

Talend Big Data Sandbox

Big Data Insights Cookbook



Table of Contents

Table of Contents	2
1 Overview	4
1.1 Setup Talend Big Data Sandbox	4
1.1.1 Pre-requisites to Running Sandbox.....	5
1.1.2 Setup and Configuration of Sandbox	5
2 Talend License and Services Status.....	6
2.1 Talend License Setup.....	6
2.2 Hortonworks Services Status	8
3 Scenario: Clickstream Insights	11
3.1 Overview	11
3.2 Clickstream Dataset	11
3.3 Using Talend Studio	13
3.3.1 Talend HDFS Puts	13
3.3.2 Talend MapReduce Review.....	14
3.3.3 Talend to Google Charts and Hive	18
4 Scenario: Twitter Sentiment Insights.....	21
4.1 Twitter Sentiment Analysis Overview	21
4.2 Twitter Data	21
4.3 Talend Processes.....	22
4.3.1 Retrieve the Data	22
4.3.2 Process and Aggregate Results	23
4.3.3 Analysis and Sentiment.....	24
5 Scenario: Apache Weblog Insights.....	25
5.1 Apache Weblog Overview.....	25
5.2 Apache Weblog Data	25
5.3 Talend Processing	26
5.3.1 Talend Filter and Load Data	26
5.3.2 Talend PIG Scripts to Process.....	28
5.3.3 Talend MapReduce to Process.....	29
6 Scenario: ETL Off-loading.....	30

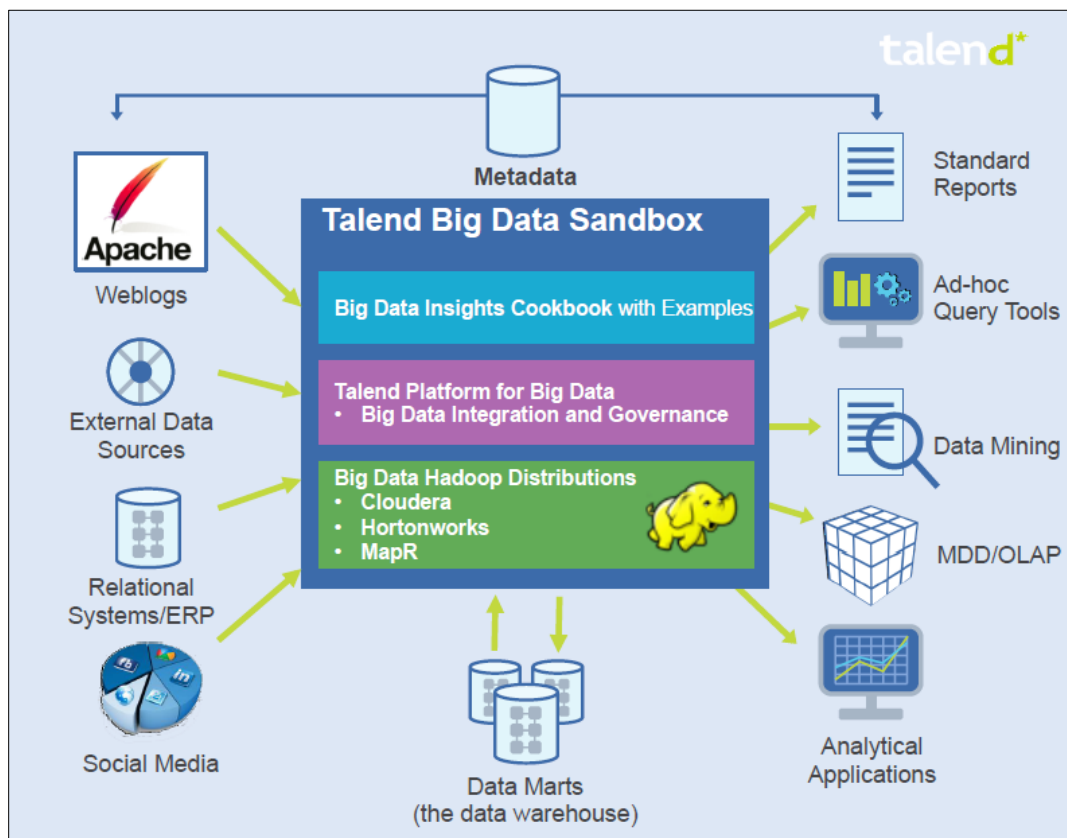
6.1	Overview	30
6.2	Data	31
6.3	Talend Process	32
6.3.1	Single-Click Execution	32
6.3.2	Step-by-Step Execution	33
6.3.3	Extended Scenario Functionality.....	38
7	Demo: NoSQL Databases	40
7.1	Hadoop Core – Hive and HBase	40
7.1.1	Hive ELT	40
7.1.2	HBase	42
7.2	Cassandra	44
7.3	MongoDB	45
8	Conclusion.....	49
9	Next Steps	49

1 Overview

The purpose of this document and associated projects is to guide you through a set of big data scenarios using the Talend Big Data Sandbox. At the end of these projects, you will have a better understanding of how Talend can be used to address your big data challenges and move you into and beyond the sandbox stage.

1.1 Setup Talend Big Data Sandbox

The Talend Big Data Sandbox is delivered as a Virtual Machine (VM). The VM includes an Apache Hadoop distribution provided by a partner such as Cloudera, Hortonworks or MapR. The VM comes with a fully installed and configured Talend Platform for Big Data development studio with several test-drive scenarios to help you see the value that using Talend can bring to big data projects. The high-level sandbox architecture looks like:



There are four scenarios in this cookbook and sandbox:

1. Analysis of clickstream data
2. Sentiment analysis on Twitter hashtags
3. Analysis of Apache weblogs
4. ETL Off-Loading

There are also basic demonstrations of several NoSQL databases for: Hive ELT, MongoDB, Cassandra and HBase. Each scenario and demonstration work independent of each other and you are free to walk through any of them as you desire.

Talend Platform for Big Data includes a graphical IDE (Talend Studio), teamwork, management, data quality, and advanced big data features. To see a full list of features please visit Talend's Website: www.talend.com/products/platform-for-big-data.

1.1.1 Pre-requisites to Running Sandbox

You will need a Virtual Machine player such as VMWare or Virtual Box. We recommend VMware Player which can be downloaded from the [VMware Player Site](http://www.vmware.com).

- Follow the VM Player install instructions from the provider
- The recommended host machine memory is 8GB
- The recommended disk space is 12GB (6GB is for the image download)

1.1.2 Setup and Configuration of Sandbox

If you have not done so already, download the Virtual Machine file at www.talend.com/talend-big-data-sandbox. You will receive an email with a license key attachment, and a second email with a list of support resources and videos.

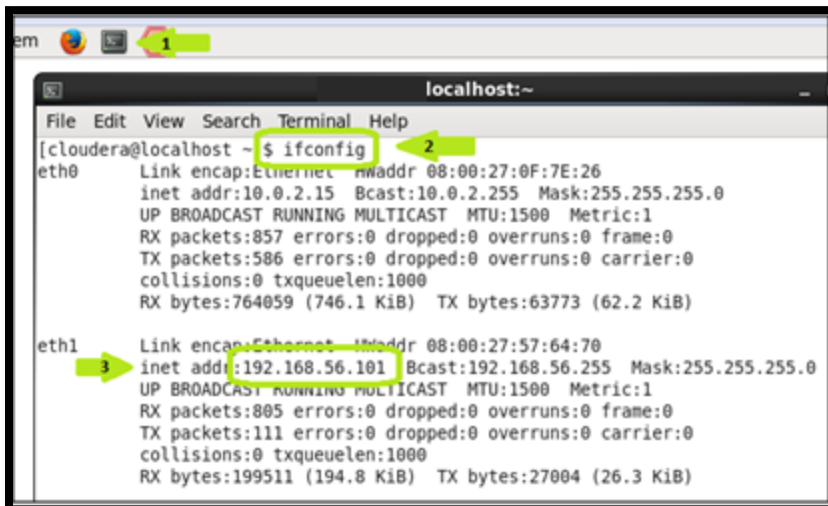
1. Open the VMware Player
2. Click on **“Open a Virtual Machine”**
 - a. Find the .ova file you downloaded
 - b. Select where you would like the disk to be stored on your local host machine:
e.g. C:/vmware/sandbox
 - c. Click on **“Import”**
3. Edit Settings if needed:
 - a. Check the setting to make sure the memory and processors are not too high for your host machine.
 - b. It is recommended to have 6GB or more allocated to the Sandbox VM and it runs very well with 8GB if your host machine can afford the memory.
4. Make sure there are two Network Adapters: **“Host Only”** and **“NAT”**. If one or both are missing add the network Adapter as follows:
 - a. Click **“Add”**
 - b. Select Network Adapter : **“Host Only”** and **“NAT”** and select **“Next”**
 - c. Select the needed Network that is missing and repeat as necessary
 - d. Once finished select OK to return to the main Player home page.
5. Start the VM

2 Talend License and Services Status

2.1 Talend License Setup

You should have been provided a license file by your Talend representative or by an automatic email from Talend Support. This license file is required to open the Talend Studio and must reside within the VM. The virtual player tools are installed on both the VMWare and VirtualBox versions so you should be able to copy and paste the file or the content of the file to the VM. If not then here is another way to place the license file on the VM:

1. Start the Virtual Machine and let it boot up completely
2. Once you see the desktop, open a Terminal Window and type the command: **ifconfig** and press <ENTER>. This command will display the IP of the Virtual Machine (i.e. – 192.168.x.x). Make note of your IP address.

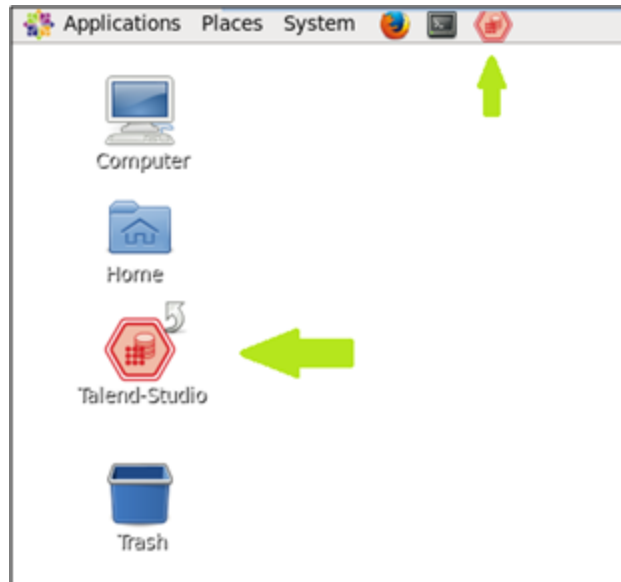


```
localhost:~
File Edit View Search Terminal Help
[cloudera@localhost ~]$ ifconfig
eth0      Link encap:Ethernet  HWaddr 08:00:27:0F:7E:26
          inet addr:10.0.2.15  Bcast:10.0.2.255  Mask:255.255.255.0
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:857 errors:0 dropped:0 overruns:0 frame:0
          TX packets:586 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:764059 (746.1 KiB)  TX bytes:63773 (62.2 KiB)

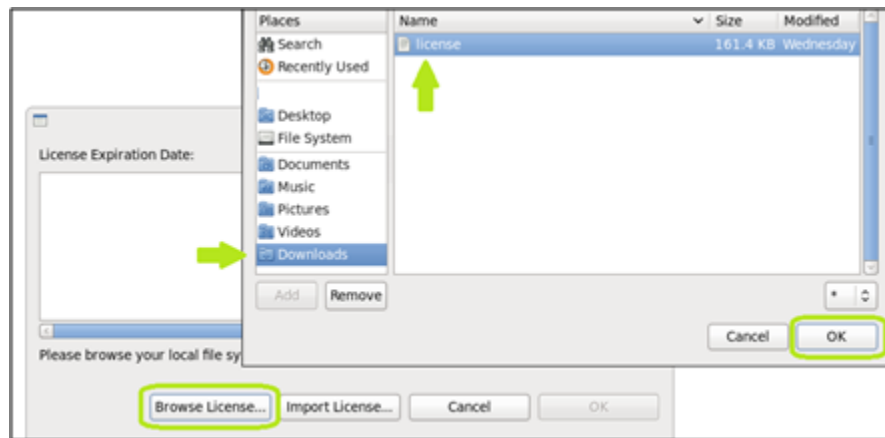
eth1      Link encap:Ethernet  HWaddr 08:00:27:57:64:70
          inet addr:192.168.56.101  Bcast:192.168.56.255  Mask:255.255.255.0
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:805 errors:0 dropped:0 overruns:0 frame:0
          TX packets:111 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:199511 (194.8 KiB)  TX bytes:27004 (26.3 KiB)
```

3. On your local PC, open an FTP client such as FileZilla and use the following credentials to connect to the VM (Note: the VM must be running in order to connect via FTP and have an IP address like 192.168.xx.xxx):
 - a. Host: <IP Address noted on VM>
 - b. Username: talend
 - c. Password: talend
 - d. Port: 22
4. Transfer the license file to the VM and save to a directory for easy reference (i.e. the Downloads directory).
5. Once the license file is successfully transferred, you can open Talend Studio. This can be done a couple different ways. On the Desktop there is a shortcut to launch the Studio or up on the top menu bar there is a Launcher icon as well. Click either to launch.

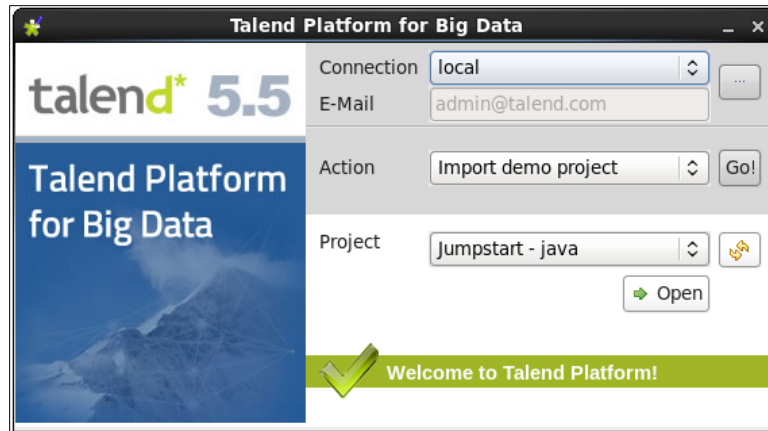
Note – If your IP address does not look like 192.168.xx.xxx then you may need to assure your network adapters are setup right on the VM Player.



6. The Talend Studio will start to launch and the first thing it will ask for is the License Key.
 - a. Click the “Browse License...” button. Then in the pop-up, specify the directory where you just saved the license file and Click OK.



7. The license will initialize and indicate the “License Expiration Date”. Click OK again and you will be presented with the project selection page. Select the **Jumpstart** project and click **Open**.



You may be prompted to Sign-in or to Register for the TalendForge Community, an online community of other Talend software users and technical experts to share tips, tricks and best practices. Additionally you will be able to view documentation and technical articles from the Talend Software Knowledgebase. We recommend you take advantage of this valuable source of information to get the most out of your Big Data journey with Talend.

Once your TalendForge registration is complete, Talend Studio will finish launching and the "Welcome Page" will appear. You can close the welcome page to see the Talend Integration perspective and the Jumpstart Sandbox projects.

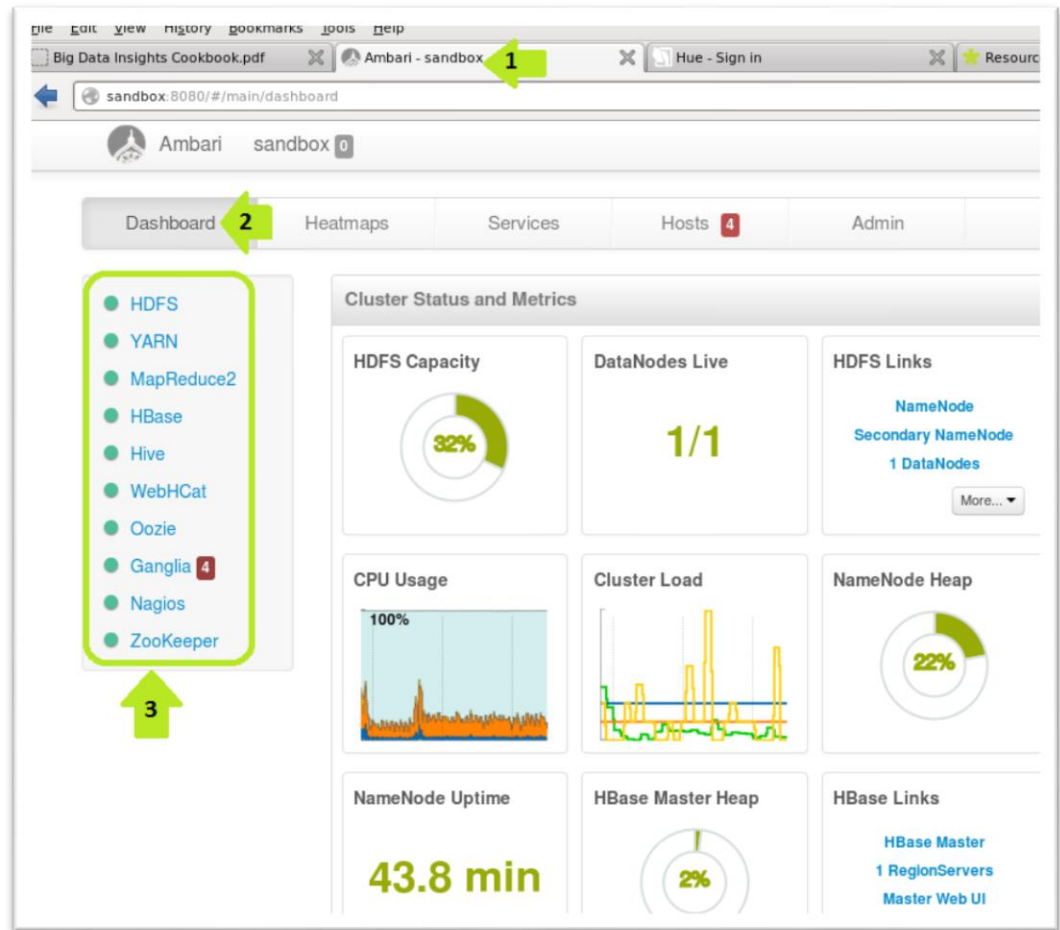
2.2 Hortonworks Services Status

Next we need to assure the Hortonworks Hadoop services are all running. Open the Browser where there should be several tabs already open. Click on either the already open tab or bookmark toolbar shortcut for "Ambari".

