Table design

VIDEOSTART_RAW

COLUMN_NAME	DATA_TYPE	PK	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
DATETIME	VARCHAR2(30 BYTE)	N	Yes	null	1	Data from raw file
VIDEOTITLE	VARCHAR2(200 BYTE)	N	Yes	null	2	Data from raw file
EVENTS	VARCHAR2(150 BYTE)	N	Yes	null	3	Data from raw file

VIDEOSTART_DLT

COLUMN_NAME	DATA_TYPE	PK	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
DATETIME	TIMESTAMP (6)	N	No	null	1	Data reformatted from VIDEOSTART_RAW. DATETIME
PLATFORM	VARCHAR2(200 BYTE)	N	No	null	2	Data derived from VIDEOSTART_RAW. VIDEOTITLE
SITE	VARCHAR2(200 BYTE)	N	No	null	3	Data derived from VIDEOSTART_RAW. VIDEOTITLE
VIDEO	VARCHAR2(200 BYTE)	N	No	null	4	Data derived from VIDEOSTART_RAW. VIDEOTITLE

FACTVIDEOSTART

COLUMN_NAME	DATA_TYPE	PK	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
DATETIME_SKEY	VARCHAR2(12 BYTE)	N	No	null	1	Data derived from DIMDATE. DATETIME_SKEY
PLATFORM_SKEY	NUMBER(38,0)	N	No	null	2	Data derived from DIMPLATFORM. PLATFORM_SKEY
SITE_SKEY	NUMBER(38,0)	N	No	null	3	Data derived from DIMSITE. SITE_SKEY
VIDEO_SKEY	NUMBER(38,0)	N	No	null	4	Data derived from DIMVIDEO. VIDEO_SKEY
DB_INSERT_TIMESTAMP	TIMESTAMP (6)	N	No	null	5	TIMESTAMP when inserting the data

DIMDATE_DLT

COLUMN_NAME	DATA_TYPE	PK	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
DATETIME	VARCHAR2(12 BYTE)	N	No	null	1	Data reformatted from VIDEOSTART_DLT. DATETIME

DIMPLATFORM_DLT

COLUMN_NAME	DATA_TYPE	PK	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
PLATFORM	VARCHAR2(200 BYTE)	N	No	null	1	Data derived from VIDEOSTART_DLT. PLATFORM

DIMSITE_DLT

COLUMN_NAME	DATA_TYPE	PK	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
SITE	VARCHAR2(200 BYTE)	N	No	null	1	Data derived from VIDEOSTART_DLT. SITE

DIMVIDEO_DLT

COLUMN_NAME	DATA_TYPE	PK	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
VIDEO	VARCHAR2(200 BYTE)	N	No	null	1	Data derived from VIDEOSTART_DLT. VIDEO

DIMDATE

COLUMN_NAME	DATA_TYPE	PK	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
DATETIME_SKEY	NUMBER(38,0)	Y	No		1	Data derived from DIMDATE_DTL. DATETIME

DIMPLATFORM

COLUMN_NAME	DATA_TYPE	PK	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
PLATFORM_SKEY	NUMBER(38,0)	Y	No		1	
PLATFORM	VARCHAR2(200 BYTE)	N	No	null	2	Data derived from DIMPLATFORM_DLT. PLATFORM

DIMSITE

COLUMN_NAME	DATA_TYPE	PK	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
SITE_SKEY	NUMBER(38,0)	Y	No		1	
SITE	VARCHAR2(200 BYTE)	N	No	null	2	Data derived from DIMSITE_DLT. SITE

DIMVIDEO

COLUMN_NAME	DATA_TYPE	PK	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
VIDEO_SKEY	NUMBER(38,0)	Y	No		1	
VIDEO	VARCHAR2(200 BYTE)	N	No	null	2	Data derived from DIMVIDEO_DLT. VIDEO

Process design

1. Load raw videostarts file into VIDEOSTART_RAW

a. Use sqlldr to load raw data into table

dos2unix video_data.csv

sqlldr ${DB_USER}/${DB_PWD}@${DB_NAME} control=video_data.ctl direct=true errors=-1

