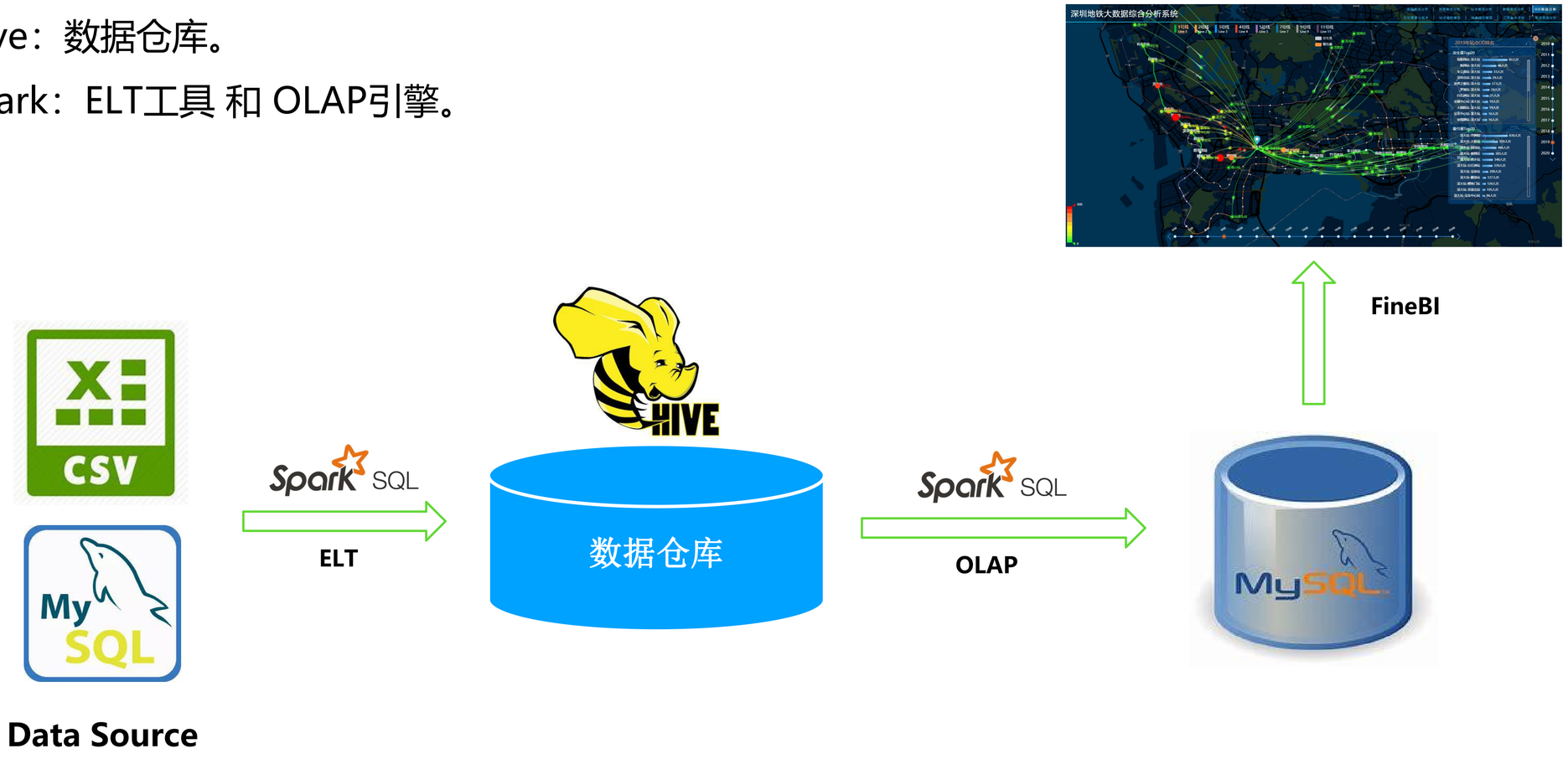
# 案例：交通大数据分析

山丘

# 项目架构

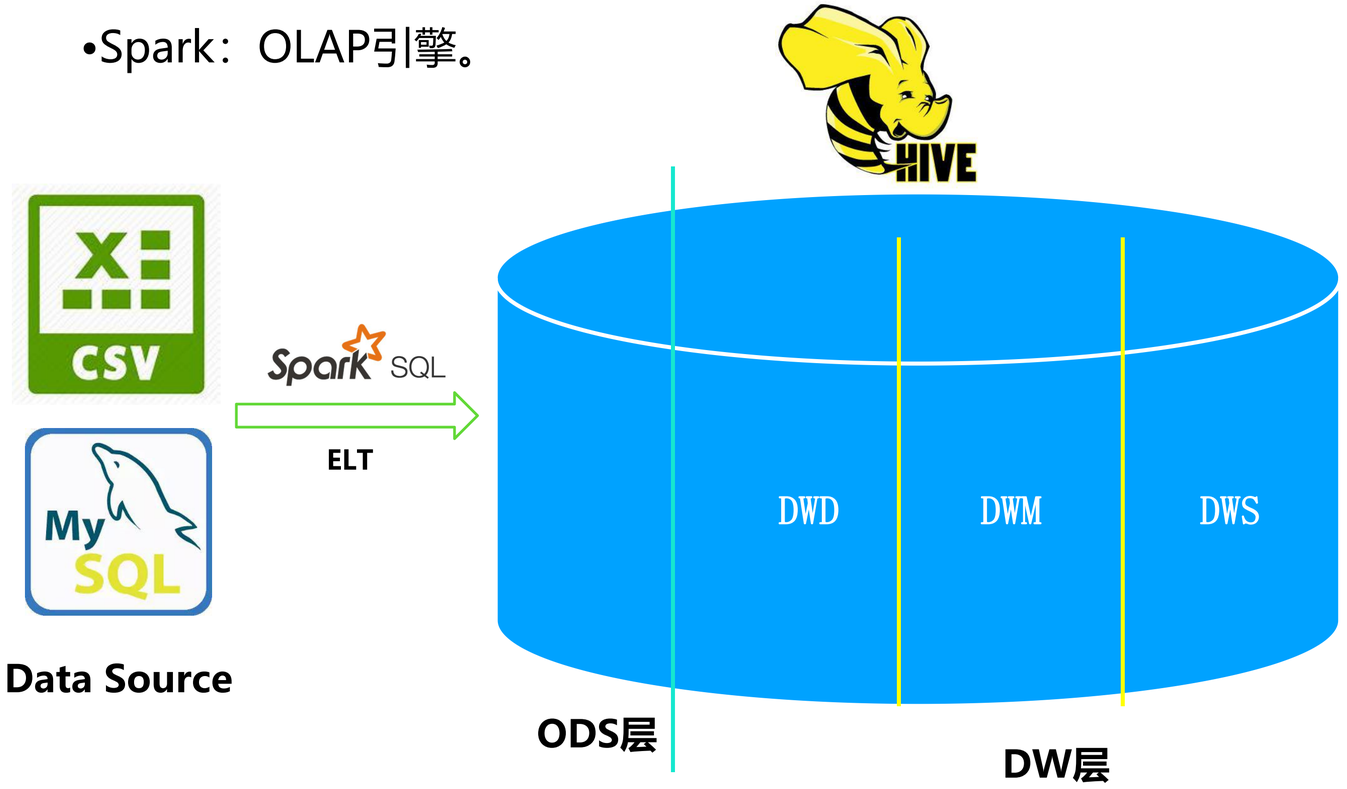
