

联合主办: MEIZU | Flyme | msup®

数据洞察

魅族技术开放日第八期

大数据平台架构技术选型 与场景运用



张逸
大眼科技联合创始人 CTO

数据采集

数据存储

数据处理



云端



数据库



文件



WEB

数据源

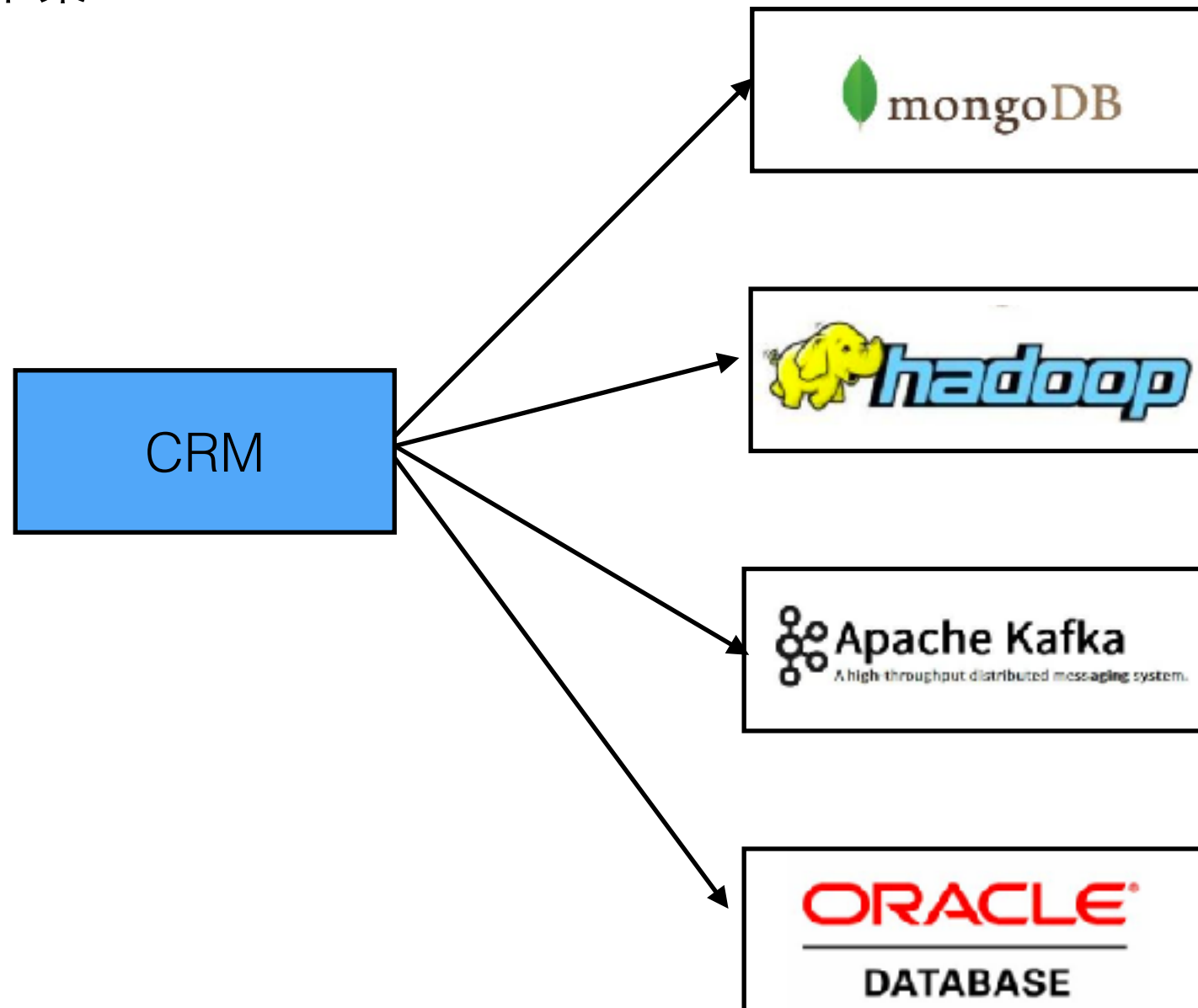
数据源的特点

来源		结构		可变性		数据量	
	内部数据		非结构化		不变可添加		大
	外部数据		结构化		修改删除		小

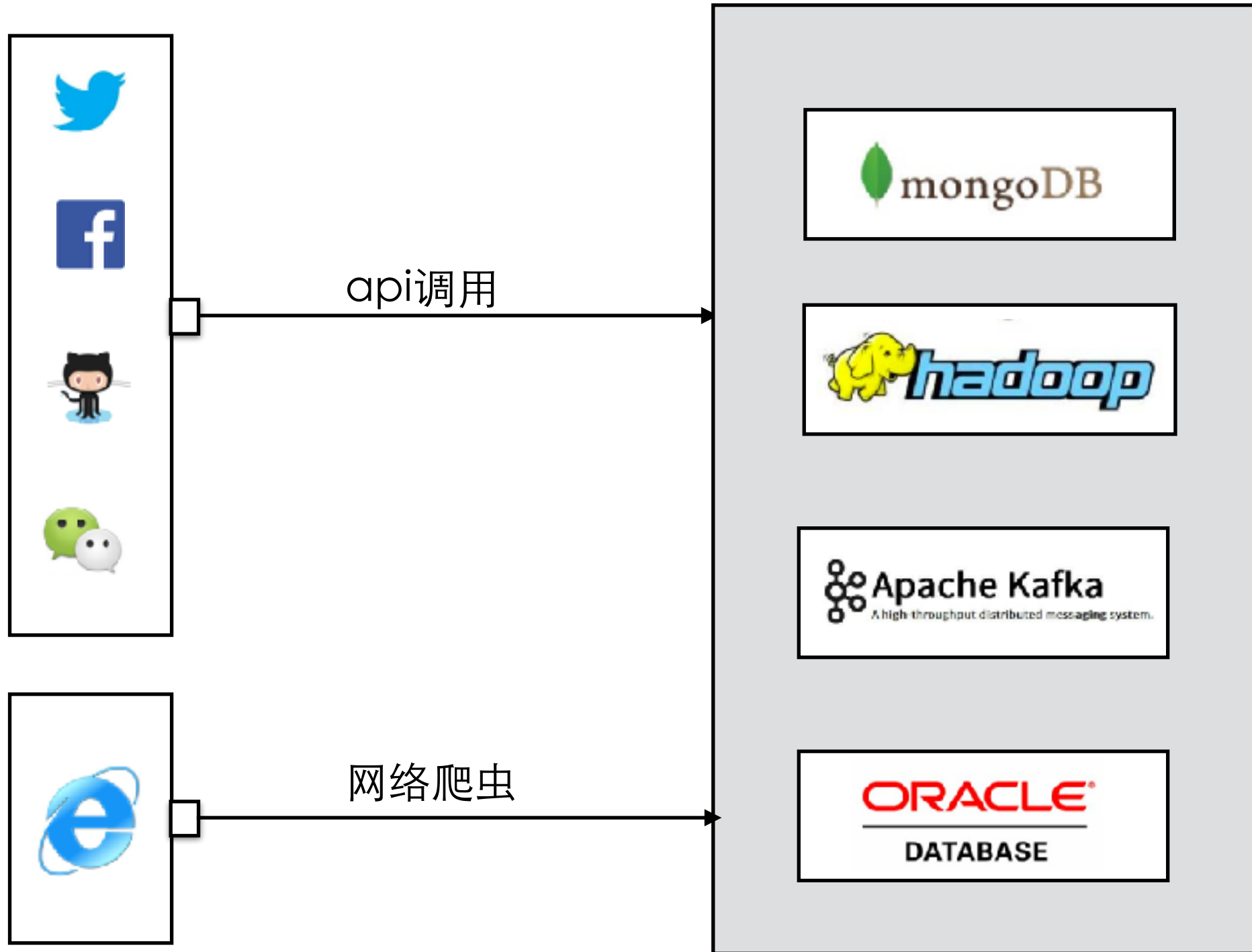
数据源的特点决定了
数据采集与数据存储的技术选型

内部数据

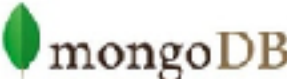
来自企业内部系统，可以采用主动写入技术(push)，从而保证变更数据及时被采集。



外部数据



❌ 非结构化数据

Document Database	Graph Databases
 Couchbase  MarkLogic  mongoDB	 Neo4j  InfiniteGraph The Distributed Graph Database
Wide Column Stores	Key-Value Databases
 redis  amazon DynamoDB  AEROSPIKE  riak	 accumulo HYPERTABLE [™]  Cassandra  APACHE HBASE Amazon SimpleDB

@cloudbit <http://www.aryannava.com>

🔗 结构化数据



不变可添加

如果数据源的数据是不变的，或者只允许添加（通常，数据分析的事实表，例如银行交易记录等），则采集会变得非常容易，同步时只需要考虑最简单的增量同步策略，维持数据的一致性也相对变得容易。



