【最佳实践】实时计算Flink在广告行业的实时数仓建设实践

行业背景

- 行业现状：
  - 广告仍然是互联网公司的主要变现手段，2019年，中国广告市场总体规模达到8674.28亿元，较2018年增长了8.54%，据统计全球互联网市值前十的公司广告收入占比高达40%，可见其重要性。AI、大数据、智能投放等创新技术的普及应用，不仅创生了一批独角兽营销平台，而且大幅拉低了广告投放门槛，拓宽了广告市场空间。
- 大数据在其行业中的作用：
  - 大数据技术的应用在改变我们生活及工作的同时，为我们寻找数据背后的客观规律提供了一种有效途径。对潜在消费群体进行深入分析，并进行定制营销基础上的现代广告营销，对数据的规模及精准度有着极高的要求，而大数据的出现无疑为其落地提供了强有力的支撑。

业务场景

类似媒体，新闻类等APP，上面有各种广告位提供给广告主。广告主投放广告，用户点击广告将实时的产生操作日志数据，对这些日志数据进行实时分析，通过每个广告位上不同广告的投放地区、广告ID、设备唯一编码等信息，可以统计点击次数、投放次数等指标，可用于制定更高效的广告投放策略，降低投放成本，提高广告收益。

技术架构

架构解析：
数据采集：该场景中，APP、Web、Server等服务上会产生大量的广告投放、用户广告点击等操作日志数据，这些日志数据被实时采集至日志服务系统（SLS），作为Flink的数据源。
实时数仓架构：该场景中，整个实时数仓构建，全部通过 Flink完成。Flink读取SLS中的原始日志数据，经过数据清洗、数据处理等操作写出到DataHub，Flink进一步读取DataHub的数据进行实时统计分析，最终输出对应的指标结果到RDS，供业务系统使用。

业务指标

- 实时数据中间层，对原始日志进行实时数据清洗
  - 获取投放主题及维度打宽
  - 获取点击主题及维度打宽
- 统计投放指标
  - 某个广告在某个省的当天投放量
  - 某个广告在某个市的当天投放量
  - 某个广告在某个投放终端的当天投放量
- 统计点击指标
  - 某个广告在某个省的当天点击量
  - 某个广告在某个市的当天点击量
  - 某个广告在某个投放终端的当天点击量
- 热门广告排行榜

业务代码

场景一：对原始日志进行实时数据清洗

投放主题

> 根据业务主题分成投放主题和点击主题，当release_status=1时为投放主题。

输入表

```
create table ods_release(
  `sid` varchar,           --投放请求ID
  exts varchar,                      --扩展信息
  device_type varchar,     --1 android| 2 ios | 9 其他
  release_status varchar,  --投放状态 1 or 2
  device_num varchar,      --设备唯一编码
  release_session varchar, --投放会话ID
  `date` date              --创建时间
) with (
  type ='sls',
...
);
```

输出表

```
create table dw_release_exposure(
  release_session varchar, -- comment '投放会话id'
```

```
  release_status varchar,  -- comment '投放状态'
  device_num varchar,      -- comment '设备唯一编码'
  device_type varchar,     -- comment '1 android| 2 ios | 9 其他'
  area_code varchar,       -- comment '地区'
  aid varchar,             -- comment '广告id'
  ct date                  -- comment '创建时间'
)with(
type='datahub',
...
);
```

业务代码

```
insert into dw_release_exposure
select
  release_session,
  release_status,
  device_num,
  device_type,
  json_value(exts,'$.area_code'),
  json_value(exts,'$.aid'),
  `date` as ct
from
ods_release
where release_status='1'
;
```

