الفصل الثاني عشر:
إدارة المناقلات والتحكم المتزامن

الكلمات المفتاحية:

نظم دعم القرار، مستودع المعطيات، فلترة المعطيات، استخلاص المعطيات، الأبعاد، مستويات التجميع، الحفر للأسفل، التجميع للأعلى، مخزن المعطيات الصغير، تكامل المعطيات، إجرائية التحليل المباشر (OLAP)، المعطيات متعددة الأبعاد، مكعب المعطيات، شرائح المعطيات، إجرائية التحليل المباشر العلاقاتية (ROLAP)، إجرائية التحليل المباشر متعددة الأبعاد (MOLAP)، بنية نجمية، الحقائق، الواصفات، مكعب فائق، جدول الحقيقة، جدول البعد، هرمية الواصفات، التنقيب عن المعطيات.

ملخص:

يركز هذا الفصل على مستودعات المعطيات، ميزاتها الأساسية، اختلافها عن قواعد المعطيات العملياتية وطريقة بنائها والبنى الأساسية فيها (البنية النجمية)، بالإضافة إلى الفوائد منها.

الأهداف تعليمية:

يهدف هذا الفصل التعريف بالمفاهيم التالية:

- تعريف مخزن المعطيات
- ميزات مخازن المعطيات:
  1. موجهة حسب الغرض منها Subject oriented
  2. متكاملة Integrated
  3. متغيرة زمنياً Time-variant
  4. غير متحركة Non-volatile
- الفرق بين مخازن المعطيات وقواعد المعطيات العملياتية
- بنية مخازن المعطيات Data warehouse architecture:
  1. جدول الحقائق (Fact table):
  2. جدول الأبعاد (Dimension table):
  3. البنية النجمية Star schema
  4. البنية البلورية Snowflake schema
- بناء مخزن المعطيات:
  5. الاستخلاص Extract
  6. المعالجة Transform
  7. التحميل Load
- استخدامات مخازن البيانات

1. مقدمة

يعتبر مخزن المعطيات حسب تعريف Bill Inmon عام 1990، هو بيانات مجمعة من مصادر متنوعة. مجمّعة حسب الغاية منها، متكاملة فيما بينها، متغيرة زمنياً وغير متحركة (non-volatile)، بمعنى أنه لا تتم عليها عمليات إدخال وتعديل وحذف (transactions)، هذه البيانات مجمعة بغرض مساعدة محلل البيانات ومتخذ القرار في استخلاص ما يطلبه من معلومات بسرعة.

2. ميزات مخازن المعطيات

يتم تخزين البيانات في قواعد المعطيات العملياتية، ضمن جداول تحكم بنيتها قواعد التنظيم (فصل 6-7)، وقواعد تصميم بنى المعطيات العلائقية (فصل 2)، وغاية هذه القواعد الوصول إلى طريقة التخزين الأفضل للمناقلات (ادخال، تعديل وحذف).

أما مخازن المعطيات، فبنى المعطيات فيها لا تتبع قواعد التنظيم والقواعد العلائقية (فمثلاً يمكن أن تحوي جداول مخزن المعطيات حقول محسوبة وبيانات مكررة)، لأن الغاية الأساسية منها هي سرعة تنفيذ الاستعلامات، و تغير بياناتها يتم من خلال عمليات الاستخلاص – التحويل – التحميل (ETL)، ولا يتم عبر المناقلات التقليدية.

أهم ما يميز مخازن البيانات، هي أربع خصائص:

- **موجهة حسب الغرض منها Subject oriented**

تُنَظَم معطيات مستودع المعطيات وتُختَزَل في عدة مجالات، لإعطاء أجوبة عن الأسئلة المختلفة للشركة (حجم المبيعات حسب الفرع، حجم المبيعات موزعة على أشهر السنة، ...)، أي أن الأسئلة التي تتعلق بإجمالي المبيعات يمكن ايجاد أجوبتها في إحدى بنى مخزن المعطيات، ويمكن أن نجد بنى أخرى للاجابة على أسئلة أخرى (مثلاً إجمالي التكاليف موزعة حسب أقسام الشركة أو زمنياً). وسنشرح بنى التخزين في مخزن المعطيات في فقرة لاحقة.

