Contents

[I/ 2.5 points: theory - 3, 4 terms (slide) 1](#_Toc92643688)

[1. Định nghĩa DW 1](#_Toc92643689)

[a. Khái niệm 1](#_Toc92643690)

[b. Một số loại DW 1](#_Toc92643691)

[c. Mục đích của Data Warehouse 1](#_Toc92643692)

[d. Lợi ích từ Data Warehouse 1](#_Toc92643693)

[e. Hướng tiếp cận Data Warehousing 1](#_Toc92643694)

[2. Thành phần của DW. 1](#_Toc92643695)

[a. Data Warehouse Database 1](#_Toc92643696)

[b. Sourcing, Acquisition, Clean-up and Transformation Tools (ETL) 1](#_Toc92643697)

[c. Metadata 1](#_Toc92643698)

[d. Query Tools 1](#_Toc92643699)

[e. Data Marts 1](#_Toc92643700)

[3. DW Tools 1](#_Toc92643701)

[a. Amazon Redshift 1](#_Toc92643702)

[b. Microsoft Azure 1](#_Toc92643703)

[c. Google BigQuery 1](#_Toc92643704)

[d. Snowflake 1](#_Toc92643705)

[4. Tìm ví dụ additive, semi, non measure 1](#_Toc92643706)

[5. Slides 2](#_Toc92643707)

[a. 3 types of Fact Table 2](#_Toc92643708)

[b. DW architecture 2](#_Toc92643709)

[c. Slowly Changing Dimension (SCD) 2](#_Toc92643710)

[d. Granularity 3](#_Toc92643711)

[e. Solving problems related to PK 3](#_Toc92643712)

[II/ 4 points: Source (database: entity relationships) >> DDS 3](#_Toc92643713)

[1. (Taken note in book**\***) 3](#_Toc92643714)

[a. Hiện trạng 3](#_Toc92643715)

[b. Nhu cầu 3](#_Toc92643716)

[c. Yêu cầu 3](#_Toc92643717)

[d. Kiến trúc cài đặt 4](#_Toc92643718)

[e. Phân tích yêu cầu 4](#_Toc92643719)

[f. Mô hình hóa 6](#_Toc92643720)

[g. Thiết kế NDS 7](#_Toc92643721)

[h. Thiết kế DDS 8](#_Toc92643722)

[2. Case study in slide 9](#_Toc92643723)

[a. Analyze 9](#_Toc92643724)

[b. Drawing 9](#_Toc92643725)

[III/ DDS >> write OLAP, MDX. 10](#_Toc92643726)

[1. OLAP 10](#_Toc92643727)

[a. OLAP Server 10](#_Toc92643728)

[b. Show the revenue of Region (Canada, Australia) in 2015 & 2016 according to each Sub-Product category: 10](#_Toc92643729)

[c. Show the revenue of Region (Canada, Australia) in February of 2015 according to each Sub-Product category: 10](#_Toc92643730)

[2. MDX queries 11](#_Toc92643731)

3 questions in 90’:

# I/ 2.5 points: theory - 3, 4 terms (slide)

## 1. Định nghĩa DW

### a. Khái niệm

DW (kho dữ liệu) là một hệ thống thông tin lưu trữ dữ liệu lịch sử (historical), dữ liệu thay đổi (commutative) từ một hay nhiều nguồn khác nhau nhằm phục vụ báo cáo, khai thác phân tích dữ liệu, và là một phần cốt lõi của Trí tuệ kinh doanh.

Ví dụ: Ngành hàng không, ngành tài chính, chăm sóc sức khỏe, chuỗi cửa hàng bán lẻ…

### b. Một số loại DW

Enterprise Data Warehouse

Operational Data Store

Data Mart

### c. Mục đích của Data Warehouse

Báo cáo và phân tích

Lưu trữ những dữ liệu trong quá khứ của doanh nghiệp

Là nền tảng để đưa ra quyết định

### d. Lợi ích từ Data Warehouse

Hiểu được xu thế của doanh nghiệp và dự đoán

Được thiết kế để xử lý tốt khối lượng dữ liệu khổng lồ

Cấu trúc của KDL giúp người dùng cuối dễ điều hướng, hiểu và truy vấn

Những truy vấn phức tạp trên các CSDL thông thường có thể dễ xây dựng và duy trì trên KDL

KDL là phương pháp hiệu quả để quản lý yêu cầu thông tin từ nhiều người dùng

KDL cung cấp khả năng phân tích khối lượng lớn dữ liệu lịch sử

### e. Hướng tiếp cận Data Warehousing

Có 2 hướng tiếp cận cơ bản:

**Top down:** Xây dựng một Data Warehouse bao quát tất cả những báo cáo cần thiết cho doanh nghiệp.

**Bottom up:** Xây dựng những Data Mart nhỏ, phục vụ việc báo cáo cho những phòng ban hay chức năng có nhu cầu.

## 2. Thành phần của DW.

### a. Data Warehouse Database

### b. Sourcing, Acquisition, Clean-up and Transformation Tools (ETL)

### c. Metadata

### d. Query Tools

### e. Data Marts

## 3. DW Tools

### a. Amazon Redshift

### b. Microsoft Azure

### c. Google BigQuery

### d. Snowflake

## 4. Tìm ví dụ additive, semi, non measure

|  |  |  |
| --- | --- | --- |
| **Loại** | **Định nghĩa** | **Ví dụ** |
| Additive | Là các tiêu chí có thể thực hiện phép tính cộng mà ý nghĩa vẫn chính xác. | Một cửa hàng bán lẻ và nếu chúng ta muốn xác định tổng doanh số đã xảy ra trong sáu tháng qua, chúng tôi có thể lấy nhóm các bản ghi của sáu tháng qua và lấy giá trị tổng hợp. Trong trường hợp này, Doanh số hoặc Số tiền bán hàng là một dữ kiện phụ, vì chúng ta có thể tổng hợp (nhóm) các bản ghi dữ kiện với các trường liên quan khác (sản phẩm, khách hàng, nhà cung cấp, v.v.) có trong bảng dữ kiện.. |
| Một công ty vận tải hành khách chúng ta muốn xác định trong 1 năm qua đã vận chuyển được bao nhiêu hành khách với bao nhiêu chuyến đi thì ta có thể cộng dồn lại. |
| Semi-additive | Là tiêu chí có thể cộng dồn nhưng phải trên tất cả các chiều | Số lượng tồn kho, có thể cộng dồn theo tiêu chí thành phố, vùng để biết số lượng tồn kho ở nơi đó. Nhưng tiêu chí này kết hợp với chiều thời gian sẽ mất ý nghĩa vì nó chỉ mang tính thời điểm. |
| Số dư trong tài khoản ngân hàng không thể tổng hợp theo chiều thời gian. |
| Non-additive | Là các tiêu chí không thể được cộng dồn.  *Ví dụ: tỉ lệ, trung bình, trung vị, số sản phẩm khác nhau.* | Tỉ lệ người đã được tiêm đủ hai mũi vaccine ở một thành phố trong một ngày.  Không thể cộng dồn khi thống kê theo tuần/tháng/năm. |
| Số lượng người dùng hoạt động hằng ngày.  Khi thống kê theo thời gian, cũng không thể cộng dồn lại |

## 5. Slides

### a. 3 types of Fact Table

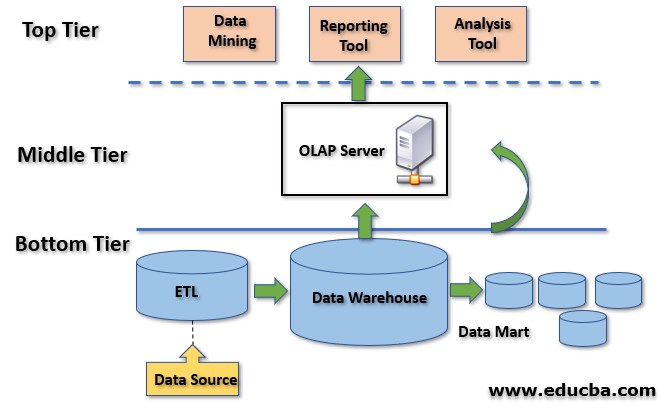
* **Transaction:** basic of our reporting (default type)
* **Periodic Snapshot Tables:** KPI (report weekly, monthly…)
* **Accumulating Snapshot Tables:** “series of steps” >> Utilize BI to determine which steps are the most problematic in their production process.

