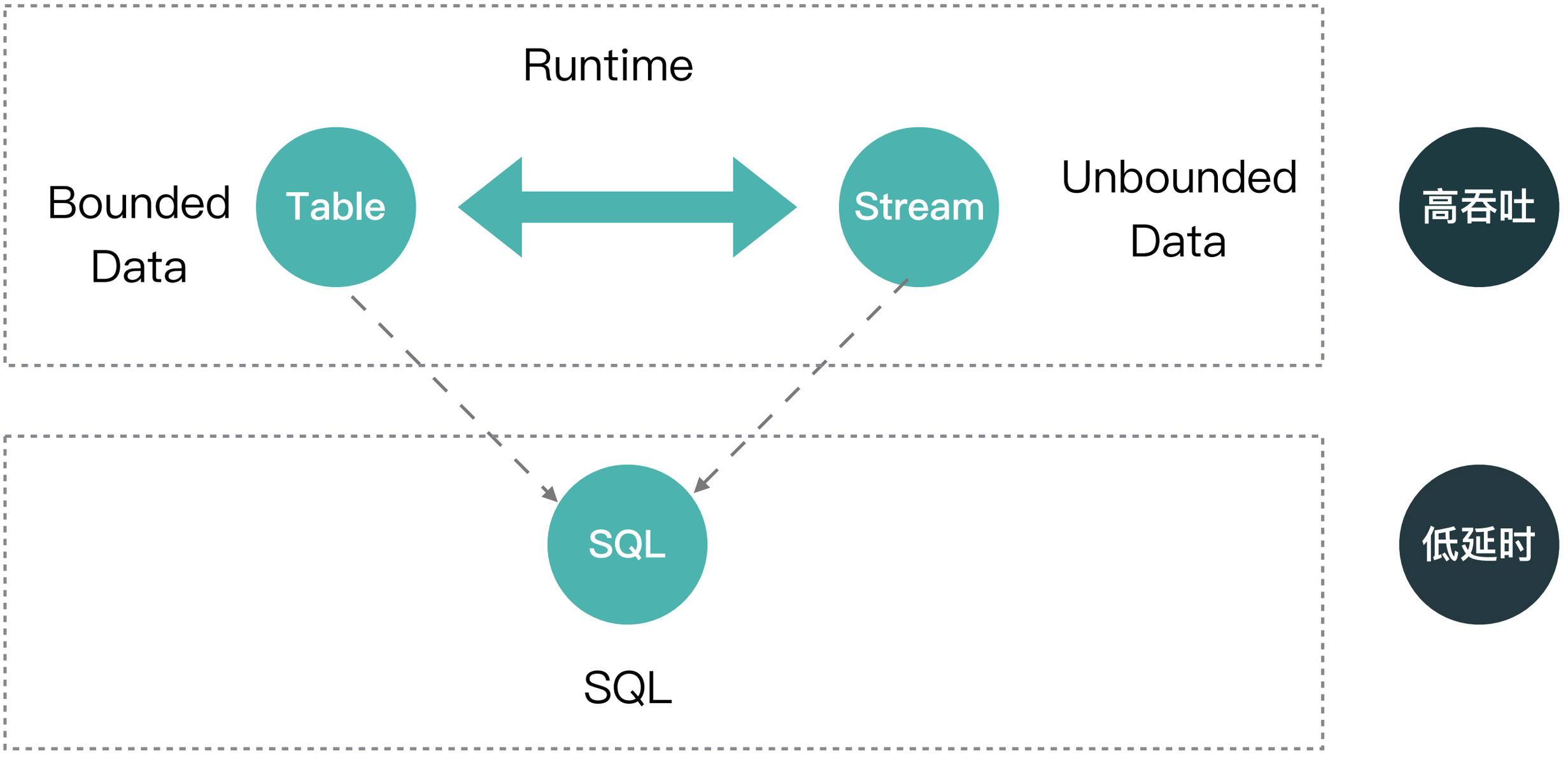
Flink

批处理及其应用

What is Apache Flink

* Apache Flink 是一个分布式大数据处理引擎
* 可对**有限数据流**和**无限数据流**进行有状态计算
* 可部署在各种集群环境
* 对各种大小的数据规模进行快速计算

