Cho schema ai4e\_test, các bảng:

* customer (là dữ liệu customer ngày 17-10-2023)
* customer\_cdc\_18 (là dữ liệu customer thay đổi ngày 18-10-2023)
* customer\_cdc\_19 (là dữ liệu customer thay đổi ngày 19-10-2023)

1. Tạo các bảng <user>\_customer\_scd\_type\_1, <user>\_customer\_scd\_type\_2, <user>\_customer\_scd\_type\_4 với dữ liệu copy hoàn toàn từ bảng customer

2. Thực hiện thao tác update dữ liệu từ bảng customer\_cdc\_18 vào các bảng <user>\_customer\_scd\_type<> với các kiểu 1, 2, 4. Riêng type 4, cần tạo thêm bảng hist với tên bảng <user>\_customer\_scd\_type\_4\_hist.

3. Thực hiện thao tác update dữ liệu từ bảng customer\_cdc\_19 vào các bảng <user>\_customer\_scd\_type<> với các kiểu 1, 2, 4. Riêng type 4, cần tạo thêm bảng hist với tên bảng <user>\_customer\_scd\_type\_4\_hist.

Thực hiện làm bài như sau:

* Tạo schema mới (theo tên học viên), và tạo các bảng kết quả trong schema này
* ví dụ:

A screenshot of a computer program

Description automatically generated

* Thực hiện thực hành, code Python/SQL

Guide SCD Homework

Thiết kế các bảng theo từng scd type và versioning dữ liệu dữ trên các technical field

1. SCD Type 1

Lưu latest version nên không cần các trường kĩ thuật cũng được. Tuy nhiên, có thể thêm cột record\_created\_time để thể hiện ngày mà bản ghi được tạo ra.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Primary key | Col 1 | Col 2 | Col 3 | record\_created\_time |
| A | BOND | 1000 | 2021-11-10 | 2023-08-01 |
| B | FUND | 200 | 2023-01-31 | 2023-08-01 |
| C | BANCA | 450 | 2022-12-12 | 2023-07-02 |

1. SCD Type 2

* SCD Type 2 lưu cả historical data (các version trước đó) => thêm một cột active thể hiện version nào là đang hoạt động. Khi đó khi sử dụng, chỉ cần filter active = 1 là có thể lấy ra các bản ghi mới nhất.
* Tuy nhiên, nhiều khi cần chạy lại dữ liệu trong quá khứ, điều kiện filter trên không thể hiện đc các bản ghi trong quá khứ => cần thêm các technical field sao cho dựa trên giá trị của các cột đó ta có thêm một filter để truy vấn đc yêu cầu trên.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Primary key | Col 1 | Col 2 | record\_created\_time | record\_updated\_time | record\_is\_active |
| A | BOND | 1000 | 2023-08-02 | 3000-12-12 | 1 |
| B | FUND | 500 | 2023-07-30 | 3000-12-12 | 1 |
| C | BANCA | 1200 | 2023-08-05 | 3000-12-12 | 1 |
| D | BOND | 120 | 2023-06-02 | 3000-12-12 | 1 |
| A | BOND | 800 | 2023-06-15 | 2023-08-02 | 0 |
| A | BOND | 1000 | 2023-05-01 | 2023-06-15 | 0 |
| B | FUND | 1000 | 2023-05-01 | 2023-07-30 | 0 |
| C | BANCA | 550 | 2023-05-01 | 2023-08-05 | 0 |
| E | BOND | 4000 | 2023-05-01 | 3000-12-12 | 1 |

Giả định ngày 2023-05-01 là ngày đầu tiên các bản ghi đc ETL

=> chỉ có 4 record A, B, C, D sau đó mới có thêm bản ghi E. Update time sử dụng một giá trị rất lớn trong tương lai (3000-12-12) hoặc một giá trị time tương tự

Giả sử có ngày 2023-06-01 (run\_data\_date), khi đó, để lấy các bản ghi active trong ngày này, chỉ cần filter:

record\_created\_time <= run\_data\_date < record\_update\_time.

