DAF Dataportal: manuale per l'utente

Release version: latest

AgID - Team Digitale

21 ago 2018

Indice

Il Dataportal è il punto di accesso al Digital & Analytics Framework (DAF) e al catalogo nazionale degli open data della Pubblica Amministrazione (PA).

Attraverso il Dataportal è possibile consultare il catalogo degli open data, leggere i risultati delle analisi dati («*data stories*») realizzate dalle Pubbliche Amministrazioni, dalla Community degli open data e dal team di data scientist del DAF, nonché contribuire all'accrescimento del sapere collettivo utilizzando gli strumenti di analisi disponibili sulla piattaforma DAF.

Il Dataportal è costituito da un"area pubblica, accessibile a tutti gli utenti del Web, e un"area privata, a uso esclusivo degli utenti registrati.

Questo manuale per l'utente illustra le principali operazioni che è possibile svolgere all'interno delle due piattaforme. La guida per gli sviluppatori, che descrive nel dettaglio le operazioni per l'installazione locale del DAF, è invece contenuta in un documento separato.

CAPITOLO 1

Data & Analytics Framework (DAF)

Il Data & Analytics Framework (DAF) è una delle attività atte a valorizzare il patrimonio informativo pubblico nazionale approvata dal Governo italiano nell'ambito del Piano Triennale per l'Informatica nella PA 2017-2019.

Per informazioni relative al che cosa è il DAF, a chi si rivolge, alla roadmap di evoluzione, all'"architettura di alto livello e ai casi d'uso che saranno realizzati durante la fase sperimentale si rimanda al Piano di Sviluppo.

1.1 DAF e Dataportal

L'intera piattaforma del DAF può essere riassunta con lo schema nella Figura 1.1, e viene descritta in dettaglio nella guida per gli sviluppatori (in inglese).

Il Dataportal costituisce il principale punto di accesso al DAF e alle sue funzionalità. L'utente finale può avere accesso ai cataloghi di dati ed eseguire analisi su di essi tramite una web app, senza dover installare alcun componente nel proprio computer.

Tuttavia, se l'utente lo desidera, può installare l'intera piattaforma localmente seguendo le istruzioni all'installazione contenute nella guida per gli sviluppatori (in inglese).

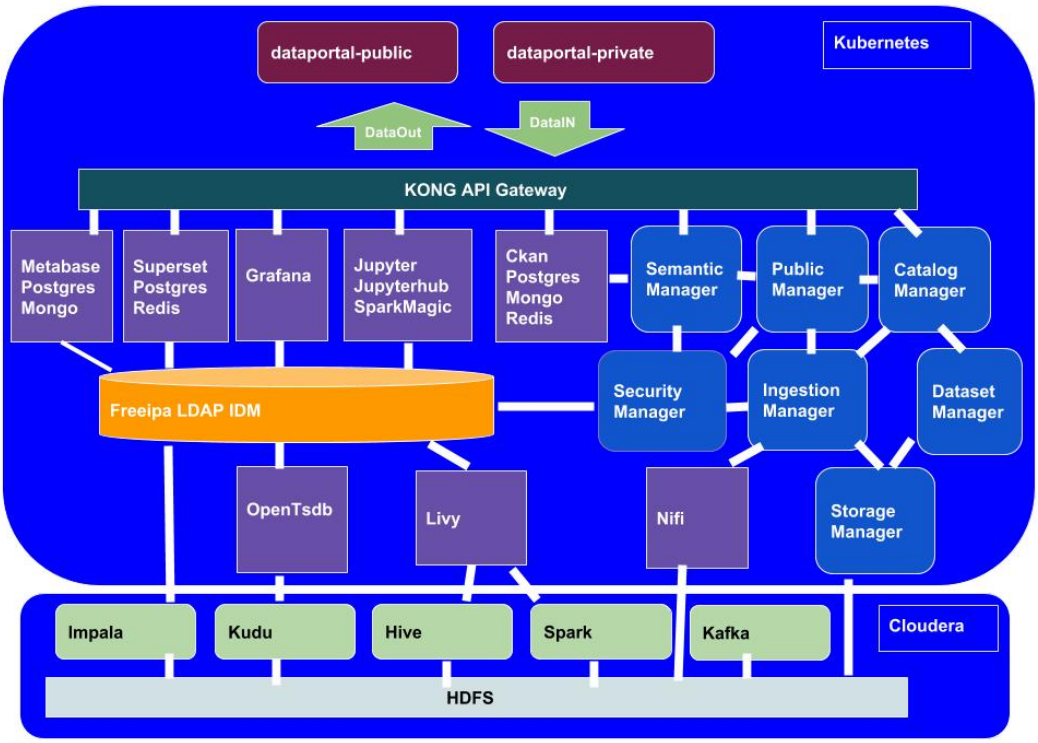


Fig. 1.1: Architettura del DAF.

CAPITOLO 2

Dataportal: area pubblica

L'area pubblica del Dataportal è disponibile all'indirizzo dataportal.daf.teamdigitale.it.

L'area pubblica rappresenta il punto di raccordo degli open data della PA e il punto di accesso al DAF. Essa costituisce l'evoluzione del sito dati.gov.it - *i dati aperti della Pubblica Amministrazione*.

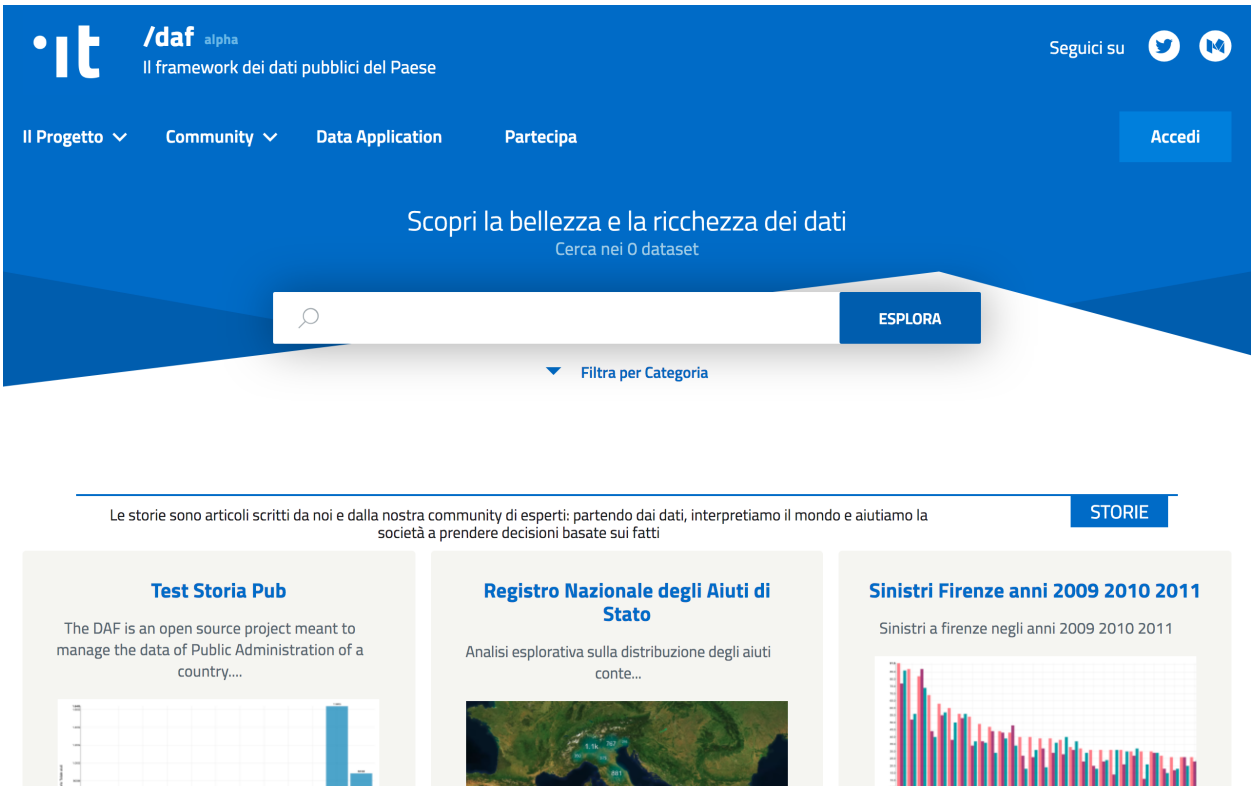


Fig. 2.1: Area pubblica del Dataportal.

Nota: Per poter accedere al catalogo degli open data presenti nella piattaforma, l'utente deve prima effettuare la registrazione al Dataportal, seguendo le stesse indicazioni per l'accesso all'area privata.

Oltre al catalogo degli open data, l'utente avrà accesso alle *data stories* condivise dagli utilizzatori del DAF, potrà ricercare e utilizzare le Data Application presenti sulla piattaforma, informarsi sulle novità provenienti dal mondo dei dati pubblici, e tanto altro.

In questa area sarà inoltre disponibile una sezione dedicata agli analisti e agli sviluppatori di Data Application contenente documentazione tecnica e schede informative sulle funzionalità del DAF.

LINK ALLE SEZIONI

- Catalogo dati
- Data stories
- Novità e documentazione

CAPITOLO 3

Dataportal: area privata

L'area privata del Dataportal è disponibile all'indirizzo dataportal-private.daf.teamdigitale.it.

Gli utenti hanno la possibilità di registrarsi per avere accesso all'area privata del Dataportal e, conseguentemente, agli strumenti di analisi messi a disposizione dalla piattaforma.

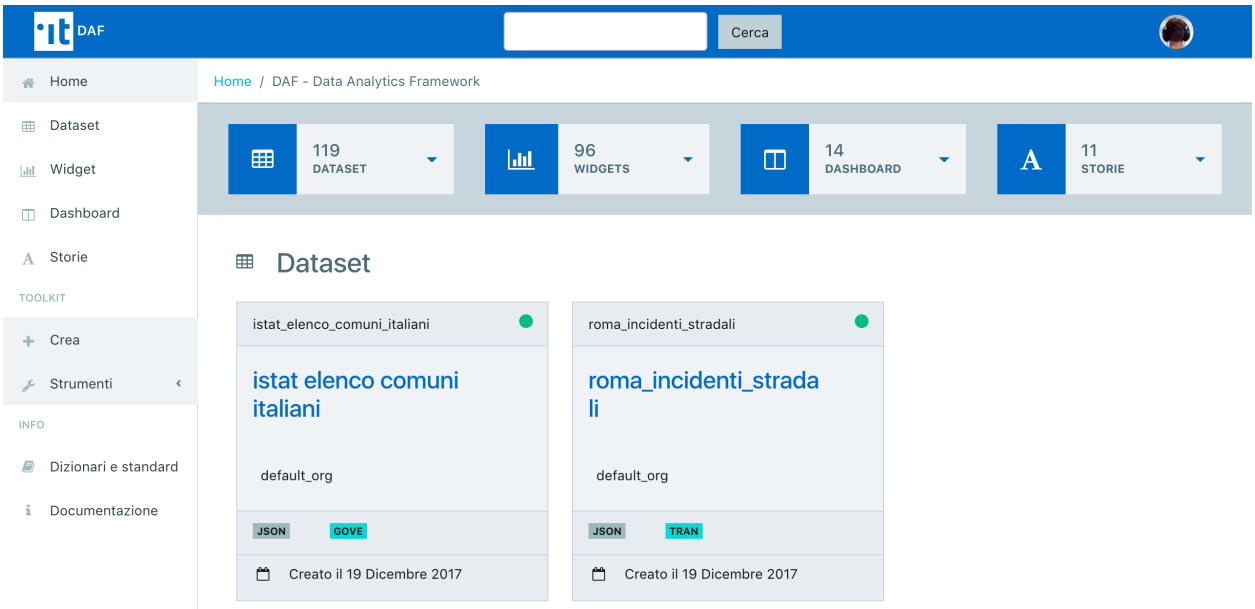


Fig. 3.1: Area privata del Dataportal.

