

CLUDERA

# Introduction to the Mm.. FLaNK Stack

Timothy Spann



# Welcome to Future of Data - Princeton



<https://www.meetup.com/futureofdata-princeton/>

From Big Data to AI to Streaming to Containers to Cloud to Analytics to Cloud Storage to Fast Data to Machine Learning to Microservices to ...



@PaasDev

## Today's Lead

Who am I?

# Data in Motion Field Engineer



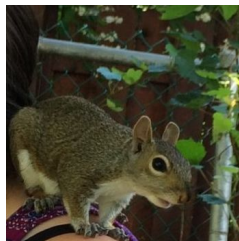
DZone Zone Leader and Big Data MVB;  
Princeton NJ Future of Data Meetup;  
ex-Pivotal Field Engineer;  
Author of Apache Kafka RefCard  
<https://github.com/tspannhw>  
<https://www.datainmotion.dev/>



@PaasDev

# Mm.. FLaNK

<https://github.com/tspannhw/MmFLaNK>



Flink

APACHE

nifi



APACHE

kafka



minifi



MINiFi  
Agent



mxnet



Store in Apache Kudu Table	
Read/Write	35.97 KB / 0 bytes
Out	0 (0 bytes)
Task/Time	9 / 00:00:00.182



mxnet

cloudera  
EDGE MANAGEMENT

APACHE NIFI  
registry

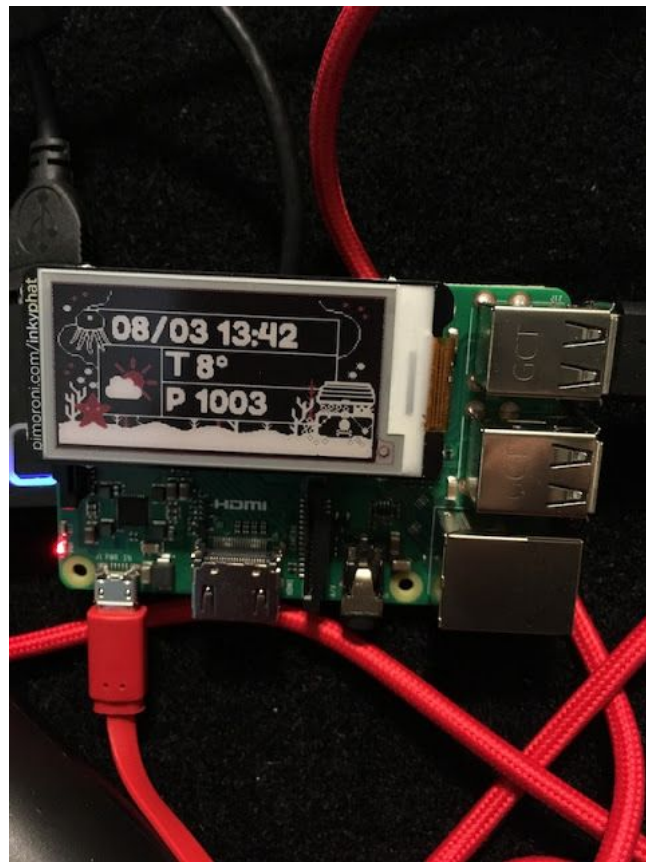


SCHEMA  
REGISTRY

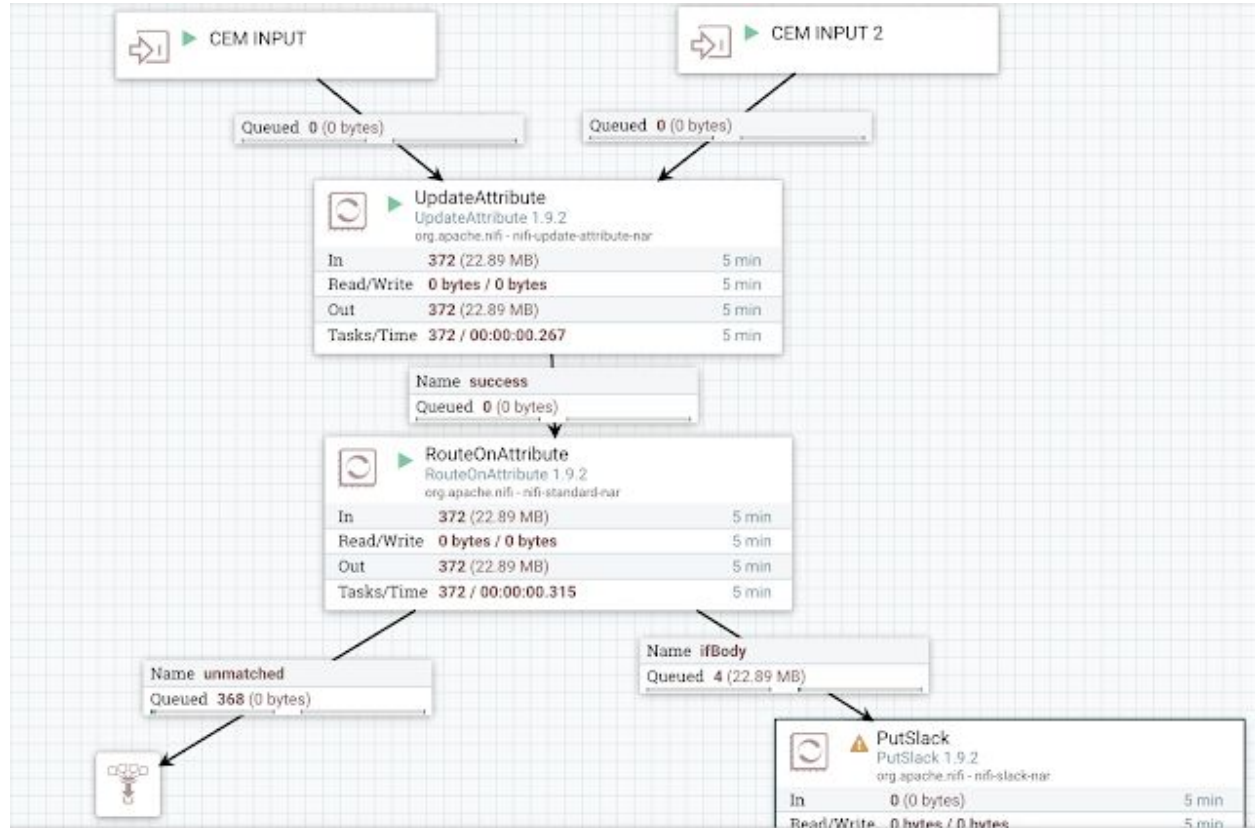
<https://www.datainmotion.dev/2019/11/introducing-mm-flank-apache-flink-stack.html>

