Fernando Amaral

BIG DATA:
UMA VISÃO GERENCIAL

São Paulo

PoloBooks

2016

Dados Internacionais de Catalogação na Publicação (CIP)

Amaral, Fernando

Big Data: uma visão gerencial / Fernando Amaral. -- 1. ed. São Paulo : PoloBooks, 2016.

122 p.; 14x21cm

ISBN: 978-85-5522-111-8

1. Ciência da computação. 2. Gerenciamento. 3. *Software*. I. Título

CDD 004

Índice para catálogo sistemático
1 Ciência da computação : 2 Gerenciamento : 3 : *Software*

BIG DATA:
Uma Visão Gerencial

1ª edição: agosto de 2016

Capa: Thiago Francisco
Revisão gramatical: Eulália Érica Dutra dos Santos
Diagramação: Anne Charlyne Raviani
Impressão: PoloPrinter

Gráfica PoloPrinter
11 . 3791.2965 11 . 98393.7000
www.poloprinter.com.br
polo.books
atendimento@poloprinter.com.br

SUMÁRIO

PREFÁCIO

Nas últimas décadas, uma das grandes preocupações da ciência da computação tem sido os processos de desenvolvimento de software: muito se pesquisou, debateu e se experimentou sobre o assunto e surgiram muitas metodologias, algumas extremamente burocráticas e outras prometendo mais agilidade e clientes mais satisfeitos. Todo esse movimento tem um motivo: criar software não é fácil. Estima-se que apenas 32% dos projetos de software são um sucesso completo[1], ou seja, um verdadeiro caos. Nenhuma destas metodologias proporcionou um milagre: desenvolver softwares continua sendo algo desafiador, de difícil sucesso completo. Mas por quê? Software, em geral, é algo complexo, abstrato, cujas tecnologias mudam rapidamente. Ainda, existe um sentimento, mesmo para quem é da área de tecnologia, de que é algo rápido e banal para ser construído ou alterado, sem demais consequências. Além disso, desenvolver softwares é ainda uma ciência imatura.

Mas o que isso tem a ver com análise de dados e *Big Data*? A indústria do software, aos poucos, percebeu que a gerência de projetos aplicada a desenvolvimento de software é um investimento com alto ROI. Claro, usar metodologia de gerência de projetos na produção de software, assim como em qualquer outra indústria, não é garantia de sucesso. Porém, comprovadamente, minimizam seus riscos e maximizam as chances de sucesso.

Da mesma forma, projetos de análise de dados e *Big Data* são, usualmente, de complexidades ainda maiores do que os pro-

1 Relatório CHAOS (2011), do *Standish Group*.

jetos enfrentados pela indústria de software com percentuais de fracasso em torno de 55%. Mas claro que o problema não é apenas a complexidade. Faz parte da construção deste índice de fracassos a imaturidade. Já há algum tempo que se fala em *Big Data*, o documento com a teoria sobre *MapReduce* foi publicado em 2004, ou seja, há mais de 10 anos; a primeira versão estável do *Hadoop* foi liberada em 2012. Mas o amadurecimento de uma nova tecnologia pode levar muitas décadas. Da mesma forma que em outras indústrias e no próprio desenvolvimento de software, a gerência de projetos adequada irá aumentar as chances de um projeto que entregue valor, dentro do planejamento estabelecido.

Este é o objetivo da obra: fornecer dicas para que projetos de *Big Data* e análise de dados, em geral, tenham maiores chances de sucesso, com entrega de valor.

QUEM DEVE LER ESTA OBRA

Embora o título fale em visão gerencial, esta obra pode ser útil para todos aqueles que atuam em projetos de *Big Data*: partes interessadas, executivos, analistas de negócio, arquitetos, desenvolvedores, cientistas de dados e, é claro, gerentes de projetos. Também é uma ótima leitura para aqueles envolvidos em vendas de produtos ou serviços de *Big Data*, seja em nível gerencial ou operacional.

Este não é um livro de conceitos ou aplicações de gerência de projetos, portanto, considero que o leitor já tem um conhecimento básico sobre o assunto. Se ainda não tem experiência ou educação na área, sugiro que procure, existe vasta oferta de conteúdo.

Esta obra é uma coletânea de boas práticas, baseadas em experiência pessoal do autor em vários anos atuando nestes tipos de projetos. Embora exista um capítulo contendo um checklist, esta obra não deve ser usada como um guia passo a passo, pois cada projeto é totalmente único com relação a objetivos, pessoas, tecnologias e dados envolvidos. A obra deve, sim, ser usada como

um manual de boas práticas, que não necessariamente devem ser sempre aplicadas, ou que podem ser aplicadas em diferente "intensidade" da aqui descrita. Cabe à equipe do projeto avaliar o que e de que forma, o que está aqui descrito deve ser aplicado.

ESTRUTURA DA OBRA

A obra está estruturada em 12 capítulos. A sugestão, obviamente, é que sejam lidos em sequência. O capítulo 12, que contém um checklist, pode ser consultado diretamente a qualquer momento durante um projeto, para se certificar de que nada foi esquecido. Vamos ver o que é estudado em cada capítulo:

- → Capítulo 1, Introdução: neste capítulo, vamos falar sobre o que é *Big Data* e Análise de Dados. Também vamos ver o que diferencia um projeto tradicional de análise de dados de um projeto de *Big Data*, e vamos entender um pouco, porque desenvolver este tipo de projeto não é uma tarefa trivial. Vamos estudar como estes projetos estão estruturados, e como eles são classificados sob diversos aspectos. Finalmente, vamos ver quais são as características ideais de um gerente de projetos para atuar em *Big Data*.
- → Capítulo 2, Antes do Projeto: talvez, você não saiba o que significa um sistema sombra, mas com certeza já conhecer diversos deles em empresas por onde passou. O capítulo 2 traz vários aspectos que devem ser observados antes de começar a implementar *Big Data*. Por exemplo, quais são os critérios e fatores de sucesso, e se o projeto entrega valor. O capítulo 2 está bastante relacionado ao capítulo 3, que trata de riscos, já que muitos dos aspectos mencionados no capítulo 2, se forem subestimados, podem pôr em risco o sucesso do projeto.
- → Capítulo 3, Riscos: ignorar a gestão de riscos em seu projeto de *Big Data* é assinar uma carta de intenção com o fracasso. Além de abordar alguns aspectos da gestão de riscos,

este capítulo traz os principais riscos que ameaçam projetos de *Big Data*. Alguns, como cronogramas não realistas e orçamentos inadequados, são comuns em projetos em várias indústrias. Outros são exclusivos de projetos de análise de dados e *Big Data*. Por exemplo, pode-se fazer um empreendimento de meses para depois descobrir que o modelo construído não terá um bom desempenho devido à qualidade dos dados.

- Capítulo 4, Escopo: podemos usar técnicas clássicas de gestão de projetos para definirmos o escopo de um projeto de *Big Data*. Vamos ver que alguns cuidados especiais devem ser tomados como, por exemplo, se o escopo está alinhado com as estratégias organizacionais ou é um projeto para atender a um anseio pessoal.
- Capítulo 5, Tempo: quando se constrói algo intangível de longa duração e cujos primeiros resultados podem levar muito tempo para serem entregues, gerenciar tempo torna-se uma tarefa complexa, porém, necessária. Neste capítulo, são apresentadas algumas dicas de gestão de tempo em projetos de *Big Data*.
- Capítulo 6, Custos: não se engane com artigos que falam em soluções milagrosas e quase de graça com "hardware comum". Não que eles não possam ser usados, mas os custos de projetos de *Big Data* não estão apenas em dois ou três servidores. Além de infraestrutura, projetos de grande porte podem levar muitos meses ou anos, envolvendo muitos custos diretos e indiretos. Neste capítulo, vamos falar de gestão de custos, indispensável para o sucesso destes tipos de projeto.
- Capítulo 7, Qualidade: mesmo que você não esteja há muito tempo na tecnologia da informação, já deve ter ouvido o jargão: "lixo entra, lixo sai". Em *Big Data*, mais do que nunca, esta é uma verdade. O gerente de projetos não deve se preocupar somente com a qualidade do produto que vai criar, mas também com os dados que serão cap-

turados e carregados para gerar informação. Este capítulo trata de aspectos da qualidade em gerência de projetos de *Big Data*.

- Capítulo 8, Recursos Humanos e Partes Interessadas: uma equipe de projetos de *Big Data* pode ser formada por um grande número de diferentes profissionais, muitos deles podem não serem fáceis de encontrar disponíveis no mercado. Este capítulo fala sobre as competências necessárias, de como motivá-las e qualificá-las. Também é assunto a evangelização de partes interessadas, um tema de extrema importância, quando estamos tratando de algo ainda um pouco obscuro: *Big Data*!
- Capítulo 9, Aquisições: a maioria dos projetos irá envolver a aquisição de produtos ou serviços. Estas aquisições devem seguir etapas de decisão de fazer ou comprar, identificação de necessidades, recebimentos de propostas, seleção de propostas, negociação e compra. Podem envolver ainda *benchmarking* entre alguns fornecedores. Neste capítulo, trazemos algumas dicas importantes sobre os processos de aquisições.
- Capítulo 10, *Big Data* Ágil: é possível usar metodologia Ágil em gerência de projetos de *Big Data*? Claro! Mas existem situações em que não é possível, e outras em que não é recomendável. Neste capítulo, estas situações são discutidas.
- Capítulo 11, Estabelecendo um Processo: paralelamente aos processos de gerência de projetos, este capítulo traz algumas etapas que devem ser executadas no desenvolvimento de projetos de *Big Data*.
- Capítulo 12, Checklist: este capítulo apresenta um checklist, com dezenas de itens divididos em várias categorias, que a equipe do projeto pode usar como referência em seus projetos, riscando o que não se aplica ou que já se tem controle, ou definindo responsáveis e datas de resolução para os demais itens.

Ao final da obra, existe ainda um glossário, onde o leitor deve consultar durante a leitura, caso encontre algum termo ou acrônimo que não tenha familiaridade.

1. INTRODUÇÃO

Na era da informação e conhecimento, analisar dados não é uma atividade qualquer para empresas e governos, é uma questão de sobrevivência. Em um mundo globalizado, e cada vez mais competitivo, em que dados eletrônicos são produzidos de forma exponencial, quem for capaz de extrair informação e conhecimento de dados de forma eficiente, venderá mais, produzirá mais, gastará menos, terá clientes mais satisfeitos, fornecedores mais eficientes e estará em conformidade com agências reguladoras e fiscalizadoras. Não somos capazes ainda de avaliar com precisão como essa nova era que está surgindo será, mas sabemos que a produção e análise de dados terão um papel fundamental.

E como se extrai informação e conhecimento de dados? Implementando projetos de *Big Data*. Um projeto, segundo definição do PMBOK (2013), é um esforço temporário para produzir um produto ou serviço exclusivo. Um projeto de *Big Data* possui estas mesmas características: é um esforço com início e fim, portanto temporário, e produz algo até então inexistente, e, portanto, exclusivo. Embora o esforço seja temporário, o resultado produzido não é. Podemos sim, por exemplo, empreender um projeto de auditoria, que uma vez encerrado, seu resultado é entregue: encerra-se o projeto e o produto criado também chega ao fim. Mas de uma maneira geral, um projeto de *Big Data* gera um produto ou serviço que vai entrar em seu ciclo de vida útil, gerando resultados por tempo indeterminado. É o que acontece, por exemplo, quando se constrói um armazém de dados. A construção é um projeto

com escopo, custos e tempo definidos. Quando pronto, o projeto termina, mas o armazém continua sendo carregado com dados históricos e gerando informação e conhecimento relevantes para a tomada de decisão.

Tudo isso para dizer que, de uma maneira geral, um projeto de *Big Data* nada mais é do que um projeto qualquer, com início e fim, escopo, orçamento, cronograma, entre outros. Então, por que uma obra específica para tratar destes projetos? Porque este tipo de projeto está entre os mais complexos e com maiores riscos de todas as indústrias. Anteriormente, comentamos que, semelhante à produção de software, projetos de *Big Data* produzem algo abstrato, não tangível, de extrema complexidade no sentido de poder envolver muitas estruturas, sistemas, formatos, pessoas e infraestrutura.

Nesta obra, vamos usar alguns termos que precisamos definir antes, de forma a não haver um entendimento ambíguo do seu significado:

- Projeto de análise de dados ou de *Big Data*: Nesta obra, vamos usar projetos de análise de dados ou de *Big Data* como uma definição genérica para qualquer tipo de projeto que envolva extrair informação e conhecimento de dados: analisar dados pode ser produzir um relatório, construir um *data mart*, analisar sentimentos em redes sociais, criar um modelo para uma análise preditiva, capturar dados de sensores para avaliar a saúde dos equipamentos na linha de produção, classificar *pentabytes* de documentos jurídicos ou, simplesmente, rodar uma consulta para ordenar as vendas no mês. De uma maneira genérica, vamos nos referir a projetos de *Big Data* e Análise de Dados como B&A. Quando a referência ao longo da obra não for genérica, usaremos um termo mais específico, como por exemplo, análises preditivas.
- Dados de origem: analisar dados requer coletar dados de algum lugar. Vamos usar dados de origem para nos referirmos a dados ou conjuntos de dados que são utilizados para popular um modelo, ser carregado num processo de extração e

transformação, ser ordenado em uma planilha eletrônica ou mesmo sobre o qual se executa uma consulta SQL.

- Dados de *Staging*: muitos processos de análises de dados possuem uma etapa intermediária, entre os dados de origem e o resultado dos dados já consolidados. Esta etapa intermediária serve para que os dados sejam transformados, somados, resumidos, juntados ou calculados. O termo *staging* na obra se refere a qualquer processo intermediário, não apenas aquele conhecido num processo de construção de um *data warehouse* tradicional.
- Dados de destino: aqui, estaremos sempre nos referindo ao resultado: o cubo, a previsão de uma análise preditiva, o relatório, o arquivo gerado, entre outros.

O PMBOK nos ensina que nem todos os seus processos são obrigatórios. Tampouco temos que aplicar os processos sempre com a mesma intensidade. Quais processos usar e em qual intensidade depende das decisões tomadas pelo gerente do projeto junto à sua equipe, e deve estar diretamente relacionada a fatores como: complexidade do projeto; elementos novos relacionados a pessoas, tecnologias ou dados de origem.

O QUE ENTENDEMOS POR *BIG DATA*

Se você está lendo esta obra, provavelmente, já ouviu e leu muito sobre *Big Data*, e não será nosso objetivo aqui apresentar este conceito. Porém, muitos entendimentos são ambíguos. Criou-se quase uma relação obrigatória entre *Big Data* e tecnologias de *MapReduce* como Hadoop. No mínimo, associa-se o termo a grandes volumes de dados, pentabytes ou mais, ignorando-se todos os demais "Vs" (velocidade, volume, variedade, veracidade e valor). Nesta obra, vamos ter uma definição mais abrangente para projetos de B&A (*Big Data* e Análise de Dados). Se o projeto envolve uma única fonte de dados relacionais para produzir dimensões em um *data mart*, não podemos considerar *Big Data*, mas em outras

hipóteses, em que existem muitas fontes de dados, ou algumas poucas, mas com, pelo menos, uma não estrutura, ou mesmo dados semiestruturados ou fontes de dados NoSQL, ou até mesmo, claro, volumes de dados além de um projeto tradicional, então sim, todos estes serão considerados projetos de *Big Data*.

Falamos em seção anterior, mas vamos repetir: cabe ao gerente de projeto decidir quais processos e com qual intensidade eles serão aplicados, conforme a complexidade do projeto. Da mesma forma, todas as recomendações e técnicas apresentadas nesta obra, específicas para a análise de dados, devem ser avaliadas pelo gerente e sua equipe, se devem ser aplicadas e com qual intensidade. Por exemplo, vamos falar da criação de protótipos ao longo da obra. Um determinado projeto com baixa complexidade, do qual já se tenha experiência prévia, tecnologia madura e testada, talvez, dispense totalmente um protótipo. Já outro projeto de média complexidade, que terá uma entrega de artefatos de visualização, pode ter protótipos gerados em uma ferramenta de maquetes para simular telas. Finalmente, um grande projeto, de longo prazo e grande complexidade, pode requerer que um protótipo funcional seja gerado para verificar a viabilidade e minimizar riscos.

PROJETOS TRADICIONAIS VERSUS PROJETOS DE *BIG DATA*

Desde a pré-história, o homem analisa dados. A análise de dados eletrônicos é um evento mais recente, com pouco mais de 70 anos. A capacidade de usar um computador para analisar dados vem mudando o mundo. Vamos pensar em um exemplo simples: auditar folha de pagamento. Tradicionalmente, a auditoria teria que ser manual, por amostragem, e mesmo assim, poderia levar semanas. Com o computador, o processo passou a ser totalmente eletrônico, a prova de erros, milhões de vezes mais rápida e de forma contínua.

Mas a análise de dados só começou a tomar força na década de 90, foi quando os grandes armazéns de dados, ou *data*

warehouse, começaram a se tornar mais comuns para apoiar as empresas a tomarem decisões. De lá para cá, armazéns de dados se popularizaram, encontrados em praticamente qualquer grande empresa.

Mas o que diferencia um projeto de análise de dados tradicional, como os popularizados nos anos 90, de um projeto de *Big Data*? Primeiro, os "Vs" que falamos na seção anterior: de velocidade, volume, variedade, veracidade e valor. Veracidade e valor já eram características de projetos tradicionais de análise de dados. Vamos entender o que os outros três "Vs" representam em relação a projetos tradicionais:

- Velocidade: a velocidade diz respeito não somente a da produção do dado em si, mas a velocidade do processamento e produção de informação e conhecimento, visto que o valor da informação é inversamente proporcional ao tempo em que ocorreu o evento que gerou o dado. Por exemplo, ocorreu uma falha em um equipamento. O operador recebe o alerta 5 segundos depois da falha: ele desliga o equipamento, identifica a falha, aperta o colar de suporte vibratório e a linha de produção volta a produzir em 10 minutos. Porém, imaginando outro cenário, se ele recebe o alerta em 10 segundos: a máquina tem que ser desmontada para troca do colar de suporte vibratório, e a linha de produção por 2 horas. Mas se por algum problema, ele só recebe o alerta em 30 segundos: o colar de suporte vibratório se rompe e causa um estrago em série na máquina. O equipamento deve ser substituído, seu custo de reparação será na casa dos 6 dígitos e a linha de produção tem que parar por 6 horas, causando um prejuízo de centenas de milhares de reais. O gráfico da *Figura 1* abaixo mostra a relação inversa entre tempo e valor da informação. Toda a informação chegará a um tempo em que seu valor é zero, isso pode ocorrer após alguns segundos ou até mesmo depois de centenas de anos. É preciso deixar claro que estamos falando de valor da informação, e não do dado: no

exemplo acima, quando o dado é extraído do equipamento, ele ainda não tem valor. Só depois que é processado, ele passa a ser informação e ter valor.

- Volume: projetos tradicionais eram construídos em armazéns de dados, contendo por volta de *terabytes* de dados. Projetos de *Big Data* são de *petabytes* ou mais.
- Variedade: projetos tradicionais carregavam dados estruturados de sistemas de operação tradicionais, em modelos relacionais, hierárquicos ou de rede. Projetos de *Big Data* processam estes tipos de dados, que na verdade, ainda são a maioria, mas também incluem dados não estruturados e semiestruturados de redes sociais, sensores, web, documentos, e-mails, etc.

Mas, além dos "Vs", existem outras diferenças significativas que devem ser consideradas em projetos de B&A. Vamos ver alguns:

Primeiro, do ponto de vista de arquitetura: projetos tradicionais têm uma arquitetura centralizada, enquanto *Big Data* é

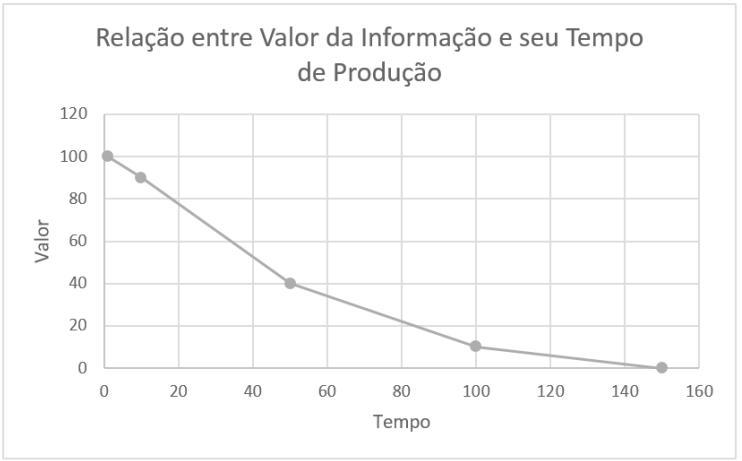

Figura 1: Relação entre valor e tempo de produção da informação.

distribuída. Se o projeto tradicional precisar crescer, este crescimento é feito de forma vertical: adiciona-se mais memória, poder de processamento e armazenamento ao servidor. Projetos de *Big Data* crescem horizontalmente, em vez de um "*upgrade*" no servidor, novos nodos, como *data nodes* ou *name nodes*, são adicionados à estrutura de B&A, geralmente, virtualizados. Projetos tradicionais trafegam mais dados do servidor para o cliente ou do servidor para fora da empresa. Em projetos de *Big Data*, o maior tráfego é entre os nós dos servidores: entre *data nodes*, entre *names nodes* e *data notes* etc.

