

SISTEMAS DE APOIO À INTELIGÊNCIA DE NEGÓCIOS

Asterio K. Tanaka

<http://www.uniriotec.br/~tanaka/SAIN>
tanaka@uniriotec.br



Projeto e Construção de Data Warehouse

baseado no Capítulo 16. Building the Data Warehouse
de Kimball, Ralph; Ross, Margy. The Data Warehouse Toolkit. John Wiley, 2002.
Agosto de 2007

- Ciclo de Vida Dimensional do Negócio
- Planejamento e Gerência de Projeto
- Definição dos Requisitos do Negócio
- Trilha Tecnológica do Ciclo de Vida
 - Projeto da Arquitetura Técnica
 - Seleção de Produtos e Instalação
- Trilha de Dados do Ciclo de Vida
 - Modelagem Dimensional
 - Projeto Físico
 - » Estratégia de Agregação
 - » Estratégia Inicial de Indexação
 - Projeto e Desenvolvimento de Data Staging
 - » Dimensões
 - » Fatos
- Trilha de Aplicações Analíticas do Ciclo de Vida
- Implantação (Deployment)
- Manutenção e Expansão
- Dez Erros Comuns a Evitar em Projetos de DW

Ciclo de Vida Dimensional do Negócio

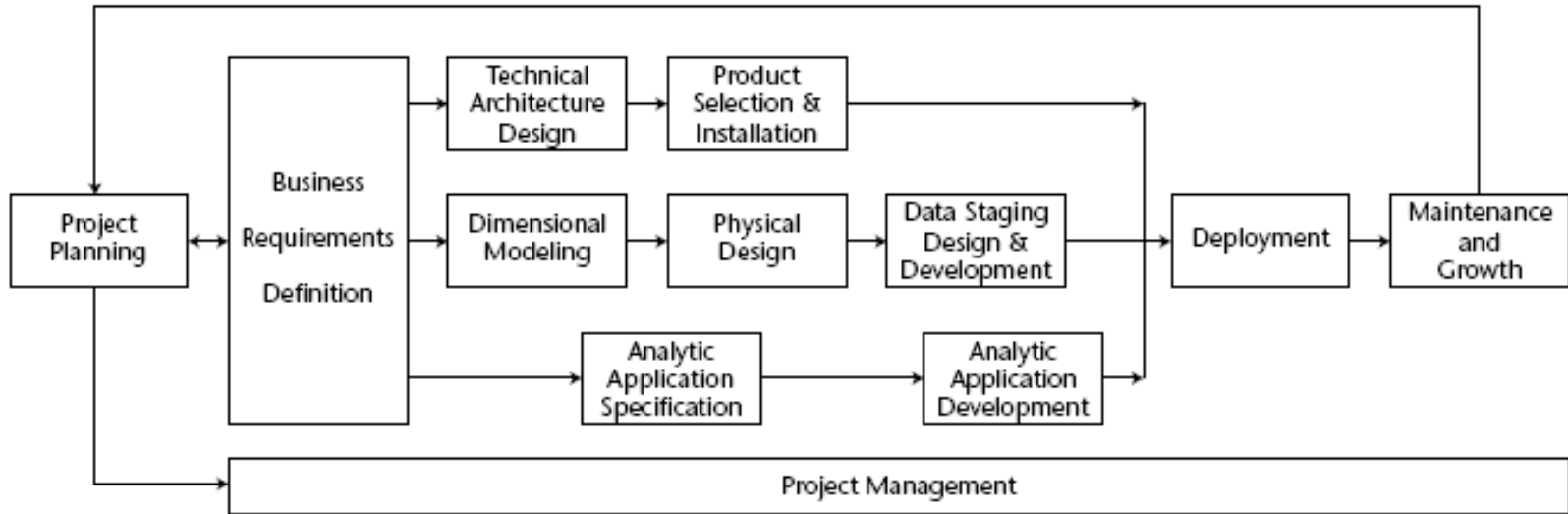


Figure 16.1 Business dimensional lifecycle diagram.

O diagrama acima encapsula as atividades mais importantes do ciclo de vida dimensional do negócio. O diagrama ilustra sequência, dependência e concorrência de tarefas. Serve como um road map para ajudar as equipes a fazer a coisa certa na hora certa. O diagrama não reflete uma linha de tempo absoluta. Embora as caixas sejam de mesmo tamanho, há uma vasta diferença em tempo e esforço requerido para cada atividade importante.

