fumsoft

Uma visão sintética e comentada do Data Management Body of Knowledge (DMBOK)

Carlos Barbieri
com colaboração de Fernanda Farinelli

Uma visão sintética e comentada do Data Management Body of Knowledge (DMBOK)

Carlos Barbieri

Com colaboração de Fernanda Farinelli

Belo Horizonte

Janeiro de 2013

Versão 01

fumsoft

FICHA TÉCNICA

Autor
Carlos Barbieri

Colaboração
Fernanda Farinelli (PRODEMGE)

Equipe técnica
Isabella Fonseca (FUMSOFT)
Claudio Filardi (FUMSOFT)
Evilene Santos (FUMSOFT)

Editoração
Pedro Ivo Brandão (FUMSOFT)
Renata Ferreira (FUMSOFT)

Projeto gráfico
Gracielle Santos (FUMSOFT)

Barbieri, Carlos.

Uma visão sintética e comentada do Data Management Body of Knowledge (DMBOK). Fumsoft - Belo Horizonte, 2013.

Outras informações podem ser obtidas pelo e-mail qualidade@fumsoft.org.br

Fumsoft
Av. Afonso Pena, 4.000, 3º andar - bairro Cruzeiro
CEP: 30.130-009 - Belo Horizonte/MG
Tel.: (31) 3281-1148
www.fumsoft.org.br

SUMÁRIO

1. UMA VISÃO SINTÉTICA E COMENTADA DO DMBOK

O objetivo desse trabalho é fornecer uma visão sintética sobre os corpos de conhecimentos do *Data Management Body of Knowledge* (DMBOK), adicionando aspectos práticos sobre gestão de dados desenvolvidos pelo autor nesta área por mais de 40 anos. Esse trabalho não tem a pretensão de substituir o documento DMBOK original, e intenciona servir somente de um guia mais rápido e comentado sobre as práticas daqueles corpos de conhecimentos. Para detalhes completos de conteúdo e de referências, os documentos DMBOK, tanto o original, quanto a sua edição brasileira, deverão ser consultados.

Esse trabalho surgiu nos cursos de pós-graduação ministrados pelo autor, originado da necessidade de se ter um texto menor e acessível aos alunos que ainda não dispunham (ou não dispõem) das referências originais. Além disso, incorpora comentários correlatos, percepções e visões do autor sobre o tema, que podem servir para o entendimento das interpretações realizadas sobre a pesquisa realizada pela *Data Management Association* (Dama) Brasil e pela Fumsoft, abordando a gestão estratégica de dados.

A Gestão de Dados (no inglês, *Data Management* ou DM), conforme o DMBOK (2009), visa controlar e alavancar eficazmente o uso dos ativos de dados e sua missão e objetivos são atender e exceder às necessidades de informação de todos os envolvidos (*stakeholders*) da empresa em termos de disponibilidade, segurança e qualidade. É uma responsabilidade tanto do setor de Tecnologia da Informação de uma empresa quanto de seus clientes internos e externos e envolve desde a alta direção, que utiliza dados na geração de informações estratégicas, até profissionais de nível operacional, que muitas vezes são responsáveis pela coleta e produção dos dados.

O DMBOK (2009) estrutura o processo de DM por meio de funções e atividades e está distribuído por dez áreas de conhecimento, conforme apresentado na Figura 1, a seguir.

Figura 1 - Áreas de conhecimento na Gestão de Dados, segundo o DMBOK

- ✓ Governança de dados
- ✓ Gerência da Arquitetura de dados
- ✓ Desenvolvimento de dados
- ✓ Gestão de operações de bancos de dados
- ✓ Gestão de Segurança de dados
- ✓ Gestão de Dados mestres e de Referência
- ✓ Gestão de Data Warehousing e BI
- ✓ Gestão de Documentos e conteúdo
- ✓ Gestão de Metadados
- ✓ Gestão de Qualidade de dados

1.1. Governança de dados

A definição de Governança de Dados (GD) é ampla e plural. É um conceito em evolução, que envolve o cruzamento de diversas disciplinas, com foco em qualidade de dados, passando por avaliação, gerência, melhoria, monitoração de seu uso, além de aspectos de segurança e privacidade associados a eles. Para tal, as empresas deverão definir objetivos organizacionais e processos institucionalizados, que deverão ser implementados dentro do equilíbrio fundamental entre TI e áreas de negócios. Através da GD, as empresas hoje também definem mecanismos para analisar os processos que se abastecem de ou produzem os dados, criando um sentido maior de qualidade conjunta entre esses dois elementos seminais (dados e processos) e contribuindo para a valorização desses ativos, através do pleno conhecimento da cadeia produtiva de informação e conhecimentos.

Segundo o DMBOK (2009), a Governança de Dados se divide em duas atividades macro, Planejamento e Controle da Gestão dos Dados, etc.

1.1.1. Planejamento da gestão de dados

O objetivo é:

- Entender as necessidades estratégicas de dados da empresa.
- Desenvolver e manter uma estratégia de dados.
- Estabelecer unidades organizacionais e papéis voltadas para dados.
- Identificar os Data Stewards.
- Estabelecer as camadas de GD e de data stewards.
- Desenvolver e aprovar Políticas, Padrões e Procedimentos de dados.
- Revisar e aprovar a Arquitetura de Dados.
- Planejar e patrocinar Projetos e Serviços de Gestão de Dados.
- Estimar o valor dos Ativos de Dados e custos associados (Riscos).

A visão sintética é:

a. Entender as necessidades estratégicas de dados:

Entender a estratégia e os dados necessários para apoiá-la. São evidentes questões como:

- Para onde vou (em temos de negócios), como vou, quando vou?
- Que dados serão necessários nesse caminho?
- Como obtê-los, como mantê-los?
- Como garantir a sua qualidade?
- Que áreas serão prioritárias no tratamento dos dados, baseado nas estratégias de negócios?
- Para que segmentos vamos caminhar? Big Data, *sentiment analysis* via redes sociais, etc.?

b. Desenvolver e manter a estratégia de dados:

Instanciar as ações para a obtenção dos dados, sua manutenção, sua qualidade, baseado nas necessidades estratégicas identificadas anteriormente.

c. Estabelecer unidades organizacionais e papéis para essas atividades de dados

Estruturas *in-business* (*data stewards*), estruturas in-TI (AD, DBA, etc.), estruturas táticas (CDO, DMO, gerencia os data stewards) e estruturas estratégicas (Comitê de GD, que define regras, tira dúvidas, resolve impasses, etc.).

d. Identificar os Data Stewards

Serão os responsáveis, dentro da área de negócios, pelo controle e uso dos dados. Nos usuários, seriam figuras com amplo domínio de conhecimento desses assuntos. Tomarão conta daquele recurso, serão os responsáveis (*liability*) e gerenciarão o seu uso.

e. Estabelecer as estruturas organizacionais (camadas) para Gestão de Dados e de data stewards

Enquanto no item ”c” há uma visão mais genérica, aqui há uma visão mais detalhe. Envolve Membros do Comitê de GD, principais Data Stewards, principais componentes do DMO, entre outros. Para as funções *in-business*, definir as áreas prioritárias (em função da estratégia) que deverão ter os seus stewards (gestores de dados). Há várias proposições possíveis de estruturas para GD. Abaixo, na figura 2, uma das proposições com as camadas e suas principais atribuições:

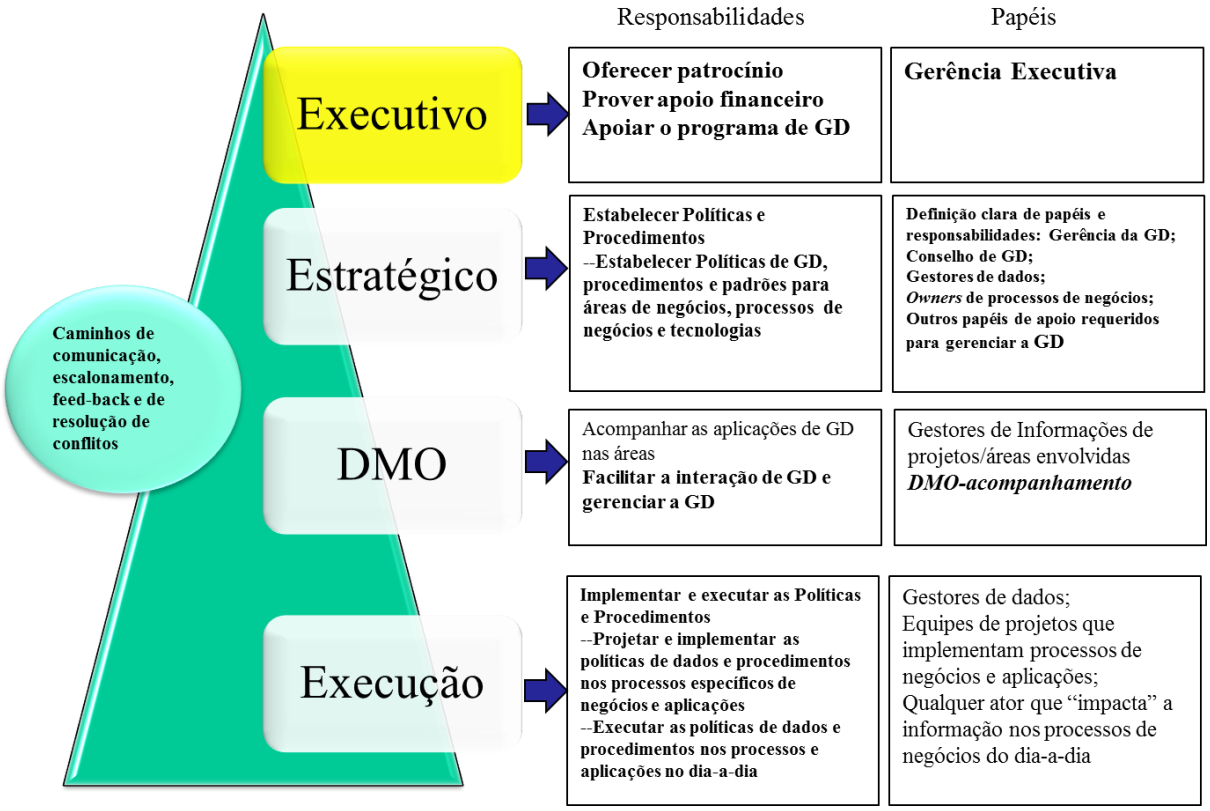

Figura 2 - Uma possível estruturação das camadas de GD

f. Desenvolver e aprovar políticas, padrões e procedimentos de gestão e governança de dados.