Log file is ***video_data.log***

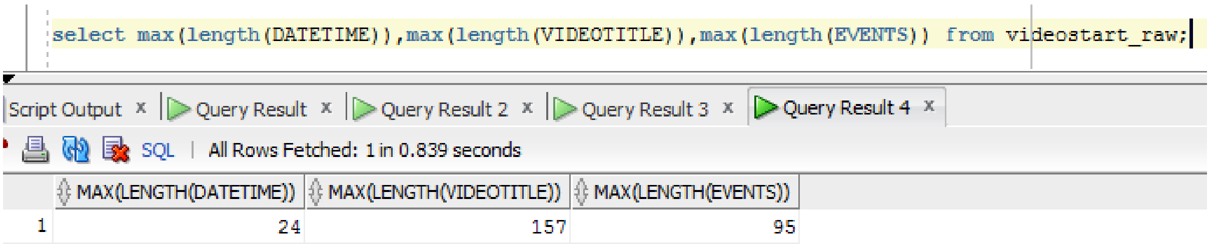
Control file is ***video_data.ctl***

Bad records are in ***video_data.csv.bad***

Inform the source data holder to see if they can revise the data in bad file. However, this is optional depending on the specific project.

b. Data auditing:

select max(length(DATETIME)),max(length(VIDEOTITLE)),max(length(EVENTS)) from videostart_raw;



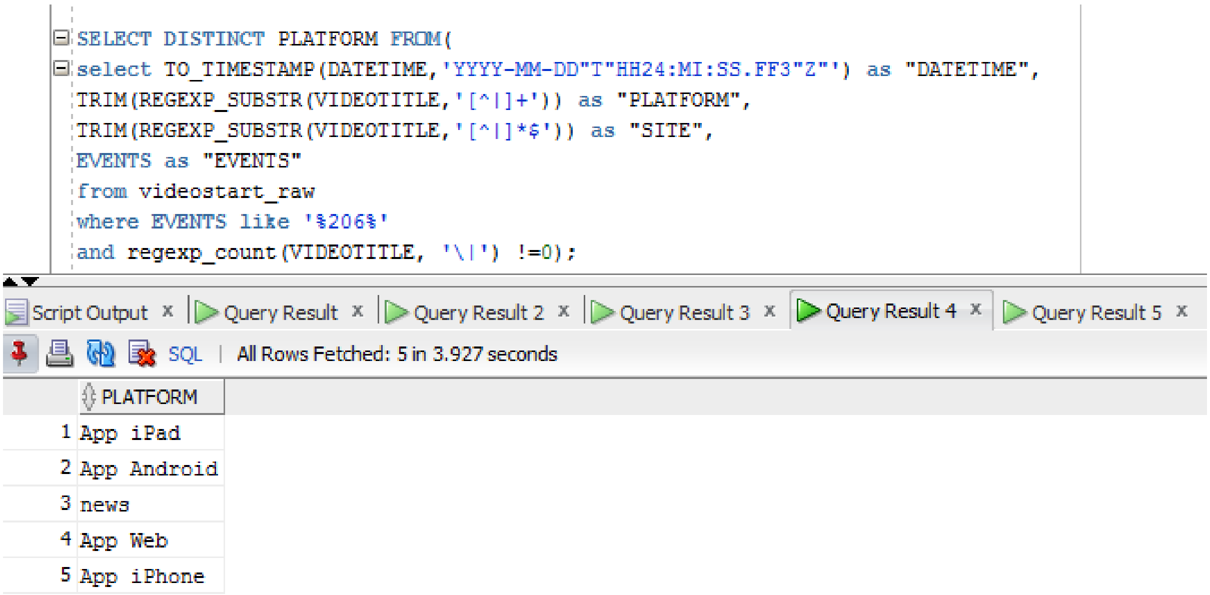
```
select max(length(DATETIME)),max(length(VIDEOTITLE)),max(length(EVENTS)) from videostart_raw;
```

