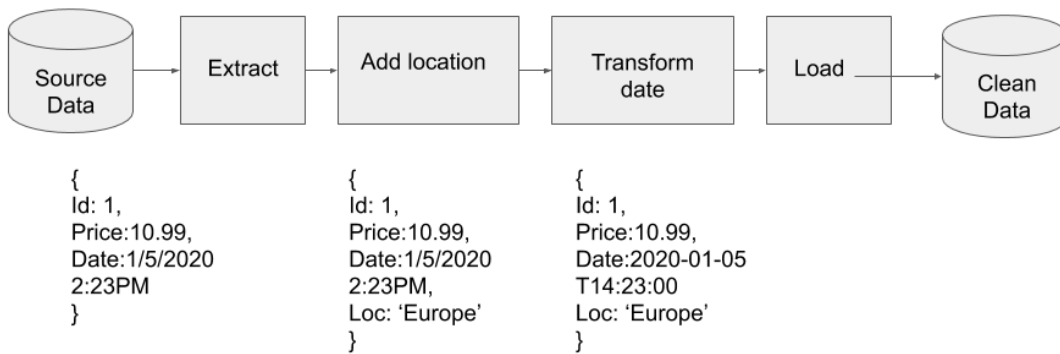


Chapter 1: What Is Data Engineering



Widget	Region
Blue	1
Green	2
Red	3

RegionID	Name
1	N.America
2	Asia
3	Europe

Widget	Region
Blue	1
Green	2
Red	3

Widget	Blue
	Green
	Red
Region	1
	2
	3

SELECT * FROM Sales


Group by Color

INSERT Data Warehouse









```
{
  {"Blue","Europe",19.95},
  {"Green","Asia",20.00},
  {"Blue","N.America",25.99},
  {"Red","Asia",14.95}
}
```

```
{
  {"Blue",2},
  {"Green",1},
  {"Red",1}
}
```

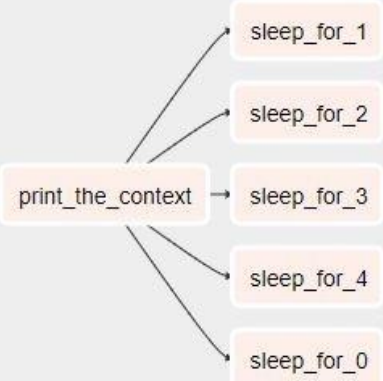
 **Airflow** DAGs Data Profiling ▾ Browse ▾ Admin ▾ Docs ▾ About ▾

DAG [example_python_operator] is now fresh as a daisy

☒ On **DAG: example_python_operator**

 **Graph View**  Tree View  Task Duration  Task Tries  Landing Times  Gantt

Base date: 2020-02-18 01:23:57 Number of runs: 25 ▾ Run: ▾ Layout: Left->Right



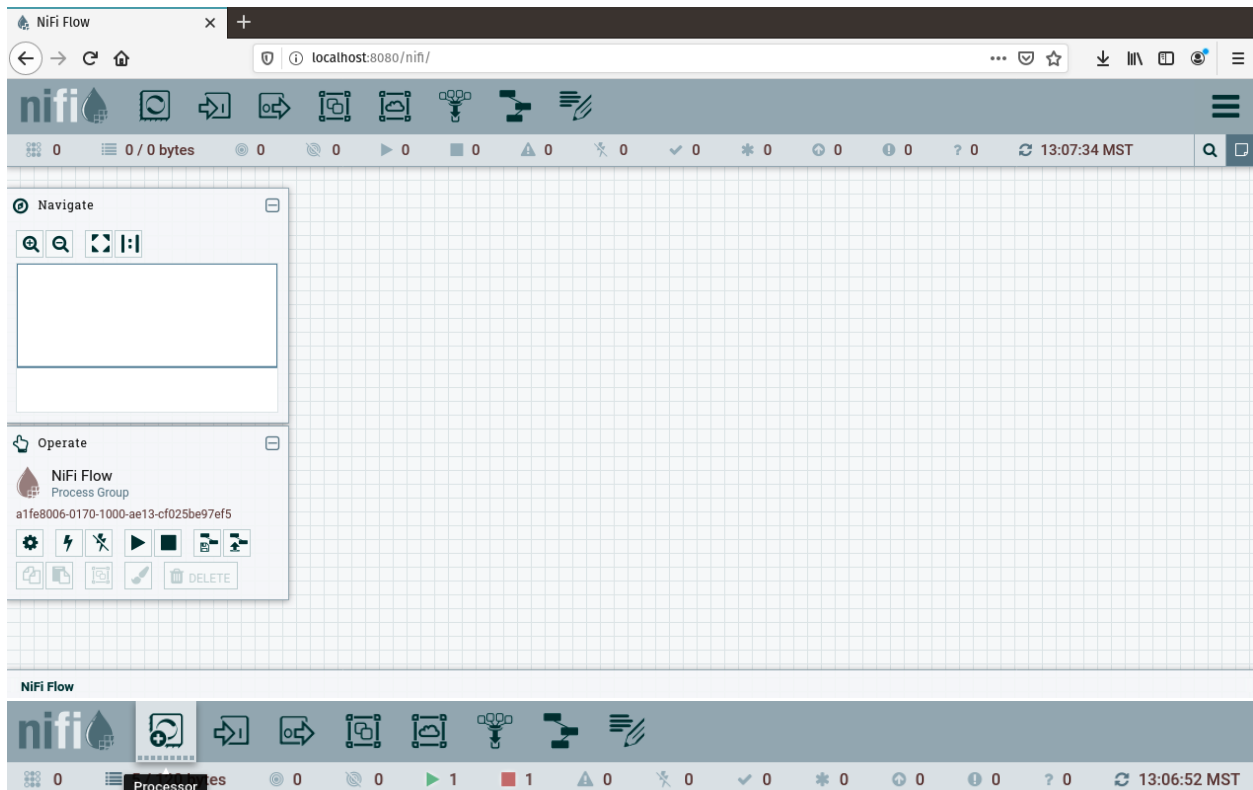
```
graph LR; A[print_the_context] --> B[sleep_for_1]; A --> C[sleep_for_2]; A --> D[sleep_for_3]; A --> E[sleep_for_4]; A --> F[sleep_for_0]
```


Chapter 2: Building Our Data Engineering Infrastructure

```
paulcrickard@pop-os:~$ sudo nifi*/bin/nifi.sh start

Java home: /usr/lib/jvm/java-1.11.0-openjdk-amd64
NiFi home: /home/paulcrickard/nifi-1.11.3

Bootstrap Config File: /home/paulcrickard/nifi-1.11.3/conf/bootstrap.conf
paulcrickard@pop-os:~$
```



NiFi Flow

localhost:9300/nifi/?processGroupId=root&componentId=ac42a400-0170-1000-8ea0-ee5ca5617099

Configure Processor

Invalid

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field

Property	Value
File Size	0B
Batch Size	1
Data Format	Text
Unique FlowFiles	false
Custom Text	This is a file from nifi
Character Set	UTF-8
filename	NifiFile.txt

CANCEL APPLY

NiFi Flow

localhost:9300/nifi/

0 1 / 24 bytes 0 0 1 1 0 0 0 0 0 0 0 0 12:59:26 MST

Navigate

Operate

NiFi Flow Process Group

a1fe8006-0170-1000-ae13-cf025be97ef5

GenerateFlowFile
GenerateFlowFile 1.11.3
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 24 bytes	5 min
Out	1 (24 bytes)	5 min
Tasks/Time	1 / 00:00:00.989	5 min

Name success
Queued 1 (24 bytes)

PutFile
PutFile 1.11.3
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

NiFi Flow

localhost:9300/nifi/

success

Displaying 2 of 2 (48.00 bytes)

	Position	UUID	Filename	File Size	Queued Duration	Lineage Duration	Penalized	
i	1	f8f4fd78-38f7-4966-b40c-1892f4...	NifiFile.txt	24.00 bytes	00:00:22.332	00:00:22.332	No	👤 👁 📄
i	2	87b095a1-f82f-48b5-a47a-129b1...	NifiFile.txt	24.00 bytes	00:00:12.312	00:00:12.312	No	👤 👁 📄

🔄 Last updated: 13:01:14 MST

NiFi Flow

localhost:9300/nifi/?processGroupId=root&componentId=ac47d551-0170-1000-e7b7-c166b78a3365

success

Displaying 2 of 2 (48.00 bytes)

	Position	UUID	Filename	File Size	Queued Duration	Lineage Duration	Penalized	
i	1	f8f4fd78-38f7-4966-b40c-1892f4...	NifiFile.txt	24.00 bytes	00:00:22.332	00:00:22.332	No	👤 👁 📄
i	2	87b095a1-f82f-48b5-a47a-129b1...	NifiFile.txt	24.00 bytes	00:00:12.312	00:00:12.312	No	👤 👁 📄

🔄 Last updated: 13:01:14 MST

FlowFile

DETAILS ATTRIBUTES

FlowFile Details

UUID
f8f4fd78-38f7-4966-b40c-1892f47e4386

Filename
NifiFile.txt

File Size
24 bytes

Queue Position
No value set

Queued Duration
00:00:46.491

Lineage Duration
00:00:46.491

Penalized
No

Content Claim

Container
default

Section
1

Identifier
1583438364746-1

Offset
72

Size
24 bytes

📄 DOWNLOAD 👁 VIEW

OK

NiFi Flow

NiFi

localhost:9300/nifi-content-viewer/?ref=http%3A%2F%2Flocalhost%3A9300%2Fnifi-api%2Fflowfile-queues%2Fac47d551-0170-1000-e7b7-c166b78a3365

View as: original

Filename: NifiFile.txt
Content Type: text/plain

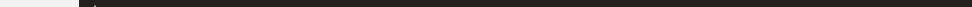
1 This is a file from nifi


```
paulcrickard@pop-os: /opt/nifioutput

paulcrickard@pop-os:/opt/nifioutput$ ls
NifiFile.txt
paulcrickard@pop-os:/opt/nifioutput$ cat NifiFile.txt; echo
This is a file from nifi
paulcrickard@pop-os:/opt/nifioutput$
```

```
paulcrickard@pop-os: ~































































































[2020-03-05 16:08:06,657] {__init__.py:51} INFO - Using executor SequentialExecutor
[2020-03-05 16:08:06,823] {scheduler_job.py:1344} INFO - Starting the scheduler
[2020-03-05 16:08:06,824] {scheduler_job.py:1352} INFO - Running execute loop for -1 seconds
[2020-03-05 16:08:06,825] {scheduler_job.py:1353} INFO - Processing each file at most -1 times
[2020-03-05 16:08:06,825] {scheduler_job.py:1356} INFO - Searching for files in /home/paulcrickard/airflow/dags
[2020-03-05 16:08:06,844] {scheduler_job.py:1358} INFO - There are 23 files in /home/paulcrickard/airflow/dags
[2020-03-05 16:08:06,845] {scheduler_job.py:1409} INFO - Resetting orphaned tasks for active dag runs
[2020-03-05 16:08:06,940] {dag_processing.py:556} INFO - Launched DagFileProcessorManager with pid: 29453
[2020-03-05 16:08:06,964] {settings.py:54} INFO - Configured default timezone <Timezone [UTC]>
[2020-03-05 16:08:07,000] {dag_processing.py:758} WARNING - Because we cannot use more than 1 thread (max_threads = 2) when using sqlite. So we set parallelism to 1.
```

The screenshot shows the Apache Airflow web interface. The browser tab is titled "Airflow - DAGs". The address bar shows the URL "localhost:8080/admin/". The interface has a dark teal header with the Airflow logo and navigation links: "DAGs", "Data Profiling", "Browse", "Admin", "Docs", and "About". The "DAGs" link is currently selected. The timestamp "2020-03-05 23:18:01 UTC" is displayed in the top right corner of the header.

DAGs

Search:

		DAG	Schedule	Owner	Recent Tasks 	Last Run 	DAG Runs 	Links
		example_bash_operator	0 0 * * *	Airflow	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>		<div><div></div><div></div><div></div></div>	        
		example_branch_dop_operator_v3	* * 1 * * * *	Airflow	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>		<div><div></div><div></div><div></div></div>	        
		example_branch_operator	@daily	Airflow	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>		<div><div></div><div></div><div></div></div>	        
		example_complex	None	airflow	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>		<div><div></div><div></div><div></div></div>	        
		example_external_task_marker_child	None	airflow	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>		<div><div></div><div></div><div></div></div>	        
		example_external_task_marker_parent	None	airflow	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>		<div><div></div><div></div><div></div></div>	        
		example_http_operator	1 day, 0:00:00	Airflow	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>		<div><div></div><div></div><div></div></div>	        
		example_passing_params_via_test_command	* * 1 * * * *	airflow	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>		<div><div></div><div></div><div></div></div>	        
		example_pig_operator	None	Airflow	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>		<div><div></div><div></div><div></div></div>	

A screenshot of a web browser window showing the Airflow DAGs page. The browser tab is titled "Airflow - DAGs". The address bar shows the URL: `localhost:8080/admin/airflow/graph?dag_id=example_bash_operator&execution_date=`. The page header includes the Airflow logo and navigation links: DAGs, Data Profiling, Browse, Admin, Docs, and About. The current time is displayed as 2020-03-05 23:23:42 UTC.

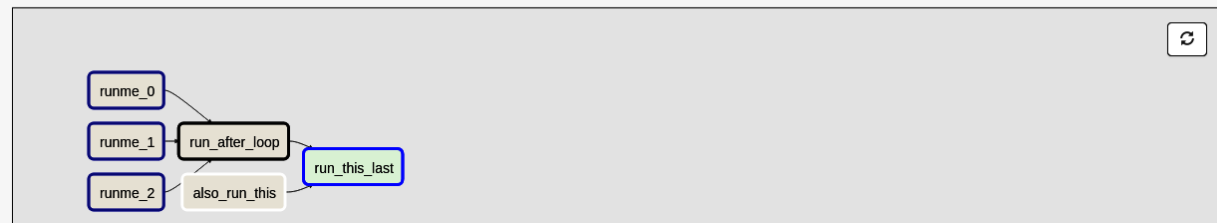
☐ Off DAG: example_bash_operator

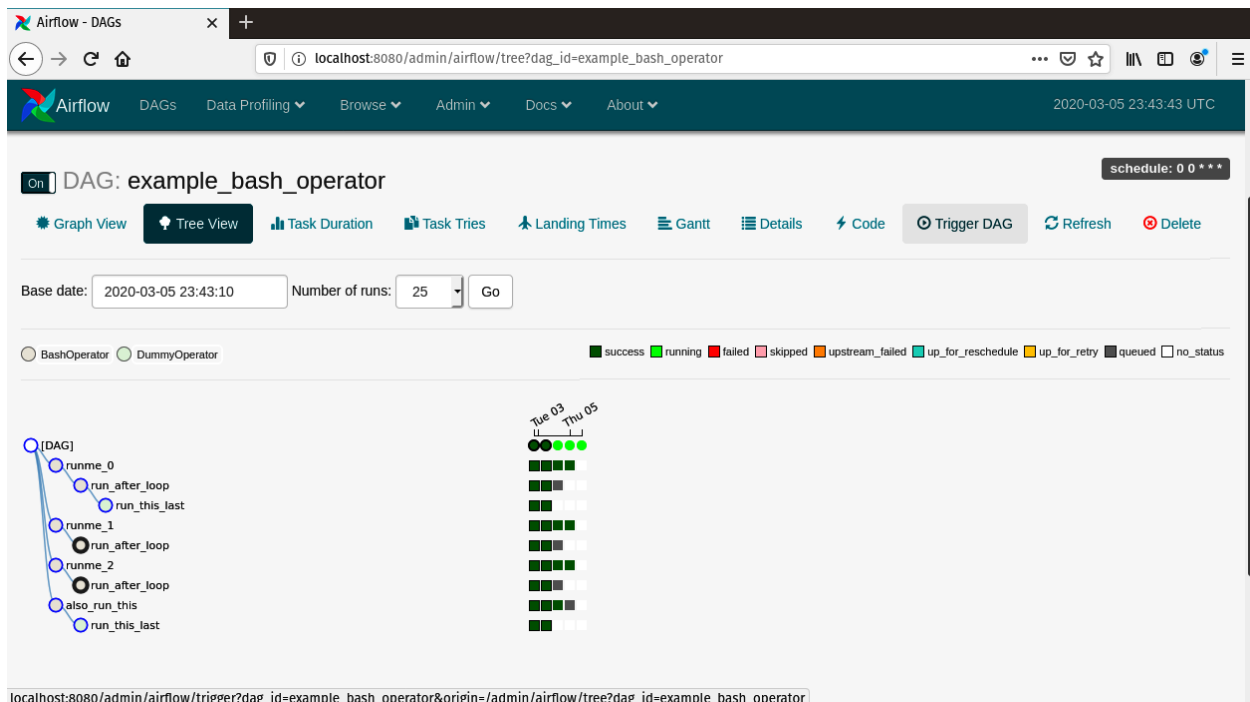
```
schedule: 0 0 * * *
```

Graph View
Tree View
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Trigger DAG
Refresh
Delete

None Base date: 2020-03-05 23:23:06 Number of runs: 25 Run: Layout: Left->Right Go Search for...

BashOperator
DummyOperator
success
running
failed
skipped
upstream_failed
up_for_reschedule
up_for_retry
queued
no_status





```
paulcrickard@pop-os: ~/airflow

# The amount of parallelism as a setting to the executor. This defines
# the max number of task instances that should run simultaneously
# on this airflow installation
parallelism = 32

# The number of task instances allowed to run concurrently by the scheduler
dag_concurrency = 16

# Are DAGs paused by default at creation
dags_are_paused_at_creation = True

# The maximum number of active DAG runs per DAG
max_active_runs_per_dag = 16

# Whether to load the examples that ship with Airflow. It's good to
# get started, but you probably want to set this to False in a production
# environment
load_examples = True

# Where your Airflow plugins are stored
plugins_folder = /home/paulcrickard/airflow/plugins

# Secret key to save connection passwords in the db
```


Airflow - DAGs

localhost:8080/admin/

Airflow DAGs Data Profiling Browse Admin Docs About 2020-03-06 00:06:45 UTC

DAGs

Search:

	?	DAG	Schedule	Owner	Recent Tasks ?	Last Run ?	DAG Runs ?	Links
No data available in table								

Showing 0 to 0 of 0 entries

« < > »

[Hide Paused DAGs](#)

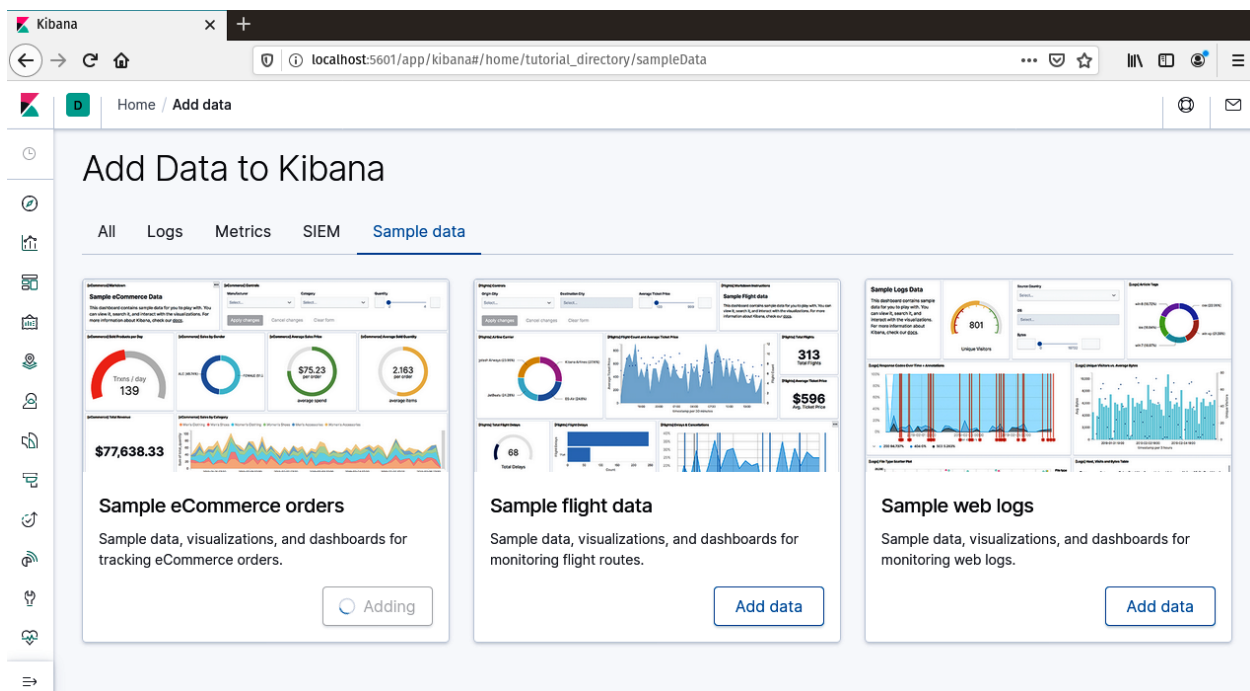
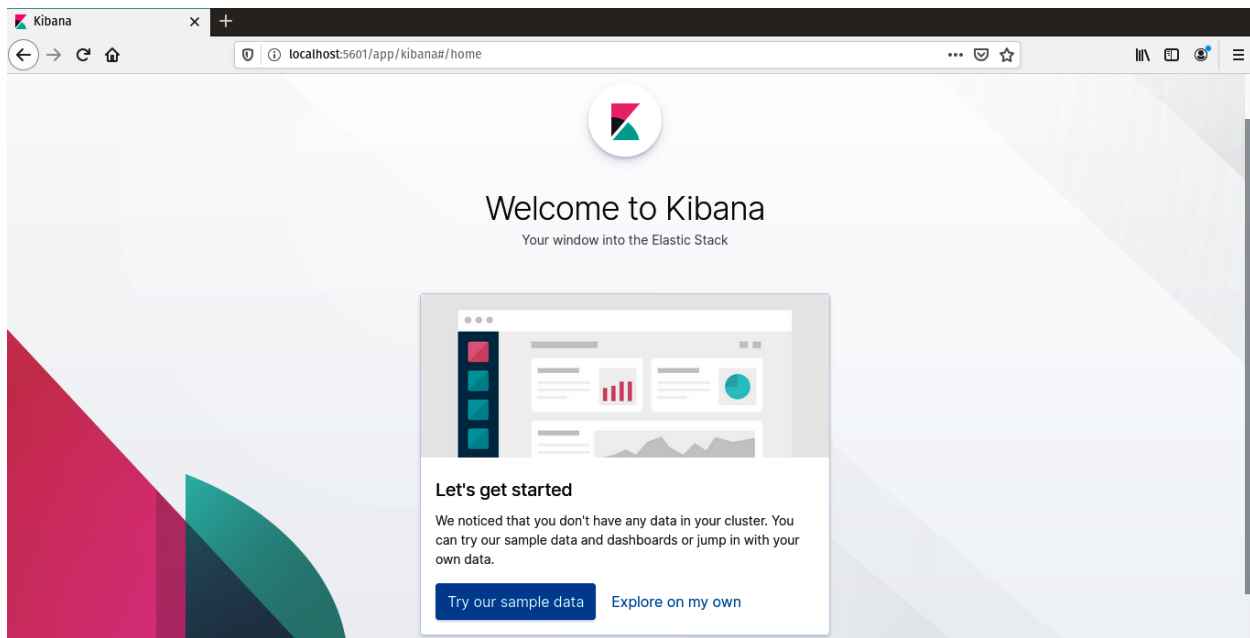
localhost:9200/

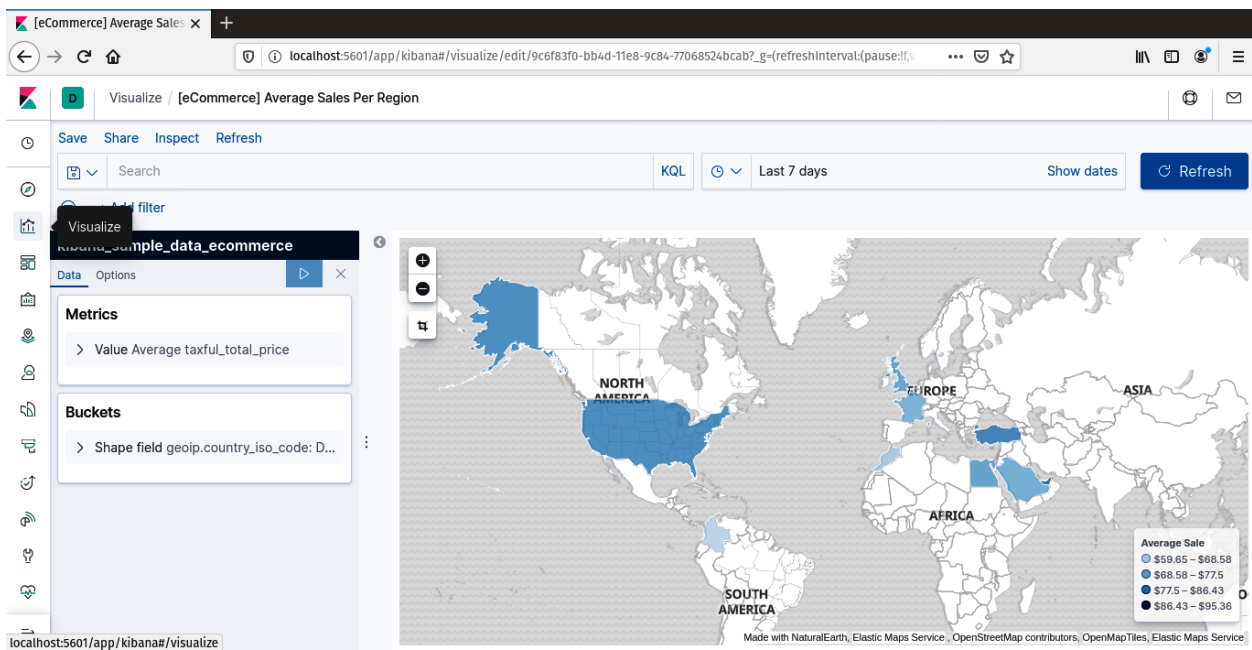
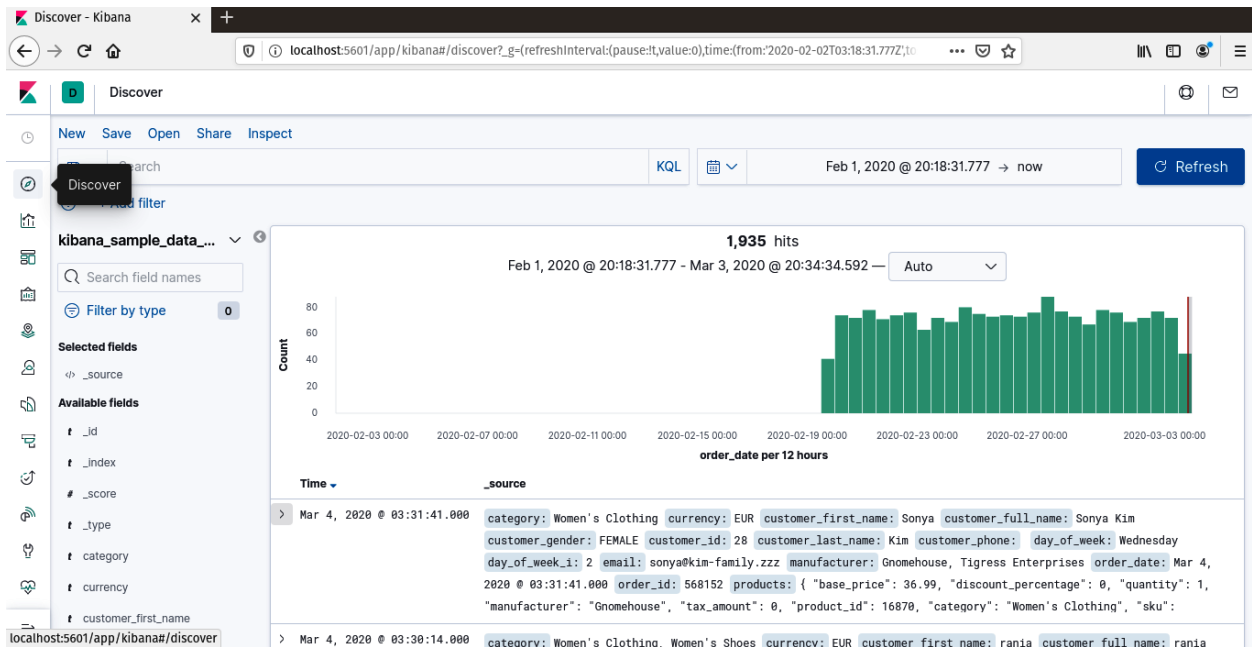
localhost:9200

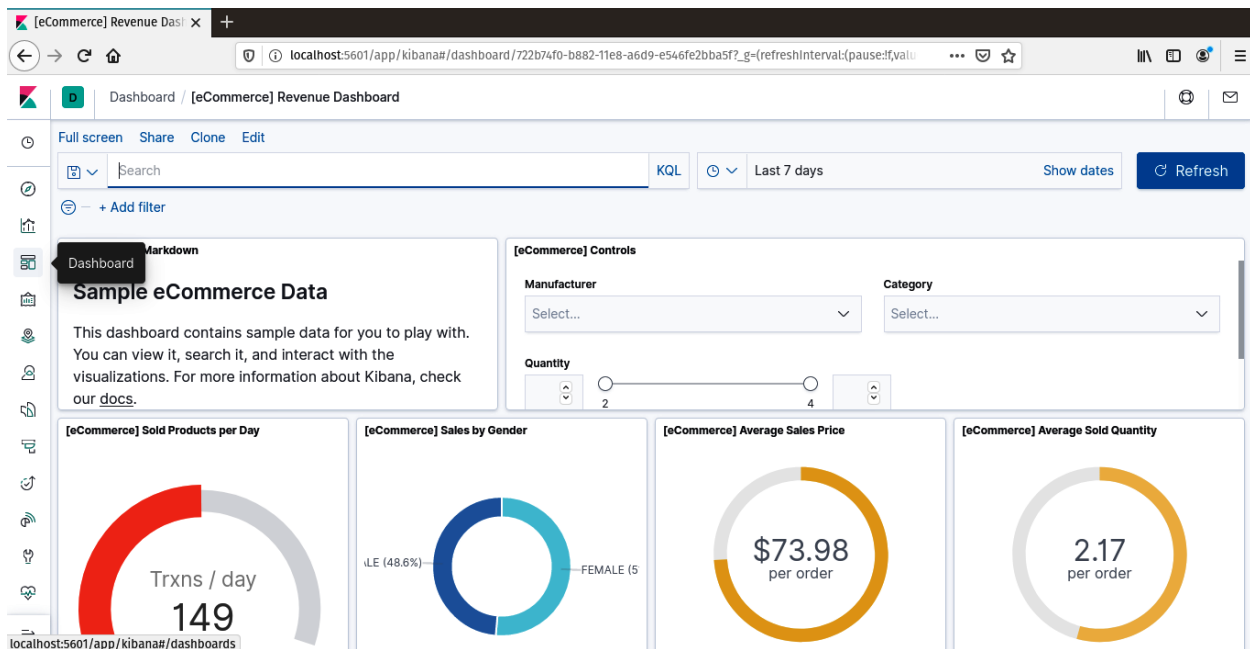
JSON Raw Data Headers

Save Copy Collapse All Expand All Filter JSON

```
name: "OnlyNode"
cluster_name: "DataEngineeringWithPython"
cluster_uuid: "hxxcUZ4dQTeMbh-VV66U4w"
version:
  number: "7.6.0"
  build_flavor: "default"
  build_type: "tar"
  build_hash: "7f634e9f44834fbc12724506cc1da681b0c3b1e3"
  build_date: "2020-02-06T00:09:00.449973Z"
  build_snapshot: false
  lucene_version: "8.4.0"
  minimum_wire_compatibility_version: "6.8.0"
  minimum_index_compatibility_version: "6.0.0-beta1"
tagline: "You Know, for Search"
```





Kibana

localhost:5601/app/kibana#/dev_tools/console

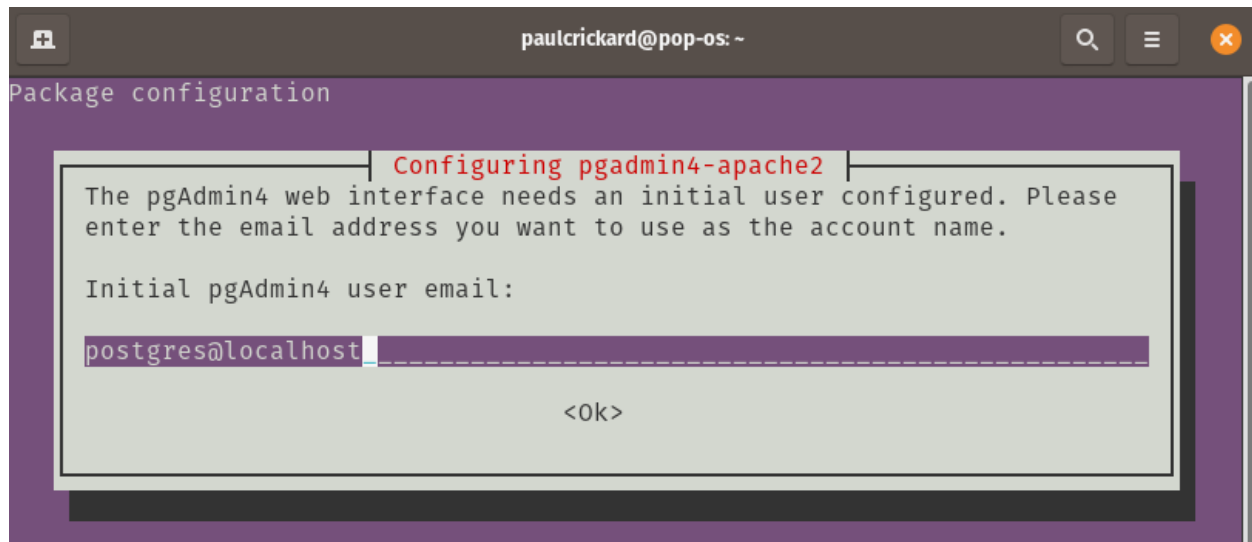
Dev Tools

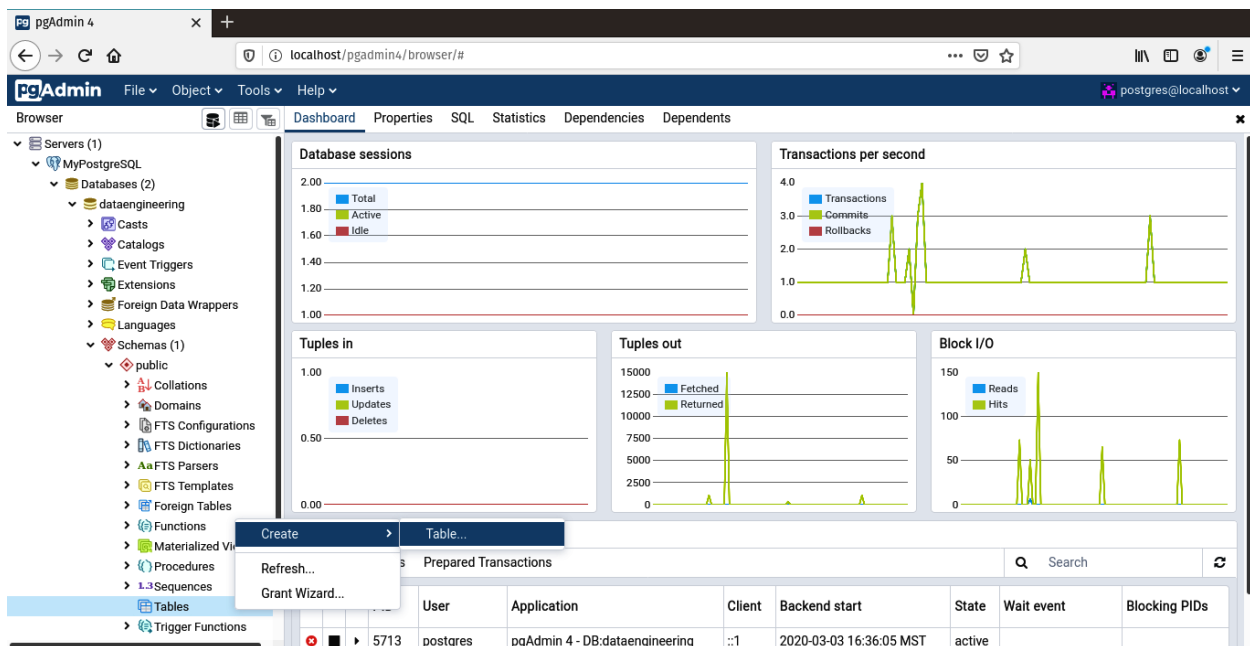
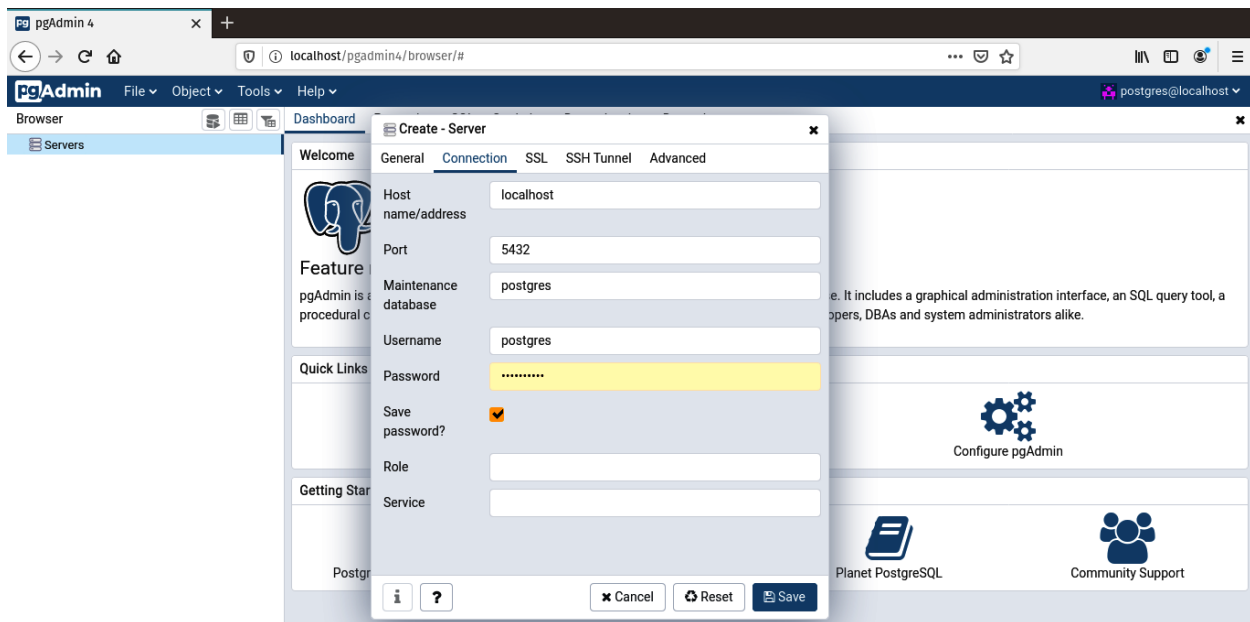
Console Search Profiler Grok Debugger

History Settings Help

```
1 PUT /test/_doc/1
2 {
3   "id":1,
4   "title":"Data Engineering With Python"
5 }
6
7 GET /test/_search
8 {
9   "query": {
10    "match": {"id":1}
11  }
12 }
```

```
1 {
2   "took" : 241,
3   "timed_out" : false,
4   "_shards" : {
5     "total" : 1,
6     "successful" : 1,
7     "skipped" : 0,
8     "failed" : 0
9   },
10  "hits" : {
11    "total" : {
12      "value" : 1,
13      "relation" : "eq"
14    },
15    "max_score" : 1.0,
16    "hits" : [
17      {
18        "_index" : "test",
19        "_type" : "_doc",
20        "_id" : "1",
21        "_score" : 1.0,
22        "_source" : {
23          "id" : 1,
24          "title" : "Data Engineering With Python"
25        }
26      }
27    ]
28  }
29 }
```



Create - Table



General **Columns** Constraints Advanced Partitions Parameters Security SQL

Inherited from table(s)

Select to inherit from...

