CENTRO DE COMPUTAÇÃO CIENTÍFICA E SOFTWARE LIVRE

Educational Open Government Data: from requirements to end users

Rudolf Eckelberg, Vytor Bezerra Calixto, Marina Hoshiba Pimentel, Marcos Didonet del Fabro, Marcos Sunyé, Leticia Peres, Eduardo Todt, Thiago Alves, Adriana Dragone, and Gabriela Schneider

Departamento de Informática
Universidade Federal do Paraná
Curitiba – PR

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

June 7th, 2018

Index

Introduction

- Laws about open government data since 2008

Legislação

por Cintia de Freitas Rodrigues Loureiro — publicado 17/10/2017 11h57, última modificação 17/10/2017 11h57

G+ Tweet

- DECRETO Nº 8.777, DE 11 DE MAIO DE 2016 – Institui a Política de Dados Abertos do Poder Executivo federal;
- DECRETO DE 15 DE SETEMBRO DE 2011 - Institui o Plano de Ação Nacional sobre Governo Aberto e dá outras providências;
- Instrução Normativa SLTI nº 4, de 12 de abril de 2012 – Institui a Infraestrutura Nacional de Dados Abertos – INDA;
- Lei nº 12.527, de 18 de novembro de 2011 – Lei de acesso à informação;
- Decreto nº 7.724, de 16 de maio de 2012 – Regulamenta a Lei nº 12.527/2011;
- Decreto nº 6.666, de 27 de novembro de 2008 – Institui a Infraestrutura Nacional de Dados Espaciais – INDE.

Introduction

- Slowly publishing open data since 2011
- Many sources: ministries, government agencies, and others
  - Data scattered and difficult to integrate

Introduction

- Build a platform
  - Using **educational data**
    - All educational stages
  - For multiple segments of society
- Educational metrics supports:
  - Planning
  - Assessment
  - Research
  - Policy formulation
  - Public information

Introduction

- Joint effort
  - Center of Educational Policies (NuPE)
  - Center for Scientific Computation and Free Software (C3SL)
  - All students and professors
    - ~ 20 people

