

Génération du jeu de test

Dans le chapitre précédent, nous avons achevé la modélisation de notre entrepôt de données. Néanmoins, pour tester et appréhender le modèle, il nous faut des données concrètes.

Lors de la phase de conception, même si nous vous engageons à vérifier la disponibilité des informations sources, nous vous incitons vivement à ne pas chercher à travailler avec les données réelles mais avec un jeu de test. Et ce, jusqu'à une validation de la modélisation.

Nous allons voir dans ce chapitre comment générer et travailler avec un jeu de test. Cette phase peut sembler lointaine des préoccupations du concepteur, elle reste cependant capitale pour la validation et la réussite de cette phase de modélisation. En effet, lors des phases de conception, il est important de faire des itérations très rapides avec les utilisateurs métier, et surtout d'être capable de leur faire voir et leur faire manipuler le fruit des décisions des dernières délibérations.

À l'issue de ce chapitre, c'est-à-dire après la création du cube, dans notre étude de cas Distrisys, nous devrions réaliser une démonstration et un atelier. Cela permettra aux utilisateurs qui ont participé à sa conception de manipuler le modèle, ce afin de le leur faire valider. L'objectif étant qu'une fois le modèle validé à l'aide de données de test, nous le mettrons de côté. Nous pourrions alors nous intéresser exclusivement au chargement des données réelles : une phase qui peut se révéler très longue.

C'est justement parce que le chargement des données réelles est souvent long et fastidieux, qu'il faut :

- S'assurer que le modèle final est validé et stable.
- Générer et travailler avec des données de test, pour permettre aux utilisateurs clés de visualiser le comportement de leur cube, et ce jusqu'à validation.

Revenons à notre étude de cas. Pour générer nos données de test, nous procéderons ainsi :

- Nous allons tout d'abord saisir des données manuellement dans toutes les dimensions.
- Nous utiliserons Excel pour générer le jeu de données des faits.
- Nous intégrerons les données Excel précédemment créées dans la table de faits SQL Server. Cette intégration se fera à l'aide de l'outil en ligne de commande BCP.

Ce processus, une fois acquis, est assez simple et rapidement reproductible à chaque modification du modèle.

Pour rappel, précédemment, nous avons saisi le contenu des tables suivantes :

- *DimProduit* : 10 lignes.
- *DimSite* : 5 lignes.
- *DimClient* : 10 lignes.

Maintenant, nous allons nous atteler à générer les données de la table de faits *FactFacture*.

Pour cela, nous allons :

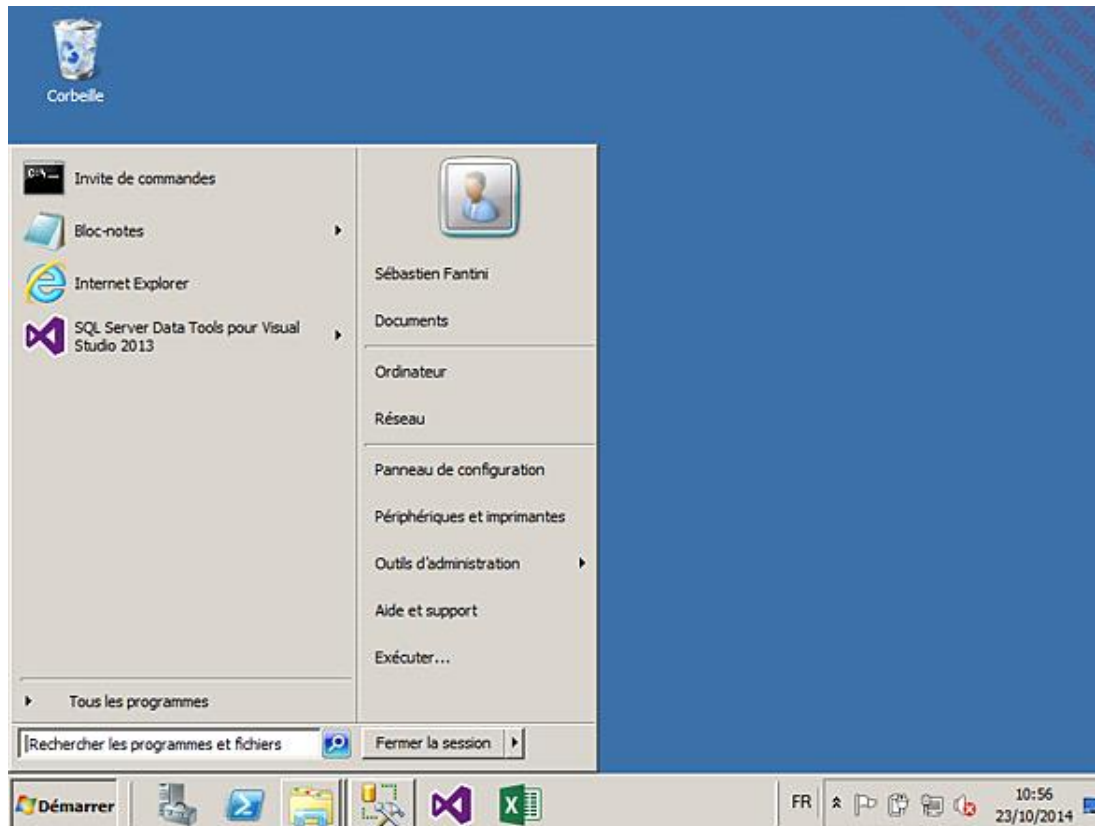
- Saisir une ligne de la table *FactFacture*.
- Exporter sous Excel le contenu et la structure de la table par l'outil BCP.
- Générer les données avec Excel.

