ProKanban.org

O Guia de Bolso

O Que Ninguém Te Contou Sobre o Kanban Pode Te Matar

Colleen Johnson, Prateek Singh
Dan Vacanti

O Guia de Bolso Kanban

O Que Ninguém Te Contou Sobre O Kanban

Pode Te Matar

Daniel Vacanti, Prateek Singh e Colleen Johnson

Tradução para o português:

Renato Barbieri

Nota do tradutor:

Fiquei muito feliz pela oportunidade que me foi dada pela ProKanban.org de poder oferecer a minha contribuição à comunidade Kanban e Ágil com essa tradução do Guia de Bolso Kanban para o português.

Procurei seguir as traduções dos principais termos utilizados no Guia Kanban (2020) para manter consistência entre os guias. Assim, WIP, Cycle Time, Work Item Age, Throughput e Service Level Expectation, todos seguem as traduções originais. Outros termos foram mantidos no seu original em inglês, pois considerei que os traduzir causaria confusão nos leitores e leitoras, como, por exemplo: upstream, downstream, swarm, mobbing, etc.

Embora SLE signifique “Expectativa de Nível de Serviço”, e, portanto, seja um termo feminino, sempre me refiro ao seu acrônimo, SLE, no masculino, apenas para manter consistência com a tradução original do Guia Kanban (2020). O mesmo ocorre com DoW (“Definição de Fluxo de Trabalho”). E uma última observação: ao ler “funcionalidade”, nesta tradução, estou me referindo ao termo em inglês “feature”.

Gostaria de agradecer à **Janée McConnell** pela oportunidade e por todo o seu apoio durante o processo de tradução, e ao trabalho de revisão de **Ingrid Andrade Nunes**, Professional Kanban Trainer da ProKanban.org.

Junte-se a nós no Slack gratuito da ProKanban.org

Sumário

Prefácio

O que é Kanban?

Parafraseando Ryan Ripley e Todd Miller, "Você já leu o Guia Kanban?"[1]

O “Guia Kanban”[2], publicado em 2020, é a referência definitiva sobre o que o Kanban realmente é:

"Kanban é uma estratégia para otimizar o fluxo de valor através de um processo que utiliza um sistema visual, baseado em um sistema puxado (pull-based system). Pode haver várias maneiras de definir valor, incluindo a consideração das necessidades do cliente, do usuário final, da organização e do ambiente, por exemplo.

O Kanban compreende as três práticas a seguir trabalhando em conjunto:

- Definir e visualizar um fluxo de trabalho
- Gerenciar ativamente itens em um fluxo de trabalho
- Melhorar um fluxo de trabalho

Em sua implementação, essas práticas Kanban são chamadas coletivamente de sistema Kanban.”

Os conceitos incluídos neste livro são baseados nesta definição de Kanban, bem como nos outros elementos do Kanban detalhados no guia. Se você ainda não leu o Guia Kanban, então deveria parar de ler este livro agora, e ler o Guia Kanban primeiro para depois voltar aqui.

Por Que o Mundo Precisa de Outro Livro sobre Kanban?

Na verdade, não precisa, mas uma verdade ainda maior é que realmente precisa. Este livro — assim como o Guia Kanban mencionado acima — procura destilar o Kanban até a sua essência mais básica para torná-lo o mais aplicável possível aos mais diversos contextos. Como tal, este livro focará no Kanban como uma estratégia para obter o fluxo e ignorará intencionalmente muitos elementos prescritivos desnecessários que você pode ter visto erroneamente adicionados ao Kanban no passado.

Isso não significa de forma alguma que você não deve estudar outras práticas Lean-Agile que possam ser usadas com o Kanban (por exemplo, Scrum, Teoria das Restrições etc.) Certamente, outras práticas podem — e na maioria dos contextos devem — ser adicionadas ao seu repertório ao implementar o Kanban. Mas isso significa que o que é definido no Guia Kanban, e o que é explicado de forma mais detalhada aqui, representa o que a comunidade mais ampla aceita como Kanban e estabelece uma linguagem comum a ser usada ao discutir a estratégia de fluxo.

A Quem este Livro se Destina

Este livro se destina a qualquer pessoa que tenha lido o Guia Kanban e gostaria de obter alguma ajuda para preencher quaisquer lacunas no seu conhecimento sobre o que o Kanban realmente é.

Escrevemos este livro para aquelas tantas pessoas que já tentaram o Kanban ou que estão a pensar em experimentá-lo nas suas vidas profissionais ou pessoais. Com esse objetivo, enfatizamos especialmente algumas justificativas para o porquê de o Kanban existir, além de esclarecer

muitos dos mitos que existem em torno do Kanban hoje em dia. Esforçamo-nos continuamente para esclarecer muitos dos equívocos sobre o Kanban, e este livro representa apenas um passo nessa jornada.

Este livro, no entanto, não representa um guia passo a passo para implementar o Kanban. Nem a referência definitiva para o Kanban. Seria impossível apresentar toda a teoria subjacente, então tudo o que tentamos fazer foi: resumir alguns conceitos importantes e dar algumas dicas sobre como começar. Para uma lista mais detalhada de referências para teorias e práticas complementares, você pode consultar a bibliografia deste livro ou, ainda melhor, visitar a página do ProKanban.org (https://www.prokanban.org). Com este livro e essas referências, você deverá ter uma compreensão bastante completa do que o Kanban realmente é.

Como Ler este Livro

Embora este livro tenha sido projetado para ser lido em ordem, nada impede você (como em qualquer livro) de pular diretamente para qualquer capítulo que pareça interessante. Em cada capítulo, tentamos fornecer referências, tanto quanto possível, para outras partes do livro onde conceitos importantes são abordados. Nesse sentido, ao ler este guia de bolso, você pode transformá-lo em uma espécie de "escolha a sua própria aventura". No entanto, se você é novo no Kanban, recomendamos fortemente que leia os capítulos em ordem para poder progredir em seu aprendizado sem perda de continuidade.

Pensamento Final

Há muito ruído por aí sobre o Kanban. Pesquise na internet e logo se perderá na cacofonia da desinformação sobre fluxo. Esperamos que este livro lhe ajude a navegar em meio a esse ruído para obter o máximo do conteúdo, que acreditamos ser o conjunto mais empolgante de práticas no mundo Lean-Agile. Então, vamos parar de enrolar e deixar você chegar a esse conteúdo interessante começando no Capítulo 1.

Divirta-se!

Notas

1. Todd Miller and Ryan Ripley, “Fixing Your Scrum” (The Pragmatic Programmers, 2020)
2. John Coleman and Daniel Vacanti, “The Kanban Guide” https://kanbanguides.org/

Prólogo – A História Secreta do Kanban e Por Que Isso Importa

O texto abaixo apareceu originalmente como um blog post acompanhado de uma apresentação[1] de Darren Davis. Foi reproduzido aqui com a permissão integral de Darren.

No Começo

No verão de 2006, eu trabalhava como Gerente de Desenvolvimento para a Corbis, uma empresa de licenciamento de mídia de propriedade de Bill Gates. A Corbis tinha uma habilidade incrível para perder dinheiro. Quando Bill passou de homem mais rico do mundo para o segundo mais rico, eu dizia que contribuí para isso por simplesmente trabalhar na Corbis. Na época, o trabalho de engenharia de sustentação estava um caos, e eu estava na equipe que tentava tirá-lo do atoleiro e torná-lo funcional. Naquele momento, a sustentação era executada como uma série de pequenos projetos, liberando correções e melhorias a cada trimestre, com um escopo fixo; era mais ou menos como pegar um monte de meias da secadora e, andando de lado, subir as escadas tentando derrubar o mínimo possível de meias pelo caminho. Nem é preciso dizer que o processo era consistentemente desanimador. A equipe havia ouvido falar de algo chamado Agile e decidiu incluir stand-ups em seu processo. Estes rapidamente degeneraram em maratonas matinais de 30 a 45 minutos e conseguiram sugar o otimismo de todos os infelizes que apareciam para as

reuniões. Estava claro que as coisas tinham que mudar, e alguns de nós começamos a discutir sobre como melhorar o processo.

