百度地图 开放平台

智慧交通与数据仓库

——实习四周串讲

李涛

2017.07.14

目录

业务相关：拥堵指数

- 计算
  - **拥堵指数**

    1.畅通速度 / 平均速度

  - **畅通速度**

    v_{ichang}, m_i

    l_i, width

  - **平均速度**

    加权总长度 / 加权总时间

$$v_{chang} = \frac{\sum_{i=1}^{n} l_i \cdot w_i}{\sum_{i=1}^{n} t_i \cdot w_i} = \frac{\sum_{i=1}^{n} l_i \cdot \frac{m_i}{l_i \cdot width}}{\sum_{i=1}^{n} \frac{l_i}{v_{ichang}} \cdot \frac{m_i}{l_i \cdot width}} = \frac{\sum_{i=1}^{n} \frac{m_i}{width}}{\sum_{i=1}^{n} \frac{m_i}{v_{ichang} \cdot width}}$$

$$w_i = \frac{m_i}{l_i \cdot width}$$

$$t_i = \frac{l_i}{v_{ichang}}$$

业务相关：拥堵指数

- 所需表
  - **畅通速度表：**lbs_map_traffic_yongdu_changtong_basic 畅通速度与平均权重
  - link权重表：lbs_map_traffic_yongdu_weight_basic
  - link长度宽度表：lbs_map_traffic_yongdu_road_network_basic
  - 当时路况表：lbs_map_traffic_yongdu_lukuang_basic
  - 加权总里程：**lbs_map_traffic_yongdu_lukuang_granularity_dwa**

目录

数据仓库

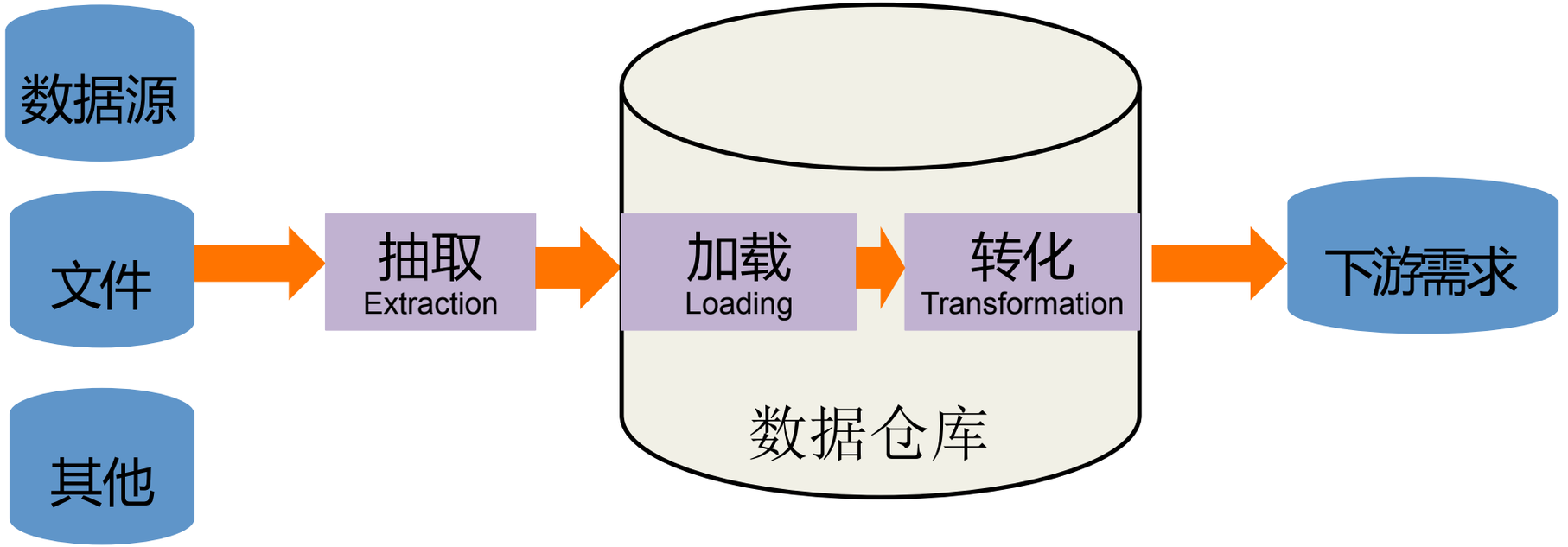
- OLTP VS OLAP
- ETL
- 研判平台数据架构
- 研判平台仓库架构
- 元数据管理
- 命名规范

OLTP VS OLAP

OLTP	OLAP
面向事务	面向主题
增删查改	查询
业务人员	决策人员
实时	历史数据

数据仓库系统的主要应用主要是联机分析处理OLAP（On-Line Analytical Processing），支持复杂的分析操作，侧重决策支持，并且提供直观易懂的查询结果。

