OVHSUMMIT

TOGETHER

Lightning Talks

#OVHSummit

OVH.com

Découvrez le Machine Learning appliqués aux logs web avec Dataiku DSS

(11 octobre 2016)

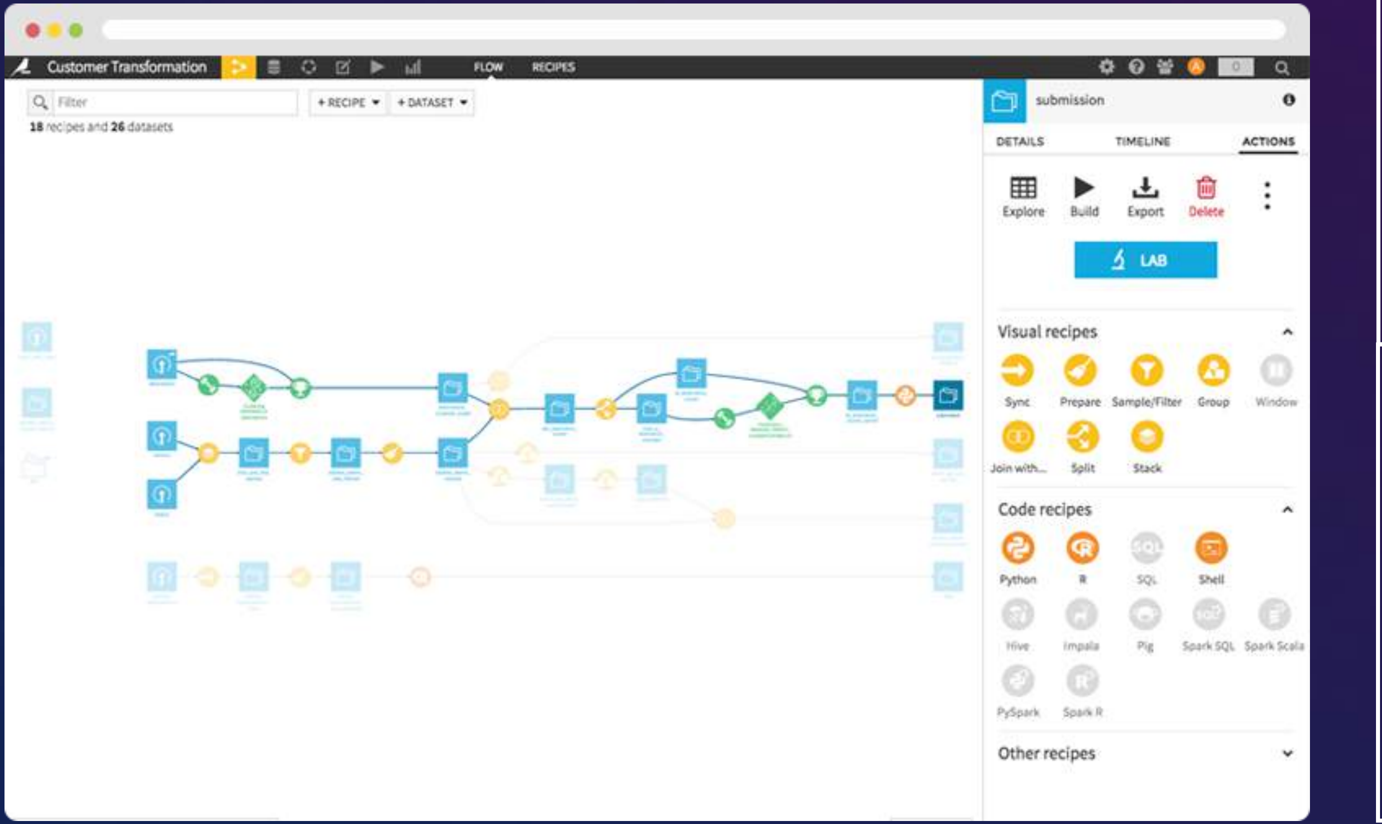
Jérémy Grèze

Data Analyst

Dataiku

Dataiku DSS

A Collaborative Data Science Platform



01010
1100

PRÉPARER

Chargez et préparez vos données

ANALYSER

Visualisez et partagez vos découvertes

MODÉLISER

Construisez vos modèles

AUTOMATISER

Ré-exécutez tous les jours…

EXECUTER

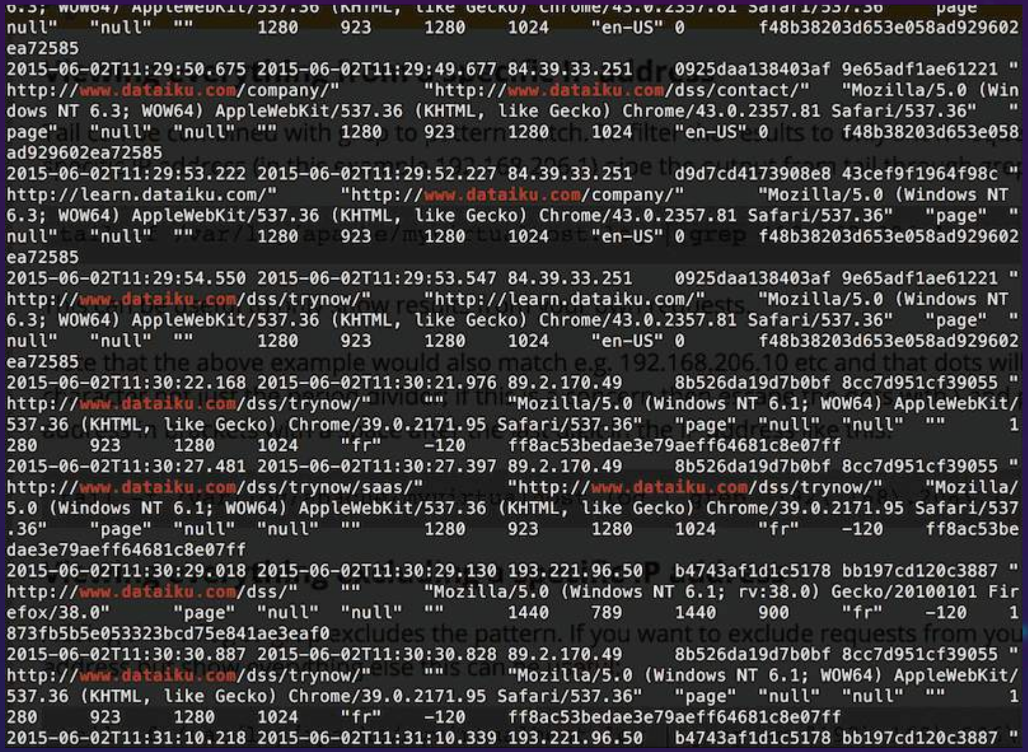
… ou en Temps Réel

MONITORER

Suivez votre production

Logs webs

- Une ligne par action sur un site web
- Des informations typiques:
  - Date (timezone?)
  - IP
  - Page/ressource
  - Identifiant?
  - …
- Deux “catégories”:
  - Logs serveurs
  - Logs Javascript

