

Data Ingestion & Distribution with Apache NiFi



by Aol.
PUBLISHERS

Agenda

Introduction to NiFi

Our use case for NiFi

Demo

Q&A

Introduction to NiFi

History & Facts

Created by : **NSA**

Incubating : **2014**

Available : **2015**

Main contributors: **Hortonworks**

Current Stable Version : **1.1.1**

Delivery Guarantees : **at least once**

Out of Order Processing : **no**

Windowing : **no**

Back-pressure : **yes**

Latency : **configurable**

Resource Management : **native**

API : **REST (GUI)**

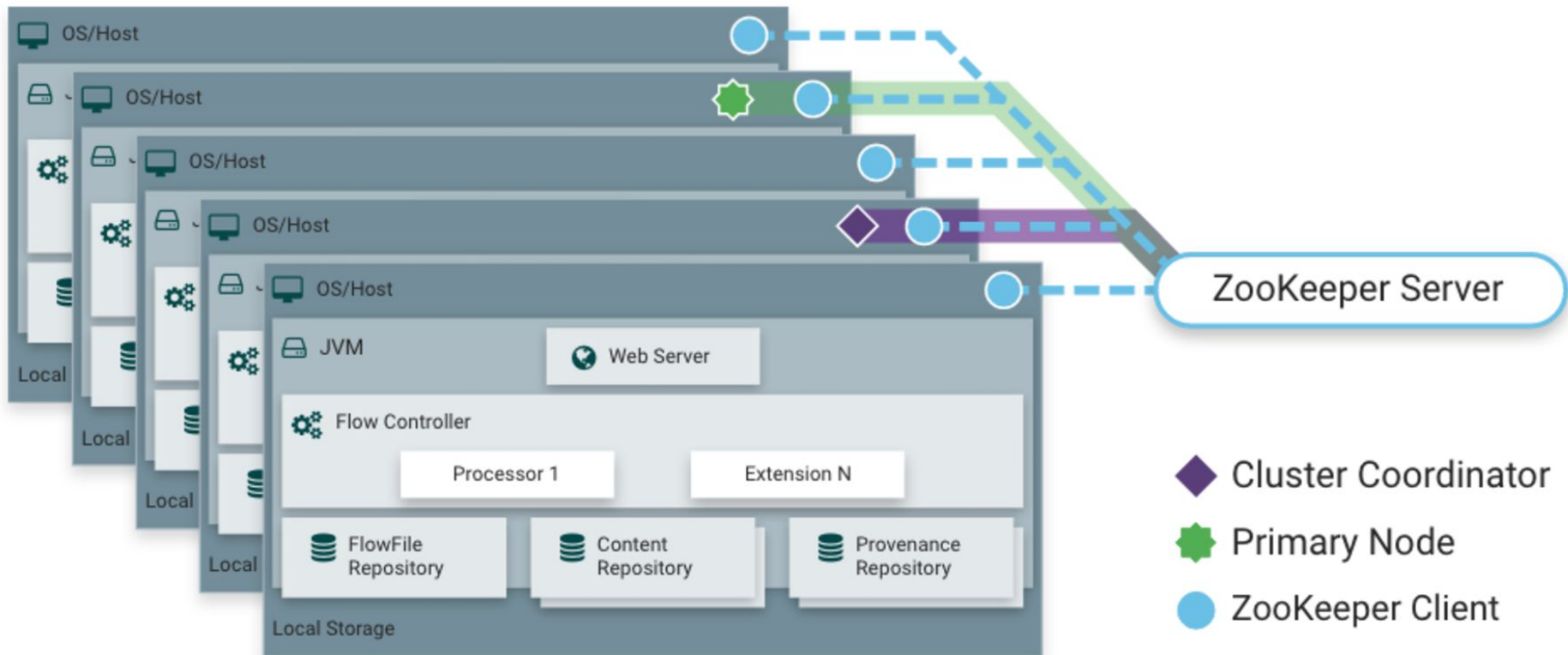
Ecosystem



Data Moving

Stream Processing

Architecture



Flow Files

Basic Abstraction

- **Pointer to content**
- **Content Attributes (key/value)**
- **Connection to provenance events**

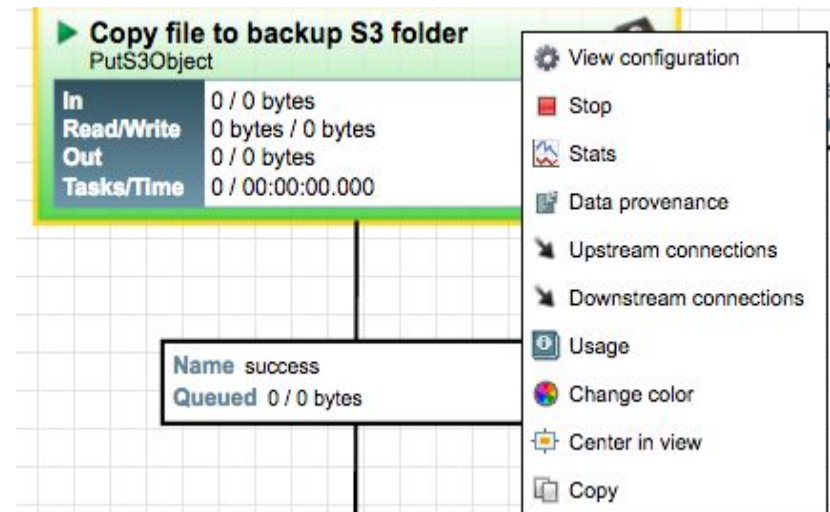
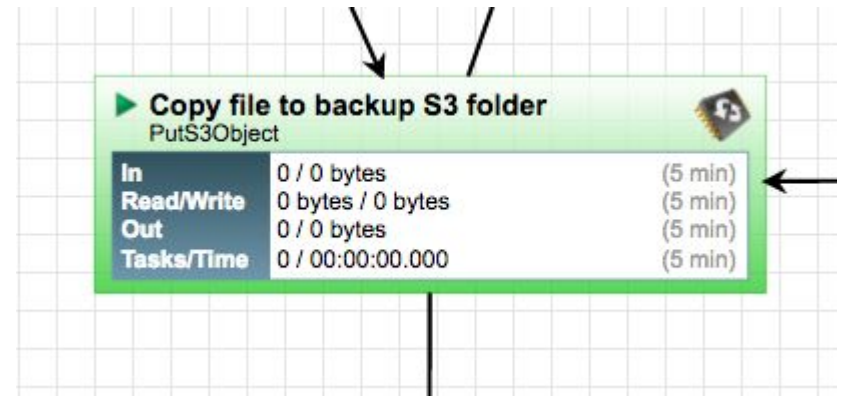
Repositories

- **FlowFile**
- **Content**
- **Provenance**
- **Immutable**
- **Copy-on-write**

Processor

Processors actually perform the work of data routing, transformation, or mediation between systems. Processors have access to attributes of a given FlowFile and its content stream.

