Analyse des données

CM6 : SQL avancé et OLAP

Mickaël Martin Nevot

V1.0.0

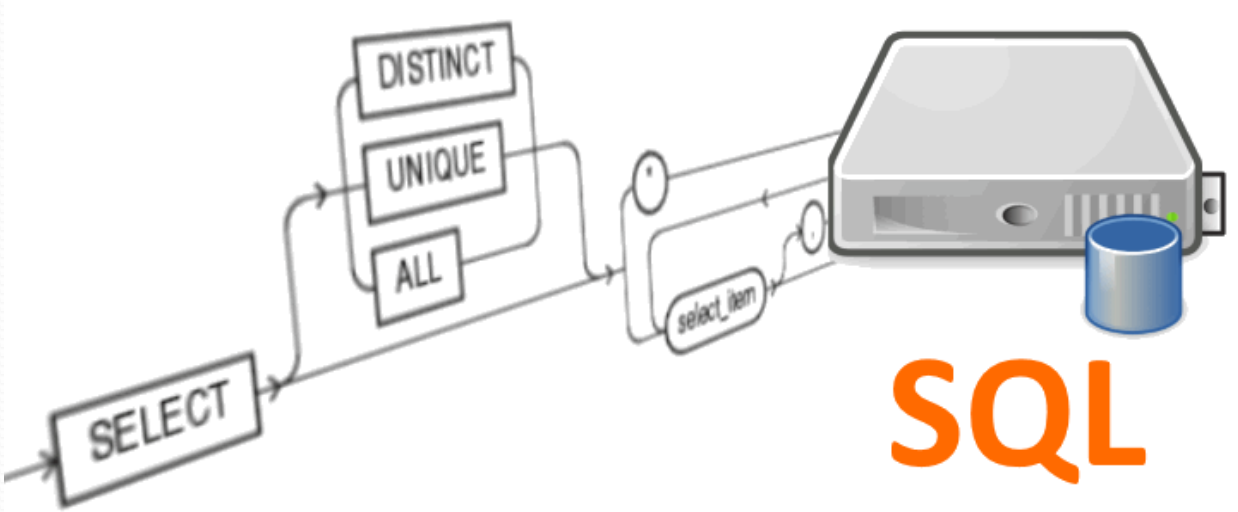
Analyse des données

Rappel : SQL

- *Structured query language* : langage de requête structurée
- **Un seul langage** général :
  - Langage de description de données (LDD)
  - **Langage de manipulation de données (LMD)**
  - Langage de description des schémas physiques (LDSP)
  - Contrôle et administration

Rappel : LMD

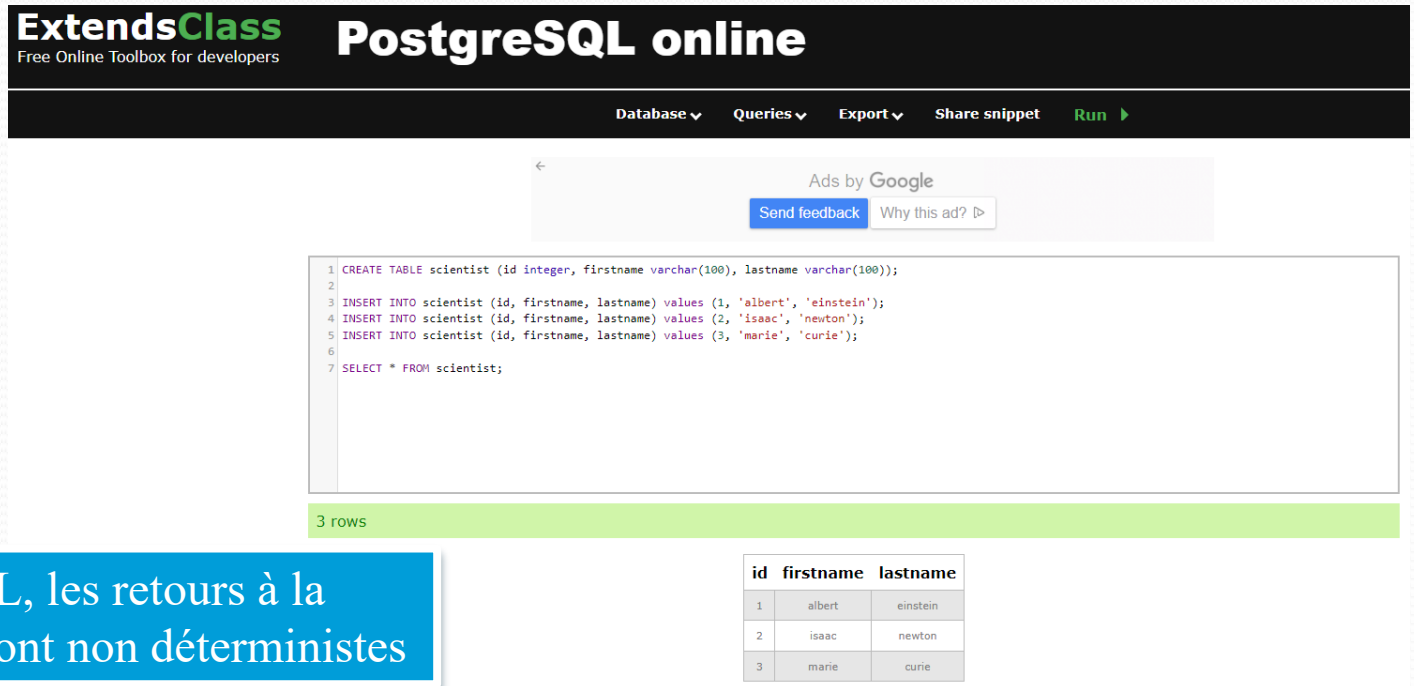
- **Langage de manipulation de données** (LMD)
- Recherche des données d'une BD
- Mise à jour des données d'une BD