二 基本的ETL



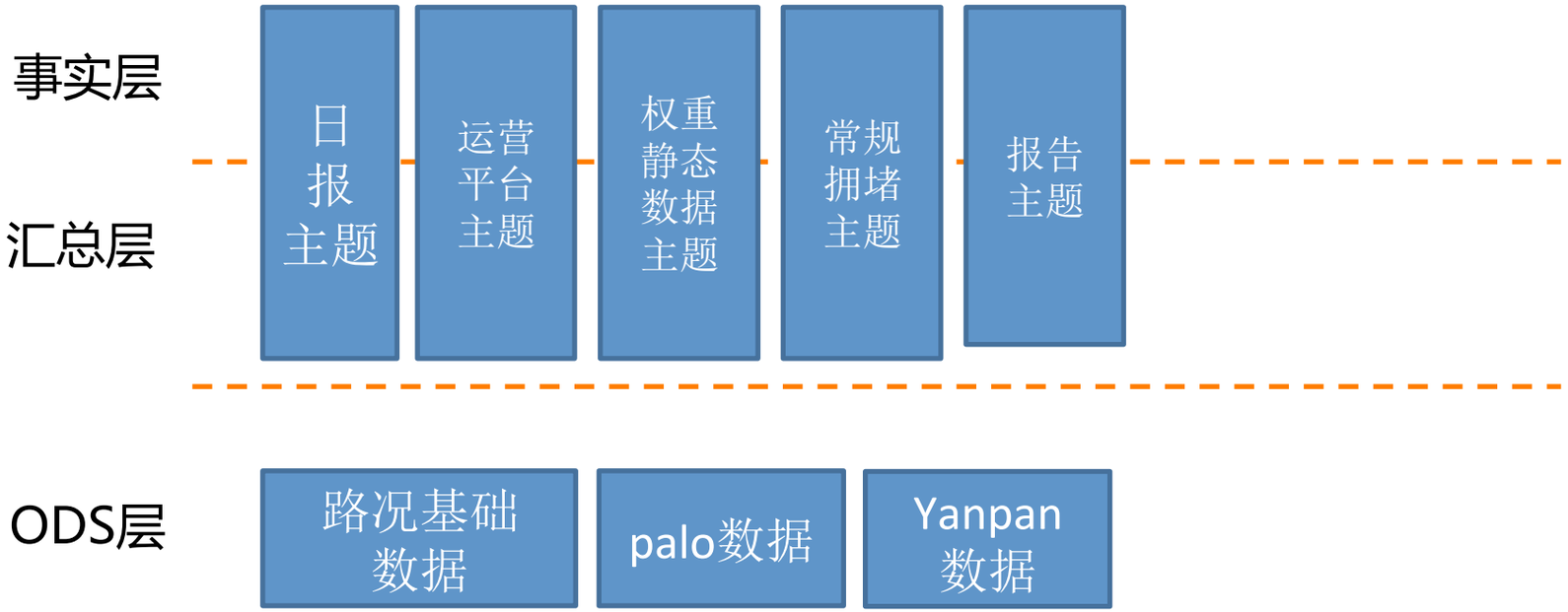
抽取：将数据从原始的业务系统中读出来，全量或增量抽取

加载：数据按计划导入到数据仓库

转化：按照规则将数据转化

每一步伴随着存储传输计算。

三 研判平台的数据架构



- 纵向上分为ODS（Operational Data Store）层，汇总层，事实层，横向上按照主题划分。
- 一般由汇总层和事实层对外提供数据服务，ODS层不对外开放。

二 研判平台的数据架构

ODS层

- 用于存储来自于不同数据源的历史数据，包

首页 / 数据浏览 / default.lbs_map_traffic_yanpan_main_basic

基本信息 | schema | 数据样例 | 数据质量 | 存储计划 | 动态信息 | 存储信息 | 数据表历史

Schemas

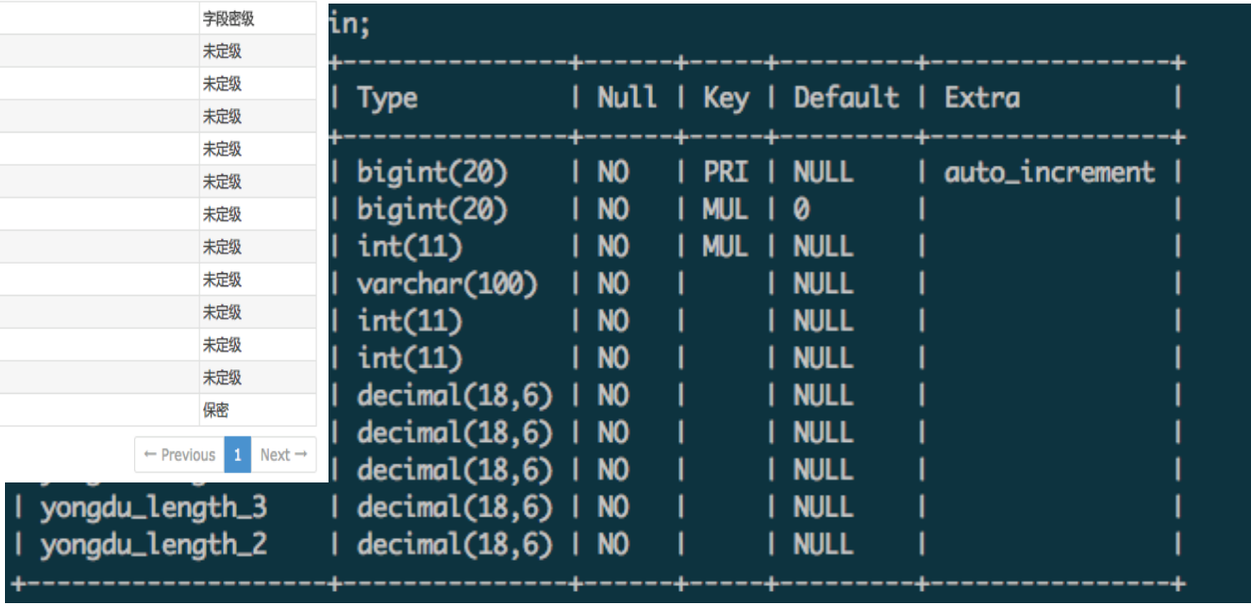
100 records per page

Search:

字段名	字段类型	字段说明	字段例子	字段密级
id	bigint	自增id		未定级
time	string	时间切片		未定级
citycode	bigint	城市编码		未定级
cityname	string	城市名称		未定级
road_type	bigint	道路类型	全部；高速；环路快速路；主干路；次干路；支路；其他	未定级
district_type	bigint	城区类型	主城区；全市	未定级
yongdu_index	double	拥堵指数		未定级
road_network_speed	double	路网的平均速度		未定级
yongdu_length_4	double	对应拥堵等级4		未定级
yongdu_length_3	double	对应拥堵等级3		未定级
yongdu_length_2	double	对应拥堵等级2		未定级
event_day	string	日期	分区	保密

Showing 1 to 12 of 12 entries

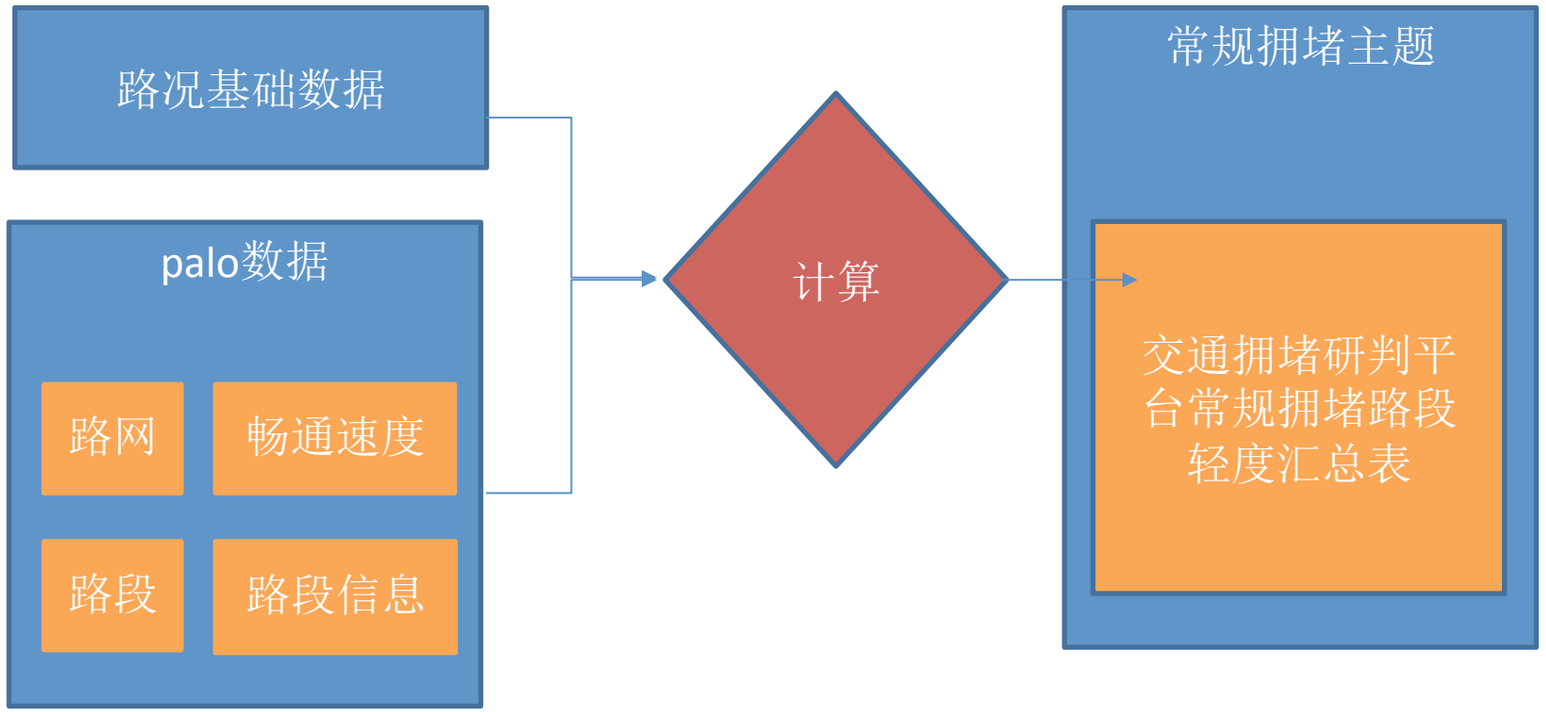
← Previous 1 Next →



```
in;
+--------------+------+-----+---------+----------------+
| Type         | Null | Key | Default | Extra          |
+--------------+------+-----+---------+----------------+
| bigint(20)   | NO   | PRI | NULL    | auto_increment |
| bigint(20)   | NO   | MUL | 0       |                |
| int(11)      | NO   | MUL | NULL    |                |
| varchar(100) | NO   |     | NULL    |                |
| int(11)      | NO   |     | NULL    |                |
| int(11)      | NO   |     | NULL    |                |
| decimal(18,6)| NO   |     | NULL    |                |
| decimal(18,6)| NO   |     | NULL    |                |
| decimal(18,6)| NO   |     | NULL    |                |
| yongdu_length_3 | decimal(18,6) | NO |  | NULL |  |
| yongdu_length_2 | decimal(18,6) | NO |  | NULL |  |
+-----------------+---------------+----+---+------+---+
```