Em projetos tradicionais, existe uma grande preocupação em só carregar dados em que, a partir de uma análise prévia, se vê valor. Estes dados são tratados e carregados em repositórios pequenos (para os padrões de *Big Data*) para apoiar decisões. Por outro lado, projetos de *Big Data* carregam grandes volumes de dados em um sistema de arquivo como o HDFS, em seu formato nativo, mesmo que em princípio não se enxergue valor. Estes dados formam os conhecidos *data lakes*, ou lagos de dados. Posteriormente, parte destes dados pode ser transformados e carregados em um *data mart* tradicional.

Outra forma que podemos olhar uma solução de *Big Data* é sob sua arquitetura básica. Neste contexto, temos quatro elementos: fontes de dados, carga, armazenamento, análise e visualização (ou apresentação). Neste aspecto, em uma solução de análise de dados clássica estão presentes estes mesmos elementos de arquitetura, o que mudam são algumas particularidades em alguns elementos. Vamos entender melhor.

Quantos às fontes de dados, podemos ter nos dois casos os mesmos elementos: dados estruturados ou não estruturados. Porém, projetos de *Big Data* têm mais presente fontes de dados não estruturadas, como já estudamos no quesito variedade. Na carga, também temos elementos semelhantes nos dois casos: a carga pode ser feita por um processo de carga, um *web service* ou mesmo uma API. No armazenamento começam as diferenças. Um projeto tradicional armazena dados em um banco de dados

relacional ou dimensional, enquanto projetos de *Big Data* armazenam dados em sistemas de arquivos distribuídos com HDFS e bancos de dados NoSQL. Na etapa de análise, em que o objetivo é transformar dados em informação, também há alguma semelhança entre projetos tradicionais e de *Big Data*: em comum, pode-se usar consultas com uma linguagem estruturada (Pig Latin ou SQL) ou um algoritmo de aprendizado de máquina. Já projetos tradicionais usam cubos, enquanto *Big Data* usa *map reduce*. Na visualização, temos novamente os mesmos elementos: painéis, relatórios, KPIs, entre outros.

O TRIÂNGULO DOS PROJETOS DE *BIG DATA*

Projetos de B&A não são criados para analisar dados, configurar clusters, contratar consultores, ou programar scripts: estes elementos são meios para chegar ao objetivo do projeto: responder perguntas. B&A são altamente complexos e envolvem um grande número de variáveis. Se pudermos agrupar estes elementos em grupos, eles seriam quatro, citados aqui em ordem decrescente de importância: pessoas, processos, dados e tecnologia.

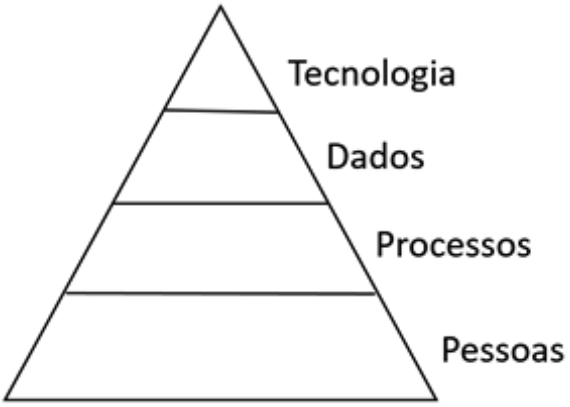

Figura 2: O Triângulo dos projetos de *Big Data*.

As pessoas são, sem dúvida, o elemento mais importante. Projetos de B&A são feitos por pessoas para pessoas, elas estarão sempre envolvidas no processo: produzem os dados de origem, configuram os sistemas de processamento, programam os scripts, definem as regras de negócios, fornecem suporte, interpretam a legislação, geram os relatórios, parametrizam os KPIs, definem os requisitos técnicos e funcionais das soluções, entre outros. A gestão de pessoas, nestes projetos, são tão ou mais importantes do que em outros projetos de tecnologia da informação.

Depois, temos os processos de negócio. Estes processos que vão definir os requisitos funcionais do projeto. Eles podem vir de pessoas, de normas internas, de necessidade gerenciais, de legislação, de conformidade ou de regras de mercado.

Temos, então, o dado, que é a matéria-prima essencial. Poderíamos fazer uma analogia de um projeto de B&A com uma máquina em uma linha de produção. De um lado, entram dados. Estes dados vão entrar incompletos, inconsistentes, imprecisos e inexatos. O primeiro processo desta linha de produção é trata-los, deixando-os com a qualidade mínima para os objetivos do projeto. Além disso, os dados podem ter suas estruturas modificadas de diversas formas. O dado ainda será processado por um algum algoritmo, com uma função matemática, estatística ou de álgebra relacional, a fim de chegar ao seu formato, em que o usuário seja capaz de extrair informação e conhecimento.

Por fim, tecnologia: projetos de dados estão cercados de tecnologia: conectores, softwares de extração, softwares de transformação, softwares de qualidade de dados, algoritmos de aprendizado de máquina, software de visualização, discos rígidos, clusters, roteadores, switches, etc.

TIPOS DE PROJETOS

Projetos de B&A podem ser classificados de muitas formas, vamos ver nesta seção algumas delas.

QUANTO AO TIPO DE NEGÓCIO

Do ponto de vista de tipo de negócio, um projeto de *Big Data* pode ser classificado de duas formas: o que busca trazer uma vantagem competitiva, ou aquele que cria um produto ou serviço de dados. Buscar uma vantagem competitiva é resolver um problema de negócio, os clássicos vender mais, gastar menos, ser mais eficiente. Um produto ou serviço é oferecer algo em forma de valorizar dados: um mecanismo de buscas, um serviço de previsão de preços de passagens áreas, um BI self-service, entre outros.

QUANTO AOS OBJETIVOS

A segunda forma de classificação destes projetos é quanto aos objetivos. Vejamos alguns:

- Negócios: é o tipo mais comum, buscam melhor eficiência e eficácia de negócios: os já mencionados vender mais, produzir mais, perder menos.
- Auditoria: buscam encontrar fraudes, desperdícios e erros. Pode ser uma iniciativa interna ou externa da organização.
- Conformidade: projetos para atender a alguma norma de conformidade. A conformidade está associada a alguma entidade reguladora ou fiscalizadora sob a qual a empresa está inserida. Alguns tipos de conformidade, com os referentes ao eSocial brasileiro, requerem a produção de dados ao invés de artefatos de visualização, em outras palavras, estes tipos de projetos geram arquivos produzidos em um layout pré-definido, ao invés de um relatório ou um KPI. Outro exemplo é a seção 404 da *Sarbanes-Oxley Act*. O *Sarbanes-Oxley Act* aplica-se a empresas que tenham ações registradas na SEC (*Securities and Exchange Comission*) dos Estados Unidos. Este ato determina que uma série de controles e procedimentos financeiros, incluindo relatórios, sejam produzidos e mantidos.

- Operação: atende requisitos operacionais como, por exemplo, controle de linha de produção, CEP (Controle Estatístico de Processos), monitoramento de sistemas de informação, controle de tráfego, entre outros.
- Pesquisa: qualquer projeto de pesquisa em áreas como medicina, biologia ou até mesmo busca de vida extraterrestre. Um exemplo é o projeto GIMPS (*Great Internet Mersenne Prime Search*), que busca números primos de *Mersenne*. É o maior projeto de computação distribuída do mundo. Neste projeto, qualquer pessoa ou organização pode baixar um software específico que usa o tempo ocioso de processamento do computador, para buscar números primos. Existem muitos projetos de computação distribuída e colaborativa no mundo, buscando cura para doenças, modelos climáticos, entre outros.

QUANTO AO TIPO DE ANÁLISE

São vários os tipos de análise que projetos de B&A podem realizar. Alguns projetos podem, inclusive, utilizar vários tipos. Por exemplo, é normal em projetos de análise preditiva iniciar com uma análise exploratória. Vamos ver os principais tipos:

- Descritivos ou Exploratórios: buscam analisar os dados para conhecê-los ou apresentar resultados consolidados, como vendas por tipo de produto ou funcionários por departamento. É fortemente baseada em visualização, produzindo gráficos como dispersão, diagrama de caixa, setores, barras, entre outros.
- Inferência: tem como objetivo apresentar prováveis fatos, a partir de uma parte menor da população dos dados. Este tipo de análise é apresentado junto com uma margem de erro.
- Analíticos: busca relações entre dados. Por exemplo, relação de vendas com o sexo do cliente, ou absentismo de acordo com o cargo do funcionário.

- Preditivos: buscam prever fatos futuros. Têm como base fatos ocorridos no passado. Também funcionam com uma margem de erro na previsão. As aplicações em análises preditivas são muitas: prever que será bom pagador, qual aluno terá melhor desempenho, quanto tempo o funcionário ficará na empresa, etc.
- Prescritivos: se o preditivo permite prever o fato, a prescritiva traz informações que permite mudar um evento futuro. Por exemplo, na seção anterior, o objetivo era prever quanto tempo o funcionário ficará na empresa, mas uma análise prescritiva pode dizer o que deve ser feito para reter o funcionário na empresa.

QUANTO À FONTE DE DADOS

A classificação quanto às fontes de dados também pode ser muito diversa. Primeiro, podemos olhar a fonte quanto ao modelo: relacional, rede, hierárquico, orientado a objetos, chave-valor, dimensional etc. Quanto ao fornecedor: *Oracle*, IBM, *Microsoft*, *Apache* etc. Quanto à origem: sistemas de produção, mídias sociais, pesquisas, benchmarking, sensores etc. Por fim, os dados podem ser estruturados, não estruturados e semiestruturados. Vamos entender melhor esta classificação:

- Estruturados: são dados organizados em linhas em colunas, como em planilhas e banco de dados relacionais. Sua estrutura é fixa e rígida: uma tabela de fornecedores, de produtos ou mesmo de clientes são exemplos.
- Semiestruturados: neste modelo, existe uma definição de estrutura, mas ela não é homogênia e não é fixa. Exemplos são arquivos XML e Json.
- Não estruturados: são dados sem qualquer tipo de estrutura definida: documentos, e-mails, sites etc.

O GERENTE DE PROJETOS ANALÍTICOS

Já estudamos que projetos de B&A são, acima de tudo, projetos como quaisquer outros no sentido de que precisam de planejamento e controle. Porém, normalmente, são mais complexos e com maior risco. Mas o que diferencia o gerente de projetos tradicionais de projetos de B&A? Coloco aqui alguns pontos que considero importantes:

- Experiência em projeto de B&A: estamos falando de projetos complexos e difíceis. Por mais que o gerente de projetos tenha anos de experiência em outras indústrias, é fundamental que grandes projetos sejam gerenciados por pessoas com experiência neste tipo de projeto. Você pode estar se perguntando, mas como adquirir experiência? Começando com projetos pequenos e de baixa complexidade ou atuando com algum gerente de projeto já experiente.
- Conhecimento Técnico: Um bom gerente de projetos analíticos deve ser curioso, gostar de tecnologia e, principalmente, de B&A. Deve ler a respeito, frequentar fóruns e eventos e estudar o assunto. Não que ele precise saber implementar um modelo de análise preditiva, mas deve entender o que é e como funciona. É fundamental saber, por exemplo, o que é um *data mart*, uma dimensão, um nó de um cluster ou um código de redução. Tais requisitos são ainda mais latentes se a empresa para a qual o gerente trabalha presta serviços implementando projetos em terceiros. Seu cliente vai querer tratar do projeto com alguém que fale a mesma linguagem e, muito além disso, junto com sua equipe, o gerente de projetos deverá atuar como consultor ao seu cliente, oferecendo as soluções que tragam os melhores resultados em termo de escopo, custo e cronograma.
- Gestão de Equipe: um projeto de B&A envolve equipes multifuncionais, com habilidades e experiências diversificadas. O gerente de projeto deve ser capaz de buscar e re-

conhecer competências, garantir que os recursos recebam os treinamentos necessários, manter a equipe motivada e garantir que cada membro da equipe doe o melhor de sua capacidade em busca dos objetivos do projeto.

2. ANTES DO PROJETO

Entendidos os conceitos e premissas fundamentais em B&A, neste capítulo, vamos analisar alguns fatores importantes que o gerente de projetos e sua equipe devem observar antes ou durante a execução do projeto. Estes elementos estão intimamente relacionados com o capítulo seguinte: riscos, pois, muitos dos fatores aqui apresentados, se forem ignorados, certamente aumentarão os riscos do projeto.

VALOR DE NEGÓCIO

Não é função do gerente de projetos criar o *Business Case* que vai justificar o projeto: normalmente, ele vai receber este documento pronto, já amplamente discutido, analisado e aprovado. Caberia ao gerente de projetos apenas acatar o plano e implementá-lo. Mas as coisas podem não ir muito bem, se simplesmente "fecharmos os olhos" e seguirmos em frente.

Um projeto qualquer deve entregar valor. Um projeto de B&A, como qualquer outro, deve entregar valor. Não é porque o projeto tem um *Business Case* bem elaborado, discutido e aprovado, que a geração de valor estará garantida (embora, teoricamente, deveria estar). Você, como gestor do projeto, quer ter em seu currículo um projeto que seja um sucesso em cumprir cronogramas, prazos e custos e que atendeu a definição de pronto, mas que não entregou valor? Que depois, ou logo depois que entrou em produção é abandonado? Ou que nem chegou a entrar em produção, pois a entrega de valor simplesmente deixou de existir ou ser rele-

vante? Claro que todo gerente de projetos tem que lidar, uma hora ou outra, com projetos que fracassam pelos mais diversos motivos. Mas não é algo que acontece pelas circunstâncias, não por omissão.

Destarte, é comum projetos de B&A serem produzidos com pouca ou nenhuma utilidade prática. Um gerente funcional que elabora um projeto com questões meramente operacionais, quando seu setor está indo para o buraco e ele não tem qualquer dado para apoiar a decisão, um auditor que quer gerar um arquivo para uma norma de conformidade que vai mudar e ser adiada dezenas de vezes antes de ser obrigatória, um executivo que quer um relatório para atender a uma questão pessoal, o CIO que ouviu falar de uma nova tecnologia e quer ter aquilo em seu currículo, entre muitos outros casos.

Um projeto de B&A não é uma brincadeira. Envolve muitos recursos financeiros e humanos, tempo, e põe em risco a reputação de pessoas. Normalmente, um projeto é aprovado em detrimento de outro, que poderia ser mais importante e de fato entregar valor. Por isso, antes de mais nada, o gerente de projetos deve analisar e garantir que o projeto entregue, de fato, valor para a organização.

CRITÉRIOS DE SUCESSO

Garantido que o projeto entregará valor, como gerentes de projetos é nosso dever também ter claro quais são os critérios que definem o sucesso de projeto. Sem critérios claros de sucesso definidos, o projeto poderá entrar numa eterna discussão de que se foi bem-sucedido ou não, pois esta passará a ser uma questão subjetiva e pessoal.

Já sabemos que não basta o projeto ser entregue em tempo, prazo e custo, aliás, é muito comum um ou mais destes três elementos falharem, e mesmo assim, o projeto ser considerado um sucesso.

Os critérios de sucesso devem ser definidos como patrocinador e usuários-chave, todos devem concordar com os mesmos

e estes devem ser documentados. Os critérios de sucesso em projetos de B&A podem ser os mais diversos, vejamos alguns dos mais comuns:

- Usuários satisfeitos
- Aumento de vendas
- Redução de fraudes
- Redução de custo operacional
- Disponibilidade
- TCO

Normalmente, mais de um critério de sucesso é definido. Estes critérios devem ser mensuráveis para serem avaliados. Por exemplo, 70% dos usuários satisfeitos, redução de fraudes em 1%, redução de custo operacional de 0,5%.

Os critérios de sucesso não devem ser confundidos com a definição de pronto. A definição de pronto é quando o projeto termina de ser implementado, devendo ser aceito e entregue, dentro dos critérios estabelecidos de escopo, performance e qualidade.

FATORES DE SUCESSO

Na seção anterior, falamos de critérios de sucesso, eles são usados para medir se o projeto foi bem-sucedido. Nesta seção, vamos falar de fatores de sucesso, que são pontos específicos que devem ser analisados para aumentar as chances de sucesso do projeto. Então, para ficar claro, critérios são métricas aplicadas quando o projeto está pronto, fatores são pontos que o gerente de projeto e sua equipe devem observar para aumentar as chances de o projeto ser bem-sucedido. Vamos ver alguns fatores de sucesso de projetos de B&A:

- Suporte do patrocinador: seu patrocinador, realmente, acredita no projeto, ou alguém o designou com a missão de fazer o projeto acontecer, mas ele não está muito interessado, tem outras prioridades? Não ter total apoio do

patrocinador é um péssimo sinal de dias ruins, como ter algum recurso-chave do seu projeto cortado sem aviso prévio. É fundamental que o patrocinador conheça, acredite e priorize o projeto dentro da organização.

- Envolvimento dos interessados: este item está, de certa forma, relacionado à falta de suporte do patrocinador. Em grandes empresas todos estão sempre muito ocupados. Se o projeto não é prioridade, as partes interessadas não vão se envolver. Se eles não vêm valor no projeto, também não vão se envolver. No capítulo de recursos humanos, falamos sobre a importância da evangelização das partes interessadas sobre aspectos de B&A, e especificamente, do projeto em andamento.
- Metas de projeto: falamos, em seção anterior, que um projeto de sucesso não é apenas aquele que entrega o projeto com mínima variação de prazo, com o escopo acordado e com os custos previstos. Mas claro, isso também é um fator de sucesso! Falamos que entregar no prazo, custo e escopo não adianta se não entregar valor. Mas pense a situação contrária, o projeto entregou valor, mas levou 18 meses a mais do que o planejado! Uma das funções mais importantes do gerente de projeto é manter o controle, para que, quando necessário, as medidas corretivas sejam tomadas para evitar um dano maior ao projeto. Se tivesse que resumir gerência de projetos em uma única palavra, a palavra seria CONTROLE.
- Orientado a negócio e não à tecnologia: entretanto, muitos projetos nascem orientados à tecnologia. Se as tecnologias estão de acordo com o que se quer construir, tudo bem, porém, se forem sub ou superestimadas, poderão levar o projeto ao fracasso. E acredite, isso acontece todo o dia! Um projeto de B&A deve antes de tudo avaliar o que se quer produzir, qual o seu objetivo. Tendo isso claro, então, avaliam-se as tecnologias existentes.

VOLUME E CRESCIMENTO

Vamos falar, posteriormente, em análise de protótipos e pilotos, quando o projeto será analisado mais profundamente do ponto de vista de arquitetura. Mas, inicialmente, alguns aspectos devem ser analisados com a equipe do projeto, como arquitetos, programadores e o pessoal de infraestrutura: qual vai ser o volume de dados? Muitas questões estão relacionadas com o volume: quantas serão as transações por minuto? Qual o volume de dados diário e mensal em *terabytes* ou *pentabytes*? Quantos *datanodes* e *namenodes* serão necessários? É preciso analisar o crescimento do sistema para os próximos anos (pelo menos, para os próximos 3 anos) e avaliar a infraestrutura necessária quanto a armazenamento, rede e processamento.

VELOCIDADE

Analisando, agora, outro "V" de *Big Data*, temos que analisar o aspecto da velocidade dos dados, não só na sua origem, mas de qual a latência possível até a sua análise concluída. Serão dados em fluxo (*streaming*) analisados em memória ou informação estática? Serão dados em tempo real, próximo ao tempo real, ou com baixa latência (dias, semanas, meses)? Dados em tempo real podem ter um requisito de milissegundos, segundos, minutos ou mesmo horas. Já falamos sobre o valor da informação: talvez, um indicativo de fraude deve ser reportado em segundos. Uma auditoria de folha de pagamento pode ser processada em 30 dias! Em certos casos, você terá que considerar, por exemplo, infraestrutura de rede de 10 gigabit *Ethernet* para suportar o *streaming*. *Spark*, irmão do *Hadoop*, é uma excelente plataforma de processamento em memória para projetos de tempo real ou próximo ao tempo real. Hoje, um servidor com 6 TB de memória pode suportar grandes volumes de dados e custar muito menos do que vários computadores commodity que comporiam alguns nós de um cluster.

VARIEDADE

Outro grande impacto no projeto é com relação à variedade. Serão dados estruturados ou não estruturados? Quais serão os fornecedores? Que tipo de dados não estruturado serão? Imagens, vídeos, e-mails? Estão em plataforma baixa ou alta? Existem conectores conhecidos e funcionais para todas as fontes? Existem dados criptografados ou em formatos proprietários? Existem dados gerados por máquinas? Serão necessários dados de dispositivos? Redes sociais? Será preciso comprar algum tipo de dados de um *data broker*[2]? As fontes de dados estão documentadas? Para piorar, a integridade referencial não existe no banco de dados e é mantida na aplicação. Se não há documentação, estão em uma estrutura legível? Será preciso fazer um processo de "engenharia reversa" sobre as fontes de dados? Imagine se entre uma das fontes de dados existe, por exemplo, um sistema em que as entidades foram geradas por um gerador de aplicação, onde as tabelas e colunas são nomeadas por um prefixo mais um número, e não há qualquer documentação? Você já imaginou o impacto em custo e tempo no projeto, se for necessário descobrir toda a estrutura de dados necessária para os objetivos do projeto usando uma ferramenta de *profiler*? Ou ainda, se a fonte de dados é algo muito antiga e específica, que não existe um conector conhecido, ou este conector só está disponível numa ferramenta, cujo custo está nos seis dígitos! Em seção posterior, falaremos ainda sobre sistemas sombra, que estão, de certa forma, relacionados com a variedade.

DISPONIBILIDADE

A indisponibilidade faz parte dos riscos de um sistema em produção, ou seja, que não é mais um projeto. Porém, devem ser planejadas na sua construção. A disponibilidade é afetada por falhas em hardware, software, falta de luz, falhas de segurança, problemas físicos ou lógicos de redes, entre outros. Deve-se ainda

2 *Data Broker* é uma empresa especializada em vender dados.

prever e distinguir entre a indisponibilidade prevista e não prevista. Alguns tipos de sistemas podem ser projetados para, eventualmente, ficarem de forma programada, fora do ar.

