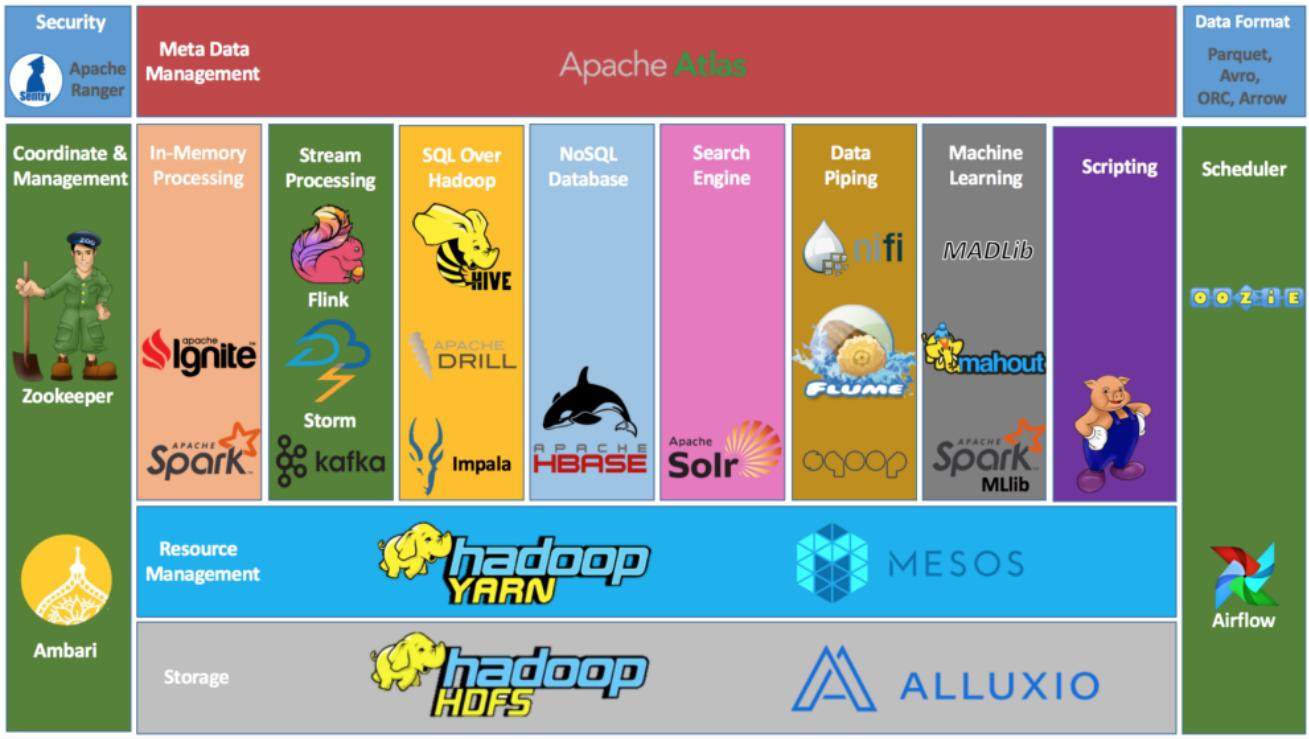
Hadoop介绍

- 狭义上:Hadoop指的是Apache一款Java开源软件,是一个大数据分析处理平台
- Hadoop HDFS:分布式文件系统.解决了海量数据存储问题
  - Hadoop Distributed File System (HDFS™)
- Hadoop==MapReduce:分布式计算框架==解决海量数据计算问题
  - parallel processing of large data sets.
- Hadoop==YARN:集群资源管理和任务调度
  - A framework for job scheduling and cluster resource management.
  - 资源指的是和程序运行相关的硬件资源
    - cpu ram内存
  - #任务调度
    - 集群资源繁忙的时候 如何分配资源给各个程序 调度
    - 调度的关键是策略：先来后到 权重
- 广义上:Hadoop指的是==Hadoop生态圈==
  - 提供了大数据的几乎所有软件
  - 采集、存储、导入、分析、挖掘、可视化、管理...



