

MULTI-SYSTEM OPTIMISATION

Nicole.Glassey@hevs.ch, Anne.LeCalve@hevs.ch, Alexandre.Cotting@hevs.ch, professeurs, Zhan.Liu@hevs.ch, doctorant, Fabian.Cretton@hevs.ch et Benjamin.Nanchen@hevs.ch adjoint scientifique à l'Institut Informatique de Gestion de la HES-SO Valais-Wallis & Fabrice.Chapuis@he-arc.ch, professeur HEG Arc

How to optimize your choice of data models?

Comment optimiser le choix de vos modèles de données ?

Smartphones, Cloud ou encore Big Data ont contribué à démultiplier la quantité de données ainsi que leur complexité. Les approches traditionnelles de gestion de données (stockage et traitement dans des bases de données relationnelles pour diffusion sur une plate-forme logicielle et/ou Web) ne sont pas toujours optimales et requièrent d'autres formes de modélisation et de diffusion de l'information.

À présent, l'évolution des données doit être appréhendée de manière globale depuis leur capture jusqu'à leur exploitation en passant par leur stockage. De nouveaux modèles de structuration des données apparaissent avec leurs spécificités propres – leurs points forts et faiblesses, leurs niches d'utilisation – pouvant apporter de gros avantages aux SI. Ces différents modèles n'ayant pour le moment encore peu été utilisés de concert, il nous apparaît de plus en plus évident qu'une utilisation combinée permettrait de couvrir de façon pertinente l'ensemble du processus de la récolte à la diffusion sur tout type de canaux. Afin de réussir ce pari, il est primordial que les choix technologiques soient pilotés par les données, les traitements nécessaires et le type de diffusion attendu.

Le cœur de MUSYOP (MULTI-SYSTEM OPTIMISATION), un projet de l'Institut Informatique de gestion de la HES-SO Valais-Wallis en collaboration avec la HEG Arc, consiste justement à étudier comment mixer plusieurs modèles de structuration des données. Il s'agit de développer une architecture multisystèmes intégrant différentes technologies de bases de données afin de stocker,

traiter et exploiter de façon optimale des données hétérogènes (structurées, non structurées ou faiblement structurées). Pour ce faire, MUSYOP propose d'aborder des réflexions telles que: est-ce qu'une analyse des besoins et des données est faite avant de considérer les technologies de structuration des données? Le recul nécessaire, tant par rapport à la complexité des données (granularité, besoins variables en capacité hardware pour les traiter ou y accéder pendant leur durée de vie, ...) que par rapport aux technologies existantes pouvant répondre à ces besoins, est-il vraiment pris? Les futurs accroissements en quantité de données sont-ils pris en compte dès le début?

D'un modèle unique à une solution multi-système

Les systèmes d'information utilisent en général un ou deux types de modèles de données. Cela n'est bien souvent pas suffisant pour trouver des solutions innovantes permettant de répondre aux besoins des entreprises. Les nouveaux modèles de structuration des données comme NoSQL, XML ou sémantique ont chacun leurs caractéristiques utiles aux SI. Par exemple, le NoSQL apparaît comme une alternative au relationnel pour gérer des grands volumes de données dont la performance en lecture et écriture est la priorité, au détriment des propriétés ACID (Atomicity, Consistency, Isolation, Durability) des bases de données relationnelles. Le NoSQL propose plusieurs modèles de stockage permettant de s'adapter finement au type de données à traiter (modèle orienté colonne, document, clé-valeur, graphe).

L'utilisation d'un seul de ces modèles de structuration de données ne peut cependant pas répondre à l'ensemble des besoins d'une application métier et couvrir de façon pertinente l'ensemble du processus de la récolte à la diffusion sur tout type de canaux. Il nous apparaît donc de plus en plus évident qu'une utilisation