Os requisitos de disponibilidade de projetos de B&A variam muito. Um sistema hospitalar pode requer alta disponibilidade, já um sistema de geração de um arquivo de conformidade não. Uma disponibilidade alta, por exemplo, de 90%, significando que o sistema poderá estar indisponível cerca de 36 dias por ano é intolerável até para a aplicação menos crítica. Já uma disponibilidade de 99,999% (conhecida como cinco noves), representa uma indisponibilidade de 5,26 minutos por ano. Os requisitos de disponibilidade vão afetar drasticamente o seu projeto em termos de tempo, infraestrutura, pessoal, fornecedores e mais ainda, em custos. A inclusão de um ou dois noves depois da vírgula na disponibilidade, pode causar um impacto significativo nos custos de seu projeto.

Se partes da solução serão contratadas de terceiros, como por exemplo, um BI como serviço, além do SLA, pode ser interessante ter em contratação para projetos críticos o RTO (*Recovery Time Objective*), em português objetivo de tempo de recuperação, que define, em caso de falha ou desastre, o tempo que o contratado tem para recuperar o sistema e o colocar de volta em funcionamento.

A alta disponibilidade significa sistemas mais tolerantes a falhas, que pode ser implementado por um "simples" *Raid* 6 ou até envolver redundância de servidores, armazenamentos, redes físicas e lógicas, replicação de dados para servidores remotos, entre outros. Do ponto de vista de bancos de dados, os grandes fornecedores de banco de dados (não só relacionais) oferecem soluções de redundância. Certos produtos têm a redundância como característica nativa, como o próprio *Hadoop*, onde *datanodes* e *namenodes* são naturalmente replicados.

VIRTUALIZAÇÃO

A virtualização não é nenhuma novidade. Projetos de B&A, normalmente, têm pelo menos uma de suas camadas virtualizada,

o que reduz custos e facilita o gerenciamento. Certas partes de um projeto, obviamente, não fazem sentido serem virtualizadas em um mesmo servidor físico. Por exemplo, *datanodes* redundantes, se estiverem virtualizados no mesmo hardware, em caso de pane de qualquer elemento deste equipamento, tornarão os dados destes nós indisponíveis.

BACKUP

A solução de backup, que será aplicado ao mesmo, depende diretamente dos requisitos de disponibilidade e segurança do projeto. Podem ser necessários, além de backups completos, backups diferenciais, de configurações, de sistemas de autenticação, entre outros. Também pode ocorrer que o backup tenha que ser feito em uma estrutura física, diferente, ou seja, feito ou levado para outras instalações.

SEGURANÇA E PRIVACIDADE

Os requisitos de segurança e privacidade podem afetar muito o projeto, não apenas no que se refere a ferramentas, mas no seu próprio desenvolvimento, por isso, devem ser considerados já no início do projeto. Dados podem ter que trafegar ou terem seu armazenamento criptografado. Requisitos funcionais podem exigir controle de acesso por linha ou por coluna. Pode ser preciso que ferramentas de *front end* suportem autenticação integrada com LDAP.

AUDITORIA

A solução a ser projetada e desenvolvida tem requisitos de auditoria de uso e acesso a dados? Certos projetos podem ter requisitos de auditoria de uso, que normalmente vêm associados a questões de segurança e privacidade bastante restritos. A auditoria pode ser necessária em mais de uma etapa: na extração, no *staging* e

na visualização. Normalmente, quando a equipe e ETL têm acesso a dados em uma fonte de origem, este acesso é restrito a certas tabelas e colunas. Uma auditoria vai se certificar de que dados aos quais não se deveria ter acesso foram extraídos. No *staging*, a auditoria registra quais dados foram atualizados. Na visualização, normalmente, a mais simples, o sistema deve registrar quais fatos ou mesmo painéis foram acessados por quem, quando e de onde.

GOVERNANÇA DE DADOS

Devem-se avaliar cuidadosamente as políticas organizacionais de governança de dados. Trata-se de uma série políticas, processos, pessoas e tecnologia de como os dados devem ser governados na organização. O projeto terá que se adequar a estas políticas, não somente quanto ao manuseio dos dados, mas quanto ao uso de sistemas e infraestrutura que serão utilizados.

CULTURA E IDIOMA

Se a empresa para qual o projeto é executado é uma multinacional, podem haver requisitos de cultura e de idioma. Todos sabemos que requisitos de idioma significam: a interface traduzida para o idioma local. Requisitos de cultura requerem que elementos como datas e moedas, sejam formatados de acordo com a cultura do país onde a ferramenta é usada.

Se o problema é apenas com a interface da ferramenta para o usuário final, uma consulta à documentação e uma provável prova de conceito podem sanar qualquer dúvida a respeito. Mas há um porém: normalmente, um objeto de visualizado no *front end* vai usar um *label* para um fato, já nomes de colunas serão exibidos como estão no banco de dados. Pode ser necessário que a ferramenta suporte ainda a tradução de elementos como nomes de dimensões e atributos.

Às vezes, os requisitos de cultura vão exigir programação ou implementação de fórmulas em medidas. Por exemplo, con-

versão de valores através de consultas em tempo real à conversão de moedas.

NOVAS TECNOLOGIAS

Se seu projeto vai usar um produto de uma tecnologia que está no mercado há décadas, você poderá ter algumas preocupações quanto a requisitos técnicos e funcionais, como suporte de volume de dados, *failover*, backup, entre outros. Porém, se está nos planos do projeto usar alguma tecnologia mais atual, digamos, com cinco anos ou menos de existência, é preciso tomar mais alguns cuidados. Vamos falar, durante a obra, da importância de pilotos e protótipos, mas aqui fica um relato de como uma nova tecnologia foi descartada através da implementação de um protótipo. Embora o produto fosse muito bem-conceituado e aparecesse como a nova sensação do mercado para projetos de B&A, na prática, apresentou vários problemas sérios: primeiro, o produto criava uma centena de processos-zumbis, pelo qual parecia que ele perdia o gerenciamento, chegando ao ponto de acabar com a memória do servidor. Não foi encontrado nenhum documento descrevendo como resolver o problema ou mesmo oferecendo um workaround. Na comunidade de usuários também não foi possível encontrar uma única pista. O problema acabou sendo resolvido, mas proporcionou um grande atraso e gerou a desconfiança de todos envolvidos no projeto, que acabou sendo implementado em outra tecnologia. Por isso, quando se tratar de produtos ou tecnologias novas, alguns cuidados podem evitar sérios problemas no projeto.

Inicialmente, avalie o TCO, Total Cost of Ownership, ou custo total de aquisição. Vamos pensar no Hadoop. Você pode avaliar implementar um projeto usando os binários gratuitos da Fundação Apache, ou optar por uma distribuição de terceiros, como Cloudera, Hortonworks, MapR ou até Microsoft. Os custos destas licenças de terceiros podem ser significativos a partir de certo número de nós, embora para projetos pequenos existam

versões gratuitas. Outros oferecem a gratuidade apenas por um tempo. Porém, para um projeto não devemos apenas avaliar o custo de licenças, mas todos os custos diretos e indiretos envolvidos, que incluem treinamento, infraestrutura, hardware, testes, crescimento, resolução de incidentes e falhas, backups, recuperação, entre muitos outros. Deixo claro que não estou advogando para o fato de que se tenha que usar uma versão não Apache original do *Hadoop*, estou dizendo apenas que todos os custos devem ser considerados.

Outro ponto importante é avaliar a documentação e a base de conhecimento produzida sobre o produto. Existem produtos muito bem documentados, outros com documentação extremamente pobre, em que mal há linhas escritas sobre configurações básicas.

Também as questões de treinamento. Normalmente, será preciso treinar vários grupos de usuários, com vários aspectos diferentes da tecnologia. Existe treinamento disponível? O treinamento é de qualidade e atende aos requisitos do produto? Já gerenciei projetos em que o caminho mais fácil foi a própria empresa desenvolver os treinamentos.

Outro fator que também talvez seja óbvio, mas que não podemos deixar de mencionar, é o suporte. Que tipo de suporte será prestado? Quais os canais de suporte? Qual a SLA?

Se você vai incorporar um produto *open source* e pretende alterar seu código, é bom verificar quais os padrões de codificação foram usados, se é que foi usado algum. O código é indentado, organizado, legível, documentado? Foram aplicadas as melhores práticas, não apenas de escrita de código, mas de arquitetura? Deve-se também avaliar se a empresa quer arcar com o ônus de manter o código fonte de um sistema, às vezes, de proporções gigantescas.

Por fim, será feita alguma transferência de tecnologia? De que forma? Quais são as garantias? Quem ficará responsável pelo quê neste processo? No capítulo 9, falaremos mais sobre os processos de aquisições.

SHADOW SYSTEMS E SILOS DE INFORMAÇÃO

Shadow Systems, ou sistemas sombra, são aqueles softwares ou planilhas que são usados na empresa de forma não oficial, normalmente, em paralelo com sistemas de operações maiores, como ERPs, CRMs, e outros. Os motivos da adoção destes tipos de sistemas são muitos, mas o mais comum, é que o sistema oficial não atende a determinada funcionalidade na forma que o usuário gostaria, o que é natural, uma vez que sistemas são genéricos, não adaptados para o processo de negócio específico da empresa onde foram implantados. Sistemas sombra são o pesadelo de muitos CIOs, e são encontrados em algum canto mesmo em corporações mais rígidas em sua governança de TI. Já silos de dados são bases de dados isoladas e não integradas com o restante da organização. Um silo de dado pode existir, às vezes, até na forma de um *data mart*, bem estruturado e com alta entrega de valor. Sistemas sombra são, normalmente, origens de silos de informação.

Entendido o que são, como estes sistemas afetam projetos de B&A? Todavia, é muito comum que os requisitos funcionais de projetos envolvam em algum ponto um ou mais sistema sombra, e isso pode se tornar um problema. Este tipo de sistema é quase como algo pirata dentro da organização, às vezes, instalados em estações de trabalho. Já vi casos de sistemas de CRM inteiros em notebooks de vendedores que, obviamente, de vez em quando, desconectavam seus equipamentos da rede e saíam em campo, levando todo o sistema junto (e não era um sistema preparado para sincronizar os dados na volta!). Não há garantia nenhuma para o processo de extração de dados, que a informação estará lá quando necessária. Pior, e bastante comum, são quando estes sistemas são nada mais do que planilhas, cheias de fórmulas, desestruturadas e sobrecarregadas de dados. Nada garante que, quando necessário, a planilha não vai estar aberta, ou que seja movida de lugar, ou ainda que sua estrutura seja deliberadamente alterada. Em todas essas situações seu processo de carga de dados, obviamente vai falhar.

Não podemos, simplesmente, determinar que sistemas sombra não façam parte do projeto. Mas podemos minimizar os riscos: buscar alternativas para a informação é o primeiro passo. Podemos ainda determinar requisitos mínimos para que este tipo de sistema seja incluído no projeto. Se nenhuma dessas hipóteses funcionar, devemos definir políticas mínimas para a manutenção do sistema e minimizar o impacto no projeto. Bloquear estruturas de planilhas e criar diretivas para que o arquivo não seja movido ou excluído, são algumas. Também é possível incluir no orçamento do projeto o desenvolvimento de sistemas centralizados para suprir o problema de planilhas.

VERSÃO ÚNICA DA VERDADE

Não é incomum que fatos de negócio estejam em mais de um sistema. Por exemplo, o valor total de horas extras da folha de pagamento do mês pode estar diferente no sistema de folha e na contabilidade. Na reunião com o CEO, o gerente financeiro acha que a meta de redução de horas extras não foi atingida, pois o seu KPI capturou dados do sistema financeiro. Já o gerente de RH, com um relatório com dados oriundos do sistema de folha, garante que a meta foi atingida. O cliente pode ter um endereço no sistema financeiro e outro no CRM, que acabou de vender um novo produto e atualizou o endereço. A cobrança da nova compra vai, então, para o endereço errado. Desde que as funcionalidades do projeto comecem a ser definidas, é importante o gerente do projeto estabelecer e documentar quais serão as origens dos fatos e medidas que farão parte do projeto. Felizmente, empresas que adotam processos de governança de dados, como MDM (*Master Data Management*), têm este risco minimizado. É claro que, mesmo com um MDM implementado, muita informação não vai ter uma representação como registro mestre.

3. RISCOS

Colocar riscos como um dos primeiros capítulos não é por acaso. Projetos de B&A são, especialmente, sujeitos a riscos. Projetos de TI são, por natureza, projetos de alto risco. Especificamente, projetos de desenvolvimento de softwares estão entre os mais difíceis em todas as indústrias. Eu me arriscaria a dizer que projetos de B&A são os mais complexos e difíceis em Tecnologia da Informação. Já falamos em falhas de projetos de desenvolvimento de software, mas projetos de tecnologia falham em torno de 25%, enquanto, o índice de falhas de projetos de B&A é o já mencionado índice de 55% Na verdade, projetos de desenvolvimento de software e de B&A têm muito em comum: envolvem o desenvolvimento de algo abstrato e não tangível. Grandes projetos de B&A, como por exemplo, a construção de um *data warehouse*, têm ainda o fator tempo como inimigo. Diferente da construção de um prédio, em que o cliente consegue ver as paredes sendo construídas, o que o cliente vai ver entregue e que lhe agregue valor ao longo da construção de um *data warehouse*? Scritpts de extração? Tabelas de fato modeladas? Fórmulas de medidas? Portanto, um gerente de projetos de B&A deve ter especial atenção à gestão de riscos durante todas as etapas do projeto.

Neste capítulo, vamos analisar riscos com ênfase em projetos de B&A. Em seção posterior deste capítulo, vamos comentar sobre os principais riscos que podem estar associados a projetos deste tipo.

RISCOS EM PROJETOS

Qualquer evento está sujeito a algum tipo de incerteza: atravessar a rua, embora pareça uma ação simples, tem as suas incertezas, você pode torcer o pé, ser atropelado, cair num buraco, desmaiar etc. O maior risco neste caso, sem dúvida, é ser atropelado. Por isso, antes de atravessarmos a rua, tomamos algumas medidas visando minimizar o risco de atropelamento: paramos, olhamos para os lados e quando acharmos que estamos seguros, atravessamos a rua. Quando vamos dirigir o carro, colocamos o cinto de segurança. Um voo possui uma série de medidas buscando minimizar riscos: treinamento da tripulação, divulgação de instruções de segurança antes da decolagem, entre outros. Alguns passageiros podem tomar medidas pessoais de minimização de riscos antes de uma viagem: fazer orações ou deixar recomendações com parentes, caso algo de ruim aconteça. Se na nossa vida tomamos medidas para minimizar riscos, porque não deveríamos tomá-las também em projeto, em que há grandes expectativas criadas, dinheiro investido e pessoas empenhadas em criar um produto ou serviço? Assim como as medidas de segurança não impedem que aviões caiam, gerenciamento de riscos não vai garantir que o projeto não fracasse, mas com certeza, vão aumentar as chances de sucesso e no mínimo, minimizar o impacto de eventos negativos. O gerenciamento de riscos é uma série de processos que tem como objetivos principais identificar riscos, analisá-los e responder a eles, eliminando ou minimizando o seu impacto no projeto. O gerenciamento de riscos deve ocorrer através de um plano, que deve conter, entre outras coisas, a equipe responsável pela gestão de riscos, a metodologia que será empregada, o orçamento destinado à gestão de riscos, os processos necessários com sua frequência de execução. Existe um consenso entre especialistas que mesmo com uma boa gerência de riscos, ainda cerca de 10% deles não serão identificados, para os quais o projeto deve ter alocado recursos de contingência.

A gestão de riscos, como um processo contínuo, deve ocorrer durante todo o projeto. Não adianta identificar riscos apenas

no início: novos riscos podem surgir. As fontes de novos riscos podem ser muitas: crises econômicas, mudanças tecnológicas, mudanças em parceiros ou fornecedores. No decorrer do projeto, riscos já identificados podem passar a ter a probabilidade de ocorrência zero, ou seja, deixarem de existir. Da mesma forma, riscos já existentes podem mudar a sua probabilidade ou o seu impacto, o que, consequentemente, deverá alterar a estratégia que existe para aquele risco. A identificação de riscos ao longo do projeto deve ser um trabalho da equipe do projeto, podendo envolver até mesmo clientes e terceiros. Muitas técnicas podem ser aplicadas, como entrevistas, análise de documentação de projetos anteriores, estudo da WBS, dinâmicas de grupo como *Brainstorming* e *Delphi*. O risco identificado não deve ser simplesmente considerado para acompanhamento, é preciso analisá-lo quanto à viabilidade da sua exatidão.

Riscos possuem fatores de risco, que são eventos que influenciam de alguma forma o risco. Por exemplo, uma crise econômica pode influenciar positivamente o risco de falta de mão de obra qualificada. O resultado do risco é o que acontece se o risco se efetiva, sem antes ter sido tomada qualquer medida visando minimizá-lo. Atrasar o projeto por falta de mão de obra qualificada é o resultado de um risco. A isso cabe uma reação ao risco: treinar um recurso em determinada tecnologia para minimizar o impacto da falta da mão de obra.

Já falamos que o risco é cercado de incerteza, pois pode ou não ocorrer. A efetivação do risco resulta na ocorrência de um problema. Por exemplo, existe o risco de falta de mão de obra qualificada. O projeto iniciou e o risco se materializou: não se conseguiu contratar mão de obra. O risco tornou-se, então, um problema.

Os riscos são divididos em cinco grupos: externos, internos, organizacionais, de negócios e puros. Externos, como devemos imaginar, são externos à organização e ao projeto. Um exemplo de risco externo é, por exemplo, uma mudança em alguma legislação que afete o projeto. Interno são os inerentes à organização. Um

exemplo clássico, e que será abordado ainda neste capítulo, são os riscos relacionados a recursos humanos. Um tipo de risco organizacional são os riscos técnicos, muito presentes em projetos de B&A. Um exemplo seria uma incompatibilidade do extrator com uma fonte de dados. Riscos de negócios são aqueles inerentes a qualquer projeto, como por exemplo, uma crise econômica. Por fim, os riscos puros, também chamados de seguráveis, são riscos como incêndios, terremotos, ferimentos etc.

Todo o risco possui duas dimensões: probabilidade e impacto. Um risco só é risco se a probabilidade de ocorrer é maior que zero e menor que cem por cento. Se a probabilidade é zero, deixa de ser um risco pois não vai ocorrer nunca. Se a probabilidade é de cem por centro, deixa de ser um risco e passa a ser um problema. Da mesma forma, o impacto em caso de ocorrência deve ser significativo, pois mesmo que a probabilidade seja de 99 por cento e o impacto de ocorrência seja insignificante, deixa de ser um risco relevante.

ANÁLISE QUALITATIVA DE RISCOS

A análise qualitativa de riscos é uma análise subjetiva e mais simples se comparada à análise quantitativa, que será estudada na próxima seção. Analisar um risco quantitativamente é identificar os riscos, avaliar sua probabilidade e seu impacto e criar um ranking para acompanhamento. Riscos no topo da lista devem ser acompanhados com mais cuidado. Já os riscos na parte inferior são os que oferecem menos perigo ao projeto, e não requerem tanta atenção. Em uma análise qualitativa, a classificação quanto à probabilidade e impacto pode usar escalas ordinais, como baixo, moderado, alto etc.

ANÁLISE QUANTITATIVA DE RISCOS

A análise quantitativa de riscos avalia numericamente a probabilidade e o impacto de riscos. Uma avaliação quantitativa, em

geral, é melhor que uma qualitativa por ser menos subjetiva e, consequentemente, mais próxima dos fatos e consequências reais do risco. O valor monetário esperado, ou VME, é uma forma de analisar riscos quantitativamente. O VME é, simplesmente, o produto da probabilidade versus o impacto financeiro do risco, podendo ser positivo ou negativo. Por exemplo, se a probabilidade de não encontrar mão de obra não qualificada é de 15% e o impacto no projeto seria de R$ 100 mil, então, o VME deste risco é de R$ 15 mil. Falamos em seção anterior que o ideal é que projetos tenham previstos os custos para contingências de riscos. Mas como chegar a um valor? Uma forma é somar o VME de todos os riscos do projeto.

ESTRATÉGIAS PARA RISCOS

Existem diversas estratégias que devem ser aplicadas na ocorrência de um risco. A melhor estratégica depende, obviamente, do tipo de risco e deve estar já desenhada no plano de gestão de riscos. Vamos ver as principais estratégias a seguir:

- Prevenir: a prevenção consiste na tomada de uma ação para que o risco seja eliminado. Por exemplo, se uma fonte de dados oferece risco ao projeto devido ao fato de não haverem extratores conhecidos, pode-se retirar os dados desta fonte do escopo do projeto, desta forma, o risco é eliminado.
- Mitigar: a mitigação consiste em reduzir o impacto e/ou probabilidade do risco no caso eventual de sua ocorrência, o que não vai eliminá-lo. A execução de um piloto, que falaremos em seção posterior, é uma técnica de mitigação.
- Transferir: a transferência consiste em passar o risco para terceiros. Aqui, também o risco não é eliminado, já que o terceiro pode falhar em criar o produto ou o serviço para o qual foi terceirizado.
- Aceitar: neste caso, não se faz nada, se o risco ocorrer, simplesmente, aceitam-se as suas consequências.

