Approach for SCD Type-I and TYPE-II using HIVE

In this approach we need to create a partition on current\_indicator column, which holds the value either Y or N.

If required we might add an additional layer of partition based on client name or source\_identifier or…etc.this layer can be top level partition.

To achieve SCD Type-I and Type-II on one single table.

**Master\_Table \_Schema:**

CREATE TABLE EMP\_MASTER (id INT,

name string,

phone\_number string,

state string,

designation string,

eff\_strt\_dt string,

eff\_end\_dt string)

partitioned by ( current\_indicator char(1) )

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Name | Phone\_Number | STATE | Designation | eff\_strt\_dt | Eff\_end\_dt | Current\_Indicator |
| 1001 | XYZ | 1112223333 | NY | Analyst | 2018-01-01 | 9999-12-31 | Y |
| 1002 | ABC | 1234567891 | CA | Senior Analyst | 2018-01-01 | 9999-12-31 | Y |
| 1003 | DEF | 7418529635 | IL | Analyst | 2018-01-01 | 9999-12-31 | Y |

**SCD Type-I Columns:**

Phone Number

**SCD Type-II Columns:**

State,Designation

**Daily Load table (Delta) Schema:**

CREATE TABLE EMP\_DELTA ( id INT,

name string,

phone\_number string,

state string,

designation string)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Name | Phone\_Number | STATE | Designation |
| 1001 | XYZ | 3692581475 | AZ | Analyst |
| 1003 | ABC | 8523697416 | CA | Senior Analyst |
| 1004 | BSE | 5648972315 | OR | Junior Analyst |

**Required Result:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Name | Phone\_Number | STATE | Designation | Eff-strt-dt | Eff-end-dt | Current\_Indicator |
| 1001 | XYZ | 1112223333 | NY | Analyst | 2018-01-01 | 2019-07-15 | N |
| 1001 | XYZ | 3692581475 | AZ | Analyst | 2019-07-16 | 9999-12-31 | Y |
| 1002 | ABC | 1234567891 | CA | Senior Analyst | 2018-01-01 | 9999-12-31 | Y |
| 1003 | ABC | 8523697416 | CA | Senior Analyst | 2018-01-01 | 9999-12-31 | Y |
| 1004 | BSE | 5648972315 | OR | Junior Analyst | 2019-07-16 | 9999-12-31 | Y |

Steps to achieve CDC using TYPE-I and TYPE-II SCD in HIVE.

**Step -1:**

Create a temporary table with same schema as our main table. And It should be dropped and recreated daily.

DROP TABLE IF EXISTS EMP\_TEMP;

CREATE TABLE EMP\_TEMP IF NOT EXISTS (id int,

name string,

phone\_number string,

state string ,

designation string,

current\_indicator char(1))

**Step -2:**

In this step we need to set aside the rows for which there is a change in SCD TYPE-II columns.

When there is change in SCD TYPE-II colu,ns e need to make the current row from master row as Not current.

Query:

INSERT INTO EMP\_MASTER ( select mstr.id,

mstr.name,

mstr.phone\_number,

mstr .state,

mstr .designation,

mstr .eff\_strt\_dt,

from\_unixtime(unix\_timestamp(current\_date()),"YYYY-MM-dd ") as eff\_end\_dt, ‘N’ as current\_indicator )

from EMP\_MASTER mstr,

join EMP\_DELTA delta

on (mstr.id = delta.id)

where concat(NVL(mstr. State,”DUMMY”),NVL(mstr.designation,”DUMMY”)) !=

concat(NVL(delta. State,”DUMMY”),NVL(delta.designation,”DUMMY”));

This will insert rows for which we have change in SCD TYPE-II columns into master table with current indicator as ‘N’ and eff\_end\_dt as the day on which file has come.

**Step - 3:**

In this step we will update all the other records which have ONLY TYPE-I col changes. If a row has both TYPE-I and TYPE-II column change then that row will not be covered here.

INSERT OVERWRITE EMP\_TEMP

SELECT delta.id,

delta.name,

delta.phone\_number,

delta .state,

delta .designation,

mstr.eff\_strt\_dt,

mstr.eff\_end\_dt,

“Y” as current\_indicator

FROM EMP\_MASTER mstr,

join EMP\_DELTA delta

on (mstr.id = delta.id)

where mstr.id NOT IN ( SELECT mstr\_inactive.id from EMP\_MASTER mstr\_inactive where mstr\_inactive.current\_indicator = "N" and mstr\_inactive.eff\_end\_dt = from\_unixtime(unix\_timestamp(current\_date()),"YYYY-MM-dd "))

**Step - 4:**

In this step we will create new row for all the keys which are set to N in step-2

INSERT INTO EMP\_TEMP

SELECT delta.id,

delta.name,

delta.phone\_number,

delta .state,

delta .designation,

from\_unixtime(unix\_timestamp(current\_date()),"YYYY-MM-dd 00:00:00") as eff\_strt\_dt,

"9999-12-31 00:00:00" eff\_end\_dt,

‘’Y” as current\_indicator

FROM EMP\_MASTER mstr,

join EMP\_DELTA delta

on (mstr.id = delta.id)

where mstr.id in IN ( SELECT mstr\_inactive.id from EMP\_MASTER mstr\_inactive where mstr\_inactive.current\_indicator = "N" and mstr\_inactive.eff\_end\_dt = from\_unixtime(unix\_timestamp(current\_date()),"YYYY-MM-dd "))

**Step -5:**

In this step we will insert new entries,which do not exist in Master table, and are present in delta file only.

INSERT INTO EMP\_TEMP

SELECT delta.id,

delta.name,

delta.phone\_number,

delta .state,

delta .designation,

from\_unixtime(unix\_timestamp(current\_date()),"YYYY-MM-dd 00:00:00") as eff\_strt\_dt,

"9999-12-31 00:00:00" eff\_end\_dt,

“Y” as current\_indicator

FROM EMP\_MASTER mstr,

RIGHT OUTER JOIN EMP\_DELTA delta

on (mstr.id = delta.id)

where mstr.id = null;

**Step -6:**

In this step we will insert new entries, which do not exist in delta file and are existing entries in Master table.

INSERT INTO EMP\_TEMP

SELECT delta.id,

delta.name,

delta.phone\_number,

delta .state,

delta .designation,

from\_unixtime(unix\_timestamp(current\_date()),"YYYY-MM-dd 00:00:00") as eff\_strt\_dt,

"9999-12-31 00:00:00" eff\_end\_dt,

“Y” as current\_indicator

FROM EMP\_MASTER mstr,

LEFT OUTER JOIN EMP\_DELTA delta

on (mstr.id = delta.id)

where delta.id = null;

**Step -7:**

Overwrite EMP\_MASTER current\_indicator = ‘Y’ partition from temp table created.

INSERT OVERWRITE EMP\_MASTER PARTITON ( current\_indicator = ‘Y’

)

SELECT id,name,phone\_number,state,designation,eff\_strt\_dt,eff\_end\_dt from EMP\_TEMP ;