

(2)

بنية مستودعات البيانات Data Warehouse Architecture

تحدد بنية مستودع البيانات كيفية جمع البيانات، تخزينها، ومعالجتها، هناك عدة نماذج لبنية مستودعات البيانات، ولكن النموذج الأكثر شيوعاً يتضمن الطبقات التالية:

1- طبقة مصادر البيانات Data Source Layer :

جميع المصادر التي يتم جمع البيانات منها، بتسويات مختلفة Structured، Unstructured

✓ قواعد بيانات تشغيلية OLTP OnLine Transaction Processing

✓ ملفات مسطحة CSV، Excel

✓ أنظمة إدارة الموارد ERP أو إدارة العلاقات مع العملاء CRM

✓ بيانات خارجية مثل بيانات الطقس أو وسائل التواصل الاجتماعي

2- طبقة التكامل Data Integration Layer

تُعرف أيضاً بـ : ETL (Extract, Transform, Load)

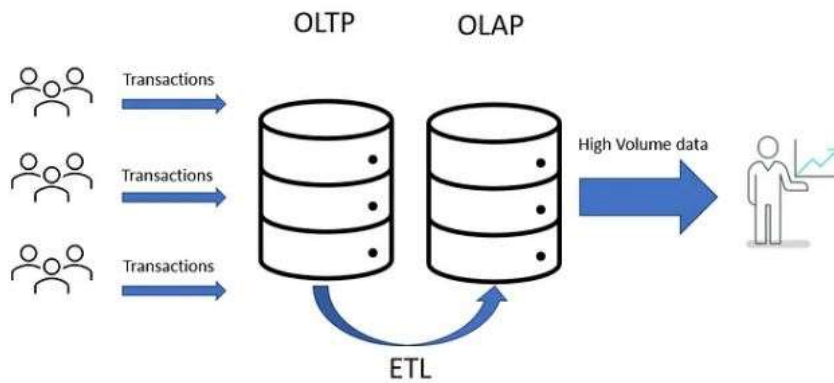
✓ الاستخلاص Extract جمع البيانات من المصادر.

✓ التحويل Transform تنظيف البيانات، توحيد التسويات، إزالة التكرارات، وإجراء التحويلات مثل

تحويل العملات .

✓ التحميل Load تخزين البيانات في مستودع البيانات.

من أدوات ETL الشائعة: Informatica، Talend، Apache NiFi.



3- طبقة التخزين Data Storage Layer

هي قاعدة البيانات المستخدمة لإدارة مستودع البيانات، وتُصمم غالباً باستخدام نماذج مثل:
 1. نموذج النجمة Star Schema أو 2. نموذج رقاقة الثلج Snowflake Schema
 وتستخدم الفهرسة والتقسيم Partitioning لتحسين جاهزية هذه الطبقة للاستعلامات التحليلية.

1- نموذج النجمة:

هو أحد أكثر نماذج تصميم مستودعات البيانات شيوعاً بسبب بساطته وأدائه العالي في الاستعلامات التحليلية، يُطلق عليه "النجمة" لأن تصميمه يشبه شكل نجمة، حيث يوجد جدول حقائق (Fact Table) في المركز مرتبط بمجموعة من جداول الأبعاد تحيط به (Dimension Tables).



جدول الحقائق (Fact Table) :

هو الجدول المركزي في نموذج النجمة، ويحتوي على البيانات الكمية أو المقاييس (Measures) التي تُستخدم في التحليل، المقاييس غالباً هي أرقام يمكن تجميعها لدعم التحليلات (مثل المجموع، المتوسط)، بشكل عام يحتوي جدول الحقائق على:

- ✓ المقاييس (Measures): مثل إجمالي المبيعات، الكمية المباعة، أو التكلفة.
- ✓ المفاتيح الأجنبية (Foreign Keys): التي تربط جدول الحقائق بجدول الأبعاد.



مثلاً في مستودع بيانات لمتجر تجزئة، قد يحتوي جدول الحقائق Sales_Fact على الحقول التالية:

- Sales_Amount (قيمة المبيعات).
- Quantity_Sold (الكمية المباعة).
- Product_ID (مفتاح أجنبي يربط بجدول المنتجات).
- Customer_ID (مفتاح أجنبي يربط بجدول العملاء).
- Time_ID (مفتاح أجنبي يربط بجدول الوقت).
- Store_ID (مفتاح أجنبي يربط بجدول المتاجر).

جداول الأبعاد (Dimension Tables):

يحتوي على البيانات الوصفية (Descriptive Data) التي توفر وصف المقاييس في جدول الحقائق، كل جدول أبعاد يمثل بُعداً (Dimension) معيناً يمكن من خلاله تحليل البيانات، مثل الوقت، المنتج، العميل، أو الموقع، بشكل عام تحتوي جداول الأبعاد على:

- ✓ مفتاح أساسي (Primary Key) الذي يربط جدول البعد مع جدول الحقائق.
- ✓ السمات (Attributes) التي توفر تفاصيل وصفية.

