

# Armazenamento Distribuído de Dados Seguros para Efeito de Sistemas de Identificação Civil

Acadêmico: Matheus Magnusson Bolo

Disciplina: Segurança Computacional

# Contexto

- WGID – IV Workshop de Gestão de Identidades Digitais.

# Contexto - Resumo

- O Registro de Identificação Civil (RIC) é um projeto do Ministério da Justiça concebido para criar uma nova cédula de identidade. Esta cédula deverá ter um chip (tecnologia smart card) para maior segurança e flexibilidade, sendo similar a um cartão bancário. Ela reunirá todos os dados da cédula de identidade atual, bem como CPF e número de título de eleitor, dentre outras informações.

# Contexto - Resumo

- Integração com o sistema de identificação de impressões digitais;
- Integração com todos os bancos de dados de identificação do Brasil.

# Contexto - Objetivos

- Criar um número único de Registro de Identidade Civil – RIC;
- Criar um órgão central coordenado com os órgãos estaduais de identificação, para a emissão, em âmbito nacional, da nova cédula de Identificação Civil com recursos modernos de segurança e de certificação digital.

# Contexto - Fase

- Este projeto foi oficialmente lançado em 2010 e, após uma fase inicial de testes em alguns municípios brasileiros, caminha para uma nova fase em que várias questões verificadas nos testes serão devidamente tratadas.

# Armazenamento Distribuído de Dados Seguros para Efeito de Sistemas de Identificação Civil

- Este trabalho conta com o apoio do Ministério da Justiça, do Ministério do Planejamento, Orçamento e Gestão, da FINEP, da Fundação de Apoio à Pesquisa do Distrito Federal e do Programa Nacional de Pós-Doutorado/CAPES in Brazil.

# Introdução

- Com o intuito de impedir a invasão e a descoberta de dados sigilosos em sistemas de identificação, uma nova abordagem é apresentada para proteger a confidencialidade dos dados.



# Introdução

- A maioria dos ataques relatados é feita diretamente na origem dos dados por pessoas pertencentes à organização ou pela falta de mecanismos que evitam a leitura direta das bases de dados remotamente por terceiros.

# Proposta


- Utilizar um cenário real (NoSQL, Google EBQ e Hadoop) para garantir a confidencialidade e integridade em bancos de dados utilizando modelos criptográficos ajustáveis aos campos das tabelas, permitindo operações de consultas diretamente nos dados cifrados.

# Big Data

- Em tecnologia da informação, o termo **Big Data** refere-se a um grande ou complexo conjunto de dados armazenados – Wikipedia.



# Google BigQuery

- Introduzido em 2011 e disponibilizado ao público em 2012, o Google desenvolveu uma solução de análises de dados que oferece um framework fácil de usar e rapidamente escalável para procurar por grandes quantidades de dados na nuvem dentro de um framework SQL tradicional – 



# Google BigQuery

- “Analisar terabytes de dados com apenas um clique de um botão”. O processo de configuração do BigQuery leva menos de 5 minutos. É só fazer o login no Google APIs Console e, em seguida, criar um novo projeto ou utilizar um existente –
- “Simply move your data into BigQuery and let us handle the hard work.” –



# Google EBQ

- “The Encrypted BigQuery client is an experimental extension of the BigQuery client. It offers client-side encryption for a subset of query types. Currently it is implemented in Python.”



# NoSQL

- “O NoSQL surgiu da necessidade de uma performance superior e de uma alta escalabilidade. Os atuais bancos de dados relacionais são muito restritos a isso, sendo necessária a distribuição vertical de servidores, ou seja, quanto mais dados, mais memória e mais disco um servidor precisa.”



