QingCloud Technologies

大数据专题科普

2022.09.07

迟连义

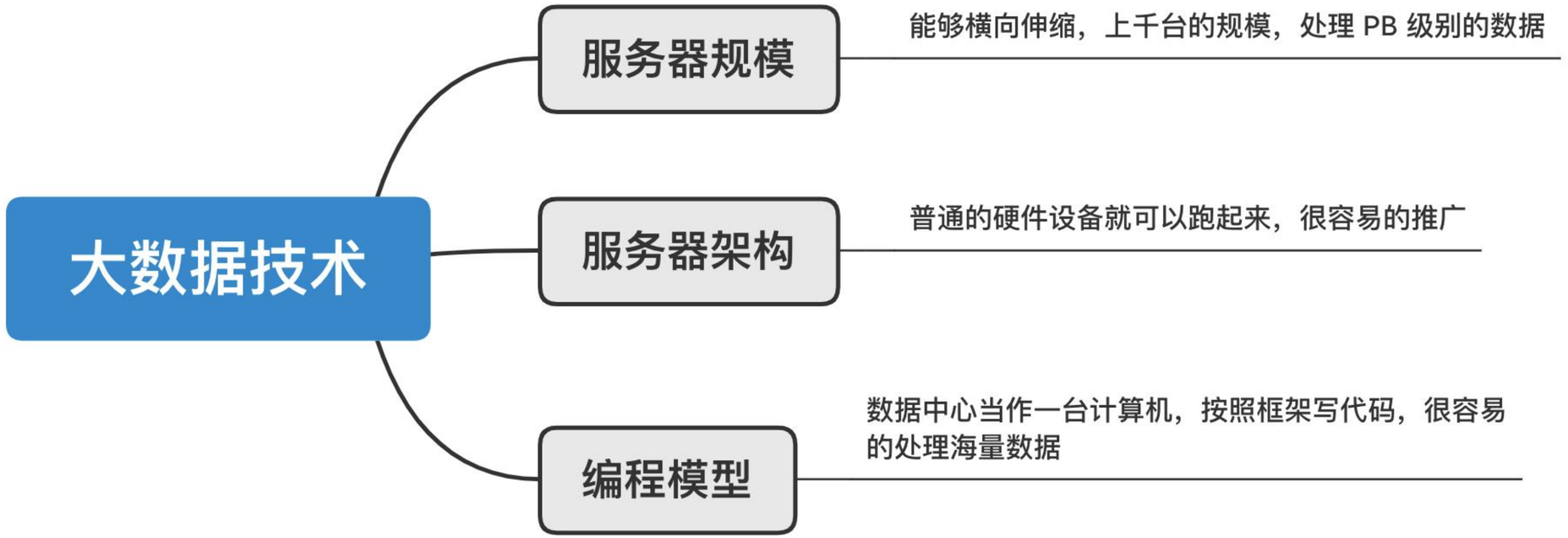
目录

- 1. 大数据核心理念与特点
- 2. 大数据起源与发展
- 3. 大数据处理架构：Lambda & Kappa
- 4. 大数据概念科普：数据仓库、数据湖、大数据平台、数据中台、湖仓一体
- 5. 大数据未来趋势和热点技术

