

# FIRY

## Outil d'analyse de l'historique de navigation de Mozilla Firefox

- Projet de fin de Master 2 BI& BIG DATA -

# FIRY

## Outil d'analyse de l'historique de navigation de Mozilla Firefox

- Projet de fin de Master 2 BI& BIG DATA -

Par Stephanya **CASANOVA MARROQUIN**

Encadrée par

**OMAR BOUSSAID** <sup>1</sup>

1: Laboratoire ERIC, Institut de Communication, Université de Lyon 2, France

# SOMMAIRE

1. **Introduction**
2. **Contexte Et Périmètre**
  - 2.1 Définition de la problématique
  - 2.2 Objectifs
3. **Travail Effectué**
4. **Conclusions**
5. **Questions**

# 1. INTRODUCTION

## 2. CONTEXTE ET PRÉ-REQUIS

## 2.1 Problématique



## 2.2 Objectifs



## 2.2.1 Objectif Général

*“Mettre en place un outil d’analyse de l’historique de navigation de Mozilla Firefox en considérant les étapes d’ingestion, préparation, raffinement et visualisation de données et les contraintes de traitement en temps réel et volume de données.”*

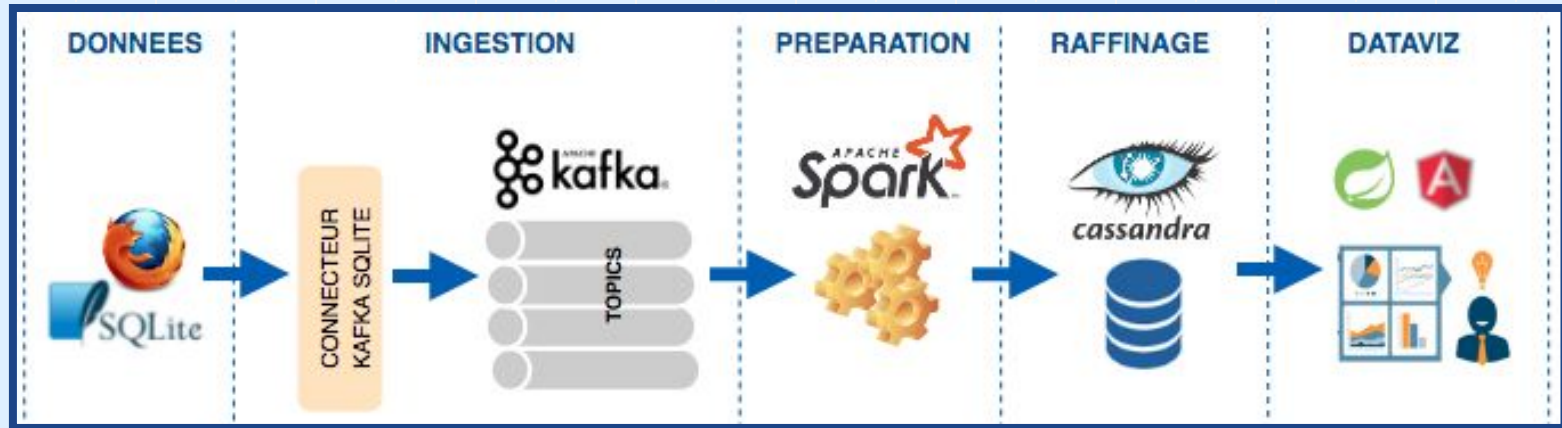


## 2.2.2 Objectifs Spécifiques

- ❖ *Identifier les axes d'analyse sur les données de navigation de Mozilla Firefox.*
- ❖ *Proposer une architecture qui satisfait les contraintes de traitement en temps réel et d'un grand volume de données.*
- ❖ *Mettre en place une application web qui permet de réaliser la “chaîne complète” : depuis la récupération des données jusqu'à la visualisation des résultats.*