Columns



	Name	Data type	Length/Precision	Scale	Not NULL?	Primary key?
	name	text			<input checked="" type="checkbox"/> Yes	<input type="checkbox"/> No
	id	integer			<input checked="" type="checkbox"/> Yes	<input checked="" type="checkbox"/> Yes
	street	text			<input checked="" type="checkbox"/> Yes	<input type="checkbox"/> No
	city	text			<input checked="" type="checkbox"/> Yes	<input type="checkbox"/> No
	zip	text			<input checked="" type="checkbox"/> Yes	<input type="checkbox"/> No



Cancel

Reset

Save

[DAGs](#)
[Data Profiling](#)
[Browse](#)
[Admin](#)
[Docs](#)
[About](#)
2020-03-13 18:55:47 UTC

The scheduler does not appear to be running. Last heartbeat was received 1 week ago.
The DAGs list may not update, and new tasks will not be scheduled.

DAGs

Search:

	DAG	Schedule	Owner	Recent Tasks	Last Run	DAG Runs	Links
--	-----	----------	-------	--------------	----------	----------	-------

Chapter 3: Reading and Writing Files

```
paulcrickard@pop-os: ~  
paulcrickard@pop-os:~$ cat mycsv.csv  
name,age  
Bob Smith,40  
paulcrickard@pop-os:~$
```

```
paulcrickard@pop-os: ~  
paulcrickard@pop-os:~$ python3  
Python 3.7.5 (default, Nov 20 2019, 09:21:52)  
[GCC 9.2.1 20191008] on linux  
Type "help", "copyright", "credits" or "license" for more information.  
>>> import pandas as pd  
>>> df=pd.read_csv('data.csv')  
>>> df.head(10)  
  
   name  age  ...      lng      lat  
0  Patrick Hendrix  23  ...  103.914462 -59.009437  
1   Grace Jackson  36  ...  170.503858  58.163167  
2   Arthur Garcia  61  ...   -39.845646  38.689889  
3   Gary Valentine  29  ...  -30.304522  81.272300  
4    Erin Mclean  23  ... -110.860085  11.476733  
5    Karen Hodges  57  ... -128.085033  23.872011  
6   Edgar Humphrey  42  ...  176.490032 -78.616711  
7  Andrew Williamson  34  ...  -72.802215  73.918076  
8   Michael Mack  57  ...   70.723959  63.915574  
9  Kristine Nielsen  77  ...  137.055928  61.351681  
  
[10 rows x 8 columns]  
>>>
```

Airflow - DAGs - Mozilla Firefox

Airflow - DAGs x +

localhost:8080/admin/

Airflow DAGs Data Profiling Browse Admin Docs About 2020-03-18 19:19:40 UTC

DAGs

Search:

	DAG	Schedule	Owner	Recent Tasks	Last Run	DAG Runs	Links
	Off MyCSVDAG	0:05:00	paulcrickard		2020-03-18 01:20		

Showing 1 to 1 of 1 entries

« < 1 > »

Hide Paused DAGs

Airflow - DAGs - Mozilla Firefox

Airflow - DAGs

localhost:8080/admin/airflow/tree?base_date=&num_runs=5&root=&dag_id=MyCSVVDAG&csrf_token=ijZmNzRmY

Airflow DAGs Data Profiling Browse Admin Docs About 2020-03-18 19:41:35 UTC

Graph View Tree View Task Duration Task Tries Landing Times Gantt Details Code Trigger DAG Refresh Delete

Base date: 2020-03-18 00:20:00 Number of runs: 5 Go

BashOperator PythonOperator

success running failed skipped upstream_failed up_for_reschedule up_for_retry queued no_status

[DAG] starting convertCSVtoJson

06 PM

Airflow - DAGs - Mozilla Firefox

Airflow - DAGs

localhost:8080/admin/airflow/tree?base_date=&num_runs=5&root=&dag_id=MyCSVVDAG&csrf_token=ijZmNzRmY

Airflow DAGs Data Profiling Browse Admin Docs About 2020-03-18 19:44:01 UTC

On DAG: MyCSVVDAG schedule: 0:05:00

Graph View Tree View Task Duration Task Tries Landing Times Gantt Details Code Trigger DAG Refresh Delete

Base date: 2020-03-18 00:35:00 Number of runs: 5 Go

BashOperator PythonOperator

Task_id: convertCSVtoJson
Run: 2020-03-18T00:15:00+00:00
Operator: PythonOperator
Started: 2020-03-18T19:42:44.659219+00:00
Ended: 2020-03-18T19:42:46.211794+00:00
Duration: 1.553Sec
State: success

success running failed skipped upstream_failed up_for_reschedule up_for_retry queued no_status

[DAG] starting convertCSVtoJson

Airflow - DAGs - Mozilla Firefox

Airflow - DAGs

localhost:8080/admin/airflow/tree?dag_id=MyCSVVDAG&num_runs=&root=

Airflow DAGs Data Profiling Browse Admin Docs About 2020-03-18 18:05:55 UTC

The scheduler does not appear to be running. Last heartbeat was 2020-03-18 18:05:55 UTC. The DAGs list may not update, and new tasks will not be created.

On DAG: MyCSVVDAG schedule: 0:05:00

Graph View Tree View Task Duration Task Tries Landing Times Gantt Details Code Trigger DAG Refresh Delete

Base date: 2020-03-18 00:50:00 Number of runs: 5 Go

BashOperator PythonOperator

[DAG] starting readCSV

starting on 2020-03-18T00:00:00+00:00

Task Instance Details Rendered Task Instances View Log

Download Log (by attempts):

1

Run Ignore All Deps Ignore Task State Ignore Task Deps

Clear Past Future Upstream Downstream Recursive Failed

Mark Failed Past Future Upstream Downstream

Mark Success Past Future Upstream Downstream

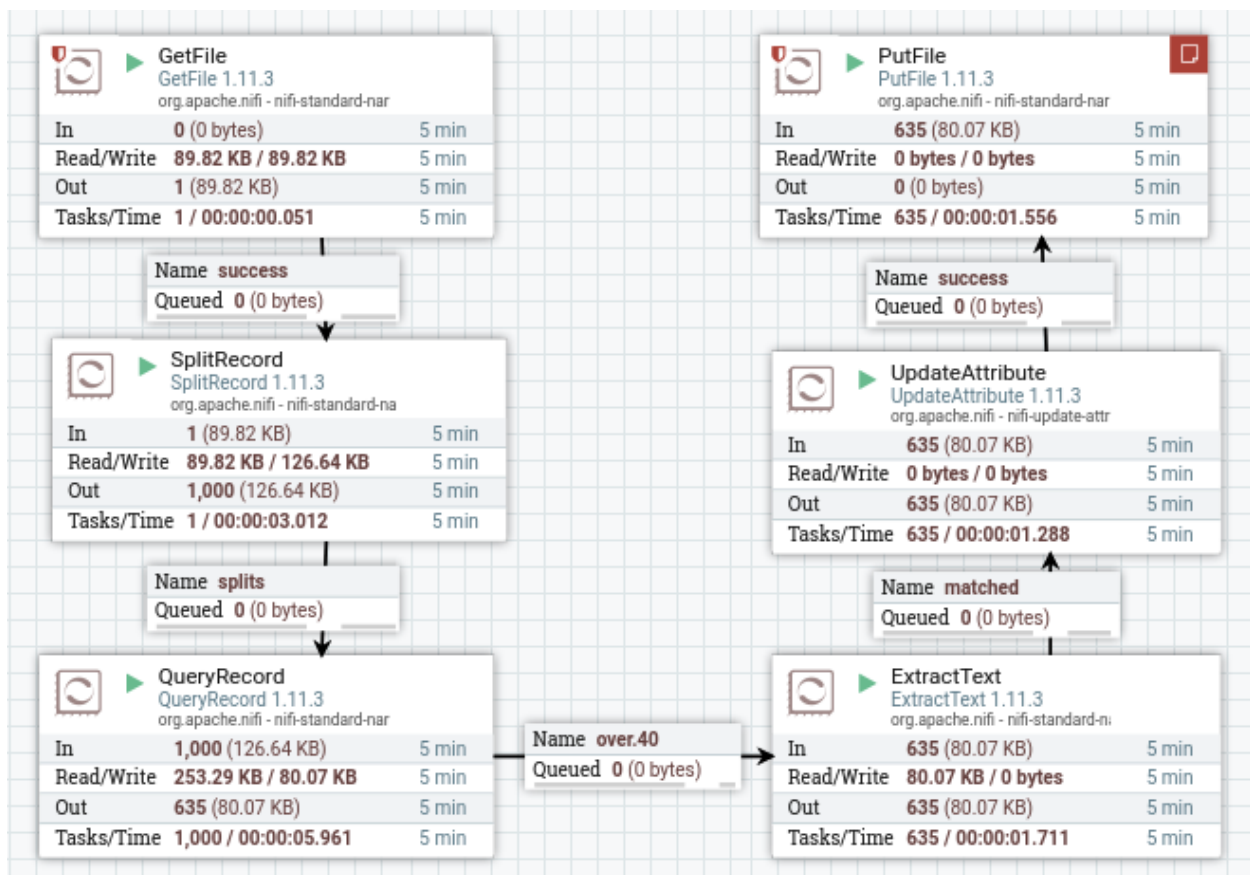
Close

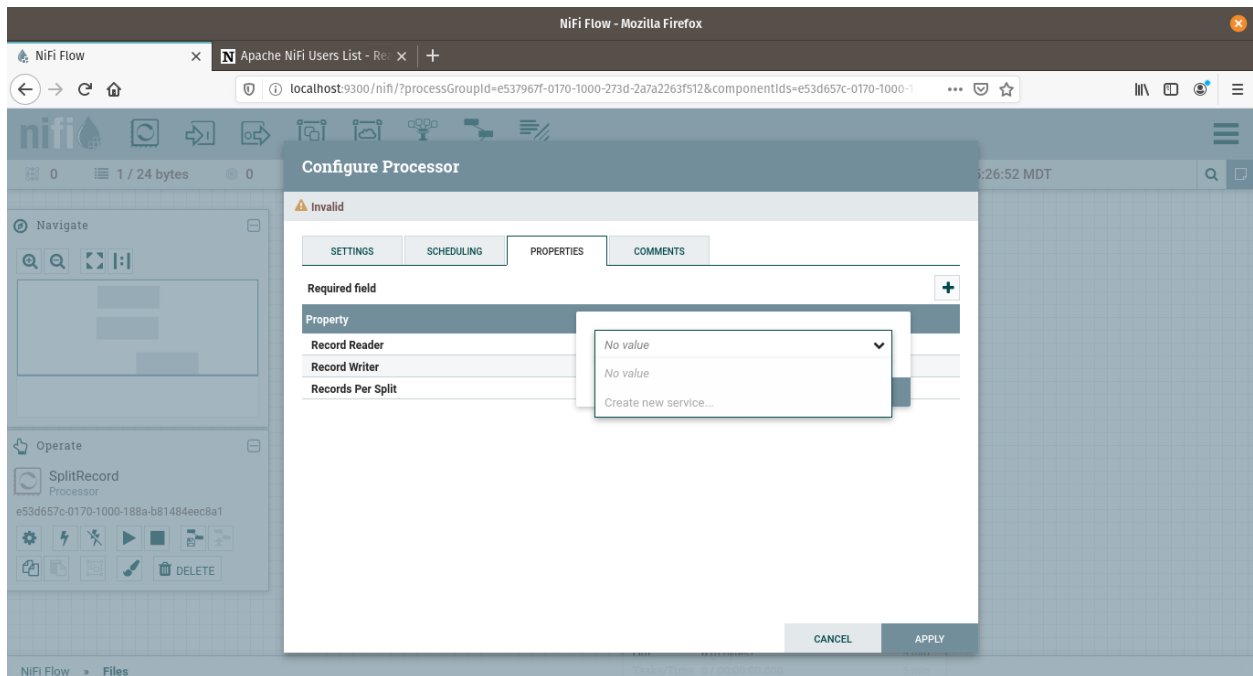
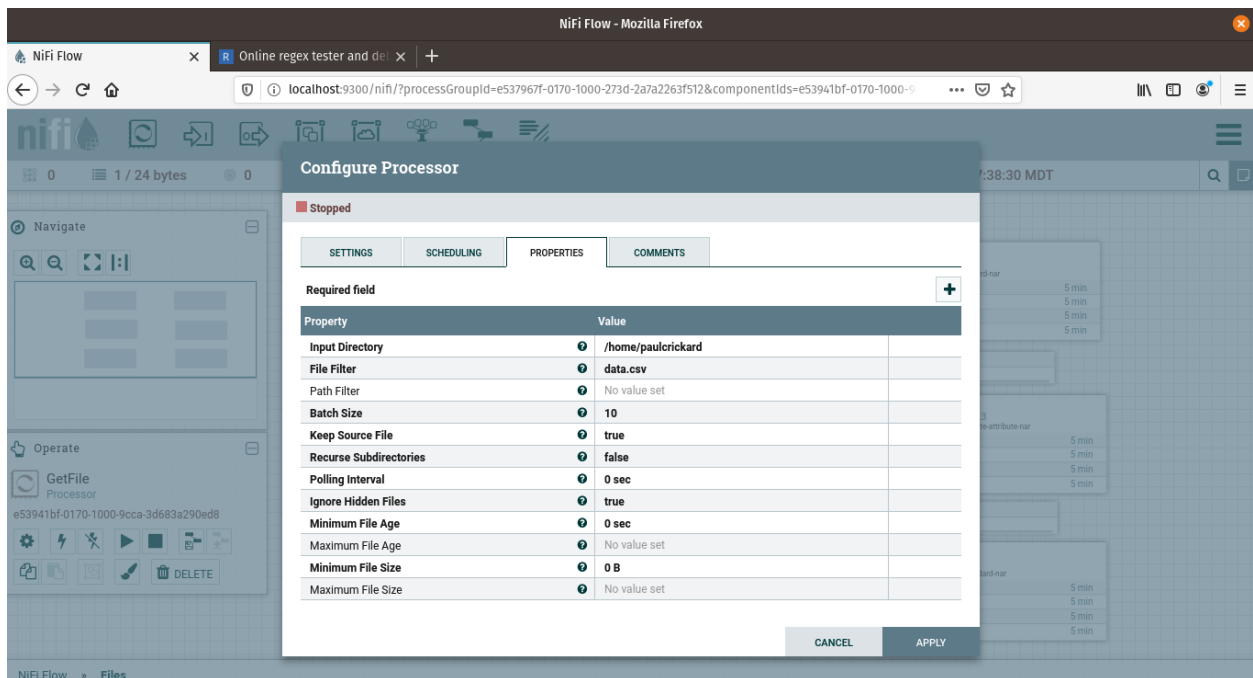

```
Airflow - DAGs - Mozilla Firefox

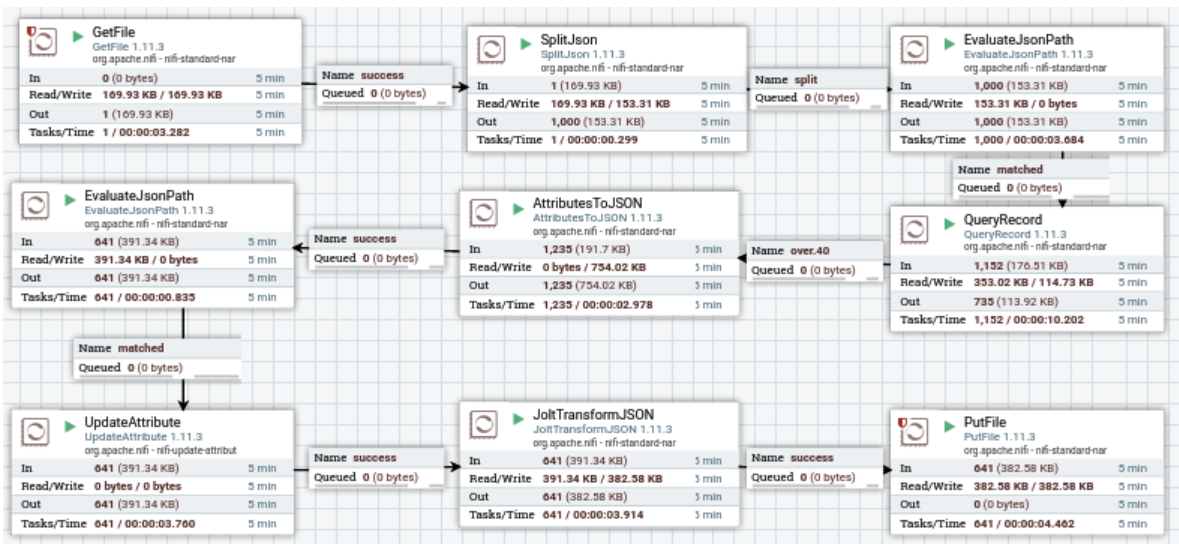
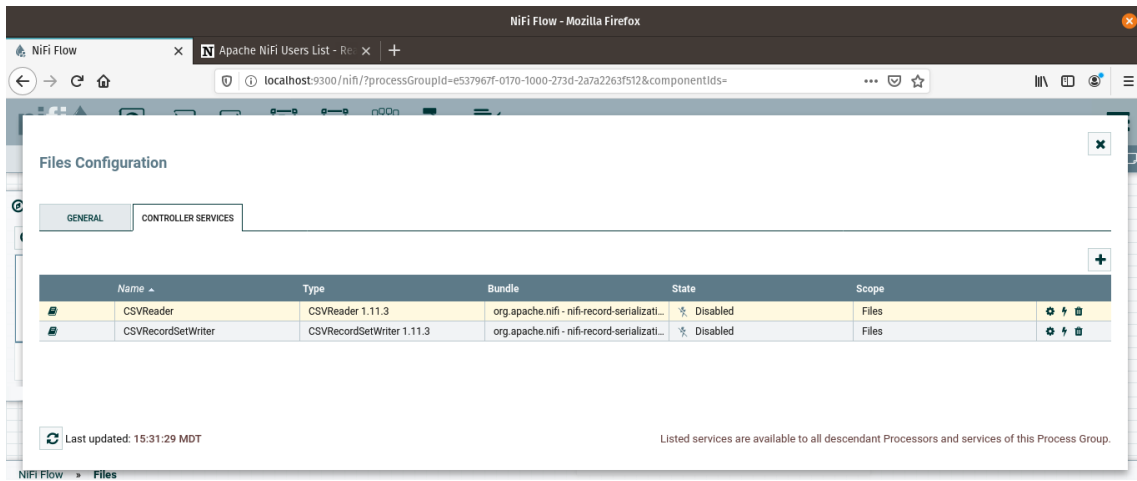
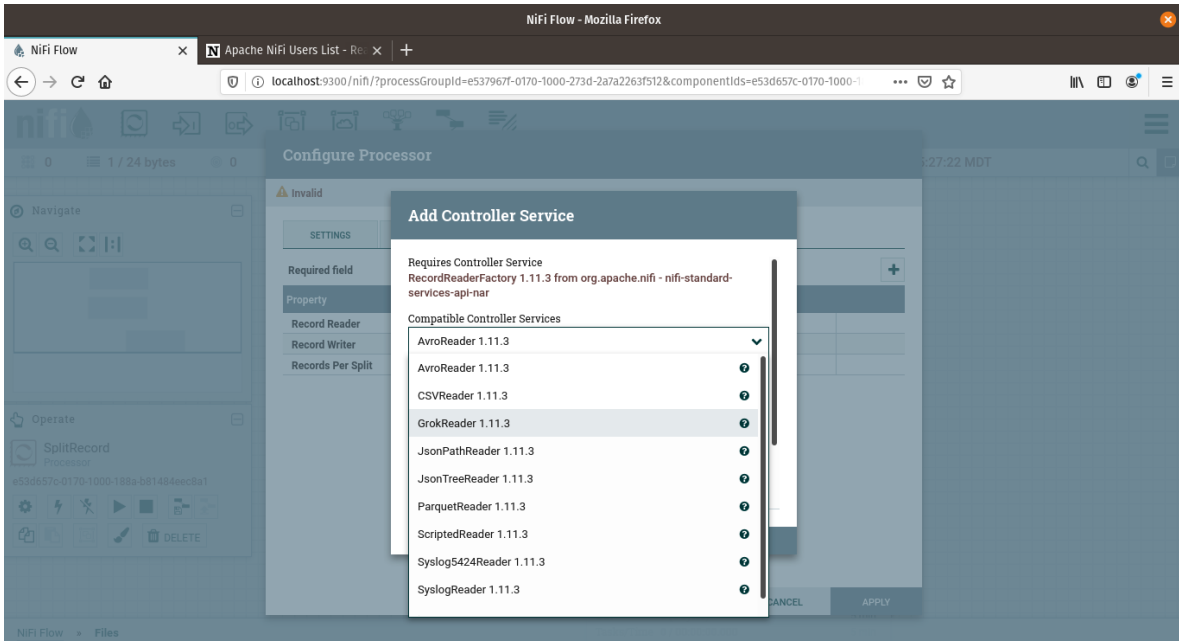
Airflow - DAGs
localhost:8080/admin/airflow/log?task_id=readCSV&dag_id=MyCSV DAG&execution_date=2020-03-18T00%3A00%3A00

Airflow DAGs Data Profiling Browse Admin Docs About 2020-03-18 18:07:26 UTC

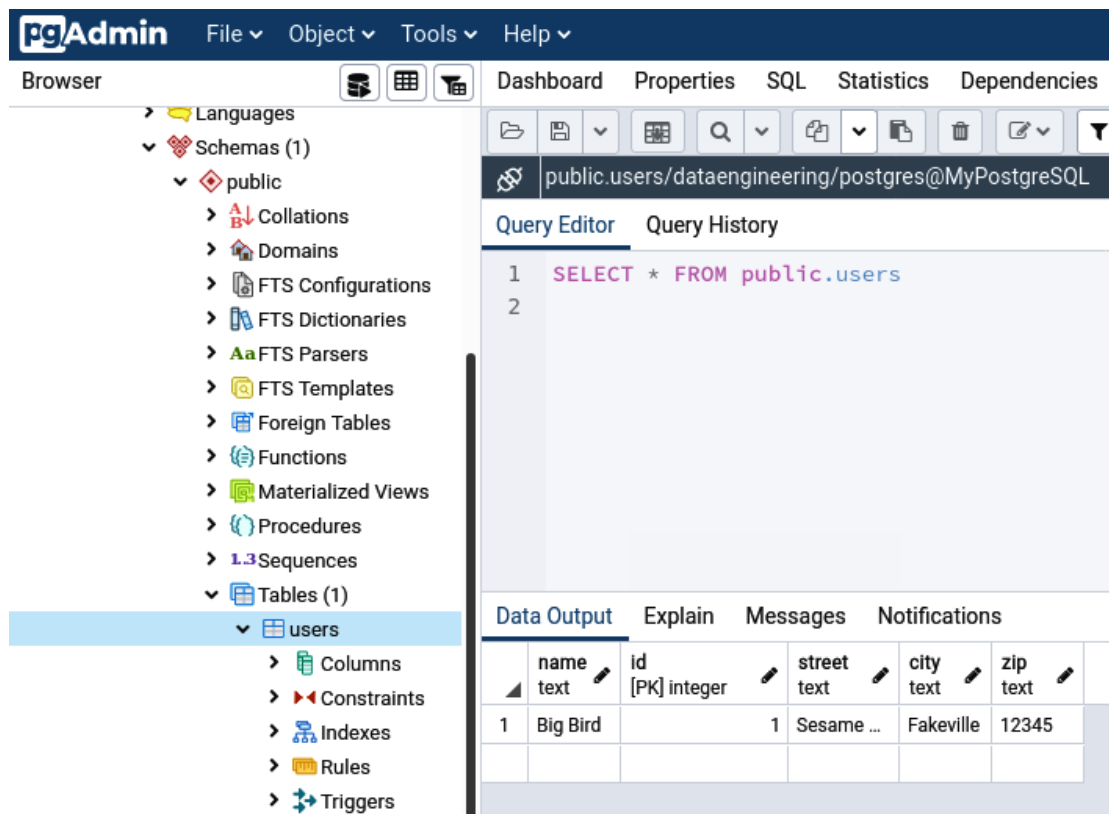
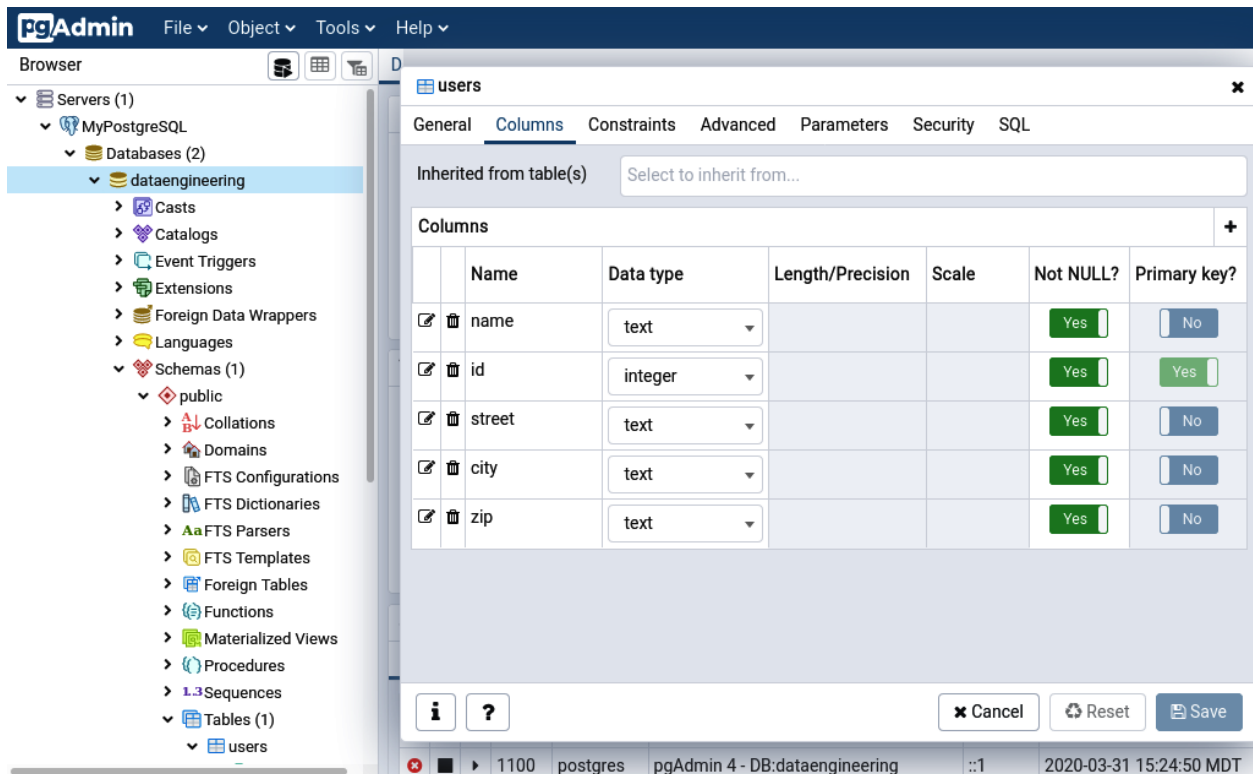
[2020-03-18 12:00:41,017] {taskinstance.py:887} INFO -
[2020-03-18 12:00:41,787] {taskinstance.py:887} INFO - Executing <Task(PythonOperator): readCSV> on 2020-03-18T00:00:00+00:00
[2020-03-18 12:00:41,792] {standard_task_runner.py:53} INFO - Started process 25895 to run task
[2020-03-18 12:00:42,169] {logging_mixin.py:112} INFO - Running %s on host %s <TaskInstance: MyCSV DAG.readCSV 2020-03-18T00:00:00+00:00 [running]> pop-os.localdomain
[2020-03-18 12:00:42,201] {logging_mixin.py:112} INFO - Grace Jackson
[2020-03-18 12:00:42,202] {logging_mixin.py:112} INFO - Arthur Garcia
[2020-03-18 12:00:42,202] {logging_mixin.py:112} INFO - Gary Valentine
[2020-03-18 12:00:42,202] {logging_mixin.py:112} INFO - Erin Mclean
[2020-03-18 12:00:42,202] {logging_mixin.py:112} INFO - Karen Hodges
[2020-03-18 12:00:42,202] {logging_mixin.py:112} INFO - Edgar Humphrey
[2020-03-18 12:00:42,203] {logging_mixin.py:112} INFO - Andrew Williamson
[2020-03-18 12:00:42,203] {logging_mixin.py:112} INFO - Michael Mack
[2020-03-18 12:00:42,203] {logging_mixin.py:112} INFO - Kristine Nielsen
[2020-03-18 12:00:42,203] {logging_mixin.py:112} INFO - Mark Grimes
[2020-03-18 12:00:42,203] {logging_mixin.py:112} INFO - Bruce Hardin
[2020-03-18 12:00:42,204] {logging_mixin.py:112} INFO - Margaret Price
[2020-03-18 12:00:42,204] {logging_mixin.py:112} INFO - Wesley Holloway
[2020-03-18 12:00:42,204] {logging_mixin.py:112} INFO - Bruce Thompson
[2020-03-18 12:00:42,204] {logging_mixin.py:112} INFO - Thomas Hall
[2020-03-18 12:00:42,205] {logging_mixin.py:112} INFO - Troy Oconnor
[2020-03-18 12:00:42,205] {logging_mixin.py:112} INFO - Brad Roman
[2020-03-18 12:00:42,205] {logging_mixin.py:112} INFO - Elizabeth Smith
[2020-03-18 12:00:42,205] {logging_mixin.py:112} INFO - Brad Morrison
[2020-03-18 12:00:42,205] {logging_mixin.py:112} INFO - David Miller
[2020-03-18 12:00:42,206] {logging_mixin.py:112} INFO - Jon Nichols
[2020-03-18 12:00:42,206] {logging_mixin.py:112} INFO - Colleen Walsh
[2020-03-18 12:00:42,206] {logging_mixin.py:112} INFO - Jessica Cooper
[2020-03-18 12:00:42,206] {logging_mixin.py:112} INFO - Brenda Oneal
[2020-03-18 12:00:42,206] {logging_mixin.py:112} INFO - Chris Johnson
[2020-03-18 12:00:42,207] {logging_mixin.py:112} INFO - Susan Fernandez
```







Chapter 4: Working with Databases



pgAdmin

File Object Tools Help postgres@localhost

Browser

public

- Collations
- Domains
- FTS Configurations
- FTS Dictionaries
- FTS Parsers
- FTS Templates
- Foreign Tables
- Functions
- Materialized Views
- Procedures
- Sequences
- Tables (1)
 - users
- Columns
- Constraints
- Indexes
- Rules
- Triggers
- Trigger Functions
- Types
- Views

postgres

- Login/Group Roles
- Tablespaces

Dashboard Properties SQL Statistics Dependencies Dependents public.users/dataengineering/postgres@MyPostgreSQL

Query Editor Query History

```
1 SELECT * FROM public.users
2
```

Data Output Explain Messages Notifications

	name text	id [PK] integer	street text	city text	zip text
1	Big Bird		1 Sesame Street	Fakeville	12345
2	Clarence Coffey		2 237 Danielle Spur	East Richard	83211
3	Margaret Alexander		3 41682 Wilson Square	Nelsonport	05479
4	Amy Gordon		4 4703 Vasquez Stream	Pattersonfort	76184
5	Sierra Smith		5 7428 Goodman Parkways Suite 306	West Ryanville	24869
6	Kevin Smith		6 575 Werner Summit	West Jenniferview	83115
7	Timothy Fitzpatrick		7 819 Boyd Glen Apt. 416	Leachhaven	72697
8	Emily Thomas		8 6721 Stephens Alley Apt. 203	Harrismouth	96156
9	Kathleen Smith		9 3371 Clay Court Suite 160	Morenoburgh	37626
10	Carrie Cruz		10 2503 Cheryl Keys	Joneshaven	22439
11	Shannon Johnson		11 3529 Amanda Drives	Jesushaven	18214
12	Elaine Molina		12 22948 Stephanie Hollow Apt. 242	Kennedyhaven	85691

Management / Index patterns

Elasticsearch

- Index Management
- Index Lifecycle Policies
- Rollup Jobs
- Transforms
- Remote Clusters
- Snapshot and Restore
- License Management
- 8.0 Upgrade Assistant

Kibana

- Index Patterns
- Saved Objects
- Spaces
- Reporting
- Advanced Settings

Index patterns ?