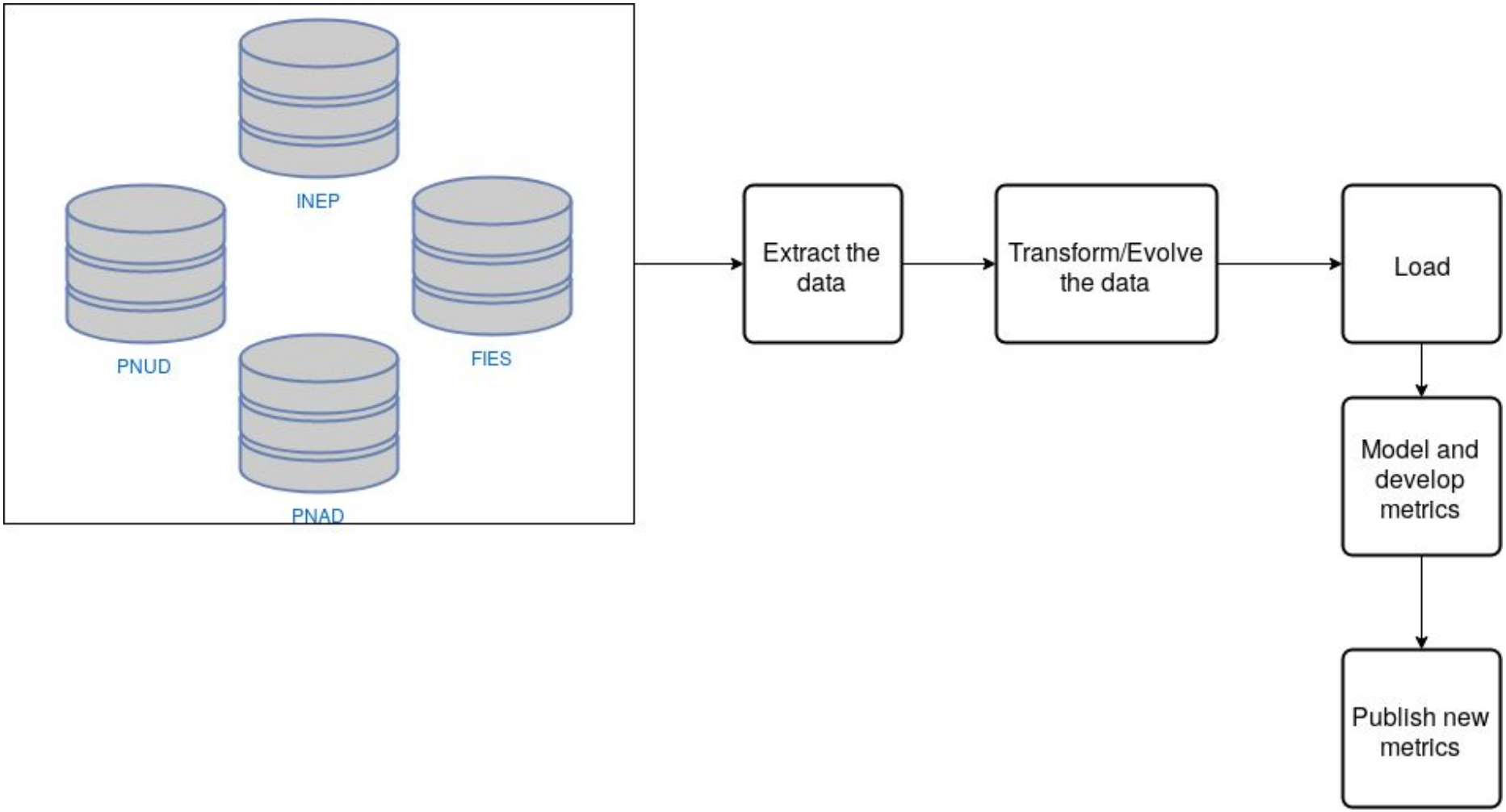
Introduction

One example: school infrastructure, historical evolution

	2013	2014	2015	2016	2017
Libraries	48%	48%	50%	49%	50%
Science labs	28%	28%	14%	14%	13%
Internet	14%	16%	24%	30%	33%
Electrical power	96%	96%	96%	97%	97%
Sewage treatment	95%	95%	95%	95%	96%

Many others: number of enrollments, teachers, sections and combinations of filters

Application Data Flow



Application Data Requirements

Open data:

- Different sources
- The data is yearly updated
  - Some smaller datasets are monthly
- Columns names and their values may evolve
  - It's necessary to preserve consistency through data transformations
- The largest datasets contains hundreds of columns (up to 170), millions of records

Currently we have more than 1 billion records used in 14 metrics

- How to model the data?
- How to evolve the data?
  - Compromise between normalization and data/code evolution

Application API Requirements

WEB API

- Needs to support every filter and dimension combination
- Generate appropriate queries
- Results used in multiple web clients
- New metrics are released with a steady pace
- Outputs need to be synchronized with any changes on the database

WEB Client:

- Used by different segments
- User friendly interface
  - Display different results (sums, counts, averages, percentages, ...)
  - Different visualizations (tables, plots, maps, ...)

Database modeling/choice

- Data is extracted and injected into a DBMS
- Every year large CSVs are made available
- Information is not normalized
- No control over input fields and values
- De-normalized views

Choosing the DBMS

TPC-H benchmark

Query	Time in ms		MonetDB performance against PostgreSQL
	PostgreSQL	MonetDB	
1	14583	741	1968.02%
3	2056	2067	99.47%
4	942	234	402.56%
5	1406	1278	110.02%
6	1665	179	930,17%
10	3446	908	379.52%
12	2567	330	777.88%
14	1677	86	1950.00%
16	2047	204	1003.43%
19	2143	136	1575.74%

Data evolution

- Mapping file relate fields' meanings and names through years
  - Schema matching tools are more complex than needed
  - 1-to-1 relationships

