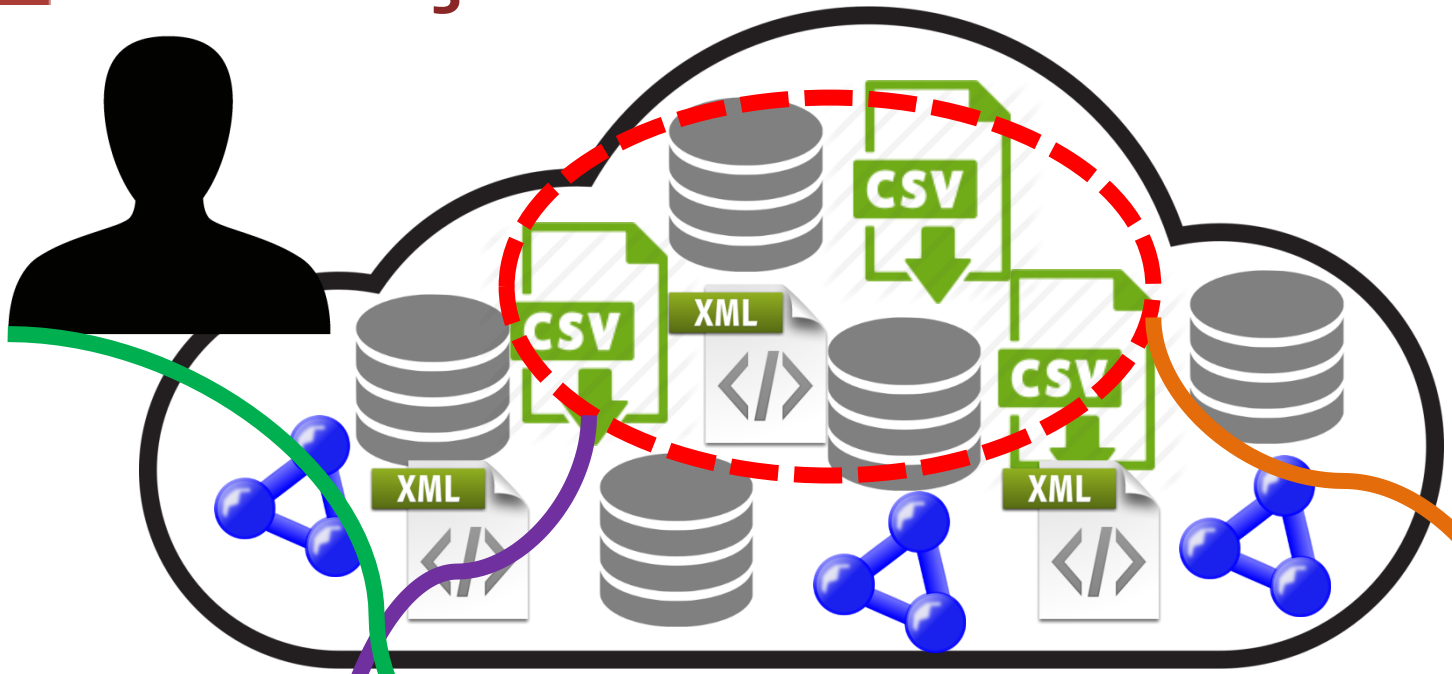
Geração de um Perfil de Qualidade para Fontes de Dados Dinâmicas

Everaldo Neto, Bernadette Lóscio, Ana Carolina Salgado
(ecsn, bfl, acs)@cin.ufpe.br
Universidade Federal de Pernambuco – UFPE
Centro de Informática - CIn

Introdução

- Identificar quais fontes de dados são mais adequadas para um determinado uso é um **grande desafio**
  - Grande número de fontes de dados disponíveis
  - **Ausência de informações sobre a qualidade dos dados**
- Considerando o contexto da Web...
  - Ambiente flexível e heterogêneo

Introdução



Tarefa exaustiva!!!

nº alto de fontes

fontes desconhecidas

heterogeneidade de formatos/tipos de dados

Introdução

- Uso de **critérios de QI** para solucionar o problema
- Uso de estratégias de **avaliação pontual**
- Fontes de dados podem ser **dinâmicas**
  - Valores podem variar de acordo com mudanças