Le résultat d'une requête relationnelle est une relation

Pour tester

- Interpréteur en ligne : https://extendsclass.com/postgresql-online.html
- Documentation : https://www.postgresql.org/docs/current



En SQL, les retours à la ligne sont non déterministes

Rappel : Recherche de données

- Sélection :
  - SELECT : définit les attributs de la relation résultante
  - FROM : spécifie les relations sur lesquelles porte la recherche
  - WHERE : indique des conditions de restriction
  - Les autres clauses seront traitées ultérieurement

Syntaxe :

```
SELECT [DISTINCT] ...
FROM ...
[WHERE ...]
[GROUP BY ...]
[HAVING ...]
[ORDER BY ...]
```

SELECT, FROM, WHERE, etc. sont des clauses

Rappel : Fonctions d'agrégation

- Agrège les tuples (un seul résultat)
- Uniquement dans clause `SELECT` (ou `HAVING`)
- Fonctions :
  - `SUM(...)` : somme
  - `AVG(...)` : moyenne algébrique
  - `MIN(...)`, `MAX(...)` : valeur min., max.
  - `COUNT(...)` : dénombrement
  - Etc.

Uniquement pour des attributs numériques

Aussi appelées fonctions de calcul intégrées

Pas de fonction d'agrégat en paramètre d'une fonction d'agrégat

Rappel : Partitionnement

- Partitionner les données afin d'effectuer des calculs par ensemble de données groupées :

```
SELECT prenom, COUNT(*)
FROM Etudiant WHERE sexe = 'M'
GROUP BY prenom;
```

On obtient autant de partitions que de valeurs distinctes dans l'ensemble d'attributs de la clause `GROUP BY`

- Partitionnement avec condition de sélection :

```
SELECT prenom, COUNT(*)
FROM Etudiant WHERE sexe = 'M'
GROUP BY prenom
HAVING COUNT (DISTINCT ide) >= 2;
```

Les agrégations s'appliquent à chaque valeur de l'ensemble de la clause `SELECT`

Règle d'or : tous les attributs non agrégés projetés dans un `SELECT` **doivent** figurer dans le `GROUP BY`, et **inversement**

Rappel : Partitionnement

R	A	B	C
	a1	b1	c1
	a1	b1	c2
	a2	b1	c1
	a2	b2	c1
	a1	b2	c1
	a2	b1	c1

```
SELECT A, B, COUNT(*)
FROM R
GROUP BY A, B;
```

T1	A	B	(*)
	a1	b1	2
	a1	b2	1
	a2	b1	2
	a2	b2	1

```
SELECT C, A, COUNT(*)
FROM R
GROUP BY C, A;
```

