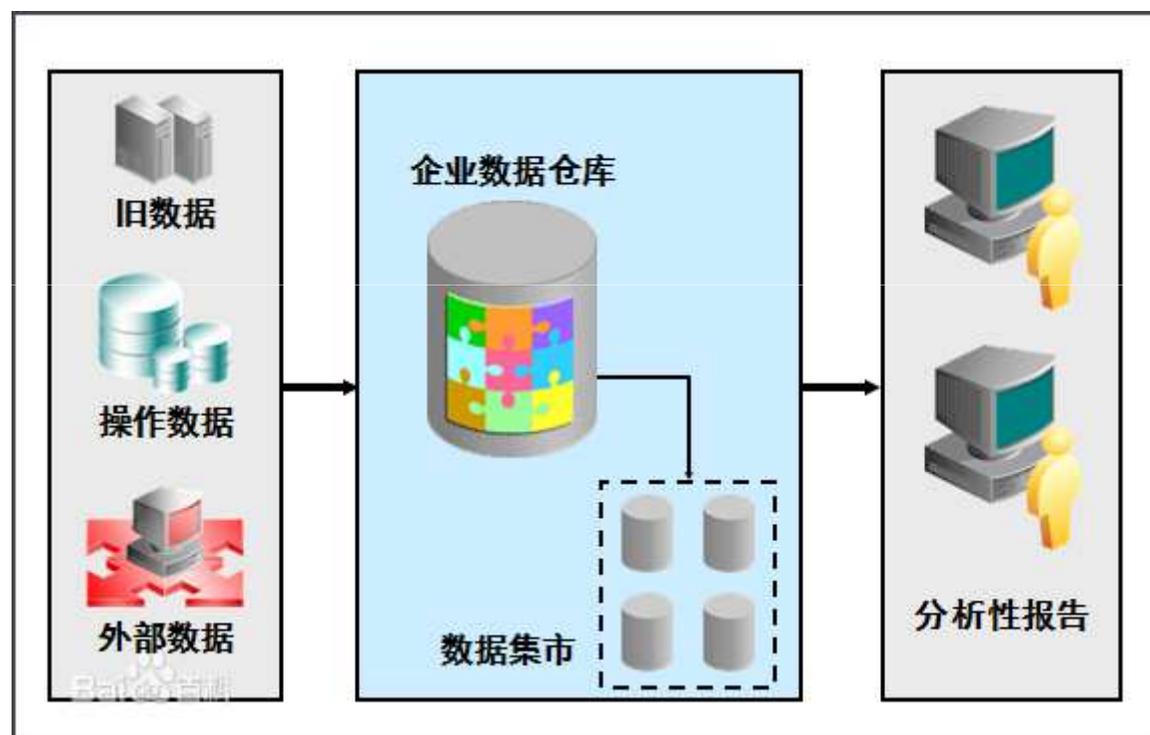


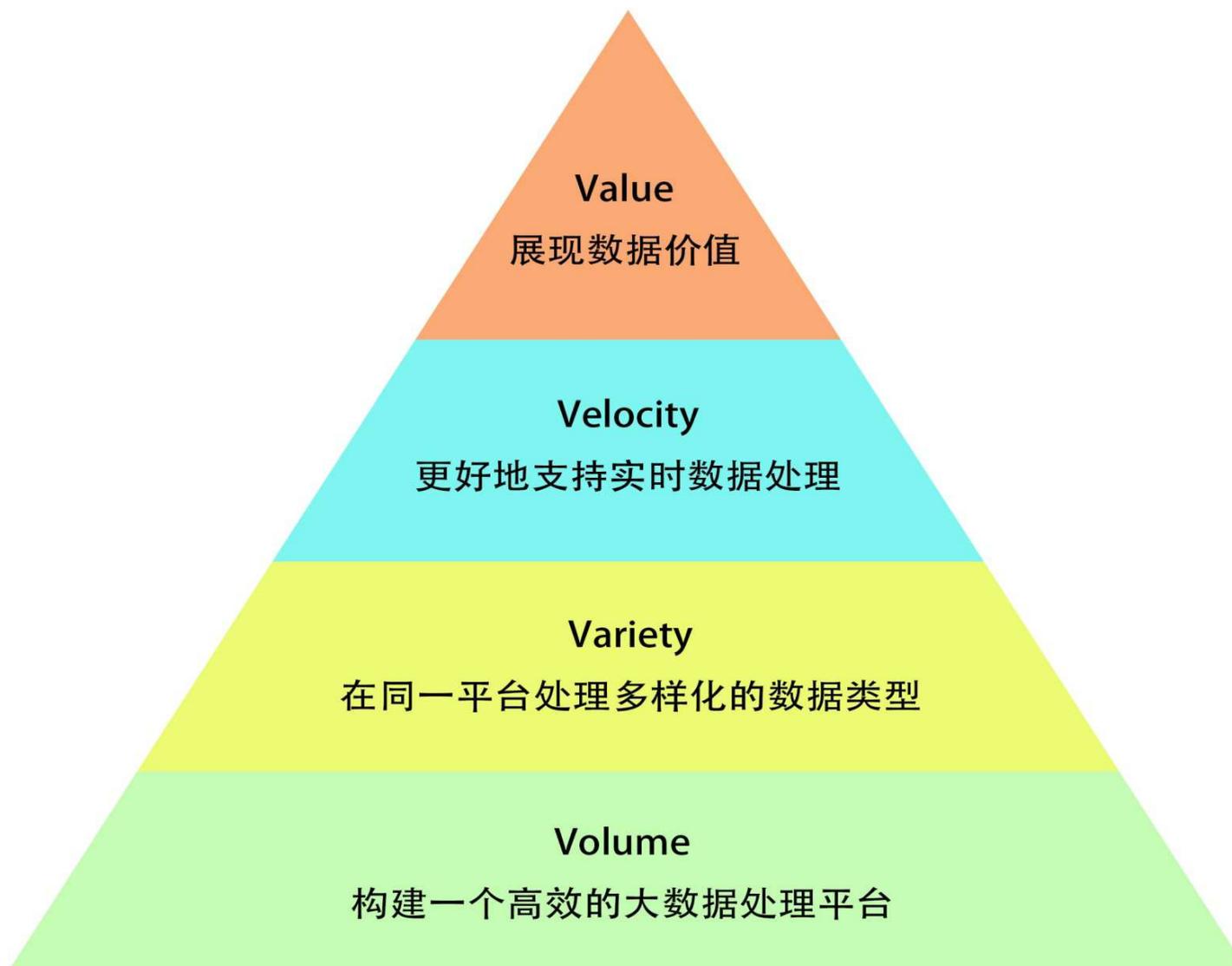
# Hbase架构简介

Report by 李修鹏

# 数据仓库



# 数据仓库



# 大数据的变革

# 数据仓库变化

- 只能支持战略决策->支持战略决策和战术决策（ tactical decision ）
  - 实时营销&个性化服务

# 实时主动数据仓库

(real-time active data warehouse)

