Open platform of civil society organizations

Challenges, architecture, techniques and technologies in development of the most important system of IPEA

Bio

Technical coordinator (Researcher II) at IPEA

MBA Professor at INFNET

MSc student at PESC/UFRJ

Research interests: Data Mining & Machine Learning in Dynamic Environments, Complex Networks, Big Data for Social Development

Summary

Introduction

Challenges

Architecture

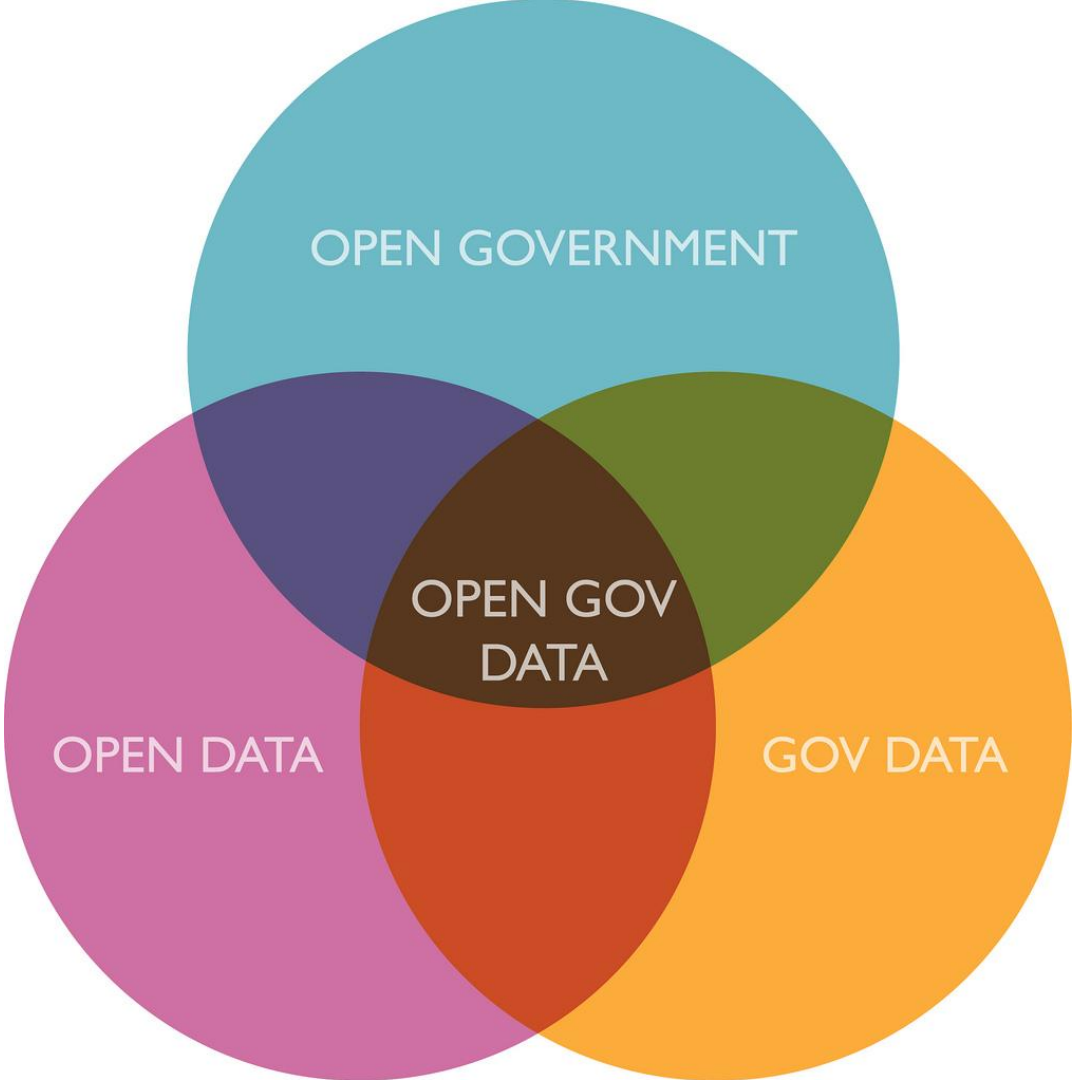
Techniques

Technologies

Current state

Future steps

Introduction


OPEN GOVERNMENT
OPEN GOV DATA
OPEN DATA
GOV DATA

IPEA

Instituto de Pesquisa Econômica Aplicada (IPEA) is a traditional Brazilian federal public foundation linked to the Ministry of Planning, Development and Management

Focus on support formulation and reformulation of public policies and Brazilian development programs

Traditional areas: Economics, stats, sociology, public policy research

New areas: Information technology, big data management, computer science

Civil society organizations

Civil society organization is the "aggregate of non-governmental organizations and institutions that manifest interests and will of citizens."

