Drag and Drop: Open Source GeoTools ETL with Apache NiFi

Andrew Hulbert – CCRI
Constantin Stanca – Hortonworks

5/16/2018

Andrew Hulbert – Principal Software Engineer @CCRI, GeoMesa Contributor: andrew.hulbert@ccri.com

Cpnstantin Stanca – Solutions Engineer @Hortonworks, Data Transformation and Analytics SME: cstanca@hortonworks.com

Executive Summary

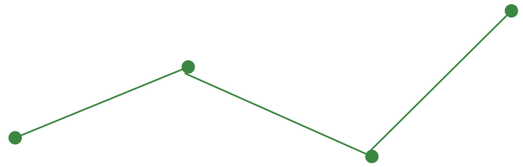
- Loading your GeoTools SimpleFeatures into your cloud or database never seems as easy as it should be. There's Twitter streams, FTP servers, inboxes, dropboxes, and all sorts of other data that you need to parse, convert to SimpleFeatures, and then ingest into your GeoTools datastore.
- GeoMesa and NiFi can provide a fully open source solution to ease the pain of ingesting data into ANY GeoTools data store. It's as easy as drag and drop! Literally!
- We'll show how real-time streaming data such as satellite AIS can be ingested and managed in real-time using ingest pipelines with your web browser without having to compile any code.

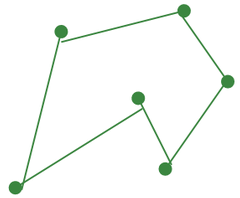
Challenges

- Ancillary:
  - Lots of data stores, tools and frameworks
  - Moving data between systems (data tracking and data loss) – convoluted data flow
  - Configuration and Monitoring
- Big Data:
  - Volume, Variety, Velocity
  - More data stores, tools and frameworks
- Spatial Data Types

Spatial Data Types

Points	Locations Events Instantaneous Positions
Lines	Road networks Voyages Trips Trajectories
Polygons	Administrative Regions Airspaces





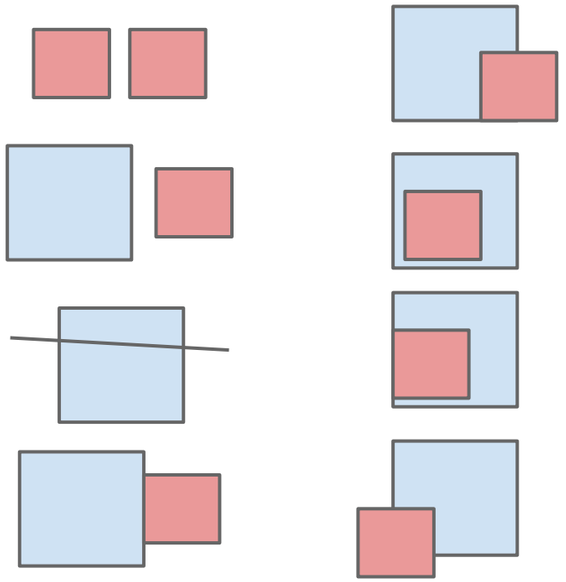
Spatial Data Relationships

equals

disjoint

intersects

touches



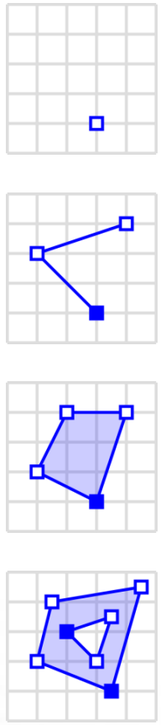
crosses

within

contains

overlaps

“Traditional” Geo-Spatial ETL



OGC
Simple Features

OGC™
Open Geospatial Consortium, Inc.

GeoTools

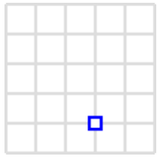
GeoServer

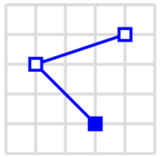
PostGIS
Spatial PostgreSQL

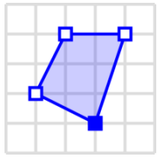
CCRi
DATA TO KNOWLEDGE

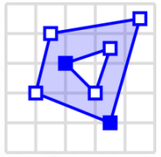
HORTONWORKS

”Cloud” Geo-Spatial ETL







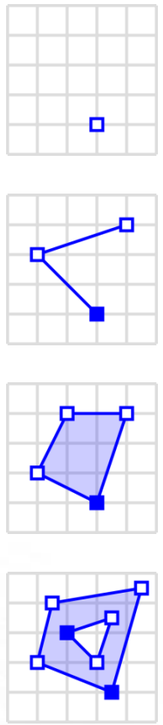


OGC
Simple Features

OGC™
Open Geospatial Consortium, Inc.

?

Solution: Apache NiFi and GeoMesa



OGC
Simple Features

OGC™
Open Geospatial Consortium, Inc.

