Earth Science Data Analytics (ESDA) [1]

Submitted by stevek on Fri, 2014-04-04 09:56  Thursday, July 10, 2014 - 14:00 to 15:30

Event: Summer Meeting 2014 [2]
Session Type: Breakout [3]
Data Management Training [5]
Discovery [6]
Information Technology and Interoperability [7]
Products and Services [8]
Student Cluster [9]

Abstract/Agenda:
Earth Science Data Analytics (ESDA) and the Data Scientist - Agenda

1. Review Telecon discussions since January Meeting
2. Discuss Use Cases / Analytics Findings, thus far... and moving forward
4. Discuss candidate cluster deliverables
4. Guest Speaker
5. Discuss next steps...i.e., How do we think it is going? Course corrections, etc.

For reference, Cluster Objectives:

- Provide a forum for ‘Academic’ discussions that allow ESIP members to be better educated and on the same page in understanding the various aspects of Data Analytics
- Bring in guest speakers to describe overviews of external efforts and further teach us about the broader use of Data Analytics.
- Perform activities that:
  --- Compile use cases generated from specific community needs to cross analyze heterogeneous data (could be ESIP members or external)
  --- Compile experience sources on the use of analytics tools, in particular, to satisfy the needs of the above data users (also, could be ESIP members or external)
  --- Examine gaps between needs and expertise
  --- Document the specific data analytics expertise needed in above collaborations
- Seek graduate data analytics/ Data Science student internship opportunities

Notes:
The ESDA Cluster, attracting a lot of interest, continues to ‘churn’ through the process of maturing their understanding and impacts of this new paradigm: Data Analytics and Data Science. Session participants were comprised of technologists and data users, with the majority of people, in attendance to ‘learn’. Thus, in the early stages of this Cluster life, we continue to emphasize learning, which will doubtlessly evolve into applying (shaping) the knowledge we gain into implementable techniques that facilitate the use and advancement of data analytics and data science.

**Our goal is to facilitate making information into knowledge**

Steve started the session by reminding participants of the ESDA Cluster Mission:

- To promote a common understanding of the usefulness of and activities that pertain to Data Analytics and, more broadly, the Data Scientist.

- To facilitate collaborations between organizations that seek new ways to better understand the cross usage of heterogeneous datasets.

- To identify gaps that, once filled, will expand collaborative activities.

And Objectives:

- To provide a forum for ‘Academic’ discussions
- Host guest speakers to provide overviews of external efforts
- Perform activities that:
  - Compile specific community use cases (analytics needs) to cross analyze heterogeneous data
  - Compile experienced sources on the use of analytics tools to satisfy the needs of the above data users
  - Examine gaps between needs and expertise
  - Document specific data analytics expertise needed
  - Seek graduate data analytics/ Data Science student internship opportunities

Next, Steve presented relevant AGU Earth Science Data Analytics Sessions, encouraging abstract submission to any of these sessions:
- Teaching Science Data Analytics Skills Needed to Facilitate Heterogeneous Data/Information Research: The Future Is Here - Session ID#: 1879

- Identifying and Better Understanding Data Science Activities, Experiences, Challenges, and Gaps Areas - Session ID#: 1809

- Advancing Analytics using Big Data Climate Information System - Session ID#: 3022

- Big Data in the Geosciences: New Analytics Methods and Parallel Algorithm - Session ID#: 3292

- Leveraging Enabling Technologies and Architectures to enable Data Intensive Science - Session ID#: 3041

- Open source solutions for analyzing big earth observation data - Session ID#: 3080

- Technology Trends for Big Science Data Management - Session ID#: 2525

And presented the clusters accomplishments and telecon guest speakers over the past 6 months.

