

# La stratégie de reprise des données

Dans un contexte économique en perpétuel mouvement, l'évolution des systèmes d'information des entreprises est permanente. Fusions, cessions, lancement de nouveaux produits ou services, adaptation au marché, optimisation, évolution de la législation... constituent autant de raisons d'adapter, de rénover ou de refondre le SI. Les projets informatiques qui en découlent sont très souvent organisés en chantiers (conception applicative, développement, technique, formation, conduite du changement...) parmi lesquels la reprise des données tient une place importante.

## Reprise ou migration

La reprise des données, appelée également **migration des données**, consiste à transférer les données d'une ou plusieurs applications existantes vers une application cible. Selon les volumes, le transfert s'effectue de façon automatisée ou par saisie manuelle. Dans la plupart des cas, le transfert des données des sources vers la cible s'accompagne de **transformations** plus ou moins complexes liées au contenu de la future application à mettre en œuvre, comme l'indique le schéma de la figure 1 :

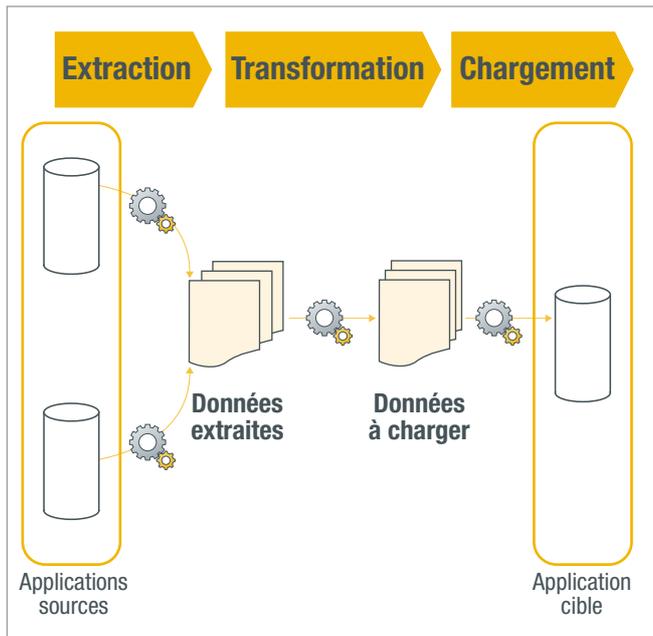


Figure 1 : Des sources à la cible

La reprise des données est le processus complet permettant de :

- collecter les données depuis des sources (bases de données, fichiers, documents) ;
- transformer ces données collectées pour les adapter à l'application cible ;
- initialiser les données (après transformation) dans l'application cible, pour que celle-ci puisse être utilisable lors de sa mise en service.

## De la collecte à l'extraction de données

Lors d'une reprise, la phase de collecte permet de récupérer les données provenant de différentes sources. Généralement, les sources se situent au sein du système informatique de l'entreprise. Cependant, dans certains cas, elles peuvent aussi se trouver à l'extérieur de l'entreprise. Ce second cas concerne des données gérées et attribuées par des organismes ou des sociétés de normalisation spécialisées dans des domaines spécifiques. Ainsi, La Poste répertorie les adresses des personnes physiques et morales situées en France. Il s'agit certainement de la meilleure source à utiliser si l'on souhaite, par exemple, mettre à jour un fichier comportant des adresses françaises lors de la reprise des données.

**Les sources** sont, en règle générale, matérialisées par des bases de données, des fichiers (Excel ou autres formats), ou des documents sous forme papier. Lorsque l'entreprise met en œuvre une application informatique concernant un domaine totalement nouveau et jamais géré jusqu'alors, il est fréquent qu'il n'y ait pas de source matérielle existante. Dans ce cas, l'information est à reconstituer ou à créer en fonction de ce qui est attendu et spécifié.

Lorsque l'opération de collecte est automatisée, on parle **d'extraction**. L'extraction des données depuis des systèmes sources est particulièrement intéressante dans le cas de volumes importants.

Une fois les sources identifiées, et les informations à extraire définies, l'extraction ne pose en général pas de problème particulier. En effet, la plupart des applications disposent d'outils de requête. À défaut, les extractions peuvent être effectuées directement depuis la base de données ou les fichiers par des outils de requête (SQL par exemple), ou encore en utilisant des outils spécialisés ETL (Extract, Transform and Load) comprenant des extracteurs capables de se connecter à de multiples sources et formats.

