FERRAMENTA PARA PROCESSAMENTO E INTEGRAÇÃO DE DADOS GOVERNAMENTAIS ABERTOS

PEDRO MAGALHÃES BERNARDO

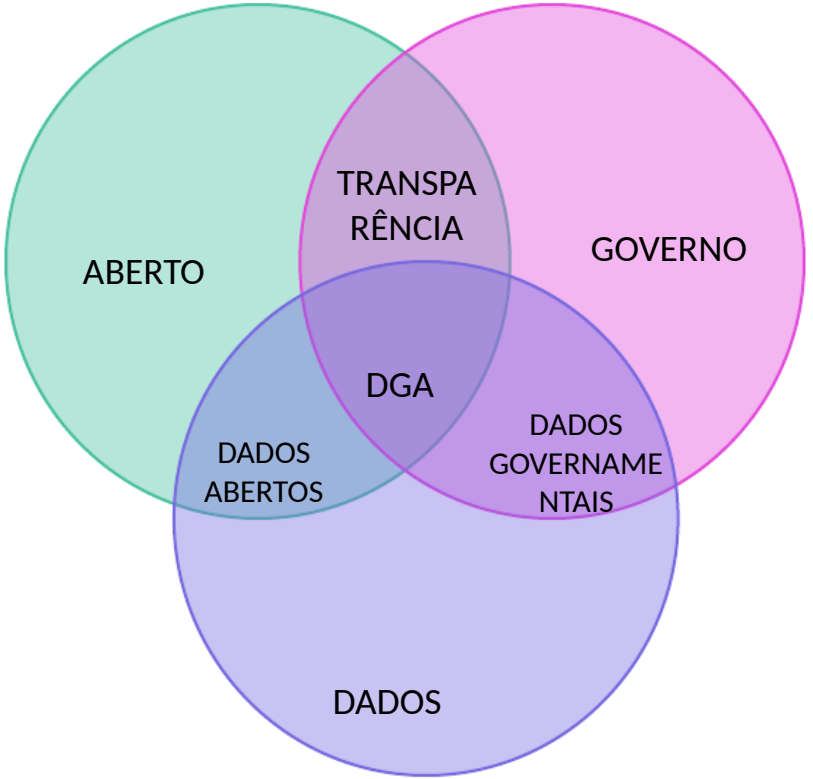
Orientador: Ismael Santana Silva

Coorientadores: Glívia Angélica Rodrigues Barbosa
Flávio Roberto dos Santos Coutinho

Contextualização

Demanda popular por mais **transparência das ações governamentais**.

Novas políticas, como a publicação de **dados de interesse público em estado bruto**, são os chamados **dados governamentais abertos**.



Contexto Brasileiro

No Brasil foi criado a **Lei nº 12.527/2011** que permite a qualquer cidadão a obtenção de **dados** e **informações** de qualquer **entidade pública**.

Motivação

A forma como os dados são disponibilizados não permite a obtenção de **informações relevantes** sem o uso de **ferramentas computacionais**.

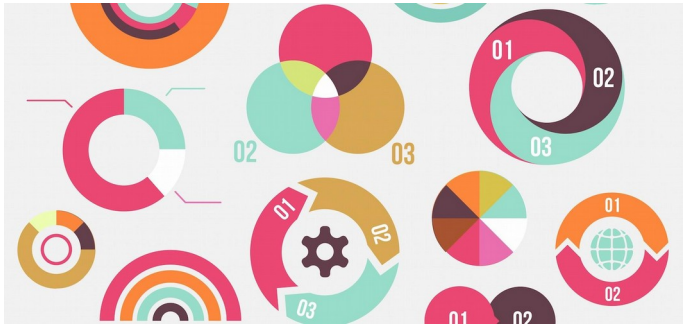
Dados são **heterogêneos**, em **diversos formatos** e em **grande volume**.

Motivação

Dois desafios:

(1) Infraestrutura capaz de **processar** e **integrar** os DGA, de forma a viabilizar a **exploração** e **análise** dessas bases de forma conjunta.

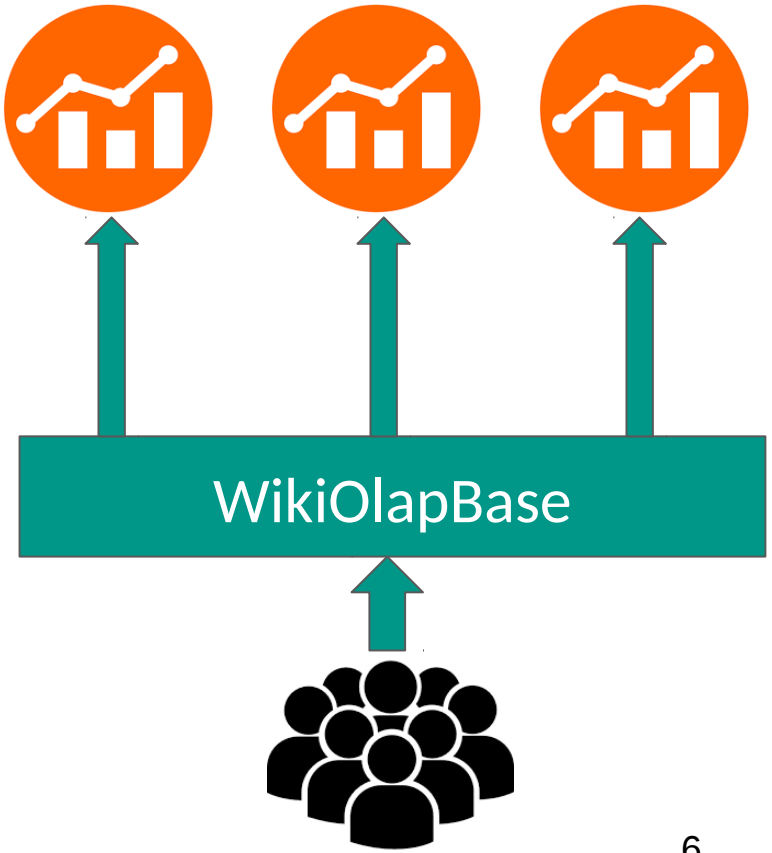
(2) Ferramenta de visualização de dados alimentada por essa infraestrutura



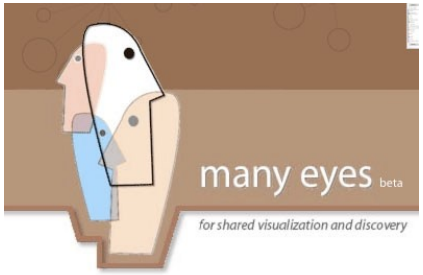
Objetivos

WikiOlapBase: uma **ferramenta colaborativa** que seja capaz de **processar** e **integrar** dados abertos.

Infraestrutura base para outras ferramentas de **análise** e **visualização** de **grandes volumes de dados**.

