数据平台：初探

郭宁@猫眼

2017/01/13

目录

- 数据平台：价值?
- 数据平台：本质?
- 数据平台：演进思路?
- 数据平台：实例

为什么需要?

演进，怎么做?

场景：数据服务能力

- **老板**：捉急呀
- **CTO**：业务势头这么猛，怎么还会捉急？
- **老板**：
  - 风口来了
  - 咱们业务发展猛，竞争对手发展的也猛，相对优势小
  - 预计有 6～18 个月的胶着期
  - 现在是关键阶段，如果决策出错，就被趋势抛弃了，捉急...
- **CTO**：
  - 提升数据能力呀
  - 对情况有准确的掌握，更大概率做出正确决策
  - 3 个月前，已经着手，构建了数据平台，形成初步数据服务能力
  - 未来 6～18 个月，根据业务需要，将进行多次迭代、完善，增强对情况的掌握

问题1：数据平台，有什么价值?

具体是指?

数据平台：价值

- **数据价值，支撑**：
    - **服务质量**：
        - 系统的监控、告警、自动降级，提升可用性；
        - 快速、准确定位服务瓶颈；
        - 提前预测服务能力（节假日）
    - **成本控制**：
        - 系统运营成本：机器、带宽的利用率，耗电量等
        - 业务运营成本：业务指标，用户停留时常、访问路径（漏斗模型）
- **利润增长**：
    - 用户体验/商业价值：用户的时间有限，如何有限时间内，尽可能多的达成交易
        - 电商的推荐
        - 电商的个性化搜索
        - 电商的广告投放
- **战略方向**：用户增长、市场突破口/竞争壁垒、业务发展的关键节点（年度计划）
- **投机/投资**：量化交易等（群体收益，忽略单次的成败）

问题2：数据平台，价值这么高，赶快做一个?

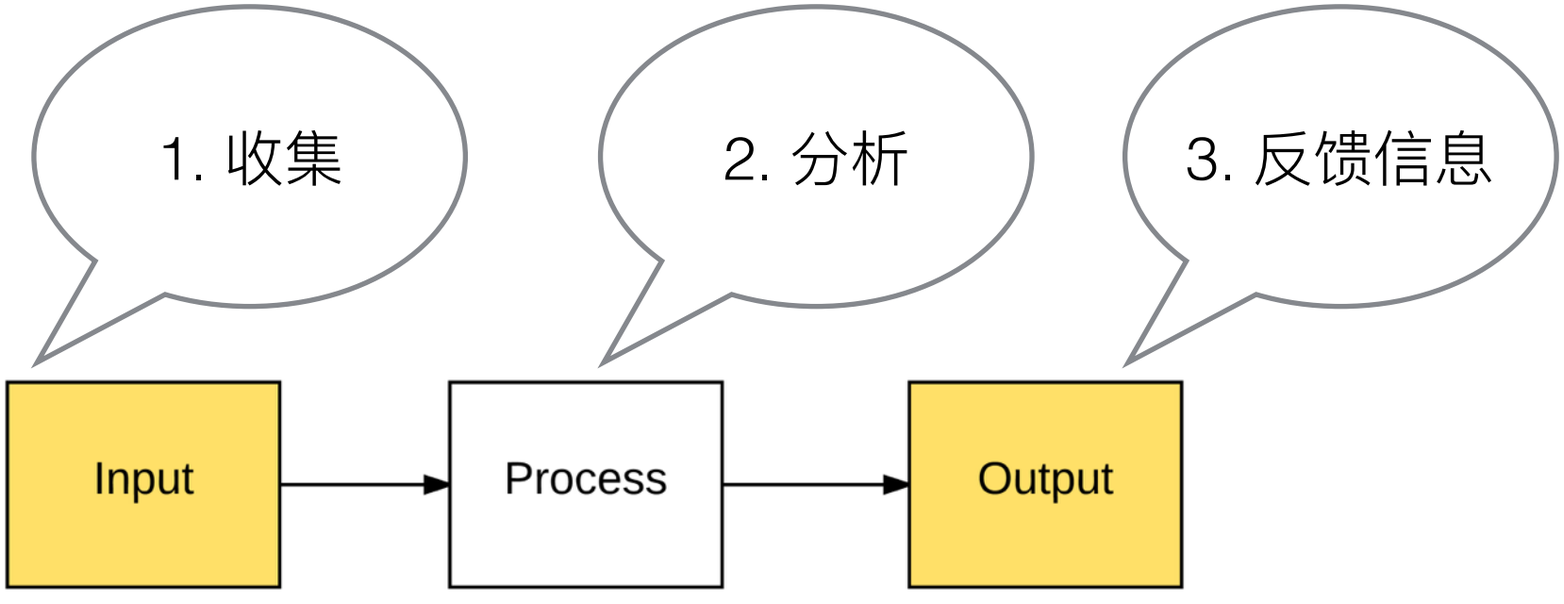
数据平台：价值

- **数据平台**，也会引入成本：
  - **人力成本**：大量的资源投入，人力。
  - **机会成本**：资源有限，投资源做「数据平台」，就势必减少其他业务上投入。
- **核心问题**：
  - **要不要做**？ trade off：收益 *vs.* 成本
  - **什么时候做**？
  - **做到什么程度**？

问题3：数据平台，核心思维?

