REPORT FROM THE SATELLITE WORKSHOP ON EXTREMELY LARGE DATABASES FOR SOUTH AMERICA

***Jacek Becla*[*1]*, Bill Howe*[2]**

[*1] *Email: becla@slac.stanford.edu*
[2] *Email: billhowe@cs.washington.edu*

ABSTRACT

Since 2007, Extremely Large Databases (XLDB) events have been organized to host influential discussions on topics relating to extreme-scale databases – holding terabytes through exabytes of data – and to bring together different communities working on Big Data challenges. The XLDB event series includes annual "main" events, typically organized at Stanford University in the USA, and "satellite" events organized outside of North America to reach out to local communities. Past satellite events have been held in Europe and Asia. An XLDB event was organized in June 2014 in Brazil to connect with the South American communities.

Keywords: analytics, database, petascale, exascale, XLDB, big data, extreme-scale

1 ABOUT THE XLDB

Extremely Large Databases (XLDB) events attempt to tackle challenges related to extreme-scale data sets. Main activities include identifying trends, commonalities, and roadblocks related to managing and analyzing extreme scale data sets, as well as facilitating the development and growth of appropriate technologies including (but not limited to) databases. XLDB attendees include a mix of data-intensive industrial and scientific users, as well as big data academic researchers, and vendors.

Since 2007, the XLDB community has met annually each fall at Stanford University in California. In addition, satellite events have been organized on different continents to connect with, and engage local communities working on data-intensive challenges. Such events were held in Edinburgh, UK, in 2011; Beijing,China, in 2012 and in Geneva, Switzerland, in 2013. The 2014 satellite event in Brazil engaged the South America data-intensive community with the XLDB community.

2 ABOUT THE WORKSHOP

The XLDB-South America workshop[1] was held on June 3-4, 2014, at Observatório Nacional in Rio de Janeiro, Brazil. Its main goals were:

- to connect the groups working on extreme-scale data challenges with the XLDB community, and
- to identify areas that need improvements, and draw up recommendations for future actions.

[1] Workshop website: http://xldb-rio2014.linea.gov.br/

The agenda can be found on the workshop website[2]. Information about past XLDB workshops, including the reports, can be found at http://xldb.org/events.

2.1 Participation

Attendance at the XLDB-South America workshop was, like all XLDB workshops, by invitation. This keeps the group small enough for interactive discussions and the represented communities balanced. Eighty-six attended, representing science and industry database users, academic database researchers, and database vendors. The mix mirrored that of the XLDB workshops in the USA. Details about the attendance can be found on the event website[3].

2.2 Structure

The workshop's structure – presentations from different communities, interleaved with panel discussions – is designed to maximize information exchange and stimulate candid, productive discussions.

3 OBSERVATIONS AND RECOMMENDATIONS

The attendees unanimously agreed that the top need for the communities working on extreme-scale data management in Brazil – as well as in South America as a whole – is an appropriate environment for better planning and information exchange. Planning in particular was raised as a big concern. While many groups are working on grand projects and use cutting-edge technologies adopted by other XLDB users, lack of proper planning prevents these groups from seeing “the big picture” and staying focused on those projects and challenges that matter most and are most strategically important for the community and their nations. One example cited multiple times was the opportunity to join the Large Synoptic Survey Telescope (LSST) project. With its construction now beginning, the LSST is perceived as one of the most significant efforts happening in astronomy worldwide. As such, it is extremely important for South American scientists to be part of this effort. While Chile is already a strong partner, Brazil has been undecided. As a result, the entire Brazilian astronomy community may miss the opportunity, if they do not get involved soon.

A significant theme throughout the workshop, addressed explicitly in a panel discussion on the first day, was the need for and potential design of a new eScience Center to facilitate collaboration in the area of data-intensive science and data-driven discovery. Such an institute could serve as a platform for discussing, planning and prioritizing nation's major data management efforts. The existing eScience Institute at the University of Washington is possibly a good model for creating a similar center in Brazil. Detailed information about the structure and activities of this institute is contained in the appendix.

[2] http://xldb-rio2014.linea.gov.br/program/
[3] http://xldb-rio2014.linea.gov.br/participants/

4 NEXT STEPS

Attendees found the XLDB South America workshop to be very informative and useful. They expressed interest in organizing similar XLDB events in South America in the future, either at the same location or elsewhere.

