



OVHcloud SUMMIT PARIS

10 OCT. 2019

W E L C O M E

Labs

Processing data at scale



Pierre Gronlier
Cloud Solution Architect
[@ticapix](#)



OVHcloud

#OVHcloudSummit

Goals

Learn how to automate infrastructure deployment using Terraform and Ansible via the OpenStack API

Show a workload usecase with Spark and Scala

Learn how to monitor your infrastructure with OVH Metrics, your business application with OVH Logs and display metrics in Grafana with OpenTSDB and Warp10

Share some tips on data partition

Having fun 😊



ANSIBLE



Scala



Setup workflow

Makefile

- Setup the tools

setup.sh

- Setup the environment

Ansible

- Setup the Software, Keys, Logging, Monitoring, ...

Terraform

- Setup the infrastructure



OVHcloud

#OVHcloudSummit

Find and count verbs

The english Wikipedia dump:

- 68 GB raw XML file
- 19 567 269 articles
- 45 927 570 tokens
- ~1% of daily tweets



OVHcloud

#OVHcloudSummit

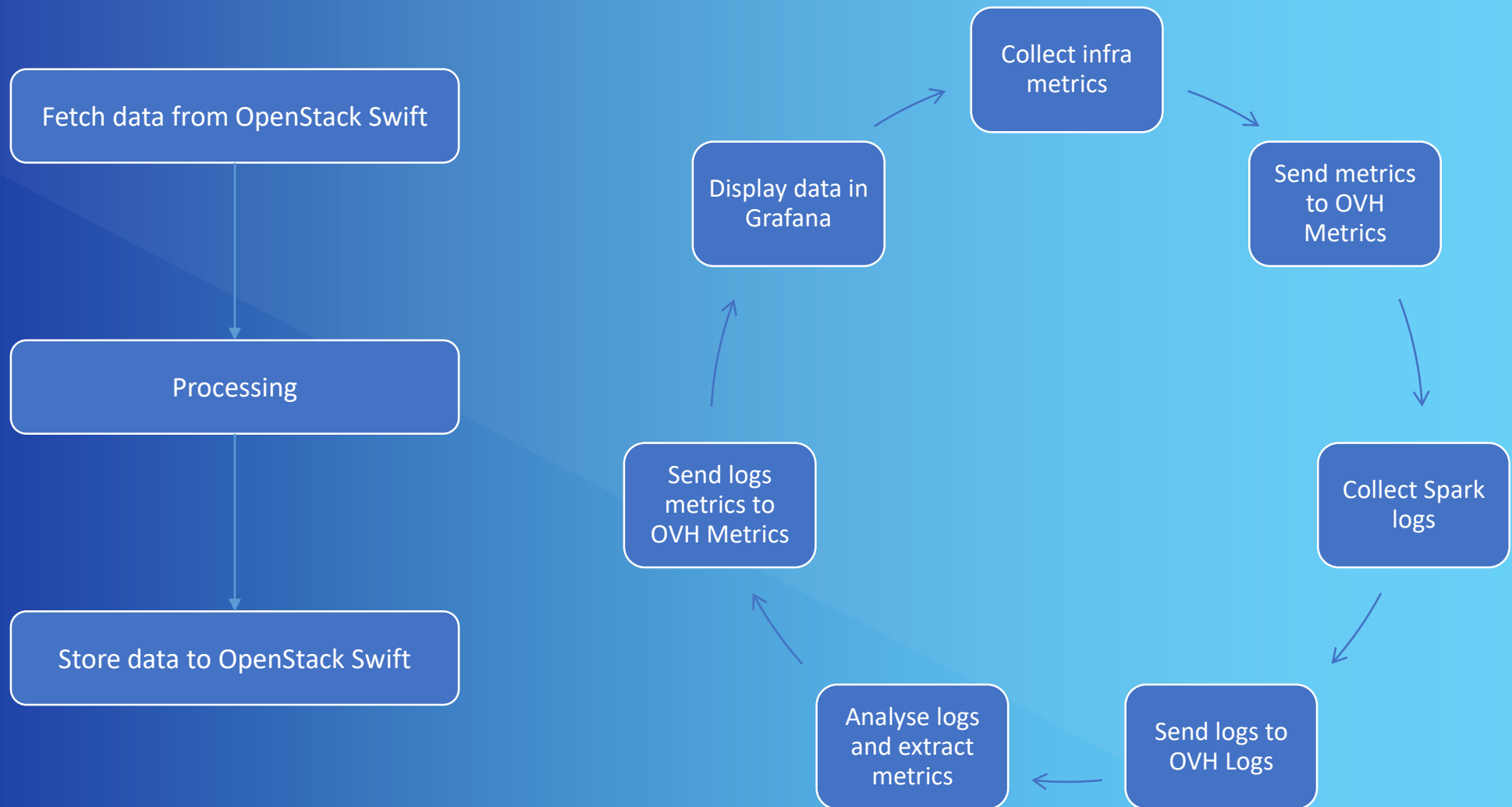
A bit of Natural Language Processing background

