3)
> text3
[1] " data science leads better decision making week news us europe middle east vivid reminder fallacy
rational decision making whether reading business economics society day seems bring load conspicuously
poor decision making jack zenger joseph folkman outlined several reasons decision makers fail includ
ing negligence lack anticipation indecisiveness isolation fake news faked facts manipulated opinions c
ause result poor decisions ii importantly can done improve decision making skills organizations custom
ers careers live time space data constantly mistaken facts taking better decisions rather crunching da
ta ultimate benchmark improving management currently producing roughly 2 5 quintillion bytes data day
data last two years previous history mankind iii klaus schwab suggests entered fourth industrial revol
ution value defined ability capture analyze vast amount data iv date little evidence revolution led be
tter decisions past data science transforming data impactful action address fundam... <truncated>
> text
[1] "How Data Science leads to better decision-making. This week's news in the US, in Europe, and in
the Middle East is a vivid reminder of the fallacy of rational decision-making. Whether we are reading
about business, economics or society, each day seems to bring its load of conspicuously poor decision
-making. Jack Zenger and Joseph Folkman have outlined several reasons why decision-makers fail includ
ing negligence, lack of anticipation, indecisiveness, and isolation. [i]. Are fake news, faked facts,
and manipulated opinions the cause or the result of poor decisions? [ii] Most importantly, what can be
done to improve our decision-making skills for our organizations, our customers, and our careers? We
live in a time and space in which the which data is constantly mistaken for facts. Taking better dec
isions, rather than crunching the data, is the ultimate benchmark for improving management. We are cu
rrently producing roughly 2.5 quintillion bytes of data each day?-?more data in th... <truncated>
```

Afin d'analyser la subjectivité du texte il faut ne garder que les mots significatifs, c'est la raison pour laquelle construit un tableau des mots importants figurant dans le texte.

```
> str_split(text3, pattern="\s+")
[[1]]
[1] ""          "data"      "science"   "leads"     "better"    "decision"  "making"
[8] "week"      "news"      "us"        "europe"    "middle"    "east"      "vivid"
[15] "reminder"  "fallacy"  "rational"  "decision"  "making"    "whether"   "reading"
[22] "business"  "economics" "society"   "day"       "seems"    "bring"     "load"
[29] "conspicuously" "poor"    "decision"  "making"    "jack"      "zenger"   "joseph"
[36] "folkman"   "outlined" "several"   "reasons"   "decision"  "makers"   "fail"
[43] "including" "negligence" "lack"     "anticipation" "indecisiveness" "isolation" "fake"
[50] "news"      "faked"    "facts"    "manipulated" "opinions"  "cause"     "result"
[57] "poor"      "decisions" "ii"       "importantly" "can"      "done"      "improve"
[64] "decision"  "making"   "skills"    "organizations" "customers" "careers"   "live"
[71] "time"      "space"    "data"     "constantly"  "mistaken" "facts"     "taking"
[78] "better"    "decisions" "rather"   "crunching" "data"     "ultimate"  "benchmark"
[85] "improving" "management" "currently" "producing" "roughly"  "2"        "5"
[92] "quintillion" "bytes"   "data"     "day"       "data"     "last"     "two"
[99] "years"     "previous" "history"  "mankind"  "iii"      "klaus"    "schwab"
[106] "suggests"  "entered"  "fourth"   "industrial" "revolution" "value"    "defined"
[113] "ability"   "capture"  "analyze"  "vast"     "amount"   "data"    "iv"
[120] "date"      "little"   "evidence" "revolution" "led"      "better"   "decisions"
[127] "past"      "data"     "science"  "transforming" "data"     "impactful" "action"
[134] "address"   "fundamental" "organizational" "challenges" "improving" "decision"  "making"
[141] "entail"    "decision" "science"  "learn"     "major"    "challenges" "effective"
[148] "management" "perceptions" "complexity" "ambiguity" "uncertainty" "environment" "take"
[155] "decisions"  "cognitive" "sciences" "taught"   "pre"      "conceptions" "prejudices"
[162] "distort"   "see"      "problem" "bound"    "ability"  "propose"   "innovative"
[169] "solutions" "management" "schools" "trained"  "recognize" "complexity" "real"
[176] "world"     "problems" "today"    "defy"     "logic"    "one"      "best"
[183] "way"       "finally"  "business" "sense"   "culprit" "isn"      "just"
[190] "decisions" "often"    "taken"    "around"  "us"      "better"   "decision"
```

Nous avons téléchargé deux documents, l'un comportant tous les mots positifs dans la langue anglaise et l'autre tous les mots négatifs.

- Fichier des mots positifs : <http://ptrckprry.com/course/ssd/data/positive-words.txt>
- Fichier des mots négatifs : <http://ptrckprry.com/course/ssd/data/negative-words.txt>

A l'aide de la fonction "match" nous vérifions l'existence des mots positifs (puis négatifs) -issus des deux fichiers- dans notre texte original. Pour un mot positif ou négatif donné, nous obtenons soit la position du mot dans le texte du départ ou bien "NA" càd la non-existence de ce mot. Nous comptons ensuite le nombre de mots positifs (en négligeant l'intensité du mot et en considérant dans ce cas simple que tous les mots sont de même poids, càd de même importance). On soustrait le nombre de mots négatifs au nombre de mots positifs et on se retrouve avec un résultat positif de 15 ce qui signifie que le document a abordé le thème de "Comment la Data Science conduit à une meilleure prise de décision" d'une manière positive.

```
> match(textbag, poswords)
 [1] 2007 NA NA 1075 200 NA 1926
[15] NA NA 1420 NA NA
[29] NA NA
[43] NA NA
[57] NA NA NA NA NA NA 971 NA NA NA NA NA NA NA NA NA NA
[71] NA NA NA NA NA NA NA NA 200 NA NA NA NA NA NA NA NA
[85] 976 NA NA
[99] NA NA
[113] NA 1077 200 NA
[127] NA 976 NA NA
[141] NA NA NA NA NA NA 490 NA NA NA NA NA NA NA NA NA NA
[155] NA 998
[169] NA 196
[183] NA 200 NA
[197] NA NA NA 1987 NA NA NA 294 445 832 200 857 NA 832
[211] NA NA NA NA NA 1533 NA NA NA NA NA NA NA NA NA NA
[225] NA NA NA NA 1533 NA NA 156 NA 200 NA NA NA NA NA
[239] NA NA NA NA NA 857 NA NA NA NA NA NA NA NA NA NA
[253] NA NA
[267] NA NA
[281] NA 1533 NA
[295] NA NA
[309] NA 1476 NA NA
[323] NA NA
[337] NA 200 NA 200
[351] NA NA 1987 NA NA
[365] NA NA NA NA NA 1987 NA NA NA NA 365 NA NA NA NA
[379] NA NA
```

```
> sum(!is.na(match(textbag, poswords)))
[1] 40
> sum(!is.na(match(textbag, negwords)))
[1] 25
```

```
> sum(!is.na(match(textbag, poswords)))
[1] 40
> sum(!is.na(match(textbag, negwords)))
[1] 25
> score <- sum(!is.na(match(textbag, poswords))) - sum(!is.na(match(textbag, negwords)))
> score
[1] 15
```

Représentation visuelle des mots-clefs (tags) les plus utilisés dans notre texte. C'est une sorte de condensé sémantique de notre document dans lequel les concepts clefs évoqués sont dotés d'une unité de taille (dans le sens du poids de la typographie utilisée) permettant de faire ressortir leur importance.



2.1.4 Applications en Python

Pour le second modèle d'analyse de sentiments sur Twitter, nous avons choisi un jeu de données de 1.600.000 tweets étiquetés positifs ou négatifs. L'ensemble de données pour la formation du modèle provient de "Sentiment140", un ensemble de données provenant de l'Université de Stanford. Plus d'informations sur l'ensemble de données peuvent être trouvées à partir du lien ci-dessous.
<http://help.sentiment140.com/for-students/>

L'ensemble de données peut être téléchargé à partir du lien ci-dessous.
[http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip\](http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip)

```
In [ ]: import warnings  
        warnings.filterwarnings('ignore') #nous évitons l'affichage des alertes
```

```
In [3]: import pandas as pd  
        import numpy as np  
        import matplotlib.pyplot as plt  
        plt.style.use('fivethirtyeight')
```

#Ceci montre un exemple du style "fivethirtyeight", qui essaie de reproduire les styles de FiveThirtyEight.com.

Tout l'enjeu du nettoyage est là : évacuer ce qui peut constituer du bruit, tout en conservant l'information pertinente. Dans le cas présent, le sens est bien conservé. Pendant le processus de nettoyage, les mots de négation sont divisés en deux parties, et le 't' après l'apostrophe disparaît lorsque nous filtrons les jetons de longueur supérieure à une syllabe. Cela fait que des mots comme «can't» finissent par être aussi semblables que «can». Cela ne semble pas une question triviale pour l'analyse des sentiments. Le deuxième problème que nous avons rencontré est que certains liens URL ne commencent pas par "http", parfois les gens collent un lien dans le formulaire "www.aaaa.com". Cela n'a pas été correctement géré lorsque nous avons défini le modèle d'adresse URL comme 'https? : // [A-Za-z0-9./]+'. Et un autre problème avec ce modèle regex est qu'il ne détecte que l'alphabet, nombre, période, barre oblique. Cela signifie qu'il échouera à attraper la partie de l'URL s'il contient un autre caractère spécial tel que " = ", " ", "e", etc. Le troisième problème est le modèle regex pour Twitter ID. Dans la fonction de nettoyage précédente, nous l'avons défini comme '@ [A-Za-z0-9] +', mais avec un peu de recherche sur google, nous avons découvert que l'ID de twitter permet également le symbole de soulignement car un caractère peut être utilisé avec ID. À l'exception du symbole de soulignement, seuls les caractères autorisés sont des alphabets et des chiffres. Ci-dessous, la fonction de nettoyage de données mise à jour.

L'ordre du nettoyage est le suivant :

1. Suppression des indicateurs d'ordre des octets ou BOM (pour l'anglais byte order mark) est une donnée qui indique l'utilisation d'un encodage unicode ainsi que l'ordre des octets, généralement situé au début de certains fichiers texte,
2. Suppression d'adresse url (modèle 'http :'),
3. Suppression de l'identifiant Twitter,
4. Suppression d'adresse url (modèle 'www.'),
5. Ecrire tout en minuscule,
6. La manipulation de la négation,
7. Supprimer des chiffres et des caractères spéciaux,
8. Tokenizing et jointure.

Avant cela, nous allons voir un exemple simple pour l'extraction de noms de pages/personnes, d'URLs, et de Hashtags :

```
In [1]: import twitter_text

# Exemple simple

txt = "RT @SocialWebMining Mining 1M+ Tweets About
#Syria http://wp.me/p3QiJd-1I"

ex = twitter_text.Extractor(txt)

print ("Screen Names:", ex.extract_mentioned_screen_names_with_indices())
print ("URLs:", ex.extract_urls_with_indices())
print ("Hashtags:", ex.extract_hashtags_with_indices())

Screen Names: [{'screen_name': 'SocialWebMining', 'indices': [3, 19]}]
URLs: [{'url': 'http://wp.me/p3QiJd-1I', 'indices': [51, 73]}]
Hashtags: [{'hashtag': 'Syria', 'indices': [44, 50]}]
```

```
In [3]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
plt.style.use('fivethirtyeight')
```

```
In [ ]: #je vais d'abord définir la fonction de nettoyage des données,
#qui sera appliquée à l'ensemble de données, (1.600.000 Tweet)
# à travers la Tokenization, stemming (racinisation ou désuffixation)/
#lemmatisation(désigne l'analyse lexicale du contenu d'un texte regroupant
```

```

# les mots d'une même famille.), mots vides et URLs.

import re
from bs4 import BeautifulSoup
from nltk.tokenize import WordPunctTokenizer
tok = WordPunctTokenizer()

pat1 = r'@[A-Za-z0-9_]+'
pat2 = r'https?://[^ ]+'
combined_pat = r'|'.join((pat1, pat2))
www_pat = r'www.[^ ]+'
negations_dic =
{"isn't":"is not", "aren't":"are not", "wasn't":"was not", "",
 "haven't":"have not", "hasn't":"has not", "hadn't":"had not", "",
 "wouldn't":"would not", "don't":"do not", "doesn't":"does not", "",
 "can't":"can not", "couldn't":"could not", "shouldn't":"should not",
 "mustn't":"must not"}
neg_pattern = re.compile(r'\b(' + '|'.join(negations_dic.keys()) + r')\b')

def tweet_cleaner_updated(text):
    soup = BeautifulSoup(text, 'lxml')
    souped = soup.get_text()
    try:
        bom_removed = souped.decode("utf-8-sig").replace(u"\ufffd", "?")
    except:
        bom_removed = souped
    stripped = re.sub(combined_pat, '', bom_removed)
    stripped = re.sub(www_pat, '', stripped)
    lower_case = stripped.lower()
    neg_handled = neg_pattern.sub
    (lambda x: negations_dic[x.group()], lower_case)
    letters_only = re.sub("[^a-zA-Z]", " ", neg_handled)
    #Pendant le processus letters_only deux lignes ci-dessus,
    # nous avons créé des espaces blancs inutiles, nous allons donc marquer
    # et joindre ensemble pour supprimer les espaces blancs inutiles
    words = [x for x in tok.tokenize(letters_only) if len(x) > 1]
    return (" ".join(words)).strip()

```

```
In [6]: cols = ['sentiment', 'id', 'date', 'query_string', 'user', 'text']
df = pd.read_csv("Jeu de données", encoding = "ISO-8859-1",
header=None, names=cols)
df['sentiment'] = df['sentiment'].map({0: 0, 4: 1})
df.head()
```

```
#J'ai d'abord commencé par abandonner les colonnes dont je n'ai pas
#besoin dans le but spécifique de l'analyse des sentiments.

#La colonne "id" est un identifiant unique pour chaque tweet
#La colonne "date" est pour l'information de la date de première
#publication du tweet
#La colonne "query_string" indique si le tweet a été collecté avec un mo
-clé de requête particulier,
#mais pour cette colonne, 100% des entrées ont la valeur "NO_QUERY"
#La colonne "utilisateur" est le nom du pseudonyme Twitter
# de l'utilisateur qui a tweet
```

Out[6] :

	sentiment	id	date	query_string	user	text
0	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scothamilton	is upset that he can't update his Facebook by ...
2	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...
3	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
4	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....

```
In [ ]: print ("Cleaning the tweets...\n")
clean_tweet_texts = []
for i in xrange(0,len(df)):
    if( (i+1)%100000 == 0 ):
        print ("Tweets %d of %d has been processed" % ( i+1, len(df) ))
    clean_tweet_texts.append(tweet_cleaner_updated(df['text'][i]))
```

```
In [47]: csv='Chemin vers mon nouveau jeu de données, après nettoyage
my_df = pd.read_csv(csv,index_col=0)
my_df.tail(19)
#pour visualiser les derniers éléments du jeu de données après nettoyage'
```

Out[47] :

		text	target
1599981		another commenting contest yay	1
1599982		figured out how to see my tweets and facebook ...	1
1599983		theri tomorrow drinking coffee talking about o...	1
1599984		you heard it here first we re having girl hope...	1
1599985		if ur the lead singer in band beware falling p...	1
1599986		too much ads on my blog	1
1599987		neveer think that you both will get on well wi...	1
1599988		ha good job that right we gotta throw that big...	1
1599989		im glad ur doing well	1

1599990	wooooo xbox is back	1
1599991	mmmm that sounds absolutely perfect but my sch...	1
1599992	recovering from the long weekend	1
1599993	NaN	1
1599994	yeah that does work better than just waiting f...	1
1599995	just woke up having no school is the best feel...	1
1599996	thewdb com very cool to hear old walt interviews	1
1599997	are you ready for your mojo makeover ask me fo...	1
1599998	happy th birthday to my boo of alll time tupac...	1
1599999	happy charitytuesday	1

Ici je crée une fonction "getLine" qui va dans le deuxième colonne (j=1, puisque python commence par 0) ensuite je prends le nombre de lignes que je souhaite (exemple : 16 premiers tweets dans le corpus)

```
In [485]: import tweepy
import numpy as np
from io import BytesIO
import textblob.exceptions
from textblob import TextBlob
import csv
fichier='Chemin vers mon nouveau jeu de données, après nettoyage'
def getLine(fichier, i, j):
    with open(fichier, 'r') as f:
        a=i
        head = [next(f) for x in range(0,a+1) ]
        reader = csv.reader(head)
        for line in reader:
            for i in range(1, reader.line_num):
                print((line[j]))
                break
```

```
In [486]: getLine(file,16,1)
```

awww that bummer you shoulda got david carr of third day to do it
is upset that he can not update his facebook by texting it and might
cry as result school today
dived many times for the ball managed to save the rest go out of bounds
my whole body feels itchy and like its on fire
no it not behaving at all mad why am here because can not see you
all over there
not the whole crew
need hug

```

hey long time no see yes rains bit only bit lol fine thanks how you
nope they did not have it
que me muera
spring break in plain city it snowing
just re pierced my ears
could not bear to watch it and thought the ua loss was embarrassing
it it counts idk why did either you never talk to me anymore
would ve been the first but did not have gun not really though
zac snyder just doucheclown
wish got to watch it with you miss you and how was the premiere

```

```

In [ ]: #Faire un test sur 3 phrases (étudier leur subjectivité,
         # les traduire et faire un split de chacune d'elles)
         # que je généraliserai sur les 18 tweets issus "clean_tweet2.csv"

```

```

In [175]: Data=["I'm happy","Je suis content,"
           "With all of the failed experts weighing in,
           "Does anybody really believe that talks and dialogue would be going on"]
           for i,text in enumerate(Data):
               blob = TextBlob(text)
               print(blob.detect_language())
               print(u"Polarity {}, Subjectivity {}".format(blob.sentiment.polarity,
               blob.sentiment.subjectivity))
               try:
                   print(u"French : {}".format(blob.translate
                   (from_lang="en-US", to='fr')))
               except textblob.exceptions.NotTranslated:
                   pass
               # end try
               print(u"Tokens : {}".format(blob.words))

en
Polarity 0.8, Subjectivity 1.0
French : je suis heureux
Tokens : ['I', "'m", 'happy']
fr
Polarity 0.0, Subjectivity 0.0
Tokens : ['Je', 'suis', 'content']
en
Polarity -0.15, Subjectivity 0.25
French : Avec tous les "experts" ratés, personne ne croit vraiment que
les pourparlers et le dialogue se poursuivra ... https://t.co/EfxwZcRmLZ
Tokens : ['With', 'all', 'of', 'the', 'failed', "'", 'experts', "'", 'weighing',
'in', 'does', 'anybody', 'really', 'believe', 'that', 'talks', 'and', 'dialogue',
'would', 'be', 'going', 'on', 'https', 't.co/EfxwZcRmLZ']

```

```
In [7]: import tweepy
import numpy as np
from io import BytesIO
import textblob.exceptions
from textblob import TextBlob
import csv
fichier='Chemin vers mon nouveau jeu de données'
def analyse_sentiments(fichier,i,j):
    with open(fichier,'r') as f:
        tab=[]
        a=i
        head = [next(f) for x in range(0,a+1) ]
        reader = csv.reader(head)
        for line in reader:
            for i in range(1, reader.line_num):
                tab[i]=tab.append(line[j])
                break

        for i,text in enumerate(tab):
            blob = TextBlob(text)
            print(blob.detect_language())
            print(u"Polarity {}, Subjectivity {}".format(blob.sentiment.polarity,
            blob.sentiment.subjectivity))
            try:
                print(u"French : {}".format(blob.translate
                (from_lang="en-US",to='fr')))
            except textblob.exceptions.NotTranslated:
                pass

            print(u"Tokens : {}".format(blob.words))
```

In [8]: analyse_sentiments(fichier,18,1)

```
en
Polarity 0.2, Subjectivity 0.45
French : awww que bummer vous devriez avoir david carr du troisième jour
pour le faire
Tokens : ['awww', 'that', 'bummer', 'you', 'shoulda', 'got', 'david', 'carr',
'of', 'third', 'day', 'to', 'do', 'it']
en
Polarity 0.0, Subjectivity 0.0
French : est contrarié qu'il ne peut pas mettre à jour son facebook en
le textant et pourrait pleurer comme résultat école aujourd'hui
Tokens : ['is', 'upset', 'that', 'he', 'can', 'not', 'update', 'his',
```

'facebook', 'by', 'texting', 'it', 'and', 'might', 'cry', 'as',
'result', 'school', 'today']
en
Polarity 0.5, Subjectivity 0.5
French : plongé plusieurs fois pour la balle a réussi à sauver le reste
sortir des limites
Tokens : ['dived', 'many', 'times', 'for', 'the', 'ball', 'managed',
'to', 'save', 'the', 'rest', 'go', 'out', 'of', 'bounds']
en
Polarity 0.2, Subjectivity 0.4
French : mon corps entier me démange et comme son feu
Tokens : ['my', 'whole', 'body', 'feels', 'itchy', 'and', 'like', 'its',
'on', 'fire']
en
Polarity -0.625, Subjectivity 1.0
French : non, il ne se comporte pas du tout pourquoi je suis ici
parce que je ne peux pas vous voir partout
Tokens : ['no', 'it', 'not', 'behaving', 'at', 'all', 'mad', 'why',
'am', 'here', 'because', 'can', 'not', 'see', 'you', 'all', 'over', 'there']
en
Polarity 0.2, Subjectivity 0.4
French : pas tout l'équipage
Tokens : ['not', 'the', 'whole', 'crew']
en
Polarity 0.0, Subjectivity 0.0
French : besoin d'un câlin
Tokens : ['need', 'hug']
en
Polarity 0.273333333333333, Subjectivity 0.5599999999999999
French : hey depuis longtemps ne vois pas oui pleut peu seulement lol
bien merci comment vous
Tokens : ['hey', 'long', 'time', 'no', 'see', 'yes', 'rains', 'bit',
'only', 'bit', 'lol', 'fine', 'thanks', 'how', 'you']
en
Polarity 0.0, Subjectivity 0.0
French : Non, ils ne l'ont pas
Tokens : ['nope', 'they', 'did', 'not', 'have', 'it']
es
Polarity 0.0, Subjectivity 0.0
French : que moi muera
Tokens : ['que', 'me', 'muera']
en
Polarity -0.21428571428571427, Subjectivity 0.35714285714285715
French : vacances de printemps dans la plaine de la ville il neige
Tokens : ['spring', 'break', 'in', 'plain', 'city', 'it', 'snowing']
en

Polarity 0.0, Subjectivity 0.0
French : juste re percé mes oreilles
Tokens : ['just', 're', 'pierced', 'my', 'ears']
en

Polarity 0.0, Subjectivity 0.0
French : ne pouvait pas supporter de le regarder et pensait que la perte de l'ua était embarrassant
Tokens : ['could', 'not', 'bear', 'to', 'watch', 'it', 'and', 'thought', 'the', 'ua', 'loss', 'was', 'embarrassing']
en

Polarity 0.0, Subjectivity 0.0
French : il compte idk pourquoi avez-vous jamais plus me parler
Tokens : ['it', 'it', 'counts', 'idk', 'why', 'did', 'either', 'you', 'never', 'talk', 'to', 'me', 'anymore']
en

Polarity 0.075, Subjectivity 0.2666666666666666
French : aurait été le premier mais n'a pas d'arme à feu pas vraiment si zac snyder juste doucheclown
Tokens : ['would', 've', 'been', 'the', 'first', 'but', 'did', 'not', 'have', 'gun', 'not', 'really', 'though', 'zac', 'snyder', 'just', 'doucheclown']
en

Polarity 0.0, Subjectivity 0.0
French : souhaite avoir à le regarder avec vous manqué et comment était la première
Tokens : ['wish', 'got', 'to', 'watch', 'it', 'with', 'you', 'miss', 'you', 'and', 'how', 'was', 'the', 'premiere']
en

Polarity 0.0, Subjectivity 0.0
French : scène de la mort hollis me blessera sévèrement à regarder sur le film wry est coupés directeurs pas maintenant
Tokens : ['hollis', 'death', 'scene', 'will', 'hurt', 'me', 'severely', 'to', 'watch', 'on', 'film', 'wry', 'is', 'directors', 'cut', 'not', 'out', 'now', 'en']

Polarity 0.0, Subjectivity 0.0
French : sur le point de déposer des taxes
Tokens : ['about', 'to', 'file', 'taxes']

Les utilisateurs de Twitter à travers le monde envoient environ 350 000 nouveaux Tweets chaque minute, créant 6 000 informations de 140 caractères par seconde. Twitter est maintenant une ressource extrêmement précieuse à partir de laquelle nous pouvons extraire des informations en utilisant des outils d'exploration de texte tels que l'analyse des sentiments.

Au sein du bavardage social généré chaque seconde, il y a de grandes quantités d'informations extrêmement précieuses qui attendent d'être extraites. Grâce à l'analyse des sentiments, nous pouvons générer des réflexions sur les réactions

des consommateurs aux annonces, des opinions sur les produits ou les marques, et même suivre l'opinion sur les événements au fur et à mesure qu'ils se déroulent. Pour cette raison, nous entendrons souvent l'analyse des sentiments appelée «extraction d'opinion».

Dans cet esprit, nous avons décidé de mettre en place un outil utile basé sur un seul script Python pour aider l'UITS à démarrer l'exploration de l'opinion publique sur Twitter.

Qu'est-ce que le script fait? En utilisant ce script, l'UITS peut rassembler des Tweets avec l'API Twitter, analyser leur sentiment avec l'API AYLIEN Text Analysis, et visualiser les résultats avec matplotlib. Le script fournit également une visualisation et enregistre les résultats pour vous dans un fichier CSV afin de faciliter le rapport et l'analyse.

Les 4 objectifs de sa création pour l'UITS : 1-Comprendre la réaction du public aux nouvelles ou aux événements sur Twitter 2-Mesurer la voix de leurs clients et leurs opinions sur eux ou leurs concurrents 3-Générer des prospects en identifiant les mentions négatives de leurs concurrents 4-La chose intéressante dans le script est que l'UITS peut rechercher ce qu'elle aime et le script va exécuter leurs tweets à travers le même pipeline d'analyse, stocker les résultats dans un fichier CSV et afficher une visualisation

```
In [4]: import sys
import csv
import tweepy
import matplotlib.pyplot as plt

from collections import Counter
from aylienapiclient import textapi

if sys.version_info[0] < 3:
    input = raw_input

## Informations d'identification Twitter
consumer_key = "dLcHBP0Qo3IwggxQwFxMQ2Y3v"
consumer_secret = "hByRFKy0DZS1K7ZkUTMAv5fXTaBcfUnxpE0XY1JRnObYBnk7MI"
access_token = "3335730178-1rySRqRerpJ2VZEMeyuNohV5HFPhSZNqPzgR0vu"
access_token_secret = "381eC8zIvsn4MJGhC1ApLEvoelIYrlDQ0Sh9IP7x5oQex"

## Informations d'identification Alyen
application_id = "2dbbb66d"
application_key = "97f9a67569cde2c9281a98877628d25a"

## configurer une instance de Tweepy
```

```
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth)

## configure une instance de l'API Text AYLIEN
client = textapi.Client(application_id, application_key)

## recherche Twitter pour quelque chose qui vous intéresse
query = input ("Quel sujet voulez-vous analyser pour cet exemple? \n")
number = input ("Combien de Tweets voulez-vous analyser? \n")

results = api.search(
    lang="en",
    q=query + " -rt",
    count=number,
    result_type="recent"
)

print("--- Recherche des Tweets \n")

## ouvrir un fichier csv pour stocker les Tweets et leurs sentiments
file_name = 'Analyse_des_sentiments_dans_{}_Tweets_sur_{}.csv'
.format(number, query)

with open(file_name, 'w', newline='') as csvfile:
    csv_writer = csv.DictWriter(
        f=csvfile,
        fieldnames=["Tweet", "Sentiment"]
    )
    csv_writer.writeheader()

    print
    ("--- Ouvre un fichier CSV pour stocker les résultats de votre analyse")

## ranger les Tweets et les envoyer à l'API AYLIEN Text
for c, result in enumerate(results, start=1):
    tweet = result.text
    tidy_tweet = tweet.strip().encode('ascii', 'ignore')

    if len(tweet) == 0:
        print('Tweet vide')
        continue
```

```

        response = client.Sentiment({'text': tidy_tweet})
        csv_writer.writerow({
            'Tweet': response['text'],
            'Sentiment': response['polarity']
        })

        print("Nombre de Tweets analysés {}".format(c))

## compter les données dans la colonne "Sentiment" du fichier CSV
with open(file_name, 'r') as data:
    counter = Counter()
    for row in csv.DictReader(data):
        counter[row['Sentiment']] += 1

    positive = counter['positive']
    negative = counter['negative']
    neutral = counter['neutral']

## déclare les variables pour le camembert
colors = ['green', 'red', 'yellow']
sizes = [positive, negative, neutral]
labels = 'Positive', 'Negative', 'Neutral'

## on utilise matplotlib pour dessiner un graphe
plt.pie(
    x=sizes,
    shadow=True,
    colors=colors,
    labels=labels,
    startangle=90
)

plt.title("L'analyse des sentiments dans {} Tweets au sujet de {}"
          .format(number, query))
plt.show()

--- Recherche des Tweets

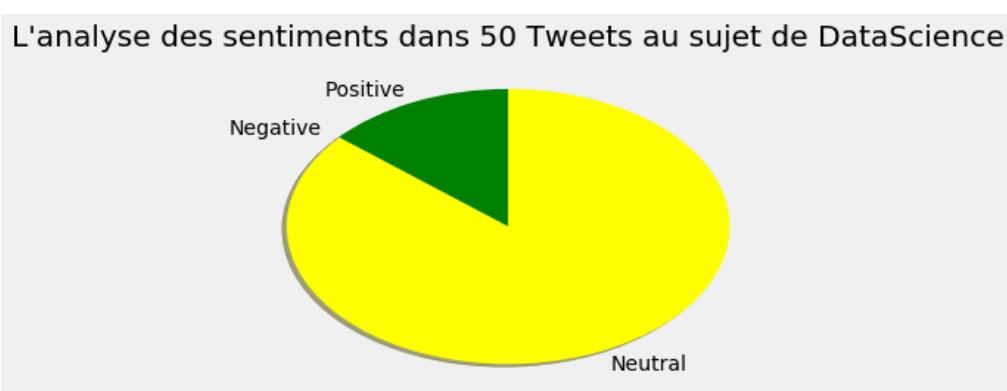
--- Ouvre un fichier CSV pour stocker les résultats de votre analyse ...