Processors can operate on zero or more Flow Files in a given unit of work and either commit that work or rollback



Processor

- Basic Work Unit
- State
- Statistics
- Settings
- Input/Output
- Provenance
- Scheduling
- Logging (bulletins)

Processor Details

Settings | Scheduling | Properties | Comments

Name
Copy file to backup S3 folder

Id
e21f4eae-3532-3a41-99be-e44f880bd215

Type
PutS3Object

Penalty duration ?
30 sec

Yield duration ?
1 sec

Bulletin level ?
WARN

Auto terminate relationships ?
failure
FlowFiles are routed to failure relationship
success
FlowFiles are routed to success relationship

Processor Details

Settings | Scheduling | Properties | Comments

Scheduling strategy ?
Timer driven

Run duration ?
00:00:00.000

Concurrent tasks ?
1

Run schedule ?
0 sec

Processor Details

Settings | Scheduling | Properties | Comments

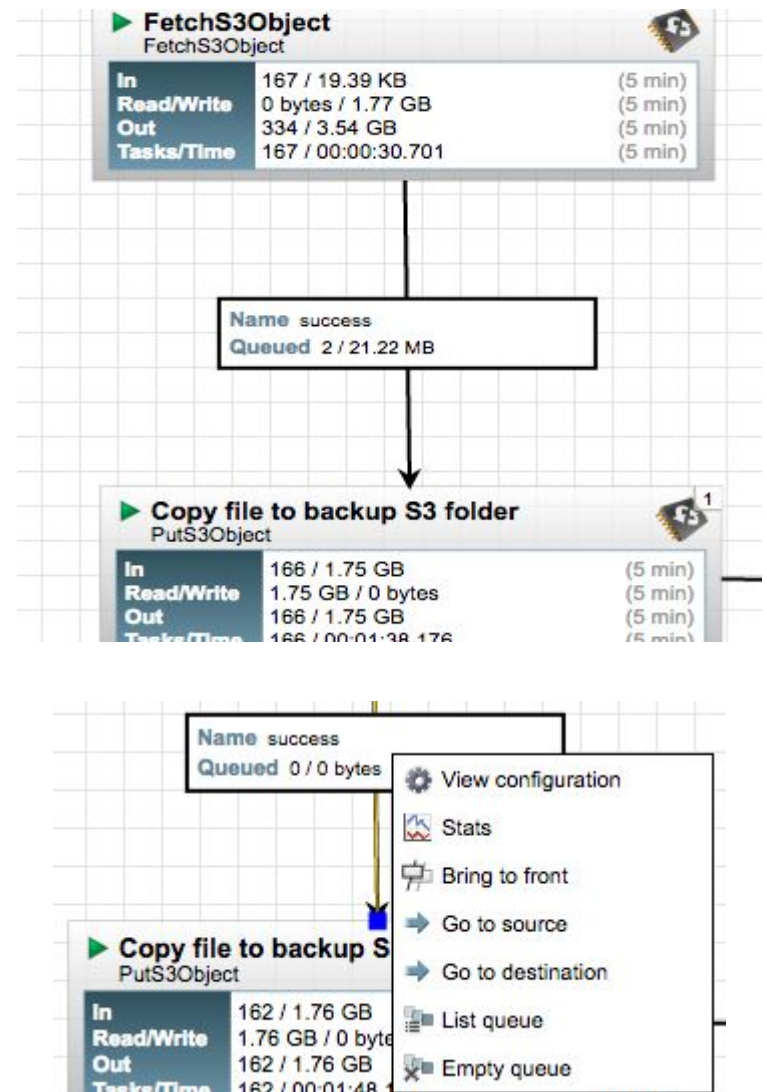
Required field

Property	Value
Object Key	`\${filename:replace("analytics-logs/kafka","analyt...
Bucket	vdb-prod-files
Access Key	Sensitive value set
Secret Key	Sensitive value set
Credentials File	No value set
AWS Credentials Provider service	No value set
Storage Class	Standard
Region	us-east-1
Communications Timeout	30 secs
Expiration Time Rule	No value set
FullControl User List	`\${c3.permissions.full_users}

Ok

Connection

Connections provide the actual linkage between processors. These act as queues and allow various processes to interact at differing rates. These queues can be prioritized dynamically and can have upper bounds on load, which enable back pressure



Connection

- Queue
- Statistics
- Settings
- Prioritization
- Details

Connection Details

Details Settings

From processor FetchS3Object FetchS3Object	To processor Copy file to backup S3 folder PutS3Object
Within group O2 Export S3 to Hadoop Test	Within group O2 Export S3 to Hadoop Test
Relationships ? failure success	

Connection Details

Details **Settings**

Name Empty string set	Prioritizers ? No value set
Id cec78641-bc05-3bd0-9e5c-147033d91c49	
FlowFile expiration ? 0 sec	
Back pressure object threshold ? 20	
Back pressure data size threshold ? 0 MB	

Available prioritizers ?

FirstInFirstOutPrioritizer

NewestFlowFileFirstPrioritizer

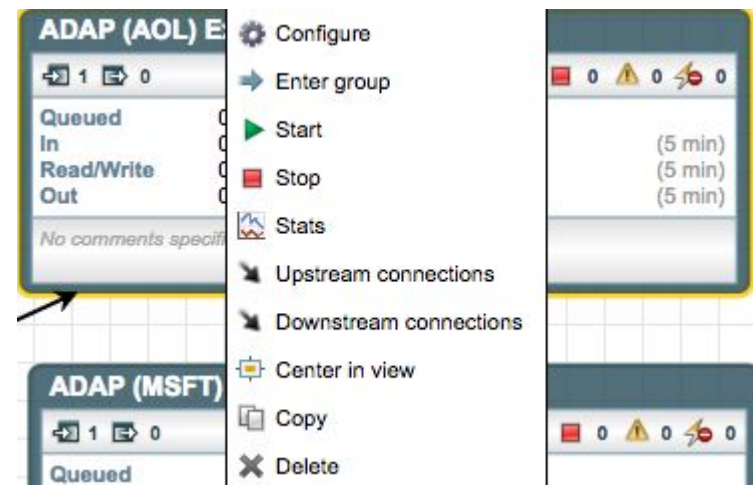
OldestFlowFileFirstPrioritizer

PriorityAttributePrioritizer

Selected prioritizers ?

