HATCH

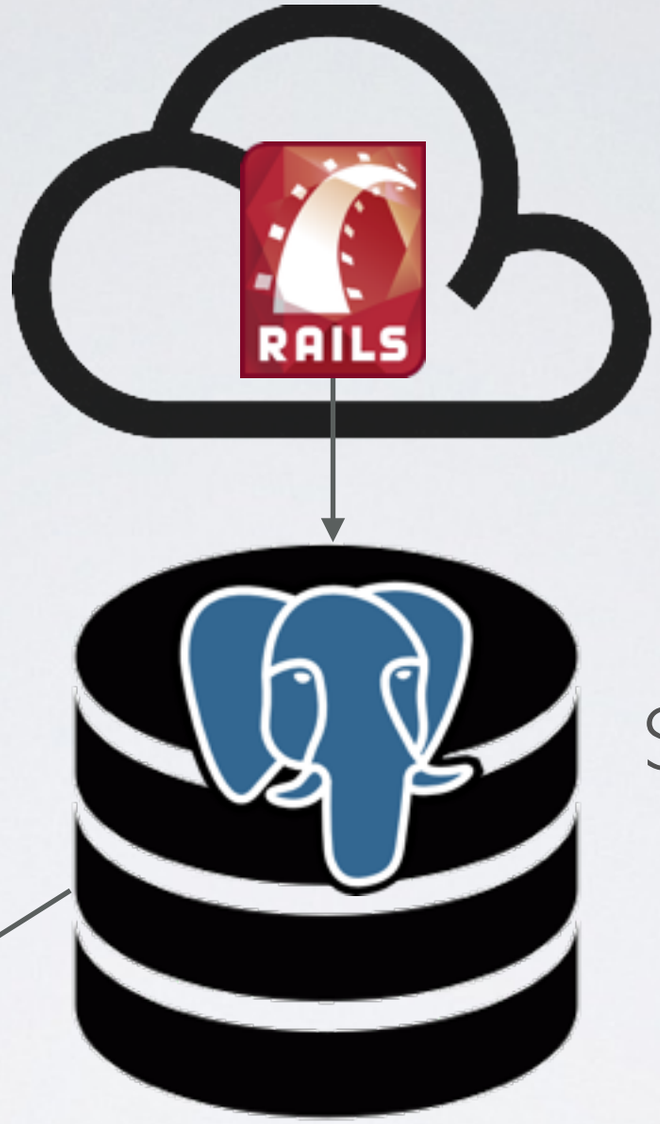
Overview

OVERVIEW

- Hatch is a cloud based solution for storing, sharing, and searching data.
- Locating data close to computational and storage resources leads to better utilization, and less requirements on researchers.
- Tools enable and encourage reduction of data redundancy.
- Sharing and searching eliminates cumbersome data collections, and improves research.

