NodeManager节点自身健康状态检测机制

每个 NodeManager 节点内置提供了检测自身健康状态的机制（详情参见 NodeHealthCheckerService）；通过这种机制，NodeManager 会将诊断出来的监控状态通过心跳机制汇报给 ResourceManager，然后ResourceManager 端会通过 RMNodeEventType.STATUS_UPDATE 更新 NodeManager 的状态；如果此时的 NodeManager 节点不健康，那么 ResourceManager 将会把 NodeManager 状态变为 NodeState.UNHEALTHY 。关于 NodeManager 的状态之间的转移请参见NodeManager生命周期介绍 。这种内置的健康检测机制主要包括以下两种：

- 健康状况检测脚本
- 本地目录健康检测

健康状况检测脚本

每个 NodeManager 上都有一个名为 NodeHealthScriptRunner 的类，其会启动一个名为 NodeHealthMonitor-Timer 的 Timer 定期执行（默认每个十分钟实行一次，由参 yarn.nodemanager.health-checker.interval-ms 指定）用户编写的用于检测 NodeManager 健康状态的脚本，这个脚本是通过参数 yarn.nodemanager.health-checker.script.path 指定的。一旦发现以下几种结果则认为节点处于不健康状态：

- 脚本输出包含以"ERROR"开头的字符串，不顾是一个还是多个都算
- 执行脚本的时候出现超时
- 执行脚本出现异常

如果执行脚本出现了 ExitCodeException 异常，并不认为 NodeManager 节点健康出现异常，因为我们的脚本可能编写错误。NodeHealthScriptRunner 类涉及到的参数如下：

- yarn.nodemanager.health-checker.script.path：这个就是 Hadoop 集群管理员编写的用于健康检测的脚本路径，如果这个属性没有设置， NodeManager 则不会启动 NodeHealthScriptRunner。
- yarn.nodemanager.health-checker.interval-ms：每个多久执行健康脚本，默认是10分钟。
- yarn.nodemanager.health-checker.script.timeout-ms：执行健康检测脚本超过了这个属性配置的时间，则认为节点不健康。默认值是20分钟。
- yarn.nodemanager.health-checker.script.opts：健康检测脚本的输入参数，如果有多个请使用空格分割。

健康检测脚本这种机制有点好处：目前的 YARN 资源管理主要是内存核CPU，其他的比如磁盘使用、系统负载、网络等并没有监控到，我们可以使用这种机制来主动告诉 ResourceManager 自身的监控状态。

下面是一个简单的健康检测的脚本（假设脚本的保存路径为/user/iteblog/check_memory_usage.sh），如果内存的使用率达到了 95% 以上，则认为此节点处于不健康的状况：

```
#!/bin/bash

echo $1
echo $2

mem_usage=$(echo `free -m | awk '/^Mem/ {printf("%u", 100*$3/$2);}'`)
echo "Usage is $mem_usage%"
if [ $mem_usage -ge 95 ] then
  echo 'ERROR: Memory Usage is greater than 95%'
else
  echo 'NORMAL: Memory Usage is less than 95%'
fi
```

然后我们可以在 NodeManager 进行如下配置：

```
<property>
  <name>yarn.nodemanager.health-checker.script.path</name>
  <value>/user/iteblog/check_memory_usage.sh</value>
</property>

<property>
  <name>yarn.nodemanager.health-checker.script.opts</name>
  <value>第一个参数 第二个参数</value>
</property>
```

本地目录健康检测

除了管理员提供的健康检测脚本之外，NodeManager
还提供了检测磁盘好坏的机制。检测的磁盘目录主要是 yarn.nodemanager.local-dirs 和 yarn.nodemanager.log-dirs 参数指定的目录，这两个目录分别用于存储应用程序运行的中间结果，比如MapReduce作业中Map Task的中间输出结果）和日志文件存放目录列表。这两个参数都可以配置多个目录，多个目录之间使用逗号分隔。如果这两个参数配置的目录不可用的比例达到一定的设置，则认为该节点不健康。某个目录可不可用的定义是：运行 NodeManager 节点的进程是否对这个目录可读、可写、可执行。如果这些条件都满足，这个目录则健康，否则该目录就被放入 failedDirs 列表里面。本地目录健康检测主要涉及到以下几个参数：

- yarn.nodemanager.disk-health-checker.interval-

ms：本地目录健康检测线程执行的频率，默认值为2分钟；
- yarn.nodemanager.disk-health-checker.enable：是否启用本地目录健康检测，默认值是启用；
- yarn.nodemanager.disk-health-checker.min-healthy-disks：正常目录数目相对于总目录总数的比例，低于这个值则认为此节点处于不正常状态，默认值为0.25。

以上两种检测机制都会随着 NodeManager 节点启动而运行，并且检测到的状态会随心跳信息发送到 ResourceManager 端，然后 ResourceManager 端会根据相关的信息得到当前节点的可用情况，一旦发现这个节点不健康，则会标记此节点的状态为 NodeState.UNHEALTHY ，此后将不会忘这个节点分配任务，直到该节点状态正常。