# 项目架构

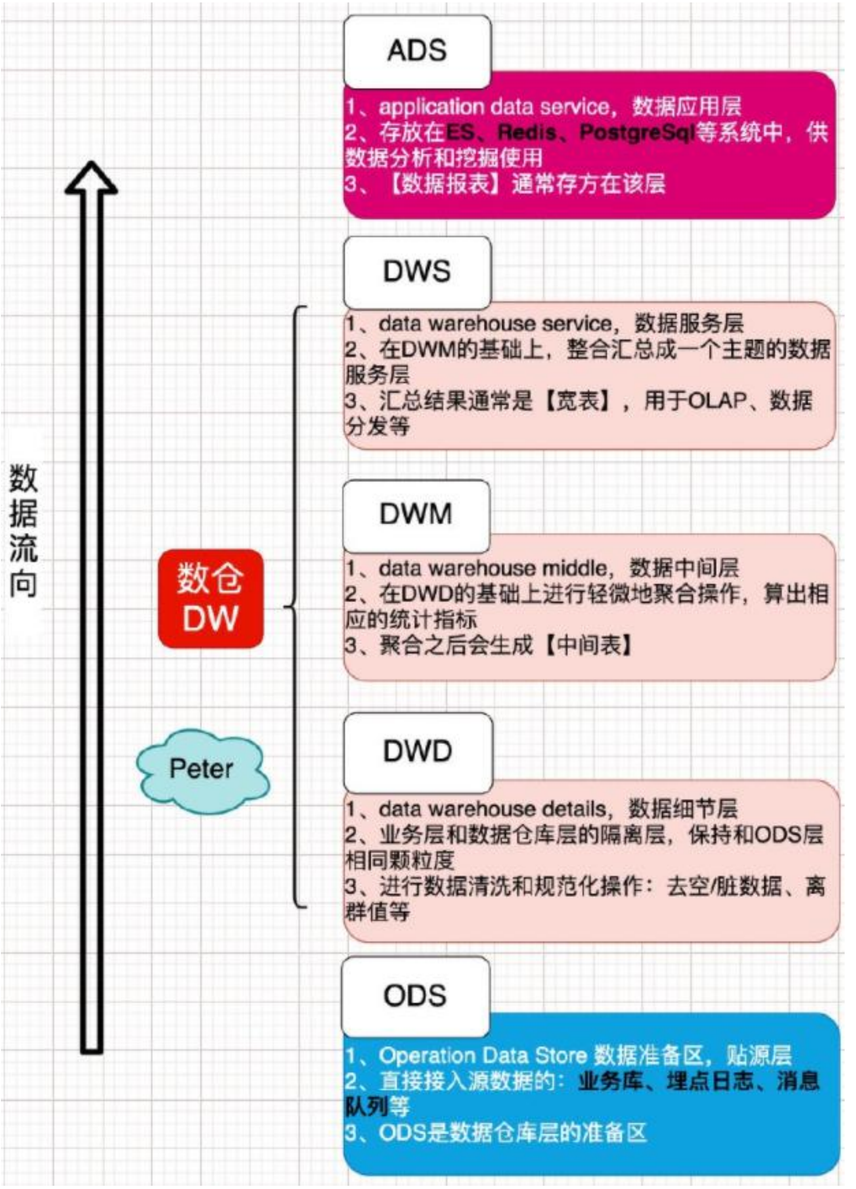
- 项目架构
  - Hive：数据仓库。
  - Spark：ELT工具 和 OLAP引擎。