+ Create index pattern

Search...

Pattern ↑

kibana_sample_data_ecommerce Default

Rows per page: 10 < 1 >

[illegible]

Discover

frompostgresql* 2,208 hits

Search field names

Filter by type 0

Selected fields

<> _source

Available fields

- # Unnamed: 0
- t _id
- t _index
- # _score
- t _type
- t city
- t name

2,208 hits

_source

- Unnamed: 0: 0 name: Big Bird city: Fakeville _id: jja8QnEBxMEH3Xr-PgFM _type: doc _index: frompostgresql _score: 0
- Unnamed: 0: 1 name: Clarence Coffey city: East Richard _id: jza8QnEBxMEH3Xr-SQfP _type: doc _index: frompostgresql _score: 0
- Unnamed: 0: 2 name: Margaret Alexander city: Nelsonport _id: kDa8QnEBxMEH3Xr-SwdU _type: doc _index: frompostgresql _score: 0
- Unnamed: 0: 3 name: Amy Gordon city: Pattersonfort _id: kTa8QnEBxMEH3Xr-TAfC _type: doc _index: frompostgresql _score: 0
- Unnamed: 0: 4 name: Sierra Smith city: West Ryanville _id: kja8QnEBxMEH3Xr-Tgc8 _type: doc _index: frompostgresql _score: 0
- Unnamed: 0: 5 name: Kevin Smith city: West Jenniferview _id: kza8QnEBxMEH3Xr-Twes _type: doc _index: frompostgresql _score: 0
- Unnamed: 0: 6 name: Timothy Fitzpatrick city: Leachhaven _id: lDa8QnEBxMEH3Xr-UAeC _type: doc _index: frompostgresql _score: 0
- Unnamed: 0: 7 name: Emily Thomas city: Harrismouth _id: lTa8QnEBxMEH3Xr-UAfX _type: doc _index: frompostgresql _score: 0
- Unnamed: 0: 8 name: Kathleen Smith city: Morenoburgh _id: lja8QnEBxMEH3Xr-UQd- _type: doc _index: frompostgresql _score: 0
- Unnamed: 0: 9 name: Carrie Cruz city: Joneshaven _id: lza8QnEBxMEH3Xr-Ugdc _type: doc _index: frompostgresql _score: 0
- Unnamed: 0: 10 name: Shannon Johnson city: Jesushaven _id: mDa8QnEBxMEH3Xr-Uwca _type: doc _index: frompostgresql _score: 0

nifi

0 1,166 / 147.51 KB 0 0 2 19 0 0 0 0 0 0 0 0 0 0 20:33:48 MDT

Navigate

Operate

postgresToelasticseach

42f4030d-0171-1000-166d-3a3bdb8d120b

ExecuteSQLRecord

ExecuteSQLRecord 1.11.3

org.apache.nifi - nifi-standard-nar

In 0 (0 bytes) 5 min

Read/Write 0 bytes / 233.05 KB 5 min

Out 4 (233.05 KB) 5 min

Tasks/Time 4 / 00:00:00.512 5 min

Name success

Queued 0 (0 bytes)

SplitText

SplitText 1.11.3

org.apache.nifi - nifi-standard-nar

In 4 (233.05 KB) 5 min

Read/Write 233.05 KB / 0 bytes 5 min

Out 4,004 (229.15 KB) 5 min

Tasks/Time 4 / 00:00:02.670 5 min

Name splits

Queued 0 (0 bytes)

PutElasticsearchHttp

PutElasticsearchHttp 1.11.3

org.apache.nifi - nifi-elasticsearch-nar

In 4,004 (229.15 KB) 5 min

Read/Write 229.15 KB / 0 bytes 5 min

Out 0 (0 bytes) 5 min

Tasks/Time 44 / 00:00:36.778 5 min

NIFI Flow > postgresToelasticseach

localhost:9300/nifi/?processGroupId=42f4030d-0171-1000-166d-3a3bdb8d120b&componentId=42f4a28f-0171-1000

Configure Processor

Invalid

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field

Property	
Database Connection Pooling Service	No value
SQL Pre-Query	No value
SQL select query	
SQL Post-Query	Create new service...
Max Wait Time	0 seconds
Record Writer	No value set
Normalize Table/Column Names	false
Use Avro Logical Types	false
Max Rows Per Flow File	0
Output Batch Size	0
Fetch Size	0

CANCEL APPLY

NIFI Flow > postgresToelasticsearch

localhost:9300/nifi/?processGroupId=42f4030d-0171-1000-166d-3a3bdb8d120b&componentId=42f4a28f-0171-1000

Controller Service Details

SETTINGS PROPERTIES COMMENTS

Required field

Property	Value
Database Connection URL	jdbc:postgresql://localhost/dataengineering
Database Driver Class Name	org.postgresql.Driver
Database Driver Location(s)	/home/paulcrickard/nifi-1.11.3/drivers/postgresql-42.2...
Kerberos Credentials Service	No value set
Database User	postgres
Password	Sensitive value set
Max Wait Time	500 millis
Max Total Connections	8
Validation query	No value set
Minimum Idle Connections	0
Max Idle Connections	8
Max Connection Lifetime	-1
Time Between Eviction Runs	-1

OK

postgresToelasticsearch Config

GENERAL CONTROLLER SERVICES

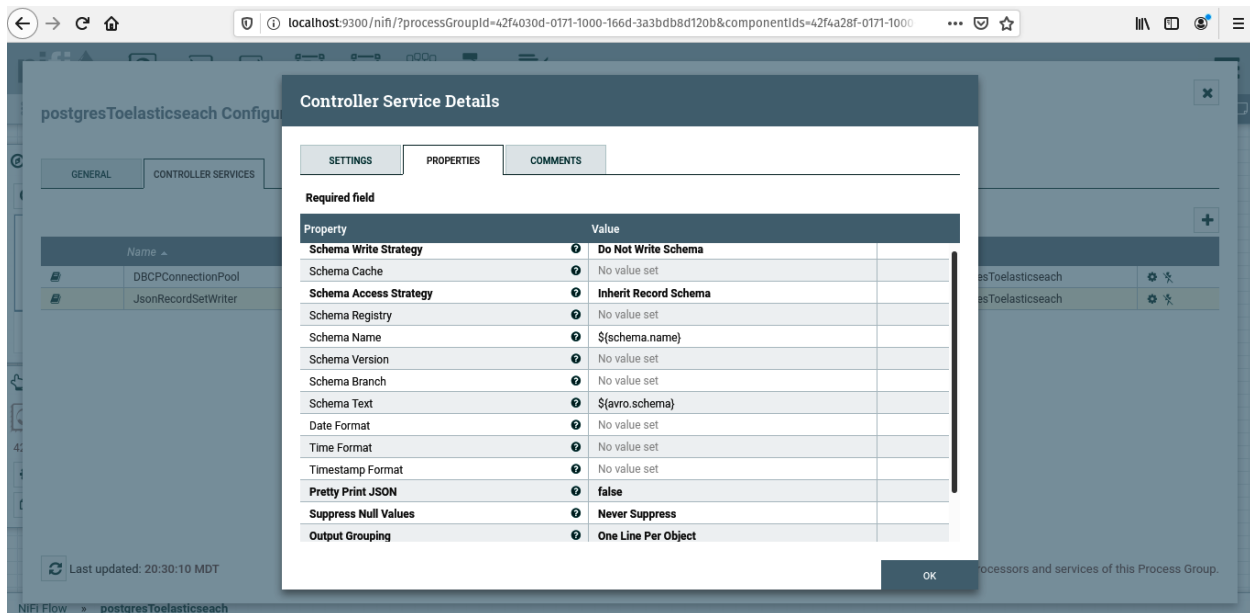
Name

DBCPConnectionPool

JsonRecordSetWriter

Last updated: 20:30:10 MDT

NIFI Flow > postgresToelasticsearch

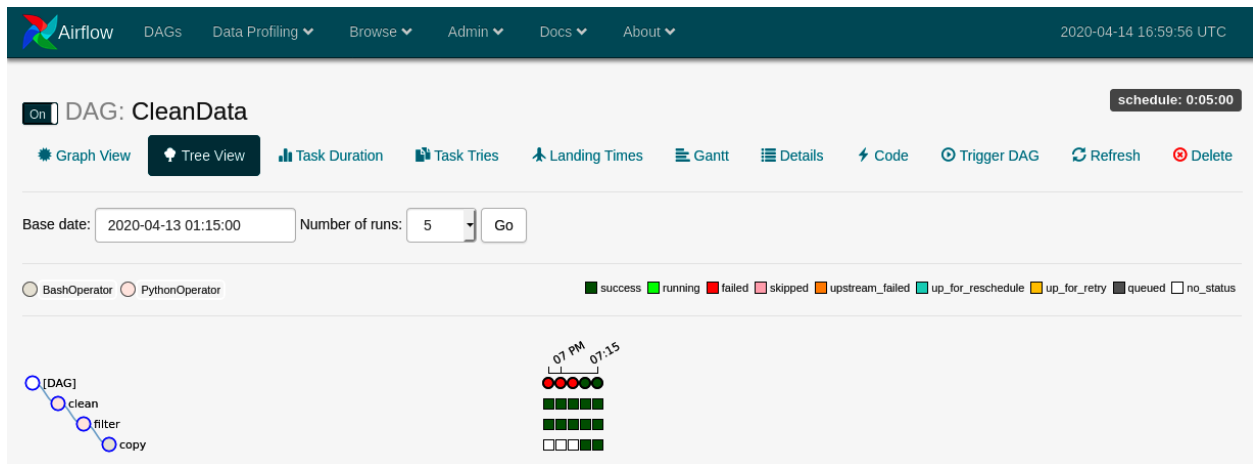


```

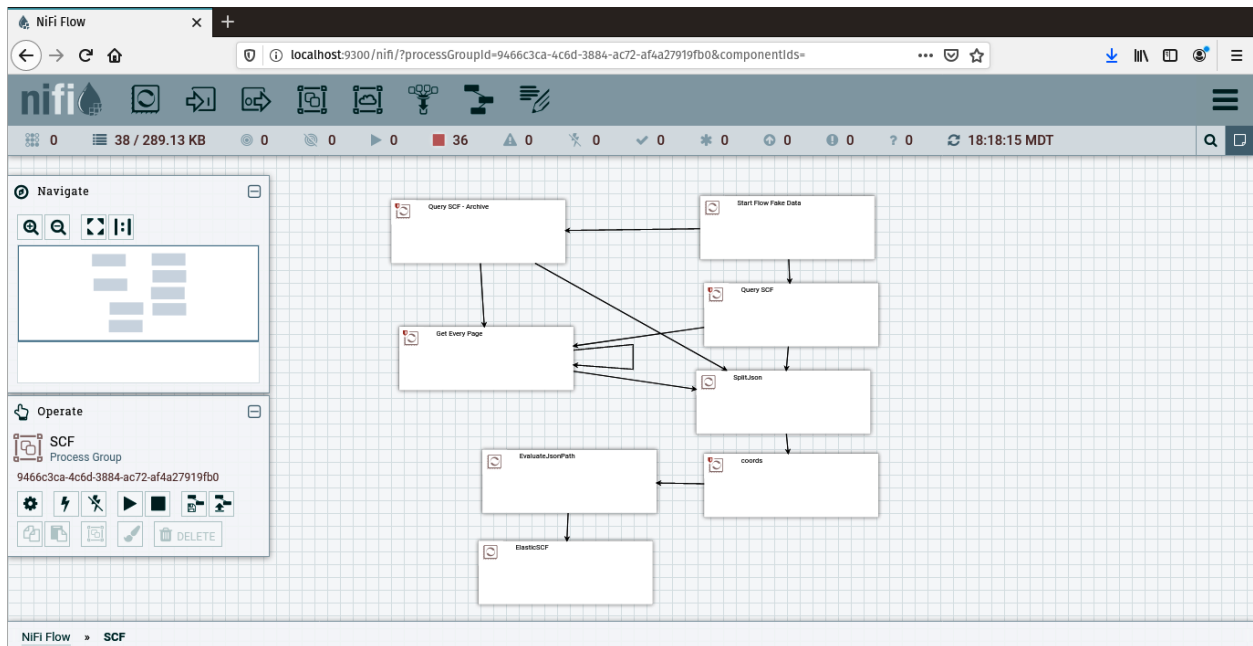
yellow open frompostgresql u3EVJEFJR-eDPtrufvSNkA 1 1 2208 0 250.4kb 250.4kb
yellow open fromnifi NBAL_aLdQ1y4iK90kDFauw 1 1 1001 0 233.8kb 233.8kb
yellow open test VwacjWq_S9a0fW6l-YbcIA 1 1 1 0 3.9kb 3.9kb
green open kibana_sample_data_ecommerce -6sdGU13T12GIjE2rU_ZlQ 1 0 4675 0 4.9mb 4.9mb
green open .kibana_task_manager_1 1nLCVInfTa6XmmifoGntXA 1 0 2 0 13kb 13kb
green open .apm-agent-configuration 51toZCyRS2yWlnLJZWF5-A 1 0 0 0 283b 283b
green open .kibana_1 g5gxKq0xR0y71FCG0rKiMw 1 0 66 5 993.3kb 993.3kb
yellow open users 8K7wOmQ4S90gxh2TqBybKw 1 1 1003 0 286.6kb 286.6kb

```


Chapter 5: Cleaning and Transforming Data



Chapter 6: Building a 311 Data Pipeline



The screenshot shows the Kibana Dev Tools console with the following content:

Console Search Profiler Grok Debugger

History Settings Help

```
1 PUT scf
2 {
3   "mappings": {
4     "properties": {
5       "coords": {
6         "type": "geo_point"
7       }
8     }
9   }
10 }
```

```
1 {
2   "acknowledged" : true,
3   "shards_acknowledged" : true,
4   "index" : "scf"
5 }
6
```


NiFi Flow

localhost:9300/nifi/?processGroupId=9466c3ca-4c6d-3884-ac72-af4a27919fb0&componentId=e483c112-d08e...

Configure Processor

Stopped

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field +

Property	Value
File Size	0B
Batch Size	1
Data Format	Text
Unique FlowFiles	false
Custom Text	No value set
Character Set	UTF-8

CANCEL APPLY

Query SCF - Archi
ExecuteScript 1.11.3
org.apache.nifi - nifi-script

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

Name suc
Queued 0

Kibana

localhost:5601/app/kibana#/management/kibana/index_pattern?_g=()

Management / Index patterns / Create index pattern

Create index pattern

Kibana uses index patterns to retrieve data from Elasticsearch indices for things like visualizations. ☐ Include system indices

Step 1 of 2: Define index pattern

Index pattern

scf*

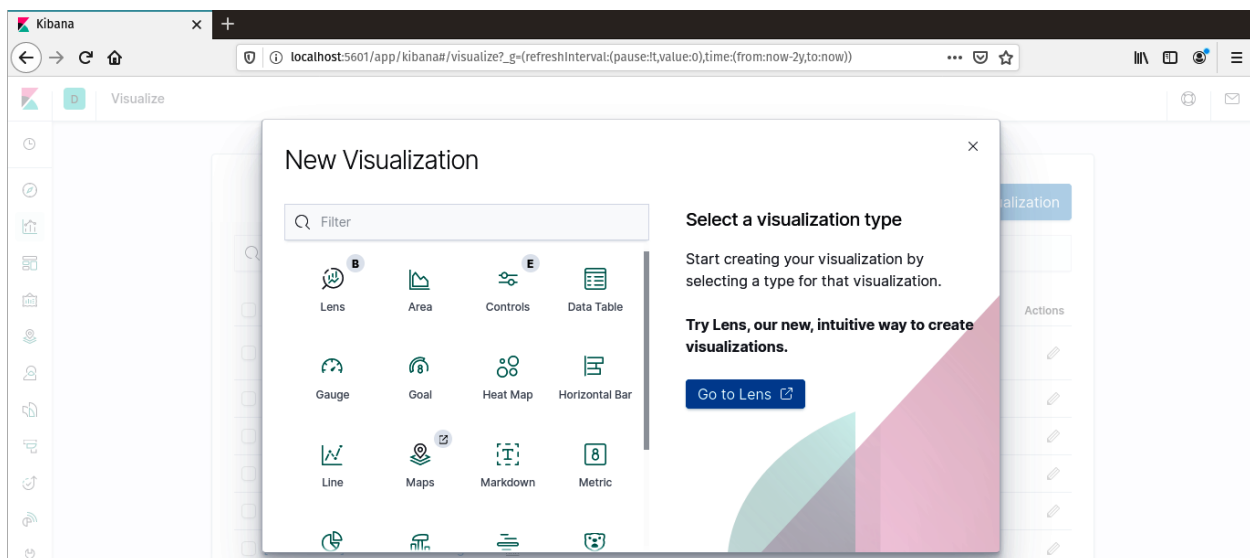
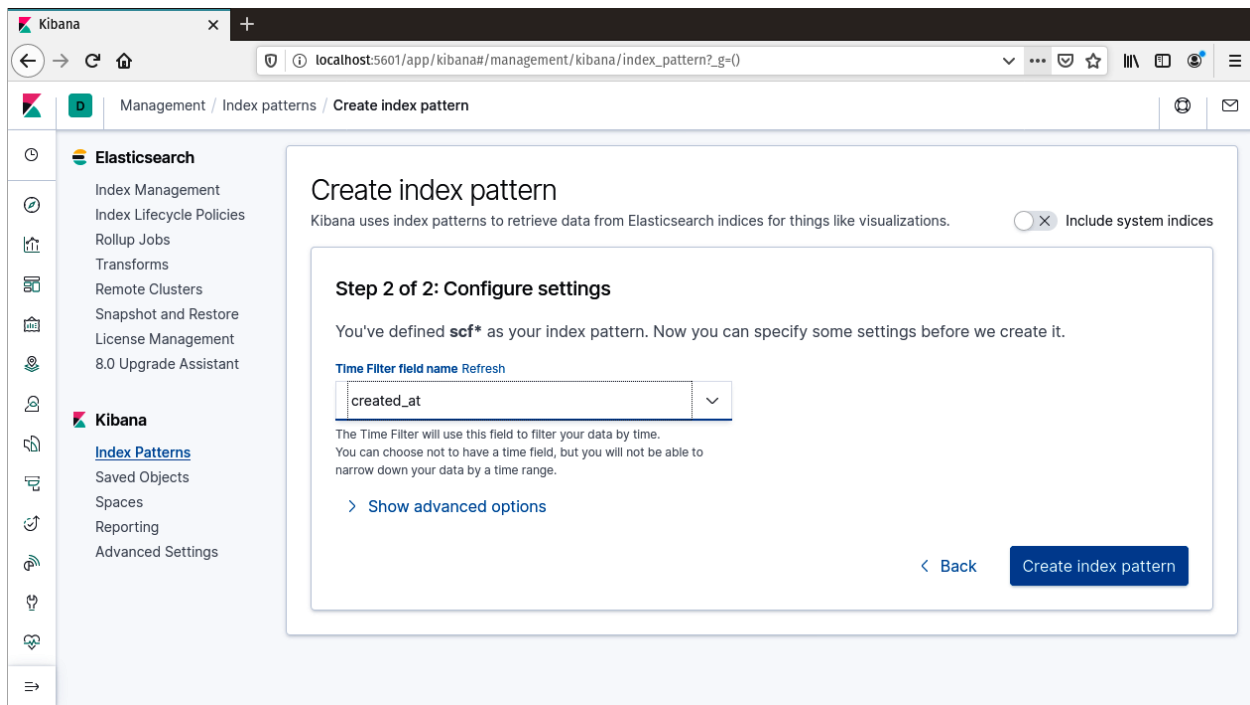
You can use a * as a wildcard in your index pattern.
You can't use spaces or the characters \, /, ?, *, <, >, |.

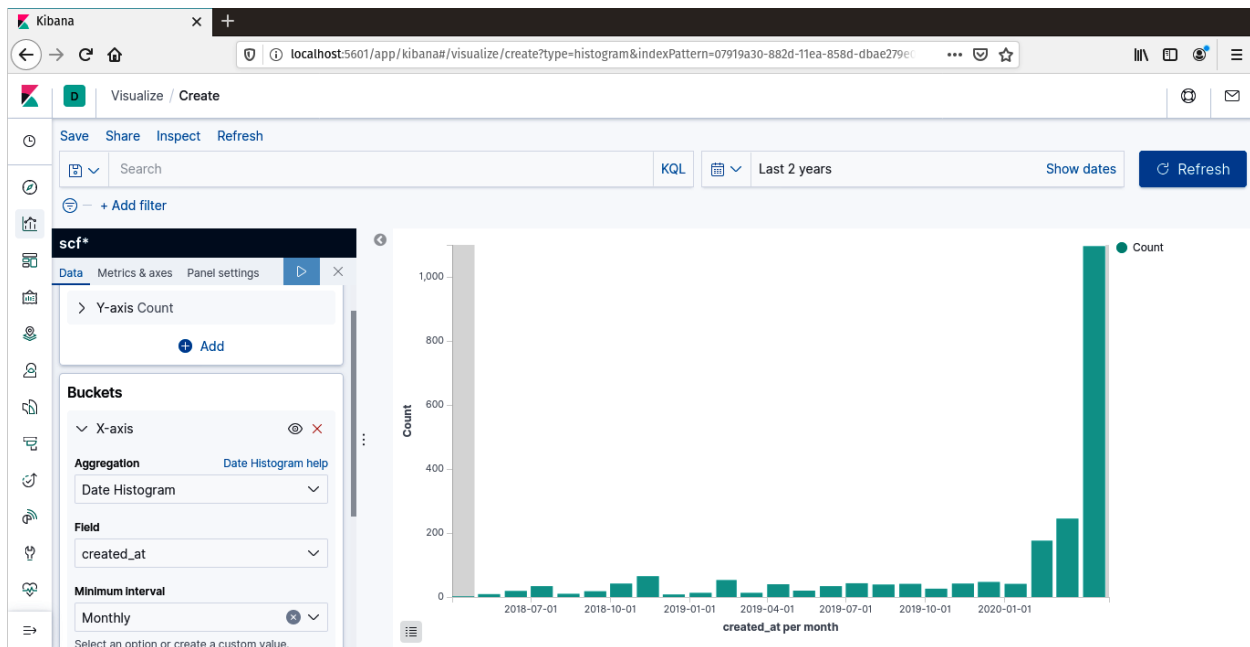
✓ **Success!** Your index pattern matches 1 index.

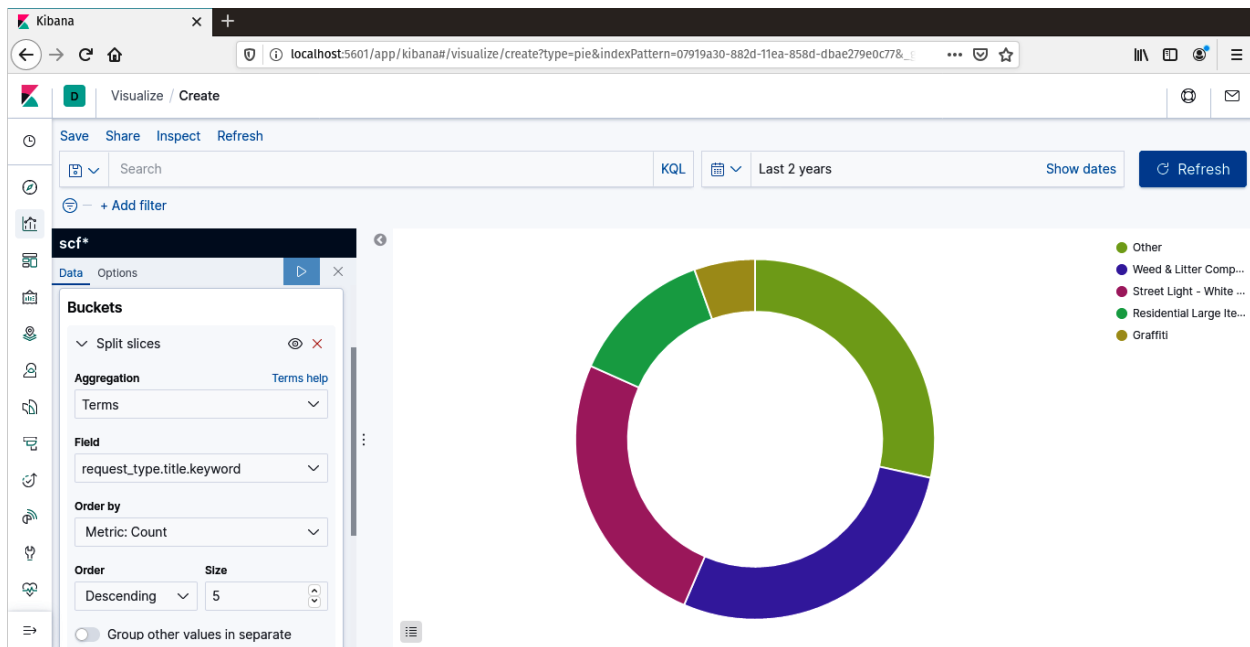
scf

Rows per page: 10

> Next step







Kibana

Visualize / Create

Save Share Inspect Refresh

Off Refresh

Markdown

311 Dashboard

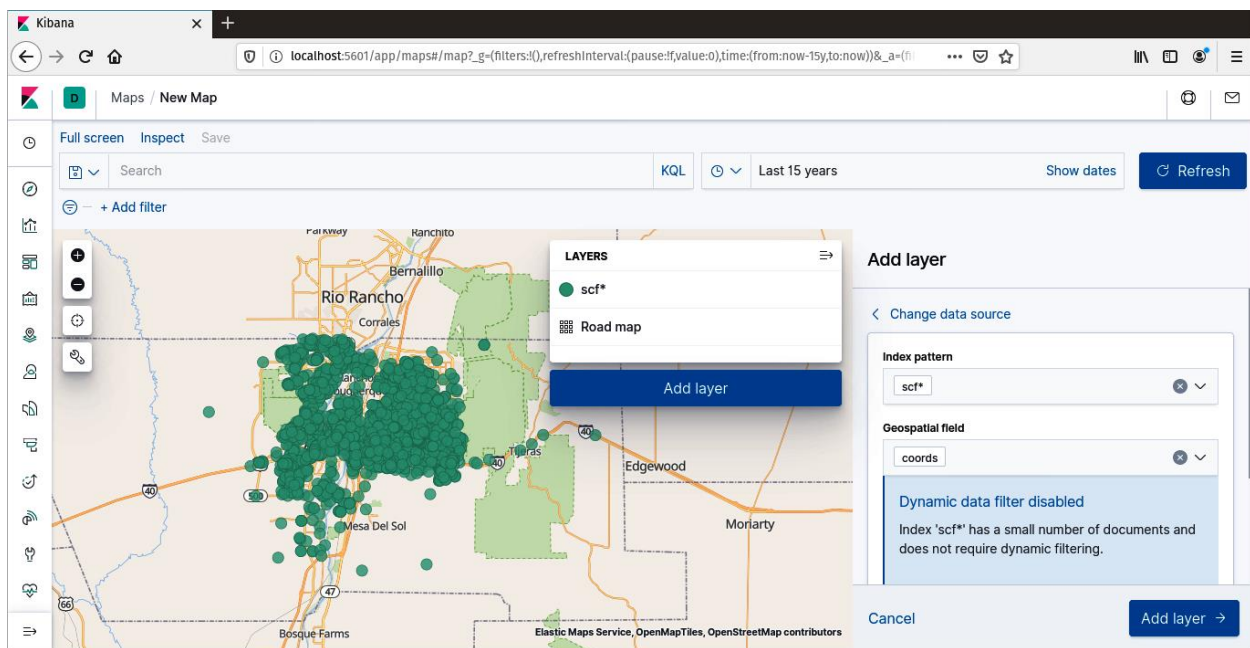
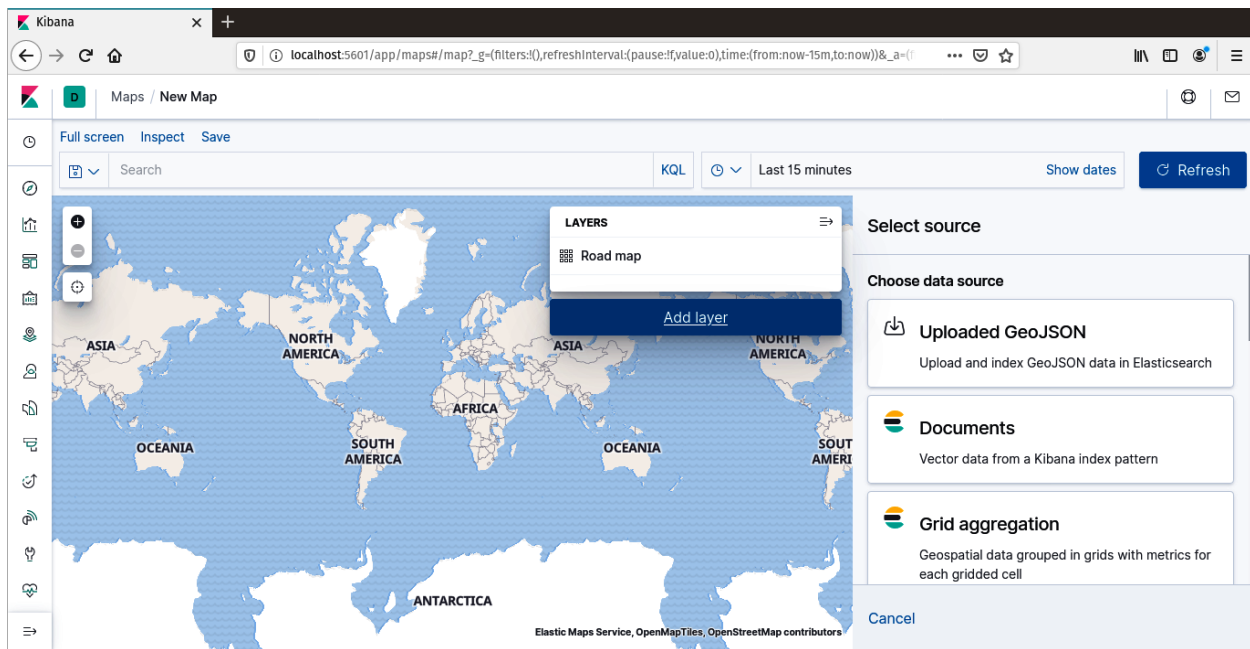
This dashboard displays data collected by Nifi at 8 hour intervals. It queries the SeeClickFix API, and was backfilled with Archived data.