Em algum momento durante o verão, um dos membros da nossa equipe, Rick Garber, ouviu uma palestra de um escocês, David Anderson, explicando como ele havia resolvido alguns dos mesmos problemas em uma das suas equipes na Microsoft. Os seus métodos, baseados na Teoria das Restrições, e em outros trabalhos de nomes como Goldratt e Deming, eliminaram a estimativa explícita do processo, e se basearam em dados para fornecer um meio probabilístico para determinar quando o software provavelmente estaria pronto. Francamente, ele me conquistou ao eliminar estimativas, mas o restante das teorias também era convincente. Ficamos pasmos ao pensar em software como uma forma de estoque que poderia ficar obsoleto, ou mesmo que a redução da carga de trabalho de alguns recursos poderia, na verdade, tornar o sistema todo mais eficaz. Vários de nós lemos o seu livro e participamos de uma série de conversas com total convicção, os olhos brilhando com zelo revolucionário, e ansiosos para mudar o mundo. Ou, pelo menos, o nosso moribundo processo de sustentação. Começamos a construir o nosso primeiro sistema baseado em Kanban.

Isso levou vários meses e estendeu-se até o outono de 2006. No meio desse processo, David ingressou na Corbis como Diretor Sênior de Engenharia de Software, e eu comecei a me reportar a ele. Também estavam na equipe, naquela época, Dominica Degrandis, Mark Grotte, Larry Cohen, Rick Garber e Steven Weiss. David nos guiou pelo restante do processo e finalmente, em novembro, chegamos a um design que tinha a sua aprovação. Treinamos nossa equipe no processo, povoamos as nossas filas e, com grande expectativa e pompa, lançamos a primeira implementação significativa de um sistema baseado em Kanban que tínhamos conhecimento.

E nada aconteceu. Por meses.

Kanban é Lançado, e Fracassa

Tenha em mente que o nosso sistema era baseado no primeiro livro de David, “Agile Management for Software Engineering”, e não em seu trabalho subsequente. Não havia exemplos práticos reais de como implementar as suas teorias no primeiro livro, e projetamos o nosso sistema muito parecido com a sua única (e menor) implementação anterior, que ele nos assegurou ter sido um tremendo sucesso. No entanto, o nosso processo estava rapidamente se mostrando inviável. Nesse ponto da sua história, o Kanban não se parecia em nada com o processo que conhecemos hoje. O que tínhamos eram cerca de 25 itens de trabalho armazenados no Team Foundation Server, organizados em uma série de aproximadamente 14 filas, com uma rede confusa de transições entre os vários estados, e quase mais nada. Segundo a teoria, uma vez que o sistema estivesse configurado, ele se autogerenciaria. As pessoas entenderiam os seus papéis, monitorariam as suas filas, fariam o seu trabalho e o passariam adiante. A cada duas semanas, qualquer trabalho que tivesse passado pelo processo e estivesse em “Pronto para Produção” seria liberado. No entanto, muito poucos itens chegavam à produção, e nenhum de nós, incluindo David, entendia por quê. As pessoas reclamavam que não tinham ideia de onde as coisas estavam no processo, os desenvolvedores e engenheiros de QA não tinham visibilidade do trabalho que estava a caminho deles, e os clientes estavam cada vez mais irritados por receberem apenas poucos resultados, frutos escassos do trabalho da nossa grande experiência. Insistimos que deveria funcionar, que projetamos a máquina de forma que as pessoas pudessem se concentrar apenas em suas áreas, sem se preocupar com outras áreas do sistema. “Apenas foque na sua fila”, nós dizíamos, “e quando o trabalho chegar, simplesmente pegue-o, faça o trabalho e passe adiante.” Mas ainda assim eles reclamavam, e os nossos clientes ficavam cada vez mais

impacientes. Dizer que estávamos caminhando para o desastre implica em muita velocidade, quando, na verdade, estávamos praticamente parando.

No início de fevereiro, o CIO (e chefe de David), Stephen Gillett, me chamou no seu escritório e disse: “Se você não consertar isso, vou ter que demitir alguém.” Eu não achava que ele queria dizer que eu pessoalmente precisava consertar, mas pensei que ele estava ameaçando me demitir se a equipe não conseguisse fazer o processo funcionar (descobri depois que ele nem estava se referindo a mim, mas essa é uma história para outro dia). Me reuni com Mark, Dominica, Rick e Larry (a equipe de liderança de David), e discutimos sobre como desbloquear as coisas. Creio que seja importante notar que não tínhamos a intenção de modificar ou “consertar” o processo. Simplesmente, e ingenuamente, pensamos ser apenas um problema de ignição, que se conseguíssemos fazer a coisa começar, ela decolaria. Com esse objetivo, decidimos começar a fazer uma reunião diária. Foi sugestão da Dominica, e todos achamos uma coisa razoável a se fazer. A nossa intenção era fazer a reunião diária por um mês, até começarmos a ver algum movimento, e então voltaríamos a conduzi-la da maneira habitual desde o início. Decidimos começar as reuniões na segunda-feira seguinte. Insisti em conduzi-las porque sabia que, se eu liderasse a reunião, ela nunca duraria mais de 15 minutos. Se eu era dogmático em relação a alguma coisa, era que as reuniões diárias deveriam ser curtas. Ainda tinha cicatrizes emocionais das reuniões diárias que o nosso processo anterior de sustentação realizava. Curiosamente, essa única mudança, destinada a ser um impulso temporário, conduziu todas as outras modificações que se tornaram o Kanban como o conhecemos hoje.

Micro Melhorias e Acidentes Felizes

Eu nunca havia conduzido uma reunião diária antes, então realmente não sabia o que fazer naquela primeira segunda-feira. Como as pessoas estavam reclamando que não tinham visibilidade sobre onde estavam as

coisas, achei que fazia sentido colocar o trabalho em um quadro branco para podermos discuti-lo. No entanto, eu não tinha uma ideia clara do formato e decidi discuti-lo com alguns dos nossos desenvolvedores. Na mesma semana, por sorte, eles estavam trabalhando com Daniel Vacanti fazendo Modelagem de Domínio de Cores, um processo de usar *Post-its* coloridos para elaborar o design de sistemas de software. Quando contei a eles que estava pensando em colocar o trabalho no quadro branco, um dos desenvolvedores, Kurt Quamme, sugeriu usar *Post-its* de cores diferentes para representar o trabalho. Pareceu uma boa ideia, então fui para casa naquele fim de semana e esbocei um plano bastante simples.

