大数据技术导论

山丘

大数据系列课程

大数据技术演变

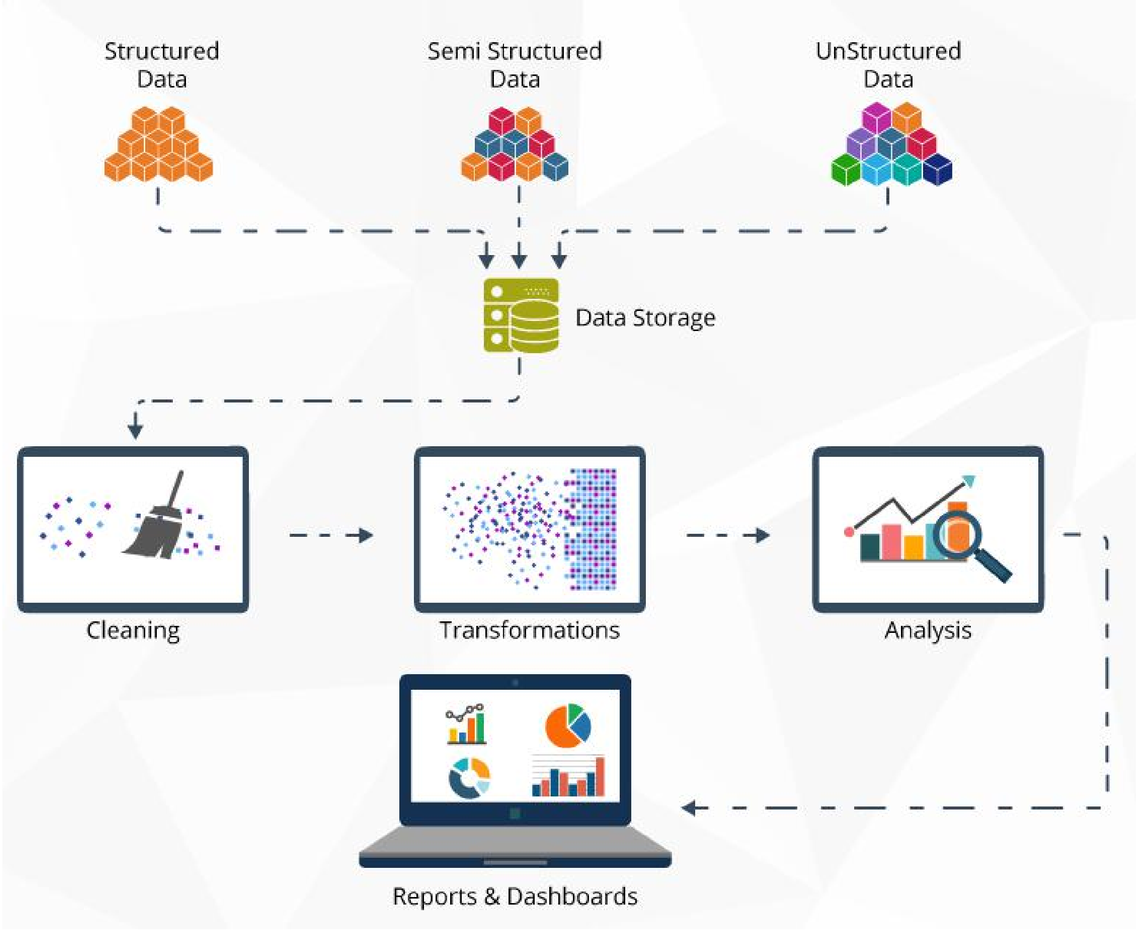
大数据技术演变

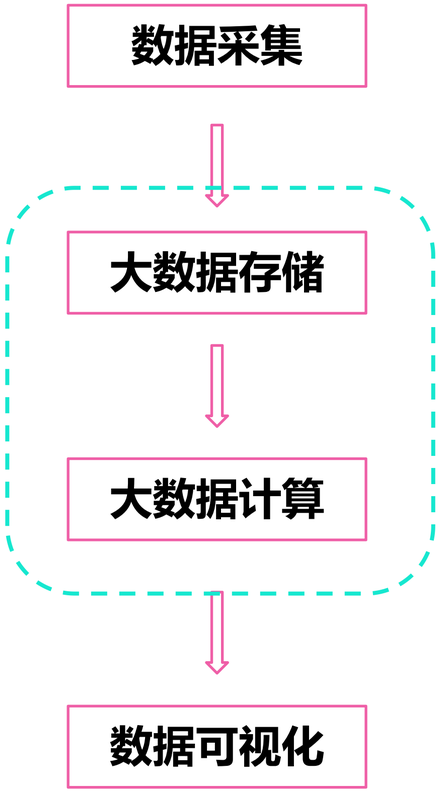
- **大数据技术定义：**
  - 是指伴随着大数据的采集、传输、存储、分析和应用的相关技术
  - 是一系列使用非传统的工具来对大量的结构化、半结构化和非结构化数据进行处理，从而获得分析和预测结果的一系列数据处理和分析技术。
  - 从数据分析全流程的角度，大数据技术主要包括数据采集与预处理、数据存储和管理、数据处理与分析、数据安全和隐私保护等机个层面的内容：