## RTADW

# 实时主动数据仓库

- RTADW要集成的数据包括实时数据和历史数据两部分。
- 主动
  - 事件、条件、动作(event-condition-action, ECA)
- 实时事件进行主动分析和处理的能力

# 数据仓库架构介绍

# 建模角度

# 基础知识

- 第二范式（**2NF**）：首先是 **1NF**，另外包含两部分内容，一是表必须有一个主键；二是没有包含在主键中的列必须完全依赖于主键，而不能只依赖于主键的一部分。
- 第三范式（**3NF**）：首先是 **2NF**，另外非主键列必须直接依赖于主键，不能存在传递依赖。即不能存在：非主键列 **A** 依赖于非主键列 **B**，非主键列 **B** 依赖于主键的情况。

# 基础知识

- 星型模式
  - 性能优势
  - 业务模型
- 雪花型模式
  - 属性众多
  - 星型模式进一步层次化,减少数据冗余

# 重复性问题 交互性问题

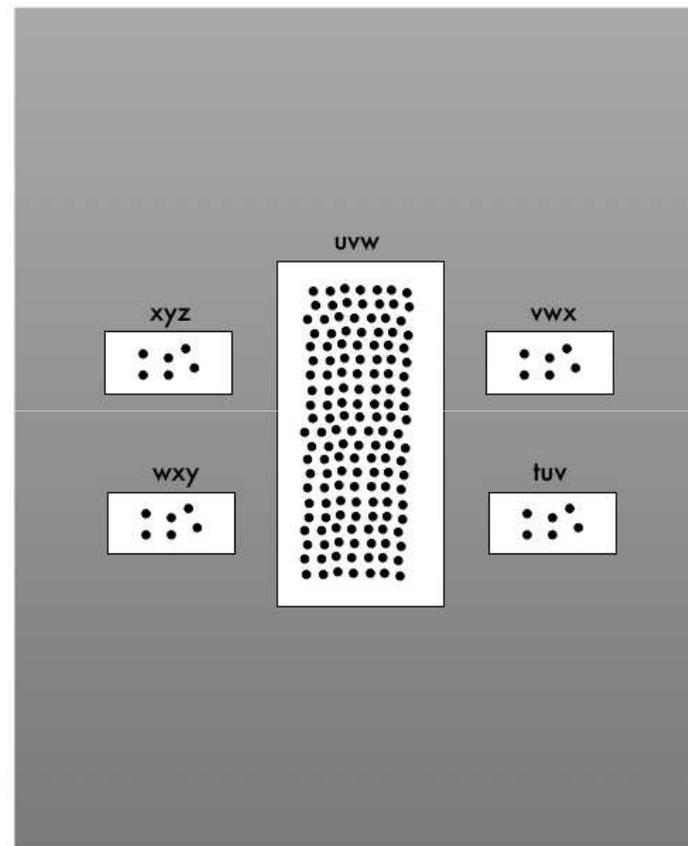
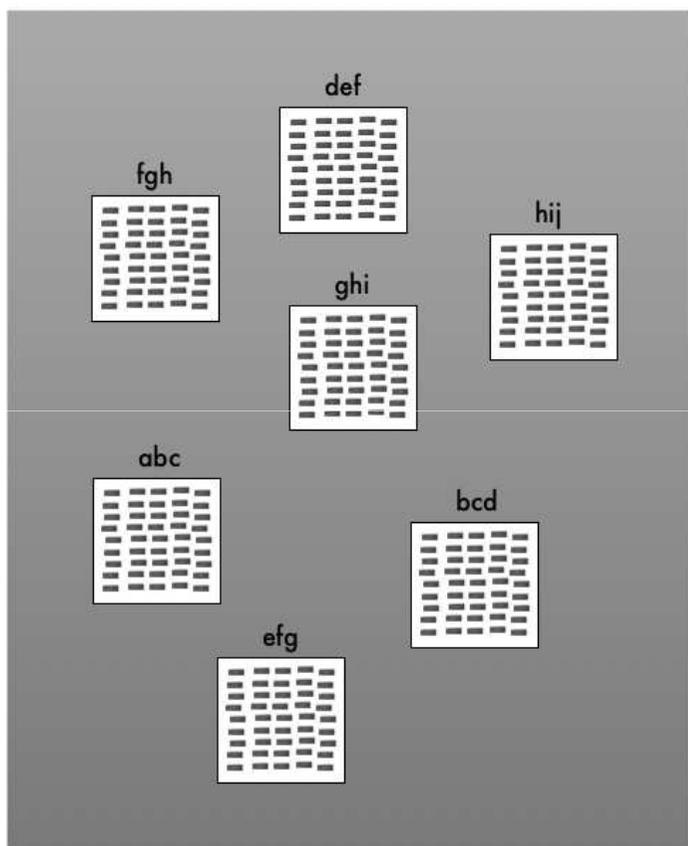
# 数据仓库vs数据集市

	数据仓库	数据集市
数据来源	遗留系统、OLTP系统、外部数据	数据仓库
范围	企业级	部门级或工作组级
主题	企业主体	部门或特殊的分析主题
数据粒度	最细的粒度	较粗的粒度
数据结构	规范化结构（第3范式）	星型模式、雪片模式或两者混合
历史数据	大量的历史数据	适度的历史数据
优化	处理海量数据	便于访问和分析快速查询
索引	高度索引	高度索引

