Clusterken: A Reliable Object-Based Messaging Framework to Support Data Center Processing

Marc Stiegler, Research Scientist
Date: 05/06/2011

Your Speaker

Marc Stiegler

Sorry to be missing

Moscow

Hadoop

- Good News
  - No Deadlock
  - No Fine Grain Data Race
  - No Error Handling for Network Partition, Node Failure
- Bad News
  - Only Spawn/Gather pattern
  - Risks with Non-Idempotent Data Mutation

Clusterken

- Good News
  - Arbitrary Patterns of Interaction across Nodes
  - No Deadlock/No Fine Grain Data Race
  - Exactly Once Message Processing
  - Transparent Checkpointing
- Consequence
  - Many of the virtues of Hadoop … and some new ones

A Fortunate Happenstance

- Are Clusterken's merits important?

- One team developing Clusterken

- One team developing a scalable pub/sub system with Hadoop
  - First developed a good spec!

- Meet the same spec using Clusterken
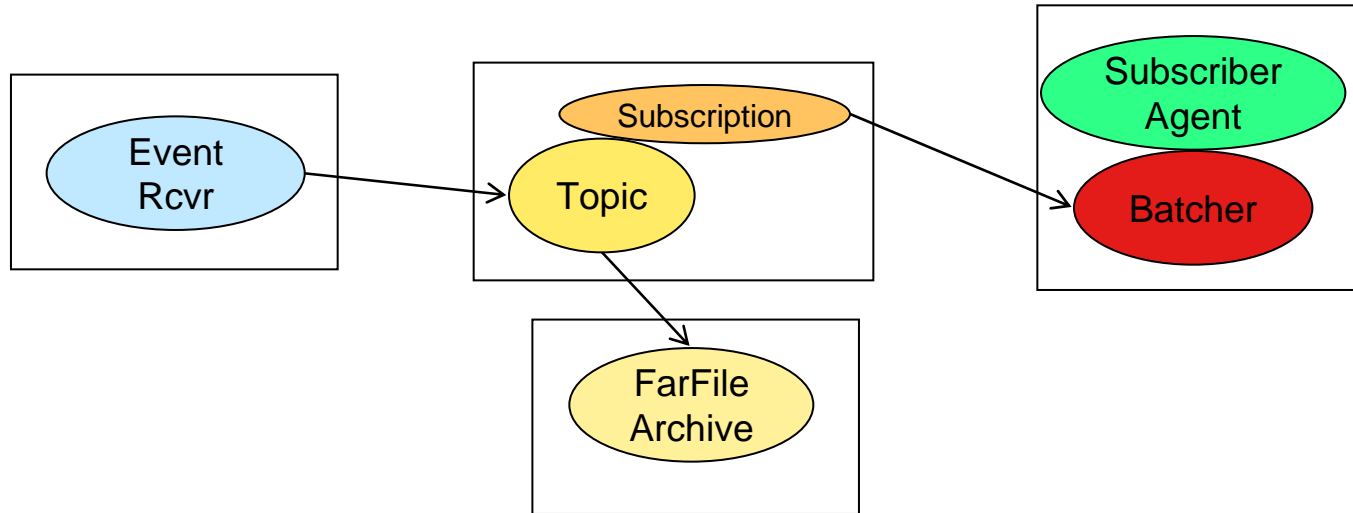
- How do they compare?

A Tale of 2 Implementations

Key Features	Hadoop	Clusterken
Effort	9 weeks	1 week
Lines of Code	1527	394
Access Control	No	Yes
Node Failure Handling	Automatic	Manual
Reliable Messaging	Limited	Automatic
Concurrent Processing	Fair Scheduler	Randomized Vats
Tuple Matching	By table join	Output validity
Event Processing	Batched	Streamed
Real-time	No	Soft
Pipeline Processing	No	Yes