- **Username:** admin
- **Password:** admin

View the Services from the home page of the Ambari Manager

1. Click on **Dashboard**

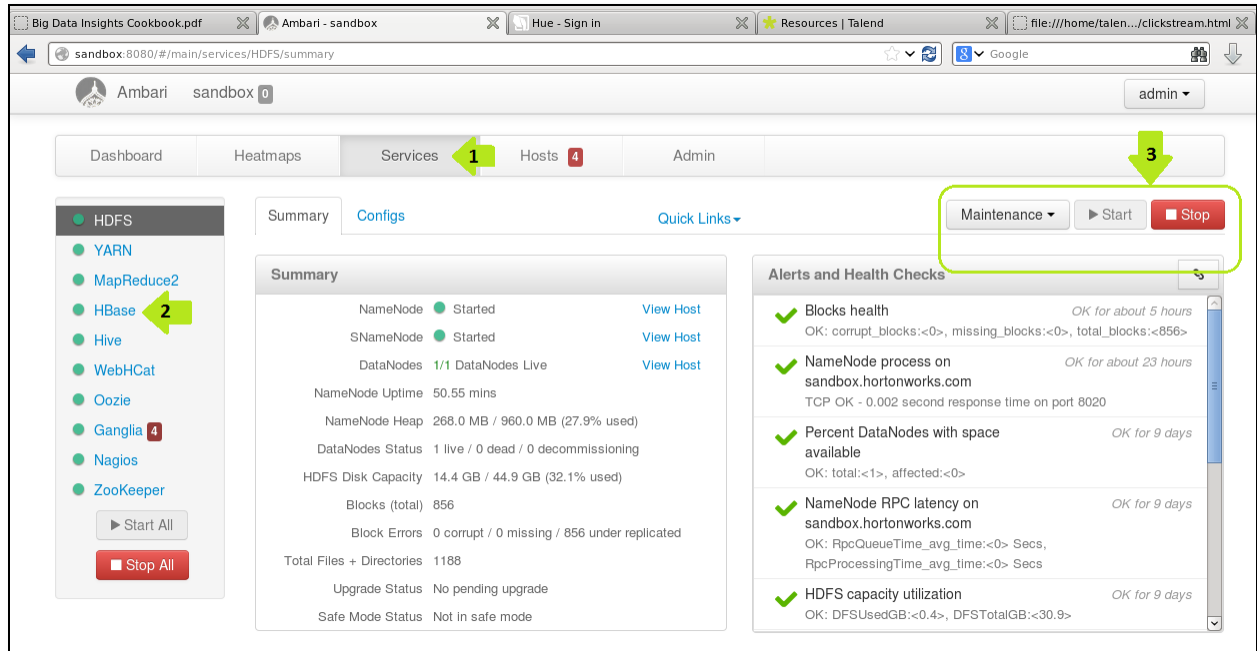


View the Services from the home page of the Ambari Manager

1. Click on **Services**
2. Click on the service in question
3. See any Red squares next to a server if there are any
4. Click on the **Stop**
5. Then click **Start** to restart the service

The service should then restart and go to a 'healthy' status.

**On Talend's Sandbox some services may have issues even after restart this will not impact any of the scenarios.*



Note: The Talend Big Data Sandbox is built on top of Hortonworks VM. If you have questions/concerns regarding the Hortonworks Sandbox, we suggest reaching out to Hortonworks technical support directly.

Note - The Root Password on the Hortonworks VM is:

- **Username:** root
- **Password:** Hadoop

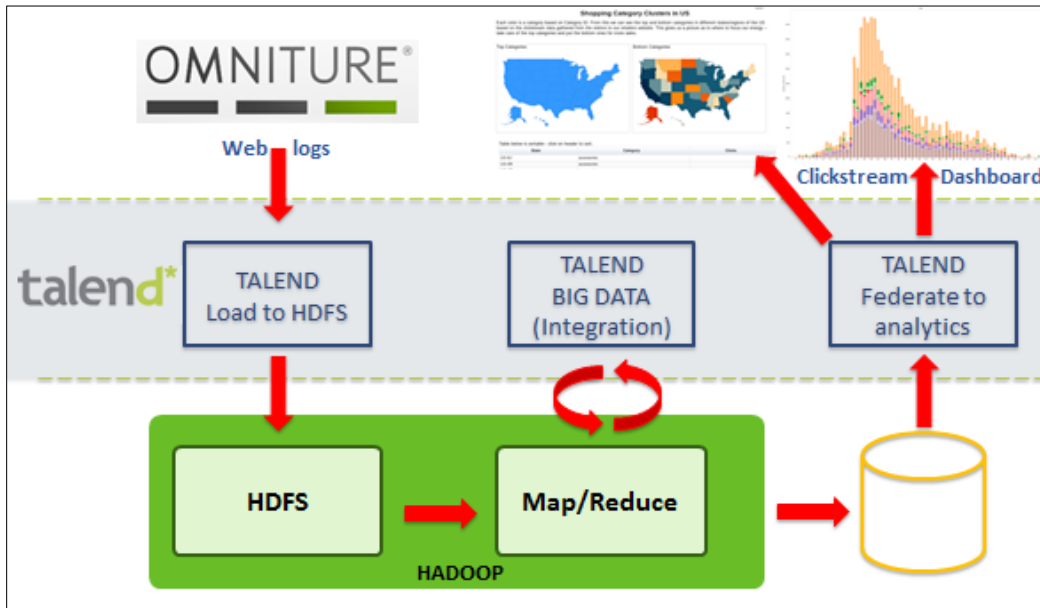
Note – The Talend user password on the Hortonworks VM is:

- **Username:** talend
- **Password:** talend

Now we can start working on the Talend Big Data Sandbox examples!

3 Scenario: Clickstream Insights

3.1 Overview



Clickstream¹ data provides insights to companies on how users are browsing their product web pages and what flow they go through to get to the end product. Omniture is one company that provides this type of data. In the example for Clickstream Insights you will load the data to HDFS and then use a Talend MapReduce job to enrich the data and calculate different results for different dashboards like a Google Chart or a Tableau Report, but any analytic tool that can connect to Hive can be used.

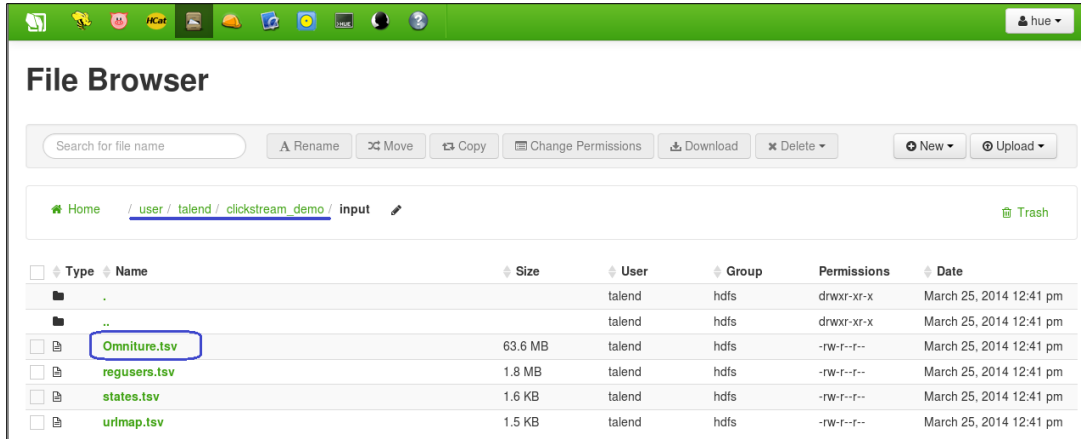
3.2 Clickstream Dataset

Clickstream log files are unstructured and can be viewed using the management console:

Hue Management Console - <http://sandbox:8000/about/>

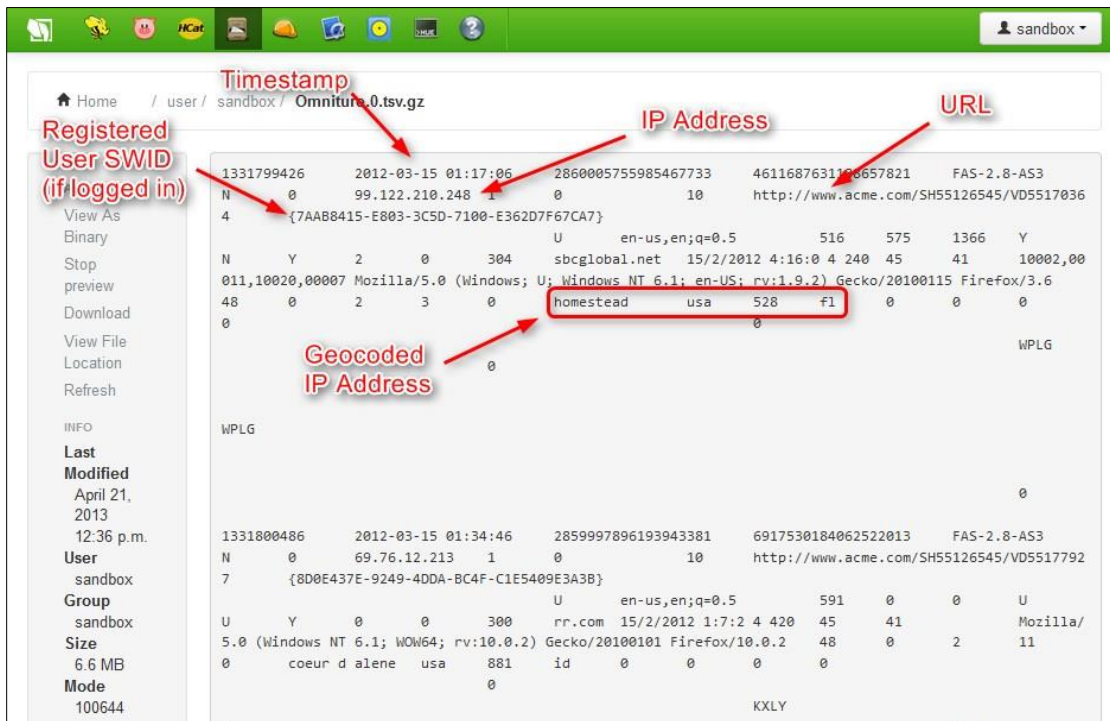
- **User** – talend
- **Password** - talend

¹ Clickstream is based on an original demo created by Hortonworks



The raw data file appears in the File Browser, and contains information such as URL, timestamp, IP address, geocoded IP address, and user ID (SWID).

The Omniture log dataset contains about 4 million rows of data, which represents five days of clickstream data. Often, organizations will process weeks, months, or even years of data.



Using HiveQL you can process and access the data, for example:

```
1. create table webloganalytics as
2.
3.     select
4.
5.         to_date(o.ts) logdate,
6.
7.         o.url,
8.
9.         o.ip,
10.
11.        o.city,
12.
13.        upper(o.state) state,
14.
15.        o.country,
16.
17.        p.category,
18.
19.        CAST(datediff(
20.            from_unixtime( unix_timestamp() ),
21.            from_unixtime(
22.                unix_timestamp(u.birth_dt, 'dd-MMM-yy')) / 365 AS
23.            INT) age,
24.
25.        u.gender_cd gender
26.
27.    from
28.
29.        omniture o
30.
31.        inner join products p on o.url =
32.        p.url
33.
34.        left outer join users u on o.swid =
35.        concat('{', u.swid , '}')
```

Keep in mind some of the limitations to Hive processing - here the actual age will not be computed exactly right as it is using a standard 365 year.

There is much more Hive coding needed to complete this example, which is not shown here.

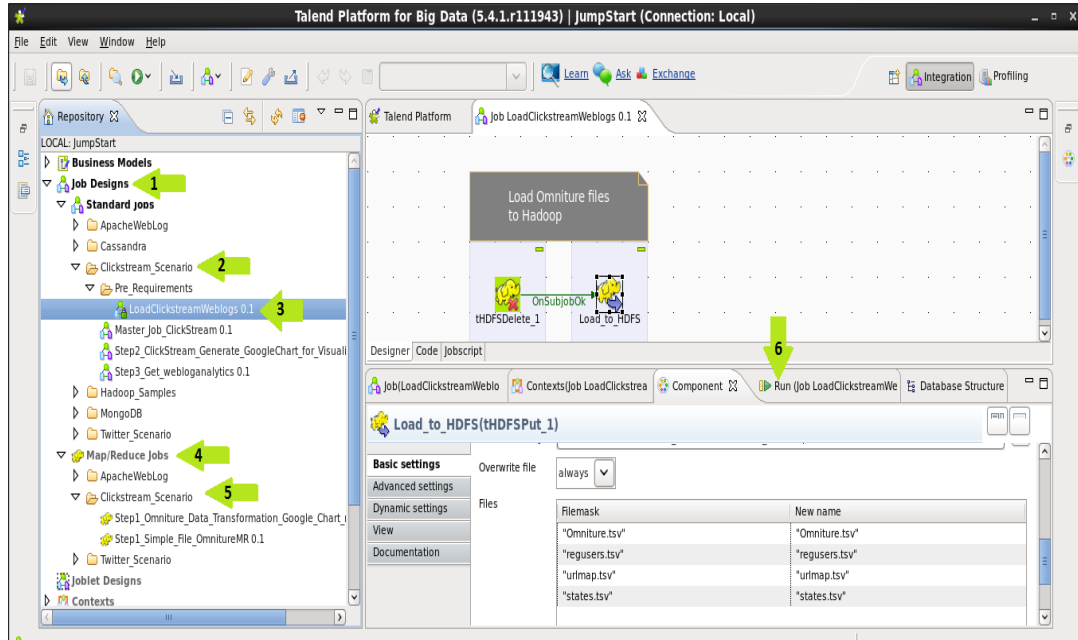
3.3 Using Talend Studio

3.3.1 Talend HDFS Puts

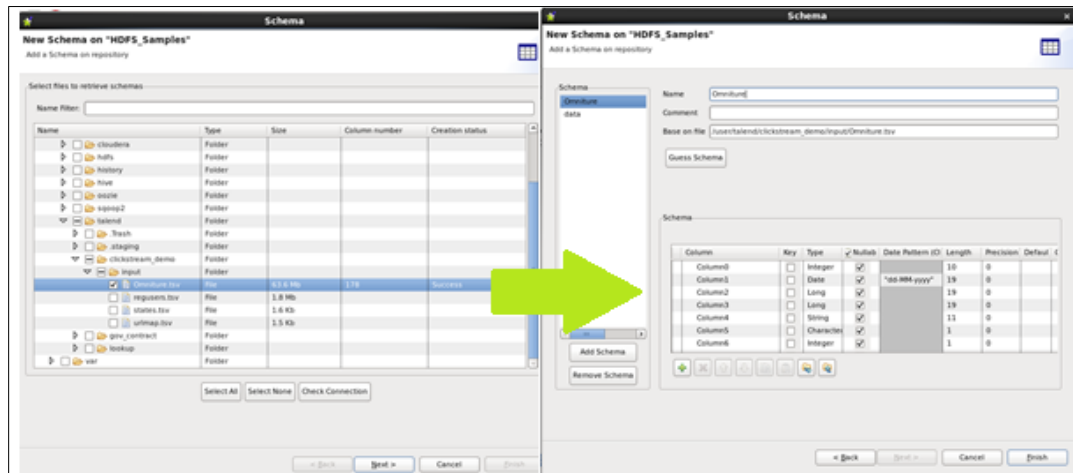
Using simple components, we can load data into HDFS for processing on HIVE or MapReduce and YARN.

Review the process in this Clickstream example for putting files directly in to HDFS.

Job Designs / Standard Jobs / Clickstream_Scenarios / Pre_Requirements / LoadWeblogs



Now that the data is in HDFS you can use Talends wizards to retrieve the file schema:



This new Schema can then be the input to the MapReduce process that will do joins to the Product URL Maps and user files in HDFS. (Also, you can use the wizard to import the URL and User schemas if needed. This is already done in the Sandbox for you.)