APACHE nifi

geomesa
open source through
LocationTech and eclipse

hadoop

APACHE Spark™

jupyter

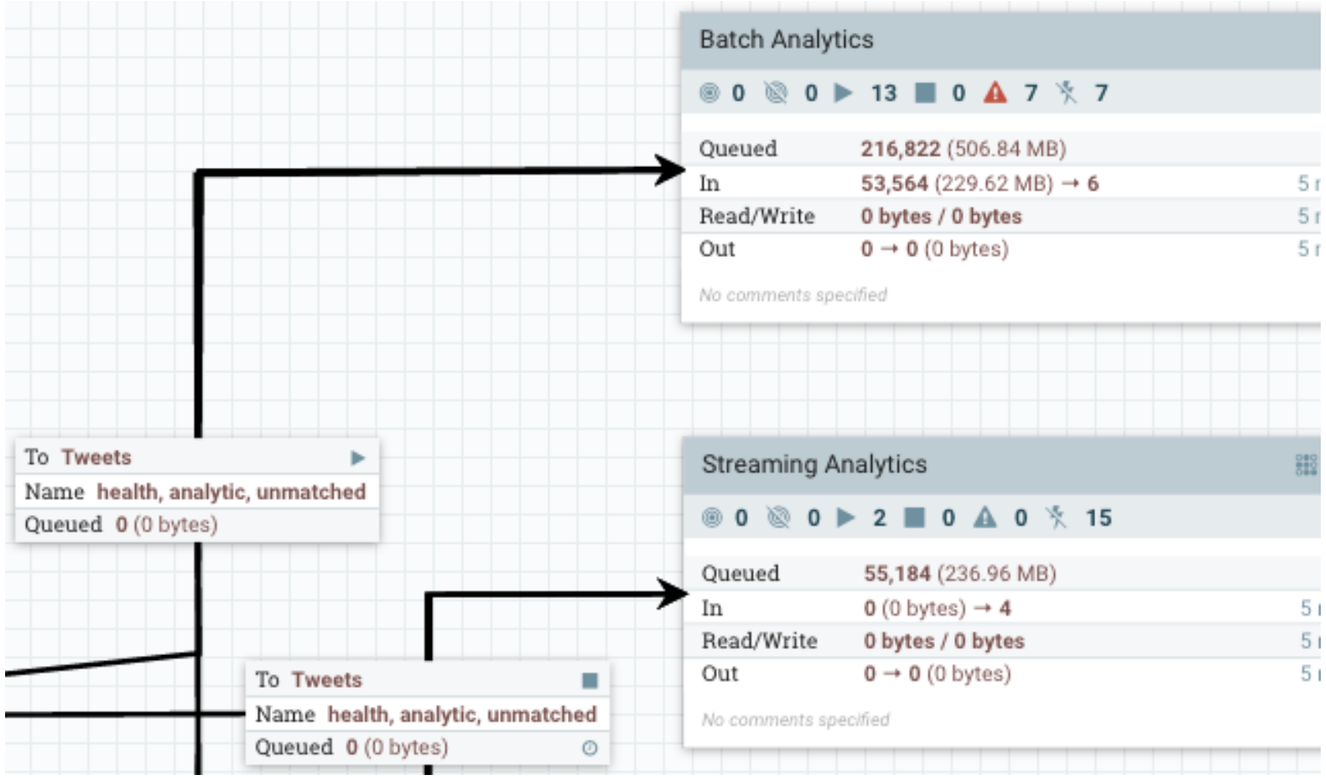
CCRi
DATA TO KNOWLEDGE

HORTONWORKS

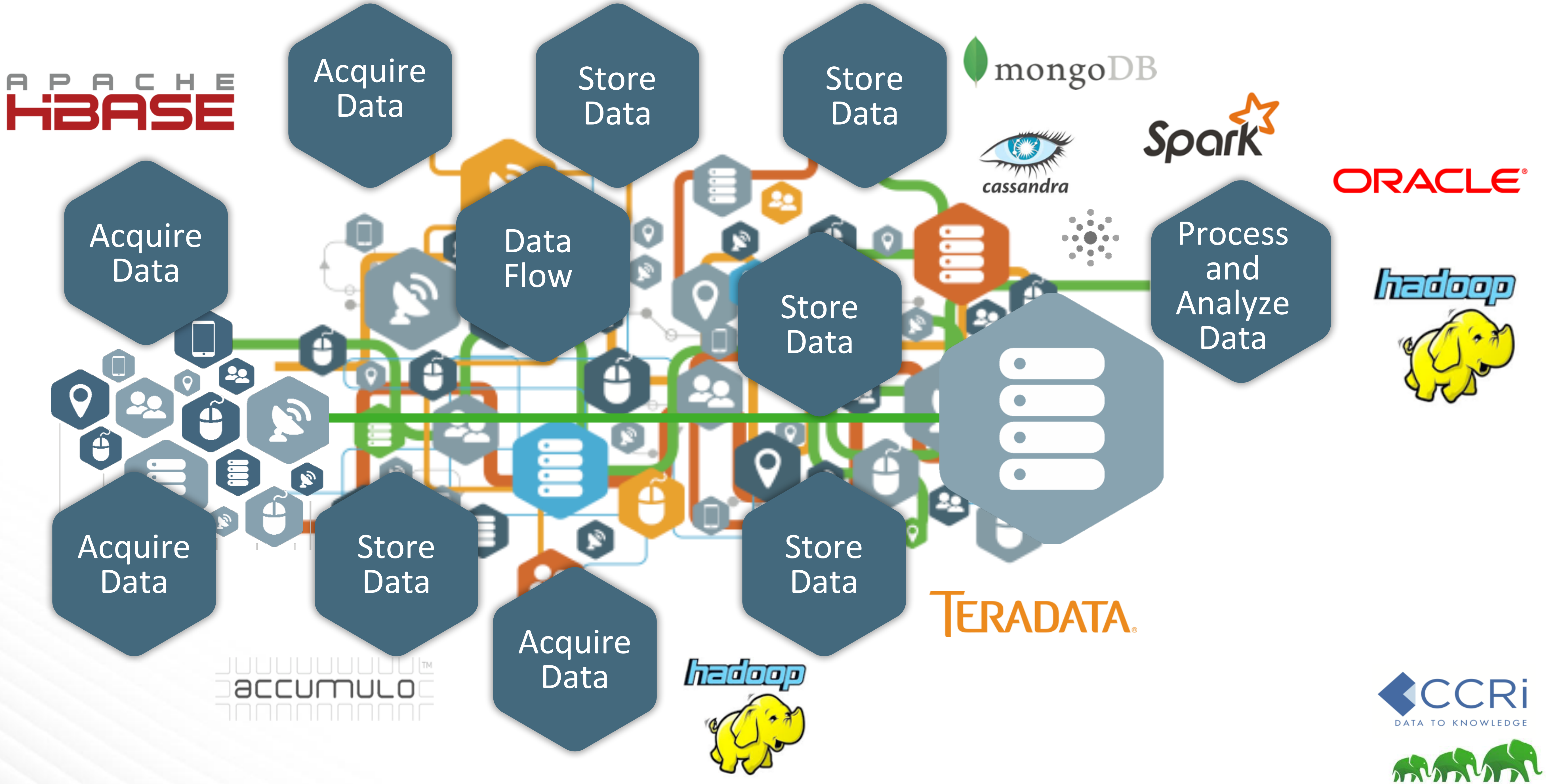
What is Apache NiFi?

An open source project dedicated to making dataflow easy. It's as easy as drag and drop! Literally!