L'idea è quella di fornire alle Pubbliche Amministrazioni e alla community di utilizzatori un insieme di strumenti per l'analisi e l'accesso ai dati contenuti nel DAF, chiaramente preservando i diritti di accesso associati alle varie utenze.

Dopo aver effettuato la registrazione e il login, nell'area privata è possibile[1]:

[1] Elenco non esaustivo e in evoluzione.

- Accedere al cruscotto utile alla registrazione di nuovi dataset. La procedura guidata facilita una completa metadatazione in conformità al profilo DCAT-AP_IT; i metadati sono arricchiti con informazioni sulla struttura del dato e relativa semantica[2].
- Creare e pubblicare *Dashboard*, ovvero un insieme di grafici creati con gli strumenti messi a disposizione dal DAF.
- Creare e pubblicare *Storie* (*data stories*), ovvero dei blog post in cui è possibile inserire nativamente risultati di analisi svolte con gli strumenti messi a disposizione dal DAF.
- Utilizzare Jupyter Notebook integrato con il cluster big data del DAF, con cui effettuare analisi esplorative iniziali fino a complessi modelli di machine learning.
- Accedere allo strumento di business intelligence (Superset) e strumenti più orientati alla data visualization (Metabase). I risultati delle analisi svolte con questi strumenti possono essere integrati nativamente nelle Data Story e nelle Dashboard.

Le pubbliche amministrazioni registrate potranno inoltre usufruire di strumenti avanzati di gestione degli utenti.

Operazioni e strumenti dell'area privata del Dataportal

3.1 Registrazione e accesso all'area privata

3.1.1 Registrazione

Accedere alla pagina https://dataportal-private.daf.teamdigitale.it direttamente o tramite il pulsante *Accedi* presente sul portale pubblico. Cliccare su *Registrati!* e fornire le informazioni richieste, quindi cliccare su *Crea Account*. In mancanza di segnalazioni (nel qual caso è necessario prima risolverle) il sistema provvederà all'invio di un link di attivazione all'indirizzo email indicato.

Nota Il nome utente può essere composto esclusivamente da caratteri alfanumerici e dai seguenti caratteri speciali: – (dash) e _ (underscore). La password deve essere formata da almeno 8 caratteri e contenere almeno una lettera maiuscola e almeno un numero.

Fig. 3.2: La schermata di login.

3.1.2 Autenticazione e home page

Accedere alla pagina https://dataportal-private.daf.teamdigitale.it direttamente o tramite il pulsante *Accedi* presente sul portale pubblico. Utilizzare l'indirizzo email e la password definiti in fase di registrazione.

[2] Al momento questa funzionalità è offerta solo alle Pubbliche Amministrazioni

Eseguito l'accesso viene proposta la home page della sezione privata, che sulla sinistra presenta un menu da quale è possibile accedere alle funzionalità e agli strumenti di analytics. Il menu può essere ridotto e riespanso cliccando sulla barra in calce allo stesso.

In primo piano compare una sintesi dei dataset, delle dashboard, dei widgets e delle storie visibili all'utente, e la possibilità di accedere rapidamente alla rispettiva sezione o al dettaglio.

Cliccando sul menu a tendina accanto al proprio nome utente, è possibile visualizzare i dettagli del proprio profilo. Oltre alle informazioni fornite in fase di registrazione, sono riportate anche indicazioni sul ruolo di cui si è in possesso e sulle organizzazioni di cui si fa parte. Dallo stesso menu è possibile anche eseguire il logout.

In qualsiasi momento è possibile ritornare alla home cliccando su *Home*, prima voce nel menu laterale.

3.1.3 Reset password

La funzionalità, disponibile dalla pagina di autenticazione, permette sia di ripristinare l'accesso alla sezione privata del portale nel caso non si ricordi la password, sia di modificare la propria password. A tal fine fornire la mail utilizzata in fase di registrazione, alla quale, se corretta, verrà inoltrato un link. Utilizzare il form per indicare la nuova password.

3.2 Gestione dei dataset

La metadatazione e il caricamento di un dataset sul DAF sono operazioni concesse solamente a specifiche utenze dell'organizzazione che assumono il ruolo di *Editor*. Tali operazioni possono essere eseguite sfruttando le funzionalità presenti nel menu *Toolkit* > *Crea* > *Nuovo dataset*.

La ricerca e l'accesso ai dataset sono funzionalità accessibili a tutti gli utenti, i dataset rilasciati in formato open data sono direttamente ricercabili dall"interfaccia pubblica del portale, i dataset caricati e appartenenti alla propria organizzazione sono accedibili tramite l"interfaccia privata del portale secondo i permessi selezionati.

3.2.1 Elenco e ricerca dei dataset

La funzionalità di elenco e ricerca è attivabile dalla voce di menu *Dataset* o attraverso il filtro di ricerca presente in alto a destra. In questo secondo caso i dataset mostrati sono quelli rispondenti alla ricerca effettuata. In entrambi i casi, sono mostrati solo i dataset sui quali l'utente ha diritti di accesso. L'elenco può essere ordinato o filtrato; è inizialmente composto dai primi 10 dataset, continuando a scorrere la lista vengono man mano caricati i successivi.

Cliccando su un dataset viene mostrata una pagina contenente:

- informazioni di dettaglio relative al dataset
- la possibilità di eseguire il download del dataset (limitato a 1000 record)
- un'anteprima del contenuto del dataset (in formato json)
- l'API endpoint per effettuare chiamate REST API
- informazioni di accesso agli strumenti di analisi: Superset e Jupyter
- eventuali slice realizzate con il dataset

3.2.2 Caricamento di un nuovo dataset

Per inserire un nuovo dataset all'interno della piattaforma DAF è necessario prima di tutto eseguire una procedura di metadatazione, così da definire le caratteristiche del flusso di ingestion e quelle del dataset stesso.

Nel caso di un flusso batch, inoltre, i file dovranno essere caricati su un canale SFTP.

Nota: la prima volta che si definisce un flusso batch è necessario effettuare un accesso preliminare all'host daf.teamdigitale.it (porta 22) utilizzando la propria chiave ssh privata e dopo aver comunicato al Team la propria chiave pubblica. Il nome utente è disponibile accedendo dal portale ai dettagli del proprio profilo.

Creazione e metadatazione del dataset

Per iniziare la procedura di caricamento di un nuovo dataset è necessario cliccare nel menu *Toolkit* > *Crea* > *Nuovo dataset*.

Il cruscotto di registrazione si compone di 3 passi.

Passo 1: Carica file e descrivi le colonne

Il primo passo permette di definire il tipo di file utilizzato per la trasmissione dei dati e la modalità di caricamento. Al momento sono supportati i formati CSV e JSON (quest'ultimo in alpha test) che possono essere conferiti in modalità batch o indicando un servizio web da interrogare periodicamente (quest'ultimo in alpha test).

In entrambi i casi, il processo di metadatazione è supportato da una procedura di inferenza su un file campione rappresentativo del dataset reale. Questo permette al sistema di inferire la struttura del dataset (campi e tipi dei campi) e di proporla all'utente. Per i flussi batch caricare il file campione mediante il *Drag and Drop*

E" necessario che il file campione rispetti alcune regole:

- Il numero di record deve esser limitato (circa 50 righe)
- I nomi dei campi non devono avere spazi al loro interno
- I nomi dei campi devono essere tutti diversi
- I nomi dei campi non devono contenere caratteri speciali (e.g. "@", "-", "(", ")") né delimitatori di stringhe
- Dimensione del file sample massima: 1 MB
- Deve essere rappresentativo del dataset al fine di migliorare il risultato del processo di inferenza: per quanto possibile deve contenere record coerenti con il tipo di dato che ci si aspetta sulla tabella finale
- I file CSV devono contenere l'header sia nel sample che in tutti i file che vengono caricati successivamente
- I file JSON devono avere i dati su una sola riga
- Encoding UTF-8 (senza BOM)

Dopo aver caricato il file verrà visualizzata una schermata con l'elenco dei campi inferiti e per ognuno un campione dei valori contenuti.

Inizialmente è necessario indicare se il dataset è privato; in caso contrario, il dataset sarà automaticamente aggiunto tra quelli fruibili liberamente al di fuori dell'organizzazione e come open data (disponibile a breve).

Successivamente, per ogni campo del dataset è possibile indicare:

- tipo del campo: scegliere il tipo di dato rappresentato (int, string, timestamp. . .)
- concetto semantico: individuare nelle ontologie installate il concetto espresso dal campo; il sistema fornisce suggerimenti man mano che viene valorizzato
- descrizione: inserire una descrizione per il campo
- tags
- obbligatorio: indicare se il campo è obbligatorio o meno
- tipo della colonna

Passo 2 - Metadati:

Indicare:

- il titolo
- una descrizione del dataset
- categoria
- licenza
- organizzazione di appartenenza (nel caso di dataset pubblico selezionare default_org)

Passo 3 - Modalità di invio

Definire:

- Dominio e sottodominio di appartenenza del dataset

al momento trascurare i seguenti campi:

- se il dataset definisce uno standard
- se il dataset segue uno standard
- il tipo di lettura del dataset
- il tipo di dataset

Al termine si viene indirizzati alla pagina di dettaglio del dataset. Si ricordi però che non è presente ancora alcun dato, quindi il download e l'anteprima non produrranno alcun risultato.

Caricamento dei dataset via SFTP

Nel caso di caricamento di dataset in modalità batch mediante canale SFTP effettuare l'accesso all'host daf.teamdigitale.it (porta 22) utilizzando la propria chiave ssh privata. Se non si dispone di un coppia di chiavi ssh, è possibile utilizzare il nome utente e la propria password (il nome utente è disponibile accedendo dal portale ai dettagli del proprio profilo); questa possibilità sarà deprecata a breve.

Caricare il/i file relativi al dataset definito in precedenza al path che sarà stato creato dal sistema, utilizzando un percorso relativo. La struttura segue la convenzione `dominio/sottodominio/dataset/`. Nel caricamento non impostare l'opzione (`-p`) di mantenimento della data di aggiornamento.

Tutti i file che verranno man mano caricati in tale cartella incrementeranno il contenuto del dataset. Un processo in ascolto si occuperà di acquisire i file al massimo entro qualche minuto, in base alle dimensioni del file.

Caricamento con interrogazione di web service

Funzionalità disponibile a breve

3.3 Pubblicazione di Dashboard e Data Story

Dashboard e data story sono due elementi chiave del DAF, che forniscono all'utente una visione di insieme degli insights e delle conclusioni tratte dall'analisi dei singoli dataset. Entrambe si possono comporre di uno o più widget. Per creare una dashboard o una data story è necessario accedere all'area privata del Dataportal.

3.3.1 Widget

Un widget è una slice importata da Superset o Metabase. Tutte le slice disponibili e visibili all'utente sono automaticamente mostrate dalla funzionalità accessibile a partire dalla voce di menu *Widget*.

3.3.2 Le mie Dashboard

Le dashboard raccolgono elementi testuali e rappresentazioni dei dati realizzate con gli strumenti di analisi quali *Superset* e *Metabase*.

