COMMENT CHOISIR SA SOLUTION DECISIONNELLE

Partie 1 : Acquisition des données

Microsoft

GUSS

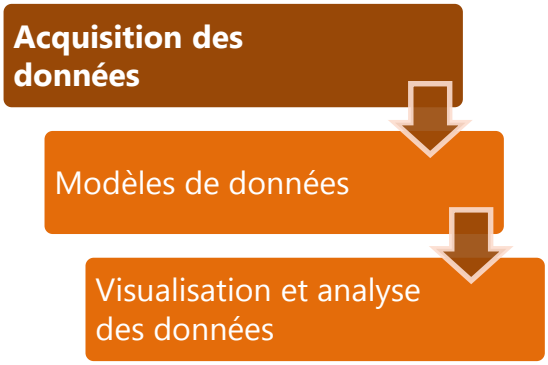
Copyright

Table des matières

Préambule

Un livre blanc en 3 parties

L'ambition de ce livre blanc est de traiter de l'ensemble des aspects d'une solution décisionnelle autour des outils ad hoc de l'offre Microsoft.



Il est découpé en 3 parties : l'acquisition des données, les modèles de données et enfin la visualisation et l'analyse des données.

Cette première partie du livre blanc est intitulée « acquisition des données » : Elle s'attachera à aborder toutes les problématiques liées à la récupération des données.

Audience

Ce livre blanc s'adresse particulièrement à des chefs ou directeurs de projets, architectes et responsables informatiques, ou des décideurs métiers qui souhaitent mettre en œuvre une solution décisionnelle.

Ce livre est destiné à un public souhaitant avoir une vision complète des briques constituantes de la plate-forme décisionnelle de Microsoft. Si vous envisagez de démarrer ou de transformer un projet de Business Intelligence (BI), ce livre blanc peut vous aiguiller dans vos choix d'outils et d'architecture.

Comment lire ce livre blanc

Ce livre ne se veut ni un cours, ni une formation sur le décisionnel, ni un guide d'implémentation pratique. Il décrit les éléments constituants une solution décisionnelle, qu'ils soient obligatoires ou optionnels, les usages associés et des pistes pour leur mise en œuvre.

Ce document a été rédigé de manière collaborative par des experts des technologies Microsoft, spécialistes des solutions décisionnelles, appartenant à différentes sociétés de conseil indépendantes de l'éditeur. Les rédacteurs sont des consultants expérimentés qui implémentent, conseillent et auditent quotidiennement des projets BI.

Toutefois, même si les sujets ont été débattus entre les rédacteurs, il subsiste forcément un biais propre à l'expérience de chaque auteur. Comme avec tout contenu éditorial, le lecteur sera juge et se fera son propre avis sur la base des éléments qui lui seront apportés par les auteurs.

Nota : l'actualité autour de la plate-forme décisionnelle chez Microsoft évolue très vite, les informations présentées ici correspondent à la situation au **30 septembre 2014.**

Les auteurs

Jean-Pierre Riehl

Responsable Data & Business Intelligence chez AZEO. Architecte, consultant, expert, chef de projet, développeur, formateur, manager, MVP, leader du GUSS mais surtout passionné par les données et la Self-Service BI.

http://blog.djeepy1.net @djeepy1

Romain Casteres

Consultant en informatique décisionnelle et en Big Data chez Dcube. Romain est MVP SQL Server, MCSE Data Platform et Business Intelligence.

http://pulsweb.fr @PulsWeb

Philippe Geiger

Consultant certifié (MCITP, MCSE), formateur certifié (MCT) et speaker (JSS, SQLSat, Techdays), Philippe travaille actuellement chez Neos-SDI en Alsace où il accompagne aussi les professionnels de l'IT que les développeurs, mais également les utilisateurs de la BI.

http://blog.pgeiger.net/ @PGeiger

Arnaud Voisin

Consultant expert pour le compte de WAISSO, spécialisalisé sur la partie décisionnelle, certifié (MCITP, MCSE), formateur certifié (MCT), Arnaud intervient aussi bien sur la partie audit, que conseil ou sur la réalisation.

http://arnaudvoisin.blogspot.fr @ArnaudVoisinSQL

La communauté SQL Server, ce sont tous les acteurs qui travaillent de près ou de loin avec SQL Server, qu'ils soient développeurs, formateurs, architectes, consultants, DBA, etc.

Le GUSS est une association loi de 1901, dirigée par le Board, composé de 9 personnes qui sont des professionnels reconnus sur SQL Server. Le GUSS fédère la communauté autour d'échanges réguliers et rassemble tous ceux qui souhaitent apprendre, partager ou tout simplement échanger sur SQL Server.

http://guss.pro @GUSS_FRANCE

C'est à ce titre que nous avons rédigé pour vous ce livre blanc.

Introduction

De plus en plus de sources de données

Tous les medias, spécialisés ou non, le répètent depuis un moment : *il y a de plus en plus de données*. Elles ne proviennent plus d'une unique application monolithique mais de sources hétérogènes : données d'entreprise, données temps réel, données sociales, données personnelles, données non-structurées, etc.

Elles arrivent également de plus en plus rapidement et dans des volumes toujours plus importants. À l'heure du Big Data et des Marketplace, les données proviennent de partout et couvrent tous les métiers.

Une des conséquences de cet état de fait est un foisonnement de formats et de protocoles. Même si des protocoles comme OLEDB ou ODBC facilitent l'interaction, on doit conjuguer avec les HTTP, JSON, REST, ODATA, HIVE, etc. La connectivité ne doit plus être un frein, on parle de plus en plus d'informatique « sans couture » (seamless).

Les utilisateurs partent d'un postulat simple : « *si je vois la donnée, je peux la récupérer et la travailler*». Ils exigent des outils qui répondent à leur attente en leur facilitant le travail.

Faire de la Business Intelligence aujourd'hui, c'est analyser des données, **d'où qu'elles viennent**, **quel que soit leur format**, **leur taille** et sans faire de compromis sur l'outil de restitution, et tout cela dans des cycles de réalisation toujours plus courts.

« Any Data, anywhere, any device »

« Any Data, anywhere, any device » (toutes les données, n'importe où, sur n'importe quel terminal), c'est le leitmotiv de Microsoft depuis quelques années. Les frontières disparaissent : Cloud, On-Premise, hybride, mobile, Big Data, Data Mining, analyse prédictive, self-service, massivement parallèle, les outils de la plate-forme doivent répondre **aux nouveaux besoins et usages des utilisateurs** de façon intégrée et homogène.

« Cloud-first, Mobile First »

En 2014, **le Cloud** est aujourd'hui une réalité omniprésente et toutes les entreprises le considèrent. Et la question qui se pose n'est pas « Cloud ou On-Premises[1] ». Aujourd'hui, le questionnement est de savoir comment j'enrichis et j'améliore mon système d'information avec les possibilités offertes par Cloud, souvent en construisant des **solutions hybrides**.

Microsoft est considéré comme un leader dans le domaine que ce soit sur l'offre IaaS, PaaS ou SaaS [2] (voir En savoir plus). C'est le fruit de son innovation (lancement de l'offre

[1] Système d'Information géré localement par l'entreprise, au sein de son Data Center ou hébergé chez un prestataire.

[2] Infrastructure as a Service, Platform as a Service, Software as a Service.

en 2008) mais également de sa philosophie globale Cloud-First. Et quand on parle de solution décisionnelle ou plus généralement de données, l'offre Cloud est même plus large que sur celle on-premises, notamment avec Power BI pour Office 365 ou la base de données en mode PaaS (SQL Azure).

L'autre principe que Microsoft met en avant, c'est **l'abolition des frontières entre les terminaux** : tablettes, téléphones, ordinateurs, laptop, phablettes, etc. Les outils doivent s'adapter au terminal et non l'inverse. On commence son travail d'analyse sur son poste de travail de bureau, on le présente sur une tablette ou un tableau tactile en réunion, on consulte ses KPI dans les transports. La **« mobilité »** est devenue une vraie composante du travail moderne.

Mode de consommation : Intégration ou Self-Service

Enfin, un autre paradigme qui est celui du **mode de « consommation » des données**. Traditionnellement, on dispose d'un cahier des charges, de spécifications, et une équipe projet « développe » les flux et à la fin vient la phase de recette des données. Cela prend du temps alors que l'industrie réclame du **Time-to-Market**. Ainsi apparaît de plus en plus la notion de **Self-Service** où l'utilisateur est placé au cœur de ses propres données.

Microsoft a su proposer une gamme d'outils, centrée autour d'Excel, permettant aux utilisateurs de gagner en autonomie. Mais cela ne remet pas en cause l'intégration de données traditionnelles, industrialisées et basées sur le DatawareHouse d'Entreprise. C'est seulement un élargissement de la **boîte à outils autour des données**.

Microsoft et la « culture de la donnée »

Le 15 avril 2014, lors d'une conférence exceptionnelle, les dirigeants de Microsoft ont partagé leur point de vue sur les données dans l'entreprise.