All Rows Fetched: 1 in 0.839 seconds

	MAX(LENGTH(DATETIME))	MAX(LENGTH(VIDEOTITLE))	MAX(LENGTH(EVENTS))
1	24	157	95

Use the result to adjust the length of column in table

c. Identify the type of PLATFORM and SITE



```
SELECT DISTINCT PLATFORM FROM(
select TO_TIMESTAMP(DATETIME,'YYYY-MM-DD"T"HH24:MI:SS.FF3"Z"') as "DATETIME",
TRIM(REGEXP_SUBSTR(VIDEOTITLE,'[^|]+')) as "PLATFORM",
TRIM(REGEXP_SUBSTR(VIDEOTITLE,'[^|]*$')) as "SITE",
EVENTS as "EVENTS"
from videostart_raw
where EVENTS like '%206%'
and regexp_count(VIDEOTITLE, '\|') !=0);
```

All Rows Fetched: 5 in 3.927 seconds

	PLATFORM
1	App iPad
2	App Android
3	news
4	App Web
5	App iPhone

d. The sql script to create the table

1_create_tables.sql

2. Clean data in Intermediate tables

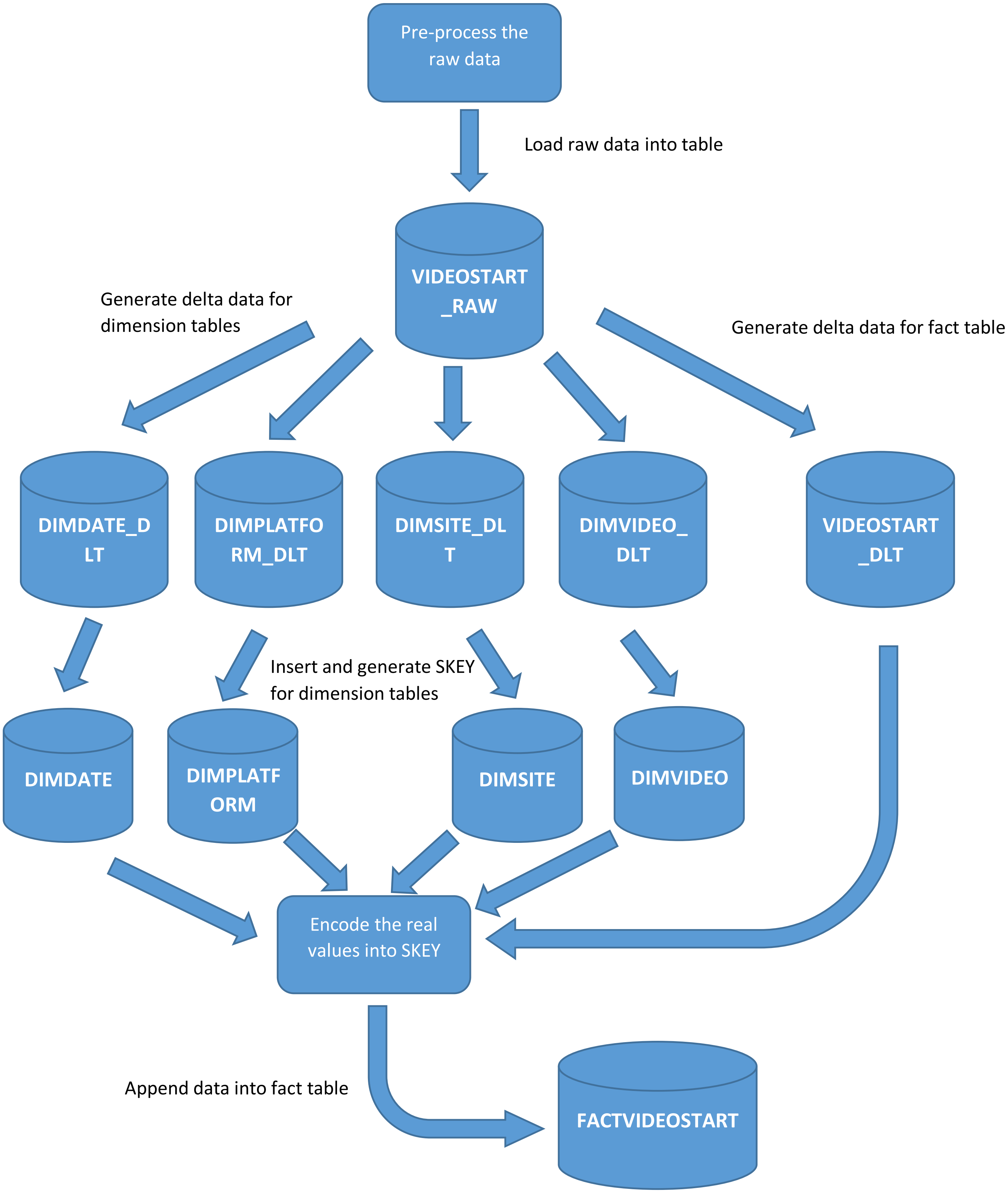
2_clean_delta_table.sql

3. **Wash data in VIDEOSTART_RAW and load into VIDEOSTART_DLT**
   3_wash_data.sql
4. **Populate DIMDATE_DLT, DIMPLATFORM_DLT, DIMSITE_DLT and DIMVIDEO_DLT**
   4_populate_dim_dlt.sql

5. **Insert delta data into staging tables - DIMDATE, DIMPLATFORM, DIMSITE and DIMVIDEO**
   5_insert_dim.sql

6. **Use VIDEOSTART_DLT, DIMDATE, DIMPLATFORM, DIMSITE and DIMVIDEO to generate output data and append the data into fact table – VIDEOSTART**
   6_append_fact.sql

On-going process workflow



NOTE:

1. **SKEY stands for surrogate key.**
2. **The current design is Dimension Type One.**
3. **If the source dimension data contains not only the PK but also some attributes, and we want to track the changes of attributes, we should use Dimension Type Two.**

One sample of Dimension Type Two
Data from 06/04/2017:

Product_ID	Product_Name	Price	Location
P001	Iphone6	750	Townhall Shop
P003	Iphone7	1000	Townhall Shop