## Profiter de l'occasion pour transformer à la volée

Le passage des données sources aux données cibles constitue **le cœur de la reprise** des données. Les raisons de **la transformation des données** sont nombreuses, et varient selon le contexte. Globalement, la transformation est le reflet des améliorations du nouveau système par rapport à l'ancien, et se traduit informatiquement par l'un ou plusieurs des cas suivants :

- **modification du modèle de données** : le modèle de données cibles peut être différent du modèle de données sources, soit parce que celui-ci présente des inconvénients auxquels on souhaite remédier (par exemple, accès à l'information compliqué), soit parce que l'application cible est un progiciel avec son propre modèle de données, présentant forcément des différences par rapport à la source. Lors de la migration, le contenu des zones des objets cibles est déterminé selon des règles de transformation qui s'appliquent au contenu de zones d'objets sources. Les règles correspondantes peuvent s'avérer complexes.
- **modification des règles de codification** de certaines zones : la modification de la codification du contenu des zones est généralement guidée par le souhait d'améliorer la codification existante. Celle-ci peut en effet présenter des défauts comme des doublons ou encore des codes significatifs qui ne collent plus à la réalité de l'entreprise. La transformation peut s'appuyer sur une simple table de correspondance, ou au contraire faire appel à un programme de transcodification plus complexe.



En pratique, la chaîne de transformation peut être plus ou moins longue. Dans le cas de transformations complexes, il est relativement fréquent d'utiliser une base de données intermédiaire dite « pivot » comme le montre la figure 2 :

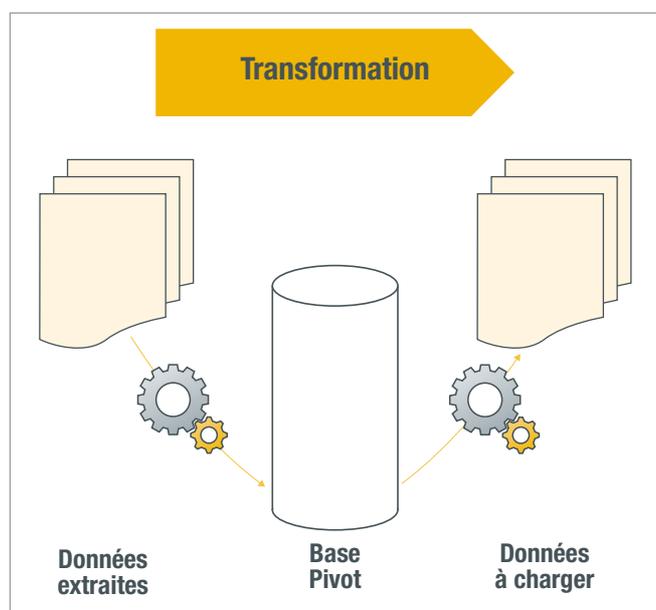


Figure 2 : Transformation avec utilisation d'une base pivot

Le modèle de données de la base pivot est généralement très proche, voire identique au modèle cible.

**L'intérêt de la base intermédiaire** est qu'elle permet de :

- centraliser les données provenant de multiples sources ;
- vérifier les transformations ;
- effectuer des traitements, avant injection dans l'application cible.

Prenons l'exemple d'une reprise de données « fournisseurs » issues de plusieurs bases de données régionales que l'on souhaite à l'avenir gérer au sein d'une seule base nationale. Sachant qu'un même fournisseur peut exister dans plusieurs bases, la base pivot permettra dans ce cas de nettoyer les données issues des différentes sources, de s'assurer que les transformations sont conformes aux spécifications, et de traiter les doublons avant injection dans la nouvelle application.

## Le chargement final

Une fois la transformation réalisée, le chargement a pour objectif d'injecter les données dans la base de données cible. Le chargement se fait avec des outils qui dépendent de l'application cible. Lorsqu'il s'agit de progiciels, les éditeurs proposent des outils dédiés au chargement des données. Cependant, dans bon nombre de cas, les outils standards doivent être adaptés en fonction des développements spécifiques réalisés pour l'application cible.

## Stratégies et méthodes de migration

Les opérations de reprise de données ne s'improvisent pas ! Elles nécessitent de la méthode, de l'organisation, et un suivi fin. En amont, et parmi les toutes premières tâches, il convient de définir un nombre de choix qui guideront le déroulement du chantier : c'est la stratégie de reprise (ou de migration).

