Uma Visão Geral e Implementação de Extração-Transformação-Carga (ETL) Processo no Data Warehouse

(Estudo de caso: Departamento de Agricultura)

Rahmadi Wijaya1 , Bambang Pudjoatmodjo2
Escola de Ciências
Aplicadas Telkom
University Bandung,
[1]Indonésia rahmadi@telkomuniversity.ac.id [2]bpudjoatmodjo@telkomuniversity.ac.id

O processo Abstract-Extraction-transformation-loading (ETL) no desenvolvimento do data warehouse executa a extração de dados de vários recursos, transforma os dados em um formato adequado e carrega-os no armazenamento do data warehouse. No processo ETL, há uma função de processo de limpeza de dados que lida com dados de redundância, inconsistência e integridade. O processo ETL moverá os dados da fonte para a camada de integração (armazenamento de dados no data warehouse). Na camada de integração, os dados podem ser agrupados em escopos menores e mais específicos para o requisito em outros repositórios chamados de data marts. O programa de relatório do data warehouse será associado a um data mart como sua fonte de dados. Nesta pesquisa, o data warehouse é construído para lidar com o processo ETL. O data warehouse cria metadados para dar suporte ao processo. A construção de metadados para processos ETL levará a programas ETL com alto grau de reutilização. A conclusão desta pesquisa é que o uso do processo ETL dinâmico (usando metadados ETL) é necessário quando o processo ETL está lidando com o sistema operacional que ainda está instável e provavelmente mudará o esquema do banco de dados. O processo ETL dinâmico também é necessário para atender ao aumento da exigência de relatór

Palavra-chave-Data Warehouse, ETL, Camada de Integração, Reutilização, Metadados.

I. INTRODUÇÃO

A chave do sucesso para uma empresa sobreviver nos últimos tempos é a capacidade de analisar, planejar e reagir às mudanças no ambiente de negócios. Essa habilidade só será cumprida se informações adequadas estiverem disponíveis no formato e estrutura dos dados adequados à tomada de decisão [10].

O Departamento de Agricultura é uma agência governamental que possui dados que devem ser analisados, como dados financeiros, dados pessoais, dados de projetos, equipamentos de gerenciamento de dados e dados agrícolas, bem como dados de exportação e importação. Além disso, o departamento de Agricultura possui diferentes sistemas de informação para processar os dados em cada domínio em suas divisões, como Sistema de Informação de Gestão de Recursos Humanos, Sistema de Informação de Gestão de Projetos, Sistema de Informação de Gestão Financeira, Sistema de Informação de Gestão de Equipamentos, Sistema de Banco de Dados de Estatísticas Agrícolas, Sistema de banco de dados de importação de exportação. Atualmente, o Departamento da Agricultura extraiu dados de cada domínio manualmente e, em seguida, usa-os como informações na tomada de decisões para o departamento [5].

A tecnologia de data warehouse é usada para superar a variedade de sistemas de dados e informações. O desenvolvimento do data warehouse é realizado em um grupo, que inclui três componentes; processo de extração, transformação e carregamento (ETL), e como componente de suporte está o data mart e o data warehouse de relatórios, onde cada componente é construído separadamente de acordo com as condições do Departamento de Agricultura.
Ao construir o data warehouse, espera-se que o processo manual possa ser automatizado [8].

A situação atual do Departamento de Agricultura na análise e processamento de seus dados é descrita a seguir: Cada sistema de informação tem seu próprio banco de dados diferente. Existe até um formato de arquivo simples (.xls), os dados estão dispersos em cada sistema de informação, ainda ocorre redundância de dados, na análise e tomada de decisão os Executivos da Secretaria da Agricultura precisam fazer uma recapitulação dos dados de cada sistema de informação.

Para melhorar o desempenho do Departamento, os executivos do Departamento de Agricultura devem dispor de dados precisos e atualizados para gerar decisões estratégicas.
Esta pesquisa discutirá sobre o planejamento de data warehouse com estudo de caso no Departamento de Agricultura [7].