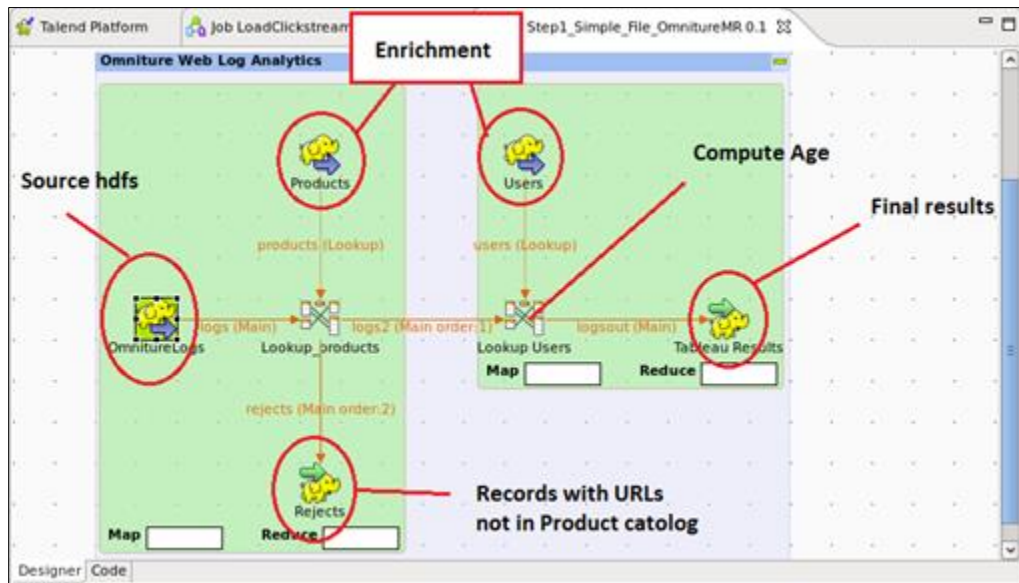
This is what you would call the 'schema on read' principle; how it allows any data type to be easily loaded to a 'data lake' and is then available for analytical processing.

3.3.2 Talend MapReduce Review

Open the following MapReduce process:

Job Designs / Map Reduce Jobs / Clickstream_Scenarios / Step1_Simple_File_OmnitureMR

This process will run completely native as MapReduce code. The first component on the left is the source file (the clickstream file with 170+ columns). It is joined to the Product HDFS file to match the URL in the log file to the known product pages on the website. Any URLs in the source log file that cannot be matched in the product catalog are rejected to a separate file. This file can be mined at a later time to make sure we are not missing new pages. Next, the data is matched to known users to determine things like ‘age’ and ‘gender’ for additional analysis. Finally the results are written to a new HDFS file.



To see how the lookups join or how you can apply logic like computing the ‘age’, double-click on the tMap labeled “Lookup Users”.

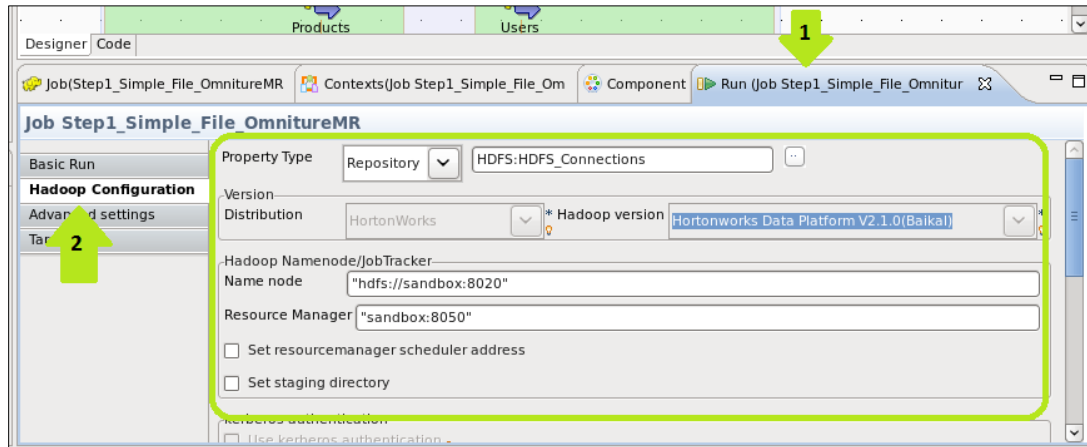
Expr. key	Column
logs2.swid	SWID
	BIRTH_DT
	GENDER_CD

Column	Key	Type	Nullab	Date Pattern	Length	Precisio	Defau	Comment
logdate		Date	✓	"dd-MM-yyyy"		0		
ip		String	✓			0		
url		String	✓			0		
swid		String	✓			0		
city		String	✓			0		
country		String	✓			0		
state		String	✓			0		

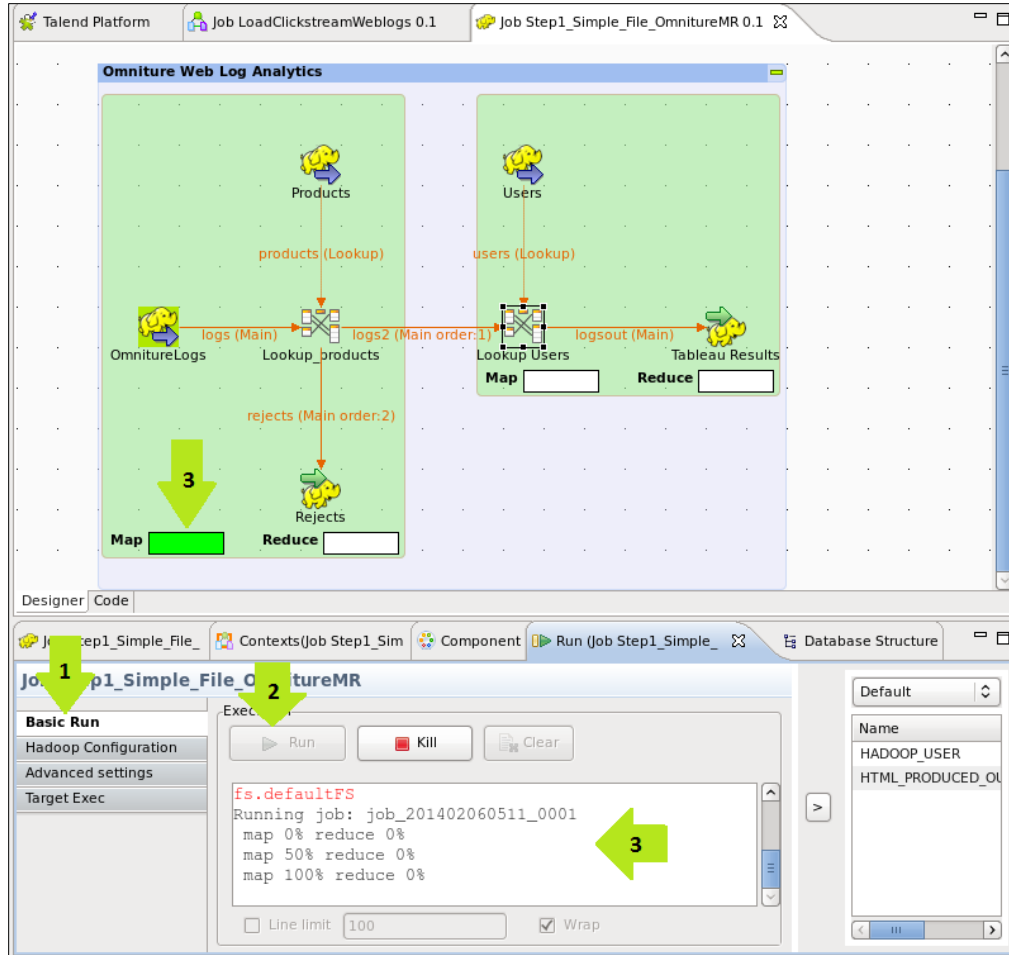
Column	Key	Type	Nullab	Date Pattern	Length	Precisio	Defau	Comment
logdate		Date	✓	"dd-MM-yyyy"		0		
ip		String	✓			0		
url		String	✓			0		
swid		String	✓			0		
city		String	✓			0		
country		String	✓			0		
state		String	✓			0		

You can run this job to view the results.

To run a process in Talend Studio you need to go to the Run tab. Then, on the left, confirm the Hadoop configuration for MapReduce processes. In the Big Data Sandbox all the jobs are using the same Hadoop metadata connections and are all configured to run so no changes should be needed.



To run the process click on the “Basic Run” menu option above the “Hadoop Configuration”, then click the Run button to start the process. You will then see the progress bars on the designer view advance to green bars as the steps complete. See below:



Once this is complete you can run the second MapReduce process.

Map Reduce Jobs / Clickstream_Scenarios/ Step1_Omniture_Data_Transformation_Google_Chart_mr

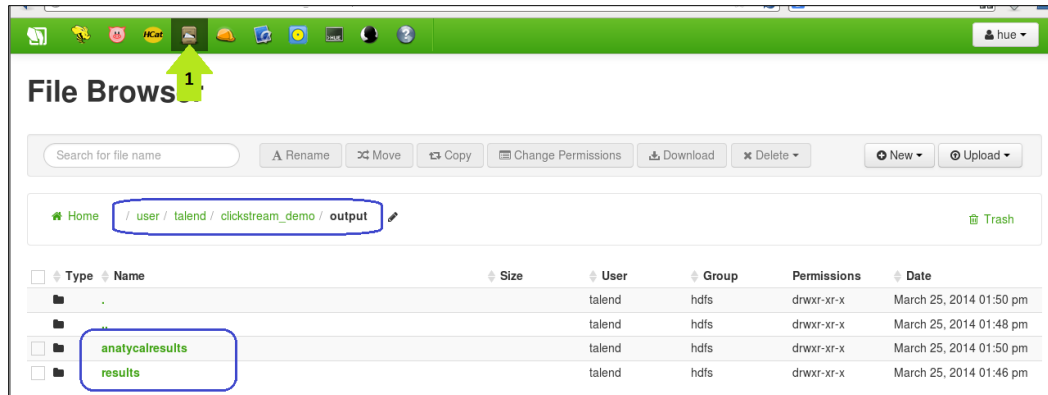
The result of this process is aggregated data indicating the product interests of different areas across the United States for visualization within a Google Chart.

Run this process in the same fashion as the first MapReduce process.

View the output data files in HUE (a browser-based web tool to view Hadoop data like HDFS and Hive). In Firefox there should be a tab already open to the <http://sandbox:8000/> location. If prompted for login, use the following credentials:

- **Username:** talend
- **Password:** talend

Open the File Browser and click on the links on the left side of the page to go to /user/talend/clickstream_demo/output



3.3.3 Talend to Google Charts and Hive

To format the results of the MapReduce processes to fit Google Charts, run the following job located under the Standard Jobs in the Clickstream_Scenario folder:

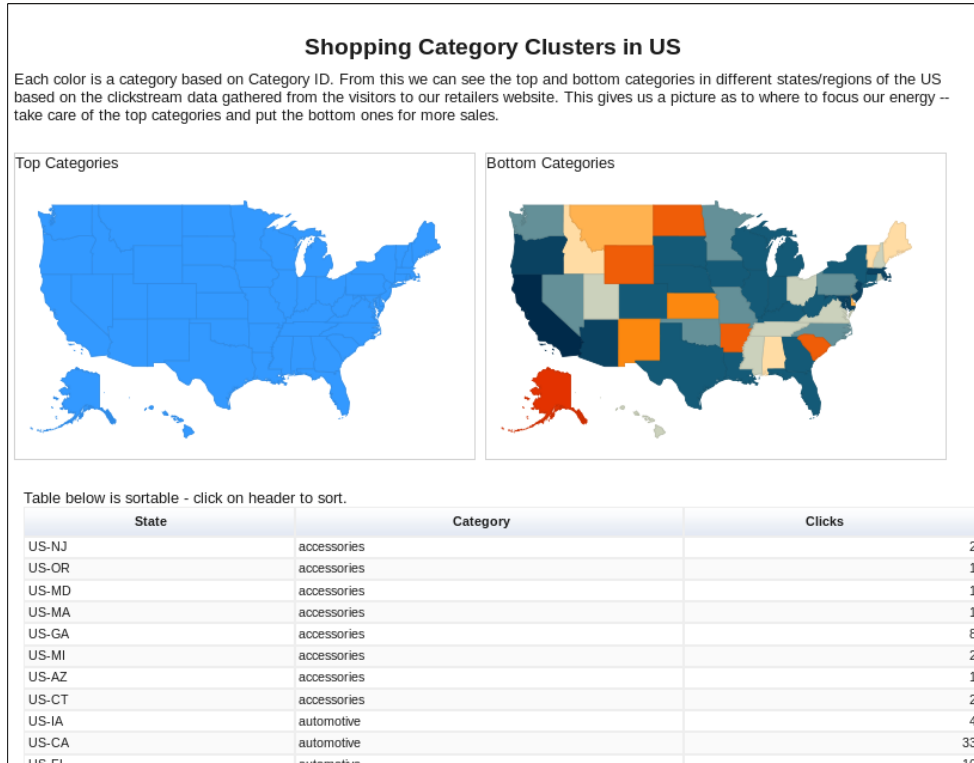
Step2_ClickStream_Generate_GoogleChart_for_Visualization

This job will read the HDFS files and put them into an html format required by the Google Charts API. You can see the result in the browser if you have internet access setup to your VM. (The Google Charts API connects to Google's website to render the results.)

To view the results in Google Charts, navigate to the following directory on the VM file system (not HDFS):

```
/home/talend/Documents/Clickstream/
```

Double-click on the clickstream.html file to open. You may need to right-click on the file, choose "Open With" and select "Firefox Web Browser".



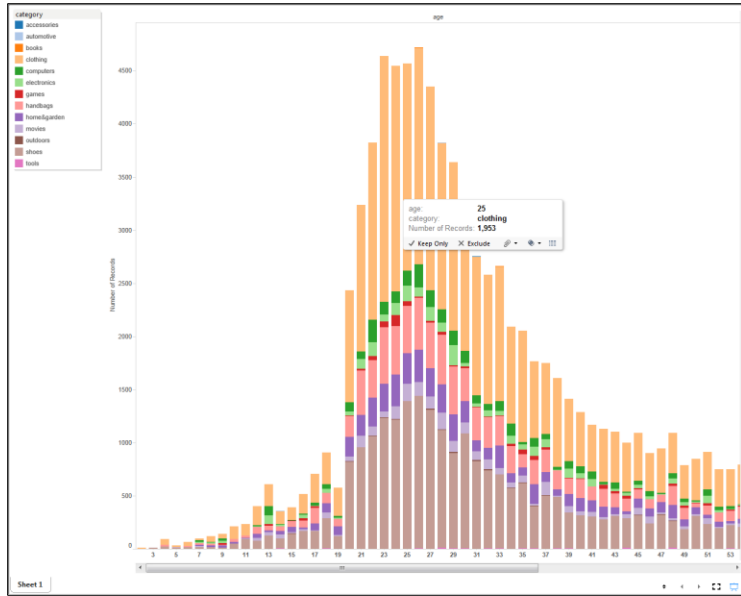
Run the final job in the Clickstream Scenario located under the Standard Jobs in the Clickstream_Scenario folder:

Step3_Get_webanalytics

This job sets the files needed for the Insights on the Click Stream logs. View the following file on the local VM file system (not HDFS):

`/home/talend/Documents/webanalytics.csv`

This file can be imported to MS Excel or other BI tools like Tableau (not included in the Big Data Sandbox) to see insights like this:



Or query the HIVE table to see results (in Hue select HIVE under the “Query Editors” dropdown. Then under My Queries use the saved query “Clickstream Analysis Data”):

The screenshot shows the Hue Query Editor interface with the following components:

- Query Editor** tab selected.
- Query Results: ClickStream Analysis Scenario** title.
- DOWNLOADS** section: Download as CSV, Download as XLS, Enable visualization, Save.
- Results** tab selected, showing a table with columns: logdate, ip, url, swid, city.
- Table Data:**

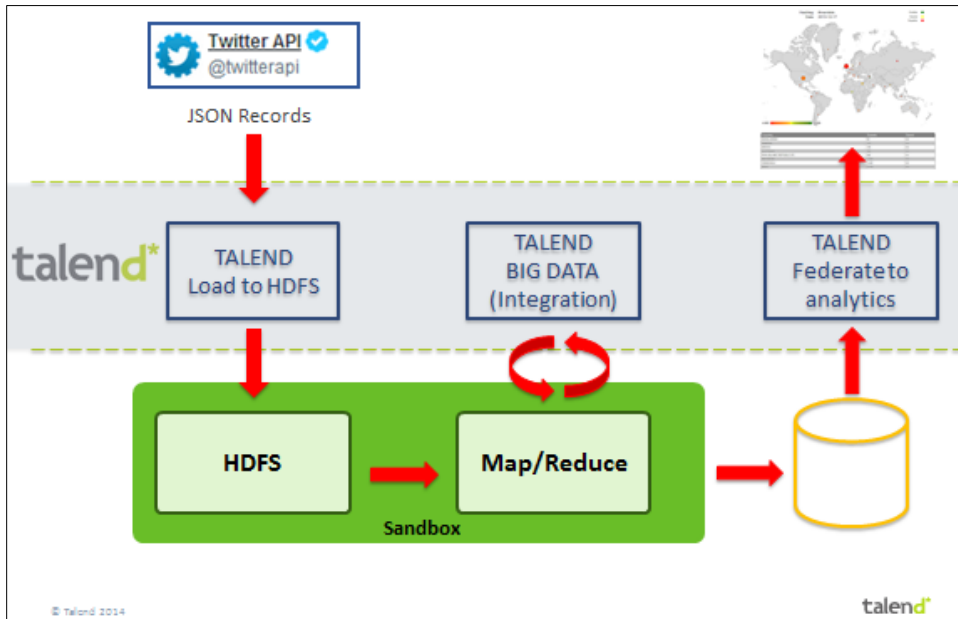
	logdate	ip	url	swid	city
0	12-03-2012	76.166.167.172	http://www.acme.com/SH55126545/VD55179433	0001BDD9-EABF-4D0D-81BD-D9EABFC0D07D	oxnard
1	12-03-2012	76.166.167.172	http://www.acme.com/SH55126545/VD55179433	0001BDD9-EABF-4D0D-81BD-D9EABFC0D07D	oxnard
2	12-03-2012	12.132.157.137	http://www.acme.com/SH55126545/VD55179433	000B90B2-92DC-4A7A-8B90-B292DC9A7A71	opelika
3	15-03-2012	24.184.60.95	http://www.acme.com/SH55126545/VD55179433	000C47AD-EBFC-CDB8-CF70-DC4C2ED5051B	brookly
4	15-03-2012	24.184.60.95	http://www.acme.com/SH55126545/VD55179433	000C47AD-EBFC-CDB8-CF70-DC4C2ED5051B	brookly
5	15-03-2012	24.184.60.95	http://www.acme.com/SH55126545/VD55179433	000C47AD-EBFC-CDB8-CF70-DC4C2ED5051B	brookly
6	15-03-2012	24.184.60.95	http://www.acme.com/SH55126545/VD55179433	000C47AD-EBFC-CDB8-CF70-DC4C2ED5051B	brookly
7	15-03-2012	24.184.60.95	http://www.acme.com/SH55126545/VD55179433	000C47AD-EBFC-CDB8-CF70-DC4C2ED5051B	brookly
8	12-03-2012	24.58.5.10	http://www.acme.com/SH55126545/VD55179433	000E15BA-EB3E-14A6-4921-0E24C052821D	ithaca
9	12-03-2012	24.58.5.10	http://www.acme.com/SH55126545/VD55179433	000E15BA-EB3E-14A6-4921-0E24C052821D	ithaca
- Next Page** button at the bottom right.

