BIG DATA IG
Big Data - Where Do We Stand?

Time: Thursday September 15th, 14:00 - 15:30 - Breakout Session 2
Location: Room Tower Court D, Sheraton Denver Downtown Hotel, Denver, Colorado, US
IG URL: https://rd-alliance.org/groups/big-data-analytics-ig.html
IG Meeting URL: https://rd-alliance.org/ig-big-data-rda-8th-plenary-meeting


Contents

Summary
Minutes

Summary

The IG Big Data session started off with a presentation of the current status of the IG together with some insights from Kwo-Sen Kuo on the common misconceptions, challenges and goals of the IG. This introduction was followed by a presentation of the Big Picture in Big Data from Lesley Wyborn, outlining the state of Big Data across multiple disciplines, with a particular highlight on the absence of communication within the IG. The final presentation by Peter Baumann touched the technical perspective of Big Data and the data cubes as a solution towards the integration of data and meta-data in a unified schema. The session concluded with an extensive Q&A with the participants as well as with the results of a targeted survey regarding the IGs membership makeup.

Minutes
Introduction

Introduction of the panel

Co-Chairs: Kwo-Sen Kuo, Lesley Wyborn, Peter Baumann.
IG RDA Secretary: Line Pouchard
Group Liaison in the Technical Advisory Board in RDA: Francoise Genova

Presentation “IG Status: Where do we stand?”

Kwo-Sen Kuo

Emphasis on the accelerated knowledge extraction from just data production. Simulated data (i.e. data produced from algorithms, simulations etc) is NOT Big Data and as such should not be the emphasis of this group. “Accelerated” means getting “real-time or near real-time” results. Batch processing is not an emphasis, but is should be accommodated.

HPC has been successful in using MPI - but it does not provide the same value for data intensive analysis (shared nothing architecture). This decision leads to optimization considerations at the application and design level.

There are pros and cons for each Big Data technology, depending on the domain. We do not have a clear picture as yet.

In any case, we need to ensure seamless and easy transition from the old way of doing things to the new way. If adoption is slow, it will fail.

Challenges in scientific data-analysis research include collaboration and reproducibility.

Goal of the IG: find the solution with the following characteristics:
1. Maximizing the ratio productivity/cost
2. As real-time as possible.
3. Addressing more than just Big Data challenges

Understanding this IGs membership makeup is an essential first step!


Presentation “IG Big Data: the Big Picture”

Lesley Wyborn

Discussion on how the National Computation Infrastructure on Australia and can integrate with the solutions in RDA.

“Big Data” is a relative term - essentially it means I have more data today than the amount I had yesterday. The definition keeps growing throughout recent history. Looking at the Gartner Hype Cycle for emerging technologies:
In 2014 Big Data was quite prominent in the cycle of emerging technologies. In 2015 it is gone - one of the fastest drop-offs in the Gartner Hype Cycle. One of the Big Data issues currently at the moment is how (e.g.) machine learning is dealing with Data.

Presented a discipline perspective of “Big Data” (see figure below). Our initial scope was the up in the top right corner. However, people in the down bottom left part do have real Big Data issues.
At the communication level, we currently have ~150 members, but there is sporadic communication in the IG. That is a reason we need this survey as well, to see that we are on track.

Finally, it is important to note that combined and integrated, the TB & PB collections are too large to move. So, managing, access and analysis is a clear issue. We need to change our focus to moving users/processing and applications to data.

Presentation “Handling Big Datacubes: RDA Array Database working group?”

Peter Baumann

In Big Data, there is still an unclear definition. These ‘V’s cannot be classed formally. So we can see the following questions: what are the Data Structures and what are the Operations on these Data Structures. And the answer to these questions leads to Data Models. The relational models work fine for meta-data, but for the actual data, not so much.

Discussion on Structural variety in Big Data. At the structure level, we have a very small set of distinct structures: sets + hierarchies + graphs + arrays.

