
	<p>REPUBLIQUE TUNISIENNE ***** MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE ***** Département : Ingénierie Informatique</p>	
---	---	---

RAPPORT

De

Projet de spécialité

Analyse des journaux avec Hive

Élaboré par : Habib AROUA

Anas NAJJAR

Imen TRABELSI

Manel TRABELSI

Encadré par : Mr. Aymen HAJ KACEM

Année Universitaire : 2018 / 2019

Remerciement

Premièrement nous remercions Dieu source de toute connaissance...

Nous tenons à présenter nos vifs remerciements et nos profondes gratitudes à notre encadrant monsieur Aymen HAJ KACEM pour sa pédagogie, sa compétence, sa modestie et son aide précieuse tout au long de ce projet.

Sommaire

Introduction Générale	1
Chapitre 1 : Cadre du projet	4
Introduction :	4
I. Présentation du projet :	4
1. L'objectif à atteindre :	4
2. Travail demandé :	4
II. Étude de l'existant :	4
III. La solution proposée :	6
Conclusion :	6
Chapitre 2 : Choix méthodologique et technologique.....	7
Introduction :	7
I. Choix Méthodologique :	7
II. Planification :	8
1. Planification Pratique :	9
2. Modélisation avec UML :	9
3. Définition des UML :	9
4. Les diagrammes de UML :	9
III. ANALYSE COMPARATIVE ENTRE L'APPROCHE TOP-DOWN ET BOTTOM-UP : 10	
1. L'approche de Bill Inmon:	10
2. L'approche de Ralph Kimball :	11
a. Le Datamart :	11
b. Le modèle de Ralph Kimball :	11
3. Comparaison entre les deux architectures :	13
IV. Comparaison entre les différents de modèles :	13
1. Modèle en étoile :	13
2. Modèle en flocon de neige :	14
V. Récapitulatif des choix :	15
1. Choix de la démarche :	16
2. Choix du type du modèle :	16
3. Choix d'outils :	16
Conclusion :	19
Chapitre 3 : Analyse et Spécification des besoins	20
Introduction :	20
I. Étude des besoins :	20

1. Présentation des acteurs :	20
2. Besoins fonctionnels :	20
3. Besoins non fonctionnels :	20
II. Diagramme de cas d'utilisation global :	20
III. Organisation du projet à l'aide de SCRUM :	22
1. Backlog du produit :	22
2. Pilotage du projet avec SCRUM :	22
Notre équipe et formé de 5 comme suit :	22
3. Étude des données sources :	23
Conclusion :	23
Chapitre 4 : Sprint 1 (Mise en place du Data Warehouse)	24
Introduction :	24
I. Conception du Data Warehouse :	24
1. Choix des mesures :	24
2. Choix de dimensions :	24
3. Table de faits :	25
4. Modélisation du Data Warehouse :	25
Conclusion :	26
Chapitre 5 : Sprint 2 (Alimentation du DW)	27
Introduction :	27
I. Backlog du sprint :	27
II. Exécution de l'ETL :	28
1. Phase d'extraction :	28
2. Phase de Transformation :	30
3. Phase de Chargement :	31
Conclusion :	33
Chapitre 6 : Sprint 3 (Mise en place d'un Dashboard)	34
Introduction :	34
I. Sprint Backlog :	34
II. Création des tableaux de bord :	35
Conclusion :	41
Conclusion Générale	42
Webographie	43

Liste des figures

Figure 1: Processus SCRUM.....	8
Figure 2: Inmon MODEL	11
Figure 3: KIMBALL model	12
Figure 4: Exemple d'architecture en étoile	14
Figure 5: Exemple d'architecture en flocon de neige.....	15
Figure 6: Architecture de HDFS	18
Figure 7: Démarche de Bi	19
Figure 8: Cas d'utilisation global.....	21
Figure 9: Scrum Master: Mr Mohamed Aymen	22
Figure 10: Team Members	23
Figure 11: New York times	28
Figure 12: The guardian.....	29
Figure 13: CNN	29
Figure 14: Contenu du fichier journal.txt sur hdfs	30
Figure 15: contenu de la base des données multidimensionnelle.....	31
Figure 16: Contenu de la table globale journal	31
Figure 17: Dimension auteur.....	32
Figure 18: Dimension pays	32
Figure 19: Dimension date	33
Figure 20: page d'accueil de l'application JAVA	35
Figure 21: menu de l'application.....	36
Figure 22: Table des articles par auteur	37
Figure 23: histogramme des articles par auteur	37
Figure 24: diagramme circulaire des articles par auteur.....	38
Figure 25: table des articles par pays	39
Figure 26: Histogramme des articles par pays	39
Figure 27: table des articles par date	40
Figure 28: Histogramme des articles par date	40

Liste des tables

Tableau 1: Comparaison entre les modeles	13
Tableau 2: Backlog du produit.....	22
Tableau 3: mesures	24
Tableau 4: Dimensions	24
Tableau 5: Table de faits	25
Tableau 6: Architecture en étoile.....	26
Tableau 7: Backlog du sprint 2	27
Tableau 8: sprint backlog du sprint 3	35

Introduction Générale

La révolution numérique a complètement bouleversé notre consommation des médias et la manière dont nous accédons à l'information. On lit la presse sur format numérique, on écoute son émission de radio en podcast et on regarde sa série télévisée préférée en différé sur Internet. D'ailleurs, l'information est un besoin impérieux et même un droit, qui peut et doit désormais être satisfait, que ce soit en kiosque, sur un téléphone portable, une tablette tactile ou un écran d'ordinateur. Qu'importe le support ! Mais, avec cette nouvelle ère vient des grands défis, parmi lesquels ou dis-on la plus importante entre elles est le fait de distinguer la bonne nouvelle de la mauvaise.

La fausse nouvelle, que l'actualité remet au goût du jour sous l'appellation de fake news ou d'infox, n'est pas, loin s'en faut, un phénomène inédit !

Il y a toujours eu des fausses nouvelles. De simples erreurs, des canulars ou, plus sérieusement, de la désinformation. Mais le phénomène se présente aujourd'hui sous un nouveau jour. À cause de la prolifération des messages que permettent les réseaux sociaux, l'effet est viral.

Ainsi, le problème avec les fausses nouvelles vient de leur viralité, et leur viralité vient de leur capacité à exploiter la communication multilatérale des médias numériques pour se propager et muter.

Grâce à cette facilité nouvelle de se propager via le web 2.0, les fausses nouvelles bénéficient d'un levier autrefois inexistant : l'appât du gain. Contrairement à l'époque où la propagande et la désinformation circulaient dans les grands médias, les algorithmes des plateformes telles que Facebook et Google attribuent une valeur monétaire à la diffusion, de sorte que le 21^e siècle a vu apparaître des personnes qui rédigent des fausses nouvelles dans le seul but de faire de l'argent.

À cela s'ajoutent les robots qui rediffusent massivement de telles informations, et ce, tant à des fins idéologiques que financières, selon leur origine.

Là où le bât blesse, c'est par la rediffusion par les internautes crédules motivés, eux, par la réaction émotive. Des recherches démontrent que le premier vecteur d'engagement sur le web - ce qui motive les gens à intervenir par des « j'aime » ou par un partage - est l'indignation et c'est précisément ce que cherchent à soulever les titres accrocheurs et les articles de désinformation.

Par contre, même lorsque l'argent est le premier motivateur, le résultat demeure le même, comme l'indiquent les chercheurs Samantha Bradshaw et Philip Howard, de l'Oxford Internet Institute au Royaume-Uni : « Les algorithmes et l'analyse des mégadonnées rendent la subversion des processus politiques plus efficace que jamais. La démocratie est menacée », affirment-ils, faisant la démonstration par leurs travaux que les élections ont été perturbées dans les grandes démocraties par les technologies tout au long de 2016 et 2017.

Ce qu'eux appellent les « nouvelles bidon » (junk news) échangées sur les réseaux sociaux ont d'ailleurs surpassé les nouvelles professionnelles dans 12 des 16 états pivots où la lutte était serrée entre Hillary Clinton et Donald Trump et les robots politiques, ces « cybersoldats », ont été largement utilisés pour amplifier certaines nouvelles. Le phénomène ne se limite pas aux États-Unis : les deux chercheurs ont identifié une telle « manipulation organisée des médias sociaux » dans 28 pays en 2017.

L'information professionnelle débordée :

L'internet vient ainsi placer les fausses nouvelles en concurrence directe avec l'information factuelle et vérifiée selon les normes journalistiques professionnelles et les réseaux sociaux, avec leurs algorithmes qui dirigent vers le lecteur des contenus similaires, aggravent le problème en créant ce qu'il est désormais convenu d'appeler des « bulles informationnelles ».

Comme le dit le sociologue Serge Proulx, « quand Facebook devient la première source d'information politique, en particulier chez les jeunes, il existe un réel danger pour que les informations qui atteignent les individus enfermés dans leur bulle ne possèdent pas le degré suffisant de précision et de diversité pour que s'enclenchent des débats fructueux entre porteurs d'arguments contradictoires ».

Crise de confiance :

Toute cette mouvance s'inscrit dans un contexte de crise des médias, souligne Normand Landry, titulaire de la Chaire de recherche du Canada en éducation aux médias de la TÉLUQ (Université du Québec). Cette crise comporte deux aspects, soit l'effondrement des revenus et du modèle d'affaires (qui s'est traduit par la fermeture de nombreux médias et des réductions de personnel importantes dans ceux qui survivent encore) et une précarisation croissante du métier de journaliste.

En parallèle, le chercheur décrit une crise de confiance d'une part croissante de la population devenue « méfiante, critique et sceptique » en raison de la croyance que les médias souffrent de biais et sont complices « d'intérêts politiques et économiques »,

une croyance souvent alimentée par les médias eux-mêmes lorsqu'ils errent sur les sentiers risqués du sensationnalisme ou maintiennent de faibles standards journalistiques.

La combinaison de la prolifération des fausses nouvelles et de cette méfiance face à l'information vérifiée mène Normand Landry à reprendre la formulation de l'IAMCR (International Association for Media and Communication Research) selon qui les fausses nouvelles « postulent que toutes les positions sont égales en raison de l'abondance de l'information en ligne (faits alternatifs, post-factuels) et menacent l'intégrité du savoir et du raisonnement scientifique ».

Dans ce contexte, plusieurs questions se posent : Comment, dans cet univers en réseau, les citoyens peuvent-ils s'assurer de la véracité des informations qui leur sont transmises ? Les mécanismes qui permettraient leur validation dans le monde des médias traditionnels semblent appartenir à une autre époque. Les journalistes doivent-ils repenser leur rôle ? Peut-on faire confiance aux Social Media qui dit multiplier les efforts pour débusquer les faussetés sur sa plateforme ? Que faut-il attendre des chercheurs ?

Dans le cadre de notre projet de spécialité de la troisième année cycle ingénieur en Génie Logiciel à Sésame, nous sommes lancés à développer une application de détection des fausses nouvelles ou fake news en se basant sur les technologies du BI et pour être précis de la manipulation du Data Warehouse (entrepôt de données).

Chapitre 1 : Cadre du projet

Introduction :

Avant d'entrer dans les détails de notre projet, nous devons l'encadrer dans son contexte. Dans ce chapitre, je commence par une présentation brève de l'entreprise et du projet et puis j'entamerai l'étude de l'existant et la découverte de ses limites ce qui me permettra de dégager la solution à mettre en œuvre.

I. Présentation du projet :

Pour cette partie, l'objectif est de développer une application de détection de Fake News en se basant sur les technologies du BI.

Notre travail est réalisé dans le cadre d'un projet de spécialité. L'application consiste en un logiciel qui donne la possibilité de vérifier si une nouvelle présente sur le net est valide ou fausse.