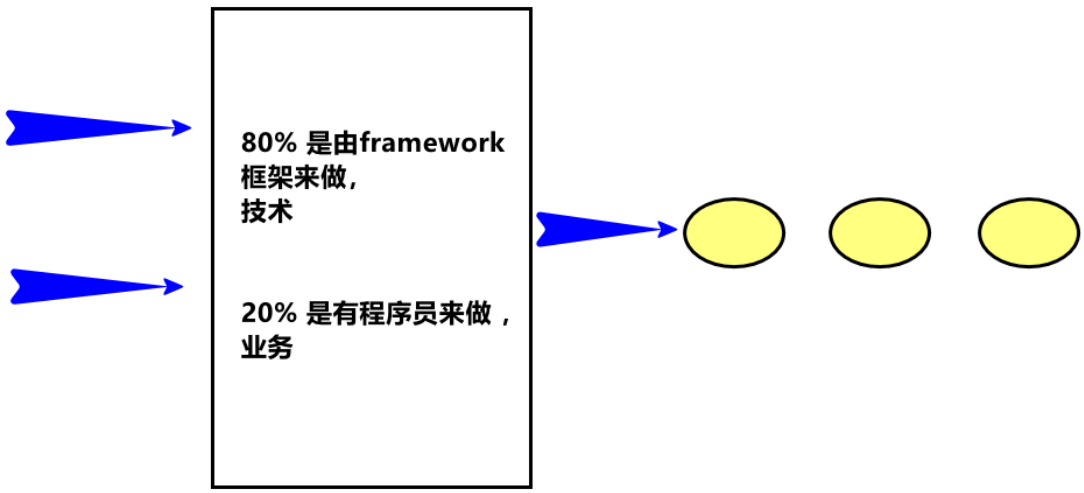
- Hadoop之父==Doug Cutting==卡大爷

Hadoop特性优点

- 分布式、扩容能力
  - 不再注重单机能力,注重集群的整体能力 ; 动态扩容、缩容
- 成本低
  - 在集群下,单机成本很低,可以是普通服务器组成集群,意味着大数据处理不一定需要超级计算机
- 高效率 并发能力
- 可靠性
- 通用性
  - hadoop区分技术和业务。

        - 业务,程序员（20%）
        - 封装了技术的细节（80%）
            - 如何存储,如何分,副本
            - 序列化和反序列化
            - 容错的机制
            - 切片的策略
            - 放置策略(机架感知策略)
            - 心跳机制
    - 用户实现相关的业务