Aqui encontramos três dos P's da GD. Políticas são as regras maiores, definidas em consenso com áreas envolvidas, todas aprovadas pelo Comitê e divulgadas. As políticas balizam as grandes direções. Padrões regulam formas de nomes, documentos, layouts, entre outros. Procedimentos são detalhes específicos de como fazer certas atividades e que se juntam a descrições de processos que serão desenvolvidos.

g. Revisar e aprovar a arquitetura de dados:

Envolve analisar a arquitetura de dados, formada em níveis conceituais gradativamente detalhados (assuntos, entidades de negócios, entidades de dados, relacionamentos, principais atributos, etc.) e mostrando a sua conexão com outras arquiteturas, como de negócios, de sistema, de processos, de tecnologia, entre outros. Observar que há um processo (corpo de conhecimento) focado em arquitetura.

h. Planejar e patrocinar projetos e serviços da Gestão de dados:

Definir os projetos mais prioritários para se começar a implementação dos conceitos de GD. Podem ser melhorias na integração de dados, na definição de arquiteturas, na segurança,

em foco de dados não estruturados, em qualidade, entre outros. Serão definidos de acordo com as prioridades estratégicas dos negócios.

i. Estimar o valor dos ativos de dados e custos associados:

Trabalhar pelo custo negativo (riscos em imagem/reputação, *compliance*, etc.). Trabalhar em valoração relativa dos dados com relação aos outros recursos de um projeto e na aferição de valores que os competidores dariam por aqueles recursos de dados.

1.1.2. Controle da gestão de dados

O objetivo é:

- Supervisionar as camadas/estruturas e papéis envolvidos com dados.
- Coordenar as atividades de Governança de Dados.
- Gerenciar e resolver “conflitos” sobre dados.
- Monitorar e garantir aderência a aspectos regulatórios (no que tange a dados).
- Monitorar e garantir a aplicação e conformidade às Políticas, Padrões Procedimentos e Arquitetura.
- Supervisionar projetos e serviços relativos à Gerência de Dados.
- Comunicar e promover os valores dos ativos de dados.

Visão sintética:

Nesse ponto, a ideia é justamente o controle do funcionamento da estrutura definida anteriormente: Envolve coordenar as atividades de GD, supervisionar as estruturas definidas para as atividades de dados, gerenciar conflitos, entre outros.

1.2. Gestão da arquitetura de dados

Segundo o DMBOK (2009), o objetivo da Gestão da Arquitetura de Dados é:

- Entender as necessidades de informação da empresa. Aqui aparece uma variante com relação à outra já mencionada anteriormente (Gestão de Dados). O foco aqui é na necessidade de informações, ou seja, algo mais elaborado e focado em negócios e derivado do anterior.
- Desenvolver e manter o modelo corporativo de dados (MCD).
- Analisar e alinhar o MCD com outros modelos de negócios.
- Definir e manter uma arquitetura de tecnologia de Dados.
- Definir e manter uma arquitetura de integração de dados.
- Definir e manter uma arquitetura de *Data Warehousing* e de *Business Intelligence.*
- Definir e manter uma taxonomia e padrões de nomes (*namespaces*) de dados para a empresa.
- Definir e manter uma arquitetura de Metadados.

A visão sintética é:

a. Entender as necessidades de informações da empresa:

Levantar as informações, criando visões de dados por áreas/assuntos (*subject areas*). Envolve a criação da visão de dados necessários em variados níveis de abstração. Os dois primeiros são focos da visão corporativa e os três últimos estão dentro da visão de aplicações:

- Modelo de Entidades de negócios por *subject area*, numa visão de alto nível, com menção das principais Entidades de Negócios daquele domínio.
- Modelo conceitual de dados: Um pouco mais detalhado que o anterior, contemplando visões de relacionamentos e dos principais atributos envolvidos.
- Modelo lógico de dados: Visão mais detalhada que a anterior, contemplando as Entidades de Dados, com seus relacionamentos e seus atributos, normalizados, numa visão relacional.
- Modelo Físico, com uma visão de implementação, dentro da restrição do SGBD/tecnologia.

- Visão do implementador, com aspectos relacionados com SQL/DDL, *Views*, etc. ou de implementações pelos SGBD ou FMS (Hadoop, NOSQL), entre outros.

b. Desenvolver e manter o modelo de dados corporativo:

Envolve a manutenção dos dois níveis anteriormente definidos, em função do desenvolvimento dos modelos da aplicação. O grande objetivo é manter a coerência do significado dos dados (Entidades, relacionamentos) para toda a empresa. Aqui começa uma parte da definição semântica das Entidades de Negócios, com extensões semânticas de classificação e agregação, se necessário.

c. Analisar e compatibilizar o MCD com outros modelos da empresa:

Aqui ao grande foco é manter a coerência entre o modelo de negócios da empresa (como grande balizador) e os modelos de dados, modelos de processos, modelos de sistemas/aplicações, modelo de tecnologia e modelo de organização. Isso significa que uma Entidade de Negócios (modelo conceitual de dados) será criada, atualizada, mantida e eliminada por *processos* implementados em *sistemas*, apoiados em *tecnologia* e sob a responsabilidade de áreas (organização). É o dado se integrando nas várias dimensões da empresa.

d. Definir e manter a arquitetura de tecnologia de dados:

Envolve um framework que contemple os elementos de tecnologias que compõem o domínio de dados da empresa. Por exemplo, os SGBD´s tradicionais, os SGBD´s envolvidos em projetos de ERP, que podem ser diferentes, outros tipos de softwares usados para tratamento de Big Data, como Hadoop e NoSQL, utilitários desses componentes, ferramentas de modelagem de dados, ferramentas de qualidade e de *profiling* de dados, ferramentas de metadados, como dicionários, catálogos, glossários, entre outros. Tudo que tangencia a tecnologia que se usa para dados.

e. Definir e manter a arquitetura de integração de dados:

Envolve uma visão acerca das ferramentas e soluções de integração de dados. Inclui o envolvimento dos sistemas/aplicativos onde os dados são gerados, transformados, consumidos, eliminados, dentro do conceito de *data lineage* (linhagem de dados). Linhagem de dados é uma espécie de visão dos dados, desde a sua origem, observando o seu ciclo de vida. Dessa forma, essa recomendação do DMBOK inclui sistemas e informações e envolve papéis que fazem manipulações (CRUD) de dados e suas transformações a fim de torná-los adequados ao uso da empresa.

f. Definir e manter a arquitetura de DW e BI:

No fundo é um detalhamento dos itens anteriores, com foco em *Business Intelligence* e *Data Warehousing*. Envolve as estruturas de armazenamento (DW, Dmarts, ODS), camadas de transformação e integração (ETL) e camadas de consumo de informações (Relatórios, OLAP, *dashboards*, estudos de inferência por técnicas de *analytics*, *data mining*, etc.).

g. Definir e manter taxonomias e nomes (namespaces) como padrões corporativos:

Envolve a estruturação de taxonomias como, por exemplo, representações abstratas de classes/subclasses, heranças, ou composição e agregação, estendendo a semântica definida nos modelos conceituais e aprimorando o seu entendimento. É uma forma de se entender os dados do ponto de vista mais de suas classificações hierárquicas e de relacionamentos semânticos. Com relação aos nomes (*namespaces*), envolve a definição de termos padrões que objetivam a consistência dos elementos entre os modelos da empresa.

h. Definir e manter uma arquitetura de metadados:

Envolve a definição do fluxo de metadados, a integração entre os variados tipos de depósitos de metadados (catálogos, dicionários, glossários, etc.). Sugere a gerência sobre como os metadados são criados nas suas fontes, controlados, integrados e acessados. Visa a garantir a coerência na referência semântica dos metadados, em todos os níveis (dados no ambiente negocial, transacional e também dado no ambiente analítico de BI) e em todas as suas fontes.

1.3. Desenvolvimento de dados

O objetivo do Desenvolvimento de Dados, de acordo com o DMBOK (2009), é projetar, implementar e manter soluções que satisfaçam as necessidades de dados da empresa. Compreende as atividades focadas em dados dentro do ciclo de desenvolvimento do sistema, incluindo a modelagem de dados, análise de requisitos de dados e projeto, implantação e manutenção de bancos de dados. A sua estrutura é:

- Modelagem de dados, análise e projeto de soluções:
    - Analisar os requisitos de informação.
    - Desenvolver e manter modelos conceituais de dados.
    - Desenvolver e manter modelos lógicos de dados.
    - Desenvolver e manter modelos físicos de dados.
- Projeto detalhado de dados:
    - Projetar(desenhar) os Bancos de dados físicos.
    - Projetar(desenhar) os produtos de informação necessários.
    - Projetar (desenhar) serviços de acesso aos dados.
    - Projetar (desenhar) os serviços de integração de dados.
- Gerência de qualidade dos modelos de dados e dos projetos derivados:
    - Desenvolver padrões para modelagem de dados e projetos.
    - Revisar (auditar) a qualidade dos modelos de dados e dos projetos de bancos de dados.
    - Gerenciar versionamento e integração de modelos de dados.
- Implementação de dados:
    - Implementar, desenvolver e testar alterações em bancos de dados.
    - Criar e manter dados para ambientes de testes.
    - Migrar e converter dados.
    - Construir e testar produtos de informação.
    - Construir e testar serviços de acesso a dados.
    - Validar requisitos de informação.
    - Preparar para a implementação (da parte) dos dados.

A visão sintética é:

a. Modelagem de dados, análise e projeto da solução:

Os itens "analisar os requisitos de informação, desenvolver/manter modelos conceituais/modelos lógicos e modelos físicos" são parte do processo tradicional de desenvolvimento de aplicações e dizem respeito ao levantamento dos requisitos (de dados e de sistemas), com o intuito de desenvolver os modelos necessários à compreensão das necessidades de informações da solução em projeto. Essa abstração de dados é construída em vários níveis, indo da visão conceitual (entidades, relacionamentos, alguns atributos), lógica (entidades, relacionamentos, atributos, com maior nível de detalhe e aspectos de normalização, domínios, chaves, etc.), física (detalhamento da abstração anterior, com incorporação de aspectos associados a índices, campos nulos, formas de armazenamento em coerência com o SGBD a ser usado, etc.). Como qualquer proposição, o DMBOK não sugere nenhuma abordagem específica, devendo a empresa centrar no "o quê" está sendo sugerido e não no "como".

b. Projeto detalhado de dados:

Projetar os Bancos de Dados físicos se relaciona com colocar as estruturas de BD de acordo com as características do SGBD em questão. Significa se preocupar com aspectos de performance, armazenamento, particionamento de dados, colunas calculadas definidas como armazenadas, entre outros. Projetar (desenhar) os produtos de informação merece uma observação anterior: Produto de informação significa tudo aqui que se pode extrair direta ou indiretamente de bancos de dados (Relatórios operacionais, gerenciais, cubos, *dashboards, scorecard,* saídas na forma de documentos XML, portal, dados para integração com outros aplicativos, etc.). Assim, esse item foca nos projetos das saídas desejadas do sistema.

Os serviços de acesso aos dados podem ser entendidos como a disposição com que os SGBD´s se encontram numa arquitetura ou topologia. Podem ser servidores "linkados", acesso por Serviços (SOA), *Message Broker* (serviços assíncronos de mensagens), dispositivos tipo ODBC, JDBC, arquitetura de bancos distribuídos, como replicação, partições, distribuição, camadas de ETL que fazem leitura de bancos de dados, entre outros.

