Earth Science Data Analytics: Bridging Tools and Techniques with the Co-Analysis of Large, Heterogeneous Datasets

Steve Kempler[1], Tiffany Mathews[2]

[1] NASA Goddard Space Flight Center, [2] NASA Langley Research Center

Thanks to the work of the Earth Science Information Partners (ESIP) Federation, Earth Science Data Analytics (ESDA) Cluster

Abstract

The continuum of ever-evolving data management systems affords great opportunities to the enhancement of knowledge and facilitation of science research.

To take advantage of these opportunities, it is essential to understand and develop methods that enable data relationships to be examined and the information to be manipulated.

This presentation describes the efforts of the **Earth Science Information Partners (ESIP) Federation Earth Science Data Analytics (ESDA) Cluster** to understand, define, and facilitate the implementation of ESDA to advance science research.

As a result of the void of Earth science data analytics publication material, **the cluster has defined ESDA along with 10 goals** to set the framework for a common understanding of tools and techniques that are available and still needed to support ESDA.

The Challenge

How Do We Facilitate the Use of Large Amounts of Heterogeneous Data?
How Can We Glean Knowledge from the Plethora of Available Information?

Scoping the Challenge

Suddenly everything became 'BIG'… or at least we came to notice

But, what's new?... what's different?... what's the problem?

- We have been managing large volumes of heterogeneous datasets for a long time
- Researchers have been analyzing this data for a long time
- Technology is accommodating our needs

What is new is the need to grow and implement the ability to efficiently analyze data and information in order to extract knowledge

Thus, it is not necessarily about 'Big', itself.

It is about the ability to examine large amounts of data of a variety of types to uncover hidden patterns, unknown correlations and other useful information.

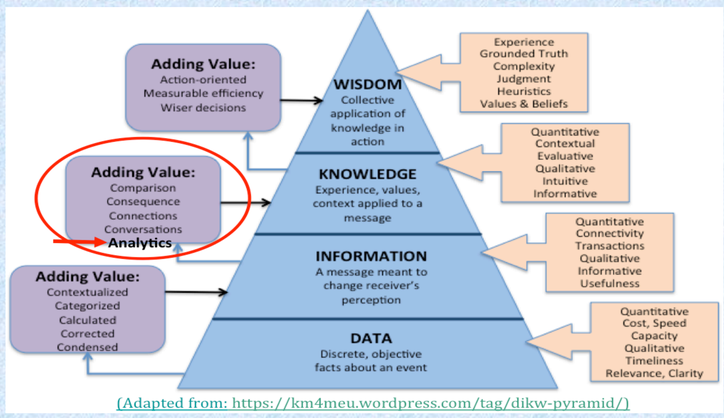
That is:

To glean knowledge from data and information

The Solution Set - Gleaning Knowledge about Earth from All Available Data and Information

On the continuum of ever evolving data management systems, we need to understand and develop ways that allow for heterogeneous data relationships to be examined, and information to be manipulated, such that knowledge can be enhanced, to facilitate science.

In short, we have a lot of heterogeneous data that we really have not provided opportunity for users to holistically 'mine'.



(Adapted from: https://km4meu.wordpress.com/tag/dikw-pyramid/)

The five types of **Data Analytics** that describe Business Analytics (descriptive, predictive, etc.), do not fit the science paradigm.

For Earth Science:

- We do not necessarily come up with the answers, but typically come up with discoveries that explain, at least for now, an answer.
- Analytics are goal oriented

ESIP Federation Earth Science Data Analytics: Definition

The process of examining, preparing, reducing, and analyzing large amounts of spatial (multi-dimensional), temporal, or spectral data encompassing a variety of data types to uncover patterns, correlations and other information, to better understand our Earth.