Scegliendo l'opzione *Dashboard*, si attiva la funzionalità che mostra tutte le dashboard create dall'utente, dagli utenti appartenenti alla sua stessa organizzazione e quelle la cui visibilità è aperta a tutti. L'elenco delle dashboard può essere filtrato, tramite le parole contenute nei titoli delle dashboard di interesse, usufruendo dell'apposita barra di ricerca.

Ogni dashboard è indicata da una *card* grazie cui si individuano subito:

- Il *titolo*;
- Il *nome utente* dell'autore;
- Il *cartellino* che la definisce come pubblica o privata;
- Il numero di *widget* di cui si compone;
- Il *dataset* cui fa riferimento.

Cliccando sul titolo si ha la possibilità di visualizzare la dashboard. Immediatamente sotto al titolo compare un tag che indica l'organizzazione cui la dashboard appartiene.

Quindi, si ha la possibilità di modificare la dashboard. Una volta effettuate le modifiche, per pubblicarla ed uscire dalla modalità *bozza*, occorre salvare e pubblicare utilizzando l'icona apposita. Le modalità di *edit* corrispondono a quelle di creazione e sono riportate nel paragrafo seguente.

3.3.3 Creazione di una nuova Dashboard

Per creare una nuova dashboard, trovandosi su *Le mie dashboard*, è sufficiente premere sull'icona con il + che si trova alla destra della barra di ricerca. Altrimenti, dal menù principale, cliccare su *Crea* scegliendo di creare una dashboard.

In entrambi i casi è chiesto di inserire delle informazioni di base sulla nuova dashboard:

- Il *titolo*;
- Il *sottotitolo*;
- Se ha *visibilità* privata o pubblica, nel caso di visibilità privata è richiesto indicare l'organizzazione di riferimento.

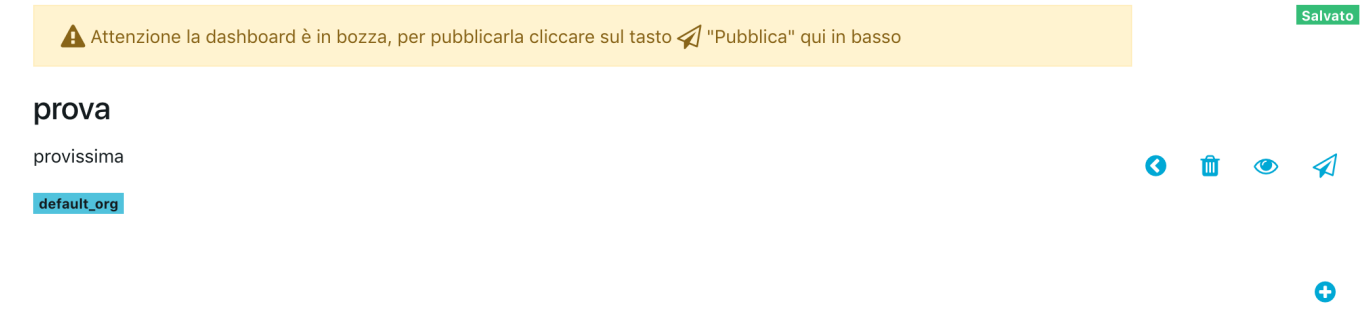
Quindi, appare la seguente schermata su cui iniziare a lavorare.

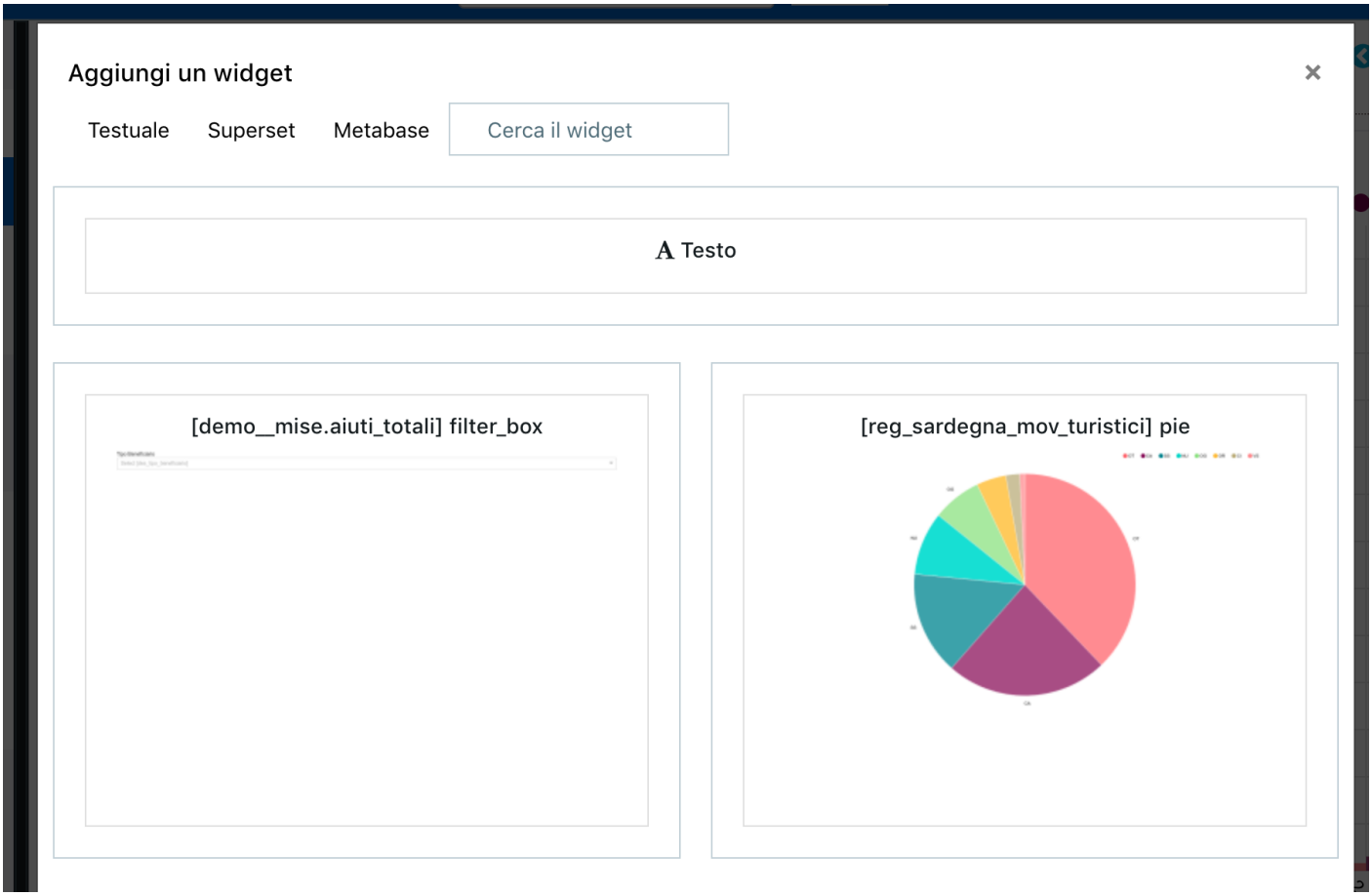
Per iniziare a popolare la dashboard si preme il tasto *Aggiungi* attivando la funzionalità di aggiungere un widget. Sono mostrati i widget a cui si ha accesso.

Come visibile dall'immagine, si ha la possibilità di scegliere la tipologia di widget (*Testo*, *Metabase*, *Superset*) e di effettuare ricerche basate sul titolo delle slices/grafici in elenco.

Al momento i widget inseriti sono tra loro indipendenti. Perciò, eventuali widget corrispondenti a slices che effettuano filtri su *Superset* non hanno lo stesso potere sulla dashboard creata sul *Dataportal*.

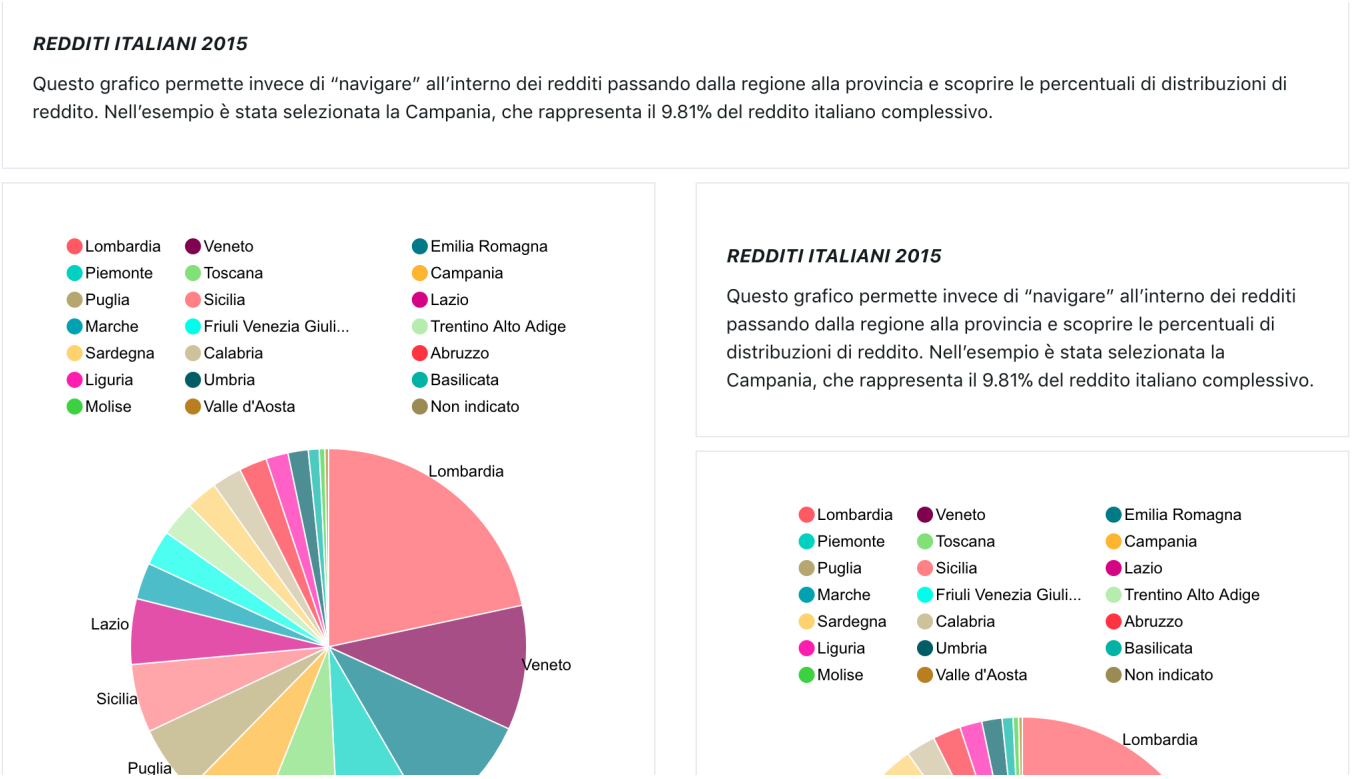
Durante il processo di creazione della dashboard è possibile visualizzarne l'*anteprima*. Invece per condividerla è necessario cliccare sul tasto *Pubblica*; è possibile condividere il lavoro con la sola organizzazione dell'utente o con tutti.


Attenzione la dashboard è in bozza, per pubblicarla cliccare sul tasto "Pubblica" qui in basso
Salvato
prova
provissima
default_org


Aggiungi un widget
Testuale
Superset
Metabase
Cerca il widget
Testo
[demo__mise.aiuti_totali] filter_box
[reg_sardegna_mov_turistici] pie

Per eliminare una dashboard entrare in modalità *modifica* della storia e premere il pulsante *Elimina*.

A seguire un esempio di dashboard.