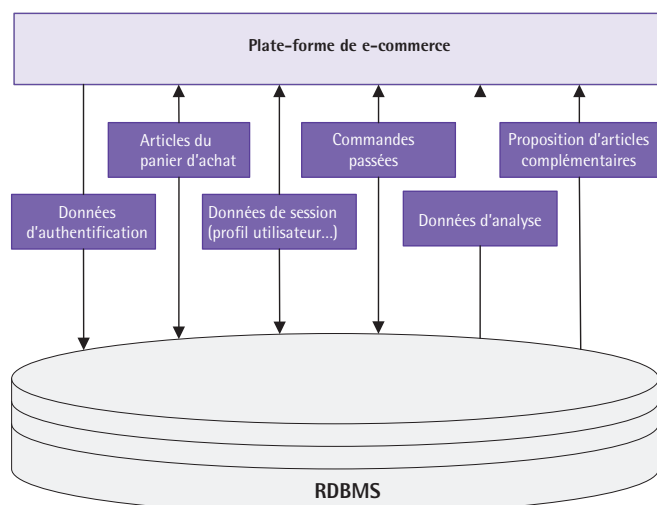
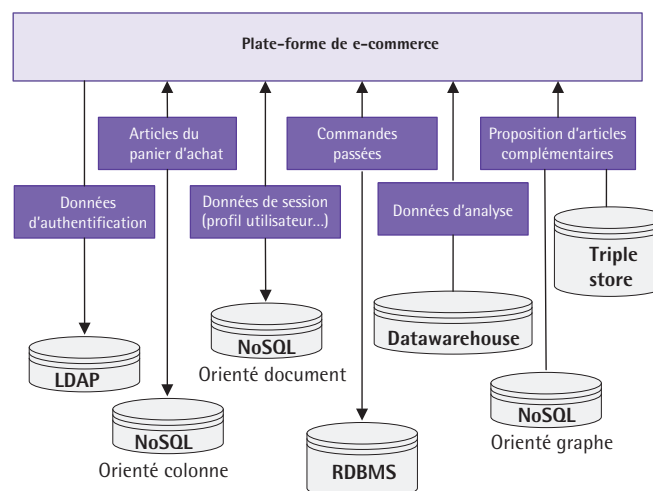


fig. 1 – d'un datastore unique à une approche multisystème



combinée (relationnel, NoSQL, XML ou sémantique) pourrait être une solution.

Prenons l'exemple d'une plate-forme de e-commerce et considérons les données utiles à ce métier comme l'authentification, le profil de l'utilisateur ou son panier d'achats, nous remarquons que les besoins en terme de consistance, de performance, de disponibilité, d'analyse et de stratégie de sauvegarde ne sont pas les mêmes [1].

Afin de résoudre cette problématique, nous proposons une approche multisystème où chaque *datastore* implémente un modèle de stockage spécifique pour traiter un type particulier de données et capable de stocker des données utilisées par plusieurs applications (fig. 1).

Les modèles au service des données

Mettre en place une architecture multisystème au service des données nécessite de choisir les modèles adéquats. Le projet MUYOP identifie les critères entrant en ligne de compte lors du choix d'une technologie. Les capacités techniques d'un modèle n'en sont qu'une des composantes, le plus important étant de comprendre le contexte du projet. Ceci permet de ressortir les critères les plus déterminants pour la décision finale. Cependant, un choix peut s'orienter vers une technologie *a priori* moins adaptée pour la représentation des données, mais répondant plus efficacement à d'autres critères comme le volume de données, la sécurité ou encore les possibilités de requêtage.

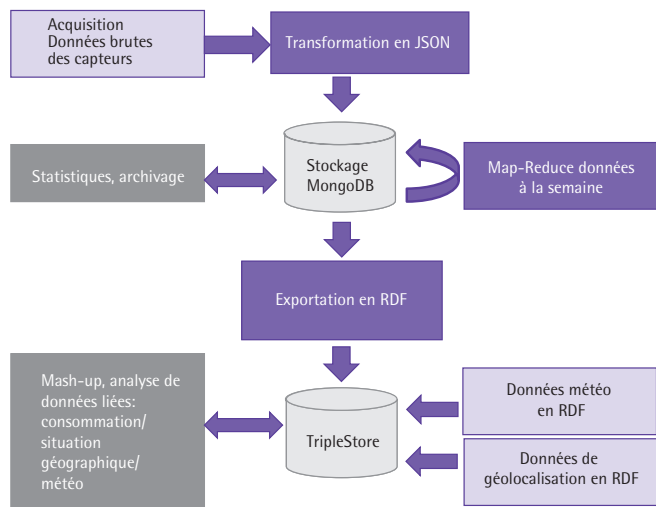


fig. 2 – Data Flow de données énergétiques

Le domaine de l'énergie illustre particulièrement bien cette problématique. Des capteurs relèvent fréquemment (heure, minute ou même seconde) des données énergétiques telles que consommation, production et stockage. Ceci implique donc une contrainte de performance en termes d'écriture et capacité de stockage lors de l'acquisition de ces informations. Une approche NoSQL-MongoDB (JSON) peut être recommandée. L'exploitation de ces données brutes est déjà possible, mais leur transformation (par exemple, journalisation ou mensualisation par un algorithme de *map-reduce*) est également envisageable afin que celles-ci répondent aux besoins d'autres applications. Par exemple, une exportation en RDF dans un triple store permet de les lier avec

