**UNISUL – UNIVERSIDADE DO SUL DE SANTA CATARINA**

**ANDERSON DE FARIA**

**DESENVOLVIMENTO DE UM DATA WAREHOUSE E IMPLANTAÇÃO DE FERRAMENTA DE ANÁLISE DE DADOS NA EMPRESA PUBLICAÇÕES ONLINE**

Florianópolis

2014

**ANDERSON DE FARIA**

**PROPOSTA DE DESENVOLVIMENTO DE UM DATA WAREHOUSE E IMPLANTACAO DE FERRAMENTA DE ANÁLISE DE DADOS NA EMPRESA PUBLICAÇÕES ONLINE**

Monografia apresentada ao Curso de

Especialização em Engenharia de Projetos de

Software da Universidade do Sul de Santa

Catarina, como requisito parcial à obtenção do

título de Especialista em Engenharia de Projetos

de Software.

Orientador: Aran Bey Tcholakian Morales

Florianópolis

2014

# Introdução

## Apresentação do problema

Nas últimas décadas houve um avanço considerável no que se diz respeito à tecnologia da informação e na maneira como as informações são armazenadas. A algumas décadas atrás a maioria das empresas armazenavam a maioria de seus dados em arquivos de papel em grandes salas e almoxarifados. Com o surgimento dos sistemas de informação as organizações começaram a automatizar seus processos e a mudar o meio de armazenamento dos dados.

Hoje podemos dizer que seria impossível grandes organizações sobreviverem sem sistemas de informações e de armazenamento dos seus dados. Grande maioria das empresas possuem softwares que controlam as operações e sistemas de armazenamento de dados para suprir a necessidades destes sistemas. Este tipo de sistema não garante informações exatas que auxiliem no processo de gestão para tomada de decisões relacionadas aos objetivos de negócio das organizações.

Com a implantação destes novos softwares as organizações passaram a ter diferentes sistemas como por exemplo um sistema financeiro, um para recursos humanos e outro para controle das operações, gerando diferentes fontes de dados. Ao perceber a necessidade de obter informações de todas as fontes de dados para auxiliar no processo gerencial, as organizações começaram a desenvolver projetos de informações chamados Data Warehouse que servem como um repositório de dados históricos para auxiliar na visão gerencial das organizações.

A empresa Publicações Online fornece um serviço de busca de intimações judiciais em nome de advogados nos diários oficiais da justiça de todos os estados brasileiros. A empresa foi fundada no ano de 2003 iniciando apenas no estado de Santa Catarina. Após alguns anos se especializou em todo o território nacional. Atualmente conta com mais de vinte mil clientes ativos em todo o Brasil.

Ao longo destes onze anos foi desenvolvido, junto com uma empresa parceira, um sistema ERP (Enterprise Resource Planning) chamado de intranet onde estão todos os módulos do sistema de informação da empresa: Administrativo, Comercial, Financeiro, Recursos Humanos, Operacional, Gerência e Tecnologia da Informação. Todas os dados da empresa estão armazenados em um único servidor, porém distribuídos em diferentes bancos de dados relacionais.

Esta organização possui um grande volume de dados armazenados ao longo destes onze anos, mas não possui ferramentas de análise destes dados de forma que possam ser extraídas informações relevantes para auxiliar a diretoria da empresa na tomada de decisões e elaborações de novas estratégias de negócio.

Os diretores da empresa desejam planejar a expansão de seus negócios, mas para isso precisam de respostas para algumas questões como: quanto a empresa cresceu no último ano, se o número de clientes aumentou no último mês em relação ao mesmo mês do ano passado, se o lucro da empresa aumentou proporcionalmente em relação ao número de clientes, qual o tipo de cliente que mais contratou o serviço no último ano, etc. De acordo com as respostas os diretores podem tomar decisões como: contratar mais funcionários para suporte, aumentar a equipe de TI, realizar campanhas para captação de clientes, etc.