**Peter Fox - Guest Speaker**

Main points (presentation attached below) and other key ideas:

- Peter spoke on Data Science and Analytics Curriculum development at Rensselaer

- Much of what is being done is relabeling and repacking- not much is being done new with new data science and analytics. Peter is working with Xinformatics: data moved to information (Data Science, Semantic eScience, Data Frameworks)

- Peter has developed GIS4Science and Data Analytics courses at RPI- there is no separate degree program in data science or informatics science – the courses are embedded in other programs (bioinformatics, physics, etc.)

- The Power in Analytics is Predictive and Prescriptive – in big data knowledge of nonparametric aspect is critical

- Students should be solving real problems with real science from the start, data science must be a skill – there is a key element of team work since data science is mainly done in groups.
- Recently published an article in the big data journal

- There are only 2 papers written on the theory of Data (one is from 1963) this makes it difficult to teach data and even more difficult for students to understand.

- It is important to distinguish between analytics and analysis – “Analysis is looking in (at the data) and Analytics is looking beyond (the data)”

- In discussing the scope of data analytics, from, for example, data discovery to science discovery, we need to be clear of the scope in which data analytics addresses. Thus we need to bound the problem being addressed. For example, science discovery utilizing data analytics discovery method assumes that the data to be studied has already been discovered (i.e., data discovery)

Peter’s 5-6 years in...

• Science and interdisciplinary from the start!
  - Not a question of: do we train scientists to be technical/data people, or do we train technical people to learn the science
    - It’s a skill/ course level approach that is needed
    - We teach methodology and principles over technology *
    - Data science must be a skill, and natural like using instruments, writing/using codes
    - Team/ collaboration aspects are key **
    - Foundations and theory must be taught ***

• Multi-disciplinary science program - PhD in Data and Web Science
• DATUM: Data in Undergraduate Math! (Bennett)
• Missing – intermediate statistics

(Presentation attached below)

Let’s start working toward ‘5-6 years in...’
What was planned for this session:

The goal of this session was to analyze the various use cases and data analytics tools/techniques thus far compiled by the cluster, by responding to the following questions:

For each use case:

1. What specifically is to be done?
2. Which analytics types is the use case attempting?
3. What classes of users is represented by this use case?

And or each tool:

1. What specifically does the tool provide?
2. Which analytics types does the tool address?
3. What classes of users would best benefit from use of this tool?

Since, as we all know in this business, one size does not fit all, it would be correct to address Data Analytics in terms of the currently defined types. Thus, we should focus on one Data Analytics type at a time, and map to these Data Analytics types, the use cases, tools/techniques, and classes of users as they apply. Hopefully, we will be able to categorize the information we are gathering (and continue to gather), and make it easier to gather more specific use cases, as well as describe the big Data Analytics picture. So, we need to:

- Map use cases to particular types of analytics
- Map analytics tools/techniques to types of data analytics
- Map types of data analytics to classes of users (user model)

Types of Data Analytics

**Descriptive Analytics:** You can quickly understand "what happened" during a given period in the past and verify if a campaign was successful or not based on simple parameters.

**Diagnostic Analytics:** If you want to go deeper into the data you have collected from users in order to understand "Why some things happened," you can use ... intelligence tools to get some insights.

**Discovery Analytics:** The use of data and analysis tools/models to discover information
**Predictive Analytics:** If you can collect contextual data and correlate it with other user behavior datasets, as well as expand user data … you enter a whole new area where you can get real insights.

**Prescriptive Analytics:** Once you get to the point where you can consistently analyze your data to predict what's going to happen, you are very close to being able to understand what you should do in order to maximize good outcomes and also prevent potentially bad outcomes. This is on the edge of innovation today, but it's attainable!