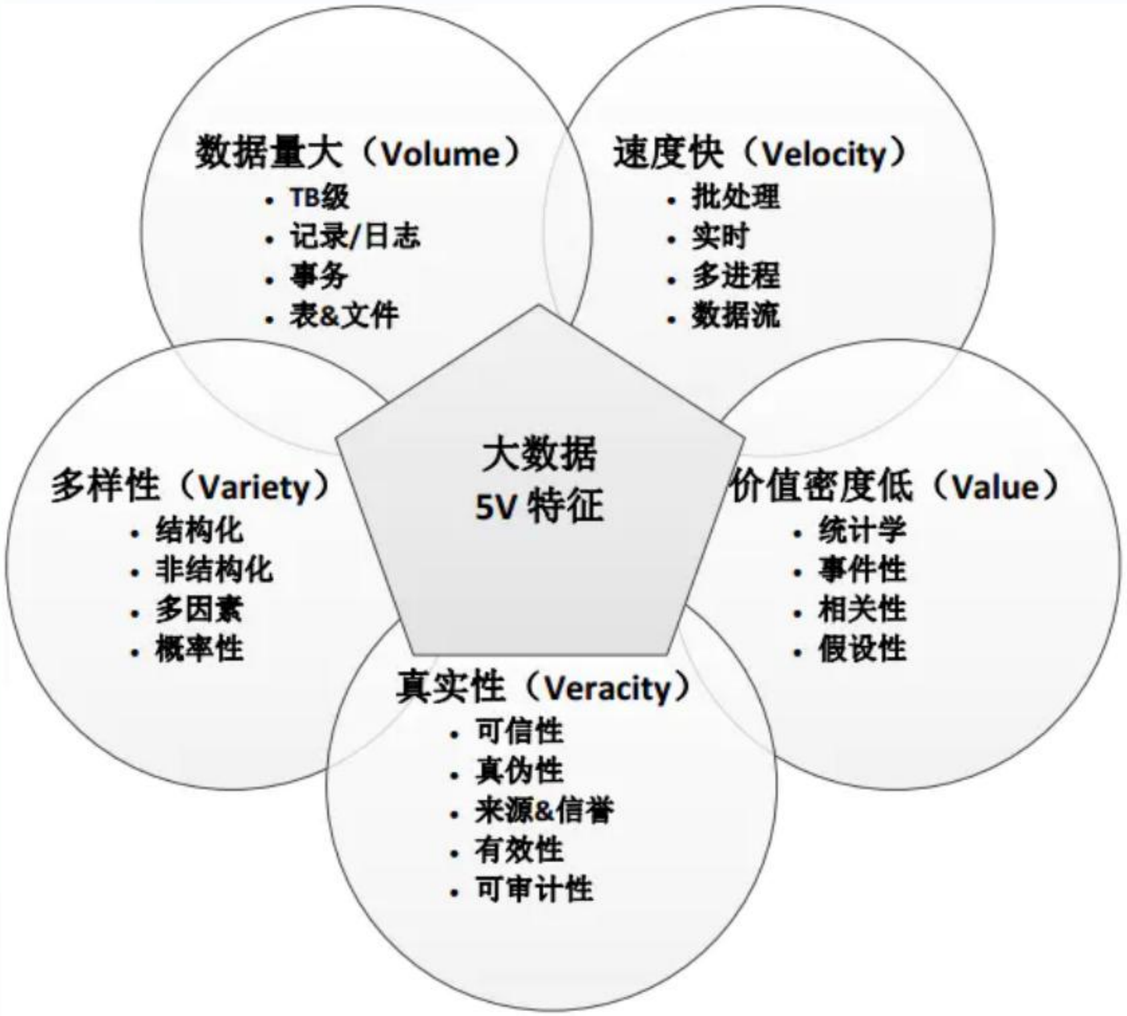
大数据核心理念

“大数据”是指传统数据处理应用软件时，不足以处理的大的或者复杂的数据集的术语。

-- Wikipedia



大数据特点



2001 年麦塔集团(META Group)分析师莱尼提出

3V：Velocity、Volume、Variety。