投放主题关联维度表

> 投放主题与地区维度表、设备维度表进行聚合，得出宽表

输入表

```
create table dw_release_exposure(
  release_session varchar, -- comment '投放会话id'
  release_status varchar,  -- comment '投放状态'
  device_num varchar,      -- comment '设备唯一编码'
  device_type varchar,     -- comment '1 android| 2 ios | 9 其他'
  area_code varchar,       -- comment '地区'
  aid varchar,             -- comment '广告id'
  ct date                  -- comment '创建时间'
)with(
type='datahub',
...
);

--dim维度表
-- (地区，省市，唯一地区编码,编码和city_id是一一对应的)
create table dim_province(
  area_code varchar,
```

```
  province_id bigint,
  province_name varchar,
  city_id bigint,
  city_name varchar,
  region_id bigint,
  region_name varchar,
 PRIMARY KEY (area_code),
 PERIOD FOR SYSTEM_TIME--定义维表的变化周期。
 )with(
    type= 'rds',
...
);

--(用户设备维度表)
create table dim_device(
  device_type varchar comment '1 android| 2 ios | 9 其他',
  device_name varchar comment '设备名字',
 PRIMARY KEY (device_type),
 PERIOD FOR SYSTEM_TIME--定义维表的变化周期。
)with(
type= 'rds',
...
);
```

输出表

```
create table dm_release_exposure(
  aid varchar,
  aid_count bigint,
  device_name varchar,
  area_code varchar,
  province_id bigint,
  province_name varchar,
  city_id bigint,
  city_name varchar,
  ct date
)with(
type='datahub',
...
);
```

业务代码

```
insert into dm_release_exposure
select
  a.aid,
  count(a.aid) aid_count,
  c.device_name,
  a.area_code,
  b.province_id,
```

```
  b.province_name,
  b.city_id,
  b.city_name,
  a.ct
from
dw_release_exposure a
join
dim_province  FOR SYSTEM_TIME AS OF PROCTIME() as b on a.area_code=b.area_code
join
dim_device  FOR SYSTEM_TIME AS OF PROCTIME() as c on a.device_type=c.device_type
group by
a.aid,
a.area_code,
a.ct
;
```

点击主题

> 根据业务主题分成投放主题和点击主题，当release_status=2时为点击主题。

输入表

```
create table ods_release(
  `sid` varchar,           --投放请求ID
  exts varchar,                        --扩展信息
  device_type varchar,     --1 android| 2 ios | 9 其他
  release_status varchar,  --投放状态 1 or 2
  device_num varchar,      --设备唯一编码
  release_session varchar, --投放会话ID
  `date` date              --创建时间
  ) with (
  type ='sls',
...
);
```

输出表

```
create table dw_release_click(
  release_session varchar,  -- comment '投放会话id'
  release_status varchar,   -- comment '投放状态'
  device_num varchar,       -- comment '设备唯一编码'
  device_type varchar,      -- comment '1 android| 2 ios | 9 其他'
  `user_id` varchar,          -- comment '用户id'
  area_code varchar,        -- comment '地区'
  aid varchar,              -- comment '广告id'
  ct date                   -- comment '创建时间'
)with(
type='datahub',
...
);
```

业务代码

```
insert into dw_release_click
select
  release_session,
  release_status,
  device_num,
  device_type,
  json_value(exts,'$.user_id') as `user_id`,
  json_value(exts,'$.area_code') as area_code,
  json_value(exts,'$.aid') as aid,
  `date` as ct
from
ods_release
where release_status='2'
;
```

点击主题关联维度表

> 点击主题与地区维度表进行聚合，得出宽表

输入表

```
create table dw_release_click(
  release_session varchar,  -- comment '投放会话id'
  release_status varchar,   -- comment '投放状态'
  device_num varchar,       -- comment '设备唯一编码'
  device_type varchar,      -- comment '1 android| 2 ios | 9 其他'
  area_code varchar,        -- comment '地区'
  aid varchar,              -- comment '广告id'
  user_id varchar,          -- comment '用户id'
  ct date                   -- comment '创建时间'
)with(
type='datahub',
...
);

--dim维度表
-- (地区，省市，唯一地区编码,编码和city_id是一一对应的)
create table dim_province(
  area_code varchar,
  province_id bigint,
  province_name varchar,
  city_id bigint,
  city_name varchar,
  region_id bigint,
  region_name varchar,
 PRIMARY KEY (area_code),
 PERIOD FOR SYSTEM_TIME--定义维表的变化周期。
```

```
 )with(
    type= 'rds',
...
);

--(用户设备维度表)
create table dim_device(
device_type varchar comment '1 android| 2 ios | 9 其他',
device_name varchar comment '设备名字',
 PRIMARY KEY (device_type),
 PERIOD FOR SYSTEM_TIME--定义维表的变化周期。
)with(
type= 'rds',
...
);
```