# MiNiFi.. FLaNK

- <https://www.datainmotion.dev/2019/03/using-raspberry-pi-3b-with-apache-nifi.html>
- <https://www.datainmotion.dev/2019/05/cloudera-edge-management-introduction.html>
- <https://www.datainmotion.dev/2019/11/running-demo-apache-flink-application.html>
- <https://www.datainmotion.dev/2019/11/learning-apache-flink-19.html>
- <https://www.datainmotion.dev/2019/10/migrating-apache-flume-flows-to-apache-42.html>

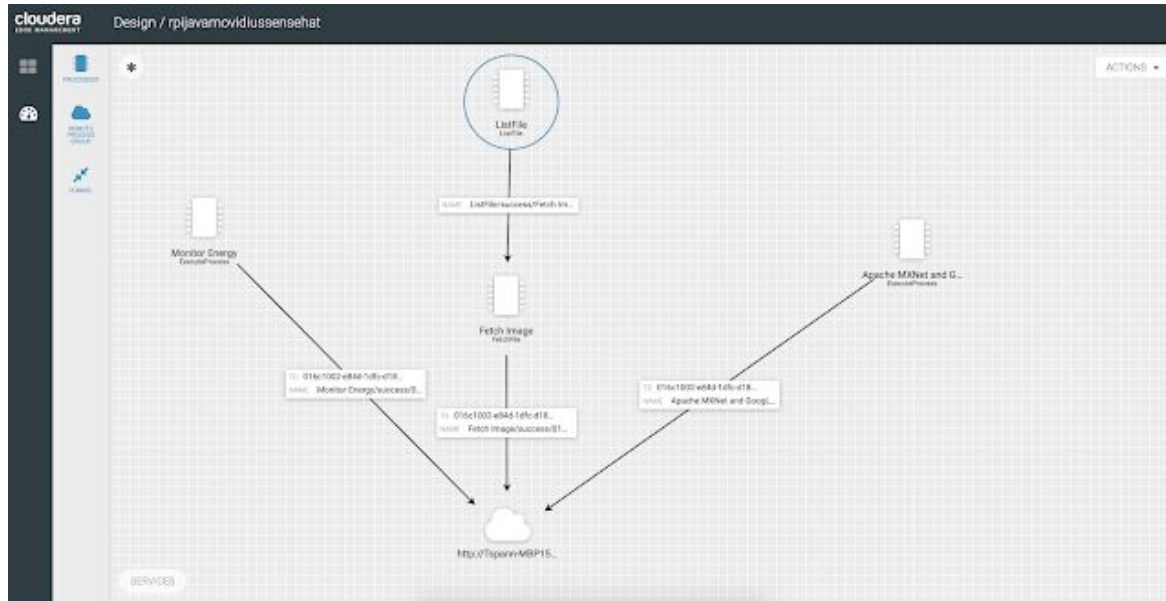


# MiNiFi.m.. FLaNK





# MiNiFi.m.. FLaNK



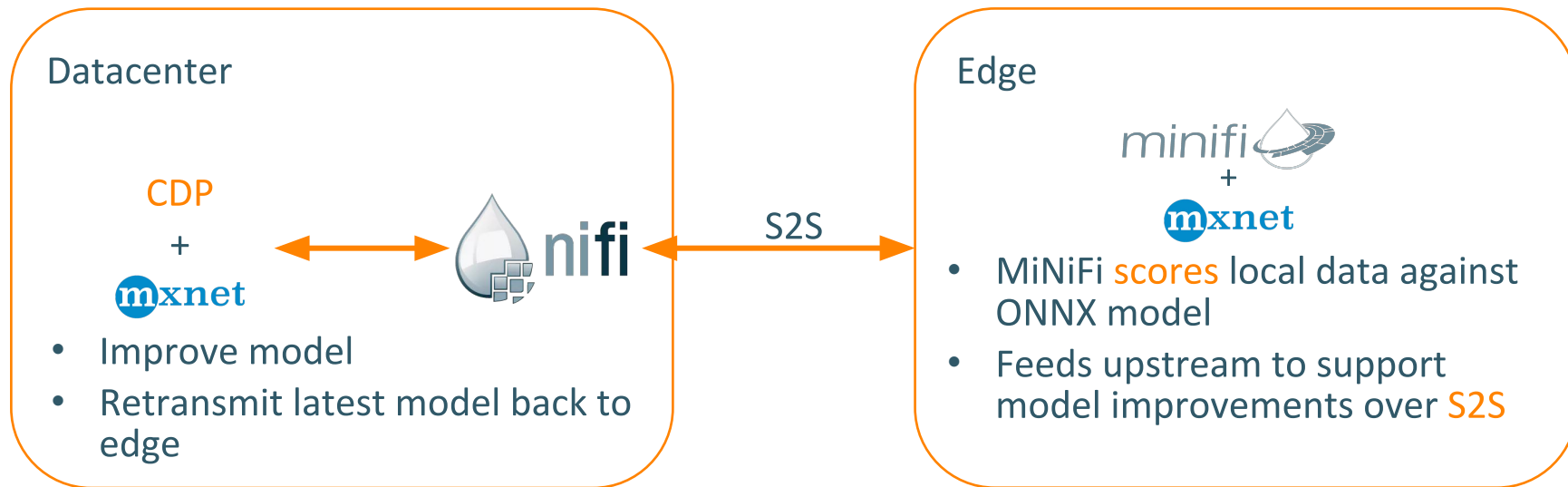
# APACHE MINIFI C++ 0.6.0

## Key New Features

Feature	Description	Apache JIRA
NiFi JNI Bindings	Ability to run any native NiFi processor via JNI. Makes all NiFi processors available to MiNiFi C++	<a href="#">MINIFICPP-740</a>
Native Python Processors	Ability to write native MiNiFi C++ processors in Python	<a href="#">MINIFICPP-750</a>
PublishKafka	Ability to write to secured Kafka instance	<a href="#">MINIFICPP-731</a>
CoAP Support	Ability to communicate with EFM server over CoAP. Drastically reduces the network impact of heartbeats	<a href="#">MINIFICPP-759</a>
Windows Support	Ability to install MiNiFi via a MSI	<a href="#">MINIFICPP-700</a>