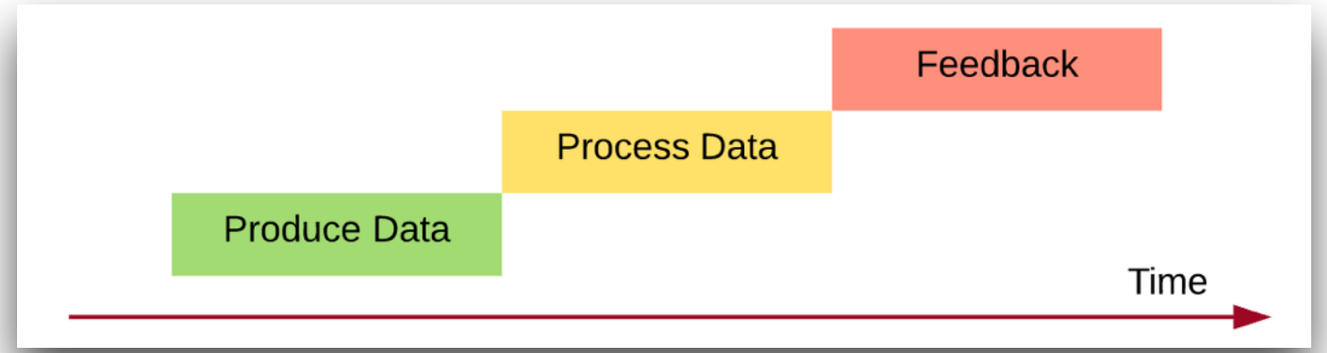
数据平台：本质

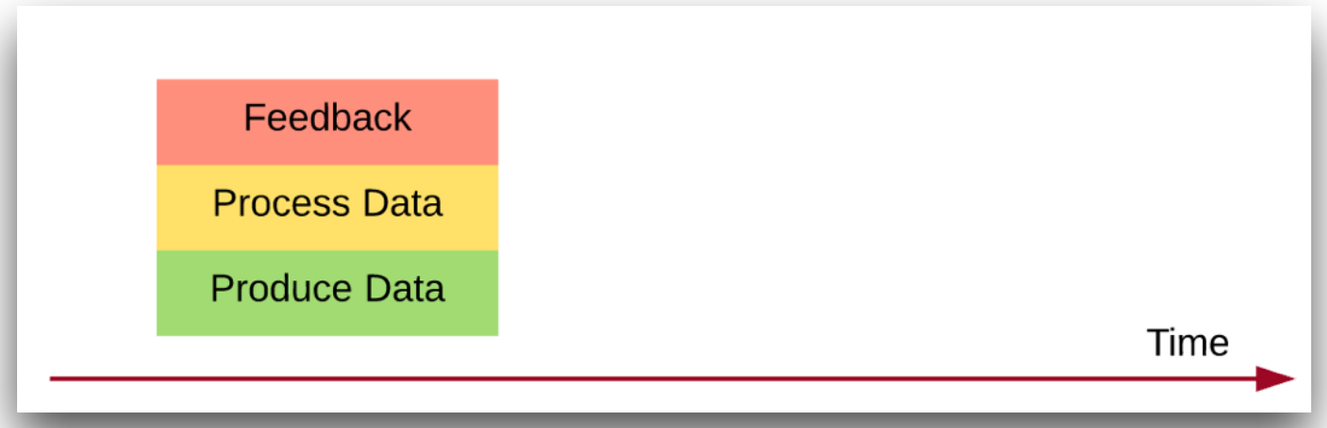
- 第一谚语：GIGO
  - Garbage In, Garbage Out
- 本质：
  - Input
  - Process
  - Output



数据平台：本质

- 时间维度，根据实时性不同，数据系统分类：
  - 批量处理：1h～10d
  - 准实时：2～10s
  - 实时：0～2s
- 备注：上述时间是指，从「产生数据」 —>「获取信息」