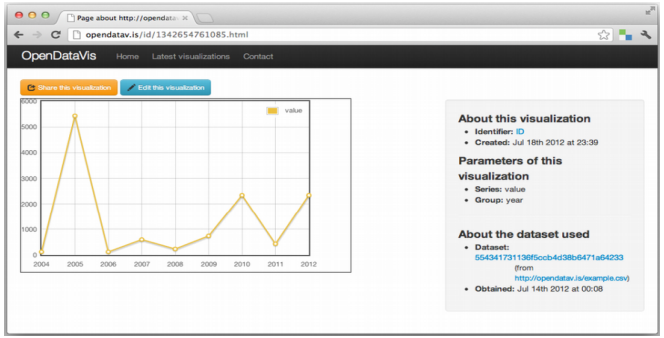


Trabalhos Relacionados



ManyEyes - Projeto desenvolvido pela IBM. Ferramenta de visualização colaborativa. (VIEGAS et al.,2007)

DataGovWiki - Projeto desenvolvido por Ding et al. (2010) – Integrar dados do site data.gov



OpenDataVis - Projeto desenvolvido por Graves e Hendler (2013) – Demonstram a importância da utilização de visualizações no contexto dos DGA.

Referência	Modelo de Dados	Forma de acesso aos dados	Formato de importação dos dados	Importação de dados por usuários	Acesso a base de dados de outros usuários	Disponibilização de metadados	Cruzamento entre dados*
OpenDataVis - Graves e Hendler (2013)	Linked Data	Interface gráfica	Não especificado	Não especificado	Não especificado	Sim	Não
Hoxha e Brahaj (2011)	Linked Data	Interface gráfica e consultas SPARQL	XML, CSV, Texto	Não	Não	Sim	Não
DataGovWIki - Ding et al. (2010)	Linked Data	Webservice SPARQL	CSV	Não	Não	Sim	Não
Many Eyes - Viegas et al. (2007)	Tabela e texto não estruturado	Interface Gráfica	Texto separado por tabulação	Sim	Sim	Sim	Não
Rivet - Tang et al. (2014)	Relacional	API Rest	CSV, MDX, e conexões SQL	Sim	Não	Sim	Não

Metodologia

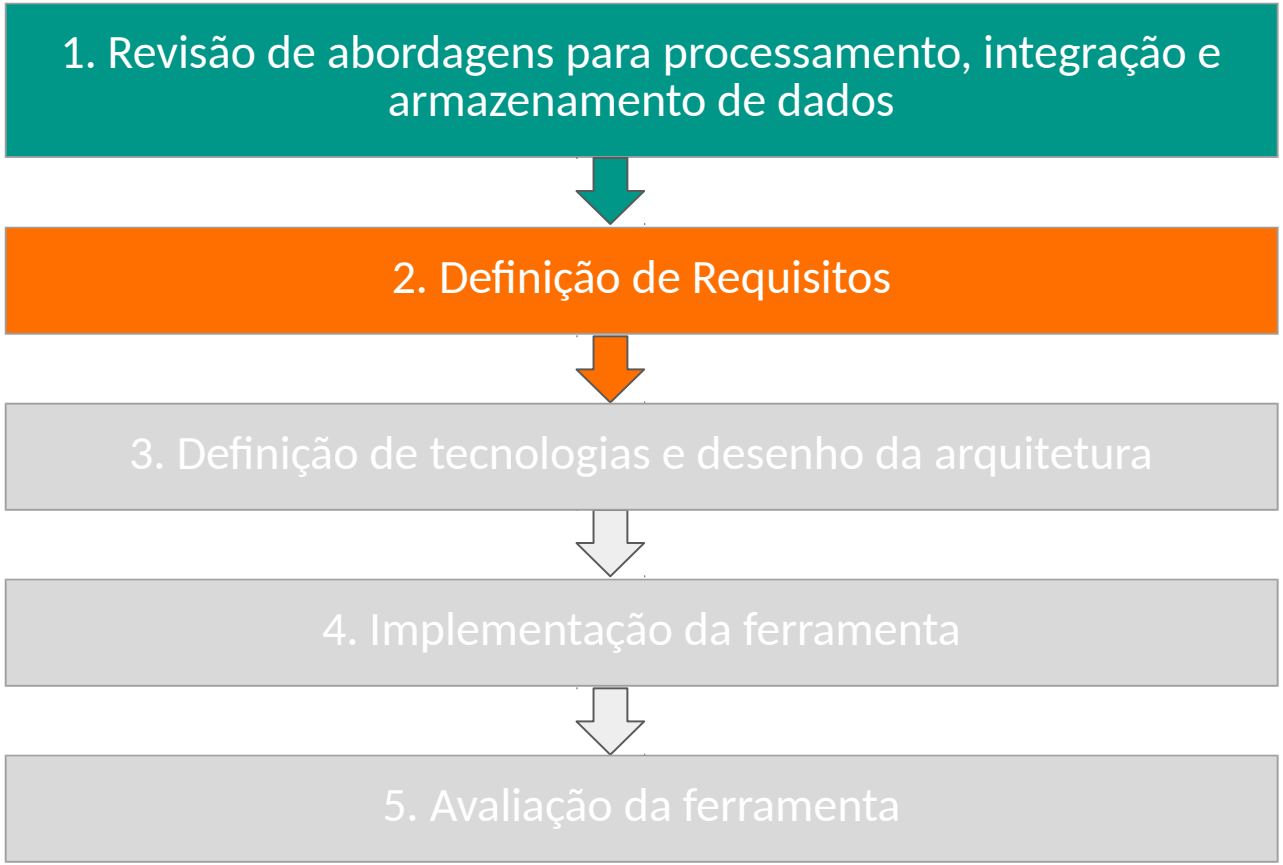
1. Revisão de abordagens para processamento, integração e armazenamento de dados

2. Definição de Requisitos

3. Definição de tecnologias e desenho da arquitetura

4. Implementação da ferramenta

5. Avaliação da ferramenta


1. Revisão de abordagens para processamento, integração e armazenamento de dados
2. Definição de Requisitos
3. Definição de tecnologias e desenho da arquitetura
4. Implementação da ferramenta
5. Avaliação da ferramenta