# 数据仓库vs数据集市

数据仓库的数据结构

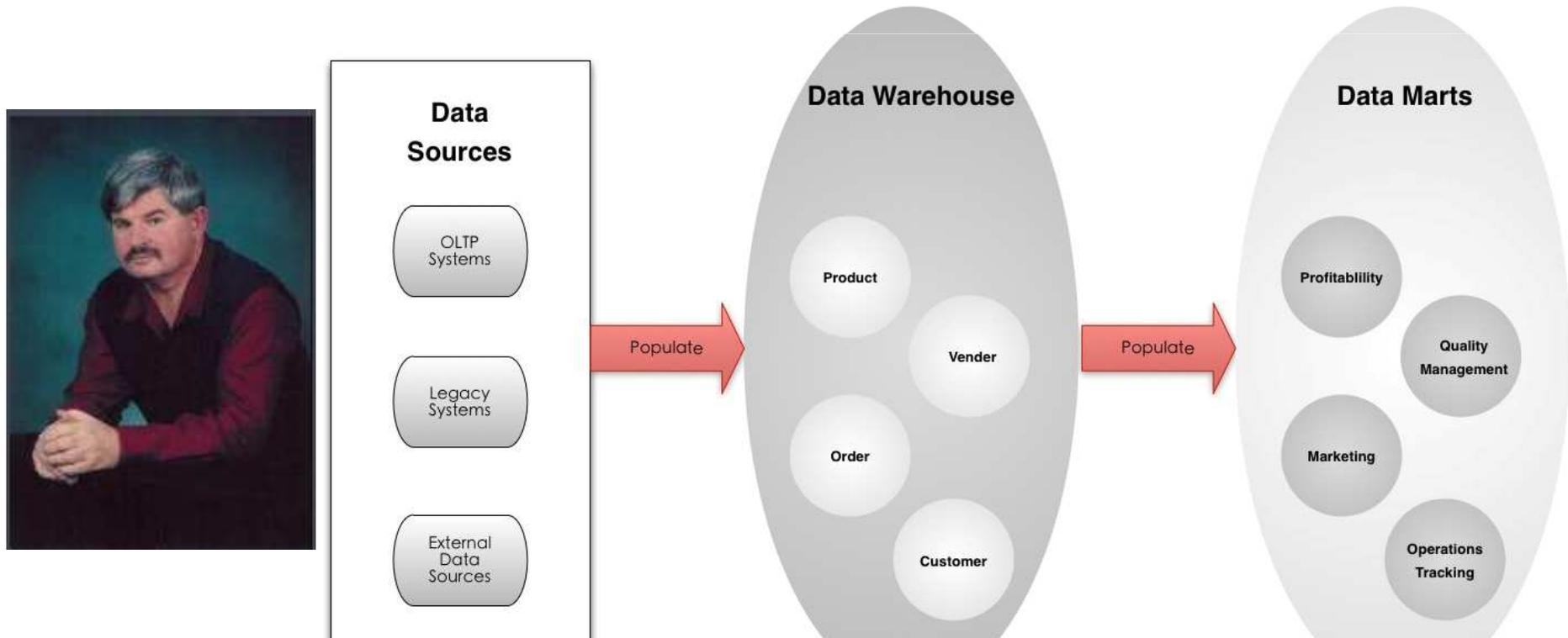
数据集市的数据结构



f_id	↓ Σ ▾ ▹ ▸	f_mostype	↓ Σ ▾ ▹ ▸	f_uid	▾ ▹ ▸	f_url	↓ Σ ▾ ▹ ▸	f_vid	↓ Σ ▾ ▹ ▸	f_lastdt	↓ Σ ▾ ▹ ▸								
69		1		b7a7c662a53bb...		7002		1090667		20140827									
f_uid	▾ ▹ ▸	f_pro	▾ ▹ ▸	f_mtype	▾ ▹ ▸	f_cv	▾ ▹ ▸	f_channelid	▾ ▹ ▸	f_province	▾ ▹ ▸	f_city	▾ ▹ ▸	f_mfov	▾ ▹ ▸	f_mfo	▾ ▹ ▸	f_webdealer	▾ ▹ ▸
%2bXZDXzRU...	1		11		2.9.0		319		18		1822		Nokia 800		NOKIA		2		

# Inmon 和 Kimball的大辩论

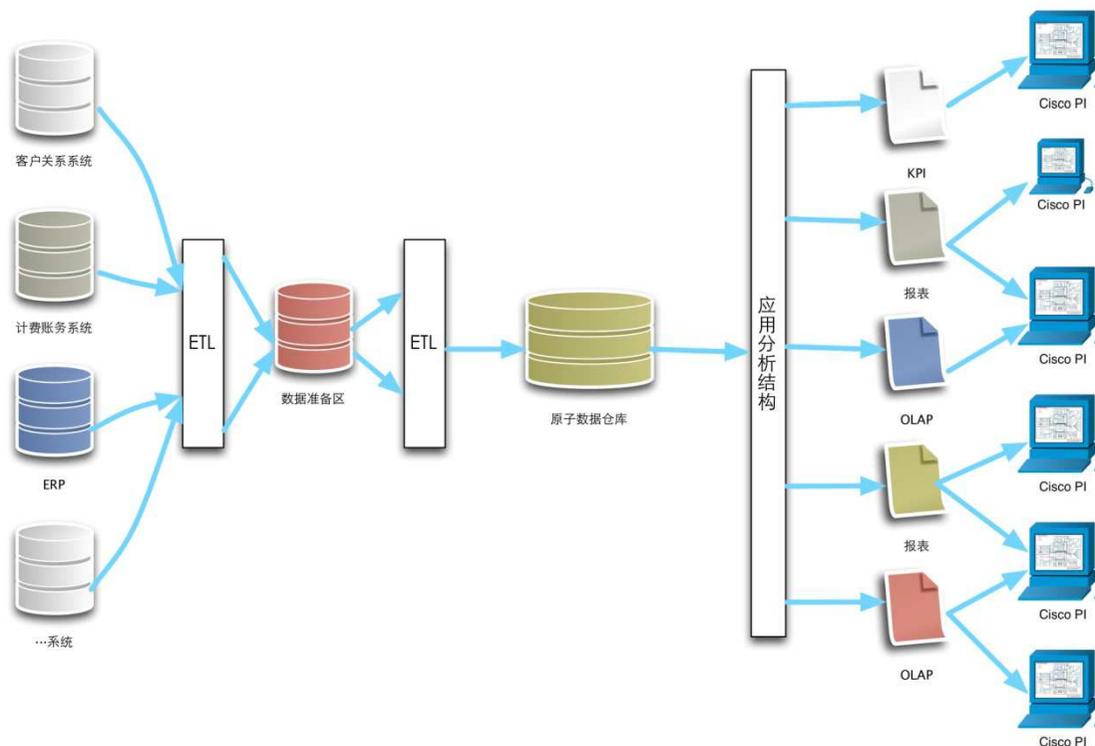
- Bill Inmon 将数据仓库定义为“一个面向主题的、集成的、随时间变化的、非易变的用于支持管理的决策过程的数据集合”