DATA STORAGE

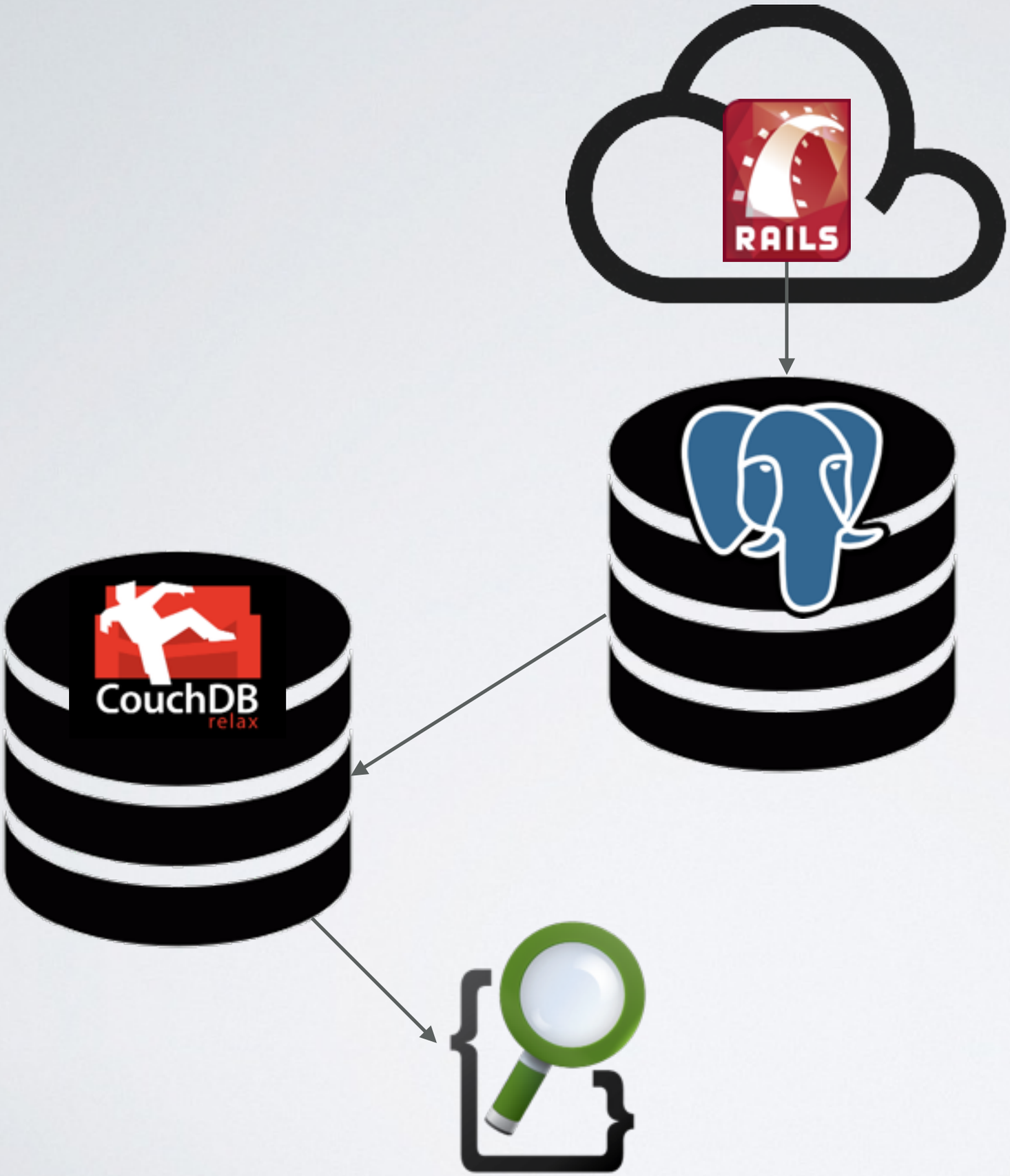
- Local data is uploaded to Hatch's cloud storage.
- Hatch filters different kinds of data into one storage type.
- Different kinds of data are now equally searchable, mergable, and sharable.
- Data still must meet their original structure and scheme requirements.



Schemeless database
for storing documents

Schemed database for
forcing Hatch
relationships

Search engine for fast
document indexing


RAILS
CouchDB
relax

HATCH HYBRID DATABASE

Schemed

- Hatch data is schemed when its needed.
- Enforcing relationships is central to Hatch's design.
- Data relationships that are know up front are kept in the schemed layer.

HATCH SCHEME

Postgresql

Active Record

Charts
- name
- id
- created_at
- updated_at

Documents
- name
- id
- user_id
- collection_id
- created_at
- updated_at

IFilters
- name
- id
- regex
- created_at
- updated_at

Collections
- name
- id
- user_id
- collection_id
- created_at
- updated_at

Feeds
- name
- id
- feed_url
- interval_val
- interval_unit
- document_id
- created_at
- updated_at

Users
- name
- id
- email
- encrypted_password
- reset_password_token
- reset_password_se
- rm_at
- remember_created_at
- sign_in_count
- current_sign_in_at
- last_sign_in_at
- current_sign_in_ip
- last_sign_in_ip
- created_at
- updated_at

CouchDB

- data
- is_search_doc

- last_id

HATCH HYBRID DATABASE

Unschemed

- Data relationships that are not know up front are kept in the unschemed layer.
- Hatch Project owners define data schemes on the Hatch API.
- Users then eventually converge their data to different project schemes.
- Project data is highly available to other members, instead of being only available once perfected.
- Projects can enforce not only data schemes, but data quality.

