



Centro de Computação Científica e Software Livre

Laboratório de Dados Educacionais

Fernando Claudecir Erd - fcerd@inf.ufpr.br
Pedro Demarchi Gomes - pdg16@inf.ufpr.br

Latinoware 2019

- Centro de Computação Científica e Software Livre (C3SL)
- Dados Abertos
- Dados Educacionais
- Laboratório de Dados Educacionais
- SimCAQ
- Arquitetura Geral
- Banco de Dados
- Mapeamento
- SimCAQ-API

C3SL: Centro de Computação Científica e Software Livre

- Grupo de pesquisa do Departamento de Informática da UFPR
- Registrado no diretório de grupos de pesquisa do CNPq

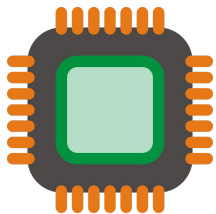
Responsabilidades:

- Gerenciamento do núcleo da rede do Departamento de Informática
- Desenvolvimento de **projetos de pesquisa** com caráter multidisciplinar direcionados à **inclusão digital** e benefício geral da sociedade

- Mais de 8 projetos em andamento
- Mais de 5 projetos ainda não divulgados

Mais informações em www.c3sl.ufpr.br/projetos/

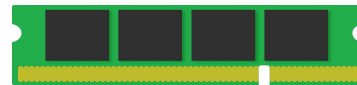
- LDE - Laboratório de Dados Educacionais
<https://dadoseducacionais.c3sl.ufpr.br>
- BIOD - Blended Integrated Open Data
<https://biod.c3sl.ufpr.br>
- Portal MEC
<https://plataformaintegrada.mec.gov.br>



Mais de 2000
núcleos de
processamento



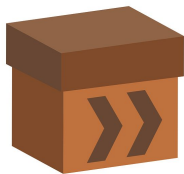
1 *Petabyte* de
armazenamento
(3,4 anos de FULL HD 24/7)



Mais de 10
Terabytes de RAM
(1000+ notebooks [8 Gb de RAM])



Rede de até 20
Gbps
(~667x conexão doméstica [30 MBps])¹



Mais de 2000 pacotes
de software



Mais de 500 pontos
de trabalho



Equipe inter e
multidisciplinar



14 professores
doutores



Média de
35 a 40 bolsistas
(graduação e pós)

¹ <https://datareportal.com>

O C3SL mantém espelhos (*mirrors*) dos principais repositórios de *software livre* (SL)

- Maior espelho não-comercial de SL do hemisfério sul e um dos maiores do mundo (tanto em repositórios hospedados quanto em fluxo de download)
- Exige investimento significativo em hardware, rede, configuração e administração
- Para manter o espelho sincronizado com as origens mesmo sob alta carga foi desenvolvido um processo de atualização dos repositórios mais eficiente que o tradicional

E-mail de contato: contato@c3sl.ufpr.br

- Dados são abertos quando qualquer pessoa pode livremente **acessá-los, utilizá-los, modificá-los e compartilhá-los** para qualquer finalidade, estando sujeito a, no máximo, a exigências que visem preservar sua proveniência e sua abertura.
- Governo
 - Dados educacionais, econômicos, folha de pagamento, etc.

Legislação

por [Cintia de Freitas Rodrigues Loureiro](#) — publicado 17/10/2017 11h57, última modificação 17/10/2017 11h57

[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.

Organizações

Banco Central do Br... (9)

Ministério da Justi... (3)

Agência Nacional de... (2)

Fundação Nacional d... (1)

Instituto Brasileir... (1)

Instituto Federal d... (1)

Instituto Federal d... (1)

Instituto Nacional ... (1)

Ministério do Desen... (1)

Secretaria de Gover... (1)

Grupos

Geografia (1)

Etiquetas

inclusão financeira (9)

Paraná

21 conjuntos de dados encontrados para "Paraná" Ordenar por: Relevância

Diagnóstico dos municípios do estado do Paraná

Diagnóstico apresenta o número de óbitos e índices vulnerabilidade social, bem como acesso a água e esgoto, dos jovens entre 15 e 29 anos.