O sistema atual possui uma grande quantidade de relatórios baseados na sua base de dados de operação mas que, na maioria das vezes, disponibilizam informações incompletas ou incorretas. Em muitas ocasiões os diretores da empresa acabam tomando decisões e elaborando estratégias de negócio que não surtem efeito, ou até mesmo prejudicam o desempenho da organização, por não terem as informações corretas.

Para ajudar a responder as perguntas indagadas pelos diretores da empresa e auxiliar no processo de gestão e tomada de decisão é interessante criar uma ferramenta BI (Business Intelligence) com um repositório de dados históricos centralizado, onde todas as informações relevantes seriam armazenadas para posteriores análises. Também é necessário o desenvolvimento de uma ferramenta de extração dos dados das bases de dados relacionais para que sejam unificados no Data Warehouse.

Mas para gerar informações corretas para os diretores da empresa, não basta apenas possuir um histórico dos dados, é necessário a implantação de uma ferramenta de análise destes dados armazenados. Esta ferramenta deve gerar relatórios e gráficos que ajudem os diretores a interpretar os fatos e responder às suas perguntas. Com os dados apresentados de forma adequada a análise das informações fica mais clara ajudando os diretores na tomada de decisão.

## Justificativa

A quantidade de dados armazenados pela empresa Publicações Online vem aumentando consideravelmente a cada ano, tornando a coleta e a análise dos dados para gerar informações cada vez mais difícil. Os diretores da empresa ficam sem saber o que está realmente acontecendo com seus clientes, e quais caminhos devem seguir.

Este projeto de pesquisa e desenvolvimento de um Data Warehouse com a implantação de uma ferramenta de extração dos dados operacionais tem como justificativa melhorar a qualidade das informações gerenciais geradas para auxiliar nas tomadas de decisão por parte dos diretores da empresa.

Em um Data Warehouse, os dados históricos são armazenados e centralizados em uma única fonte de dados possibilitando uma análise mais precisa e gerando informações mais concretas. A extração dos dados da base de dados atual não seria tão complicada já que os dados atuais estão armazenados no mesmo servidor.

Com um Data Warehouse definido e alinhado com as necessidades de negócio da empresa, os dados armazenados ficariam mais claros e prontos para a implantação de uma ferramenta de análise dos dados possibilitando a geração de informações relevantes para auxiliar os gestores da empresa na tomada de decisão.

Com as informações adequadas, os diretores da empresa poderão verificar onde a empresa se encontra no momento e quais caminhos poderão seguir para expandir seus negócios de maneira organizada e efetiva.

## Objetivos

### Objetivo Geral

Esta pesquisa tem como objetivo principal o desenvolvimento de um Data Warehouse e a carga de seus dados para a empresa Publicações Online baseado nos questionamentos feitos pelos diretores da organização.

### Objetivos específicos

1. Pesquisar conceitos e maneiras de modelar, desenvolver e implantar um Data Warehouse.
2. Definir junto aos diretores da empresa um assunto e suas medidas necessárias para auxiliar no processo gerencial e na tomada de decisão.
3. Modelar e criar o banco de dados dimensional baseados nas questões levantadas pelos diretores.
4. Desenvolver um processo par efetuar a carga dos dados no banco de dados dimensional criado.

## Estrutura do trabalho

No primeiro capítulo são apresentadas uma contextualização do problema, a justificativa e os objetivos desta pesquisa.

No segundo capítulo serão abordadas os principais conceitos e definições relacionados à BI (Business Intelligence) como: Sistemas de informação, Sistemas OLPT e SAG, Data Warehouse, modelagem dimensional, ETL e ferramentas OLAP.

No terceiro capítulo serão apresentadas as questões levantadas pelos diretores da empresa, o fato escolhido com suas respectivas medidas o que a empresa possui de dados relacionados ao assunto determinado.

No quarto capítulo será apresentada a modelagem dimensional da tabela fato junto com suas respectivas medidas e a criação da base de dados.

No quinto capítulo será apresentada a solução para realização da carga dos dados originários do banco de dados relacional para o banco de dados dimensional (Data Warehouse).

