Background

• NASA has historically focused on systematic capture and stewardship of data for observational systems
  – Limited use of advance computational technologies to support scientific inferences

• Increasing “big data” era is driving needs to
  – Scale computational and data infrastructures
  – Support new methods for deriving scientific inferences
  – Shift towards integrated data analytics
  – Apply computational and data science across the lifecycle

• NASA Advanced Information Systems Technology (AIST) program initiated a study of needed data and computational science techniques across the data lifecycle and have made some key recommendations
Data and Computational Science Across the Data Lifecycle

- Architectural considerations/tradeoffs for integrating the entire data lifecycle

- Onboard
  - Enable data reduction and triage close to the sensor/instrument
  - Manage bandwidth for communicating results

- Scalable Data Management
  - Capturing well-architected and curated data repositories based on well-defined data/information architectures
  - Architecting automated pipelines for data capture

- Scalable Data Analytics
  - Access and integration of highly distributed, heterogeneous data
  - Novel statistical approaches for data integration and fusion
    - Including sampling strategies
  - Computation applied at the data sources
  - Algorithms for identifying and extracting interesting features and patterns
Data Lifecycle Model for NASA Space Missions

Emerging Solutions
- Onboard Data Products
- Onboard Data Prioritization
- Flight Computing

(1) Too much data, too fast; cannot transport data efficiently enough to store

Observational Platforms /Flight Computing

Emerging Solutions
- Low-Power Digital Signal Processing
- Data Triage
- Exa-scale Computing

(2) Data collection capacity at the instrument continually outstrips data transport (downlink) capacity

Ground-based Mission Systems

Emerging Solutions
- Distributed Data Analytics
- Advanced Data Science Methods
- Scalable Computation and Storage

(3) Data distributed in massive archives; many different types of measurements and observations

Massive Data Archives and Big Data Analytics

Concept included in OCT TA-11 Roadmap (2015): 11.4.1 – Mission, Science, Engineering Data Lifecycle
NASA Earth Science Data Pipeline Today:
Constructing Scientific Archives

NASA Earth Observation System Distributed Active Archive Centers
Capture of well-curated repositories > 10 PB
Emerging Challenges as Data Increases

• Reproducibility
• Uncertainty management
• Data fusion (including distributed data)
• Data reduction
• Data movement
• Data visualization
• Cost
• Performance
## Driving Use Cases

<table>
<thead>
<tr>
<th>Use Case</th>
<th>Description</th>
<th>Data Science Challenge</th>
<th>Enabling Mission/ Capability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Climate Modeling</td>
<td>Formulate hypotheses from observed empirical relationships; Simulate current and past conditions under those hypotheses using climate models; Test hypotheses by comparing simulations to observations; Evaluate uncertainty of predictions originated from statistical sampling of models and observations.</td>
<td>Highly distributed data sources; fusion of different observations; moving computation to the data; data reduction</td>
<td>CMIP6 will move towards exascale archives requiring new approaches to evaluating models relative to observational data.</td>
</tr>
<tr>
<td>Satellite Missions</td>
<td>Missions such as NI-SAR and SWOT will generate massive observational data. However, they are have different architectural patterns including compute intensive, data intensive, heterogeneous, etc.</td>
<td>Massive data rates, data movement challenges, computational scalability, archiving and distribution; onboard processing for data reduction/analysis; high-volume data transfer for ground processing</td>
<td>NI-SAR and SWOT require new approaches for computation, data movement, data archiving and distribution, analytics.</td>
</tr>
<tr>
<td>Applications - Hydrology</td>
<td>Understanding groundwater dynamics on a regional scale using measurements from satellite, airborne and in-situ measurements. Compare against predictive models.</td>
<td>Distributed computation; highly distributed data sources; data fusion of multiple products; massive new satellite observations.</td>
<td>Integration of data from PALSAR-2, Sentinel, Grace-FO, ASO, and SMAP. Scale to support NI-SAR and SWOT. Comparison against models. Requires new architectural approaches for distributed data analytics.</td>
</tr>
<tr>
<td>(Central Valley of California)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Airborne Missions</td>
<td>Airborne missions tend to be much more agile and on-demand. Integrating this into a data ecosystem provides new opportunities to quickly generate and understand various measurements.</td>
<td>On-demand architectures; distributed data sources; on-the-fly data processing; onboard processing for data reduction/analysis; high-volume data transfer for ground processing</td>
<td>Current missions such as CARVE and Airborne Snow Observatory; Future such as proposed EVI-3 and ASO follow-on missions</td>
</tr>
</tbody>
</table>
• Flight Computing, Ground Systems, Archiving/Distribution, and Analytics are not architected into a scalable big data system...