- **متكاملة Integrated**

مصدر البيانات في مخزن المعطيات هو قواعد بيانات قد تكون متنافرة ومن أنواع مختلفة (MS access db, Oracle db, Excell sheets)، وقبل إدخالها إلى مخزن المعطيات تمر بمراحل من التنظيف والتصحيح والإكمال، وفي النهاية يجب مكاملتها قبل أن تصبح جزءاً من مخزن المعطيات.

- **متغيرة زمنياً Time-variant**

تتم تغذية مخزن المعطيات بالبيانات عادةً بشكل دوري، وغالباً ما تحمل بعداً زمنياً، للإجابة على الأسئلة التي تعكس التطور الزمني (تزايد المبيعات، تقليص التكاليف ..).

- **غير متحركة Non-volatile**

تدخل البيانات إلى مخزن المعطيات بتحميلها من المصدر (Bulk load)، ولا تدخل التسجيلات إفرادياً عبر المناقلات، وغالباً لا يتم حذف البيانات أو تعديلها بعد أن تصبح جزءاً من المخزن، والعمليات التي تتم عليها هي الاستعلام فقط.

3. الفرق بين مخازن المعطيات وقواعد المعطيات العملياتية

ثلاثة اختلافات أساسية (الفترة الزمنية، مستويات التجميع، الأبعاد):

- **الفترة الزمنية**

تغطي المعطيات العملياتية فترة صغيرة من الزمن، فالمناقلات تحتاج لسجلات المبيعات والفواتير اليومية والكميات المباعة والمخزنة، أما بالنسبة لنظم دعم القرار فلا تهتم بفاتورة معينة أو بمشتريات زبون معين، إنما بعمليات الشراء التي حدثت في شهر أو سنة، أو مثلاً بمشتريات نمط معين من الأشخاص.

- **مستويات التجميع**

يتم تجميع المعطيات في نظم دعم القرار في مستويات متعددة من معطيات فردية تقريباً إلى معطيات كلية شاملة. حيث يمكن مثلاً للمدير أن يرى المبيعات بحسب القطاع أو بحسب المدينة داخل القطاع أو بحسب المتجر داخل المدينة، وتدعى عمليات طلب معلومات تفصيلية أكثر بالحفر للأسفل (drill down)، أم طلب معلومات مُجمَّعة فتدعى التجميع للأعلى (roll up).

- **الأبعاد**

يتعامل نظام دعم القرار مع أبعاد متعددة للمعطيات، مثلاً إذا أردنا أن نعرف المبيعات التي حدثت في قطاع معين وخلال شهر معين فلدينا بعدين للمعطيات (بعد للمنطقة) و(بعد للزمن).

4. بنية مخازن المعطيات Data warehouse architecture

تعتبر بنى التخزين في قواعد المعطيات جداول تخضع لقواعد التنظيم، أما في مخازن المعطيات فتختلف بنى التخزين قليلاً، فهي جداول إلا أنها ليست بالضرورة من الشكل النظامي الثالث، ويتم تصميمها وفق إحدى بنيتين: البنية النجمية (Star schema) والبنية البلورية (Snowflake schema)، وفيما يلي شرح لكلا البنيتين وتطبيق على مثال يعكس تحليل مبيعات شركة.

قبل شرح البنية النجمية والبنية البلورية ، يجب أن نعرف أن كلا البنيتين يعتمد على نوعين من الجداول، هما جدول الحقائق، وجدول الأبعاد.

جدول الحقائق (Fact table)

يتألف جدول الحقائق من نوعين من الحقول: مفاتيح مستوردة من جداول الأبعاد، وقياسات تتضمن قيم (غالباً ما تكون رقمية)، وهذه القياسات قد تكون مفصّلة أو مجمعة.

مثال عن المفاتيح المستوردة: رقم الفرع، الشهر، العام، صنف المنتج، وغيرها.

مثال عن القياسات: القيمة الإجمالية للمبيعات، كمية المبيعات، وغيرها.

