

TP3

Twitter Analytics with Hadoop

What's the difference between TP2 and TP3 ?

- With the TP2, you work with a data Warehouse. Data come from a cloudant noSQL (couchDB) DB. The data process was done with R program. Several matrix of solving are needed to produce an output visualization.
- With TP3, you don't use a DB. The Twitter stream is captured into a file on a HDFS. Further, functions from BigInsights are applied on the file to extract data to work with. After, final data are analyzed with hadoop to produce a dataViz.

Goal : Try to find trends into tweets.

Solution : Make an application with node JS, node RED, IBM HDFS, IBM insights

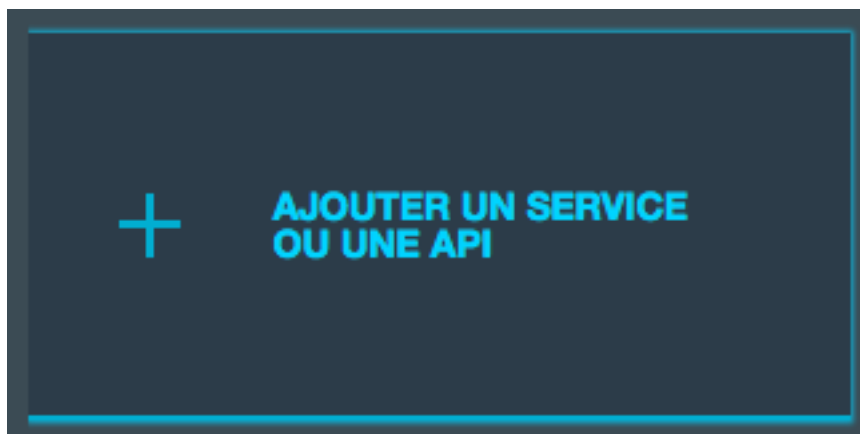
You need :

- 10 minutes to build an application to produce dataViz solution ! with BigInsights
- Non stop twitter public stream stored into a file an a HDFS.
- No infrastructure to deploy !
- No code to produce !
- That's bluemix solution ! Let's start !

1. Create a nodeRED starter application

Search in boilerplates !

2. Add a service



Add IBM Analytics Hadoop



3. Add twitter service

Go to nodeRED flow editor, and add social function « twitter »

Edit twitter in node

Twitter ID: @sgagneur

Search: all public tweets

for: comma-separated words, @ids, #tags

Name: Twitter public stream

Tip: Use commas without spaces between multiple search terms. Comma = OR, Space = AND.
The Twitter API WILL NOT deliver 100% of all tweets.
Tweets of who you follow will include their retweets and favourites.

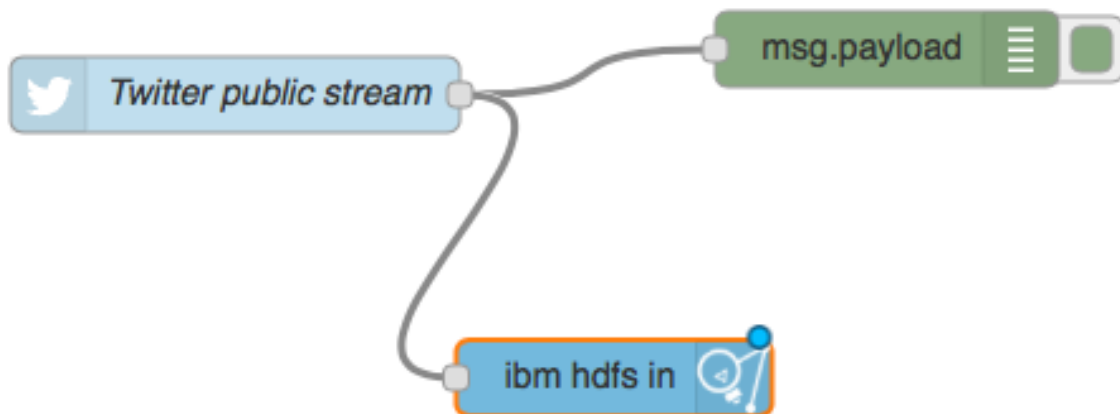
Ok Cancel

« for » textbox can store comma-separated words, @id, #tags, so you can write words, counts, or hashtags that you want retrieve into tweets. It's just like a first filter.

You can connect debug function to see the twitter public stream

Add, IBM hdfs from storage section :

Now, your flow editor must look like :



The IBM HDFS component must have following properties :

The screenshot shows a dialog box titled 'Edit ibm hdfs node'. It contains the following fields and options:

- Filename:** A text input field containing '/twitterPublicStream/stream'.
- Append newline ?** A checked checkbox.
- Overwrite complete file ?** An unchecked checkbox.
- Name:** A text input field containing 'Name'.

Below these fields is a yellow warning box with the text: **Bounded Service:** Provide a valid filename with filepath. Check the info tab, to get more information about each of the fields.

At the bottom right are 'Ok' and 'Cancel' buttons.

Be careful : the filename must start with / !!!!! it's VERY IMPORTANT.
The « stream » file will be stored in the following folder : /usr/biblumix/twitter Public/.

4. Use Hadoop

Go to your app.