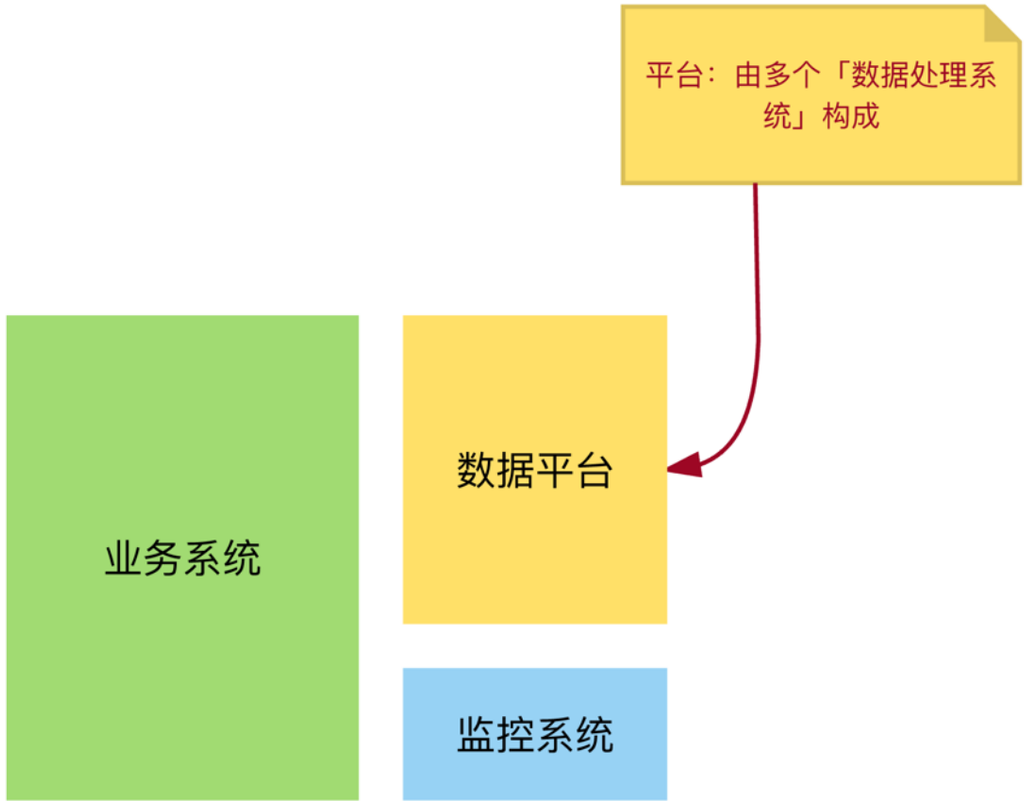


问题4：数据平台，怎么做?

