Chương I

1. BI là gì / Mục đích/ Lợi ích?

BI là đề cập tới các ứng dụng, các kỹ thuật và các tiến trình thu thập, lưu trữ và phân tích dữ liệu để giúp các người làm kinh doanh đưa ra quyết định tốt hơn.

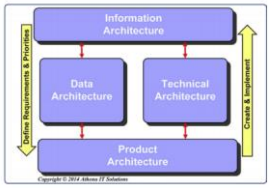
BI là sử dụng Kho dữ liệu (Data warehouse) để phân tích hiệu suất kinh doanh. Các BI tool nên cung cấp cái nhìn sâu sắc hướng dữ liệu.

Mục đích chính của BI là Phân tích, khai phá tri thức với các công cụ và phương pháp cho phép các doanh nghiệp đưa ra quyết định hiệu quả và kịp thời.

Lợi ích:

- Tăng khả năng kiểm soát thông tin của doanh nghiệp một cách chính xác
- Phân tích, khai phá tri thức giúp doanh nghiệp có thể dự đoán về xu hướng của giá cả dịch vụ, hành vi khách hàng, phát hiện khách hàng tiềm năng - đề ra các chiến lược kinh doanh phù hợp.

2. BI Framework.

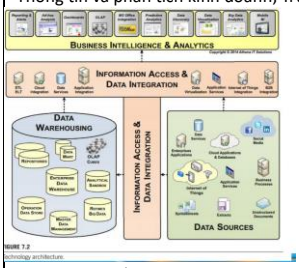


4 thành phần (component) chính:

• A data warehouse (DW): data source, Operational data stores, Data marts, Meta data
• Business analytics: Công cụ thao tác, khai thác, phân tích dữ liệu trong DW.
• Business performance management: Để theo dõi và phân tích hiệu suất.
• User interface: dashboard.

4 tầng kiến trúc:

- Information architecture: xác định "What, Where, Who, Why" cho BI hoặc các ứng dụng phân tích.
+ What: quy trình nghiệp vụ/chức năng sẽ hỗ trợ, loại phân tích nào sẽ cần, loại quyết định nào bị ảnh hưởng
+ Who: (nhân viên, khách hang, nhà cung ứng, các bên liên quan khác...) sẽ truy cập
+ Where: dữ liệu được tích hợp vào đâu, dùng ở đâu trong ứng dụng
+ Why: tại sao cần xây dựng giải pháp BI? Các yêu cầu kỹ thuật và nghiệp vụ nào liên quan?
- Data architecture:
• Giúp bạn hiểu rõ hơn về dữ liệu.
• Cung cấp hướng dẫn quản lý dữ liệu từ quá trình thu thập ban đầu trong hệ thống nguồn đến việc sử dụng thông tin của người kinh doanh.
• Hướng dẫn cách thu thập, tích hợp, nâng cao, lưu trữ và phân phối dữ liệu cho những người kinh doanh sử dụng dữ liệu đó để thực hiện công việc của họ.
• Cung cấp cấu trúc để phát triển và triển khai quản trị dữ liệu.
- Technology & product architecture: Có 4 lớp
• Thông tin và phân tích kinh doanh, Truy cập thông tin và tích hợp dữ liệu, Kho dữ liệu, Nguồn dữ liệu.



3.OLTP và OLAP/DW

- OLTP:
• Operational databases (hệ thống kế toán, quản lý bán hàng/sinh viên hệ thống...), ERP, SCM, CRM, ...
• Mục tiêu: thu thập dữ liệu, quy trình được xác định rõ ràng, hầu như không có thay đổi trong quy trình => dữ liệu đồng thời (hỗ trợ nhiều người dùng), toàn vẹn dữ liệu, hiệu suất đọc/ghi.
- OLAP => Kho dữ liệu
• Mục tiêu: hỗ trợ các quyết định trung hạn và dài hạn (phân tích bán hàng theo vùng, theo thời gian xác định, theo sản phẩm, nhóm sản phẩm...);