You could use the HIVE ODBC to connect and pull this data into a BI tool now as well.

4 Scenario: Twitter Sentiment Insights

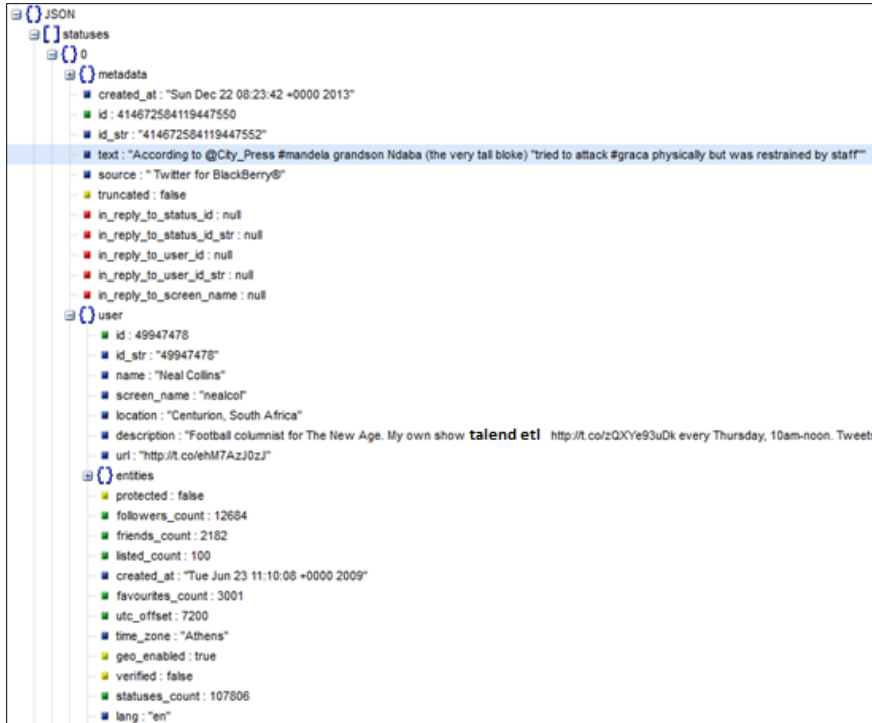
4.1 Twitter Sentiment Analysis Overview

In this scenario Talend has taken on the popular big data use case of social media analysis. Here you will stream all Tweets related to an entered #hashtag value for a brief period of time and then provide analysis on the hashtag sentiment and geolocations. You will get exposure to Ingesting, Formatting, Standardizing, Enriching and Analyzing Tweets within Hadoop (HDFS or other storage + MapReduce computation) to capture the sentiment based on geographic regions of the world.



4.2 Twitter Data

We are using the standard Twitter API and a tRESTClient component to ingest Tweets based on a hashtag, entered as a context variable into the job. The data from Twitter, and indeed other popular social media, sites typically return JSON records. Below is an example that will be parsed and analyzed:

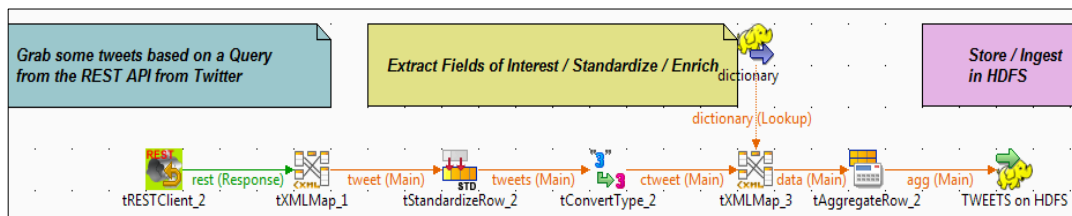


4.3 Talend Processes

The Sentiment Analysis scenario is a 3 steps process: (be sure the Pre-requirements process has been run as well).

4.3.1 Retrieve the Data

The purpose of the Step1_Import_Tweets process is to query Twitter. This job is using the tRestClient and the Rest API from Twitter to capture Tweets about a given hashtag or Keyword.



In this Job there are a few operations such as extracting from each Tweet only the valuable information, standardizing the TEXT (the message) of those tweets and applying a first layer of transformation. The Process is using a dictionary of positive, negative, and neutral words to determine the sentiment of the tweet as well.

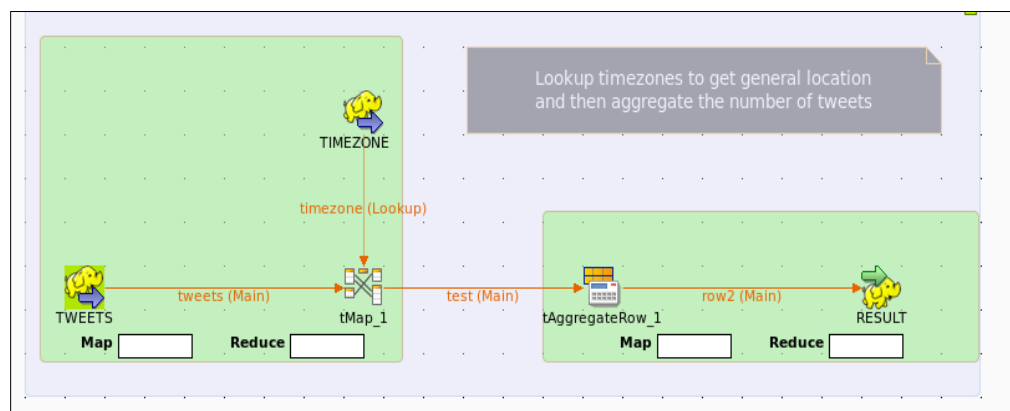
When you run the Job a prompt will pop up with the question about the Hashtag; feel free to use any hashtag.

***Note if the process does not complete and fails on the tRestClient_2 make sure the VM has access to the internet otherwise this scenario will fail as it is querying Twitter live.*

To view the twitters in the HDFS use the Hue HDFS file browser to see the output of this process:

4.3.2 Process and Aggregate Results

Aggregation of the tweets and enrichment occur in the MapReduce process. Within this MapReduce process it is adding the geo-location data into the data as well as determining the number of followers and re-tweets each tweet had based on the user info from Twitter.



Once the process has completed you can find the results in HDFS at `/user/talend/bigdata_demo/result/tweet_analysis/part-00000`

This data has been prepared for the final step of loading the sentiment to a format that Google Charts can then display in the form of a heat map on a global scale.

Home / user / talend / bigdata_demo / result / tweet_analysis / part-00000

ACTIONS

[View As Binary](#)

[Edit File](#)

[Download](#)

[View File](#)

[Location](#)

[Refresh](#)

INFO

Last Modified
March 25, 2014
2:56 p.m.

User
talend

Group
hdfs

Size
589 bytes

Mode
100644

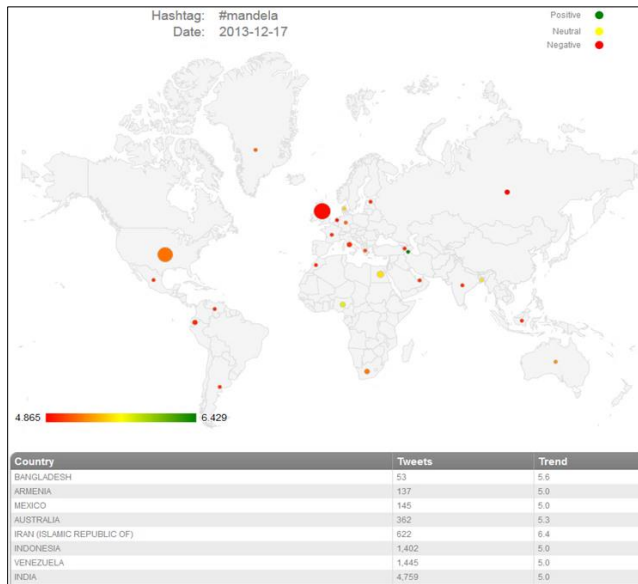
First Block
Previous Block
Next Block
Last Block

hashtag	date	country	count	trend
#talend	2014-03-25	ECUADOR	74231	6.0
#talend	2014-03-25	FRANCE	304282	6.0301657
#talend	2014-03-25	GUAM	2391	6.3636365
#talend	2014-03-25	INDIA	99470	6.014493
#talend	2014-03-25	IRAN (ISLAMIC REPUBLIC OF)	2392	6.0
#talend	2014-03-25	JAPAN	2391	6.75
#talend	2014-03-25	RUSSIAN FEDERATION	151	6.0
#talend	2014-03-25	SLOVAKIA	2393	6.0
#talend	2014-03-25	SOUTH AFRICA	150	5.9333334
#talend	2014-03-25	SPAIN	150	6.1875
#talend	2014-03-25	UNITED KINGDOM	5903	6.205128
#talend	2014-03-25	UNITED STATES	14190	6.2349095
#talend	2014-03-25	UZBEKISTAN	2715	6.2857146

First Block
Previous Block
Next Block
Last Block

4.3.3 Analysis and Sentiment

This final step is a process to generate the Google Chart HTML page. This page will be saved on a local file and use the Google Charts APIs to show the sentiment of the tweets across the globe. This is a basic process of reading the data from HDFS and putting it in an analytical tool. In this case it is doing the final formatting changes needed for Google Charts.

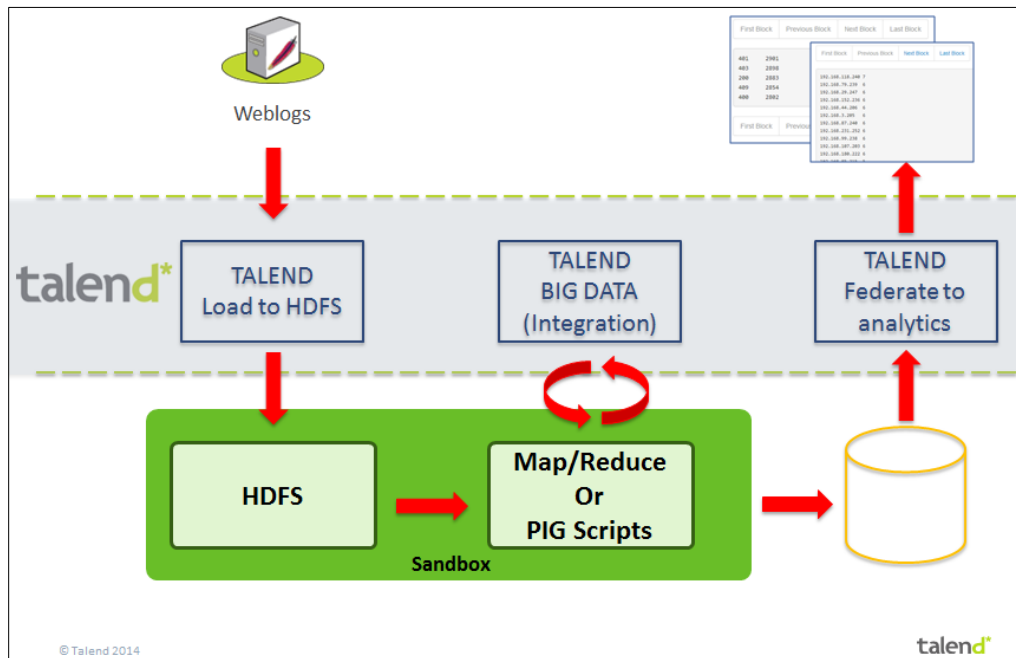


With a simple 3 step process you can now start seeing trends for popular hashtags related to your products on a regular basis as well as if they are using positive or negative tones against your brands.

5 Scenario: Apache Weblog Insights

5.1 Apache Weblog Overview

Building processes that capture and analyze billions of web traffic records can be very challenging. In this example, Talend demonstrates taking large volumes of weblog data and before loading to HDFS doing some basic filtering to first get some of the noise records out before processing and doing aggregates on the logs.



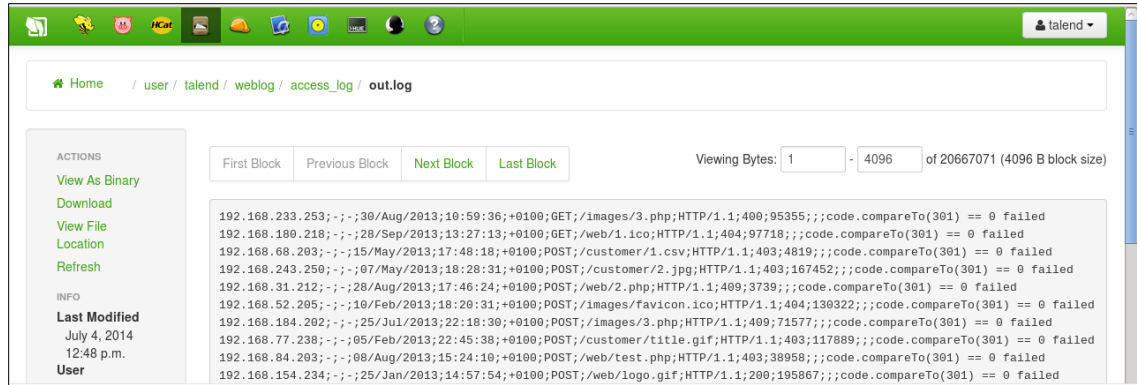
Once the data is in HDFS we will do data extractions of the IPs and return codes using PIG components to show the different functionality available in Talend for PIG. You can then also go to the MapReduce process and see how to accomplish the very same process done in PIG but with MapReduce. This will allow you to compare and contrast different styles of Hadoop processing first hand.

Finally there is an output process that puts the data to the Talend console to demonstrate the results from both PIG and MapReduce.

5.2 Apache Weblog Data

Talend offers a component that will read native Apache weblog files. In this scenario the Apache weblog component is used to read a file that is generated using the Talend Data Generator component called tRowGenerator. This component is a very useful way to simulate and test processes in development and QA.

