1. Apresentação

Este documento tem por objetivo tornar pública a síntese de resultados de provas de conceito realizadas nos meses de maio e junho correntes com soluções inscritas na Chamada de Propostas e Audiência Pública SLTI 001/2005.

Como se verá ao longo do texto, as seis provas realizadas permitiram alcançar plenamente os objetivos do processo, tanto na forma da obtenção de indícios de inconsistências e duplicidades em amostra de cadastros sociais, que seriam invisíveis às rotinas atualmente empregadas; quanto no que tange ao refinamento das especificações para o processo de aquisição dos recursos.

Para fornecer a visão global do processo, o documento se organiza em tópicos que tratam, das condições preparatórias à realização das provas, dos resultados obtidos e das conclusões derivadas.

2. Condições Preparatórias à Realização das Provas

a) Objetivo: Em consonância com termos da Chamada de Propostas e Audiência Pública SLTI 001/2005[1], as provas buscaram permitir a análise das condições de uso de metodologias e softwares na identificação probabilística e tratamento de inconsistências em acervos referentes a cadastros sociais, com vistas a confirmar sua aplicabilidade ao ambiente e subsidiar o refinamento de requisitos constantes da Chamada.

b) Escopo: Para permitir efetiva visualização da aplicabilidade dos recursos ao ambiente dos cadastros sociais, as provas foram realizadas sobre amostra de dados referentes a programas de transferência de renda, informações de remuneração derivadas de dados previdenciários e registros de óbitos. Tal contexto foi entendido como subconjunto expressivo da realidade dos cadastros e propício à identificação de evidências de pagamentos indevidos decorrentes de problemas de qualidade de dados (por duplicidade de cadastramento, sub declaração de renda ou falecimento não registrado), úteis para gestores públicos.

c) Segurança de acesso a Programas e Dados/Ambiente de Processamento: Visando manter isolamento do ambiente de cada prova e manutenção de sigilo sobre as informações, optou-se pela utilização do cluster instalado no Departamento de Integração de Sistemas da SLTI/MP em modalidade que impediu acesso à web e à rede local, e o controle de acesso a dispositivos de gravação. Termos de manutenção de sigilo foram assinados por todos os profissionais que tiveram envolvimento direto com o ambiente de testes. O uso do cluster foi decisivo também para analisar uma das condições previstas na Chamada: emprego de computação em grade. Foi disponibilizado ambiente com 20 servidores com 2 processadores *HyperThread* cada, sendo que 4 equipamentos com configurações similares ficaram disponíveis para uso em casos de problemas no hardware originalmente alocado.

[1] Publicada no DOU de 28/1/2005, Seção 3, página 147 e retificada no mesmo veículo e seção, em 9/02/2005 à página 59 e em 18/02/2005 à página 75.

d) Quantitativo de Registros: Para que o processo pudesse ser realizado em período de referência de 5 dias úteis, definiu-se, em acordo com as instituições que apresentaram propostas, que o quantitativo total de registros seria da ordem de 1,5 milhão. Para que a amostra tivesse "semântica" adequada, os dados foram extraídos dos cadastros correspondentes a uma região metropolitana do país. Considerado o quantitativo referente às diversas regiões metropolitanas, foi escolhida a mais próxima de 1,5 milhões: Belo Horizonte.

e) Responsabilidades das Equipes de Governo e das Instituições Participantes: Para tornar o processo o mais equânime possível, não foi feita qualquer instalação de software por parte da equipe de governo, que se limitou a apoiar tecnicamente os testes quanto às características dos equipamentos, entrega dos dados e de sua documentação básica (estrutura, exemplos de registros e estatísticas de preenchimento e consistência).

- A documentação das bases foi enviada previamente às instituições participantes, visando permitir que parte das atividades de customização das ferramentas ao contexto dos testes fosse realizada antes mesmo do início das provas.
- As atividades a cargo das instituições participantes incluíram: instalação da infra-estrutura de software (sistema operacional, software de comunicação no ambiente do cluster e sistema gerenciador de banco de dados), instalação da ferramenta e customização do software às condições de realização das provas, processamento de etapas preparatórias e de pareamento das bases, gravação e apresentação de resultados e, finalmente, a desinstalação de todos os conteúdos armazenados no ambiente do cluster.
- A entrega das bases aos membros das instituições participantes foi realizada pela equipe de governo logo após conclusão da instalação da ferramenta. Para garantir que os conteúdos entregues em cada caso seriam idênticos, chaves de identificação foram geradas para cada arquivo. Os arquivos recebidos pela equipe de governo com os resultados dos processamentos foram também objeto de geração de chaves de identificação e armazenados em ambiente controlado quanto a acesso, para posterior avaliação da eficácia dos resultados.