# 数据仓库架构

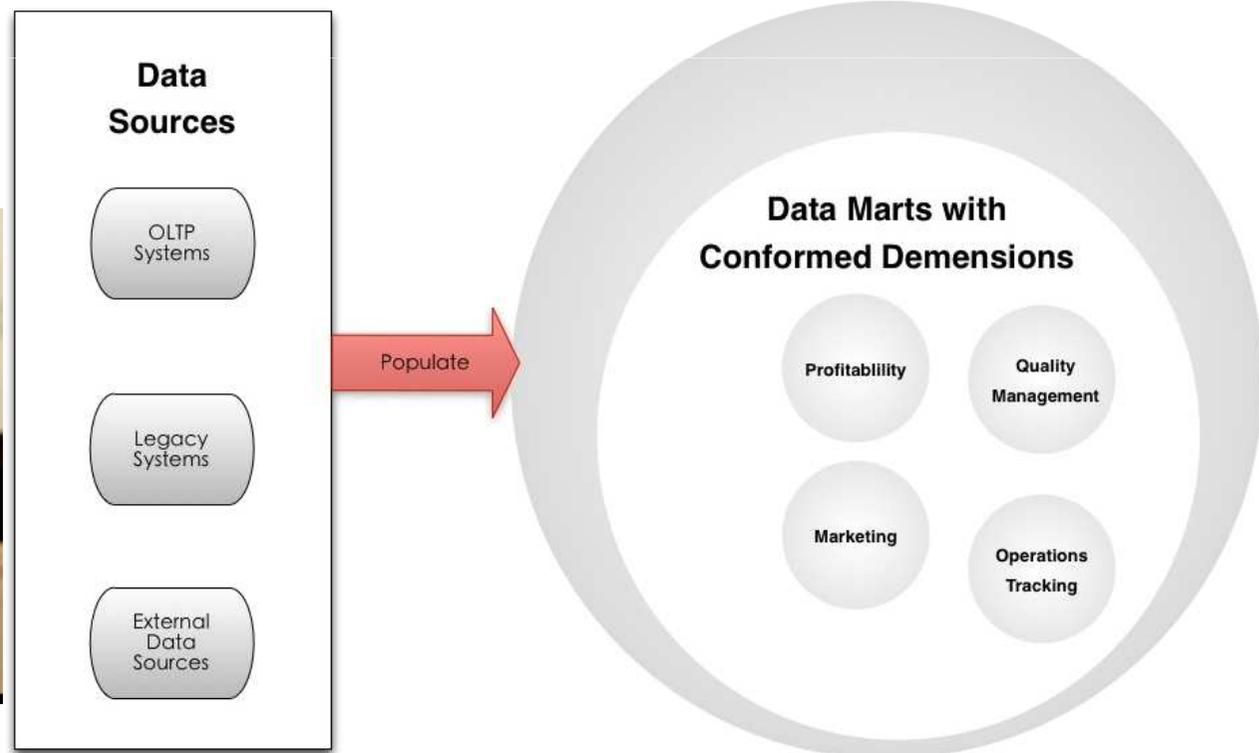
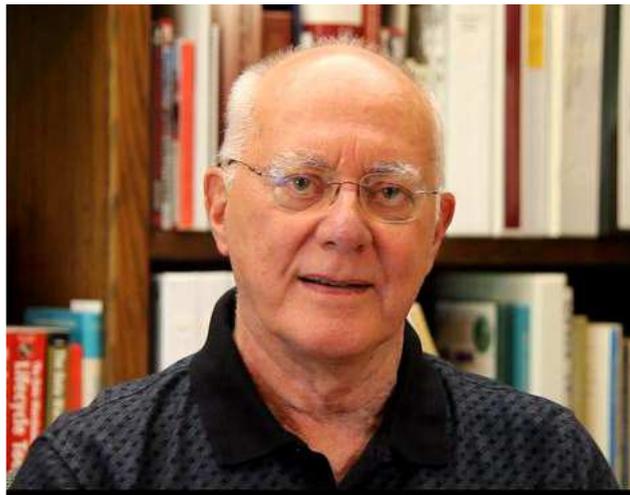
- 集中式架构