1. L'objectif à atteindre :

Le but de notre projet est de développer un logiciel pour :

- Alimenter l'entrepôt de données de toutes les nouvelles valides sur le net et provenant des sources de confiance vérifiées.
- Analyser les news obtenus avec des requêtes spécifiques.
- Donner une présentation graphique des résultats de ses analyses.
- Permettre à l'utilisateur du logiciel de vérifier si une nouvelle dont il doute de sa véracité est valide ou fausse.

2. Travail demandé :

Nos tâches consistent à :

- Créer un entrepôt de données.
- Alimenter l'entrepôt des données avec des news provenant de sources de confiance.
- Analyser les news avec des requêtes.
- Présenter quelques analyses graphiquement.

II. Étude de l'existant :

Les médias et journalistes, malgré leur situation financière précaire dans l'environnement actuel, représentent l'un de quatre axes de lutte contre les Fake News. Les chercheurs Florian Sauvageau et Simon Thibault constatent qu'en réaction aux fausses nouvelles, les médias ont multiplié les rubriques de vérification des faits. Certaines fausses nouvelles, « aussi loufoques qu'incroyables » deviennent virales et c'est cette popularité qui « permet de penser que les médias n'ont sans doute pas tort de

croire que leur réfutation est utile et fait partie de la nécessaire lutte aux fausses nouvelles ».

Plus encore, ils notent sans complaisance que le public est préoccupé par « la qualité médiocre du journalisme (erreurs factuelles, simplification à outrance, titres trompeurs) », d'où la nécessité pour eux, comme le souligne le sociologue Serge Proulx, de revoir les pratiques et standards journalistiques, d'instaurer des mesures serrées de validation et de procéder à une révision des politiques éditoriales pour mieux distinguer les opinions et les faits. En se basant sur ces mesures de validation et vérifications, des journaux se sont distingués des autres puisque le pourcentage de propagation des Fake News chez ceux-ci sont minimales. Pourtant, les internautes poursuivent et fassent encore confiance à des sources considérablement moins fiables et ainsi se fassent avoir à chaque fois suite à leurs croyances en des nouveautés qui manquent la moindre de crédibilité.

Plusieurs développeurs ont essayé de développer une solution optimale de détection de Fake News mais peu de logiciels ont eu un succès notable ou ont vu le jour pour plusieurs raisons (financières etc.).

Parmi les logiciels qui ont eu du succès est un logiciel pour automatiser la détection des « Fake News » surnommé « détecteur de baratin », il est financé par les milliardaires George Soros et Pierre Omidyar et a été testé et mis en disposition des utilisateurs en octobre 2017. « *Détecteur de baratin* ». Tel est l'irrévérencieux surnom du logiciel développé par une équipe de chercheurs de l'ONG londonienne spécialiste du « fact-checking », Full Fact. L'objectif ? Mettre à la disposition des journalistes un programme de détection automatisée et en temps réel de fausses informations. Le logiciel « *scanne les sous-titres d'émissions d'actualité diffusées en direct, des retransmissions de sessions parlementaires, mais aussi les articles de journaux* », rapporte le journal britannique « The Guardian », jusqu'à identifier les « *affirmations correspondant aux faits vérifiés qui sont dans sa base de données* ». Il s'avère en outre capable de faire apparaître à l'écran « *les informations confirmées ou infirmées au fil du discours d'un politicien* ». Ce logiciel est performant mais il comporte des inconvénients tels que le coût très élevé et la difficulté de vérifier et valider toutes les nouveautés vu leur quantité importante qui s'améliore à chaque minute donc il y a une grande possibilité que ces nouveautés viennent aux mains des utilisateurs finaux sans vérification depuis d'autres sources. Cette solution pourra être optimale seulement au sein des journaux qui ne vont partager les nouveautés qu'après la vérification et la validation et ça explique son usage qui est limité aux journaux seulement.

III. La solution proposée :

Notre solution consiste à créer un entrepôt de données et l'alimenter depuis les journaux les plus fiables et qui mènent des vérifications serrées sur leurs nouveautés ce qui minimise le pourcentage de passer une nouveauté fautive dans leurs plateformes. Suite à l'alimentation du DW (Data Warehouse), on va faire des analyses multiples sur les nouveautés obtenues et enfin on va montrer les résultats de ces analyses sur des chartes graphiques. Cette application sera la base d'une application web qui sert à donner aux utilisateurs la possibilité de tester si une nouveauté est vraie ou fautive. Les avantages de notre solution sont : Le fait qu'elle soit peu coûteuse, permet de fournir des réponses instantanées et qu'elle soit à la portée de tous les utilisateurs de l'internet et aussi qu'elle couvre tous les types de nouveautés et dans le monde entier.

Conclusion :

Dans ce chapitre nous avons présenté le cadre général de notre projet en donnant une présentation générale du projet, en déterminant la problématique et en proposant une solution envisagée pour faire face à la situation courante. Dans le chapitre suivant on va présenter la méthodologie et les technologies utilisés.

Chapitre 2 : Choix méthodologique et technologique

Introduction :

Nous avons permis de prendre connaissance et d'appliquer un ensemble de nouvelles méthodologies de travail auxquelles nous n'avions jamais été confrontés.

Avant d'aborder l'analyse et la conception de notre application, nous allons présenter notre choix des méthodologies de travail et des technologies nécessaires pour mener à terme de réalisation de notre application.

I. Choix Méthodologique :

Un projet informatique, quelle que soit sa taille et la portée de ses objectifs, nécessite la mise en place d'un planning organisationnel tout au long de son cycle de vie. C'est ainsi qu'est apparue la notion de méthode.

Une méthode, dans le contexte informatique, peut être définie comme une démarche fournissant une méthodologie et des notations standards qui aident à concevoir des logiciels de qualité.

Malgré la diversité des méthodes d'analyse et de conception, il est possible de les classer en quatre catégories :

- ✓ Les méthodes cartésiennes ou fonctionnelles : SADT
- ✓ Les méthodes systémiques : MERISE, AXIAL.
- ✓ Les méthodes objet : OMT,
- ✓ Les méthodes/méthodologies agiles : SCRUM, etc.

Pour atteindre les objectifs, les méthodes agiles partagent pour la plupart un ensemble de technique. L'une des particularités des méthodes agiles est de considérer le groupe projet comme une équipe plus qu'une somme de personnes. Les notions de rôles et de hiérarchie sont réduites à leur strict minimum et, c'est l'esprit de groupe qui est favorisé. Ce groupe doit partager un but commun : celui de réussir le projet.

Scrum :

La méthodologie Scrum est une méthodologie agile, créée en 2002, dont le nom est un terme emprunté au rugby qui signifie « la mêlée ». Elle s'appuie sur le découpage des projets en itérations encore nommées « Sprints ». Un Sprint peut avoir une durée qui varie généralement entre deux semaines et un mois. Scrum signifie mêlée au rugby.

Scrum utilise les valeurs et l'esprit du rugby et les adapte aux projets de développement. Comme le pack lors d'un ballon porté au rugby, l'équipe chargée du développement travaille de façon collective, soudée vers un objectif précis. Comme un demi de mêlée, le Scrum Master aiguillonne les membres de l'équipe, les repositionne dans la bonne direction et donne le tempo pour assurer la réussite du projet.

Scrum est issu des travaux de deux des signataires du Manifeste Agile, Ken Schwaber et Jeff Sutherland, au début des années 1990. Il appartient à la famille des méthodologies itératives et incrémentales et repose sur les principes et les valeurs agiles.

Le plus souvent, les experts de Scrum, même ses fondateurs, le décrivent comme un cadre ou un patron de processus orienté gestion de projet et qui peut incorporer différentes méthodes ou un langage et méthodologie de conception.

La méthodologie est une démarche organisée rationnellement pour aboutir à un résultat. Parmi les différentes méthodologies existantes, nous pouvons citer le modèle en cascade utilisée souvent dans les simples projets dont les besoins sont clairs et bien définis dès le début, le modèle en Y utiliser pour le développement des applications mobiles, ainsi que le processus unifié et les méthodologies agiles (Scrum & extreme programming) caractérisées par leurs souplesses et utilisées dans des grands projets. Pour bien conduire notre projet et nous assurer du bon déroulement des différentes phases, nous avons opté Scrum comme une méthodologie de conception et de développement.

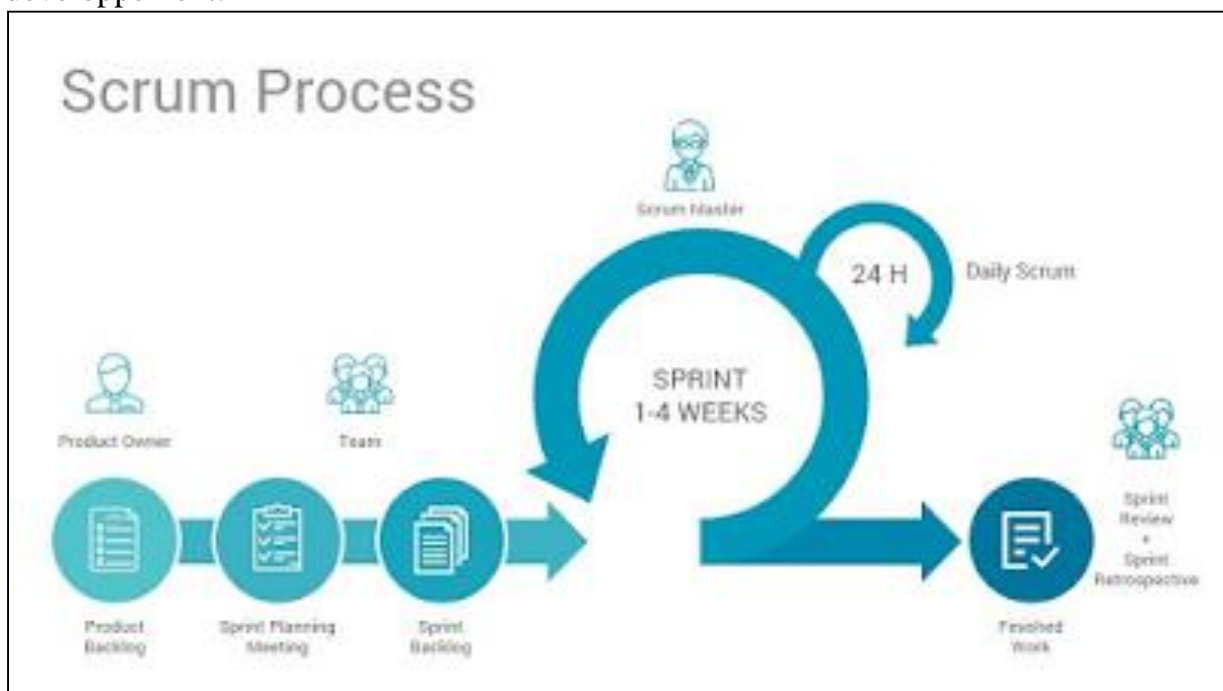


FIGURE 1: PROCESSUS SCRUM

Le choix de Scrum comme une méthodologie de pilotage pour notre projet s'est basé sur les atouts de ce dernier. Il se résumé comme suit :

- Plus de souplesse et de réactivité.
- La grande capacité d'adaptation au changement grâce à des itérations courtes.
- Et la chose plus importante, c'est que Scrum rassemble les deux cotés théorique et pratique et se rapproche beaucoup de la réalité.

II. Planification :

SCRUM est un processus de gestion du projet vise à organiser les exigences d'affaires, établir le coût et le calendrier précis du projet (y compris une liste des livrables et leurs dates de livraison), planifier l'organisation du travail et obtenir l'autorisation des

gestionnaires. Dans SCRUM la planification se fait par niveau, chaque niveau correspondant à un Sprint. Une réunion des collaborateurs s'effectue généralement sur 8h et en deux temps.

Dans SCRUM l'équipe choisit, à partir du Backlog de produit, les éléments qu'elle s'engage à finir. Une fois le Backlog de Sprint est créé, les tâches sont identifiées et estimées (1-16 heures). Finalement on peut alors lancer le Sprint.