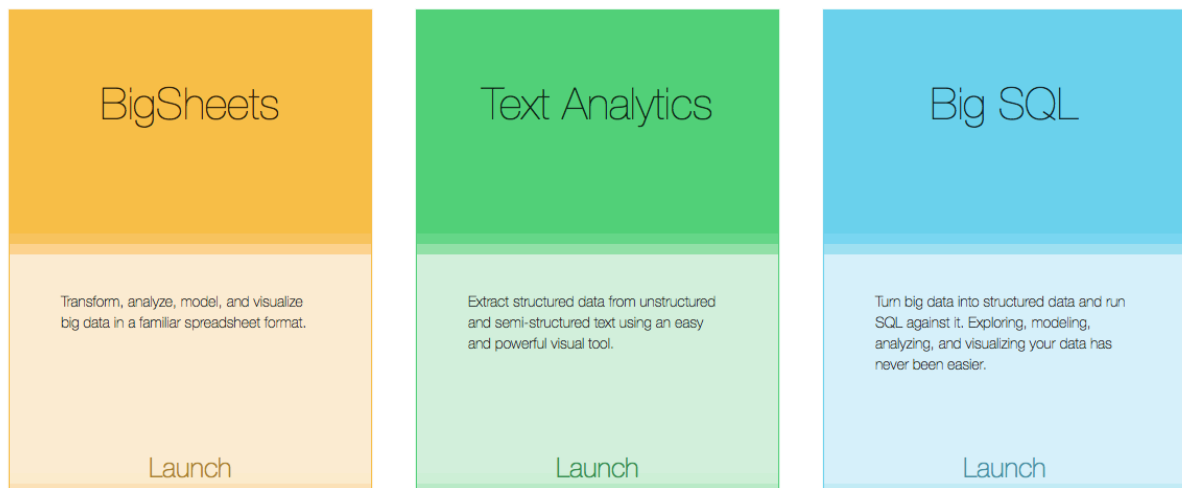
Press Launch Button (select Analytics for Apache Hadoop service)



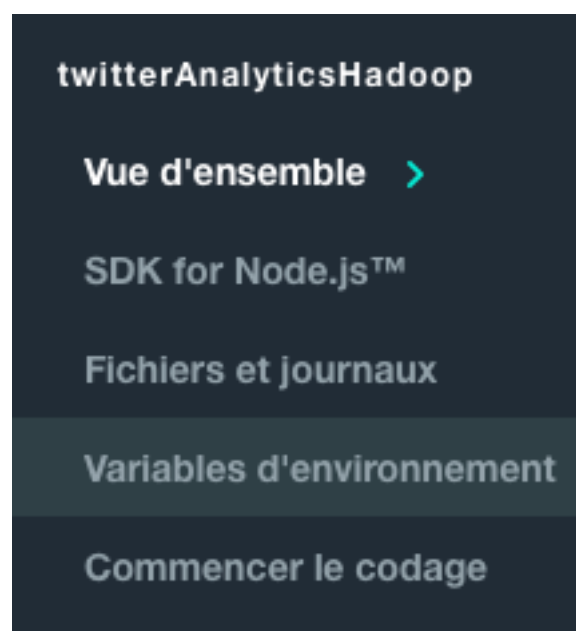
After launch, BigInsights Start page must appear !

If start page can't be open, you must solve the problem.

5. How to open BigInsights Start page ?



If a problem exists go to VCAP variables and read available URL for BigInsights :



Available URL to go to BigInsights :

"BigInsightsHomeUrl": <https://bi-hadoop-prod-2081.services.dal.bluemix.net:8443/gateway/default/BigInsightsWeb/index.html>

Put ID and Password, you can read them in VCAP values :

"userid": "biblumix",
"password": "balbalfafe",

Or in the start page on the service.

Extend License

Your license will expire on **2015-11-18 UTC**. Extend your license by 14 day(s).

EXTEND

Credentials

Username: **biblumix**

Password: **u8S0xLgnc**

6. Manage HDFS

Open BigInsights start page :

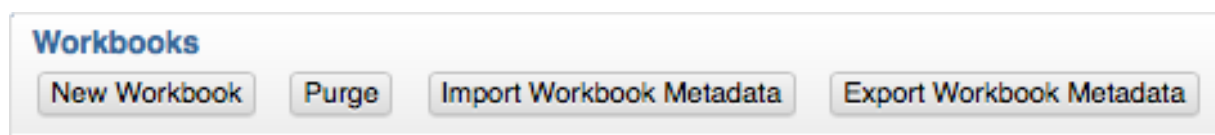
Launch BigSheets :

BigSheets

Transform, analyze, model, and visualize
big data in a familiar spreadsheet format.

Launch

create a new workbook :
















Select your file in HDFS :

New Workbook

Name:

Description:



DFS Files **Catalog Tables**

- ▼  hdfs://mn01.services.dal.bluemix.net:8020/
 - ▶  app-logs
 - ▶  apps
 - ▶  biginsights
 - ▶  iop
 - ▶  mapred
 - ▶  mr-history
 - ▶  tmp
 - ▼  user
 - ▶  ambari-qa
 - ▼  biblumix
 - ▼  twitterPublicStream
 -  stream

Now, you can see data from twitter :

✓ Ready	
	Header
1	trivial007: RT JP_Courtois: Find out how #heal
2	RT @SteerMark More blue red green dots beh
3	RT @ultrarecords: .@Hardwell shares an @es
4	Episerver Digital Experience Cloud product co
5	What do LLLT's mean for the future of the lega
6	[post] Moving Test Environments to the #Cloud
7	The biggest cloud security challenges of 2016
8	RT @LibreActu: 6 links that will show you wha
9	Particle Photon (or Core) on SAP HANA Cloud
10	RT @RHarbridge: #Microsoft has the best Pub

Save the workbook with a name :


TwitterPublicStream
 | Delete

No description


Owner: biblumix Created: 09/11/15 19:16 Last visited: 09/11/15 19:16




7. Analyze

We will use a Watson/NLP (Natural Language Processing). It's a leverage for extracting company names out of the tweets.