PDF

Massa d' Água Região Hidrográfica Paraná

Acesse os Metadados O dado compreende todas as massas d'água da Região Hidrográfica Paraná, classificadas basicamente segundo a tipologia: naturais e artificiais e o domínio...

HTML Esri REST GeoJSON CSV KML ZIP

Massa d' Água Região Hidrográfica Paraná

Acesse os Metadados O dado compreende todas as massas d'água da Região Hidrográfica Paraná, classificadas basicamente segundo a tipologia: naturais e artificiais e o domínio...

HTML Esri REST GeoJSON CSV KML ZIP

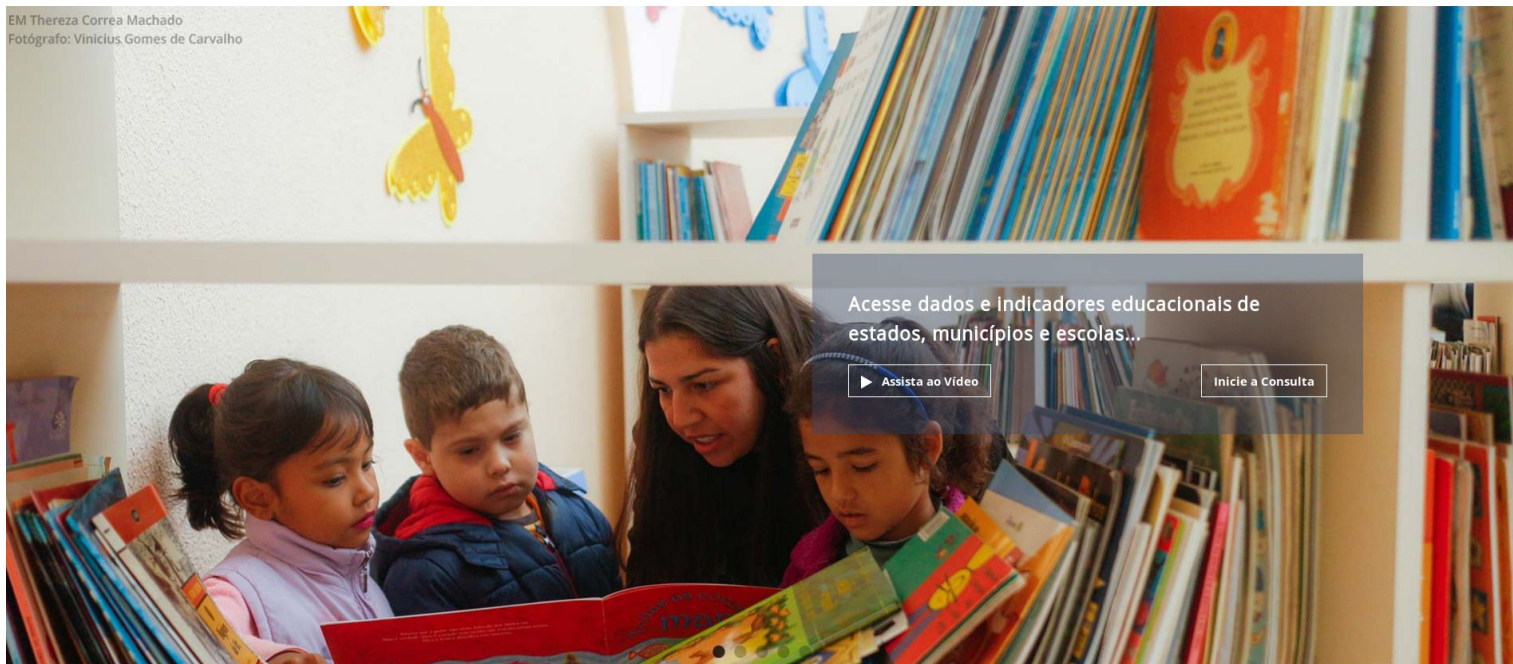
- Muitos dados abertos disponíveis
- Muitas fontes
- Ministérios, agências governamentais, universidades, órgãos públicos.
- Muitos formatos: CSV, PDF, JSON, XLS, HTML, DOCX, etc.
- Muitos arquivos diferentes tamanhos.
- Dados educacionais: arquivos com mais de 4Gb

