Hybrid Architectures

Supporting data science analytics for multiple missions with streaming, interactive and batch analytics components

© 2015 by Cray. Published by The Aerospace Corporation with permission
Challenge: evaluating new analytics methods

- Many organizations have mature, established workflows for their core processes
- These workflows tend to be rigid, having been developed over the years to well-defined requirements
- New analytics methods must be efficiently researched, prototyped and evaluated to be of value to the enterprise
- This exploration is best performed outside of existing workflows
- Hybrid architectures are the best high-level environment for enabling this exploration
What is a hybrid architecture?

- Using the right tool for the job
- Having hammers, screw drivers and paint brushes
- Leveraging a set of complementary computing resources
Analytics classes

Streaming – process each data element as it arrives in near real-time
- Apache Storm
- Spark Streaming
- IBM InfoSphere Streams

Batch – process all of the data at once
- Hadoop MapReduce
- MPI

Interactive – provide human-interactive response times to user queries
- Apache Jena
- Apache HBase/Hive
Key considerations

- Time to first solution is more important than overall run time. “Fail fast” to validate analytics approaches and value.
- Measure performance improvements by order of magnitude.
- “Data gravity” – Only move the data once
- Machine-to-machine interfaces for data flow
  - Start general and “inefficient”; get more efficient as needed
  - SPARQL/RDF
  - Spark/RDDs
## Example workflows

<table>
<thead>
<tr>
<th>Workflow name</th>
<th>ETL</th>
<th>Analytics and architecture</th>
<th>User display</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social media analysis</td>
<td>Parse social media transactions</td>
<td>Batch – perform graph clustering</td>
<td>Present results in an interactive display</td>
</tr>
<tr>
<td>Computer network analysis</td>
<td>Parse transactional and enrichment data sources</td>
<td>Interactive – support multiple exploratory graph operations</td>
<td>Link chart and tabular display</td>
</tr>
<tr>
<td>Topic trending and identification</td>
<td>Parse Wikipedia</td>
<td>Batch – perform latent semantic indexing/SVD</td>
<td>Present tabular results and highlight changes</td>
</tr>
<tr>
<td>Key cyber-terrain identification</td>
<td>Parse transactional computer network information</td>
<td>Streaming – build histograms and perform change detection</td>
<td>Send email alerts of significant events</td>
</tr>
<tr>
<td>Threat fusion and intelligence</td>
<td>Parse open-source and transactional information</td>
<td>Batch and interactive – identify attributes of interest</td>
<td>Prepare visual and tabular summary results as static displays</td>
</tr>
</tbody>
</table>
Key elements of a hybrid architecture

- Workflow manager
- Batch engine
- Streaming engine
- Interactive engine
- Results store
- Visualization/user display

- Kepler/ArcGIS ModelBuilder
- BDAS/Spark
- Hive/HBase
- Tableau
- Gephi/Cytoscape
- Thin client
Contact information

Louis Hackerman
lhackerman@cray.com | 301-910-6416

Eric Dull
edull@cray.com | 408-771-3174