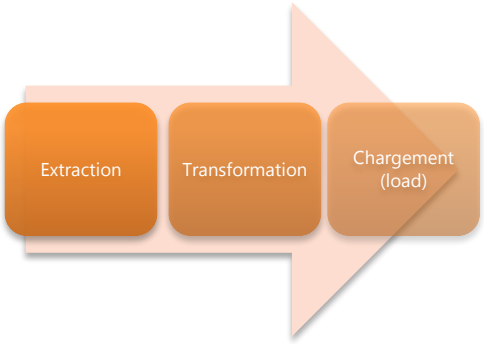
Satya Nadella, PDG de Microsoft, a exposé la vision d'une plate-forme construite pour une ère où l'intelligence est partout. Il a souligné l'importance d'une « Culture de la Donnée » encourageant la curiosité, l'action et l'expérimentation par l'ensemble des acteurs de l'entreprise. Pour Microsoft, la donnée doit être au cœur des réflexions quotidiennes car, bien exploitée, elle est créatrice de valeur. L'orientation de Microsoft est de mettre en place des solutions technologiques mettant les données à la portée de tous !

Pour visualiser la conférence : http://bit.ly/1oCFmnd.

EIM ou la gestion de l'information de l'entreprise

La gestion de l'information d'entreprise (EIM pour Enterprise Information Management) fournit un éventail de solutions qui permettent aux organisations d'acquérir, d'évaluer la crédibilité et de s'assurer de la cohérence de leurs données.

Les **outils de type ETL**[3] constituent la colonne vertébrale de l'acquisition de données. L'ETL est l'acronyme d'Extract-Transform-Load.



L'ETL est donc une méthode généralement admise pour assurer l'acquisition des données dans un environnement à usage décisionnelle :

1. L'extraction : il s'agit **d'acquérir les données où qu'elles soient et quel que soit leur format** : Les outils exécutant les tâches d''extraction doivent être capable de disposer de protocoles et de format de données le plus exhaustif possible.
2. La transformation : À partir des données brutes obtenues à l'étape d'extraction, cette étape est de **transformer les données pour les rendre ensuite exploitables** en fonction des usages attendues dans les étapes suivantes de la plate-forme décisionnelles (étape *Modèles d'analyse et Visualisation* abordées dans les parties suivantes du livre blanc). Parmi les transformations habituelles, il est possible de citer la suppression des données incohérentes, la mise en forme des données, la concaténation, l'agrégation des données issus de plusieurs sources, la recherche de clés (au sens contraintes d'intégrité des bases de données).
3. Le chargement : Une fois les données extraites et transformées, elles doivent être conservées pour être utilisées ultérieurement (base de données, entrepôt de données ou toute autre démarche).

En complément des outils de type ETL, il en existe deux autres :

1. **Outil de nettoyage des données** : La qualité de la donnée issue d'système d'information quelconque est primordiale pour un système décisionnel : une mauvaise qualité de données ne permettra pas une prise de décision pertinente.
2. MDM (ou Master Data Management) : Sous le concept de « **gestion des données de référence** », il s'agit pour l'entreprise de bénéficier d'un référentiel qui contient toutes les données importantes de l'entreprise.

Ainsi, à travers la gestion des informations d'entreprise, il s'agit d'accéder à toutes les données de l'entreprise en s'assurant que **ces données sont disponibles, cohérentes et de qualité**.

En complément de ces outils, une fonction dans l'entreprise prend toute son importance il s'agit du **data steward** qui sera décrit plus loin dans ce document.

[3] Il existe des variations autour de la notion d'ETL, comme ELT où la démarche est légèrement différente : l'objectif est de stocker le plus de données possibles issues de l'extraction avec le minimum de modification, la transformation a lieu plus tard, au moment de son utilisation. Dans ce cas, l'idée est donc de conserver le plus de données possibles dans l'idée que leurs usages futurs ne sont pas encore connus.

Les outils disponibles

SSIS, l'acquisition des données d'entreprise

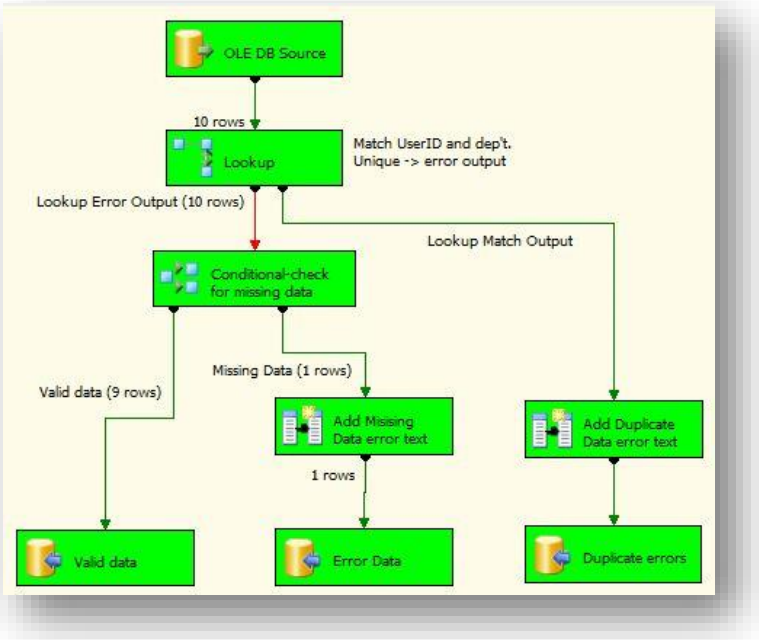
SQL Server Integration Services (SSIS) est l'outil d'acquisition de **données d'entreprises** de Microsoft. Livré avec SQL Server, il est composé d'un moteur de chargement et un environnement de développement basé sur Visual Studio (SSDT[4], anciennement BIDS[5]).

SSIS, c'est avant tout un ETL puissant, capable d'intégrer de très haute volumétrie de manière performante. Avec la version 2008, il était déjà possible d'intégrer 1To octet de donnée en moins de 30 minutes (cf : http://technet.microsoft.com/en-us/library/dd537533(v=sql.100).aspx)

Cet ETL ne doit pas être confondu avec un EAI comme Biztalk ou comme les Messages Broker (MSMQ) qui fonctionnent en mode continu avec les messages comme unité de travail, SSIS est, le plus souvent, lancé selon une planification (par exemple une fois par jour) et fonctionne sur des volumes de données pouvant être importants.

SSIS est d'abord un outil de workflow. Ce workflow est orchestré par le flux de contrôle (« Control Flow ») dont le rôle premier est de préparer le flux de données (récupération de fichier sur FTP, déplacement de fichier, …), mais aussi de déclencher un ensemble d'actions post-intégration (envoi de mail par exemple) et de gérer les dépendances entre les différentes tâches.

Les deux principales tâches d'intégration de données sont la tâche SQL et la tâche de flux de données (« Data Flow »). Cette dernière est une tâche qui permet de traiter des données dans un buffer, principalement en mémoire. Le flux de données fonctionne selon 3 étapes : acquisition des données sources, transformation, intégration dans la destination.



Les opérations de transformation les plus courantes du flux de données sont le tri, l'agrégation, la jointure (qui peuvent être faite dès la phase d'acquisition) et le lookup.

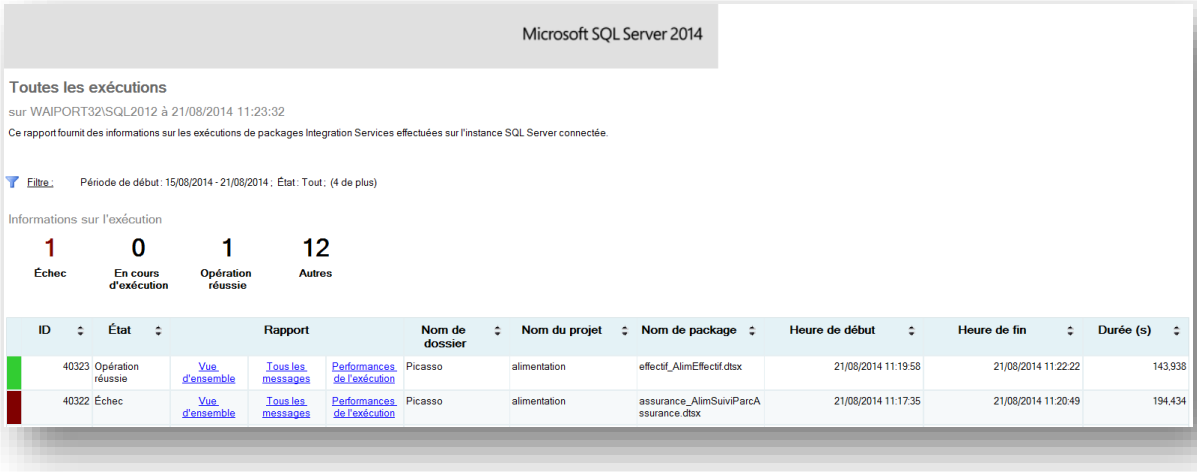
Par le biais du flux de données, il est possible de traiter les données en lots ou en mode ligne (dans ce dernier mode, le traitement est nettement moins performant).