### 3. TRAVAIL EFFECTUÉ

# Chaîne de valorisation des données de navigation



Architecture

## 3.1 Exploration des données et Définition du modèle d'analyse



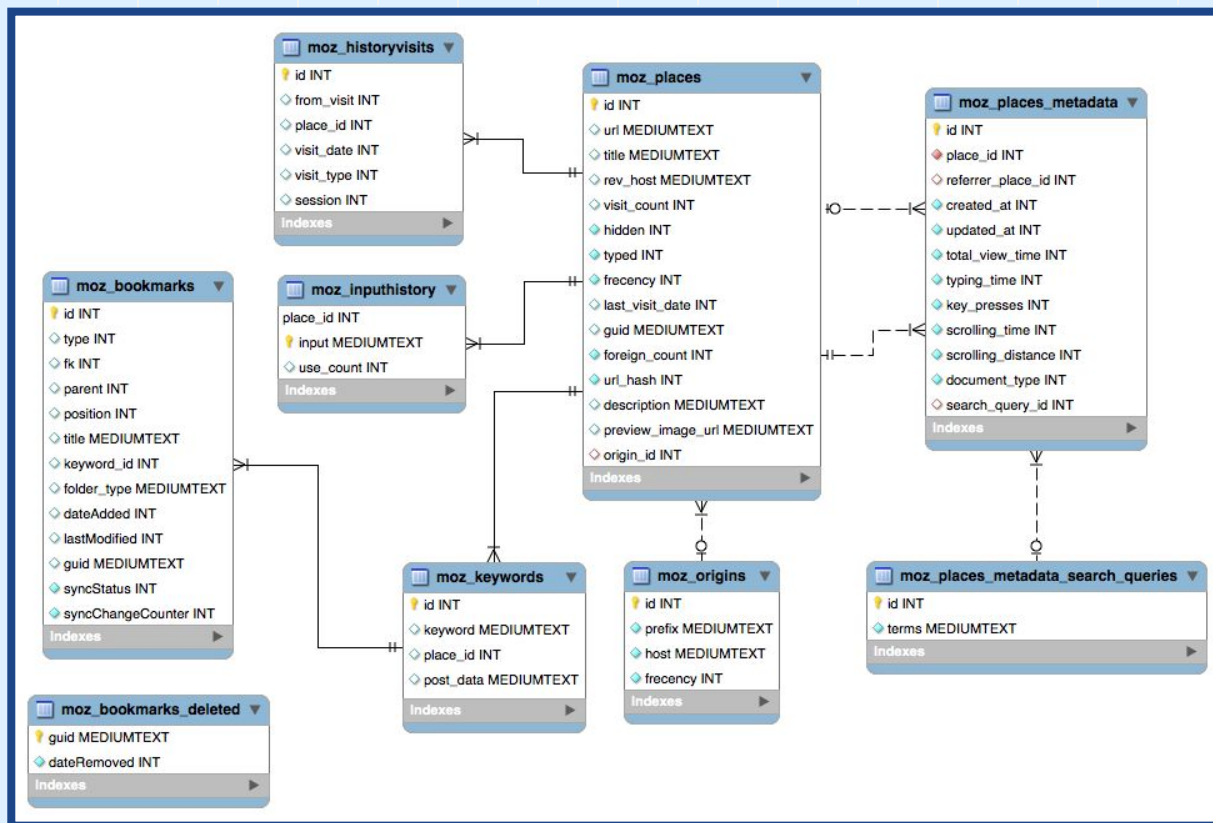
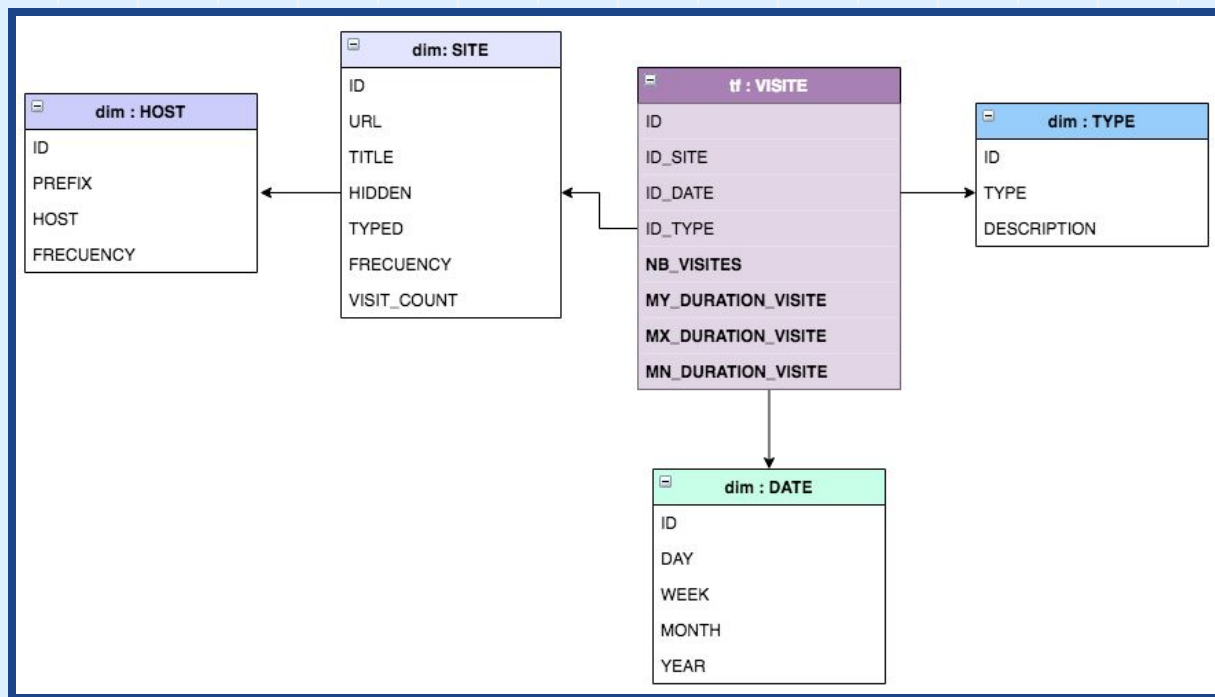
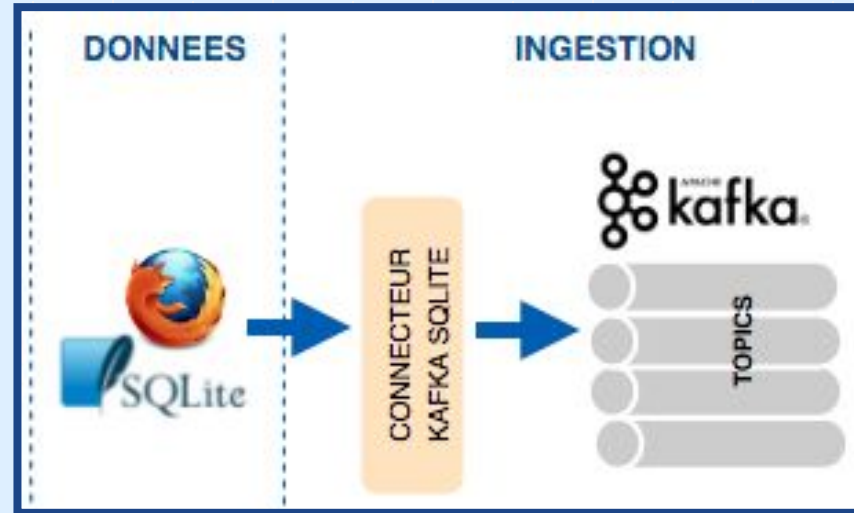