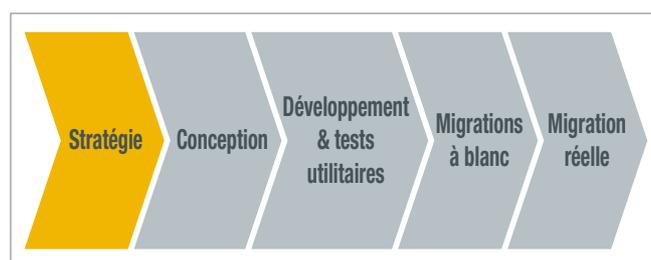


Figure 3 : La stratégie de reprise, première étape essentielle d'un projet de migration

Plus précisément, **la stratégie** a pour objectif de :

- Définir le périmètre applicatif de la reprise des données ;
- Identifier les objets qui feront l'objet d'une reprise, ainsi que les éléments de volumétrie qui permettront de choisir le mode de reprise approprié (automatique ou manuel) ;
- Identifier les référentiels sources de chaque objet cible ;
- Définir les outils de reprise par objet ;
- Définir les rôles et les responsabilités ;
- Décrire les principes directeurs et les recommandations à suivre au cours des phases de conception et de réalisation ;
- Planifier les travaux de construction et d'exécution des reprises.

## FACTEURS CLÉS DE SUCCÈS D'UN PROJET DE REPRISE DE DONNÉES

Communiquer sur le rôle essentiel des données dans la réussite d'une nouvelle application

Anticiper la complexité de ce type de projet

Ne pas sous-estimer la charge des différentes étapes (conception, réalisation, tests, validation...)

Considérer la reprise sous un angle à la fois fonctionnel et technique

Impliquer les utilisateurs métier dès la conception

Prévoir et s'assurer que le chantier de reprise des données travaille en étroite collaboration avec les chantiers fonctionnels

Prévoir des étapes de contrôle et de validation des données migrées

Maîtriser et appliquer une méthodologie solide (stratégie, conception, développement des outils, tests des outils, migrations à blanc, contrôle des données...)

Mener des actions de nettoyage des données dans les systèmes sources

Utiliser des plates-formes et des outils de reprise éprouvés

### Des succès mitigés sur une opération sous-estimée

Force est de constater que trop de projets rencontrent des difficultés dues à la reprise des données avec pour conséquence des dépassements de temps et de budget, et au final beaucoup d'efforts pour un faible résultat.

Le Standish Group indique ainsi que **sur 10 projets de migration des données, 1 termine dans les temps et avec le budget prévu**, 6 finissent en retard et en dépassement de budget, 3 sont abandonnés.

L'une des principales raisons des difficultés est liée à l'enjeu de la reprise des données qui n'est pas suffisamment perçue à sa juste valeur. Il en résulte un sous-dimensionnement général, parfois critique, des tâches correspondantes.

Aussi, les projets de reprise sont trop souvent réduits à leur composante technique. Bien qu'essentiel, l'aspect technique n'est pas suffisant. Il est fréquent de constater que la migration ne s'appuie que très partiellement sur la **connaissance métier** des acteurs de l'entreprise. Cette lacune se traduit assez régulièrement par un résultat décevant en termes de qualité des données de l'application cible.

Par ailleurs, **le réflexe d'industrialisation** est souvent absent de ce type de projet. Il n'est pas rare de tout réinventer, alors que l'on dispose aujourd'hui de plates-formes dédiées à la migration des données et de l'expérience de ressources techniques maîtrisant ces outils.

Enfin, certaines difficultés trouvent leur origine dans l'organisation même du projet. Ainsi, l'équipe en charge de la reprise des données n'est pas toujours informée des choix effectués par l'équipe responsable de la solution applicative cible, ce qui a un impact direct sur la qualité des spécifications de la reprise des données, et donc sur le résultat.

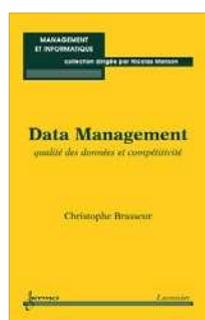
Néanmoins, les choses évoluent, et la mise en œuvre de bonnes pratiques permet d'aboutir à d'excellents résultats. ■



**Christophe Brasseur**,  
Consultant Manager  
chez Capgemini

Christophe Brasseur est ingénieur ESTP et titulaire d'un DESS de gestion de l'IAE de Montpellier. Il a participé à de nombreux projets en France et à l'étranger en tant que directeur de projet ou consultant, principalement dans le secteur des services publics, du pétrole, de l'énergie, des utilities, et de la chimie. Il est actuellement Consultant Manager chez Capgemini.

<http://mydatanews.blogspot.com>



Christophe Brasseur :  
« Data Management, qualité des données et compétitivité »,  
éditions Hermès-Lavoisier, Paris, 2005.