A cor mais comum do *Post-it* era amarela, então usamos essa cor para representar solicitações de recursos. Para representar “bugs”, usamos azul, porque, bem, para ser honesto, tanto “bugs” quanto “azul” começam com a letra b (azul em inglês é “blue”). E isso foi praticamente tudo para o primeiro quadro. Cheguei cedo ao trabalho na segunda-feira, desenhei um conjunto muito rudimentar de filas, escrevi alguns *Post-its* e esperei as pessoas chegarem. Havíamos deixado claro que se algo fosse atribuído a você, você era obrigado a participar da reunião diária, mas naquela primeira manhã a maioria da equipe compareceu. Talvez tenha sido porque as pessoas estavam curiosas para ver o que estava acontecendo, ou porque queriam sentir que faziam parte da equipe, mas é provável que tenha sido porque o quadro estava configurado em uma área bastante pública, diretamente atrás da minha mesa, e durante a reunião diária era difícil para as pessoas nas proximidades fazerem qualquer outra coisa que não fosse participar dela. Seja qual for o motivo, tivemos um grande grupo naquela primeira manhã, e continuaria a ser uma reunião grande e inclusiva durante a maioria do tempo em que estive lá.

O primeiro problema, e o mais urgente, que estávamos tentando resolver (na verdade, naquele momento, o único problema que estávamos tentando resolver), era descobrir por que o trabalho não estava se movendo pelas filas. Na reunião diária, esse era o nosso foco. Parece óbvio agora que focar em problemas de bloqueio é o principal objetivo de uma reunião diária do

Kanban, mas na época era uma ruptura radical em relação ao estilo mais ortodoxo de “Todo Mundo Pode Falar”, comum ao Scrum, o tipo de abordagem que passa em círculo e pergunta às pessoas o que fizeram ontem e o que planejam fazer hoje. Eu não estava tentando ser radical, simplesmente não sabia fazer de outra maneira. Com o tempo, como grupo, refinamos um pouco o processo, estabelecendo algumas perguntas padrão:

- Há algo te bloqueando que não está no quadro?
- Há alguém trabalhando ativamente para resolver os problemas no quadro?
- Você precisa de algo da gerência para resolver o problema?

Sempre mantivemos a reunião abaixo de 15 minutos, muitas vezes abaixo de 10, mas descobrimos que as pessoas “ficavam depois da aula” e discutiam questões específicas em grupos de 2 ou 3. De certa forma, aprendemos uma regra básica das reuniões: não desperdice o tempo das pessoas, e discuta apenas questões diante de todo o grupo quando todo o grupo precisa ouvir. Também tínhamos a sensação de que a equipe estava mais focada, mais energizada do que quando as antigas reuniões diárias se arrastavam por muito tempo.

Não me recordo quem sugeriu usar *Post-its* rosa-choque para identificar problemas de bloqueio, nem exatamente quando isso aconteceu, mas foi muito cedo. Gostaria de saber para poder comprar uma bebida para essa pessoa. Foi outra inovação simples, mas brilhante, que nos ajudou a encontrar um dos principais benefícios do Kanban contemporâneo: o poder inerente de visualizar o trabalho em andamento. Não inventamos o conceito, de forma alguma, e se tivéssemos lido sobre o trabalho de Edward Tufte, talvez tivéssemos chegado a essas ideias muito mais cedo, mas à medida que a equipe iterava sobre a forma e o conteúdo do quadro, ficamos cada vez mais conscientes da sua importância. Ao colocar um *Post-it* rosa na funcionalidade ou “bug” que estava bloqueado, isso rapidamente

comunicava, até ao observador mais casual, que algo estava errado. A codificação por cores e outras pistas nos permitiram obter diferentes níveis de detalhes do mesmo quadro. Você poderia observar de longe e ter uma ideia da saúde geral do sistema, onde estavam os gargalos, qual seria o tamanho do lote da próxima versão, ou poderia observar mais de perto e ver que o mesmo desenvolvedor estava atribuído a vários itens e precisava ser reorientado para uma única tarefa. O quadro também se tornou o ponto focal de discussões e planejamentos para várias pessoas da equipe. O nosso analista de testes, Tom Utterback, e o nosso engenheiro de *build*, Doug Buros, usavam o quadro para planejar como os *builds* seriam movidos para QA. Começamos a coletar todos os itens liberados colando-os na parede, no lado direito do quadro. No início era uma espécie de piada, mas rapidamente se tornou uma propaganda e um lembrete do nosso sucesso como equipe. Todas essas coisas são óbvias agora e amplamente utilizadas, mas na época estávamos apenas inventando, experimentando, descobrindo o seu valor conforme avançávamos.

Bem-Vindos à “Cúpula do Trovão”, o Processo de Priorização

O processo de seleção e priorização do trabalho era feito em uma reunião semanal dos vice-presidentes de várias partes da empresa. Marketing, Finanças, Vendas, Imagem, todos compareciam às segundas-feiras, por uma hora, para selecionar os próximos itens a serem trabalhados. A reunião era conduzida por Diana Kolomiyets, que também desempenhava um papel fundamental em manter as coisas fluindo através das filas e gerenciar as liberações. A reunião passou por várias iterações, como tudo o mais, mas eventualmente se estabeleceu em um sistema de múltiplas votações. Cada representante tinha três votos. A reunião começava revendo os novos pedidos da semana e depois solicitava indicações. Indicar um item para ser trabalhado era, essencialmente, um pedido de apoio, e havia uma quantidade justa de troca de favores nessas sessões. Uma vez que um

grupo de itens candidatos fosse determinado, todos votariam, com os itens mais votados sendo adicionados à nossa fila de trabalho (chamávamos Pronto para Engenharia). O mecanismo era bastante simples, e ficou estabelecido desde o início que se você não participasse da reunião, havia pouca chance de que o seu item pessoal fosse adicionado à fila de trabalho. Enquanto estávamos construindo o processo inicial, antes de quaisquer modificações que realmente fizessem o sistema funcionar, mantivemos contato regular com os nossos clientes. Um deles, Drew McLean, era o vice-presidente do departamento de imagem. Ele era um ex-fuzileiro naval e havia trabalhado grande parte da sua carreira na área de manufatura, mais recentemente na Boeing. Ele entendia bem as teorias, mas insistiu que incluíssemos uma Bala de Prata, uma capacidade de acelerar um pedido. Resistimos, pensando que tudo se tornaria uma Bala de Prata, mas é difícil resistir a um ex-fuzileiro naval que sabe o que quer. Incluímos uma cláusula para acelerar um pedido, mas adicionamos duas regras:

- Pode haver apenas um único pedido urgente em todo o sistema em qualquer momento. Se já houver um pedido urgente, um novo não poderá ser adicionado até que o anterior seja liberado para produção.
- A decisão de marcar um item como Urgente tinha que ser um consenso dos vários interessados de todos os grupos de clientes (Marketing, Vendas, Finanças, Imagem etc.)

Ficamos surpresos ao descobrir, com o tempo, que esse mecanismo era raramente era invocado. Quando os itens eram marcados como urgentes, eles tinham um efeito negativo muito perceptível sobre o resto do sistema. As coisas precisavam esperar pela Bala de Prata, o que levava diretamente a prazos de entrega mais longos para os itens regulares. Por causa disso, não era fácil fazer com que todos concordassem que um determinado item era tão mais importante do que os outros, a ponto de merecer passar na frente de todos.

Outro efeito interessante ao alcançar um fluxo melhor e uma melhor vazão, foi com os nossos SLAs. Inicialmente, decidimos por um SLA de 21 dias, baseado puramente em uma suposição. Como não tínhamos dados para começar, precisávamos começar em algum lugar e 21 dias parecia um número defensável. Concordamos em revisar esse número ao longo do tempo à medida que obtivéssemos mais dados. Nunca atingimos esse SLA de 21 dias, nem mesmo uma vez. Aumentamos esse número para 28, mas o melhor que conseguimos foi em torno de 31 dias. Estranhamente, no entanto, os nossos clientes não pareciam se importar. Como estávamos movendo itens pelo sistema com bastante regularidade e eles tinham total visibilidade de onde os itens estavam, eles pareciam bem satisfeitos em deixar o sistema funcionar sem nos pressionar muito sobre o SLA. Também pode ser que historicamente o sistema foi tão disfuncional por tanto tempo que acabamos estabelecendo expectativas muito baixas.