II. ARMAZÉM DE DADOS

A. Projeto de

Lei de Definição Em armazenamento de dados não definido como coleta de dados para apoiar os processos de tomada de decisão da administração. Tem várias características [8]:

- Orientado ao assunto
- Integrado
- Não volátil
- Tempo variável

978-1-4799-7752-9/15/$31,00 ©2015 IEEE

B. Arquitetura de armazenamento de dados

O data warehouse é um sistema de arquitetura aberta e pode ser arquitetado de várias maneiras diferentes, dependendo das necessidades específicas dos requisitos do sistema. Figura 1. descreve um exemplo de arquitetura de data warehouse. A arquitetura mostra a diferença entre data warehouse e banco de dados em sistema operacional. No data warehouse, os dados são usados para fins específicos, geralmente como ferramentas de análise para aplicações em Sistema de Informação Executiva (EIS) ou Sistema de Apoio à Decisão (DSS). Por outro lado, bancos de dados operacionais são geralmente utilizados em aplicações transacionais que farão operações de leitura/escrita no banco de dados[10].

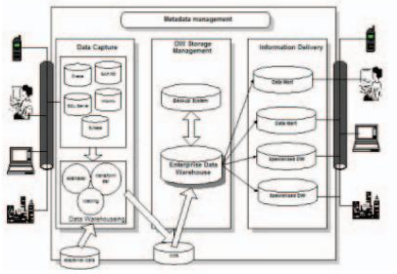

Figura 1. Arquitetura de armazenamento de dados

A arquitetura na Fig. 1 é construída sobre propósitos específicos, que são [8]:

1. A entrada de dados para data warehouse não é apenas do sistema interno. A entrada de dados deve acomodar recursos externos.

2. As informações armazenadas no data warehouse podem ser especializadas em vários data warehouses mais específicos (data mart), de modo que vários processos adicionais na arquitetura são necessários para preencher os dados do data warehouse em algum data mart.

3. A aplicação na camada do usuário se desenvolve em vários modelos, por exemplo: sistema baseado na web, baseado em desktop e até mesmo baseado em dispositivos móveis

C. Acesso do usuário O acesso do usuário é um componente que define como um usuário acessa o data warehouse. Normalmente, o usuário usa software de Business Intelligence para definir consulta e análise no acesso ao data warehouse. Abaixo estão listas de exemplos de Software de Business Intelligence:

1. Sistema de Apoio à Decisão 2. Sistemas de Informação Executiva 3. OLAP (Online Analytical Processing)
4. Ferramentas de Mineração de Dados

III. ETL (EXTRAÇÃO, TRANSFORMAÇÃO, CARREGAMENTO)

Durante o processo de ETL, os dados de várias fontes serão extraídos e integrados no data warehouse periodicamente.
A extração é um processo para identificar e recuperar todos os dados relevantes das fontes. O papel da transformação é limpar os dados e integrar diferentes esquemas para definir esquema no data warehouse. Enquanto isso, o carregamento é um processo para mover fisicamente os dados do sistema operacional para o data warehouse.

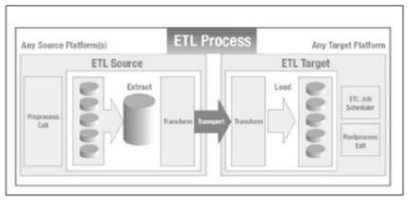

Fig.2. Processo ETL

A. Conceito ETL É necessário definir o escopo ETL analisando cada tabela de destino (suas dimensões e fatos) no início do processo de arquitetura ETL. É necessário buscar o comportamento de cada tabela alvo; onde está sua fonte e que tipo de processo de negócios depende dela [8].

1. Metadados No desenvolvimento inicial do data warehouse, a integração é formada pela criação de um programa ETL específico para a estrutura do banco de dados de origem e do banco de dados do data warehouse.

Com o passar do tempo, descobriu-se que esses programas ETL específicos estão essencialmente fazendo o mesmo processo. Muitos programas de bloco são reutilizáveis para outro processo ETL. Neste ponto, são desenvolvidas ferramentas ETL que podem fazer a integração automatizada de dados ao data warehouse.

