

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche
Scientifique
Université Abderrahman mira - Bejaia -
Faculté des Sciences exactes
Département d'Informatique



MÉMOIRE DE MASTER PROFESSIONNEL

En
Informatique

Option

Administration et sécurité des réseaux

Thème Intégration d'un cluster automatisé

**Elastic Search pour l'optimisation des recherches
dans un environnement Hadoop**

Présenté par: Mlle.ATROUN Ilhem
Mlle.BOUSSEKINE Kenza

Déposé le 28 octobre 2020 devant le jury composé de:

Co_Encadreur	M. N.ELSAKAAN	Doctorant LMD	U.A/Mira Béjaia.
Encadreur	M. A.TARI	Professeur	U. A/Mira Béjaia.
Examinatrice	Mme S.Bennai	Doctorante LMD	U. A/Mira Béjaia.
Examineur	M. K.Amroun	Maître de conférences A	U. A/Mira Béjaia.

Promotion 2019-2020.

- *Remerciements* -

Nous exprimons notre profonde reconnaissance à
notre Encadreur

M. ELSAKAAN Nadim

pour sa patience, son infinie sympathie, sa
disponibilité, malgré ses nombreuses responsabilités,
et ses judicieux conseils qui nous ont été d'une aide
précieuse.

Nous remercions également et particulièrement notre
Encadreur principal M. Tari, qui était notre enseignant
avant tout, pour l'expérience estudiantine qui nous a
permis d'avoir.

Nous tenons à témoigner de notre pleine gratitude
aux personnes de Cevital qui nous ont accompagnés
au long de notre stage au sein de cette entreprise et
qui nous ont accordés leur temps pour nous aider à
réaliser ce travail.

M. Ghidouche Athmane et M. Touloum Nadir.

Nous remercions chacun des examinateurs pour
l'intérêt porté à notre travail en acceptant de le
corriger et de l'enrichir avec leurs propositions.

Enfin, nous adressons nos plus sincères
remerciements à tous nos proches et amis qui nous
ont soutenus et encouragés tout au long de notre
cursus universitaire et de la réalisation de ce mémoire.

Dédicace

A **Mes chers parents**, pour tout leurs sacrifices, leurs encouragements, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études pour me voir réussir.

A **mes soeurs et mon frère** pour leurs encouragements permanents et leur soutien moral.

A **toute ma famille** pour leur présence dans toutes les circonstances et leur soutien permanent durant mon parcours universitaire.

A **mes amis** qui ont contribué, de près ou de loin à la réalisation de ce travail et à l'accomplissement de toutes ses années d'études qui resteront gravés à jamais.

A ma formidable binôme et amie **Ilhem** qui, sans elle, rien de tout ça n'aurait été possible.

Que ce travail soit le fruit de vos vœux tant allégés. Merci d'être toujours là pour moi.

Kenza

Dédicace

je dédie ce modeste travail accompagné d'un profond amour à celle qui m'a arrosée de tendresse, et l'espoir à ma source d'amour **ma mère**.

A mon support dans la vie **mon père**.

A mes très chères et adorables soeurs . puisse dieu vous donner amour, bonheur et réussite. **Sara & Imen**.

A la meilleure personne que je connaisse , ma soeur, mon amie, ma confidente, à toi ma chère binome qui a toujours été la pour moi dans le meilleur et le pire **Kenza**.

A ceux qui m'ont toujours aidé et encouragé, qui étaient toujours à mes côtés, qui m'ont accompagné durant mon chemin d'études supérieures.

A tous ceux qui, d'une manière ou d'une autre, ont contribué à l'élaboration de ce mémoire ainsi qu'à la réussite de cette formidable année universitaire, mes aimables amis à vous **Mira & Nadim**

A tous ceux qui compte pour moi et qui n'ont pas pu être cités ici.

Ilhem

Intégration d'un cluster automatisé Elastic
Search pour l'optimisation des recherches dans
un environnement Hadoop

Sommaire

Introduction générale	5
1 Le système d'information dans le monde du travail	7
1 Introduction	7
2 Première partie: Parc informatique et ses composants	8
2.1 Parc materiel	8
2.2 Parc logiciel	8
3 Deuxième partie: Les applications de gestions orientées métiers	8
3.1 L'urbanisation des SI	8
3.1.1 Les motivations de l'urbanisation	9
3.1.2 Objectifs	9
3.2 Les services web	9
3.3 ERP	10
3.4 Gestion de la Relation Client	12
3.5 Gestion de la Production Assistée pas Ordinateur	13
3.6 Gestion de la Maintenance Assisté par Ordinateur	13
3.6.1 Fonctionnalités de la GMAO	14
3.7 Transport Management System	14
3.7.1 Fonctionnalités	14
4 Troisième partie : Le mouvement NoSQL au sein de l'entreprise	15
4.1 NoSQL	15
4.2 HADOOP	16
4.2.1 Les avantages d'Hdoop	16
4.2.2 Architecture	16
4.2.3 HDFS	17
4.2.4 MapReduce	17
4.2.5 PIG et HIVE	18
4.3 Spark	18
4.3.1 Resilient Distributed Dataset	19
4.4 Kafka	20
4.4.1 Architecture et concepts de Kafka	20
4.4.2 les API's de Kafka	20
4.5 ELK Stack(Elasticsearch, Logstash, Kibana)	21
4.5.1 Elasticsearch	21

4.5.2	Logstash	22
4.5.3	Kibana	22
4.6	Kubernetes	23
4.7	Docker	24
5	Conclusion	24
2	Introduction du cas d'étude	25
1	Introduction	25
2	Présentation de la société d'accueil Cevital	25
2.1	Organigramme du groupe Cevital	26
2.1.1	Les missions	27
2.1.2	Les activités	27
2.1.3	Les objectifs	27
2.2	L'informatique dans Cevital	27
2.3	Présentation de l'organisme d'accueil	27
3	Etat des lieux	29
3.1	Présentation de l'infrastructure réseau	29
3.2	Présentation du GMAO de Cevital	30
3.3	Présentation de l'ERP Sage	31
3.4	Présentation des processus Métiers gérant la production d'énergie	31
3.4.1	L'unité d'énergie et utilités	31
3.5	Présentation du TMS de Numilog	32
4	Conclusion	34
3	Mise en place de la solution	35
1	Introduction	35
2	Infrastructure	36
3	Installation de Fedora sur VM Ware	36
4	Installation de IntelliJ IDEA	38
5	Installation de SPARK	40
6	Installation d'Elasticsearch	41
7	Installation de Kibana	44
4	Déploiement et exploitation de la solution	46
1	Introduction	46
2	Extraction des données de l'UPE	46
3	Stockage via Spark	46
4	Restitution des Données via elasticsearch et kibana	47
5	Conclusion	48
	Conclusion générale	49

Table des Figures

1.1	facteurs d'urbanisation	9
1.2	Service web	10
1.3	Fonctionnalités d'ERP	11
1.4	Comparatif des ERP	12
1.5	Fonctionnalités de CRM	13
1.6	comparaison de TMS	15
1.7	les logos d'SGBD	16
1.8	Apache Spark.	19
1.9	Architecture d'Apache Kafka.	21
1.10	les services de Kubernetes	23
2.1	organigramme du groupe Cevital	26
2.2	organigramme de la direction Système d'information	28
2.3	Infrastructure réseau de Cevital.	29
2.4	Interface de Coswin 8i.	30
2.5	Organigramme du pole d'énergie	32
2.6	Transport Mangement System	33
3.1	Infrastructure	36
3.2	Lancement de Fedora.	37
3.3	connexion sur Fedora.	37
3.4	fin d'installation d'IntelliJ	38
3.5	démarrage d'IntelliJ IDEA	38
3.6	configuration ItelliJ	39
3.7	Installation de Scala	39
3.8	Installation de SPARK	40
3.9	emplacement SPARK	40
3.10	configuration SPARK	41
3.11	Variable d'environnement SPARK	41
3.12	Installation d'Elasticsearch	42
3.13	Démarrage d'Elasticsearch	42
3.14	Interface d'Elasticsearch	43
3.15	En-tête Elasticsearch.	43
3.16	Démarrage de Kibana	44
3.17	Interface de Kibana	44

TABLE DES FIGURES

4.1	stockage via spark	47
4.2	restitution de donnees via kibana	48

Introduction générale

La révolution numérique a accompagnée depuis des années le développement de différents secteurs d'activités industriels et plus particulièrement l'informatique, qui est devenu indispensable dans tout les domaines. Effectivement, le développement informatique permet de bénéficier de programmes spécialisés. Chaque client et chaque entreprise possède ces propres spécifications par rapport aux produits (application ou logiciel) qu'il voudrait développer et le développement informatique existe pour, justement, permettre de réaliser des programmes unique à chaque demande. Pour celà, il faudra bien comprendre et analyser les besoins de cette entreprise et tout ce qui se rapporte au programme à développer mais pas que ! Le développeur devra aussi prendre compte des spécifications reliées à l'emplacement de l'entreprise, son secteur, avec qui ils travaillent, dans quel domaine et leur capacité d'accès à internet.

Le marché de l'emploi a été façonné par l'évolution importante des technologies et la concurrence qui s'est installée en masse. L'insertion des outils informatique est devenu incontournable dans tout les domaines, ainsi le monde industriel devient moins complexe et ses machines plus fluides et efficaces, notamment dans le domaine de l'agro-alimentaire, pour une entreprise international telle que Cevital, celà a changé le cour des choses en permettant une meilleure organisation, une meilleure réactivité et une meilleure productivité.

Cevital est un groupe familial de plusieurs sociétés. c'est la première entreprise privée algérienne à avoir investi dans des secteurs d'activités diversifiés dont l'industrie agro-alimentaire, qui compte une raffinerie d'huile et une autre de sucre. Cette société ayant une très grande ampleur internationale industrielle et économique possède une direction du système d'information qui est l'une des directions les plus importantes et des plus convoitée étant donné la quantité de données à gérer et le repos de tout les secteurs sur l'informatique.

En effet, Cevital met en place une stratégie novatrice et une infrastructure adéquate pour analyser et gérer le flux de données énorme qui constitue une sources d'informations précieuse à véhiculer en toute sécurité et en temps réel. Le Big Data est une technologie conçu spécialement pour des sociétés comme Cevital, ses architectures sont très complexes et gigantesques en termes de volume et d'espace occupé et la gestion de tout celà demande une conception bien étudié en choisissant les technologies et les logiciels adéquats aux besoins de l'entreprise à court terme et à long terme en combinant efficacité, rapidité et rentabilité.

L'objectif principal de ce travail est

- Comprendre les besoins métiers et numérique des entreprises industrielles.
- Etudier le modèle de gestion et restitution des masses de données générées par l'activité de l'entreprise.
- Découvrir et se familiariser avec un nouveau concept informatique.

Nous allons donc, programmer en Scala sur IntelliJ, introduire les fichier de données à traiter, nous allons utilisé le moteur de recherche Elasticsearch pour le stockage et la recherche et Enfin Kibana pour la visualisation des résultats.

Dans notre mémoire, Nous aborderons le parc informatique et ses composants, les applications de gestion orientées métiers, le mouvement NOSQL au sein de l'entreprise et les technologies utilisées dans le Big Data dans un premier lieu. Ensuite nous ferons une introduction au cas d'étude et tout ce qui concerne la société d'accueil en passant par l'état des lieux. Nous présenterons la mise en place de la solution et enfin, le déploiement et l'exploitation de la solution proposée.

Chapitre 1

Le système d'information dans le monde du travail

1 Introduction

L'environnement des entreprises est caractérisé par le développement technologique, la complexité des relations, l'augmentation des opportunités et la multiplication des risques. Ce contexte a entraîné un accroissement important du volume des informations, ce qui handicape l'entreprise c'est le problème d'obtenir rapidement des informations utiles afin de répondre à des situations en temps réel.

A présent, la croissance d'une entreprise passe forcément par un grand volume d'activité et donc d'une grande quantité d'information à gérer et dont il faudra tirer le meilleur parti pour une bonne gestion.

Les entreprises, grandes ou petites soient-elles ont adopté des technologies de plus en plus sophistiquées et leurs données ne cessent de croître et de devenir plus complexes par leurs différence de format et de contexte, il est donc plus difficile de les gérer avec un seul et unique algorithme, c'est pour cela que la révolution du Big Data a été sollicité et ce qui décrit ce concept sont les 3 vecteurs dont nous allons parlé [8] :

- Le Volume : Quand on dit Big Data, on parle forcément de volume et étant donné les différentes informations à stocker, celles-ci atteignent des proportions inimaginable! On aura donc besoin d'énormes espaces de stockage, c'est le VOLUME du Big Data.
- La Vitesse : C'est la mesure de vitesse de circulation des fichiers. Le problème que rencontraient les différentes entreprises est le manque ou l'absence de traitement des données par manque de rapidité, l'intégration du Big Data a révolutionné ça en temps réel.
- La Variété : Vu l'utilisation des données très différentes compte tenu de leurs formats, leurs types et leurs traitements, le Big Data relève le défi de résoudre ces inconvénients.

En réalité le Big Data est un moyen de mieux comprendre et gérer les données existantes avant de n'être simplement qu'un grand nombre de données [2].

2 Première partie : Parc informatique et ses composants

Un parc informatique désigne l'ensemble d'ordinateurs ou autre matériel qui composent l'infrastructure informatique d'une entreprise relié en réseau (imprimante, fax, serveur...)