### b. DW architecture

Single-tier

Two-tier

Three-tier: this is the most widely used. It consists of the Top, Middle and Bottom Tier



### c. Slowly Changing Dimension (SCD)

Is a technique used to store the historical value of dimension attributes.

* Type 0: Fixed values (change -> fault)
* Type 1: Overwrite old values
* Type 2: Add new row and inactive old rows (set state = 0)
* Type 3: Add new column (using in case you know accurately the amount of columns. Ex: name, nationality…)

SSIS supports Type 0 + 1 + 2.

### d. Granularity

Every measurement has a grain, which is the level of detail in the measurement of an event.

***Granularity*** *is determined by its data source, indicates the level of detail stored in a table.*

Ex:

* “Each **time** record” represents a day.
* “Each **product** record” represents a product.
* “Each **organization** record” represents one branch.

Then, the grain of a sales Fact Table with these dimensions is sales per **product** per **day** per **branch**.

Additional details:

* Too high level >> limit the ability of users to obtain additional detail.
* Too low level >> results in an exponential increase in the size requirement of the DW.

### e. Solving problems related to PK

PK may be **inconsistent**, **incompatible**, **changing**…

* *Solutions*: Surrogate Keys (SK), source system’s natural key…

# II/ 4 points: Source (database: entity relationships) >> DDS

(dim, fact, granularity, types of attribute, calculated attribute > fomular)

## 1. (Taken note in book**\***)

### a. Hiện trạng

Khoa CNTT bao gồm nhiều giảng viên

* Mỗi giảng viên được phân công dạy nhiều lớp, nhiều môn, nhiều chương (cao đẳng, chất lượng cao, hoàn chỉnh, từ xa ...)
* Mỗi cuối kỳ, sinh viên tiến hành đánh giá các giảng viên (rất tốt, tốt, bình thường, chưa tốt, tệ) qua các tiêu chí cách dạy, sự nhiệt tình, công bố tài liệu/quy định rõ ràng

### b. Nhu cầu

Giáo vụ cần thống kê tình hình giảng dạy của tất cả giảng viên với mỗi môn trong mỗi học kỳ của mỗi năm học

Giáo vụ cần đánh giá giảng viên cuối mỗi học kỳ (khen thưởng, kỷ luật) thông qua các số liệu (thống kê feedback). Tức là thống kê số lượng bao nhiêu sv đánh giá điểm 1, 2, 3, 4, 5 của một giảng viên dạy một môn học, trong một học kỳ của một năm, so với tổng số sv của lớp học.

### c. Yêu cầu

* Cấp quản lý lãnh đạo Khoa có thể nắm bắt được tình hình chung của hoạt động giảng dạy chung của các giáo viên và học viên đến tình hình riêng ở từng lớp/ môn/ bộ môn
* Hệ thống có thể hiển thị các hình, biểu đồ cho phép in được, có thể xuất ra excel hoặc csv..
* Phục vụ phân tích các chiến dịch khảo người học qua việc xem các đo lường của mỗi chiến dịch sau khi gởi cho SV: số lượng thông điệp được gởi của mỗi kênh (mobile message, email, post...), số thông điệp phân phối thành công, thất bại, tỷ lệ mở, tỷ lệ bỏ qua, than phiền, đánh spam, ...
* Phục vụ phân loại người học cho các chiến dịch quảng cáo cho Khoa/ Trung tâm tin học,... dựa vào quyền thông tin, đặc trưng địa lý, thông tin SV, sở thích, ...

### d. Kiến trúc cài đặt

Mô hình NDS + DDS

Fact table thiết kế theo mô hình sao

### e. Phân tích yêu cầu

*5.1. Giáo vụ cần thống kê (****xem*** *các feedback) tình hình giảng dạy của mỗi giảng viên theo chiều: môn học, thời gian: học kỳ của mỗi năm học*

Sự kiện: khi một sinh viên đánh giá một giảng viên

Bối cảnh sự kiện:

* Ai: Sinh viên
* Cái gì: Feedback
* Khi nào: cuối học kỳ của năm học

Đo lường (dữ kiện): Số điểm hài lòng

Dữ liệu có sẵn từ nguồn: Điểm hài lòng

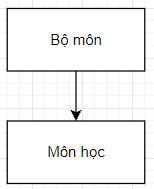
Giá trị phải tính toán: không có

Thuộc tính có ở nhiều nguồn -> lưu thêm SOR\_NK để xác định dòng dữ liệu của nguồn nào đổ vào

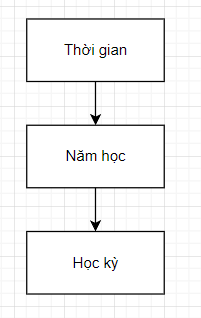
Thuộc tính Create\_timestamp và Update\_timestamp xác định thời gian insert và update dòng dữ liệu đó trong KDL

Các chiều liên quan sự kiện cần phân tích:

* Chiều giảng viên: không cần lưu giá trị cũ vì giảng viên nghỉ việc -> không cần phân tích
* Chiều môn học: lưu giá trị cũ



* Chiều thời gian (năm -> học kỳ)



* Chiều sinh viên: lưu giá trị cũ

*5.2. Giáo vụ cần đánh giá giảng viên cuối mỗi học kỳ (khen thưởng, kỷ luật) thông qua các số liệu (Tức là thống kê* ***số lượng*** *bao nhiêu sv đánh giá điểm 1, 2, 3, 4, 5 của một giảng viên dạy một môn học, trong một học kỳ của một năm, so với tổng số sv của lớp học.)*

Sự kiện: khi một sinh viên đánh giá một giảng viên

Bối cảnh sự kiện:

* Ai: Sinh viên
* Cái gì: Feedback
* Khi nào: cuối học kỳ

Đo lường: Số lượng sinh viên đánh giá điểm 1,2,3,4,5, Điểm theo % số sinh viên/tổng số sinh viên

Dữ liệu có sẵn từ nguồn: Điểm đánh giá

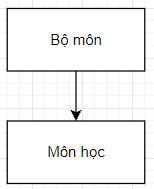
Giá trị phải tính toán: % từng mức điểm (1, 2, 3, 4, 5) = Số sv đánh giá mức điểm đó/tổng số sv của lớp đó.

Thuộc tính có ở nhiều nguồn -> lưu thêm SOR\_NK để xác định dòng dữ liệu của nguồn nào đổ vào (Souce\_ID)

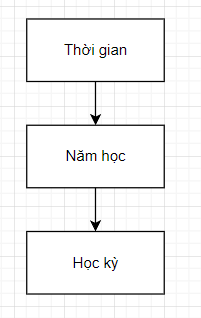
Thuộc tính Create\_timestamp và Update\_timestamp xác định thời gian insert và update dòng dữ liệu đó trong KDL

Các chiều liên quan sự kiện cần phân tích:

* Chiều giảng viên: không cần lưu giá trị cũ vì giảng viên nghỉ việc -> không cần phân tích
* Chiều môn học: lưu giá trị cũ



* Chiều thời gian (năm -> học kỳ)



* Chiều sinh viên: lưu giá trị cũ

### f. Mô hình hóa

* Fact table

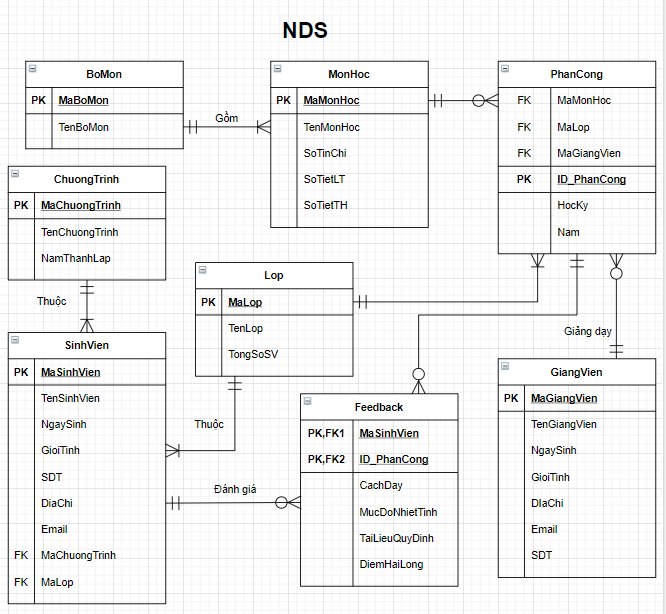
Các giá trị có sẵn: Cách dạy, mức độ nhiệt tình, tài liệu quy định, điểm hài lòng,...