Nombre de Tweets analysés 1
Nombre de Tweets analysés 2
Nombre de Tweets analysés 3
Nombre de Tweets analysés 4

```

Nombre de Tweets analysés 5
Nombre de Tweets analysés 6
Nombre de Tweets analysés 7
Nombre de Tweets analysés 8
Nombre de Tweets analysés 9
Nombre de Tweets analysés 10
Nombre de Tweets analysés 11
Nombre de Tweets analysés 12
Nombre de Tweets analysés 13
Nombre de Tweets analysés 14
Nombre de Tweets analysés 15
Nombre de Tweets analysés 16
Nombre de Tweets analysés 17
Nombre de Tweets analysés 18
Nombre de Tweets analysés 19
Nombre de Tweets analysés 20
Nombre de Tweets analysés 21
Nombre de Tweets analysés 22
Nombre de Tweets analysés 23
Nombre de Tweets analysés 24
Nombre de Tweets analysés 25
Nombre de Tweets analysés 26

Nombre de Tweets analysés 50



«Plus que jamais les données sont nombreuses et nécessitent d'être «bien traitées» pour en tirer tout le potentiel décisionnel», souligne Thierry Vallaud, responsable data mining et décisionnel de Socio Logiciels. Il faut faire ressortir les marges d'erreur, les significativités, les biais éventuels. Avec la généralisation du numérique, l'imagerie produit des volumes de données de plus en plus importants. En parallèle, les capacités de calcul des ordinateurs autorisent des traitements d'une complexité toujours plus grande.

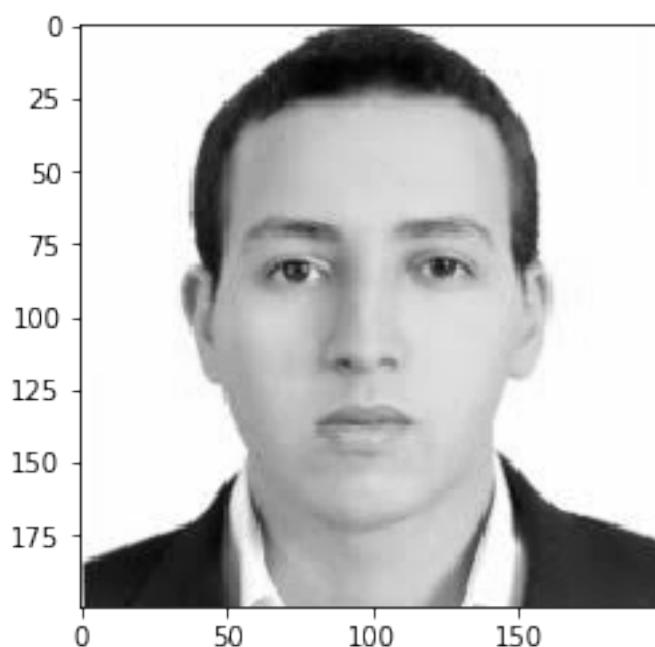
Pour en venir au sujet de la détection des visages et l'analyse des sentiments, voyons tout d'abord comment former un système pour détecter les visages. Si nous voulons construire un système d'apprentissage automatique, nous devons d'abord extraire les caractéristiques de toutes les images. Dans notre cas, les algorithmes d'apprentissage automatique utiliseront ces fonctionnalités pour apprendre à quoi ressemble un visage. Nous utilisons les fonctionnalités de Haar pour construire nos vecteurs de caractéristiques. Les fonctionnalités de Haar sont des sommes simples et des différences de correctifs à travers l'image. Nous faisons cela à plusieurs tailles d'image pour nous assurer que notre système est invariable à l'échelle.

OpenCV fournit un cadre de détection de visage agréable. Nous avons juste besoin de charger le fichier en cascade et de l'utiliser pour détecter les visages dans une image. Voyons voir comment le faire :

```
In [1]: import cv2
test1 = cv2.imread('moad2.jpg')
#convertir l'image en couleur en image grise car le détecteur de visage
#openCV attend des images grises
gray_img = cv2.cvtColor(test1, cv2.COLOR_BGR2GRAY)
```

```
In [3]: import matplotlib.pyplot as plt
plt.imshow(gray_img, cmap='gray')
```

```
Out[3]: <matplotlib.image.AxesImage at 0x19a4e97cf0>
```

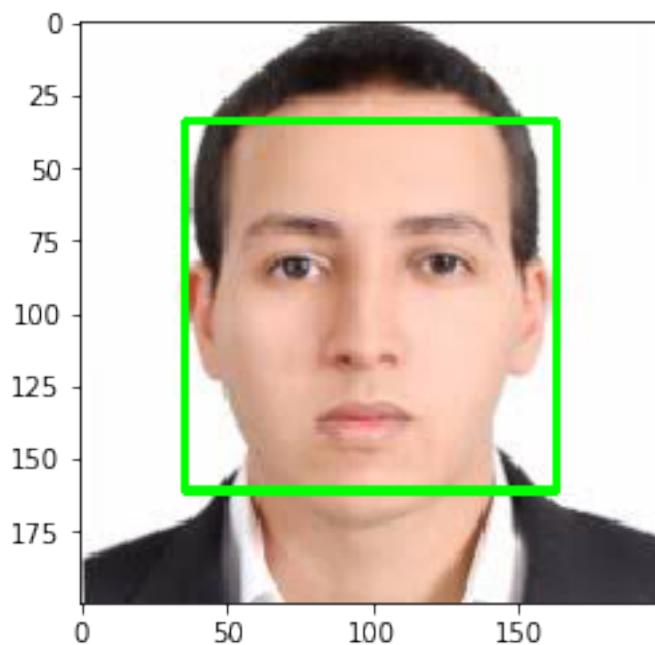


```
In [4]: 2 # charger le fichier d'entraînement du classifieur  
#pour la reconnaissance faciale  
haar_face_cascade = cv2.CascadeClassifier  
('haarcascade_frontalface_alt.xml')
```

```
In [5]: import numpy as np  
faces = haar_face_cascade.detectMultiScale(gray_img, scaleFactor=1.1  
, minNeighbors=5);  
test2=np.copy(test1)  
for (x, y, w, h) in faces:  
    cv2.rectangle(test2,(x, y),(x+w, y+h),(0, 255, 0), 2)  
plt.imshow(cv2.cvtColor(test2, cv2.COLOR_BGR2RGB))  
#Afficher le nombre de visages trouvées  
print('Faces found: ', len(faces))  
cv2.imwrite('test2.jpg',gray_img)
```

Faces found: 1

Out[5]: True

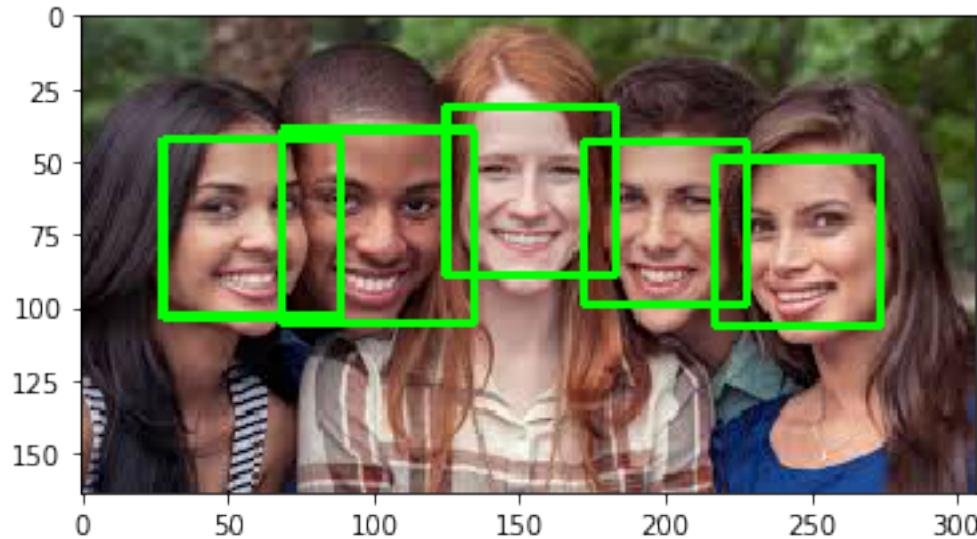


```
In [6]: test3 = cv2.imread('faces4.jpg')  
gray_img = cv2.cvtColor(test3, cv2.COLOR_BGR2GRAY)  
haar_face_cascade = cv2.CascadeClassifier  
('haarcascade_frontalface_alt.xml')
```

```
faces = haar_face_cascade.detectMultiScale(gray_img, scaleFactor=1.1,
minNeighbors=5);
test4=np.copy(test3)
for (x, y, w, h) in faces:
    cv2.rectangle(test4,(x, y),(x+w, y+h),(0, 255, 0), 2)
plt.imshow(cv2.cvtColor(test4, cv2.COLOR_BGR2RGB))
print('Faces found: ', len(faces))
cv2.imwrite('test4.jpg',gray_img)
```

Faces found: 5

Out[6]: True



Nous avons créé un mini-dataset d'images pour entraîner notre modèle : une vingtaine d'images pour chaque sentiments (Alors que pour une étude plus développée, nous devrions avoir plus d'un million d'images pour chaque sentiment.)



```
In [ ]: import cv2, glob, random, math, numpy as np, dlib
        from sklearn.svm import SVC

emotions = ["peur", "joie", "neutralité", "tristesse"]
#liste des sentiments
clahe = cv2.createCLAHE(clipLimit=2.0, tileSize=(8,8))
detector = dlib.get_frontal_face_detector()
predictor = dlib.shape_predictor("Moad_dataset.dat")
clf = SVC(kernel='linear', probability=True, tol=1e-3)

def get_files(emotion):
    files = glob.glob("dataset\\%s\\* %s" %emotion)
    random.shuffle(files)
    training = files[:int(len(files))]

    prediction=['moad2.jpg']
    #nous spécifions l'image que nous voulons analyser
    return training, prediction
```

```

def get_landmarks(image):
    detections = detector(image, 1)
    for k,d in enumerate(detections):
        #Pour toutes les instances de visage détectées individuellement
        shape = predictor(image, d)
        # Dessiner des repères faciaux avec la classe de prédicteur
        xlist = []
        ylist = []
        for i in range(1,68):
            xlist.append(float(shape.part(i).x))
            ylist.append(float(shape.part(i).y))

        xmean = np.mean(xlist)
        #Trouver les deux coordonnées du centre de gravité
        ymean = np.mean(ylist)
        xcentral = [(x-xmean) for x in xlist]
        #Calculer le centre de distance des autres
        #points dans les deux axes
        ycentral = [(y-ymean) for y in ylist]

        if xlist[26] == xlist[29]:
            #Si les coordonnées x de l'ensemble sont les mêmes, l'angle est 0,
            #distinguer ce cas pour empêcher l'erreur 'division par 0'
            anglenose = 0
        else:
            anglenose
            = int(math.atan((ylist[26]-ylist[29])/(xlist[26]-xlist[29]))*180
                  /math.pi)

            #point 29 est le bout du nez, le point 26 est le sommet du nez

        if anglenose < 0:
            # Obtenir un décalage en trouvant comment le nez de bridge
            # devrait être tourné pour devenir perpendiculaire
            # Obtenir un décalage en trouvant comment le nez de bridge
            # au plan horizontal
            anglenose += 90
        else:
            anglenose -= 90

    landmarks_vectorised = []
    for x, y, w, z in zip(xcentral, ycentral, xlist, ylist):
        landmarks_vectorised.append([x])
        #Ajouter les coordonnées relatives au centre de gravité
        landmarks_vectorised.append([y])
        # Obtenir la distance euclidienne entre chaque point

```

```

# et le point central (la longueur du vecteur)
meannp = np.asarray((ymean,xmean))
coornp = np.asarray((z,w))
dist = np.linalg.norm(coornp-meanp)

#Obtenir l'angle que le vecteur décrit par rapport a l'image,
#corrigé le décalage si le nez n'est pas parfaitement
# dans la bonne position
anglerelative
= (math.atan((z-ymean)/(w-xmean))
)*180/math.pi
- anglenose
landmarks_vectorised.append(dist)
landmarks_vectorised.append(anglerelative)

if len(detections) < 1:
    landmarks_vectorised = "error"
    #Si aucun visage n'est détecté, considérer cela comme une erreur
    #et le signaler à l'utilisateur.
return landmarks_vectorised

def make_sets():
    training_data = []
    training_labels = []
    prediction_data = []
    prediction_labels = []
    training = []
    prediction = []
    for emotion in emotions:
        training, prediction = get_files(emotion)
        #Annexe des données à la liste d'entraînement et de prédiction,
        #et génère des 7 étiquettes (entraînement sur 7 sentiments)
        for item in training:
            image = cv2.imread(item)
            gray = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
            clahe_image = clahe.apply(gray)
            landmarks_vectorised = get_landmarks(clahe_image)
            if landmarks_vectorised == "error":
                pass
            else:
                training_data.append(landmarks_vectorised)
                training_labels.append(emotions.index(emotion))

        for item in prediction:
            image = cv2.imread(item)
            gray = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)

```

```
clahe_image = clahe.apply(gray)
landmarks_vectorised = get_landmarks(clahe_image)
if landmarks_vectorised == "error":
    pass
else:
    prediction_data.append(landmarks_vectorised)
    #append le tableau d'images à la liste

prediction_labels.append(emotions.index(emotion))
#des données d'entraînement

return training_data, training_labels, prediction_data, prediction_labels
```



```
probam1 = np.zeros((4,10))
probam2 = np.zeros((1,4))

accur_lin = []

for i in range(0,10):
    print("Making sets %s" %i)
    # Créer des ensembles par échantillonnage aléatoire 80/20%
    training_data, training_labels, prediction_data, prediction_labels =
make_sets()

npar_train = np.array(training_data)
# Transformer l'ensemble d'apprentissage en un tableau numpy
# pour le classifieur
npar_trainlabs = np.array(training_labels)
print("training SVM linear %s" %i)
#Utilisation de la technique d'apprentissage
clf.fit(npar_train, training_labels) #Machine à vecteurs de support"

print("getting accuracies %s" %i)
# Utilisez la fonction score () pour obtenir la précision
npar_pred = np.array(prediction_data)
pred_lin = clf.score(npar_pred, prediction_labels)
print ("linear: ", pred_lin)
accur_lin.append(pred_lin) #Ajouter la précision dans une liste
proba=clf.predict_proba(prediction_data)
print ("proba: ", proba)
probam1[:,i]=proba[1,:]
probam2=proba[1,:]+probam2
```

```

proba=probam2/10
p1=round(proba[0,0],2)
p2=round(proba[0,1],2)
p3=round(proba[0,2],2)
p4=round(proba[0,3],2)
print("Mean value lin svm: %.3f" %np.mean(accur_lin))
# Obtenir la précision moyenne des 10 exécutions

frame=cv2.imread('moad2.jpg')

cv2.putText(frame, "Peur: {}".format(p1), (10, 30),
            cv2.FONT_HERSHEY_SIMPLEX, 0.7, (233, 236, 18), 2)
cv2.putText(frame, "Joie: {:.2f}".format(p2), (10, 60),
            cv2.FONT_HERSHEY_SIMPLEX, 0.7, (233, 236, 18), 2)
cv2.putText(frame, "Neutralité: {}".format(p3), (10, 90),
            cv2.FONT_HERSHEY_SIMPLEX, 0.7, (233, 236, 18), 2)
cv2.putText(frame, "Tristesse: {:.2f}".format(p4), (10, 120),
            cv2.FONT_HERSHEY_SIMPLEX, 0.7, (233, 236, 18), 2)

cv2.imshow("Frame", frame)
cv2.imwrite('resulat_neutralité_moad.png',frame)
cv2.waitKey(0)
cv2.destroyAllWindows()

```

```

In [5]: import matplotlib.pyplot as plt
        import matplotlib.image as mpimg
        from PIL import Image

def merge_images2(file1, file2, file3, file4):

    """Fusionner 4 images en une seule
    : paramètre file1: chemin vers le premier fichier image
    : paramètre file2: chemin vers le deuxième fichier image
    : paramètre file3: chemin vers le troisième fichier image
    : paramètre file4: chemin vers le quatrième fichier image
    : return: l'objet Image fusionné
    """

    image1 = Image.open(file1)
    image2 = Image.open(file2)
    image3 = Image.open(file3)
    image4 = Image.open(file4)

    (width1, height1) = image1.size

```

```
(width2, height2) = image2.size  
(width3, height3) = image3.size  
(width4, height4) = image4.size  
  
result_width = max(width1 , width2 , width3 , width4)  
result_height = height1 + height2 + height3 + height4  
  
result = Image.new('RGB', (result_width, result_height))  
result.paste(im=image1, box=(0,0))  
result.paste(im=image2, box=(0,height1))  
result.paste(im=image3, box=(0,height1+height2))  
result.paste(im=image4, box=(0,height1+height2+height3))  
return result
```