Apache Weblog data will be loaded into HDFS and a HCatalog in the next step, but first here is a view of that raw file:



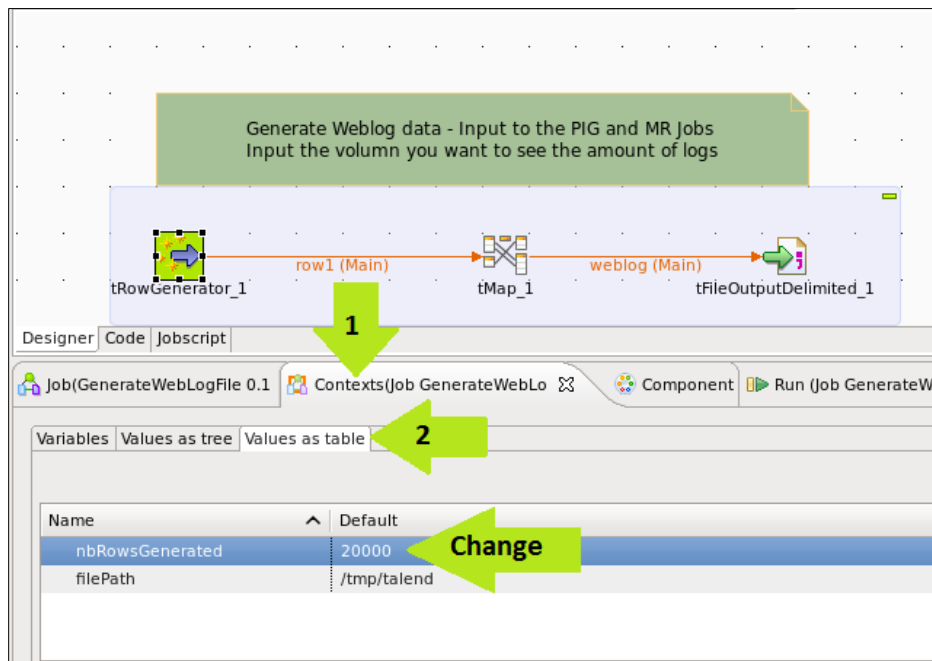
5.3 Talend Processing

5.3.1 Talend Filter and Load Data

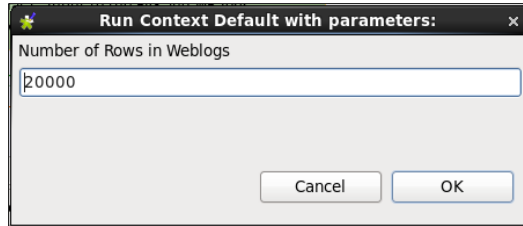
The job is setup to process 20,000 records. If you want to use more data you can do one of two things:

- Modify the settings on the context variable called “nbRowsGenerated”. This can be found in the job:

/Standard Jobs / ApacheWeblogs / GenerateWebLogFile



- As a simple alternative, you can update the default value on the Pop-up when the job is executed.

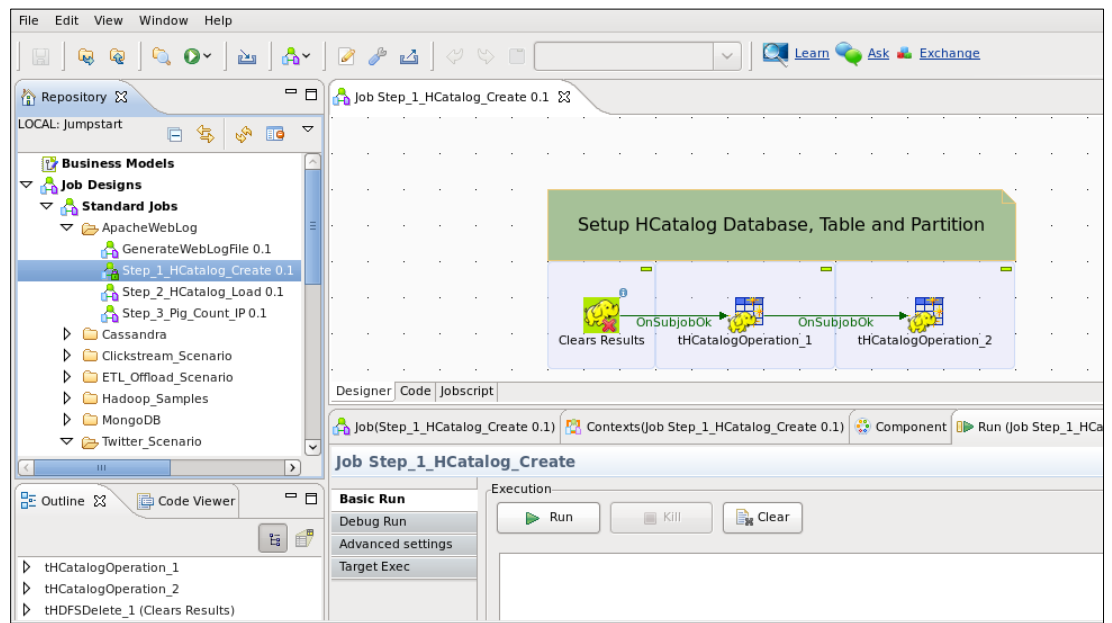


****Note:** You do not have to change this number in either location if you don't care about how many rows are processed.

Now run the process that reads the weblog data as it gets generated into HDFS and notice the filter on the code value of 301. In this example we are showing the ease of using Talend to limit the data that is processed into your Data Lake or Hadoop system.

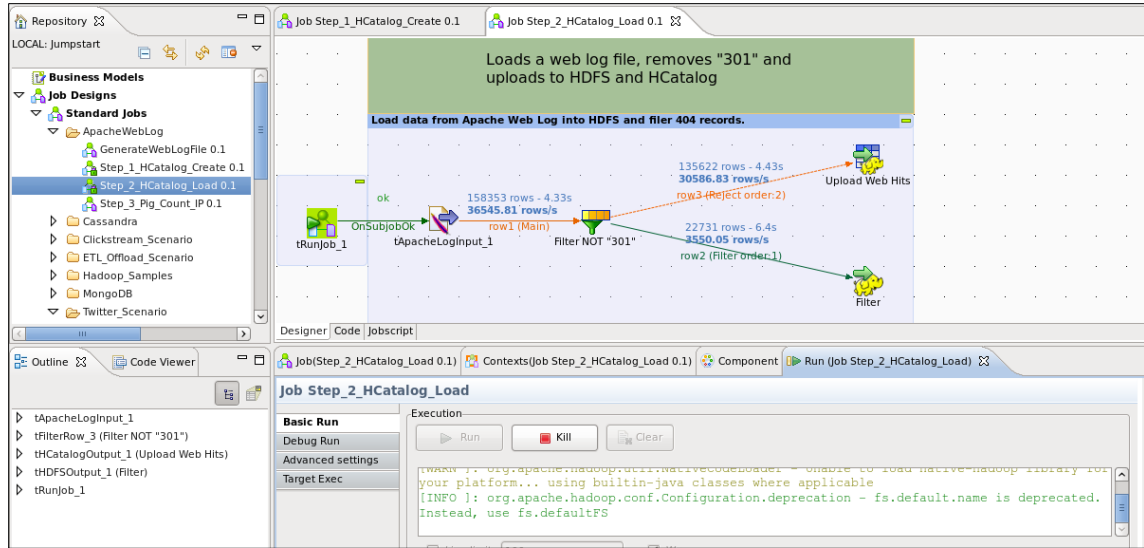
Run the following two processes in the ApacheWeblog folder:

1. Step_1_HCatalog_Create

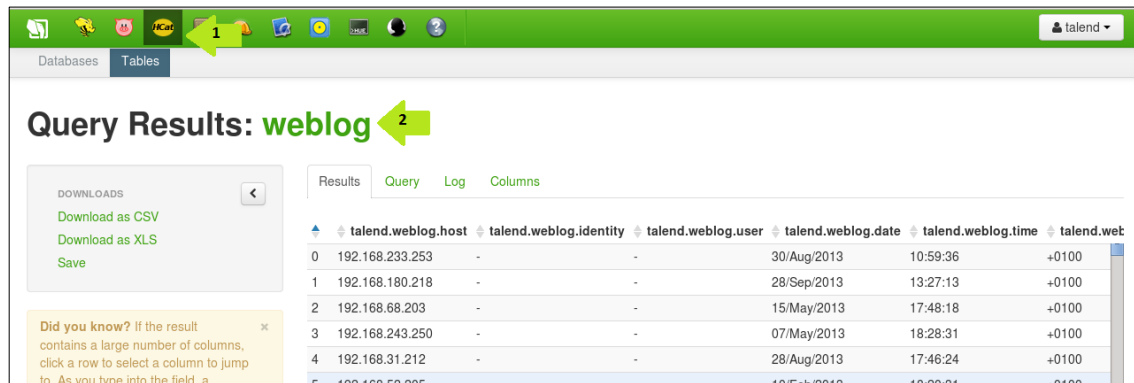


2. Step_2_HCatalog_Load

The Step 1 process does some house cleaning and sets up HCatalog tables. Step 2 loads the weblog data into the HCatalog tables. Now you can view the data either through HCatalog browsing or directly on the HDFS.



In the Hue browser you can find the data in /user/talend/weblog/access_log/out.log. You can also view the data directly in HCatalog by navigating to H under the “Query Editors” dropdown. Once in the Hive Interface there is a link for “My Queries” and in there is a query saved to see the Weblog raw data loaded to Hive.



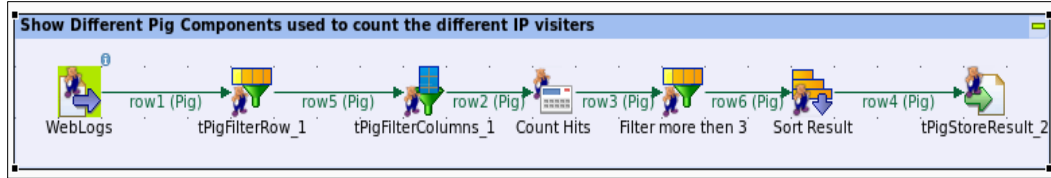
5.3.2 Talend PIG Scripts to Process

In the Weblog PIG analysis we have a process doing basic aggregations on the IP address data. This shows Pig with HCatLoader and basic column and row filtering of the data. Then, using PIG functions, it performs a count on a specific attribute of the data. There is a Map Reduce example to show how these counts the unique IP address the same as Pig.

In the Standard Jobs/ApacheWebLog folder:

Step_3_Pig_Count_IP

This job will count the visits by a unique IP address.



Results can be found on HDFS:

`/user/talend/weblog/Pig_apache_ip_cnt`

Talend Studio provides a quick and easy way to take advantage of PIG without needing a full understanding of the PIG Latin language to perform analysis on your data. By providing this built-in functionality, Talend is helping reduce the skill-level challenges many employers face. Even if your data scientist is already using PIG in house, you can still use Talend to enhance and take full advantage of any existing User Defined Functions (UDFs) in use today as well as speed up the development of new PIG data analysis.

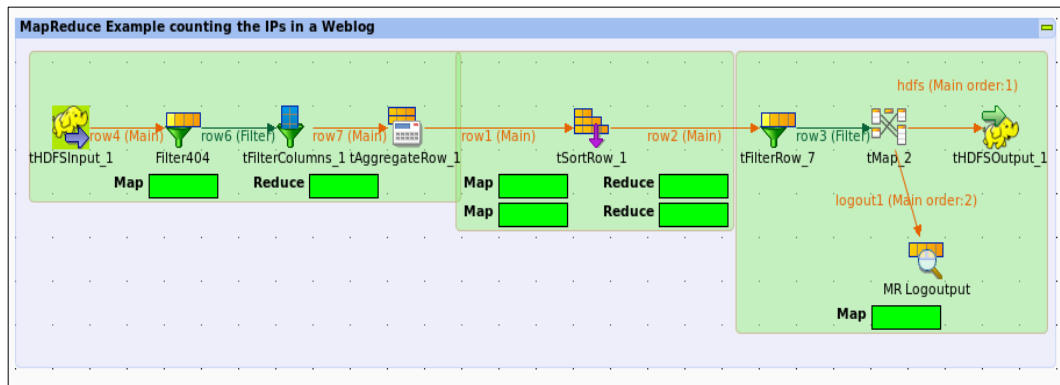
5.3.3 Talend MapReduce to Process

For comparison you can do a similar unique IP count using a MapReduce process to see the differences between a MapReduce process and a PIG process. The results and processing are very similar to PIG because ultimately PIG uses MapReduce under the covers. Again, this is only to provide a comparison of the two different methods of doing analysis on data in HDFS.

Process is found in the MapReduce Jobs /ApacheWebLog:

Step_4_MR_Count_IP

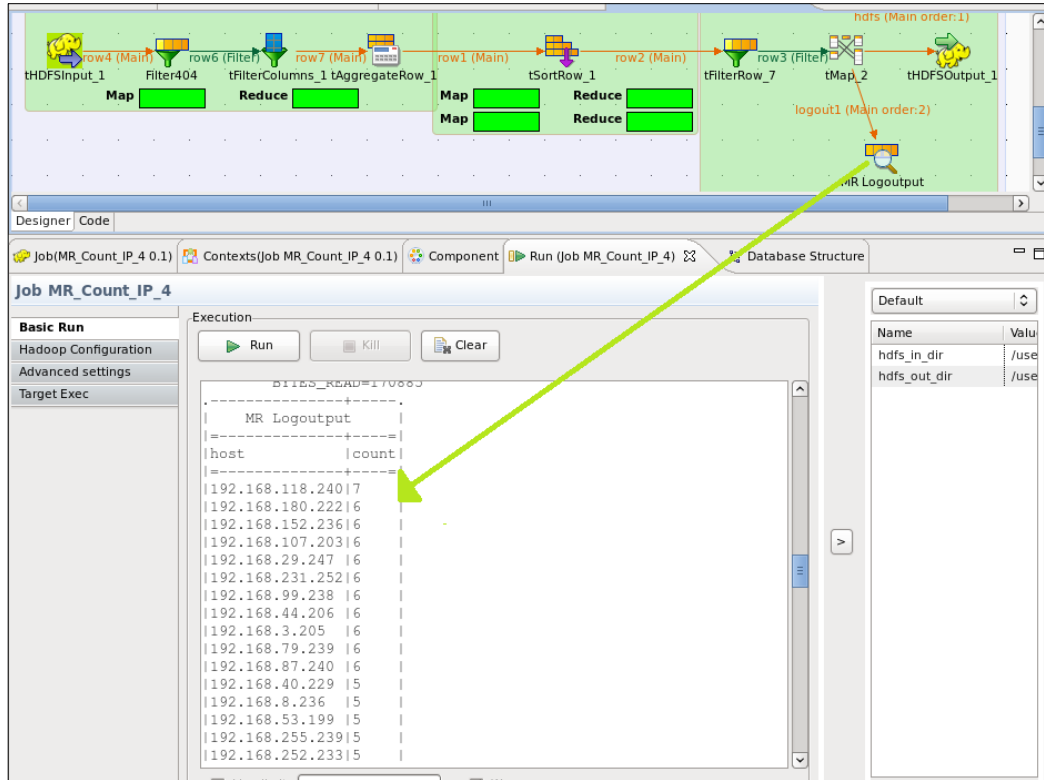
This job will count the visits by a unique IP address.



Results can be found on HDFS:

`/ user/ talend/mr_apache_ip_out/`

Also, the MapReduce process shows the output of the MapReduce job in the Talend Studio Run tab console:

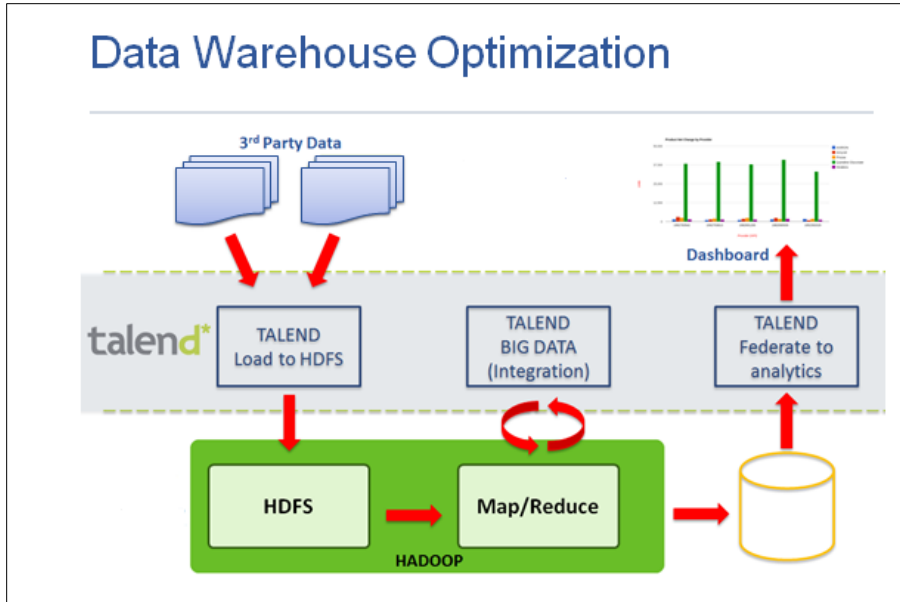


With all the different big data technologies and platforms/projects available in Hadoop environments you will find there are many ways to achieve the same results. The technology and platform you choose to solve a particular problem will often depend on the situation and the data. With Talend, the solution can be based on the right technology and not on whether you have the right skills for PIG or MapReduce or some other technology. Because Talend makes all the solutions equally the same level of skill to implement, your IT department is already equipped with the right tools to give your business the right answers from you unique data sources.

6 Scenario: ETL Off-loading

6.1 Overview

Processing large volumes of 3rd party data has always been a challenge in the old world of ETL, where it would take just as long to un-compress the data as it did to load into a data warehouse. The long execution times usually resulted in the trade-off of up-front data quality analysis which resulted in costly errors later on. These errors resulted in additional hours and sometimes days to back out and restore the data warehouse to a more stable state. In this scenario we will look at how Talend Big Data can help optimize your data warehouse by off-loading the ETL overhead to Hadoop and HDFS while minimizing the time-to-value for your business.



With Talend Big Data and Hadoop, the data quality analysis can be done before data loading takes place, without the need to un-compress the data files and in a fraction of the time necessary to load the data. Ultimately this will save costs as well as ensure valuable data remains in high quality, thereby allowing the business lines to react faster to changing market trends and to make quicker business decisions.

6.2 Data

This scenario comes ready-to-run with the data already staged for execution. The compressed data files will be loaded into HDFS with a simple tHDFSPut component. Even while the files remain compressed, the raw data can be viewed within the Hue File Browser.