Depois da aplicação de uma estratégica para lidar com um risco, podem ocorrer duas coisas: permanecer ainda algum risco ou surgirem novos. A continuidade do risco é chamada de risco residual. Isto ocorre porque nem sempre conseguimos eliminar o risco totalmente. Por exemplo, podemos mitigar o risco de não ter recursos qualificados oferecendo treinamento à equipe. Mas o treinamento pode não ser eficiente e o risco persiste. Os riscos que surgem em consequência de aplicarmos uma estratégia de gestão de riscos são conhecidos como riscos secundários. Imagine a seguinte situação: seu projeto tem o risco de atrasar por falta de mão de obra qualificada, sua empresa faz um grande investimento e treina toda a sua equipe. Neste momento, com a equipe qualificada, existe o risco de eles abandonarem o projeto para ir trabalhar na empresa concorrente, que faz uma oferta melhor.

Quando recebemos um projeto devemos planejá-lo e sair fazendo? Não, projetos de B&A podem ser inviáveis, do ponto de vista técnico ou de negócio. Podemos fazer uma análise de viabilidade, examinando os pontos de riscos, mitigando-os ou mesmo eliminando-os. Às vezes, é preciso uma abordagem "*hands on*" para tratar destes riscos, como vamos ver na seção a seguir.

ANÁLISE DE VIABILIDADE: PROVA DE CONCEITO, PROTÓTIPO E PILOTO

Na próxima seção, vamos estudar os principais riscos que afetam projetos de B&A, porém, antes, vamos estudar algumas técnicas bastante eficientes na mitigação de riscos nesses tipos de projetos.

A gestão de riscos refere-se a processos que ocorrem durante todo o projeto, por uma razão óbvia: os riscos estarão sempre presentes, mesmo que a sua probabilidade e impacto variem ao longo do projeto. Porém, algumas técnicas podem ser executadas no início do projeto. Na maioria dos casos, é importante que estas etapas sejam elaboradas mesmo antes de entregarmos aos

envolvidos no projeto uma dimensão de seu tempo e custo. Isso porque, como veremos, estas etapas são usadas na mitigação de riscos que podem, inclusive, dependendo do seu resultado, afetar diretamente o projeto como um todo.

Escrito por Carlos Alberto Júlio, o livro A arte da estratégia: pense grande, comece pequeno e cresça rápido, tem no título a aplicação perfeita para a estratégia que deve ser adotada em projetos de B&A. Começar pequeno é uma excelente estratégica de sucesso. E uma excelente estratégia de começar pequeno, é com provas de conceito, protótipos e pilotos.

Prova de conceito é o desenvolvimento de um pequeno projeto, com uma pequena fração do que está destinado do projeto, e por isso, é desenvolvido em um curto espaço de tempo, com poucos recursos e com escopo bem restrito, cujo objetivo é validar um ou mais conceitos e, consequentemente, minimizar riscos.

Um protótipo, diferente da prova de conceito, vai tentar simular todo o produto, ou pelo menos, grande parte dele. Pode ser feito com o sistema não oficial, que ainda poderá ser desenvolvido. Um protótipo pode ser feito com vários níveis de intensidade, e nem sempre significa implementar código. Por exemplo, para um projeto, cuja entrega final sejam painéis de visualização de dados, um protótipo pode ser estes painéis desenhados em uma ferramenta como *Balsamiq* ou *Visio*, para que os usuários possam analisar se aquilo é de fato o que eles esperam receber ao final do projeto. Em outros casos, pode ser necessário desenvolver de fato os extratores e produzir os painéis, de forma não estruturada e usando uma base de dados de desenvolvimento. Depois de aprovados, os protótipos, em muitos casos, podem ser até descartados, para se dar início ao desenvolvimento do produto oficial.

Já em um Piloto, todo o sistema oficial a ser implantado é utilizado e testado. Além de ser um processo de minimização de riscos, certas organizações exigem que sistemas sejam amplamente testados e validados em um ambiente paralelo, em que não exista ligação lógica ou física com sua infraestrutura oficial, para

validar o produto em termos de segurança, desempenho e conformidade com as políticas da organização.

Nem sempre será necessária a implementação de provas de conceito, protótipos e pilotos, esta decisão é do gerente de projetos com sua equipe. É importante lembrar que tais processos requerem orçamento e tempo. Estes processos podem nos dar as informações vitais para decidir se o projeto deve continuar ou não, isso porque se pode concluir que o projeto vai custar mais caro do que o orçamento disponível para o mesmo (por exemplo, pode-se chegar à conclusão que, para o projeto prosseguir, será preciso adquirir uma ferramenta de extração com preço em seis dígitos) ou mesmo que o projeto simplesmente é inviável (por exemplo, certos dados de origem vitais para o projeto estão criptografados).

PRINCIPAIS RISCOS EM PROJETOS DE *BIG DATA*

Agora, que já sabemos o que são riscos e como eles devem ser gerenciados ao longo de um projeto de B&A, vamos estudar os principais riscos que afetam projetos de B&A.

MEDIDAS

As medidas são um elemento problemático em muitos projetos. Medidas mais comuns vão, simplesmente, somar a ocorrência de um fato em determinado período, de acordo com o nível de granulidade. Porém, muitas vezes, a equação resultante da medida é bastante complexa. Outro agravante da medida é que, muitas vezes, não há um valor de comparação no sistema transacional de origem. Temos aqui um paradoxo: quando há um valor de comparação no sistema de origem, não se conhece a fórmula e não se consegue chegar ao mesmo valor. Quando não há, não temos parâmetro para checar se o cálculo está correto. Outro grande problema com medidas é que, como falamos na seção anterior, sua fórmula de cálculo não é conhecida por pessoas da empresa.

Ocorre também de não haver um consenso entre colaboradores de um mesmo departamento sobre como calcular o valor. Chega a ocorrer o absurdo de, na mudança de um gestor de um departamento, este concluir que as medidas estão erradas e solicitar que todo o processo de construção dos artefatos seja refeito para chegar aos cálculos de seu entendimento. Sobre isso, falaremos ainda no capítulo de escopo, da missão do gerente de projetos de avaliar se o que está sendo modelado e construído deve atender às necessidades imediatas de um gestor, ou deve ser algo voltado para o departamento, algo impessoal que sobreviverá a uma eventual mudança nos cargos de direção da empresa.

MODELOS PREDITIVOS

Projetos de análise preditiva dependem da criação de um modelo. Para quem não é familiarizado com este tipo de projeto, vai aqui uma breve explanação. Análises preditivas usam dados do passado para prever o futuro. Para fazer a previsão, estes dados históricos são submetidos a um algoritmo que os processa e produz um modelo. O modelo nada mais é do que uma referência para prever dados futuros: uma vez construído o modelo, o sistema preditivo não precisa mais dos dados históricos nem do algoritmo, tudo o que ele vai precisar é das informações a serem previstas, que são submetidas ao modelo que tem como saída a previsão. Vamos imaginar um exemplo prático. Seu projeto é criar um modelo de previsão de fraudes em solicitação de crédito em uma instituição financeira. Seu sistema terá que carregar dados históricos de transações passadas onde houve ou não fraude, estes dados são submetidos ao modelo que aprende o padrão da ocorrência ou não de fraude. Então, novas solicitações de crédito são executadas, antes da efetivação, o sistema submete os dados da solicitação, como por exemplo, idade, sexo, renda, quantidade de filhos, se tem casa própria ou não etc. para o modelo, que informa se aquela transação é uma provável fraude ou não, baseado no que aprendeu com os dados históricos.

Porém, existe um detalhe: um modelo nunca é perfeito: ele terá uma taxa provável de acertos, ou seja, ele não vai acertar sempre. No exemplo acima, existem solicitações de crédito verificadas que o modelo vai classificar como fraudes e transações fraudulentas que ele irá classificar como boas. Essa taxa pode ser medida durante a construção do modelo, de forma que, já com o modelo em produção, tenha-se uma ideia precisa do quanto o projeto vai acertar.

As criações de modelos apresentam alguns problemas relacionados a riscos. O primeiro é que os envolvidos no projeto podem esperar taxas de acerto irrealistas. Por exemplo, 75% de acerto é uma taxa comum e considerada de sucesso em muitos projetos de B&A, mas para os envolvidos pode ser considerada inexequível e inadmissível devido às perdas que serão ocasionadas. É preciso, já no início do projeto, deixar claro com os envolvidos quais são as expectativas de sucesso do projeto com relação à sua precisão em acertar as previsões. Outro problema é que, simplesmente, pode não ser possível criar um modelo com uma taxa de acerto mínima. São dois motivos principais que a inviabilizam: primeiro, e mais comum, é que não existem atributos suficientes nos sistemas de origem para produzir o modelo, ou, se os atributos existem, eles não estão preenchidos. Voltando ao exemplo da aprovação de crédito: vimos que precisamos de dados históricos para criar o modelo. E se estes dados não existem (não foram coletados ou nem sequer eram previstos no sistema transacional)? É preciso um conjunto mínimo de atributos para que o modelo possa funcionar. O segundo impeditivo mais provável para a não criação de um modelo de sucesso é quanto à qualidade dos dados. Isso pode colocar em risco qualquer tipo de projeto de B&A, não só os preditivos, por isso, este assunto é tratado na próxima seção.

O maior problema com relação a modelos ineficientes é que a equipe do projeto pode vir a descobrir isto tarde demais: quando já se passaram meses de projetos, já se gastou boa parte do orçamento e as expectativas dos envolvidos estão ainda

maiores. Para minimizar estes riscos, um piloto ou uma análise de viabilidade técnica prévia pode, em pouco tempo e sem gastar muito dinheiro, nos mostrar que o projeto é inviável. Um piloto poderia consistir em extrair "manualmente" dados dos atributos transacionais e submetê-los a um algoritmo para avaliar a precisão. Algoritmos de seleção de atributos também poderiam ser utilizados para identificar quais atributos seriam os mais relevantes. Uma abordagem mais arriscada, porém, mais rápida e de menor custo seria uma análise de viabilidade técnica, onde, sem extração de dados, se analisaria a existência de atributos relevantes e sua frequência de preenchimento.

QUALIDADE DE DADOS

A qualidade dos dados é um problema crônico e global. Costumo sempre citar *English* [2009], que relata que as perdas pela falta de qualidade de dados em apenas 122 organizações superaram perdas em mais de US$ 1 trilhão. Sistemas de origem, em geral, no melhor caso, são pensados em manter as operações: o dado é coletado, processado, produz uma saída e sua vida útil neste sistema praticamente acabou. No pior caso, são usados mecanismos de entrada arcaicos que sequer dispõem de recursos para minimizar problemas de qualidade. O resultado disso são décadas de dados que por, muitas vezes, são inúteis ou economicamente inviáveis de serem aproveitados em qualquer tipo de processo analítico. Existem, sim, muitas e excelentes ferramentas de qualidade de dados que minimizam este problema. Mas não há milagre, alguns casos apenas uma intervenção manual registro a registro agregariam valor aos dados, o que pode tornar o projeto inviável economicamente e temporalmente.

O gerente de projeto deve garantir que os dados de origem sejam analisados quanto à sua qualidade, para que eventuais problemas possam ser considerados nos custos e cronogramas do projeto. A qualidade dos dados é mais um fator que pode, em muitos casos, inviabilizar o projeto.

EXPECTATIVAS DE TEMPO NÃO REALISTAS

Outro problema talvez mais grave que cronogramas não realistas são expectativas de tempo não realistas. Uma parte interessada, nova no assunto, pode imaginar que, em algumas semanas, terá todo o tipo de informação financeira à sua disposição. Grandes projetos de B&A podem levar muitos meses, ou até mesmo anos. Construir um *data mart* pode facilmente envolver projetos em horas na casa de cinco dígitos. Um projeto com uma solução de *MapReduce*, só na configuração dos nós, pode levar muitas semanas. Quando o projeto depende ainda de fornecedores, tema de seção posterior, os prazos podem ser ainda mais dilatados. É importante que o gerente de projetos, desde o início, gerencie as expectativas com relação a tempo. A preposição de entregas parciais e constantes, quando possível, como pregam as metodologias ágeis, podem ser uma forma de minimizar falsas expectativas e desconfiança de partes interessadas.

ORÇAMENTO INADEQUADO

No capítulo de custos vamos falar sobre ROI de projetos de análises de dados, onde veremos que, em geral, ele é muito bom. Na introdução desta obra, falamos brevemente que investimentos em projetos de B&A são uma questão de sobrevivência para as organizações. Mesmo assim, o tema orçamento é muito delicado, e o motivo é simples: em geral, estes tipos de projetos não são baratos! Projetos de B&A são de alto valor agregado e envolvem mão de obra, que além de escassa, é altamente especializada e qualificada, e isso tem um custo. Assim como qualquer outro projeto, não há projeto de B&A sem analistas de dados, administradores de bancos de dados, designers, cientistas de dados, programadores, analistas de negócios, entre outros. Depois vem a tecnologia envolvida, hardware e software. Certo, o projeto pode ser na nuvem? Sim, mas nuvem não significa de graça, certos tipos de negócio, como instituições financeiras, esta é uma possibilidade

totalmente descartada. Projetos podem envolver um grande ecossistema de softwares que podem custar pequenas fortunas. Construir um cluster *MapReduce* ou mesmo um *data warehouse appliance*, mesmo com o chamado "Commodity hardware"[3], pode requerer um grande investimento.

O gerente de projetos deve, logo que possível, eliminar falsas ilusões no sentido do que pode ser entregue com o orçamento existente ou qual o orçamento necessário para onde se quer chegar. A melhor forma de se fazer isso é demonstrando os custos aproximados envolvidos no projeto, como mão de obra, terceirização, software e hardware, mesmo que no início do projeto isto só seja possível em altíssimo nível.

FALTA DE COMPROMETIMENTO

O desejo de alguns nem sempre é a vontade de todos. Falta de comprometimento, porém, é uma causa comum na dificuldade e no fracasso de projetos de B&A. O dano mais comum ocorre em estruturas matriciais, onde o recurso humano está subordinado a um gerente departamental e é alocado paralelamente para atender ao projeto. Como já mencionamos, na vida corporativa, todos estão sempre muito ocupados. Para piorar, às vezes, o recurso necessário é o especialista no assunto, que praticamente possui conhecimento exclusivo no tema dentro da organização. É aquele que nunca tira férias. Claro que o recurso vai priorizar o departamento funcional, que é onde está sua estabilidade e continuidade, ao invés de um projeto passageiro do qual ele se vê como mero coadjuvante e que, muitas vezes, sequer vai se beneficiar diretamente. Não há aqui uma falta de tempo, mas sim, uma priorização de atividades de acordo com sua avaliação de relevância. É o que todos nós fazemos na nossa vida profissional e privada: usamos nosso tempo dando prioridade ao que entendemos ser mais importante.

3 Commodity hardware é como é chamado o equipamento comum, de uso geral e de baixo custo.

No caso do parágrafo anterior, que é o caso mais comum, não há dolo. Mas existem casos em que há, ou seja, em que indivíduos que deveriam colaborar com o projeto, propositalmente, tentam de alguma forma o sabotar. Por quê? Os motivos são diversos. Às vezes, um projeto do qual tinha mais interesse foi abandonado pelo projeto em andamento, medo de que o projeto diminua a sua visibilidade na organização (aquela pessoa que faz as consultas no banco de dados para produzir relatórios para todos os diretores), entre muitos outros. Mas o caso mais preocupante ocorre quando, em projetos de B&A relacionados à auditoria ou fraudes, um possível fraudador se sente ameaçado.

Como o gerente de projetos deve agir? No primeiro caso, em que há o conflito de tempo, o gerente deve logo no início do projeto negociar com o gerente funcional a disponibilidade do recurso x horas por semana ou mês. Devem ser estabelecidos os mecanismos que vão garantir a disponibilidade do recurso sempre que combinado, e qual será a ação caso, em algum momento, ele não esteja disponível. Porém, aqui há um fator crítico: muitas vezes, o gerente funcional não está a par do projeto, não é envolvido ou interessado, ou até mesmo pode se sentir, de alguma forma, ameaçado pelo projeto. Entra aqui, então, o termo de abertura: se o gerente do projeto seguiu devidamente os processos básicos de gerência de projetos, ele tem um termo de abertura, assinado pelo patrocinador do projeto, que dá a ele autonomia para alocar os recursos necessários quando necessários. O gerente funcional neste ponto já recebeu o termo de abertura e já está ciente da alocação do recurso e de sua priorização em relação a outras atividades. Outra atividade importante é evangelização de partes interessadas, que será discutida no capítulo 8.

No segundo caso, em que há uma suspeita em que o projeto possa levar a revelar um caso de fraude ou erro e o recurso tenta prejudicar o projeto, o gerente de projeto deve, com o apoio do patrocinador do projeto, tomar as medidas para imediatamente afastá-lo das atividades relacionados ao projeto.

RECURSOS HUMANOS

Aqui, também temos um tipo de risco que afeta muitas indústrias, e especialmente, projetos de tecnologia da informação. Mas projetos de B&A são mais suscetíveis a este tipo de risco, os motivos são muitos: projetos de B&A requerem mão de obra altamente especializada, que não é encontrada em abundância no mercado. Também as tecnologias e os projetos de B&A são mais recentes do que, por exemplo, projetos de desenvolvimento de software, e para piorar, a tendência é que a demanda seja crescente. Por outro lado, noto que existe um crescente interesse em programadores de sistemas e de banco de dados em se qualificarem para se tornarem analistas de dados ou cientistas de dados.

Então, nesta perspectiva, uma estratégia interessante para mitigar riscos relacionados a recursos humanos é capacitar desenvolvedores de sistema e administradores de banco de dados. Claro que a curva de aprendizado varia muito em qual competência na ciência de dados se quer qualificar o colaborador. Mas por exemplo, considero que a curva de aprendizado de um programador de banco de dados para um analista de BI é muito pequena. Já para um programador de sistemas para a mesma função de analista de BI é média. Um programador de banco de dados pode ser treinando em 40 horas, um programador de sistemas, em 80 horas. Claro que existem muitas variáveis, estes números servem apenas como um parâmetro.

FORNECEDORES

É bem provável que um projeto de B&A tenha um ou mais fornecedores envolvidos, sejam de soluções de software, equipamento de hardware ou alguma modalidade de software e/ou infraestrutura como serviços ou até mesmo de alguma empresa que fornece dados. Equipamentos não entregues, *bugs* críticos em soluções, APIs não documentadas, dados incorretos, são alguns dos muitos problemas que podem ser encontrados e que, sem dúvida,

podem colocar o projeto em risco. Grandes fornecedores podem parecer mais confiáveis, mas estes também têm seus riscos: podem colocar o seu chamado em uma fila de atendimento de dias.

Minimizar riscos com fornecedores não é tarefa fácil, teremos um capítulo inteiro para tratar disso, mas algumas dicas para o gerente de projeto não tão óbvias quanto verificar o histórico do fornecedor, é investir em expertise *in house*, de forma a ter um suporte interno, minimizando as dependências do fornecedor.

SEGURANÇA E PRIVACIDADE

Na era da informação, a segurança dos dados é uma preocupação constante de empresas, governos e pessoas. Ouvimos, todo dia, notícias de escândalos relacionados a acesso indevido ou vazamento de informações. Projetos em que não há qualquer requisito com relação à privacidade da informação, ou seja, todo dado é do domínio público, não são a regra.

Os riscos aos projetos relacionados à segurança de dados acontecem em três cenários: primeiro, eles estarão em um ambiente seguro, onde, se não forem dados públicos, será preciso algum tipo de permissão ou privilégio para acessá-los. Você, como gerente de projetos, deve garantir que os trâmites internos ou legais necessários para tal acesso sejam disponibilizados. Na segunda etapa, estes dados estarão em um ambiente de *staging*, onde serão acessados por consultores. Se houverem restrições de acesso a dados nesta etapa, o projeto se complica mais ainda. É preciso garantir que as ferramentas de desenvolvimento permitam que dados sejam mascarados ou substituídos em um ambiente intermediário para depois, no ambiente de produção, voltarem à sua forma original. Por fim, na etapa de produção, os dados podem estar sujeitos a uma série de restrições. Em um ambiente perfeito, mas que como falamos não é a regra, os dados são públicos e qualquer um com uma URL pode acessar. Num ambiente um pouco mais restritivo, um usuário terá que ter uma conta válida em um controlador de domínio para, a partir daí, acessar todo e qualquer dado.

Num projeto um pouco mais restritivo, usuários estarão associados a grupos, que estarão vinculados às tabelas, fato que poderão ou não ler seus dados. Mas a situação pode se complicar. Os requisitos de segurança podem requerer uma granulidade horizontal ou vertical, ou um misto das duas. Na granulidade horizontal, determinados usuários ou grupos de usuários tem acesso apenas a determinados registros das tabelas fatos. Por exemplo, um gerente pode não ter acesso a linhas com salários de diretores. Na granulidade vertical, certos usuários ou grupos de usuários podem não ter acesso a determinadas colunas de tabelas fatos, como a coluna onde consta a renda do cliente, por exemplo.

Em uma boa gestão de riscos de segurança de informação, o gerente do projeto deve, o mais cedo possível, identificar quais são os requisitos de segurança de dados do projeto em cada etapa, qual vai ser o impacto de configuração e manutenção destes requisitos, e acima de tudo, se as ferramentas tecnológicas do ecossistema do projeto suportam os requisitos definidos em cada etapa. Esta analise pode concluir que os requisitos são, por exemplo, inviáveis com as ferramentas existentes, ou mesmo que o projeto não poderá ser implementado da forma como foi pensado, requerendo adequações.

4. ESCOPO

Se você está gerenciando um projeto de engenharia civil, seus interessados estarão vendo o progresso. Mesmo muito antes de se levantar uma parede, já podem ver a construção do tapume, a energia provisória sendo ligada, o depósito de água sendo instalado. Já logo em seguida, começa-se a criar as estacas, vigas, colunas e lajes. A partir daí, o projeto passa a ficar cada vez mais tangível e de se aproximar cada vez mais do objeto final planejando entre o cliente, patrocinador, gerente de projeto e toda a equipe.