二 研判平台的数据架构

汇总层



- 一般只做数据清洗，保留原粒度
- 数据融合，按照主题属性进行预加工
- 汇总层逻辑比较复杂，可进行更细层次的划分。

二 研判平台的数据架构

百度地图 开放平台

事实层

首页 / 数据浏览 / default.lbs_map_traffic_yanpan_daily_report_majorcity_fact

基本信息 | schema | 数据样例 | 数据质量 | 存储计划 | 动态信息 | 存储信息 | 数据表历史

Schemas

100 records per page

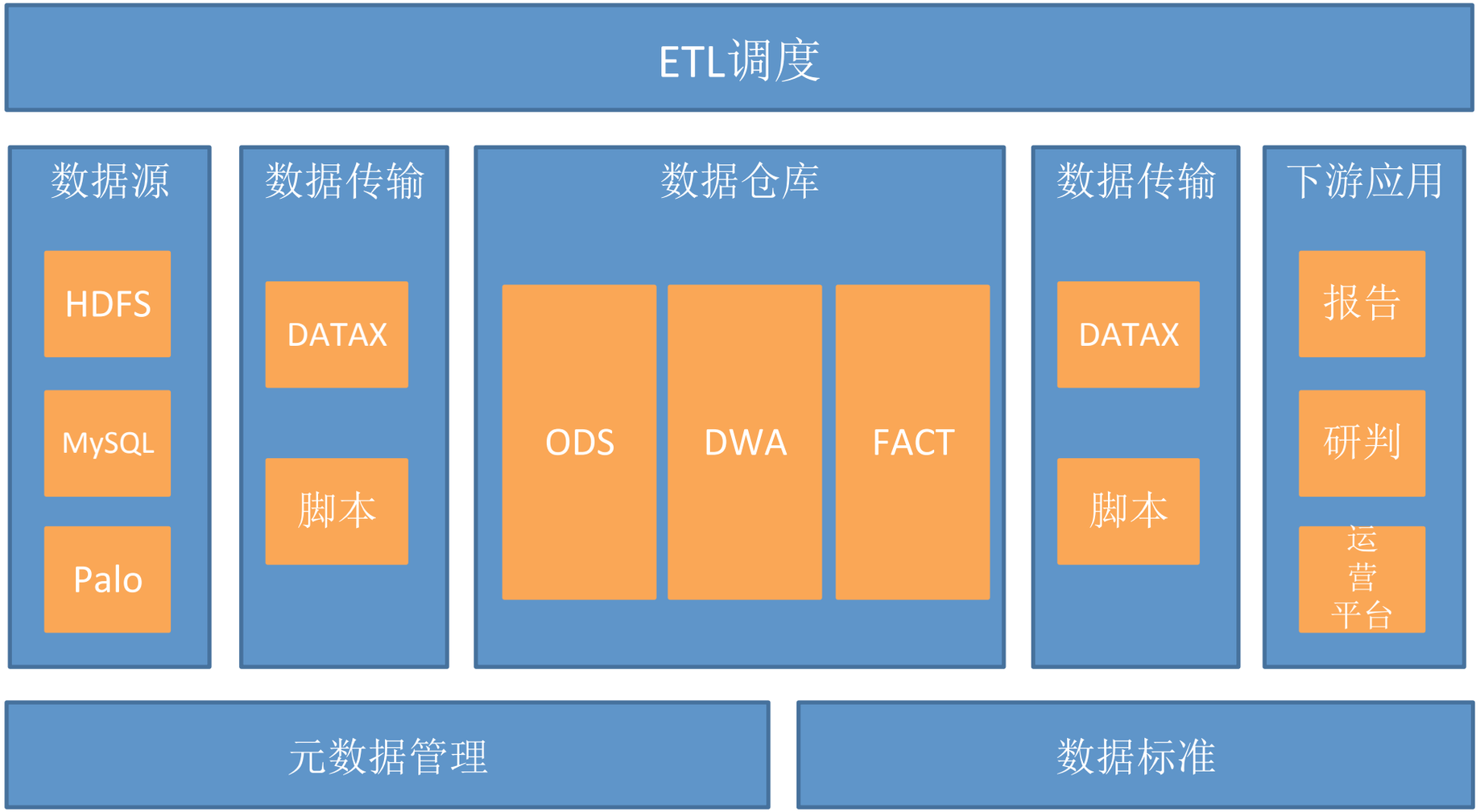
Search:

字段名	字段类型	字段说明		
citycode	int	城市编码		
road_type	int	道路类型		
district_type	int	区域类型		
datatime	string	数据时间		
congest_interval_info	string	分时段拥堵指标		
congest_period_info	string	拥堵时段时长信息		
congest_hour_info	string	拥堵高峰小时信息		
curve	array<string>	当前拥堵指标曲线		
curve30	array<string>	近30拥堵指标曲线		
congest_road_num_curve	array<string>	拥堵路段条数曲线		
max_day_total_road	int	本日拥堵路段总计		
max_time_period	string	本日最拥堵时间段		未定级
time_road_count_avg	float	本日最拥堵时间段平均拥堵路段数		未定级
event_day	string	日期		保密

据均由事

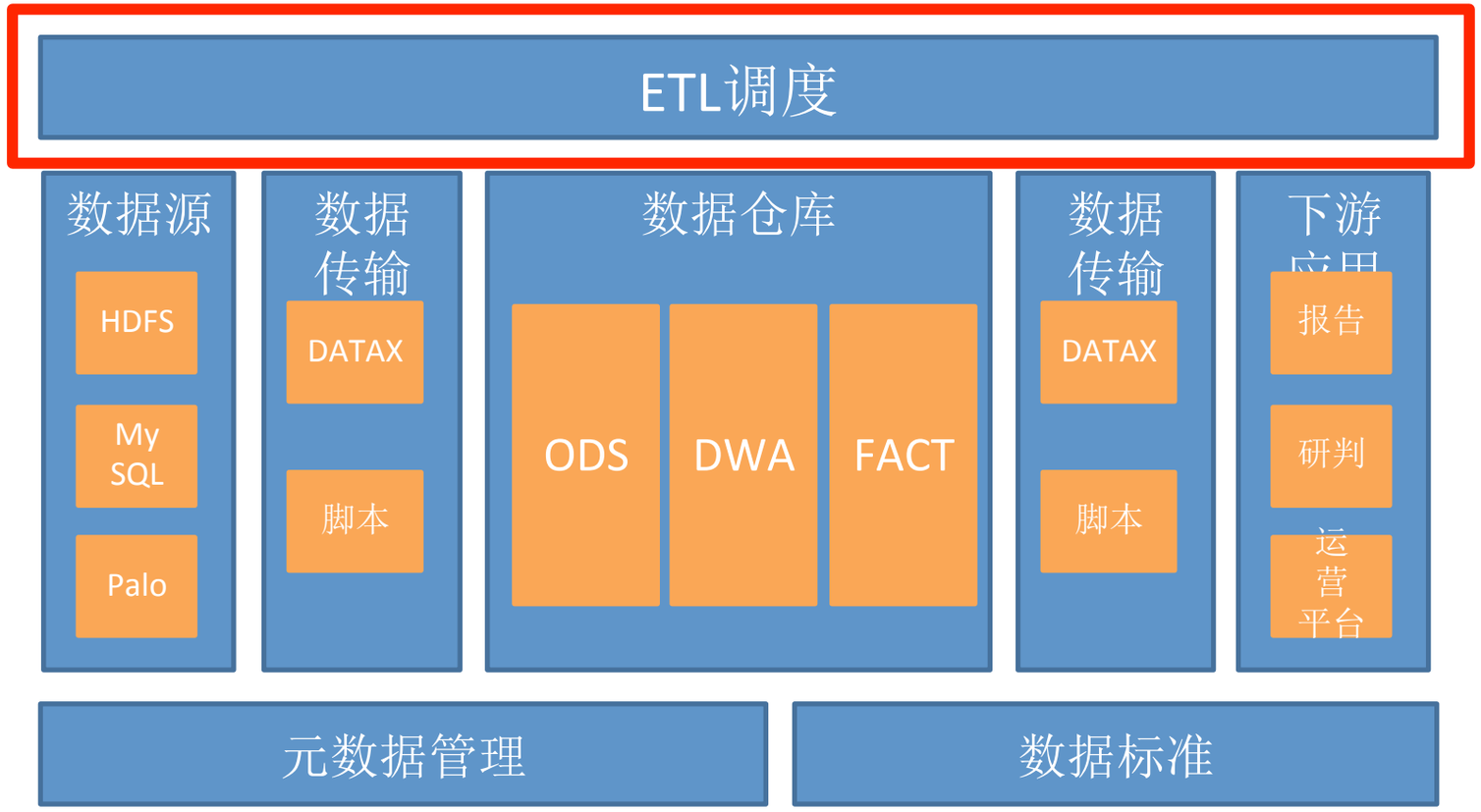
```
+-----------------------+--------------------+------+-----+---------+----------------+
| Field                 | Type               | Null | Key | Default | Extra          |
+-----------------------+--------------------+------+-----+---------+----------------+
| id                    | bigint(20)         | NO   | PRI | NULL    | auto_increment |
| citycode              | int(10) unsigned   | NO   | MUL | 0       |                |
| road_type             | tinyint(8) unsigned| NO   |     | 0       |                |
| district_type         | int(11) unsigned   | NO   |     | 0       |                |
| datatime              | bigint(20)         | NO   |     | 0       |                |
| congest_interval_info | text               | NO   |     | NULL    |                |
| congest_period_info   | text               | NO   |     | NULL    |                |
| congest_hour_info     | text               | NO   |     | NULL    |                |
| curve                 | text               | NO   |     | NULL    |                |
| curve30               | text               | NO   |     | NULL    |                |
| congest_road_num_curve| text               | NO   |     | NULL    |                |
| max_day_total_road    | int(11) unsigned   | NO   |     | 0       |                |
| max_time_period       | varchar(128)       | NO   |     | 0       |                |
| time_road_count_avg   | float(9,4)         | NO   |     | 0.0000  |                |
+-----------------------+--------------------+------+-----+---------+----------------+
```