4V：3V + Value，总体价值大但价值密度低。

5V：4V + Veracity

大数据起源与发展

谷歌的“三驾马车”开启了大数据时代，并指明了大数据的发展方向。

为什么是 谷歌？解决什么问题？

搜索引擎

1. 存储整个互联网的内容
2. 构建倒排索引
3. 解决随机读写和热点更新

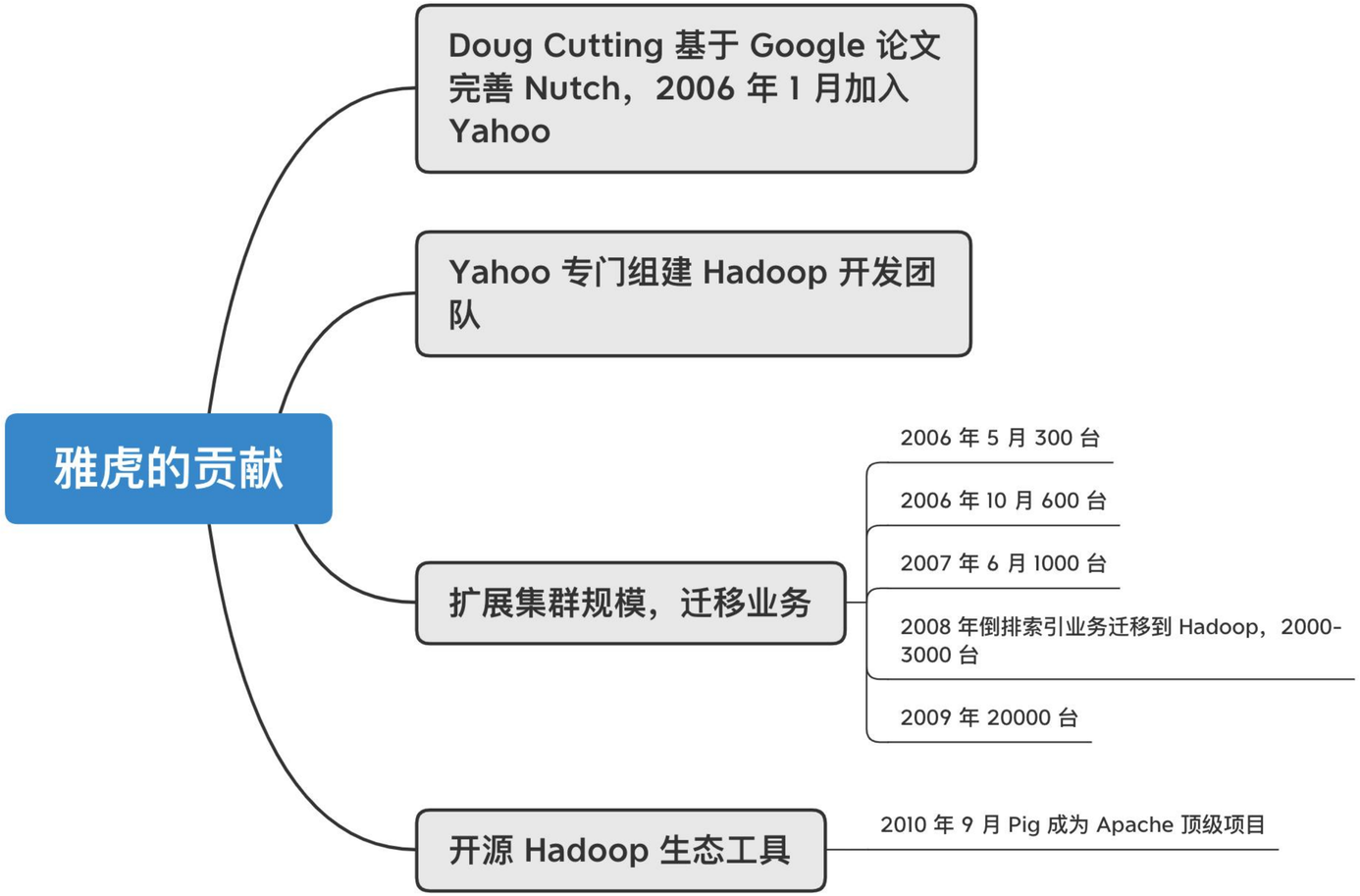
谷歌是大数据领域的先驱者和引领者

DataFlow：现代流计算的基石。

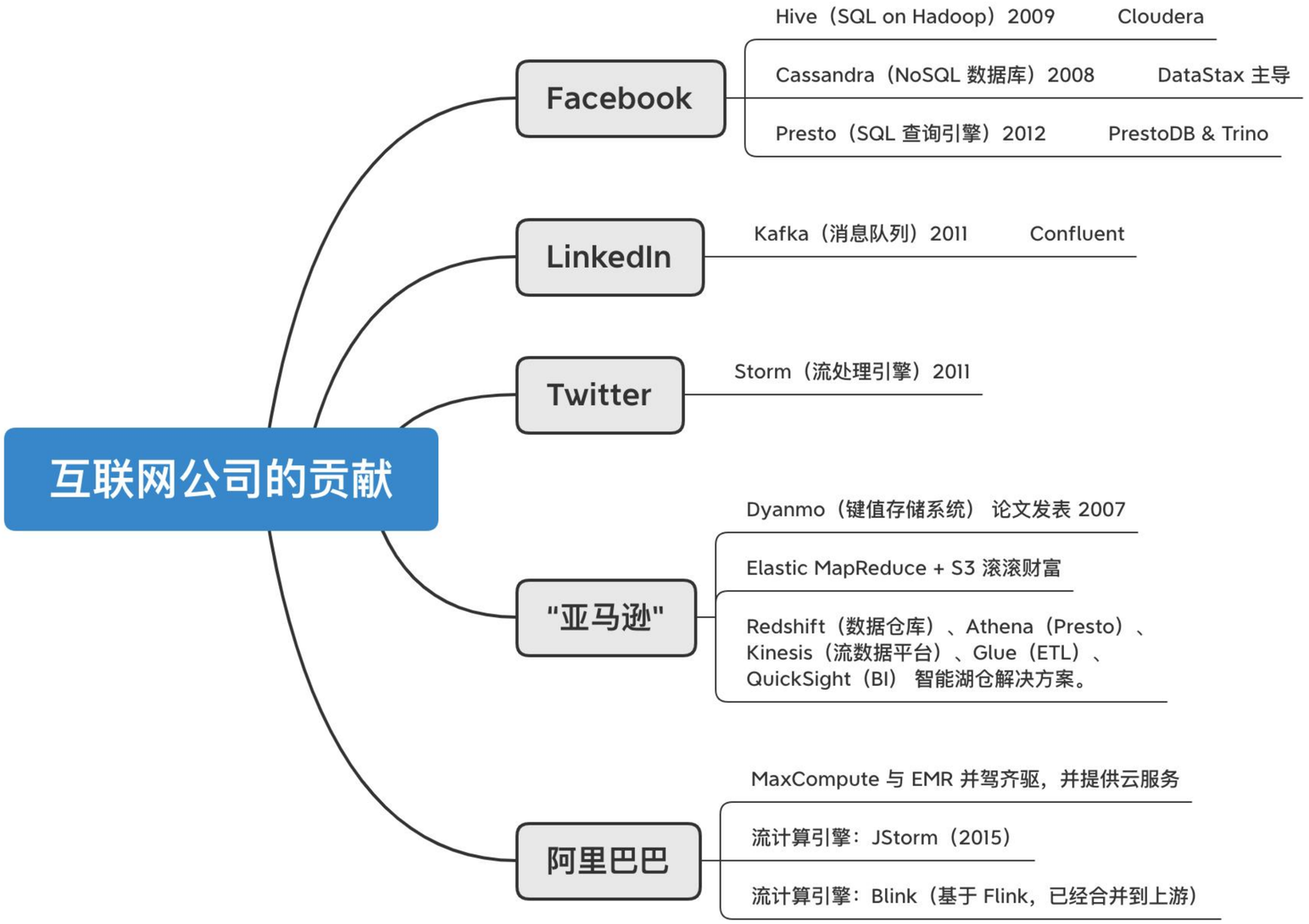
Kubernetes：容器编排的事实标准。

大数据起源与发展