T2	C	A	(*)
	c1	a1	2
	c1	a2	3
	c2	a1	1

OLAP : agrégation étendue

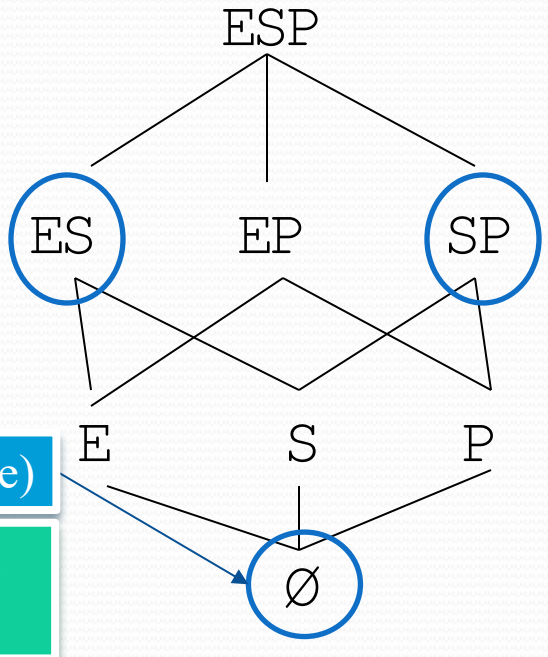
- `GROUP BY` ne construit **qu'une seule partition** des résultats
- Possibilité d'y adjoindre des opérateurs afin de visualiser **plusieurs partitions** en même temps

OLAP *(online analytical processing)* : traitement analytique en ligne (couramment utilisé en informatique décisionnelle) permet l'analyse sur-le-champ d'informations selon plusieurs axes, dans le but d'obtenir des rapports de synthèse

OLAP : sélection de partitions

```
-- Donner les durées totales des stages en entreprise effectués d'une façon générale ; par
-- étudiant et par société ; et enfin par société et par tuteur.
SELECT ide, ids, idp, SUM(duree) AS dureet
FROM Convention
GROUP BY GROUPING SETS((), (ide, ids), (ids, idp));
```

ide	ids	idp	dureet
			41
17	8		4
8	21		5
8	8		3
17	34		6
12	13		6
8	34		6
15	13		5
15	8		6
	13	12	11
	8	2	3
	8	4	6
	8	7	4
	21	19	5
	34	53	12



Pas de partition (partition vide)

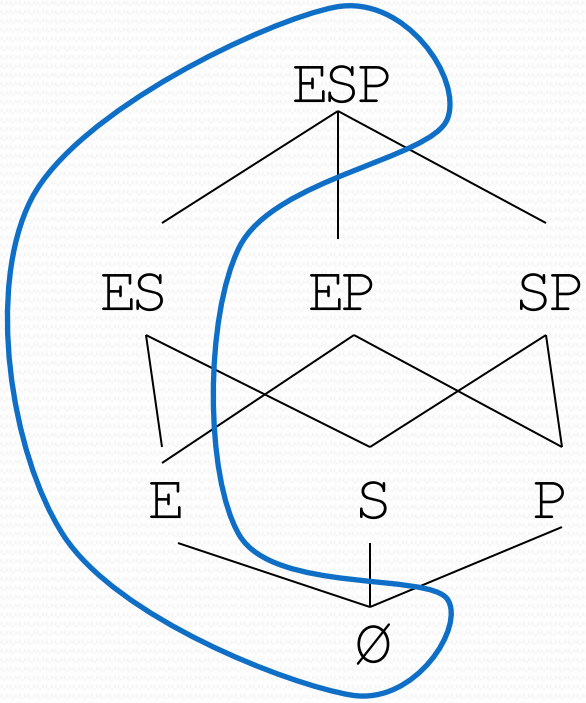
Les résultats obtenus forment un treillis de données

E : ide
S : ids
P : idp
Ø : absence de partition

OLAP : chemin de partitions

```
-- Donner les durées totales des stages en entreprise effectués d'une façon générale ; par
-- étudiant ; par étudiant et par société ; et enfin par étudiant, par société et par tuteur.
SELECT ide, ids, idp, SUM(duree) AS dureet
FROM Convention
GROUP BY ROLLUP(ide, ids, idp);
```