为什么Flink能做批处理

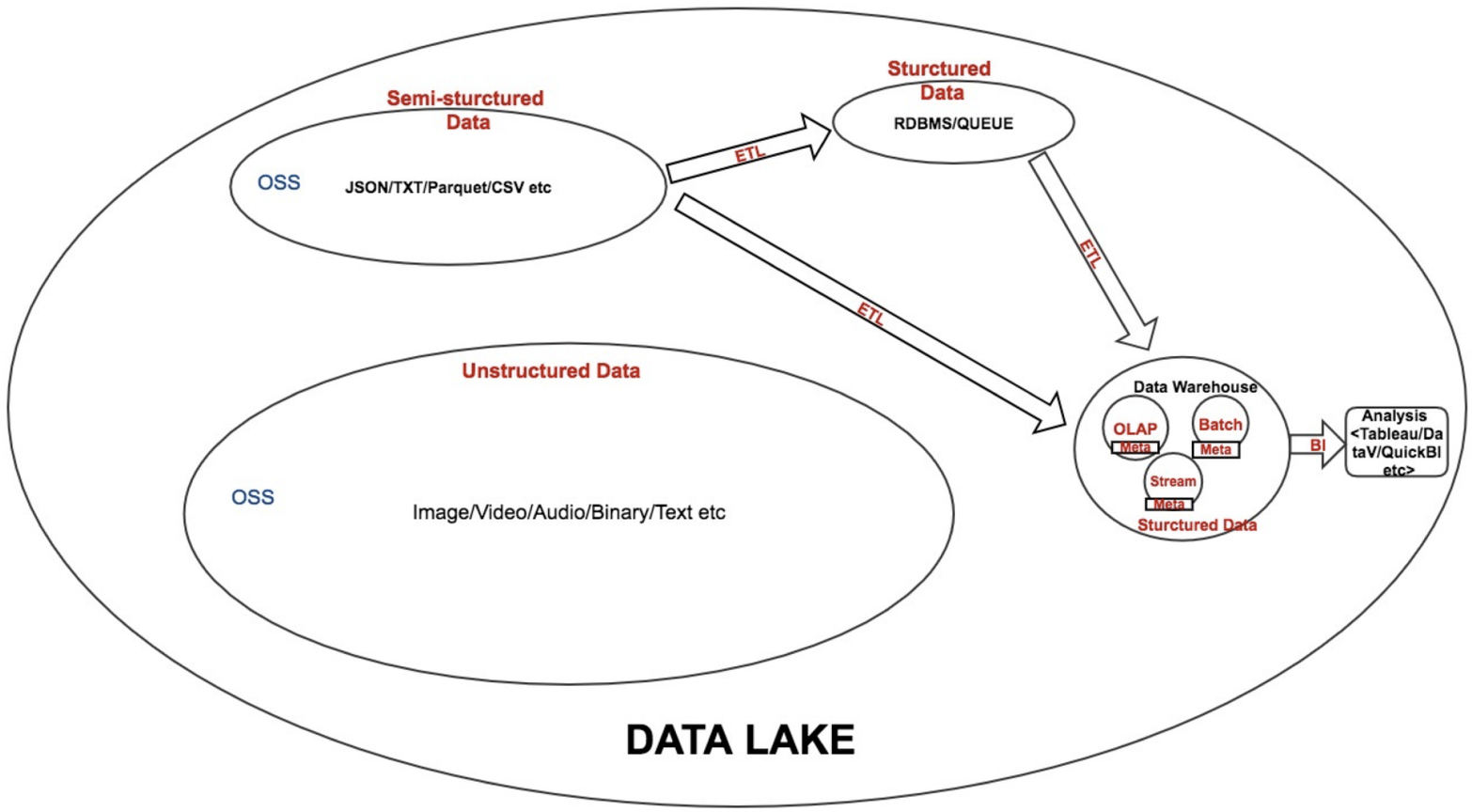


Hive vs. Spark vs. Flink Batch

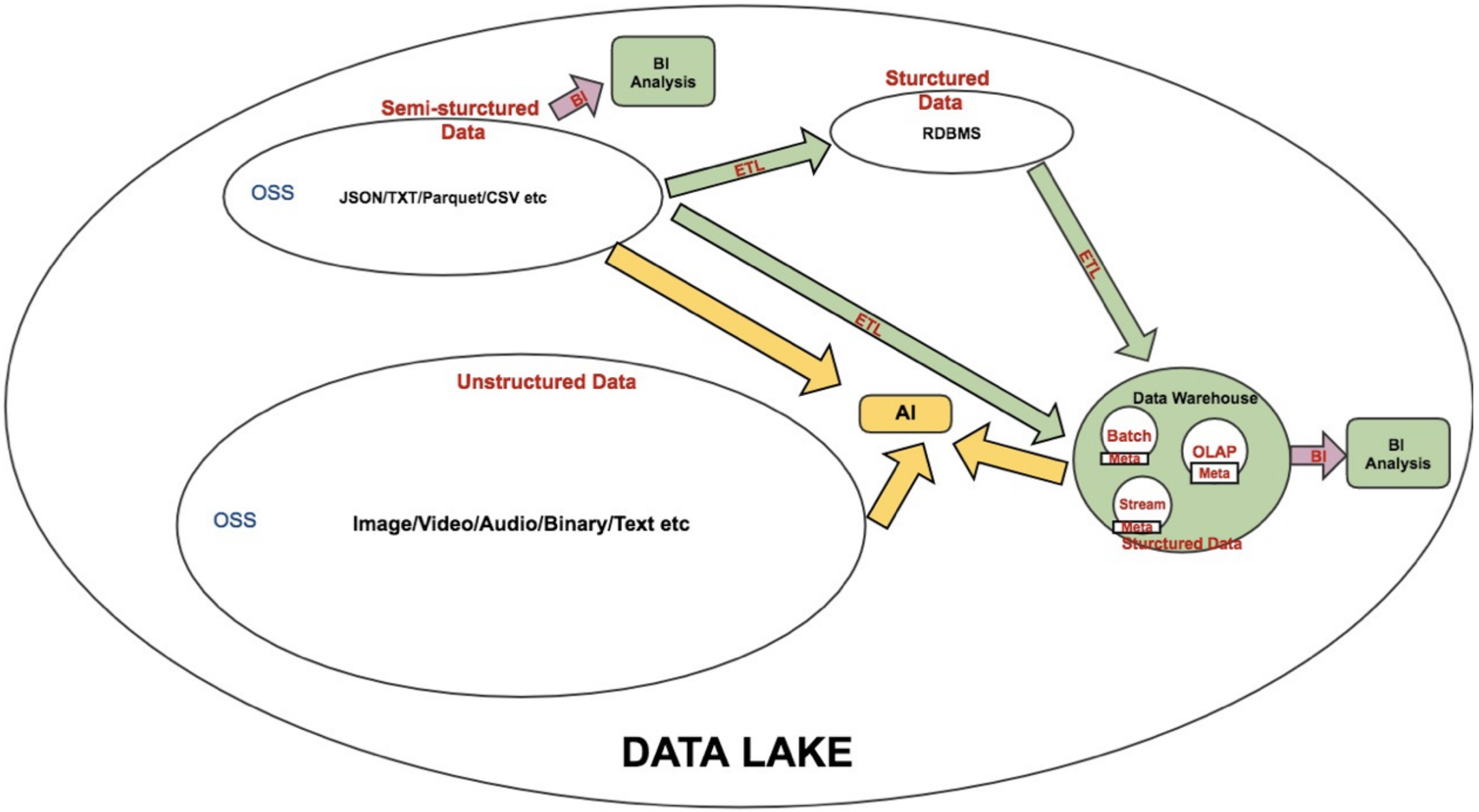
	Hive/Hadoop	Spark	Flink
模型	MR	MR(Memory/Disk)	Pipeline
吞吐	TB-PB	TB-PB	未经大规模生产验证
性能	一般(分钟小时级别)	快(秒级)	优秀 x2
稳定性	好	一般	已在阿里内部验证
API	差(MR)	最丰富 (RDD/DataSet/DataFrame) Python/Scala/R/Java	丰富 (TableAPI) Scala/Java
SQL	HiveSQL	SparkSQL	ANSI SQL
易用性	一般	易用	一般
工具/生态	一般	丰富	一般

