DIP: A Dataset and Process Management System for Big Data

Supervisor: prof. Yannis Velegrakis

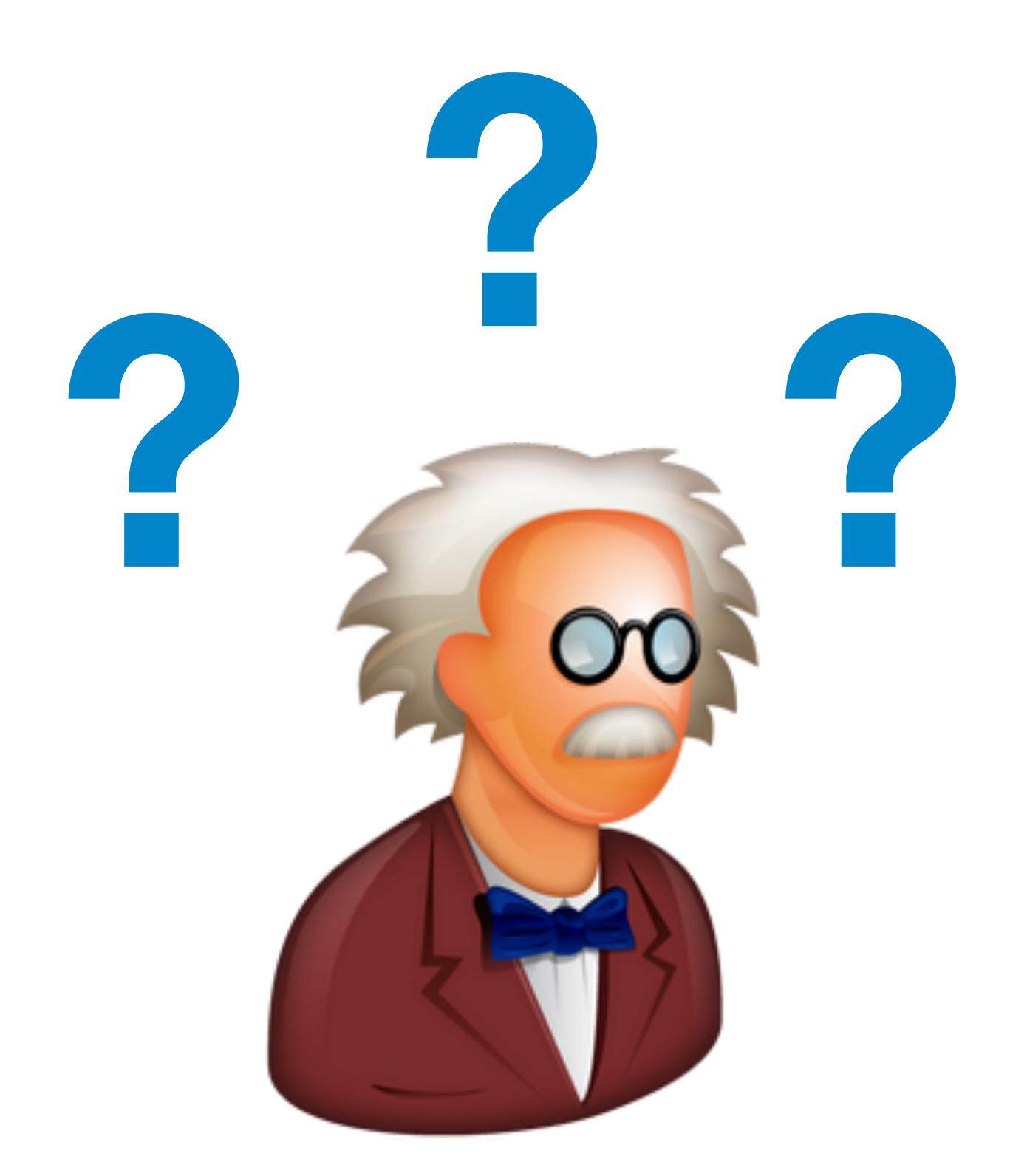
Student: Martin Brugnara



Motivating example



Private Datacenter



herb@corbett.com 1973-06-07 1990-09-26 1994-01-03 1990-01-17 lanette@gervasi.com shemika@huneke.com 1985-01-29 Columb is Agudelo columbus@agudelo.com 1991-08-11 jordon@seibold.com Hope Mcfarren hope@mcfarren.com 1961-01-08 Aide Shield aide@shield.com 1951-12-28 Ramoni a Castello ramonita@castello.com 1964-10-25 Conrad Koziol conrad@koziol.com 1968-04-21 Norma Clymer norma@clymer.com 1969-12-04 4 Rina Swope rina@swope.com Vertie Jaqua vertie@jaqua.com Joyce Creighton joyce@creighton.com 1970-11-05 Chanell Mcfarland chanelle@mcfarland.com 1956-05-04 Paulette Kiddy paulette@kiddy.com 1957-10-22 Christin : Thrasher christine@thrasher.com 1984-08-15 Dacus shonna@dacus.com 1977-10-10 Alameda carylon@alameda.com Loveday marisha@loveday.com Brescia dede@brescia.com d sec. this new oppose, so sec. havens opposed theel, error's d return bitheddaethi, use, tall mee, se, teutes) COMMONDATION Position has been used assessment of the beautiful control of the co for one create (SSE mone) edicardresterts: "sels.Tic. and tet. talt name atrians stead, errors of the fee.Sarusoft

