Département de génie logiciel et des TI

RAPPORT D'ÉTAPE

LOG 792. Projet de fin d'études
Département de génie logiciel et des TI

EndoMine
Projet de développement d'outils de forage de données de résultats de tests patients - endocrinologie, métabolisme et épidémologie clinique du JGH

Auteurs
Anton Zakharov
ZAKA12038406
David Lauzon
LAUD01028300

Professeur superviseur
Alain April

Date
29 Octobre 2012

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 CHANGEMENT
1.0	2012/10/28		A	Document Complete	

Table des matières

1. PROBLÉMATIQUE ET CONTEXTE

Actuellement le département de diagnostic médical du JGH possède une base de données de production Oracle contenant toutes les données diagnostics. Pour des raisons de sécurité des données et de performance, une base de données Access a été créée et elle est actuellement utilisée pour effectuer des extractions et des fonctions de « *data mining* ». Cette approche comporte des limites et le Dr. Eintrachtt aimerait avoir des propositions de solutions pour l'avenir d'un environnement de « *data mining* » à grande échelle. Parallèlement, le département d'endocrinologie du JGH aimerait avoir un outil de « *data mining* » pour effectuer des recherches spécifiques à son domaine d'expertise.

2. OBJECTIFS DU PROJET

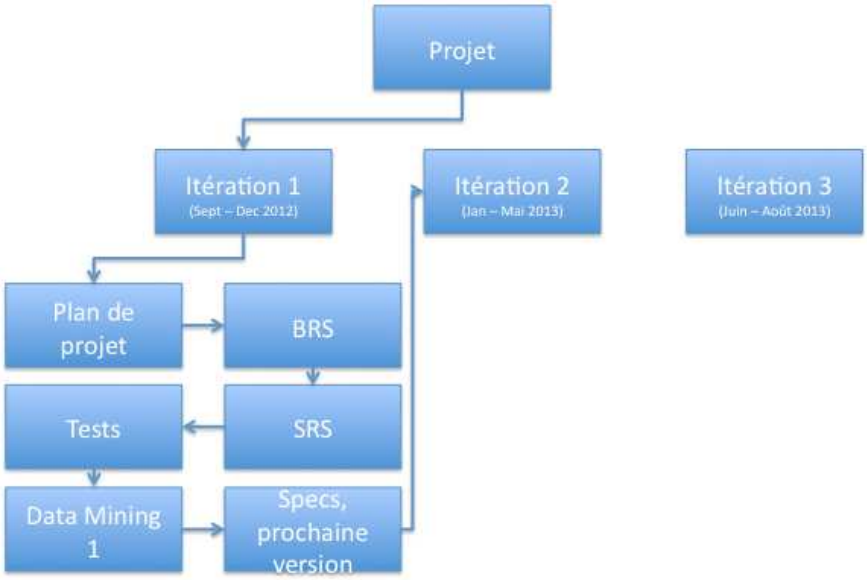
Notre projet contient plusieurs aspects :

- Adapter le scope du projet dépendant l'impact qu'il aurait sur différents intervenants.
- La mise en place d'une base de données exploitable par des outils de « *data mining* » qui remplacerait la base de données Access;
- Le choix ou le développement d'outils (à base de logiciels libres) de « *data mining* » afin de permettre l'étude des données d'une manière interactive. Nous avons décidé d'explorer les différentes technologies de l'écosystème «Hadoop » étant donnée leurs coût de mise à l'échelle peu élevée. Nous allons aussi explorer quelques solutions propriétaires.
- Reproduction de la B.D. de production Oracle actuelle dans l'environnement de « *data mining* ». Synchronisation entre la B.D. de production et l'environnement de « data mining » et de la base de données de *staging*.
- Création d'une interface permettant de : 1) faire la recherche dans l'environnement de « *data mining* » et 2) En affichage graphique des données (style tableau ou visuel).

3. MÉTHODOLOGIE

Vu la portée et la nature du projet l'approche itérative par phase (création, élaboration, construction, transition) du processus unifié (U.P.) sera utilisée. Les itérations et tâches exactes seront décrites dans l'Annexe A. Par contre certaines parties (comme par exemple reproduction de la BD de production) n'ont pas une définition exacte du temps, car nous n'avons aucune idée pour le moment quelle technologie nous allons utiliser. Ces questions seront résolues au fur et à mesure. Chaque itération prendrait 2 semaines (environ 15h de travail par personne par semaine). Un bilan serait fait à la fin de chaque évaluation, et les tâches seront réévaluées selon le temps disponible et l'évaluation des besoins du client.

Voici le schéma des livrable proposés :



4. SOMMAIRE DES TRAVAUX RÉALISÉS ET RECOMMANDATIONS

4.1 Sommaire des travaux réalisés

Pour la facile de présentation du projet nous allons diviser cette section en 3 parties : Définition des besoins du client, exploration des différents technologies et test effectuées sur différentes technologies.

Après nous allons décrire les installations effectuées.

4.1.1 Définition des besoins du client

Étant donné que le projet est réalisé pour un client réel, la création d'un document de vision décrivant les différents intervenants et leurs besoins est essentielle. Le document de vision complet avait été ajouté à l'annexe D.

Les besoins des clients pouvant être divisé en 3 projets avec une architecture différente, nous avons décidé de se concentrer sur le projet répondant au Dr. Eintrachtt. Ce dernier voudrais pouvoir faire de recherches rapidement sur la base de donnée avec un générateur de requêtes de type Ms Access. Voici un exemple :

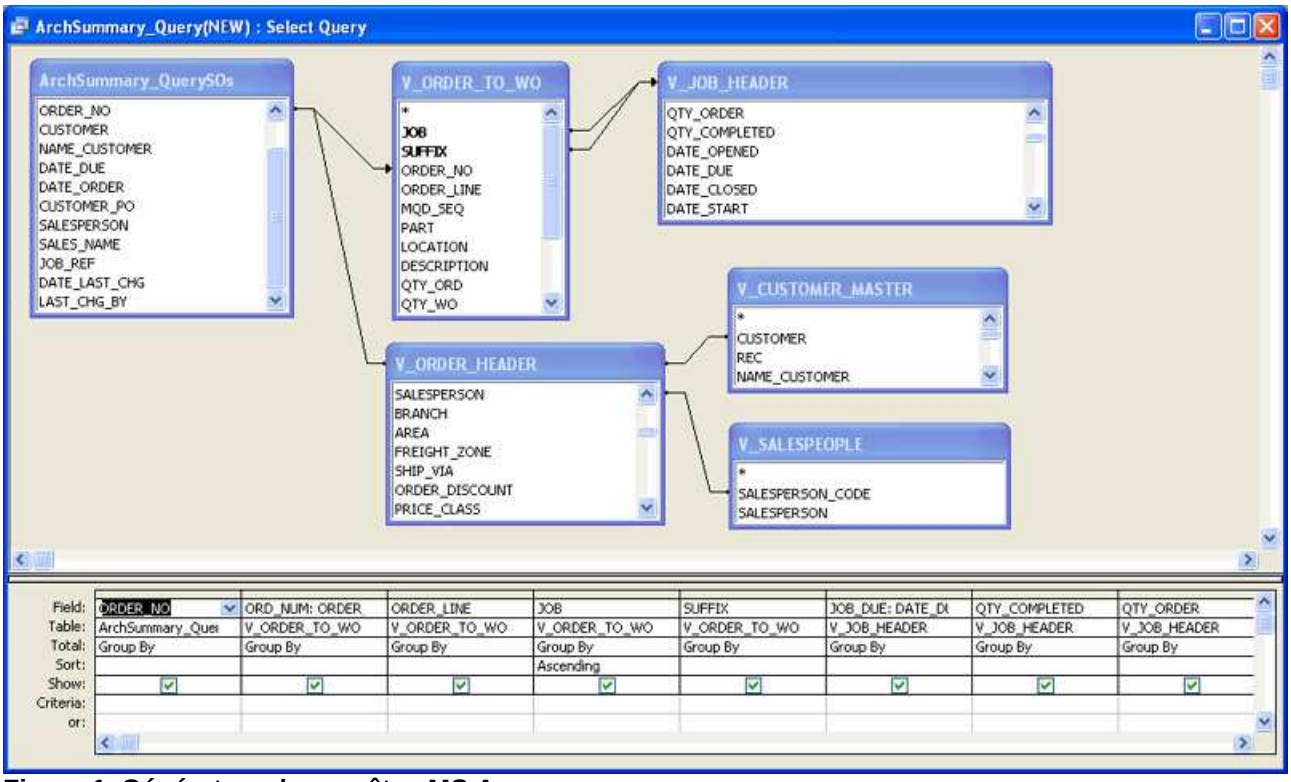


Figure 1. Générateur de requêtes MS Access.

Le nombre de tables dans la base de données est énorme. Comme nous ne faisons qu'un prototype, importer toutes les tables avec toutes les relations dépasse le scope de ce projet. Nous avons donc extrait et formaté les tables que le Dr. Einthracht allait utiliser.

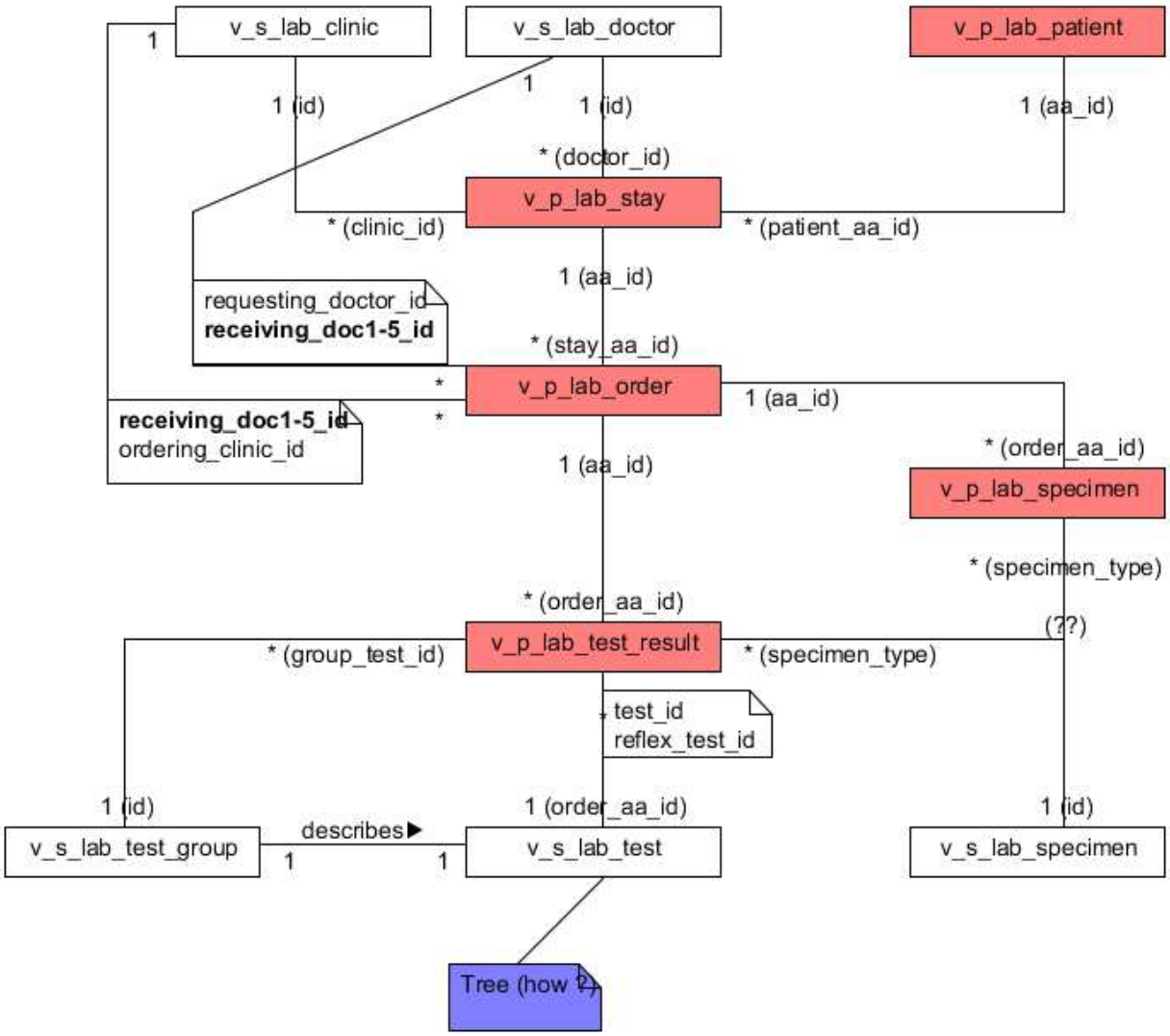


Figure 2. Relations entre les tables principales.

