HORTONWORKS®

NiFi Crash Course

Speaker name: Nathan Gough

Job title: Senior Software Engineer

Goals for today’s crash course:

- Understand some typical data problems
- Understand what NiFi is and how it solves those problems
- Understand the core NiFi concepts
- Hands on experience using NiFi

Before We Start

- What is your current experience with NiFi?
- Does everyone have the Hortonworks Sandbox set up on their machine?

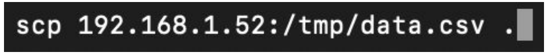
USBs available to copy these to your machine or:

https://www.cloudera.com/downloads/hortonworks-sandbox/hdf.html

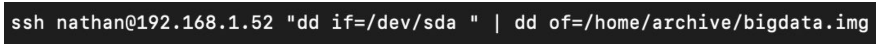
https://www.virtualbox.org/wiki/Downloads

Moving Data Is Easy

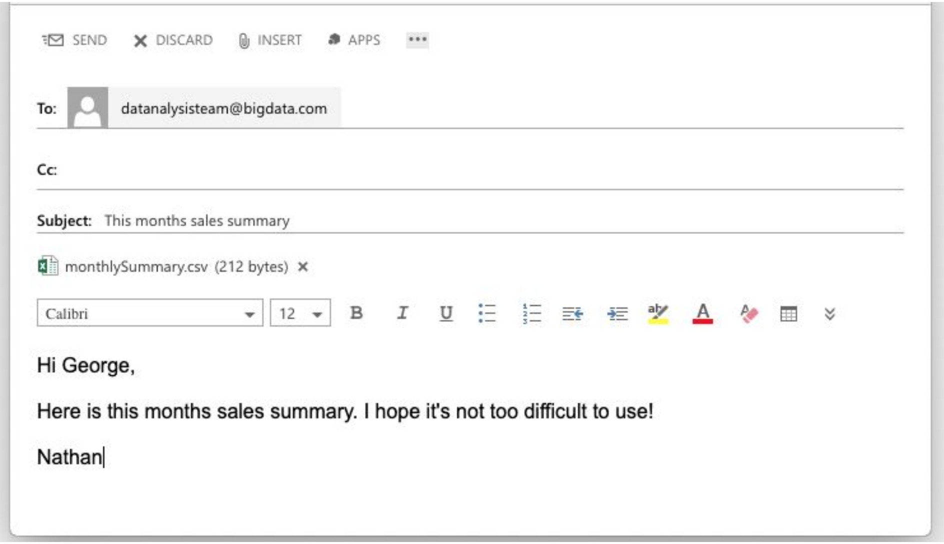
- Just send it, copy it, carry it



```
scp 192.168.1.52:/tmp/data.csv .
```



```
ssh nathan@192.168.1.52 "dd if=/dev/sda " | dd of=/home/archive/bigdata.img
```



To: datanalysisteam@bigdata.com

Cc:

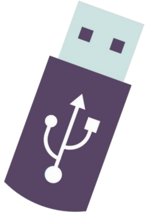
Subject: This months sales summary

monthlySummary.csv (212 bytes) ×

Hi George,

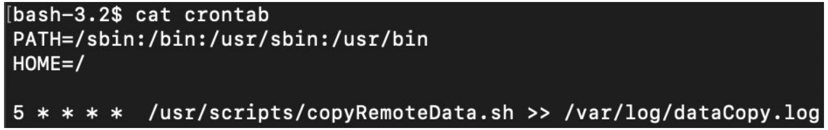
Here is this months sales summary. I hope it's not too difficult to use!

