NoSQL approaches in GnpIS

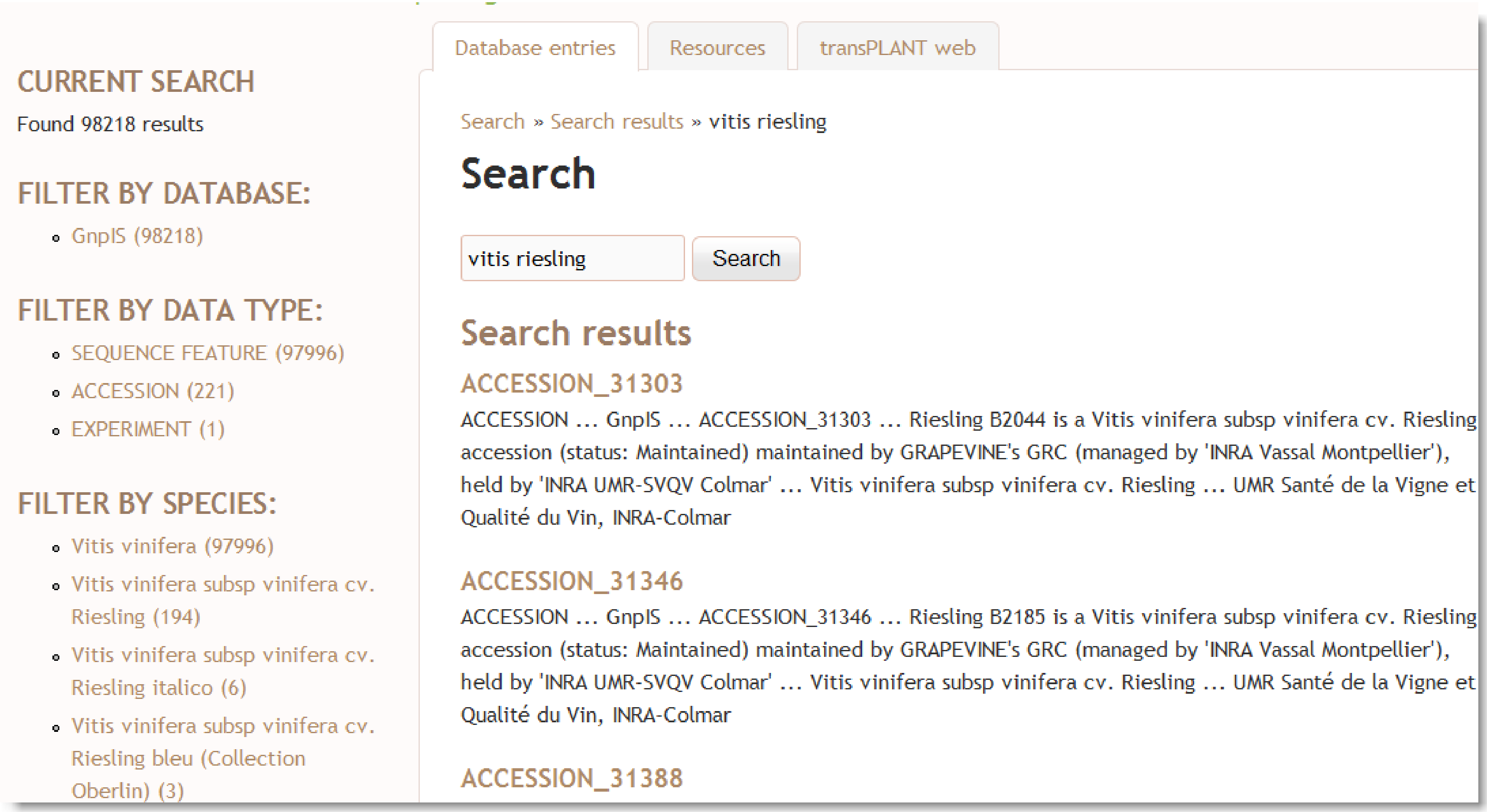
Elasticsearch @ URGI Feedback

Phenomic Data Management & Sharing

- Data Discovery
  - Links to datasets, by metadata
  - Loose integration (keywords)
  - Data model
    - Simple
    - Easy to implement and feed
  - Ex:
    - GnpIS Google-like search
    - transPLANT search
    - WheatIS
    - IFB-Elixir
- Data set building, Data Mining
  - Combine data from different sources
  - Strong integration
  - Data model
    - Rich
    - Complex to model, implement and feed
  - Ex: GnpIS.Ephesis

Data discovery system

- GnpIS, transPLANT, WheatIS searches
- Google like, full text, filters
- http://www.transplantdb.eu



CURRENT SEARCH

Found 98218 results

FILTER BY DATABASE:

- GnpIS (98218)

FILTER BY DATA TYPE:

- SEQUENCE FEATURE (97996)
- ACCESSION (221)
- EXPERIMENT (1)

FILTER BY SPECIES:

- Vitis vinifera (97996)
- Vitis vinifera subsp vinifera cv. Riesling (194)
- Vitis vinifera subsp vinifera cv. Riesling italico (6)
- Vitis vinifera subsp vinifera cv. Riesling bleu (Collection Oberlin) (3)

Search » Search results » vitis riesling

Search

vitis riesling

Search results

ACCESSION_31303

ACCESSION … GnpIS … ACCESSION_31303 … Riesling B2044 is a Vitis vinifera subsp vinifera cv. Riesling accession (status: Maintained) maintained by GRAPEVINE's GRC (managed by 'INRA Vassal Montpellier'), held by 'INRA UMR-SVQV Colmar' … Vitis vinifera subsp vinifera cv. Riesling … UMR Santé de la Vigne et Qualité du Vin, INRA-Colmar

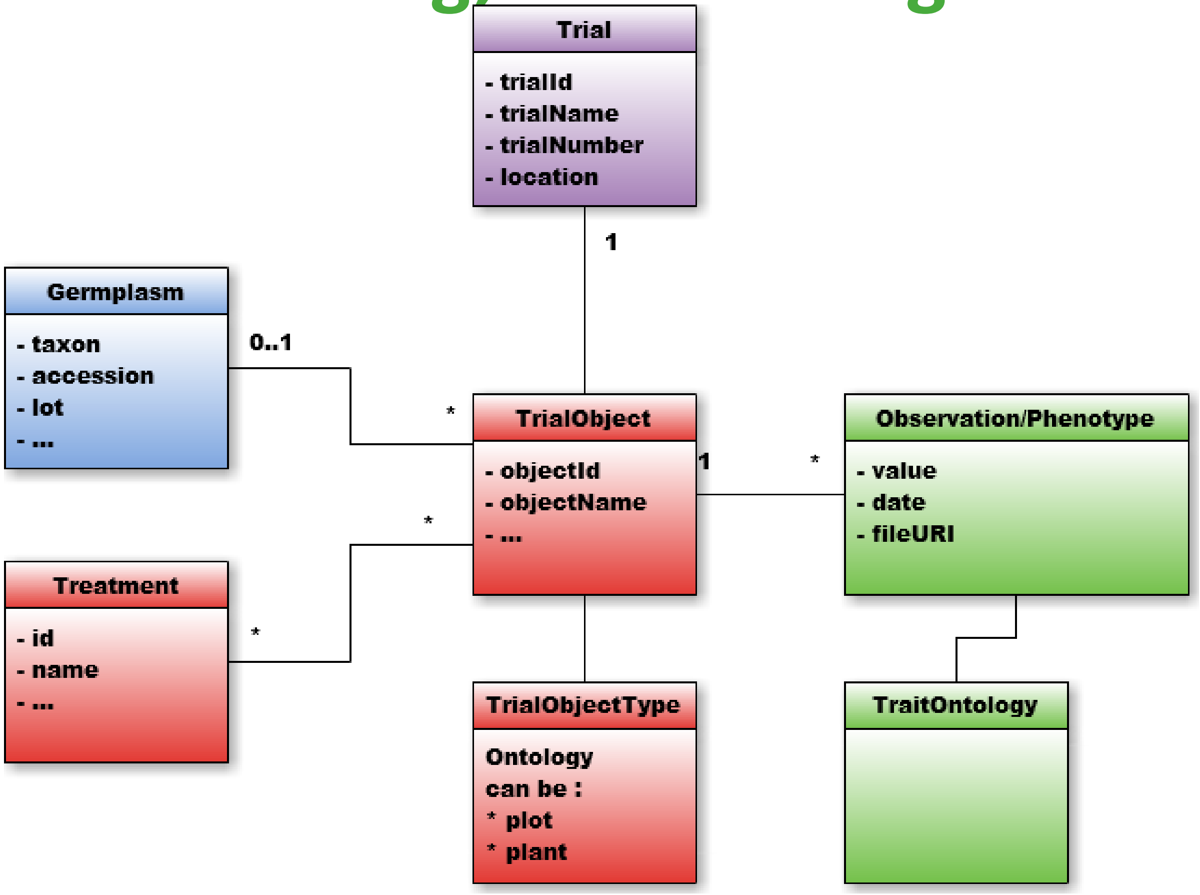
ACCESSION_31346

ACCESSION … GnpIS … ACCESSION_31346 … Riesling B2185 is a Vitis vinifera subsp vinifera cv. Riesling accession (status: Maintained) maintained by GRAPEVINE's GRC (managed by 'INRA Vassal Montpellier'), held by 'INRA UMR-SVQV Colmar' … Vitis vinifera subsp vinifera cv. Riesling … UMR Santé de la Vigne et Qualité du Vin, INRA-Colmar

ACCESSION_31388

Data Set Building, Data Mining

- Phenotyping data
- Conceptual model



Trial
- trialId
- trialName
- trialNumber
- location

Germplasm
- taxon
- accession
- lot
- ...