- Importer le contenu du fichier Excel dans la table *FactFacture* avec l'outil BCP.

→ Commencez donc par saisir au moins une ligne dans la table de faits *FactFacture* :

DateFacturation_FK	Site_FK	Produit_FK	Client_FK	PrixCatalogue	Remise	CA	Marge	CoutDirectMatiere	CoutDirectMainOeuvre	CoutIndirect	Quantite	NumFacture
20140101	1	1	1	600,00	50,00	550,00	250,00	100,00	150,00	50,00	1,00	P1
20140101	1	1	1	600,00	50,00	550,00	250,00	100,00	150,00	50,00	1,00	P1

→ Ouvrez l'**Invite de commandes** :



→ Positionnez-vous à la racine de votre disque C. Pour cela, tapez dans l'invite de commandes l'instruction suivante :

```
cd\
```

→ Tapez la ligne de commande suivante, puis appuyez sur la touche [Entrée] :

```
bcp DistriSysDW.dbo.FactFacture out "FactFacture.csv" -T -c -t";"
-S"Serveur\Instance"
```

➤ L'option **-S** encadre le nom de votre serveur et instance SQL Server.

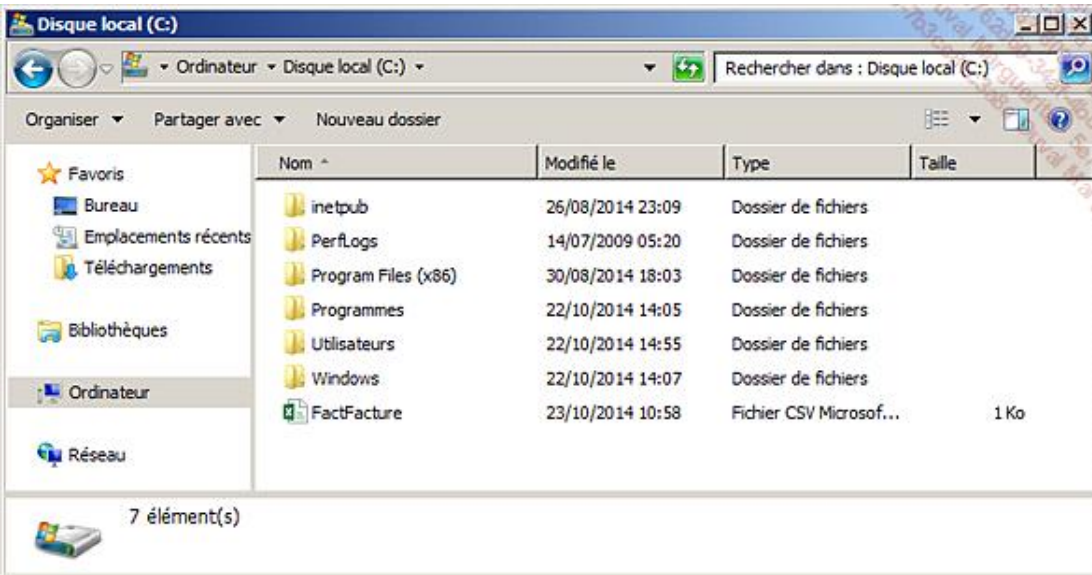
```
Administrateur : Invite de commandes
Microsoft Windows [version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. Tous droits réservés.

C:\Users\sfantini>cd\

C:\>bcq DistrisysDW.dbo.FactFacture out "FactFacture.csv" -I -c -t";" -S"Serveur1"
Dénarrage de la copie...
1 lignes copiées.
Taille du paquet réseau (octets) : 4096
Heure (ms) Total : 1 Moyenne : (1000.00 lignes par seconde)
C:\>
```

