

CURSO SOBRE MANUSEIO DE BASES DE DADOS DO GOVERNO FEDERAL NO SOFTWARE “R”

Introdução à base de dados

Professores:

Marcelo Prudente

Rafael Giacomini

Conceitos

Passado



Presente



Conceitos

Dados



- são um conjunto de valores ou ocorrências em um estado bruto com o qual são obtidas informações com o objetivo de adquirir benefícios¹.
- Existem dois tipos de dados: estruturados e não estruturados.
 - Os dados estruturados, que são dados formatados, organizados em tabelas - linhas e colunas - e são facilmente processados, geralmente é utilizado um sistema gerenciador de banco de dados para armazenar esse tipo de dado, um exemplo são os dados gerados por aplicações empresariais.
 - Os dados não estruturados não possuem uma formatação específica e são mais difíceis de serem processados. Por exemplo, mensagens de email, imagens, documentos de texto, mensagens em redes sociais.

¹SHRIVASTAVA; SOMASUNDARAM (2009). Armazenamento e Gerenciamento de Informações: Como armazenar, gerenciar e proteger informações digitais. São Paulo: Bookman.

Conceitos

Metadados (dados sobre os dados)



- Descrição das componentes semânticas e sintáticas de uma informação, necessárias para sua compreensão e para o seu manuseio computacional.
- Os metadados fornecem informações sobre os dados e sobre os processos de produção e uso dos dados.

Conceitos

Banco/base de dados



- É uma coleção de dados inter-relacionados, representando informações sobre um domínio específico”, ou seja, sempre que for possível agrupar informações que se relacionam e tratam de um mesmo assunto, posso dizer que tenho um banco de dados¹.
- São um conjunto de arquivos relacionados entre si com registros sobre pessoas, lugares ou coisas. São coleções organizadas de dados que se relacionam de forma a criar algum sentido (Informação) e dar mais eficiência durante uma pesquisa ou estudo².

¹KORTH, H.F. e SILBERSCHATZ, A.; Sistemas de Bancos de Dados, Makron Books, 2a. edição revisada, 1994.

²https://pt.wikipedia.org/wiki/Banco_de_dados

Conceitos

Banco/base de dados - Propriedades

- Uma base de dados é uma coleção de dados logicamente relacionados, com algum significado. Associações aleatórias de dados não podem ser chamadas de bases de dados;
- Uma base de dados é projetada, construída e preenchida com dados para um propósito específico. Ela tem um grupo de usuários e algumas aplicações pré-concebidas para atendê-los;
- Uma base de dados representa algum aspecto do mundo real, algumas vezes chamado de “mini-mundo”. Mudanças no mini-mundo provocam mudanças na base de dados.

Conceitos

Tabelas

- é um conjunto de dados dispostos em número infinito de colunas e número ilimitado de linhas (ou tuplas)¹.
 - As colunas são tipicamente consideradas os campos da tabela, e caracterizam os tipos de dados que deverão constar na tabela (numéricos, alfanuméricos, datas, coordenadas, etc).
 - O número de linhas pode ser interpretado como o número de combinações de valores dos campos da tabela, e pode conter linhas idênticas, dependendo do objetivo. A forma de referenciar inequivocamente uma única linha é através da utilização de uma chave primária.

¹[https://pt.wikipedia.org/wiki/Tabela_\(banco_de_dados\)](https://pt.wikipedia.org/wiki/Tabela_(banco_de_dados))

Conceitos

Tabelas

• Exemplo

Colunas

Campo/Variáveis

- Atributos
- Características

Formatos

- Sequência de caracteres
- Numérico
- Data

Linhas

Registro

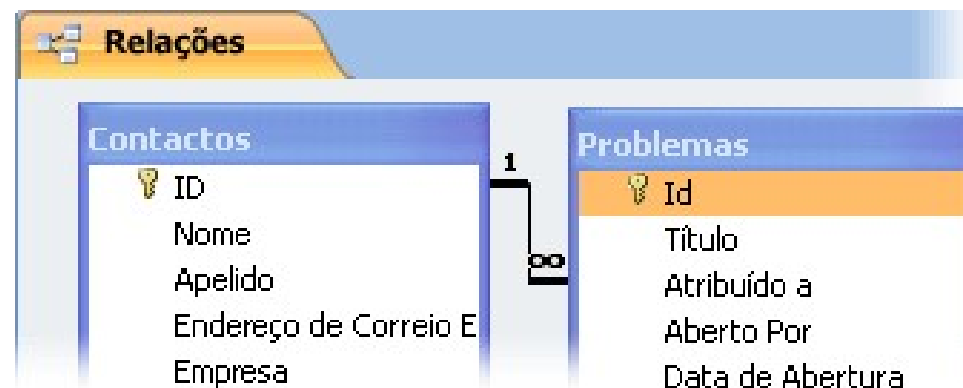
- Pessoa
- Domicílio
- Empresa
- Propriedade
- Turma
- ...