No sexto capítulo será apresentada a conclusão do trabalho, pontos positivos e negativos do desenvolvimento de um data warehouse de acordo com o assunto levantando junto aos diretores.

# Revisão Bibiolgrafica

Para proporcionar um embasamento teórico sobre BI (Business Inteligence) serão abordados nestes capitulo os seguintes assuntos: Sistemas de informação, Sistemas OLPT e SAG, Data Warehouse, modelagem dimensional, ETL e ferramentas OLAP.

## Sistemas de informação

Sistema é um grupo de componentes relacionados trabalhando para a obtenção de um objetivo comum.

Segundo LAUDON (1998), um Sistema de Informação consiste em um conjunto de componentes que se relacionam entre si trabalhando juntos para alcançar um objetivo. A coleta, o processamento, o armazenamento e a distribuição da informação são realizados com o intuito de auxiliar o planejamento, o controle e a análise das informações dentro das organizações.

Os sistemas de informação armazenam informações de coisas do interesse da organização auxiliando seus colaboradores na execução de suas tarefas, análises, tomadas de decisões e resolução de problemas. Para isso são realizadas cinco atividades descritas abaixo:

* Entrada: é a coleta de dados de dentro ou fora da organização realizada através da interface com o usuário;
* Processamento: engloba a conversão dos dados para uma forma que seja útil para a organização. São realizadas algumas operações como soma, comparação, classificação e atualização;
* Saída: envolve a disponibilização da informação processada para uma pessoa ou processo dentro da organização como pro exemplo relatórios gráficos e para impressão;
* Armazenamento: consiste em guardar a informação de maneira organizada para posterior consulta;
* Controle: onde são controladas todas as atividades do sistema, acesso ao sistema, segurança do sistema, etc.

De acordo com LAUDON (1998 ), “Um sistema de informação é uma parte integrante de uma organização e é um produto de três componentes: tecnologia, organizações e pessoas.” As organizações são compostas por pessoas que desempenham diferentes funções. Estas pessoas necessitam dos sistemas de informação para auxiliar na realização das tarefas. As pessoas provêm dados de entrada para os sistemas e recebem os dados de saída que foram processados pelo sistema de informação. A tecnologia é o meio como os dados são enviados, processados e disponibilizados para as pessoas. Dentro da tecnologia temos os seguintes componentes:

* Hardware: equipamento físico onde se realiza as atividades de entrada, processamento e saída. Exemplos: computadores, impressoras, tablets, etc.;
* Software: são instruções programadas para controlar o funcionamento do hardware e realizar tarefas especificas para cada sistema de informação. Exemplos: editores de texto, programas de envio de e-mail, driver de placa de vídeo, etc.;
* Armazenamento: engloba tanto a parte física do armazenamento de dados como fitas magnéticas e discos quanto os softwares que controlam estes dispositivos. Exemplos: cd-rom, pendrive, disco óptico, sistema de gerenciamento de banco de dados (SGBD), etc.;
* Comunicação: consiste nos meios físicos e softwares que possibilitam a comunicação entre sistemas de informação. Exemplos: cabo de rede, fibra ótica, placa de rede wireless, programas de compartilhamento de arquivos, etc.

Existem vários tipos de Sistemas de Informação como:

* Sistemas de Processamento de Transações: on-line ou em lote, grande processamento de dados;
* Sistemas de Controle de Processo: sensores e processos utilizados nas indústrias;
* Sistemas colaborativos: e-mail, videoconferência, possibilitam a iteração entre uma equipe, grupo de estudo ou trabalho;
* Sistemas de Informação Gerencial (SAG): sistemas que disponibilizam informações gerencias como relatórios;
* Sistemas de Apoio à Decisão: simulação que auxiliam os gestores na tomada de decisão;
* Sistemas Especialistas: baseados no conhecimento realizando tarefas especializadas de uma determinada área;

### Dado, informação e conhecimento

Segundo LAUDON (1998), os dados podem ser definidos como fatos brutos sem caráter informativo. No caso de sistemas de informação os dados são registros dos processamentos e são armazenados para consultas. A informação é um conjunto de dados formatados e organizados de uma maneira que seja útil e apresente algum significado para o usuário. O conhecimento são estratégias utilizadas por pessoas ou máquinas para criar, armazenar, interpretar informações sobre um determinado assunto.