---标识着数据仓库架构已经进入比较成熟的时期



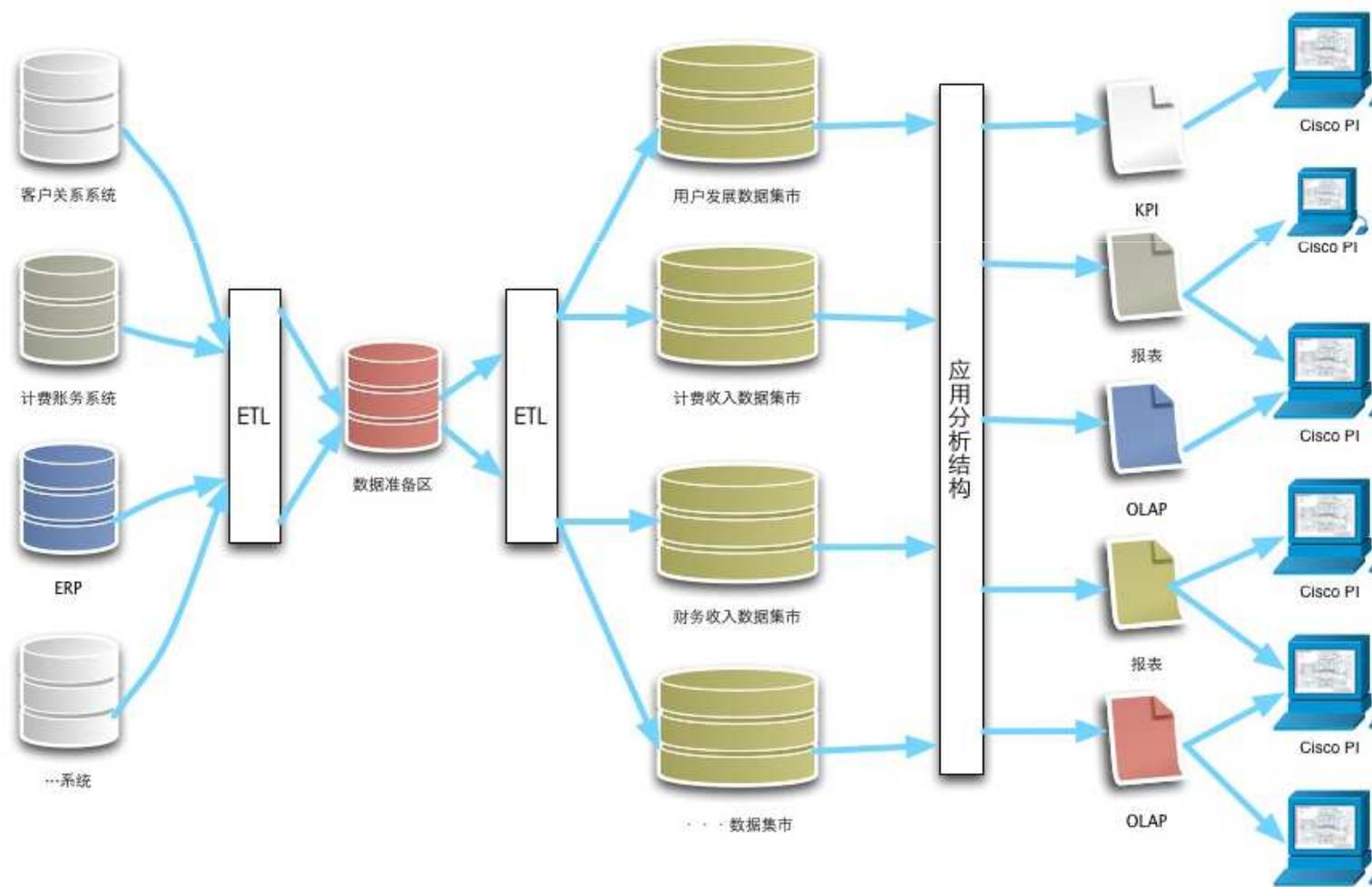
# Inmon 和 Kimball的大辩论

- Ralph Kimball 说“数据仓库仅仅是构成它的数据集市的联合”，他认为“可以通过一系列维数相同的数据集市递增地构建数据仓库”



# 数据仓库架构

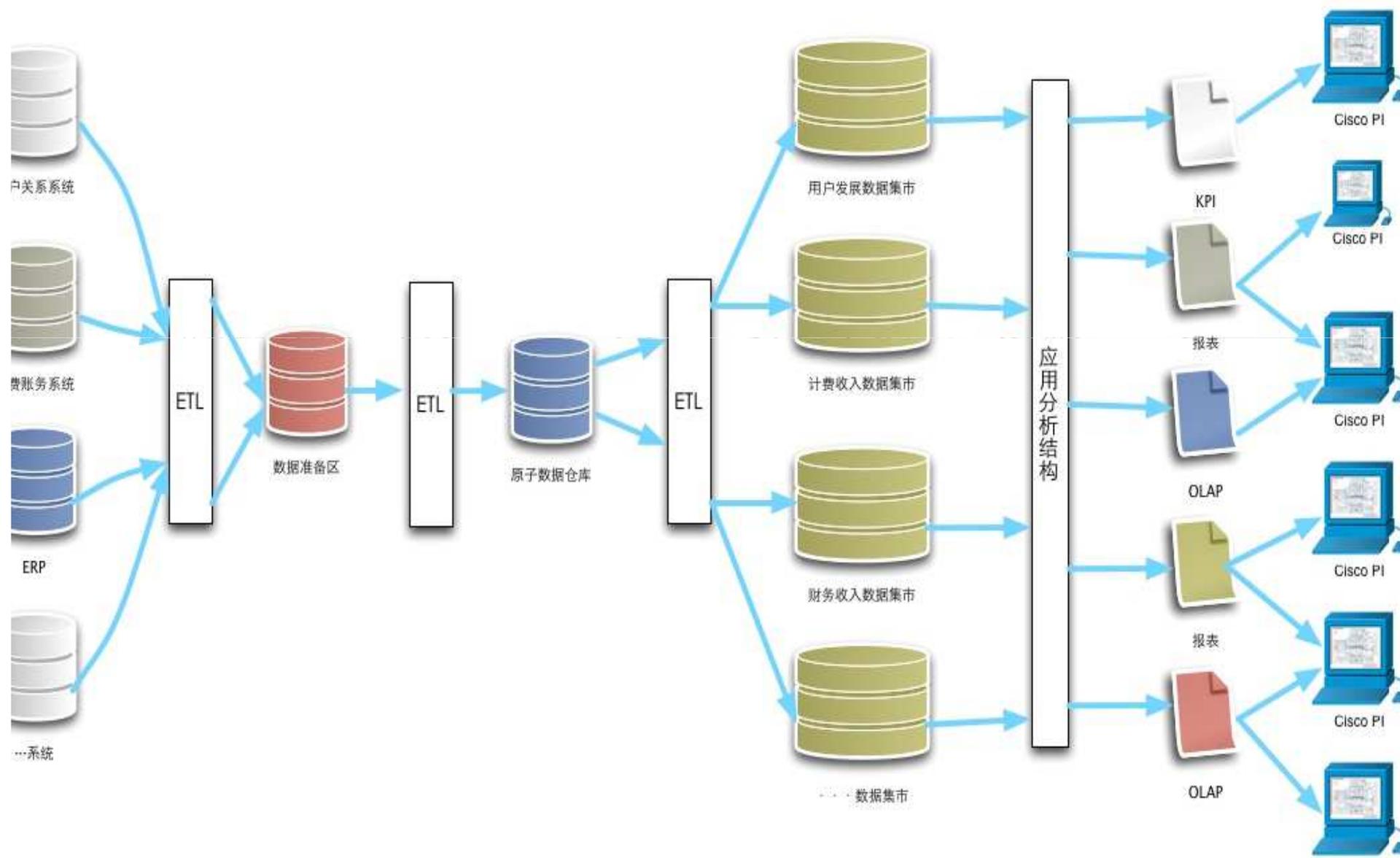
## • 总线架构



# 数据仓库架构

- 独立的数据集市架构
  - 去哪儿 事业部一个数据集市
  - 不是企业内一致的数据，产生信息孤岛
- 联邦式数据仓库架构
  - 原有独立数据集市的数据交换