Hadoop的特点： 通用性



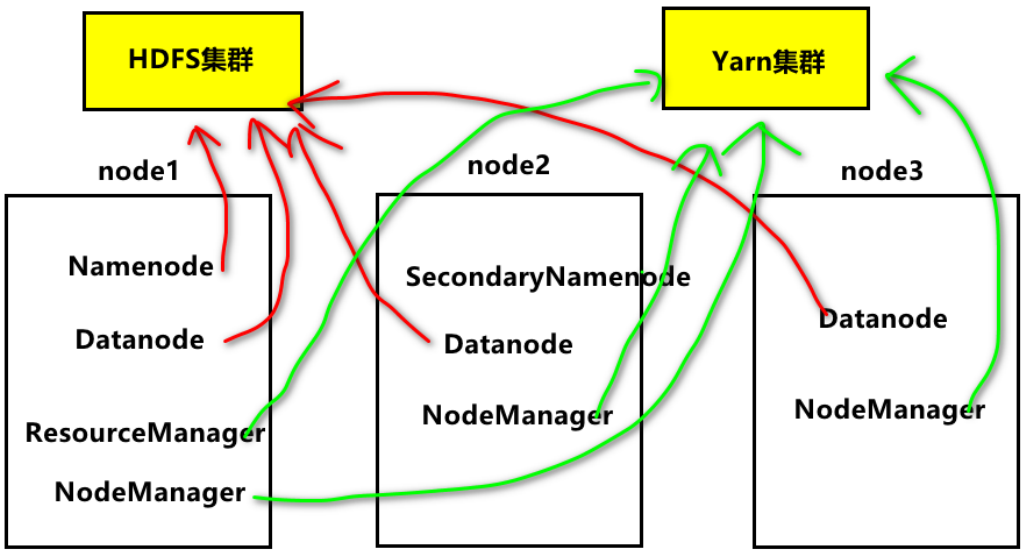
Apache Hadoop集群搭建

发行版本

- 官方版本 --版本新,兼容性欠缺
    - https://archive.apache.org/dist/hadoop/common/hadoop-3.3.0/
- 商业版本 -- 版本旧,稳定性好,收费,技术支持
    - 著名:Cloudera、hotonWorks、MapR、FusionInsight、星环科技大数据
    - cloudera | 90%商业 集群 组件/框架
- 版本变化
    - 1.x
        - 只有hdfs mapreduce . 架构过于垃圾,性能不高,当下企业中没人使用了
    - 2.x
        - HDFS:分布式文件系统
            - 解决海量数据存储问题
        - MapReduce: 分布式计算框架
            - 解决海量数据计算问题
        - yarn: 分布式集群资源管理系统
            - 分配硬件资源给程序,调度任务
    - 3.x

    - 架构和2一样 性能做了优化
- Hadoop集群
    - 通常是由hdfs集群 和 yarn集群 . 两个集群都是标准的 主从架构 集群
    - 两个集群逻辑上分分离,物理上在一起

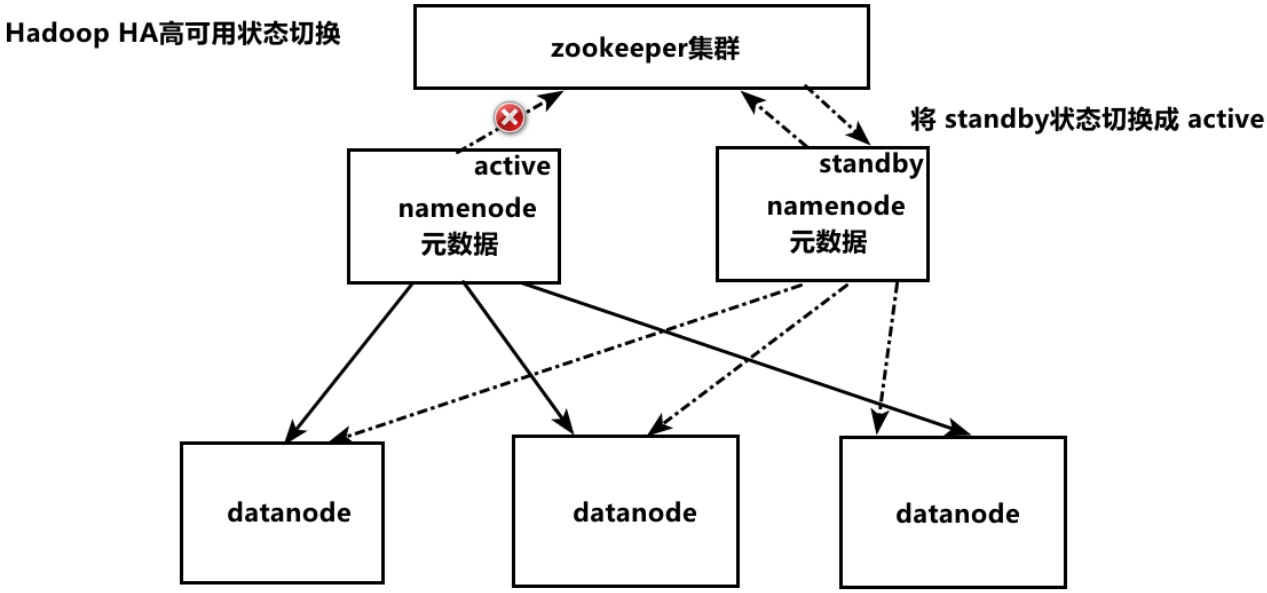
物理上在一起，逻辑上分离
Hadoop集群 分为两种集群： HDFS集群（存储）、Yarn集群（资源的管理 CPU、内存）
HDFS集群角色： Namenode（名字节点）、Datanode(数据节点)、SecondaryNamenode（秘书）
YARN集群角色：ResourceManager（资源管理老大）、NodeManager（节点小弟）