TrialObject
- objectId
- objectName
- ...

Observation/Phenotype
- value
- date
- fileURI

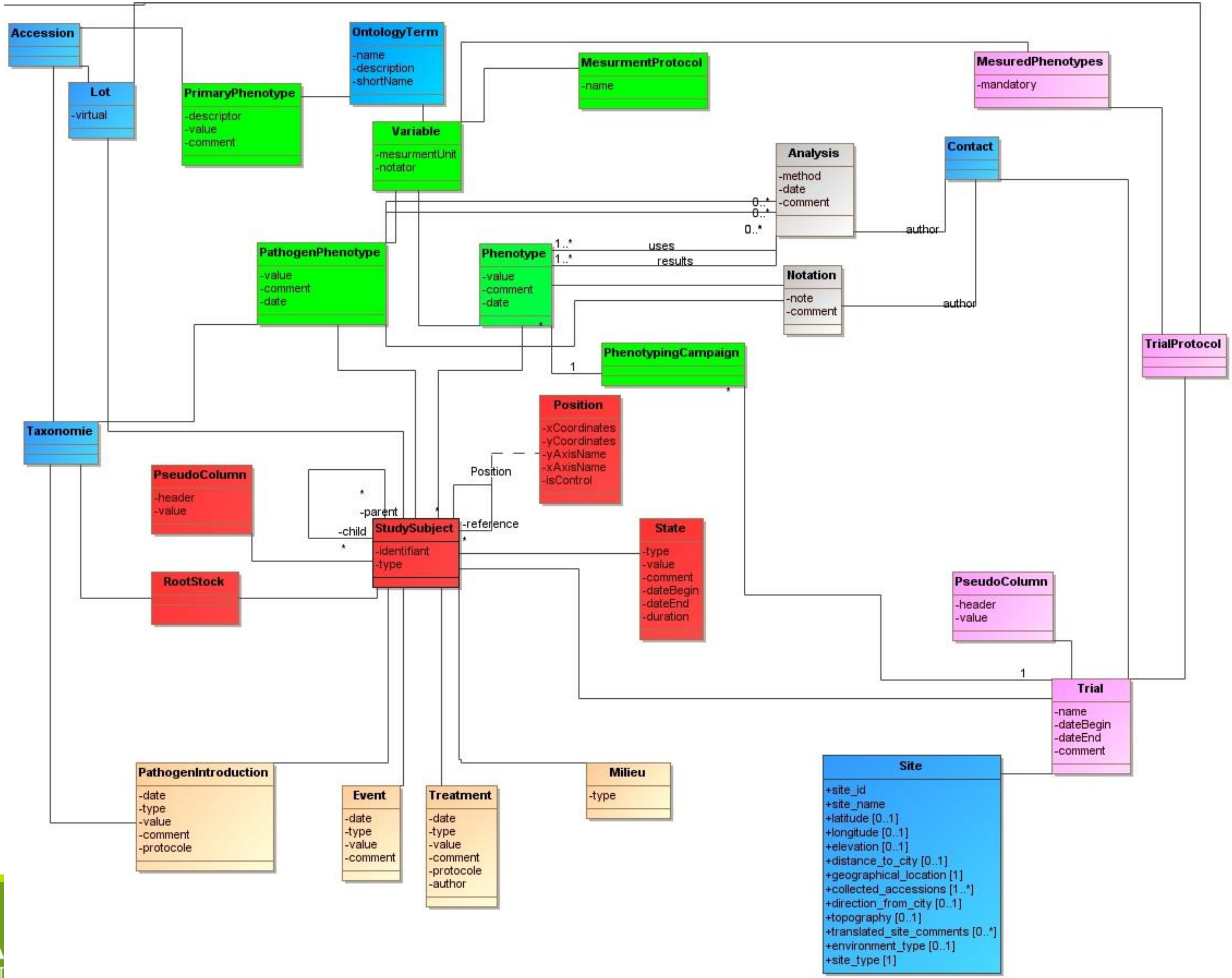
Treatment
- id
- name
- ...

TrialObjectType

Ontology
can be :
* plot
* plant

TraitOntology

GnpIS.Ephesis Implementation



Accession

Lot
-virtual

PrimaryPhenotype
-descriptor
-value
-comment

OntologyTerm
-name
-description
-shortName

Variable
-mesurmentUnit
-notator

MesurmentProtocol
-name

MesuredPhenotypes
-mandatory

Analysis
-method
-date
-comment

Contact

PathogenPhenotype
-value
-comment
-date

Phenotype
-value
-comment
-date

Notation
-note
-comment

PhenotypingCampaign

TrialProtocol

Taxonomie

Position
-xCoordinates
-yCoordinates
-yAxisName
-xAxisName
-isControl

PseudoColumn
-header
-value

StudySubject
-identifiant
-type

State
-type
-value
-comment
-dateBegin
-dateEnd
-duration

RootStock

PseudoColumn
-header
-value

Trial
-name
-dateBegin
-dateEnd
-comment

PathogenIntroduction
-date
-type
-value
-comment
-protocole

Event
-date
-type
-value
-comment

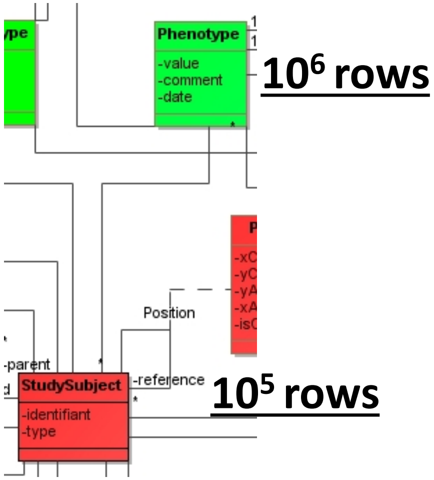
Treatment
-date
-type
-value
-comment
-protocole
-author

Milieu
-type

Site
+site_id
+site_name
+latitude [0..1]
+longitude [0..1]
+elevation [0..1]
+distance_to_city [0..1]
+geographical_location [1]
+collected_accessions [1..*]
+direction_from_city [0..1]
+topography [0..1]
+translated_site_comments [0..*]
+environment_type [0..1]
+site_type [1]

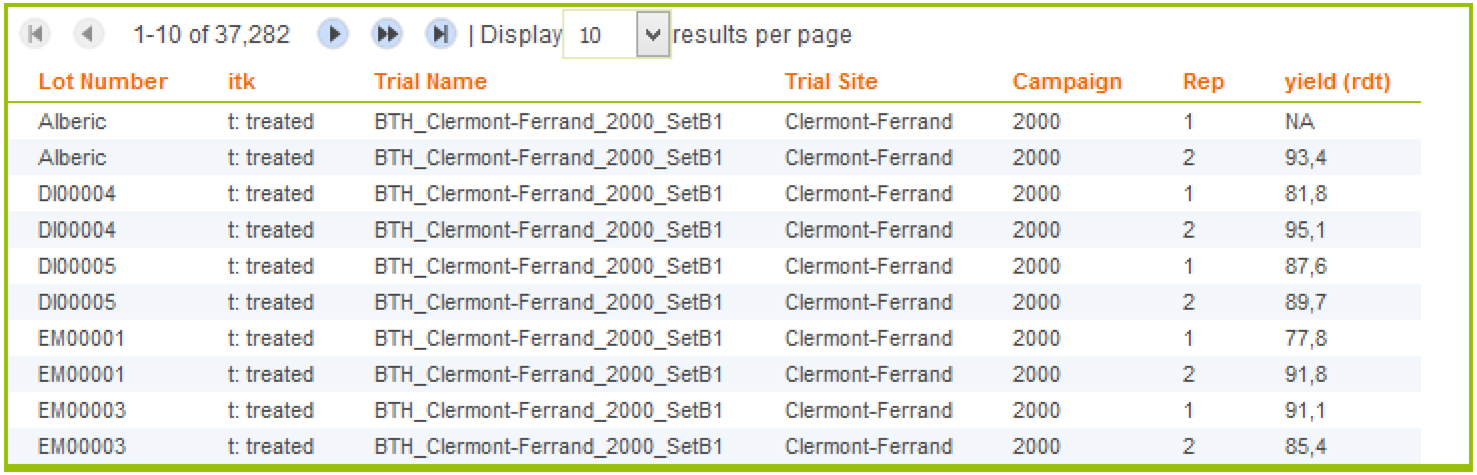
GnpIS.Ephesis V1 Performances

- StudySubject **join** Phenotype
  - Too slow, too big
- Solution
  - Denormalisation, agregation