Data in dimension table:

Product_SKEY	Product_ID	Product_Name	Price	Location	Current_Flag	Start_Date	End_Date
111	P001	Iphone6	800	Townhall Shop	Y	31/12/2016	31/12/9999
112	P002	Iphone6Plus	900	Townhall Shop	Y	20/01/2017	31/12/9999

Add new product (P003) and update product (P001) in dimension table:

Product_SKEY	Product_ID	Product_Name	Price	Location	Current_Flag	Start_Date	End_Date
111	P001	Iphone6	800	Townhall Shop	N	31/12/2016	05/04/2017
112	P002	Iphone6Plus	900	Townhall Shop	Y	20/01/2017	31/12/9999
113	P003	Iphone7	1000	Townhall Shop	Y	06/04/2017	31/12/9999
114	P001	Iphone6	750	Townhall Shop	Y	06/04/2017	31/12/9999

Yellow part is update, and red part is insertion.
When there is a new record coming in, we generate a new record with new SKEY, Current_Flag = 'Y', Start_Date = Current_Date, End_Date = 31/12/9999
When there is a updated record coming in, we also generate a new record with new SKEY Current_Flag = 'Y', Start_Date = Current_Date, End_Date = 31/12/9999; and at same time we need to update the old record in dimension table with Current_Flag = 'N', End_Date = Current_Date – 1

Therefore, when we populate new records into fact table, we need to put a filter such as Current_Flag = 'Y' in order to get the correct SKEY; if we want to track the history data in dimension table for certain days or certain period, we need to put a time range filter such as EVENT_DATE(or CONTACT_DATE) between Start_Date and End_Date

For example, if in fact table we see a transaction like customer purchased product(P001) on 01/04/2017, by looking at product dimension table, we could find the price that customer paid at that moment was 800 not 750, although 750 is the current price of P001