2.1 Parc matériel

Les matériels dont il est question ici sont bien évidemment les solutions qui peuvent être considérés comme étant des machines, dénommées hardware dans le jargon informatique, à savoir les ordinateurs et tous ses composants internes mais également les matériels accessoires, les scanner, imprimante, webcams, les périphériques tels que les disques durs et graveurs externes ainsi que les clés USB, les commutateurs réseaux communément connus sous la qualification de switch, les différents câbles, les modems nécessaires à la connexion internet, les vidéoprojecteurs et les onduleurs.

2.2 Parc logiciel

Pour ce qui est des programmes ou software, ce sont toutes les applications informatiques qui peuvent être considérées comme étant des logiciels et qui sont nécessaires au fonctionnement des machines (ordinateurs) ainsi que du réseau. En réalité, il n'existe que deux types des logiciels : les programmes applicatifs et ceux qualifiés de système[?]. Dans le contexte de la gestion d'un parc informatique, ils peuvent être subdivisés en trois catégories, à savoir :

- Les softs destinés aux activités professionnelles si l'on ne cite que les outils bureautiques .
- Les programmes dédiés à l'opérationnalité des machines tels que les systèmes d'exploitation ou les anti-virus .
- Les logiciels conçus pour le réseau mais également pour la gestion du parc.

3 Deuxième partie : Les applications de gestions orientées métiers

3.1 L'urbanisation des SI

L'urbanisation des SI transforme progressivement le système d'information afin de le simplifier et de l'optimiser par rapport aux besoins de l'entreprise. L'urbanisation consiste à se doter d'un plan d'ensemble et de règles de constructions pour simplifier et piloter ces changements.

3.1.1 Les motivations de l'urbanisation

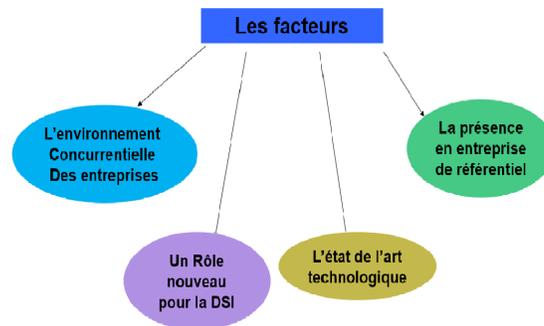


FIGURE 1.1 – facteurs d'urbanisation

3.1.2 Objectifs

La démarche d'alignement stratégique et d'urbanisation d'un Système d'Information vise à répondre à 4 objectifs majeurs [3] :

- Aligner les tâches du Système d'Information sur la stratégie globale de l'entreprise, afin de renforcer la valeur d'usage du Système d'Information et en faire un atout pour l'Entreprise
- Rendre le Système d'Information « lisible » pour favoriser le dialogue et la communication avec les directions métiers, les Maîtrises d'ouvrage (MOA) et les Maîtrises d'œuvre (MOE)
- Disposer d'un Système d'Information agile, capable d'évoluer rapidement et en toute sécurité tant d'un point de vue métier que technique
- Contribuer à la maîtrise des coûts informatiques [3]

3.2 Les services web

C'est une technologie qui permet à des applications de communiquer à distance par le biais d'Internet indépendamment des plates-formes et des langages sur lesquelles elles reposent. Un service Web est un composant logiciel encapsulant des fonctionnalités métier de l'entreprise, et accessible via des protocoles standards basés sur XML et HTTP/SMTP. La figure() représente l'interaction entre le client et le service web.

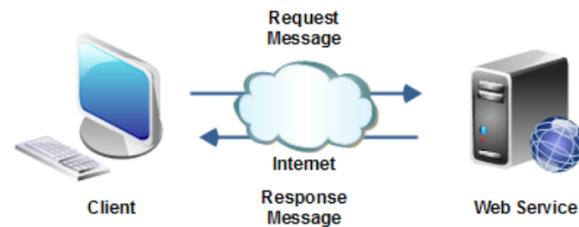


FIGURE 1.2 – Service web

Il existe probablement autant de définitions des Web Services que d'entreprises qui les créent, mais presque toutes ces définitions ont ceci en commun :

- les Web Services proposent aux utilisateurs du Web des fonctionnalités pratiques grâce à un protocole Web standard (dans la plupart des cas, le protocole utilisé est SOAP)
- les Web Services offrent un moyen de décrire leurs interfaces suffisamment en détail pour permettre à un utilisateur de créer une application cliente capable de converser avec eux cette description est généralement fournie dans un document XML nommé WSDL (Web Services Description Language)
- les Web Services sont inscrits afin que les utilisateurs potentiels puissent les trouver facilement. Ceci est possible grâce à UDDI (Universal Discovery Description and Integration).

3.3 ERP

Le terme ERP vient de l'anglais « Enterprise Resource Planning ». ERP a été traduit en français par l'acronyme PGI (Progiciel de Gestion Intégré) et se définit comme un groupe de modules relié à une base de données unique. L'ERP est un progiciel qui permet de gérer l'ensemble des processus opérationnels d'une entreprise en intégrant plusieurs fonctions de gestion : solution de gestion des commandes, solution de gestion des stocks, solution de gestion de la paie et de la comptabilité, solution de gestion e-commerce... dans un système. Autrement dit, l'ERP représente la « colonne vertébrale » d'une entreprise. Pour être qualifiée de « Progiciel de Gestion Intégré », une solution logicielle ERP doit couvrir au moins deux principes fondamentaux qui sont les suivants [10] :

- Construire des applications informatiques sous forme de modules indépendants mais parfaitement compatibles sur une base de données unique et commune.
- L'usage d'un moteur de Workflow permet de définir l'ensemble des tâches d'un processus et de gérer leur réalisation dans tous les modules du système qui en ont besoin.

Fonctionnalités de l'ERP

- La gestion comptable et financière.
- La gestion des stocks.
- La gestion des ressources humaines.
- La gestion des fournisseurs (ERP fournisseurs grande distribution).

- La gestion de la vente.
- La gestion de la distribution.
- La gestion de l'e-commerce. Dans la figure() nous distinguons les différentes fonctionnalités de l'ERP.

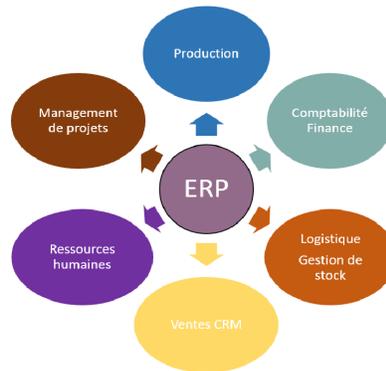


FIGURE 1.3 – Fonctionnalités d'ERP

En complément d'un ERP, et dans le cadre d'une démarche de satisfaction client, de nombreuses entreprises se dotent aussi d'un logiciel de CRM. Cet outil leur permet de mettre en commun et de maximiser la connaissance d'un client donné et, ainsi, de mieux comprendre, anticiper et gérer ses besoins[10].

Comparaison entre les ERP : Parmi les éditeurs les plus connus Sap, Sage et Odoo se livrent une concurrence acharnée par le biais des solutions de gestion commerciale qu'ils proposent. Afin de contribuer à vous apporter une vision plus claire des logiciels de chacun des trois poids lourds de la gestion commerciale, voici un comparatif des différentes fonctionnalités qu'offrent les logiciels Sap, Sage et Odoo :

ERP	ODOO	SAGE	SAP
Meilleur pour	Une suite d'applications open source entièrement intégrée et personnalisable, est destinée aux PME	La Distribution, Vente (détail et gros), Transport, Technologie, Médial... Est destinée aux TPE	Entreprise (Grand Comptes) et filiales cherchant à se développer dans l'avenir sur un noyau numérique avec des processus et des opérations commerciales entièrement intégrés
Evaluation	Très bien	Bien	Très bien
Utilisation	Facile à utiliser	Très facile	Interface conviviale avec des fonctionnalités ambitieuses
Service client	Excellent support client et facilité d'accès		Entièrement intégré à tous les modules critiques
Caractéristiques et Fonctionnalités	Facile à mettre en œuvre et à comprendre	Optimisation des activités tout en mobilité, ultra simple	Fiable et stable, gain de temps, grande efficacité et efficience
Mode	Gratuit	Payant	Payant
Nombre de salariés	10 à 500	50 à 100	+ 500

FIGURE 1.4 – Comparatif des ERP

3.4 Gestion de la Relation Client

GRC en français , ou CRM, Customer Relationship Management en Anglais, est un concept préconisant la centralisation au sein d'une base de données de toutes les interactions entre une entreprise et ses clients. Cela permet de mettre en commun et de maximiser la connaissance d'un client donné et, ainsi de mieux comprendre, anticiper et gérer ses besoins [5]. Il permet :

- D'améliorer votre performance commerciale.
- De connaître avec exactitude votre portefeuille client.
- D'organiser votre force commerciale.
- De gagner du temps dans la recherche d'un client ou d'une information pour vous consacrer davantage à votre activité commerciale.



FIGURE 1.5 – Fonctionnalités de CRM

3.5 Gestion de la Production Assistée pas Ordinateur

La GPAO permet d'améliorer la planification et le suivi des flux physiques et financiers de l'activité de production. Elle permet de piloter la production pour fabriquer des produits conformes aux besoins des clients, en actionnant plusieurs leviers pour abaisser les coûts : Minimiser les stocks et respecter les délais-clients.

- Synchroniser les approvisionnements avec le plan de fabrication.
- Optimiser l'utilisation des surfaces de stockage, des machines, et des équipes de travail. . .
- Opter pour une meilleure gestion ou valorisation des stocks.
- Fiabiliser les dates d'engagement, réagir efficacement aux modifications externes.
- Suivre finement les engagements et les achats, la réalisation des commandes, et les écarts entre coûts de production standards et réels.
- Décrire à un endroit unique les données-produits.

3.6 Gestion de la Maintenance Assisté par Ordinateur

La GMAO permet pour sa part d'assurer le maintien en bon fonctionnement des équipements industriels en planifiant et exécutant la maintenance préventive qui se définit comme étant une mise en oeuvre avant la panne pour éviter l'endommagement des machines, en effectuant la maintenance curative qui est appliquée lorsque le matériel est en panne et consiste à changer celui-ci partiellement ou intégralement selon la gravité de la panne, ou en prenant en charge la maintenance prédictive qui prévoit un programme pour la gestion des problèmes en effectuant la maintenance au bon moment, avec la meilleure économie de moyens employés (pièces de rechange, consommables, main d'œuvre, etc.). Elle offre aussi la possibilité de gérer l'entretien d'un produit complexe chez un client comme un bateau, une machine, un avion, un train. . . Le tableau ci-dessous donne un éclairage sur la ventilation des coûts des outils de GPAO et de GMAO.

3.6.1 Fonctionnalités de la GMAO

- la gestion des équipements.
- la gestion de la maintenance.
- la gestion de la mise en sécurité des installations pour les travaux de maintenance.
- la gestion des stocks de l'équipement de maintenance
- la gestion des achats des pièces détachées.
- la gestion du personnel et le planning.
- la gestion des coûts et budget.
- Indicateurs de clés de performance (key performance indicators : KPI). [6].

Parmi les solutions de GMAO qui permettent la planification des actions de maintenance pour faciliter l'installation des interventions industrielles, nous avons Coswin qui aide à réduire les coûts d'achats et les volumes de stocks, à améliorer la productivité du personnel de maintenance tout en restant conforme aux réglementations en vigueur.

3.7 Transport Management System

Le TMS, ou logiciel de gestion du transport, est un outil d'aide à la gestion du transport. Le TMS répond principalement aux besoins de traçabilité des livraisons et d'optimisation du transport (schémas et affectation des transporteurs). L'ensemble permet d'améliorer l'organisation du transport. Ces améliorations se traduisent par une réduction du budget de transport.

3.7.1 Fonctionnalités

Les TMS ont trois fonctions principales

- l'analyse / l'aide à la décision : Un TMS permet de planifier les expéditions et les tournées, il contribue à définir les schémas de transport les plus efficaces (choix et nombre de plates-formes de rupture/massification).
- le suivi opérationnel des tournées de transport : suivi des expéditions en temps réel (traçabilité), remontée d'alertes en cas de souci, édition des documents de facturation, de booking (réservation des moyens de transport) et des dossiers de litige.
- le reporting : Enfin, les TMS permettent de générer automatiquement des indicateurs de clés de performances logistique (KPI). A titre d'exemple, les TMS sont en mesure de produire d'indicateurs du type : respect des délais de livraison, taux de remplissage des camions, taux d'occupation des moyens de transport, émissions CO2, etc.

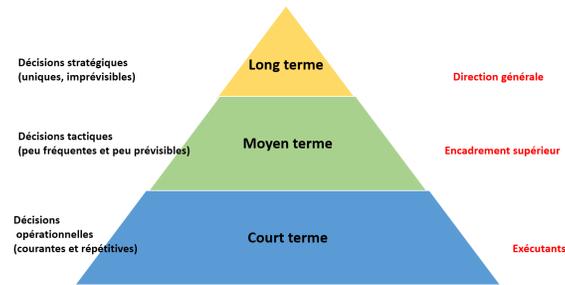


FIGURE 1.6 – comparaison de TMS

4 Troisième partie : Le mouvement NoSQL au sein de l'entreprise

4.1 NoSQL

Appelée également « Not Only SQL » (pas seulement SQL), la base de données NoSQL est une approche de la conception des bases et de leur administration particulièrement utile pour de très grands ensembles de données distribuées.

