Data Distribution Patterns with Apache NiFi

Hortonworks

Bryan Bende – Member of Technical Staff

March 2016

Overview

- Frequently Asked Question:
  - *"How do I distribute data across my NiFi cluster?"*

- Answer:
  - *"It depends…"*

- Typically based on whether the data is being pushed or pulled

Pushing to Listeners

- Processors listening for incoming data on each node in the cluster

- Load balancer sitting in front of the cluster pointing at listeners

- Data producers push data to load balancer which distributes data across the cluster

- Same approach works for ListenSyslog, ListenUDP, and HandleHttpRequest

Hortonworks
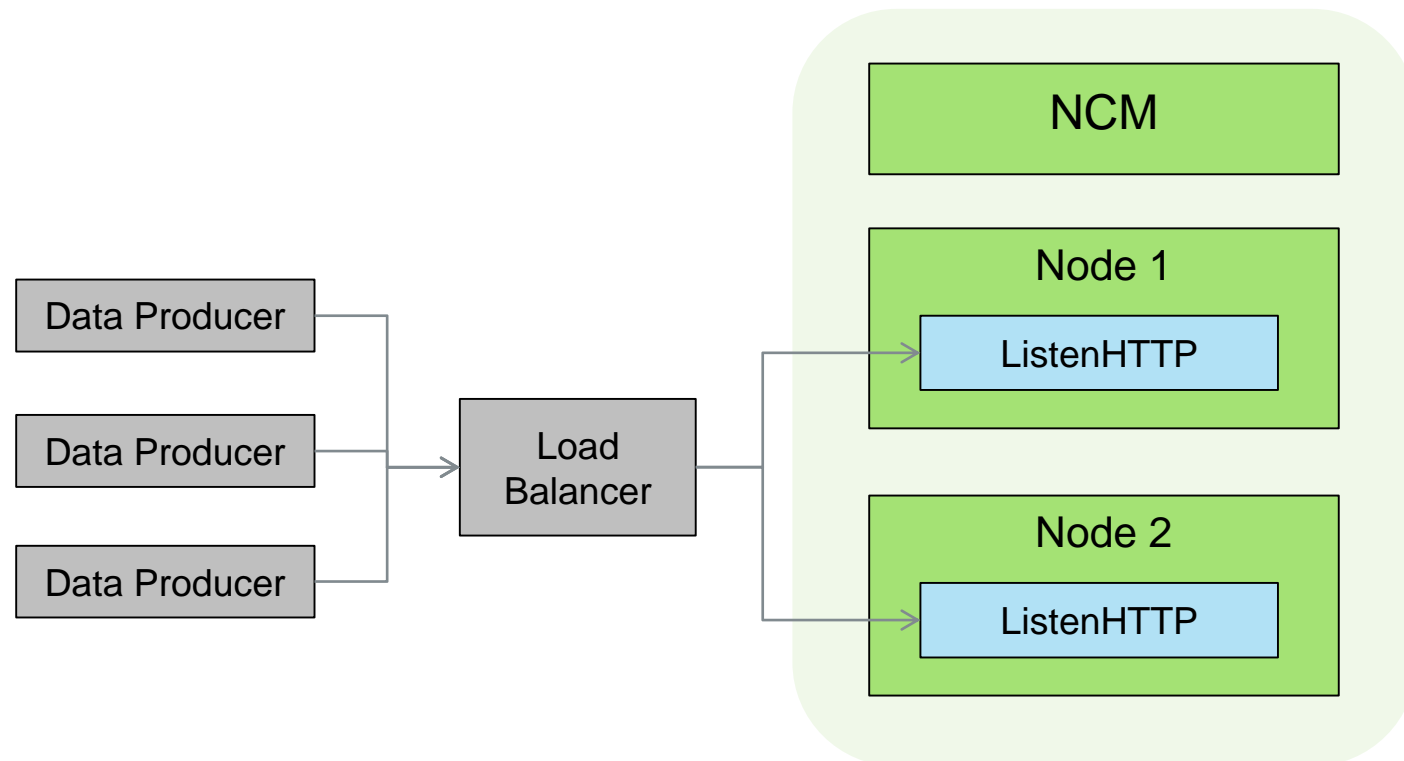
Pulling on All Nodes

- Works when data source ensures each request gets a unique piece of data

- Kafka sees each GetKafka processor as the same client

- Each GetKafka processor gets different data

NCM

Node 1

GetKafka

Kafka Topic

Node 2

GetKafka

Hortonworks
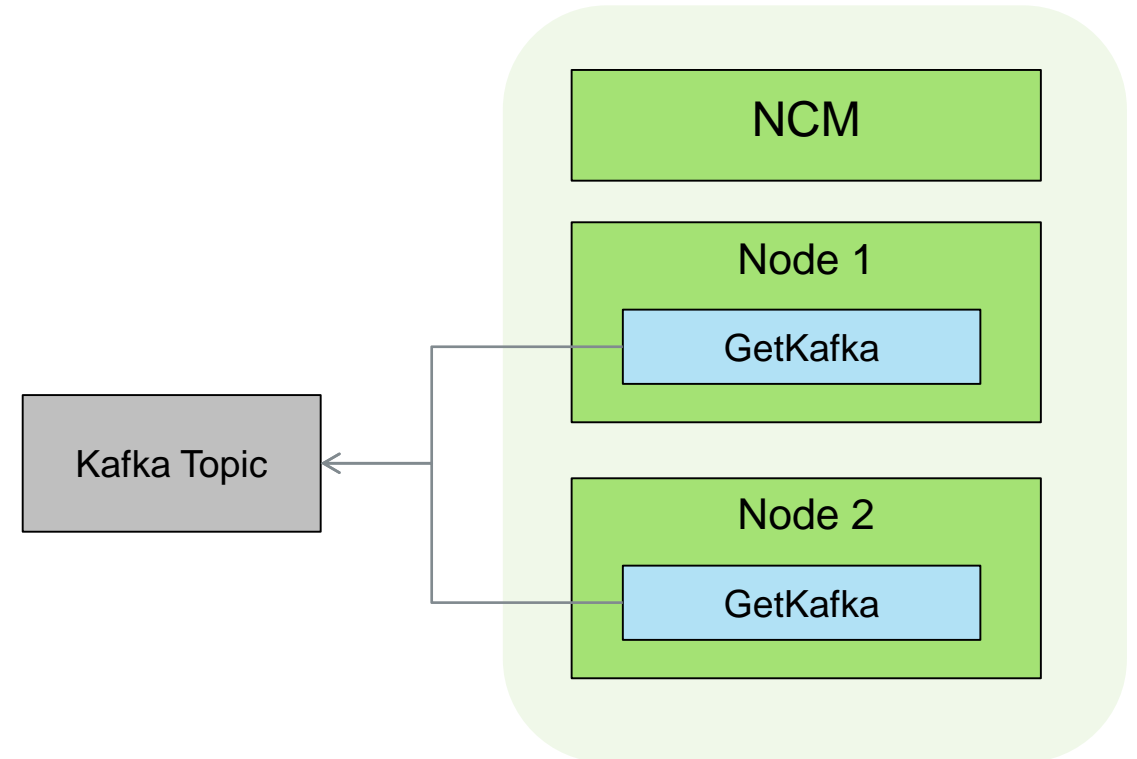
Pulling With List & Fetch Operations

