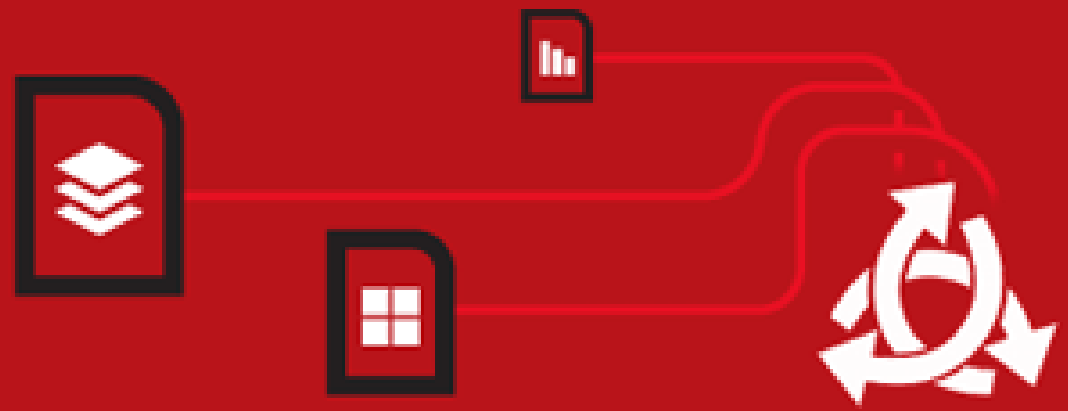
Les journées
SQL Server 2014



Un événement organisé par GUSS

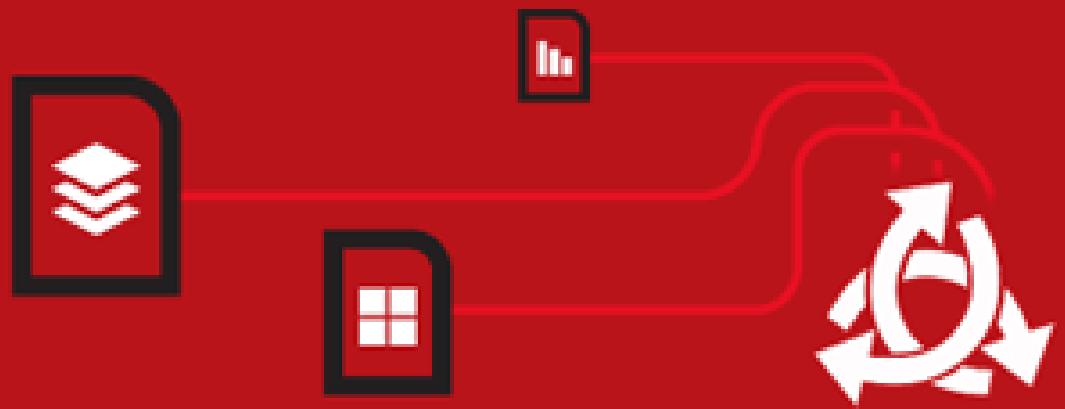
#JSS2014

Les journées

SQL Server 2014

Les statistiques

Arian Papillon, MVP SQL Server
Frédéric Brouard, MVP SQL Server



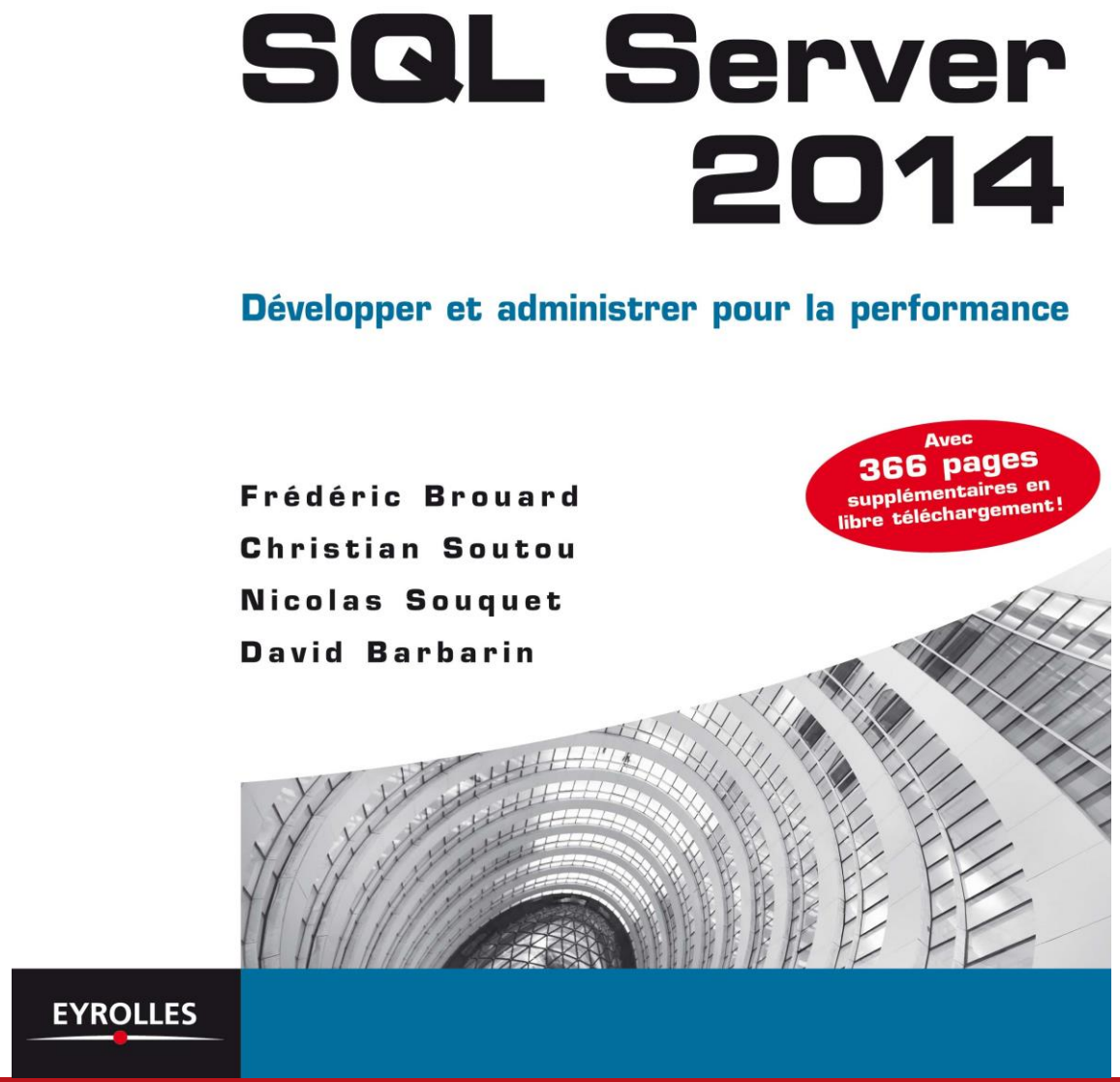
Merci à nos sponsors

GUSS

Les journées SQL Server

- Arian Papillon
  - a.papillon@datafly.fr
- Frédéric Brouard
  - sqlpro@sqlspot.com

Parution : 31/12/2014



1 232 pages !

- **Partie I. Création des objets et manipulation des données**
- **Partie II. Programmation avancée : Transact-SQL, vues, .NET**
- **Partie III. Gestion des bases de données : stockage, sécurité, sauvegarde…**
- **Partie IV. Maintien des performances**
- **Partie V. Administration du serveur**

Agenda

- A quoi servent les statistiques
- La création des statistiques
- Comment sont-elles utilisées par l'optimiseur
- La maintenance des statistiques
- Cas particuliers
- Bonnes pratiques

A quoi servent les statistiques

Il y a les mensonges, les gros mensonges…
et les statistiques !

Benjamin Disraéli, première ministre anglais,
lorsqu'on lui présenta les statistiques vers 1870

Requête et contexte

- Table des employés d'une maternité, requête :

```
SELECT * FROM Employés
WHERE Salaire > 2000
AND Sexe = 'femme'
```

- Quel critère prendre en priorité ?

Même requête, autre contexte

- Même requête dans une mine :

```
SELECT * FROM Employés
WHERE Salaire > 2000
AND Sexe = 'femme‘
```

- Et maintenant ?

