大讲台
dajiangtai.com

大数据集群规划及部署

— 集群规划

主讲人：小马哥

目录

大数据集群规模规划

大数据集群规模规划

- 规模规划的目的：规划集群的存储容量及集群（服务器资源）规模

- 集群资源有以下种类：
  - 磁盘存储容量
  - CPU总核（线程）数
  - 物理内存总量
  - 单机网络带宽、汇聚层总网络带宽
  - 其他选配资源，如GPU等

- 规模规划的原则：根据计算需求，为集群配置合理的硬件规模，不多配，不少配

大数据集群规模规划

集群规模取决于数据量及计算复杂度两个因素，最终规划值为以下每种估算方式得到的最小集群规模的最大值。

•容量需求

估算相对容易且准确。大多数案例可以通过容量需求来决定集群规模。

•计算需求

估算相对困难，没有统一的模型，变量较多。要准确地估算计算资源，只能通过小规模线上任务测试，并结合经验进行合理估算。

•其他资源限制

例如需要执行机器学习应用，则可能对CPU/GPU、磁盘IO等资源有特殊要求，且可能产生单节点资源需求有下限、机器配置异构等情况，在估算集群规模时需满足此类需求

大数据集群规模规划

- 规模规划的一些假设：
  - 假设集群的计算节点都使用相同的CPU、内存、磁盘，即集群所有计算节点是同构的。异构集群在管理上会有一定难度，因此我们建议一次采购需要配同一厂家的同型号服务器，多次采购尽量使用同型号服务器
  - 假设当前最经济的配置为单节点CPU56-64核（2路28/32核，并且忽略不同主频、架构的CPU在单核性能上的差距）、硬盘为48T（4T*12）。随着硬件的发展，在某一特定时间，最有性价比的CPU、磁盘一般就几个型号，请事先咨询供应商

大数据集群规模规划

1 容量需求估算

•数据副本数

HDFS默认使用3副本。如使用Hadoop3.0后的核心版本（如CDH6.x的发型版），则可以利用纠删码特性，使副本数最低降低到1.4左右。我们暂不考虑纠删码特性。

•压缩算法

Snappy、gzip…

•数据膨胀率

若存储于HBase，则和column数目以及rowkey长度等因素都有直接关系；单column存储通常膨胀率在15%以内

大数据集群规模规划

- 附加数据存储空间

如元数据、缓存数据等，通常在20%以内

- 预留临时存储空间

通常预留20%至30%的临时空间供MapReduce、Spark等组件使用

- 考虑数据增长率

一般根据业务规模增长确定，如年增长率为20%

大数据集群规模规划

案例：某大型企业日志分析平台容量估算
该Hadoop集群需要保留3个月的原始日志记录，原始日志采用snappy压缩。汇聚层（如DW、RPT等）的数据量为原始数据量的1/4，永久保留，不压缩。

数据类型	日数据量	保留策略	预计需存储数据量	HDFS物理存储容量
原始日志	4T	3个月	360T	360T*3(副本数)*0.3(压缩比)/70%(非临时空间比例)/50%(磁盘利用率)=926T
汇聚层数据	1T	全量（先存1年）	360T	360T*3(副本数)/70%(非临时空间比例)/50%(磁盘利用率)=3086T

大数据集群规模规划

假设该企业决定采购单台12块数据盘，每块4TB，则需要服务器台数=（926+3086）/48=84台。这84台全部为存储/计算节点，再配上适量的管理节点、边缘节点等，则总服务器需求在90-95台左右。

2 计算需求估算

•Map任务通常可线性扩展

将单个Map任务使用的CPU、内存以及IO资源成比例增加，一般可以获得相同倍数的性能提升

•Reduce任务通常不可线性扩展

Reduce任务数目（如count distinct）以及数据偏斜可能造成系统瓶颈，需要通过对真实负载进行测试来发现

大数据集群规模规划

- 宽依赖join线性扩展受网络限制

Spark中如果两个RDD分区数和分区方式一致，则连接时为窄依赖，否则为宽依赖。宽依赖一定会造成shuffle，如果以单纯增加计算节点/分区数的方式试图加速宽依赖join，则瓶颈可能出现在网络IO上。

- 机器学习任务的计算资源更为复杂

CPU时钟速度、缓存、内存及IO等只是计算能力中的一个维度。采用CPU+GPU集群工作模式，每个节点内采用CPU+GPU异构模式，在小规模集群下测试benchmark，以此推算计算能力

大数据集群规模规划

案例：某大型企业日志分析平台容量估算
如果该平台将要上线的所有Hive/Spark等任务已经开发好，则可以使用1/10数据量对所有任务进行一次试跑。假设在测试过程中不可并行部分、数据偏斜的影响不显著，网络IO离负载上限很远，且最少使用总计400核的计算节点可以在预期时间内完成计算任务，则如果决定采购单台2路32核的服务器（共64核），对计算节点的真实需求可估算为400*10/64=63台