Planejamento e Gerência de Projeto (1)

- **Cinco fatores de sucesso (em ordem de importância)**
 - Patrocinador forte do negócio (visionário e convencido da importância do DW) – de preferência, mais de um patrocinador
 - Motivação organizacional compelidora
 - Viabilidade técnica e de recursos, mas principalmente viabilidade de dados
 - Relacionamento entre as organizações do negócio e da TI
 - Cultura analítica atual da organização (para tomada de decisões)
- **Se o projeto não estiver pronto para começar, tipicamente por falta de patrocínio forte**
 - Análise e priorização de requisitos do negócio (alto nível, BPM)
 - ou
 - Prova de conceito (demonstração das capacidades potenciais do DW)

Planejamento e Gerência de Projeto (2)

- **Escopo**

- Uma vez confortável com o “pronto” da organização, é hora de estabelecer os limites de um projeto inicial.
- Escopo requer input da TI e do negócio; deve ser significativo para o negócio e gerenciável para a TI.
- Evitar a “lei do muito”: compromisso muito firme, com prazo muito curto, envolvendo muitos sistemas fontes e muitos usuários em muitos locais, com requisitos de análise muito diversos.

- **Justificativa**

- Estimativa de benefícios e custos associados com o DW.
- TI responsável pelos custos (hardware, software, esforço de desenvolvimento)
- A área do negócio deve determinar os benefícios
- Se você tiver dificuldade em estimar os benefícios, é um sintoma de que está focado em patrocinador ou problema errado.

Planejamento e Gerência de Projeto (3)

- **Staffing (vide atribuições no livro)**

“Um plano é tão bom quanto as pessoas que irão implementá-lo.”

- **Papéis do lado do negócio**

- » Patrocinador ou Representante do patrocinador (em grandes organizações);
- » Líder de projeto do negócio (em geral é o próprio patrocinador ou representante, se disponível para interação diária com o gerente do projeto);
- » Usuários do negócio.

- **Recursos técnicos que entendam do negócio ou recursos do negócio que entendam de tecnologia**

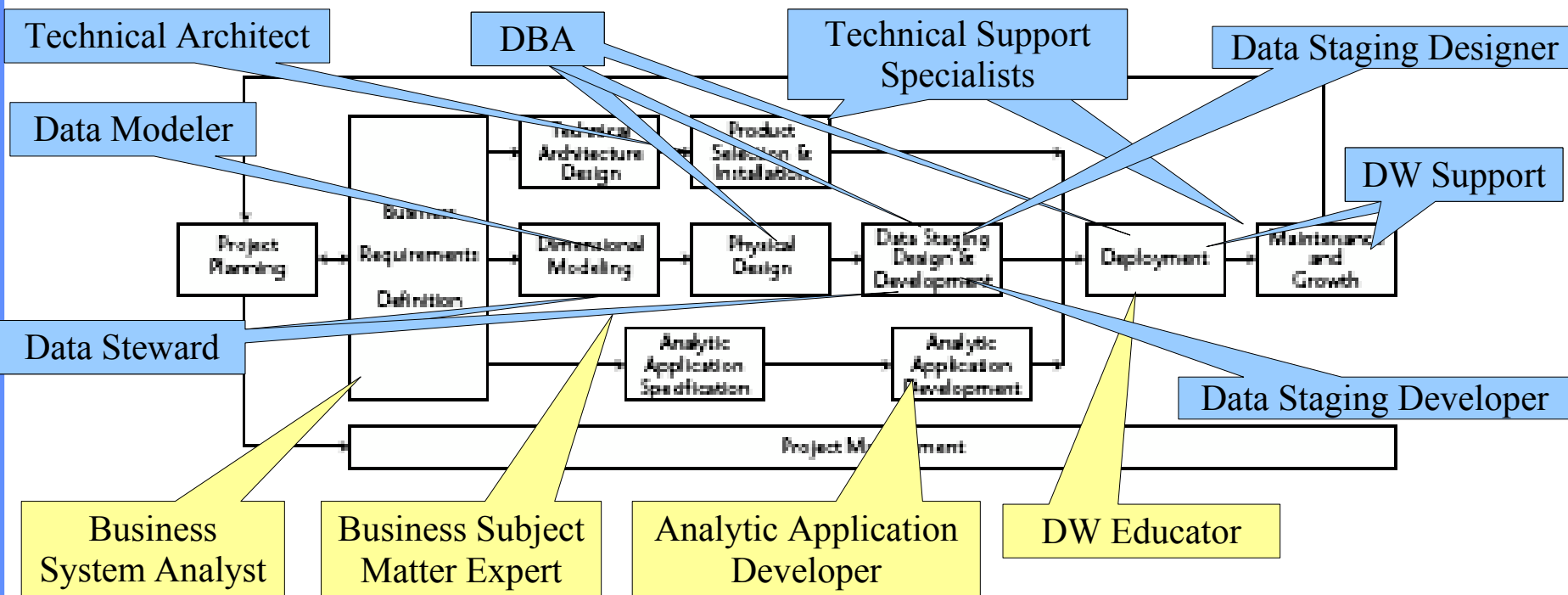
- » Analista de sistemas do negócio;
- » Especialista do assunto do negócio;
- » Desenvolvedor de aplicação analítica;
- » Educador de data warehouse