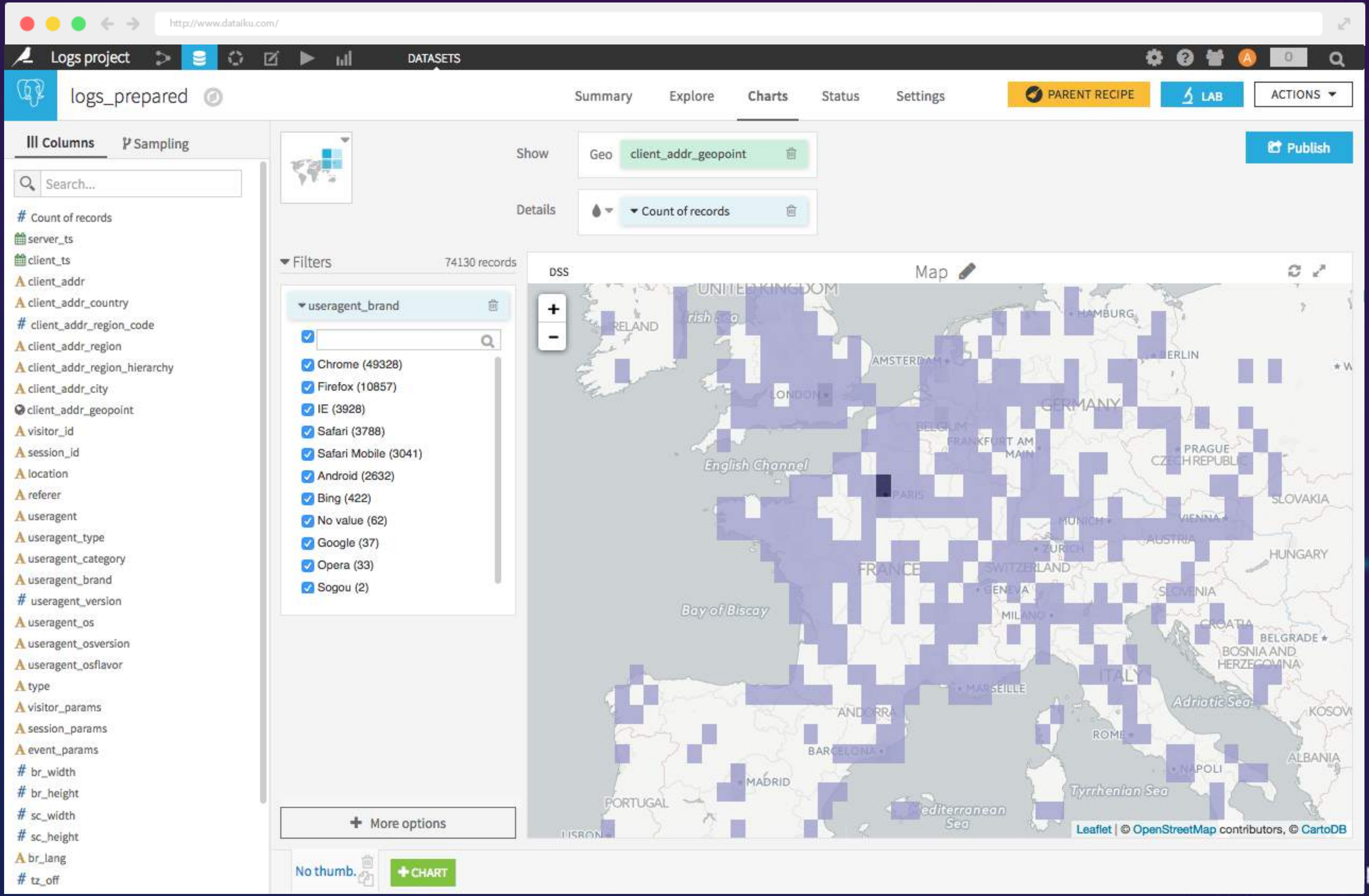


```
6.3; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/43.0.2357.81 Safari/537.36"    page
null"   "null"  ""      1280    923     1280    1024    "en-US" 0       f48b38203d653e058ad929602
ea72585
2015-06-02T11:29:50.675 2015-06-02T11:29:49.677 84.39.33.251    0925daa138403af 9e65adf1ae61221 "
http://www.dataiku.com/company/"        "http://www.dataiku.com/dss/contact/"   "Mozilla/5.0 (Win
dows NT 6.3; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/43.0.2357.81 Safari/537.36"   "
page"   "null"  "null"  ""      1280    923     1280    1024    "en-US" 0       f48b38203d653e058
ad929602ea72585
2015-06-02T11:29:53.222 2015-06-02T11:29:52.227 84.39.33.251    d9d7cd4173908e8 43cef9f1964f98c "
http://learn.dataiku.com/"      "http://www.dataiku.com/company/"       "Mozilla/5.0 (Windows NT
6.3; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/43.0.2357.81 Safari/537.36"   "page"  "
null"   "null"  ""      1280    923     1280    1024    "en-US" 0       f48b38203d653e058ad929602
ea72585
2015-06-02T11:29:54.550 2015-06-02T11:29:53.547 84.39.33.251    0925daa138403af 9e65adf1ae61221 "
http://www.dataiku.com/dss/trynow/"     "http://learn.dataiku.com/"     "Mozilla/5.0 (Windows NT
6.3; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/43.0.2357.81 Safari/537.36"   "page"  "
null"   "null"  ""      1280    923     1280    1024    "en-US" 0       f48b38203d653e058ad929602
ea72585
2015-06-02T11:30:22.168 2015-06-02T11:30:21.976 89.2.170.49     8b526da19d7b0bf 8cc7d951cf39055 "
http://www.dataiku.com/dss/trynow/"     ""      "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/
537.36 (KHTML, like Gecko) Chrome/39.0.2171.95 Safari/537.36"   "page"  "null"  "null"  ""      1
280     923     1280    1024    "fr"    -120    ff8ac53bedae3e79aeff64681c8e07ff
2015-06-02T11:30:27.481 2015-06-02T11:30:27.397 89.2.170.49     8b526da19d7b0bf 8cc7d951cf39055 "
http://www.dataiku.com/dss/trynow/saas/"        "http://www.dataiku.com/dss/trynow/"    "Mozilla/
5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/39.0.2171.95 Safari/537
.36"    "page"  "null"  "null"  ""      1280    923     1280    1024    "fr"    -120    ff8ac53be
dae3e79aeff64681c8e07ff
2015-06-02T11:30:29.018 2015-06-02T11:30:29.130 193.221.96.50   b4743af1d1c5178 bb197cd120c3887 "
http://www.dataiku.com/dss/"    ""      "Mozilla/5.0 (Windows NT 6.1; rv:38.0) Gecko/20100101 Fir
efox/38.0"      "page"  "null"  "null"  ""      1440    789     1440    900     "fr"    -120    1
873fb5b5e053323bcd75e841ae3eaf0
2015-06-02T11:30:30.887 2015-06-02T11:30:30.828 89.2.170.49     8b526da19d7b0bf 8cc7d951cf39055 "
http://www.dataiku.com/dss/trynow/"     ""      "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/
537.36 (KHTML, like Gecko) Chrome/39.0.2171.95 Safari/537.36"   "page"  "null"  "null"  ""      1
280     923     1280    1024    "fr"    -120    ff8ac53bedae3e79aeff64681c8e07ff
2015-06-02T11:31:10.218 2015-06-02T11:31:10.339 193.221.96.50   b4743af1d1c5178 bb197cd120c3887 "
```