Agora, vamos pensar em um projeto de construção de um armazém de dados. O que os interessados vão ver durante o seu desenvolvimento? Máquinas virtuais configuradas? Scritps de carga programados e testados? Códigos de mapeamento e redução codificados? Processos de carga prontos e rodando redondos durante as madrugadas? E mesmo ao final do projeto, como justificar que algumas dezenas de tabelas fatos levaram 18 meses para ser construídas?

Tudo isso para dizer que o trabalho de um gerente de projetos analíticos não é fácil, nem quando o assunto é definir o escopo. Se estamos falando de projetos não tangíveis e de uma longa curva de entrega para usuários finais, é fundamental que o escopo seja definido apropriadamente.

GERENCIAMENTO DE ESCOPO

São funções do gerente do projeto e sua equipe: levantar o trabalho a ser feito e garantir que o trabalho necessário para

desenvolver o projeto está identificado; não deixar que mudanças sem uma justificativa afetem o projeto e se estas mudanças forem inevitáveis, controlá-las, manter o projeto no foco, dentro do que foi definido pelo termo de abertura, e, como em todo projeto, evitar que trabalho não acordado seja desenvolvido, ou seja, o cliente deve receber aquilo que foi planejado no projeto.

As atividades de gerenciamento de escopo envolvem ainda a criação de um ou mais documentos de escopo do projeto. Um ou mais porque pode haver uma declaração de escopo preliminar, criada e aprovada durante a iniciação do projeto, e um documento posterior mais completo, a declaração de escopo. Uma declaração de escopo deve conter o que deve ser entregue e ainda contar com uma aprovação formal do cliente. A gerência de escopo também não é uma atividade de execução exclusiva no início do projeto: o escopo deve ser monitorado durante toda a sua execução, no sentido de que o que foi planejado está sendo executado, inclusive, quando se trata de mudanças aprovadas. Deve ser ainda elaborada uma WBS (*Work Breakdown Structure*) da qual falaremos um pouco mais em seção posterior.

REQUISITOS

Definir os requisitos em um projeto de B&A é também um processo complexo e cercado de riscos. Quando pensamos em grandes projetos, em que o resultado só poderá ser avaliado meses depois, é sempre importante usar técnicas como pilotos e protótipos, já estudados no capítulo 3. Técnicas clássicas de levantamento de requisitos podem ser utilizadas neste tipo de projeto, vamos ver algumas delas:

- Entrevistas: a entrevista é a técnica clássica de coleta de requisitos, é bastante utilizada pela sua informalidade. Entrevistas funcionam melhor com um agendamento prévio, de forma que o entrevistado saiba com antecedência qual será o tema da entrevista. Também é fundamental que o entrevistador construa um roteiro básico a ser seguido, nada impedindo

que novas perguntas sejam feitas no decorrer da entrevista. É interessante anotar ou gravar a conversa. De forma geral, não é necessário formalizar a entrevista em uma ata, isso porque o levantamento de requisitos vai gerar o escopo do projeto, que será revisado e assinado pelos principais envolvidos. Um grande erro que gerentes de projetos e analistas de negócios inexperientes cometem é de tenderem a usar apenas esta técnica como levantamento de requisitos e acharem que está tudo entendido e esclarecido após meia hora de conversa.

- Questionários: hoje ferramentas online gratuitas e facilmente acessíveis são muito utilizadas para a criação de questionários. Alguns usuários ainda preferem este tipo de abordagem para o levantamento de requisitos, pois têm tempo de pensar e elaborar de forma mais apropriada suas respostas. Porém, como toda técnica, existem algumas limitações: não é possível elaborar novas perguntas baseadas nas respostas dos entrevistados. Outro problema grave é que grande parte dos envolvidos não vai responder ao questionário, a não ser que você encontre um bom motivador que leve a grande maioria a respondê-lo.
- *Brainstorming*: *brainstorming* é uma técnica executada em grupos, onde o objetivo é gerar o maior número de ideias que deverão estimular o surgimento de novas ideias. Algumas regras, normalmente, adotadas nesta técnica: todos os presentes devem participar; devem-se evitar críticas ou julgamentos; a reunião deve ser documentada.
- *Delphi*: o *brainstorming* pode fazer com que boas ideias não sejam expostas, pois, mesmo não podendo haver críticas, os participantes podem ficar constrangidos a expor ideias que pensam ser ridículas, mas que, na verdade, não são. *Delphi* é outra técnica realizada em grupo, que tem como princípio básico o anonimato. Primeiramente, o tema é exposto, e todos expõem ideias em papéis, de forma anônima. As ideias são expostas e debatidas, o que pode levar a novas rodadas, até haver um consenso.

- Prototipagem: aqui, a prototipagem serve para validar com os envolvidos as interfaces com os usuários, normalmente, objetos de visualização como painéis, relatórios e KPIs. Neste caso, a prototipagem é um desenho estático da aplicação desenhada em uma ferramenta especializada como *Visio* ou *Balsamic*, de forma que o usuário, antecipadamente, pode prever o que vai ser entregue.

WBS

O WBS (*Work Breakdown Structure*), no português EAP (Estrutura Analítica do Projeto) é um artefato importantíssimo, mesmo em projetos de B&A. O WBS é um diagrama hierárquico, muito embora, diferente de um diagrama de rede, as atividades não estão relacionadas pela dependência, mas sim, pelos entregáveis. Tudo que o projeto vai entregar, deve estar no EAP, inclusive, artefatos de gerência de projetos. Deve ser criado pela equipe do projeto sobre orientação do gerente do projeto.

O primeiro nível do WBS deve ser único e deve conter o nome do projeto, em níveis intermediários devem conter fases ou assuntos que compõe o projeto. No último nível, as atividades que devem ser executadas, conhecidas como pacotes de trabalho, que vão servir como base para a construção do cronograma.

PARA QUEM É O PROJETO?

O CEO, entendendo as necessidades de soluções analíticas para a saúde dos negócios, aprova um grande orçamento para projetos analíticos. Um diretor de um departamento específico, vê, então, a chance de implementar o tal painel que ele semanalmente tinha que, depois que recebesse alguns dados do departamento de TI, montar no Excel. O projeto é executado, e alguns semanas depois, o diretor tem painel implementando em uma ferramenta altamente moderna com todos os processos automatizados. Dois meses depois, o diretor vai embora da empresa. O novo Diretor,

após assumir suas funções, até conhece o painel e o consulta uma vez, mas não enxerga muita utilidade, além de discordar como as medidas que foram calculadas. O painel nunca mais é utilizado.

Talvez, pareça uma história exagerada, mas não é. Um grande problema em projetos analíticos são projetos de grande porte que são implementados para responder perguntas pessoais que certos usuários têm em determinado momento. Para isso, existem relatórios ad hoc e ferramentas self-service. Em geral, projetos têm que ser implementados visando ao departamento e à corporação, não necessidades passageiras ou pessoais. É função do gerente do projeto garantir que, por exemplo, um projeto de construção de um *data mart* financeiro seja modelado e construído para atender ao departamento financeiro e não às luxúrias de um ou dois diretores.

5. TEMPO

Como já comentamos, em capítulos anteriores, um dos problemas de projetos de B&A, é que, normalmente, são projetos longos, e que as entregas de fato vão ser tangíveis ao final, às vezes, depois de meses ou anos. Tempo é um elemento crítico, que merece especial atenção da equipe do projeto.

DEFINIÇÃO DAS ATIVIDADES E MARCOS

Uma etapa crítica em projetos é definir as atividades. Tornar o projeto mais facilmente gerenciável e minimizar riscos, significa identificar todas as atividades que devem ser desenvolvidas. Às vezes, atividades que aparentemente são simples, e por isso, são subestimadas e sequer se planeja tempo para elas, complicam e atrasam o projeto. Por exemplo, uma ferramenta de visualização Web já foi configurada várias vezes em um servidor de Internet, porém, por alguma questão técnica, no servidor do projeto, ela não roda e se perdem dias em sua configuração. Também se devem definir marcos. Recordando o conceito, marcos são eventos importantes do projeto, que têm uma data prevista para ocorrer, esta data depende da dependência das atividades com o marco. Por exemplo, ter a infraestrutura de testes configurada e testada, pode ser um marco. Os marcos são importantes para definir etapas importantes do projeto.

ESTIMATIVA DE DURAÇÃO

Um grande erro na estimativa de duração, em atividades de B&A, é que a equipe tende a estimar as atividades pelo tempo ideal. O tempo ideal, de certa forma, não é uma estimativa errada, porém, ela supõe que tudo vai dar certo conforme planejado. Estimar com tempo ideal, é subestimar os riscos inerentes ao projeto, pois, na prática, uma infinidade de coisas que afetam a duração das atividades dá errado ao longo do projeto: recursos ficam doentes, analistas de negócio são convocados para reuniões mais urgentes que o projeto, servidores caem, redes ficam lentas, fornecedores atrasam, *bugs* misteriosos aparecem. Então, estimar um tempo de reserva para cada atividade, não é superestimar, é apenas estimar!

Quanto tempo de reserva estimar, quando ela poderá variar para mais ou para menos? Você pode fazer uma estimativa mais rápida e imprecisa, simplesmente, atribuindo um percentual de acordo com a natureza da atividade. Por exemplo, atividades de negócio, como entrevistas, levantamentos de requisitos, documentações, entre outras, devem receber entre 5 e 7% de tempo de reserva. Atividades de caráter técnico já conhecidas, de 10 a 15%, e atividades técnicas em que a equipe tem pouco ou nenhum conhecimento, ou seja, que nunca foram feitas, ou que foram feitas poucas vezes antes, de 14 a 20%.

Uma forma mais científica de apurar reservas, mas ainda assim com agilidade, é calcular o desvio padrão da estimativa. Por exemplo, se usamos a técnica de estimativa PERT, que prevê uma estimativa pessimista (P), uma otimista (O) e outra mais provável (M), a fórmula para esta estimativa é:

$$\frac{(P + 4M + 0)}{6}$$

Estimada a atividade, estima-se o seu desvio padrão, que se dá pela fórmula abaixo:

$$\frac{(P - 0)}{6}$$

Dessa forma, a estimativa para a atividade vai ser PERT mais ou menos o desvio padrão. Vamos ver um exemplo prático: se a atividade teve uma estimativa otimista de 41, mais provável de 60 e otimista de 89, seu PERT é de 61,66 e o desvio padrão é de 8. Dessa forma, a atividade deve variar entre 53,66 e 69,66.

TÉCNICAS DE ESTIMATIVA

As técnicas de estimativa em projetos de B&A podem ser as mesmas usadas tradicionalmente em projetos de TI:

- Analogia: se em projetos anteriores já foram configurados nós Hadoop, o histórico de duração dos projetos anteriores pode ser uma boa base. Claro que, normalmente, não se trata de simplesmente copiar a duração anterior. Por exemplo, pode ser que se tenham técnicos mais experientes, ou que o número de nós não seja o mesmo.
- Opinião Especializada: as estimativas cabem à equipe, e não ao gerente de projetos. Na opinião especializada, usa-se a experiência da equipe na execução de atividades anteriores. Pode ser usada em conjunto com uma estimativa de três pontos, como PERT, que estudamos em seção anterior.
- *Delphi* ou Brainstorminig: algumas técnicas de decisão em grupo também podem ser usadas. *Delphi* é uma técnica que preza pelo anonimato. *Brainstorming* busca incentivar as opiniões de todos, evitando que haja críticas. Estudamos estas técnicas no capítulo 4, Escopo.

Técnicas ágeis de estimativa, como *Planing Pokker* e *Ideal Day*, também podem ser usadas em projetos de B&A e podem ser bem eficientes. No *Planing Pokker*, é usado um baralho especial, a atividade é descrita e a equipe vira a carta com a estimativa, dessa forma, não há risco de os membros da equipe estimarem em função do que um colega estimou. Em seguida, o grupo pode

discutir a respeito da estimativa. Por exemplo, alguém estimou 5 horas a mais que o restante da equipe para certa atividade, ele pode explicar que a atividade deve ter um processo de mudança de codificação dos caracteres, que não é muito simples de ser feito. O restante do grupo admite que não havia considerado este fato e concordam que as suas estimativas estavam abaixo do que de fato realmente seria necessário para desenvolver a atividade.

6. CUSTOS

O material publicitário de fornecedores de soluções de B&A, falam em projetos fáceis, rápidos e baratos, graças às maravilhas tecnológicas que eles criaram. A verdade é que grandes projetos de B&A não são baratos: podem ter um custo com seis ou mais dígitos. Por isso, uma boa gestão de custos, do início ao fim do projeto, é uma atividade fundamental, que deve ser exercida com bastante profissionalismo pelo gerente de projetos e sua equipe.

ROI

O ROI deve fazer parte de qualquer projeto, mas em especial, em projetos de longa duração e altos custos, ele se faz mais importante, para minimizar desconfianças e justificar investimentos. O cálculo do ROI, assim como *Business Case* descrito no capítulo 2, pode não ser função do Gerente de Projetos, e sim, do patrocinador. Porém, se o projeto já foi aprovado e autorizado pelo patrocinador do projeto sem um ROI, é recomendável que, se possível, o gerente de projetos o faça.

TCO

TCO (Total cost of ownership), ou custo total de propriedade, é quanto a solução vai custar para ser adquirida e implantada. Supondo que o seu projeto de B&A seja um produto de mercado, um software. Quando a área comercial for vendê-lo, provavelmen-

te, o potencial cliente não vai olhar apenas o custo da licença, mas custos de manutenção, de treinamento, de recuperação de falhas e desastres, crescimento, entre outros. Não é incomum produtos terem um preço de licenciamento menor, mas um TCO é bem superior a um concorrente do mercado.

ANÁLISES FINANCEIRAS EXTERNAS

É comum, principalmente em grandes projetos, empresas fazerem a sua própria estimativa de custos. Às vezes, uma empresa externa é contratada para estimar. Em outros casos, embora não seja apresentada uma estimativa paralela, a estimativa do projeto é analisada e pode ser amplamente questionada por especialistas na área. Por isso, é fundamental que todos os custos do projeto estejam muito bem fundamentados, pois, se você fez uma boa estimativa, provavelmente, ela será a mais cara. Vou dar um exemplo: a sua estimativa do projeto prevê a criação de uma prova de conceito, ao custo de R$ 20 mil. Provavelmente, a estimativa paralela não terá a tal prova de conceito, por isso, já em seu orçamento, é imprescindível que ela esteja justificada:

> *"Prova de Conceito: Foi identificado que uma das fontes de dados para o projeto "Sistema de Monitoramento Contínuo de Fraudes Financeiras" foi desenvolvida utilizando um banco de dados DataFlex. Considerando que: a) não existe conector nativo para esta fonte de dados; b) não há na equipe ou na empresa histórico anterior de extração destas fontes de dados; c) sabe-se, historicamente, que a extração deste tipo de fonte de dados não é um processo trivial; d) os requisitos de extração, para o projeto em questão, são de certa complexidade: alta frequência e volume. Portanto, para minimizar riscos de a extração inviabilizar o projeto já quando considerável volume de aporte financeiro foi realizado, optou-se por fazer a prova de conceito, que será composta pelas seguintes atividades..."*

Também outro problema encontrado em estimativas paralelas é que elas são feitas com o tempo ideal, aquele que não prevê os imprevistos naturais que ocorrem em projetos e de que já falamos. Convencer executivos que, com razão, estão preocupados em gastar o menos possível no projeto, de que tempo de reserva deve ser estimado, pode não ser uma tarefa fácil. Uma das formas é apresentar documentações históricas de projetos anteriores, além de evangelizá-los sobre o conceito de "tempo ideal" usado nas estimativas. Um último problema com análises externas é que os estimadores desconhecem de fato a natureza das atividades. Por exemplo, são acostumados a estimar desenvolvimento de software e não um projeto de *MapReduce*. Cabe a você, como gerente de projetos, demostrar documentos de outros projetos, mostrando o tempo de atividades semelhantes.

TIPOS DE CUSTO

Primeiro, vamos recordar algumas definições. Custos fixos são aqueles que, independente do que se está produzindo, não variam. Já custo variável tem variação em relação à quantidade do que se produz. Um projeto de B&A vai ter custos fixos semelhantes a outros tipos de projeto: luz, aluguel, salário de pessoal administrativo, entre outros. Já os custos variáveis serão salários de desenvolvedores, arquitetos e designers, fornecedores, entre outros. Custos também são classificados como diretos e indiretos. Os custos diretos são aqueles relacionados diretamente ao desenvolvimento do projeto. São custos diretos o de software e equipamento de infraestrutura. São custos indiretos, material de escritório, depreciação etc.

ESTIMATIVAS DE CUSTOS

Existem várias técnicas que podem ser utilizadas para estimar custos. Custos que tratam da aquisição de produtos, ou serviços de terceiros, podem ser consultados através de solicitações

de propostas ou licitações. Falaremos mais sobre aquisições no capítulo 9. Ainda se pode avaliar projetos anteriores com atividades ou aquisições semelhantes. Por fim, custos podem ser estimados baseados na estimativa de tempo das atividades, com técnicas como as estudadas no capítulo passado.

Assim, como há estimativa de tempo, estimar custos requer que riscos sejam considerados, ou seja, que valores adicionais sejam contabilizados para quando coisas não saírem como planejado ou situações imprevistas ocorram, são as reservas de contingência. Por exemplo, se uma atividade está estimada em R$ 20 mil, existe uma probabilidade de ocorrência de um risco de 5%, seria o ideal ter alocado ao projeto R$ 1 mil para contingências. O total de valores para contingências seria a soma do total de todas as atividades. Estávamos falando até agora de riscos conhecidos. Estima-se que pelo menos 10% dos riscos do projeto são impossíveis de prever. O ideal é que também se estimem os valores de contingência para este tipo de risco.

RECURSOS EXISTENTES

Pode parecer óbvio, mas antes de sair em busca de uma solução para um projeto de B&A, precisamos olhar o que temos dentro de casa. Não somente no que diz respeito a um software ou tecnologia já licenciada, mas também a pessoas que possuam expertise em produtos de mercado. Às vezes, comprar uma solução da qual já se tem um corpo técnico qualificado, pode sair mais barato do que adquirir outra solução, mesmo que esta outra solução não tenha custo de licenciamento, pois, lembrando mais uma vez, devemos olhar sempre o TCO.

OPEN SOURCE

É difícil, hoje em dia, pensar em implementar uma solução de B&A sem considerar soluções *open source*: esta é uma área em que projetos *open souce* existem em abundância, são dezenas deles.

Vamos ver alguns:

- Ecossistema *Hadoop*: o *Hadoop* é quase um sinônimo para *Big Data*. Mantido pela Fundação Apache, trata-se de uma solução de análise de dados distribuída, inspirada no modelo *MapReduce*, e operando em conjunto com um sistema de arquivos distribuídos HDFS. Ao redor do *Hadoop* orbitam diversos outros projetos com finalidades específicas, como gerenciamento, acesso a bandos de dados relacionais, execução de consultas com linguagem declarativa, entre outros.
- R: o R é uma poderosa ferramenta de estatística, análise, mineração de dados e visualização de linha de comando. É extensível através de pacotes que são desenvolvidos por terceiros e incorporados à ferramenta. Existem milhares de pacotes desenvolvidos, gratuitos e prontos para o uso. O sucesso do R é tanto, que ele tem sido incorporado dentro de ferramentas de análise de dados de grandes players do mercado, com a *Oracle* e *Microsoft*.
- *Spark*: enquanto o *Hadoop* é um sistema de análise de dados em batch de computação distribuída, o *Spark* é um produto de análise de dados em memória, por isso, ideal para análise em tempo real ou próximo ao tempo real. Também é mantido pela fundação Apache.
- MongoDB: um banco de dados NoSQL orientado a documentos, de fácil instalação e operação.

Supondo que existam duas ferramentas semelhantes no mercado (você já fez um benchmark com ambas), uma open source e outra proprietária. Quando optar por uma solução open source? Esta é uma pergunta complexa, e portanto, a resposta não é simples. Devemos avaliar dezenas de fatores, entre eles, a criticidade do projeto e orçamento. Em uma solução *open source*, haverá um suporte técnico com uma SLA para resolver problemas e falhas? Você possui expertise interna para prestar suporte a solução? Existem cases conhecidos e de sucesso no uso da solução? De-

vem-se considerar ainda empresas que fornecem uma distribuição de uma solução *open source* com suporte técnico.

Todas estas questões devem ser avaliadas com cuidado. Lembre-se de que já falamos em seções anteriores, o custo de licença não representa o custo total de aquisição e uso de uma solução, devemos sempre avaliar seu TCO.

Existem empresas cujas políticas de governança trazem restrições a adoção de produto *open source*. Por exemplo, elas podem permitir que uma solução *open source* seja incorporada, apenas se houver uma empresa que ofereça licenças com suporte técnico. Você fazendo parte ou não de uma empresa com este tipo de restrição, este é um fator a considerar: adquirir um produto *open source* licenciado por uma empresa que mantém suporte técnico especializado de um produto *open source*. Muitas destas empresas ainda fornecem versões melhoradas destes produtos.

BIG DATA COMO SERVIÇO

Um projeto de B&A significa que será preciso capacidade de armazenar e processar dados. Esta capacidade pode existir na organização, pode ser incorporada ou pode ser comprada de terceiros. Quando ela é contratada, temos BdaaS (*Big Data* as a Service), ou *Big Data* como serviço. De fato, são vários elementos que, em conjunto ou separadamente, constituem BdaaS:

- IaaS, *Infraestructure* as *a Service*, ou Infraestrutura como Serviço, são oferecidos serviços computacionais, normalmente, através de máquina virtuais, incluindo serviços de segurança e backup e facilidade para escalar a aplicação (normalmente, pode acontecer até de forma automática).
- SaaS, Software *as a Service* ou Software como Serviço, oferece software específico, que pode incluir um gerenciador de banco de dados. A infraestrutura que executa o software é transparente para o usuário e já vem instalada.
- PassS, acrônimo de *Plataform as a Service* ou plataforma como serviço, é o serviço mais amplo, temos toda a plataforma de

serviços à disposição para implementar o projeto, que pode incluir servidores, rede, armazenamento, sistema operacional, banco de dados, entre outros. Pode ser, por exemplo, uma plataforma com *Hadoop* ou uma de suas distribuições, já com *nodos*, *firewall*, links, armazenamento configurados e prontos e para a utilização.