Open your workbook :

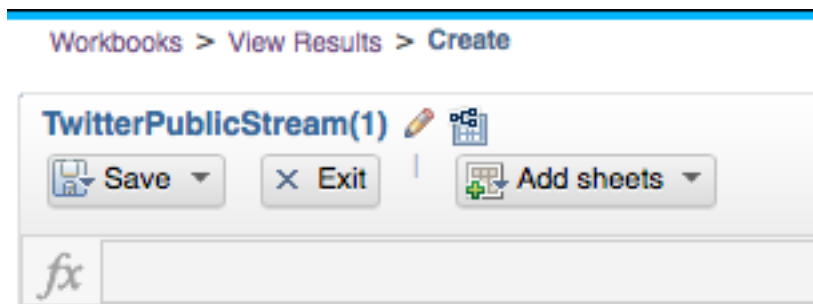
Workbooks > View Results

TwitterPublicStream 

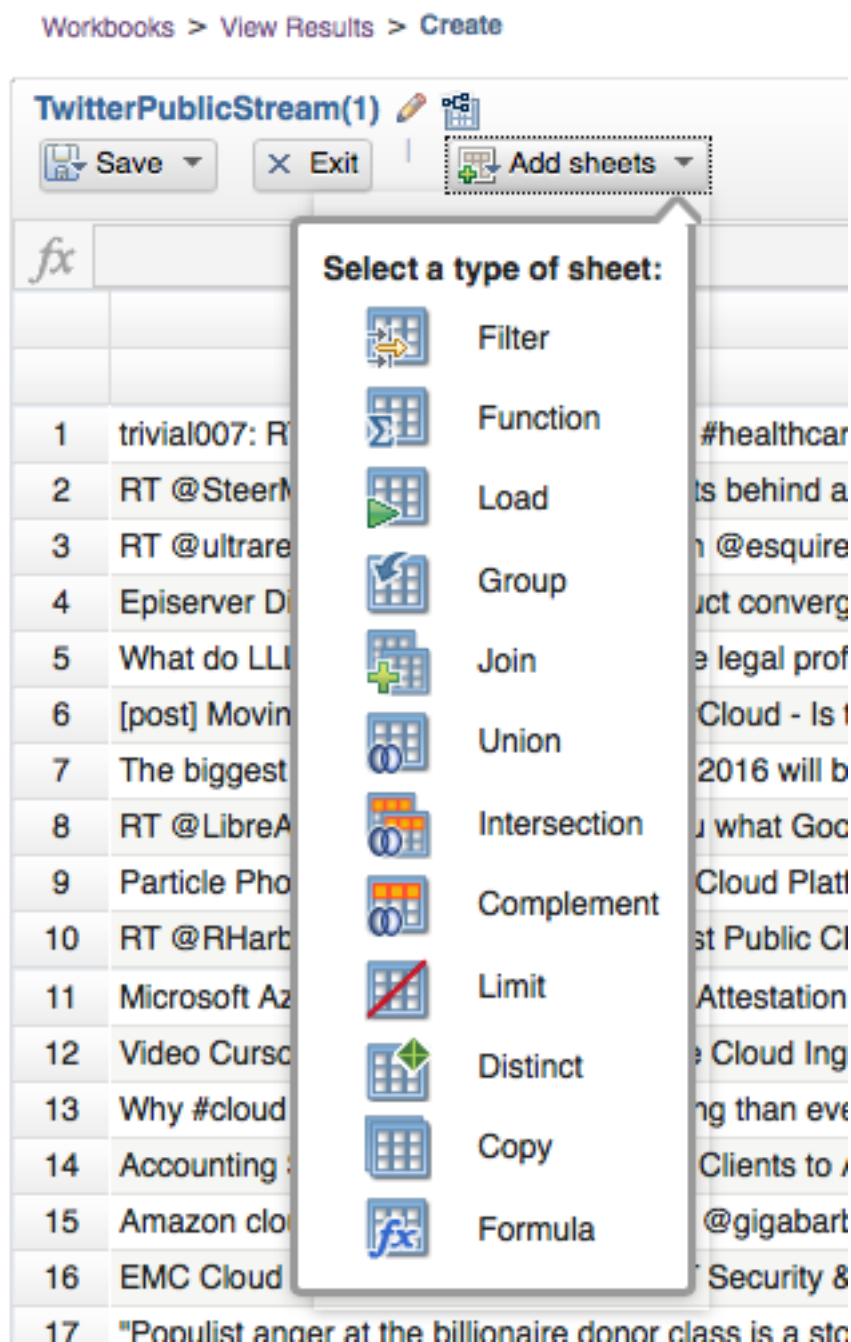
 Delete
 |
  Add chart
 |
 TwitterPublicStr... :
  Build new workbook

