An Analysis of Earth Science Data Analytics Use Cases

Analyzing how Earth Science Data Analytics (ESDA) are used reveals foundation for ESDA goals categorization

Abstract
Earth Science Data Analytics (ESDA) is the process of examining large amounts of data of various types to uncover hidden patterns, unknown correlations and other useful information. It can include Data Preparation, Data Reduction, and Data Analysis. Through work associated with the Earth Science Information Partners (ESIP) Federation, a collection of use cases have been analyzed for the purpose of extracting the types of Earth Science data analytics employed, and requirements for tools and techniques yet to be implemented, based on use case needs. ESIP generated use case template. ESDU use cases, use case types, and preliminary use case analysis (this is a work in progress) will be presented.

Goals of Earth Science Data Analytics

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 coarse data preparation (e.g., subsetting, data mining, transformations, recover data)
5. To intercompare data (e.g., 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 (e.g., Special kind of conclusion)
9. To perform data preparation
10. To calibrate data

Use Cases Template

a. Use Case Title
b. Author/Company/Email - author of the use case
c. Actors/Stakeholders/Project URL and their roles and responsibilities
d. Use Case Goal - What is the goal of the Earth science data analytics?
e. Use Case Description - This yields more details regarding how data analytics is utilized.
f. Current technical requirements to take into account that may impact needed data analytics. These can include:
   - Data Source (distributed/centralized)
   - Volume (size)
   - Velocity (e.g., real time)
   - Variability - bringing distributed heterogeneous data together
   - Accuracy (Robustness Issues) / Data Quality
   - Data Analytics Challenges (Gaps): Identifying known data analytics challenges, roadblocks, areas needing attention
   - Type of User - Taken from the ESIP Use Analysis Study, types of user performing use case.
   - Dominant Data Analytics Skills Needed - Skills needed to perform use case analytics
   - Science Research Areas - NASA Earth science research areas (http://science.nasa.gov/earth-science/focus-areas/)
   - Societal Benefit Areas – GEO or NASA Applications (http://esipfed.sciences.nas.nasa.gov/)
   - Potential for and/or issues for generalizing this use case
   - More Information and relevant URLs (e.g. who to contact or where to go for more information)

Conclusions (thus far, with our limited number of use cases):
I. For Earth Science, defining results oriented Data Analytics types is 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
II. Most ESDU use cases tend to focus on data intercomparison, deriving new products, forecasting/predicting, and deriving conclusions
III. Most use cases were not identified to glean knowledge from data/information. Perhaps some use cases were not recognized as such (more analysis needed)
IV. Distributed data sources, and data heterogeneity are persistent characteristics
V. Velocity issues are not significant (thus far)
VI. ESDU challenges provide interesting problems for data analytics tool/technique developers to ponder

1. If any, use case 5 and 6 provides the true Big Data problem

Earth Science Data Analytics Goals

Other Significant Earth Science Data Analytics Considerations

<table>
<thead>
<tr>
<th>Use Cases</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Sources</td>
<td>Volume</td>
<td>Velocity</td>
<td>Variety</td>
<td>Veracity</td>
<td>Visualization</td>
<td>Specialized s/w</td>
<td>Current data analytics tools applied</td>
<td>Data Analytics Challenges</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MERIS Analytics Services: Climate Analytics-as-a-Service</td>
<td>Distributed</td>
<td>Large</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MUSTANG QA: Ability to detect seismic instrumentation problems</td>
<td>Distributed</td>
<td>Centralized</td>
<td>Small</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inter-calibrations among datasets</td>
<td></td>
<td></td>
<td>Centralized</td>
<td>Huge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inter-comparisons between multiple model or data products</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sampling Total Precipitable Water Vapor using AIRS and MERRA</td>
<td>Distributed</td>
<td>Collocated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using Earth Observations to Understand and Predict Infectious Diseases</td>
<td></td>
<td></td>
<td>Centralized</td>
<td>Large</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CREATE-IP: Collaborative REAnalysis Technical Environment - Intercomparison Project</td>
<td>Distributed</td>
<td>Up to one PB</td>
<td>Different</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The GSSTF Project (MEASUREs-2006)</td>
<td>Distributed</td>
<td>&gt;1 TB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Science- and Event-based Advanced Data Service Framework at GES DISC</td>
<td></td>
<td></td>
<td>Distributed</td>
<td>Diverse data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk Analysis for environmental issues</td>
<td></td>
<td></td>
<td>Distributed</td>
<td>Large</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aerossil Characterization</td>
<td></td>
<td></td>
<td>Distributed</td>
<td>Huge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reconstructing Sea Ice Extent from Early Nimbus Satellites</td>
<td></td>
<td></td>
<td>Distributed</td>
<td>Large of records</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOE-BER AmeriFlux and FLUXNET Networks</td>
<td></td>
<td></td>
<td>Distributed</td>
<td>Diverse data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOE-BER Subsurface Biogeochemistry Scientific Focus Area</td>
<td></td>
<td></td>
<td>Distributed</td>
<td>Diverse data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Climate Studies using the Community Earth System Model at DOE's MERRS center</td>
<td>Distributed</td>
<td></td>
<td>Up to 30 PB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Radar Data Analysis for CrIS/IS</td>
<td></td>
<td></td>
<td>Distributed</td>
<td>&gt;10 PB per year</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UF/UF Data Processing, Data Product Delivery, and Data Service</td>
<td></td>
<td></td>
<td>Distributed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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