Projetar os serviços de integração de dados representa a preservação de certos conceitos fundamentais do ambiente transacional, como ACID (Atomicidade, Consistência, Isolamento e Durabilidade) (ELMASRI, 2000). Os conceitos de Atomicidade estão associados a mecanismos ou serviços que garantem a execução conjunta ou integrada de comandos sob o mesmo escopo transacional, sacramentando todas as ações ou desfazendo-as completamente. A consistência é um conceito que garante os estados de consistência inicial

ou final dos dados alterados pela transação. Os serviços de isolamento garantem que as transações executadas em paralelo não sofrerão ou interferirão nas outras, simulando um ambiente exclusivo de recursos, quando na realidade eles são compartilhados. O conceito de durabilidade se expressa nos serviços que garantem a manutenção do estado alcançado pela transação, depois que ela foi terminada, mesmo que ainda alguma intercorrência possa afetar os dados atualizados por ela.

Além disso, também devem ser considerados os conceitos de integração numa visão mais ampla. Envolve, dessa forma, a definição de camadas de integração, como ETL, de persistência, etc.; e do uso de *Stored Procedures*, como camada de ações essenciais de dados como ADD, MOD e DEL de linhas /registros.

c. Gerência da qualidade do modelo de dados e do projeto de Bancos de Dados:

Envolve a definição e verificação de padrões a serem usados nos modelos, incluindo nomes de entidades, de atributos, de relacionamentos, simbologias de entidades, relacionamentos, atributos, cardinalidade, entre outros. A revisão é a verificação dessas aderências feitas por trabalhos de QA (Quality Assurance) ou por revisões por pares (VER/VAL), garantindo a compatibilidade entre os requisitos iniciais (de dados) e os elementos implementados. Inclui também a gerência de versionamento, é parte da gerência de configuração, garantindo a integridade de modelos, com controles de versionamento, controles de alterações (quem fez, porque, quando, e o que?), entre outros.

d. A implementação de dados:

Está diretamente associada com o desenvolvimento, implementação e testes (das partes de dados), dentro do contexto geral de teste dos sistemas. Os testes se referem aos elementos definidos anteriormente (Bancos de Dados e outros produtos de dados, serviços de dados, integração de dados, etc.). O conceito de validação de requisitos de dados é aplicado aqui com a avaliação dos pontos levantados na forma de requisitos (de informação) e a análise de sua devida transformação em produtos do sistema. Também se relaciona com migração, preparação e conversão de dados dentro do contexto do projeto, incluindo aspectos de implantação.

1.4. Gestão de operações de dados

O objetivo da Gestão de Operações de Dados, segundo o DMBOK (2009) é planejar, controlar a apoiar os ativos de dados ao longo do seu ciclo de vida, indo desde a criação e aquisição (obtenção) até o arquivamento final (*archiving*) e eliminação (*purge*). A estrutura é:

- Suporte a Bancos de dados:
    - Implementar e controlar ambientes de Bancos de Dados
    - Obter dados originados de fontes externas.
    - Planejar para Recuperação de dados (*Recovery*).
    - Realizar *Backup* e *Recovery* de Bancos de Dados.
    - Estabelecer níveis de serviços relacionados à performance de Bancos de Dados.
    - Monitorar e ajustar aspectos de performance de Bancos de Dados.
    - Planejar a retenção de dados.
    - Arquivar, reter e eliminar dados.
    - Suportar aspectos de Bancos de Dados especializados.
- Gerência de tecnologia de dados:
    - Entender os requisitos tecnológicos de dados.
    - Definir arquiteturas tecnológicas de dados, já mencionada anteriormente na função Gerência da Arquitetura de dados como “Definir e manter uma arquitetura de Bancos de Dados”.
    - Avaliar tecnologias de dados.
    - Instalar e administrar tecnologias de dados.
    - Controlar e acompanhar aspectos de licenças de tecnologia de dados.
    - Suportar o uso e as dúvidas (pendências) sobre tecnologia de dados.

A visão sintética é:

a. Suporte a bancos de dados:

- Implementar e controlar ambientes de Bancos de Dados: significa ter controles sobre os possíveis diversos ambientes de SGBD´s da empresa, suas várias instâncias, tanto de SGBD quanto de tecnologias correlatas, gerência e conhecimento de parâmetros e afinamento de SGBD e correlatos, controle de sua conectividade (com outros SGBD em ambientes distribuídos, ou com outras camadas), ajuste e afinamento de outras camadas correlatas que dialogam com o SGBD e controle do ambiente de *data storage* usado pelos SGBD´s e produtos correlatos.
- Obter dados originados de fontes externas: Controle de aquisição dos dados obtidos de fontes externas (na forma de CD, DVD, EDI, XML, RSS, etc.), como vem (licenciados ou contratos) de onde vem (fontes), com qual periodicidade chegam, dados de contrato com fornecedores, SLA com o fornecedor, entre outros; e registro no modelo lógico/conceitual de dados.
- Planejar para recuperação de dados (*recovery*): Planejamento das atividades de *backup* e *recovery*, com definição de procedimentos, periodicidades, tipos de *backup*s (integral, incremental), tipos de mídia destino, SLA definido para tempos máximos de recuperação, tipos de perdas, tipos de recuperação, tipos de arquivos a serem resguardados (BD, Logs, cópias lógicas, cópias físicas, etc.).
- Realizar *Backup* e *Recovery* de Bancos de Dados: Instanciação do plano definido acima, com a criação das *backup*s, logs de transações, estratégias de imagens em discos RAID, etc.
- Estabelecer níveis de serviços relacionados à performance de Bancos de dados: SLA definido para a tempo de resposta de BD, de algumas transações chaves, de jobs batchs de apoio, como ETL, de tempo de recuperação de BD, de interrupções físicas, lógicas, etc.
- Monitorar e ajustar aspectos de performance de Bancos de Dados: Realizar acompanhamento proativo (preventivamente) e reativo (após incidentes). Envolve aspectos de gerência de tempo de resposta ao usuário, provenientes das mais variadas causas-raiz (problemas de codificação de SQL, comandos, falhas de projetos de bancos, ausência de indexações corretas, problemas provenientes de desatualização de estatísticas usadas pelo otimizador de pesquisas, etc.). Associado a conceitos de processos do ITIL, MPS-SV, entre outros, para controle de incidentes e problemas.
- Planejar a retenção de dados: Planejar a forma, tempo e tipo de retenção, transferência para mídias secundárias e de eliminação de dados, de acordo com políticas de Governança de dados e/ou aspectos regulatórios.

- Arquivar, reter e eliminar dados: Instanciação do Plano de retenção de dados definido anteriormente.
- Suportar aspectos de Bancos de Dados especializados: Entender que hoje há uma grande variedade de sistemas gerenciadores de bancos de dados e de FMS (*File Management Systems*), cada qual disponível para certos tipos de necessidades: OODBMS (SGBD para Bancos orientados a objetos), XML, NOSQL (para dados semi ou não estruturados), *Hadoop/Map Reduce* para armazenamentos de petavolumes, Bancos de dados de armazenamentos dimensionais, entre outros. Para detalhes sobre Bancos de Dados NoSQL veja (SADALAGE, 2014).

b. Gerência de Tecnologia de Dados:

- Entender os requisitos tecnológicos de dados: Como em qualquer solução a ser desenvolvida, entender primeiramente o problema posto, quais as limitações das tecnologias colocadas e existentes, quais os requisitos específicos de HDW para aquela tecnologia de dados, quais os requisitos de sistema operacional, os de conectividade, as habilidades do "*peopleware*" envolvido, as implicações de custo e se há equivalentes no domínio de softwares livres.
- Definir arquiteturas tecnológicas de dados (já mencionadas anteriormente na função Gerência da Arquitetura de dados como "Definir e manter uma arquitetura de Bancos de Dados"): Significa que tipo de software será necessário em cada camada relacionada com dados: BD Convencionais, BD especiais (Georreferenciados, XML, NOSQL para big data, FMS, Bancos de Dados multidimensionais, etc.) e que outras camadas são necessárias: integração, ferramentas de modelagem, BI, ETL, virtualização de servidores, suites de testes, camadas para geração de dados, entre outros.
- Avaliar tecnologias de dados: Inclui a análise de alternativas tecnológicas de dados. Isso pode ser feita por um processo de Gerência de Decisão (GDE), no modelo MPS.BR ou DAR (CMMI), envolvendo a definição de critérios e pesos para a melhor opção, benchmarks, visitas, provas de conceito, etc.
- Instalar e administrar tecnologias de dados: Na realidade, é a instanciação dos dois últimos itens anteriormente discutidos.
- Controlar e acompanhar aspectos de licenças de tecnologia de dados: Considerar a importância do controle de licenças de uso de SGBD, de ferramentas de BI, de ferramentas de integração, de geradores de relatórios,

de cubos, de mining, entre outros; visando preservar aspectos legais e de controle de custo.

- Suportar o uso e as dúvidas (pendências) sobre tecnologia de dados: Aqui estão concentradas as ações de apoio, suporte e resolução de incidentes associados à camada de dados, com processos, por exemplo, do modelo ITIL, ou MPS-SV, com detalhamento de níveis de apoio. Passa também pela escolha adequada e pelo treinamento ministrado visando à utilização daquela tecnologia de dados.

1.5. Gestão da segurança de dados

Segundo o DMBOK (2009), o objetivo da gestão da segurança de dados é planejar, desenvolver e executar as políticas de segurança e procedimentos a fim de prover a adequada autenticação, acesso e auditoria de dados e informações. A estrutura é:

- Entender as necessidades de segurança de dados e os requisitos regulatórios associados.
- Definir Política de segurança de dados.
- Definir Padrões de segurança de dados.
- Definir Procedimentos e controles de segurança de dados.
- Gerenciar usuários, *passwords* e membros de grupos de usuários.
- Gerenciar visões e permissões de acesso aos dados.
- Monitorar autenticação de usuários e comportamento de acesso.
- Classificar o grau de confidencialidade das informações.
- Auditar a segurança dos dados.

