doi:10. 16055/j. issn. 1672-058X. 2020. 0001. 007

基于 Ambari 的 Hadoop 集群快速部署研究

吴丽杰, 张璐璐, 张 婷
(安徽粮食工程职业学院,合肥 230011)

摘 要:Hadoop 作为处理大数据的一个优秀分布式计算框架,在企业应用非常普通。然而 Hadoop 集群部署需要考虑各组件的兼容性、编译问题及繁琐的组件参数配置,初学者往往耗时几天也不能部署成功。Ambari 是一种支持 Hadoop 集群部署、监控和管理的开源工具。针对 Hadoop 集群部署的复杂性,提出基于 Ambari 工具部署 Hadoop 集群各组件的实践方法并讨论了快速部署的若干要点及重要步骤;通过 Ambari 工具,完成了 Hadoop 生态圈最小化集群大部分常用组件的快速部署,如 HDFS、HBase、Hive、Pig、Oozie、Zookeeper、Sqoop、Spark、Storm、Kafka、Flume 等;项目实践表明:利用 Ambari 工具能够在 8 h 内部署完毕 Hadoop 集群,相比较传统手工部署方式,Ambari 工具极大提高了 Hadoop 集群部署的效率及成功率。

关键词:Hadoop;Ambari;HDP;大数据;快速部署

中图分类号:TP311 **文献标志码**:A **文章编号**:1672-058X(2020)01-0042-07

0 引 言

在我国近期的政府机构改革中,有很多省市建立了大数据管理局。大数据管理局的成立表明我国政府已充分认识到大数据的重要性,专职机构的设立也能真正敦促各级各部门更重视大数据建设。随着大数据时代的到来,大数据应用也日趋成熟,如大数据智慧交通、大数据智慧教育、大数据精准营销[1-2]等应用,国内很多高校新设立了大数据专业,来适应社会发展的需要。大数据营销在国内的电商平台应用尤为普遍。在近期国内高校组织的大数据与人工智能应用类竞赛中,如 2018 年全国大学生信息安全竞赛安徽省赛,将大数据平台部署、大数据预处理、大数据分析、大数据可视化以及综合应用作为竞赛内容。

Hadoop 是一种基于 MapReduce 数据并行处理及 HDFS 分布式数据存储的分布式计算框架[3-4],在企业应用非常普遍。Hadoop 经过多年发展,已经成为大数据平台的一个庞大的生态圈,包含了数据存储、数据分析、图计算模型、流计算模型、访问接口、消息队列等众多组件。基于 Hadoop 生态圈,可以构建海量存储、高性能、安全可靠、易扩展的云存储等系统[5]。由于 Hadoop 的组件及版本很多且环境构建复杂,初学者往往要花费好几天时间也不能部署成功,造成刚接触这门学科的学生容易对平台部署产生畏惧心理。然而学习大数据这类实践性很强的学科,部署平台是学习的基本要求。纸上谈兵终觉浅,只有部署好平台才能已最快的方式熟悉大数据各组件,如 HDFS、HBase、Hive、Zookeeper、Sqoop、Spark 等。

利用 Ambari 则可以快速部署 Hadoop 集群。讨

收稿日期:2019-05-31;**修回日期**:2019-07-05.

作者简介:吴丽杰(1983—),女,山东聊城人,讲师,硕士,从事计算机应用研究.

论了基于 Ambari 的 Hadoop 集群部署在教学及项目实践过程中的重要步骤及快速部署要点。

1　Hadoop 版本及集群管理方式

Hadoop 发展了多年,也涌现了众多版本,但对于初学者而言,那些需要进行复杂环境部署的版本并不合适,特别是大数据开发专业,应该将学习的重点放在 Hadoop 等应用开发,而不是把研究的重点放在环境部署上面。

1.1　主流 Hadoop 版本

目前 Hadoop 的主流版本除了 Apache 的社区版 Hadoop 之外,还有遵从 Apache 开源协议的第三方发行版本。由于 Apache Hadoop 版本管理、集群的部署及升级、添加组件繁琐,第三方发行版本,如市场占有率较高的 Cloudera 的 CDH, Hortonworks 的 HDP, MapR 的 MapR 产品等,则解决了 Apache Hadoop 的版本管理、兼容性等问题并在大量的生产环境进行了检验,比 Apache Hadoop 的安全性、稳定性均有提高。