- INEP é um grande produtor de dados
- Produz: Censo Escolar, Educação Superior, Enade, ENEM, etc.
- Arquivo de matrículas da educação básica = ~4Gb
 - + 50 milhões de linhas, por ano
 - Tabelas com mais de 100 colunas

- Plataforma usando dados educacionais
 - Todos os estágios (creche até ensino superior)
 - Múltiplos segmentos interessados
- Indicadores educacionais para
 - Planejamento
 - Pesquisa
 - Criação de políticas públicas
 - Publicização da informação

- Objetivo
 - Facilitar o acesso e o uso de dados e indicadores para a formulação, implementação, monitoramento e avaliação de políticas educacionais
 - Estimular pesquisas que visem à elaboração de indicadores educacionais
 - Potencializar pesquisa a partir do uso de dados e indicadores disponibilizados em séries temporais e em diferentes níveis de desagregação e formatos
- Tecnologia
 - React
 - Redux
 - SASS

EM Thereza Correa Machado
Fotógrafo: Vinicius Gomes de Carvalho



Acesse dados e indicadores educacionais de
estados, municípios e escolas...

▶ Assista ao Vídeo

Inicie a Consulta

Consulta de Indicadores

1 SELECIONE A LOCALIDADE



2 SELECIONE O PERÍODO



3 MONTE SUA CONSULTA



Selecione as informações para visualizar os resultados nas **linhas** e **colunas** da tabela*:

Tipo de escola que o aluno concluiu o Ensino Médio



Coluna: selecione uma variável



	Colunas	
Linhas		

LIMPAR CONSULTA

MOSTRAR RESULTADO

REFINE SUA CONSULTA



LIMPAR FILTROS

REFINAR

NÚMERO DE MATRÍCULAS

Educação Superior

Baixar

Número de Matrículas por Tipo de escola que o aluno concluiu o Ensino Médio - UNIVERSIDADE FEDERAL DA INTEGRAÇÃO LATINO-AMERICANA, 2018

Tipo de escola que o aluno concluiu o Ensino Médio	Total
Pública	2.442
Privada	1.179
Não classificado	8
Total	3.629

Fonte: Elaborado pelo Laboratório de Dados Educacionais a partir dos Microdados do Censo de Educação Superior/INEP 2018

NÚMERO DE PROFESSORES

Educação Básica

 Baixar

Número de Professores, dependência administrativa (estadual) por Formação do professor - FOZ DO IGUAÇU, 2018

Formação do professor	Total
Ensino Médio	4
Superior	3
Superior com licenciatura	42
Especialização	878
Mestrado	55
Doutorado	4
Total	986

Fonte: Elaborado pelo Laboratório de Dados Educacionais a partir dos microdados do Censo Escolar/INEP 2018

Nota: Um(a) professor(a) pode ser contado(a) mais de uma vez, se atuar em mais de uma unidade de agregação: regiões, unidades da federação, municípios, área de localidade, dependência administrativa e etapa/modalidade. Portanto, o total representa o número de professores em unidades de agregação diferentes.

NÚMERO DE MATRÍCULAS

Educação Básica

 Baixar

Número de Matrículas por Dependência Administrativa - FOZ DO IGUACU, 2013 a 2018

Dependência Administrativa	2013	2014	2015	2016	2017	2018
Federal	449	398	344	776	755	659
Estadual	31.477	31.211	30.132	27.792	27.644	27.881
Municipal	24.548	24.577	24.064	25.643	26.962	26.363
Privada	12.492	13.520	13.032	13.027	12.888	13.168
Total	68.966	69.706	67.572	67.238	68.249	68.071

Fonte: Elaborado pelo Laboratório de Dados Educacionais a partir dos microdados do Censo Escolar/INEP 2013 - 2018