Process Group







Specific set of processes and their connections, which can receive data via input ports and send data out via output ports. In this manner, process groups allow creation of entirely new components simply by composition of other components



Templates

Templates tend to be highly pattern oriented and while there are often many different ways to solve a problem, it helps greatly to be able to share those best practices. Templates allow subject matter experts to build and publish their flow designs and for others to benefit and collaborate on them

- **XML Based**
- **Reusable unit**
- **Versioning (versioning with Git)**

Data/Time ▾	Name	Description	
12/01/2016 21:31:20 UTC	ProductionPipeline-20161201	Empty string set	 X
11/17/2016 23:59:10 UTC	ProductionPipeline_20161117	Empty string set	 X
09/09/2016 04:29:26 UTC	General S3 - Kafka Processing	Empty string set	 X
09/09/2016 04:29:05 UTC	General S3 - Hadoop Processing	Empty string set	 X
09/09/2016 04:28:39 UTC	ADAP Processing	Empty string set	 X
09/09/2016 04:28:15 UTC	sFTP Processing	Empty string set	 X

Data Provenance

NiFi automatically records, indexes, and makes available provenance data as objects flow through the system even across fan-in, fan-out, transformations, and more. This information becomes extremely critical in supporting compliance, troubleshooting, optimization, and other scenarios

NiFi Flow Data Provenance

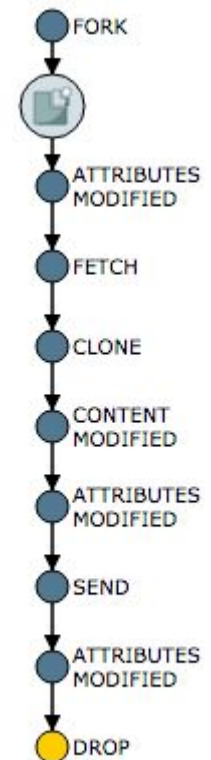
Oldest event available: 01/03/2017 17:41:41 UTC

Last updated: 17:42:16 UTC

Filter by component name
Displaying 1,000 of 1,000

Showing the most recent 1,000 of 4,353,925 events, please refine the search. Search

	Data/Time	Type	FlowFile Uuid	Size	Component Name	Component Type	Node	
1	01/04/2017 17:42:13.971 UTC	RECEIVE	c2164242-eeeb-4a19-aa41-1f2f6371756f	821 bytes	Get SQS QoE Pipeline	GetSQS	10.99.78.0:8080	🔍 ➔
2	01/04/2017 17:42:13.971 UTC	RECEIVE	9fa062f9-032b-42e0-b5f6-aefc8691ba91	820 bytes	Get SQS QoE Pipeline	GetSQS	10.99.78.0:8080	🔍 ➔
3	01/04/2017 17:42:13.626 UTC	DROP	ddeb0dc6-809e-4da0-8450-d73741a42...	26.59 MB	DeleteS3Object	DeleteS3Object	10.99.78.0:8080	🔍 ➔
4	01/04/2017 17:42:13.601 UTC	ATTRIBUTES_MODIFIED	ddeb0dc6-809e-4da0-8450-d73741a42...	26.59 MB	Copy file to backup S3 folder	PutS3Object	10.99.78.0:8080	🔍 ➔
5	01/04/2017 17:42:13.601 UTC	SEND	ddeb0dc6-809e-4da0-8450-d73741a42...	26.59 MB	Copy file to backup S3 folder	PutS3Object	10.99.78.0:8080	🔍 ➔
6	01/04/2017 17:42:13.222 UTC	DROP	280bdaa0-526d-4a8d-b85d-16aa421fc1...	551.9 KB	PutHDFS	PutHDFS	10.99.78.0:8080	🔍 ➔
7	01/04/2017 17:42:13.222 UTC	ATTRIBUTES_MODIFIED	280bdaa0-526d-4a8d-b85d-16aa421fc1...	551.9 KB	PutHDFS	PutHDFS	10.99.78.0:8080	🔍 ➔
8	01/04/2017 17:42:13.221 UTC	SEND	280bdaa0-526d-4a8d-b85d-16aa421fc1...	551.9 KB	PutHDFS	PutHDFS	10.99.78.0:8080	🔍 ➔
9	01/04/2017 17:42:13.091 UTC	DROP	b5f3d5ef-2671-45d4-b679-1ba0724bc9...	551.9 KB	DeleteS3Object	DeleteS3Object	10.99.78.0:8080	🔍 ➔
10	01/04/2017 17:42:13.071 UTC	SEND	b5f3d5ef-2671-45d4-b679-1ba0724bc9...	551.9 KB	Copy file to backup S3 folder	PutS3Object	10.99.78.0:8080	🔍 ➔
11	01/04/2017 17:42:13.071 UTC	ATTRIBUTES_MODIFIED	b5f3d5ef-2671-45d4-b679-1ba0724bc9...	551.9 KB	Copy file to backup S3 folder	PutS3Object	10.99.78.0:8080	🔍 ➔



Data Provenance

- Details
- Attributes
- Content

Provenance Event		
Details	Attributes	Content
Time	01/03/2017 19:41:18.576 UTC	
Event Duration	No value set	
Lineage Duration	00:00:00.296	
Type	DROP	
FlowFile Uuid	27fa0855-87a8-4876-a677-128bf59bf17f	
File Size	1.13 MB	
Component Id	0af51e00-2d36-3acd-b618-a4b626882836	
Component Name	PutHDFS	
Component Type	PutHDFS	
Node Address	10.99.84.132:8080	
Details	Auto-Terminated by success Relationship	
		Parent FlowFiles (0) No parents
		Child FlowFiles (0) No children