Para resumir, metadados são dados sobre dados. Especialmente, metadados ETL são dados sobre o processo ETL. Definir metadados ETL é necessário para construir programas ETL com alta reusabilidade [8].

2. Extração

A extração identifica todas as fontes relevantes e extrai os dados da maneira mais eficiente possível. Esse processo está passando por um arquivo ou banco de dados, usando vários critérios na seleção de dados, encontrando dados adequados e, em seguida, transportando os dados para um arquivo ou outro banco de dados.

A captura de dados de alteração (CDC) é um elemento importante na análise de extração. Quase todas as transações em dados têm carimbos de data/hora. No entanto, nem todos os dados de dimensão no sistema de origem possuem carimbos de data/hora devido à sua tendência de não depender de um evento.
Portanto, é muito difícil implementar o CDC em dados dimensionais[8].

3. Transformação

Transformação é um processo de manipulação de dados do sistema de origem para outro formato em data warehouse ou data mart para torná-los uma informação significativa. As funções de transformação que podem ser usadas estão listadas abaixo[8]:

4. Carregamento

O carregamento moverá os dados transformados para o data warehouse. Existem duas estratégias de carregamento na camada de integração; estratégia de carregamento para tabela de dimensões e estratégia de carregamento para tabela de fatos.

978-1-4799-7752-9/15/$31,00 ©2015 IEEE

B. Estratégia de Carregamento para Tabela de Dimensões

Existem três estratégias para carregar a tabela de dimensões. Cada estratégia aborda a alteração da dimensão de dados de maneira diferente, dependendo da maneira como lida com os dados da dimensão antiga. Em todas as estratégias, se não houver registro com a mesma chave natural, um novo registro será adicionado. A chave natural é um atributo de dimensão que diferencia exclusivamente o registro de dimensão (não é uma chave substituta). Se houver registro com a mesma chave natural, então: •

Estratégia 1: O histórico dos dados não é armazenado. Se o valor de entrada ocorreu na tabela de dimensões com base no valor da chave natural, o registro será atualizado.

- Estratégia 2: Colunas críticas são colunas importantes na tabela de dimensões que devem ser reservadas. Se nos dados de entrada for encontrada uma ou mais colunas na categoria de coluna crítica e o valor nos dados de entrada for diferente da coluna correspondente na tabela de dimensões (com base no valor da chave natural), o registro correspondente será expirado e o novo registro com novo chave substituta será adicionada. Se não houver similaridade na coluna crítica, ela será tratada como na estratégia 1, o registro correspondente será atualizado.

- Estratégia 3: Sua estratégia em lidar com alterações de valor na coluna crítica é bastante semelhante à estratégia 2. A diferença é que, na estratégia 3, para cada coluna crítica existem colunas diferentes que são colocadas em cada registro para armazenar o valor do dado atual e n dados anteriores valor.
  Quando houver alteração no valor da coluna crítica, todos os valores serão deslocados e o último valor (o mais antigo) será excluído e o valor mais recente será adicionado

C. Estratégia de carregamento para tabela

de fatos O carregamento da tabela de fatos consistirá apenas na adição de novos dados. Cada novo registro será conectado com sua dimensão com base no valor da chave natural. A chave substituta correspondente na tabela de dimensão é obtida das informações da chave natural da entrada do registro [8].

4. ESTUDO DE CASO

Nesta pesquisa, o Sistema de Informação Executiva (EIS), geralmente o Sistema de Apoio Executivo (ESS) será implementado no Departamento de agricultura. O ministério tem muitas divisões que operam em diferentes campos. O EIS é usado para integrar dados de muitas fontes de banco de dados dispersas na divisão de produção.
O EIS será utilizado para aumentar a eficiência e eficácia da tomada de decisão dos executivos [10].

Secretaria de Agricultura não possui banco de dados integrado entre suas divisões. A base de dados é gerenciada separadamente e localmente em cada divisão, e possui formato de dados não informados. Isso leva a uma dificuldade de tomada de decisão global para os executivos, os diretores de alto escalão. Os dados dispersos irão incomodar o executivo em fazer análises para extrair os dados em informações desejadas [3].