1. Planification Pratique :

Nous avons choisi de diviser notre projet sur 4 sprints 2-4 semaines pour chacun :
Sprint 0 : L'analyse des besoins (spécifications fonctionnelles et non fonctionnelles etc.).

- Sprint 1 : Modélisation du Data Warehouse.
- Sprint 2 : Alimentation du Data Warehouse.
- Sprint 3 : Analyse des données.
- Sprint 4 : Mise en place des tableaux de bord.

2. Modélisation avec UML :

Vu l'importance cruciale de la modélisation dans le cycle de vie de n'importe quelle application, il fallait utiliser une méthode de modélisation qui s'adapte le mieux à nos besoins et à nos exigences qui sont entre autres : L'ouverture, la réutilisabilité, la modularité et l'extensibilité. Pour répondre à ces exigences, nous avons choisis de modéliser avec le langage de modélisation UML qui s'adapte parfaitement à la modélisation des applications à base d'objets et qui offre grâce à ses différents diagrammes une grande souplesse permettant la modélisation de différents aspects de l'application. Notre choix de ce langage se justifie aussi par le fait qu'UML est devenu un standard de modélisation adopté pour toutes les applications à aspect orienté objet.

3. Définition des UML :

UML (Unified Modeling Language) : Se définit comme un langage de modélisation graphique et textuel destiné à comprendre et décrire des besoins, spécifier, concevoir des solutions et communiquer des points de vue. UML unifie à la fois les notations et les concepts orientés objet. Il ne s'agit pas d'une simple notation, mais les concepts transmis par un diagramme ont une sémantique précise et sont porteurs de sens au même titre que les mots d'un langage, c'est pour ça qu'UML est présenté parfois comme une méthode alors qu'il ne l'est absolument pas. UML unifie également les notations nécessaires aux différentes activités d'un processus de développement et offre, par ce biais, le moyen d'établir le suivi des décisions prises, depuis la définition des besoins jusqu'au codage.

4. Les diagrammes de UML :

Ci-dessous une présentation rapide du diagramme UML qui va être utilisé pour ce projet :

Le diagramme des cas d'utilisation : Il représente la structure des fonctionnalités nécessaires aux utilisateurs du système. Il est normalement utilisé lors des étapes de capture des besoins fonctionnels et techniques.

III. ANALYSE COMPARATIVE ENTRE L'APPROCHE TOP-DOWN ET BOTTOM-UP :

Quand on parle de la conception du data Warehouse pour le projet, les deux méthodes les plus connues sont l'approche introduite par Bill Inmon (Top-down) et celle de Ralph Kimball (Bottom-up). Chacune d'entre elles a ses propres caractéristiques. Dans les paragraphes qui vont suivre, nous nous intéresserons de plus près à chaque approche pour faire une étude comparative entre les deux approches.

1. L'approche de Bill Inmon:

D'après Inmon l'entrepôt de données fournit un cadre logique pour une prestation de business intelligence et gestion d'affaire.

Selon la théorie et l'approche de Bill Inmon l'entrepôt de données est :

- **Orienté vers le sujet** : les données de l'entrepôt de données sont organisées selon les sujets (table de fait) et les dimensions sont les axes d'analyse de ces derniers.
- **Non volatile** : les données de l'entrepôt de données ne sont jamais écrasées ou supprimées une fois engagées, les données sont statiques. Elles sont conservées pour les analyses futures.
- **Intégré** : l'entrepôt de données contient des données en provenance de la plupart ou la totalité des systèmes opérationnels et ces données sont rendues compatibles.
- **Varient sur le temps (Time-Variant)** : Pour un système d'exploitation, les données mémorisées contiennent la valeur actuelle.

Le data warehouse :

Il s'agit d'un entrepôt de données centralisé, qui a pour tâche d'enregistrement d'organiser des informations dans un entrepôt de données à partir d'un organisme ou d'une société.

Dans ce cas le DW est une architecture de base relationnelle formée d'une approche dite <<Top-Down>> qui provient d'un modèle de données d'entreprise défini d'avance. Le data warehouse représente une vue centralisée et organisée sur toutes les informations, ce qui met en avant la massification d'un data warehouse.

Cette figure illustre le processus du data warehouse selon l'approche de Bill Inmon:

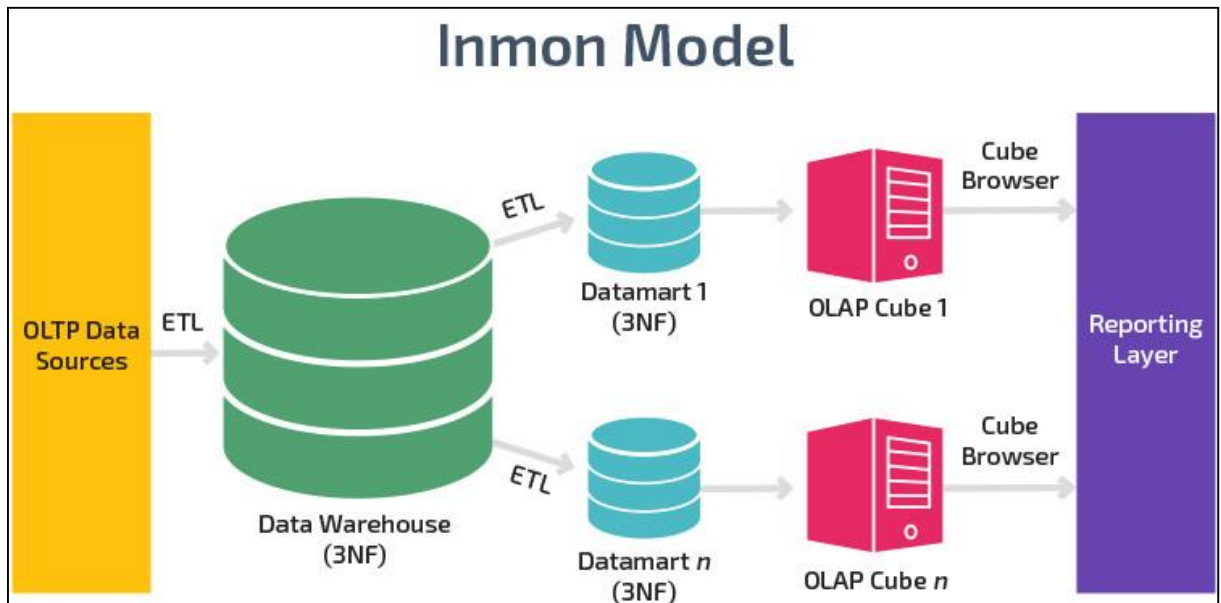


FIGURE 2: INMON MODEL

2. L'approche de Ralph Kimball :

Une approche du data warehouse peut être prise en compte, c'est celle de Ralph Kimball selon qui l'entrepôt de données est l'union des datamarts cohérents entre eux grâce aux dimensions conformes.

a. Le Datamart :

Un datamart est une vue spécifique, orientée sujet d'une organisation. C'est un sous-ensemble d'un Data Warehouse destiné à fournir des données aux utilisateurs et souvent spécialisé vers un groupe ou un type d'affaire. Il est composé de tables détaillées ou agrégées reliées entre elles, il est possible de trouver plusieurs Datamart dans une seule organisation.

b. Le modèle de Ralph Kimball :

Le modèle de Ralph est dimensionnel, il revient donc de définir les tables de faits et les dimensions. Les tables de faits comprennent des faits numériques et des

mesures dont les dimensions forment les axes d'analyseur lesquels nous nous basons pour étudier les données multidimensionnelles.

Pour concevoir un bon modèle dimensionnel il faut passer par quatre étapes :

- Choisir le processus à modéliser
- Définir la granularité du processus
- Choisir les dimensions
- Identifier le fait

L'objectif de Ralph Kimball est de rendre les données accessibles à l'utilisateur grâce à la présentation harmonieuse des informations.

Voici l'architecture de Kimball :

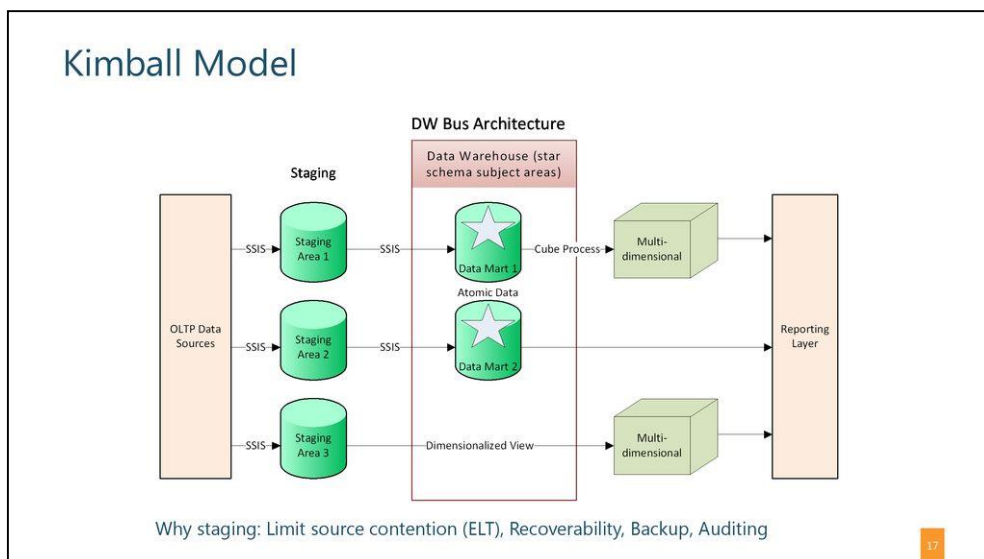


FIGURE 3: KIMBALL MODEL

3. Comparaison entre les deux architectures :

Caractéristique	L'approche Bottom-Up	L'approche Top-Down
Partisan	Bill Inmon	Ralph Kimball
Conception	Datawarehouse	Data Mart
La structure de l'architecture	Un datawarehouse regroupe toutes les données de l'entreprise	Un datamart est centré sur un sujet ou un métier particulier
La complexité de la méthode	Assez simple	Simple
Outils	Traditionnel (modèle entité relation)	Modélisation dimensionnelle qui part de la modélisation relationnelle
Autre	-Flexibilité -orienté donnés -Longue vie	-Restrictive -Orienté projet -Courte vie

TABLEAU 1: COMPARAISON ENTRE LES MODELES

IV. Comparaison entre les différents de modèles :

Avant de choisir le type de modèle que nous allons concevoir, il faut faire une étude comparative entre les types existants afin de connaître le modèle le plus approprié pour le projet.

1. Modèle en étoile :

Il s'agit d'un modèle multidimensionnel qui est formé d'une table de faits liés par des dimensions comme le montre la figure ci-dessous. Pour choisir le modèle le plus adéquat, nous allons présenter les avantages et les inconvénients de chacun.