Provenance Event		
Details	Attributes	Content
Attribute Values Only show modified <input type="checkbox"/>		
absolute.hdfs.path	/raw_data/visible	
previous	No value set	
filename	analytics-logs-hadoop-Nginx-2017-01-03-19-2017-01-03-194101_Nginx-lua-prod-ap-southeas...	
filename.0	analytics-logs/hadoop/Nginx/2017-01/03/19/2017-01-03-194101_Nginx-lua-prod-ap-southeas...	
filename.1	analytics-logs/hadoop/Nginx/2017-01/03/19/2017-01-03-194101_Nginx-lua-prod-ap-southeas...	
hash.algorithm	md5	
hash.value	0f061bbe5a936cb1fd8573b97462d24	
mime.type	application/octet-stream	
path	./	
s3.bucket	vdb-prod-files	
s3.etag	2707d41a0f6c58b54abab23f1fd93506	
s3.sseAlgorithm	AES256	
s3.version	null	
sqs.ApproximateFirstReceiveTimesta...	1483472478285	
sqs.ApproximateReceiveCount	1	
sqs.message.id	659bc816-7727-4dc1-8bfe-57f6b0e68a16	
sqs.receipt.handle	AQEBL/R2xInsgsYDmzcqRKxW6CuwBND/3XcckBTkqM5OI7k26kbGitiUITBX4YKV7UpahpiTPV/Np...	
sqs.SenderId	AIDAJVEQ32BJMF27H2JKW	

Details	Attributes	Content
Input Claim		
Container	default	
Section	102	
Identifier	1483472478280-1698918	
Offset	936	
Size	1.13 MB	
Output Claim		
Container	default	
Section	102	
Identifier	1483472478280-1698918	
Offset	936	
Size	1.13 MB	
Replay		
Content is no longer available in Content Repository		

Controller Service

Controller Service allows developers to share functionality and state across the JVM in a clean and consistent manner

- **No scheduling**
- **No connections**
- **Used by Processors, Reporting Tasks, and other Controller Services**

Add Controller Service

Filter
Displaying 12 of 12

Type	Tags
AWSCredentialsProviderControllerService	credentials, provider, aws
CouchbaseClusterService	database, couchbase, connection, nosql
DBCPCConnectionPool	database, pooling, dbcp, jdbc, connection, store
DistributedMapCacheClientService	cluster, cache, distributed, state, map
DistributedMapCacheServer	cluster, server, cache, key/value, distributed, map
DistributedSetCacheClientService	cluster, cache, set, distributed, state
DistributedSetCacheServer	server, cache, set, distributed, distinct
HBase_1_1_2_ClientService	client, hbase
HiveConnectionPool	hive, database, pooling, dbcp, jdbc, connection, ...

AWSCredentialsProviderControllerService

Defines credentials for Amazon Web Services processors. Uses default credentials without configuration. Default credentials support EC2 instance profile/role, default user profile, environment variables, etc. Additional options include access key / secret key pairs, credentials file, named profile, and assume role credentials.

Available on **Node**

Cancel Add

Reporting Tasks

Provides a capability for reporting status, statistics, metrics, and monitoring information to external services

Add Reporting Task

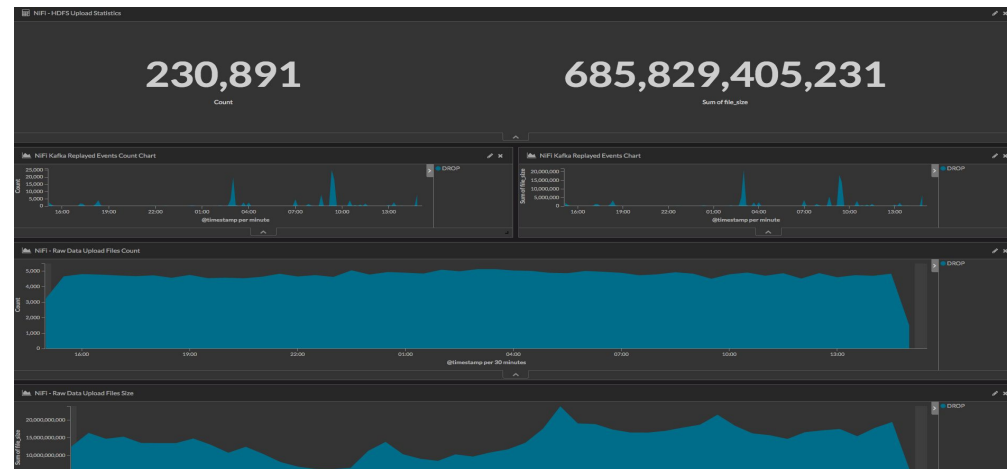
Tags: ambari disk elasticsearch ganglia garbage collection gc heap http jvm lineage log memory metrics monitor monitoring provenance repo reporting site site to site stats storage tracking warning

Filter
Displaying 8 of 8

Type	Tags
AmbariReportingTask	ambari, metrics, reporting
ControllerStatusReportingTask	stats, log
ElasticsearchProvenanceReporter	provenance, elasticsearch
HttpProvenanceReporter	provenance, http
MonitorDiskUsage	disk, repo, warning, storage, monitoring
MonitorMemory	jvm, memory, warning, monitor, heap, gc, garba...
SiteToSiteProvenanceReportingTask	lineage, site, provenance, tracking, site to site
StandardGangliaReporter	stats, ganglia

ElasticsearchProvenanceReporter
A provenance reporting task that writes to Elasticsearch

Available on:



- ElasticSearchProvenanceReporter and DataDogReportingTask

Extensibility

- Ready to use maven template
- Well defined interface for each component
- Classloader Isolation (.nar files)
- Great documentation for developers

Statistics

- **200+ built in Processors**
- **10+ built Control Services**
- **10+ built in Reporting Tasks**