3.3.4 Le mie storie

Una Storia, detta anche Data Story, è una modalità di presentazione, simile ad un blog post, che alterna parti testuali a grafici/widget realizzati tramite Superset.

Per visualizzare le proprie storie si sceglie, dal menù principale, la sezione *Storie*. Questa mostra tutte le storie create dall'utente, dagli utenti appartenenti alla sua stessa organizzazione e quelle la cui visibilità è aperta a tutti. Analogamente a quanto accade per *Le mie dashboard*, l'elenco mostrato può essere filtrato usufruendo dell'apposita barra di ricerca.

Ogni storia è indicata da una *card* su cui appaiono:

- Il *titolo*;
- Il *nome utente* dell'autore;
- Il *cartellino* che la definisce come pubblica o privata.

3.3.5 Creazione di una nuova storia

Per creare una storia, trovandosi su *Le mie storie*, è sufficiente premere sull'icona con il + che si trova alla destra della barra di ricerca. Oppure, dal menù principale, cliccare su *Crea* scegliendo di creare una storia.

Quindi si inseriscono le informazioni relative alla storia quali:

- *Titolo*;

- *Privata* o *pubblica*.

Una volta inserite queste informazioni appare la seguente schermata su cui iniziare a lavorare.

Si è nella modalità di editing della storia. Pertanto è possibile:

- Modificare il *titolo*;
- Inserire il *sottotitolo*;
- Lavorare al contenuto della storia.

Il contenuto della storia è popolato inserendo i widget con le stesse modalità indicate per la dashboard. Per eliminare una dashboard occorre entrare in modalità *modifica* della storia e premere il pulsante *Elimina*. A seguire un esempio di data story.

3.4 Amministrazione di utenti e organizzazioni

3.4.1 Aggiunta e rimozione di utenti

3.4.2 Aggiunta e rimozione di organizzazioni

3.5 Modalità di impiego dell'API REST

A ogni dataset caricato sul DAF è associato un API endpoint che permette l'interazione con lo stesso in via programmatica. L'endpoint è mostrato nella pagina di dettaglio del dataset, e assume la forma:

`https://api.daf.teamdigitale.it/dataset-manager/v1/dataset/<dataset_logical_url>`

È possibile richiamarlo previa autenticazione (basic) utilizzando l'indirizzo email associato al proprio account.

3.5.1 Download del dataset

È possibile effettuare lo scarico del dataset (bulk download) con una richiesta `HTTP GET` alla URL ad esso associata.

> Ad esempio: https://api.daf.teamdigitale.it/dataset-manager/v1/dataset/daf%3A%2F%2Fdataset%2Fdefault_org%2FGOVE__amministrazione%2Fistat_elenco_comuni_italiani

Di default la chiamata restituisce i primi 1.000 record del dataset in formato JSON. E" comunque possibile specificare i seguenti parametri in *query string*:

- `limit` per limitare il numero di record restituiti, indicare un valore numerico non superiore al massimo consentito (1.000)
- `format` per specificare il formato dei dati restituiti, valori possibili: `json` (default) e `csv`
- `method` per specificare il metodo di download, valori possibili `quick` (default) e `batch`. La prima è soggetta alla limitazione sul numero di record restituiti, la seconda invece nel caso di dataset molto grandi restituisce una redirect per lo scarico dell'intero file

3.5.2 Recupero del dataset schema

Richiamando (in `HTTP GET`) l'endpoint col suffisso `/schema`, si possono ottenere informazioni sullo schema del dataset.

Ad esempio: https://api.daf.teamdigitale.it/dataset-manager/v1/dataset/daf%3A%2F%2Fdataset%2Fdefault_org%2FGOVE__amministrazione%2Fistat_elenco_comuni_italiani/schema

3.5.3 Interrogazione del dataset

Richiamare l'endpoint in `HTTP POST` aggiungendo alla url il suffisso `/search`.

Ad esempio: https://api.daf.teamdigitale.it/dataset-manager/v1/dataset/daf%3A%2F%2Fdataset%2Fdefault_org%2FGOVE__amministrazione%2Fistat_elenco_comuni_italiani/search

L'API è ottimizzata per dataset con una struttura colonnare (csv), il supporto ai json è al momento consente l'interazione solo con gli elementi di primo livello.

Una query può essere composta di un insieme di clausole, scelte tra le seguenti supportate:

1. select
2. where
3. groupBy
4. having
5. limit

Le clausole rappresentano le proprietà complesse dell'oggetto json che deve essere inserito nel corpo della richiesta, che quindi assume la forma mostrata di seguito:

```
{
        "select": ...,
        "where": ...,
        "groupBy": ...,
        "having": ...,
        "limit": ...
}
```

Di seguito vengono dettagliate le singole proprietà.

SELECT

Permette di indicare i campi (colonne) di interesse del dataset ed eventualmente l'alias. La proprietà `select` accetta come valore una lista di oggetti del tipo: `{ "name": "col2", "alias": "alias_col2" }`. Come valore di `name` è ammesso l'uso del carattere `*`. La lista può essere vuota, in tal caso si assume l'estrazione di tutte le colonne (come con l'uso del `*`)

Esempio: selezione della colonna `col1` e della colonna `col2` con alias `alias_col2`

```
{
        "select": [
                { "name": "col1" },
                { "name": "col2", "alias": "alias_col2" }
        ]
}
```

è equivalente a

```
SELECT col1, col2 AS alias_col2
FROM table
```

E" anche possibile rappresentare una colonna contenente un valore costante utilizzando un oggetto del tipo `{ "value": 1, "alias": "one_alias" }`, con alias sempre opzionale.

Esempio:

```
{
        "select": [
                { "value": "string" },
                { "value": 1, "alias": "one_alias" }
        ]
}
```

è equivalente a

```
SELECT 'string', 1 AS one_alias
FROM table
```

WHERE

Permette di indicare condizioni di filtro mediante l'uso di operatori logici e di confronto. La proprietà `where` accetta come valore un oggetto composto da uno o più operatori, a seconda della complessità del filtro, scelti tra quelli:

- **logici**
  - unario di negazione `not`: accetta come valore un oggetto
  - `and` e `or`: accettano come valore una lista di oggetti
- **binari di confronto: accettano come valore un oggetto composto dalle proprietà `left` e `right` relativi ai due elementi di**
  - `gt` (maggiore strettamente di)
  - `gte` (maggiore o uguale di)
  - `lt` (minore strettamente di)
  - `lte` (minore o uguale di)
  - `eq` (uguale a)
  - `neq` (non uguale a)

Gli operatori logici possono essere innestati tra di loro per creare filtri complessi. E" possibile confrontare tra loro due colonne o una colonna con una costante.

Esempio:

```
{
        "where": {
                "not": {
                        "and": [
                                {
                                        "or": [
                                                {"gt": { "left": "col1", "right":
↪"col2" }},
                                                {"eq": { "left": "col3", "right":␣
↪false }}
                                        ]
                                },
                                {
```

(continues on next page)

(continua dalla pagina precedente)

```
                                        "neq": { "left": "col4", "right": "'string'" }
                                }
                        ]
                }
        }
}
```

è equivalente a

```
SELECT *
FROM table
WHERE NOT (
                col1 > col2
         OR col3 == false
        AND col4 <> 'string'
)
```

Nota: effettuare l'escaping delle costanti testuali con `\"` o usare il delimitatore `'`

GROUP BY

Permette di indicare condizioni di raggruppamento per il calcolo di valori aggregati. La proprietà `groupBy` accetta come valore una lista non vuota di oggetti del tipo: `{"name": "col2"}`. Gli operatori di aggregazione attualmente supportati sono `min`, `max`, `count`, `sum` e `avg`.

Esempio:

```
{
        "select": [
                { "name": "col1" },
                { "name": "col2" },
                {
                        "max": { "name": "col3"}, "alias": "max_col3" }
                },
                {
                        "count": { "name": "*" }
                }
        ],
        "groupBy": [
                { "name": "col1" },
                { "name": "col2" }
        ]
}
```

è equivalente a

```
SELECT col1, col2, MAX(col3) AS max_col3, COUNT(*)
FROM table
GROUP BY col1, col2
```

HAVING

Permette di specificare clausole having se presente un raggruppamento. La proprietà `having` accetta come valore una lista non vuota di oggetti del tipo impiegato per esplicitare condizioni di filtro.

Esempio:

```
{
        "select": [
                { "name": "col1" },
                { "name": "col2" },
                {
                        "max": { "name": "col3", "alias": "max_col3" }
                },
                {
                        "count": { "name": "*" }
                }
        ],
        "groupBy": [
                { "name": "col1" },
                { "name": "col2" }
        ],
        "having": [
                {
                        "gt": { "left": "max_col3", "right": 50 }
                }
        ]
}
```

è equivalente a

```
SELECT col1, col2, MAX(col3) AS max_col3, COUNT(*)
FROM table
GROUP BY col1, col2
HAVING max_col3 > 50
```

LIMIT

Permette di specificare il numero massimo di elementi restituiti. La proprietà `limit` accetta un valore numerico.

Esempio:

```
{
        "limit": 5
}
```

è equivalente a

```
SELECT *
FROM table
LIMIT 5
```

CAPITOLO 4

Data science

Il Dataportal offre un ambiente integrato per svolgere task di data science come analisi esplorative, costruzione di grafici e dashboard e sviluppo di modelli statistici e di machine learning.

Gli strumenti di data analysis e data visualization messi a disposizione dal DAF sono i seguenti:

4.1 Jupyter Notebook: strumento per data science

4.1.1 Introduzione

Jupyter Notebook è un'applicazione Web open source che permette di creare e condividere documenti testuali interattivi, contenenti oggetti quali equazioni, grafici e codice sorgente eseguibile. Jupyter è diventato uno standard de-facto per data scientist perché offre la possibilità di realizzare, documentare e condividere analisi di dati all'interno di un framework che supporta[1]:

- operazioni di *data cleaning & trasformation*, simulazioni numeriche, modellazione statistica, machine learning e altro;
- l'esecuzione di applicazioni Scala e Python su piattaforme big data, grazie all'integrazione con Apache Spark.

A fronte di queste caratteristiche, Jupyter è stato scelto per essere uno degli strumenti messi a disposizione degli utenti del DAF. L'obiettivo è quello di favorire l'analisi dei dati presenti nel DAF stesso e la pubblicazione su Web dei relativi risultati.

Questa breve guida è pensata per fornire le informazioni basilari per configurare un notebook Jupyter utile per effettuare analisi sui dataset contenuti nel DAF utilizzando la capacità di calcolo di un cluster Spark. Pertanto, qualora si intenda eseguire analisi gestibili su un singolo laptop, si consiglia di scaricare i dataset sulla propria macchina e realizzare le analisi localmente. In tal caso si consideri che le analisi locali possono essere realizzate, oltre che con framework convenzionali quali R e Python (pandas + scikit-learn), anche con un"installazione locale di Jupyter Notebook.