数据平台：演进思路

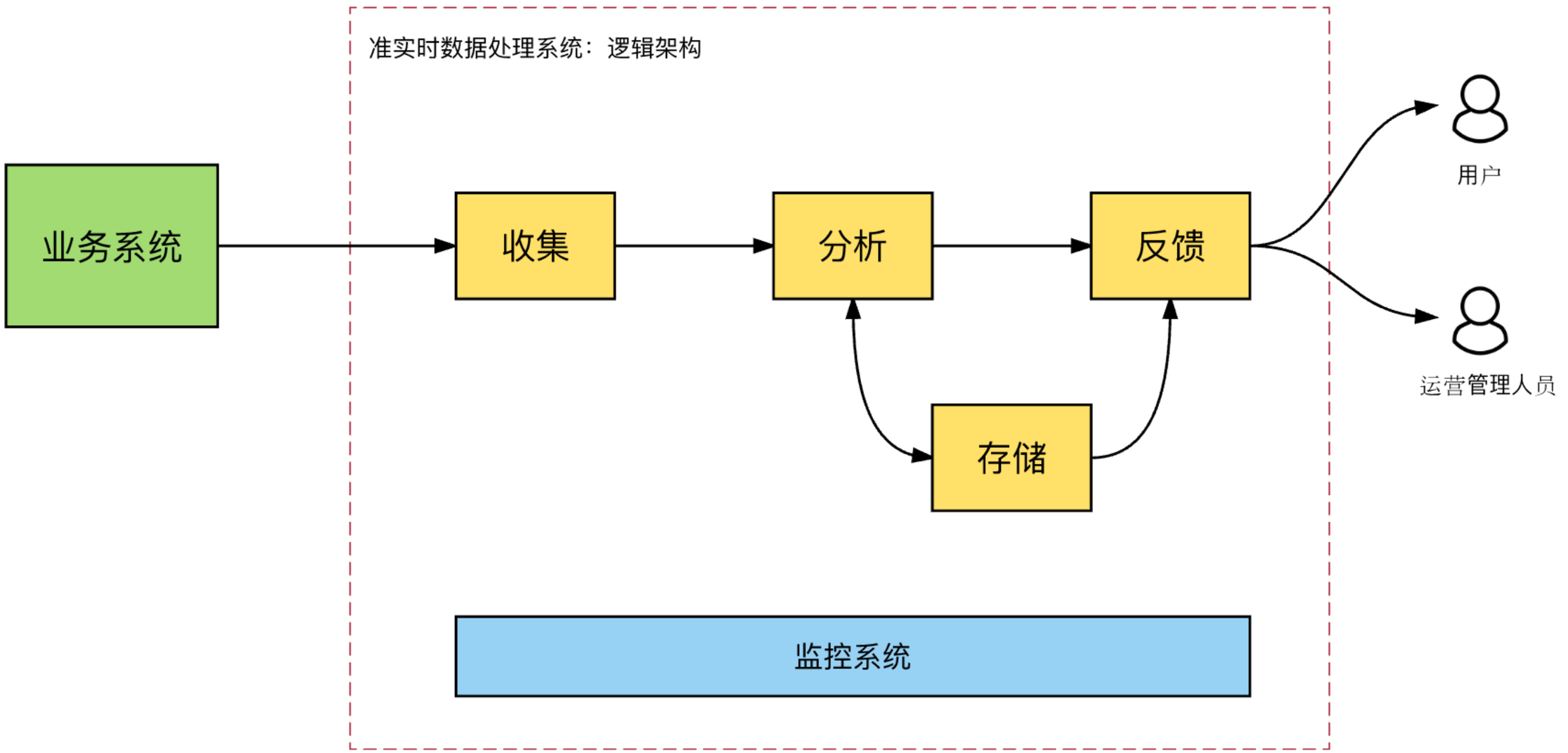
- 演进思路：
  - 自动化脚本 → 系统 → 平台

- 设计数据平台，考虑 2 方面：
  - 对内
  - 对外