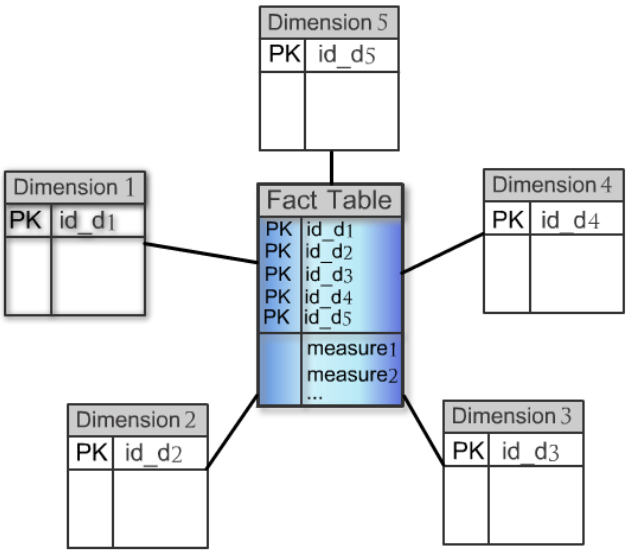
جدول الأبعاد (Dimension table)

قد يكون جدول الأبعاد بسيطاً (مفتاح وقيمة واحدة نصية أو تاريخ)، وقد يمثل بنية هرمية (مفتاح، سنة، شهر، يوم مثلاً)، وفي الحالتين يصدّر جدول الأبعاد مفتاحه لجدول الحقائق ويكون جزءاً من مفتاحه الأساسي المركب.

وتمثل جداول الأبعاد عادةً توزع جغرافي أو تصنيف لمنتج، وهي غالباً الأبعاد التي تطلب الاحصاءات مجمّعة حسبها.

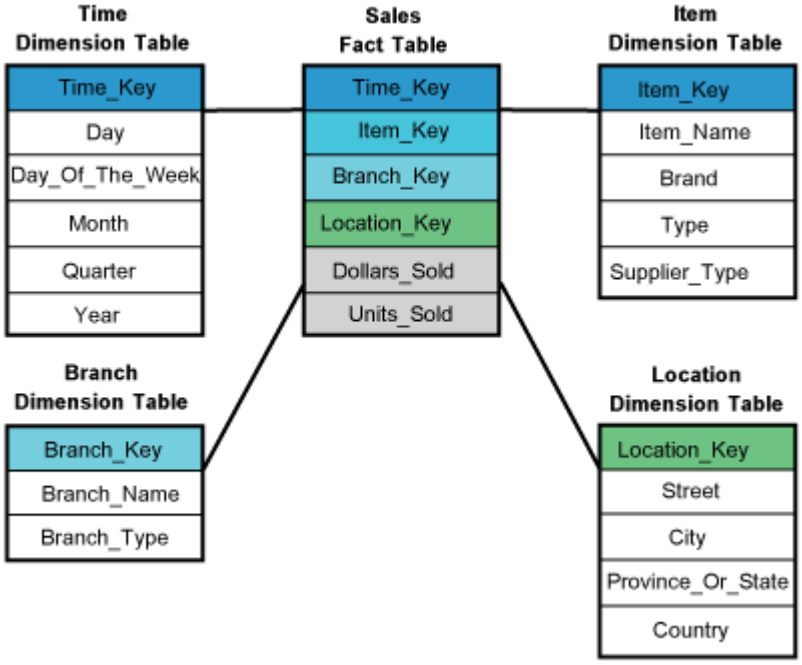
البنية النجمية (Star schema)

البنية النجمية هي بنية بسيطة، تشبه نجمة مركزها هو جدول الحقائق، ورؤوسها هي جداول الأبعاد:



أهم ميزات البنية النجمية:

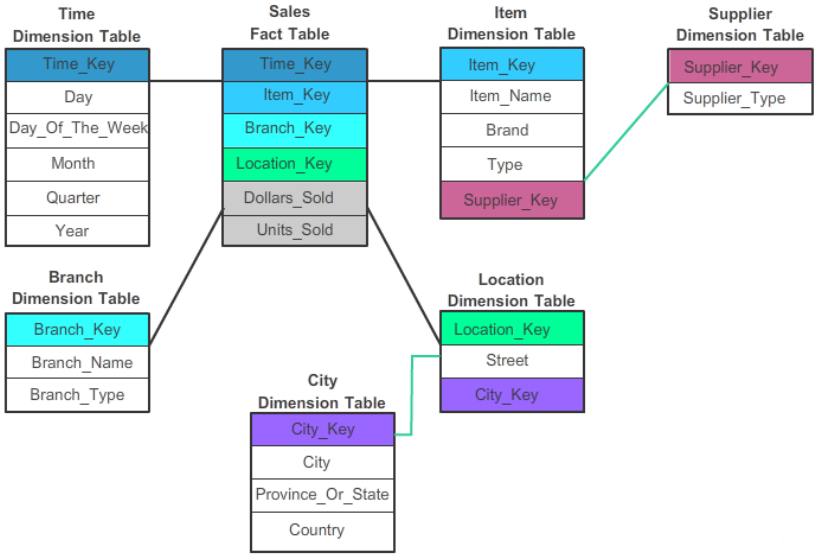
- سهلة الفهم وبسيطة نسبياً.
- الاستعلامات المبنية على جدول الحقائق بسيطة التركيب (عدد قليل من الجداول للربط) وسريعة التنفيذ.
- قد يزداد حجم جداول الأبعاد كثيراً (جداول غير منظمة)، وقد يستهلك تحميلها بالمعطيات وقتاً كبيراً.
- تعتبر البنية النجمية الأوسع انتشاراً والأكثر استخداماً في مخازن المعطيات.



البنية البلورية (Snowflake schema)

تكون جداول الأبعاد في البنية النجمية غير منظمة واحتمال التكرار فيها وارد، فمثلاً في الشكل السابق يمكن أن نجد في البعد المكاني (Location)، تكرار للقيمة دمشق بعدد الشوارع المدخلة أسماؤها.

بينما تكون جداول الأبعاد في البنية البلورية منظمة، ومثل الجدول Location تتم تجزءته إلى أكثر من جدول كما في الشكل التالي:



ويتم بناء الجداول بنوعيها (الحقائق والأبعاد) في مخزن المعطيات والتخاطب معها من خلال توسعة للغة SQL تدعى DMQL، وتدعمها معظم نظم ادارة قواعد المعطيات العلائقية (Oracle, SQL Server).

سنعرض هنا بعض الأمثلة على لغة DMQL:

البنية النجمية

Sales fact table

```
define cube sales star [time, item, branch, location]:

dollars sold = sum(sales in dollars), units sold = count(*)
```

Dimension tables

```
define dimension time as (time key, day, day of week, month, quarter, year)

define dimension item as (item key, item name, brand, type, supplier type)

define dimension branch as (branch key, branch name, branch type)

define dimension location as (location key, street, city, province or state, country)
```

==========

البنية البللورية

Sales fact table

```
define cube sales snowflake [time, item, branch, location]:

dollars sold = sum(sales in dollars), units sold = count(*)
```

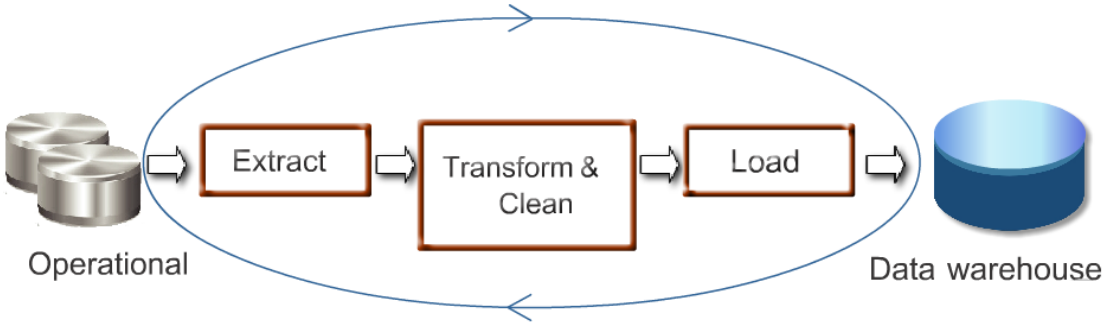
Dimension tables

```
define dimension time as (time key, day, day of week, month, quarter, year)

define dimension item as (item key, item name, brand, type, supplier
(supplier key, supplier type))
define dimension branch as (branch key, branch name, branch type)

define dimension location as (location key, street, city (city key, city, province or state, country))
```

5. بناء مخازن المعطيات

تبدأ عملية بناء مخزن المعطيات باستخلاص البيانات من مصادرها، تعديلها وتصحيحها ومكاملتها، وأخيراً إدخالها إلى مخزن المعطيات (ETL)، ويمكن تكرار العملية من بدايتها بشكل دوري (شهرياً مثلاً)، ويستخدم مخزن البيانات بعد بنائه وتغذيته بالمعطيات، كقاعدة معطيات تبنى عليه أدوات الاسترجاع والعرض وتحليل البيانات.