Les tables commençant par **v_p** contiennent les données particulières, alors que les **v_p** servent à décrire l'information dans les tables **v_p**. Par exemple, **v_s_lab_test** contient le nom du test et les caractéristiques propres servant à classifier un test par rapport aux autres. **V_p_lab_test_result** contient des données comme les dates des tests réalisés, ainsi que les valeurs reliées aux tests.

Voici la description des tables :

- **V_p_lab_patient** : est une table vide. Une clé anonymisée est utilisée pour rendre chaque patient unique.
- **V_p_lab_stay** : contient les séjours de patients dans l'hôpital.
- **V_p_lab_order** : contient la commande de tests effectués durant un séjour particulier.
- **V_p_lab_test_result** : contient les résultats des tests.

- **V_p_lab_specimen** : contient les spécimens physiques pris pour faire des tests.
- **V_s_lab_test_group** : décrit les composantes d'un test particulier.
- Les tables restantes n'ont pas besoins d'être décrits pour le moment.

Une fois les tables et les relations particulières effectuées, nous avons pris un soin de définir la taille de chaque table à partir d'un échantillon (de 1 jour de données). Voici les résultats obtenus :

Data Type	Table	vp_lab_stay	vp_lab_order	vp_lab_specimen	vp_lab_test_result	vs_lab_specimen	vs_lab_test	vs_lab_test_group	TOTAL
		34	**75**	**19**	**62**	**11**	**145**	**47**	**393**
	Bytes								
Number(5)	2	0	1	0	0	0	4	1	6
Number(10)	4	7	9	7	7	0	10	1	41
Number(14)	6	2	2	2	2	1	1	0	10
									0
double	8	0	0	0	0	0	11	0	11
									0
datetime	3	2	2	2	2	0	0	0	8
									0
char(1)	1	1	7	0	1	0	7	1	17
varchar(1)	1	8	33	4	40	5	56	0	146
varchar(2)	2	0	1	0	1	0	0	0	2
varchar(3)	3	1	0	1	1	1	2	0	6
varchar(4)	4	0	0	0	0	0	1	0	1
varchar(5)	5	8	18	3	7	0	13	41	90
varchar(7)	7	1	0	0	0	0	0	0	1
varchar(10)	10	0	0	0	0	1	0	0	1
varchar(11)	11	3	2	0	0	0	28	0	33
varchar(15)	15	0	0	0	1	1	3	0	5
varchar(23)	23	1	0	0	0	1	0	1	3
varchar(30)	30	0	0	0	0	0	1	0	1
varchar(39)	39	0	0	0	0	1	0	0	1
varchar(59)	59	0	0	0	0	0	2	2	4
varchar(79)	79	0	0	0	0	0	4	0	4
varchar(239)	239	0	0	0	0	0	1	0	1
mediumtext	1024	0	0	0	0	0	1	0	1
									0
Fixed Size		46	56	46	46	6	142	6	348

Variable Size		115	154	22	96	95	2218	347	3047
Max Size per row		161	210	68	142	101	2360	353	3395
(magnitude)		1	2	4	36				
per hour		84	140	367	3,048				3,639
per day		2,014	3,366	8,810	73,150	741	4,321	1,243	87,340
per week		14,098	23,562	61,670	512,050				611,380
per month		60,420	100,980	264,300	2,194,500				2,620,200
per year		735,110	1,228,590	3,215,650	26,699,750				31,879,100

Table 1. Row count statistiques

Comme on peut voir la majorité des tables v_p ont au moins 1 million de rangées par année. Faire les jointures sur des rangées de cette taille serait très difficile. Vue la taille des données, il va devenir très compliqué d'effectuer des jointures entre les grosses tables.

4.1.2 Exploration des différents technologies

Étant donné que le problème d'une grande quantité des données n'as pas de solution fixe, nous allons décrire les différentes méthodes analysées. Les bases de données Relationnelles, NoSQL Open Source et technologies propriétaires seront décrites.

4.1.2.1 Base de données relationnelles

Oracle :

Avant de réinventer la roue, nous avons pensé à utiliser les bases de données existantes. En théorie la base de donnée Oracle, présentement en production devrait permettre de faire des requêtes nécessaires à Dr. Eintracht. Il nous faudrait les données au complet, ainsi que des requêtes spécifiques du Dr. Eintracht pour pouvoir le tester.

Cette solution va couter une licence Oracle et ne pourrait pas vraiment être mise à niveau si le nombre de données augmente.

4.1.2.2 Base de données NoSQL Open Source

Les bases de données NoSQL dont on va présenter vont être bâties en utilisant le système de gestion de fichier distribué «Hadoop». Présenter l'architecture de «Hadoop» vas au delà du scope du présente document, mais si vous êtes intéressé voici un lien qui la décrit assez sommairement : http://en.wikipedia.org/wiki/Apache_Hadoop .

De base les données sont stockées dans des fichiers qui sont automatiquement répliqués sur plusieurs disques. MapReduce est un *framework* utilisé pour créer des tâches fonctionnent en parallèle et permettant de faire des opérations sur les fichiers stockés dans «Hadoop». Durant l'étape « **Map** », le nœud principal (master) prends les données d'entrées, les divises en problèmes plus petits, et les distribue aux nœuds de travail. Durant l'étape «**Reduce**» le nœud principal collecte les résultats de tous les sous-problèmes et les combine dans un résultat. Ce résultat est une réponse au problème posé initialement.(http://en.wikipedia.org/wiki/MapReduce).

Hive

Hive est une infrastructure de «data warehouse» construite en top de « Hadoop ». Hive permet de faire de l'analyse, des résumés et des requêtes dans «Hadoop» (**http://en.wikipedia.org/wiki/Apache_Hive)** avec un langage ressemblant de beaucoup à SQL. Ceci fait en sorte que c'est une solution très simple à implémenter.

Le problème avec Hive est qu'il est qu'il utilise les fonctions MapReduce qui sont très lentes à starter (aux alentour de 20 secs !!! selon nos test). De plus les jointures qui peuvent être faites sont limitées. De plus si jamais la structure des données de base change, les données devraient être réimportées d'oracle. Hive ne supporte pas (facilement) les opérations d'update ou d'effacement des données.

Hbase

HBase est une infrastructure par-dessus Hadoop, mais qui n'utilise pas MapReduce. HBase store les tables par colonnes (alors que traditionnellement ils sont stockées par rangées), ce qui luis permet de réduire grandement le temps de recherche (car on ne lits que les informations des colonnes dont on à besoin). HBase supporte les updates et les suppressions naturellement (avec son système de Versionnement). Par contre il n'y a aucun moyen simple et efficace de faire des jointures avec HBase.

Voici la structure d'une entrée dans HBase :

Figure 3. HBase entrée. (http://www.larsgeorge.com/2009/10/hbase-architecture-101-storage.html)

Faire des recherches en connaissant la valeur d'une clé, où d'un préfixe de clé est très rapide. Tous les autres types de recherche vont demander de faire un scan complet de la table, qui est plus inefficace dans HBase que dans Hive.

Il serait très difficile d'utiliser cette solution directement pour le Dr.Eintracht, car ce dernier à besoin de rechercher en utilisant comme critère plusieurs colonnes sans ordre prédéfini. Par contre il pourrait être difficile si le problème du Dr. Eintracht puisse être réduit à un schéma en étoile (http://en.wikipedia.org/wiki/Star_schema).

Impala (Cloudera)

Le problème principal de Hive est le « *start up cost* » à cause de l'utilisation du MapReduce. Impala est un projet séparé développé par Cloudera. L'objectif est d'être un engin de recherche en temps réel pour «Hadoop» (http://www.theregister.co.uk/2012/10/24/cloudera_hadoop_impala_real_time_query/) . En gros on peut le voir comme Hive, mais plus rapide (Important à noter il n'as pas toutes les possibilités de Hive mais dans notre cas d'utilisation on peut considérer comme tel. Voici le lien du papier de google sur lequel il est basé : http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//pubs/archive/38125.pdf).

Le projet est encore en stage beta. Pour cette raison nous allons éviter de l'utiliser à part si on n'a complètement pas le choix.

Pentaho :

Pentaho est un outil open source, pouvant intégrés divers sources de données ainsi que de faire des solutions de BI (Business Intelligence). Nous avons essayé les fonctions d'intégration est ils marchent plutôt bien. Nous n'avons pas encore eu le temps de tester les solutions BI (car nous ne savons pas si le problème de Dr. Eintracht peut être résumé à ces derniers). Le GUI pour les solutions BI n'est pas gratuit dans les majeures parties des cas.

4.1.2.3 Technologies Propriétaires

Datameer
À la conférence de Hadoop World, on a eu une démonstration d'un développeur de DataMeer. Le logiciel prend une approche de chiffrier et permet de faire des jointures, ainsi que des prévisualisations rapides (sur des sous-ensembles de données). Hadoop roule en back end.

- La version "Personal" (300$/an, 1-node) devrait suffire pour le prototype.
- La version "Enterprise" permet d'utiliser un cluster Hadoop (approx. 9000$/an).

Vertica, Greenplum, Teradata :

Des bases de données avec stockage par colonne. Toutes ces base de données sont chères, mais peuvent répondre au besoin de Mise à l'échelle.

4.1.3 Test Effectués sur différentes technologies

4.1.3.1 Installation de Hadoop et HBase :

Nous avons utilisé la distribution CDH4.1 de Cloudera car elle est stable et que l'installation est plus facile. De plus nous avons utilisé la version 1 de MapReduce (Mrv1), car la version 2 (Yarn) n'est pas encore considérée comme stable.

Voir **Annexe C** pour les instructions d'installation.

Le package Scoop est utilisé pour importer les données relationnelles dans HBase. Nous avons utilisé une base de données MySQL comme source de données initiale, mais ça devrait marcher avec la base de données Oracle.

4.2 Recommandations

Étant donné les données disponibles, nous recommandons d'avoir les requêtes exactes de Dr. Eintracht. S'il est possible d'avoir une solution modélisé en tant que schéma en étoile ou autre modèle dimensionnel, nous allons utiliser Pentaho. Si ce n'est pas possible nous n'aurions pas d'autre choix que d'utiliser Hive avec les pertes de performance que ça encours.

Dans ce cas le cas d'utilisation de Hive les données des tables V_P seront dé normalisés en une seule table, puis partitionnés par semaine. L'outil Visual SQL Query Builder pourrait être utilisé pour faire le Front End.

5. LIVRABLES ET PLANIFICATION

5.1 Description des artéfacts

Nom de l'artefact	Description
Plan de projet	Plan décrivant brièvement le projet, l'architecture / technologie et les personnes ressources, permettant d'initier le dialogue avec le client.
Échéancier (plan de travail)	Plan décrivant la distribution des tâches dans le temps. Les tâches sont sujettes à changement.
Proposition de projet	Document décrivant la proposition du projet, incluant la description, la proposition, les risques et l'allocation du temps.
Schéma relationnel de production actuel	Schéma relationnel des tables utiles de la base de données de production. Pour des raisons de confidentialité, il est possible que certaines tables soient renommées ou retirées du schéma.
Document de Vision (B.R.S.)	Document donnant plus de détails que la proposition du projet. Inclut la description et la portée. Inclut aussi le besoins du client, les caractéristiques du système ainsi que les demandes matérielles du système. Permet la discussion entre divers parties prenantes (« *stake holders* »).
Structure de stockage des données	Définit la structure des fichiers dans Hadoop (HDFS, Hbase, Hive, etc.).
Schéma architectural	Décrit les vues architecturales du système de « *data mining* ».
Rapport d'étape	Rapport intermédiaire pour le PFE.
Exigences systèmes (S.R.S.)	Exigences détaillées du système.
Scénarios de test / Cas de test	Décrit les différents cas et scénarios de test pour confirmer la validité du système pour le client.
Prototype de "data mining"	Prototype fonctionnel utilisable par le client.
Spécifications de prochaine version	Spécifications pour les personnes désirant poursuivre le projet avec le client.
Rapport final d'équipe	Rapport final d'équipe à remettre à la fin du projet.
Rapport final - DL	Rapport final de David Lauzon à remettre à la fin du projet.
Rapport final - AZ	Rapport final d'Anton Zakharov à remettre à la fin du projet.