Segundo LEME FILHO (2004), os dados são simples fatos, simples de organizar, apresentável e fácil de transferir. A informação são dados com alguma semântica, contextualizado por sistemas, é necessário algum esforço para transferir e apresentados por relatórios gráficos. O conhecimento é a informação com valor agregado sobre determinado assunto, é de difícil organização e apresentação, muito difícil de transferir e em sua apresentação depende de interpretação.

A informação e o conhecimento estão sendo cada vez mais importantes nas organizações para que seus objetivos sejam alcançados. Informações e conhecimento sobre os clientes e quais as suas preferências, por exemplo, se tornam cada vez mais importante pelo fato da empresa poder oferecer produtos e aumentar suas vendas.

## Sistemas OLPT e SAG

### OLPT - Online Transation Processing

OLPT são sistemas de processamento de transações responsável por monitorar e executar tarefas rotineiras do dia a dia de uma organização. Manipulam dados atuais, armazenados em detalhe e sem necessidade de histórico. O modelo de armazenamento em banco de dados é o Modelo Relacional.

Segundo THOMSEN (2002), OLPT são sistemas que auxiliam a parte operacional de uma organização, utilizados por usuários que executam as tarefas do dia a dia. Tem como objetivo lidar com a maior quantidade de dados, processando mais transações e suportar o maior número de usuários concorrentes. Um exemplo de sistema é um sistema de reserva de uma companhia aera, onde é necessário assegurar que muitos usuários possam acessar o sistema ao mesmo tempo garantindo a consistência dos dados processados.

### SAG - Sistemas de apoio a gerência

Sistemas SAG conhecidos também como sistemas OLAP (Online Analytical Processing), são sistemas voltados para a análise de um grande volume de dados e informações fornecendo suporte para a parte gerencial de uma organização. Lidam com dados sumarizados históricos de um determinado assunto e utilizam o Modelo Dimensional para armazenamento dos dados.

De acordo com TURBAN (2003), com o aumento da capacidade computacional, foi possível desenvolver sistemas que realizam tarefas mais analíticas do que simplesmente processar transações. Estes sistemas também conhecidos com MISs (management information systems), tem como objetivo de fornecer informações sobre o andamento das atividades da empresa para assegurar que sejam executadas de forma eficiente. SAG normalmente geram relatórios periodicamente sobre produtividade, eficácia, etc. Estes relatórios são gerados a partir de dados processados consultados no banco de dados da organização e armazenados em formato histórico para futuras consultas. Os sistemas SAG são utilizados para dar suporte a decisões rotineiras, prever tendências, planejar, monitorar e controlar atividades de uma empresa possibilitando os gerentes a detectarem problemas em seu estágio inicial.

### Diferenças entre sistemas OLPT e SAG

Afirma INMON (1997), que os sistemas OLPT exigem um tempo de resposta rápido em relação aos SAG. O desempenho das organizações dependem diretamente do tempo de resposta dos sistemas OLPT que coordenam as operações. Nos sistemas SAG o tempo de resposta não é um fator crítico mas não menos importante.

Em se tratando de usuários do sistema, os sistemas OLPT são utilizados pela parte operacional da empresa em processos repetitivos e diários dando suporte as operações diárias. Já os sistemas SAG são utilizados por gerentes e diretores dando suporte a tomada de decisões estratégicas que exigem análise e conhecimento do negócio da empresa.