	V0101	UF	V0102	V0103	V0301	V0302	V3031	V3032	V3033	V8005	V0401	V0402	V0403
1	2015	11	11000015	1	1	2	27	2	1992	23	1	1	1
2	2015	11	11000015	3	1	4	4	5	1992	23	1	1	1
3	2015	11	11000015	4	1	4	4	1	1980	35	1	1	1
4	2015	11	11000015	4	2	2	5	6	1981	34	2	2	1
5	2015	11	11000015	4	3	4	8	4	2004	11	3	3	1
6	2015	11	11000015	4	4	4	1	12	2007	7	3	3	1
7	2015	11	11000015	4	5	4	6	12	2010	4	3	3	1
8	2015	11	11000015	4	6	4	8	9	1997	18	5	5	1
9	2015	11	11000015	5	1	2	3	1	1969	46	1	1	1
10	2015	11	11000015	5	2	4	1	1	1934	81	4	4	1
11	2015	11	11000015	6	1	4	3	6	1944	71	1	1	1
12	2015	11	11000015	7	1	4	8	12	1967	47	1	1	1
13	2015	11	11000015	7	2	2	28	3	1991	24	3	3	1
14	2015	11	11000015	8	1	4	23	3	1987	28	1	1	1
15	2015	11	11000015	8	2	2	25	6	1965	50	2	2	1
16	2015	11	11000015	8	3	2	17	3	2014	1	3	3	1
17	2015	11	11000015	9	1	4	22	6	1981	34	1	1	1
18	2015	11	11000015	9	2	2	25	9	1980	35	2	2	1
19	2015	11	11000015	9	3	2	27	6	1999	16	3	3	1
20	2015	11	11000015	9	4	4	27	12	2003	11	3	3	1

Conceitos

Chave

- As tabelas relacionam-se umas as outras através de chaves.
- Uma chave é um conjunto de um ou mais atributos que determinam a unicidade de cada registro.



¹https://pt.wikipedia.org/wiki/Banco_de_dados_relacional

Conceitos

Banco de dados relacional

- é um banco de dados que modela os dados de uma forma que eles sejam percebidos pelo usuário como tabelas, ou mais formalmente relações¹.
- A arquitetura de um banco de dados relacional pode ser descrita de maneira informal ou formal.
 - Na descrição informal estamos preocupados com aspectos práticos da utilização e usamos os termos tabela, linha e coluna.
 - Na descrição formal estamos preocupados com a semântica formal do modelo e usamos termos como relação (tabela), tupla (linhas) e atributo (coluna).

¹https://pt.wikipedia.org/wiki/Banco_de_dados_relacional

Conceitos

Banco de dados relacional

- Exemplo - PNAD

Dicionário dos dados

Posição Inicial	Tamanho	Código de variável	Quesito		Categorias	
		Nº	Descrição	Tipo	Descrição	
PESQUISA BÁSICA						
PARTE 1 – IDENTIFICAÇÃO E CONTROLE						
1	4	V0101		Ano de referência		
5	2	UF	2	Unidade da Federação	11	Rondônia
					12	Acre
					13	Amazonas
					14	Roraima
					15	Pará
					16	Amapá
					17	Tocantins
					21	Maranhão
					22	Piauí
					23	Ceará
					24	Rio Grande do Norte
					25	Paraíba
					26	Pernambuco
					27	Alagoas
					28	Sergipe
					29	Bahia
					31	Minas Gerais
					32	Espírito Santo
					33	Rio de Janeiro
					35	São Paulo
					41	Paraná
					42	Santa Catarina
					43	Rio Grande do Sul
					50	Mato Grosso do Sul
					51	Mato Grosso
					52	Goiás
					53	Distrito Federal
5	8	V0102	2	Número de controle	As 2 primeiras posições são o código da Unidade da Federação	
13	3	V0103	3	Número de série		

Tabela 1 - Domicílio

	V0101	UF	V0102	V0103	V0104	V0105	V0106	V0201	V0202	V0203	V0204
1	2015	11	11000015	1	1	1	1	1	4	1	1
2	2015	11	11000015	2	6	-	-	-	-	-	-
3	2015	11	11000015	3	1	1	1	1	4	1	1
4	2015	11	11000015	4	1	6	4	1	2	1	1
5	2015	11	11000015	5	1	2	2	1	2	1	1
6	2015	11	11000015	6	1	1	1	1	2	1	1
7	2015	11	11000015	7	1	2	2	1	2	1	1
8	2015	11	11000015	8	1	3	2	1	2	2	1
9	2015	11	11000015	9	1	5	4	1	2	2	1
10	2015	11	11000015	10	6	-	-	-	-	-	-
11	2015	11	11000015	11	1	3	3	1	2	1	1
12	2015	11	11000015	12	1	1	1	1	2	2	1
13	2015	11	11000015	13	5	-	-	-	-	-	-
14	2015	11	11000015	14	1	2	2	1	4	1	1
15	2015	11	11000015	15	6	-	-	-	-	-	-
16	2015	11	11000015	16	1	3	2	1	4	1	1
17	2015	11	11000023	1	1	1	1	1	2	1	1
18	2015	11	11000023	2	1	4	3	1	2	1	1
19	2015	11	11000023	3	1	5	5	1	2	1	1
20	2015	11	11000023	4	1	2	2	1	2	1	1