雅虎的无私奉献和鼎力支持，成就了 Hadoop 繁荣生态。



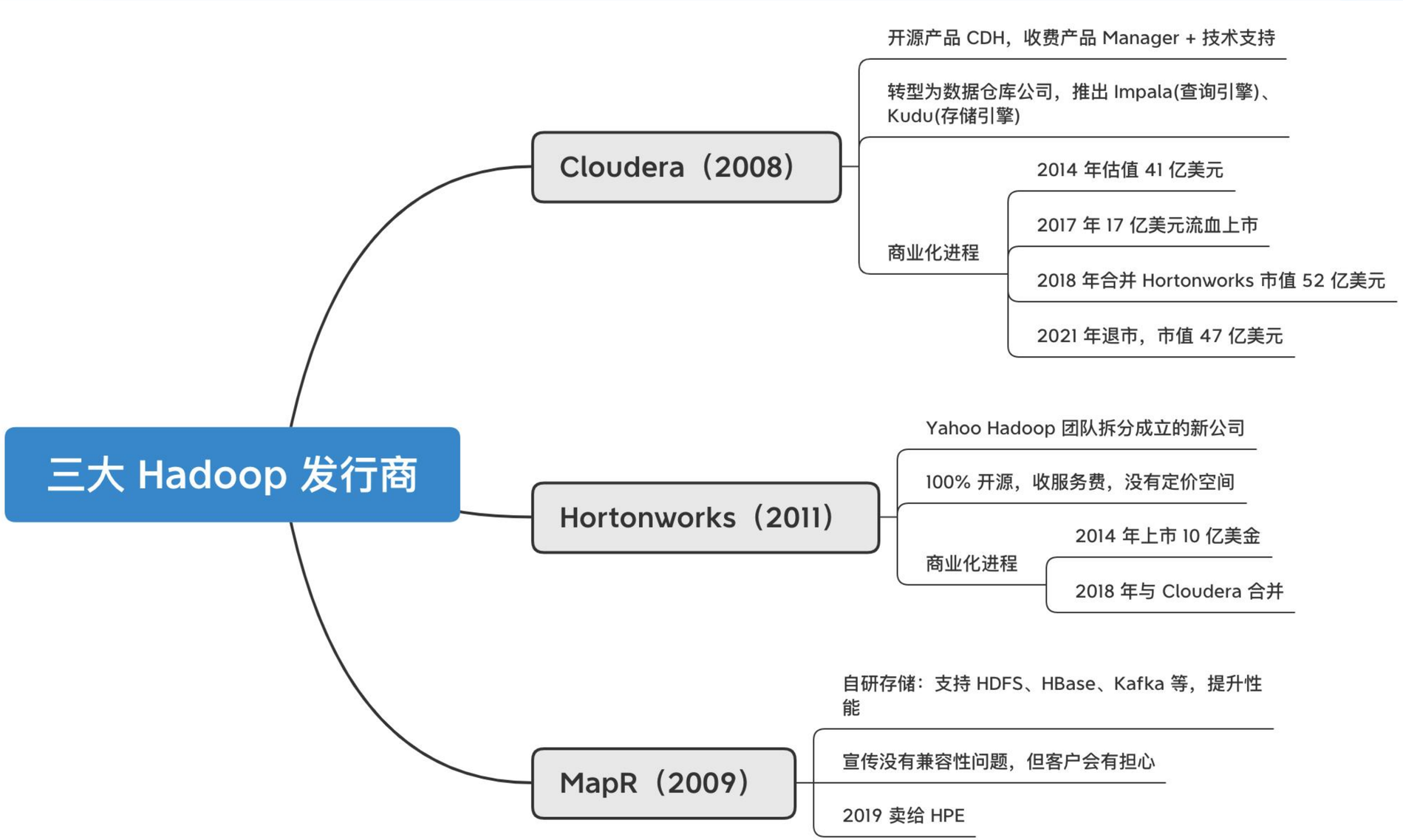
大数据起源与发展



不自建轮子，抱团取暖，
共同发展

要贡献回开源，否则就不
可能有完善可用的生态。

大数据起源与发展

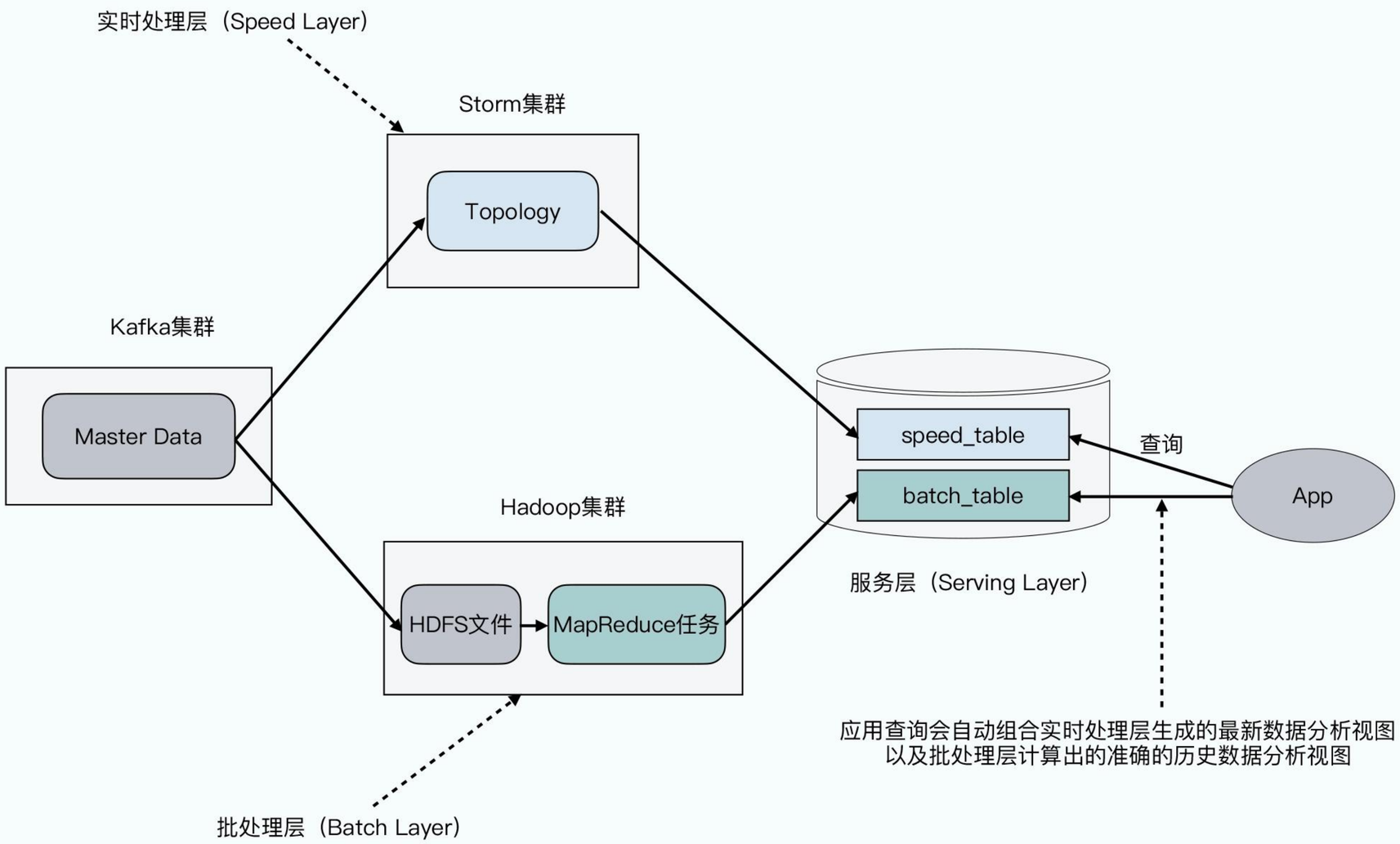