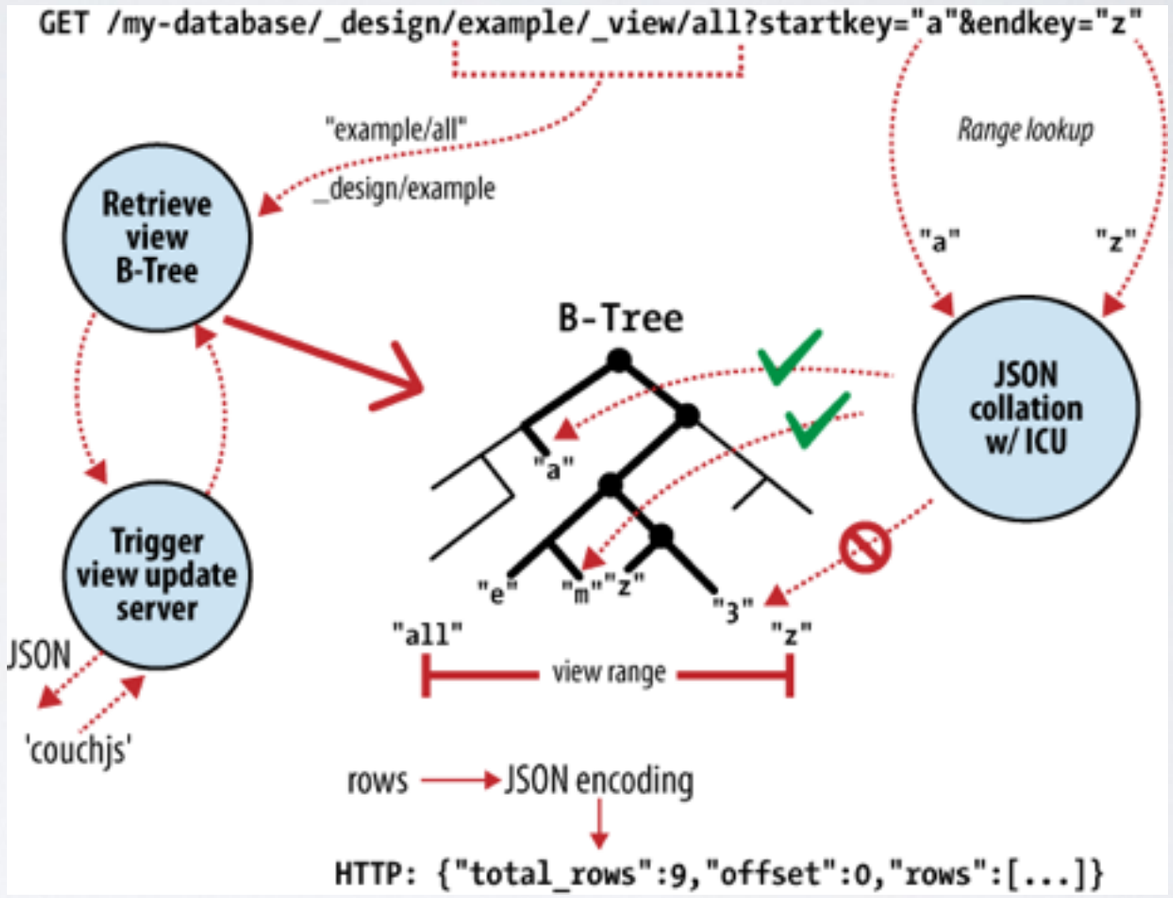
HATCH SCHEMELESS

CouchDB

Field	Value
_id	"Document-93"
_rev	"9249-b28048a4e69c5cbc17badd089f97757a"
data	[{ "date": "Sat Apr 28 07:11:34 PDT 2012", "1": 1357, "id": 9241 }, { "date": "Sat Apr 28 07:11:34 PDT 2012", "1": 1357, "id": 9242 }, { "date": "Sat Apr 28 07:11:34 PDT 2012", "1": 1357, "id": 9243 }, { "date": "Sat Apr 28 07:11:34 PDT 2012", "1": 1357, "id": 9244 }, { "date": "Sat Apr 28 07:11:34 PDT 2012", "1": 1357, "id": 9245 }, { "date": "Sat Apr 28 07:11:34 PDT 2012", "1": 1357, "id": 9246 }, { "date": "Sat Apr 28 07:11:34 PDT 2012", "1": 1357, "id": 9247 }]
last_id	9247