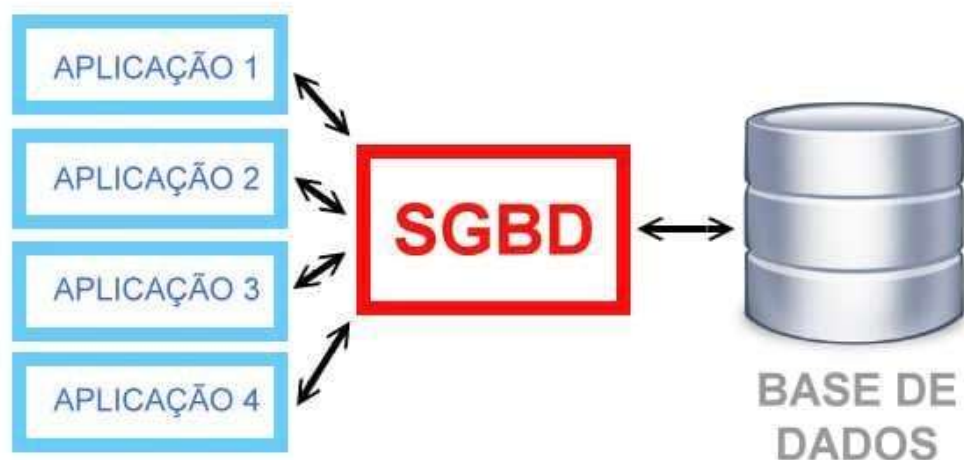
Tabela 2 - Pessoa

	V0101	UF	V0102	V0103	V0301	V0302	V0303	V0302	V0303	V8005	V0401	V0402	V0403
1	2015	11	11000015	1	1	2	27	2	1992	23	1	1	1
2	2015	11	11000015	3	1	4	4	5	1992	23	1	1	1
3	2015	11	11000015	4	1	4	4	1	1980	35	1	1	1
4	2015	11	11000015	4	2	2	5	6	1981	34	2	2	1
5	2015	11	11000015	4	3	4	8	4	2004	11	3	3	1
6	2015	11	11000015	4	4	4	1	12	2007	7	3	3	1
7	2015	11	11000015	4	5	4	6	12	2010	4	3	3	1
8	2015	11	11000015	4	6	4	8	9	1997	18	5	5	1
9	2015	11	11000015	5	1	2	3	1	1969	46	1	1	1
10	2015	11	11000015	5	2	4	1	1	1934	81	4	4	1
11	2015	11	11000015	6	1	4	3	6	1944	71	1	1	1
12	2015	11	11000015	7	1	4	8	12	1967	47	1	1	1
13	2015	11	11000015	7	2	2	28	3	1991	24	3	3	1
14	2015	11	11000015	8	1	4	23	3	1987	28	1	1	1
15	2015	11	11000015	8	2	2	25	6	1965	50	2	2	1
16	2015	11	11000015	8	3	2	17	3	2014	1	3	3	1
17	2015	11	11000015	9	1	4	22	6	1981	34	1	1	1
18	2015	11	11000015	9	2	2	25	9	1980	35	2	2	1
19	2015	11	11000015	9	3	2	27	6	1999	16	3	3	1
20	2015	11	11000015	9	4	4	27	12	2003	11	3	3	1

Conceitos

Sistema de Gerenciamento de Banco de Dados (SGBD)

- Conjunto de programas e ferramentas utilizadas para configurar, atualizar e manter um banco de dados.



Exemplos



Tipos de bases de dados

➤ Cadastros e registros administrativos

- São dados individuais (sobre pessoas, empresas, transações comerciais, etc.) produzidos por instituições (como governo e empresas) com vistas ao agir administrativo.

Exemplo: SIGEPE, CADÚNICO, RAIS

- Características comumente encontradas para análise de dados:
 - Informações sigilosas (acesso restrito)
 - Problemas como duplicação, ausência de informação completa, ausência de documentação
 - Disponibilidade em tempo real

Tipos de bases de dados

➤ Censos

- Tipo de levantamento que obtém informações de todas as pessoas de um grupo.

Exemplo: Censo Demográfico, Censo Escolar, Censo agropecuário

- Características comumente encontradas para análise de dados:
 - As informações obtidas em Censos e/ou pesquisas não amostrais, para estarem conformes com a legislação, devem ser **desidentificadas** e tratadas em áreas suficientemente grandes para não permitirem a revelação do informante. Pesquisas econômicas apresentam elevado número de informações únicas e/ou representativas que inviabilizam a divulgação de microdados;
 - Bases muito grandes
 - Exatidão das respostas
 - Poucas variáveis

Tipos de bases de dados

➤ Pesquisas amostrais

- Levantamento que escolhe aleatoriamente algumas pessoas da população, as quais representam as respostas de um todo da população através de inferências estatísticas.

Exemplo: PNAD, Pesquisa Nacional de Saúde – PNS, Pesquisas Eleitorais

- Características comumente encontradas para análise de dados:
 - Em pesquisas amostrais é viável a disponibilização de arquivos de microdados para uso público sem comprometer o sigilo da informação, suprimindo-se as informações geográficas menores do que a das áreas de ponderação (utilizadas para expansão dos dados)
 - Dados já trabalhados (imputação)
 - Farta documentação