L'exportation s'est déroulée avec succès.

→ Explorez le disque C :



Le fichier *FactFacture.csv* a été généré.

→ Ouvrez ce fichier avec Excel :

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	20120101	1	1	1 600.00	50.00	550.00	250.00	100.00	150.00	50.00	1.00	F1	
2													
3													
4													
5													
6													
7													
8													
9													
10													
11													
12													
13													
14													
15													
16													
17													

Vous pouvez constater que l'on retrouve la ligne préalablement insérée manuellement dans la table *FactFacture*.

Nous allons maintenant utiliser les fonctions tirer et aléatoire de Excel afin de créer un jeu de test, que nous réintégrerons dans la table *FactFacture*.

Après modification manuelle de la colonne A du fichier Excel, nous obtenons ceci :

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	20140101	1	1	1 600.00	50.00	550.00	250.00	100.00	150.00	50.00	1.00	F1				
2	20140201															
3	20140301															
4	20140401															
5	20140501															
6	20140601															
7	20140701															
8	20140801															
9	20140901															
10	20141001															
11	20141101															
12	20141201															
13	20150101															
14	20150201															
15	20150301															
16	20150401															
17	20150501															
18	20150601															
19	20150701															
20	20150801															
21	20150901															
22	20151001															
23	20151101															
24	20151201															

La colonne A représentant la table de dimension *Temps*, nous saisissons les mois que nous souhaitons voir apparaître dans notre cube.

Dans cet exemple, nous avons saisi 24 mois, de janvier 2014 à décembre 2015, les mois étant tous représentés par le premier jour du mois. Dans notre exemple, nous considérons que détailler les ventes par jour n'a que peu d'intérêt. Cela ne sera pas toujours le cas. Dans certains cas, par exemple, afin de démontrer que vous savez calculer une moyenne des 30 derniers jours, il vous faudra forcément générer des données détaillées au jour.

La colonne B représente le site de la ligne de fait. La dimension *Site* dispose de 5 membres avec un identifiant allant de 1 à 5. Nous utilisons alors une fonction d'Excel générant une valeur aléatoire entre 1 et 5 :

```
=ALEA.ENTRE.BORNES(1;5)
```

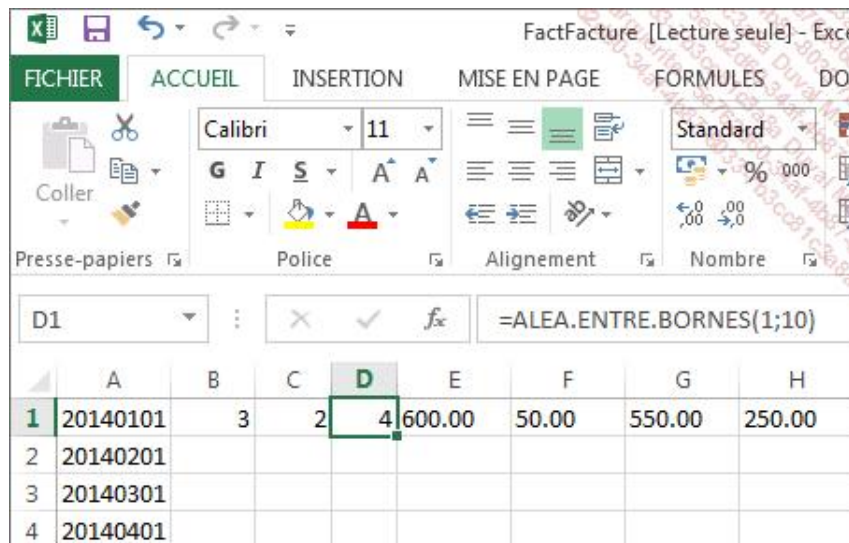
	A	B	C	D	E	F	G	H
1	20140101	5	1	1	600.00	50.00	550.00	250.00
2	20140201							
3	20140301							
4	20140401							

