**Problem Statement :**

We have employees related tables in MySQL.

Schema\_name : labuser\_database

Tables :

* + Employees\_table (employee details like id, name, email,phone,salary,department etc.)
  + Department\_table (department details like department id, name, location etc.)
  + Locations\_table (location info, street name, city, pin, state country)
  + Jobs\_table (job metadata, job id, job name etc)
  + Regions\_table (region id, region name)
  + Countries\_table (country id, country name)

We need to perform below tasks:

* Import all these tables to hive tables. Hive DDLs can be inferred from table schemas in MySQL.
* There is a delta file present at local edge node path which has changes in employee information as on current date.
* Create a new table (employee\_scd) which is a slowly changing dimension (type 2) implementation using employees\_table as base table and delta file as a delta table.
* Scd table should have additional columns, start\_date, end\_date and active\_flag.
* Start\_date and end\_date behave as per concept of SCD type 2. And active flag can be 0 or 1 depending on the which row in table contains latest information of the employee.
* Based on changes present in the delta file (on columns email, phone\_number, job\_id, department\_id, manager\_id, salary, commission\_pct) maintain the change history using start\_date, end\_date and active flag columns with old information being closed with end\_date = (current\_Date -1)
* Create a final table by joining all these tables (join conditions as per ERD) to get latest information of an employee.
* Send the latest information from hive to a mysql table.

**Scd type 2 example :**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| id | name | dept\_id | manager\_id | start\_date | end\_date | active\_flag |
| 1 | abc | 10 | 111 | 2019-01-01 | 9999-12-31 | 1 |
| 2 | def | 10 | 111 | 2019-01-01 | 2019-12-08 | 0 |
| 2 | def | 20 | 111 | 2019-12-09 | 9999-12-31 | 1 |

**Employee 1 was present in employee\_table but there was no change available in delta file hence it has only record which is the latest info and active flag is 1.**

**For employee 2, department got changed from 10 to 20 as per information in delta file. and if we load SCD using base employee and delta file then it will have two records with maintaining the history.**

**Start date of the latest record will be current date and end date of previous record will be current date -1**

**Employee\_scd table structure:**

drop table if exists edureka\_dw.employees\_scd;

CREATE TABLE edureka\_dw.employees\_scd

( employee\_id int

, first\_name VARCHAR(20)

, last\_name VARCHAR(25)

, email VARCHAR(25)

, phone\_number VARCHAR(20)

, hire\_date varchar(12)

, job\_id VARCHAR(10)

, salary decimal

, commission\_pct decimal(2,2)

, manager\_id int

, department\_id int

, start\_date string

, end\_date string

, active\_flag char(1)

, row\_insertion\_dttm string

) ;

**final table with latest employee details structure:**

CREATE table edureka\_dw.employee\_details\_latest

(employee\_id int,

job\_id varchar(10),

manager\_id int,

department\_id int,

location\_id int,

country\_id char(2),

first\_name varchar(20),

last\_name varchar(25),

salary decimal,

commission\_pct decimal(2,2),

department\_name varchar(30),

job\_title varchar(35),

city varchar(30),

state\_province varchar(25),

country\_name varchar(40),

region\_name varchar(25)

);