f) Bases Utilizadas:
   - CADÚNICO: 1.011.952 registros, gerados a partir de extração do Cadastro Único, considerando desnormalização de tabelas referentes à pessoa e ao domicílio;
   - GFIP: 500.100 registros, gerados a partir de extração do Cadastro de Informações Sociais – CNIS e das Guias de Recolhimento Previdenciário e Trabalhista – GFIP, desnormalizadas de forma a representar em único registro as remunerações referentes a 18 meses para cada cidadão/vínculo.

g) Tratamento de Informações de Óbitos: Considerado o prazo escasso para realização das provas, o tratamento de informações de óbitos foi deixado como opção de cada instituição. As instituições que se dispuseram a realizar tal etapa de processamento receberam informações extraídas do SIM - Sistema de Informações de Mortalidade – num total de 13.073 registros, em formato que omitiu toda informação especialmente sigilosa. Como por exemplo, causa da morte.

h) Indicação de Maturidade do Software no Tratamento de Grandes Volumes: Para permitir à equipe de governo, no posicionamento quanto à maturidade das ferramentas no tratamento de grandes volumes de dados, foi definido que além das experiências anteriores de uso do software expressas nas propostas, fosse feito às instituições questionamento quanto ao interesse/disposição em processar, no mesmo período de

referência (5 dias úteis), versão da base GFIP com quantitativo cerca de três vezes superior ao previamente acordado 1.270.338 registros. Tal versão da base, doravante referida como GFIP Expandida, corresponde em verdade ao quantitativo total de registros extraídos para a região metropolitana. A versão de cerca de 500 mil registros foi produzida a partir de extração da amostra original, considerando como filtro as maiores remunerações (e conseqüentemente maiores probabilidades de pagamentos indevidos por sub declaração de renda).

i) Avaliação de Eficácia dos Resultados: a análise sobre a amostra feita comparando os resultados obtidos nas provas e aqueles gerados por simulação das rotinas atualmente empregadas no tratamento de cadastros. A simulação, realizada pela equipe de governo, tomou com critério de pareamento entre registros a ocorrência de coincidência absoluta entre nome da pessoa e da mãe, data e UF de nascimento e entre ao menos um documento de identidade. Condição definida com base em levantamentos realizados pelo Grupo de Trabalho "Integração de Cadastros Sociais" (outubro/2003 a setembro/2004).

j) Instituições Participantes: A seleção das instituições participantes teve como critérios o envio de proposta no prazo da Chamada e o aceite das definições preliminares, constantes dos tópicos anteriores[2]. Das sete propostas apresentadas no prazo, apenas uma não chegou a ser testada, tendo em vista o não posicionamento da instituição proponente quanto às condições de realização dos testes.

2 Síntese dos Resultados

2.1 Cronograma e Custos de Realização

As seis provas foram realizadas entre os dias 2 de maio e 17 de junho corrente, envolveram 9 instituições[3] e 21 profissionais. A SLTI/MP responsabilizou-se por despesas de deslocamento e hospedagem para sete profissionais que atenderam às condições definidas na Chamada de Propostas e Audiência Pública, com despesa total da ordem de R$ 10,5 mil.

Tabela 1 – Período de Realização das Provas de Conceito

Instituição (ões)	Período	Dias de Uso do Ambiente
FITEC	02/05 a 09/05/2005	08
GODIGITAL	12/05 a 20/05/2005	07
UNICAMP/IPS	23/05 a 30/05/2005	08
UFMG/UFAM	30/05 a 03/06/2005	05
LACTEC/UFPR	06/06 a 10/06/2005	05
SOFTEK	13/06 a 17/06/2005	05
Visão Geral		
Mínimo: 05 dias	Médio: 6,5 dias	Máximo: 08 dias

2.2 Etapas de Processamento

Consideradas as especificidades de cada ferramenta foram observadas diferenças de encadeamento e nomenclatura das etapas. Todas as soluções, no entanto, apresentaram passos referentes à carga dos dados e produção de estatísticas, preparação das bases para

[2] Com exceção do item (g), que foi apresentado às instituições, apenas durante a realização das provas.
[3] 3 propostas foram apresentadas por mais de uma instituição.

