Analyzing Streaming Data with Cayuga

David Lifka
Streaming Data Analysis Example Applications

- Monitoring the stock market
- Sensor data
  - Oceanography
  - Seismology
  - Climatology (wind, precipitation, temperature)
- Satellite data
- Traffic & weather mashups
- Systems administration event log management
Cornell Database Group’s Cayuga Project

• Project web site
  – http://www.cs.cornell.edu/bigredata/cayuga/

• Designed as a data aggregation point to “monitor” multiple pub/sub data sources whose data have a high degree of complexity

• Provides a SQL-like event language for users to specify the events/conditions they are interested in
Challenge: Managing Ranger System Logs

- Ranger: one of largest HPC systems in the world
  - 3,936 compute nodes
  - 62,976 processing cores
  - 123 TB aggregate RAM
  - 72 I/O servers
  - 1.73PB shared disk space (31.5TB local)
  - Full-CLOS Infiniband interconnect

- Ranger produces millions of lines of event log data per day
  - Opensm (Infiniband subnet manager)
  - Sun Grid Engine job scheduling logs
  - Lustre parallel filesystem logs
  - Hardware diagnostic data
Predicting and Preventing System Failures

• Goal: Providing better system availability and reliability

• Identify early warnings and trends that allow systems administrators to prevent system failures or minimize their impact

• Analyzing Ranger system logs in real-time (streaming data) is essential to identify warnings or errors as soon as possible

• Correlating data in one event log with another is necessary to distinguish between expected and unexpected error messages

• Cayuga: Merge multiple system logs into a single time-stamp ordered data stream to gain insight into overall system health
Examining Ranger Infiniband Logs

- Ranger users the OpenFabrics Alliance Infiniband Subnet Manager
  - http://www.openfabrics.org

- The OpenSM produces ~300 million lines of messages per day!
  - Relevant error messages or warnings are buried in amongst millions of information messages…

- The first task is finding the needles in the haystack
  - http://www.cac.cornell.edu/~lifka/Ranger/Ranger.htm
Cayuga: Putting the Pieces Together
Workshop Example

- Cayuga Server
  - CayugaConfig.xml
  - OpenSMSchema.xml
  - OpenSM_query.txt

- Sun Grid Engine Logs
- Lustre Logs
- Node Hardware Logs
- Machine Room Environment Logs

- OpenSM Logs

- SendLogt oCayuga.py
- SendLogt oCayuga.prop

Event Handler.py