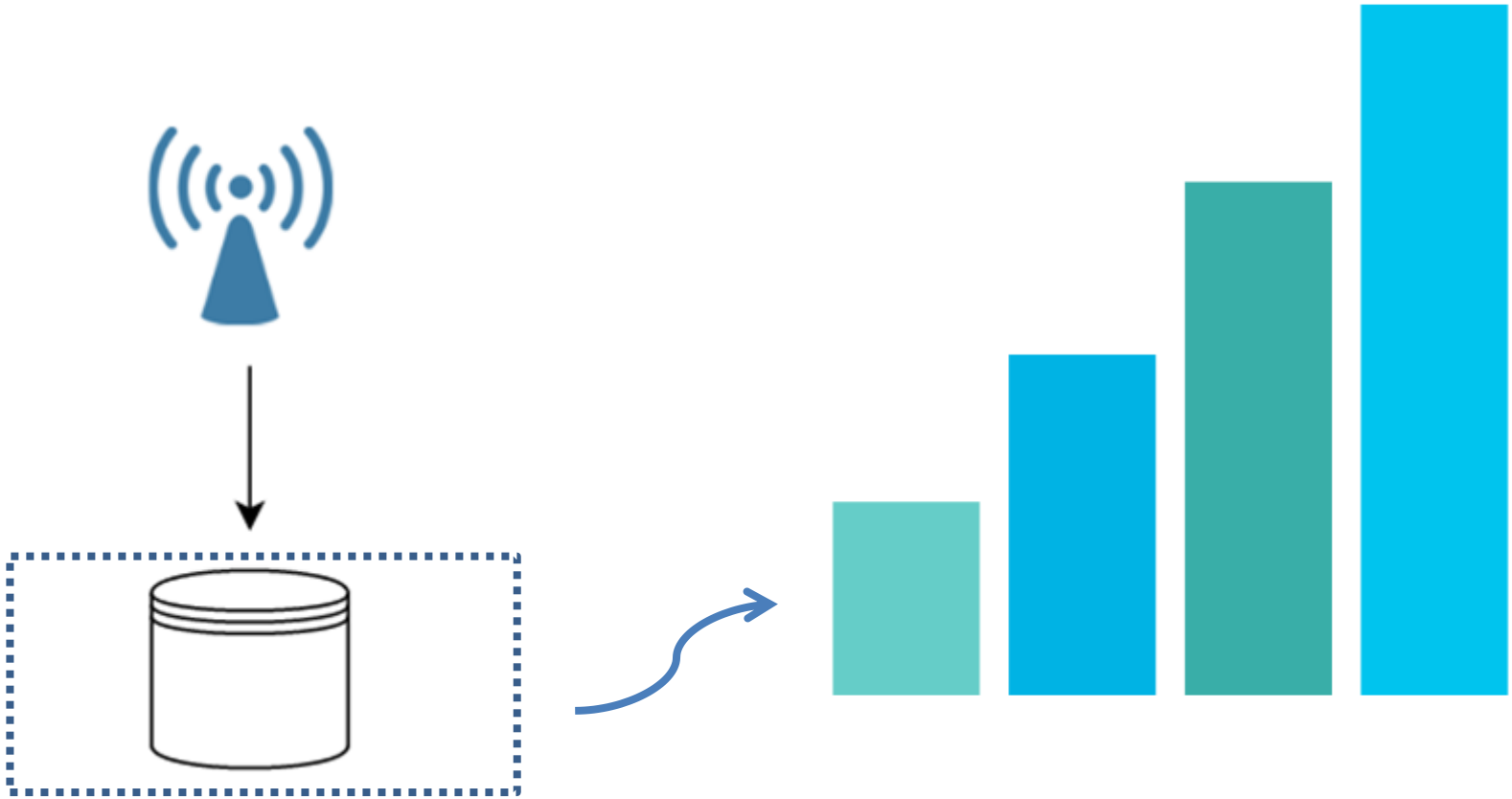
Proposta

- **Avaliar a qualidade** das fontes de dados
  - Conjunto de critérios de QI
  - Considerando o aspecto dinâmico das fontes de dados
- Gerar um **Perfil de Qualidade** para as fontes de dados
  - Atualizado periodicamente
  - Utilizado em diversos contextos

Roteiro

- Definição do Problema
- Perfil de Qualidade
- Geração do Perfil de Qualidade
- Experimentos
- Conclusão

Definição do Problema



SBBD2016

CIn.ufpe.br

Questão de Pesquisa

Como avaliar a qualidade de uma fonte de dados, que sofre atualização de inserção de dados com uma frequência elevada, considerando que:

(i) a qualidade dos dados poderá sofrer atualizações ao longo do tempo, e

(ii) um grande volume de dados será gerado a partir das frequentes inserções?

Possível Solução

- Avaliação com um **viés incremental**, realizada de **maneira contínua**
  - Combinar resultados de avaliações
  - Diminuir o custo de processamento sem perder precisão dos resultados

Roteiro

- Definição do Problema
- Perfil de Qualidade
- Geração do Perfil de Qualidade
- Experimentos
- Conclusão