# MiNiFi C++ - MiNiFi IoT AI



<https://www.datainmotion.dev/2019/08/rapid-iot-development-with-cloudera.html>

The screenshot displays the AWS Step Functions console interface. At the top, a header bar shows system metrics: 709 / 8.16 MB, 0, 38, 1, 547, 232, 0, 3, 0, 0, 0, 0, 1, and a timestamp of 15:34:12 EST. Below the header, a workflow diagram is visible, showing a sequence of tasks and their state transitions. The tasks include:

- GetHTTP** (GetHTTP 1.8.0): A task that reads from a standard input and writes to a standard output. It has a state transition to **Name success**.
- RouteOnAttribute** (RouteOnAttribute 1.8.0): A task that reads from a standard input and writes to a standard output. It has a state transition to **Name success**.
- InferenceProcessor** (InferenceProcessor 1.8.0): A task that reads from a standard input and writes to a standard output. It has a state transition to **Name failure**.
- Resizemage** (Resizemage 1.8.0): A task that reads from a standard input and writes to a standard output. It has a state transition to **Name Response**.
- PutFile** (PutFile 1.8.0): A task that reads from a standard input and writes to a standard output. It has a state transition to **Name success**.
- PutS3Object** (PutS3Object 1.8.0): A task that reads from a standard input and writes to a standard output. It has a state transition to **Name success**.

The diagram shows the flow of data and state transitions between these tasks, with state transitions labeled as **Name success**, **Name failure**, and **Name Response**. The console interface at the top provides system metrics and a timeline for the workflow execution.

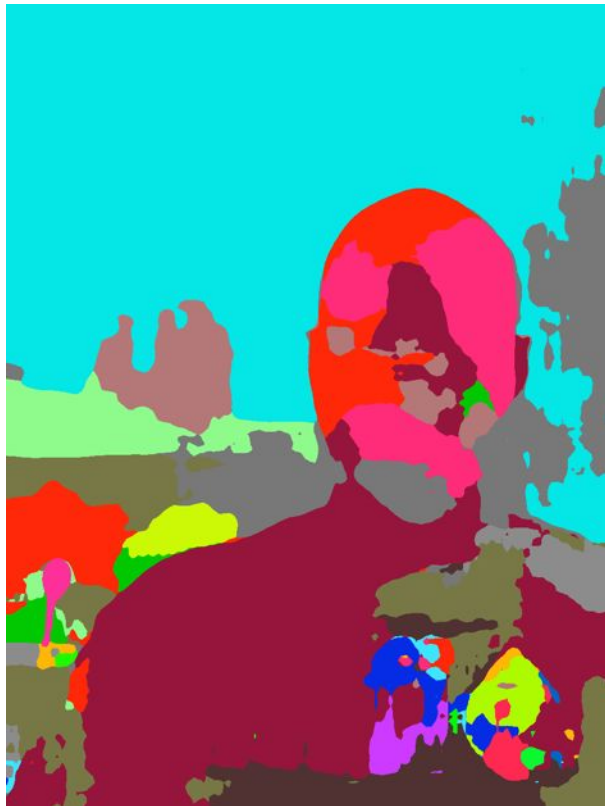
This is a beta, community release by me using the new beta Java API for Apache MXNet.

<https://community.hortonworks.com/articles/229215/apache-nifi-processor-for-apache-mxnet-ssd-single.html>

<https://www.youtube.com/watch?v=Q4dSGPvgXSA>

# MXNetm.. FLaNiFiKafka

- <https://www.slideshare.net/bunkertor/apache-deep-learning-101-apachecon-montreal-2018-v031>
- <https://www.slideshare.net/bunkertor/apache-deep-learning-202-washington-dc-dws-2019>
- <https://www.slideshare.net/bunkertor/apache-deep-learning-201-barcelona-dws-march-2019>

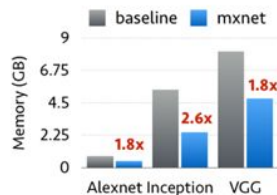




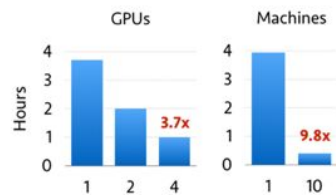
*Portable*



*Efficient*



*Scalable*



- Cloud ready
- Python, C++, Scala, R, Julia, Matlab, MXNet.js and Perl Support
- Experienced team (**XGBoost**)
- AWS, Microsoft, NVIDIA, Baidu, Intel
- Apache Incubator Project
- Run distributed on YARN and Spark
- In my early tests, faster than TensorFlow. (Try this yourself)
- Runs on Raspberry PI, NVidia Jetson Nano and other constrained devices

<https://github.com/apache/incubator-mxnet/tree/1.3.1/example>

[https://mxnet.incubator.apache.org/how\\_to/cloud.html](https://mxnet.incubator.apache.org/how_to/cloud.html)

[https://elinux.org/Jetson\\_Nano](https://elinux.org/Jetson_Nano)

[https://gluon-cv.mxnet.io/api/model\\_zoo.html](https://gluon-cv.mxnet.io/api/model_zoo.html)



- Great documentation
- Crash Course
- **Gluon (Open API), GluonCV, GluonNLP**
- **Keras (One API Many Runtime Options)**
  - Great **Python** Interaction. Java and Scala APIs!
- Open Source Model Server Available
  - **ONNX (Open Neural Network Exchange Format)** Support for AI Models
- Now in Version 1.5.1!
- Rich Model Zoo!
- Math Kernel Library and NVidia CUDA Optimizations
- TensorBoard compatible