OLTP	OLAP
Application orientated: Table và views được tối ưu hóa để làm cho ứng dụng chạy nhanh hơn.	Subject orientated. Table là được mô phỏng theo các khái niệm kinh doanh và được thiết kế cho khả năng sử dụng.
Non integrated. Dữ liệu cho khác nhau các ứng dụng kinh doanh (như tài chính so với tiếp thị) thường được lưu trữ trên nhiều hệ thống.	Tích hợp. Tất cả dữ liệu liên quan đến một chủ đề cụ thể (như Khách hàng) được lưu trữ cùng nhau.
Volatile. Dữ liệu được cập nhật mỗi lần giao dịch xảy ra. Hồ sơ được chỉnh sửa tại chỗ trong cơ sở dữ liệu.	Non-volatile. Dữ liệu hiếm khi được cập nhật hoặc bị xóa. Họ gần như luôn luôn chỉ được thêm vào.
Little summary data. Dữ liệu là chuẩn hóa để tối ưu hóa cho hiệu suất. Không có nơi lưu trữ các giá trị roll-up.	Multiple granularity with summaries. Dữ liệu được tóm tắt tại mức độ chi tiết khác nhau để cung cấp thời gian phản hồi thích hợp cho số lượng lớn khối lượng dữ liệu giao dịch.
Non-time variant. Chứa dữ liệu mà đại diện cho trạng thái hiện tại của doanh nghiệp.	Time variant. Giữ dữ liệu cho một số khoảng thời gian để tăng trưởng hữu ích có thể thực hiện những so sánh.

Chương 2: Kho dữ liệu

1. Định nghĩa

Kho dữ liệu là một hệ thống: lấy dữ liệu và hợp nhất dữ liệu định kỳ từ hệ thống nguồn thành kho lưu trữ dữ liệu theo chiều hoặc chuẩn hóa. Nó thường lưu giữ nhiều năm lịch sử và được truy vấn về thông tin kinh doanh hoặc các hoạt động phân tích khác. Nó thường được cập nhật theo đợt, không phải mỗi khi giao dịch xảy ra trong hệ thống nguồn.

Kho dữ liệu là một hệ thống trích xuất, làm sạch, tuân thủ (tải) và cung cấp dữ liệu nguồn vào kho lưu trữ dữ liệu thứ nguyên, sau đó hỗ trợ và thực hiện truy vấn và phân tích nhằm mục đích ra quyết định.

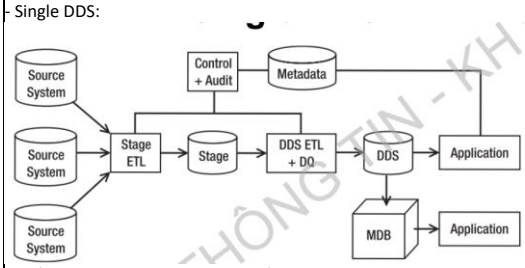
"DW là một tập hợp dữ liệu theo định hướng chủ đề, tích hợp, thay đổi theo thời gian và không thay đổi để hỗ trợ quá trình ra quyết định của ban quản lý".

2. Đặc điểm

- Truy xuất dữ liệu - Việc truy xuất dữ liệu được thực hiện bởi một tập hợp các quy trình được biết đến rộng rãi như hệ thống ETL.
- Hợp nhất dữ liệu: Tính khả dụng của dữ liệu, Phạm vi thời gian, So khớp.
- Lịch sử: Hầu hết các hệ thống giao dịch đều lưu trữ một số lịch sử, nhưng hệ thống kho dữ liệu lưu trữ lịch sử rất lâu.
- Định kỳ: Việc truy xuất, tổng hợp dữ liệu không chỉ diễn ra một lần; chúng xảy ra nhiều lần và thường đều đặn, chẳng hạn như hàng ngày hoặc vài lần trong ngày.
- Cập nhật theo đợt: người dùng không thể cập nhật, xóa dữ liệu trong kho dữ liệu. Dữ liệu DW được cập nhật bằng cơ chế tiêu chuẩn gọi là ETL vào những thời điểm nhất định.
- Hướng thực thể: Dữ liệu chứa trong kho dữ liệu chủ yếu liên quan đến các thực thể chính cần phân tích
- Tích hợp: Dữ liệu có nguồn gốc từ các nguồn khác nhau được tích hợp và đồng nhất khi chúng được tải vào kho dữ liệu
- Time-variant: Tất cả dữ liệu được nhập vào kho dữ liệu đều được gắn nhãn khoảng thời gian mà chúng tham chiếu.
- Non-volatile: Dữ liệu sau khi được nhập vào kho dữ liệu, người dùng không thể thay đổi, cập nhật dữ liệu.