- Perform "list" operation on primary node

- Send results to Remote Process Group pointing back to same cluster

- Redistributes results to all nodes to perform "fetch" in parallel

- Same approach for ListFile + FetchFile and ListSFTP + FetchSFTP
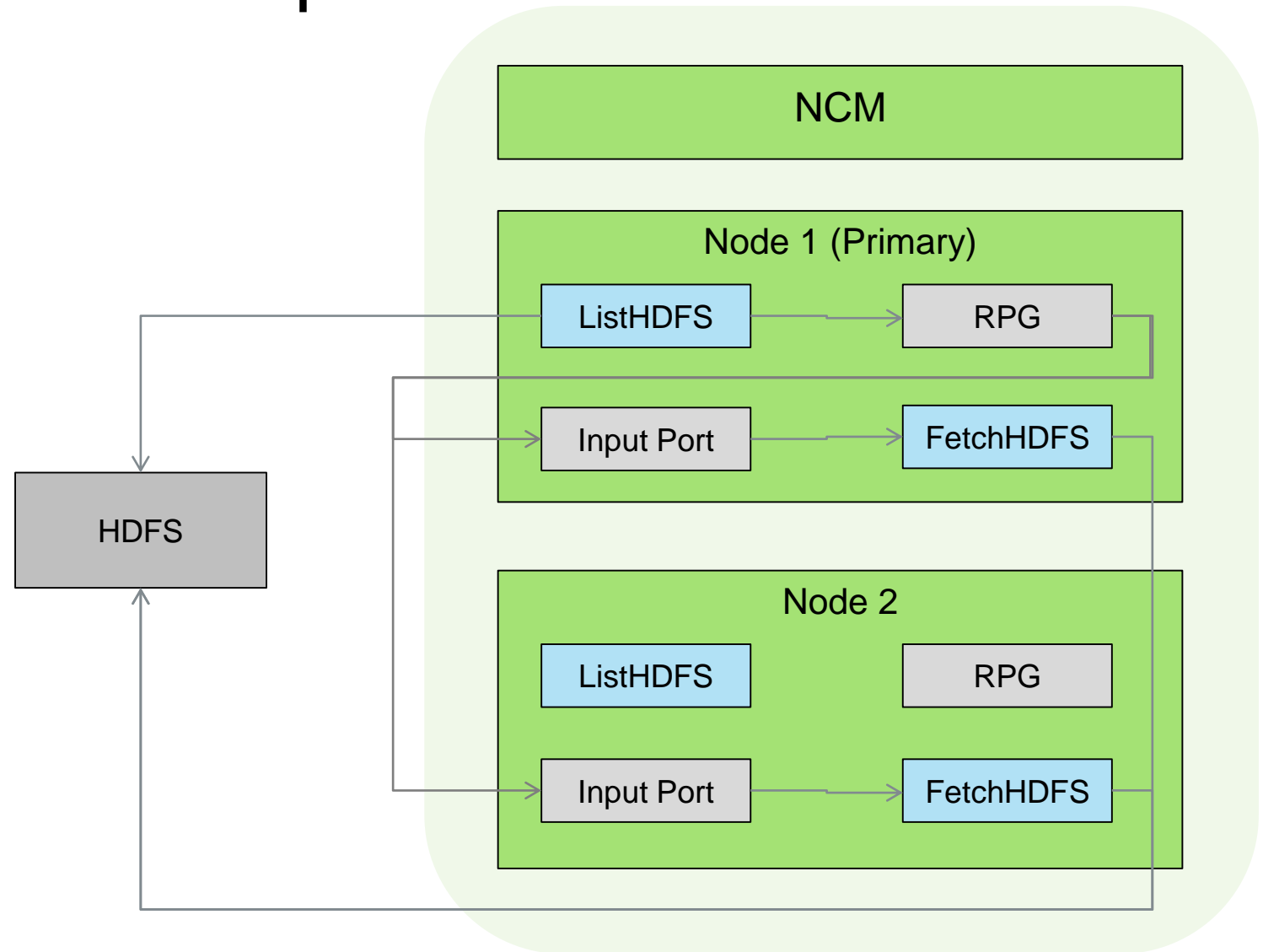
Site-To-Site Push

- Site-To-Site makes a direct connection between NiFi instances

- Either side can be a cluster or standalone instance

- In a push scenario, the source connects a Remote Process Group to an Input Port on the destination