Introduction Summary

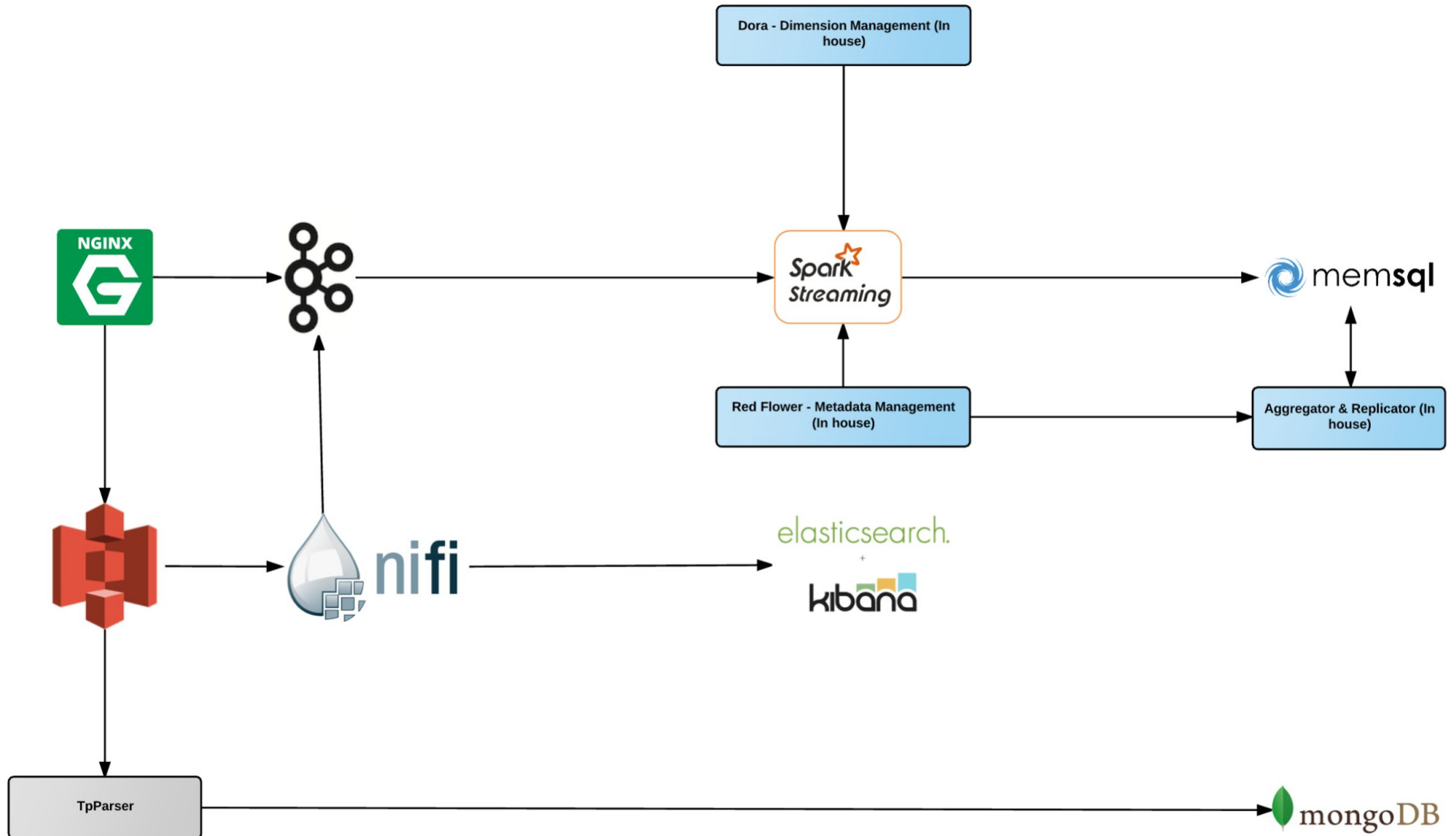
- **Processor**
- **Connection**
- **Processing Group**
- **Template**
- **Controller Service**
- **Reporting Task**

Our use case for NiFi

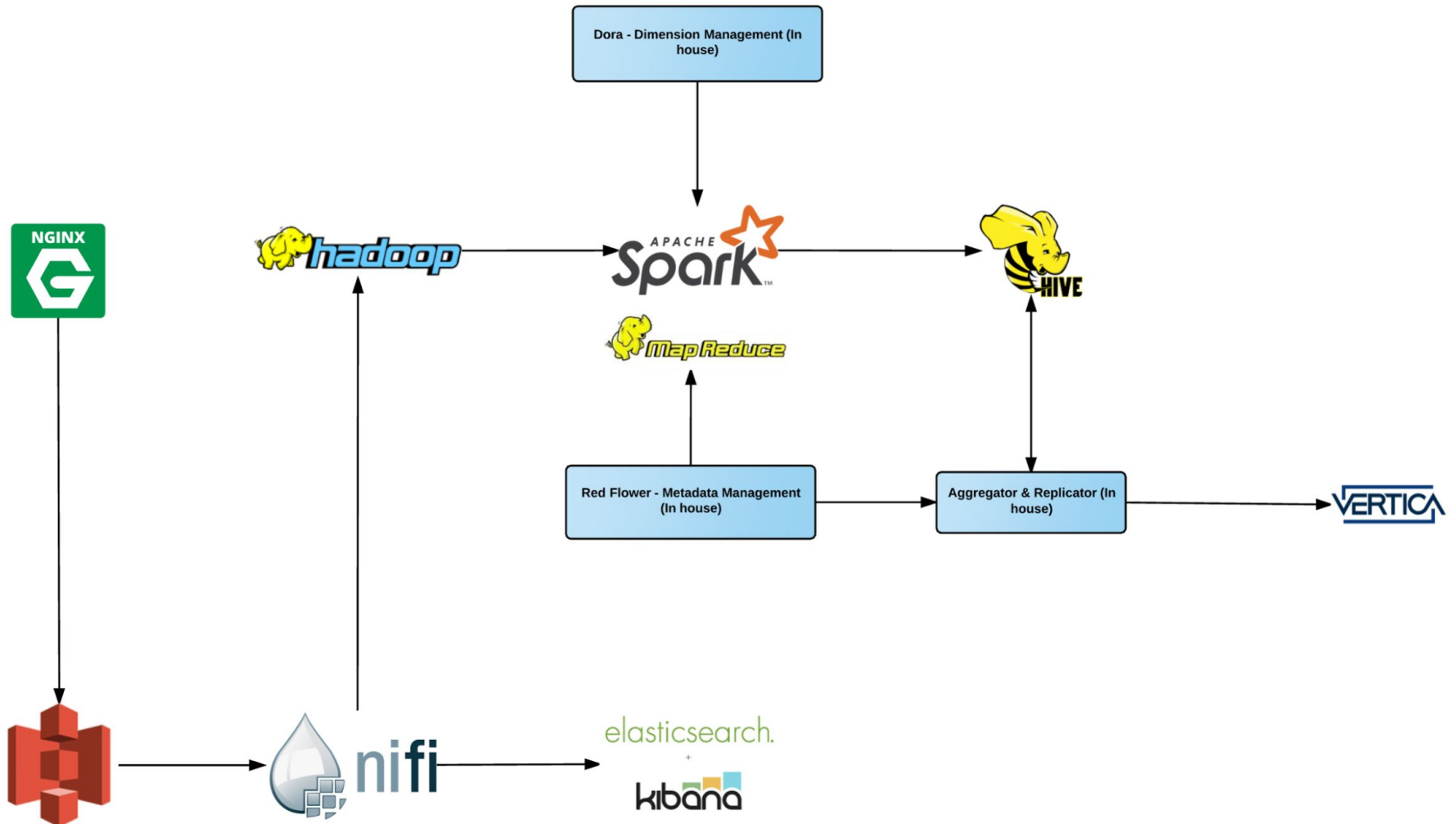
What was before

- Inhouse built file collector
- Footprint of 10 server
- Hard to manage, scale, extend

