Data-Intensive Computing

The world is awash with digital data from social networks, blogs, business, science, and engineering. Data-intensive computing facilitates understanding of complex problems that must process massive amounts of data. Through the development of new classes of software, algorithms, and hardware, data-intensive applications can provide timely and meaningful analytical results in response to exponentially growing data complexity and associated analysis requirements. This emerging area brings many challenges that are different from traditional high-performance computing.

This reference for computing professionals and researchers describes the dimensions of the field, the key challenges, the state of the art, and the characteristics of likely approaches that future data-intensive problems will require. Chapters cover general principles and methods for designing such systems and for managing and analyzing the big data sets of today that live in the cloud, and describe example applications in bioinformatics and cyber-security that illustrate these principles in practice.

Ian Gorton is a Laboratory Fellow in Computational Sciences and Math at Pacific Northwest National Laboratory (PNNL), where he manages the Data Intensive Scientific Computing group and was the Chief Architect for PNNL’s Data Intensive Computing Initiative. Gorton is a Senior Member of the IEEE Computer Society and a Fellow of the Australian Computer Society.

Deborah K. Gracio joined Pacific Northwest National Laboratory in 1990 and is currently the Director of the Computational and Statistical Analytics Division and of the Data Intensive Computing Research Initiative. Since joining the laboratory, she has led the research, development, and management of multiple cross-disciplinary, multi-laboratory projects focused in the basic sciences and national security sectors.
Data-Intensive Computing
Architectures, Algorithms, and Applications

Edited by
IAN GORTON
Pacific Northwest National Laboratory

DEBORAH K. GRACIO
Pacific Northwest National Laboratory
Contents

List of Contributors

1 Data-Intensive Computing: A Challenge for the 21st Century
Ian Gorton and Deborah K. Gracio

2 Anatomy of Data-Intensive Computing Applications
Ian Gorton and Deborah K. Gracio

3 Hardware Architectures for Data-Intensive Computing Problems: A Case Study for String Matching
Antonino Tumeo, Oreste Villa, and Daniel Chavarría-Miranda

4 Data Management Architectures
Terence Critchlow, Ghaleb Abdulla, Jacek Becla, Kerstin Kleese-Van Dam, Sam Lang, and Deborah L. McGuinness

5 Large-Scale Data Management Techniques in Cloud Computing Platforms
Sherif Sakr and Anna Liu

6 Dimension Reduction for Streaming Data
Chandrika Kamath

7 Binary Classification with Support Vector Machines
Patrick Nichols, Bobbie-Jo Webb-Robertson, and Christopher Oehmen

8 Beyond MapReduce: New Requirements for Scalable Data Processing
Bill Howe and Magdalena Balazinska

v
## Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Let the Data Do the Talking: Hypothesis Discovery from</td>
<td>235</td>
</tr>
<tr>
<td></td>
<td>Large-Scale Data Sets in Real Time</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Christopher Oehmen, Scott Dowson, Wes Hatley, Justin</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Almquist, Bobbie-Jo Webb-Robertson, Jason McDermott,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ian Gorton, and Lee Ann McCue</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Data-Intensive Visual Analysis for Cyber-Security</td>
<td>258</td>
</tr>
<tr>
<td></td>
<td>William A. Pike, Daniel M. Best, Douglas V. Love, and</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shawn J. Bohn</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Index</td>
<td>287</td>
</tr>
</tbody>
</table>
List of Contributors

Ghaleb Abdulla Lawrence Livermore National Laboratory
Justin Almquist Pacific Northwest National Laboratory
Magdalena Balazinska University of Washington
Jacek Becla Stanford University
Daniel M. Best Pacific Northwest National Laboratory
Shawn J. Bohn Pacific Northwest National Laboratory
Daniel Chavarría-Miranda Pacific Northwest National Laboratory
Terence Critchlow Pacific Northwest National Laboratory
Scott Dowson Pacific Northwest National Laboratory
Ian Gorton Pacific Northwest National Laboratory
Deborah K. Gracio Pacific Northwest National Laboratory
Wes Hatley Future Point Systems
Bill Howe University of Washington
Chandrika Kamath Lawrence Livermore National Laboratory
Sam Lang Pacific Northwest National Laboratory
Anna Liu National ICT Australia (NICTA), University of New South Wales
Kerstin Kleese-Van Dam Pacific Northwest National Laboratory
Douglas V. Love Pacific Northwest National Laboratory
Lee Ann McCue Pacific Northwest National Laboratory
List of Contributors

Jason McDermott Pacific Northwest National Laboratory
Deborah L. McGuinness Rensselaer Polytechnic Institute
Patrick Nichols Pacific Northwest National Laboratory
Christopher Oehmen Pacific Northwest National Laboratory
William A. Pike Pacific Northwest National Laboratory
Sherif Sakr National ICT Australia (NICTA), University of New South Wales
Antonino Tumeo Pacific Northwest National Laboratory
Oreste Villa Pacific Northwest National Laboratory
Bobbie-Jo Webb-Robertson Pacific Northwest National Laboratory