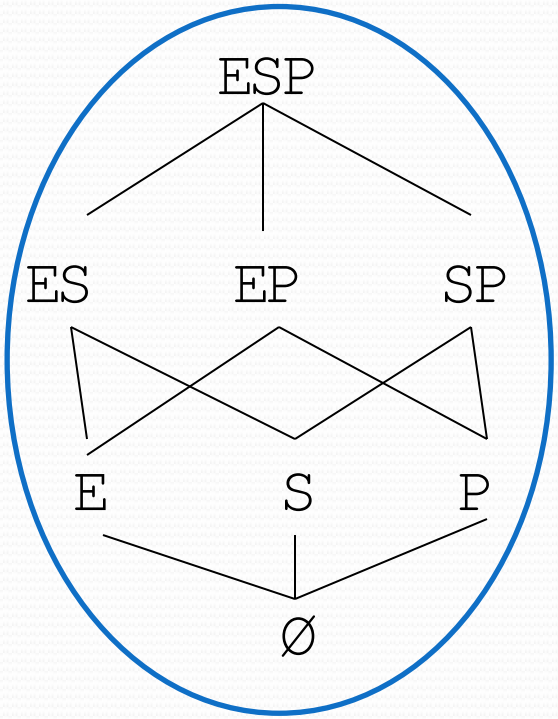
ide	ids	idp	dureet
			41
15	13	12	5
8	21	19	5
17	8	7	4
8	8	2	3
15	8	4	6
8	34	53	6
12	13	12	6
17	34	53	6
17	8		4
8	21		5
8	8		3
17	34		6
12	13		6
8	34		6
15	13		5
15	8		6
15			11
17			10
12			6
8			14



E : ide
S : ids
P : idp
Ø : absence de partition

OLAP : cube (toutes les partitions)

```
-- Donner les durées totales des stages en entreprise effectués d'une façon générale ; par
-- étudiant ; par société ; par tuteur ; par étudiant et par société ; par étudiant et par
-- tuteur ; par société et par tuteur ; et enfin par étudiant, par société et par tuteur.
SELECT ide, ids, idp, SUM(duree) AS dureet
FROM Convention
GROUP BY CUBE(ide, ids, idp);
```