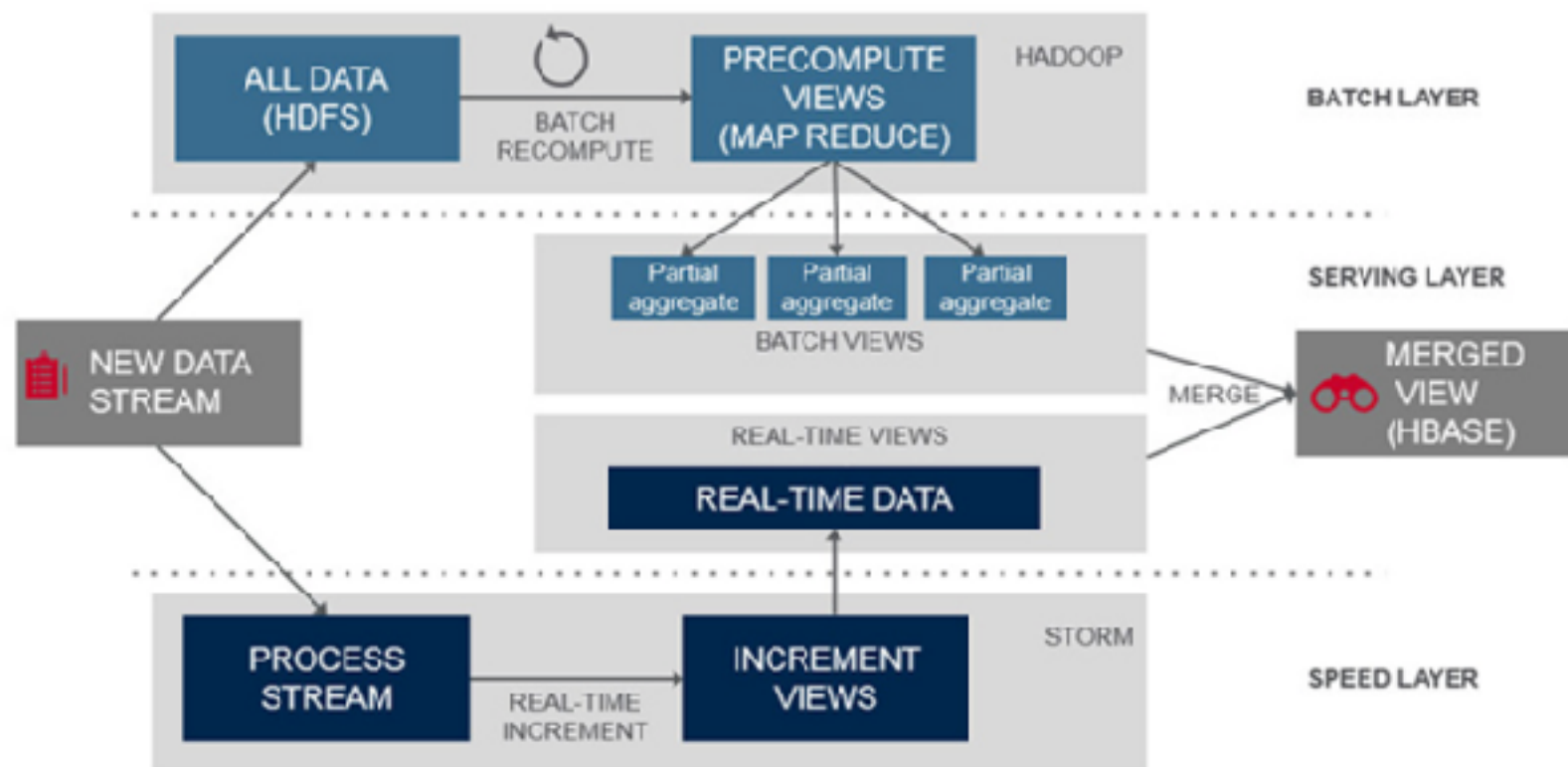
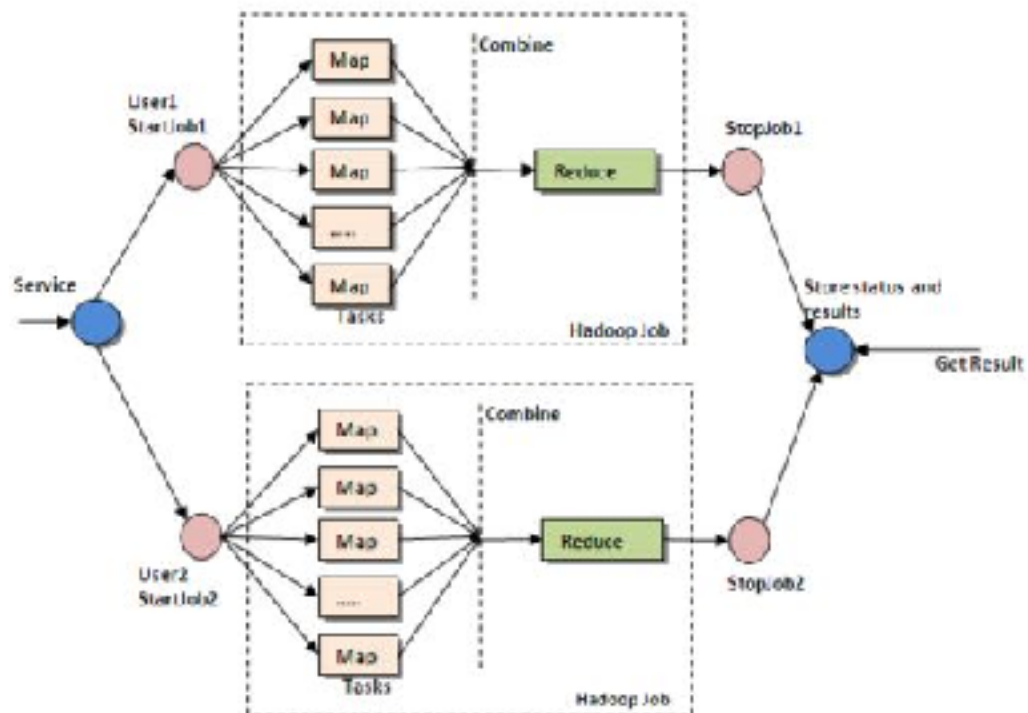