Source name	Variable name	Description	Database name
NU_ANO_CENSO	ANO	Census year	ano_censo
ID_MATRICULA	CEBMA002N0	Enrollment id	id
CO_PESSOA_FISICA	CEBMA003N0	Student id	cod_aluno
NU_DIA	CEBMA004N0	Student birthday	nasc_dia

Data evolution

- Infrastructure
  - Years without broadband internet
  - Future years will have new information

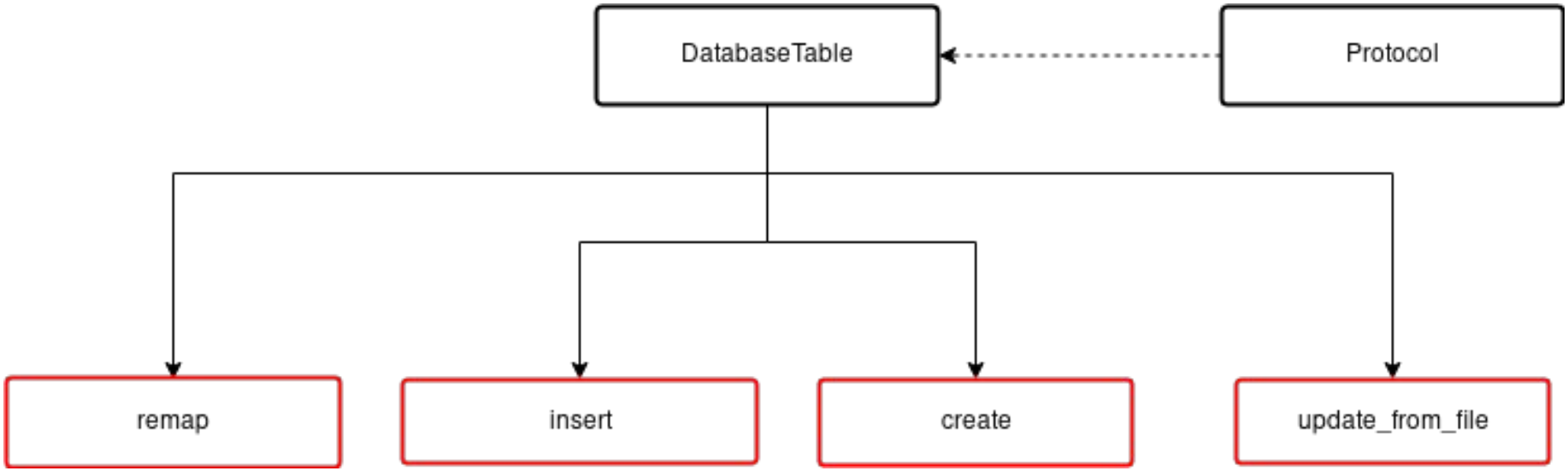
	2009	2010	2011	2012	2013
Broadband Internet	???	65%	68%	73%	73%

	2019	2020	2021	2022	2023
Pool	???	30%	27%	28%	28%

	2019	2020	2021	2022	2023
Dial-up internet	15%	8%	???	???	???

Data evolution CLI

- CLI tool to manage the input mapping
  - Python 3 and SQLAlchemy



Mapping protocol

Name	Standard name	Data type	2014	2015
ANO	NU_ANO_CENSO	INT	NU_ANO_CENSO	NU_ANO_CENSO
CEBES002N0	CO_ENTIDADE	INT	PK_CO_ENTIDADE	CO_ENTIDADE
CEBES003N0	NO_ENTIDADE	VARCHAR(256)	NO_ENTIDADE	NO_ENTIDADE

```
~CASE WHEN (cod_escolaridade = 1 OR cod_escolaridade = 2) THEN 1

WHEN (cod_escolaridade = 3) THEN 2 WHEN (cod_escolaridade = 4 OR
cod_escolaridade = 5) THEN 3

WHEN (cod_escolaridade = 6 AND "ID_DOUTORADO" = 1) THEN 8

WHEN (cod_escolaridade = 6 AND "ID_MESTRADO" = 1) THEN 7

WHEN (cod_escolaridade = 6 AND "ID_ESPECIALIZACAO" = 1) THEN 6

WHEN (cod_escolaridade = 6 AND ("ID_LICENCIATURA_1" = 1 OR
"ID_LICENCIATURA_2" = 1 OR "ID_LICENCIATURA_3" = 1)) THEN 5

WHEN (cod_escolaridade = 6) THEN 4 END
```