<https://onnx.ai>   <http://mxnet.incubator.apache.org/>   <http://gluon.mxnet.io/>   <https://gluon-nlp.mxnet.io>

`pip3.7 install -U keras-mxnet`   `pip3.7 install gluonnlp`   `pip3.7 install gluoncv`

`pip3.7 install mxnet-mkl>=1.5.1 --upgrade`   `pip3.7 install --upgrade mxnet`



# Instance Segmentation: Mask RCNN with GluonCV

Mask RCNN model trained on COCO dataset with ResNet-50 backbone

```
net = model_zoo.get_model('mask_rcnn_resnet50_v1b_coco',  
pretrained=True)
```



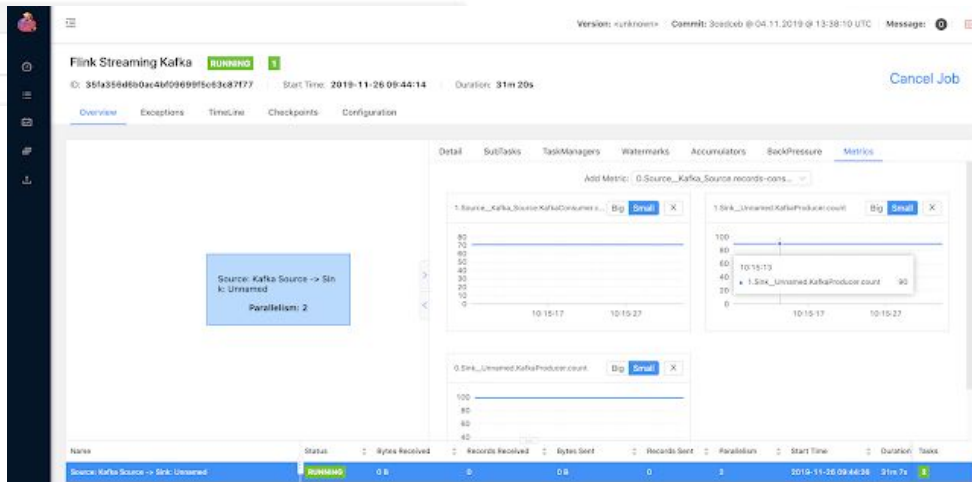
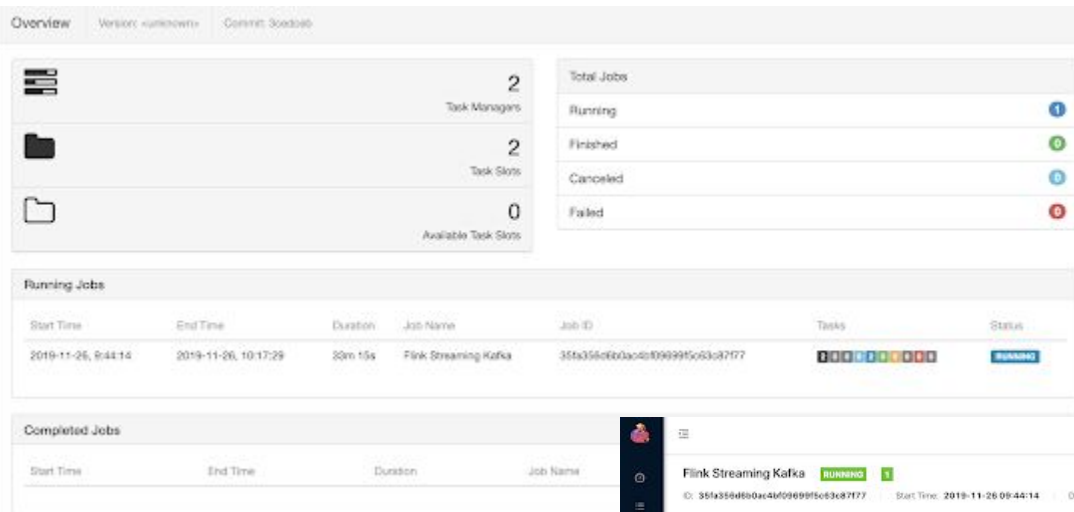
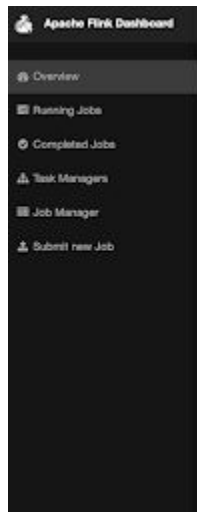
<https://arxiv.org/abs/1703.0687>

[https://github.com/matterport/Mask\\_RCNN](https://github.com/matterport/Mask_RCNN)


[https://gluon-cv.mxnet.io/build/examples\\_instance/demo\\_mask\\_rcnn.html](https://gluon-cv.mxnet.io/build/examples_instance/demo_mask_rcnn.html)



# FLINK



FLINK

<https://github.com/tspannhw/MmFLaNK/blob/master/loTKafka.java>


## RUNNING Applications

- Cluster
- About Nodes
- Node Labels
- Applications
- NEW
- SAVING
- SUBMITTED
- ACCEPTED
- RUNNING
- FINISHED
- FAILED
- KILLED
- Scheduler
- Tools

Cluster Metrics										
Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores		
7	0	1	0	3	3 GB	6 GB	0 B	3		

Cluster Nodes Metrics					
Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Final
1	0	0	0	0	0

User Metrics for drwho									
Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Containers Pending	Containers Reserved	Memory Used	Memory Pending	Memory Reserved
0	0	0	0	0	0	0 B	0 B	0 B	0 B

Scheduler Metrics			
Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Fair Scheduler	[memory-mb (unitsMB), vcores]	<memory:1024, vCores:1>	<memory:6144, vCores:32>

Show: 20 : entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime
application_1574708736290_0007	root	loT	Apache Flink	root:users.root	0	Tue Nov 26 13:24:54 -0500 2019	Tue Nov 26 13:24:55 -0500 2019

Showing 1 to 1 of 1 entries



- Cluster
- About Nodes
- Node Labels
- Applications
- NEW
- SAVING
- SUBMITTED
- ACCEPTED
- RUNNING
- FINISHED
- FAILED
- KILLED
- Scheduler
- Tools