Hortonworks 的 HDP 作为市场占有率前几名的 Hadoop 大数据平台,提供了 Hadoop 生态圈的所有关键组件,如图 1 所示[6],非常适合大数据专业学生学习使用。

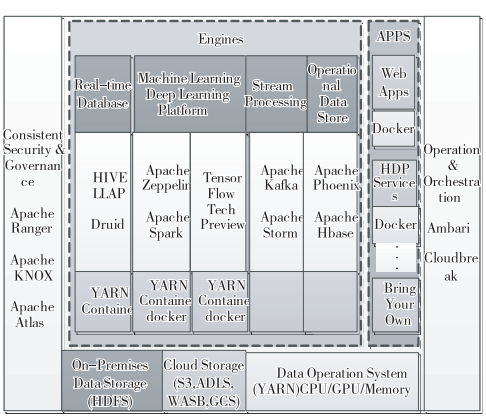

图 1　HDP 架构图

Fig. 1　Architecture diagram of HDP

1.2　主流 Hadoop 集群管理方式

Apache Hadoop 采用手工方式管理集群。方式虽然方便初学者对组件和集群管理的理解,但是需要初学者考虑很多的细节,需自行解决组件之间的兼容性问题。采用手工方式在部署环节就可能浪费几天时间去解决 Storm、Spark、Flume 等各组件的兼容性及编译问题。

第三方 Hadoop 发行版本的 CDH 和 HDP 采用工具方式管理集群。CDH 采用 Cloudera Manager 工具,HDP 采用 Ambari 工具。这类工具考虑了部署过程中所需的全部细节,大大提高了 Hadoop 集群部署的效率。通常初学者也可在 1 d 内部署好 Hadoop 集群。集群工具还提供了管理、监控、诊断、配置修改的工具,管理员可以快速定位问题,使 Hadoop 集群的运维工作更加简单、快捷。

2　Ambari 概述

Ambari 是 Apache Software Foundation 中的一个顶级项目。Ambari 是部署及监控 Hadoop 生态圈各组件的集群管理工具,支持的大数据平台组件包括 Hbase、Hive、Zookeeper、Sqoop、Atlas、Storm、Spark、YARN 等。用户通过 Ambari WEB 界面就可以判断出各服务器或服务中哪个组件发生异常,大大节省了管理人员的时间及人力成本。Ambari 支持用户根据需求选择所需要的服务并且可以选择每个服务安装在哪个节点上,从而达到更好的资源分配和性能要求[7-8]。

2.1　Ambari 架构

Ambari 主要由 Ambari Server 和 Ambari Agent 组成,如图 2 所示。Ambari Server 通知 Ambari Agent 安装对应的软件;通过 Ganglia 收集指标和 Nagios 系统预警来进行 Hadoop 集群的监控。Ambari 提供 RESTful API 接口,使集成商拥有工具提供的管理和监控 Hadoop 组件的能力。另外工具提供 WEB 界面,来管理各组件的配置、部署及监控[9]。

2.2　Ambari 优势

第三方 Hadoop 发行版本 CDH 和 HDP 采用的集群管理工具分别为 Cloudera Manager 和 Ambari。

二者均对 Hadoop 的部署及监控提供众多功能,而 Ambari 的优势主要如下:Ambari 为开源项目,支持二次开发;Ambari 支持 Redis、Apache Kylin、ElasticSearch 等服务集成;Ambari 支持创建自己的视图,添加自定义服务。

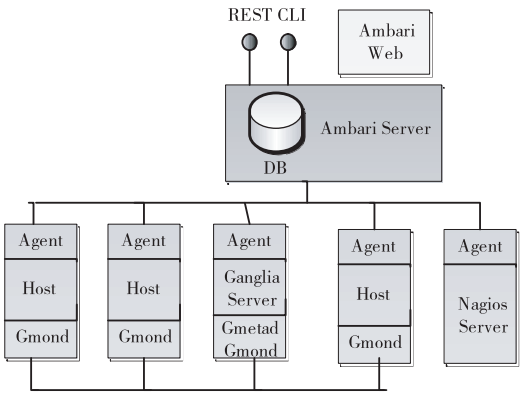

图 2 Ambari 架构图

Fig. 2 Architecture diagram of Ambari

2.3 Ambari 劣势

Ambari 只能管理 Hortonworks 的 HDP,不支持 Apach Hadoop 及其他第三方 Hadoop 发行版本。

考虑教学中各类大数据组件学习的需要及快速部署需求,采用 Ambari 来进行 Hadoop 集群的部署。

3 Hadoop 集群部署

利用 Ambari 工具快速部署 1 个最小化 Hadoop 集群环境,部署阶段主要包括几个要点[10]:

3.1 部署节点规划

通过 VMware 虚拟机软件部署 3 个虚拟机,分别为 Ambari Server 节点、Hadoop 节点 1 和 Hadoop 节点 2。宿主机基本配置如下:64 位 Windows Server 2008 R2 Enterprise+ Intel E5620 2.4 Ghz(16 核) + 16 GB 内存+600GB SCSI 硬盘。虚拟机上分别部署 Ambari Server 服务器(OS:Centos6.7 X86-64 位,IP:192.168.40.126,主机名:hadoop11.chase.com)、Hadoop 节点 1 服务器(OS:Centos6.7 X86-64 位,IP:192.168.40.127,主机名:hadoop12.chase.com)和 Hadoop 节点 2 服务器(OS:Centos6.7 X86-64 位,IP:192.168.40.128,主机名:hadoop13.chase.com),具体部署如图 3 所示。

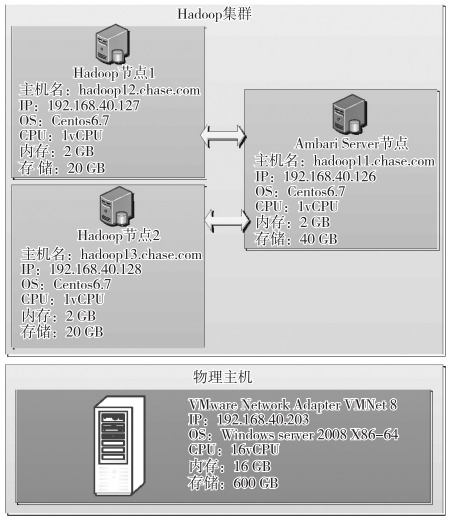

图 3 Hadoop 集群部署图

Fig. 3 The deployment diagram of Hadoop cluster

3.2 环境准备

(1) 组件兼容性版本选择及下载。根据 Hortonworks 官网提供的信息选择部署节点操作系统对应的 Ambari、HDP、JDK、DB 等组件的兼容版本。项目使用 Ambari2.6.2 + HDP2.6.5 + JDK8 + Mysql5.7 及 mysql-connector-java-5.1.40.jar,考虑到 HDP 等软件包过大,通过 Yum 等在线安装方式经常超时,采用从官网下载软件包。

(2) 配置 SSH 无密码登录。由于部署过程中,各节点需要互相访问,通过复制各节点.ssh/id_rsa.pub 至其他节点的.ssh/authorized_keys 文件中,实现任意两节点的无密码登录。

(3) 配置 NTP、iptables、selinux 服务。集群中的各节点需要安装并启动 ntp 服务以保证集群时间的一致。为了方便集群节点的互相通信,关闭各节点的 iptables、selinux 服务。

(4) 制作本地源。在 Ambari Server 节点上执行#yum install httpd - y #service httpd start 命令安装 Apache HTTP 并启动服务,修改默认目录为/bigdata,并将下载的 Ambari、HDP、JDK、MySQL 安装包上传到/bigdata/ambari 目录。访问 http://192.168.40.216/ambari,查看 HTTPD 服务是否可用,如图 4 所示。

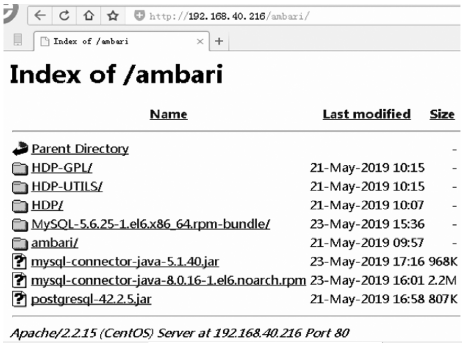

图 4　查看 HTTPD 服务

Fig. 4　Check the HTTPD service

使用#yum install　#yum-utils createrepo 命令安装 yum 源工具并配置 Ambari、HDP、HDP-UTILS 的本地源文件中的 baseurl 和 gpgkey 为本地路径。如 baseurl = http://192. 168. 40. 216/ambari/ambari/centos6/2. 6. 2. 0 - 155, gpgkey = http://192. 168. 40. 216/ambari/ambari/centos6/2. 6. 2. 0 - 155/RPM-GPG-KEY/RPM-GPG-KEY-Jenkins。