大数据集群规模规划

结论：结合存储需求估算和计算需求估算，我们可以看到这个案例是一个存储密集型的场景，取两者最大值84台作为存储/计算节点需求即可。

思考：上述例子单机使用4T*12，如果使用8T*12，则按存储估只需要42台，与计算一起两者取小只需要63台，一下可以节约1/4节点数，如果你的领导问你这个问题，该如何回答？

生产环境部署规划

生产环境部署规划

- 硬件部分：
  - Hadoop集群根据不同的计算需求，通常可分为IO密集型和CPU密集型两类。IO密集型的计算任务有数据导入导出、ETL、索引、分组等。CPU密集型的计算任务有数据挖掘、机器学习等。不同的计算需求适合于配置不同的硬件，每个企业的预算、集群规模、现有硬件（如果搭建Hadoop需要利用现有硬件）也不尽相同。
  - 以目前的生产环境为例，如果使用自建机房搭建集群，一般会采购PC服务器作为集群节点（通常大小为2U），安装在机架上（标准机架为42U，一般不会安装超过14台服务器），机架内部（接入层）至少要保证千兆以太网连接（推荐万兆），机架与机架之间（汇聚层）至少要保证万兆以太网连接。

生产环境部署规划

- Hadoop集群也可安装在虚拟机或公有云上，CDH对此有良好的支持，选择硬件时，可参照物理机搭建集群的配置，并适当地考虑数据交换成本等额外因素。（CDP即CDH7开始支持混合云部署方式，但由于CDP只有付费license，我们不做推荐）

生产环境部署规划

- 按照节点在集群中角色的不同，我们一般会分为四类节点：
  - 管理节点：主要用于运行重要的管理进程，如NameNode，ResourceManager等。
  - 工具节点：主要用于非Hadoop管理进程的其他进程，如Cloudera Manager，Hue等。
  - 边缘节点：用于运行集群的客户端、Flume等数据采集进程、FTP服务等。
  - 工作节点：主要用于运行各种分布式计算进程，如nodemanager，impalad等。

生产环境部署规划

- 对于前三类节点，推荐配置：
  - 2路6核以上的CPU，主频至少2GHz；
  - 64-512GB内存，具体取决于负载多重，如NamaNode可以多配一些；
  - 4-8个1TB以上的SAS或SATA硬盘，一般OS、ZooKeeper存储目录等可以用裸盘，NameNode的fsimage、数据库数据文件等盘建议用RAID 1或RAID10。

生产环境部署规划

- 对于工作节点，推荐配置：
  - 2路6核以上的CPU，主频至少2GHz，如果为CPU密集型集群，可选择2路12核及以上CPU；
  - 64-512GB内存，具体取决于集群部署的角色，如果只运行Hadoop核心组件，则64或128GB一般够用，如果混合部署Impala、Spark等内存计算组件，则至少配置256或512GB（也可如下估算，CPU密集型——CPU:内存为1:4，IO密集型或内存计算——CPU:内存为1:8或1:16）；
  - 4-24个2TB以上的SAS或SATA硬盘，一般2U服务器内插硬盘个数不超过8个，可以通过背板扩展卡扩展到16甚至24个。虽然Hadoop也支持异构存储，但一般不需要使用SSD硬盘，除非对IO有特别高的需求；
  - 柜顶交换机（接入层）使用千兆或万兆的，机架之间的核心交换机（汇聚层）至少也要是万兆的，保证异机架节点之间的带宽至少为千兆。如果预算充裕，可以进一步考虑网卡bond、交换机堆叠等部署策略，进一步提升带宽。

生产环境部署规划

- 具体品牌/型号的选择，以某电商网站查到的某品牌服务器为例，这个配置（1颗金牌5218，128G，8*12T）的报价大约是5w块钱。如果作为计算节点使用，我们最好再加1颗5218，并将内存扩到256G，这样一台大概是6w块钱。管理节点/工具节点则不需要这么多的核、内存和磁盘，可以灵活选配其他型号。

生产环境部署规划

- 接入层/汇聚层交换机

¥9000.00

新华三（H3C）三层网管多速率企业级核心交换机 S6520-24S-SI 24口万兆 #H3C

9条评价

¥26100.00

华三（H3C) S6520X-30QC-EI 24万兆SFP+光口三层核心交换机 【企业放心购

0条评价

生产环境部署规划

- 软件部分：
  - 集群软件，目前CDH和CM均有两个大版本，5.x和6.x。5.x最新版本为5.16.2（均有免费license），6.x最新版本为6.3.x（但是license情况比较复杂，6.3.3后只提供收费版）。我们课程使用6.2.x版本进行演示，但实际生产中小马哥更推荐使用5.16.2

生产环境部署规划

- 6.2.x的CM/CDH支持的操作系统如下表，但如果需要安装CDSW（Cloudera Data Science Workbench），则需要RHEL/CentOS7以上系统。