The screenshot shows the Hue File Browser interface. The breadcrumb path is "Home / user / talend / Product_demo / Input / Product_Monthly_Current_Part-0000.gz". On the left, there are "ACTIONS" (View As Binary, Stop preview, Download, View File, Location, Refresh) and "INFO" (Last Modified: July 4, 2014 1:24 p.m.). The main area displays a table of data with columns for NPI_NUMBER, PRACTITIONER_TYPE_DESC, CREDENTIAL_DESC, PostalCode, Product_NDC, Product_Name, and monthly counts (MONTH_1 to MONTH_18).

NPI_NUMBER	PRACTITIONER_TYPE_DESC	CREDENTIAL_DESC	PostalCode	Product_NDC	Product_Name	MONTH_1	MONTH_2	MONTH_3	MONTH_4	MONTH_5	MONTH_6	MONTH_7	MONTH_8	MONTH_9	MONTH_10	MONTH_11	MONTH_12	MONTH_13	MONTH_14	MONTH_15	MONTH_16	MONTH_17	MONTH_18
1366430530	Physician	DOCTOR OF MEDICINE	12926	0002-1200	Amyvid	77	83	69	93	122													
38	53	142	92	137	66	73	59	96	44	128	53	66											
1497734883	Advanced Practice Nurse	NURSE PRACTITIONER	79248	0002-1975	AXIRON	41	109	142	95														
33	110	120	84	29	75	126	44	72	79	88	132	25	147										
1841275971	Physician	DOCTOR OF MEDICINE	02321	0002-3004	Prozac	34	137	40	95	95													
132	38	39	60	53	114	61	63	96	103	59	75	38											
1265424378	Physician	DOCTOR OF MEDICINE	36800	0002-1200	Amyvid	100	116	64	106	43													
51	68	127	100	110	62	35	32	29	65	117	52	38											
1760466379	Advanced Practice Nurse	NURSE PRACTITIONER	45747	0002-1407	Quinidine Gluconate																		
35	84	47	96	49	133	123	69	71	144	37	31	50	104	46									
295																							

The final output is a set of report files that can be federated to a visualization tool.

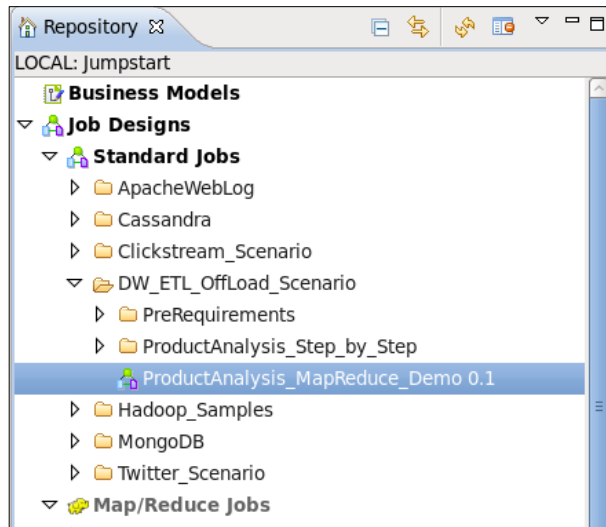
To further explore the capabilities of Talend within this scenario, you have the flexibility of generating a completely different set of input data files. Execute the scenario again with the new data files and review the resulting reports to see if you can find the inaccuracy within the data.

6.3 Talend Process

6.3.1 Single-Click Execution

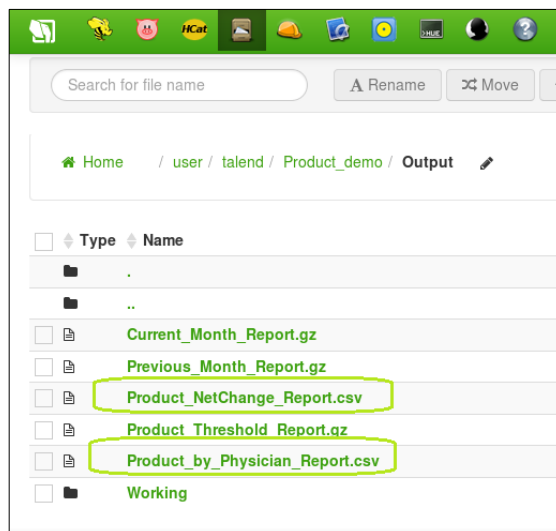
The quickest and easiest way to see the value of the ETL Off-loading scenario is to execute the Single-Click job in Talend Studio. This job shows how the individual steps of the process can be combined into a single execution plan of individual Standard and MapReduce jobs for a "single-click" execution stream.

Execute Standard Job DW_ETL_OffLoad_Scenario / ProductAnalysis_MapReduce_Demo



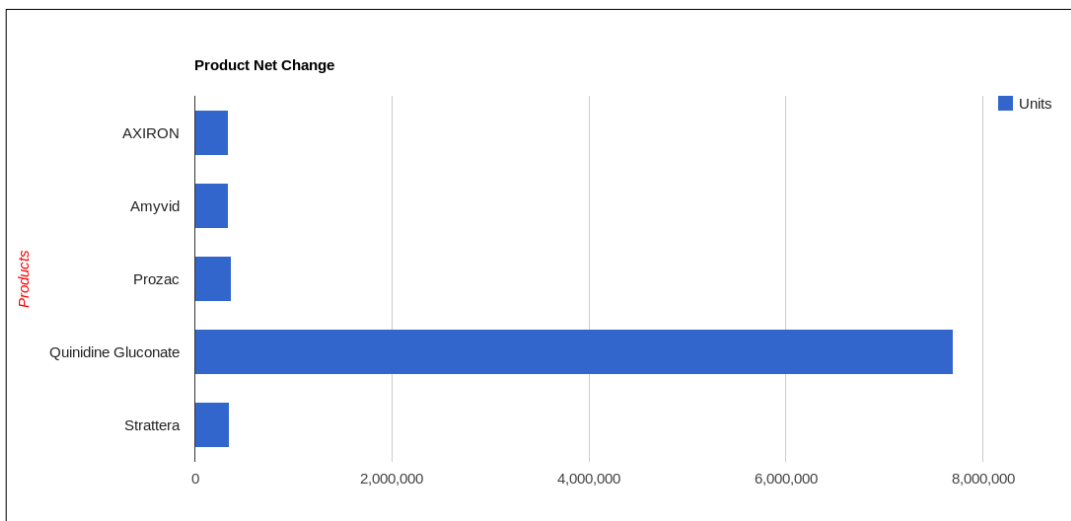
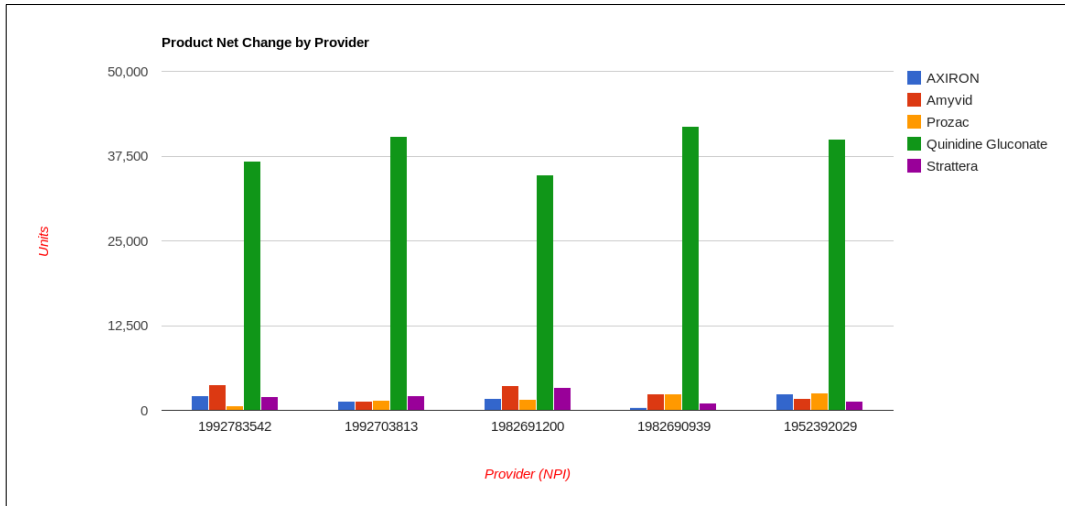
The raw report files will be on HDFS:

```
/user/talend/Product_demo/Output
```



Additionally, the graphical representations of the raw report files have been saved to the following location on the local VM:

/home/talend/Documents/Product_Analysis/Reports



With the Single-Click job all the compressed data files were moved to HDFS, aggregated using MapReduce and compared to the previous months aggregate file. Finally, reports were generated and formatted according to the Google Charts API and saved to the local VM for viewing within a web browser.

6.3.2 Step-by-Step Execution

The Step-by-Step execution of the ETL Off-loading scenario produces the same results as the Single-Click execution but offers deeper insight into the simplicity of using Talend Big Data connected to a partner Hadoop distribution.

To ensure the demo environment is clean we must first run the Standard Job in DW_ETL_OffLoad_Scenario / PreRequirements:

PreStep_3_HDFS_File_Cleanup

This job resets the demo environment and cleans up the directories from any previous execution. It should be run between every execution of this scenario.

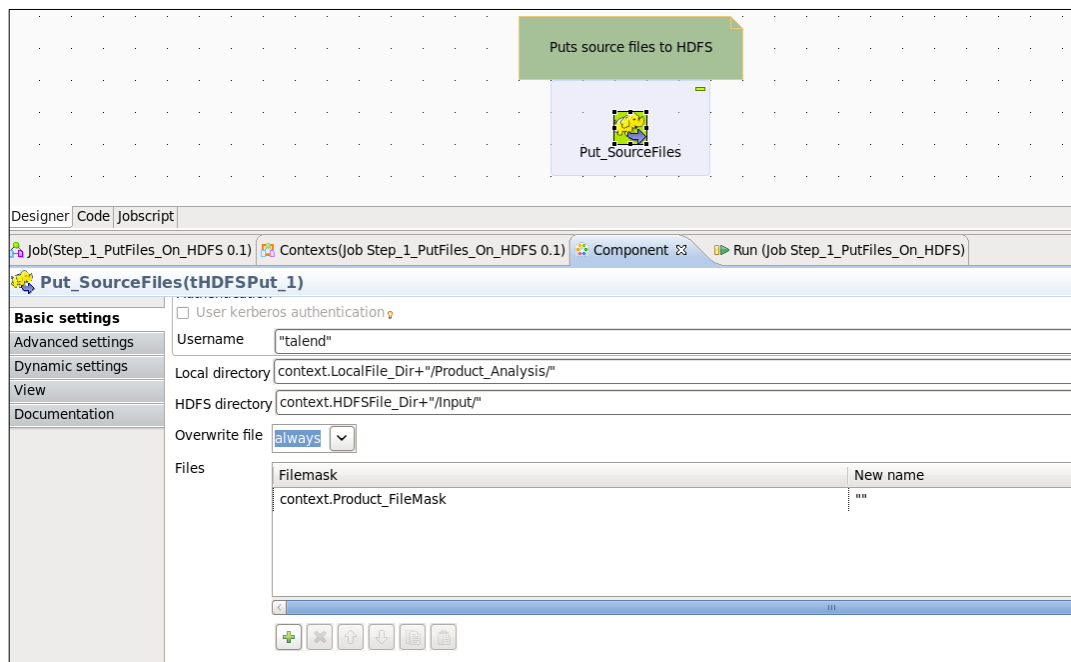
Now let's run the Step-by-Step Execution.

In Standard Jobs DW_ETL_OffLoad_Scenario / ProductAnalysis_Step_by_Step execute:

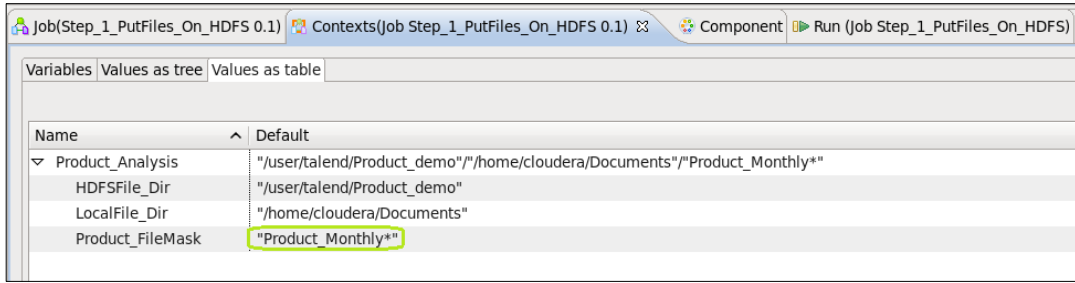
Step_1_PutFiles_On_HDFS

This simple, one-component job moves all compressed, raw data files to the distributed file system.

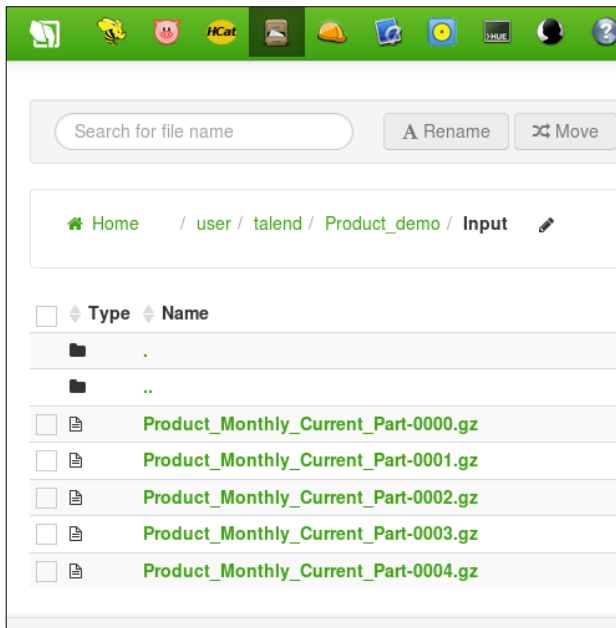
In configuring this component, we identify the source directory of the input files as well as the target HDFS directory. Finally we specify the files to be loaded to HDFS. In this case we are making use of Talend's context variables for consistency and reference within other jobs throughout the scenario.



Talend allows the flexibility of using a standard wild card in the filemask specification which enables the job to select all files at once to load to HDFS without the need of generating multiple iterations of the same job.



The result of this particular job is the files matching the filemask description and residing in the identified local directory are transferred to HDFS in the specified location.

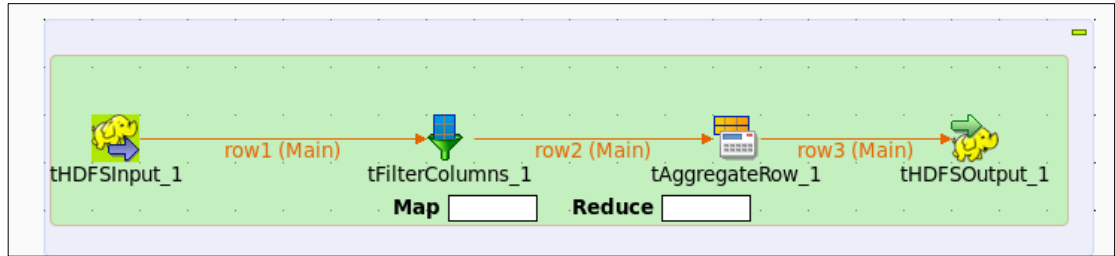


Next, in the MapReduce Jobs under the Product_DW_ETL_OffLoad_Scenario folder, execute:

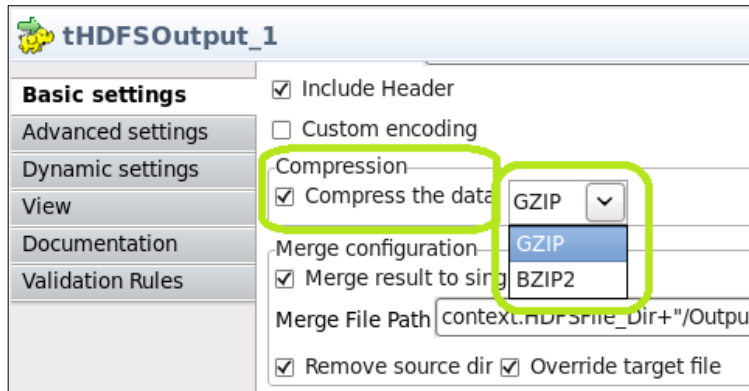
Step_2_Generate_MonthlyReport_mr

This basic but powerful job takes all the compressed input files just loaded to HDFS and with the power of Hadoop and MapReduce, aggregates the massive amount of data, thereby condensing it into something more manageable for QA analysis.

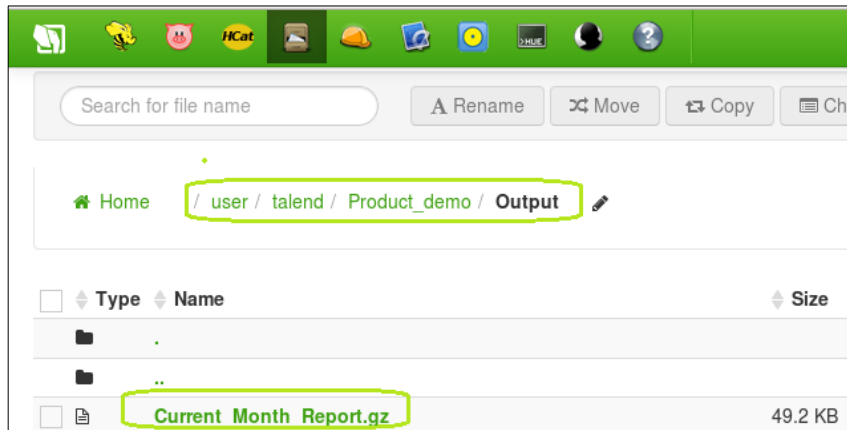
By specifying a folder in the tHDFSInput component, Hadoop will process every file within the folder – compressed files, uncompressed files, or a mixture of both types.