Table definition

```
{
    "pairing_description": "Schools table",
    "data_source": "Microdados do Censo Escolar/INEP (arquivo
Escolas)",
    "primary_keys": ["year", "id"],
    "foreign_keys": [
        {
            "keys": ["city_id"],
            "reference_columns": ["id"],
            "reference_table": "cities"
        },
        {
            "keys": ["state_id"],
            "reference_columns": ["id"],
            "reference_table": "state"
        },
        {
            "keys": ["region_id"],
            "reference_columns": ["id"],
            "reference_table": "region"
        }
    ]
}
```

table_definition\school.json

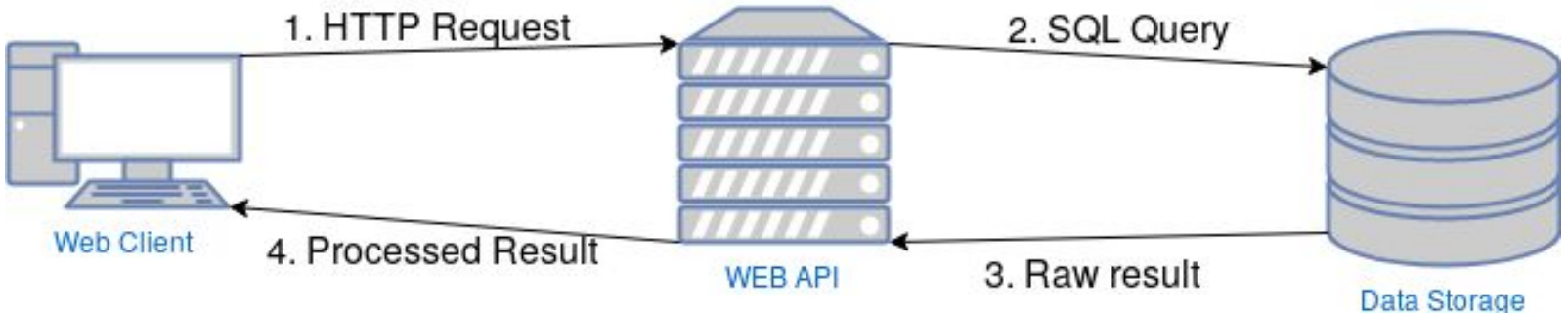
Metrics description

- Domain experts at NuPE provides complete descriptions for metrics (aka. indicators)

Name	Dimensions & filters	Periodicity	How to calculate	Free text explanation	Source
Enrollments count	• City, State, Region • Class period • Year	Yearly	Count total enrollments	Number of students enrolled	School census/INEP