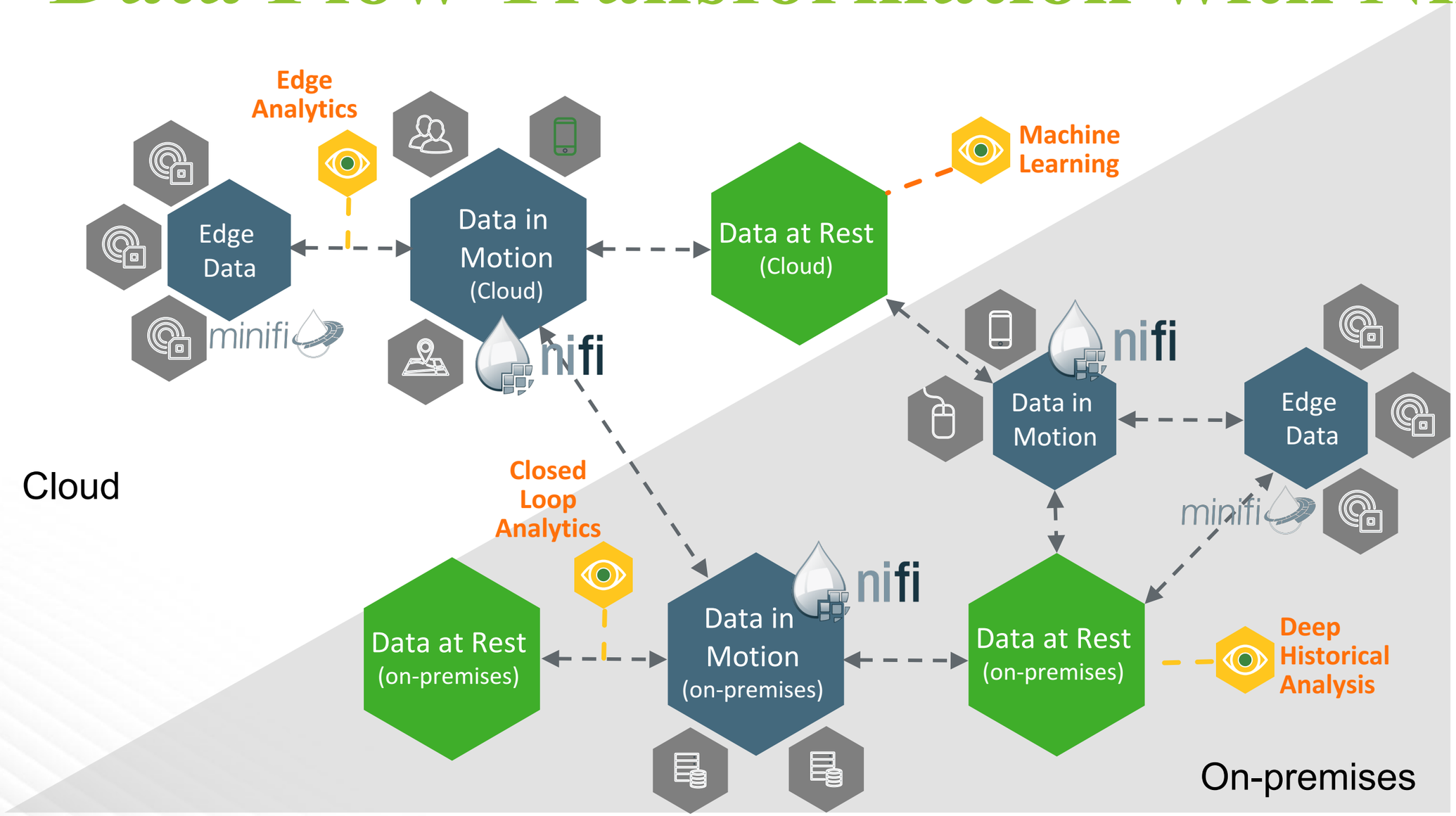
- Powerful and scalable directed graphs of data routing, transformation, and system mediation logic.
- Web-based user interface
- Highly configurable
- Data Provenance
- Designed for extension
- Secure