The tHDFSOutput component also allows you to specify whether to compress the output. Talend currently supports two types of compression – GZIP and BZIP2.



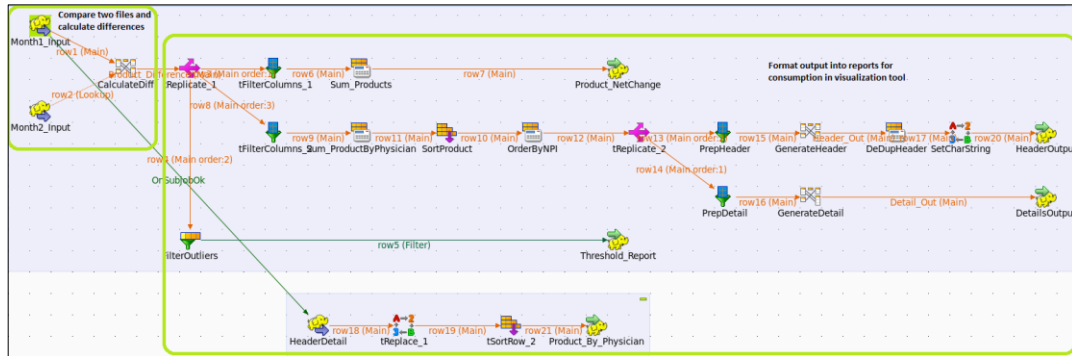
At this point in the process, you have just generated the Current_Month_Report.gz file to go along with the Previous_Month_Report.gz file. These files will be compared in the next step of the process. Notice both files are still compressed.



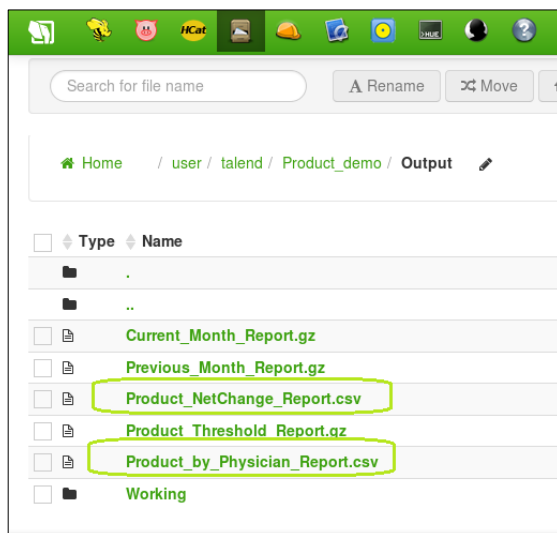
Now run the final two jobs located in Standard Jobs under DW_ETL_OffLoad_Scenario / ProductAnalysis_Step-by_Step:

- Step_3_Month_Over_Month_Comparison**
- Step_4_GoogleChart_Product_by_Unit**

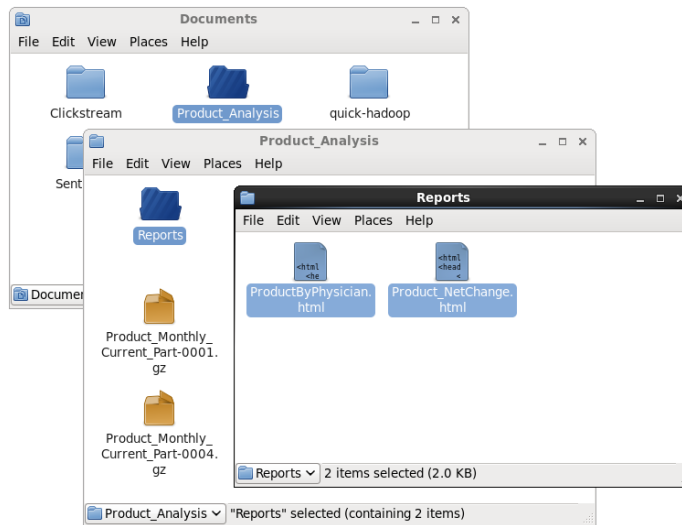
Step_3_Month_Over_Month_Comparison will compare the current month aggregate data with the previous month's aggregate data and format the output into .csv reports that can be federated to a visualization tool of choice.



Output of Step 3 will contain the two .csv report files along with a third compressed file that could be analyzed further with additional Talend and MapReduce jobs.



Step_4_GoogleChart_Product_by_Unit uses the .csv files from Step 3 and integrates them into the Google Charts API for easy viewing within a web page. You can find the files on the local VM as standard HTML documents that can be opened within any web browser.



6.3.3 Extended Scenario Functionality

The ETL Off-loading scenario also allows further exploration of the power of Talend Big Data with Hadoop and MapReduce by allowing Sandbox users to generate their own data sets.

As always, before each execution of the ETL Off-loading scenario, users must execute the following Standard Job in the DW_ETL_OffLoad_Scenario / PreRequirements folder:

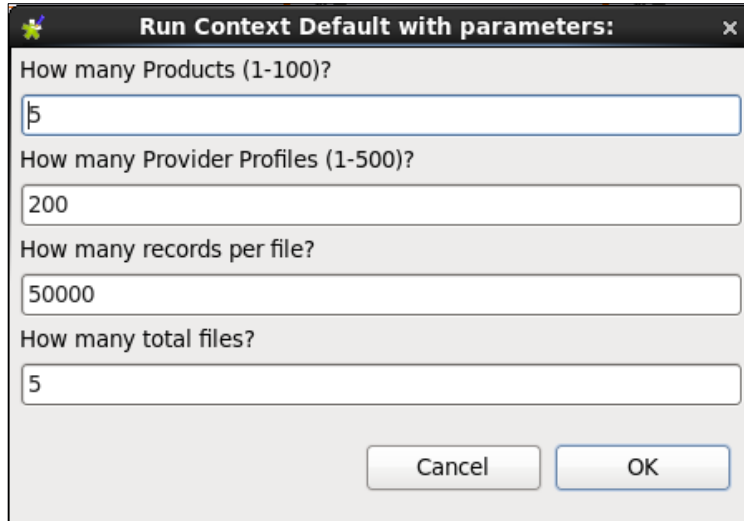
PreStep_3_HDFS_File_Cleanup

Once this job is complete, users can explore the process of generating new data sets to run through the scenario. To generate custom data sets, execute the Standard Job in the DW_ETL_OffLoad_Scenario / PreRequirements folder:

PreStep_1_Generate_Mock_Rx_Data

This is the main job used to generate multiple sets of files for the previous and current months as used in the scenario.

When executing this job, the user will be prompted for input to determine the size of the data set and how many files to process. Default values are provided but can be modified within the guidelines of the job.



Run Context Default with parameters:

How many Products (1-100)?
5

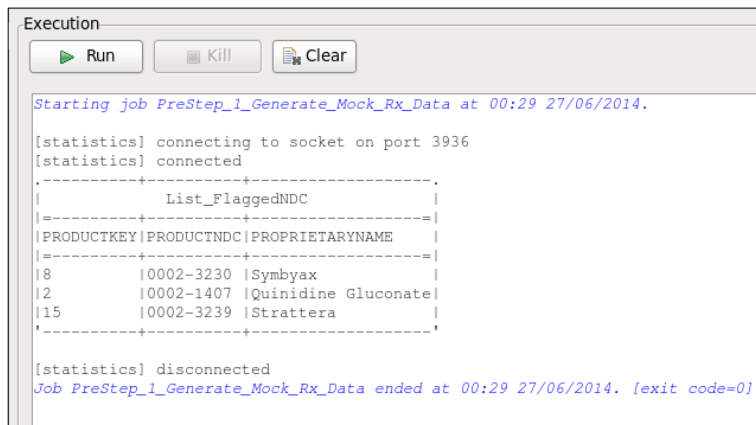
How many Provider Profiles (1-500)?
200

How many records per file?
50000

How many total files?
5

Cancel OK

When the job is complete, the new data files will reside in the `.../Documents/Product_Analysis/Staging_Data/` directory. Additionally, the Execution Output of the job will identify the specific drug(s) that will stand out as inaccurate within the data reports.



```

Execution
[Run] [Kill] [Clear]

Starting job PreStep_1_Generate_Mock_Rx_Data at 00:29 27/06/2014.

[statistics] connecting to socket on port 3936
[statistics] connected

-----+-----
|                               |
|-----+-----+-----|
|PRODUCTKEY|PRODUCTNDC|PROPRIETARYNAME|
|-----+-----+-----|
|8          |0002-3230 |Symbiyax        |
|2          |0002-1407 |Quinidine Gluconate|
|15         |0002-3239 |Strattera       |
|-----+-----+-----|

[statistics] disconnected
Job PreStep_1_Generate_Mock_Rx_Data ended at 00:29 27/06/2014. [exit code=0]

```

Make note of these to ensure the results match your expectations.

To initialize the environment with the newly generated data files, execute the Standard Job found in `DW_ETL_OffLoad_Scenario / PreRequirements`:

PreStep_2_PrepEnvironment

This job compresses the newly generated data files and establishes the initial comparison file. Further, this job will clean up any data from previous runs.

When the two jobs have completed, you are now ready to complete the ETL Off-loading demo using either the Single-Click method or Step-by-Step method as outlined above.

7 Demo: NoSQL Databases

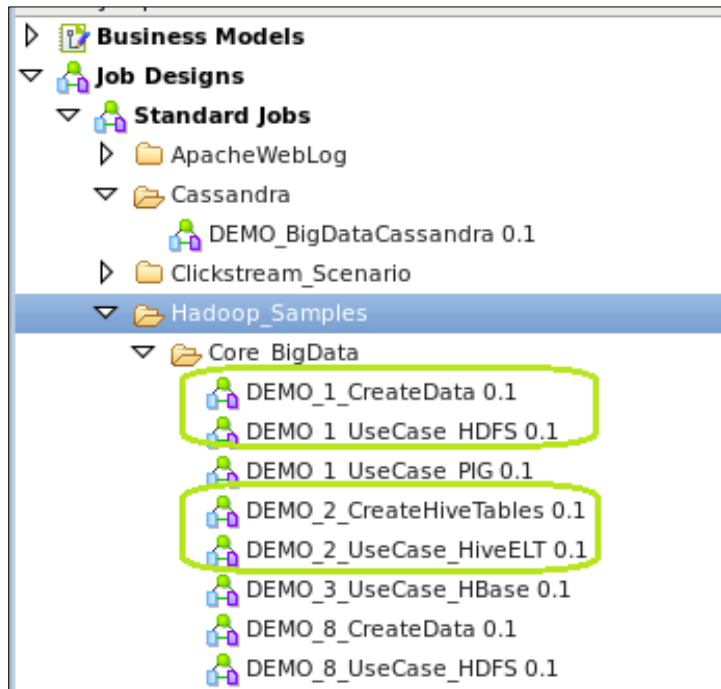
In the Big Data Sandbox environment there are some simple examples on how to use a few of the NoSQL databases available today. Here we will show you how Talend can be used to read and write to HBase, Cassandra, MongoDB and Hive. Within the Hive example we also demonstrate a simple Extract Load and Transform (ELT) example to show a push-down type of process using Talend on Hive.

7.1 Hadoop Core – Hive and HBase

7.1.1 Hive ELT

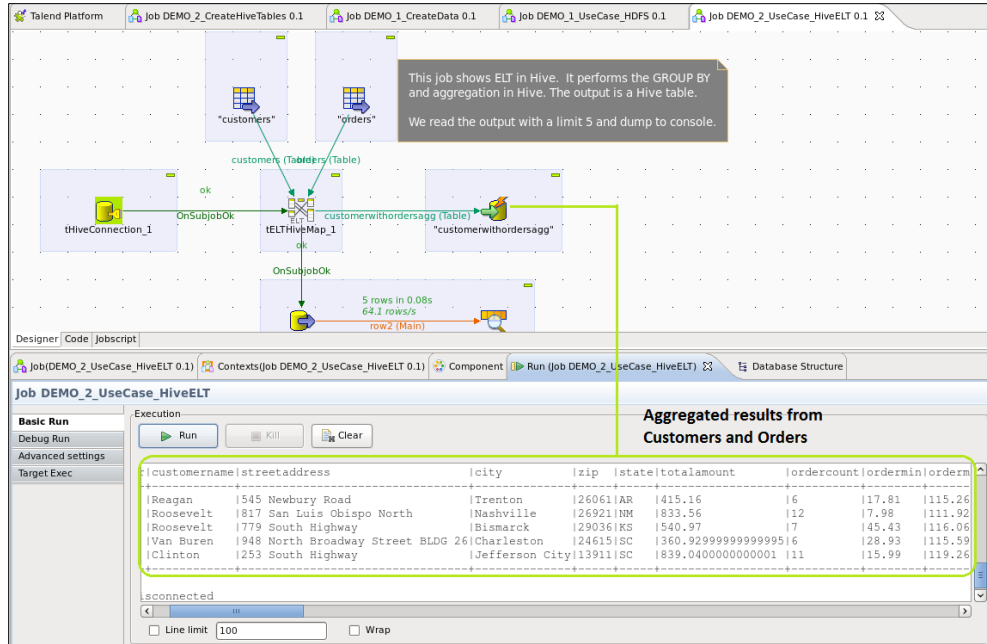
First we will start out with a simple Hive ELT example. Prior to running the ELT process on Hive you need to set up the data by running the following jobs in the order listed:

DEMO_1_CreateData
DEMO_1_UseCase_HDFS
Demo_2_CreateHiveTables



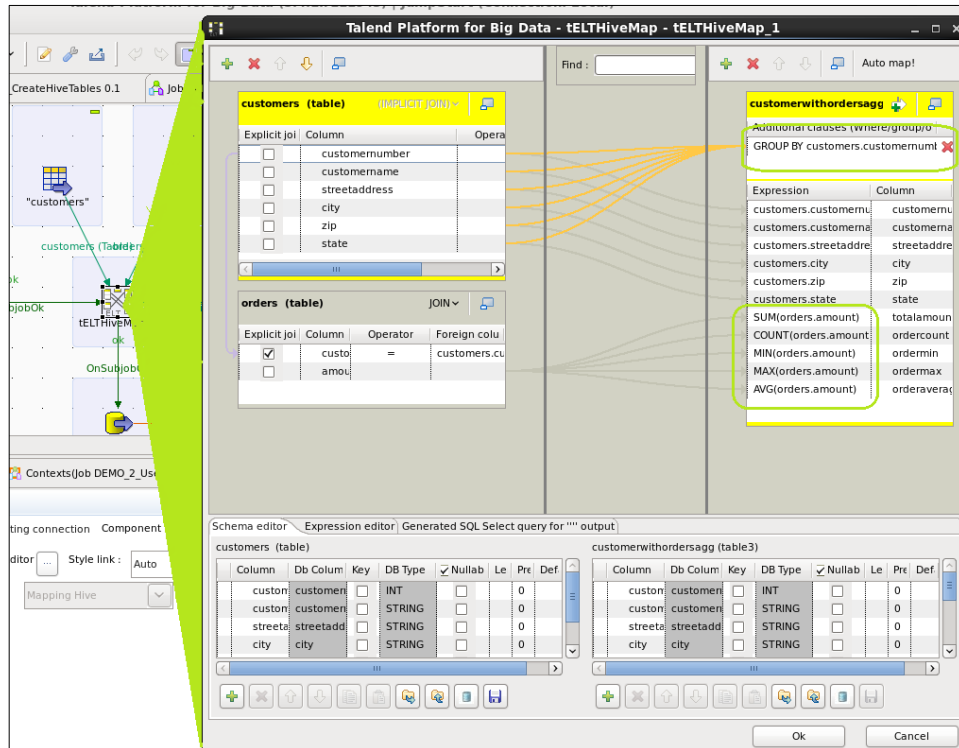
These are great examples of creating and loading data to Hive.

DEMO_2_UseCase_HiveELT



This shows an example of how you can build Hive ELT processing that will take advantage of all the Hive processing power without the data leaving the Hadoop/Hive environment.

In this example the tables "Customer" and "Orders" are being joined and values from the orders table are being computed and saved in the Hive table "customerwithordersagg". Examples of the computed values are the total amount of all orders, the number of orders, and the min and max order per customer.



If you have used any of the ELT functions on the other Talend components, e.g. Oracle or MySQL, you will see that Hive works very similar to those other RDBMS ELT components.

