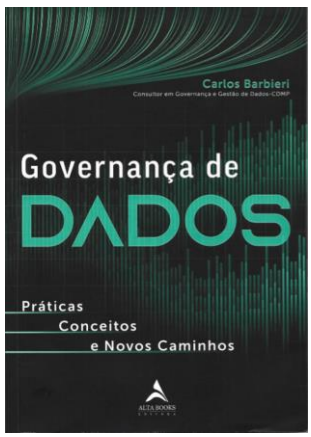
Governança de Dados-Práticas-Conceitos e Novos caminhos

GD 2.0

Carlos Barbieri

Parte-10-GD Ofensiva e DataOps



GD 2.0-Governança de dados

Parte 10-GD 2.0- GD Ofensiva e DataOps

1-Introdução:

Um dos pontos a se observar na Informática, seja ela dos anos 1980 ou dos anos 2020, é que os movimentos que surgem vêm sempre alinhavados com conceitos já existentes, alguns preservados na íntegra, outros claramente melhorados e acoplados com os novos insights na forma de fazer a computação rodar. Isso traz grande vantagem no entendimento das tecnologias fresquinhas e inovadoras, quando você percebe que as novidades têm, embutidas no seu "core", muitas raízes do antigamente. Só para citar: A inteligência artificial foi gestada em torno dos anos 1950, com Alan Turing, o pai da computação, no Dartmouth College, com a criação de um campo de pesquisa para máquinas inteligentes. Hoje a IA, com ML e redes neurais é um grande "must" e traz fortes elementos da sua origem, da metade dos anos 1950. No entorno disso tudo, observamos tecnologias mais evoluídas, softwares mais específicos nos seus objetivos e uma miríade de novos produtos rodando em máquinas mais rápidas e com linguagens mais especializadas. Mas a metáfora da receita culinária aplicada na computação continua pertinente. A computação sempre teve um grande paralelismo com a culinária, por assim dizer. Vejamos: Uma receita é formada basicamente de passos e ingredientes, visando a produção de algo, que queremos que seja bom. Os passos da receita são, no fundo, passos de processos que, no mundo dos bytes, se transformam em programas codificados. Os ingredientes da culinária, por sua vez, transpostos para o mesmo conceito, são os "dados" que abastecem os programas, lembrando também que uma receita executada, dá origem a produtos (novos elementos, que podem ser considerados dados também). A receita do delicioso pudim de pão, tem como ingrediente principal (de entrada) o pão, que, diga-se de passagem, já foi produto de saída de outra receita. Puras divagações na linha de "pipeline" de master chef. Na culinária, o objetivo sempre buscado (mas nem sempre conseguido) é fazer um ótimo prato, com bons ingredientes, num tempo adequado, sempre com o objetivo final de satisfação (do cliente ou do "chef"). Na computação, é exatamente igual: O objetivo final é sempre fazer um pedaço de código rodar, com menor propensão a erros e com rapidez, para atender demandas, cada vez mais urgentes das áreas de negócios. Hoje, felizmente, a computação caminha para alcançar a maturidade da culinária e perceber que a qualidade dos ingredientes da receita (os dados) é fundamental para a qualidade do produto gerado (o prato final, o resultado). A importância dos dados e de sua qualidade hoje é inquestionável. Os metadados, que são elementos fundamentais na qualidade dos dados, também começam a sair do patamar de "patinho feio" e ganham importância. E eles, claramente, se encaixam na metáfora da culinária. Costumo falar que os metadados são aquelas plaquinhas que ficam ao lado dos "réchauds" nos restaurantes self-services. De nada adianta você ter feito um ótimo prato (com receita consagrada, com os processos/procedimento bem elaborados), com excelentes ingredientes(que são os dados com qualidade), produzindo um resultado muito bom, se visualmente/conceitualmente não se sabe o que aquilo significa. Você olha cuidadosamente aquele prato interessante, com borbulhas cheirosas, imerso em um molho espesso e sedutor, mas fica em dúvida: É peixe, frango, filé ou bacalhau ? Isso é tão fundamental no consumo de um belo prato, quanto saber, nos dias de hoje, o que os dados significam. E estando no ambiente dito SELF, isso torna-se fundamental. Hoje, nas novas soluções de BI, Analytics e Data Ops, o conceito de ***self service*** torna-se fundamental e os metadados se tornam essenciais nessa equação. Assim, as receitas, tanto da culinária, quanto as de um processo de pipeline, se casam. Tudo isso, para destacar esse alinhavo