[4] SQL Server Data Tools
[5] Business Intelligence Development Studio

Les principales fonctionnalités de SSIS

Pour rappel, l'objectif de SSIS, c'est d'automatiser et d'**industrialiser** l'échange de données. Les fonctions principales couvertes par l'outil sont :

- **La connexion de sources hétérogènes** : SSIS offre une solution de consolidation entre des sources diverses et variées grâce à des providers intégrés (ADO, OLEDB, CSV, XML, …).
- **Le chargement d'entrepôt de données** : SSIS intègre des composants permettant d'optimiser le chargement de votre entrepôt. Le CDC (Change Data Capture) permet de ne récupérer que les données modifiées. Ce composant est idéal dans le cas du chargement incrémental. Le SCD (Slowly Changing Dimension ou Chargement de dimension à variation lente) permet de charger des dimensions de type 1, 2 ou 3 (classification Kimball).
- **De garantir la cohérence de données et la fiabilité** : L'utilisation combinée des transactions et des checkpoints permettent à la fois de modifier les données qu'en cas de succès total et de reprendre le package à un point donné afin d'éviter de recommencer une série de traitements déjà effectués.
- **De faciliter l'exploitation et la maintenance** : Couplé avec l'Agent SQL, les packages SSIS héritent de toutes les fonctionnalités de planification de son exécution (ordonnancement, alertes, notifications, …) Des fonctionnalités sont prévues pour l'analyse et la compréhension d'erreur. Les Data Tap permettent de visualiser des données à un endroit du flux pendant l'exécution en production. De plus, un reporting détaillé de suivi d'exécution est fourni.



- **La gestion des rejets et erreurs :** Une des fonctionnalités les plus importantes d'un outil comme SSIS est de pouvoir rediriger des lignes en erreurs selon des règles de gestion ou bien des vérifications de contraintes. Un mécanisme de recyclage des rejets peut être mis en place pour augmenter l'intégration des données.

Les nouveautés apportées par la version 2012

La version 2012 a apporté bon nombre d'innovations et améliorations à SSIS, notamment

par l'introduction du mode projet et du catalogue SSIS associé. Les principales améliorations sont :

- **La journalisation** : Prête à l'emploi avec un lot de reporting faisant état des statuts d'exécution, de leur contexte et même d'informations plus fines au niveau des flux de données.
- **Le paramétrage simplifié** : Deux niveaux supplémentaires de paramétrage ont été apportés, le niveau package et le niveau projet. Ceci permet de séparer ce qui est variable de ce qui est paramétrable et de gagner en maintenabilité et lisibilité.
- **Le versionnement** : Le mode projet apporte un nouveau grain de déploiement, le projet. Ce dernier permet d'historiser les versions des projets et des packages qui les constituent. Ceci permet de faire un retour arrière en cas de bug après une mise en production par exemple.
- **La contextualisation de l'exécution** : Dans le catalogue SSIS, les projets, rangés dans des dossiers, peuvent être associés à des environnements permettant de changer le paramétrage et de contextualiser l'exécution. Cela peut être très utile pour modifier le comportement d'un jeu de packages.

Aller plus loin avec SSIS

SSIS propose de base un certain nombre de fonctionnalités. Il existe cependant différents moyens d'étendre celles-ci grâce notamment à des API qui permettent de manipuler SSIS par programmation, aussi bien pour les développeurs que les administrateurs.

Il est possible de développer ses propres composants SSIS : tâche du Control Flow, connexion, source, transformation ou encore destination. Il existe des composants tout prêts, gratuits ou payants, d'éditeur ou open source. Par exemple, on peut citer des tâches de compression, des sources et destinations SharePoint, une destination XML, etc. Pour développer plus rapidement, il est possible d'utiliser les composants Scripts pour implémenter du code C# ou VB.NET directement au sein du Control Flow ou du Data Flow.

Il est également possible de générer des packages complets grâce à des API, que ce soit celle de SSIS ou via des frameworks tiers comme EzAPI.

Enfin, l'administration peut elle aussi être automatisée, soit en T-SQL via le catalogue SSIS, soit en Powershell grâce à un modèle objet. Ainsi des tâches telles que le déploiement de projet, l'export vers un autre environnement, la création d'environnement ou l'exécution de package peuvent être automatisées.

Utilisation de BIML

BIML est un méta-langage, en XML, qui permet de modéliser tous les objets d'un projet BI Microsoft. Pour SSIS il permet de définir des templates de packages réutilisables et ainsi d'obtenir des gains de productivité.

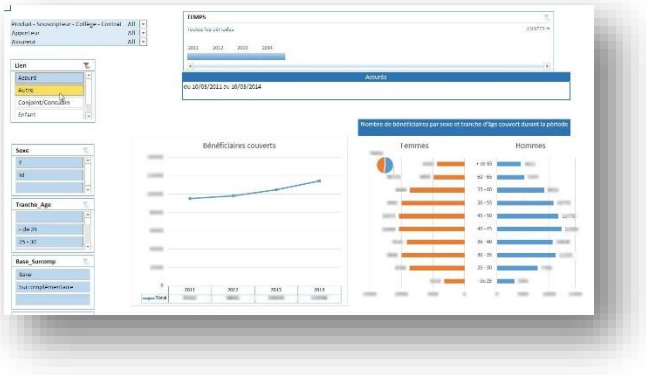
Power Query, l'acquisition de données en mode Self-Service

Dans les principes de la Self-Service BI, tout utilisateur doit pouvoir récupérer facilement et rapidement des données afin de les exploiter dans ses outils quotidiens. Acquérir des données ne doit pas être une tâche dévolue aux équipes informatiques.

L'outil le plus utilisé est sans aucun doute Excel, 1 milliard d'utilisateurs selon Microsoft. C'est l'outil de prédilection pour travailler avec des jeux de données, que l'on soit contrôleur de gestion, chef de projet, décideur, contremaître, etc. Tout le monde a un jour utilisé une feuille de calcul dans Excel pour manipuler des tableaux de données, récupérés ou créés manuellement.

Malheureusement, la façon d'injecter des données dans Excel reste souvent peu conventionnelle, à base de copier-coller et de formules de calcul comme le RECHERCHEV, pour souvent n'y faire que quelques modifications mineures.

Power Query est le complément à Excel qui permet de répondre aux besoins de récupération de données et à leur façonnage en offrant une boîte à outils riche et puissante.

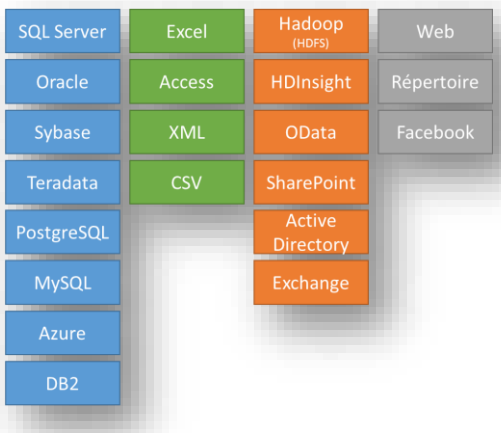


Power Query s'occupe de la plomberie entre la source des données et son exploitation que ce soit dans une simple feuille Excel ou dans un modèle Power Pivot. Le complément garde le lien vers la source et permet de rafraîchir les données en réappliquant toutes les manipulations.

De nombreux connecteurs de données

Power Query permet de se brancher à de très nombreuses sources de données toutes très différentes mais répondant aux besoins des utilisateurs.

On compte bien évidemment les classiques bases de données que l'on retrouve dans toutes les entreprises : SQL Server, SQL Azure, Oracle, DB2, MySQL, Sybase, PostgreSQL, Teradata sans oublier le Big Data avec HDFS et HDInsight.



Il y a encore la possibilité de se connecter à toutes sortes de fichiers comme des fichiers textes, Excel, Access ou encore XML. L'outil permet même de consolider un ensemble de fichiers se trouvant dans un répertoire.

On peut également récupérer des données

dans diverses sources que l'on trouve dans une entreprise, comme des listes SharePoint, un annuaire Active Directory ou encore des éléments dans la messagerie Exchange.

Enfin, Power Query permet de se connecter à des sources externes comme des flux OData[6], des données se trouvant dans des pages Web (tables HTML) ou encore des réseaux sociaux comme Facebook.

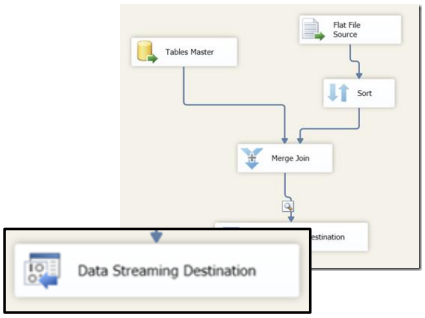
Il y a donc tous les connecteurs permettant aux utilisateurs de récupérer les données dont ils ont besoin, que ce soit des données d'équipe, des données d'entreprise ou des données externes.