5.2 Planification

Voir Annexe A.

6. RISQUES

Les changements par rapport à la proposition de projet sont en **gras**.

Risque	Impact	Probabilité	Mitigation / atténuation

R1. La portée (« *scope* ») du projet est très ambitieuse.	**Faible**	**Moyen**	L'emphase du projet sera sur la documentation précise; ce qui facilitera les prochaines équipes qui travailleront sur le projet. Le projet sera réalisé en plusieurs courtes itérations afin de pouvoir obtenir régulièrement de la rétroaction de la part du client. Le prototype réalisé de répondra qu'à un seul besoin du client; ce qui fournira une preuve de concept et offrira un minimum de fonctionnalités au client. **Nous avons diminué le scope afin de répondre au besoin du Dr. Eintracht seulement.**
R2. La confidentialité des données soit compromise.	Élevé	Faible	Afin de réduire le risque que les données confidentielles sur les patients soient divulgués, les informations personnelles sur les patients seront omises du système «Hadoop». Seul le Dr. Shaun aura accès aux données confidentielles.
R3. Les technologies utilisées sont mal connues.	**Moyen**	**Faible**	Lire la documentation et faire des exercices. De plus nous considérons la possibilité d'assister à la conférence «Hadoop World» en Octobre, ce qui inclue une activité de formation. **Nous avons une bonne idée des limitations des technologies.**
R4. L'écosystème «Hadoop» ne répond pas aux besoins du client.	**Moyen**	**Moyen**	Nous allons commencer à s'informer sur «Hadoop» le plus tôt possible. De plus nous allons collaborer étroitement avec le client pour nous assurer qu'il comprend la solution et qu'il reçoit la formation nécessaire. Nous nous somment informés. Les besoins du **Dr. Shawn ne pourrons pas être résolus entièrement par aucun système sans ajout d'argent / machines supplémentaires.**
R5. L'importation des données est inexacte.	Élevé	Faible	Nous allons développer un bon jeu de test, pour s'assurer que les données sont bien importées. **Nous allons probablement skipper cette partie.**

R6. Le système développé est moins performant que la solution Oracle existante.	**Moyen**	**Élevé**	Des discussions importantes avec le client seront réalisées afin de s'assurer que le système développé soit plus performant que le système existant. **Avec une seule machine c'est certains. Les systèmes distribués sont faits pour marcher avec plusieurs machines. Donc à court terme c'est inévitable.**
R7. Les besoins du client sont mal compris.	Élevé	Faible	Nous avons planifié des rencontres hebdomadaires avec le client, en plus de réaliser plusieurs documents précisant le projet de manière itérative (Vision, SRS, Cas d'utilisation, etc.). Chaque document sera présenté au client afin de collecter une rétroaction permettant d'ajuster le projet aux besoins du client. **Les besoins sont biens compris mais ne sont pas réalisables tels quels. Par contre la performance pourrait être améliorée si Dr. Shawn nous présente les besoins les plus courants de façon plus spécifique.**

7. RÉFÉRENCES CONSULTÉES

La liste des références ayant été consultées pour la réalisation du travail, la recherche de technologies existantes, et l'étude de faisabilité est placée à l'**Annexe B**.

8. TABLE DES MATIÈRES DU RAPPORT

Introduction
- Problématique et contexte
- Objectifs du projet
- Méthodologie

Sommaire des travaux réalisés et recommandations
- Sommaire des travaux réalisés
- Définition des besoins du client
- Exploration de différentes technologies
- Test Effectués sur différentes technologies
- Choix technologique effectué.
  - Description de l'architecture.
  - Description d'installation des technologies.
  - Description du déploiement de la solution technologie chez le client.
  - Référence pour modifier du schéma / importation des nouvelles données.

ANNEXE A : PLAN DE TRAVAIL RÉVISÉ

*Indiquez les changements et ajouts en **gras**.*

#	Commence	Termine	Efforts Estimés*	Tâches/Jalon	Livrable(s)/ Artéfacts	Responsable(s)
	13/09/2012	***26/09/2012***		***Phase d'Inception - Itération I1***		
1.1	14/09/2012	14/09/2012	1	Rencontre – au JGH avec le client		DL
1.2	14/09/2012	14/09/2012	1.5	Rencontre – professeur superviseur		DL
1.3	14/09/2012	20/09/2012	1	Définition de la fiche de renseignements	Fiche de renseignements	DL
1.4	18/09/2012	28/09/2012	2	Planification du projet	Plan de projet	DL
2.1	20/09/2012	20/09/2012	4	Rencontre – au JGH avec le client		DL,AZ
2.2	20/09/2012	24/09/2012	4	Identification et répartition des tâches du projet par itération	Échéancier	DL
2.3	20/09/2012	24/09/2012	3	Définition de la proposition de projet	Proposition de projet	DL,AZ
	27/09/2012	***10/10/2012***		***Phase d'Inception - Itération I2***		
3.1	27/09/2012	27/09/2012	3	Rencontre – au JGH avec le client		DL,AZ
3.2	27/09/2012	03/10/2012	6	Analyse du système existant	Schéma relationnel de production actuel	DL,AZ
3.3	27/09/2012	03/10/2012	5	Recherche sur le forage de données à grande échelle (part 1)		DL
4.1	04/10/2012	04/10/2012	1.5	Rencontre – au JGH avec le client		DL,AZ
4.2	04/10/2012	10/10/2012	5	Identifier les problèmes du système existant		DL,AZ
4.3	04/10/2012	10/10/2012	8	Identifier les besoins et caractéristiques du nouveau système, et les prioriser		DL,AZ
4.4	04/10/2012	10/10/2012	5	Recherche sur le forage de données à grande échelle (part 2)		DL
4.5	01/10/2012	12/10/2012	5	Élicitation des besoins d'affaire.	Document de	AZ

					Vision (B.R.S.)	
	11/10/2012	***21/10/2012***		***Phase d'Élaboration - Itération E1***		
5.1	11/10/2012	11/10/2012	1.5	Rencontre – au JGH avec le client		DL,AZ
5.2	11/10/2012	17/10/2012	4	Concevoir les cas d'utilisations		DL,AZ
5.3	11/10/2012	17/10/2012	5	Identifier les exigences fonctionnelles, non-fonctionnelles, et contraintes de conception		AZ
5.4	11/10/2012	17/10/2012	5	Établir une stratégie de miroir de la BD de production actuelle		DL
5.5	11/10/2012	17/10/2012	5	Recherche sur le forage de données à grande échelle (part 3)		DL
6.1	18/10/2012	18/10/2012	1.5	Rencontre – au JGH avec le client		DL,AZ
6.2	18/10/2012	21/10/2012	5	Dénormaliser le schéma relationnel et sélection des champs importants		DL,AZ
6.3	18/10/2012	21/10/2012	5	Définir le schéma des relations Hadoop	Structure de stockage des données	DL,AZ
6.4	18/10/2012	21/10/2012	10	Concevoir une architecture pour le "data mining"	Schéma architectural	DL,AZ
6.5	18/10/2012	21/10/2012	5	Design des interface graphiques (recherche, résultats, etc.)		AZ
6.6	18/10/2012	21/10/2012	5	Recherche sur le forage de données à grande échelle (part 4)		DL
	22/10/2012	***07/11/2012***		***Phase d'Élaboration - Itération E2***		
7.1	22/10/2012	22/10/2012	1.5	Rencontre – au JGH avec le client		DL,AZ
7.2	22/10/2012	22/10/2012	1	Rencontre – professeur superviseur		DL,AZ
7.3	23/10/2012	25/10/2012	18	Formation/Conférence Hadoop World		DL,AZ
7.4	27/09/2012	29/10/2012	3	Réévaluation des objectifs du projet	Rapport d'étape	DL,AZ
7.5	15/10/2012	31/10/2012	5	Définition des exigences du client	Exigences systèmes (S.R.S.)	AZ

8.1	01/11/2012	01/11/2012	1.5	Rencontre – au JGH avec le client		DL,AZ
8.2	01/11/2012	09/11/2012	5	Élaboration des scénarios de test / cas de tests	Scénarios de test / Cas de test	DL,AZ
8.3	01/11/2012	07/11/2012	10	Extraction/Conversion des données de BD production vers BD "miroir"		AZ
8.4	01/11/2012	07/11/2012	20	Installation de l'éco-système Hadoop		DL
	08/11/2012	***21/11/2012***		***Phase de Construction - Itération C1***		
9.1	08/11/2012	08/11/2012	1.5	Rencontre – au JGH avec le client		DL,AZ
9.2	08/11/2012	14/11/2012	10	Implémentation de la structure Hadoop pour stocker les données		DL
9.3	08/11/2012	14/11/2012	5	Importer les données dans Hadoop		DL
9.4	08/11/2012	14/11/2012	5	Tester l'intégrité des données et des relations		DL
10.1	15/11/2012	15/11/2012	1.5	Rencontre – au JGH avec le client		DL,AZ
10.2	15/11/2012	21/11/2012	20	Implémentation de l'engin de recherche (part 1)		AZ
	22/11/2012	***05/12/2012***		***Phase de Construction - Itération C2***		
11.1	22/11/2012	22/11/2012	1.5	Rencontre – au JGH avec le client		DL,AZ
11.2	22/11/2012	22/11/2012	1	Rencontre – professeur superviseur		DL,AZ
11.3	22/11/2012	28/11/2012	5	Implémentation de l'engin de recherche (part 2)		AZ
11.4	22/11/2012	28/11/2012	5	Tester l'engin de recherche		DL
11.5	22/11/2012	28/11/2012	10	Implémentation de l'affichage et exportation des résultats		DL
11.6	22/11/2012	28/11/2012	5	Interface avec le système de visualisation des données		DL
11.7	22/11/2012	28/11/2012	15	Tester le prototype de "data mining"		AZ
12.1	29/11/2012	29/11/2012	1.5	Rencontre – au JGH avec le client		DL,AZ
12.2	29/11/2012	30/11/2012	2	Démonstration d'un prototype de "data mining"	Prototype de "data mining"	DL,AZ
12.3	29/11/2012	05/12/2012	2	Manuel d'utilisation	Manuel	AZ

					d'utilisation	
	06/12/2012	***14/12/2012***		***Phase de Transition - Itération T1***		
13.1	06/12/2012	06/12/2012	1.5	Rencontre – au JGH avec le client		DL,AZ
13.2	03/12/2012	14/12/2012	4	Documentation sur la continuation du projet	Spécifications de prochaine version	DL,AZ
13.3	06/12/2012	10/12/2012	2	Présentation oral (David Lauzon)		DL
13.4	06/12/2012	10/12/2012	2	Présentation oral (Anton Zakharov)		AZ
13.5	06/12/2012	12/12/2012	4	Rapport final d'équipe	Rapport final d'équipe	DL,AZ
13.6	06/12/2012	12/12/2012	5	Rapport final individuel (David Lauzon)	Rapport final - DL	DL
13.7	06/12/2012	12/12/2012	5	Rapport final individuel (Anton Zakharov)	Rapport final - AZ	AZ
	13/09/2012	***14/12/2012***		***JALON 1 : Prototype de data mining***		

* En heures.
** En heures. Efforts réels.