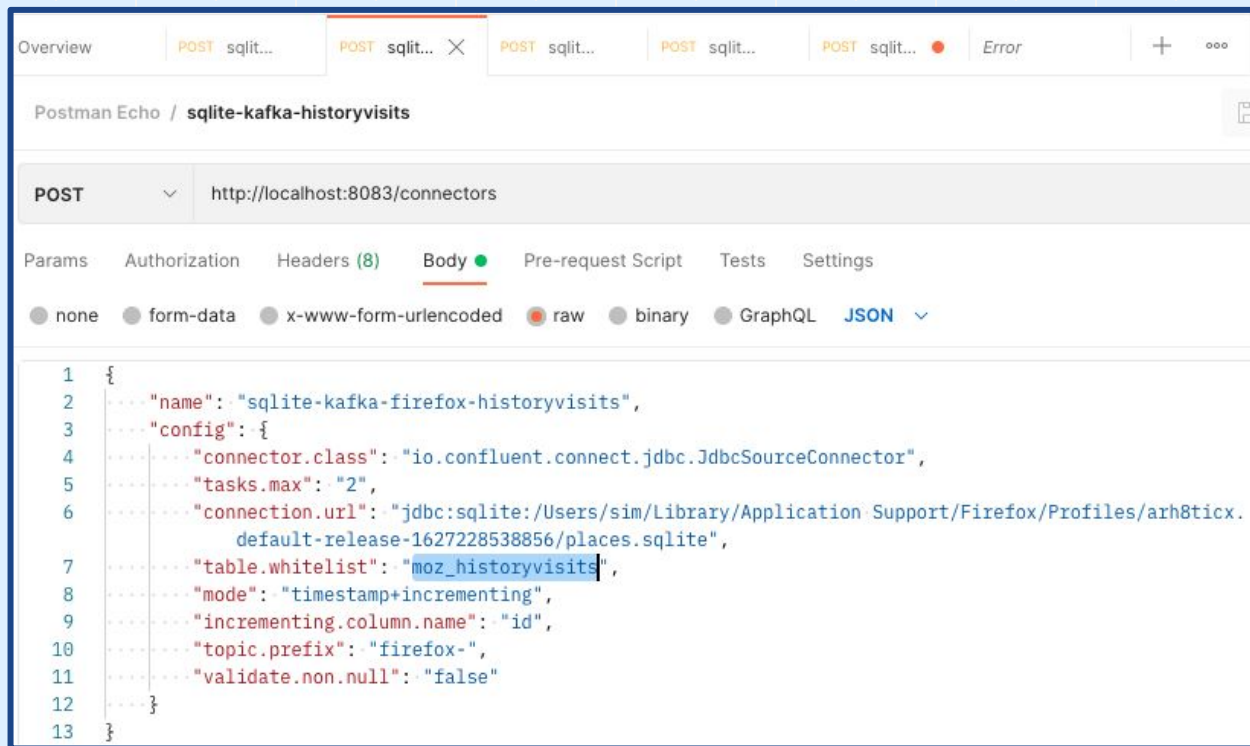
Diagramme E-R places.sqlite



Modèle d'analyse

## 3.2 Ingestion de données





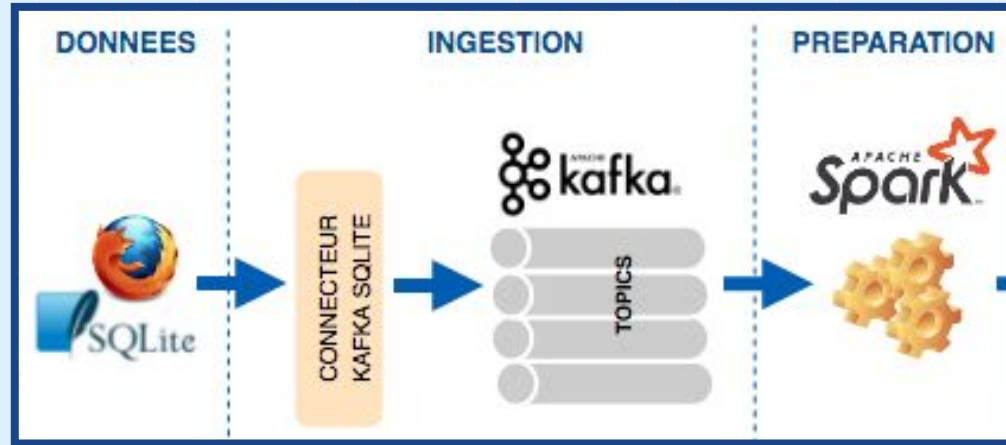
Création du connecteur sqlite - kafka



```
Last login: Wed Aug 25 14:09:10 on ttys005
[→ kafka_2.13-2.8.0 bin/kafka-console-consumer.sh --topic firefox-moz_historyvisits]
--from-beginning --bootstrap-server localhost:9092
{"schema":{"type":"struct","fields":[{"type":"int32","optional":false,"field":"id"},
{"type":"int32","optional":true,"field":"from_visit"},{"type":"int32","optional":true,
"field":"place_id"},{"type":"int32","optional":true,"field":"visit_date"},{"type":"int32","optional":true,"field":"visit_type"},{"type":"int32","optional":true,"field":"session"}],
"optional":false,"name":"moz_historyvisits"},"payload":{"id":1,"from_visit":0,"place_id":1,"visit_date":-1417447786,"visit_type":1,"session":0}}
```