Flink Batch应用 – 数据湖

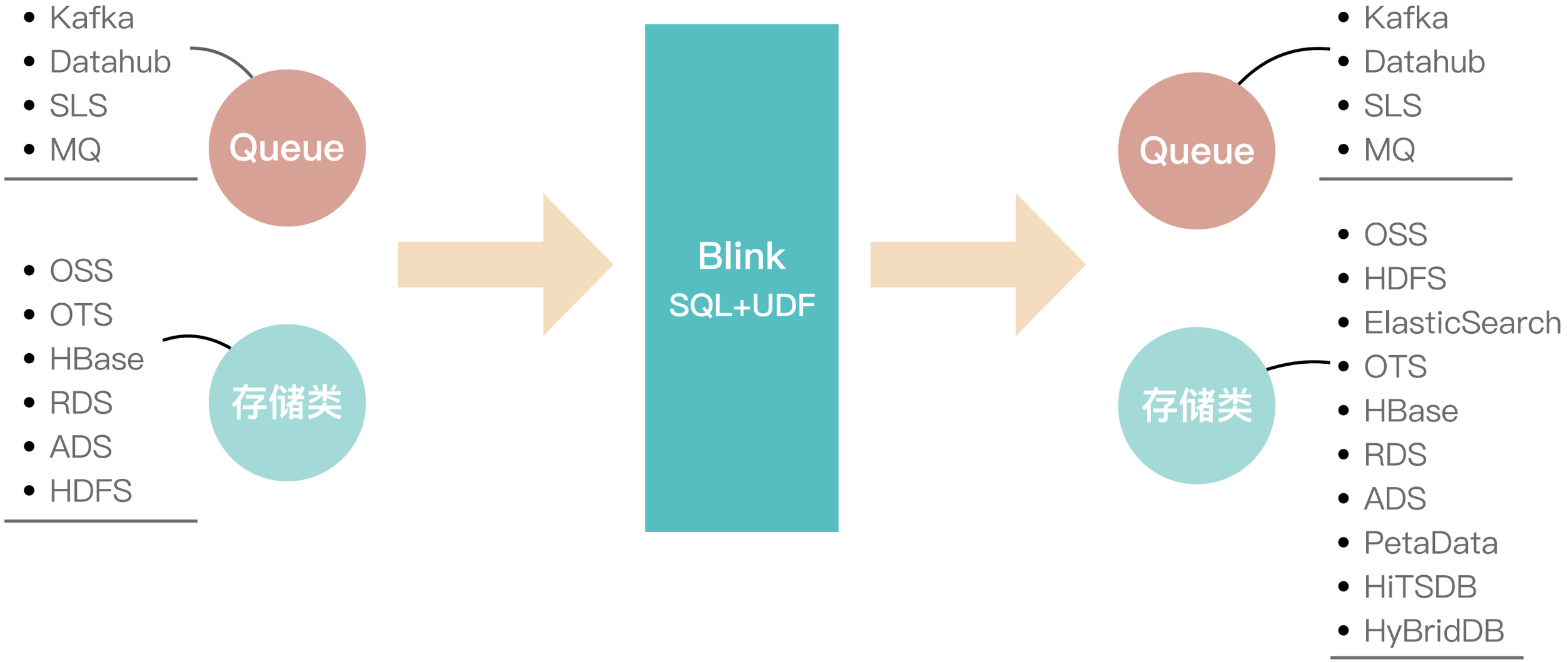
Data Lake vs. Data Warehouse