# 项目架构

- 项目架构
  - Hive：数据仓库。
  - Spark：OLAP引擎。

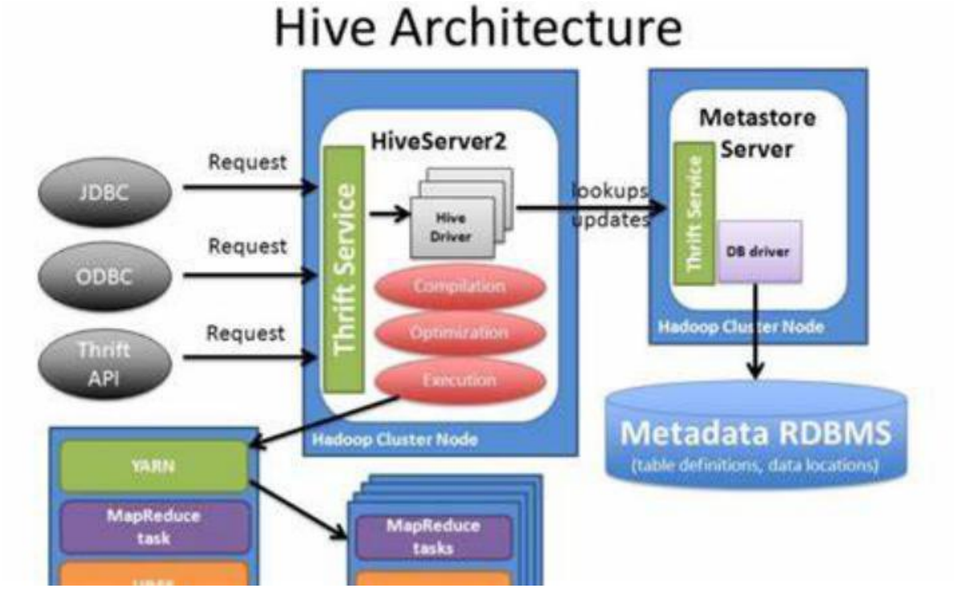


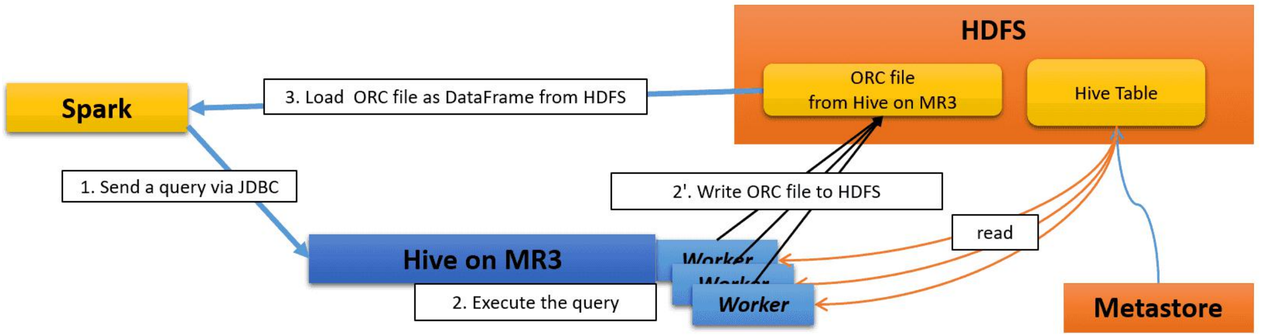




数据运营层ODS：Operation Data Store 数据准备区，也称为贴源层。

# 项目架构

- Spark整合Hive
  - 将配置文件hive-site.xml拷贝到 $SPAR_HOME/jars/ 目录下。



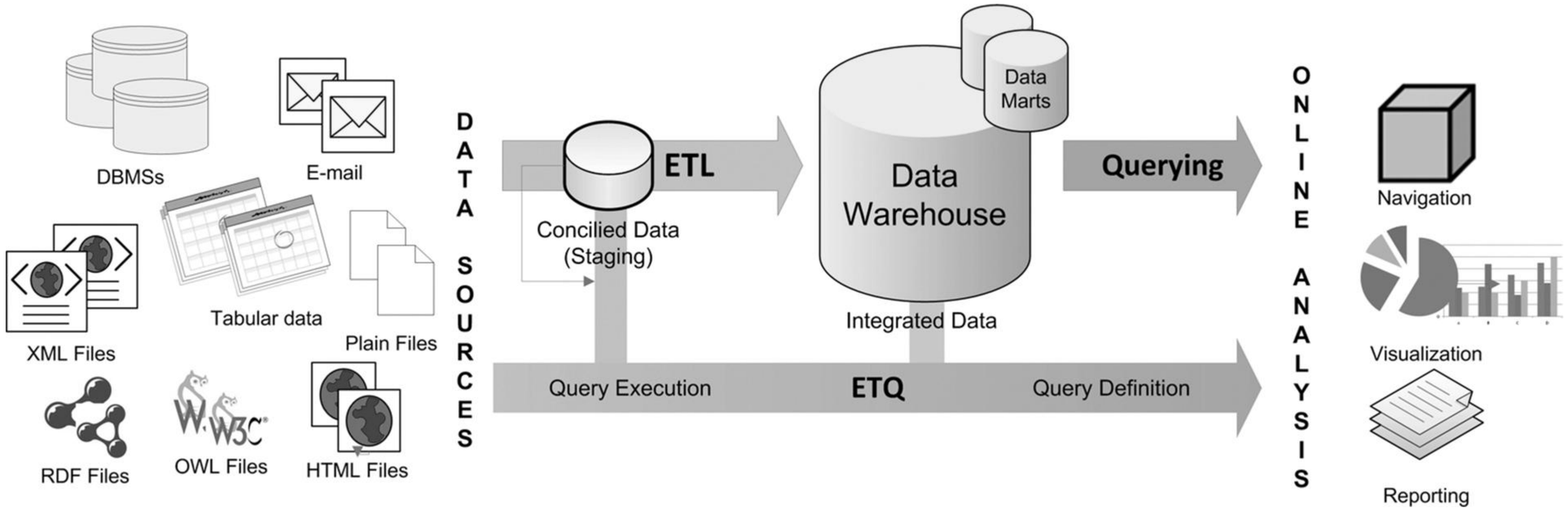