Então, Aonde Você Quer Chegar?

Por que tudo isso é importante? Acredito que seja importante por uma variedade de motivos, alguns tão pequenos a ponto de quase serem mesquinhos, mas outros que acredito terem implicações muito mais amplas. Primeiro, os pequenos motivos: acho importante que as pessoas que realmente fizeram o trabalho recebam algum crédito. Sei como a história funciona e sei que, na maioria das vezes, a atenção é focada em um indivíduo como responsável por um movimento, mesmo que esse movimento não tivesse ocorrido sem as contribuições de um grupo muito maior. É mais fácil rotular alguém como o “Pai do Kanban” do que apontar continuamente todos os tios e tias. Esta é uma pequena tentativa de corrigir isso, ou pelo menos adicionar alguns nomes ao registro. As pessoas mencionadas acima desempenharam um papel na evolução do Kanban, mas certamente há outras que esqueci também. Também é importante porque ilustra o abismo que frequentemente existe entre teoria e prática. Amo teoria, realmente amo, mas apenas até certo ponto, se for eficaz, na

prática. No entanto, o Kanban, como o conhecemos na indústria hoje, não existiria se um grupo de pessoas não tivesse sido especificamente encarregado de fazê-lo funcionar. As soluções que desenvolvemos funcionaram para nós naquele momento, naquele contexto. Não acho que ninguém estivesse tentando criar uma metodologia, estávamos apenas tentando fazer funcionar para não sermos demitidos. Em relação às inovações que fizeram o nosso sistema funcionar, ninguém estava orientando a equipe, aprovando mudanças no processo ou decidindo quais inovações tentar em seguida. Tudo o que fizemos foi conduzido pela equipe e avaliado exclusivamente com base em sua eficácia ou não. Qualquer pessoa que lhe disser o contrário está tentando lhe vender algo.

A razão maior pela qual creio que isso é importante é porque acredito que, como indústria, muitas vezes o nosso foco está no lugar errado. Parece que nos concentramos em aprender uma metodologia como Scrum ou Kanban, entender as cerimônias e artefatos que definem essas abordagens e nos tornar tão proficientes quanto possíveis nelas. Procuramos fora das nossas organizações por consultores ou coaches que possam vir e nos ajudar a aprender essas técnicas e aplicá-las aos nossos desafios, frequentemente únicos e peculiares. Enormes quantias são gastas em justificativas para várias metodologias como forma de adicionar legitimidade à opinião de alguém sobre o que devemos ou não estar fazendo, com um crescente senso de ortodoxia sobre o que é “ágil” e o que não é. Mas é minha opinião que uma equipe Scrum ou uma equipe Kanban tem muito menos valor para qualquer organização do que uma equipe de pensadores ágeis, pessoas que têm um bom entendimento das teorias por trás do ágil, mas são capacitadas para questionar tudo, experimentar, falhar, aprender e seguir. Confesso que não sou fã de Scrum, mas gosto de Kanban e o usei com grande eficácia em várias organizações até o momento. Por mais que eu goste, sei que não durará para sempre. Outras ideias devem surgir, e surgirão provando ser mais eficazes em fornecer valor aos nossos clientes. É evolução, e é implacável. Por melhor que uma ideia seja, não serão os teóricos que demonstrarão o seu valor. Caberá aos praticantes da arte da

engenharia de software, pessoas nas trincheiras todos os dias, descobrindo como fazer essas ideias funcionarem no mundo real.

Para fazer isso, é necessário ter uma mente ágil.

— Darren Davis

Notas

1. Darren Davis, “The Secret History of Kanban and Why it Matters” https://www.youtube.com/watch?v=7VT_Wqs4cRg&ab_-channel=ConfEngine

Capítulo 1 - O Elemento Mais Importante do Kanban

“Como você ensinaria alguém a ler um Diagrama de Fluxo Cumulativo?” (Alerta de Spoiler: o CFD definitivamente NÃO é a parte mais importante do Kanban.)

Era por volta de 2010 e eu tinha acabado de conhecer Frank Vega em um encontro de Kanban. Honestamente, talvez eu tenha cruzado com o Frank antes disso, mas em 2010 foi a primeira vez que tive a chance de conversar de verdade com ele. Naquela época, Frank era uma das poucas pessoas na comunidade de Kanban que realmente sabia do que estava falando (na minha opinião, é claro). Em uma chamada no Skype, logo após aquele encontro, Frank fez uma pergunta bastante inofensiva: “Como você ensinaria alguém a ler um Diagrama de Fluxo Cumulativo?” (Essa era uma pergunta completamente retórica, já que Frank sabia muito bem — melhor do que qualquer outra pessoa, na verdade — como ler um CFD). Nos dez anos ou mais desde que ouvi essa pergunta, grande parte da minha carreira se baseou em tentar encontrar uma resposta inteligente. A minha busca levou-me à seguinte conclusão:

A maioria do que se tornou a doutrina do Kanban está simplesmente incorreta. A doutrina do Kanban, tal como existe hoje, baseia-se principalmente em rumores e mal-entendidos e raramente se baseia em ciência, muito menos em fatos. Os primeiros dias de como a comunidade falava sobre Diagramas de Fluxo Cumulativo (não apenas como lê-los, mas sua importância em geral) foram um exemplo perfeito desse triunfo da ignorância.

Tenha paciência comigo por um momento enquanto eu explico.

No cerne do que faz um Diagrama de Fluxo Cumulativo (CFD) funcionar reside uma relação conhecida com o nome de Lei de Little. O Dr. Little até mesmo utiliza um CFD em uma das provas da sua lei homônima[1] (observe: ele o chamou de diagrama de Chegadas/Partidas Cumulativas, como mostrado na Figura 1.1):