A visão sintética é:

a. Entender as necessidades de segurança de dados e os requisitos regulatórios associados:

Os requisitos regulatórios normalmente estão associados com modelos do tipo SOX, Basiléia-II, Clerp-Act of Australia, etc.

b. Definir política de segurança de dados:

São as regras e diretrizes maiores que devem ser seguidas pela empresa, em termos de segurança da informação. São normalmente definidas por administradores de segurança juntamente com gestores de dados e auditores de segurança externa ou interna. Deverá ser aprovada pelo Conselho de Governança de Dados (GD).

c. Definir padrões de segurança de dados:

Os padrões de segurança gravitam em torno de: padrões de senhas, grupos de usuários, padrões de criptografia, guia para acessos externos, protocolos de transmissão pela internet, requisitos de documentação das informações de segurança, padrões de acesso remoto,

procedimentos para relato de incidentes de segurança, padrões de armazenamento e acesso de dados para portáveis e descarte (eliminação) desses tipos de equipamentos.

d. Definir procedimentos e controles de segurança de dados:

Significa, para o DMBOK, estabelecer um grau de detalhe sobre as diversas atividades, tanto de planejamento, operação quanto de controle da gestão de segurança de dados.

e. Gerenciar usuários, *passwords* e membros de grupos de usuários:

Analisar os diversos usuários, *passwords*, grupos de usuários, privilégios de usuários e/ou de grupos, tendo uma fotografia correta dessas entidades e de suas modificações no contexto da segurança de dados.

f. Gerenciar visões e permissões de acesso aos dados:

Envolve a aplicação dos conceitos de *opt in* e *opt out*, ou seja, as atividades de se garantir privacidade e segurança de dados por conceder especificamente permissões, através de definições explícitas (*opt in*) ou por se restringir ações específicas, dentro de um leque amplo de alternativas concedido por *default* (*opt out*). Também os conceitos de visões (*views*) de bancos de dados são elementos considerados nesse ponto.

g. Monitorar autenticação de usuários e comportamento de acesso:

Representa o acompanhamento dos acessos, com a observação de comportamentos atípicos ou de riscos, que deverão ser reportados aos envolvidos. Isso alimenta as futuras alterações de planos, projetos e políticas de segurança de dados. Alguns tipos de dados mais sensíveis poderão ser monitorados em tempo real, com possibilidade de alertas e mensagens para administradores, gestores, imediatamente ao seu acontecimento.

h. Classificar o grau de confidencialidade das informações:

Classificar o grau de confidencialidade, definindo hierarquias como: informação para acesso geral (todos podem ver); informações somente para uso interno (somente colaboradores podem acessar as informações que poderão ser mostradas ou discutidas no âmbito externo da empresa, porém não copiadas); informações confidenciais (não devem ser compartilhadas por toda empresa); informações confidenciais restritas (somente aberta para certos colaboradores previamente identificados com o status “devem saber”); informações confidenciais registradas (poucos têm acesso e há a estrita necessidade de assinatura de documento de responsabilidade pelo seu uso ou conhecimento).

i. Auditar a segurança dos dados:

Objetiva a realização de sessões de auditoria com o propósito de analisar, validar, aconselhar e recomendar políticas, padrões e atividades relacionadas à gerência de

segurança de dados. Pode ser interna ou externa, porém feitas por pessoas sem nenhum envolvimento nos processos em auditoria.

1.6. Gestão de dados mestres e de referência

O objetivo da Gestão de dados mestres e de referência é planejar, implementar e controlar atividades para garantir consistência de dados Mestres e de Referência. Os dados Mestres são os dados fundamentais de uma empresa e envolvem clientes, fornecedores, colaboradores, contas, locais, entre outros. Os dados de Referência são dados relacionados com códigos, como estado, país, status de um pedido, entre outros, e servem como elementos para categorizar/classificar outros dados. (DMBOK, 2009).

Os dois Mestres e Referências servem como input para os dados transacionais: Num pedido, por exemplo, que representa um dado do tipo Transacional (normalmente associado a uma data) você tem dados Mestres (clientes e produtos entregues, vendedor envolvido, etc.), dados de Referência, como o status do pedido, CEP padrão do fornecedor, entre outros, e no conjunto formam os dados Transacionais do Pedido. Essa classificação de tipos de dados está mais detalhada adiante. A estrutura é:

- Entender as necessidades de integração de dados Mestres e de Referência.
- Identificar Fontes e contribuintes (*contributors*) de dados Mestre e de Referência.
- Traçar a linhagem do dado, para identificar a suas Fontes originais e temporárias (BD, FMS, processo, área organizacional, pessoas, papéis/indivíduos envolvidos).
- Definir e manter a arquitetura de integração de dados.
- Implementar soluções de gerência de Dados Mestres e de Referência.
- Definir e manter regras de “match” para os dados replicados.
- Definir os “*Golden Records*”.
- Definir e manter hierarquias e afiliações (conceitos de MDM).
- Planejar e implementar integrações das novas fontes de dados.
- Replicar e distribuir Dados Mestres e de Referência.
- Gerenciar alterações de Dados Mestres e de Referência.

Algumas considerações iniciais: Os dados, há muito, são usados por diferentes áreas dentro de uma empresa, de formas também diferentes. As áreas de Venda, Fornecedores, Manufatura, etc. veem os dados de Vendas, por exemplo, de forma diferente, cada uma com uma visão ou conjunto de atributos específicos, algumas inclusive com definições diferentes para a mesma entidade ou informação.

Os dados podem ser vistos como enquadradas dentro de três tipos: Os Mestres, os de Referência e os Transacionais:

Os mestres são aqueles dados centrais da empresa, com certa característica de imutabilidade. Representam entidades de negócios vitais da empresa, como cliente, fornecedores, empregados, locais, entre outros. Os dados Mestres são de domínios mais amplos, com variações semânticas como pessoas (Física, Jurídica), locais, elementos geográficos, etc.

Os dados de referência representam elementos com características mais voltadas para codificação de valores, como código e descrição, por exemplo. Servem para categorizar outros dados. Representam tipos de dados que servem de referência para algum contexto, como CEP, códigos geográficos (cidade, estado, etc.), códigos contábeis, lista de valores de certos domínios, entre outros. Uma das áreas onde são muito usados é na Saúde. Os códigos internacionais de doença (CID) e os códigos de tratamentos (*Current Procedural Terminology* - CPT) são fundamentais nas estatísticas e estudos de doenças e as ações realizadas de tratamento. O CID está na versão 9, com a previsão da incorporação do CID-10 para outubro de 2013. Os dados de referência possuem relacionamento entre eles (o atributo código-CEP e o atributo descrição-Local) e entre si (códigos entre eles-CEPs da mesma raiz). Outro exemplo: Considere a trinca cod-produto, desc-produto, cod-produto-pai. Esses elementos representam uma referência de código para descrição e uma relação de hierarquia de cod-produto-pai para código-produto. Os dados de referência tendem a ser mais imutáveis, pois representam codificações que tendem a permanecer.

Ambos (dados “mestres” e dados de “referência”) são insumos para a produção de dados tipos “transacionais”. Por exemplo, um “cliente” comprando “produtos” em “locais” da minha empresa, gera transação de compras (que podem ter os seus dados próprios, como data, tipo de desconto daquela compra, etc.).

Os dados Mestres e de Referência devem ser geridos pela empresa de forma centralizada, envolvendo os gestores de dados da(s) área(s) afim(ins). Sua gerência envolve a criação, integração, manutenção uso e descarte. Contempla também a definição das versões abrangentes (que englobem todos os seus atributos) e, preferencialmente únicas, que representem a verdade dos dados (*golden records*). Buscam, na essência, a entidade com seus atributos e valores mais íntegros, atuais e associados ao negócio.

Os DMR (Dados Mestres e de Referências) são implementados por diversas ferramentas como ETL, integração de dados, ODS (*Operational Data Store*) para armazenamento das versões únicas, ferramentas de *profiling* e *cleasing*, para a descoberta de duplicatas, entre outras.

Os dados mestres podem ser classificados em alguns domínios: "partes (*parties*)", que representam organizações, indivíduos, seus papéis, como clientes, empregados, fornecedores, vendedores, entre outros. Na visão de segurança podem ser: cidadãos, testemunhas, vítimas. No domínio saúde podem ser: clientes, provedores de serviços, estes classificados em médicos, convênios, etc. No domínio educação, podem ser: aluno, professor, inspetor, diretor, etc. Há dados Mestres também nos domínios de clientes, ambiente financeiro, produtos ou serviços e localização, dentre outros.

A visão sintética segundo o DMBOK é:

a. Entender as necessidades de integração de dados Mestres e de Referência:

Significa ter os devidos controles para compatibilizar os dados que estão replicados e usados em diferentes sistemas/aplicativos. Normalmente as causas-raiz de problemas de qualidade de dados revelam esse aspecto. A solução Master Data management (MDM) é complexa e, como tal, exige solução gradativa e incremental. A sugestão é entender a necessidade e o uso daquele dado em estudo, nas diversas aplicações/sistema da empresa.

b. Identificar Fontes e contribuintes (*contributors*) de dados Mestre e de Referência:

Depois de entendido, é importante traçar a linhagem do dado, para identificar a suas fontes originais e temporárias (BD, FMS, processo, área organizacional, pessoas, papéis/indivíduos envolvidos).

c. Definir e manter a arquitetura de integração de dados:

Já mencionada anteriormente na função Gerência da Arquitetura de dados como "Definir e manter uma arquitetura de integração de dados", a arquitetura de solução de MDM passa por topologias parecidas com as de Bancos de dados distribuídos e/ou replicados e busca evitar o aparecimento de "silos" de dados ou arquivos isolados e personalizados para atender aplicações específicas.

d. Implementar soluções de gerência de Dados Mestres e de Referência:

Passa por definição de soluções que permitam o uso compartilhado do dado Mestre ou de Referência, na sua forma "*golden record*" por variadas aplicações OLTP ou de BI, mantendo a sua integridade. As topologias devem permitir leituras diretas dos DM (dados mestres ou de referência) ou o seu uso em sistemas através de replicações controladas (síncronas ou assíncronas).

e. Definir e manter regras de "*match*" para os dados replicados:

Trabalhar padrões para que se possa identificar duas ocorrências como sendo do mesmo objeto. Conforme citado anteriormente, dois registros de cliente com nome Carlos Barbieri são considerados o mesmo objeto (Carlos Barbieri)?

Tal atividade passa por técnicas de identificação de elementos duplicados, através de regras de inferência de similaridade, por técnicas de eliminação de duplicação de registros iguais e por técnicas de consolidação que podem ser:

- *Match-merge:* Essas técnicas consistem no batimento (match) das várias ocorrências multiplicadas e a produção de um registro abrangente que as represente.
- *Match-Link:* Técnicas que definem um registro, com apontadores para outros, sem consolidação física em um único elemento.

f. Definir os "*Golden Records*":

"*Golden Records*" significa o conceito de fonte única, íntegra e confiável, que procura garantir a verdade sobre os dados. Por exemplo, um único cadastro lógico de clientes, formado por informações advindas de várias fontes físicas.

g. Definir e manter hierarquias e afiliações (conceitos de MDM):

As hierarquias e afiliações complementam as informações de relacionamentos dos dados Mestres, mostrando, por exemplo, dois registros mestres de clientes, relacionados como Todo-Parte (um cliente é parte de um outro cliente, ou seja faz parte do mesmo grupo, ou é afiliada, ou seja tem um relacionamento com a outra empresa). Também há o relacionamento conhecido como "É um tipo de". As duas classificações de dados são muito aplicadas no conceito de objetos (Todo-Parte ou composição e agregação) e (É um tipo de definindo tipos e subtipos), adotados em Modelagem de Classes e objetos.

h. Planejar e implementar integrações das novas fontes de dados:

Nesse ponto, o DMBOK se preocupa com a chegada de novas fontes de dados que deverão ser integradas ao ambiente já existente. Isso envolve: analisar as requisições feitas de integração, a complexidade e custo de sua integração e avaliar a qualidade dos dados entrantes.

i. Replicar e distribuir Dados Mestres e de Referência:

Esse ponto versa sobre a arquitetura definida para a solução de MDM implementada. Poderá ser via bancos distribuídos, ou através de replicações.

j. Gerenciar alterações de Dados Mestres e de Referência:

Esse ponto, de crucial importância, deverá ser considerado com cuidado, pois os dados do ambiente MDM estarão compartilhados e as suas alterações implicam controles mais rigorosos, a fim de evitar impactos e rupturas em sistemas em funcionamento. Passa por: criar e receber pedidos de alteração, identificar áreas interessadas; avaliar impactos em função da solicitação, aceitar ou rejeitar a solicitação ou transferir a decisão para o âmbito da Governança de Dados (GD), comunicar a decisão às partes interessadas, realizar as mudanças, caso aprovada, comunicar as partes interessadas acerca das mudanças.