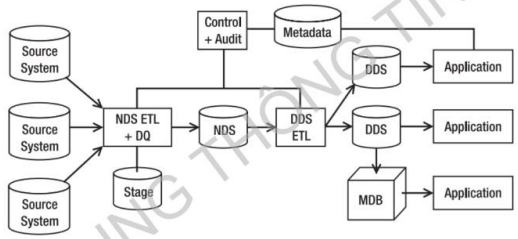
3. Data flow architecture

- Single DDS:



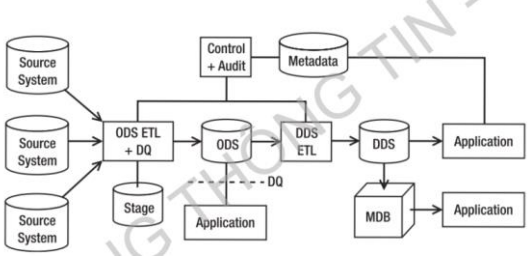
+ Kiến trúc DDS (lưu trữ dữ liệu chiều) đơn có kho dữ liệu giai đoạn và DDS
+ Stage: là nơi bạn lưu trữ tạm thời các dữ liệu đã trích xuất từ hệ thống cửa hàng trước khi xử lý tiếp.
+ Control + Audit: quản lý các quy trình ETL và ghi lại kết quả thực hiện ETL
+ Meta data: chứa các định nghĩa về dữ liệu, hệ thống liên quan, thông tin kiểm toán…
+ DQ Database: lưu trữ dữ liệu xấu do tường lửa dữ liệu phát hiện, thông báo cho người chịu trách nhiệm về chất lượng dữ liệu (DQ)
+ DW application: đọc dữ liệu trong DDS và chuyển tới người dùng.
+ Stage, DDS và các quy trình làm sạch dữ liệu được kết hợp thành một ETL.
+ DDS trong kiến trúc DDS đơn là kho lưu trữ dữ liệu chính. Nó chứa một bộ dữ liệu hoàn chỉnh trong kho dữ liệu bao gồm tất cả các phiên bản và tất cả dữ liệu lịch sử -> ETL package phức tạp, khó khăn khi tạo DDS khác.

- NDS + DDS:



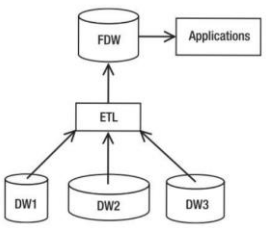
+ NDS là kho lưu trữ dữ liệu chính, nghĩa là NDS chứa các bộ dữ liệu hoàn chỉnh, bao gồm tất cả dữ liệu lịch sử giao dịch và tất cả các phiên bản lịch sử của dữ liệu chính -> Dạng 3NF trở lên -> ETL sang DDS đơn giản.
+ NDS là kho lưu trữ dữ liệu nội bộ, nghĩa là người dùng cuối hoặc ứng dụng của người dùng cuối không thể truy cập được. – MDB: csdl đa chiều

- ODS + DDS:



+ The ODS (operational data store) + DDS architecture has stage, ODS, and DDS data stores.
+ ODS dạng chuẩn 3NF hoặc hơn. Giống như NDS, ODS chứa dữ liệu hiện tại, không chứa dữ liệu lịch sử. Trong kiến trúc ODS + DDS, bạn chỉ có một DDS. ODS là kho lưu trữ dữ liệu kết hợp → người dùng cuối và ứng dụng của người dùng cuối có thể truy cập → ODS có thể cập nhật.
+ Chỉ có 1 DDS cho mục đích hỗ trợ khách hàng

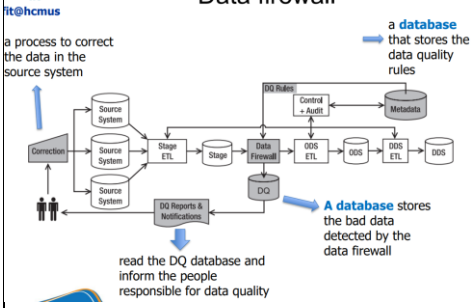
- Fedarate data warehouse



- Là một kiến trúc được sử dụng để tích hợp các kho dữ liệu không đồng nhất nhằm cung cấp một phiên bản sự thật duy nhất trong toàn bộ tổ chức
- Được dùng trong các tổ chức lớn, việc kinh doanh được mở rộng ra nhiều quốc gia, nhiều khu vực, phân bổ các bộ phận IT vào từng quốc gia, khu vực để phát triển hệ thống.
- Phân loại theo: vị trí địa lý, chức năng kinh doanh, pháp nhân kinh doanh.

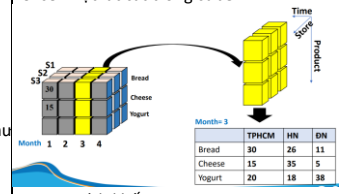
Chương 3: Data Modeling
1. Các bước Data Modeling
• Bước 1: Chọn thuộc tính cần thống kế
có bao nhiêu viên Paracetamol và Diclofenac được bán ra từ một cửa hàng MedPlus mỗi ngày
• Bước 2: Xác định mức độ chi tiết, What, Who, Where, How
Cửa hàng MedPlus bán 1.000 viên Paracetamol vào một ngày cụ thể thì độ hạt là hàng ngày và 10.000 viên vào một tháng cụ thể thì độ hạt sẽ là hàng tháng.
• Bước 3: Xác định các bảng chiều và các thuộc tính của bảng chiều.
Mua sắm", "Thuốc" và "Ngày"
• Bước 4: Xác định bảng Fact
2. Thực hiện mô hình hoá DDS theo mô tả.
- Phân tích yêu cầu đề bài -> Suy ra cả bảng Dim
- Thiết kế DDS, bảng Fact
• Bối cảnh, event
• Dim
+ Phân cấp chiều, Quản lý khoá
• Mô hình hóa Fact: Quản lý khoá (NK, SK), Chi tiết dữ kiện, Measure: có sẵn/ cần tính toán nạp vào, Loại dl (additive, semi-….)
• Mô hình hóa Dim: SK hay NK (chủ yếu là SK), chiều thay đổi loại?
• Schema (star, snowflake, chòm sao….)
3. Các phương pháp lưu giá trị thay đổi của chiều.
- 3 loại thay đổi:
+ Loại 1: ghi đè, cập nhật giá trị mới, không lưu giá trị cũ
+ Loại 2: lưu lại giá trị lịch sử, có cột status để lưu trạng thái dòng dữ liệu, bất hoạt dòng cũ bằng 1 thuộc tính trạng thái.
+ Loại 3: số giá trị thay đổi là cố định, biết trước để xác định số cột lưu trữ sẵn.
4. Cái loại Measure
- Additive Facts: có thể được tổng hợp theo bất kỳ chiều nào liên quan đến bảng Fact.
- Semi-additive Facts: có thể được tính toán theo một số chiều, nhưng không phải tất cả
- Non-Additive Facts: việc tính toán không sử dụng giá trị từ bảng chiều vì không có ý nghĩa.
• Measure bảng Fact: Là giá trị đo lường (measure) thực sự của 1 hoạt động kinh doanh như: tiền lời từ việc bán hàng, hay số lượng đặt hang
• Mỗi sự đo lường đều có 1 tính chất hạt(grain) – chỉ cấp độ chi tiết trong việc đo lường 1 sự kiện (đơn vị đo lường, tiề tệ, số dư tài khoản cuối ngày,...)
– Tính chất hạt của tiền tệ có thể là: dollar amount, hoặcchi tiết hơn: xu (cents)
– Tính chất này được quyết định bởi nguồn dữ liệu
• Tính hạt (GRANULARITY)
– Dùng để chỉ mức chi tiết được lưu trữ trong bảng fact.
– Tính hạt càng cao -> giới hạn khả năng lấy thông tin mức chi tiết, không drill down xuống mức thấp hơn được.
– Tính hạt thấp -> mở rộng kích thước của kho dữ liệu so với nhu cầu, thiết kế nhiều hơn và tốn chi phí roll up
5. Bài tập ETL
- Phân tích yêu cầu đề bài -> Suy ra cả bảng Dim
- Thiết kế DDS, bảng Fact
– Sự kiện:
• Khi 1 khách hàng mua 1 sản phẩm
– Bối cảnh sự kiện:
• Ai: khách hang
• Ở đâu: cửa hàng, lãnh thổ bán hàng
• Cái gì: sản phẩm
• Khi nào: ngày mua hang
– Đo lường (dữ kiện): Số lượng, đơn giá, giá trị.
• Các giá trị có sẵn từ nguồn: Quantity, unit_price, unit_cost
• Các giá trị phải tính toán: Sales_value, sales_cost, margin
+ sales_value = unit_price x quantity
+ sales_cost = unit_cost x quantity
+ Margin = sales_value – sales_cost
• Cấp chi tiết dữ liệu (độ mịn)
+ Đơn vị nhỏ nhất xảy ra sự kiện: Một dòng trong fact tương ứng mỗi item được bán.