INTEROPERABILIDADE DE BASES DE DADOS

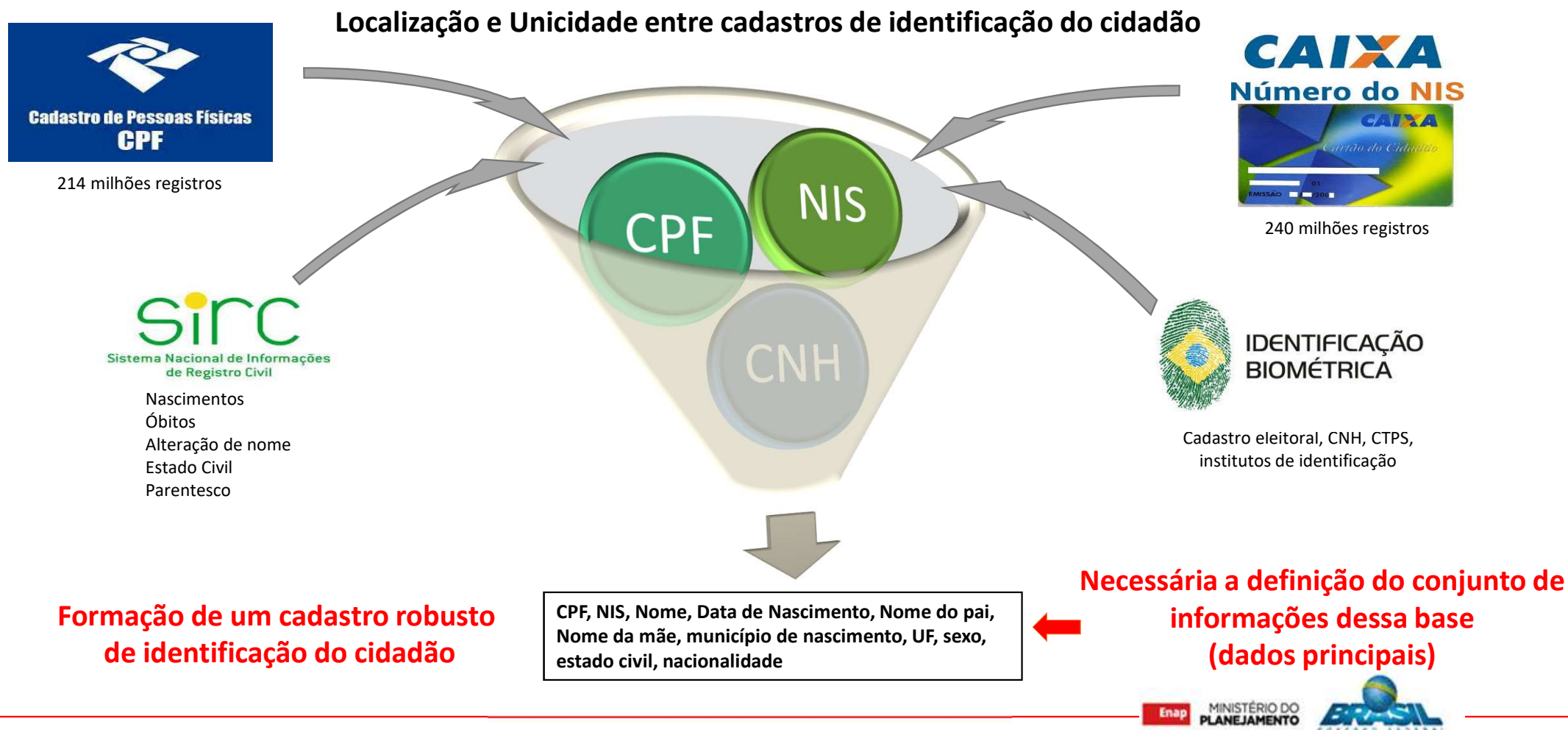
Sistema integrado de informações sobre a identificação e renda das famílias beneficiárias para fins de concessão, avaliação e manutenção dos programas

Modelo de Longo Prazo - Etapas

- 1 - Formação de um Sistema de Interligação de Dados Cadastrais que inclua todos os cidadãos, em que seja garantido o máximo de confiabilidade em termos de unicidade;
- 2 - Utilização deste Sistema para qualificar os demais cadastros e registros administrativos;
- 3 – Interligação entre os demais cadastros e registros administrativos;
- 4 – Viabilização de conexão online entre todas as bases de dados (interoperabilidade de dados) e utilização de soluções cognitivas para aperfeiçoar os programas federais e dar suporte às decisões governamentais;
- 5 – Disponibilizar para pesquisa (academia) uma base desidentificada com diversas variáveis socioeconômicas obtidas dos registros administrativos interligados.

Sistema de Interligação de Dados Cadastrais – Modelo de Longo Prazo

Bases de dados com informação de identificação do cidadão



Cadastro Nacional de Identificação do Cidadão - CNIC

Biometria/foto



Cadastro de identificação

CPF, NIS, Nome, Data de Nascimento, Nome do pai, Nome da mãe, município de nascimento, UF, sexo, estado civil, nacionalidade

Dados principais (lista exemplificativa)