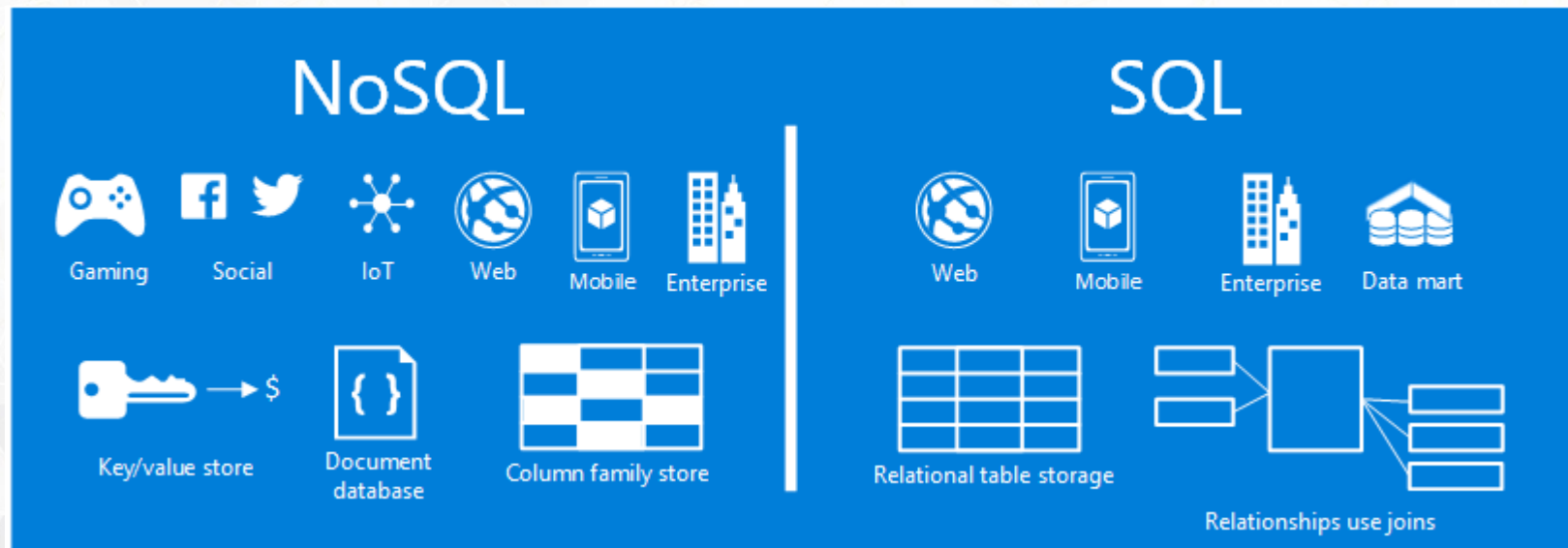
# NoSQL

- Grandes utilizadores deste conceito: Google, Redes Sociais.
- “Atualmente 90% dos sites podem usar sem problemas os bancos de dados tradicionais, para os outros 10% é aconselhável o uso do NoSQL.”