A Figura 3 mostra uma classificação de dados, contemplando DMR (Dados Mestres e Referenciais) e outros dados como históricos, temporários e condicionais, enquadrados em dimensões origem, forma, definição e gênese.

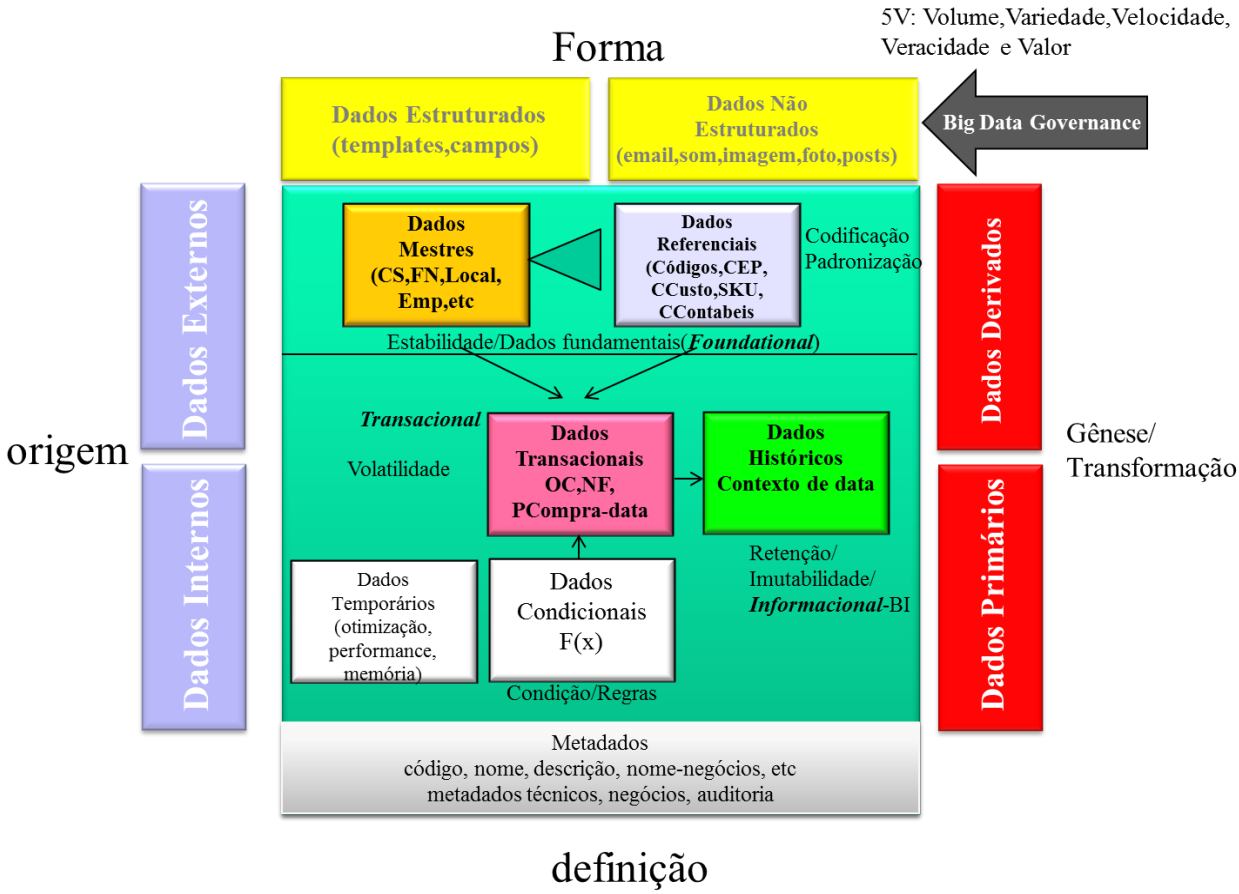

Figura 3 - Classificação de Dados

1.7. Gestão de *data warehousing* e *business intelligence*:

O objetivo da Gestão de *data warehousing* (DW) e *business intelligence* (BI) (DMBOK, 2009) é planejar, implementar e controlar processos para prover dados de suporte à decisão e apoio a colaboradores envolvidos em produção de relatórios, consultas e análises.

A estrutura é:

- Entender as necessidades de informações analíticas (BI).
- Definir e manter a arquitetura de DW e de BI (já mencionada anteriormente na função Gerência da Arquitetura de dados como "Definir e manter uma arquitetura de DW e de BI").
- Implementar os DW e DataMarts.
- Implementar as ferramentas de BI e de Interface para usuários.
- Processar os dados para o ambiente de BI.
- Monitorar e ajustar os processos de DW.
- Monitorar e ajustar as atividades e aspectos de performance de BI.

A visão sintética é:

a. Entender as necessidades de informações analíticas (BI):

Os requisitos de BI são revestidos de certas diferenças quando comparados com os requisitos tradicionais de sistemas transacionais. O fornecedor de requisitos, que vocaliza as necessidades e os problemas de BI pertence a outro patamar gerencial, normalmente na camada mais tática e estratégica. Isso demanda uma observação mais apurada sobre as necessidades e problemas (requisitos de negócios e de usuários), além de técnicas mais efetivas de comprometimento das partes, como protótipos, provas de conceito, entre outros. O levantamento das necessidades de negócios sugere a captura de metadados (significado dos dados, dos processamentos, de indicadores, etc.). Nesse momento, é importante a observação comparativa dos dados demandados contra os dados existentes.

b. Definir e manter a arquitetura de DW e de BI:

Passa pelas diferentes alternativa de escolas: Visão de EDW (Bill Inmonn) e de DMarts evolutivos e integrados (Ralph Kimball), com todos os elementos que formam uma arquitetura de DW e BI: sistemas transacionais fomentadores dos dados, camada de ETL, camada de armazenamento com *DataWarehouse*, *DataMarts*, ODS,etc, camada de ferramentas para produção de informações, camada de *profiling* e *cleansing* dos dados, etc.

c. Implementar os DW e Data Marts:

Nesta atividade o DMBOK foca na materialização gradativa de DW e BI, em projetos evolutivos e integrados.

d. Implementar as ferramentas de BI e de interface:

Passa pelos tipos de ferramentas necessários para se alcançar os objetivos. São ferramentas de *dashboards*, ferramentas de visualização de dados, ferramentas de relatórios, OLAPs (de cubos), de análise preditiva*/mining*, entre outros, que formarão o arsenal de aplicativos para atender as necessidades de informação da empresa.

e. Processar os dados para o ambiente de BI:

Relaciona-se com as atividades de ETL (Extração, Transformação e Carga) de dados, atividades de *cleansing* e integração de dados, considerando áreas intermediárias, como *staging*, depósitos intermediários como ODS, etc.

f. Monitorar e ajustar os processos de DW:

Passa pelos processos de monitoração de performance de bancos dos dados que compõem a estrutura do DW, processos de *backup/recovery*, processos de *archiving*, etc.

g. Monitorar e ajustar as atividade e aspectos de performance de BI:

Passa por atividades de monitoração de tempos de respostas de aplicativos de BI, número de usuários de BI por unidade de tempo, entre outros. Esses dois últimos elementos interferem diretamente no SLA de serviços de BI.

1.8. Gestão de documentos e conteúdo:

O objetivo é planejar, implementar e controlar atividades para armazenar, proteger e acessar dados encontrados em arquivos eletrônicos e registros físicos (texto, gráficos, imagens, áudio e vídeo), ou seja, o foco em dados não estruturados, não armazenados em sistemas relacionais (DMBOK, 2009). Há duas grandes subfunções: Gerência de Documentos e de Registros e Gerência de Conteúdo.

A estrutura de atividades desta função e suas subfunções é descrita abaixo:

- Gerência de Documentos e de Registros
  - • Planejar a gerência de Documentos e de Registros;
  - • Implementar Sistemas de Gerência para Aquisição, Armazenamento, Acesso e controle de Documentos e Registros;
  - • *Backup* e Recuperação de Documentos e Registros;
  - • Retenção e eliminação de Documentos e Registros;
  - • Auditar a Gerência de Documentos e Registros.
- Gerência de Conteúdo
  - • Definir e manter taxonomia corporativa para documentos e conteúdo (Já mencionada anteriormente na função Gerência da Arquitetura de dados como "Definir e manter uma taxonomia e padrões de nomes (namespaces) de dados para a empresa");
  - • Documentar/indexar Metadados sobre informações de Conteúdo;
  - • Prover acesso e recuperação de Conteúdos;
  - • Estabelecer Governança sobre qualidade de Conteúdos.

1.8.1. Gerência de Documentos e de Registros:

Essa gerência se fundamenta em dois pilares: o primeiro fala sobre a Gerência do documento em si e o outro fala do seu conteúdo. O primeiro se preocupa com o documento como se fora um objeto e o outro com a sua estruturação e conteúdo. O conceito de Big Data, que envolve os diferentes tipos de dados semi ou não estruturados, não foi (ainda) contemplado diretamente no DMBOK, devendo ser foco das próximas versões do modelo. Esse corpo de conhecimento, embora não explicite o termo Big Data, trata

fundamentalmente dos dados não estruturados (DNE) como: arquivos (em formato não relacional), gráficos, imagens, textos, relatórios, formulários, vídeo, som, entre outros. Os novos tipos de dados como posts de Linkedin, posts de Facebook, etiquetas de RFID, dados biométricos e dados gerados por comunicação máquina-máquina (M2M), como monitoração de pacientes, medidores inteligentes de energia elétrica, dados de localização (GPS), etc. não foram diretamente considerados nesse contexto, embora a sua governança se revista dos mesmos conceitos. Esses dados (DNE) constituem cerca de mais de 70% dos dados existentes hoje num ambiente corporativo e, portanto, passam a merecer a atenção, até porque muitas regulações oficiais assim exigem. Aspectos regulatórios como Sarbannes Oxley (SOX*), E-Discovery amendments to Federal rule of civil procedures*, Canada Bill´s 190, dentre outros, são exigências presentes em muitos segmentos da indústria.

A visão sintética é:

a. Planejar a gerência de documentos e registros:

Nesta atividade o DMBOK foca nos processos, técnicas e tecnologias que visam o controle e a organização dos documentos e registros, quer seja em meio eletrônico ou papel. Nesta ponto devem ser considerados o planejamento dos diferentes sistemas de controle de documentos e registros: sistemas de bibliotecas, sistemas de controle de emails, sistemas de controle de documentos na forma eletrônica ou em microficha. Devem ser planejados os seguintes pontos: armazenamento, integridade, segurança, qualidade do conteúdo, formas de indexação e de acesso e guias gerais para a sua gerência. O planejamento deve considerar aspectos dos vários pontos do ciclo de vida do documento, da sua criação ao descarte/eliminação, passando pela sua classificação (taxonomia), indexação, arquivamento e uso e recuperação.