L'optimiseur

- L'optimisation statistique des requêtes est basée sur le coût
  - Moins il y a de données à traiter, plus c'est rapide
- Pour élaborer un plan d'exécution efficace, l'optimiseur doit donc estimer le nombre de lignes traitées par chaque opérateur : c'est **l'estimation de la cardinalité**.
  - Cardinalité : nombre de lignes traitées par un opérateur de requête

La estimación de la cardinalitad*

* L'estimation de la cardinalité

- L'estimation des cardinalités affectera le plan d'exécution
  - Les modes d'accès aux données et le choix (ou non) d'un index
  - Le choix des opérateurs : certains sont meilleurs pour traiter peu de lignes, d'autres pour beaucoup de lignes
  - La taille de la mémoire nécessaire pour l'exécution
  - Le parallélisme
- SQL Server se sert des **statistiques** pour faire cette estimation
  - Il peut aussi s'aider des contraintes (check, unique, primary key, foreign key – *optimisation sémantique*)

Un peu de vocabulaire

- Prédicat
  - Expression logique qui renvoie VRAI, FAUX ou INCONNU
  - Exemple : évaluation d'une clause WHERE, HAVING, ON (join)
- Densité
  - S'applique à une colonne ou un jeu de colonnes
  - Mesure le taux de doublons
  - Formule : 1/nombre de valeurs uniques
- Sélectivité
  - S'applique à un prédicat
  - Mesure le pourcentage de la table qui correspond au prédicat

Statistique

- Une statistique (sur une ou plusieurs colonnes d'une table) contient des informations permettant d'évaluer le nombre de lignes retournées pour un prédicat (une valeur ou une plage de valeurs)
- Elle contient plusieurs informations :
  - Nombre de lignes total
  - Taille moyenne de la colonne
  - Densité moyenne
  - Histogramme de répartition des valeurs, basé sur un échantillon
  - Statistiques de résumé de chaînes (pour colonne caractère)
- Ces informations sont capturées au moment de la création ou de la mise à jour.

Que contiennent les statistiques ?

Entête

Name	Updated	Rows	Rows Sampled	Steps	Density	Average key length	String Index	Filter Expression	Unfiltered Rows
nix$client$nom	nov 22 2014 3:25PM	1018774	1018774	200	0,6233087	29,72929	YES	NULL	1018774

All density	Average Length	Columns
1,576183E-06	25,72929	Nom
9,815719E-07	29,72929	Nom, IdClient

Vecteur de densité

RANGE_HI_KEY	RANGE_ROWS	EQ_ROWS	DISTINCT_RANGE_ROWS	AVG_RANGE_ROWS
A. Scott Wright	0	1	0	1
Adam Khan	4788	7	2940	1,628571
Adrienne Rush	3543	6	2133	1,661041
Albert Walter	5471	6	3396	1,611013
Alexandra Hall	4742	6	2943	1,611281
Alfred Nguyen	3558	7	2254	1,578527
Alicia Howard	2974	6	1835	1,620708
Allan Owens	5426	6	3389	1,601062

Histogramme

Comment les voir ?

- SSMS
- DBCC SHOW_STATISTICS
  - Ex : DBCC SHOW_STATISTICS (MaTable, _WA_Sys_00000002_2A4B4B5E)
- Vues système, DMV's, fonction
  - Vues sys.stats, sys.stats_columns
  - DMV sys.dm_db_stats_properties(object_id,stat_id)
    - A partir de SQL 2008 R2 SP2 / SQL 2012 SP1
  - Fonction stats_date(object_id, stats_id)

Démo

Création des statistiques

Les sympathisant de droite ont une espérance de vie de 5 années supérieures à ceux de gauche…

Sondage anglais début des années 2000

MORALITÉ : pour vivre plus vieux, votez à droite !

Création des statistiques

- Les statistiques sont créées :
  - Automatiquement pour chaque index, à la création
  - Automatiquement sur les colonnes non indexées qui sont interrogées
    - si l'option auto_create_statistics est activée pour la base de données
    - statistiques mono-colonne uniquement
    - _WA_Sys_xxxxx
  - Ou manuellement
    - Instruction CREATE STATISTICS
    - Procédure sp_createstats

CREATE STATISTICS

- Les statistiques peuvent être :
  - Multi-colonnes (jusqu'à 32 colonnes)
  - Filtrées : clause WHERE
  - Incrémentales (SQL 2014) : mises à jour par partition
- On peut choisir l'échantillonnage :
  - FULLSCAN : scanne toute la table
  - ou SAMPLE : en pourcentage ou en nombre de lignes

```
CREATE STATISTICS statistics_name
ON { table_or_indexed_view_name } ( column [ ,...n ] )
    [ WHERE <filter_predicate> ]
    [ WITH
        [ [ FULLSCAN | SAMPLE number { PERCENT | ROWS }
        [ [ , ] NORECOMPUTE ]
        [ [ , ] INCREMENTAL = { ON | OFF } ]
    ] ;
```

sp_createstats

- Crée une statistique sur chaque colonne de chaque table (sauf si statistique déjà existante)
  - Peu souvent utile, car la création automatique peut suffire à faire ce travail

```
EXEC sp_createstats
    [    [ @indexonly =    ] { 'indexonly'    | 'NO' } ]
    [ , [ @fullscan =      ] { 'fullscan'     | 'NO' } ]
    [ , [ @norecompute =   ] { 'norecompute'  | 'NO' } ]
    [ , [ @incremental =   ] { 'incremental'  | 'NO' } ]
```