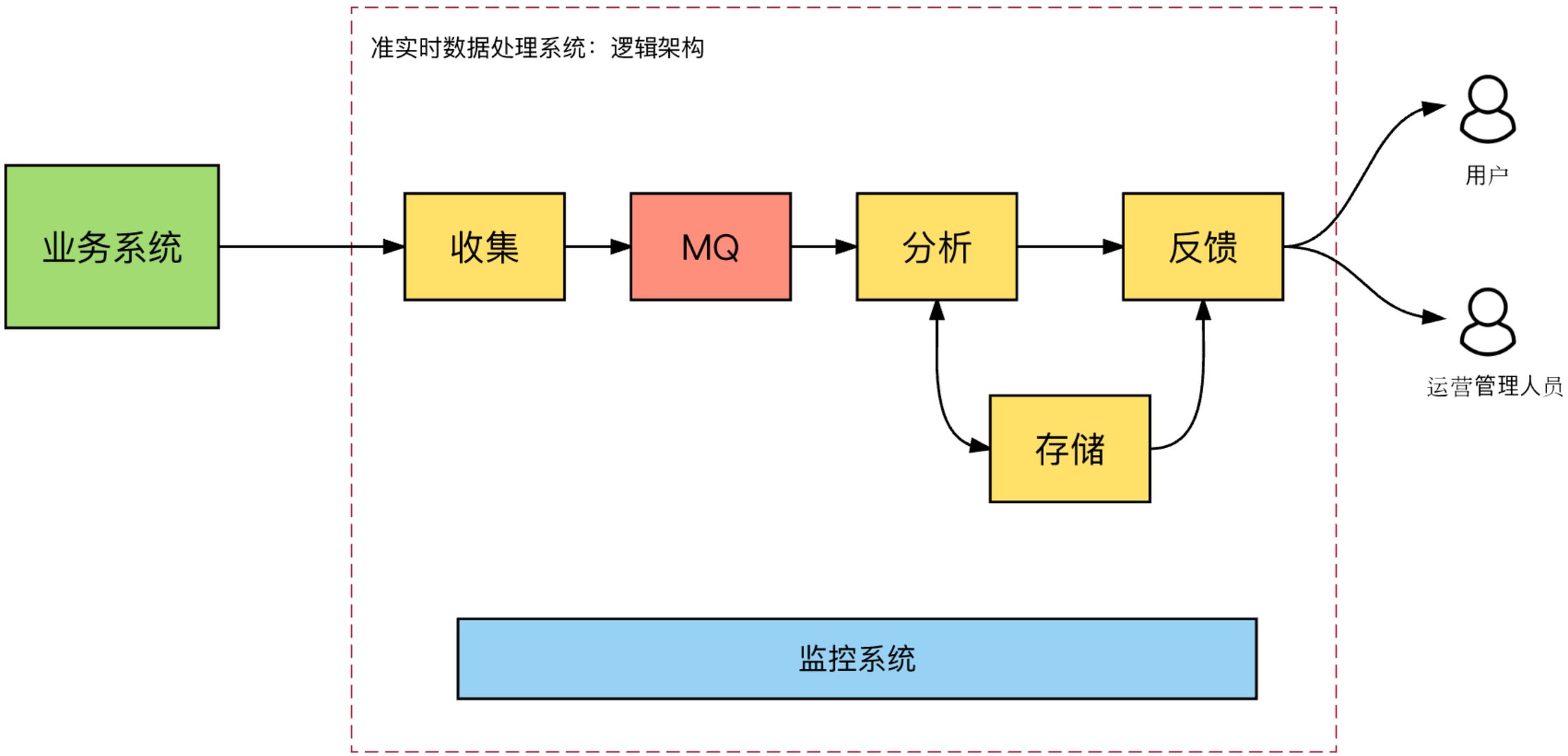


- Note：
  - 数据平台自身的监控、维护、升级。

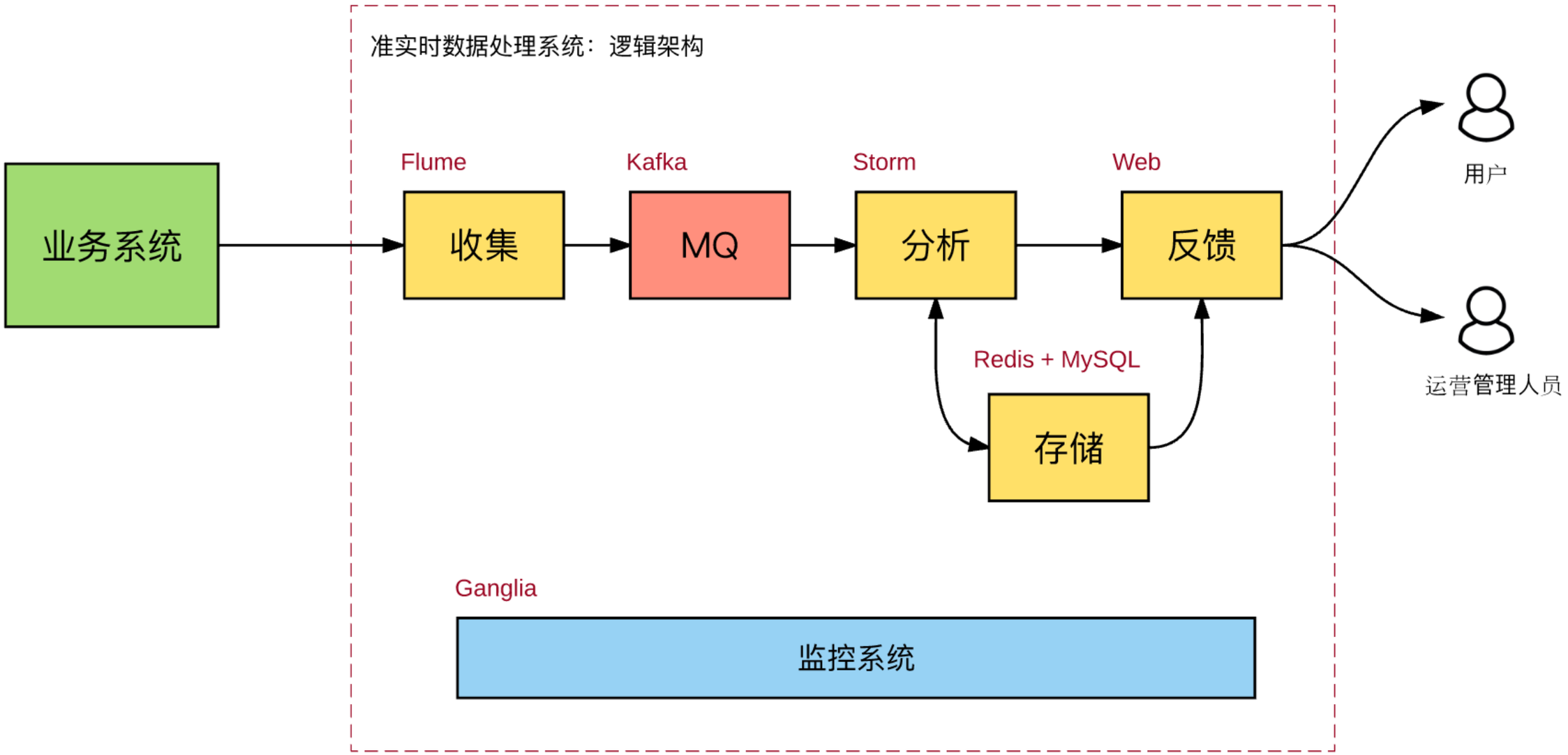