Các giá trị cần phải tính toán: không có

Cấp chi tiết của dữ liệu: mỗi dòng dữ liệu trong bảng Fact tương ứng với một đánh giá của sinh viên

* Thiết kế chiều: Các chiều liên quan đến phân tích: Chiều giảng viên, chiều môn học, chiều sinh viên, chiều thời gian.

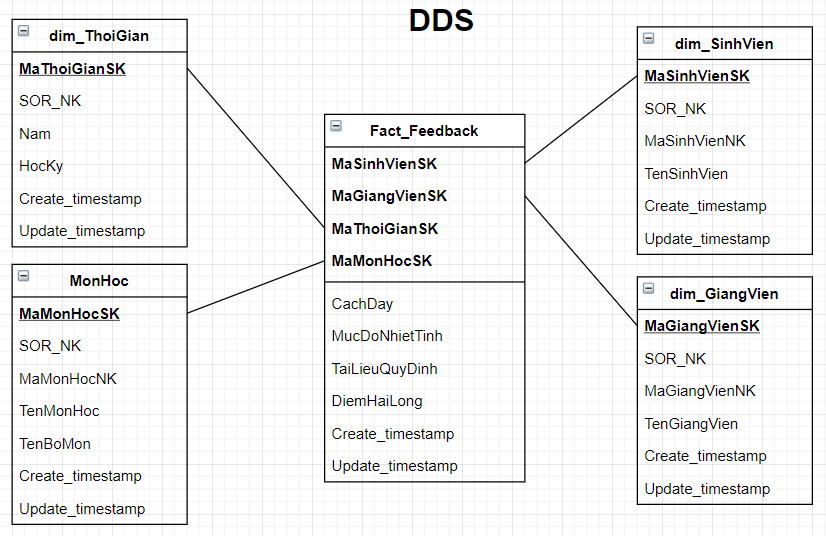
### g. Thiết kế NDS

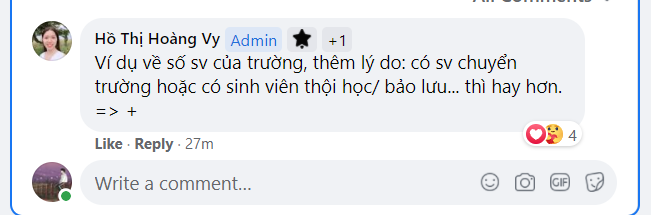


Đánh giá dạng chuẩn cho bảng và lược đồ trong NDS

|  |  |  |
| --- | --- | --- |
| STT | Tên bảng | Dạng chuẩn |
| 1 | Chương trình | BCNF |
| 2 | Môn học | BCNF |
| 3 | Giảng viên | BCNF |
| 4 | Lớp | BCNF |
| 5 | Sinh viên | BCNF |
| 6 | Phân công | 3NF |
| 7 | Feedback | 3NF |
| 8 | Bộ môn | BCNF |
| Cả lược đồ | | 3NF |

### h. Thiết kế DDS





## 2. Case study in slide

### a. Analyze

User would like to analyze product sales. He would like to acquire revenue, costs, profit according to $.

* Event:

When a customer buy a product.

* Event context: (-> Find out Dimension tables)

Who: customer

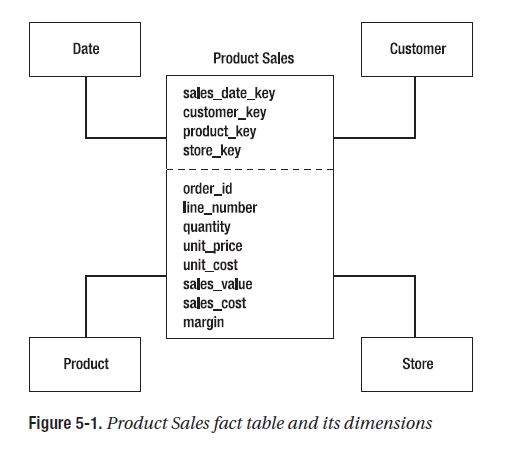
Where: shope, area

What: product

When: purchasing date.

* Measures: amount, cost, value (-> Add to Fact Tables)

### b. Drawing



* Available values: Quantity, unit\_price, unit\_cost.
* Calculation values: Sales\_value, sales\_cost, margin.
* Sales\_value = unit\_price \* quantity
* Sales\_cost = unit\_cost \* quantity
* Margin = sales\_value – sales\_cost
* Granularity (Least unit of happening event): 1 row in fact is representing each sold item.
* Related dimensions: Date, Product, Customer, Store.
* SCD (Old values of dimension):
* Product dimension: no need >> SCD 1
* Customer dimension: need to >> SCD 2
* Store dimension: knowing accurately the changing times >> SCD 3
* Data hierachy:

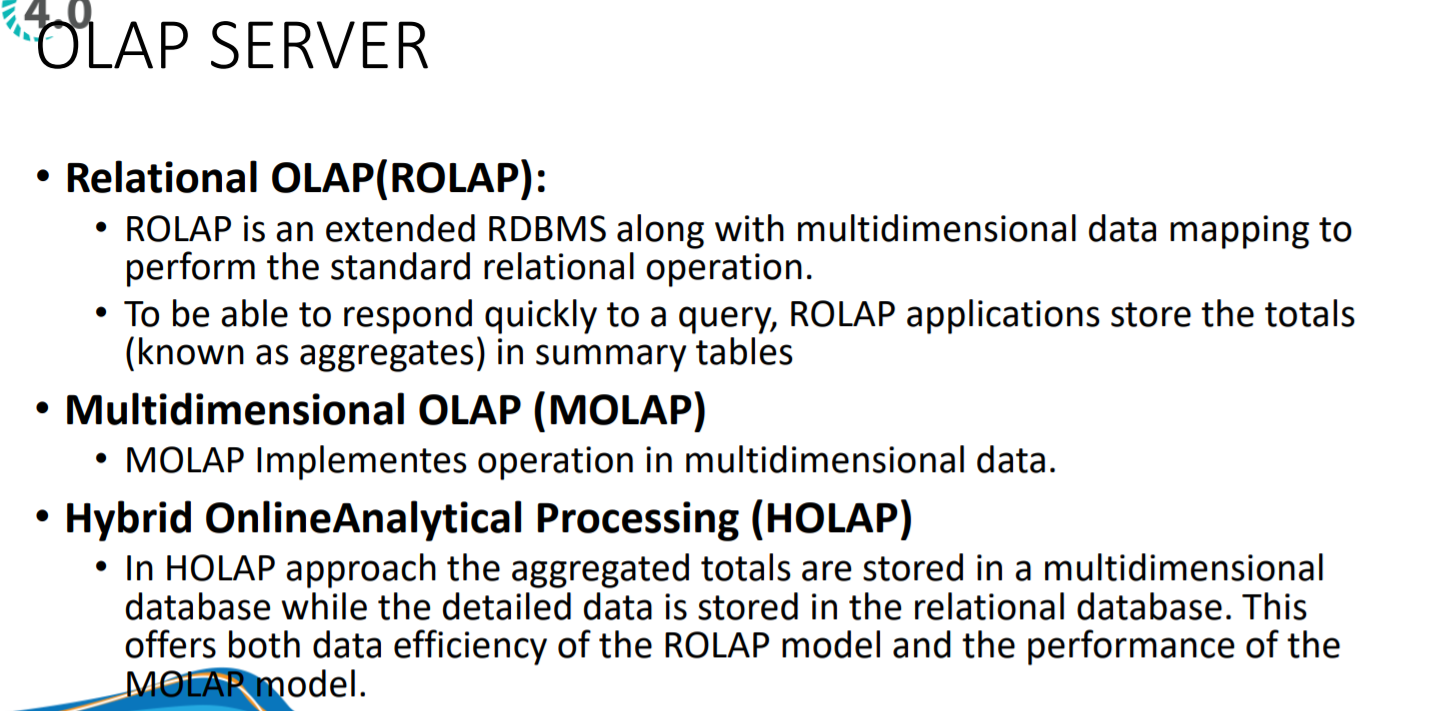
*\*According to each Date, Store, Customer, Product*

# III/ DDS >> write OLAP, MDX.

***Dice:*** *filter data more than 2 dimensions*

## 1. OLAP

### a. OLAP Server



### b. Show the revenue of Region (Canada, Australia) in 2015 & 2016 according to each Sub-Product category:

> Các phép OLAP thực hiện:

- Chiều Location: rollup từ Store lên Region;

- Chiều Date: rollup từ Day -> Year;

- Chiều Product: rollup từ Product lên Sub-Product;

- Dice Region theo (Canada & Australia) và chiều Date theo Year = 2015 and Year = 2016.

