

Tratamento e armazenamento de informações de bases de dados heterogêneas: com relativa distância entre elas

Maurilio Aparecido
Benevento ¹

Dr. Carlos Alberto
Chaves ^{2,2}

Dr. Eduardo
Hidenori Enari ^{3,3}

Álvaro Azevedo Cardoso,
PhD ^{4,4}

maurilio@mbtwo.com.br

carlos.chaves@unitau.br

enari@unitau.br

alvaroacardoso@yahoo.com.br

1 Universidade de Taubaté (UNITAU), Curso de Engenharia Mecânica – Produção, Taubaté, SP, Brasil.

2 Universidade de Taubaté (UNITAU), Curso de Engenharia Mecânica – Produção, Taubaté, SP, Brasil.

3 Universidade de Taubaté (UNITAU), Curso de Engenharia Mecânica – Produção, Taubaté, SP, Brasil.

4 Universidade de Taubaté (UNITAU), Curso de Engenharia Mecânica – Produção, Taubaté, SP, Brasil.

RESUMO

Este artigo tem a finalidade de contribuir com um case de integração de informações coletadas de bases de dados heterogêneas localizadas em locais espalhados pelo Brasil observando os aspectos da segurança da informação. Para concentrar tais informações são necessárias a utilização de ferramentas de extração de dados e posterior normalização das informações para a conformidade dos dados gerenciais. As informações a disposição em qualquer momento, principalmente após o fechamento dos faturamentos das filiais distantes é sem dúvida o desejo de qualquer organização, sendo exatamente o foco deste trabalho, viabilizar todas as informações possíveis de forma integrada. Ter filiais trabalhando com base de dados distintas é o retrato comum de muitas empresas, por inúmeras razões isto não tem uma resolução trivial, entretanto, nesta contribuição veremos uma maneira que facilita a concentração das bases de dados de tal maneira que qualquer gerente ou executivo obtém informações no dia seguinte ao faturamento de qualquer lugar utilizando a internet.

Palavras-Chave: Integração de dados, Extração de dados, Normalização de dados.

1. Introdução

Esta contribuição tem a finalidade de demonstrar como é possível criar um único repositório de dados de diversos bancos de dados distintos e em localidades distantes com observância às questões de segurança da informação consideradas na ISO IEC 27001:2005.

A utilização de conceitos básicos de tecnologia e de BI “Business Intelligence”, são determinantes para o fechamento deste trabalho, ocorrido numa empresa de grande porte nacional com algumas filiais espalhadas pelo território brasileiro.

Nós vivemos numa época de muita informação e pouca cultura no tratamento dela, isto nos remete a uma vertente de pensamento que é muito clara: “A construção do conhecimento é a junção de fragmentos e ligações que adquirimos ou observamos com o passar do tempo”.

Diante disto, existe apenas uma direção correta para olharmos, - o passado -, a história, o início das construções e o estudo de suas permanências para enfim compreender qual seria a melhor opção na resolução de problemas.

Ir ao passado justifica o porquê deste trabalho e esclarece a situação que muitas empresas brasileiras estão vivendo no seu dia-a-dia de operação.

1.1 Histórico

Como breve introdução sobre o assunto de sistemas operacionais, temos a colocação inicial dos Professores da UFRGS sobre o assunto “O sistema operacional é uma camada de software colocado entre o hardware e os programas que executam tarefas para os usuários”, (OLIVEIRA et al, 2001).

A história nos mostra que existiam muitas pessoas empenhadas em criar um sistema operacional. O primeiro sistema operacional CTSS (Compatible Time-Sharing System), criado dentro do MIT (Massachusetts Institute of Technology), por (CORBATÓ, 1961), foi o precursor do projeto MAC proposto por (LICKLIDER, 1962), resultando no projeto MULTICS (MULTIplexed Information and Computing Service), este era o nome idealizado para o segundo sistema operacional.

Estes projetos foram muito bem visto por órgãos governamentais e fabricantes de computadores de grande porte e empresas do setor de tecnologia, pois havia uma enorme expectativa acadêmica e evidentemente interesses de outras pessoas e organizações. Isto culminou em ajudas de diversos setores, tais como a ARPA (Advanced Research Projects Agency), agência subordinada ao Departamento de Defesa dos EUA, contribuindo com 2 milhões de dólares por ano, durante 8 anos, especificamente para o desenvolvimento do MULTICS. No mesmo período, o Bell Labs e a GE (General Eletric), contribuíram de forma semelhante.

Foi uma contribuição fundamental para o desenvolvimento do MULTICS. Já em 1969 ele foi disponibilizado para uso, tendo como principais clientes a Força Área Americana, a General Motors e a Ford. O projeto foi tão bem sucedido que a última cópia funcional deste sistema operacional foi desativada no ano de 2000, no Quartel General do Comando Marítimo Canadense.

A questão ainda não estava totalmente resolvida, pois o MULTICS não tinha algumas funcionalidades importantes, tais como time-sharing, portabilidade, etc. Ainda esperava-se um sistema operacional com tais características, e não demorou muito para que a idéia de dois programadores do Bell Labs (THOMPSON e RITCHIE, 1969) chamada de UNICS

(UNiplexed Information and Computing Service), nascesse de um projeto pessoal, culminando no UNIX (nome sugerido por um amigo chamado Brian Kernighan).

Este foi o esforço definitivo para todos os sonhos. Os fragmentos estavam sendo ligados e consolidados em um conhecimento que mudaria toda a história da computação, de centralizada para distribuída, da conectividade de grande porte para microinformática, das redes de computadores, de outros sistemas operacionais pagos e não pagos, de linguagens de programação de alto nível e diversos bancos de dados.

1.2 O resultado da criação do Sistema Operacional UNIX

Toda essa abertura de possibilidades difundiu a criação de inúmeras versões de sistemas operacionais baseadas no UNIX, e outras diversas versões da Linguagem C, criada também por (THOMPSON e RITCHIE, 1969).