Educação Básica

Número de Matrículas	Período Disponível: 2013 - 2018	Ficha Técnica	CONSULTAR
Número de Escolas	Período Disponível: 2013 - 2018	Ficha Técnica	CONSULTAR
Número de Turmas	Período Disponível: 2013 - 2018	Ficha Técnica	CONSULTAR
Número de Professores	Período Disponível: 2013 - 2018	Ficha Técnica	CONSULTAR
Número de Auxiliares Docentes	Período Disponível: 2013 - 2017	Ficha Técnica	CONSULTAR
Número de Funcionários	Período Disponível: 2013 - 2018	Ficha Técnica	CONSULTAR
Taxa de Atendimento	Período Disponível: 2004 - 2015	Ficha Técnica	CONSULTAR
População Fora da Escola	Período Disponível: 2007 - 2015	Ficha Técnica	CONSULTAR
Taxa de Matrícula Líquida	Período Disponível: 2013 - 2015	Ficha Técnica	CONSULTAR
Taxa de Matrícula Bruta	Período Disponível: 2013 - 2015	Ficha Técnica	CONSULTAR

Educação Superior

Número de Matrículas	Período Disponível: 2010 - 2018	Ficha Técnica	CONSULTAR
Número de Instituições de Educação Superior	Período Disponível: 2010 - 2018	Ficha Técnica	CONSULTAR
Número de Cursos	Período Disponível: 2010 - 2018	Ficha Técnica	CONSULTAR
Número de Docentes da Educação Superior	Período Disponível: 2010 - 2018	Ficha Técnica	CONSULTAR

- Esforços conjuntos
 - Especialistas do domínio: Laboratório de Dados Educacionais
 - Especialistas em computação: C3SL/UFPR
- Equipe
 - ~15 integrantes, entre alunos e professores
 - Comunicação é importante

Simulador de Custo Aluno-Qualidade (SimCAQ)

- Sistema gratuito e disponível na internet que estima o custo da oferta de ensino em condições de qualidade nas escolas públicas de educação básica, ou seja, o Custo-Aluno Qualidade (CAQ).
- Objetivos
 - Oferecer suporte ao processo de elaboração/adequação e monitoramento/avaliação dos Planos Estaduais e Municipais de Educação, visando a articulação das metas educacionais locais com as metas do Plano Nacional de Educação (PNE)
 - Fomentar pesquisas e publicações sobre financiamento da educação, custos educacionais e oferta da educação básica em condições de qualidade;
 - Promover a interação entre pesquisadores, gestores das redes de ensino, governos, profissionais da educação e sociedade civil para debater os desafios para o cumprimento da meta 20 do PNE 2014-2024.

Simulador de Custo Aluno-Qualidade (SimCAQ)

- Implementado em Angular
- Front-end desenvolvido pelo Laboratório Media Lab/UFG
- Rotas de indicadores são usadas em conjunto com o Dados Educacionais
- Possui rotas específicas

Simulador de Custo Aluno-Qualidade (SimCAQ)

Quanto custa uma educação pública de qualidade?

Custo-aluno Qualidade

Estime o Custo-Aluno Qualidade (CAQ) de cada etapa/modalidade

Acessar

O que entendemos por uma educação de qualidade

[Leia mais](#)

Orçamento educacional

Estime o orçamento necessário para financiar a educação na sua localidade

Simular

Simulador de Custo Aluno-Qualidade (SimCAQ)

Resultado

Você está consultando os valores do CAQ para BRASIL / 2019

Os resultados do SimCAQ decorrentes dos parâmetros de qualidade (PQR) propostos pela equipe do projeto. Desse modo, não são decisões do governo federal, dos governos estaduais ou municipais. Por consequência, não vinculam responsabilidade de repasses de recursos financeiros por parte dos entes federativos.

Etapa	Área da localidade	Turno	CAQ (R\$)	Valor-aluno Fundeb (R\$) ⁽³⁾	
			2019	Menor ⁽¹⁾	Maior ⁽²⁾
				2019	2019
Creche	Urbana	Parcial	10.622	3.724	4.917
		Integral	19.132	4.210	5.558
	Rural	Parcial	17.174	3.724	4.917
		Integral	25.450	4.210	5.558
Pré-Escola	Urbana	Parcial	6.324	3.400	4.489
		Integral	11.147	4.210	5.558
	Rural	Parcial	8.570	3.400	4.489
		Integral	17.352	4.210	5.558
Ensino Fundamental - anos iniciais	Urbana	Parcial	5.690	3.239	4.275
		Integral	7.666	4.210	5.558
	Rural	Parcial	7.833	3.724	4.917
		Integral	11.697	4.210	5.558

Simulador de Custo Aluno-Qualidade (SimCAQ)

Você está elaborando o orçamento educacional para FOZ DO IGUAÇU / 2019

Os resultados do SimCAQ decorrentes dos parâmetros de qualidade (PQR) propostos pela equipe do projeto. Desse modo, não são decisões do governo federal, dos governos estaduais ou municipais. Por consequência, não vinculam responsabilidade de repasses de recursos financeiros por parte dos entes federativos.