identificação de pares e pareamento. Para simplificar a apresentação dos resultados, foi adotada denominação geral, conforme tabela a seguir:

Tabela – Estatística de Processamento das Etapas

Etapas	Iniciados	Concluídos com Sucesso
1. Instalação do ambiente e ferramenta	6 em 6	6 em 6
2. Carga e Estatísticas das Bases	6 em 6	6 em 6
3. Limpeza/Padronização das Bases	6 em 6	6 em 6
4. Deduplicação CADÚNICO	3 em 6	3 em 6
5. Pareamento CADÚNICO/GFIP	3 em 6	3 em 6
6. Gravação de Resultados	3 em 6	3 em 6
7. Desinstalação dos Softwares e Bases	6 em 6	6 em 6
Informações Adicionais sobre Etapas Processadas		
2 Provas deram tratamento específico a Endereço 1 Prova utilizou solução incompatível com Cluster	1 Prova utilizou GFIP Expandida 1 Prova processou base de Óbitos	

2.3 Ocorrências Especiais

Durante a realização das provas ocorreram imprevistos que, em diferentes medidas, causaram impacto ao processo. Merecem registro, as seguintes ocorrências especiais:

- Em duas situações o acesso físico às instalações foi interrompido, sendo que em uma delas a causa foi a necessidade de evacuação do prédio, por questões de segurança, e em outra ocorreram problemas na rede elétrica. Cada situação gerou atraso de cerca de 4 horas à prova realizada no período correspondente, que foram compensados por extensão do tempo de uso do ambiente após o expediente normal;
- Houve uma ocorrência de atraso na entrega de bases por parte da equipe de governo, superior a 24 horas após a conclusão da instalação da ferramenta. Tal atraso referente à amostra do SIM[4] teve como desdobramento a impossibilidade de que uma das instituições que demonstraram interesse em tratar dados de óbitos, não conseguisse efetivar tal tratamento;
- Embora tenham sido utilizados nos testes apenas microcomputadores novos, em todas as provas, com exceção de uma, ocorreram problemas no hardware. Em geral estes problemas causaram apenas atraso no cronograma de processamento, e, foram resolvidos com a alocação de equipamentos reserva ao cluster. No entanto, em uma das provas os problemas de hardware se associaram à falta de conectividade entre o ambiente de processamento e as estações clientes, determinando a impossibilidade de conclusão dos trabalhos no prazo estipulado;
- Nos três casos em que as provas não atingiram a etapa final de processamento, a equipe de governo ofereceu condições para estender o prazo. Uma das instituições participantes não teve interesse em usar a extensão de prazo. As demais utilizaram prazos adicionais, mas ainda assim, cancelamentos nas versões das ferramentas instaladas impediram a obtenção de pares.

[4] Como houve dificuldade para obter a base SIM diretamente com o Ministério da Saúde, foi utilizado um subconjunto de dados referentes a Belo Horizonte, ano de 2003, gentilmente enviados pela Secretaria Municipal de Saúde. Mesmo com esta decisão, ocorreu atraso na geração da base simplificação.

2.4 Desempenho das Ferramentas

Os tempos de processamento obtidos em cada prova não são diretamente comparáveis, tanto pelo fato de que o número de equipamentos utilizados no cluster variou conforme decisão das instituições participantes, quanto por que o escopo das etapas processadas em dada prova, em geral não apresentava correspondência absoluta com as demais. Neste sentido, opta-se por apenas ilustrar os tempos de processamento verificados, o que pode ser visto no quadro a seguir, que apresenta dois grupos de resultados.

Tabela 3 –Resultados Expressivos quanto a Tempo de Processamento

Processamento	Prova 1 01 equipamento	Prova 2 04 equipamentos
1. Limpeza das bases	-	00:24:00
2. Deduplicação CADÚNICO	03:00:00	00:10:14
3. Deduplicação GFIP	02:30 :00	00:15:25
4. Pareamento CADÚNICO/GFIP	04:00:00	-
5. Pareamento CADÚNICO/GFIP Expandida	-	00:24:02
6. Pareamento SIM/GFIP	-	00:15:25

Deve-se destacar ainda que em algumas das provas, as instituições participantes repetiram processamento de dada etapa com número crescente de servidores, de forma a estabelecer comparativo de performances decorrentes de processamento paralelo. Este tipo de procedimento, que não foi imposto, é ilustrado no gráfico a seguir. Note-se que para volume pequeno de registros a opção por processamento paralelo não apresenta vantagem expressiva, uma vez que o tempo de execução de sub tarefas de distribuição e controle do processamento supera os ganhos do paralelismo. No entanto, à medida que o volume de registros aumenta, é possível observar sensível melhoria no tempo total de processamento.