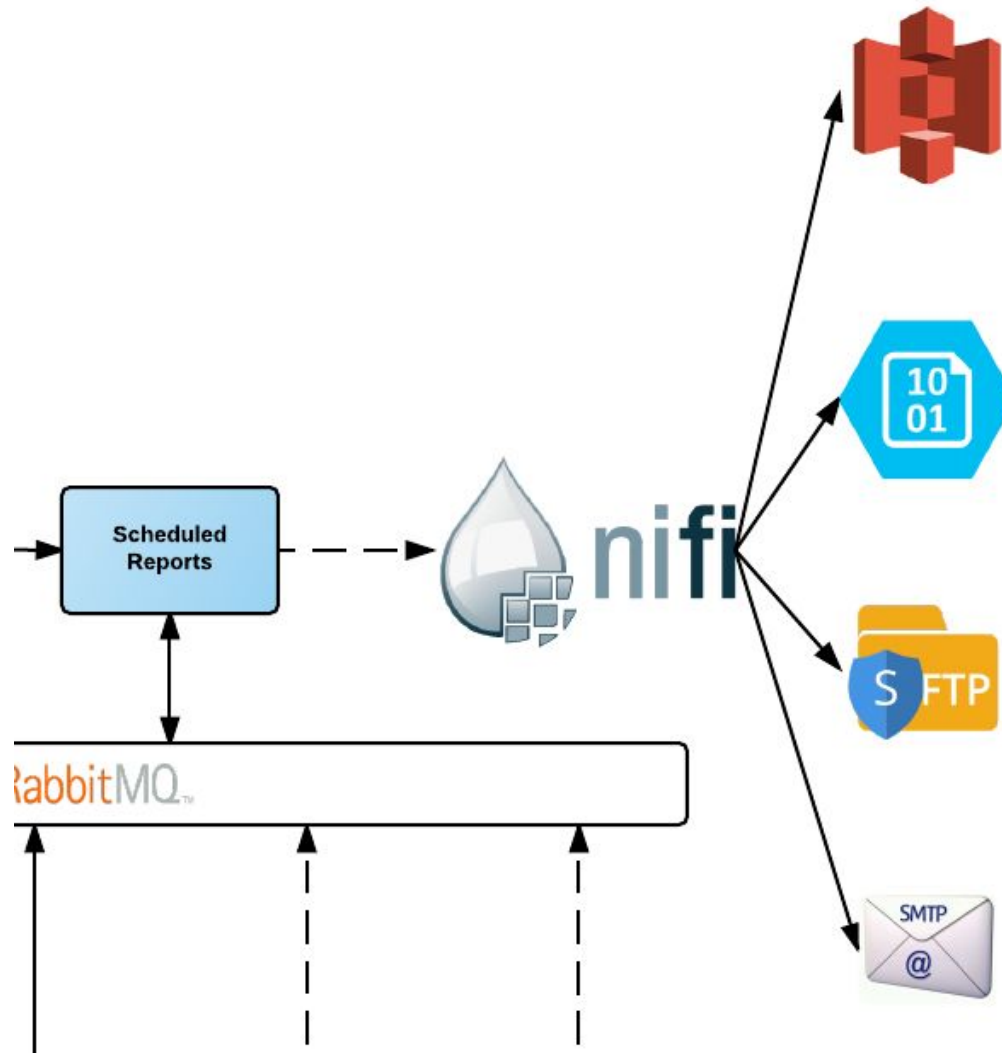
DWH Real Time



DWH Batch



Reports Distribution



Statistics

250K

Files Ingested Daily

30K

Files Exported Daily

20TB

Data Ingested Daily

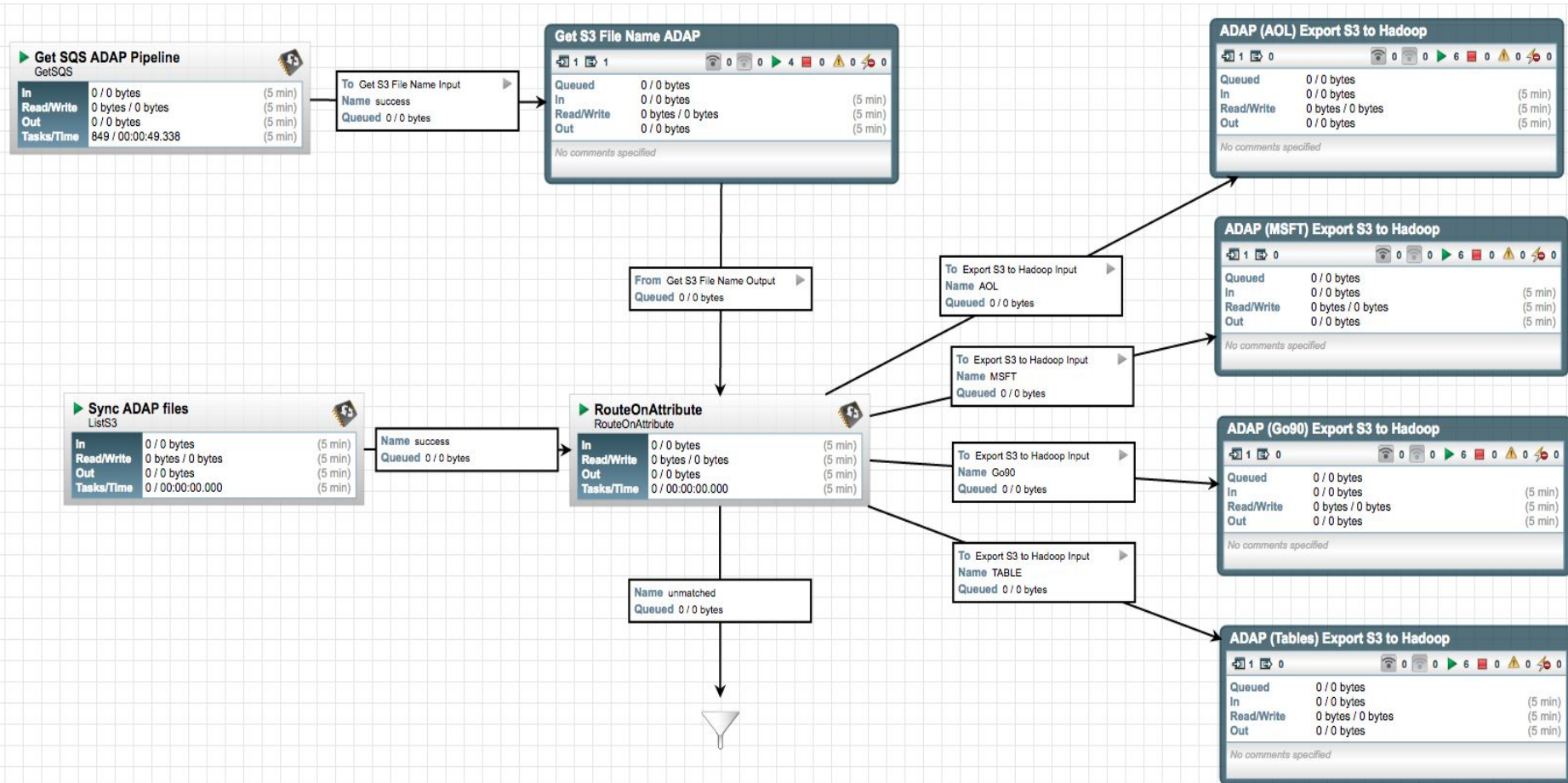
1 TB