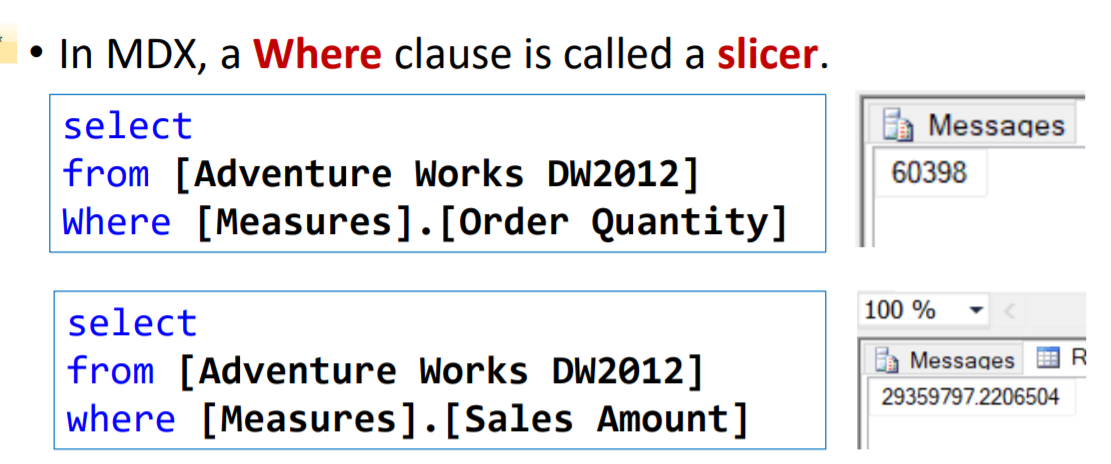
### c. Show the revenue of Region (Canada, Australia) in February of 2015 according to each Sub-Product category:

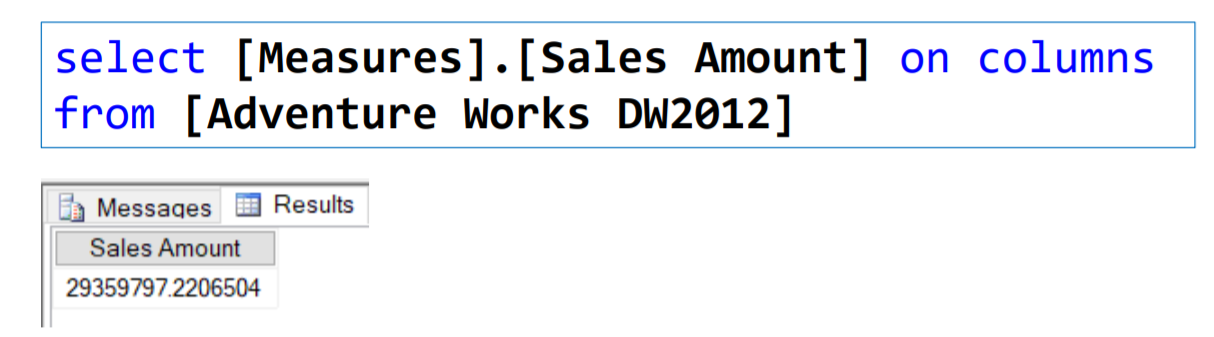
- Chiều Date: rollup lên Year;

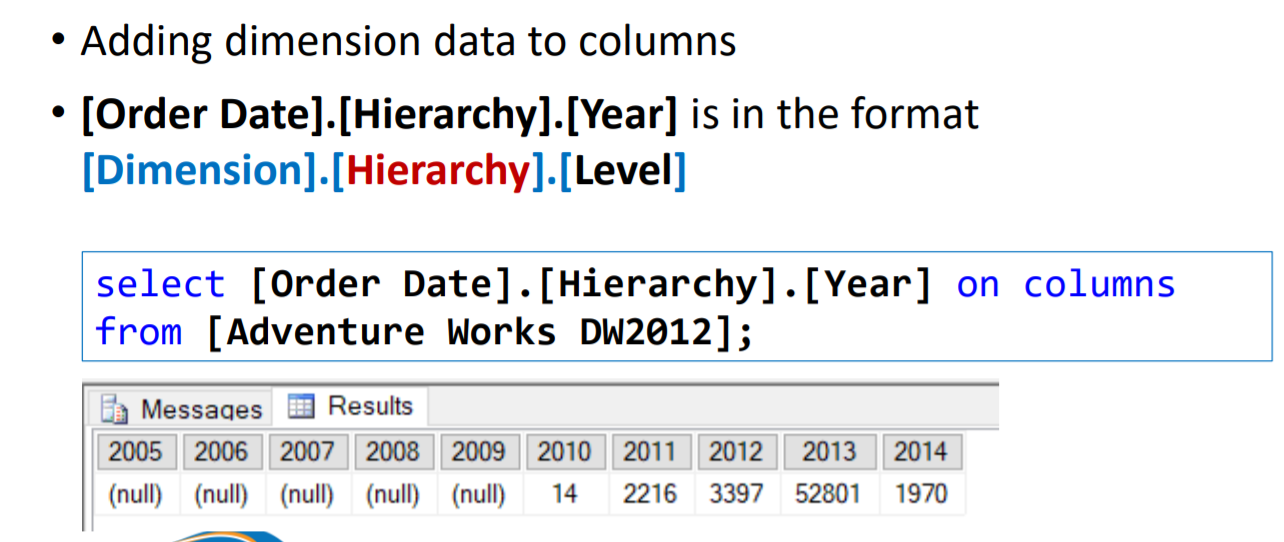
- Drilldown Month;

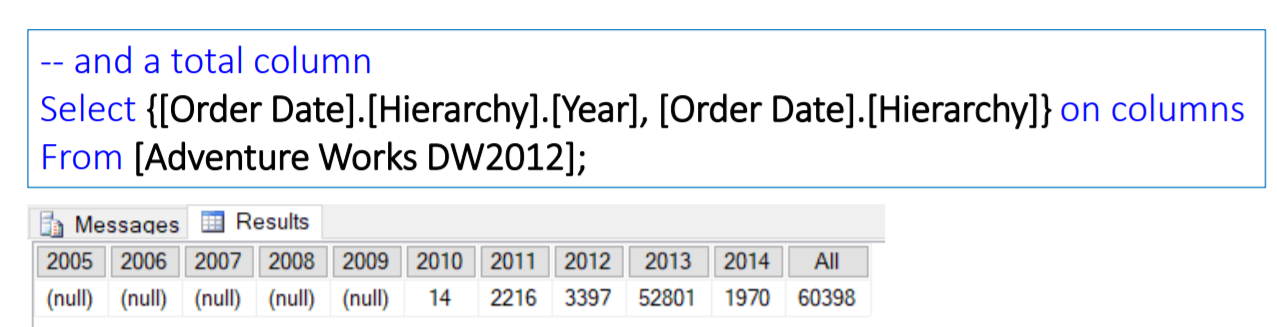
- Dice Region theo (Canada, Australia) và chiều Date theo Year = 2015 and Month = 2.