Várias versões de UNIX foram produzidas. Muitas empresas passaram a vender máquinas projetadas para o uso com o UNIX, dentre elas a SUN Microsystem, a SGI, a Hewlett-Packard, a NCR e a IBM. Em paralelo, na Universidade de Berkeley, um trabalho constante introduziu melhorias e versatilidade. Várias versões foram desenvolvidas pelo Bell Labs, inclusive a famosa **System V**, que criou um estilo seguido por vários sistemas atuais. Outra importante versão, também muito utilizada, é a BSD (Berkeley Software Distribution), uma linha Unix, desenvolvida em Berkeley (FILHO, 2006).

Naturalmente muitos outros sistemas foram criados e comercializados durante toda a década de 1970-2000. Como era de se esperar muitas empresas instaladas no Brasil de capital nacional ou estrangeiro, se expandiram pelo território nacional, usando as tecnologias existentes e variadas.

Em 1978-1980 surgiram os primeiros computadores pessoais e com eles o processamento local com um sistema operacional portátil para microcomputador, mas não foi somente isto, houve a possibilidade de ligação em rede destas máquinas, criando a saída de empresas de grande e médio porte dos mainframes IBM, surgindo assim o conceito de downsizing em tecnologia.

Na década de 90 tivemos o surgimento do conceito de Tecnologia Web e as possibilidades de negócios utilizando os recursos da rede mundial, mas não no Brasil, pois comercialmente isto aconteceu somente em 1995. Com esta abertura da Tecnologia Web muitos negócios e sistemas que eram utilizados localmente foram expandidos para diversas empresas utilizando a internet.

É importante lembrar, que a abertura de mercado ocorrida em 1989-1990, trouxe ao mercado nacional uma grande parte destas complexidades de tecnologia.

Em relato da linha do tempo é importante para representar claramente a evolução das arquiteturas de sistemas e de computadores, (Figura 1) e (Figura 2).

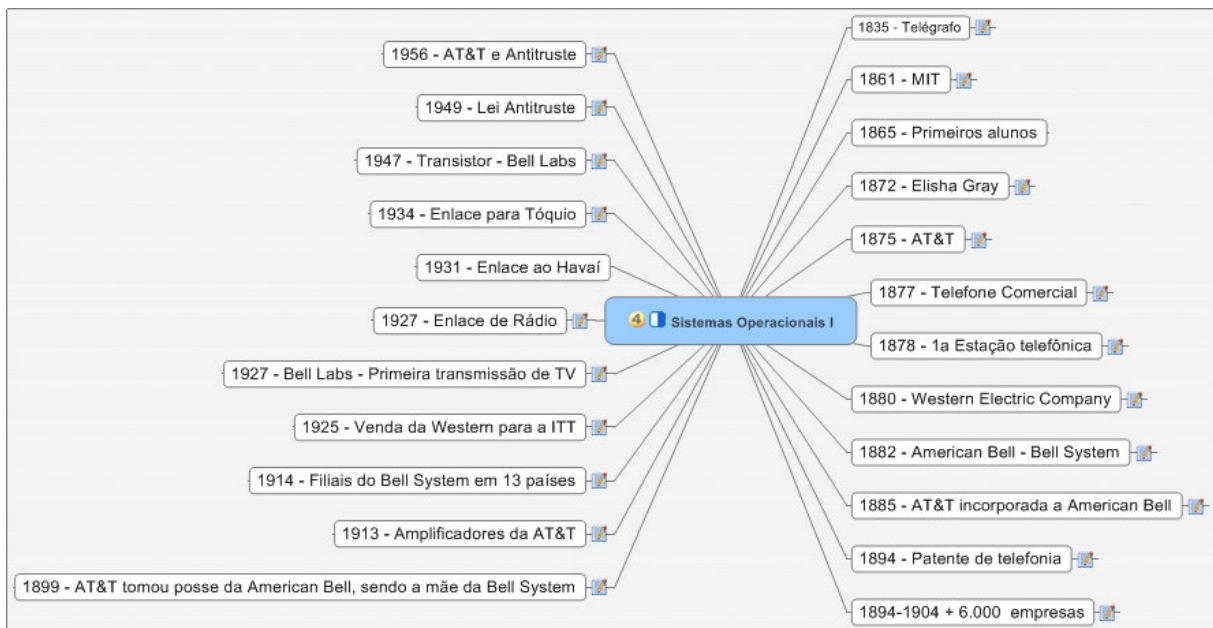


Figura 1 - Mapa da linha do tempo 1 apresentada acima – Elaborado pelo autor.