Data about these organizations must to be open for society

Brazilian government is investing in open data and new ways to allow research in this data

One of the most expected systems at IPEA since this one will be the baseline to get info about all organizations of the country

Development team

Effort to build this system is divided in two teams located at Brasília and Rio

The project has around 15 members originated from many areas: Sociology, Stats, Computer Science and others

In the systems development team: DBA, designers, devops, developers and testers

Almost all members are graduated

Challenges

Big data

Process and visualize hundreds of thousands of points in the view:

- ~400 millions of registers

Integrate many different databases from other governmental domains

- The system is linked with 6 different big datasets

The ecosystem must be “Data Lake” ready

- The API is only a single independent application that will use a data from a big repository

BI, machine learning and data mining

Real time analytics

- Helping decision makers and researchers through the data

Automatic classification

- Organizations must be labeled and it means high costs to do it

Financial patterns

- Patterns helps to understand the growth of a NGO
- Fraud detection
- Financial linking between organizations

Open data concerns

Code and data must follow the open data plan

- Brazilian government demands that systems follows an open data plan

Needs strong security policies

- A governmental system must be ready to suffer cyber attacks

Data privacy

- Certain data attributes cannot be published or leaked

Continuous deployment

Get code from repository automatically

Automatic tests in front-end and back-end

Each system must be isolated

Build must be automatic

How achieve success using all things above using a simple solution but mature enough to use inside an big institution ?