## Application application\_1574708736290\_0006

Kill Application

User:	root
Name:	loT
Application Type:	Apache Flink
Application Tags:	flink
Application Priority:	0 (Higher integer value indicates higher priority)
YarnApplicationState:	RUNNING: AM has registered with RM and started running.
Queue:	root:users.root
FinalStatus Reported by AM:	Application has not completed yet.
Started:	Tue Nov 26 14:44:06 +0000 2019
Launched:	Tue Nov 26 14:44:07 +0000 2019
Finished:	N/A
Elapsed:	34mins, 5sec
Tracking URL:	ApplicationMaster
Log Aggregation Status:	NOT_START
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	6046040 MB-seconds, 6100 vcore-seconds
Aggregate Preempted Resource Allocation:	0 MB-seconds, 0 vcore-seconds

Show: 20 : entries

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app
appattempt_1574708736290_0006_000001	Tue Nov 26 09:44:06 -0500 2019	http://princeton01.feld.hortonworks.com:8042	Logs 0	0

Showing 1 to 1 of 1 entries



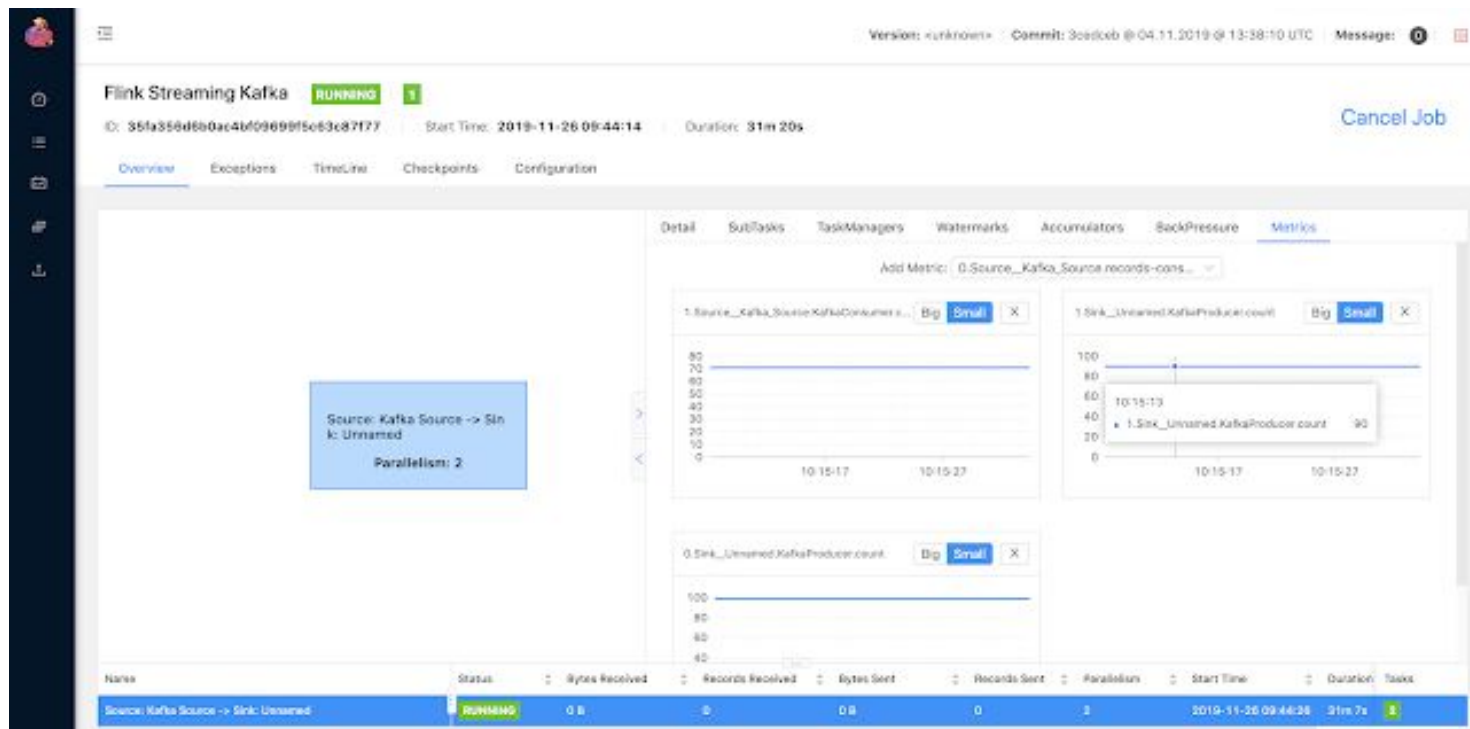
# FLINK

```
dera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/lib/hadoop/libexec/.../hadoop-yarn/.../hadoop-yarn-registry-3.8.0-cdh6.3.2.jar:/opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/lib/hadoop/libexec/.../hadoop-yarn/.../hadoop-yarn-client-3.8.0-cdh6.3.2.jar:/opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/lib/hadoop/libexec/.../hadoop-yarn/.../hadoop-yarn-server-common-3.8.0-cdh6.3.2.jar:/opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/lib/hadoop/libexec/.../hadoop-yarn/.../hadoop-yarn-server-applicationhistoryservice-3.8.0-cdh6.3.2.jar:/opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/lib/hadoop/libexec/.../hadoop-yarn/.../hadoop-yarn-server-resourcemanager-3.8.0-cdh6.3.2.jar:/opt/cloudera/parcels/FLINK-1.9.0-csa1.0.0-cdh6.3.2/bin/.../lib/flink/opt/cloudera/*jar:/etc/hadoop/conf:/etc/hbase/conf
19/11/26 18:25:02 INFO zookeeper.ZooKeeper: Client environment:java.library.path=/usr/java/packages/lib/amd64:/usr/lib64:/lib64:/lib:/usr/lib
19/11/26 18:25:02 INFO zookeeper.ZooKeeper: Client environment:java.io.tmpdir=/tmp
19/11/26 18:25:02 INFO zookeeper.ZooKeeper: Client environment:java.compiler=<NA>
19/11/26 18:25:02 INFO zookeeper.ZooKeeper: Client environment:os.name=Linux
19/11/26 18:25:02 INFO zookeeper.ZooKeeper: Client environment:os.arch=amd64
19/11/26 18:25:02 INFO zookeeper.ZooKeeper: Client environment:os.version=3.10.0-327.22.2.el7.x86_64
19/11/26 18:25:02 INFO zookeeper.ZooKeeper: Client environment:user.name=root
19/11/26 18:25:02 INFO zookeeper.ZooKeeper: Client environment:user.home=/root