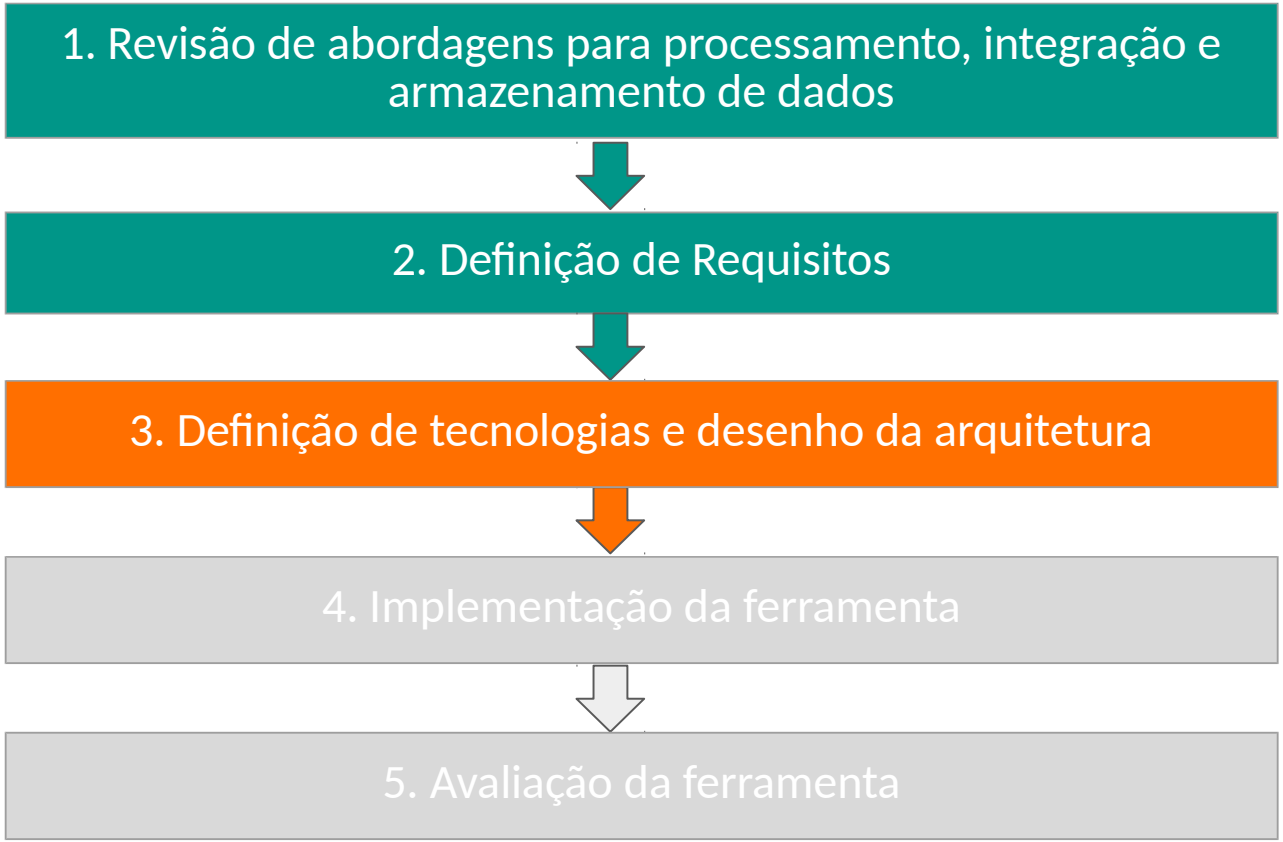
Requisitos

Objetivo: Definir as funcionalidades e características do software proposto.

Como: Reunião de *brainstorming* no dia 29 de Abril de 2016 com três especialistas, que possuem mais de oito anos de experiência na área de processamento e análise de dados.

Resultado: Lista de requisitos para a ferramenta.


Requisitos
A ferramenta deve permitir a importação de dados por parte do usuário
A ferramenta deve manter o significado dos dados originais
A ferramenta deve permitir a definição de metadados
A ferramenta deve permitir o relacionamento entre conjuntos diferentes
A ferramenta deve permitir o armazenamento em larga escala
A ferramenta deve permitir o acesso a base de dados integrada


1. Revisão de abordagens para processamento, integração e armazenamento de dados
2. Definição de Requisitos
3. Definição de tecnologias e desenho da arquitetura
4. Implementação da ferramenta
5. Avaliação da ferramenta

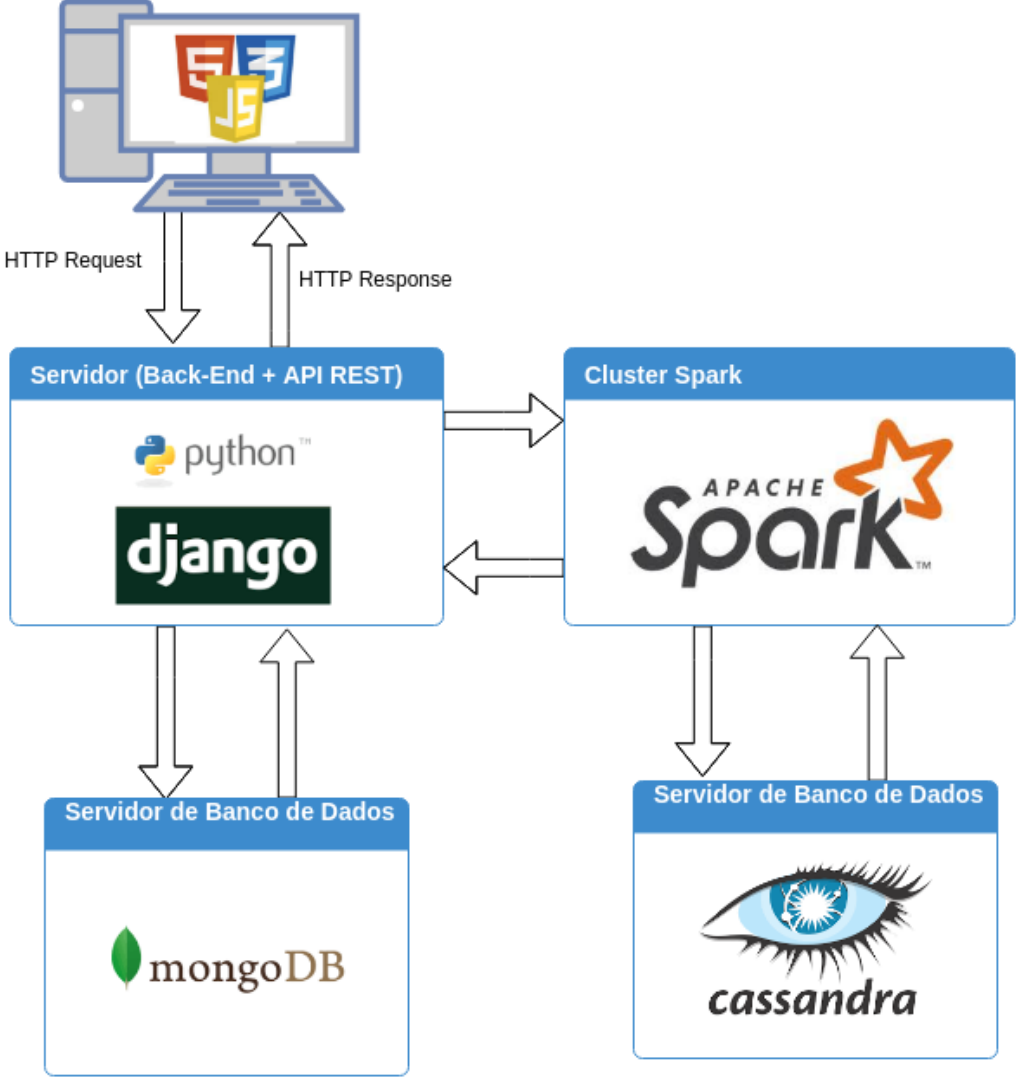
Tecnologias e Arquitetura

Objetivo: Definir linguagem de programação, modelo de dados e SGBDs, forma de acesso a dados e outras decisões de projeto.

Como: A partir da revisão bibliográfica e dos requisitos levantados.

Resultado: Desenho da arquitetura do sistema.

Arquitetura



Front-End

Objetivo: Interface que permite ao usuário enviar e caracterizar um conjunto de dados.

HTML5+CSS+JS: Linguagens padrão para desenvolvimento *front-end*.

Servidor (*Back-End* + API REST)

Objetivo: Processar dados e comunicar com os servidores de bancos de dados.