Cadastro de endereços

Instituição de um cadastro de endereços padronizado

Código familiar/domiciliar

Endereço

Dados complementares

Cada cidadão fica vinculado a um endereço do cadastro de endereços

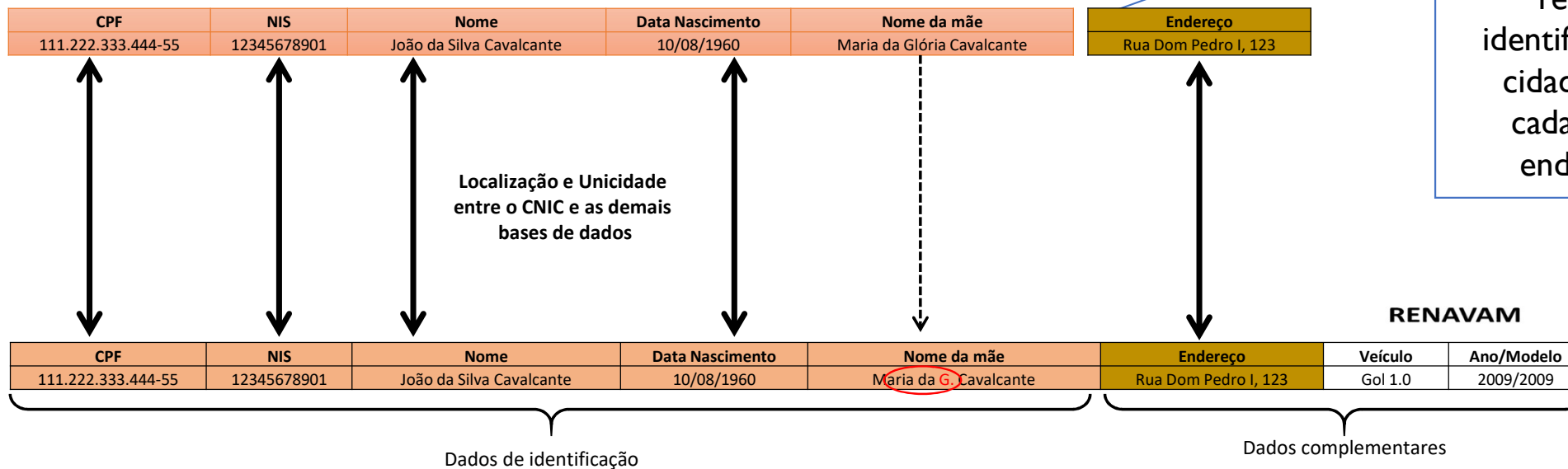
➡ Endereço pelo qual se relaciona com o Estado

Possibilidade de atualização do endereço por meio de um portal do cidadão na Internet

Sistema de Interligação de Dados Cadastrais – Modelo de Longo Prazo

- ✓ Qualificação dos demais registros administrativos por meio de confrontação com o Cadastro Nacional e vinculação dos dados de identificação do cidadão

Exemplo: Cadastro Nacional de Identificação do Cidadão - CNIC

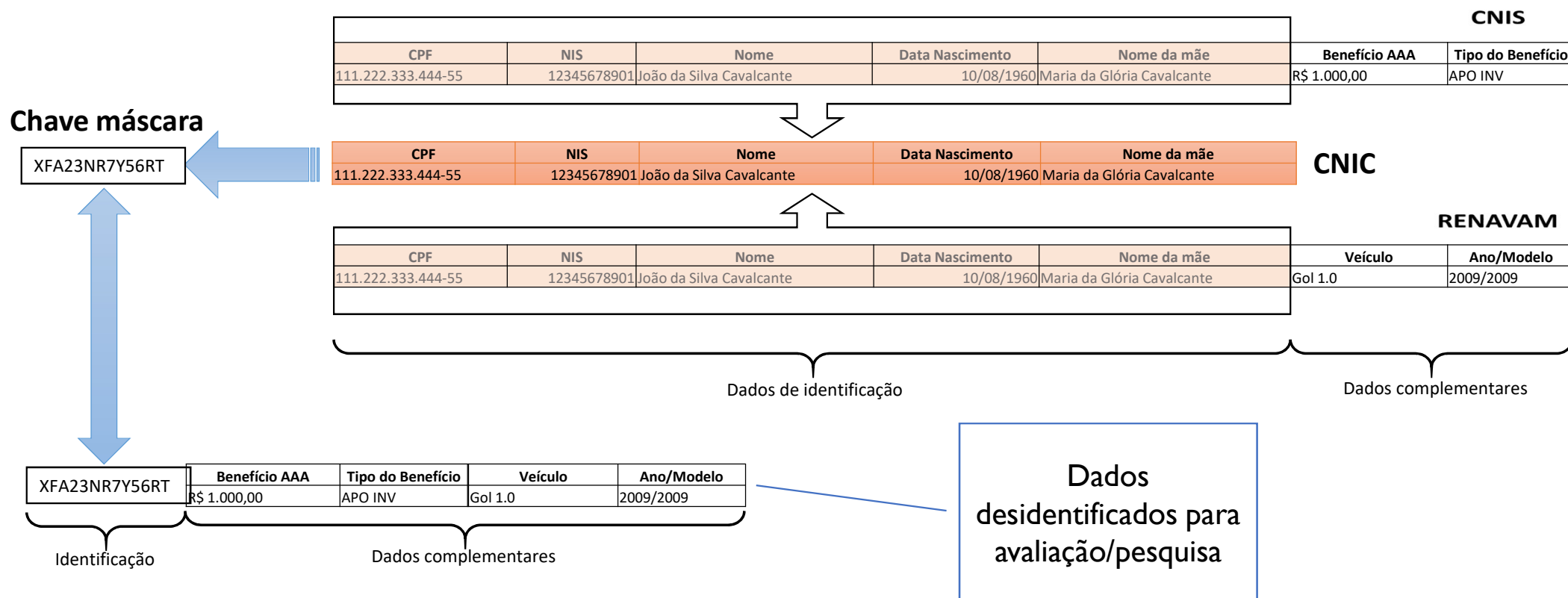


Demais cadastros se “alimentam” do CNIC no que se refere a identificação do cidadão e do cadastro de endereços

Os dados de identificação são vinculados a uma chave primária em comum, que pode ser o CPF e/ou o NIS. Assim, na prática, o CNIC seria o responsável pelas variáveis de identificação do cidadão em todas as bases de dados a ele vinculadas.