Pub/Sub Core Elements

Publisher → Topic

Subscription

Topic → Archive

Subscriber

Batcher

Subscription → Batcher
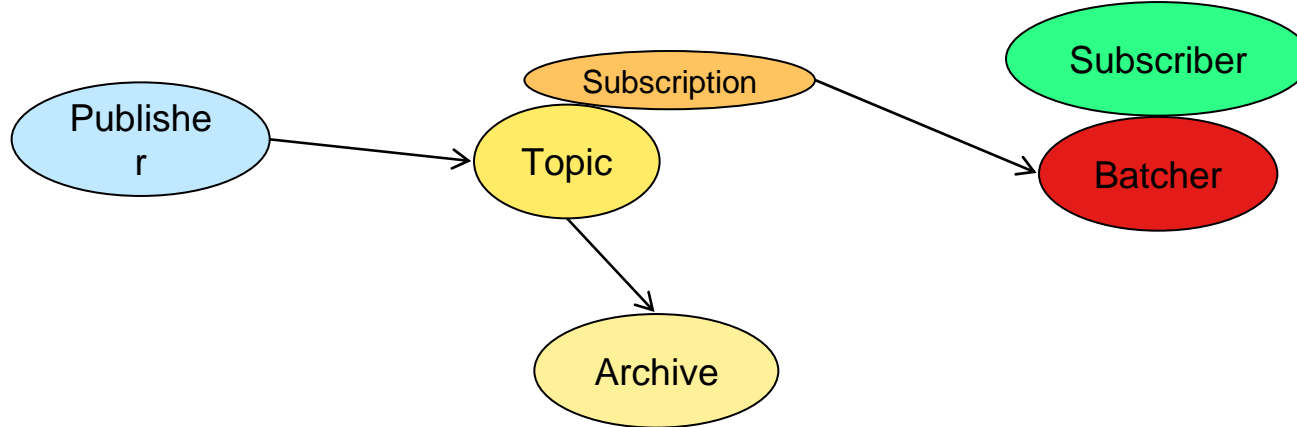
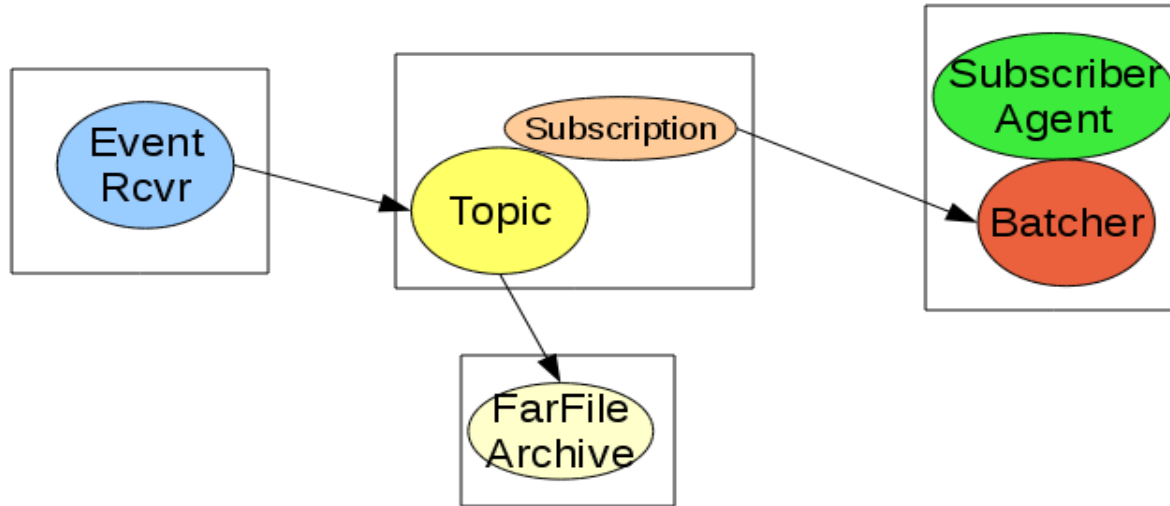# Clusterken Object Reference Graph