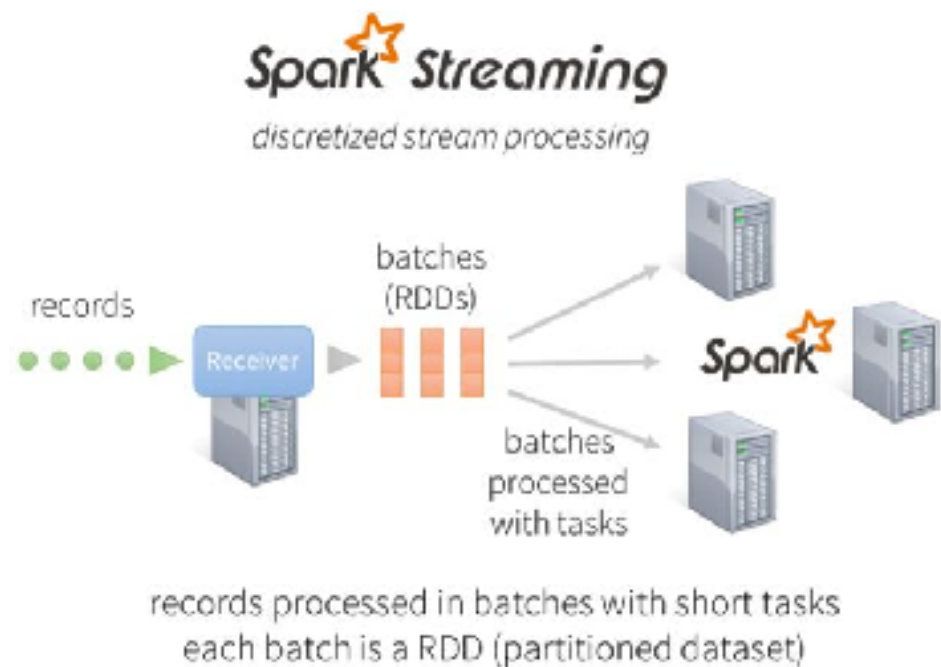
可修改可删除