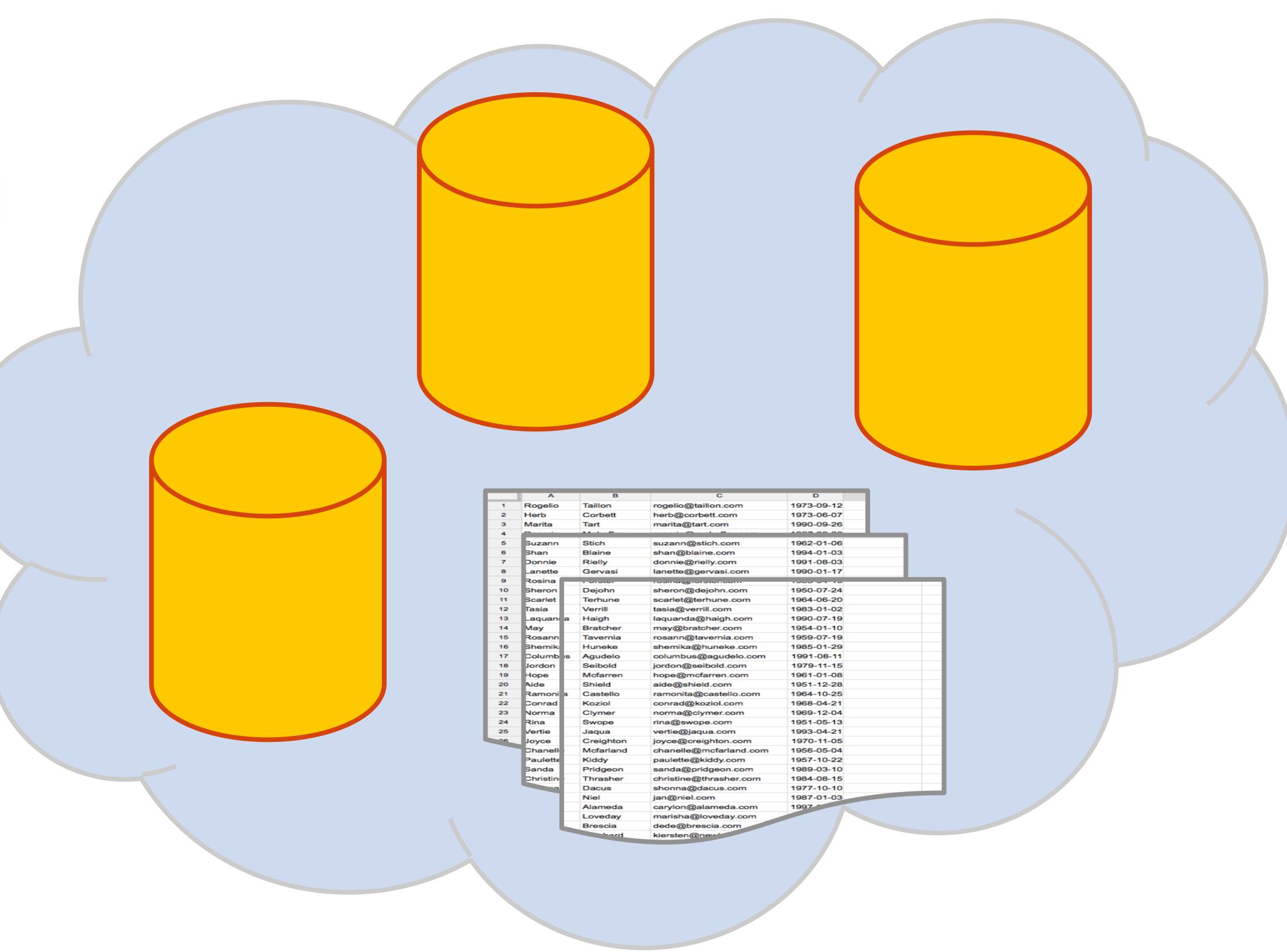
Amazon Cloud

Where is that dataset

Where does it come from

PROVENANCE

Gov. open data portal

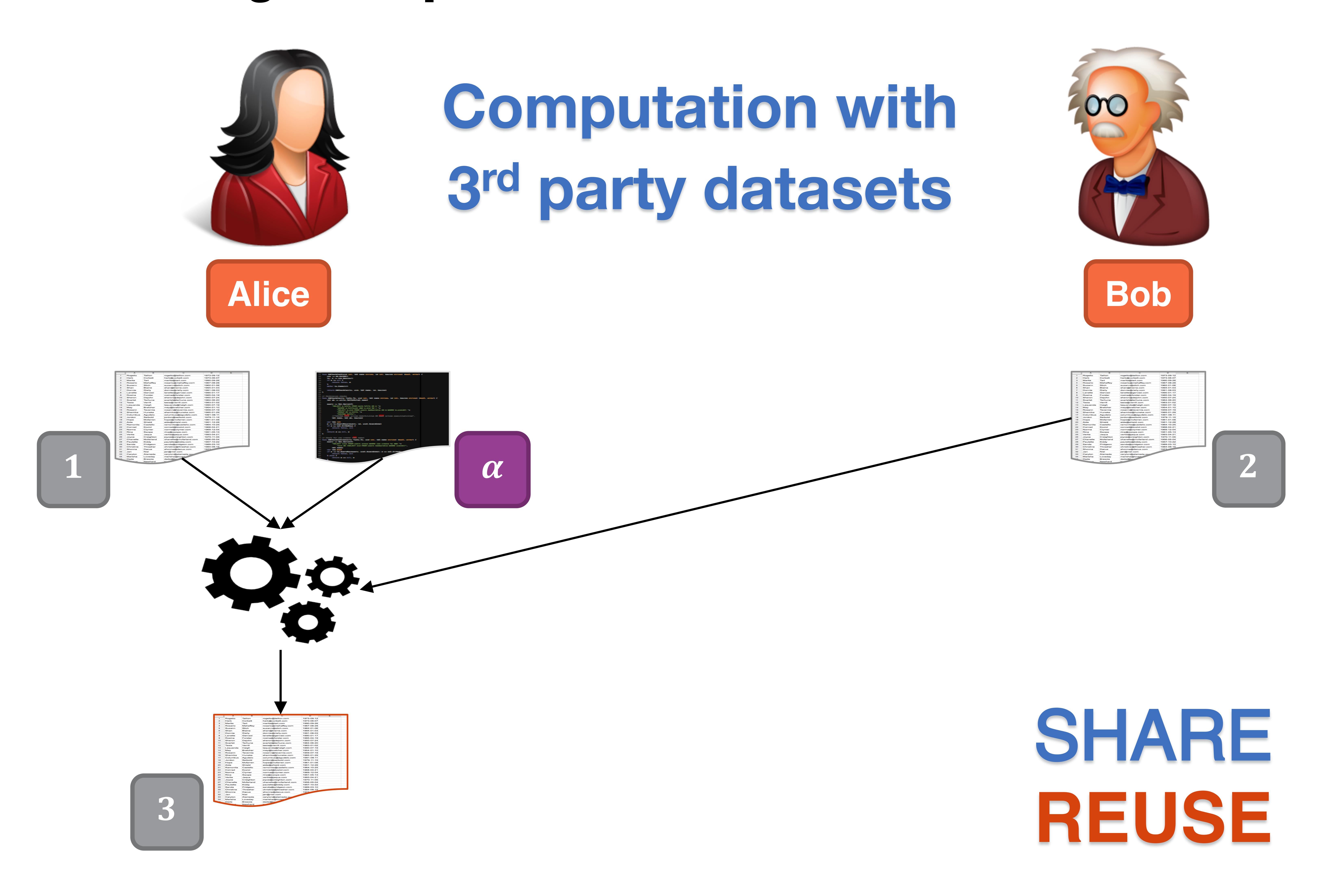




Researcher Workstation



Motivating example





Challenges

Storage

- Data
- Metadata

Accessing

- Technical
- e Legal

Processing

- Scheduling
- Reproducibility
- Validation



Existing systems

Data Stores

- ·CKAN
- Amazon Public Dataset repository