La colonne C représente le produit de la ligne de fait. La dimension *Produit* dispose de 10 membres avec un identifiant allant de 1 à 10. Nous utilisons alors la même astuce que précédemment, mais avec une valeur aléatoire entre 1 et 10 :

	A	B	C	D	E	F	G	H
1	20140101	5	1	1	600.00	50.00	550.00	250.00
2	20140201							
3	20140301							
4	20140401							

La colonne D représente le client de la ligne de fait. La dimension *Client* dispose de 10 membres avec un identifiant de 1 à 10.

→ Nous générerons donc un nombre aléatoire entre 1 et 10 :

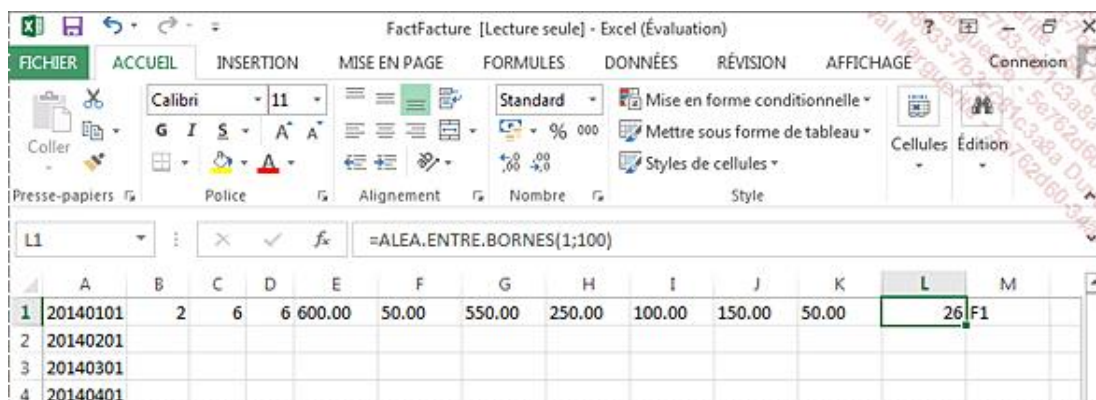


The screenshot shows the Excel interface with the formula bar containing `=ALEA.ENTRE.BORNES(1;10)`. The spreadsheet has columns A through H. The data in the first four rows is as follows:

	A	B	C	D	E	F	G	H
1	20140101	3	2	4	600.00	50.00	550.00	250.00
2	20140201							
3	20140301							
4	20140401							

Les colonnes E à L représentent les mesures. Nous pouvons utiliser la même fonction aléatoire et bien évidemment, toutes les fonctions dont dispose Excel, afin de générer des valeurs les plus cohérentes possibles pour le cas à traiter.

→ Nous commençons donc par générer la *quantité* de produits vendus : un nombre aléatoire entre 1 et 100 (valeurs totalement arbitraires) :



The screenshot shows the Excel interface with the formula bar containing `=ALEA.ENTRE.BORNES(1;100)`. The spreadsheet has columns A through L. The data in the first four rows is as follows:

	A	B	C	D	E	F	G	H	I	J	K	L
1	20140101	2	6	6	600.00	50.00	550.00	250.00	100.00	150.00	50.00	26
2	20140201											
3	20140301											
4	20140401											

→ Puis nous définissons la valeur *prix* de vente de la ligne de fait : on considère ici que nos produits sont vendus en prix catalogue entre 1 et 150 euros. Nous multiplions cette valeur par la quantité de produit vendus.