19/11/26 18:25:02 INFO zookeeper.ZooKeeper: Client environment:user.dir=/opt/demo
19/11/26 18:25:02 INFO zookeeper.ZooKeeper: Initiating client connection, connectString=princeton0.field.hortonworks.com:2181 sessionTimeout=60000 watcher=org.apache.flink.shaded.curator.org.apache.curator.ConnectionState@252f626c
19/11/26 18:25:02 WARN zookeeper.ClientCnxn: SASL configuration failed: javax.security.auth.login.LoginException: No JAAS configuration section named 'Client' was found in specified JAAS configuration file: '/tmp/jaas-672186626981691890.conf'. Will continue connection to Zookeeper server without SASL authentication, if Zookeeper server allows it.
19/11/26 18:25:02 INFO zookeeper.ClientCnxn: Opening socket connection to server princeton0.field.hortonworks.com/172.26.226.24:2181
19/11/26 18:25:02 INFO zookeeper.ClientCnxn: Socket connection established to princeton0.field.hortonworks.com/172.26.226.24:2181, initiating session
19/11/26 18:25:02 ERROR curator.ConnectionState: Authentication failed
19/11/26 18:25:02 INFO zookeeper.ClientCnxn: Session establishment complete on server princeton0.field.hortonworks.com/172.26.226.24:2181, sessionId = 0x16ea3f3556d1042, negotiated timeout = 60000
19/11/26 18:25:02 INFO state.ConnectionStateManager: State change: CONNECTED
19/11/26 18:25:03 INFO rest.RestClient: Rest client endpoint started.
19/11/26 18:25:03 INFO leaderretrieval.ZooKeeperLeaderRetrievalService: Starting ZooKeeperLeaderRetrievalService /leader/rest_server_lock.
19/11/26 18:25:03 INFO leaderretrieval.ZooKeeperLeaderRetrievalService: Starting ZooKeeperLeaderRetrievalService /leader/dispatcher_lock.
19/11/26 18:25:03 INFO cli.CliFrontend: Job has been submitted with JobID ce629db2f62f18ddb15f2d5466488b99
Job has been submitted with JobID ce629db2f62f18ddb15f2d5466488b99
19/11/26 18:25:03 INFO rest.RestClient: Shutting down rest endpoint.
19/11/26 18:25:03 INFO rest.RestClient: Rest endpoint shutdown complete.
19/11/26 18:25:03 INFO leaderretrieval.ZooKeeperLeaderRetrievalService: Stopping ZooKeeperLeaderRetrievalService /leader/rest_server_lock.
19/11/26 18:25:03 INFO leaderretrieval.ZooKeeperLeaderRetrievalService: Stopping ZooKeeperLeaderRetrievalService /leader/dispatcher_lock.
19/11/26 18:25:03 INFO impls.CuratorFrameworkImpl: backgroundOperationsLoop exiting
19/11/26 18:25:03 INFO zookeeper.ZooKeeper: Session: 0x16ea3f3556d1042 closed
19/11/26 18:25:03 INFO zookeeper.ClientCnxn: EventThread shut down for session: 0x16ea3f3556d1042
```