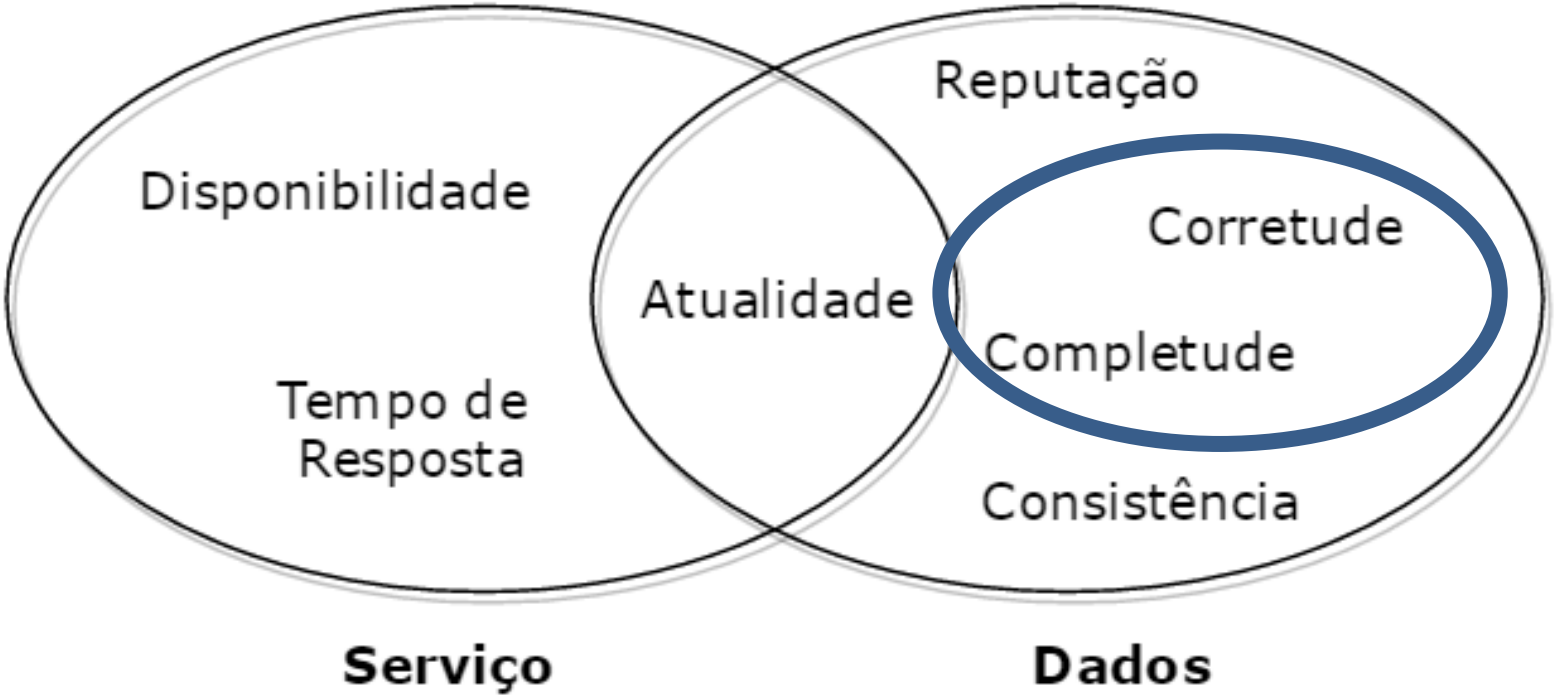
Perfil de Qualidade

- Conjunto de metadados de qualidade
- Inspirado no conceito de ***Data profiling***
- Facilitador para **consumidores** e **produtores** de dados
  - Seleção de Fontes
  - Acompanhar evolução das fontes de dados
  - Adaptação de consulta *quality-aware*

Definições Preliminares

- Conjunto de Critérios de Qualidade
  - **$Q = \{Q_1, ..., Q_n\}$**
- Conjunto de Valores dos Critérios de Qualidade
  - **$CQ = \{(Q_1, v_1), ..., (Q_n, v_n)\}$**
- Medida Global de Qualidade
  - **MGQ**
- Perfil de Qualidade
  - **PQ = (CQ, MGQ)**
- Frequência de Atualização do Perfil de Qualidade
  - **λ**