Python+Django:

- Linguagem de programação popular
- Suporte a outras ferramentas
- *Don’t repeat yourself (DRY)*

Banco de Dados - Metadados

Objetivo: Armazenamento dos metadados

Modelo: Orientado a documentos

Porque:

- Metadados não possuem estrutura definida.

Banco de Dados – Dados Brutos

Objetivo: Armazenar dados brutos que foram enviados pelos usuários.

Modelo: Família de colunas.

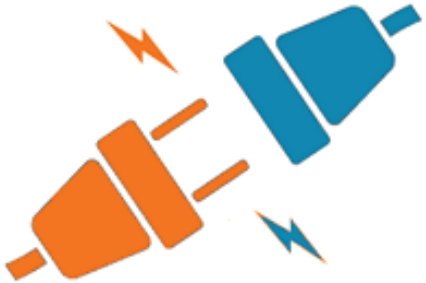
Porque:

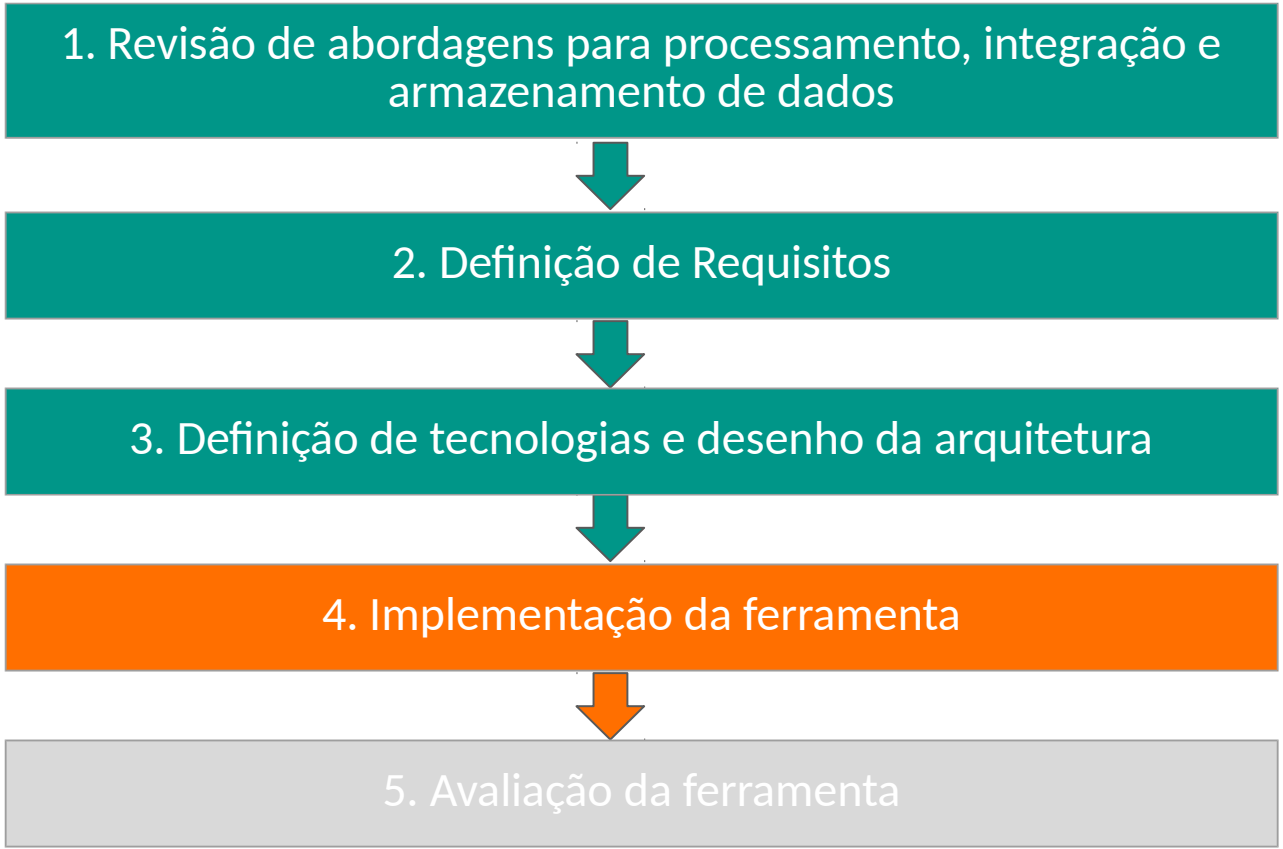
- Modelo mais próximo da realidade.
- Dados não normalizados.
- Mais adequado para operações OLAP.

Spark + Cassandra

Plataforma para processamento distribuído em memória. Permite realização de operações complexas como *join* e *order by* em qualquer fonte de dados.

Banco de dados de família de colunas, distribuído. Permite escalabilidade horizontal. Não possui suporte a operações como *join* e *order by*.




1. Revisão de abordagens para processamento, integração e armazenamento de dados
2. Definição de Requisitos
3. Definição de tecnologias e desenho da arquitetura
4. Implementação da ferramenta
5. Avaliação da ferramenta