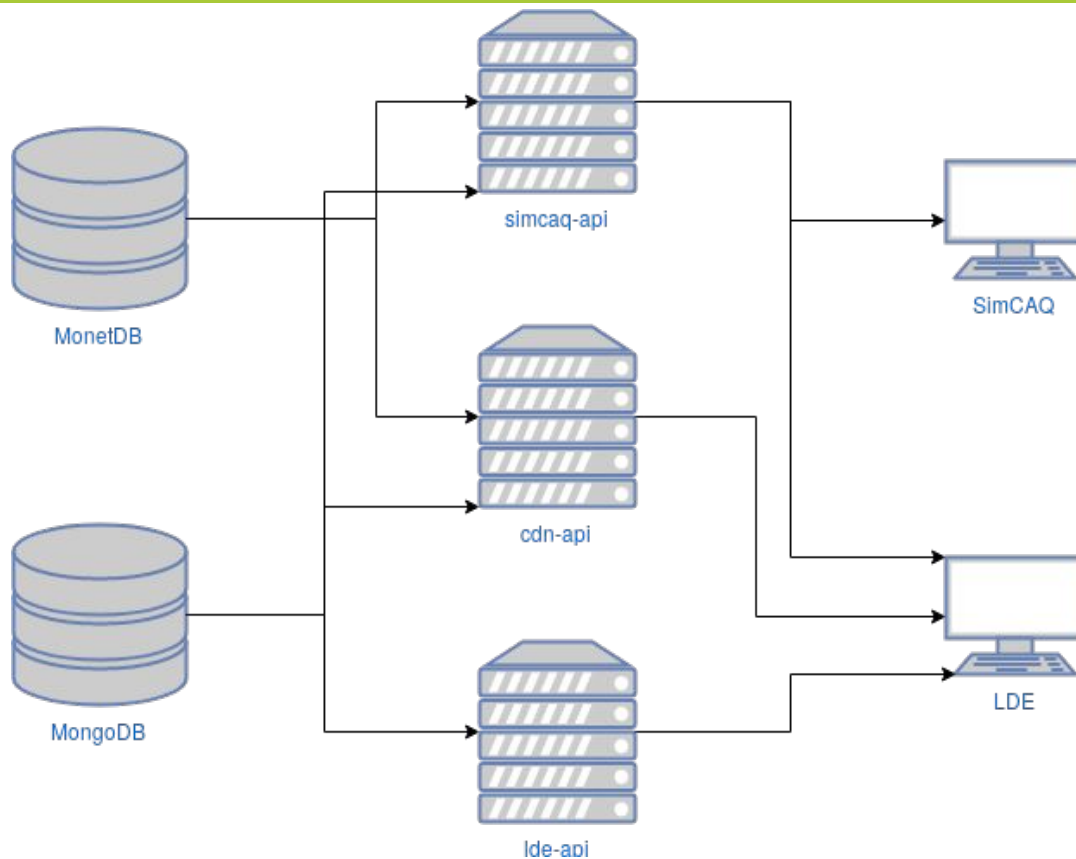
Dimensão da oferta	Atual	Projeção	Variação 2018-2019 (1)
	2018	2019	
Número de matrículas	53.934	53.934	0 %
Número de turmas	2.210	2.369	7,19 %
Número de salas	1.339	1.516	13,22 %
Número de professores ⁽²⁾	2.097	2.400	14,45 %
Número de auxiliares	224		

Notas:

⁽¹⁾ Variação de valores entre o diagnóstico e o projetado.