Critérios e Métricas de Qualidade



Completude

- Grau em que os dados de uma fonte são completos para uma determinada tarefa

- É calculada por meio de duas métricas: **Densidade** e **Cobertura**

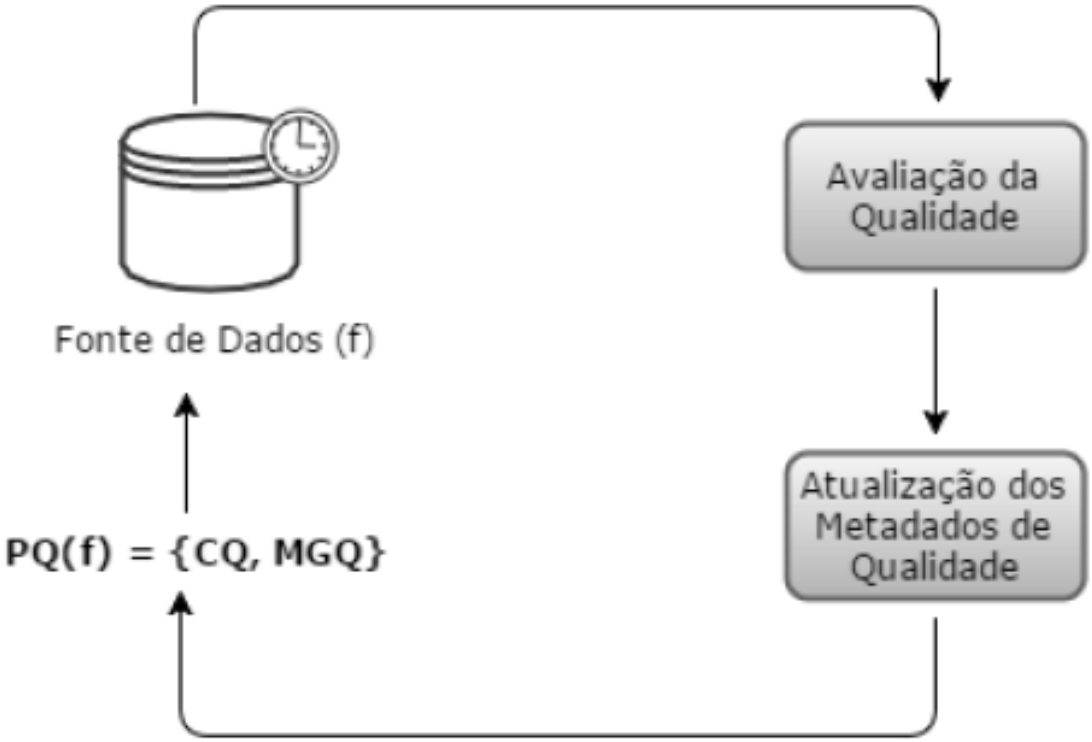
Corretude

- Grau em que os dados de uma fonte estão livres de erro
- Avaliar corretude pode ser uma **tarefa difícil**
  - Avaliação subjetiva
  - Depende do domínio de dados
  - Não existe uma medida que caracterize corretude de maneira geral
- Uso de **regras de validação**

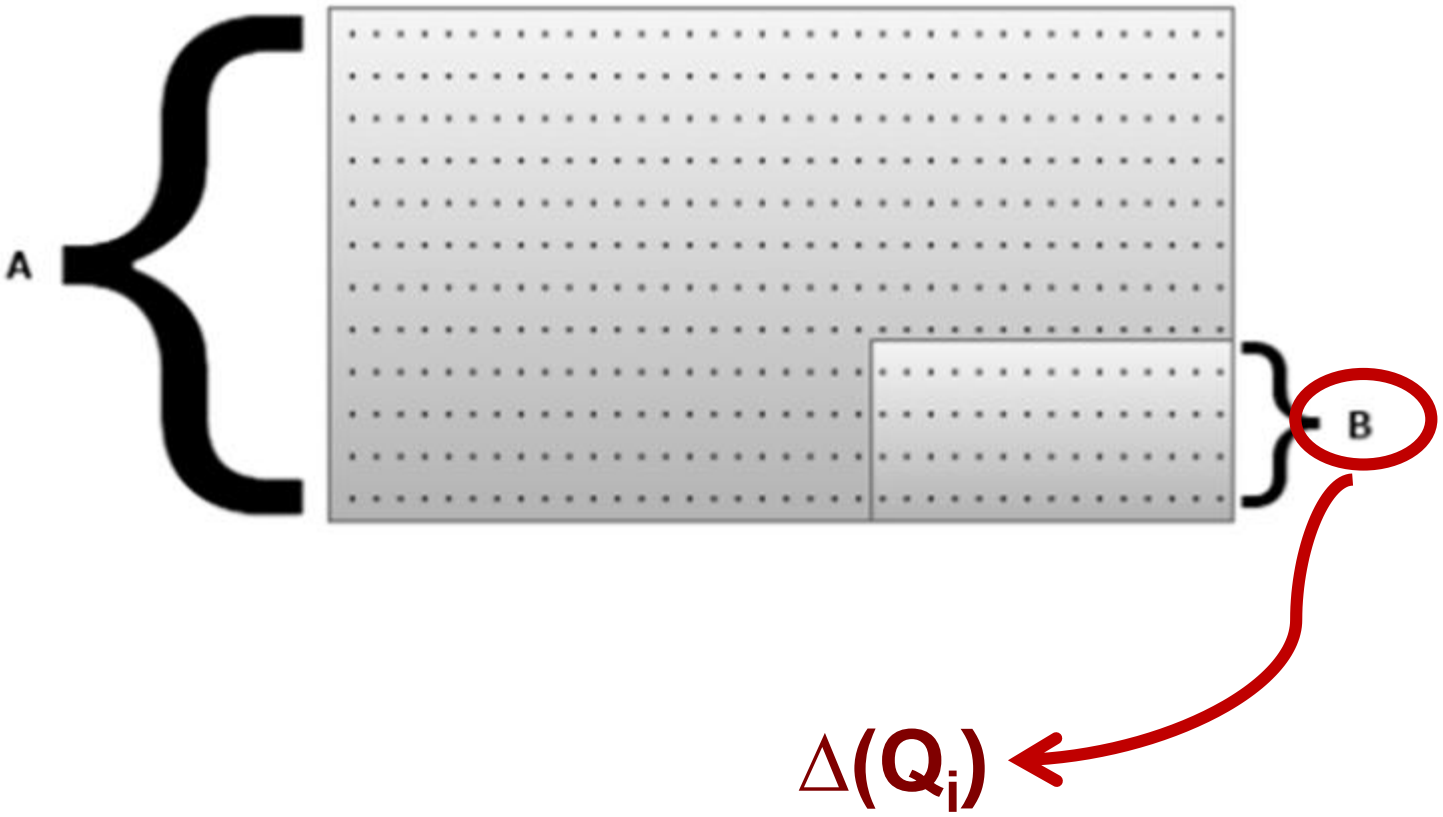
Roteiro

- Definição do Problema
- Perfil de Qualidade
- Geração do Perfil de Qualidade
- Experimentos
- Conclusão