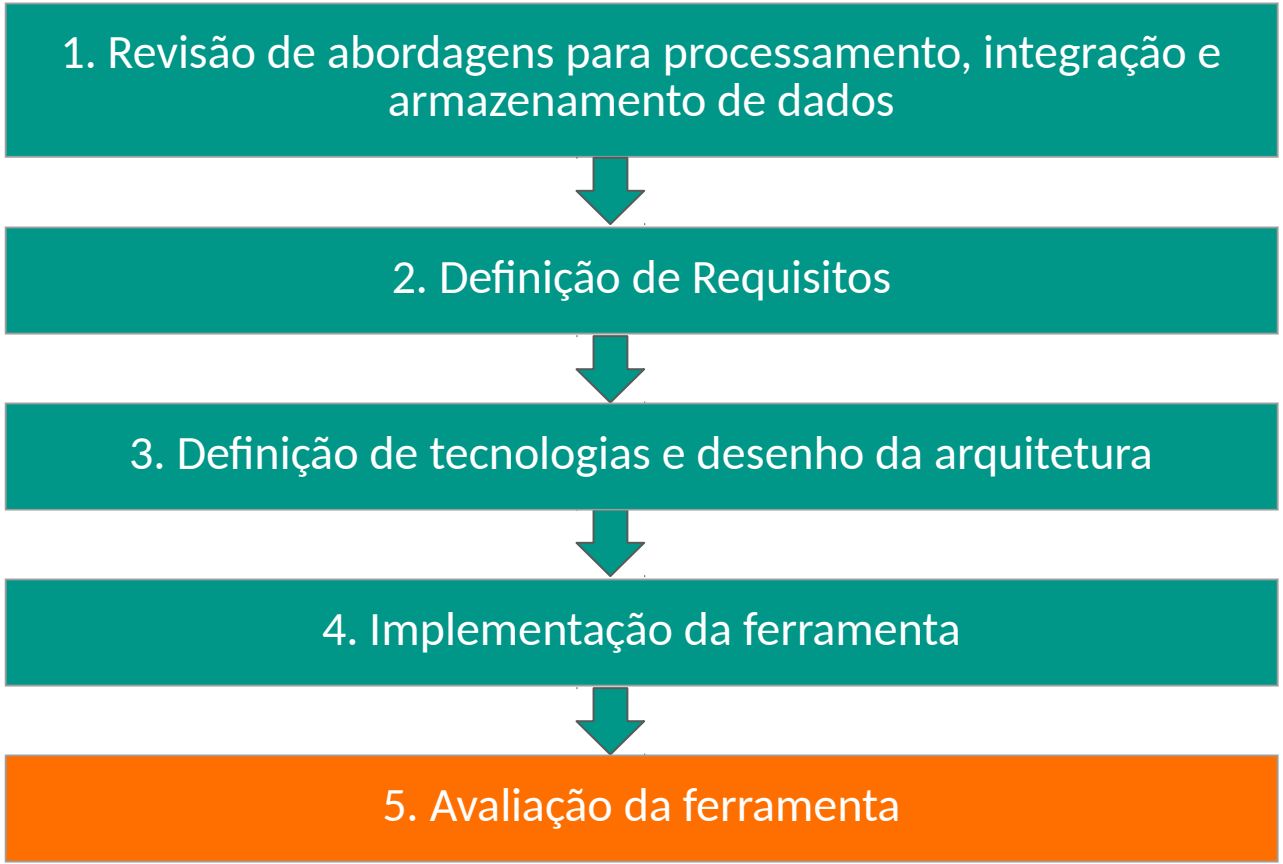
WikiOlapBase

Dois módulos:

(1) responsável por receber, caracterizar e integrar conjuntos de dados;

(2) permite acesso ao repositório integrado por meio de uma API REST.

Link para vídeo de execução:
https://www.youtube.com/watch?v=OGebiusMAlU


1. Revisão de abordagens para processamento, integração e armazenamento de dados
2. Definição de Requisitos
3. Definição de tecnologias e desenho da arquitetura
4. Implementação da ferramenta
5. Avaliação da ferramenta

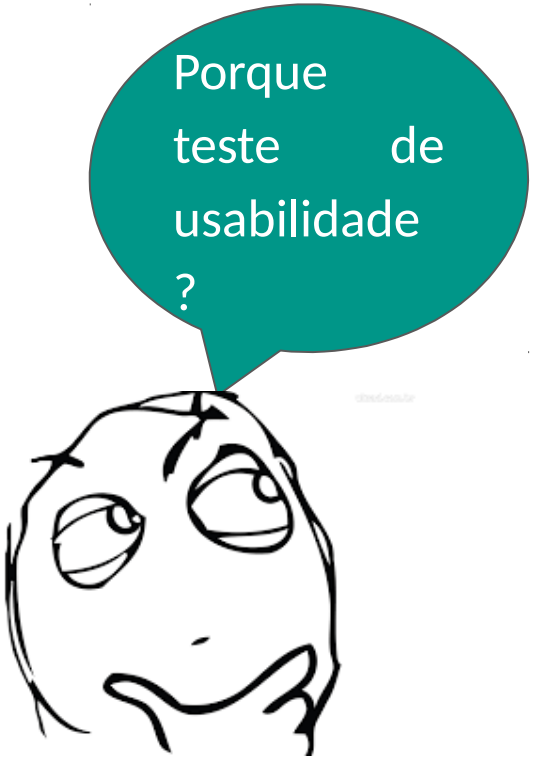
Avaliação

Objetivo: Avaliar a adequação ao uso da ferramenta.

Como: Teste de usabilidade de acordo com a metodologia proposta por Barbosa e Silva (2010)

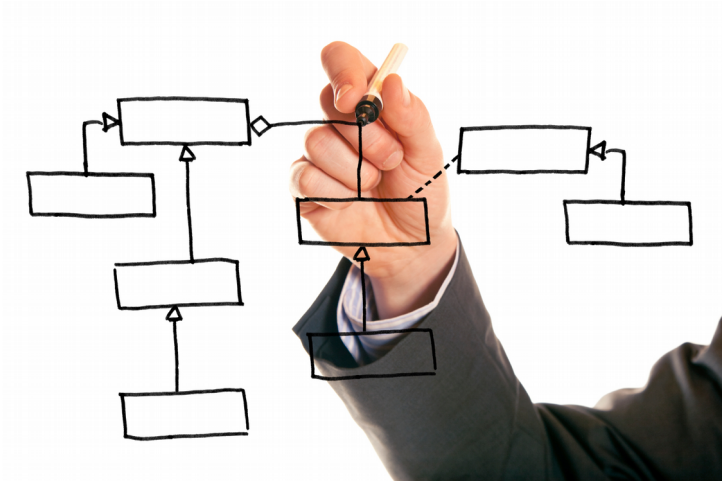
Resultado: Indicadores que caracterizam a adequação ao uso da ferramenta

TESTE DE USABILIDADE



Dois fatores fundamentais para o sucesso da ferramenta: (1) adequada para utilização por parte do público alvo; (2) permitir colaboração entre usuários.

Avaliação - Metodologia



Definições: Métricas de usabilidade/colaboração. Tarefas a serem executadas

Execução: Observar tarefas executadas. Questionário pós-teste.

Avaliação: Verificação das medidas de usabilidade/colaboração.

Avaliação

- Testes realizados com 6 usuários entre 27 de setembro de 2016 e 29 de setembro de 2016.
- Segundo Nielsen (2000), testes de usabilidade devem ser executados por 3 a 5 usuários.
- 10 tarefas que geraram 3 cenários