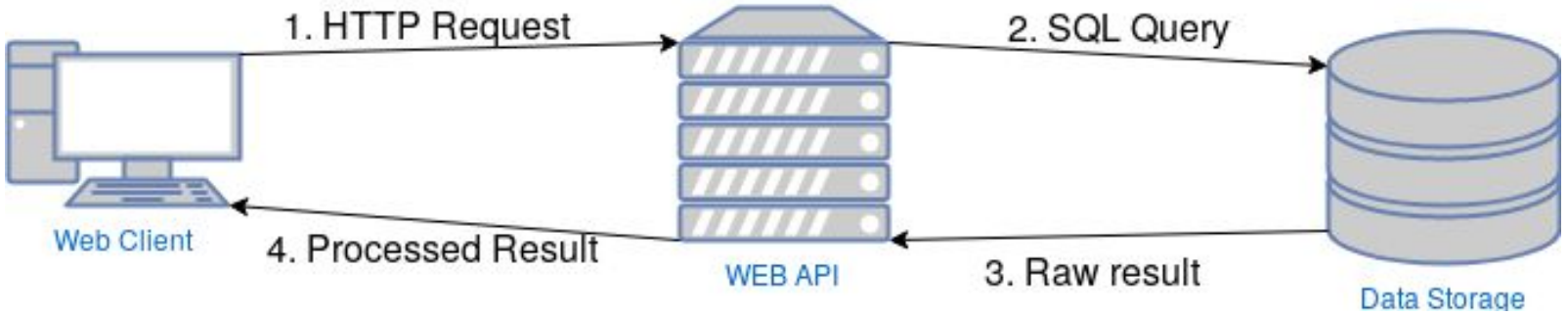
WEB API

- NodeJS + Express
- Provide results to multiple clients
- Metrics definition
  - Need to support data evolution

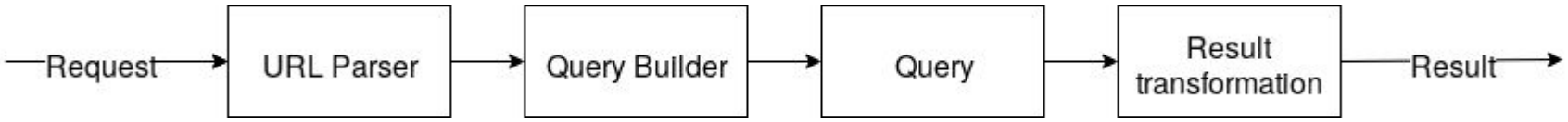


WEB API

- Request to a metric URL
  - `HTTP GET site.com/api/enrollment?filter=min_year:2016`
- Translate request to SQL queries
- Query the database
- Process results and return to client



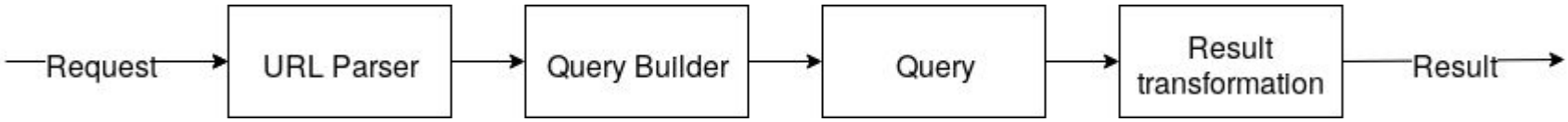
API flow



1. **Request**
   a. https://simcaq.c3sl.ufpr.br/api/v1/enrollment?dims=school_year,state&filter=min_year:2016,max_year:2016
2. **Parser:** identify filters and dimensions
   a. Dimensions: school_year and state
   b. Filters: min_year (2016) and max_year (2016)
3. **Query builder:** builds the query using squel.js
   a. Default query:

```
req.sql.field('COUNT(*)', 'total')
.field("'Brasil'", 'name')
.field('matricula.ano_censo', 'year')
.from('matricula')
.group('matricula.ano_censo')
.order('matricula.ano_censo')
.where('matricula.tipo<=3');
```

Query builder: dimensions definition

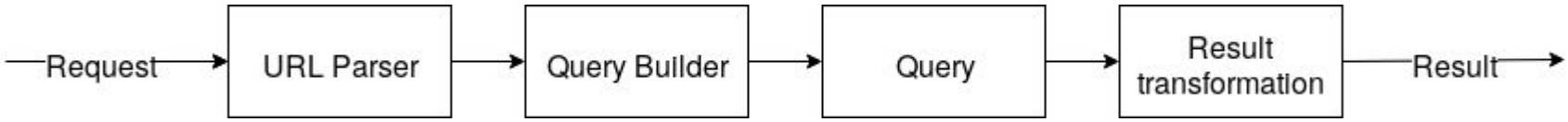


3. **Query builder**
   b. Add dimensions and filters

```
{
    name: 'state',
    table: 'estado',
    tableField: 'nome',
    resultField: 'state_name',
    where: {
            relation: '=',
            type: 'integer',
            field: 'id'
    },
    join: {
            primary: 'id',
            foreign: 'estado_id',
            foreignTable: 'matricula'
    }
}
```

```
{
    name: 'min_year',
    table: 'matricula',
    tableField: 'ano_censo',
    resultField: 'year',
    where: {
        relation: '>=',
        type: 'integer',
        field: 'ano_censo'
    }
}
```