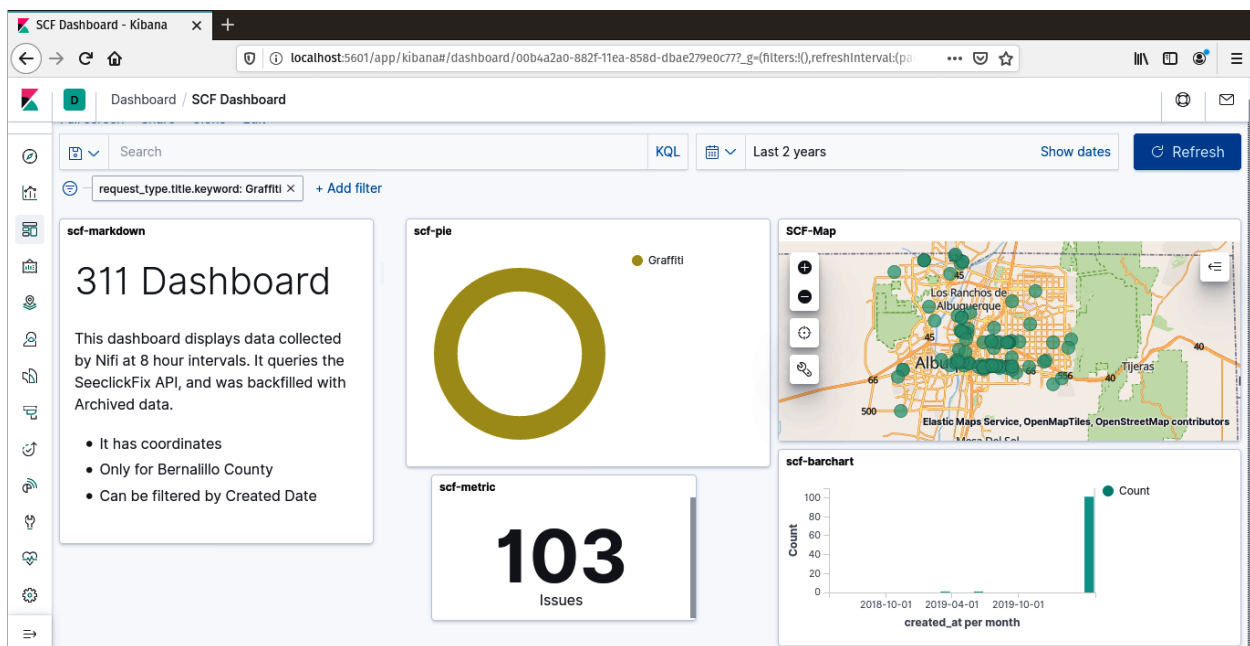
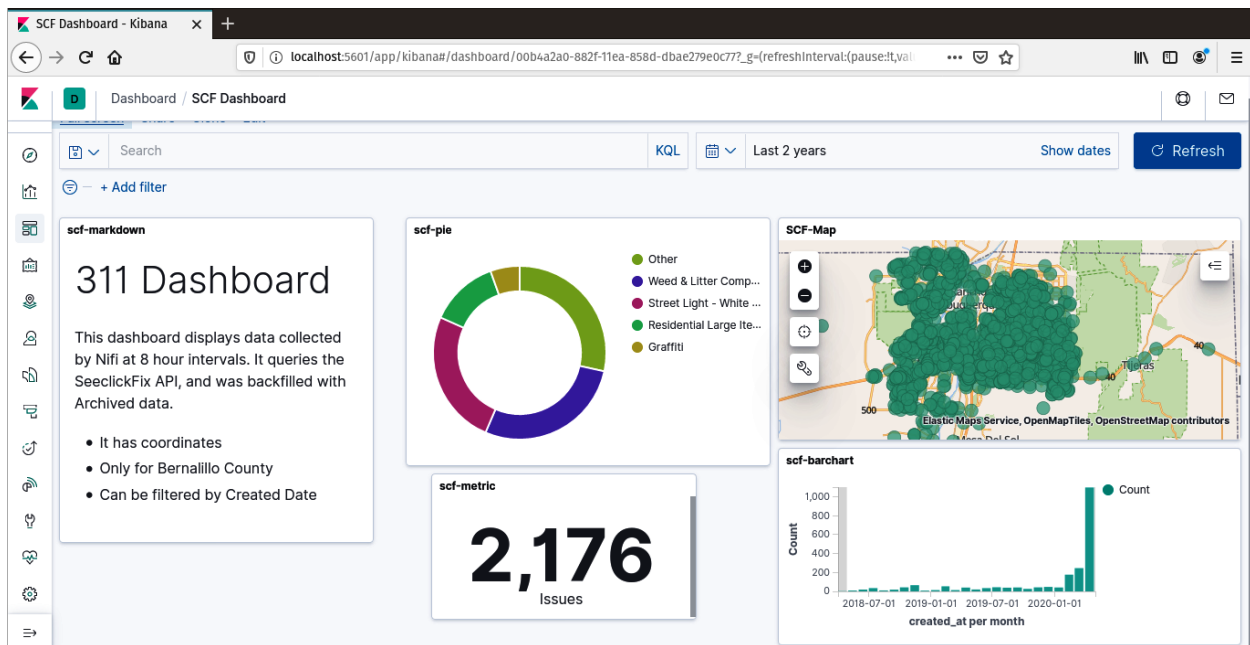
- * It has coordinates
- * Only for Bernalillo County
- * Can be filtered by Created Date

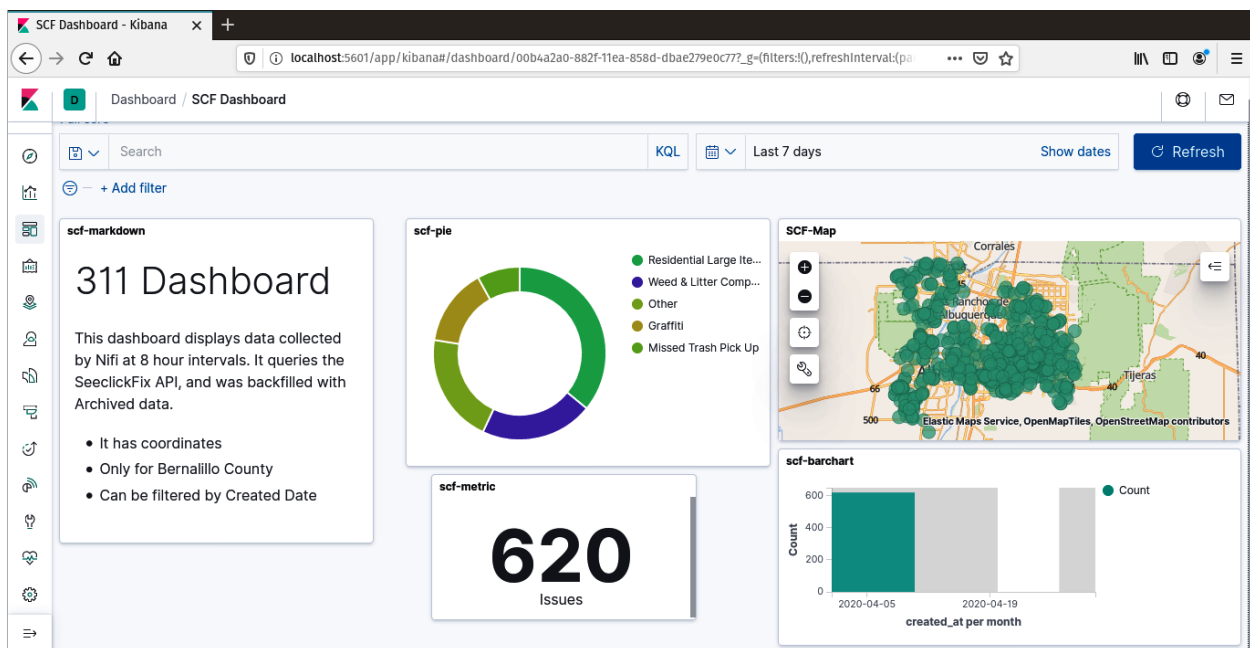
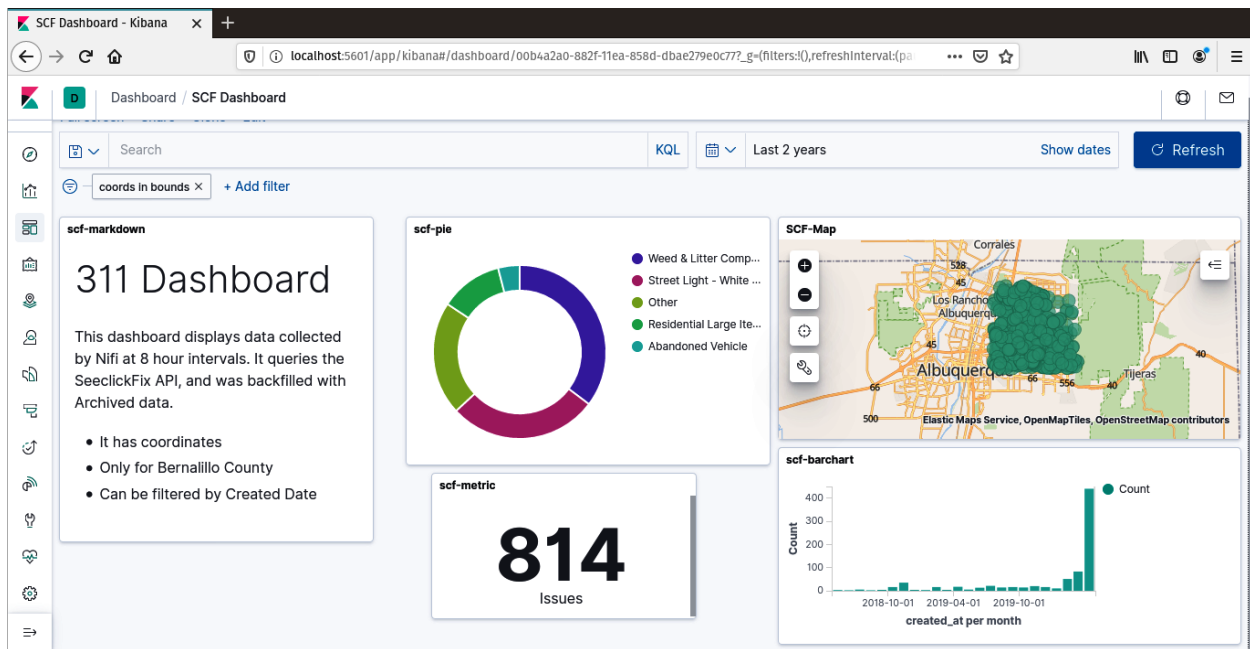
311 Dashboard

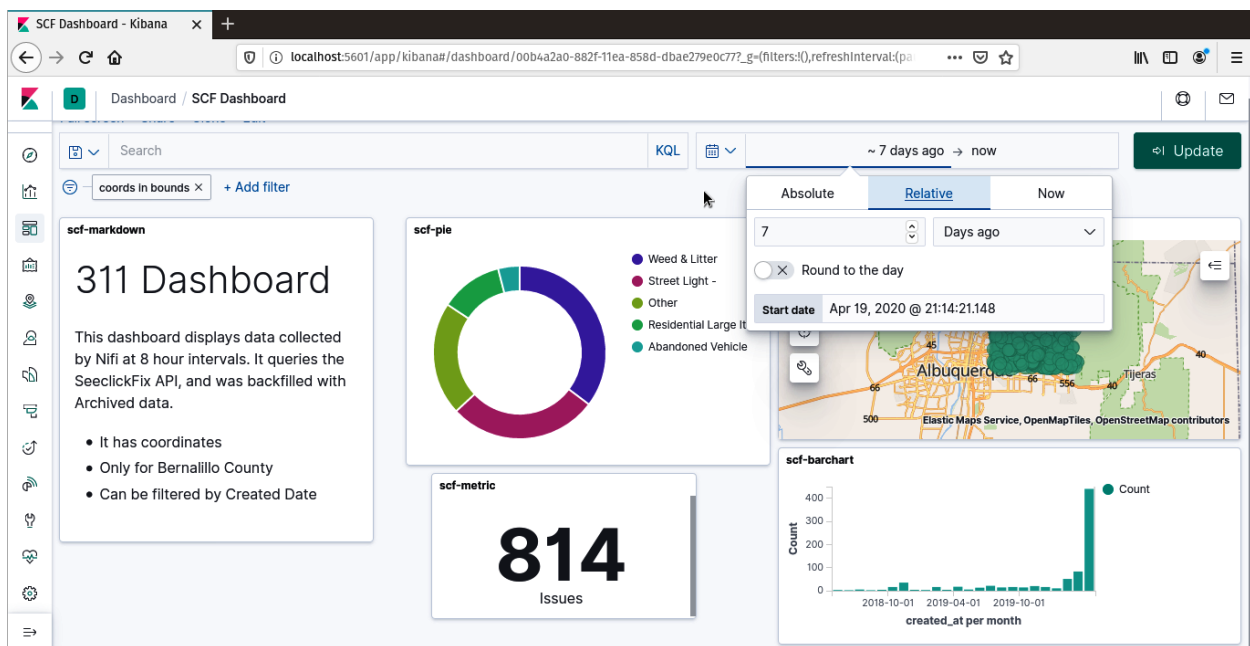
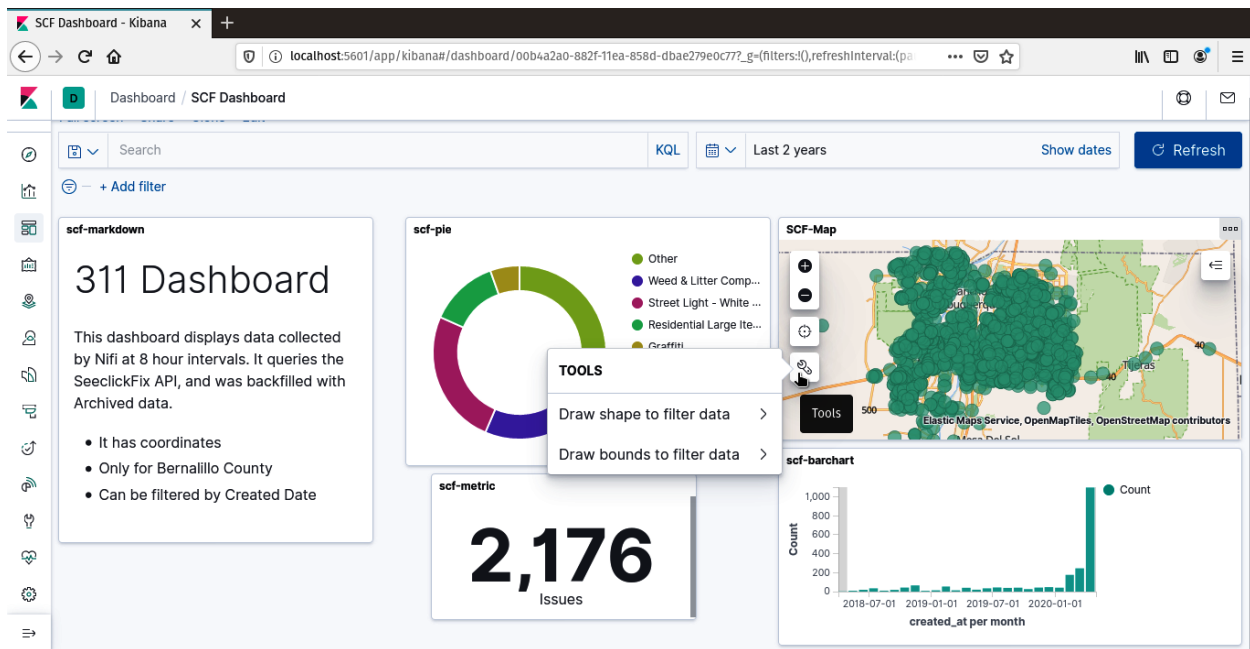
This dashboard displays data collected by Nifi at 8 hour intervals. It queries the SeeClickFix API, and was backfilled with Archived data.

- It has coordinates
- Only for Bernalillo County
- Can be filtered by Created Date

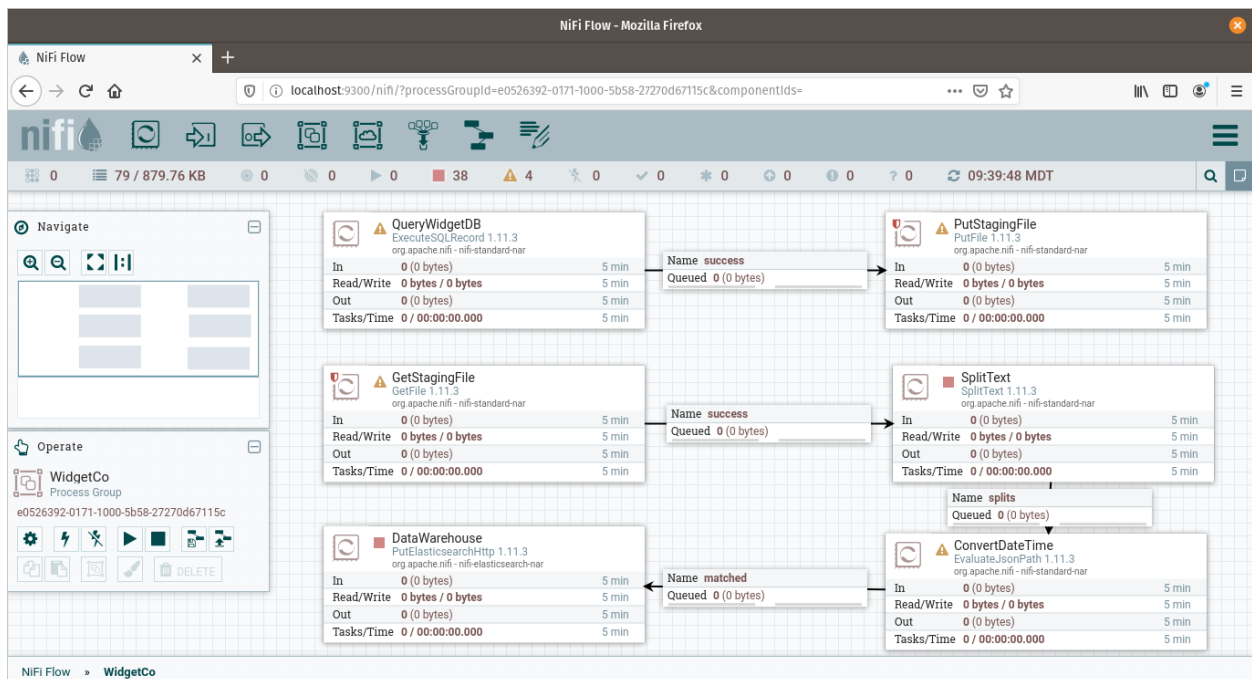
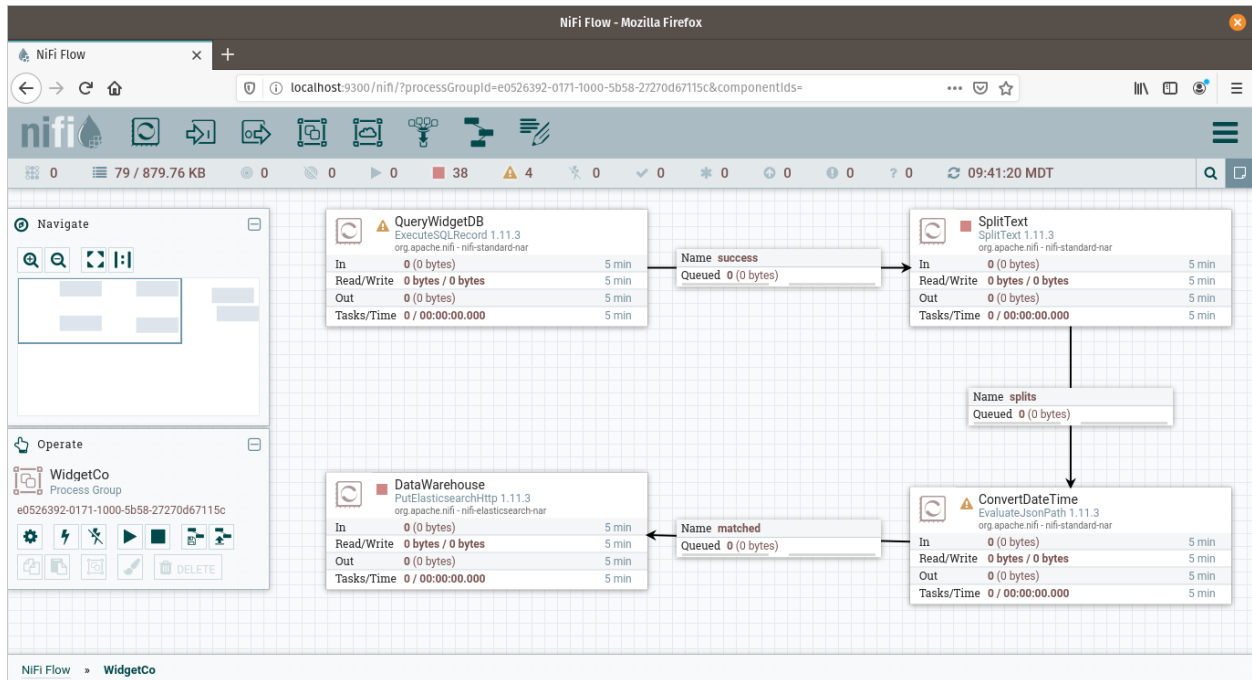


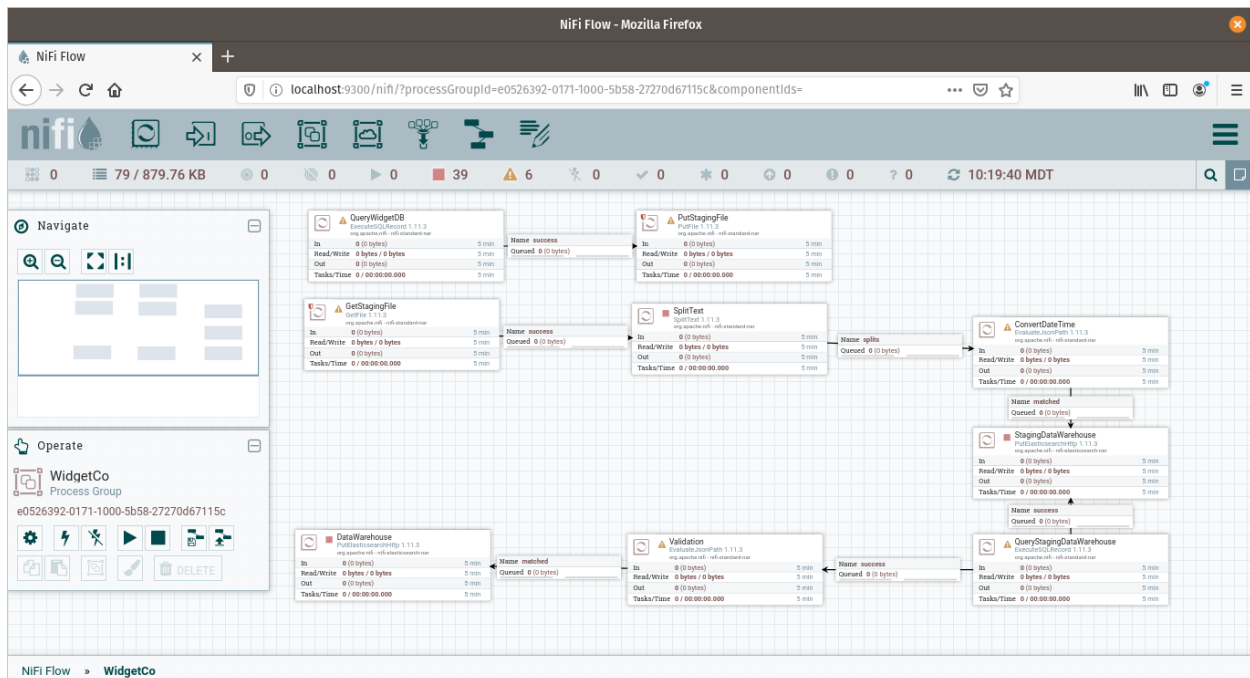






Chapter 7: Features of a Production Data Pipeline





Home | Great Expectations - Mozilla Firefox

https://greatexpectations.io

great_expectations

Documentation Blog Community

Greetings! Have any questions about using Great Expectations? Join us on [Slack](#)


Welcome to Great Expectations

Always know what to expect from your data

Great Expectations helps data teams eliminate pipeline debt, through data testing, documentation, and profiling.

[Join us on GitHub](#) ★ 1725

We're open source. Get involved!



great_expectations

Home / people.validate / 20200504T195933.614976Z / 6f1eb7a06079eb9cab8de404c6faab62

Expectation Validation Result

Evaluates whether a batch of data matches expectations.

Actions

Validation Filter:

Show All

Failed Only

How to Edit This Suite

Show Walkthrough

Table of Contents

Overview

Expectation Suite: people.validate

Status: ✓ Succeeded

Statistics

Evaluated Expectations	11
Successful Expectations	11
Unsuccessful Expectations	0
Success Percent	100%

Show more info...

Table-Level Expectations

Status	Expectation	Observed Value
✓	Must have between 900 and 1100 rows.	1000
✓	Must have exactly 8 columns.	8

Status	Expectation	Observed Value
✓	Must have between 900 and 1100 rows.	1000
✓	Must have exactly 8 columns.	8
✓	Must have these columns in this order: name, age, street, city, state, zip, lng, lat	['name', 'age', 'street', 'city', 'state', 'zip', 'lng', 'lat']

age

Status	Expectation	Observed Value
✓	values must never be null.	100% not null
✓	minimum value must be between 17 and 19.	18
✓	maximum value must be between 79 and 81.	80
✓	mean must be between 49.151 and 51.151.	50.151

edit_people.validate - Jupyter Notebook - Mozilla Firefox

edit_people.validate - Jupyter X +

localhost:8888/notebooks/edit_people.validate.ipynb

jupyter edit_people.validate (unsaved changes) Logout

File Edit View Insert Cell Kernel Help Not Trusted Kernel

Edit Your Expectation Suite

Use this notebook to recreate and modify your expectation suite:

Expectation Suite Name: people.validate

We'd love it if you [reach out to us on the Great Expectations Slack Channel](#)

```
In [ ]: from datetime import datetime
import great_expectations as ge
import great_expectations.jupyter_ux
from great_expectations.data_context.types.resource_identifiers import (
    ValidationResultIdentifier,
)

context = ge.data_context.DataContext()

# Feel free to change the name of your suite here. Renaming this will not
# remove the other one.
expectation_suite_name = "people.validate"
suite = context.get_expectation_suite(expectation_suite_name)
suite.expectations = []

batch_kwargs = {
    "datasource": "files_datasource",
    "path": "/home/paulcrickard/peoplepipeline/people.csv",
    "reader_method": "read_csv",
}
```

Table Expectation(s)

```
In [ ]: batch.expect_table_row_count_to_be_between(max_value=1100, min_value=900)
```

```
In [ ]: batch.expect_table_column_count_to_equal(value=8)
```

```
In [ ]: batch.expect_table_columns_to_match_ordered_list(
    column_list=["name", "age", "street", "city", "state", "zip", "lng", "lat"]
)
```

age

```
In [ ]: batch.expect_column_values_to_not_be_null("age")
```

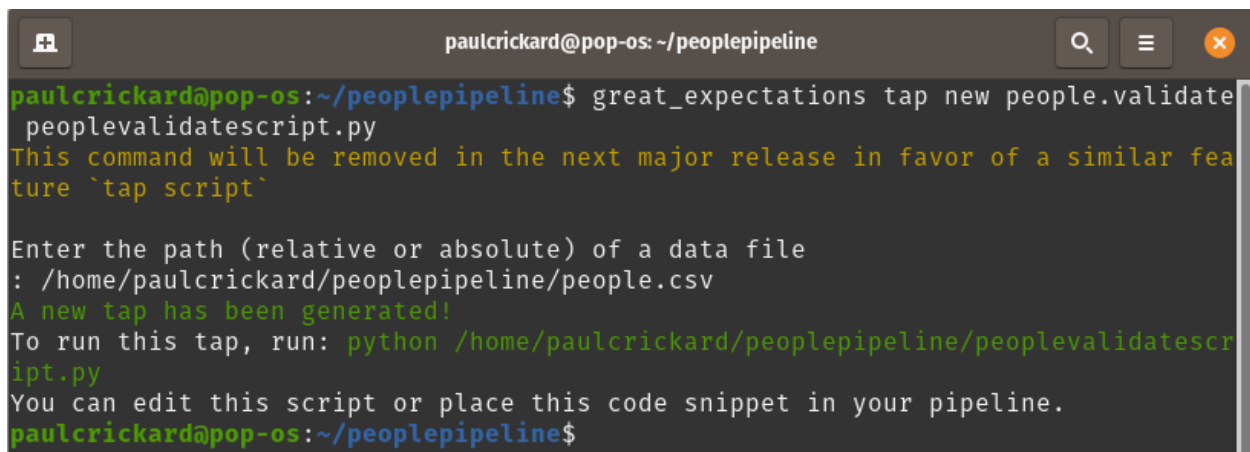
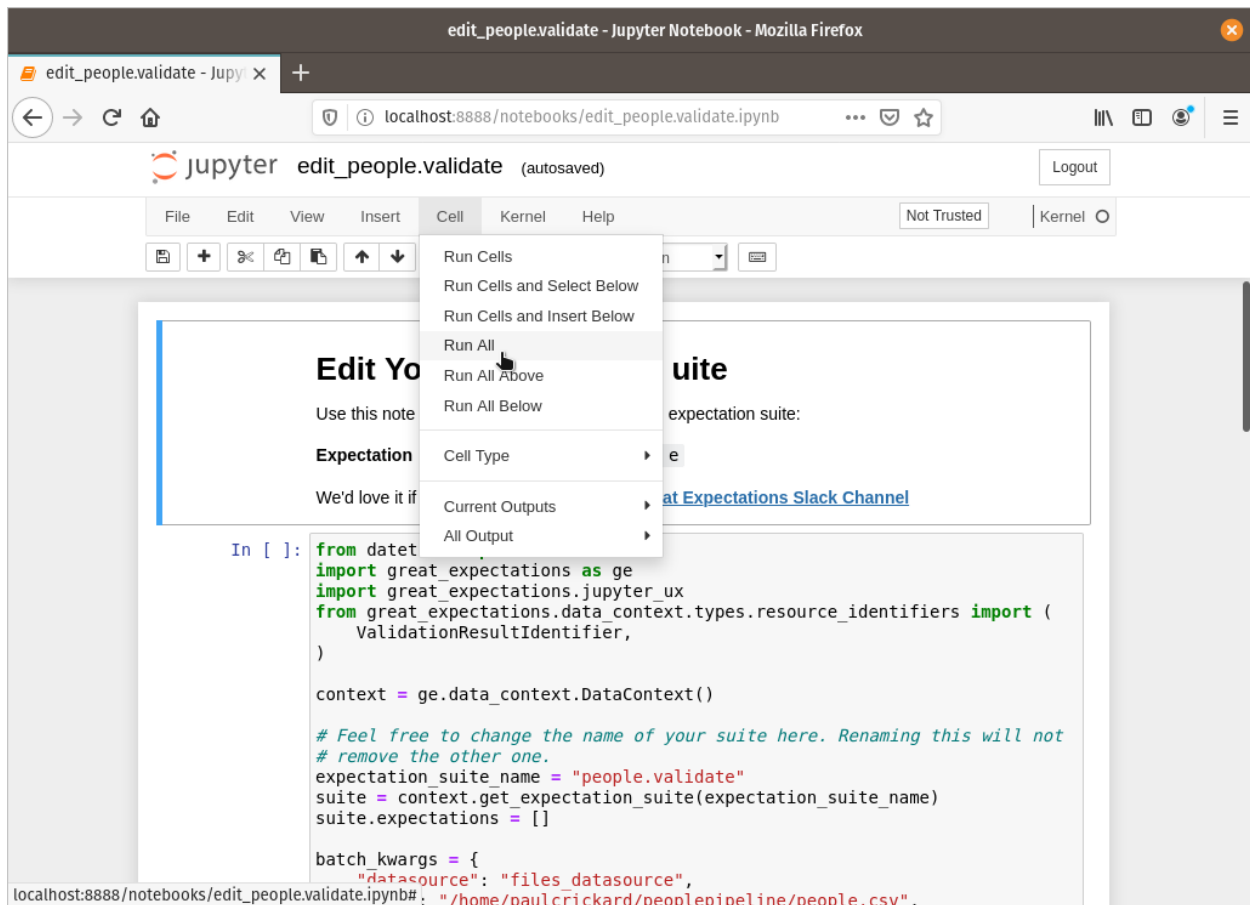
```
In [ ]: batch.expect_column_min_to_be_between("age", max_value=19, min_value=17)
```

```
In [ ]: batch.expect_column_max_to_be_between("age", max_value=81, min_value=79)
```

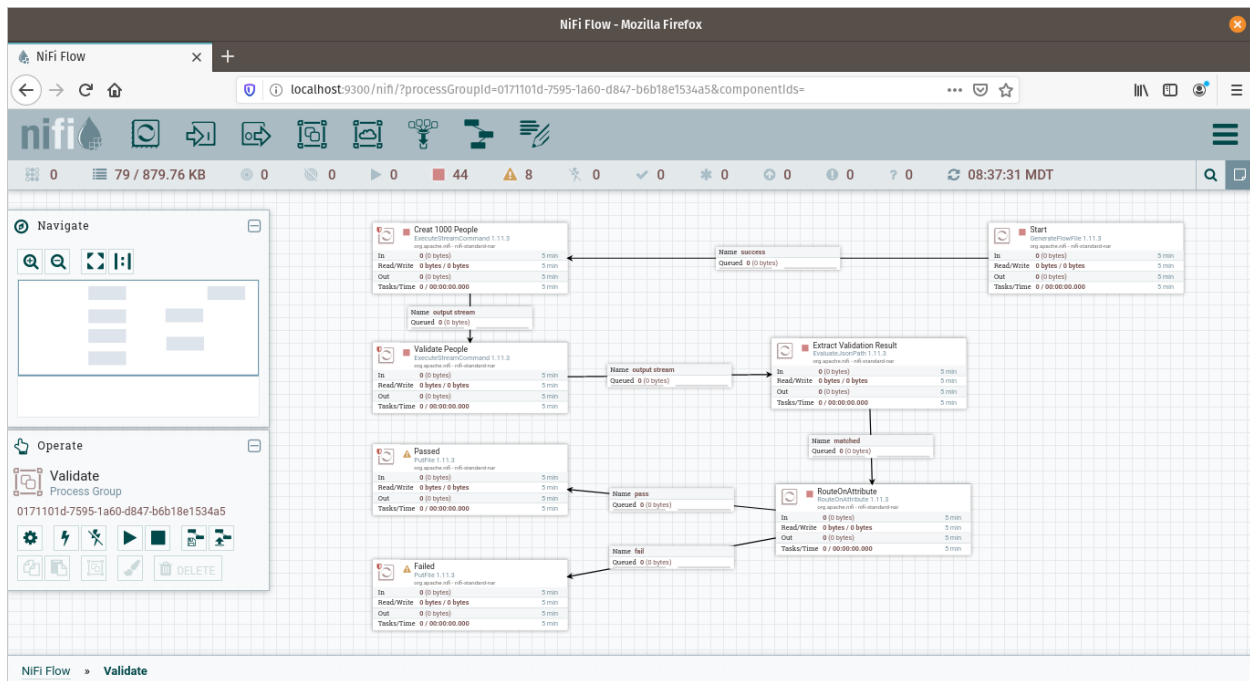
```
In [ ]: batch.expect_column_mean_to_be_between("age", max_value=51.151, min_value=49.151)
```


```
In [ ]: batch.expect_column_median_to_be_between("age", max_value=52.0, min_value=50.0)
```

```
In [ ]: batch.expect_column_quantile_values_to_be_between(
    "age",
    quantile_ranges={
        "quantiles": [0.05, 0.25, 0.5, 0.75, 0.95],
        "value_ranges": [[21, 23], [34, 36], [50, 52], [64, 66], [76, 78]],
    },
)
```

```
paulcrickard@pop-os:~/peoplepipeline$ python3 peoplevalidatescript.py
Validation Succeeded!
```



View as: original

1
2 {"status": "complete"}



View as: original

1
2 {"result": "pass"}

Typical Workflow — great_e ×

edit_people.validate - Jupy ×


Data Docs created by Great ×

+

← → ↺ 🏠

file:///home/paulcrickard/peoplepipeline/great_expectations/uncommitted/data_docs/local_site/index.html

⋮ 📄 ⌕ 🌐 ☰

 great_expectations

Data Docs autogenerated using Great Expectations.