Conclusion

• This concludes the slideware part of the presentation
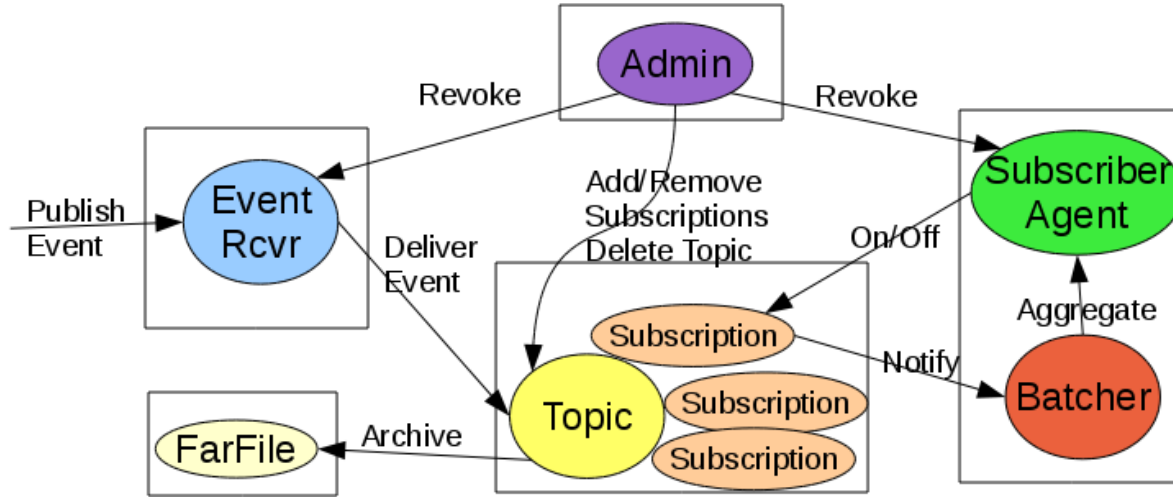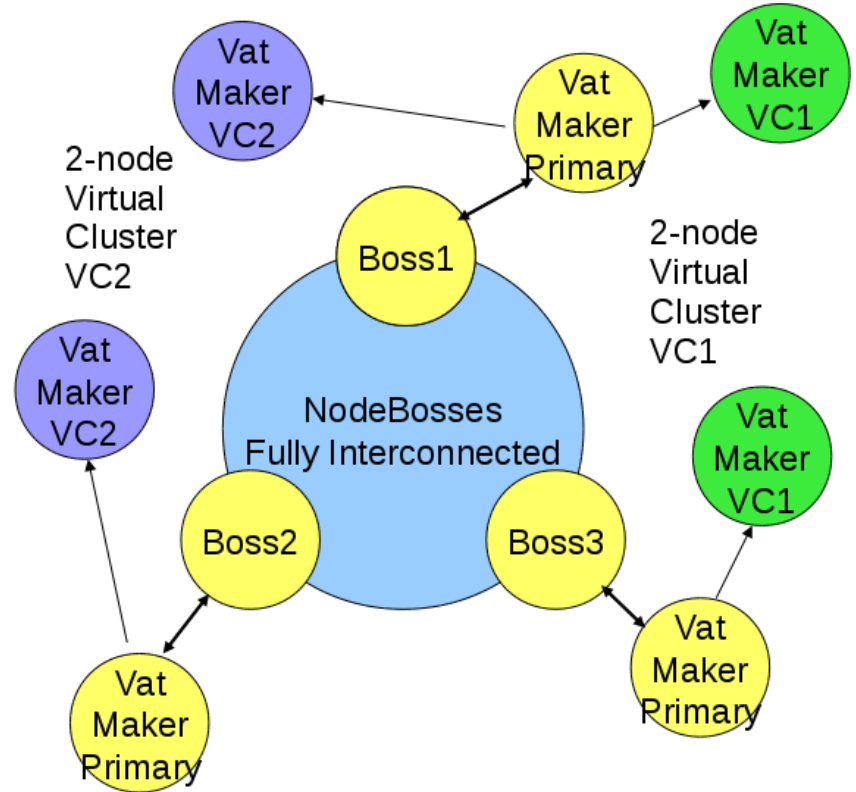
Backups

Clusterken Object Reference Graph

Pub/Sub System With Admin

Virtual Subclusters

Virtual Subclusters