Avaliação - Cenários

Enviar

1. Enviar um conjunto de dados e gerar uma visualização a partir do mesmo

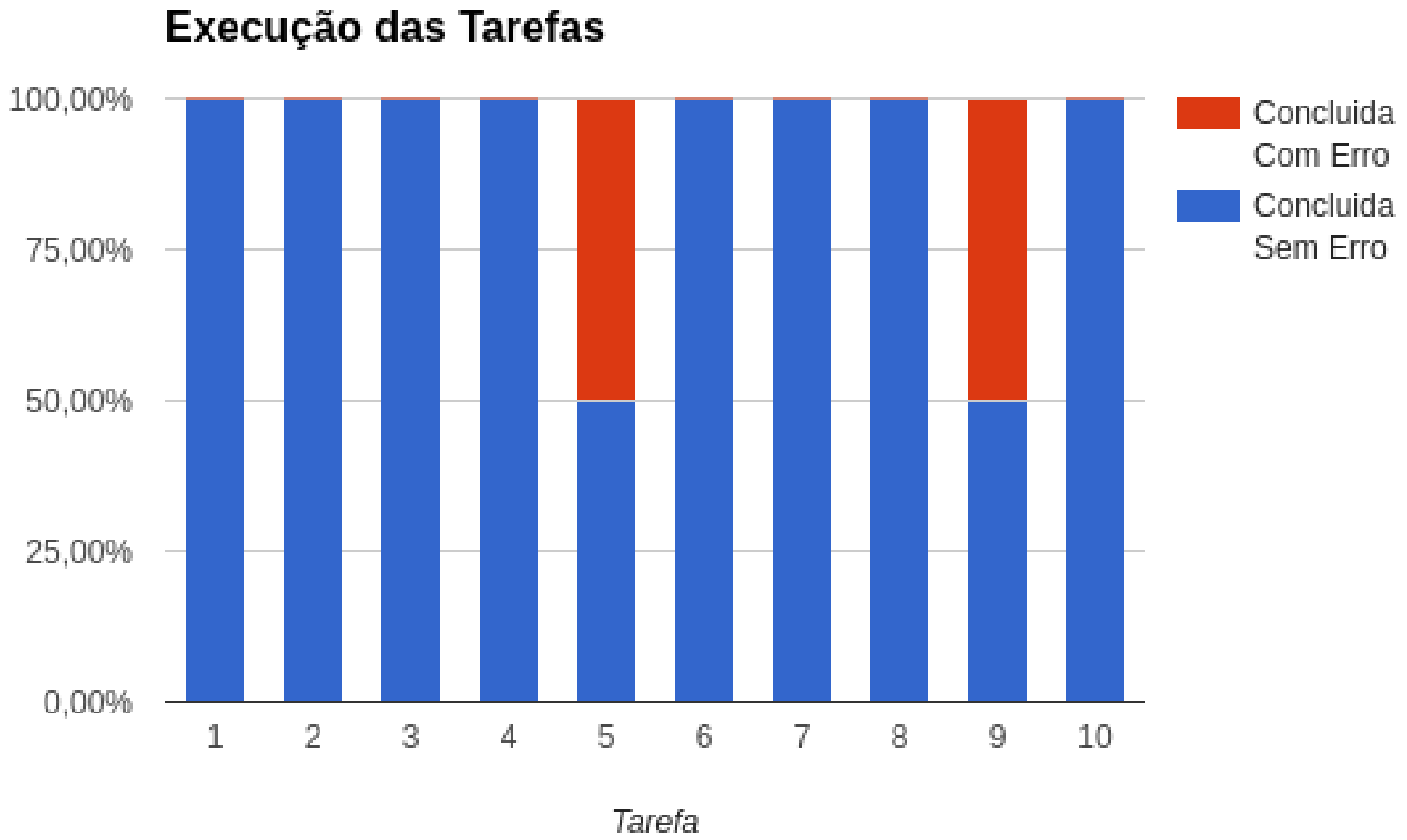
Enviar e Cruzar

2. Enviar um conjunto de dados e fazer o cruzamento do mesmo com outro conjunto já presente no repositório

Cruzar

3. Utilizar dois conjuntos de dados já existentes no repositório e gerar uma visualização a partir deles