O planejamento passa também pela definição das políticas e procedimentos para esses diversos momentos do ciclo de vida do documento, bem como pela definição das unidades organizacionais (UO) que deverão ser as suas gestoras (*stewards*). Essas políticas deverão, entre outros pontos, definir aspectos de responsabilidade dos gestores, políticas de retenção em conformidade com as regulações existentes, aspectos de circulação e distribuição interna e externa, entre outros.

b. Implementar sistemas de gerência para aquisição, armazenamento, acesso e controle:

Aqui o DMBOK está focando na implementação de sistemas de software que apoiem esses pontos. Passa por sistemas de Gerência de Conteúdo (ECM), com documentos criados via eletrônica, scanner ou OCR. Devem permitir facilidades de indexação por palavras chaves ou por elementos do conteúdo (aqui as duas partes desse processo DMBOK se encontram). Deverá haver metadados que bem caracterizem aqueles documentos, como data de criação, data de revisão, nome do criador/responsável, entre outros. As referências bibliográficas,

associadas eventualmente ao documento formam uma parte de suas informações estruturadas. O sistema deverá permitir o controle de versionamento de documentos, com garantia de GCO (configuração), via *check-in* e *check-out* e comparações de versões, além de possibilidade de entendimento do seu fluxo (*work flow* dos documentos). As facilidades para pesquisa deverão contemplar palavras chaves, buscas via *drill-down*, etc.

c. Backup e recuperação dos documentos e registros:

Aqui o foco é na manutenção da integridade dos documentos, com um plano de risco associado às suas eventuais perdas. O plano de *backup/recovery* aponta aspectos de frequência de cópia, alternativas de *backup* passivas, como *cold-site,* ou ativas, como *hot-site*, além de políticas e procedimentos para mitigação.

d. Retenção e eliminação de documentos:

Aqui o foco é nos aspectos de retenção dos dados no ambiente principal até serem transferidos para uma mídia secundária. Deverão ser considerados aspectos legais, fiscais e valores históricos do documento. Um ponto importante a ser considerado é a garantia da compatibilidade do documento com relação à versão do sistema de gerência e do sistema operacional onde este funciona. Isso é importante no caso de recuperações de documentos que ao serem trazidos para o ambiente atual de software, podem apresentar problemas de compatibilidade de versão. Aspectos de privacidade e de retenção de dados pessoais também devem ser considerados neste item.

e. Auditar a Gerência de Documentos e Registros:

Envolve aspectos de controle, visando à aplicação das políticas, procedimentos e regras definidas pela Governança dos dados. Envolve periodicidade de auditorias e observação de vários aspectos, como: local de armazenamento, confiabilidade, precisão, classificação e indexação, acesso e recuperação, métodos de eliminação (*disposition*), segurança e confidencialidade, percepção e entendimento organizacional sobre a gerência de documentos, entre outros.

1.8.2. Gerência de Conteúdo:

Esta gerência está relacionada à ciência da informação e a gerência de conhecimentos e trata fundamentalmente de aspectos de entendimento e classificação de conteúdos de documentos, via aplicação de conceitos de taxonomia. No fundo, é prover uma forma de documentação e entendimento da arquitetura de conteúdo através de elementos constituintes, relacionamentos (links), atributos e instâncias. São normalmente estruturadas

via uma visão ontológica (conhecimento de ser ou entes), com taxonomias da seguinte forma: achatada (com os elementos listados em sequência, sem uma estruturação entre eles), hierárquica (com uma organização dos elementos apresentando certa forma de subordinação), na forma de *facets* ou estrelas (com os elementos dispostos numa forma de radial, dispostos em torno de um ponto central, como comumente encontrado nos mapas mentais) e de rede (misturando a hierarquia com *facets*).

A gerência de conteúdo também foca nos aspectos de indexação e documentação dos metadados, visando à facilidade de se localizar e identificar certo documento. Quando se fala de dados não estruturados (DNE), há que se considerar o aspecto característico de diversidade desses elementos, buscando-se soluções mais evoluídas para tal. Muitas delas, agora é que estão sendo desenvolvidas, como indexação de arquivos de áudio, de imagens (nesse caso, considerando cores, formas, texturas, disposição de elementos, etc.), reconhecimentos faciais, entre outros. Uma vez definidos os mecanismos de indexação e recuperação, teremos as facilidades para prover acesso e recuperação.

O DMBOK (2009) também foca no estabelecimento de governança sobre esses tipos de dados não estruturados. O tema sobre a governança desses novos ativos já começa a ser discutido e será, certamente, incluído nas próximas versões do modelo DMBOK. Até lá, muito já se diz e escreve sobre isso, numa nova capa denominada "*Big Data Governance*". O livro mais recente que toca no tema, é de Sunil Soares e se chama "*Big Data Governance: an emerging imperative*", lançado em novembro de 2012. O foco é justamente a adoção e adaptação da Governança de dados, digamos tradicional, para ser aplicada na Governança de Big Data.

1.9. Gestão de metadados:

O objetivo é planejar, implementar e controlar atividades que viabilizem um fácil acesso aos metadados integrados e de qualidade (DMBOK, 2009).

A estrutura de atividades desta função é descrita abaixo:

- Entender os requisitos de Metadados.
- Definir a arquitetura de Metadados.
- Desenvolver e manter os padrões de Metadados.
- Implementar um ambiente gerenciado de Metadados.
- Criar e manter Metadados.
- Integrar Metadados.
- Gerenciar Repositórios de Metadados.
- Distribuir e entregar Metadados.
- Consulta, Relatórios e Análises sobre Metadados.

A visão sintética é:

a. Entender os requisitos de metadados:

De início é importante entender o que são os metadados, além da definição clichê de “dados sobre os dados”. Os metadados estão para os dados assim como os dados estão para as “coisas”/entidades colocadas sob os processos de um sistema computadorizado. Por exemplo, os objetos, os eventos, as transações e os relacionamentos são as “coisas” controladas num computador, através de sistemas. Assim, os dados definem esses objetos, da mesma forma como os metadados definem os dados. Assim, a gerência de metadados é um processo que controla a sua criação (quando se define, se entende e se documenta aquilo que está sendo objeto do processo), o seu armazenamento (se estrutura, se carrega e se cuida para que seja acessado com facilidade e rapidez), a integração (quando dois ou mais metadados sobre o mesmo o objeto, diferentemente definidos em tempos distintos, por unidades organizacionais distintas, não estão consistentes) e o seu controle (quando se procura mantê-los com qualidade e sobre os quais se define métricas, no sentido de que não se controla aquilo que não se mede).

Um conceito simples e metafórico de metadado é aquela plaquinha que fica ao lado dos “*rechauds*”, nos restaurantes de comida à quilo, indicando o nome do prato, detalhes da sua

composição complementar, a sua localização. Também quando se pensa num catálogo de biblioteca, entende-se com sentido mais computacional o conceito de metadados, ou seja, aqueles elementos que ajudam a entender os objetos, a sua composição, o seu relacionamento, a sua localização, entre outros. O porquê de se gerenciar os metadados? Os metadados aumentam o valor da informação estratégica lhe dando expressão, detalhes, conhecimentos. Isso reduz, de certa forma, o custo do aprendizado, pois as informações sobre os dados estão sempre mais claras. Isso também reduz o tempo gasto na busca pelo entendimento de certos objetos, regras, fórmulas, traduzindo em maior efetividade no seu uso, ou no desenvolvimento de sistemas em torno ou que usam aquele conceito. Assim, os metadados melhoram a comunicação entre a área de negócios e a área que processa a informação (TI). Uma razoável gerência de metadados reduz a redundância acerca daquele conceito, minimizando erros de interpretação que podem ser transformados em falhas graves de sistemas ou produtos.

b. Arquitetura de metadados:

Como a arquitetura dos dados, a de metadados também pode ser centralizada ou descentralizada, dependendo de como os repositórios (DD, Catálogos, etc.) de metadados estão dispostos. Normalmente, os produtos de desenvolvimento de software têm um catálogo próprio (ferramentas Case, SGBD) ou uma área específica onde eles são mantidos. A arquitetura tenta colocar ordem nessa dispersão.

A centralizada impõe as vantagens de um controle mais rigoroso e de menor conflito, visando à criação de uma estrutura única e consolidadora. Tem como desvantagem, por outro, o trabalho de se consolidar os metadados oriundos de várias fontes para colocá-los num único depósito. A descentralizada tem a vantagem de se economizar nos gastos de integração, não havendo persistência centralizada, porém com um custo de busca integrada, em vários depósitos, para se resolver as consultas solicitadas. Uma arquitetura mista envolve a parte da descentralização, com as buscas dinâmicas e outra parte de definição centralizada no catálogo único, onde são colocadas outras definições de metadados, acrescidas aos existentes, para se ajustar adequadamente as definições de negócios da empresa.

c. Desenvolver e manter padrões de metadados:

Os metadados são basicamente de dois tipos: negócios e técnicos. Os metadados de negócios tem o objetivo de documentar os elementos de negócios, centrando num patamar mais conceitual. Envolvem definições de processos de negócios, sistemas, aplicações e aplicativos, regras de negócios, formas de cálculos, algoritmos, linhagem de dados, modelos conceituais e lógicos de dados, aspectos de qualidade de dados e de conceitos de gestores de (meta) dados e das unidades organizacionais responsáveis por eles. Os metadados do ambiente negocial também envolvem regras de CRUD de dados, definição de owners de

dados (UO responsáveis por eles), regras de compartilhamento de dados, papéis e definições sobre os gestores de dados, áreas de assunto, entre outros. Um ponto emergente sobre metadados é a sua definição para DNE, resultante do fenômeno Big Data. Sua visão particular sugere a definição de metadados descritivos (definição, catálogos, etc.), metadados estruturais (formato de áudio, vídeo, email, XML, etc.) e metadados administrativos (direitos de acesso, planos de integração, etc.).

Há padrões formais definidos para os metadados. Os principais são: *Case Definition Interchange Facility* (CDIF), usado para facilitar a troca de metadados entre ferramentas de desenvolvimento, *Dublin Core Metadata Initiative* (DCMI), ISO-11179, que versa sobre definição de padrões e especificações para elementos de dados, e *Common Warehouse Metadata Model* (CWM). Há também sugestões de métricas para se controlar os metadados, como, por exemplo: cobertura de metadados (MD) existentes no escopo desejado (número de objetos já definidos com MD/número de objetos estimados no domínio em análise). Também o grau de cobertura de documentação dos MD (o quanto, em cobertura, os MD estão documentados, sugerindo a completude de sua definição).

d. Implementar um ambiente de metadados:

A implementação de um ambiente de metadados deverá ser revestida de todo cuidado, devendo-se optar por uma abordagem evolutiva e incremental, com estabelecimento de pilotos para verificar conceitos, aderências e adesões.

e. Outros pontos do processo:

O desafio de se criar e manter metadados é muito grande. Daí a ainda baixa incidência de implementação nas empresas. Normalmente se tem modelos isolados oriundos das ferramentas adquiridas, sendo a sua integração um dos grandes desafios. A devida definição de uma arquitetura funcional, prática e que mostre retornos é o grande lance da gestão de metadados. A instanciação dessa gerência se dará pelo gerenciamento adequado dos diversos repositórios, que possam produzir, distribuir e entregar os metadados na forma de consultas, relatórios e análise, no momento exigido e com a devida consistência. Os desafios de metadados são (quase) os mesmos que sempre enfrentamos na área de dados. Aliás, não poderia ser diferente, pois estamos falando dos dados sobre os dados. Um problema, na sua meta referência.

