弹性 MapReduce

常见问题

产品文档

腾讯云

版权所有：腾讯云计算（北京）有限责任公司

【版权声明】

©2013-2019 腾讯云版权所有

本文档著作权归腾讯云单独所有，未经腾讯云事先书面许可，任何主体不得以任何形式复制、修改、抄袭、传播全部或部分本文档内容。

【商标声明】

腾讯云

及其它腾讯云服务相关的商标均为腾讯云计算（北京）有限责任公司及其关联公司所有。本文档涉及的第三方主体的商标，依法由权利人所有。

【服务声明】

本文档意在向客户介绍腾讯云全部或部分产品、服务的当时的整体概况，部分产品、服务的内容可能有所调整。您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或模式的承诺或保证。

常见问题

最近更新时间：2019-08-26 10:25:34

如何查看任务日志？

您可以登录任意一台 EMR 服务器执行以下命令查看任务日志：

```
yarn logs -applicationId application_1507732460084_0057
```

> 注意：
>
> - 需以 Hadoop 用户身份执行该命令。
> - 如果是其他用户的任务可以添加参数 -appOwner username。

如需查看任务异常原因可通过以下命令实现：

```
yarn logs -applicationId application_1507732460084_0057|grep -A20 Exception
```

如何调整集群计算资源？

集群计算资源由 yarn-site.xml 中的以下两项配置决定：

```
<property>
<name>yarn.nodemanager.resource.cpu-vcores</name>
<value>4</value>
</property>
<property>
<name>yarn.nodemanager.resource.memory-mb</name>
<value>14745</value>
</property>
```

默认情况下 cpu-vcores 等于机器的 CPU 核数，memory-mb 等于机器内存的91%，可以根据实际情况作出调整，如果设置太大则存在机器宕机的风险。

如何处理任务执行时内存溢出？

提交 MR 任务或者通过 Hive 执行 SQL 脚本时发生内存溢出可以通过设置以下参数处理：

- `set mapreduce.map.java.opts=-Xmx4096m;`
- `set mapreduce.reduce.java.opts=-Xmx4096m;`

版权所有：腾讯云计算（北京）有限责任公司

可以根据计算需要调整内存参数，如果是 Hive 也可写在 `~/.hiverc` 文件下，提交的时候会自动执行。

如何预估集群规模？

假设您的一次运算以 SQL 执行为例，如果想要在确定的时间里查询到结果需要的 vcore 为64个，内存为128GB，业务要求一次要支持10个并发，那么需要的资源为 vcore 640个，内存1280GB，假设采用24核48GB的设备，那么需要的计算设备量为：1280 / 48约等于40台。

如何设置 Hive 的 fetch 查询？

Hive 默认查询如下：

```
select * from tablename where a=' 1' limit 10;
```

默认查询不会启动计算任务，您可以通过添加 `set hive.fetch.task.conversion=none` 参数开启分布式查询。

如何选择集群存储介质？

EMR 集群支持如下存储介质，普通本地盘、SSD 本地盘、普通云硬盘，SSD 云硬盘以及对象存储 COS，您可以根据实际需要来选择存储介质：

- 如果您的应用场景是大规模数据仓库分析，对时延不是那么敏感，建议您使用 COS 作为底层存储。
- 如果您非常熟悉 HDFS 而且使用 COS 迁移成本过高，您也可以使用普通云盘。
- 如果您的应用是海量列式数据库 Hbase，需要高效写入和查询，建议您使用本地 SSD 盘或者 SSD 云硬盘。

版权所有：腾讯云计算（北京）有限责任公司