- **Papéis do lado da organização de TI ou consultoria externa**

- » Gerente do projeto;
- » Arquiteto técnico;
- » Especialistas de suporte técnico;
- » Modelador de dados;
- » Administrador de Banco de Dados;
- » Coordenador de metadados;
- » Administrador de dados;
- » Projetista de “data staging”;
- » Desenvolvedor de “data staging”;
- » Suporte do Data Warehouse

Pessoas no Ciclo de Vida Dimensional

É comum que uma pessoa exerça mais de um papel, em uma fase ou ao longo de todo o ciclo de vida do projeto. A alocação de pessoas aos papéis depende da maturidade da organização, da magnitude e do escopo do projeto, e disponibilidade, capacidade e experiência dos indivíduos.



Necessariamente, ao longo de todo o ciclo de vida:

Business Driver, Business Project Lead, Project Manager, Metadata Coordinator

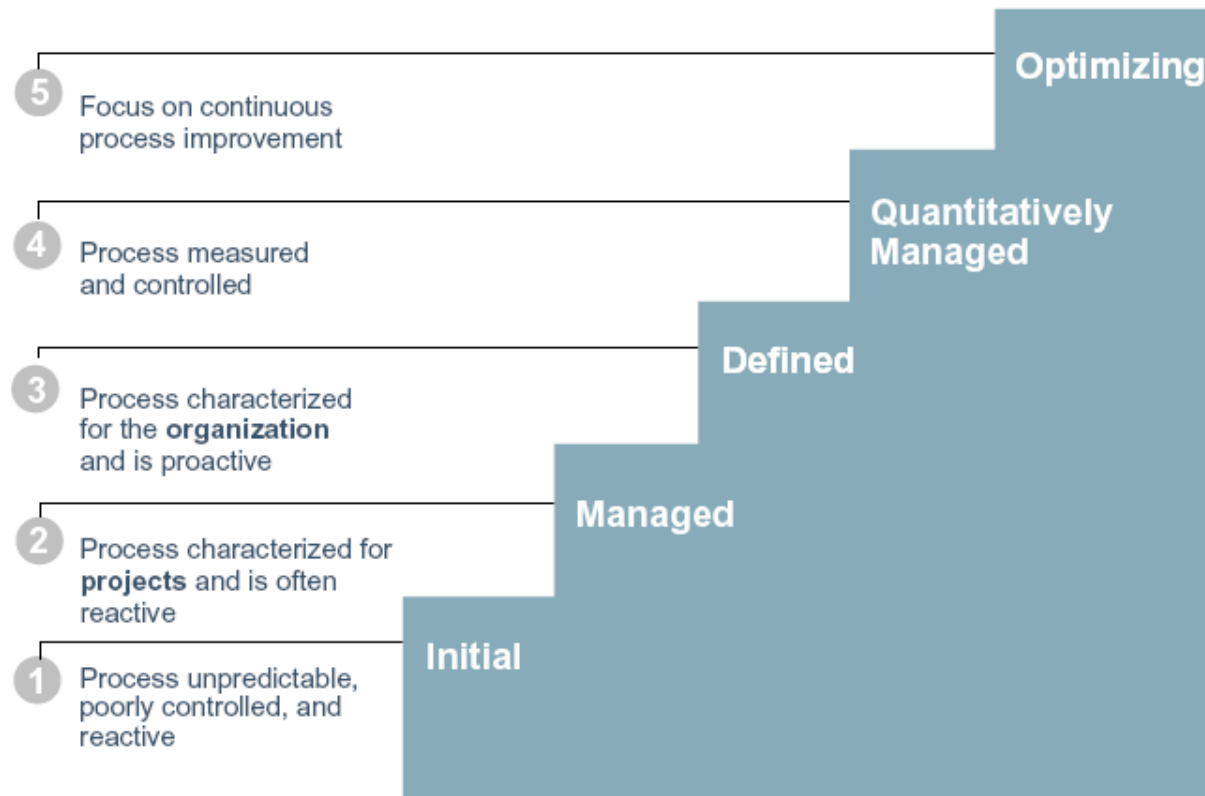
Desejavelmente, ao longo de todo o ciclo de vida:

Business Sponsor, Business Users

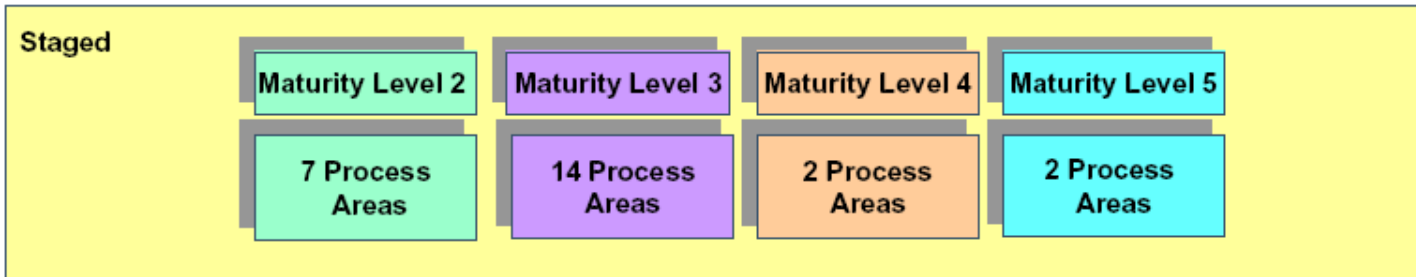
Relacionando Recursos Humanos à Maturidade das Organizações

Organizações maduras	Organizações imaturas
Papéis e responsabilidades bem definidos	Processo improvisado
Existe base histórica	Não existe base histórica
É possível julgar a qualidade do produto	Não há maneira objetiva de julgar a qualidade do produto
A qualidade dos produtos e processos é monitorada	Qualidade e funcionalidade do produto sacrificadas
O processo pode ser atualizado	Não há rigor no processo a ser seguido
Existe comunicação entre o gerente e seu grupo	Resolução de crises imediatas

The Maturity Levels



Staged View of CMMI



- Requirements Management
- Project Planning
- Project Monitoring and Control
- Supplier Agreement Management
- Measurement and Analysis
- Process and Product Quality Assurance
- Configuration Management

- Requirements Development
- Technical Solution
- Product Integration
- Verification
- Validation
- Organizational Process Focus
- Organizational Process Definition

- Organizational Training
- Integrated Project Management
- Risk Management
- Integrated Teaming
- Integrated Supplier Management
- Decision Analysis and Resolution
- Organizational Environment for Integration

- Organizational Process Performance
- Quantitative Project Management

- Organizational Innovation and Deployment
- Causal Analysis and Resolution

CMM sob a perspectiva dos dados

Nível 1 - O nível inicial

- Não há regras ou procedimentos sobre gerência de dados; múltiplos arquivos e bancos de dados; múltiplos formatos; redundância; mudanças “on the fly”; geralmente não há um grupo de gerência de dados; caso exista, apenas aplicam as mudanças requisitadas de acordo com as necessidades.
- Qualidade de dados depende das habilidades do pessoal de desenvolvimento. Frequentemente, projetos são cancelados ou, pior ainda, produzem dados incorretos e relatórios inválidos.
- Aproximadamente 30 a 50% das organizações operam no Nível 1.

(Craig S. Mullins, The Data Administration Newsletter, 19??)

CMM sob a perspectiva dos dados

Nível 2 - O nível repetível

- Política de gerência de dados: como e quando as estruturas de dados são criadas, mudadas e gerenciadas. Em geral, a política não é institucionalizada, ficando nas mãos de um grupo ou pessoa, com função de ABD.
- O sucesso depende das habilidades dos ABDs. Porém, não há muito esforço em documentar e capturar o significado de negócio dos dados. Pouca diferenciação entre modelos lógicos e físicos. Preocupação maior com os aspectos técnicos da gerência de dados.
- Aproximadamente 15% a 20% das organizações operam no Nível 2.

(Craig S. Mullins, The Data Administration Newsletter, 19??)

CMM sob a perspectiva dos dados

Nível 3 - O nível definido

- Política de gerência de dados documentada e estabelecida como um componente central do ciclo de desenvolvimento de aplicações. Há um entendimento do significado de negócio dos dados, através de uma função de AD. Dados são tratados como recursos corporativos.
- O sucesso depende da interação das funções de AD e ABD e do uso apropriado de ferramentas, para criar modelos de dados, automatizar processos de ABD, e pro-ativamente monitorar e sintonizar o desempenho dos BDs.
- Aproximadamente 10% a 15% das organizações operam no Nível 3.