三 仓库架构


ETL调度
数据源
HDFS
MySQL
Palo
数据传输
DATAX
脚本
数据仓库
ODS
DWA
FACT
数据传输
DATAX
脚本
下游应用
报告
研判
运营平台
元数据管理
数据标准

三 仓库架构

ETL调度模快



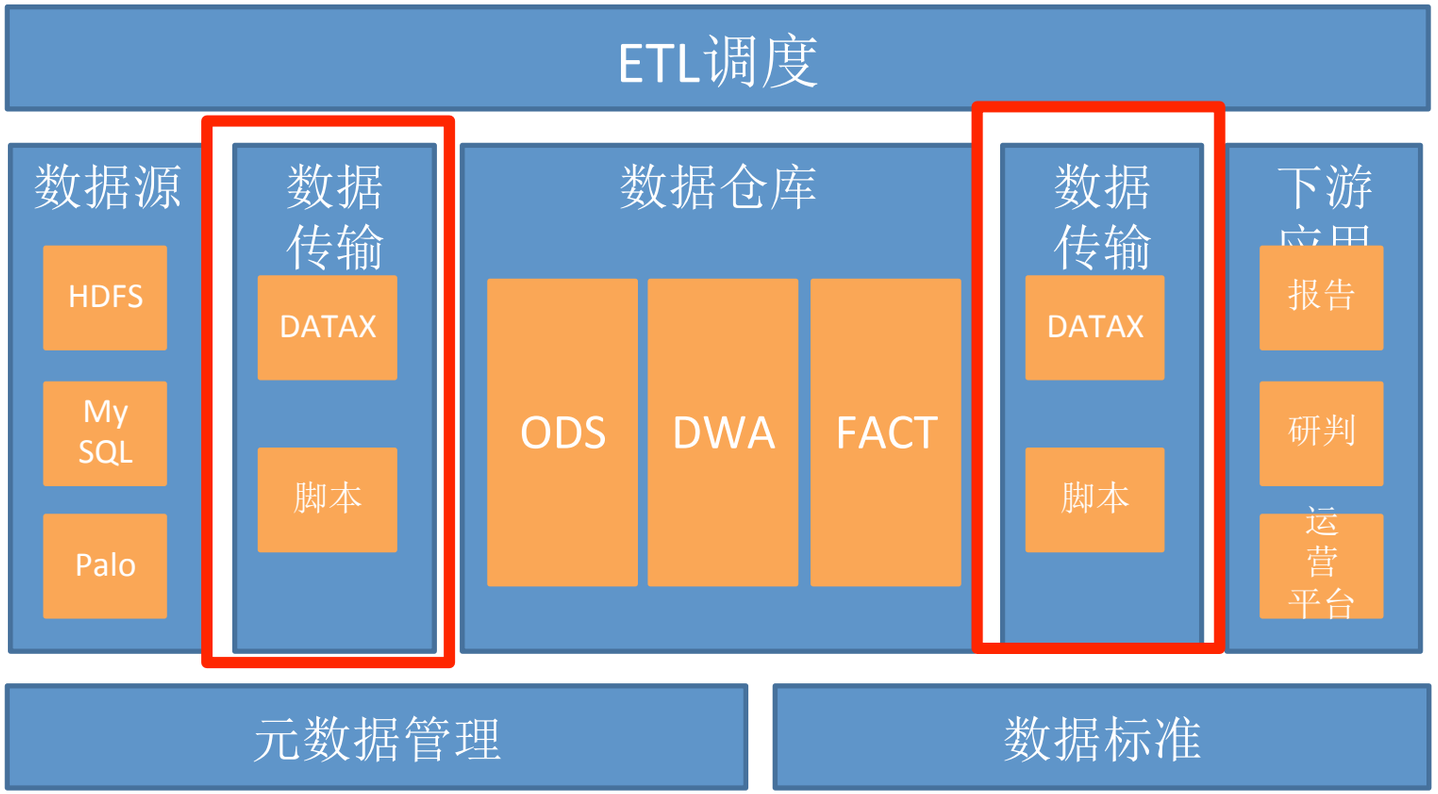
调度工具为公司提供的bigdata平台调度模块

主要提供的功能包括：

- 依赖关系配置和检查
- 任务启动
- 失败重试和报警
- 调度频度配置

三 仓库架构

数据传输模块

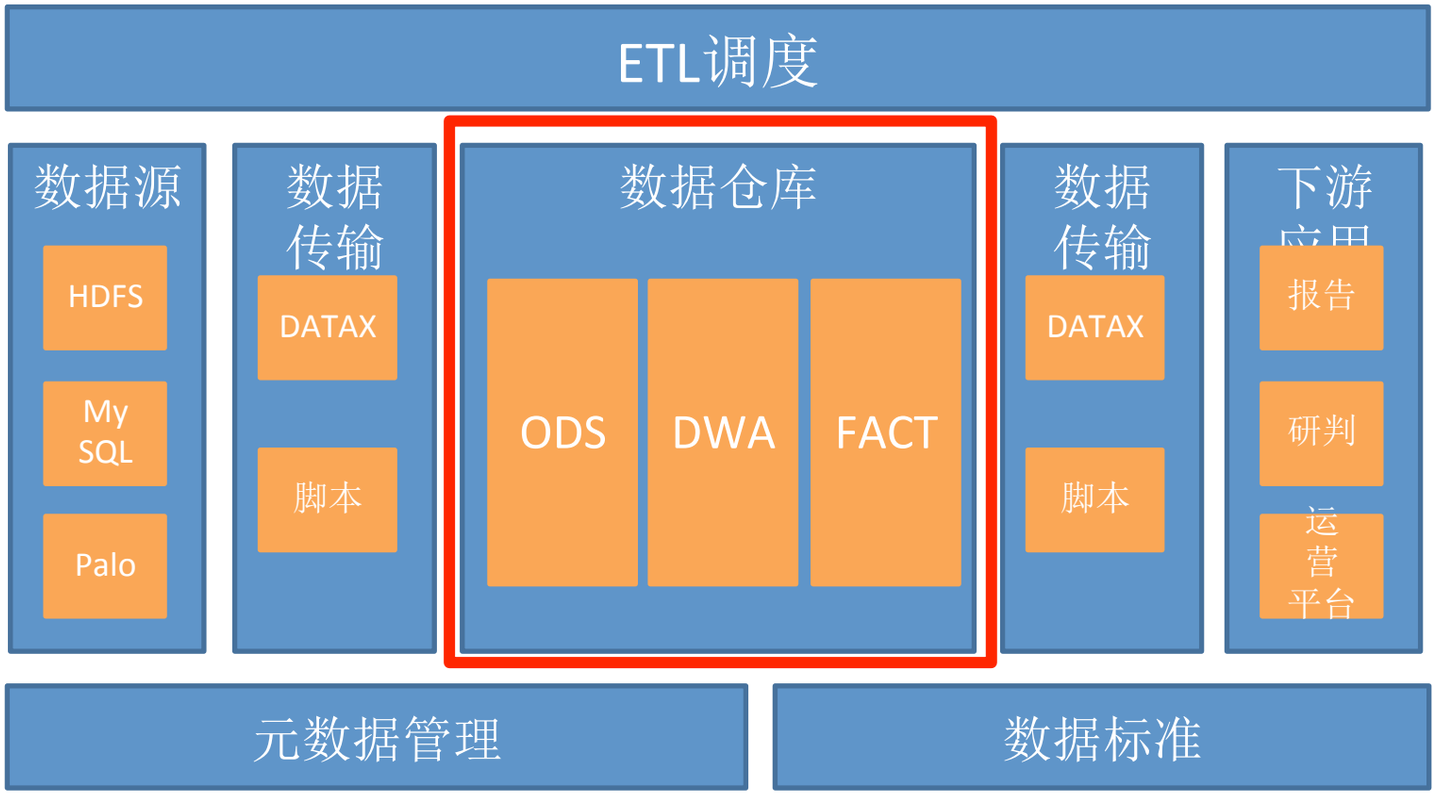


数据传输模块主要负责数据传输，采用不落地的方式。

目前采用的数据传输模块使用datax和脚本两种方式。

三 仓库架构

数据仓库模块



- 数据仓库采用yq01-Heng集群的存储，Quota为880T。目前研判平台占用350T左右。
- 数据仓库中数据计算重要采用HQL+Streaming的方式处理。

四 元数据管理_两个平台——udw平台

UDW平台（http://udw.baidu.com/nimitz/dataTree/tree）

UDW平台提供表创建和查看的基本功能，且可以保证平台上查看的schema与真实数据表一致

字段名	字段类型	字段说明	字段例子
id	bigint	自增id	
time	string	时间切片	
citycode	bigint	城市编码	
cityname	string	城市名称	
road_type	bigint	道路类型	全部；高速；环路快速路；主干路；次干路；支路；其他
district_type	bigint	城区类型	主城区；全市
yongdu_index	double	拥堵指数	
road_network_speed	double	路网的平均速度	
yongdu_length_4	double	对应拥堵等级4	
yongdu_length_3	double	对应拥堵等级3	
yongdu_length_2	double	对应拥堵等级2	
event_day	string	日期	分区