Data Distributed Reports

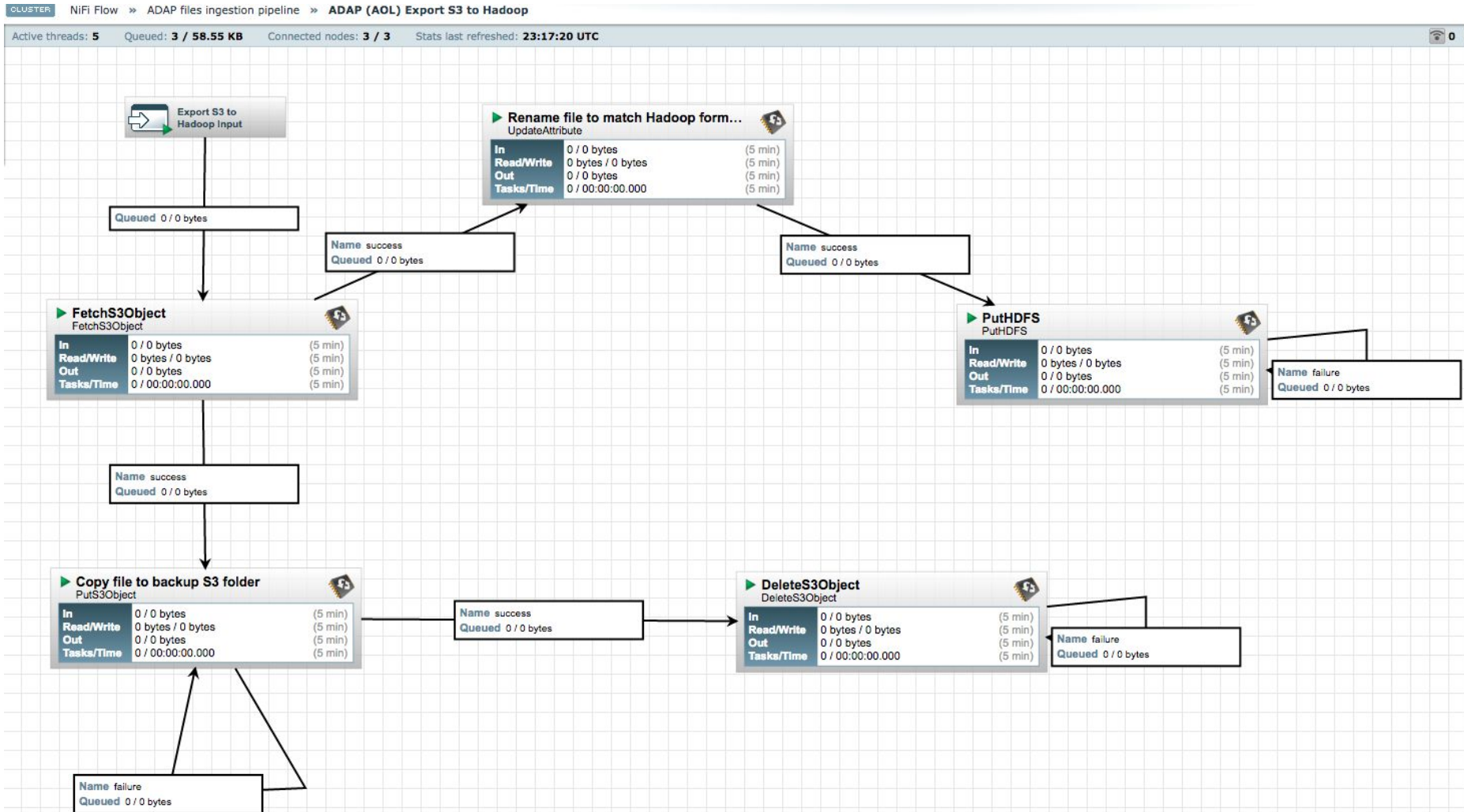
Near Real Time Data Availability

Minimum Interval :1 min

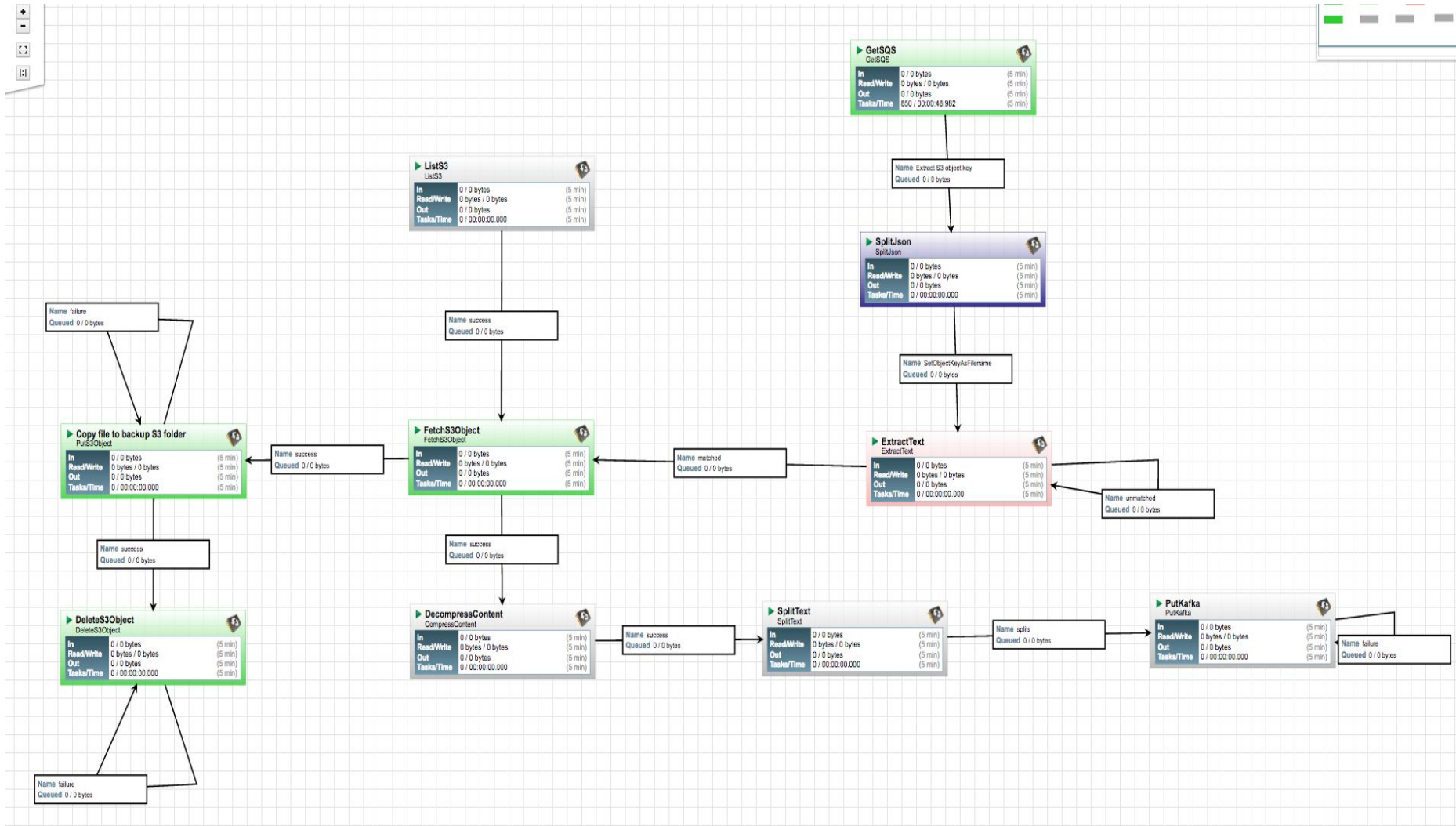
AWS - Hadoop Ingestion



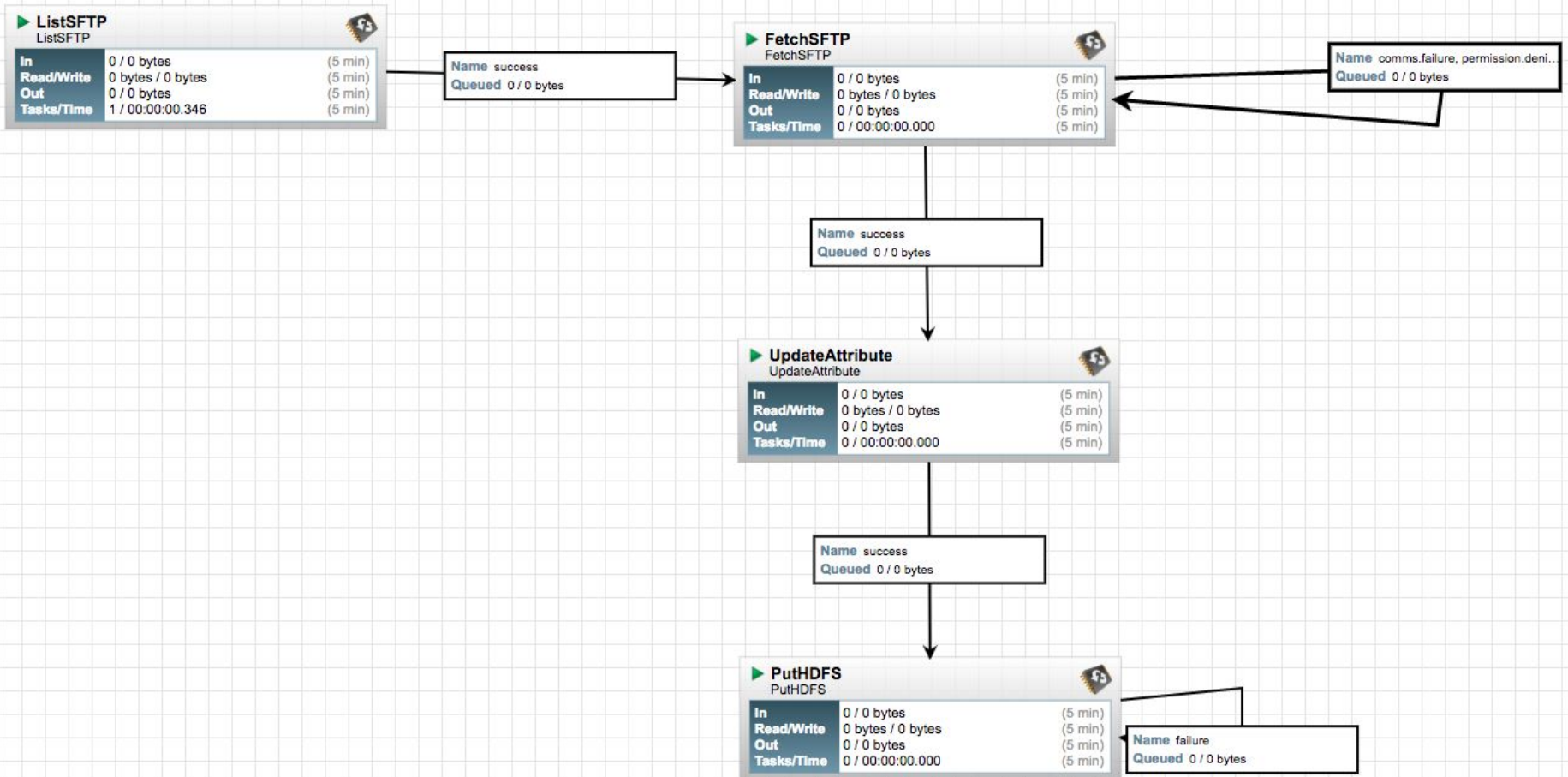
AWS - Hadoop Ingestion



Kafka Reprocessing



sFTP - HDFS Ingestion



**Let's break
something ;)**

Use Cases Summary

- **Web User Interface**
- **Configurable**
- **Scalable**
- **Easy to Manage**
- **Designed for Extension**

Q & A

THANK
YOU