A small group of workshop representatives will follow up with the Brazilian government regarding possibilities of organizing a nationwide eScience center.

ACKNOWLEDGMENTS

The XLDB-SouthAmerica event was sponsored by Huawei and SGI.

The event was organized by a committee consisting of:

- Antonio Divino Moura – Instituto Nacional de Meteorologia
- Fabio Porto – Laboratório Nacional Computação Científica
- Harry Westfahl – Laboratório Nacional de Luz Sincrotron
- Jacek Becla – SLAC / Stanford University
- **Luiz Nicolaci da Costa – Laboratório Interinstitucional de e-Astronomia (chair)**
- Michael Stanton – Rede Nacional de Ensino e Pesquisa
- Sergio Novaes – Universidade Estadual Paulista

with help from the local organizing committee:

- Angelo Fausti Neto – LIneA
- Leandro Ciuffo – Rede Nacional de Ensino e Pesquisa
- Marcio A.G. Maia – Observatório Nacional & LIneA
- Ricardo L.C. Ogando – Observatório Nacional & LineA

The report was edited by Michael Ross.

APPENDIX – THE UW ESCIENCE INSTITUTE AS A POTENTIAL MODEL

A Generational Shift: Motivating a National eScience Center

Rapid advances in technology are transforming nearly every field from “data-poor” to “data-rich” – not only in the sciences, engineering, and medicine, but also in the social sciences. The ability to extract knowledge from this abundance of heterogeneous, noisy, and massive data – “data-intensive discovery” or “data science” – lies at the heart of 21st century discovery.

Data-intensive discovery is now referred to as “The Fourth Paradigm” of scientific inquiry, augmenting observation/experiment, theory, and simulation. Data-intensive discovery relies more on *intellectual infrastructure* – new *methods*, new *tools*, new *partnerships*, and new types of *people* – than on physical infrastructure. Today, most researchers – even the very best – are not well versed in data-intensive discovery methodologies. Until recently, for example, one could be a world-class oceanographer without expertise in data science. No longer! Oceanography, like many other fields, is becoming an *information* field, through rapid advances in chemical, physical, and biological sensors, highly capable teleoperated and autonomous vehicles, and cabled observatories. Similarly, the discovery 50 years ago of a universal genetic code transformed our understanding of biology. But within the past five years, technological advances have radically improved scientists' ability to read, understand, modify, transfer and re-write this code. These new technologies are transforming biology and medicine into information fields, and have huge potential to inform new treatments and even engineer new organisms for a wide variety of applications in both human health and basic research. Neuroscience and neuroengineering are being revolutionized by new sensors and new sensing techniques. Social science researchers can watch relationships evolve in real-time through the Facebook graph, get immediate updates about individuals' opinions, feelings and actions through Twitter, and conduct large-scale surveys automatically online through Amazon's Mechanical Turk rather than relying only on small focus groups.

The UW eScience Institute

The mission of the University of Washington eScience Institute[4] is to be a leader both in developing new data science methodologies and in putting these advances to work in a broad range of fields, creating cross-campus partnerships that drive a "virtuous cycle" in which advances in data science methodologies drive new discoveries, which in turn stimulate the creation of new data science methodologies. The Institute was established with modest core funding from the state in 2008, but in the past year a number of major awards have been received to amplify the effort:

- UW, UC Berkeley, and NYU were chosen from a field of 15 leading universities as partners in a five-year $37.8 million data science initiative funded by the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation.
- UW received a $2.8 million award from the National Science Foundation to create a novel interdisciplinary Ph.D. program in data science.
- The team has been awarded $9.3 million from the Washington Research Foundation for data science efforts.
- Under the Provost's Initiative in Data-Intensive Discovery, a number of "half faculty slots" have been made available to incent the recruiting of top faculty to contribute to the effort.

The programs at the UW eScience Institute are organized into six working groups: Career Paths, Education and Training, Software and Tools, Reproducibility and Open Science, Working Spaces and Culture, and Ethnography and Evaluation.

1) Career Paths for Data Scientists

The Institute aims to produce and advance the careers of "pi-shaped" researchers – researchers with one leg in domain science and one leg in a technical methodology discipline.

The Institute has three researcher roles: *Postdoctoral data science fellows*; *research scientists*, who split their time between domain research and eScience programs, and d*ata scientists*, who perform software development, engineering, and applied research activities across multiple disciplines.

