推荐序

由 Google 公司研发的 Google 文件系统和 MapReduce 编程模型以其 Web 环境下处理大规模海量数据的特有魅力，在学术界和工业界引起了非同小可的反响。以此为开端，学术界不断涌现出针对海量数据处理、立足于 MapReduce 的研究成果。而在工业界，大量类似于 Google 文件系统、采用类 MapReduce 编程模型的系统也得到了广泛的部署和应用。

今天，在像互联网应用、科学数据处理、商业智能数据分析等具有海量数据需求的应用变得越来越普遍时，无论是从科学研究还是从应用开发的角度来看，掌握像 Google 文件系统和 MapReduce 编程模型这样的技术已成为一种趋势。在这样的背景下，实现了 MapReduce 编程模型的 Hadoop 开源系统就成为大家一种自然而又合理的选择。

MapReduce 编程模型之所以受到欢迎并迅速得到应用，在技术上主要有三方面的原因。首先，MapReduce 所采用的是无共享大规模集群系统。集群系统具有良好的性价比和可伸缩性，这一优势为 MapReduce 成为大规模海量数据平台的首选创造了条件。

其次，MapReduce 模型简单、易于理解、易于使用。大量数据处理问题，包括很多机器学习和数据挖掘算法，都可以使用 MapReduce 实现。

第三，虽然基本的 MapReduce 模型只提供一个过程性的编程接口，但在海量数据环境、需要保证可伸缩性的前提下，通过使用合适的查询优化和索引技术，MapReduce 仍然能够提供相当好的数据处理性能。

显然，要真正掌握 MapReduce 编程技术，需要对上述技术有一个较为深入的了解，也需要熟悉支撑 MapReduce 的运行环境及系统的部署要求。非常令人兴奋的是，Hadoop 开源项目负责人 Tom White 所写的 *Hadoop: The Definite Guide* 一书为我们解决了这一问题。