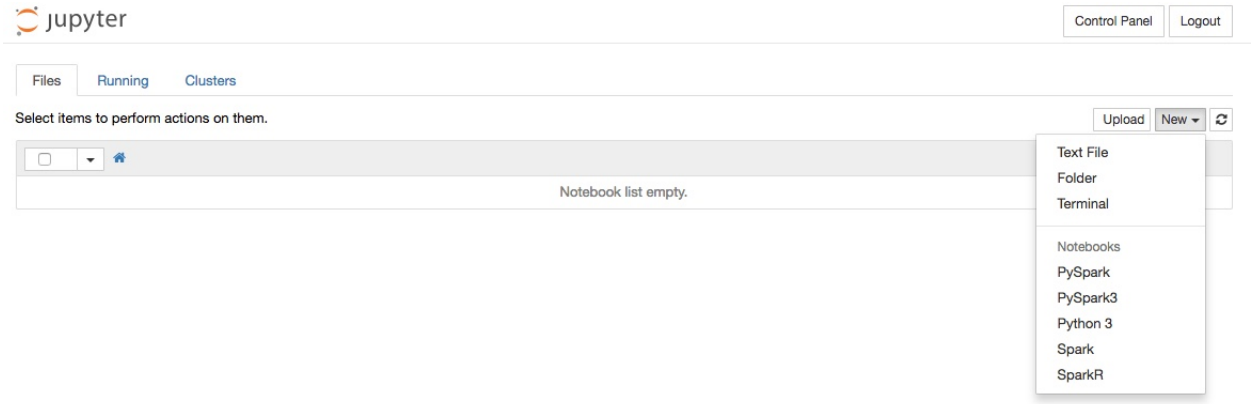
[1] Per una panoramica completa sulle funzionalità offerte da Jupyter e per l'elenco completo dei linguaggi di programmazione da esso supportati, far riferimento al sito ufficiale.

4.1.2 Creazione e configurazione di un notebook

Per accedere alla console Jupyter del DAF collegarsi a questo indirizzo. Una volta effettuato il login con le credenziali fornite per il Dataportal, si avrà accesso alla propria home nella quale è riportato l'elenco dei notebook associati all'utenza (al primo accesso l'elenco sarà vuoto).

Per iniziare una nuova analisi è necessario creare un nuovo notebook scegliendo l'ambiente di sviluppo (*kernel*). A tal fine cliccare sul pulsante "new" in alto a destra e selezionare una delle seguenti voci dal menù a tendina:

- "PySpark": Notebook Python 2 integrato con Spark
- "PySpark3": Notebook Python 3 integrato con Spark
- "Python 3": Notebook Python
- "Spark": Notebook Scala integrato con Spark
- "SparkR": Notebook R integrato con Spark



Una volta effettuata la selezione, Jupyter provvederà a creare un notebook del tipo selezionato e a visualizzarlo in una nuova scheda del browser.

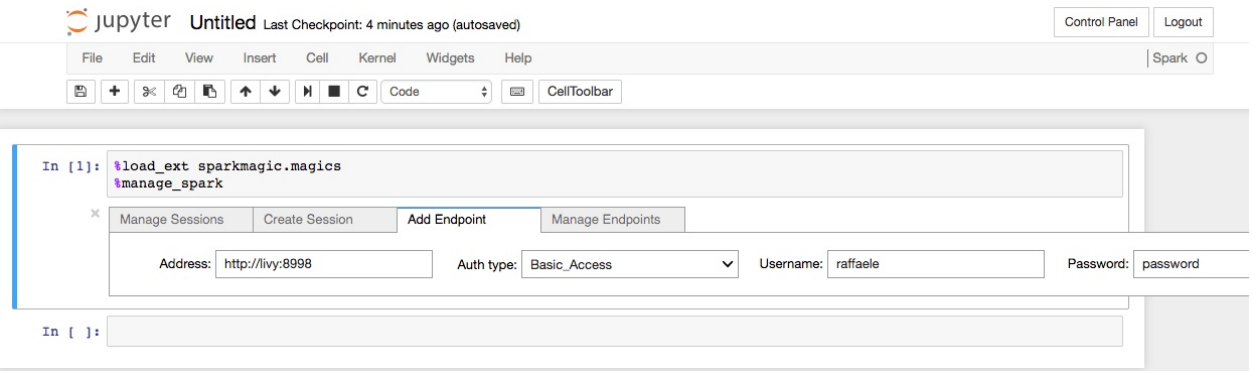
La prima operazione da eseguire è la configurazione della connessione verso il cluster Spark del DAF. A tal fine è necessario configurare l'accesso al server Livy, specificando l'endpoint del server, il tipo di autenticazione e le credenziali di accesso[2].

Per fare ciò è sufficiente copiare e incollare i seguenti comandi nel primo quadrante disponibile, ed eseguirli premendo la combinazione di tasti *[shift] + [invio]*:

```
%load_ext sparkmagic.magics
%manage_spark
```

L'esecuzione dei comandi produrrà l'apertura di un pannello come mostrato in figura:

[2] Al momento l'autenticazione presso il server è necessaria ogni volta che si crea un nuovo notebook. Nelle prossime release del DAF, sarà fornito uno script che permetterà di effettuare questa operazione solo una volta.



Cliccare nel tab "Add Endpoint" e inserire le seguenti informazioni:

- Address: http://livy:8998
- Auth type: Basic_Access
- Username: *<il vostro username>*
- Password: *<la vostra password>*

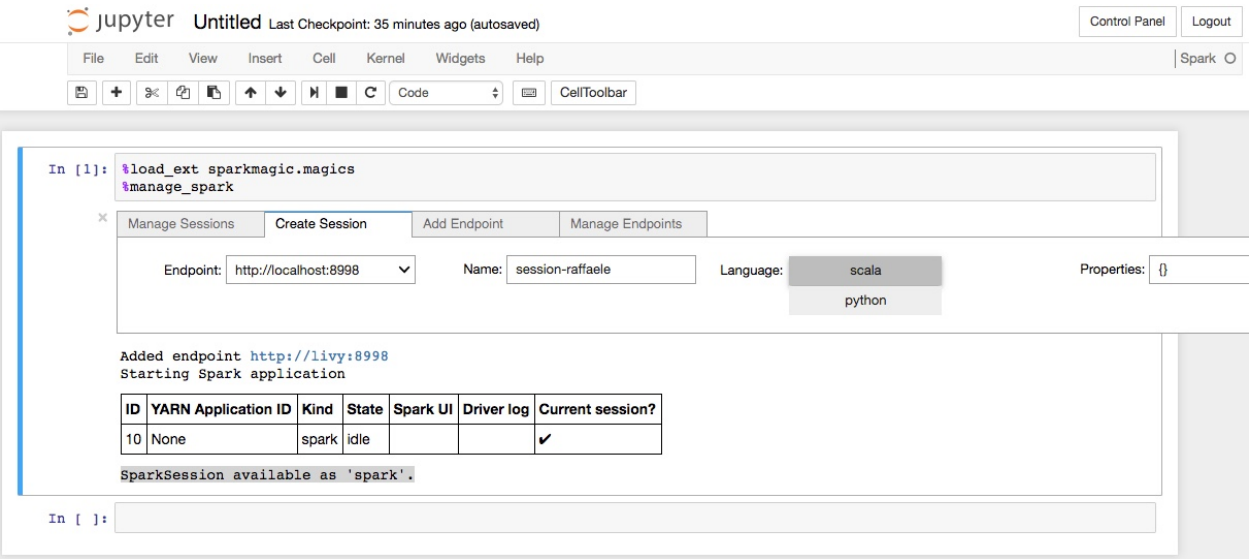
Cliccando il pulsante "Add Endpoint" la configurazione sarà aggiunta al notebook.

Dopo aver configurato correttamente la connessione al server Livy sarà possibile creare una *SparkSession*. Per fare ciò:

- selezionare il tab "Create Session";
- scegliere nel menu a tendina "Endpoint" l'endpoint appena configurato (http://livy:8998);
- dare un nome alla sessione che si sta creando valorizzando il campo "Name";
- selezionare il linguaggio che si intende usare nel notebook (ad esempio "scala");
- cliccare il pulsante "Create Session".

Il sistema impiegherà qualche decina di secondi per attivare il collegamento con il cluster Spark. Al termine dell'operazione il notebook è pronto per caricare dataset e realizzare analisi.

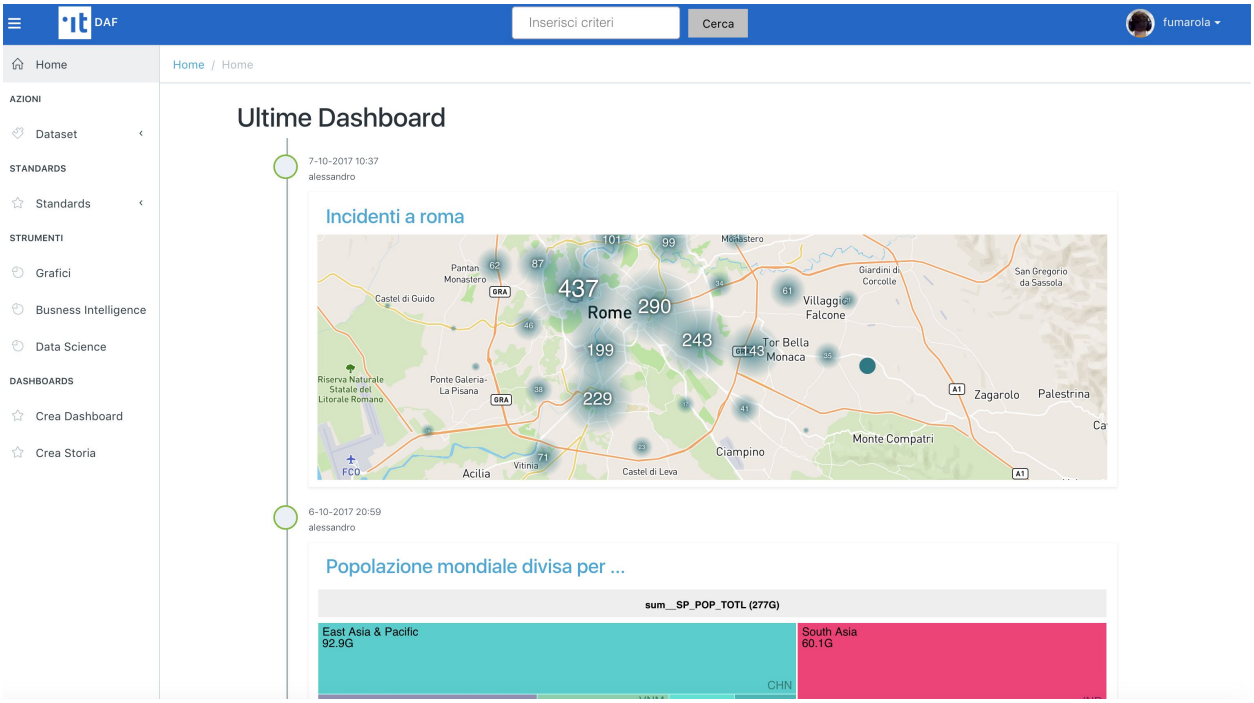
Se tutto è andato andato a buon fine, sarà visualizzata una tabella con le informazioni prodotte da YARN (il gestore delle risorse del cluster Spark) e la seguente frase conclusiva «SparkSession available as "spark".», così come mostrato nella figura successiva.