Os metadados técnicos já estão mais associados a elementos de desenvolvimento e implementação, como BD, atributos, modelos físicos de dados, tabelas, campos, *triggers*, aspectos de armazenamento (*storage*), padrões de acesso, frequência e tempo de execução de relatórios e consultas, entre outros. Há também, dentro dos metadados técnicos, uma visão mais operacional, que envolve: necessidades de recursos relativos à operação de TI; informações sobre movimentações de dados (ETL, por exemplo), como transformações e erros; sistemas fontes e *targets*; frequência de *jobs*, erros de "*schedule*"; dados sobre

*backup*s e *recovery*; informações de controles de auditoria, regras de arquivamento e retenção de dados, entre outros.

A Gestão de Metadados se mostra, há muito tempo, como a parte da gestão estratégica de dados com maiores lacunas, dentre todas. Os metadados podem ser considerados como um dos temas mais falados e menos implementados no mundo dos dados. O metadado é como aquela placa que identifica "comida a quilo", que fica ao lado dos *rechauds*. Sem a perfeita identificação dos pratos oferecidos, você não sabe o que está consumindo. Poucas empresas se preocupam com uma arquitetura de metadados, afora aqueles que são produzidos automaticamente pelos SGBD´s para abrigar informações físicas sobre tabelas, campos, índices, *triggers*, entre outros. Mas isso é muito pouco, e nesse particular a Gestão Estratégica de Dados terá muito trabalho pela frente. Algumas empresas, na busca do resgate dos dados e de seus metadados escondidos no ambiente legado, têm adotado técnicas de engenharia reversa, visando o seu levantamento. A Figura 4 mostra, esquematicamente, um fluxo simplificado usando essa abordagem.

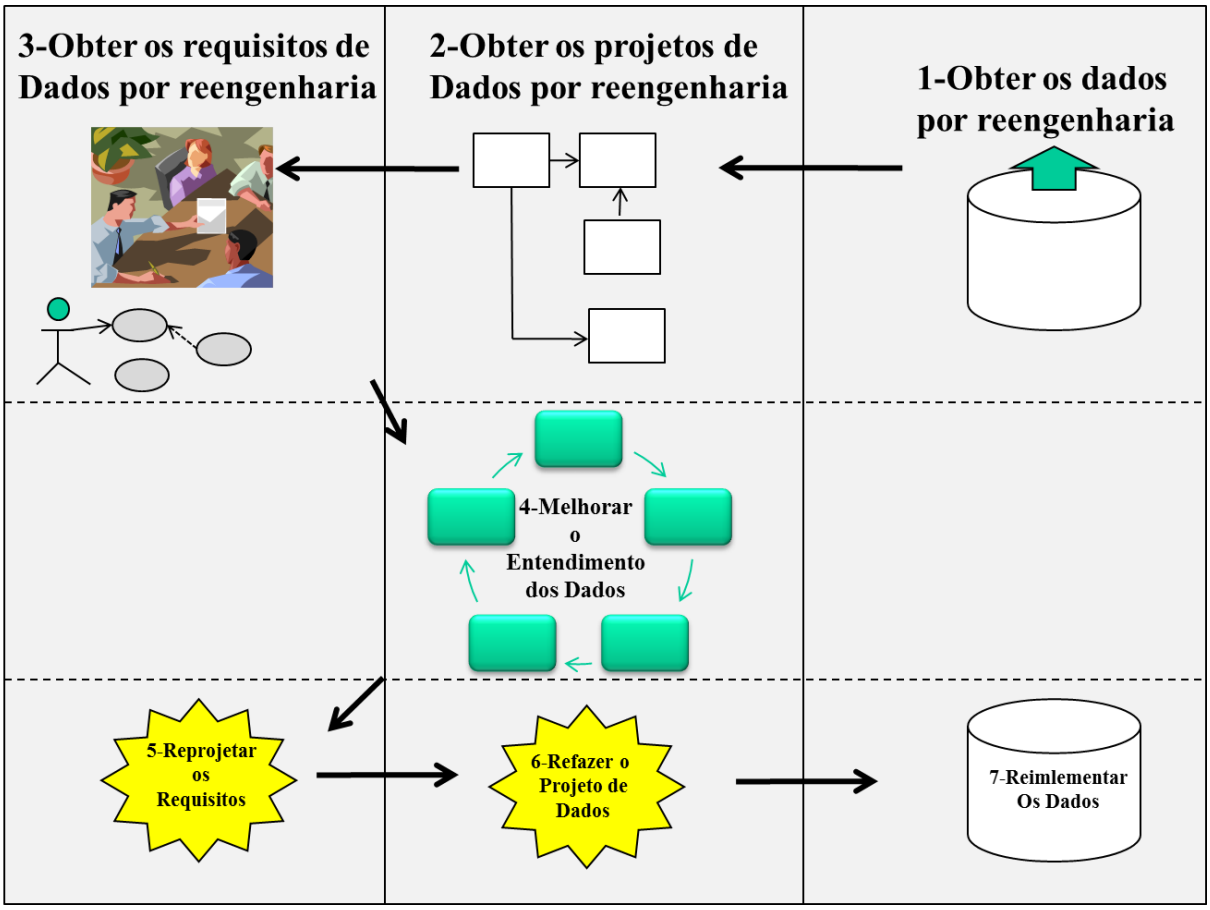

Figura 4 - Engenharia reversa para resgate de dados e metadados

1.10. Gestão de qualidade de dados

O objetivo é planejar, implementar e controlar atividades que apliquem técnicas de gerência de qualidade de dados para medir, avaliar, melhorar e garantir a adequação dos dados ao seu uso pretendido.

A estrutura de atividade desta função é:

- Desenvolver e promover aspectos de conscientização sobre Qualidade de Dados.
- Definir requisitos de Qualidade de Dados.
- Estabelecer processos de "*profiling*", análise e avaliação de Qualidade de Dados.
- Definir métricas para Qualidade de Dados.
- Definir regras de negócios para Qualidade de Dados.
- Testar e validar os requisitos de Qualidade de Dados.
- Definir e avaliar níveis de serviços de Qualidade de Dados.
- Medir e monitorar continuamente a Qualidade de Dados.
- Gerenciar as pendências de Qualidade de Dados.
- Corrigir os defeitos de Qualidade de Dados.
- Projetar e implementar procedimentos operacionais de Gerência de Qualidade de Dados.
- Monitorar os procedimentos operacionais e a performance da Gerência de Qualidade de Dados.

A visão sintética é:

a. Desenvolver e promover aspectos de conscientização sobre Qualidade de dados:

Aqui a grande questão é vender a importância da qualidade de dados nas empresas. É preciso difundir a importância dos conceitos, seja por mecanismos diretos ou indiretos. Os diretos seriam através de palestras, consultorias ou assemelhados. Os meios indiretos são através de exemplos acerca dos efeitos negativos da qualidade de dados nas empresas. No fundo, a ideia é mostrar arranhões na reputação, problemas com normas regulatórias, entre outros. Mostrar também que o problema não é (somente) do domínio da TI, mas principalmente um problema da esfera de negócios. A qualidade de dados deve ser um dos

elementos fundamentais do arco da Governança de Dados da empresa, que define política, padrões, procedimentos, papéis, programas e projetos dentre outros itens do seu escopo, visando tratar e preservar ao ativo "dado". A realização de um trabalho inicial de *profiling* dos dados mais importantes da empresa, mostrando os resultados preocupantes com relação à qualidade dos dados é algo a ser fortemente pensado, pois serve como "*start-up*" para todo o processo de convencimento material sobre os problemas de dados.

b. Definir requisitos de Qualidade de Dados:

Os requisitos de qualidade de dados são definidos diretamente em função das necessidades da empresa. Há que se pensar nos processos críticos da empresa, suas regras de negócios, seus dados consumidos e produzidos e o impacto da qualidade dos dados na sua execução, tanto como *input* quanto *output*. Esse é o início de tudo. Os requisitos de qualidade de dados passam por vários domínios, que podem variar de acordo com os autores. O mostrado pelo DMBOK é:

- Precisão (*accuracy*) ou como as "coisas"/entidades da vida real estão corretamente representadas.
- Completude (*completeness*): O quão completos estão os dados (todos os atributos? Faltam alguns? Todos os essenciais? Alguns acessórios?) exigidos na execução daquele processo de negócio.
- Consistência (*consistency*): Se refere à integridade cruzada entre duas ou mais fontes que armazenam o mesmo dado. Há coerência entre esses dados que habitam fontes diferentes? A coerência existe no mesmo contexto ou em contextos diferentes?
- Atualidade (*currency*): O quanto os dados estão atualizados e representam o estado corrente e mais atual.
- Precisão numérica (*accuracy*): Representação de valores no grau de precisão necessária, como casas decimais para dados numéricos.
- Disponibilidade (*availability*): O dado é disponibilizado no momento de sua necessidade?
- Unicidade (*uniquiness*): O fato de haver representação única de certa entidade, sem ambiguidade ou sentidos diferentes.

c. Estabelecer processos de "*profiling*", análise e avaliação de Qualidade de Dados:

Refere-se à necessária fotografia inicial do estado dos dados de certa(s) área(s) de assunto ou domínio(s) da empresa. Deve ser uma das primeiras ações para se estabelecer as "*baselines*" dos processos de melhoria de qualidade de dados da empresa. Permite criar as primeiras métricas e a definição dos objetivos a serem alcançados em função delas. É como se fosse a analise laboratorial solicitada por um médico para melhor diagnosticar o estado do paciente e iniciar o seu tratamento.

d. Definir métricas para Qualidade de Dados:

As métricas deverão ser definidas para a avaliação do estado atual e da evolução dos tratamentos de qualidade dos dados. As métricas, como todas as medidas definidas em processos de qualidade deverão:

- Ser atreladas a objetivos bem definidos.
- Responder a questões associadas a esses objetivos.
- Ser medidas definidas com clareza, que apontem elementos quantificáveis associáveis a objetivos de negócios, com formulações claras (como medir), valores definidos para análise (como analisar), com faixas aceitáveis e não aceitáveis (como interpretar), plano de ações no caso de discrepâncias, frequência de medição (quando medir), entre outros.

e. Definir regras de negócios para Qualidade de Dados:

Implica na análise das regras de negócios fundamentais dos processos e na descoberta dos dados que podem implicar em eventuais quebras de conformidade delas. Esses dados deverão ser observados na sua qualidade justamente para garantir a conformidade da regra com o processo. Por exemplo, a regra de negócios que define que nenhum colaborador com oito horas diárias de turno de trabalho poderá ganhar menos que o salário mínimo. Um campo de um arquivo enviado ao INSS contendo um valor abaixo desse estabelecido implica numa quebra de conformidade do processo (admissão, por exemplo), com as regras de negócios definidas.

f. Testar e validar os requisitos de Qualidade de Dados:

Nesse ponto, o DMBOK sugere que haja processo de verificação inicial (*data profiling*, por exemplo) e verificação constante e recorrente, a fim de que os dados sejam sempre avaliados nos seus domínios de qualidade.

g. Definir e avaliar níveis de serviços de Qualidade de Dados:

Nesse item, é sugerida a definição de níveis de serviços de qualidade de dados, o que deverá ser garantido por medições e verificações constantes. Os níveis de SLA são o compromisso firmado sobre qualidade da área gestora dos dados com os seus usuários. Os itens subsequentes, relativos a medir e monitorar continuamente, gerenciar as pendências e corrigir os defeitos são consequências diretas desse compromisso de nível de serviços.

h. Projetar e implementar procedimentos operacionais de Gerência de Qualidade de dados e monitorar os procedimentos operacionais e a performance da Gerência de Qualidade de Dados:

São, no fundo, a materialização do processo de Garantia de Qualidade dos Dados. Todo processo deverá ser constantemente avaliado a fim de se obter melhorias nos seus procedimentos, políticas e resultados.

