Un petit rappel sur l'opération de regroupement (GROUP BY)

Il arrive fréquemment que l'on souhaite regrouper des données possédant certaines similitudes. Considérons l'exemple d'une table recensant, pour chaque employé d'une entreprise, le nombre d'heures qu'il a effectuées tel ou tel jour de cette semaine :

Nom	Date	Heures
Marc	13/02/2012	8
Jean	13/02/2012	7
Marc	14/02/2012	7
Jean	14/02/2012	6
Marc	15/02/2012	8
Jean	15/02/2012	4
Marc	16/02/2012	6
Jean	16/02/2012	5
Marc	17/02/2012	8
Jean	17/02/2012	6

Remarquez qu'il existe une forte redondance sur le champs *Nom*, plusieurs enregistrements ayant une valeur similaire pour celui-ci. Il semble alors judicieux de regrouper en un tous les enregistrements concernant un même employé pour pouvoir ensuite obtenir plus d'informations sur les groupes résultant. Pour cela, on utilise, en SQL, le mot-clef *GROUP BY* en spécifiant sur quel(s) critère(s) les groupes doivent être créés.

Dans notre exemple, la commande

SELECT C.Nom
FROM HeuresTravail C
GROUP BY C.Nom

renverrait quelque chose qui peut s'apparenter à cela :

Nom	Champs	hybride
Marc	*Date*	*Heures*
	13/02/2012	8
	14/02/2012	7
	15/02/2012	8
	16/02/2012	6
	17/02/2012	8
Jean	*Date*	*Heures*
	13/02/2012	7
	14/02/2012	6
	15/02/2012	4
	16/02/2012	5
	17/02/2012	6

Une fois la sélection des données effectuée, nous avons demandé au SGBD de regrouper les données concernant un même employé. Notez la présence du « champs hybride » composé de plusieurs sous-champs. Typiquement, un SGBD ne sait pas afficher un tel champs ; ainsi, une requête telle que

```
SELECT C.Date
FROM HeuresTravail C
GROUP BY C.Nom
```

n'est pas correcte puisque, une fois le regroupement effectué, le champs *Date* n'existe plus réellement ! En revanche, il est toujours possible de sélectionner le champs *Nom* puisqu'il reste un champs atomique. C'est pourquoi **tout ce qui est sélectionné doit impérativement être un critère de regroupement**. Remarquez par ailleurs que le fait de regrouper retire les doublons dans les enregistrements ; sélectionner les noms des groupes est alors équivalent à la requête suivante.

```
SELECT DISTINCT C.Nom
FROM HeuresTravail C
```

Les données regroupées au sein d'un groupe n'étant plus accessibles, on peut alors se demander l'utilité du regroupement puisqu'il semble nous faire perdre de l'information. La motivation principale est de se servir d'informations sur les groupes pour pouvoir trier sur des critères qu'un *WHERE* classique ne peut pas considérer. Ainsi, on peut créer des groupes suivant certains champs clefs, et ne garder que les clefs dont les groupes associés respectent telle ou telle contrainte.

L'affinement sur les groupes s'effectue en utilisant le mot-clef *HAVING*, qui est l'équivalent du *WHERE* mais réservé aux groupes. Ainsi, un *HAVING* ne peut pas être utilisé sans *GROUP BY*. Attention : comme spécifié plus haut, il n'est pas possible d'utiliser un critère portant sur l'un des sous-champs du champs hybride puisqu'une fois les groupes créés, ils n'existent plus pour le SGBD. Les seules possibilités sont alors de restreindre sur l'un des champs ayant permis à réaliser le regroupement, ou sur l'une des propriétés des groupes obtenues par le biais d'une *fonction d'agrégation* (*SUM*, *COUNT*, *MIN*, *MAX*, *AVG*, etc.).

Par exemple, la requête suivante permet d'obtenir la liste des employés ayant travaillé plus de 35h lors de la semaine considérée.

```
SELECT C.Nom
FROM HeuresTravail C
GROUP BY C.Nom
HAVING SUM(C.Heures) >= 35
```