NoSQL est particulièrement utile lorsqu'une entreprise doit accéder, à des fins d'analyse, à de grandes quantités de données non structurées ou de données stockées à distance sur plusieurs serveurs virtuels du Cloud. Il faut savoir qu'il existe plusieurs types de bases de données "NoSQL"

- Les bases clef/valeur : ce type consiste à stocker les informations sous forme de clef/valeur, elles sont très performantes compte tenu de leur simplicité (exemple : redis, Membase, voldemort...)
- Les bases orientées colonnes : les données sont sauvegardées sous forme de lignes avec des colonnes, où le nombre de celles-ci peuvent être différents d'une ligne à une autre (exemple : HBase, Cassandra).
- Les bases orientées document : les données sont stockées sous forme d'objet XML ou JSON et la récupération se fait de manière hiérarchique (les plus connus sont : CouchDB, RavenDB, MangoDB).
- Les bases orientées graphe : sont faites pour les relations complexes car les données sont sous forme de nœuds et de relations (exemple : Neo4J).



FIGURE 1.7 – les logos d’SGBD

4.2 HADOOP

C'est un framework Java open source géré par Apache Software Foundation et sous licence Apache 2.0. qui peut être considéré comme un système de traitement de données évolutif pour le stockage et le traitement par lot de très grande quantité de données. Les données sont stockées sur des serveurs standard peu coûteux configurés en clusters.

Le système de fichiers distribué Hadoop supporte des fonctionnalités de traitement concurrent et de tolérance aux incidents. Développé par Doug Cutting et Michael J. Cafarella, Hadoop utilise le modèle de programmation MapReduce pour stocker et récupérer plus rapidement les données dans ses nœuds.

4.2.1 Les avantages d’Hdoop

Hadoop est une solution miracle pour l’analytique et le Big Data, aide à relever le défi de l’énormité des big data grâce à ses nombreuses qualités :

- Robuste : si un nœud de calcul tombe, ses tâches sont automatiquement réparties sur d’autres nœuds. Les blocs de données sont également répliqués
- Virtualisation : ne plus se reposer directement sur l’infrastructure physique (baie de stockage coûteuse), mais choisir la virtualisation de ses clusters Hadoop
- Cout modéré :il optimise les coûts via une meilleure utilisation des ressources présentées.
- Vitesse : Le système de fichiers distribué, les traitements concurrents et le modèle MapReduce permettent d’exécuter les requêtes les plus complexes en quelques secondes.
- Souple : car il répond à la caractéristique de variété des données en étant capable de traiter différents types de données.

4.2.2 Architecture

Hadoop se compose principalement de

- Hadoop Distributed File System (HDFS) :Un système de fichiers distribués qui fournit

- un accès haut-débit aux données de l'application.
- Hadoop YARN : Un Framework pour la planification des tâches et la gestion des ressources du cluster.
- Hadoop MapReduce : Un système basé sur YARN pour le traitement parallèle des gros volumes de données.
- Hadoop Common : Les utilitaires communs qui supportent les autres modules d'Hadoop.

4.2.3 HDFS

Le HDFS est le système de fichiers distribué d'Hadoop. C'est une implémentation du GFS, le système de fichiers distribué créé par Google. Il a été conçu et installé sur un cluster Hadoop afin de gérer le stockage distribué d'un grand volume de données sur les nœuds du cluster de façon complètement tolérante aux pannes.

NameNode, DataNode et Secondary NameNode sont les trois types de nœud que HDFS définit. Ces types sont caractérisés par :

- NameNode :
 - Responsable de la distribution et de la réplication des blocs.
 - Serveur d'informations du Hdfs pour le client Hdfs.
 - Stocke et gère les métadonnées.
 - Comporte la liste des blocs pour chaque fichier (dans le cas de lecture).
 - Contient la liste des DataNodes pour chaque bloc (dans le cas de l'écriture).
 - Logs toute métadonnée et toute transaction sur un support persistant.
 - Lectures/écritures et Créations/suppressions.
 - Démarre à partir d'une image d'HDFS (fsimage).
- DataNode :
 - Stocke des blocs de données dans le système de fichier local .
 - Maintenir des métadonnées sur les blocs possédés (ex : CRC).
 - Serveur de bloc de données et de métadonnées pour le client hdfs.
- Secondary NameNode :
 - Télécharger régulièrement les logs sur le NameNode.
 - Crée une nouvelle image en fusionnant les logs avec l'image HDFS.
 - Renvoie la nouvelle image au NameNode.

4.2.4 MapReduce

MapReduce est un modèle de programmation parallèle disponible dans les environnements Hadoop, proposé par Google. Ce modèle fournit un cadre à un développeur afin de simplifier la vie à ce dernier, d'écrire une fonction Map et une fonction Reduce et de faciliter les traitements concurrents en divisant les péta-octets de données en volumes plus petits et en les traitant en parallèle sur des serveurs standard dédiés à Hadoop. Le programme MapReduce peut se résumer à deux fonctions Map () et Reduce () [11] :

- MAP : subdivise le problème par le nœud principal, en sous-problèmes qui sont soumis

aux différents noeuds qui peuvent à leur tour le subdiviser davantage, c'est à dire découper les données d'entrée en plusieurs fragments, et faire exécuter l'opération MAP à chaque machine du cluster sur un fragment distinct ?

- REDUCE : Les noeuds remontent le résultat du traitement aux noeuds les ayant sollicités. Ces derniers remontent le résultat construit à leur tour, et ainsi de suite jusqu'au noeud principal.

4.2.5 PIG et HIVE

Bien que Pig et HIVE soient dotés de fonctionnalités similaires, ils peuvent être plus ou moins plus efficaces dans différents scénarios.

- PIG : est utile dans la phase de préparation des données, car il peut exécuter très facilement des jointures et requêtes complexes. Par ailleurs, il fonctionne très bien avec différents formats de données, y compris semi-structurées et non structurées.
- HIVE : fonctionne bien avec les données structurées, et il est donc plus efficace pour les opérations de data warehousing. Hive est utilisé côté serveur du cluster.

Les chercheurs et programmeurs ont tendance à utiliser Pig côté client du cluster, alors que les utilisateurs de Business Intelligence (par exemple, les analystes de données) constatent généralement que Hive est mieux adapté.

4.3 Spark

Apache Spark est une infrastructure de traitement de données dédiée au Big Data, il est sollicité pour sa rapidité de traiter les grands ensembles de données de manière totalement distribuée, pour sa simplicité d'usage et sa polyvalence.

Ces particularités offrent une puissance de calcul massive pour permettre le traitement de grands jeux de données en temps réel. Spark prend en charge une grande partie de la programmation grâce à son API facile à utiliser.

Spark est facile à mettre en oeuvre, il permet de développer des applications en Java, Scala, Python et R et prend en charge SQL. Depuis de nombreuses années, Hadoop est considéré comme la principale technologie de traitement de données Big Data, c'est en effet une solution très efficace pour le traitement de larges ensembles de données. Pour les calculs, MapReduce est effectivement très performant mais se trouve moins pratique pour les cas d'usage nécessitant des calculs multi-pass et des algorithmes mais dans la pratique cette approche est très lente, en plus, Hadoop inclue en général des clusters difficile à configurer et à manipuler. Par contre Apache Spark permet aux programmeurs de développer d'énormes données complexes en utilisant des patterns DAG, il permet aussi d'effectuer différentes tâches avec les mêmes données et de déployer des applications sur un cluster SIMR, YARN ou Apache Mesos [9].

4.3.1 Resilient Distributed Dataset

Le concept Resilient Distributed Dataset se trouve au cœur de Spark, c'est une abstraction de programmation qui représente une collection d'objets qui peuvent être divisés sur un cluster informatique.

Les RDD peuvent être créés à partir de n'importe quel fichier de base, une grande partie de l'API Spark core est basée sur ce concept, permettant une cartographie traditionnelle et réduisant les fonctionnalités, mais en fournissant également une prise en charge intégrée pour la jonction des ensembles de données.

L'exécution de Spark se fait de manière distribuée en combinant un processus principal de pilote qui divise une application Spark en tâches et les répartit entre de nombreux processus exécutants qui font le travail. Ces derniers sont classés selon les besoins de l'application.

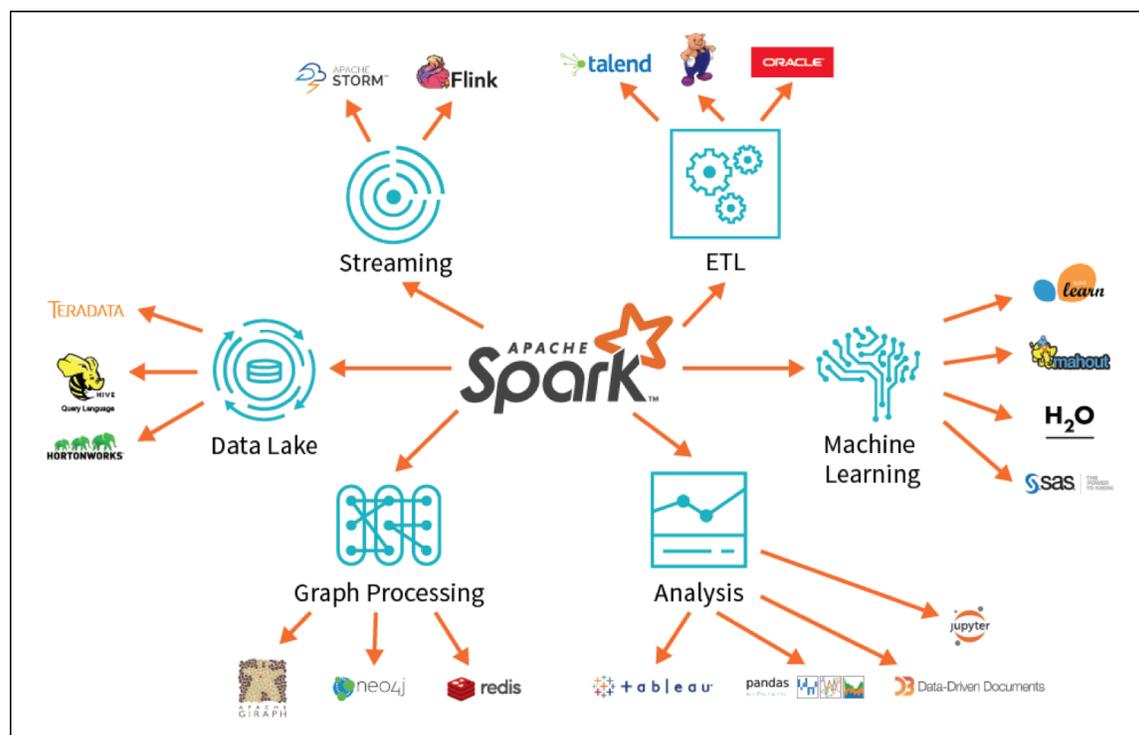


FIGURE 1.8 – Apache Spark.

4.4 Kafka

Kafka est une plate-forme de messagerie et de streaming distribuée, c'est un logiciel open source, il a été créé, à la base, comme une file d'attente pour les messages destinés à la plateforme LinkedIn et vu l'empeteur du développement qu'il a apporté, il a été transformé en une puissante plate-forme de streaming qui effectue des fonctionnalités

- Optimiser et traiter les flux de données échangés entre le destinataire et la source et les enregistrer de façon continue.
- Stocke les flux de données durablement et tolère les pannes en proposant des solutions pour les éventuels problèmes
- Propose une file d'attente de messages et assure la bonne réception de ces derniers
- obtient des données de manière totalement fiable entre les systèmes ou les applications en construisant des pipelines de données en temps réel.

4.4.1 Architecture et concepts de Kafka

Apache Kafka est exécuté en tant que cluster ou grappe de serveurs, les différents nœuds qu'on y trouve (Broker) stockent les flux de données dans des rubriques ou topics, chaque donnée se compose d'une clé, d'une valeur et d'un horodatage ce qui permet d'avoir une très bonne disponibilité et un rapide accès en lecture. Les applications qui éditent les données dans un cluster Kafka sont des producteurs (producers) et celles qui lisent les données d'un cluster sont des consommateurs (consumers), ces derniers ont accès à la bibliothèque Java, appelée Kafka Stream, lors du traitement des flux de données.

4.4.2 les API's de Kafka

Les API, Application Programming Interface, sont des ensembles de protocoles qui facilitent l'intégration de logiciels d'applications.

- L'API Producer : permet aux applications d'envoyer des flux de données sur un cluster Kafka pour les traiter dans les topics
- L'API consumer : donne aux consommateurs un accès lecture aux données des topics.
- L'API Stream : permet à une application de fonctionner comme processeur de flux, consommant un flux d'entrée à partir d'un ou plusieurs topics et produisant un flux de sortie à un ou plusieurs topics de sortie.
- L'API Connector : permet de mettre en place des consumers et des producers qui connectent des topics aux applications ou aux systèmes de données existants
- L'API Admin : permet d'administrer et d'inspecter facilement tout les objets Kafka.

Zookeeper Zookeeper est un Logiciel utilisé pour maintenir les informations de configuration, la dénomination et pour fournir une synchronisation flexible, distribuée. C'est un composant essentiel de Kafka, il agit comme un système centralisé et il garde une trace de l'état des nœuds, des sujets et des partitions de Kafka. Zookeeper permet à plusieurs clients d'effectuer simultanément des lectures et des écritures et agit comme un service de configuration partagé au sein du système.