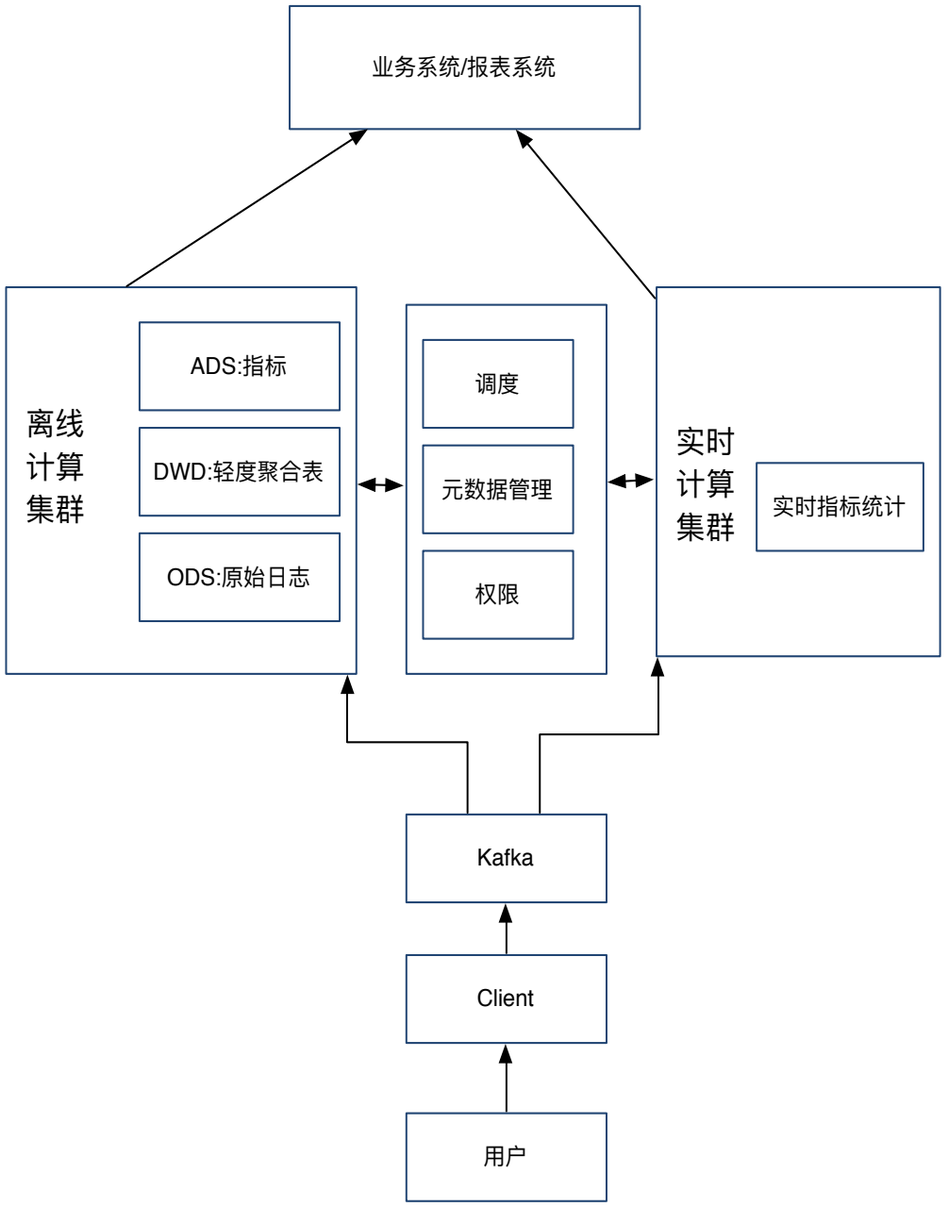
Flink Batch应用 – 数据湖

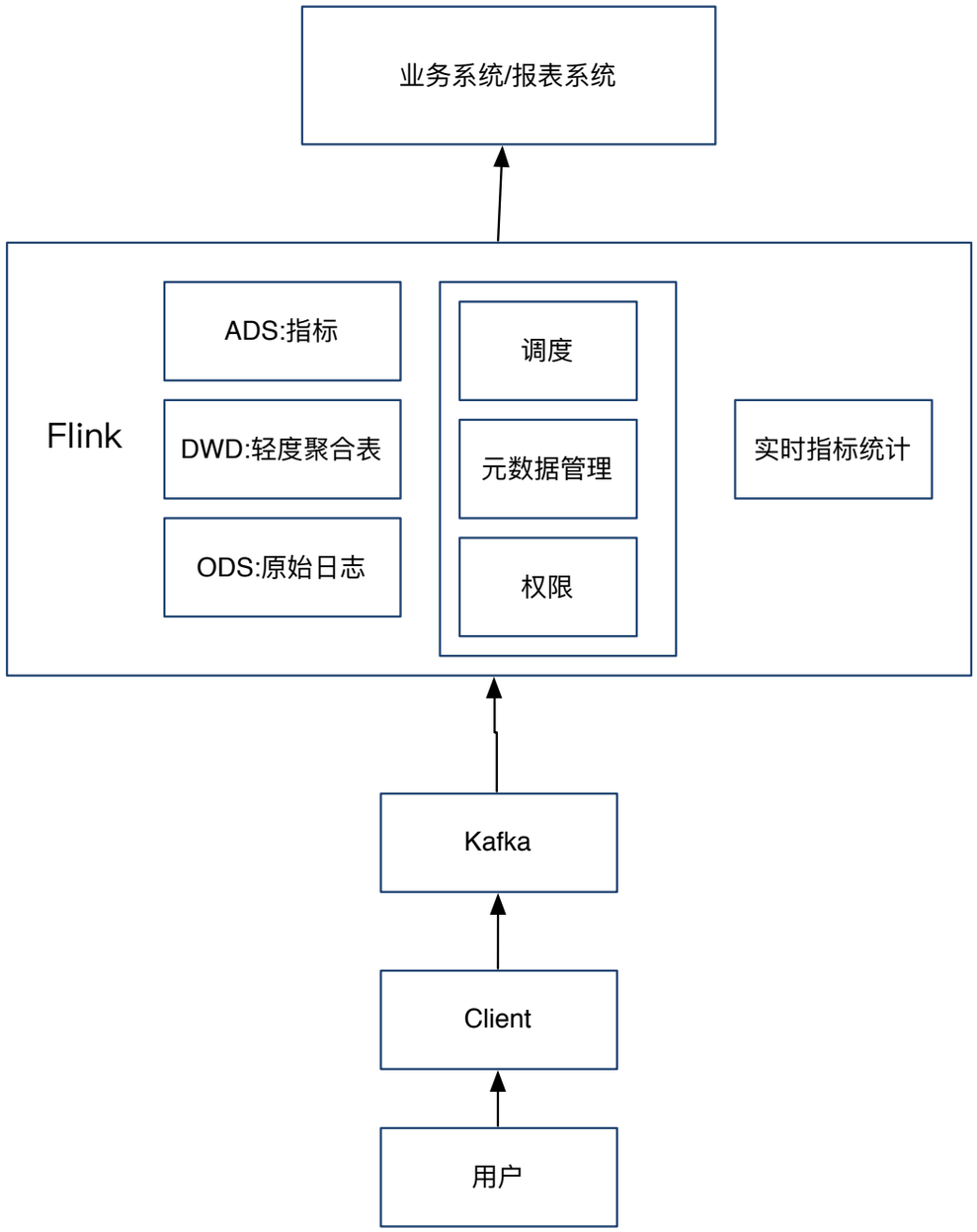


Flink Batch应用 – 数据湖