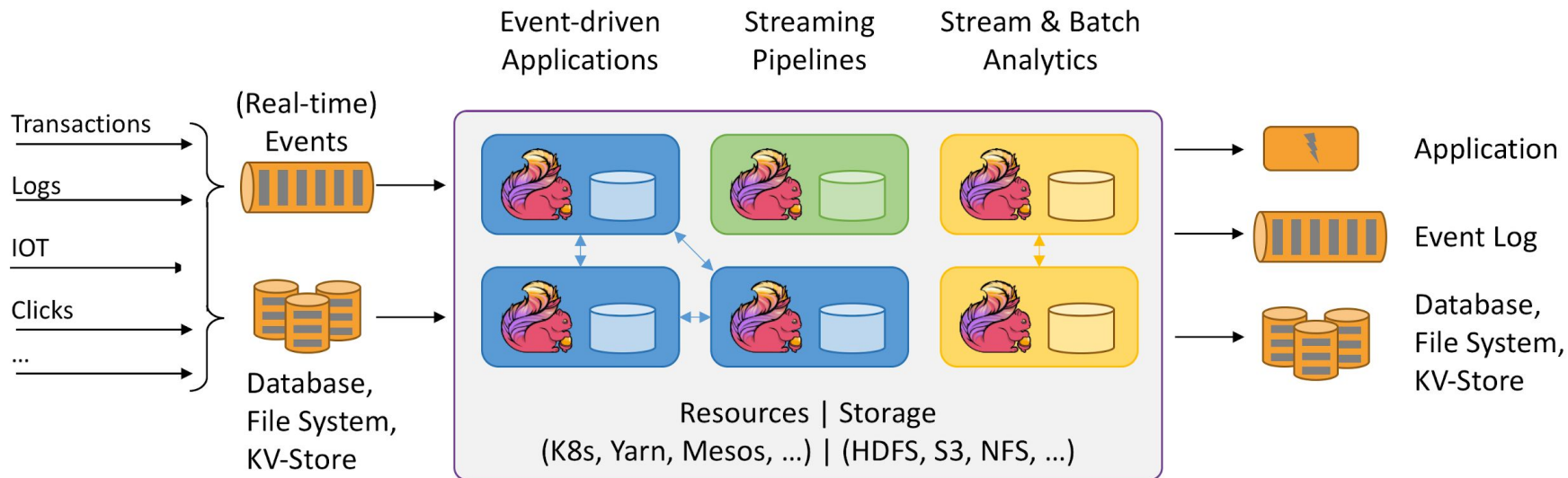
# Flink

# FLINK



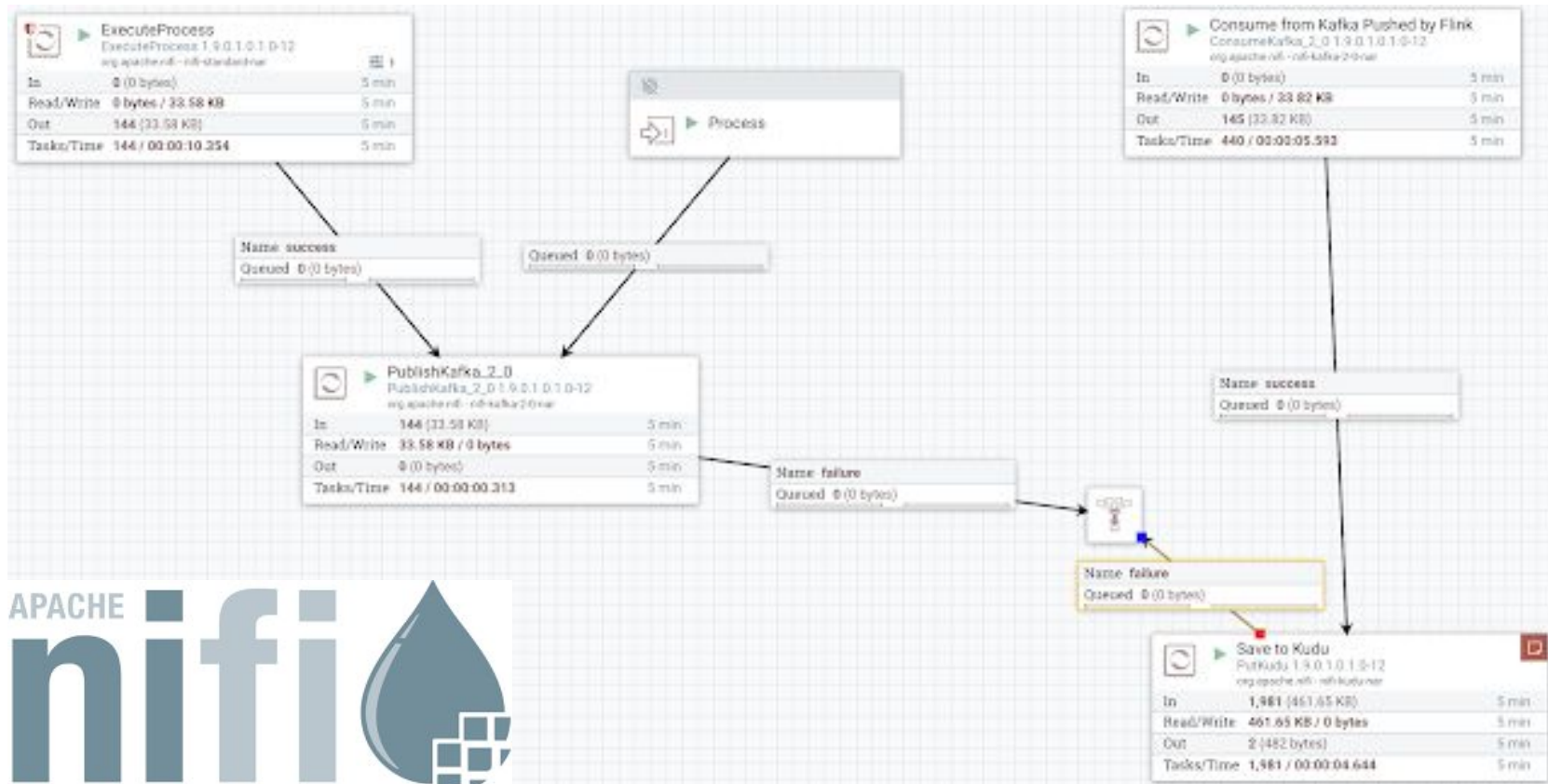
# Flink

# Flink is a Distributed Data Processing System



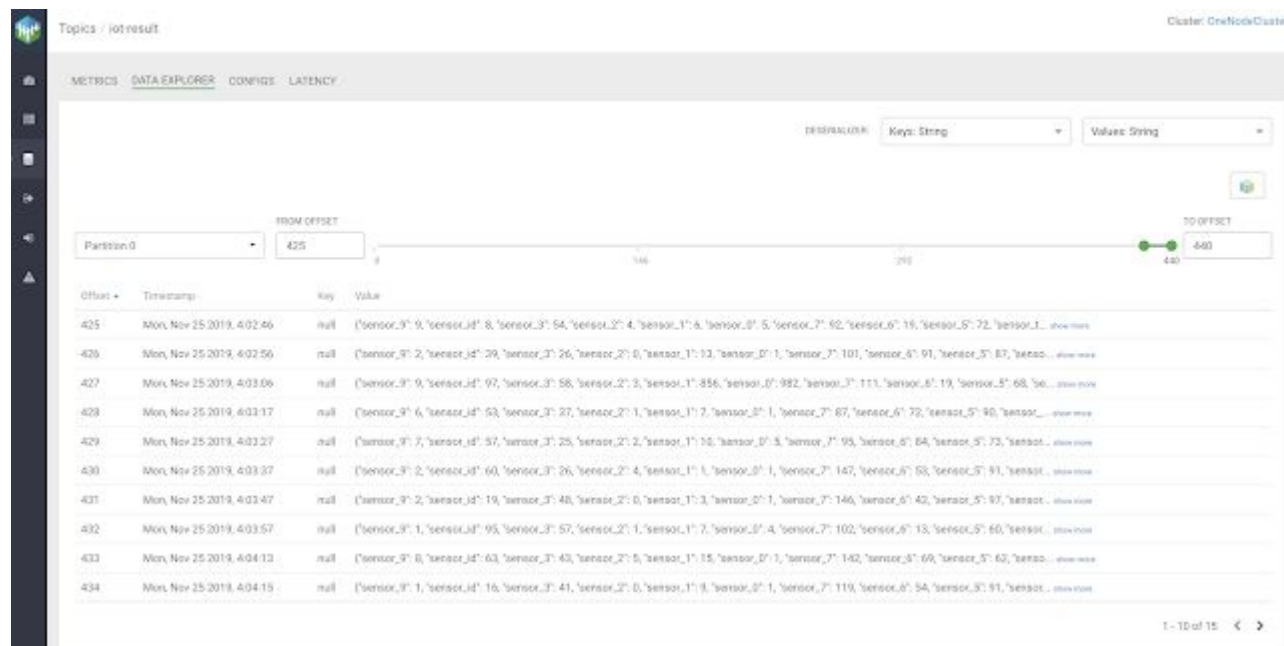


# FLaNiFiK





# FLaNiFiKafka



# FLaNiFiKafka

filename  
3cac0145-ae8e-4dfd-ad9e-ce084f39ea9a

kafka.offset  
1760

kafka.partition  
0

kafka.topic  
iot-result

path  
./

uuid  
3cac0145-ae8e-4dfd-ad9e-ce084f39ea9a

Cluster: OneNodeCluster

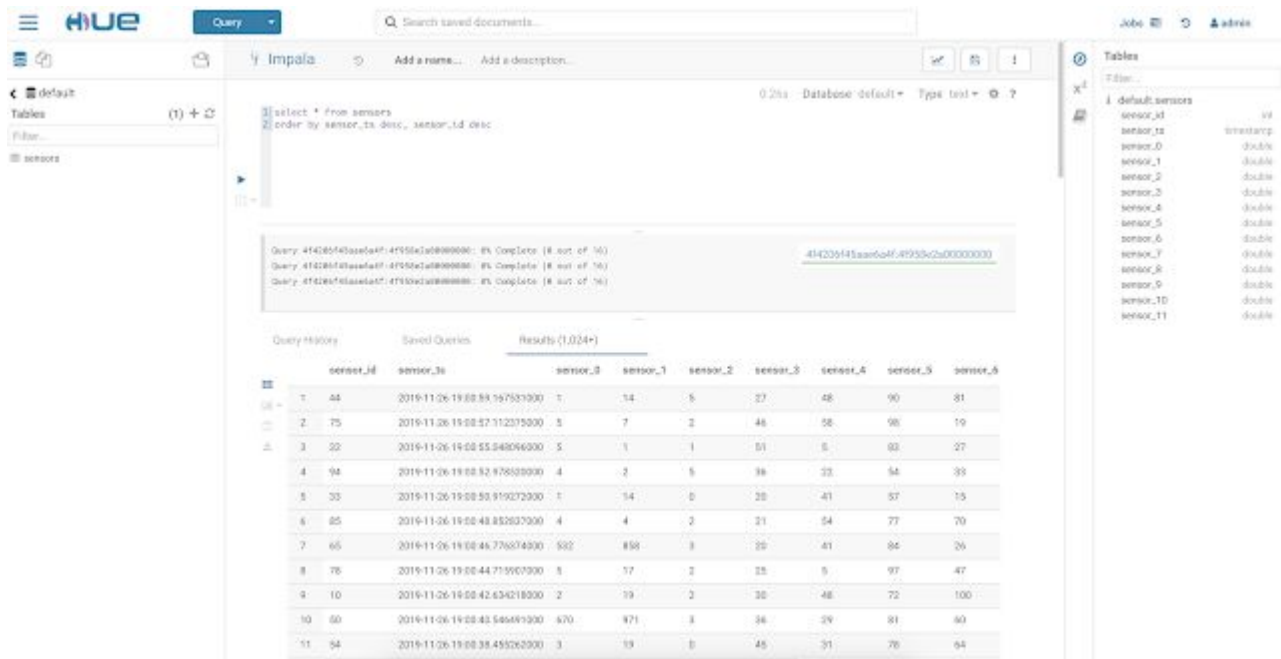
Total Bytes In	Total Bytes Out	Produced Per Sec	Fetches Per Sec	In Sync Replicas	Out Of Sync	Under Replicated	Offline Partitions
2 MB	2 MB	0	222	67	0	0	0

Topics (18)

NAME	DATA IN	DATA OUT	MESSAGES IN	CONSUMER GROUPS
✓ wiki-result	0B	0B	0	0
✓ iot-result	764 KB	292 KB	2.5k	1
✓ iot	1 MB	1 MB	3.9k	1
✓ flink-alerts	0B	0B	0	0



# FLaNiFiKudu



The screenshot displays the Hue web interface for Hive. The main panel shows a query execution window with the following SQL query:

```
1 select * from sensors
2 order by sensor_ts desc, sensor_id desc
```