Spark访问Hive的过程

# 数据仓库与OLAP

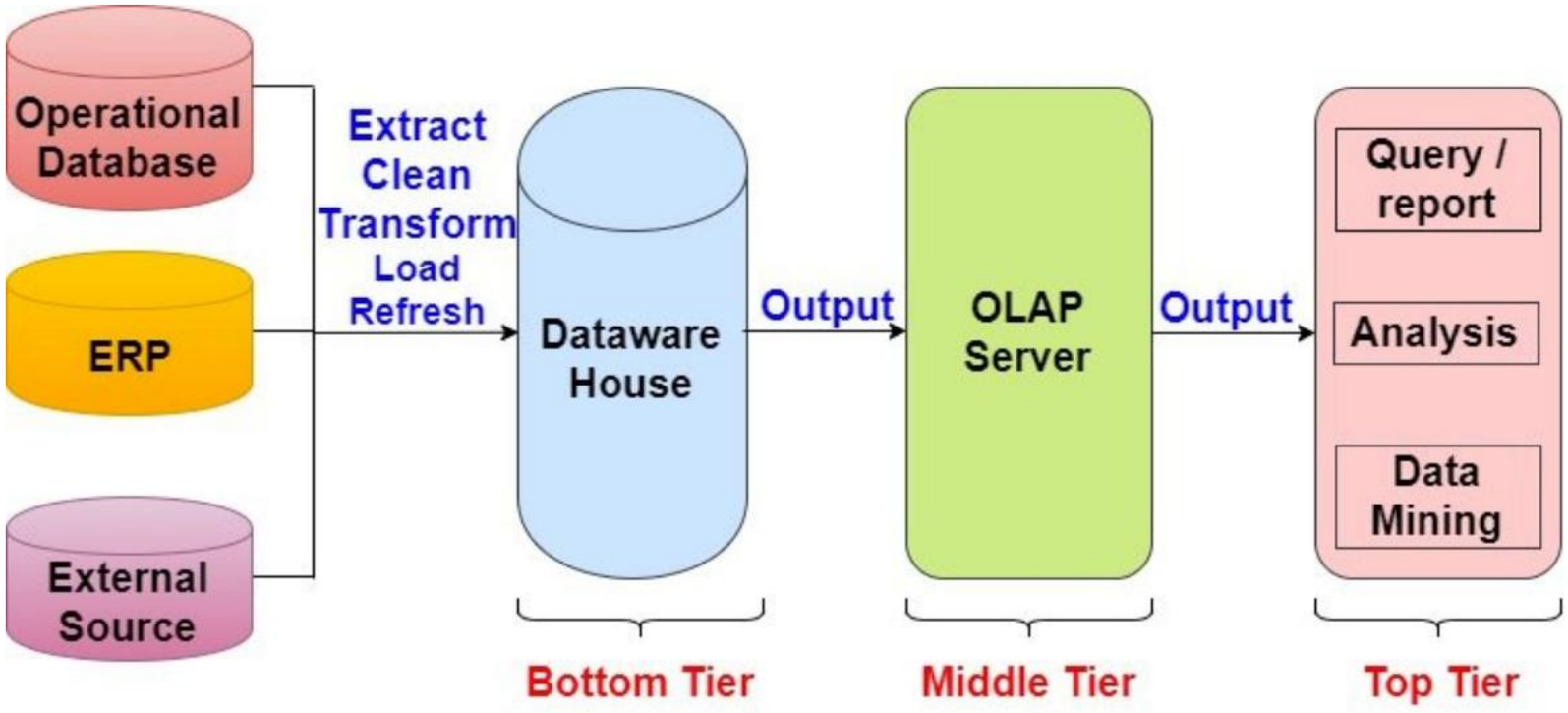
# 数据仓库与OLAP

- 数据仓库
  - 数据仓库体系结构图。



# 数据仓库与OLAP

- 数据仓库
  - 数据仓库与OLAP(联机分析处理)。



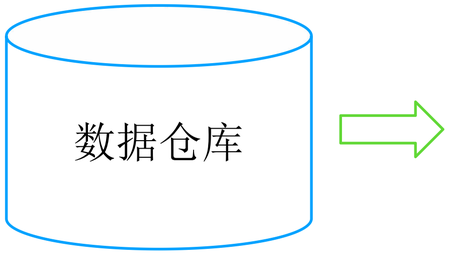
# 数据仓库与OLAP

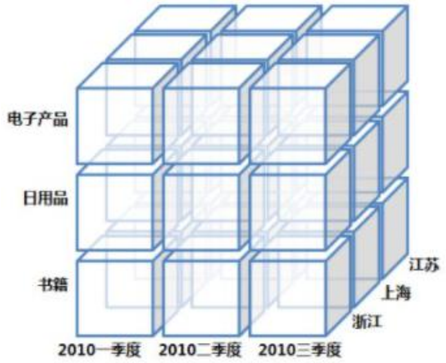
- OLTP与OLAP
  - OLTP：联机事务处理。
  - OLAP：联机分析处理。

| | OLTP System<br>Online Transaction Processing | OLAP System<br>Online Analytical Processing |
|---|---|---|
| 业务目的 | 处理业务，如订单、合同等 | 业务支持决策 |
| 面向对象 | 业务处理人员 | 分析决策人员 |
| 主要工作负载 | 增、删、改 | 查询 |
| 主要衡量指标 | 事务吞吐量 | 查询响应速度（QPS） |
| 数据库设计 | 3NF 或 BCNF | 星型/雪花模型 |

