 Département de génie logiciel et des TI	PROJET LOG/GTI 792	DOCUMENT NO.	DATE 2016-09-11	VERSION 1.0
	TITRE Ticksmith Analytics		PAGE 1	PAGES 9


PROPOSITION
Projet de fin d'études
Département de génie logiciel et des TI

Ticksmith Analytics

Auteurs
Philippe Grenier-Vallée
GREP02028906

Professeur superviseur
Alain April

Date
16 Septembre 2016

 Département de génie logiciel et des TI	PROJET LOG/GTI 792	DOCUMENT NO.	DATE 2016-09-11	VERSION 1.0
	TITRE Ticksmith Analytics		PAGE 2	PAGES 9

Suivi des changements

*A – Ajouté M – Modifié S – Supprimé

NUMÉRO DE VERSION	DATE aaaa/mm/jj	NUMÉRO DE FIGURE, TABLE OU SECTION	A* M S	BRÈVE DESCRIPTION DU CHANGEMENT	NUMÉRO DE DEMANDE DE CHANGEMENT
1	2016/09/11	Sections 1 à 8, Annexe A	A	Brouillon	
2	2016/09/16	Sections 1 à 8, Annexe A	M	Version finale	



 Département de génie logiciel et des TI	PROJET	DOCUMENT NO.	DATE	VERSION
	LOG/GTI 792		2016-09-11	1.0
	TITRE		PAGE	PAGES
	Ticksmith Analytics		3	9

TABLE DES MATIÈRES

1.	Définition, acronymes et abréviations	4
2.	Problématique et contexte	4
2.1	Analyse de données financières massives.....	4
2.2	Utilisation de services Big Data élastiques.....	4
3.	Objectifs du projet.....	5
3.1	Analyse de données financières massives.....	5
3.2	Utilisation de services Big Data élastiques.....	5
4.	Méthodologie.....	5
4.1	Développement de la plateforme extensible de statistiques financières.....	5
4.2	Conception de l'approche pour déploiement sur les services EMR d'AWS	6
4.3	Conception de l'interface utilisateur de la plateforme de statistiques.....	6
5.	Livrables et planification	6
5.1	Description des artefacts.....	6
5.2	Planification	7
6.	Risques	7
7.	Techniques et outils.....	7
8.	Références	8
	ANNEXE A : Plan de travail.....	9

 Département de génie logiciel et des TI	PROJET	DOCUMENT NO.	DATE	VERSION
	LOG/GTI 792		2016-09-11	1.0
	TITRE		PAGE	PAGES
	Ticksmith Analytics		4	9

1. DÉFINITION, ACRONYMES ET ABRÉVIATIONS

Apache Impala	Base de données utilisées par Hadoop et Spark.
Apache Spark	Moteur rapide et qui peut être mis à l'échelle (Big Data) pour le traitement de données.
Apache Parquet	Format de stockage de données colonnaire.
Avro	Format de stockage de données linéaire de données utilisant JSON.
AWS	Amazon Web Services; un ensemble de services web offerts par Amazon.
Docker	Logiciel libre d'automatisation de création d'images Linux et de déploiement d'applications.
EMR	Elastic Map Reduce, un des services faisant partie des AWS permettant de faire de faire rouler des scripts Map Reduce, Sparks et autre sur une grappe de la taille de notre choix.
JSON	Standard ouvert de format de données textuelles.
Scala	Langage de programmation basé sur Java avec un support pour la programmation fonctionnelle et supporté par Apache Spark.
S3	Système d'entrepotage avec haute redondance et une grande tolérance à la faute faisant partie des AWS.

2. PROBLÉMATIQUE ET CONTEXTE

2.1 Analyse de données financières massives

Présentement, *Ticksmith* offre les services d'une plateforme pour obtenir des données financières à grande échelle. Or, il revient aux clients de faire certaines analyses de ces données extraites des marchés boursiers.


Donc, puisque les clients veulent avoir des résultats d'analyses sur ces données massives (Big Data), ils doivent utiliser leur matériel et leurs outils statistiques ou bien payer pour d'autres services pour y arriver.

Ce problème touche principalement les clients de *Ticksmith*. Malgré qu'ils peuvent obtenir des données des marchés financiers très simplement avec les services déjà offerts, ils doivent tout de même effectuer certaines analyses suivantes par eux-mêmes. Il pourrait donc être intéressant pour eux d'avoir une solution qui pourrait subvenir à cette demande.