```
In [ ]: merge_images('resultat_joie.png','resultat_neutralité.png',  
'resultat_peur.png','resultat_tristesse.png')
```

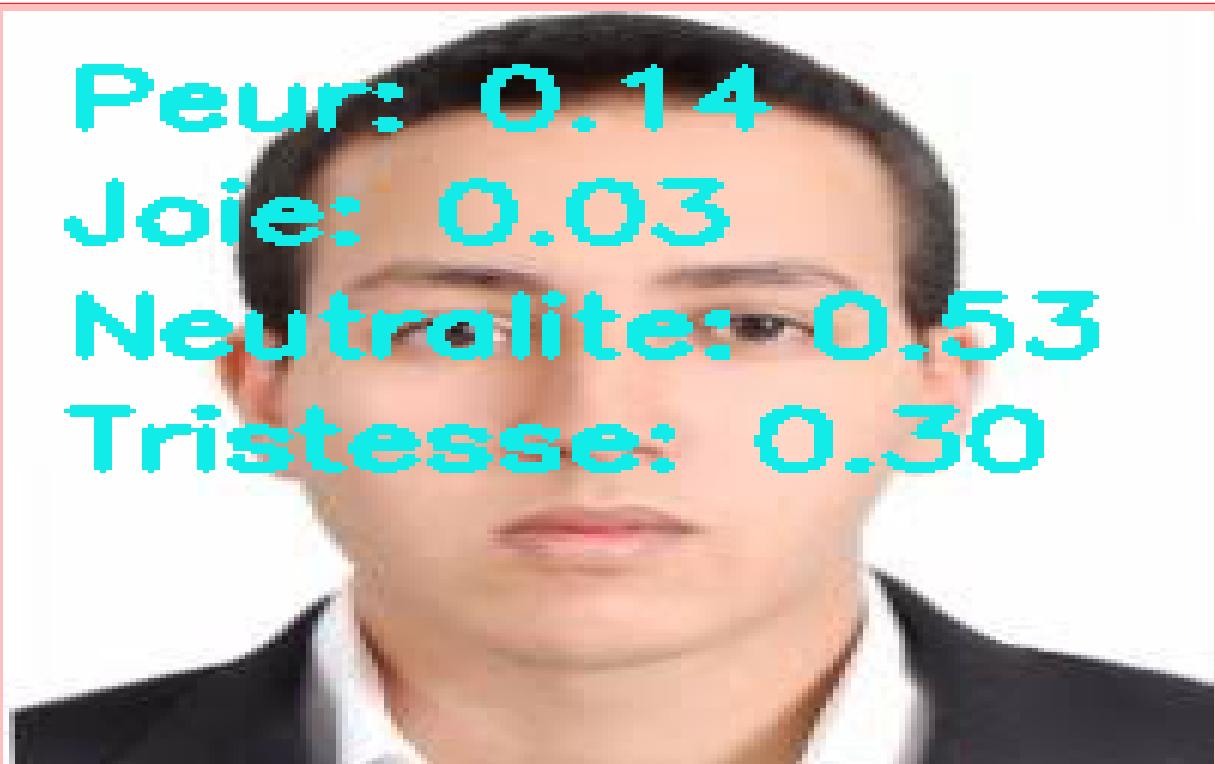
Résultat le plus probable : Joie

Peur: 0.0
Joie: 0.99
Neutralité: 0.0
Tristesse: 0.00



Résultat le plus probable : Neutralité

Peur: 0.14
Joie: 0.03
Neutralité: 0.53
Tristesse: 0.30

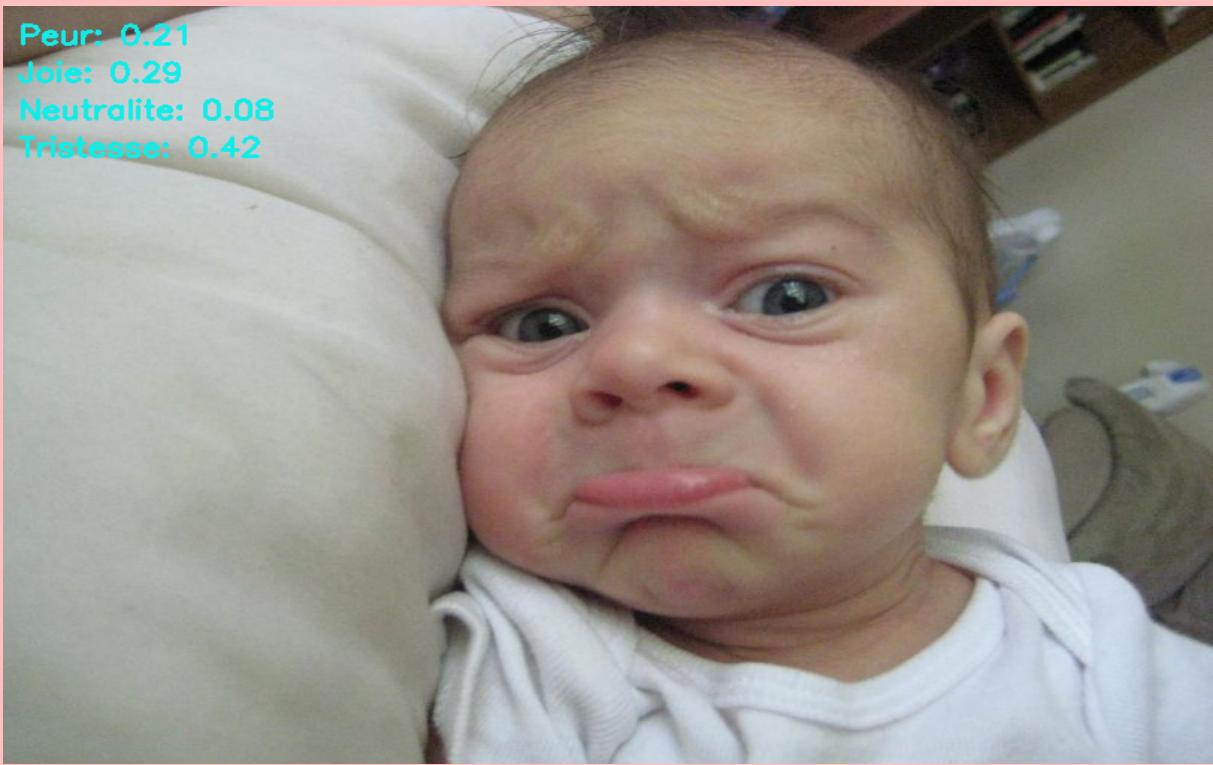


Résultat le plus probable : Peur

Peur: 0.6
Joie: 0.13
Neutralité: 0.09
Tristesse: 0.17



Résultat le plus probable : Tristesse



2.1.5 Machine learning et résolution de cas probables

Prédire si un client va quitter l'entreprise

Dans notre premier cas d'étude nous nous intéresserons à la base de données "Churn-Analysis" qui contient 10.000 lignes d'informations sur les clients d'une très grande Banque. Cette banque a observé ses clients pendant 6 mois et a noté quel client est parti et quel client est resté ; c'est pourquoi la case finale (qui est notre variable dépendante que nous cherchons à prédire) à une valeur :

$$x = \begin{cases} 0 & \text{si le client est resté} \\ 1 & \text{sinon} \end{cases}$$

The screenshot shows a Microsoft Excel spreadsheet titled "Churn Analysis.xlsx - Excel (Échec de l'activation du produit)". The data is organized into 10,000 rows and 14 columns. The columns are labeled: RowNum, CustomerID, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfPrts, HasCrCard, IsActiveM, EstimatedExit, and Exited. The "IsActiveM" column contains binary values (0 or 1), which corresponds to the variable we are trying to predict. The "Exited" column contains the actual outcome, either 0 (client remained) or 1 (client left). The "EstimatedExit" column is likely a predicted value from a machine learning model.

Nous allons nous appuyer sur le logiciel Tableau pour combiner et visualiser différents paramètres de notre jeu de données. Tableau Software est une société de logiciel américaine dont le siège se trouve à Seattle. Elle conçoit une famille de produits orientés visualisation de données.

Nous commençons par connecter Tableau au fichier Excel.

Tableau - Classeur2

Fichier Données Fenêtre Aide

Connexions Ajouter

Churn Analysis Excel

Feuilles

Churn Analysis Nouvelle union

Churn Analysis (Churn Analysis)

Connexion Direct Extraire Filtres 0 Ajouter

Trier les champs Ordre de la source de données

Afficher les alias Afficher les champs masqués 1000 lignes

Geo...	Abc Churn Analysis Geography	Abc Churn Analysis Gender	# Age	# Tenure	# Balance	# Num Of Prod...	# Has Cr Card	# Is Active Member	# Estimated Salary	# Exited
619	France	Female	42	2	0,00	1	1	1	101 349,88	1
508	Spain	Female	41	1	83 807,86	1	0	1	112 542,58	0
502	France	Female	42	8	159 660,80	3	1	0	113 931,57	1
599	France	Female	39	1	0,00	2	0	0	93 826,63	0
850	Spain	Female	43	2	125 510,82	1	1	1	79 084,10	0
645	Spain	Male	44	8	113 755,78	2	1	0	149 756,71	1
822	France	Male	50	7	0,00	2	1	1	10 062,80	0
376	Germany	Female	29	4	115 046,74	4	1	0	119 346,88	1

Source de données Feuille 1

Nous allons ensuite faire une Map pour voir d'où viennent nos clients. On déplace "Geography" vers la fenêtre du milieu après avoir changé son type de "ABC" à "Country/Region". On remarque l'apparition de deux autres paramètres "Longitude" et "Latitude".

Tableau - Classeur2

Fichier Données Feuille de calcul Tableau de bord Histoire Analyse Carte Format Fenêtre Aide

Données Analyse

Churn Analysis (Churn A... Pages

Dimensions

- Customer Id
- Gender
- Geography
- Row Number
- Surname
- Noms de mesures

Colonnes Longitude (générée)

Lignes Latitude (générée)

Filtres

Repères

Longitude Latitude

Feuille 1

Royaume-Uni Pays-Bas Allemagne Tchéquie Autriche Italie Espagne Portugal France Suisse

Longitude (générée) Latitude (générée)

OpenStreetMap contributors

Montre-moi

1 géo Dimension

0 ou plus Dimensions

0 à 2 Mesures

Pour les cartes de symboles, essayer

1 géo Dimension

0 ou plus Dimensions

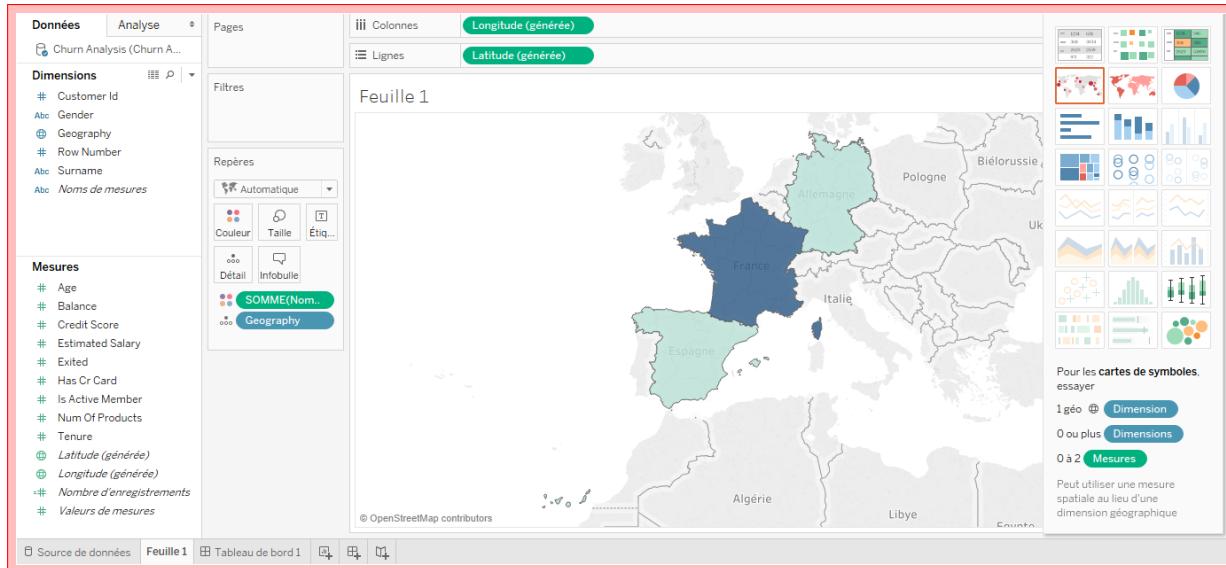
0 à 2 Mesures

Peut utiliser une mesure spatiale au lieu d'une dimension géographique

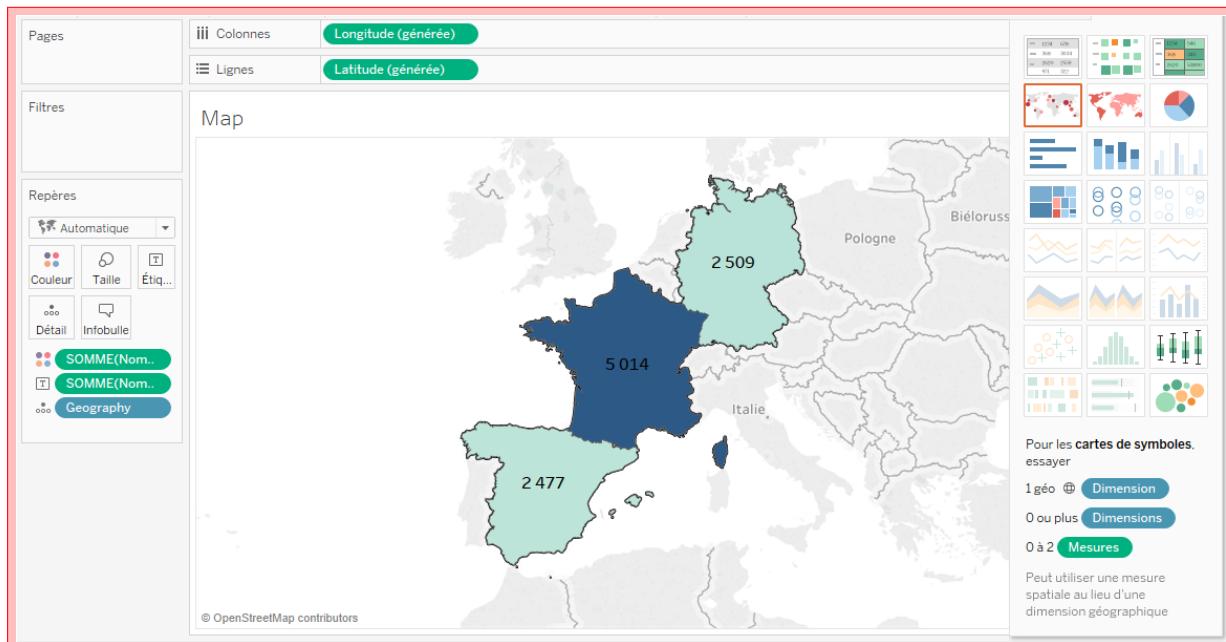
Source de données Feuille 1 Tableau de bord 1 ligne par 1 colonne

Maintenant, nous voulons supprimer les points et colorier les pays pour les différencier. En outre, nous voulons indiquer sur la Map combien il y a de clients dans chaque pays. Puisque dans notre dataset, chaque ligne correspond à un client, alors nous pouvons utiliser le "number of records" c'est-à-dire "le nombre total d'observations". Nous prenons "number of records" et nous le déplaçons vers

"couleur". Le contraste des couleurs indique qu'il y a plus de clients en France et qu'il y a à-peu-près le même nombre de clients en Allemagne et en Espagne.



Pour créer un label, c'est-à-dire faire apparaître le nombre de clients dans chaque pays, nous prenons à nouveau "number of records" et nous le déplaçons cette fois-ci vers "Label". Nous constatons alors que la majorité des clients viennent de la France (presque la moitié) tandis que l'Allemagne et l'Espagne ont le même nombre de clients. C'était une approche pour visualiser nos données, dans la suite nous allons commencer à faire du Data Mining.

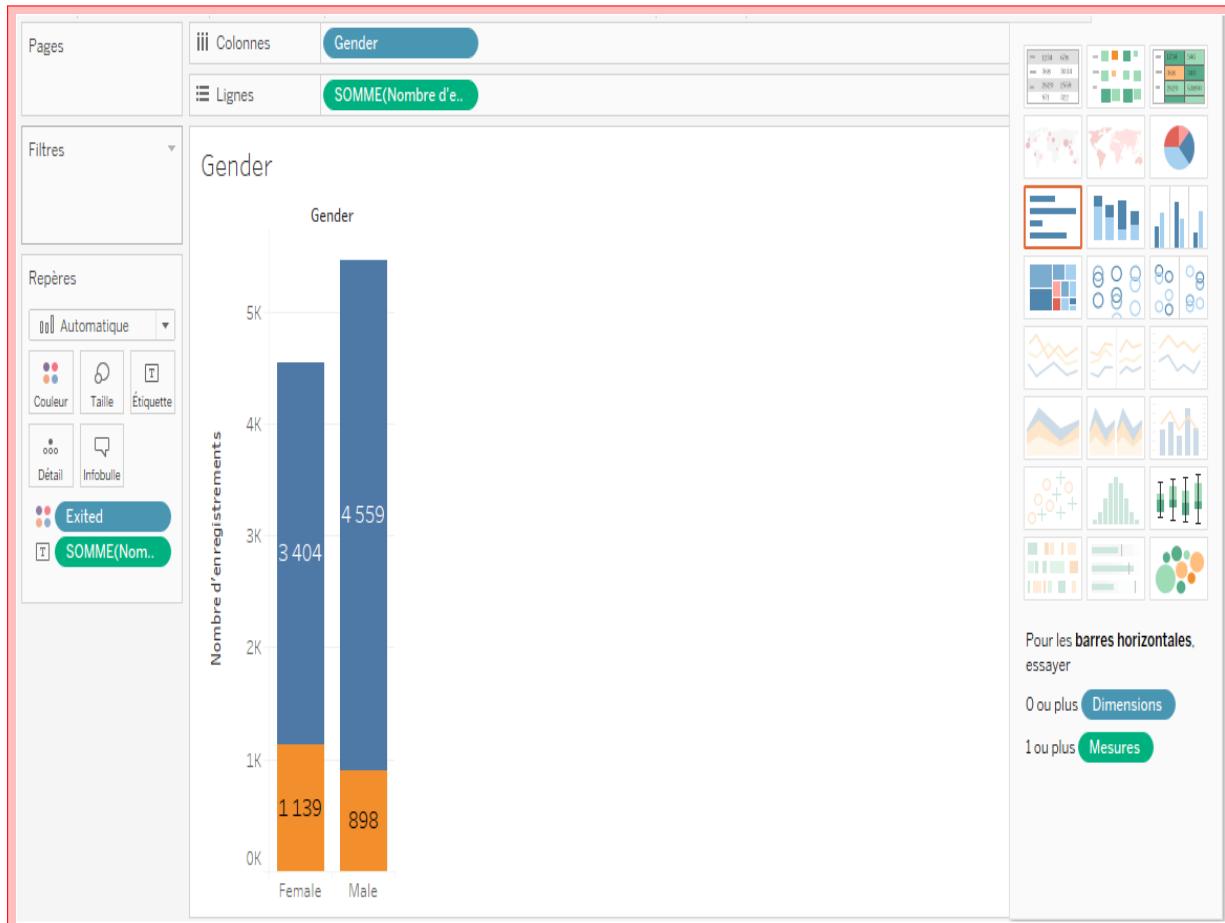


Visualiser un A-B test avec Tableau

Notre variable dépendante est "Exited" vaut 1 si le client à quitté la banque et 0 sinon. Nous la glissons donc dans "Colonnes" (car Tableau l'avait considérée comme variable numérique dépendante et l'a placée dans les "Mesures"; pour nous cette variable est plutôt "catégorique" : "est-ce que le client est parti" ou "est-ce qu'il est resté".)

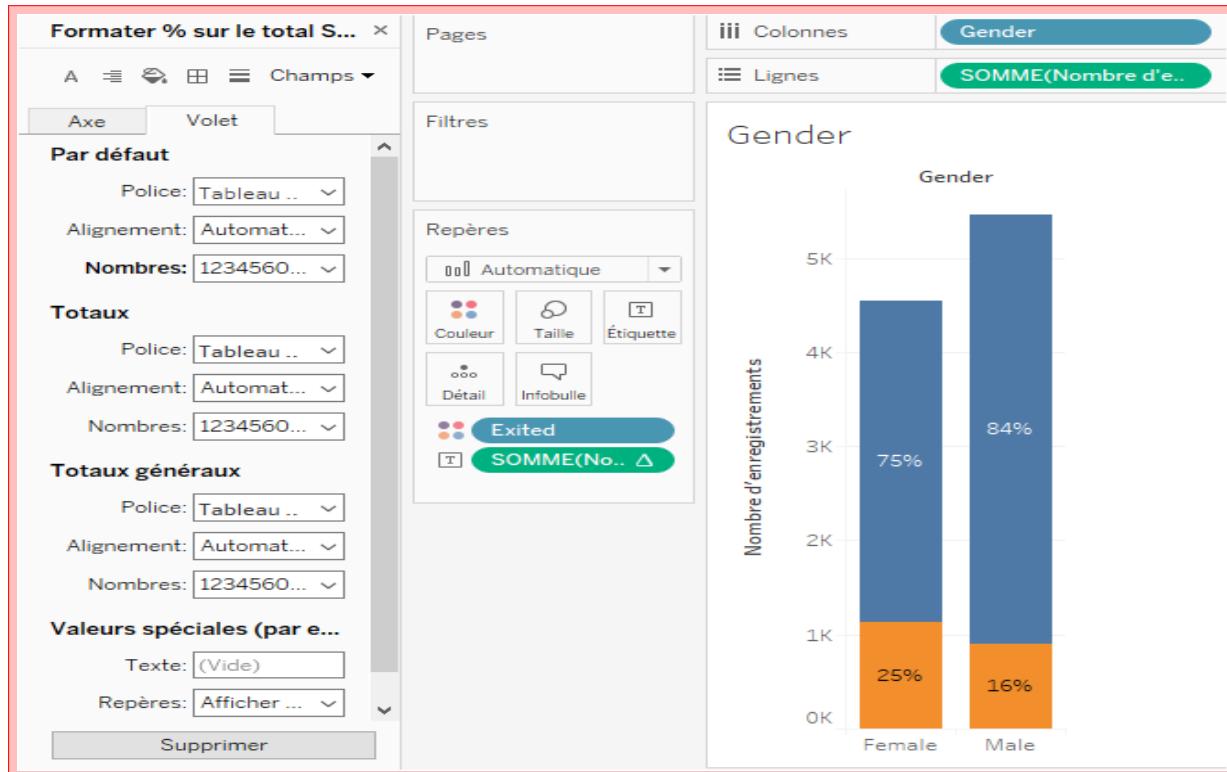
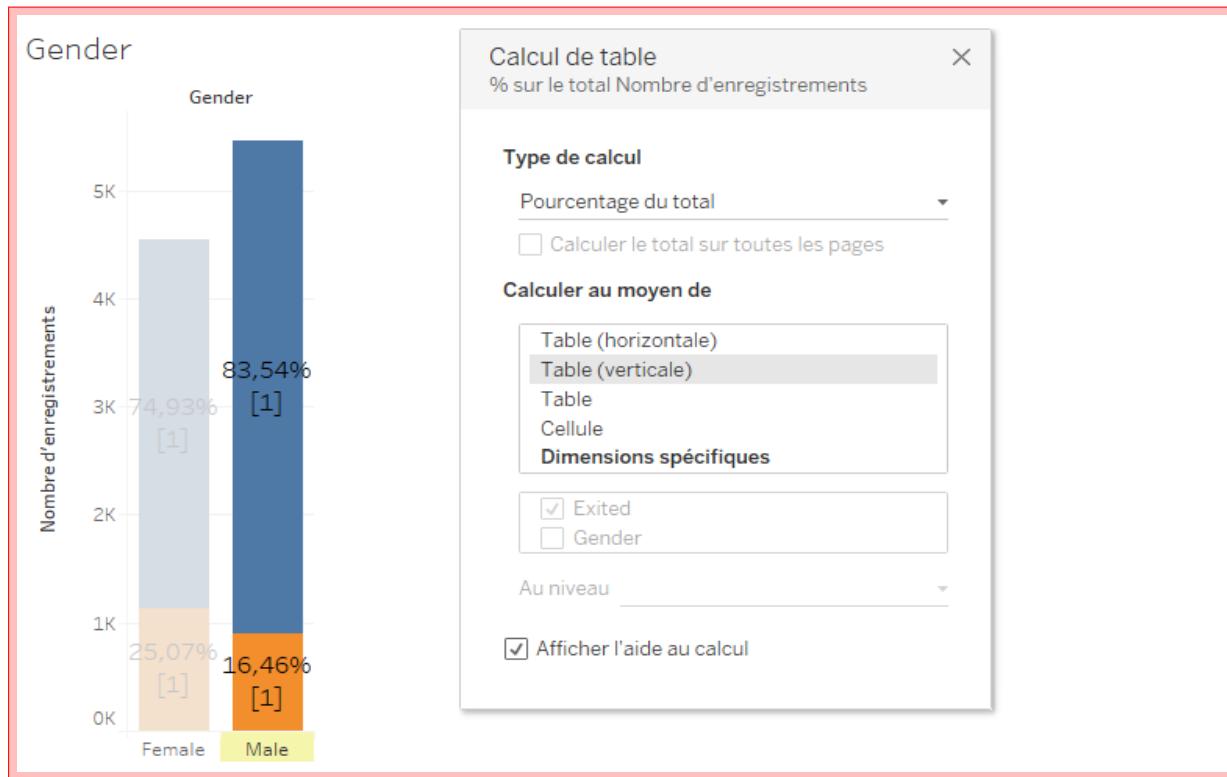
A-B test pour le genre homme/femme :

Si l'on maintient tout le reste constant et si l'on prend un client homme et une cliente femme. Lequel des deux est le plus susceptible de quitter la banque ? Tout d'abord nous prenons le genre "Gender" et nous le glissons dans "Colonnes" (car Tableau l'avait considéré comme variable numérique dépendante et l'a placée dans les "Mesures") et nous prenons la variable "Exited" et nous la plaçons dans "couleur". On obtient le graphe ci-dessous :



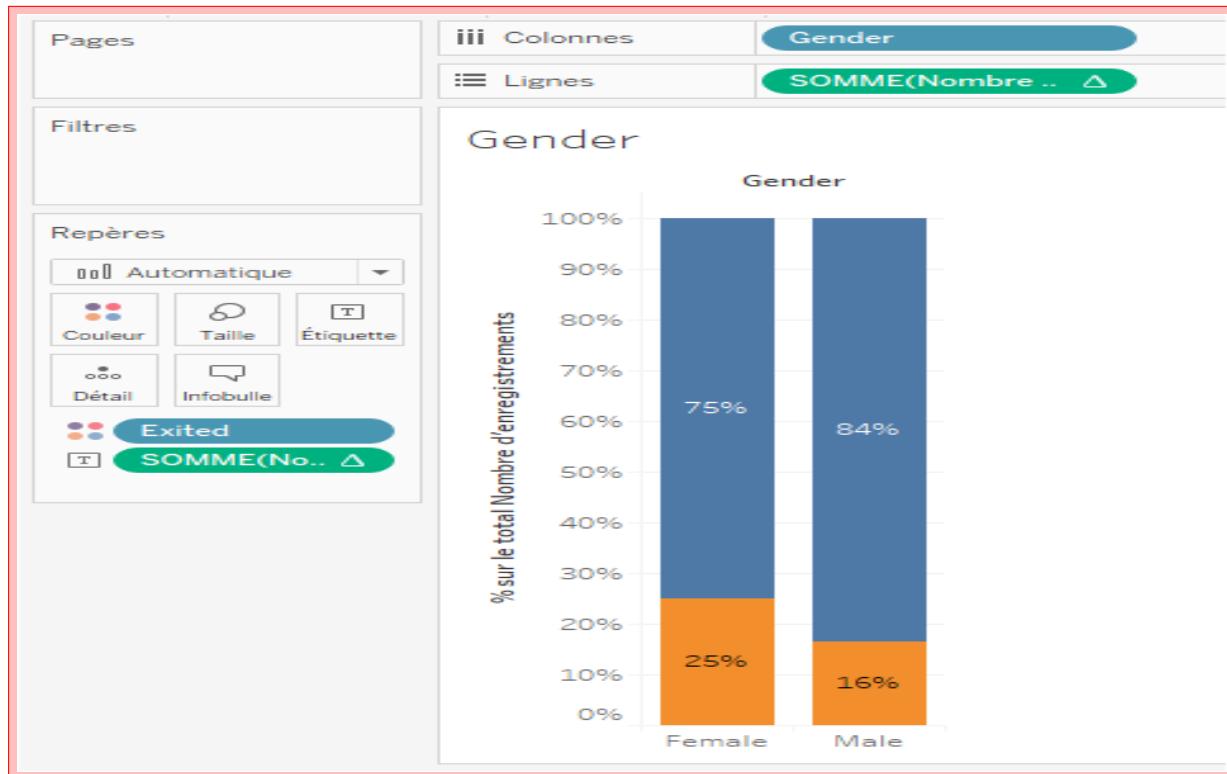
On constate que parmi les femmes une grande proportion est partie tandis que parmi les hommes une plus petite proportion est partie. Mais cela est loin d'être suffisant pour comprendre ce qui se passe. On commence par ajouter le "number of records" comme "Label". Au lieu de voir les valeurs en absolu nous aimeraisons les voir en pourcentage. C'est-à-dire nous voulons voir quel pourcentage de clientes femmes sont parties et quel pourcentage de clients hommes sont partis, afin de pouvoir les comparer directement puisque jusqu'à présent le nombre

total de femmes est différent du nombre total d'hommes.



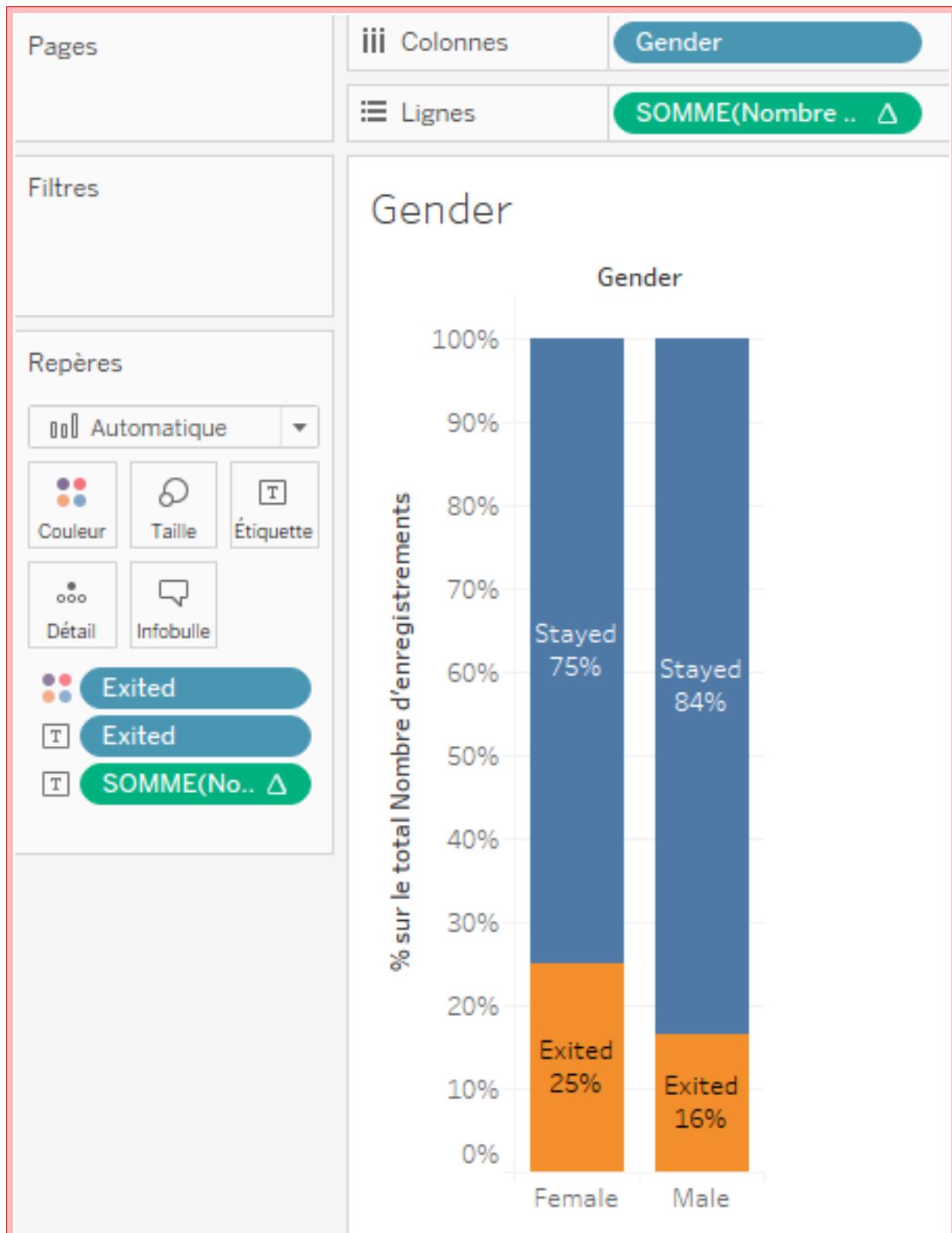
Pour convertir ces valeurs absolues en pourcentages nous allons dans le menu principal de "SUM NUMBER OF RECORDS" et nous sélectionnons "ajouter une table de calcul". Ensuite dans "type de calcul" nous choisissons "Pourcentage"

du total". Nous changeons également la valeur par défaut "table horizontale" en "table verticale" en vue d'avoir le pourcentage pour chaque colonne. Après formatage du "Label" avec zéro décimales, nous obtenons le graphe ci-dessus. Afin de rendre notre graphe plus consistant, nous remplaçons le nombre d'enregistrements par celui qu'on vient tout juste de créer.

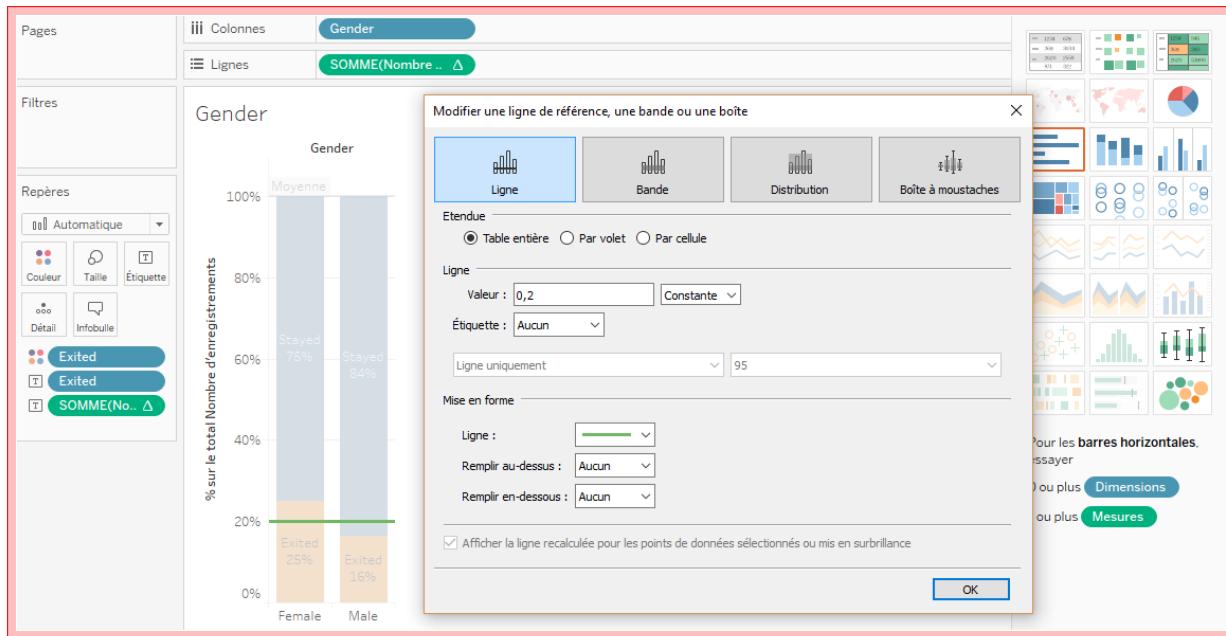


On voit que le pourcentage d'hommes qui sont partis de la banque est de 16 % alors que celui des femmes parties de la banque est de 25 %.

Donc on peut conclure que les clientes femmes sont plus susceptibles de quitter la banque que les clients hommes, tout le reste étant égal. On ajoute aussi les aliases "Stayed" et "Exited" à la place de "0" et "1" respectivement. Ce n'est pas une étude complète car nous n'avons pas fait de tests de significations statistiques cependant c'est une approche rapide et visuelle qui nous aide à obtenir rapidement des résultats et ne pas perdre de temps avec des variables non pertinentes.



Dans ce dernier graphe nous avons ajouté une ligne de référence pour constituer un benchmark qui est équivalent - dans notre cas - au pourcentage de clients qui ont quitté la banque (20 %) dans cet échantillon de 10.000 personnes. Ce qui signifie que le "churn rate" c'est-à-dire le taux moyen de départs des clients est de 20 %.



A nouveau, les clientes femmes sont plus susceptibles de quitter la banque que les clients en moyenne et les clients homme sont moins susceptibles de quitter la banque que les clients en moyenne. Par la suite nous avons répliqué cet A-B test pour inclure et combiner d'autres paramètres.

1. Pour "Geography" :

En Allemagne, les clients sont nombreux à quitter la banque avec un taux moyen de départs de 32 %. Quant à la France et l'Espagne leur taux de départ est inférieur au taux moyen de départ (qui est de 20 %), 16 % pour la France et 17 % pour l'Espagne. L'on peut dire que quelque chose se passe en Allemagne.

- Compétiteur agressive : qui fait que les clients quittent la banque pour le rejoindre.
- De nouvelles lois et réglementations : qui s'accordent moins bien avec les offres de la banque.

2. Pour "Has credit Card" :

On observe qu'il n'y a pas de grande différence entre le taux de départ des clients qui n'ont pas de carte de crédit et le taux de départ de ceux qui ont une carte de crédit puisqu'ils sont respectivement égaux à 21 % et 20 %. Donc avoir une carte de crédit n'a pas réellement d'impact sur le choix de rester ou quitter la banque.

- 0 : signifie que le client n'a pas de carte de crédit.
- 1 : le contraire.

3. Pour "Is Active Member" :

"Être actif" dépend des propres critères de la banque qui peuvent être par exemple :

- Est-ce que le client s'est connecté au moins une fois à son compte bancaire le mois dernier.
- Est-ce que le client a fait au moins une opération bancaire, etc.

Ici, nous avons changé les aliases : au lieu de 0 c'est No c'est à dire client non actif et 1 pour dire l'inverse.

Parmi les clients non actifs 27 % ont quitté la banque et pour les clients actifs 14 % ont quitté la banque. Celà confirme nos intuitions puisqu'il est évident que les clients non actifs sont plus susceptibles de quitter la banque que les clients actifs. En effet, le fait d'être actif signifie que le client est plus satisfait de sa banque donc aura tendance à y rester.

4. Pour "Number of products" :

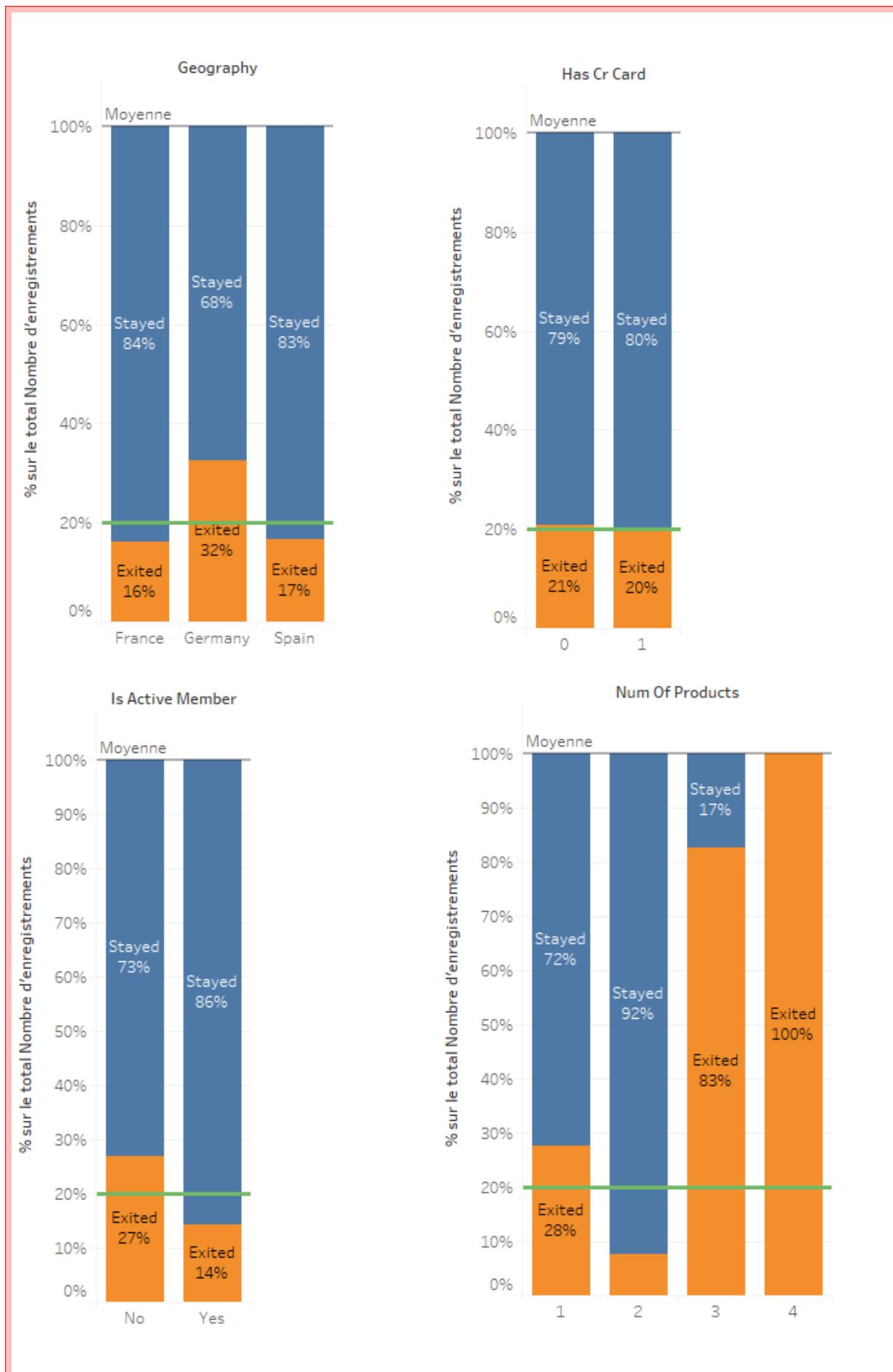
On observe quelques anomalies. En effet, on tend à penser que plus le client achète de produits plus il sera susceptible de rester dans la banque. C'est bien ce qu'on observe dans les deux premières barres : celui qui a un produit est plus susceptible de quitter la banque que celui qui a deux produits. Mais quand le nombres de produits est de 3 ou 4 on observe un énorme taux de départ à tel point que 100 % des clients qui ont acheté 4 produits ont quitté la banque.

Supposition :

- Effets de hasard : lorsque la banque a sélectionné ses clients dans cet échantillon alors par un pur effet de hasard elle a pris très peu de clients qui ont 4 produits et qui sont tous partis de la banque. Chose qui est facilement vérifiable : On prend "Number of records" et on le glisse dans "Label", on obtient le graphe ci-dessous. Ce qui valide notre supposition, en effet on peut observer sur les deux premières barres qu'il y a eu beaucoup de clients sélectionnés pour 1 ou 2 produits dans cet échantillon. En revanche, il y a eu peu de clients dans les barres 3 et 4 (266 et 60 clients, respectivement).

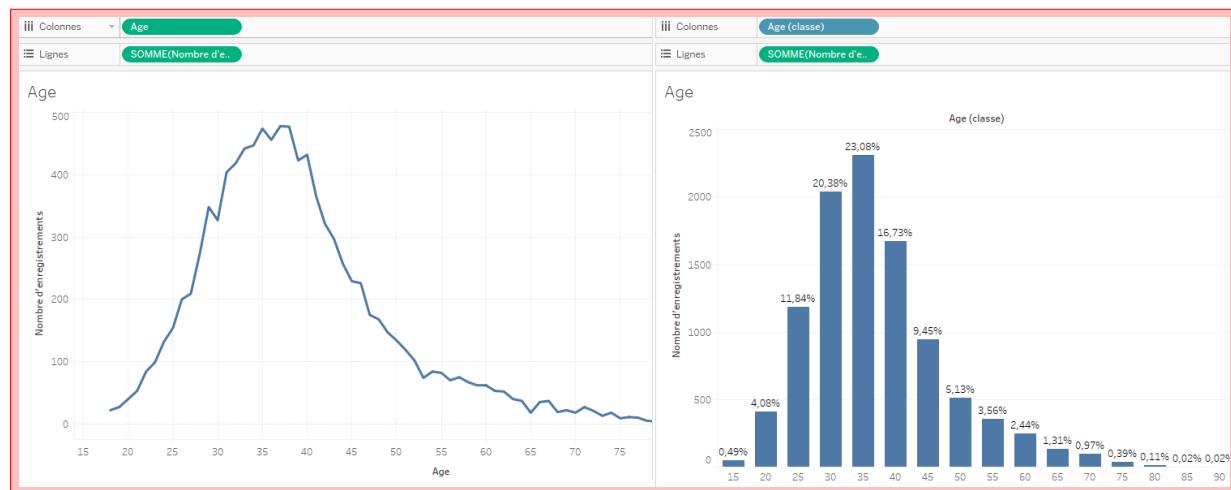
Conclusion :

Dans 1 et 2 beaucoup de clients sont sélectionnés, donc on peut faire des conclusions statistiques. En revanche, pour 3 et 4 peu de clients sont sélectionnés, donc on ne doit pas tirer de conclusions hâtives.



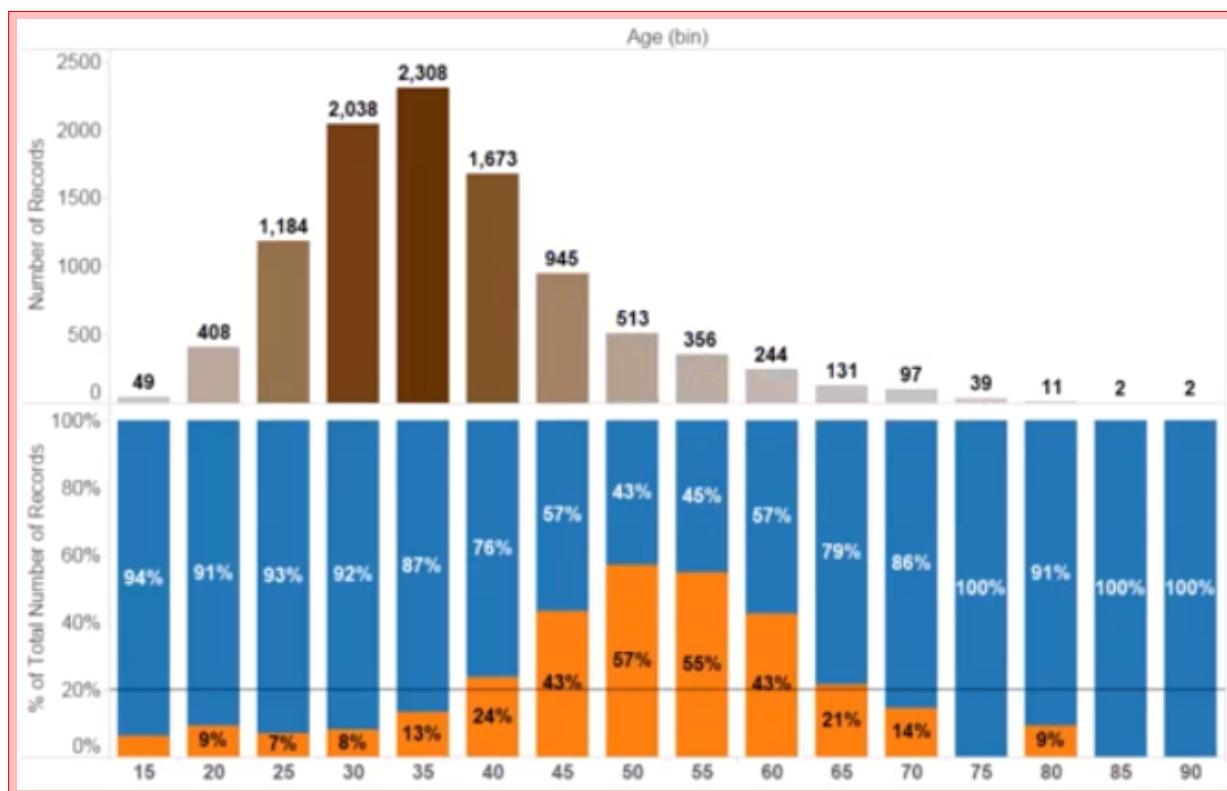
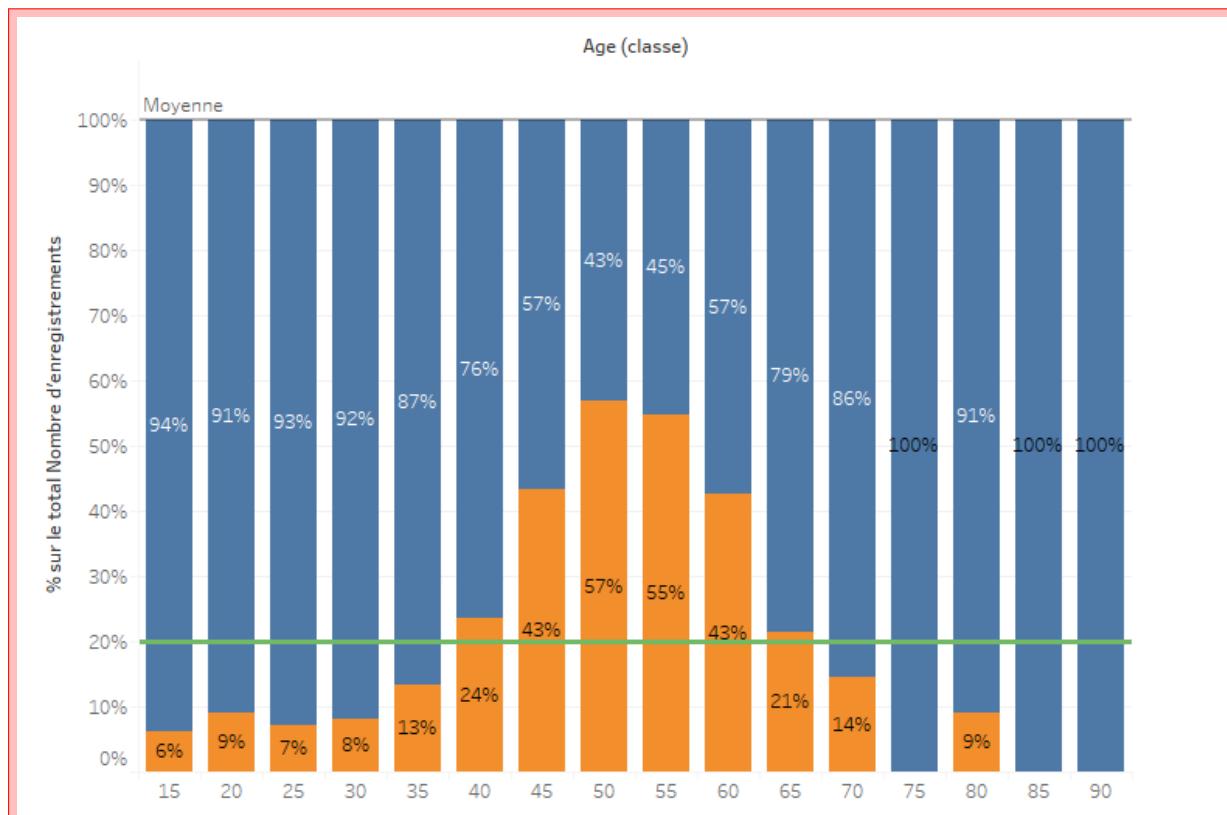
En ce qui concerne la distribution continue des âges de nos clients : la courbe à gauche nous donne pour n'importe quel age le nombre de clients qui ont cet age. Si dans le dataset les ages n'étaient pas arrondis alors on aurait eu des personnes ayant l'âge de 35,5 ce qui provoquera des irrégularités. (plein de pics avec de fortes variations) Bien que nos âges soit arrondis nous voulons nous débarrasser des petites irrégularités que l'on observe sur notre courbe. Il existe un moyen très efficace il s'agit des "bins" ou "classes" qui consistent à regrouper les observations en différentes tranches. Dans le graphe, à droite, nous les avons regroupées par tranches de 5 ans. Cette nouvelle distribution est discrète tandis que la distribution précédente était continue. Puisque notre échantillon compte 10.000 clients, il est assez représentatif de la population totale des clients et par conséquent cette distribution l'est également.

On voit que la tranche la plus dominante (en pourcentage) est celle de 35-39 ans et que la deuxième tranche la plus dominante est celle de 30-34 ans. Après en prenant plus de recul on voit que la plupart des clients ont un âges compris entre 25-40 ans. Ce qui paraît très cohérent. Après quelques recherches nous avons trouvé qu'il s'agit d'une "right skewed distribution". C'est le terme technique d'une distribution étalée sur la droite, en effet elle a une longue queue à droite. Les distributions asymétriques à droite sont également appelées "distributions asymétriques positives" / "positive-skew distributions". C'est parce qu'il y a une longue queue dans la direction positive sur la ligne numérique. La moyenne est également à la droite du pic.



Ici, nous allons créer un nouveau A-B test de la variable catégorique Age(bin). Ce dernier nous donnera dans chaque tranche d'âge la proportion de clients qui quitte la banque et celle qui reste dans la banque. Ainsi on obtiendra la tranche d'âge dans laquelle les clients sont les plus susceptibles de partir. On obtient alors le test de classification ci-dessous. On remarque que pour les tranches d'âge dans les extrémités, les clients sont peu à partir puisqu'on voit en effet que les proportions des clients qui quittent la banque est bien en dessous de la moyenne de 20 %. En revanche dans les tranches d'âges du milieu les clients sont plus nombreux à partir. Conclusion : les clients agés de 45-60 ans sont plus susceptibles à quitter la banque.

Maintenant il est question de comprendre pourquoi dans les tranches d'âges les moins élevés et les plus élevés les clients sont peu susceptibles de partir et d'analyser pourquoi ces tranches d'âges de 45-60 ans sont celles ou les clients sont le plus susceptibles de partir.

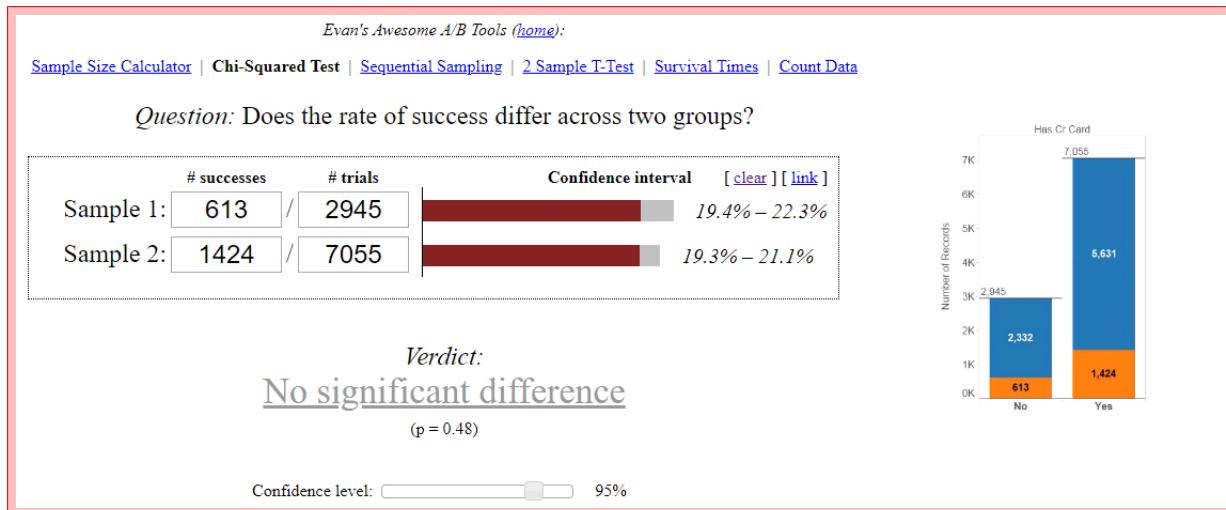


Le test du Chi-deux

En se servant du site : www.evanmiller.org/ab-testing/chi-squared.html nous avons fait correspondre "sample 1" aux femmes et "sample 2" aux hommes pour la variable "Gender". Ensuite "successes" correspond à 1 (c'est-à-dire lorsque le client part de la banque) et "trials" correspond aux nombres d'observations de "sample 1" et "sample 2" c'est-à-dire pour "sample 1" au nombre total de femmes et pour "sample 2" cela correspond au nombre total d'hommes. On obtient le verdict : "Sample 1 is more successful". Ce dernier signifie que sur la population totale les femmes sont plus susceptibles de quitter la banque que les hommes. $p - value < 0.001$: signifie que la variable indépendante "Gender" a un effet statistiquement très significatif sur la variable dépendante "Exited". Ce qui implique que sur la population totale et pas seulement sur l'échantillon les femmes sont plus susceptibles de quitter la banque que les hommes.



En procédant de la même manière sur la variable "Has credit card" on obtient le verdict "No significant difference". Dans ce cas, la p-value est très élevée ; largement supérieur à 5 % ce qui confirme que la variable indépendante "Has Credit Card" n'a pas d'effet statistiquement significatif sur la variable dépendante "Exited". Donc comme nous l'avons prédit la variable Has Credit Card n'a pas d'influence sur le taux de sortie de la banque. Ainsi voilà comment l'on peut faire un A-B test statistique avec deux catégories.



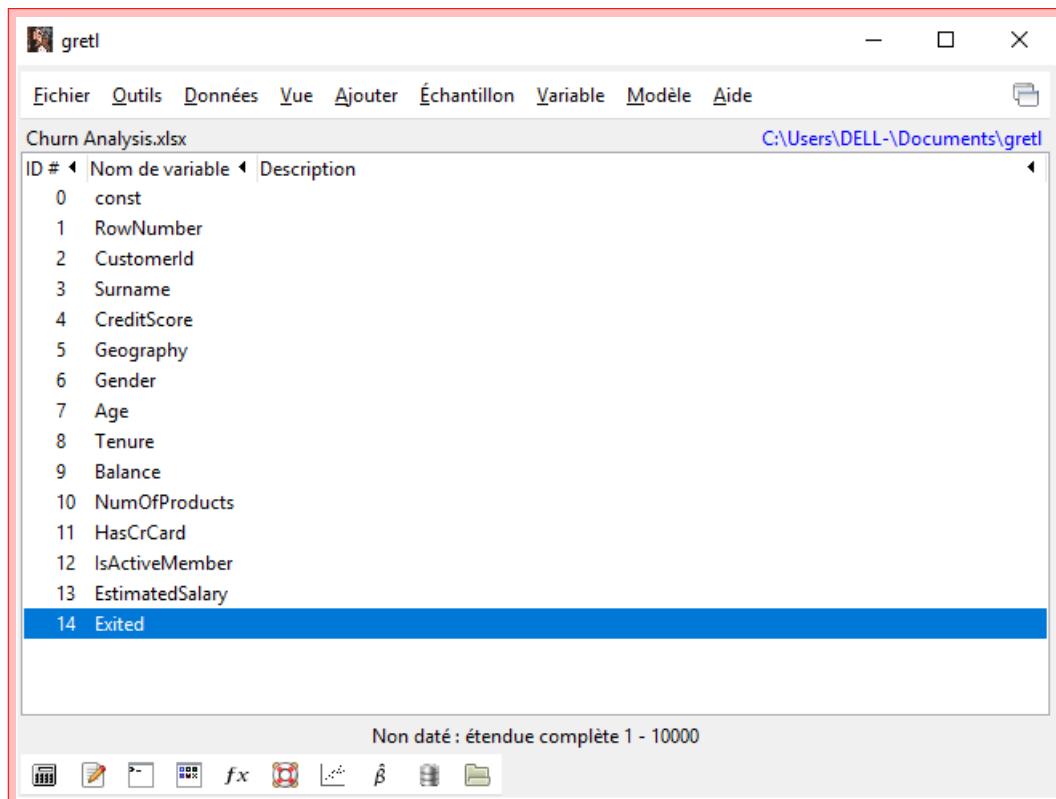
Pour un test statistique Chi-deux pour n'importe quel nombre de catégories. - Cas de 3 variables : l'idée est de mettre de coté l'une des variables, pour comparer les deux autres catégories. Si on fait un A-B test statistique avec les variables A et B et qu'on obtient une différence significative des taux de départs pour A et B alors cela voudrait dire que la catégorie de A et B à une influence significative sur la variable Exited. Donc, si on ajoute une variable C cela ne va rien changer sur le ABC test.

La segmentation géo-démographique

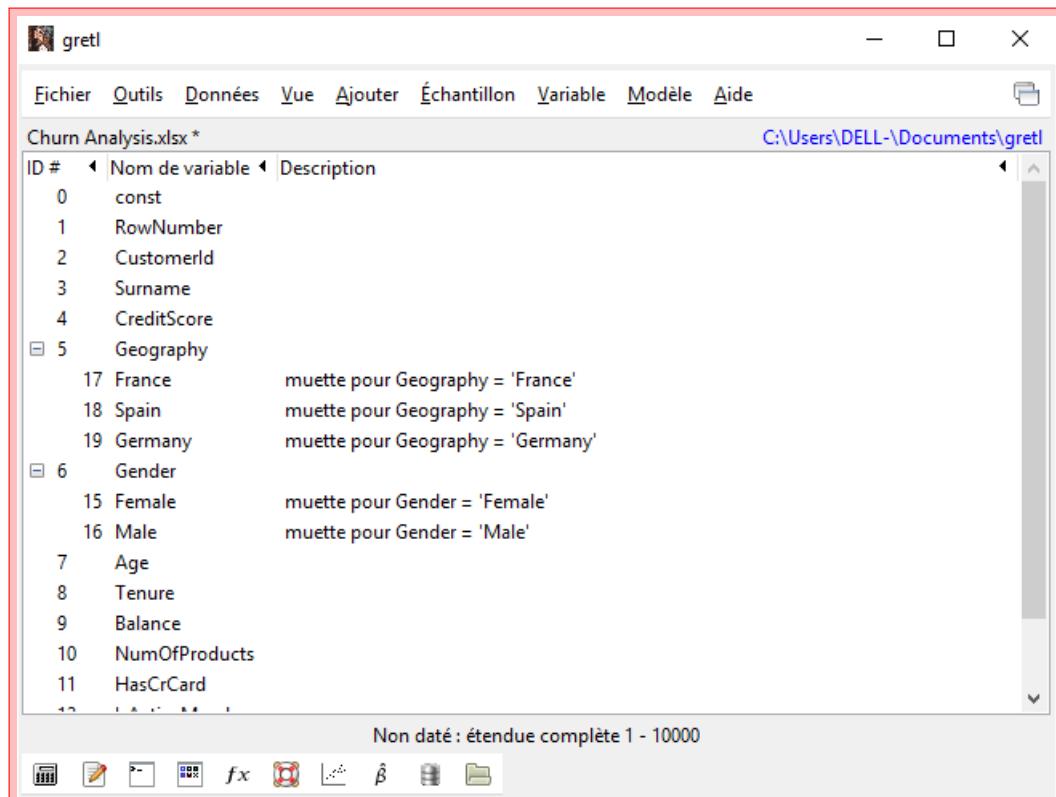
Le système de segmentation géodémographique, aussi appelé analyse typologique, est une technique de classification statistique multivariée servant à classer une population en sous-groupes d'individus les plus similaires possibles en appliquant plusieurs comparaisons quantitatives de caractéristiques multiples. Les différences au sein de tout groupe doivent être inférieures à la différence entre les groupes. Cette méthode est souvent utilisée pour la segmentation des consommateurs et le positionnement de marque.

Nous allons construire un modèle robuste de segmentation géo-démographique pour le même jeu de données en utilisant la régression logistique. Tout simplement parce que ce qu'on voudrait prédire est un résultat binaire : 0 si le client reste dans la banque et 1 sinon.

Nous ouvrons notre dataset avec Gretl (GNU Regression, Econometrics and Time Series Library) qui est un logiciel de statistiques qui peut être utilisé en ligne de commande ou au travers d'une interface graphique.



Ce dernier nous affiche un message disant qu'il a trouvé 3 variables non numériques à savoir Surname, Geography et Gender qui sont bien des variables catégoriques. Nous les convertirons en variables muettes, mis à part "Surname" qui n'influencera en rien notre modèle.



Après avoir spécifié les variables dépendantes et indépendantes de la régression logistique, on obtient la première itération du modèle de segmentation géodémographique.

- Pour le premier bloc il nous donnent les informations générales sur les variables (coefficients, écart type, z-score et p-value).
- Le deuxième bloc nous donne des informations sur le modèle dans sa globalité y compris le R^2_{ajust} .
- Le troisième et dernier bloc en bas à gauche nous donne la matrice de confusion.

	coefficients	erreur std.	z	p. critique
const	-3,92076	0,245354	-15,98	1,76e-057 ***
CreditScore	-0,00069329	0,000280345	-2,384	0,0171 **
Age	0,0727060	0,00257551	28,23	2,52e-175 ***
Tenure	-0,0159491	0,00935487	-1,705	0,0882 *
Balance	2,63707e-06	5,14213e-07	5,128	2,92e-07 ***
NumOfProducts	-0,101523	0,0471342	-2,154	0,0312 **
HasCrCard	-0,0446764	0,0593395	-0,7529	0,4515
IsActiveMember	-1,07544	0,0576856	-18,64	1,43e-077 ***
EstimatedSalary	4,80699e-07	4,73663e-07	1,015	0,3102
Female	0,526483	0,0544884	9,699	3,04e-022 ***
Spain	0,0352178	0,0706379	0,4986	0,6181
Germany	0,774714	0,0676740	11,45	2,41e-030 ***
Moy. var. dép.	0,203700	Éc. type var. dép.	0,402769	
R2 de McFadden	0,153161	R2 ajusté	0,150787	
Log de vraisemblance	-4280,678	Critère d'Akaike	6585,355	
Critère de Schwarz	8671,879	Hannan-Quinn	8614,643	
Nombre de cas 'correctement prédis' = 8103 (81,0%)				
f(beta*x) à la moyenne des variables indépendantes = 0,135				
Test du ratio de vraisemblance: Chi-deux(11) = 1548,43 [0,0000]				
Prédit	0	1		
Actuel 0	7666	297		
1	1600	437		
Constante mise à part, la probabilité critique est la plus élevée pour la variable 18 (Spain)				

Nous allons tenter d'améliorer cette première itération avec la méthode "backward elimination" ou élimination descendante qui est une procédure de sélection de variables au cours de laquelle toutes les variables sont introduites dans l'équation, puis éliminées une à une. La variable ayant la plus petite corrélation partielle avec la variable dépendante est la variable dont la suppression est étudiée en premier. Si elle répond aux critères d'élimination, elle est supprimée. Une fois la première variable éliminée, l'élimination de la variable suivante restant dans l'équation et ayant le plus petit coefficient de corrélation partielle est étudiée. La procédure prend fin quand plus aucune variable de l'équation ne satisfait aux critères de suppression. Notre priorité sera donc de nous débarrasser des variables non significatives en regardant les p-values et le nombre d'étoiles pour chaque variable.

La dernière ligne du premier modèle nous indique clairement qu'il faut supprimer la variable muette Spain avec une p-value de 0.6181 très élevée par rapport aux autres variables du modèle rapport à 5 % (ce qui justifie l'absence totale d'étoiles). En la supprimant, on obtient notre deuxième itération.