技术层面	功能
数据采集	利用ETL工具将分布的、异构数据源中的数据如关系数据、平面数据文件等，抽取到临时中间层后进行清洗、转换、集成，最后加载到数据仓库或数据集市中，成为联机分析处理、数据挖掘的基础；或者也可以把实时采集的数据作为流计算系统的输入，进行实时处理和分析
数据存储和管理	利用分布式文件系统、数据仓库、关系数据库、NoSQL数据库、云数据库等，实现对结构化、半结构化和非结构化海量数据的存储和管理
数据处理与分析	利用分布式并行编程模型和计算框架，结合机器学习和数据挖掘算法，实现对海量数据的处理和分析；对分析结果进行可视化呈现，帮助人们更好地理解和分析数据
数据隐私和安全	在从大数据中挖掘潜在的巨大商业价值和学术价值的同时，构建隐私数据保护体系和数据安全体系，有效保护个人隐私和数据安全

大数据技术演变

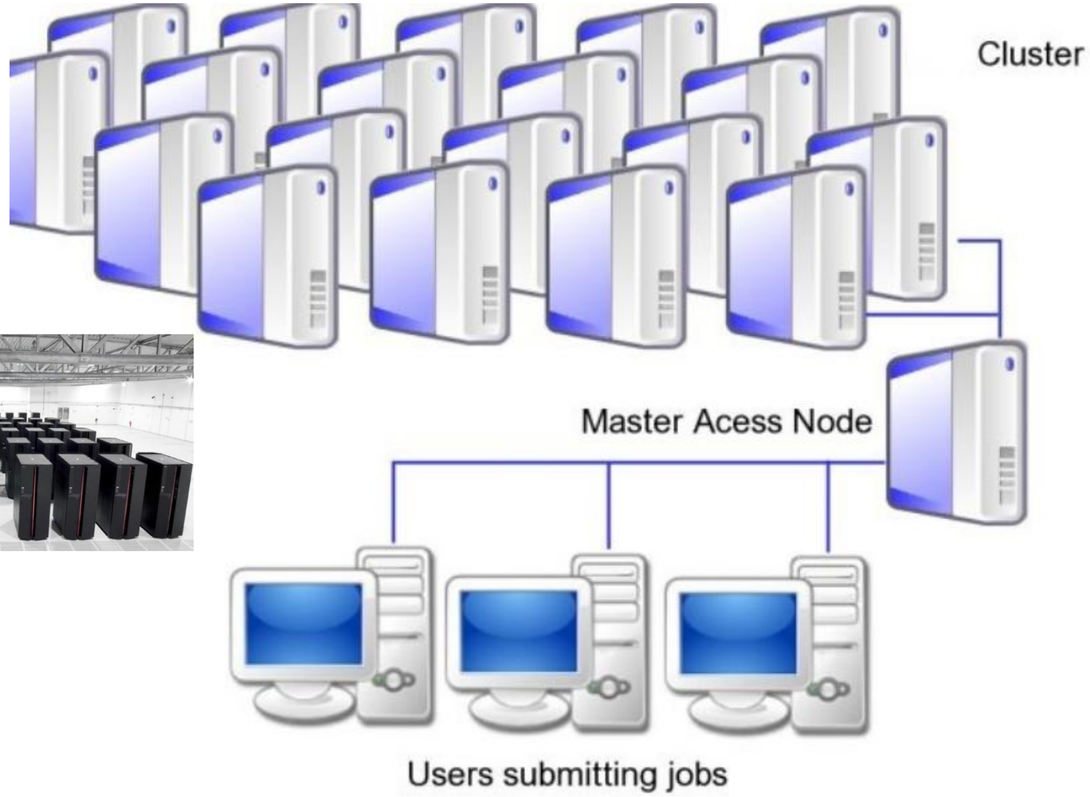
- 大数据的处理/分析流程：

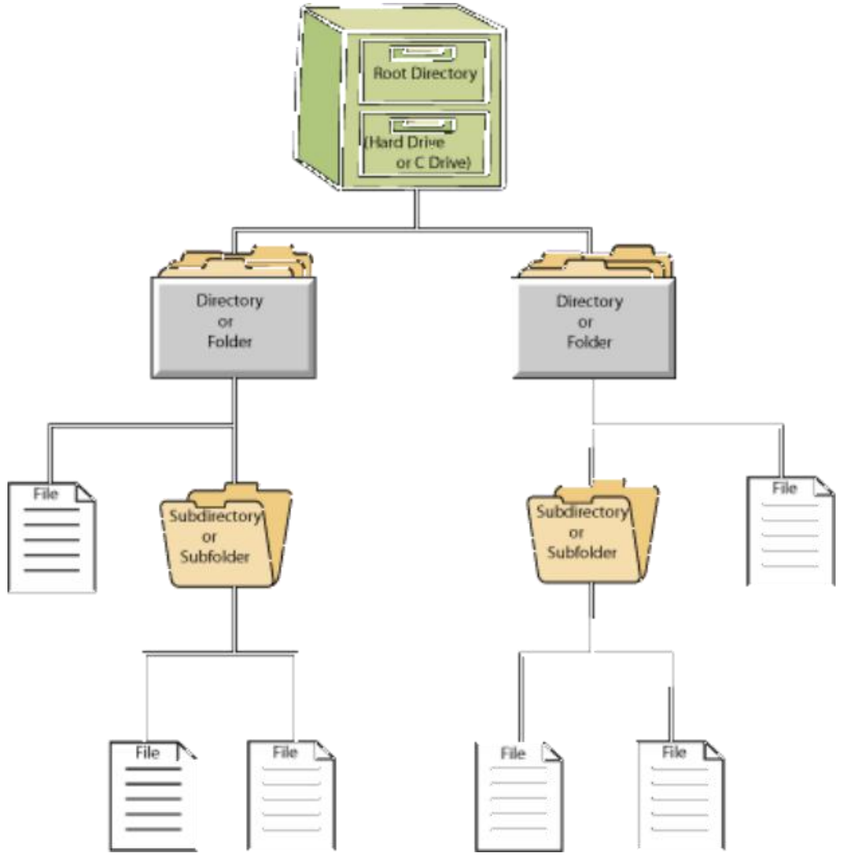




大数据技术演变

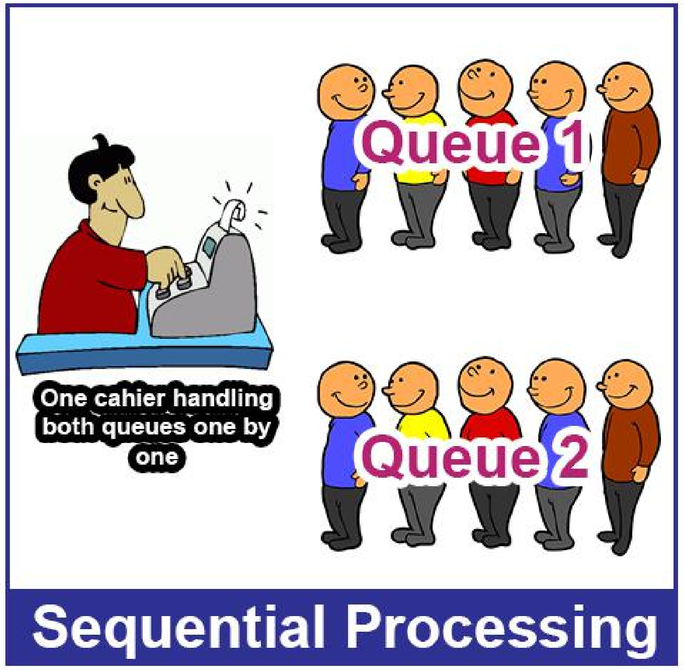
- **如何存储大数据？**
  - 分布式存储 & 分布式文件系统(DFS)