Exemple de message récupéré depuis une topic

## 3.3 Préparation Des Données



```
{id,url,title,rev_host,visit_count,hidden,t,typed,frecency,description,location,guid,preview_image_url,last_visit_date,url_hash,origin_id,foreign_count} => {id,url,title,visit_count,hidden,t,typed,frecency,description,last_visit_date,origin_id}
```

1

## Tâche de filtrage des attributs

```
cqlsh:cassandra@firy> select * from visit where id='7649';
```

id	duration	from_visit	place_id	visit_date	visit_date_simple	visit_type
7649	0	0	3739	1630650660550	2021-09-03	1

(1 rows)  
cqlsh:cassandra@firy>

## Tâches de transformation de données

### Convert epoch to human-readable date and vice versa

 [batch convert]

Supports Unix timestamps in seconds, milliseconds, microseconds and nanoseconds

Assuming that this timestamp is in **milliseconds**:

**GMT:** Friday 3 September 2021 06:31:00.550

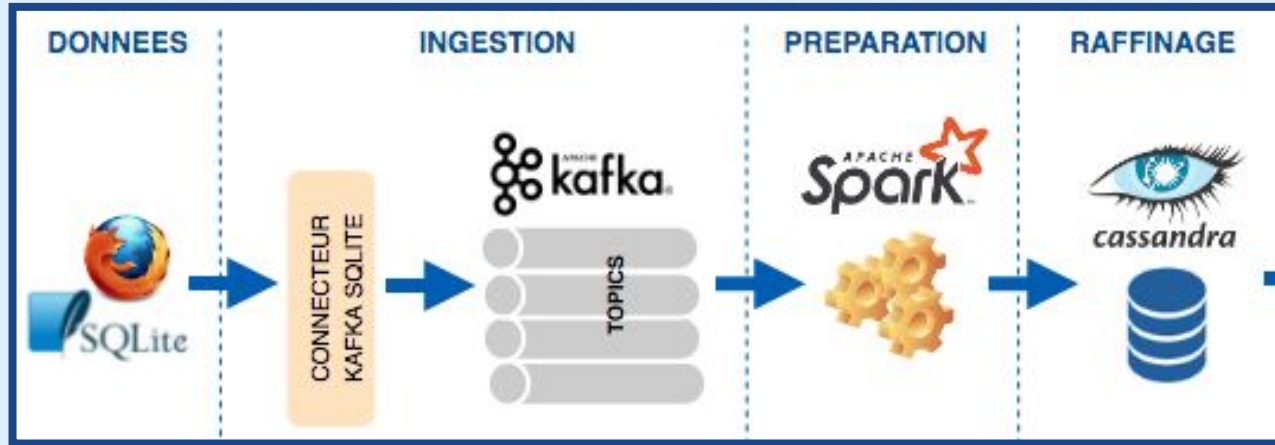
**Your time zone:** vendredi 3 septembre 2021 08:31:00.550 GMT+02:00