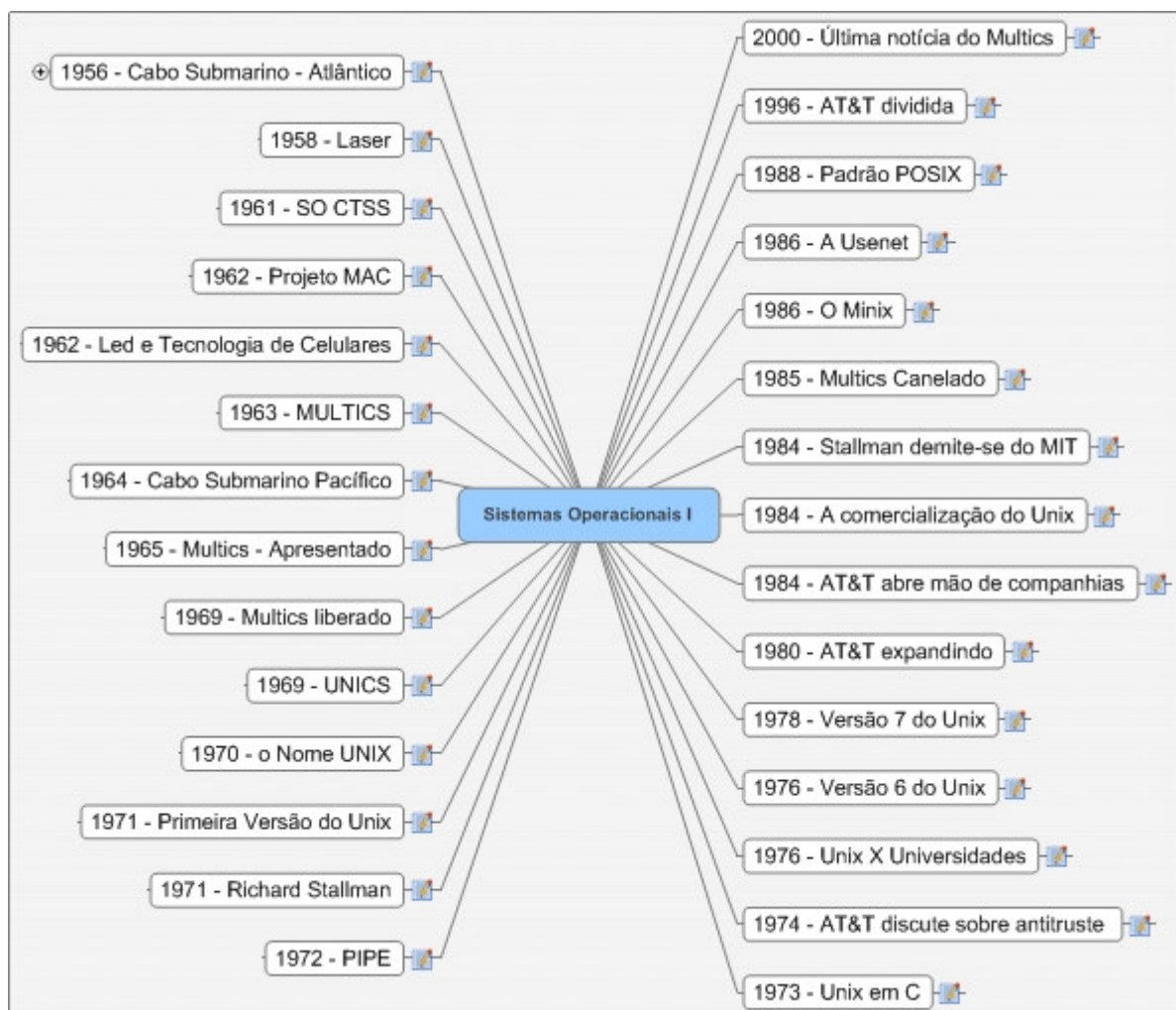


Figura 2 - Mapa da linha do tempo 2 apresentada acima – Elaborado pelo autor

1.3 O problema dos “Legados”

É evidente que o advento dos sistemas operacionais, impulsionaram a comercialização de milhares de produtos diferentes, para atender demandas de diferentes clientes.

É comum encontrarmos empresas com inúmeros sistemas operacionais, sistemas de contas a pagar, sistemas contábeis, sistemas de compras, sistemas de estoque e de faturamento, todos diferentes para cada filial da empresa. Adotados normalmente pela dificuldade de suporte da matriz ou pelo tempo de resposta para atender algumas urgências locais. Enfim, este é o começo do legado de muitas empresas, caracterizando um dos maiores problemas a serem resolvidos.

O saldo disto é o crescimento organizacional, com inúmeras complexidades tecnológicas, espalhadas pela empresa. De tal forma que para gerar um relatório é necessário ler muitas bases de dados e outros tipos de sistemas de arquivos, e ao final tratar em um programa como planilha eletrônica. O que torna o processamento muito mais lento e em alguns casos causando lentidão em toda a rede.

A linha do tempo demonstrada anteriormente deixa muito claro a dificuldade que uma organização enfrenta ao adotar tecnologias para o seu parque de TI, não sendo nunca um trabalho trivial, isto considerando somente o item Sistema Operacional.

Abaixo uma demonstração do trabalho com sistemas legados, pois eles são a principal razão da consolidação de informações em um único repositório, (Figura 3).