    - HDFS集群: 解决了海量数据存储,分布式存储系统
        - 主角色: namenode(NN)
        - 从角色: datanode(DN)
        - 秘书 : secondarynamemode(SNN)
    - YARN集群: 集群资源管理 任务调度
        - 主角色: resourcemanager(RM)
        - 从角色: nodemanager (NM)

Hadoop部署模式、集群规划

- 单机模式:
    - 一台机器,所有的角色在一个Java进程中运行. 适合体验
- 伪分布式:
    - 一台机器,每个角色单独的java进程.适合测试
- 分布式 cluster
    - 多台机器,每个角色运行在不同的机器上,生成测试都可以
- 高可用(持续可用) 集群 HA
    - 在分布式的模式下,给主角色设置备份角色,实现了容错的功能,解决了单点故障,保证集群持续可用性



- Hadoop集群的规划
  - 根据软件和硬件的特性 合理的安排,各个角色在不同的机器上
    - 有冲突的尽量不部署在一起
    - 有工作依赖尽量部署在一起
    - nodemanager 和datanode是基友

```
node1: namenode  datanode
    | resourcemanager  nodemanger
node2:                      datanode   secondarynamenode
    |                             nodemanger
node3:                      datanode
    |                         nodemanger
```

    - 如果后续需要扩容hadoop集群,应该增加哪些角色呢?

```
node4:  datanode  nodemanger
node5:  datanode  nodemanger
node6:  datanode  nodemanger
.....
```

Hadoop源码编译

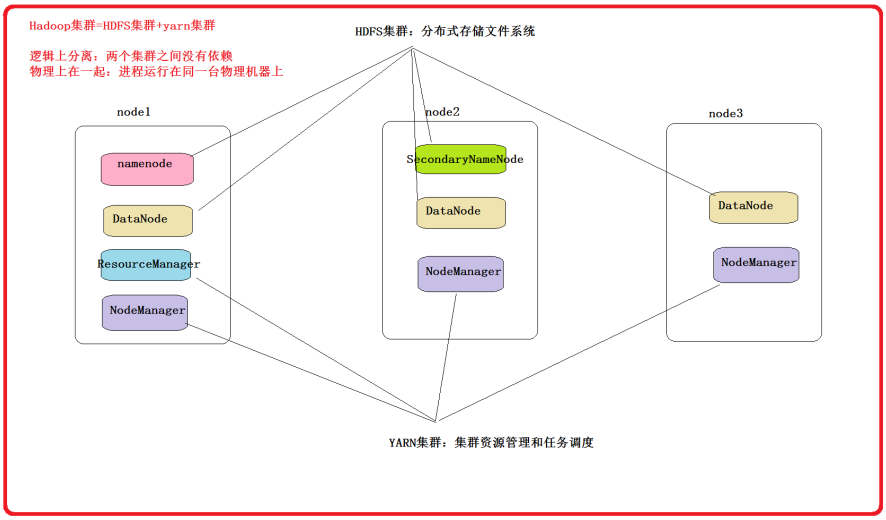
Index of /dist/hadoop/common/hadoop-3.3.0