将 Ambari server 节点上配置好的 ambari. repo、hdp. repo、hdp. gpl. repo 分发到 Hadoop 节点 1 和 Hadoop 节点 2,并执行如下命令#yum clean all #yum makecache #yum list 完成本地源配置。

(5) 安装 Ambari 服务。在 Ambari Server 节点上执行#yum install ambari-server 命令,安装 Ambari sever 服务。配置节点的 Java 环境并远程分发到其他节点。安装 MySQL 数据库并启动服务。拷贝 mysql-connector-java-5. 1. 40. jar 至/usr/share/java 目录并添加 server. jdbc. driver. path =/usr/share/java/mysql-connector-java. jar 至/etc/ambari-server/conf/ambari. properties 文件。执行如下命令完成 Ambari 安装初始化:

ambari-server setup --jdbc-db = mysql --jdbc-driver=/usr/share/java/mysql-connector-java. jar

执行如下命令创建 MySQL ambari 数据库、用户并设置权限,如需安装 Hive、oozie 组件,同样需要创建对应的数据库及用户。

执行#ambari-server setup 进行 Ambari Server 的安装,如图 5 所示。Ambari 默认使用 PostgreSQL 作为内置数据库,鉴于项目学习的需要,选择使用率最广的开源数据库 MySQL,根据安装提示即可完成 Ambari 服务的安装。

```
Checking JDK...
Do you want to change Oracle JDK [y/n] (n)? y
[1] Oracle JDK 1.8 + Java Cryptography Extension (JCE) Policy Files 8
[2] Oracle JDK 1.7 + Java Cryptography Extension (JCE) Policy Files 7
[3] Custom JDK
==============================================================================
Enter choice (1): 3
WARNING: JDK must be installed on all hosts and JAVA_HOME must be valid on all
WARNING: JCE Policy files are required for configuring Kerberos security. If yo
Path to JAVA_HOME: /opt/java/jdk1.8.0_112
Validating JDK on Ambari Server...done.
Checking GPL software agreement...
Completing setup...
Configuring database...
Enter advanced database configuration [y/n] (n)? y
Configuring database...
==============================================================================
Choose one of the following options:
[1] - PostgreSQL (Embedded)
[2] - Oracle
[3] - MySQL / MariaDB
[4] - PostgreSQL
[5] - Microsoft SQL Server (Tech Preview)
[6] - SQL Anywhere
[7] - BDB
==============================================================================
Enter choice (3): 3
Hostname (ambari): hadoop11.chase.com
Port (5432): 3306
Database name (ambari):
Username (ambari):
Enter Database Password (bigdata):
Re-enter password:
Configuring ambari database...
Configuring remote database connection properties...
WARNING: Before starting Ambari Server, you must run the following DDL against
Proceed with configuring remote database connection properties [y/n] (y)? y
Extracting system views...
.....
Adjusting ambari-server permissions and ownership...
Ambari Server 'setup' completed successfully.
```