d'autres informations telles que des données de géolocalisation ou des données météorologiques. Ceci permet de nouvelles possibilités d'analyse et de requêtage.

Une méthodologie simple pour choisir ses modèles

Afin d'accompagner les entreprises dans leur démarche pour déterminer les modèles adéquats, nous proposons une méthodologie simple associant plusieurs outils, parmi lesquels un questionnaire élaboré permettant d'appréhender le contexte du projet, une matrice conceptuelle et une matrice chromatique aidant à pondérer les forces et les faiblesses pour chaque modèle en fonction d'un ensemble de critères préalablement déterminés.

Le questionnaire ouvert permet d'identifier les points clés pour les choix technologiques. Premièrement, il s'agit de déterminer les connaissances techniques de nos interlocuteurs ainsi que les technologies utilisées au sein de l'entreprise. Puis, les particularités du projet (contraintes, impératifs de sécurité, mise en production) sont étudiées. Dans un troisième temps, toutes les caractéristiques des données sont prises en compte (type, acquisition, mise à jour, traitement, représentation, volume, liaisons, schéma, requêtage, stockage, évolution prévue, accès). Finalement, une discussion ouverte sur les expériences de nos interlocuteurs nous permet de cerner au mieux leurs besoins. Cette étape permet d'obtenir une représentation schématique du besoin client (architecture haut-niveau et data-flow des services). Elle sert également à définir une *Vision Box* et un *Product Backlog* (thèmes/services).

Une matrice conceptuelle des données (fig. 3) s'intéressant plus particulièrement à la nature des informations traitées par l'application a été pensée. Celle-ci fait suite au questionnaire proposé au client et a pour but de regrouper dans différentes catégories l'ensemble des données traitées par l'application.

La matrice chromatique réalisée représente les différents critères de choix par rapport aux modèles de données. Elle permet d'indiquer la *force* de chaque modèle pour chacun des critères (fig. 4). Cinq catégories principales (données, schéma, requête, système et sécurité) permettent de comparer quatre modèles (NoSQL, RDBMs, XML et sémantique). La matrice chromatique prend en compte des critères liés à la performance et à la disponibilité du système, à la volumétrie, aux différents types de requêtes pouvant s'exécuter sur le système. En confrontant celle-ci avec les données récoltées lors des entretiens, il est possible d'effectuer des choix en vue de l'élaboration de l'architecture multisystèmes.

En sus de ces trois outils notre méthodologie a permis également d'explorer d'autres pistes de réflexion telles que:

- une matrice table de décision (dans laquelle les couples conditions ou critères de choix, actions ou choix du modèle, deviennent les règles de la table de décision),
- ainsi qu'un raisonneur OWL DL (intégrant une ontologie basée sur la logique de description pouvant jouer un rôle similaire à un système expert).

Une plate-forme multisystème !

Cette première phase d'analyse nous conforte dans l'idée que l'utilisation simultanée de plusieurs modèles est une piste des plus