FactFacture [Lecture seule] - Excel (Évaluation)

FICHIER Enregistrer (Ctrl+S) INSERTION MISE EN PAGE FORMULES DONNÉES

Calibri 11 Standard

Coller Presse-papiers Police Alignement Nombre

E1 : $=ALEA.ENTRE.BORNES(1;150)*L1$

	A	B	C	D	E	F	G	H	I
1	20140101	2	1	1	4350	50.00	550.00	250.00	100.00
2	20140201								
3	20140301								
4	20140401								

→ Ensuite, nous définissons la remise de la ligne de fait : on considère ici que la remise se situe entre 2 % et 30 %.

Pour éviter toute erreur lors de l'utilisation avec le BCP, on arrondit la valeur afin qu'il n'y ait pas de valeurs après la virgule.

FactFacture [Lecture seule] - Excel (Évaluation)

FICHIER ACCUEIL INSERTION MISE EN PAGE FORMULES DONNÉES RÉVISION

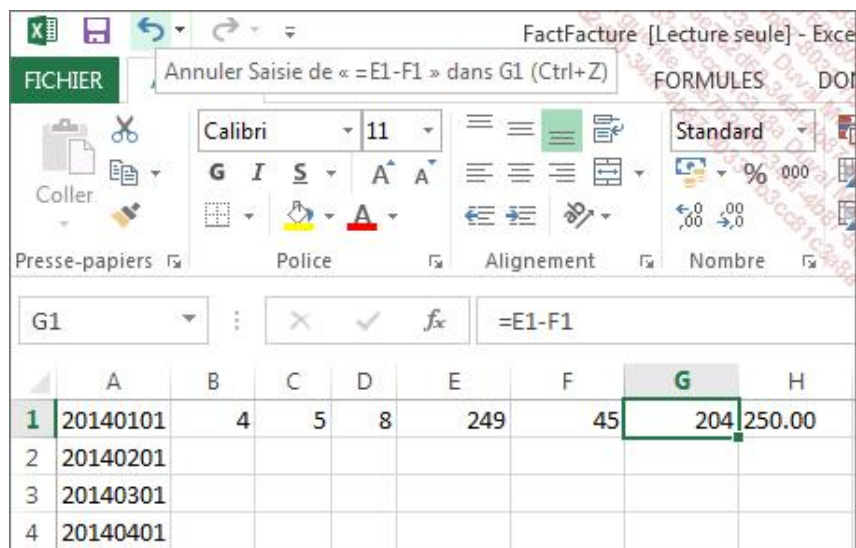
Calibri 11 Standard

Coller Presse-papiers Police Alignement Nombre Style

F1 : $=ARRONDI(E1 * ALEA.ENTRE.BORNES(2;30)/100;0)$

	A	B	C	D	E	F	G	H	I	J
1	20140101	2	3	1	1350	297	550.00	250.00	100.00	150.00
2	20140201									
3	20140301									
4	20140401									