图 5　Ambari Server 安装

Fig. 5　The installation of Ambari Server

3.3 集群的安装、配置和部署

3.3.1 启动 Ambari Server

执行#ambari-server start,登录 http://192.168.40.216:8080/,进入 Ambari 管理界面,如图6所示。

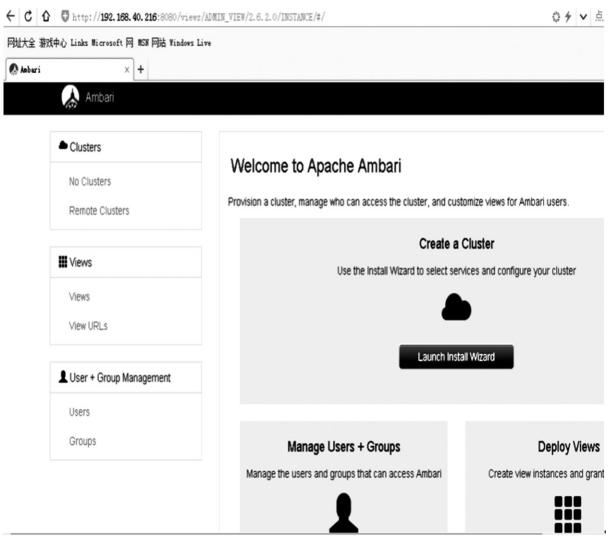

图6 Ambari 管理界面

Fig. 6 The management interface of Ambari

3.3.2 配置本地源及其他安装选项

根据向导创建集群名称,配置 HDP 2.6.5 的本地源路径并完成部署节点的注册,如图7所示。

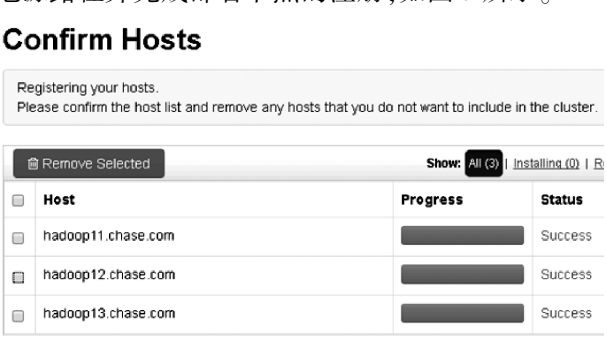

图7 节点注册

Fig. 7 Registering nodes

3.3.3 选择组件服务并指定主节点

鉴于初学者学习需要,选择 HDP 2.6.5 的所有组件,包含 Hive、HBase、Pig、Flume、Spark、Mahout 等。将 Kafka、Accumulo、Falcon、Oozie 等服务指定到待部署的节点上。考虑到节点的性能,组件的选择尽量均匀,防止单个节点负载过高造成服务无法启动。

