

一文看懂 Cloudera 对 CDH/HDP/CDP 的产品支持策略



michaelli

恒生电子 大数据架构师

前言

熟悉大数据业界的小伙伴们都知道，Cloudera 在跟HortonWorks 合并后，便推出了新一代大数据平台 CDP，并正在逐步停止原有的大数据平台 CDH 和 HDP。

但具体到何时会彻底停止对 CDH 和 HDP 的维护，CDH/HDP 停止维护后现有用户该如何应对，以及 CDP 跟 CDH/HDP 的异同，或多或少都有些模式。

笔者希望通过本文，能彻底回答清楚小伙伴们的上述问题。以下是正文。

CDH/HDP 各版本的关键时间点

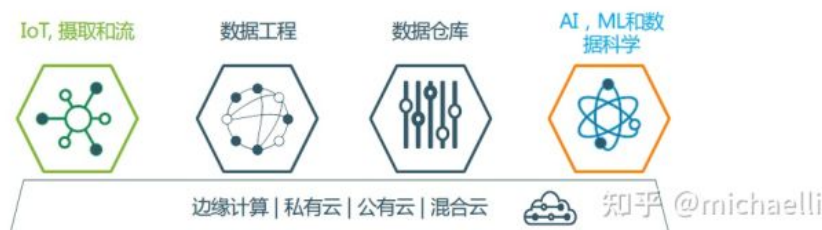
1. 下图简要介绍了CDH和HDP侧重的技术领域：

- 可以看到CDH和HDP都重点涵盖了数据工程和数据仓库场景
- 同时CDH对AI,ML和数据科学场景有侧重
- 而HDP对IoT数据摄取和流场景有侧重

2019年1月3日，Cloudera和Hortonworks正式合并



- Cloudera (代表产品CDH – Cloudera Distribution of Hadoop)
- Hortonworks (代表产品HDP – Hortonworks Data Platform)



1. 下图简要介绍了目前市面上主流CDH和HDP版本的关键时间点

- HDP2.x和CDH5.x对应的是hadoop2.x
- HDP3.x和CDH6.x对应的是hadoop3.x
- 目前官方已经停止了对HDP2.x和CDH5.X的技术支持
- 官方对HDP3.x和CDH6.x的技术支持也都会在最近半年停止
- 具体来讲，HDP3.x的最新版HDP3.1会在2021/12月停止技术支持
- 具体来讲，CDH6.x的最新版CDH6.3会在2022/3月停止技术支持

CDH和HDP所有版本在最近半年停止技术支持，不会再有新特性增强

Cloudera支持的生命周期策略，包括产品版本生命周期的重要信息，包括预计每个版本宣布支持终止的预计日期。

Release	GA Date	EoS Date
Cloudera Enterprise 6.3	Jul-2019	Mar-2022
Cloudera Enterprise 6.2	Mar-2019	Mar-2022
Cloudera Enterprise 6.1	Dec-2018	Dec-2021
Cloudera Enterprise 6.0	Aug-2018	Aug-2021
Cloudera Enterprise 5.16	Nov-2018	Dec-2020
Cloudera Enterprise 5.15	Jun-2018	Dec-2020
Cloudera Enterprise 5.14	Jan-2018	Dec-2020
HDP 3.1	Dec-2018	Dec-2021
HDP 3.0	Jul-2018	Jul-2021
HDP 2.6	Apr-2017	Dec-2020
HDP 2.5	Aug-2016	Aug-2019

<https://www.cloudera.com/legal/policies/support-lifecycle-policy.html>

CLUSTERA Confidential - Restricted

CDP 介绍

1. CDP 可以认为是将原来的 CDH/HDP 融合在了一起，具体融合方式如下图所示，关键点是：

- 淘汰了竞争的技术
- 融合了重叠的技术
- 保留了互补的技术
- 升级了共享的技术
- 并增加了某些新功能

CDH/HDP与CDP的组件比较

Cloudera Runtime—由CDH和HDP的精华创建

淘汰竞争的技术

Ambari	Cloudera Manager
Sentry	Ranger
Cloudera Director	Cloudbreak
Hive-on-Spark	Hive on Tez
Pig	

融合重叠的技术

Navigator + Atlas + DSS
WXM + DAS
BDR + DLM
Hue + DAS Lite

保留互补的技术

Impala	ORC
Parquet	Hive ACID
Kudu	Zeppelin
CDSW	Phoenix
NiFi	Livy
Knox	

升级共享的技术

Hadoop 3.1	Spark 2.4/3.1
Hive 3.1	Oozie 5.1
HBase 2.2	Accumulo 2
Kafka 2.5	Sqoop 1.4
Solr 8.4	Zookeeper 3.5

Cloudera Data Platform(CDP)