De acordo com THOMSEN (2002), um resumo das diferenças entre as atividades dos sistemas OLPT e SAG pode ser definido de acordo com a tabela abaixo:

|  |  |
| --- | --- |
| **OLPT** | **SAG** |
| Acessado frequentemente | Acessado com menos freqüência |
| Operações previsíveis | Operações menos previsíveis |
| Pequena quantidade de dados carregados a cada consulta | Grande quantidade de dados carregados a cada consulta |
| Consulta principalmente dados brutos | Consulta principalmente dados sumarizados |
| Utilização de dados recentes | Utilização de dados históricos, recentes e projeções para o futuro |
| Pouca ou nenhuma derivação complexa dos dados | Muitas derivações complexas dos dados |

Table 1 - Operações sistemas OLPT e SAG (Fonte olap solutions, thomsen pag 13)

De acordo com INMON (1997), são apresentadas algumas diferenças no que se diz respeito aos dados entre os sistemas OLPT e SAG na tabela abaixo:

|  |  |
| --- | --- |
| **OLPT** | **SAG** |
| Dados primitivos/operacionais | Dados derivados |
| Baseados em aplicações | Baseados em assuntos de negócio |
| Detalhados | Resumidos |
| Exatos em relação ao momento do acesso | Representam valores históricos |
| Atendem a comunidade funcional | Atendem a comunidade gerencial |
| Podem ser atualizados | Não são atualizados |
| Processados repetitivamente | Processados de forma huerística |
| Performance fundamental | Performance atenuada |
| Acessado uma unidade por vez | Acessado um conjunto por vez |
| Voltados para transações | Voltados para análise |
| Pequena quantidade de dados utilizada em um processo | Grande quantidade de dados utilizado em um processo |
| Atendem às necessidades diárias | Atendem às necessidades gerenciais |
| Alta probabilidade de acesso | Baixa probabilidade de acesso |

Table 2 - Diferencas OLPT e SAG (adaptacao, Fonte Inmon- como construir data warehouse pag 18)

## Data Warehouse

### Banco de dados

Como foi apresentado anteriormente, uma dos componentes de um sistema de informação é a tecnologia de armazenamento dos dados. Todo sistema de informação possui uma forma de armazenar seus dados.

O início do mundo computacional consistia em aplicações individuais executadas em arquivos mestres que por sua vez eram armazenados em fitas magnéticas sendo acessadas de maneira seqüencial. Com o aumento significativo destas aplicações surgiram vários problemas como: sincronização dos dados, manutenção dos programas e grande quantidade de fita necessária. As aplicações evoluíram mas continuavam cada uma a ter seu próprio arquivo de dados. Com evolução da tecnologia de armazenamento de dados chegou-se a uma abordagem denominada banco de dados, onde uma grande quantidade de dados é armazenada em um mesmo local para ser acessado por diferentes programas.

Fazer figura no EA – turban pagina 135

Figure 1 - ambiente tradicional dos arquivos

FAzer figura no EA turban pagina 137

Figure 2 – Sistema gerenciador de banco de dados

Segundo LAUDON (1998), banco de dados pode ser definido como ”uma coleção de dados organizados de tal forma que possam ser acessados e utilizados por muitas aplicações diferentes”. Desta maneira os dados ficam centralizados minimizando a redundância, isolamento e inconsistência como acontecia no sistema de armazenamento em arquivos.

De acordo com TURBAN (2003), “um banco de dados, que é um grupo lógico de arquivos relacionados, pode evitar vários problemas associados a um ambiente tradicional de arquivos”. O banco de dados possui um sistema que gerencia o acesso aos arquivos de dados chamado SGBD (Sistema de gerenciamento de banco de dados), que funciona como uma interface entre o banco de dados e as outras aplicações que necessitam acessar os dados.

### Chegando ao Data Warehouse

Com o surgimento dos sistemas SAG, houve a necessidade de armazenar uma grande quantidade de dados sumarizados, históricos com grande valor de informação para realização de análises auxiliando nas decisões gerenciais das organizações. Estes dados são então armazenados em uma base de dados diferente dos sistemas OLPT denominada data warehouse.

Segundo INOMN (1997), o data warehouse é a base principal dos sistemas SAG e pode ser definido como “um conjunto de dados baseado em assuntos, integrado, não-volátil, e variável em relação ao tempo, de apoio às decisões gerenciais.

### Objetivos

De acordo com KIMBALL (2002), um data warehouse possui alguns objetivos descritos abaixo.