• Problems across the data lifecycle:
  – Data Generation: Limited onboard computing (or computing at sensor) for planning
  – Data Triage: Limited onboard triage and processing
  – Data Compression: Limited intelligent data reduction
  – Data Transport: Dependent on bandwidth capabilities; challenges in moving and distributing massive data
  – Data Processing: Ground systems and ground processing have limited support for dynamic workflows, scaling to large-scale environments (clouds, HPC), integrating intelligent discovery algorithms, etc. Processing disconnected from science analysis.
  – Data Archiving: Scaling the capture, management and distribution of data; distributed archives; limited computational capabilities; different models, formats, representations of data.
  – Visualization: Limited visualization capabilities for massive data; challenges in presenting massive data to users
  – Data Analytics: Limited analytics services; generally tightly coupled to DAACs; limited cross-archive, cross-agency integration; limited capabilities in data fusion; statistical uncertainty; provenance of the results

Copyright 2016 by California Institute of Technology. All Rights Reserved.
Future: Enabling Scalable, Data-Intensive Science

- Towards the systematic analysis of massive data -
# Computational Capability Needs and Gaps Across Lifecycle*

<table>
<thead>
<tr>
<th>System</th>
<th>2015</th>
<th>2025</th>
<th>Application to Earth Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Onboard</td>
<td>Limited onboard computation including data triage and data reduction. Investments in new flight computing technologies for extreme environments.</td>
<td>Increase onboard autonomy and enable large-scale data triage to support more capable instruments. Support reliable onboard processing in extreme environments to enable new exploration missions.</td>
<td>Onboard computation for airborne missions on aircraft; new flight computing capabilities deployed for extreme environments; use of data triage and reduction for high volume instruments on satellites.</td>
</tr>
<tr>
<td>Ground Systems</td>
<td>Rigid data processing pipelines; limited real-time event/feature detection. Support for 500 TB missions.</td>
<td>Increase computational processing capabilities for mission (100x); Enable ad hoc workflows and reduction of data; Enable real-time triage, event and feature detection. Support 100 PB scale missions.</td>
<td>Future mission computational challenges (e.g., NI-SAR); support more agile airborne campaigns; increase automated detection for massive data streams (e.g., automated tagging of data).</td>
</tr>
<tr>
<td>Archive Systems</td>
<td>Support for 10 PB of archival data; limited automated event and feature detection.</td>
<td>Supportexascale archives; automated event and feature detection. Virtually integrated, distributed archives.</td>
<td>Turn archives into knowledge-bases to improve data discovery. Leverage massively scalable virtual data storage infrastructures.</td>
</tr>
<tr>
<td>Analytics</td>
<td>Limited analytics services; generally tightly coupled to DAACs; limited cross-archive, cross-agency integration; limited capabilities in data fusion; statistical uncertainty; provenance of the results</td>
<td>Analytics formalized as part of the mission-science lifecycle; Specialized Analytics Centers (separate from archives); Integrated data, HPC, algorithms across archives; Support for cross product data fusion; capture of statistical uncertainty; virtual missions.</td>
<td>Shift towards automated data analysis methods for massive data; integration of data across satellite, airborne, and ground-based sensors; systematic approaches to addressing uncertainty in scientific inferences; focus on answering specific science questions.</td>
</tr>
</tbody>
</table>

*Derived from OCT TA-11 Roadmap (2015)*
# Proposed Technology Areas

<table>
<thead>
<tr>
<th>Technology Name</th>
<th>Data Lifecycle Area (s)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Data Architecture Earth Science Remote Sensing</td>
<td>Cross-Cutting</td>
<td>Definition of a scalable data big data lifecycle architecture for earth observing systems identifying how Big Data can scale from onboard computing to data analysis to increase science yield.</td>
</tr>
<tr>
<td>Big Data Information Models and Semantics</td>
<td>Cross-Cutting</td>
<td>Advanced semantic technologies for defining, deriving, and integrating heterogeneous ontologies and information models as applied across the entire data lifecycle (onboard, ground-based operations, archives, analysis)</td>
</tr>
<tr>
<td>Onboard data science methods for data triage</td>
<td>Data triage</td>
<td>Onboard data science methods for real-time event detection, and planning.</td>
</tr>
<tr>
<td>Onboard data science methods for data reduction</td>
<td>Data Compression</td>
<td>Onboard data science methods for data reduction.</td>
</tr>
<tr>
<td>Massive Data Movement Technologies</td>
<td>Data Transport</td>
<td>Massive data movement technologies for ground-based networks from operations through analysis.</td>
</tr>
<tr>
<td>Real-time ground-based data science methods</td>
<td>Data Processing</td>
<td>Real-time ground-based data science methods for data reduction and real-time event detection for massive data streams as part of the data lifecycle architecture.</td>
</tr>
<tr>
<td>Open source data processing frameworks</td>
<td>Data Processing</td>
<td>Open source data processing and workflow frameworks that can massively scale to computational infrastructures (HPC, public cloud, etc.) handling large data streams, products, including near-real time constraints, as part of the data lifecycle architecture.</td>
</tr>
<tr>
<td>Reusable data science methodologies for missions and science</td>
<td>(1) Data Processing; (2) Data Analytics</td>
<td>Development of reusable data science methodologies for analysis of data on the ground as part of the data lifecycle architecture. This includes on-demand data analytics for massive data repositories.</td>
</tr>
</tbody>
</table>
## Proposed Technology Areas (2)