de conceitos, atuais, históricos e metafóricos, costurados em threads, como estabelecem o Modelo ágil, Devops , DataOps e Governança de dados.

2-Devops:

O conceito de DevOps foi um movimento que chegou atrelado à maturidade da proposta Ágil, que tinha o seu maior oxigênio dedicado à parte inicial do ciclo de desenvolvimento de sistemas. A agilidade era mais concentrada na parte de Requisitos, distribuídos pelas histórias nos Sprints, com a interação com os PO´s etc. O conceito de Devops, de forma geral, chegou para complementar, propondo maior agilidade e controle na fase final do ciclo de desenvolvimento de software (Desenvolvimento e Operação), com introdução e consolidação de conceitos de integração contínua, maior automatização dos processos de testes, e coleta de indicadores, que sempre foram propostos pela Engenharia de Software tradicional. Tudo para acompanhar a fábrica de software e melhorar a qualidade de seus produtos. Grandes nomes da indústria abraçaram a ideia de Devops e produtos foram desenvolvidos para apoiar essa parte do ciclo de desenvolvimento e operação Outro ingrediente de bom senso trazido nesse pacote foi a comunicação, fundamental de ser estabelecida entre o Desenvolvimento e a Operação, levando, dessa forma, os conceitos de engenharia de software, para mais próximo da turma da infra. Abordagens de Gerência de configuração (com controles de versão, baselines, fluxos paralelos, integração etc.), e de certos controles estatísticos desembarcaram também na operação, permitindo maior automação dos processos, com a detecção dos seus erros mais incidentes. Por exemplo, alterações(não governadas) em códigos, como causas-raízes de erros no ciclo de vida de dados e processos. A integração se dá também com a área de negócios, criando maior sinergia entre os que desenvolvem e operam, e aqueles que consomem os produtos dos sistemas em operação. Seria você, por exemplo, que adora pão italiano, conversando com o padeiro da Trigopane de frente da sua casa. A computação in Cloud surgiu nesta equação, melhorando a rapidez, a elasticidade e a escalabilidade das camadas de Software e Hardware e algumas mensagens se tornaram fortes, como a necessidade de integração de processos e de dados. Tanto na terra(on-premise), quanto no céu(in-cloud)...

3-DataOps:

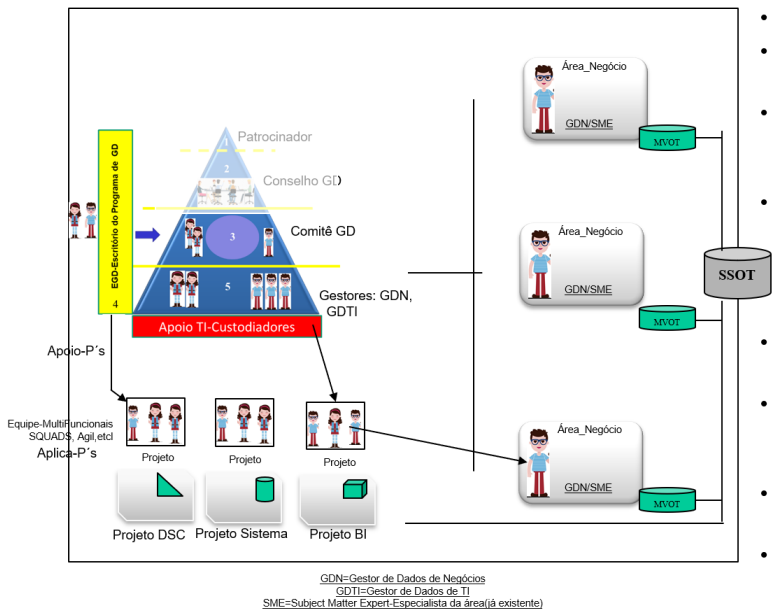
O conceito de DataOps, claro, bebeu dos princípios do Devops e nasceu no meio do tufão de Big Data, IOT etc, produzindo dados em volumes assustadores. Isso tem até levado os conselhos internacionais a repensar as escalas definidas para os volumes de bytes. Os terabytes (10^{12}), com 1 trilhão de bytes, já se aproximam dos minúsculos pen-drives e as Comissões internacionais já estão buscando novas escalas para depois do Yottabytes(10^{24}), o chamado septilhão de bytes. Com essa inundação de dados, o mundo dos negócios começou a se mexer de forma diferente. Surgiram novas possibilidades de produtos e de melhorias de outros já existentes, o uso da análise preditiva e prospectiva sobre informações, decisões quase imediatas sobre assuntos críticos, capazes de definir o “go/nogo” de uma decisão fundamental de negócios. Tudo isso definiu uma necessidade óbvia: As empresas precisam adaptar as suas receitas de desenvolvimento de soluções de sistemas, digamos mais tradicionais, (como pensado pelo Devops) mas também aplicar a mesma receita em sistemas para dados (Analytics, BI, IA etc). Assim nasceu o DataOps. Nele há princípios filosóficos semelhantes aos dos Agilistas, que convenhamos são básicos (ou deveriam ser), como foco em entrega de valor, rapidez de resposta, equipes múltiplas, interações diárias (como o daily Scrum) etc. Outros são mais focados em certas culturas, como auto-organização, equipes autossustentadas etc. Há conceitos fundamentais que são sugeridos visando a resolução dos problemas que brotam nesses domínios. O CookBook da Data Kitchen (1), empresa com foco em DataOps, por exemplo aponta alguns itens, sobre os quais faço as minhas considerações:

- Os requisitos nem sempre são claros. Há usuários que sabem que precisam de algo, mas não sabem bem o quê. Todos conhecemos essa história;
- Os dados continuam sendo produzidos em silos. Esse desafio vem desde os primeiros sistemas construídos e pensou-se que a introdução das técnicas de bancos dados, nos anos 60/70 seria solução. Não foi;
- Os dados para essa camada não chegam nos formatos mais adequados. Os dados que chegam na boca do pipeline, vêm de variadas fontes e em diferentes formatos. Por isso, o conduto terá diversas etapas para acertar dados, corrigir formatos, enriquecer campos e consolidar e integrar informações. Aliado a isso, em função da ***velocidade*** de chegada e do ***volume*** de dados começamos a ter a necessidade de automatizar os controles, dentro das camadas dos pipelines, visando a minimização de erros, com detecções precoces de problemas de qualidade. Soma-se a isso também, e associado ao outro "v" dos Big Data (a ***variedade)***, os dados são provenientes de diferentes fontes, sistemas, origens etc. Isso clama por preocupações de integração e interoperabilidade de dados (a fatia nova do DMBoK2 lembra isso) e faz com que a velha virtualização seja repensada, de forma mais otimizada e enxuta. O problema agora se junta com a velocidade requerida para se estabelecer esses ambientes heterogêneos de execução. Essa espécie de virtualização de ambientes de 2ª geração, denominada Containers(softwares e agregados, isolados com baixíssimo acoplamento), dockers e kubernetes, aparece para controlar e orquestrar esses espaços de processamento, buscando melhorias de "práticas" , porém em contextos mais complexos. O resumo da visão "dataopista" é que a manutenção e o controle manual do fluxo(pipeline) das ferramentas de ETL são custosas e com margem para erros. Dessa forma, os v´s do Big Data gritam que há espaço para se automatizar boa parte do caminho, minimizando intervenções manuais e os erros que delas se originam;
- A percepção de que a qualidade dos dados arruína qualquer projeto começa a piscar com intensidade. Aqui entra o grande mote da GD. Ter controle sobre a Qualidade de dados. Há um ditado "mineirinho" que diz que "de cano sujo não sai água limpa". Isso serviu de lema para que as melhorias de processos de software fossem adotadas, nas mais de 100 empresas onde a consultoria Fumsoft em MPS.BR(que liderei por 14 anos) trabalhou. Por outro lado, nessa dicotomia dados e processos, houve o encaixe da turma de dados, que acrescentava: "Mesmo que o cano esteja limpo, se entrar água suja, sairá água suja". O clássico "Garbage in, Garbage out". Ou seja, os dados críticos que chegam na boca do pipeline deverão ter qualidade. As camadas do pipeline podem corrigir erros pequenos de formatos, conteúdos com erros de padrão, mas dificilmente detectam erros de conceitos em dados mestres, referenciais que são o "core" de uma empresa. Assim, a GD defensiva(já discutida) também entra em cena;
- Finalmente, a percepção, mostrada por pesquisas de fontes de certa referência, indica que há urgência na modificação da forma que se aplica hoje, na criação de ***sistemas analíticos***. Verdade, e daí a corrida em busca de cientistas de dados, engenheiros de dados. Mas isso também não está sendo suficiente, segundo as fontes apontadas. Há que se ter um ecossistema mais integrado de envolvidos da ***área de negócios***, da ***TI*** e da ***Governança***. Aí chegamos na Governança de dados ofensiva.