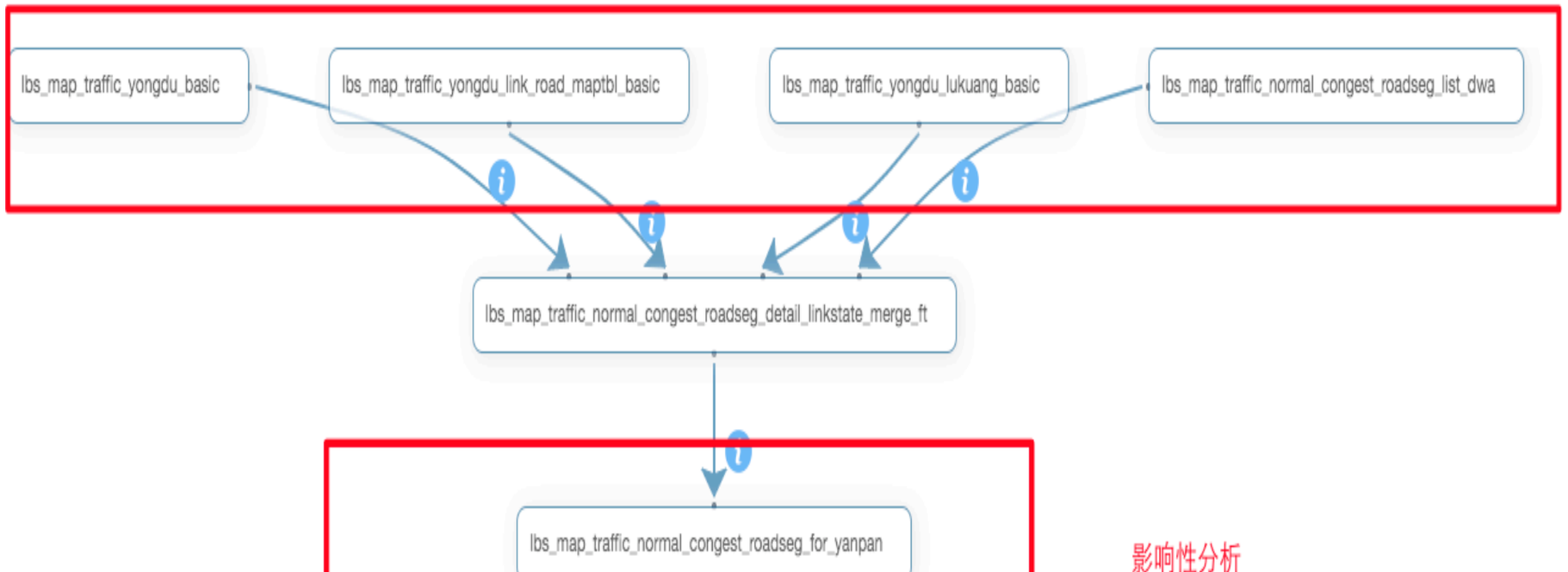
```
[lbs-di@magi]> desc lbs_map_traffic_yanpan_main_basic;
Executing command:  desc lbs_map_traffic_yanpan_main_basic
Query: 3 Time taken: 1.052 seconds
QUERY SUCCEEDED, FETCHING RESULTS...
================================================================
# col_name          # data_type      # comment
id                  bigint          自增id
time                string          时间切片
citycode            bigint          城市编码
cityname            string          城市名称
road_type           bigint          道路类型
district_type       bigint          城区类型
yongdu_index        double          拥堵指数
road_network_speed                  double          路网的平均速度
yongdu_length_4 double              对应拥堵等级4
yongdu_length_3 double              对应拥堵等级3
yongdu_length_2 double              对应拥堵等级2
event_day           string          日期
12 rows fetched.
================================================================
```

四 元数据管理——bigdata平台

bigdata平台（http://bigdata.baidu.com/ls/?grtab=gr&tab=gr）

提供血缘分析和依赖分析服务，通过血缘分析和影响性分析，快速定位数据结构或者内容改动造成的影响。

我们的作业采用作业组与目标表同名的方式，通过数据组关系即可反映数据表关系



五 命名规范

作业组

一

作

表名相

各

全部 | 已启用调度 | 未启用调度 | lbs_map_traffic | 申请权限 | 删除 | 作业交接 | normal_congest | 高级搜索

作业组名	创建者 [用户组]	调度策略	操作	启用调度
lbs_map_traffic_***normal_congest***_roadseg_for_yanpan	wanghao46 [lbs-jiaotong]	例行：0 2 1 * *	编辑 手动执行 执行记录	开
lbs_map_traffic_***normal_congest***_roadseg_detail_linkstate_merge_ft	wanghao46 [lbs-jiaotong]	例行：0 2 1 * *	编辑 手动执行 执行记录	开
lbs_map_traffic_***normal_congest***_roadseg_detail_curve_merge_fact	wanghao46 [lbs-jiaotong]	例行：0 2 1 * *	编辑 手动执行 执行记录	开
lbs_map_traffic_***normal_congest***_roadseg_list_dwa	wanghao46 [lbs-jiaotong]	例行：0 4 1 * *	编辑 手动执行 执行记录	开
lbs_map_traffic_yanpan_***normal_congest***_roadseg_detail_dwa	wanghao46 [lbs-jiaotong]	例行：0 2 1 * *	编辑 手动执行 执行记录	开

每页 10 条 | 共 5 条记录

第一页 上一页 1 下一页 最后一页

目录

根据实际需求提取相关数据

- 北京绿波带数据提取
- 成都某些路段拥堵指数平均速度获取
- 北京绿波带优化前后结果对比及排序

- 具体工作
  - 了解业务逻辑，熟悉Hadoop平台，写Hive；
  - 相关工作写成比较通用Python脚本，比如结果表生成、文件读写、结果排序。

二 Midmif路网数据重新划分

- 背景
  - 原路网midmif文件之间交错严重，比如原济南数据中8%属于泰安聊城等，而24%的实际济南数据存在与德州滨州文件下；
  - 区域划分过老，11年之前的划分标准。
- 大致流程
  - ① midmif文件的解析与配对
  - ② midmif文件的按实际按格式重新写入。
  - ③ 根据实际行政划分各种调整。

Thank You!

百度地图 开放平台