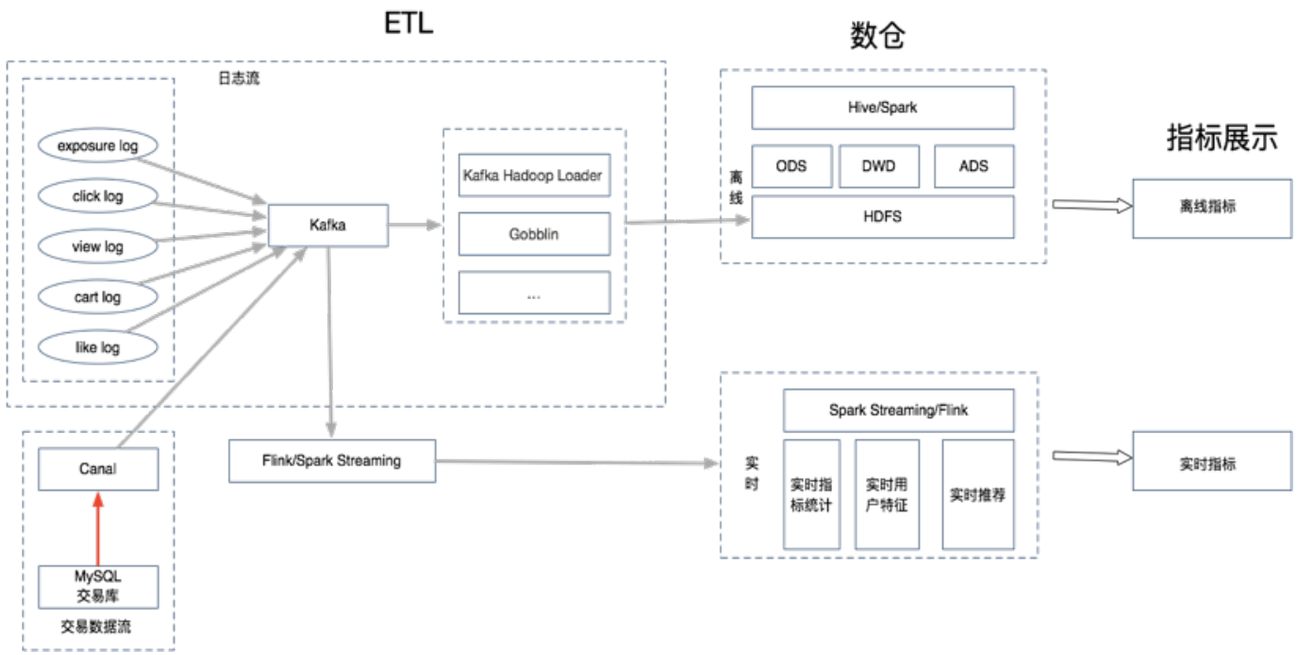


Flink Batch应用 – 数仓

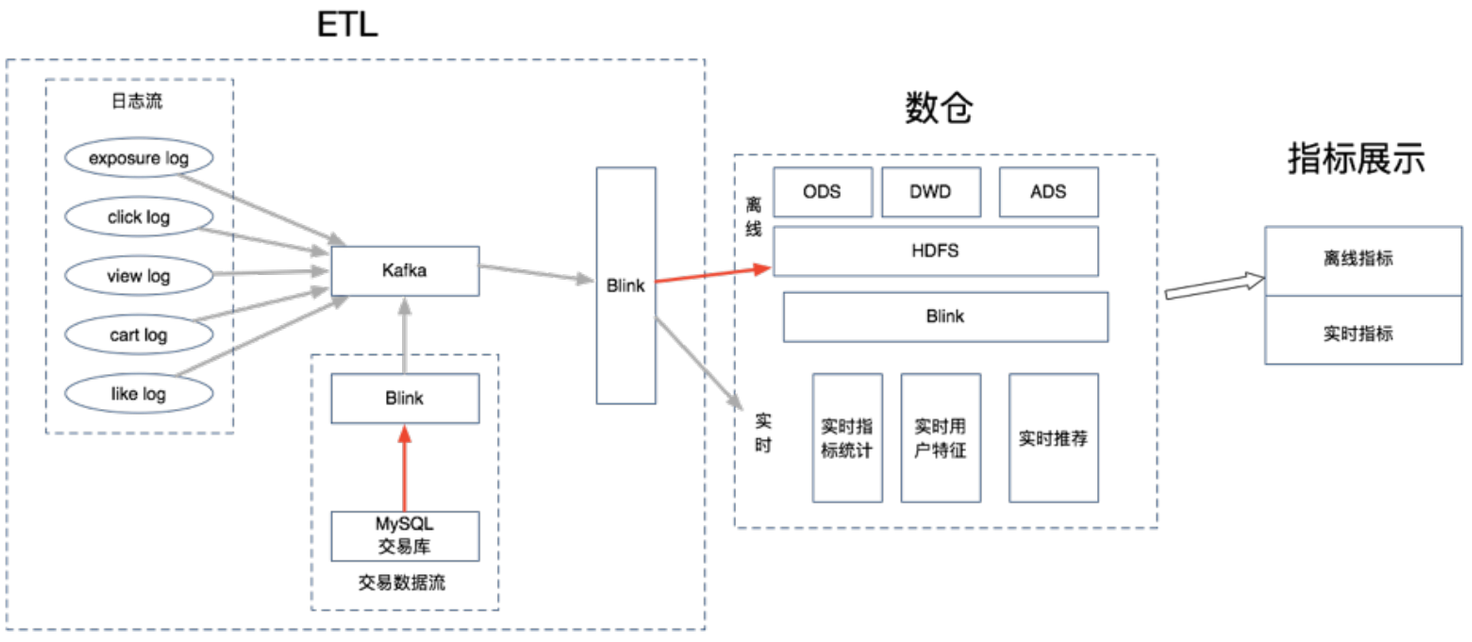