10^6 rows

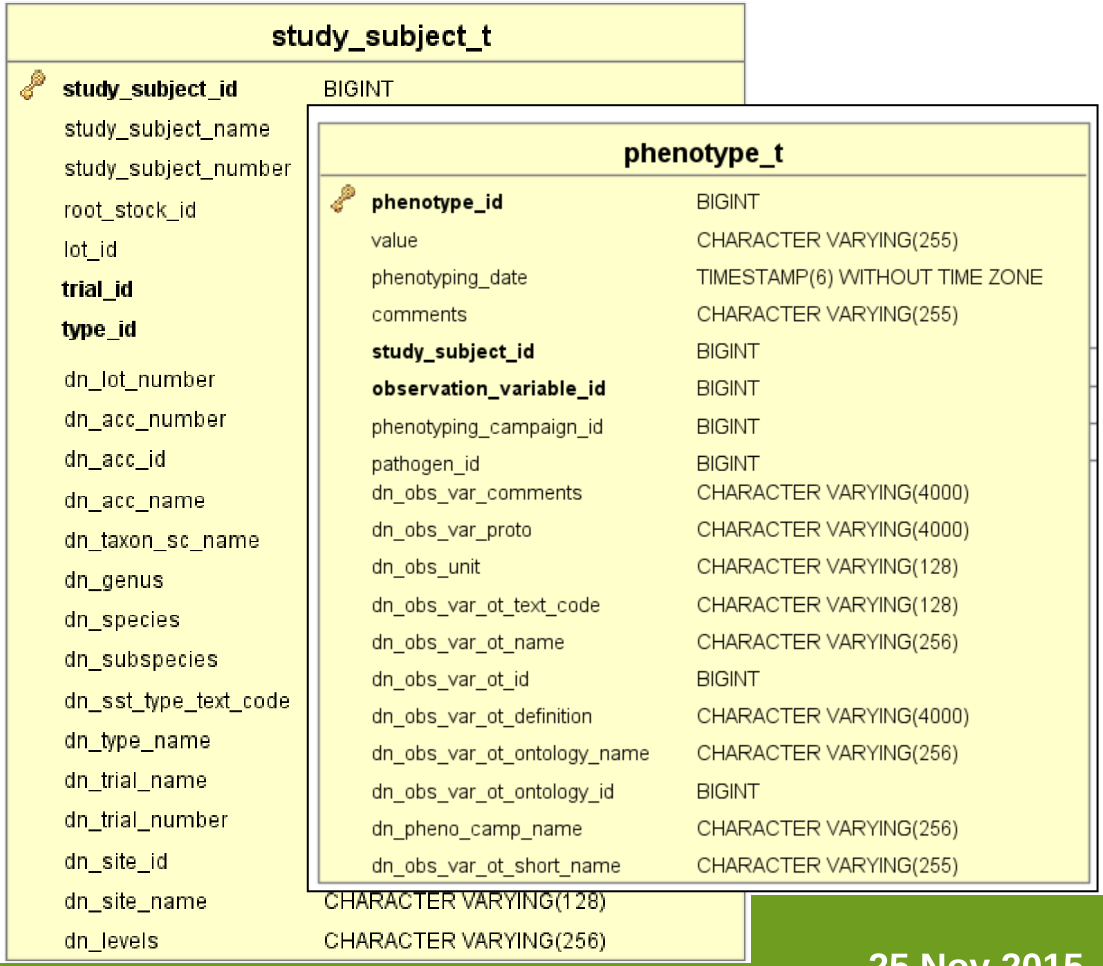
10^5 rows



Lot Number	itk	Trial Name	Trial Site	Campaign	Rep	yield (rdt)
Alberic	t: treated	BTH_Clermont-Ferrand_2000_SetB1	Clermont-Ferrand	2000	1	NA
Alberic	t: treated	BTH_Clermont-Ferrand_2000_SetB1	Clermont-Ferrand	2000	2	93,4
DI00004	t: treated	BTH_Clermont-Ferrand_2000_SetB1	Clermont-Ferrand	2000	1	81,8
DI00004	t: treated	BTH_Clermont-Ferrand_2000_SetB1	Clermont-Ferrand	2000	2	95,1
DI00005	t: treated	BTH_Clermont-Ferrand_2000_SetB1	Clermont-Ferrand	2000	1	87,6
DI00005	t: treated	BTH_Clermont-Ferrand_2000_SetB1	Clermont-Ferrand	2000	2	89,7
EM00001	t: treated	BTH_Clermont-Ferrand_2000_SetB1	Clermont-Ferrand	2000	1	77,8
EM00001	t: treated	BTH_Clermont-Ferrand_2000_SetB1	Clermont-Ferrand	2000	2	91,8
EM00003	t: treated	BTH_Clermont-Ferrand_2000_SetB1	Clermont-Ferrand	2000	1	91,1
EM00003	t: treated	BTH_Clermont-Ferrand_2000_SetB1	Clermont-Ferrand	2000	2	85,4

GnpIS.Ephesis V2 Performances

- Study Subject
  - No join
- Phenotype: select N+1
- Ok now
- Not for the Future



study_subject_t

study_subject_id	BIGINT
study_subject_name	
study_subject_number	
root_stock_id	
lot_id	
trial_id	
type_id	
dn_lot_number	
dn_acc_number	
dn_acc_id	
dn_acc_name	
dn_taxon_sc_name	
dn_genus	
dn_species	
dn_subspecies	
dn_sst_type_text_code	
dn_type_name	
dn_trial_name	
dn_trial_number	
dn_site_id	
dn_site_name	CHARACTER VARYING(128)
dn_levels	CHARACTER VARYING(256)

phenotype_t

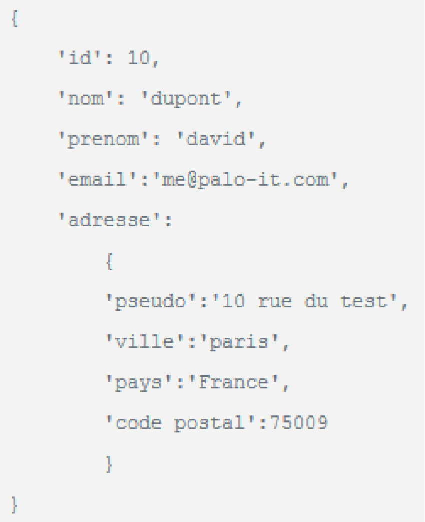
phenotype_id	BIGINT
value	CHARACTER VARYING(255)
phenotyping_date	TIMESTAMP(6) WITHOUT TIME ZONE
comments	CHARACTER VARYING(255)
study_subject_id	BIGINT
observation_variable_id	BIGINT
phenotyping_campaign_id	BIGINT
pathogen_id	BIGINT
dn_obs_var_comments	CHARACTER VARYING(4000)
dn_obs_var_proto	CHARACTER VARYING(4000)
dn_obs_unit	CHARACTER VARYING(128)
dn_obs_var_ot_text_code	CHARACTER VARYING(128)
dn_obs_var_ot_name	CHARACTER VARYING(256)
dn_obs_var_ot_id	BIGINT
dn_obs_var_ot_definition	CHARACTER VARYING(4000)
dn_obs_var_ot_ontology_name	CHARACTER VARYING(256)
dn_obs_var_ot_ontology_id	BIGINT
dn_pheno_camp_name	CHARACTER VARYING(256)
dn_obs_var_ot_short_name	CHARACTER VARYING(255)

NoSQL Modelisation

- http://blog.palo-it.com/2013/06/10/modelisation-dun-schema-dune-base-de-donnees-nosql/
- NoSQL Document
- From Database to documents
  - **Denormalisation and agregation**
  - **3 methods**