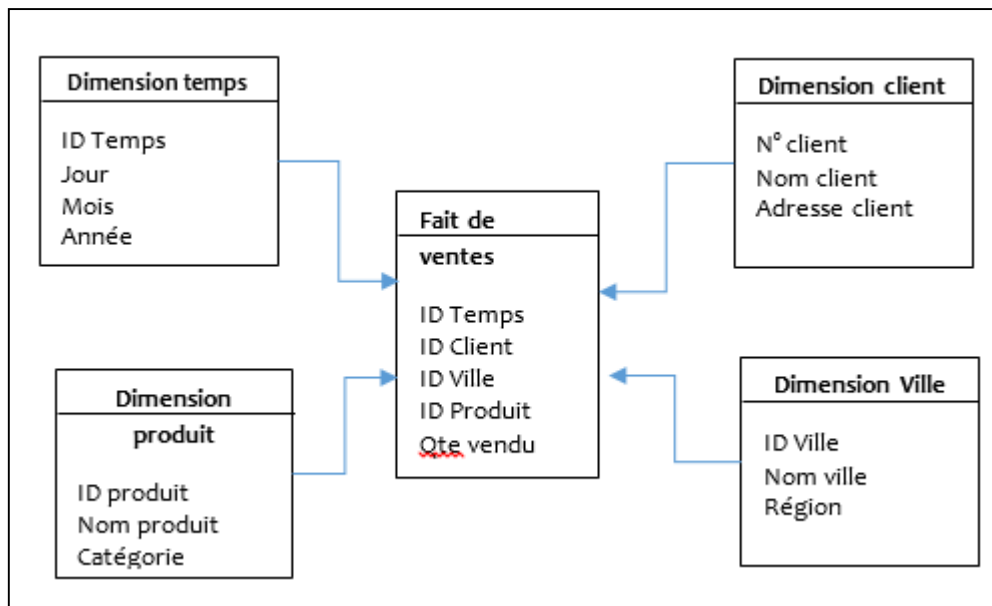


FIGURE 4: EXEMPLE D'ARCHITECTURE EN ETOILE

Avantages :

- Les requêtes sont plus simples car les jointures du schéma en étoile sont plus simples que celle du schéma normalisé
- La logique de reporting est simplifiée dans un schéma en étoile, telle que la génération de rapports sur une période ou un intervalle de temps
- La performance des applications est améliorée et amène à une lecture simple du rapport

Inconvénients :

- Il y a une redondance dans les dimensions
- La mise à jour peut entraîner des anomalies sur les données

2. Modèle en flocon de neige :

La seule différence par rapport au modèle en étoile c'est le changement des dimensions. Dans le schéma en étoile, elles sont normalisées.

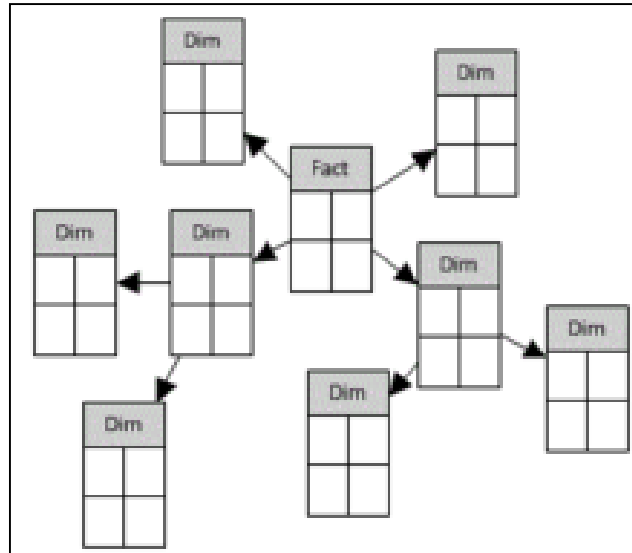


FIGURE 5: EXEMPLE D'ARCHITECTURE EN FLOCON DE NEIGE

Avantages :

- Il y a une économie de stockage grâce à la normalisation des attributs

Inconvénients :

- Difficulté de navigation à cause des nombreuses jointures
- Puisque l'objectif est de stocker plus, ceci implique une augmentation du coût au niveau temporel. La performance lors de la navigation sera également mauvaise

V. Récapitulatif des choix :

Toutes ces comparaisons mènent à un choix de méthodologie appropriée à notre projet. Dans les paragraphes qui suivent, nous allons expliquer nos choix.

1. Choix de la démarche :

En se basant sur la comparaison entre l'approche Top-Down de Bill Inmon et l'approche Bottom-Up de Ralph Kimball, nous avons choisi celle de Bill Inmon étant donné que nous ne disposons pas de dimensions qui sont en relation l'une avec l'autre.

2. Choix du type du modèle :

Les deux modèles présentés sont intéressants, mais nous allons opter pour le modèle en étoile puisque nous n'avons pas besoin de normaliser les dimensions existantes.

3. Choix d'outils :

Ci-suit nous allons parler des outils qui vont nous faciliter tout le processus :

HDFS :

HDFS (Hadoop Distributed File System) est le composant de Hadoop en charge du stockage des données dans un cluster Hadoop.

HDFS se démarque d'un système de fichiers classique pour les principales raisons suivantes :

- HDFS est optimisé pour maximiser les débits de données. La taille d'un bloc de données est ainsi de 64 Mo dans HDFS contre 512 octets à 4 Ko dans la plupart des systèmes de fichiers traditionnels, ce qui permet de réduire le seek time. (Il est toutefois possible d'augmenter la taille d'un bloc à 128 Mo ou à 256 Mo en fonction des besoins)
- HDFS est un système de gestion de fichiers du type Write Once Read Many (WORM) : on y écrit une fois le fichier, puis on y accède plusieurs fois.
- HDFS fournit un système de réplication des blocs dont le nombre de réplications est configurable (3 par défaut). Pendant la phase d'écriture, chaque bloc correspondant au fichier est répliqué sur des nœuds distincts dans le cluster, ce qui contribue à en garantir la fiabilité et la disponibilité au moment de lecture de données. Si un bloc est indisponible sur un nœud, des copies de ce bloc seront disponibles sur d'autres nœuds.
- HDFS s'appuie sur le système de fichier natif de l'OS pour présenter un système de stockage unifié reposant sur un ensemble de disques et de systèmes de fichiers hétérogènes.

Le fonctionnement de HDFS est assuré par deux types principaux de daemons :

- Le NameNode (nœud maitre – master node):

Dans un cluster Hadoop, le NameNode gère l'espace de noms, l'arborescence du système de fichiers et les métadonnées des fichiers et des répertoires. Il centralise la localisation des blocs de données répartis dans le cluster.

Quand un client sollicite Hadoop pour récupérer un fichier, c'est via le Namenode que l'information est extraite. Ce Namenode va indiquer au client quels sont les Datanodes qui contiennent les blocs.

Le NameNode reçoit régulièrement un « battement de cœur » et un Blockreport de tous les DataNodes dans le cluster afin de s'assurer que les datanodes fonctionnent correctement. Un Blockreport (ou rapport de bloc) contient une liste de tous les blocs d'un DataNode.

En cas de défaillance du datanode, le NameNode choisit de nouveaux datanodes pour de nouvelles répliquations de blocs de données, équilibre la charge d'utilisation des disques et gère également le trafic de communication des datanodes.

- Le DataNode. (Nœud Slave - Slave Node).

Voici un schéma de l'architecture de HDFS :

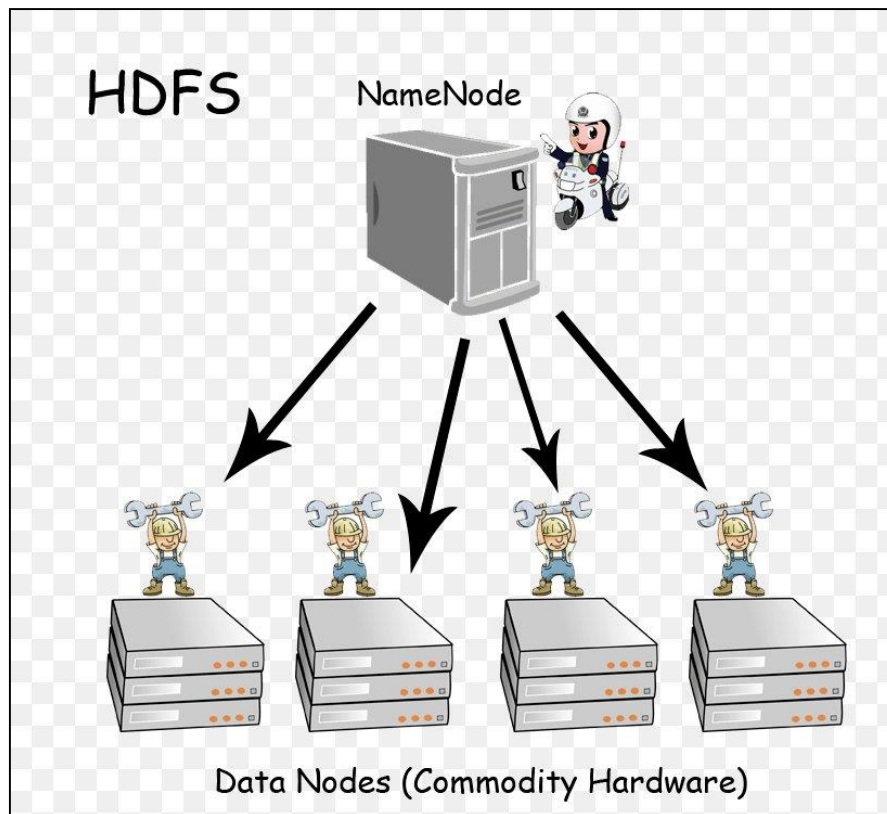


FIGURE 6: ARCHITECTURE DE HDFS

Hive :

Hive est un projet de logiciel d'entrepôt de données construit sur Hadoop pour fournir des requêtes et des analyses de données. Hive propose une interface de type SQL pour interroger les données stockées dans diverses bases de données et systèmes de fichiers qui s'intègrent à Hadoop.

Java SWING :

Swing est une boîte à outils de widget graphique pour Java. Il fait partie des classes Java Foundation d'Oracle - une API pour fournir une interface utilisateur graphique pour les programmes Java. Swing a été développé pour fournir un ensemble de composants d'interface graphique plus sophistiqué que le précédent Window Toolkit abstrait.

Le schéma ci-dessous représente le processus de la réalisation du projet BI avec les outils choisis :

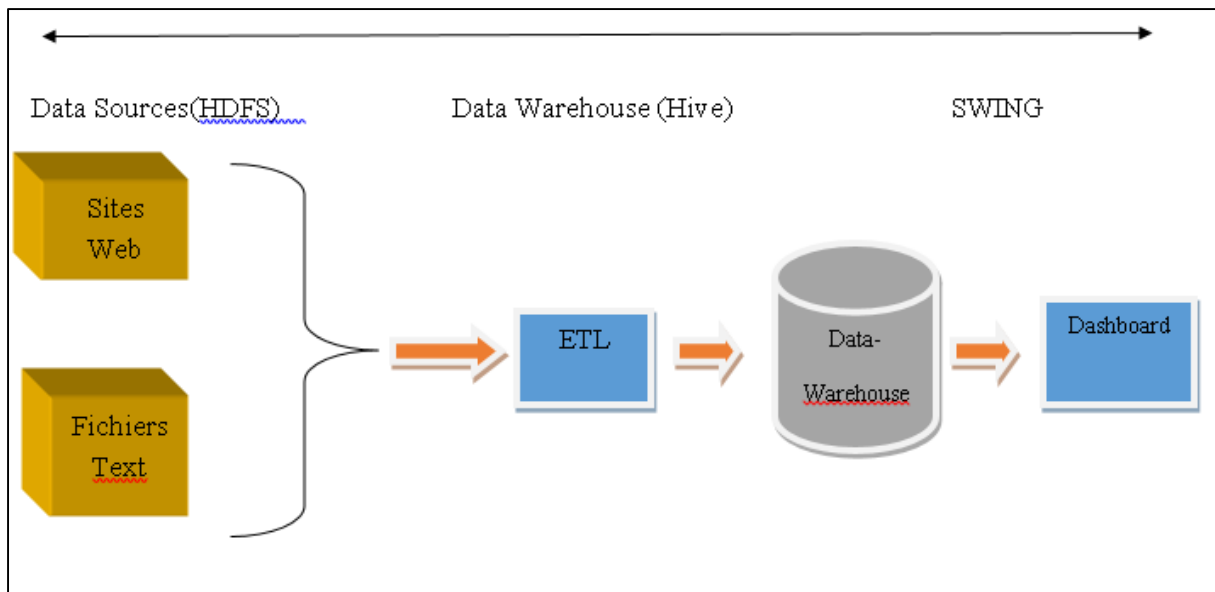


FIGURE 7: DÉMARCHE DE BI

Conclusion :

Dans ce chapitre nous avons présenté le planning du projet en liaison avec nos tâches, la méthodologie et les technologies utilisés tout en déterminant la cause de nos choix. Dans le chapitre suivant on va présenter l'analyse et la spécification des besoins.