Convoluted Data Flow



Data Flow Transformation with NiFi

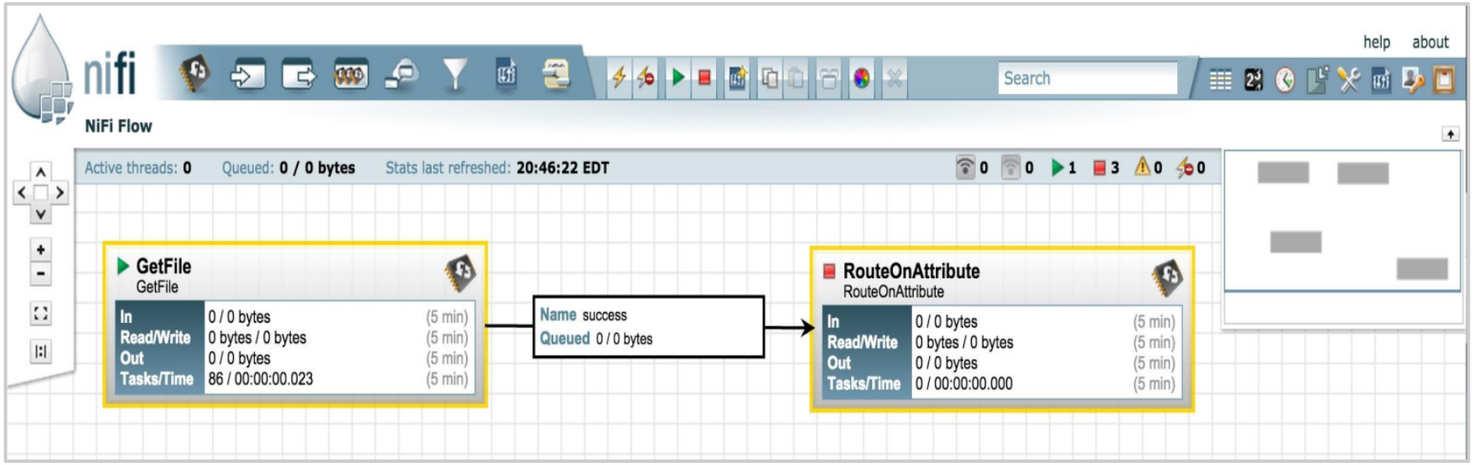


200+ Processors

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

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

NiFi – UI and Terms

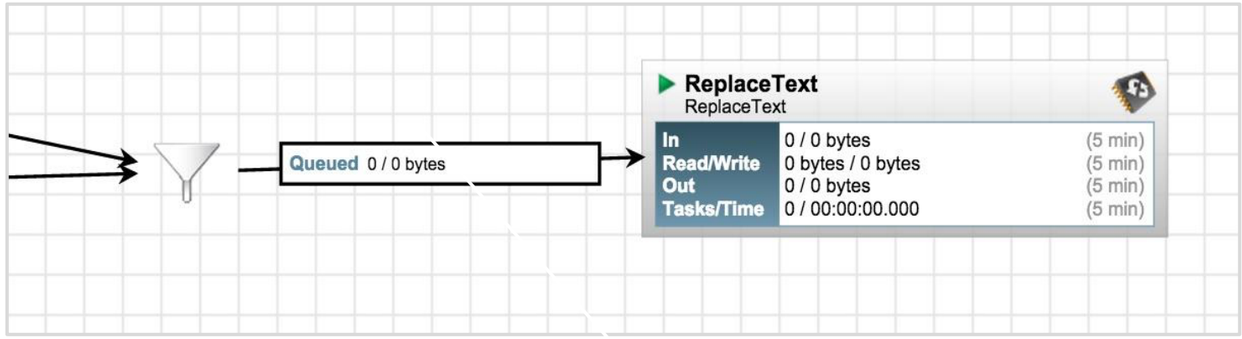


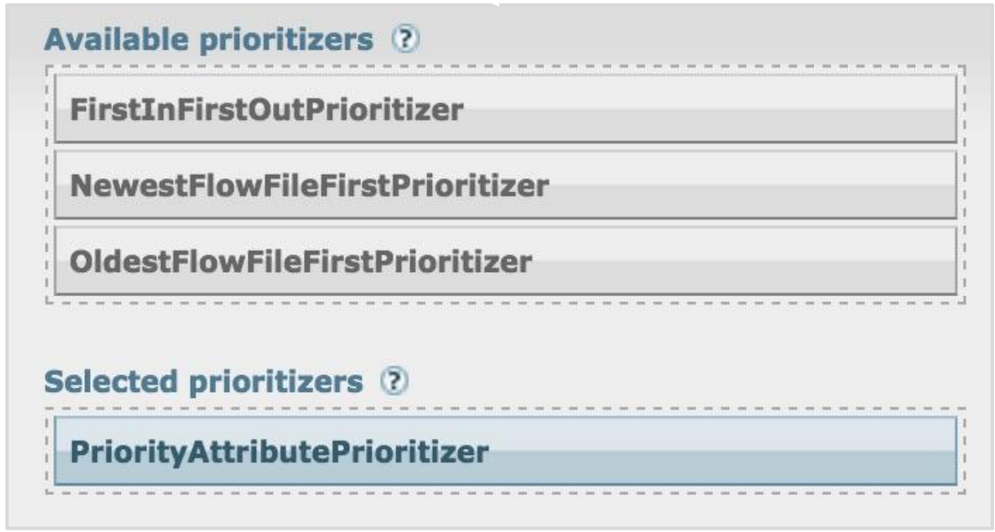
NiFi Terminology	
FlowFile Unit of data moving through the system Content + Attributes (key/value pairs)	**Connection** Links between processors Queues that can be dynamically prioritized
Processor Performs the work, can access FlowFiles	**Process Group** Set of processors and their connections Receive data via input ports, send data via output ports

- Drag and drop processors to build a flow
- Start, stop, and configure components in real time
- View errors and corresponding error messages
- View statistics and health of data flow
- Create templates of common processor & connections

NiFi – Queue Prioritization

- Configure a prioritizer per connection
- Determine what is important for your data – time based, arrival order, importance of a data set
- Funnel many connections down to a single connection to prioritize across data sets
- Develop your own prioritizer, if needed





NiFi – Provenance

NiFi Flow Data Provenance

Oldest event available: 07/29/2015 14:08:06 EDT

Filter | by component name

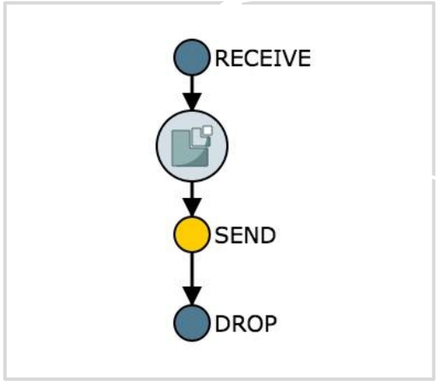
Displaying 1,000 of 1,000

Last updated: 21:12:00 EDT

Showing the most recent 1,000 of 62,293 events, please refine the search. Search

	Date/Time	Type	FlowFile Uuid	Size	Component Name	Component Type	
	07/29/2015 16:21:34.368 EDT	DROP	3b9f20bc-031e-4af8-ad8a-fedce...	158 bytes	PutSolrContentStream	PutSolrContentStream	
	07/29/2015 16:21:34.367 EDT	SEND	3b9f20bc-031e-4af8-ad8a-fedce...	158 bytes	PutSolrContentStream	PutSolrContentStream	
	07/29/2015 16:21:34.366 EDT	DROP	6f5036bc-1768-476d-9b6d-1f83...	2.15 KB	PutSolrContentStream	PutSolrContentStream	

- Tracks data at each point as it flows through the system
- Records, indexes, and makes events available for display
- Handles fan-in/fan-out, i.e. merging and splitting data
- View attributes and content at given points in time