# NoSQL vs SQL

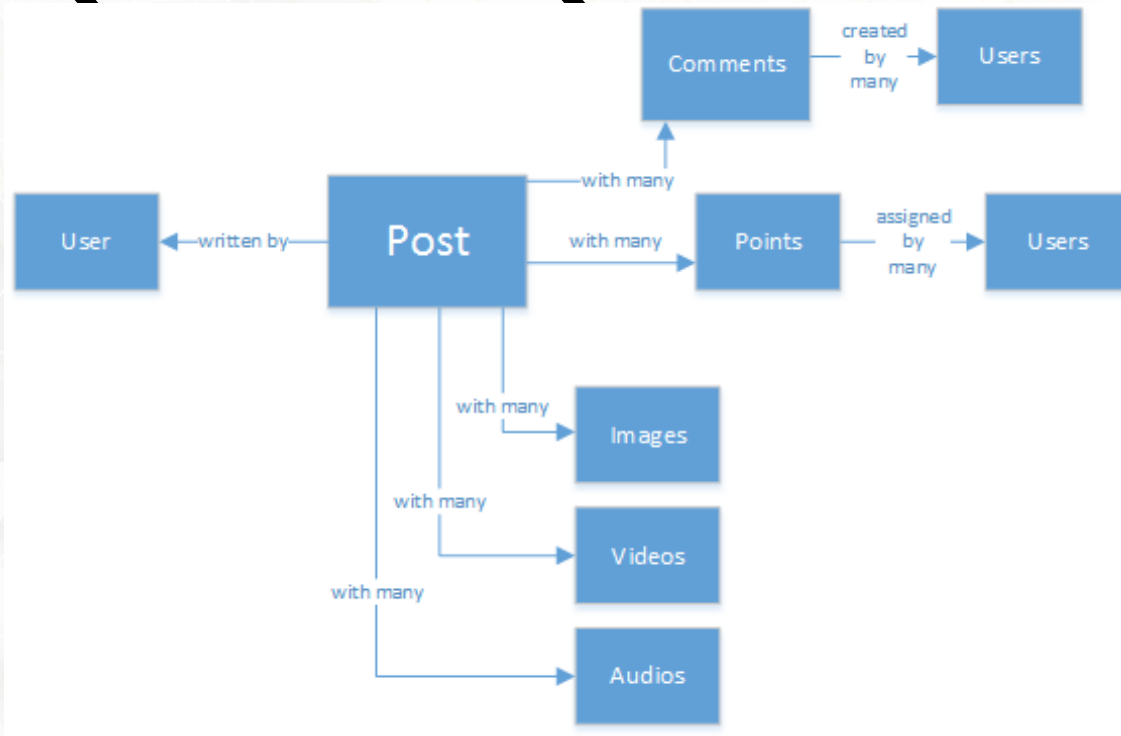


# NoSQL vs SQL

- A seguir, veremos um exemplo caso estejamos criando um novo site de envolvimento social utilizando SQL, onde os usuários podem comentar sobre as postagens e dar pontos (curtidas) para classificá-las.



# NoSQL vs SQL



# NoSQL vs SQL

- Até aqui, tudo bem, mas agora considere a estrutura de uma única postagem e como exibi-la. Será executada uma consulta com oito junções de tabela apenas para recuperar o conteúdo de um post.



# NoSQL vs SQL

- Uma opção seria o NoSQL. Transformando a postagem em um documento JSON, e armazenando-o no Banco de Dados de Documentos, um serviço de banco de dados de documento NoSQL do Azure, aumentando o desempenho e recuperando a postagem inteira com apenas uma consulta e sem junções.



# NoSQL vs SQL

```
{
  "id": "ew12-res2-234e-544f",
  "title": "post title",
  "date": "2016-01-01",
  "body": "this is an awesome post stored on NoSQL",
  "createdBy": "User",
  "images": ["http://myfirstimage.png", "http://mysecondimage.png"],
  "videos": [
    {"url": "http://myfirstvideo.mp4", "title": "The first video"},
    {"url": "http://mysecondvideo.mp4", "title": "The second video"}
  ],
  "audios": [
    {"url": "http://myfirstaudio.mp3", "title": "The first audio"},
    {"url": "http://mysecondaudio.mp3", "title": "The second audio"}
  ]
}
```

# Hadoop

- “O Hadoop é usado para aplicações analíticas de dados massivos. Criado em 2005 pela Yahoo!, pode ser considerado uma das maiores invenções de data management desde o modelo relacional. Hoje é um dos projetos da comunidade Apache e vem sendo adotado por empresas que precisam tratar volumes massivos de dados não estruturados.”



# Hadoop

- O Hadoop, na prática, é uma combinação de dois projetos separados:
- HMR (Hadoop MapReduce);
- HDFS (Hadoop Distributed File System).





# Hadoop

- É um projeto OpenSource com licenciamento Apache, permitindo a criação de distribuições específicas, como o Amazon Elastic MapReduce, que permite às empresas tratarem dados massivos sem demandar aquisição de servidores físicos. O usuário escreve a aplicação Hadoop e roda em cima da nuvem da Amazon.



# Hadoop

- Algumas empresas que contribuem com seu desenvolvimento e o utilizam:



- Algumas empresas que o utilizam:

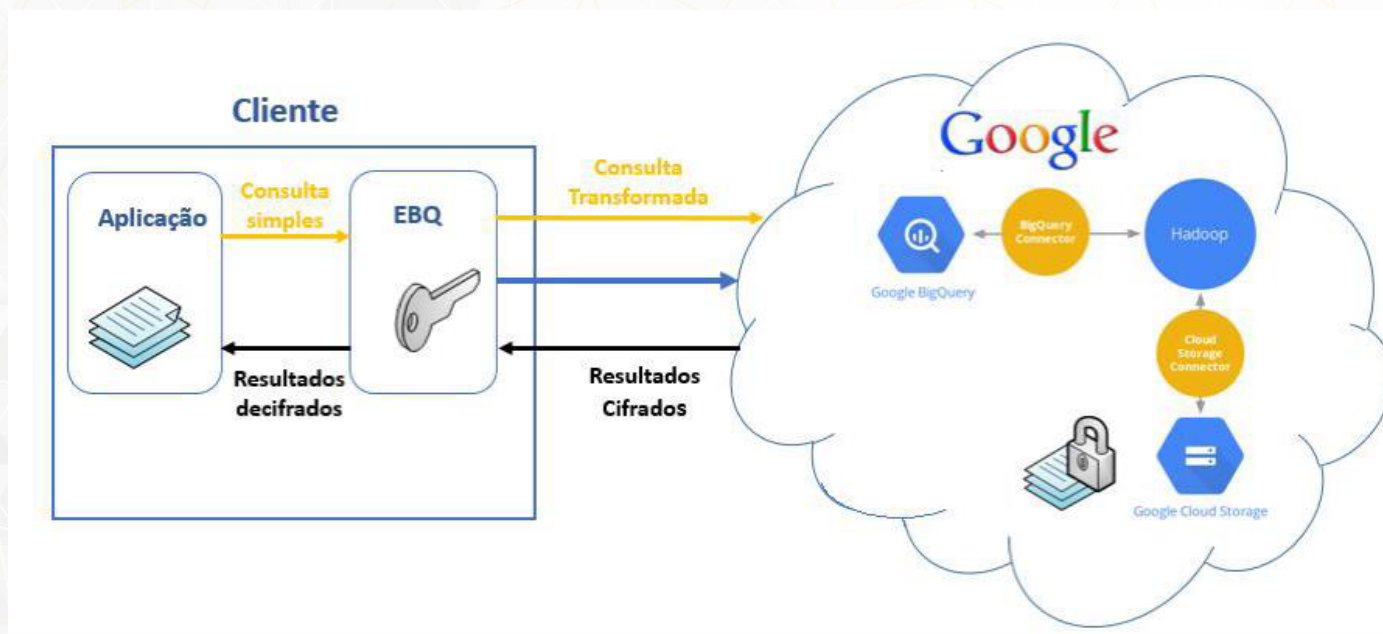


# Hadoop

- “Recentemente, uma pesquisa mostrou que pelo menos 20 empresas da lista da Fortune 1000 assumiram publicamente que usam Hadoop de alguma forma.”
- “Uma pesquisa pelo termo Hadoop no Google Trends já aponta um crescimento significativo no interesse pela tecnologia.”



# Proposta de Armazenamento de Dados Seguros



# Proposta de Armazenamento de Dados Seguros

- A consulta de dados criptografados utiliza, para cada tipo de criptografia definido, um conjunto de criptossistemas que trate esses dados adquadamente, sem a perda de confidencialidade. Os quais serão vistos a seguir.

# Proposta de Armazenamento de Dados Seguros

- **Algoritmo probabilístico:** a cifra é dita probabilística quando para valores diferentes existirem cifras diferentes com grande probabilidade. Apesar de garantir integridade e confidencialidade, não é possível realizar manipulações com os dados cifrados, exceto o comando SELECT.

# Proposta de Armazenamento de Dados Seguros

- **Algoritmo determinístico:** o modelo da criptografia é dito determinístico quando é gerado a mesma cifra para a mesma mensagem em claro. Esse esquema permite a realização de consultas de igualdade, isto é, pode realizar o comando SELECT com predicados de igualdade, junções de igualdade, COUNT.

# Proposta de Armazenamento de Dados Seguros

- **Busca por palavras:** nesse esquema é calculada a função hash de todas as sequências possíveis de palavras. Em seguida, os hashes são mantidos em um campo e separados por espaços; e assim pode ser usada uma cláusula WHERE com checagem de conteúdo (CONTAINS) com palavras-chaves inteiras, mas não aparece em consultas SELECT simples.



# Proposta de Armazenamento de Dados Seguros

- **Busca probabilística por palavras:** consiste em buscar uma palavra com o retorno de todas as posições em que ela aparece no texto em claro. Permite consultas com o uso da cláusula WHERE e atributo CONTAINS ou LIKE.