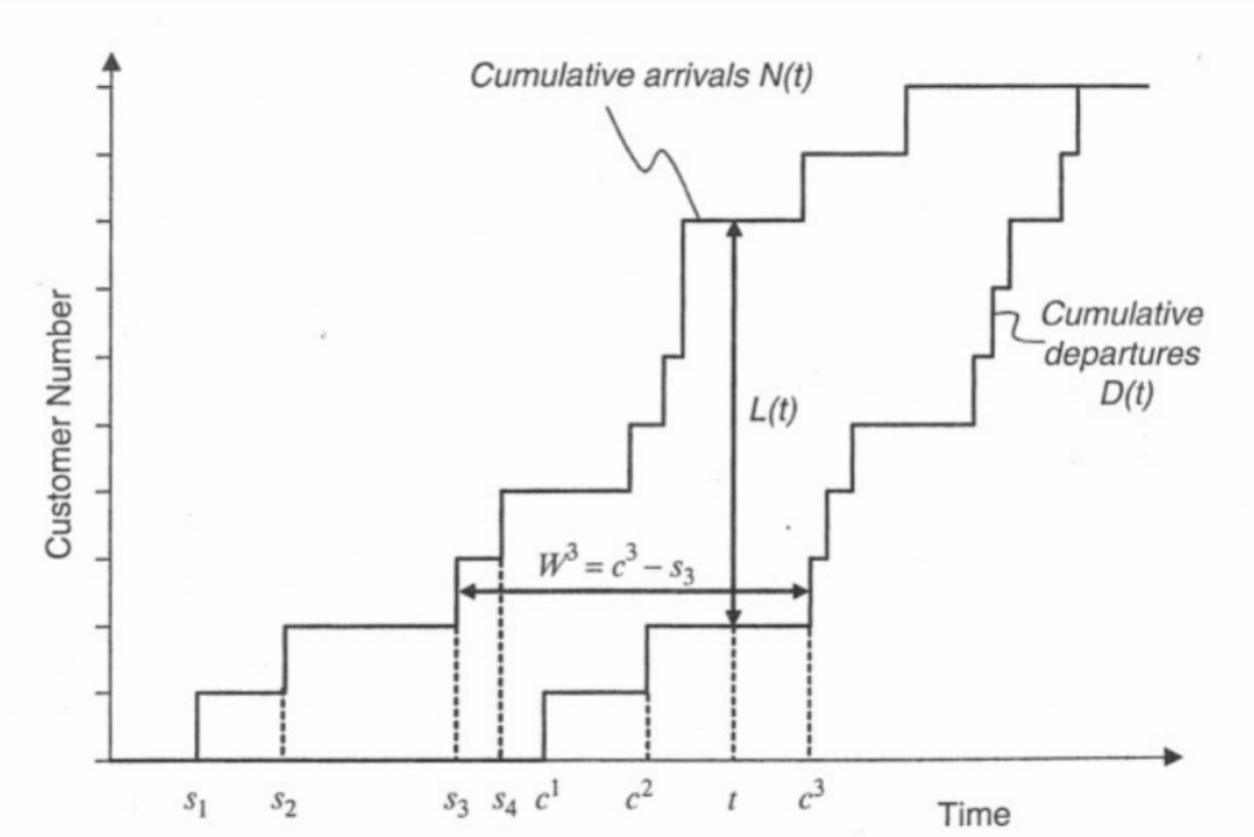

Figura 1.1: Um Diagrama de Chegadas/Partidas Cumulativas

Menciono a Lei de Little porque essa equação muitas vezes é apresentada como a justificativa matemática fundamental e irrefutável para o Kanban. E é. Só que não pelos motivos que você pensa. A pergunta de Frank me obrigou a uma imersão profunda na Lei de Little (LL) e foi o impulso para eu descobrir por que todos deveríamos realmente nos importar com essa

equação simples (Spoiler: a própria Lei de Little também **NÃO** é a parte mais importante do Kanban).

Permita-me dizer primeiro que uma explicação completa da LL está além do escopo deste livro (para uma discussão mais detalhada, consulte "Actionable Agile Metrics for Predictability"[1]); no entanto, precisarei dedicar alguns momentos para resumir alguns dos seus pontos relevantes.

A LL pode ser expressa como CT = WIP / TH, onde CT é o Tempo Médio de Ciclo do seu processo, WIP é o Trabalho em Andamento Médio do seu processo e TH é a Vazão Média do seu processo.

A LL é exata no seu cálculo, e esta equação pode ser aplicada a qualquer sistema de fluxo. Antes de explicar essa afirmação mais detalhadamente, façamos uma pausa por um segundo e tentar uma experiência. Se você conhece a TH, CT e WIP médios do seu processo (por exemplo, no último mês), gostaria que você inserisse esses números na equação LL agora. Experimente várias permutações diferentes. Talvez primeiro divida o seu WIP pela TH e veja se obtém o seu CT. Em seguida, tente multiplicar o seu CT pela TH para ver se obtém o seu WIP, etc. O que você observa? A minha suposição é que os seus números não estão saindo exatamente como você esperaria ou como previsto pela LL. Não apenas eles provavelmente estão incorretos, mas em alguns casos provavelmente estão consideravelmente errados.

O que está acontecendo aqui? O cálculo da LL é de fato exato, mas ele é exato apenas em contextos nos quais um conjunto específico de pressupostos é cumprido. Esses pressupostos são (para o período sob observação):

1. A taxa média de saída deve ser igual à taxa média de entrada.

2. Todos os itens que entram no sistema devem ser finalizados e sair do sistema.
3. A quantidade de WIP é aproximadamente a mesma no início e no final do intervalo de tempo sob observação.
4. A idade média do WIP não está aumentando nem diminuindo.
5. Unidades consistentes são usadas para a medição de TH, CT e WIP.

[**Nota**: A suposição #5 é fornecida aqui apenas para fins de completude, pois esta última suposição é trivial. Tudo o que estamos dizendo é que se você quiser medir CT em dias, então TH precisa ser medido por dia e o WIP médio deve ser medido por dia. Misturar unidades seria um erro básico (por exemplo, CT em semanas e TH em pontos de história), mas isso deveria ser intuitivamente óbvio. E se alguém na sua equipe tiver dificuldade com este ponto, então você tem problemas maiores do que como aplicar melhor a LL.]

Como os seus números calculados não saíram como previsto pela LL, isso nos diz que o seu processo violou explicita ou implicitamente uma, ou mais das suposições da LL pelo menos uma vez, e provavelmente em vários pontos ao longo do período que você escolheu para o cálculo. O efeito líquido da violação das suposições da LL é que você desestabilizou o seu processo — como evidenciado pela equação não funcionando.

A estabilidade do sistema (do ponto de vista da LL) é muito importante porque é impossível otimizar um processo que é inerentemente instável. A sua experiência lhe diz isso. Quão fácil é otimizar um processo em que o número de coisas com as quais você está trabalhando aumenta a cada dia? O quanto é fácil otimizar um processo em que todas as coisas com as quais você trabalha são bloqueadas por dependências em outras equipes? As suposições da LL, portanto, servem como um guia poderoso para as políticas que devemos implementar para prevenir que o nosso processo se desestabilize.

Sempre que você está aplicando o Kanban ao seu contexto, você se preocupa com todas as 5 suposições da Lei de Little (quer você saiba ou não).

E dessas cinco, há uma suposição que rege todas elas — como prometido pelo título deste capítulo.

O Elemento Mais Importante do Kanban: Idade do Item de Trabalho

Um entendimento completo do que significa violar cada uma das suposições da Lei de Little é fundamental para a otimização do seu processo de entrega. Então, gastaremos um minuto para examinar cada uma delas com um pouco mais de detalhe.

A primeira coisa a observar sobre as suposições é que #1 e #3 são logicamente equivalentes. Não tenho certeza porque o Dr. Little menciona separadamente cada uma delas, porque nunca vi um caso em que uma seja cumprida e a outra não. Portanto, acredito que podemos classificá-las com segurança como equivalentes. Mas, ainda mais importante, você notará o que Little **não** está dizendo aqui com as suposições #1 ou #3. Ele não faz nenhum julgamento sobre a quantidade real de WIP que precisa estar no sistema. Ele não diz que menos WIP é melhor ou que mais WIP é pior. Na verdade, Little não se importa com nenhum dos dois casos. Tudo o que importa é que o WIP seja estável ao longo do tempo. Portanto, embora terem chegadas correspondentes a saídas (e, portanto, WIP inalterado ao longo do tempo) seja importante, isso não nos diz *nada* sobre se temos muito WIP, pouco WIP ou a quantidade certa de WIP. As suposições #1 e #3, portanto, embora importantes, podem ser descartadas como *as* mais importantes.