✓ Ready

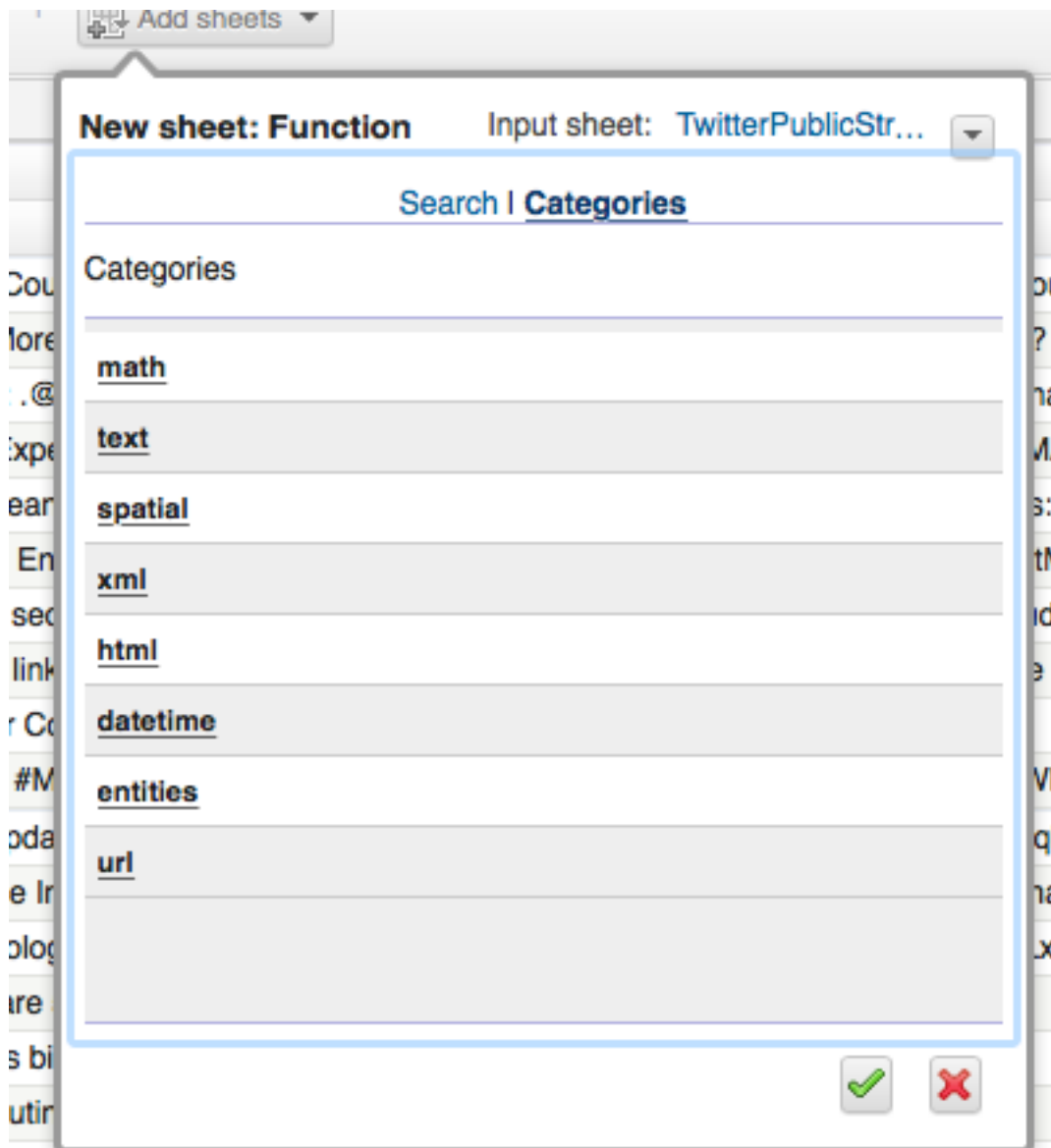
Buid a new workBook :



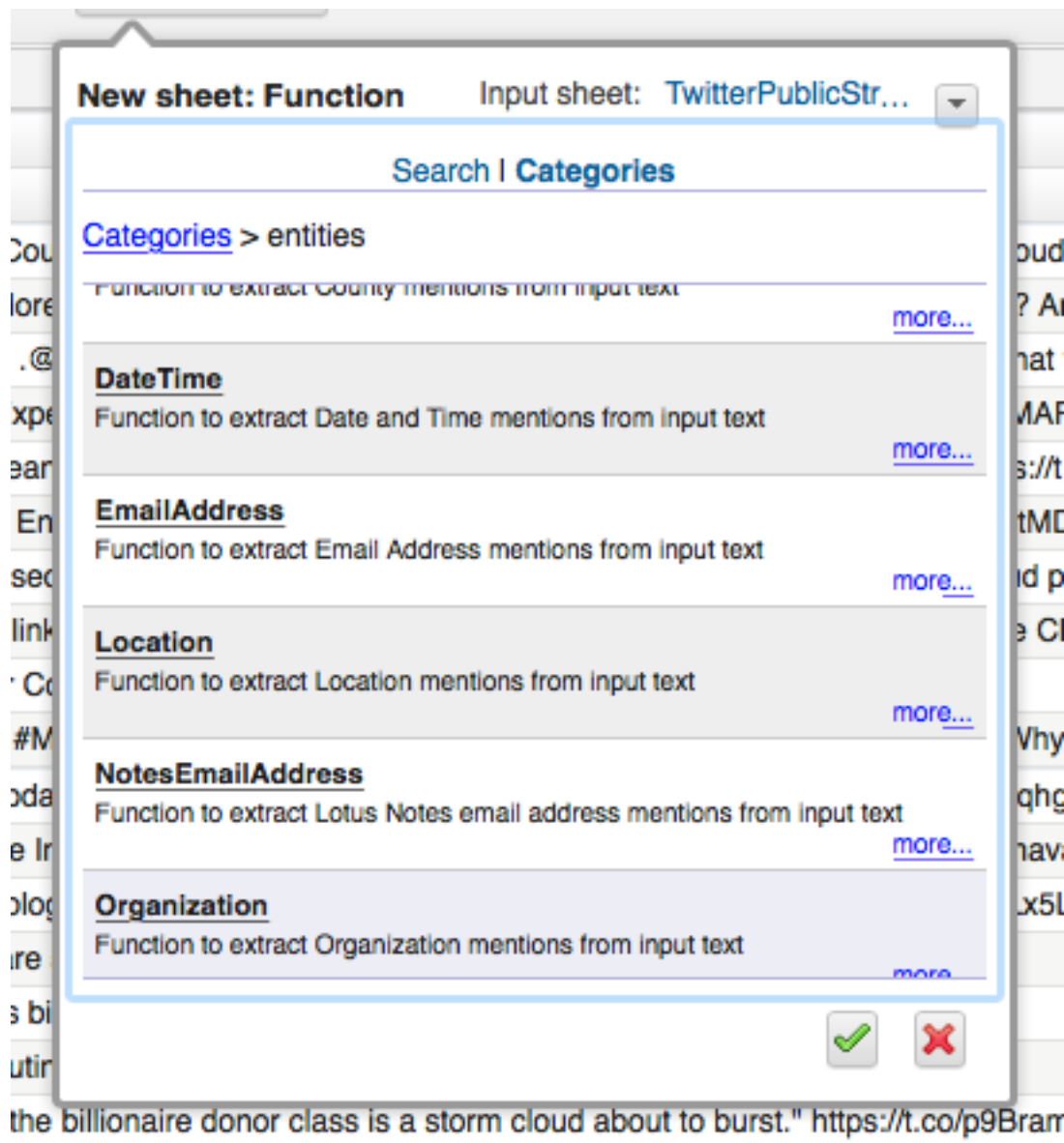
Add new sheets :