Toujours de nouvelles sources de données

L'équipe qui développe Power Query met le produit à jour tous les mois, ajoutant des sources de données, des transformations et enrichissant les interfaces.
L'une des dernières sources de données est la possibilité de se connecter à des univers SAP BI et ainsi capitaliser sur des univers sémantiques existants.

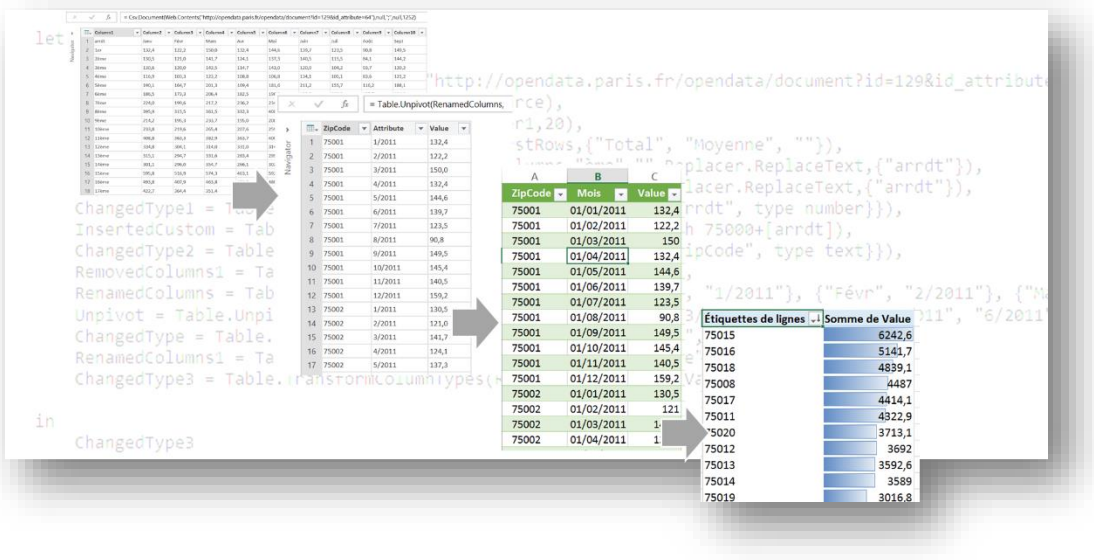
Opérations de transformations

De nombreuses transformations sont disponibles pour façonner les données pour obtenir le format de sortie attendu. Correction de valeurs, regroupement, dédoublonnage, pivot, calculs, etc., l'outil permet de manipuler les lignes et les colonnes du jeu de données.



Il est possible de mixer entre eux les jeux de données créés avec Power Query en effectuant des jointures ou des unions. Cela ouvre le champ des possibilités de façonnage du résultat et permet de croiser des sources différentes.

Toutes les opérations nécessaires à un outil d'ETL sont accessibles d'un clic dans Excel.

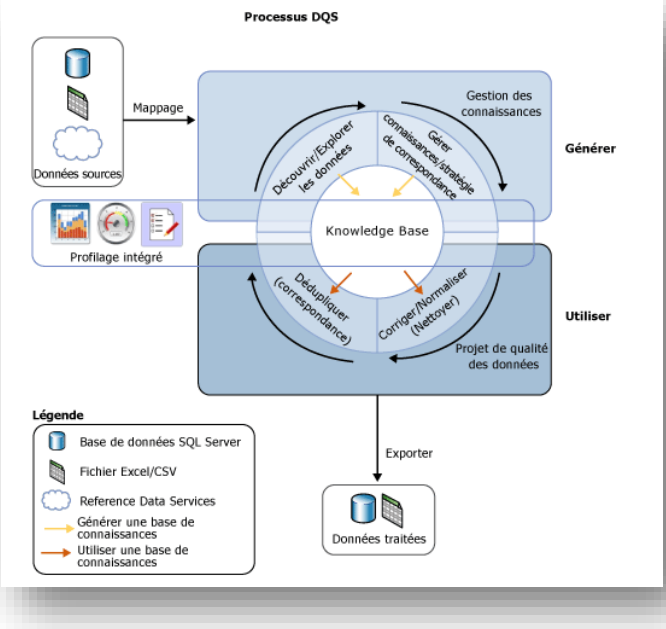


[6] OData : L'Open Data Protocol est un protocole permettant le partage de données.

L'outil permet même de développer vos propres fonctions de transformation avec un langage de programmation, le langage M.

La qualité des données

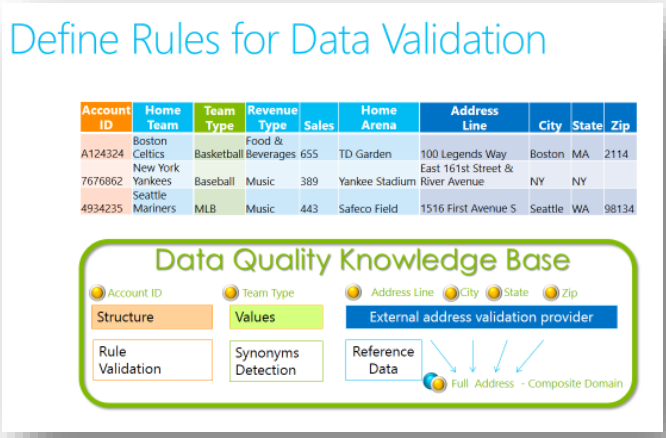
Acquérir de la donnée est une chose mais s'assurer de sa valeur est un prérequis à toute utilisation. L'un des fondements d'un système décisionnel est de fournir des chiffres corrects.



Fiabilité, pertinence, complétude, précision, exactitude sont des composantes de données « de qualité » mais souvent difficiles à évaluer. En revanche, il est aisé de dire qu'une donnée est de mauvaise qualité. C'est pour cela que le traitement de la qualité des données s'appuie sur des outils de correction dans lesquels le processus humain est central.

Microsoft propose 2 outils complémentaires pour adresser ces problématiques :

- **Data Quality Services (DQS)** : traite la qualité des données
- **Master Data Services (MDS)** : met en place des référentiels exploitables par l'ensemble du système d'information



Ces outils travaillent de concert et s'interfacent également avec les autres briques décisionnelles Microsoft.

Data Quality Services

Data Quality Services (DQS) est l'outil de la plate-forme SQL Server permettant de travailler la qualité des données.

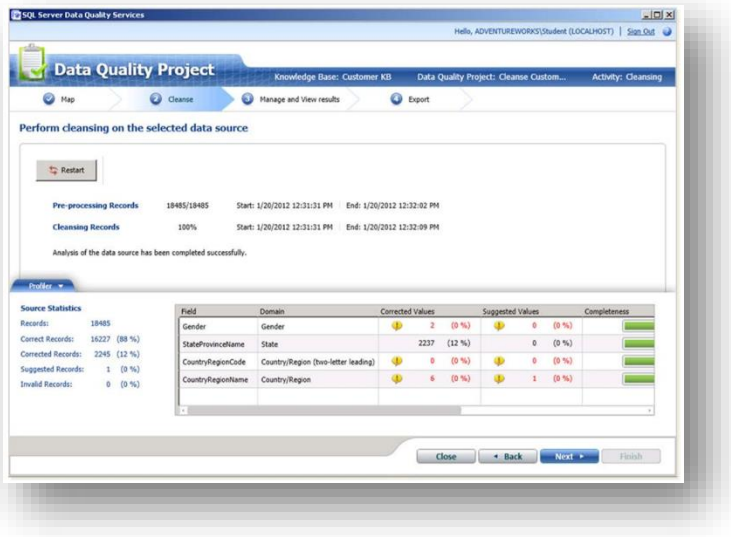
Au centre de l'outil, une base de connaissances contenant toutes les règles afférentes à la qualité. Son utilisation est basée sur l'apprentissage et la base s'enrichit itérativement au fur et à mesure des intégrations de données.

Pour chaque élément métier (on parle de domaine) à vérifier comme le nom d'une ville, un type, le nom d'un magasin, etc., on définit un ensemble de règles à respecter. On trouve les règles suivantes :

- Condition de validation (type, taille, masque de saisie, etc.)

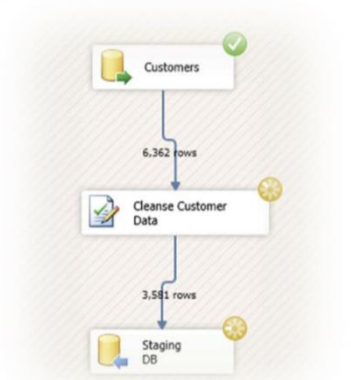
- Valeurs connues ou de référentiels existants (cf. Azure Data Market dans Gérer les données d'entreprise)
- Corrections automatiques, synonymes
- Détection de valeurs incorrectes ou inconnues
- Règles de similarité pour détecter des doublons
- Composition de règles entre domaines

Ces règles peuvent être créées et maintenues par des utilisateurs métiers comme par des administrateurs. DQS dispose d'un outil client pour gérer la base de connaissance.