(Craig S. Mullins, The Data Administration Newsletter, 19??)

CMM sob a perspectiva dos dados

Nível 4 - O nível gerenciado

- Gerenciamento de metadados permite ao grupo de gerência de dados (AD e ABD) catalogar e manter metadados de estruturas de dados corporativas. Sabe-se que dados existem onde. O grupo envolve-se em todos os esforços de desenvolvimento (no nível lógico sempre; no nível físico quando necessário).
- O sucesso depende do suporte da alta administração à máxima “dados são recursos corporativos”. Ferramentas avançadas são necessárias para gerenciar metadados, qualidade de dados e as bases de dados.
- Aproximadamente 5% a 10% das organizações operam no Nível 4.

(Craig S. Mullins, The Data Administration Newsletter, 19???)

CMM sob a perspectiva dos dados

Nível 5 - O nível “otimizado”

- Organizações neste nível usam as práticas desenvolvidas nos níveis anteriores para continuamente melhorar o acesso a dados, qualidade de dados e desempenho dos bancos de dados. Nenhuma mudança é introduzida sem primeiro ser escrutinada pelo grupo de gerência de dados e documentada no repositório de metadados.
- Deve haver uma preocupação contínua na melhoria do processo de gerência de dados com uso de ferramentas avançadas.
- Menos de 5% das organizações operam no Nível 5.

(Craig S. Mullins, The Data Administration Newsletter, 19??)

Sobre Administração de Dados e Metadados

- ABD: função operacional (técnica)
- AD: função tática (gerencial)
- CTO/CIO: funções estratégicas (direção)
- Não basta administrar bancos de dados; mais que isso, é fundamental administrar os dados e metadados da organização.
 - Gerência de Dados/Metadados →
 - Gerência de Informações →
 - Gerência de Conhecimento →
 - Inteligência Competitiva

Planejamento e Gerência de Projeto (4)

- **Desenvolvendo e mantendo o plano do projeto**
 - Envolve identificar todas as tarefas necessárias para implementar o DW (*The Data Warehouse Lifecycle Toolkit* inclui uma lista de cerca de 200 tarefas).
- **As chaves para o planejamento e gerência de projeto de DW incluem:**
 - 1. Ter um patrocinador de negócio sólido
 - 2. Equilibrar alto valor e viabilidade para definir o escopo
 - 3. Trabalhar com a melhor equipe possível para desenvolver um plano de projeto detalhado
 - 4. Ser um excelente gerente de projeto para motivar, gerenciar e comunicar para cima, para baixo e através da organização

Definição dos Requisitos do Negócio (1)

- **Pré-planejamento do projeto**

- **Escolher o fórum**

- » Técnicas primárias: entrevistas e sessões facilitadas
 - » Pesquisa não é ferramenta razoável porque respondentes só respondem o que foi perguntado
 - » Melhor uma abordagem híbrida

- **Identificar e preparar a equipe de requisitos**

- » Entrevistador líder + “escrivão” + pessoa para “gravar” sem gravador
 - » Um ou dois membros da equipe de projeto dependendo do número de entrevistados
 - » Dever de casa da equipe: pesquisar sobre o assunto

- **Selecionar, agendar e preparar os representantes do negócio**

- » Cobertura horizontal de pessoas do negócio através da organização, para permitir a formulação da matriz de DW.
 - » Cobertura vertical para garantir patrocínio estratégico ao longo do levantamento de requisitos.

Definição dos Requisitos do Negócio (2)

- **Coletar os requisitos do negócio**
 - **Lançamento**
 - » Introdução deve ser uma mensagem centrada no negócio, não em tecnologia.
 - **Fluxo de entrevistas**
 - » Lembre o seu papel na entrevista; ouça e absorva como uma esponja
 - » Mantenha um fluxo conversacional; não mergulhe muito rapidamente (ou tire cópias de elementos de dados potenciais)
 - » Verifique terminologia de comunicação e captura precisamente porque a maioria das organizações usa terminologia inconsistentemente
 - » Estabeleça uma parceria com o entrevistado; use o vocabulário dele(a)
 - **Wrap up**
 - » Cada entrevistado deve ser perguntado sobre o seu critério de sucesso para o projeto. Cada critério deve ser mensurável.
 - **Conduzindo entrevistas centradas em dados**
 - » Enquanto se foca no entendimento dos requisitos de negócio, é importante intercalar sessões com especialistas de dados de sistemas fontes, para avaliar a viabilidade dos requisitos de negócio.
- **Documentação pós-coleta e follow up**
 - **Documentação não é atividade favorita, mas é crucial.**
 - **Prioritização e consenso**