A suposição #2 é frequentemente ignorada. No seu trabalho, com que frequência você começa algo, mas nunca o completa? O meu palpite é que o número de vezes que isso aconteceu com você nos últimos meses é maior que zero. Mesmo assim, enquanto essa suposição é novamente de importância crucial, é geralmente a exceção e não a regra. A menos que você se encontre em um contexto em que está sempre abandonando mais trabalho do que completa (caso em que você tem problemas muito maiores do que a LL), essa suposição também não será a razão dominante pela qual você tem um fluxo de trabalho subótimo.

O que nos deixa com a suposição #4. Permitir que itens envelheçam é arbitrariamente o maior fator para que você não seja eficiente, eficaz, nem previsível na entrega de valor ao cliente. Dito de outra forma, se você planeja adotar o Kanban (ou se já está praticando o Kanban), ***o aspecto mais importante ao qual você deve prestar atenção é não deixar que os itens de trabalho envelheçam desnecessariamente!***

Mais do que limitar o WIP, mais do que visualizar o trabalho, mais do que encontrar gargalos (o que, na verdade, não é realmente uma coisa do Kanban), a única pergunta a se fazer sobre o seu sistema Kanban é: você está permitindo que os itens envelheçam desnecessariamente?

Por Que Você Deveria se Importar com o Envelhecimento?

Antes de entrarmos no envelhecimento, precisamos dar um passo atrás e primeiro falar sobre Tempo de Ciclo (CT). A maioria das pessoas pensa que o motivo pelo qual o Kanban enfatiza tanto o CT é para pressionar as equipes ágeis a fazerem mais coisas mais rapidamente. Nada poderia estar

mais longe da verdade. O motivo pelo qual o Kanban se preocupa com o CT é porque o CT representa o tempo até o feedback do cliente.

Veremos em um capítulo posterior que, até que um item de trabalho esteja realmente nas mãos do cliente, esse item representa apenas valor hipotético. O valor só pode ser determinado pelos próprios clientes e essa determinação só pode ser feita após a entrega do item. Assim, o CT é realmente uma medida do “tempo até o feedback validado”.

No entanto, o CT em si só pode ser calculado quando o item é realmente concluído. Antes disso, tudo o que sabemos é a idade do item. Esse processo de envelhecimento começa imediatamente assim que o trabalho começa. Além disso, os itens de trabalho continuarão a envelhecer até serem entregues ao cliente. Portanto, quanto mais os itens envelhecem, mais tempo adiamos o valioso feedback do cliente.

Esse feedback atrasado aumenta as chances de algo dar errado na entrega. Talvez o ambiente de negócios mude, talvez os requisitos do cliente mudem, talvez uma pandemia global se instale — é impossível saber o que pode acontecer e que vai mudar as necessidades de um cliente. Mas o que sabemos é que uma idade mais longa representa um risco maior. E o risco final é que passemos muito tempo trabalhando em algo que acaba não sendo valioso.

Como o meu amigo e colega Prateek Singh gosta de dizer, “tudo se resume a descobrir o quão errado você está o mais rápido possível.” Ao permitir que os itens envelheçam desnecessariamente, você não está apenas sabotando a sua capacidade de entregar, mas também está sabotando a sua capacidade de entregar o que os seus clientes realmente querem.

Como Evitar que os Itens de Trabalho Envelheçam

Então, se o envelhecimento é tão prejudicial, como podemos evitar que isso aconteça?

Uma pergunta que adoro fazer nos meus workshops é “quais são as duas maneiras mais eficazes de evitar que os itens envelheçam desnecessariamente?” Esta pergunta deixa geralmente os participantes perplexos, pois eles tendem a recorrer aos dogmas ensinados anteriormente. Você receberá respostas como “diminuir o WIP” ou ”remover bloqueios”, ou algo semelhante. Mas, como acabamos de ver, essas respostas não levam necessariamente a uma idade mais curta.

A primeira maneira de evitar que os itens envelheçam é terminá-los. É tão simples assim. Se um item é concluído, ele não está mais envelhecendo. Podemos então iniciar o processo para obter o feedback do cliente.

A segunda (e provavelmente ainda melhor) maneira de evitar que os itens envelheçam é não os começar. Quantas vezes você e a sua equipe são pressionados a iniciar o trabalho quando ainda não estão prontos, apenas para parecer que estão progredindo? Do ponto de vista da LL, isso é absolutamente o pior que se pode fazer.

Agora vamos juntar tudo isso. Se você terminar o trabalho o mais rápido possível e não começar o trabalho até estar pronto para fazê-lo, o que você acabou de fazer? Você acertou, você acabou de controlar o Trabalho em Andamento.

A verdadeira razão para controlar o WIP é prevenir o envelhecimento desnecessário.

Podemos levar essa lógica um passo adiante e afirmar que todas as práticas do Kanban podem ser derivadas do princípio básico de que não queremos que os itens envelheçam desnecessariamente. Por que visualizar o trabalho? Para podermos ver onde o trabalho está se acumulando e os itens estão envelhecendo desnecessariamente. Por que marcar o trabalho como bloqueado? Para podermos, ver onde o fluxo não está acontecendo e os itens estão envelhecendo desnecessariamente. Por que implementar políticas de puxar? Para que não se permita que alguns furem a fila, o que faria com que outros itens envelhecessem desnecessariamente. E assim por diante.

Todas as práticas do Kanban podem ser derivadas da motivação singular de não querer que os itens envelheçam desnecessariamente.

Por último, mas muito importante de ser mencionado, é como evitar que os itens envelheçam muito: se um item está demorando muito para fluir, então o maior culpado provavelmente é que o item é muito grande. Uma das primeiras coisas que você deve considerar para um item “parado” é encontrar maneiras criativas de dividi-lo em vários itens menores. Tenha em mente que a ideia aqui não é dividir os itens apenas para tornar nossos números melhores. Pelo contrário, queremos encontrar maneiras de quebrar um trabalho grande e valioso em vários itens menores — mas ainda valiosos! – de trabalho. Em termos de fluxo, o que realmente estamos falando é do tamanho do lote. Muitas vezes, você pode estar trabalhando em um único item — uma única história, um único épico, uma única funcionalidade, seja lá o que for — mas, na verdade, está trabalhando em vários itens pequenos disfarçados de um item grande (as estratégias para dividir o trabalho serão exploradas com mais detalhes no Capítulo 3). O melhor sinal que você tem de que algo pode ser muito grande e, portanto, pode precisar ser dividido, é a sua idade. Ignorar a idade é muito arriscado.

Conclusão

Se você não está prestando atenção ao envelhecimento, você está perdendo a única razão real para fazer Kanban.

Em outras palavras, se o Kanban se trata de otimizar a entrega de valor ao cliente, então como você sabe realmente o quão ótimo você está? A resposta não está nos limites de WIP, nos CFDs, na Eficiência de Fluxo, na gestão de mudanças, ou em qualquer outra besteira que você possa ter sido alimentado até agora. A resposta está na sua capacidade de saber se os seus itens estão envelhecendo desnecessariamente ou não. Tudo o mais no Kanban deve ser subordinado a esse único objetivo.

No entanto, um item que está envelhecendo, por si só, não é necessariamente uma coisa ruim. A realidade é que todos os itens devem envelhecer em algum grau antes de poderem ser entregues. A pergunta que devemos fazer, portanto, é quanto tempo de envelhecimento é demais?