Nathan



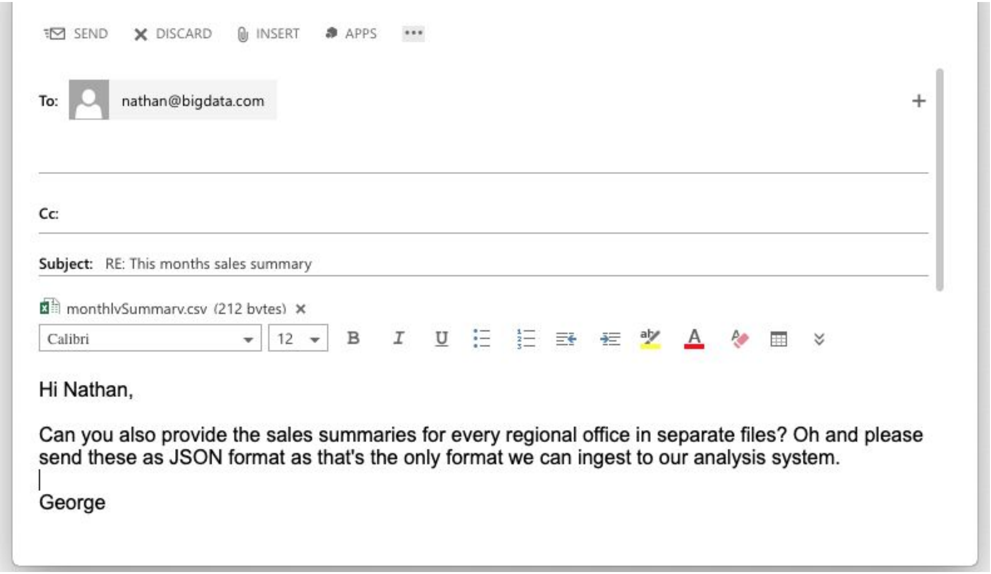
Moving Data Is Harder

- But people always ask for more



```
bash-3.2$ cat crontab
PATH=/sbin:/bin:/usr/sbin:/usr/bin
HOME=/

5 * * * *  /usr/scripts/copyRemoteData.sh >> /var/log/dataCopy.log
```



To: nathan@bigdata.com

Cc:

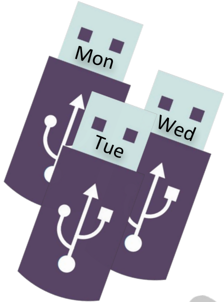
Subject: RE: This months sales summary

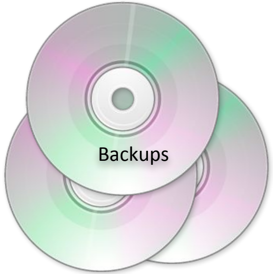
monthlySummary.csv (212 bytes) ×

Hi Nathan,

Can you also provide the sales summaries for every regional office in separate files? Oh and please send these as JSON format as that's the only format we can ingest to our analysis system.