HATCH SCHEMELESS

CouchDB



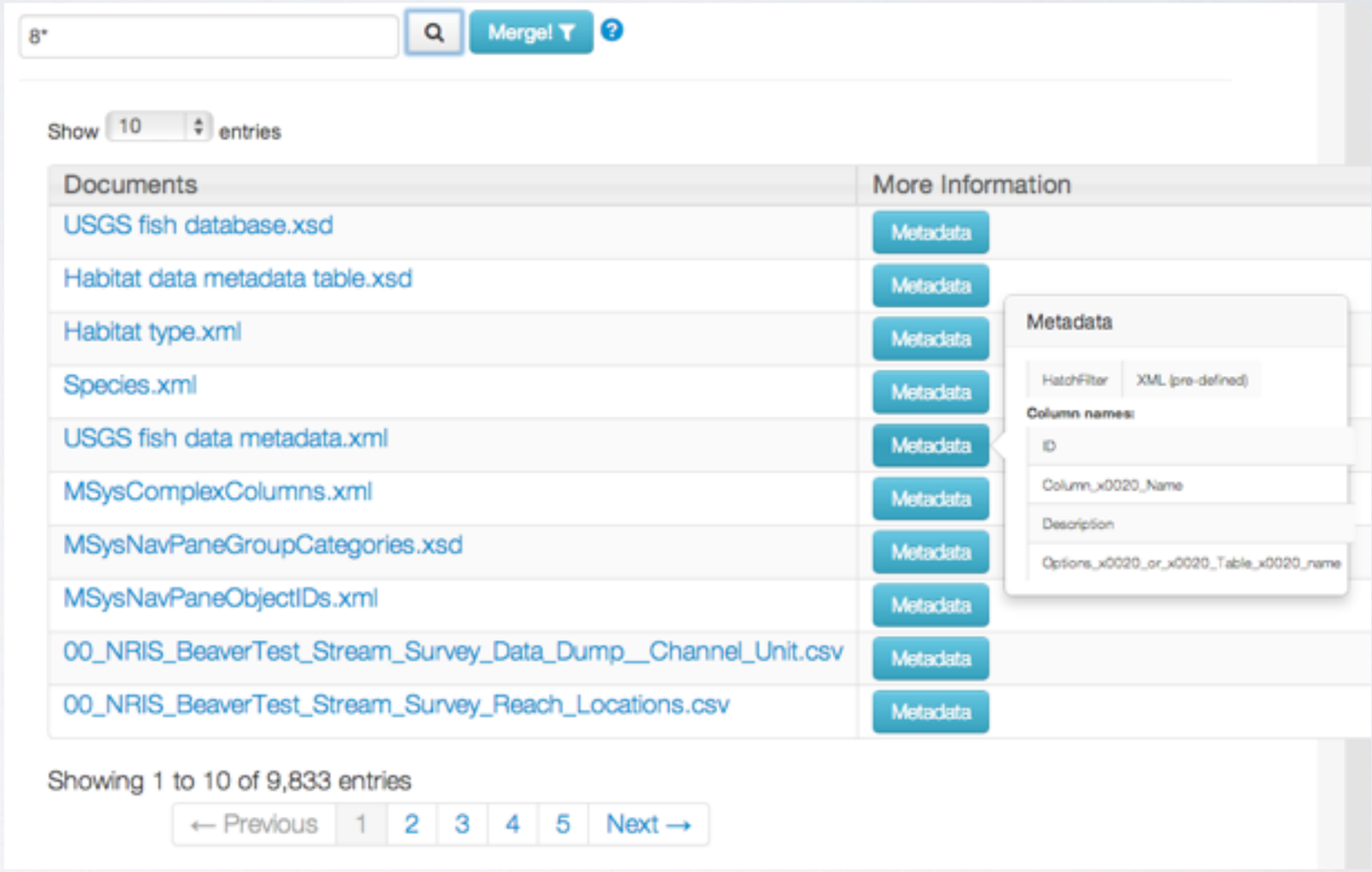
HATCH SEARCH

Elasticsearch

- The Hatch search engine automatically indexes incremental changes in the database.
- The Lucene query syntax is familiar to existing tools (i.e. Google).
- Search also finds relationships between data, some of which the user may not know about.
- Search can merge different document types together if the core data is related.

HATCH SEARCH

Elasticsearch



HATCH SEARCH

Elasticsearch - Merging

Search results:

Column names in common for merged documents: Subwatershed, Stream, Site, WSU, RKM, Date, Year, Length, Width

Save as...

Year:2004

Merge!

Show 10 entries

Subwatershed	Stream	Site	WSU	RKM	Date	Year	Length	Width
Methow	Beaver	UBR1	6	843.057.005	8/11/2004 0:00:00	2004	14	3
Methow	Beaver	UBR1	6	843.057.005	8/11/2004 0:00:00	2004	9	4
Methow	Beaver	UBR1	6	843.057.005	8/11/2004 0:00:00	2004	28	4
Methow	Beaver	UBR1	6	843.057.005	8/11/2004 0:00:00	2004	46	1
Methow	Beaver	UBR1	6	843.057.005	8/11/2004 0:00:00	2004	14	4
Methow	Beaver	UBR1	6	843.057.005	8/11/2004 0:00:00	2004	29	4
Methow	Beaver	UBR1	6	843.057.005	8/11/2004 0:00:00	2004	16	4