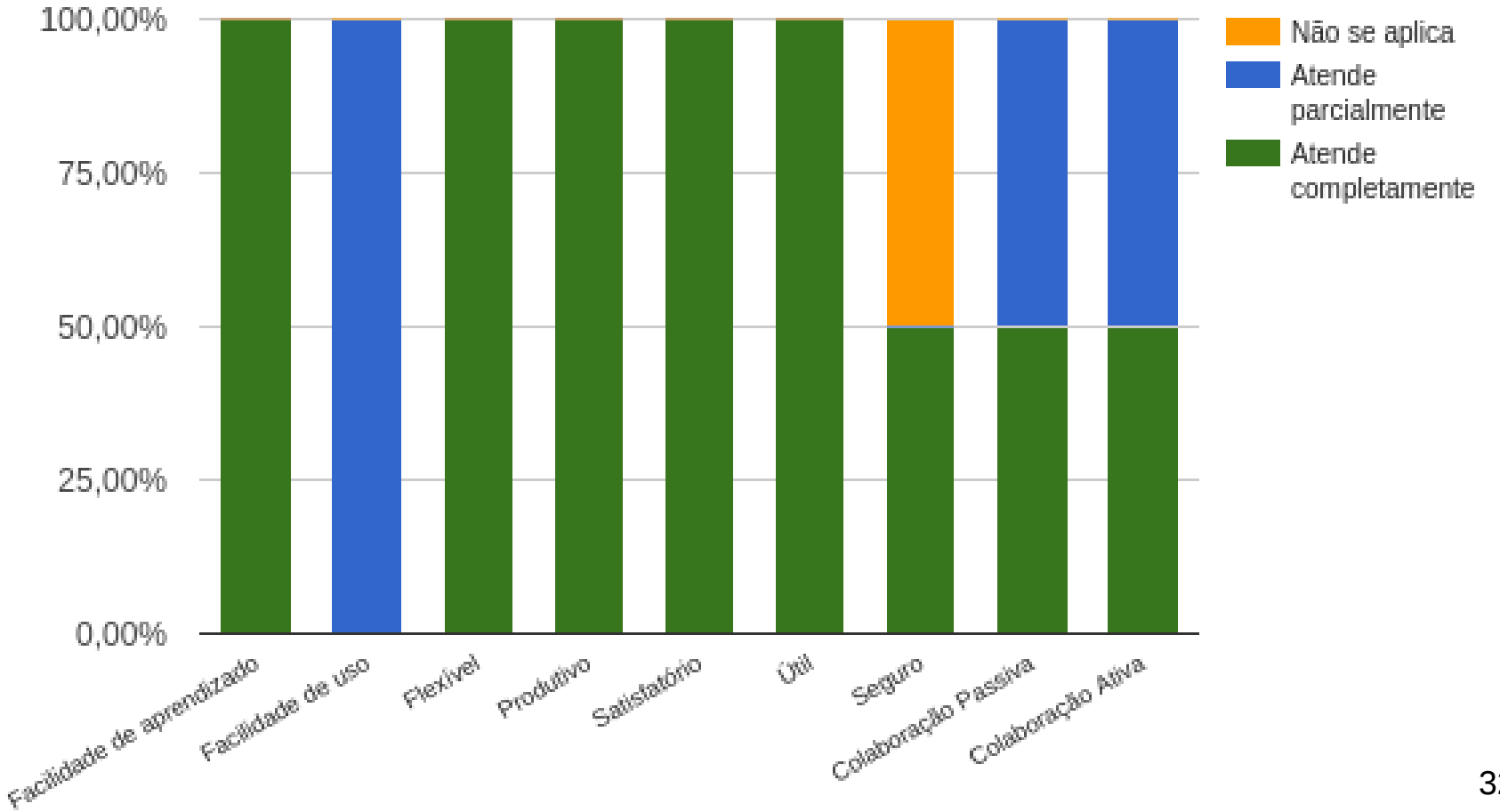
Avaliação - Resultados



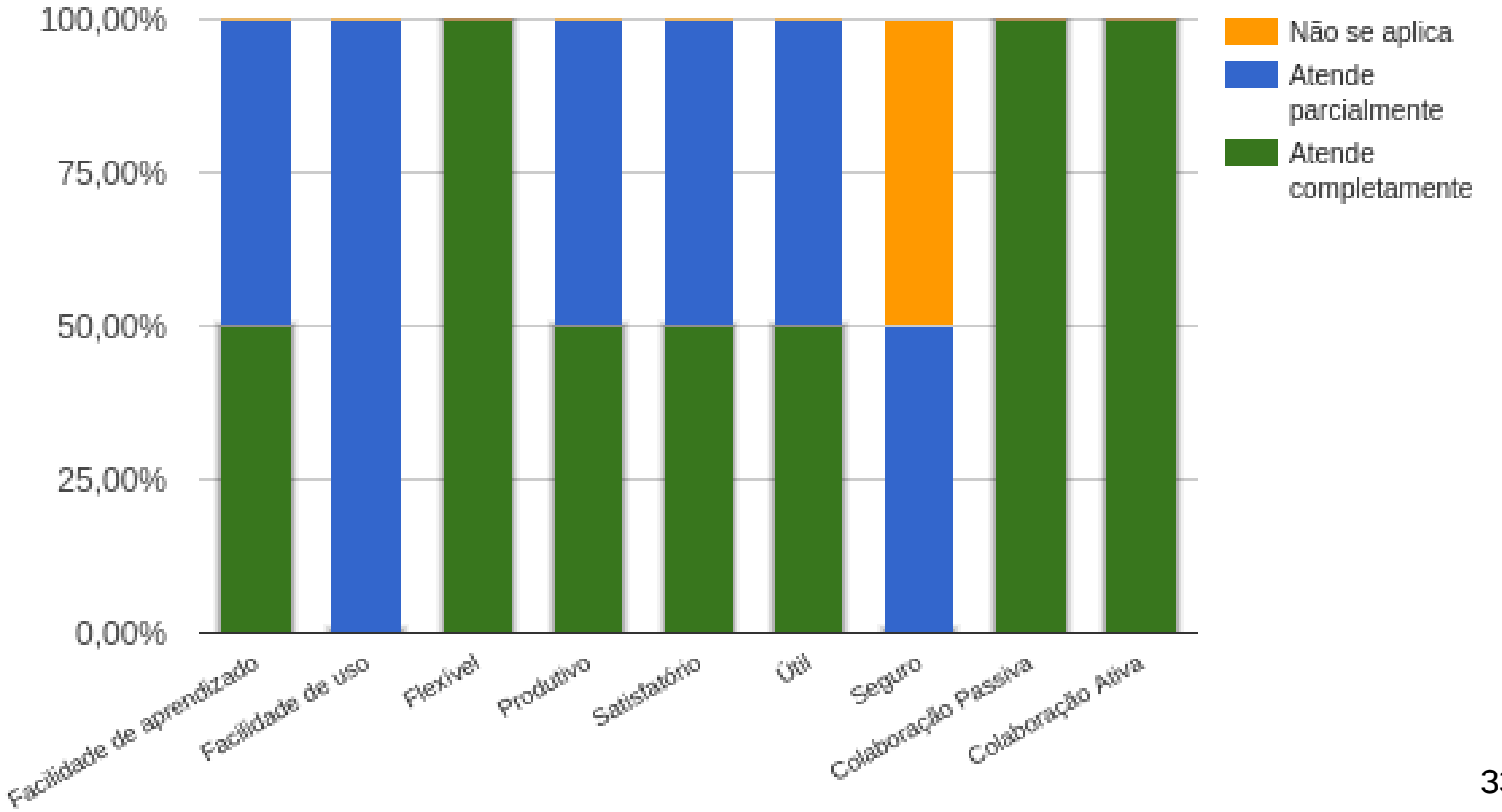
Avaliação – Grau de Adequação

- 7 princípios de usabilidade de Nielsen (1994).
- 2 princípios de colaboração definidos: colaboração ativa e colaboração passiva.
- Separado por cenários.