```

gretl : modèle 2

Fichier Édition Tests Sauvegarder Graphiques Analyse LaTeX

Modèle 2: Logit, utilisant les observations 1-10000
Variable dépendante: Exited
Écarts type basés sur la matrice hessienne

      coefficient   erreur std.      z      p. critique
-----
const          -3,91097    0,244526    -15,99    1,41e-057 *** 
CreditScore     -0,000666615  0,000280294   -2,378    0,0174 ** 
Age             0,0727230    0,00257536    28,24    2,00e-175 *** 
Tenure          -0,0159766    0,00935423   -1,708    0,0876 * 
Balance         2,63733e-06  5,14201e-07    5,129    2,91e-07 *** 
NumOfProducts   -0,101288    0,0471276   -2,149    0,0316 ** 
HasCrCard       -0,0449303    0,0593378   -0,7572   0,4489 
IsActiveMember -1,07519     0,0576828   -18,64    1,53e-077 *** 
EstimatedSalary 4,81342e-07  4,73649e-07    1,016    0,3095 
Female          0,528343     0,0544870    9,697    3,11e-022 *** 
Germany         0,762937     0,0633614   12,04    2,16e-033 *** 

Moy. var. dép.      0,203700  Éc. type var. dép.      0,402769
R2 de McFadden    0,153137  R2 ajusté           0,150961
Log de vraisemblance -4280,802 Critère d'Akaike      8583,603
Critère de Schwarz 8662,917   Hannan-Quinn        8610,451

Nombre de cas 'correctement prédis' = 8100 (81,0%)
f(beta'*x) à la moyenne des variables indépendantes = 0,135
Test du ratio de vraisemblance: Chi-deux(10) = 1548,18 [0,0000]

Prédit
      0      1
Actuel 0  7665  298
      1  1602  435

Constante mise à part, la probabilité critique est la plus élevée pour la variable 11 (HasCrCard)

```

A nouveau , Gretl indique que la prochaine variable qu'on doit supprimer est Has Credit Card avec la p-value la plus élevée mais également plus grande du seuil de significativité de 5 %. Donc le fait d'avoir une carte de crédit ou pas n'a aucun impact significatif sur le taux de départ des clients de la banque. De même pour EstimatedSalary dans le quatrième modèle, qui n'influence en rien le départ des clients.

