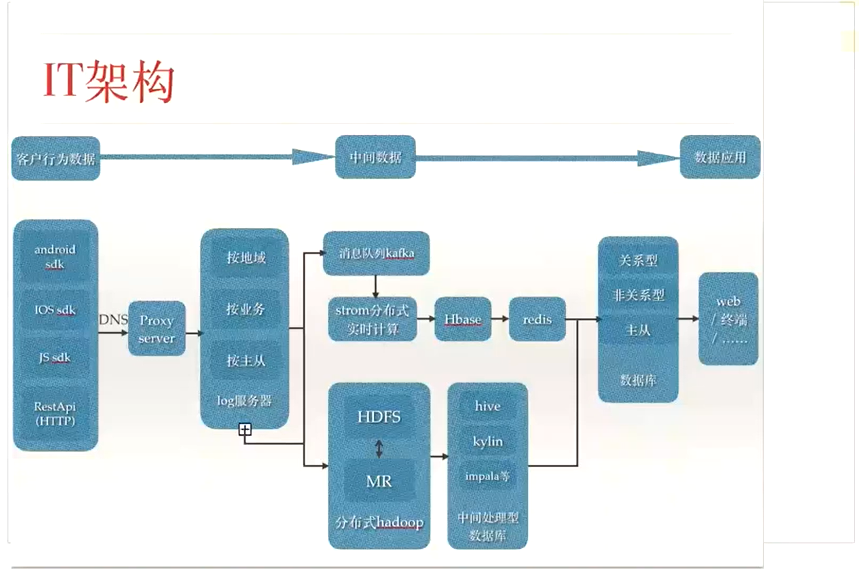
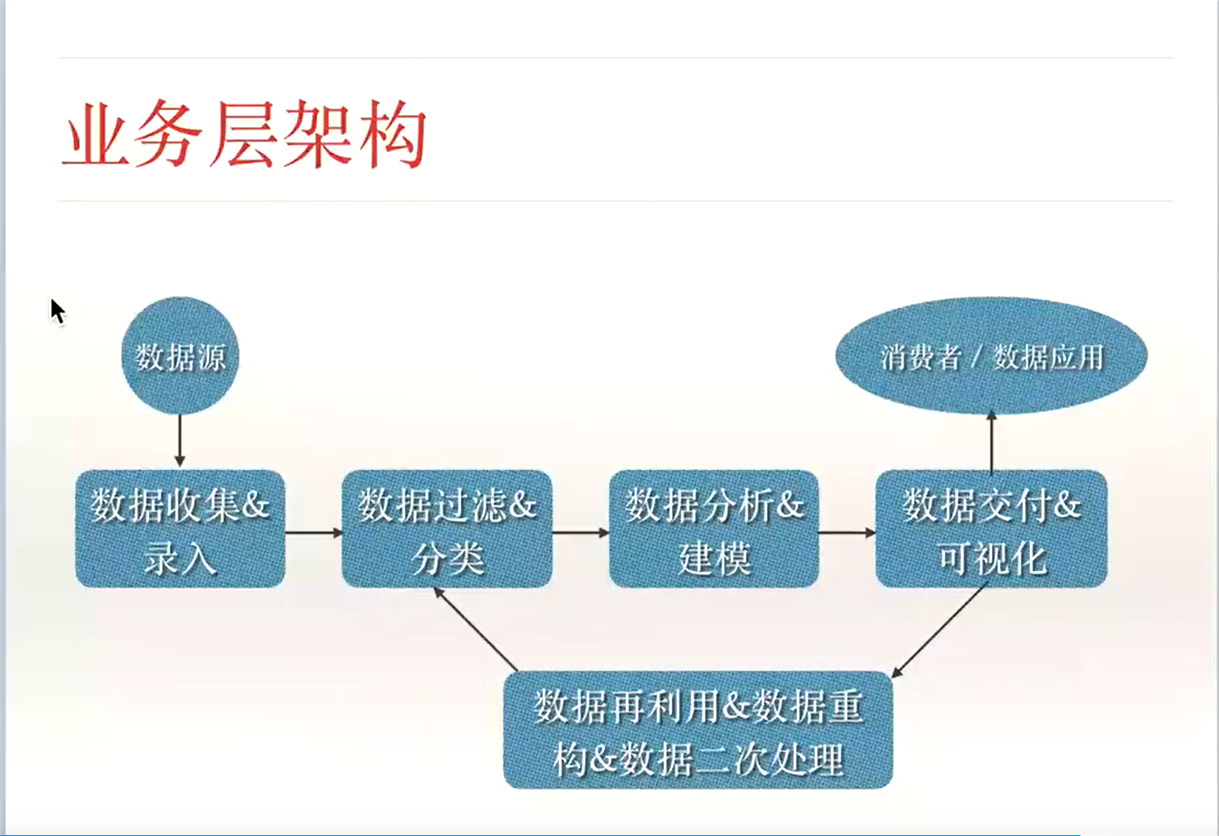
1. 完成Hadoop架构图及组件讲解