Além disso, o banco de dados geralmente é usado para armazenar dados detalhados. Não é suficiente com a exigência de dados dos executivos. Eles geralmente precisam de um resumo dos dados, apresentados em uma visão gráfica como um diagrama para ajudá-los na tomada de decisão [10].

Problemas relevantes que possivelmente serão levantados na implementação do EIS no departamento de agricultura, relacionados com a quantidade de divisão que tem, são [10]:

1. Devido à falta de padronização do formato dos dados, provavelmente ocorrerá incompatibilidade entre os bancos de dados.
2. Extraindo informações. De vários dados no data warehouse, o EIS deve ser capaz de extrair dados brutos de várias fontes disponíveis, além de fornecer informações úteis para os executivos, geralmente para prever tendências em cinco anos. É necessário determinar o formato da solicitação na extração de informações e, em seguida, combiná-lo com os dados no data warehouse. Por exemplo, que a solicitação seja informações gerais de todas as divisões, e o formato de dados para solicitação de informações é: somatório de despesas no mês atual, segmentação de clientes, despesas operacionais por mês e alocação de riscos de custos.
3. Mantenha a precisão dos dados. Deve-se garantir que os dados acessados sejam os necessários.
4. Escolher o momento certo para produzir e armazenar dados no data warehouse. Por exemplo, determinar as frequências de coleta de dados, pode ser feito uma vez por mês, duas vezes por mês ou em qualquer outro intervalo.

Determinar o tempo expirado em manter os dados. A quantidade de dados armazenados pode atingir um número enorme; portanto, o período expirado no armazenamento de dados deve ser determinado. Ele pode economizar capacidade de memória e melhorar o desempenho da velocidade EIS.

A. Análise de data warehouse no departamento de agricultura

A construção do data warehouse no departamento de agricultura é apenas a fase inicial para o desenvolvimento do data warehouse integrado no departamento de agricultura.

O objetivo da construção do data warehouse, portanto, é trazer a necessidade de análise geral para os executivos do departamento de agricultura. Esta pesquisa construirá um sistema dinâmico que se adapte às novas exigências dos executivos por meio da construção do data warehouse.

Nos últimos tempos, os executivos do departamento de agricultura estão acessando dados de relatórios por meio do Centro de Dados e Informações do Departamento de Agricultura *(Pusat Data dan Informasi Dinas Pertanian,* PUSDATIN). Mesmo tendo o privilégio de acessar diretamente os dados no sistema operacional sob sua divisão, os dados são dados transacionais, portanto, são muito detalhados e não podem ser usados como referência de análise [7].

B. Requisitos do sistema para armazenamento de dados no Departamento de Agricultura

A realização do processo, a extração de dados operacionais e sua apresentação como relatórios, gerará problemas se for feito manualmente. Combinar dados que envolvem muitas fontes é um processo complexo, portanto, vulnerável a falhas. A complexidade do processo aumenta se ele contiver muitos dados redundantes [7].

Os pontos fracos do sistema atual são [7]:

1. Acesso

   Os executivos estão acessando o relatório através do Centro de Dados e Informações do Departamento de Agricultura

978-1-4799-7752-9/15/$31,00 ©2015 IEEE

(Pusat Data dan Informasi Dinas Pertanian, PUSDATIN). Embora tenham acesso direto ao sistema operacional, os dados do sistema não são adequados o suficiente para serem usados como referência de análise.

2. Tempo

Caso ocorra uma transação no sistema operacional, os executivos não conseguem obter o relatório sobre ela em pouco tempo.
Os executivos enfrentarão uma longa burocracia para conseguir o relatório.

3. Formato

O formato do relatório geralmente é definido pelos executivos no início da solicitação do relatório. Depois de receber o relatório em um formato, os executivos devem realizar exatamente os mesmos procedimentos para obter informações em outro formato.
Assim, é difícil para os executivos visualizar as informações sob várias perspectivas, como planilhas ou gráficos.