```

gretl: modèle 3

Fichier Édition Tests Sauvegarder Graphiques Analyse LaTeX

Modèle 3: Logit, utilisant les observations 1-10000
Variable dépendante: Exited
Écarts type basés sur la matrice hessienne

      coefficient   erreur std.      z      p. critique
-----
const          -3,94435    0,240579    -16,40    2,07e-060 *** 
CreditScore    -0,000664033 0,000280270   -2,369   0,0178   ** 
Age            0,0727303    0,00257516    28,24    1,73e-175 *** 
Tenure          -0,0161505   0,00935127   -1,727   0,0842   * 
Balance         2,64543e-06  5,14070e-07   5,146    2,66e-07 *** 
NumOfProducts   -0,101333    0,0471228    -2,150   0,0315   ** 
IsActiveMember -1,07438     0,0576668    -18,63   1,81e-077 *** 
EstimatedSalary 4,81783e-07  4,73661e-07   1,017    0,3091 
Female          0,528489     0,0544853    9,700    3,02e-022 *** 
Germany         0,761879     0,0633445    12,03    2,55e-033 *** 

Moy. var. dép.      0,203700   Éc. type var. dép.  0,402769 
R2 de McFadden    0,153080   R2 ajusté        0,151102 
Log de vraisemblance -4281,088 Critère d'Akaike  8582,175 
Critère de Schwarz 8654,279   Hannan-Quinn    8606,582 

Nombre de cas 'correctement prédis' = 8111 (81,1%)
f(beta'x) à la moyenne des variables indépendantes = 0,135
Test du ratio de vraisemblance: Chi-deux(9) = 1547,61 [0,0000]

Prédit
      0      1
Actuel 0  7673  290
      1  1599  438

Constante mise à part, la probabilité critique est la plus élevée pour la variable 13 (EstimatedSalary)

gretl: modèle 4

Fichier Édition Tests Sauvegarder Graphiques Analyse LaTeX

Modèle 4: Logit, utilisant les observations 1-10000
Variable dépendante: Exited
Écarts type basés sur la matrice hessienne

      coefficient   erreur std.      z      p. critique
-----
const          -3,89591    0,235717    -16,53    2,31e-061 *** 
CreditScore    -0,000666426 0,000280263   -2,378   0,0174   ** 
Age            0,0727016    0,00257462    28,24    2,01e-175 *** 
Tenure          -0,0159836   0,00934933   -1,710   0,0873   * 
Balance         2,65326e-06  5,13979e-07   5,162    2,44e-07 *** 
NumOfProducts   -0,100475    0,0471176    -2,132   0,0330   ** 
IsActiveMember -1,07509     0,0576636    -18,64   1,41e-077 *** 
Female          0,528981     0,0544804    9,710    2,74e-022 *** 
Germany         0,762059     0,0633400    12,03    2,43e-033 *** 

Moy. var. dép.      0,203700   Éc. type var. dép.  0,402769 
R2 de McFadden    0,152978   R2 ajusté        0,151197 
Log de vraisemblance -4281,605 Critère d'Akaike  8581,210 
Critère de Schwarz 8646,103   Hannan-Quinn    8603,176 

Nombre de cas 'correctement prédis' = 8115 (81,2%)
f(beta'x) à la moyenne des variables indépendantes = 0,135
Test du ratio de vraisemblance: Chi-deux(8) = 1546,57 [0,0000]

Prédit
      0      1
Actuel 0  7676  287
      1  1598  439

```

Après avoir créé 4 modèles, nous allons commencer à regarder le R^2_{ajust} qui dans le premier modèle vaut 0.1507, pour le deuxième vaut 0.1509, le troisième avec une valeur de 0.1511 et enfin le quatrième qui vaut à-peu-près 0.1512. On sait que lorsque le R^2_{ajust} augmente cela signifie que la nouvelle équipe de variables

prédit mieux la variable dépendante que l'ancienne équipe de variables. Même chose pour l'"accuracy" qui a augmenté progressivement en passant du premier modèle au quatrième. Ce qui, jusqu'à maintenant, nous montre qu'on n'a pas exclut à tort une variable. Cependant, dans ce quatrième modèle Gretl ne nous indique plus quelle variable à la plus grande p-value car en effet toutes les variables ont au moins une étoile qui montre leur importance (significativité) pour notre modèle pour le seuil de 10 %.

```

gretl : modèle 5
Fichier Édition Tests Sauvegarder Graphiques Analyse LaTeX

Modèle 5: Logit, utilisant les observations 1-10000
Variable dépendante: Exited
Écarts type basés sur la matrice hessienne

      coefficient   erreur std.      z      p. critique
-----
const        -3,97602    0,231165   -17,20   2,66e-066 *** 
CreditScore  -0,000666050  0,000280207   -2,377  0,0175 **  
Age          0,0726871    0,00257415    28,24   2,04e-175 *** 
Balance      2,65176e-06   5,13865e-07    5,160   2,46e-07  *** 
NumOfProducts -0,100984   0,0470915    -2,144  0,0320 **  
IsActiveMember -1,07180   0,0576145    -18,60   3,04e-077 *** 
Female       0,530571    0,0544653     9,741   2,01e-022 *** 
Germany      0,760829    0,0633294     12,01   3,01e-033 *** 

Moy. var. dép.      0,203700   Éc. type var. dép.  0,402769 
R2 de McFadden    0,152689   R2 ajusté           0,151106 
Log de vraisemblance -4283,067 Critère d'Akaike    8582,134 
Critère de Schwarz 8639,817   Hannan-Quinn      8601,660 

Nombre de cas 'correctement prédis' = 8112 (81,1%) 
f(beta'*x) à la moyenne des variables indépendantes = 0,135 
Test du ratio de vraisemblance: Chi-deux(7) = 1543,65 [0,0000] 

Prédit
      0      1
Actuel 0  7676  287
      1  1601  436

gretl : modèle 6
Fichier Édition Tests Sauvegarder Graphiques Analyse LaTeX

Modèle 6: Logit, utilisant les observations 1-10000
Variable dépendante: Exited
Écarts type basés sur la matrice hessienne

      coefficient   erreur std.      z      p. critique
-----
const        -3,99238    0,232615   -17,16   5,02e-066 *** 
CreditScore  -0,000674380  0,000280215   -2,407  0,0161 **  
Age          0,0726405    0,00257405    28,22   3,29e-175 *** 
NumOfProducts -0,0954940   0,0475089    -2,010  0,0444 **  
IsActiveMember -1,07253   0,0575976    -18,62   2,17e-077 *** 
Female       0,528301    0,0544440     9,704   2,91e-022 *** 
Germany      0,746303    0,0650378     11,47   1,76e-030 *** 
Log_Balance   0,0690313   0,0139553     4,947   7,55e-07  *** 

Moy. var. dép.      0,203700   Éc. type var. dép.  0,402769 
R2 de McFadden    0,152501   R2 ajusté           0,150919 
Log de vraisemblance -4284,015 Critère d'Akaike    8584,029 
Critère de Schwarz 8641,712   Hannan-Quinn      8603,554 

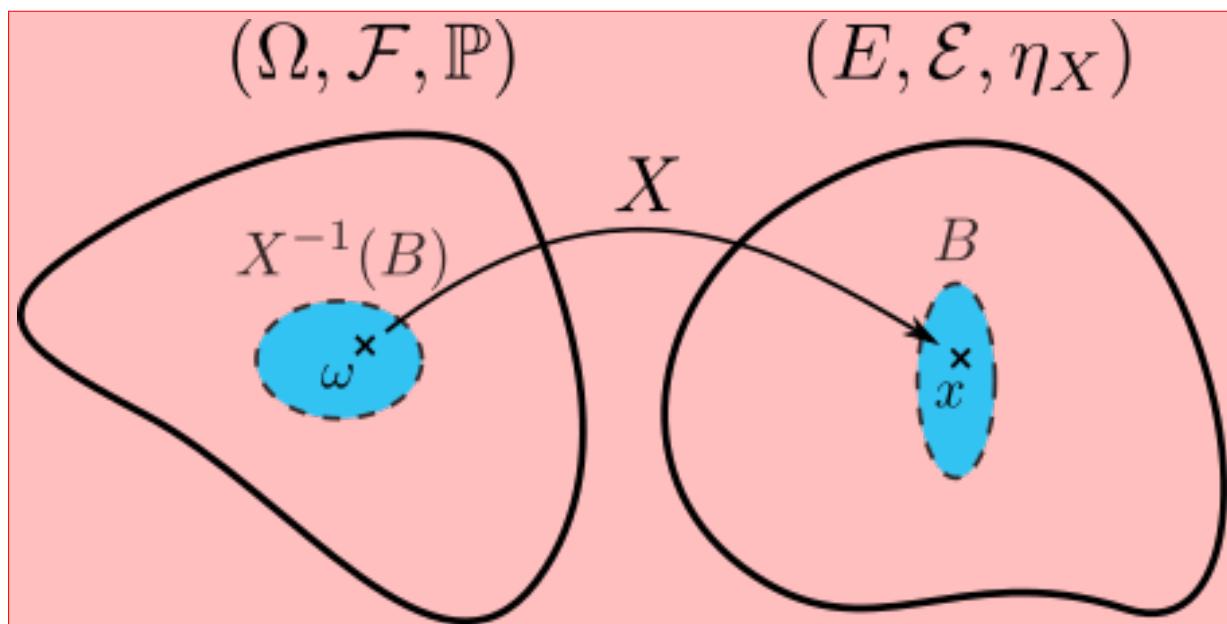
Nombre de cas 'correctement prédis' = 8114 (81,1%) 
f(beta'*x) à la moyenne des variables indépendantes = 0,135 
Test du ratio de vraisemblance: Chi-deux(7) = 1541,75 [0,0000] 

Prédit
      0      1
Actuel 0  7681  282
      1  1604  433

```

Transformation de variables indépendantes

Transformer une variable consiste en une opération arithmétique qui vise à construire une nouvelle variable à partir de la variable d'origine, de sorte que la distribution de la nouvelle variable soit différente de celle de la variable d'origine et plus conforme à certaines caractéristiques, tout en préservant l'ordre des valeurs de la variable d'origine. Elle sert à transformer des variables indépendantes de telle sorte que la nouvelle combinaison de variables obtenue prédisse mieux la variable dépendante. Dans certains cas toutefois, la relation transformée apporte des avantages pour l'interprétation et/ou offre une meilleure description du phénomène sous examen. La transformation la plus courante pour une variable dépendante est de prendre son logarithme, le plus souvent le logarithme naturel, i.e. en utilisant la base e, qui vaut approximativement 2.71828. Nous avons choisi "la Balance" avec la transformation logarithme en base 10. On rajoute +1 au cas où il y a des soldes égaux à zéro.



gretl : ajout de variable

Entrer la formule pour une nouvelle variable

Aide
Annuler
Valider

gretl : modèle 6						
Fichier Édition Tests Sauvegarder Graphiques Analyse LaTeX						
Modèle 6: Logit, utilisant les observations 1-10000						
Variable dépendante: Exited						
Écarts type basés sur la matrice hessienne						
	coefficient	erreur std.	z	p. critique		

const	-3,99238	0,232615	-17,16	5,02e-066	***	
CreditScore	-0,000674380	0,000280215	-2,407	0,0161	**	
Age	0,0726405	0,00257405	28,22	3,29e-175	***	
NumOfProducts	-0,0954940	0,0475089	-2,010	0,0444	**	
IsActiveMember	-1,07253	0,0575976	-18,62	2,17e-077	***	
Female	0,528301	0,0544440	9,704	2,91e-022	***	
Germany	0,746303	0,0650378	11,47	1,76e-030	***	
Log_Balance	0,0690313	0,0139553	4,947	7,55e-07	***	
Moy. var. dép.	0,203700	Éc. type var. dép.	0,402769			
R2 de McFadden	0,152501	R2 ajusté	0,150919			
Log de vraisemblance	-4284,015	Critère d'Akaike	8584,029			
Critère de Schwarz	8641,712	Hannan-Quinn	8603,554			
Nombre de cas 'correctement prédis' = 8114 (81,1%)						
f(beta'x) à la moyenne des variables indépendantes = 0,135						
Test du ratio de vraisemblance: Chi-deux(7) = 1541,75 [0,0000]						
Prédit						
	0	1				
Actuel 0	7681	282				
1	1604	433				
gretl : modèle 7						
Fichier Édition Tests Sauvegarder Graphiques Analyse LaTeX						
Modèle 7: Logit, utilisant les observations 1-10000						
Variable dépendante: Exited						
Écarts type basés sur la matrice hessienne						
	coefficient	erreur std.	z	p. critique		

const	-3,91258	0,237164	-16,50	3,84e-061	***	
CreditScore	-0,000674866	0,000280272	-2,408	0,0160	**	
Age	0,0726550	0,00257451	28,22	3,24e-175	***	
NumOfProducts	-0,0950198	0,0475374	-1,999	0,0456	**	
IsActiveMember	-1,07578	0,0576458	-18,66	1,01e-077	***	
Female	0,526721	0,0544591	9,672	3,97e-022	***	
Germany	0,747595	0,0650515	11,49	1,44e-030	***	
Log_Balance	0,0690263	0,0139592	4,945	7,62e-07	***	
Tenure	-0,0158791	0,00934627	-1,699	0,0893	*	
Moy. var. dép.	0,203700	Éc. type var. dép.	0,402769			
R2 de McFadden	0,152787	R2 ajusté	0,151006			
Log de vraisemblance	-4282,570	Critère d'Akaike	8583,141			
Critère de Schwarz	8648,034	Hannan-Quinn	8605,107			
Nombre de cas 'correctement prédis' = 8127 (81,3%)						
f(beta'x) à la moyenne des variables indépendantes = 0,135						
Test du ratio de vraisemblance: Chi-deux(8) = 1544,64 [0,0000]						
Prédit						
	0	1				
Actuel 0	7687	276				
1	1597	440				

En comparant les deux modèles, celui avec la "Balance" et l'autre avec la "Log-Balance", on remarque que le coefficient de balance est très petit, tandis que la "log-balance" a un coefficient non négligeable par rapport aux autres, et donc c'est un meilleur coefficient. En effet, regardons comment cela a impacté notre modèle : On voit que dans le modèle avant la transformation $R^2_{ajust} = 0.151197$ et dans le nouveau modèle après la transformation $R^2_{ajust} = 0.151006$, donc R^2_{ajust} a diminué ce qui signifie que peut être le second modèle est un petit peu moins bien que le premier (modèle 5). Mais nous allons tout de même opter pour ce deuxième modèle pour les raisons qui vont suivre.

La variable "Balance" étant mesurée en milliers de dollars ce qui signifie que l'ajout d'une unité équivaut à mille dollars. Pour le scénario 1 supposons qu'on a 1.000 dollars dans le compte bancaire. Que se passerait-il lorsque la balance augmente d'une unité ?

Cela nous fera 1.000 dollars de plus et au final on aura 2.000 dollars. Maintenant pour le second scénario on voudrait comprendre l'effet de l'augmentation d'une unité sur un compte ayant à la base 10.000 dollars. Dans ce cas le compte bancaire passerait de 10.000 à 11.000 dollars. Soit une augmentation de 100 % dans le premier cas et une augmentation de 10 % dans le second. On conclut que plus le compte bancaire est élevé à la base moins l'augmentation de ce solde d'une unité sera significative.

En faisant exactement la même chose avec la "log-balance" : si il y a une augmentation d'une unité pour cette variable transformée alors cela implique que pour la variable "Balance" elle sera multipliée par 10. Cette deuxième approche a un effet beaucoup plus consistant que pour la première approche. Parce que avec la première approche, quand on considère le compte bancaire d'un client donné c'est-à-dire quand on essaie de segmenter ce client par son solde bancaire, plus le solde du client est élevé moins l'augmentation de son solde d'une unité aura un effet significatif. Par conséquent, plus le compte initial du client est élevé moins on remarque l'effet de l'augmentation d'une unité. On ne pourra donc pas segmenter deux clients qui ont des soldes très différents. Parce que pour ces deux clients l'augmentation de leurs soldes d'une unité signifiera deux choses totalement différentes. Tandis que pour la deuxième approche celle avec la "log-balance" l'augmentation d'une unité pour le solde bancaire reviendra à multiplier par 10 le solde bancaire, peu importe le solde initial. Et donc là l'effet est beaucoup plus consistant sur l'ensemble de la population des clients. En conclusion, la transformation est une technique très puissante qui ne restreint pas notre régression logistique à un segment particulier de clients mais à la totalité de la population des clients. Ainsi, en faisant le choix d'ajouter la tenure le meilleur modèle est le modèle 7.

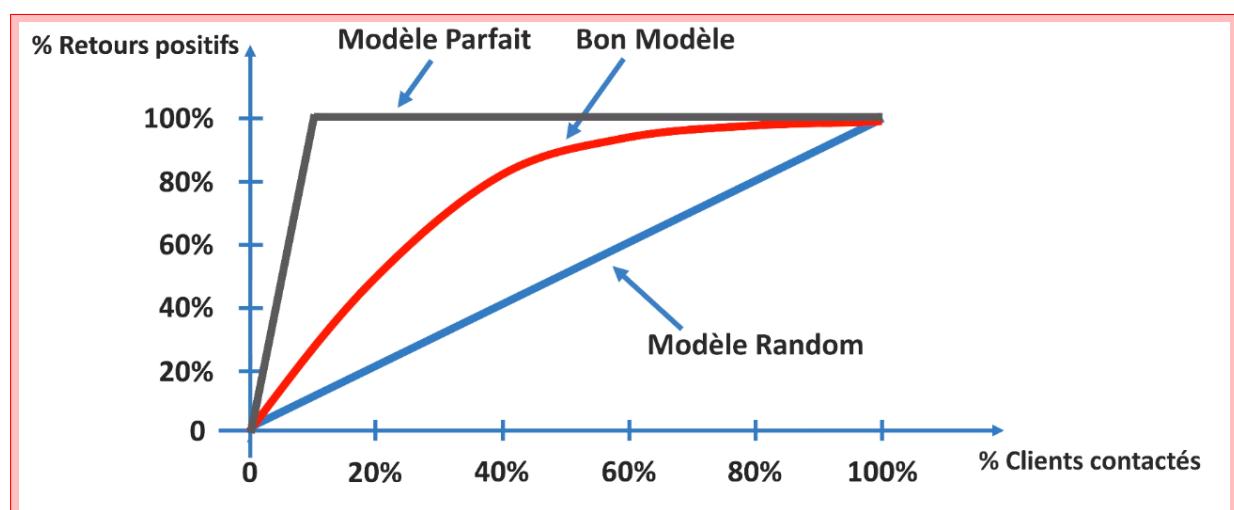
Balance (en 1000 dollars)	$\log_{10}(Balance + 1)$
$Bal_2 = Bal_1 + 1 \text{ unité}$	$\log_{10}(Bal_2) = \log_{10}(Bal_1) + 1 \text{ unité}$
$Bal_2 = Bal_1 + 1.000 \text{ dollars}$	$Bal_2 = Bal_1 * 10$
Scénario 1 : $Bal_1 = 1.000$	Scénario 1 : $Bal_1 = 1.000$
$Bal_2 = 1.000 + 1.000 = 2.000$	$Bal_2 = 1.000 * 10 = 10.000$
Scénario 2 : $Bal_1 = 10.000$	Scénario 2 : $Bal_1 = 10.000$
$Bal_2 = 10.000 + 1.000 = 11.000$	$Bal_2 = 10.000 * 10 = 100.000$

Influence de la transformation $\log_{10}()$

Evaluation de notre modèle - Courbe CAP

The cumulative accuracy profile (CAP) est utilisé en science des données pour visualiser le pouvoir discriminant d'un modèle. Le CAP d'un modèle représente le nombre cumulé de résultats positifs le long de l'axe y par rapport au nombre cumulé correspondant d'un paramètre de classification le long de l'axe des x. Le CAP est distinct de la Receiver Operating Characteristic (ROC), qui représente le taux des vrais-positifs par rapport au taux des faux-positifs. Il y a trois courbes importantes dans l'analyse CAP :

- La droite bleue qui représente la selection au hasard des clients.
- La courbe rouge de notre modèle logistique pour faire la segmentation géo-démographique.
- La courbe en gris qui est associée au modèle absolument parfait. C'est-à-dire le modèle où l'on a déjà les réponses à l'avance. Dans notre cas c'est savoir quels clients vont quitter la banque sans jamais se tromper.

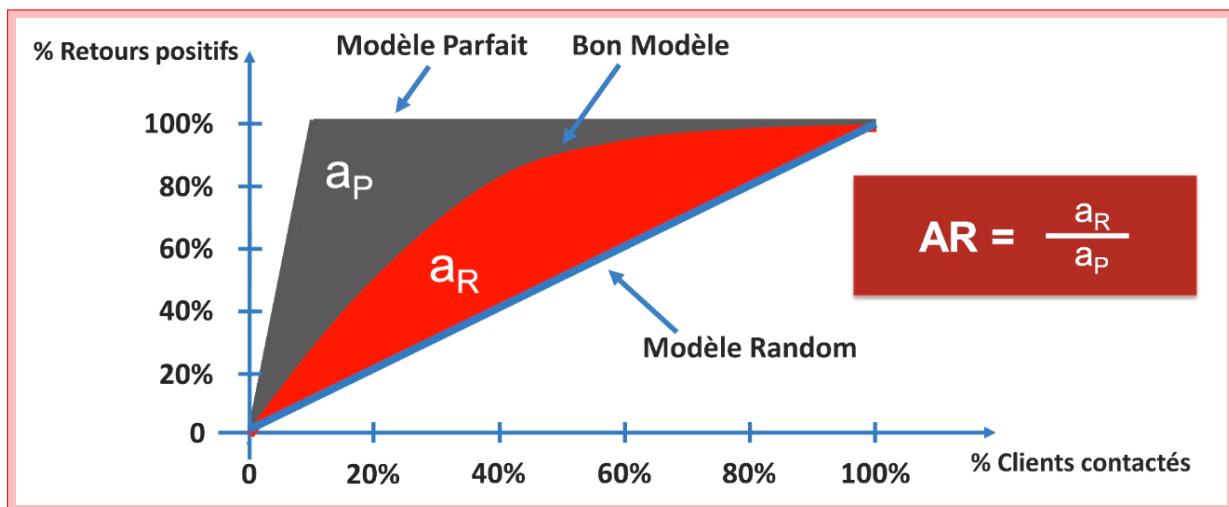


1. Approche qualitative :

Il paraît assez intuitif de dire que plus la courbe rouge est proche de la courbe grise plus le modèle est performant et plus la courbe rouge est proche de la bleue moins notre modèle sera performant.

2. Approche quantitative :

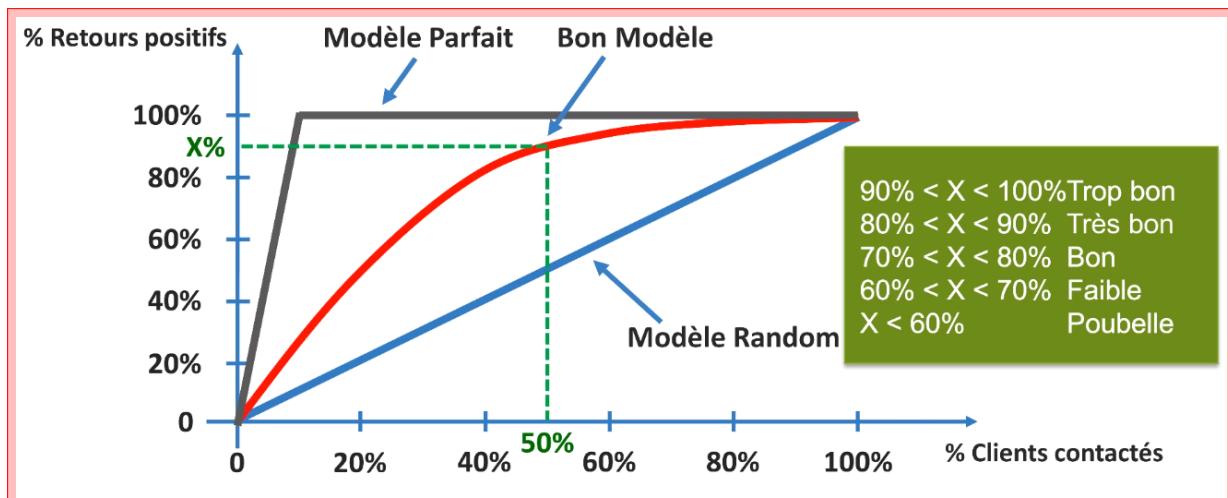
Se résume à prendre l'aire entre la courbe rouge et la droite bleue a_R et la diviser par l'aire entre la courbe grise et la droite bleue a_P . Ce ratio est clairement compris entre 0 et 1 puisque a_R reste plus petite que a_P . Plus le ratio est proche de 1, plus les deux courbes Rouge et Grise seront proches et plus le modèle sera proche du modèle absolument performant et donc notre modèle sera robuste. A l'inverse, plus le ratio est proche de 0, plus les deux courbes celle bleue et l'autre rouge seront proches et donc moins le modèle est performant. Cependant, calculer ces deux aires n'est pas toujours facile. C'est pourquoi nous optons pour la troisième approche.



3. Approche géométrique :

On projette l'abscisse 50 % sur la courbe rouge et on projette à nouveau ce point de la courbe rouge sur l'axe vertical des retours positifs, X. Ce point projeté (X % des retours positifs) c'est le pourcentage des clients qui ont acheté notre produit après avoir sélectionné la moitié des clients qui ont les plus grandes probabilités de répondre favorablement à l'offre et d'acheter le produit. De ce pourcentage X on peut établir une règle pour quantifier le CAP :

- Si X est inférieur à 60 % on peut trouver mieux.
- Entre 60 et 70 %, le modèle est considéré comme faible ou moyen.
- Pour un X compris entre 70 et 80, le modèle est performant et c'est l'intervalle qu'on doit viser.
- Si on obtient un X compris entre 80 et 90 %, ce modèle est très bon.
- Enfin pour un X entre 90 et 100 %, il y a deux cas : soit le modèle est trop bon ou bien il y a un problème de surapprentissage ou surajustement ou surinterprétation (en anglais « overfitting »). Le surapprentissage s'interprète comme un apprentissage « par coeur » des données. Il résulte souvent d'une trop grande liberté dans le choix du modèle. Un modèle surajusté est un modèle statistique qui contient plus de paramètres que ne peuvent le justifier les données.

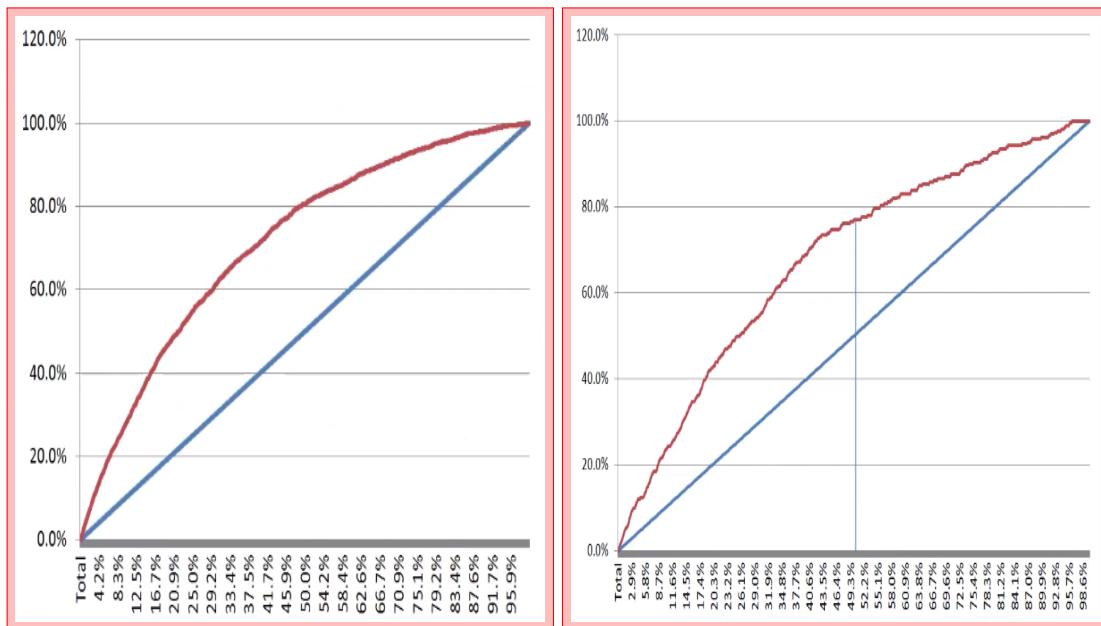


Pour visualiser la courbe CAP de notre modèle, on commence par créer une nouvelle variable, P_{Hat} , qui englobe toutes les valeurs prédictes par le modèle 7 c'est-à-dire les prédictions des probabilités de quitter la banque.

	Exited	prédition
1	1,000000	0,124048
2	0,000000	0,159205
3	1,000000	0,326060
4	0,000000	0,226189
5	0,000000	0,156311
6	1,000000	0,246175
7	0,000000	0,098720
8	1,000000	0,293328
9	0,000000	0,116342
10	0,000000	0,036933
11	0,000000	0,123881
12	0,000000	0,060543
13	0,000000	0,169997
14	0,000000	0,098908
15	0,000000	0,066085
16	0,000000	0,219464
17	1,000000	0,700780
18	0,000000	0,031634
19	0,000000	0,226262
...

	A	B	C	D	E	F	G	H	I	J	K
1											
2	Stats										
3	Total Exited	2037									
4	Total Records	10000									
5	Exit Ratio	20.4%									
6											
7											
8	RowNumber	Exited	P_Hat	Total Sel	Total Sele	Random S	Random S Model	Model Sel	Model Select	Percent	
9	4816	0	0.93067	1	0.0%	0.2037	0.0%	0	0.0%		
10	3532	1	0.92831	2	0.0%	0.4074	0.0%	1	0.0%		
11	9588	0	0.9233	3	0.0%	0.6111	0.0%	1	0.0%		
12	7500	1	0.90618	4	0.0%	0.8148	0.0%	2	0.1%		
13	9556	1	0.90446	5	0.1%	1.0185	0.1%	3	0.1%		
14	8489	1	0.89224	6	0.1%	1.2222	0.1%	4	0.2%		
15	7630	1	0.88286	7	0.1%	1.4259	0.1%	5	0.2%		
16	7693	0	0.8762	8	0.1%	1.6296	0.1%	5	0.2%		
17	9748	1	0.87429	9	0.1%	1.8333	0.1%	6	0.3%		
18	4464	1	0.87183	10	0.1%	2.037	0.1%	7	0.3%		
19	7009	1	0.86191	11	0.1%	2.2407	0.1%	8	0.4%		
20	8157	0	0.85975	12	0.1%	2.4444	0.1%	8	0.4%		
21	4436	1	0.857	13	0.1%	2.6481	0.1%	9	0.4%		
22	7814	1	0.84206	14	0.1%	2.8518	0.1%	10	0.5%		
23	4560	1	0.83991	15	0.2%	3.0555	0.2%	11	0.5%		
24	417	1	0.83736	16	0.2%	3.2592	0.2%	12	0.6%		

A l'aide d'Excel, nous créons la courbe du modèle de regression logistique appliquée à la segmentation géo-démographique après avoir ordonné les probabilités pour que les clients quittent ou non la banque dans l'ordre décroissant. La première ligne Excel correspond alors au client le plus susceptible de partir tandis que la dernière ligne correspond au client le moins susceptible de quitter la banque.



En comparant les deux courbes CAP (celle du training set et celle du test set) nous remarquons que le modèle est un petit peu moins performant sur le test set que sur le training set puisque la projection de l'abscisse 50 % sur la courbe

rouge a pour ordonnée une valeur plus de 80 % pour le training set et moins de 80 % pour le test set. On peut ainsi dire que la performance du modèle a baissé de 3 %. La deuxième remarque est que la courbe rouge est plus lisse dans le training set tandis qu'il y a beaucoup d'irrégularités dans celle du test set. En effet, la cause principale de ces irrégularités est que le training set contient 10 fois plus d'observations que le test set. Donc il y a moins de points pour construire la courbe du test set et par conséquent elle est plus rugueuse. De manière générale puisque le modèle était construit sur le training set alors c'est tout à fait normal d'avoir une légère différence de performance entre les deux modèles puisque l'apprentissage était basé sur la compréhension des corrélations du training set. Une autre explication serait cette différence soit due aux données biaisées, c'est-à-dire la distribution des 1000 observations (dans le test set) peut être plus dense à droite ou à gauche. Malgré cela, le modèle du test set reste assez performant puisque son score de performance est de 78 %. En effet, cela veut dire que lorsqu'on contacte 50 % de la totalité des clients, le modèle trouve 78 % des clients qui ont effectivement quitté la banque. Conclusion : notre score est compris entre 70 et 80 % ce qui signifie que notre modèle a un bon pouvoir prédictif sur le test set.

2.1.6 Chatbot et marketing, la combinaison gagnante

Avec l'accord de l'entreprise UITS, nous avons pris l'initiative de créer deux sortes de Chatbots, l'une à intégrer dans le site de l'Union It Services pour l'assistance à distance des clients effectifs et l'autre dans les réseaux sociaux pour marketer les formations et services de l'entreprise visant les clients potentiels. En ce qui concerne le premier chatbot, Nous nous sommes servi de RiveScript pour créer une bot conversationnel (interface robotisée pour répondre aux questionnements des clients) qui est un langage de script simple pour les chatbots avec une syntaxe conviviale et facile à apprendre.

Chatbot est un robot logiciel pouvant dialoguer avec un individu ou consommateur par le biais d'un service de conversations automatisées effectuées en grande partie en langage naturel. Le chatbot utilise à l'origine des bibliothèques de questions et réponses. Les chatbots peuvent répondre à une logique de marketing « relationnel » ou avoir une vocation de support client en avant vente ou après vente, il s'agit alors de bot conversationnels. Ils sont également utilisés pour prendre directement des commandes. On parle alors de chatbot transactionnel et de commerce conversationnel. Nous avons utilisé les tutos de Rivescript pour créer le fichier "brain.rive". Nous avons également intégré p5.speech qui est une bibliothèque JavaScript qui fournit un accès simple et clair aux "API Web Speech Recognition" et "Speech Recognition", permettant de créer facilement des sketches capables de dialoguer et d'écouter. Quant au second chatbot, Nous avons utilisé DialogFlow qui est avant tout une interface qui va nous permettre d'utiliser l'intelligence de Google. DialogFlow contient l'API Cloud Natural Language qui permet de reconnaître des phrases envoyées par l'utilisateur. Avec les phrases récupérées et un peu de machine learning, Google reconnaît la phrase, et lance en adéquation une action proposée par notre configuration.

Nous avons opté pour Firebase afin d'héberger notre second ChatBot. Firebase est un ensemble de services d'hébergement pour n'importe quel type d'application (Android, iOS, Javascript, Node.js, Java, Unity, PHP, C++ ...). Il propose d'héberger en NoSQL et en temps réel des bases de données, du contenu, de l'authentification sociale (Google, Facebook, Twitter et Github), et des notifications, ou encore des services tel qu'un serveur de communication temps réel.

Pourquoi construire un Chatbot pour votre entreprise? Par définition, un chatbot est un programme informatique qui interagit avec l'utilisateur via un texte. Chatbot est également connu comme talkbot, chatterbot, Bot, bot de messagerie instantanée, agent interactif ou Entité de conversation artificielle. Les Chatbots sont conçus pour avoir une interaction conviviale via une interface graphique qui a la capacité de capturer ou de suggérer automatiquement du texte. L'idée de chatbot a été inventée en 1960, cependant, il a fallu près de 50+ ans pour obtenir les chatbots dans le flux principal. Nous vivons à une époque où le trafic Internet des bots a dépassé le trafic généré par les humains en 2016. En même temps, les Chatbots sont équipés de l'intelligence artificielle et de l'apprentissage automatique qui peuvent être programmés pour apprendre automatiquement et répondre aux utilisateurs via une application de messagerie sur une

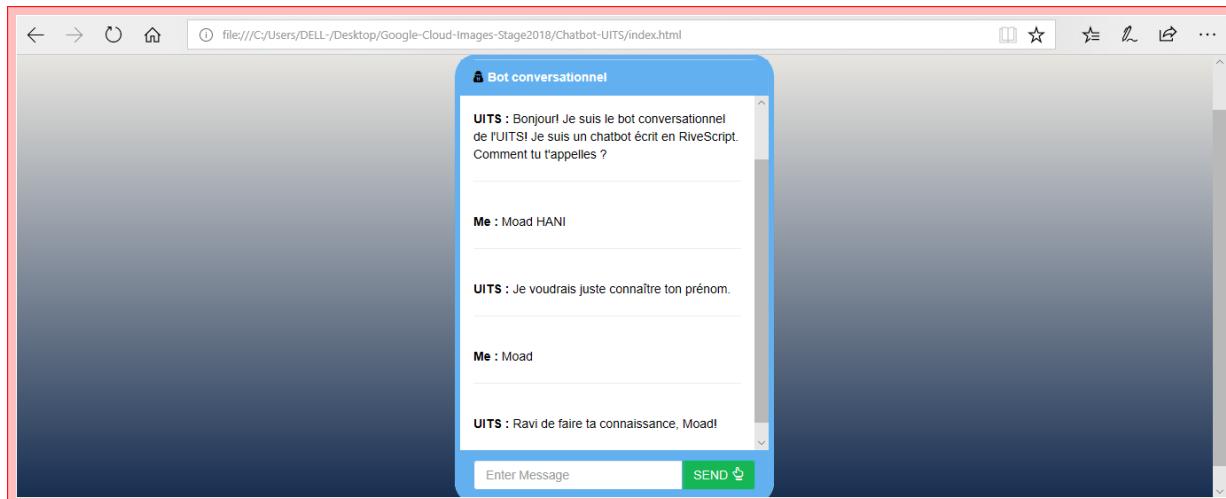
base de 24x7. Cette technologie révolutionnaire à venir permet aux entreprises d'avoir un robot communiquant avec les consommateurs et les prospects afin de déterminer leurs besoins et de les servir. Si le bot peut aider à répondre à des besoins spécifiques, un humain n'aura même pas besoin de s'impliquer!

Chatbots sont personnalisés et intégrés avec des applications de chat populaires (comme Facebook Messenger, WhatsApp, Slack, Instagram, Télégramme, Snapchat, Twitter, SMS) et divers CRM (comme Salesforce, Oracle, SAP, etc.). Les entreprises sont présentes là où les clients sont; que ce soit une interaction personnelle, un courriel, un texte mobile ou un média social. Les facteurs qui favorisent l'adoption des chatbots sont les suivants :

- 1. Accessibilité 24x7 sans perte d'efficacité
- 2. Capacité évolutive de gérer plusieurs transactions
- 3. Lancement rapide de diverses campagnes marketing basées sur le profil de l'utilisateur et le marché
- 4. Satisfaction client améliorée
- 5. Opportunité pour l'automatisation améliorée et l'intégration de bout en bout avec le reste de l'écosystème
- 6. Toujours sur le canal de vente et le soutien à la clientèle
- 7. Capacité d'apprentissage continu par l'interaction et la disponibilité des connaissances des clients
- 8. Intégration avec les médias sociaux
- 9. Meilleures génération et surveillance.
- 10. Capacité d'interagir dans plusieurs langues à l'échelle mondiale
- 11. Meilleure fidélité à la marque

Premier chatbot : Assister à distance les clients effectifs.

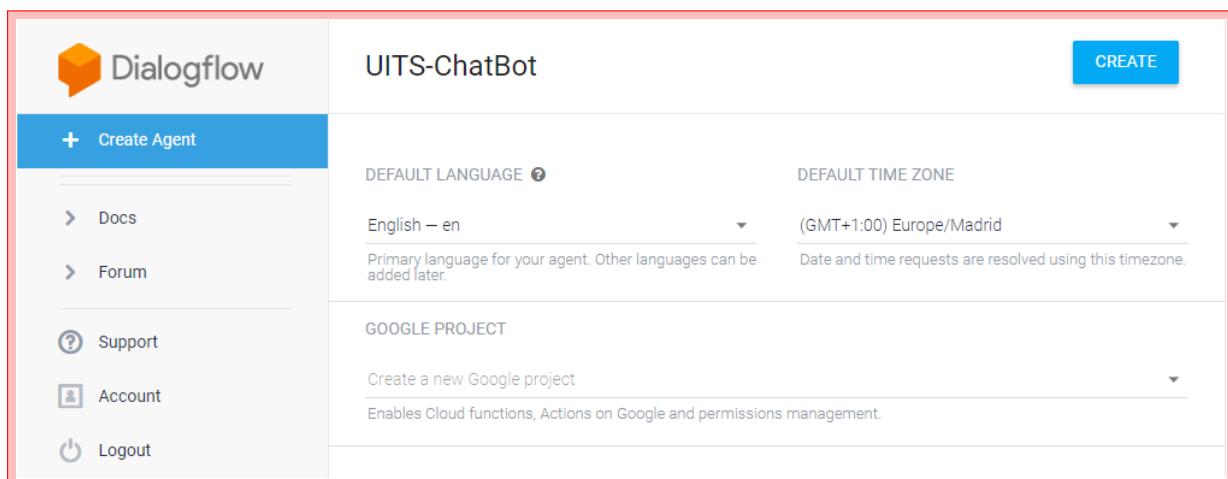
Nous avons utilisé les tutos de Rivescript pour créer le fichier "brain.rive". Nous avons également intégré p5.speech qui est une bibliothèque JavaScript qui fournit un accès simple et clair aux "API Web Speech Recognition" et "Speech Recognition", permettant de créer facilement des bots capables de dialoguer et d'écouter.



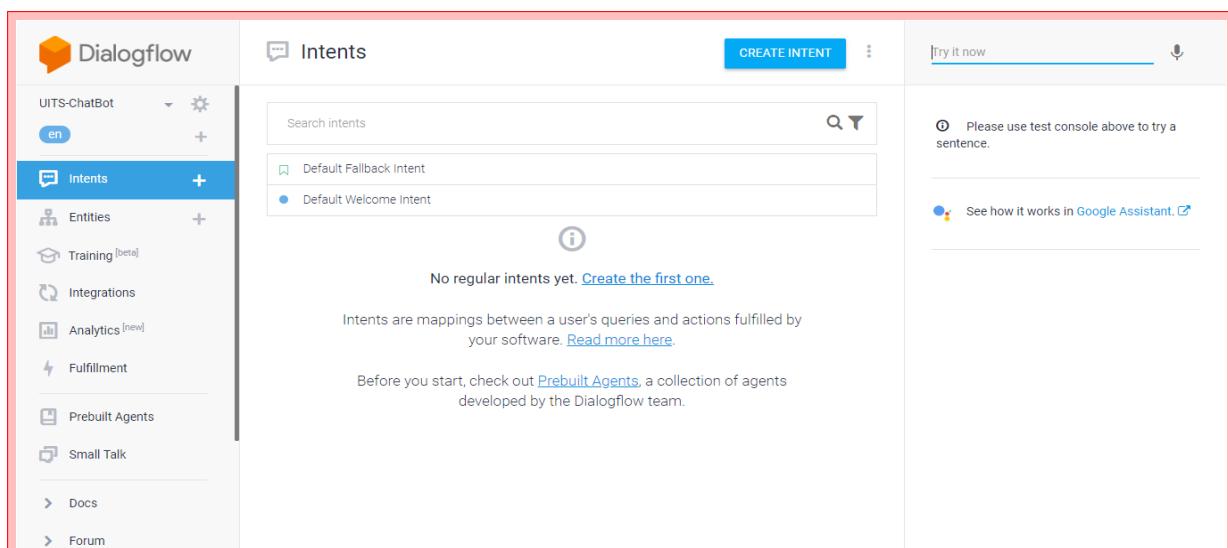
Second chatbot : Marketer les formations et services de l'UITS auprès des clients potentiels.

Dirigeons-nous vers Dialogflow et dans le coin supérieur droit, cliquons sur "Aller à la console".

Nous sommes invité à nous connecter avec notre compte Google. Nous nous connectons et autorisons Dialogflow à afficher et gérer nos données sur les services Google Cloud Platform. Acceptons les termes et nous devrions être accueilli avec une page de démarrage initial.



Console de dialogue Dialogflow : Nous Cliquons sur le bouton "Créer un agent". Dans Dialogflow, un agent désigne le chatbot que l'application iOS utilisera pour communiquer par voie hertzienne afin de recevoir des réponses. Nous remplissons le nom de l'agent (par exemple, UITS-Chatbot) et cliquons sur le bouton "Créer" pour continuer. Dialogflow va créer l'agent pour nous. Nous devrions avoir 2 intentions par défaut : une Default Welcome Intent et Default Fall-back Intent. Dans le volet de gauche, nous devrions voir les onglets pour les Intents et les entités.



