测试度量

在性能测试基准中，经常以吞吐量、响应延迟作为度量指标。在软件运行的可靠性测试中，已有的方法通常使用故障平均时间来衡量系统的可靠性[1]。由于大数据系统运行在多节点的集群环境下，系统的可靠性无法使用常规的故障平均时间来衡量。同时，Gillies等人[2]也指出，由于每个系统以及系统的不同部分的目的是不同的，因此评估可靠性的方式也会有变化。

本文根据在2.2.1节中给出的大数据系统应用的可靠性定义，首先将以下故障或软件错误称为系统不可靠的评测标准：

1. 性能出现异常。如，用户给定的时间内，系统出现无响应、假死等现象。
2. 资源使用异常。如，出现I/O异常、内存溢出、磁盘异常等现象。
3. 计算结果异常。如，计算结果错误、计算数据丢失或重复计算等现象。

由上述系统不可靠的评测标准，本文将大数据系统的可靠性度量指标定义为：在用户给定的系统配置、以及指定的应用领域（如SQL、Graph、Machine Learning等）内，系统不可靠的频率，即。

参考文献

1. Rosenberg, Linda, Ted Hammer, and Jack Shaw. "Software metrics and reliability." 9th International Symposium on Software Reliability Engineering. 1998.
2. Gillies, Alan. Software quality: theory and management. Lulu. com, 2011.