# Hybrid Architectures

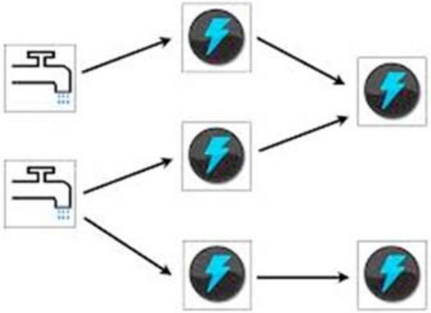
## Supporting data science analytics for multiple missions with streaming, interactive and batch analytics components

# 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**

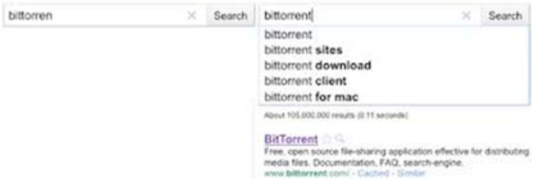
- Apache Storm
- Spark Streaming
- IBM InfoSphere Streams

**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



bittorrent
bittorrent **sites**
bittorrent **download**
bittorrent **client**
bittorrent **for mac**

About 105,000,000 results (0.11 seconds)

BitTorrent
Free, open source file-sharing application effective for distributing media files. Documentation, FAQ, search-engine.
www.bittorrent.com/ - Cached - Similar

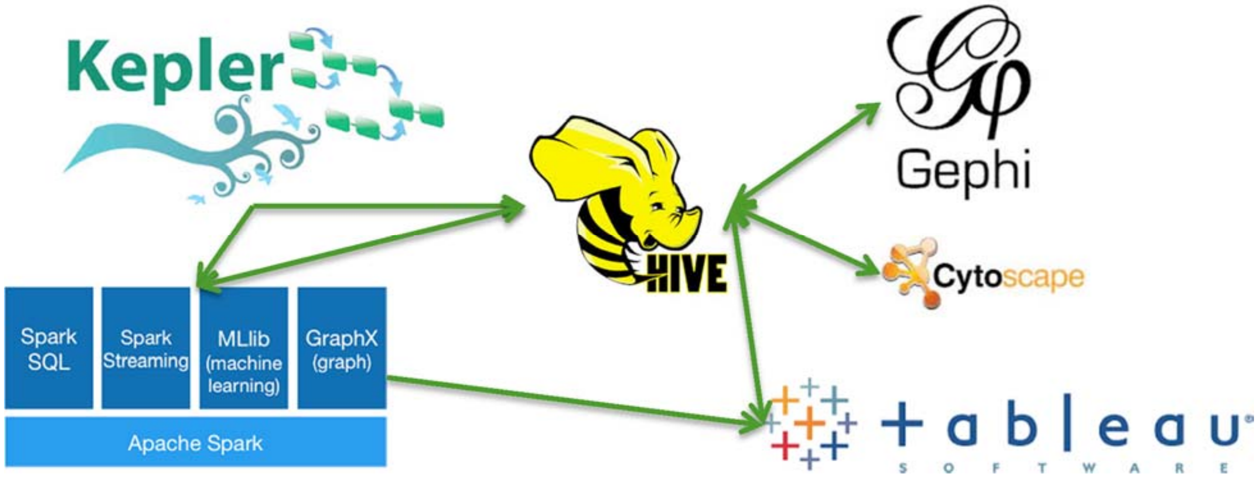
# Key considerations

- **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

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

# Key elements of a hybrid architecture



| | |
|---|---|
| ● Workflow manager | • Kepler/ArcGIS ModelBuilder |
| ● Batch engine | • BDAS/Spark |
| ● Streaming engine | • Hive/HBase |
| ● Interactive engine | • Tableau |
| ● Results store | • Gephi/Cytoscape |
| ● Visualization/user display | • Thin client |

# Contact information

**Louis Hackerman**

lhackerman@cray.com | 301-910-6416

**Eric Dull**

edull@cray.com | 408-771-3174