Technologies

Databases and ETL process

mongoDB®

redis

2016/10/21

Back-end


Lumen


R

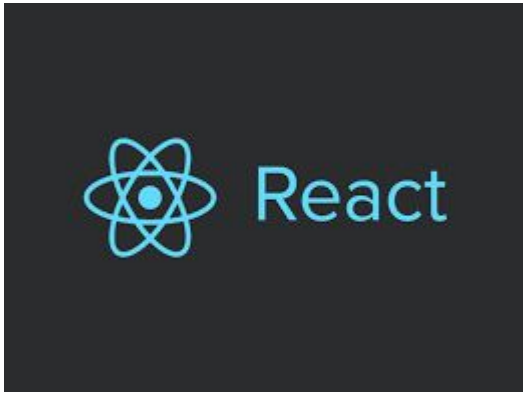

php
V7


laravel


saiku
CUTTING EDGE OPEN SOURCE ANALYTICS


ckan

Front-end



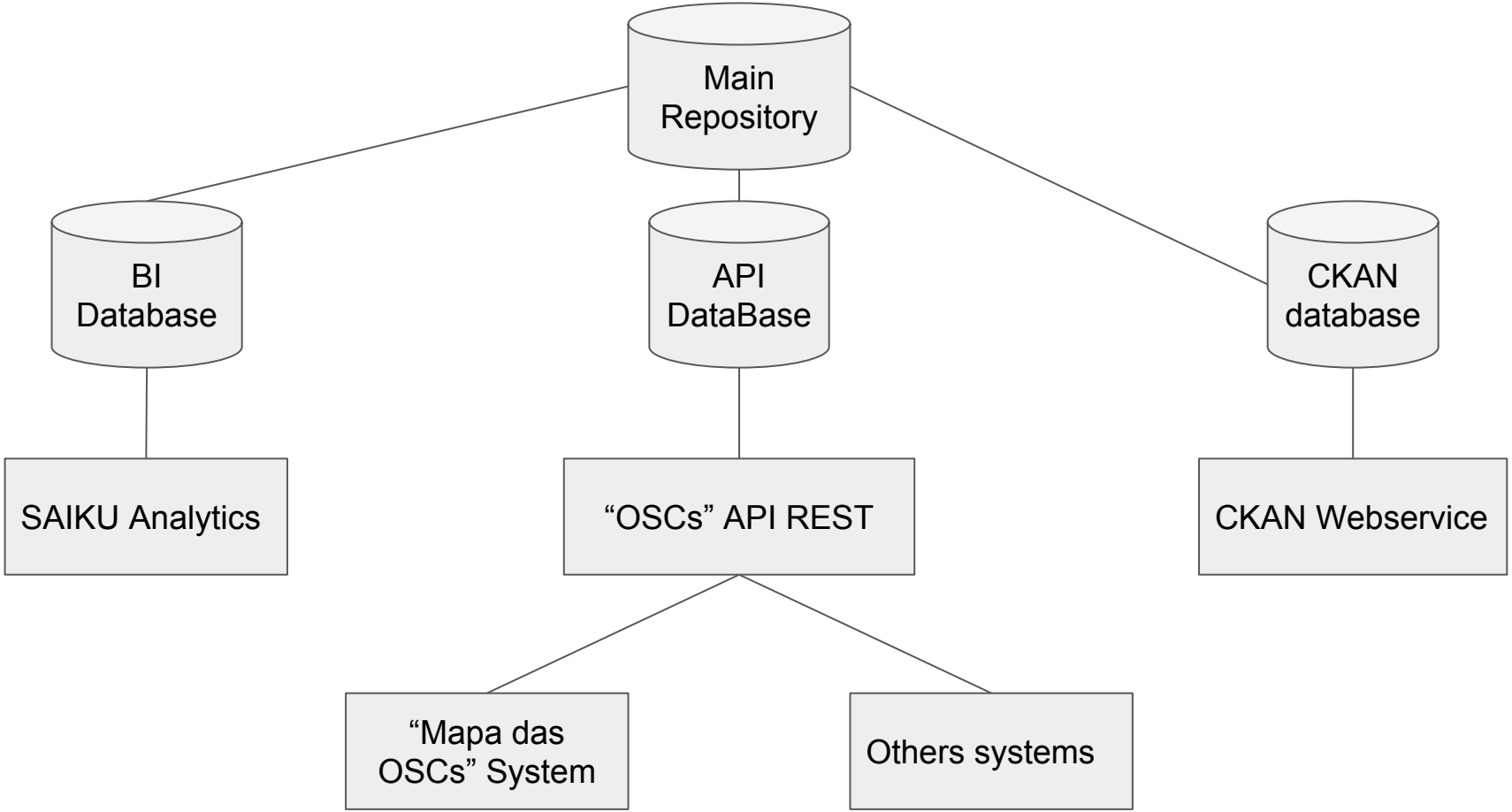
2016/10/21

Tests and automation

Jenkins

Dependency Manager for PHP

Architecture


Main Repository
BI Database
API DataBase
CKAN database
SAIKU Analytics
“OSCs” API REST
CKAN Webservice
“Mapa das OSCs” System
Others systems

2016/10/21

Database architecture

Main repository

- MongoDB
- Repository only for reading
- Raw data

BI database

- Postgres
- OLAP cube for analytics (SAIKU)