Provenance Event

Details | Attributes | Content

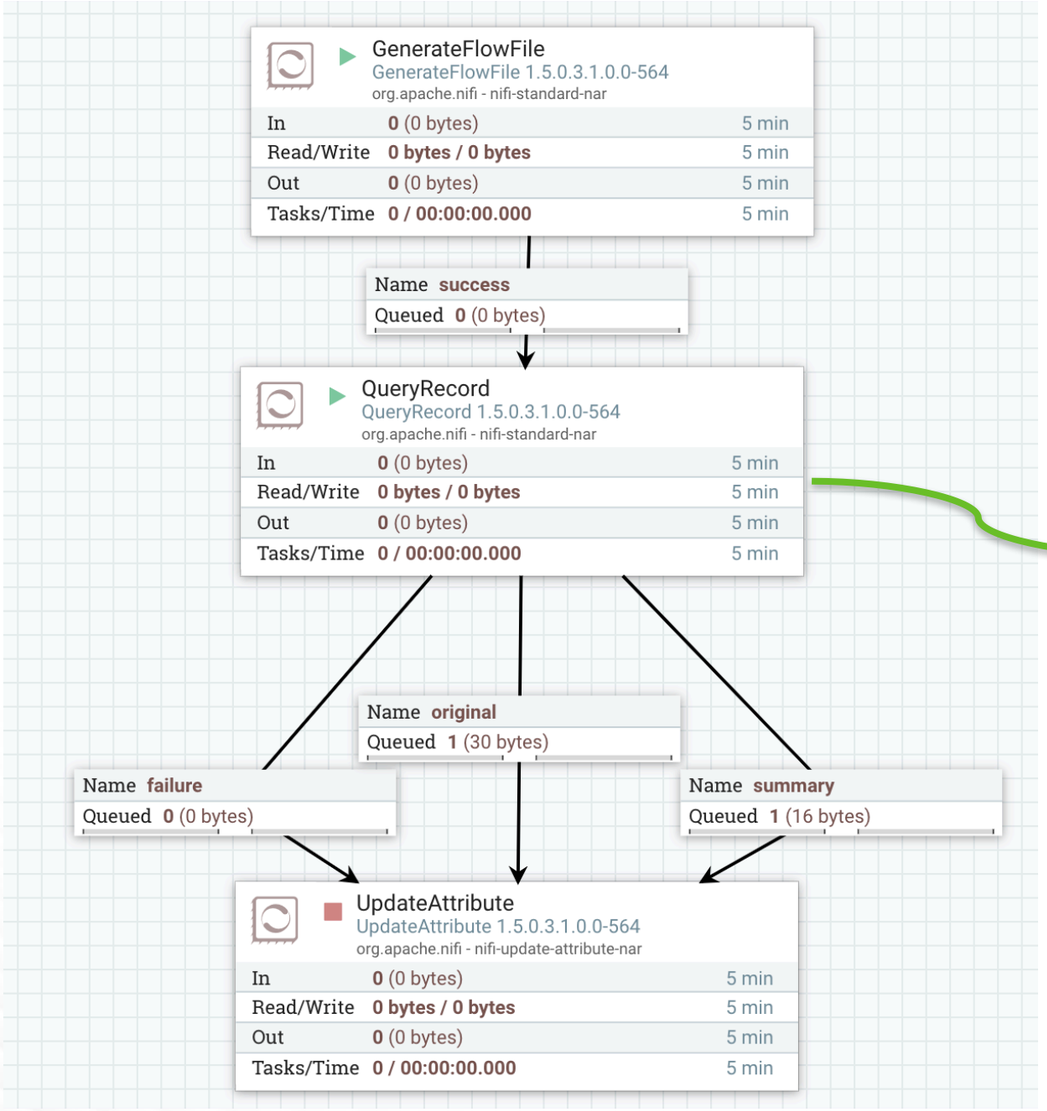
Time	07/29/2015 16:21:34.367 EDT
Event Duration	00:00:00.001
Lineage Duration	00:00:00.117
Type	SEND
FlowFile Uuid	3b9f20bc-031e-4af8-ad8a-fedcef4e0099
File Size	158 bytes
Component Id	fa7b551f-c405-4fde-b004-0b0d69c03472
Component Name	PutSolrContentStream
Component Type	PutSolrContentStream
Transit Uri	solr://http://localhost:8984/solr/chronicle
Details	*No value set*