Geração do Perfil de Qualidade



Etapa 1 – Avaliação da Qualidade



Etapa 2 – Atualização dos Metadados de Qualidade

$$Q_i.v_{i_{(new)}} = w * \Delta(q_i) + (1 - w) * (Q_i.v_i)$$

$$MGQ_{(new)} = \sum_{i=1}^{|Q|} w_i * Q_i.v_i$$

Estrutura do Perfil de Qualidade

URL
Última atualiza
Última modifica
Volume de Dad
Critérios de QI
Medida Global

```
{
    "url_fonte": "http://fontededados.com.br/acesso",
    "ultima_atualizacao": "20/05/2016 17:00:01",
    "geracao_perfil": "20/05/2016 17:15:06",
    "volume_dados": 12.395,
    "criterios_qi": [{
         "nome_criterio": "Completude",
         "valor_criterio": 0.89
    }, {
         "nome_criterio": "Corretude",
         "valor_criterio": 0.95
    }],
    "medida_global": 0.92
}
```

fonte de dados.

dos.

liado.

SBBD2016

CIn.ufpe.br

Roteiro

- Definição do Problema
- Perfil de Qualidade
- Geração do Perfil de Qualidade
- Experimentos
- Conclusão

Experimentos

- **Cenário**
  - Domínio Meteorológico
  - Dados provenientes de Estações Meteorológicas
  - Regras de Validação (Baba et al., 2014);
    - Validação de Limites (VL);
    - Validação Lógica (VLG);
    - Validação Limites do Período (VLP);
- **Fontes de Dados**
  - APAC e ITEP
  - Período 01/01/2013 a 31/12/2014
    - 12.255 instâncias (APAC) e 12.265 instâncias (ITEP)

Avaliação Experimental

- Simular o aspecto dinâmico das fontes
  - Implementada a funcionalidade **Carga de Dados**
  - Conjunto de dados divididos em lotes

Lote de Dados					
Fonte de Dados	t_1	t_2	t_3	t_4	**Volume Total**
APAC	3.098	3.040	3.240	2.977	12.255
ITEP	3.086	3.070	3.138	2.971	12.265