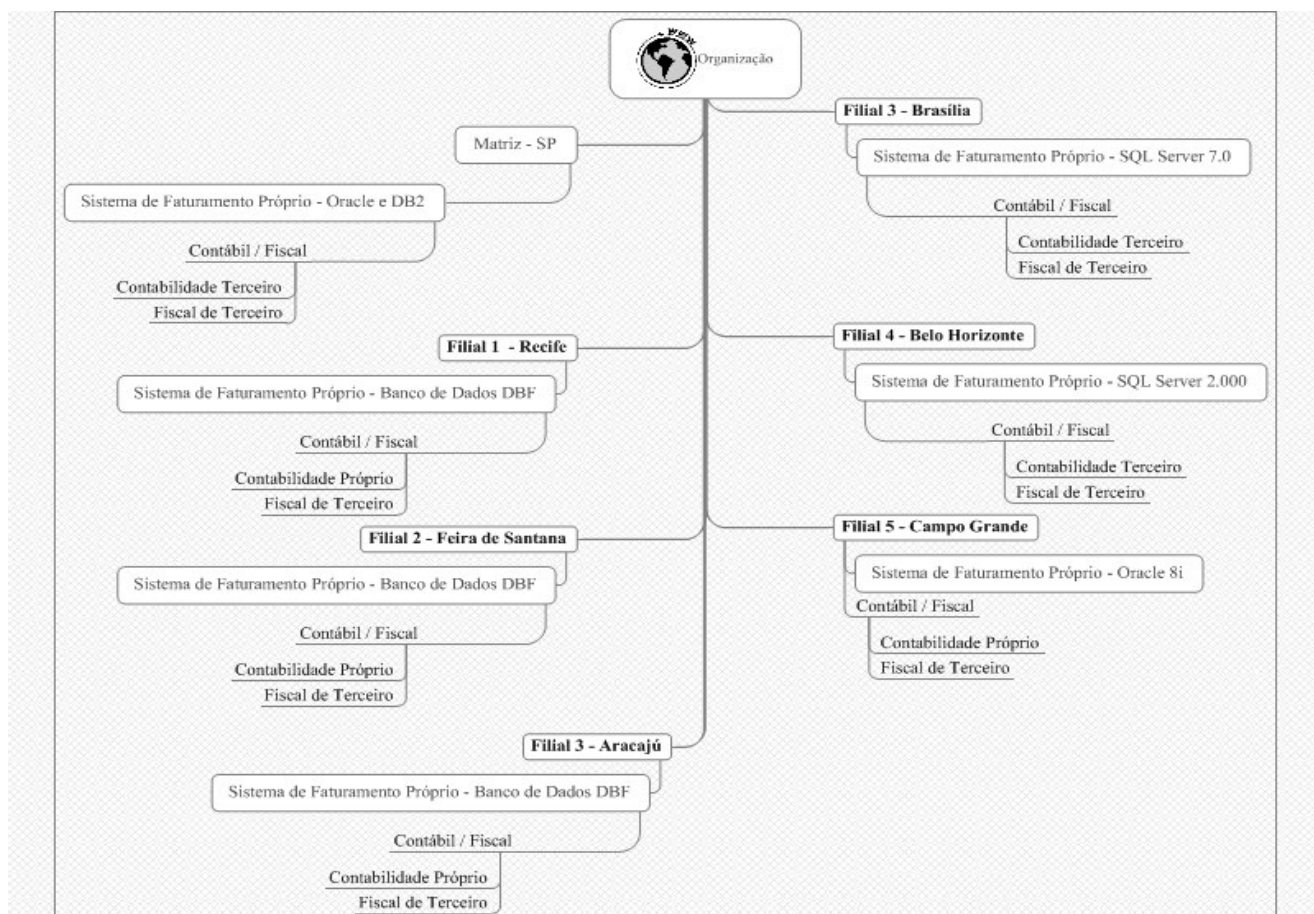


Figura 3 - Demonstração do legado existente na organização – Elaborado pelo autor

2. Desenvolvimento do Projeto

O quadro demonstrado, figura exatamente a realidade de muitas empresas nacionais, do primeiro, segundo e terceiro setor.

A escolha de um modelo de resolução de problemas pode não ser adequada nesta problemática, visto que existem muitas complexidades e uma delas pode não ser compatível com a solução imaginada.

O recurso mais óbvio para este problema é adotar ferramentas de extração de dados ou “Datamining”, seguindo o modelo de primeiro estruturar um Data Warehouse, para receber todos os dados e depois populá-lo com as informações das filiais, (Figura 4).

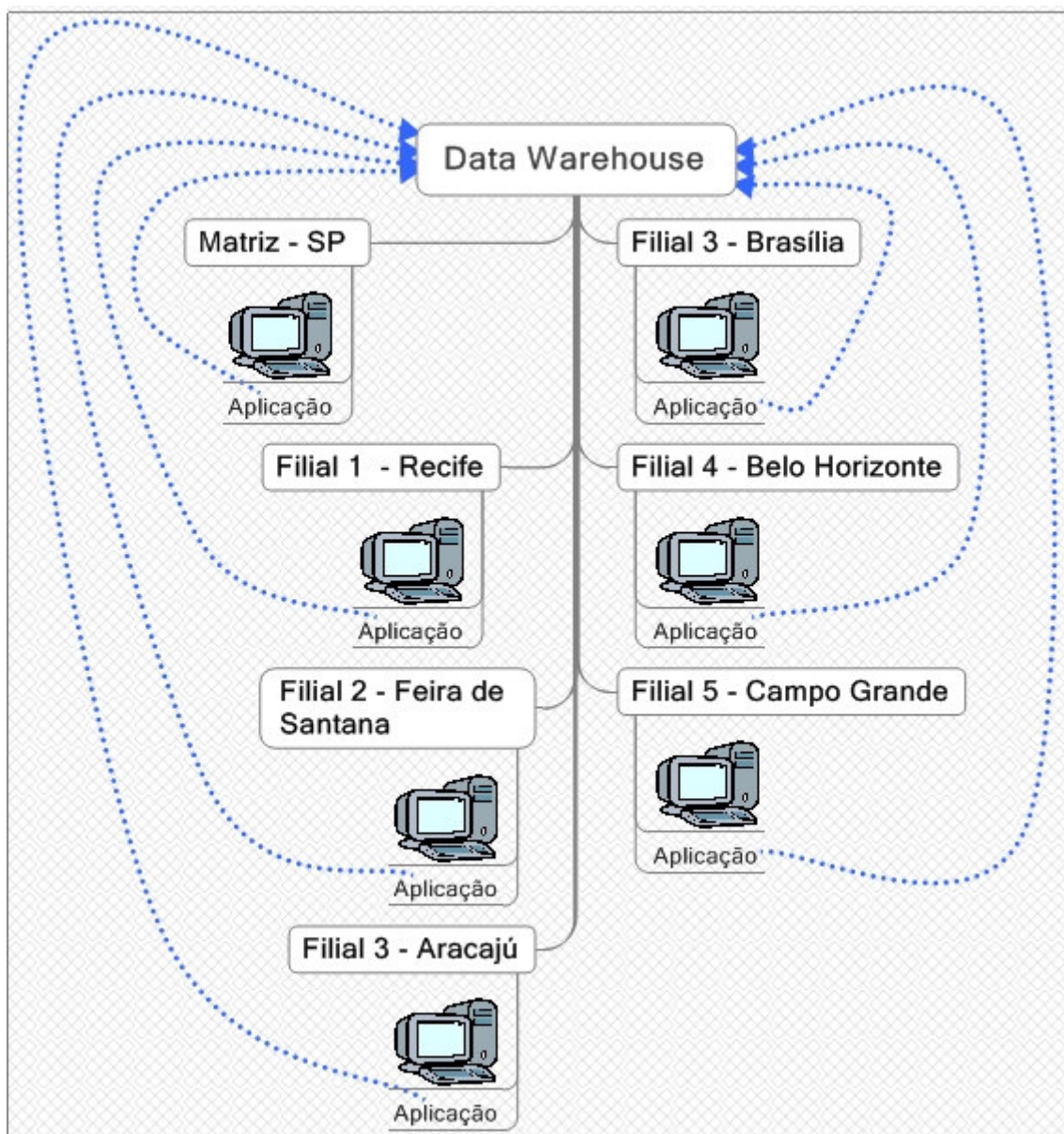


Figura 4 - Modelo da extração de informações para o DW – Elaborado pelo autor

2.1 Integração de Dados

A integração aqui trata da mineração de dados feita em todas as formas de bases de dados das filiais, envolvendo qualquer meio de arquivo onde possa ter informação para ser cruzada e armazenada em um Data Warehouse centralizado na matriz.

Busca-se neste trabalho de integração a normalização dos dados das bases de dados heterogêneas, armazenando-as conforme o padrão adotado no escopo do projeto.