# 数据仓库与OLAP

- 数据仓库与OLAP的关系
  - 数据仓库是一个包含企业历史数据的大规模数据库，这些历史数据主要用于对企业的经营决策提供分析和支持。
  - OLAP是与数据仓库密切相关的工具产品。OLAP的源数据通常存储在数据仓库中。
  - 在OLAP系统中，客户能够以多维视觉图的方式，搜寻数据仓库中存储的数据。- 特点：多维数据分析。
  - OLAP服务工具利用多维数据集和数据聚集技术对数据仓库中的数据进行处理和汇总，用联机分析和可视化工具对这些数据进行评价，将复杂的分析查找结果快速地返回用户。




多维数据集



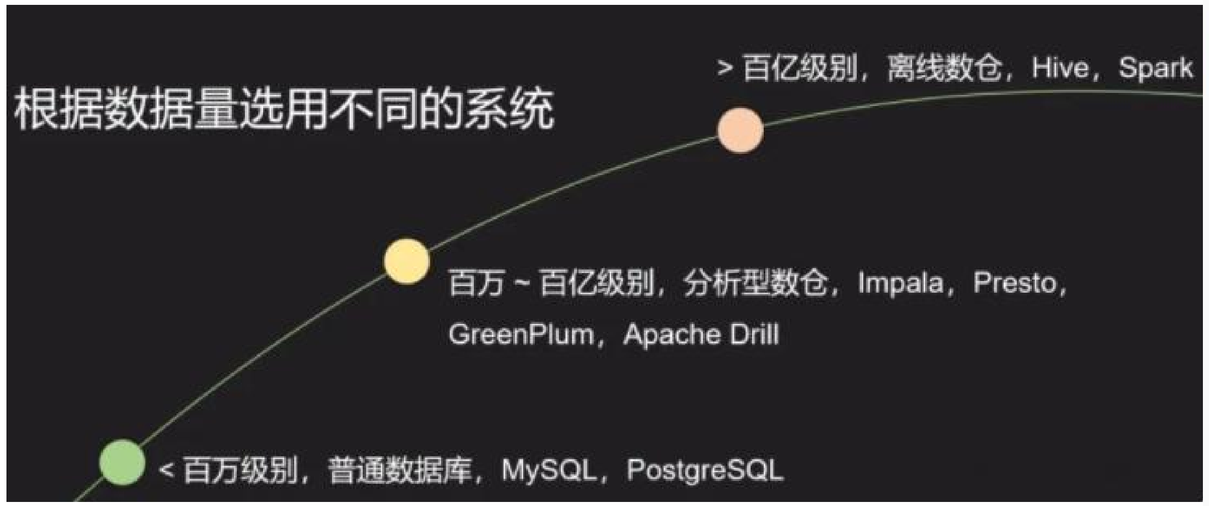


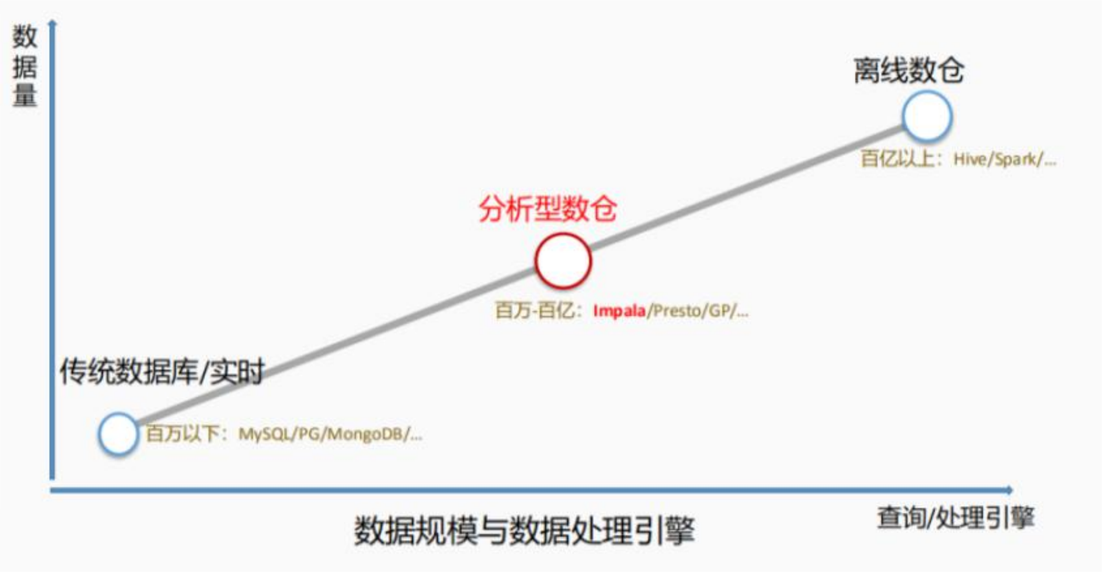
OLAP服务工具

OLAP结果