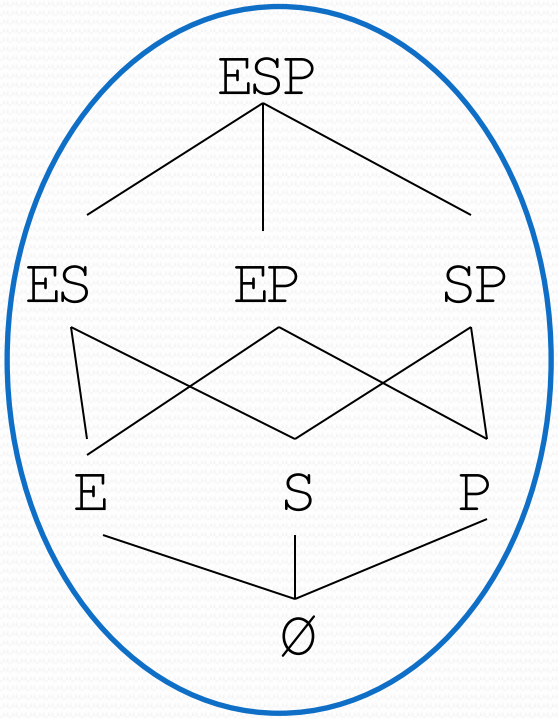


45 résultats

E : ide
S : ids
P : idp
Ø : absence de partition

OLAP : cube iceberg

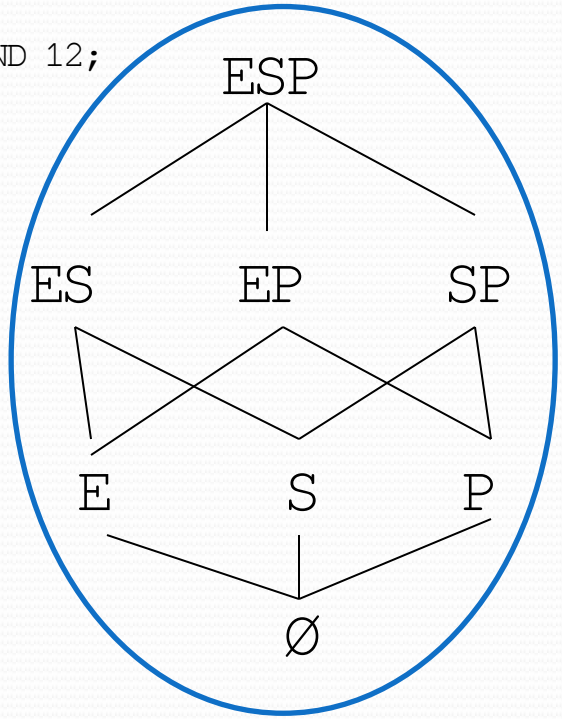
```
-- Donner les durées totales des stages de sept mois ou plus en entreprise effectués d'une
-- façon générale ; par étudiant ; par société ; par tuteur ; par étudiant et par société ;
-- par étudiant et par tuteur ; par société et par tuteur ; et enfin par étudiant, par
-- société et par tuteur.
SELECT ide, ids, idp, SUM(duree) AS dureet
FROM Convention
GROUP BY CUBE(ide, ids, idp)
HAVING SUM(duree) >= 7;
```



ide	ids	idp	dureet
			41
15			11
17			10
8			14
	13	12	11
	34	53	12
	34		12
	13		11
	8		13
		53	12
		12	11

OLAP : cube intervallaire

```
-- Donner les durées totales des stages de sept à douze mois en entreprise effectués d'une
-- façon générale ; par étudiant ; par société ; par tuteur ; par étudiant et par société ;
-- par étudiant et par tuteur ; par société et par tuteur ; et enfin par étudiant, par
-- société et par tuteur.
SELECT ide, ids, idp, SUM(duree) AS dureet
FROM Convention
GROUP BY CUBE(ide, ids, idp)
HAVING SUM(duree) BETWEEN 7 AND 12;
```