**Relative:** 4 minutes ago

2

```
private Dataset<Row> calculateDurationVisite(final Dataset<Row> visits)
{
    WindowSpec windowSpec = Window.orderBy("visit_date");
    return visits.withColumn("duration",
        visits.col("visit_date")
        .minus(when((lag("visit_date",1).over(windowSpec)).isNull(), 0)
        .otherwise(lag("visit_date", 1).over(windowSpec))));}
```

## 3.4 Raffinage Des Données



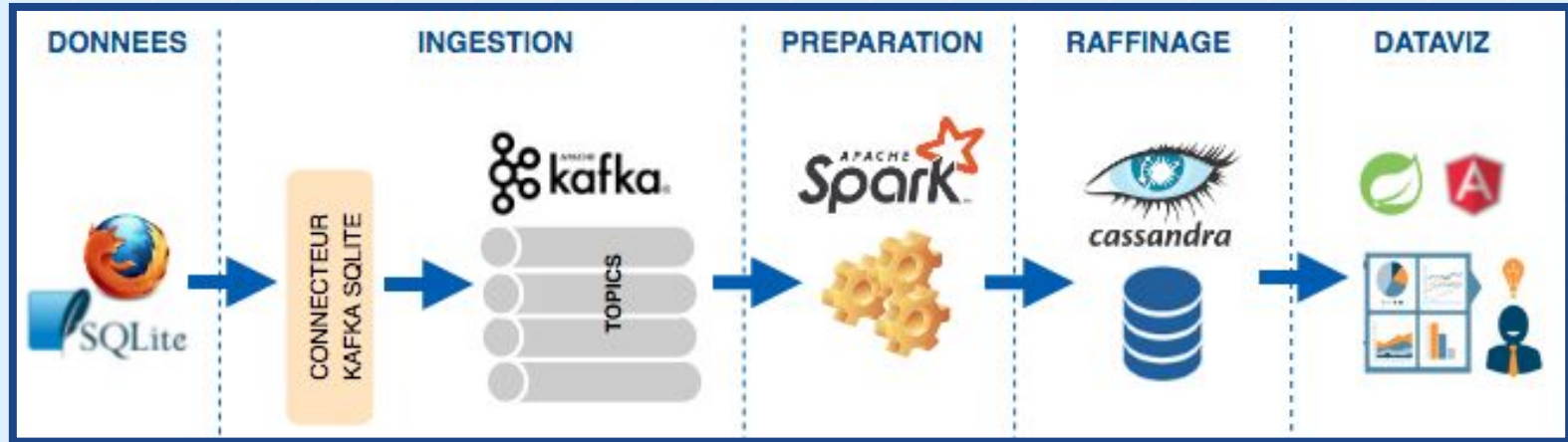
```
visits.groupBy("place_id", "visit_type", "visit_date_simple")  
      .agg( count("id").alias("nb_visits"));
```

```
visits.groupBy("place_id", "visit_type", "visit_date_simple")  
      .agg( avg("duration").alias("dur_mean_vis"));
```

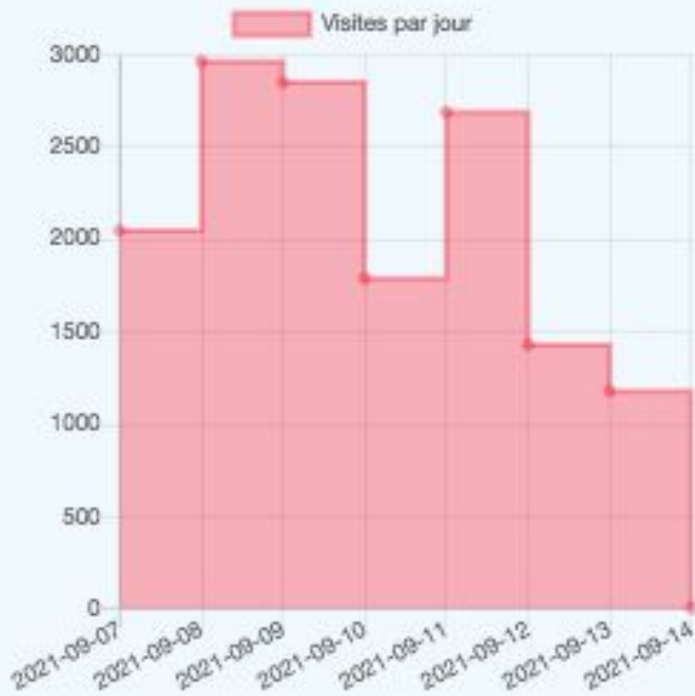
```
visits.groupBy("place_id", "visit_type", "visit_date_simple")  
      .agg( min("duration").alias("dur_min_vis"));
```

```
visits.groupBy("place_id", "visit_type", "visit_date_simple")  
      .agg( max("duration").alias("dur_max_vis"));
```