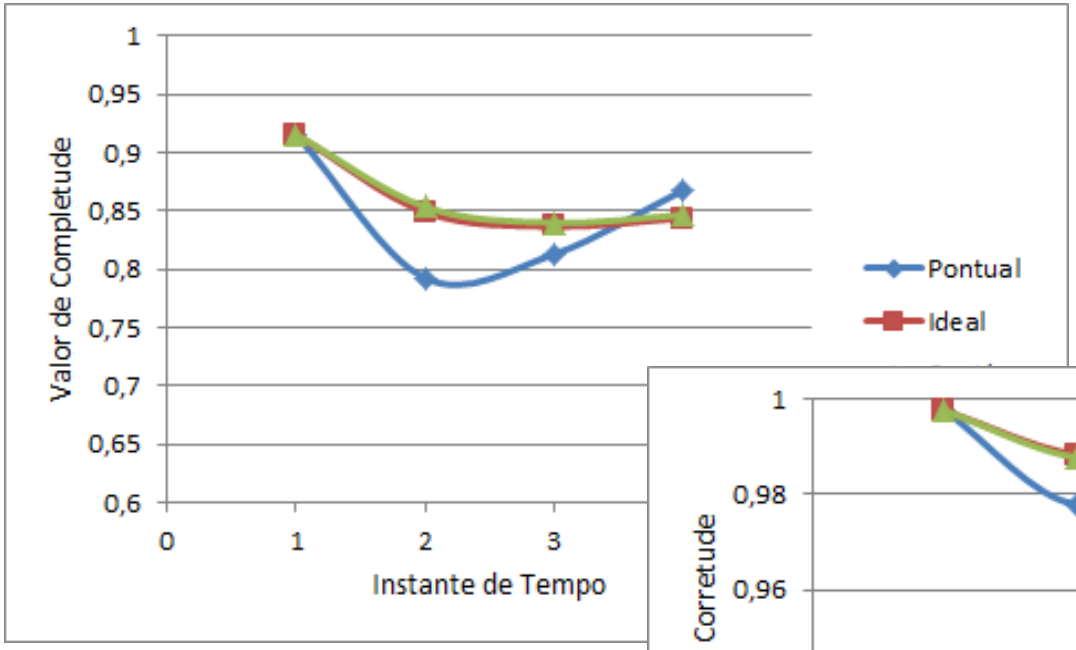
Avaliação Experimental

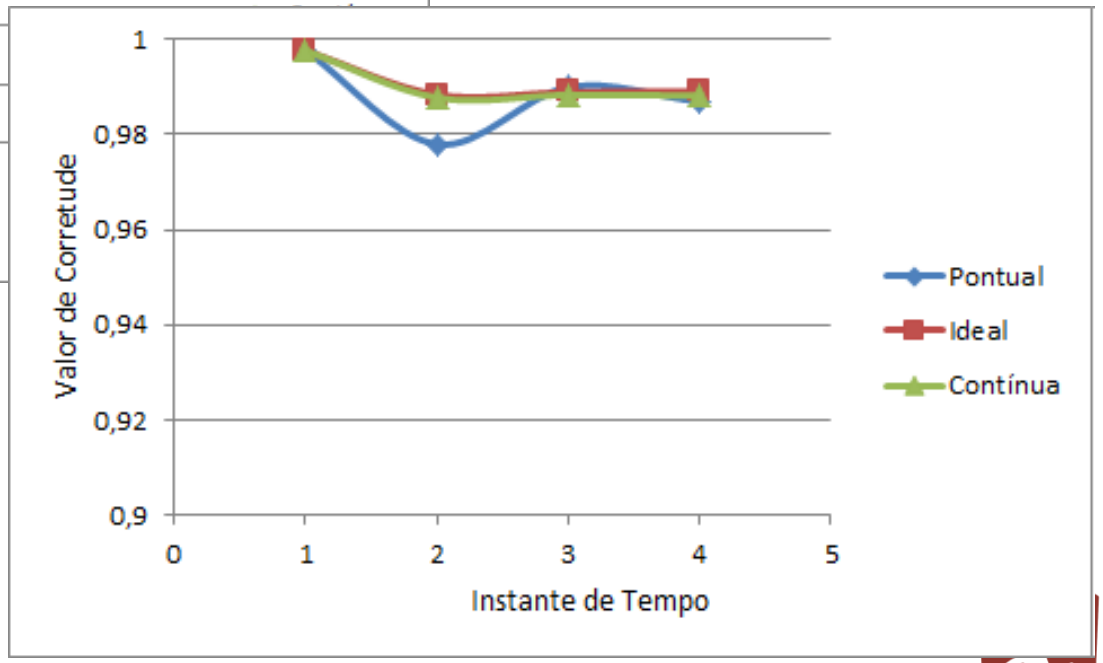
- Para **avaliar a estratégia de avaliação contínua**, implementamos três (3) estratégias para realizar a avaliação da qualidade das fontes
  - **Avaliação Pontual**
  - **Avaliação Ideal**
  - **Avaliação Contínua**

Hipóteses

- H1 – A estratégia contínua é mais eficiente que a pontual, quando comparada com a estratégia ideal
- H2 – Há pouca perda de precisão nos valores calculados pela estratégia contínua, quando comparada com a estratégia ideal
- H3 – A estratégia contínua produz uma redução no tempo de execução da avaliação da qualidade quando comparada com a estratégia ideal