# 数据仓库与OLAP

- 有哪些类型的OLAP数仓?
  - 按数据量划分
  - 我们可以基于数据量来选择不同类型的数仓，如下图所示：

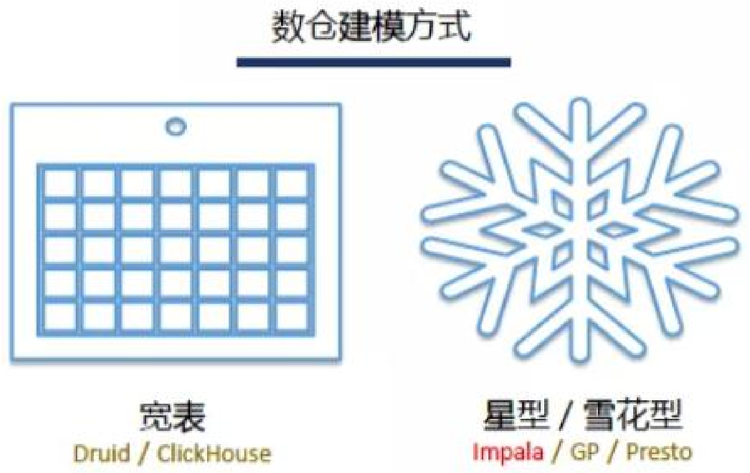


# 数据仓库与OLAP

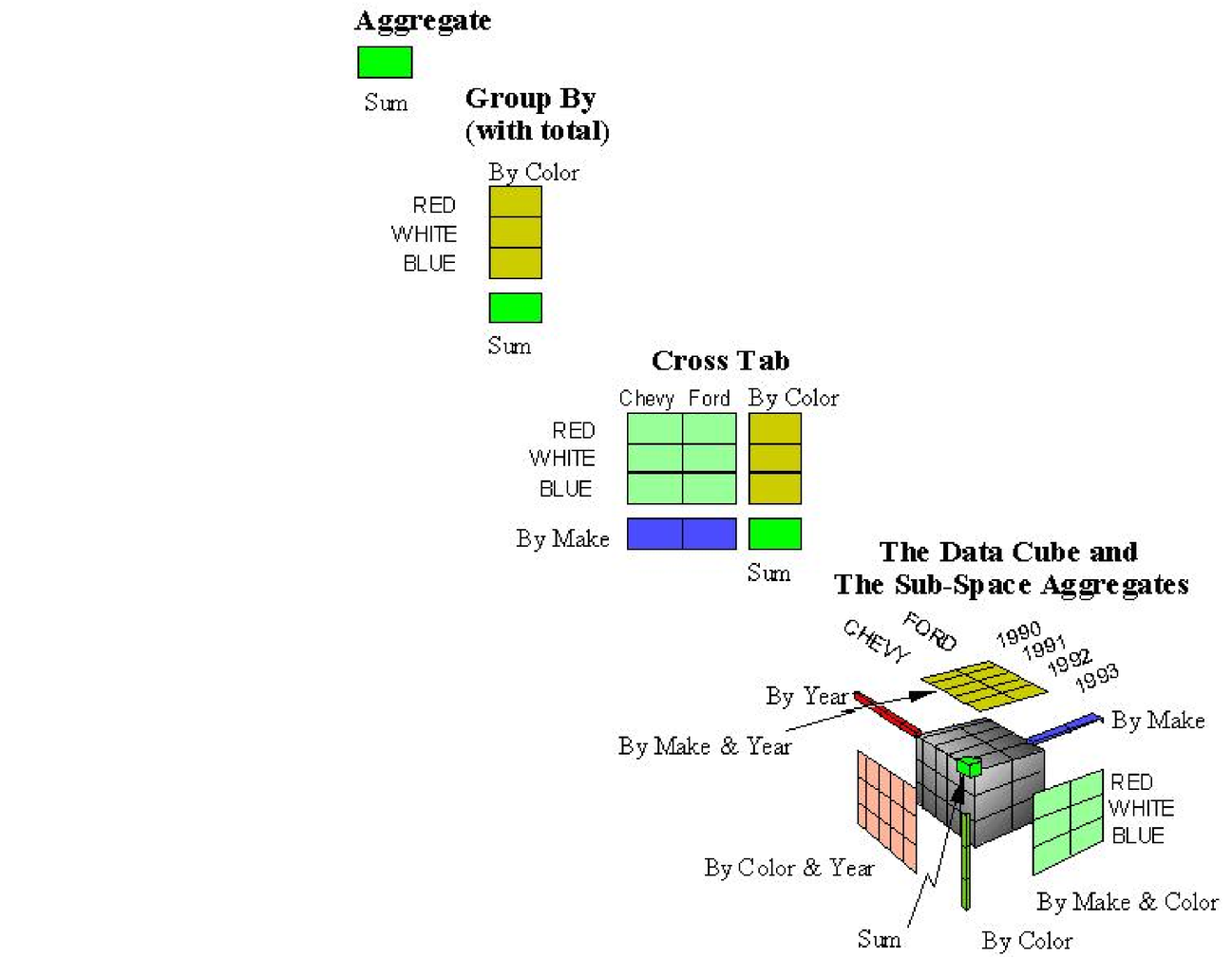
- 有哪些类型的OLAP数仓?
  - 按建模类型划分
  - 根据维基百科对OLAP的介绍，一般来说OLAP根据建模方式可分为MOLAP、ROLAP和HOLAP 3种类型
  - MOLAP：传统的数仓。大数据领域代表性开源产品是 Kylin，支持在百亿规模的数据集上进行亚秒级查询。
  - ROLAP：与MOLAP相反，ROLAP无需预计算，直接在构成多维数据模型的事实表和维度表上进行计算。