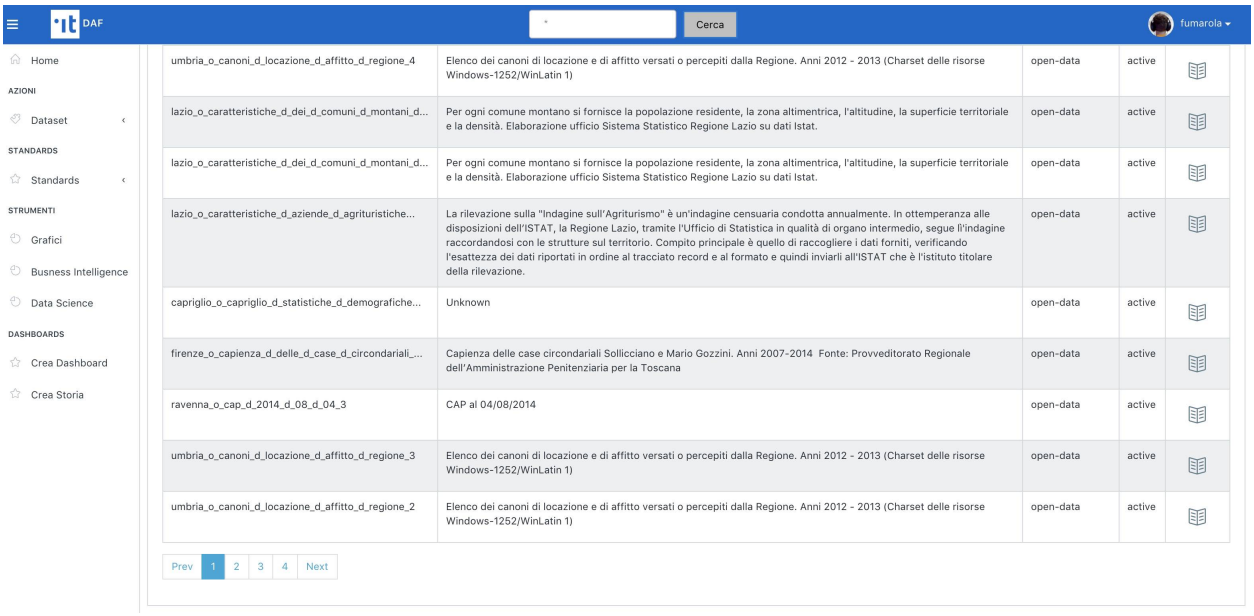
4.1.3 Caricare un dataset

Una volta attivata con successo una sessione Spark, l'ambiente è pronto per iniziare l'attività di analisi. La prima cosa da fare è selezionare il dataset da analizzare nel Dataportal.

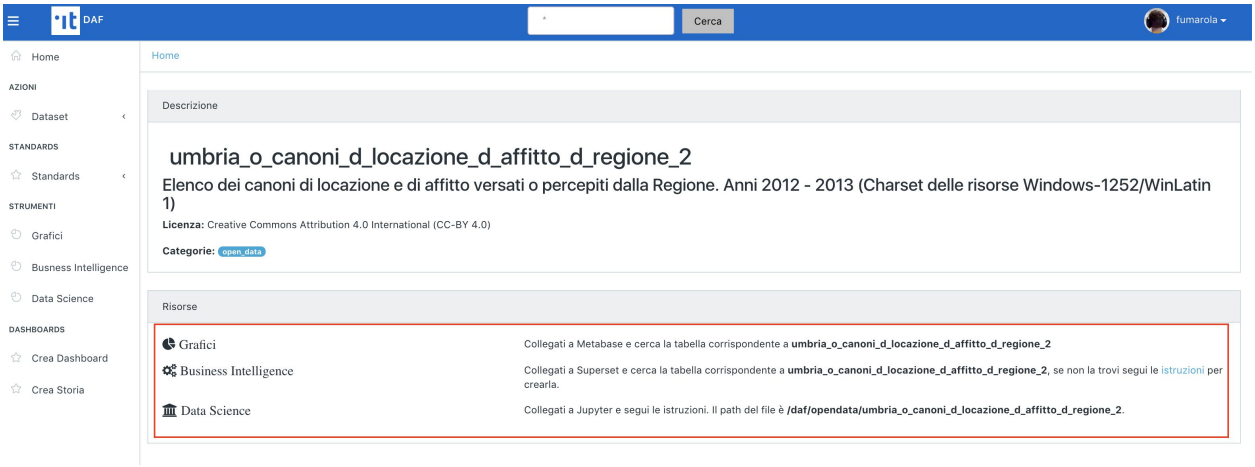
Nella barra di ricerca in alto al centro è possibile inserire del testo per fare ricerca dei dataset indicizzati.



In alternativa, inserendo "*" verranno visualizzati tutti i dataset indicizzati dal DAF.



Premendo l'icona a destra di ogni risultato si accede alla scheda informativa del dataset (immagine di seguito), dove potete individuare le *uri* per accedere al dataset selezionato utilizzando gli strumenti di data science del DAF.



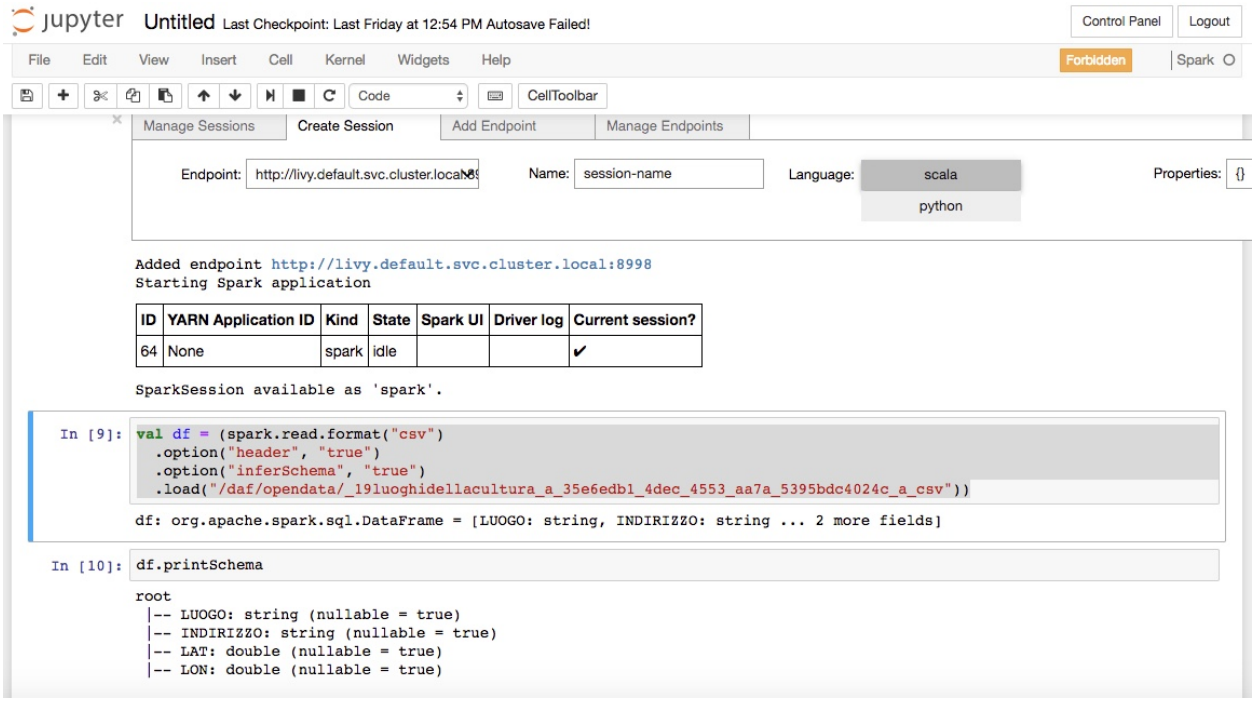
Per creare uno Spark DataFrame con il dataset che si vuole analizzare, procedendo come segue:

- Identificare il dataset nel Dataportal e tenere traccia del path e del formato indicati nella scheda informativa (vedi sopra). Nel caso in esempio, si è scelto il dataset "abitazioni_d_ritirate_d_e_d_ultimate_d_per_d_numero_d_stanze_d_anno_d_2013_d_4e3f6d25_63".
- Eseguire nel notebook il seguente comando, che chiede a Spark di creare un nuovo DataFrame contenente il dataset specificato:

```
val df = (spark.read.format("csv")
     .option("header", "true")
     .option("inferSchema", "true")
     .load("/daf/opendata/abitazioni_d_ritirate_d_e_d_ultimate_d_per_d_numero_d_
↪stanze_d_anno_d_2013_d_4e3f6d25_63")
)
```

- [Optional] Eseguite il seguente comando per ottenere lo schema del dataframe:

```
df.printSchema
```



A questo punto il notebook è pronto all'uso. Per approfondimenti sull'utilizzo di Spark e di Jupyter Notebook si rimanda alla documentazione ufficiale.

4.1.4 Risorse utili

- http://spark.apache.org/docs/latest/sql-programming-guide.html
- http://jupyter.org/

Note

4.2 Superset: strumento per data visualization

Superset è un'applicazione Web open-source di Business Intelligence, sviluppata da AirBnB, con cui è possibile creare grafici ("slices" nel gergo di Superset), dashboard ed eseguire query SQL. Superset è stato integrato con il DAF per offrire agli utenti la possibilità di creare le proprie analisi e condividerle con la community in modalità self-service.

Questa sezione è pensata per fornire indicazioni utili alla creazione di tabelle e grafici. Per approfondimenti sulle funzionalità e sulle modalità d'uso di Superset, si rimanda alle documentazione suggerita nella sezione "Risorse Utili" di questa guida.

4.2.1 Utilizzare Superset su un dataset di interesse

Per accedere a Superset direttamente da dataset presente all'interno del dataportal è necessario: * accedere alla scheda del dataset * cliccare sul link della sezione "Superset" dal tab "Strumenti" Verrai automaticamente indirizzato su Superset nella modalità di creazione di un widget

4.2.2 Proprietà del dataset

Prima di iniziare a costruire degi slicer è buona regola controllare che i campi del dataset e le metriche proposte siano conformi al tipo di informazione che si intende comunicare.

Ogni dataset è definito da un elenco di:

- colonne: dati dalle misure quantitative (es: numero di abitanti, reddito, percentuale di disoccupazione) e dalle dimensioni di analisi (es.nome comune, fascia di reddito, tipologia di azienda). Le dimensioni consentono di determinare il metodo di raggruppamento dei dati provenienti dalla misura (es. mostrami la media della misura «reddito» per fascia di reddito e per comune)
- metriche: Superset mette a disposizione per ogni campo diverse metriche che è possibile avere a disposizione inserendo la spunta sulla metrica di interesse. Stà all'utente valutare quali inserire e su quali campi abilitarle.

Colonne: .. image:: img_superset/checkMetriche.png

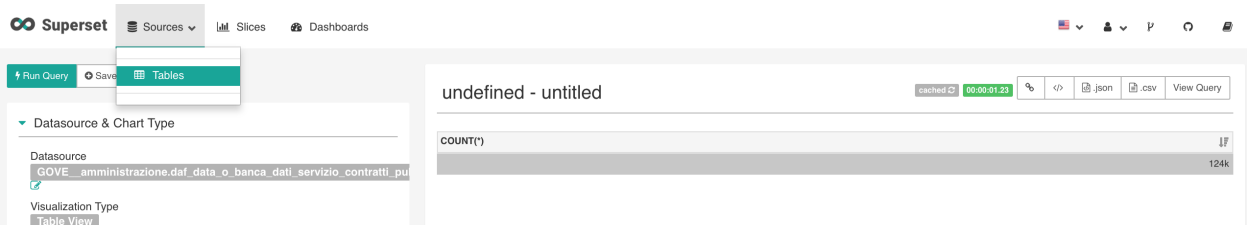
Metriche: .. image:: img_superset/metriche.png

4.2.3 Configurazione di un dataset

Un utente può accedere alla homepage di Superset direttamente dal Dataportal grazie al menù laterale sinistro della sezione privata del Dataportal (sezione: Toolkit - Strumenti - Superset) oppure atterrare comodamente alla schermata di crezione di uno slicer attraverso la scheda del dataset mostrata in precedenza.