Ao adotar qualquer meio de integração sempre surge a preocupação de como resolver a questão de bases de dados distintas. Não poderia ser diferente, visto que a parte fundamental do trabalho de integração é justamente entender o funcionamento de todos os bancos de dados envolvidos, para viabilizar os acessos e medir os riscos, e viabilizar um meio adequado para a transmissão de dados entre as filiais.

2.2 Problemas comuns

- a) Redigitação de diversas movimentações das filiais para o sistema da matriz;
- b) diversos atrasos no fechamento do dia;
- c) inconsistências nos dados após consolidação na matriz;
- d) atrasos nos envios das informações por indisponibilidades dos sistemas das filiais;
- e) desenvolvimentos paralelos para normalizar as informações antes de inseri-las;
- f) falta de informações na matriz de cadastros novos nas filiais;
- g) Demora excessiva para alterações de informações.

Este quadro poderia ser explorado de forma detalhada em um documento mais extenso; mesmo assim os itens acima demonstram exatamente a extensão dos problemas causados pelos sistemas legados.

Inevitavelmente cada item referido fere o tripé fundamental da informação:

Confidencialidade, Integridade e Disponibilidade. Assim a empresa que vive desta forma não confia na informação disponibilizada, podendo afirmar que “uma informação ruim, é melhor não te-la”.

2.3 Modelo inicial para integração

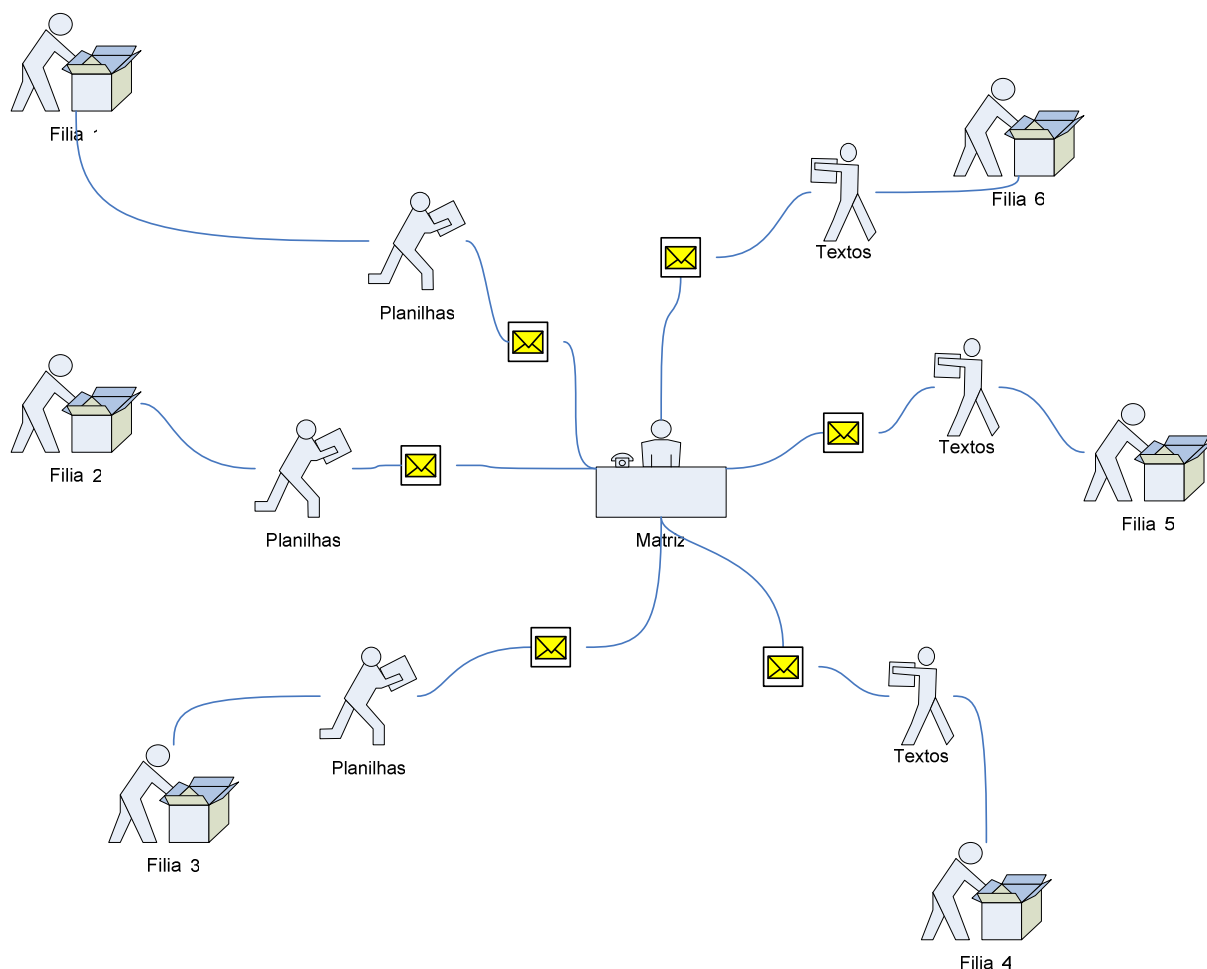


Figura 5 - Modelo inicial para integração dos dados – Elaborado pelo autor

O meio de transmissão de dados para o envio de arquivos neste modelo da (Figura 5) geralmente é utilizado o email, pois as informações trafegadas nos arquivos são consolidadas, portanto, não levam os detalhes do dia-a-dia. Este ponto é de fundamental importância para entender o objetivo de uma integração, pois com a criação de um repositório é possível armazenar detalhes das movimentações do dia-a-dia de cada filial.

2.4 Trabalhando com FTP

Como início da integração das informações temos a utilização do meio FTP (File Transfer Protocol) como meio de armazenamento das bases de dados, textos e planilhas em um ponto comum à todas as filiais. Trata-se de um diretório onde todos os operadores das filiais têm permissão para gravar as movimentações de faturamento e fechamentos, (Figura 6).