فيما يلي سنفصّل مراحل بناء مخزن المعطيات، مع الأمثلة:



استخلاص المعطيات Extraction

تكون البيانات في مخزن المعطيات مجمعة لتخدم غرضاً معيناً، كتحليل مبيعات شركة مثلاً، وهنا نجد أن البيانات المطلوبة قد تكون موزعة في أكثر من مصدر (قواعد بيانات عملياتية، جداول الكترونية Excel sheets وغيرها)، وقد تكون قواعد البيانات العملياتية مختلفة من حيث النوع والبنى داخلها.

وهنا يتم عادةً توحيد صيغة البيانات المستخلصة لنقلها إلى المخزن، إما عن طريق مراحل وسيطة (ملفات نصية أو قواعد معطيات)، وإما عن طريق التحكم بتعليمات Select التي يتم من خلالها استخلاص البيانات من المصدر، ويمكن دائماً استخدام أدوات ETL التي توفرها نظم إدارة قواعد المعطيات العلائقية.

معالجة المعطيات Transformation

غالباً ما تكون البيانات المستخلصة من المصادر بحاجة إلى معالجة: **تصحيح** و**استكمال** و**تنظيف**، لتتلاءم وتكون قابلة للنقل إلى الحقول المقابلة في مخزن المعطيات.

التصحيح: قد تكون البيانات مخزنة بأنماط لا تتوافق مع نمط الجهة التي ستصب فيها، يمكن مثلاً أن نجد التاريخ في أحد مصادر فواتير الزبائن مخزن على شكل نص، أو قد نجد فروع الشركة في أحد المصادر مرمزة ومدخلة في فهارس بينما نجدها في مصادر أخرى مدخلة بشكل نصي، وهنا لا بد من تصحيح بيانات المصادر المختلفة وتوحيد صيغتها لتسهيل نقلها إلى مخزن المعطيات.

الاستكمال: في كثير من الأحيان يمكن أن تكون بيانات المصدر ناقصة، فمثلاً إذا كان أحد المصادر هو قاعدة بيانات لمبيعات أحد الفروع في الشركة، فمن الممكن أن لا نجد اشارة للفرع في الفواتير كونها تخص فرعاً واحداً، وهنا لا بد من إضافة رمز أو الرقم المعرف للفرع لكل فاتورة قبل نقل بياناتها لمخزن المعطيات.

التنظيف: يمكن أن يتضمن المصدر بيانات خاطئة (فاتورة غير مستكملة – اسم الزبون غير مدخل مثلاً)، والتنظيف هنا يعني حذف هذه الادخالات أو اسناد قيم افتراضية للمدخلات الناقصة أو المدخلة خطأً.

وتعتبر المعالجة الأهم للبيانات قبل نقلها إلى مخزن المعطيات، هي توحيد صيغتها.

وفيما يلي مثال عن مصدرين لمبيعات فرعين لشركة، ونموذج عن عمليات التصحيح اللازم تطبيقها عليها:

مبيعات المنطقة الجنوبية (المصدر الأول)

المبيعات مدخلة بالليرة السورية والاجمالي متضمن في الفاتورة

Invoice

Id	Date	City	Total
1	1/1/2015	Damascus	50000
2	15/1/2015	Sweida	40000
3		Damascus	15000
4	4/2/2015	Daraa	2500
5	1/3/2015	Damascus	10000

==========

مبيعات حمص (المصدر الثاني)

المبيعات مدخلة بالدولار

Invoice

Id	Date	Client
1	1/1/2015	X client
2	15/2/2015	Y client

Item

Id	Invoice	Mat	Qty	Price
1	1	M1	2	15
2	1	M2	5	22
3	2	M2	10	22
4	1	M3	6	30

==========

مخزن البيانات

Sales

Branch	Month	Total
Damascus	Jan	50000
Damascus	Mar	10000
Sweida	Jan	40000
Daraa	Feb	2500
Homs	Jan	99000
Homs	Feb	66000

