数据仓库规范

1. 新数据库

*

blued_mainland	blued 国内业务数据
blued_overseas	blued 国际业务数据
blued_buffer_mainland	blued 国内业务中间表（表不能删，但是可以删数据）
blued_buffer_overseas	blued 国际业务中间表（表不能删，但是可以删数据）
blued_common	blued 包含 ods,dws 等不区分国内国际的总数据以及无法区分国际国内的维度表。
blued_apm_mainland	blued 国内 apm 数据和业务分离
blued_apm_overseas	blued 国际 apm 数据和业务分离
tmp	临时数据（表可以随便删，包括 data 和 schema）
catch_overseas	catch 国际业务数据
xiaogege_mainland	小哥哥直播国内业务数据

2. 表名命名规则

重要：对所有表添加备注。

2.1 SD (Source Data Layer) -ODS 层（操作数据层）- 不进 BDP

*

格式：sd_业务数据表名_数据来源/sde_业务数据表名_数据来源
注：若数据源为 mysql 或 redis 等业务数据库，则采用 sd 开头。
　　若数据源为 kafka 日志，则将 json 展开，采用 sde 开头。
例：国内用户消费表，业务库 mysql 中为 users_orders,在数据仓库中表名为 blued_mainland .sd_users_orders_mysql;
　　当 sd 层数据为日志来源，则将外部 json 串展开，私有属性 extra 字段依旧保留 json 串存储，存成 orc 格式;

2.2 DD(Detail Data Layer)-DWD 层（细节数据层）- 不进 BDP

*

格式：dd_业务名

2.3 MD (Middle Data Layer) -DWS 层（轻级汇总数据层）- 不进 BDP

*

格式：fact_md_业务名_时间维度
例：国内访问日志的 md 层表名为 blued_mainland .fact_md_access_log_daily；其中 access_log 为业务名，daily 为日期纬度，代表按日维度聚合；

2.4 UA (User Level Aggregation Layer)- DM 层（数据集市层）- 用户级数据（百万~千万/天），可以进 BDP，数据保留不超过 3 个月

*

格式：fact_ua_业务名_时间维度_高级聚合维度首字母缩写
例：国内充值业务 ua 层表名为 blued_mainland.fact_ua_users_exchange_daily_upfs;
其中 users_exchange 是业务名，daily 代表按日维度，upfs 分别是 uid,platform_id,from_id,status 首字母;

2.5 PA (Presentation Layer)-非用户级（万/天）可以进 BDP，理论上支持全时间段数据

*

格式：fact_pa_业务名_时间维度

2.6 DIM 层

*

格式：dim_业务名

3. 建表 location

*

TODO：集群迁移后不应该放到/hive 下 格式：/产品名（例如 blued,catch）/数据库名字/表名 例：blued_mainland.fact_pa_push_switch_daily 的 location 为 /blued/blued_mainland/fact_pa_push_switch_daily

4. 文件格式

*

优先用 ORC 格式,压缩格式 zlib; 特殊情况需要文本格式，用 tab 分割，也就是 tsv 文件

5. 约定数据格式及用语

5.1 日期

*

命名：date 格式：date 例如：2019-01-01

5.2 日期分类

*

命名：granularity 格式：string 例如：day，week，month 备注：不是每个表必有，week 存当周周一日期，month 存当月 1 号日期。

5.3 事件

*

命名：event 格式：string

6. partition 顺序

原则上应以数据最终存储的目录分层来考虑合理的 partition 顺序

*

例如，有一个页面的日、周、月的 pv，partition 顺序为“granularity/date”更合理

目前推荐 partition 顺序如下

*

granularity/date/event

7. 存储集群名称

feature 新集群: feature apm 集群: apm