数据迁移

大数据迁移指引

产品文档

腾讯云

大数据迁移指引

最近更新时间：2020-10-28 14:55:30

大数据接入方案提供完整的大数据平台数据接入流程和方式。

大数据迁移场景

腾讯云提供了托管 Hadoop 集群的产品（EMR），同时用户也可以选择在腾讯云上自建 Hadoop 集群，可根据自身的实际情况，选择 CVM 或者 CPM。

真实情况下的迁移场景有：

- 本地 Hadoop 集群迁移至腾讯云 EMR。
- 本地 Hadoop 集群迁移至腾讯云自建集群。
- 第三方云 Hadoop 集群迁移至腾讯云 EMR。
- 第三方云 Hadoop 集群迁移至腾讯云自建集群。

大数据迁移方式

普通迁移

可将本地 HDFS 中的数据通过迁移工具（例如 Distcp 等）迁移至目标环境。该方法比较通用，适用于多数场景下的大数据应用（例如实时计算）。

迁移可参考以下流程：

1. 打通源和目标的网络连接
   如果源是本地自建 Hadoop 集群或者第三方云，建议搭建专线连接到目标。源和目标都在腾讯云的情况下，如果源和目标在同一 VPC 网络则可自由拷贝。如果源和目标不在同一 VPC 则需要先建立对等网络。
2. 使用工具执行迁移
   事先确认源和目标为相同的版本。可全量迁移，也可选择指定文件的迁移。

   > ⚠ 注意：
   > 迁移过程中源如果有写入操作可能会导致迁移失败。

3. 验证迁移结果，完成迁移

计算存储分离的迁移方式

可将本地 HDFS 中和实时计算关系不大的数据迁移至 COS，然后配置数据从 COS 读取，可以很大降低存储成本。适用于离线计算场景。迁移工具可以参考腾讯云提供的迁移工具（HDFS_TO_COS）。

需要注意的问题：

1. 请确保填写的配置信息正确，包括 AppID、密钥信息、bucket 和 region 信息，以及机器的时间和北京时间一致（如相差1分钟左右是正常的）。如果相差较大，请设置机器时间。
2. 请保证对于 DateNode，拷贝程序所在的机器也可以连接。因 NameNode 有外网 IP 可以连接，但获取的 block 所在的 DateNode 机器是内网 IP，无法连接。因此建议同步程序放在 Hadoop 的某个节点上执行，保证对 NameNode 和 DateNode 皆可访问。
3. 权限问题，在当前账户使用 hadoop 命令下载文件，确认是否正常，再使用同步工具同步 hadoop 上的数据。
4. 对于 COS 上已存在的文件，默认进行重传覆盖，除非用户明确的指定`-skip_if_len_match`，当文件长度一致时则跳过上传。
5. cos path 都认为是目录，最终从 HDFS 上拷贝的文件都会存放在该目录下。