Definição dos Requisitos do Negócio (3)

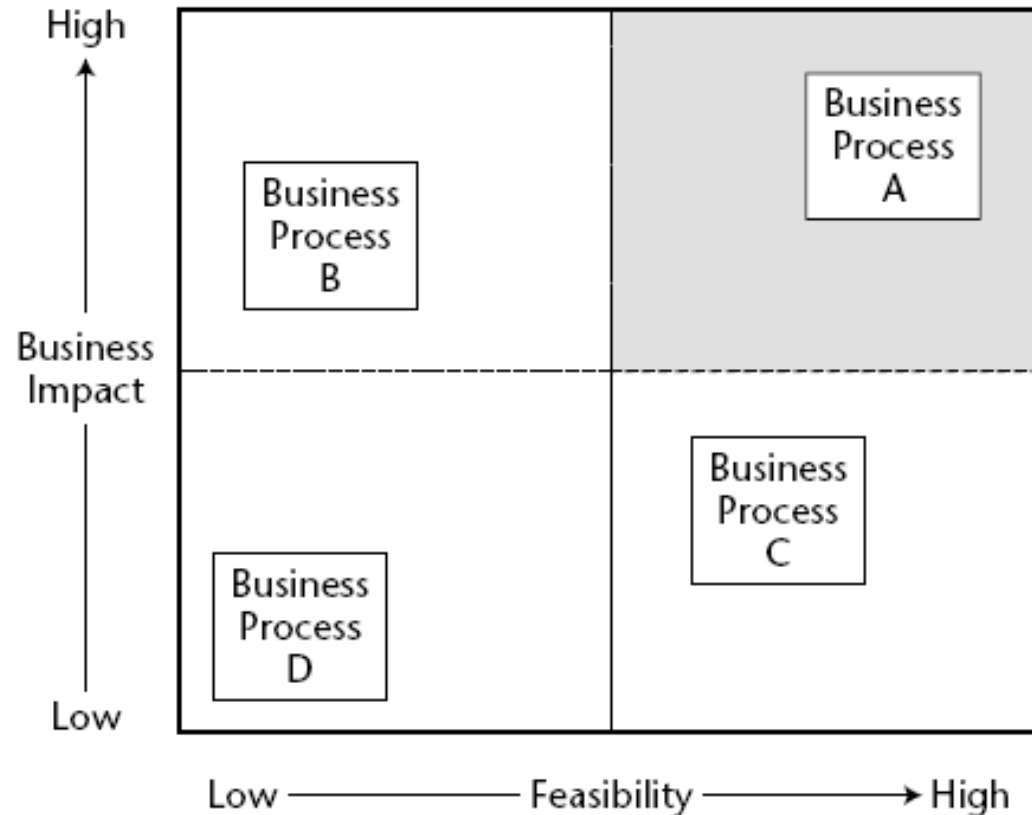


Figure 16.2 Prioritization quadrant analysis.

Projetos que merecem atenção imediata são os de alto impacto no negócio. Enquanto a equipe de projeto de DW estiver focado no processo de negócio A, outras equipes de TI devem estar estudando as limitações de viabilidade do processo de negócio B.

Trilha Tecnológica do Ciclo de Vida (1)

- **Projeto da arquitetura técnica**
 - **Processo de Oito Passos para criar a Arquitetura Técnica**
 - » Estabeleça uma força tarefa de arquitetura
 - » Colete requisitos relacionados a arquitetura
 - » Documente requisitos de arquitetura
 - » Desenvolva um modelo arquitetural de alto nível
 - » Projete e especifique os subsistemas
 - » Determine as fases de implementação da arquitetura
 - » Documente a arquitetura técnica
 - » Revise e finalize a arquitetura técnica

Trilha Tecnológica do Ciclo de Vida (2)

- **Seleção de Produtos e Instalação**
 - Entenda o processo de compras corporativas
 - Desenvolva uma matriz de avaliação de produtos
 - Conduza uma pesquisa de mercado
 - Reduza as opções a uma lista pequena e realize avaliações detalhadas
 - Conduza uma prototipação, se necessário
 - Selecione um produto, instale um trial e negocie

Trilha de Dados do Ciclo de Vida (1)

- **Modelagem Dimensional**

- Seguindo a definição dos requisitos de negócio, desenhar ou visitar a matriz de barramento de DW
- Escolher uma linha da matriz implica no passo 1 do roteiro de 4 passos:
 - 1) Selecionar o processo de negócio
 - 2) Declarar o grão do processo de negócio
 - 3) Escolher as dimensões
 - 4) Identificar os fatos
- Os demais passos devem ser atacados com as técnicas de modelagem dimensional já vistas.

