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

The ESIP ESDA Cluster Members, Prepared by Steven Kempler
Steven.J.Kempler@nasa.gov

http://wiki.esipfed.org/index.php/Earth_Science_Data_Analytics

Mission:

To promote a common understanding of the usefulness of, and activities that pertain to, Data Analytics and more broadly, the Data Scientist. Thus will be done through:

- Facilitation of collaborations to better understand the cross usage of heterogeneous datasets
- Accommodation of data analytics expertise, current and future needs
- Identification of gaps that, once filled, will further collaborative activities.

Objectives

- Provide a forum for 'Academic' discussions that provides ESIP members a better understanding of the various aspects of Earth Science Data Analytics
- Bring in guest speakers to describe 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
  - Compile 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 ESIP)
- 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

- Wo Chang: **NIST Big Data Public Working Group & Standardization Activities** - 2/20/14
- Brand Niemann: **Sorting out Data Science and Data Analytics** - 3/20/14
- John' Schnase: **MERRA Analytic Services (MERRA/AS)** - 3/20/14
- Bamshad Mobasher: **Data Analytics Masters Program at DePaul University Overview** - 3/20/14
- Joan Aron: **Data Analytics Needs Scenario** - 4/17/14
- Rudy Husar: **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** - 1/7/15]
- Dave Bolvin: **From Many, One (or creating one great precipitation data set from many good ones)** - 1/7/15
- David Gallaher: **Reconstructing Sea Ice Extent from Early Nimbus Satellites** - 1/7/15
- Thomas Hearty: **Sampling Total Precipitable Water Vapor using AIRS and MERRA** - 1/7/15
- Radina Soebiyanto: **Using Earth Observations to Understand and Predict Infectious Diseases**- 1/7/15
- Tiffany Mathews**: Promising data analytics technologies** - 1/7/15

Other References

- Education for Data Scientists
- Data Analytics (an exemplary Data Analytics course)
- Data Science (an exemplary Data Science course)
- Introduction to Data Science (an exemplary on-line course)
- RDA Big Data Analytics Interest Group Charter
- NIST Big Data Program
- Schnase: MERRA Analytic Services paper
- Ralph Kahn, "Why we need huge datasets of Earth observations…"

Events and Activities [edit]

2015-06-18: Fourteenth Telecon
2015-05-21: Thirteenth Telecon
2015-04-16: Twelfth Telecon
2015-03-19: Eleventh Telecon
2015-02-26: Tenth Telecon
2015-02-05: Ninth Telecon
2015-01-07: January, 2015 ESIP Meeting notes (Washington), ESDA 201 Session
2015-01-07: January, 2015 ESIP Meeting notes (Washington), ESDA 101 Session
2014-11-20: Eighth Telecon
2014-10-23: Seventh Telecon
2014-08-21: Sixth Telecon
2014-07-10: July, 2014 ESIP Meeting notes (Frisco)
2014-06-26: Fifth Telecon
2014-05-22: Fourth Telecon
2014-04-17: Third Telecon
2014-03-20: Second Telecon
2014-02-20: First Telecon
2014-01-09: Initial ESIP Meeting notes

Archive

Active Collaborations [edit]

Gathering Use Cases...

Gathering Analytics Tools/Techniques...

Use Case Information Needed Working Spreadsheet...

Resources [edit]

Presentations

Other References

Get Involved [edit]

- **Earth Science Data Analytics Discussion Forum**
- **Email List:** ESIP-ESDA
- **Telecons:**
  - Third Thursday of each month (3 - 4 p.m. EST)
  - Next, after Summer ESIP Meeting: August 20, 2015, 3-4 EST
  - WebEx: https://esipfed.webex.com/ , 23136782
  - Telecon: 1-877-668-4493, 23136782#
- **Cluster Contacts:** Steve Kempler, Tiffany Mathews

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, OpenCyc, Powerset, True Knowledge, WolframAlpha, myGrid, UV-CDAT, ClimatePipes, MIIC II, CtrazyEgg/Heat Maps