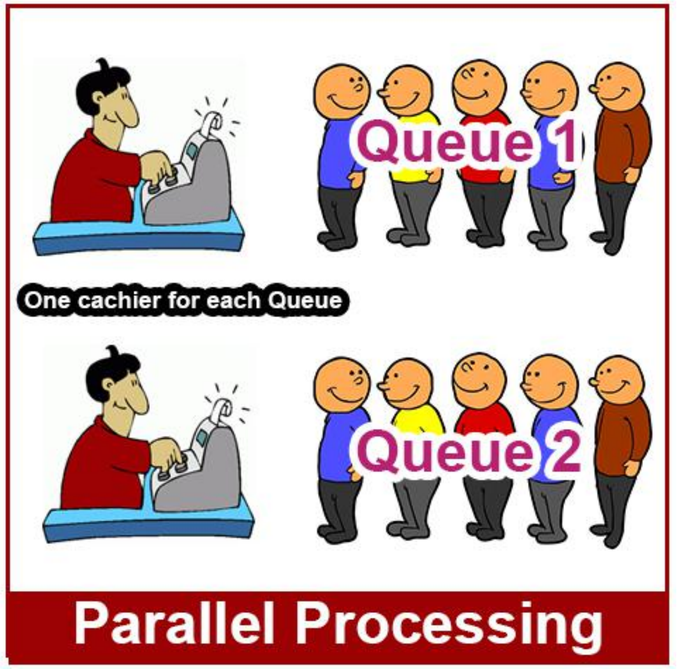




大数据技术演变

- **如何处理大数据？**
  - 顺序处理 vs 并行处理（计算）
  - 顺序处理：移动数据； 并行处理：移动计算。



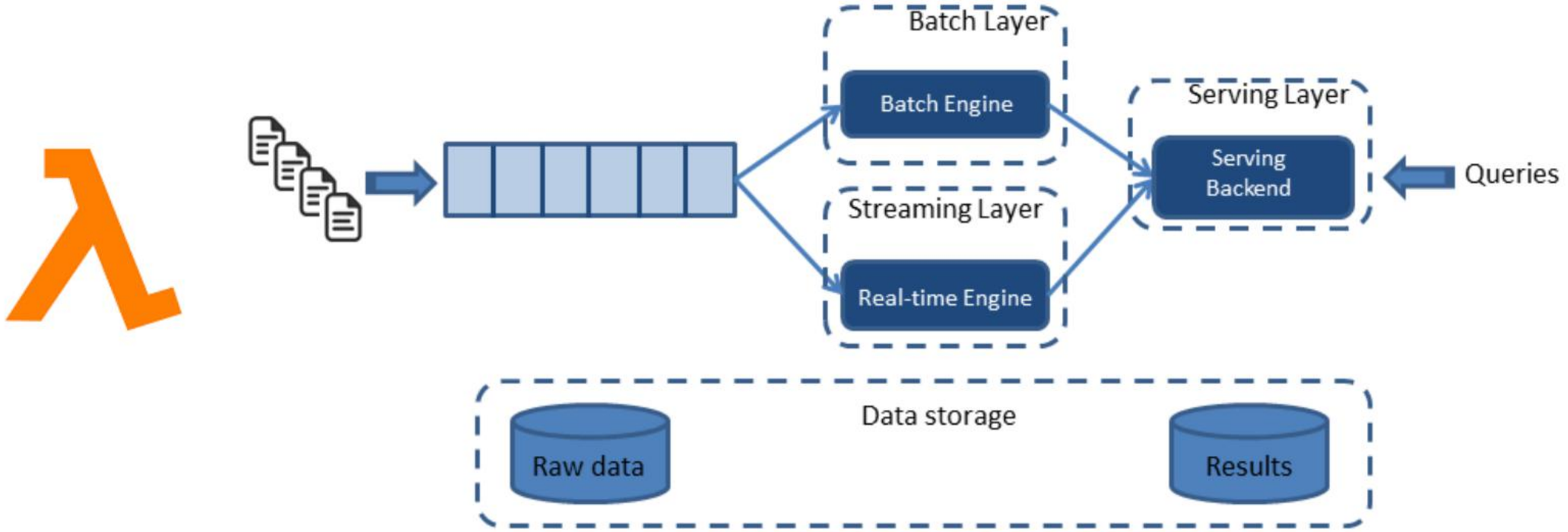


大数据技术演变

企业如何构建大数据处理技术平台?

