

Chapter 9

Scalable Data Mining, Archiving, and Big Data Management for the Next Generation Astronomical Telescopes

Chris A. Mattmann

California Institute of Technology, USA

Andrew Hart

California Institute of Technology, USA

Luca Cinquini

California Institute of Technology, USA

Joseph Lazio

California Institute of Technology, USA

Shakeh Khudikyan

California Institute of Technology, USA

Dayton Jones

California Institute of Technology, USA

Robert Preston

California Institute of Technology, USA

Thomas Bennett

SKA South Africa Project, South Africa

Bryan Butler

*National Radio Astronomy Observatory
(NRAO), USA*

David Harland

*National Radio Astronomy Observatory
(NRAO), USA*

Brian Glendenning

*National Radio Astronomy Observatory
(NRAO), USA*

Jeff Kern

*National Radio Astronomy Observatory
(NRAO), USA*

James Robnett

National Radio Astronomy Observatory (NRAO), USA

ABSTRACT

Big data as a paradigm focuses on data volume, velocity, and on the number and complexity of various data formats and metadata, a set of information that describes other data types. This is nowhere better seen than in the development of the software to support next generation astronomical instruments including the MeerKAT/KAT-7 Square Kilometre Array (SKA) precursor in South Africa, in the Low Frequency Array (LOFAR) in Europe, in two instruments led in part by the U.S. National Radio Astronomy Observatory (NRAO) with its Expanded Very Large Array (EVLA) in Socorro, NM, and Atacama Large Millimeter Array (ALMA) in Chile, and in other instruments such as the Large Synoptic Survey Telescope (LSST) to be built in northern Chile. This chapter highlights the big data challenges in constructing data management systems for these astronomical instruments, specifically the challenge of integrating legacy science codes, handling data movement and triage, building flexible science data portals and user interfaces, allowing for flexible technology deployment scenarios, and in automatically and rapidly mitigating the difference in science data formats and metadata models. The authors discuss these challenges and then suggest open source solutions to them based on software from the Apache Software Foundation including Apache Object-Oriented Data Technology (OODT), Tika, and Solr. The authors have leveraged these solutions to effectively and expeditiously build many precursor and operational software systems to handle data from these astronomical instruments and to prepare for the coming data deluge from those not constructed yet. Their solutions are not specific to the astronomical domain and they are already applicable to a number of science domains including Earth, planetary, and biomedicine.

1. INTRODUCTION

The next generation of astronomical telescopes including MeerKAT/KAT-7 in South Africa (Jonas 2009), the Low Frequency Array (LOFAR) in Europe (De Vos, 2009), the Expanded Very Large Array (EVLA) in Socorro, New Mexico (Perley, 2011), the Atacama Large Millimeter Array (ALMA) in Chile (Wootten, 2003) and eventually over the next decade the cross-continental Square Kilometre Array (SKA) (Hall, 2004), and the Large Synoptic Survey Telescope (LSST) in northern Chile (Tyson, 2002) will generate unprecedented volumes of data, stretching from the near terabyte (TB) of data/day range for EVLA on the lower bounds to the 700 TB of data per second range for the SKA. These ground-based instruments will push the boundaries of *Big Data* (Lynch, 2008) (Mattmann, 2013) in several dimensions shown in Table 1. Table 1 represents the common challenges that users, educators, scientists, and other discipline users face when leveraging astronomical data, namely its size (volume, velocity); variety of

formats (complexity); the geographically distributed nature of these telescopes, and the limitations in bandwidth that prevents the wide dissemination of the information throughout the world's users who desire access to it. Big data is the buzzword of the day, used to define data sets so large and complex that traditional data management systems have difficulties handling them. There are three main challenges when dealing with big data: the amount of data collected (volume), the speed at which data must be analyzed (velocity), and the array of different data formats that is collected (complexity).

Engineering the data management, data mining, and archiving systems for these world-wide science assets is a complex (computer) scientific task in its own right, especially considering most of these telescopes are under construction from different funding agencies in the U.S. and abroad, each with different priorities and with different scientific end-user communities. Furthermore, each of the telescopes and their science foci have engendered highly complex data mining chal-

24 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/scalable-data-mining-archiving-and-big-data-management-for-the-next-generation-astronomical-telescopes/85457

Related Content

The Expert's Opinion

Mohammad Dadashzadeh (1991). *Journal of Database Administration* (pp. 38-40).

www.irma-international.org/article/expert-opinion/51097

Conceptual Modeling for XML: A Myth or a Reality

Sriram Mohanand Arijit Sengupta (2009). *Database Technologies: Concepts, Methodologies, Tools, and Applications* (pp. 527-549).

www.irma-international.org/chapter/conceptual-modeling-xml/7930

An Investigation of the Impact of Organization Size on Data Quality Issues

G. Daryl Nord, Jeretta Horn Nordand Hongjiang Xu (2005). *Journal of Database Management* (pp. 58-71).

www.irma-international.org/article/investigation-impact-organization-size-data/3337

Towards an Ontology for Information Systems Development—A Contextual Approach

Mauri Leppänen (2007). *Contemporary Issues in Database Design and Information Systems Development* (pp. 1-36).

www.irma-international.org/chapter/towards-ontology-information-systems-development/7019

Methodology Evaluation Framework for Component-Based System Development

Ajantha Dahanayake, Henk Soland Zoran Stojanovic (2003). *Journal of Database Management* (pp. 1-26).

www.irma-international.org/article/methodology-evaluation-framework-component-based/3288