Database architecture

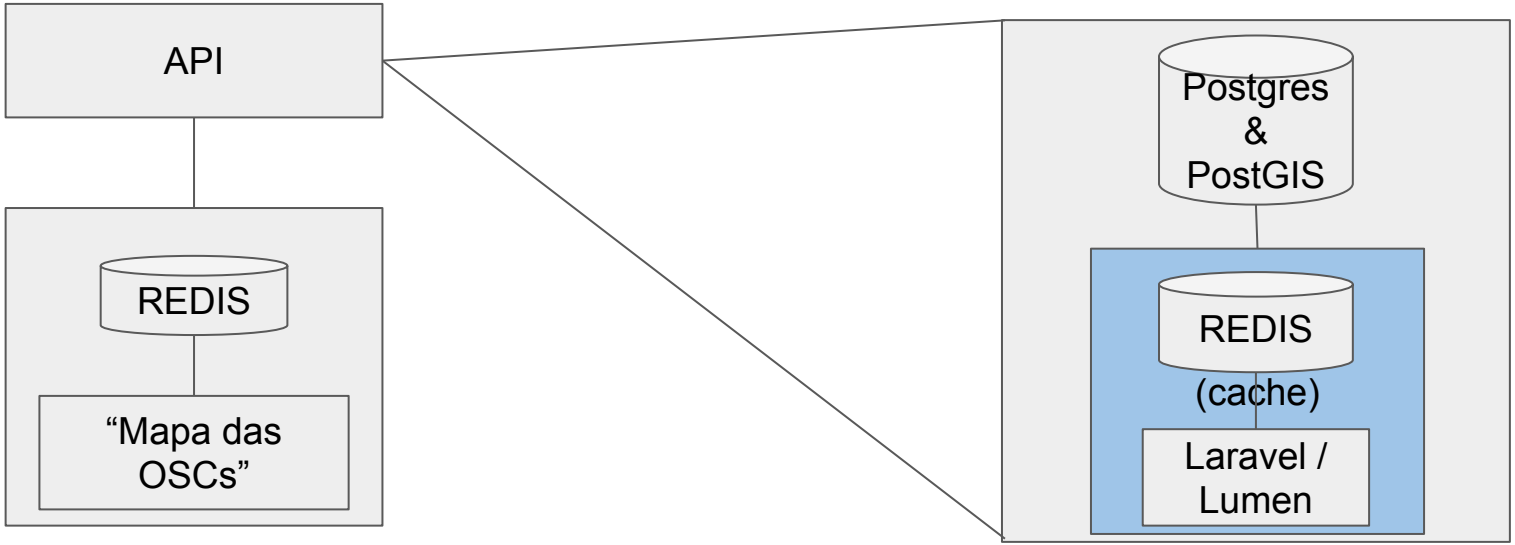
CKAN database

- Postgres
- Feeds the CKAN webservice

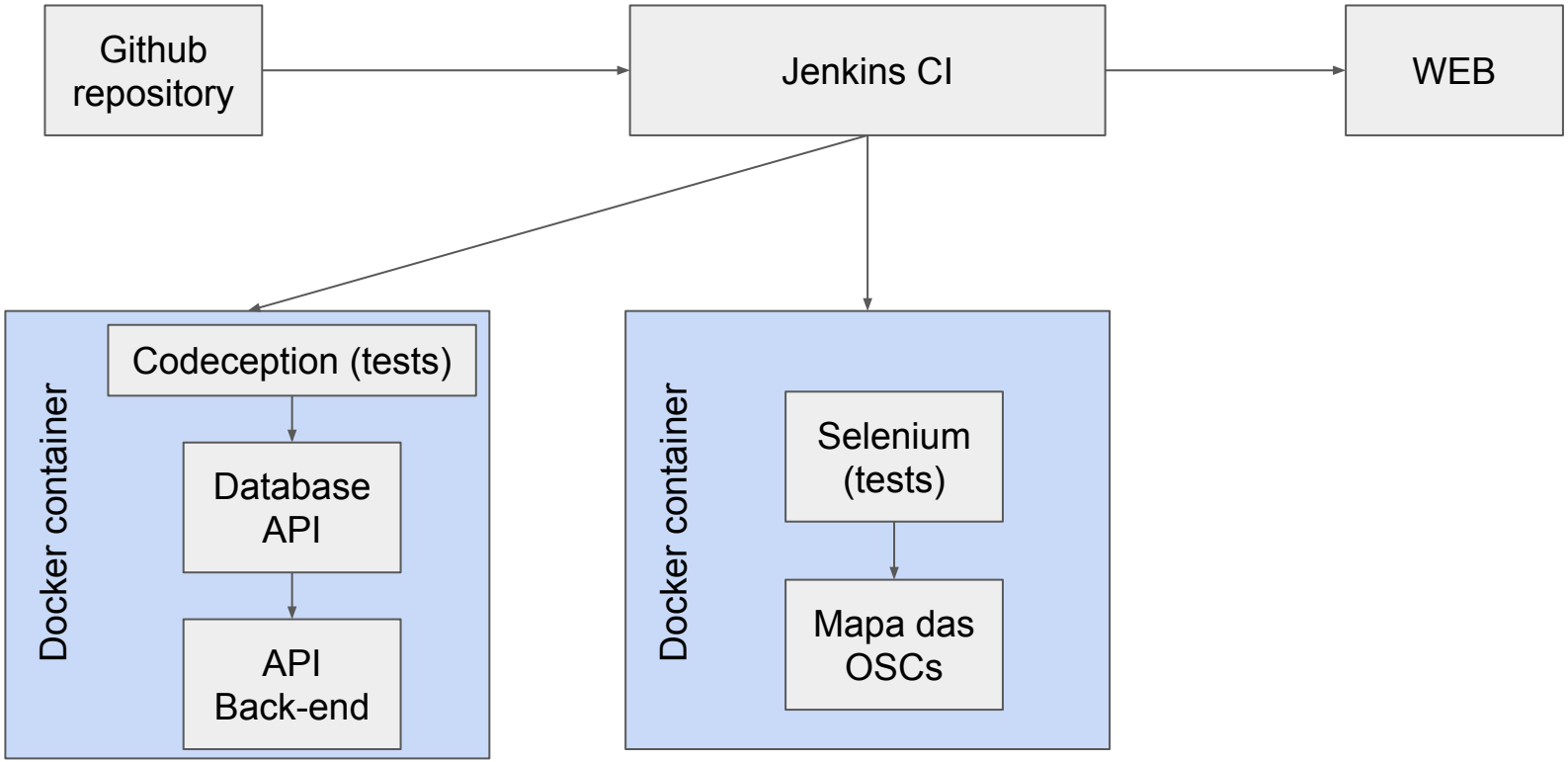
Database API

- Postgres & PostGIS
- Clean database for IPEA applications concern about CSOs

API architecture



Deployment



2016/10/21

Techniques

Database

All SQL inside functions

R-Tree indexing at geometry bounds

Materialized views serving the API

Triggers updating views and cache

Databases are isolated from each other

Back-end

Data sent by chunks:

- Big chunks of data sent by PHP using functions flush(); and sleep();

All queries pass through a middleware before to go the API:

- Search becomes a key and response becomes a value. These info are stored inside REDIS. Only new queries are sent to API.
- API also has his own cache (REDIS). Only new queries are sent to database.

Front-end

Transform almost everything in web components using React.js

```
var DropdownMenu = React.createClass({
  renderList: function(){
    var elems = [];
    for (var i=0; i< this.props.submenu.length; i++) {
      var l = this.props.submenu[i];
      //se titulo vir vazio entende-se que devemos adicionar um separador
      if (l.text=='') elems.push(<li className="divider"></li>);
      else elems.push(<li><a href={l.link}>{l.text}</a></li>);
    }
    return elems;
  },
  render: function () {
    return (<div>
            <a href="#" className="dropdown-toggle" data-toggle="dropdown" role="button" aria-expanded="false">{this.props.titulo}<
            <ul className="dropdown-menu" role="menu">{this.renderList()}</ul>
            </div>);
  }
});
```

Front-end

Require.js for transform components in modules

```
require(['react', 'jsx!components/Util'], function (React) {

  require(['componenteBlocoDeTexto'], function(BlocoTexto){
    function BlocoDeTexto(titulo, formato){
      this.titulo = titulo;
      this.formato = formato;
    }

    var csv = "O formato CSV(Comma Separated Values) é um dos formatos mais utilizados para a troca de dados entre
    var xls = "Formato padrão do Microsoft Excel, o XLS(XmL Spreadsheet) tem uma qualidade razoável e é simples de
    var xml = "O formato XML (eXtensible Markup Language) é popularmente utilizado para a tranferência de dados bin
    var json = "O formato JSON (JavaScript Object Notation) é um formato de transferência de dados que apresenta ca
    var titulos = ["CSV", "XLS", "XML", "JSON"];
    var formatos = [csv, xls, xml, json];

    var blocosDeTexto = [];
    for (var i=0; i<titulos.length; i++){
      blocosDeTexto.push(new BlocoDeTexto(titulos[i], formatos[i]));
    }

    BlocoTexto = React.createFactory(BlocoTexto);
    ReactDOM.render(BlocoTexto({dados:blocosDeTexto}), document.getElementById("bloco_texto_formato_dados"));
  });


  require(['componenteDropdown'], function(Dropdown){
    var arquivosRetornados, arquivosEnviados;
    arquivosRetornados = arquivosEnviados = ["XML", "JSON", "CSV"];
    var periodicidade = ["Dia(s)", "Semana(s)", "Mês(es)"];

    Dropdown = React.createFactory(Dropdown);

    ReactDOM.render(Dropdown({list: arquivosRetornados}), document.getElementById("arquivo_retornado_dropdown"));
    ReactDOM.render(Dropdown({list:periodicidade}), document.getElementById("periodicidade_dropdown"));
    ReactDOM.render(Dropdown({list:arquivosEnviados}), document.getElementById("tipo_arquivo_dropdown"));
   });
});
```