Chapitre 3 : Analyse et Spécification des besoins

Introduction :

La spécification des besoins constitue une étape essentielle dans le cycle de développement de chaque système informatique, puisqu'elle permet de déterminer les fonctionnalités de l'application.

I. Étude des besoins :

L'objectif de la phase de spécification des besoins consiste à définir en détails l'ensemble des fonctionnalités offertes par le système ainsi que les acteurs. Les besoins dégagés ont été répartis en deux groupes fonctionnels et non fonctionnels.

1. Présentation des acteurs :

Ils sont définis comme étant les utilisateurs directs de l'application. Dans le cadre de notre projet, et à son état d'avancement actuel, nous ne pouvons que désigner un seul acteur :

L'analyste : Il peut mener des analyses spécifiques sur l'ensemble des données du Data Warehouse et les afficher graphiquement sur le Dashboard.

Le spécialiste de l'ETL : Il a comme rôle de collecter les données et les stocker dans le Data Warehouse après traitement et transformation.

Le décideur : Il impose certains critères avant de commencer une ou plusieurs analyses.

2. Besoins fonctionnels :

- Appliquer des analyses sur les données du Data Warehouse.

3. Besoins non fonctionnels :

- Clarté : des tableaux clairs lisibles et faciles à interpréter
- L'efficacité : l'analyse du tableau de bord doit donner un résultat à un pourcentage près des attentes des décideurs
- Scalabilité : la solution doit être extensible pour pouvoir suivre le rythme de l'ajout des nouveautés en termes de données et historique de donnée.

II. Diagramme de cas d'utilisation global :

L'étude approfondie de la spécification des besoins permet de dégager plusieurs cas d'utilisation. Un cas d'utilisation décrit une utilisation du système par un acteur particulier. Ce qui revient à présenter les besoins fonctionnels de façon formelle. Les cas d'utilisation permettent de structurer les besoins des utilisateurs et les objectifs correspondants d'un système. Ils centrent l'expression des exigences du système sur ses utilisateurs.

Le diagramme de cas d'utilisation est un schéma qui montre les cas d'utilisations reliés par des associations à leurs acteurs.

Dans notre application nous avons réparti les cas d'utilisation selon les acteurs cités précédemment.

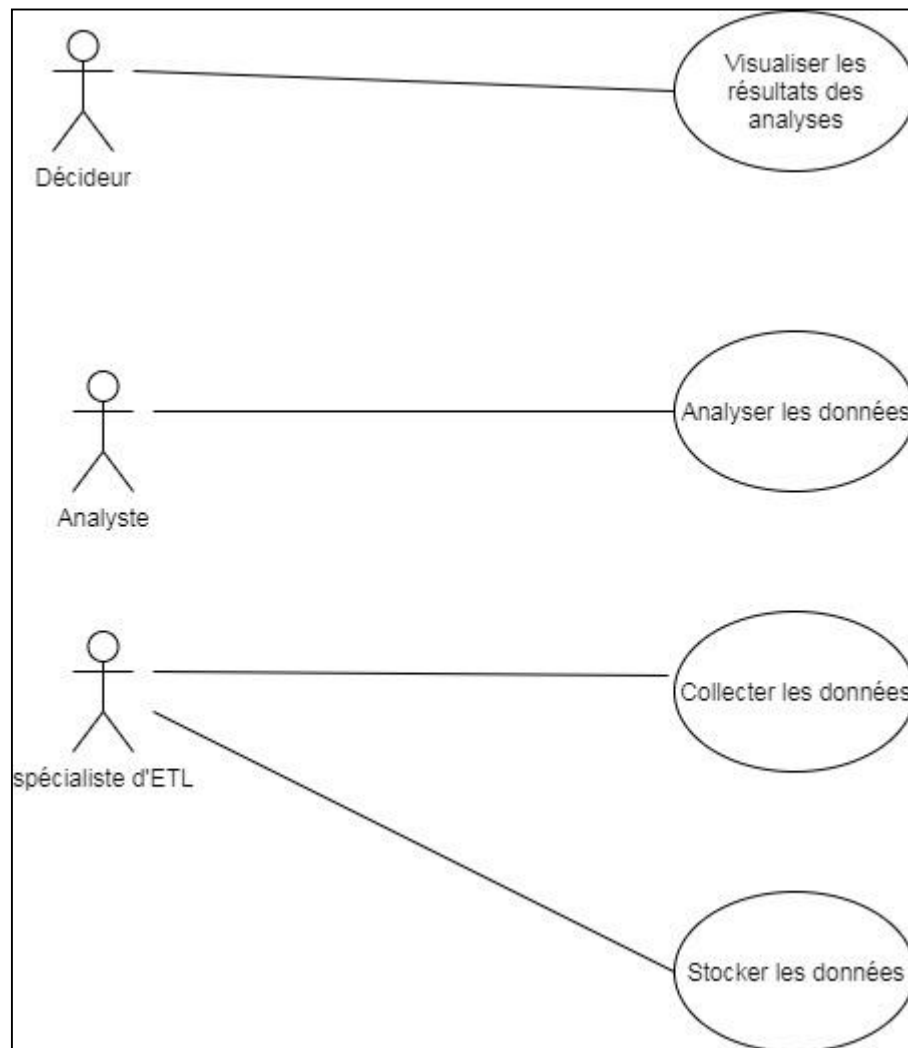


FIGURE 8: CAS D'UTILISATION GLOBAL

III. Organisation du projet à l'aide de SCRUM :

1. Backlog du produit :

ID	User Stories	Priorité
1	En tant que décideur je veux avoir une vue globale sur les journaux.	élevé
2	En tant que décideur je veux avoir le pouvoir de filtrer les journaux par pays.	élevé
3	En tant que décideur je veux avoir le pouvoir de filtrer les journaux par Auteur.	élevé
4	En tant que décideur je veux avoir le pouvoir de filtrer les journaux par Date.	élevé
5	En tant qu'analyste j'analyse les données.	élevé
6	En tant que spécialiste d'ETL je stocke les données .	moyenne
7	En tant que spécialiste d'ETL je collecte les données .	moyenne

TABLEAU 2: BACKLOG DU PRODUIT

2. Pilotage du projet avec SCRUM :

Notre équipe est formée de 5 comme suit :

- Le Scrum Master représenté par Mr Mohamed Aymen BEL HAJ KACEM
- Les team members représentés par Anas NAJJAR , Habib AROUA , Manel TRABELSI et Imen TRABELSI.

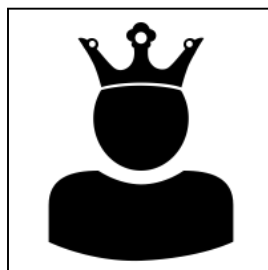


FIGURE 9: SCRUM MASTER: MR MOHAMED AYMEN

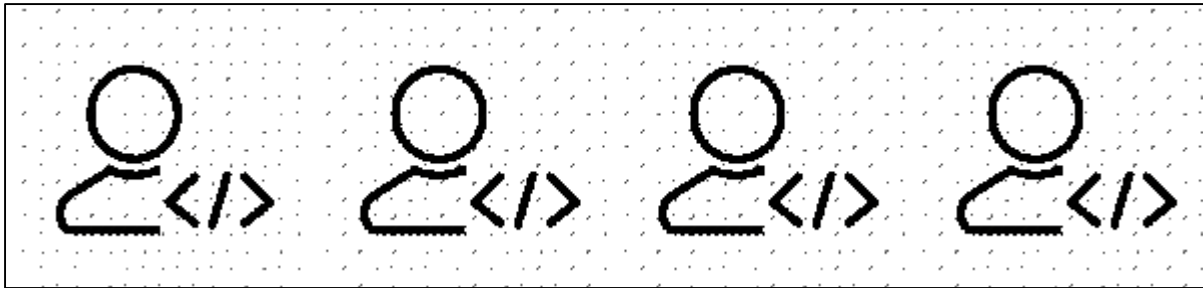


FIGURE 10: TEAM MEMBERS

3. Étude des données sources :

Après la détermination des besoins fonctionnels et non fonctionnels du projet, il faut passer à la collecte des données. Dans cette partie nous allons étudier les données sources collectées du système d'information, avec lesquels nous allons alimenter notre entrepôt de données.

Dans ce cas nous avons eu recours à des données sous format texte dans le fichier journal.txt depuis le HDFS, ses données proviennent des sites des journaux numériques suivants :

- <https://www.bbc.com/news>
- <https://www.france24.com/en/>
- <https://www.aljazeera.com/>
- <https://edition.cnn.com/>
- <https://www.nytimes.com>
- <https://www.independent.co.uk>
- <https://www.theguardian.com>

Conclusion :

Dans ce chapitre nous avons présenté la spécification des besoins tout en déterminant la cas d'utilisation global et Le backlog du produit. Dans le chapitre suivant on va présenter le Sprint 1.

Chapitre 4 : Sprint 1 (Mise en place du Data Warehouse)

Introduction :

Le sprint 1 contient les objectifs de la construction du Data Warehouse qui va permettre par la suite, l'élaboration du tableau de bord et des rapports.

I. Conception du Data Warehouse :

1. Choix des mesures :

Afin de préparer la table de fait, il faut déterminer les mesures des éléments recueillis et collectés pendant la phase de préparation des données sources. Il faut à ce moment préciser les mesures que nous allons utiliser, représentées dans le tableau suivant :

Mesures	Nom
	Titre
	Contenu

TABLEAU 3: MESURES

2. Choix de dimensions :

Les dimensions représentent les thèmes qui vont servir à analyser les données. En effet, la table de dimension contient des attributs et une clef primaire indépendante des autres attributs.

Notre projet contient les dimensions suivantes :

Dimension	Nom relative à la dimension
Date_Dim	Dimension Date
Country_Dim	Dimension Pays
Author_Dim	Dimension Auteur

TABLEAU 4: DIMENSIONS

3. Table de faits :

La table de faits contient les données observables(faits) que nous allons étudier, selon les dimensions.