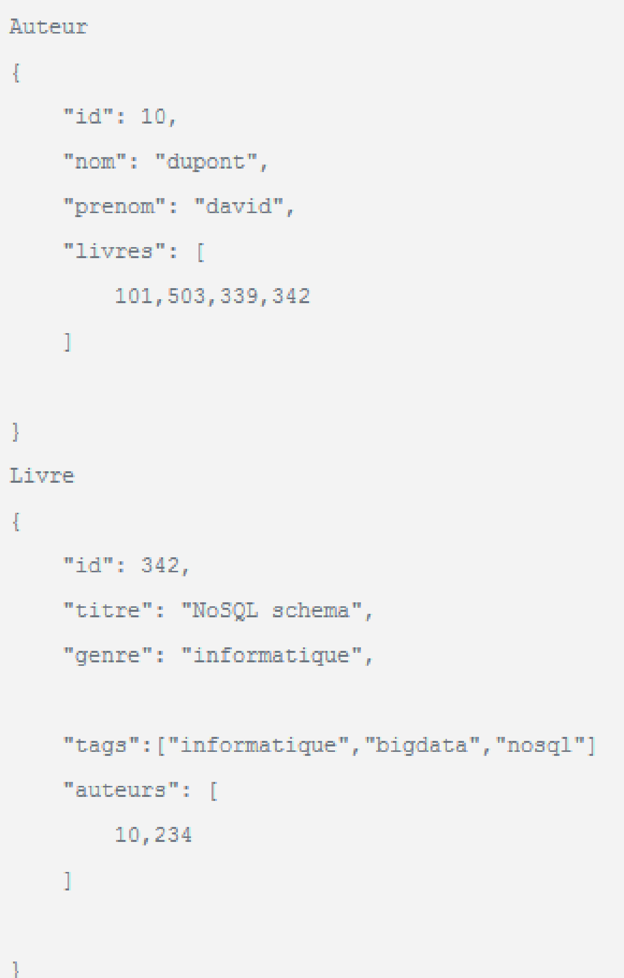
Agregations

- Document interweaving



```
{
    'id': 10,
    'nom': 'dupont',
    'prenom': 'david',
    'email':'me@palo-it.com',
    'adresse':
        {
        'pseudo':'10 rue du test',
        'ville':'paris',
        'pays':'France',
        'code postal':75009
        }
}
```

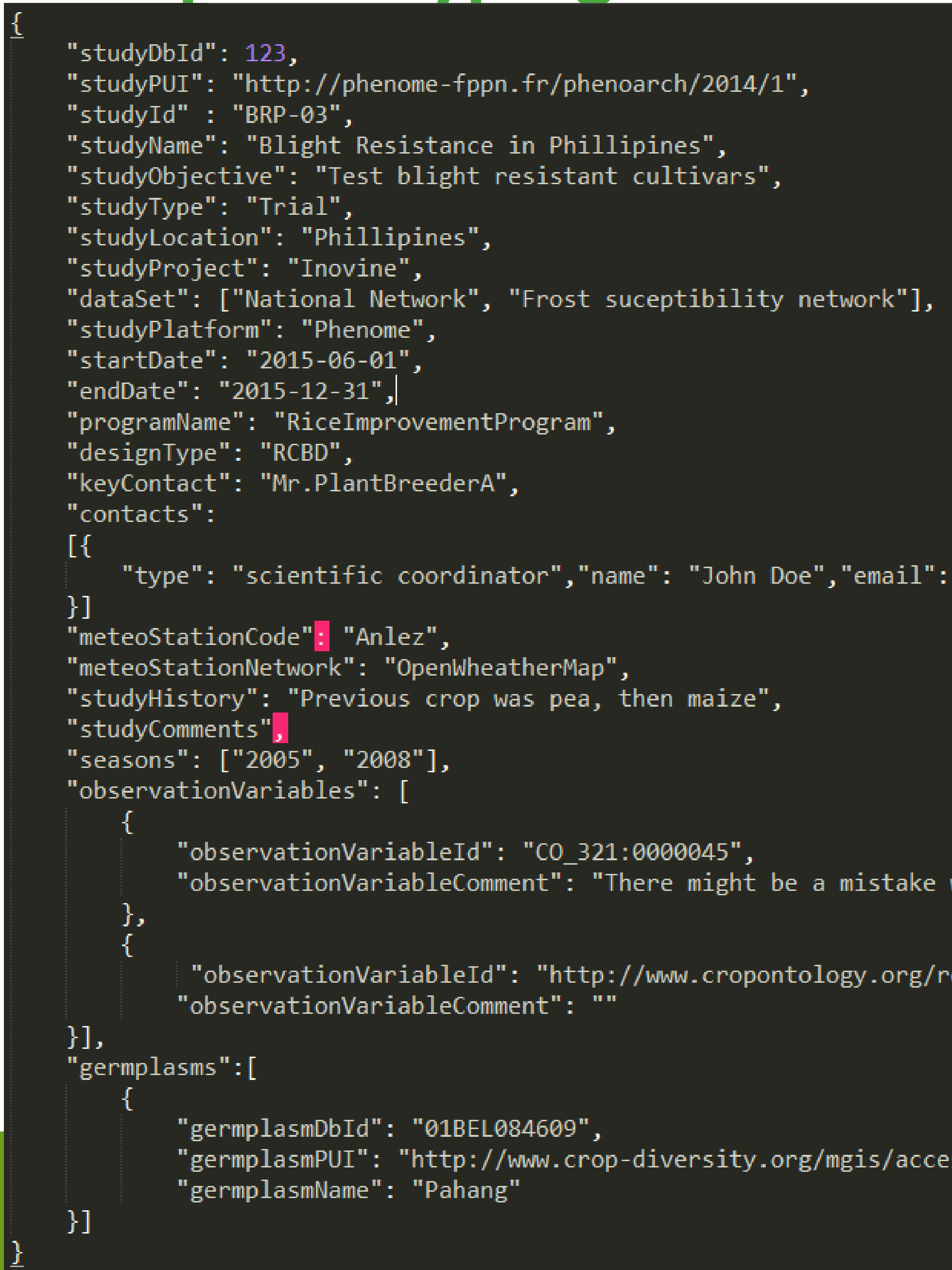
- Field duplication
  - **Interweaved document as attributes of root document**
- Correlated documents (ie foreign Key)



```
Auteur
{
    "id": 10,
    "nom": "dupont",
    "prenom": "david",
    "livres": [
        101,503,339,342
    ]

}
Livre
{
    "id": 342,
    "titre": "NoSQL schema",
    "genre": "informatique",

    "tags":["informatique","bigdata","nosql"]
    "auteurs": [
        10,234
    ]

}
```

Application to phenotyping

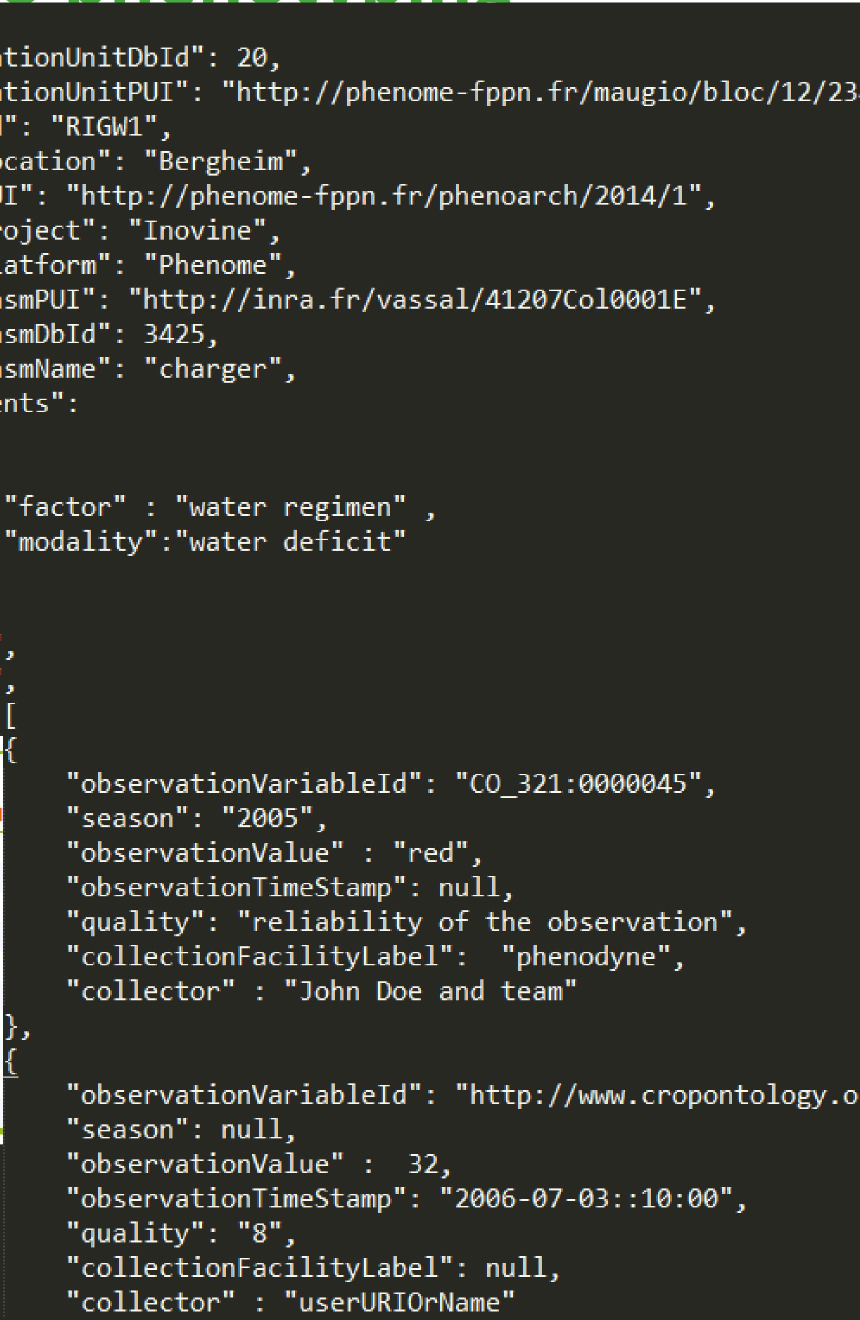
- Breeding API
  - http://docs.brapi.apiary.io
- Study aka Trial