- Tokenisation
- Part of Speech Tagging
- Lemmatization vs Stemming



Stemming	Lemmatization
adjustable → adjust	was → (to) be
formality → formaliti	better → good
formaliti → formal	meeting → meeting
airliner → airlin △	

Runtime workflow



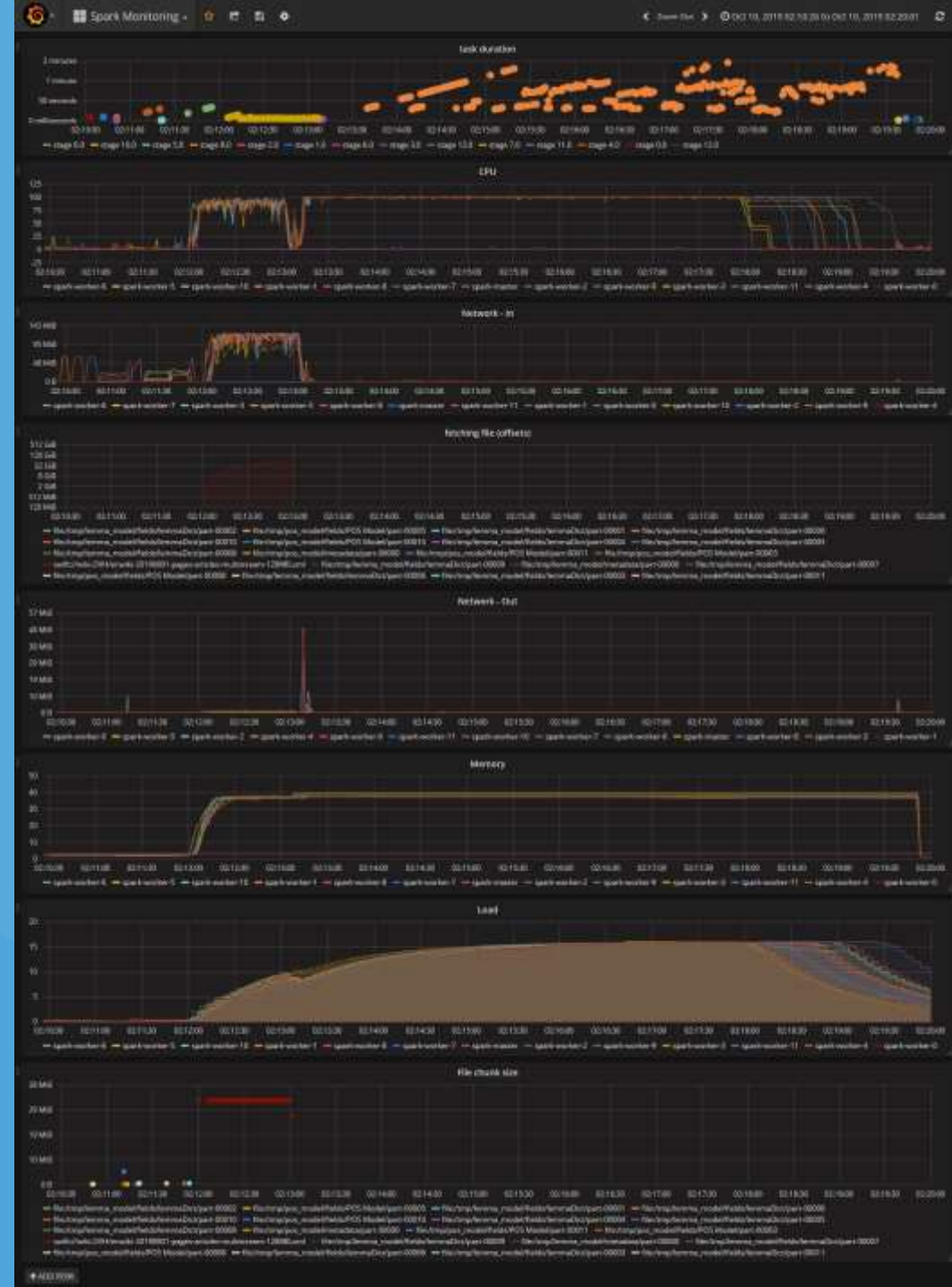
Result sample

Compute cluster of 12x c2-60:

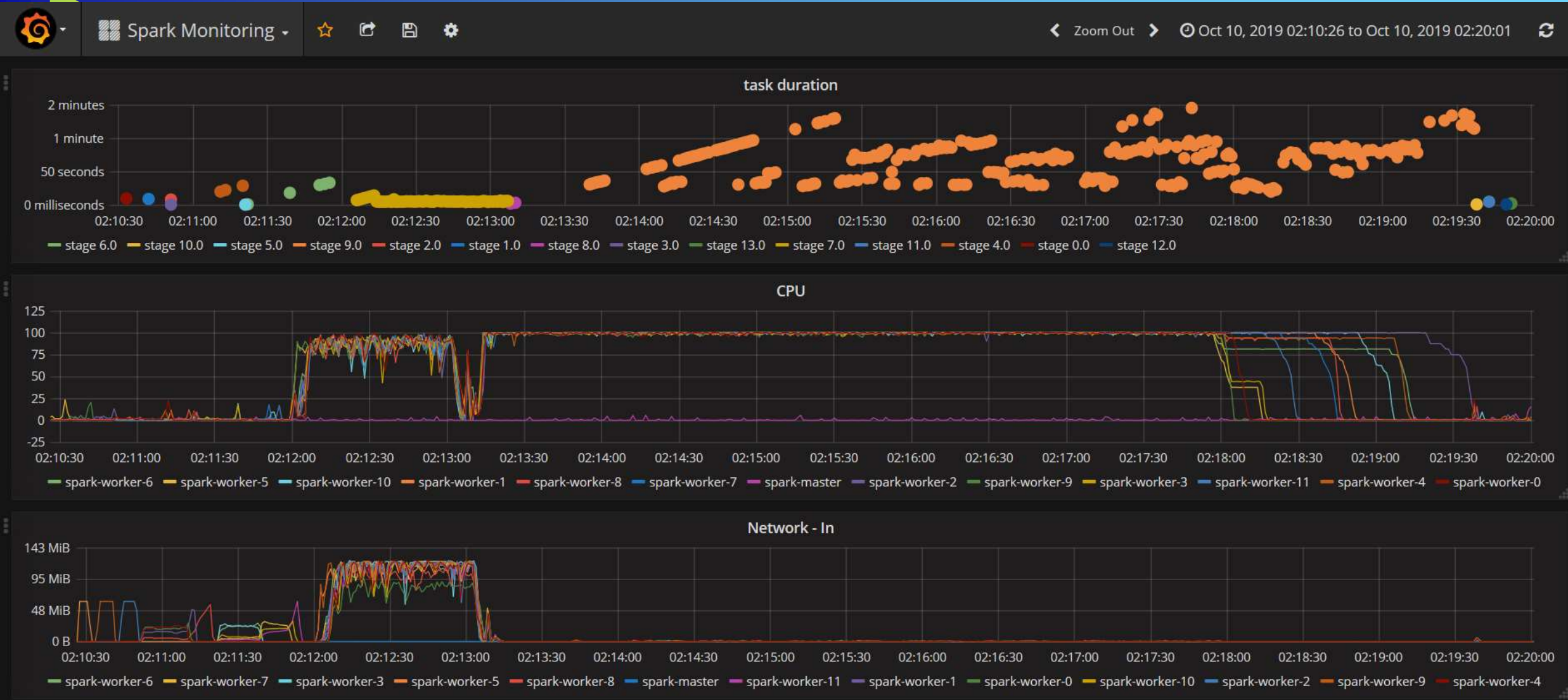
- 720 GB RAM
- 192 vCPU @3Ghz
- 12Gbps bandwidth

Storage on Openstack Swift

Total time: ~8minutes (load + compute)



Result sample





Your
TURN 😊



OVHcloud

#OVHcloudSummit



OVHcloud is an ecosystem helping you to
build your applications

And it's simple

Questions ?



OVHcloud

#OVHcloudSummit

The lab

- Smaller cluster
- Pre-created OVH account
- Pre-compiled spark-2.4.4-hadoop3.2 package
- Smaller Wikipedia dump



OVHcloud

#OVHcloudSummit

The lab

- Connect to your OVH manager
- Subscribe to a free OVH Metrics plan
- Subscribe to a free OVH Logs plan



OVHcloud

#OVHcloudSummit

METRICS_WRITE_TOKEN

The screenshot shows the OVHcloud interface for managing metrics write tokens. Red circles and arrows highlight the navigation path: from the 'Cloud' tab in the top navigation bar, to the 'Metrics' section in the left sidebar, then to the 'hcgagihblhfli' account, and finally to the 'Tokens' sub-tab. The 'Tokens' page displays a table of existing tokens. The first token is highlighted with a red circle, and its full value is shown in a tooltip. The second token is also highlighted with a red circle.