Una volta entrati in Superset dal menù Strumenti, si accede alla lista delle dashboard create nell'ambito dell'organizzazione a cui afferisce l'utente che ha effettuato l'accesso e della community. Superset adotta infatti un approccio self-service: questo significa che ciascun utente, da un lato, potrà provvedere alla creazione dei grafici e delle dashboard; dall'altro potrà riutilizzare le tabelle, i grafici e le dashboard creati dalla community. Chiaramente quest'ultima possibilità impone anche una certa disciplina nell'uso dello strumento, per evitare di interferire con le analisi fatte dagli altri utenti dell'organizzazione.

Il primo step nell'uso di Superset consiste nell'individuazione dei dataset oggetto di analisi. Nella homepage selezionare la voce "Tables" dal menu a tendina "Sources".



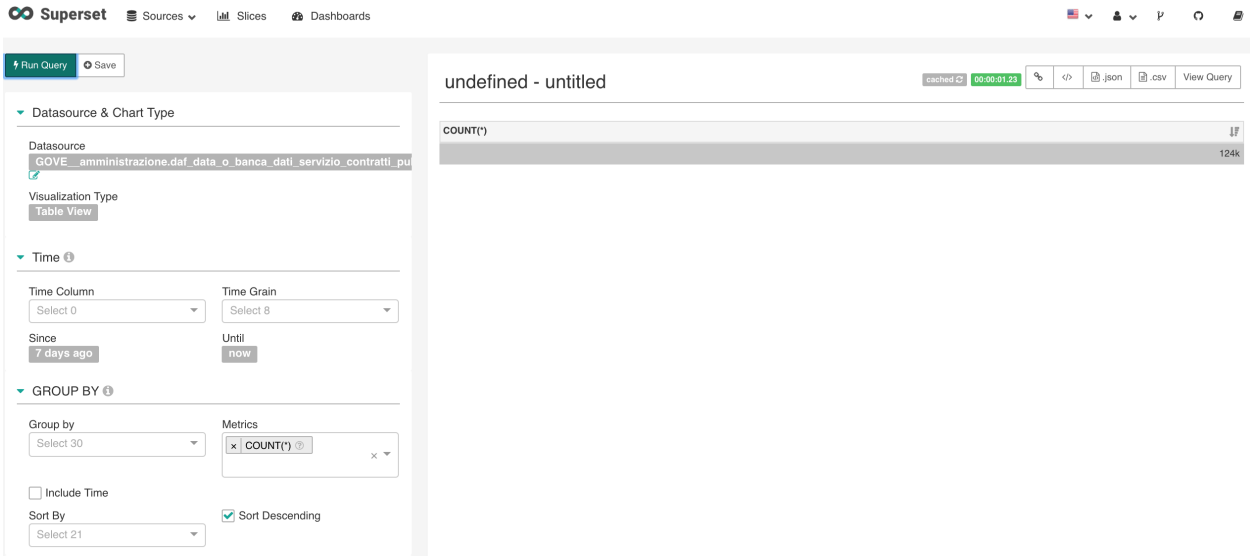
Scorrendo la lista è possibile trovare e selezionare la tabella che contiene il dataset di interesse, anche aiutandosi, all'occorrenza, con i filtri attivabili cliccando sul bottone in alto a destra "Add Filter".

Per accedere ai campi della tabella cliccare sull'icona «Edit»

Dopo aver selezionato la tabella verrà presentato un elenco in cui ogni riga è un campo del dataset. Per ogni campo, se necessario è possibile ridefinire le proprietà elencate e creare nuovi campi derivati. Per apportare le modifiche cliccare sull'icona «edit» a sinistra del campo

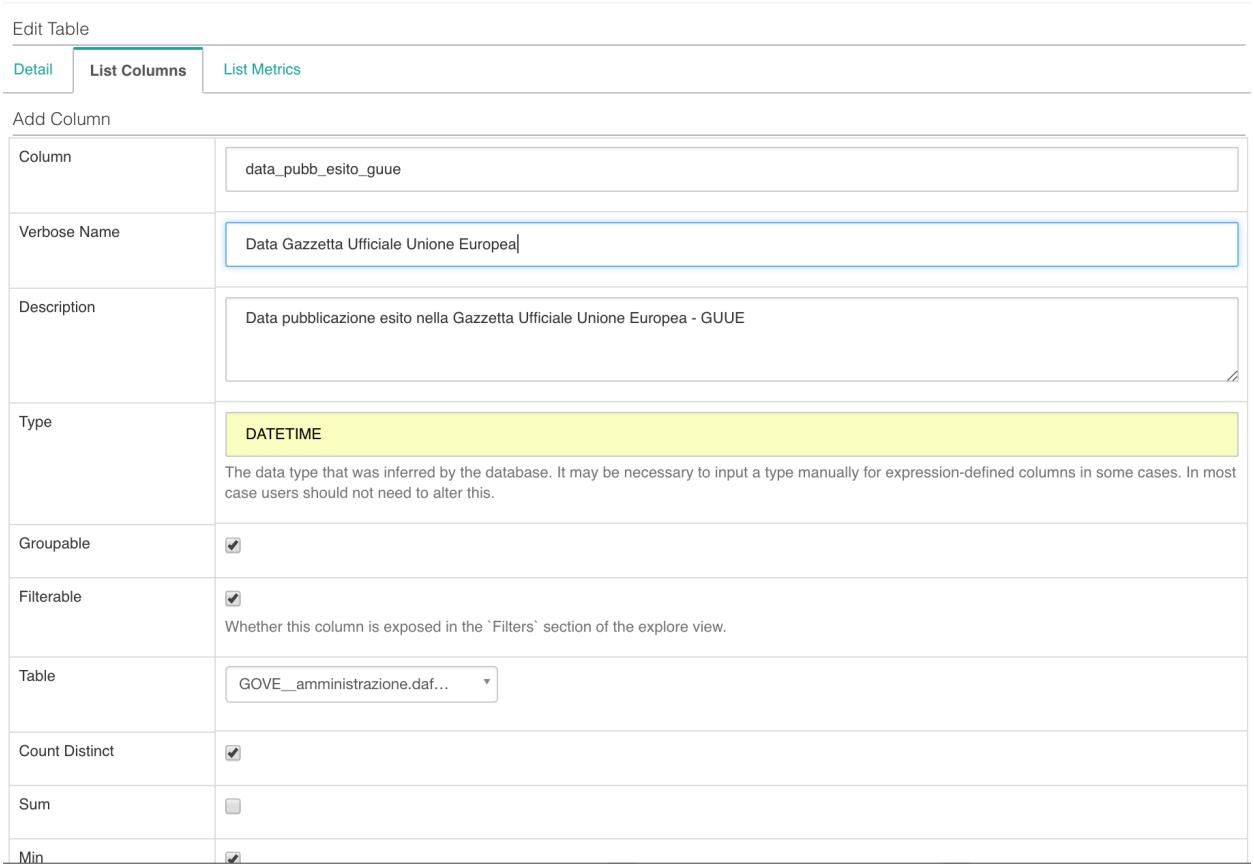
4.2.4 Controllo del dataset

Se si è in possesso di un'utenza editor e si è appena caricato il dataset sul dataportal, può essere utile controllare rapidamente se il numero di record caricati corrisponde a quelli attesi. E" possibile effettuare questo primo test selezionando nel campo Metrics "COUNTS (*)" e cliccando su "Run query".



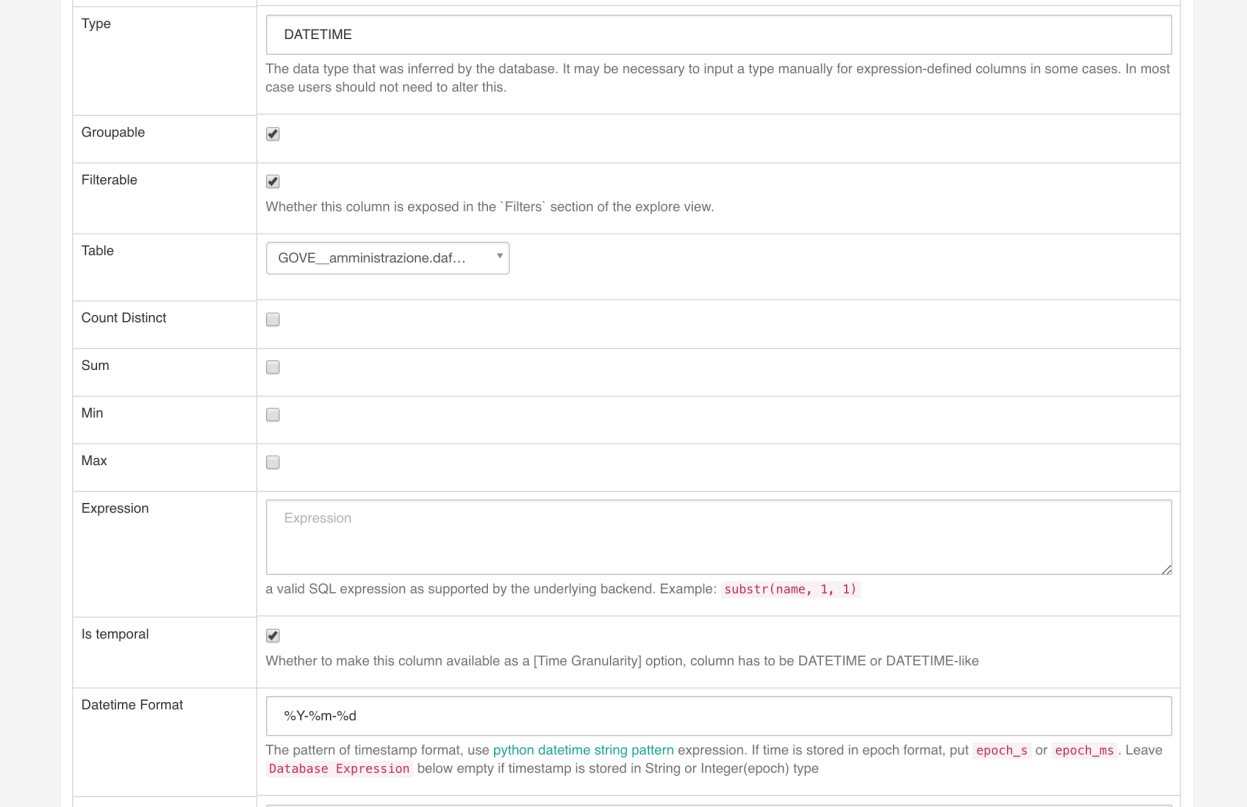
4.2.5 Definizione e modifica di un campo

Cliccando sul nome di un campo verrà presentata una scheda con le caratteristiche del campo. E' possibile indicare un "Verbose name" che sarà l'etichetta mostrata quando il campo verrà usato nei widget, una "description" che sarà mostrata come tooltip.