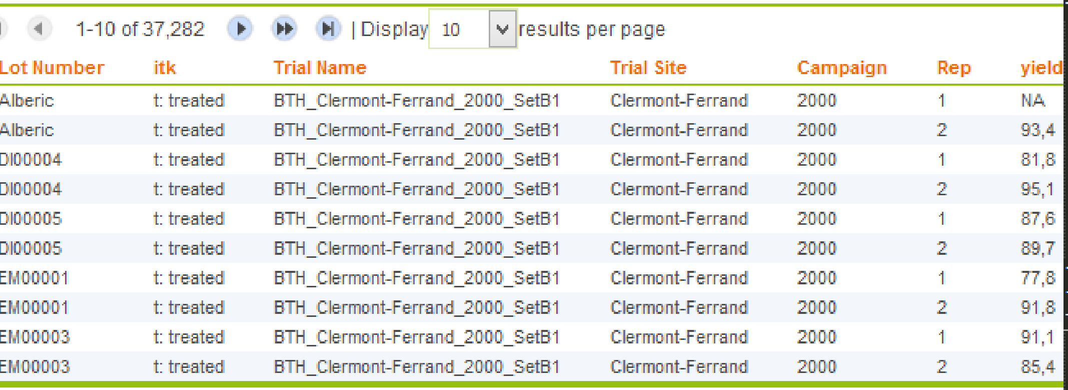
```
{
    "studyDbId": 123,
    "studyPUI": "http://phenome-fppn.fr/phenoarch/2014/1",
    "studyId" : "BRP-03",
    "studyName": "Blight Resistance in Phillipines",
    "studyObjective": "Test blight resistant cultivars",
    "studyType": "Trial",
    "studyLocation": "Phillipines",
    "studyProject": "Inovine",
    "dataSet": ["National Network", "Frost suceptibility network"],
    "studyPlatform": "Phenome",
    "startDate": "2015-06-01",
    "endDate": "2015-12-31",
    "programName": "RiceImprovementProgram",
    "designType": "RCBD",
    "keyContact": "Mr.PlantBreederA",
    "contacts":
    [{
        "type": "scientific coordinator","name": "John Doe","email":
    }]
    "meteoStationCode": "Anlez",
    "meteoStationNetwork": "OpenWheatherMap",
    "studyHistory": "Previous crop was pea, then maize",
    "studyComments",
    "seasons": ["2005", "2008"],
    "observationVariables": [
        {
            "observationVariableId": "CO_321:0000045",
            "observationVariableComment": "There might be a mistake
        },
        {
             "observationVariableId": "http://www.cropontology.org/r
            "observationVariableComment": ""
    }],
    "germplasms":[
        {
            "germplasmDbId": "01BEL084609",
            "germplasmPUI": "http://www.crop-diversity.org/mgis/acce
            "germplasmName": "Pahang"
    }]
}
```

Application to phenotyping

- Breeding API
- Phenotypes
- One document = one data matrix line
- Elasticsearch nested documents



```
{
        "observationUnitDbId": 20,
        "observationUnitPUI": "http://phenome-fppn.fr/maugio/bloc/12/23
        "studyId": "RIGW1",
        "studyLocation": "Bergheim",
        "studyPUI": "http://phenome-fppn.fr/phenoarch/2014/1",
        "studyProject": "Inovine",
        "studyPlatform": "Phenome",
        "germplasmPUI": "http://inra.fr/vassal/41207Col0001E",
        "germplasmDbId": 3425,
        "germplasmName": "charger",
        "treatments":
        [
            {
                "factor" : "water regimen" ,
                "modality":"water deficit"
            }
        ],
        "X" : "",
        "Y" : "",
        "data": [
                {
                        "observationVariableId": "CO_321:0000045",
                        "season": "2005",
                        "observationValue" : "red",
                        "observationTimeStamp": null,
                        "quality": "reliability of the observation",
                        "collectionFacilityLabel":  "phenodyne",
                        "collector" : "John Doe and team"
                },
                {
                        "observationVariableId": "http://www.cropontology.o
                        "season": null,
                        "observationValue" :  32,
                        "observationTimeStamp": "2006-07-03::10:00",
                        "quality": "8",
                        "collectionFacilityLabel": null,
                        "collector" : "userURIOrName"
```



1-10 of 37,282 | Display 10 results per page

Lot Number	itk	Trial Name	Trial Site	Campaign	Rep	yield
Alberic	t: treated	BTH_Clermont-Ferrand_2000_SetB1	Clermont-Ferrand	2000	1	NA
Alberic	t: treated	BTH_Clermont-Ferrand_2000_SetB1	Clermont-Ferrand	2000	2	93,4
DI00004	t: treated	BTH_Clermont-Ferrand_2000_SetB1	Clermont-Ferrand	2000	1	81,8
DI00004	t: treated	BTH_Clermont-Ferrand_2000_SetB1	Clermont-Ferrand	2000	2	95,1
DI00005	t: treated	BTH_Clermont-Ferrand_2000_SetB1	Clermont-Ferrand	2000	1	87,6
DI00005	t: treated	BTH_Clermont-Ferrand_2000_SetB1	Clermont-Ferrand	2000	2	89,7
EM00001	t: treated	BTH_Clermont-Ferrand_2000_SetB1	Clermont-Ferrand	2000	1	77,8
EM00001	t: treated	BTH_Clermont-Ferrand_2000_SetB1	Clermont-Ferrand	2000	2	91,8
EM00003	t: treated	BTH_Clermont-Ferrand_2000_SetB1	Clermont-Ferrand	2000	1	91,1
EM00003	t: treated	BTH_Clermont-Ferrand_2000_SetB1	Clermont-Ferrand	2000	2	85,4

GnpIS,Ephesis V3 performances

- 700 000 observation units
- 3 * 10^6 phenotypes / observations
- Elasticsearch implementation
- Response time
  - https://urgi.versailles.inra.fr/ephesis/ephesis/viewer.do#dataResults/trialSetIds=6,5,7
  - 770 Trials, 49 658 documents subset, 6 variables each
  - One page, Ajax call
    - 300 – 600 ms
  - Full export
    - 23 s
- Genotyping :

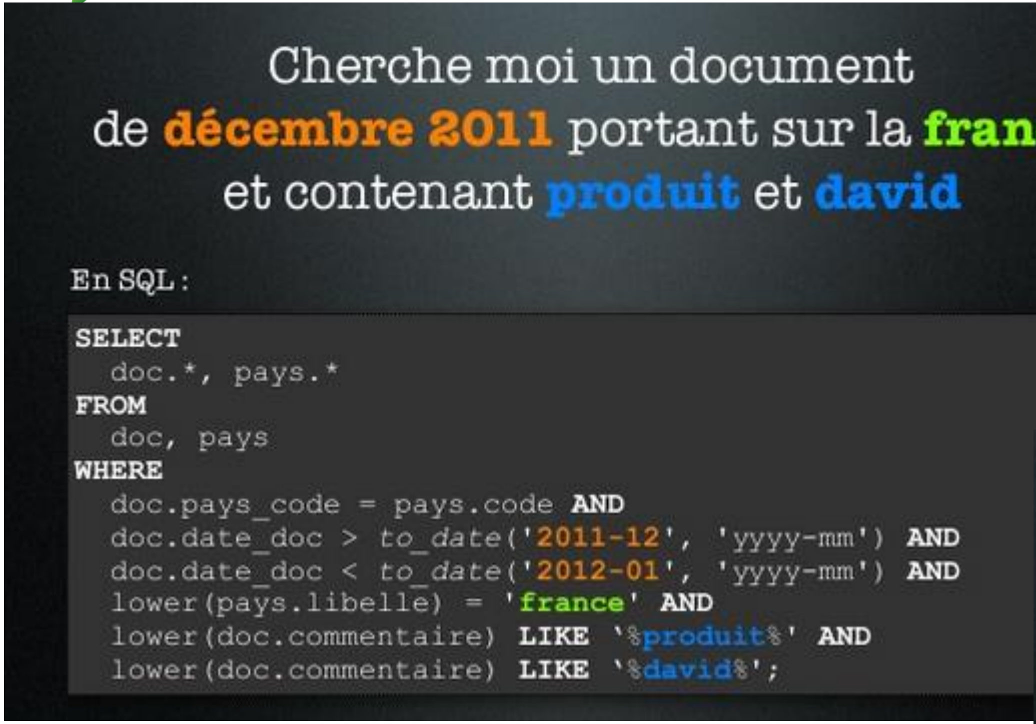
Elasticsearch

- David Pilato
  - http://david.pilato.fr/
  - Evangelist at elastic and creator of the Elastic French Speakers User Group.
- http://fr.slideshare.net/*
- Elasticsearch
  - https://www.elastic.co/
  - Créé en 2010
  - International
  - Applications et adoption croissante