衰败的原因：

1. Hadoop 生态技术复杂，成本高。

2. 未能及时和正确的理解云商业，被云厂商抢夺了市场。

大数据处理架构：Lambda vs Kappa



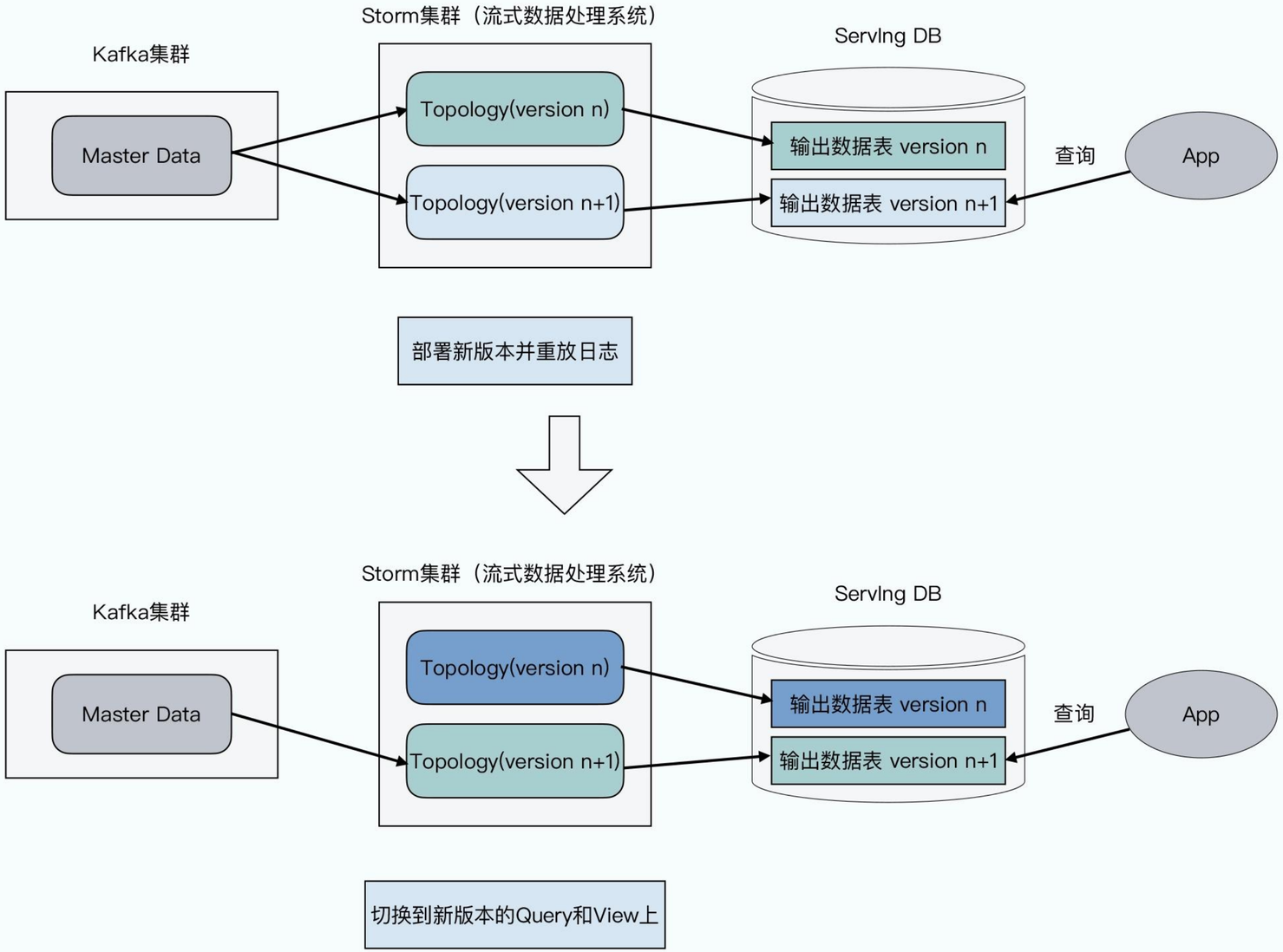
Lamda 架构，把大数据的批处理和实时数据结合在一起，变成一个统一的架构。