Sistema de Interligação de Dados Cadastrais – Modelo de Longo Prazo

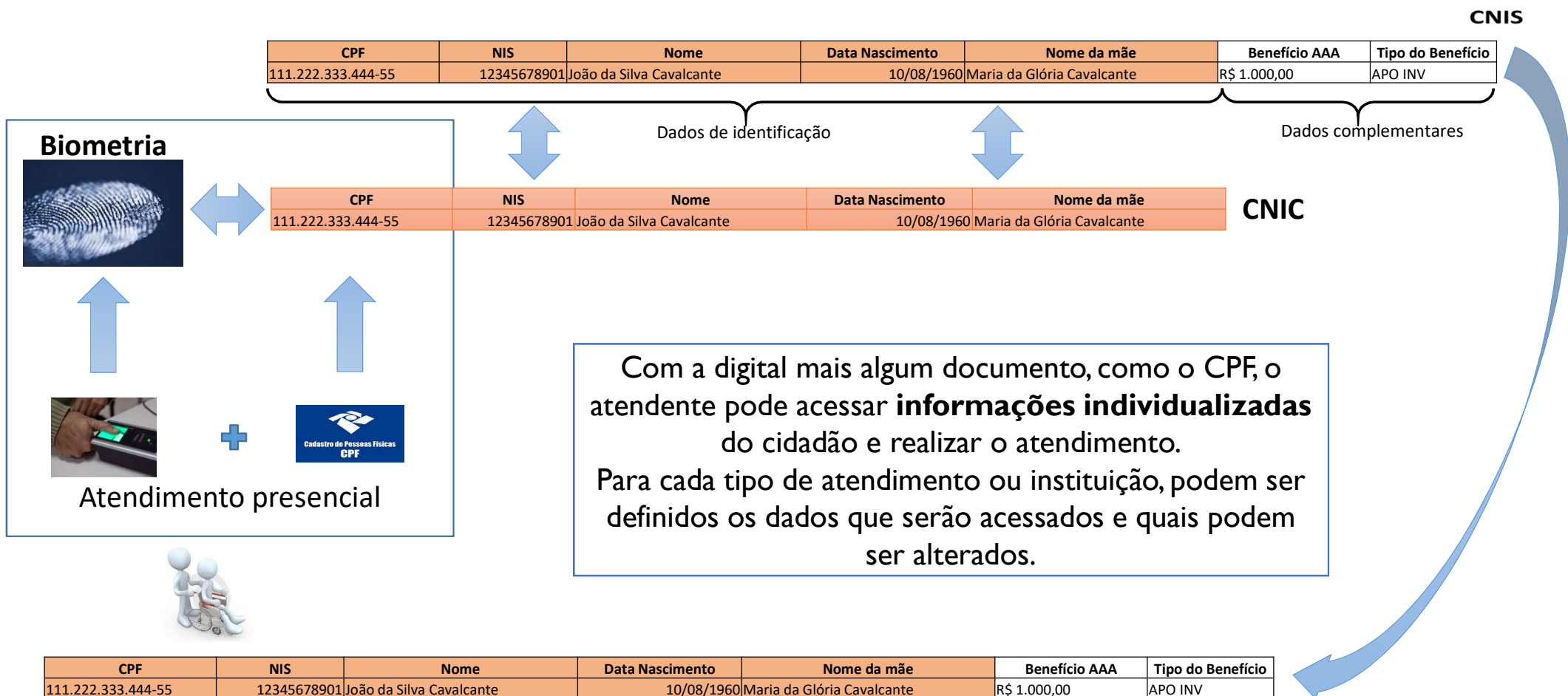
✓ Extração de dados para avaliação/pesquisa



Avaliar possibilidade de acesso a dados protegidos por sigilo bancário/fiscal dessa forma. Com as bases de dados já interligadas, o usuário não precisa ter acesso aos dados de identificação do cidadão.

Sistema de Interligação de Dados Cadastrais – Modelo de Longo Prazo

✓ Fluxo dos dados em atendimento presencial



CPF	NIS	Nome	Data Nascimento	Nome da mãe	Benefício AAA	Tipo do Benefício
111.222.333.444-55	12345678901	João da Silva Cavalcante	10/08/1960	Maria da Glória Cavalcante	R\$ 1.000,00	APO INV

Avaliação Contínua de Programas Sociais



Definição de regra do negócio e avaliação de programas

MINISTÉRIO DO PLANEJAMENTO, DESENVOLVIMENTO E GESTÃO



Verificação dos rendimentos (trabalho e benefícios sociais)

CPF
NIS
SIRC
SIM
SIAB
CNH

Cadastros dos Programas Sociais

Renda declarada

Interligação dos Dados Cadastrais



Identificação do cidadão
Composição familiar

Verificação de renda

Preditor de renda

Renda Verificada

Preditor de renda

Destino

Baixa

Baixa

Permanece

Baixa

Elevada

Prioridade na averiguação

Elevada

Baixa

Bloqueio – averiguação

Elevada

Elevada

Cancelamento

SIAPÉ →

CNIS
Benefícios previdenciários,
RAIS, GPS, GFIP

Avaliação de programas

BIG DATA

Relacionamentos
Endereço

IR
Renavam
CNPJ
Plano de saúde
CNIR
Financiamento
imobiliário

Automatização da verificação das condicionantes para acesso às políticas públicas