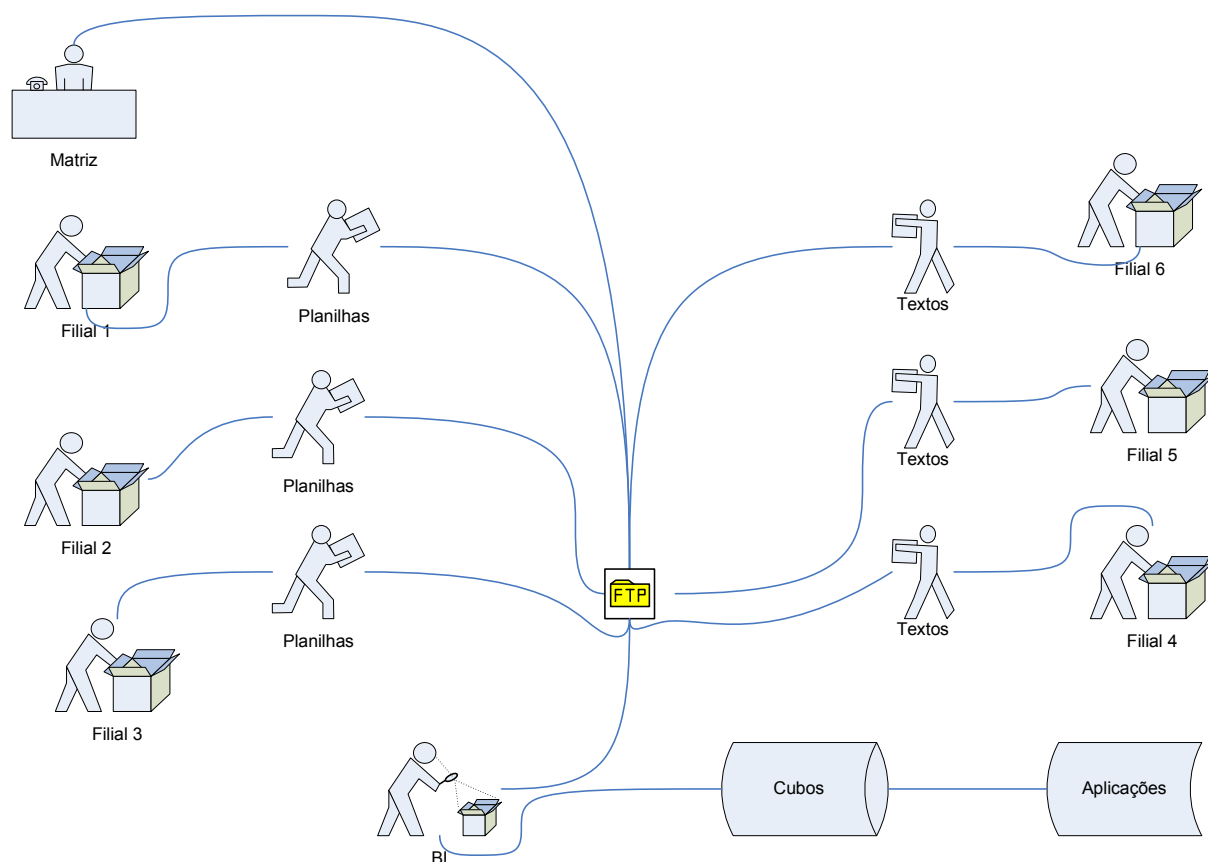


Figura 6 - Gravação dos arquivos em FTP – Elaborado pelo autor

O uso de FTP é mais adequado do que o email, utilizando o email tem-se limitações que poderiam atrapalhar as cargas diárias, pois alguns arquivos podem não chegar no destino, ou seja, na matriz. Por exemplo, anexos poderiam ser barrados por programas de email ou proxy implantados em algumas filiais. Também não seria mais possível usar o EMAIL, pois a integração de dados visa trazer os detalhes das movimentações, representando arquivos muito mais pesados.

A criação de um local para concentrar arquivos de todas as filiais, com nomenclatura correta para cada rotina e dia do envio, torna-se a melhor solução inicialmente. Cada arquivo depositado no diretório de FTP, inserido pelos operadores das filiais após o encerramento do faturamento do dia, deveria basicamente ser composto por:

- a) Arquivo de vendas;
- b) Cadastro de Clientes;
- c) Produtos;
- d) Fornecedores;
- e) Estoques;
- f) Transferências;
- g) Compras;
- h) Contas a pagar;

i) Contas a receber.