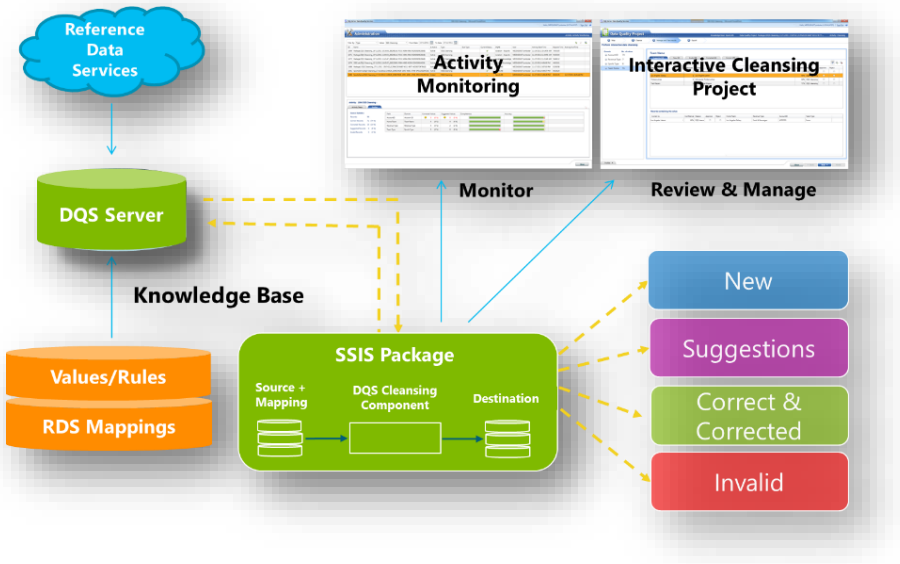
Cet outil a également un autre usage pour les utilisateurs.

DQS est focalisé sur l'évaluation de la qualité de vos données (le « scoring »). L'outil permet donc de passer au crible un jeu de données et de ressortir tout ce qui ne va pas vis-à-vis de sa base de connaissances, donnant ainsi un indice de confiance sur vos données. Il permet également de rechercher des similarités ou des doublons sur les données.



Évidemment, cet usage est industrialisable grâce à un composant SSIS permettant d'utiliser la puissance de DQS au sein d'un ETL.

DQS s'intègre parfaitement dans la plateforme de données SQL, offrant également tous les outils permettant sa gestion quotidienne, que ce soit par les utilisateurs ou par les services informatiques.



Gestion de référentiels avec MDS

Dans une solution décisionnelle, il y a toujours un besoin d'avoir quelques tables de référence. Évidemment, on pense à certaines dimensions normalisées au sein d'un département ou de l'entreprise, mais on a aussi régulièrement besoin de tables plus techniques comme des tables de transcodage ou des tables de mapping par exemple.

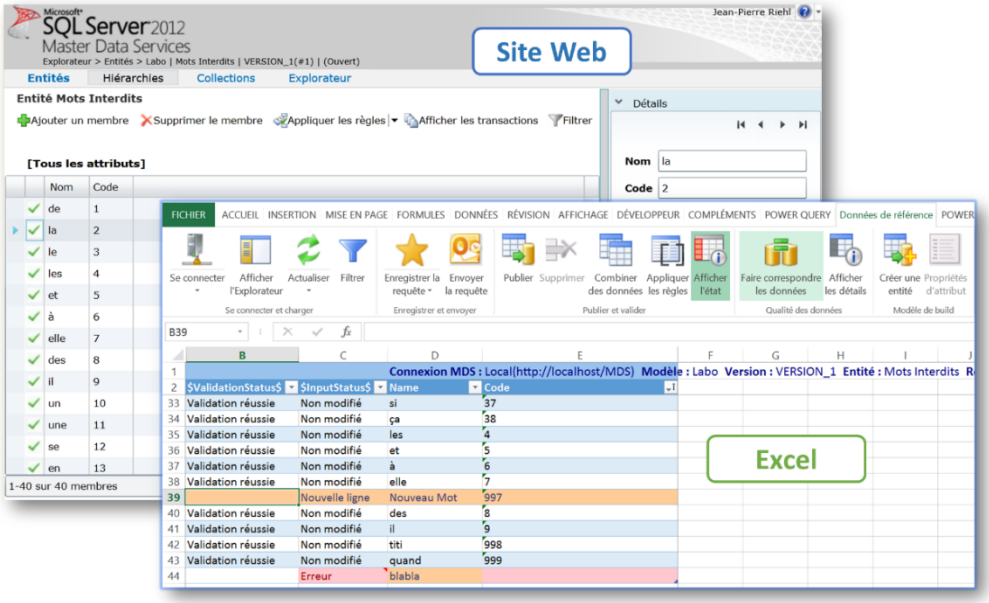
Ces tables de référence sont par essence maintenues pour et par les utilisateurs métiers car elles représentent en partie la configuration de leur activité vis-à-vis d'un système décisionnel.

La principale difficulté est de mettre à disposition des interfaces de saisie pour que les utilisateurs puissent faire vivre ces « référentiels ».

Dans la plate-forme SQL Server, c'est Master Data Services qui répond à ces besoins. MDS est l'outil de gestion des données de références de Microsoft et propose :

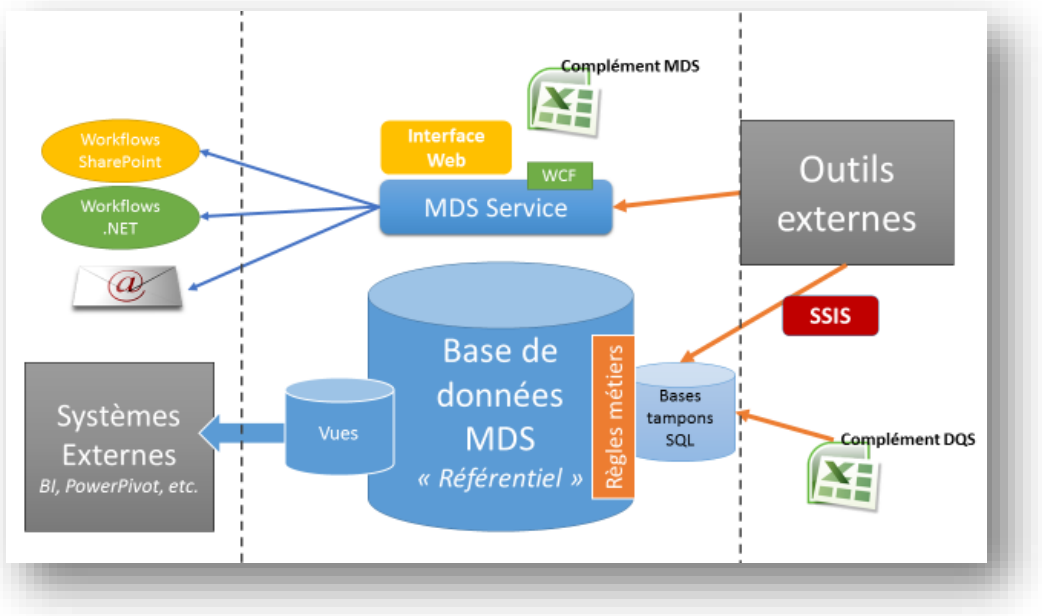
- La création et la gestion de référentiels
- La création et la gestion de règles de validation
- La saisie ou l'importation des données de référence
- La création et la gestion de hiérarchies
- La création de vues métiers à destination des applications tierces
- Le branchement des opérations de base sur des workflows

Master Data Services est accessible via un site web assurant une bonne partie des opérations de gestion. Mais il existe également un complément qui permet de manipuler les données des référentiels directement dans Excel.



En plus des interfaces utilisateurs, MDS s'intègre parfaitement dans le système d'information par le biais d'API. D'autres briques du SI peuvent interagir avec MDS via :

- L'appel de services WCF
- Le branchement avec des Workflows .NET (Workflow Fundation) ou SharePoint
- Le chargement par lot dans des tables tampons, par exemple SSIS
- L'export de vues de références SQL

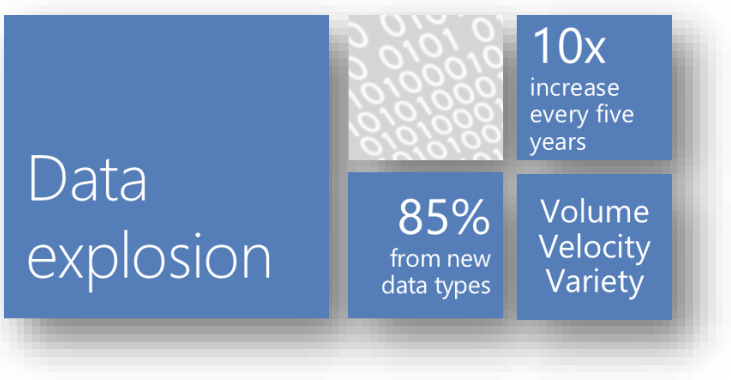


Toutefois, même si Master Data Services se classe parmi les outils de Master Data Management (MDM), cette catégorie d'outils implique obligatoirement une démarche d'entreprise associée pour gérer les données de références au niveau de l'ensemble de la société. MDS apporte la première pierre à la construction de cette démarche et favorise sa mise en œuvre dans l'entreprise.