Data cleaning (dates, IPs…)

Exemples:

- Parsing dates
- Géolocalisation IP
- Nettoyage des valeurs
- Extraire composants d'URL
- Enrichissement user-agent
- Filtrage
- Catégorisation



- # Count of records
- server_ts
- client_ts
- A client_addr
- A client_addr_country
- # client_addr_region_code
- A client_addr_region
- A client_addr_region_hierarchy
- A client_addr_city
- client_addr_geopoint
- A visitor_id
- A session_id
- A location
- A referer
- A useragent
- A useragent_type
- A useragent_category
- A useragent_brand
- # useragent_version
- A useragent_os
- A useragent_osversion
- A useragent_osflavor
- A type
- A visitor_params
- A session_params
- A event_params
- # br_width
- # br_height
- # sc_width
- # sc_height
- A br_lang
- # tz_off

useragent_brand

- ☑ Chrome (49328)
- ☑ Firefox (10857)
- ☑ IE (3928)
- ☑ Safari (3788)
- ☑ Safari Mobile (3041)
- ☑ Android (2632)
- ☑ Bing (422)
- ☑ No value (62)
- ☑ Google (37)
- ☑ Opera (33)
- ☑ Sogou (2)

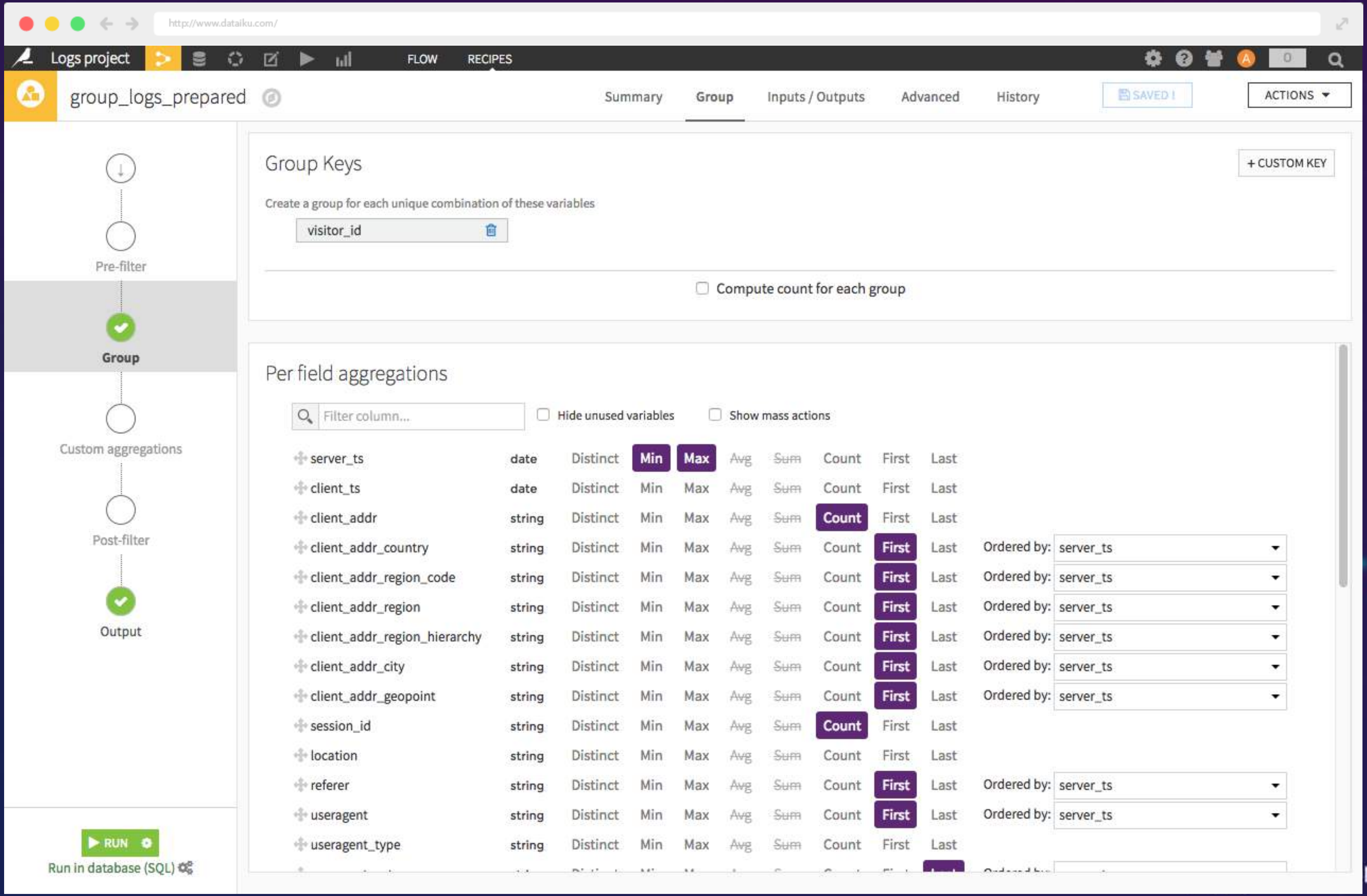
Réduction de dimension (*group by*)

Agrégation au niveau de l’utilisateur:

- Cookie
- Identifiant
- IP
- Hash IP + user-agent

Recette visuelle SQL de *group by*.