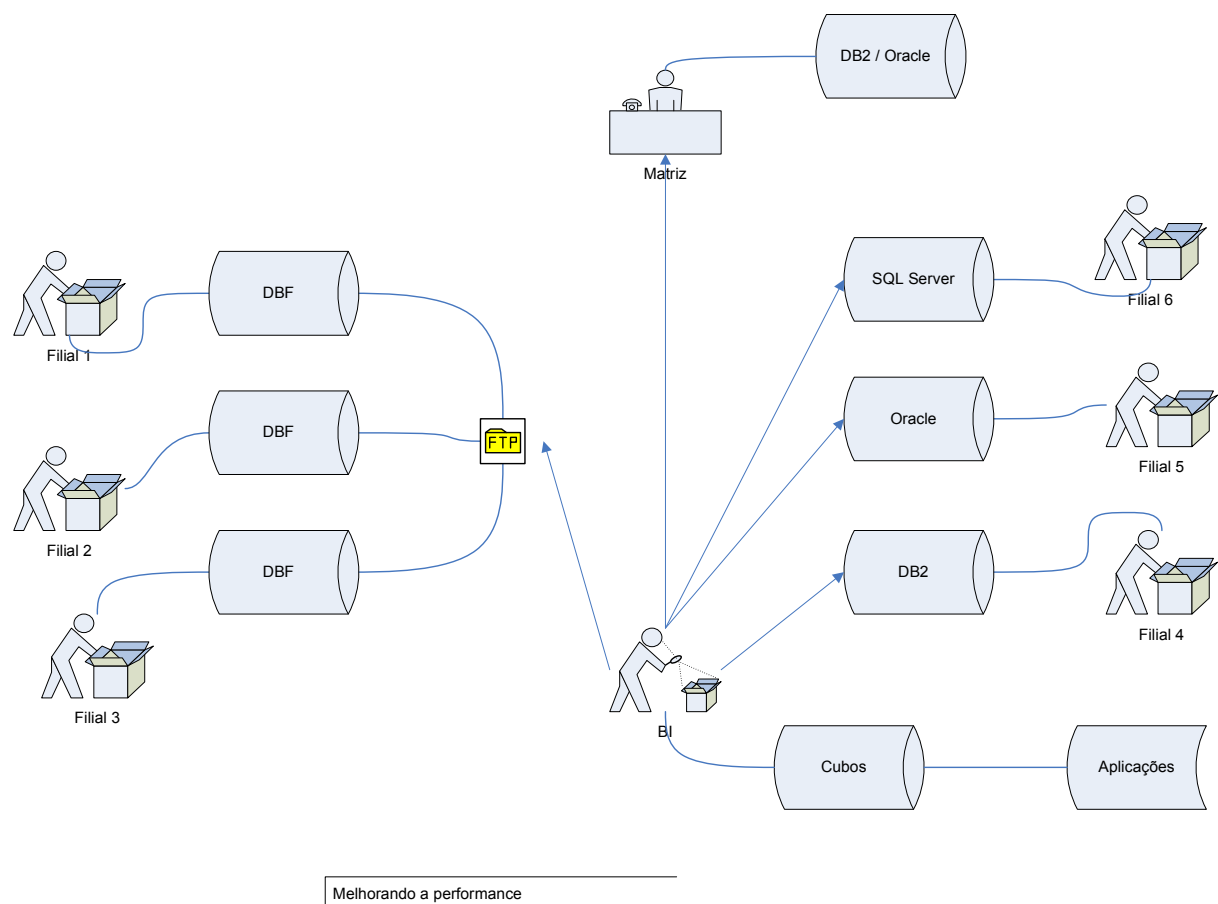
Como dito anteriormente o FTP inicialmente poderia ser uma solução definitiva para a integração total, mas infelizmente isso poderia causar um consumo de banda da rede e da transmissão de dados muito grande devido o tamanho dos arquivos.

As cargas ainda são problemáticas, principalmente quando em uma empresa o resultado de um dia de faturamento poderia ultrapassar 800.000 registros. Isto é, cada filial colocando no FTP um arquivo com esta quantidade de registros, e após isto serem carregados numa base do aplicativo do BI, isto resultaria numa queda drástica de performance, acarretando também mais demora na liberação da informação pelo BI.

O trabalho de integração muitas vezes reserva para os profissionais alguns problemas muito particulares, que devem ser tratados imediatamente, sem perder a informação do dia. Uma vez que a informação é disponibilizada não se tem mais a possibilidade de reestudar todo o projeto para implementar mudanças demoradas.

2.5 Lendo da fonte

Quando os executivos percebem que a informação está disponível e que está passando por processos de validações contínuas, começam a utilizar todo o seu potencial. Neste ponto não se tem mais a possibilidade de buscar informações em outro lugar, aumentando muito a responsabilidade do projeto de integração com BI. Qualquer mudança pensando em melhorias deve ser estudada e feito testes paralelos para evitar paralisações no uso do aplicativo, neste ponto o trabalho requer muita experiência por parte da equipe, pois qualquer tentativa sem uma ótima base de conhecimento poderia indisponibilizar a informação, fazendo com que a empresa tome decisões erroneamente, (Figura 7).



Diante dos fatores de atraso comentados, surge a necessidade de melhores formas de carga, como a exemplificada na (Figura 7, representando a opção ideal para atender uma demanda alta de arquivos para a realização de cargas para o sistema de BI. Observe que o sistema de BI realiza os acessos diretamente em cada base de dados, usando uma conexão ODBC comum do sistema operacional ou do fabricante do banco de dados.

A velocidade da carga aumenta exponencialmente, mesmo que algumas bases de dados fiquem no FTP, como os DBFs, isto não diminui a performance alcançada com os acessos diretos. Outro aspecto importante é o ganho quando houver alguma recarga durante o dia, ou seja, em operação, isto poderia ser feito sem perder as informações anteriores e ainda ter a posição atual da organização.

Respeitar a integridade de bases de dados mais antigas, mesmo sendo possível acessar essas bases de dados diretamente, o melhor é não concorrer com o sistema integrado nos acessos às bases, pois o risco de truncar a base de dados é muito alto. Por esta razão é que os DBFs são mantidos no FTP.

Lembrando que diversos trabalhos internos de cada filial ainda possuem uma complexidade alta e muitas vezes são paralelos aos sistemas integrados. Sabe-se que o fechamento mensal, até a composição do balanço, muitas informações são tratadas para serem inseridas no sistema; se não for assim alguns fechamentos não ocorrem ou não batem. Um exemplo bastante claro é quando se tem um sistema de folha de pagamento à parte do sistema integrado ou um sistema fiscal diferente do sistema contábil. Essa integração pode nunca ocorrer, pois o pessoal de TI raramente tem expertise em sistemas contábeis ou de recursos humanos. E, este é exatamente o ponto deste artigo, mostrar que a integração pode ser realizada e bem sucedida em pontos como este. Muito trabalho, conceitos de sistemas de bancos de dados, regras de negócios e de BI são essenciais para o sucesso de uma integração.

Imagine um fluxo de caixa que precise das informações de diversas filiais, cada uma mandando de um jeito, com nomes de campos e colunas diferentes, com formatações de valores distintos. Este geralmente é o meio normal de se formar um fluxo de caixa e com algumas horas de trabalho diário, com muito otimismo fecha-se o dia anterior com uma média de 7 horas de trabalho posterior ao fechamento.

2.6 Normalização

Como não existe nenhuma bibliografia que ajude a entender os sistemas legados em funcionamento nas empresas, e como foram criados tais bancos de dados, o trabalho de normalização geralmente toma toda a atenção do trabalho. Deve-se iniciar um estudo de todas as bases de dados existentes nas filiais e buscar uma compreensão absoluta das aplicações.

O projetista precisa estudar cada tabela, cada arquivo texto, cada planilha, cada DBF, cada campo, outros arquivos usados e porque são utilizados. Fazer as ligações para continuar alimentando os processos para não quebrar as regras existentes. Até encontrar uma forma de automatizar tudo sem interromper o fluxo de informações.