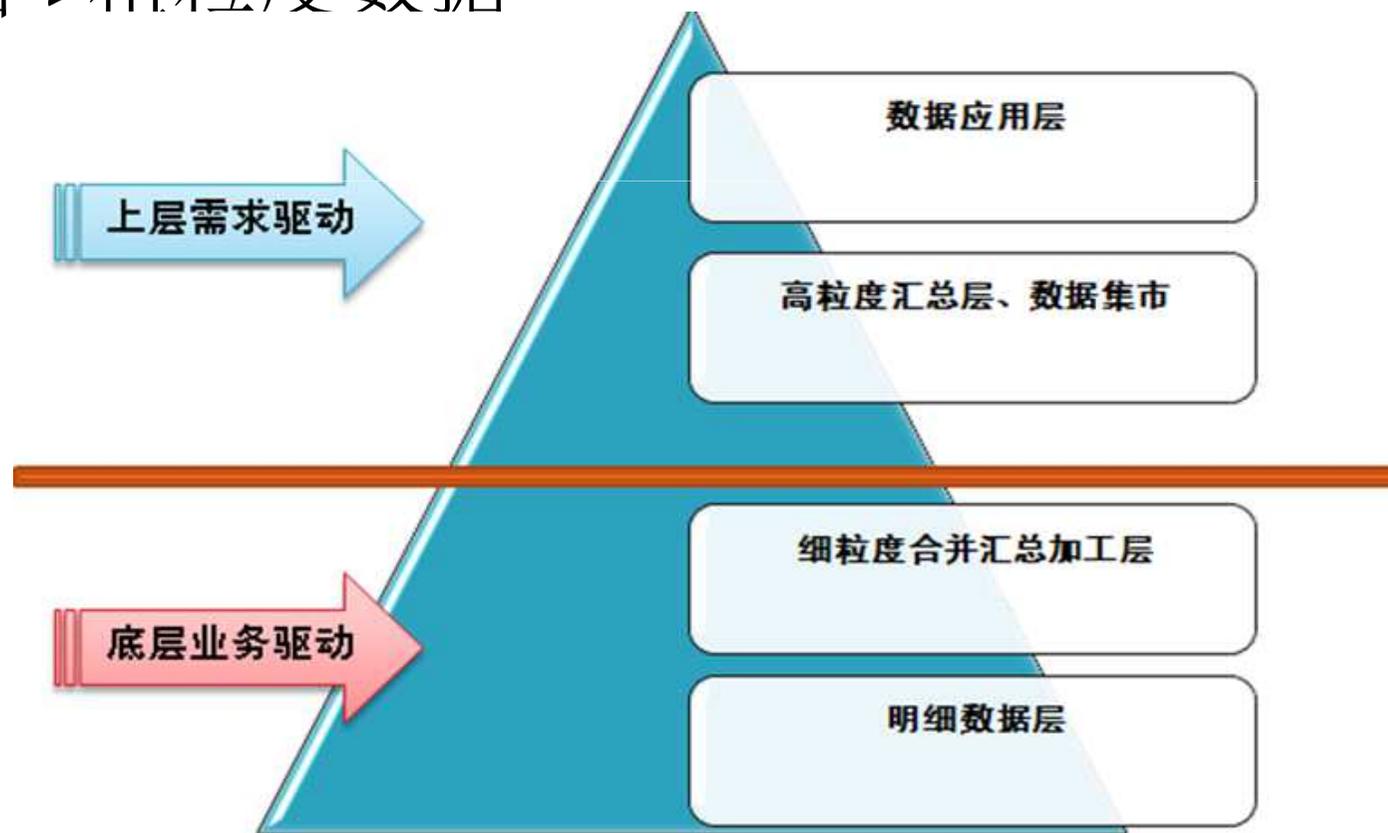
# 数据仓库架构



# HOW WE DO

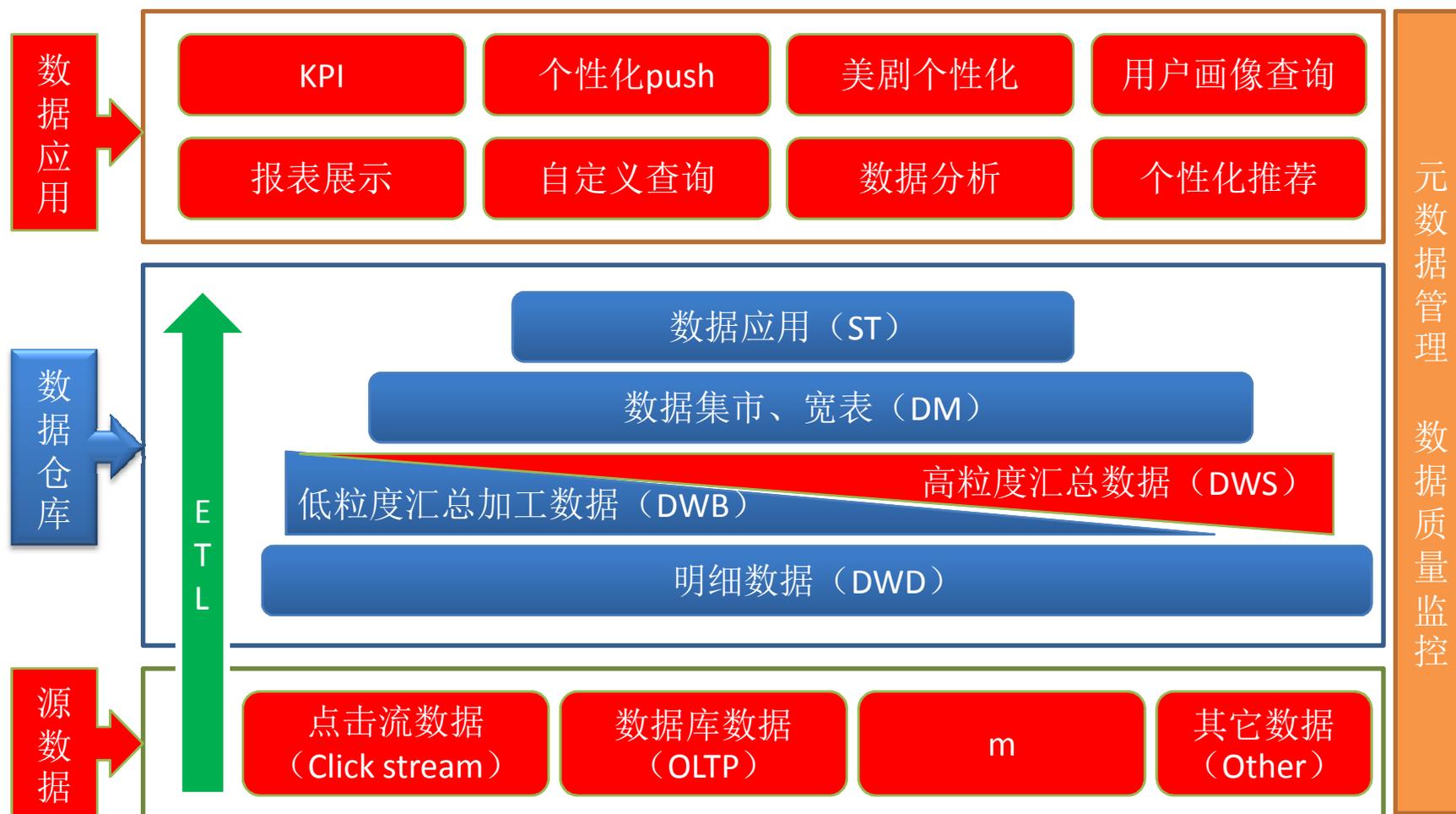
# 需求

- 数据挖掘->细粒度数据
- 统计数据->粗粒度数据





# 建立数据仓库架构

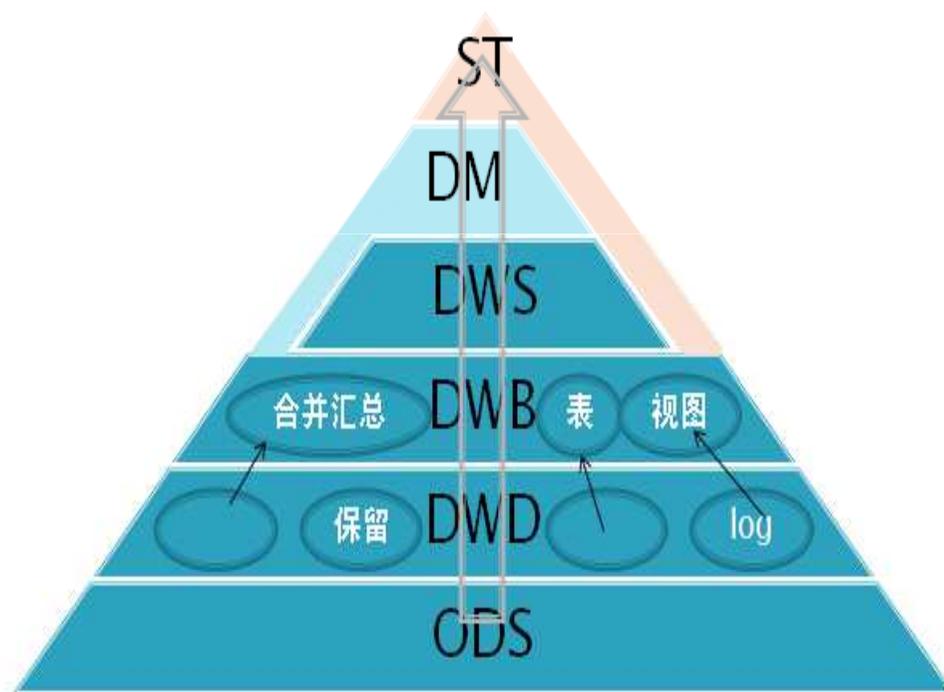


# DW五层模型架构介绍

❖ DW五层模型是按照EDW各个应用层次的需求进行分层细化而来的，每个层次满足不同的应用。

❖ 分为以下5层：

1. ODS 数据准备层
2. DWD 数据明细层
3. DW(B/S) 数据汇总层
4. DM 数据集市层
5. ST 数据应用层





# DW五层模型架构介绍

	数据来源及建模方式	服务领域	数据ETL过程描述
ST层	数据来自DW层，采用维度建模，星型架构	前端报表展现，主题分析，KPI报表	从DW层的数据进行粗粒度聚合汇总；如按年、月、季、天对一些维度进行聚合生成业务需要的事实数据
DM层	数据来自DW层，采用维度建模，星型架构	数据挖掘，自定义查询，应用集市	从DW层的数据进行粗粒度聚合汇总；按业务需求对事实进行拉宽形成宽表
DW层	数据来自DWD层，是DW事实层，采用维度建模，星型架构，这一层可细分为dwb和dws	为EDW提供各种统计汇总数据	从DWD层进行轻度清洗，转换，汇总聚合生成DW层数据，如字符合并，Cv,uid,日期，mtype，合并；用代理键取代维度；按各个维度进行聚合汇总
DWD层	数据来自ODS层，是DW明细事实层,数据模型是ODS一致	为EDW提供各主题业务明细数据	根据ODS增量数据进行merge生成全量数据，不做清洗转换，保留原始全量数据
ODS层	数据准备区，数据来源是各业务系统的源数据，物理模型和业务系统模型一致。	为其它逻辑层提供数据，为统一数据视图子系统提供数据实时查询	通过移动视频dc中心平台，把业务数据抽取落地成文本文件，再装载到数据仓库ODS层，不做清洗转换

# Demo 数据

	f_platform integer	f_uid character varying	f_day integer	f_playlistid bigint	f_hour integer	f_vid bigint	f_catecode integer	f_videolength integer	f_site smallint	f_speed integer	f_suspend integer	f_backward integer	f_slide integer	f_playtime integer
i3	6	1314e0a1de033d9:	20141113	419	13	7260	100	7069	1	3	1	2	2	4320
i4	6	6c24d0b7360bba4:	20141113	7275034	20	2018559	115	404	1	3	0	2	0	417
i5	6	41796368399cd9f:	20141113	8327823	10	2076963	101	3600	1	2	0	4	0	141
i6	6	57c609cd5c27700:	20141113	7140110	21	2017181	101	2436	1	4	0	2	0	2198