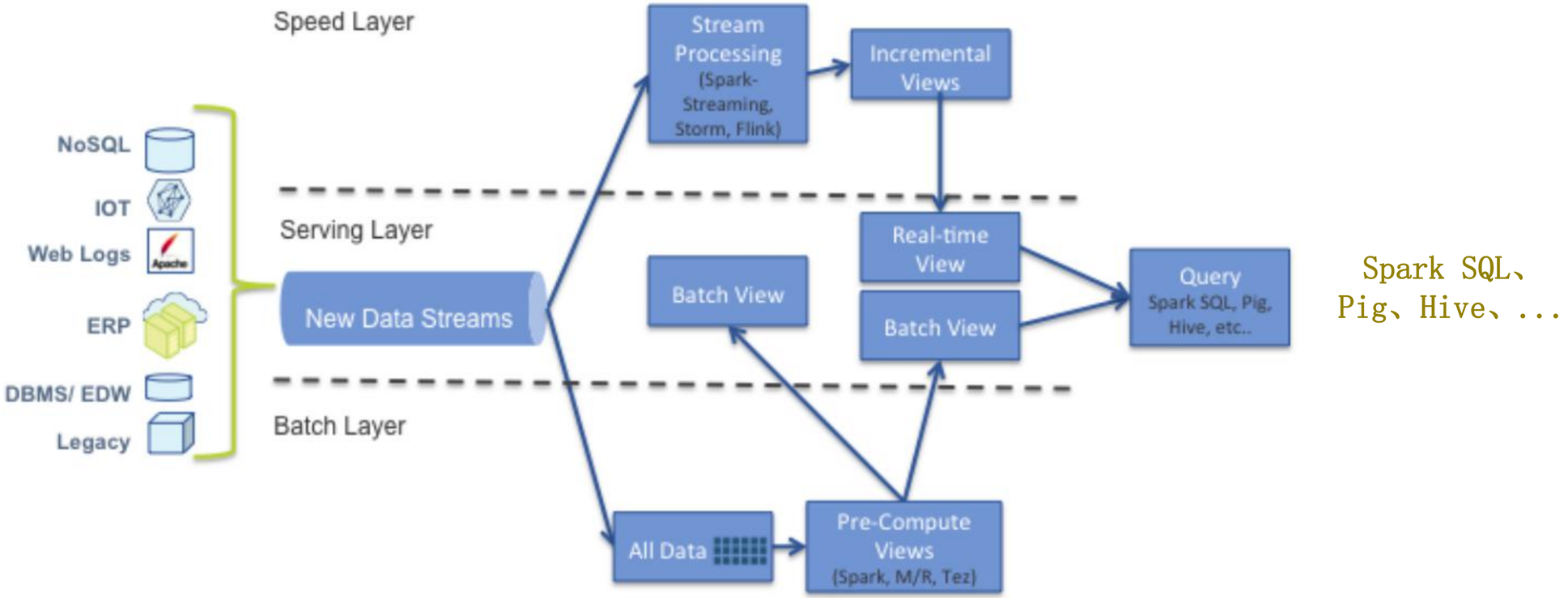
大数据技术演变

- 大数据Lambda体系架构
  - **Lambda架构是一种数据处理架构，通过使用批处理和流处理方法来处理大量数据(即“大数据”)。- Nathan Marz**