Statistiques dupliquées

- Lors de la création d'un index, la création des statistiques ne tient pas compte de l'existence préalable de statistiques de colonnes.
- Nettoyez les statistiques en double…

Comment l'optimiseur utilise les statistiques

73% des accidents de la route surviennent à proximité du domicile

Source : INRETS, http://www.roulons-autrement.com/videos/voir/les-accidents-de-la-route-surviennent-a-proximite-du-domicile-0404

Conclusion : allez habiter chez les autres !

Expérimentations

- Ici nous regardons comment l'optimiseur utilise les statistiques
  - Egalité, inégalité, prédicats multiples, jointure…
- Avec SQL 2014, certains modes de calcul pour l'estimation des cardinalités sont modifiés
  - est-ce réellement mieux ?

Démo

Egalité avec une constante

```
WHERE Nom = 'Constante'
```

- A la compilation, l'optimiseur connaît la valeur recherchée
- Il se sert de l'histogramme

Egalité avec une variable

```
WHERE Nom = @Variable
```

- A la compilation, l'optimiseur ne connaît pas la valeur recherchée
  - Valeur inconnue -> OPTIMIZE FOR UNKNOWN
- Il se sert de la densité
- Le SQL dynamique fonctionne mieux !

Inégalité

```
WHERE Colonne > | < | <>| BETWEEN
```

- Avec valeurs connues (constantes)
  - se sert de l'histogramme pour calculer l'estimation
- Sans valeur connues (variables)
  - La densité ne l'aide pas
  - Calcul générique
    - Une borne : 30 % de la table
    - Deux bornes :
      - SQL 2012 : 9 % de la table
      - SQL 2014 : ??? - nouveau calcul !

Absence de statistiques

- Egalité : formule générique
  - SQL 2012 : √(NbLignes * √(NbLignes))
  - SQL 2014 : √(NbLignes)
- Inégalité : calcul « à la louche »
  - 30% de la table… !

Prédicats multiples (plusieurs colonnes)

- Combine les cardinalités des prédicats
  - Avant SQL 2014, le calcul de l'optimiseur considère que les prédicats sont indépendants
  - Cela peut provoquer des sous-estimations si les prédicats sont corrélés : WHERE Ville='Paris' AND CodeDept = '75'
- A partir de SQL 2014, le mode de calcul a évolué

Jointure

- Combine les cardinalités des différentes tables
  - Utilise l'histogramme ou la densité
- Une mauvaise estimation peut dégrader l'ensemble du plan d'exécution
  - Attention si la distribution des données est mal répartie
  - Les jointures successives héritent de la mauvaise estimation
- Pour résoudre le problème on peut parfois utiliser des statistiques filtrées

Statistique filtrée

- Intéressante si couvrant une requête
  - Cas du booléen : *commande active*
  - Critère fréquemment demandé : *devise euro*
  - …

La gestion des statistiques

Les médecins font-il mourir ?

En 1982 en Israël, lors d'une grève des médecins de plusieurs semaines, on constata une baisse sensible des décès !

Source : Joseph Klatzmann – Attention statistiques !

Les statistiques sont-elles exactes ?

- *Les statistiques, c'est comme le bikini : ça donne des idées mais ça cache l'essentiel (Coluche)*

Echantillonnage par défaut

- L'échantillon par défaut dépend de la taille de la table :
  - Si < 8MB : FULLSCAN
  - Si > 8MB : selon un algorithme qui réduit l'échantillonnage selon le nombre de lignes de la table
- Plus la table est grosse, plus l'échantillon est grossier…

Distribution inégale

- Les statistiques, même à jour, peuvent être imprécises :
  - Il n'y a que 200 étapes (maximum) dans l'histogramme
  - La densité est calculée sur toute la table