4-GD Ofensiva:

Com o desenvolvimento dos novos cenários de negócios, onde entraram Big Data e dados não estruturados, exploração de dados de mídias sociais, IA para buscar novos caminhos de marketing e vendas, IoT e Inteligência geográfica, com Uber etc e a chegada dos drones de entrega, uma nova forma de GD começou a ser pensada em termos controle dos dados. Nasceu a GD Ofensiva.

- A GD Ofensiva não nasceu diretamente para substituir a outra (Defensiva), mas para ser uma opção paralela em empresas cujo foco, além da defesa dos dados, também visa o seu consumo intensivo e sua elaboração dentro dos novos caminhos de negócio, com a agilidade que o mercado demanda. A GD Ofensiva, ganhou, conforme as referências consultadas e em observações em algumas empresas onde tive oportunidade de discutir o assunto, essa nova feição de uma GD mais ágil e menos controladora;
- A GD Ofensiva otimiza as aplicações analíticas com modelos de aplicações, transformações, simulações e enriquecimento e as palavras chaves são flexibilidade, competitividade e cooperação ;
- O GD Ofensiva favorece no sentido de buscar espaço para brigas em novos campos de batalhas dos dados, saindo da defesa para o ataque, com maior rapidez, enquanto a outra fica atrás garantindo a defesa, caso, por exemplo, uma agência reguladora apareça para vasculhar seus movimentos que devem seguir "compliances" rigorosos. Enquanto a GD Defensiva prima pelo controle de qualidade dos dados sensíveis e críticos da organização (via SSOT-Single Source of Truth), a GD Ofensiva permite o uso de versões múltiplas de dados derivados da fonte principal(MVOT-Multiple versions of Truth), porém com cuidados imprescindíveis, como a necessidade de reconciliações periódicas;
- Resumo: No fundo, observa-se que as empresas poderão ter as duas personalidades de GD, que deverão trabalhar num ponto da balança, onde o equilíbrio buscado entre a rapidez na produção e o consumo dos dados seja alinhado com a necessidade de dados controlados e com qualidade. Os dados mestres e mais críticos deverão ser trabalhados em conjunto com dados mais operacionais como vendas, compras, acessos de clientes, pesquisas no site da empresa, variações de mercado etc. Fazendo-se uma projeção simplificada, com o Spotify (Serviço digital de músicas), por exemplo, observaríamos: Os dados considerados "mais" mestres e referenciais (Assinantes, Tipos de planos, Músicas, Artistas, Compositores, Editoras e Anunciantes) deverão ser analisados, em conjunto, com os dados flutuantes e operacionais gerados na rede, na forma de execução de músicas, podcasts criados e ouvidos, playlists definidos e compartilhados, anúncios veiculados e que chegam, aos bilhões de bytes, na forma de streaming etc. É importante observar que essa dicotomia entre qualidade buscada pelo controle mais rigoroso e a liberdade necessária para a maior rapidez no consumo e nas tomadas de decisão, nos persegue há muitos anos....A figura 01 ilustra os conceitos de GD Ofensiva.



