# Earth Sciences Data Analytics (ESDA) Cluster – Work in Progress

**Mission:**

To promote a common understanding of the usefulness of, and activities that pertain to, Data Analytics and more broadly, the Data Scientist; facilitate collaborations to better understand the cross usage of heterogeneous datasets and to provide accommodating data analytics expertise, now and as the needs evolve into the future; identify gaps that, once filled, will further collaborative activities.

**Objectives**

- Provide a forum for “Academic” discussions that provides ESDA members a better understanding of the various aspects of Earth Science Data Analytics
- Bring in outside speakers to describe external efforts, and further teach us about the broader use of Data Analytics.
- Perform activities that:
  - Collect use cases generated from specific community needs to cross analyze heterogeneous data
  - Collect sources of analytics tools, in particular, to satisfy the needs of the above data users
  - Examine gaps between needs and sources
  - Examine gaps between needs and community expertise
- Document specific data analytics expertise needed to perform Earth science data analytics
- Seek graduate data analytics Data Science student internship opportunities

**Agenda Highlights**

- Analytics and Data Scientist...in the Federation
- Other Activity Briefings: RDA, NIST
- Compiling use cases, analytics tools (internal and external to ESDA)
- Various guest speakers
- Cluster Information Sharing Website
- Describe/Demonstrate UV CDAT and ClimatePipes visualization analytics tools
- Use Case Information Needed Template
- Defining, describing, and applying 5 Data Analytics Types
- Acquiring Use Case
- Planning Summer/2015 ESDA Sessions:
  - Yesterday, in case you missed it: Teaching Science Data Analytics Skills, and the Earth Science Data Scientist
  - Tomorrow, 10:30, don’t miss it: The Need for Earth Science Data Analytics to Facilitate Community Resilience (and other applications)

**Presentations**

- Wei-Chang: NIST Big Data Public Working Group & Standardization Activities - 2/20/14
- Brand Nierman: Sizing out Data Science and Data Analytics - 2/20/14
- Mark Schenna: MERRA Analytics Services - 3/24/14
- Samuel Mobasher: Data Analytics Masters Program at DePaul University - 3/24/14
- Jan Ares: Data Analytics Needs Scenario - 4/17/14
- Rash Dicker: User-Oriented Data Analytics and Tools using the Federated Data System DataFed - 4/17/14
- Tiffany Mathews: Atmospheric Science Data Center Sample Analytics Use Cases - 4/17/14
- Steve Kempler: Analytics and Data Scientists, Earth Science Data Analytics 101 - 5/17/14
- Gene Bovin: From Maps: One (or creating one great precipitation dataset set from many good ones) - 5/17/15
- David Guilder: Reconstructing Sea Ice Extent from Early Nimbus Satellites - 7/15/15
- Thomas Hearty: Sampling Total Precipitable Water Vapor using AERIS and MERRA - 7/15/15
- Radia Sobolay: Using Earth Observations to Understand and Predict Infectious Diseases - 7/15/15
- Tiffany Mathews: Promising data analytics technologies - 7/15/15

**Other References**

- Education for Data Scientists
- Data Analytics (as required) Data Analytics course
- Data Science (as exemplary Data Science course)
- Introduction to Data Science (as exemplary Data Science course)
- RDA Big Data Analytics Interest Group Charter
- NIST Big Data Program
- Schenna: MERRA Analytics Services paper
- Ralph Kahn, “Why we need huge datasets of Earth observations...”

---

**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

---

**Use Case Template**

- Use Case Title
- Author/Company/Email
- Actors/Stakeholders/Project URL and their roles and responsibilities
- Use Case Goal → Earth Science Data Analytics TYPES! (see below)
- Use Case Description
- Current technical considerations to take into account that may impact needed data analytics
- Data Analytics tools applied
- Data Analytics Challenges (Gaps)
- Type of User
- Research Areas
- Societal Benefit Areas
- Potential for and/or issues for generalizing this use case (e.g. for ref. architecture)
- More information and relevant URLs (e.g. who to contact or where to go for more information)

**Analytics Tools/Techniques Examined (to mention a few)**

- Dryad, MapReduce, Hadoop, OpenMP, PowerNet, PowerLaw, True Knowledge, WolframAlpha, myGrid, UV CDAT, ClimatePipes, MIBIC II, CrazyEgg/Heat Maps

---

**Types of Earth Science Data Analytics**

<table>
<thead>
<tr>
<th>Types of Earth Science Data Analytics</th>
<th>Use Cases</th>
<th>Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Analytics (as required) Data Analytics course</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Science (as exemplary Data Science course)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Introduction to Data Science (as exemplary Data Science course)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RDA Big Data Analytics Interest Group Charter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIST Big Data Program</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schenna: MERRA Analytics Services paper</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ralph Kahn, “Why we need huge datasets of Earth observations...”</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Use Case Goals:**

- To calibrate data
- To validate data (quality; note it does not have to be via data intercomparisons)
- To perform course data reduction (e.g., subsetting, data mining)
- To intercompare data (i.e., any data intercomparison; Could be used to better define validation/quality)
- To derive new data product
- To intercompare data from different sources
- To glean knowledge from data and information
- To incorporate phenomena (i.e., Special kind of correlation)
- To derive conclusions (i.e., that do not easily fall into another type)
- To derive analytics tools
- To discover new data

**Use Case Conclusion:**

For Earth Science, deriving results oriented Data Analytics types are more appropriate for categorizing Earth science data analytics...

- They accommodate Earth science use cases which are typically results oriented
- They invite better defined data analytics tools and techniques that address user goals

- Most Earth science data analytics use cases tend to focus on data intercomparison, deriving new products, forecasting/predicting, and deriving correlations
- No use cases were identified to glean knowledge from data information. Perhaps some use cases were not recognized as such
- Distributed data sources, and data heterogeneity are persistent characteristics...
- Velocity issues are not
- Earth science data analytics challenges provide interesting problems for data analytics tool/technique developers to ponder

- If any, use case 5.16 provides the true Big Data problem

---

**ESIP ESDA Cluster Members, Prepared by Steven Kempler**

**Steven J. Kempler, nasa.gov**

---

**Data Case Study:**

<table>
<thead>
<tr>
<th>Data Case Study</th>
<th>Earth Science Data Analytics</th>
<th>Use Case Goals</th>
<th>Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earth Monitoring</td>
<td>Earth System Science</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSFC</td>
<td>Earth System Science</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NASA</td>
<td>Earth System Science</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOAA</td>
<td>Earth System Science</td>
<td></td>
<td></td>
</tr>
<tr>
<td>USGS</td>
<td>Earth System Science</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nasa</td>
<td>Earth System Science</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESIP</td>
<td>Earth System Science</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Analytics Tools/Techniques Examined (to mention a few)**

- Dryad, MapReduce, Hadoop, OpenMP, PowerNet, PowerLaw, True Knowledge, WolframAlpha, myGrid, UV CDAT, ClimatePipes, MIBIC II, CrazyEgg/Heat Maps

---

**ESIP Earth Sciences Data Analytics (ESDA) Cluster – Work in Progress**