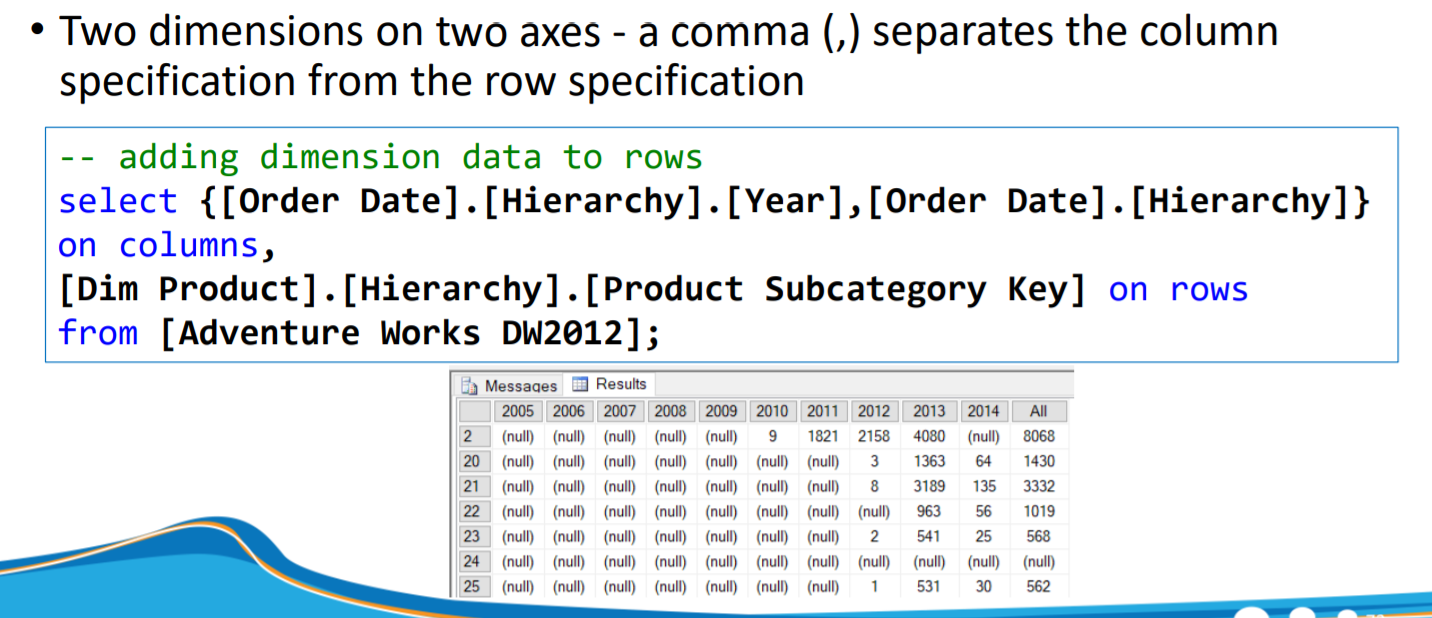
## 2. MDX queries

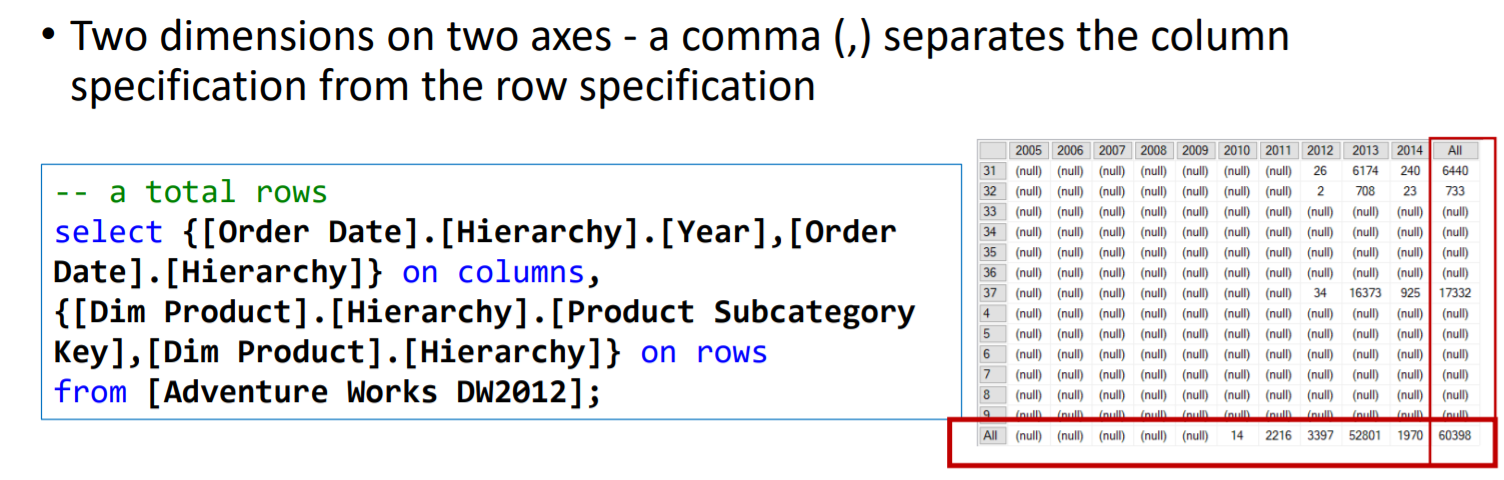


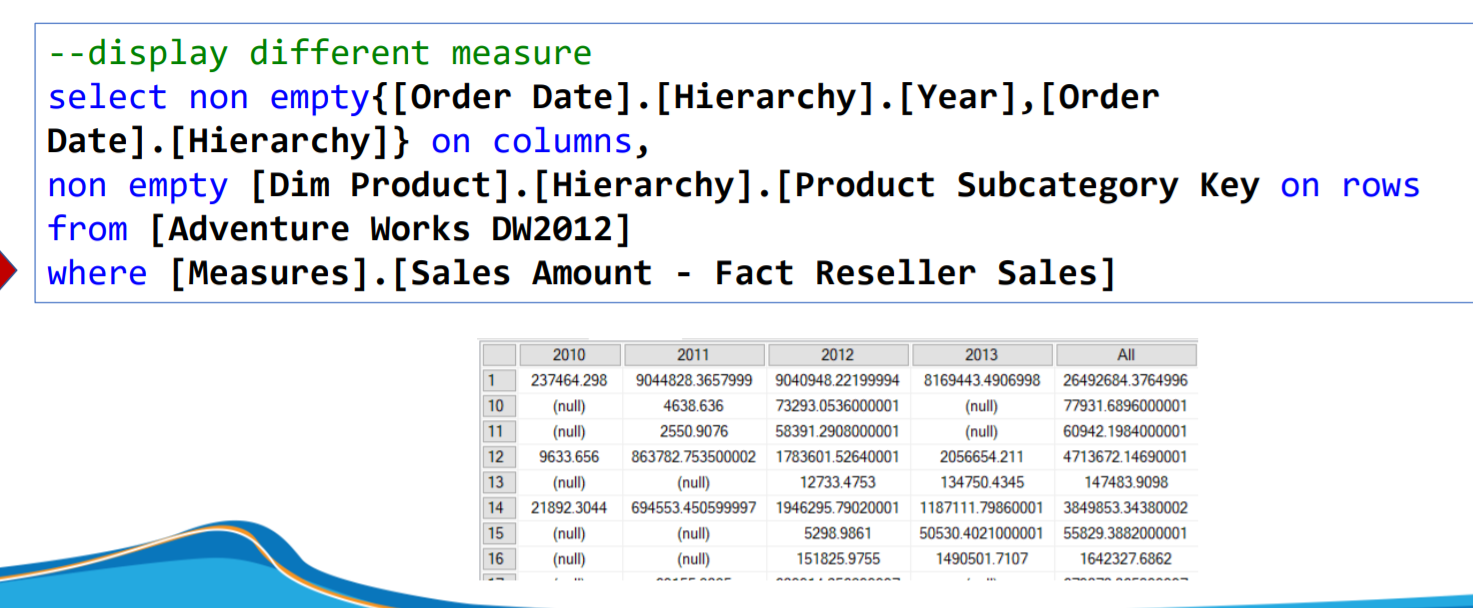


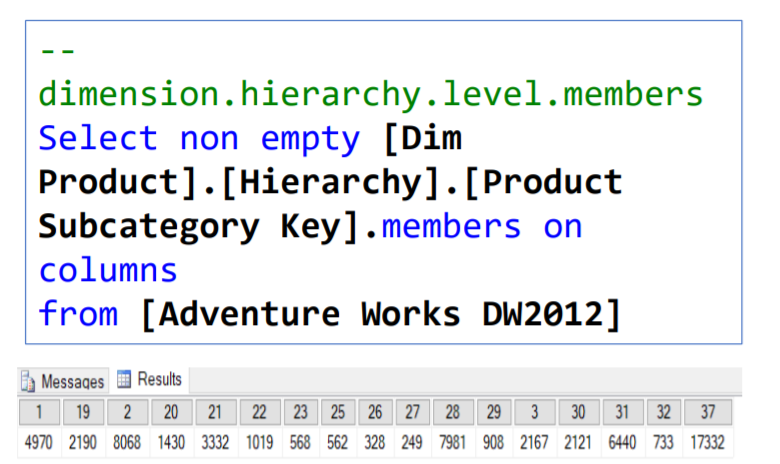