Chương 4: ETL
1. Các phương pháp ETL
- **Incremental extract (rút trích lũy tiến):**
+ Sử dụng cột create_at, update_at để lưu thời gian.
+ Mỗi khi hàng trong bảng thay đổi (chèn/cập nhật), dấu thời gian được cập nhật.
+ Status: trạng thái của dòng ghi nhận dòng đã xoá(không xoá thực sự)
+ Logic trích rút lũy tiến sử dụng LSET và CET
LSET: thời điểm dữ liệu được trích xuất lần cuối.
CET là thời điểm thực thi ETL packages (rút trích hiện tại)
Các bước:
1. Lấy thông tin LSET được lưu trong metadata
2. Lấy CET: thời gian khởi động ETL package
3. Rút trích dữ liệu:
select * from order_header where (created >= LSET and created < CET) or (last_updated >= LSET and last_update < CET).
4. Cập nhật LSET = CET
Giải pháp 3: sử dụng thuộc tính tự tăng (identity), last updated, order status
Trường hợp dò tìm thêm mới
– Lấy ID được rút trích sau cùng nhất (LSEI) từ csdl metadata
– Lấy max(orderID) từ bảng hoá đơn, gán vào CEI (ID rút trích hiện tại)
– Lấy tập các dòng nằm giữa LSEI và CEI như sau:
– Select * from order_header where order_id >= LSEI and order_id < CET
– Gán LSET mới = CET
– Trường hợp dò tìm cập nhật: tương tự giải pháp 1 (sử dụng last updated).
- **Fix ranged**:
Giải pháp: rút trích 1 số lượng chính xác các dòng hoặc theo 1 khoảng thời gian cụ thể dựa vào ràng buộc nghiệp vụ.
– Không thể rút trích toàn bộ bảng vì khối lượng quá lớn
– Không thể rút trích incremental do: Không có thuộc tính nhãn thời gian, nhãn thời gian không tin cậy. Thuộc tính tự tăng không tin cậy. Không thể cài trigger trên bảng source
- **Whole table:**
• Tình huống: Kích thước bảng nhỏ, Không có nhãn thời gian, Không có thuộc tính tự tăng, Không có ràng buộc nghiệp vụ
• Giải pháp: Rút trích toàn bộ dữ liệu nguồn
2. PP lưu stage
Ba cách tiếp cận:
1. Giữ dữ liệu của ngày hôm trước trong cùng một bảng
2. Giữ mỗi ngày trong một bảng riêng
3. Chỉ sử dụng một bảng và cắt bớt bảng mỗi lần trước khi tải
3. Data Firewall
Tường lửa dữ liệu là chương trình kiểm tra dữ liệu đến, tương tự như khái niệm tường lửa trong mạng → đảm bảo chất lượng dữ liệu
- Về mặt vật lý, nó là một gói SSIS hoặc một thủ tục lưu sẵn
- Đặt tường lửa dữ liệu giữa giai đoạn và
- từ chối dữ liệu (không tải nó vào DW),
- cho phép dữ liệu (tải nó vào DW)
- sửa dữ liệu (sửa dữ liệu trước khi tải vào DW)