2.2 Utilisation de services Big Data élastiques

Les services de Big Data élastiques sont très récents dans le domaine. Ils permettent un déploiement facile d'une solution évolutive (« scalable ») dans une grappe de serveurs. Ce qui a de particulièrement intéressant avec une telle approche est qu'il est possible de choisir la taille de la grappe au moment où on veut s'en servir. Auparavant, il fallait louer une grappe de serveur d'une taille prédéterminée, ce qui limite la flexibilité de performance.

Ces technologies étant très récentes, peu de compagnies en font l'utilisation. Il pourrait être intéressant pour *Ticksmith* d'expérimenter avec ces services grâce à ce projet.

 Département de génie logiciel et des TI	PROJET LOG/GTI 792	DOCUMENT NO.	DATE 2016-09-11	VERSION 1.0
	TITRE Ticksmith Analytics		PAGE 5	PAGES 9

3. OBJECTIFS DU PROJET

3.1 Analyse de données financières massives

La solution est d'offrir une plateforme extensible de génération d'analytiques et de statistiques intégrées aux services présentement offerts. Ceci permettrait d'avoir une solution tout en un. Les analytiques sont des données traitées à partir des données brutes qui serviront aux calculs de statistiques (par exemple : nombre de transaction pour une action dans une journée). Les statistiques seront souvent des agrégats d'analytiques sur une période de temps. Ces dernières devront être représentées visuellement dans des graphiques. Il faudra aussi permettre aux clients d'extraire des données traitées (analytiques) pour ensuite faire des analyses propriétaires. Un point important sera d'assurer l'extensibilité de la solution pour faciliter l'ajout de techniques de traitement de données et d'analyses par la suite.

Pour y arriver, des données prêtes à être traitées dans le format *Avro* seront reçues (cette étape est faite par les services existants de *Ticksmith*) puis seront analysées par la solution de ce projet utilisant *Apache Spark*. Finalement, les résultats seront exportés sous le format *Parquet*.

3.2 Utilisation de services Big Data élastiques


Une approche pour atteindre cet objectif serait d'utiliser les services flexibles d'AWS qui chargent un prix en fonction de l'utilisation faite. En d'autres mots, le client d'AWS ne paie pas lorsqu'il n'utilise pas les machines. Aussi, AWS offre une solution d'entreposage (S3) de données avec un haut niveau de redondance et qui permet de retourner à des anciennes versions de fichiers si nécessaire. Plusieurs compagnies ont déjà reporté des économies substantielles lorsqu'ils ont migré leurs services de Big Data de serveurs traditionnels vers les services d'Amazon. De plus, il sera nécessaire d'utiliser un conteneur (Docker) pour augmenter la transportabilité de la solution offerte par le projet.

4. MÉTHODOLOGIE

Le projet proposé ici suit une série d'étapes qui seront effectuées itérativement. Chacune de ces étapes verront leur cycle de développement complété avant de passer à la prochaine et suivent l'ordre d'importance aux yeux de la compagnie partenaire du projet.

4.1 Développement de la plateforme extensible de statistiques financières

En tout premier lieu, il sera nécessaire de faire le développement complet de la plateforme de statistiques financières. Ce projet n'a pas de raison d'être sans cette partie. Pour débiter, une rencontre aura lieu avec un étudiant des cycles supérieurs en finances pour déterminer les analytiques et statistiques qui seront essentiels à intégrer dans la plateforme. Ensuite, il sera nécessaire de faire des recherches sur le fonctionnement d'Apache Spark pour déterminer comment l'utiliser correctement pour atteindre les objectifs de ce projet. Ensuite, il faudra concevoir l'architecture de la plateforme qui démontrera son aspect extensible. Ensuite, il y aura l'implémentation de la plateforme elle-même et des séries de tests locaux pour déterminer son bon fonctionnement. Dans ces tests, il sera de la première importance de vérifier que l'algorithme utilisé ne présente pas d'entonnoir séquentiel, il faut préserver une parallélisation en tout moment. Finalement, la documentation pour l'utilisation sera écrite.

 Département de génie logiciel et des TI	PROJET LOG/GTI 792	DOCUMENT NO.	DATE 2016-09-11	VERSION 1.0
	TITRE Ticksmith Analytics		PAGE 6	PAGES 9

4.2 Conception de l'approche pour déploiement sur les services EMR d'AWS

Malgré le fait que cette étape est plutôt petite comparativement à la dernière, elle n'est pas moins essentielle. Pour y arriver, il sera nécessaire de faire l'analyse des options offertes par les services d'*Elastic Map Reduce* offerts par les *Amazon Web Services*. Ce sont parmi les services les plus populaires à l'instant permettant de faire des calculs sur des grappes de serveurs conçues pour des données massives. Ces services viennent aussi avec du support pour *Apache Spark* nativement, ce qui est important pour ce projet. Aussi, comme mentionné précédemment, ces services viennent avec une solution d'entreposage de données avec redondance pour assurer aucune perte de données.