Navigation Path:

- Cloud (Top Navigation Bar)
- Metrics (Left Sidebar)
- hcgagihblhfli (Left Sidebar)
- Tokens (Sub-tab)

Tokens Table:

Name ↑	Labels	Permission	Token
-		Write	JLgzOnJNEws6e7iSMw52iJKgD_o8rvyrkx7SzMQr557pAbrLqhWG5eshGlyu...
-		Read	1VCGmY7PRhhQDraXKAqvaZtLZd7wJr3cKUISpHG_PoFJgc_Aa7y.9R9TKiEr ...

Annotations:

- Red circle around 'Cloud' in the top navigation bar.
- Red circle around 'Metrics' in the left sidebar.
- Red circle around 'hcgagihblhfli' in the left sidebar.
- Red circle around 'Tokens' in the sub-tab navigation.
- Red circle around the first token value in the table.
- Red circle around the second token value in the table.

LDP_TOKEN

The screenshot shows the OVHcloud Logs Data Platform (LDP) interface. The top navigation bar includes 'MyOVH', 'Web', 'Dedicated', 'Cloud' (highlighted), 'Telecom', and 'Sunrise'. The left sidebar shows 'Order', 'Servers', 'Metrics', 'Logs Data Platform', and 'All accounts' with 'ldp-ep-81592' selected. The main content area is titled 'ldp-ep-81592' and has a sub-menu with 'Home', 'Data stream' (highlighted), 'Dashboards', 'Data-gathering tools', 'Index', 'Alias', and 'Roles'. The 'Data stream' section explains that streams classify logs in real time and mentions 'write tokens'. A table shows 1/1 data streams used. A context menu is open for the 'spark-logs' stream, with 'Copy the write token' highlighted. A red arrow points from this option to a three-dot menu icon at the bottom right of the table.

EN Notifications 2 Need help? Hello pierre Your account: Go to the new interface

New: Discover the new Public Cloud interface

ldp-ep-81592

Home Data stream Dashboards Data-gathering tools Index Alias Roles

Data stream

With streams, you can classify your logs into precise categories in real time, as soon as they arrive on the Logs Data Platform. If you have several streams, you can easily based on their source (the application or server they originated from), and create different levels of confidentiality for different streams authorized people only. Here, the tokens are write tokens, which can be used to direct logs to the appropriate stream. Generally, having multiple streams for your logs, since you no longer need to exclude the results from another system or application.

1/1 data streams used ([increase quota](#))

+ Add data stream

Name ↑	Description	Archives	Alerts	Associated option	Indexation	Latest modification	
spark-logs	spark-logs	-	-	-	Active	10/07/2019	...

1 - 1 of 1 results

Context menu options: Edit, Graylog access, Copy the write token, Monitor in real time, Manage alerts, Archives, Delete


The lab

- Create an Openstack user, save the password and download the openrc.sh file
- Pick a region SBG5 or WAW1 and edit your openrc.sh file
- Create a key: `openstack keypair create mykey > mykey.pem`
- Create a bastion: `openstack server create --key-name mykey --image 'Debian 9' --flavor b2-15 --network Ext-net bastion`
- Get bastion ip: `openstack server show bastion`
- Connect to bastion: `ssh -i mykey.pem debian@<ip>`
- Follow steps on <https://github.com/ticapix/ovh-demo-spark/>


Grafana 1/2





#OVHcloudSummit

 Data Sources ▾


Edit data source

Name	SparkMonitoring OpenTSDB 	Default	<input type="checkbox"/>
Type	OpenTSDB ▾		

HTTP settings

URL	https://opentsdb.gra1.metrics.ovh.net 
Access	direct ▾ 

HTTP Auth

Basic Auth	<input checked="" type="checkbox"/> With Credentials 	<input type="checkbox"/>
------------	--	--------------------------

Basic Auth Details

User	token
Password

OpenTSDB settings

Version	<=2.1 ▾
Resolution	second ▾


Save & TestDeleteCancel

[Docs](#) | [Support Plans](#) | [Community](#) | Grafana v4.6.3 (commit: Baadf1af8)




Grafana 2/2






#OVHcloudSummit

 Data Sources

Edit data source


Name	SparkMonitoring Warp10 	Default 
Type	Warp10 	

HTTP Address


URL	https://warp10.gra1.metrics.ovh.net 
Access	direct  

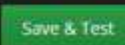


Execution constants

This constants can be used in every templating or query.
Register your constant name and value and prefix your constant name with "\$" in your queries/templating.
example: `$token`

 token > d9YM2p880pRZCoeFi_1nb.M7IXnIOjQYkNLZ7ZM020eO3CSbww4

Add a new constant

Name	<input type="text"/> 
Value	<div><div></div></div>

[Docs](#) | [Support Page](#) | [Community](#) | Grafana v4.8.3 (commit: 8a8f1af)