

图 6-23 2017 大数据产业全景图

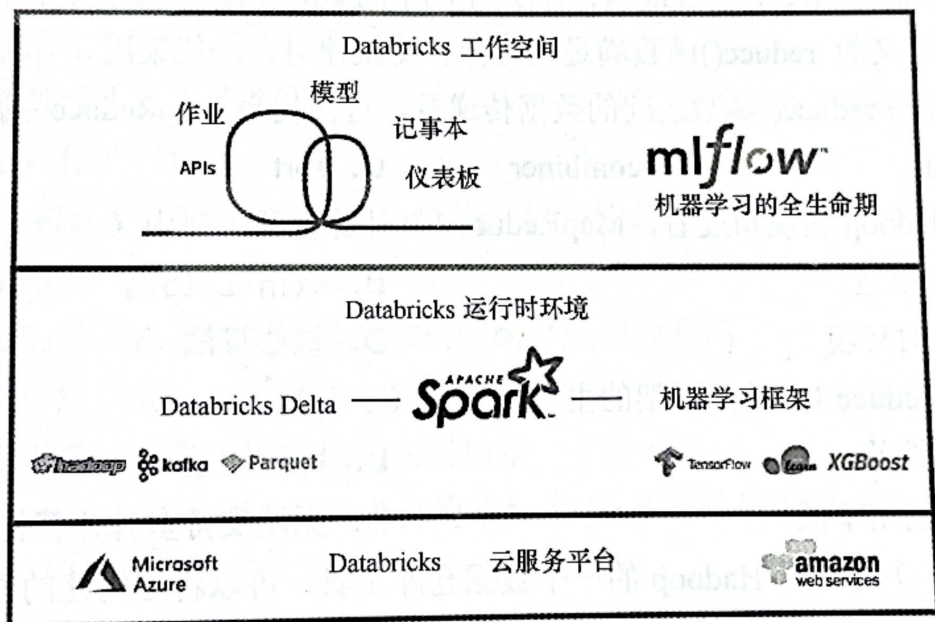


图 6-24 基于 Databricks 的统一分析平台

习 题

一、选择题

1. 《信息技术—大数据技术参考模型》中给出了(), 描述了大数据的参考框架, 包括角色、活动和功能组件及它们之间的关系。
 - A. 大数据技术参考模型
 - B. 大数据术语体系
 - C. 大数据流程
 - D. 大数据生命周期



扫描全能王 创建

2. () 计算框架源自一种分布式计算模型, 计算过程分为两个阶段——Map 阶段和 Reduce 阶段, 并分别以两个函数 map() 和 reduce() 进行抽象。

- A. Spark B. Tez C. MapReduce D. Flink

3. Google 于 2003—2008 年发表的 3 篇论文在云计算和大数据技术领域产生了深远影响, 通常被称为 Google 三大云计算技术的包括 ()。

- A. BigTable B. GFS C. MapReduce D. Spark

4. 当用户程序调用 MapReduce 框架时, 将输入文件分成 M 个 ()。

- A. 切片 B. 分片 C. 分块 D. 分区

5. 以下特征中, 不属于 MapReduce 的特征是 ()。

- A. 以主从结构的形式运行 B. key-value 类型的输入、输出
C. 容错机制的简单性 D. 数据存储位置的多样性

6. 通常, “落伍者” 是影响 MapReduce 总执行时间的主要影响因素之一。为此, MapReduce 中采用 ()。

- A. 推测性执行的任务备份机制 B. 惰性计算
C. 急性/热情计算 D. 分布式计算

7. 在 MapReduce 中, 当 map() 函数产生的中间 key 值的重复数据会占很大的比重, 而且用户自定义的 reduce() 函数满足结合律和交换律时, 一般采用 () 函数来降低 map() 函数与 reduce() 函数之间的数据传递量, 进而提高 MapReduce 的处理速度。