Name	Last modified	Size	Description
Parent Directory		-	
CHANGELOG.md	2020-07-15 17:05	376K	
CHANGELOG.md.asc	2020-07-15 17:05	819	
CHANGELOG.md.sha512	2020-07-15 17:05	153	
RELEASENOTES.md	2020-07-15 17:05	26K	
RELEASENOTES.md.asc	2020-07-15 17:05	819	
RELEASENOTES.md.sha512	2020-07-15 17:05	156	
hadoop-3.3.0-aarch64.tar.gz	2020-07-15 17:19	478M	
hadoop-3.3.0-aarch64.tar.gz.asc	2020-07-15 17:19	819	
hadoop-3.3.0-aarch64.tar.gz.sha512	2020-07-15 17:19	168	
hadoop-3.3.0-rat.txt	2020-07-15 17:05	2.0M	
hadoop-3.3.0-rat.txt.asc	2020-07-15 17:05	819	
hadoop-3.3.0-rat.txt.sha512	2020-07-15 17:05	161	
hadoop-3.3.0-site.tar.gz	2020-07-15 17:33	40M	
hadoop-3.3.0-site.tar.gz.asc	2020-07-15 17:33	819	
hadoop-3.3.0-site.tar.gz.sha512	2020-07-15 17:33	165	
hadoop-3.3.0-src.tar.gz	2020-07-15 17:05	32M	
hadoop-3.3.0-src.tar.gz.asc	2020-07-15 17:05	819	
hadoop-3.3.0-src.tar.gz.sha512	2020-07-15 17:05	164	
hadoop-3.3.0.tar.gz	2020-07-15 17:30	478M	
hadoop-3.3.0.tar.gz.asc	2020-07-15 17:30	819	
hadoop-3.3.0.tar.gz.sha512	2020-07-15 17:30	160	

