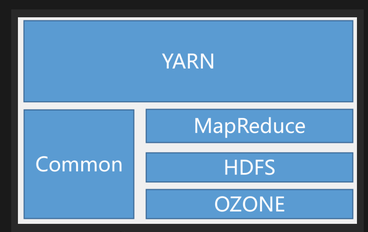
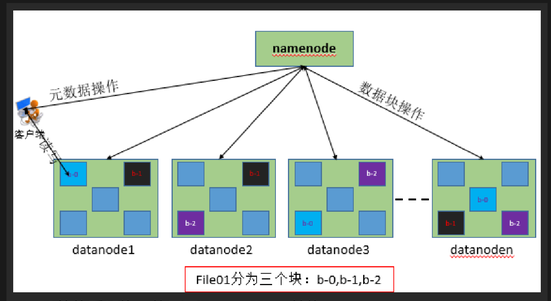
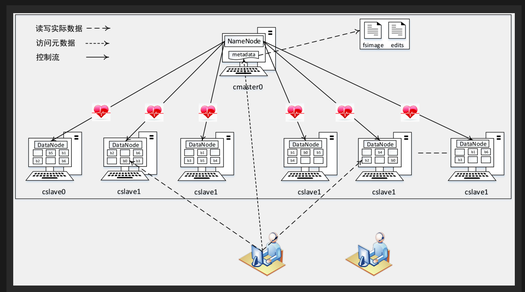
Hadoop3.0

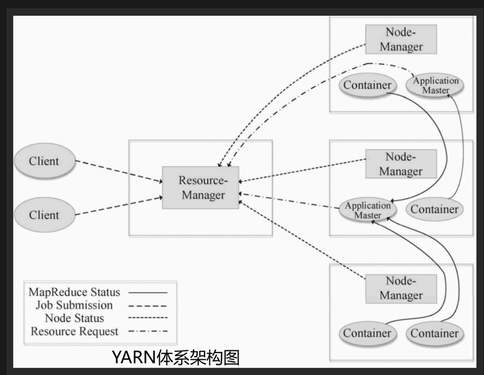
- Hadoop是Google云计算的开源实现
- Hadoop组成 掌握
  - Common — 支持Hadoop其他模块的公共组件
  - HDFS — Hadoop使用的分布式文件系统
  - YARN — 分布式资源管理和任务调度器
  - Hadoop MapReduce — 分布式计算框架
  - Ozone — 分布式对象存储系统

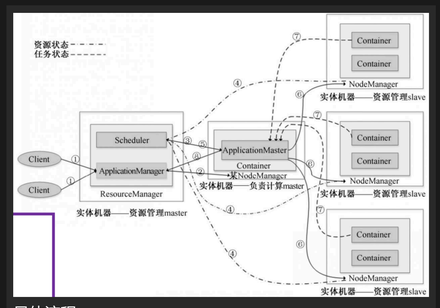


Hadoop3.0构成

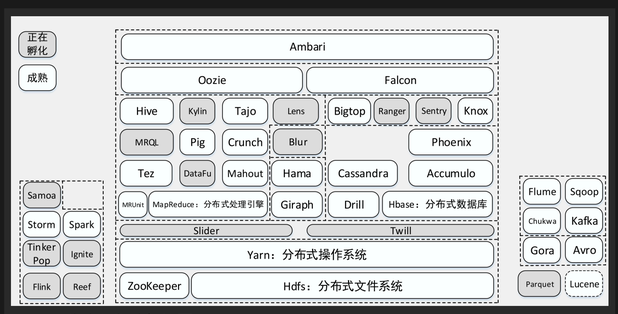
- HDFS
  - Hadoop使用的分布式文件系统
    - 

      HDFS的体系架构，使用了Master/Slave结构
      - 主从架构能提供高可扩展性并简化架构设计
      - 主进程是NameNode，负责存储集群的元数据从而管理存储空间
      - 从进程是DataNode，负责储存数据
    - 体系架构 理解
      - 

        逻辑架构，展示了数据流动，是对体系架构的逻辑抽象
        - 用户访问元数据进行读写操作
        - NameNode发送控制流信号给DataNode
        - DataNode将数据发给用户
    - 拓扑
      - 一般拓扑 — 单NameNode节点
      - 商业拓扑 — 双NameNode节点
- YARN
  - 体系架构 理解
    - 

      YARN体系架构图
      分布式操作系统YARN的架构图，同样使用Master/Slave架构
      - 主进程是ResouceManager，管理整个集群的资源仲裁
      - 从进程是NodeManager，管理本地计算机资源
  - 执行流程 了解
    - 

      具体流程
    - 相关实体
      - Client — 客户端，用户使用的应用
      - RM
        - ResourceManager，主进程，资源仲裁者
        - Scheduler — RM中的资源仲裁模块
        - AppManager — RM中的应用程序操作模块，选定、启动一些AppMaster
      - NM
        - NodeManager，从进程，管理本机资源，执行任务的地方
        - Container
          - 容器，存放本机资源
          - AppMaster — 任务执行、监管中心

- Hadoop发展史 练习题
  - 1.0
    - 实现了MapReduce与HDFS
    - 对其他计算框架支持差，计算性能存在瓶颈，于是需要升级
  - 2.0
    - 引入了YARN资源管理框架，将资源与计算解耦，支持非MapReduce应用
    - NameNode的单点故障问题仍然存在，为了提高可用性与资源利用率，需要继续升级
  - 3.0
    - 引入多NameNode、纠删码从而提高了容错与存储效率
    - 运维、配置相较前两版复杂了很多
- Hadoop Common 了解
  - 公共组件，为Hadoop其他模块提供公用API，降低Hadoop设计复杂性，减少模块间耦合性，增强Hadoop健壮性
  - 功能接口
    - 公共编程接口
    - 微集群功能 — 为用户提供快速体验Hadoop集群的功能
    - 公平呼叫队列 — 解决传统消息队列先进先出造成的新请求长时间等待而无响应的问题
    - Hadoop本地库
    - 用户代理 — 提供模拟的权限供用户执行一些功能
    - 机架感知
    - 安全服务 练习题
      - 认证服务
        - 使用Kerberos统一集群认证，为用户提供安全性保障
        - 以HTTP认证提供Web页面的访问安全性保障
      - 权限与访问控制
        - ACL
        - 服务间Token
      - 存在问题
        - 配置复杂，用户手动配置出错会造成漏洞
        - ACL可能被滥用导致错误给出极高权限于非预料用户
- Hadoop 访问接口 了解
  - Web接口
    - 配置项接口 — 配置Hadoop文件用的
    - 浏览器接口 — 访问页面用的
  - HDFS Shell接口 — 执行命令用的
  - WebHDFS REST API接口 — 网页通信用的
  - HDFS Java API接口 — 提供与HDFS交互方式里的编程方式，可以通过Java API编程进行文件上传等操作
- Hadoop生态组件 了解 练习题
  - 

    Hadoop生态的整体架构图
  - HDFS — 分布式存储
  - YARN — 分布式操作系统
  - MapReduce — 分布式处理算法
  - ZooKeeper — 分布式锁服务
  - 分布式数据库
    - HBase
    - Cassandra
  - 高层语言
    - Pig
    - Hive
  - 机器学习库
    - Mahout
    - Giraph
  - 数据传输工具
    - Flume
    - Avro
    - Chukwa
    - Sqoop
  - 集群管理工作 — Ambari
  - 组件间版本依赖处理工具 — BigTop