Total d'heures David : 207,5
Total d'heures Anton : 185
Grand total : 392,5

ANNEXE B : RÉFÉRENCES

Auteur	Date de l'article	Titre de l'article	Date de consul-tation	URL
Oracle		Java SE 6 Downloads	8 octobre 2012	http://www.oracle.com/technetwork/java/javasebusiness/downloads/java-archive-downloads-javase6-419409.html
Charles Toepfer	4 mai 2012	Install Oracle Java JDK 7 in Ubuntu / Mint 12	8 octobre 2012	http://www.iokom.com/drupal/node/9
Cloudera		Before You Install CDH4 on a Single Node	8 octobre 2012	https://ccp.cloudera.com/display/CDH4DOC/Before+You+Install+CDH4+on+a+Single+Node#BeforeYouInstallCDH4onaSingleNode-SupportedOperatingSystemsforCDH4
Cloudera		Deprecated Properties	8 octobre 2012	http://archive.cloudera.com/cdh4/cdh/4/hadoop/hadoop-project-dist/hadoop-common/DeprecatedProperties.html
Cloudera		CDH Version and Packaging Information	8 octobre 2012	https://ccp.cloudera.com/display/DOC/CDH+Version+and+Packaging+Information#CDHVersionandPackagingInformation-CDHVersion4.1.0Packaging
Cloudera		CDH4 Installation	8 octobre 2012	https://ccp.cloudera.com/display/CDH4DOC/CDH4+Installation
Cloudera		Installing CDH4 on a Single Linux Node in Pseudo-distributed Mode	8 octobre 2012	https://ccp.cloudera.com/display/CDH4DOC/Installing+CDH4+on+a+Single+Linux+Node+in+Pseudo-distributed+Mode
Cloudera		HBase Installation	8 octobre 2012	https://ccp.cloudera.com/display/CDH4DOC/HBase+Installation#HBaseInstallation-InstallingHBase
Cloudera		ZooKeeper Installation	8 octobre 2012	https://ccp.cloudera.com/display/CDH4DOC/ZooKeeper+Installation#ZooKeeperInstallation-InstallingtheZooKeeperServerPackage
Apache Sofware Foundation	19 février 2010	HDFS File System Shell Guide (hadoop 0.20.2)	8 octobre 2012	http://hadoop.apache.org/docs/r0.20.2/hdfs_shell.html
Apache Sofware Foundation		Apache HBase Book	8 octobre	http://archive.cloudera.com/cdh4/cdh/4/hbase/book.html

Auteur	Date de l'article	Titre de l'article	Date de consul-tation	URL
			2012	
Nathan	12 septembre 2011	HBase Error – assignment of -ROOT- failure – Cant connect to web interface	8 octobre 2012	http://blog.nemccarthy.me/?p=110
Dominic Williams	28 janvier 2010	Quick install HBase in "pseudo distributed" mode and connect from Java	8 octobre 2012	http://ria101.wordpress.com/2010/01/28/setup-hbase-in-pseudo-distributed-mode-and-connect-java-client/
Alex Baranau	9 avril 2012	HBaseWD: Avoid RegionServer Hotspotting Despite Sequential Keys	9 octobre 2012	http://blog.sematext.com/2012/04/09/hbasewd-avoid-regionserver-hotspotting-despite-writing-records-with-sequential-keys/
Lars George	novembre 2011	Hadoop World 2011: Advanced HBase Schema Design	9 octobre 2012	http://www.slideshare.net/cloudera/hadoop-world-2011-advanced-hbase-schema-design
Lars George	june 2012	Berlin Buzzwords June 2012: Advanced HBase Schema Design	9 octobre 2012	http://www.slideshare.net/larsgeorge/hbase-advanced-schema-design-berlin-buzzwords-june-2012
Oracle		Using Connector/ODBC with Microsoft Access	11 octobre 2011	http://dev.mysql.com/doc/refman/5.5/en/connector-odbc-examples-tools-with-access.html
Oracle		Configuring a Connector/ODBC DSN on Windows	11 octobre 2011	http://dev.mysql.com/doc/refman/5.5/en/connector-odbc-configuration-dsn-windows.html
Oracle		MySQL: Download Connector/ODBC	11 octobre 2011	http://dev.mysql.com/downloads/connector/odbc/
Justin Grégoire	4 mars 2010	MySQL ODBC 32 vs 64 bit (answer from Justin Grégoire)	11 octobre 2011	http://stackoverflow.com/questions/2381906/mysql-odbc-32-vs-64-bit
Google		Google Refine	12 octobre 2012	http://code.google.com/p/google-refine/
OpenTSDB		Open Time Series Database	12 octobre 2012	http://opentsdb.net/

Auteur	**Date de l'article**	**Titre de l'article**	**Date de consul-tation**	**URL**
Ian Varleys	Août 2009	No Relation: The Mixed Blessings of Non-Relational Databases	14 octobre 2012	http://ianvarley.com/UT/MR/Varley_MastersReport_Full_2009-08-07.pdf
Ikai Lan	25 janvier 2011	App Engine datastore tip: monotonically increasing values are bad	14 octobre 2012	http://ikaisays.com/2011/01/25/app-engine-datastore-tip-monotonically-increasing-values-are-bad/
Michael Stack	20 février 2011	HBASE-3551 Issue: Loaded hfile indexes occupy a good chunk of heap; look into shrinking the amount used and/or evicting unused indices	14 octobre 2012	https://issues.apache.org/jira/browse/HBASE-3551
Gaojinchao	24 mai 2011	HBase, mail # user - a question storefileIndexSize	14 octobre 2012	http://search-hadoop.com/m/hemBv1LiN4Q1/a+question+storefileIndexSize&subj=a+question+storefileIndexSize
Sreejith P. K.	24 mars 2011	HBase, mail # user - Stargate+hbase	15 octobre 2012	http://search-hadoop.com/m/nvbiBp2TDP/Stargate%252Bhbase&subj=Stargate+hbase
Matteo Bertozzi	9 février 2011	HBase I/O: HFile	15 octobre 2012	http://th30z.blogspot.ca/2011/02/hbase-io-hfile.html?spref=tw
Apache Sofware Foundation		HFile Source Code	15 octobre 2012	http://hbase.apache.org/xref/org/apache/hadoop/hbase/io/hfile/HFile.html
Michael Stack	13 février 2009	HBASE-1200 Issue: Add bloomfilters	15 octobre 2012	https://issues.apache.org/jira/browse/HBASE-1200
Wikipedia	12 octobre 2012	Bloom filter	15 octobre 2012	http://en.wikipedia.org/wiki/Bloom_filter
khan	15 mai 2012	StackOverflow: HBase MemStore and Garbage Collection	15 octobre 2012	http://stackoverflow.com/questions/10596717/hbase-memstore-and-garbage-collection
leon	5 août 2011	StackOverflow: Where does HBase store all the row keys?	15 octobre 2012	http://stackoverflow.com/questions/6956400/where-does-hbase-store-all-the-row-keys

Auteur	Date de l'article	Titre de l'article	Date de consul-tation	URL
Lars George	12 octobre 2009	HBase Architecture 101 - Storage	15 octobre 2012	http://www.larsgeorge.com/2009/10/hbase-architecture-101-storage.html
Panks	22 janvier 2012	StackOverflow: How to Scan HBase Rows efficiently	15 octobre 2012	http://stackoverflow.com/questions/8961989/how-to-scan-hbase-rows-efficiently
AlexBaranau	3 août 2012	Re: How to query by rowKey-infix	15 octobre 2012	http://permalink.gmane.org/gmane.comp.java.hadoop.hbase.user/28109
Lars George	mai 2012	HBaseCon: HBase Coprocessors	15 octobre 2012	http://www.slideshare.net/cloudera/3-h-base-coprocessors-hbase-con-may-2012
Jimbojw	22 mai 2008	Understanding HBase and BigTable	16 octobre 2012	http://jimbojw.com/wiki/index.php?title=Understanding_Hbase_and_BigTable
Hypertable		HYPERTABLE VS. HBASE PERFORMANCE EVALUATION II	16 octobre 2012	http://hypertable.com/why_hypertable/hypertable_vs_hbase_2/
William Kang	10 octobre 2010	grokbase: [HBase-user] Hbase internally row location mechanism	16 octobre 2012	http://grokbase.com/t/hbase/user/10ab7vvfzy/hbase-internally-row-location-mechanism
Lars Hofhansl	27 janvier 2012	Scanning in HBase	16 octobre 2012	http://hadoop-hbase.blogspot.ca/2012/01/scanning-in-hbase.html
Quora	4 février 2011	How are bloom filters used in HBase?	16 octobre 2012	http://www.quora.com/How-are-bloom-filters-used-in-HBase
Culvert	2012	Culvert: A Robust Framework for Secondary Indexing	16 octobre 2012	https://github.com/booz-allen-hamilton/culvert
Apache Software Foundation		HBase API Docs: Package org.apache.hadoop.hbase.coprocessor	16 octobre 2012	http://hbase.apache.org/apidocs/org/apache/hadoop/hbase/coprocessor/package-summary.html
Lars Hofhansl	7 Octobr	Musings on Secondary Indexes	16 octobre	http://hadoop-hbase.blogspot.ca/2012/10/musings-on-secondary-indexes.html

Auteur	Date de l'article	Titre de l'article	Date de consul-tation	URL
	e 2012		2012	
Yonatan	2 Juin 2011	how to design Hbase schema?	16 octobre 2012	http://stackoverflow.com/questions/375194/how-to-design-hbase-schema
Cloudera		Installing the Sqoop RPM or Debian Packages	16 octobre 2012	https://ccp.cloudera.com/display/CDH4DOC/Sqoop+Installation#SqoopInstallation-installRPM
Cloudera		Scoop User Guide: 7.2.11. Importing Data Into HBase	16 octobre 2012	http://archive.cloudera.com/cdh4/cdh/4/sqoop/SqoopUserGuide.html#_importing_data_into_hbase
Apache Software Foundation		Sqoop Developer's Guide v1.4.1-cdh4.1.0		http://archive.cloudera.com/cdh4/cdh/4/sqoop/SqoopDevGuide.html
J J Singh	2 mars 2012	Sqoop installation tutorial	16 octobre 2012	http://jugnu-life.blogspot.ca/2012/03/sqoop-installation-tutorial.html
Becky Benton	4 octobre 2011	Google Groups: Cloudera Forum: Unable to load com.mysql.jdbc.Driver		https://groups.google.com/a/cloudera.org/forum/?fromgroups=#!topic/sqoop-user/pwdahVfAAAc
Oracle		MySQL: Download Connector/J	16 octobre 2012	http://www.mysql.com/downloads/connector/j/
Apache Sofware Foundation		HBase Book : 16.1. Using existing ZooKeeper ensemble	16 octobre 2012	http://hbase.apache.org/book/zookeeper.html
N Keywal	13 avril 2012	Hbase-user mailing list: Zookeeper available but no active master location found	16 octobre 2012	http://mail-archives.apache.org/mod_mbox/hbase-user/201204.mbox/%3CCAPcDmSviyQXUG8u5dPmarjakFddndyCbm6Pcgd8T6AR-aErw3g@mail.gmail.com%3E
Jimmy Xiang, Cloudera	18 juin 2012	Cloudera Developer Center: HBase Write Path	19 octobre 2012	http://www.cloudera.com/blog/2012/06/hbase-write-path/
Alex Baranau	16 juillet 2012	Configuring HBase Memstore: What You Should Know	19 octobre 2012	http://blog.sematext.com/2012/07/16/hbase-memstore-what-you-should-know/

Auteur	Date de l'article	Titre de l'article	Date de consul-tation	URL
Alex Holmes	24 septembre 2012	How partitioning, collecting and spilling work in MapReduce	19 octobre 2012	http://grepalex.com/2012/09/24/map-partition-sort-spill/
Cade Metz (Wired)	21 août 2012	Google's Mind-Blowing Big-Data Tool Grows Open Source Twin	19 octobre 2012	http://www.wired.com/wiredenterprise/2012/08/googles-mind-blowing-big-data-tool-grows-open-source-twin/?utm_source=Contextly&utm_medium=RelatedLinks&utm_campaign=Previous
Dj Walker-Morgan (h-online)	21 août 2012	Apache to Drill for big data in Hadoop	19 octobre 2012	http://www.h-online.com/open/news/item/Apache-to-Drill-for-big-data-in-Hadoop-1671686.html
Tomer Shiran	9 août 2012	Apache Drill : Interactive Analysis of Large-Scale Datasets	19 octobre 2012	http://wiki.apache.org/incubator/DrillProposal?action=AttachFile&do=view&target=Drill+slides.pdf
Apache Software Foundation	9 août 2012	Apache Incubator Wiki : Drill Proposal	19 octobre 2012	http://wiki.apache.org/incubator/DrillProposal
Wikipedia		Pentaho	19 octobre 2012	http://en.wikipedia.org/wiki/Pentaho
Pentaho		Video Pentaho Presentation MapReduce	19 octobre 2012	http://www.pentaho.com/resources/videos/67/pentaho-mapreduce-a-major-league-baseball-use-case/
Pentaho		Pentaho Community Edition (CE) : Community Wiki Home	19 octobre 2012	http://wiki.pentaho.com/display/COM/Community+Wiki+Home
Pentaho		Pentaho Community Edition (CE) : Latest Stable Builds	19 octobre 2012	http://wiki.pentaho.com/display/COM/Latest+Stable+Builds
Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, Theo Vassilakis Google, Inc.	2010	Dremel: Interactive Analysis of WebScale Datasets	19 octobre 2012	http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//pubs/archive/36632.pdf