特点：

批处理解决流处理吞吐低和不准确的问题。

但需要两套框架和代码。并计算两次。

大数据处理架构：Lambda vs Kappa



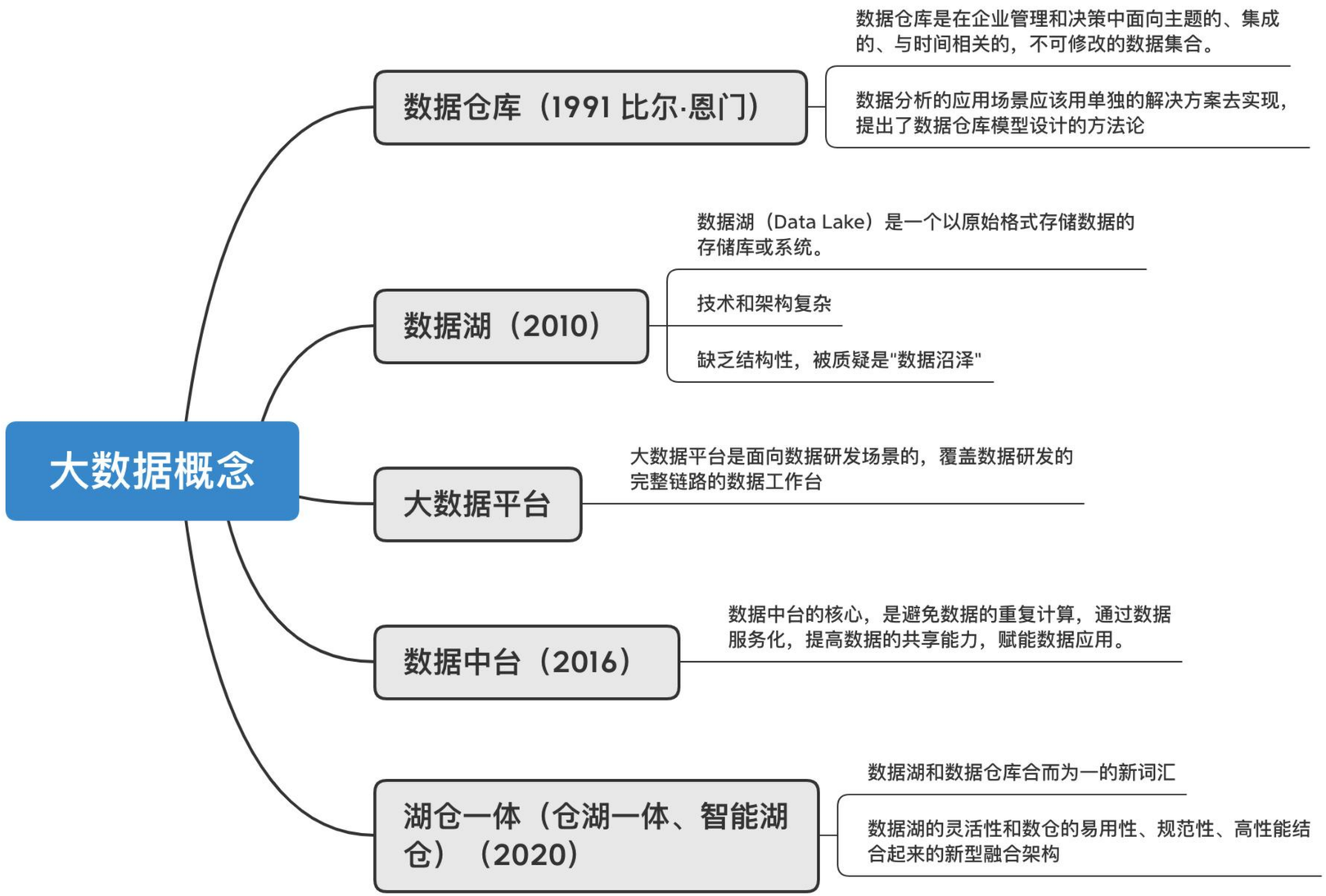
Kappa 架构去掉了 Lambda 架构的批处理层，而是在实时处理层，支持了多个视图版本。