Elasticsearch

- Lucene based
- Search engine
  - Not presented as a NoSQL database
    - NoSQL document search engine
    - BUT : it has backup systems
  - Not for long term storage
    - Aggregated documents, query oriented
  - Could be used as NoSQL DB ?
    - Storage
- HTTP, REST, JSON
- Distributed, Scalable, Cluster and Cloud ready

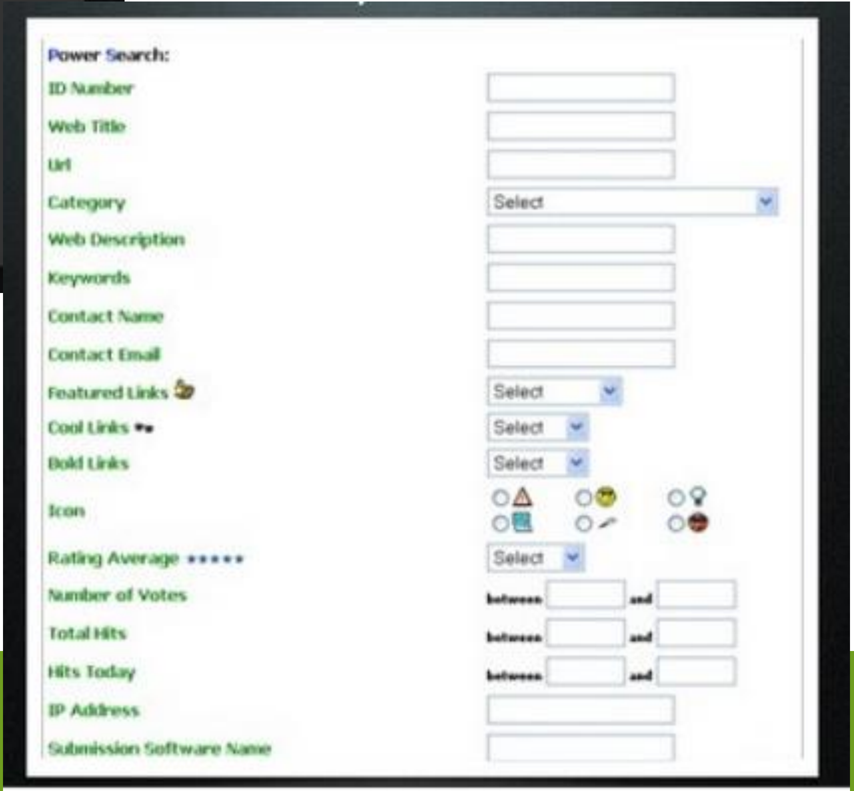
ES vs SQL



Cherche moi un document
de **décembre 2011** portant sur la **france**
et contenant **produit** et **david**

En SQL :

```
SELECT
  doc.*, pays.*
FROM
  doc, pays
WHERE
  doc.pays_code = pays.code AND
  doc.date_doc > to_date('2011-12', 'yyyy-mm') AND
  doc.date_doc < to_date('2012-01', 'yyyy-mm') AND
  lower(pays.libelle) = 'france' AND
  lower(doc.commentaire) LIKE '%produit%' AND
  lower(doc.commentaire) LIKE '%david%';
```

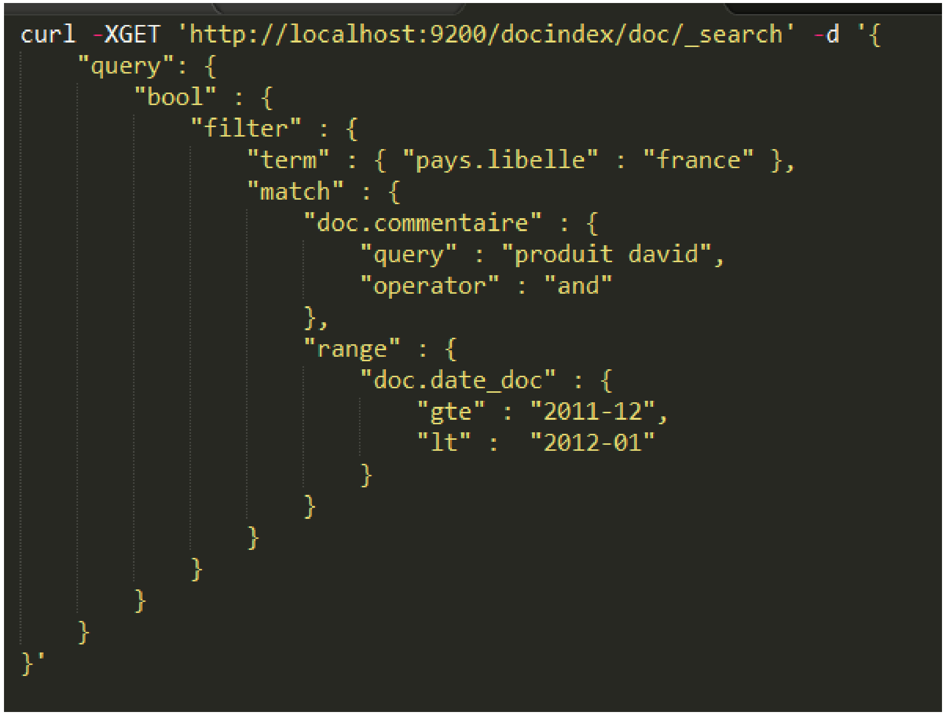


Power Search:

ID Number	
Web Title	
Url	
Category	Select
Web Description	
Keywords	
Contact Name	
Contact Email	
Featured Links	Select
Cool Links	Select
Bold Links	Select
Icon	○ ○ ○ ○ ○ ○
Rating Average ★★★★★	Select
Number of Votes	between and
Total Hits	between and
Hits Today	between and
IP Address	
Submission Software Name	

http://fr.slideshare.net/dadoonet/elasticsearch-esme-sudria

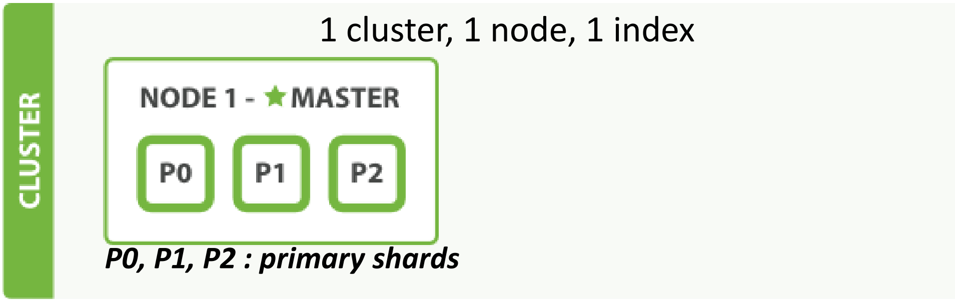
Elasticsearch



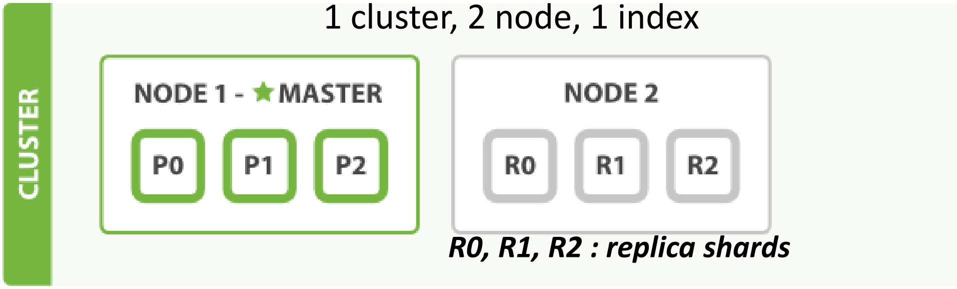
```
curl -XGET 'http://localhost:9200/docindex/doc/_search' -d '{
    "query": {
        "bool" : {
            "filter" : {
                "term" : { "pays.libelle" : "france" },
                "match" : {
                    "doc.commentaire" : {
                        "query" : "produit david",
                        "operator" : "and"
                    },
                    "range" : {
                        "doc.date_doc" : {
                            "gte" : "2011-12",
                            "lt" :  "2012-01"
                        }
                    }
                }
            }
        }
    }
}'
```