Exécution *in-database* possible: MySQL, PostgreSQL, Vertica, Hive (Hadoop ou Spark)…



Group Keys

+ CUSTOM KEY

Create a group for each unique combination of these variables

visitor_id

□ Compute count for each group

Per field aggregations

Filter column...

□ Hide unused variables

□ Show mass actions

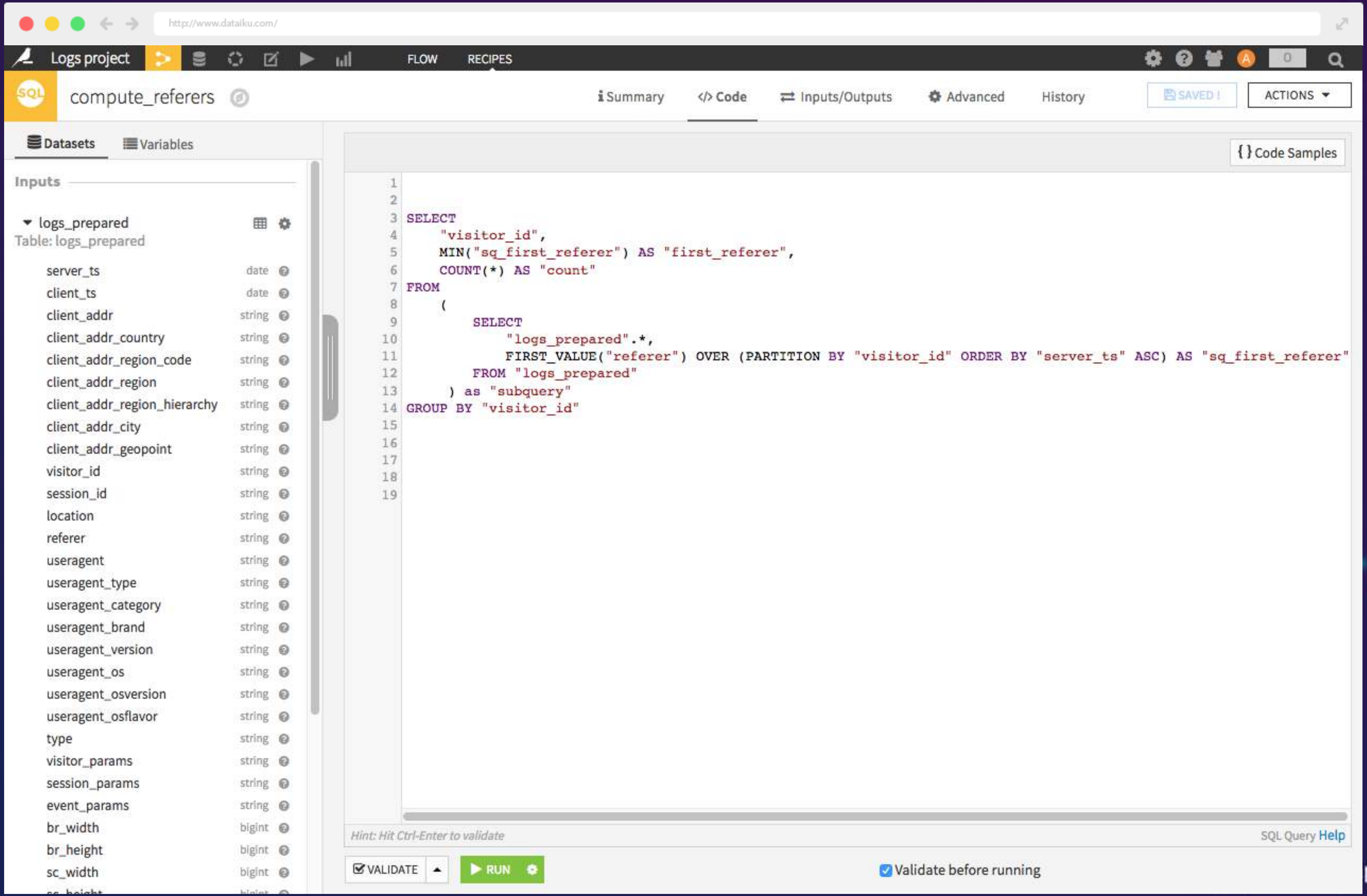
server_ts	date	Distinct	[✓ Min]	[✓ Max]	~~Avg~~	~~Sum~~	Count	First	Last	
client_ts	date	Distinct	Min	Max	~~Avg~~	~~Sum~~	Count	First	Last	
client_addr	string	Distinct	Min	Max	~~Avg~~	~~Sum~~	[✓ Count]	First	Last	
client_addr_country	string	Distinct	Min	Max	~~Avg~~	~~Sum~~	Count	[✓ First]	Last	Ordered by: server_ts
client_addr_region_code	string	Distinct	Min	Max	~~Avg~~	~~Sum~~	Count	[✓ First]	Last	Ordered by: server_ts
client_addr_region	string	Distinct	Min	Max	~~Avg~~	~~Sum~~	Count	[✓ First]	Last	Ordered by: server_ts
client_addr_region_hierarchy	string	Distinct	Min	Max	~~Avg~~	~~Sum~~	Count	[✓ First]	Last	Ordered by: server_ts
client_addr_city	string	Distinct	Min	Max	~~Avg~~	~~Sum~~	Count	[✓ First]	Last	Ordered by: server_ts
client_addr_geopoint	string	Distinct	Min	Max	~~Avg~~	~~Sum~~	Count	[✓ First]	Last	Ordered by: server_ts
session_id	string	Distinct	Min	Max	~~Avg~~	~~Sum~~	[✓ Count]	First	Last	
location	string	Distinct	Min	Max	~~Avg~~	~~Sum~~	Count	First	Last	
referer	string	Distinct	Min	Max	~~Avg~~	~~Sum~~	Count	[✓ First]	Last	Ordered by: server_ts
useragent	string	Distinct	Min	Max	~~Avg~~	~~Sum~~	Count	[✓ First]	Last	Ordered by: server_ts
useragent_type	string	Distinct	Min	Max	~~Avg~~	~~Sum~~	Count	First	Last	

Enrichissement en SQL

Objectif: construire des variables au niveau agrégé (visiteur)

Exemples:

- Récupération du premier *referer* avec une *Window Function*
- Enrichissement avec des données à partir du CRM avec un *Left Join*



logs_prepared
Table: logs_prepared

server_ts	date
client_ts	date
client_addr	string
client_addr_country	string
client_addr_region_code	string
client_addr_region	string
client_addr_region_hierarchy	string
client_addr_city	string
client_addr_geopoint	string
visitor_id	string
session_id	string
location	string
referer	string
useragent	string
useragent_type	string
useragent_category	string
useragent_brand	string
useragent_version	string
useragent_os	string
useragent_osversion	string
useragent_osflavor	string
type	string
visitor_params	string
session_params	string
event_params	string
br_width	bigint
br_height	bigint
sc_width	bigint

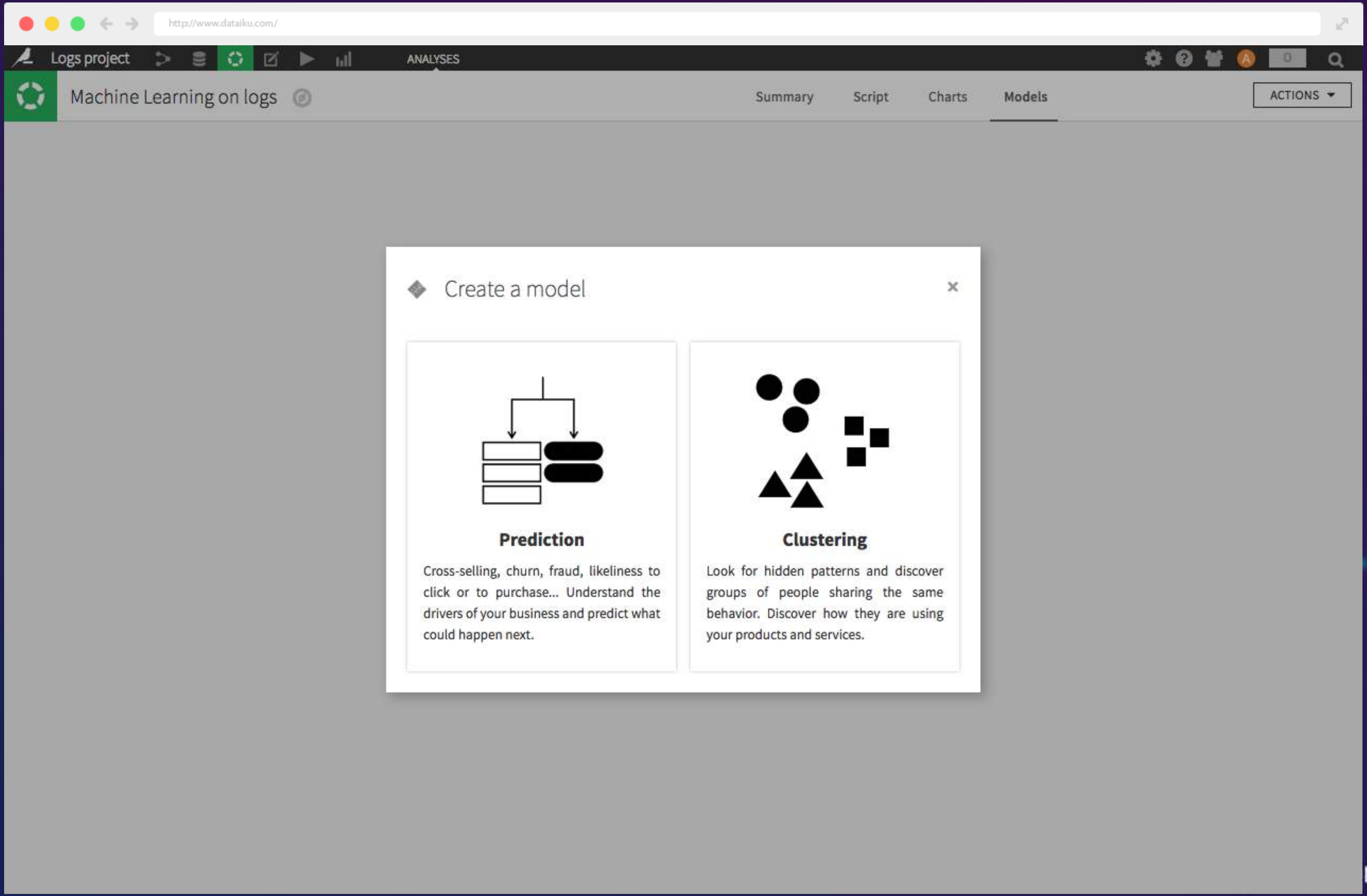
```sql


SELECT
    "visitor_id",
    MIN("sq_first_referer") AS "first_referer",
    COUNT(*) AS "count"
FROM
    (
        SELECT
            "logs_prepared".*,
            FIRST_VALUE("referer") OVER (PARTITION BY "visitor_id" ORDER BY "server_ts" ASC) AS "sq_first_referer"
        FROM "logs_prepared"
     ) as "subquery"
GROUP BY "visitor_id"
```

Prêt pour le Machine Learning ?!

Prérequis: des données agrégées, propres et enrichies

Machine Learning:

- Supervisé: pour prédire une variable cible
- Non-supervisé: pour trouver des groupes similaires (*clustering*)



Prediction

Cross-selling, churn, fraud, likeliness to click or to purchase... Understand the drivers of your business and predict what could happen next.

