Chukwa: A scalable log analysis framework on top of Hadoop

Motivation

- Logs are on remote machine
- Hard to collect/access logs from thousand of machines
- Hard to correlate information from different system
- Unable to extract useful information from terabytes of data
- No easy way to detect failures on thousand of machines

Chukwa Goals

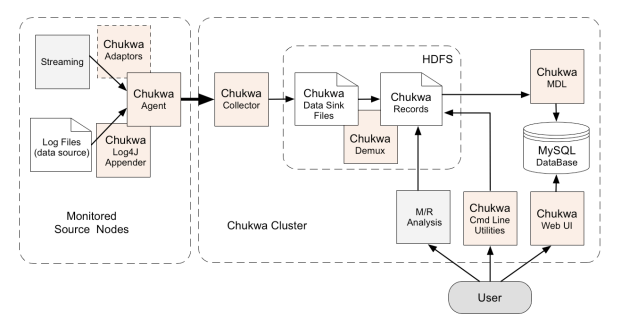
- Collect
  - Arbitrary log files (unknown format)
  - Known log files (well define format)
  - Handle log rotation
  - Latency should be in minutes but not in hours
  - Scale to large cluster
- Store large volume of data : all data in one place
- Advanced log analytics and data mining
- Reporting framework

Audience

- Application owners
- Performance engineers
- End users
- Grid ops

Advantages

- Saclable & light log collection pipeline
- Scalable log processing pipeline
- All your data in one place
- Cross system analysis
- Native M/R & Pig integration
- Open source – Apache 2.0
  http://hadoop.apache.org/chukwa/



Easy to:

- Collect application logs: log4j integration
- Collect new source of data by implementing the Adaptor interface
- Extract additional information by using an existing parser or by extending or writing your own.

Data access

- Pig or M/R to mine/extract useful information
- View Generation
- Data aggregation
- Down sampling
- RDBMS support

Alerting System

- Rule based event alert across multiple subsystem built on top of Pig
- Built in integration with Nagios

HICC

- Reporting framework
- User drag and drop customizable dashboard
- Common graph component
- Common table wizard

Hadoop integration Built in processors for:

- Hadoop metrics collection
- JobHistory
- JobConf
- GridUptime plugin
- Anomaly detection
  - Machine learning
  - Swim lanes
  - HDFS data corruption
- Metering