Notre table de faits se présente comme le montre la table suivante :

Les clés étrangères	Fk_Date	ID Date de l'article
	Fk_Country	ID pays de l'article
	Fk_Author	ID Auteur de l'article
Mesures	Nom	Nom de l'article
	Titre	Titre de l'article
		Contenu de l'article

TABLEAU 5: TABLE DE FAITS

4. Modélisation du Data Warehouse :

Après la définition des mesures, des dimensions et de la table de faits, nous allons finalement procéder à la modélisation du schéma conceptuel de notre entrepôt de données qui se présente comme suit :

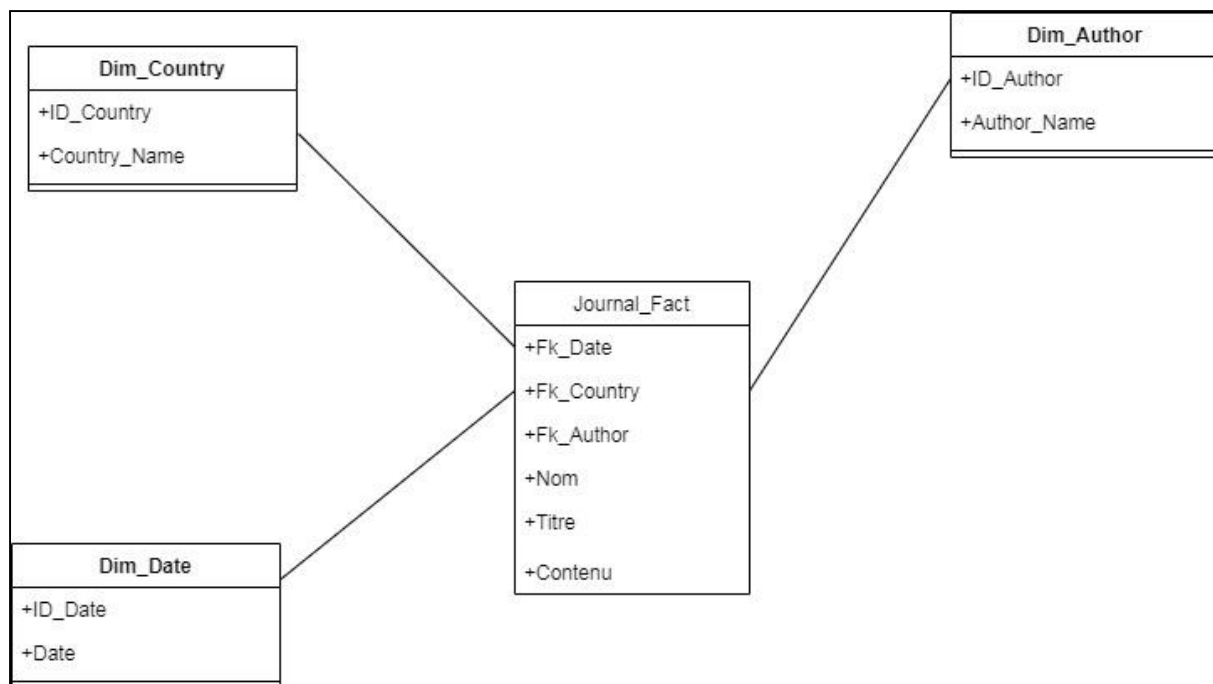


TABLEAU 6: ARCHITECTURE EN ÉTOILE

Conclusion :

En guise de conclusion, nous pouvons dire que l'objectif mis au point au premier sprint est atteint grâce à la modélisation du Data Warehouse. Dans le chapitre suivant, nous allons commencer l'alimentation du Data Warehouse.

Chapitre 5 : Sprint 2 (Alimentation du DW)

Introduction :

Le sprint 2 contient comme objectif d'alimenter le Data Warehouse avec les articles des journaux après une suite de traitements.

I. Backlog du sprint :

Voici le backlog du sprint :

ID	User Stories	Priorité	ID_Tache	Tache	Durée (en heures)
6	En tant que spécialiste d'ETL je stocke les données .	moyenne	1	Je stocke les données dans la base des données Hive	1/2
			2	Je stocke les données depuis les tables HIVE dans les tables des dimensions et de faits dans la base des données multi-dimensionnelle	1/2
7	En tant que spécialiste d'ETL je collecte les données .	moyenne	1	Je prépare la liste des articles depuis les sites web	4
			2	Je copie les articles dans le fichier texte du HDFS.	2

TABLEAU 7: BACKLOG DU SPRINT 2

II. Exécution de l'ETL :

La technologie ETL (extraction, transformation, chargement) est un élément important de la Business Intelligence (BI) d'aujourd'hui, car les données provenant de sources différentes peuvent être réunies au même endroit pour analyser. L'ETL forme presque la totalité du travail, elle est formée de trois étapes :

- **La phase d'Extraction** consiste à collecter les données en provenance d'un ou plusieurs sources
- **La phase de transformation** consiste à reformater et à transformer les données
- **La phase de chargement (loading)** consiste à transférer les données transformées vers la Data Warehouse

1. Phase d'extraction :

Lors de cette phase, nous allons d'abord extraire les données depuis plusieurs sites comme par exemple les sites web suivantes :



FIGURE 11: NEW YORK TIMES

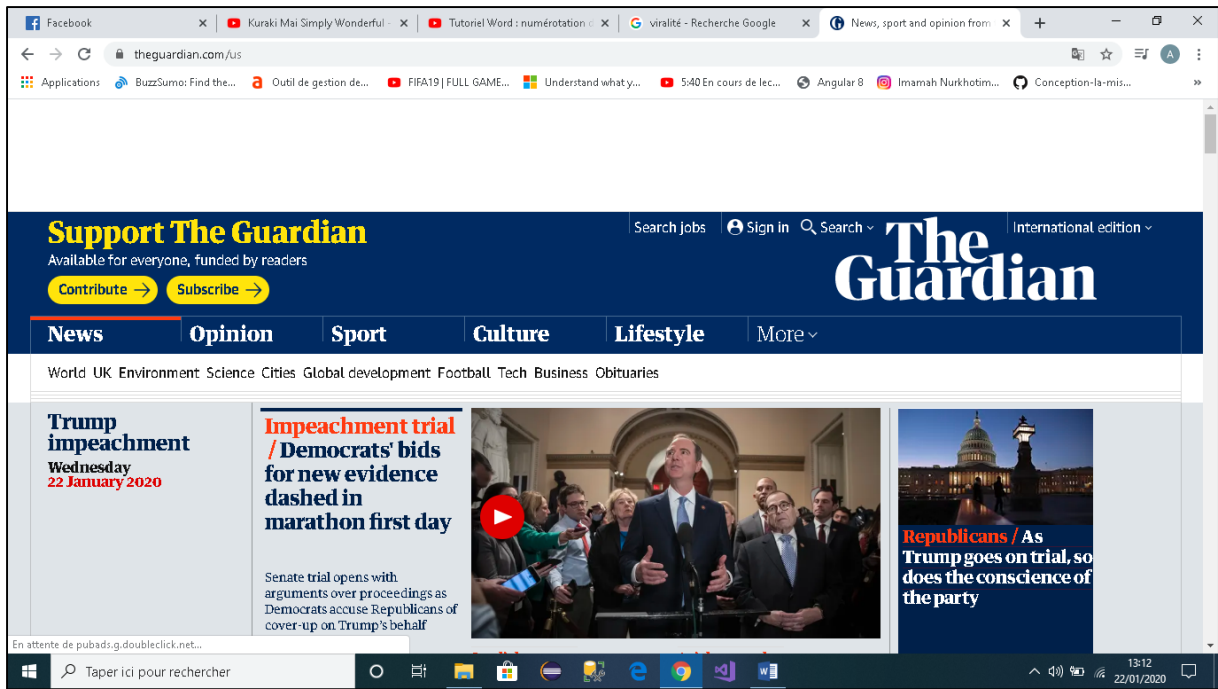


FIGURE 12: THE GUARDIAN

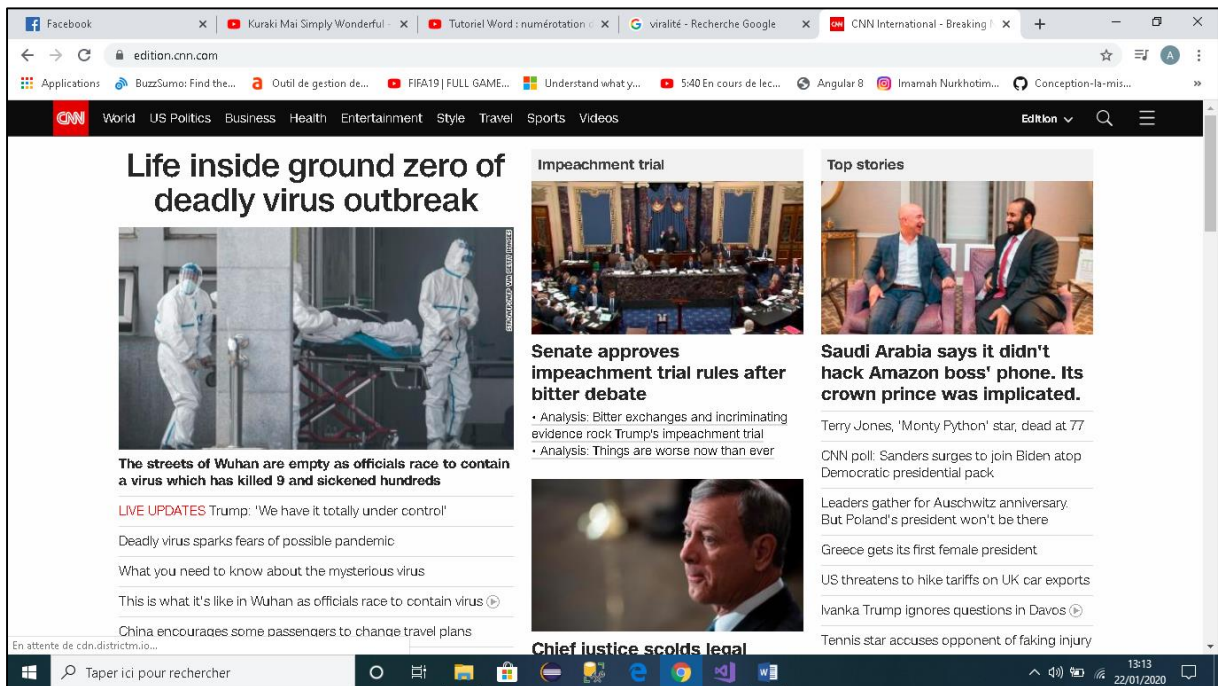
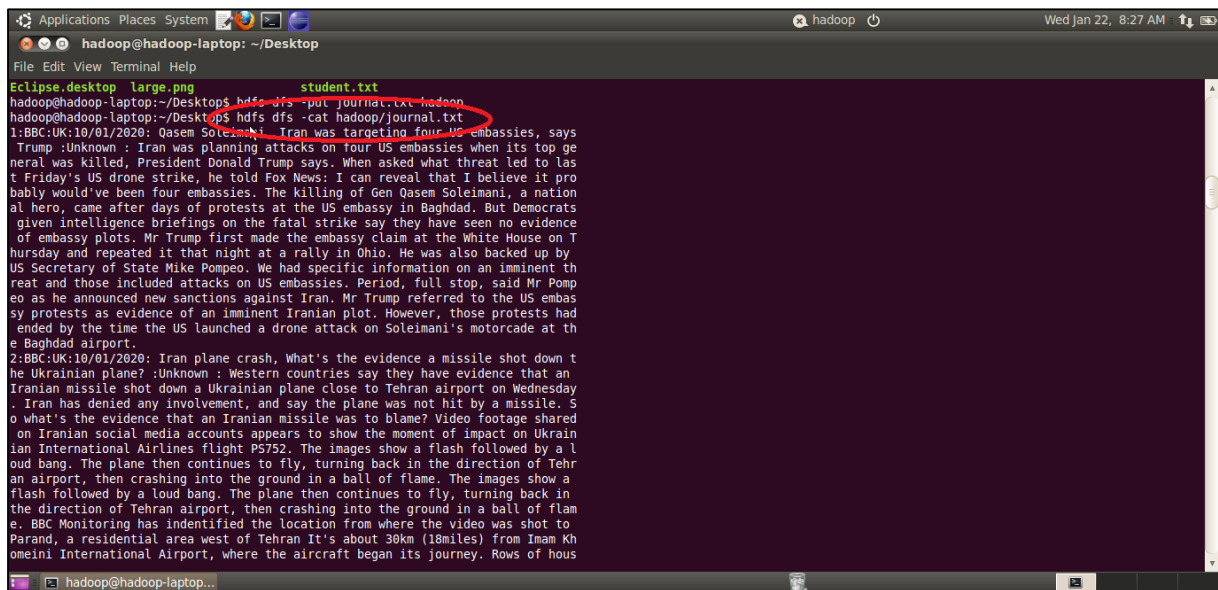