数据源的数据有些可能会修改或删除，尤其是许多维表经常需要变动。要对这样的数据进行分析处理，最简单的办法就是采用直连形式。如果要进行数据采集，就要考虑同步问题。



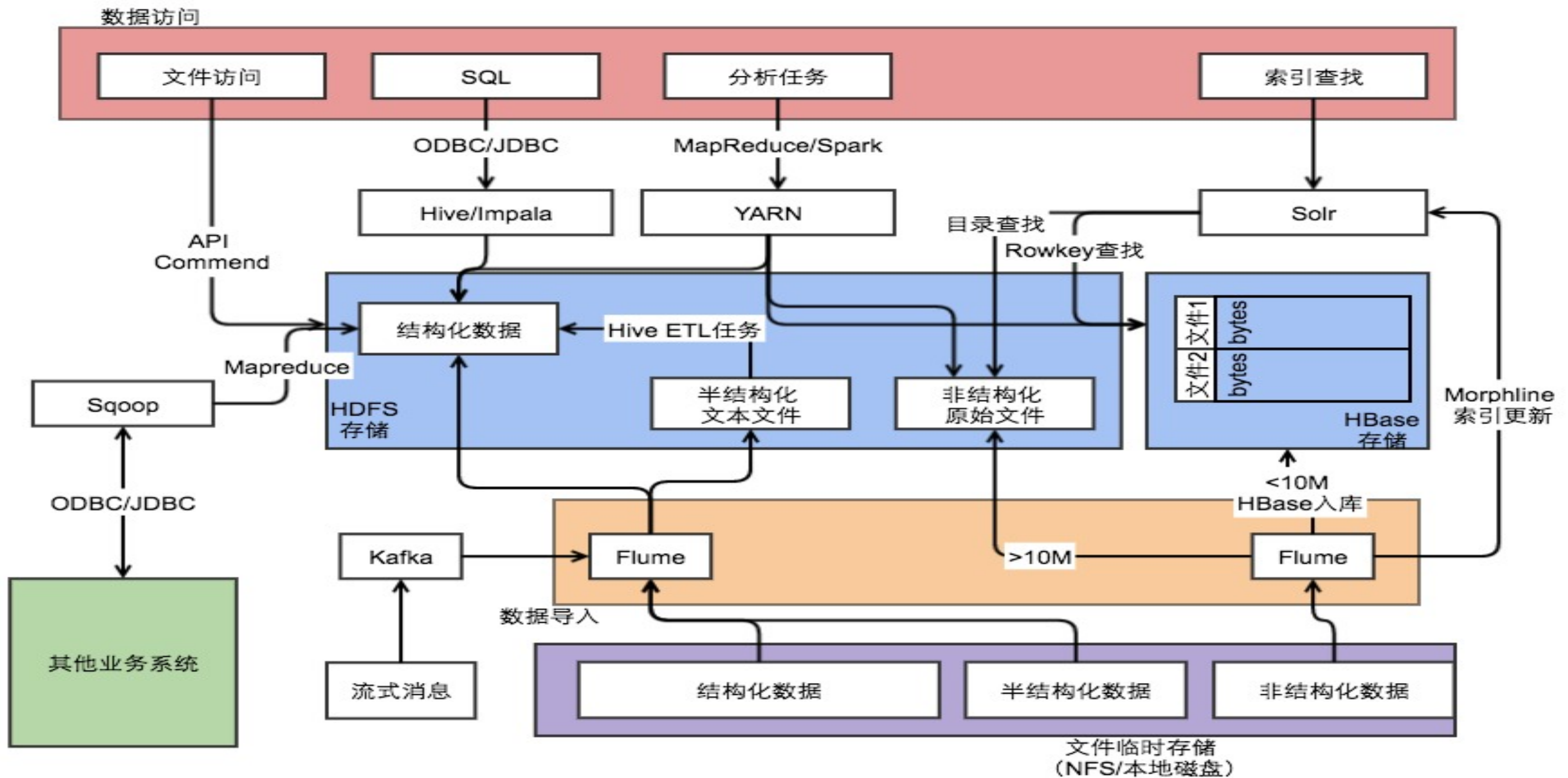


大数据量