Auteur	Date de l'article	Titre de l'article	Date de consul-tation	URL
Akbar Ahmed	24 mai 2012	Install Pentaho BI Server 4.5 on Ubuntu 12.04 LTS Desktop	19 octobre 2012	http://akbarahmed.com/2012/05/24/install-pentaho-bi-server-4-5-on-ubuntu-12-04-lts-desktop/
Akbar Ahmed	29 mai 2012	Install Kettle 4.3.0 on Ubuntu 12.04 LTS	19 octobre 2012	http://akbarahmed.com/2012/05/29/install-kettle-4-3-0-on-ubuntu-12-04-lts/
Sébastien Dejean	8 décembre 2006	Installing a Pentaho demo server on Ubuntu 6.10 Server Edition	19 octobre 2012	http://ubuntu-pentaho.blogspot.ca/
codek	20 novembre 2008	Pentaho Forums: Kettle repository - use and how to create?	19 octobre 2012	http://forums.pentaho.com/showthread.php?65955-Kettle-repository-use-and-how-to-create
bizcubed	12 Sept 2012	Joining Data Sources Performing Lookups and Joins in Pentaho Data Integration	19 octobre 2012	http://www.youtube.com/watch?v=na6yRrhX5yo
Pentaho		Configure Pentaho for Cloudera CDH4	19 octobre 2012	http://wiki.pentaho.com/display/BAD/Configure+Pentaho+for+Cloudera+CDH4
Wikipedia		Classpath (Java)	19 octobre 2012	http://en.wikipedia.org/wiki/Classpath_(Java)
Mark Hall, John Paz	17 février 2012	HBase Input	19 octobre 2012	http://wiki.pentaho.com/display/EAI/HBase+Input
James Dixon	août 2010	Pentaho Hadoop Series: Big Data Analytics: Part 1 - 5	21 octobre 2012	http://www.pentaho.com/resources/videos/25/hadoop-series-part-1-big-data-architecture/
Quora		What are the advantages of Hadoop over distributed RDBMS?	21 octobre 2012	http://www.quora.com/What-are-the-advantages-of-Hadoop-over-distributed-RDBMS
IBM		Informix Features and Benefits	21 octobre 2012	http://www-01.ibm.com/software/data/informix/feature.html?S_CMP=rnav

Auteur	Date de l'article	Titre de l'article	Date de consul-tation	URL
Glenn Engstrand	13 novembre 2011	OLAP Versus Big Data	21 octobre 2012	http://datawarehouse.ittoolbox.com/groups/strategy-planning/dw-projectmanagement/olap-versus-big-data-4508739
DataMeer		DataMeer: Enterprise: Analytics at the speed of business	28 octobre 2012	http://www.datameer.com/enterprise/index.html
Marcel Kornacker & Justin Erickson	24 octobre 2012	Cloudera Impala: Real-Time Queries in Apache Hadoop, For Real	28 octobre 2012	http://blog.cloudera.com/blog/2012/10/cloudera-impala-real-time-queries-in-apache-hadoop-for-real/
Cloudera		Cloudera Impala 1.0 Beta Documentation	28 octobre 2012	https://ccp.cloudera.com/display/IMPALA10BETADOC/Cloudera+Impala+1.0+Beta+Documentation
legrand_legrand		SQLeo Visual Query Builder		http://sqleo.sourceforge.net/index.html
Tomer	11 mai 2012	Announcing the MapR Hive ODBC Driver		http://www.mapr.com/blog/269?Itemid=78
Apache Software Foundation		Hive JDBC Driver		https://cwiki.apache.org/Hive/hivejdbcinterface.html
Rahul Patodi	16 mars 2011	Hue Features	28 octobre 2012	http://www.technology-mania.com/2011/03/hue-features.html
Sachini Jayasekara	29 septembre 2012	Using Different Reporting Frameworks with WSO2 Business Activity Monitor	28 octobre 2012	http://wso2.org/library/articles/2012/09/using-different-reporting-frameworks-wso2-business-activity-monitor
The Eclipse Foundation		New and Notable Features within BIRT 3.7	28 octobre 2012	http://www.eclipse.org/birt/phoenix/project/notable3.7.php
xianglei	27 septembre 2012	phpHiveAdmin : Big data to Drive, Make easier for Hive	28 octobre 2012	http://www.phphiveadmin.net/
yahelc	26 janvier 2012	StackOverflow: JavaScript Boolean Search Query Builder Interface Library?	28 octobre 2012	http://stackoverflow.com/questions/9022033/javascript-boolean-search-query-builder-interface-library

Auteur	Date de l'article	Titre de l'article	Date de consul-tation	URL
salk31		RedQueryBuilder - JavaScript SQL Query Builder UI	28 octobre 2012	http://redquerybuilder.appspot.com/
Developer Extensions		Ext Grid Query Builder Example	28 octobre 2012	http://www.developerextensions.com/index.php/extjs-grid-query-builder
martinorth	23 mai 2012	Sencha: Ext.ux.window.VisualSQLQueryBuilder	28 octobre 2012	http://www.sencha.com/forum/showthread.php?208444-Ext.ux.window.VisualSQLQueryBuilder
martinorth		Visual SQL Query Builder	28 octobre 2012	http://www.cfsolutions.de/qb/
Ramesh Nair		Squel.js: Lightweight Javascript for building SQL query strings	28 octobre 2012	http://hiddentao.github.com/squel/
Servoy		Servoy: Query Builder	28 octobre 2012	http://wiki.servoy.com/display/public/DOCS/Query+builder;jsessionid=7D0E08E4433B09432DC71E4F5584DE72

ANNEXE C : INSTALLATION DE HADOOP ET HBASE (DANS LINUX)

```
==========================================================
Oracle JAVA JDK Installation
==========================================================
cd Downloads/
chmod a+x jdk-6u34-linux-x64.bin
sudo ./jdk-6u34-linux-x64.bin
sudo mv jdk1.6.0_34/ /usr/lib/jvm/
ls -l /usr/lib/jvm
sudo ln -s /usr/lib/jvm/jdk1.6.0_34 /usr/lib/jvm/java-6-oracle
sudo update-alternatives --install /usr/bin/java java /usr/lib/jvm/java-6-oracle/jre/bin/java 2
sudo update-alternatives --install /usr/bin/javac javac /usr/lib/jvm/java-6-oracle/bin/javac 1
sudo update-alternatives --install /usr/bin/javaws javaws /usr/lib/jvm/java-6-oracle/bin/javaws 1
sudo update-alternatives --config java
sudo update-alternatives --install /usr/lib/mozilla/plugins/mozilla-javaplugin.so mozilla-javaplugin.so /usr/lib/jvm/java-6-
oracle/jre/lib/amd64/libnpjp2.so 1
sudo mkdir -p /opt/google/chrome/plugins
sudo update-alternatives --install /opt/google/chrome/plugins/chrome-javaplugin.so chrome-javaplugin.so /usr/lib/jvm/java-6-
oracle/jre/lib/amd64/libnpjp2.so 1
sudo update-alternatives --config chrome-javaplugin.so
sudo update-alternatives --config mozilla-javaplugin.so
sudo vi /etc/profile.d/java.sh
                                        #!/bin/bash
 export JAVA_HOME=$( dirname $( dirname $( dirname $( readlink -e /usr/bin/java ) ) ) )
                                    export JDK_HOME=$JAVA_HOME
                                  export JRE_HOME=$JAVA_HOME/jre
export PATH=$JAVA_HOME/bin:$PATH
sudo chmod +x /etc/profile.d/java.sh
source /etc/profile.d/java.sh
java -version


==========================================================
Make sure the JAVA_HOME environment variable is set for the root user
==========================================================
sudo reboot
```

[CTRL]+[ALT]+[F1]
login as root
env | grep JAVA_HOME

==========
fix localhost problem
==========
Remove entries relating to 127.0.1.1 <COMPUTER NAME> from /etc/hosts
Make sure /etc/hostname matches the value in /etc/hosts

==========
CDH4 Installation
==========
# Default data dir: /var/lib/hadoop-hdfs/cache/hdfs/dfs/data
sudo -u hdfs hdfs namenode -format
# Start HDFS services
sudo service hadoop-hdfs-namenode start
sudo service hadoop-hdfs-secondarynamenode start
sudo service hadoop-hdfs-datanode start
# Create Hadoop directory structure
sudo -u hdfs hadoop fs -mkdir /tmp
sudo -u hdfs hadoop fs -chmod -R 1777 /tmp
sudo -u hdfs hadoop fs -ls /
sudo -u hdfs hadoop fs -mkdir /var
sudo -u hdfs hadoop fs -mkdir /var/lib
sudo -u hdfs hadoop fs -mkdir /var/lib/hadoop-hdfs
sudo -u hdfs hadoop fs -mkdir /var/lib/hadoop-hdfs/cache
sudo -u hdfs hadoop fs -mkdir /var/lib/hadoop-hdfs/cache/mapred
sudo -u hdfs hadoop fs -mkdir /var/lib/hadoop-hdfs/cache/mapred/mapred
sudo -u hdfs hadoop fs -mkdir /var/lib/hadoop-hdfs/cache/mapred/mapred/staging
sudo -u hdfs hadoop fs -chmod 1777 /var/lib/hadoop-hdfs/cache/mapred/mapred/staging
sudo -u hdfs hadoop fs -chown -R mapred /var/lib/hadoop-hdfs/cache/mapred
sudo -u hdfs hadoop fs -ls -R /
# Start Map Reduce services
sudo service hadoop-0.20-mapreduce-jobtracker start

sudo service hadoop-0.20-mapreduce-tasktracker start

==========
Map Reduce Test
==========
# Create a directory structure in HDFS for current user to run mapreduce jobs
sudo -u hdfs hadoop fs -mkdir /user/david
sudo -u hdfs hadoop fs -chown david /user/david
sudo -u hdfs hadoop fs -ls -R /
# Grep all config that starts with dfs from the hadoop config
hadoop fs -mkdir input
hadoop fs -put /etc/hadoop/conf/*.xml input/
hadoop fs -ls input
hadoop jar /usr/lib/hadoop-0.20-mapreduce/hadoop-examples.jar grep input output 'dfs[a-z.]+'
hadoop fs -ls
hadoop fs -ls output
hadoop fs -cat output/part-00000 | head

==========
HBase Installation in Standalone mode
==========
sudo apt-get install hbase
dpkg -L hbase

# Increase maximum number of open files for hdfs/hbase user
sudo vi /etc/security/limits.conf
hdfs - nofile 32768
hbase - nofile 32768
sudo vi /etc/pam.d/common-session
session required pam_limits.so

# Increase maximum number of files that can be served by a DataNode
sudo cp /etc/hadoop/conf/hdfs-site.xml /etc/hadoop/conf/hdfs-site.xml.initial
sudo vi /etc/hadoop/conf/hdfs-site.xml
<property>

```
<name>dfs.datanode.max.xcievers</name>
<value>4096</value>
</property>
```

```
# Restart HDFS services
sudo service hadoop-hdfs-datanode stop
sudo service hadoop-hdfs-secondarynamenode stop
sudo service hadoop-hdfs-namenode stop
sudo service hadoop-hdfs-namenode start
sudo service hadoop-hdfs-secondarynamenode start
sudo service hadoop-hdfs-datanode start

# Install HBase Master
sudo apt-get install hbase-master
sudo service hbase-master start

# Install HBase REST Interface
sudo apt-get install hbase-rest

sudo cp /etc/hbase/conf/hbase-site.xml /etc/hbase/conf/hbase-site.xml.initial
sudo vi /etc/hbase/conf/hbase-site.xml
```

```
<property>
<name>hbase.rest.port</name>
<value>60050</value>
</property>
```

```
sudo service hbase-rest restart
```

```
============================================================
Configuring HBase in Pseudo-distributed Mode
------
Pseudo-distributed mode differs from standalone mode in that each of the component processes (e.g. HBase Master, Region Server
and ZooKeeper peer) run in a separate JVM.
============================================================
# Stop standalone HBase Master
sudo service hbase-master stop
```

```
# Set host to matches fs.default.name or fs.defaultFS in core-site.xml
sudo vi /etc/hbase/conf/hbase-site.xml
                                        <property>
                                    <name>hbase.cluster.distributed</name>
                                        <value>true</value>
                                        </property>
                                        <property>
                                    <name>hbase.rootdir</name>
                                <value>hdfs://localhost:8020/hbase</value>
                                        </property>

# Create user for HBase
sudo -u hdfs hadoop fs -mkdir /hbase
sudo -u hdfs hadoop fs -chown hbase /hbase

# Installing the ZooKeeper Server Package and Starting ZooKeeper on a Single Server
sudo apt-get install zookeeper-server
sudo service zookeeper-server init
sudo service zookeeper-server start

# Tell HBase to use separate JVM for ZooKeeper in
# /etc/hbase/conf.dist/hbase-env.sh
export HBASE_MANAGES_ZK=false

# Starting HBase Master
sudo service hbase-master start

# Installing and Starting a HBase RegionServer
sudo apt-get install hbase-regionserver
sudo service hbase-regionserver start

# Verifying the Pseudo-Distributed Operation
sudo jps
sudo /usr/lib/jvm/jdk1.6.0_34/bin/jps
                                        32694 Jps
                                    30674 HRegionServer
                                        29496 HMaster
```