George





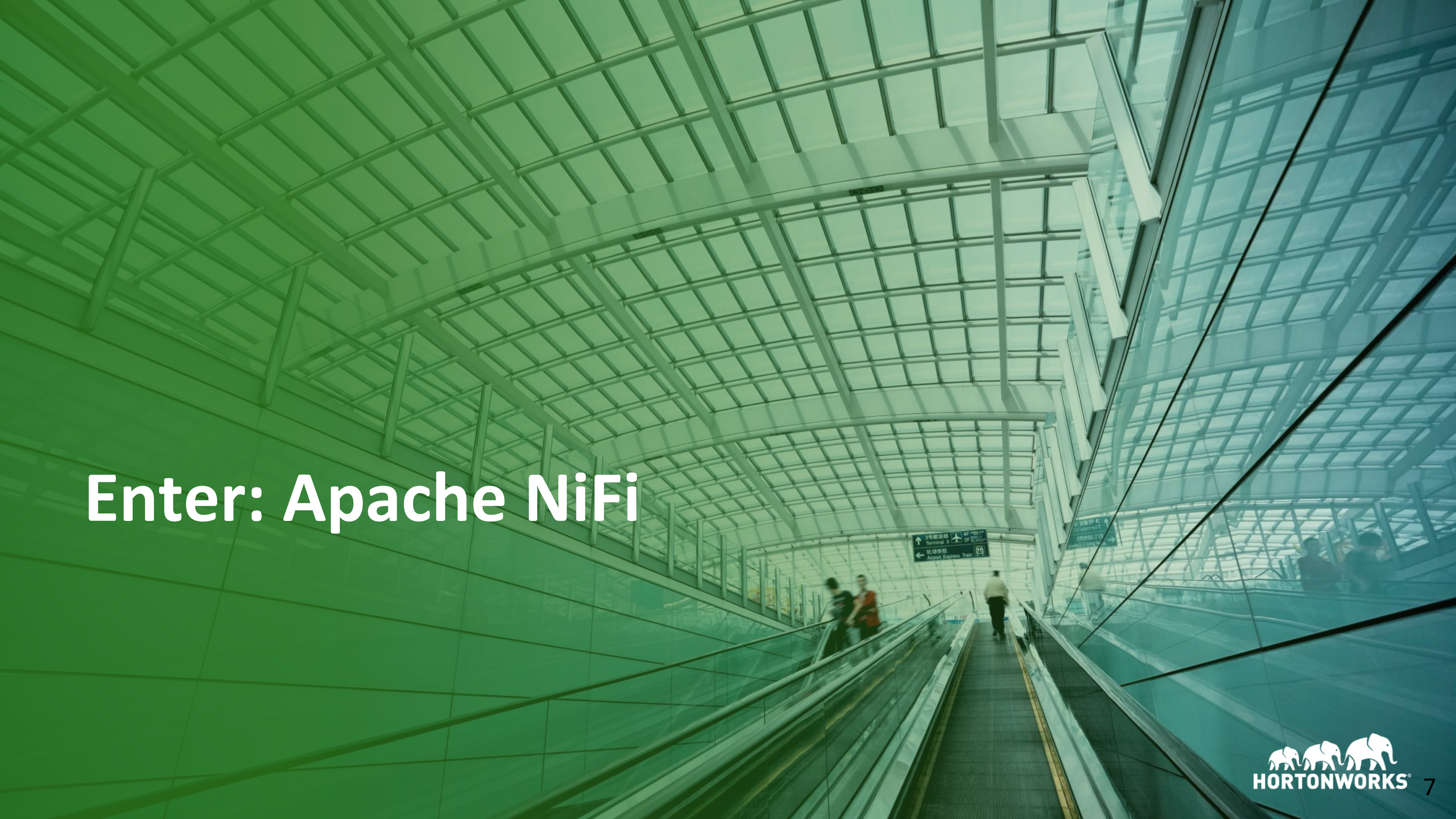
Moving Data Is Hard

- Many possible sources
- Many possible formats
- Many possible destinations

{ }
CSV
CSV
{ }
{ }
01110
10110
00110
01101
01110
10110
00110
01101

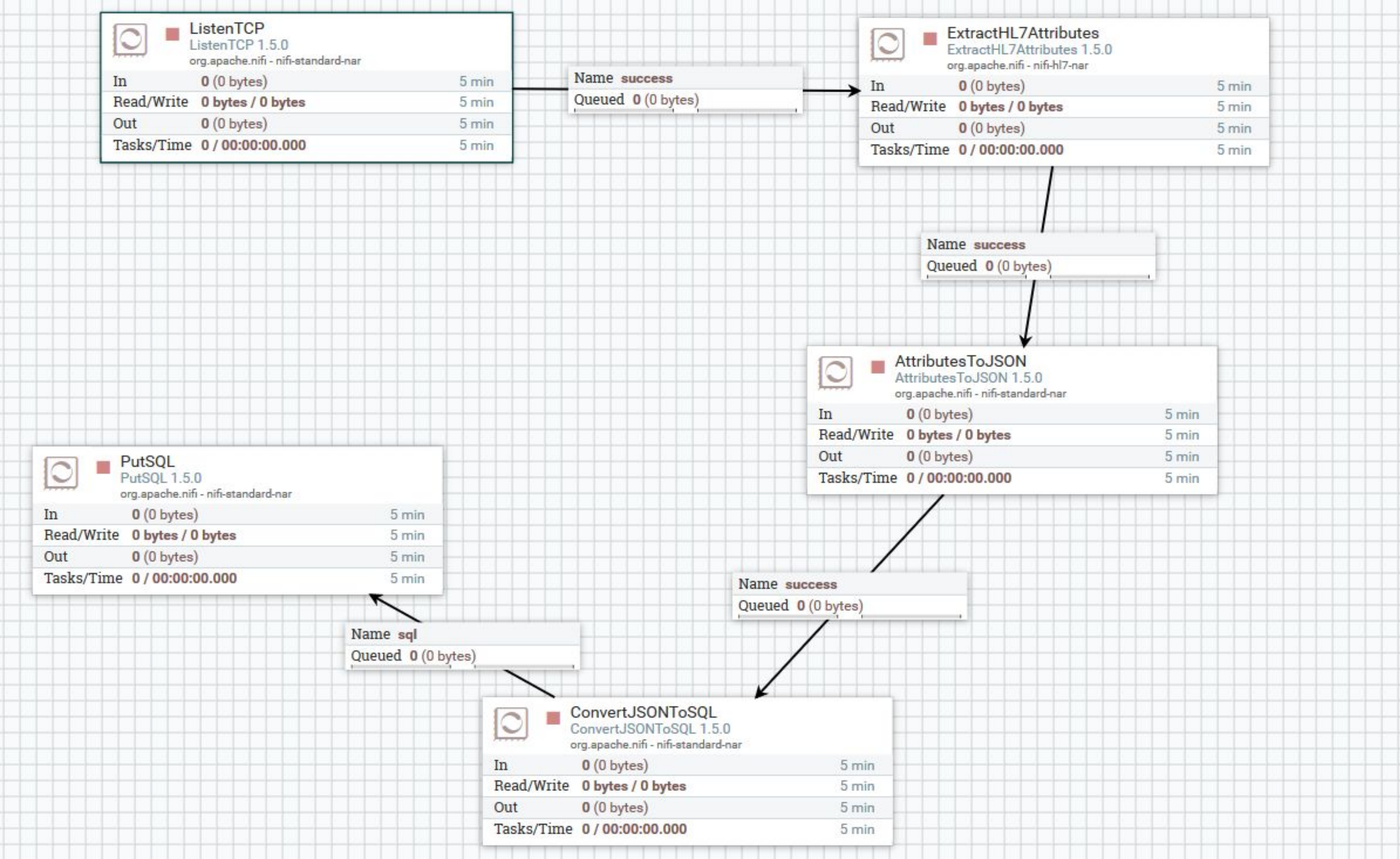
HORTONWORKS®

Enter: Apache NiFi



What is Apache NiFi?