src的是源码包

官方编译好的安装包
其他软件带有bin关键字是安装包

Hadoop集群的概念



Q：为什么没有MapReduce集群？哪里去了？

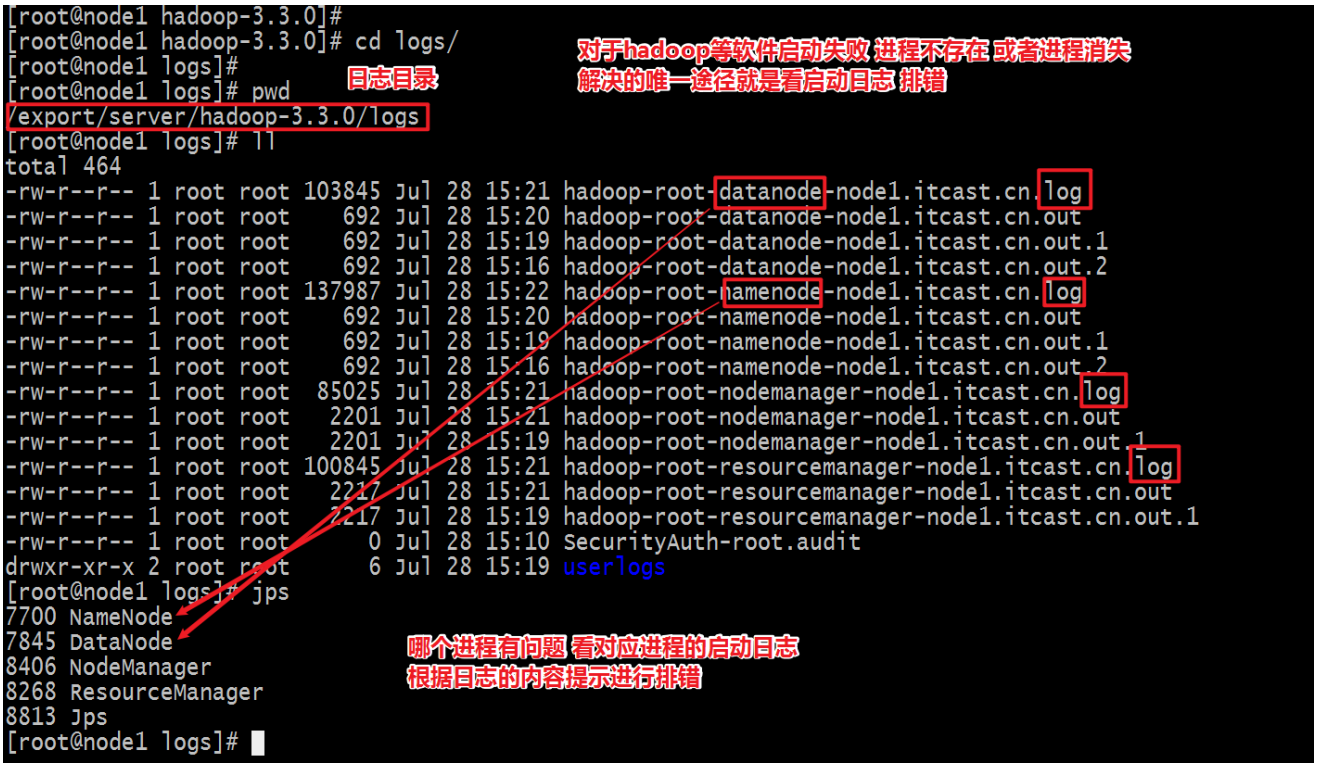
MapReduce是什么？

根本就没有MapReduce集群这一说法，MapReduce是计算程序，本质是代码（可以使用java C python）不同的语言来编写。

最终MapReduce程序是运行在yarn上面 处理HDFS数据的。

hdfs yarn是物理层面的组件
MapReduce是代码层面的组件

启动日志


[root@node1 hadoop-3.3.0]#
[root@node1 hadoop-3.3.0]# cd logs/
[root@node1 logs]#
[root@node1 logs]# pwd
/export/server/hadoop-3.3.0/logs
[root@node1 logs]# ll
total 464
-rw-r--r-- 1 root root 103845 Jul 28 15:21 hadoop-root-datanode-node1.itcast.cn.log
-rw-r--r-- 1 root root 692 Jul 28 15:20 hadoop-root-datanode-node1.itcast.cn.out
-rw-r--r-- 1 root root 692 Jul 28 15:19 hadoop-root-datanode-node1.itcast.cn.out.1
-rw-r--r-- 1 root root 692 Jul 28 15:16 hadoop-root-datanode-node1.itcast.cn.out.2
-rw-r--r-- 1 root root 137987 Jul 28 15:22 hadoop-root-namenode-node1.itcast.cn.log
-rw-r--r-- 1 root root 692 Jul 28 15:20 hadoop-root-namenode-node1.itcast.cn.out
-rw-r--r-- 1 root root 692 Jul 28 15:19 hadoop-root-namenode-node1.itcast.cn.out.1
-rw-r--r-- 1 root root 692 Jul 28 15:16 hadoop-root-namenode-node1.itcast.cn.out.2
-rw-r--r-- 1 root root 85025 Jul 28 15:21 hadoop-root-nodemanager-node1.itcast.cn.log
-rw-r--r-- 1 root root 2201 Jul 28 15:21 hadoop-root-nodemanager-node1.itcast.cn.out
-rw-r--r-- 1 root root 2201 Jul 28 15:19 hadoop-root-nodemanager-node1.itcast.cn.out.1
-rw-r--r-- 1 root root 100845 Jul 28 15:21 hadoop-root-resourcemanager-node1.itcast.cn.log
-rw-r--r-- 1 root root 2217 Jul 28 15:21 hadoop-root-resourcemanager-node1.itcast.cn.out
-rw-r--r-- 1 root root 2217 Jul 28 15:19 hadoop-root-resourcemanager-node1.itcast.cn.out.1
-rw-r--r-- 1 root root 0 Jul 28 15:10 SecurityAuth-root.audit
drwxr-xr-x 2 root root 6 Jul 28 15:19 userlogs
[root@node1 logs]# jps
7700 NameNode
7845 DataNode
8406 NodeManager
8268 ResourceManager
8813 Jps
[root@node1 logs]#
日志目录
对于hadoop等软件启动失败 进程不存在 或者进程消失
解决的唯一途径就是看启动日志 排错
哪个进程有问题 看对应进程的启动日志
根据日志的内容提示进行排错