大数据技术演变

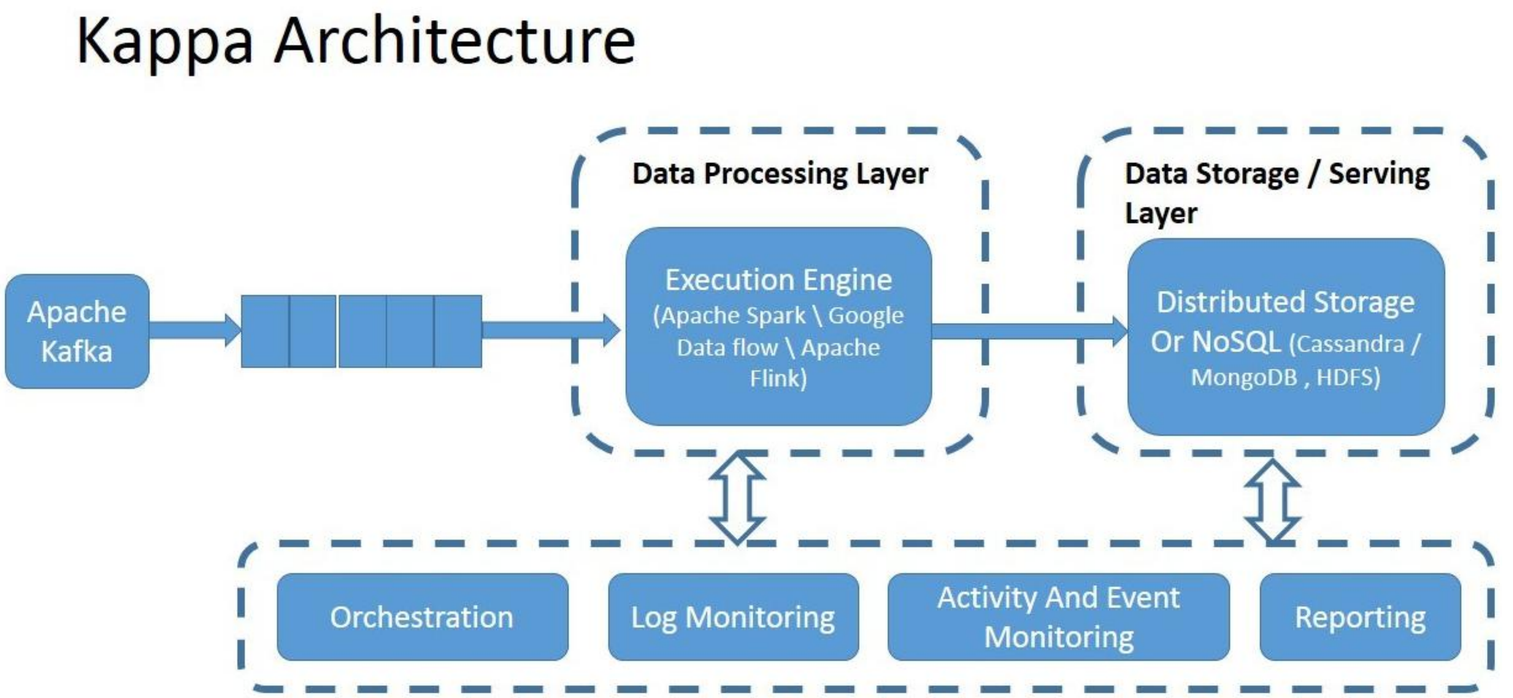
- 大数据Lambda体系架构
  - Lambda架构实现技术

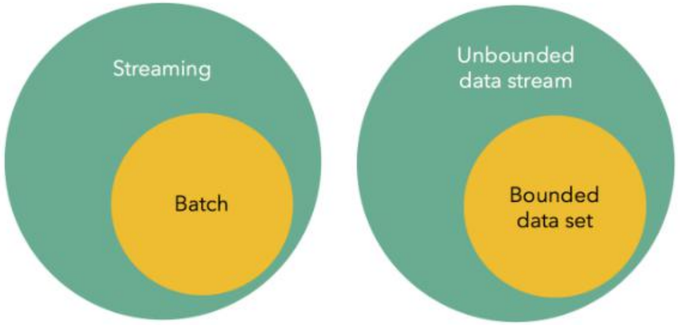


大数据技术演变

- Kappa体系架构—Lambda架构的简化
  - Kappa体系结构系统是去掉批处理系统的体系结构。
  - 为了取代批量处理，数据只需通过流系统快速地输入

Jay Kreps





大数据技术演变

- **Lambda架构实现进化路线**
  - 批处理 => 批流结合 => 流批结合

代际划分	计算引擎	特点
第一代	Hadoop 承载的 MapReduce	将计算分为两个阶段，分别为 Map 和 Reduce
第二代	Tez 以及更上层的 Oozie	支持 DAG 的框架，大多还是批处理的任务
第三代	Spark	Job 内部的 DAG 支持，以及强调的实时计算
第四代	Flink	对流计算的支持，以及更一步的实时性上面