28781 DataNode
28422 NameNode
30348 QuorumPeerMain

# Installing and Starting the HBase Thrift Server
sudo apt-get install hbase-thrift

Verifying HBase shell
hbase shell

status 'detailed'

==========
Testing HBase
==========
$ sudo -u hbase hbase shell
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 0.92.1-cdh4.1.0, rUnknown, Sat Sep 29 11:55:59 PDT 2012

hbase(main):001:0> list
TABLE
0 row(s) in 0.5590 seconds

hbase(main):002:0> create 'test', 'cf'
0 row(s) in 1.2210 seconds

hbase(main):004:0> list
TABLE
test
1 row(s) in 0.0400 seconds

hbase(main):006:0> put 'test', 'row1', 'cf:a', 'value1'
0 row(s) in 0.0900 seconds

hbase(main):007:0> put 'test', 'row2', 'cf:b', 'value2'

```
0 row(s) in 0.0310 seconds

hbase(main):008:0> put 'test', 'row3', 'cf:c', 'value3'
0 row(s) in 0.0280 seconds

hbase(main):009:0> scan 'test'
ROW                  COLUMN+CELL
 row1                column=cf:a, timestamp=1349751879062, value=value1
 row2                column=cf:b, timestamp=1349751884145, value=value2
 row3                column=cf:c, timestamp=1349751892350, value=value3
3 row(s) in 0.0780 seconds

hbase(main):011:0> get 'test', 'row1'
COLUMN               CELL
 cf:a                timestamp=1349751879062, value=value1
1 row(s) in 0.0150 seconds

hbase(main):012:0> disable 'test'
0 row(s) in 2.1320 seconds

hbase(main):013:0> list
TABLE
test
1 row(s) in 0.0610 seconds

hbase(main):016:0> drop 'test'
0 row(s) in 1.4340 seconds

hbase(main):017:0> list
TABLE
0 row(s) in 0.0520 seconds


==========================================================
Scoop Installation
==========================================================
sudo apt-get install sqoop
```

sqoop help
sqoop version
sqoop import
http://archive.cloudera.com/cdh4/cdh/4/sqoop/SqoopUserGuide.html
http://archive.cloudera.com/cdh4/cdh/4/sqoop/SqoopDevGuide.html

Install MySQL JDBC Driver
tar -xzf ~/Downloads/mysql-connector-java-5.1.22.tar.gz
sudo cp mysql-connector-java-5.1.22/mysql-connector-java-5.1.22-bin.jar /usr/lib/sqoop/lib/
chmod a+r /usr/lib/sqoop/lib/mysql-connector-java-5.1.22-bin.jar

Should we set the connector path somewhere ??
#export SQOOP_HOME="/home/hadoop/software/sqoop-1.3.0"
#export PATH=$PATH:$SQOOP_HOME/bin

==========
Scoop Test
==========
ERROR sqoop.Sqoop: Got exception running Sqoop: java.lang.RuntimeException: Could not load db driver class: com.mysql.jdbc.Driver

Test JDBC connector and connection
java -cp .:/usr/lib/sqoop/lib/mysql-connector-java-5.1.22-bin.jar jdbctest 'jdbc:mysql://localhost/endomine_access?user=root&password=...'

Initial test (hbase key != sql key)
sqoop import -libjars /usr/lib/sqoop/lib/mysql-connector-java-5.1.22-bin.jar --connect jdbc:mysql://localhost/endomine_access --username root -P --table vp_lab_order --split-by AA_ID --hbase-table vp_lab_order --column-family d --hbase-row-key ID --hbase-create-table

list
get 'vp_lab_order', "H3010002"
get 'vp_lab_order', "H3010002", 'd:AA_ID'
disable 'vp_lab_order'
drop 'vp_lab_order'

ANNEXE D : DOCUMENT DE VISION

(attaché à la page suivante)

EndoMine

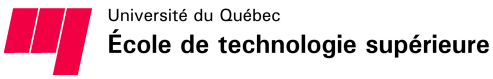
Projet de développement d'outils de forage de données de résultats de tests patients - endocrinologie, métabolisme et épidémiologie clinique

Projet #2012-076

Version	:	0.9.1
Date d'émission	:	4 octobre 2012
Date de révision	:	11 octobre 2012

Vision

Auteurs :



David Lauzon
Anton Zakharov

Présenté à :

Département d'endocrinologie

Table des matières

Liste des tableaux

Table des figures

TABLE 1: Historique des Révisions

Date	Version	Description	Auteur
27 sept. 2012	v0.1	Sections 2.2, 2.3, 3.4, 3.7, 4.1	Anton Zakharov
27 sept. 2012	v0.2	Sections 1.3, 3.2, 3.3, 4.1, 4.3	David Lauzon
3 oct. 2012	v0.2.1	Formattage et mise en page	David Lauzon
3 oct. 2012	v0.3-preview	Section 1.2 et 3.7	Anton Zakharov
5 oct. 2012	v0.4	Section 4.2, 4.4, 4.5, 5-7	Anton Zakharov
7 oct. 2012	v0.5	Section 1.3-1.5, 3.2-3.4, 8-11	David Lauzon
8 oct. 2012	v0.5.1	Revue des sections à David	Anton Zakharov
9 oct. 2012	v0.9-rc1	Revue des sections à Anton	David Lauzon
11 oct. 2012	v0.9.1	Corrections suggérées par Fodil (2.1, 4.4, 4.5, 9.3).	David Lauzon

1 Introduction

1.1 Objectif

Le but de ce document est de collecter, analyser, et définir les besoins et caractéristiques de haut niveau du système EndoMine. Il se concentre sur les fonctionalités recherchées par les parties prenantes, et explique pourquoi ces besoins existent. Les documents de cas d'utilisations (UC) et les spécifications des exigences logicielles (SRS) détaillent comment EndoMine satisfait ces besoins.

1.2 Portée

Ce document de vision porte sur le développement et l'intégration du système de forage de données EndoMine. L'outil permettrait aux chercheurs du JGH de réaliser du forage de données à grande échelle sur la banque de données de tests biomédicaux. Les requêtes pourraient être faites à partir d'une interface utilisateur, tout en respectant les règles de sécurité et de confidentialité de l'hôpital.

La figure 1 présente les relations entre les principaux concepts dans le domaine du client. Une étoile signifie "plusieurs". Par exemple, la relation entre *Test Order* et *Test* se lit comme suis : "1 commande de tests peut comporter plusieurs tests", et la relation entre *Test* et *Test Result* se lit : "un même test n'a qu'un seul résultat". Les diagnostics (en jaune) et son lien avec le séjour n'existe pas encore dans le contexte du département de biochimie médicale, mais il est attendu qu'ils seront ajoutés en cours de projet. Se référer a la section 1.3 (Définitions, acronymes et abbréviations) pour l'explication de ces concepts et autres terminologie mentionné dans ce document.

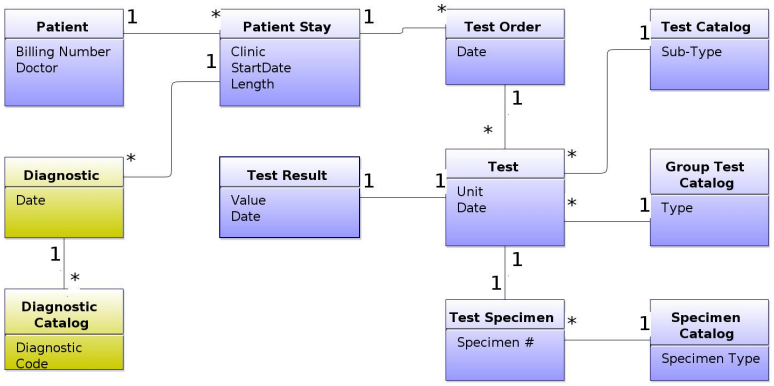


FIGURE 1 – Modèle du domaine

1.3 Définitions, acronymes et abréviations

BD, DB	Base de données / banque de données / *database*
Diagnostic	Associe le séjour d'un patient à un code de diagnostic.
Diagnostic Catalog	Décris un diagnostic (ex : diabète).
ÉTS	École de Technologie Supérieure

GÉLOG	Software Engineering Research Laboratory
Group Test Catalog	Décris le type général d'un test (ex : Testostérone).
JGH / HGJ	Jewish General Hospital / Hôpital général juif
Patient	Personne ayant un dossier (ou fiche) à l'hôpital JGH.
Patient Stay	Séjour d'un patient à l'hôpital associé à une clinique en particulier (ex : endocrinologie).
Specimen Catalog	Décris la méthode du test (ex : urine, sang, etc.)
SQL	Structured Query Language
Test	Décris le test à effectuer sur un patient. Contient des informations sur les unités du résultat de test (ex : g/mL).
Test Catalog	Décris le type spécifique d'un test (ex : Testostérone-B, Testostérone-C).
Test Order	Commande d'un ou plusieurs tests effectuée à une date précise pour un séjour en particulier.
Test Result	Contient les informations relatives au résultat du test (ex : quantité de mL de glucose).
Test Specimen	Fait référence à l'échantillon de test (ex : l'éprouvette de sang).

1.4 Références

Voir le Plan de Projet pour la liste des artéfacts livrables de ce projet.
Les autres références citées se trouvent à la fin dans la section 10.2 intitulée Bibliographie.

1.5 Langue de rédaction

La langue de rédaction principale des documents sera le français. Toutefois, lorsque approprié, l'anglais pourrait être utilisé pour s'assurer la compréhension de l'ensemble des utilisateurs d'EndoMine.

2 Positionnement

2.1 Énoncé du problème

2.1.1 Problème 1

Le problème de	l'extraction de statistiques complètes et précises sur l'utilisation des ressources médicales (*scanner*, microscope, tout ce qui ce trouve dans les laboratoires) est compliquée, voire impossible.
Cela affecte	les gestionnaires du laboratoire biomédical
dont l'impact est	la difficulté de prendre des décisions objectives d'achat et d'allocation de matériels.
Une bonne solution serait	d'ajouter les diagnostics médicaux et les relier aux tests médicaux dans le système de forage de données. Cela permettrait de prouver que des tests demandés par des médecins sont effectués inutilement, en vérifiant les tests avec les diagnostics.

2.1.2 Problème 2

Le problème de	forer des données à grande échelle temporelle (données échelonnées sur plusieurs années) prend trop de temps et ralentis le système pour tous les utilisateurs.
Cela affecte	les chercheurs et les utilisateurs d'équipements médicaux dans les laboratoires de biochimie
dont l'impact est	que les chercheurs doivent soit : a) limiter la quantité de données recherchées ou b) attendre que les résultats d'une recherche complète soient disponibles.
Une bonne solution serait	un système de forage de données rapide dont le traitement d'une recherche intensive soit quasi transparente pour les autres utilisateurs non concernés. De plus, la solution devrait s'adapter facilement à une quantité de données et à un nombre d'utilisateurs grandissants.