Data Preparation
Preparing heterogeneous data so that they can be jointly analyzed

Data Reduction
Correcting, ordering and simplifying data in support of analytic objectives

Data Analysis
Applying techniques/methods to derive results

Earth Science Data Analytics: Goals

- To validate data
- To perform coarse data preparation
- To calibrate data
- To assess data quality
- To intercompare datasets
- To tease out information
- To glean knowledge
- To derive conclusions
- To forecast/ predict/model
- To derive new analytics tools

ESDA Techniques and Tools Identified and Described (http://wiki.esipfed.org/index.php/Earth_Science_Data_Analytics)

Data Preparation

- Aggregation
- Bias Correction
- Classification
- Clustering; Hierarchical
- Coordinate Transformation
- Data Mining
- Exponential Differentiation
- Factor Analysis
- Filtering
- Format Conversion
- Imputation
- Map Reduce
- Normalization/ Transformation
- Outlier Removal
- Ratios
- Sensitivity Analysis
- Smoothing; Exponential
- Spatial Interpolation

Data Reduction

- Aggregation
- Anomaly Detection
- Bias Correction
- Classification
- Clustering; Hierarchical
- Data Fusion
- Data Mining
- Factor Analysis
- Filtering
- Machine Learning
- Neural Networks
- Outlier Removal
- Ratios

Data Analysis

- Anomaly Detection
- Bayesian Techniques
- Bivariant Regression
- Constrained Variational Analysis
- Correlation/Regression Analysis
- Factor Analysis
- Fourier Analysis
- Gaussian Distribution
- Graph Analytics
- Imputation
- Linear/Non-linear Regression
- Machine Learning
- Monte Carlo method
- Multi-variate time series analysis
- Pattern Recognition
- Principal Component Analysis
- Spectral Analysis
- Temporal Trend Analysis

Samplings

ESDA Techniques Descriptions

Bayesian Techniques	Bayesian analysis, a method of statistical inference that allows one to combine prior information about a population parameter with evidence from information contained in a sample to guide the statistical inference process. Bayesian Synthesis.
Bivariant Regression	The simplest form of regression is bivariate regression, in which one variable is the outcome and one is the predictor.
Classification	The problem of identifying to which of a set of categories a new observation belongs
Clustering; Hierarchical Clustering	An approach to organize objects into a classification and can be accomplished utilizing various methods, including statical techniques.
Constrained Variational Analysis	A field of mathematical analysis that deals with maximizing or minimizing functionals, which are mappings from a set of functions to the real numbers
Coordinate Transformation	Put data into a different coordinate system
Correlation Analysis/ Regression Analysis	Correlation and regression analysis are related in the sense that both deal with relationships among variables. The correlation coefficient is a measure of linear association between two variables. Correlation computes the value of the Pearson correlation coefficient, r. Its value ranges from -1 to +1.
Data Fusion	The process of integration of multiple data and knowledge representing the same real-world object into a consistent, accurate, and useful representation.

ESDA Tools Descriptions

Hadoop	Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware
Java	Java is a general-purpose computer programming language that is concurrent, class-based, object-oriented,[13] and specifically designed to have as few implementation dependencies as possible.
Javascript	A high level interpreted language used by most websites and browsers.
MATLAB	MATLAB is a multi-paradigm numerical computing environment and fourth-generation programming language.
Minitab	Minitab is a statistics package developed at the Pennsylvania State University
MySQL	MySQL is an open-source relational database management system (RDBMS);
Parallel NetCDF	Parallel NetCDF is a library providing high-performance parallel I/O while still maintaining file-format compatibility with Unidata's NetCDF, specifically the formats of CDF-1 and CDF-2.
Perl	A high level interpreted scripting language frequently used on UNIX computers. It is frequently used to wrap other programs together.
PHP	A scripting language designed for web development. It can be used to create CGI (Common Gateway Interface) executable for web pages.
PostgreSQL	PostgresSQL is an open-source relational database management system (RDBMS);
Python	Python is a widely used general-purpose, high-level programming language. Its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer lines of code than would be possible in languages such as C++ or Java. The language provides constructs intended to enable clear programs on both a small and large scale. (Wikipedia)