i7	6	2096946f5b4ec15:	20141113	8327823	15	2076963	101	3600	1	186	0	2	0	1
i8	6	e293c41c57c272e:	20141113	6996011	21	2017578	101	3532	1	112	1	34	0	1683

f_uid character varying	f_mtype integer	f_first_catecode integer	f_second_catecode integer	f_countview integer	f_goodview integer	f_badview integer	f_playtimetate numeric(14,8)
fcc796a29291d6f96ed:	6	126	126108	2	0	2	0.00000000
fcc796a29291d6f96ed:	6	129	129109	1	1	0	1.00000000
fcc796a29291d6f96ed:	6	189	189105	1	0	1	0.00000000
fcc796a29291d6f96ed:	6		-1	14	6	8	
fcc796a29291d6f96ed:	6	124	124105	5	0	5	0.03846154
fcc796a29291d6f96ed:	6	130	130106	1	1	0	0.12500000
fcc796a29291d6f96ed:	6	129	129118	2	0	2	0.00000000

# 消息 历史

er varying	f_frequency integer	f_onesvs integer	f_onemy integer	f_oneplaytimevs integer	f_oneplaytimemy integer	f_twovs integer	f_twomy integer	f_twoplaytimevs integer	f_twoplaytimemy integer	f_threevs integer	f_t inte
i95dccf3c5219:	2	0	0	0	0	0	0	0	0	0	0
if7669b4ab0ea:	8	0	0	0	0	0	0	0	0	0	0
la72a4bc25fef:	7	15	28	7047	199	0	0	0	0	0	0
fd109290cc8e:	6	0	0	0	0	0	0	0	0	0	0

f_uid character varying	f_pro character	f_mtype character varying	f_cv character va	f_channelid character var	f_province character vai	f_city character varying	f_mfov character varying	f_mfo character varying	f_webdeak character v
2bJEZOyqy86CpXVPOq:	1	11	2.9.0	319	22	2205	Lumia 800	NOKIA	1
2benkBHP4Qao12qTzm:	1	11	2.9.0	319	18	1822	Nokia 710	NOKIA	3