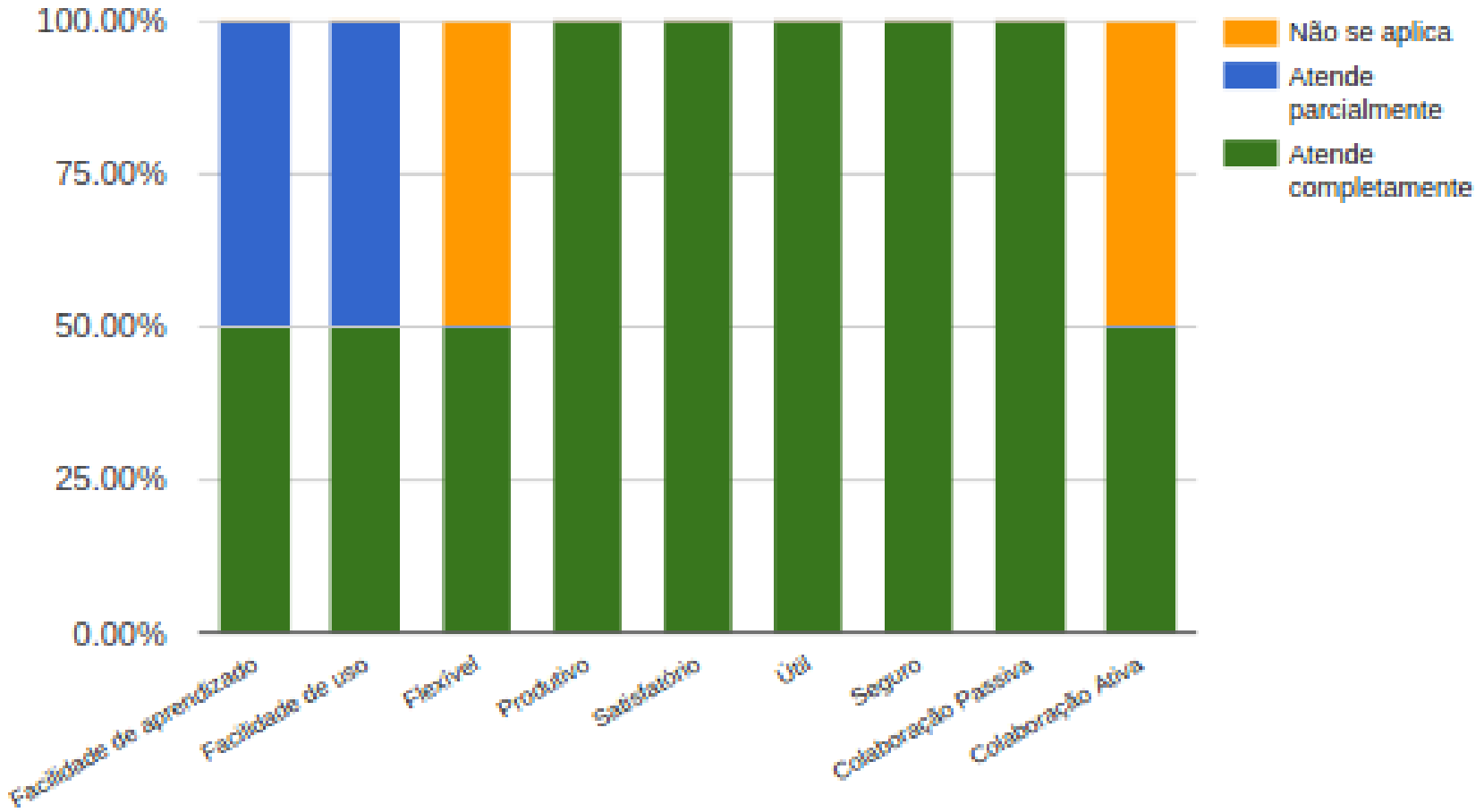
Grau de Adequação – Cenário 1: Enviar



Grau de Adequação – Cenário 2: Enviar e Cruzar

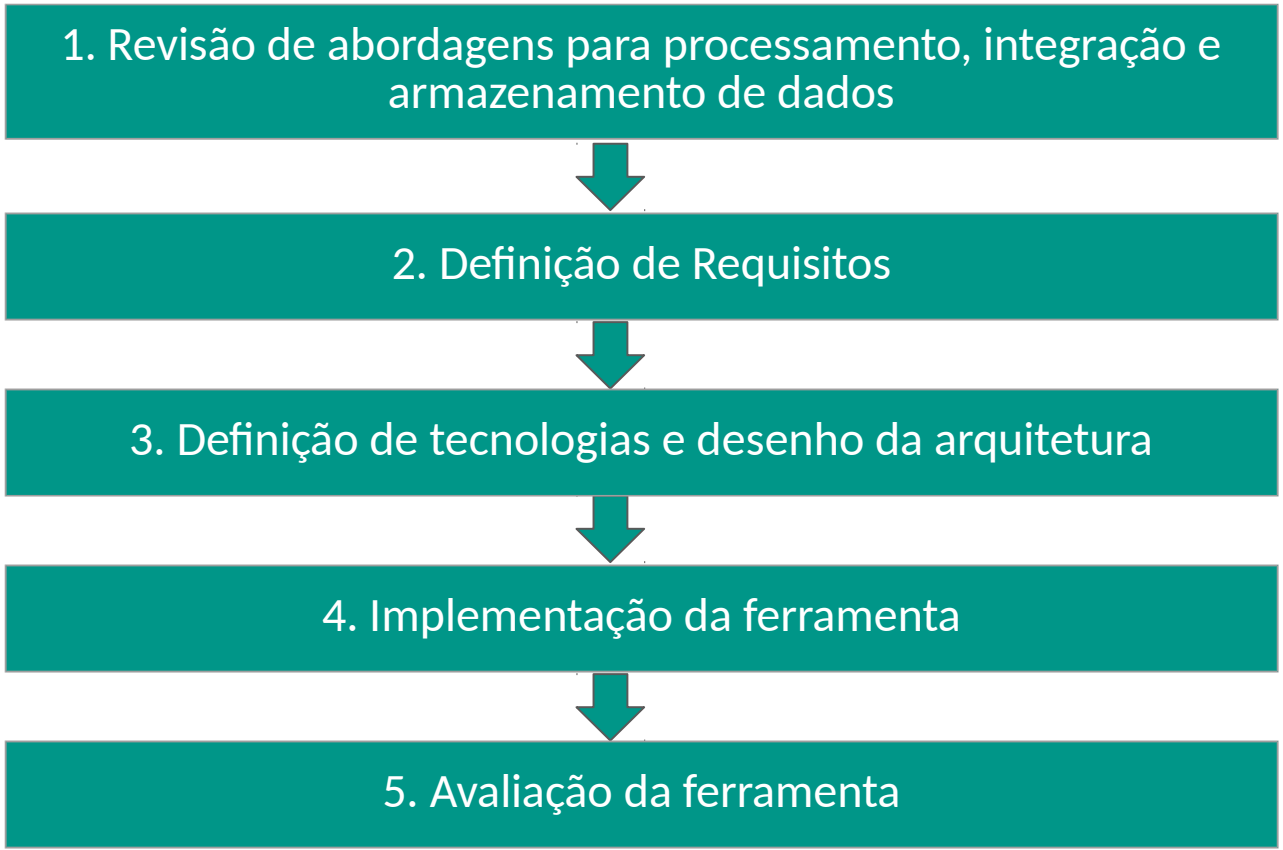


Grau de Adequação – Cenário 3: Cruzar



Avaliação - Conclusão

- Nenhum princípio deixou de ser atendido.
- Ferramenta adequada ao uso.
- *Feedback* positivo por parte dos usuários em relação a ideia por trás da ferramenta e do fluxo de execução.


1. Revisão de abordagens para processamento, integração e armazenamento de dados
2. Definição de Requisitos
3. Definição de tecnologias e desenho da arquitetura
4. Implementação da ferramenta
5. Avaliação da ferramenta

Conclusões

Os resultados da avaliação mostram que a ferramenta é **útil**, **satisfatória**, **adequada ao uso** e permite a **integração de dados** de forma **colaborativa**.

O estudo de ferramentas similares presente neste trabalho permite explorar diferentes abordagens e técnicas para criação de ferramentas para **integração de dados.**

Contribuições

Contribuições Práticas

- Adicionar elementos de colaboração no processo de integração de dados
- Disponibilização do código fonte

Contribuições Científicas

- Vantagens e desvantagens de diferentes abordagens para processamento, integração e armazenamento de dados

Trabalhos Futuros

Segunda fase - Ferramenta de **visualização** que consome a API

Evolução da ferramenta:

- Sistema de **cadastro/autenticação** de usuários
- Outros **formatos** de **arquivos**
- Estender API para permitir outras **operações**
- Utilização de URIs **semânticas** para identificação de ***tags*** - e.g.: schema.org

Trabalhos Futuros

Benchmarking de desempenho

Análise da **infraestrutura** para **disponibilizar** a ferramenta para o **público**

Referências

- BARBOSA, S.; SILVA, B. S. D. Interação Humano-Computador. São Paulo: Elsevier, 2010.
- DING, L. et al. Data-gov wiki: Towards linking government data. In: . [S.l.: s.n.], 2010
- VIEGAS, F. B. et al. Manyeyes: A site for visualization at internet scale. IEEE Transactions on Visualization and Computer Graphics, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 13, n. 6, p. 1121–1128, nov. 2007. ISSN 1077-2626. Disponível em: <http://dx.doi.org/10.1109/TVCG.2007.70577>
- GRAVES, A.; HENDLER, J. Visualization tools for open government data. In: Proceedings of the 14th Annual International Conference on Digital Government Research. New York, NY, USA: ACM, 2013. (dg.o '13), p. 136–145. ISBN 978-1-4503-2057-3. Disponível em: <http://doi.acm.org/10.1145/2479724.2479746>.
- HOXHA, J.; BRAHAJ, A. Open government data on the web: A semantic approach. In: IEEE. Emerging Intelligent Data and Web Technologies (EIDWT), 2011 International Conference on. [S.l.], 2011. p. 107–113
- TANG, D. et al. Design choices when architecting visualizations. Information Visualization, Palgrave Macmillan, v. 3, n. 2, p. 65–79, jun. 2004. ISSN 1473-8716. Disponível em: <http://dx.doi.org/10.1057/palgrave.ivs.9500067>.
- NIELSEN, J. Why You Only Need to Test with 5 Users. 2000. [Online; acesso 23-Outubro-2016]. Disponível em: <https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/>.
- NIELSEN, J. Usability inspection methods. In: ACM. Conference companion on Human factors in computing systems. [S.l.], 1994. p. 413–414.

Referências

- Código da ferramenta disponível em: https://github.com/pedromb/wikiolapbase
- Documentação da API disponível em: http://docs.stormwind.apiary.io/

FERRAMENTA PARA PROCESSAMENTO E INTEGRAÇÃO DE DADOS GOVERNAMENTAIS ABERTOS

Obrigado!