O data warehouse deve proporcionar fácil acesso às informações da organização com dados de fácil entendimento, intuitivos e claros. Ferramentas de acesso devem ser de fácil usabilidade, devem possibilitar o usuário realizar diferentes combinações dos dados e devem retornar resultados claros e com consultas rápidas.

As informações disponibilizadas pelo data warehouse devem ser confiáveis e concretas. Os dados devem consistentes e estarem disponíveis para a empresa apenas após serem filtrados e revisados.

O data warehouse deve estar preparado para mudanças pois as necessidades das áreas de negócio da empresa estão sempre sujeitas a mudanças.

O data warehouse deve ser seguro, pois é onde todos os dados e informações vitais das organizações estão armazenados.

O data warehouse dever dispor de informações que ajudem os gestores na tomada de decisão. Segundo TURBAN (2003), o data warehouse “é um sistema de gerenciamento de banco de dados multidimensional ou relacional, elaborado para oferecer suporte para a tomada de decisões de gerenciamento”.

Além de todos os objetivos acima, o data warehouse deve ser deve ser aceito pelos diretores e gerentes das organizações, pois serão os principais usuários do sistema.

### Características

Uma das principais características é que o data warehouse é baseado em assuntos de negócios de uma organização. Enquanto uma base de dados de um sistema operacional armazena dados sobre produtos, lojas e registros de vendas por exemplo, um data warehouse armazena dados relacionados à histórico de vendas e clientes.

De acordo com INMON (1997) uma das principais características é que o data warehouse deve ser integrado, pois os dados são armazenados de forma diferente nos sistemas operacionais de uma organização e ao serem transferidos para o data warehouse, estes dados devem ser transformados e as inconsistências devem ser removidas e se tenha um único padrão estabelecido. Um exemplo seria uma data em um formato dia-mês-ano e no formato ano-mês-dia em diferentes sistemas operacionais.

Outra característica consistem em que o data warehouse é não volátil. Afirma MACHADO (2010) os dados são mantidos para se ter um histórico e que após a carga dos dados o data warehouse possui apenas operações de consultas, ao contrário dos banco de dados operacionais onde são efetuadas com grande freqüência operações de consulta, inserção, remoção e atualização. Por este motivo pode-se afirmar que o data warehouse é a memória da organização.

O data warehouse é variável em relação ao tempo. Afirma INMON (1997) que “o horizonte de tempo válido para o data warehouse é significativamente maior do que o dos sistemas operacionais”. Os dados em um data warehouse representam dados operacionais em um determinado momento.

### Componentes

O data warehouse possui os seguintes componentes: sistemas operacionais de origem, data staging area, dados apresentáveis e ferramentas de acesso aos dados apresentados.

--Figura do Kimball pagina 9

**Sistemas operacionais (OLPT) ou sistemas legados**

Segundo KIMBALL (2002) os sistemas operacionais priorizam o desempenho e disponibilidade de processamento. Como vimos anteriormente os sistemas operacionais mantém um volume pequeno de dados históricos que registram e capturam as transações da empresa. Os dados possuem formatos diferentes neste sistemas. Todos os dados para compor o data warehouse são provenientes dos sistemas operacionais da empresa.

**Data Staging Area ou Operational Data Storage (ODS)**

Área que abrange tudo desde a extração dos dados dos sistemas legados, passando pela transformação dos dados e a transferência dos dados para a apresentação. Segundo MACHADO (2010), “representa um armazenamento intermediário dos dados, facilitando a integração dos dados do ambiente operativo antes da sua atualização no data warehouse”.

KILBALL (2002), afirma que a data staging area é composta por uma área de armazenamento e também por processos de extração, transformação e carga dos dados no ambiente do data warehouse chamado de ETL (Extract-Trasnformation-Load). Afirma também que a data staging area não deve ser acessada pelos usuários do data warehouse, pois os dados ainda não estão no formato adequado para consulta.

**Dados apresentáveis ()**

**Ferramentas de acesso aos dados**

...

## Modelagem dimensional

...

## ETL

...

## Ferramentas OLAP

...