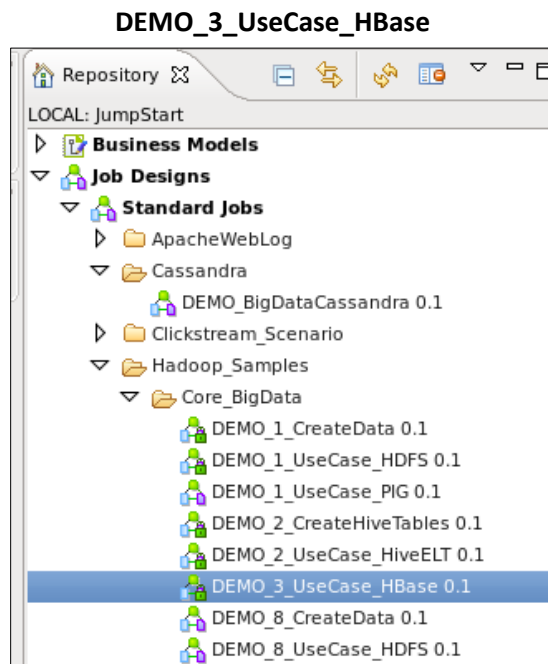
7.1.2 HBase

HBase is a non-relational, distributed database modeled after Google's BigTable and is good at storing sparse data. HBase is considered a key-value columnar database and it runs on top of HDFS. It is used mostly when you need random, real-time read/write access.

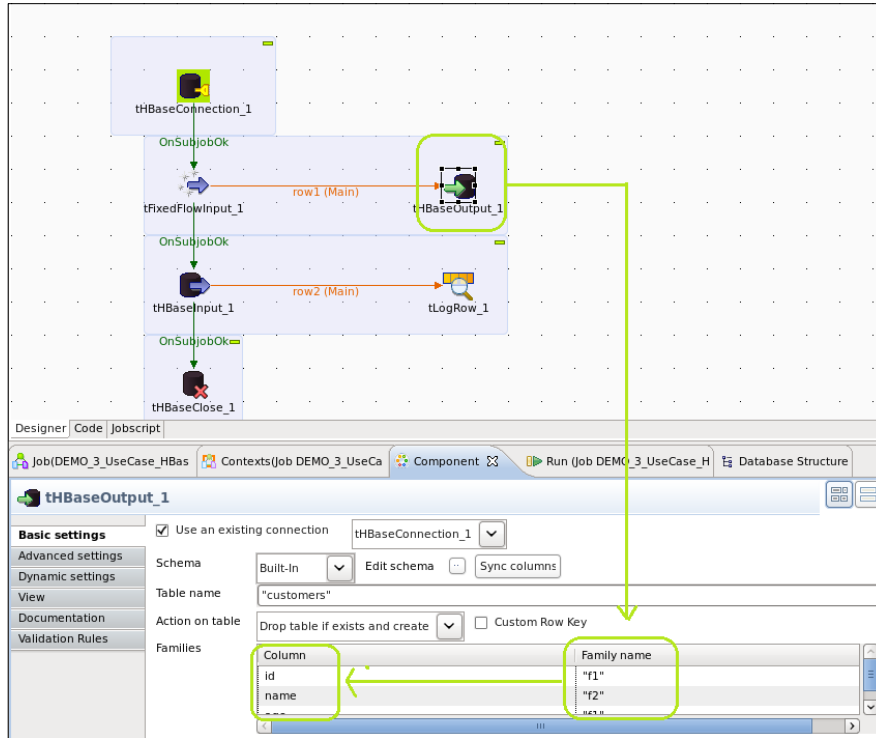
The goal of HBase is to handle billions of rows * millions of columns. If your relational table looks like below (data missing in columns), it is considered "sparse" and a good candidate for HBase.

	Col A	Col B	Col C	Col D	Col E
Row 01	Val1A				
Row 02	Val2A	Val2B	Val2C	Val2D	Val2E
Row 03	Val3A		Val3C		Val3E

In the Big Data Sandbox you will see the following example of loading and reading from an HBase database:

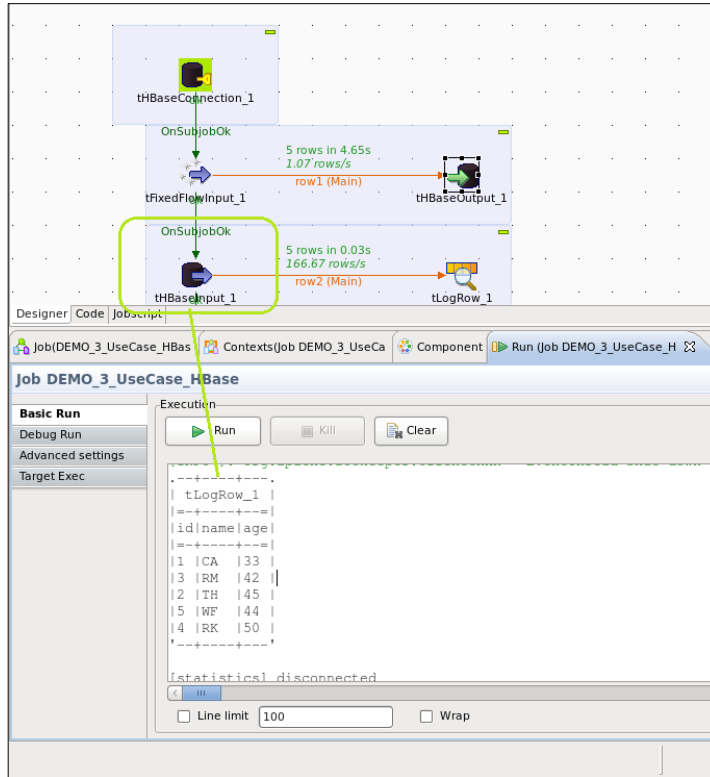


The HBase example shows how to setup the column families and load data to the database as well as read the data back out:



The Columns need to be assigned to a Family name as shown above, "F1" and "F2". The Family names are defined on the "Advanced settings" tab of the `tHBaseOutput_1` component.

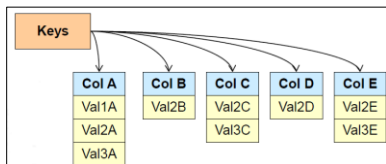
The data is loaded into the HBase database on the Big Data Sandbox VM and then read back based on the query in the `tHBaseInput` component.



The advantage of using Talend for NoSQL databases like HBase is that Talend Studio gives you a consistent and easy way to interact with all the databases. You need to understand the NoSQL database you are working with, but then Talend makes it easy to achieve the core functions like creating tables, loading data and reading results.

7.2 Cassandra

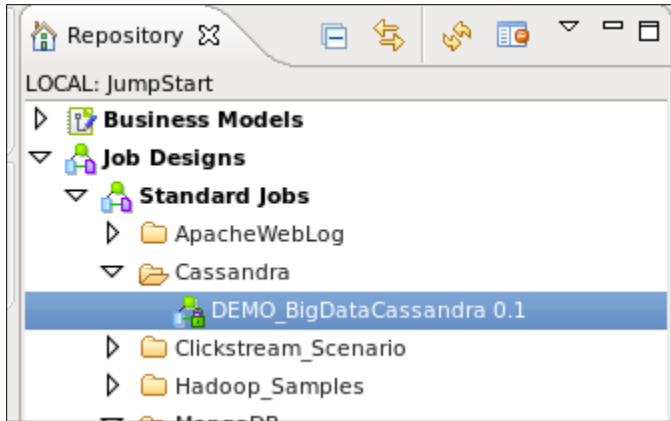
Cassandra is an Apache distributed database designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure. Column storage such as Cassandra, stores data tables as sections of columns of data rather than as rows of data. This is good for finding or aggregating large sets of similar data. Column storage serializes all data for one column contiguous on disk (resulting in very quick read of a column). Organization of your data REALLY matters in columnar storage.



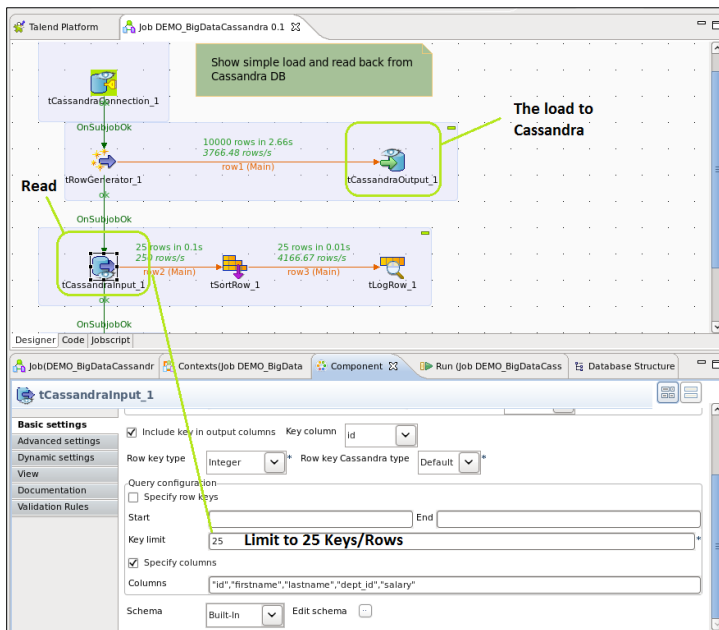
There are no restrictions on number of columns. One row in relational may be many rows in columnar.

The Sandbox VM has Cassandra installed and configured to allow you to see how Talend can load and manage data with Cassandra. In the Talend Studio you will find a simple example in the Jumpstart project. In the Standard Jobs /Cassandra folder:

DEMO_BigDataCassandra



This process generates a sample of 10k employee records and loads that data into a new column family in a Cassandra store.



The last step is to read back the data and display the first 25 records from the database.

7.3 MongoDB

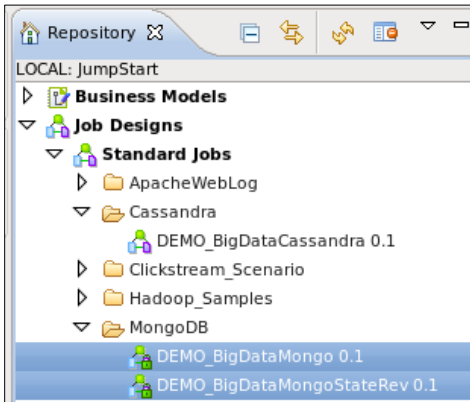
MongoDB is best used as a document storage database. MongoDB stores documents that encapsulate and encode data in some standard format (including XML, YAML, and JSON as well as

binary forms like BSON, PDF and Microsoft Office documents). Different implementations offer different ways of organizing and/or grouping documents.

Documents are addressed in the database via a unique key that represents that document. The big feature is the database offers an API or query language that allows retrieval of documents based on their contents. Below is an example of how these databases store the content.

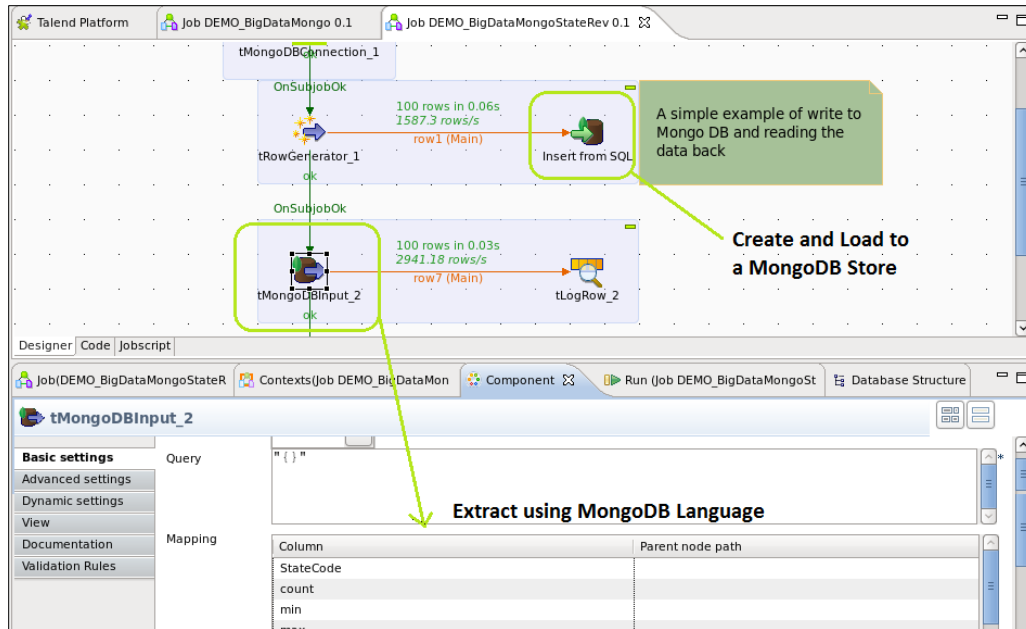
The Sandbox VM has MongoDB installed and configured for demonstration purposes. In the Talend Studio you will find a simple example in the Jumpstart project. In the Standard Jobs/ MongoDB folder:

DEMO_BigDataMongoStateRev
DEMO_BigDataMongo



The first example is a simple process that generates a list of US states and revenues and loads into a MongoDB table. Then the following step demonstrates how to extract the data from MongoDB.

DEMO_BigDataMongoStateRev



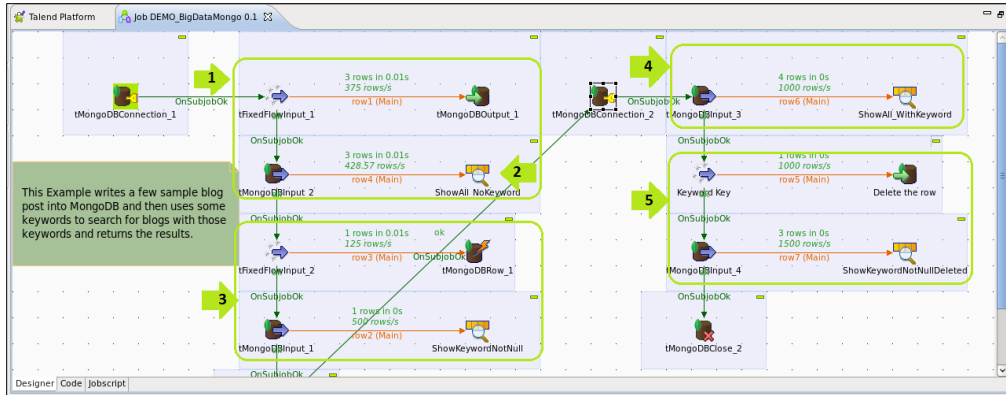
In this example you can see a simple process that creates and loads data to MongoDB and then reads the states revenue back to the Talend Studio console.

The second example has a more detailed process of how to use MongoDB, how flexible the schema is, and how Talend handles the “schema on read” and “schema on write”. This example is writing blog posts with titles and then later keywords as columns/keys.

1. The first step writes blog posts with only 2 keys “Title” and “Content”.
2. Then it will read that data with that schema and display on the console.
3. Next the job will add a new record with three keys, “Title”, “Keywords” and “Content”.
4. Then there will be a series of reads showing different queries with “keyword” and without the “Keyword”
5. Finally it shows deleting any records with a null “Keyword”

The following job demonstrates this example:

DEMO_BigDataMongo



In this first sample, blogs are loaded to the MongoDB with a schema of just “title” and “content”:

title	content
How to crack a safe	In this blog post we will discuss manipulation of an group 2 S&G combination lock...
Sugru Form & Fix	Surgu is a new silicon based putty that hardens but remains flexible....
Innova vs Discraft	Disc golf showdown between the two premier makers of flying discs....

Then a new record is added but it also adds another key called “keyword”. Now a total of 4 records exist in the table with some records having the “keyword” attribute while others do not.

ShowAll_WithKeyword		
title	keyword	content
Fintails Forever	Heckflosse	Mercedes 190 and 200 series from the 1960s...
How to crack a safe	null	In this blog post we will discuss manipulation of an group 2 S&G combination lock...
Innova vs Discraft	null	Disc golf showdown between the two premier makers of flying discs....
Sugru Form & Fix	null	Surgu is a new silicon based putty that hardens but remains flexible....

This is another example of how Talend can help take the complexity away from all the different technologies and help you become big data proficient. By leveraging Talend’s ability to handle complex data types like XML and JSON and combining it with NoSQL database technologies like MongoDB, your integration experts will quickly begin providing you the big value from your big data initiatives.

8 Conclusion

With all the different big data technologies and Hadoop platforms/projects that are available you will find that there are many ways to address your big data opportunities. Our research has found that companies need to overcome five key hurdles for big data success: obtaining big data skills, identifying big data integration opportunities, building a big data infrastructure (security, scalability, data quality, privacy ...), governing big data processes, and showing success for continued funding.

Talend addresses these challenges with the most advanced big data integration platform, used by data-driven businesses to deliver timely and easy access to all their data. Talend equips IT with an open, native and unified integration solution that unlocks all your data to quickly meet existing and emerging business use cases.

How?

First, with Talend you can leverage in-house resources to use Talend's rich graphical tools that generate big data code (PIG, MapReduce, Java) for you. Talend is based on standards such as Eclipse, Java, and SQL, and is backed by a large collaborative community. So you can up-skill existing resources instead of finding new resources.

Second, Talend is big data ready. Unlike other solutions that bolt on big data, Talend provides native support for Hadoop, MapReduce and NoSQL with over 800 connectors to all data sources. Talend's native Hadoop data quality solution delivers clean and consistent data at infinite scale.

And third, Talend lowers operations costs. Talend's zero footprint solution takes the complexity out of integration deployment, management, and maintenance. And a usage based subscription model provides a fast return on investment without large upfront costs.

9 Next Steps

We hope that this set of projects has given you a better understanding of how you can start addressing your big data opportunities using Talend. Being a new technology, big data has many challenges – Talend can help. We provide a broad set of integration products and services to quickly ramp up your team, including big data assessments, big data training and support. An appropriate next step would be to discuss with your Talend sales representative your specific requirements and how Talend can help “Jumpstart” your big data project into production.