lambda架构

一个典型的数据加载架构





数据存储的技术选型

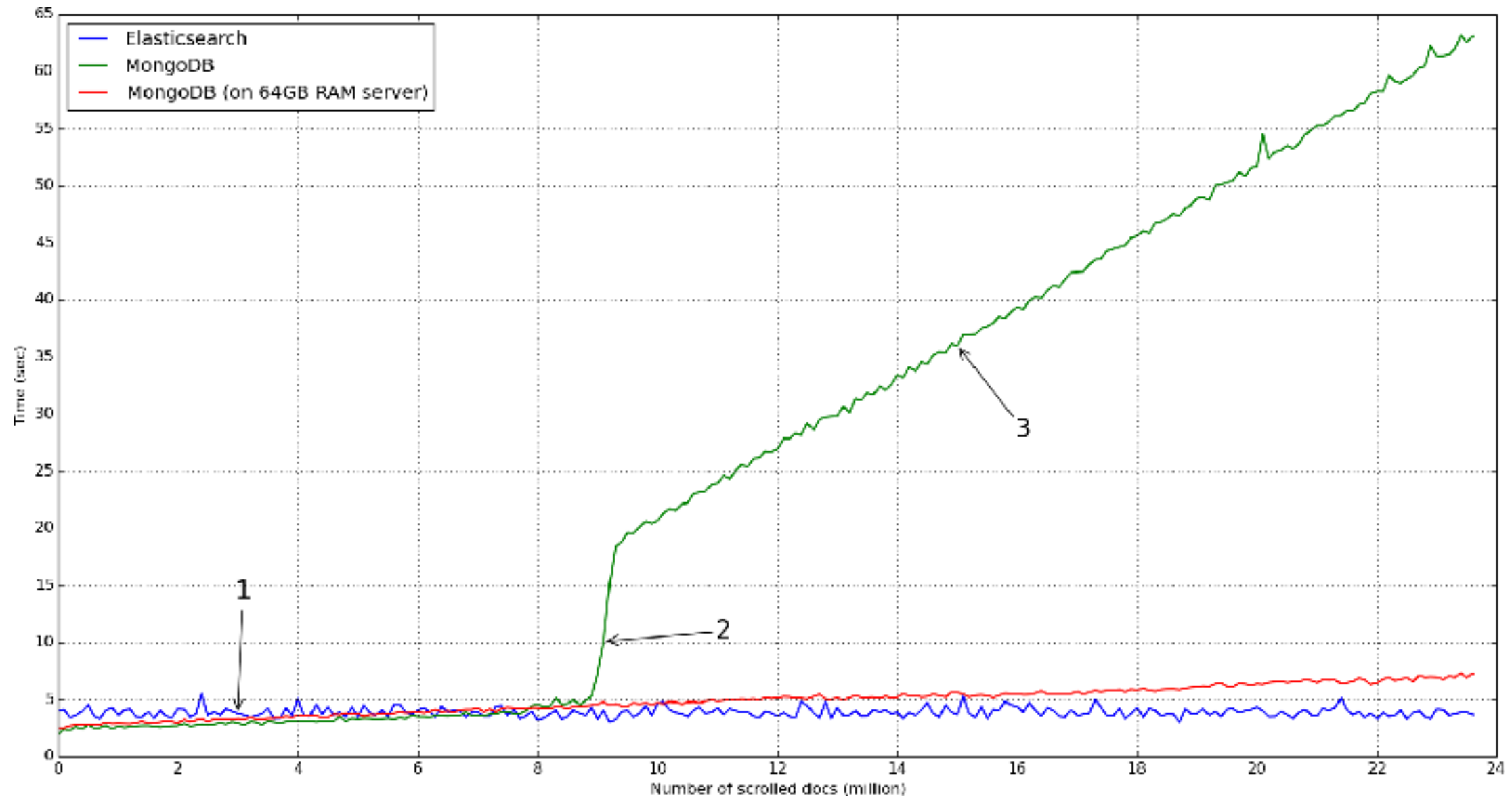
- 取决于数据源的类型与数据的采集方式
- 取决于采集后数据的格式与规模
- 取决于分析数据的应用场景