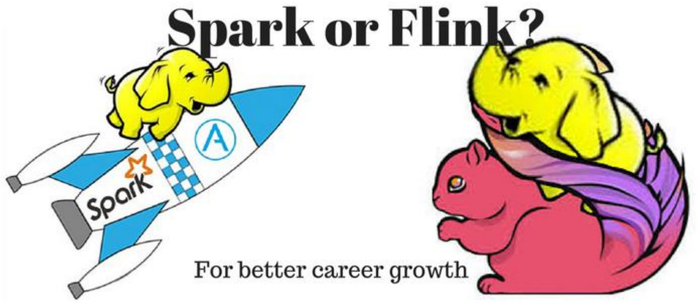
大数据技术

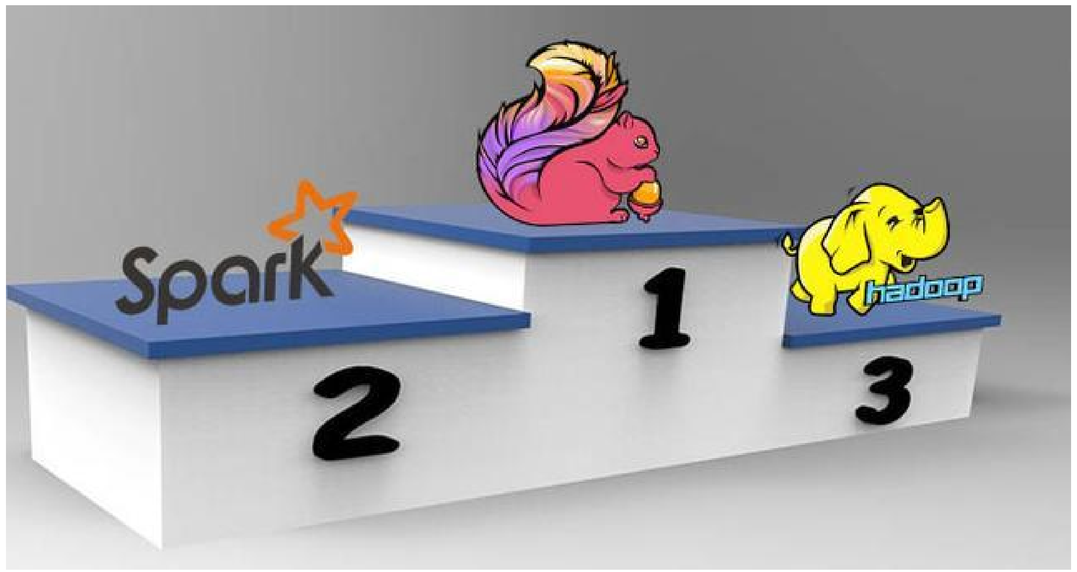
- **大数据计算模式及其代表产品：**

大数据计算模式	解决问题	代表产品
批处理计算	针对大规模数据的批量处理	MapReduce、Spark等
流计算	针对流数据的实时计算	Flink、Storm、Kafka、S4、Flume、Streams、Puma、DStream、Super Mario、银河流数据处理平台等
图计算	针对大数据图结构数据的处理	Pregel、GraphX、Giraph、PowerGraph、Hama、GoldenOrb等
查询分析计算	大规模数据的存储管理和查询分析	Dremel、Hive、Cassandra、Impala等

Flink简介

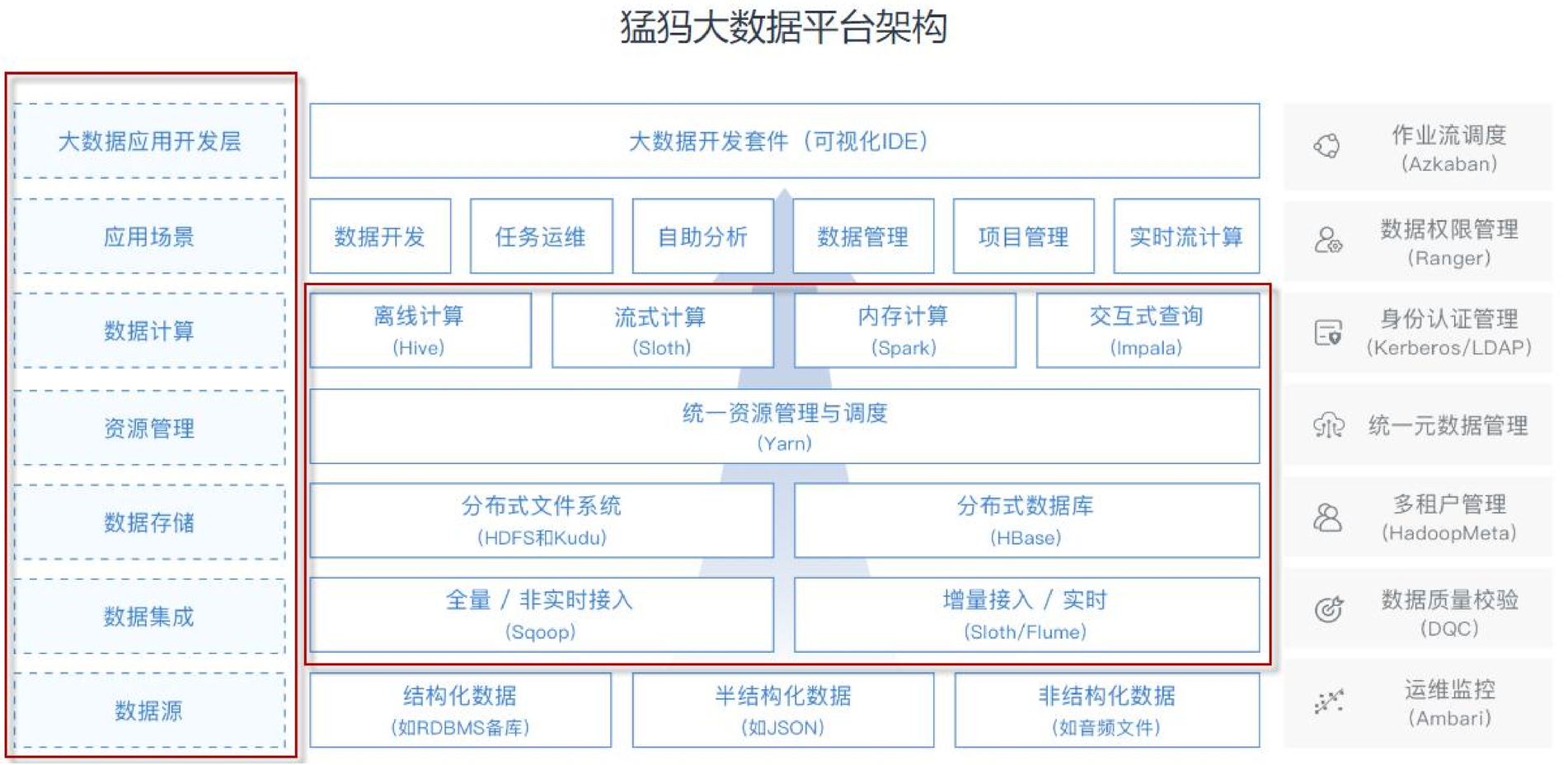
- Flink vs Spark vs Hadoop：
  - 实时计算，首选 Flink
  - 大数据离线分析，首选 Spark
  - 大数据机器学习，首选 Spark