- GD Ofensiva, com tom mais liberal
- Ênfase no consumo, liberdade, tempo real e rapidez de soluções
- Foco mais nos dados transacionais da empresa(movimentos de Consumidores, Clientes, Marketing, Vendas etc)
- Pirâmide mais discreta, com atuação predominante em casos mais críticos no Conselho de Owners e Patrocinadores
  - Gestores resolvem nos grupos /equipes de projetos
  - Participação nos Squads, grupos de Agile, etc
- Aplicada em situações de empresas com necessidade de maior velocidade na produção e consumo dos dados
- Maiores riscos de silos, qualidade, linhagem de dados, fontes dispersas(mais silos-MVOT), exigindo reconciliação
- As duas abordagens(Def e Ofe) podem ser adotadas em contextos diferentes, na mesma organização
- Equilíbrio entre controle e qualidade dos dados versus facilidade de consumo e maior rapidez e competitividade

Figura 01-Governança de dados Ofensiva

No livro de Smith e Basu (2), são claras as convergências entre o DataOps e uma forma de governança de dados ofensiva, que atue com maior flexibilidade em sistemas analíticos, mas garantam as premissas de manutenção de um dado confiável e com qualidade. Veja alguns princípios definidos:

- Estabelecer um framework de GD, com metodologia(P´s-Políticas, Procedimentos e Padrões), garantindo a usabilidade e proteção da informação ;
- Estabelecer a supervisão sobre “compliance”, criando regras, procedimentos e guias para garantir a privacidade e segurança da informação. Estender esse conceito (tornando-o contínuo) saindo do “data at rest” e incluindo também os “data in motion”;
- A GD contínua também é responsável por estabelecer e supervisionar aderência às regras, guias e procedimentos para a criação dos Metadados organizacionais. Isso envolve todas as áreas da Gerência de Metadados, incluindo a documentação de ativos de dados, o acompanhamento da responsabilidade e “accountability” organizacional, o estabelecimento do(s) Glossário(s) de negócios, o acompanhamento da linhagem de dados e a coleta e uso de metadados operacionais e seu uso para objetivos de auditoria e de governança ;
- Definição de guias para reutilização de dados;
- Confirmação dos papéis e responsabilidades para criação, manutenção e gestão(stewardship) dos metadados e provê orientações para aplicação e uso de ferramentas relacionadas aso softwares(de dados).