- NiFi is a ‘flow based programming’ tool for collecting, transforming and distributing data between systems
- It gets data from A to B (or more likely A_1 .. A_n to B_1 .. B_n)
- Manipulates the data too



ListenTCP
ListenTCP 1.5.0
org.apache.nifi - nifi-standard-nar

In	**0** (0 bytes)	5 min
Read/Write	**0 bytes / 0 bytes**	5 min
Out	**0** (0 bytes)	5 min
Tasks/Time	**0 / 00:00:00.000**	5 min

Name **success**
Queued **0** (0 bytes)

ExtractHL7Attributes
ExtractHL7Attributes 1.5.0
org.apache.nifi - nifi-hl7-nar

In	**0** (0 bytes)	5 min
Read/Write	**0 bytes / 0 bytes**	5 min
Out	**0** (0 bytes)	5 min
Tasks/Time	**0 / 00:00:00.000**	5 min

Name **success**
Queued **0** (0 bytes)

AttributesToJSON
AttributesToJSON 1.5.0
org.apache.nifi - nifi-standard-nar

In	**0** (0 bytes)	5 min
Read/Write	**0 bytes / 0 bytes**	5 min
Out	**0** (0 bytes)	5 min
Tasks/Time	**0 / 00:00:00.000**	5 min

Name **success**
Queued **0** (0 bytes)

ConvertJSONToSQL
ConvertJSONToSQL 1.5.0
org.apache.nifi - nifi-standard-nar

In	**0** (0 bytes)	5 min
Read/Write	**0 bytes / 0 bytes**	5 min
Out	**0** (0 bytes)	5 min
Tasks/Time	**0 / 00:00:00.000**	5 min

Name **sql**
Queued **0** (0 bytes)

PutSQL
PutSQL 1.5.0
org.apache.nifi - nifi-standard-nar

In	**0** (0 bytes)	5 min
Read/Write	**0 bytes / 0 bytes**	5 min
Out	**0** (0 bytes)	5 min
Tasks/Time	**0 / 00:00:00.000**	5 min

Why is it useful?

- NiFi can send and receive data using a range of methods
- NiFi can handle any data format (image/video, JSON, XML, binary, etc)
- NiFi is easy to configure, used by technical and non-technical people
- NiFi is efficient and allows high throughput
- NiFi is reliable and protects against data loss

Integration with other systems

- FTP
- SFTP
- HL7
- UDP
- XML
- ⋮
- HTTP
- WebSocket
- Email
- HTML
- Image
- Syslog
- AMQP

AVRO
kafka
HIVE
elastic
FLUME
Spark
Solr
EVENT HUBS
STORM
hadoop
SQL
Couchbase
amazon web services S3
Flink
cassandra
minifi
slack
DynamoDB
amazon web services
splunk>
mongoDB
amazon Kinesis

Hash	Encrypt	GeoEnrich
Merge	Tail	Scan
Extract	Evaluate	Replace
Duplicate	Execute	Translate
Split	Fetch	Convert

⋮

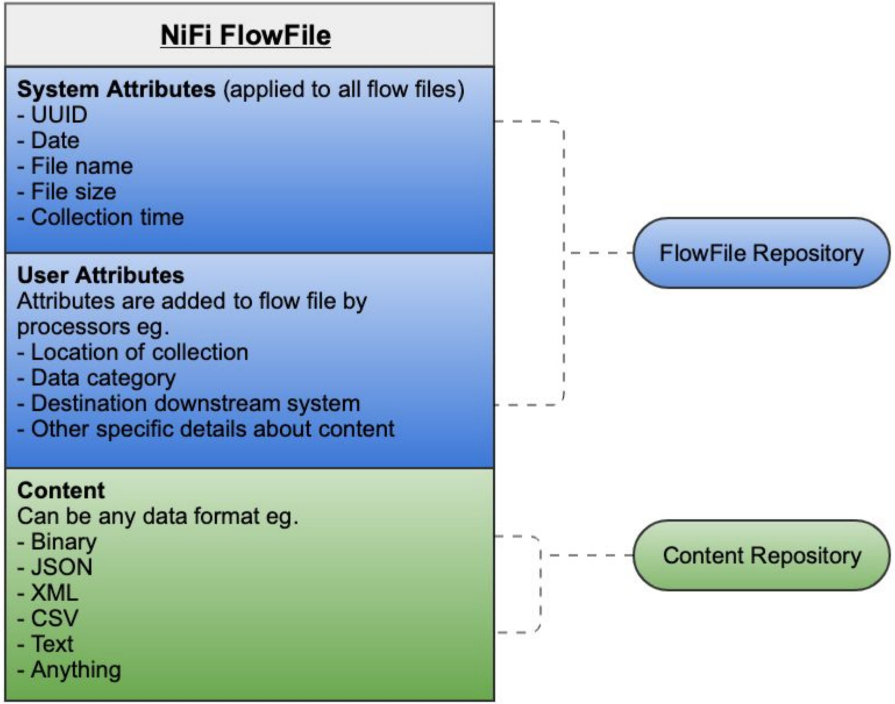
Parse Records	Convert Records
Route Text	Distribute Load
Route Content	Generate Table Fetch
Route Context	Jolt Transform JSON
Control Rate	Prioritized Delivery