Illustrated in attached file: 'ESDA page images', below

---

**Subset of Users from the ESDSWG User Needs WG User Model (most likely to utilize Data Analytics)**

Table in attached file: 'ESDA page images', below

---

**What actually was discussed in this session:**

Overall, most participants were attending to learn about Data Analytics. Informatics and physical science users were also present. Thus, the discussion evolved away from the above plans, and to drawing out a better understanding of what Data Analytics is and means. Nevertheless, the conversation was very interesting, and hopefully brought us further down the road to becoming more familiar and comfortable with the data analytics paradigm and lingo. Discussion highlights:

**Data Analytics Types:**

- In other words:
  - hindsight - descriptive
  - insight - discovery
  - foresight - predictive

- How does Discovery Analytics differ from Data Discovery? When discussing Data Analytics we must bound the problem. Thus, Data Discovery uses tools and services to locate data for further use (science, applications, etc.) Discovery Analytics draws upon tools/techniques to discover unknown information/knowledge (i.e., multi-dataset analysis). The assumption is the desired data is already acquired.
Use Cases:

- The comment was made that our use cases need flushing out to be more useful. This is a good suggestion and one to follow-up on.

- Suggestion: Set pre-conditions and triggers - otherwise we're going to be pursuing unbounded case studies.

- Suggested Use Cases:
  - The model verification use case: NSIDC permafrost modeler needs to be able to access temp profiles over a long period in the permafrost (Some are accessibly and some are not). A lot of the data lacks consistency.
  - Leverage analytics to be able to identify images and select what models would best help extract the appropriate data
  - Use analytics to identify what other datasets can be used to help characterize and remove background noise (process and change using both sets of data in different processing)
  - Peter spends about 80% of his time munging/cleaning the data, getting them in a format that their tools can use- if there were analytics tools to help clean up the data so more time can be spent working with the results than getting the results.
  - Different take…using large heterogeneous datasets, but also the work involved prior to using the large heterogeneous datasets. One user would like to use large heterogeneous datasets but faces challenges: So much data available; Difficulty in knowing what data and services opportunities exist; Understanding the data (either too much documentation or not enough)

Data Analytics Comments and Thoughts:

- Many people have tried large data synthesis once but not more than once – tools to help facilitate this (that can deal with a lot of different data, e.g. unit conversions) Also with heterogeneous data sets, when one instrument does not collect things the way they said they did this is problematic – tools that could help with this would be very helpful.

Actions:
1. Flush out use cases: Bound the issue, be specific. Seek additional individuals who are facing issues utilizing large heterogeneous datasets
2. Further define Data Analytics types: Per type: Issues, Potential solutions, exemplary situations, user classes, other
3. Initiate some of the above planned mapping
4. In December, have 2 ESDA sessions. One can be entitled: ‘Earth Science Data Analytics 101’
Earth Science Data Analytics (ESDA)
Published on Commons (http://commons.esipfed.org)

Attachments/Presentations:
- Fox_DataScienceAnalytics20140710.ppt [12]
- ESDA page images.pptx [13]

Session Leads:
Name: Steve Kempler [14]
Organization(s): Goddard DAAC [15]

Name: Tiffany Mathews [16]
Organization(s): Booz Allen Hamilton [17]
Email: tiffany.j.mathews@nasa.gov [18]

Presenters:
Name: Peter Fox [19]
Organization(s): Rensselaer Polytechnic Institute [20]

Name: Steve Kempler [14]
Organization(s): Goddard DAAC [15]

Participants:
Tom, Narock, Liping Di, David Gallahe, Yuan Ho, Jenny Fox, Ward Fleri, Nic Weber, Eric Kiln, Ruth Duerr, Heather McCullough, Stephan Klene, Sara Graves, Peter Fox, John Porter, Doug Lindholm, Josh Young, Ed Johnson, Thomas Harris, Karl Benedict

Creative Common License: Creative Commons Attribution 3.0 License
Teaser: Data Science: “the sexiest job of the 21st century” yeah
Accepted:
Keywords: analytics data-scientist big-data techniques/methodologies [21]

Source URL: http://commons.esipfed.org/node/2352

Links:
[16] http://commons.esipfed.org/node/1555
[18] mailto:tiffany.j.mathews@nasa.gov