# Proposta de Armazenamento de Dados Seguros

- **Criptografia homomórfica:** criptografia que permite lidar com tipos específicos de cálculos em cifras e gerar um resultado também cifrado que, quando decifrado, corresponde ao resultado de operações realizadas em textos em claro.

# Análise e Resultados

- Foi aplicado modelos criptográficos que atendessem todos os tipos criptográficos em estudo, avaliando o desempenho e o custo operacional adicionados por cada criptossistema.

# Análise e Resultados

- Foi utilizado um ambiente com um computador com processador core I5 1,8GHz, 6GB de RAM e Linux Ubuntu 14.04; executando a ferramenta EBQ conectada ao Google BigQuery.

# Análise e Resultados

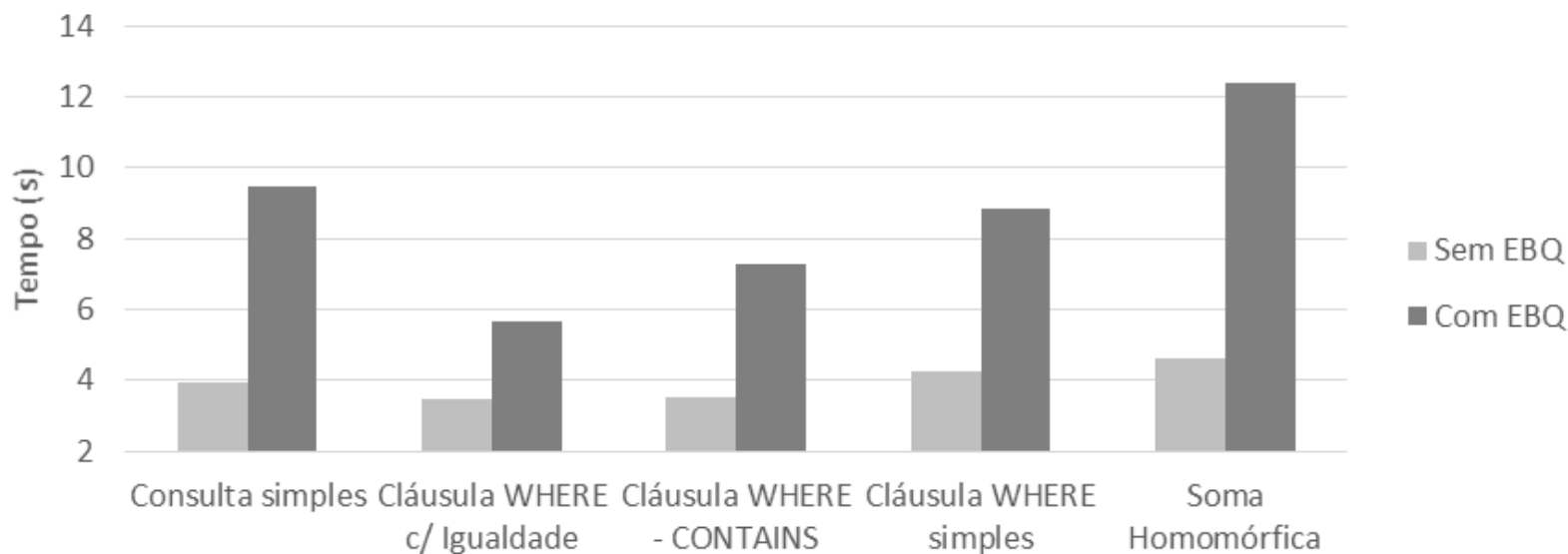
- Foi utilizado utilizado um esquema orientado a colunas, tendo como tema um cadastro pessoal que possui um número de identificação e diversas famílias de colunas, aonde um tipo de criptografia foi escolhido para cada coluna de forma a permitir a análise de sua influência no desempenho da consulta. Realizando a inserção com dois volumes de entrada (5000 e 50000 registros).

# Análise e Resultados

- Foi avaliado o tempo de resposta de cada consulta EBQ, usando como comparação uma tabela sem criptografia com os mesmos registros. Cada processo de consulta foi repetido 50 vezes, de forma a obter um tempo médio, uma vez que a maior parte do processamento é realizada nos servidores do Google em um ambiente não-controlado em que não se pode garantir a expectativa de execução correta e homogênea.