```
C:\Users\DELL->npm install -g firebase-tools
C:\Users\DELL-\AppData\Roaming\npm\firebase -> C:\Users\DELL-\AppData\Roaming\npm\node_modules\firebase-tools\bin\firebase
> @google-cloud/functions-emulator@1.0.0-beta.4 postinstall C:\Users\DELL-\AppData\Roaming\npm\node_modules\firebase-tools\
> node scripts/upgrade-warning

If you're using the Emulator via the Firebase CLI, you can
disregard this message.

If you're upgrading @google-cloud/functions-emulator, these
are the recommended upgrade steps:

1. Stop the currently running emulator, if any:
   functions stop

2. Uninstall the current emulator, if any:
   npm uninstall -g @google-cloud/functions-emulator

3. Install the new version of the emulator:
   npm install -g @google-cloud/functions-emulator

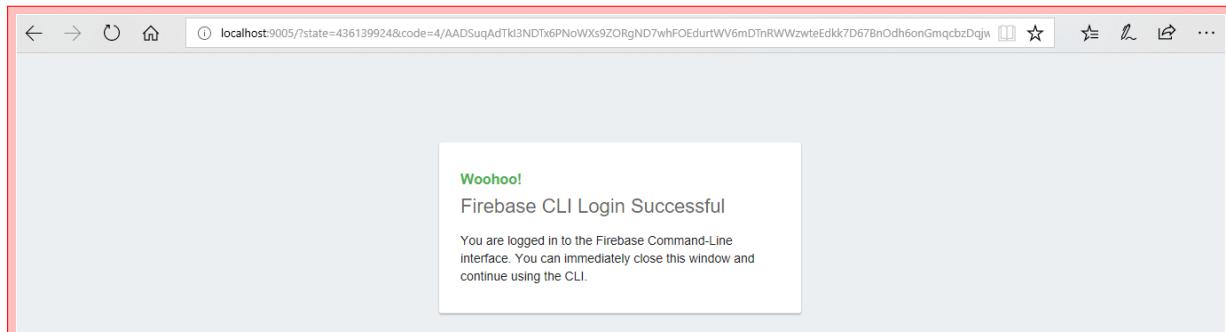
If you have trouble after upgrading, try deleting the config
directory found in:
   ~/.config/configstore/@google-cloud/functions-emulator

Then restart the emulator. You can also check for any renegade
Node.js emulator processes that may need to be killed:
   ps aux | grep node

+ firebase-tools@3.18.4
added 533 packages from 269 contributors in 328.56s

Update available 5.6.0 → 6.0.0
Run npm i -g npm to update
```

Nous avons opté pour Firebase afin d'héberger notre ChatBot. Firebase est un ensemble de services d'hébergement pour n'importe quel type d'application (Android, iOS, Javascript, Node.js, Java, Unity, PHP, C++ ...). Il propose d'héberger en NoSQL et en temps réel des bases de données, du contenu, de l'authentification sociale (Google, Facebook, Twitter et Github), et des notifications, ou encore des services tel qu'un serveur de communication temps réel.



```
C:\Users\DELL->cd "C:\Users\DELL-\Desktop\Google-Cloud-Images-Stage2018\Firebase Functions"
C:\Users\DELL-\Desktop\Google-Cloud-Images-Stage2018\Firebase Functions>firebase init functions
#####
##  ## ######  ## ##### ##  ##  ##  ## ######  ## #####
#####  ## ##### ##  ##  ## ##### ## ##### ## #####
##  ##  ##  ##  ##  ##  ##  ##  ##  ##  ##  ##  ##  ##
##  ##  ##  ##  ##  ##  ##  ##  ##  ##  ##  ##  ##  ##

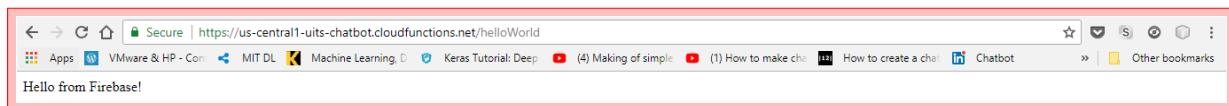
You're about to initialize a Firebase project in this directory:
  C:\Users\DELL-\Desktop\Google-Cloud-Images-Stage2018\Firebase Functions
? Are you ready to proceed? (Y/n)
```

Voici l'étape cruciale du déploiement de l'App par défaut.

```
C:\Users\DELL-\Desktop\Google-Cloud-Images-Stage2018\Firebase Functions>firebase deploy
== Deploying to 'uits-chatbot'...
i  deploying functions
Running command: npm --prefix "$RESOURCE_DIR" run lint
> functions@ lint C:\Users\DELL-\Desktop\Google-Cloud-Images-Stage2018\Firebase Functions\functions
> eslint .

+  functions: Finished running predeploy script.
+  functions: ensuring necessary APIs are enabled...
+  functions: all necessary APIs are enabled
+  functions: preparing functions directory for uploading...
+  functions: packaged functions (50.24 KB) for uploading
+  functions: functions folder uploaded successfully
+  functions: creating function helloworld...
+  functions[helloworld]: Successful create operation.
Function URL (helloworld): https://us-central1-uits-chatbot.cloudfunctions.net/helloworld
+ Deploy complete!

Project Console: https://console.firebaseio.com/project/uits-chatbot/overview
```



Une action correspond à l'étape que prendra l'application lorsqu'une "intent" spécifique a été déclenchée par l'entrée d'un utilisateur. Les actions peuvent avoir des paramètres pour extraire des informations des demandes des utilisateurs et apparaîtront dans le format suivant dans une réponse JSON :

"Action" : "nom_action"

"Nom_paramtre" : "valeur_paramtre"

Le nom de l'action et ses paramètres sont définis dans la section Action d'une "intent". Par exemple, si nous créons une application pour l'envoi de messages, la section Action aura le nom de l'action, ainsi que toutes les valeurs de paramètres définies automatiquement ou ajoutées manuellement.

Les paramètres sont des éléments généralement utilisés pour connecter des mots dans la réponse d'un utilisateur, à des entités. Dans les réponses JSON à une requête, les paramètres sont renvoyés au format suivant :

"Nom_paramtre" : "valeur_paramtre"

Action and parameters 

ChatBot

REQUIRED 	PARAMETER NAME 	ENTITY 	VALUE	IS LIST 	PROMPTS 
<input checked="" type="checkbox"/>	number	@sys.number	Snumber	<input checked="" type="checkbox"/>	Professional tr...
<input checked="" type="checkbox"/>	email	@sys.emai	Semail	<input checked="" type="checkbox"/>	Please tell me ...
<input checked="" type="checkbox"/>	given-name	@sys.give	Sgiven-na	<input checked="" type="checkbox"/>	Define prompts...
<input checked="" type="checkbox"/>	TrainingType	@Training	STrainingT	<input checked="" type="checkbox"/>	Define prompts...
<input type="checkbox"/>	Enter name	Enter entity	Enter value	<input type="checkbox"/>	—

+ New parameter

Les paramètres apparaissent dans deux zones différentes. Dans la section Phrases d'apprentissage, les paramètres liés aux entités connues seront mis en surbrillance (annotés) après l'ajout d'un exemple. En cliquant sur un mot annoté dans un exemple, nous découvrirons une table avec des données sur l'entité choisie.

Une table des paramètres dans l'intent, se trouve dans la section Action et Paramètres de la page "d'intents". Ce tableau représente tous les paramètres utilisés dans tous les exemples d'expressions.

Training phrases 

Add user expression

Search training phra 

“ We want to take part in the professional training of <u>CCNA RS</u>	X	
PARAMETER NAME	ENTITY	RESOLVED VALUE
TrainingType	@TrainingType	CCNA RS
“ My email is abc@abc.com		
“ For <u>3</u> persons, please !		
“ My name is <u>John</u>		

Nous devons diversifier les réponses pour chaque type de questions pour renforcer l'effet du réalisme. (Ici, le bot choisira une des trois propositions).

Prompts for "number"

NAME	ENTITY	VALUE
number	@sys.number	\$number

PROMPTS

- 1 Professional training for how many persons ?
- 2 For how many people you want the professional training ?
- 3 How many people you are ?
- 4 Enter a prompt variant

Close

Nous rédigeons la dernière phrase pour clôturer la conversation.

Responses ?

DEFAULT GOOGLE ASSISTANT +

Text response ? trash

- 1 You will be notified that your application form has been processed. All the necessary information concerning the Professional Training will be sent to your email. Have a nice day !
- 2 Enter a text response variant

Add Responses

Set this intent as end of conversation ?

Déploiement dans Slack

Slack fonctionne à la manière d'un chat IRC organisé en canaux correspondant à autant de sujets de discussion. La plateforme permet également de conserver une trace de tous les échanges (« Slack » est l'acronyme de « Searchable Log of All Conversation and Knowledge »), permet le partage de fichiers au sein des conversations et intègre en leur sein des services externes comme GitHub, Dropbox, Google Drive ou encore Heroku pour centraliser le suivi et la gestion d'un projet. Un robot peut également répondre automatiquement à certaines requêtes de l'utilisateur et s'améliore au fil du temps grâce à des algorithmes d'apprentissage. Slack dispose de clients natifs sur la plupart des plateformes mobiles (iOS, Android, Windows Phone) ainsi que sur macOS, Windows, Linux, et via un navigateur internet.

Créer un espace de travail

Veuillez confirmer votre adresse e-mail afin de créer un tout nouvel espace de travail.

moad11320347@gmail.com **Confirmer**

Quelle URL souhaitez-vous pour votre espace de travail Slack ?

Choisissez l'adresse que vous utiliserez pour vous connecter à Slack.

URL de votre espace de travail (lettres, chiffres et tirets seulement)

UITS-Chatbot .slack.com

Créer un espace de travail →

Faculté des Scie... Aujourd'hui

Customer Hello

Dialogflow Bot APPLICATION 8 h 17 Greetings! How Can I Help You?

Customer 8 h 18 We want to take part in the professional training of CCNA RS.

Dialogflow Bot APPLICATION 8 h 18 Professional training for how many persons ?

Customer 8 h 18 For 5 persons, please.

Dialogflow Bot APPLICATION 8 h 18 Please tell me your email for further contact

Customer 8 h 19 David.H@gmail.com

Dialogflow Bot APPLICATION 8 h 19 What is the given-name?

Customer 8 h 19 My name is David

Dialogflow Bot APPLICATION 8 h 19 You will be notified that your application form has been processed. All the necessary information concerning the Professional Training will be sent to your email. Have a nice day !

Envoyer un message à @Dialogflow Bot

HANI MOAD •

Modifier le profil

Améliorations : Plus l'ordre se rapprochera d'une question ouverte, plus le processus de réponse sera compliqué. Avec la technologie actuellement disponible sur le marché, les seront parfaits pour répondre à des demandes directes, formulées simplement, mais toutes les requêtes des clients ne respectent pas forcément ce critère. C'est la raison pour laquelle nous avons décidé de combiner python et rivescript pour que le bot conversationnel puisse répondre à plusieurs questions d'ordre général et par conséquent puisse passer pour un bot INTELLIGENT!!

1. Date, mois, heure et culture :



```

C:\Invite de commandes - python rivescript ./eg/brain
Microsoft Windows [version 10.0.17134.112]
(C) 2018 Microsoft Corporation. Tous droits réservés.
C:\Users\DELL->cd C:\Users\DELL\Desktop\Google-Cloud-Images-Stage2018\Last Project\rivescript\rivescript
C:\Users\DELL\Desktop\Google-Cloud-Images-Stage2018\Last Project\rivescript>python rivescript ./eg/brain
.....: RiveScript Interpreter (Python)
.....: Library Version: v1.14.9
.....: Type '/quit' to quit.
.....: Type '/help' for more options.
:
Using the RiveScript bot found in: ./eg/brain
Type a message to the bot and press Return to send it.

You> Quel mois sommes nous ?
Bot> on est en Juin

You> Qui est Ronaldo ?
Bot> Cristiano Ronaldo dos Santos Aveiro, couramment appelé Cristiano Ronaldo et surnommé CR7, né le 5 février 1985 à Funchal sur l'île de Madère, est un footballeur international portugais. Considéré comme l'un des meilleurs joueurs du monde, il remporte le Ballon d'or en 2008, 2013, 2014, 2016 et 2017. Auteur de plus de 650 buts en carrière, il est le meilleur buteur de l'histoire de la Ligue des champions, en coupes d'Europe, du Real Madrid, du derby madrilène, de la Coupe du monde des clubs et de la sélection portugaise, dont il est le capitaine depuis 2007.

```

Les scripts de cette première amélioration :

```

> object annee python
    import datetime
    dat = datetime.date.today()
    LAnnee = dat.year
    return LAnnee
< object

> object mois python
    import datetime
    Mois=['Janvier','Fevrier','Mars','Avril','Mai','Juin','Juillet','Aout','Septembre','Octobre','Novembre','Decembre']
    dat = datetime.date.today()
    mois = dat.month
    LeMois = Mois[mois-1]
    return LeMois
< object

> object jour python
    import time
    a = time.strftime('%a') # ex: Mon
    b = time.strftime('%d') # ex: 6
    c = a.replace("Sun","Dimanche").replace("Mon","Lundi").replace("Tue","Mardi").replace("Wed","Mercredi").replace("Thu","Jeudi")
    .replace("Fri","Vendredi").replace("Sat","Samedi")
    jour = c+" "+b
    return jour
< object

> object heure python
    import time
    Heure = time.strftime('%H heures %M minutes et %S secondes',time.localtime())
    return Heure
< object

```

```

> object add python
    a, b = args
    return int(a) + int(b)
< object

> object Multiplication python
    d, c = args
    return int(d) * int(c)
< object

+ what is \# plus \#
- <star1> + <star2> = <call>add <star1> <star2></call>

+ combien font \# fois \#
- <star1> * <star2> = <call>Multiplication <star1> <star2></call>

```

```

//ex: on est quel jour UITS-Chatbot
//ex: en quelle annee sommes nous s'il te plait
//ex: en quel mois sommes nous
//ex: il est quelle heure là.....