ROLAP



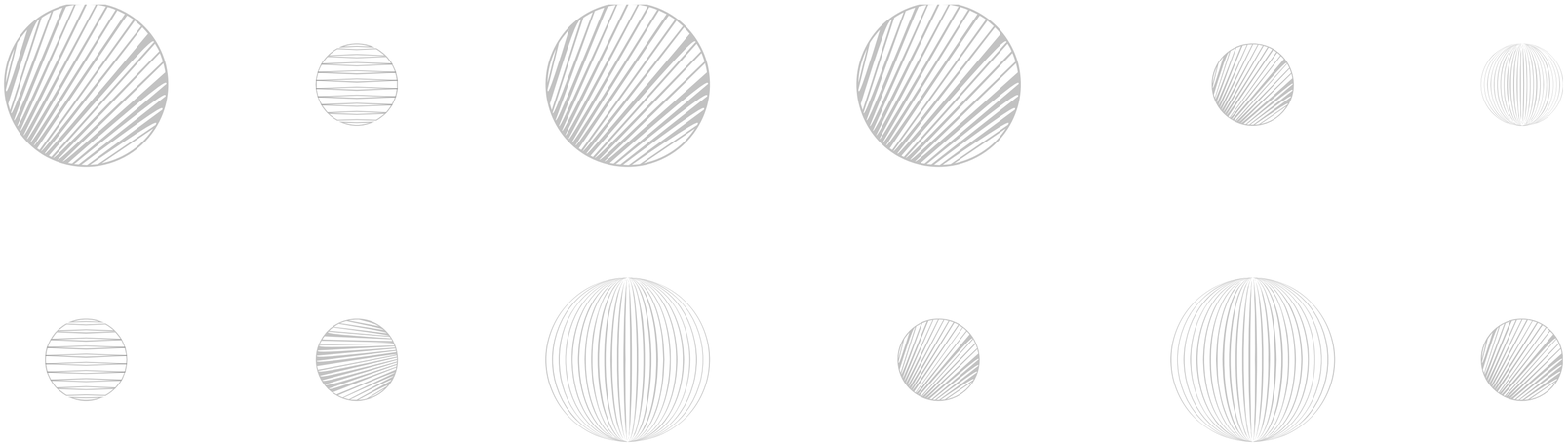
# 数据仓库与OLAP

- 多维数据集
  - 数据立方体




数据仓库的构建包括一系列的数据预处理过程：

- 数据清理
- 数据集成
- 数据变换



THANKS