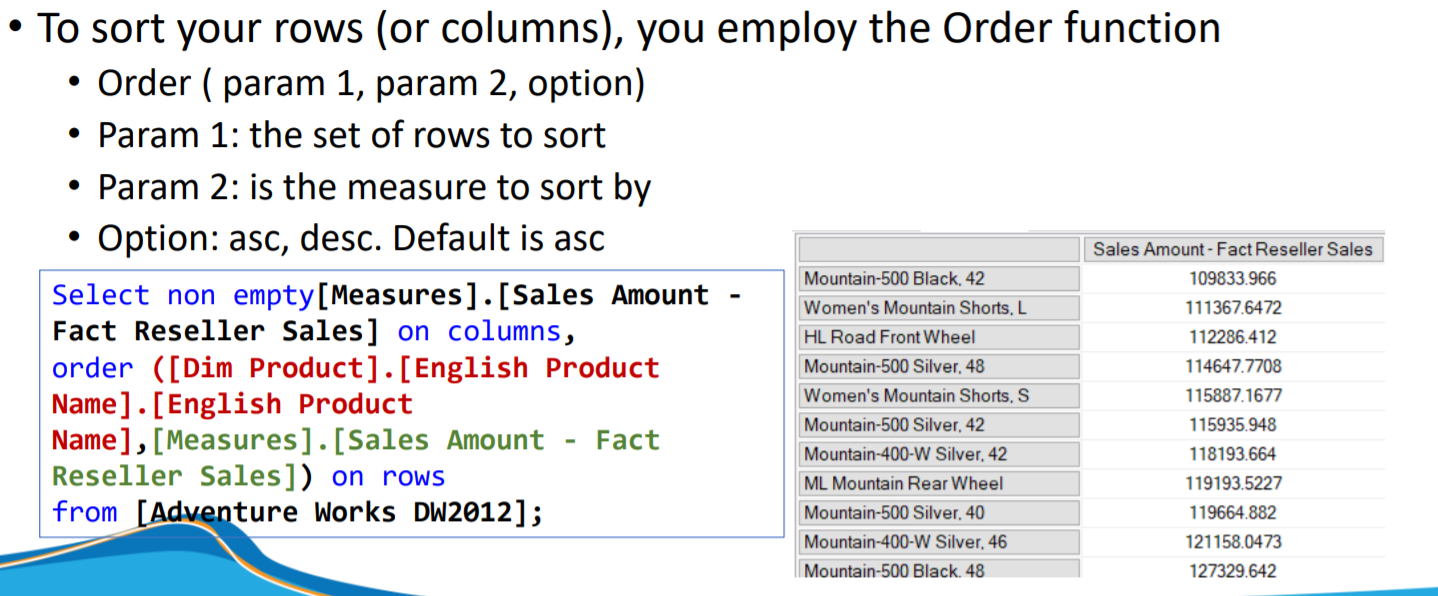


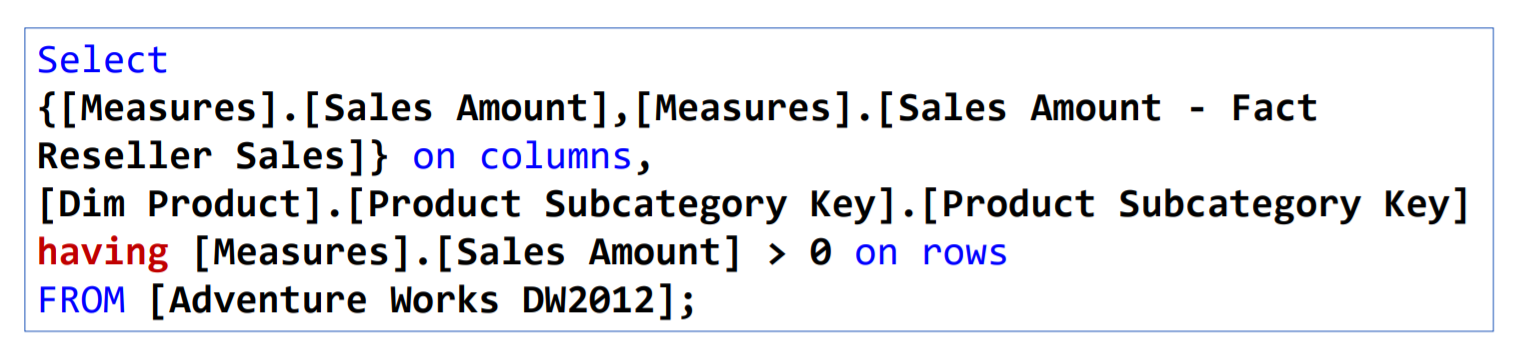


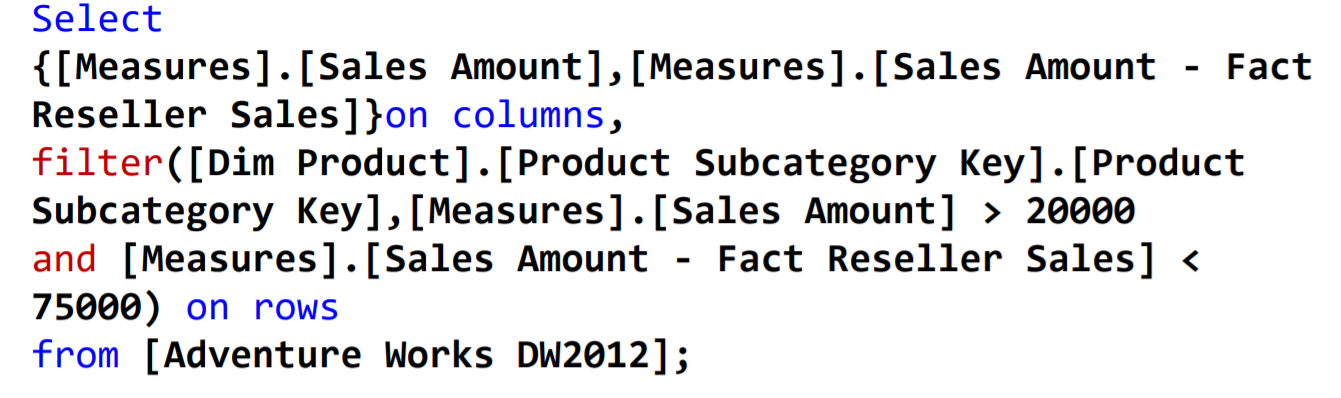


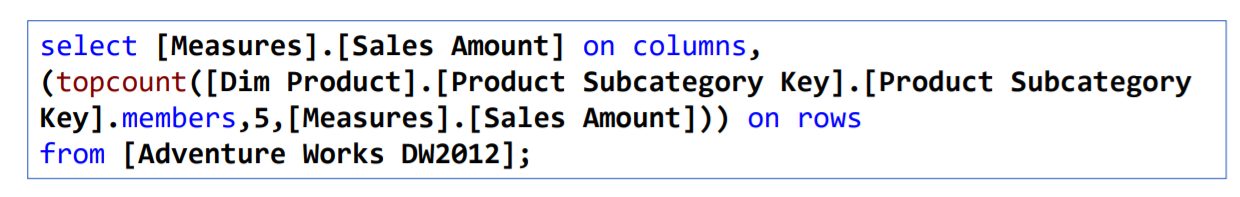


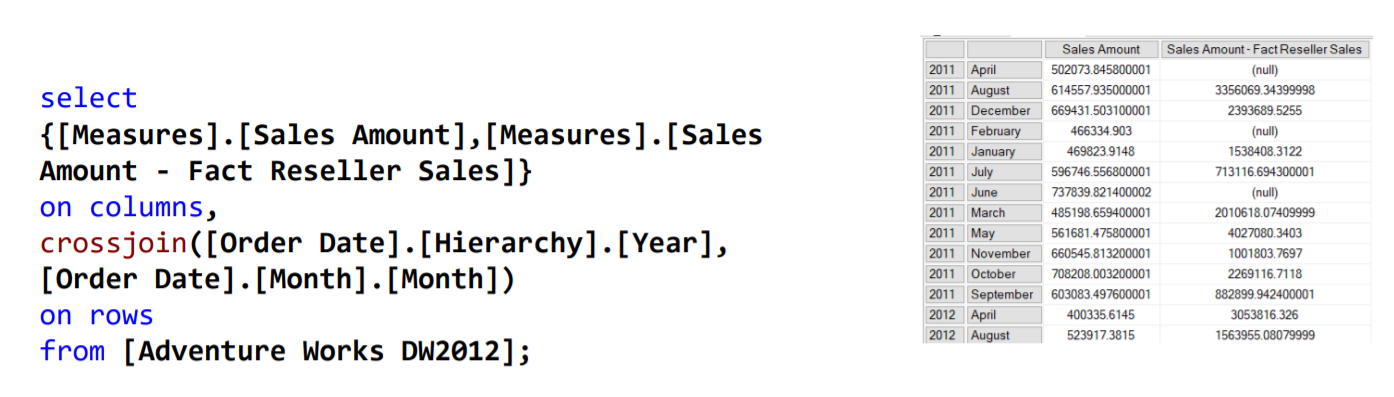