Mise à jour automatique des statistiques

- La mise à jour automatique des statistiques se déclenche lorsque le nombre de modifications dépasse 20% des lignes de la table (+500 lignes), avec au moins 500 modifications
  - L'échantillonnage par défaut est utilisé
  - La requête (compilation) déclenche la mise à jour :
    - Immédiate
    - ou asynchrone si l'option de base de données AUTO_UPDATE_STATISTICS_ASYNC est activée (édition Entreprise)
- Le traceflag 2371 permet de rendre dynamique le seuil de déclenchement pour les grosses tables
  - Plus il y a de lignes, plus le seuil est bas

Mise à jour manuelle des statistiques

- UPDATE STATISTICS

```
UPDATE STATISTICS table_or_indexed_view_name
    [ {{ index_or_statistics__name } | ( { index_or_statistics_name } [ ,...n ] )}]
    [    WITH
        [ FULLSCAN | SAMPLE number { PERCENT | ROWS } | RESAMPLE
              [ ON PARTITIONS ( { <partition_number> | <range> } [, …n] ) ]
        ]
        [ [ , ] [ ALL | COLUMNS | INDEX ]
        [ [ , ] NORECOMPUTE ]
        [ [ , ] INCREMENTAL = { ON | OFF } ]
    ] ;
```

- sp_updatestats
  - Pour toutes les tables (internes et utilisateur) de la base
  - Met à jour les statistiques dès lors qu'une ligne a été modifiée
  - RESAMPLE ou échantillon par défaut

Contrôler la mise à jour automatique

- Options de base de données
  - AUTO_UPDATE_STATISTICS ON|OFF
  - AUTO_UPDATE_STATISTICS_ASYNC ON|OFF
- Option NORECOMPUTE
  - CREATE STATISTICS, UPDATE STATISTICS
- Option STATISTICS_NORECOMPUTE
  - CREATE INDEX, ALTER INDEX
- sp_autostats

```
sp_autostats [ @tblname = ] 'table_or_indexed_view_name'
    [ , [ @flagc = ] 'stats_value' ]
    [ , [ @indname = ] 'statistics_name' ]
```

Cas particuliers

Statistiquement, chaque humain est doté de deux yeux, un cœur, deux bras, un cerveau, deux jambes…

… et d'un testicule !

Tables In Memory – Hekaton (2014)

- Aucune mise à jour automatique des statistiques
  - Les statistiques doivent être mises à jour manuellement avec les options FULLSCAN ou RESAMPLE et NORECOMPUTE
  - A la création de la table et des index, les statistiques sont créées mais non peuplées
  - La création automatique de statistiques de colonne fonctionne, mais pas de mise à jour automatique par la suite.
- Les procédures stockées en code natif doivent être recréées après mise à jour des statistiques

ColumnStore index

CREATE COLUMNSTORE INDEX…

- créé une entrée non renseignée de statistiques :

Name	Updated	Rows	Rows Sampled	Steps	Density	Average key length	String Index	Filter Expression	Unfiltered Rows
X_CTP_ITF_JOR_PRS_VAL_CS	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

All density	Average Length	Columns

RANGE_HI_KEY	RANGE_ROWS	EQ_ROWS	DISTINCT_RANGE_ROWS	AVG_RANGE_ROWS

statistiques de cardinalité par segments :

```
SELECT * FROM sys.column_store_row_groups
```

statistiques de volume et d'entrée dans les "dictionnaires"

```
SELECT * FROM sys.column_store_dictionaries
```

XML

Pas de statistiques

- Mais entrées fantômes dans la vue sys.stats
- Pensez à rendre votre index « sélectif » :
  - sys.sp_db_selective_xml_index

