

什么是大数据

大数据是指无法在一定时间范围内用传统的计算机技术进行处理的海量数据集。

对于大数据的测试则需要不同的工具、技术、框架来进行处理。

大数据的体量大、多样化和高速处理所涉及的数据生成、存储、检索和分析使得大数据工程师需要掌握极其高的技术功底。

需要你学习掌握更多的大数据技术、Hadoop、Mapreduce等等技术。

大数据测试策略

大数据应用程序的测试更多的是去验证其数据处理而不是验证其单一的功能特色。

当然在大数据测试时，功能测试和性能测试是同样很关键的。

对于大数据测试工程师而言，如何高效正确的验证经过大数据工具/框架成功处理过的至少百万兆字节的数据将会是一个巨大的挑战。

因为大数据高效的处理测试速度，它要求测软件工程师具备高水平的测试技术才能应对大数据测试。

我们来看下大数据处理的三个特性：

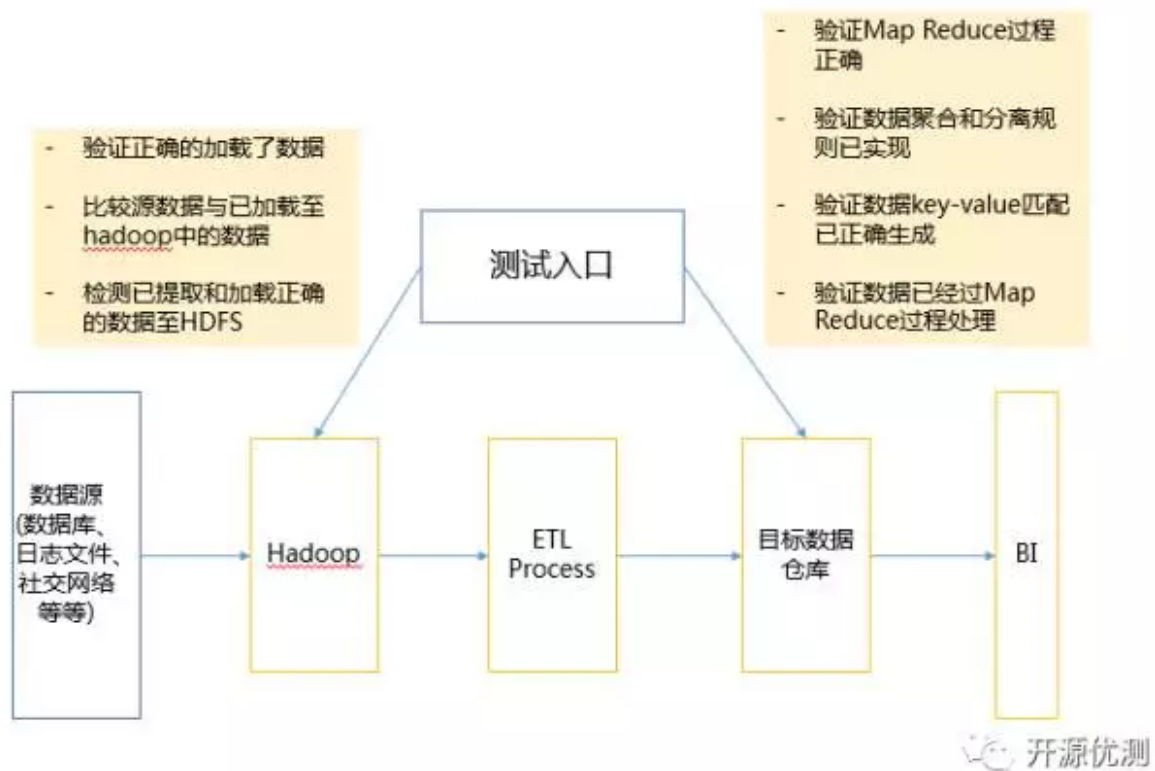
- 大批量
- 实时性
- 可交互

另外，数据质量也同样是大数据测试的一个重要维度。

因此在进行应用程序测试之前，必须确保数据质量，并且考虑把数据质量作为数据库测试的一部分。涉及数据的各种特性的检验，例如一致性、准确性、重复性、连贯性、有效性及完整性等等。

大数据应用测试步骤

下面我们一起来看看大数据应用的测试过程是怎么样的。



整体而言，大数据测试大体可以分为三大步骤：

- 步骤一，数据预处理验证

在进行大数据测试时，首先要预hadoop前验证数据的准确性等等。

1. 我们数据来源可能是关系数据库、日志系统、社交我那个落等等，所有我们应该确保数据能正确的加载到系统中
2. 我们要验证加载的数据和源数据是一致的
3. 我们要确保正确的提取和加载数据至hdfs中

- 步骤二，Map Reduce验证 在进行大数据测试时，第二个关键步骤是“Map Reduce”验证。在本阶段，我们主要验证每一个处理节点的业务逻辑是否正确，并验证在多个运行后，确保：

1. Map Reduce过程工作正常
2. 数据聚合、分离规则已经实现
3. 数据key-value关系已正确生成
4. 验证经过map reduce后数据的准确性等特性

- 步骤三，结果验证 在本阶段主要验证在经过大数据工具/框架处理后，生成的最终数据的成果。

主要验证：

1. 验证数据转换规则是否正确应用
2. 验证数据的完整性和是否成功持久化到目标系统
3. 验证无数据损坏

架构测试

Hadoop处理海量数据是非常的消耗资源的，良好的架构是确保大数据项目成功的基础。糟糕的涉及会导致性能急剧的下降，进而使得系统无法满足我们的需要，因此我们需要，或是说至少在Hadoop环境下进行性能测试、故障恢复测试，以应改进效率和应对可能的最糟糕的情况。