Workflow Managers

- Azkaban
- Ocie

Mixed Solutions

- Vistrails
- Hue

NO execution













Data Management

- Data portal
- Metadata managing
- Access Control

Workflow Design & Processing

- Operation scheduling
- Process Execution
- Support for different environments

Sharing & Provenance

- Workflow reutilization
- Data Provenance

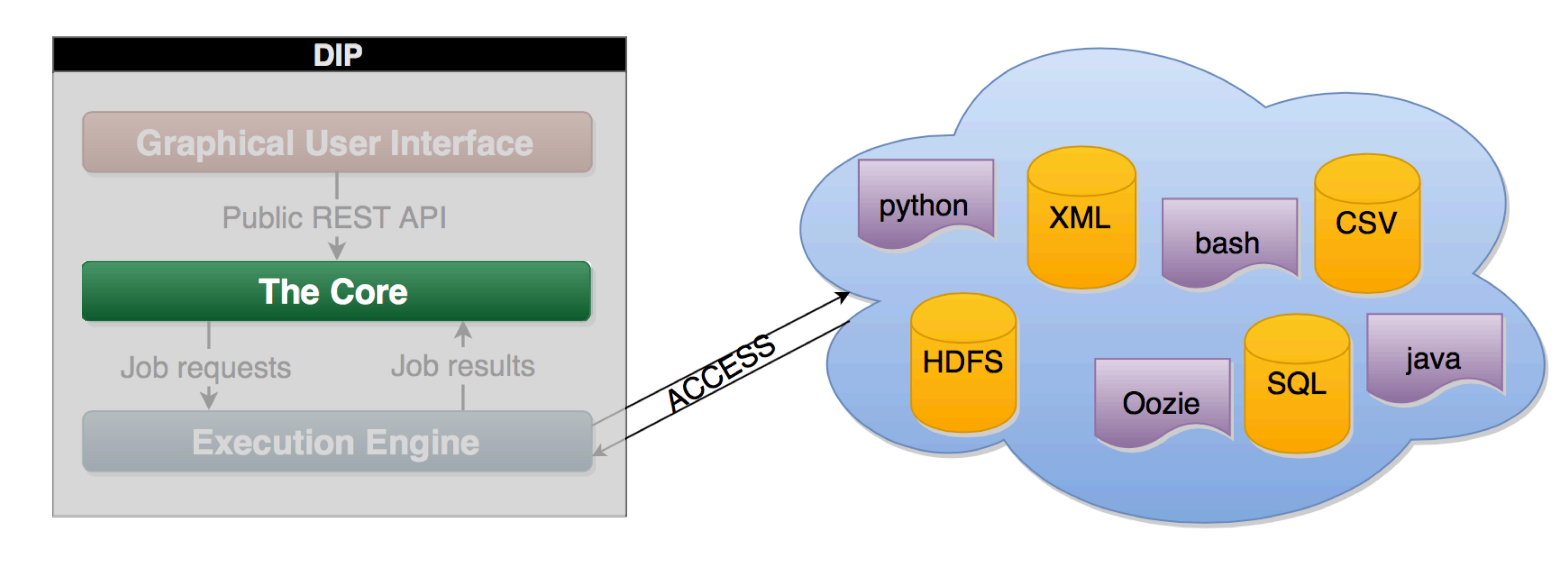


Our Solution: DIP

Built bottom up from scratch.