Types of Earth Science Data Analytics

1. To calibrate data
2. To validate data (quality) (note it does not have to be via data intercomparison)
3. To perform course 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

Current Conclusions

- For Earth Science, defining 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 conclusions
- 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

As use cases are added and updated conclusions are expected to change

	Types of Earth Science Data analytics											Other Significant Earth Science Data Analytics Considerations							Current data analytics tools applied	Data Analytics Challenges
Use Cases	1	2	3	4	5	6	7	8	9	10	11	Data sources	Volume	Velocity	Variety	Veracity	Visualization	Specialized s/w		
1 MERRA Analytics Services: Climate Analytics-as-a-Service										√		Distributed					For Mapping		Cloudera MapReduce	
2 MUSTANG QA: Ability to detect seismic instrumentation problems		√	√					√				Centralized	100's TB --> PB		Uniform	Problematic		scheduler, SQL	R, Matlab, Excel, PQLX	Large ds; erroneous data
3 Inter-calibrations among datasets	√	√		√																MIICII, XML
4 Inter-comparisons between multiple model or data products				√								Centralized	Huge		Heterogeneous		To Identify event			
5 Sampling Total Precipitable Water Vapor using AIRS and MERRA		√		√								Co-located			Heterogeneous		To detect differences		Sampling, Gridding	
6 Using Earth Observations to Understand and Predict Infectious Diseases								√	√			Distributed	Large		Heterogeneous		Data exploration, findings	db, math/stat modeling	Regression Modeling; Machine Training; Neural Network; R	Data heterogeneity; data/results validation
7 CREATE-IP - Collaborative Reanalysis Technical Environment - Intercomparison Project				√								Distributed	up to 1 PB		Different formats	Depends on input	WMS, UV-CAT, ArcGIS		Anomaly correction	Volume; Data heterogeneity
8 The GSSTF Project (MEaSUREs-2006)					√							Distributed			Heterogeneous	Depends on input				Large data inputs/outputs
9 Science- and Event-based Advanced Data Service Framework at GES DISC					√					√		Distributed			Diverse data					
10 Risk analysis for environmental issues								√				Distributed			Diverse data					Determine model output suitability
11 Aerosol Characterization				√					√			Distributed	Huge		Heterogeneous	Part of analysis	Customized	Developed as needed		Reliable pattern recognition
12 Creating One Great Precipitation Data Set From Many Good Ones					√							Distributed		Near real time	Diverse data	Can be a problem		Intercomparison; morphing	Kalman filtering technique	Intercalibrate datasets to produce best data
13 Reconstructing Sea Ice Extent from Early Nimbus Satellites	√										√	Single source	Large # of records			Very problematic				Unreadable tapes = not automated
14 DOE-BER AmeriFlux and FLUXNET Networks *						√			√			Distributed			Diverse data		Graphs and 3D surfaces	EddyPro, python, Matlab, neural networks	Data mining, interpolation, fusion, R	Translation across diverse datasets
15 DOE-BER Subsurface Biogeochemistry Scientific Focus Area *								√				Distributed			Diverse data	Very problematic	To understand data	PFLOtran, postgres, NEWT	Data mining, interpolation, fusion	Translation across diverse datasets
16 Climate Studies using the Community Earth System Model at DOE's NERSC center *								√	√	√		Distributed	up to 30 PB	42 GBytes/sec	Diverse data		To understand data	PIO, NCL, NCO, parallel NetCDF	Data reduction; analysis near archive	A true Big Data problem
17 Radar Data Analysis for CReSIS *						√						Single source	~0.5 PB per year			Needs analysis		Matlab, MapReduce, MPI, GIS	Signal/Image processing	Immature image processing algorithms
18 UAVSAR Data Processing, Data Product Delivery, and Data Service *					√	√						Centralized			2 main types		GIS	ROI_PAC, FGeoServer, GDAL		Human inspection needed

*** - Borrowed, with permission, from NIST Big Data Use Case Submissions [http://bigdatawg.nist.gov/usecases.php]**

s/w = software; ds = dataset; db = database