## 3.5 Visualisation



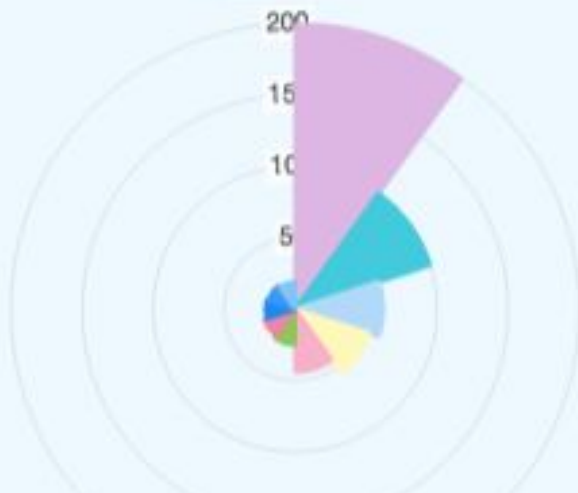
## Nombre de visites par jour



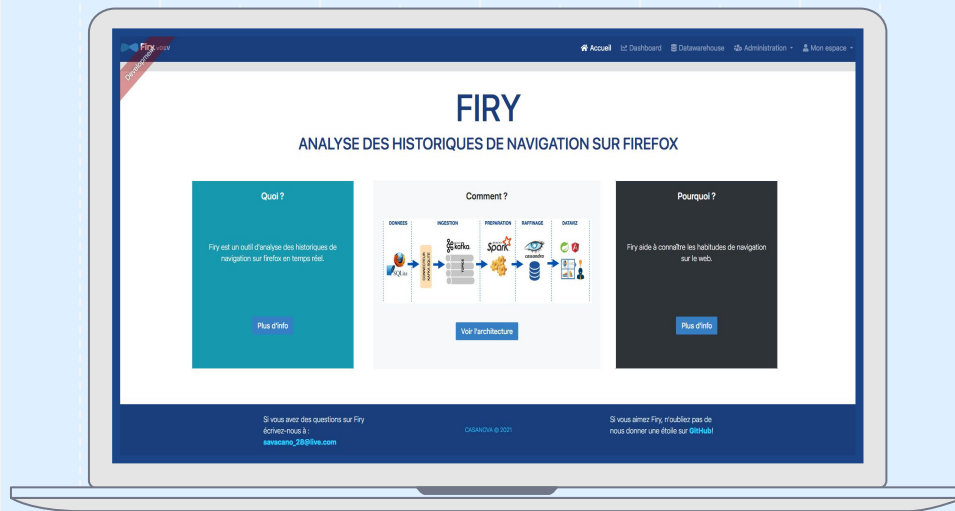




## Mots plus fréquents sur la navigation



## 3.6 Application WEB : FIRY



 Firy vDEV  
Development

Accueil Dashboard Datawarehouse Administration Mon espace

# FIRY

## ANALYSE DES HISTORIQUES DE NAVIGATION SUR FIREFOX

### Quoi ?

Firy est un outil d'analyse des historiques de navigation sur firefox en temps réel.

Plus d'info

### Comment ?



Diagram illustrating the data processing pipeline:

- DONNEES**: Sources (SQLite, Connecteur Kafka, Kafka, Spark, Cassandra, DataViz)
- INGESTION**: Ingestion process
- PREPARATION**: Preparation process
- RAFFINAGE**: Refinement process
- DATAVIZ**: Data visualization

Voir l'architecture

### Pourquoi ?

Firy aide à connaître les habitudes de navigation sur le web.