<table>
<thead>
<tr>
<th>Federated data access</th>
<th>Data Archives</th>
<th>Federation of data access from distributed repositories as part of the data lifecycle architecture, moving towards on-demand distributed data analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Massive Data Distribution</td>
<td>Data Archives</td>
<td>Massive data distribution for large-scale repositories and archives including methods for data reduction, computation, etc., as integrated, on-demand data analytics.</td>
</tr>
<tr>
<td>Intelligent search and mining</td>
<td>(1) Data Archives; (2) Data Analytics</td>
<td>Provide methods for intelligent search and mining of massive data. This may include integration of on-demand analytics to perform deep searches.</td>
</tr>
<tr>
<td>Visualization of massive data sets</td>
<td>Visualization</td>
<td>Visualization of massive data sets including data reduction methods that are driven by domain.</td>
</tr>
<tr>
<td>On-demand distributed data analytics</td>
<td>Data Analytics</td>
<td>On-demand data analytics that can integrate data from archives, repositories, etc., applying data science methods (data reduction, fusion, feature detection, etc.) provided through a computational infrastructure</td>
</tr>
<tr>
<td>Distributed data analytics</td>
<td>Data Analytics</td>
<td>Analysis of data across distributed archives to support Earth system science</td>
</tr>
<tr>
<td>Uncertainty Quantification; Measurement Science</td>
<td>Data Analytics</td>
<td>Management of uncertainty in scientific inferences as part of a measurement science strategy for data fusion and data science</td>
</tr>
<tr>
<td>Open source data management/science frameworks</td>
<td>(1) Data Archives; (2) Data Analytics</td>
<td>Open source data management/science frameworks that can massively scale to handle and manage large data streams, products, including near-real time constraints, as part of the data lifecycle architecture, for archiving and analytics as part of a big data cyber-infrastructure.</td>
</tr>
<tr>
<td>Computational Infrastructures</td>
<td>(1) Data Processing; (2) Data Analytics</td>
<td>Computational Infrastructures to scale data analytics using HPC and public cloud. This includes on-demand massive HPC and storage for integration to drive analytics.</td>
</tr>
</tbody>
</table>
Cloud Computing: Enabling the Data Ecosystem

- A platform for Data and Computational Science
  - Ground systems
  - Archive systems
  - Data analytics

- Delivering data and computational services
  - APIs for data access
  - PGEs for data processing
  - Algorithms for data integration within and across systems
  - Algorithms for reduction, classification, event detection, etc

- Scalability on-demand

Copyright 2016 by California Institute of Technology. All Rights Reserved.
Selecting Cloud Topologies for Scalability

- DAWN (Distributed Analytics, Workflows and Numeric) is a model for simulation and optimization of system architectures for intensive data processing.
- Particularly suited to analyze the deployment of a processing pipeline on the Cloud:
  - Can predict application performance as a function of allocated Cloud resources.
  - Can “score” different Cloud topologies (for same resources) based on performance.

**EDRN Example:** duration of CPTAC data processing pipeline versus number of processing nodes.

-> **Conclusion:** allocate 18 nodes, no more gain after that.

**Climate Example:** centralized vs distributed architecture for comparing models and observations as a function of network speed.

-> **Conclusion:** distributed architecture is more efficient, more so for slower network and less powerful servers.
Key Recommendations

• Shift from ad hoc investments across the mission and science data lifecycle to an *integrated architecture* where technology investments fit into a broader capability to enable earth system science.
  – Big Data Architectures should be *modeled and assessed* overall to address and plan technology capabilities and improvements to ensure that architectural support for science activities can scale and meet performance, cost, and uncertainty goals.
  – Architectures should enable *flexible and transparent tradeoffs* of where to compute including improved integration of HPC and data infrastructures.

• **Formalize** *data analytics* as a first class capability across the data lifecycle
  – Shift from a stewardship model to a *data-driven discovery model* where both stewardship and data discovery are enabled through a systematic computational infrastructure.
  – Data discovery methods should be applied *across the entire data lifecycle* to support scalable science activities at each point, sometimes automated, from onboard computing, to data processing and archive, to analysis and discovery.

• Computation and data science should play an important role in *planning new missions* including identification of how data, algorithms, and computation are to be integrated to improve overall data discovery, reproducibility and uncertainty management.
  – New capabilities should improve *reproducibility* of derived scientific results.
  – Derived scientific inferences should be *measurable and quantifiable*.
Acknowledgements

- NASA AIST Program
- JPL Data Science Working Group
- NASA OCT TA-11 Roadmap Team

See: http://ieee-bigdata-earthscience.jpl.nasa.gov/references for more details

Copyright 2016 by California Institute of Technology. All Rights Reserved.