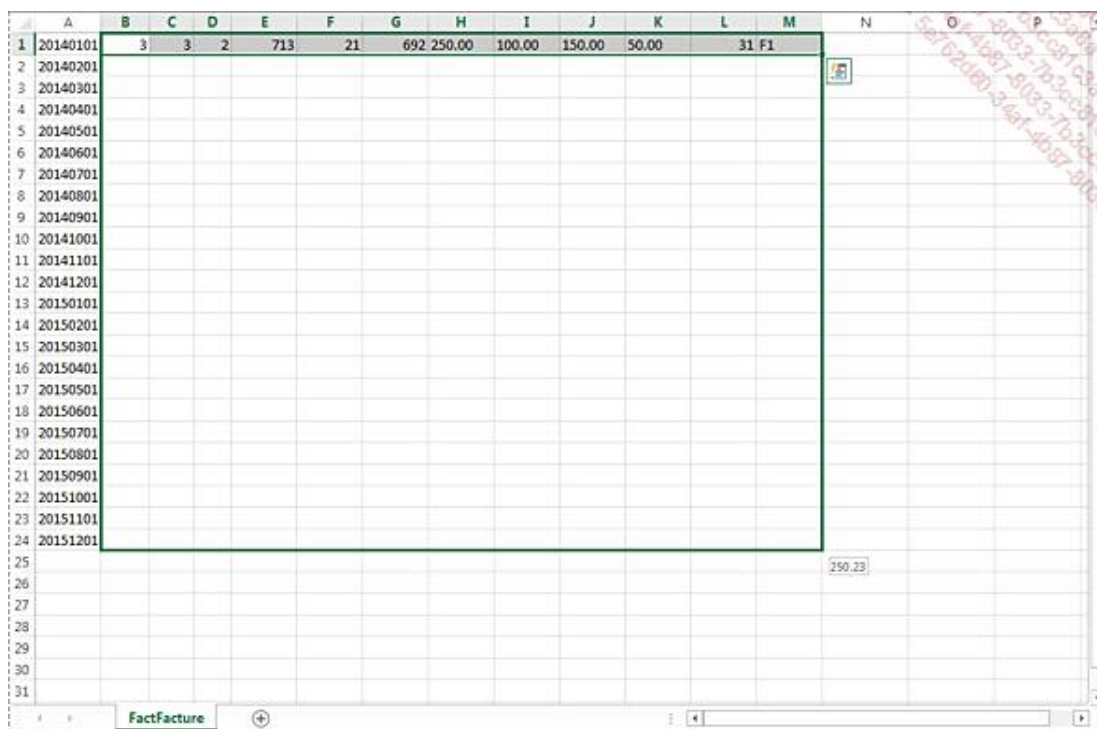
→ La colonne G est le CA, il est donc égal au prix de vente auquel on retranche la remise :



Et nous procédons ainsi de suite pour l'ensemble des autres mesures. Ce travail est plus ou moins détaillé. L'objectif étant surtout que les chiffres affichés n'entravent pas la compréhension de votre modèle

Par exemple, une marge supérieure au CA, serait incomprise par les utilisateurs qui rejetteraient alors aussitôt le modèle que vous leur proposerez.

- Une fois les règles de calcul aléatoires fixées pour chaque colonne, il vous suffit alors de tirer ces formules sur l'ensemble des mois :



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	20140101	2	10	8	1711	188	1523	250.00	100.00	150.00	50.00	59	F1
2	20140201	5	6	7	2401	528	1873	250.01	100.01	150.01	50.01	49	F2
3	20140301	3	3	10	4250	893	3357	250.02	100.02	150.02	50.02	34	F3
4	20140401	5	6	5	180	49	131	250.03	100.03	150.03	50.03	3	F4
5	20140501	2	7	5	498	85	413	250.04	100.04	150.04	50.04	83	F5
6	20140601	1	9	8	603	12	591	250.05	100.05	150.05	50.05	9	F6
7	20140701	3	1	7	8100	1782	6318	250.06	100.06	150.06	50.06	100	F7
8	20140801	3	8	9	1976	138	1838	250.07	100.07	150.07	50.07	38	F8
9	20140901	1	10	7	5452	1199	4253	250.08	100.08	150.08	50.08	58	F9
10	20141001	5	10	1	8832	707	8125	250.09	100.09	150.09	50.09	69	F10
11	20141101	2	5	6	7257	2032	5225	250.10	100.10	150.10	50.10	59	F11
12	20141201	5	5	3	2312	324	1988	250.11	100.11	150.11	50.11	17	F12
13	20150101	3	1	3	5110	153	4957	250.12	100.12	150.12	50.12	70	F13
14	20150201	5	3	1	4611	415	4196	250.13	100.13	150.13	50.13	53	F14
15	20150301	5	4	10	660	73	587	250.14	100.14	150.14	50.14	33	F15
16	20150401	1	7	4	4752	950	3802	250.15	100.15	150.15	50.15	44	F16
17	20150501	5	9	9	9715	291	9424	250.16	100.16	150.16	50.16	67	F17
18	20150601	4	7	7	1185	249	936	250.17	100.17	150.17	50.17	79	F18
19	20150701	5	2	6	2100	210	1890	250.18	100.18	150.18	50.18	28	F19
20	20150801	1	1	5	252	28	224	250.19	100.19	150.19	50.19	63	F20
21	20150901	4	5	6	9504	1521	7983	250.20	100.20	150.20	50.20	66	F21
22	20151001	5	4	5	564	45	519	250.21	100.21	150.21	50.21	12	F22
23	20151101	3	3	3	600	90	510	250.22	100.22	150.22	50.22	40	F23
24	20151201	5	10	6	4089	1186	2903	250.23	100.23	150.23	50.23	47	F24