Clustering

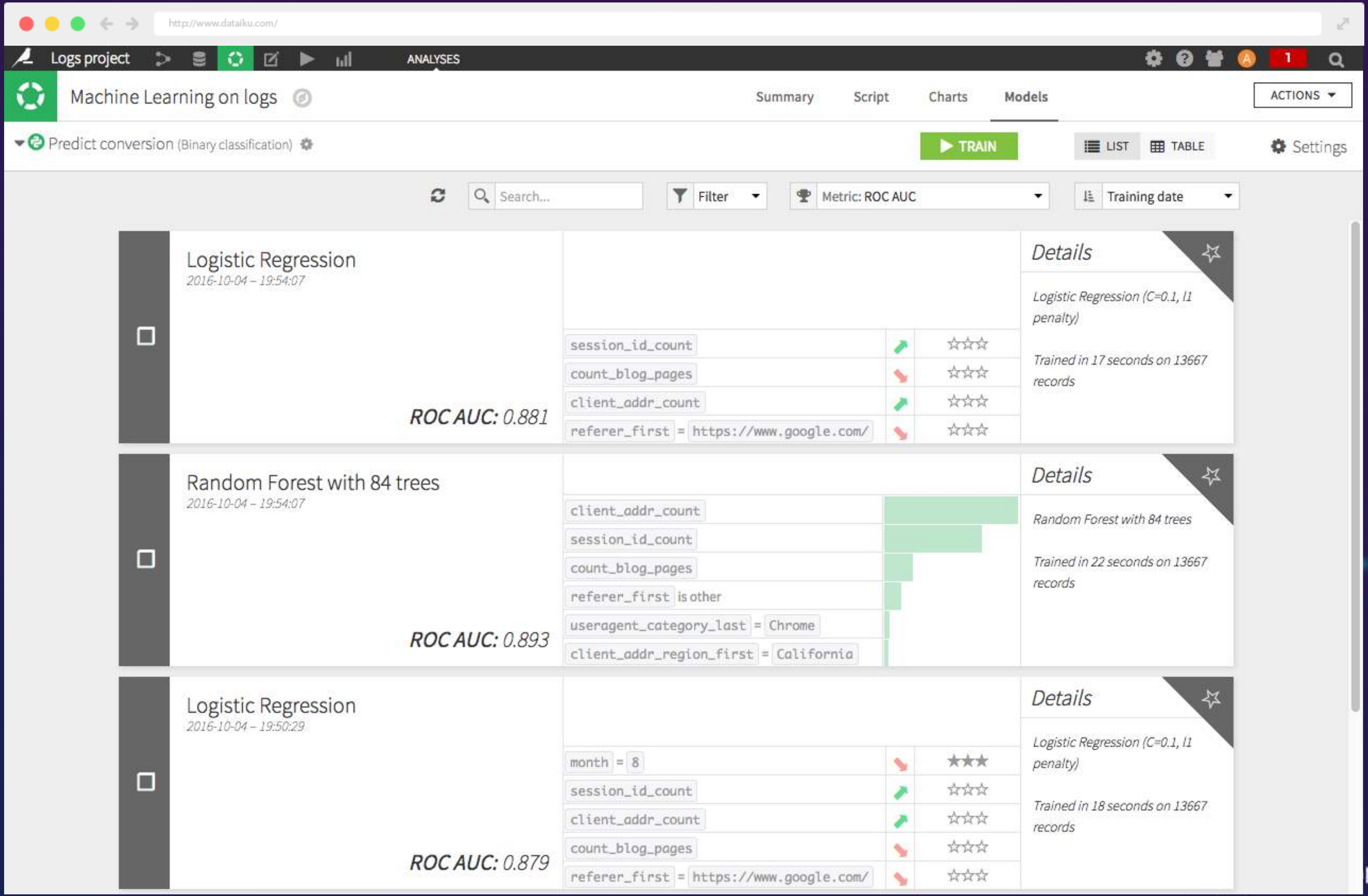
Look for hidden patterns and discover groups of people sharing the same behavior. Discover how they are using your products and services.

Prédiction

Exemple: prédire une conversion

Etapes:

- Choix des variables et types (numériques, catégorielles, …)
- Choix des algorithmes (linéaires, arbres…)
- Phase d'entrainement
- Etude des résultats



Logistic Regression
2016-10-04 – 19:54:07
ROC AUC: 0.881

session_id_count
count_blog_pages
client_addr_count
referer_first = https://www.google.com/

Details
Logistic Regression (C=0.1, l1 penalty)
Trained in 17 seconds on 13667 records

Random Forest with 84 trees
2016-10-04 – 19:54:07
ROC AUC: 0.893

client_addr_count
session_id_count
count_blog_pages
referer_first is other
useragent_category_last = Chrome
client_addr_region_first = California

Details
Random Forest with 84 trees
Trained in 22 seconds on 13667 records

Logistic Regression
2016-10-04 – 19:50:29
ROC AUC: 0.879

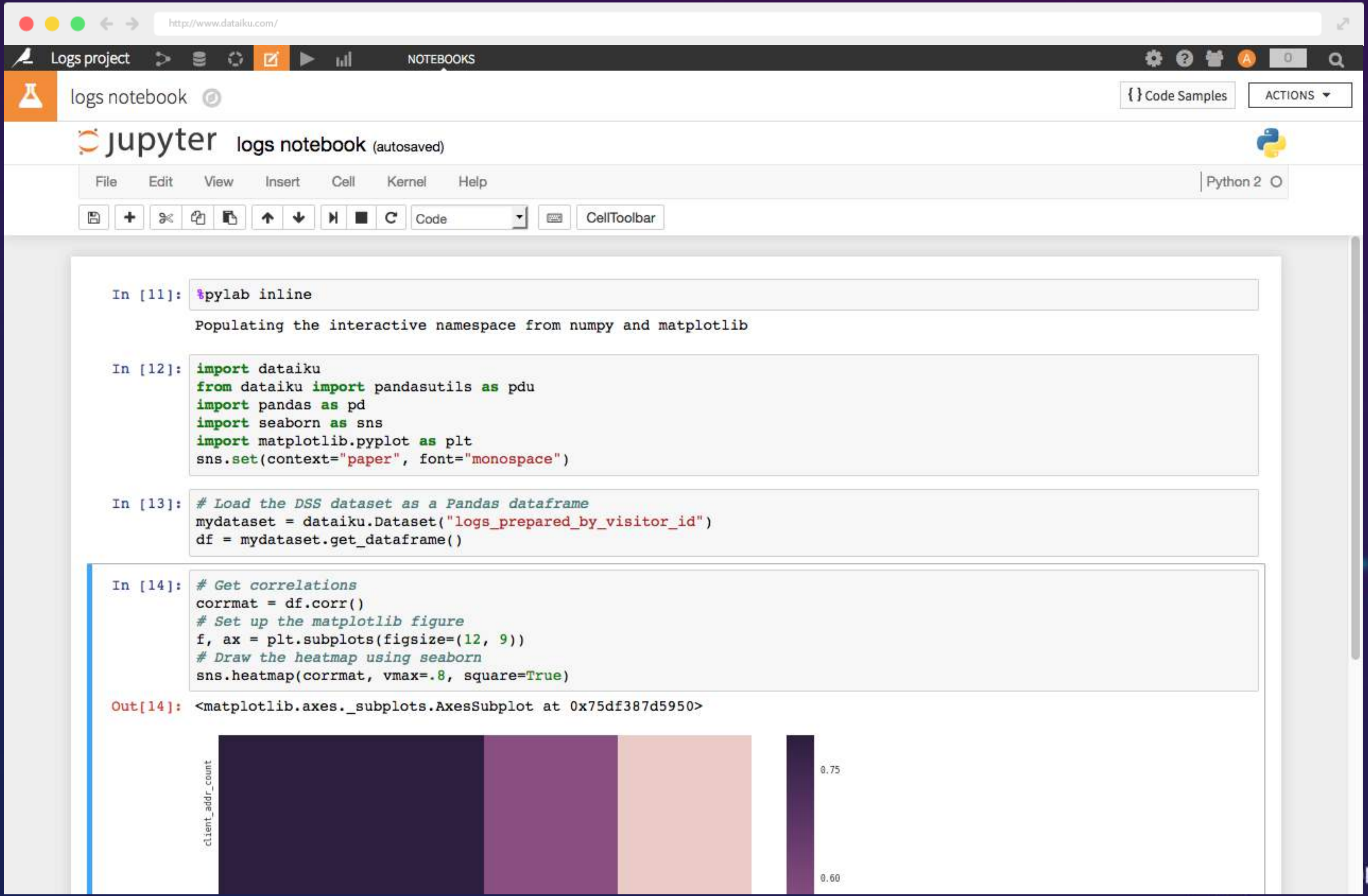
month = 8
session_id_count
client_addr_count
count_blog_pages
referer_first = https://www.google.com/