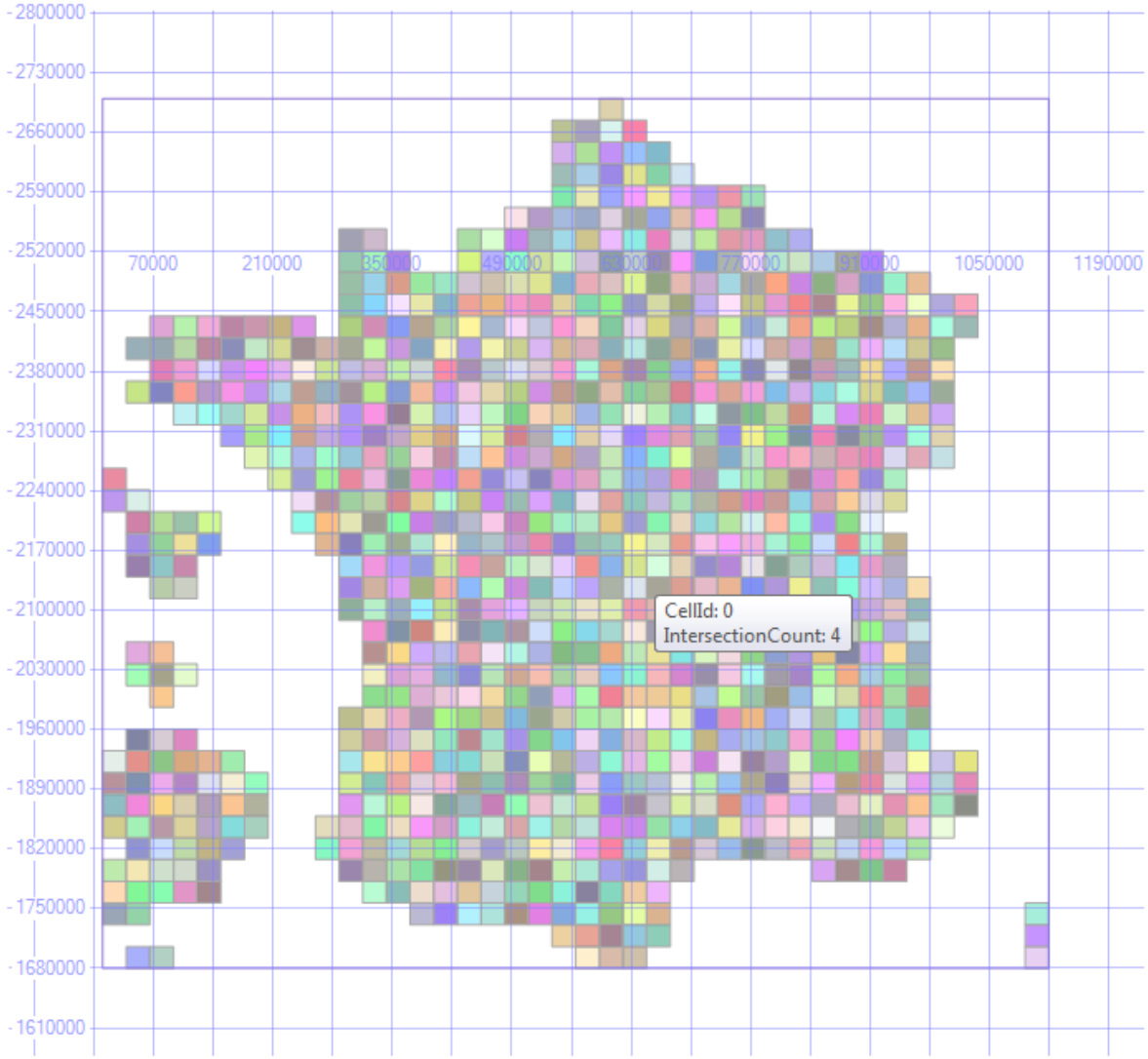
Spatial

- Statistiques collectées sur un « plan » ou une aire géodésique
- Visualisable par :

  sp_help_spatial_

  [geometry | *geography]*

  _histogram



FULLTEXT

Recherches « plain texte » :

- Pas de statistiques

Recherches sémantiques :

- Les statistiques sont incluses dans la base *semanticsdb*

Bonnes pratiques

la proportion de divorces parmi les ménages qui partagent les tâches domestiques de manière équitable est environ 50% plus élevée que chez ceux où l'essentiel du travail est accompli par la femme

Source : Thomas Hansen, institut Nova à Oslo 2012

Bonnes pratiques

- Conservez AUTO_CREATE_STATISTICS et AUTO_UPDATE_STATISTICS à ON (par défaut)
  - Testez AUTO_UPDATE_STATISTICS_ASYNC si la mise à jour automatique pose des problèmes
- Eliminez les statistiques dupliquées
- Planifiez régulièrement un job de mise à jour des statistiques avec un fort échantillonnage ou un FULLSCAN

Bonnes pratiques

- Utilisez les statistiques filtrées et/ou multicolonnes pour résoudre des problèmes d'estimation et de distribution de valeurs

Questions & Réponses

Téléchargez la présentation

- Blog Datafly : http://blog.datafly.pro/post/les-statistiques
- Blog SQLPro : http://blog.developpez.com/sqlpro/p/ms-sql-server/les-statistiques
- Site GUSS : www.guss.pro

mssql.fr
L'actualité technique MS SQL Server en France (et ailleurs)


GUSS