Big Data

Comme énoncé dans la présentation du document, le nombre de sources à analyser augmente, la fréquence à laquelle les données sont générées ne cesse aussi d'augmenter et leurs formats sont de moins en moins structurés. En parallèle les coûts de stockage ont bien diminué tout comme les ordinateurs de calculs.

Là où hier on détruisait les données qui n'apportaient aucune intelligence à l'instant T, aujourd'hui nous pouvons nous permettre de les sauvegarder et de les analyser plus tard. Cette tendance a pour but de démocratiser l'utilisation de nouvelles technologies dites de Big Data permettant la gestion et l'analyse de données jusqu'à présent difficilement exploitable.



Gartner "By 2015 businesses that build a modern information management system will outperform their peers financially by 20 percent."

HDInsight

HDInsight est une plate-forme Apache Hadoop (Framework Open Source Big Data) disponible dans un environnement Windows. Il a été développé en partenariat avec Hortonworks et se base sur Hortonworks Data Platform (HDP). HDInsight permet le traitement de gros volumes de données structurées et non structurées que les systèmes traditionnels de bases de données relationnelles ne peuvent généralement pas supporter pour de nombreuses raisons (volumétrie, types de données, temps de traitement, prix, …).

HDInsight est disponible en deux versions : dans le cloud Azure en mode PaaS et la version On-Premise (à demeure). Depuis peu, HDInsight est également disponible dans l'Appliance[7] massivement parallèle APS.

HDInsight est composé de :

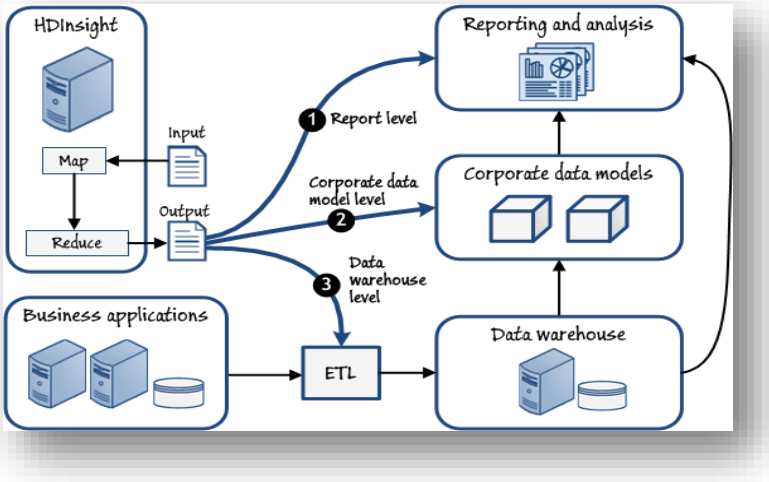
- **Hadoop** : Le Framework open-source est composé de plusieurs modules :
  - Un système de fichiers distribué appelé **HDFS** (Hadoop Distributed File System) répartissant et répliquant sur N nœuds les données découpées en bloc.
  - **Map Reduce**, un modèle de programmation permettant le traitement de données en parallèle. Dans la première étape (MAP), un problème se décompose en de nombreux sous problèmes envoyés aux serveurs de traitement. Dans la deuxième étape (REDUCE), les résultats de l'étape précédente sont combinés pour créer les résultats définitifs du problème d'origine.
  - Des couches d'abstraction au modèle de programmation Map Reduce, comme par

[7] Une Appliance est un produit matériel intégrant le logiciel, permettant de répondre à un besoin par une solution clé en main.

exemple :
  - **Hive** un système d'entrepôt de données,
  - **PIG** une plate-forme d'analyse et de traitement de données proche d'un ETL,
  - **Oozie** un système de workflow, un planificateur pour gérer les jobs,
  - **Sqoop** une application en ligne de commande permettant de transférer des données entre une base de données relationnelle et Hadoop et vice versa
  - **HBase** un système de stockage de données NoSQL (en colonnes) permettant d'accéder aux big data en mode lecture/écriture, de façon aléatoire et en temps réel.
- Scripts **PowerShell** permettant de déployer et de configurer un cluster Hadoop en quelques minutes (environ 10 minutes).
- Des fonctions de programmation pour différents langages, notamment **.NET** et **Java**. Les développeurs .NET peuvent même exploiter la puissance HDInsight via **LINQ** (LINQ to Hive).
- Le pilote **Hive ODBC** permettant de se connecter au cluster HDInsight.

Une nouvelle brique du SI ?

Apache Hadoop n'est pas un substitut à une base de données, le schéma ci-contre présente un exemple d'intégration au sein d'un système d'information décisionnelle.



À l'issu des traitements dans HDInsight, il en ressort un jeu de données intégrable dans un datawarehouse via un ETL. Ce jeu de données peut aussi venir enrichir des modèles tabulaires ou multidimensionnelles ou encore directement des rapports.

Les capacités de HDInsight à traiter en parallèle, de gros volumes de données, structurées ou non viennent donc en complément des autres outils décisionnels de Microsoft. Par exemple HDInsight s'intègre naturellement avec Power Pivot, Power Query, dans SQL Server via la création d'un serveur lié, …

Acquisition des données

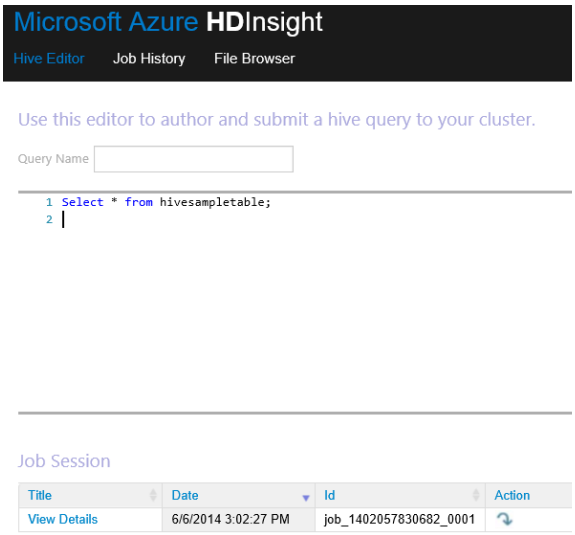
La première étape dans l'utilisation de HDInsight est l'intégration des données dans le système de fichier distribué (HDFS).

Contrairement au SGBD traditionnel, dans Hadoop aucun schéma n'est appliqué lorsque les données sont intégrées. Le schéma sera défini lors de l'exploitation des données. L'idée est d'intégrer les données dans leurs formats d'origine avec le maximum d'information sans se préoccuper de leurs tailles (principe du Data Lake). HDInsight se charge de stocker de façon efficiente et d'exploiter ces données a posteriori en fonction des besoins.

Voici une liste non exhaustive de façons d'intégrer des données dans HDFS :

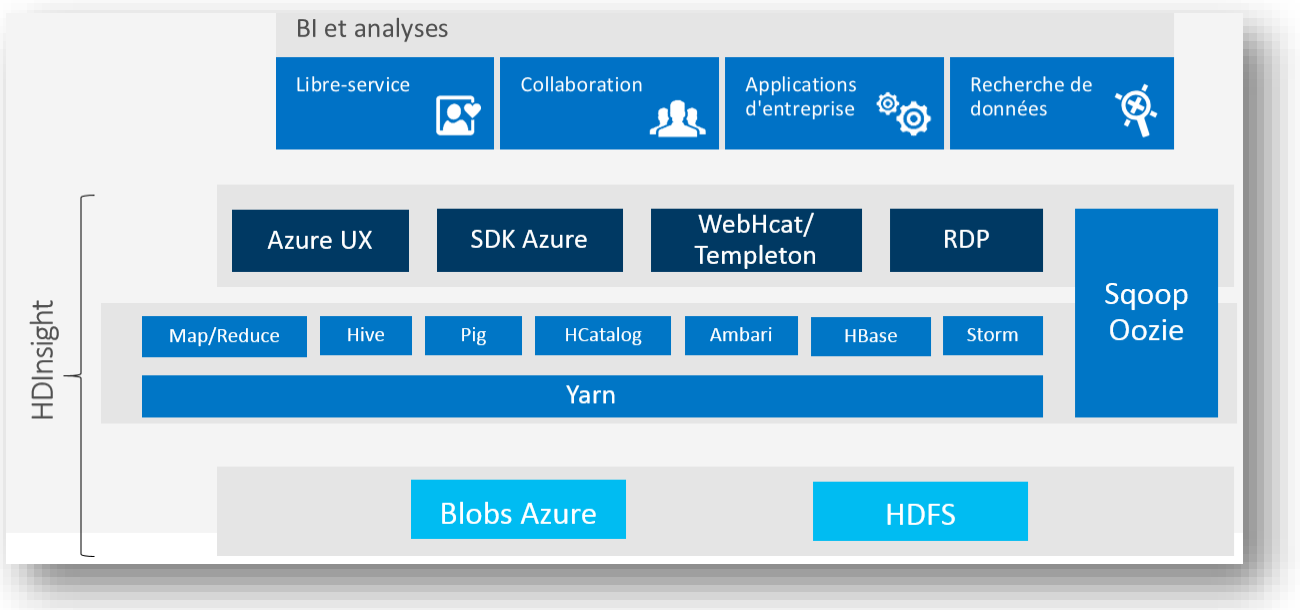
- Importation de données directement dans le Blob Storage Azure via **AzCopy**, **PowerShell Azure**, **Azure Storage Explorer**, …
- Importation de données locales au cluster via des commandes **Hadoop**
- Importation de données depuis une base de données **SQL Server** via **Sqoop**
- Importation de données à la volée via **Flume** et **Storm**