输出表

```
create table dm_release_click(
  aid varchar,
  aid_count bigint,
  device_name varchar,
  area_code varchar,
  province_id bigint,
  province_name varchar,
  city_id bigint,
  city_name varchar,
  ct date
)with(
type='datahub',
...
);
```

业务代码

```
insert into dm_release_click
select
  a.aid,
  count(a.aid) aid_count,
  c.device_name,
  a.area_code,
  b.province_id,
  b.province_name,
  b.city_id,
  b.city_name,
  a.ct
from
dw_release_click a
join
dim_province  FOR SYSTEM_TIME AS OF PROCTIME() as b
```

```
on a.area_code=b.area_code
join
dim_device  FOR SYSTEM_TIME AS OF PROCTIME() as c on
a.device_type=c.device_type
group by
a.aid,
a.area_code,
a.ct
;
```

场景二：统计投放指标

某个广告在某个省的当天投放量

> 以aid和province_name分组，统计某个广告在某个省的当天投放量

输入表

```
create table dm_release_exposure(
  aid varchar,
  aid_count bigint,
  device_name varchar,
  area_code varchar,
  province_id bigint,
  province_name varchar,
  city_id bigint,
  city_name varchar,
  ct date
)with(
type='datahub',
...
);
```

输出表

```
--某个广告在某个省的当天投放量
CREATE TABLE ads_release_exposure_pro (
    aid                       VARCHAR,
    aid_count                 BIGINT,
    province_name             VARCHAR,
  ct                        DATE,
    primary key(aid,province_name,ct)
) WITH (
    type= 'rds',
...
);
```

业务代码

```
insert into ads_release_exposure_pro
select
  aid,
  sum(aid_count) as aid_count,
  province_name,
  ct
from
dm_release_exposure
group by
aid,
province_name,
ct
;
```

某个广告在某个市的当天投放量

> 以aid和city_name分组，统计某个广告在某个市的当天投放量

输入表

```
create table dm_release_exposure(
  aid varchar,
  aid_count bigint,
  device_name varchar,
  area_code varchar,
  province_id bigint,
  province_name varchar,
  city_id bigint,
  city_name varchar,
  ct date
)with(
type='datahub',
...
);
```

输出表

```
CREATE TABLE ads_release_exposure_city (
    aid                   VARCHAR,
    aid_count             BIGINT,
    city_name             VARCHAR,
  ct                    DATE,
    primary key(aid,city_name,ct)
) WITH (
    type= 'rds',
...
);
```

业务代码

```
insert into ads_release_exposure_city
select
  aid,
  sum(aid_count) as aid_count,
  city_name,
  ct
from
dm_release_exposure
group by
aid,
city_name,
ct
;
```

某个广告在某个投放终端的当天投放量

> 以aid和device_name分组，统计某个广告在某个用户客户端上的当天投放量

输入表

```
create table dm_release_exposure(
  aid varchar,
  aid_count bigint,
  device_name varchar,
  area_code varchar,
  province_id bigint,
  province_name varchar,
  city_id bigint,
  city_name varchar,
  ct date
)with(
type='datahub',
...
);
```

输出表

```
CREATE TABLE ads_release_exposure_device (
    aid                     VARCHAR,
    aid_count               BIGINT,
    device_name             VARCHAR,
  ct                      DATE,
    primary key(aid,device_name,ct)
) WITH (
    type= 'rds',
...
);
```

业务代码

```
insert into ads_release_exposure_device
select
  aid,
  sum(aid_count),
  device_name,
  ct
from
dm_release_exposure
group by
aid,
device_name,
ct
;
```