- If pushing to a cluster, Site-To-Site takes care of load balancing across the nodes in the cluster
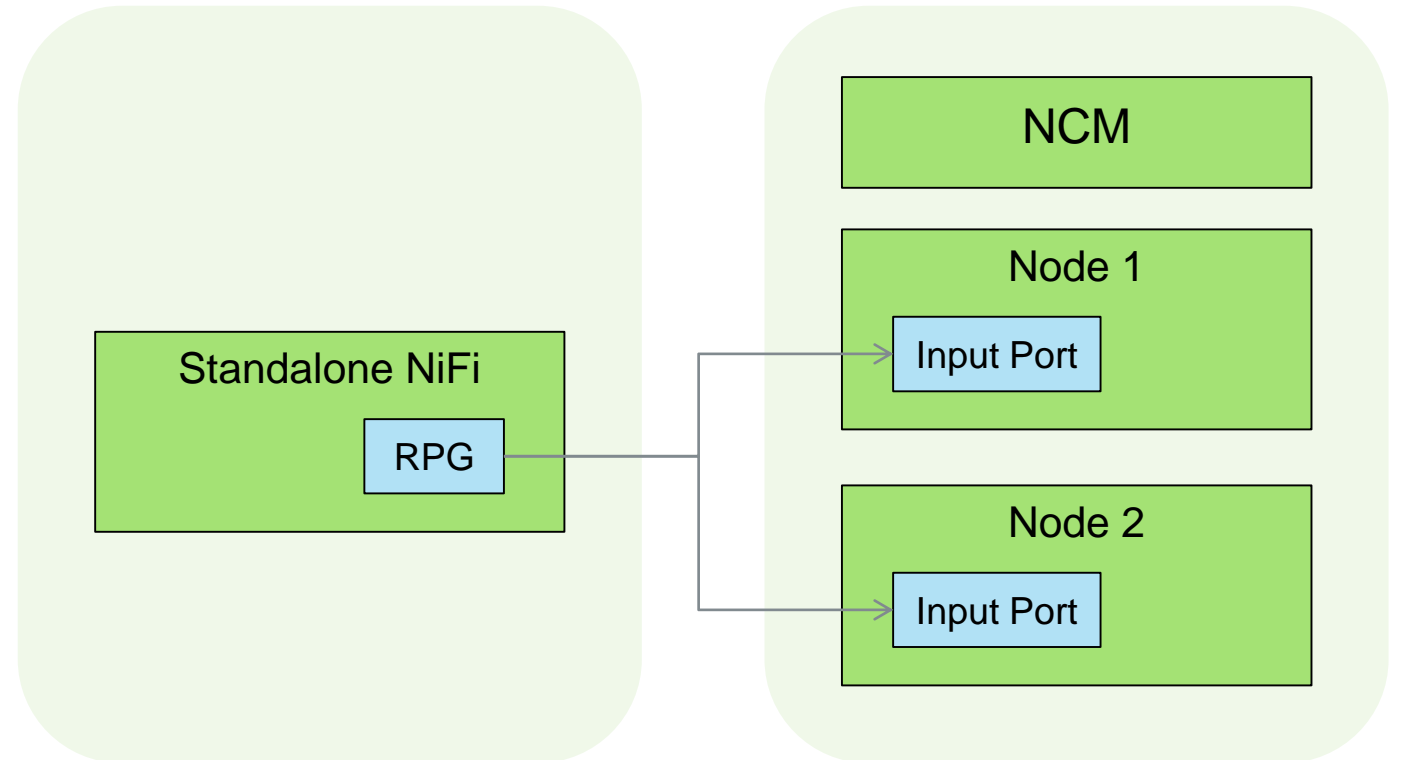
Site-To-Site Pull

- In a pull scenario, the destination connects a Remote Process Group to an Output Port on the source

- Each node will pull different data from the source

- If the source was a cluster, each node would pull from each node in the source cluster

