Monitoring Hadoop and Datameer

The Hadoop configuration can be fine-tuned to optimize cluster performance. To manage a cluster effectively, you need to monitor the hardware and system performance to catch issues before they become problems.

You can monitor Java applications such as Hadoop and Datameer with JMX and monitor the underlying OS with SNMP to watch the machines CPU activity levels, memory usage, network traffic levels, disk IO, and so on. There are a variety of applications available that provide real-time monitoring, and alerts.

Some tools commonly used for monitoring include:

- Nagios - an open source computer system and network monitoring software application. It watches hosts and services, alerting you when something goes wrong and again when it get better. It can be used for tasks including a ping check (is the machine there), disk percentage full, that the swap level is less than 85%, and so on http://www.nagios.org/ For additional information, see: Monitoring Hadoop and Datameer using Nagios
- Cacti - gathers system information and provides a graphical view for monitoring LAN-sized installations. It has templates which make it easier to scale to monitor a large number of data sources http://www.cacti.net/
- JMX - Java Management Extensions (JMX) is a Java technology that supplies tools for managing and monitoring applications, system objects, and devices. It pulls information from Hadoop and monitors Java applications. http://java.sun.com/javase/technologies/core/mntr-mgmt/javamanagement/
- Ganglia - a scalable distributed monitoring system for high-performance computing systems such as clusters and Grids. It is based on a hierarchical design targeted at federations of clusters. Learn how to install it at http://ganglia.sourceforge.net/
- Hyperic - an open source systems monitoring, server monitoring, and IT management tool. http://www.hyperic.com/

In a cluster, the most vital machines are the NameNode, SecondaryNameNode, and the Job Tracker. If the slave servers go down, there is built-in redundancy so their configuration isn't as vital.

To learn about best monitoring practices, see http://www.cloudera.com/blog/2009/11/hadoop-world-monitoring-best-practices-from-ed-capriolo/

Learn more:

- http://www.jointhegrid.com/hadoop/ (Hadoop, cacti-jtg is designed for graphing JMX attributes from Hadoop)