

COMMISSION DE STATISTIQUE et
COMMISSION ÉCONOMIQUE POUR
L'EUROPE



Distr.
GÉNÉRALE

CES/SEM.43/4 (Summary)
14 janvier 2000

CONFÉRENCE DES STATISTICIENS
EUROPÉENS

FRANÇAIS
Original : ANGLAIS

Séminaire sur les systèmes intégrés
d'information statistique et les questions
connexes (ISIS 2000)

(Riga, Lettonie, 29-31 mai 2000)

Thème I : Entreposage des données et mise
en place et utilisation des bases de données
statistiques dans un environnement de réseaux

**ÉVALUATION DE LA SOLUTION OFFERTE PAR LA CRÉATION
D'UN ENTREPÔT POUR LA GESTION DES DONNÉES D'ENTREPRISE
DANS UN SERVICE NATIONAL DE STATISTIQUE**

Communication sollicitée

de l'Office fédéral Suisse de la statistique¹

RÉSUMÉ

I. PRINCIPES

1. D'après Sundgren, l'entrepôt d'entreprise va devenir la clé de voûte de l'architecture d'un service national de statistique. C'est le réceptacle de toutes les microdonnées finales (registres des observations finales) et de toutes les macrodonnées valables (y compris les indicateurs) ainsi que des métadonnées correspondantes. L'entrepôt doit intégrer dans ses fonctions les moyens d'exécuter la majorité des opérations de transformation des microdonnées en macrodonnées. Au niveau des métadonnées,

¹ Établie par Georges Fleuti.

l'entrepôt doit englober et maîtriser la terminologie officielle des univers/populations, des objets/variables ainsi que de toutes les grandes classifications, la description des opérations de transformation et l'inventaire des sources, activités et produits statistiques. Il doit servir de source pour une base de données en ligne à laquelle ont accès (en payant) des utilisateurs extérieurs (par le biais par exemple de l'Internet) avec un moteur de recherche et un logiciel de navigation adéquats, y compris pour l'exportation de données.

II. EXCLUSION

2. L'entrepôt n'englobe pas les opérations de traitement des données en amont des registres des observations finales, à l'exception des interfaces pour l'échange de données et de métadonnées entre le réceptacle central et l'univers de la production. La gestion des registres des unités (par exemple les registres d'entreprises) et des microdonnées géocodées ou des agrégats définis par quadrillage est elle aussi exclue. Par ailleurs, nous ne prenons pas en compte les divers moyens de création d'un produit destiné à être diffusé à partir de l'entrepôt.

III. FONCTIONS

3. En tant que clé de voûte, l'entrepôt doit englober les principales fonctions nécessaires en statistique afin de réduire au minimum le travail de routine. La plupart des fonctions doivent être régies par des métadonnées. Le chargement des données brutes provenant de sources diverses dans une base de données relationnelle pour les microdonnées doit être aussi simple que possible et comprendre aussi peu de travail de routine que possible. Ces microdonnées doivent alors être transformées par agrégation automatique en cubes de macrodonnées dans une base de données multidimensionnelle. Les cubes-indicateurs qui associent des éléments de différents cubes de macrodonnées constituent le troisième niveau de l'entrepôt. L'une des fonctions essentielles est l'accès à toutes les données - micro et macro - dans le cas de requêtes spéciales. Les microdonnées doivent être accessibles en ligne pour les utilisateurs internes qui veulent sélectionner des sous-ensembles ou construire des macrodonnées ad hoc.

IV. CONDITIONS REQUISES

4. Le système doit offrir plusieurs langages, à la fois pour la commodité de l'utilisateur et pour les métadonnées. Les interfaces à haute performance et faciles d'emploi sont inévitables. Le système doit offrir la possibilité d'application sur diverses plates-formes et garantir une bonne performance de toutes les fonctions utilisées.

V. POLITIQUE

5. Pour que l'entrepôt d'un service de statistique fonctionne bien et de façon efficace, il est absolument indispensable d'appliquer une politique rigoureuse pour garantir qu'aucun produit désigné comme statistique officielle, quel qu'en soit le mode de diffusion, ne contienne de figure ni de notion en rapport avec les figures qui n'apparaissent pas dans la partie macro de l'entrepôt. Le mode de structuration et de définition des macrodonnées à inclure dans l'entrepôt devra satisfaire à certaines

règles et certains critères et ne pourra être décidé unilatéralement. Il en ira de même de la définition des métadonnées correspondantes. L'accès à l'entrepôt est strictement limité aux utilisateurs internes.

VI. BANC D'ESSAI

6. Un banc d'essai est une bonne occasion d'apporter la preuve que la fonction proposée existe ou n'existe pas réellement et de montrer comment fonctionnent différents systèmes de base de données à toutes les étapes appropriées dans le cadre de l'entrepôt pour une plate-forme donnée.

7. Les éléments essentiels pour le banc d'essai peuvent se répartir en cinq grandes catégories :

- L'architecture,
- Le système de base de données (microdonnées, macrodonnées, métadonnées),
- Les fonctions de sélection des microdonnées et des macrodonnées,
- Les mêmes métadonnées pour les microdonnées et les macrodonnées,
- La transformation des microdonnées en macrodonnées par agrégation.

8. Le banc d'essai organisé par l'Office fédéral suisse de la statistique en 1999 dans le cadre du projet d'entrepôt a donné des résultats surprenants. Comme plate-forme commune pour les essais, le Centre Compaq Benchmark de Valbonne (France) a fourni un Alpha-Serveur comportant quatre processeurs de 64 bits, avec 1 téraoctet d'espace disque en RAID-5. Le système d'exploitation utilisé était un Tru64-UNIX. Les trois partenaires - ORACLE, SAS et MSI - ont accepté de concourir et ont obtenu quatre journées entières pour mener à bien les opérations demandées.

9. Les données initiales fournies sur des tableaux bidimensionnels en code ASCII étaient tirées du recensement de la population de 1990 et du commerce extérieur ainsi que d'une enquête par sondage sur les loyers des logements municipaux. Les données ont ensuite été reproduites au moyen d'un algorithme spécial (50 ans) afin d'obtenir plus de 2,2 milliards d'enregistrements à stocker sur 106 gigaoctets d'espace disque.

10. Les critères utilisés pour comparer les résultats étaient l'espace nécessité, la durée d'exécution pour les traitements par lots et le temps de réponse pour les requêtes spéciales. Les différences entre les trois systèmes étaient considérables; ORACLE et SAS étaient de plusieurs (jusqu'à 10) facteurs plus lents que WIDAS de MSI. Les résultats sont spécifiques de l'entrepôt d'un service de statistique qui n'est pas axé sur des transactions mais plutôt sur des demandes de consultation qu'il n'est possible de normaliser que jusqu'à un certain point.