Todo o trabalho de normalização não interfere nos sistemas já implantados pela organização, toda mudança é feita dentro do sistema de BI, é nele que acontece toda a programação necessária para cruzar as correspondências entre os dados das filiais com a matriz.

2.7 Benefícios

Para entendermos os benefícios é preciso mensurar o valor da informação para a organização, o quanto ela é preciosa para a organização. Ou poderíamos questionar quanto tempo a organização sobrevive sem ter a informação.

A tomada de decisão nos tempos atuais não é mais uma arte e tornou-se uma ciência precisa.

Um executivo em posse de informações pode decidir em qualquer parte do mundo se deve ou não fechar um negócio. Ao buscar informações do seu último faturamento consolidado. Pode saber por filial, poderia até descer a um nível de detalhe, enfim ele decide quais medidas devem ser tomadas para que o negócio funcione melhor.

A organização moderna trabalha em função do cliente; a convergência para o cliente, como foco principal, tem sido a principal razão dos aumentos de vendas de sistemas de ERPs que oferecem níveis de análise melhor elaboradas para atender o cliente. Agregando valor ao seu produto, visando a satisfação do cliente. Quem poderia responder, qual produto da minha empresa agrega mais valor ao meu cliente? Esta é umas das perguntas que se responde com um projeto de integração, não é só numa visão consolidada, poderia ser por produtos regionais para clientes locais. A compreensão da curva ABC é absoluta, pode-se virar esse cubo de todos os lados para tomar decisões e preservar totalmente a margem do produto.

3. Conclusão

Existem inúmeros produtos de Sistema de Apoio a Decisão, na verdade temos visto até o SAP com funcionalidades voltadas a este assunto, alguns outros divulgam explicitamente que possuem um aplicativo de BI embutido no ERP. Hoje o foco desta discussão está em outro aspecto, a Forrester Research Inc., fez um recente estudo e diz que a maioria das grandes organizações tem entre 5 a 15 diferentes ferramentas de BI sendo usadas internamente. Isto indica que teremos que tratar de padronizações de sistemas de BI além das integrações das bases de dados discutidas nesta contribuição.

De qualquer forma percebemos que essas ferramentas são essenciais para a tomada de decisão, principalmente para as empresas que utilizam bases de dados heterogêneas, e que têm a dificuldade de transmissão de dados entre as filiais.

Mas qual é, exatamente, o tamanho desse mercado? De acordo com o IDC, as vendas mundiais de soluções de BI atingiram 3,9 bilhões de dólares em 2003, com aumento de 6,2% em relação a 2002. E deverão continuar crescendo a taxas médias de 4,5% ao ano, até 2008. No Brasil, os dados do IDC indicam que o mercado de BI, em 2003, chegou a 49,5 milhões de dólares, 6,8% maior do que em 2002. Foi uma das principais prioridades de investimento das empresas brasileiras no ano passado, segundo o relatório Brazil: Business Intelligence Tools 2004. A estimativa do IDC é que a taxa de crescimento do mercado seja mantida em 7,1% ao ano até 2008. A empresa E-Consulting leva em conta em suas projeções a receita total, incluindo hardware, software e serviços, como o desenvolvimento e a implantação de projetos. (INFO, 04/05/2007).

Sob esse ponto de vista, a E-Consulting estima que o mercado brasileiro atingiu 5,6 bilhões de reais em 2004, o que representará crescimento de 522% em relação a 2000. Só a divisão de consultoria da Brasoftware, que utiliza tecnologia Microsoft em seus projetos, registrou aumento de 120% na procura de aplicações e serviços na área de BI entre os meses de setembro de 2003 e de 2004. Para 2005, a E-Consulting previu mais 23% de crescimento na

receita total de aplicações de Business Intelligence, que poderá chegar a 6,9 bilhões de reais, (INFO, 04/05/2007).

4. REFERÊNCIAS

KAPLAN, R.S. & NORTON, D.P. *Mapas Estratégicos, Convertendo ativos intangíveis em resultados tangíveis*. Ed. Campus 6a Edição, 2005

FILHO, J.E.M. *Descobrendo o LINUX, entenda o sistema operacional GNU-LINUX*, 2006.

RÔMULO, S.O., et all, Sistemas Operacionais – Série Livros Didáticos Instituto de Informática UFRGS, Ed. Sagra Luzzatto 2ª Edição, 2001.

THOMPSON, M.A. Windows Server 2003 – Administração de Redes, 1ª Edição, São Paulo: Ed. Érica, 2003.

BADDINI, F. C., Windows Server 2003 – Implementação e Administração, 1ª Ed. São Paulo: Ed. Érica, 2003.

HACKATHORN, R. D., Web farming for the data warehouse, California – São Francisco, Morgan Kaufmann Publishers Inc., 1999.

GIL, A. L., Segurança em Informática – Ambientes Mainframe e de Microinformática, Segurança Empresarial e Patrimonial, 200 Questões sobre Segurança., 2ª Edição, São Paulo: Ed. Atlas, 1998.

NOVO, R., STARLIN, G., Segurança completa contra Hackers, 1ª Edição, Rio de Janeiro: Ed. Book Express, 2000.

TANEMBAUM, A. S., Operating systems : design and implementation, Englewood Cliffs, N.J. : Prentice-Hall, ©1987.

TANEMBAUM, A. S., Sistemas Operacionais Modernos, 2ª Edição, Ed. Pearson Brasil, 2003.

MACHADO, F.N.R., Aprenda as técnicas de Data Warehouse com exemplos e de forma simples e objetiva!, 1ª Edição, São Paulo: Ed. Érica, 2004.

INMIN, W. H., Como Construir o Data Warehouse, Rio de Janeiro:Ed. Campus, 1999.

INMIN, W. H., Como usar o Data Warehouse, Ed. IBPI, 1997

GONÇALVES, M., Extração de Dados para Data Warehouse, Ed. Axcel, 2003.

Crescimento e Conceitos do BI.

http://info.abril.com.br/canal/edicoes/1/conteudo_133618.shtml (04/05/2007)