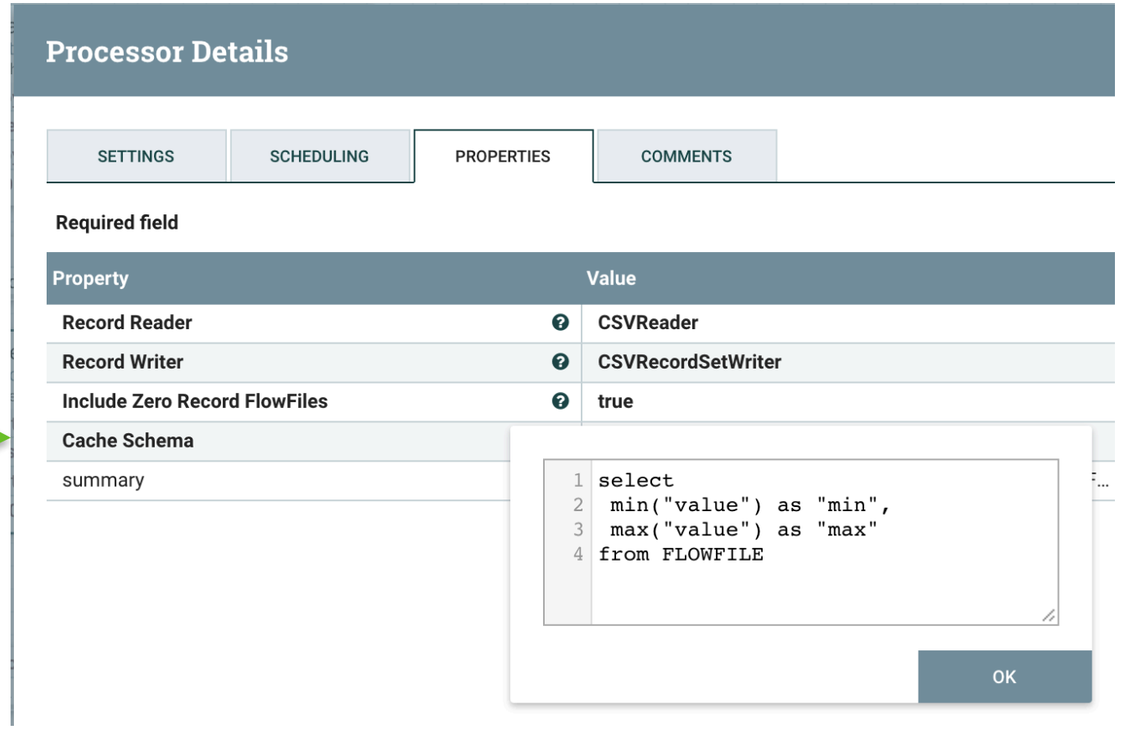
NiFi Demo


CCRi
DATA TO KNOWLEDGE

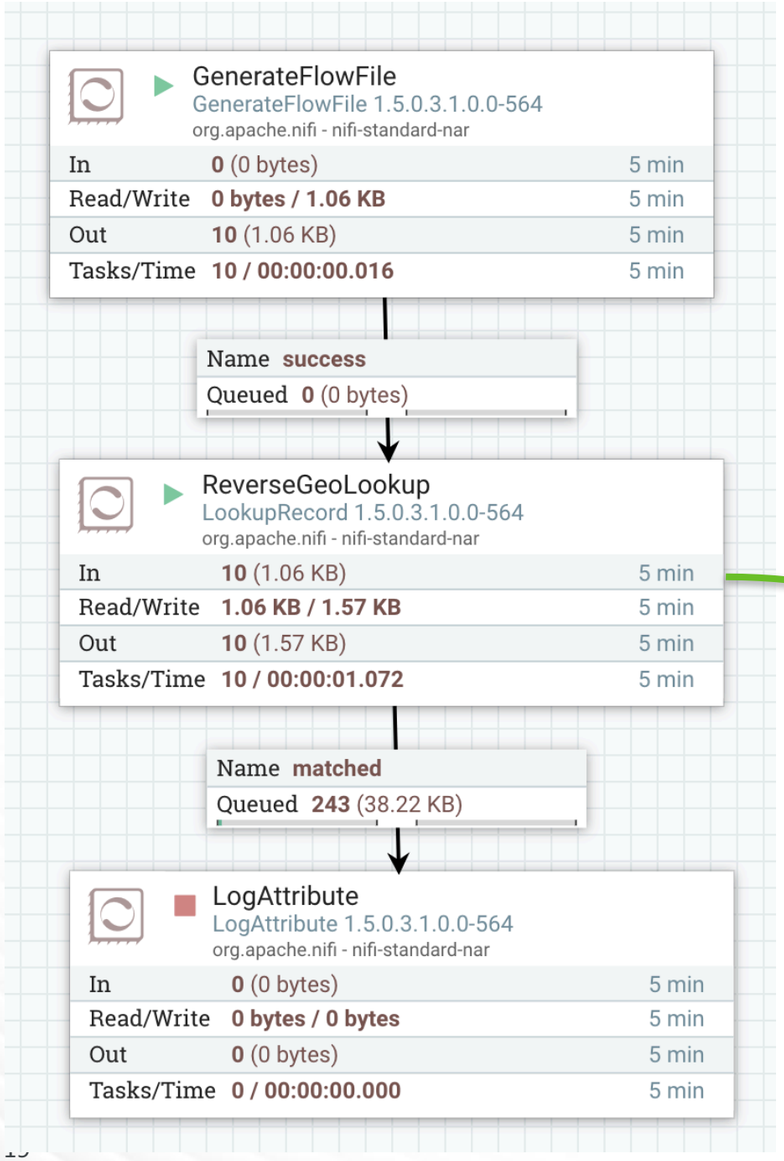

HORTONWORKS®
POWERING THE FUTURE OF DATA™

SQL with NiFi Record Processors





Reverse GeoLookup



Processor Details

SETTINGS | SCHEDULING | PROPERTIES | COMMENTS

Required field

Property	Value
Record Reader	**CSVReader**
Record Writer	**CSVRecordSetWriter**
Lookup Service	**ScriptedLookupService**
Result RecordPath	/location
Routing Strategy	**Route to 'matched' or 'unmatched'**
Record Result Contents	**Insert Entire Record**
lat	/lat
lng	/lng

CCRi DATA TO KNOWLEDGE

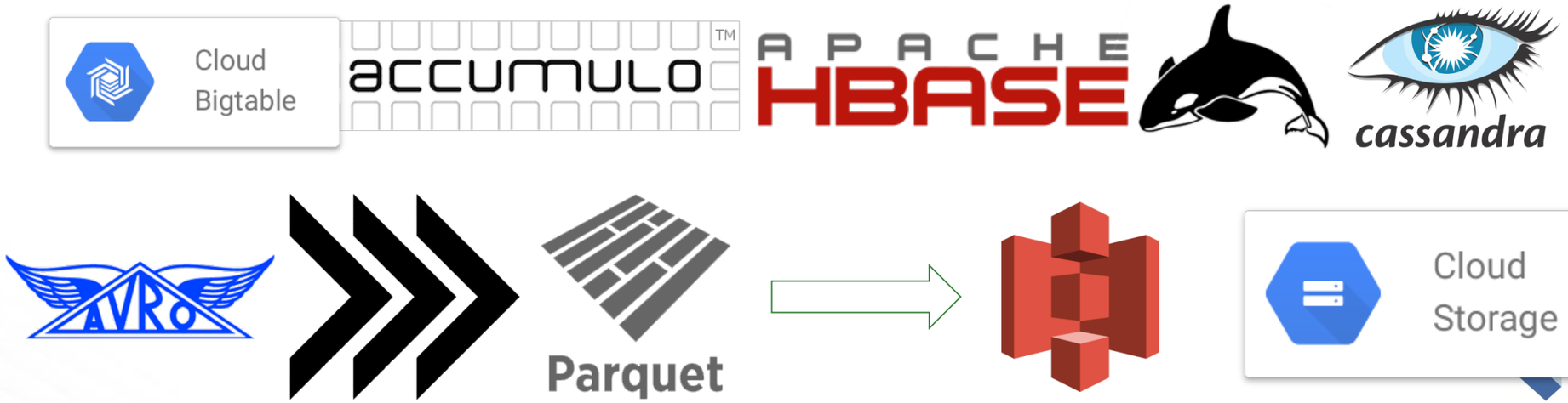
HORTONWORKS

What is GeoMesa?

A suite of tools for persisting, querying, analyzing, and streaming spatio-temporal data at scale

What is GeoMesa?

A suite of tools for **persisting**, querying, analyzing, and streaming spatio-temporal data at scale



What is GeoMesa?