Slide borrowed from Andy LoPresto's NiFi Crash Course

NiFi Core Concepts

NiFi FlowFile

- A data object flowing through NiFI
- FlowFile attributes **describe** the Content
- The attributes (key/value pairs) are passed through the NiFi ‘flow’ and exist in JVM memory
- Content (file bytes) is stored on disk and accessed only as required



NiFi FlowFile

System Attributes (applied to all flow files)
- UUID
- Date
- File name
- File size
- Collection time

User Attributes
Attributes are added to flow file by processors eg.
- Location of collection
- Data category
- Destination downstream system
- Other specific details about content

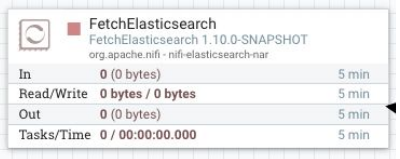
Content
Can be any data format eg.
- Binary
- JSON
- XML
- CSV
- Text
- Anything

FlowFile Repository

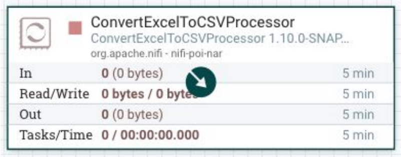
Content Repository

NiFi Processor

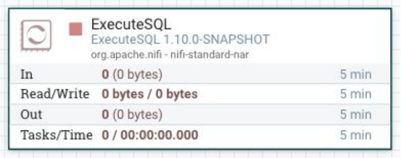
- An operation that can be applied to a FlowFile
- Operations for flowfile attributes and content
- Over 290 processors currently available
- Can be grouped to create new ‘black box’ processing
- New processors can be written using Java



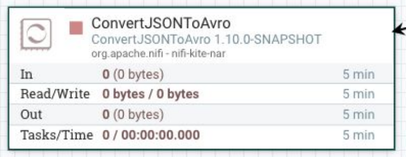
```
FetchElasticsearch
FetchElasticsearch 1.10.0-SNAPSHOT
org.apache.nifi - nifi-elasticsearch-nar
In          0 (0 bytes)              5 min
Read/Write  0 bytes / 0 bytes        5 min
Out         0 (0 bytes)              5 min
Tasks/Time  0 / 00:00:00.000         5 min
```



```
ConvertExcelToCSVProcessor
ConvertExcelToCSVProcessor 1.10.0-SNAP...
org.apache.nifi - nifi-poi-nar
In          0 (0 bytes)              5 min
Read/Write  0 bytes / 0 bytes        5 min
Out         0 (0 bytes)              5 min
Tasks/Time  0 / 00:00:00.000         5 min
```



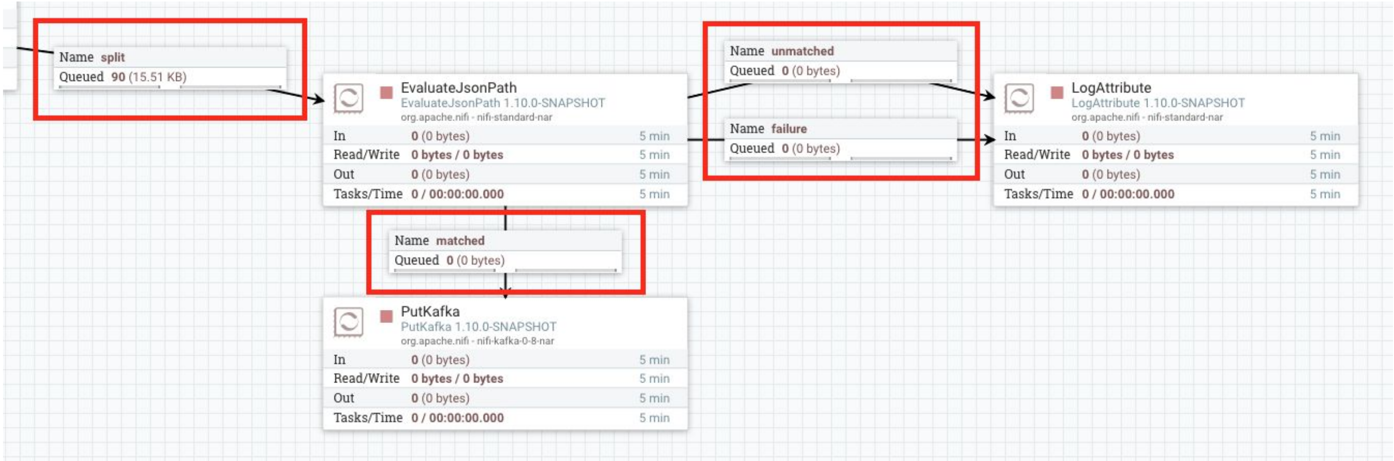
```
ExecuteSQL
ExecuteSQL 1.10.0-SNAPSHOT
org.apache.nifi - nifi-standard-nar
In          0 (0 bytes)              5 min
Read/Write  0 bytes / 0 bytes        5 min
Out         0 (0 bytes)              5 min
Tasks/Time  0 / 00:00:00.000         5 min
```



```
ConvertJSONToAvro
ConvertJSONToAvro 1.10.0-SNAPSHOT
org.apache.nifi - nifi-kite-nar
In          0 (0 bytes)              5 min
Read/Write  0 bytes / 0 bytes        5 min
Out         0 (0 bytes)              5 min
Tasks/Time  0 / 00:00:00.000         5 min
```