Elasticsearch administration concepts

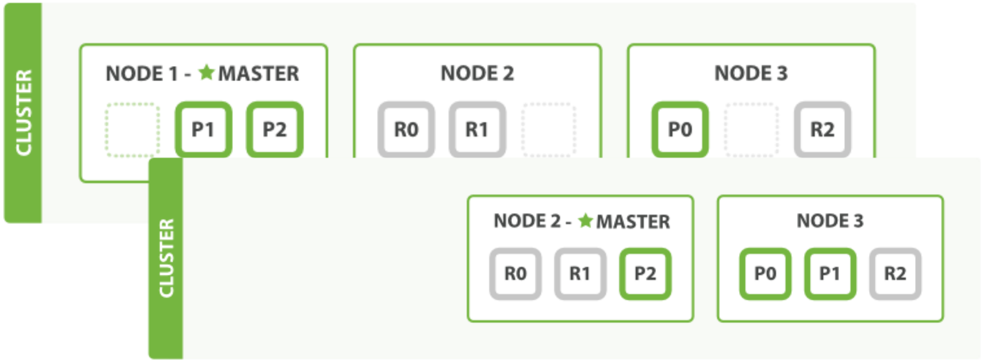
- URL HTTP : ES_instance/index/type/documents
- node :
  - Running instance of elasticsearch, belongs to a cluster.
- cluster :
  - one or more nodes with same cluster name
  - Single master node
    - chosen automatically
    - replaced if fails.
- shard
  - single Lucene instance
  - low-level “worker”
  - managed automatically by elasticsearch.
  - primary and replica shards.


1 cluster, 1 node, 1 index

P0, P1, P2 : primary shards


1 cluster, 2 node, 1 index

R0, R1, R2 : replica shards

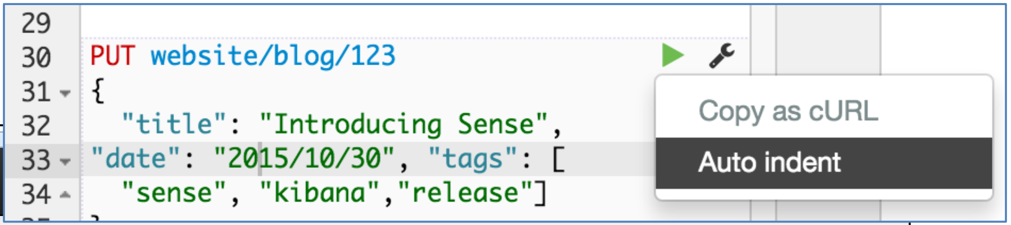


Elasticsearch developer concepts

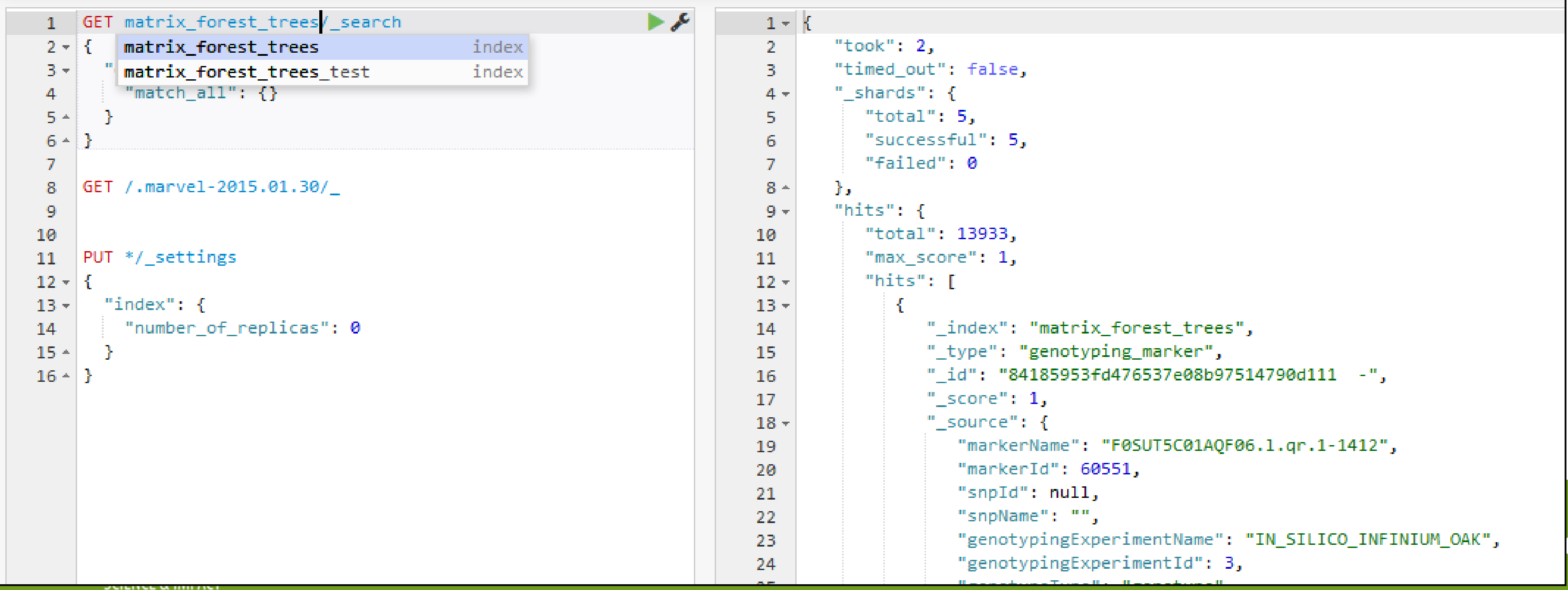
- type
  - like a *table* in a relational database.
- index
  - like a *database* in a relational database.
  - has a mappings which defines multiple types.
  - types
    - Study
      - http://localhost:9200/phenoindex/studytype/_search?
    - Phenotypes
      - http://localhost:9200/phenoindex/phenotype/_search?
- Alias
  - like a SDGB view on multiple indices (possibly filtered via a query)

Development

- Very good query DSL/API documentation
  - https://www.elastic.co/guide/en/elasticsearch/reference/current/search.html
- Sense
  - Free, integrated into marvel (https://www.elastic.co/products/marvel)
  - https://www.elastic.co/guide/en/sense/current/sense-ui.html
  - query execution
  - autocompletion



```
PUT website/blog/123
{
  "title": "Introducing Sense",
"date": "2015/10/30", "tags": [
  "sense", "kibana","release"]
```



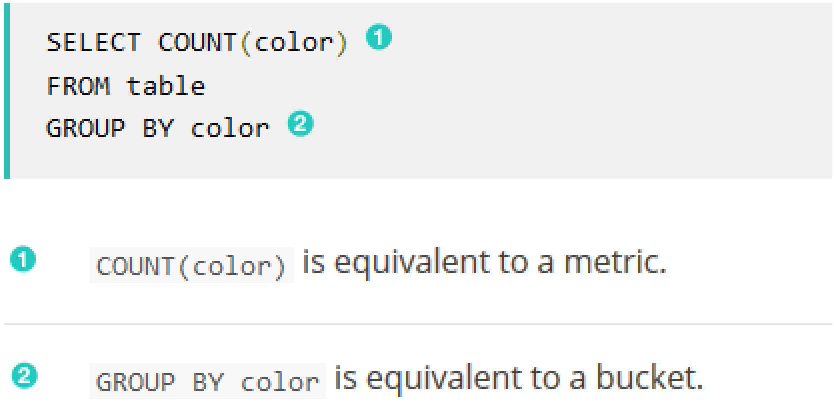
Server 10.0.0.128:9200

```
GET matrix_forest_trees/_search
{
  "
    "match_all": {}
  }
}

GET /.marvel-2015.01.30/_


PUT */_settings
{
  "index": {
    "number_of_replicas": 0
  }
}
```

```
{
   "took": 2,
   "timed_out": false,
   "_shards": {
      "total": 5,
      "successful": 5,
      "failed": 0
   },
   "hits": {
      "total": 13933,
      "max_score": 1,
      "hits": [
         {
            "_index": "matrix_forest_trees",
            "_type": "genotyping_marker",
            "_id": "84185953fd476537e08b97514790d111  -",
            "_score": 1,
            "_source": {
               "markerName": "F0SUT5C01AQF06.1.qr.1-1412",
               "markerId": 60551,
               "snpId": null,
               "snpName": "",
               "genotypingExperimentName": "IN_SILICO_INFINIUM_OAK",
               "genotypingExperimentId": 3,
```

Agregations, ie Facets

- Analytic queries
- Two main concepts:
  - ***Buckets*** **Collections of documents that meet a criterion**
  - ***Metrics*** **Statistics calculated on the documents in a bucket**