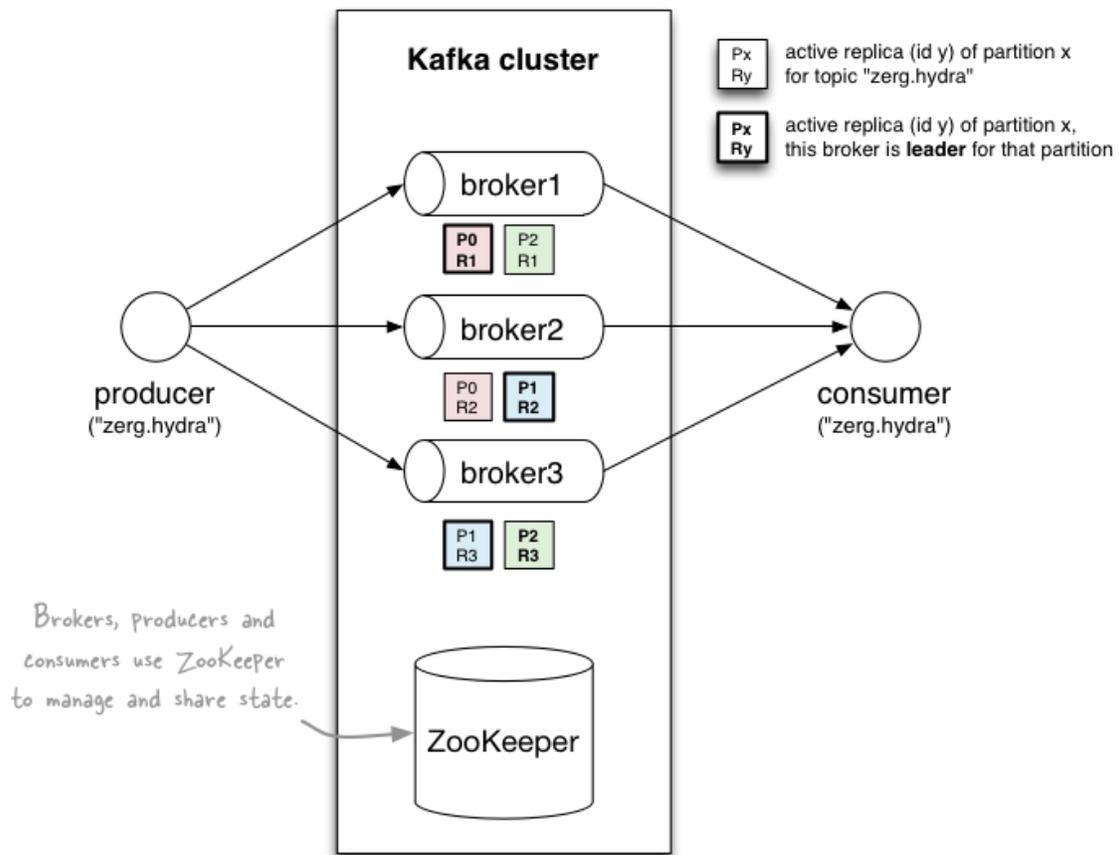


FIGURE 1.9 – Architecture d'Apache Kafka.

4.5 ELK Stack(Elasticsearch, Logstash, Kibana)

ELK Stack est l'ensemble de trois logiciels open source, à savoir : Elasticsearch, Logstash et Kibana. Ils sont tous développés, gérés et maintenus par la société Elastic. Elk est conçu pour permettre aux utilisateurs d'exploiter des données de n'importe quelle source, dans n'importe quel format et d'étudier ces données en temps réel.

4.5.1 Elasticsearch

C'est un moteur distribué de stockage de recherche et d'analyse de contenu, développé en Java, il stocke les données en format JSON, il n'a donc pas besoin de support de stockage. Il est basé sur un moteur de recherche Lucene, il possède une plateforme d'exploration et de visualisation de données pour effectuer les analyses sur les données qu'il stocke de manière centralisée. Il offre

un déploiement simple, une fiabilité maximale et une gestion facile et permet l'exécution d'une recherche rapide des documents en temps quasi réel. Le fait qu'il utilise le JSON pour le stockage rend l'exploitation des données possible à partir de n'importe quel langage de programmation. Toute interaction d'Elasticsearch se fait à travers son API REST permettant d'envoyer des requêtes HTTP. Elasticsearch est un moteur NoSQL orienté document au même titre que MangoDB ou RavenDB et il fournit toutes les fonctionnalités de stockage distribué que ce type de moteur offre [1].

Les concepts d'Elasticsearch

- Le noeud : C'est un processus applicatif exécuté sur une machine (fait référence à une instance d'Elastic Search).
- Le cluster : C'est un ensemble d'un ou plusieurs noeuds, il fournit l'indexation collective et la distribution des requêtes de recherche à travers tous ses noeuds.
- L'index : C'est un ensemble de documents JSON.
- Le type : C'est un ensemble de documents qui partagent un ensemble de champs communs, présents dans le même index.
- L'indice : C'est le container logique des types, il introduit les notions d'index ElasticSearch et index Lucene.
- Le document : C'est un ensemble de champs ou propriétés définis de façon spécifique dans le format JSON, chaque document appartient à un type et réside dans un index. à chaque document est associé un id unique (UID).
- La partition : (Shard) est une partie de l'index, ses derniers sont horizontalement partitionnés en Shards.
- La réplique : les index et les partitions sont répliqués à travers les noeuds du cluster pour accroître la haute disponibilité des traitements en cas de panne, elles permettent aussi de paralléliser les traitements de recherche de contenu dans le cluster et cela accroît la performance d'Elasticsearch.

4.5.2 Logstash

C'est un moteur de collecte de données avec des capacités de pipelining en temps réel. Logstash peut unifier dynamiquement les données de sources différentes et normaliser les données dans n'importe quelle destination. Logstash a innové la collecte de journaux mais ses capacités ne se limitent pas à ça, il permet d'enrichir et de transformer tout type d'événement avec un large éventail de plug-ins d'entrée, de filtre et de sortie [4].

4.5.3 Kibana

C'est une plate-forme d'analyse et de visualisation qui fonctionne avec Elasticsearch. Kibana est utilisé pour l'interaction avec les données d'Elasticsearch, il fonctionne en streaming en effectuant les opérations en temps réel. Cette plate-forme offre aux utilisateurs la possibilité d'effectuer une analyse de données avancée et consulter les données dans différents graphiques, les rapports générés peuvent représenter les données sous forme de barres, de lignes, de diagrammes de dispersion ou de graphiques à secteurs. Le logiciel génère des rapports PDF à la demande ou dans les délais et fournit un tableau de bord flexible et dynamique, son installation est très facile, elle

est conçu spécialement pour les gros volumes de données et simplifie tout ce qui se rapporte à leur utilisation.

Fonctionnalités principales de Kibana :

- Le logging et l'analyse de logs
- Les indicateurs d'infrastructure et le monitoring de conteneurs
- Le monitoring des performances applicatives (APM)
- L'analyse et la visualisation de données géospatiales
- L'analyse de la sécurité
- L'analyse de données métier

On cite parmi les avantages de Kibana, l'analyse de données non structurées et le profit de la grande puissance du moteur d'Elasticsearch.

4.6 Kubernetes

Kubernetes est une plateforme open source créée par Google en 2015, permet l'automatisation du déploiement et de la gestion à l'échelle des applications conteneurisées. Kubernetes peut fonctionner avec n'importe quel type de conteneurs et se charge de la gestion de l'orchestration de ces derniers.

Kubernetes doit pouvoir s'intégrer aux services de mise en réseau, de stockage, de sécurité, de télémétrie, afin de fournir une infrastructure de conteneurs complète.

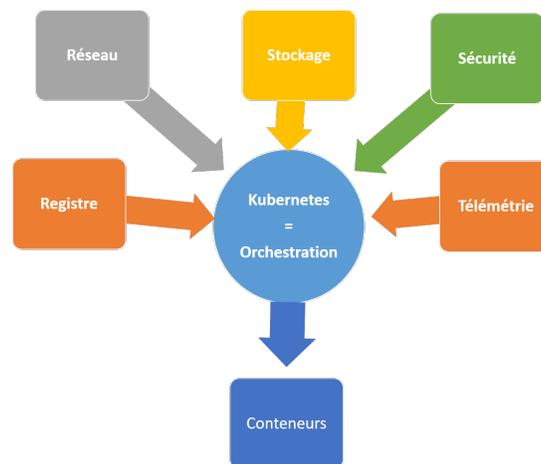


FIGURE 1.10 – les services de Kubernetes

4.7 Docker

Docker est un logiciel libre open source qui permet de créer des environnements (appelés conteneurs). C'est une plateforme de virtualisation par conteneurs qui permet de concevoir, tester et deployer facilement des applications.

C'est une technologie dotée d'un noyau Linux, pour l'utiliser sur Windows ou Mac OS X, il est nécessaire de faire appel à un logiciel de machine virtuelle.

La clé du succès de Docker est : la possibilité de décrire formellement comment construire le conteneur et d'avoir un format d'image standardisé.

Docker se base sur deux extensions du noyau Linux, à savoir :

1. Cgroups : qui permet de diviser les ressources d'une machine afin de contrôler la consommation de celles-ci par processus.
 2. Namespaces : qui permet de voiler les ressources aux processus qui n'elles utilisent pas. Il offre l'isolation nécessaire pour la création d'un conteneur.
- Un conteneur est un système de fichiers sur lequel s'exécutent des processus.
 - Les deux outils sont complémentaires, Docker gère le cycle de vie des conteneurs, les développe, les déploie et donc les itère plus vite tandis que Kubernetes gère leur orchestration et permet de les runner sans danger [7].

5 Conclusion

Dans ce chapitre, nous avons abordé la définition du Big data et ses caractéristiques en premier lieu, puis, nous avons présenté en quoi consistent le parc informatique et ses composants (parc matériel et parc logiciel).

Ensuite, dans la deuxième partie de ce chapitre, nous avons parlé des applications de gestion orientées métiers qui comportent les points essentiels sur lesquels nous nous sommes attardé pour mener à bien notre projet.

Enfin, la dernière partie, est le mouvement NoSQL au sein de l'entreprise dans laquelle nous définissons le NoSQL et les environnements et les plates-formes qui nous ont permis de réaliser ce travail.

Notre prochain chapitre se portera sur l'introduction au cas d'étude et tout ce qui est lié à l'entreprise.

Chapitre 2

Introduction du cas d'étude

1 Introduction

La clé de la réussite et de l'épanouissement de chaque entreprises est liée à son histoire, son organisation et aux différents acteurs qui la dirigent.

Dans ce chapitre nous allons faire une présentation générale sur l'entreprise ; historique, organigramme et les différents services de l'entreprise, en vue de mieux la comprendre et répondre à ces exigences en leur proposant des solutions adaptées à leur besoin afin qu' ils puisse bénéficier d'une meilleure gestion des ressources et la terminer avec les généralité sur la virtualisation et ces types ainsi que les logiciels de virtualisation.

2 Présentation de la société d'accueil Cevital

Cevital est un groupe familial de vingt-cinq sociétés, réparties dans cinq secteurs d'activités : L'Industrie Métallurgique, l'Information et la Communication, la Distribution Automobile, le Transport Terrestre et Maritime, l'Industrie Agroalimentaire. CEVITAL est parmi les entreprises qui ont vu le jour dès l'entrée du pays dans l'économie de marché. Disposant de technologies de pointe.

Cevital possède deux raffineries : une d'huile et l'autre de sucre. La raffinerie d'huile alimentaire a été mise en chantier en Mai 1998, en Aout 1999 elle est rentré en production, plus tard en 2000, la raffinerie du sucre est mise en chantier, elle n'est devenue fonctionnelle qu'en 2002.

Un autre produit est mis en chantier en 2000 et en production en 2001, c'est la margarine. Une deuxième raffinerie de sucre de 3000 T, de plus le silo sucre blanc 80000 T et le silo sucre roux 150000 T, une unité d'eau minéral Lalla Khadîdja, et une autre unité de Cojek a El Kseur. Enfin, une station de cogénération.

2.1 Organigramme du groupe Cevital

Voici le schéma général du groupe Cevital, dont chaque direction a pour but d'assurer le bon fonctionnement de chaque partie du groupe comme le montre cette figure :

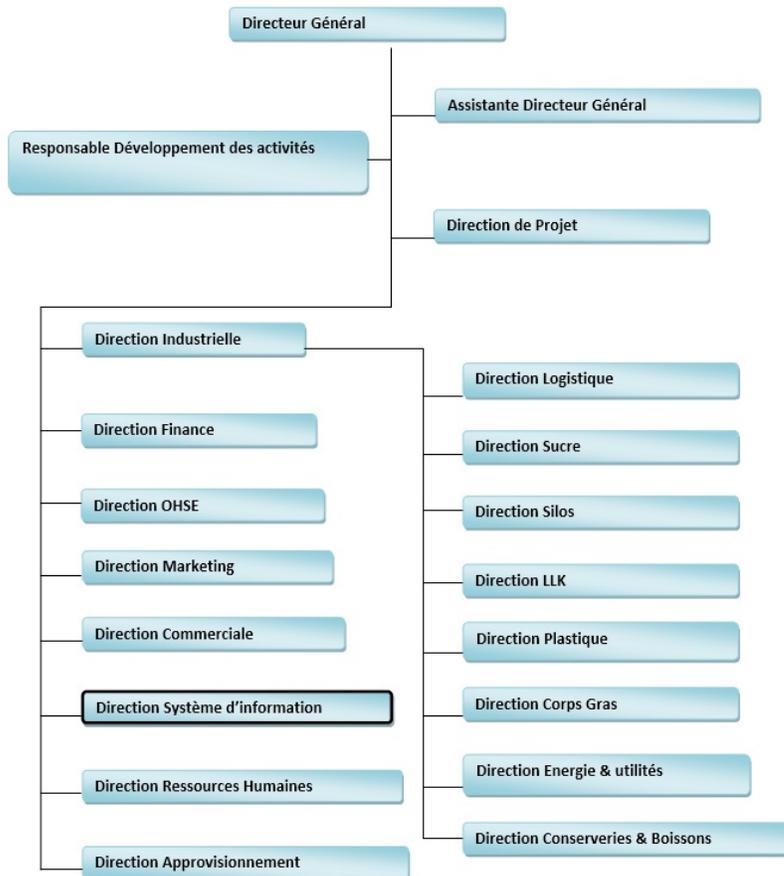


FIGURE 2.1 – organigramme du groupe Cevital

2.1.1 Les missions

L'entreprise a pour missions principales de développer la production et d'assurer la qualité du conditionnement des huiles, des margarines et du sucre à des prix nettement plus compétitifs et cela dans le but de satisfaire le client et de le fidéliser.