==========

ملاحظات:

- الفاتورة رقم 3 (المنطقة الجنوبية) لم يتم ادخال التاريخ لها، وهي حالة يتم تنظيفها باستثناء الادخالات التي لا تتضمن تاريخ.
- الأسعار مدخلة في المصدر الأول بالليرة السورية وفي المصدر الثاني بالدولار، وهي حالة بحاجة معالجة (يمكن تحويل الأسعار في المصدر الثاني إلى الليرة السورية بضربها بوسطي سعر 300)
- المدينة في المصدر الثاني غير مدخلة (استكمال)، يمكن تصحيحها بإضافة "حمص" في التسجيلات التي يتم استخلاصها من المصدر الثاني.
- يمكن نقل بيانات كل مصدر إلى محطة وسيطة (بعد معالجتها) قبل نقلها إلى مخزن البيانات.
- البيانات في مخزن المعطيات مجمعة حسب الشهر والمدينة.

==========

```
Select date, city, sum(total) from Invoice where not (isnull(date))
and not (isnull(city))
Group by city,date;

Select date, 'Homs', sum(y) from
(
Select invoice.id as x, date, 'Homs', sum(qty*price*300) as y
From invoice, item
Where invoice.id=item.invoice
Group by date
)
Group by date;
```

- البيانات الناتجة عن تعليمتي SQL السابقتين هي من الشكل:

date	city	Total

وقبل نقلها لمخزن المعطيات يجب تحويل التاريخ إلى شهر، والتجميع حسب الأشهر.

تحميل المعطيات Load

يعني تحميل البيانات نقلها بعد استخلاصها ومعالجتها إلى مخزن المعطيات، ويتم تحميل البيانات بشكل دوري (كل شهر مثلاً)، ويمكن أن يتم التحميل إما من المصدر مباشرة أو من المحطات الوسيطة.

تتوفر في معظم نظم إدارة قواعد المعطيات العلائقية (Oracle, MS Sql server, ...)، أدوات الاستخلاص والمعالجة وتحميل البيانات، وغالباً ما تكون هذه الأدوات بيانية وسهلة الاستخدام.

6. استخدامات مخازن البيانات

تعتبر أهم التطبيقات التي تبنى على مخازن المعطيات هي نظم دعم القرار، وهي تعتمد على استخلاص المعلومات (حجوم كبيرة من المعطيات)، وعرضها بشكل مقروء (نصوص، جداول، مخططات بيانية)، بحيث تسهل عمل متخذ القرار وتعطيه المعلومات اللازمة وبشكل سريع ليبني قراره على معلومات دقيقة.

تحتاج نظم دعم القرار إلى حجوم كبيرة من المعطيات (عينة احصائية كبيرة)، لتكون المؤشرات والحقائق المبنية عليها، أقرب ما يكون إلى الدقة.

من أهم المجالات التي تخدم فيها مخازن المعطيات في توجيه القرار والمساعدة على اتخاذه:

- تخطيط الانتاج، بالاعتماد على البيانات والاحصاءات التي تتعلق بحجم المبيعات وتوزعها الزمني على مدار العام.
- تحليل الزبائن، من حيث أفضليات الشراء ومواعيدها.
- التحليل العملياتي كإدارة العلاقة مع الموردين (Supply chain management)، ومع الزبائن والموزعين.

7. المراجع

- http://www.tutorialspoint.com/dwh/dwh_data_warehousing.htm

8. تدريبات

- مخازن المعطيات هي قواعد معطيات علائقية تخضع لقواعد التنظيم ؟

1. صح.
2. خطأ.

الإجابة: (2) (ليس بالضرورة أن تكون الجداول في مخزن المعطيات من أي شكل نظامي)

- تصميم مخزن المعطيات مقاد بالمعطيات ؟

1. صح.
2. خطأ.

الإجابة: (2) (تصميم مخزن المعطيات مقاد بالعمليات)

- لغة التخاطب مع محتوى مخازن المعطيات هي DMQL؟

1. صح.
2. خطأ.

الإجابة: (1)

- تكرار المعطيات غير وارد في مخزن المعطيات ؟

1. صح.
2. خطأ.

الإجابة: (2) (تكرار المعطيات وارد في مخازن المعطيات، وخاصة في البنية النجمية)