FIGURE 13: CNN

Ensuite, l'ensemble des articles obtenus dans les sites vont être copiés dans le fichier texte journal.txt qui se trouve dans le HDFS comme le montre la figure suivante :



```
Applications Places System | hadoop | Wed Jan 22, 8:27 AM
hadoop@hadoop-laptop: ~/Desktop
File Edit View Terminal Help
Eclipse.desktop large.png student.txt
hadoop@hadoop-laptop:~/Desktop$ hdfs dfs -put journal.txt hadoop
hadoop@hadoop-laptop:~/Desktop$ hdfs dfs -cat hadoop/journal.txt
1:BBC:UK:10/01/2020: Qasem Soleimani: Iran was targeting four US embassies, says
Trump :Unknown : Iran was planning attacks on four US embassies when its top ge
neral was killed, President Donald Trump says. When asked what threat led to las
t Friday's US drone strike, he told Fox News: I can reveal that I believe it pro
bably would've been four embassies. The killing of Gen Qasem Soleimani, a nation
al hero, came after days of protests at the US embassy in Baghdad. But Democrats
given intelligence briefings on the fatal strike say they have seen no evidence
of embassy plots. Mr Trump first made the embassy claim at the White House on T
hursday and repeated it that night at a rally in Ohio. He was also backed up by
US Secretary of State Mike Pompeo. We had specific information on an imminent th
reat and those included attacks on US embassies. Period, full stop, said Mr Pomp
eo as he announced new sanctions against Iran. Mr Trump referred to the US embas
sy protests as evidence of an imminent Iranian plot. However, those protests had
ended by the time the US launched a drone attack on Soleimani's motorcade at th
e Baghdad airport.
2:BBC:UK:10/01/2020: Iran plane crash, What's the evidence a missile shot down t
he Ukrainian plane? :Unknown : Western countries say they have evidence that an
Iranian missile shot down a Ukrainian plane close to Tehran airport on Wednesday
. Iran has denied any involvement, and say the plane was not hit by a missile. S
o what's the evidence that an Iranian missile was to blame? Video footage shared
on Iranian social media accounts appears to show the moment of impact on Ukrain
ian International Airlines flight PS752. The images show a flash followed by a l
oud bang. The plane then continues to fly, turning back in the direction of Tehr
an airport, then crashing into the ground in a ball of flame. The images show a
flash followed by a loud bang. The plane then continues to fly, turning back in
the direction of Tehran airport, then crashing into the ground in a ball of flam
e. BBC Monitoring has identified the location from where the video was shot to
Parand, a residential area west of Tehran It's about 30km (18miles) from Imam Kh
omeini International Airport, where the aircraft began its journey. Rows of hous
```

FIGURE 14: CONTENU DU FICHIER JOURNAL.TXT SUR HDFS

2. Phase de Transformation :

Suite à la phase d'extraction, nous allons découper ses données par article et les stocker dans une table globale dans une base de données HIVE multidimensionnelle qui contiendra aussi les tables des dimensions et la table de faits, dont le type des données est non relationnel et ainsi il n'y a pas de relations entre ces tables, pourtant ils peuvent se relier par références, voici la base de données :

```

hadoop@hadoop-laptop:~$ hive
Logging initialized using configuration in file:/home/hadoop/Training/CDH4/hive-0.8.1-cdh4.0.0/conf/hive-log4j.properties
Hive history file=/tmp/hadoop/hive_job_log_hadoop_202001220832_780660801.txt
hive> use pds ;
OK
Time taken: 6.707 seconds
hive> show tables ;
OK
author dim
country dim
date dim
journal
journal_fact
Time taken: 0.791 seconds
hive>

```

FIGURE 15: CONTENU DE LA BASE DES DONNÉES MULTIDIMENSIONNELLE

et voici le contenu du table globale journal:

```

hive> select * from journal ;
OK
1  BBC  UK  10/01/2020  Qasem Soleimani, Iran was targeting four US embassies, says Trump  Unknown  Iran was planning attacks on four US em
bassies when its top general was killed, President Donald Trump says. When asked what threat led to last Friday's US drone strike, he told Fox News
2  BBC  UK  10/01/2020  Iran plane crash, What's the evidence a missile shot down the Ukrainian plane?  Unknown  Western countries say t
hey have evidence that an Iranian missile shot down a Ukrainian plane close to Tehran airport on Wednesday, Iran has denied any involvement, and say the plane was not h
it by a missile. So what's the evidence that an Iranian missile was to blame? Video footage shared on Iranian social media accounts appears to show the moment of impact
on Ukrainian International Airlines flight PS752. The images show a flash followed by a loud bang. The plane then continues to fly, turning back in the direction of Te
hran airport, then crashing into the ground in a ball of flame. The images show a flash followed by a loud bang. The plane then continues to fly, turning back in the di
rection of Tehran airport, then crashing into the ground in a ball of flame. BBC Monitoring has indentified the location from where the video was shot to Parand, a resi
dential area west of Tehran It's about 30km (18miles) from Imam Khomeini International Airport, where the aircraft began its journey. Rows of housing blocks, a construc
tion site and a storage tank, which all appear in the video, match a Google Earth image of the location, according to BBC Monitoring.
3  BBC  UK  11/01/2020  Sultan Qaboos of Oman, Arab world's longest-serving ruler, dies aged 79  Unknown  The sultan deposed his father i
n a bloodless coup with British support in 1970 and set Oman on a path to development, using its oil wealth. Widely regarded as popular, he was also an absolute monarch
and any dissenting voices were silenced. No cause of death has been confirmed. His cousin Haitham bin Tariq Al Said has been sworn in as his successor. The former cult
ure and heritage minister took the oath of office on Saturday after a meeting of the Royal Family Council, the government said. The sultan is the paramount decision-mak
er in Oman. He also holds the positions of prime minister, supreme commander of the armed forces, minister of defence, minister of finance and minister of foreign affai
rs. Last month Sultan Qaboos - who had no heir or designated successor - spent a week in Belgium for medical treatment, and there were reports he was suffering from can
cer. With great sorrow and deep sadness... the royal court mourns His Majesty Sultan Qaboos bin Said, who passed away on Friday, a court statement said earlier, announci
ng three days of national mourning. Images showed a crowd of men gathered outside the Sultan Qaboos Grand Mosque in the capital, Muscat, where the casket had been take
n and prayers were being held.
4  BBC  UK  10/01/2020  Qasem Soleimani Why his killing is good news for IS jihadists  Jeremy Bowen  The Islamic State (IS) group has welcomed the d
eath of Iranian general Qasem Soleimani, the head of the elite Quds Force. In a statement, it described the general's demise as an act of divine intervention that benef
itted jihadists. However, it made no mention at all of the US, which carried out the deadly drone strike against Soleimani in Baghdad on 3 January. President Donald Tru
mp's decision to assassinate Gen Soleimani set off a chain of consequences - one of the first was on the unfinished war against jihadists. Almost immediately the US-led
coalition fighting IS suspended operations in Iraq. The US and its allies announced that their main job was now defending themselves. From a military point of view, th
ey probably had no choice. Iran and the militias it sponsors here in Iraq have sworn vengeance for the killings caused by the missile fired by a US drone at Soleimani's

```

FIGURE 16: CONTENU DE LA TABLE GLOBALE JOURNAL

3. Phase de Chargement :

Enfin, ses données vont être transités dans les tables dimensions et dans la table fait que nous avons déjà créer, voici la composition du Data Warehouse (qui est la meme base des données)

Et voici le contenu des tables dimensions :

```
Applications Places System hadoop Wed Jan 22, 8:35 AM
hadoop@hadoop-laptop: ~
File Edit View Terminal Help
22 UK
23 UK
24 UK
Time taken: 0.213 seconds
hive> select * from author_dim ;
OK
1 Unknown
2 Unknown
3 Unknown
4 Jeremy Bowen
5 Unknown
6 James Brown
7 James Brown
8 James Brown
9 James Brown
10 James Brown
11 Martin Chris
12 Martin Chris
13 Alex Lin
14 Alex Lin
15 Amanda Jackson
16 Joshua Berlinger
17 Joshua Berlinger
18 Jennifer Hansler
19 Mike Conte
20 Thomas Gibbons-Neft
21 Michel Beaubien
22 Unknown
23 Andrew Brwen
24 Adam Forrest
Time taken: 0.195 seconds
hive>
```

FIGURE 17: DIMENSION AUTEUR

```
Applications Places System hadoop Wed Jan 22, 8:35 AM
hadoop@hadoop-laptop: ~
File Edit View Terminal Help
22 18/01/2020
23 18/01/2020
24 18/01/2020
Time taken: 0.205 seconds
hive> select * from country_dim ;
OK
1 UK
2 UK
3 UK
4 UK
5 France
6 France
7 France
8 France
9 Qatar
10 Qatar
11 Qatar
12 Qatar
13 USA
14 USA
15 USA
16 USA
17 USA
18 USA
19 USA
20 USA
21 USA
22 UK
23 UK
24 UK
Time taken: 0.213 seconds
hive>
```

FIGURE 18: DIMENSION PAYS

```
Applications Places System | hadoop@hadoop-laptop: ~
File Edit View Terminal Help
date_dim
journal
journal_fact
Time taken: 0.31 seconds
hive> select * from date_dim ;
OK
1      10/01/2020
2      10/01/2020
3      11/01/2020
4      10/01/2020
5      11/01/2020
6      10/01/2020
7      08/01/2020
8      01/01/2020
9      11/01/2020
10     11/01/2020
11     13/01/2020
12     13/01/2020
13     13/01/2020
14     13/01/2020
15     13/01/2020
16     13/01/2020
17     13/01/2020
18     13/01/2020
19     13/01/2020
20     13/01/2020
21     18/01/2020
22     18/01/2020
23     18/01/2020
24     18/01/2020
Time taken: 0.305 seconds
hive>
```

FIGURE 19: DIMENSION DATE

Conclusion :

Dans ce chapitre, nous avons terminé l'extraction, la transformation et le chargement des données. Dans le chapitre suivant, nous allons aborder la mise en place du Dashboard.

Chapitre 6 : Sprint 3 (Mise en place d'un Dashboard)

Introduction :

Après avoir déterminé les besoins des décideurs, nous allons procéder à la génération des tableaux de bord demandés. Au cours de ce chapitre, nous allons donc mettre en place les tableaux de bord extraits des données qui résident dans le DW et faire les analyses nécessaires.

I. Sprint Backlog :

ID	User Stories	Priorité	ID_Tache	Taches	Durée (par heure)
2	En tant que décideur je veux avoir le pouvoir de filtrer les journaux par pays.	élevé	1	Connexion du Data Warehouse avec le programme JAVA.	1/2
			2	Détermination des graphiques à utiliser.	2
			3	Génération du table de bord de l'analyse par pays.	1/2
3	En tant que décideur je veux avoir le pouvoir de filtrer les journaux par Auteur.	élevé	1	Connexion du Data Warehouse avec le programme JAVA.	1/2
			2	Détermination des graphiques à utiliser.	2
			3	Génération du table de bord de l'analyse par Auteur.	1/2

4	En tant que décideur je veux avoir le pouvoir de filtrer les journaux par Date.	élevé	1	Connection du Data Warehouse avec le programme JAVA.	1/2
			2	Détermination des graphiques à utiliser.	2
			3	Génération du table de bord de l'analyse par Date	1/2

TABLEAU 8: SPRINT BACKLOG DU SPRINT 3

II. Création des tableaux de bord :

Après la connexion de notre base des données multidimensionnelles avec notre programme JAVA, nous avons analysé les données comme a été cité dans le sprint backlog avec la librairie SWING, et ci-suit on trouve les résultats des différentes analyses :