3.3.4 指定从节点和客户端并修改配置信息

根据每个节点的实际情况,选择哪些客户端运行在从节点上。客户端包括 HDFS、YARN、MapReduce、Sqoop、Slider 等。

根据系统的提示信息进行组件配置信息的修改,如图8所示,需要修改 Hive 的账号密码并测试连接。

3.3.5 组件安装和测试

确认组件的部署信息后,即可进行已选择组件的安装和测试,如图9所示。安装过程中,如有问题,系统会针对问题给出提示信息,解决问题后可进行重新安装。

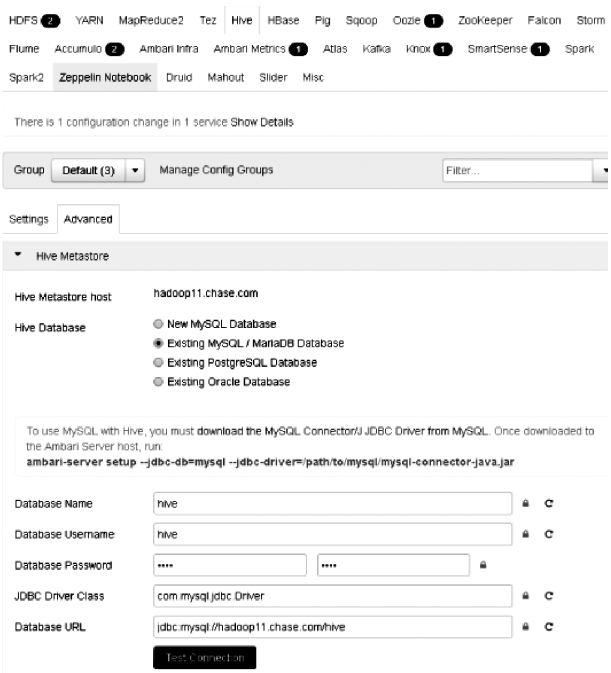

图8 组件配置信息修改

Fig. 8 The modification of component configuration information

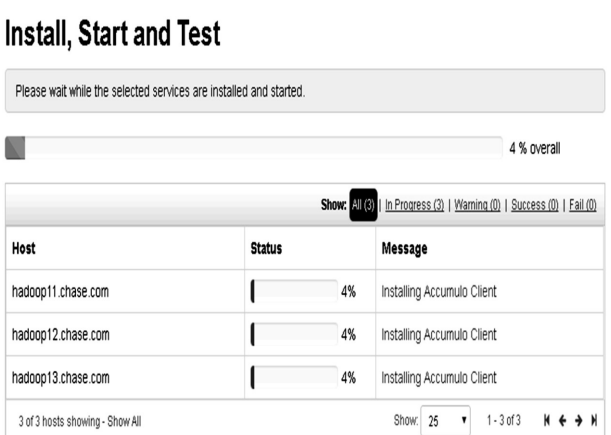

图9 组件安装过程

Fig. 9 The process of component installation

3.4　部署确认

成功部署 Hadoop 集群后,可通过 Ambari 管理界面查看组件的运行情况,如图 10 所示。如某些组件服务没有启动,也可以通过界面查看对应组件服务启动失败原因,解决后单独启动。

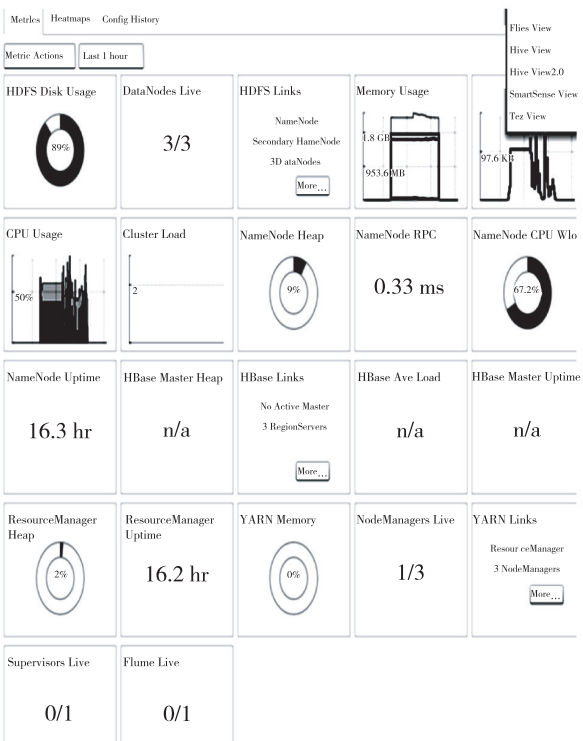

图 10　Ambari 监控界面

Fig. 10　The monitoing interface of Ambari

4　结　语

讨论了基于 Ambari 的 Hadoop 集群快速部署的若干要点及重要步骤。部署过程中,利用 Ambari 部署工具仅仅 3 h 左右完成了 2 节点 Hadoop 生态圈大部分常用组件的部署。SSH 无密码登录、本地源制作、安装 Ambari 服务、JDK 及 MySQL 等环境配置也仅耗时 4 h 左右,可见利用 Ambari 工具能极大提高部署 Hadoop 集群的效率及成功率。

在实际生产环境中,也可以通过 Ambari Blueprints 的声明定义一个集群,勿需通过 Ambari 的安装向导,实现自动化部署 Hadoop 集群[9]。项目实践表明使用 Ambari 部署 Hadoop 有如下缺点:Ambari 作为 Apache 的一个开源项目,稳定性不如 Cloudera Manager;部署的宿主机内存要求较高,不然 Ambari Server 节点启动不了;屏蔽很多细节,妨碍初学者对 Hadoop 生态圈各组件的理解。但此种基于 Ambari 的 Hadoop 集群快速部署方式省去了通常需要耗时长达几天且复杂的 Storm、Spark、Kafka、Flume 等各组件的兼容性及编译问题及繁琐的组件参数配置等,非常适合在高校教学环境、大规模 Hadoop 节点部署中使用,让学生尽快进入大数据平台,提高大数据预处理、大数据分析、大数据可视化以及综合应用等操作的实践能力。

参考文献(References):

[1] 施芬. 企业大数据精准营销的接受意愿影响因素分析:基于整合 UTAUT 模型与 4C 理论[J]. 重庆工商大学学报(社会科学版),2019,36(1):62—71
SHI F. Analysis of Impacting Factors of Consumers Willingness to Accept Precise Marketing of Big Data: Based on Integrated UTAUT Model and 4C Theory [J]. Journal of Chongqing Technology and Business University (Social Sciences Edition), 2019, 36 (1): 62—71 (in Chinese)