场景三：统计点击指标

某个广告在某个省的当天点击量

> 以ct和aid、provice_name分组，统计某个广告在某个省的当天点击量

输入表

```
create table dm_release_click(
  aid varchar,
  aid_count bigint,
  device_name varchar,
  area_code varchar,
  province_id bigint,
  province_name varchar,
  city_id bigint,
  city_name varchar,
  ct date
)with(
type='datahub',
...
);
```

输出表

```
CREATE TABLE ads_release_click_pro (
  aid                   VARCHAR,
  aid_count             BIGINT,
  province_name         VARCHAR,
  ct                    DATE,
  primary key(aid,province_name,ct)
) WITH (
```

```
  type= 'rds',
...
);
```

业务代码

```
insert into ads_release_click_pro
select
  aid,
  count(aid) as aid_count,
  province_name,
  ct
from
dm_release_click
group by
aid,
province_name,
ct
;
```

某个广告在某个市的当天点击量

> 以ct和aid、city_name分组，统计某个广告在某个市的当天点击量

输入表

```
create table dm_release_click(
aid varchar,
aid_count bigint,
device_name varchar,
area_code varchar,
province_id bigint,
province_name varchar,
city_id bigint,
city_name varchar,
ct date
)with(
type='datahub',
...
);
```

输出表

```
CREATE TABLE ads_release_click_city (
  aid                   VARCHAR,
  aid_count             BIGINT,
  city_name             VARCHAR,
  ct                    DATE,
```

```
  primary key(aid,city_name,ct)
) WITH (
  type= 'rds',
...
);
```

业务代码

```
insert into ads_release_click_city
select
aid,
count(aid) as aid_count,
city_name,
ct
from
dm_release_click
group by
aid,
city_name,
ct
;
```

某个广告在某个投放终端的当天投放量

> 以aid和device_name分组，统计某个广告在某个用户客户端上的当天投放量

输入表

```
create table dm_release_click(
  aid varchar,
  aid_count bigint,
  device_name varchar,
  area_code varchar,
  province_id bigint,
  province_name varchar,
  city_id bigint,
  city_name varchar,
  ct date
)with(
type='datahub',
...
);
```

输出表

```
CREATE TABLE ads_release_click_device (
  aid                     VARCHAR,
  aid_count               BIGINT,
```

```
  device_name             VARCHAR,
  ct                      DATE,
    primary key(aid,device_name,ct)
) WITH (
  type= 'rds',
...
);
```

业务代码

```
insert into ads_release_click_device
select
  aid,
  sum(aid_count),
  device_name,
  ct
from
dm_release_exposure
group by
aid,
device_name,
ct
;
```

场景四：热门广告排行榜

> 以ct和aid分组，计算当天每个广告的总点击量，对广告ID进行topn排序，得到点击次数最多的三个广告作为最热门广告。根据按天维度的时间字段（ct）和广告ID（aid）分组，计算每天每个广告的总点击量，根据广告ID对点击量进行topn排序，统计得到每天点击次数最多的三个广告，用于数据大屏中的热门广告排行榜。

输入表

```
create table dm_release_click(
aid varchar,
aid_count bigint,
area_code varchar,
province_id bigint,
province_name varchar,
city_id bigint,
city_name varchar,
ct date
)with(
type='datahub',
...
);
```

输出表

```
CREATE TABLE ads_release_click_dtclick (
  Ranking              BIGINT,
    aid                  VARCHAR,
    ct                   DATE,
  aid_count            BIGINT,
  primary key(aid,ct)
) WITH (
    type= 'rds',
...
);
```

业务代码

```
INSERT INTO ads_release_click_dtclick
SELECT
Ranking,
aid,
ct,
aid_count
FROM (
  SELECT *,
     ROW_NUMBER() OVER (PARTITION BY `ct` ORDER BY aid_count desc) AS Ranking
  FROM (
        SELECT
       `ct` AS `ct`,
        COUNT(aid) AS aid_count,
        aid
        FROM  dm_release_click
        GROUP BY `ct`,aid
    )a
)
WHERE Ranking <= 3
```