Actions

Validation Filter:

Show All

Failed Only

Show Walkthrough

Data Docs | local_site

Expectation Suite

Validation Results (run_id)

people.validate

- ✗ 20200505T145722.862661Z
- ✗ 20200505T145528.788769Z
- ✓ 20200505T144948.314177Z
- ✓ 20200505T144559.641868Z
- ✓ 20200505T021322.913702Z
- ✓ 20200505T020553.174834Z
- ✓ 20200505T012823.671147Z
- ✓ 20200505T011831.865106Z
- ✓ 20200504T224236.055072Z
- ✓ 20200504T204504.344872Z
- ✓ 20200504T195933.614976Z

Typical Workflow — great_e ×

edit_people.validate - Jupy ×


Data documentation compi ×

+

← → ↺ 🏠

file:///home/paulcrickard/peoplepipeline/great_expectations/uncommitted/data_docs/local_site/validations/people.validate/20200505T145722.862661Z/6f1eb7a06079eb9cab8de404c6faab62

⋮ 📄 ⌕ 🌐 ☰

 great_expectations

Home / people.validate / 20200505T145722.862661Z / 6f1eb7a06079eb9cab8de404c6faab62

Actions

Validation Filter:

Show All

Failed Only

✍ How to Edit This Suite

Show Walkthrough

Table of Contents

[Overview](#)

[Table-Level Expectations](#)

[age](#)

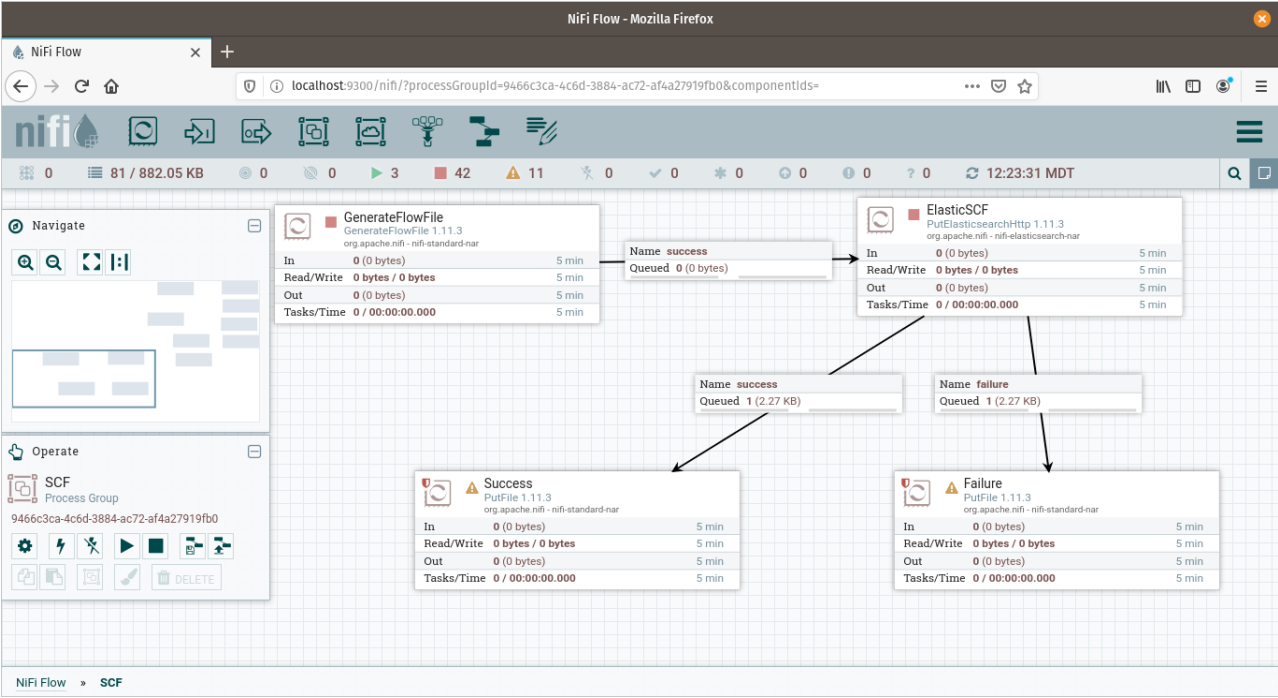
[name](#)

age

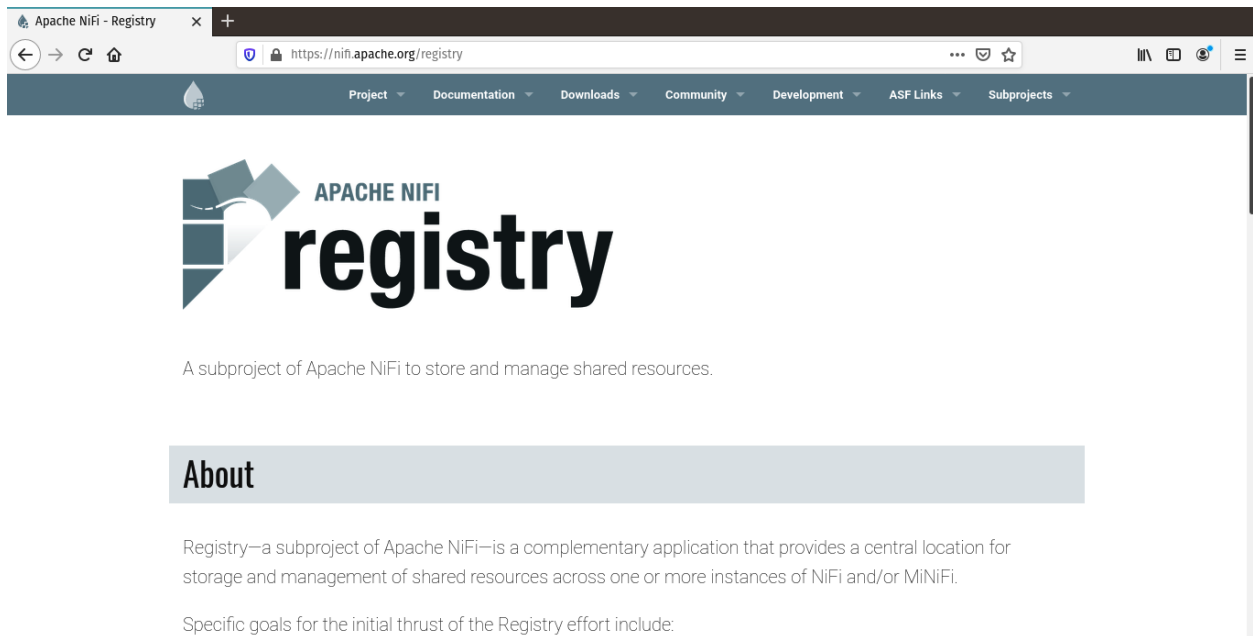
Status	Expectation	Observed Value
✓	values must never be null.	100% not null
✗	minimum value must be between 17 and 19.	1
✗	maximum value must be between 79 and 81.	100
✓	mean must be between 49.151 and 51.151.	51.111
✓	median must be between 50.0 and 52.0.	51

name

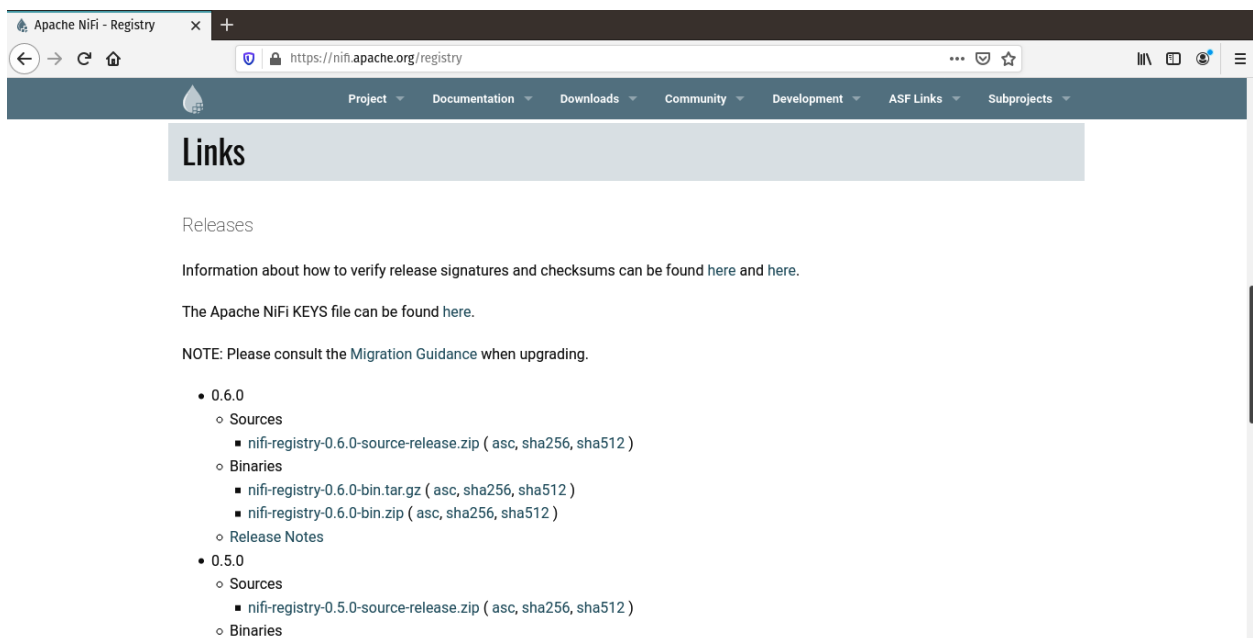
Status	Expectation	Observed Value
✓	values must never be null.	100% not null



Chapter 8: Version Control Using the NiFi Registry

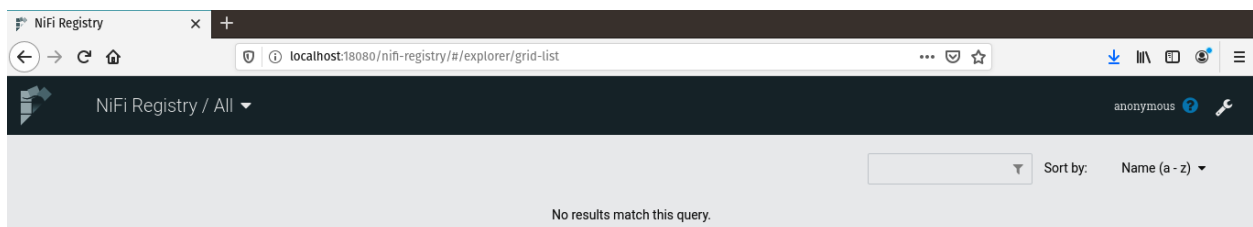


The screenshot shows the Apache NiFi Registry website. The browser tab is "Apache NiFi - Registry" and the address bar shows "https://nifi.apache.org/registry". The navigation bar includes links for Project, Documentation, Downloads, Community, Development, ASF Links, and Subprojects. The main content area features the Apache NiFi Registry logo, which consists of a stylized 'N' made of blue and grey squares, followed by the text "APACHE NIFI" in a small font and "registry" in a large, bold, black font. Below the logo, a subtitle reads: "A subproject of Apache NiFi to store and manage shared resources." A section titled "About" is highlighted in a light blue box. The text under "About" states: "Registry—a subproject of Apache NiFi—is a complementary application that provides a central location for storage and management of shared resources across one or more instances of NiFi and/or MiNiFi. Specific goals for the initial thrust of the Registry effort include:"

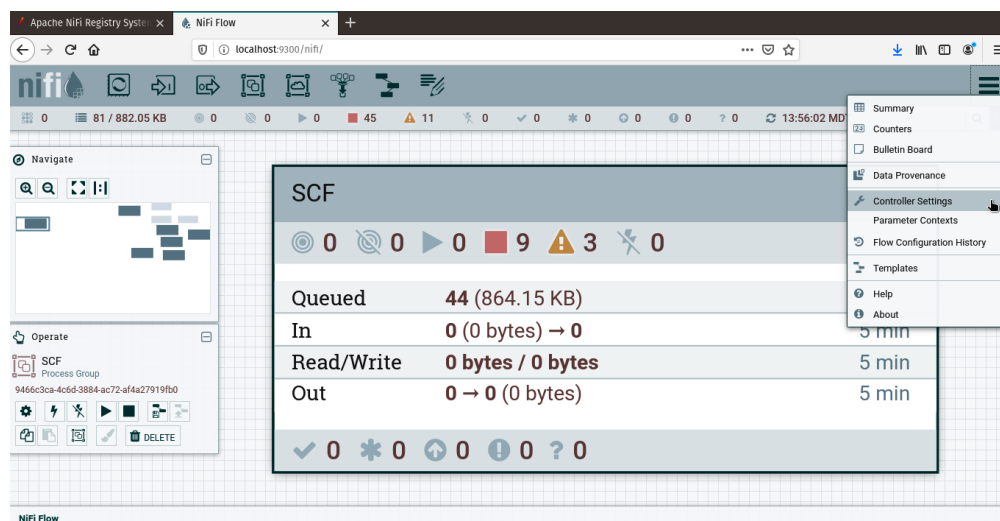
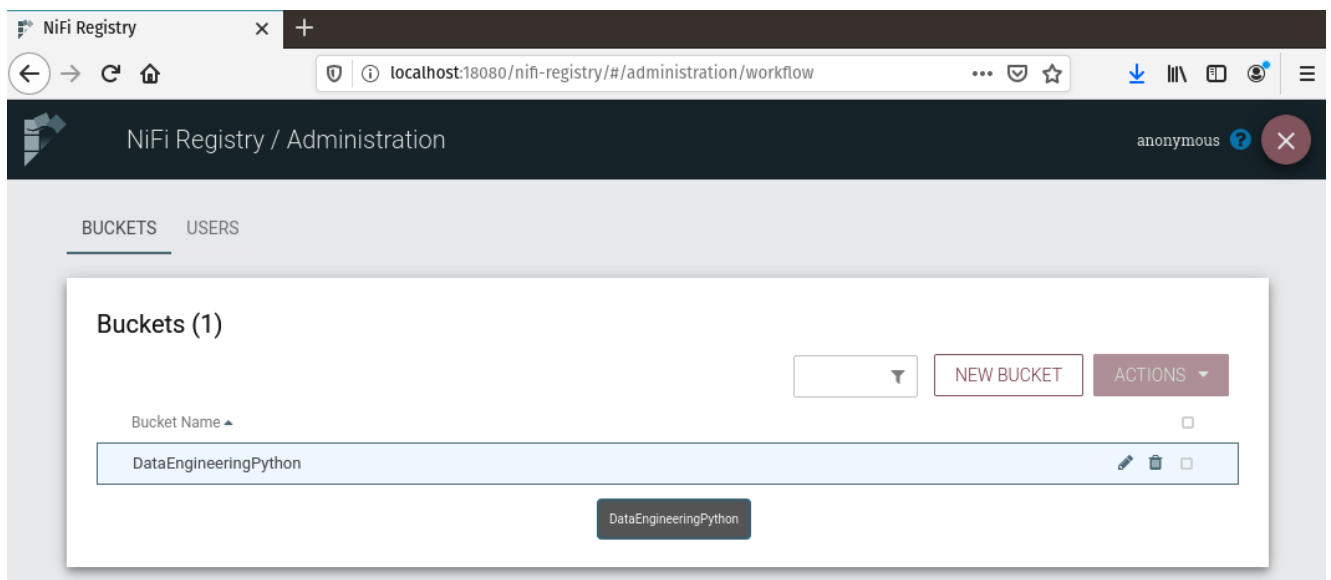
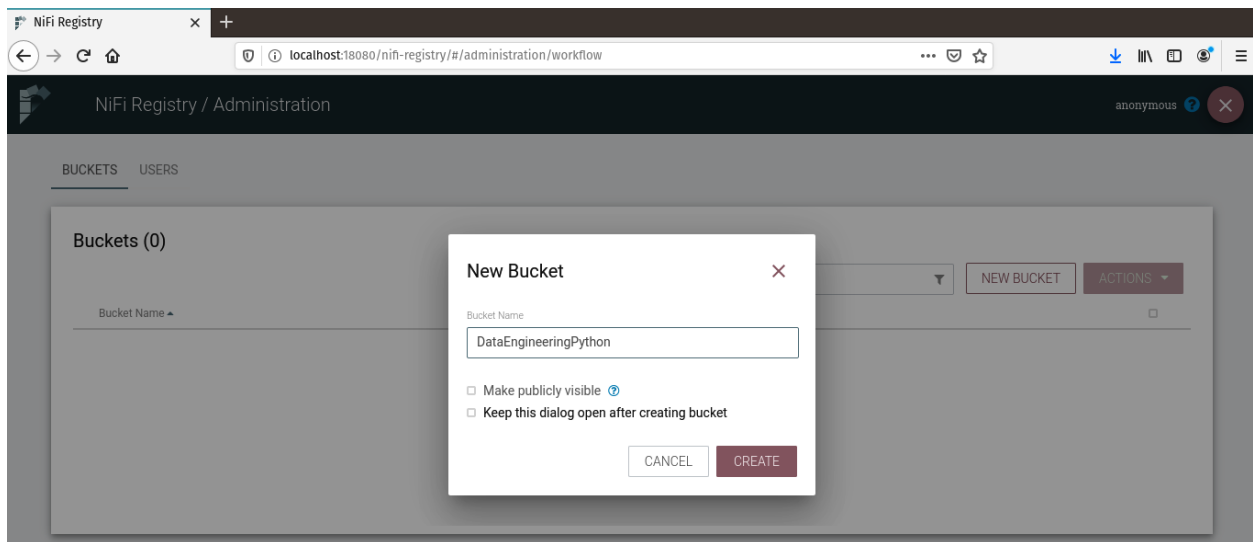


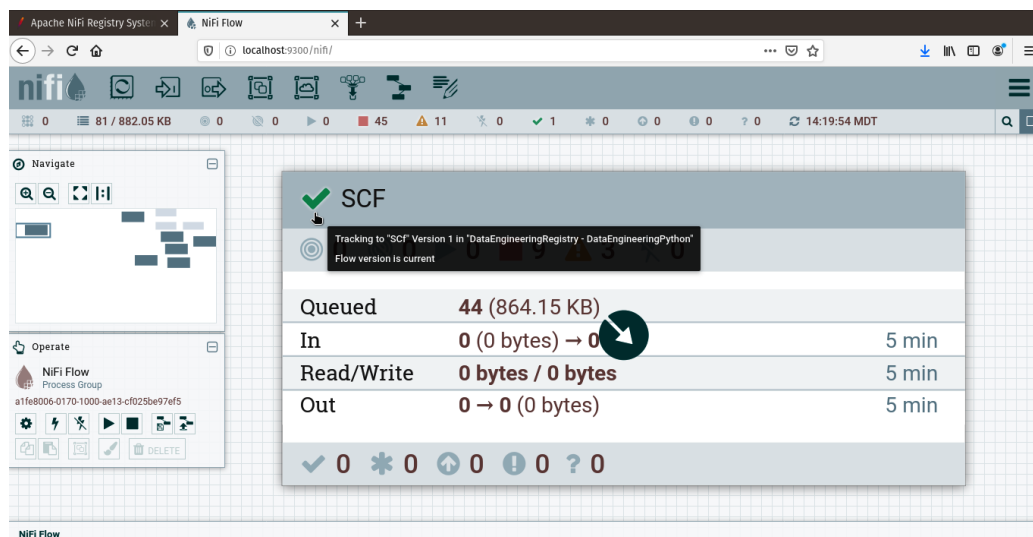
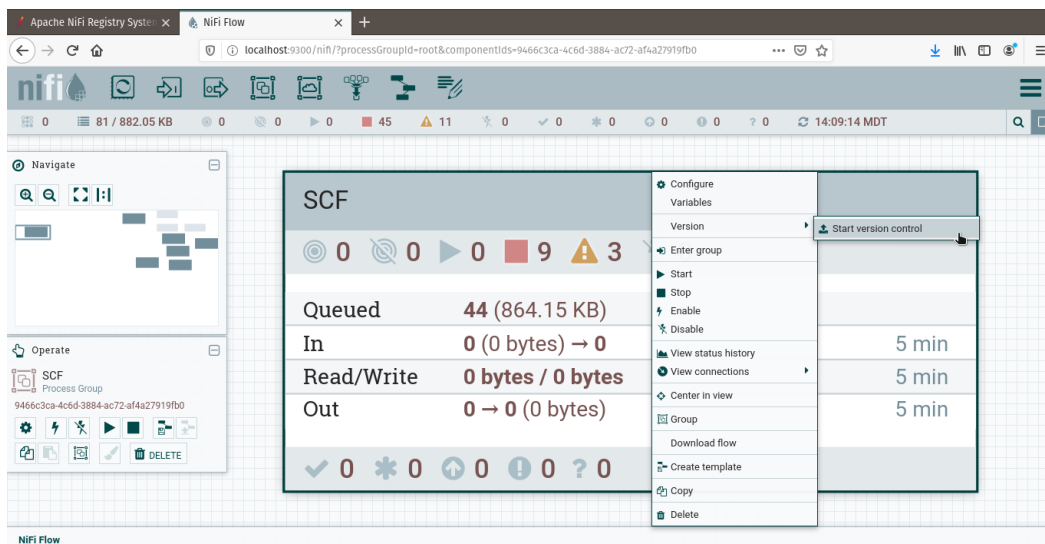
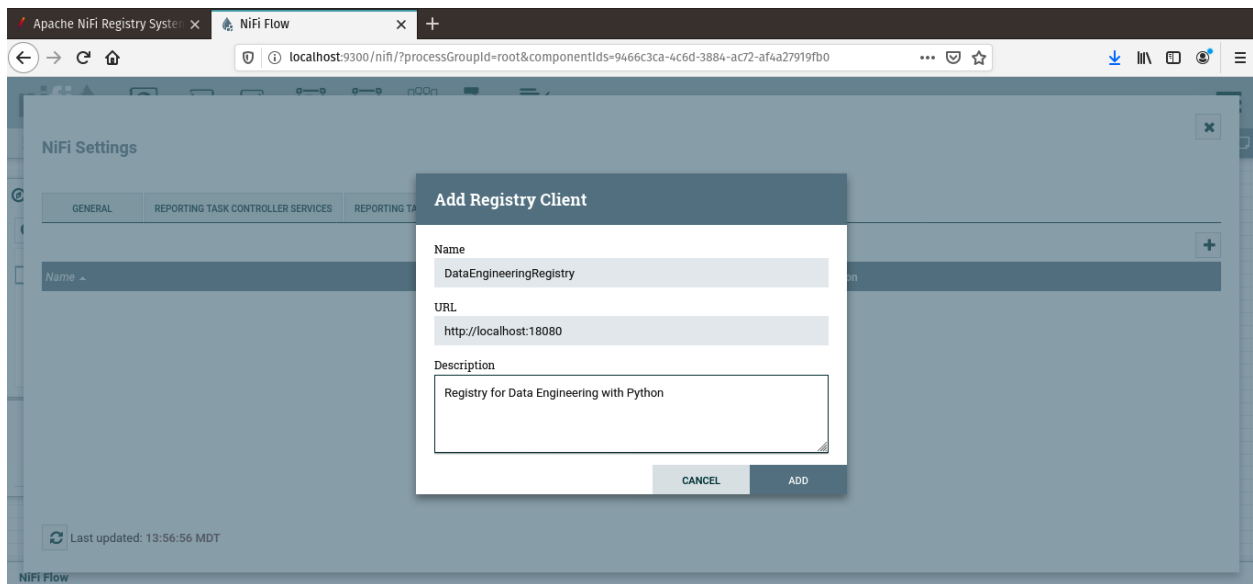
The screenshot shows the "Links" section of the Apache NiFi Registry website. The browser tab is "Apache NiFi - Registry" and the address bar shows "https://nifi.apache.org/registry". The navigation bar is the same as in the previous screenshot. The "Links" section is highlighted in a light blue box. Below the heading, there is a "Releases" section. The text under "Releases" states: "Information about how to verify release signatures and checksums can be found [here](#) and [here](#). The Apache NiFi KEYS file can be found [here](#)." A note follows: "NOTE: Please consult the [Migration Guidance](#) when upgrading." Below the note, there is a list of releases:

- 0.6.0
 - Sources
 - [nifi-registry-0.6.0-source-release.zip](#) (asc, sha256, sha512)
 - Binaries
 - [nifi-registry-0.6.0-bin.tar.gz](#) (asc, sha256, sha512)
 - [nifi-registry-0.6.0-bin.zip](#) (asc, sha256, sha512)
 - [Release Notes](#)
- 0.5.0
 - Sources
 - [nifi-registry-0.5.0-source-release.zip](#) (asc, sha256, sha512)
 - Binaries



The screenshot shows the NiFi Registry web interface. The browser tab is "NiFi Registry" and the address bar shows "localhost:18080/nifi-registry/#/explorer/grid-list". The navigation bar includes the NiFi Registry logo and the text "NiFi Registry / All". The main content area is a grid view. At the bottom, there is a search bar and a "Sort by: Name (a - z)" dropdown menu. The text "No results match this query." is displayed in the center of the grid.





NiFi Registry

localhost:18080/nifi-registry/#explorer/grid-list

NiFi Registry / All

Sort by: Name (a - z)

SCf - DataEngineeringPython Flow

VERSIONS 1

BUCKET IDENTIFIER
c8e69fca-7f08-44a1-967c-32a1a5709404

CHANGE LOG [CHANGE LOG](#)

Version 1 - 13 minutes ago by anonymous

FLOW IDENTIFIER
7aa48258-ce5c-44d8-ae7a-6e67d9dbaa4c

First commit
May-15-2020 at 2:10 PM

DESCRIPTION
Chapter 6: SeeClickFix API to Elasticsearch

ACTIONS

Apache NiFi Registry System

NiFi Flow

localhost:9300/nifi/?processGroupId=9466c3ca-4c6d-3884-ac72-af4a27919fb0&componentId=1a1026fd-0172

0 81 / 882.05 KB 0 0 0 46 11 0 0 1 0 0 0 14:39:46 MDT

Navigate

Operate

Multiple components selected

DELETED

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 2 (528.77 KB) 5 min

Split.Json 1.11.3
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 split
Queued 36 (65.75 KB) 5 min

coords
ExecuteScript 1.11.3
org.apache.nifi - nifi-scripting-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

Write 0 bytes / 0 bytes 5 min

Evaluate.JsonPath
Evaluate.JsonPath 1.11.3
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 1 (2.3 KB) 5 min

Name matched
Queued 0 (0 bytes) 5 min

ElasticSCF
PutElasticsearchHttp 1.11.3
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

Name matched
Queued 0 (0 bytes) 5 min

New DataWarehouse
PutElasticsearchHttp 1.11.3
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

NiFi Flow

SCF

Apache NiFi Registry System

NiFi Flow

localhost:9300/nifi/

0 81 / 882.05 KB 0 0 0 46 11 0 0 1 0 0 0 14:45:53 MDT

Navigate

Operate

NiFi Flow

Process Group
a1f68006-0170-1000-ae13-cf025be97af5

SCF

Tracking to 'SCF' Version 1 in 'DataEngineeringRegistry - DataEngineeringPython'
Local changes have been made

Queued 44 (864.15 KB)

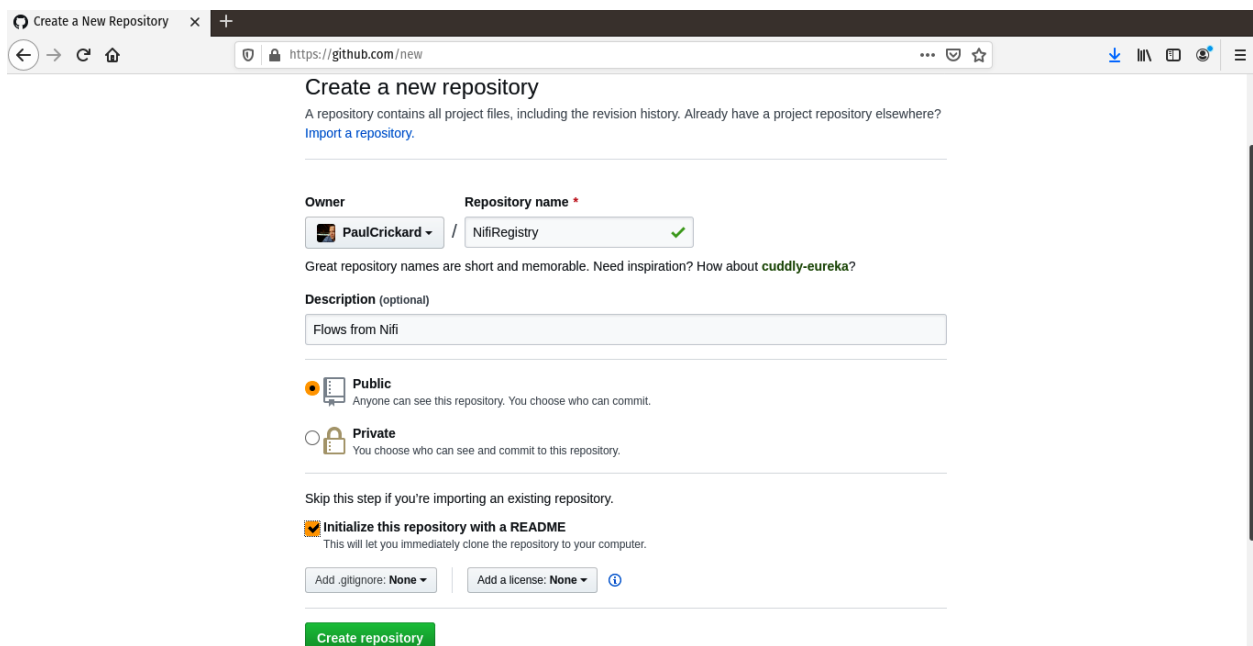
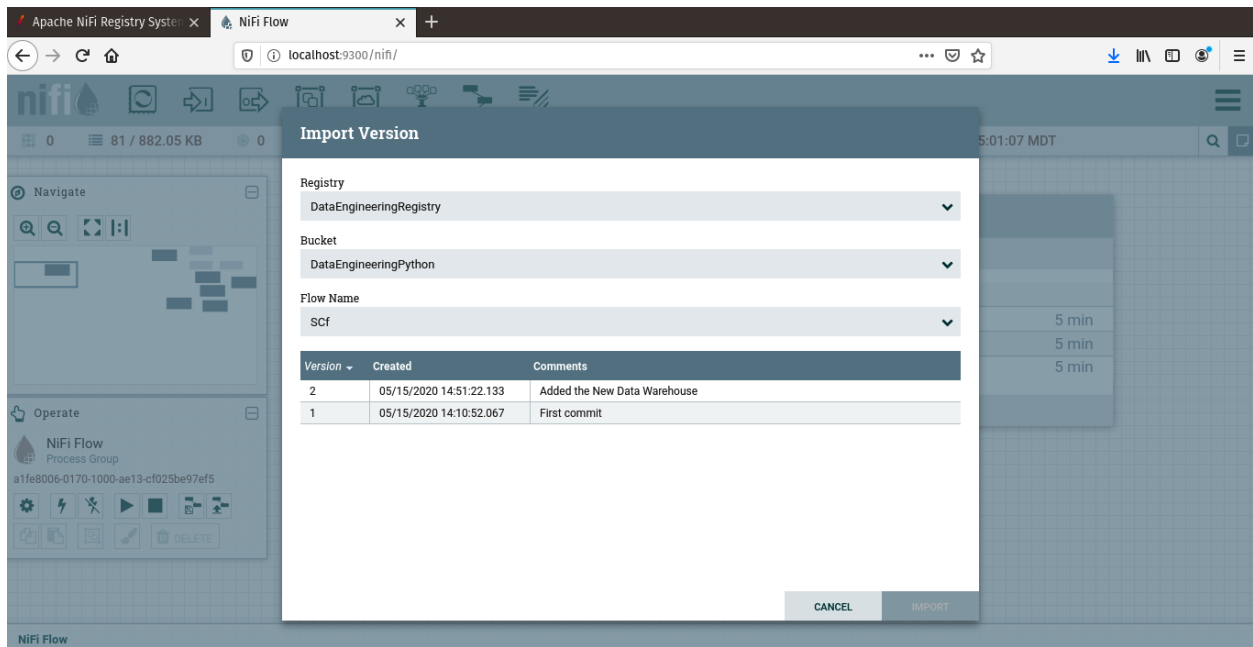
In 0 (0 bytes) → 0 5 min

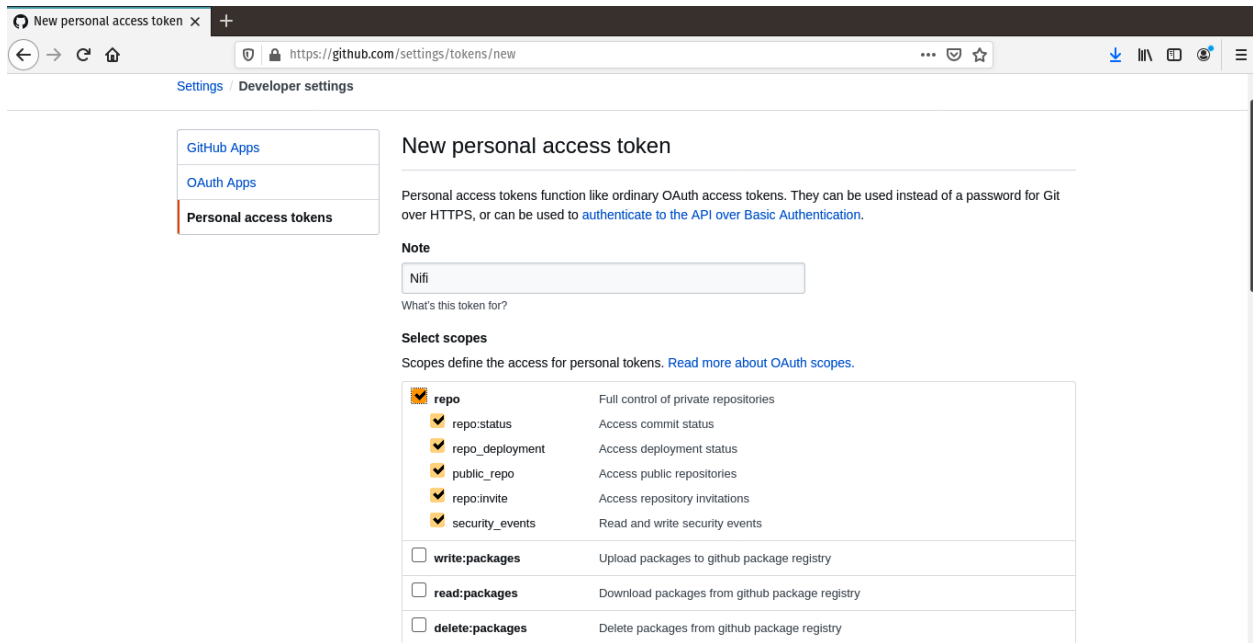
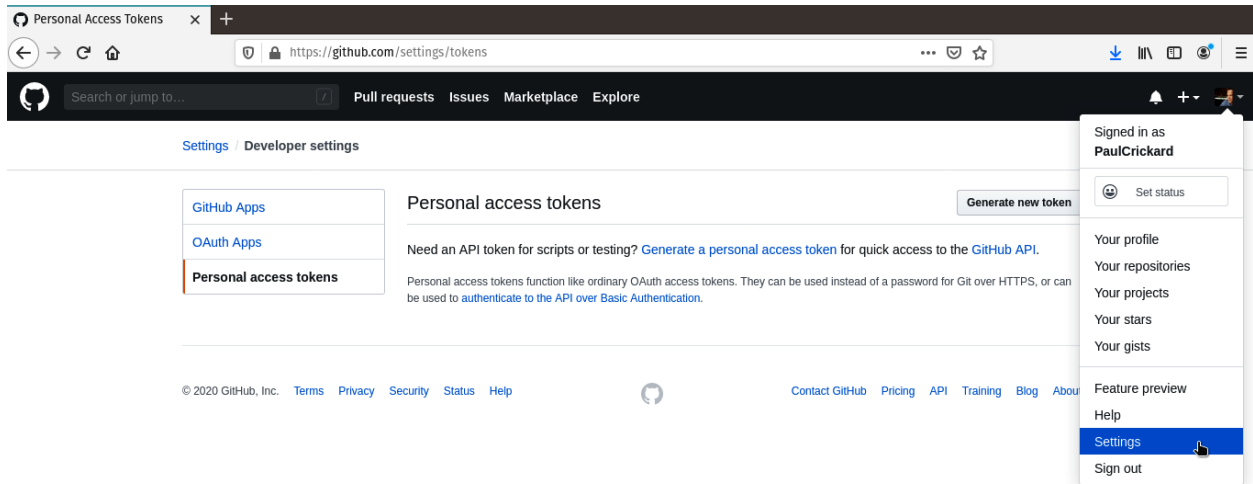
Read/Write 0 bytes / 0 bytes 5 min

Out 0 → 0 (0 bytes) 5 min

✓ 0 * 0 ↑ 0 ! 0 ? 0

NiFi Flow





```
paulcrickard@pop-os:~$ git clone https://github.com/PaulCrickard/NifiRegistry.git
Cloning into 'NifiRegistry'...
remote: Enumerating objects: 3, done.
remote: Counting objects: 100% (3/3), done.
Unpacking objects: 100% (3/3), done.
remote: Total 3 (delta 0), reused 0 (delta 0), pack-reused 0
```


Open

providers.xml
~/nifi-registry-0.6.0/conf

Save

```
limitations under the License.
-->
<providers>

  <!-- NOTE: The providers in this file must be listed in the order defined in providers.xsd which is the following:
    1) Flow Persistence Provider (Must occur once and only once)
    2) Event Hook Providers (May occur 0 or more times)
    3) Bundle Persistence Provider (Must occur once and only once)
  -->

  <!--
    <flowPersistenceProvider>
      <class>org.apache.nifi.registry.provider.flow.FileSystemFlowPersistenceProvider</class>
      <property name="Flow Storage Directory">./flow_storage</property>
    </flowPersistenceProvider>
  -->

  <flowPersistenceProvider>
    <class>org.apache.nifi.registry.provider.flow.git.GitFlowPersistenceProvider</class>
    <property name="Flow Storage Directory">/home/paulcrickard/NifiRegistry</property>
    <property name="Remote To Push">origin</property>
    <property name="Remote Access User">paulcrickard</property>
    <property name="Remote Access Password">YOUR TOKEN</property>
  </flowPersistenceProvider>

  <!--
    <flowPersistenceProvider>
      <class>org.apache.nifi.registry.provider.flow.DatabaseFlowPersistenceProvider</class>
    </flowPersistenceProvider>
  -->

  <!--
    <eventHookProvider>
      <class>org.apache.nifi.registry.provider.hook.ScriptEventHookProvider</class>
      <property name="Script Path"></property>
      <property name="Working Directory"></property>
    </eventHookProvider>
  -->

  -->
</providers>
-->
-->
```

XML Tab Width: 8 Ln 35, Col 59 INS

NiFi Registry

localhost:18080/nifi-registry/#/explorer/grid-list

NiFi Registry / All anonymous

Sort by: Name (a - z)

SCf - DataEngineeringPython
Flow

VERSIONS
3

ACTIONS

BUCKET IDENTIFIER
c8e69fca-7f08-44a1-967c-32a1a5709404

FLOW IDENTIFIER
7aa48258-ce5c-44d8-ae7a-6e67d9dbaa4c

DESCRIPTION
Chapter 6: SeeClickFix API to Elasticsearch

CHANGE LOG

Version 3 - a minute ago
by anonymous

Added a third warehouse
May-15-2020 at 4:25 PM

Version 2 - 2 hours ago
by anonymous

Version 1 - 2 hours ago
by anonymous

NifiRegistry/DataEngineeri

+

←

→

↺

🏠

https://github.com/PaulCrickard/NifiRegistry/tree/master/DataEngineeringPython

💡 Recommendation

...