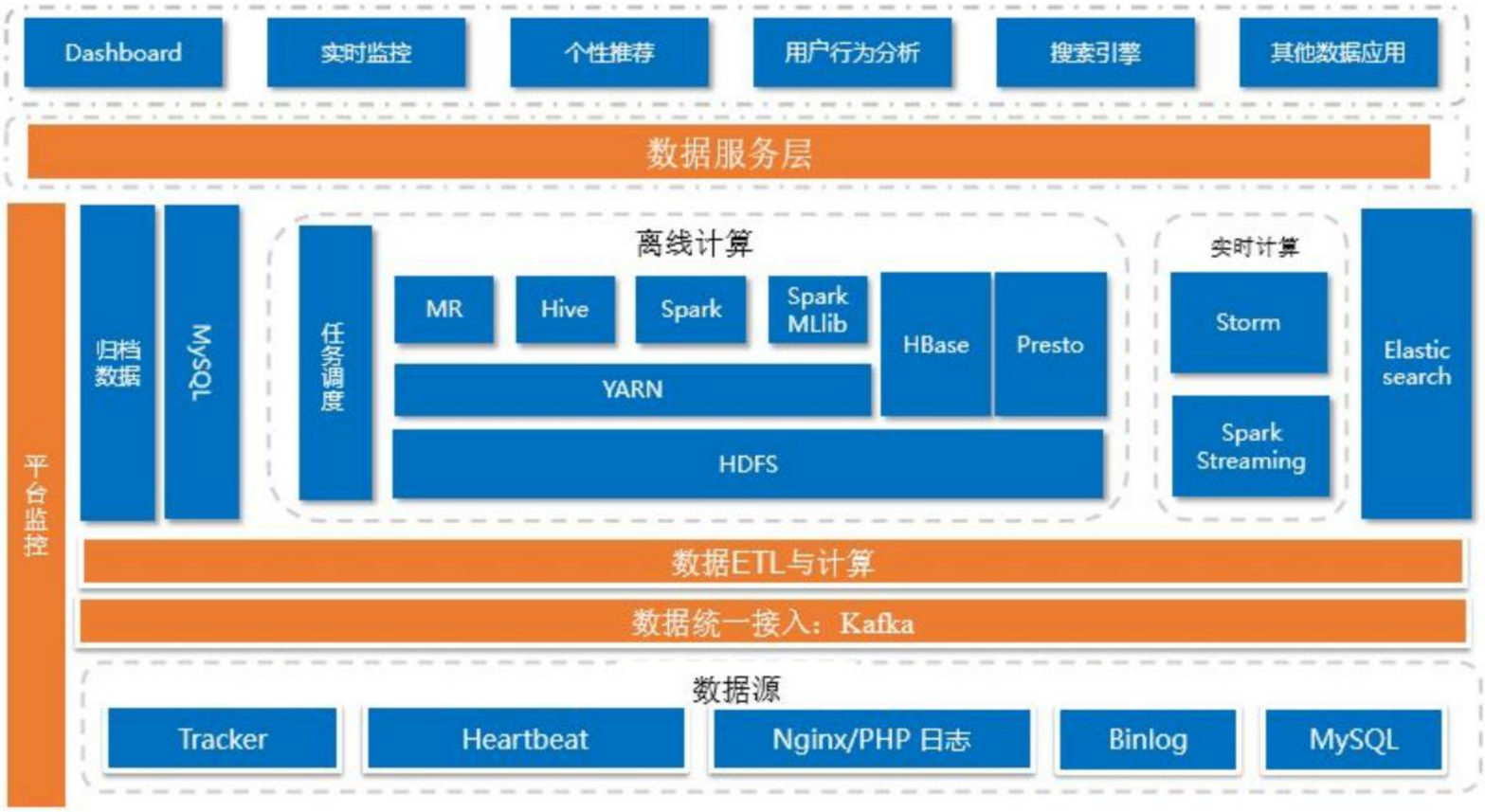
大数据技术应用

- 网易猛犸大数据平台架构：



大数据技术应用

- 斗鱼大数据平台架构：



大数据岗位方向：

大数据运维

大数据开发

大数据分析