DW(data warehousing) 数据仓库

EDW(enterprise data warehousing) 企业数据仓库

BI(business interlligence) 商业智能

ODS(Operational Data Store): 操作数据存储

DM(Data Market): 数据集市

ETL(Extract Transformation and Load)

获取——转换——加载

事实表

维护表

清洗

合并

复制

ETL系统

前端

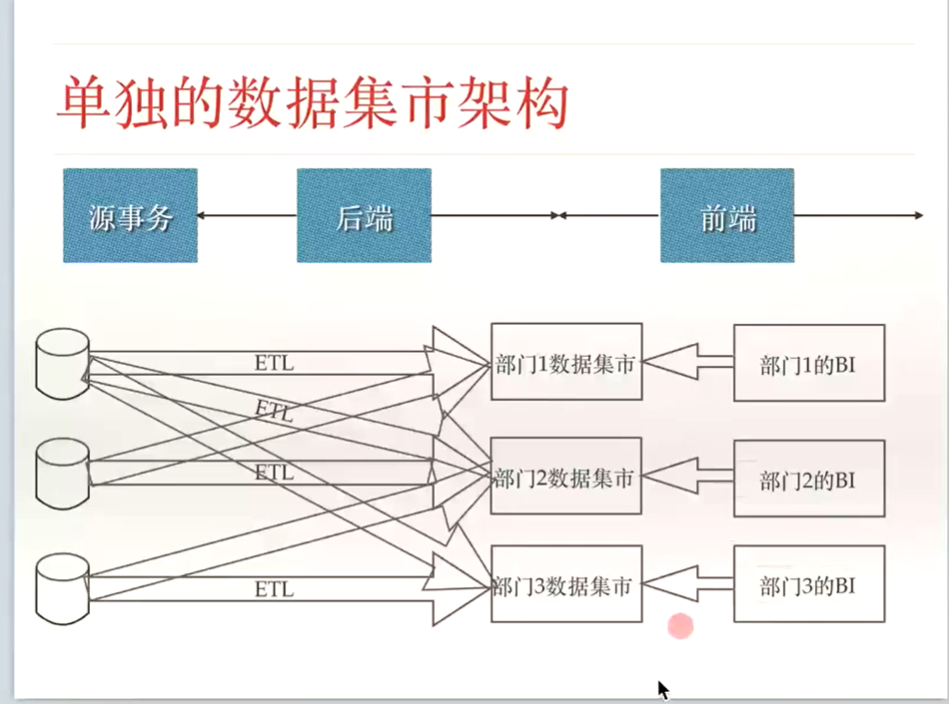
后端

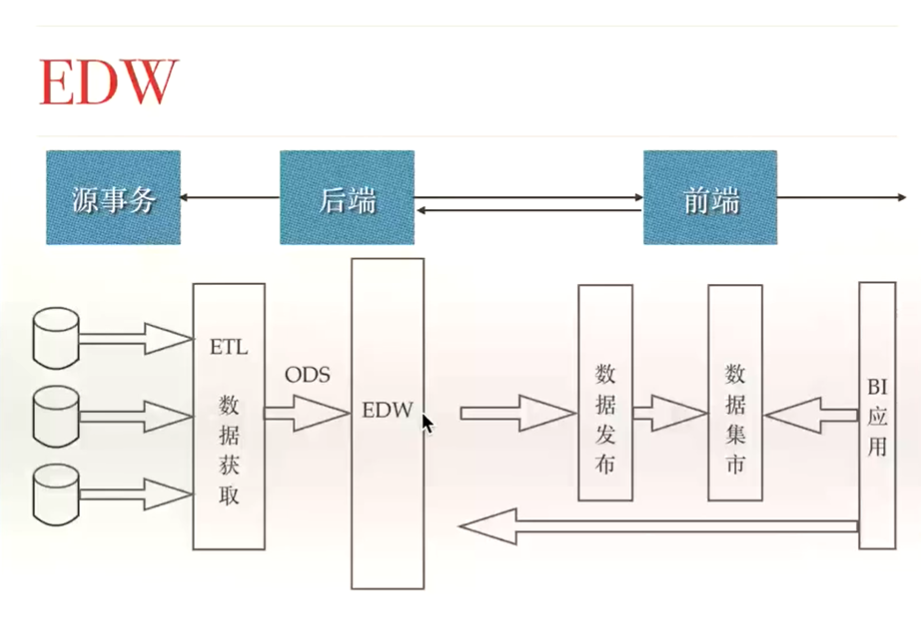
源事务

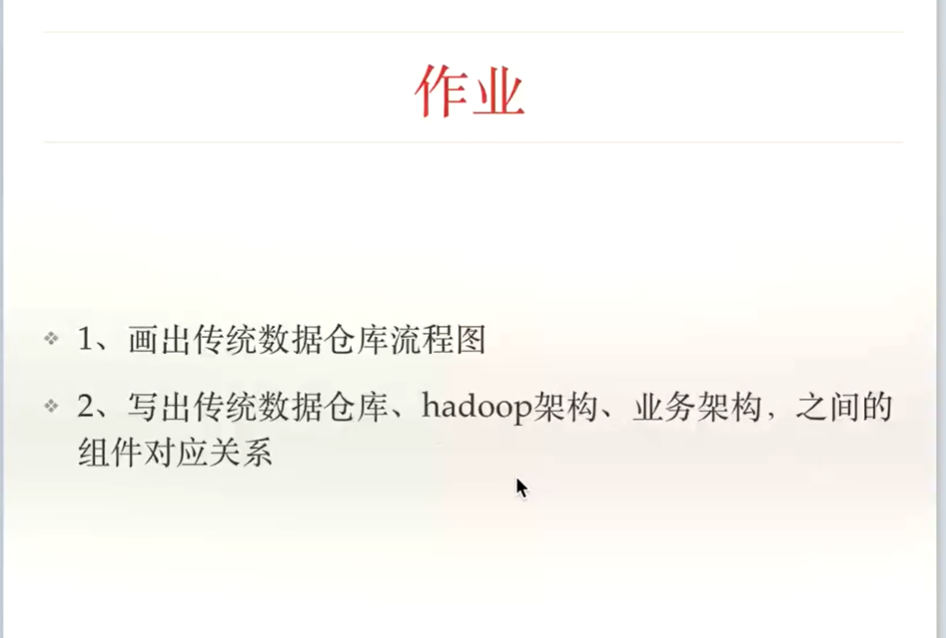
BI应用

**数**据仓库

ODS







## 流处理 streaming

原生处理流：意味着所有输入的记录一旦到达即会一个接着一个进行处理。一个消息传递过来，立刻处理。

Storm

微批处理：把输入的数据按照某种预先定义的时间间隔（典型的是几秒钟）分成短小的批量数据，流经流处理系统。

Spark Streaming

交互式查询 交互式sql

流处理与批处理之间

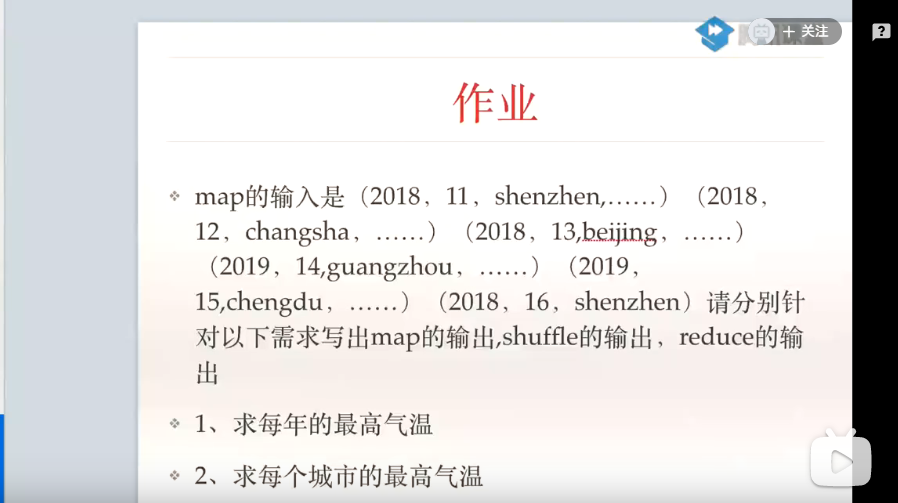
1. 分布式数据仓库 MPP