Choose Functions :



Choose entities :




Choose Organization :

New sheet: Function Input sheet: **TwitterPublicStr...** ▼

* Sheet Name:



Sheet1

Organization 

Function to extract Organization mentions from input text

Fill in parameters:
text*

Parameters Carry over (0)


 

You can name the Sheet :

New sheet: Function Input sheet: [TwitterPublicStr...](#)

* Sheet Name:

Organization



Organization 

Function to extract Organization mentions from input text

Fill in parameters:
text*

Header

Parameters Carry over (0)

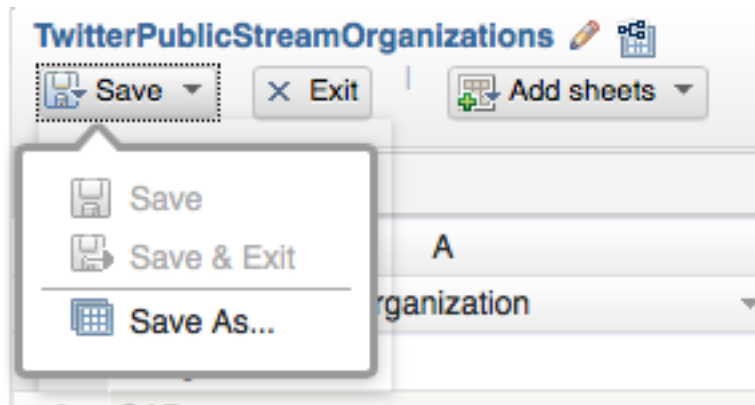
 

select Header as "Fill in parameter" and accept by clicking on the green hook

And voilà ! An nice job from Watson !

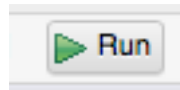
<i>fx</i>	
	A
	organization ▼
1	Google
2	SAP
3	Microsoft
4	Microsoft
5	Microsoft
6	Adobe
7	EMC
8	Adobe
9	Microsoft
10	Microsoft
11	Verizon
12	Apple
13	Accenture
14	Trans-Pacific Partnership
15	Juniper Networks
16	Verizon
17	SYNNEX
18	Microsoft
19	Akamai Technologies
20	SPSS
21	Oracle
22	Oracle

Click on Save&Exit, or Save and Exist

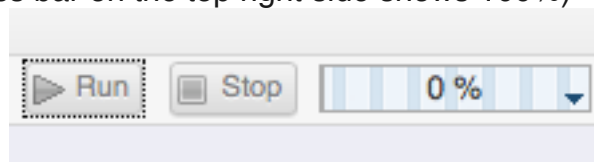


Save & exit will cause a MapReduce job. It will start to analyze all collected tweets on HDFS.

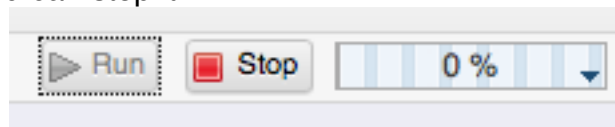
When you press exit button only, you must click run button for updating data and start MapReduce job.



(wait until the progress bar on the top right side shows 100%)