Il sera donc nécessaire de faire une analyse des options offertes pour ensuite concevoir un script de déploiement. Pour faciliter le déploiement, la solution sera mise dans un conteneur (*Docker*). Pour faire la conception du conteneur, une analyse des étapes de configuration de la machine sera faite. Aussi, il faudra effectuer des tests similaires à ceux de la dernière étape pour s'assurer que les calculs eux-mêmes fonctionnent bien sur ces serveurs distants. Ici, la documentation sera certainement le livrable le plus important pour cette étape pour s'assurer que *Ticksmith* soit en mesure d'utiliser les services *EMR* avec leurs paramètres.


4.3 Conception de l'interface utilisateur de la plateforme de statistiques

Un prototype statique a déjà été produit pour donner une idée de ce à quoi il est attendu du projet. L'interface sera créée en utilisant principalement *amCharts*. Évidemment, il sera nécessaire de faire des recherches sur l'utilisation de cette librairie. Ensuite, il faudra présenter des prototypes dynamiques pour s'assurer que le fonctionnement plait à *Ticksmith*. L'implémentation complète et la création d'une suite de tests seront évidemment les étapes qui suivent. Finalement, des tests complets utilisant toute la solution du projet seront conçus et effectués.

5. LIVRABLES ET PLANIFICATION

5.1 Description des artefacts

Nom de l'artefact	Description
Fiche de renseignements	Fiche de renseignements contenant les détails de base du projet.
Proposition de projet	Document contenant les détails de la planification du projet.
Schéma architectural	Schéma décrivant la conception du code de la plateforme extensible de statistiques financières.
Document de vision	Document détaillant plus précisément la planification du projet.
Conteneur <i>Docker</i>	Conteneur permettant un déploiement facile de la solution et qui automatise l'installation de l'image de la machine.
Rapport d'étape	Document faisant un retour sur le déroulement du projet à mi-chemin de la session.
Code de la plateforme et documentation	Code source de la plateforme de statistiques ainsi que la documentation permettant l'utilisation correcte de celle-ci.

 Département de génie logiciel et des TI	PROJET LOG/GTI 792	DOCUMENT NO.	DATE 2016-09-11	VERSION 1.0
	TITRE Ticksmith Analytics		PAGE 7	PAGES 9

Nom de l'artefact	Description
Série de tests locaux	Tests locaux permettant de prouver le bon fonctionnement de la plateforme de statistiques telle que livrée.
Scripts de déploiement et documentation	Scripts permettant de déployer la solution sur une grappe EMR d'AWS.
Série de tests sur la grappe	Tests sur la grappe permettant de s'assurer que la solution fonctionne correctement et respecte les niveaux de performance envisagés.
Prototype dynamique	Prototype basé sur le prototype statique présenté avec le projet qui permet un certain niveau d'interaction. Celui-ci permet de tester et de détecter des changements nécessaires à apporter pour l'implémentation finale de l'interface.
Code de l'interface utilisateur	Code représentant l'interface utilisateur en fonction des prototypes statiques et dynamiques.
Rapport final	Rapport final du projet contenant tous les détails du déroulement du projet.
Présentation	Présentation vulgarisée présentée devant les autres étudiants de PFE en génie logiciel.

5.2 Planification

Voir Annexe A.

6. RISQUES

Risque	Impact	Probabilité	Mitigation / atténuation
Limitation budgétaire pour les tests sur les grappes EMR.	Haut	Haut	S'assurer d'avoir fait tous les tests locaux possibles avant de passer aux tests sur les grappes EMR ainsi que le script de déploiement fasse exactement ce qu'il est prévu avant de passer aux tests.
Délais sur l'échéancier dû aux manques de connaissances des technologies utilisées par ce projet.	Haut	Medium	Il est toujours possible de prendre plus de temps que ce qui était prévu pour arriver aux dates prévues. Il sera nécessaire de contacter des personnes ressources pour aider si un mur est atteint.
Problèmes avec l'interface utilisateur suite à son implémentation selon les employés de <i>Ticksmith</i>	Medium	Faible	Il sera essentiel de faire valider l'interface le plus tôt possible avec un prototype dynamique.