2.1.2 Les activités

Lancé en Mai 1998, le complexe Cevital a débuté son activité par le conditionnement en Décembre 1998, en Février 1991, les travaux de génie civil de la raffinerie ont débuté. Cette dernière est devenue fonctionnelle, en Août 1999. L'ensemble des activités de Cevital est concentré sur la production et la commercialisation des huiles végétales, de margarine et de sucre se présente comme suite :

- Raffinage d'huile 1600 T/J pouvant passer après extension à 1800 T/J.
- Production de margarine de capacité 600 T/J.
- Fabrication d'emballage en PET (9600 unités/h).
- Stockage céréales.
- Electrolyseur (par mesure de sécurité doit être déplacé hors Cevital).
- Extension de la sucrerie.
- Savonnerie.
- Minoterie.
- Hydroélectrique d'huile.

2.1.3 Les objectifs

Les objectifs visés par Cevital peuvent se présenter comme suit :

- Encouragement des agricultures par des aides financières pour la production locale de graines oléagineuses.
- Importation de graines oléagineuses pour l'extraction directe des huiles brutes.
- Diversification de ses produits et sa diffusion sur tout le territoire national.
- Modernisation de ses installations et adoption de nouvelles démarches de gestion technique afin d'augmenter le volume de sa production.
- Positionner ses produits sur le marché étranger par leurs exportations.
- Optimisation de ses offres d'emploi sur le marché du travail.

2.2 L'informatique dans Cevital

Cevital est parmi les entreprises possédant une direction informatique et donne une grande importance au domaine de l'informatique.

2.3 Présentation de l'organisme d'accueil

Notre étude se focalise au niveau du groupe Cevital de Béjaïa où nous avons effectué notre stage, dans la direction de l'informatique réseau et télécom.

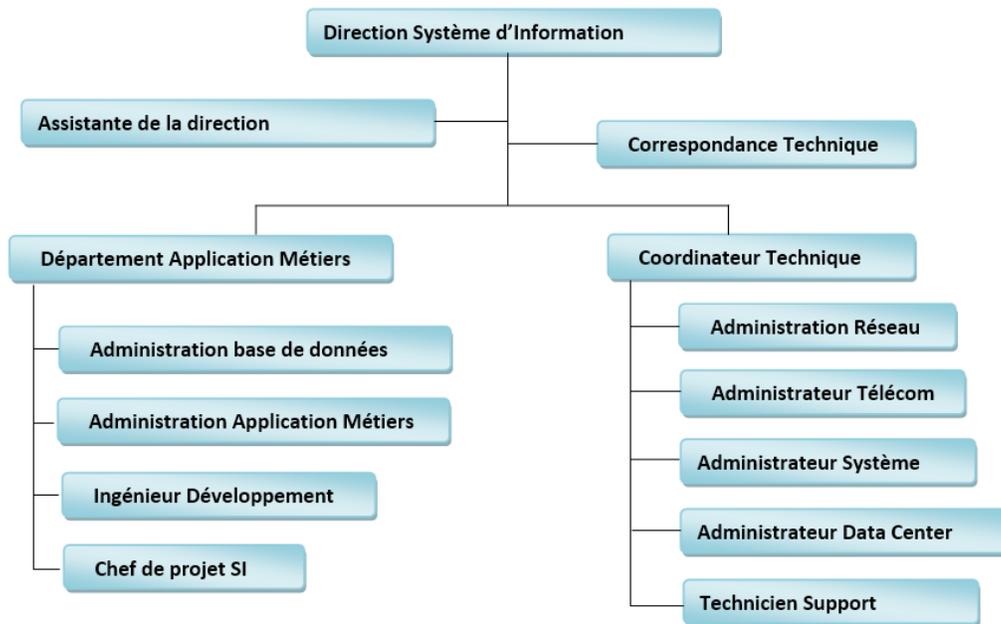


FIGURE 2.2 – organigramme de la direction Système d'information

3 Etat des lieux

3.1 Présentation de l'infrastructure réseau

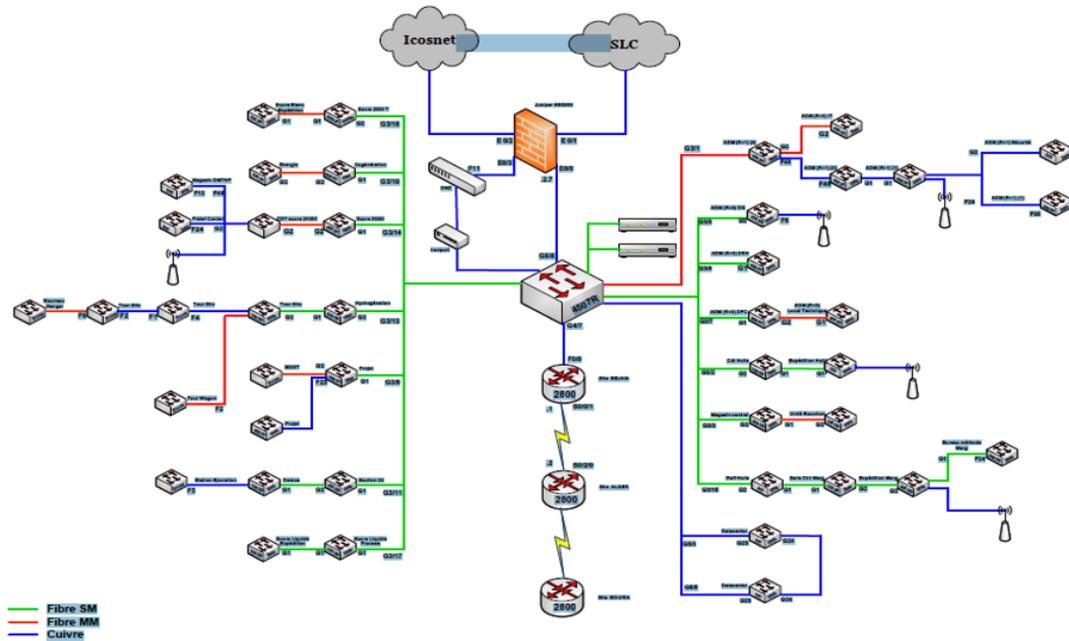


FIGURE 2.3 – Infrastructure réseau de Cevital.

3.2 Présentation du GMAO de Cevital

Cevital a opté pour l'utilisation de Coswin 8i comme solution pour la GMAO (gestion de la Maintenance Assistée par Ordinateur) permettant d'optimiser la gestion et le pilotage de la maintenance ainsi que la performance des équipements. Cette solution aide à réduire les coûts d'achats et les volumes des stocks, à améliorer la productivité du personnel de maintenance tout en restant conforme aux réglementations en vigueur.

Coswin 8i est intégralement architecturé Web et exploite tous les avantages offerts par les dernières technologies (JSF, HTML5, Ajax). Les utilisateurs peuvent y accéder via un navigateur Internet standard de n'importe où et à tout moment, ce qui permet de minimiser le trafic réseau et d'éviter l'achat de matériel informatique coûteux.

Coswin 8i peut ainsi être implémenté dans des configurations très différentes, de la plus simple à la plus complexe, en réseau local ou étendu. Coswin 8i est très intuitif, ce qui permet une prise en main rapide et une exploitation efficace. En effet, le logiciel bénéficie d'une ergonomie innovante, qui allie performance, design, convivialité et simplicité d'utilisation afin d'apporter un maximum de confort à l'utilisateur, parmi ses avantages, la facilité d'utilisation de son interface.

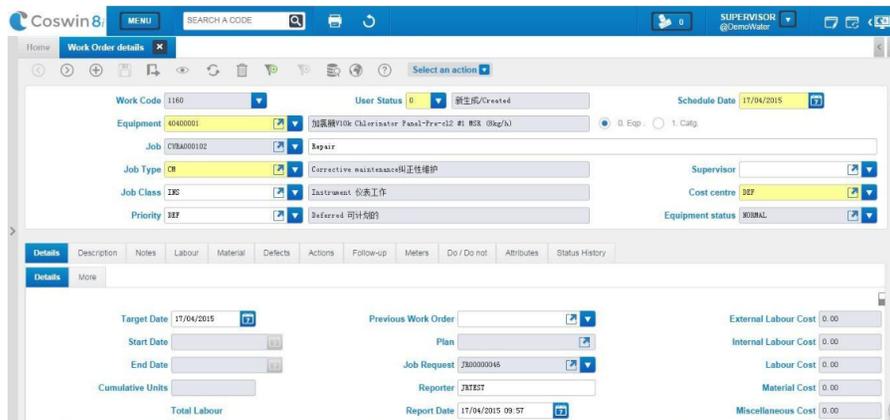


FIGURE 2.4 – Interface de Coswin 8i.

Coswin 8i est doté de trois nouveaux modules qui incluent les nouvelles innovations de l'in-

dustrie.

- Coswin IoT permet d'analyser les données des équipements connectés : recueillir des informations en temps réel, analyser les performances, prédire les opérations de maintenance à venir.
- Coswin BIM étend les capacités de l'outil à centraliser les données de vos actifs en y ajoutant les maquettes numériques.
- Coswin SIG vous permet d'accéder rapidement aux données de vos équipements et les géolocaliser.

De plus, Coswin 8i est entièrement personnalisable par profil utilisateur (vocabulaire, mise en forme des écrans, ajout de nouveaux champs, définition de liens avec des applications externes, navigation en mode contextuel).

3.3 Présentation de l'ERP Sage

Cevital utilise l'ERP Sage, c'est une solution logiciel de planification et de contrôle des ressources de l'entreprise. ERP veut dire : Entreprise Ressources Planing, ce système permet d'accéder aux données de l'entreprise pertinentes depuis un emplacement centralisé

L'ERP Sage est utilisé dans cette société car il aide à développer et faire croître l'entreprise en rassemblant toutes les données liées aux tâches quotidiennes dans une seule et unique base de donnée.

Modules de Sage

- La logistique inclut la gestion des nomenclatures, des plans de production, l'analyse des coûts de fabrication et la gestion des stocks.
- La comptabilité, qui inclut la comptabilité générale et analytique, la gestion de trésorerie et la communication bancaire.
- les ressources humaines
- Le marketing, CRM et ventes.

3.4 Présentation des processus Métiers gérant la production d'énergie

3.4.1 L'unité d'énergie et utilités

L'unité d'énergie sert essentiellement à produire de l'électricité à l'aide d'un système d'installation qui se compose d'une chaudière, dont le rôle de produire de la vapeur sèche qui va se transformer en électricité en fonction d'autres éléments.

La direction d'énergie est constituée de deux départements qui sont :

- Département électricité (production et distribution de l'énergie électrique) :
 - Le poste 60KV
 - poste 30KV
 - La cogénération
- Département chaufferie (production et distribution de la vapeur)

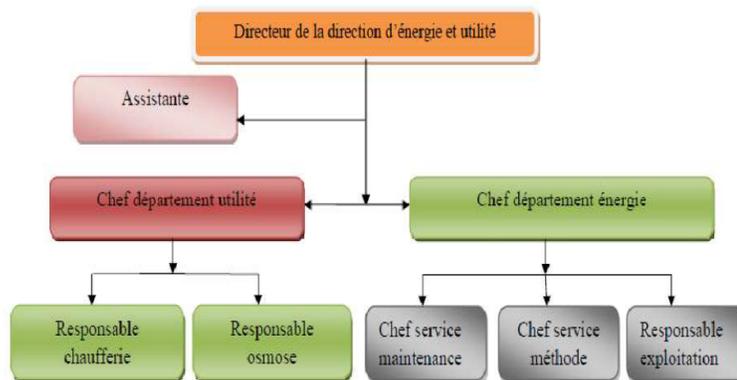


FIGURE 2.5 – Organigramme du pôle d'énergie

3.5 Présentation du TMS de Numilog

Transwide est un TMS transport (Transport Management System) conçu par une société belge à l'attention de nombreuses entreprises disposant d'une chaîne logistique complexe comme Cevital qui l'utilise d'ailleurs depuis 3ans . Cette solution permet de faciliter la gestion du transport, l'affectation des fournisseurs, d'offrir une traçabilité totale des livraisons et d'optimiser leur schéma. Les chargeurs, transporteurs et commissionnaires de transport sont les premiers concernés par l'utilisation de ce type de solution. Elle s'adresse donc aux entreprises dont le cœur de métier est d'assurer le transport et la logistique pour le compte de leurs clients. Dans l'ensemble, ce logiciel permet de réduire les coûts de transport. En vue d'une meilleur traçabilité Cevitale utilise la solution Skipper.