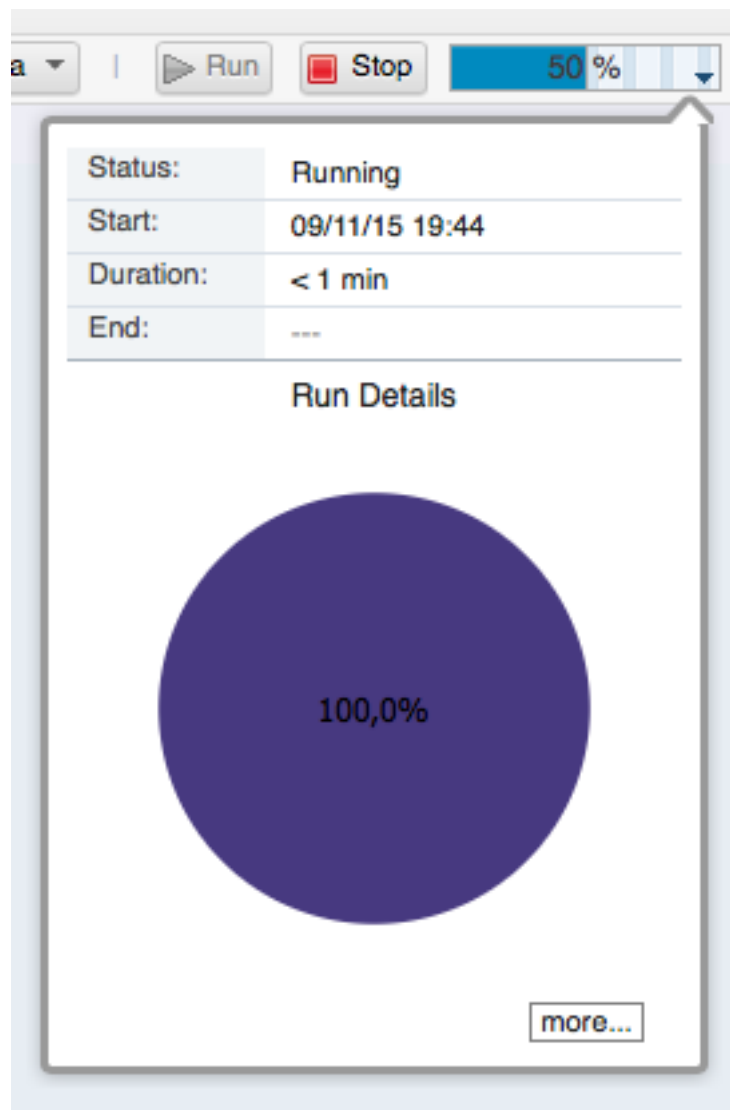


After starting run, you can stop it :

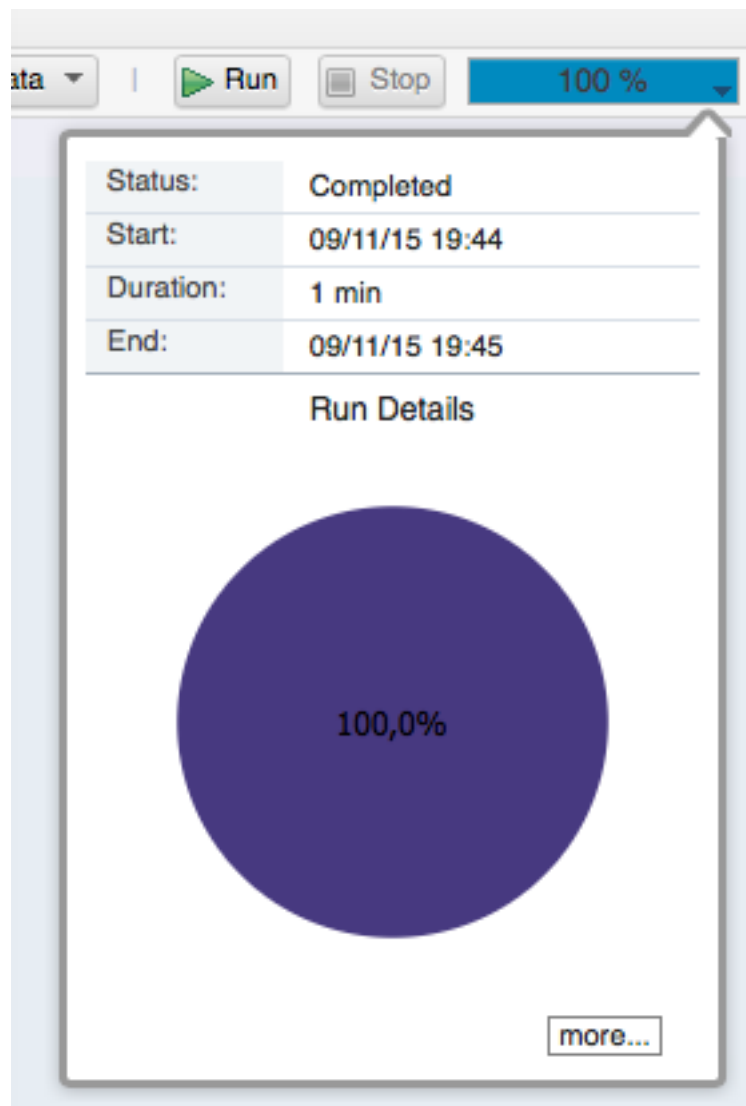


or wait for the end of the job :

You can check job during progress :



Job is ended :

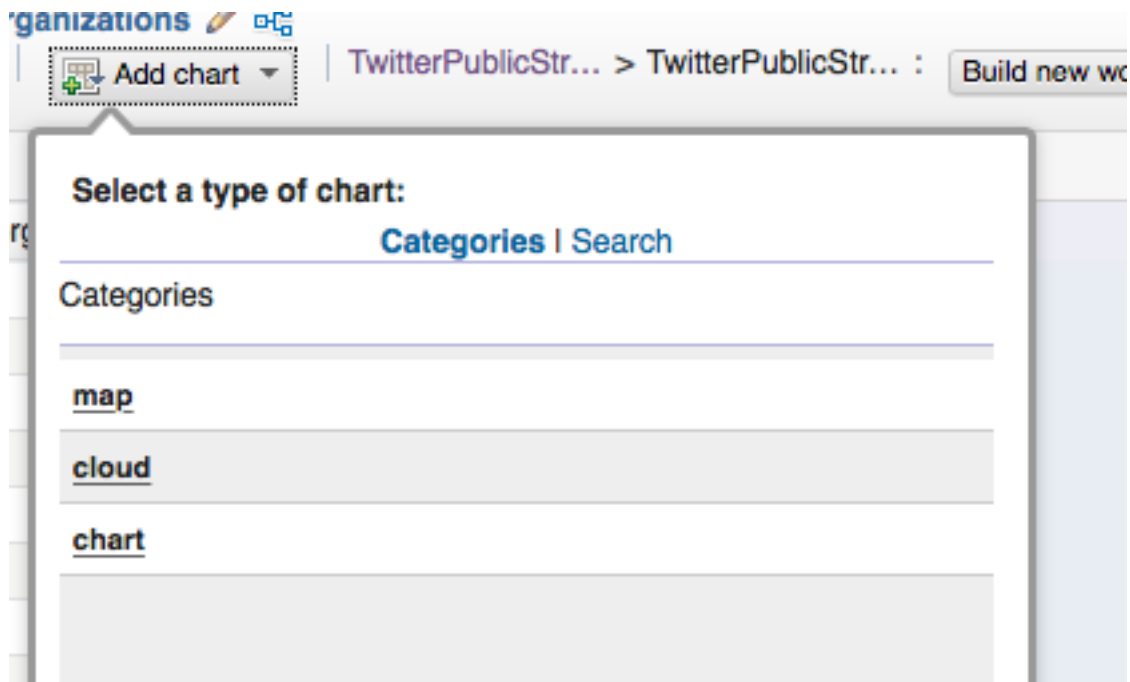


you can run it once again !