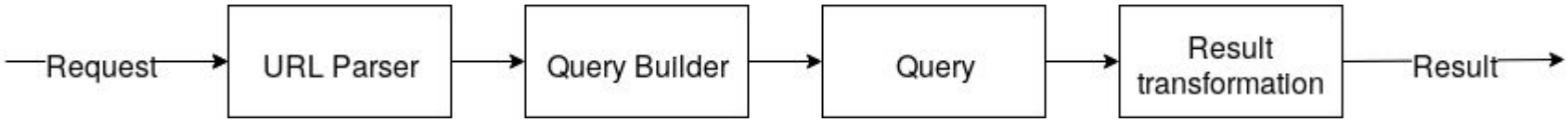
Building the queries



3. **Query builder**
   c. Build final query

```
SELECT
    matricula.serie_ano_id AS "school_year_id",
    estado.nome AS "state_name",
    COUNT(*) AS "total",
    'Brasil' AS "name",
    matricula.ano_censo AS "year"
FROM matricula INNER JOIN
    estado ON (matricula.estado_id=estado.id)
WHERE (matricula.ano_censo >= 2016 )
    AND (matricula.ano_censo <= 2016 )
    AND (matricula.tipo<=3)
GROUP BY matricula.serie_ano_id, estado.nome,
    matricula.ano_censo
ORDER BY matricula.serie_ano_id ASC, estado.nome ASC,
    matricula.ano_censo ASC;
```

WEB API response



4. **Query:** queries the database
5. **Result transformation:** light processing on the data to be returned
6. **Returns:** JSON, XML, CSV

```
{
    “result”: [
        {
            “school_year_id”:11,
            “state_name”: “Acre”,
            “total”: 257,
            “name”: “Brasil”,
            “year”: 2016,
            “school_year_name”: “Creche - Menor de 1 ano”
        },
        ...
    ]
}
```

WEB Interface

- React and Redux

Client only needs to :

- Know how to request the API

```
https://SITE/api/METRIC?dims=FIRST,SECOND,THIRD,N
    &filters=FILTER1:VALUE1, FILTER2:VALUE2,
FILTERN:VALUEN
    &format=JSON (or CSV or XML)
```

- How to display the result
- Response tables adapt to the dimensions and filters

Por uma educação pública gratuita e de qualidade para todos/as.

Consulta de Indicadores

1 SELECIONE A LOCALIDADE

2 SELECIONE O PERÍODO

Observação: para consulta em série histórica só poderá selecionar uma variável de resultado (etapa 3)