大数据平台 特征

相同的业务数据会以多种不同的表现形式，存储在不同类型的数据库中，形成一种poly-db的数据冗余生态。



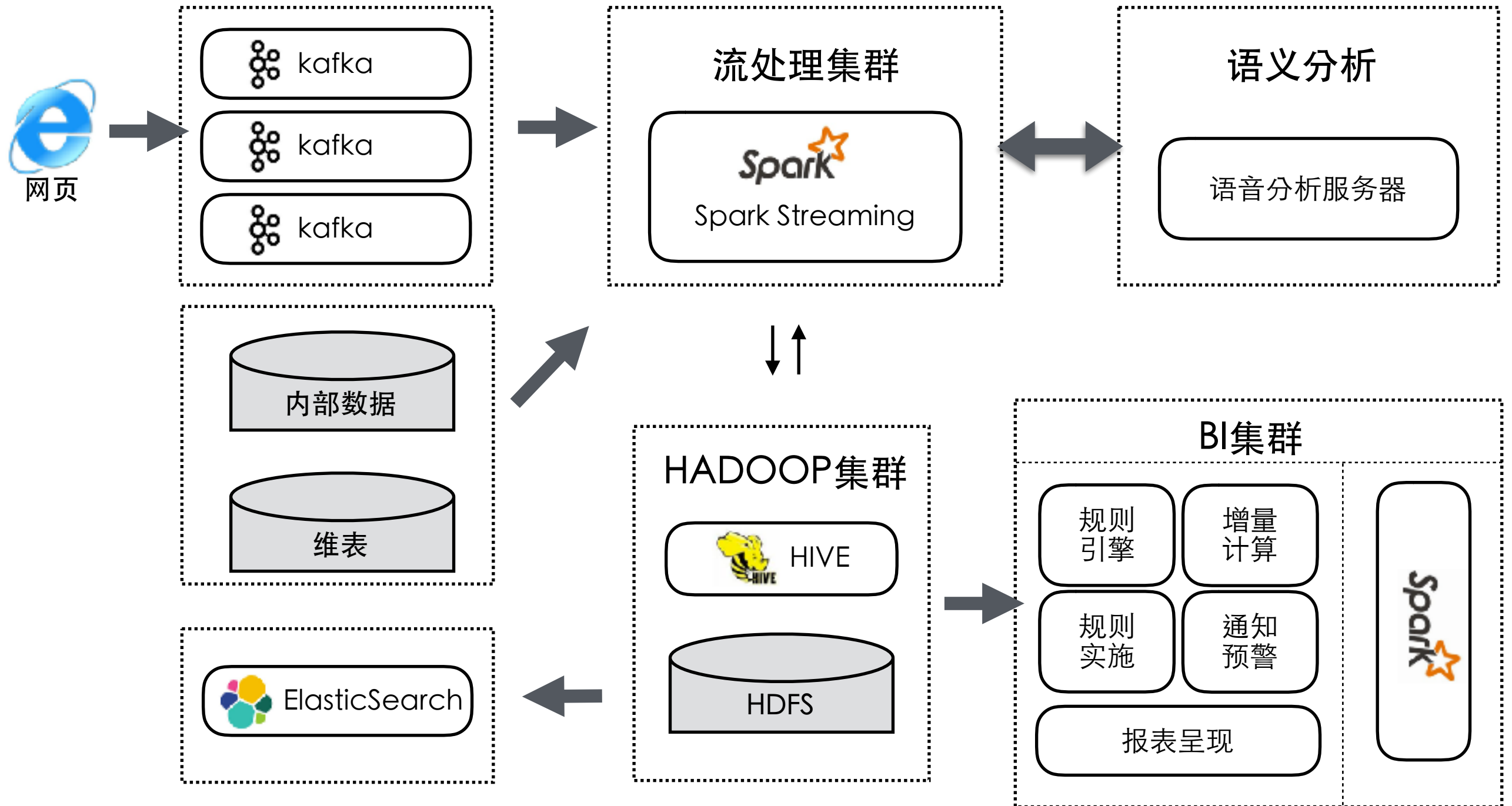
场景一：舆情分析



针对某手机品牌的舆情分析。客户提出的需求是能够对舆情数据进行全文本搜索。舆情数据最高可能达到70亿条，而全文本搜索的性能指标要求响应时间控制在10s以内。

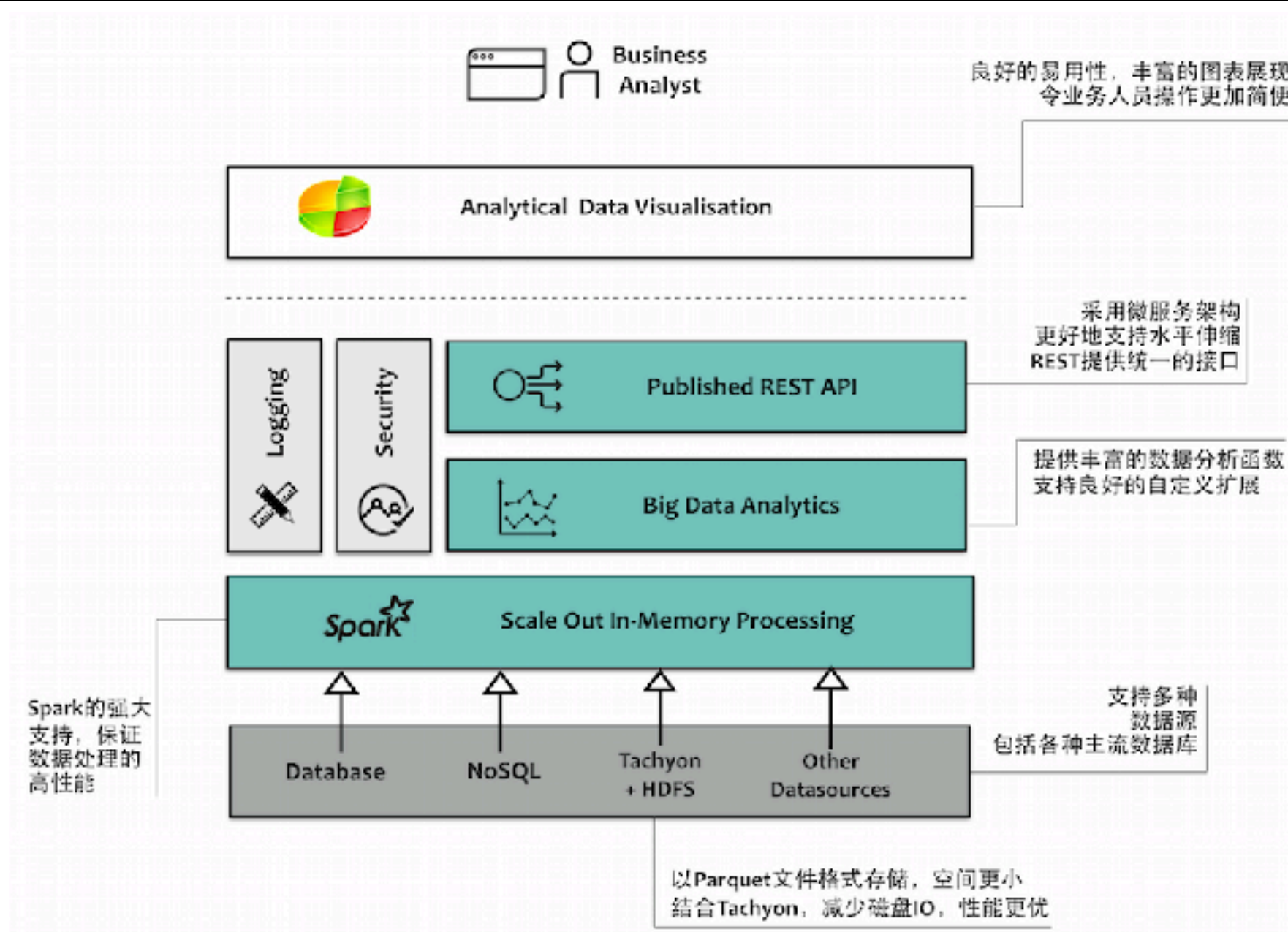


场景一：舆情分析