Standalone NiFi

Output Port

NCM

Node 1

RPG

Node 2

RPG

Hortonworks
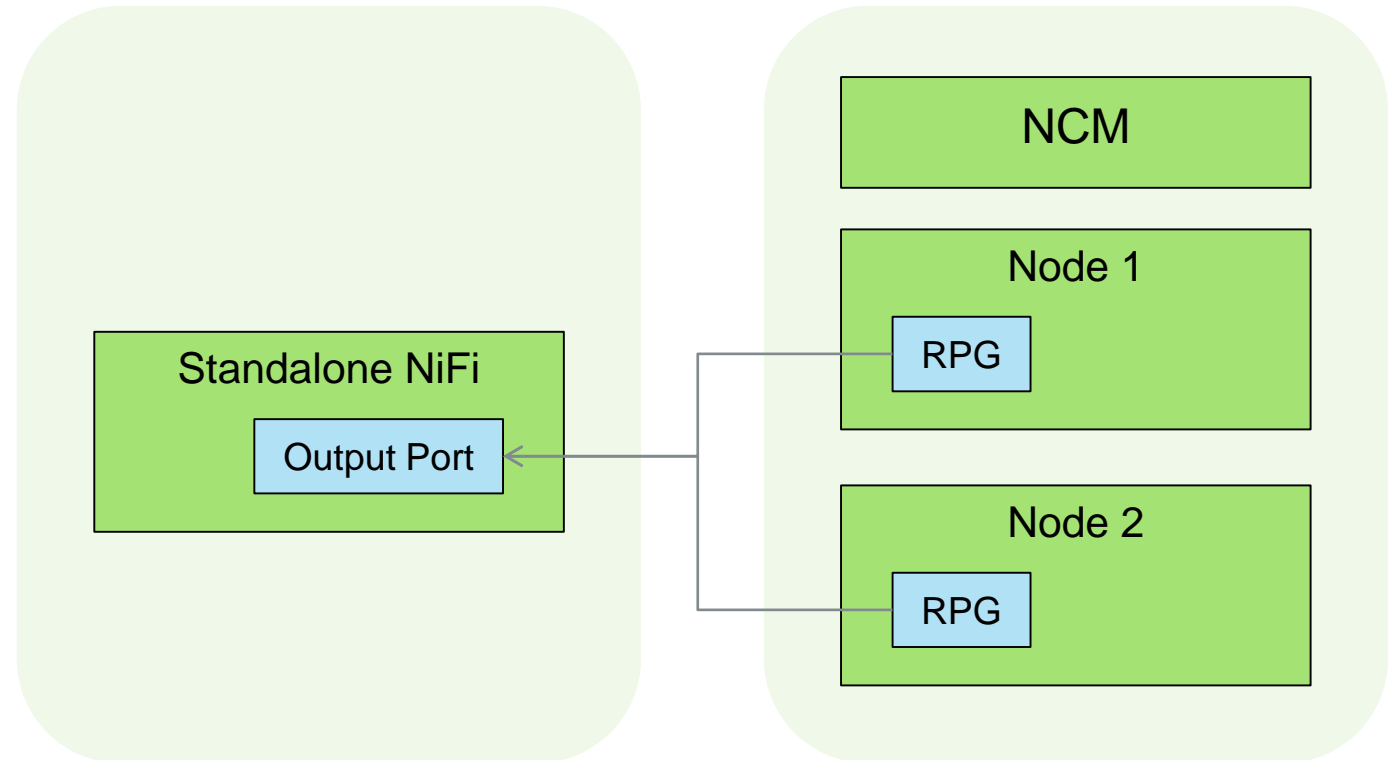
Summary

Several mechanism for distributing data across a cluster…

- Push to Listeners with a load balancer in front

- Pull from a source that provides a queue

- Pull with a List + Fetch approach

- Push with Site-to-Site

- Pull with Site-to-Site

Contact Info

- bbende@hortonworks.com

- Twitter - @bbende