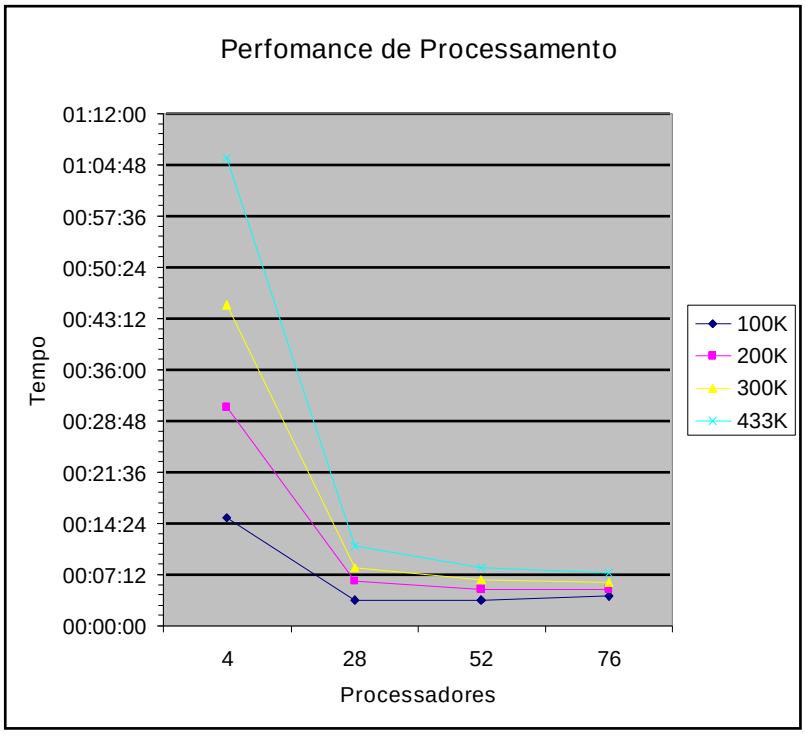

Figura 1 –Variação de Performance por nº de Processadores/Quantidade de Registros.

2.5 Pares Obtidos

A análise de eficácia das ferramentas foi feita com base em comparação do quantitativo de pares obtidos durante as provas e a simulação de processamento das rotinas atuais sobre a amostra (conforme definido no item 2.i).

A comparação entre os resultados desta simulação e os melhores resultados obtidos durante as provas encontram-se na tabela a seguir:

Tabela 4 –Resultados Expressivos quanto a Pares Obtidos

Bases	Melhor Resultado Obtido nas Provas	Simulação de Rotinas Atuais	%
DUPLICIDADE CADÚNICO	66.430	52.733	125,97
CADÚNICO/GFIP	36.546	25.862	141,31
CADÚNICO/GFIP Expandida	98.090	37.425	262,09
SIM/GFIP	156	1	15.600

3 Conclusões

O contato direto com as metodologias e softwares testados em ambiente similar ao dos cadastros sociais foi extremamente útil para refinar a elaboração de especificação dos requisitos do projeto. Vários elementos da especificação preliminar constante da Chamada de Propostas e Audiência Pública SLTI 0001/2005 puderam ser revistos, destacando-se:

- A importância de combinar algoritmos de busca por similaridade de registros (além do pareamento probabilístico, outros métodos devem ser agregados para que se tenha resultados de qualidade);
- A necessidade de ter o processo de identificação e tratamento de inconsistências gerenciado. Neste sentido, os princípios de Gestão de Qualidade de Dados se mostraram especialmente aplicáveis ao problema.

Por outro lado, e mesmo ressalvadas que as condições do experimento não são evidentemente idênticas às do ambiente real, há elementos para antever com relativa certeza que a extensão do uso de ferramentas similares às utilizadas nos testes aos ambientes de tratamento de informações sociais contribuirá expressivamente na melhoria da qualidade dos dados. Mesmo considerando que processos de auditoria precisem ser associados a grande parte das situações apontadas, é racional imaginar que a extensão de soluções similares às analisadas ao conjunto do acervo poderia resultar em expressiva melhoria na gestão de benefícios sociais.

Esta compreensão tem sido debatida com os órgãos intervenientes ao projeto, estabelecendo condições objetivas para iniciar procedimentos referentes à contratação dos recursos.