Alguns serviços de BdaaS são bem populares e fornecidos por gigantes da TI, como o *Amazon Elastic MapReduce* e HD*Insight* da *Microsoft*, que oferecem *Hadoop*, *Spark* e R em um ambiente na nuvem. Contratar um serviço de B&A na nuvem não significa uma solução monolítica. Às vezes, partes da solução podem ser *on primeses* e outros *on cloud*. Por exemplo, se seu projeto envolve um streaming de dados de alto volume, colocá-lo em uma nuvem de terceiros para processamento pode ser inviável devido a limitações em taxa de transferência de rede. Pode-se ter, então, uma solução de carga e processamento *on primeses* e a visualização *on cloud*, com um serviço de BI na nuvem como o Power BI da Microsoft.

7. QUALIDADE

Normalmente, a maior preocupação da equipe do projeto é a entrega final, que pode ser na forma de relatórios, painéis, dados formatados (um arquivo de conformidade, por exemplo), entre outros. De fato, este é um elemento importante. Porém, o gerente de projetos e a equipe devem ter em seu radar outros aspectos da qualidade. Primeiro, não é apenas a qualidade do que é produzido, mas da matéria-prima do projeto: o dado. De nada vai adiantar todo o esforço da equipe em fazer do projeto um sucesso, se os dados não forem bons: o projeto só será tão bom, quanto os dados forem. Falaremos mais sobre qualidade de dados na próxima seção.

Já olhando para o que será entregue, claro que o objeto final é importante, porém, deve-se zelar pela qualidade de todos os artefatos e processos que forem criados durante o projeto, pois estes também são parte da entrega. Como artefatos, nós nos referimos a conectores, documentação técnica, como diagramas, escopos, documentos de manutenção, de recuperação de desastre, de manutenções preventivas, padrões de codificação, entre outros. Como processos, referimo-nos a extratores, procedimentos de extração, transformação e carga: mais do que rodando e livre de erros, eles devem estar otimizados e amplamente testados.

Também é importante que os critérios de qualidade sejam definidos entre a equipe do projeto e o cliente, e que todos tenham claros estes critérios. Vamos ver diversos aspectos da qualidade em projetos de B&A.

O QUE É QUALIDADE DE DADOS

Quando pensamos em qualidade de dados, normalmente, lembramo-nos de problemas como um endereço cadastrado duplicado. Por exemplo, um iniciando com "AV." e outro com "Avenida". Embora este seja um indicativo de dados sem qualidade, qualidade de dados é um conceito mais profundo. Primeiro, a qualidade é relativa ao seu propósito, aos seus usuários e à temporalidade dos dados. Por exemplo, certos dados podem estar atendendo aos requisitos de qualidade no sistema transacional, mas precisam de tratamento para carga no *data warehouse* para atender os requisitos deste negócio. Também a qualidade não está relacionada apenas a como o dado está registrado fisicamente no banco de dados ou sistema de arquivos, mas a outros quesitos não físicos. Por exemplo, temporalidade: todo dado tem um valor que se degrada com um tempo. Supondo que ocorra uma fraude financeira. O valor da informação é muito maior se os dados apontarem a fraude no dia da ocorrência do que seis meses depois. Outro exemplo é quanto à credibilidade, que se refere à confiabilidade dos dados. A maioria dos gestores, acostumados à tomada de decisão baseada em dados, já sofreram com dados não confiáveis: o valor das vendas está correto? O imposto a ser pago é de fato este? A provisão de caixa não está errada?

QUALIDADE DE DADOS NO CONTEXTO DE *BIG DATA*

Em projetos de análise de dados tradicionais, as fontes de dados de projetos eram internas, a maioria estruturados, provenientes de banco de dados relacionais ou mainframes. Neste contexto, os padrões tradicionais de governança de dados e as ferramentas técnicas de qualidade de dados ainda se aplicam. Hoje, existem duas diferenças essenciais entre projetos clássicos e de *Big Data* quanto à qualidade de dados: primeiro, que dados externos as organizações são agora utilizadas em muitos projetos: são da-

dos de redes sociais, de *data brokers*, de fontes públicas, celulares entre outras. Dessa forma, não há como a empresa impor políticas de governança e qualidade de dados, tudo o que pode ser feito é um trabalho de limpeza de dados no processo de transformação. Outra mudança é que cada vez mais dados não estruturados estão presentes nestes projetos, este é o elemento variedade dos "Vs" que conceituam *Big Data*. Muito embora a maior quantidade de dados eletrônicos no mundo seja não estruturada, a maior parte das fontes de dados para projetos de B&A ainda são de dados estruturados, especialmente, bancos de dados relacionais.

Dados em pequena quantidade podem ser arrumados manualmente, por rotinas de programação ou por ferramentas especializadas. Mas no contexto de grandes volumes de dados não estruturados, precisamos usar técnicas não tradicionais para limpeza, como algoritmos especializados. Quando o assunto é mineração de texto, os próprios algoritmos são capazes de eliminar o que é lixo daquilo que tenha valor semântico. Por exemplo, em mídias sociais, um algoritmo de aprendizado de máquina é capaz de detectar que tuítes contendo a palavras "sux" podem estar relacionadas a um sentimento negativo, bem como "rox" a um sentimento positivo.

TESTES EM *BIG DATA*

Testar projetos que envolvam grandes volumes de dados e *MapReduce*, requer que o pessoal de qualidade e testes sejam capacitados neste novo conceito, pois o processo é exponencialmente mais complexo. Não é o escopo desta obra testes. Porém, deixo aqui algumas observações já que este tema está intimamente relacionado com qualidade. Os testes em projetos de B&A devem ser executados sob duas perspectivas: funcional e infraestrutura.

Do ponto de vista funcional, os testes devem verificar se as regras de negócio estão corretamente implementadas. Falamos no capítulo de riscos sobre medidas, que não se referem especificamente sobre medidas de um *data warehouse*, mas a cálculos

gerados pelas aplicações. Medidas, em especial, podem se tornar um pesadelo para testadores, pela sua complexidade, pela falta de documentação, pela falta de consenso entre pessoas de negócio. Aqui, também estão os testes gerais de saídas geradas pelo sistema, como painéis, relatórios ou arquivos de dados.

Do ponto de vista de arquitetura, o ambiente deve ser testado sobre desempenho, escalabilidade e confiabilidade. Deve-se verificar tempo de processamento, uso de memória, uso de CPU e taxa de transferência de dados. Também se devem testar sistemas de recuperação de falhas: testes de *failover*. Por exemplo, algum nó pode ser desligado para verificar se o sistema consegue se recuperar.

Já falamos bastante sobre a qualidade dos dados de origem, ou seja, daqueles que serão extraídos de fontes diversas para originar o projeto. Pessoalmente, não considero a avaliação dos dados de origem como uma função da equipe de testes. Esta é uma etapa de viabilidade técnica, que deve ser executada previamente ao início do desenvolvimento do projeto.

FALTA DE QUALIDADE

Falando da qualidade do produto do projeto como um todo (e não apenas da qualidade dos dados), esta tem um custo alto: pilotos, planejamentos, provas de conceitos, ambientes de teste, testes etc. Porém, a falta de investimento em qualidade tem um custo bem maior. Uma boa equipe de qualidade é capaz de achar a equação certa entre investimentos em qualidade e defeitos na solução. Primeiro, é preciso estar ciente de que não existe projeto de B&A perfeito, e eventualmente, defeitos irão ocorrer. Isso não significa que se deva entregar um projeto cheio de problemas para a equipe de produção resolver. Também não significa que devesse investir quatro vezes o valor do projeto em testes unitários e/ou de regressão. Isso porque a relação entre investimento em qualidade e defeitos é inversamente proporcional, mas não é linear. Em outras palavras, se você investir 10 em qualidade, vai remover

5 defeitos. Mas se investir 20 em qualidade, ao invés de remover 10 defeitos, removerá 6. Devemos buscar a equação ideal entre investimento e a qualidade desejada.

Quando se cria um script ou o código Java de uma função de redução, junto com as funcionalidades, o desenvolvedor está inserindo defeitos dentro do código. A abordagem mencionada logo acima, de deixar defeitos para serem solucionados em produção, é um aspecto da falta de inspeção. Na inspeção, o defeito já foi incluído dentro do produto do projeto, a última chance de tirá-lo do usuário final, é inspecionando o que foi feito, no caso de um projeto de B&A, com testes. Se a inspeção falhar, o sistema em algum momento vai falhar. O processo mais barato para o projeto, e aqui barato não se refere apenas a questões financeiras, é prevenção. Na prevenção, o defeito não é incluído dentro do produto do projeto. Não incluído, pode-se investir menos em inspeção e ter um produto final mais confiável. Como se aplica prevenção em projetos de B&A? Boas práticas de desenvolvimento, treinamento, checklists, testes unitários, revisões de código, *pair programming*, entre outras. Em termos de arquitetura, pilotos, protótipos, *benchmarking*, testes de taxa de transferência, de carga, de *failover* etc.

Como já falamos, se a qualidade tem um custo, a falta de qualidade é ainda mais cara. Produtos sem qualidade afetam a produtividade, causam retrabalho, aumentam os riscos, trazem incertezas para a equipe envolvida no projeto destroem a motivação.

EXCEDER AS EXPECTATIVAS

A metodologia de gestão de projetos clássica usa o termo g*old plating* para se referir à funcionalidade entregue que não estava previamente acordada. Afinal, sobrou tempo no projeto, por que não fazer uma nova versão do painel agora com os dados sumarizados? Minha experiência pessoal é da opinião de não realizar entregas supérfluas como sugere a gestão clássica: fazer o que foi acordado bem feito, nada mais. Por quê? Primeiro, que é perda de tempo e dinheiro. Também entregas adicionais podem causar uma

impressão errada ao cliente ou patrocinador. Imagine a seguinte situação: toda vez que quer uma funcionalidade ou alterar algo previamente acordado, a equipe do projeto inicia um controle de mudanças. Ainda se estas mudanças requerem mais tempo, terá que ser feito um ajuste no orçamento. Se sua empresa está vendendo serviços, você terá que fazer um ajuste no preço do projeto. Se você está contratando serviços, terá que pagar mais e também ajustar o seu orçamento. Se, de repente, a equipe começa a adicionar itens não combinados ao projeto, pode parecer que é fácil e grátis mudar ou adicionar escopo. Isto se agrava porque, para um usuário final, normalmente, parece simples adicionar mais uma medida no produto final, quando, na verdade, isto vai causar um efeito em cascata em todo o processo de produção daquela informação.

Outro problema são os possíveis efeitos colaterais. Às vezes, um desenvolvedor que, na melhor intenção, resolve adicionar uma funcionalidade ao produto, não conhece a arquitetura do projeto como um todo, e por isso, não é capaz de dimensionar o impacto, por exemplo, de simplesmente adicionar um novo atributo no processo de carga para exibir no painel. Pode ocorrer, por exemplo, que todo o processo de transformação de dados tenha que ser revisado, além de atualizações em layouts e medidas.

8. RECURSOS HUMANOS E PARTES INTERESSADAS

Neste capítulo, vamos falar de recursos humanos em dois grupos: primeiro, a equipe do projeto, que se refere apenas aos recursos alocados diretamente no desenvolvimento do projeto de B&A, o que exclui pessoas interessadas como patrocinadores e usuários finais; o segundo, são as partes interessadas, especialmente, os usuários finais, que deverão se beneficiar diretamente das entregas do projeto. No primeiro grupo, o desafio é montar uma equipe adequada ao projeto proposto e prover toda a capacitação necessária. Para o segundo grupo, é preciso não só vender o projeto, mas evangelizar os usuários no consumo de informação e decisões baseadas em dados.

MONTAGEM DA EQUIPE

A montagem da equipe significa definir os papéis no projeto e quando as especialidades serão necessárias. Nem sempre o gerente de projetos pode escolher os membros da equipe, normalmente, eles já estão contratados na empresa, podem estar alocados em outros projetos ou ainda podem exercer alguma atividade funcional. Esta etapa também pode significar a contratação de recursos especializados. Sabemos que a tecnologia de informação carece de mão de obra especializada. Em projetos de B&A, a situação é ainda mais complicada, estima-se que nos próximos anos, para cada quatro vagas de cientista de dados, haverá apenas um profissional qualificado no mercado.

NEGOCIAÇÃO DE RECURSOS

Como falamos em seções anteriores, recursos necessários ao projeto, muitas vezes, já fazem parte do quadro de colaboradores da empresa e desempenham uma atividade funcional. Pode ser o analista fiscal, o analista de recursos humanos, o consultor do ERP, o coordenador de produção, entre outros. Em uma empresa com estrutura matricial, este consultor vai estar subordinado a um gerente ou diretor de área. Para complicar, geralmente, estes recursos, especialistas em suas áreas, estão sempre bastante ocupados, e vão priorizar as atividades da empresa em detrimento projeto. Por isso, é fundamental que o gerente de projeto negocie com os gestores das áreas a alocação dos recursos, dentro de um planejamento previamente proposto e aprovado por ambos.

RESPONSABILIDADES

É importante que as responsabilidades em projeto de B&A estejam definidas e claras. Vamos começar pelo patrocinador. Para começar, ele é o principal responsável por aprovar os principais planos do projeto e, não menos importante, dar ao gerente de projeto a autoridade necessária para que possa tomar as decisões necessárias para o bom andamento do projeto.

A equipe do projeto, além de suas tarefas específicas relacionadas à execução das atividades do projeto, deve auxiliar o gerente de projetos em estimar o desenvolvimento das entregas e em construir o cronograma.

Finalmente, o gerente de projetos deve planejar todo o trabalho necessário para desenvolver o serviço ou produto que seja objeto do projeto, coordenar recursos, garantir que os objetivos do projeto sejam atingidos e que ele entregue valor.

PLANO DE TREINAMENTO

Também, no capítulo de gestão de riscos, falamos da possibilidade de treinar recursos que, em alguns casos, já estariam

pré-qualificados. Hoje, existem muitos ambientes virtuais especializados em treinamentos relacionados a B&A, como R, *Hadoop*, estatística, *Big Data* etc. Muitos destes ambientes são totalmente gratuitos, ou oferecem versões certificadas pagas. A esmagadora maioria do treinamento disponível online é em inglês. No Brasil, começam, aos poucos, a surgir cursos e até pós-graduações em *Big Data* ou temas relacionados. Existem centros de treinamento nas grandes cidades que já ministram cursos de distribuições *Hadoop*, mineração de dados, entre muitos outros. De qualquer forma, deve fazer parte do projeto um plano de qualificação, a fim de aprimorar as habilidades da equipe.

Também é interessante a equipe frequentar eventos de B&A. O mais proeminente é o *Strata + Hadoop World*, que ocorre quatro vezes ao ano em diversas partes do mundo, mas que o Brasil, entretanto, não está incluído. Felizmente, aos poucos, alguns eventos relacionados ao tema têm ocorrido no nosso país, muitos deles onde *Big Data* não é o tema principal, mas já ocupa posição de destaque.

EVANGELIZAÇÃO

Democratizar a informação, a tomada de decisão baseada em dados, habituar os gestores a consumirem produtos e serviços de B&A. Contudo, muitos projetos, apesar de serem bem-sucedidos, esbarram nas questões citadas acima. Junto com projetos de B&A, pode-se propor programas de evangelização em tecnologia relacionadas. Produzir informação de qualidade a partir de dados é um desafio cada vez mais complexo. Para quem consome informação digital, a curva é inversa. Se há alguns anos, um gestor para conseguir uma informação de qualidade duvidosa, tinha que ter um programador de banco de dados à disposição, hoje, ferramentas de BI self-service, são cada vez mais comuns e acessíveis. Do ponto de vista visual, as ferramentas de visualização evoluíram de forma surpreendente. Hoje, com um pouco de treinamento e quase nenhum talento, é possível ao usuário final construir painéis

ricos, interativos e altamente agradáveis visualmente. Por estes motivos, considere propor em seu projeto um plano de evangelização, que para este público-alvo não requer muito tempo nem grandes orçamentos.

COMPETÊNCIAS E RESPONSABILIDADES

Desenvolver projetos de B&A requer a participação de uma equipe multidisciplinar. São pessoas com conhecimento especializado e de alto valor agregado. Nesta seção, vamos ver as principais especialidades que compõem estes projetos. A relação não é exaustiva, mas sim, apresenta as especialidades mais comuns:

- Gerente de Projetos: já estudamos as competências necessárias a um gerente de projetos de B&A no capítulo 2. Sobre responsabilidades, são as já conhecidas e comuns a gerentes de projetos que atuam em outras indústrias: gerenciar custos, escopo, cronograma, alocar equipe, gerir riscos, gerenciar partes interessadas entre outros.
- Arquiteto: o arquiteto é responsável por estabelecer a arquitetura da solução. Pode ser um arquiteto de softwares, quando o projeto entrega um produto de desenvolvimento de software. O arquiteto deve definir padrões e formas de desenvolvimento, criar *frameworks*, definir boas práticas etc.
- Analista de Negócios: o analista de negócios deve atuar junto às áreas de negócio da organização onde o projeto se desenvolve, em busca de oportunidades de negócio. Atua também no projeto na função de elicitação de requisitos e levantamento de escopo.
- Cientista de Dados: o cientista de dados é o profissional com conhecimento mais abrangente em B&A, dessa forma, é capaz de propor e implementar soluções mais adequadas para cada tipo de problema dos projetos de B&A. Embora a referência a profissão esteja um pouco associada a um certo modismo, não há dúvida de que este profissional tem um papel fundamental nos projetos de B&A.

- Estatístico e/ou Minerador de Dados: em projetos de inferência ou preditivos, o projeto deve contar com um estatístico ou minerador de dados. Um minerador de dados vai definir atributos relevantes para construir o modelo, treiná-lo e testá-lo, avaliar o seu desempenho, além de buscar os melhores algoritmos.
- DBA: já mencionamos que, muito embora o maior volume de dados existente no mundo (e o que mais cresce) são os dados não estruturados, as principais fontes de dados de projetos de B&A ainda são de dados estruturados, principalmente, banco de dados relacionais. Por isso, este tipo de profissional é importante também em projetos de B&A.
- Integrador e/ou Especialista em ETL: este profissional é também muito requisitado em sistemas de ingestão de grandes volumes de dados já em produção. O especialista em ETL estabelece arquiteturas de integração e ingestão de dados, modelos de transformação, rotinas de atualização e processos de monitoramento.
- Desenvolvedor: aqui, o termo refere-se a um desenvolvedor que codifica scripts com PL/SQL, funções de mapeamento e redução ou o desenvolvedor de software, uma vez que é comum projetos terem em seus escopos sistemas ou algoritmos customizados, aplicativos móbile ou ainda portais de acesso e gestão de usuários.
- Analista de Infraestrutura: instalar sistemas operacionais, preparar máquinas virtuais, configurar roteadores, nós, clusters. O analista de infraestrutura é muito importante em projetos de B&A, cuja arquitetura é horizontal e depende altamente de comunicação servidor para servidor. Deve participar do projeto desde o início, planejando a arquitetura ideal, configurando os ambientes de desenvolvimento, testes e homologação, e muitas vezes, ainda preparando o ambiente de produção e garantindo o seu ajuste ideal, para termos o melhor desempenho.

9. AQUISIÇÕES

Independente se o seu projeto será cloud, *on premise* ou híbrido, se partes da solução serão baseadas em software *open source*, grandes projetos, normalmente, requerem a aquisição de vários elementos: *storages*, servidores, software de ETL, produtos de visualização, de qualidade de dados, consultoria especializada, entre outros. O mais comum é que a sua solução não seja atendida por um único fornecedor. Fornecedores vão tentar todo o tipo de estratégia para vender o seu produto, prometendo funcionalidades que, às vezes, ainda estão na prancheta do arquiteto-chefe. Por isso, costumo dizer que o gerente de projetos e a sua equipe estão "sozinhos" na missão de selecionar as melhores aquisições possíveis dentro dos requisitos e orçamento do projeto.

Separamos o processo de seleção de fornecedores em seis etapas: identificar a necessidade, planejar e especificar, receber propostas e esclarecer dúvidas, avaliar e selecionar propostas, negociação, procedimento de compra e contrato. Claro que cada organização tem ainda seus próprios processos e trâmites. Mas antes de vermos as etapas, vamos falar da decisão de fazer ou comprar.

FAZER OU COMPRAR

É bastante comum em fóruns relacionados a *Big Data* encontrar um diagrama chamado de *Big Data Landscape*[4]. Ela exibe uma miríade de fornecedores de produtos e serviços de B&A, agrupados

4 Se você nunca viu esta imagem, faça uma rápida pesquisa em um mecanismo de busca por *Big Data Landscape*.

por assunto. É bem provável que tudo o que você precisar para o seu projeto, alguém já fez, pelo menos, de forma parcial ou similar.