性能测试是一个复杂的工作，它贯穿整个测试周期，需要关注内存、CPU、网络等等指标。故障恢复测试则是验证数据处理过程中可能出现的故障，为做好意外的恢复做好相应的应对措施。

性能测试

大数据性能测试主要包含以下几个部分：

- 数据提取、存储效率

在本阶段，我们主要验证大数据应用从源数据中提取、加载数据的效率。

一是验证单位时间内数据的提取、加载效率。

二是验证数据持久化至mongodb等库的效率等等

- 数据处理

在本阶段，我们验证map reduce任务的执行效率，重点关注的是数据处理的效率。当然这个过程可能也会涉及到数据的持久化相关指标，例如存储至HDFS读写效率等等，同样也会涉及在内存中处理效率，即我们的处理算法效率等等

- 子组件性能

大数据处理，一般都会需要综合利用各种组件来辅助处理，所以我们也是需要关注这些辅助组件的性能

性能测试策略

大数据应用性能测试涉及海量的结构化和非结构化的数据，与我们平时所面对的业务系统有所不同，所以我们需要针对大数据应用制定特定的测试策略，以应对海量的数据。

根据上图性能测试执行过程一般是这样的：

1. 在性能测试前需要先初始化大数据集群环境
2. 梳理和设计大数据性能测试场景
3. 准备大数据性能测试脚本
4. 执行并分析测试结果（如果指标异常，则调优相应的组件并重新测试）
5. 优化配置

性能测试基础准备

在大数据性能测试时，需要准备相关的基础工作，如下：

- 数据准备，我们需要在不同的节点准备什么量级数据？
- 日志预估，在测试过程中，可能会生成多大的日志，日志的可能增量是什么样的？

- 并发，在测试时，可能会有多少线程并发读和写？
- 超时设置，应对设置怎样的连接超时？查询超时？写超时等等？
- JVM参数，如何设置最优的jvm参数，heap size、GC机制等等
- Map Reduce，我们应该选择什么样的sort、merge等算法？
- 消息队列，消息队列长度会怎么样？等等

必备的测试环境

大数据测试不同于常规的应用测试，你应该具备以下一些基础环境：

- 拥有足够的存储设备来存储和处理大数据
- 拥有集群来做分布式节点和数据处理
- 至少拥有足够的cpu、内存来确保有高性能的处理基础

大数据测试的挑战

对于从事大数据测试的软件测试工程师而言，与传统的测试工作相对比，我们可能面临的以下几个可能的挑战：

- 自动化

自动化测试是从事大数据测试必备的技术，但自动化测试工具可能并不具备处理测试过程所引发的异常的能力，意味着现有工具可能并不适用，编程能力将是更好的一种技能。

- 虚拟化 当前业内大规模使用虚拟化技术，但虚拟机的延迟有可能造成大数据实时测试处理的异常。

对大数据而言，管理影像信息也将是一个巨大的问题。

- 海量数据集
 1. 需要验证的数据量巨大，而且需要更快的处理速度
 2. 需要有效的自动化测试手段
 3. 需要尽可能的跨平台

大数据性能测试的挑战

对于从事大数据性能测试，与传统性能测试相比较，我们要面临是样的挑战呢，可能以下几个方面：

- 技术的多样化，复杂化，面对不同的大数据解决方案，我们可能需要掌握不同的技术和定制不同的测试解决方案

- 无通用的工具，目前业界暂无通用的标准的大数据性能测试工具，这意味着我们需要根据大数据应用解决方案技术，要自行开发或整合多种相关工具才可能解决问题
- 测试环境复杂化，因为海量的数据，我们所需要测试环境亦会更加复杂，所消耗的基础成本会更高
- 监控解决方案，目前有的监控解决方案有限，但通过整合不同的监控工具，大致可能拥有一套相对可行的监控解决方案
- 诊断方案，由于大数据应用所涉及的技术、环境复杂性，对于问题的诊断调优，我们需要根据实际情况来进行开发定制

从上面几个方面来看，从事大数据性能测试所要面临的问题是相对复杂的，尤其对当下国内的测试工程师而言，要走的路还很长，很艰难。

小结

- 随着大数据工程和数据分析逐步的进入新的阶段，大数据测试将成为必然，也必定成为未来的一个热门的职业方向
- 大数据处理必须是批量的，实时的、可交互的
- 大数据应用测试的三大阶段：
 1. 数据验证
 2. Map Reduce 验证
 3. 数据处理结果验证
- 架构测试也是非常重要的一个测试类型，糟糕的架构可能直接导致您的大数据项目的失败
- 性能测试三大节点：
 1. 数据提取、存储效率
 2. 数据处理效率
 3. 子组件工作效率
- 大数据测试不同于传统的测试，不仅仅是类型、策略的不同，工具等具体技术都会有区别
- 大数据因其复杂性，其测试所面临的挑战也会不同于传统的测试
- 大数据性能测试将会是软件测试工程师进一步艰难攻克的目标之一