Front-end

Dependencies are managed

with require.js config file

```
require.config({
  baseUrl: "js/",
  paths: {
    "react": "libs/react-15.3.1/react-with-addons.min",
    "jsx": "libs/jsxcompiler/jsx",
    "text": "libs/require-2.3.2/text",
    "JSXTransformer": "libs/jsxcompiler/JSXTransformer",
    "babel": "libs/babel-core/5.8.24/browser.min",
    "jquery": "libs/jquery-3.1.0/jquery-3.1.0.min",
    "jquery-ui": "libs/jquery-ui-1.12.0/jquery-ui",
    "bootstrap":"libs/bootstrap-3.3.7/bootstrap.min",
    "d3": "libs/nv-d3/d3.v3",
    "nv.d3":"libs/nv-d3/nv.d3",
    "nv.d3.lib":"libs/nv-d3/nv.d3.lib",
    "stream":"libs/nv-d3/stream-layers",
    "tablesaw":"libs/tablesaw-3.0/tablesaw",
    "tablesaw-init":"libs/tablesaw-3.0/tablesaw-init",
    "datatables.net": "libs/DataTables/DataTables-1.10.12/js/jquery.dataTables.min",
    "datatables-responsive":"libs/DataTables/Responsive-2.1.0/js/dataTables.responsive.min",
    "leaflet": "libs/leaflet-0.7.7/leaflet",
    "leafletCluster": "libs/leaflet-0.7.7/cluster.min",
    "google": "libs/google",
    "rotas":"rotas"
  },
  shim: {
      'jquery-ui': ['jquery'],
      'bootstrap': ['jquery-ui'],
      'd3': ['bootstrap'],
      'nv.d3': ['d3'],
      'stream':['nv.d3'],
      'nv.d3.lib': ['stream'],
```

Current State

“Mapa das OSCs”