4. Integridade

Embora as informações desejadas tenham sido obtidas, a precisão dos dados ainda é questionável. O processo manual envolvendo enormes dados redundantes é suscetível a falhas.

O data warehouse oferece privilégio para os executivos obterem os relatórios desejados em um período de tempo relativamente curto. Usando um aplicativo de relatório compatível com data warehouse, os executivos podem ver as informações em vários formatos.

O data warehouse do Departamento de Agricultura terá três domínios; produção, finanças e recursos humanos. O data warehouse é construído para facilitar o acesso a informações de dados importantes no departamento de processos de negócios da agricultura pelos executivos para a tomada de decisões estratégicas. O sistema carrega dados das três fontes de domínio, extrai-os em um único banco de dados e mostra o resultado como relatórios para seu usuário.

Os dados são obtidos por meio do sistema operacional que possui vários formatos de banco de dados de dados e de origem. O sistema operacional utiliza Sistema de Gerenciamento de Banco de Dados (SGBD), bem como planilha eletrônica (Microsoft Excel). O processo de extração é agendado pelo administrador do data warehouse. Antes disso, o cronograma é feito pelos executivos, operador do sistema operacional e administrador do data warehouse. Além disso, o administrador tem a responsabilidade de manter o banco de dados no data warehouse, como backup, recuperação e executar o processo de extração (incidentalmente) se houver uma falha no processo de extração.

A programação padrão para o processo de extração é uma vez por mês para todos os dados do sistema operacional. Se houver uma alteração no requisito, o administrador pode alterar a programação para atender à alteração.

Privilégio deste administrador, o usuário não tem este privilégio. O usuário é conectado ao sistema por meio de ferramentas OLAP como uma interface de usuário, conforme declarado anteriormente [8].

O data warehouse usa DBMS SQL Server 2000 para seu banco de dados. Ele também usa o tipo de dados padrão do DBMS, como char, varchar, number e datetime. Para o recurso, ele usa o recurso padrão DDL (Data Definition Language) para definir banco de dados e tabela e chave estrangeira.

Serviço de transformação de dados (DTS) com script ActiveX
A tarefa no VBScript é usada para o processo ETL.

A fonte de dados será extraída, transformada e carregada na camada de integração. Em seguida, os dados obterão mais extração em vários data marts. O aplicativo de relatório está conectado ao data mart com os dados necessários [2].

Este documento é focado exclusivamente no processo no fluxo 1 (processo ETL da fonte de dados para a camada de integração). Os processos de extração para data mart e a utilização de dados em data mart estão além do escopo deste documento.

Recebendo os dados extraídos do sistema operacional, o sistema de data warehouse fará um processo de transformação simples, incluindo o processo de limpeza e integração.

O sistema possui dicionário para possíveis valores de dados e correção de dados para falha de valor de dados.

O processo de integração também é definido no dicionário como referência de valor de informação para outro dado, se houver uma ocorrência [2].

C. Análise do sistema para armazenamento de dados do Departamento de Agricultura

A fim de estar preparado para um maior desenvolvimento como data warehouse integrado no Departamento de Agricultura, este sistema inicial é desenvolvido considerando o crescimento dinâmico dos requisitos. A mudança de requisitos é acomodada sem alterar a estrutura do sistema, bem como os programas usados pelo sistema. Isso apenas mudará a configuração do sistema.

Este sistema tem uma fraqueza no processo de integração.
O processo não é feito completamente para todos os dados no sistema. O motivo dessa condição é, apesar da característica de dados de origem (sistema operacional possui apenas parte dos dados).

Há também uma mesma informação que é armazenada como assunto diferente neste sistema. Assim, um processo de integração adicional pode ser desenvolvido no sistema sem perturbar todo o sistema.

D. Fluxo do Processo ETL

O fluxo do processo para projetar o ETL pode ser descrito da seguinte forma:

a. Tabelas identificadas dos dados de origem para carregar os dados necessários para a métrica e a dimensão que foram definidas com base nos requisitos do usuário.
b. Os dados são movidos para a área de preparação, de modo que o processo ETL que será feito não perturbe o processo de transação OLTP que ainda está em execução.
Esta pesquisa utiliza o SQL Server como área de staging, sendo que o nome do banco de dados é “pertanian”.
c. Os dados armazenados na tabela de fato e dimensão são transformados em dados multidimensionais que podem ser visualizados de qualquer ponto de vista, podendo ser acessados pelo cliente para fins de análise.