FIGURE 2.6 – Transport Mangement System

Fonctionnalités de Transwide

- planification et optimisation des transports
- intégration de données dans le système ERP
- recherche de transporteurs.
- gestion des contrats et des tarifs.
- demandes de cotations 'spot.
- sélection intelligente des transporteurs.
- calcul du kilométrage et de l'horaire d'arrivée prévue.
- communication des ordres de transport
- aide à la gestion de transport maritime par conteneurs.
- réservation de créneaux horaires pour les chargements/déchargements des transporteurs.

En raison de quelques insuffisances et problèmes technique tel que la saisie de données manuellement au niveau de poste de sécurité (plein d'erreurs intentionnées et malintentionnées), la faible flexibilité et le cout qui est cher (10 mille £ par mois), Cevital viens de mettre en place un nouveau TMS qui est en cours de configuration afin d'assurer le bon fonctionnement et une meilleure gestion de son TMS.

4 Conclusion

Le premier point de ce chapitre est porté sur le présentation de la société d'accueil "Cevital", ce dernier comporte l'organigramme de l'entreprise, l'informatique dans Cevital et la présentation de l'organisme d'accueil.

Ensuite, en deuxième position vient l'état des lieux, à partir de là nous avons présenter l'infrastructure réseau de l'entreprise, puis les applications de gestion orientées métier utilisé dans la société de Cevital et les processus métier gérant les ordres de transferts.

Le chapitre trois, intitulé : "mise en place de la solution" aura pour contenu des captures d'écrans qui schématisent toutes les installations et les configurations des application et logiciels que nous avons utilisé.

Chapitre 3

Mise en place de la solution

1 Introduction

Dans ce chapitre, nous avons intégré des captures d'écrans des installations des logiciels et plates-formes que nous avons utilisé dans notre projet.

Nous avons commencé par IntelliJ IDEA, l'environnement de développement que nous avons choisi, pour y installer Scala qui est le langage de programmation que nous avons jugé le plus adéquat pour notre cas. Ensuite, nous avons installé le framework Apache Spark et effectué les configurations nécessaires pour son fonctionnement. Enfin, pour l'implémentation et l'exploitation de la solution nous avons installé Elasticsearch et Kibana.

2 Infrastructure

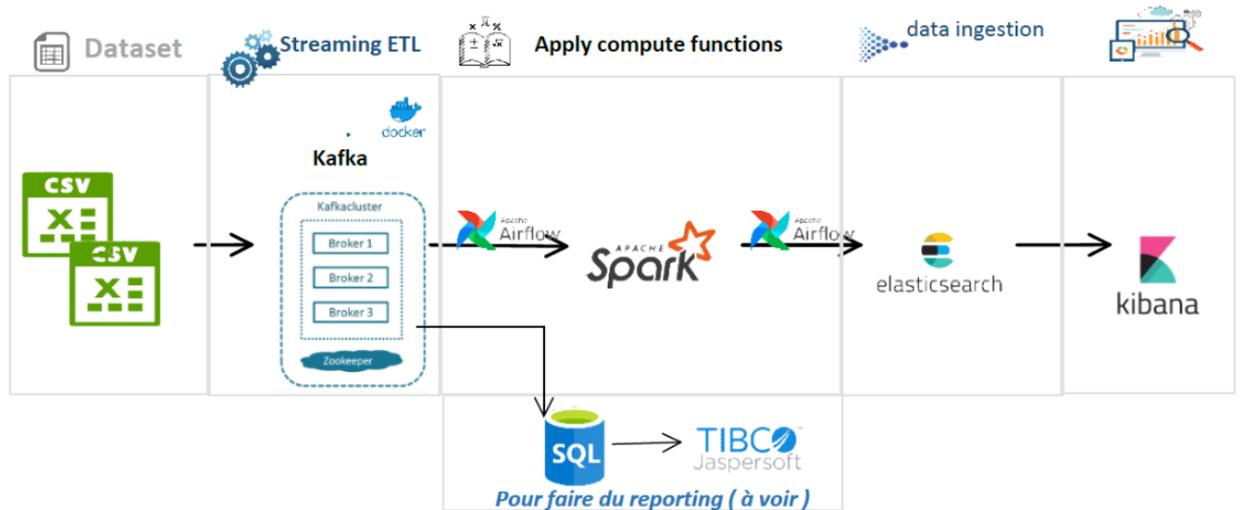


FIGURE 3.1 – Infrastructure

Cette infrastructure présente les différentes couches par lesquelles nous sommes passé pour réaliser notre projet.

En premier lieu, nous avons extrait les données en format CSV, ensuite nous avons décidé d'utiliser Apache Kafka pour programmer sous Scala mais finalement nous avons opté pour IntelliJ IDEA comme environnement de développement, dans la troisième étape nous avons choisi d'utiliser Spark comme framework. Les deux dernières couches consistent respectivement à récupérer les données et les stocker via Elasticsearch et enfin créer une visualisation à l'aide de Kibana.

3 Installation de Fedora sur VM Ware

Nous avons installé Fedora sur une machine virtuelle pour faire notre projet dessus afin que celui ci soit plus universel et pour qu'il soit à la portée de n'importe quelle entreprise.

4 Installation de IntelliJ IDEA

Après avoir visité la page d'intelliJ IDEA, télécharger le programme et effectuer les configurations nécessaires (emplacement, association de fichier...), Nous avons coché la case pour démarrer le programme ensuite cliqué sur finish pour finir l'installation.

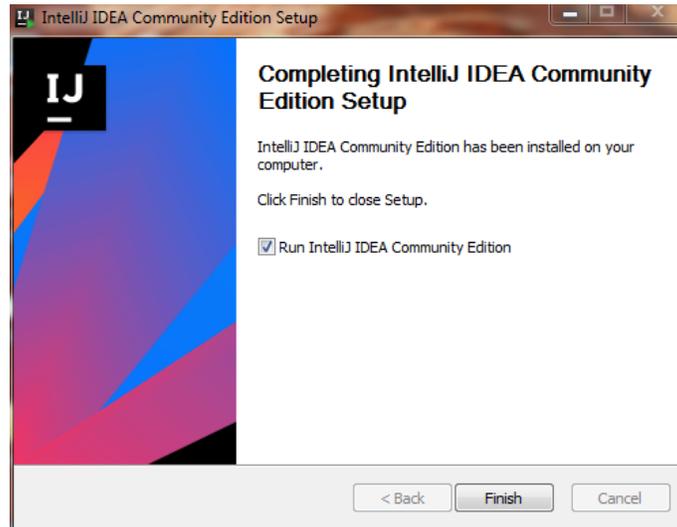


FIGURE 3.4 – fin d'installation d'IntelliJ



FIGURE 3.5 – démarrage d'IntelliJ IDEA

Nous vérifions que IntelliJ IDEA est bien installé et fonctionnel.

Mise en place de la solution

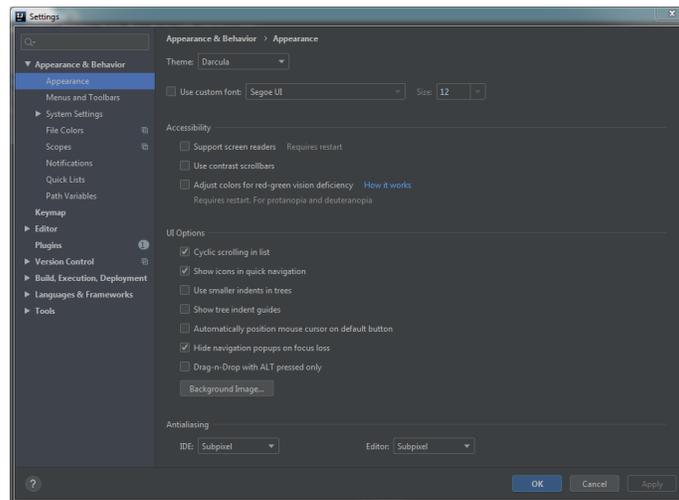


FIGURE 3.6 – configuration IntelliJ

Cette image nous permet de voir les configurations qui ont été apportées au programme IntelliJ IDEA après l'installation.

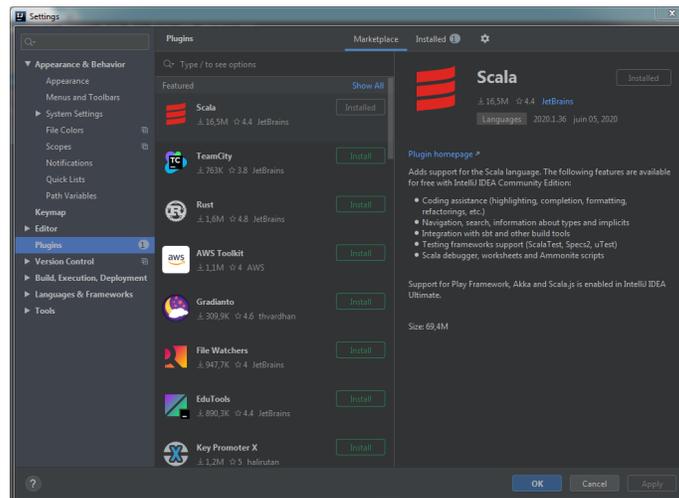
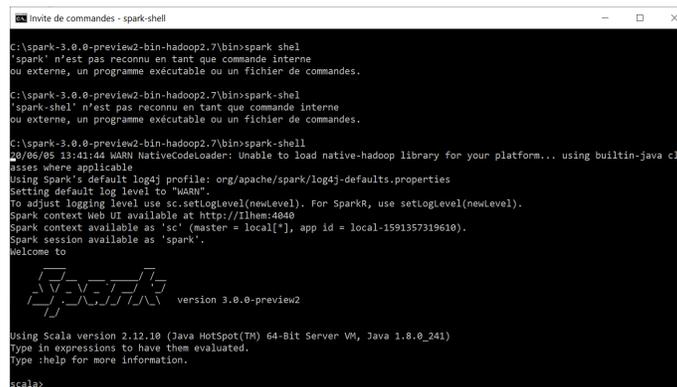


FIGURE 3.7 – Installation de Scala

Nous allons développer en Scala. Pour ça, nous téléchargeons le plugin correspondant pour IntelliJ.

5 Installation de SPARK

Pour installer Spark et arriver à cette étape, nous avons d'abord dû installer Java. Ce n'est qu'après que nous avons installé Spark et obtenu ce résultat.



```
C:\spark-3.0.0-preview2-bin-hadoop2.7\bin>spark shell
'spark' n'est pas reconnu en tant que commande interne
ou externe, un programme exécutable ou un fichier de commandes.

C:\spark-3.0.0-preview2-bin-hadoop2.7\bin>spark-shell
'spark-shell' n'est pas reconnu en tant que commande interne
ou externe, un programme exécutable ou un fichier de commandes.

C:\spark-3.0.0-preview2-bin-hadoop2.7\bin>spark-shell
20/06/05 13:41:44 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context web UI available at http://11hem4080
Spark context available as 'sc' (master = local[*], app id = local-1591357319610).
Spark session available as 'spark'.
Welcome to

 version 3.0.0-preview2

Using Scala version 2.12.10 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_241)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

FIGURE 3.8 – Installation de SPARK

Ceci montre que Spark a correctement été installé.

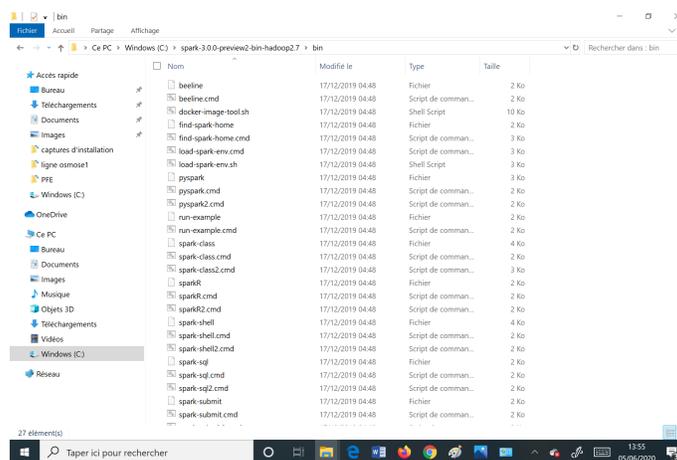


FIGURE 3.9 – emplacement SPARK

Nous avons ajouté Spark comme variable d'environnement et le chemin de celui-ci.

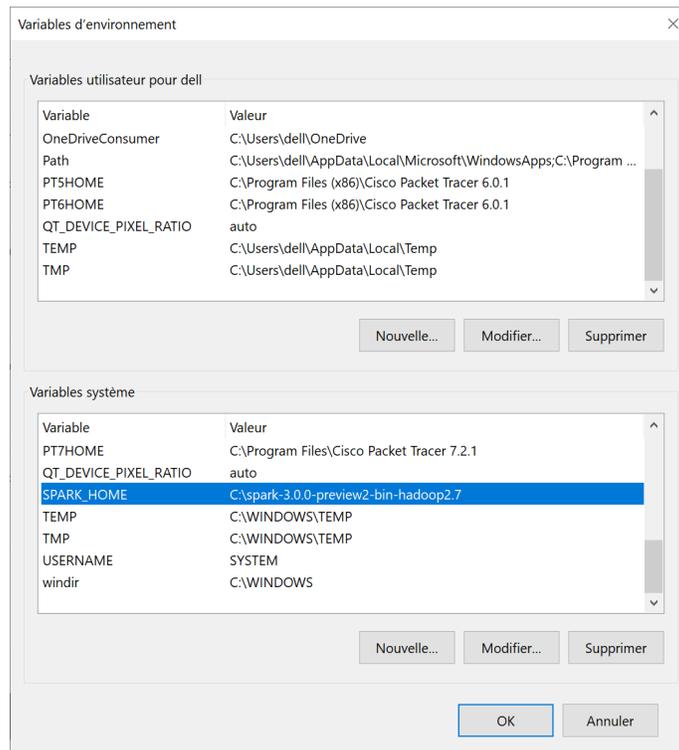


FIGURE 3.10 – configuration SPARK

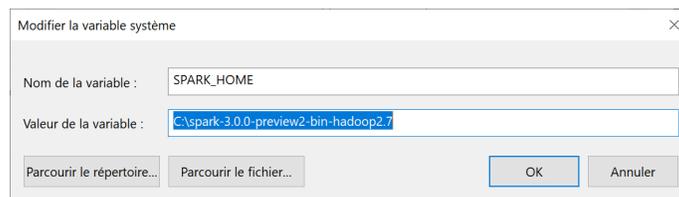


FIGURE 3.11 – Variable d'environnement SPARK

6 Installation d'Elasticsearch

Ceci est le résultat de l'installation d'Elasticsearch. Après avoir téléchargé le logiciel à partir du site, nous attendons la fin de l'installation et dans l'invite de commande, nous accédons au dossier BIN d'Elasticsearch pour le démarrage et les deux images ci-dessous indique que l'installation a été faite avec succès.

Mise en place de la solution

```
Invite de commandes - elasticsearch
Microsoft Windows [version 10.0.18362.778]
(c) 2019 Microsoft Corporation. Tous droits réservés.