Plus d'info

Si vous avez des questions sur Firy écrivez-nous à : [savacano\\_28@live.com](mailto:savacano_28@live.com)

CASANOVA © 2021

Si vous aimez Firy, n'oubliez pas de nous donner une étoile sur [GitHub!](#)

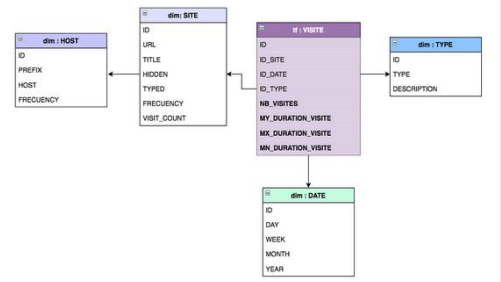




Accueil Dashboard Datawarehouse Administration Mon espace

# DATAWAREHOUSE

## FIRY : ANALYSE DES HISTORIQUES DE NAVIGATION SUR FIREFOX



```

graph LR
    dim_HOST[dim:HOST] --> fact_VISITE[fact:VISITE]
    dim_SITE[dim:SITE] --> fact_VISITE
    dim_TYPE[dim:TYPE] --> fact_VISITE
    dim_DATE[dim:DATE] --> fact_VISITE

```

Modèle de Données

**Informations**

Dernière mise à jour : 2021-09-04 14:20:00


**Types de Visites (dim : TYPE)**

Id	Type	Description
1	Transition_link	Ce type de transition signifie que l'utilisateur a suivi un lien et a obtenu une nouvelle fenêtre de niveau supérieur.
2	Transition_typed	Ce type de transition signifie que l'utilisateur a tapé l'URL de la page dans la barre d'URL ou l'a sélectionné dans les résultats de la saisie semi-automatique de la barre d'URL.
3	Transition_bookmark	Cette transition est définie lorsque l'utilisateur a suivi un signet pour accéder à la page.
4	Transition_embed	Ce type de transition est défini lorsqu'un certain contenu interne est chargé. Cela est vrai pour toutes les images d'une page et le contenu de l'iframe. C'est également vrai pour tout contenu dans un cadre, que l'utilisateur ait cliqué ou non sur quelque chose pour y accéder.
5	Transition_redirect_permanent	Défini quand la transition est une redirection permanente.
6	Transition_redirect_temporary	Défini quand la transition est une redirection temporaire.
7	Transition_download	Défini lorsque la transition est un téléchargement.
8	Transition_framed_link	L'utilisateur a suivi un lien depuis une modal.
9	Transition_reload	La page a été rechargée.

Si vous avez des questions sur Firy, écrivez-nous à : [savacano\\_28@live.com](mailto:savacano_28@live.com)

CASANOVA © 2021

Si vous aimez Firy, n'oubliez pas de nous donner une étoile sur [GitHub!](#)

 **Firy** v0.0.1

Accueil Dashboard Datawarehouse Administration

Development

### dashboard-resource

Dashboard Resource

**POST** `/api/firy-dashboard/add-figure` addFigure

**GET** `/api/firy-dashboard/data-chart-search-subject` getDataSearchSubject

**GET** `/api/firy-dashboard/data-chart-sites-with-duration` getDataSitesWithDuration

**GET** `/api/firy-dashboard/data-chart-visit-by-day` getDataVisitByDay

**GET** `/api/firy-dashboard/data-chart-visits-by-type` getDataVisitsByType

### data-ware-house-resource

Data Ware House Resource

**GET** `/api/firy-datawarehouse/data-from-table` getDataFromTable

**GET** `/api/firy-datawarehouse/dim-site` getDataFromTable

### firy-kafka-resource

Firy Kafka Resource

**GET** `/api/firy-kafka/consume` consume



## 4. CONCLUSION



# 5. QUESTIONS ?



# MERCI!