两个最好的开源数据分析平台(CDH/HDP)融合在一起并增加了新功能

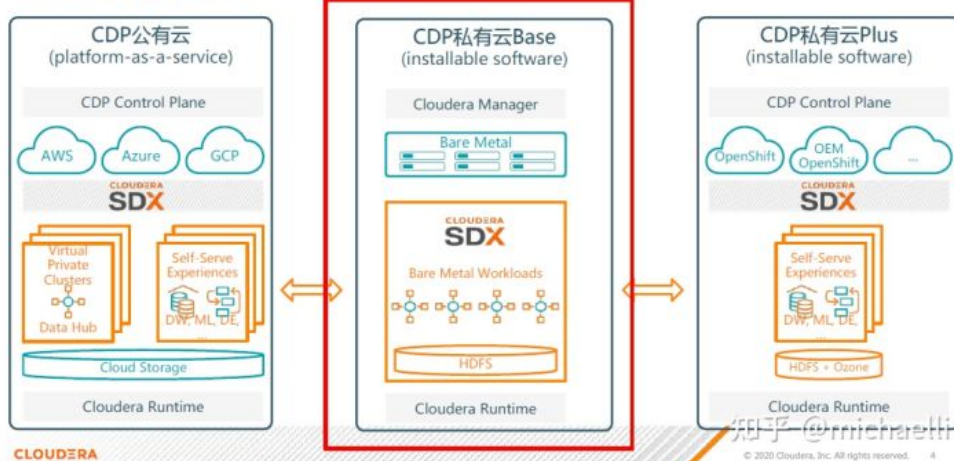
此升级引入了多项新功能（和产品）、增强功能、错误修复，并提供具有云功能的下一代企业分析平台



2. CDP 对应不同场景，推出了三大版本：

- 对应公有云场景的 CDP public cloud,目前三大公有云厂商 aws, gcp, azure都已经提供了支持;
- 对应私有云场景的 CDP private cloud plus, 其计算集群使用了 docker 和 k8s 相关技术;
- 对应数据中心场景的 CDP private cloud base, 其实对应的就是原来场景的 CDH 和 HDP;
- 三大版本底层对应的是同样的 cloudera runtime,其实质就是大数据各个具体组件, 如 hdfs/hive/spark等等。

CDP三大部署版本



CDP的主要技术优势

从CDH/HDP到CDP



Cloudera 的新许可证政策

Cloudera 的新许可证政策如下图所示，其关键点是：

- Cloudera 所有产品都将开源 (至少承诺的是如此，包括原来CDH中不开源的cloudera manager等也将开源)，其开源模型类似 Red Hat开源模型
- Cloudera 所有产品的二进制文件和源代码，都需要订阅，（即不再提供社区版，只提供企业版，都需要付费，不再有免费的午餐！）
- 自2019年11月以来，访问产品的二进制文件需要订阅和 paywall credentials (即没有paywall credentials 的话,将不再能从 cloudera 官网下载 parcel/rpm 包)
- 2021年1月后，扩展的Paywall将包括平台的早期版本，包括所有版本的CDH/HDP/HDF等（即目前所有版本的CDH/HDP/HDF/CDP，从官网下载 parcel/rpm 包,都需要有 paywall credentials）

CLUSTERA 的新许可证政策

我们的许可政策类似于成熟的Red Hat开源模型。该模型在市场上广为人知，并被全世界的企业所接受。

已有产品的许可证将在整个21财年发生变化

自2019年11月以来，访问产品的二进制文件需要订阅和paywall credentials

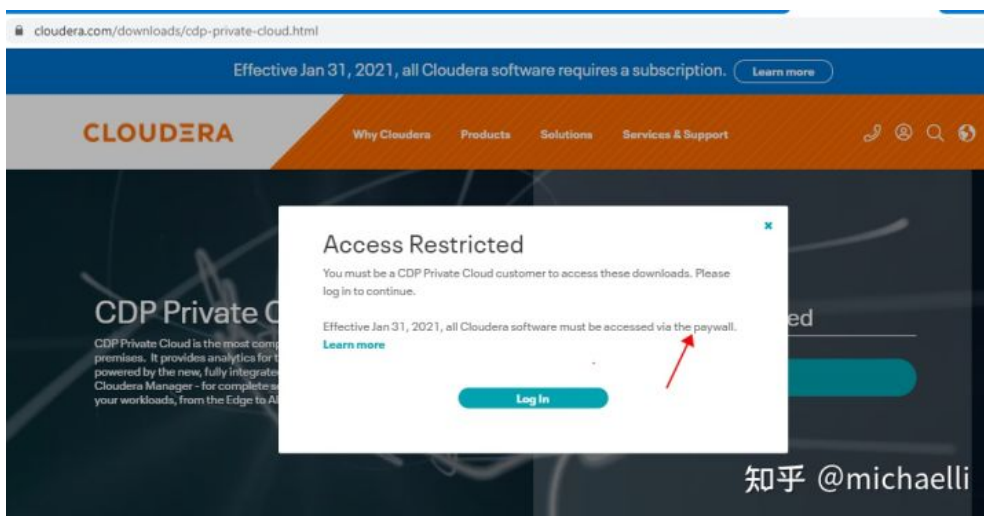
从2019年11月后，Cloudera再发布的产品不再提供社区版（免费版），只有企业版（需要订阅才能正常使用）。