Details
Logistic Regression (C=0.1, l1 penalty)
Trained in 18 seconds on 13667 records

#OVHSummit

Un peu de Python ou R?

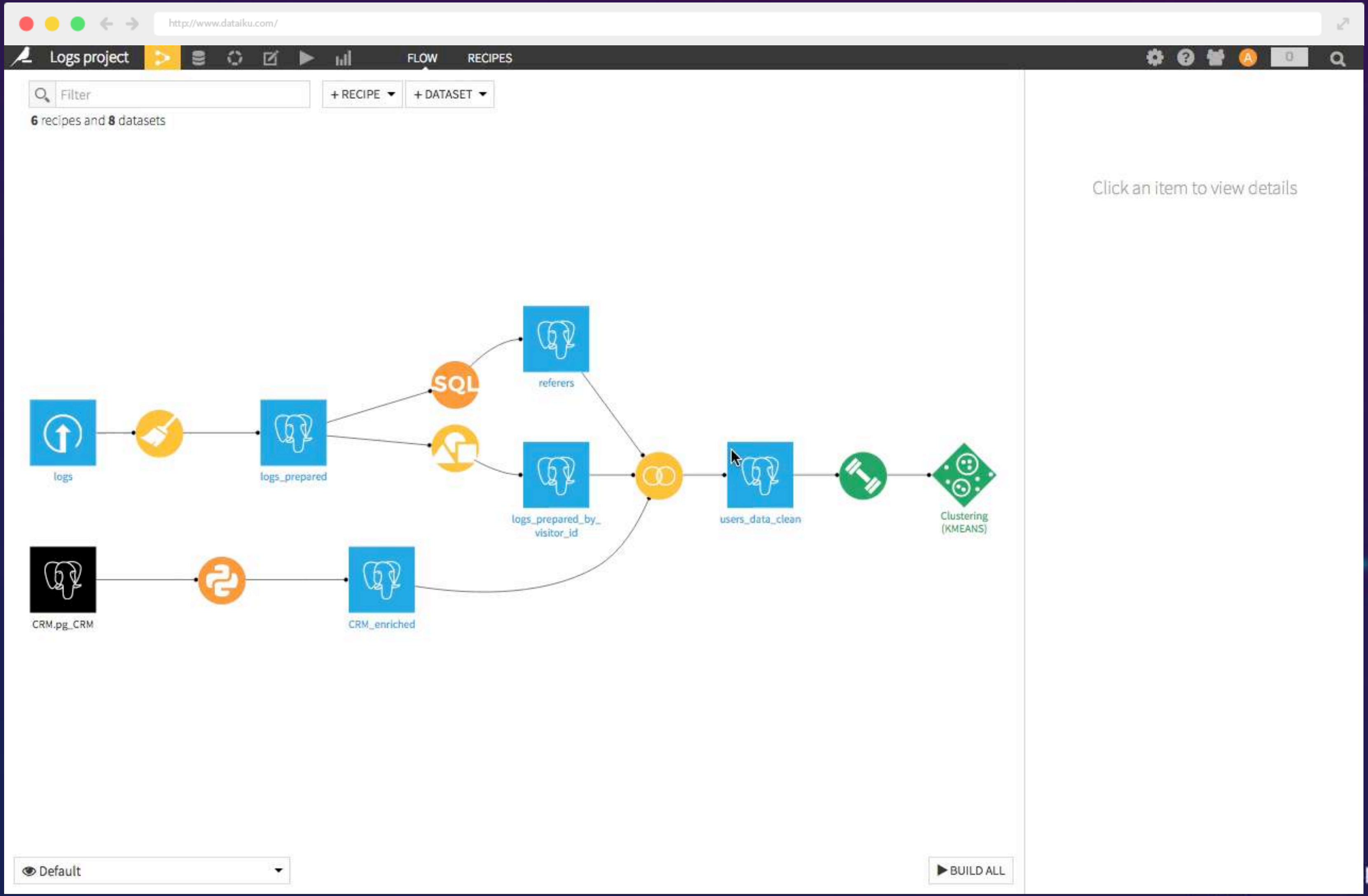
- Pour aller plus loin que l'interface visuelle dans la préparation de données
- Pour intégrer des éléments extérieurs (API, …)
- Pour utiliser des librairies de Machine Learning



```
In [11]: %pylab inline

         Populating the interactive namespace from numpy and matplotlib

In [12]: import dataiku
         from dataiku import pandasutils as pdu
         import pandas as pd
         import seaborn as sns
         import matplotlib.pyplot as plt
         sns.set(context="paper", font="monospace")

In [13]: # Load the DSS dataset as a Pandas dataframe
         mydataset = dataiku.Dataset("logs_prepared_by_visitor_id")
         df = mydataset.get_dataframe()

In [14]: # Get correlations
         corrmat = df.corr()
         # Set up the matplotlib figure
         f, ax = plt.subplots(figsize=(12, 9))
         # Draw the heatmap using seaborn
         sns.heatmap(corrmat, vmax=.8, square=True)

Out[14]: <matplotlib.axes._subplots.AxesSubplot at 0x75df387d5950>
```

Le « flow » pour l’industrialisation

- Exécuter l’ensemble de la préparation des données et la modélisation en quelques clics
- Possible de déployer sur une infrastructure de production dédiée (*batch & real time scoring*)



6 recipes and 8 datasets

Click an item to view details

m

Les cas d’usages classiques avec les logs

- Optimisation des conversions
- Segmentation des comportements
- Travail sur la recommandation
- Calculs de score de satisfaction client
- Détection comportements suspects
- …

vente privee

captain train

PagesJaunes

showroom prive.com

Bla Bla Car

Aramisauto

Nos clients (80+)