- A. filter B. combiner C. sort D. filter

8. 从 Hadoop 实现角度看, MapReduce 1.0 计算框架主要由 () 部分组成。

- A. 编程模型 B. 数据处理引擎
C. 运行时环境 D. 核心算法

9. MapReduce 1.0 计算框架的主要缺点是 ()。

- A. 扩展性差 B. 可靠性差
C. 资源利用率低 D. 无法支持多种计算框架

10. () 是基于 Hadoop 的一个数据仓库工具, 可以将结构化的数据文件映射为一张数据库表, 并提供简单的查询功能。

- A. Hive B. HBase C. Flink D. Spark

11. 以下特征中, 不属于 Spark 的特征是 ()。

- A. 支持惰性计算 B. 内存计算
C. 不支持交互式处理 D. 支持图计算

12. Spark 的技术架构可以分为 () 3 层。

- A. 资源管理层 B. Spark 核心层 C. 服务层 D. 搜索层

13. Spark 提供了较为灵活的集群管理模式, 包括 ()。

- A. Standalone 模式 B. Mesos 模式
C. YARN 模式 D. Message 模式



14. 在 Spark 中, () 是一个容错的、并行的数据结构, 允许用户显式地将数据存储在磁盘和内存中, 并能控制数据的分区。

- A. RDD B. PDD C. KDD D. CDD

15. Spark 中的一个关键问题是如何选择 RDD 序列化时机。通常, 在 () 情况下, 可以考虑对其进行序列化处理。

- A. 在完成成本比较高的操作之后
B. 在执行容易失败的操作之前
C. 当 RDD 被重复使用或者计算其代价很高时
D. 对计算速度的要求很高时

16. () 指父 RDD 的每个分区都只被子 RDD 的一个分区所依赖。

- A. 窄依赖 B. 宽依赖
C. 长依赖 D. 短依赖

17. Storm 创始人内森·马兹 (Nathan Marz), 结合自己在 Twitter 和 BackType 从事大数据处理的工作经验, 提出了一种大数据系统参考架构—— ()。

- A. Storm+架构 B. MapReduce+架构
C. Lambda 架构 D. Twitter+架构

18. 在大数据时代, 传统数据库的优点是 ()。

- A. 善于处理大数据的读写操作 B. 适用于数据模型不断变化的应用场景
C. 频繁操作的代价低 D. 产品成熟度高

19. 以下描述中, 不属于 NoSQL 数据库优势的是 ()。

- A. 易于数据的分散存储与处理
B. 数据的频繁操作代价低以及数据的简单处理效率高
C. 数据一致性高
D. 适用于数据模型不断变化的应用场景

20. NoSQL 数据库中对数据管理目的, 尤其是数据一致性保障问题的认识发生了变化, 而这些变化以以下哪两个重要理论为依据 ()。

- A. CAP 理论 B. BASE 原则
C. LA 架构 D. 奥康姆剃刀原则

21. Analytics 3.0 的主要特点有 ()。

- A. 引入嵌入式分析
B. 重视行业数据, 而不只是企业内部数据
C. 以产品与服务的优化为主要目的
D. 注重规范性分析

22. Analytics 1.0 的主要特点有 ()。

- A. 分析活动滞后于数据的生成 B. 重视结构化数据的分析
C. 以对历史数据的理解为主要目的 D. 注重描述性分析



23. Analytics 2.0 的主要特点有 ()。

- A. 分析活动几乎与数据的生成同步, 强调数据分析的实时性
- B. 重视非结构化数据的分析
- C. 以决策支持为主要目的
- D. 注重解释性分析和预测性分析

二、调研与分析题

1. 调查并对比分析大数据技术领域的国际顶级会议及学术期刊。
2. 调查分析基于 Spark 的大数据分析的现状与趋势。
3. 调查并对比分析常用 NoSQL 工具。
4. 调查并对比分析常用大数据分析工具。
5. 调查并对比分析常用大数据管理工具。
6. 调查并对比分析常用 Spark 与 MapReduce 的区别。
7. 结合自己的专业领域, 调研该领域常用的大数据技术与工具。