Busque um Perfil

Organização | Município | Estado | Região

vasco

Buscar

Mapa das OSCs

Indicadores

Editais Públicos

Atlas do IDH Municipal

Marco Regulatório das OSCs

Extrator de Dados

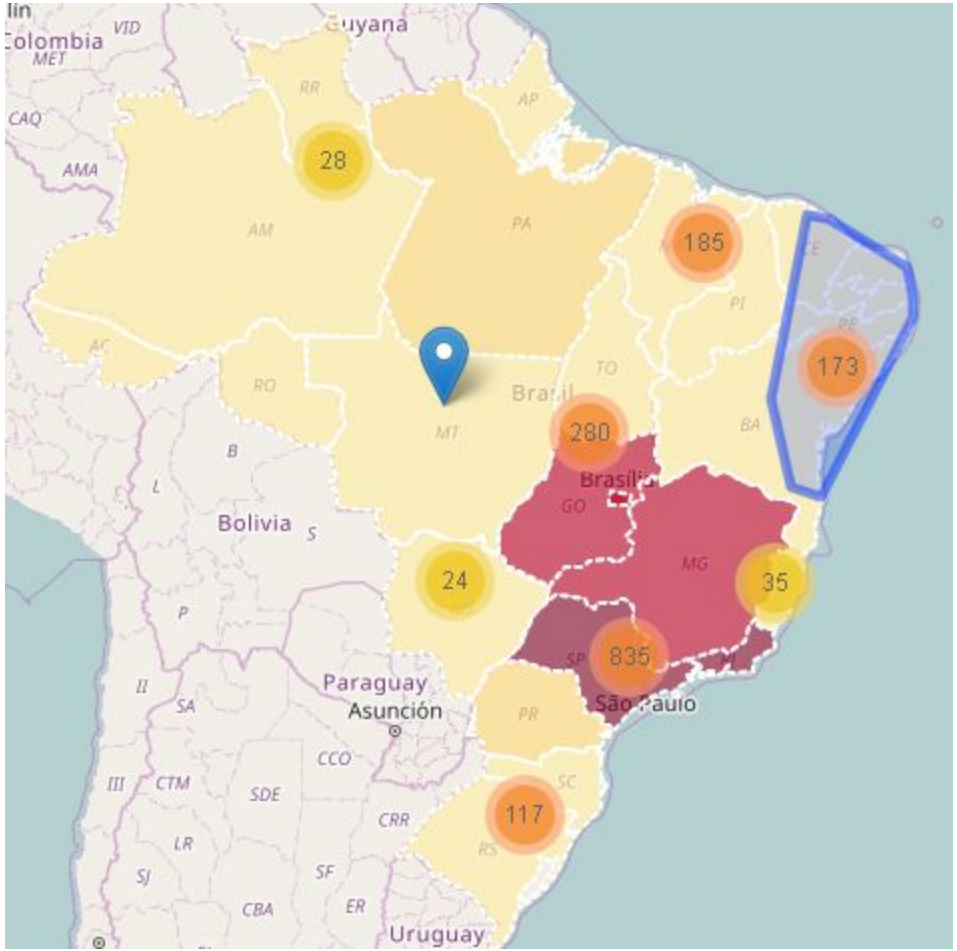
Map

Heatmap

Clustering

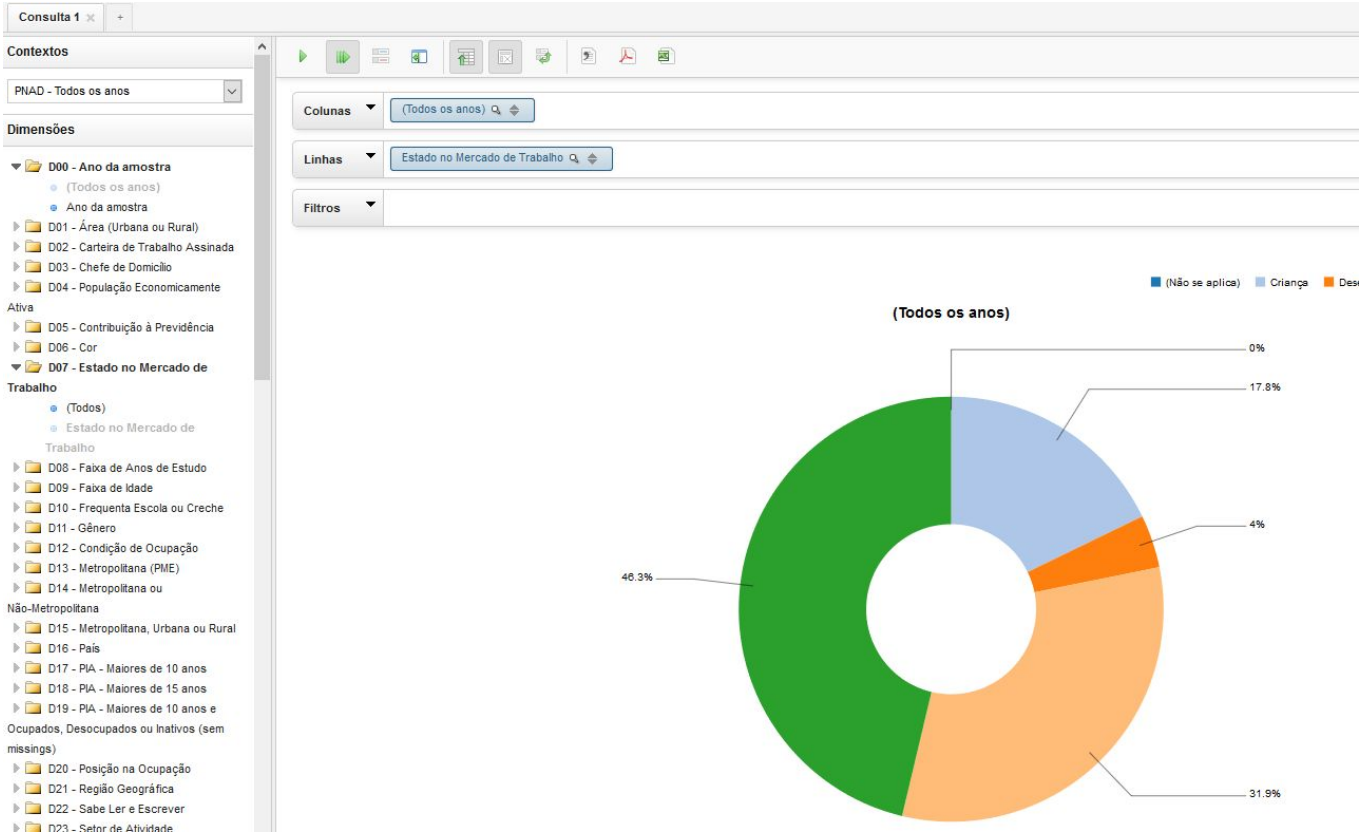
Cache

Data sent by chunks



2016/10/21

BI


Consulta 1
Contextos
PNAD - Todos os anos
Dimensões
D00 - Ano da amostra
(Todos os anos)
Ano da amostra
D01 - Área (Urbana ou Rural)
D02 - Carteira de Trabalho Assinada
D03 - Chefe de Domicílio
D04 - População Economicamente Ativa
D05 - Contribuição à Previdência
D06 - Cor
D07 - Estado no Mercado de Trabalho
(Todos)
Estado no Mercado de Trabalho
D08 - Faixa de Anos de Estudo
D09 - Faixa de Idade
D10 - Frequenta Escola ou Creche
D11 - Gênero
D12 - Condição de Ocupação
D13 - Metropolitana (PME)
D14 - Metropolitana ou Não-Metropolitana
D15 - Metropolitana, Urbana ou Rural
D16 - País
D17 - PIA - Maiores de 10 anos
D18 - PIA - Maiores de 15 anos
D19 - PIA - Maiores de 10 anos e Ocupados, Desocupados ou Inativos (sem missings)
D20 - Posição na Ocupação
D21 - Região Geográfica
D22 - Sabe Ler e Escrever
D23 - Setor de Atividade
Colunas
(Todos os anos)
Linhas
Estado no Mercado de Trabalho
Filtros
(Não se aplica)
Criança
(Todos os anos)
0%
17.8%
4%
46.3%
31.9%