📧

☆

⬇️

🔍

📄

👤

☰

Read the guide

📁 PaulCrickard / NifiRegistry

👁️ Unwatch

1

★ Star

0

🍴 Fork

0

🔗 Code

🔔 Issues 0

🔀 Pull requests 0

🔧 Actions

📁 Projects 0

📖 Wiki

🔒 Security 0

📊 Insights

⚙️ Settings

Branch: master

NifiRegistry / DataEngineeringPython /

Create new file

Upload files

Find file

History

👤 root Added a third warehouse

Latest commit 50fe8faf 1 minute ago


..

📄 SCf.snapshot Added a third warehouse 1 minute ago

📄 bucket.yml Added a third warehouse 1 minute ago

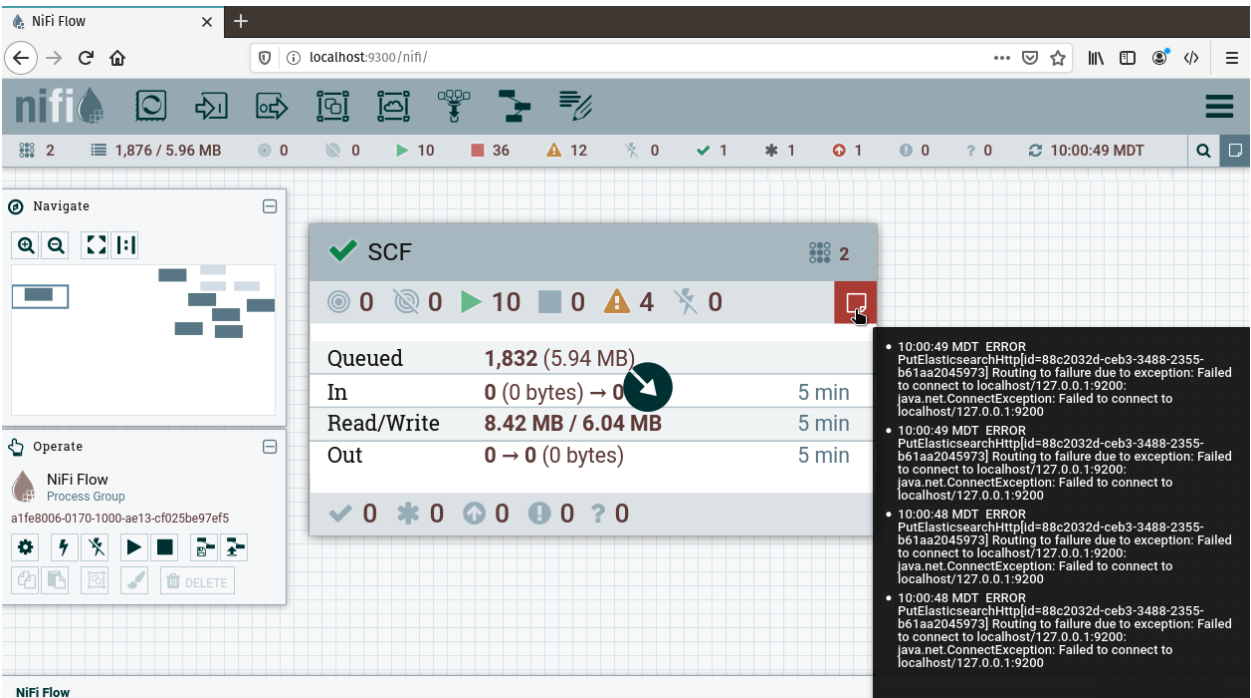
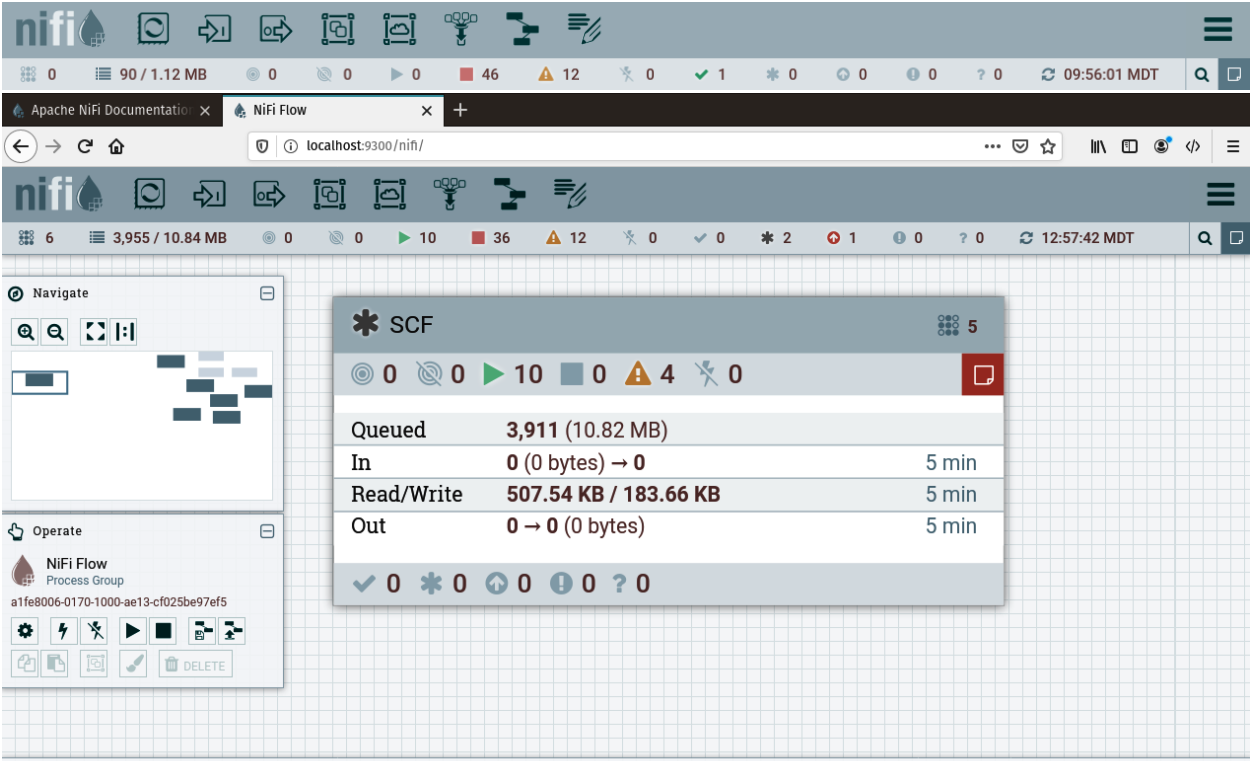
© 2020 GitHub, Inc.

[Terms](#) [Privacy](#) [Security](#) [Status](#) [Help](#)



[Contact GitHub](#) [Pricing](#) [API](#) [Training](#) [Blog](#) [About](#)

Chapter 9: Monitoring and Logging Data Pipelines



NiFi Bulletin Board

Filter

by message

13:12:21 MDT **ERROR** [88c2032d-ceb3-3488-2355-b61aa2045973](#)
PutElasticsearchHttp[Id=88c2032d-ceb3-3488-2355-b61aa2045973] Routing to failure due to exception: Failed to connect to localhost/127.0.0.1:9200: java.net.ConnectException: Failed to connect to localhost/127.0.0.1:9200

13:12:21 MDT **ERROR** [88c2032d-ceb3-3488-2355-b61aa2045973](#)
PutElasticsearchHttp[Id=88c2032d-ceb3-3488-2355-b61aa2045973] Routing to failure due to exception: Failed to connect to localhost/127.0.0.1:9200: java.net.ConnectException: Failed to connect to localhost/127.0.0.1:9200

13:12:24 MDT **ERROR** [88c2032d-ceb3-3488-2355-b61aa2045973](#)
PutElasticsearchHttp[Id=88c2032d-ceb3-3488-2355-b61aa2045973] Routing to failure due to exception: Failed to connect to localhost/127.0.0.1:9200: java.net.ConnectException: Failed to connect to localhost/127.0.0.1:9200

13:12:24 MDT **ERROR** [88c2032d-ceb3-3488-2355-b61aa2045973](#)
PutElasticsearchHttp[Id=88c2032d-ceb3-3488-2355-b61aa2045973] Routing to failure due to exception: Failed to connect to localhost/127.0.0.1:9200: java.net.ConnectException: Failed to connect to localhost/127.0.0.1:9200

13:12:24 MDT **ERROR** [88c2032d-ceb3-3488-2355-b61aa2045973](#)
PutElasticsearchHttp[Id=88c2032d-ceb3-3488-2355-b61aa2045973] Routing to failure due to exception: Failed to connect to localhost/127.0.0.1:9200: java.net.ConnectException: Failed to connect to localhost/127.0.0.1:9200

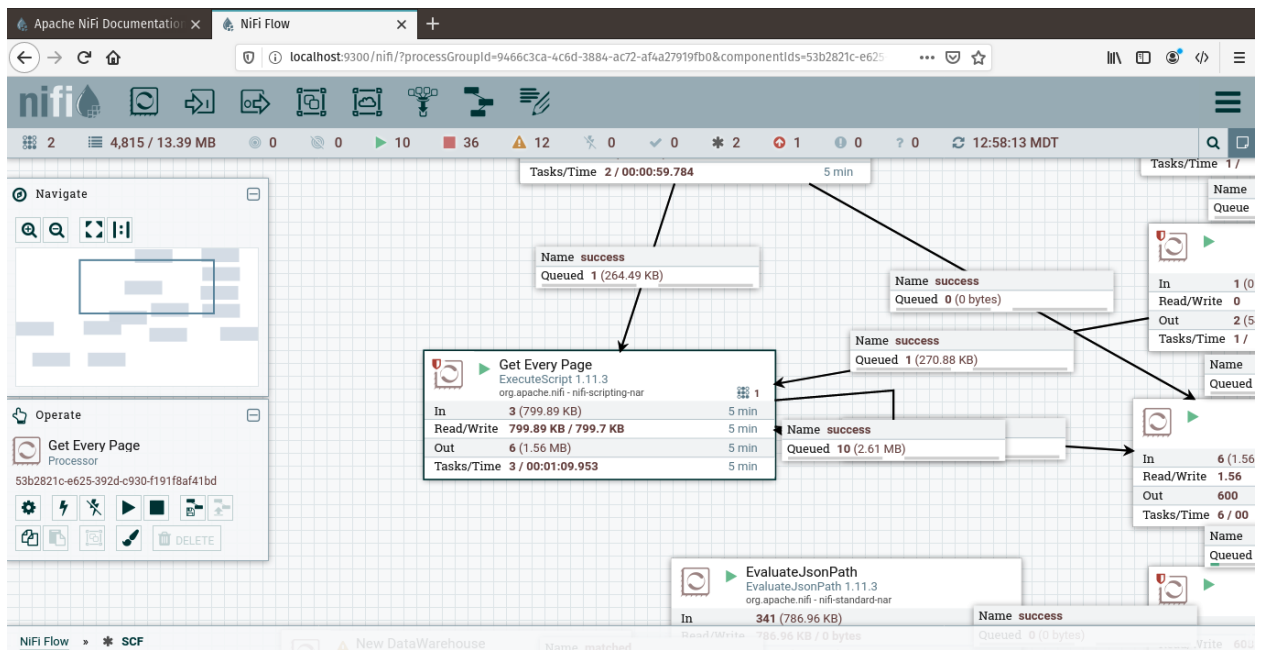
13:12:25 MDT **ERROR** [88c2032d-ceb3-3488-2355-b61aa2045973](#)
PutElasticsearchHttp[Id=88c2032d-ceb3-3488-2355-b61aa2045973] Routing to failure due to exception: Failed to connect to localhost/127.0.0.1:9200: java.net.ConnectException: Failed to connect to localhost/127.0.0.1:9200

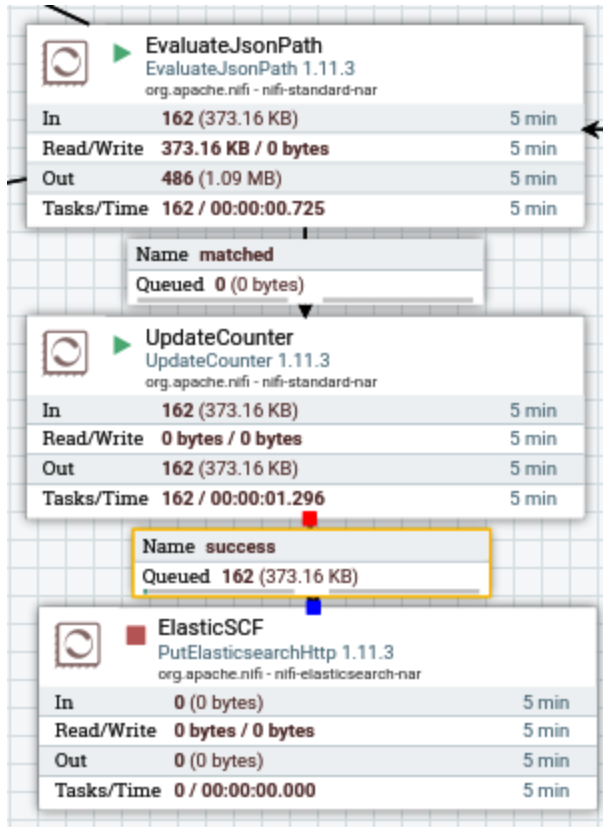
13:12:25 MDT **ERROR** [88c2032d-ceb3-3488-2355-b61aa2045973](#)
PutElasticsearchHttp[Id=88c2032d-ceb3-3488-2355-b61aa2045973] Routing to failure due to exception: Failed to connect to localhost/127.0.0.1:9200: java.net.ConnectException: Failed to connect to localhost/127.0.0.1:9200

Auto-refresh

Last updated: 13:12:25 MDT

Clear





Apache NiFi Documentation | NiFi Flow

localhost:9300/nifi/?processGroupId=9466c3ca-4c6d-3884-ac72-af4a27919fb0&componentIds=295f179f-0172-1

4:05:12 MDT

Processor Details

Running STOP & CONFIGURE

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field

Property	Value
Counter Name	SCFtoElasticsearch
Delta	1

OK

coords
ExecuteScript 1.11.3
org.apache.nifi - nifi-scripting-nar

In	161 (287.22 KB)	5 min
Read/Write	287.22 KB / 370.78 KB	5 min
Out	161 (370.78 KB)	5 min
Tasks/Time	161 / 00:00:30.480	5 min

NiFi Flow > SCF

Apache NiFi Documentation x NiFi Flow x +

localhost:9300/nifi/?processGroupId=9466c3ca-4c6d-3884-ac72-af4a27919fb0&componentIds=295f179f-0172 ...

NiFi Counters

Displaying 2 of 2

Filter by name

Context	Name	Value
All UpdateCounter's	SCFtoElasticsearch	162
UpdateCounter (295f179f-0172-1000-ee63-c25c545f224e)	SCFtoElasticsearch	162

Last updated: 14:04:28 MDT

NiFi Flow » SCF

NiFi Rest Api-1.11.4 x NiFi Flow x +

localhost:9300/nifi/?processGroupId=295e705c-0172-1000-0b19-0d3330892f8e&componentIds=...

Add Reporting Task

Source all groups

Displaying 15 of 15

Type	Version	Tags
AmbariReportingTask	1.11.3	ambari, metrics, reporting
AzureLogAnalyticsProvenanceReportingTask	1.11.3	provenance, log analytics, reporting
AzureLogAnalyticsReportingTask	1.11.3	metrics, log analytics, reporting
ControllerStatusReportingTask	1.11.3	stats, log
DataDogReportingTask	1.11.3	datadog, metrics, reporting
MetricsReportingTask	1.11.3	metrics, reporting
MonitorDiskUsage	1.11.3	disk, repo, warning, storage, m...
MonitorMemory	1.11.3	jvm, memory, warning, monitor...
PrometheusReportingTask	1.11.3	prometheus, metrics, time seri...
ScriptedReportingTask	1.11.3	lua, python, jython, js, execute...
SiteToSiteBulletinReportingTask	1.11.3	site, restricted, bulletin, site to ...

AmbariReportingTask 1.11.3 org.apache.nifi - nifi-ambari-nar

Publishes metrics from NiFi to Ambari Metrics Service (AMS). Due to how the Ambari Metrics Service works, this reporting task should be scheduled to run every 60 seconds. Each iteration it will send the metrics from the previous iteration, and calculate the current metrics to be sent on next iteration. Scheduling this reporting task at a frequency other than 60 seconds may produce ...

CANCEL ADD

NiFi Rest Api-1.11.4 x NiFi Flow x +

localhost:9300/nifi/?processGroupId=295e705c-0172-1000-0b19-0d3330892f8e&componentIds=...

Configure Reporting Task

SETTINGS PROPERTIES COMMENTS

Required field

Property	Value
Threshold	1%
Directory Location	/home/paulcrickard/nifi-1.11.3
Directory Display Name	MyDrive

CANCEL APPLY

Browser tabs: NiFi Rest Api-1.11.4, NiFi Flow

Address bar: localhost:9300/nifi/?processGroupId=295e705c-0172-1000-0b19-0d3330892f8e&componentId=

Top bar: nifi, 6,627 / 18.19 MB, 0, 0, 0, 48, 12, 0, 0, 2, 1, 0, 0, 14:17:08 MDT

Left sidebar:

- Navigate
 - Search, View, Full Screen, Help
- Operate
 - reporting Process Group
 - 295e705c-0172-1000-0b19-0d3330892f8e
 - Settings, Run, Stop, Play, Pause, Refresh, Delete

Warning message: 14:17:07 MDT WARNING MonitorDiskUsage[d=01721003-179f-195f-9-be-27f0f068b38e] MyDrive exceeds configured threshold of 1%, having 14.27 GB / 449.09 GB (3.18%) used and 434.82 GB (96.82%) free

Bottom bar: NIFI Flow > reporting

Browser tabs: NiFi Flow, Slack API: Applications | Slack

Address bar: https://api.slack.com/apps

Header: slack api, Search, Documentation, Tutorials, Your Apps

Left sidebar:

- Start learning
- Authentication
- Surfaces
- Block Kit
- Interactivity
- Messaging
- APIs
- Enterprise
- Reference

Content area:

Your Apps

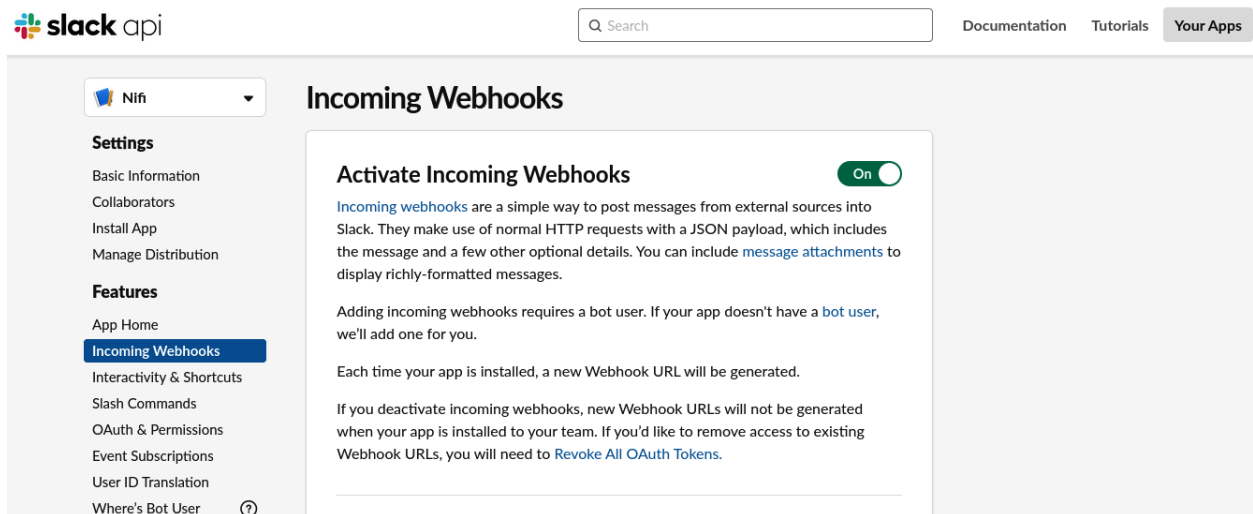
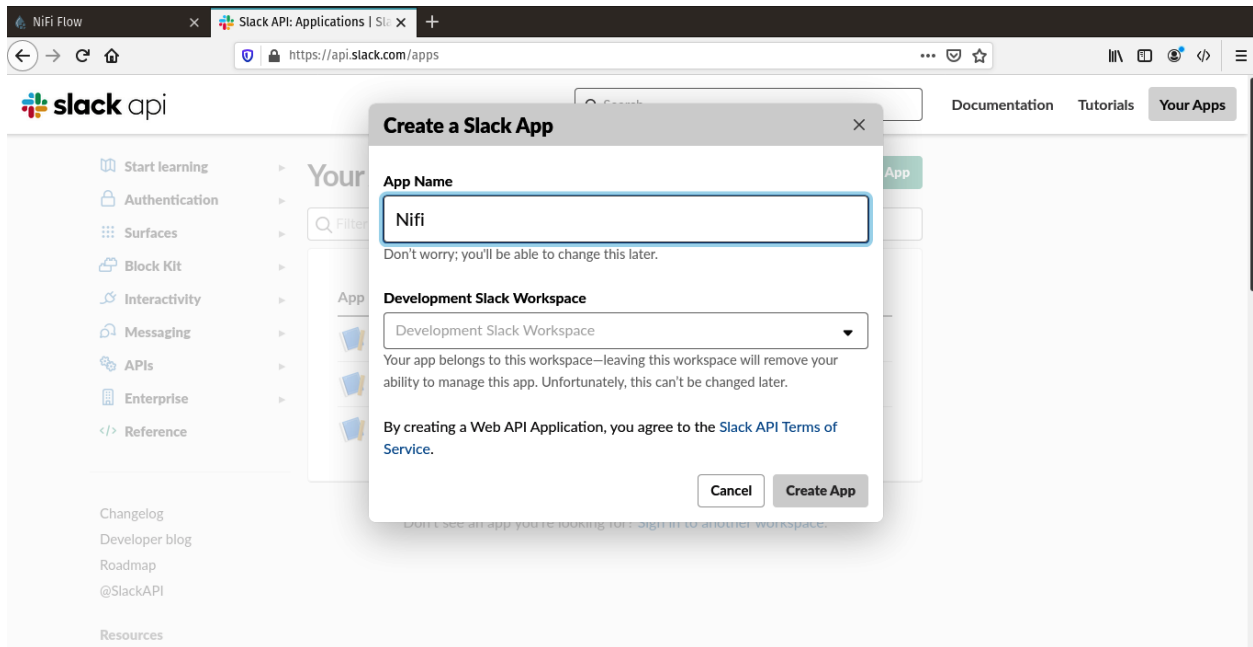
Create New App

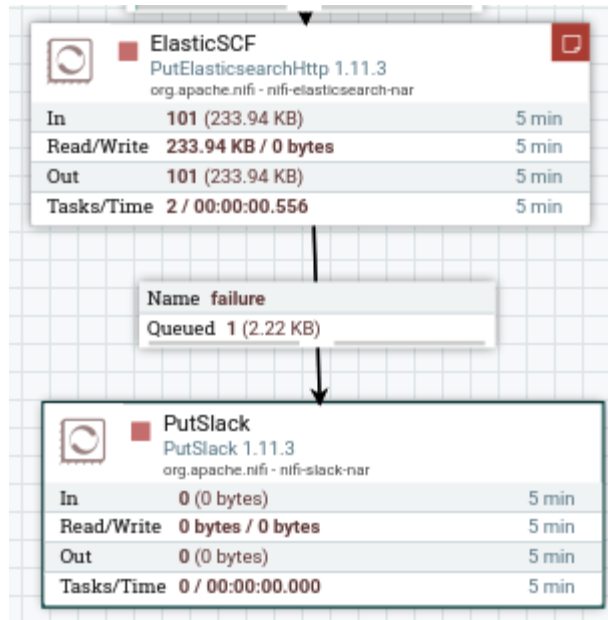
Filter apps by name or workspace

App Name	Workspace	Distribution Status
Nifi	DA2ND	Not distributed
CMSBot	DA2ND	Not distributed
NifiDM	DA2ND	Not distributed

Don't see an app you're looking for? [Sign in to another workspace.](#)

Footer: Changelog, Developer blog, Roadmap, @SlackAPI, Resources





NiFi Flow

localhost:9300/nifi/?processGroupId=9466c3ca-4c6d-3884-ac72-af4a27919fb0&componentId=01721005-179f-...

Processor Details

Running STOP & CONFIGURE

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field

Property	Value
Webhook URL	<input type="text" value="Sensitive value set"/>
Webhook Text	<input type="text" value="\${id:append(: Record failed Upsert Elasticsearch)}"/>
Channel	<input type="text" value="No value set"/>
Username	<input type="text" value="Nifi"/>
Icon URL	<input type="text" value="No value set"/>
Icon Emoji	<input type="text" value="No value set"/>

OK



Paul Crickard 2:52 PM

added an integration to this channel: [Nifi](#)

New



Nifi APP 3:03 PM

7781528: Record failed Upsert Elasticsearch

Chapter 10: Deploying Your Data Pipelines

The screenshot shows the NiFi Flow console interface. The top navigation bar includes the NiFi logo, a search bar, and a status bar with various icons and a timestamp of 11:51:20 MDT. The main workspace displays a pipeline with two processors:

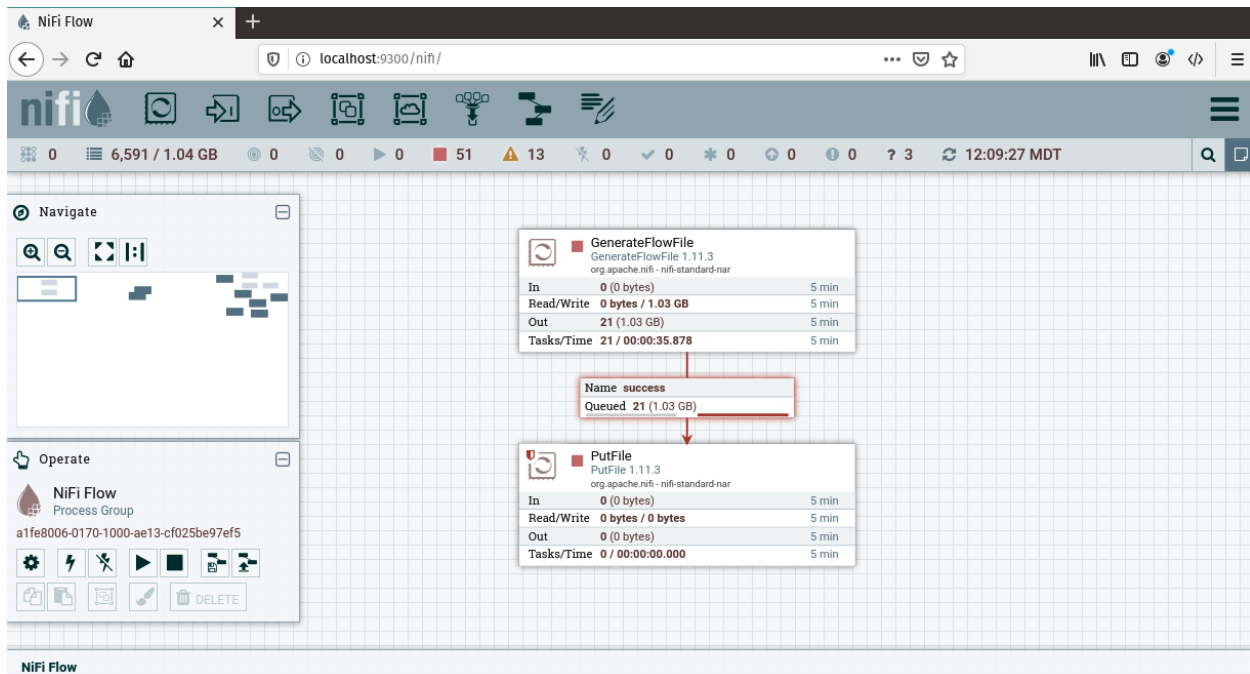
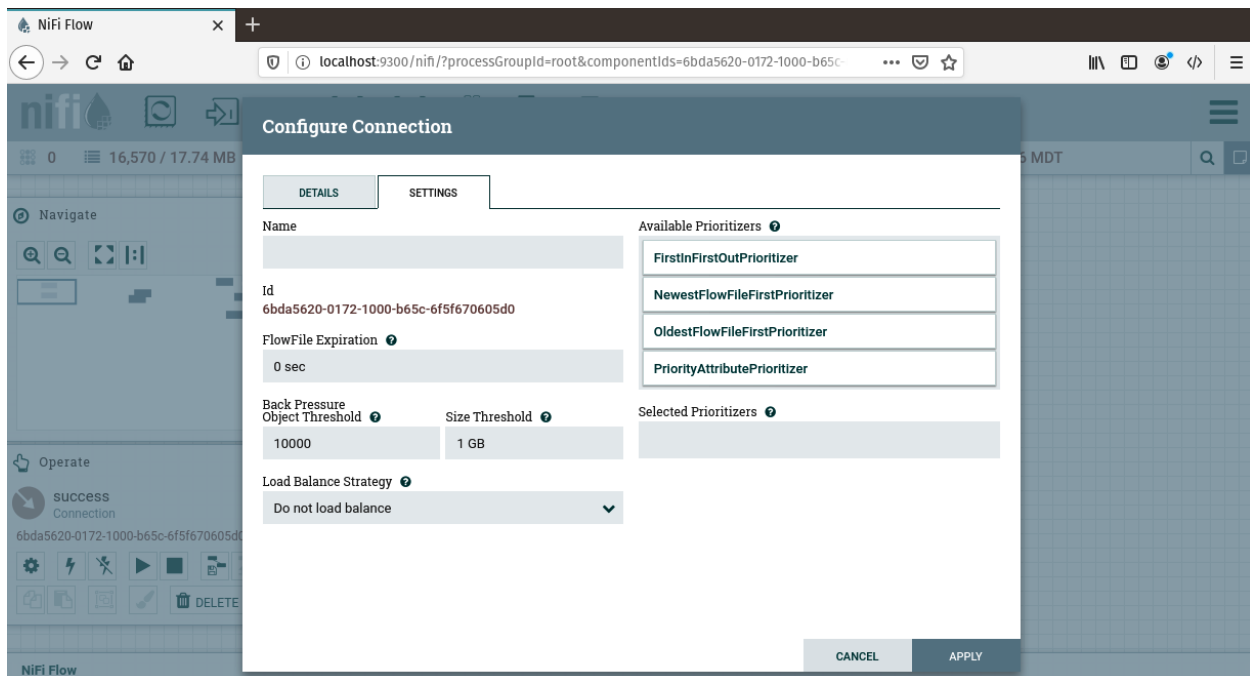
- GenerateFlowFile** (GenerateFlowFile 1.11.3, org.apache.nifi - nifi-standard-nar):
 - In: 0 (0 bytes)
 - Read/Write: 0 bytes / 0 bytes
 - Out: 0 (0 bytes)
 - Tasks/Time: 0 / 00:00:00.000
- PutFile** (PutFile 1.11.3, org.apache.nifi - nifi-standard-nar):
 - In: 0 (0 bytes)
 - Read/Write: 0 bytes / 0 bytes
 - Out: 0 (0 bytes)
 - Tasks/Time: 0 / 00:00:00.000

A flow arrow connects the output of the GenerateFlowFile processor to the input of the PutFile processor. The left sidebar shows the 'Operate' tab with a 'PutFile Processor' selected.

The screenshot shows the NiFi Flow console interface after the pipeline has been executed. The top navigation bar includes the NiFi logo, a search bar, and a status bar with various icons and a timestamp of 11:58:40 MDT. The main workspace displays the same pipeline as the previous screenshot, but with updated data:

- GenerateFlowFile** (GenerateFlowFile 1.11.3, org.apache.nifi - nifi-standard-nar):
 - In: 0 (0 bytes)
 - Read/Write: 0 bytes / 0 bytes
 - Out: 10,000 (0 bytes)
 - Tasks/Time: 10,000 / 00:00:22.426
- PutFile** (PutFile 1.11.3, org.apache.nifi - nifi-standard-nar):
 - In: 0 (0 bytes)
 - Read/Write: 0 bytes / 0 bytes
 - Out: 0 (0 bytes)
 - Tasks/Time: 0 / 00:00:00.000

A flow arrow connects the output of the GenerateFlowFile processor to the input of the PutFile processor. The left sidebar shows the 'Operate' tab with a 'NiFi Flow Process Group' selected.



NiFi Flow

localhost:9300/nifi/?processGroupId=6bfa2f2a-0172-1000-82b6-cb1f17a7fce7&component=...