Web

Banques

Assurances

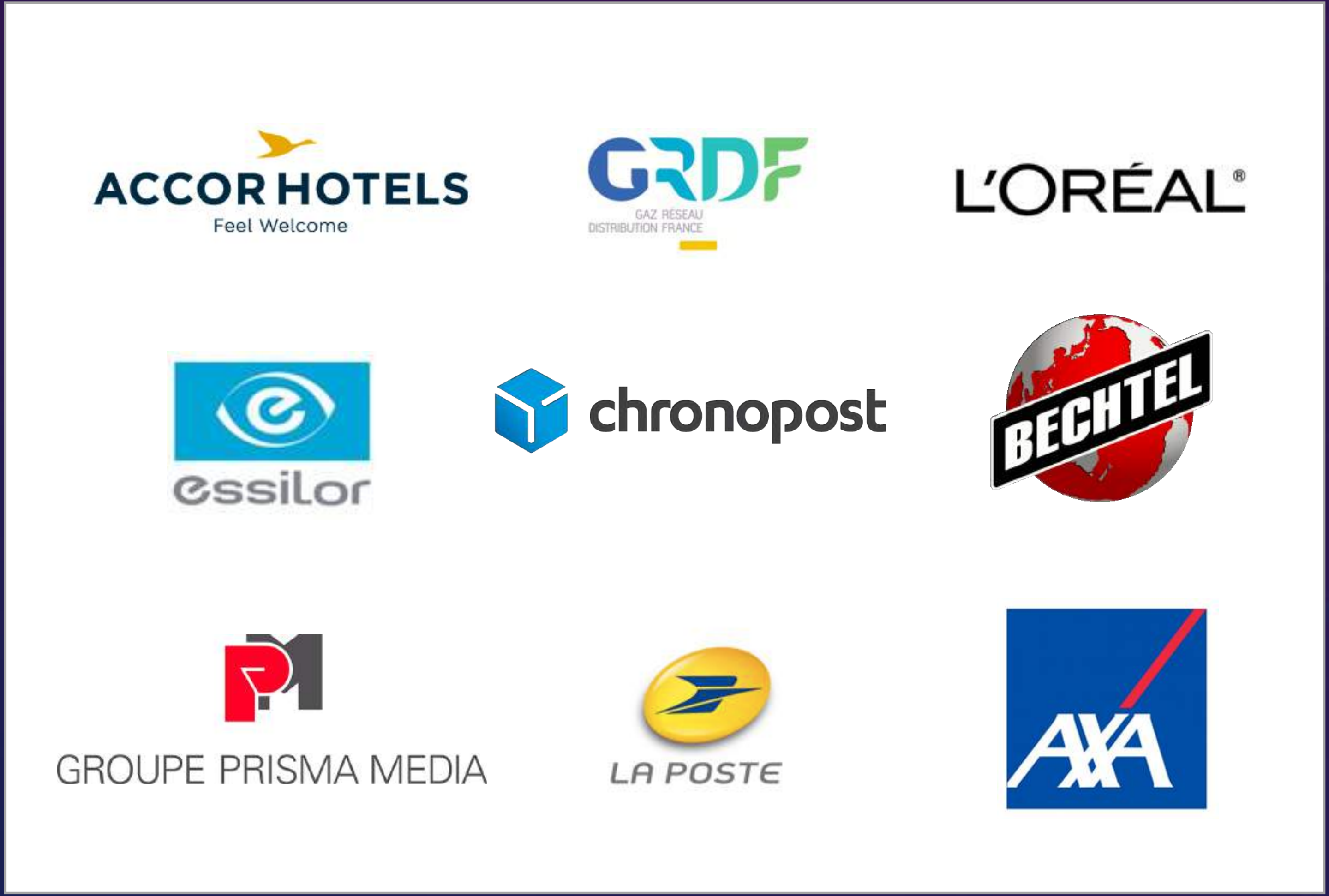
Industrie

Infrastructure

Santé

Médias

Jeux vidéo



Essayez Dataiku DSS

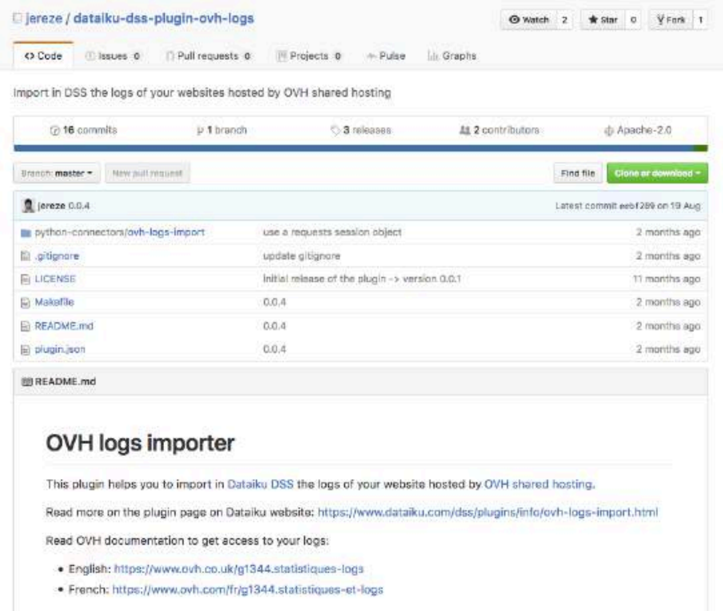
www.dataiku.com/try

Free Edition

Enterprise Edition

Bonus:

- Un plugin pour récupérer les logs OVH mutu
  github.com/jereze/dataiku-dss-plugin-ovh-logs
- Un tracker web open-source qui scale
  github.com/dataiku/wt1



Import in DSS the logs of your websites hosted by OVH shared hosting

jereze 0.0.4 — Latest commit eebf289 on 19 Aug

python-connectors/ovh-logs-import	use a requests session object	2 months ago
.gitignore	update gitignore	2 months ago
LICENSE	Initial release of the plugin -> version 0.0.1	11 months ago
Makefile	0.0.4	2 months ago
README.md	0.0.4	2 months ago
plugin.json	0.0.4	2 months ago

README.md

OVH logs importer

This plugin helps you to import in Dataiku DSS the logs of your website hosted by OVH shared hosting.

Read more on the plugin page on Dataiku website: https://www.dataiku.com/dss/plugins/info/ovh-logs-import.html

Read OVH documentation to get access to your logs:

- English: https://www.ovh.co.uk/g1344.statistiques-logs
- French: https://www.ovh.com/fr/g1344.statistiques-et-logs

Jérémy Grèze

@jeremy_gr

#OVHSummit

OVH.com