مثلاً قد يحتوي جدول أبعاد المنتج Product_Dimension على الحقول التالية:

- Product_ID (مفتاح أساسي).
- Product_Name (اسم المنتج).
- Category (الفئة، مثل إلكترونيات).
- Brand (العلامة التجارية).
- Unit_Price (سعر الوحدة).

مثال عملي:

لتحليل مبيعات متجر وحساب إجمالي المبيعات حسب السنة وفئة المنتج، نستخدم جدول الحقائق Sales_Fact الذي يحتوي على المبيعات والكميات، بالإضافة لجدول الأبعاد التالية:

- Product_Dimension (المنتجات).
- Time_Dimension (التاريخ، الشهر، السنة).
- Customer_Dimension (العملاء).

وبالتالي سيكون الاستعلام على الشكل التالي:

```
SELECT t.Year, p.Category, SUM(f.Sales_Amount) as Total_Sales
FROM Sales_Fact f
JOIN Time_Dimension t ON f.Time_ID = t.Time_ID
JOIN Product_Dimension p ON f.Product_ID = p.Product_ID
GROUP BY t.Year, p.Category;
```



2- نموذج رقاقة الثلج:

هو نسخة أكثر تعقيداً من نموذج النجمة، حيث يتم تطبيع Normalization جداول الأبعاد إلى جداول فرعية إضافية، جدول الحقائق في نموذج رقاقة الثلج مشابه لنموذج النجمة، حيث يحتوي على المقاييس والمفاتيح الأجنبية التي تربطه بجدول الأبعاد.

الفرق يكمن في بنية جداول الأبعاد التي يتم تقسيمها إلى جداول فرعية مترابطة باستخدام مفاتيح أساسية وأجنبية، مما يقلل من تكرار البيانات ولكن تصبح معالجة الاستعلامات أكثر تعقيداً وأقل توافقية مع OLAP

مثال: في نموذج النجمة، جدول أبعاد المنتج قد يحتوي على جميع السمات (مثل Product_Name، Category، Brand)، في نموذج رقاقة الثلج، يتم تقسيم هذا الجدول إلى:

- Product_Dimension:
 - Product_ID (مفتاح أساسي).
 - Product_Name.
 - Category_ID (مفتاح أجنبي يربط بجدول الفئات).
 - Brand_ID (مفتاح أجنبي يربط بجدول العلامات التجارية).
- Category_Dimension:
 - Category_ID (مفتاح أساسي).
 - Category_Name.
- Brand_Dimension:
 - Brand_ID (مفتاح أساسي).
 - Brand_Name.

وبالتالي يصبح الاستعلام كما يلي:

```
SELECT y.Year, c.Category_Name, SUM(f.Sales_Amount) as Total_Sales
FROM Sales_Fact f
JOIN Product_Dimension p ON f.Product_ID = p.Product_ID
JOIN Category_Dimension c ON p.Category_ID = c.Category_ID
JOIN Time_Dimension t ON f.Time_ID = t.Time_ID
JOIN Year_Dimension y ON t.Year_ID = y.Year_ID
GROUP BY y.Year, c.Category_Name;
```

4- طبقة الوصول والتحليل Data Access and Analysis Layer :

توفر هذه الطبقة للمستخدمين واجهات الوصول إلى البيانات والدوات اللازمة للتحليل، مثل:

✓ أدوات OLAP لتحليل البيانات متعدد الأبعاد

✓ أدوات تصور البيانات مثل Tableau أو Power BI

✓ لغات الاستعلام مثل SQL

✓ لوحات المعلومات Dashboards والتقارير



5- طبقة العرض Presentation Layer :

- ✓ تقدم النتائج لأصحاب القرار عبر تقارير، رسوم بيانية، أو تطبيقات ويب.
- ✓ تضمن سهولة الاستخدام والوصول السريع إلى الرؤى.

أنواع بنية مستودع البيانات

تُعد بنية مستودع البيانات Data Warehouse Architecture إطاراً يحدد كيفية جمع البيانات، تخزينها، معالجتها، وتحليلها لدعم اتخاذ القرار في المؤسسات.

تختلف البنية بناءً على احتياجات المؤسسة، حجم البيانات، وتعقيد التحليلات المطلوبة، الأنواع الرئيسية هي:

- ✓ البنية أحادية الطبقة: مستودع بيانات واحد يخدم جميع الأغراض وهذا نادر الاستخدام .
- ✓ البنية ثنائية الطبقة: فصل بين مستودع البيانات وطبقة التحليل.
- ✓ البنية ثلاثية الطبقة: تشمل طبقة وسيطة Data Marts لتلبية احتياجات أقسام معينة، حيث يتم استخراج جزء من البيانات من المستودع وتحميله إلى Data Marts مخصصة لأقسام معينة، تُنفذ الاستعلامات التحليلية على Data Marts، مما يقلل الضغط على المستودع الرئيسي.