978-1-4799-7752-9/15/$31,00 ©2015 IEEE

E. Análise dos requisitos do usuário

Nesta fase, o banco de dados do Departamento de Agricultura é dividido em 3 domínios; produção, finanças e recursos humanos. A Fig.3 mostra o processo ETL para esta etapa.

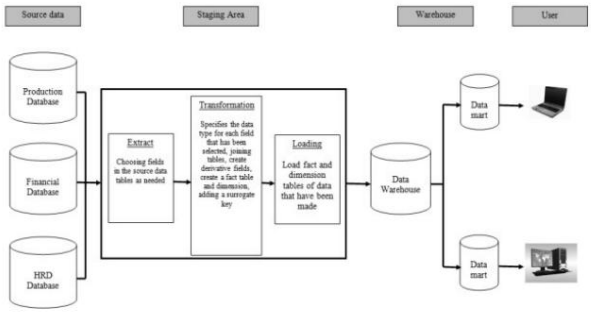

Fig. 3. Processo ETL

V. CONCLUSÃO

A fim de estar preparado para um maior desenvolvimento como data warehouse integrado no Departamento de Agricultura, este sistema inicial é desenvolvido considerando o crescimento dinâmico dos requisitos. A mudança de requisitos é acomodada sem alterar a estrutura do sistema, bem como os programas usados pelo sistema. Isso apenas mudará a configuração do sistema.

O processo ETL para armazenamento de dados no Departamento de Agricultura é organizado para ser realizado em tempo real, não envolverá uma tabela temporária como intermediária. Os dados são extraídos, transformados e carregados diretamente no banco de dados. Esse resultará em um processo ETL mais rápido, mas, por outro lado, consumirá recursos tanto no sistema operacional quanto no sistema de data warehouse. Portanto, o processo ETL é escalonado quando o sistema operacional não está ocupado com sua transação.

O processo de transformação faz um processo simples, que está corrigindo o valor dos dados usando a tabela do dicionário. Depois de transformados, os dados são carregados diretamente no banco de dados do data warehouse.

REFERÊNCIAS

[1]. Fathansyah, Ir., 2002, Buku Teks Ilmu Komputer Basis Data, Informatika, Bandung

[2]. Fowler, Martin, 2004, UML Distilled Edisi 3 Panduan Singkat Bahasa Pemodelan Objek Standar, 2005, Andi, Yogyakarta

[3]. http://www.cert.or.id/~budi/courses/ec7010/dikmenjur 2004/supawi-report.pdf

[4]. http://www.iwaysoftware.com/products/images/etl_chart_s m4.gif [5]. http://www.mcrit.com/ASSEMBLING/assemb_central/WhatESS.htmhttp://www.ptct.com/EIS.html

[6]. http://www.utminers.utep.edu/mmahmood/cis5311dtmba/slides/chapter02.ppt

[7]. Ibrahim, Nugroho Setyabudhi, Takariyana Heni A., 2004, Perancangan Data Warehouse Pada Pusat Data dan Informasi Pertanian, Tesis Magister Manajemen Informasi Universitas Bina Nusantara, Jacarta

[8]. Inmon, WH, 2002, Building The Data Warehouse, Terceira Edição, John Wiley and Sons, Inc., Nova York

[9]. Pressman, Roger S., Ph.D., 2002, Rekayasa Perangkat Lunak Pendekatan Praktisi (Buku II), Andi, Yogyakarta

[10]. Turban, Efram, Jay E. Aronson e Ting Peng Liang, 2005, Sistemas de Apoio à Decisão e Sistemas Inteligentes (Sistem Pendukung Keputusan dan Sistem Cerdas), Edisi 7 Jilid 1, Andi, Yogyakarta

978-1-4799-7752-9/15/$31,00 ©2015 IEEE