8. DataViz

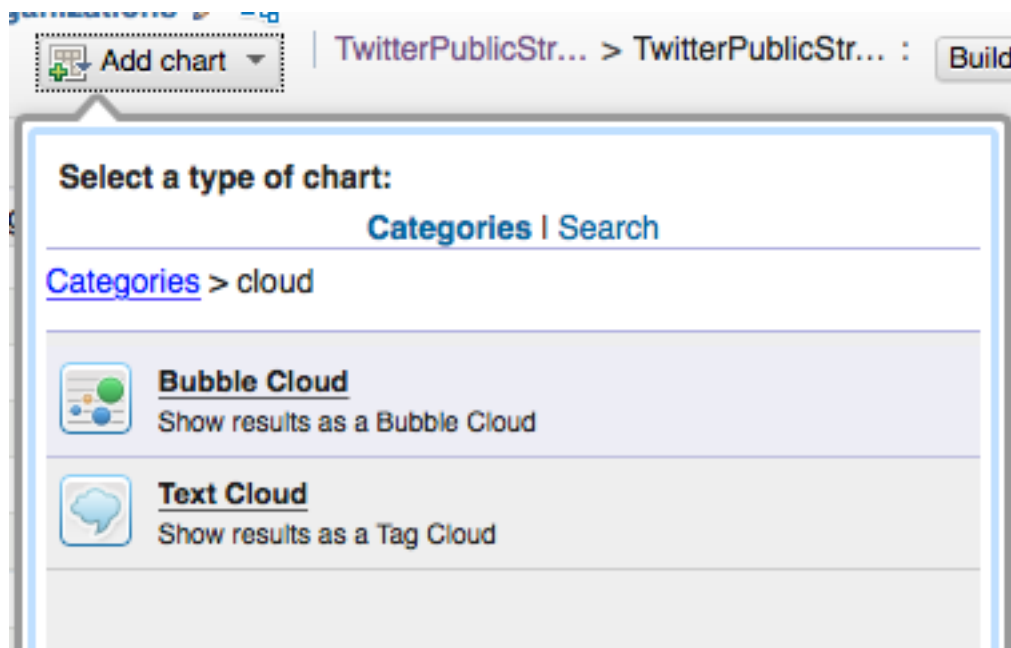
You can now visualize your data from mapReduce.

Select add chart :



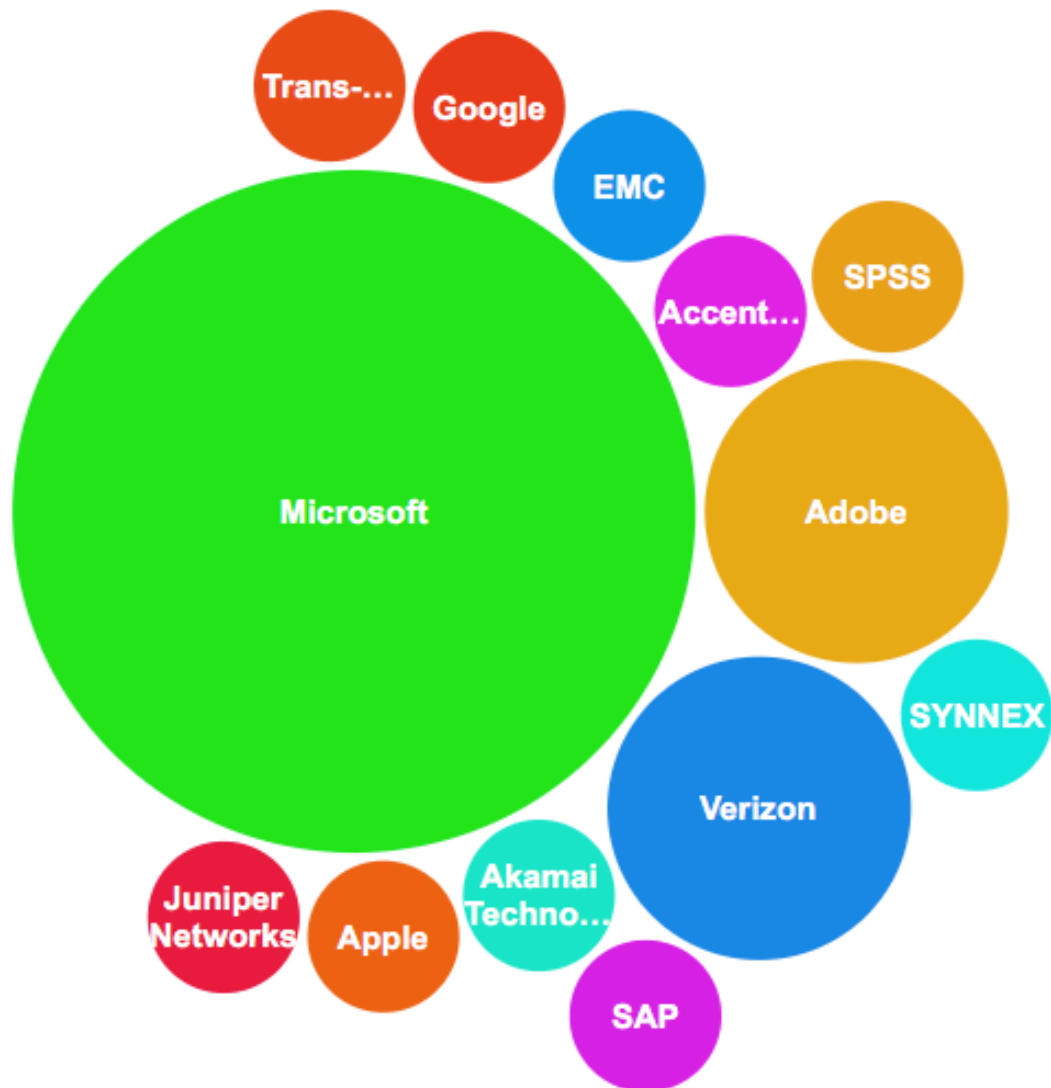
add cloud chart to see nice bubbles !