FIGURE 20: PAGE D'ACCEUIL DE L'APPLICATION JAVA

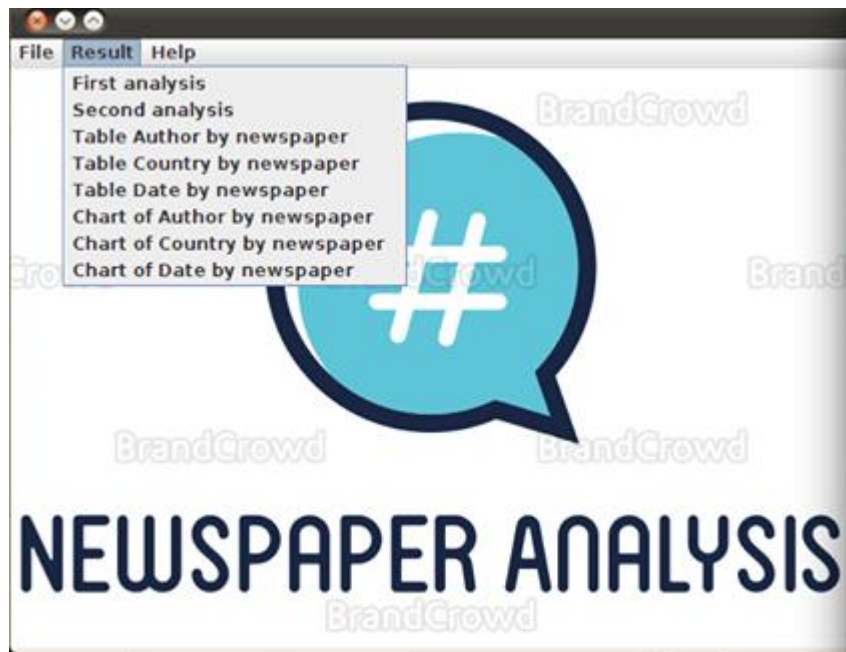


FIGURE 21: MENU DE L'APPLICATION

Author	Count
Adam Forrest	1
Alex Lin	2
Amanda Jackson	1
Andrew Brwen	1
James Brown	5
Jennifer Hansler	1
Jeremy Bowen	1
Joshua Berlinger	2
Martin Chris	2
Michel Beaubien	1
Mike Conte	1
Thomas Gibbons-Neff	1
Unknown	5

FIGURE 22: TABLE DES ARTICLES PAR AUTEUR

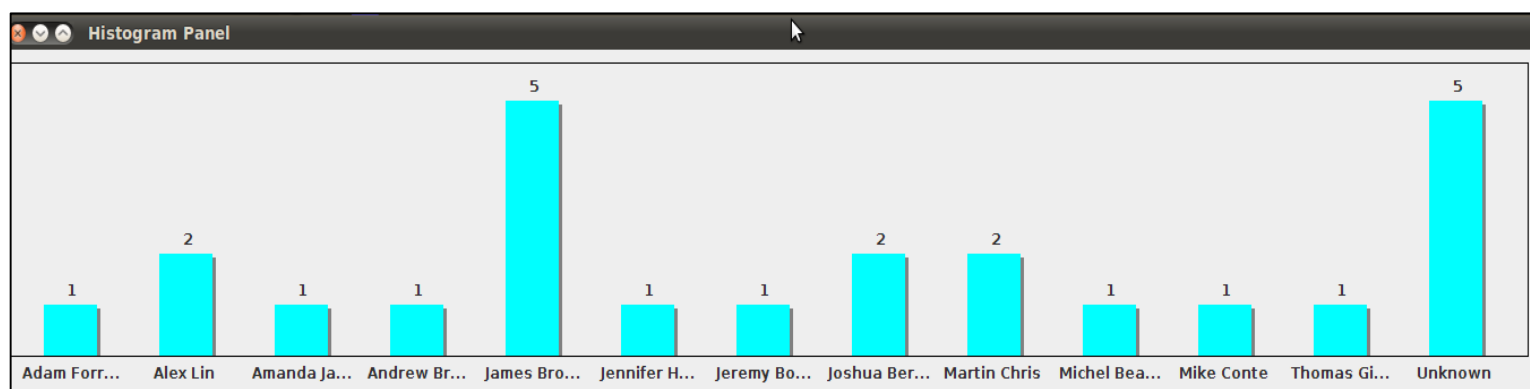


FIGURE 23: HISTOGRAMME DES ARTICLES PAR AUTEUR

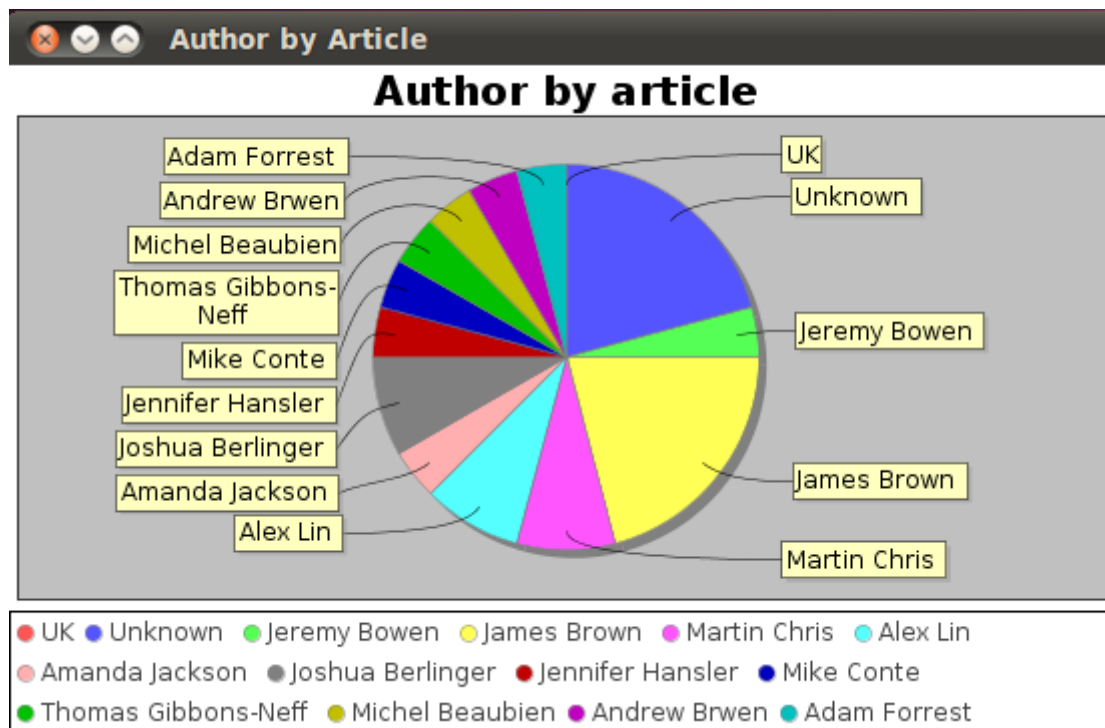


FIGURE 24: DIAGRAMME CIRCULAIRE DES ARTICLES PAR AUTEUR

Country	Count
France	4
Qatar	4
UK	7
USA	9

FIGURE 25: TABLE DES ARTICLES PAR PAYS

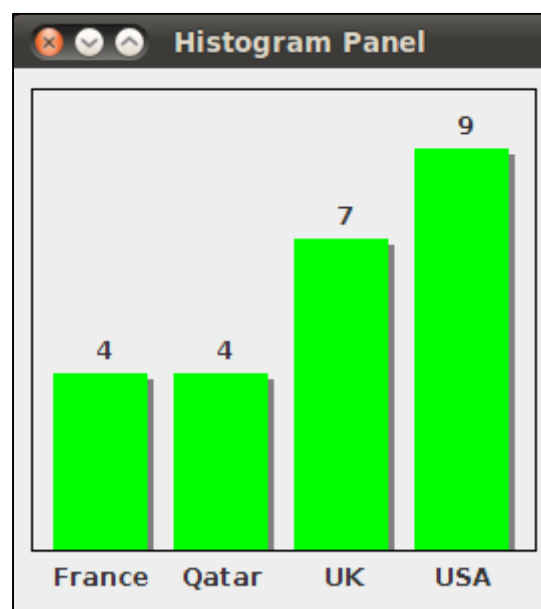


FIGURE 26: HISTOGRAMME DES ARTICLES PAR PAYS

Date	Count
13/01/2020	4
18/01/2020	1
01/01/2020	1
08/01/2020	1
10/01/2020	4
11/01/2020	4
13/01/2020	6
18/01/2020	3

FIGURE 27: TABLE DES ARTICLES PAR DATE

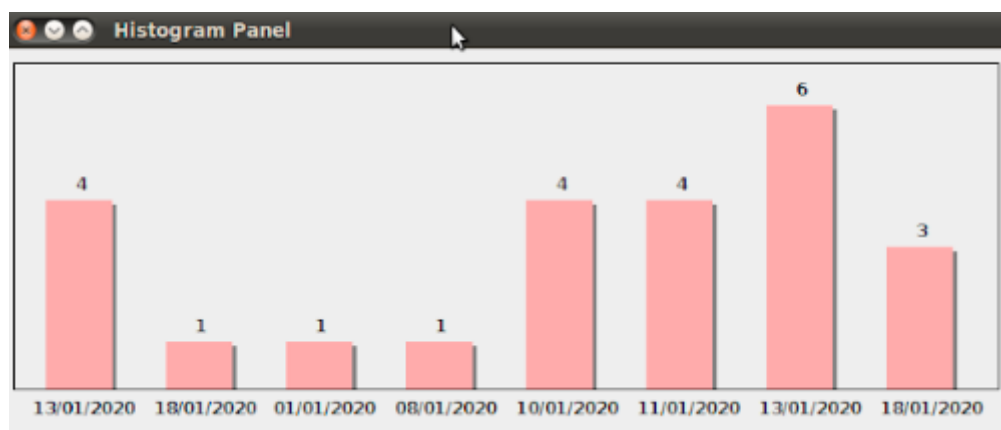


FIGURE 28: HISTOGRAMME DES ARTICLES PAR DATE

Conclusion :

Dans ce chapitre, nous avons présenté les tableaux de bord conçus grâce à la librairie SWING de JAVA et qui répondent aux besoins spécifiques des analyses demandées.

Conclusion Générale

Ce stage représente pour nous une expérience enrichissante dans la mesure où il nous a permis d'enrichir nos compétences techniques et de découvrir de plus près notre domaine dans le milieu académique.

Grâce à ce projet de spécialité, nous avons appris à mener un projet BI du début à la fin.

Le début était difficile, étant donné que nous étions dans l'obligation de se familiariser avec les nouvelles technologies.

Ensuite, nous avons élaboré notre plan d'exécution avec Scrum. Nous avons commencé par la compréhension du produit. Puis, nous avons fait une répartition de tâches, ce qui nous a permis de terminer le travail à temps. Non seulement ce travail nous a permis de s'adapter à la méthodologie Scrum mais aussi de la modeler efficacement en fonction de notre projet.

Pour finir, nous avons mis au point un data warehouse qui nous a mené à développer les tableaux de bord en fonction des analyses demandées.

Pour un temps futur, nous pouvons améliorer cette application par l'inclure dans un site web à la disposition de tous les utilisateurs de l'internet ou aussi nous pouvons développer un logiciel complet qui se basera sur notre DW et qui appliquera automatiquement toutes les analyses.

Webographie

<https://cwiki.apache.org/confluence/display/Hive/GettingStarted#GettingStarted-CompileHivePriorto0.13onHadoop0.20>

<https://cwiki.apache.org/confluence/display/Hive/GettingStarted>

<https://www.quora.com/What-is-main-differences-between-hive-vs-pig-vs-sql>

<https://cwiki.apache.org/confluence/display/Hive/LanguageManual>

<https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Cli>

<https://www.lebigdata.fr/apache-hive-definition>

<https://www.edureka.co/community/53170/why-is-hive-called-as-data-warehouse>

<https://www.lepoint.fr/dossiers/societe/rumeurs-complots-et-fake-news-comment-lutter/>

<https://la-rem.eu/2018/03/une-future-loi-pour-lutter-contre-les-fake-news/>

<https://la-rem.eu/2018/03/une-future-loi-pour-lutter-contre-les-fake-news/>

<https://www.udemy.com/course/fundamentals-of-business-intelligence-bi/>