C:\Users\dell>cd\

C:\>cd C:\elasticsearch-7.6.2\bin

C:\elasticsearch-7.6.2\bin>elasticsearch
future versions of Elasticsearch will require Java 11; your Java version from [C:\Program Files\Java\jdk1.8.0_241\jre] does not meet this requirement
Warning: with JDK 8 on Windows, Elasticsearch may be unable to derive correct ergonomic settings due to a JDK issue (JDK-8074459). Please use a newer version of Java.
Warning: MaxDirectMemorySize may have been miscalculated due to JDK-8074459. Please use a newer version of Java or set MaxDirectMemorySize explicitly.
[2020-06-05T14:22:55,433][INFO ][o.e.e.NodeEnvironment ] [ILHEM] using [1] data paths, mounts [{"Windows (C:)}], net usable_space [440.4mb], net total_space [117.4gb], types [NTFS]
[2020-06-05T14:22:55,464][INFO ][o.e.e.NodeEnvironment ] [ILHEM] heap size [990.7mb], compressed ordinary object pointers [true]
[2020-06-05T14:22:55,945][INFO ][o.e.n.Node ] [ILHEM] node name [ILHEM], node ID [VmF8CCHBQwuagfi9d8wQIQ], cluster name [elasticsearch]
[2020-06-05T14:22:55,945][INFO ][o.e.n.Node ] [ILHEM] version[7.6.2], pid[13116], build[default/zip/ef48eb35cf30adf4db14086e8aabd07ef6fb113f/2020-03-26T06:34:37.794943Z], OS[Windows 10/10.0/amd64], JVM[Oracle Corporation/Java HotSpot(TM) 64-Bit Server VM/1.8.0_241/25.241-b07]
[2020-06-05T14:22:55,945][INFO ][o.e.n.Node ] [ILHEM] JVM home [C:\Program Files\Java\jdk1.8.0_241\jre]
[2020-06-05T14:22:55,945][INFO ][o.e.n.Node ] [ILHEM] JVM arguments [-Des.networkaddress.cache.ttl=60, -Des.networkaddress.cache.negative.ttl=10, -XX:+AlwaysPreTouch, -Xss1m, -Djava.awt.headless=true, -Dfile.encoding=UTF-8, -Djna.nosys=true, -XX:-OmitStackTraceInFastThrow, -Dio.netty.noUnsafe=true, -Dio.netty.noKeySetOptimization=true, -Dio.netty.recycler.maxCapacityPerThread=0, -Dio.netty allocator.numDirectArenas=0, -Dlog4j.shutdownHookEnabled=false, -Dlog4j2.disable.jmx=true, -Djava.locale.providers=COMPAT, -Xms1g, -Xmx1g, -XX:+UseConcMarkSweepGC, -XX:CMSInitiatingOccupancyFr...
```

FIGURE 3.12 – Installation d'Elasticsearch

Elasticsearch est en cours d'installation.

```
Invite de commandes - elasticsearch
[settings]
[2020-06-05T14:23:22,788][INFO ][o.e.n.Node ] [ILHEM] initialized
[2020-06-05T14:23:22,792][INFO ][o.e.n.Node ] [ILHEM] starting ...
[2020-06-05T14:23:28,804][INFO ][o.e.t.TransportService ] [ILHEM] publish_address {127.0.0.1:9300}, bound_addresses {127.0.0.1:9300}, {[::1]:9300}
[2020-06-05T14:23:29,320][WARN ][o.e.b.BootstrapChecks ] [ILHEM] the default discovery settings are unsuitable for production use; at least one of [discovery.seed_hosts, discovery.seed_providers, cluster.initial_master_nodes] must be configured
[2020-06-05T14:23:29,336][INFO ][o.e.c.c.Coordinator ] [ILHEM] cluster UUID [FsaaFEaDQgzzGw-K6o38Q]
[2020-06-05T14:23:29,358][INFO ][o.e.c.c.ClusterBootstrapService] [ILHEM] no discovery configuration found, will perform best-effort cluster bootstrapping after [3s] unless existing master is discovered
[2020-06-05T14:23:29,505][INFO ][o.e.c.s.MasterService ] [ILHEM] elected-as-master ([1] nodes joined){ILHEM}{VmF8CCHBQwuagfi9d8wQIQ}{m0xQTQIRRCsJvN3Kkj7Svw}{127.0.0.1}{127.0.0.1:9300}{dilm}{ml.machine_memory=8466104320, xpack.installed=true, ml.max_open_jobs=20} elect leader, _BECOME_MASTER_TASK_, _FINISH_ELECTION_, term: 18, version: 239, delta: master node changed {previous [], current [{ILHEM}{VmF8CCHBQwuagfi9d8wQIQ}{m0xQTQIRRCsJvN3Kkj7Svw}{127.0.0.1}{127.0.0.1:9300}{dilm}{ml.machine_memory=8466104320, xpack.installed=true, ml.max_open_jobs=20}}
[2020-06-05T14:23:29,621][INFO ][o.e.c.s.ClusterApplierService] [ILHEM] master node changed {previous [], current [{ILHEM}{VmF8CCHBQwuagfi9d8wQIQ}{m0xQTQIRRCsJvN3Kkj7Svw}{127.0.0.1}{127.0.0.1:9300}{dilm}{ml.machine_memory=8466104320, xpack.installed=true, ml.max_open_jobs=20}}, term: 18, version: 239, reason: Publication{term=18, version=239}
[2020-06-05T14:23:30,219][INFO ][o.e.l.LicenseService ] [ILHEM] license [138ae6c0-3baa-4ea6-87ae-d8f3c1113764] mode [basic] - valid
[2020-06-05T14:23:30,229][INFO ][o.e.x.s.s.SecurityStatusChangeListener] [ILHEM] Active license is now [BASIC]; Security is disabled
[2020-06-05T14:23:30,249][INFO ][o.e.g.GatewayService ] [ILHEM] recovered [5] indices into cluster_state
[2020-06-05T14:23:32,214][INFO ][o.e.c.r.a.AllocationService] [ILHEM] Cluster health status changed from [RED] to [YELLOW] (reason: [shards started [[.apm-agent-configuration][0]]]).
[2020-06-05T14:23:34,180][INFO ][o.e.h.AbstractHttpServerTransport ] [ILHEM] publish_address {127.0.0.1:9200}, bound_addresses {127.0.0.1:9200}, {[::1]:9200}
[2020-06-05T14:23:34,180][INFO ][o.e.n.Node ] [ILHEM] started
```

FIGURE 3.13 – Démarrage d'Elasticsearch

à la fin de l'installation l'invite de commandes indique qu'Elasticsearch est prêt à être utilisé.

Mise en place de la solution

```
localhost:9200 x +
localhost:9200
{
  "name": "ILHEM",
  "cluster_name": "elasticsearch",
  "cluster_builder": "F5aaFE40qzz0n-K6o39Q",
  "version": {
    "number": "7.6.2",
    "build_flavor": "default",
    "build_type": "zip",
    "build_hash": "ef48eb35cf30aeafdb1a886e8ebd07ef6fb13f",
    "build_date": "2020-03-26T08:34:37.794943Z",
    "build_snapshot": false,
    "lucene_version": "8.4.0",
    "minimum_wire_compatibility_version": "6.8.0",
    "minimum_index_compatibility_version": "6.0.0beta1"
  },
  "tagline": "You Know, for Search"
}
```

FIGURE 3.14 – Interface d'Elasticsearch

Une fois le elasticsearch lancé, nous y accédant à partir du navigateur. Nous avons là les informations concernant le logiciel installé, dont le nom de l'hôte sur laquelle il a été installé, la version installée et tout ce qui ce rapporte à celle-ci et enfin le slogan d'Elasticsearch.

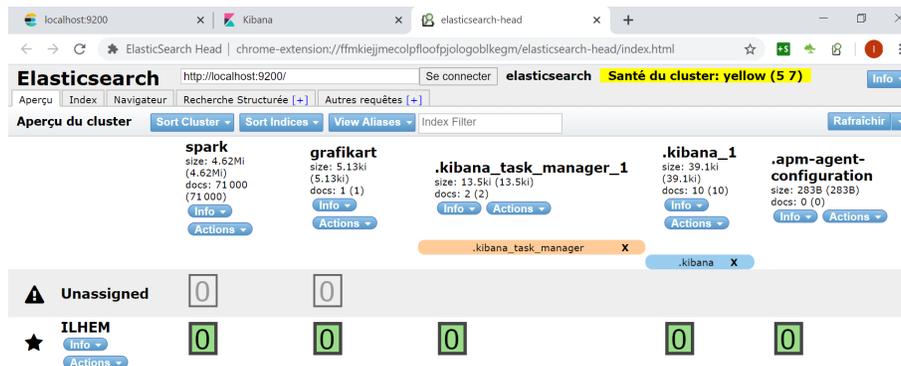
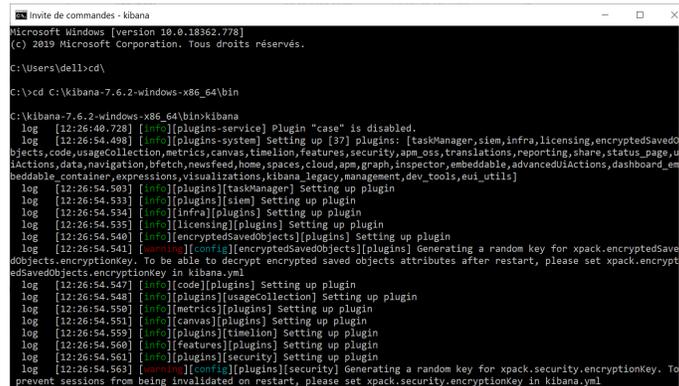


FIGURE 3.15 – En-tête Elasticsearch.

Cette page désigne l'en-tête d'Elasticsearch.

7 Installation de Kibana

Enfin, nous avons installé Kibana, un logiciel qui complète Elastic search, il nous permet de visualiser les données et de créer des dashboards des diagrammes des données que nous étudions.



```
Invite de commandes - kibana
Microsoft Windows [version 10.0.18362.778]
(c) 2019 Microsoft Corporation. Tous droits réservés.