Result list

Resultado da Consulta

Visualizar por Mapa | Visualizar por Lista

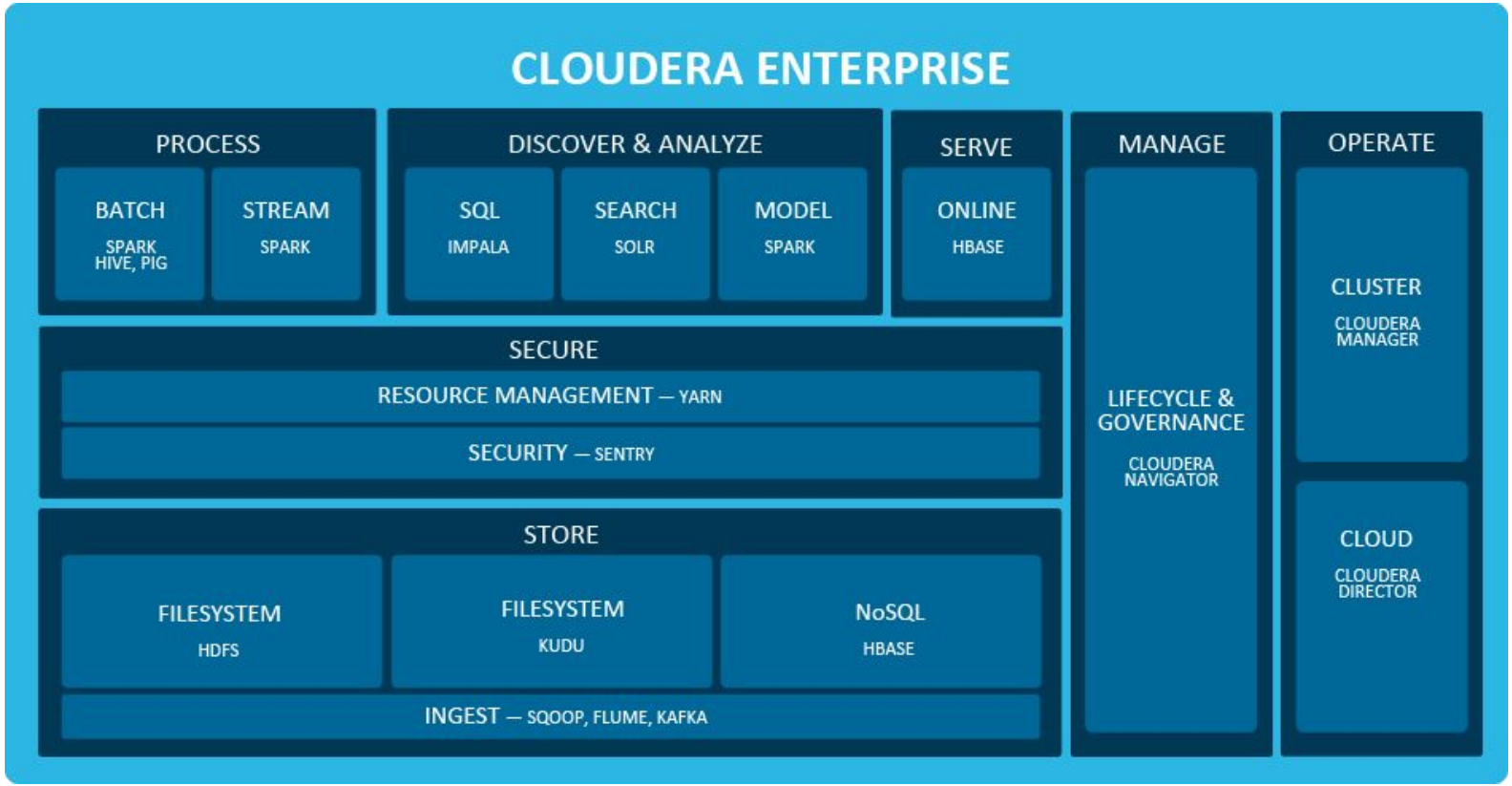
Show 10 entries Search:

	NOME DA OSC	CNPJ	NATUREZA JURÍDICA	ENDEREÇO	DETALHAR
Logotipo da OSC	Vasco Da Gama F C	87558110000151	Associação Privada	AV OSVALDO ARANHA, 1448, CIDADE ALTA, Bento Gonçalves, 95700000	Detalhar
Logotipo da OSC	Gremio Dram E Rectivo Vasco Da Gama F C	46529723000183	Associação Privada	RUA JOSE BERNARDO PINTO, 486, VILA GUILHERME, São Paulo, 2055000	Detalhar