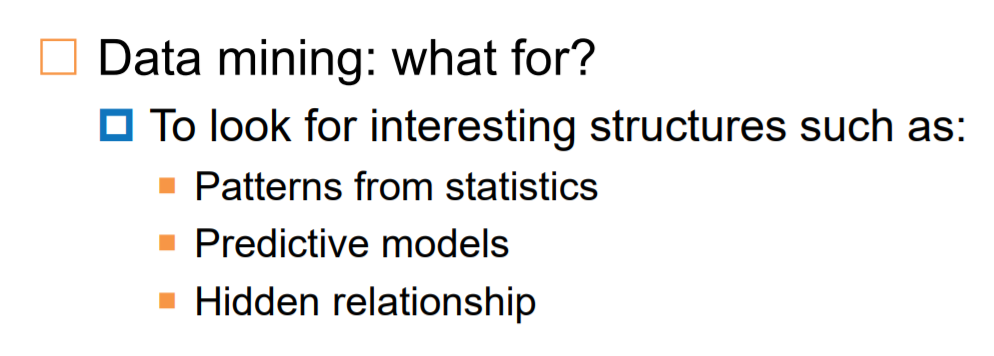


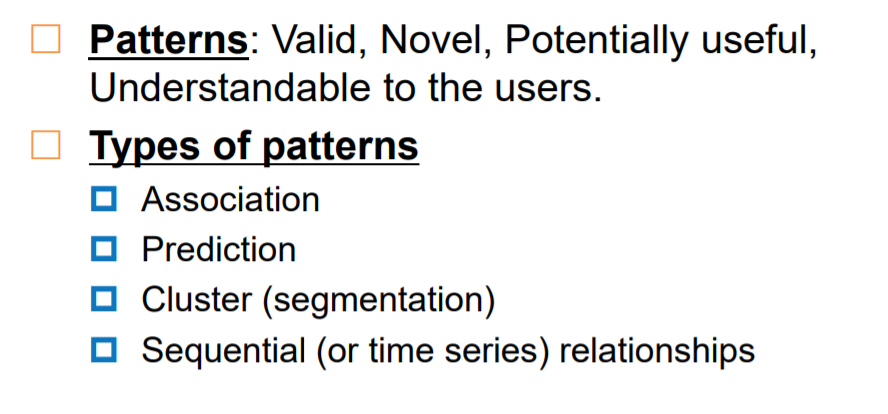


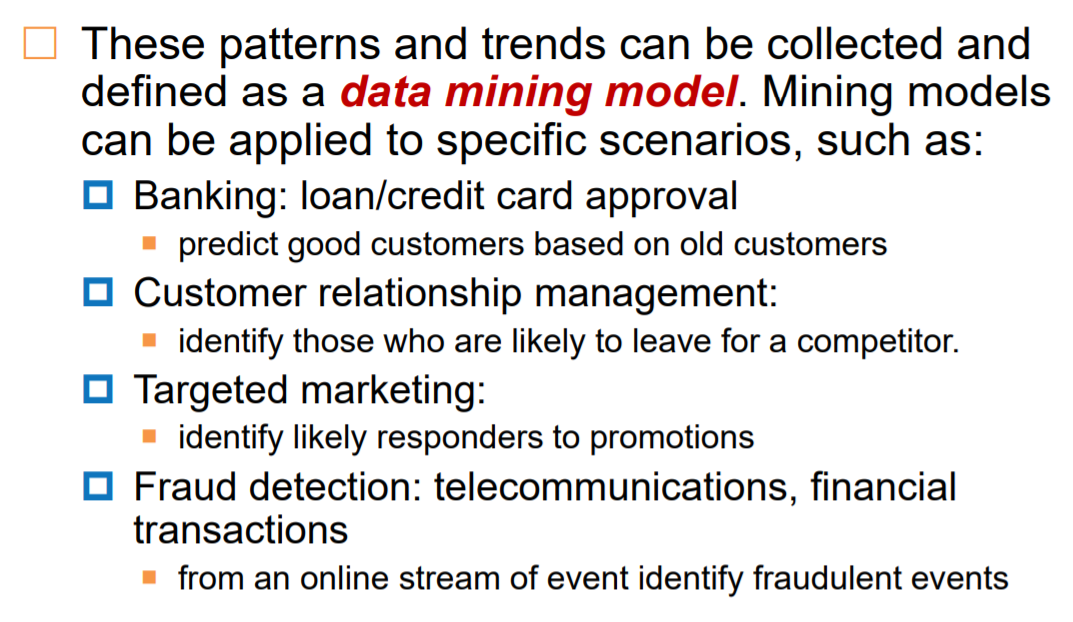




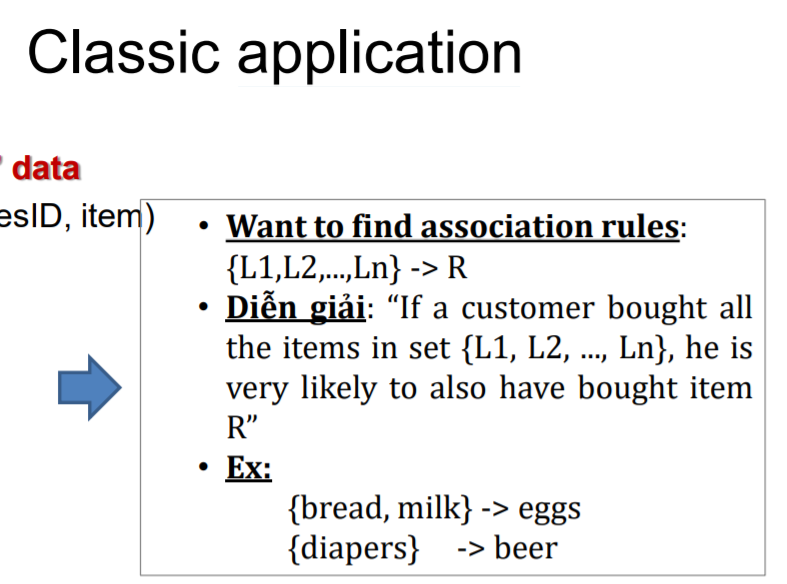
3. Mining

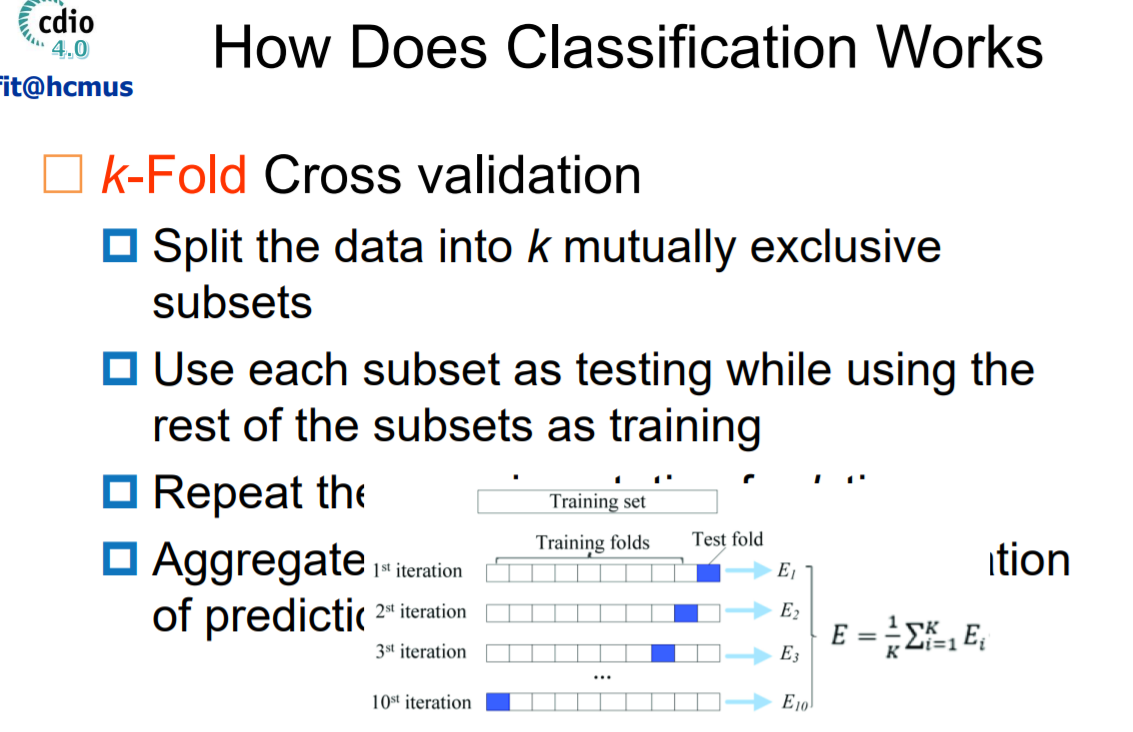