Requisitos do negócio

-
- 1. Processo de negócio
 - 2. Grão
 - 3. Dimensões
 - 4. Fatos

Realidade dos dados

Dez Erros Comuns a Evitar em Modelagem Dimensional (1)

- **Erro 10:** Colocar atributos de texto usados para restrições e agrupamento numa tabela de fatos.
- **Erro 9:** Limitar atributos descritivos verbosos em dimensões para economizar espaço.
- **Erro 8:** Separar hierarquias e níveis de hierarquia em dimensões múltiplas.
- **Erro 7:** Ignorar a necessidade de cuidar de mudanças em atributos de dimensões.
- **Erro 6:** Resolver todos os problemas de desempenho de consultas adicionando mais hardware.

Dez Erros Comuns a Evitar em Modelagem Dimensional (2)

- **Erro 5:** Usar chaves operacionais ou “inteligentes” para junções de tabelas de dimensão com tabela de fatos.
- **Erro 4:** Negligenciar a declaração e depois a consistência com o grão da tabela de fatos.
- **Erro 3:** Projetar o modelo dimensional baseado em um relatório específico.
- **Erro 2:** Esperar que usuários consultem dados de nível atômico mais baixo num formato normalizado.
- **Erro 1:** Falhar em conformar fatos e dimensões através de diferentes data marts.

Trilha de Dados do Ciclo de Vida (2)

- **Projeto Físico**

- **Estratégia de Agregação**

- » Todo DW deve conter tabelas de agregação pré-calculadas e pré-armazenadas. Dadas as regras estritas sobre evitar granularidade misturada em tabela de fatos, cada agregação distinta de tabela de fatos deve ocupar sua própria tabela de fatos física.
 - » Quando agregamos fatos, ou eliminamos dimensionalidade ou associamos os fatos a uma dimensão maior (roll up). Essas tabelas de dimensão agregadas deveriam ser versões encolhidas das dimensões associadas com a tabela de fatos granular. Desta forma, tabelas de dimensão agregadas conformam com as tabelas de dimensão base.
 - » Como é impraticável armazenar todas as combinações possíveis de agregações, é preciso considerar fatores para a estratégia de agregação:
 - Padrões de acesso de usuários do negócio
 - Distribuição estatística dos dados
 - » Estratégia depende de ferramentas de navegação em agregados.
 - » Regra comum: espaço reservado para tabelas agregadas deveria ocupar duas vezes o espaço em disco consumido pela tabela base.

Trilha de Dados do Ciclo de Vida (3)

- **Projeto Físico**

- **Estratégia Inicial de Indexação**

- » Administradores de BD ficam aflitos quando sabem que tabelas de dimensão frequentemente têm mais de um índice apenas.
 - » Tabelas de dimensão têm um índice sobre a chave primária e, adicionalmente, recomenda-se um índice em árvore B sobre colunas de alta cardinalidade usada para restrições. Índices bit map devem ser colocados sobre todos os atributos com média e baixa cardinalidade.
 - » Por outro lado, tabelas de fatos são os elefantes do DW, portanto temos que indexá-las com mais cuidado. A chave primária da tabela de fatos é quase sempre um subconjunto das chaves estrangeiras. Tipicamente colocamos um único índice concatenado sobre as dimensões primárias da tabela de fatos. Como muitas consultas dimensionais são condicionadas sobre a dimensão Data, a chave estrangeira de Data deve ser a cabeça do índice.
 - » Estratégia de indexação depende do SGBD, e não é muito diferente de outros tipos de bancos de dados.
 - » Tabelas de fatos grandes geralmente são particionadas por Data, com os dados segmentados por Mês, Trimestre, Ano em diferentes partições, embora apareçam aos usuários como uma única tabela.

Trilha de Dados do Ciclo de Vida (3)

- **Projeto e Implementação de Data Staging**

- **Data Staging de Dimensões**

- 1) Extrair dados dimensionais dos sistemas fonte operacionais**

- Extração para um arquivo de saída, acompanhado de estatísticas para auditoria.

- 2) Limpar valores de atributos**

- Situações a tratar: validação de nomes e endereços, valores descritivos inconsistentes, decodificações faltando, códigos sobrecarregados com significados múltiplos no tempo, dados inválidos, dados faltando.

- 3) Gerenciar atribuições de chaves surrogate**

- Deve ser mantida uma tabela de correspondência mestre na área de Data Staging chave surrogate – chave operacional, com outros dados de tempo e perfil (vide esquema na figura 16.3)

- 4) Gerar imagens para carga de linhas de dimensões e publicar dimensões revisadas**

- Uma vez que a tabela de dimensão reflita a extração mais recente e devidamente revisada, ela é publicada para todos os data marts que utilizam essa dimensão.