- [ ] Ano Específico
- [x] Série Histórica

2013

2016

3 MONTE SUA CONSULTA

LIMPAR CONSULTA MOSTRAR RESULTADO

NÚMERO DE MATRÍCULAS

Baixar

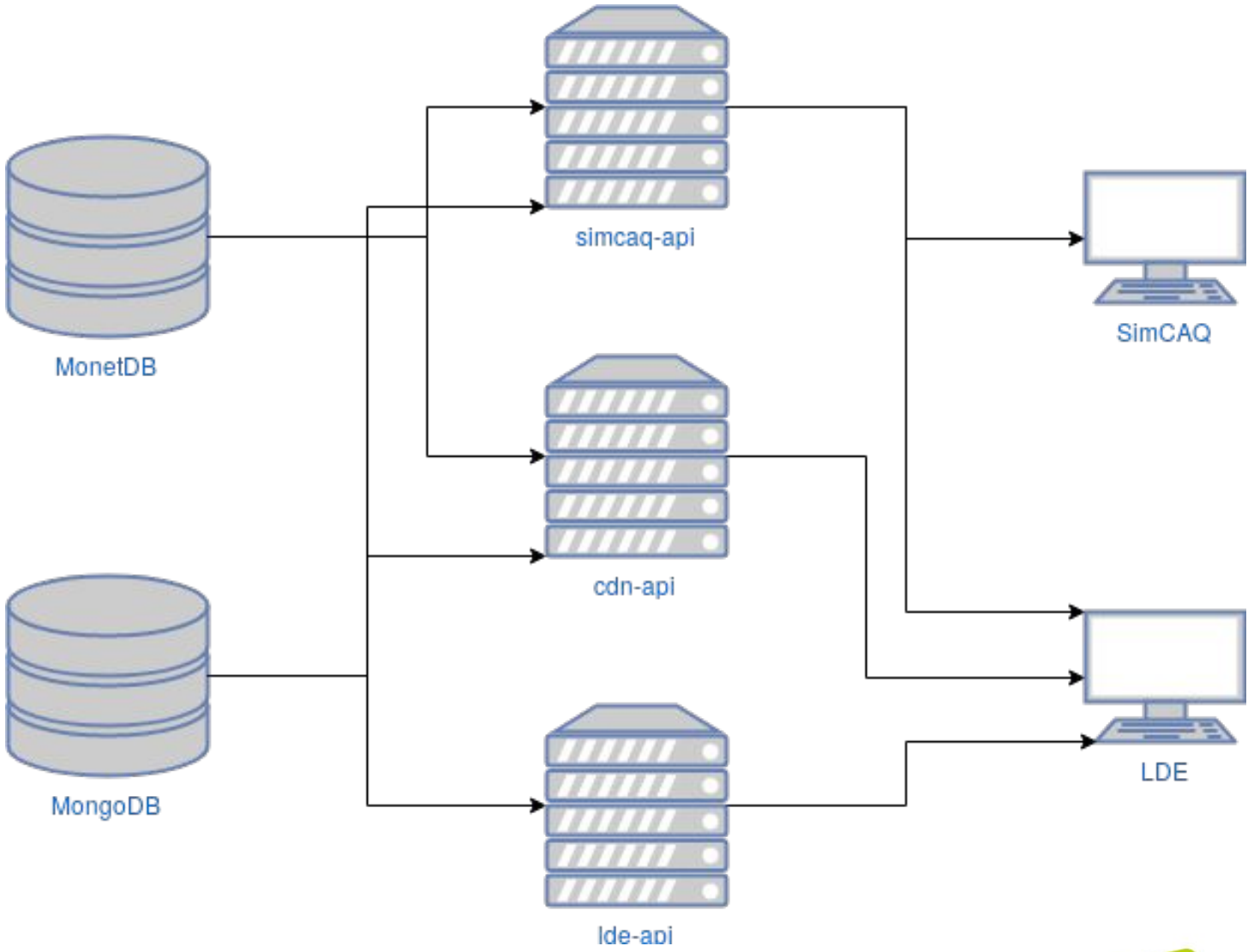
Acesso à Educação Básica

Número de Matrículas por Dependência Administrativa - Brasil, 2013 a 2016

Dependência Administrat...	2013	2014	2015	2016
Federal	290.796	296.745	376.230	392.565
Estadual	17.926.568	17.294.357	16.548.708	16.595.631
Municipal	23.215.052	23.089.488	22.813.842	22.846.182
Privada	8.610.032	9.090.781	9.057.732	8.983.101
Total	50.042.448	49.771.371	48.796.512	48.817.479

Fonte: Elaborado pelo Laboratório de Dados Educacionais/UFPR a partir dos microdados do Censo Escolar/INEP 2013 - 2016

Platform architecture



Conclusion and Lessons Learned

- Dealing with open government data is difficult
  - Communication between different teams
  - Database modelling and evolution
- A single metric could be simple, several aren't
  - Constant release of new metrics
  - Constant update of all metrics
- Input dependency diminishes the freedom of data modeling
  - No guarantee of format or specification between years
  - Each release needs data matching/mapping
  - Easier to develop a CSV based mapping format than an automatic solution
- With a column store there's no need of intermediary tables
- Such approach is viable and it is currently deployed
  - http://dadoseducacionais.c3sl.ufpr.br
    - https://dadoseducacionais.c3sl.ufpr.br/#/indicadores

www.c3sl.ufpr.br