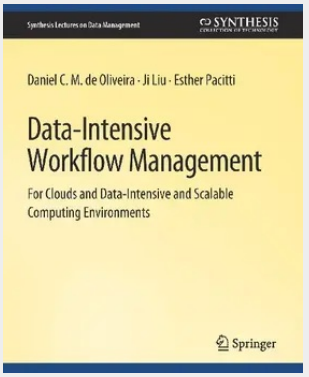
de Oliveira / Liu / Pacitti

Data-Intensive Workflow Management



64,19 €
59,99 € (zzgl. MwSt.)

Lieferfrist: bis zu 10 Tage

Artikelnummer: 9783031007446
Medium: Buch
ISBN: 978-3-031-00744-6
Verlag: Springer
Erscheinungstermin: 13.05.2019
Sprache(n): Englisch
Auflage: Erscheinungsjahr 2019
Serie: Synthesis Lectures on Data Management
Produktform: Kartoniert
Gewicht: 348 g
Seiten: 161
Format (B x H): 191 x 235 mm

Workflows may be defined as abstractions used to model the coherent flow of activities in the context of an in silico scientific experiment. They are employed in many domains of science such as bioinformatics, astronomy, and engineering. Such workflows usually present a considerable number of activities and activations (i.e., tasks associated with activities) and may need a long time for execution. Due to the continuous need to store and process data efficiently (making them data-intensive workflows), high-performance computing environments allied to parallelization techniques are used to run these workflows. At the beginning of the 2010s, cloud technologies emerged as a promising environment to run scientific workflows. By using clouds, scientists have expanded beyond single parallel computers to hundreds or even thousands of virtual machines.

More recently, Data-Intensive Scalable Computing (DISC) frameworks (e.g., Apache Spark and Hadoop) and environments emerged and are being used to execute data-intensive workflows. DISC environments are composed of processors and disks in large-commodity computing clusters connected using high-speed communications switches and networks. The main advantage of DISC frameworks is that they support and grant efficient in-memory data management for large-scale applications, such as data-intensive workflows. However, the execution of workflows in cloud and DISC environments raise many challenges such as scheduling workflow activities and activations, managing produced data, collecting provenance data, etc.

Several existing approaches deal with the challenges mentioned earlier. This way, there is a real need for understanding how to manage these workflows and various big data platforms that have been developed and introduced. As such, this book can help researchers understand how linking workflow management with Data-Intensive Scalable Computing can help in understanding and analyzing scientific big data.

In this book, we aim to identify and distill the body of work on workflow management in clouds and DISC environments. We start by discussing the basic principles of data-intensive scientific workflows. Next, we present two workflows that are executed in a single site and multi-site clouds taking advantage of provenance. Afterward, we go towards workflow management in DISC environments, and we present, in detail, solutions that enable the optimized execution of the workflow using frameworks such as Apache Spark and its extensions.

Workflows may be defined as abstractions used to model the coherent flow of activities in the context of an in silico scientific experiment. They are employed in many domains of science such as bioinformatics, astronomy, and engineering. Such workflows usually present a considerable number of activities and activations (i.e., tasks associated with activities) and may need a long time for execution. Due to the continuous need to store and process data efficiently (making them data-intensive workflows), high-performance computing environments allied to parallelization techniques are used to run these workflows. At the beginning of the 2010s, cloud technologies emerged as a promising environment to run scientific workflows. By using clouds, scientists have expanded beyond single parallel computers to hundreds or even thousands of virtual machines.More recently, Data-Intensive Scalable Computing (DISC) frameworks (e.g., Apache Spark and Hadoop) and environments emerged and are being used to execute data-intensive workflows. DISC environments are composed of processors and disks in large-commodity computing clusters connected using high-speed communications switches and networks. The main advantage of DISC frameworks is that they support and grant efficient in-memory data management for large-scale applications, such as data-intensive workflows. However, the execution of workflows in cloud and DISC environments raise many challenges such as scheduling workflow activities and activations, managing produced data, collecting provenance data, etc.Several existing approaches deal with the challenges mentioned earlier. This way, there is a real need for understanding how to

Kundenservice Verlag Dr. Otto Schmidt KG
Neumannstraße 10, 40235 Düsseldorf | kundenservice@otto-schmidt.de | 0221-93738-997 (Inland)