A resposta é a Expectativa de Nível de Serviço ou SLE. Os SLEs são mais um daqueles tópicos sobre os quais você provavelmente não ouviu falar muito. O SLE é tão fundamental, na verdade, que merece o seu próprio capítulo, e esse capítulo vem imediatamente a seguir…

Notas

1. Daniel Vacanti, “Actionable Agile Metrics for Predictability” (ActionableAgile Press, 2014)
2. Little, J. D. C., and S. C. Graves. “Little’s Law.” D. Chhajed,

T. J. Lowe, eds. Building Intuition: Insights from Basic Operations Management Models and Principles (Springer Science

+ Business Media LLC, New York, 2008)

3. Little, J. D. C., and S. C. Graves. “Little’s Law.” D. Chhajed,

T. J. Lowe, eds. Building Intuition: Insights from Basic Operations Management Models and Principles (Springer Science

+ Business Media LLC, New York, 2008)

Capítulo 2 - Expectativas de Nível de Serviço

Talvez você já tenha ouvido um mito sobre o Kanban que segue mais ou menos assim:

“Como o Kanban não tem *timeboxes*, os itens podem levar o tempo que precisarem para terminar.”

Ou talvez você tenha ouvido algo assim:

“Não podemos usar o Scrum porque não conseguimos terminar os itens em duas semanas. Então usamos o Kanban, porque o Kanban não exige que os itens terminem em duas semanas.”

Deixando de lado os equívocos sobre o Scrum na segunda declaração por um momento, ambas as citações acima estão incorretas na sua avaliação do Kanban. Itens de trabalho no Kanban não têm permissão para ficar em andamento para sempre e terminar apenas quando tivermos tempo para trabalhar neles. Isso é a antítese do fluxo. Fluxo implica movimento ou progresso. E se os itens estão apenas sentados e envelhecendo, então não há fluxo. Sem fluxo, não há Kanban.

Então, não, no Kanban os itens não podem levar o tempo que quiserem para terminar. Mas qual é a solução do Kanban para este problema? Bem, como sempre, é útil olhar as coisas do ponto de vista dos nossos clientes. Qual é a primeira pergunta que os nossos clientes nos farão assim que começarmos a trabalhar em algo para eles? Se você respondeu “Quando

estará pronto?”, então você ganha um prêmio. Concordando ou não, essa é uma pergunta razoável para nossos clientes fazerem. E precisamos de uma maneira de fornecer uma resposta a eles.

Se você pensar sobre isso, o que os nossos clientes realmente estão nos pedindo é uma previsão do futuro. Portanto, qualquer resposta que dermos a eles seria o equivalente a uma previsão. Uma coisa engraçada sobre o futuro, no entanto, é que ele tem esse hábito desagradável de estar cheio de incertezas. Apesar do que algumas pessoas possam dizer, ninguém pode prever o futuro com 100% de certeza. Quando a incerteza está envolvida em qualquer empreendimento, uma abordagem probabilística é justificada.

Por exemplo, antes de eu lançar esta moeda, me diga com 100% de certeza que ela cairá exatamente cara. Obviamente, você não pode dar 100% de certeza antes do lançamento, mas o que você pode dizer é: há uma chance de 50% de ser cara (e uma chance de 50% de ser coroa). Como outro exemplo, antes de eu rolar este dado de 6 lados, me diga com 100% de certeza que obterei exatamente um 3. Novamente, 100% de certeza não existe, mas sei que tenho cerca de 17% de chance de rolar um 3.

O mesmo princípio se aplica ao nosso trabalho. Uma vez que começo a trabalhar em um item, é impossível dizer com 100% de certeza exatamente quanto tempo levará para que esse item seja concluído. Mas o que posso fazer, é olhar para dados históricos para chegar a uma afirmação probabilística sobre quanto tempo deveria levar (por exemplo, “85% de chance de ser concluído em 12 dias ou menos”). A propósito, outra palavra para “afirmação probabilística sobre o futuro” é “previsão”.

Juntando tudo isso, quando os nossos clientes perguntam “quando estará pronto?”, precisamos responder a eles com uma previsão. No Kanban, a afirmação probabilística sobre quanto tempo levará para os itens individuais

serem concluídos, uma vez que tenham sido iniciados, é conhecida como Expectativa de Nível de Serviço ou SLE.

Segundo o Guia Kanban: “O SLE é uma previsão de quanto tempo deveria levar um único item de trabalho do início ao fim. O SLE em si tem duas partes: um período decorrido e uma probabilidade associada a esse período (por exemplo, "85% dos itens de trabalho estarão concluídos em oito dias ou menos"). O SLE deve ser baseado no tempo de ciclo histórico, e uma vez calculado, deve ser visualizado no quadro Kanban. Se não existirem dados históricos de tempo de ciclo, escolha o seu melhor palpite até que existam dados históricos suficientes para um cálculo adequado do SLE.”

O SLE serve duas funções no Kanban. Primeiro, ele fornece uma previsão de conclusão para os itens de trabalho assim que eles começam. Segundo o SLE nos ajuda a responder à pergunta que fizemos no final do último capítulo, ou seja, “quanto tempo de envelhecimento é demais?”.

Como calcular um SLE para o seu processo é detalhado extensivamente em AAMFP (Actionable Agile Metrics for Predictability) e WWIBD (When Will it Be Done?). Se você não está familiarizado com a derivação de um SLE, recomendamos que se familiarize com esse conceito antes de prosseguir, pois a nossa atenção aqui será focada nas praticidades de como usar o SLE uma vez calculado — especialmente em relação ao envelhecimento.

Percentis como Disparadores de Intervenção

À medida que os itens envelhecem, ganhamos informações sobre eles. Os percentis no nosso gráfico de dispersão funcionam como *checkpoints* perfeitos para examinar novas informações. Usaremos esses *checkpoints* para ser o mais proativos possível e garantir que o trabalho seja concluído de maneira oportuna e previsível.

Como isso funciona? Vamos falar primeiro sobre o 50.º percentil. E vamos presumir, para esta discussão, que a nossa equipe está usando um SLE de percentil 85. Uma vez que um item permanece em progresso até um ponto em que a sua idade é a mesma que o Tempo de Ciclo da linha do 50.º percentil, podemos dizer algumas coisas. Primeiro, podemos dizer que, por definição, este item agora é maior do que metade dos itens de trabalho que vimos antes. Isso pode nos dar motivo para pausa. O que descobrimos sobre este item que pode exigir que tomemos alguma ação? Precisamos nos concentrar nele? Devemos dividi-lo? Precisamos agilizar a remoção de um bloqueio? A urgência dessas perguntas se deve à segunda coisa que podemos dizer quando a idade de um item atinge o 50.º percentil. Quando inicialmente trouxemos o item de trabalho para o nosso processo, ele tinha 15% de chance de violar o seu SLE (essa é a definição exata de usar o 85.º percentil como um SLE). Agora que o item atingiu o 50.º percentil, a chance de ele violar o seu SLE dobrou de 15% para 30%. Lembre-se, quanto mais velho um item fica, maior a probabilidade de ele envelhecer ainda mais. Mesmo que isso não cause preocupação, pelo menos deve gerar uma conversa. É disso que se trata a previsibilidade acionável.

Quando um item envelheceu até a linha do 70.º percentil, sabemos que ele é mais antigo do que mais de dois terços dos outros itens que vimos antes. E agora a sua chance de não cumprir o seu SLE saltou para 50%. Lance uma moeda. As conversas que estávamos tendo anteriormente (por exemplo, trabalhar em pares, “swarm”, dividir o item) agora devem se tornar ainda mais urgentes. E elas devem continuar a ser urgentes à medida que a idade desse item de trabalho se aproxima cada vez mais do 85.º percentil. A última coisa que queremos é que esse item viole o seu SLE — mesmo que neste exemplo saibamos que isso vai acontecer 15% das vezes. Queremos garantir que fizemos tudo o que pudemos para evitar que uma violação ocorresse. O motivo para isso é que apenas porque um item violou o seu SLE, isso não significa que de repente devemos relaxar. Ainda precisamos concluir esse trabalho. Algum cliente em algum lugar está esperando que o seu valor seja entregue.