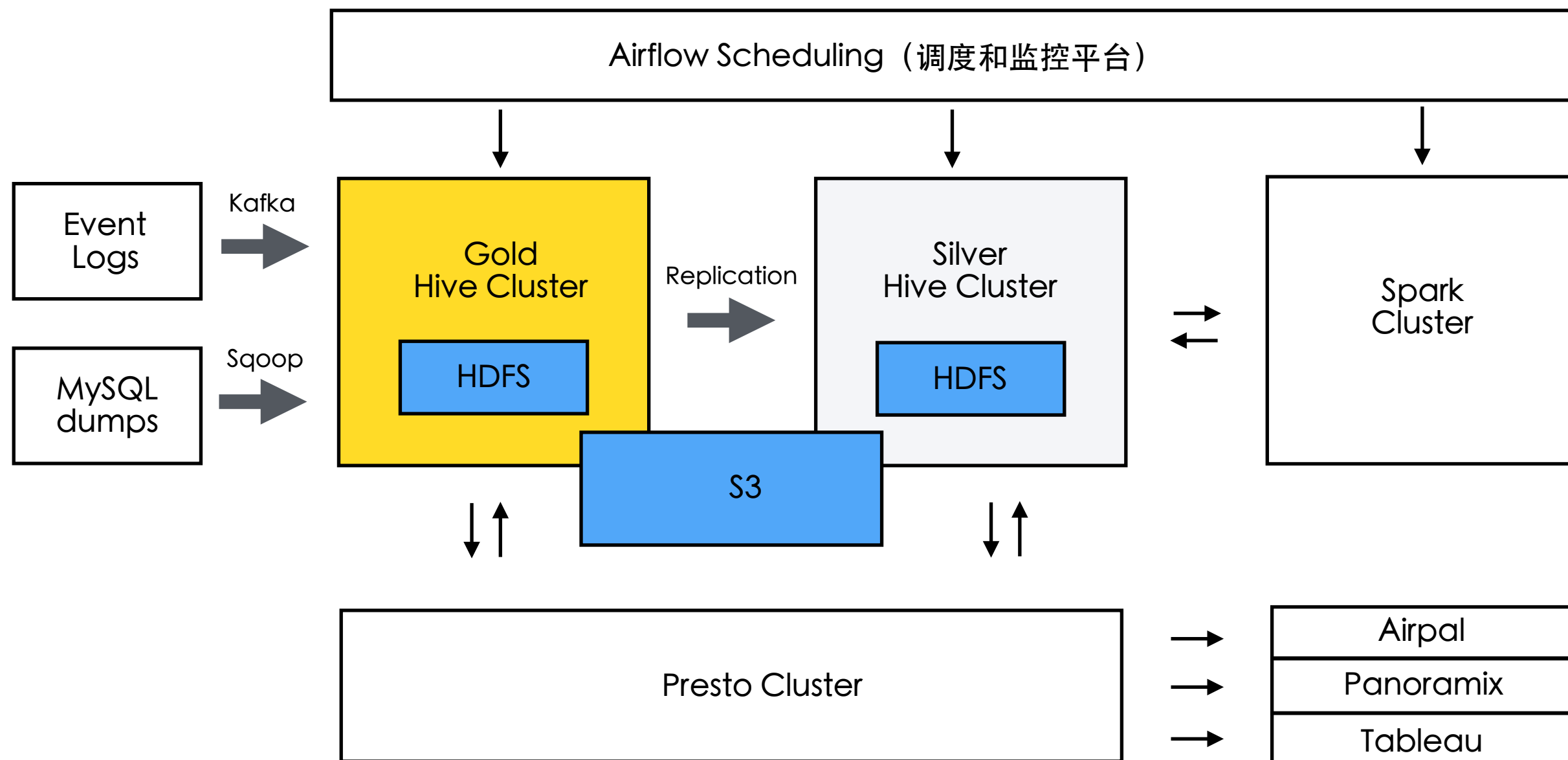
场景二：商业智能产品



BI产品主要针对数据集进行的数据分析以聚合运算为主。我们既要满足大数据量的水平可伸缩，又要满足高性能的聚合运算。选择Parquet列式存储，可以同时满足这两个需求。



场景三：Airbnb的大数据平台



数据源不同，采集方式也不一样。日志数据通过发送Kafka事件，而线上数据则通过Sqoop同步。数据存储选择HDFS集群，然后通过Presto对Hive表执行即席查询。S3是一个独立的存储系统。

数据处理的分类