8,354 / 17.74 MB

12:47:17 MDT

Navigate

Operate

Write Data
Process Group
6bfa2f2a-0172-1000-82b6-cb1f17a7fce7

IncomingData
Queued 0 (0 bytes)

EvaluateJsonPath
EvaluateJsonPath 1.11.3
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)

UpdateAttribute
UpdateAttribute 1.11.3
org.apache.nifi - nifi-update-attribute-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)

PutFile
PutFile 1.11.3
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

NiFi Flow > Write Data

NiFi Flow

localhost:9300/nifi/

8,354 / 17.74 MB

12:48:59 MDT

Navigate

Operate

NiFi Flow
Process Group
a1fe8006-0170-1000-ae13-cf025be97ef5

Create Connection

DETAILS **SETTINGS**

From Output: FromGenerateData

To Input: IncomingData

Within Group: Generate Data

Within Group: Write Data

CANCEL ADD

0 0 / 0 bytes 0 0 0 0 2 1 0 0 0 1 0 0 ? 0 15:29:35 MDT

PRODUCE

postgresToelasticsearch

Queued	0 (0 bytes)	
In	0 (0 bytes) → 0	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 → 0 (0 bytes)	5 min

NiFi Flow

0 0 / 0 bytes 0 0 0 0 2 1 0 0 0 1 0 0 ? 0 15:30:05 MDT

Change Version

Registry
TheNiFiRegistry

Bucket
DataEngineeringPython

Flow Name
postgresSQLtoElasticsearch

Current Version
1

Version	Created	Comments
2	05/31/2020 15:29:17.227	Changed batch size to 1,000
1	05/31/2020 14:40:23.642	

CANCEL CHANGE

NiFi Flow

0 0 / 0 bytes 0 0 0 0 2 1 0 1 0 0 0 0 ? 0 15:22:25 MDT

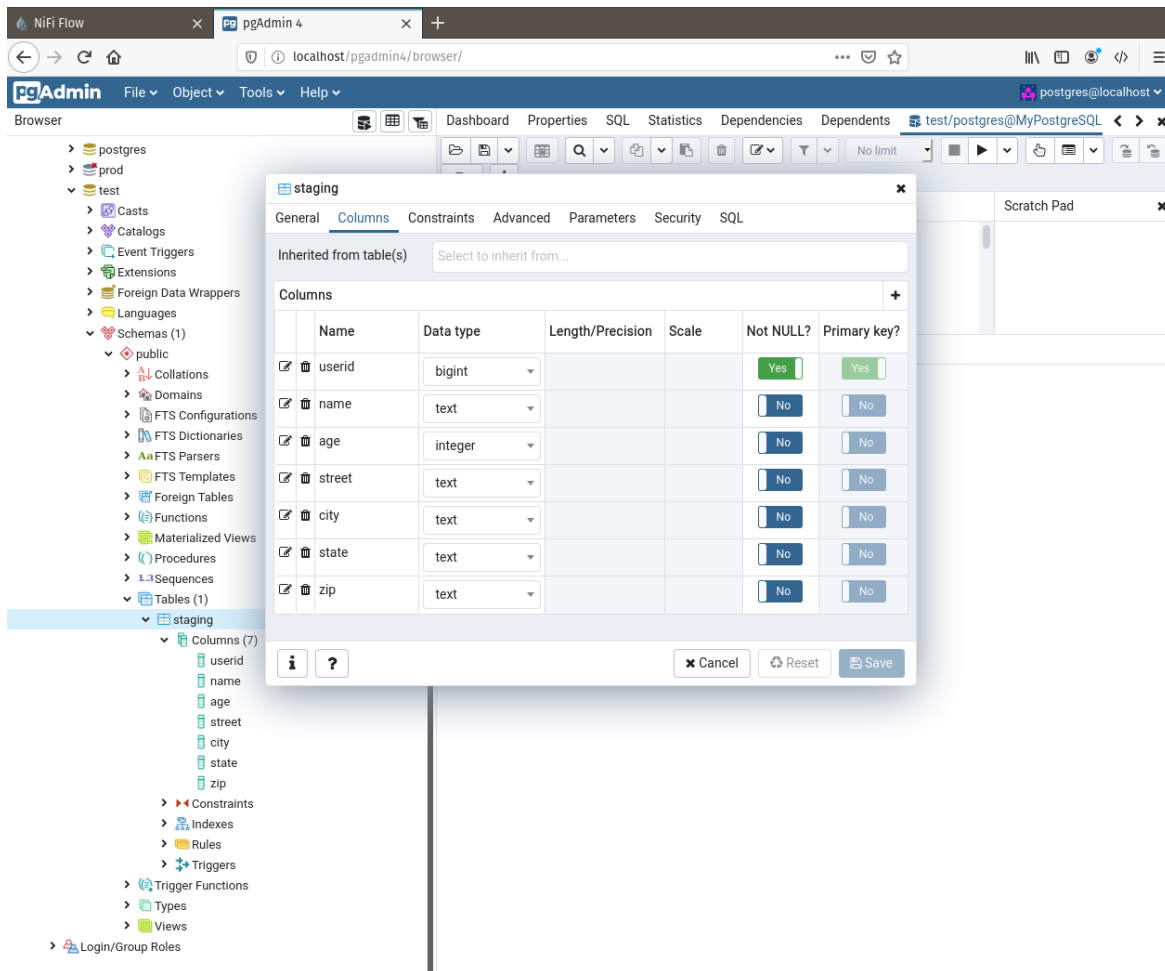
PRODUCE

postgresToelasticsearch

Queued	0 (0 bytes)	
In	0 (0 bytes) → 0	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 → 0 (0 bytes)	5 min

NiFi Flow

Chapter 11: Building a Production Data Pipeline



pgAdmin 4

localhost/pgadmin4/browser/

pgAdmin

FileObjectToolsHelp

postgres@localhost

Browser

DashboardPropertiesSQLStatisticsDependenciesDependents

Servers (1)

- MyPostgreSQL
 - Databases (4)
 - dataengineering
 - postgres
 - prod
 - test
 - Casts
 - Catalogs
 - Event Triggers
 - Extensions
 - Foreign Data Wrappers
 - Languages
 - Schemas (1)
 - public
 - Collations
 - Domains
 - FTS Configuration
 - FTS Dictionaries
 - FTS Parsers
 - FTS Templates
 - Foreign Tables
 - Functions
 - Materialized Views
 - Procedures
 - Sequences
 - Tables (1)
 - staging
 - Columns
 - Constraints
 - Indexes
 - Rules
 - Triggers
 - Trigger Functions
 - Types
 - Views
 - Login/Group Roles
 - Tablespaces

Create

Refresh...

Count Rows

Delete/Drop

Drop Cascade

Reset Statistics

Import/Export...

Maintenance...

Scripts

- CREATE Script
- DELETE Script
- INSERT Script
- SELECT Script
- UPDATE Script

Truncate

Backup...

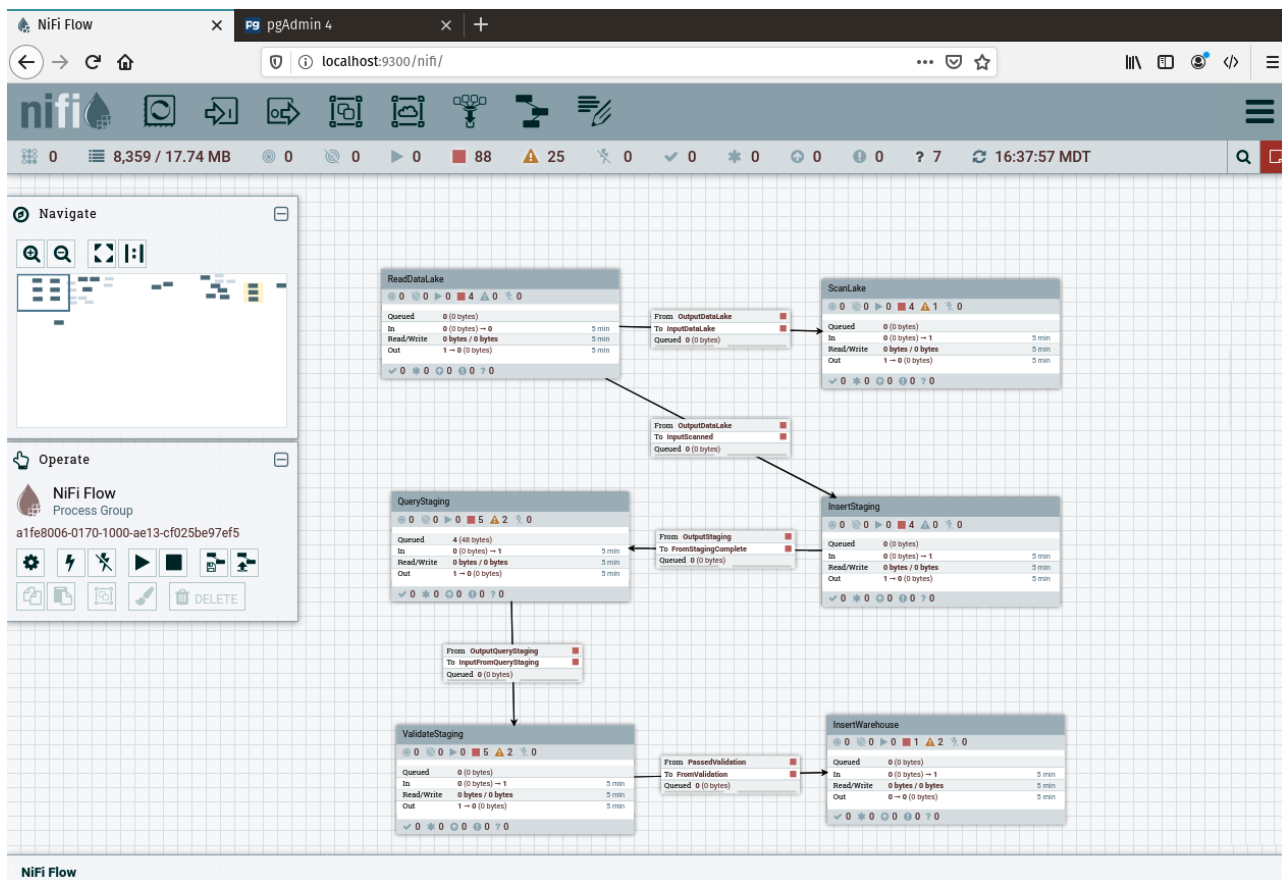
Restore...

View/Edit Data

Query Tool...

Properties...

Type	Name	
Primary Key	public.staging_pkey	auto



Configure Processor

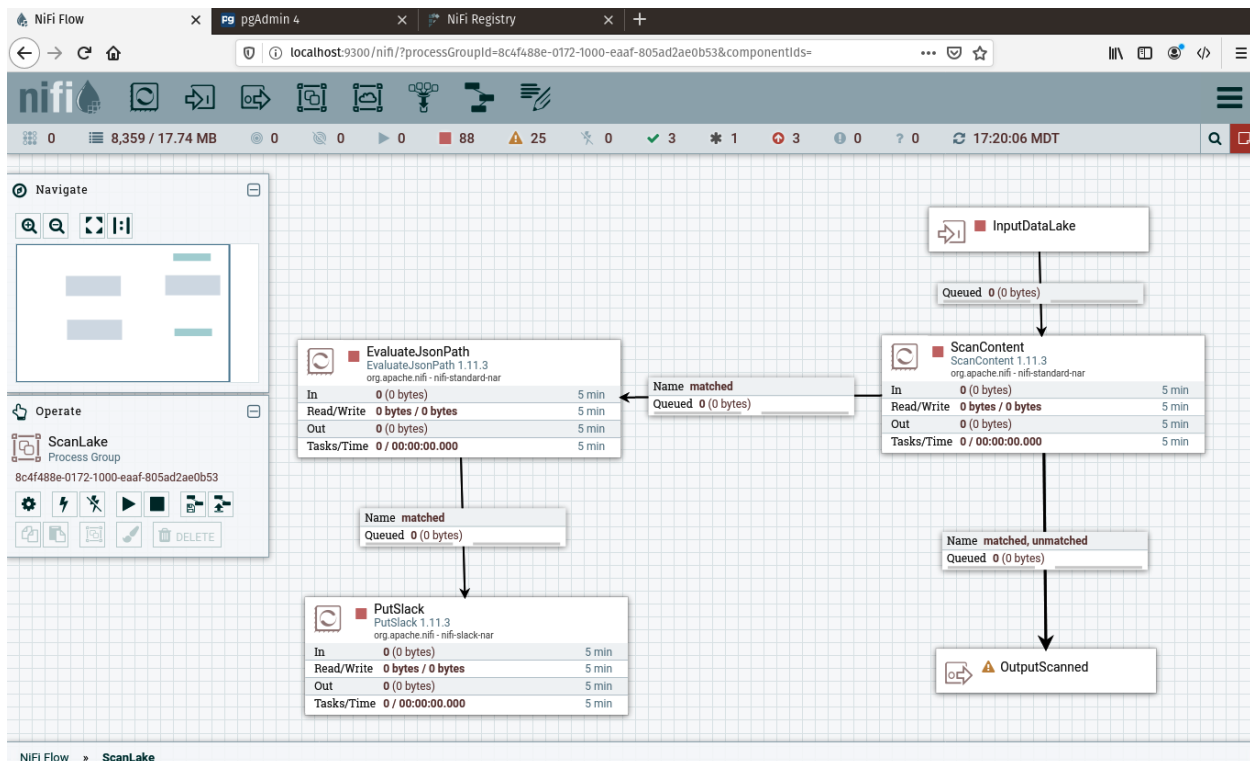
Stopped

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field +

Property	Value
Destination	flowfile-attribute
Return Type	auto-detect
Path Not Found Behavior	ignore
Null Value Representation	empty string
age	\$.age
city	\$.city
name	\$.name
state	\$.state
street	\$.street
userid	\$.userid
zip	\$.zip

CANCEL APPLY



```

paulcrickard@pop-os: ~/staging

What data would you like Great Expectations to connect to?
1. Files on a filesystem (for processing with Pandas or Spark)
2. Relational database (SQL)
: 2

Which database backend are you using?
1. MySQL
2. Postgres
3. Redshift
4. Snowflake
5. other - Do you have a working SQLAlchemy connection string?
: 2

Give your new data source a short name.
[my_postgres_db]: stagingtable

Next, we will configure database credentials and store them in the `stagingtable`
`section
of this config file: great_expectations/uncommitted/config_variables.yml:

What is the host for the postgres connection? [localhost]:
What is the port for the postgres connection? [5432]:
What is the username for the postgres connection? [postgres]:
What is the password for the postgres connection?:

```


edit_staging.validation (unsaved changes)

Logout

File Edit View Insert Cell Kernel Help

Not Trusted Kernel

```
context = ge.data_context.DataContext()

# Feel free to change the name of your suite here. Renaming this will not
# remove the other one.
expectation_suite_name = "staging.validation"
suite = context.get_expectation_suite(expectation_suite_name)
suite.expectations = []

batch_kwargs = {
    "datasource": "stagingtable",
    "limit": 1000,
    "schema": "public",
    "table": "staging",
}
batch = context.get_batch(batch_kwargs, suite)
batch.head()
```

Create & Edit Expectations

Add expectations by calling specific expectation methods on the `batch` object. They all begin with `.expect_` which makes autocompleting easy using tab.

You can see all the available expectations in the [expectation glossary](#).

Table Expectation(s)

In []:

batch.expect_table_row_count_to_be_between(max_value=0, min_value=0)

In []:

batch.expect_table_column_count_to_equal(value=7)

In []:

batch.expect_table_columns_to_match_ordered_list(
 column_list=["userid", "name", "age", "street", "city", "state", "zip"]
)

Column Expectation(s)

No column level expectations are in this suite. Feel free to add some here. They all begin with `batch.expect_column....`

Save & Review Your Expectations

Status	Expectation	Observed Value
✔	Must have exactly 7 columns.	7
✔	Must have these columns in this order: <code>userid</code> , <code>name</code> , <code>age</code> , <code>street</code> , <code>city</code> , <code>state</code> , <code>zip</code>	<code>['userid', 'name', 'age', 'street', 'city', 'state', 'zip']</code>

Chapter 12: Building an Apache Kafka Cluster



HOME
INTRODUCTION
QUICKSTART
USE CASES
DOCUMENTATION
PERFORMANCE
POWERED BY
PROJECT INFO
ECOSYSTEM
CLIENTS
EVENTS
CONTACT US
APACHE

Download

@apachekafka

PUBLISH & SUBSCRIBE

Read and write streams of data like a messaging system.
[Learn more »](#)

PROCESS

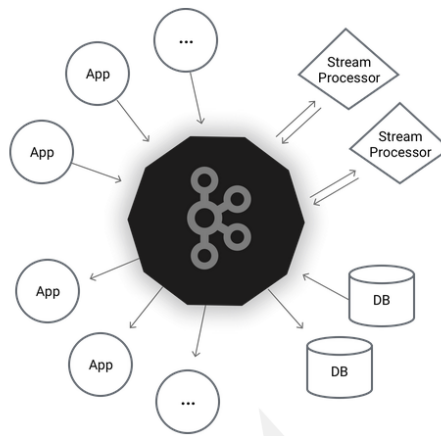
Write scalable stream processing applications that react to events in real-time.
[Learn more »](#)

STORE

Store streams of data safely in a distributed, replicated, fault-tolerant cluster.
[Learn more »](#)

LATEST NEWS

KAFKA SUMMIT 2020
AUG 24 - AUG 25, 2020
AK RELEASE 2.5.0
APRIL 15, 2020
AK RELEASE 2.4.1
MARCH 12, 2019
AK RELEASE 2.2.2
DECEMBER 1, 2019
AK RELEASE 2.3.1
OCTOBER 24, 2019



Welcome to Apache ZooKeeper?

Apache ZooKeeper is an effort to develop and maintain an open-source server which enables highly reliable distributed coordination.

What is ZooKeeper?

ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. All of these kinds of services are used in some form or another by distributed applications. Each time they are implemented there is a lot of work that goes into fixing the bugs and race conditions that are inevitable. Because of the difficulty of implementing these kinds of services, applications initially usually skip on them, which make them brittle in the presence of change and difficult to manage. Even when done correctly, different implementations of these services lead to management complexity when the applications are deployed.

Learn more about ZooKeeper on the ZooKeeper Wiki.

Getting Started

Start by installing ZooKeeper on a single machine or a very small cluster.

1. Learn about ZooKeeper by reading the documentation.
2. Download ZooKeeper from the release page.

Getting Involved

Apache ZooKeeper is an open source volunteer project under the Apache Software Foundation. We encourage you to learn about the project and contribute your expertise. Here are some starter links:

1. See our How to Contribute to ZooKeeper page.
2. Give us feedback: What can we do better?
3. Join the mailing list: Meet the community.

Copyright © 2010-2020 The Apache Software Foundation, Licensed under the Apache License, Version 2.0.

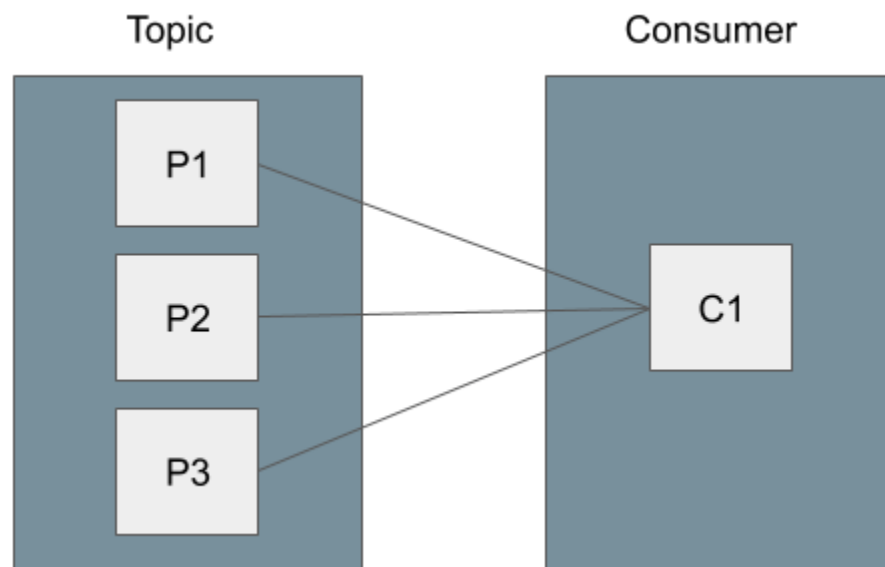
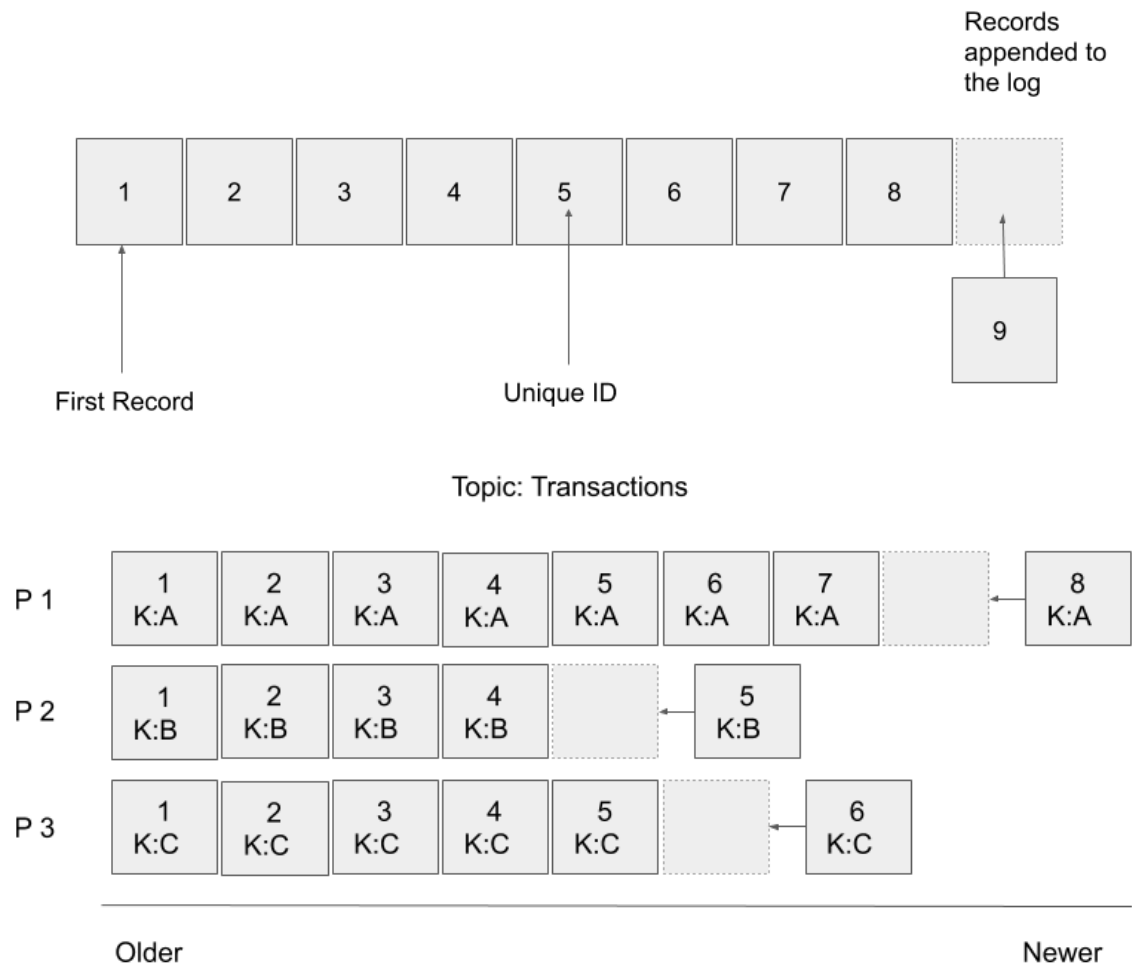
Apache ZooKeeper, ZooKeeper, Apache, the Apache feather logo, and the Apache ZooKeeper project logo are trademarks of The Apache Software Foundation.

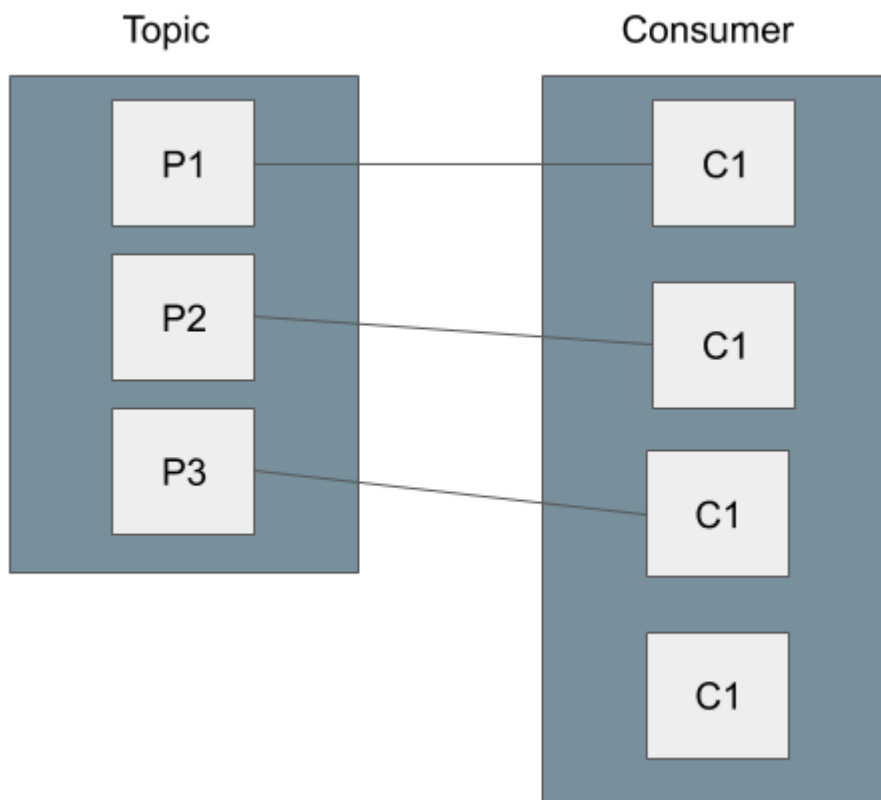
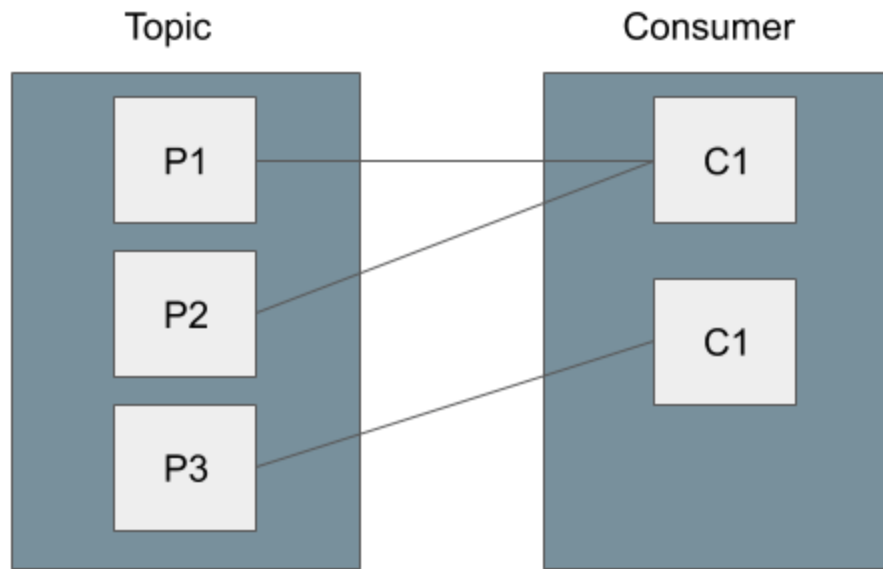


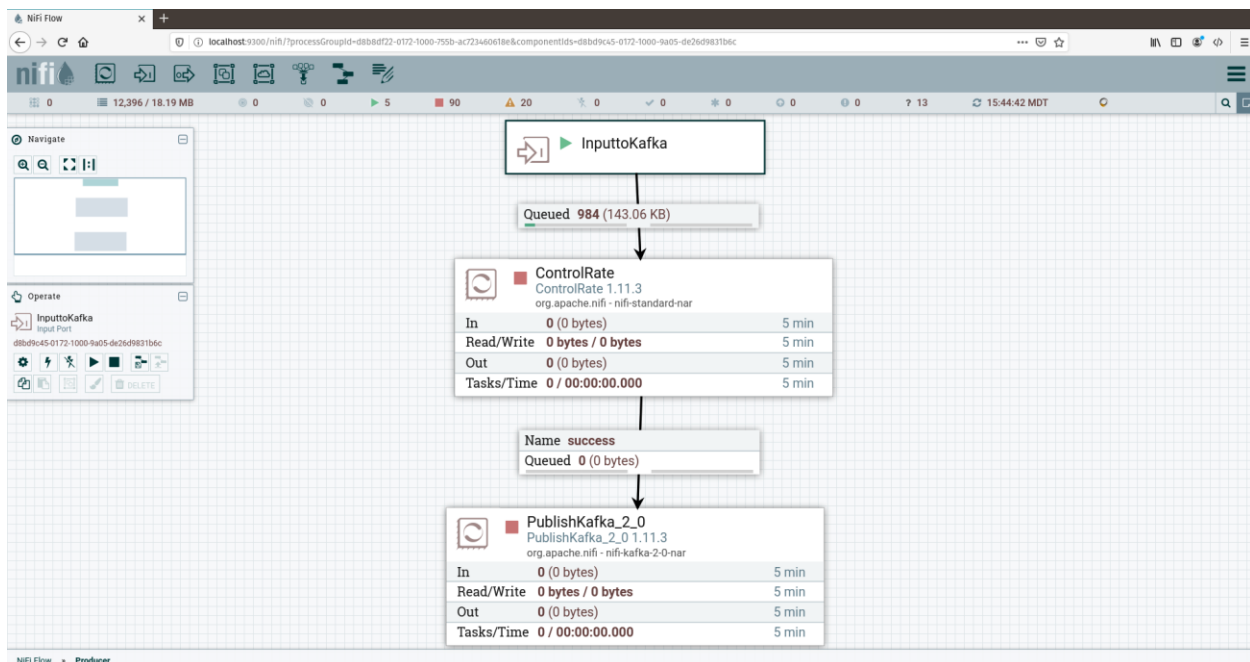
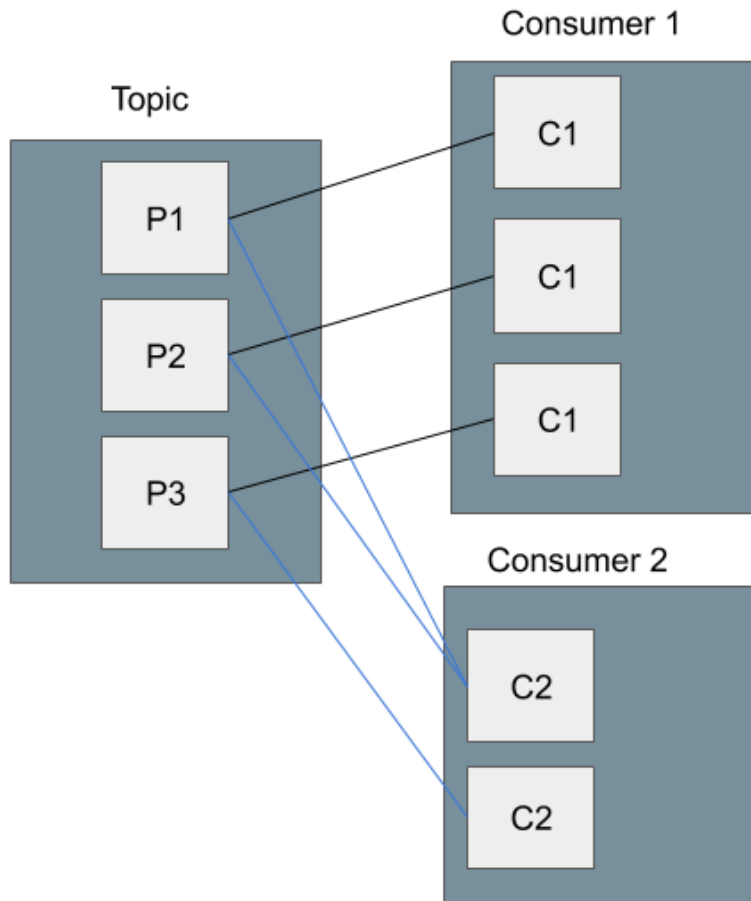
```
paulcrickard@pop-os: ~/kafka_1
paulcrickard@pop-os:~/kafka_1$ bin/kafka-console-producer.sh --broker-list localhost:9092,localhost:9093,localhost:9094 --topic dataengineering
>first message
>second message
>new message
>Hello from Kafka
>
```

```
paulcrickard@pop-os:~/kafka_1$ bin/kafka-console-consumer.sh --bootstrap-server localhost:9092,localhost:9093,localhost:9094 --topic dataengineering --from-beginning
first message
second message
first message
second message
new message
Hello from Kafka
```


Chapter 13: Streaming Data with Kafka







Create Connection

DETAILS

SETTINGS

From Output

OutputDataLake



To Input

InputtoKafka



Within Group

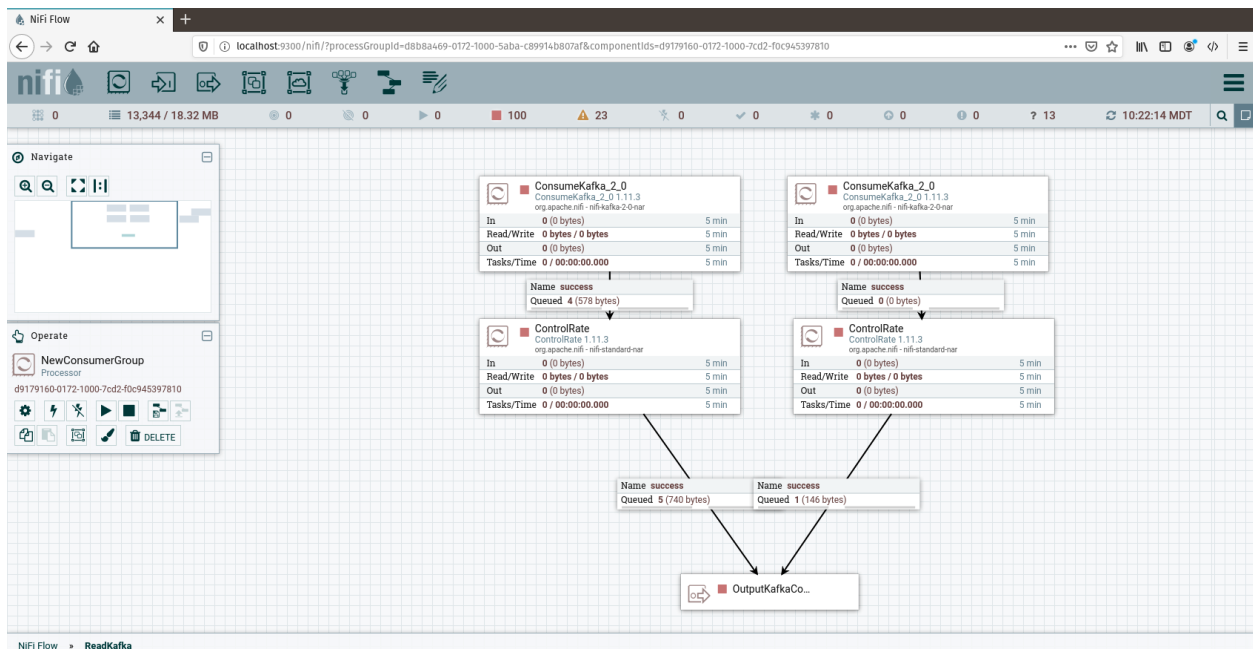
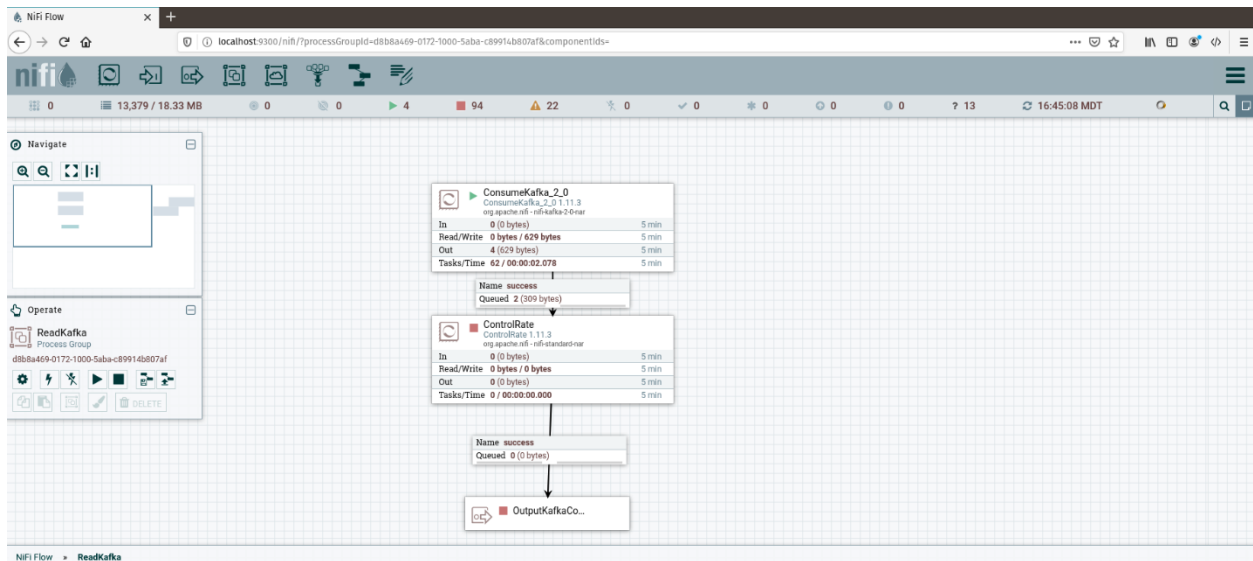
ReadDataLake