7. TECHNIQUES ET OUTILS

Lucid Chart

Système de création de schémas incluant les standards UML.

Scala IDE


Environnement de développement spécialisé pour le langage de programmation Scala.

 Département de génie logiciel et des TI	PROJET LOG/GTI 792	DOCUMENT NO.	DATE 2016-09-11	VERSION 1.0
	TITRE Ticksmith Analytics		PAGE 8	PAGES 9

Sublime Text	Éditeur de texte.
Apache Maven	Outil de gestion et d'automatisation de la production de logiciels.
Docker	Logiciel libre d'automatisation de création d'images Linux et de déploiement d'applications.
Google Drive / Doc	Outil web permettant le partage de documents qui facilitera la communication entre les parties prenantes.
Microsoft Word	Logiciel de traitement de texte
Git / github	Outil de gestion de versions pour le code qui sera créé.

8. RÉFÉRENCES

- Document officiel de présentation du projet (« Ticksmith Analytics »)
- Notes personnelles des réunions avec la compagnie partenaire et le directeur du projet.
- Sites web (visités le 2016-09-16) :
 - o Amazon. Amazon EMR. <https://aws.amazon.com/emr/>
 - o Amazon. Amazon S3. <https://aws.amazon.com/s3/>
 - o Apache. Apache Spark™ is a fast and general engine for large-scale data processing. <http://spark.apache.org/>
 - o Apache. Welcome to Apache Maven. <http://maven.apache.org/>
 - o Apache. Apache Avro™ 1.8.1 Documentation. <https://avro.apache.org/docs/current/>
 - o Apache. Apache Parquet. <https://parquet.apache.org/documentation/latest/>
 - o Docker. What is Docker? <https://www.docker.com/what-docker>
 - o Scala. What is Scala? <http://www.scala-lang.org/what-is-scala.html>

 Département de génie logiciel et des TI	PROJET LOG/GTI 792	DOCUMENT NO.	DATE 2016-09-11	VERSION 1.0
	TITRE Ticksmith Analytics		PAGE 9	PAGES 9

ANNEXE A : PLAN DE TRAVAIL

Le tableau suivant présente la planification pour la réalisation des tâches ou artefacts décrits précédemment.

#	Commence	Termine	Efforts estimés*	Tâches/Jalon	Livrable(s)/Artéfacts
1	07-09-16	14-09-16	0.5	Remise de la fiche de renseignements	Fiche de renseignements
2	31-08-16	31-08-16	2	Rencontre – superviseur et compagnie	
2.1	14-09-16	14-09-16	2	Rencontre hebdomadaire – superviseur et compagnie	
2.x	xx-xx-16	xx-xx-16	2	Rencontre hebdomadaire – superviseur et compagnie	
3.	11-09-16	16-09-16	6	Analyse et définition du projet	Proposition de projet
4.	11-09-16	20-09-16	6	Analyse complète du projet	Document de vision
5.	14-09-16	27-09-16	8	Conception architecturale de la plateforme	Schéma architectural
6.	27-09-16	10-10-16	8	Conception du conteneur <i>Docker</i>	Conteneur <i>Docker</i>
7.	14-10-16	21-10-16	6	Revue détaillée du déroulement du projet	Rapport d'étape
8.	03-09-16	21-10-16	45	Recherches et implémentation de la plateforme de statistiques	Code de la plateforme et documentation
8.1	03-09-16	28-10-16	10	Conception et implémentation des tests formels locaux et ajustements de la plateforme	Série de tests locaux
9.	29-10-16	11-11-16	25	Recherches et conception des scripts de déploiement sur EMR	Scripts de déploiement et documentation
9.1	29-10-16	18-11-16	5	Conception et implémentation des tests formels sur les grappes EMR et ajustements de la plateforme et des scripts.	Série de tests sur la grappe
10.	19-11-16	26-11-16	10	Conception du prototype dynamique de l'interface utilisateur et vérification avec <i>Ticksmith</i>	Prototype dynamique
10.1	19-11-16	02-12-16	25	Implémentation de l'interface utilisateur	Code de l'interface utilisateur
10.2	03-12-16	09-12-16	10	Vérification de l'interface avec <i>Ticksmith</i>	
11.	21-11-16	15-12-17	20	Remise du rapport final	Rapport final
12.	03-12-19	17-12-19	10	Présentation – Création et pratique de la présentation	Présentation

* En heures