特点：

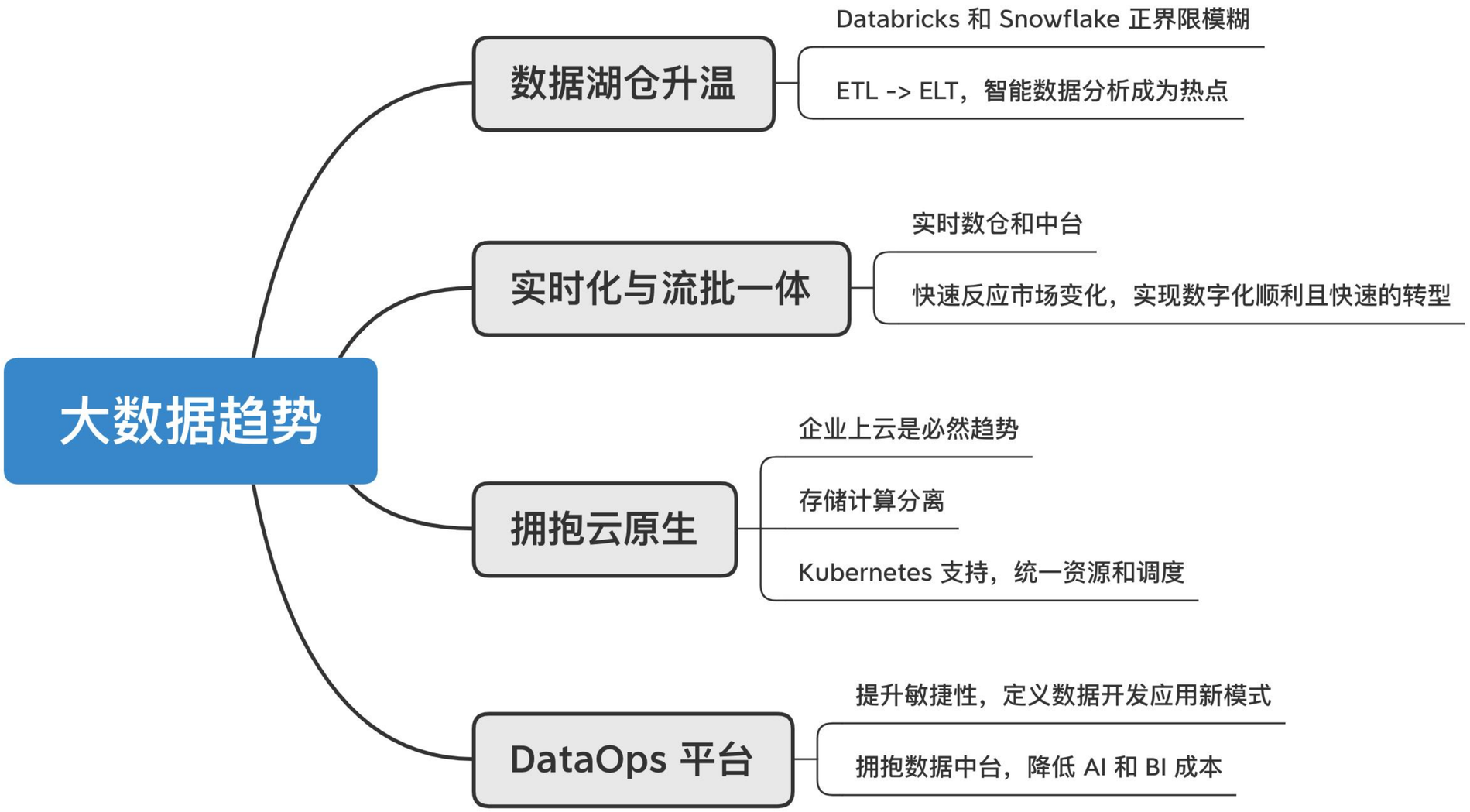
流批一体

Storm 逐步被 Spark（2010）（Databricks）和 Flink（2014）（Data Artisans）所取代。

大数据概念科普


大数据概念
数据仓库（1991 比尔·恩门）
数据仓库是在企业管理和决策中面向主题的、集成的、与时间相关的，不可修改的数据集合。
数据分析的应用场景应该用单独的解决方案去实现，提出了数据仓库模型设计的方法论
数据湖（2010）
数据湖（Data Lake）是一个以原始格式存储数据的存储库或系统。
技术和架构复杂
缺乏结构性，被质疑是"数据沼泽"
大数据平台
大数据平台是面向数据研发场景的，覆盖数据研发的完整链路的数据工作台
数据中台（2016）
数据中台的核心，是避免数据的重复计算，通过数据服务化，提高数据的共享能力，赋能数据应用。
湖仓一体（仓湖一体、智能湖仓）（2020）
数据湖和数据仓库合而为一的新词汇
数据湖的灵活性和数仓的易用性、规范性、高性能结合起来的新型融合架构

大数据未来趋势和热点技术



Thanks!