Remarque : Azure HDInsight prend en charge le système HDFS et le stockage d'objets blob Azure pour stocker des données. La force d'HDInsight dans Microsoft Azure est de pouvoir commissionner et dé-commissionner des clusters ou des nœuds de clusters Hadoop à la volée. C'est pour cela qu'il est préférable de tirer parti de Blob Storage Azure qui lui est permanent, géo-redondés et sécurisé. HDInsight sait utiliser ce stockage comme un stockage HDFS de façon transparente et par défaut. Cette architecture permet de dé-corréler le stockage (et par conséquent l'acquisition des données) des unités de calcul.



L'onglet par défaut du Dashboard HDInsight est l'éditeur Hive. D'autres onglets permettent d'accéder à l'historique des travaux et à l'explorateur de fichiers

HDInsight : Solution Hadoop avec Microsoft

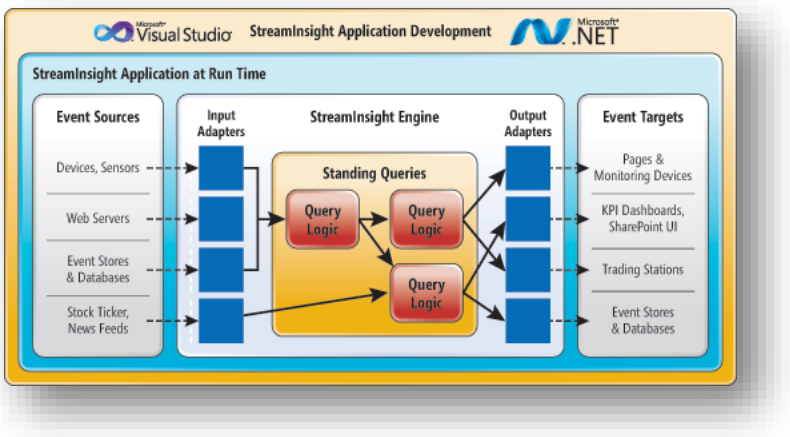


StreamInsight

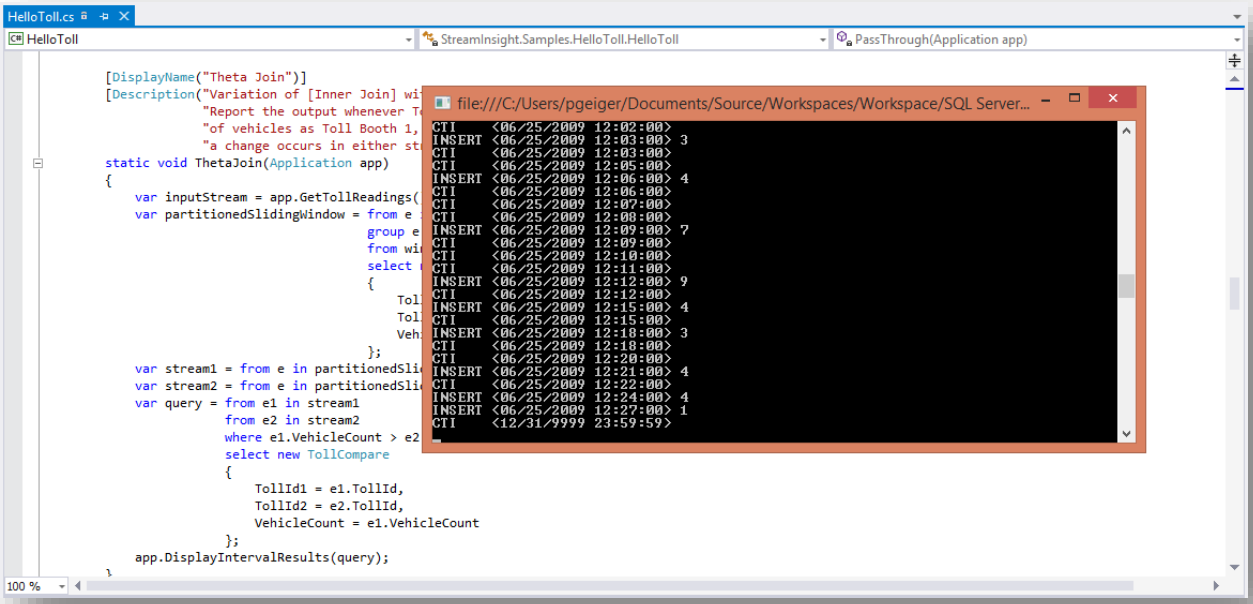
StreamInsight est une plate-forme prévue pour développer et déployer des applications de traitement **des événements complexes (CEP).**

Son architecture de traitement de flux à haut débit et la plate-forme de développement basée sur le **Framework .NET** donnent les moyens d'implémenter rapidement des applications de traitement d'événements performantes.

Les données des sources de flux d'événements proviennent généralement de programmes de fabrication, d'applications financières ou de services d'analyse Web et d'analyse opérationnelle.



Pour bien comprendre le rôle que peut jouer StreamInsight, le mieux est de prendre un exemple. Cet exemple est tiré du milieu industriel, et plus spécifiquement d'une machine-outil. Cette dernière, en complément de sa fonction principale, dispose d'un certain nombre de capteurs qui mesurent toute une série de grandeurs de fonctionnement (vitesses, températures, débits, alarmes, consommation d'énergie, etc.) Ces mesures arrivent en flux continu à intervalles plus ou moins rapprochées. Même s'il est possible de stocker toutes ces mesures au rythme où elles arrivent dans une base de données, le traitement de celles-ci (qui sont maintenant des données) restent à faire. Pour comprendre ces données, il faut probablement les corréler entre elles pour comprendre l'état de fonctionnement de la machine-outil. Le rôle de StreamInsight est justement de proposer une plate-forme logicielle qui accède à toutes ces mesures en flux continu, qui les combinent entre elles, et qui restituent enfin des données plus exploitables car plus synthétiques.



Autre exemple encore plus concret : un four industriel dispose d'un capteur de température qui fournit une valeur toutes les secondes. L'utilisateur n'a pas besoin de disposer de toutes ces valeurs. L'application développée avec StreamInsight va donc calculer pour chaque heure, la moyenne des températures sur cette heure (ou sur une journée si tel est le besoin) ainsi que la plus grande et la plus petite des mesures de températures. Ainsi, l'utilisateur disposera de données plus synthétiques et donc plus facilement utilisables.

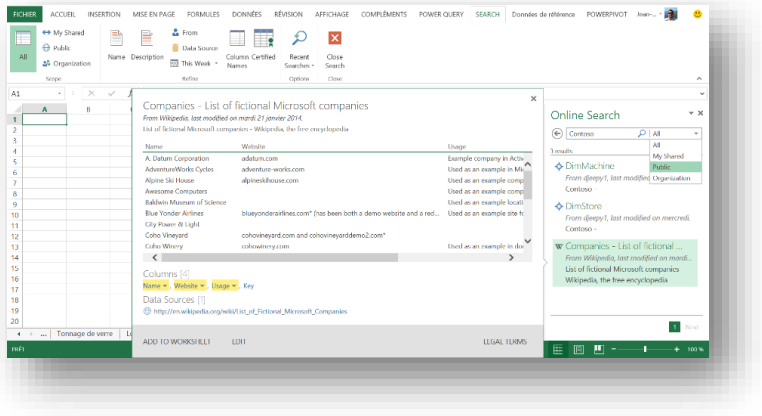
Quelle est la place de StreamInsight dans une architecture décisionnelle ? En fait, une plate-forme logicielle exploitant la technologie StreamInsight trouve sa place dans le cadre de l'acquisition de données. À partir d'une série d'événements complexes, l'objet est d'obtenir des données exploitables qui seront ensuite jointes aux autres données issues des autres systèmes d'informations.