实例：准实时数据系统



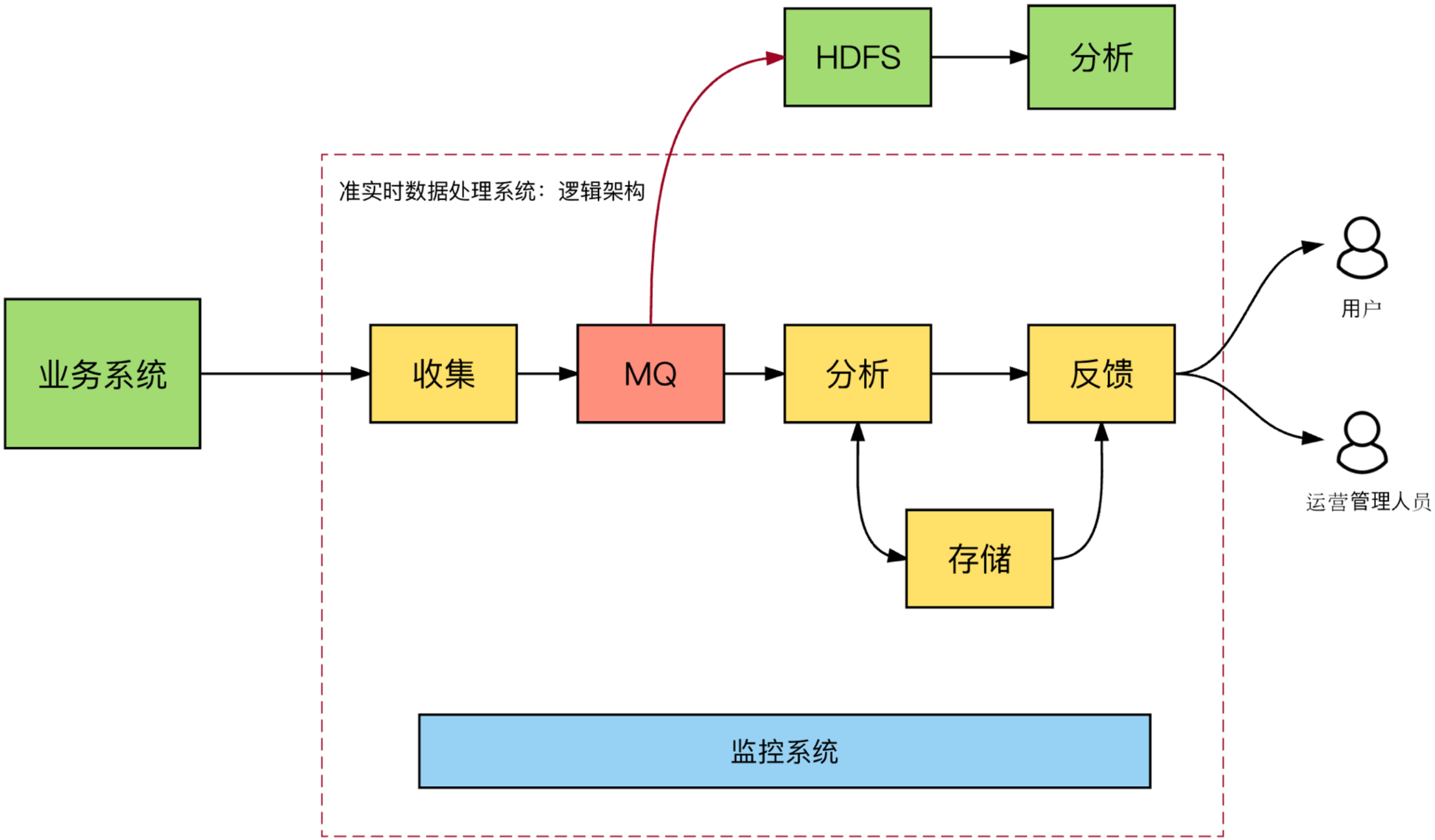
实例：准实时数据系统



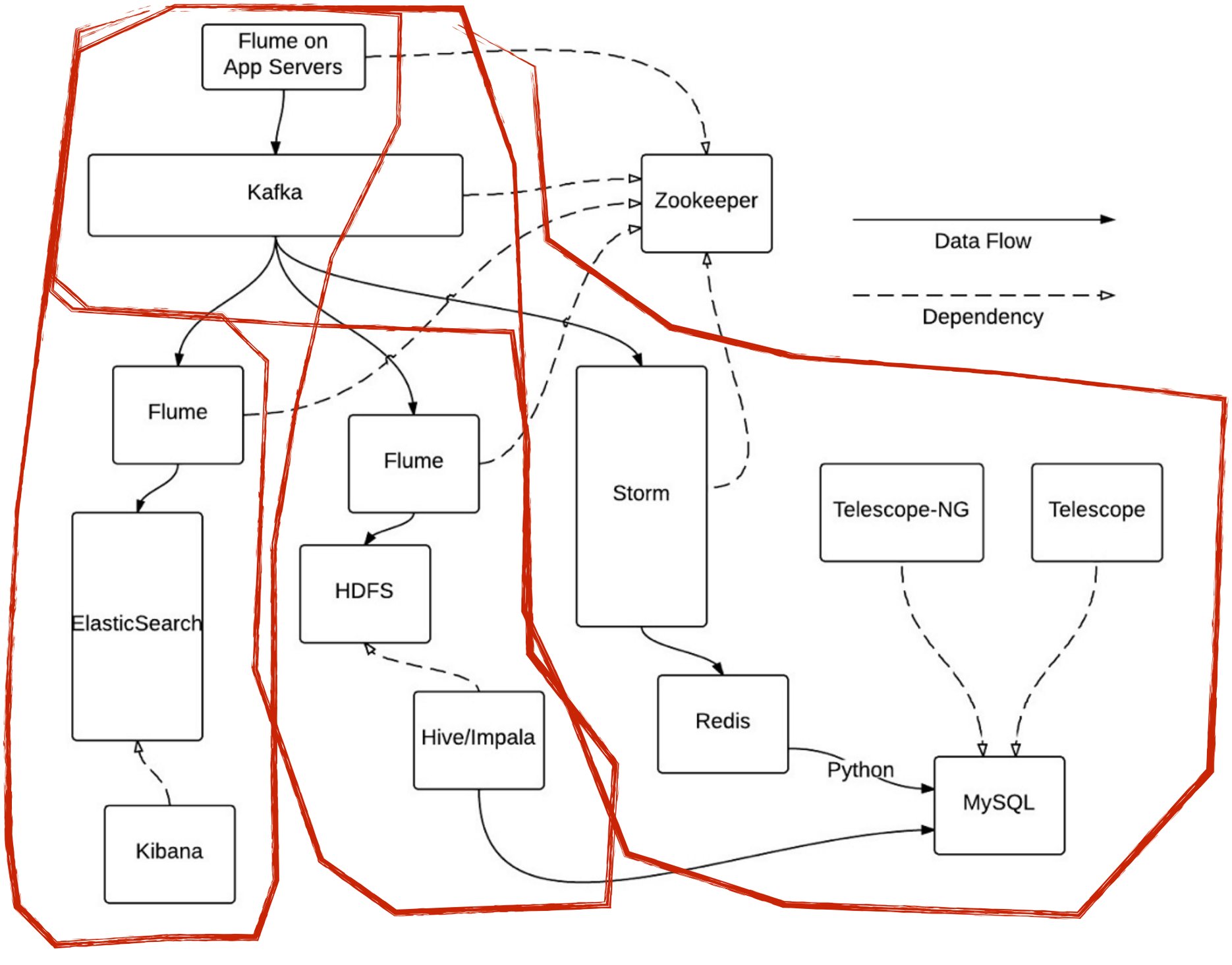
实例：准实时数据系统


准实时数据处理系统：逻辑架构
Flume
Kafka
Storm
Web
业务系统
收集
MQ
分析
反馈
用户
运营管理人员
Redis + MySQL
存储
Ganglia
监控系统