+ [en] quelle annee (sommes nous|on est) [*]
- {random}Nous sommes|on est{/random} en <call>annee</call>

+ [en] quel mois (sommes nous|on est) [*]
- {random}Nous sommes|on est{/random} en <call>mois</call>

+ * quelle heure est il [*]
- il est <call>heure</call>

```

2. Opérations mathématiques :

```

You> What is 1245 + 165468685 ?
Bot> 1245 + 165468685 = 165469930
You> Combien font 1363532 fois 565656 ?
Bot> 1363532 * 565656 = 771290056992
You>

```

Le script de cette deuxième amélioration :

```

> object add python
  a, b = args
  return int(a) + int(b)
< object

> object Multiplication python
  d, c = args
  return int(d) * int(c)
< object

+ what is \# plus \#
- <star1> + <star2> = <call>add <star1> <star2></call>

+ combien font \# fois \#
- <star1> * <star2> = <call>Multiplication <star1> <star2></call>

```

3. Chercher une information intéressante et pouvoir la transmettre :

You: Dis moi quelque chose
Bot: Boirargues est un quartier de la commune française de Lattes au sud-est de Montpellier dans l'Hérault, ses habitants sont les Boirarguais. En 1999, ce quartier comptait plus de 1 800 habitants.

-- Géographie --
Boirargues se situe sur la rive gauche de la Lironde qui sépare Lattes centre et Boirargues.
You: Dis moi quelque chose
Bot: Boopsoidea inornata est une espèce de poisson marin de la famille des Sparidae. C'est la seule espèce du genre Boopsoidea.
You: Dis moi quelque chose
Bot: En droit des obligations en France, l'opposabilité du contrat aux tiers est un principe selon lequel le contrat crée une situation juridique que les tiers ne peuvent ignorer et qu'ils doivent même respecter en tant que fait juridique. Un contrat est supposé ne pas nuire aux tiers et ne leur profiter que dans des cas légalement déterminés. Cependant, le principe reste l'effet relatif des contrats, c'est-à-dire, l'inopposabilité des contrats aux tiers.
You: Dis moi quelque chose
Bot: Turdoides affinis

Le Cratérope affin ou Cratérope à bec jaune Turdoides affinis est une espèce de passereau appartenant à la famille des Leiothrichidae.
Le Cratérope affin vit dans le sud du Sri Lanka et en Inde. Il habite la brousse et les cultures.

You: Dis moi quelque chose
Bot: L'écomusée de la bourrine du Bois-Juquaud est une structure muséale française située sur la commune de Saint-Hilaire-de-Riez, dans le département de la Vendée et la région des Pays-de-la-Loire.
Constitué autour d'une bourrine de rive, habitation traditionnelle du Marais breton, l'écomusée est dédié à la présentation et à la description de savoir-faire et de pratiques coutumières des pays maraîchins du nord-ouest vendéen aux XIXe et XXe siècles.

-- Description --

-- Localisation --
L'écomusée se situe au nord-ouest de la Vendée.

Les scripts de cette troisième amélioration :

```

//ex: recherche sur internet einstein

+ (qui est) *
- <call>WikiPerson <star2></call>

> object WikiPerson python
    import wikipedia
    search = ""
    wikipedia.set_lang("fr")
    wordSearch = (args)
    wordSearch2 = str(wordSearch)
    search = wordSearch2.encode('utf8')
    FirstSearch = wikipedia.search((search), results=1)
    modWord2 = str(FirstSearch)
    Search = wikipedia.summary(modWord2, sentences=3)
    return Search
< object

```

```
+ [UICTS-Chatbot] (dis|raconte) moi [une chose|quelque chose][UICTS-Chatbot]
- <call>WikiRandom</call>

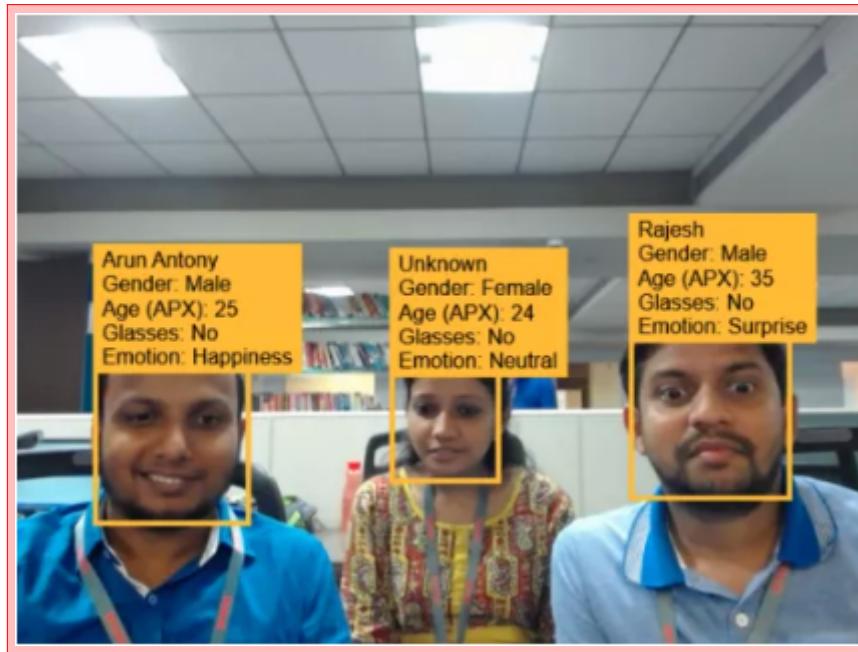
> object WikiRandom python
    import wikipedia
    wikipedia.set_lang("fr")
    wordSearch = wikipedia.random(pages=1)
    search = wordSearch.encode('utf8')
    try :
        Search = wikipedia.summary(search, sentences=3)
    except :
        a = 'Désolé, je ne sais pas quoi te dire'
        return (a)
    try:
        cleaning = Search.replace('()',').replace(')','').replace('/',', ') # remove parenthesis from text for better speaking
    except :
        pass
    try:
        a = cleaning.split('[')[0] # remove phonetics
        b = cleaning.split('[')[1]
        cleaned = a+b
        return (cleaned)
    except :
        try :
            return (cleaning)
        except :
            try :
                return (Search)
            except :
                a = 'Désolé, je ne sais pas quoi te dire'
                return (a)
< object
```

Le nombre de plateformes dont les marques ont besoin pour maintenir leur présence numérique est en constante augmentation. Cela a donné lieu à plusieurs campagnes de marketing qui ont été menées à travers différents canaux la nécessité de les garder régulièrement à jour peut être une tâche difficile. La taille de chacune de ces audiences a aussi considérablement augmenté. Beaucoup de marques ont une présence mondiale. En utilisant les Chatbots, les professionnels du marketing peuvent relier leurs comptes et les mettre à jour via un message depuis n'importe quelle plateforme Slack, SMS, Facebook Messenger, etc. Il est important pour les marketeurs de garder constamment un œil ouvert sur leurs indicateurs de performance de campagne pour repérer les différents changements. Cette donnée représente l'évolution des tendances par rapport à leur cible et donne aux marques un réel aperçu de la réaction de leurs consommateurs à certains messages et campagnes publicitaires. Etre attentif aux tendances peut aider à optimiser la stratégie de la marque. Les chatbots peuvent rendre cela plus facile et automatique. Les Chatbot aident à la fois au droit de réponses des consommateurs ainsi qu'au traitement instantané de la demande. Le nombre croissant d'entreprises apportant leurs produits et services en ligne a entraîné une vague de « droit de réponse » des consommateurs. Les utilisateurs attendent une réponse instantanée des marques à tous moments de la journée et à travers plusieurs fuseaux horaires. Bien que l'interaction manuelle avec un vaste public individuel peut être difficile, ne pas appliquer la personnalisation ou ne pas répondre dans le temps opportun peut également entraîner la perte de clients et une mauvaise expérience pour les consommateurs. Les bots permettent de s'engager avec le client dans un temps donné, au moins jusqu'à ce qu'un humain puisse intervenir si nécessaire.

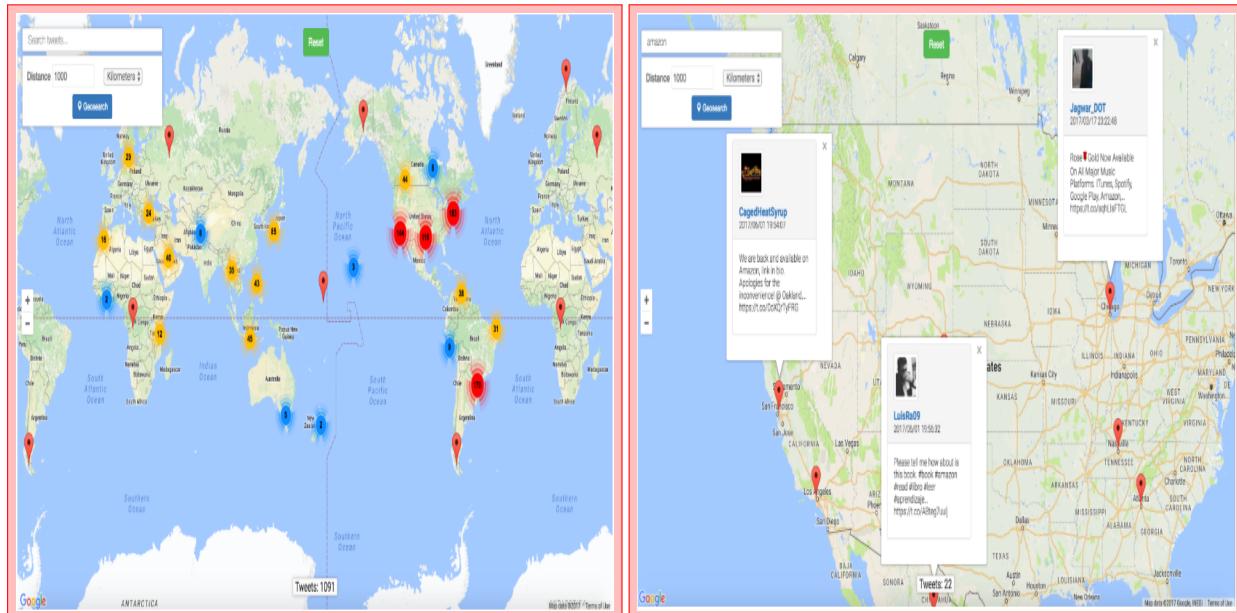
Conclusion et perspectives

Le Big Data et la Data Science à travers les réseaux sociaux est un sujet nouveau et passionnant. Nous avons vu, à travers nos différents exemples concrets comment les outils de Big Data vont aider les équipes de marketing digital, en l'occurrence L'UTS, à mieux comprendre les attentes des clients, notamment avec la croissance permanente des données. Le Big Data est devenu un outil essentiel pour l'aide à la décision stratégique en lui fournissant des données pertinentes. Il est à ce titre un élément différentiateur pour les entreprises, qui va leur permettre de développer des avantages concurrentiels. Nous avons vu par la suite le rapport entre le Big Data et les réseaux sociaux pour le marketing. Ces derniers étant des lieux d'échanges où les utilisateurs cherchent à émettre de l'information qui permet de réaliser de la veille économique afin d'avoir des retours sur une marque, mais aussi de surveiller la réputation de celle-ci par le « Département Marketing » grâce à des outils qualitatifs de Big Data et à des savoir-faire d'éminents Data Scientists, afin d'améliorer l'efficience du processus de vente d'une entreprise. En effet, il y a de plus en plus d'intervenants qui produisent de plus en plus de données sur les réseaux sociaux. Il est de ce fait difficile de trouver des informations pertinentes pour guider les décideurs dans leurs choix stratégiques et prévoir en continu les réactions de la concurrence qui pourrait déstabiliser le management. L'acquisition et le traitement de ces données a un coût, d'où la nécessité de s'interroger sur le bien-fondé d'une démarche Big Data. Ces hypothèses ont été validées, si je puis dire, à travers cet exposé. En effet, il ressort de nos explications que les projets Big Data au sein des entreprises sont essentiels à leur développement. Nonobstant l'investissement en matériel, logiciel et RH, il s'avère que l'enjeu vaut la chandelle.

Bien que l'aspect sécuritaire n'a pas été abordé dans la mise en oeuvre de nos architectures Big Data, il ne demeure pas moins intéressant et requiert un développement spécifique à chaque application. L'on peut aussi élargir l'expérience de la détection d'humeur dans les expressions faciales à d'autres aspects plus élaborés tels que la détection du genre, de l'âge, des émotions et l'analyse et la mémorisation faciales.



En ce qui concerne l'analyse des Tweets, il est possible de l'enrichir par une application de géo-localisation des tweets en temps réel.



Il faut savoir qu'il y a deux types de commentaires dans les réseaux sociaux : les commentaires négatifs constructifs et les commentaires négatifs inoffensifs. Les commentaires négatifs inoffensifs, pourraient toucher une opinion non flatteuse (décor, l'emplacement...) mais sans impact sur le commerce. Il n'y a pas besoin de répondre à ce message et surtout pas le supprimer. Il est à prendre en considération. Les commentaires négatifs constructifs peuvent créer une crise. Un mauvais service, un mauvais accueil... Il faut, dans ce cas, agir au plus vite en répondant au commentaire sur le réseau social concerné. Il faut surtout démontrer le professionnalisme en restant courtois. Lorsqu'on réagit à une crise, on le fait dabord

pour ceux qui nous regardent. Le but est de démontrer que nous gérons la crise et ensuite on dirige la personne lésée en mode privé soit par messagerie, par courriel ou au téléphone. Aux yeux de tous, on démontre notre professionnalisme et on évite ainsi à notre entreprise une argumentation publique.

Enfin, ce travail n'a pas la prétention de faire un exposé exhaustif sur ce domaine encore en friche et moi-même n'étant encore, aujourd'hui, qu'en phase d'apprentissage et de remise en cause. Cette ébauche n'est qu'approximative; Cela ressemble à un cas d'école pratiqué sur la startup Union IT Services. Malgré que je suis en terrain inconnu, la passion et le courage ne me manquent pas pour y aller de plein pied dans mon futur projet professionnel.

Annexes

Création de fichier en Rivescript pour bien informer les clients sur les services de L'UITS

```
! version = 2.0

+ salut
- Bonjour! Je suis le bot conversationnel de l'UITS! Je suis un chatbot écrit en RiveScript.\s
^ Comment tu t'appelles ?
+ _

- <set name=<formal>> Ravi de faire votre connaissance, <get name>!
^ Voulez-vous voir notre site, si oui :
^ <a href="http://www.union-it-services.com/"> cliquez ici </a>
^ si vous voulez voir notre page facebook
^ <a href="https://www.facebook.com/unionitervices"> c'est ici qu'il faut cliquer. </a>

+ *
- Je ne suis pas sûr de te comprendre pleinement.
- S'il vous plaît continuez.
- C'est intéressant. Continuez s'il vous plaît.
- Raconter-moi plus à ce sujet.

+ [*] formations [*]
- Voici nos formations : <a href=""><button class="button button2">
^ Catalogue des formations </button></a>
^ Quelle votre adresse email ?
+ _
- <set email=<formal>> Nous vous contacterons à cette adresse , <get email>!
^ c'est quoi votre numéro de téléphone ?
+ _
- <set number=<formal>> Super, c'est bien enregistré! Nous vous contacterons bientôt à ce numéro, <get
^ number>!

+ [*] merci [*]
- je vous en prie! Si vous voulez contacter Mr Khalid Katkout via gmail, il vous donnera plus
^ d'informations à ce sujet. <a href="mailto:moad11320347@gmail.com"><button class="button button1">
^ Send Email </button></a>
```

```
+ [*] merci [*]
- je vous en prie! Si vous voulez contacter Mr Khalid Katkout via gmail, il vous donnera plus
^ d'informations à ce sujet. <a href="mailto:moad11320347@gmail.com"><button class="button button1">
^ Send Email </button></a>

+ [*]@gmail.com
- nous vous contacterons à cette adresse, <star>!
^ c'est quoi votre numéro de téléphone ?

+ (05|06) [*]
- super, c'est bien enregistré! Nous vous contacterons bientôt à ce numéro <star>

+ mon numéro de téléphone est #
- on va vous appellez à <star>.

- quel est votre email ?
+ get email

+ *
- votre email est: "<star>" - Est-ce correct ? <set email=<star>>

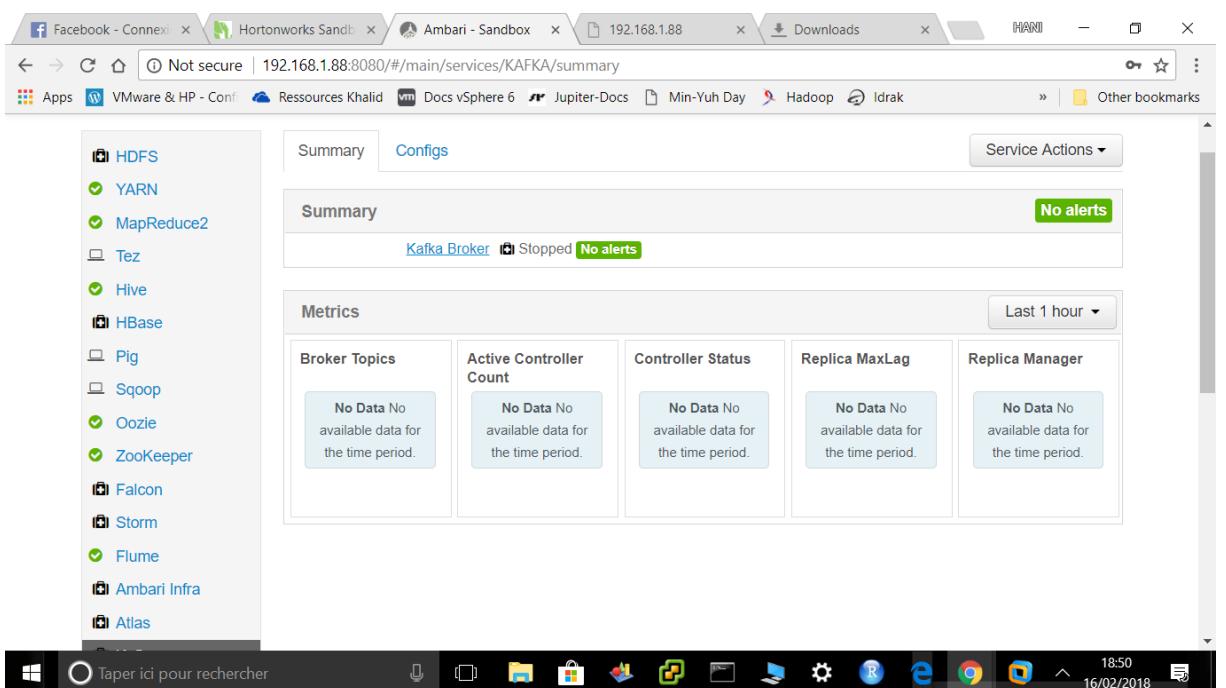
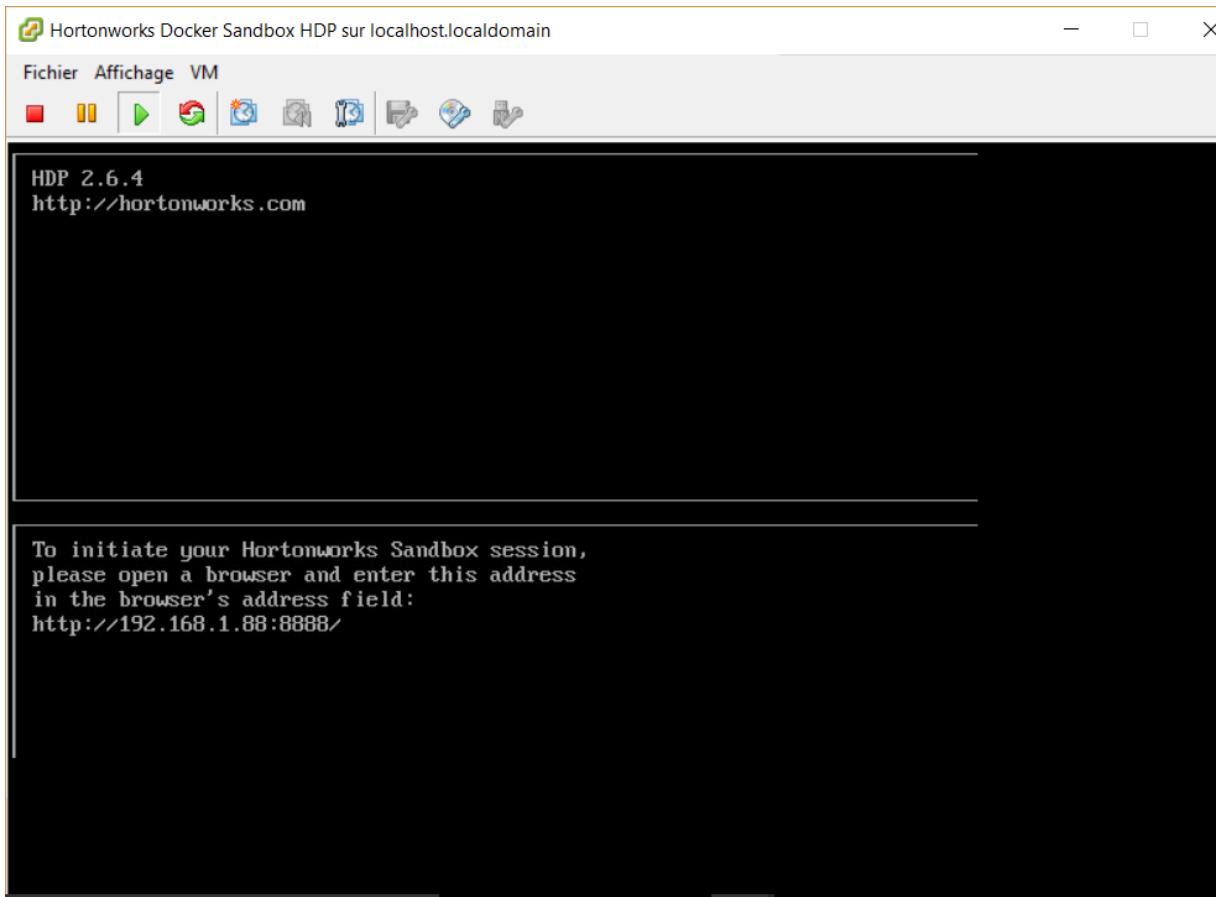
+ oui
- parfait, votre email est pris pour <get email>.

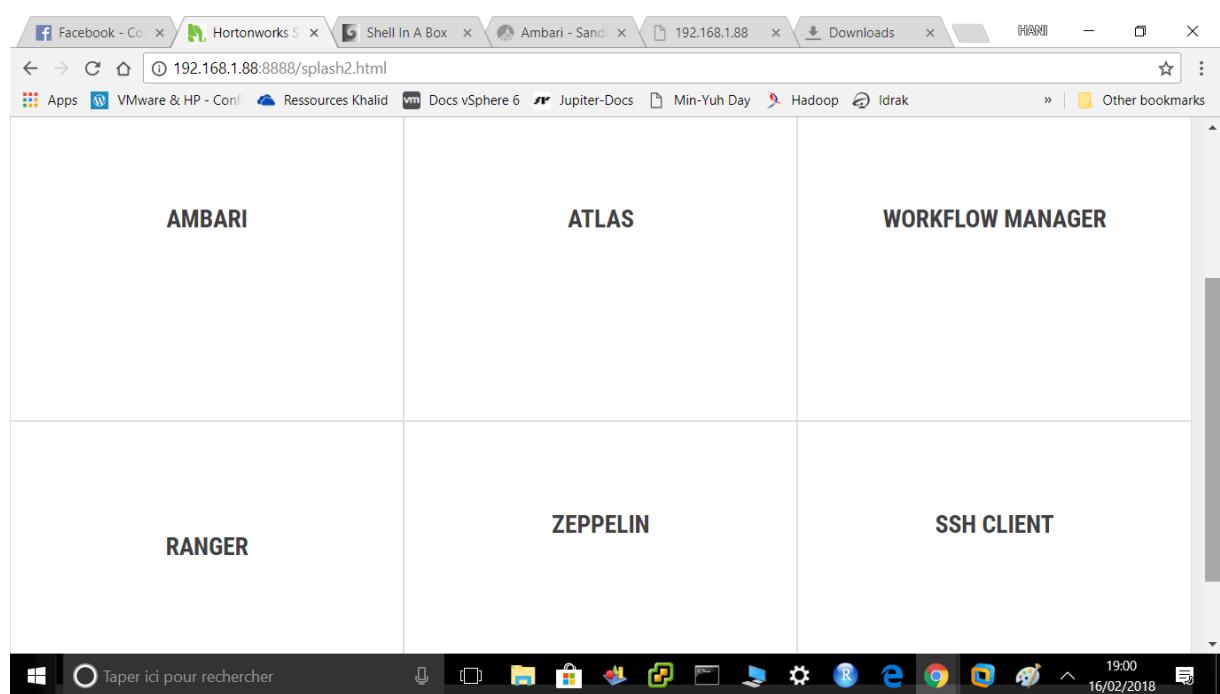
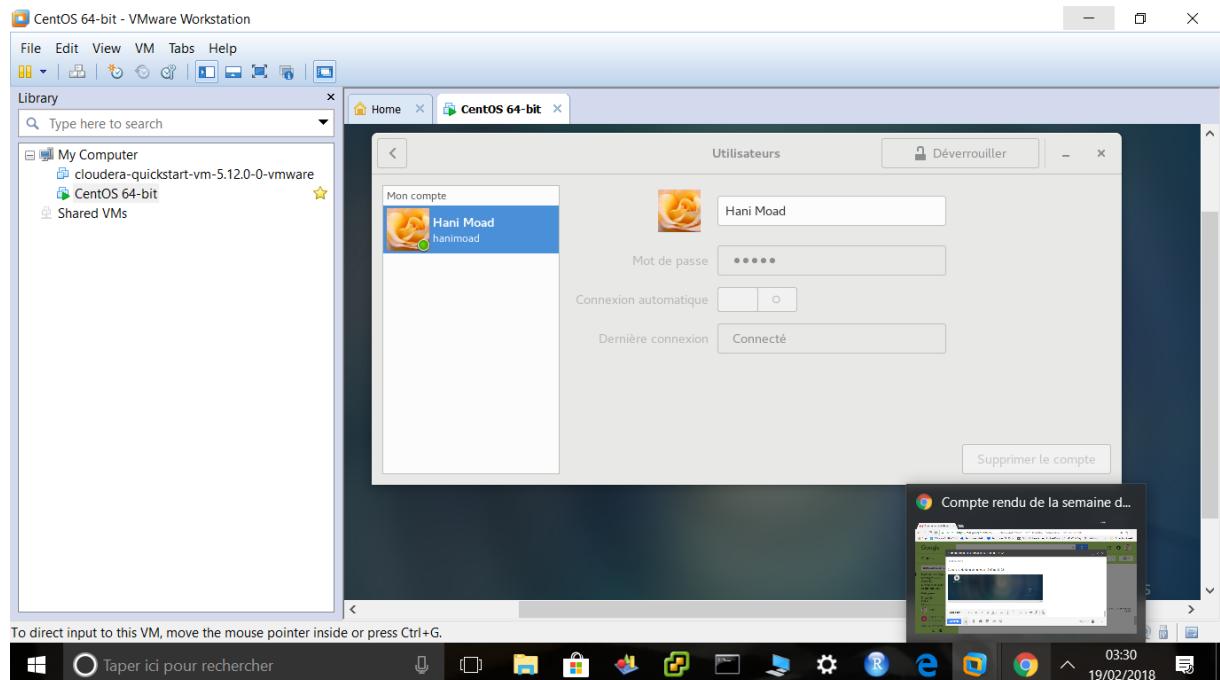
+ no
- okay, réessayez. Exemple: youremail@email.com <set email>

+ hello
- Hey there! <send> How are you?