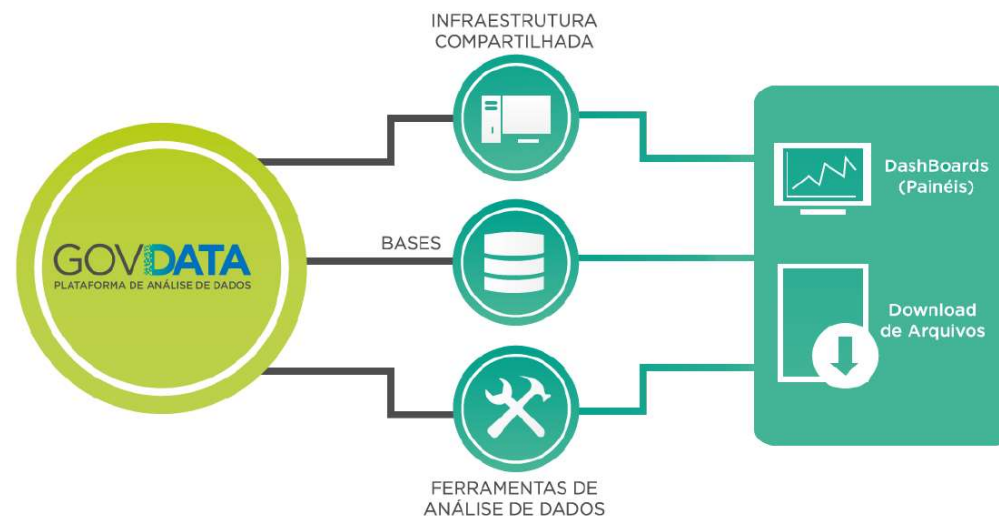
MINISTÉRIO DO PLANEJAMENTO



Desafio - Composição Familiar

- A maioria dos programas sociais utilizam a **renda domiciliar per capita** como um dos critérios de seleção de beneficiários.
- Isso exige a identificação e a verificação da renda de todos integrantes do domicílio.
- Esse processo é um dos maiores desafios do modelo de interoperabilidade de dados.
- O CadÚnico é a principal fonte de informação de composição familiar, mas a informação é declaratória e de difícil verificação.
- Estudo do IPEA aponta estruturas familiares diferentes entre o CadÚnico e a PNAD, sugerindo uma subdeclaração de integrantes.
- A utilização do **cadastro de endereços** (quando padronizado) pode ser uma alternativa.

O que é?



- É a plataforma de análise de dados do Governo que permite aos órgãos do SISP* acesso a diversas bases de dados, para a geração de informações estratégicas com a utilização de ferramentas de descoberta e mineração de dados, e de análises estatísticas e cognitivas.

*Sistema de Administração dos Recursos de Tecnologia da Informação

Benefícios

- Centralização no acesso às principais bases de dados do Governo Federal;
- Tempestividade na entrega e utilização de dados;
- Reuso de dados e de análises;
- Custo de investimento e manutenção da infraestrutura concentrado no Serpro e Dataprev;
- Solução computacionalmente eficiente para o cruzamento de grande quantidade de dados;

Bases de Dados Disponíveis*

- **Siape** - Sistema Integrado de Administração de Recursos Humanos
- **Sigepe** - Sistema de Gestão de Pessoas do Governo Federal
- **Siorg** - Sistema de Informações Organizacionais do Governo Federal
- **Comprasnet** - Sistema de Compras do Governo Federal
- **SCDP** - Sistema de Concessão de Diárias e Passagens
- **CPF** - Cadastro de Pessoa Física
- **CNPJ** - Cadastro Nacional de Pessoa Jurídica
- **CadUnico** - Cadastro Único Social
- **BPC** - Benefício de Prestação Continuada
- **Renavam** - Registro Nacional de Veículos Automotores
- **Renach** - Registro Nacional de Carteira de Habilitação

- **CNIS** - Cadastro Nacional de Informações Sociais, composto pelas 5 bases de dados abaixo:

- Maciça
- GFIP
- Segurado Especial
- Contribuinte Individual
- CNIS Pessoa Física
- **RAIS** - Relação Anual de Informações Sociais
- **Caged** - Cadastro Geral de Empregados e Desempregados
- **Sisobi** - Sistema Informatizado de Controle de Óbitos
- **Siafi** - Sistema Integrado de Administração Financeira do Governo Federal

*O acesso depende da autorização dos detentores

GOV DATA

GOV DATA

PLATAFORMA DE ANÁLISE DE DADOS

LABORATÓRIO
DE BIG DATA

Hue

DESCOBERTA
DE DADOS

Qlik Sense
MicroStrategy
Spotfire


ANÁLISE
ESTATÍSTICA

RStudio

CATÁLOGO
DE DADOS

CKAN

Hue (Hadoop User Experience) é uma interface web, de código aberto, que suporta o Apache Hadoop e seu ecossistema, licenciado sob a licença Apache v2.


Hive Metastore Manager

Captura de Janela
Browser de ficheiros
🔍
📄
🔍
🔍
🔍

📁

🔍

📄

🔍

📁

🔍

📄

🔍

hive

Bases de dados (16) 📄 + ↻

cadunico

caged

cpf

degustacao_010

mp_01

mp_02

mp_03

mp_04

mp_05

mp_06

mp_20

rais

siape

sibe

sisben

tutorial

Bases de dados > mp_05

🗨️ No comment.