2. CONCLUSÕES

Há, hoje no mercado, diversos *frameworks* sobre Governança de Dados, conforme discutidos no Blog do Barbi (Carlos Barbieri), em *posts* ao longo de 2012. O *framework* da Dama é certamente o mais completo e detalhado, pois envolve a Governança de Dados e todas as áreas associadas a ela. A trajetória de uma empresa em direção à Gestão de Dados (*Data Management*) requer muitos cuidados, exatamente pelas características fluídas deste elemento dentro da empresa, não bastando somente a adoção de um *framework* de referência.

A Fumsoft, por meio de seu setor de Qualidade, o qual coordeno, adquiriu ao longo desses últimos dez anos, uma sólida experiência em implementações de processos, cuja tônica do desafio é essencialmente a mesma exigida em empresas que queiram melhorar a sua gestão de dados. É preciso haver uma forte mudança cultural. Há que se buscar uma maturidade gradativa de dados, da mesma forma com que as empresas buscam a maturidade em processos, trilhando os caminhos do MPS.BR e/ou CMMI. No livro BI2 - Modelagem e Qualidade (Barbieri, 2011) foram apresentados e discutidos modelos de maturidade de dados, alguns dos quais centrados nas práticas consagradas de maturidade de software. Na última edição da *Data Management Conference – Latin America* (DMC Latam), em agosto de 2012, foi apresentada uma visão sobre níveis de maturidade em dados, conforme a Figura 5.

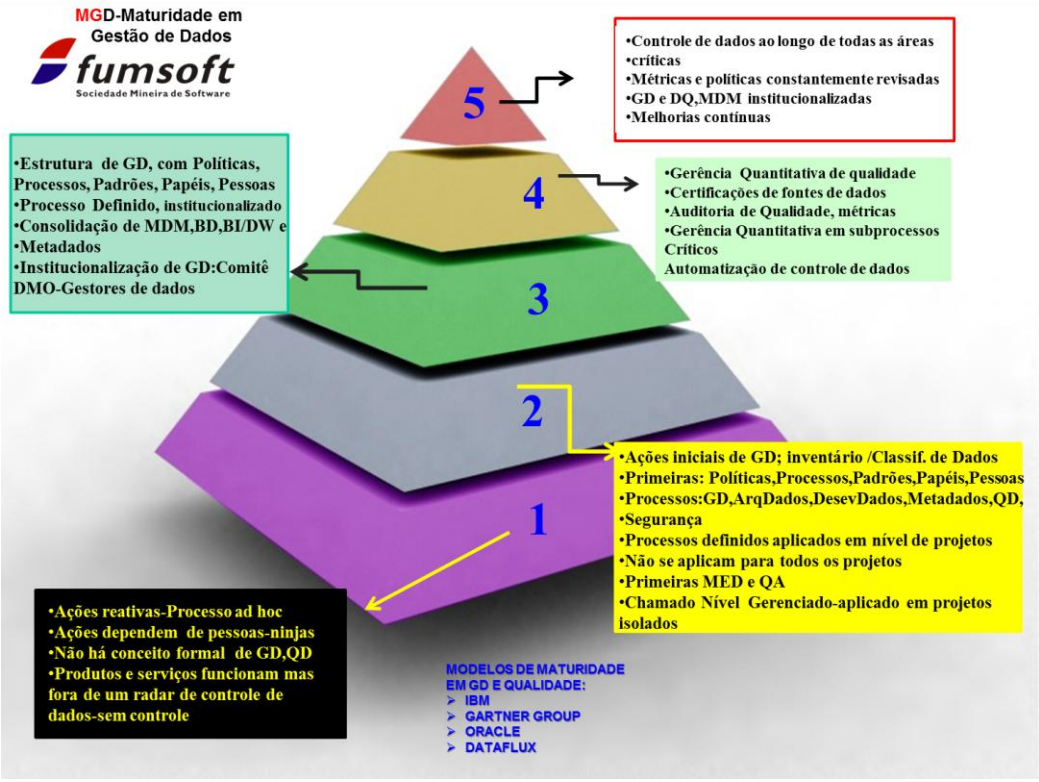

Figura 5 - Camadas de maturidade em dados

Além da maturidade, as empresas pretendentes a desenvolver ações de gestão de dados deverão ter o claro entendimento dos passos desse caminho. No curso de pós-graduação na PUC-MG, desenvolvemos com os alunos, na forma de comunicação visual de ideias, os conceitos fundamentais de Governança e Gestão de Dados, sintetizados nos nove P´s, conforme a Figura 6. Pense neles e veja em que camada de maturidade a sua empresa se encontra. Bom desafio!

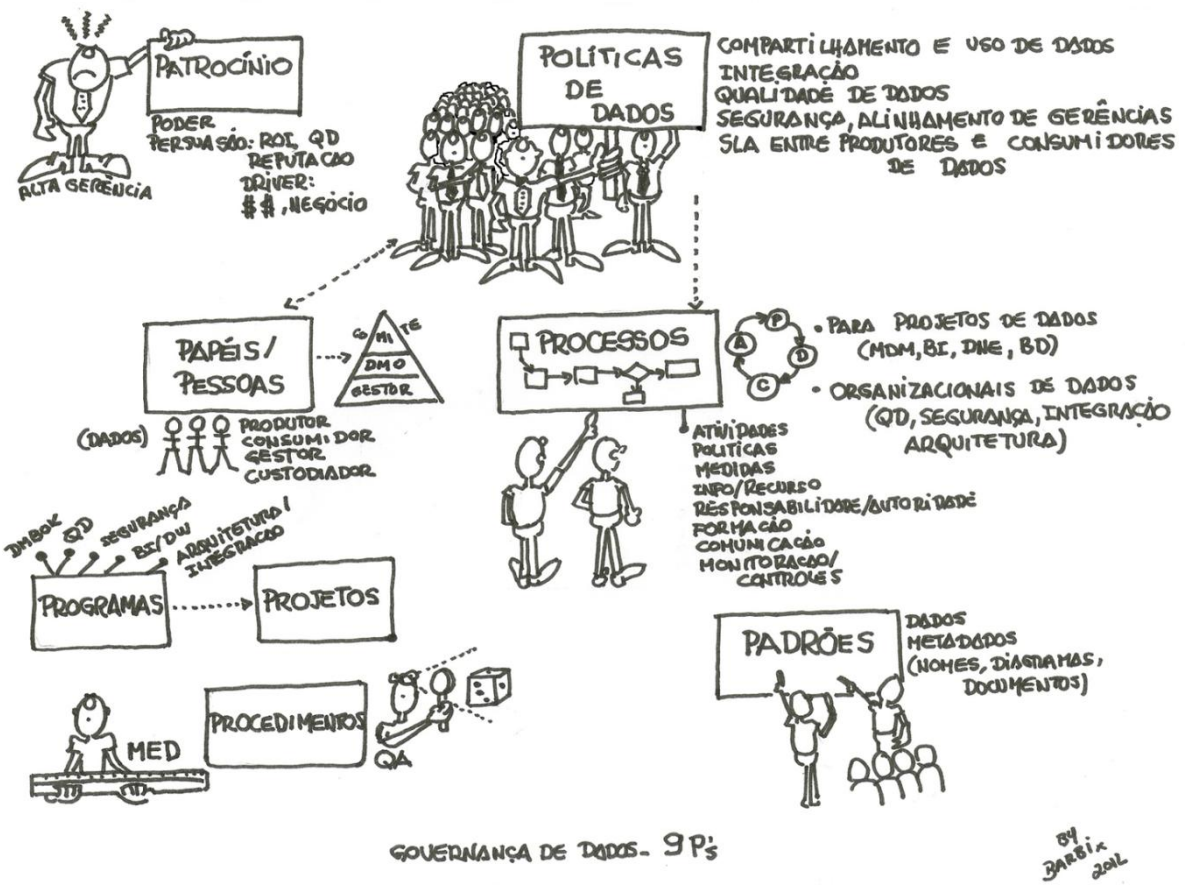

Figura 6 - Nove P´s da Gestão e Governança de Dados

3. REFERÊNCIAS BIBLIOGRÁFICAS

BARBIERI, C. *BI2 – **Business Intelligence - Modelagem e Qualidade***. Elsevier, 2011.

BARBIERI, C. Posts sobre Governança de Dados, Big Data, entre outros. Disponível em http://blogdobarbi.blogspot.com

DMBOK. MOSLEY, M. ; BRACKETT, M.; EARLEY, S. HENDERSON, D. ***DAMA Guia para o corpo de conhecimento em gerenciamento de dados***. Technics Publications, versão brasileira 2012.

DMBOK. MOSLEY, M.; BRACKETT, M.; EARLEY, S.; HENDERSON, D. ***The DAMA Guide to The Data Management Body of Knowledge: DAMA*** - DMBOK Guide. 1. ed. Estados Unidos: Technics Publications, 2009.

ELMASRI. R. ; NAVATHE. S. ***Fundamental of Data Base Systems***: Addison Wesley, 2000.

SADALAGE P.; FOWLER, M. ***NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence***. Addison-Wesley, 2013.

SERPRO. ***Modelo Global de dados - Integração de dados e processos.*** Disponível em http:// http://modeloglobaldados.serpro.gov.br/. Acesso em 22 de junho de 2012.

SOARES, S. ***Big Data Governance: An Emerging Imperative***. Mc Press, 2012.

SOFTEX - ASSOCIAÇÃO PARA PROMOÇÃO DA EXCELÊNCIA DO SOFTWARE BRASILEIRO. ***MPS.BR – Guia de Implementação – Parte 5: Fundamentação para Implementação do Nível C do MR-MPS:2009***, 2009. Disponível em: http://www.softex.br