A tendência primária em um projeto é procurar algo pronto, afinal, o risco é menor: já foi desenvolvido, e em muitos casos, testados por centenas ou milhares de usuários pelo mundo. Além disso, a compra, normalmente, vai ter um preço menor, pois os custos de desenvolvimento são rateados entre todos os clientes, e muitas vezes, já foram até pagos nos vários anos que a ferramenta pode estar no mercado. Até agora, estávamos falando de algo já pronto. Mas se não existe ainda no mercado nada parecido? Neste caso, também em algumas situações pode ser melhor comprar: se o fornecedor tem expertise no assunto, quando você não tem capacidade de implementar, quando você não tem tempo de montar uma equipe para desenvolver.

E quando é melhor fazer? Primeiro, se você não está prestando um serviço, mas criando um produto novo no mercado e tem preocupação com propriedade intelectual, é provável que a melhor opção seja fazer, para manter o controle sobre o processo (embora nada impeça de contratar o desenvolvimento, ou pelo menos parte dele). Também é interessante fazer quando você tem a mão de obra à disposição e ela está ociosa; ou quando o custo de fazer é muito menor; ou quando não existe nenhum fornecedor capaz de fazer o que você precisa.

Também é comum a contratação de serviços especializados: um consultor tributário, um especialista em recursos humanos, uma equipe de desenvolvimento de sistemas, entre outros. Em princípio, a lei permite a terceirização, desde que o serviço não esteja relacionado diretamente com as atividades-fim da empresa contratante.

Vamos ver, agora, as etapas recomendadas na compra de produtos ou contratação de serviços.

IDENTIFICAR A NECESSIDADE

Esta parece uma etapa lógica, porém, que apresenta certo risco. Primeiro, a necessidade pode ser superestimada ou subes-

timada. Por exemplo, um fornecedor de solução de *MapReduce* oferece o produto gratuitamente até 50 nós. A partir desta quantidade, os valores são bastante representativos e sobem exponencialmente. Se você superestimar o projeto, não avaliando opções, poderá estar comprando um produto que estará superestimado, com um custo elevado e subutilizado. Da mesma forma, a subestimativa pode levar a comprar algo que não atenda ao projeto, ou que rapidamente vai trazer ao produto problemas técnicos.

PLANEJAR E ESPECIFICAR

Identificada apropriadamente a necessidade, ela precisa ser especificada pela equipe técnica do projeto, para que os fornecedores possam compreender a demanda e oferecer uma solução compatível e sob medida. A falha em especificar, além de poder acabar em uma contratação inadequada, pode limitar o número de fornecedores. Em empresas públicas, o planejamento significa iniciar os trâmites legais, com montagem de edital e termos de referência técnica.

RECEBER PROPOSTAS E ESCLARECER DÚVIDAS

Especificado o que ser quer adquirir, é hora de solicitar e receber as propostas. Tanto em empresas privadas quanto em públicas, é comum receber representantes dos fornecedores para esclarecer dúvidas. As propostas devem ser mantidas em confidencialidade. Pode ocorrer ainda de aparecerem poucos interessados, ou mesmo fornecedores de destaque na área não se interessarem. Neste caso, é preciso rever as condições da contratação e as especificações. Um caso curioso em um projeto que estava atuando: não houve nenhuma proposta para fornecimento de serviços de data center em resposta a uma RFP. Revisando a especificação na RFP, constatava como pré-requisito data center Certificado TIER IV. Como este não era de fato um pré-requisito para o projeto, a RPF não foi enviada para nenhum fornecedor que tinha o data

center com esta certificação, que obviamente custaria muito além do orçamento do projeto.

AVALIAR E SELECIONAR PROPOSTAS

Recebidas as propostas e esclarecidas as dúvidas, é hora de avaliar as propostas. Claro que preço é importante, porém, diversos outros fatores devem ser analisados, dependendo do tipo de solução a ser adquirida: disponibilidade, escalabilidade, segurança, taxa de transferência, transferência de tecnologia, suporte técnico, treinamento, só para citar alguns.

Se possível, deve-se definir um critério de classificação, que deve envolver critérios técnicos e preços. Certames licitatórios, normalmente, apresentam uma fórmula composta por estes dois elementos, o fornecedor que obtiver o melhor score na aplicação da fórmula, é o vencedor do certame.

BENCHMARKING

Dentro do processo de selecionar propostas, é possível ainda realizar provas de conceito. Também é possível executar *benchmarking* com características técnicas fundamentais dos produtos, para verificar qual apresenta o melhor desempenho. Algumas organizações independentes realizam *benchmarks* em produtos de dados, como bancos de dados relacionais, de apoio à decisão e *Big Data*, e disponibilizam os resultados publicamente. Uma destas organizações é a TPC, que pode ser encontrada em www.tpc.org.

NEGOCIAÇÃO

Selecionada a melhor proposta, é hora da negociação. Claro que preço e condições de pagamento fazem parte de qualquer negociação, porém, diversos outros aspectos podem ser negociados. Por exemplo, se o produto não tem uma funcionalidade importante, como um conector nativo de SAP que não faz extrações

diferenciais, pode-se negociar com o fornecedor para que, se o contrato for estabelecido, esta funcionalidade seja incluída, por exemplo, em 180 dias. Também podem ser necessárias funcionalidades relativas a políticas da organização, como as condições de SLA, questões de propriedade intelectual, políticas de confidencialidade, entre outras.

PROCEDIMENTO DE COMPRA E CONTRATO

Encerrada a negociação, é hora de efetivá-la, executando os processos de compra que vão chegar à efetivação do contrato, que normalmente, já tem uma minuta conhecida.

FORNECEDORES DE "MILAGRES"

Um assunto que merece menção são as promessas de alguns fornecedores de soluções de *Big Data*. Se o leitor frequenta eventos sobre o tema ou tem o costume de ler material destes fornecedores, logo entenderá do que eu estou falando. Estes fornecedores oferecem soluções milagrosas, baratas, rápidas de implementar e com retorno de investimento garantido, verdadeiros milagres. Até aqui, nada anormal, é uma estratégia de venda da mesma forma. Um cirurgião plástico fala para a sua cliente o quão bela ela vai ficar depois do procedimento. Ele não vai mencionar o risco da anestesia, a dor e as complicações do pós-operatório, ou ainda que ela corre o risco de ficar com o rosto deformado depois de alguns anos.

Porém, os fornecedores de *Big Data* usam como comparativo com suas soluções milagrosas o *data warehouse* tradicional, que, segundo eles, é caro, difícil de implementar e de escalabilidade complexa, enquanto suas soluções são baratas, rápidas, de sucesso garantido e de implantação simples.

O custo e risco de um projeto é diretamente proporcional ao seu tamanho e complexidade. Projetos de *data warehouse* tradicionais podem custar caro, levarem anos para serem concluídos,

serem um sucesso absoluto ou fracassar totalmente. Mas o mesmo é válido para projeto de *Big Data*. Grandes projetos de *Big Data* são caros, de alto risco e requerem um longo período de implementação, com um agravante: enquanto projetos e fornecedores tradicionais de *data warehouse* têm a sua tecnologia e metodologia já relativamente maduros, pois estão por aí há quase 30 anos, alguns fornecedores de *Big Data* estão no mercado há menos de 5 anos.

Outro erro cometido é sugerir que o *data warehouse* clássico acabou, e que só se implementa análise de dados com soluções de *MapReduce*. A criação de modelos e sistemas fora do mundo relacional/dimensional, como bancos de dados NoSQL ou *Hadoop*, representa a resposta para problemas de dados que antes não existiam na proporção em que existem hoje. Para novos problemas, foram criadas novas soluções. Um banco de dados relacional ainda é a melhor solução para sistemas que requerem integridade e controle de transação. O *data warehouse* tradicional ainda é a melhor solução para dados de qualidade e estruturados. É preciso, em qualquer processo de seleção de tecnologia ou fornecedor, muita cautela, e sempre buscar uma solução para um problema de negócio e não a implementação de uma tecnologia. Às vezes, a solução para o seu problema pode já existir dentro de casa.

10. *BIG DATA* ÁGIL

Neste capítulo, vamos falar como metodologias Ágeis podem ser aplicadas em projetos de B&A. Também vamos ver quando projetos Ágeis não são a melhor escolha, e ainda que benefícios projetos de B&A Ágeis podem trazer ao cliente e aos usuários do projeto.

ÁGIL VERSUS TRADICIONAL

Nesta seção, vamos rapidamente recordar as principais diferenças entre gerenciar um projeto de forma tradicional (ou *Waterfall*) e Ágil. A principal diferença é que enquanto na gestão tradicional faz-se um grande esforço de planejamento no início do projeto e se procura seguir este plano até o final, na metodologia Ágil ocorrem pequenas interações também conhecidas como *sprints*, onde são feitas entregas menores e de forma mais frequente. Ao final do ciclo, o projeto é reavaliado e a equipe decide o que será feito na próxima interação. Desta forma, é entregue valor ao cliente mais rapidamente e com mais frequência.

VANTAGENS DE PROJETOS ÁGEIS EM *BIG DATA*

As vantagens de se desenvolver B&A com metodologia Ágil são as mesmas que encontramos em projetos de desenvolvimento de software. Supondo que o seu projeto vai coletar dados de sensores e produzir vinte relatórios de controle estatístico de processo (CEP) para toda a linha de produção de uma grande

indústria que opera 24 horas. Como cada equipamento é de um fornecedor diferente, cada relatório (Carta de controle) é um processo de construção à parte, tanto pela estrutura dos dados como pelas informações (quais sensores) devem compor o relatório. Ou seja, para cada equipamento é preciso estudar os modelos de dados em busca da informação que são relevantes para os objetivos do relatório, criar conexões, extratores, procedimentos de transformação, rotinas de análise estatística, e finalmente, os relatórios.

Agora, imagine o seguinte: você escreve o escopo dos vinte relatórios, produz mais alguns documentos (por exemplo, um protótipo de cada um em uma ferramenta de criação de diagramas), apresenta um cronograma e um orçamento. O cliente aprova, você inicia a produção de todos os vinte relatórios, o que dura alguns meses. Mesmo este tempo não sendo surpresa, pois o cliente recebeu um cronograma, ele fica impaciente, afinal, o tempo está passando, o orçamento está indo e ele não viu nada de concreto, além disso, ele tem medo de que o principal e mais caro equipamento da linha de produção pare por falta de controle. Ao final do prazo, você apresenta para ele todos os relatórios prontos. Por mais que previamente o cliente tenha aprovado os modelos de relatórios, só vendo os mesmos sendo gerados em produção é que ele se dá conta de que algumas coisas tinham que ser diferentes. O cliente pede algumas correções e alterações, bem como a inclusão de alguns novos atributos. Você inicia o processo integrado de mudança, faz todas as alterações na documentação, apresenta as alterações no orçamento, o cliente aprova, você altera os relatórios e apresenta o resultado, e assim, o ciclo ocorre mais algumas vezes até o cliente ficar satisfeito e o projeto ser encerrado com sucesso.

Agora, imaginemos outra situação: você conversa com o cliente e rapidamente esboça o primeiro relatório do equipamento mais importante em um papel. Desenvolve em um curto espaço de tempo e apresenta ao cliente. Ele pede algumas alterações, você altera, ele acha que está bom e considera o relatório pronto. Você está pronto para a próxima iteração. O cliente muda as prioridades e pede que, o que era para ser o último relatório, agora, seja o

próximo. Você parte para o novo desenvolvimento, mas o cliente já está satisfeito e motivado, pois já pode ler a carta de controle do equipamento mais importante. Baseado no feedback do primeiro relatório, os outros são desenvolvidos mais rapidamente. Aos poucos, todos os relatórios são entregues.

No segundo caso, temos o cliente satisfeito por entregas contínuas: relatórios são entregues com frequência. Ele enxerga valor no projeto de forma mais rápida, tem sua confiança mantida. Também não há aquele "mal-estar" causado pelas mudanças de escopo, que existe nos processos de mudança na metodologia tradicional.

Mas será que sempre é possível entregar valor desde cedo e sempre? Depende. A entrega de valor pode ser relativa. Vamos imaginar a construção de um *data warehouse*. O cliente vai enxergar valor quando ele puder fazer um *drill down* no cubo em seu notebook. Porém, até se chegar ao cubo, pode ter se passado meses de projeto: quando as conexões foram criadas e os dados carregados em *staging*, quando os scripts forem desenvolvidos, quando as tabelas relacionais foram transformadas em fatos e dimensões, quando as expressões DAX foram escritas. Em todas estas etapas, já se estava entregando valor ao projeto, embora para o cliente, estas entregas sejam intangíveis.

QUANDO NÃO USAR ÁGIL

Nos últimos anos, atuei mais prestando serviços em projetos analíticos a terceiros, ou seja, não foram projetos implementados dentro de minha própria empresa. Nestes casos, normalmente, temos que nos adequar à metodologia utilizada pelo cliente. Neste aspecto, ou seja, atendendo exigências contratuais, as práticas clássicas têm sido utilizadas na grande maioria dos projetos. Também alguns projetos foram prestados para empresas públicas. Estas empresas tem uma grande preocupação com órgãos de fiscalização como tribunais de contas. Por isso, as contratações e a metodologia de trabalho são mais rígidas e tradicionais, exigindo

uma definição detalhada e total do escopo e dos custos antes mesmo do projeto começar.

Nas situações do parágrafo anterior, uma metodologia tradicional é usada por uma exigência, seja ela de governança corporativa ou por questões de conformidade. Mas quando o projeto é na própria organização ou o cliente possui cultura de gestão Ágil de projetos, devemos sempre usar esta metodologia? A resposta é, em geral sim, porém, em algumas situações, seria recomendado usar processos clássicos de gerência de projetos ao invés de Ágeis. Vamos elencar algumas delas:

- Se o cliente ou questões de conformidade não permitem, como no exemplo já mencionado das empresas públicas.
- Se os requisitos do projeto precisam estar definidos já no início.
- Se o patrocinador precisa entender o projeto todo antes de aprovar o orçamento. Não temos a chance de evangelizar todos dos benefícios de usar uma metodologia Ágil em projetos. Algumas pessoas precisam entender o projeto todo, o que será entregue ao final, como condição para aprovar o orçamento.
- Quando o projeto é muito grande, e a equipe de projeto, aqueles envolvidos diretamente no desenvolvimento são muitos. Não posso dizer um número exato, pois vários fatores influem, mas me parece que mais de 15 pessoas já torna o projeto grande.
- Se grande parte da sua equipe está dispersa em várias regiões ou países. A metodologia Ágil funciona bem em equipes pequenas, alocadas em um mesmo ambiente físico.

11. ESTABELECENDO UM PROCESSO

Neste capítulo, é apresentada uma série de etapas que devem ser observadas na gerência de projetos de B&A. Algumas destas etapas já foram discutidas durante a obra, mas aqui são apresentadas de forma estruturada e consolidada. Antes de entrarmos nas etapas, vamos primeiro a algumas observações que vão ajudá-lo a compreendê-las melhor.

Embora as etapas estejam dispostas em uma sequência lógica, não significa que etapas não possam ser executadas em paralelos. Em outras palavras, em geral, a etapa seguinte pode ser executada enquanto a etapa anterior ainda não foi encerrada, embora em alguns casos, seja recomendado que a etapa seja encerrada antes de prosseguir. Por exemplo, seria prudente termos convicção de que o projeto é viável tecnicamente, antes de começar o desenvolvimento. Porém, enquanto o desenvolvimento ocorre, já poderão estar executados testes.

Estas etapas dizem respeito ao desenvolvimento do produto ou serviço de B&A, e não a etapas de gerência de projetos, que deverão ser seguidas conforme a metodologia de gestão de projetos que você adotar (ex: PMBOK, *Scrum* etc).

Vamos, então, ver as etapas:

- *Business case*: discutido no capítulo 2. O projeto deve ter uma justificativa, atender a um requisito de negócio e entregar valor para a organização. Como falamos, normalmente, o gerente de projetos vai receber o *Business Case* pronto.
- Entrega de Valor: mesmo que o *Business Case* tenha uma seção que trata da entrega de valor do projeto, o gerente de

projeto deve garantir que, de fato, o empreendimento vai entregar valor. Ao longo da obra, falamos de vários projetos que não entregaram valor à organização: o diretor criou algo para benefício próprio, ou o objeto do projeto já não mais fazia sentido, ou ainda alguém queria testar alguma tecnologia de última geração.

- Viabilidade Técnica: é um grande risco o projeto parar por questões técnicas: a fonte de dados é inacessível, o *stream* de dados é muito elevado, o projeto é viável, mas vai custar muito mais do que o valor empenhado para o mesmo. A viabilidade técnica pode ser analisada de diversas formas, um simples estudo de um consultor técnico, ou com implementação de provas de conceito ou protótipos.
- Projeto: o projeto é o planejamento do que será construído, bem como as rotinas de monitoramento e controle do seu andamento. Inclui atividades como definir escopo, criar cronogramas, WBS, entre outros.
- Desenvolvimento: o desenvolvimento é a fase mais longa do projeto. Nessa etapa, os artefatos serão desenvolvidos.
- Testes: Trata das atividades de garantia de qualidade, discutidas durante a obra.
- Entrega e Aceite: Baseada na definição de pronto e do escopo do projeto, a entrega é a aceitação formal do que foi desenvolvido.
- Implantação: Projetos de B&A, normalmente, são desenvolvidos em um ambiente de desenvolvimento. Conforme partes são entregues, são instaladas em um ambiente de homologação, onde passam por testes mais especializados. Nesses ambientes, os conectores estão apontados para bases de dados não oficiais, que podem ser cópias dos dados de produção com alguns dias de atraso. Provavelmente, em seu projeto estarão incluídas atividades de implantação, que é a instalação de todos os artefatos do projeto em seu ambiente definitivo, que é a passagem do ambiente de homo-

logação para a produção. O ambiente de produção é mais robusto e mais restrito do que os demais ambientes.

- Transição: A transição é a passagem do projeto para a equipe que vai mantê-lo rodando. Normalmente, membros da equipe que desenvolveram continuam a mantê-lo, cuidando também das questões evolutivas.

Na Figura-3, podemos ver estas etapas na forma de um fluxograma.

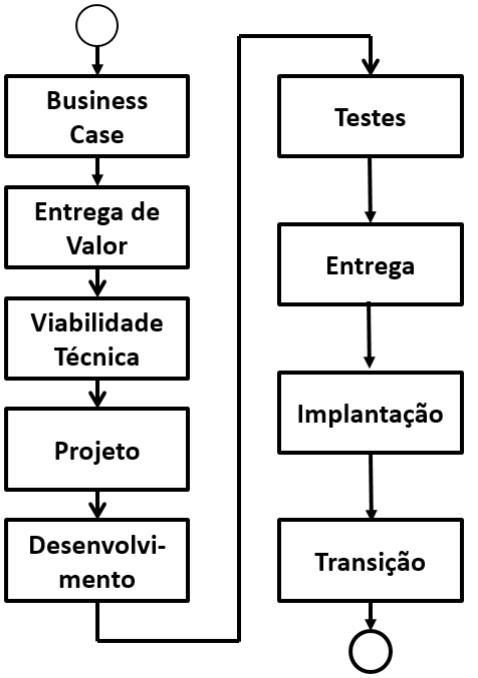

Figura 3: O processo de implementação de B&A.

O QUE UM PROJETO DE *BIG DATA* ENTREGA?

Para a maioria dos envolvidos, projetos de B&A entregam apenas o resultado final, sejam eles painéis, relatórios, processos

de integração de dados ou arquivos de conformidade. Porém, na verdade, a maioria dos projetos entrega um grande volume de artefatos. Classificamos estes artefatos em quatro grupos: documentos, treinamentos, solução e ambientes, conforme Figura-4.

Figura 4: Entregas de um projeto de B&A.

Vamos ver primeiro, os documentos. O tipo de documentação que o projeto vai gerar depende muito do tipo do projeto, mas é fato que a produção de certos documentos é parte do escopo do projeto e deve ser considerado nos custos e tempo de desenvolvimento. Muito embora devamos evitar criar documentação desnecessária, que na prática, não terá qualquer utilidade, devemos nos lembrar de que projetos de B&A são complexos por essência, e que, amanhã, talvez, as pessoas que o desenvolveram ou o mantêm operando, não estarão mais na organização. Sem um mínimo de documentação de configurações e estruturação da solução, ela corre o risco de simplesmente parar.

Podemos separar os documentos em dois grupos: os documentos que auxiliam o desenvolvimento do projeto e os documentos que são utilizados para manter a solução funcionando, que chamamos de documentos de operação, os quais também são desenvolvidos no projeto e fazem parte deste.

Como documentos do projeto, temos *use cases*, plano de testes, casos de testes, interface com usuários, maquetes de tela, entre outros.

Como documentos de operação, temos requisitos de segurança e privacidade, manutenção preventiva, administração do sistema, políticas de backup, arquitetura física e lógica e processos de contingência.

Entre os treinamentos, o ideal é que eles sejam desenvolvidos conforme o público do projeto. Geralmente, eles são três:

- Usuário final: é o treinamento para aqueles que vão consumir o resultado do projeto. Pode parecer um treinamento simples, mas nem sempre. Por exemplo, se o projeto construiu um *data warehouse,* que o usuário final irá consumir por ferramentas de BI self-service, ele deverá ter treinamento prévio no produto.
- Administração e manutenção: treinamento para a equipe que vai manter o sistema em operação após a sua transição para produção. Deve conter rotinas de manutenção preventiva, corretiva, configurações e de aumento de capacidade, para quando a solução precisar "escalar"
- *Help Desk*: treinamento para a equipe que vai dar o suporte ao usuário da solução. Dependendo do tipo do projeto, pode ser necessário diferentes tipos de treinamento, conforme o nível (1, 2 ou 3) do atendimento.

A solução a ser entregue é o objetivo principal do projeto. Corresponde a extratores, scripts, códigos de mapeamento e redução, arquivos de configuração, entre outros. Estas entregas permitem, por exemplo, que o sistema seja replicado e instalado

em outro ambiente, bem como que a manutenção corretiva e evolutiva seja mantida.

