The Need for Earth Science Data Analytics to Facilitate Community Resilience (and other applications) [1]

Submitted by stevek on Mon, 2015-04-27 12:39  Thursday, July 16, 2015 - 10:30 to 12:00
Event: Summer Meeting 2015 [2]
Session Type: Breakout [3]
Collaboration Area: Climate Education Working Group [5]
Energy and Climate [6]
Information Technology and Interoperability [7]

Abstract/Agenda:
Data Analytics is the process of examining large amounts of data of a variety of types to uncover hidden patterns, unknown correlations and other useful information. In an information science/technology field previously unexplored, the ESIP Earth Science Data Analytics (ESDA) Cluster has made progress in understanding and focusing data analytics pertaining to Earth science, defining the Use Case template, and interpreting ESDA types, based on Use Cases, and not on known data analytics types better suited for the business world.

This cluster session will review our current work (for new participants), followed by discussion on the extent of social, economic, and environmental issues, as well as science research, in which the advancement of Earth science data analytics have had an impact. The goal of this discussion is to gain sufficient information to categorize how Earth science data analytics has come to be used in our society, and identify use cases that exemplify this. Discussions of tools and techniques that yield solutions would also be nice.

Notes:

The focus of this session:

- Review our current work (for new participants)
- Discuss and finalize Earth Science Data Analytics definition (published definitions targeting the Business world do not exactly fit for Earth science)
- Discuss and finalize Earth science Data Analytics types (published types targeting the Business world do not exactly fit for Earth science)
- Discuss/Collect Use Cases pertaining to the utilization of Earth science data (analytics) in addressing social, economic, and environmental issues

In better understanding Earth Science Data Analytics (ESDA), we have spent much effort defining ESDA, and better understanding the different types of data analytics in terms of how the analytics are used. This will ultimately enable us to better identify/define Earth science data analytics tools and techniques that can directly serve Earth science data research and analysis.

In this session, 18 ESIP members participated in this session possessing expertise/interests in Science (~3 participants), Data Management (~5), Engineering (~3), Data/Information Science (~5), being a future Data Scientist (~2).

Introduction

As an introduction for new ESDA Cluster members, the following highlights were presented:
ESDA Cluster Goal: To understand where, when, and how ESDA is used in science and applications research through speakers and use cases, and determine what Federation Partners can do to further advance technical solutions that address ESDA needs. Then do it.

Ultimate Goal:

To Glean Knowledge about Earth from All Available Data and Information

ESDA Cluster – What we have done

- 14 Telecons
- 6 face-to-face sessions
- 16 ‘guest’ presentations
- Created an ESDA specific use case template
- Gathered 18 use cases
- Settled/Focused on Data Analytics definition
- Refocused on Earth science data analytics definition *
- Settled/Focused on 5 Data Analytics types
- Refocused on 11 Earth science data analytics types *
- Acquiring Use Case *
- Describe/Demonstrate UV CDAT and ClimatePipes visualization analytics tools

* - Subjects of today’s discussion

Definition of Earth science data analytics

Goal: A definition we want to stamp ‘ESIP’ on?

Earth Science Data Analytics Definition:

- The process of examining large amounts of data of a variety of types to uncover hidden patterns, unknown correlations and other useful information, involving one or more of the following:
  - Data Preparation – Preparing heterogeneous data so that they can ‘play’ together
  - Data Reduction – Smartly removing data that do not fit research criteria
  - Data Analysis – Applying techniques/methods to derive results

Discussion:

- Let’s not worry about the word “Big” what is big now will not be “big” later

- What a domain scientist does analysis within their specific field, whereas data preparation and reduction are more tools to expedite these processes.

  - However, science analysis can include these (preparation and reduction) – Data preparation and Reduction analytics can include the “pre-science” work

  - (We need to be careful in what we call “science” though there is preparation work that scientists do, it is not the actual science)

- Suggested: decouple the data analysis part with the end goal, where we decouple preparation, reduction, and analysis, why not go further in decoupling.

Question then came up regarding the goals of performing data analytics, which actually addressed the second part of today’s session. Thus with the definition of ESDA in mind, the different types of
ESDA were presented, with the intention of returning to discussion of the ESDS definition

**Definition of Earth science data analytics types**

Why is it important to identify Data Analytics Types?

*To better identify key needs that tools/techniques can be developed to address.*

Basically, once we can categorize different types of Data Analytics, we can better associate existing and future Data Analytics tools and techniques that will help solve particular problems.

The most documented set of Types of Data Analytics include: Descriptive, Diagnostic, Discoveritive, Predictive, and Prescriptive (please see session presentation for their descriptions). After unsuccessfully attempting to classify, by type, the 18 gathered (so far) use cases, it became evident that the above data analytics types are not applicable for ESDA. This was also questioned at the January, 2015 Cluster Meeting, where we concluded that ESDA types should be goal oriented.

