Deduplication using Hive ACID and Merge

Setup

```
CREATE TABLE emp (
empno    BIGINT,
ename    STRING,
job      STRING,
mgr      BIGINT,
updated_date date,
sal      INT,
deptno   INT
);
INSERT INTO emp VALUES
(7902,'FORD','ANALYST',7566,'1981-12-03',3000,20),
(7934,'MILLER','CLERK',7782,'1982-10-23',1300,10),
(7934,'MILLER','CLERK',7782,'1983-07-14',1350,20),
(7934,'MILLER','CLERK',9782,'1984-01-13',1400,10);

--Housekeeping
SELECT * FROM emp;
TRUNCATE table emp;
```

Deduplication process

Before

```
SELECT * FROM emp WHERE empno=7934;
```

```
7902    FORD    ANALYST 7566    1981-12-03  3000    20
7934    MILLER  CLERK   7782    1982-10-23  1300    10
7934    MILLER  CLERK   7782    1983-07-14  1350    20
7934    MILLER  CLERK   9782    1984-01-13  1400    10
```

Run ETL

```
MERGE INTO emp USING (
SELECT
EMPNO,
ENAME,
SAL,
UPDATED_DATE,
ROW_NUMBER() OVER ( PARTITION BY EMPNO ORDER BY UPDATED_DATE DESC ) as rnk
FROM emp
) old_data
ON emp.EMPNO=old_data.EMPNO
AND emp.UPDATED_DATE=old_data.UPDATED_DATE
WHEN MATCHED AND old_data.rnk > 1 THEN DELETE;
```

After

```
7902    FORD    ANALYST 7566    1981-12-03  3000    20
7934    MILLER  CLERK   9782    1984-01-13  1400    10
```