Mô phỏng bảng qua các version. Ngày 2023-05-01:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Primary key | Col 1 | Col 2 | record\_created\_time | record\_updated\_time | record\_is\_active |
| A | BOND | 1000 | 2023-05-01 | 3000-12-12 | 1 |
| B | FUND | 1000 | 2023-05-01 | 3000-12-12 | 1 |
| C | BANCA | 550 | 2023-05-01 | 3000-12-12 | 1 |
| E | BOND | 4000 | 2023-05-01 | 3000-12-12 | 1 |

Tiếp theo, đến ngày 2023-06-02 có thêm 1 bản ghi đc insert:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Primary key | Col 1 | Col 2 | record\_created\_time | record\_updated\_time | record\_is\_active |
| D | BOND | 120 | 2023-06-02 | 3000-12-12 | 1 |
| A | BOND | 1000 | 2023-05-01 | 3000-12-12 | 1 |
| B | FUND | 1000 | 2023-05-01 | 3000-12-12 | 1 |
| C | BANCA | 550 | 2023-05-01 | 3000-12-12 | 1 |
| E | BOND | 4000 | 2023-05-01 | 3000-12-12 | 1 |

Ngày 2023-06-15, có thêm một bản ghi key A đc update:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Primary key | Col 1 | Col 2 | record\_created\_time | record\_updated\_time | record\_is\_active |
| D | BOND | 120 | 2023-06-02 | 3000-12-12 | 1 |
| A | BOND | 800 | 2023-06-15 | 3000-12-12 | 1 |
| A | BOND | 1000 | 2023-05-01 | 2023-06-15 | 0 |
| B | FUND | 1000 | 2023-05-01 | 3000-12-12 | 1 |
| C | BANCA | 550 | 2023-05-01 | 3000-12-12 | 1 |
| E | BOND | 4000 | 2023-05-01 | 3000-12-12 | 1 |

Tương tự khi ngày 2023-07-30 bản ghi B đc update, 2023-08-02 bản ghi A đc update, 2023-08-05 bản ghi C đc update.

* Mở rộng, để đánh dấu loại cdc (insert, update, delete -> ta có thể thêm một cột mới record\_change\_type)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Primary key | Col 1 | Col 2 | record\_created\_time | record\_updated\_time | record\_is\_active | record\_change\_type |
| A | BOND | 1000 | 2023-08-02 | 3000-12-12 | 1 | CDC\_UPDATE |
| B | FUND | 500 | 2023-07-30 | 3000-12-12 | 1 | CDC\_UPDATE |
| C | BANCA | 1200 | 2023-08-05 | 3000-12-12 | 1 | CDC\_UPDATE |
| D | BOND | 120 | 2023-06-02 | 3000-12-12 | 1 | CDC\_INSERT |
| A | BOND | 800 | 2023-06-15 | 2023-08-02 | 0 | CDC\_UPDATE |
| A | BOND | 1000 | 2023-05-01 | 2023-06-15 | 0 | CDC\_INSERT |
| B | FUND | 1000 | 2023-05-01 | 2023-07-30 | 0 | CDC\_INSERT |
| C | BANCA | 550 | 2023-05-01 | 2023-08-05 | 0 | CDC\_INSERT |
| E | BOND | 4000 | 2023-05-01 | 3000-12-12 | 1 | CDC\_INSERT |

1. SCD Type 4

* SCD Type 4 cần sử dụng 2 bảng, bảng main lưu latest version (tương tự SCD type 1) còn historical data thì sử dụng bảng hist => khi update dữ liệu cần update đồng thời vào bảng main và hist. Bảng main khá đơn giản, thiết kế tương tự scd type 1
* Bảng hist, cần thêm các technical field -> sử dụng tương tự như scd type 2, tuy nhiên không cần cột thể hiện giá trị active vì đã có bảng main.

Ex:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Primary key | Col 1 | Col 2 | record\_created\_time | record\_updated\_time | record\_change\_type |
| A | BOND | 1000 | 2023-08-02 | 3000-12-12 | CDC\_UPDATE |
| B | FUND | 500 | 2023-07-30 | 3000-12-12 | CDC\_UPDATE |
| C | BANCA | 1200 | 2023-08-05 | 3000-12-12 | CDC\_UPDATE |
| D | BOND | 120 | 2023-06-02 | 3000-12-12 | CDC\_INSERT |
| A | BOND | 800 | 2023-06-15 | 2023-08-02 | CDC\_UPDATE |
| A | BOND | 1000 | 2023-05-01 | 2023-06-15 | CDC\_INSERT |
| B | FUND | 1000 | 2023-05-01 | 2023-07-30 | CDC\_INSERT |
| C | BANCA | 550 | 2023-05-01 | 2023-08-05 | CDC\_INSERT |
| E | BOND | 4000 | 2023-05-01 | 3000-12-12 | CDC\_INSERT |

=> query bảng hist tương tự bảng cdc type 2. Latest version => query bảng main