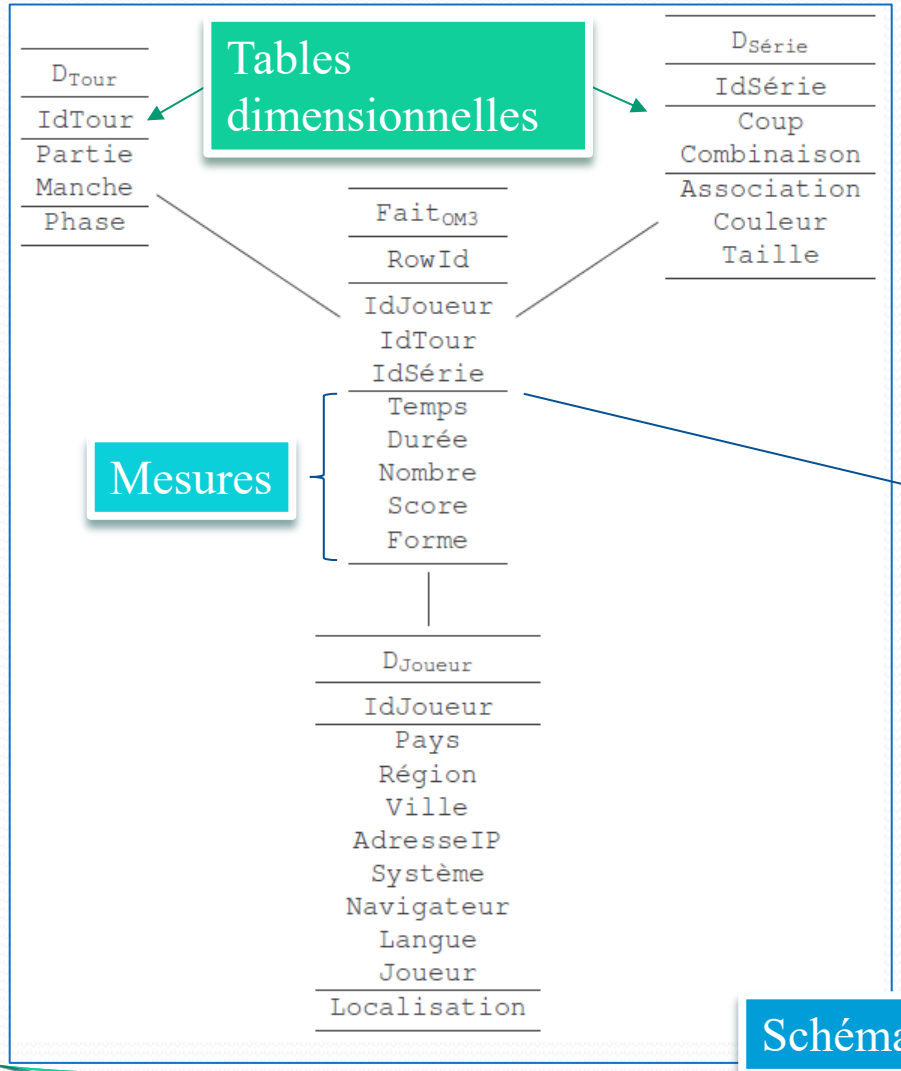


ide	ids	idp	dureet
15			11
17			10
	13	12	11
	34	53	12
	34		12
	13		11
		53	12
		12	11

Analyse multidimensionnelle

- Cube de données (hypercube OLAP) :
  - Union des résultats des requêtes agrégatives de partitionnement (avec `GROUP BY`) sur toutes les combinaisons possibles des attributs dimensions
  - Un (pré-)calcul permet des réponses rapides à toutes requêtes utiles mais crée aussi beaucoup d'opérations, et donc d'espace de stockage

Entrepôt de données



RowId	IdJ	IdT	IdS	Temps	Durée	Nombre	Score	Forme
1	J_1	P_1	A_1	6.32	2.85	3.5	700	0.5
2	J_1	P_1	A_2	18.9	1.95	3.83	2300	0.5
3	J_2	P_2	A_3	26.39	1.7	3.43	2400	0.71
4	J_2	P_2	A_4	4.1	2.07	3	300	0.5
5	J_2	P_2	A_5	7.38	3.68	3	600	0.5
6	J_3	P_3	A_6	2.14	2.15	3	300	1
7	J_3	P_3	A_7	56.04	2.25	3.17	3800	0.5

Table des faits (transactionnelle)