Finalmente, o quarto tipo de entrega do projeto são os *deploys*, ou instalações nos mais diversos ambientes que a solução terá durante o seu desenvolvimento. O mais comum é termos quatro ambientes:

- Desenvolvimento: é o ambiente onde os desenvolvedores criam script, configuram extratores, codificam mapeamentos e reduções etc. Nesse ambiente são feitos os testes unitários. Teoricamente, é único ambiente em que a solução pode ser alterada, com exceção é claro de configurações.
- Testes: este ambiente recebe atualizações periódicas do ambiente de desenvolvimento para testes. Os tipos de testes incluem a execução de casos de testes e testes de regressão.
- Homologação: aqui, a solução é implantada quando já se tem uma entrega concreta, que já passou pelas etapas de testes e teoricamente está estável. Sendo aprovada neste ambiente, a solução está pronta para o ambiente de produção. Algumas empresas fazem do ambiente de homologação uma cópia fiel do ambiente de produção, a fim de reproduzir com a maior fidelidade possível as variáveis que serão encontradas em produção. Como o ambiente de produção, em geral, é um ambiente muito mais robusto, essa opção só é viável em projetos de soluções bastante críticas.
- Produção: nesse ambiente, ocorre a entrega da solução para consumo pelo usuário final.

AVALIANDO E MONITORANDO A SOLUÇÃO

Uma vez implantada, a solução precisa ser monitorada diariamente. Existem diversos produtos de monitoramento de performance nativos de sistemas de produção de informação, ou um produto deste tipo pode ser incluído no orçamento de seu projeto.

Também é necessário avaliar a performance da solução. É natural que nos primeiros meses de produção tenham que ser fei-

tos ajustes na arquitetura para remover gargalos e correções de erros. Porém, é preciso que todos os problemas sejam monitorados e documentados, e que se possa acompanhar a evolução da performance e estabilidade da solução.

Algumas métricas que podem ser acompanhadas são:

- Tempo de indisponibilidade: mede o tempo, em minutos, que a solução ficou indisponível em certo período (por exemplo, em um mês). Deve-se ter um entendimento de qual parte da solução é considerada. Por exemplo, pode haver um processo de carga para *staging* noturno, com um processamento pela manhã. Uma falha no processo de *staging* ocorre, mas é corrigida antes do processamento. Este não é considerado tempo de indisponibilidade, já que o produto final não foi afetado.
- Tempo de carga ou transformação: é normal em uma solução que o tempo de carga oscile ou se degrade ao longo do tempo. Os motivos podem ser vários, questões de arquitetura (mau planejamento de índices), de aumento de volume de dados, ou de infraestrutura, principalmente, de banda de rede.
- Tempo de resposta em processamento: uma vez carregados os dados, haverá um tempo de processamento, etapa onde dados são transformados em informação e o valor é gerado. Embora menos comum, pode nesta etapa também ocorrer degradação do tempo de processamento, além de eventuais *bugs* que não foram detectados em ambientes de testes (datas inválidas, divisão por zero, falta de memória etc.).
- Número de paradas: esta métrica deve medir quantas vezes a solução ficou indisponível para o usuário. Tem relação com a métrica de tempo e de número de erros, pois uma parada pode levar apenas alguns minutos como pode levar horas. De outra forma, o sistema pode não ter nenhuma parada, mas os seus dados podem estar inconsistentes devido a um erro em uma medida.

- Número de erros apontados: mede o número de erros (*bugs*) da solução. Tem relação com o número de erros corrigidos.
- Número de erros corrigidos: mede o número de erros que, uma vez apontados, são corrigidos e atualizados na solução em produção.

12. CHECKLIST

Este último capítulo traz de bônus um checklist de projetos de B&A. Como todo checklist, seu uso é simples, basta passar pelos pontos e verificar se o item já está em conformidade, ou mesmo, se não se aplica ao projeto. Periodicamente, a lista deve ser revisada. Para um controle mais efetivo, é interessante colocar prazos e responsabilidades na lista.

DIVERSOS

Item	Atendido	Não se aplica	Responsável	Prazo
Os critérios de sucesso do projeto estão claros?				
O projeto entrega valor?				
O projeto está em conformidade com as normas de governança de dados da organização?				
As tecnologias a serem utilizadas são maduras e estáveis?				

ARQUITETURA

Item	Atendido	Não se aplica	Responsável	Prazo
A infraestrutura necessária ao projeto é conhecida?				
Os requisitos de *failover* foram especificados?				
Os requisitos de *clusters* foram identificados?				
A capacidade de armazenamento é conhecida?				
Existe uma estimativa inicial de *nodos* necessários ao projeto?				
As taxas de transferência necessárias foram estimadas?				
A infraestrutura planejada suporta as taxas de transferência estimadas?				
O projeto do sistema prevê o seu crescimento?				
Os requisitos de disponibilidade foram planejados?				
Quais partes da solução serão *on-primeses* e *on cloud?*				
Os requisitos de virtualização foram identificados?				
Os processos e equipamentos de backup estão definidos?				
O desenvolvimento dos manuais de implantação e manutenção da solução está planejado?				

REQUISITOS TÉCNICOS

Item	Atendido	Não se aplica	Responsável	Prazo
Todas as fontes de dados do projeto foram identificadas?				
É possível acessar todas as fontes de dados que devem fazer parte do projeto?				
As fontes de dados possuem documentação apropriada?				
Os requisitos para a ferramenta de extração de dados estão especificados?				
A latência de produção de informação é suficiente para produzir o valor necessário?				
Os requisitos para a ferramenta de processamento de dados estão especificados?				
Os requisitos para a ferramenta de B&A estão especificados?				
A ferramenta de visualização atende aos requisitos do projeto?				
Será necessária a visualização em dispositivos móveis?				
As ferramentas suportam os requisitos de segurança e privacidade?				

REQUISITOS FUNCIONAIS

Item	Atendido	Não se aplica	Responsável	Prazo
As regras de negócios são conhecidas e existe um consenso na incorporação delas?				
Os critérios de segurança são conhecidos?				
Os critérios de privacidade são conhecidos?				
Os requisitos de cultura e idioma da solução estão claros?				

AQUISIÇÕES

Item	Atendido	Não se aplica	Responsável	Prazo
Todas as necessidades de aquisições são conhecidas?				
O orçamento é adequado para as aquisições necessárias?				
O tempo para entregas pelos fornecedores é conhecido e atende aos requisitos do projeto?				
As especificações do que será adquirido estão claras?				
O processo de aquisição da empresa é conhecido?				
Os critérios de seleção de fornecedores estão estabelecidos?				
Existe um comitê para seleção de ofertas?				
A empresa está preparada para responder questões aos fornecedores?				

QUALIDADE

Item	Atendido	Não se aplica	Responsável	Prazo
Existe um plano para avaliar a qualidade dos dados?				
Existe um plano para limpar dados, se necessário?				
Existe um plano de testes?				
Os requisitos de qualidade do projeto são conhecidos, consensuais e factíveis?				

CUSTOS

Item	Atendido	Não se aplica	Responsável	Prazo
O orçamento do projeto é conhecido?				
O orçamento do projeto é realista?				
Foi planejado orçamento para contingências?				
Existe um ROI do projeto e ele é realista?				
Os custos diretos e indiretos do projeto são conhecidos?				

RECURSOS HUMANOS E PARTES INTERESSADAS

Item	Atendido	Não se aplica	Responsável	Prazo
Todas as competências necessárias para o projeto foram identificadas?				
Existem no mercado profissionais com as competências identificadas?				
Existe um plano de capacitação para usuários finais?				
Existe um plano de capacitação para administradores e operadores?				
Existe um plano de capacitação para *help desk*?				
Existe um plano de evangelização de partes interessadas?				
Todas as partes interessadas foram identificadas?				

SUPORTE AOS USUÁRIOS

Item	Atendido	Não se aplica	Responsável	Prazo
Os canais de suporte aos usuários foram estabelecidos?				
A SLA do produto está em acordo com as políticas da organização?				
A infraestrutura de operação do suporte está especificada?				
O sistema de controle de chamados já foi adquirido ou existe na organização?				

PRODUÇÃO E MONITORAMENTO

Item	Atendido	Não se aplica	Responsável	Prazo
O processo de transição de desenvolvimento para produção está claro?				
Estão definidas as ferramentas e técnicas de avaliação de performance em produção?				
Existem ferramentas para monitorar a solução em produção?				

GLOSSÁRIO

A

Ágil: um conjunto de metodologias de desenvolvimento de software, que busca entregar valor em iterações, ou ciclos, mais rápidos. Ao final de cada iteração, pode-se avaliar o que foi construído e decidir quais serão as próximas etapas. As metodologias ágeis são baseadas em 4 valores:

Indivíduos e interações mais que processos e ferramentas;
Software funcional mais que documentação abrangente;
Colaboração do cliente mais que negociação de contratos;
Responder a mudanças mais que seguir um plano.

B

BI Self-Service: são ferramentas de BI, permitem ao usuário a produção de informação de apoio à tomada de decisão, mesmo que estes não conheçam técnicas de análise de dados. Por exemplo, se o gerente quer saber quantas vendas de produto X foram realizadas em determinado período. Em uma ferramenta de BI tradicional, este gerente teria que construir uma consulta SQL e executá-la contra um banco de dados. Em uma ferramenta self-service, o gerente pode perguntar, usando a escrita ou a fala, usando linguagem natural. Internamente, a ferramenta irá construir a consulta, executá-la e apresentar o resultado.

Brainstorming: dinâmica de grupo que reúne um grupo de pessoas para buscar a solução a algum problema. Durante a reunião, todos devem opinar. A ideia é reunir o maior número possível de sugestões sobre o assunto. O objetivo da reunião é buscar um consenso sobre a solução do problema-tema da reunião.

Business case: o caso de negócio é um documento que deve apresentar as justificativas para um empreendimento. Deve mostrar os benefícios esperados, sua entrega de valor, bem como seus custos e riscos. Deve mostrar ainda o que ocorrerá se o empreendimento não for feito.

C

CEP: acrônimo para controle estatístico de processos, é uma ferramenta de qualidade usada em produção. O objetivo é gerar informação para detectar defeitos, prevenindo paradas ou inoperâncias no processo produtivo, consequentemente, colaborando com a melhor produtividade.

Certificado TIER IV: TIER é um conjunto de certificações para data centers, mantido pelo *Uptime Institute*. O nível IV é o mais sofisticado. Nesse nível, entre as várias exigências, uma delas é a de disponibilidade em 99,995%.

Cloud Computing: uso de recursos computacionais, como processamento e armazenamento, em computadores acessados através da internet. O acesso a programas instalados nestes computadores pode ser feito de qualquer lugar do mundo. Usa-se a computação na nuvem, principalmente, como uma forma de redução de custos, uma vez os recursos são providos por empresas especializadas, que provêm os serviços computacionais em grande escala, a centenas ou milhares de usuários.

Commodity: aqui o termo refere-se a computadores com características de processamento e armazenamento comuns e padronizadas. Uma vez atendendo a estas características, estes equipamentos podem ser adquiridos de qualquer fornecedor, normalmente, sendo então, o preço o fator decisivo para a compra.

CRM: *Customer Relationship Management*, ou Gestão de Relacionamento com o Cliente, é um sistema que cuida de toda a interação da empresa com o cliente, normalmente, desde o estágio de prospecção até o pós-vendas.

D

Data broker: empresa que vende dados consolidados, normalmente, dados de consumidores, divididos por segmentos. Na maioria das vezes, estes dados são utilizados para campanhas de marketing.

Data Lake: é um depósito de dados, onde estes ficam em seu formato original, sem tratamento, até terem uma utilidade. Projetos de análise de dados tradicionais armazenam dados tratados e com valor. Projetos de *Big Data* tendem a armazenar todo o dado coletado, e à medida que são necessários, são tratados e carregados no *data warehouse*.

Data mart: um armazém de dados departamental para apoio à tomada de decisão. Normalmente, o *data mart* armazena informação tratada e de qualidade.

Data nodes: em uma solução *MapReduce* como o *Hadoop*, *data nodes* são computadores, normalmente, máquinas virtuais, que armazenam dados em um sistema de arquivo distribuído HDFS.

Data warehouse: um armazém de dados corporativo, formado por vários armazéns de dados departamentais (ver *data mart*).

DAX: acrônimo para *Data Analysis Expression*, é um conjunto de funções para *Microsoft Analysis*, *Excel Power Pivot* e *Services e Power* BI.

Delphi: dinâmica de grupo para encontrar uma solução ou melhor alternativa para um problema. Sua principal característica é que a identidade de quem emite a opinião é mantida em sigilo. Através de várias interações, busca o consenso sobre o assunto.

Drill down: significa aprofundar uma análise em busca de mais detalhes. Normalmente, a operação é feita com cliques de mouse ou toques em uma ferramenta de visualização de dados. Por exemplo, um gráfico pode mostrar a venda por Estado. Um clique sobre o Estado pode trazer um detalhamento das vendas naquele Estado, agora, por cidade.

E

ETL: acrônimo para *Extract, Trasform and Load*, ou extrair, transformar e carregar. Este é o processo fundamental na construção de um *data mart*, quando as informações devem ser extraídas de uma fonte de dados de origem, transformadas, limpas, e finalmente, carregadas em seu repositório de destino onde serão consumidas.

F

Failover: uma vez que um equipamento falhe, seja ele um servidor ou componente de rede, *failover* é a troca automática para um equi-

pamento redundante, preparado e disponível para assumir em caso de falha.

G

Gold Plating: entrega supérflua, é qualquer funcionalidade a mais adicionada a um projeto, sem que a mesma esteja acordada entre as partes durante o planejamento do projeto. A gestão de projetos clássica classifica a entrega supérflua como uma prática não recomendada de gestão de projetos.

Granulidade: representa o nível de detalhe de uma informação em um armazém de dados analítico. Por exemplo, com menor granulidade, a informação de vendas pode mostrar vendas por cidade. Já com grão maior, as vendas podem ser mostradas por Estado.

H

Hadoop: sistema de análise de dados distribuído, mantido pela Fundação Apache e baseado no modelo *MapReduce*.

HDFS: acrônimo *para Hadoop Distributed File System*, é um sistema de gerenciamento de arquivos distribuídos.

K

KPI: acrônimo para *Key Performance Indicator*, ou indicador-chave de performance, é uma ferramenta visual que permite avaliar o desempenho de uma meta, dentro de limites de performance previamente estabelecidos. O KPI ainda permite que o usuário veja as tendências futuras, bem como a performance passada do indicador.

L

LDAP: acrônimo para *Lightweight Directory Access Protocol*, é um protocolo para manter e compartilhar informações sobre usuários, redes e sistemas, usando protocolo IP.

M

MapReduce: modelo de processamento de grandes volumes de dados em paralelo, usando computação distribuída em um *cluster*. É for-

mado por duas partes principais: o mapeamento, que descobre e processa os dados, e a redução, que sumariza os resultados.

MDM: acrônimo para *Master Data Managment*, é o conjunto de processos e ferramentas para gestão de dados-mestre. Seu objetivo principal é que a organização possua uma única versão oficial de dados importantes para a organização, como por exemplo, dados dos clientes.

Medidas: em um processo de análise de dados, a medida é um dado consolidado, numérico, que foi processado através de uma simples operação aritmética ou de uma fórmula mais complexa, como por exemplo, uma expressão DAX.

Metodologia Ágil: ver Ágil.

Modelos: um modelo é uma referência, criada por um algoritmo e um conjunto de dados históricos, que um processo de aprendizado de máquina usa para classificar um novo elemento.

N

Name nodes: em uma solução *MapReduce* como o *Hadoop*, *name nodes* são os *nodos* de redes responsáveis por manter metadados.

NoSQL: termo genérico para se referir a um grupo de tecnologias de banco de dados não relacionais. Existem quatro tipos de bancos de dados NoSQL: documentais, chave-valor, orientados à coluna e de grafos.

O

On premise: o termo refere-se a uma solução instalada e rodando nas próprias instalações da empresa, no oposto a uma solução *on cloud*, que pode estar fisicamente em qualquer parte do mundo.

P

Planing Pokker: método de estimativa de esforço para atividades, em que se usa um baralho. Os envolvidos apresentam uma carta com a estimativa ao mesmo tempo, dessa forma, não há influência na estimativa entre os membros da equipe. Após a estimativa inicial, a equipe pode discutir os motivos que a levou a tal estimativa, podendo-se chegar a um consenso.

PMBOK: acrônimo para *Project Management Body of Knowledge*, é o guia de referência para gestão de projetos editado e mantido pelo PMI, *Project Managment Institute*.

PMI: *Project Managment Institute* é uma instituição americana e sem fins lucrativos, que mantém certificações e metodologias em gestão de projetos. Mantém também o PMBOK, guia de referência para a gestão de projetos.

Profiler: ferramenta de monitoramento de sistemas de banco de dados. Pode ser utilizada para auditar o uso, encontrar *bugs* ou inspecionar processos ou consultas que são executados contra o sistema.

R

Raid 6: *Raid* é acrônimo para *Redundant Array of Independent Disks*, ou conjunto redundante de discos independentes. É um conjunto de discos de armazenamentos de dados que operam em grupo, buscando desempenho ou segurança. O *Raid* 6 é um conjunto de discos redundantes, onde são mantidas cópias dos dados, para o caso de falha de algum disco.

RFP: acrônimo para *Request For Proposal*, é a requisição formal, através de um documento de especificação, para que fornecedores de um determinado produto ou serviço apresentem propostas para fornecê-lo.

ROI: acrônimo para *Return Of Investiment*, ou retorno de investimento, mostra a relação entre o que foi investido no projeto e o valor ganho (ou perdido) com este.

RTO: acrônimo para *Recovery Time Objective*, ou objetivo de tempo de retorno, é o tempo esperado entre a ocorrência de um problema de parada de um sistema e sua volta à operação.

S

Sarbanes-Oxley Act: a A Lei *Sarbanes-Oxley*, de 2002, é uma lei americana que estabelece uma série de controles e procedimentos a empresas, com ações negociadas na NYSE, buscando a redução de eventos de fraude e aumentando a segurança para investidores.

Scrum: é um *framework* para gerenciamento de Ágil de projetos de software**.**

Shadow Systems: sistemas sombra, são sistemas não oficiais em uma empresa, normalmente, instalados em servidores departamentais ou estações de trabalho, que buscam oferecer funcionalidades não fornecidas por sistemas corporativos "oficiais".

Silos de Informação: são repositórios de informação isolados dentro de uma empresa, normalmente, a nível departamental. Podem existir na forma de *data marts*, que são formais no departamento, mas não possuem integração ou comunicação com armazéns de dados corporativos. Muitas fontes de silos de informação são sistemas sombra.

SLA: do inglês *Service Level Agreement*, ou Acordo de Nível de Serviço, é um contrato entre um prestador e um tomador de serviços, que estabelece as diretivas básicas de como este serviço deve ser prestado. Uma SLA deve conter, por exemplo, tempo máximo entre o registro de um incidente e o início do atendimento.

Spark: sistema *open source*, mantido pela Fundação Apache, de análise de dados de memória.

Sprints: em metodologia Ágil, um *sprint* é uma iteração em que valor é agregado ao produto que está sendo desenvolvido. O tempo de uma iteração pode variar de uma semana a um mês.

Staging: em um processo de ETL, *staging* é um estado intermediário do dado, depois de extraído da origem, mas antes de ser arquivado no banco de dados analítico de destino.

Storage: termo genérico que se refere a produto, geralmente, formado por software mais hardware, destinado ao armazenamento de dados.

T

TCO: do inglês *Total Cost of Ownership*, ou custo total de aquisição, refere-se ao custo total de adoção de um produto ou sistema, contemplando custos diretos e indiretos.

V

Virtualização: é a simulação de uma plataforma, que pode ser hardware, sistema operacional e rede. Separa os componentes lógicos (softwares e sistemas operacionais) do meio físico (hardware). Dessa for-

ma, um mesmo equipamento pode virtualizar muitos ambientes lógicos, que podem ser gerenciados em conjunto.

VME: Valor Monetário Esperado é uma análise matemática que inclui a incerteza de um processo em função dos riscos associados ao mesmo.

W

Waterfall: na metodologia de desenvolvimento de software clássica, *waterfall* é o modelo em que o processo ocorre em etapas em que, em princípio, devem ser totalmente finalizadas antes de se iniciar a próxima. Por exemplo, o desenvolvimento só ocorre depois que todo o planejamento estiver concluído. Mudanças são possíveis, mas devem ocorrer como um novo processo dentro das atividades de criação do produto ou serviço.

WBS: acrônimo para *Work Breakdown Structure* ou Estrutura Analítica de Projetos, é um diagrama hierárquico que deve mostrar toda a atividade necessária para desenvolver um produto ou serviço. É uma técnica de gerenciamento de escopo de projetos, muito embora o WBS seja também usado para estimar custos, tempo e até riscos de um projeto.

REFERÊNCIAS

A Guide to the Project Management Body of Knowledge: (PMBOK® Guide). Newtown Square, PA: PMI, 2013.

English, Larry P. *Information Quality Applied*: Best Practices for Improving Business Information, Processes, and Systems. Indianapolis, IN: Wiley, 2009.

"Manifesto for Agile Software Development." *Manifesto for Agile Software Development*. N.p., n.d. Web. 29 May 2016.

ESTA OBRA FOI IMPRESSA PELA GRÁFICA
POLOPRINTER, EM SÃO PAULO, NO INVERNO DE
2016. COM SEU MIOLO IMPRESSO NO PAPEL OFF SET
75 GRAMAS E SUA CAPA IMPRESSA EM SUPREMO
ALTO ALVURA 250 GRAMAS. A TIPOLOGIA UTILIZADA
FOI GARAMOND E BODONI BT.