4.2.6 Definizione di un Campo con formato data

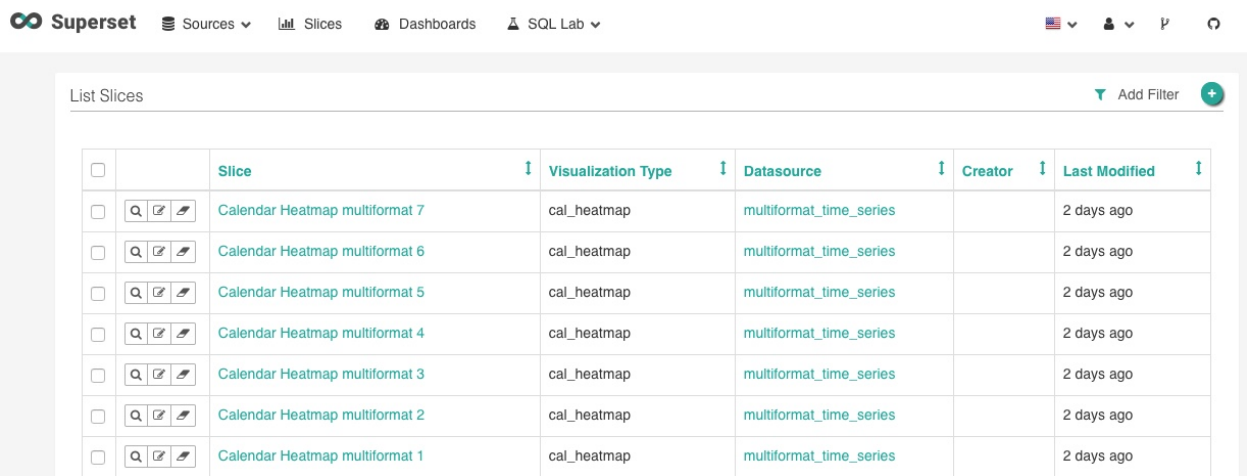
Per definire un campo che indica una data e che nella tabella originaria è impostato a STRING o con altre tipologie di dati è sufficiente cliccare sul campo di interesse e modificare il campo TYPE a DATETIME esplicitando il Datetime format utilizzato nel dataset.



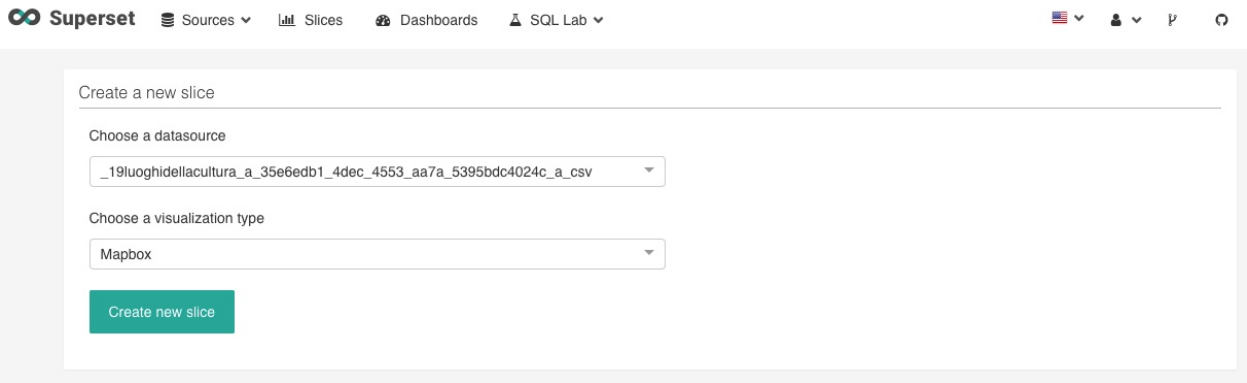
4.2.7 Creazione di un grafico ("Slice")

Il prossimo step è creare un grafico con il dataset importato. Come appena visto, utilizzando "SQL Editor", il dataset contiene informazioni georeferenziate sui luoghi della cultura di Matera, per cui potrebbe essere una buona idea graficarle su una mappa utilizzando l'integrazione con MapBox.

A tal fine, cliccare sul menu "Slices" posto nella barra in alto: comparirà l'elenco di slices già creati. Anche in questo caso è possibile utilizzare filtri per cercare il grafico di interesse. Per crearne uno nuovo, cliccare il bottone "+" in alto a destra.



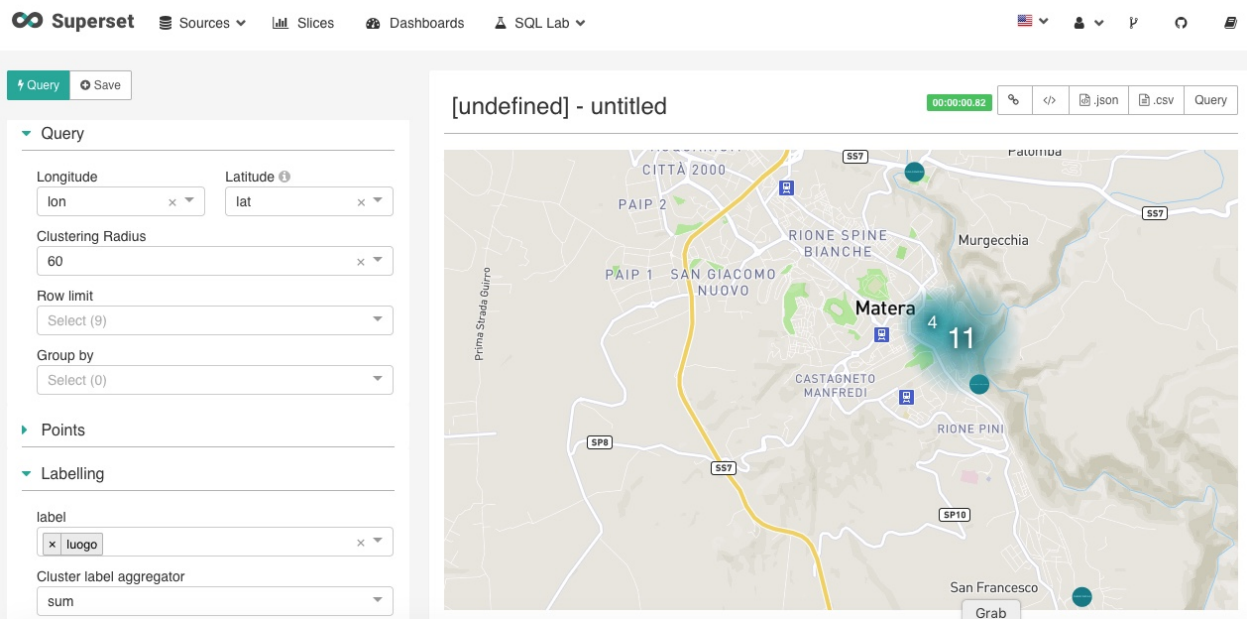
Comparirà un menu dove si dovrà indicare la tabella da cui prendere i dati ("datasource"), e il tipo di visualizzazione da utilizzare. Per graficare su una mappa selezionare "Matbox". Infine cliccare sul bottone in basso a sinistra "Create New Slice".



L'ultimo step consiste nel configurare il grafico. Nel caso in esempio vanno impostate le seguenti informazioni:

- latitudine e longitudine
- "label", ovvero il testo che comparirà all'interno del punto disegnato sulla mappa
- "Viewpoint" con le informazioni di latitudine, longitudine e zoom di default che verranno utilizzate per la visualizzazione iniziale.

Infine, eseguire la query che aggiorna il grafico, cliccando sul bottone "Query" in alto a sinistra, e salvare la "slice" cliccando sul pulsante "Save".

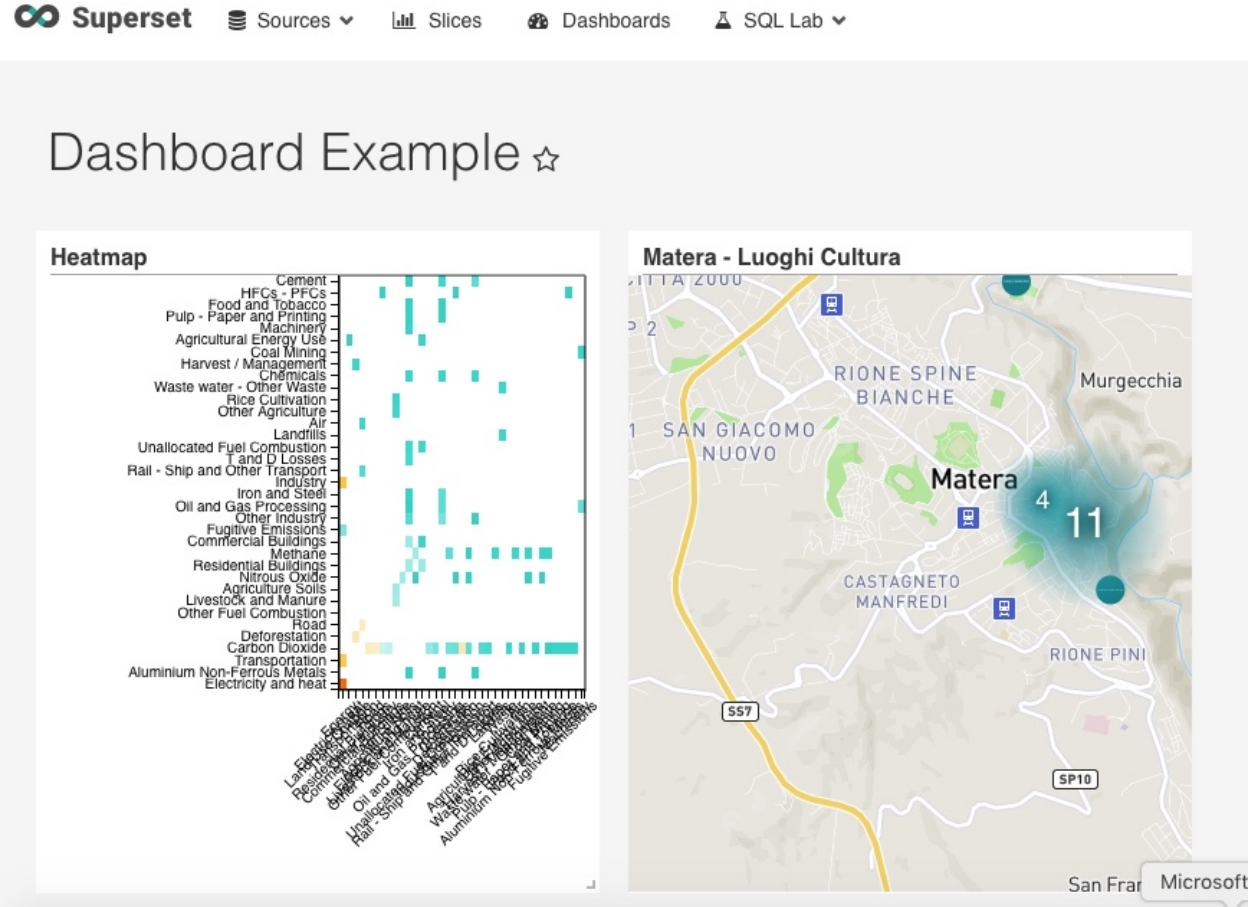


4.2.8 Creazione di una dashboard

Le dashboard sono aggregazioni di "slices" utili a tenere su un unico pannello i grafici di interesse. Per creare una dashboard in Superset occorre:

- Cliccare sul menu "Dashboard" in alto

- Cliccare sul pulsante "+" in alto a sinistra
- Compilare i campi con le informazioni utili per la dashboard, come "Title", "Slug"(per rendere richiamabile tramite un URL la dashboard), "Slices" (in cui elencare le slices che si vuole utilizzare nella dashboard; nel caso in esempio abbiamo selezionato "Matera - Luoghi Cultura" e "Heatmap") e "Owners" (in cui indicare chi può contribuire alla dashboard).
- Cliccare il bottone "Save" in basso a sinistra.



4.2.9 Risorse utili

Superset è un tool molto potente e, conseguentemente, complesso. Per utilizzare in pieno le sue funzionalità si rimanda a guide specifiche sul tema, di cui si riportano alcuni riferimenti non esaustivi.

- https://superset.incubator.apache.org/
- http://de.straba.us/2017/08/15/creare-dashboard-con-superset/

4.3 Metabase: strumento di data visualization

Nota: Sezione in fase di creazione.