Flink Batch应用 – 数仓

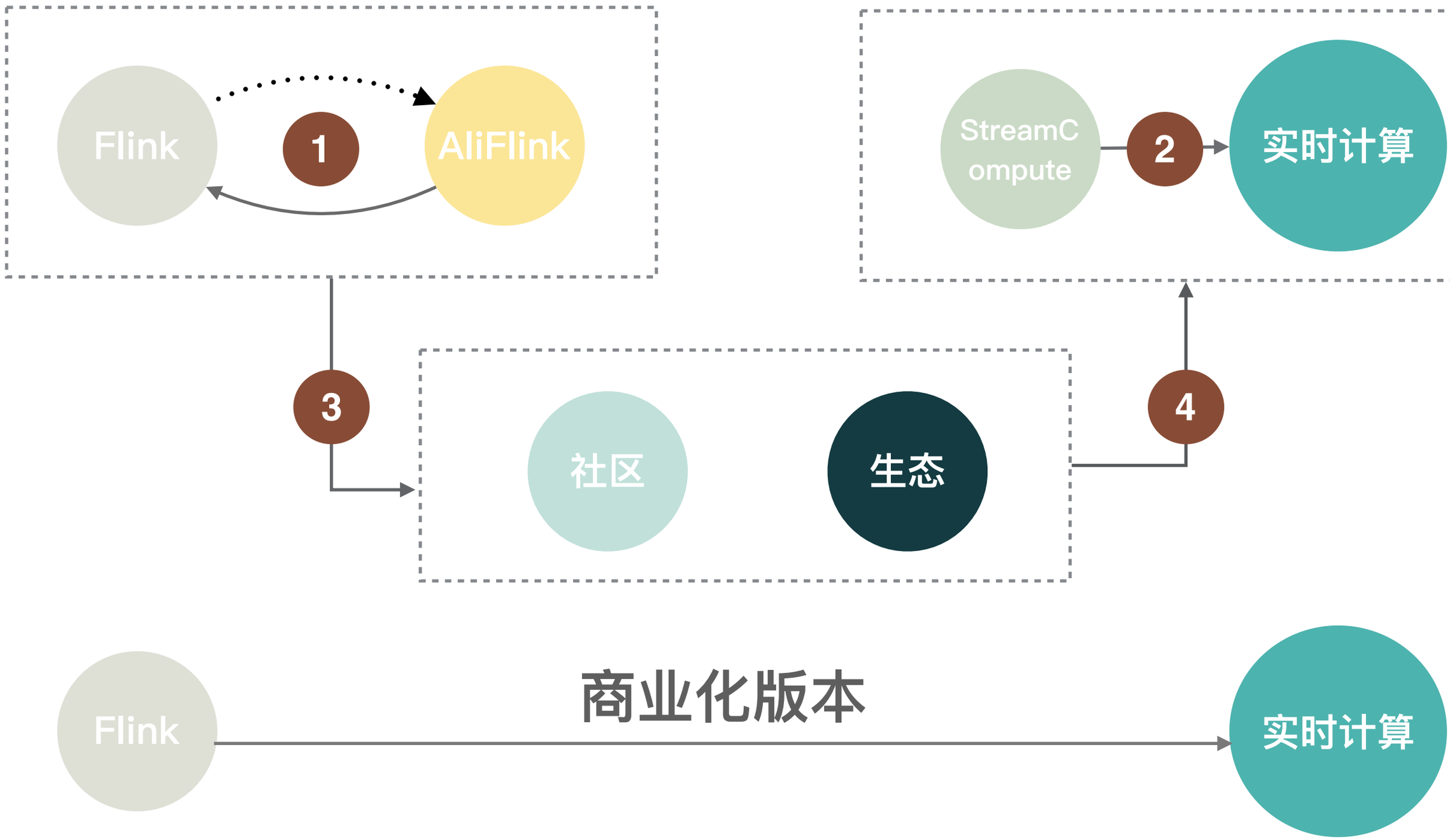


简化架构



方便运维

Flink社区规划


Flink
1
AliFlink
StreamC
ompute
2
实时计算
3
4
社区
生态
Flink
商业化版本
实时计算

阿里云实时计算产品方向

全功能大数据处理能力

存储计算分离架构

高性能

全托管架构

Thanks