Processor Connections

- Direct the flow of flowfiles between processors
- Operates like a queue



Name split
Queued 90 (15.51 KB)

Name unmatched
Queued 0 (0 bytes)

Name failure
Queued 0 (0 bytes)

EvaluateJsonPath
EvaluateJsonPath 1.10.0-SNAPSHOT
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

LogAttribute
LogAttribute 1.10.0-SNAPSHOT
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name matched
Queued 0 (0 bytes)

PutKafka
PutKafka 1.10.0-SNAPSHOT
org.apache.nifi - nifi-kafka-0-8-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

NiFi Flow Controller

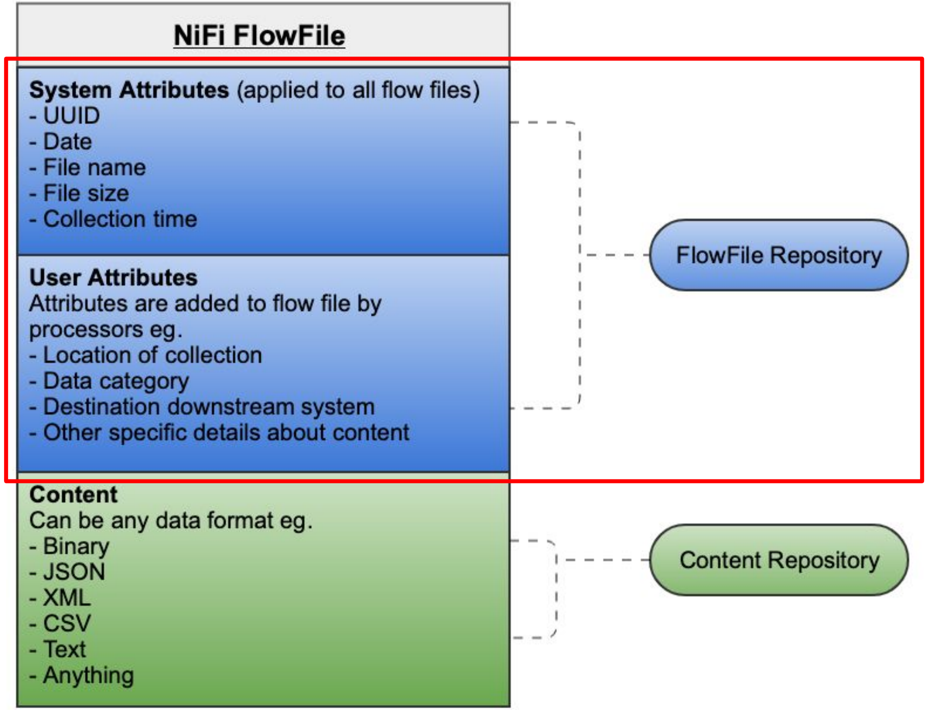
- The NiFi ‘brain’
- Maintains the knowledge of how processors are connected
- Shares and allocates execution time between processors
- Facilitates the exchange of files between processors across connections

NiFi Controller Services

- Common services usable by processors
- Examples include:
  - SSL Context Service
  - Distributed Map Cache
  - Record Readers/Writers