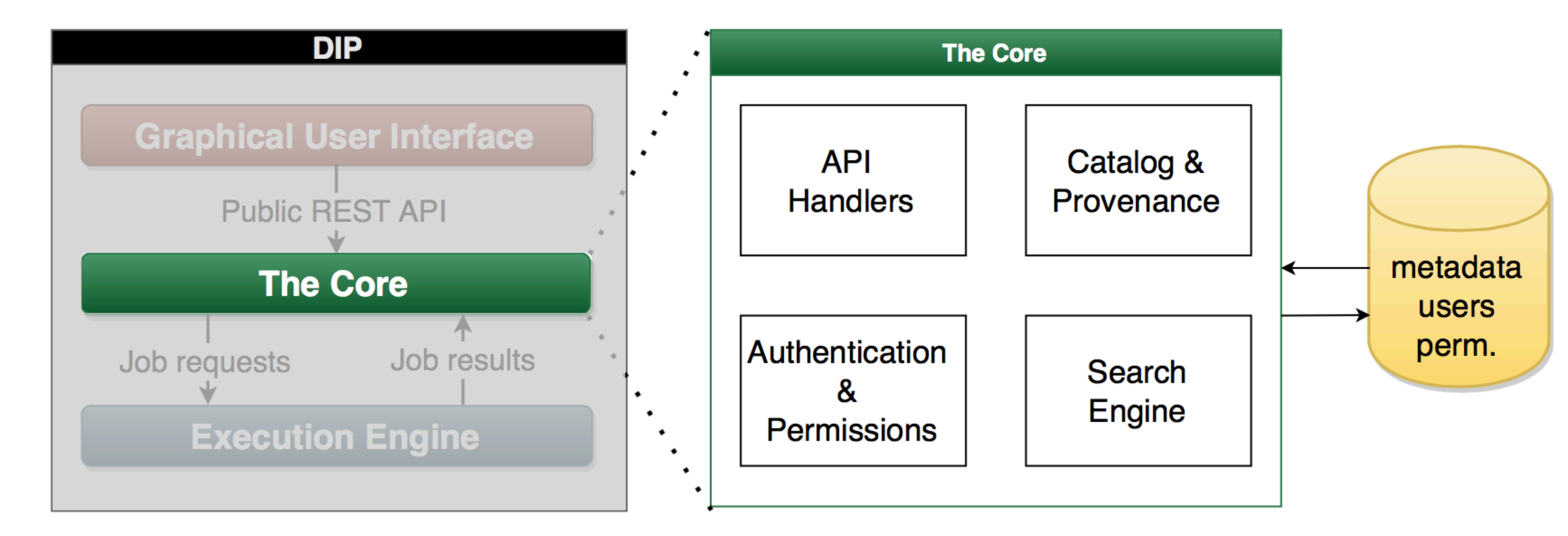


Architecture



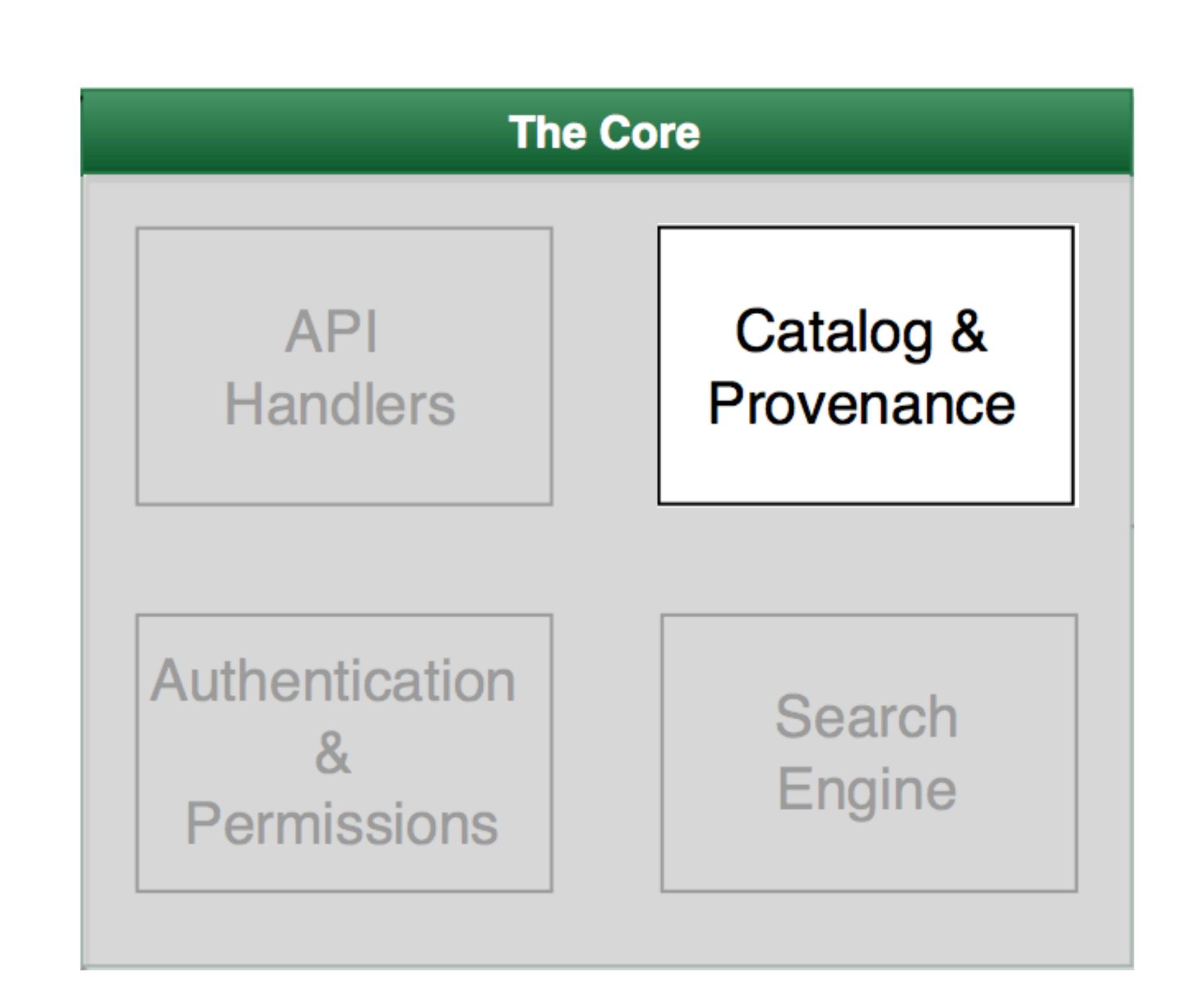


The Core





Catalog & Provenance



Orchestrator for the concept of:

Metadata

Job Serialization

Provenance Graph



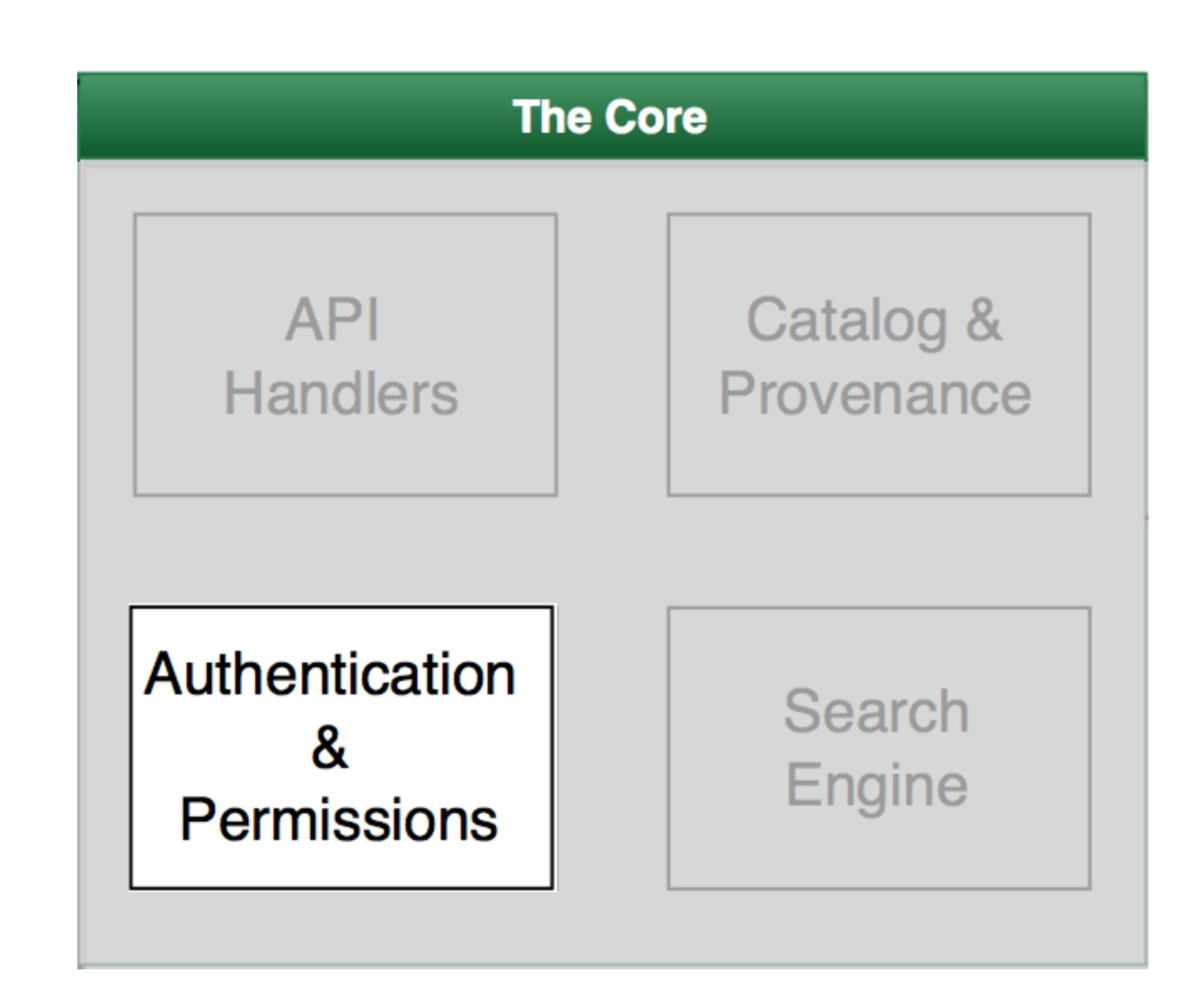
Authentication & Permissions

Request Authentication

- HTTP Basic Authentication