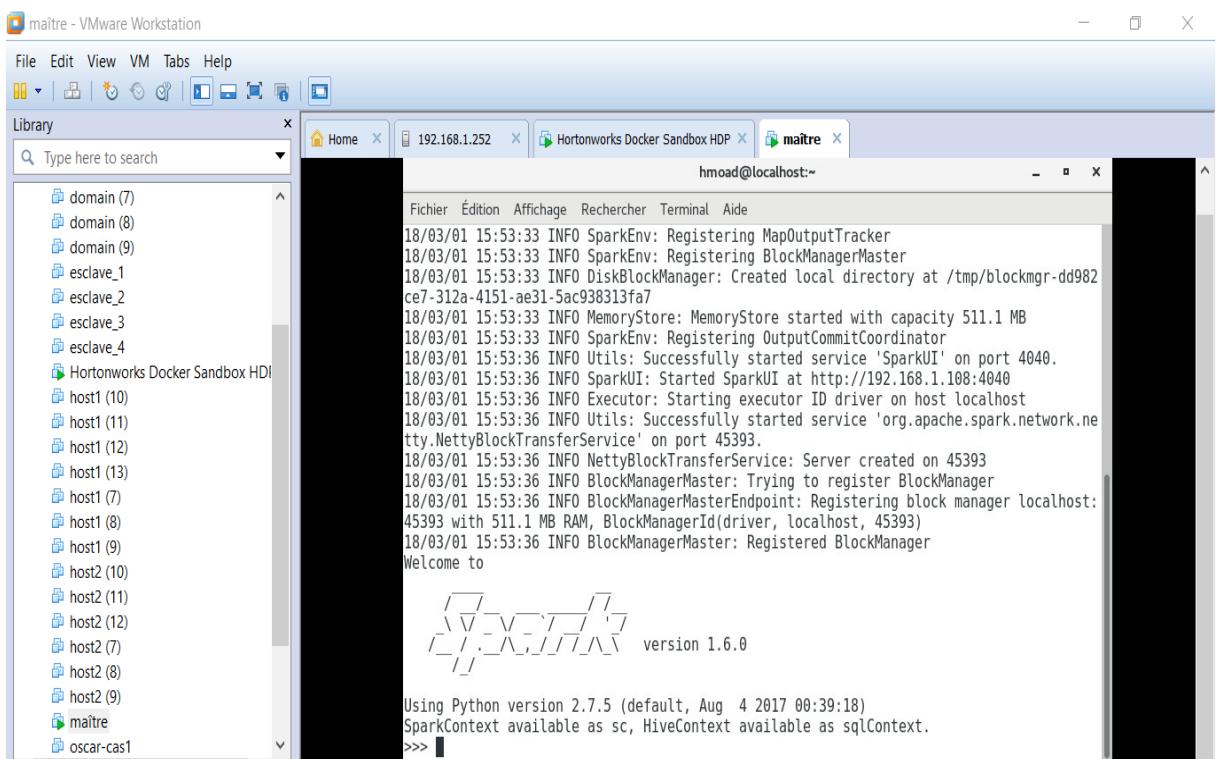
+ * inscrire *
- OK, merci de nous indiquer votre email, numéro de téléphone .{topic=inscription}
```

Installation et configuration de Hortonworks sur VMware dans un serveur à L'UTS





Installation et configuration d'un cluster Spark multi-noeuds sur le Data Center de L'UITS



Installation et configuration d'un cluster Hadoop multi-noeuds sur le Data Center de L'UTS

The screenshot shows the Hadoop Web UI interface. At the top, there's a navigation bar with links for Applications, Emplacements, Navigateur web Fire..., and a timestamp ven. 12:29. Below the bar, the title is "All Applications - Mozilla Firefox". The address bar shows "localhost.localdomain3:". The main content area features the Hadoop logo. On the left, a sidebar menu includes "Cluster" (About, Nodes, Node Labels, Applications, NEW, NEW_SAVING, SUBMITTED, ACCEPTED, RUNNING, FINISHED, FAILED, KILLED), "Scheduler", and "Tools". The main panel displays "Cluster Metrics" with a table showing 0 for all metrics: Apps Submitted, Apps Pending, Apps Running, Apps Completed, and Capacity. It also shows "Cluster Nodes Metrics" with 1 Active Node and 0 Decommissioning Nodes. Under "Scheduler Metrics", it shows a Capacity Scheduler with [MEMORY] as the Scheduling Resource Type. A table below lists 20 entries for Application ID, User, Name, Application Type, Queue, Application Priority, Start Time, Finish Time, and Status.

The screenshot shows the Hadoop Overview page. The header has a green background with the word "Hadoop" and a navigation bar with tabs: Overview (which is active), Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, and Utilities (with a dropdown arrow).

Overview 'localhost:9000' (active)

Started:	Fri Feb 23 12:44:56 +0000 2018
Version:	2.9.0, r756ebc8394e473ac25feac05fa493f6d612e6c50
Compiled:	Mon Nov 13 23:15:00 +0000 2017 by arsuresh from branch-2.9.0
Cluster ID:	CID-72dd4882-c83f-4a6f-886e-37370bdb50c7

Configuration des Pare-feux dans l'installation de Cloudera sur Google Cloud

Détails de la règle de pare-feu

Réseau: default

Priorité: 65534

Action en cas de correspondance: Autoriser

Filtres sources: Plages d'adresses IP: 10.128.0.0/9

Protocoles et ports: tcp:0-65535, udp:0-65535, icmp

Requête/Réponse REST équivalente

Détails de la règle de pare-feu

Direction: Entrée

Action en cas de correspondance: Autoriser

Cibles: Toutes les instances du réseau

Filtre source: Plages d'adresses IP: 0.0.0.0/0

Plages d'adresses IP sources: 0.0.0.0/0

Deuxième filtre source: Aucun

Protocoles et ports: Protocoles et ports spécifiés: tcp:0-65535, udp:0-65535, icmp

Enregistrer

Règles de pare-feu

Nom	Type	Cibles	Filtres	Protocoles/Ports	Action	Priorité	Réseau
default-allow-internal	Ingress	Appliquer à tous	Plages d'adresses IP: 0.0.0.0/0	tcp:0-65535, udp:0-65535, 1 autres	Allow	65534	default
default-allow-ssh	Ingress	Appliquer à tous	Plages d'adresses IP: 0.0.0.0/0	tcp:22	Allow	65534	default
default-allow-https	Ingress	https-server	Plages d'adresses IP: 0.0.0.0/0	tcp:443	Allow	1000	default
default-allow-icmp	Ingress	Appliquer à tous	Plages d'adresses IP: 0.0.0.0/0	icmp	Allow	65534	default
default-allow-rdp	Ingress	Appliquer à tous	Plages d'adresses IP: 0.0.0.0/0	tcp:3389	Allow	65534	default
default-allow-http	Ingress	http-server	Plages d'adresses IP: 0.0.0.0/0	tcp:80	Allow	1000	default

Explication de la méthode de Viola et Jones

La méthode de Viola et Jones consiste à balayer une image à l'aide d'une fenêtre de détection de taille initiale 24px par 24px (dans l'algorithme original) et de déterminer si un visage y est présent. Lorsque l'image a été parcourue entièrement, la taille de la fenêtre est augmentée et le balayage recommence, jusqu'à ce que la fenêtre fasse la taille de l'image. L'augmentation de la taille de la fenêtre se fait par un facteur multiplicatif de 1,25. Le balayage, quant à lui, consiste simplement à décaler la fenêtre d'un pixel. Ce décalage peut être changé afin d'accélérer le processus, mais un décalage d'un pixel assure une précision maximale.

Cette méthode est une approche basée sur l'apparence, qui consiste à parcourir l'ensemble de l'image en calculant un certain nombre de caractéristiques dans des zones rectangulaires qui se chevauchent. Elle a la particularité d'utiliser des caractéristiques très simples mais très nombreuses.

Il existe d'autres méthodes mais celle de Viola et Jones est la plus performante à l'heure actuelle. Ce qui la différencie des autres est notamment :

- L'utilisation d'images intégrales qui permettent de calculer plus rapidement les caractéristiques
- La sélection par boosting des caractéristiques
- La combinaison en cascade de classifieurs boostés, apportant un net gain de temps d'exécution

Une caractéristique est une représentation synthétique et informative, calculée à partir des valeurs des pixels. Les caractéristiques utilisées ici sont les caractéristiques pseudo-haar. Elles sont calculées par la différence des sommes de pixels de deux ou plusieurs zones rectangulaires adjacentes.

Penons un exemple. Voici deux zones rectangulaires adjacentes, la première en blanc, la deuxième en noire :



(1) (2)

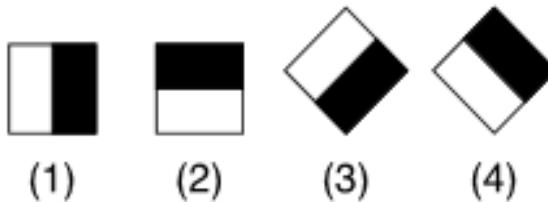
Les caractéristiques seraient calculées en soustrayant la somme des pixels noirs à la somme des pixels blancs.

Les caractéristiques sont calculées à toutes les positions et à toutes les échelles dans une fenêtre de détection de petite taille, typiquement de 24x24 pixels ou

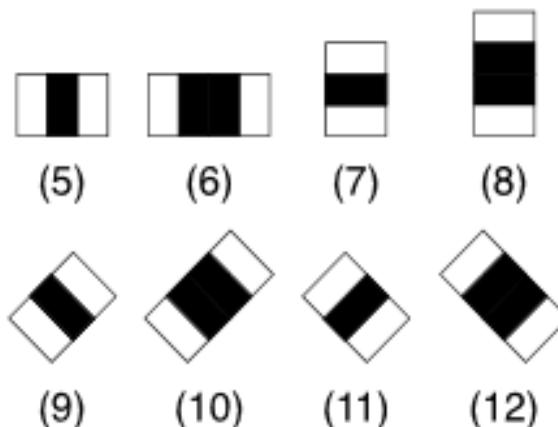
de 20x15 pixels. Un très grand nombre de caractéristiques par fenêtre est ainsi généré, Viola et Jones donnant l'exemple d'une fenêtre de taille 24 x 24 qui génère environ 160 000 caractéristiques.

L'image précédente présente des caractéristiques pseudo-haar à seulement deux caractéristiques mais il en existe d'autres, allant de 4 à 14, et avec différentes orientations.

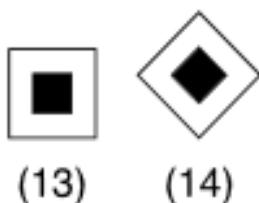
Caractéristiques de bord



Caractéristiques de ligne



Caractéristiques centre-pourtour



Malheureusement, le calcul de ces caractéristiques de manière "classique" coûte cher en terme de ressources processeur, c'est là qu'interviennent les images intégrales. Les images intégrales permettent de gagner du temps quant au calcul des caractéristiques. Il s'agit d'une image construite à partir de l'image d'origine, et de même taille qu'elle. Elle contient en chacun de ses points la somme des pixels situés au-dessus et à gauche du pixel courant. Le pixel rouge est égale à la somme de tous les pixels bleu, soient ceux à gauche et au dessus. Prenons un exemple. Nous souhaitons calcuer la sommes des pixels de la zone rectangulaire ABCD suivante :

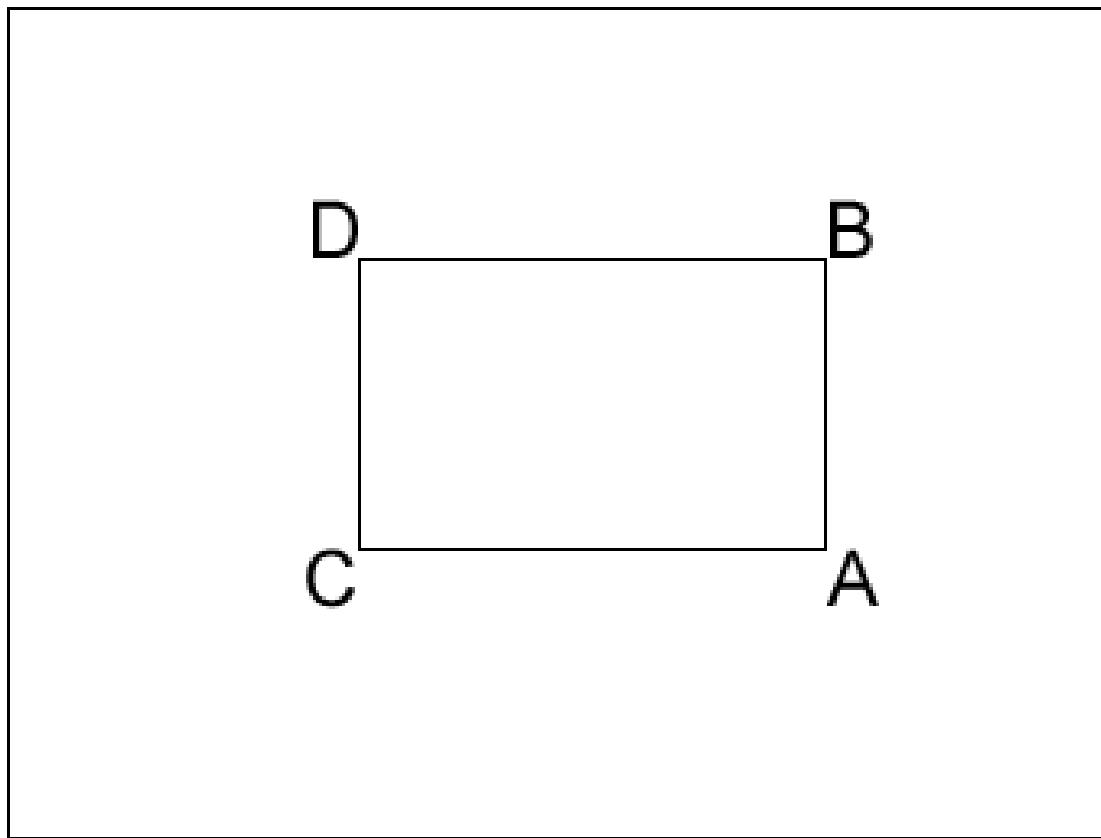
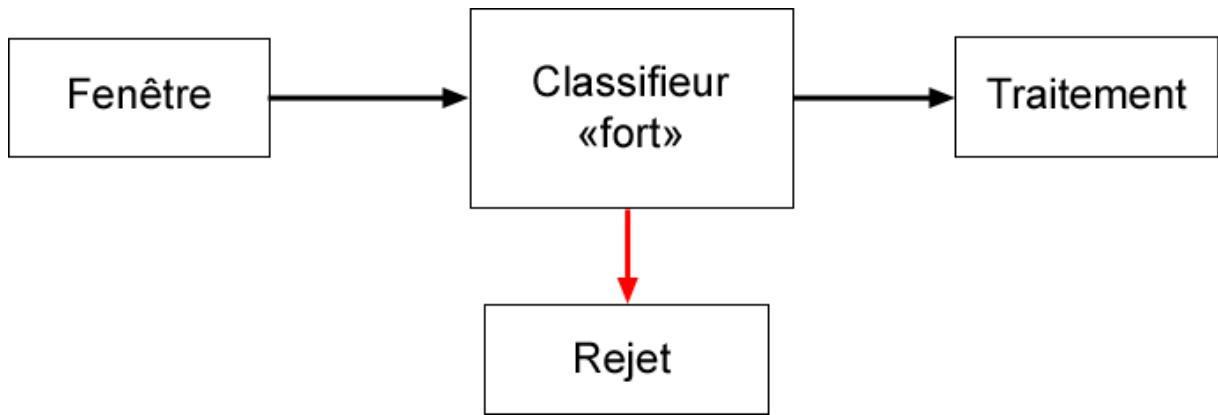
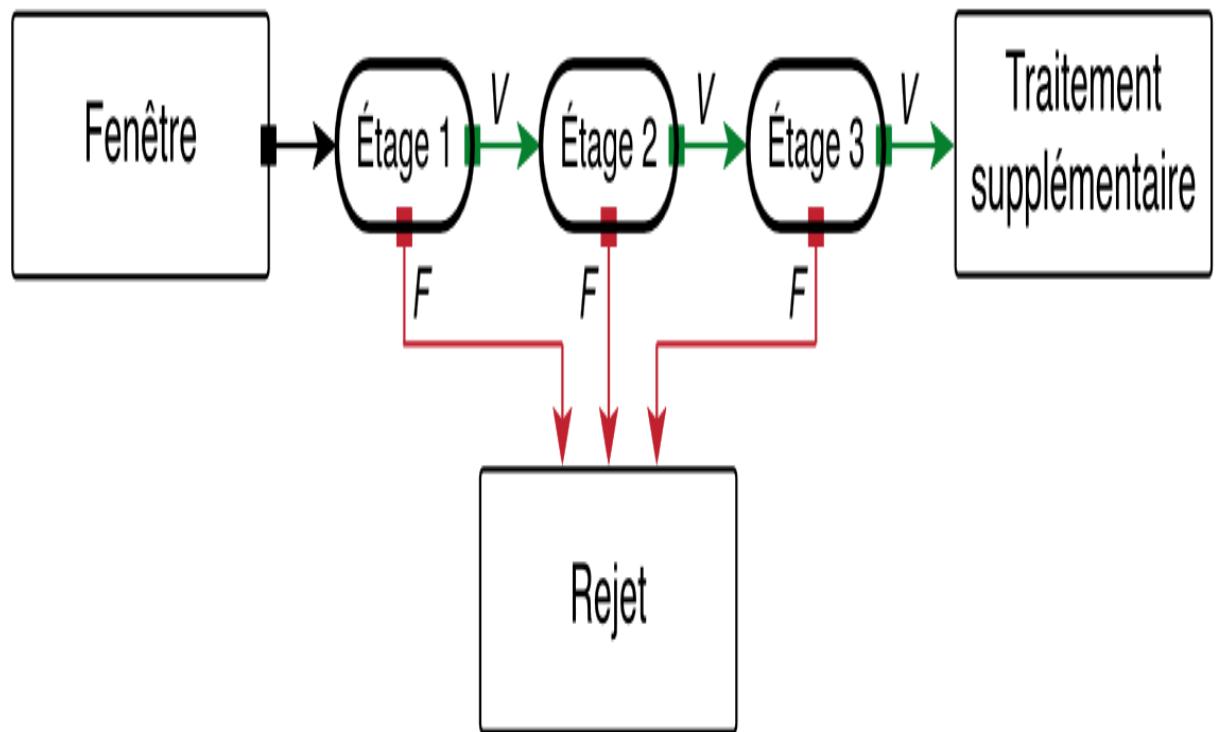


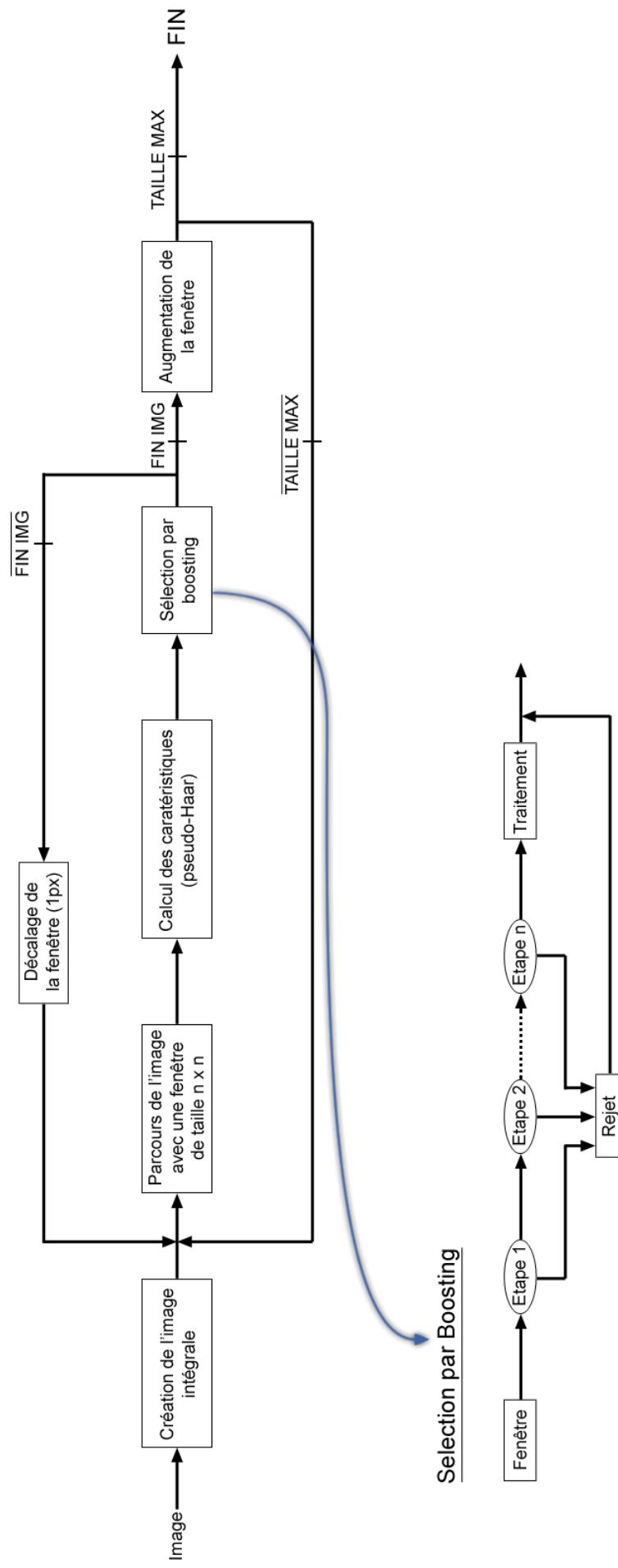
Image intégrale : calcul d'une zone rectangulaire Grâce à l'image intégrale, nous connaissons la valeur de la sommes des pixels en chacun des quatres points. Il suffit donc de faire : $A-B-C+D$. En seulement trois opérations nous avons réussi à calculer notre somme de pixels! Ainsi, on est en mesure de trouver la somme de pixels de n'importe quelle zone rectangulaire de l'image en seulement 3 opérations et 4 accès à l'image intégrale (un accès par point). Une caractéristique pseudo-Haar à deux rectangles peut alors être déterminée en seulement 6 accès (2 points sont partagés) à l'image, et une caractéristique à 3 rectangles en seulement 8 accès. Nous arrivons maintenant à la dernière partie concernant la théorie : la sélection par boosting. La sélection par boosting consiste à utiliser plusieurs classificateurs "faibles" mis en cascade plutôt que d'utiliser un seul classifieur "fort". En effet, avec un seul classifieur dit "fort" qui se présenterait de la sorte :



Il faudrait attendre que le classifieur est analysé toute la fenêtre afin de savoir si un visage est présent dans l'image ou non. Une mise en cascade de classificateurs dont le critère de sélection serait moins sévère se présenterait de la sorte :



Ainsi dès que l'un des étages estime qu'il n'y a pas de visage, la fenêtre est rejetée et l'algorithme passe à la suite ce qui permet un gain de temps considérable. Voici un beau schéma fonctionnel de l'algorithme :



Bibliographie

Auteurs	Articles, Livres
[Bahrainian and Dengel, 2013]	Bahrainian, S.-A. and Dengel, A. (2013). Sentiment analysis using sentiment features. In Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on, volume 3, pages 2629. IEEE.
[Beck et al., 2001]	Beck, K., Beedle, M., Van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeries, R., et al. (2001). Manifesto for agile software development.
[Biber, 1993]	Biber, D. (1993). Representativeness in corpus design. <i>Literary and linguistic computing</i> , 8(4) :243257.
[Cruz, 2013]	Cruz, C. D. (2013). Genes : a software package for analysis in experimental statistics and quantitative genetics. <i>Acta Scientiarum. Agronomy</i> , 35(3) :271276.
[Cruz et al., 2013]	Cruz, F. L., Troyano, J. A., Enríquez, F., Ortega, F. J., and Vallejo, C. G. (2013). 'long autonomy or long delay?'the importance of domain in opinion mining. <i>Expert Systems with Applications</i> , 40(8) :31743184.
[Dave et al., 2003]	Dave, K., Lawrence, S., and Pennock, D. M. (2003). Mining the peanut gallery : Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web, pages 519528. ACM.

[Davidov et al., 2010]	Davidov, D., Tsur, O., and Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in twitter and amazon. In Proceedings of the fourteenth conference on computational natural language learning, pages 107116. Association for Computational Linguistics.
[Deslandres, 2015]	Deslandres, L. (2015). Management de l'expérience client. Pearson.
[Fano and Wintringham, 1961]	Fano, R. M. and Wintringham, W. (1961). Transmission of information. Physics Today, 14 :56.
[Ganapathibhotla and Liu, 2008]	Ganapathibhotla, M. and Liu, B. (2008). Mining opinions in comparative sentences. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, pages 241248. Association for Computational Linguistics.
[González-Ibáñez et al., 2011]	González-Ibáñez, R., Muresan, S., and Wacholder, N. (2011). Identifying sarcasm in twitter : a closer look. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers-Volume 2, pages 581586. Association for Computational Linguistics.
[Haddi et al., 2013]	Haddi, E., Liu, X., and Shi, Y. (2013). The role of text pre-processing in sentiment analysis. Procedia Computer Science, 17 :2632.
[Hatzivassiloglou and McKeown, 1997]	Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics, pages 174181. Association for Computational Linguistics.

[Hatzivassiloglou and Wiebe, 2000]	Hatzivassiloglou, V. and Wiebe, J. M. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In Proceedings of the 18th conference on Computational linguistics Volume 1, pages 299305. Association for Computational Linguistics.
[Hu and Liu, 2004]	Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 168177. ACM.
[Jindal and Liu, 2006]	Jindal, N. and Liu, B. (2006). Identifying comparative sentences in text documents. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pages 244251. ACM.
[Jivani et al., 2011]	Jivani, A. G. et al. (2011). A comparative study of stemming algorithms. Int. J. Comp. Tech. Appl, 2(6) :19301938.
[Kaji and Kitsuregawa, 2007]	Kaji, N. and Kitsuregawa, M. (2007). Building lexicon for sentiment analysis from massive collection of html documents. In EMNLPCoNLL, pages 10751083.
[Kang et al., 2012]	Kang, H., Yoo, S. J., and Han, D. (2012). Sentilexicon and improved naïve bayes algorithms for sentiment analysis of restaurant reviews. Expert Systems with Applications, 39(5) :6000 6010.
[Kaplan and Haenlein, 2010]	Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. Business horizons, 53(1) :5968.
[Pang and Lee, 2005]	Pang, B. and Lee, L. (2005). Seeing stars : Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the 43rd annual meeting on association for computational linguistics, pages 115124. Association for Computational Linguistics.

[Pang et al., 2002]	Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? : sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pages 7986. Association for Computational Linguistics.
[Sinclair, 1996]	Sinclair, J. (1996). Preliminary recommendations on corpus typology. EAGLES Document TCWG-CTYP/P
[Wiebe et al., 1999]	Wiebe, J. M., Bruce, R. F., and O'Hara, T. P. (1999). Development and use of a gold-standard data set for subjectivity classifications. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, pages 246253. Association for Computational Linguistics.
[Wilson et al., 2005]	Wilson, T., Homann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Rilo, E., and Patwardhan, S. (2005). Opinionnder : A system for subjectivity analysis. In Proceedings of hlt/emnlp on interactive demonstrations, pages 3435. Association for Computational Linguistics.
[Xue and Zhou, 2009]	Xue, X.-B. and Zhou, Z.-H. (2009). Distributional features for text categorization. Knowledge and Data Engineering, IEEE Transactions on, 21(3) :428442.
[Yu and Hatzivassiloglou, 2003]	Yu, H. and Hatzivassiloglou, V. (2003). Towards answering opinion questions : Separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of the 2003 conference on Empirical methods in natural language processing, pages 129136. Association for Computational Linguistics.

Résumé — À l’ère où la ressource la plus précieuse n’est plus le pétrole, mais les données, les synthèses créées par les Data Scientists, deviennent des solutions et des stratégies réalisables pour les entreprises. Dans la vie réelle de l’informatique décisionnelle, les faits sont très importants, mais l’opinion joue également un rôle crucial car elle peut influencer le processus de prise de décision. Aujourd’hui, de nombreuses sources d’informations non structurées telles que les réseaux sociaux et les blogs, qui croissent avec des volumes et vitesses exponentielles, constituent une mine d’or gratuite pour les entreprises en vue de réaliser leurs enjeux stratégiques et de se démarquer de la concurrence. L’Union IT Services s’est-elle aussi lancée, entre-autres, dans l’exploitation des opinions à travers l’analyse des sentiments, afin de compléter les moyens traditionnels déjà employés dans la collecte et l’analyse de l’opinion pour la fidélisation de la clientèle. C’est dans ce contexte que s’inscrit le présent projet dont l’objet est de mettre en évidence les opinions exprimées par les internautes sur les réseaux sociaux. Ainsi nous avons procédé à une étude bibliographique se rapportant aux fondements de la Data Science, de Big Data et d’analyse des sentiments, ce qui nous a permis de comparer et mettre en oeuvre les trois architectures Big Data les plus utilisées dans le marché mondial et de proposer des solutions comblant les problèmes existants ou probables (Implémentation d’un modèle de régression logistique pour la segmentation géo-démographique, afin de prédire le taux de départ des clients dans une entreprise). Outre cela, ce rapport fait l’objet de la création de nouvelles connaissances à partir de données structurées (Création et déploiement de bots conversationnelles avec Rivescript et DataFlow), et non structurées (Analyse des sentiments dans 1.6 million de tweets à l’aide de R et Python et détection d’humeur dans les expressions faciales moyen-nant l’algorithme de Haar et celui des machines à vecteurs de support ou séparateurs à vaste marge).

Mots clés : Architectures Big Data, Administration Réseaux, Analyse des sentiments, Twitter, User-Generated Content, Apprentissage Supervisé, Traitement automatique du langage naturel, Déploiement, NoSQL, Algorithmes de Machine Learning, Application Programming Interface.