Table des faits récapitulative

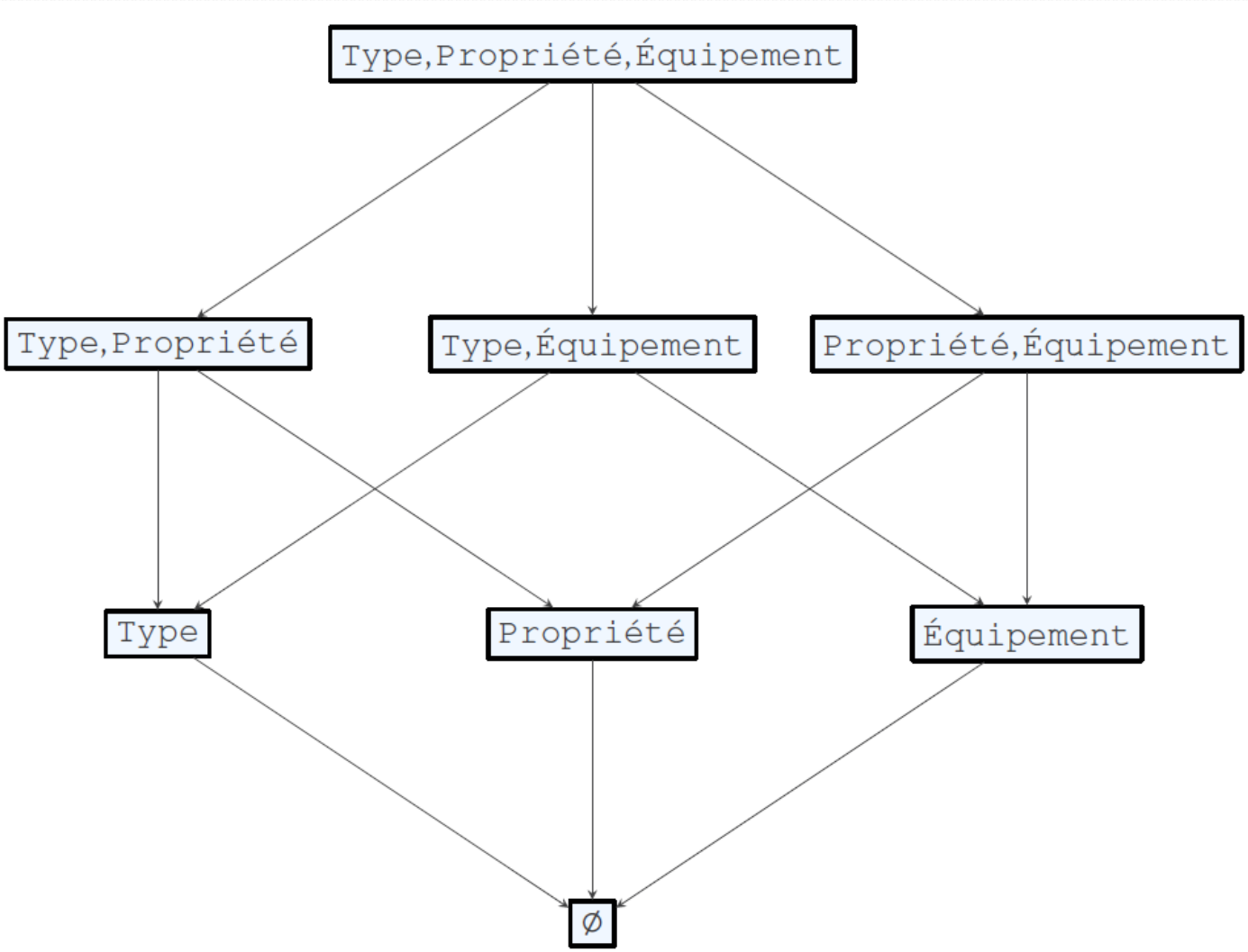
Avec agrégation de données, ou mesure

RowId	Type	Propriété	Équipement	Quantité[a]
1	*Féticheur*	*Vitesse*[b]	*Bastion*[c]	400
2	*Féticheur*	*Vitesse*	*Marche*[d]	100
3	*Sorcière*	*Chance*[e]	*Marche*	100
4	*Sorcière*	*Vitesse*	*Bastion*	300
5	*Croisé*	*Chance*	*Bastion*	200

a. Nombre arrondi de personnages dans le classement pour l'Europe.
b. Vitesse d'attaque.
c. Le Bastion de la volonté.
d. La Marche sans fin.
e. Chance de coup critique.

Jeu vidéo Diablo III : Reaper of Souls

Exemple de cuboïdes



Exemple de cuboïdes détaillés

Type	Propriété	Équipement	Compagnon	Quantité
Féticheur	Vitesse	Bastion	Enchanteresse	400
Féticheur	Vitesse	Marche	Templier	100
Sorcière	Chance	Marche	Enchanteresse	100
Sorcière	Vitesse	Bastion	Templier	300
Croisé	Chance	Bastion	Enchanteresse	200

T	P	É	SUM(Q)
Fétcheur	Vitesse	Bastion	400
Fétcheur	Vitesse	Marche	100
Sorcière	Chance	Marche	100
Sorcière	Vitesse	Bastion	300
Croisé	Chance	Bastion	200

T	P	SUM(Q)
Fétcheur	Vitesse	500
Sorcière	Chance	100
Sorcière	Vitesse	300
Croisé	Chance	200

T	É	SUM(Q)
Fétcheur	Bastion	400
Fétcheur	Marche	100
Sorcière	Marche	100
Sorcière	Bastion	300
Croisé	Bastion	200

P	É	SUM(Q)
Vitesse	Bastion	700
Vitesse	Marche	100
Chance	Marche	100
Chance	Bastion	200

T	SUM(Q)
Fétcheur	500
Sorcière	400
Croisé	200

P	SUM(Q)
Vitesse	800
Chance	300

É	SUM(Q)
Bastion	900
Marche	200

SUM(Q)
1100

Crédits

Auteur

Mickaël Martin Nevot
mmartin.nevot@gmail.com

- Laurent Carmignac

Carte de visite électronique

Relecteurs

Cours en ligne sur : www.mickael-martin-nevot.com