[2] 沈贵庆. 大数据分析在高校智慧教育中的应用研究[J]. 现代电子技术,2019,42(4):97—100
SHEN G Q. Application Research of Big Data Analysis in College Wisdom Education [J]. Modern Electronics Technique,2019,42(4):97—100(in Chinese)

[3] 薛志云,何军,张丹阳,等. Hadoop 和 Spark 在实验室中部署与性能评估[J]. 实验室研究与探索,2015,34(11):77—81
XUE Z Y,HE J,ZHANG D Y,et al. The Deployment and Performance Evaluation of Hadoop and Spark in Laboratory Environment[J]. Research and Exploration in Laboratory,2015,34(11):77—81(in Chinese)

[4] 孟永伟,黄建强,曹腾飞,等. Hadoop 集群部署实验的设计与实现[J]. 实验技术与管理,2015,32(1):145—149
MENG Y W,HUANG J G,CAO T F,et al. Design and Implementation of Deploying Hadoop Cluster Experiment [J]. Experimental Technology and Management,2015,32(1):145—149(in Chinese)

[5] 许鑫,时雷,何龙,等. 基于 NoSQL 数据库的农田物联网云存储系统设计与实现 [J]. 农业工程学报,2019,35(1):172—179
XU X,SHI L,HE L,et al. Design and Implementation of Cloud Storage System for Farmland Internet of Things Based on NoSQL Database [J]. Transactions of the Chinese Society of Agricultural Engineering, 2019, 35

(1):172—179(in Chinese)

[6] Hortonworks Corporation. Hortonworks Data Platform [EB/OL]. https://hortonworks com/products/data-platforms/hdp,2019—05—18

[7] 李可,李昕. 基于 Hadoop 生态集群管理系统 Ambari 的研究与分析[J]. 软件,2016,37(2):93—97
LI K, LI X. The Analysis and Study of Cluster Management System Ambari Based on Hadoop Eco Cluster[J]. Computer Engineering & Software, 2016, 37(2):93—97(in Chinese)

[8] 于金良,朱志祥,李聪颖. Hadoop 平台的自动化部署与监控研究[J]. 实验室研究与探索,2016,44(12):2457—2461
YU J L, ZHU Z X, LI C Y. Hadoop Platform Automated Deployment and Monitoring [J]. Computer & Digital Engineering,2016,44(12):2457—2461(in Chinese)

[9] 李杰,刘广钟. Hadoop 分布式集群的自动化容器部署研究[J]. 计算机应用研究,2016,33(11):3404—3407
LI J, LIU G Z. Research on Automatic Deployment of Hadoop Distributed Cluster [J]. Application Research of Computers,2016,33(11):3404—3407(in Chinese)

Research on Rapid Deployment of Hadoop Cluster Based on Ambari

WU Li-jie, ZHANG Lu-lu, ZHANG Ting

(Anhui Vocational College of Grain Engineering, Hefei 230011, China)

Abstract: As an excellent distributed computing framework to deal with big data, Hadoop is very popular in enterprises. However, the deployment of Hadoop cluster needs to consider the compatibility of each component, compilation problems and tedious component parameter configuration, and beginners often cannot deploy successfully even in several days. Ambari is an open source tool that supports Hadoop cluster deployment, monitoring and management. In view of the complexity of Hadoop cluster deployment, this paper puts forward the practical method of deploying each component of Hadoop cluster based on Ambari tool, and discusses some key points and important steps of rapid deployment. Through Ambari tool, the rapid deployment of most common components of Hadoop ecosphere minimization cluster has been completed, such as HDFS, HBase, Hive, Pig, Oozie, Zookeeper, Sqoop, Spark, Storm, Kafka, Flume and so on. Project practice shows that Hadoop cluster can be deployed within 8 hours by using Ambari tools. Compared with the traditional manual deployment, Ambari tools greatly improve the efficiency and success rate of Hadoop cluster deployment.

Key words: Hadoop; Ambari; HDP; big data; rapid deployment

责任编辑:田 静

引用本文/Cite this paper:

吴丽杰,张璐璐,张婷. 基于 Ambari 的 Hadoop 集群快速部署研究[J]. 重庆工商大学学报(自然科学版),2020,37(1):42—48

WU L J, ZHANG L L, ZHANG T. Research on Rapid Deployment of Hadoop Cluster Based on Ambari[J]. Journal of Chongqing Technology and Business University (Natural Science Edition),2020,37(1):42—48