👤 hive (USER)

📁 Localização

⚙️ GovData:TipoOferta: 3

GovData:Orgao: MP

GovData:Descricao: Base Própria 0

TABELAS

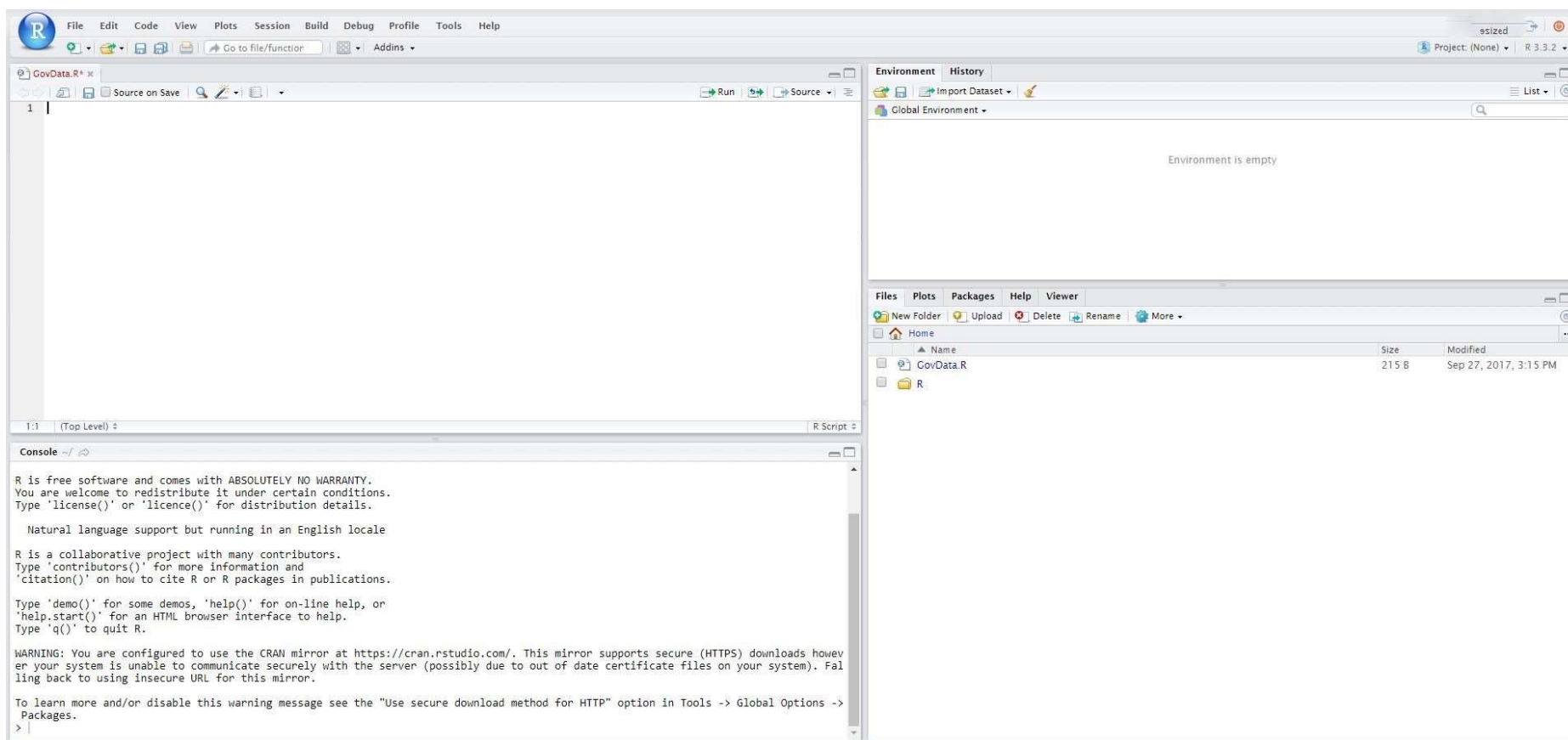
Search for a table...

👁️ View

🔍 Consultar

<input type="checkbox"/>	Nome da tabela	Comment	Type
<input type="checkbox"/>	i bat_siapexbpc_pen		📄
<input type="checkbox"/>	i bat_siapexbpc_pen_1		📄
<input type="checkbox"/>	i bat_siapexbpc_serv		📄
<input type="checkbox"/>	i bat_siapexbpc_serv_1		📄
<input type="checkbox"/>	i bat_siapexbpc_serv_2		📄
<input type="checkbox"/>	i bpc		📄
<input type="checkbox"/>	i bpc_cpf_nok		📄
<input type="checkbox"/>	i bpc_cpf_ok		📄
<input type="checkbox"/>	i bpc_siape_cpf_ok		📄
<input type="checkbox"/>	i bpc_siape_pen_ok		📄
<input type="checkbox"/>	i bpc01		📄
<input type="checkbox"/>	i cpf_bpc_nok_siape		📄
<input type="checkbox"/>	i cpf_bpc_ok_siape_pen_nok		📄
<input type="checkbox"/>	i cpf_sdpa		📄
<input type="checkbox"/>	i cpf_sdpa_nok		📄
<input type="checkbox"/>	i cpf_sdpa_ok		📄

O **RStudio** é uma IDE, de código aberto, para a linguagem de programação de gráficos e cálculos estatísticos R.



GOV DATA



O **Qlik Sense** é uma plataforma para a análise de dados. Com o Qlik Sense você pode analisar os dados e fazer suas próprias descobertas.



A **MicroStrategy** foi concebida para permitir às organizações implementar rapidamente aplicações sofisticadas de análise e segurança em grande escala.



O **Spotfire** é um software analítico para exploração de dados. Permite a descoberta de informações críticas de grande valor estratégico para o seu negócio.