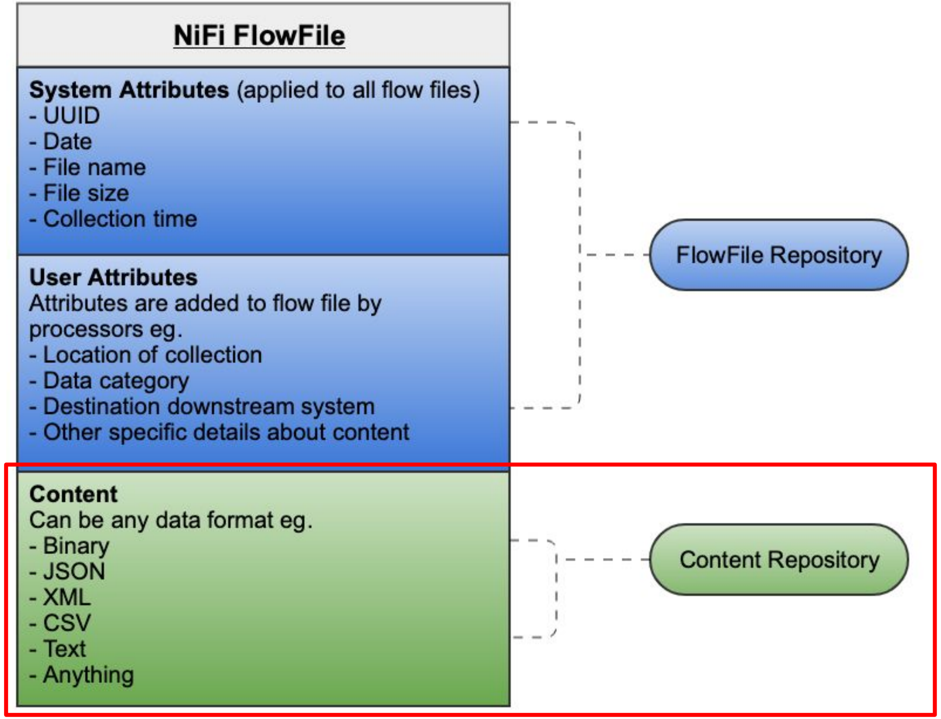
FlowFile Repository

- Maintains the flowfile attributes as key/value pairs in JVM memory (fast)
- Stores flowfile location in flow
- Pointer to flowfile's content
- Uses a write-ahead log on disk for data resilience



Content Repository

- Stores the file content (raw bytes) on disk
- Copy-on-write, read only as required
- Content is managed by NiFi
- Follow best practices guide for setup



NiFi FlowFile

System Attributes (applied to all flow files)
- UUID
- Date
- File name
- File size
- Collection time

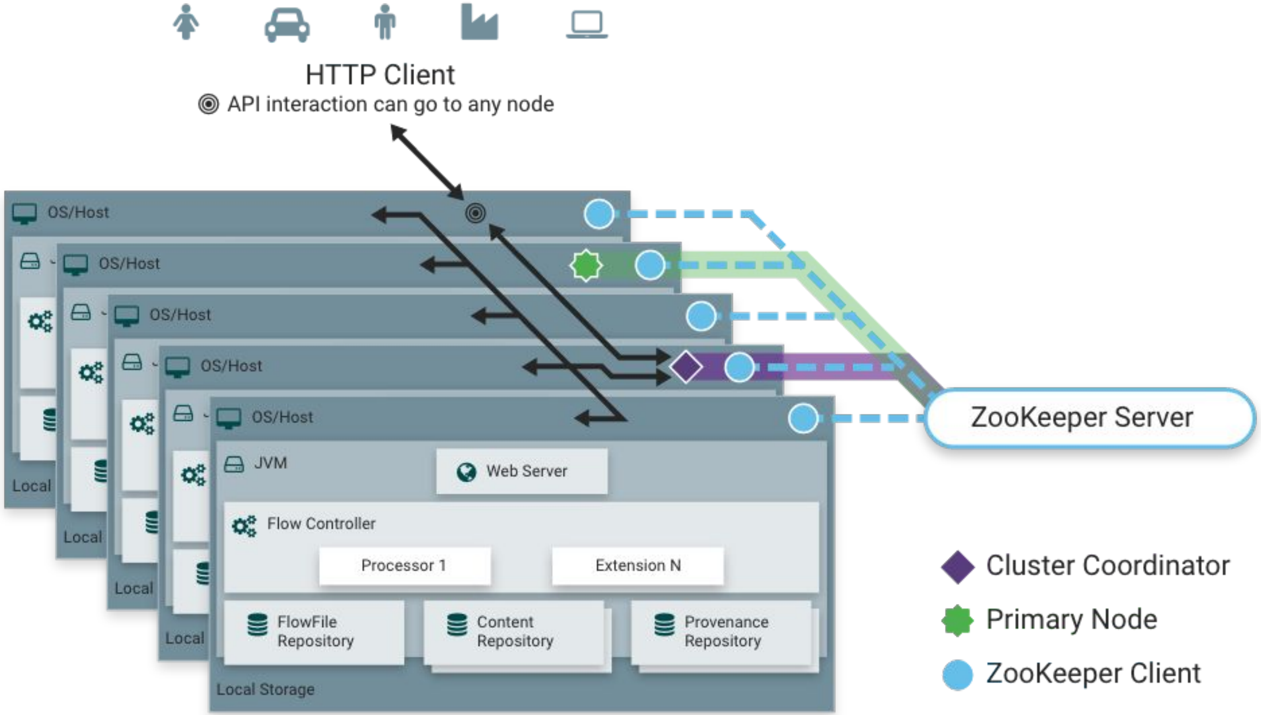
User Attributes
Attributes are added to flow file by processors eg.
- Location of collection
- Data category
- Destination downstream system
- Other specific details about content