Select Bubble cloud chart to finish work :

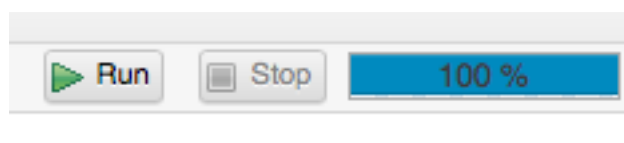


And you visualize a nice bubble chart due to Hadoop and mapReduce :

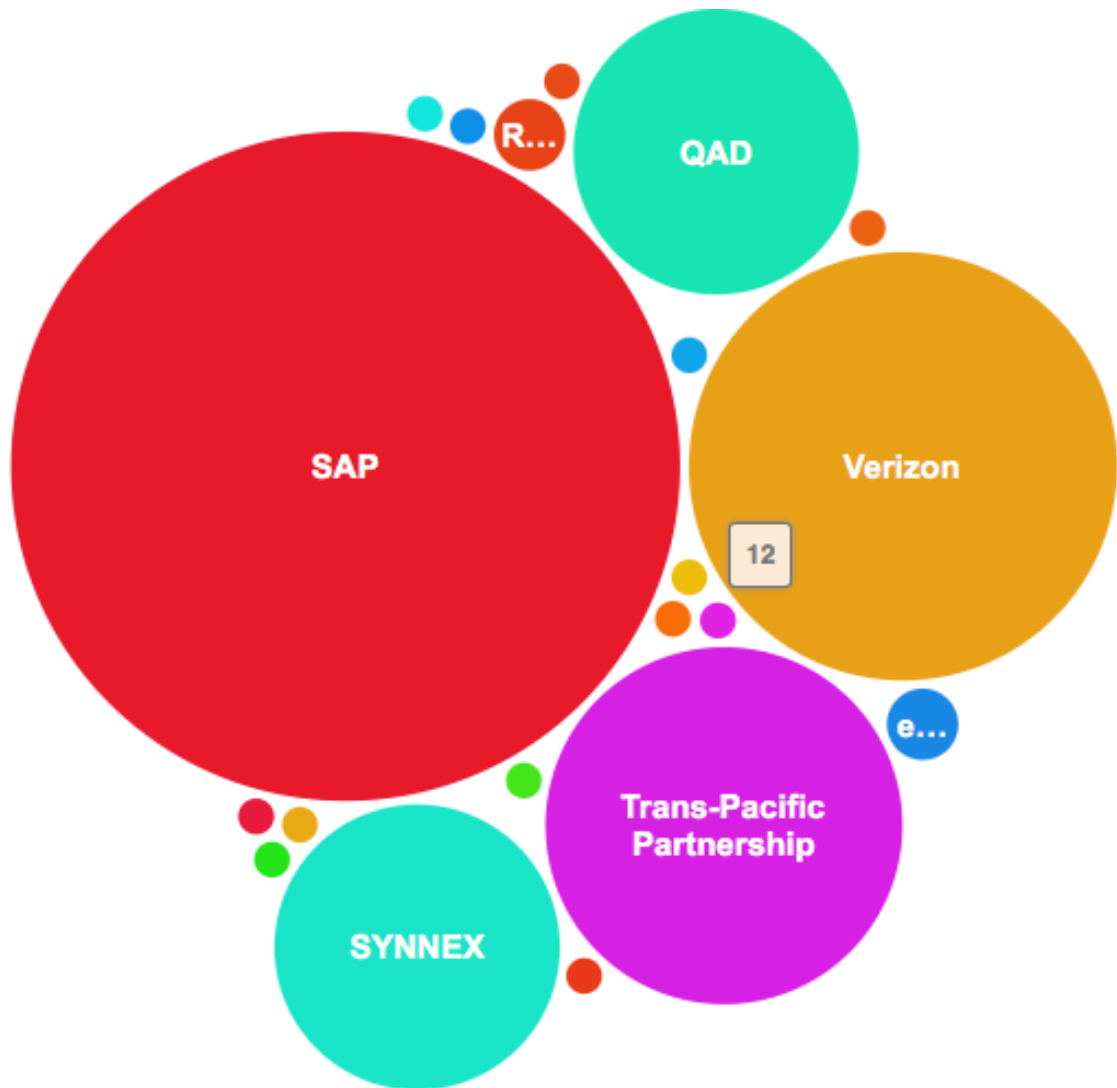
Before the process, the bubble chart looks like to :



process !



After the process :



9. New exercise

Try to produce a text cloud as new work :

The result can be the following :