2.1.3 Problème 3

Le problème de	la limite de 1GB de résultats de MS Access ne permet pas de rechercher l'ensemble des données disponible (référence : [1]).
Cela affecte	les chercheurs utilisant MS Access
dont l'impact est	l'obligation de limiter la quantité de données traitées lors d'une même recherche.
Une bonne solution serait	un système de forage de données permettant de faire des recherches sur l'ensemble des données de SoftLab, tout en conservant une faciilité d'utilisation comme MS Access.

2.2 Positionnement du produit

Pour	les chercheurs et gestionnaires des départements d'endocrinologie et diagnostique médicale de l'hopital général Juif
qui	veulent une solution efficace pour faire des recherches dans leur banque de données.
EndoMine	est une solution de forage de données
qui	est efficace, rapide, et permet une mise à l'échelle (*scaling*) à faible coût.
Contrairement à	une base données relationnelle traditionnelle, onéreuse, surchargée et lente,
notre produit	est facile à utiliser, sauve du temps précieux dans la réalisation de recherches à grande échelle, tout en conservant un faible coût de possession (TCO).

3 Descriptions des intervenants et des utilisateurs

3.1 Résumé des intervenants (*stakeholders*)

TABLE 3 – Résumé des intervenants (*stakeholders*)

Nom	Description	Responsabilités
STK1. Dr. Elizabeth Mac Namara	Chief Medical Biochemistry (JGH)	Pilote (sponsor : states long term requirements).
STK2. Dr. Shaun Eintracht	Medical Biochemist, Dept. of Diagnostic Medecine (JGH)	Pilote (will operate and use the resulting data mining system).
STK3. Dr. Mark Trifiro	Chief Endocrinology and Metabolism (JGH)	Client (sponsor : states long term requirements).
STK4. Dr. Sami Suissa	Chief Clinical Epidemiology (JGH)	Client (sponsor : states long term requirements).
STK5. Dr. Alain April	Directeur du GÉLOG (ÉTS)	Provide Information Technology solutions and guide ÉTS students.
STK6. Chris Polykandriotis	IT Specialist (JGH)	Help with IT issues.

3.2 Résumé des utilisateurs

TABLE 4 – Résumé des utilisateurs

Nom	Description	Responsabilités	Intervenant
USR1. Utilisateur privilégié	Utilisateur ayant accès à la base de données de production Oracle, et au système EndoMine. Il s'agit des utilisateurs du département de biochimie du JGH.	– Faire des recherches spécifiques. – Produire des rapports de recherche – Gérer l'utilisation de ressources médicales.	STK1 (Dr. Mac Namara), STK2 (Dr. Eintracht)
USR2. Utilisateur restreint	Utilisateur externe au département de biochimie ayant accès seulement au système EndoMine (donc ils n'ont pas accès aux informations confidentielles des patients). Il s'agit des utilisateurs des autres départements du JGH.	– Faire des recherches spécifiques.	STK3 (Dr. Trifiro), STK4 (Dr. Suissa)

3.3 Environnement utilisateur

– Poste de travail ayant un accès intranet au serveur d'EndoMine.

3.4 Principaux besoins des intervenants et utilisateurs

TABLE 5: Besoins

Besoin	Priorité	Préoccupations	Solution actuelle	Solution proposée
N1. Confidentialité des données	Critique	Informations confidentielles protégées contre l'accès de personnes non autorisées.	Mesures de sécurité inconnues.	Voir la section 4.2 ('Sécurité et Confidentialité').
N2. Forage de données n'interférant pas avec la collection de données	Critique	Comme le forage est traité directement sur l'environnement de collection des données, des pannes surviennent bloquant temporairement l'accès à tous les utilisateurs.	Aucune	Séparation distincte des environnements de production et de forage de données.
N3. Outil simple pour créer des requêtes	Critique	Courte formation pour apprendre l'outil. Aucune connaissance de SQL requise.	Dr. Eintracht utilise MS Access Query Builder	Interface similaire à MS Access Query Builder. (voir la Figure 3 de l'10.2)
N4. Mise à l'échelle facile du système	Important	Accommoder un nombre de données et d'utilisateurs grandissant, sans que les coûts ne deviennent exponentiels.	Incapable de répondre à ce besoin.	Extensibilité horizontale graduelle en ajoutant d'autres ordinateurs au système. La configuration à modifier serait minimale.
N5. Forage de données sans limite de taille	Important	Pouvoir rechercher toutes les données disponibles.	La mémoire de MS Access limite la taille des recherches. [1]	L'architecture d'EndoMine n'as pas cette restriction.

3.5 Alternatives et Compétition

A notre connaissance il n'existe pas de solution clé en main pour faire directement du forage de données sur la structure de la base de données actuelle. Une implémentation spécifique aux besoins du JGH est donc nécessaire.

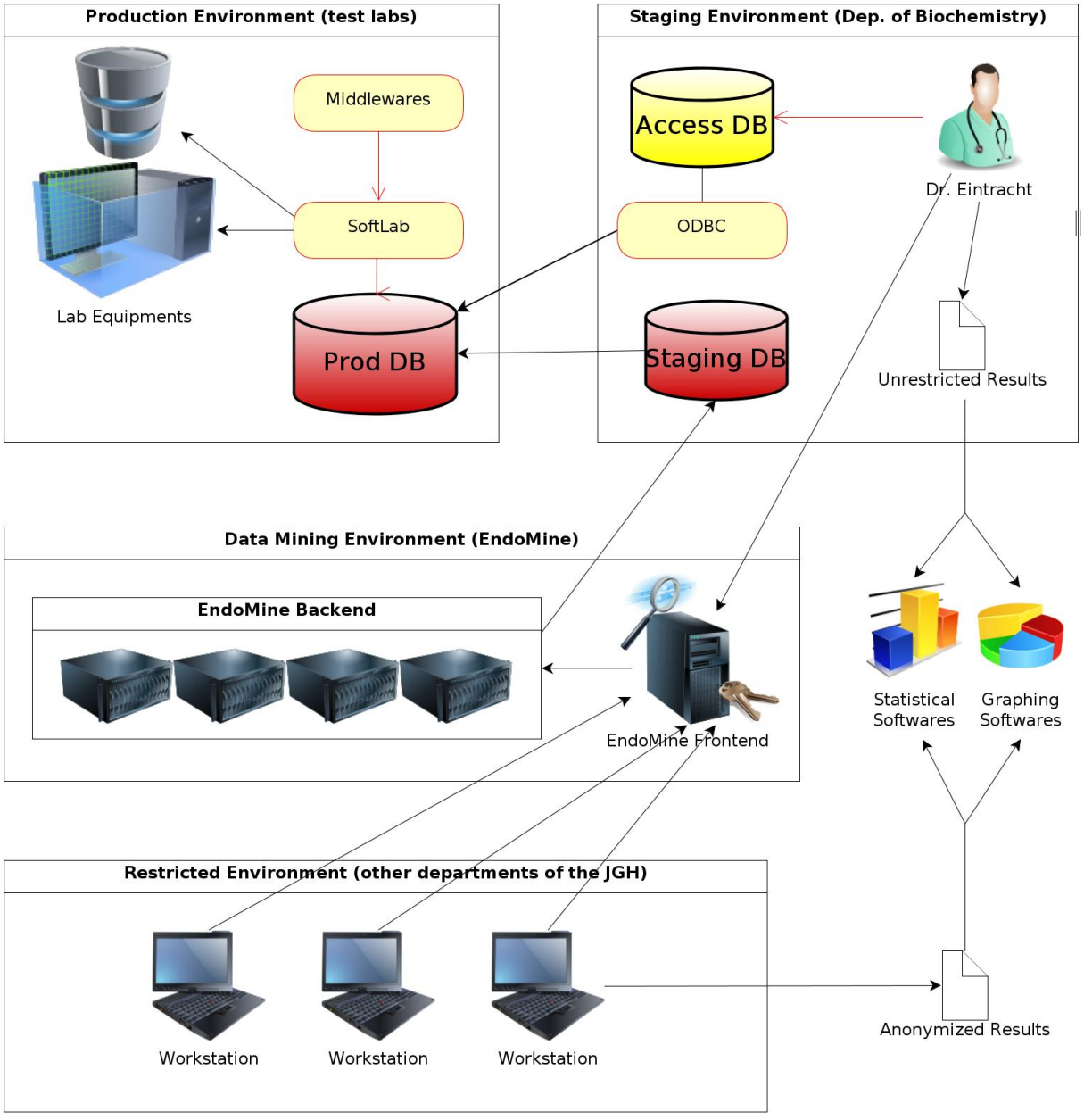


FIGURE 2 – Perspective du produit

4 Vue d'ensemble du produit

4.1 Perspective du produit

La figure 2 situe EndoMine en perspective avec les autres produits mis en place dans l'environnement du JGH.

- Le *Production Environment* représente l'environnement où les données sont collectées et la BD Oracle de production est populée par le logiciel SoftLab.
- La *Staging DB* est une copie miroir de la BD de production dont les informations confidentielles ont été anonymisées. **Seul le département de biochimie a accès au *Staging Environment*.**
- Les utilisateurs effectueront le forage de données sur le système EndoMine, et pourront par la suite importer leur résultats dans des logiciels de statistiques et de graphiques.

4.2 Sécurité et Confidentialité

Cette section décrit les différentes mesures à prendre pour protéger la confidentialité des données :
- L'accès au système EndoMine serait restreint selon les protocoles de sécurité du JGH.
- Les informations confidentielles (noms des patients, numéro de RAMQ, numéro d'hôpital, etc.) seront absentes du système EndoMine.
- EndoMine contiendra la clef anonymisée des patients, mais la correspondance entre la clef anonymisée et la fiche d'hôpital du patient ne sera pas accessible par le système EndoMine.
- Voir la section 4.4 ('Hypothèses et dépendances') pour plus d'informations.

4.3 Principaux avantages

TABLE 6: Avantages EndoMine

Bénéfices pour le client	Caractéristiques correspondantes
N1. Confidentialité et sécurité des données	HYP01, HYP03, HYP04, FEA05, FEA10, VC01, VC02, VC03
N2. Forage de données n'interférant pas avec la collection de données	HYP02, FEA05, FEA06, VC05
N3. Outil simple pour créer des requêtes	FEA07, FEA08, FEA09, FEA11, FEA12, VC04
N4. Mise à l'échelle facile du système	FEA01, FEA02, FEA03, FEA04, VC04
N5. Forage de données sans limite de mémoire	FEA02, FEA03, FEA11

4.4 Hypothèses et dépendances

HYP01 BD de *staging*

La BD de *staging* sera fournie par le JGH, et sera déja anonymisée (voir la section 4.2).

HYP02 Synchronisation

La BD de staging sera synchronisée automatiquement une fois par jour (ou toute autre fréquence jugée acceptable par le client).

HYP03 Anonymisation

Les scripts pour anonymiser / désanonymiser seront mis à la disposition du Dr. Eintracht par le JGH.

HYP04 Prototype

Lors du prototype, seul le Dr. Eintracht aura accès au système EndoMine installé au JGH.

HYP05 Conception, développement, et test

Les ressources de l'ÉTS doivent avoir accès aux données anonymisées et au système pilote pour les besoins de conception, développement et test.

4.5 Licences et installation

1. EndoMine sera implémenté avec logiciels libres ayant prouvé leur efficacité sur des sites tels que Facebook, Amazon et Yahoo. Donc aucune license à acheter.
2. Le code développé pour le logiciel EndoMine sera la propriété commune du JGH, de l'ÉTS et des développeurs ayant travaillé sur EndoMine. Les propriétaires du logiciel pourront installer une copie du logiciel sur d'autres serveurs et y effectuer des modifications.
3. Les données provenant de la base de données de production Oracle, ainsi que les résultats de recherches réalisés avec EndoMine sont la propriété exclusive du JGH.
4. Le système EndoMine sera installé sur le serveur du JGH par les développeurs de l'ÉTS.
5. Aucune installation requise sur les postes de travails des chercheurs.

5 Caractéristiques (*features*) du produit

FEA01 Configuration minimale pour ajouter d'autres machines.

L'objectif est de monter le système de telle façon que le coût de configuration / stabilité avec l'ajout d'une nouvelle machine parallèle soit minimal.

FEA02 Code supportant le parallélisme.

En lien avec FEA01, le forage de données devrait pouvoir exploiter le parallélisme du système de machines distribués.

FEA03 Distribution automatique des données pour supporter parallelisme

Une fois une nouvelle machine installée, le système devrait balancer automatiquement (le plus possible) la charge de traitement sur la nouvelle machine.