Tabela de Correspondência para controle de chave surrogate

Master Dimension Cross-Reference Table	
Surrogate Dimension Key	← If combining data from multiple sources, there would be additional columns for the other operational sources.
Operational Source Key	
Dimension Attributes 1-N	
Dimension Row Effective Date	
Dimension Row Expiration Date	
Most Recent Dimension Row Indicator	
Most Recent Cyclic Redundancy Checksum (CRC)	

Figure 16.3 Fields for the Staging Master Dimension Cross-Reference Table

Trilha de Dados do Ciclo de Vida (4)

- **Projeto e Implementação de Data Staging**
 - **Data Staging de Tabelas de Fatos (vide detalhes no livro)**
 - 1) **Extrair dados de fatos do sistema fonte operacional**
 - 2) **Receber dimensões atualizadas das autoridades de dimensões**
 - 3) **Separar os dados de fatos por granularidade conforme requerido**
 - 4) **Transformar os dados de fatos conforme requerido**
 - 5) **Substituir chaves operacionais por chaves surrogate**
 - 6) **Adicionar chaves adicionais para contextos conhecidos**
 - 7) **Assegurar qualidade dos dados da tabela de fatos**
 - 8) **Construir ou atualizar tabelas de fatos de agregação**
 - 9) **Carregar os dados**
 - 10) **Alertar os usuários**

Trilha de Aplicações Analíticas do Ciclo de Vida

- **Especificação de aplicações analíticas**
- **Desenvolvimento de aplicações analíticas**

Implantação (Deployment)

- **Consider the following for an effective education program:**
 - Understand your target audience; don't overwhelm.
 - Don't train the business community early prior to the availability of data and analytic applications.
 - Postpone the education (and deployment) if the data warehouse is not ready to be released.
 - Gain the sponsor's commitment to a "no education, no access" policy.

Manutenção e Expansão

- **Suporte**
- **Educação**
- **Suporte técnico**
- **Suporte de programa**

Dez Erros Comuns a Evitar em Projetos de Data Warehouse (1)

- **Erro 10:** Aceitar a premissa de que os responsáveis pelos sistemas fontes mais relevantes da organização são muito importantes e ocupados para gastar tempo com a equipe de DW.
- **Erro 9:** Após a equipe de DW ter sido acordada, marcar uma reunião para discutir comunicações com os usuários de negócio, se o orçamento permitir.
- **Erro 8:** Assegurar para o pessoal de suporte do DW escritórios agradáveis no prédio da TI, que fica próximo dos usuários de negócio, e providenciar um número de telefone de suporte de DW com várias opções de menu.
- **Erro 7:** Treinar cada usuário em cada característica da ferramenta de acesso a dados na primeira aula de treinamento, adiar o treinamento sobre conteúdo de dados porque a aula usa dados falsos (os dados reais não estarão prontos nos próximos dois meses) e declarar sucesso ao término da primeira aula de treinamento já que o DW foi disponibilizado para os usuários de negócio.

Dez Erros Comuns a Evitar em Projetos de Data Warehouse (2)

- **Erro 6:** Assumir que os usuários de negócio vão naturalmente gravitar em direção a dados robustos e desenvolver suas próprias killer applications analíticas.
- **Erro 5:** Antes de implementar o DW, fazer uma análise completa descrevendo todos os possíveis ativos de dados da empresa e todos os usos desejados de informação, e evitar a ilusão sedutora de desenvolvimento iterativo, que é somente uma desculpa para não fazer certo da primeira vez.
- **Erro 4:** Não aborrecer os executivos seniors de sua organização com o DW até que você tenha o implementado e possa apontar para um sucesso significativo.

Dez Erros Comuns a Evitar em Projetos de Data Warehouse (3)

- **Erro 3:** Encorajar os usuários de negócio a lhe dar feedback contínuo ao longo do ciclo de desenvolvimento sobre novas fontes de dados e métricas chaves de desempenho que eles gostariam de acessar, e assegurar a inclusão desses requisitos na release em desenvolvimento.
- **Erro 2:** Concordar em entregar um data mart centrado em cliente de alto perfil, idealmente lucratividade de cliente ou satisfação de cliente, como seu primeiro produto.
- **Erro 1:** Não conversar com os usuários de negócio; ao invés disso, confiar em consultores ou especialistas internos para lhe dar interpretação dos requisitos de usuários do DW.