The query execution status is shown as "0.21s Database default Type test ?". Below the query, the execution progress is displayed as "Query 414235f45a86a4f4f9550a2a0000000: 8% Complete (8 out of 10)".

The results table shows 11 rows of sensor data, ordered by sensor\_ts desc and sensor\_id desc. The columns are: sensor\_id, sensor\_ts, sensor\_3, sensor\_1, sensor\_2, sensor\_3, sensor\_4, sensor\_5, and sensor\_6.

	sensor_id	sensor_ts	sensor_3	sensor_1	sensor_2	sensor_3	sensor_4	sensor_5	sensor_6
1	44	2019-11-26 19:00:59.197531000	1	14	5	27	48	90	81
2	75	2019-11-26 19:00:57.312075000	5	7	2	48	58	98	19
3	32	2019-11-26 19:00:55.948066000	5	1	1	51	5	82	27
4	94	2019-11-26 19:00:52.876838000	4	2	5	38	32	54	33
5	33	2019-11-26 19:00:50.919272000	1	14	0	28	41	57	15
6	85	2019-11-26 19:00:48.852837000	4	4	2	21	54	77	70
7	65	2019-11-26 19:00:46.776874000	582	858	3	20	41	84	26
8	76	2019-11-26 19:00:44.719567000	5	17	2	25	5	97	47
9	10	2019-11-26 19:00:42.634218000	2	19	2	30	48	72	100
10	60	2019-11-26 19:00:40.546491000	670	971	3	36	29	81	60
11	54	2019-11-26 19:00:38.455262000	3	19	0	45	31	76	64

The right sidebar shows a list of tables, including "default:sensors".





THAN YOU