⁽²⁾ O resultado da projeção do número de professores necessários depende do padrão de jornada de trabalho semanal adotado. Inicialmente, o simulador faz projeções considerando uma jornada de 40 horas semanais. Se necessário, edite o resultado e altere esse parâmetro de modo a expressar a jornada dos docentes da rede em análise.

Arquitetura Geral



- Dado é extraído e incluído em um SGBD
- Todos os anos novos CSV são disponibilizados
- Dado não é normalizado
- Não há controle no nome dos campos e compatibilidade dos dados

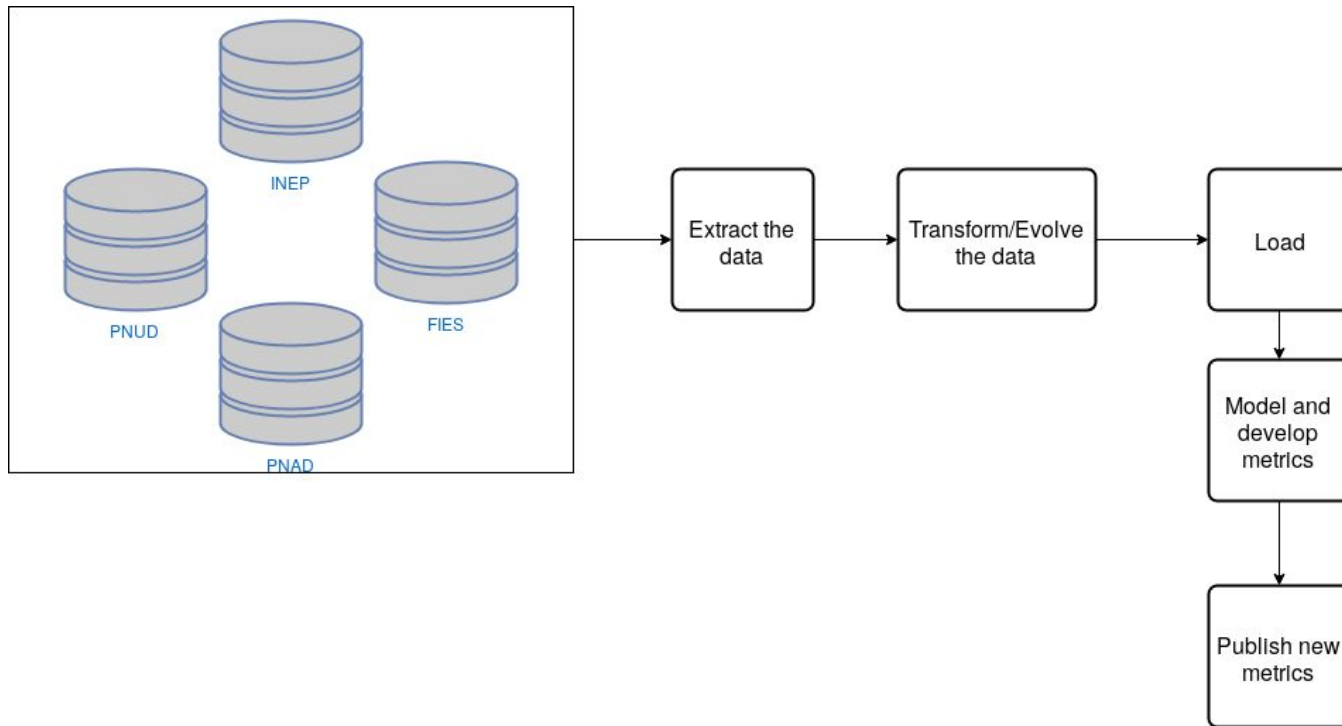


PostgreSQL



Consulta	Tempo em ms		Desempenho do MonetDB comparado com 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%