2021年1月后，扩展的Paywall将包括平台的早期版本，包括所有版本的CDH/HDP/HDF等

- 所有产品源均将获得OSI批准的许可证(Apache软件许可证(ASL)或Aferro通用公共许可证(AGPL))的许可。
- 访问从Cloudera获得的所有产品的二进制文件和源代码都需要订阅协议
- 由Apache Software Foundation托管的我们为之贡献的所有开源项目将继续是ASF管辖的项目。
- 我们的源代码贡献模型保持不变，并且仍然是第三方项目的第一个上游。第三方项目存储库的公共资源仅在上游。

CLUSTERA

知乎 @michaelli



使用遗留CDH/HDP系统的小伙伴们该何去何从

概括起来，使用遗留CDH/HDP系统的小伙伴们，有以下选择：

1. 继续使用原有版本的CDH/HDP：在指定日期之后，Cloudera官方不再对原有版本的CDH/HDP提供技术支持，这仅仅意味着Cloudera官方不会再对原有版本提供新特性增强，也不再对原有版本提供BUG修复，但客户原有的大数据平台仍然是能够正常提供服务的（这点不同于星环的TDH，TDH在许可证到期之后，整个集群中的服务就不能再重启，不能在正常提供服务了）；
2. 考虑市面上其他供应商的大数据平台，如星环的TDH，或基于开源apache版本自行封装。不过需要注意，星环的TDH是闭源的，其一些参数跟开源的并不兼容，有 vendor lock in 的风险；
3. 按照Cloudera的建议，在合适的时机，升级到 CDP平台。

如何从 CDH/HDP 迁移到CDP?

Cloudera 提供了一系列工具帮助大家尽量平滑地从CDH/HDP升级到CDP。

1. Cloudera官方建议的升级方式有四种方式：

- 原地升级
- 拷贝升级/迁移升级
- 滚动式拷贝升级/迁移升级
- 迁移到公有云

从传统发行版到CDP的四种升级和迁移路径

升级和迁移路径	描述
原地升级	<ul style="list-style-type: none">将现有的旧版CDH或HDP集群直接升级到CDP的过程。涉及计划内的停机时间、并且需要所有租户之间进行协调一致
拷贝升级	<ul style="list-style-type: none">在第二套硬件上配置了一个新的未开发的CDP PvC Base集群，配置复制过程来复制数据、元数据和治理策略等，将工作负载部署到新集群上进行测试验证，迁移完成后在就集群上禁用对应的工作负载。最大程度的减少单个工作负载的停机时间，同时提供直接的回滚机制。
滚动式拷贝升级	<ul style="list-style-type: none">对典型拷贝升级的修改。将旧集群的容量用于新集群。工作负载和数据迁移将重复多次，直到所有租户、工作负载和数据完全迁移到CDP环境为止。
迁移到公有云	<ul style="list-style-type: none">和滚动式拷贝升级路径非常类似。数据复制目标从HDFS改到了云对象存储。

<https://blog.cloudera.com/the-four-upgrade-and-migration-paths-to-cdp-from-legacy-distributions/>

CLOUDERA

知乎 @michaelli
© 2020 Cloudera, Inc. All rights reserved. 33

1. 不考虑迁移到共有云的话，大家可以主要考量两种迁移方式，即原地升级和拷贝升级/迁移升级，其优缺点如下图所示：

两种到达CDP的升级方式

原地升级	迁移升级
<ul style="list-style-type: none">❑ 卸载旧的CDH/HDP，保留原有数据，然后安装新版CDP并升级❑ 优点<ul style="list-style-type: none">● 不需要任何额外硬件资源❑ 缺点<ul style="list-style-type: none">● 服务停机时间相对较长● 需要多重验证❑ 目标客户<ul style="list-style-type: none">● 没有富余的主机资源用做中间集群● 能够接受较长的服务停机时间	<ul style="list-style-type: none">❑ 安装新的CDP集群，然后将现有数据拷贝至新集群，将新的CDP集群切换为生产集群❑ 优点<ul style="list-style-type: none">● 没有数据丢失风险● 较短的服务停机时间❑ 缺点<ul style="list-style-type: none">● 额外的硬件资源● 需要迁移数据● 整体升级周期较长❑ 目标客户<ul style="list-style-type: none">● 有富余的主机资源（新节点或者缩编现有节点）用做中间集群● 要求极短的服务宕机时间

CLOUDERA

知乎 @michaelli

各种迁移方式的详细对比，和操作手册，Cloudera官网有消息描述，笔者在这里就不再赘述了。

！关注不迷路~ 各种福利、资源定期分享！欢迎小伙伴们扫码添加明哥微信，后台加群交流学习。

发布于 2021-08-26 00:03

cloudera

大数据

Hortonworks