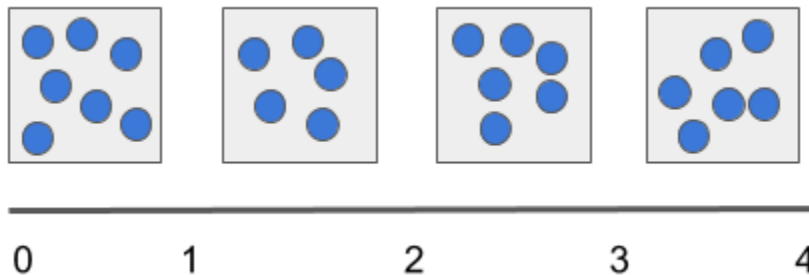
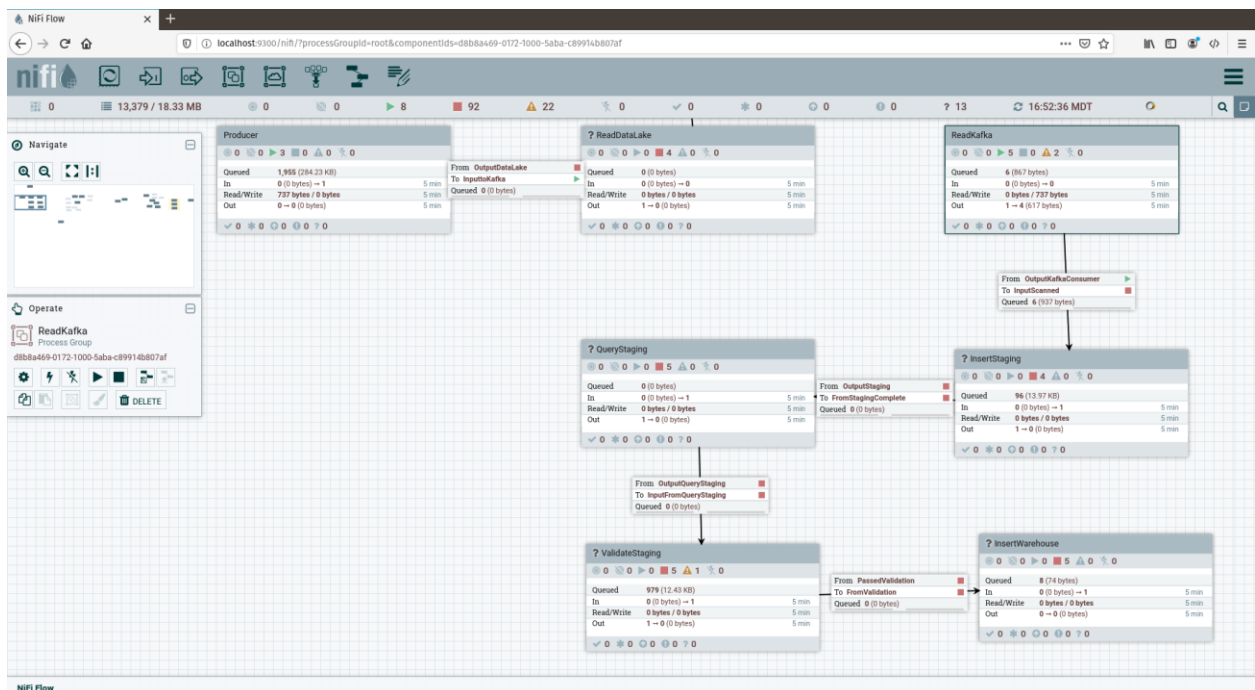
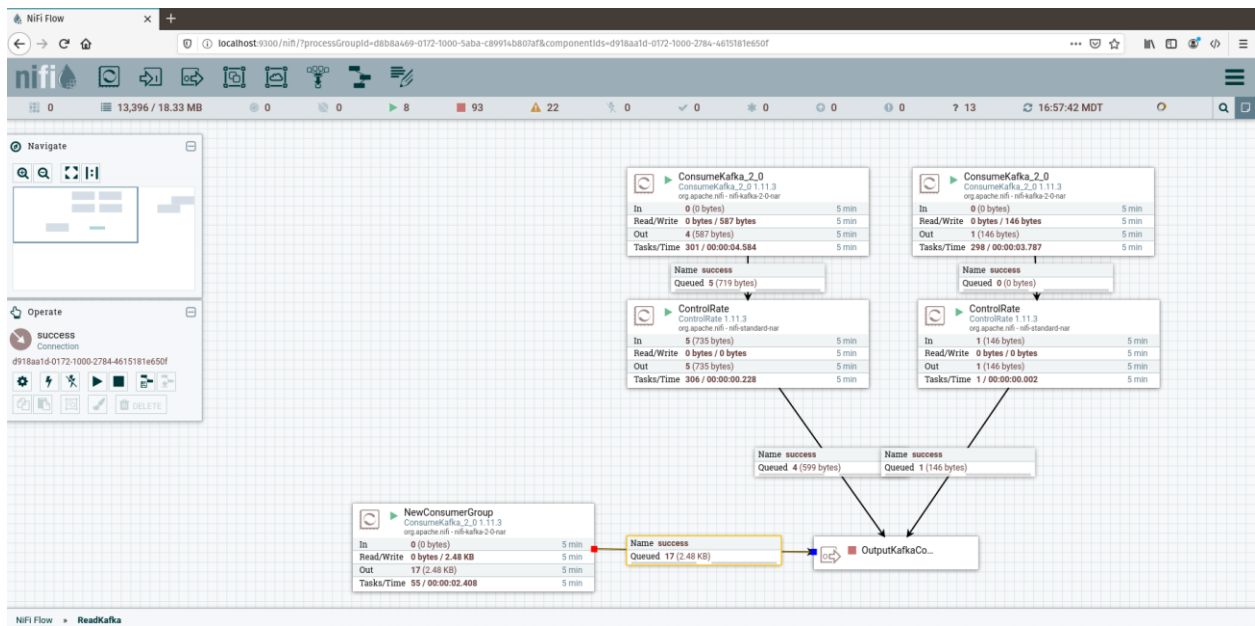
Within Group

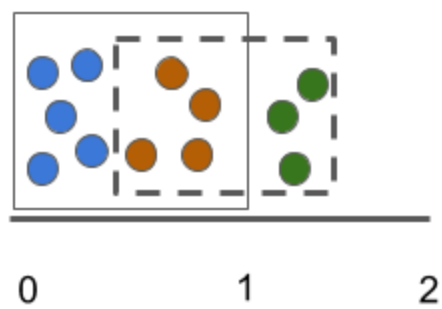
Producer

CANCEL

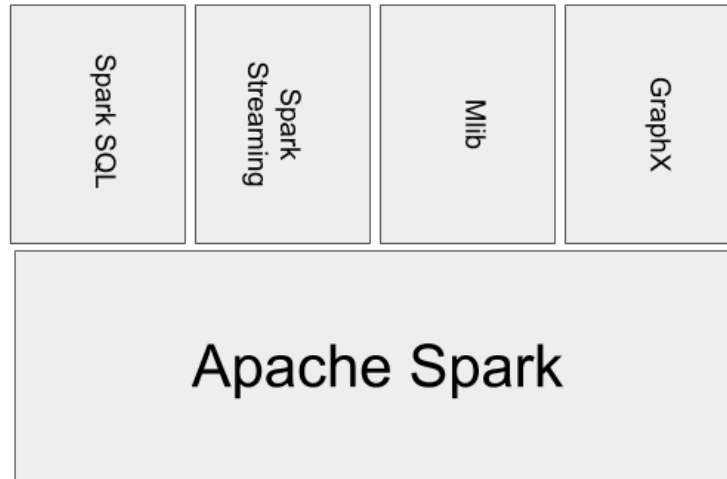
ADD








Chapter 14: Data Processing with Apache Spark



Apache Spark™ - Unified Analytics Engine

spark.apache.org



Lightning-fast unified analytics engine

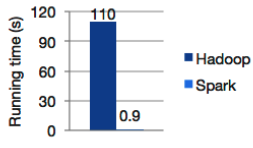
Download Libraries Documentation Examples Community Developers Apache Software Foundation

Apache Spark™ is a unified analytics engine for large-scale data processing.

Speed

Run workloads 100x faster.

Apache Spark achieves high performance for both batch and streaming data, using a state-of-the-art DAG scheduler, a query optimizer, and a physical execution engine.



Tool	Running time (s)
Hadoop	110
Spark	0.9

Logistic regression in Hadoop and Spark

Ease of Use

Write applications quickly in Java, Scala, Python, R, and SQL.

```
df = spark.read.json("logs.json")
df.where("age > 21")
  .select("name.first").show()
```

Spark's Python DataFrame API
Read JSON files with automatic schema inference

Generality

Combine SQL, streaming, and complex analytics.

Spark SQL

Spark Streaming


MLlib (machine learning)

GraphX (graph)

Latest News

- Spark 3.0.0 released (Jun 18, 2020)
- Spark+AI Summit (June 22-25th, 2020, VIRTUAL) agenda posted (Jun 15, 2020)
- Spark 2.4.6 released (Jun 05, 2020)
- Spark 2.4.5 released (Feb 08, 2020)

[Archive](#)



APACHECON @home Sep 29 - Oct 1, 2020

[Download Spark](#)

Built-in Libraries:

- [SQL and DataFrames](#)
- [Spark Streaming](#)
- [MLlib \(machine learning\)](#)
- [GraphX \(graph\)](#)

[Third-Party Projects](#)

Download Apache Spark™

- Choose a Spark release: **3.0.0 (Jun 18 2020)**
- Choose a package type: **Pre-built for Apache Hadoop 2.7**
- Download Spark: [spark-3.0.0-bin-hadoop2.7.tgz](#)
- Verify this release using the 3.0.0 [signatures](#), [checksums](#) and [project release KEYS](#).

Note that, Spark 2.x is pre-built with Scala 2.11 except version 2.4.2, which is pre-built with Scala 2.12. Spark 3.0+ is pre-built with Scala 2.12.

Latest News

Spark 3.0.0 released (Jun 18, 2020)
Spark+AI Summit (June 22-25th, 2020, VIRTUAL) agenda posted (Jun 15, 2020)
Spark 2.4.6 released (Jun 05, 2020)
Spark 2.4.5 released (Feb 08, 2020)

[Archive](#)

Spark Master at spark://pop-os.localdomain:7077

URL: spark://pop-os.localdomain:7077
Alive Workers: 1
Cores in use: 2 Total, 0 Used
Memory in use: 2.7 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (1)

Worker Id	Address	State	Cores	Memory	Resources
worker-20200628195324-10.0.0.148-9911	10.0.0.148:9911	ALIVE	2 (0 Used)	2.7 GiB (0.0 B Used)	

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

```
Welcome to
Spark version 3.0.0

Using Python version 3.7.5 (default, Nov 20 2019 09:21:52)
SparkSession available as 'spark'.
>>>
```


Get Started with PySpark x Home Page - Select or create a notebook x Spark - Jupyter Notebook x Examples | Apache Spark x +

localhost:8888/notebooks/Spark.ipynb

Jupyter Spark Last Checkpoint: an hour ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Help Trusted Python 3

```
In [1]: import findspark
findspark.init()

In [18]: import pyspark
from pyspark.sql import SparkSession
spark=SparkSession.builder.master("spark://pop-os.localdomain:7077").getOrCreate()

In [21]: import random
NUM_SAMPLES=1

def inside(p):
    x, y = random.random(), random.random()
    return x*x + y*y < 1

count = spark.sparkContext.parallelize(range(0,NUM_SAMPLES)).filter(inside).count()
print("Pi is roughly {}".format(4.0 * count / NUM_SAMPLES))

Pi is roughly 4.0

In [22]: spark.stop()
```

Get Started with PySpark x Home Page - Select or create a notebook x Spark - Jupyter Notebook x Examples | Apache Spark x Spark Master at spark://pop-os.localdomain:7077 x pyspark.sql module - PySpark x +

localhost:8080

Spark Master at spark://pop-os.localdomain:7077

URL: spark://pop-os.localdomain:7077
Alive Workers: 1
Cores in use: 2 Total, 2 Used
Memory in use: 2.7 GiB Total, 1024.0 MiB Used
Resources in use:
Applications: 1 Running, 2 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (1)

Worker Id	Address	State	Cores	Memory	Resources
worker-20200629125219-10.0.0.148-9911	10.0.0.148:9911	ALIVE	2 (2 Used)	2.7 GiB (1024.0 MiB Used)	

Running Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20200629195112-0002	(kill) Pi-Estimation	2	1024.0 MiB		2020/06/29 19:51:12	paulcrickard	RUNNING	4 s

Completed Applications (2)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20200629194952-0001	pyspark-shell	2	1024.0 MiB		2020/06/29 19:49:52	paulcrickard	FINISHED	14 s
app-20200629180816-0000	pyspark-shell	2	1024.0 MiB		2020/06/29 18:08:16	paulcrickard	FINISHED	34 min

```
+-----+-----+-----+-----+-----+-----+-----+
|_c0|_c1|_c2|_c3|_c4|_c5|_c6|_c7|
+-----+-----+-----+-----+-----+-----+-----+
|name|age|street|city|state|zip|lng|lat|
|Patrick Hendrix|23|5755 Jonathan Ranch|New Sheriland|Wisconsin|60519|103.914462|-59.0094375|
|Grace Jackson|36|2502 Stewart Plaz...|Ramirezville|Arizona|91946|170.503858|58.1631665|
|Arthur Garcia|61|627 Liu Brooks|Freemanhaven|Kansas|97783|-39.845646|38.689889|
|Gary Valentine|29|9682 Theresa Vist...|Allenborough|Oregon|81537|-30.304522|81.2722995|
+-----+-----+-----+-----+-----+-----+-----+
```

only showing top 5 rows

root

```
-- _c0: string (nullable = true)
-- _c1: string (nullable = true)
-- _c2: string (nullable = true)
-- _c3: string (nullable = true)
-- _c4: string (nullable = true)
-- _c5: string (nullable = true)
-- _c6: string (nullable = true)
-- _c7: string (nullable = true)
```


name	age	street	city	state	zip	lng	lat
Patrick Hendrix	23	5755 Jonathan Ranch	New Sheriland	Wisconsin	60519	103.914462	-59.0094375
Grace Jackson	36	2502 Stewart Plaz...	Ramirezville	Arizona	91946	170.503858	58.1631665
Arthur Garcia	61	627 Liu Brooks	Freemanhaven	Kansas	97783	-39.845646	38.689889
Gary Valentine	29	9682 Theresa Vist...	Allenborough	Oregon	81537	-30.304522	81.2722995
Erin Mclean	23	9349 Williams Lan...	East Markmouth	Ohio	4300	-110.860085	11.476733

only showing top 5 rows

root

```
-- name: string (nullable = true)
-- age: integer (nullable = true)
-- street: string (nullable = true)
-- city: string (nullable = true)
-- state: string (nullable = true)
-- zip: integer (nullable = true)
-- lng: double (nullable = true)
-- lat: double (nullable = true)
```


Chapter 15: Real-Time Edge Data – Kafka, Spark, and MiNiFi

[Project](#)[Documentation](#)[Downloads](#)[Community](#)[Development](#)[ASF Links](#)[Apache NiFi](#)

A subproject of Apache NiFi to collect data where it originates.

About

MiNiFi—a subproject of Apache NiFi—is a complementary data collection approach that supplements the core tenets of NiFi in dataflow management, focusing on the collection of data at the source of its creation.

Specific goals for the initial thrust of the MiNiFi effort comprise:

- Small size and low resource consumption
- Central management of agents
- Generation of data provenance (full chain of custody of information)
- Integration with NiFi for follow-on dataflow management

Perspectives of the role of MiNiFi should be from the perspective of the agent acting immediately at, or directly adjacent to, source sensors, systems, or servers.

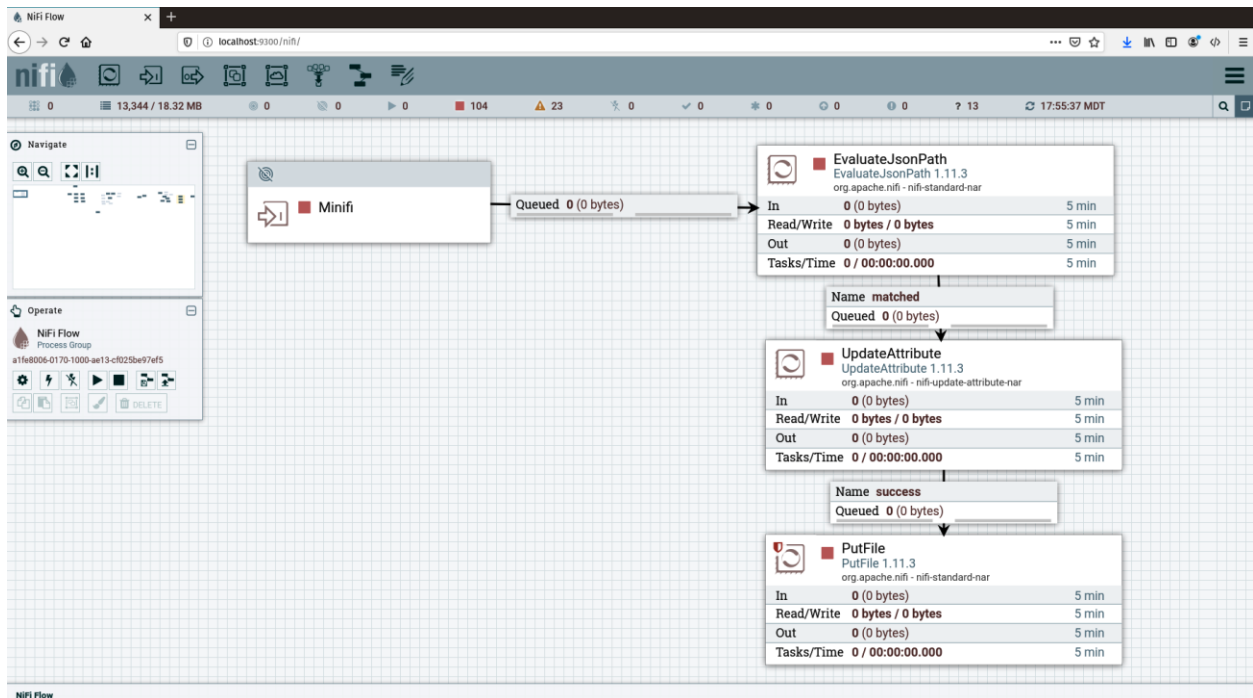
```
if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi

export JAVA_HOME=/usr/lib/jvm/java-1.11.0-openjdk.amd64
export SPARK_HOME=/home/paulcrickard/spark3
export PATH=$SPARK_HOME/bin:$PATH

export MINIFI_HOME=/home/paulcrickard/minifi
export PATH=$MINIFI_HOME/bin:$PATH
```



```
# Site to Site properties
nifi.remote.input.host=
nifi.remote.input.secure=false
nifi.remote.input.socket.port=1026|
nifi.remote.input.http.enabled=true
nifi.remote.input.http.transaction.ttl=30 sec
nifi.remote.contents.cache.expiration=30 secs
```



The screenshot shows the 'Configure Remote Process Group' dialog box in the NiFi Flow console. The dialog is for a process group named 'Minifi Flow' with ID '01721015-ec63-1c0b-7fed-73ccd6920c83'. The 'URLs' field is set to 'http://localhost:9300'. The 'Transport Protocol' is set to 'HTTP'. The 'Local Network Interface' is set to 'Local Network Interface'. The 'HTTP Proxy Server Hostname' and 'HTTP Proxy Server Port' are empty. The 'HTTP Proxy User' and 'HTTP Proxy Password' are empty. The 'Communications Timeout' is set to '30 sec' and the 'Yield Duration' is set to '10 sec'. The 'APPLY' button is highlighted.

NiFi Flow

localhost:9300/nifi/7processGroupId-01721002-ec63-1c0b-cfa7-c49844f60820&componentId=

0 13,344 / 18.32 MB 1 0 0 105 23 0 0 0 0 0 13 19:43:46 MDT

Navigate

minifitask
Process Group
01721002-ec63-1c0b-cfa7-c49844f60820

GenerateFlowFile
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

To Minifi
Name success
Queued 0 (0 bytes)

NiFi Flow
http://localhost:9300/nifi
Sent 0 (0 bytes) → 1 5 min
Received 0 → 0 (0 bytes) 5 min
07/05/2020 19:43:30 MDT

NiFi Flow > minifitask

```
paulcrickard@pop-os: ~/minifi-toolkit/bin

paulcrickard@pop-os:~/minifi-toolkit/bin$ ./config.sh transform /home/paulcrickard/Downloads/minifitask.xml /home/paulcrickard/minifi-templates/config.yml

Java home: /usr/lib/jvm/java-1.8.0-openjdk-amd64
MiNiFi Toolkit home: /home/paulcrickard/minifi-toolkit

No validation errors found in converted configuration.
paulcrickard@pop-os:~/minifi-toolkit/bin$
```

NiFi Flow

10.0.0.63:8081/nifi/

1 2 / 92 bytes 1 0 1 4 0 0 0 0 0 0 0 14:48:47 MDT

Navigate

NiFi Flow
Process Group
25d62348-0173-1000-7e9d-0035461cb2a5

FromMinifi
Queued 2 (92 bytes)

EvaluateJsonPath
EvaluateJsonPath 1.9.0
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 matched
Queued 0 (0 bytes)

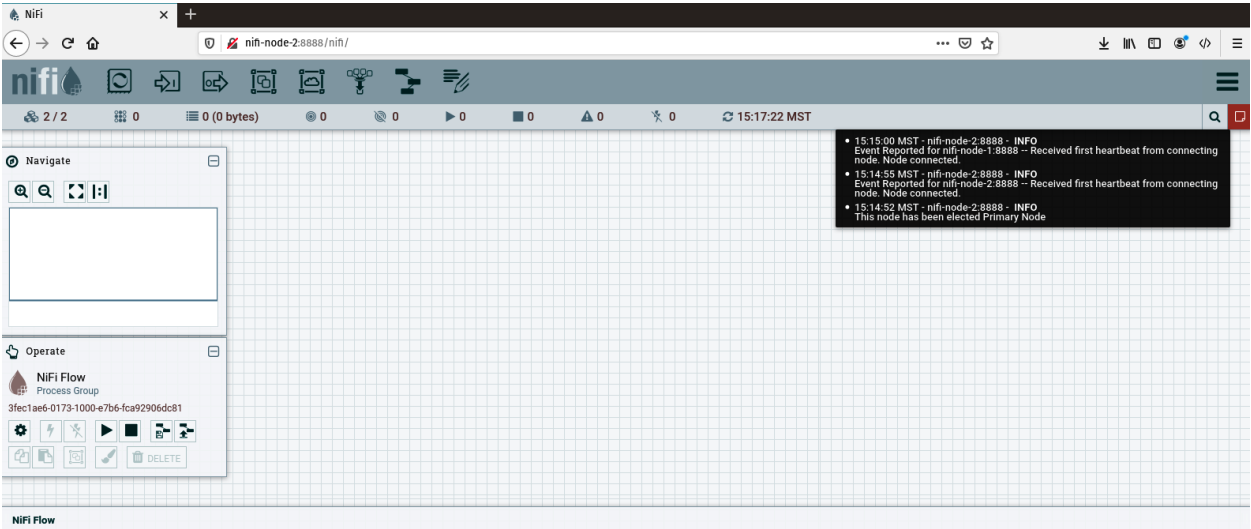
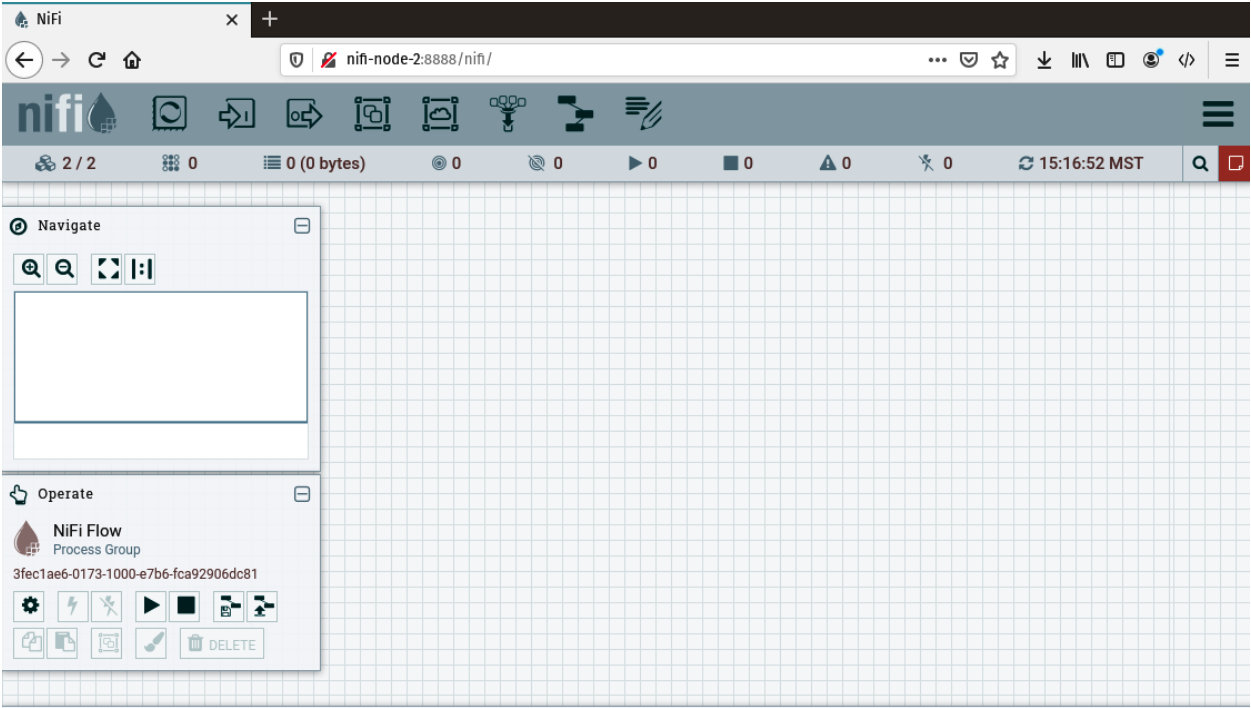
UpdateAttribute
UpdateAttribute 1.9.0
org.apache.nifi - nifi-update-attribute-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name success
Queued 0 (0 bytes)

PutFile
PutFile 1.9.0
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

Minifi-DataPipeline
1 0 0 1 0 0 0
Queued 0 (0 bytes) 5 min
In 0 (0 bytes) → 0 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 → 0 (0 bytes) 5 min
0 0 0 0 0 0 0

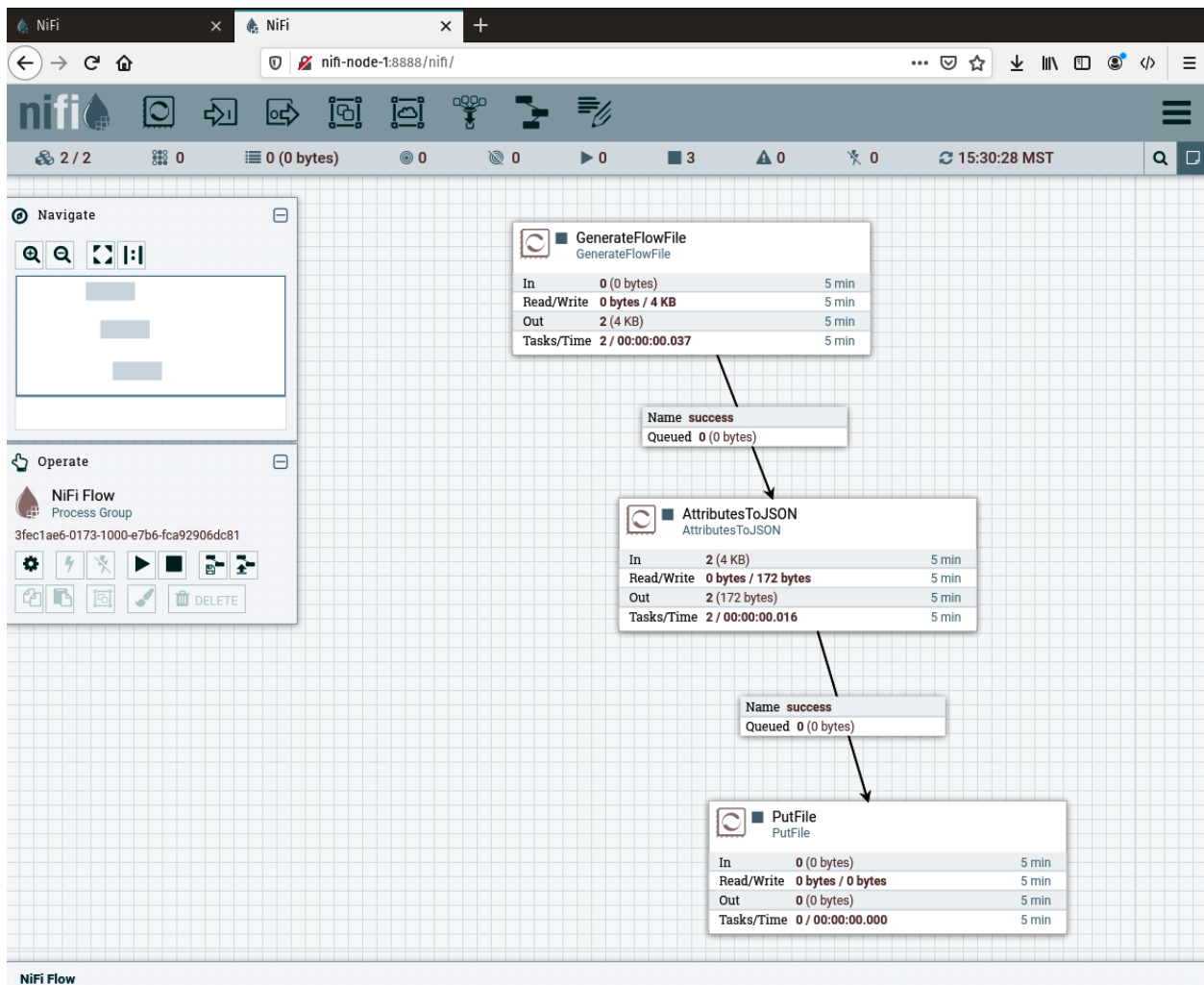
Appendix



NiFi Cluster

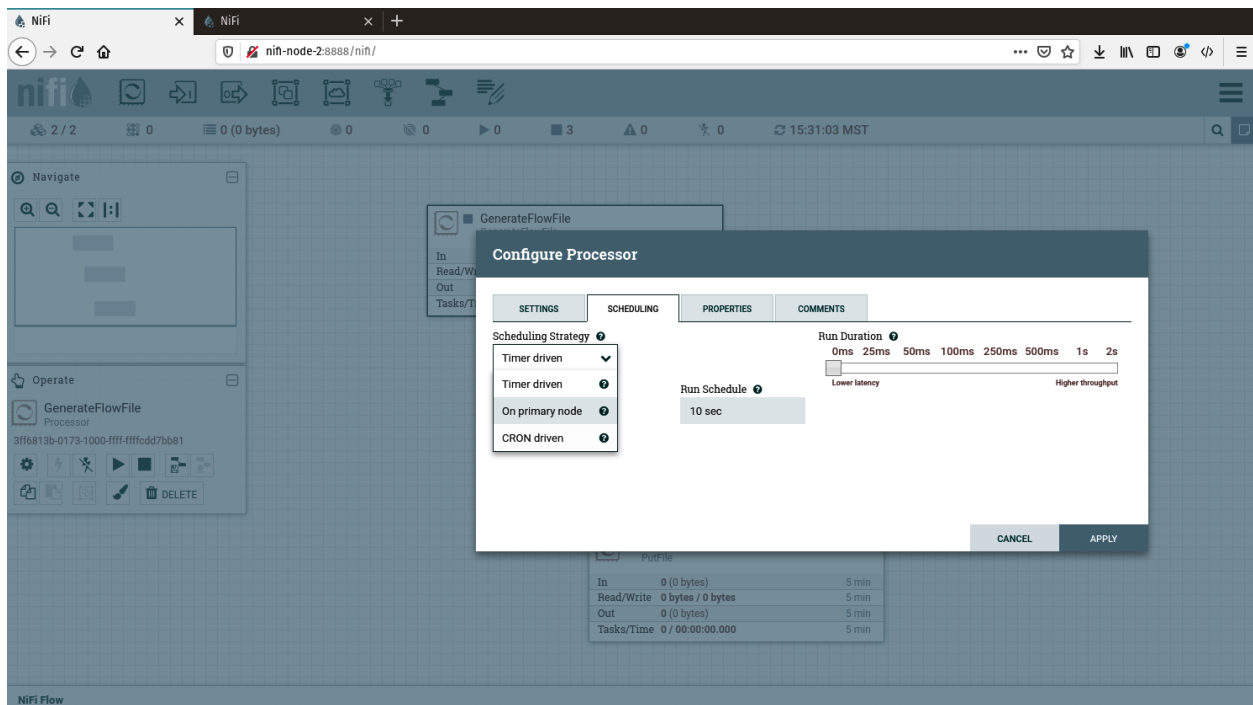
Displaying 2 of 2

Filter		by address				
Node Address	Active Thread Count	Queue / Size	Status	Uptime	Last Heartbeat	
nifi-node-1.8888	0	0 / 0 bytes	CONNECTED	07/11/2020 16:11:47 MDT	07/11/2020 16:18:51 MDT	🔌
nifi-node-2.8888	0	0 / 0 bytes	CONNECTED, PRIMARY, COORDINATOR	07/11/2020 16:14:48 MDT	07/11/2020 16:18:47 MDT	🔌



```
paulcrickard@pop-os: ~/output

paulcrickard@pop-os:~/output$ ls
4274490619179 4314534837026 4354551784292 4394578909080 data.txt
4284523756564 4324537815176 4364556613360 4404582691395
4294527247783 4334542190236 4374560985786 4414585033417
4304531894134 4344546106426 4384567644867 4424589700344
paulcrickard@pop-os:~/output$ cat 4354551784292
{"path": "./", "filename": "4354551784292", "uuid": "556872b1-f434-47f1-a610-5893f3192441"}paulcrickard@pop-os:~/output$
```

NiFi Cluster

Displaying 2 of 2

Filter	by address						
	Node Address	Active Thread Count	Queue / Size	Status	Uptime	Last Heartbeat	
❏	nifi-node-1.8888	0	4 / 2.25 KB	CONNECTED	07/11/2020 16:11:47 MDT	07/11/2020 16:38:55 MDT	🔄
❏	nifi-node-2.8888	0	4 / 2.25 KB	CONNECTED, PRIMARY, COORDINATOR	07/11/2020 16:14:48 MDT	07/11/2020 16:38:54 MDT	🔄
🔄 Last updated: 16:38:55 MDT							

NiFi Cluster

Displaying 2 of 2

Filter	by address							
Node Address		Active Thread Count	Queue / Size	Status	Uptime	Last Heartbeat		
❏	nifi-node-1.8888	0	8 / 10.25 KB	CONNECTED	07/11/2020 17:00:55 MDT	07/11/2020 17:06:41 MDT	🔄	
❏	nifi-node-2.8888	0	5 / 4.25 KB	CONNECTED, PRIMARY, COORDINATOR	07/11/2020 16:54:18 MDT	07/11/2020 17:06:41 MDT	🔄	
🔄 Last updated: 17:06:41 MDT								


NiFi Cluster

Displaying 2 of 2

Filter

by address

Node Address	Active Thread Count	Queue / Size	Status	Uptime	Last Heartbeat
nifi-node-1.8888			DISCONNECTED	No value previously set	No value previously set
nifi-node-2.8888	0	11 / 16.25 KB	CONNECTED, PRIMARY, COORDINATOR	07/11/2020 16:54:18 MDT	07/11/2020 17:07:36 MDT


 Last updated: 17:07:38 MDT


NiFi Cluster

Displaying 2 of 2

Filter

by address

Node Address	Active Thread Count	Queue / Size	Status	Uptime	Last Heartbeat
nifi-node-1.8888	0	13 / 20.25 KB	CONNECTED	07/11/2020 17:00:55 MDT	07/11/2020 17:08:32 MDT
 nifi-node-2.8888	0	13 / 20.25 KB	CONNECTED, PRIMARY, COORDINATOR	07/11/2020 16:54:18 MDT	07/11/2020 17:08:32 MDT

 Last updated: 17:08:35 MDT