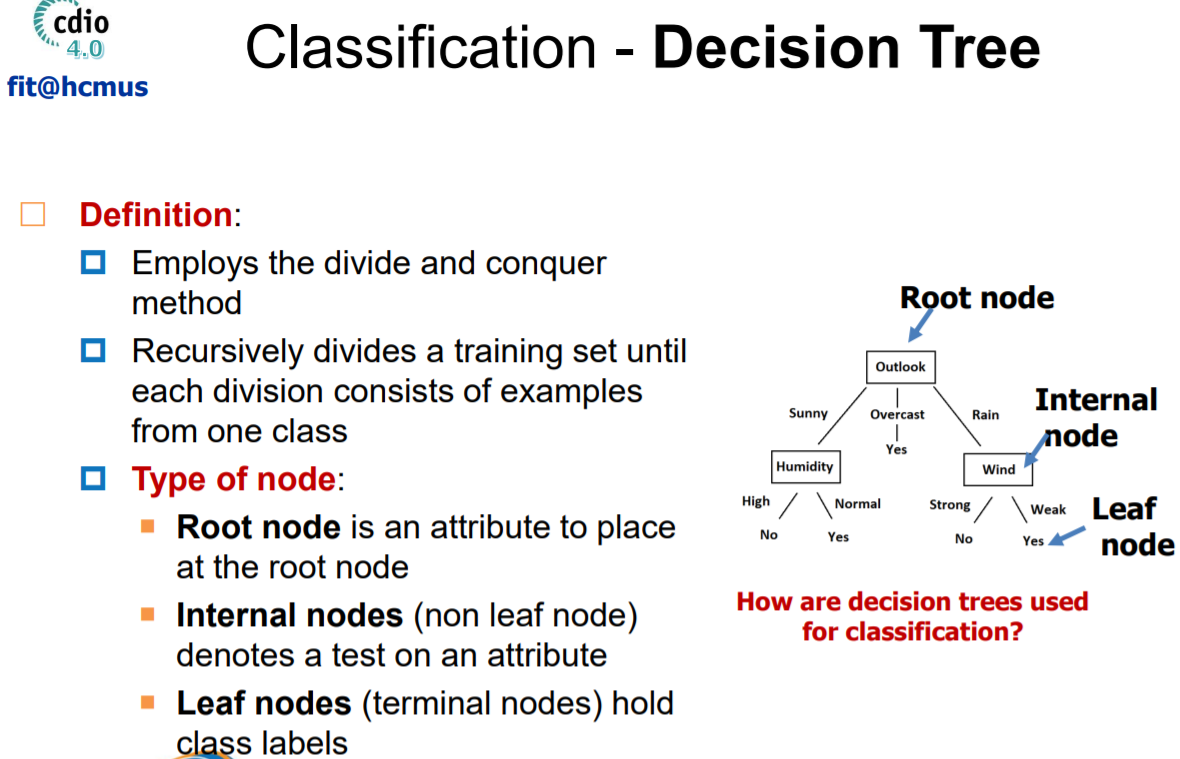


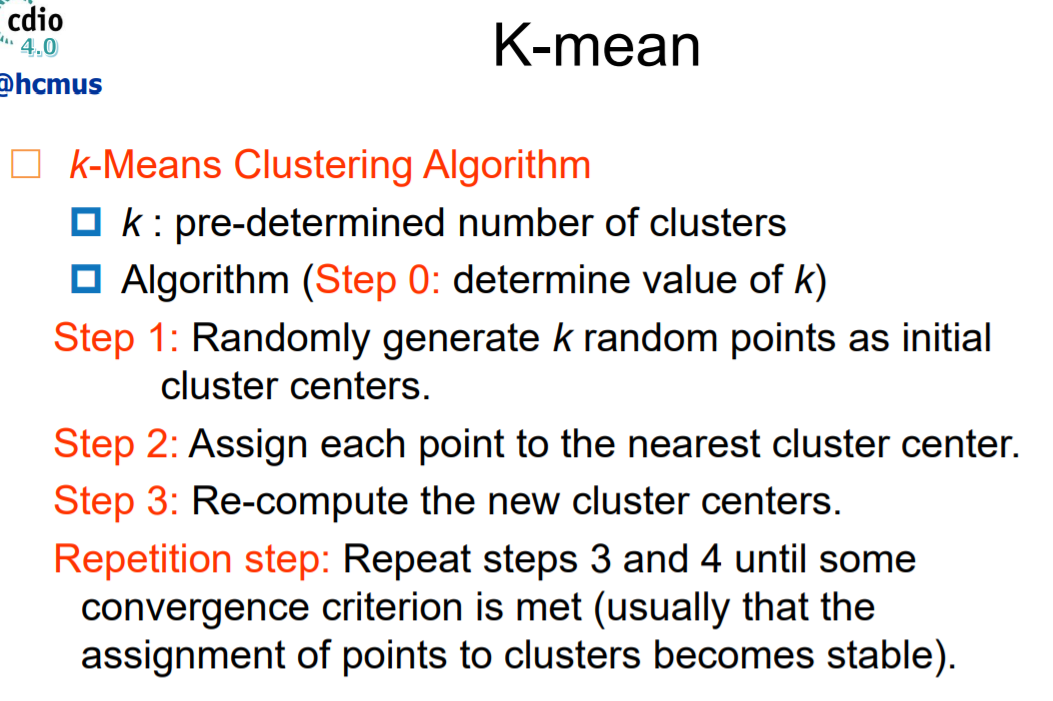


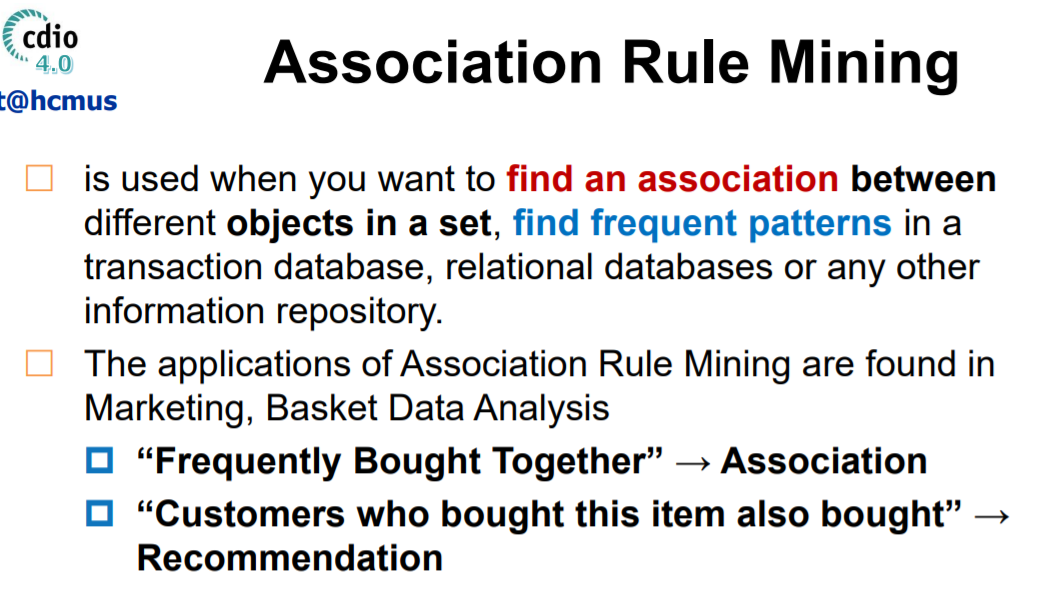












---------------------END---------------------