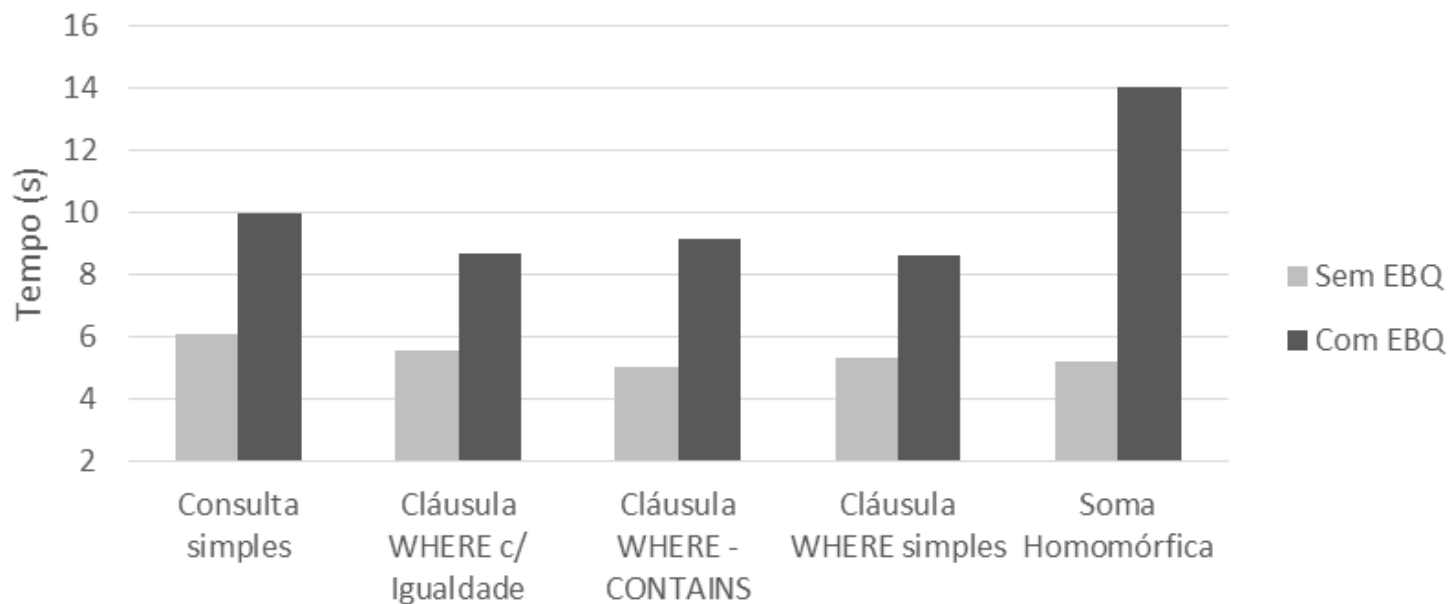
# Análise e Resultados

Tempo Gasto em consultas: 5000 registros



# Análise e Resultados

Tempo Gasto em Consultas - 50000 registros










# Conclusões

- Verifica-se, portanto, que a aplicação de consultas diretas sobre grandes volumes de dados criptografados armazenados em ambiente distribuído – como é o caso do EBQ – adiciona um atraso ao intervalo de resposta. Porém, à medida que o volume de dados cresce, essa variação se estabiliza e esse gasto extra de tempo e de memória não se torna tão relevante, levando em conta a velocidade com que as operações são efetuadas.

# Bibliografia

- [https://pt.wikipedia.org/wiki/Big\\_data](https://pt.wikipedia.org/wiki/Big_data) 
- <http://imasters.com.br/tecnologia/redes-e-servidores/analise-de-dados-na-nuvem-duas-boas-opcoes-de-big-data-nosql-para-as-pmes/?trace=1519021197&source=single> 
- <http://imasters.com.br/artigo/21026/banco-de-dados/conhecendo-o-hadoop?trace=1519021197&source=single>  
- <https://cloud.google.com/bigquery/what-is-bigquery> 
- <https://azure.microsoft.com/pt-br/documentation/articles/documentdb-nosql-vs-sql/>