2. Sql on Hadoop -> hive

批处理batch

批处理方式：1.传统ETL -> MPP技术 2. MR

Map Reduce

Input->map->shuffle->reduce->output



DB设计

测试关注：DB设计是否合理，是否可涵盖业务场景

1. 元数据
2. 维度表 事实表

元数据的概念：描述数据的数据。

元数据分类

技术类元数据：存储类型数据：表，表结构等

计算类型数据：hive的Job日志，计算任务，任务调度，日志信息等

数据质量和运维相关元数据：如人工监控，运维报警，数据质量，故障等

业务类元数据：业务类型或者场景划分贴大标签

维度

事实

维度设计

维度建模基本设计方法：

1. 选择维度，确定维度
2. 确定主维表
3. 确定从维表
4. 确定维度属性/识别元数据

事实表

|  |  |  |  |
| --- | --- | --- | --- |
| 商品ID | 订单ID | 商品售价 | 商品数量 |
| X001 | 1001 | 100 | 2 |
| X002 | 1001 | 200 | 2 |
| X001 | 1002 | 100 | 1 |

事实设计

1. 尽可能包含所有与业务过程相关的事实
2. 只选择与业务过程相关的事实
3. 同一张表中只能有一种粒度，且粒度外的其他属性字段要和粒度保持统一级
4. 单位统一，null值处理

建模思想



