专题：面向大数据处理的数据流计算技术

Dataflow Computing Techniques for Big Data Processing

客座编辑

周傲英（1965- ），男，博士，华东师范大学副校长、“智能+”研究院院长、数据科学与工程学院教授。现任第七届国务院学位委员会学科评议组成员，中国计算机学会会士，上海市计算机学会副理事长，《计算机学报》《大数据》期刊副主编。曾入选“长江学者计划”特聘教授，曾获国家杰出青年基金项目资助，主要研究方向为数据库、数据管理、数据驱动的计算教育学，以及教育科技（EduTech）、物流科技（LogTech）等基于数据的应用科技。

于戈（1962- ），男，博士，东北大学计算机学院教授、博士生导师，中国计算机学会会士。现任中国计算机学会信息系统专业委员会主任、数据库专业委员会委员、系统软件专业委员会委员，《计算机学报》《软件学报》《计算机研究与发展》等期刊编委。曾获得“教育部跨世纪人才基金”和“中国高校青年教师奖”。主要研究方向为分布式数据库系统、数据科学与大数据管理、区块链技术与应用等。

导读

数据流（data flow）是麻省理工学院（MIT）的Jack B. Dennis教授在20世纪70年代提出的一种计算机体系架构，这在当时是很大胆的想法。此前，冯•诺依曼在1946年提出的以存储程序和顺序执行为主要特征的体系结构是人们唯一的选择。相对于数据流，传统的体系结构被归为控制流（control flow）一类。与控制流相比，数据流计算有天然的并行性，这使得它在早期超级计算机的发展历史上产生了重要的影响。虽然数据流计算机至今没有成为主流，但是在大数据时代，计算机有史以来的“以计算为中心”真正转变成“以数据为中心”，数据流由于其自身的特点将重新焕发迷人的魅力。在我们承担的国家重点研发计划项目“面向异构体系结构的高性能分布式数据处理技术与系统”中，数据流是最重要的一个关键词，从面向用户的编程模型和工具到大数据处理的计算模型，再到GPU能力的充分发挥；从计算机集群资源管理到分布式缓存等数据管理，数据流计算的思想和技术是贯穿其中的一条主线。通过两年来的深入研究和比较，尤其是在系统开发和应用实践的过程中，项目组对于数据流在大数据处理中的应用有了较为深刻的认识，我们把涉及数据流计算关键技术的5篇文章汇集成“面向大数据处理的数据流计算技术”专题，以飨读者，恳请批评指正。

湖南大学邹骁锋等人将传统软件工程的面向数据流分析设计方法与当前流行的大数据处理平台的数据流编程模型的结构定义和模型参考进行了比较，给出了面向大数据处理的可视化数据流编程工具的基本框架和编程模式。

华东师范大学毕倪飞等人的文章介绍了大数据处理中的数据流计算模型，包括用以直观描述复杂的数据处理逻辑的执行引擎层面的数据流图，以及实现批、流统一处理的统一编程层面的数据流编程模型，分析了Spark批处理和Flink流计算中数据流图和数据流编程模型的具体实现。

西北工业大学汤小春等人的文章讨论了数据流编程模型在大数据处理领域应用带来的计算作业类型复杂化的问题，探讨了如何保证各种数据流计算作业对集群资源的共享使用，研究了数据流计算环境下的集群资源管理和调度。

东北大学袁旭初等人的文章讨论了数据流计算环境下的数据缓存问题。在Google Dataflow、Flink、Spark、TensorFlow等异构/分布式数据流计算系统中，算子和数据不再统一存在于单机内存，容易造成数据堆积或者算子闲置等问题。设计面向数据流的缓存系统，通过消息队列系统进行支持是未来的方向之一。

国防科技大学苏华友等人的文章从数据流模型的角度分析了英伟达GPU的体系结构以及CUDA编程模型，阐述了数据流模型在GPU软硬件系统中的应用，并分析了如何将数据流计算思想和GPU应用于大数据处理。

面向大数据处理的数据流计算技术具有广阔的发展前景。以上5篇文章自顶向下系统地介绍了数据流计算的关键技术，可以建构支持大数据分布式处理的全栈式数据流计算框架。但由于本专题篇幅有限，难以涵盖数据流计算技术的各个方面，期待通过分享我们的基本认识和实践经验，推动数据流计算技术在大数据应用领域更深入地开发和应用。