تنفيذ مستودعات البيانات Data Warehouse Implementation

يتطلب تنفيذ مستودع البيانات تخطيطاً دقيقاً وتنسيقاً بين فرق متعددة وتشمل المراحل الرئيسية:

1- تحديد المتطلبات Requirements Gathering

- ✓ فهم احتياجات المؤسسة، مثل الأسئلة التحليلية التي يجب الإجابة عنها مثل "ما هي المنتجات الأكثر مبيعاً؟".
- ✓ تحديد مصادر البيانات والمستخدمين النهائيين مثل (المدرء، المحللون) .

2- تصميم النموذج Data Modeling

- ✓ اختيار نموذج البيانات Star Schema أو Snowflake Schema
- ✓ تصميم جداول الحقائق والأبعاد بناءً على المتطلبات.
- مثال: جدول حقائق للمبيعات يحتوي على (القيم المبيعات، الكمية)، وجداول أبعاد (للمنتج، الوقت، والمنطقة).



3- تطوير عملية ETL

- ✓ تصميم عمليات الاستخلاص، التحويل، والتحميل.
- ✓ معالجة التحديات مثل البيانات المفقودة، تنسيقات غير متوافقة، أو حجم البيانات الكبير.
- ✓ اختبار عمليات ETL لضمان الدقة.

4- اختيار التقنية المناسبة Technology Selection

- ✓ قواعد بيانات مثل Oracle, Microsoft SQL Serve
- ✓ أدوات ETL مثل Talend أو Apache Airflow
- ✓ أدوات تحليل مثل Power BI أو QlikView
- ✓ النظر في الحلول السحابية مثل Amazon Redshift أو Google BigQuery

5- التنفيذ والاختبار Implementation and Testing

- ✓ تحميل البيانات الأولية إلى المستودع.
- ✓ اختبار الأداء مثل سرعة الاستعلامات، والدقة مثل التحقق من البيانات .
- ✓ معالجة أخطاء البيانات مثل التكرارات أو أخطاء التحويل.

6- النشر والصيانة Deployment and Maintenance

- ✓ نشر المستودع للمستخدمين النهائيين.
- ✓ إعداد جداول زمنية لتحديث البيانات يومياً، أسبوعياً .
- ✓ مراقبة الأداء وإجراء التحسينات مثل إضافة الفهرسة .

التحديات

- ⊗ التكلفة: بناء مستودع بيانات مكلف بسبب الأجهزة، البرمجيات، والموارد البشرية.
- ⊗ إدارة البيانات الضخمة: التعامل مع كميات هائلة من البيانات.
- ⊗ الأمان: حماية البيانات الحساسة باستخدام التشفير وإدارة الأدونات.
- ⊗ التكامل: ضمان توافق المستودع مع الأنظمة الحالية.



من مستودعات البيانات إلى التنقيب في البيانات From Data Warehousing to Data Mining

توفر مستودعات البيانات أساساً قوياً لتحليل البيانات، لكنها غالباً لا تكفي لاكتشاف الأنماط المخفية أو التنبؤ بالاتجاهات. هنا يأتي دور التنقيب في البيانات Data Mining ، وهو عملية استخدام الخوارزميات المتقدمة لاكتشاف (الأنماط، العلاقات، أو التنبؤات) في البيانات الكبيرة.

العلاقة بين مستودعات البيانات والتنقيب

- توفر مستودعات البيانات بيئة منظمة ومتكاملة لتخزين البيانات التاريخية، مما يجعلها جاهزة للتحليل.
- يستخدم التنقيب البيانات المخزنة في المستودع لتطبيق تقنيات مثل التعلم الآلي والإحصاء لاكتشاف رؤى جديدة.

كيفية الانتقال إلى التنقيب

1. إعداد البيانات:

- تنظيف وتحويل البيانات باستخدام عمليات ETL
 - اختيار مجموعة البيانات ذات الصلة بناءً على هدف التنقيب (مثل تحليل سلوك العملاء).
- #### 2. اختيار تقنيات التنقيب:
- تصنيف البيانات Classification إلى فئات (عميل محتمل الشراء أو عميل غير محتمل).
 - تجميع البيانات Clustering المتشابهة مثلاً تجميع العملاء بناءً على أنماط الشراء.
 - قواعد الارتباط Association Rules اكتشاف العلاقات مثل "الذين يشترون الحليب يشترون الخبز"
 - كشف الانحرافات Anomaly Detection تحديد البيانات غير الطبيعية مثل الاحتيال المالي .
 - التنبؤ Prediction بالقيم المستقبلية مثل التنبؤ بالمبيعات .

3. تطبيق الخوارزميات:

- تطبيق خوارزميات التنقيب Neural Networks ،K-Means ،Decision Trees.
- استخدام أدوات التطوير مثل Weka ،RapidMiner ،R ، Python Scikit-learn