- Token
- Cookie

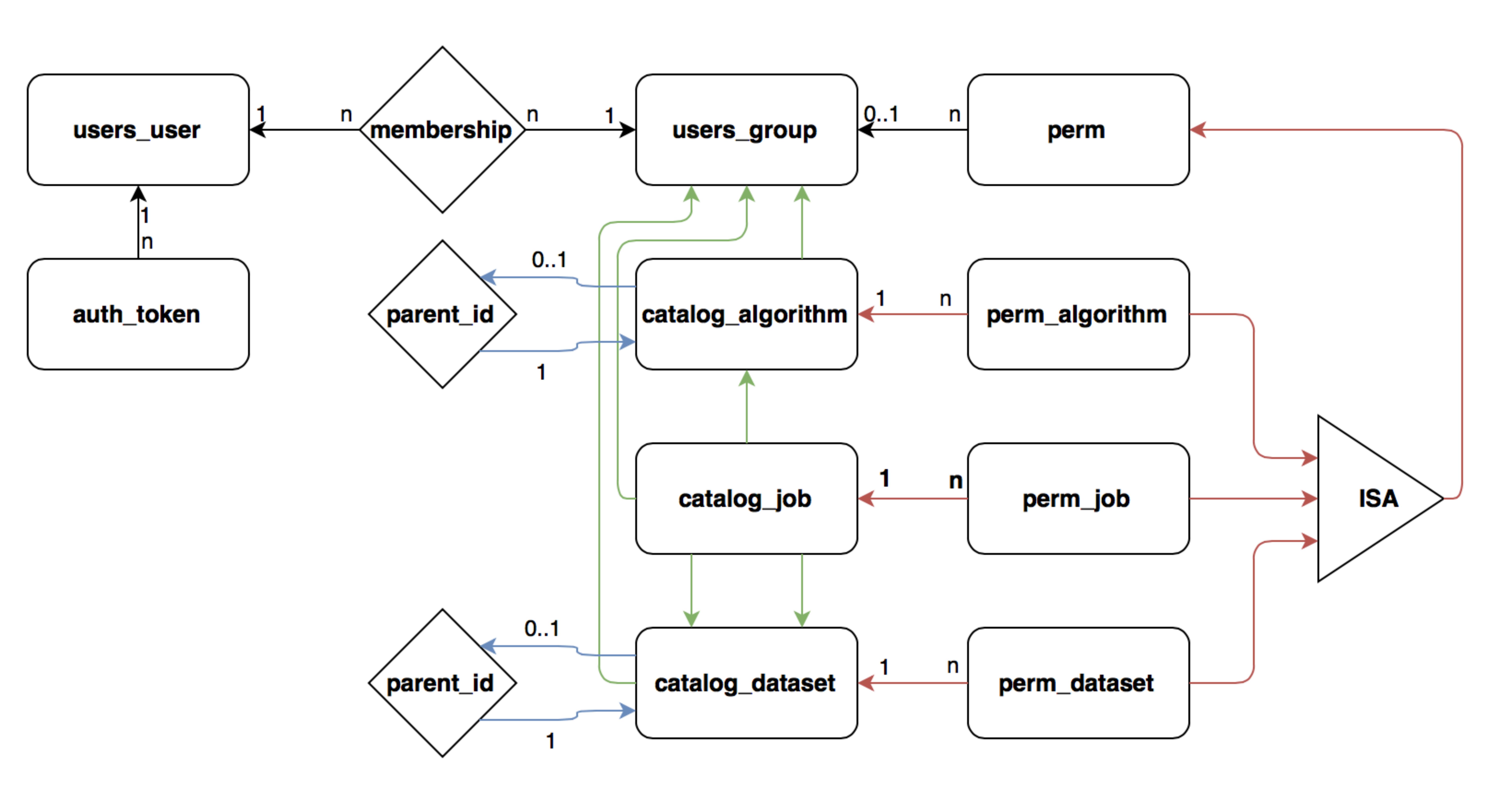
Users & Groups

Superuser



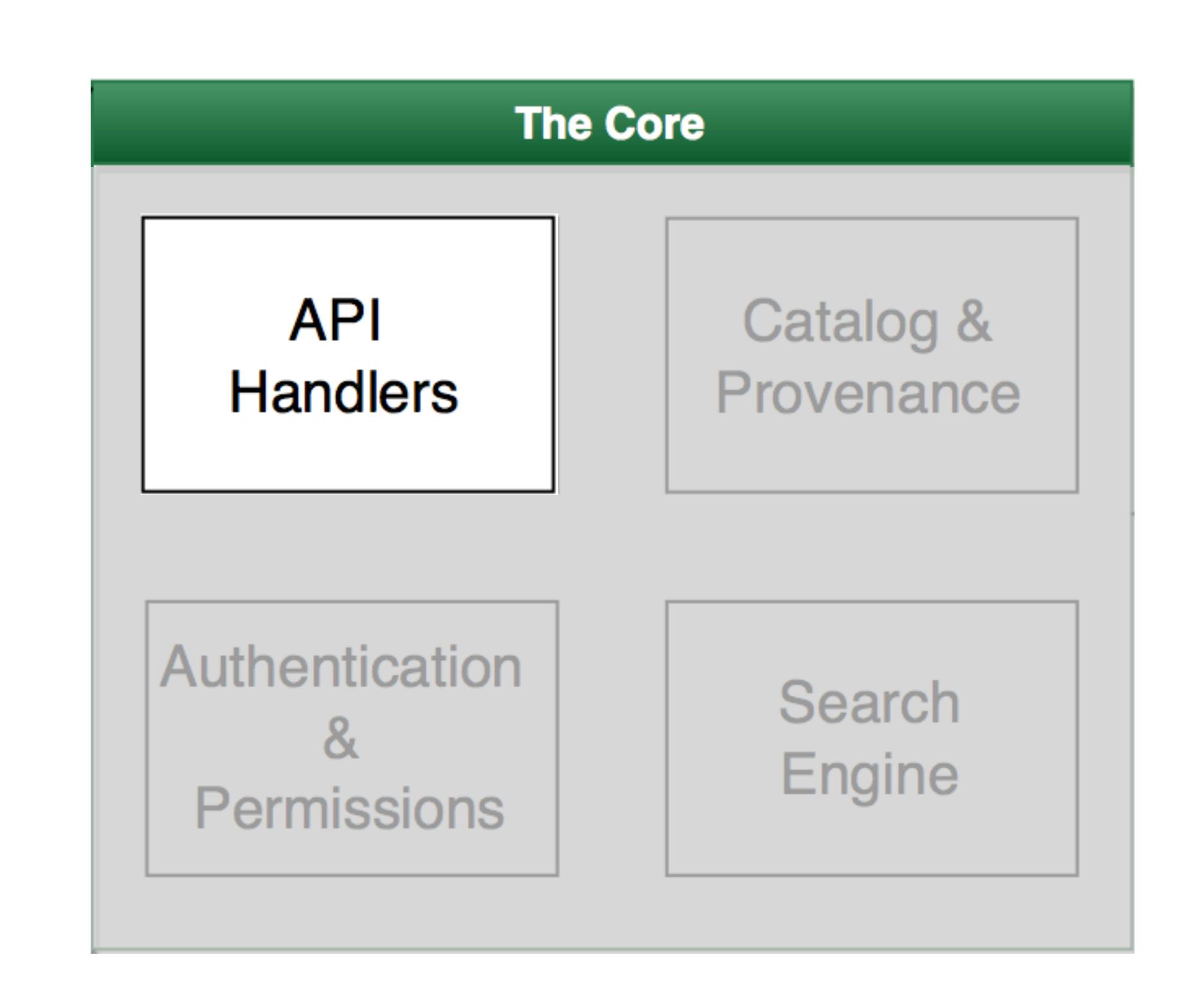


ER diagram





API Handlers



RESTful