- Afin de rendre le plus réaliste possible notre document, vous devez obtenir un nombre de lignes suffisamment important. Démultipliez alors ces 24 lignes par copier-coller jusqu'à obtenir le nombre de lignes souhaitées.
- Pour finir, enregistrez le document sous format Excel afin de conserver vos formules. Puis enregistrez ce fichier également au format CSV (séparateur point-virgule) à son emplacement originel (remplacez le fichier précédent). Enfin, fermez Excel.

Le fichier Excel de démonstration et le fichier CSV en résultant sont téléchargeables sur la page Informations générales.

- À l'invite de commande, tapez la commande qui permet d'intégrer les données du fichier CSV dans la table *FactFacture* :

```
bcp DistrisysDW.dbo.FactFacture in "FactFacture.csv" -T -c -t";"
-S"Serveur\Instance"
```

```

Administrateur : Invite de commandes
Microsoft Windows [version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. Tous droits réservés.

C:\Users\sfantini>cd\

C:\>bcp DistrisysDW.dbo.FactFacture in "FactFacture.csv" -I -c -t";" -S"Serveur1"
Dénarrage de la copie...
1000 lignes envoyées vers SQL Server. Total envoyé : 1000

1908 lignes copiées.
Taille du paquet réseau (octets) : 4096
Heure (ms) Total : 1781 Moyenne : (1071.31 lignes par seconde)
C:\>

```

La commande s'est exécutée avec succès. Nous venons de réintégrer 1908 lignes à notre table de faits *FactFacture* :

	DateFacturation_FK	Site_FK	Produit_FK	Client_FK	PrixCatalogue	Remise	CA	Marge	CoutDirectMatiere	CoutD
▶	20140101	3	2	2	53167,00	6380,00	46787...	14036...	8188,00	18996,
	20140201	3	1	10	62426,00	18104,00	44322...	13297...	9308,00	12720,
	20140301	2	5	9	7462,00	1866,00	5596,00	1343,00	1701,00	2212,0
	20140401	3	5	8	14781,00	4434,00	10347...	2069,00	2566,00	4967,0
	20140501	5	1	9	45323,00	1360,00	43963...	15827...	8159,00	15475,
	20140601	1	10	4	49764,00	14432,00	35332...	11660...	9469,00	12309,
	20140701	1	2	2	9176,00	2753,00	6423,00	1670,00	1568,00	2472,0
	20140801	2	3	4	18864,00	4527,00	14337...	5018,00	3541,00	4939,0
	20140901	2	7	9	48393,00	3871,00	44522...	12466...	11220,00	18592,
	20141001	2	6	5	10368,00	1348,00	9020,00	3157,00	1818,00	3283,0
	20141101	3	2	3	24960,00	5741,00	19219...	4805,00	4612,00	8504,0
	20141201	1	4	10	16275,00	3743,00	12532...	3634,00	3470,00	4894,0
	20150101	4	10	10	11640,00	2910,00	8730,00	2706,00	2048,00	2651,0
	20150201	2	9	9	26350,00	5797,00	20553...	5960,00	4378,00	8610,0
	20150301	4	8	4	63540,00	2542,00	60998...	16469...	17366,00	18702,
	20150401	3	4	2	43335,00	1300,00	42035...	9668,00	10681,00	13270,

La réexécution de la commande BCP n'efface pas le contenu de la table, mais ajoute d'autres lignes aux lignes déjà existantes. Cela vous permet ainsi de créer des jeux de tests de plusieurs millions de lignes, si nécessaire.

Suite à cette section, vous devrez donc être en mesure de créer vos propres jeux de test parfaitement personnalisés à votre problématique et vous permettant ainsi de vous assurer la validité de votre modélisation.

Dans la prochaine section nous allons apprendre à construire un cube. Le cube va nous permettre d'explorer facilement les données que nous venons de charger.