A suite of tools for persisting, **querying**, analyzing, and streaming spatio-temporal data at scale

GeoTools

GeoServer

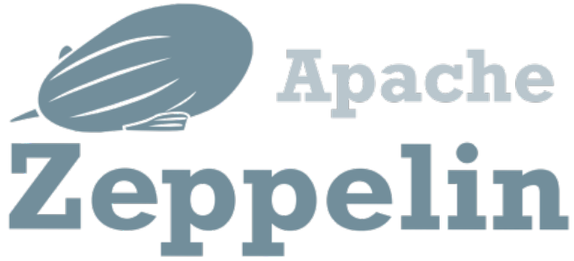
APACHE Spark™

CCRi DATA TO KNOWLEDGE

HORTONWORKS®

What is GeoMesa?

A suite of tools for persisting, querying, **analyzing**, and streaming spatio-temporal data at scale



CCRi
DATA TO KNOWLEDGE

HORTONWORKS

What is GeoMesa?

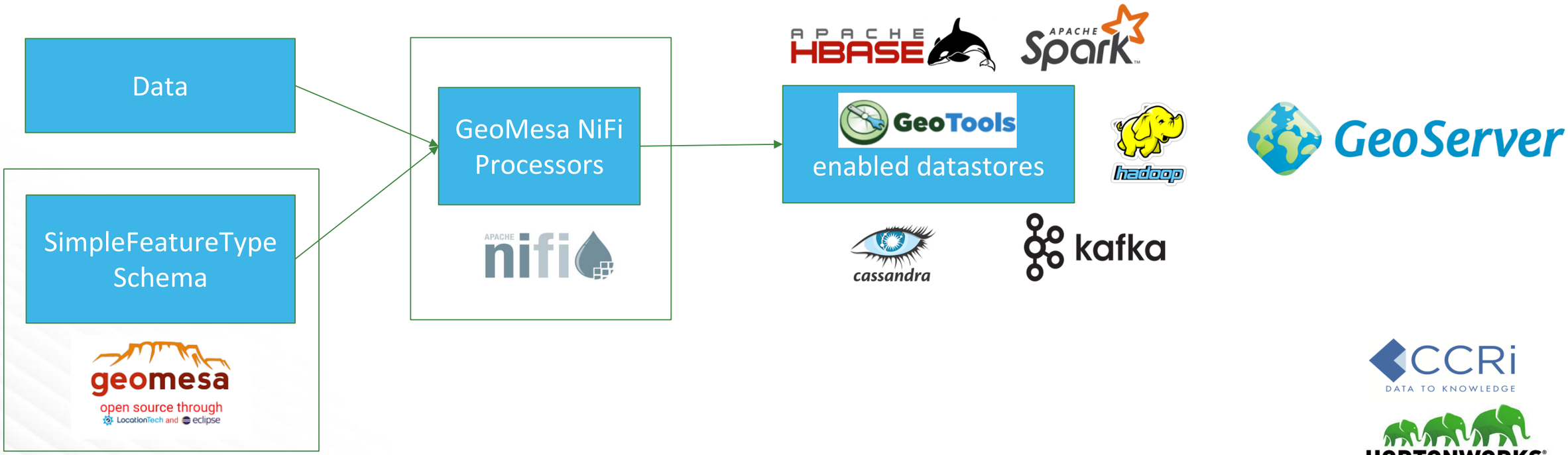
A suite of tools for persisting, querying, analyzing, and **streaming** spatio-temporal data at scale

CCRi
DATA TO KNOWLEDGE

HORTONWORKS

GeoMesa NiFi

- GeoMesa-NiFi allows you to ingest data into GeoMesa straight from NiFi by leveraging custom processors.
- NiFi allows you to ingest data into GeoMesa from every source GeoMesa supports and more.



GeoMesa NiFi Processors

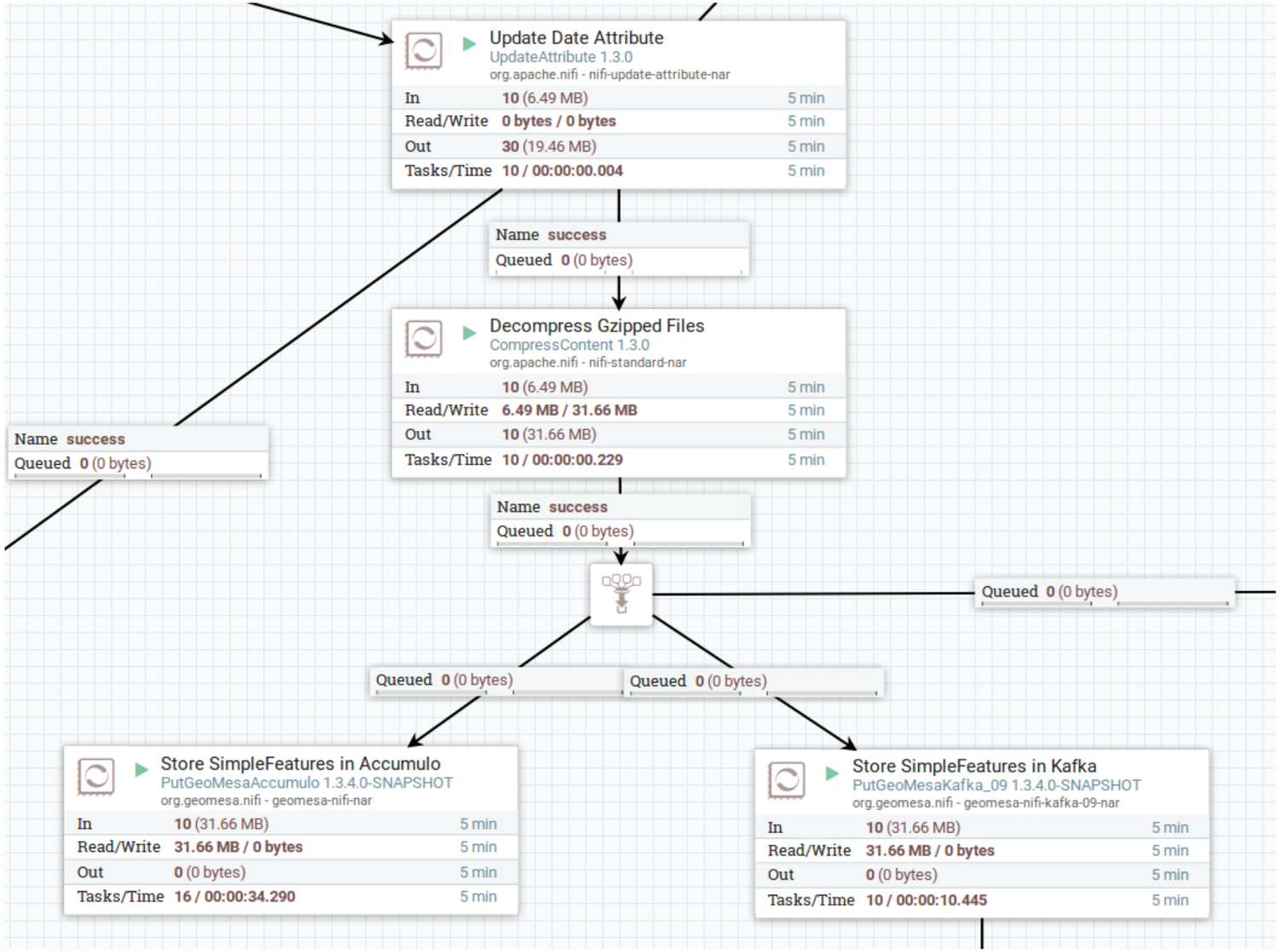
- **PutGeoMesaAccumulo**: Ingest data into a GeoMesa Accumulo datastore with a GeoMesa converter or from geoavro
- **PutGeoMesaHBase**: Ingest data into a GeoMesa HBase datastore with a GeoMesa converter or from geoavro
- **PutGeoMesaFileSystem**: Ingest data into a GeoMesa File System datastore with a GeoMesa converter or from geoavro
- **PutGeoMesaKafka**: Ingest data into a GeoMesa Kafka datastore with a GeoMesa converter or from geoavro
- **PutGeoTools**: Ingest data into an arbitrary GeoTools datastore using a GeoMesa converter or avro
- **ConvertToGeoAvro**: Use a GeoMesa converter to create geoavro