HTTP Status Code

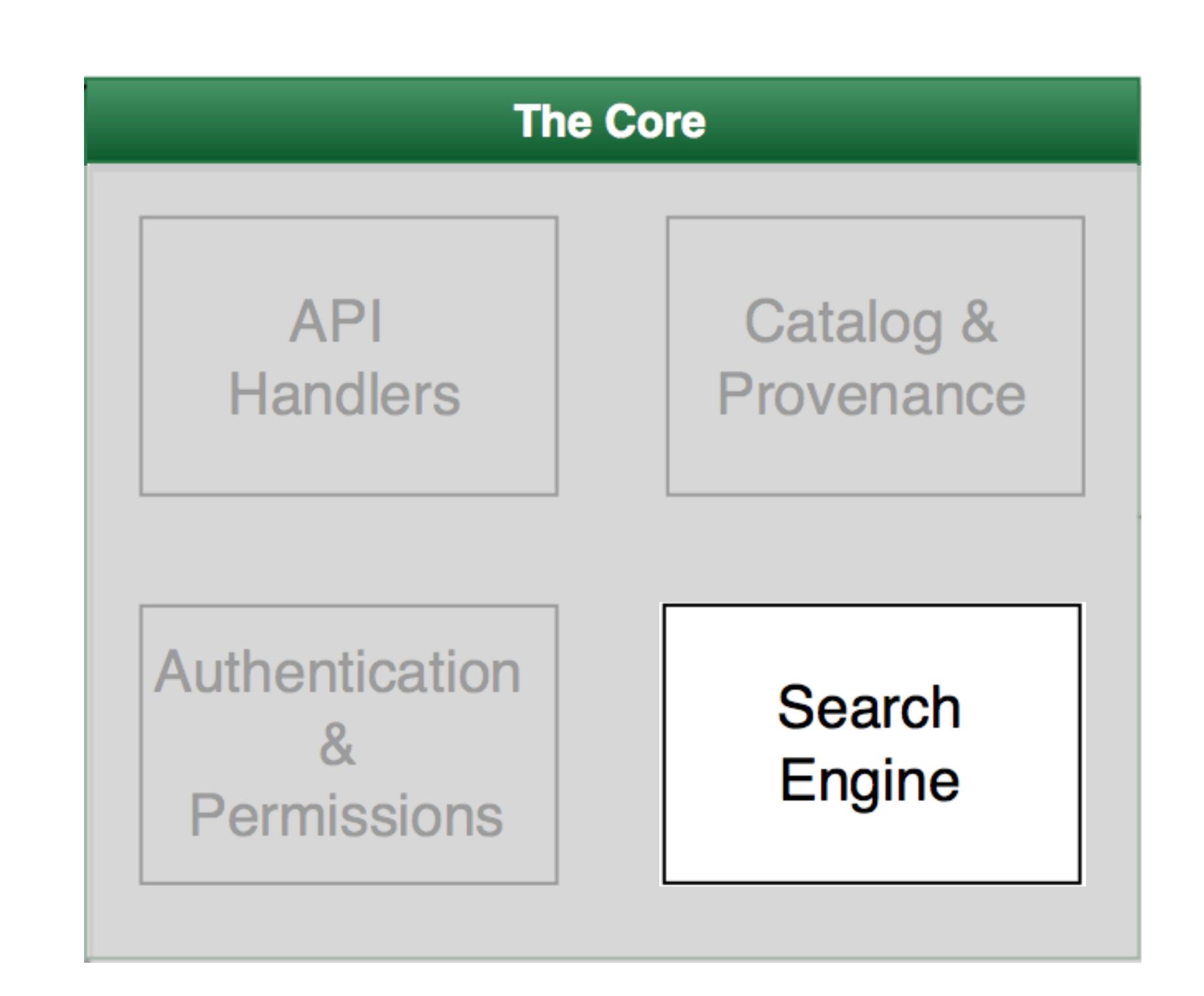
```
GET /dataset/42 HTTP/1.1
```

```
[{
    "jid": 23,
    "owner_id": 4,
    "inputs": [2,4,5,12],
    ...
},
...]
```