Content
Can be any data format eg.
- Binary
- JSON
- XML
- CSV
- Text
- Anything

Provenance Repository

- Stores the history/lineage of flowfiles that transit a NiFi flow
- Details what processors a flowfile has passed through and what changes occurred
- Useful for troubleshooting
- Potentially large data requirements over time

Clustering


HTTP Client
API interaction can go to any node
OS/Host
OS/Host
OS/Host
OS/Host
OS/Host
JVM
Web Server
Flow Controller
Processor 1
Extension N
FlowFile Repository
Content Repository
Provenance Repository
Local
Local
Local
Local
Local Storage
ZooKeeper Server
Cluster Coordinator
Primary Node
ZooKeeper Client

Clustering

- ZooKeeper server manages state of cluster
- ZooKeeper will elect the cluster coordinator and primary node
- Cluster coordinator handles:
  - ‘Source of truth’ for the flow.xml, replicates to other nodes
  - Node connection/disconnection
- Primary node can be used to execute tasks on a single node

Security in NiFi

Security in NiFi

- The NiFi UI, API and individual processors can (and should) be secured with TLS
  - NiFi provides a toolkit which can be used to generate server (and client certificates) with required fields
- Encryption/Decryption Processor
- Encrypted Provenance Repository implementation

User authentication and authorisation

- User authentication methods
  - X.509 Certificates
  - LDAP
  - Apache Knox
  - Kerberos
  - OpenID Connect
- User authorisation
  - Configurable within NiFi, fine grain control
  - Apache Ranger

Apache NiFi Project

Extending and integrating with NiFi

- NiFi is open source under Apache 2.0 License
  - Commits through GitHub: https://github.com/apache/nifi
  - Submit features/bug requests: https://issues.apache.org/jira/projects/NIFI/issues
- Write your own processors (Java)
- Write your own controller services (Java)
- Execute script processor
  - Ruby, Python, Groovy, Lua, others
- REST API and NiFi CLI

Related Apache Products

- MiNiFi
  - Small collection agent, focused on data collection at the source (edge nodes)
  - Java (50MB) or C++ (3.2MB) versions
- NiFi Registry
  - Like Git for NiFi - version control of data flows
  - Promote flows through Dev, QA and Prod environments
  - Now stores versioned processor (NARs) bundles
- Schema Registry
  - Store and shares data schemas

Some URLs

Demo code can be found at:

https://github.com/thenatog/nificrashcourse-2019

For NiFi help, email this mailing list:

users@nifi.apache.org

Trucking Demo:
https://hortonworks.com/tutorial/nifi-in-trucking-iot-on-hdf/section/3/

Questions?

Special thanks to Andy LoPresto
Special thanks to Per Liedman for the Leaflet Realtime plugin and example code

Demo

Image Sources

USB Icon Image, no changes under Creative Commons:
https://icon-icons.com/icon/pen-pendrive-usb/73780
Compact Disc Image, no changes under Creative Commons:
https://icon-icons.com/icon/compact-disc-disk-cd-dvd/1080
Nike Dunk Shoes Image, no changes, free for commercial use:
https://icon-icons.com/icon/nike-dunk-shoes-shoes-sports/55308
World Planet Earth
https://openclipart.org/detail/205429/world-planet-earth
Store Icon
https://icon-icons.com/icon/store-market/54371#64
Factory Icon
https://icon-icons.com/icon/factory/99134#64
Office Building
https://openclipart.org/detail/216806/office-building
Text Hex
https://icon-icons.com/icon/text-hex/92814#64
CSV
https://icon-icons.com/icon/csv/3633#48

USB Icon Image, no changes under Creative Commons:
https://icon-icons.com/icon/pen-pendrive-usb/73780
Home Server Icon
https://icon-icons.com/icon/home-server-computer-database/55232#64
Cargo Ship
https://icon-icons.com/icon/cargo-freighter-ship-boat-transport/54882#72
JSON Icon
https://icon-icons.com/icon/application-json/92733#48
Plane Icon
https://icon-icons.com/icon/airplane-plane-grey-transport-vehicle-vehicles/54909#64
Database Icon
https://icon-icons.com/icon/database-data/19664
Apache NiFi Example Flow
https://i.stack.imgur.com/I6bVb.png

HORTONWORKS®