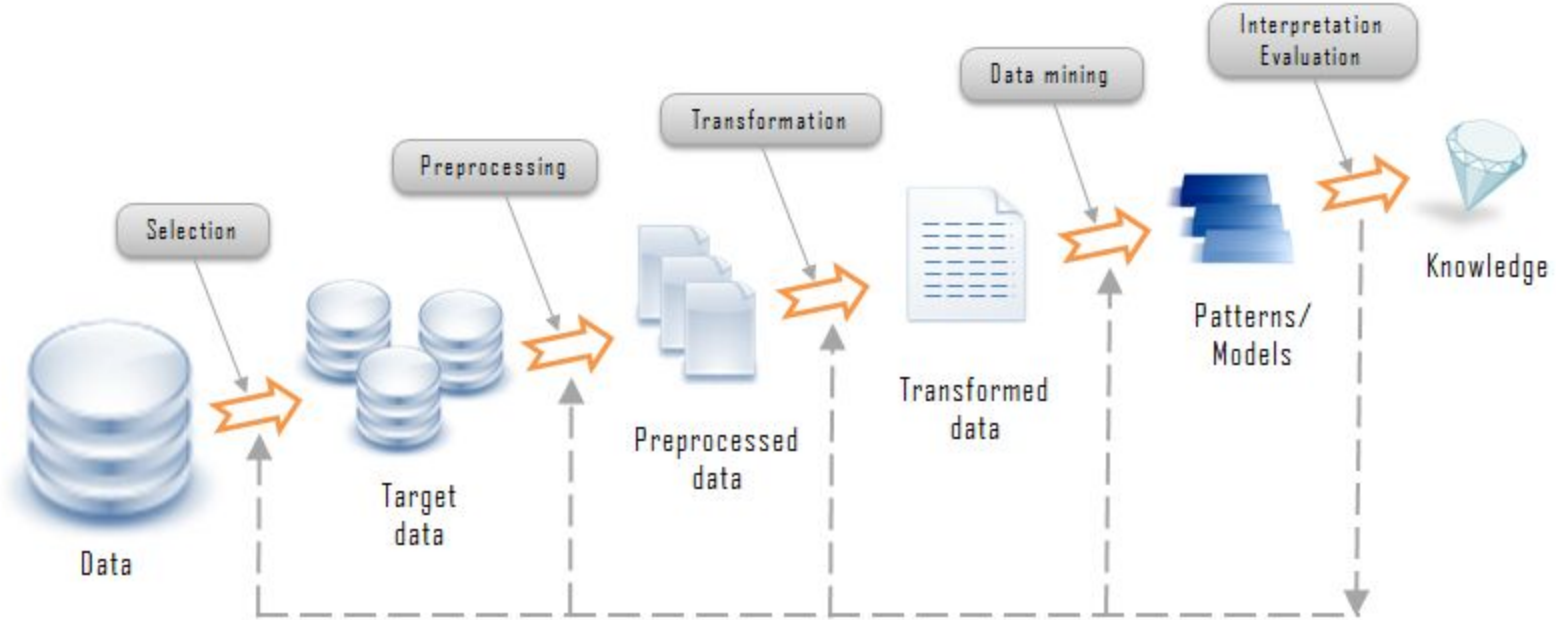
2016/10/21

Future Steps

Big data full stack



Data mining & Knowledge discovery



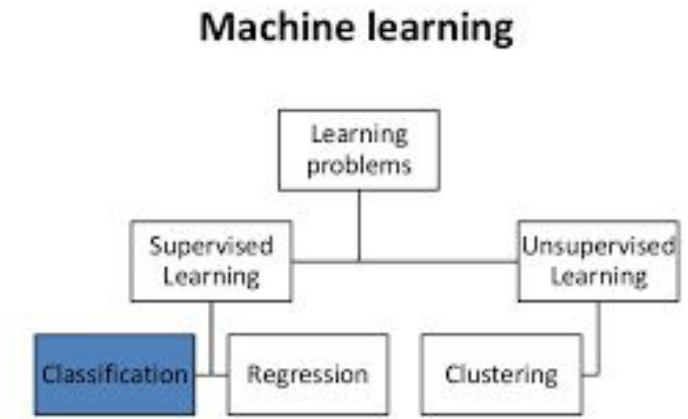
Applied machine learning & Information retrieval

Data classification (labeling)

Text searching and extraction

Time series prediction

Learning on graphs



16/10/21

Get in Touch

www.raulferreira.com.br

https://br.linkedin.com/in/raulsenaferreira

raul.ferreira@ipea.gov.br

raulsf@cos.ufrj.br

Open calls:
http://www.ipea.gov.br/portal/index.php/?option=com_content&view=article&id=23255&Itemid=5


THANK YOU
GRACIAS
ARIGATO
SHUKURIA
BOLZÏN
MERCI
BÏYAN
DANKSCHEEN
JUSPAXAR
TASHAKKUR ATU
SUKSAMA
EKHMET
MEHRBANI
GRAZIE
KOMAPSUMNIDA
GOZAIMASHITA
EFCHARISTO
YAQHANYELAY
SHUKRIA
TINGKI
SPASSIBO
SNACHALHUYA
NUHUN
CHALTU
WABEEJA
MAITEKA
YUSPAGARÌTAM
HUI
DHANYABAAD
AINHA
ATTO
MERSI
SPASIBO
DENKAUJA
NENACHALHIYA
UNALCHÉESH
GUI
HATUR
EKÒJU
SIKOMO
MAKETAU
MINMONCHAR
MAAKE
SANCO
MERASTAWHY
GAEJTHO
LAH
PALDIES
TAVTAPUCH
MEDAWAGSE
AGUYJE
FAKAAUE
BAIKA