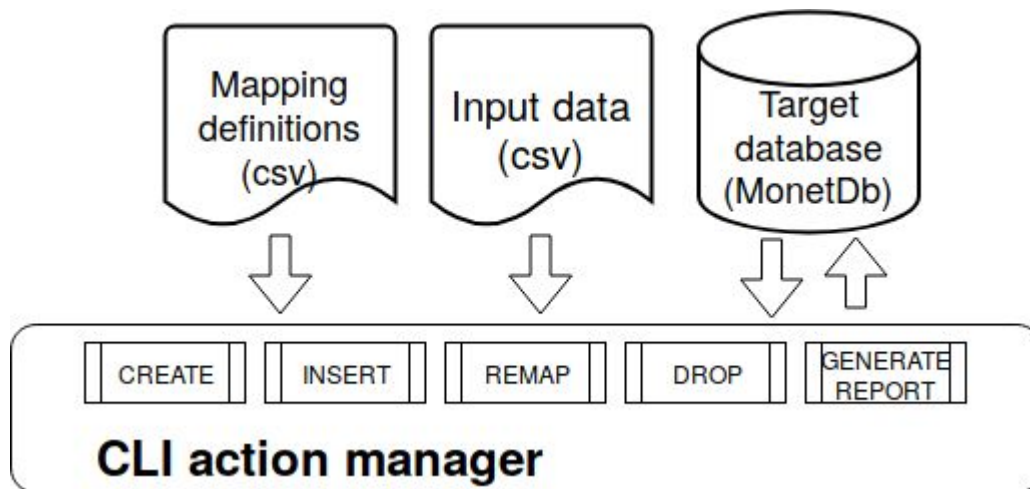
Fluxo para análise dos dados abertos



- Várias fontes
- Dado atualizado anualmente
 - Algumas fontes são mensais
- Nomes das colunas e seus valores variam
 - Dado deve ser mantido consistente ao longo do tempo
- Tabelas com mais de 150 colunas e milhões de registros
- Como modelar os dados?
- Como evoluir os dados?
 - Compromisso entre normalização e evolução/manutenção

- Dados de-normalizados
- Mapeamentos diretos
 - NACIONALIDADE <- [2013-2017] NACIONALIDADE
- Mapeamentos com traduções
 - NECESSIDADES_ESPECIAIS <- [2013-2014] TEM_NECESSIDADE
 - NECESSIDADES_ESPECIAIS <- [2015-2017] NECISSIDADE_ESP
 - REGIAO <- [2013-2015] não-disponível
 - REGIAO <- [2016-2017] REGIAO
- Evolução dos dados
 - PROFISSIONALIZANTE <-[2013-2014]
 - WHEN TIPO_ESTUDO between 30 and 40 THEN 1
 - WHEN TIPO_ESTUDO between 41 and 50 THEN 2

- Mapeamento através de scripts
- Mapeamentos em CSV
 - Subconjunto de SQL - expressões CASE
- CLI (Command Line Interface) fácil de usar
- Diferentes bancos podem ser plugado



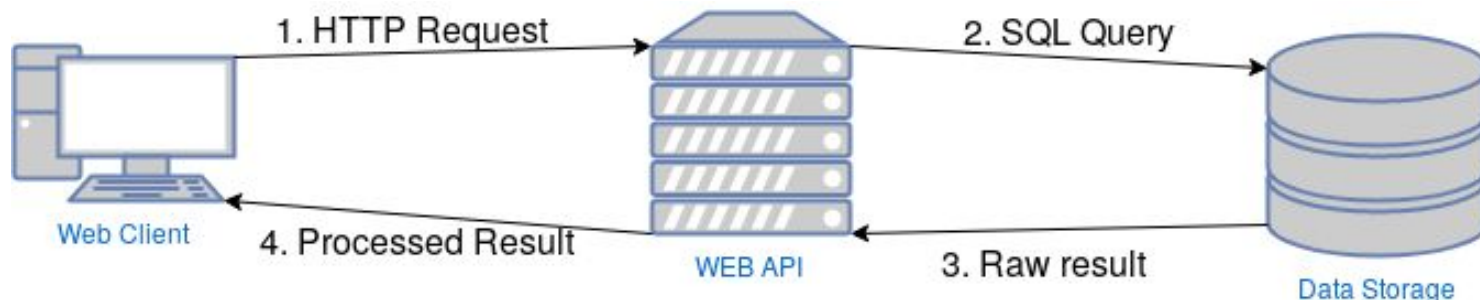
Nome	Nome padrão	Tipo de dado	2014	2015
ANO	NU_ANO_CENSO	INT	NU_ANO_CENSO	NU_ANO_CENSO
CEBES002N0	CO_ENTIDADE	INT	PK_CO_ENTIDADE	CO_ENTIDADE
CEBES003N0	NO_ENTIDADE	VARCHAR(256)	NO_ENTIDADE	NO_ENTIDADE

```

~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

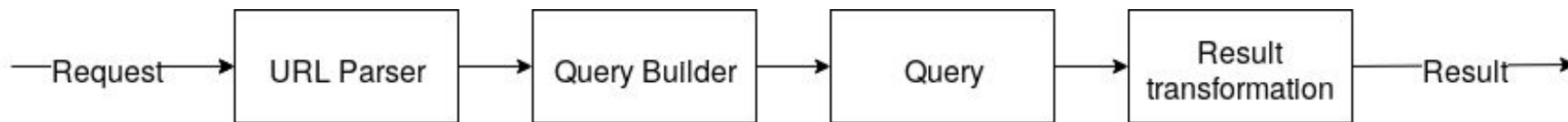
```