# Demo 数据

f_day integer	f_time_type character varying	f_platform character varying	f_position character varying	f_uv bigint	f_cc bigint	f_site character varying
2014120	day	android	app my like currentplay	320	1224	short
2014120	day	android	app my like flowcolumn	7424	9067	all
2014120	day	iphone	app my like tagcolumn v	70	99	long
2014120	01	android	app my like currentplay	8	28	short
2014120	01	iphone	app my like tagselect s	19	19	all
2014120	02	iphone	app my like flowcolumn	29	31	all
2014120	02	android	app my like rcpage load	100	122	all

f_channeled_id	f_channeled_name	f_channeled_type	f_modified_time
1000022103	美剧 - 推荐 [算法推...	2	2014-11-24 11:30:03
1000060007	流式详情页选集	2	2014-11-24 11:30:03
1000030002	直播推荐页	2	2014-11-24 11:30:03

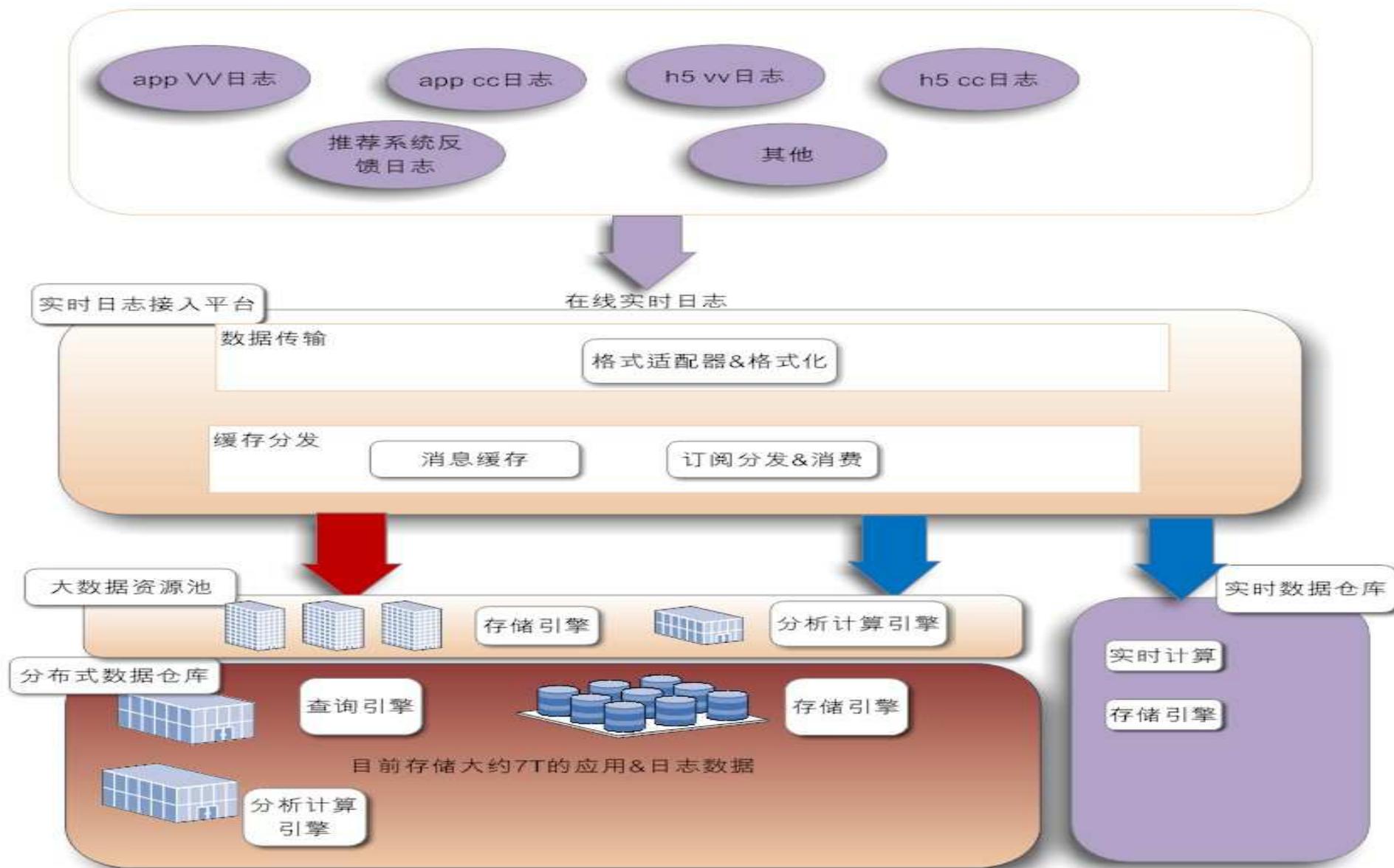
f_album_id	f_uid_count	f_album_view_count	f_avg_vv
5766709	7105	26796	1218.00
7038181	149607	286996	28699.60



# Demo 数据

```
uid
history(用户观影记录 之前徐佳出的):
    type, playlistid, vid, playpercent, playtime, td, lastdt;.....
searchinfo(用户搜索信息):
    keyword, lastdt, vid, playlistid;.....
actionInfo(用户行为数据, 目前可以先不用):
    vid, lastdt;.....
attention(用户关注数据):
    f_playlistid, f_lastwatchtime;
viewdetailinfo(用户视频播放情况):
    vid, playlistid, catecode, videlength, day, hour, speed, suspend, backward, slide, playtime; .....
feature:
    用户平台信息
timecatecodeinfo(用户对类别的喜好数据):
    hour, f_first_cate_code, f_second_cate_code, f_countview, f_goodview, f_badview, f_playtimetate
ugcviewdetailinfo(用户视频播放情况):
    vid, catecode, videlength, day, hour, speed, suspend, backward, slide, playtime; .....
ugctimecatecodeinfo(用户对类别的喜好数据):
    hour, f_cate_code, f_countview, f_goodview, f_badview, f_playtimetate
|
```

# dm数据仓库架构



# The next to do

- 实时窗口
- 主动决策
  - 应用内部通知消息数据
- 继续改进和优化现有宽表的物理实现

播放情况						播放行为			
VID	视频名称	视频专辑ID	播放来源	视频播放时间	播放结束时间	行为ID	行为名称	行为发生时间	tag
1908151	匆匆那年第2集	6906306	搜索结果-专辑	20	2014/12/05 11:44:20	21004	全网搜索点击剧集	2014/12/05 11:43:50	匆匆那年
						10001	点击搜索结果	2014/12/05 11:43:50	匆匆那年
						7012	进入中间页	2014/12/05 11:43:50	0
						7018	剧集点选	2014/12/05 11:43:50	0
						9082	播放器解码类	2014/12/05	0

**Thanks!**  
**Q&A**