Gérer les données d'entreprise

Recherche et partage de données, une vision collaborative de l'acquisition des données

Self-Service ne veut pas dire seul. Après avoir construit un jeu de données (une requête), un utilisateur pourra souhaiter le partager avec ses collègues, afin qu'eux-mêmes puissent à leur tour l'utiliser, directement ou pour en créer de nouveaux.

Power Query, dans l'offre Power BI, permet de répondre à ce besoin. Dans la philosophie même de Power BI, il y a la notion de construction d'un catalogue de données, partagé par l'ensemble de l'entreprise.



D'un côté, il est possible, depuis Excel, de partager son jeu de données dans un référentiel qui se trouve dans le tenant Power BI, dans le Office 365.

C'est en fait la requête et ses métadonnées (description, mots clés, colonnes, etc.) que l'on partage ; les données ne sont incluses que si on le précise.

Tout est indexé dans un « catalogue » disponible à l'ensemble des utilisateurs de votre entreprise, modulo les permissions.

De l'autre côté, les utilisateurs peuvent rechercher parmi ce catalogue et retrouver les jeux de données (les requêtes) partagés par leurs collègues afin de les utiliser dans leurs classeurs Excel. Power Query offre une interface de recherche dans le Catalogue mais pas seulement.

Le moteur de recherche de Power BI ne se limite pas aux jeux de données partagées par les collaborateurs, il peut également indexer les

Passerelle de gestion des données

Les fonctionnalités collaboratives de Power BI se trouvent sur le Cloud alors qu'une grande partie des données d'entreprise résident sur ses serveurs, On-Premise.

Cela implique de concevoir des architectures hybrides faisant le lien entre ces 2 mondes.

La passerelle de gestion des données (Data Management Gateway en anglais) est la brique qui permet de faire ce lien.

Elle offre aux administrateurs IT le moyen de rendre accessible des sources de données internes au SI vers le tenant Power BI de l'entreprise. Évidemment, la connexion et le transfert de données sont sécurisés.

Cette brique s'ajoutent aux multiples façons de concevoir des architectures hybrides entre le SI d'entreprise et le Cloud Microsoft, que ce soit Azure ou Office365 (AD, réseau, etc.)

bases de données ou autres sources locales, mais aussi des sources externes comme Wikipédia ou Azure Data Market.

L'objectif est de faciliter le travail d'un « analyste » en lui permettant de trouver les données dont il a besoin, qu'elles soient dans son entreprise ou à l'extérieur.

Microsoft Azure Marketplace

Trouver la bonne donnée est souvent une véritable quête. Avant d'avoir des jeux de données nettoyés et façonnés, il faut souvent trouver des données sources fiables.

À l'intérieur de l'entreprise, cela ne pose généralement pas de problème mais quand l'entreprise ne possède pas la donnée, il faut aller la chercher quelque part. Et même si Power Query propose un moteur pour rechercher dans les données publiques, trouver son bonheur reste difficile parmi la montagne de données disponibles.

Et pourtant, c'est le métier de certaines entreprises ou d'organismes institutionnels que de proposer des données qualifiées et à forte valeur ajoutée (Census, Eurostat, OCDE, INSEE, etc.).

Microsoft propose une plate-forme leur permettant de mettre ces données à disposition et même de les monétiser. Cette plate-forme est Microsoft Azure Marketplace, elle est une source de données disponible facilement depuis les outils Microsoft.

On y trouve de nombreux jeux de données en provenance de grands acteurs comme Dun & Barnstreet mais aussi d'acteurs de niche proposant des données très spécialisées.

Un nouveau rôle, le Data Steward

Avec d'un côté des utilisateurs qui partagent leurs requêtes (jeux de données) et d'un autre des utilisateurs qui recherchent de l'information, on peut vite, dans une entreprise, se retrouver perdu avec un Catalogue surchargé.

Requêtes en double, requêtes similaires, requêtes trop compliquées ou trop lourdes, mélange entre différents services de l'entreprise, non-différenciation des sources réputées fiables, et des requêtes personnelles à la marge. Toutes ces questions sont légitimes et cette nouvelle façon d'acquérir des données implique de nouvelles façons de travailler.

La discipline à mettre en œuvre pour bien gérer ces outils d'acquisition et de partage, c'est la **Gouvernance**. Et pour la mettre en place, un nouveau rôle voit le jour : le Data Steward, ou en français, le gestionnaire des données.

Son rôle est de veiller à ce que les données partagées ou disponibles répondent aux besoins des utilisateurs. Il corrige, il qualifie, il contrôle, il accompagne les utilisateurs. Il fait en sorte que les bonnes données soient disponibles au bon moment pour la bonne personne. Il permet de gérer et maximiser le capital de données de l'entreprise.

Comment le mettre en place ? Doit-on mettre une équipe de Data Stewards transverses à l'entreprise ? Ou bien un Data Steward dans chaque équipe métier ? Doit-on privilégier un nouveau recrutement ou bien une formation professionnelle des équipes existantes ? Est-ce une responsabilité de l'IT ou bien des services métiers ?

Les réponses à ces questions doivent venir de l'organisation de l'entreprise. Le gestionnaire des données doit correspondre au mode de fonctionnement des utilisateurs et également être aligné à la stratégie de l'entreprise.

Quels outils pour quel usage ?

Les outils cités dans ce livre blanc

SSIS (SQL Server Integration Services) : est un outil d'extraction, de transformation et de chargement (ETL). Parce qu'il fonctionne en mémoire, SSIS offre des performances inégalées sur le serveur sur lequel celui-ci fonctionne

Power Query (complément d'Excel) : offre sur le poste de l'utilisateur, des fonctions d'ETL puissantes. PowerQuery accède à toutes sortes de sources de données et après les traitements nécessaires, les met à la disposition de l'utilisateur dans une feuille de calcul.

DQS (Data Quality Services) : est une plate-forme serveur qui offre à partir de bases de connaissance, les outils nécessaires à traiter les données internes de l'entreprise en vue d'en améliorer la qualité.

MDS (Master Data Services) : est une plate-forme serveur qui centralise les données de l'entreprise afin de le partager ensuite sur toutes les autres plates-formes de l'entreprise

HDInsight : est une composante de Microsoft Azure. Parce que le coût du stockage et du calcul n'a jamais été aussi faible, HDInsight permet de stocker toutes les données dans leur format d'origine, mais aussi de traiter autant que nécessaire et a posteriori ces données afin d'en obtenir une synthèse exploitable par d'autres outils décisionnels.

StreamInsight : est un Framework de développement .Net qui permet de lire des flux des événements complexes dans le but de fournir en sortie des données structurées pouvant être rendu accessible par d'autres outils décisionnels.

Outils	Type d'usage	On premise	Cloud	Profil	Outillage	Licensing
SSIS	ETL	Oui	IaaS	Développeur	SQL Server Data Tools	SQL Server
Power Query	ETL self-service	Oui	Non applicable	Utilisateur métier Data-steward	Complément Excel	Excel
Data Catalog	Magasin de données d'entreprise	Non	SaaS	Data-steward	Power BI pour Office 365	Power BI pour Office 365
Azure Marketplace	Magasin de données public	Non	PaaS	Développeur Data-steward utilisateur métier	Multiple	À l'usage
DQS	Qualité	Oui	IaaS	Développeur Data-steward	Client spécifique	SQL Server
MDS	MDM	Oui	IaaS	Développeur Data-steward	Client spécifique	SQL Server
HDInsight	Big Data	Oui	PaaS	Développeur	Scripts	Microsoft Azure
Azure Storm	Flux d'événements (CEP)	Non	PaaS	Développeur	Scripts	Microsoft Azure
StreamInsight	Flux d'événements (CEP)	Oui	IaaS	Développeur	Visual Studio	SQL Server

En savoir plus

Conférence de Satya Nadella sur la culture de la donnée vue par Microsoft : http://bit.ly/1oCFmnd

Microsoft dans le carré des leaders du Cloud : http://www.journaldunet.com/solutions/cloud-computing/gartner-magic-quadrants-du-cloud.shtml
http://blogs.microsoft.com/blog/2014/05/30/the-power-of-and/

EIM : http://bit.ly/VcekHc.

EIM Datasheet : http://bit.ly/1oCFRO5

MDM Architecture hub : http://bit.ly/1r59NoP

Microsoft Azure Marketplace : http://bit.ly/1A7Fu1p

MSDN Big Data France : http://blogs.msdn.com/b/big_data_france/

MSDN Machine Learning France : http://blogs.msdn.com/b/mlfrance/

Mises à jour Azure : http://azure.microsoft.com/fr-fr/updates/

Vos commentaires nous aideront à améliorer la qualité de nos livres blancs.

Envoyez vos commentaires.