HTTP/1.1 200 OK



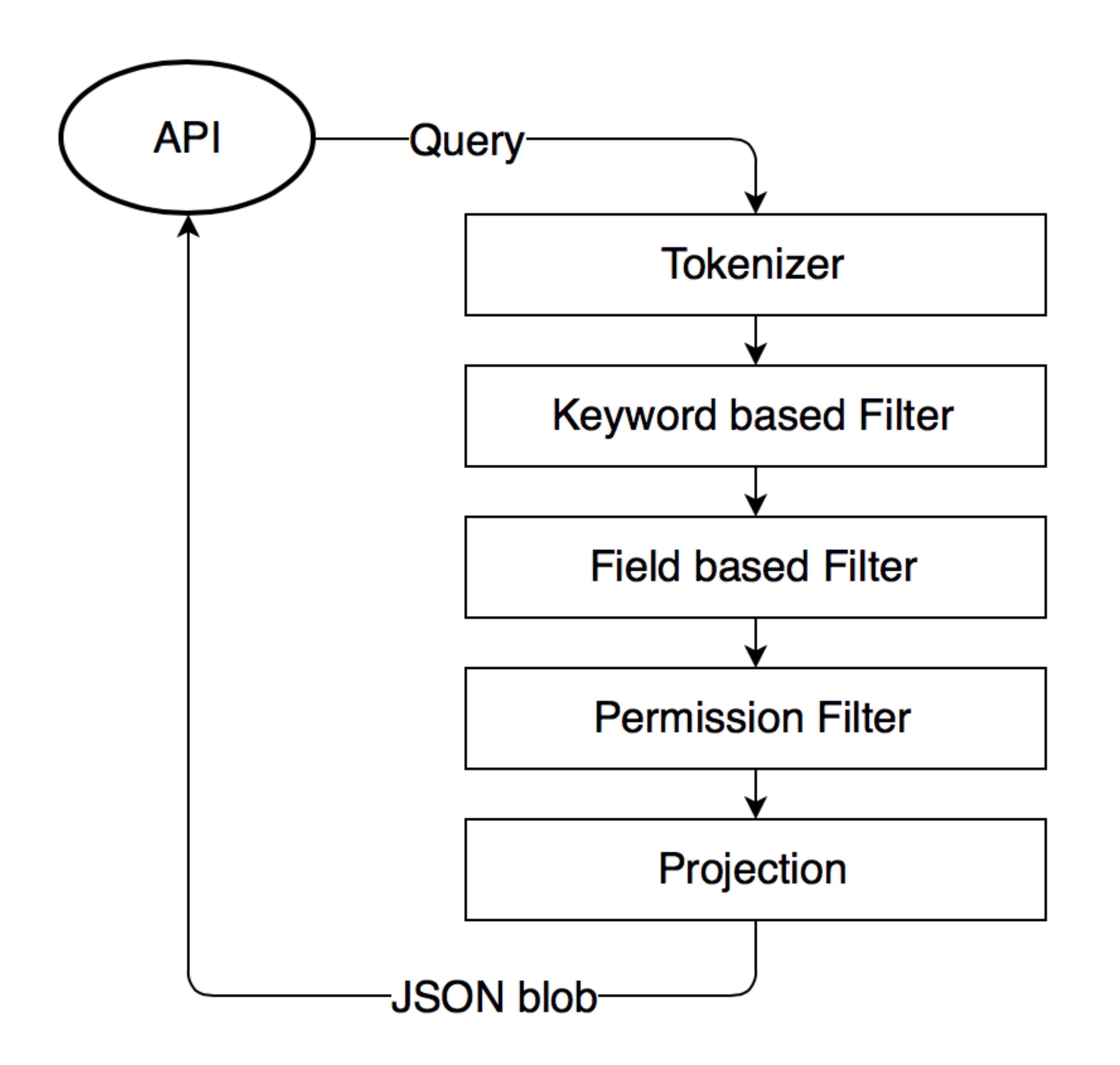
Search Engine



Indexes

Query Parsing

Record Filtering





Used Technologies

- GO 1.4.2
 - absoludity/goforms
 - jmoiron/sqlx
 - drone/routes
- PostgreSQL 9.4.1
 - JSON Type



Conclusion

Data Storage

Workflow Management

Sharing (Access Control)

Provenance Tracking



Thank you!

Questions?



Back-up slides



Why GO?

Concurrency primitives

Fast compilation and deploy

Built-in features

Standard library



PostgreSQL JSON type

Native support

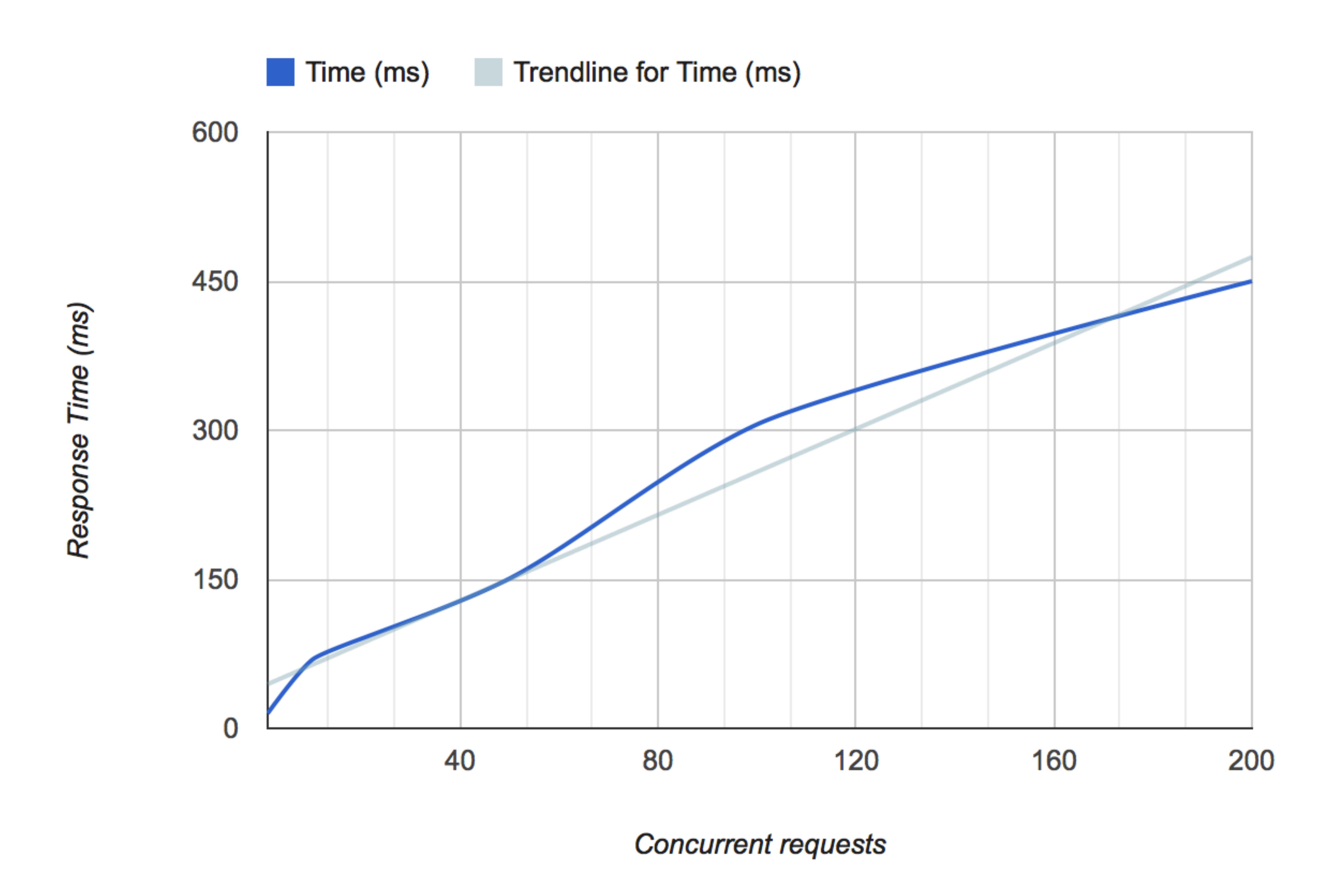
JSON validity verification

Indexes

Built-in management functionalities



Performance :: Load Test





Workflow