- NodeJs + Express
- Requisição para uma métrica
- Tradução para consulta SQL -> tradução direta
- Requisição é enviada
- Retorno em JSON ou CSV



- Filtros e combinação de dimensões
- Geração de consultas
- Reutilização por diferentes clientes web
- Novos indicadores lançados periodicamente
- Ex: `https://SITE/api/METRIC?dims=FIRST,SECOND,THIRD,N
&filters=FILTER1:VALUE1,FILTER2:VALUE2,FILTERN:VALUEN
&format=JSON (or CSV or XML)`

- Fluxo de execução



1. Requisição

- a. [https://simcaq.c3sl.ufpr.br/api/v1/enrollment?dims=education_level, state&filter=min_year:2018,max_year:2018](https://simcaq.c3sl.ufpr.br/api/v1/enrollment?dims=education_level,state&filter=min_year:2018,max_year:2018)

2. Parser: identifica filtros e dimensões

- a. Dimensões: education_level and state
- b. Filtros: min_year (2018) and max_year (2018)

3. Parser: validação e construção da query

```
query.addField({  
  name: 'dims',  
  field: true  
}).addValue({  
  name: 'education_level',  
  table: 'matricula',  
  tableField: 'etapa_ensino_id',  
  resultField: 'education_level_id',  
  where: {  
    relation: '=',  
    type: 'integer',  
    field: 'etapa_ensino_id'  
  }  
})
```

4. Builder: construção da query (SQL)

SELECT

```
matricula.etapa_ensino_id AS "etapa_ensino_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 >= 2018)

```
AND (matricula.ano_censo <= 2018 )
```

```
AND (matricula.etapa_ensino_id=1)
```

GROUP BY matricula.etapa_ensino_id, estado.nome,
matricula.ano_censo

ORDER BY matricula.etapa_ensino_id **ASC**, estado.nome **ASC**,
matricula.ano_censo **ASC**;

5. Executa a query
6. Retorna o resultado em JSON, XML ou CSV

```
{
  "result": [
    {
      "education_level_id": 1,
      "state_name": "Acre",
      "total": 11749,
      "name": "Brasil",
      "year": 2018,
      "school_year_name": "Educação Infantil - Creche"
    },
    ...
  ]
}
```

- Disponibilização de indicadores criados a partir de dados abertos é uma tarefa difícil
 - Muitas fontes de informação e formatos
 - Comunicação entre especialistas é **ESSENCIAL**
 - Não há garantia de consistência ao longo dos anos
 - Modelagem e evolução dos dados deve ser simples, para permitir atualização futura
 - Problema do “atualizado hoje”
- Uma única métrica pode ser simples, várias não
 - Indicadores novos, novos dados
- API REST separada do front-end



<https://dadoseducacionais.c3sl.ufpr.br>



<https://simcaq.c3sl.ufpr.br>



<https://gitlab.c3sl.ufpr.br/simcaq>



<https://www.c3sl.ufpr.br/>



Centro de Computação Científica e Software Livre

Obrigado

Fernando Claudécir Erd - fcerd@inf.ufpr.br
Pedro Demarchi Gomes - pdg16@inf.ufpr.br