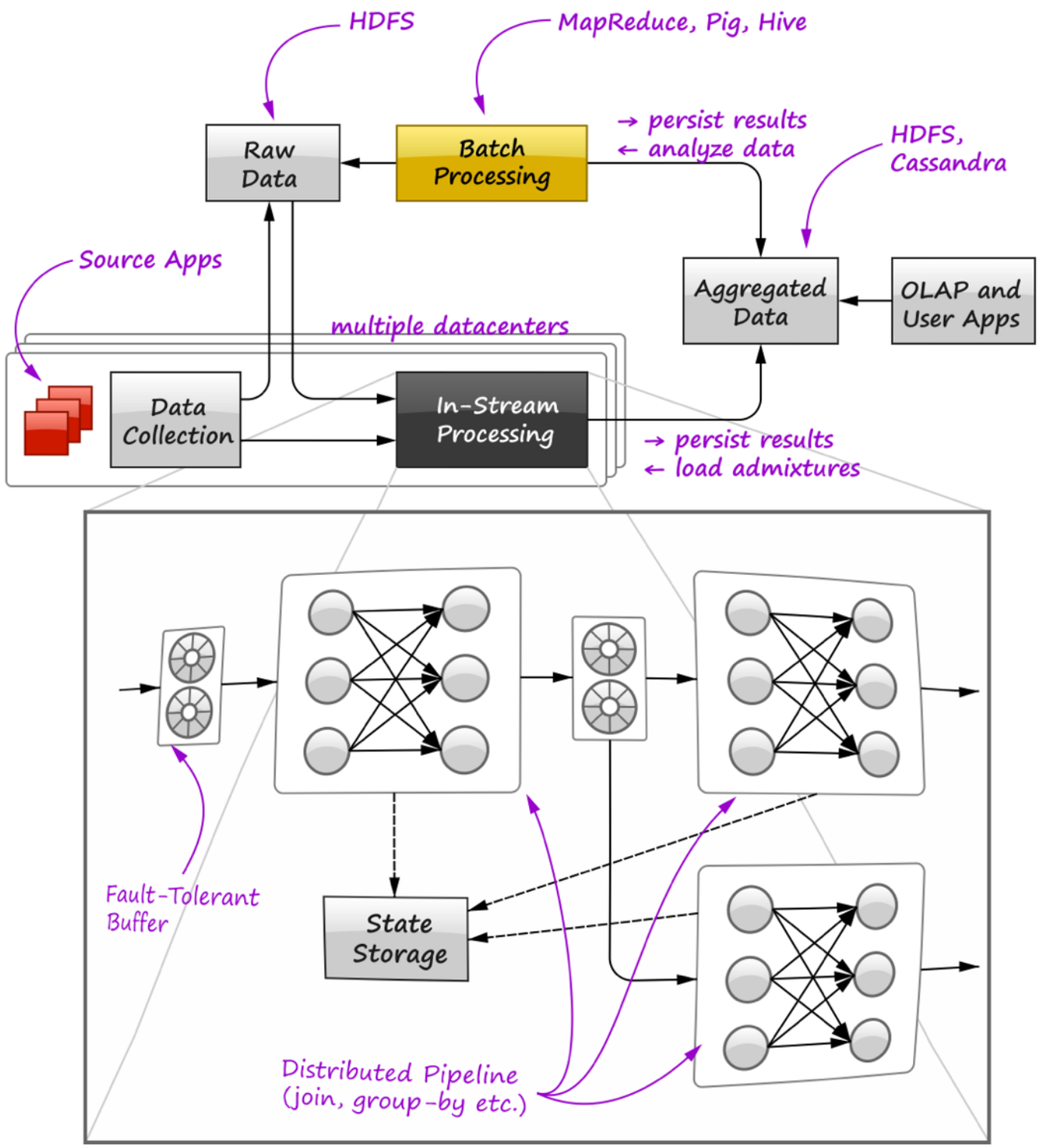
实例：准实时数据系统



实例：唯品会—日志平台

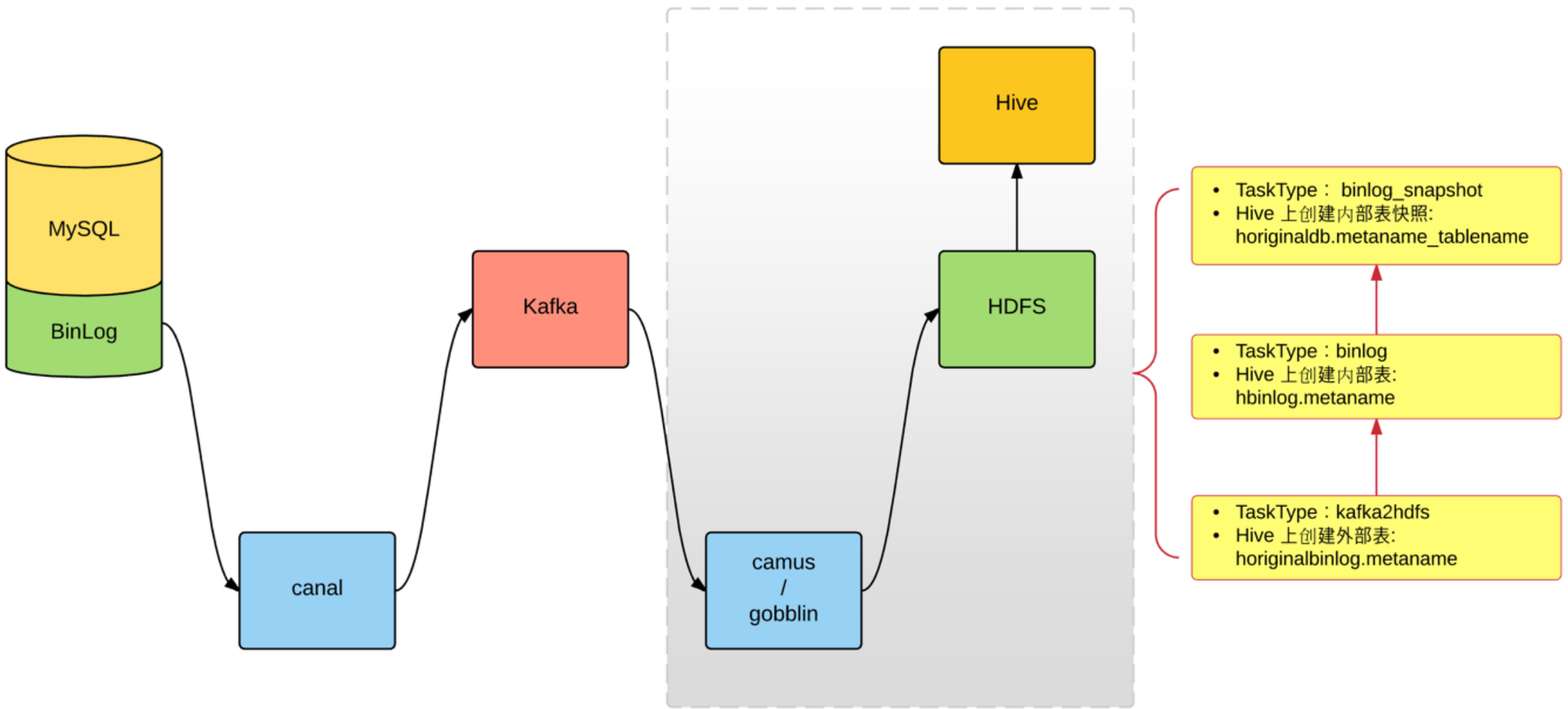

Flume on
App Servers
Kafka
Zookeeper
Data Flow
Dependency
Flume
Flume
Storm
Telescope-NG
Telescope
ElasticSearch
HDFS
Hive/Impala
Redis
Python
MySQL
Kibana

实例：Grid Dynamics—数据平台



问题5：前面都是日志送入 MQ，有没有收集 MySQL 增量数据的方案?

实例：美团—数据平台组


MySQL
BinLog
Kafka
canal
Hive
HDFS
camus
/
gobblin
TaskType： binlog_snapshot
Hive 上创建内部表快照:
horiginaldb.metaname_tablename
TaskType：binlog
Hive 上创建内部表:
hbinlog.metaname
TaskType：kafka2hdfs
Hive 上创建外部表:
horiginalbinlog.metaname

其他

- 数据治理：
  - 数据规范
    - 数据质量监控
- DW & BI： 数据仓库与商业智能
- 数据可视化

参考资料

1. http://flume.apache.org/
2. http://storm.apache.org/
3. https://redis.io/
4. https://dirtysalt.github.io/in-stream-big-data-processing.html
5. Mining of Massive Datasets - Stanford InfoLab 2010