A figura 02 ilustra o casamento dos conceitos de GD Ofensiva com DataOps

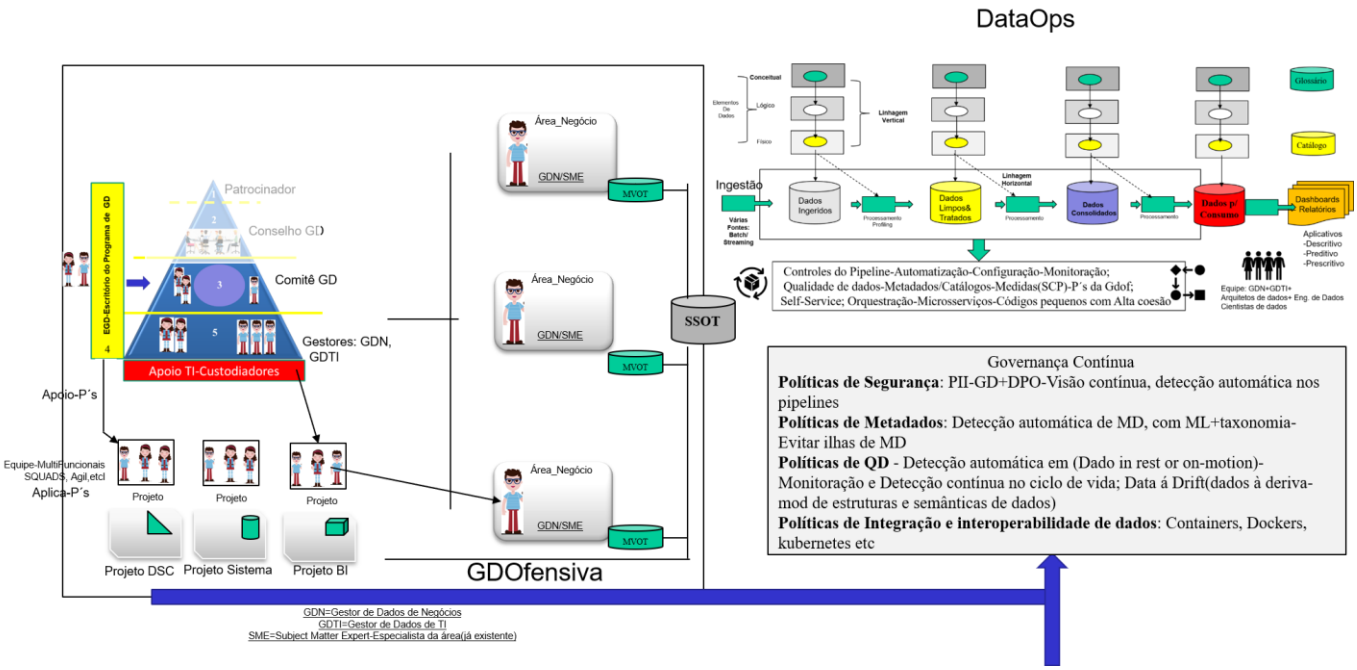


Figura 02-Visão da GD Ofensiva com DataOps

Conclusão:

1-Os dados, independentemente da sua categoria, continuam a merecer o devido controle de sua qualidade, dos seus metadados, da integração/interoperabilidade, segurança e (num futuro bem próximo) da ética;

2-Os dados considerados Mestres(Clientes, Produtos, Pessoas, Locais, etc), os referenciais (códigos diversos, como cep, código de serviços, código de doenças, códigos de profissão etc.) que são fundamentais nos alicerces das empresas e estão envolvidos com intensidade em obrigações de "compliances", estarão mais fortemente na lupa da visão mais defensiva da GD ;

3-Os dados mais transacionais, como movimentos de vendas, atendimentos, visitas a museus, visitas a sites, compras efetuadas, sinais de sensores IOT, otimização de trajetos etc, que nascem e/ou chegam na empresa honrando os 5V do Big Data, estarão mais na lupa da GD Ofensiva, com forte base oferecida pelo DataOps;

4-Por fim, caberá à empresa, com relação à Governança de seus dados, a visão de considerar os dois caminhos, no tempo certo, nos recursos disponíveis e na adequada dimensão de suas necessidades de negócios. Somente isso e isso não é pouco...

Referências:

1- DATAKITCHEN. The Datakitchen Cookbook. 2019. Disponível em: https://datakitchen.io/dataops-cookbook-second-edition.html, acesso em : 20 jul 2020.

2- Schmidt,J. ; Basu, K. *DataOps The Authorative Edition.* Panther Publishing, 2019. [E-Book, Kindle].

3- Atwal, H. *Practical DataOps-Delivering Agile Data Science at Scale.* Apress Publishing, 2020. [E-Book, Kindle].