Four basic principles of data structures. Supporting them with canonical methods, we can facilitate a lot of functions.
At the Service Quality level, we can introduce different levels of granularity: we can have images (for example) at surface. Introducing the time, we can have very concrete data cubes.

An example of a possible querying language was given through the introduction of elevation to png image files. The RASDAMAN framework supports several array methods.

There is a standard extension to RDA for multi-dimensional arrays, which will be formally introduced in the near future. Having a common storage space of data and meta-data (datacubes) can bridge the age-old gap of integration.

The Array Database WG is a spinoff of the BigData IG. The goal is to support both technologies and decision makers when deciding Big Data services.

Collaboration plans:
- Standardization (ISO “SQL”)
- Related Activities (EUDAT, EarthServer)

Who should participate: as many as possible! We need use-case providers. We need data and queries. We need service providers (host storage and compute facilities). The outcome will beN public “datacube” services sufficiently large and publicly accessible.
Follow up in the Data Array WG tomorrow. Data cubes is not panacea; however is a very interesting solution to several problems.

Questions and Answers Session

Q: Can you clarify when you say array management systems.
A: Take existing solutions and play with them (that’s one approach). But at some point you will find something missing. We embrace both developer and user side.

Q: How do you preserve data cubes in long term.
A: Interesting questions and currently an open issue. This is something that we need to discuss all together

Comment: Let’s see today’s databases. Data is owned and noone is using and it’s consequently put on tapes. And the data you want quickly are on SSD drives. The maturity of Database Systems can handle different hierarchies in storing data.

Comment: From the logical perspective, one advantage of relations databases is that they are aware of the internal structure (schema).

Q: How datacubes work with web services and computational models:
A: How can we squeeze datacubes through the web. Big Data cannot inherently do that. But code can move to data. In the end you want to extract something focussed - either specific or aggregated. So essentially you want something small. Web services can do that, working on datacubes.

Comment: A database has inherent analytical capabilities. So you can use these facilities to do that.

Q: A typical use case. A scientist has their own tools. How does your language communicating with datacubes work with those.
A: Databases have created their own “english” based query languages (SQL the case is study).So how to cover a database is a way that is most invisible.

Comment: Given the large volumes and the time the tools will require to perform the analytics. So you need a a-synchronous way of communication.

Q: How big a corpus of data would need to be for a valid use case for the WG.
A: At the max limit, It will be constraint by the dataspace we get (providers needed). A lower limit is exceeding the main memory capacities (~10TB). Otherwise it doesn’t make sense to use databases.

Comment: Again we are reducing the Big Data to volume and it’s definitely more complex than that.

Comment: It would be great to see web-services to get the results directly.
Comment: Radar data publicly accessible NMR-NMS. Problems, data and infrastructures need to come together to get interesting results.

Q: It sounds like a good native format for computational output formats rather than actual data.
A: We are working with EU weather data, so it definitely fits that. But we are exploring several options

Comment: Other the climate/earth sciences, we have several different disciplines that can benefit; life sciences, cosmology, financial. The most accommodating we can be, the bigger the impact.

Comment: Definition of Shared Nothing Architecture: Think of a cluster of computers and have a data cube. You divide it into small cubes and you distribute to the individual computers. Each individual computer processes each own data independently. So asking a question, each computer needs to anaylsis and respond only for their own specific subset. It is extremely well suited for embarrassingly parallel problems.

Discussion on the results of the survey (biggest groups earth and climate sciences). Less on the funding level (just 4 funded answers).

Comment: The survey is a very good idea in order to have more focus.

We are looking at the application / domain level. Are there any RDA IG/WG that we should connect with.
- The VRE groups
- The meta-data groups

Comment: We should create more structures ideas within the Interest Group and then start the connection with the Working Groups. It can help mature things and discussions.

Close of session and invitation to participate to the Array WG, tomorrow (Sep 16th, 2016) at 11:00 in Tower Room B.