Operating System	Version (bold=new)
RHEL-compatible	
RHEL/CentOS/OL with RHCK kernel	**7.7**, 7.6, 7.5, 7.4, 7.3, 7.2 6.10, 6.9, 6.8
Oracle Linux (OL)	**7.6**, 7.4, 7.3, 7.2 (UEK default) 6.10 (UEK default)
SUSE Linux Enterprise Server	
SLES	**12 SP4***, 12 SP3, 12 SP2
Ubuntu	
Ubuntu	**18.04** LTS (Bionic) 16.04 LTS (Xenial)

生产环境部署规划

- 6.2.x的CM/CDH支持数据库有MySQL、MariaDB、PostgreSQL、Oracle等，具体版本清单见如下页：

https://docs.cloudera.com/documentation/enterprise/6/release-notes/topics/rg_database_requirements.html

其中MySQL支持的版本见下表，注意最新的8.0版本没有说官方支持，不要用。

MySQL Support across Cloudera Enterprise 6 Releases

MySQL Version	Cloudera Enterprise 6.x
5.1 (default for RHEL/CentOS/OEL 6)	✔
5.5 (default for Debian 8.9)	✔
5.6	✔
5.7 (default for Ubuntu 16.04, 18.04 LTS)	✔

生产环境部署规划

- 6.2.x的CM/CDH不论使用OracleJDK还是OpenJDK，均只能使用1.8版本。至于小版本号的支持可参见以下文档，建议使用OracleJDK 8u181

https://docs.cloudera.com/documentation/enterprise/6/release-notes/topics/rg_java_requirements.html

Cloudera Enterprise Version	Supported Oracle JDK	Supported OpenJDK
5.3 -5.15	1.7, 1.8	none
5.16 and higher 5.x releases	1.7, 1.8	1.8
6.0	1.8	none
6.1	1.8	1.8
6.2	1.8	1.8
6.3	1.8	1.8, 11.0.3 or higher

生产环境部署规划

- 角色划分：
- 对于生产集群，还有一个重要的工作是角色划分，即为每个节点设置运行的进程。因为只有工作节点才真正承担分布式计算任务，管理节点、工具节点、边缘节点完全不承担计算任务或只承担非分布式的任务，因此在100个节点以上的中大规模集群中，我们希望计算节点的占比尽可能高。
- 但是三类非计算节点的个数也不是越少越好，尤其是管理节点上的进程都非常重要，通常会将其分散到多个节点上，以防止节点失效产生严重影响。比如，如果一个节点上既有HDFS的NameNode又有HBase的HMaster，该节点故障的话，即使两者都配置了高可用，也会造成一段时间内两个角色的元数据服务都不可用，影响比较大，因此像此类重要进程尽量单独设置节点，或和ZooKeeper这样稍次要的角色合设。
- 根据经验，中大型集群一般使用5%-10%的节点作为非工作节点，并依据这些节点上运行进程的CPU、内存、IO使用特性和HA要求，来合理地进行划分。

实验环境部署规划

实验环境部署规划

- 硬件部分，小马哥演示时会使用阿里云来进行部署，ECS配置：
  最低2C8G*1+2C4G*4，推荐2C8G*5
- 同时我也会提供使用本机VMware配置虚拟机的方式来配置主机的方式，这种方式需要你的电脑至少有16G内存，否则是虚不出来的。虚拟机配置：
  最低2C6G*1+1C1.5G*4
- 使用虚拟机或云主机后，网络带宽问题就不需要特别关注了。为了实验的方便，小马哥还会让第1个节点可以访问公网。各位学员如果在生产环境部署集群，又不允许访问公网的话，后面会给大家提供解决方案（事实上，我们的安装方案也不依赖公网下载，只是方便大家SSH和访问管理页面而已）。

实验环境部署规划

- 本讲首先进行主机环境的准备。操作系统选用CentOS7.7，JDK选择Oracle JDK8，元数据库选择Mysql5.7。具体小版本选择，参见后续演示。
- 下一讲开始集群安装。小马哥演示会安装6.2.1版本的CM，接着安装6.2.1版本的CDH，使得如果学员有想法的话，可以尝试将CM和CDH升级到6.3.x。安装的角色只选择Hadoop core、Hive、Hue、ZooKeeper、HBase，其余组件留作后续演示增加角色。

实验环境部署规划

- 角色划分方面，由于演示集群的总节点数很少，不可避免有大量角色合设。最终分配方案如下（CM：Cloudera Manager；NN：NameNode；RM：ResourceManager；ZK：ZooKeeper；SNN：SecondaryNameNode；HS2：HiveServer2；DN：DataNode；NM：NodeManager；M：HBase Master；RS：RegionServer）：
  - hadoop1（2核8G）：CM、NN、RM、Hue
  - hadoop2：SNN、HS2、M
  - hadoop3：DN、NM、ZK、RS
  - hadoop4：DN、NM、ZK、RS
  - hadoop5：DN、NM、ZK、RS

THANKS