Resultados - APAC



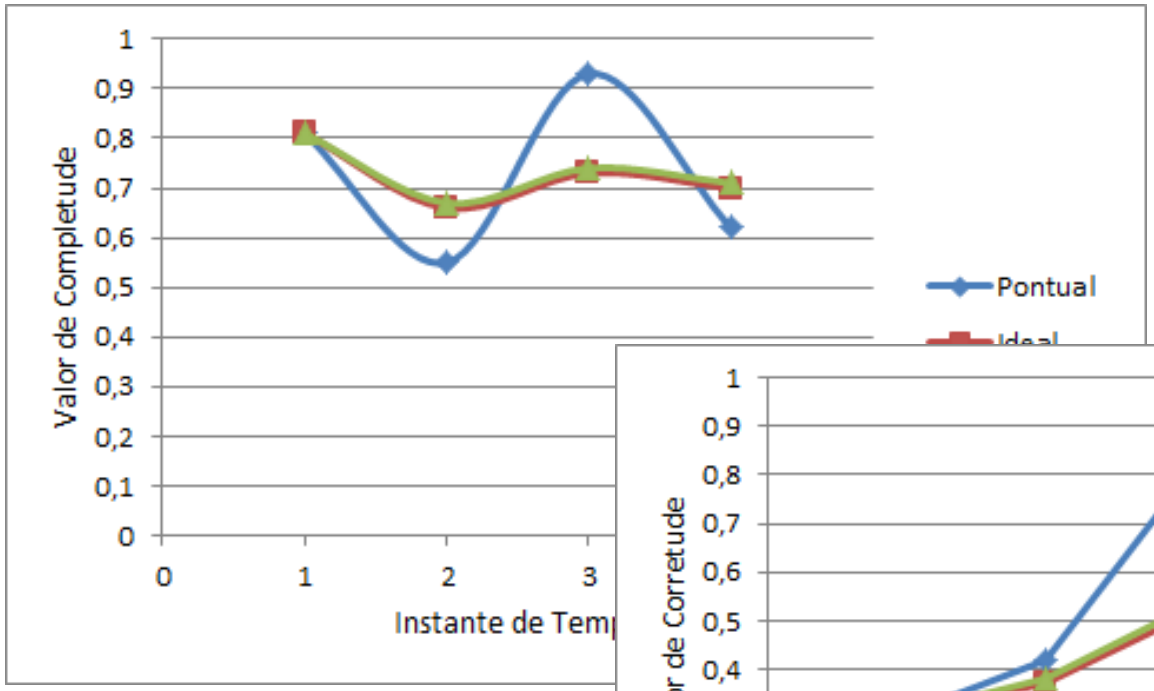


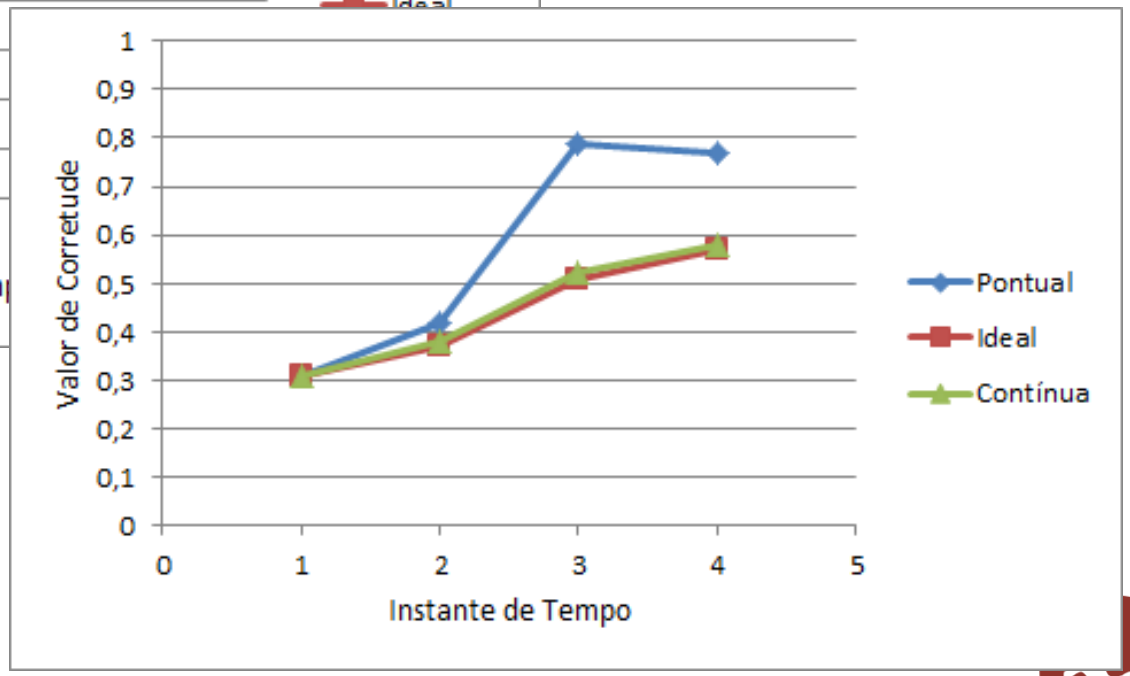
SBBD2016

Centro de Informática UFPE

CIn.ufpe.br

Resultados - ITEP



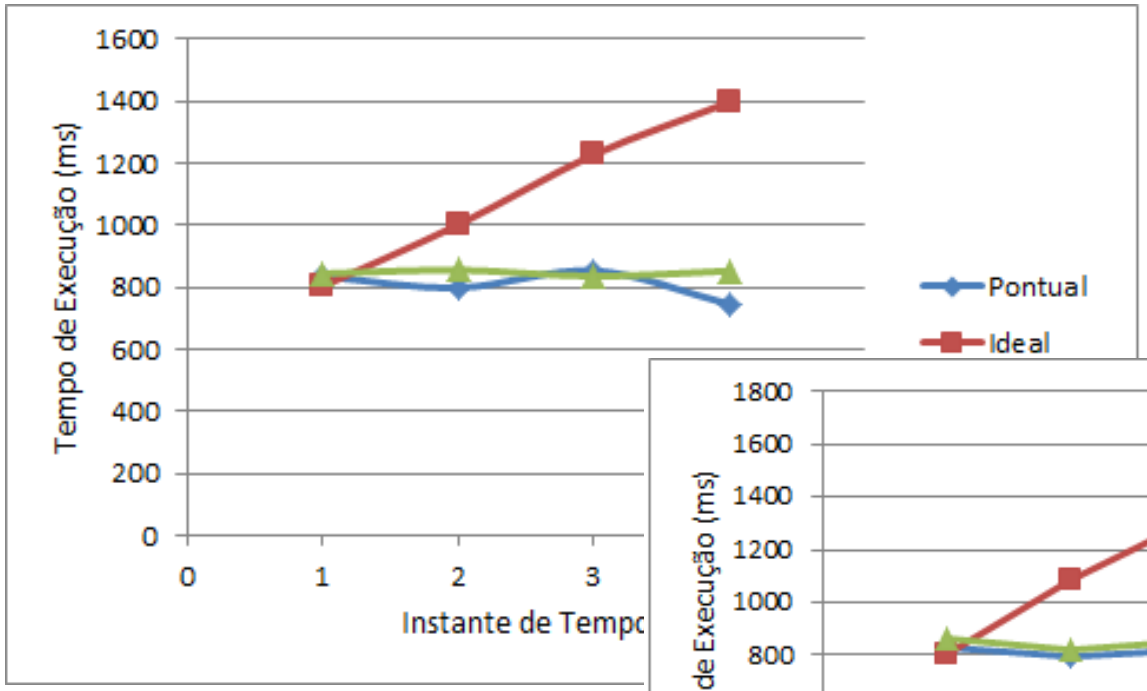


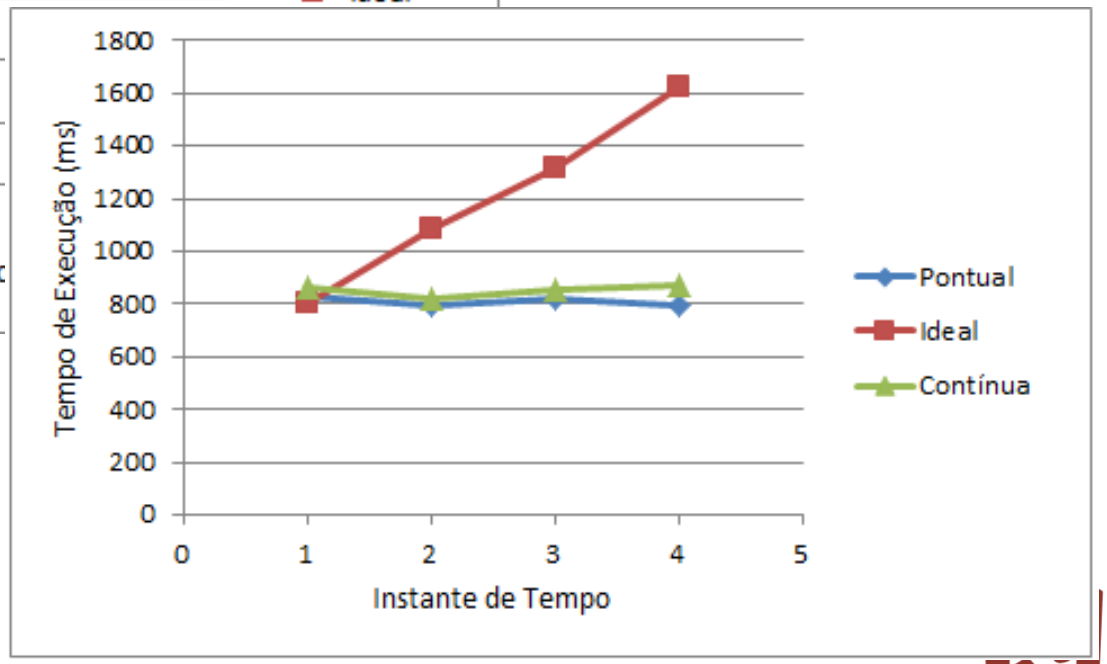
SBBD2016

Centro de Informática UFPE

CIn.ufpe.br

Tempo de Execução





Discussões

- A estratégia Pontual se mostrou ineficiente (H1)
  - **Valores distantes do esperado (ex.: ITEP t_3 +23,39%)**
- A estratégia Contínua alcançou resultados próximos ao esperado (H2)
  - **Valores com média de diferença de +/- 1 a 2% (ex.: ITEP t_3 +1,36%)**
- A estratégia Contínua conseguiu reduzir consideravelmente o tempo de execução da avaliação da qualidade (H3)
  - **No instante t_4, onde o volume de dados é maior, uma redução de -63% (APAC)/ -86% (ITEP)**
- A estratégia Contínua se apresentou como a melhor solução quando avaliado o *trade-off* entre precisão e desempenho

Roteiro

- Definição do Problema
- Perfil de Qualidade
- Geração do Perfil de Qualidade
- Experimentos
- Conclusão

Conclusão

- Proposição da Geração do Perfil de Qualidade
- Avaliação da qualidade considerando o aspecto dinâmico das fontes de dados
- Realização de experimentos com fontes de dados reais
- **Trabalhos Futuros**
  - Ampliar os experimentos com fontes de dados de outros domínios
  - Utilizar critérios relacionados à qualidade do serviço

Geração de um Perfil de Qualidade para Fontes de Dados Dinâmicas

Everaldo Neto, Bernadette Lóscio, Ana Carolina Salgado
(ecsn, bfl, acs)@cin.ufpe.br
Universidade Federal de Pernambuco – UFPE
Centro de Informática - CIn