C:\Users\dell>cd\

C:\>cd C:\kibana-7.6.2-windows-x86_64\bin

C:\kibana-7.6.2-windows-x86_64\bin>kibana
log [12:26:40.728] [info] [plugins-service] Plugin "case" is disabled.
log [12:26:54.498] [info] [plugins-system] Setting up [37] plugins: [taskManager,siem,infra,licensing,encryptedSavedObjects,code,usageCollection,metrics,canvas,timelion,features,security,apm_oss,translations,reporting,share,status_page,actions,data,navigation,fetch,newsFeed,home,spaces,cloud,apm,graph,inspector,embeddable,advancedUIActions,dashboard_embeddable,container,expressions,visualizations,kibana_legacy,management,dev_tools,eui_utils]
log [12:26:54.503] [info] [plugins][taskManager] Setting up plugin
log [12:26:54.533] [info] [plugins][siem] Setting up plugin
log [12:26:54.538] [info] [infra][plugins] Setting up plugin
log [12:26:54.535] [info] [licensing][plugins] Setting up plugin
log [12:26:54.540] [info] [encryptedSavedObjects][plugins] Setting up plugin
log [12:26:54.541] [warning] [config][encryptedSavedObjects][plugins] Generating a random key for xpack.encryptedSavedObjects.encryptedKey. To be able to decrypt encrypted saved objects attributes after restart, please set xpack.encryptedSavedObjects.encryptedKey in kibana.yml
log [12:26:54.547] [info] [code][plugins] Setting up plugin
log [12:26:54.548] [info] [plugins][usageCollection] Setting up plugin
log [12:26:54.550] [info] [metrics][plugins] Setting up plugin
log [12:26:54.551] [info] [canvas][plugins] Setting up plugin
log [12:26:54.559] [info] [plugins][timelion] Setting up plugin
log [12:26:54.560] [info] [features][plugins] Setting up plugin
log [12:26:54.561] [info] [plugins][security] Setting up plugin
log [12:26:54.563] [warning] [config][plugins][security] Generating a random key for xpack.security.encryptedKey. To prevent sessions from being invalidated on restart, please set xpack.security.encryptedKey in kibana.yml
```

FIGURE 3.16 – Démarrage de Kibana

Quand le logiciel est installé, nous y accedons de la même manière qu'Elasticsearch : à partir de l'invite de commande. Quand le programme démarre l'accès se fera sur le navigateur.

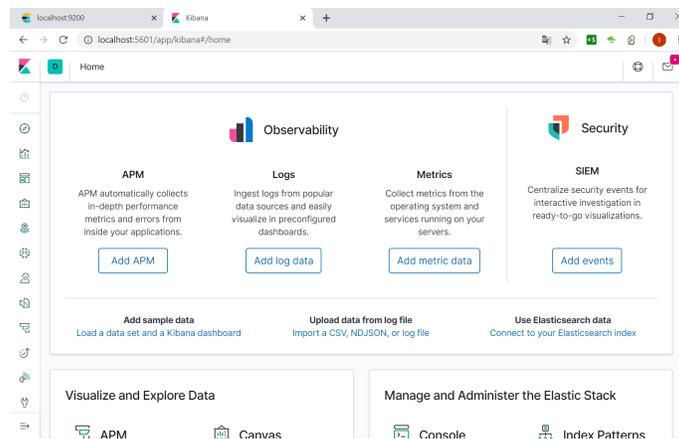


FIGURE 3.17 – Interface de Kibana

C'est l'interface principale de Kibana, que nous obtenons après la fin de son installation.

Conclusion :

Toutes les installations faites avec succès, Nous passons maintenant ,dans le prochain chapitre, à l'exploitation de toutes ces ressources pour le déploiement de la solution.

Chapitre 4

Déploiement et exploitation de la solution

1 Introduction

Ce dernier chapitre se portera sur l'exploitation des données pour le déploiement de la solution. En effet, Après tout les logiciels bien installés, Nous extrayons les données sous format de fichiers CSV, ensuite, sur IntelliJ, nous mettons en place un programme avec Scala afin de pouvoir récupérer les fichiers sur elasticsearch et enfin, réaliser une visualisation sur Kibana.

2 Extraction des données de l'UPE

Nous avons extrait les données de l'unité de production énergétique depuis le logiciel TIA PORTAL sous format de fichiers csv

3 Stockage via Spark

Une fois que nous avons extrait les fichiers csv, nous les avons stocké via le moteur d'analyse Spark qui est ultra-rapide pour le big data et le machine learning pour pouvoir les analyser et les traiter sur Elasticsearch.

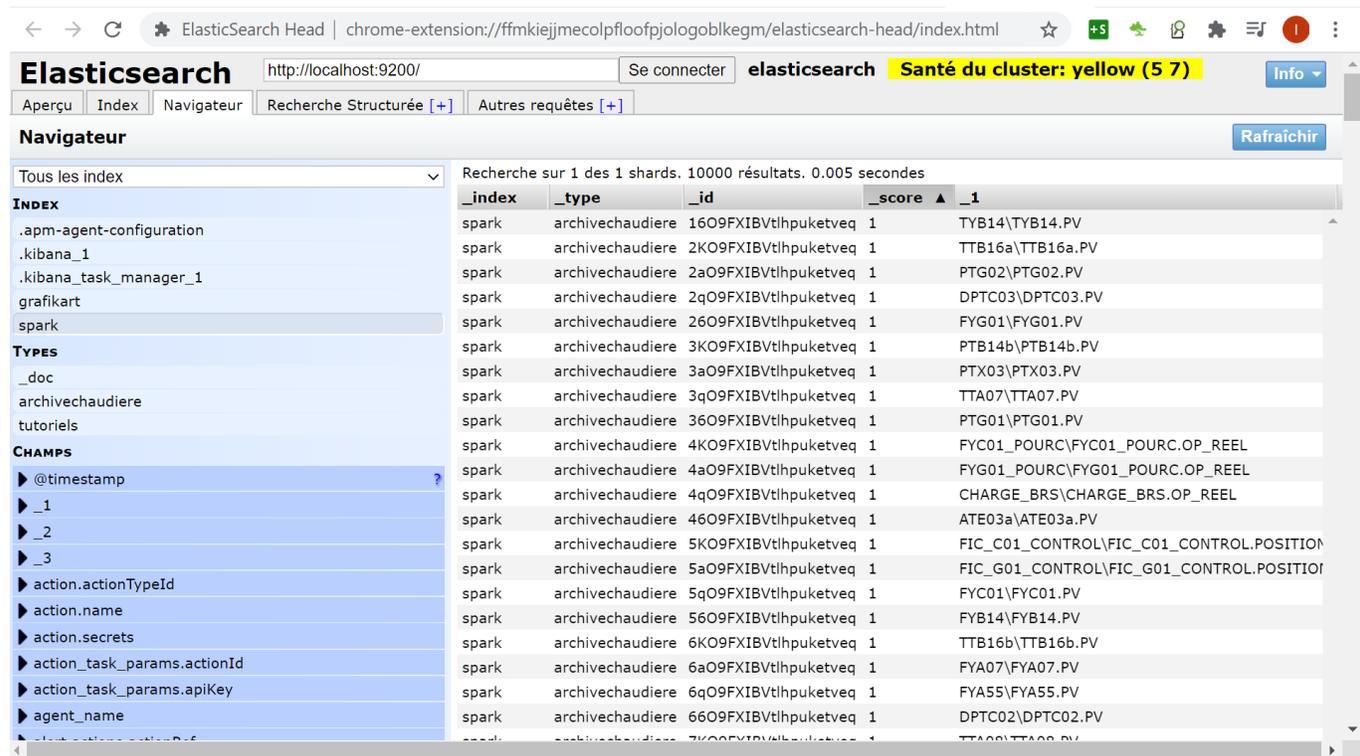


FIGURE 4.1 – stockage via spark

4 Restitution des Données via elasticsearch et kibana

Pour une bonne visualisation de nos données elasticsearch, nous avons opté pour Kibana qui nous permettra d'analyser, indexer et rechercher avec des performances proches de l'instantané .

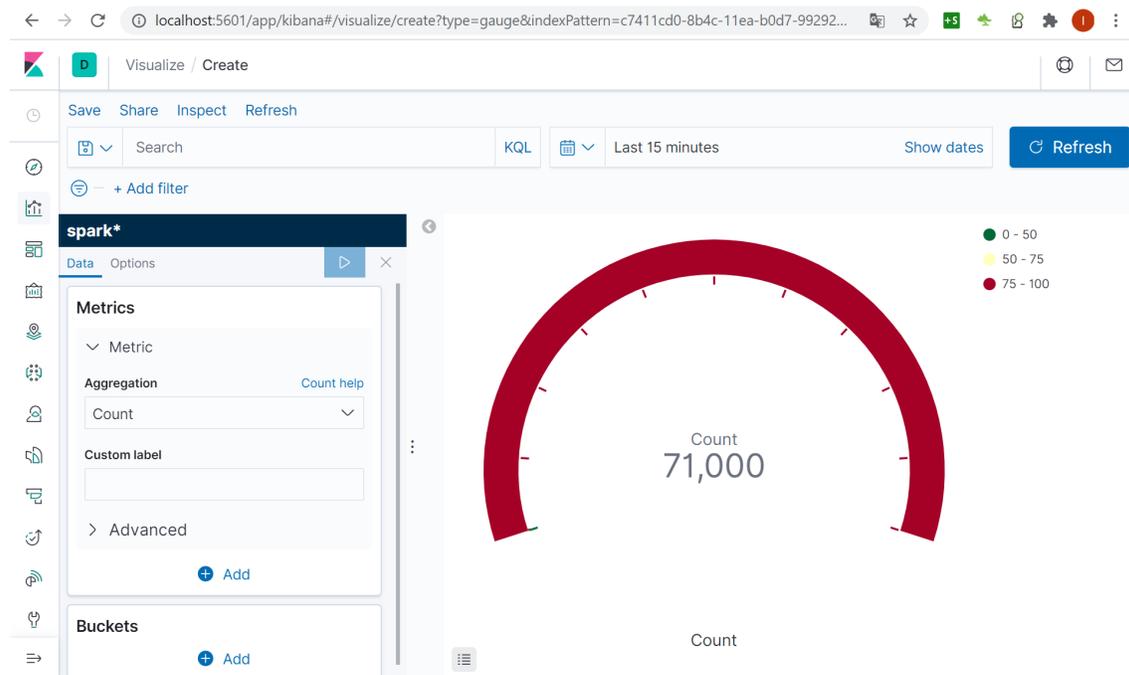


FIGURE 4.2 – restitution de données via kibana

5 Conclusion

Les fichiers ont pu être récupérés et ont été visualisés sur Kibana. Toutefois, par cause de beaucoup de contraintes, nous n'avons malheureusement pas pu achever notre travail et avoir des résultats concrets, d'où la brièveté de ce dernier chapitre. Nous allons tout de même nous surpasser afin de finir ce que nous avons commencé.

Conclusion générale

L'avenir du "business", du commerce et des entreprises internationales est le Big Data, en effet, les technologies qu'il propose sont indispensables pour la gestion et la manipulation de grandes masses de données des grandes entreprises tel que Cevital. celles-ci rencontrent d'innombrables difficultés à travailler avec des logiciels primitifs non dédiés au Big Data compte tenu de leur importante quantité de systèmes à gérer.

Notre étude est portée sur l'élaboration d'une solution pour optimiser le traitement des grandes données pour l'entreprise Cevital, au niveau de l'unité d'énergie. Notre proposition consiste à faciliter la manipulation de toutes sortes de données afin de voir et d'analyser en temps réel le déroulement et l'état de chaque équipement.

Il s'est d'abord agi pour nous dans ce travail de faire une analyse des données et des besoins de l'entreprise, ensuite, ayant un large choix, il a fallu connaître les avantages et les inconvénients de chaque logiciel et chaque plate forme à utiliser pour enfin savoir la démarche à prendre pour faire une mise en place irréprochable du projet. Nous avons donc étudié en premier lieu le parc informatique et ses composants (parc matériel, parc logiciel), les applications de gestion orientées métiers, ce point concerne : l'urbanisation des SI, les services WEB, l'ERP (Enterprise Resource Planning), le CRM (Customer Relationship Management), la GPAO(Gestion de Production Assistée par Ordinateur), la GMAO(Gestion de la Maintenance Assistée par Ordinateur) et le TMS(Transport Management System) Le troisième et dernier point de ce chapitre est porté sur le mouvement NoSQL au sein de l'entreprise, nous avons défini le NoSQL proprement dit pour ensuite nous élargir sur les infrastructure et logiciels et les plates-formes les plus utilisées et connues dans le monde du Big Data. Durant notre Stage au sein de la société Cevital, nous nous sommes dévoués pour rassembler toutes les informations sur les mécanismes et les programmes nécessaires de celle-ci afin d'apporter notre savoir faire et notre soutien technique pour l'amélioration de leur gestion des données.

Nous avons donc opté pour IntelliJ idea comme IDE pour l'implémentation du projet sous le langage Scala. L'analyse des données a été réalisé sous Apache Spark, un framework très fiable qui travaille en temps réel sur la totalité des données. Pour le stockage et la recherche nous avons utilisé le moteur distribué Elasticsearch et enfin, pour la visualisation des résultats, nous nous sommes penché sur Kibana, une plate-forme qui interagit directement avec Elasticsearch en analysant les données et en offrant des résultats quasi immédiat en forme de dashboard.

Tout au long de cette expérience, nous avons exploité toutes nos connaissances pour pouvoir créer le meilleur, dans les limites de nos espérances, pour notre client, qui a mis toute sa confiance en nous. Ce travail nous a permis d'acquérir une expérience unique et très enrichissante, de connaître le monde professionnel et de nous intéresser à un domaine dans lequel nous ne nous sommes jamais aventuré auparavant.

Cependant, pour la confiance que notre client a mis en nous, nous espérons parfaire notre projet en l'améliorant pour apporter une solution révolutionnaire à Cevital dans un avenir très proche.

Bibliographie

- [1] Manuj Aggarwal. *ElasticSearch Log Stash KibanaELK. 1 : Learn ElasticSearch*. Packt Publishing, 2018.
- [2] Syed Muhammed Fahad Akhtar. *Big Data Architect's Handbook*. Packt, 2018.
- [3] Alexandre Bonnet. *L'urbanisation des systèmes d'informations*, 2016.
- [4] Saurabh Chhaged. *Learning ELK Stack*. Packt Publishing, 2015.
- [5] Antoine Delers. *La gestion de relation client*. 50 minutes.fr, 2015.
- [6] J-P.Nernier. *Maintenace et GMAO :Tableau de bord, Organisation,Procédures*. Dunod, 2017.
- [7] Johann Kohler Pierre-Yvec cloux, Thomas Garlot. *Docker, Pratique des architectures à base de conteneurs*. Dunod, 2016.
- [8] M.Mmorel J.-L.Raffaelli P.Lemberger, M.Batty. *Big Data et Machine Learning*. Dunod, 2015.
- [9] Jillus Quddus. *Maching Learning with Appacch Spark*. Packt publishing, 2018.
- [10] Ray. *Entreprise Resource Planning*. Tata McgrawHill Education, 2011.
- [11] Tom White. *Hadoop The Definitive Guide*. O'Reilly Media, 2009.