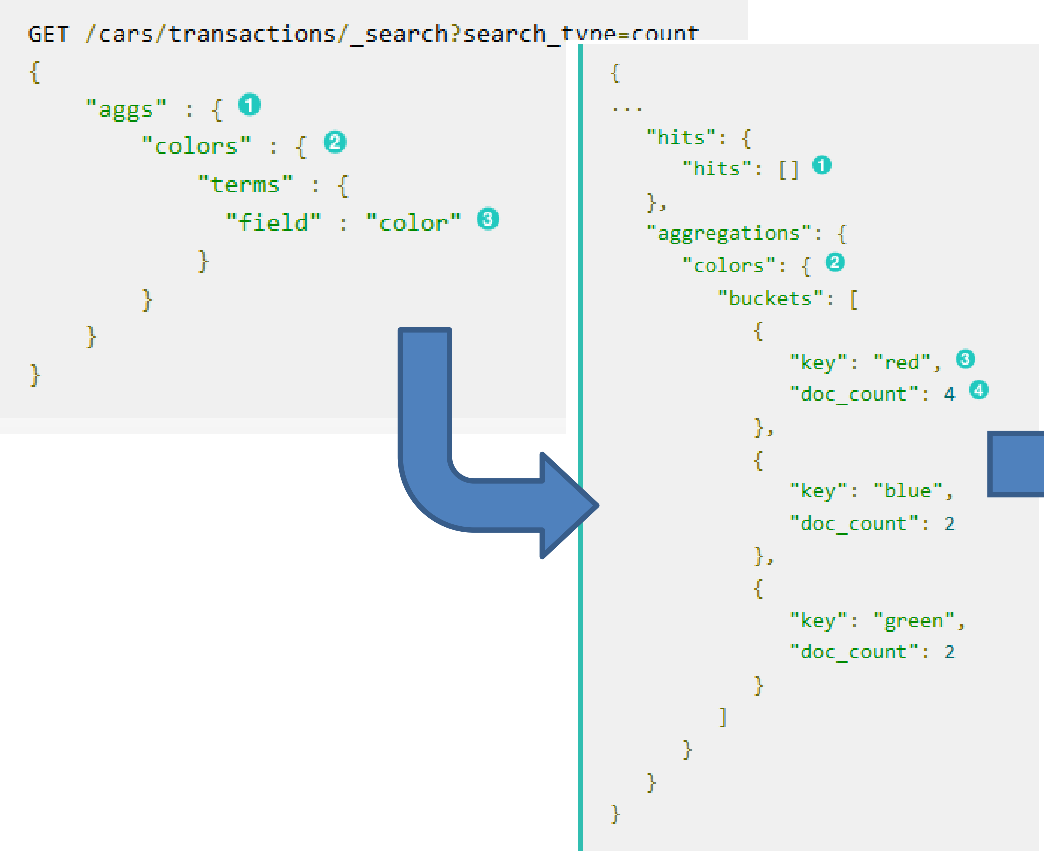


```
SELECT COUNT(color) ❶
FROM table
GROUP BY color ❷
```

❶ `COUNT(color)` is equivalent to a metric.

❷ `GROUP BY color` is equivalent to a bucket.

Aggregation example

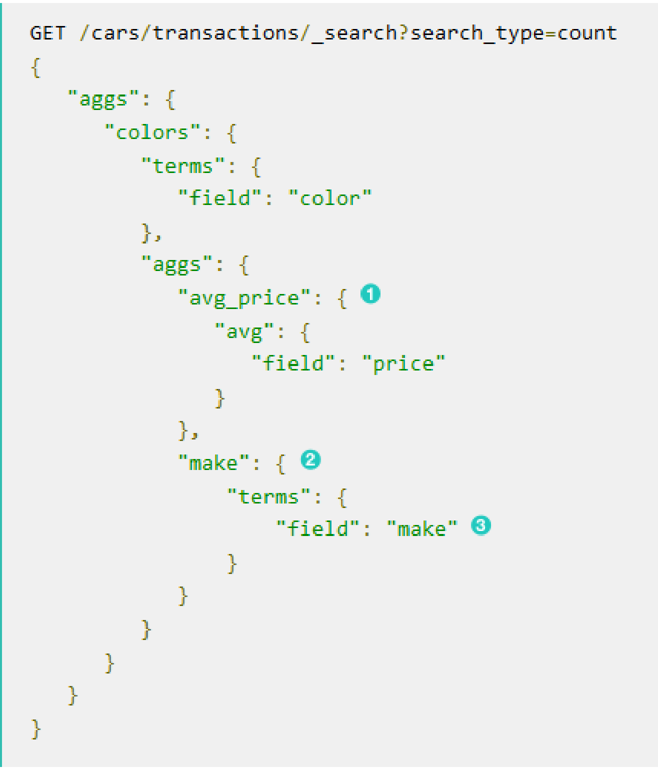


```
GET /cars/transactions/_search?search_type=count
{
    "aggs" : { ❶
        "colors" : { ❷
            "terms" : {
              "field" : "color" ❸
            }
        }
    }
}
```

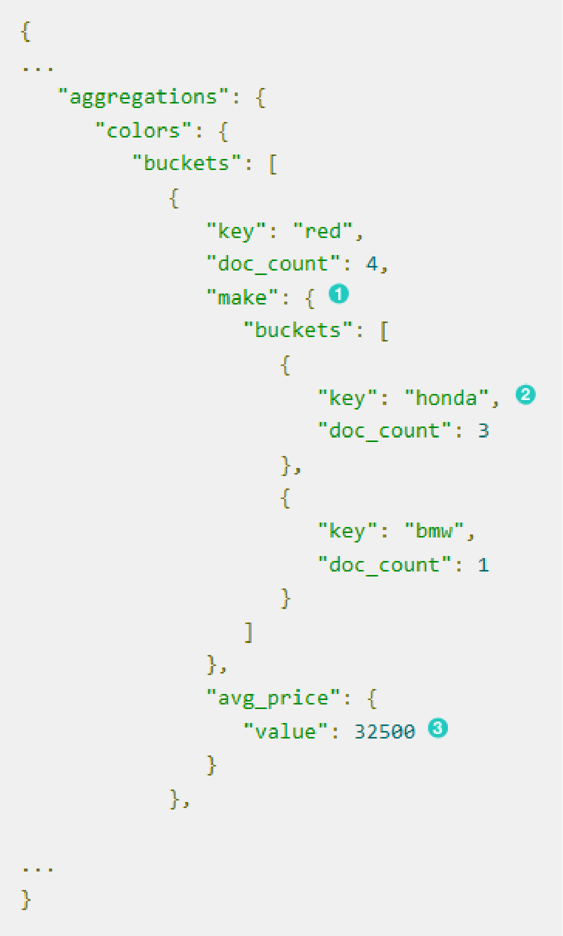
```
{
...
   "hits": {
      "hits": [] ❶
   },
   "aggregations": {
      "colors": { ❷
         "buckets": [
            {
               "key": "red", ❸
               "doc_count": 4 ❹
            },
            {
               "key": "blue",
               "doc_count": 2
            },
            {
               "key": "green",
               "doc_count": 2
            }
         ]
      }
   }
}
```

1/ Display on result page as clickable facet
2/ on click on « red » filter color : red

Aggregation : Avg and imbrication



```
GET /cars/transactions/_search?search_type=count
{
   "aggs": {
      "colors": {
         "terms": {
            "field": "color"
         },
         "aggs": {
            "avg_price": { ❶
               "avg": {
                  "field": "price"
               }
            },
            "make": { ❷
                "terms": {
                    "field": "make" ❸
                }
            }
         }
      }
   }
}
```



```
{
...
   "aggregations": {
      "colors": {
         "buckets": [
            {
               "key": "red",
               "doc_count": 4,
               "make": { ❶
                  "buckets": [
                     {
                        "key": "honda", ❷
                        "doc_count": 3
                     },
                     {
                        "key": "bmw",
                        "doc_count": 1
                     }
                  ]
               },
               "avg_price": {
                  "value": 32500 ❸
               }
            },

...
}
```

Administration

- Monitoring : Marvel
  - Gratuit dans ES 2.0
- Installation
  - Repository DEB and RPM
  - Unzip and run (development instance, tomcat like)
- Configuration YAML

Elasticsearch architecture

- Scalability
  - Automatic index data distribution accross shards
  - Adding nodes to increase number of shards
- Shard Replication
  - Option
  - Index by index configuration
  - Data availability
  - Query performances

Elasticsearch@ URGI

- Version 1.7.3
- Data :
  - 65 Gb
  - 600 millions documents, including nested
- Cluster Prod (same for dev)
  - 2 nodes: VM, 16 Go RAM, 2To, 8 CPU
- Each nodes
  - Unlimited number of indices
- Backup
  - Weekly, retention 8 weeks
  - On the fly, without service interruption, incremental
  - Via a REST query (API HTTP)
- Good data security
  - Protection against index corruption
  - Near NoSQL DB state

Insertion

- Generate JSON
  - Talend
  - Java
- Insert
  - Logstash
  - CURL

Thank you