4. Meta Data
1. Data definition and mapping metadata: Chứa ý nghĩa của từng cột thực tế và thứ nguyên cũng như nguồn gốc của dữ liệu. Các cột trong bảng kho dữ liệu phục vụ các mục đích khác nhau.
2. Data structure metadata:
3. Source system metadata
4. ETL process metadata: mô tả cấu trúc của các bảng trong mỗi Store, từng data flow trong các quy trình ETL.
5. Data quality metadata
6. Audit metadata:
7. Usage metadata

Chương 5: OLAP
1. Cú pháp MDX
- select {[Order Date].[Hierarchy].[Year],[Order Date].[Hierarchy]} on columns,
[Dim Product].[Hierarchy].[Product Subcategory Key] on rows
from [Adventure Works DW2012];
--hidding null
select non empty{[Order Date].[Hierarchy].[Year],[Order Date].[Hierarchy]} on columns,
non empty {[Dim Product].[Hierarchy].[Product Subcategory Key], [Dim Product].[Hierarchy]} on rows
from [Adventure Works DW2012];
--display different measure
select non empty{[Order Date].[Hierarchy].[Year],[Order Date].[Hierarchy]} on columns,
non empty [Dim Product].[Hierarchy].[Product Subcategory Key on rows
from [Adventure Works DW2012]
where [Measures].[Sales Amount - Fact Reseller Sales]
-- dimension.hierarchy.level.members
Select non empty [Dim Product].[Hierarchy].[Product Subcategory Key].members on columns
from [Adventure Works DW2012]
--combine two level
Select non empty{[Dim Product].[Hierarchy].[Product Subcategory Key].members,
[Dim Product].[Hierarchy].[ALL]} on columns
from [Adventure Works DW2012]
• To sort your rows (or columns), you employ the Order function
Order (param 1, param 2, option): Param 1: the set of rows to sort, Param 2: is the measure to sort by
• Let's filter the product subcategories to hide those with a null (or zero)
Filter (param_1, param_2): The first parameter for Filter is the set of members you wish to filter.
The second parameter is a Boolean test that returns true or false for each member of the set.
2. OLAP Operational
- Roll up/ Drill up: Đi lên (ví dụ day -> month)
- Drill down: Đi xuống (month -> day)
- Slice: một lát cát trong cube



- Dice – trích khối con:

- Pivot: là quá trình bạn xoay cái Cube dữ liệu để xem tất cả các mặt các khía cạnh mà nó mô tả. Ví dụ với dữ liệu dân số, bạn sẽ muốn Pivot dữ liệu theo ngày tháng, theo thành phố, theo giới tính…
3. Phép OLAP.
- Chiều …: roll-up/drill-down.
- Slice khi cần lấy một dữ liệu của thể ở 1 chiều.
- Measure là gì?