GeoMesa-NiFi Demo

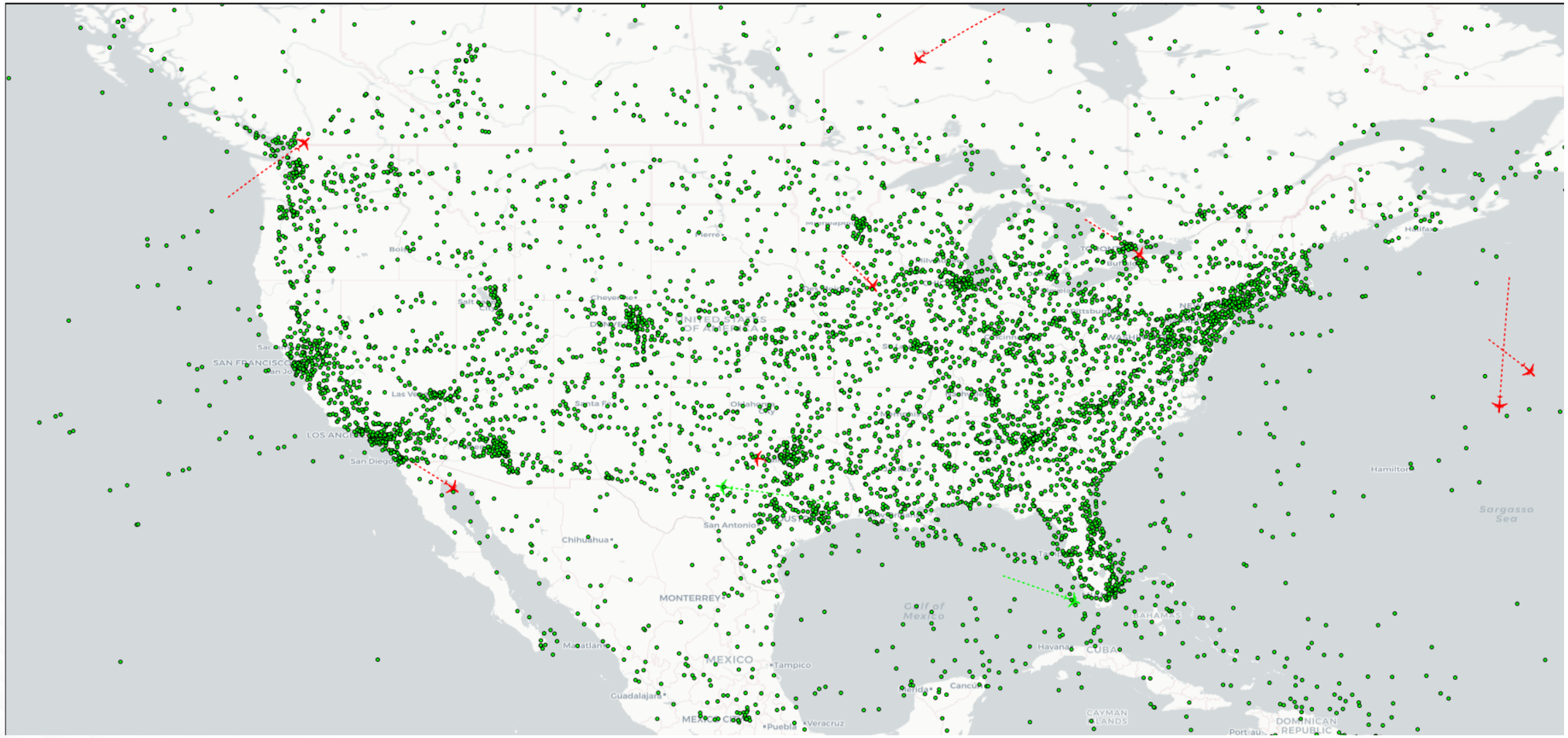
CCRi
DATA TO KNOWLEDGE

HORTONWORKS®
POWERING THE FUTURE OF DATA™

NiFi-GeoMesa Data Flow



Map in Real-Time



Resources:

- GeoMesa Project: http://www.geomesa.org/
- Geomesa-NiFi: https://github.com/geomesa/geomesa-nifi
- NiFi Project: https://nifi.apache.org/
- NiFi Overview and Tutorials: https://hortonworks.com/apache/nifi/

Thank you!