It was next realized that, from our Use Case Template (where we describe our use cases), the item: ‘Use Case Goals’, can in fact, specify the ESDA types. Coming into the session, the proposed ESDA types include:

1. To calibrate data
2. To validate data (quality) (note it does not have to be via data intercomparison)
3. To perform coarse data reduction (e.g., subsetting, data mining)
4. To intercompare data (i.e., any data intercomparison; Could be used to better define validation/quality)
5. To derive new data product
6. To tease out information from data
7. To glean knowledge from data and information
8. To forecast/predict phenomena (i.e., Special kind of conclusion)
9. To derive conclusions (i.e., that do not easily fall into another type)
10. To derive analytics tools
11. To recover/rescue data

**Goal:** *A set of ESDA types we want to stamp ‘ESIP’ on?*

The rest of the session was used to discuss if this is the correct approach, and if this is the correct list. Discussion included:

- Updating list to fix overlapping and missing ESDA types from the above list
- Qualify types with descriptions that clearly specify their differences; Use examples
- Suggestion made to map ‘ISO/DIS 19119 Geographic information – Services’ to ESDA types
- Observations: There appears to be a pattern that there is not, as typical, a research question to start with. It appears that scientists query the data then come up with a hypothesis based on the data not data based on the hypothesis.
- Additional observation: Seeking new signatures - hypothesis come from measurements. It is still an experiment but the data is from past occurrences
- Do these observations address a type of ESDA that is not included in our list?

- Logistics: An outcome of this activity should be a white paper that is presented to the ESIP community, describing cluster work leading to the definition of ESDA and ESDA types, for ESIP endorsement.

**Session ESDA Types Results**

As a result of today’s cluster session, the following list of ESDA types were derived, provided for additional refining, as necessary

1. To calibrate data
2. To validate data (note it does not have to be via data intercomparison)
3. To assess data quality
4. To perform course data preparation (e.g., subsetting, data mining, transformations, recover data)
5. To intercompare data (i.e., any data intercomparison; Could be used to better define validation/quality)
6. To tease out information from data
7. To glean knowledge from data and information
8. To forecast/predict phenomena (i.e., Special kind of conclusion)
9. To derive conclusions (i.e., that do not easily fall into another type)
10. To derive new analytics tools

**Commercial Break (going to AGU?):**

- **IN004. Advanced Information Systems to Support Climate Projection Data Analysis** - Gerald L Potter, Tsengdar J Lee, Dean Norman Williams, and Chris A Mattmann
- **IN009. Big Data Analytics for Scientific Data** - Emily Law, Michael M Little, Daniel J Crichton, and Padma A Yanamandra-Fisher
- **IN010. Big Data in Earth Science - From Hype to Reality** - Kwo-Sen Kuo, Rahul Ramachandran, Ben James Kingston Evans. and Mike M Little
- **IN012. Computing Big Earth Data** - Michael M Little, Darren L. Smith, Piyush Mehrotra, and Daniel Duffy
- **IN023. Geophysical Science Data Analytics Use Case Scenarios** - Steven J Kempler, Robert R Downs, Tiffany Joi Mathews, and John S Hughes
- **IN034. New approaches for designing Big Data databases** - David W Gallaher and Glenn Grant
- **IN039. Partnerships and Big Data Facilities in a Big Data World** - Kenneth S Casey and Danie Kinkade
- **IN049. Towards a Career in Data Science: Pathways and Perspectives** - Karen I Stocks, Lesley A Wyborn, Ruth Duerr, and Lynn Yarmey

**Notes:**

Link to download WebEx session
The Need for Earth Science Data Analytics to Facilitate Community Resilience (and other applications)

Published on Commons (http://commons.esipfed.org)

recording here: https://drive.google.com/open?id=0B8HxD_REKRt2ZUtXay1VUHRaZ00 [8]

Link to download WebEx recording player here: https://drive.google.com/open?id=0B8HxD_REKRt2RGc1VUtRmppeW8 [9]

**Actions:**
Cluster (Steve lead) – Provide more detailed descriptions of ESDA types

Cluster (all) – Based on ESDA data type descriptions, review and further refine, as necessary, ESDA types

Cluster (all) – Review and refine ESDA definition

Cluster (all) - Contribute Use Cases

**Attachments/Presentations:**
[ESDA Sessions 0715 just Thursday.pptx](#) [10]

**Session Leads:**
Name: Steve Kempler [11]
Organization(s): Goddard DAAC [12]

Name: Tiffany Mathews [13]
Organization(s): Booz Allen Hamilton [14]
Email: tiffany.j.mathews@nasa.gov [15]

**Presenters:**
Name: Steve Kempler [11]
Organization(s): Goddard DAAC [12]

**Notes takers:**
Name: Tiffany Mathews [13]
Organization(s): Booz Allen Hamilton [14]
Email: tiffany.j.mathews@nasa.gov [15]

Name: Steve Kempler [11]
Organization(s): Goddard DAAC [12]

**Participants:**
Sean Barberie, Liping Di, Meixia Deng, Hampapuram Ramapriyan, Pam Mlynczak, Charles Thompson, Toni Rosati, Donna Scott, Amanda Leon, Aleksandar Jelenak, Ward Fleri, Tiffany Mathews, Nikunj Oza, Ellen Johnson, Brian Wilson, Walt Baskin, Don Middleton, David Arctur

**Creative Common License:** Creative Commons Attribution 3.0 License

**Teaser:** Have a data analysis or application problem you are trying to solve?

**Accepted:**

**Keywords:** Data Analytics [16]
Use Cases [17]
Research [18]
Applications [19]

**Source URL:** http://commons.esipfed.org/node/7998

**Links:**
The Need for Earth Science Data Analytics to Facilitate Community Resilience (and other applications)

Page 6 of 6