FEA04 Supporter plusieurs requêtes en parallèles

Plusieurs requêtes d'un même (ou plusieurs) utilisateur(s) devraient pouvoir être exécutés en parallèle.

FEA05 Environnement de forage de données distinct des sources de données en production

L'utilisation du système développe ne devrait avoir aucun impact sur le contenu ou le fonctionnement des sources de données en production.

FEA06 Synchronisation des données automatique entre les sources de données et le système EndoMine

EndoMine devrait pouvoir se synchroniser automatiquement avec les sources de données (dans l'environnement de *staging*) à intervalles régulières configurables.

FEA07 Générateur de requêtes intégré

Un générateur de requêtes permettant de faire des requêtes d'une maniére interactive et simple devrait être fournis. EndoMine s'inspirera de *Microsoft Access Query Builder* (voir la Figure 3 de l'10.2).

FEA08 Recherche par filtrage incrémentiel de la requête originale

Modification d'une requête en ajoutant des conditions de recherche, qui ne faisait pas partie de la requête originale. C'est-à-dire de pouvoir filtrer les résultats d'une recherche avec des nouveaux critères, sans que le système aille besoin de ré-exécuter la requête originale. Effectuer un filtrage secondaire devrait prendre une fraction du temps de la requête originale.

Par exemple, si on recherche les résultats de Glucose en 2010, on devrait pouvoir ajouter les résultats de Glucose en 2009 et/ou ajouter les résultats de Fructose.

FEA09 Exportation des résultats de recherche

Les résultats de recherche devraient pouvoir être exportés selon au moins 1 format d'échange de fichier d'un logiciel d'analyse statistique.

FEA10 Exportation des clés anonymisées

Permettre de facilement exporter seulement l'ensemble unique des clés anonymisées des patients présents dans les résultats d'une recherche.

FEA11 Ajout de champs supplémentaire à des résultats de recherche

À partir de résultats de recherche, il doit être possible d'ajouter des champs qui n'étaient pas inclus dans la sélection. Par exemple, on pourrait vouloir ajouter le champs sexe et âge du patient. Tous les champs disponibles dans EndoMine devraient pouvoir être ajoutés de cette façon.

FEA12 Trier les résultats de recherche

Pouvoir trier les résultats de recherche selon n'importe quel champ inclus dans dans la recherche.

6 Contraintes

VC01 Accessibilité

Le système EndoMine ne doit pas être accessible à l'extérieur du réseau du JGH.

VC02 Confidentialité

Les utilisateurs restreints ne peuvent accéder aux informations confidentielles sur les patients. Voir la section 4.2 (Sécurité et Confidentialité).

VC03 Règles du JGH

Les règles de sécurité informatique et de confidentialité du JGH doivent être respectées.

VC04 Coût

La réalisation et l'installation du prototype doit être de faible coût.

VC05 Modification

La base de données de production ne doit pas être modifiée par EndoMine.

7 Gammes de qualité

Le forage de données doit être au minimum deux fois plus rapide que le système actuel.

8 Attributs des caractéristiques

Le tableau suivant permet :

- Au client de prendre connaissance des efforts et risques associés au projet.
- Aux développeurs de jauger les bénéfices et priorité que le client associe à chaque caractéristique.
- Et d'identifier les caractéristiques les plus susceptibles de changer dans le futur.

La légende des valeurs possibles pour chacune des colonnes est présenté à l'10.2.

Table 7: Attributs des caractéristiques

Caractéristiques	État	Bénéfice	Effort	Risque	Stabilité	Priorité
FEA01. Configuration minimale pour ajouter d'autres machines	Proposé	Moyen	Moyen	Moyen	Élevé	Important
FEA02. Code supportant le parallélisme	Proposé	Moyen	Faible	Faible	Élevé	Important
FEA03. Distribution automatique des données pour supporter parallelisme	Proposé	Moyen	Faible	Faible	Élevé	Important
FEA04. Supporter plusieurs requêtes en parallèles	Proposé	Élevé	Faible	Faible	Élevé	Important
FEA05. Environnement de forage de données distinct des sources de données en production	Proposé	Élevé	Moyen	Faible	Moyen	Critique
FEA06. Synchronisation des données automatique entre les sources de données et le système EndoMine	Proposé	Élevé	Moyen	Moyen	Faible	Important
FEA07. Générateur de requêtes intégré	Proposé	Élevé	Élevé	Élevé	Faible	Critique
FEA08. Recherche par filtrage incrémentiel de la requête originale	Proposé	Moyen	Élevé	Élevé	Faible	Critique
FEA09. Exportation des résultats de recherche	Proposé	Élevé	Moyen	Faible	Faible	Critique
FEA10. Exportation des clés anonymisées	Proposé	Moyen	Faible	Faible	Moyen	Important
FEA11. Ajout de champs supplémentaire à des résultats de recherche	Proposé	Moyen	Élevé	Élevé	Faible	Critique
FEA12. Trier les résultats de recherche	Proposé	Moyen	Moyen	Faible	Moyen	Utile

9 Autres exigences du produit

9.1 Exigences du système

ES1. Réquis minimaux de l'ordinateur exécutant la recherche :
- Fureteur Firefox 8 ou équivalent.
- 2 GB de mémoire vive (RAM).
- Suite de bureautique Office installée (optionnel).
- Logiciel(s) d'analyse(s) statistique(s) (optionnel).

9.2 Exigences de performance

? ? ? SVP. Précisez si cette section s'applique.

10 Exigences de documentation

10.1 Manuel de l'utilisateur

EMU1. Un manuel d'utilisateur simple et complet, permettra à l'utilisateur d'apprendre rapidement le fonctionnement de l'interface de création de requêtes.

10.2 Guides d'installation, de configuration, et fichier à lire

EGCF1. Documenter les configurations requises pour installer et configurer une machine supplémentaire pour le système EndoMine.

EGCF2. Documenter les configurations requises pour ajouter une nouvelle source de données ou table supplémentaire (si le temps le permet).

11 Bibliographie

[1] *Access 2010 specifications*. Microsoft. Consulté le 3 octobre 2012, à `http://office.microsoft.com/en-us/access-help/access-2010-specifications-HA010341462.aspx`

[2] *ArchSummary Query*. 10 fév. 2010. Jessica Iannone. Consulté le 3 octobre 2012, à `http://jessica-iannone.com/img/interactive/ArchSummQry.jpg`

Microsoft Access Query Builder

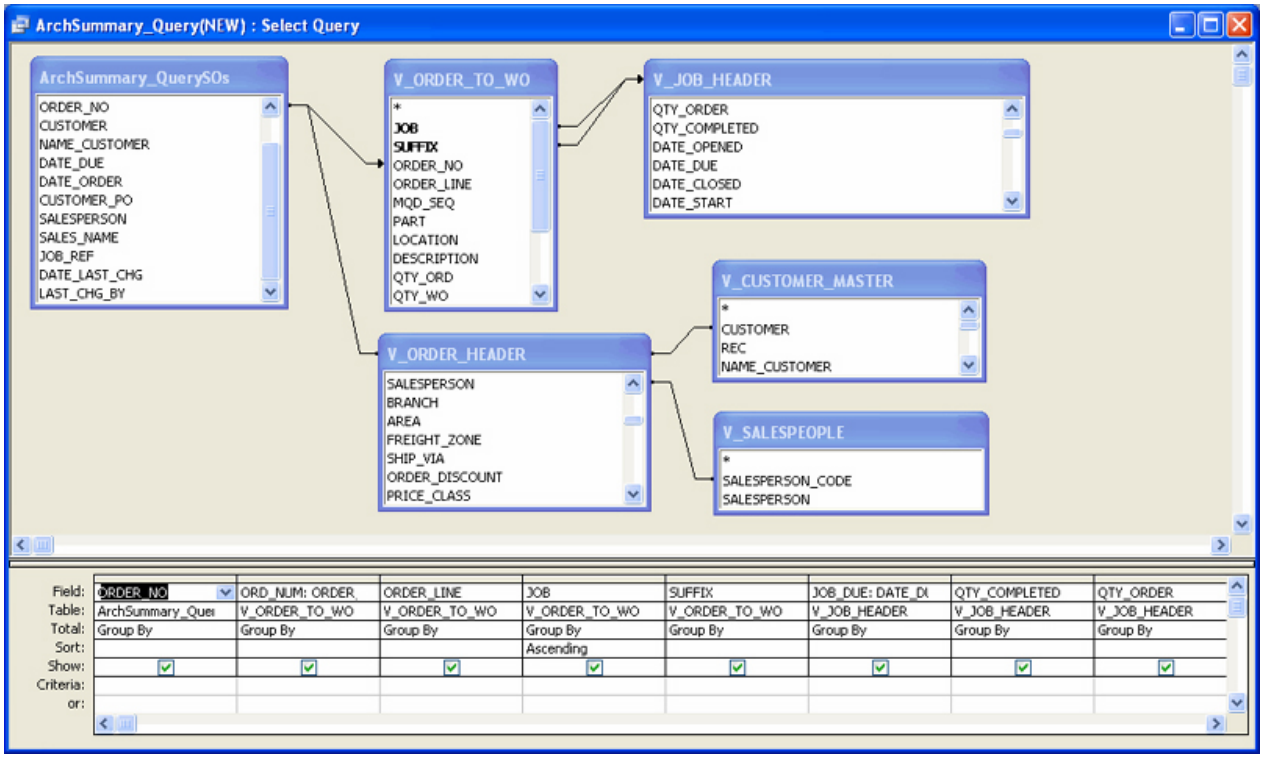


FIGURE 3 – Exemple d'une requête construite avec le Microsoft Access Query Builder (image tirée d'un siteweb : [2])

Attributs des caractéristiques

Table 8 – Légende : État des caractéristiques

Proposé	La caractéristique est proposée, mais n'a pas encore été approuvée par les parties prenantes.
Approuvé	La caractéristique est approuvée par les parties prenantes.
Incorporé	La caractéristique est incluse dans le produit.

Table 9 – Légende : Bénéfice des caractéristiques

Faible	La caractéristique apporte peu de valeur ajoutée au produit et n'est pas nécessaire à son bon fonctionnement.
Moyen	La caractéristique apporte une valeur ajoutée additionnelle au produit, mais n'est pas critique à son bon fonctionnement.
Élevé	La caractéristique apporte une valeur ajoutée importante au produit et est essentielle à son bon fonctionnement ou à la réalisation de ses tâches.

Table 10 – Légende : Effort des caractéristiques

Faible	La réalisation de la caractéristique nécessite un effort de moins de 20 heures-personnes.
Moyen	La réalisation de la caractéristique nécessite un effort entre 20 et 40 heures-personnes.
Élevé	La réalisation de la caractéristique nécessite un effort de plus de 40 heures-personnes.

Table 11 – Légende : Risque des caractéristiques

Faible	La technologie utilisée et la méthode d'implémentation sont connues et bien maîtrisées.
Moyen	La technologie utilisée est récente ou la méthode d'implémentation nécessite une attention particulière.
Élevé	La technologie utilisée est nouvelle et peu éprouvée ou la méthode d'implémentation est complexe et demande une analyse plus complète.

Table 12 – Légende : Stabilité des caractéristiques

Faible	Les exigences concernant la caractéristique ont de fortes chances de changer ou le bon fonctionnement de la caractéristique a un impact critique sur le fonctionnement général du système et peut compromettre son exécution.
Moyen	Les exigences concernant la caractéristique sont susceptibles de changer ou le bon fonctionnement de la caractéristique a un impact sur le fonctionnement général du système sans toutefois compromettre son exécution.
Élevé	Les exigences concernant la caractéristique ont peu de chance de changer et le bon fonctionnement de la caractéristique n'a pas d'impact sur le fonctionnement général du système.

Table 14 – Légende : Priorité des caractéristiques

Utile	La caractéristique apporte des fonctionnalités accessoires au système. Son inclusion dans le produit a peu d'impact sur la satisfaction du client et sur l'utilisation du système.
Important	La caractéristique apporte des fonctionnalités supplémentaires au système. Son inclusion dans le produit peut influencer la satisfaction du client, mais son absence n'empêche pas l'utilisation du système.
Critique	La caractéristique est primordiale au fonctionnement du système. Il est nécessaire de l'inclure en priorité dans le produit pour assurer la totale satisfaction du client et son absence pourrait empêcher l'utilisation du système.