The postdoctoral fellows pursue domain science research agendas, but are hired for their expertise and achievements in both science and in computational methodology – software, data management, statistical methods, visualization. Fellows receive two-year appointments, with a conditional third year depending on progress. Each fellow is dual-mentored by faculty in both a methodology domain (computer science, statistics, or applied math) and a science domain.

[4] http://data.uw.edu/

Data scientist positions are analogous to industry data science positions, emphasizing expertise in data management, statistical analysis, and visualization.

2) Education and training

Data science training is required at all levels – undergraduate students, graduate students, postdocs, research staff, and faculty – and spans basic skills to sophisticated computational thinking. It requires tailoring to meet the needs, languages, and mathematical/computing backgrounds of a variety of disciplines, as well as determining commonalities across those fields. Avoiding siloed approaches that waste resources and discourage interdisciplinary learning and discovery requires new instructional approaches that are flexible yet keep sight of a common core of skills, knowledge and language.

Key activities include

- **A Big Data PhD Track:** The Institute is working towards an interdisciplinary Ph.D. program in Big Data. A first step is offering a specialized Big Data track as part of existing Ph.D. degrees in, initially, six participating departments: Astronomy, Oceanography, Chemical Engineering, Genome Sciences, Computer Science & Engineering, and Statistics.

- **Technical Bootcamps for researchers:** The Institute has run multiple short-format bootcamps to teach programming and data analysis at both introductory and intermediate levels. Advanced bootcamps are in the works. For several of the introductory bootcamps, the Institute has partnered with Software Carpentry[5], a unit in the Mozilla Science Lab[6].

- **Online courses:** The Institute has launched a new massively open online course, called Introduction to Data Science, that attracted more than 100,000 students, with 10,000 completing all assignments in the first offering. Most of the participants online tend to come from industry. A new course being designed is informing curriculum design at the undergraduate and graduate levels.

3) Software and Tools

For data science to flourish, it is crucial to develop an ecosystem of tools and software environments that are organic, sustainable, reusable, extensible, and easy to translate across problem domains. Today’s tools, software environments, and the processes by which these are created are distracting from the science that should be the focus.

Key activities include

- **A Data Science Incubation Program[7]**: The Institute's data scientists collaborate with domain scientists on short-term exploratory software projects. Researchers – students, postdocs, or faculty – submit one-page proposals that are reviewed for science impact, technology impact, interest among the team, and balance across fields. The proposer agrees to physically come to the Institute's incubator two days a week to work alongside the data science team..

[5] http://software-carpentry.org/

[6] https://wiki.mozilla.org/ScienceLab

[7] http://data.uw.edu/incubator/

- **Shared Cyberinfrastructure:** The Institute's SQLShare[8] and Myria[9] projects are examples of shared cyberinfrastructure informed by cross-cutting science requirements. Their designs are informed by common requirements distilled from the incubator projects and long-term collaborations. These efforts also attract external funding from NSF and industry, where Institute requirements overlap with computer science research and industry goals.

4) Supporting reproducibility and open science

The pace of knowledge acquisition in science is impeded by the difficulty researchers have in building upon each other’s work. To establish a community around reproducibility, the Institute organized monthly campus-wide meetings on reproducibility and open science, which regularly attracts dozens of participants from more than eight campus departments. It also conducted a workshop this Spring with more than 100 participants. The goals of this new community are to develop and promote best practices and tools around scientific reproducibility, and advance a culture that values and rewards reproducible research.

5) Working Spaces and Culture

Institute researchers foresee that data and data science will be “the great unifier” in the coming decades, enabling seemingly disparate disciplines to share common problems and solutions. Essential ingredients are physical collaboration spaces, as well as virtual spaces, where researchers from different fields can interact. To this end, the Institute partnered with the UW libraries, the Physics and Astronomy departments, and the College of Arts and Sciences to repurpose an underused library space on campus as a Data Science Studio where researchers across campus can collaborate with Institute staff.

6) Ethnography and evaluation

The Institute has a focused effort to monitor, measure, evaluate and its progress through both quantitative and ethnographic approaches. Institute staff, researchers, and third-party services are dedicated to such self-assessment activities, which are deemed to be critical to the Institute's success.

[8] http://escience.washington.edu/sqlshare
[9] http://myria.cs.washington.edu/