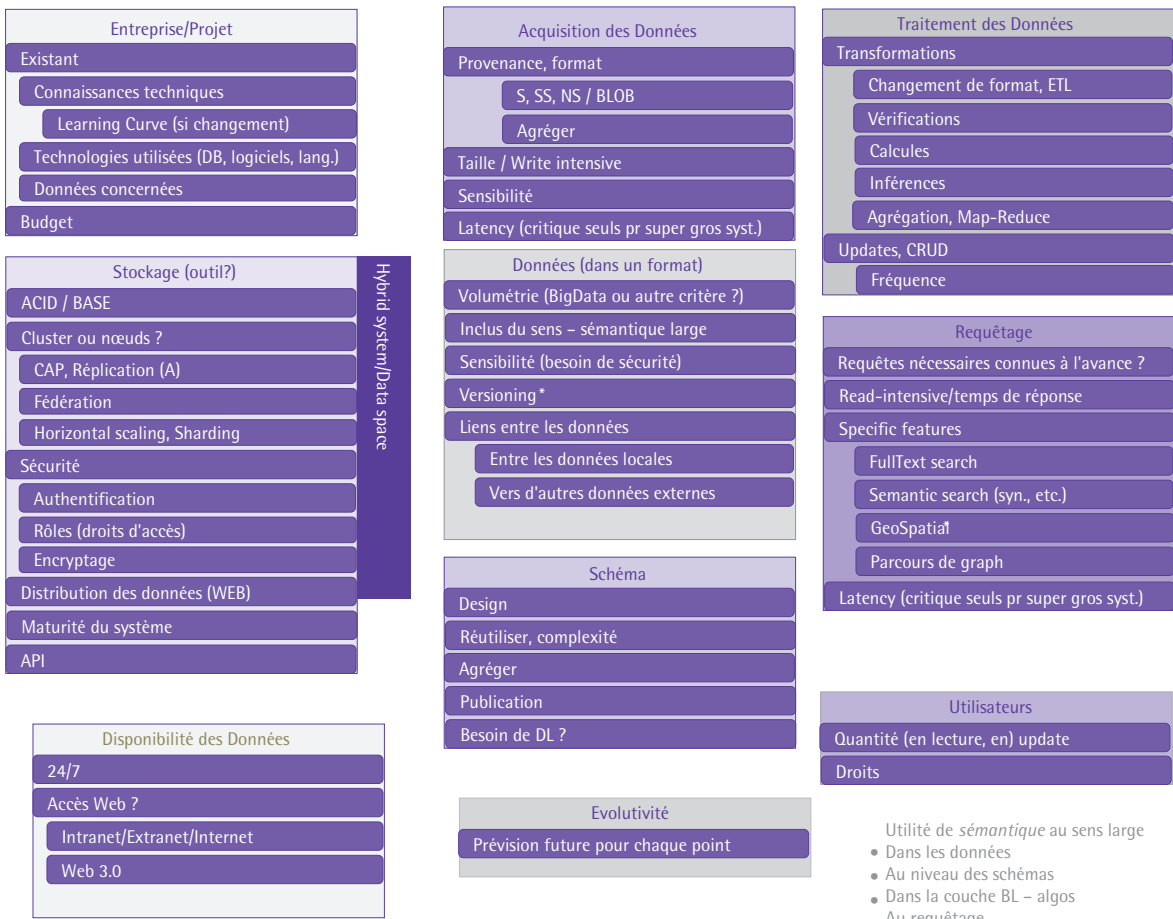


fig. 3 – matrice conceptuelle des données

Caractéristiques / Critères			Modèles						
			Clé-valeur	Orienté Colonne	Orienté Document	Orienté Graphe	RDBMs	XML	Web sémantique
Données	Nature	Données structurées							
		Données semi-structurées							
		Données non-structurées							
	Relation	Cardinalité (1,*)							
		Relations locales (intégrité référentielle)							
		Relations externes vers d'autres sources (URL)							
	Versionning								
Schéma	Avec / Sans	< 100G							
		Big Data							
		Schema-based							
Requête	Performance	Schema-less							
		Troughput (requêtes par h/m/s)							
	Transactionnel	ACID							
		Non ACID							
	Analyse	Requêtes analytiques, cube							
Système	Disponibilité	Map-reduce							
		Réplication							
	Cluster	Partitions (sharding)							
Sécurité	Authentification								
	Droits d'accès (rôles)								
	Encryption								

fig. 4 – matrice chromatique

intéressantes. L'objectif principal du projet MUSYOP est de créer un framework qui combine différentes bases de données, tels que bases de données relationnelles, NoSQL, sémantique et XML. La suite du projet va nous permettre de monter et tester une plateforme multisystème intégrant ces modèles. Il s'agira plus particulièrement de surmonter des défis techniques spécifiques à cette architecture comme la cohabitation de SGDB différents, leur communication tout en respectant la structure et le flot des données. Les données pourront alors transiter d'un système à l'autre. Afin d'avoir une vue uniforme des données, le prochain défi reposera donc sur la mise en place d'un module de requêtage multisystème unifié.

Référence

[1] SADALAGE Pramod. J. & FOWLER Martin. NoSQL Distilled. 2012, pp. 136-137. ■