Dimensionamento Adequado (Right Sizing)

Mentimos para você anteriormente quando dissemos que o SLE serve duas funções no Kanban. Na verdade, ele serve três. A terceira função é auxiliar em uma prática conhecida como dimensionamento adequado. Existe um mito insistente do Kanban de que todos os itens a fluir pelo seu processo precisam ter o mesmo tamanho. Afinal, essa é a única maneira de garantir que os Limites de WIP façam sentido, certo? Errado. Não há nada na teoria do fluxo que exija que todos os itens que fluem por um sistema controlado por WIP sejam exatamente do mesmo tamanho. Na verdade, há toda uma teoria da variação que reconhece que não apenas os itens não precisam ser exatamente do mesmo tamanho, mas que também não há nada que você possa fazer para torná-los todos exatamente do mesmo tamanho – mesmo que você quisesse. Ou seja, a variação no tamanho dos itens de trabalho *sempre existirá*.

Portanto, a consequência da variação é que precisamos projetar um sistema que possa lidar elegantemente com o tamanho variável dos itens que eventualmente entrarão no nosso sistema. Mas existem limites para a quantidade de variação com que podemos lidar. Para ilustrar essa ideia, Frank Vega adora usar o exemplo de um triturador de madeira (qualquer pessoa que tenha visto o filme Fargo sabe exatamente do que estamos falando). Pense no que acontece quando você tenta enfiar um galho de árvore muito grande no triturador de madeira (à la Fargo). No mínimo, esse galho ficará preso. Na pior das hipóteses, esse galho quebrará o seu triturador. Da mesma forma, o que aconteceria se você pegasse um monte de serragem e jogasse toda essa serragem no triturador? Isso também entupiria as coisas. Mas esses são casos extremos. O triturador de madeira seria capaz de lidar razoavelmente com qualquer coisa entre serragem e um pequeno tronco de árvore. Qualquer galho que o triturador de madeira consiga manipular sem dificuldade é considerado do tamanho certo.

O mesmo é verdadeiro para o seu processo — o tamanho certo será a faixa de resultados possíveis, conforme ditado pela confiança percentual que você escolheu para o seu SLE. Por exemplo, se você estiver usando o percentil 85 como seu SLE e, no seu processo, o percentil 85 for de 12 dias ou menos, então o tamanho certo para os itens fluírem no seu sistema é de 12 dias ou menos.

Conclusão

Uma vez que você pegue o jeito de gerenciar o trabalho por idade, você terá percorrido cerca de 80% do caminho para poder otimizar o fluxo. Existem alguns detalhes que ainda precisamos abordar (ou seja, o restante deste livro), mas nada disso terá significado a menos que você compreenda o conceito de idade dos itens de trabalho!

Lembre-se, no entanto, é impossível para nós saber o quão grande é um item antes de começarmos a trabalhar nele. Conforme o trabalho progride, precisamos comparar continuamente a idade deste item com o nosso SLE, usando as linhas de percentil como disparadores, conforme discutido anteriormente. Apenas porque você achou que algo estava do tamanho certo quando começou não significa que realmente está. A única maneira de realmente saber é monitorar o envelhecimento de cada item que flui pelo seu sistema.

Mas monitorar é apenas metade da batalha — e nem mesmo a parte mais importante. A outra metade da batalha é agir assim que você tiver a informação de que um item está demorando muito para ser concluído. A melhor informação do mundo será inútil a menos que uma ação seja tomada. E é por isso que temos a prática Kanban #2, “Gerenciamento Ativo

dos Itens em Andamento”. Felizmente, falaremos sobre esse tópico em seguida.

Capítulo 3 - Gerenciamento Ativo de Itens em um Fluxo de Trabalho

Uma vez trabalhei com uma equipe que havia projetado um quadro Kanban com um limite global de Trabalho em Andamento de 9. Eles também haviam adicionado uma raia de urgente no seu quadro (um grande erro, como qualquer pessoa que acompanha o meu trabalho sabe), mas pelo menos definiram um Limite de Trabalho em Andamento de 1 para essa raia de urgente. Em uma certa manhã, antes da nossa reunião diária, cheguei ao quadro e vi que havia 32 itens na raia de urgente. Este é um exemplo clássico de falta de gerenciamento ativo dos itens em um fluxo de trabalho.

Qual é o ponto de definir um fluxo de trabalho, estabelecer limites de WIP, concordar com políticas de puxar, etc., se você simplesmente vai ignorá-los? A segunda prática do Kanban, “Gerenciar Ativamente Itens em um Fluxo de Trabalho”, é onde a teoria do fluxo encontra a prática. O Kanban trata de fluxo e o fluxo trata de movimento, e o movimento não acontece por si só. Como disse Deming, um sistema precisa de gerenciamento[1], e sistemas baseados em fluxo não são diferentes.

O gerenciamento ativo de itens em um fluxo de trabalho pode assumir várias formas, incluindo, mas não se limitando a:

- Controle de WIP
- Evitar que os itens de trabalho se acumulem em qualquer parte do fluxo de trabalho

- Garantir que os itens de trabalho não envelheçam desnecessariamente, usando o SLE como referência
- Desbloquear itens de trabalho bloqueados

Como discutido no Capítulo 1, a sua principal ferramenta para o gerenciamento ativo de WIP é monitorar a idade. Ou seja, a única maneira de sabermos, com um nível razoável de certeza, se os itens não estão fluindo como esperamos, é observar a idade. Mas, como também vimos no Capítulo 2, a melhor informação do mundo não serve para nada a menos que você faça algo com ela.

Segue uma análise mais detalhada de algumas ações que você pode tomar quando perceber que os itens estão demorando muito para serem concluídos. Esta lista não é de forma alguma exaustiva, mas são pontos ótimos para começar.

Trabalho em Par, Swarm e Mobbing

Os itens de trabalho frequentemente revelam complexidades, antes desconhecidas, à medida que avançam no fluxo de trabalho. Essa complexidade pode fazer com que eles envelheçam mais do que outros itens. Um item que envelheceu a ponto de se destacar no contexto do fluxo da equipe merece alguma atenção especial. Isso pode se dar na forma de vários membros da equipe ajudando nesse item (cuidado com a Lei de Brooks). Muitas vezes isso significa reduzir o WIP para ajudar o item que está envelhecendo a progredir. Pessoas que estão concluindo o próximo item devem ser solicitadas a ajudar no item de trabalho envelhecido em vez de pegar itens novos. O ato de reduzir o WIP para abaixo do número de membros da equipe é conhecido por vários nomes — Trabalho em Par,

“Swarm” e “Mobbing”, para citar alguns. Para facilitar a referência, chamaremos todas essas práticas de trabalho em conjunto.

Existem três principais maneiras pelas quais o trabalho em conjunto pode ajudar a controlar a idade:

- Completando tarefas *downstream* mais cedo – À medida que um item envelhece em uma etapa anterior, podemos possibilitar um fluxo mais rápido nas etapas posteriores. Podemos executar etapas nas fases posteriores mais cedo para que a tarefa não continue envelhecendo desnecessariamente uma vez que tenha passado pela etapa atual.
- Dividindo tarefas do item de trabalho entre os membros da equipe – Se o próprio item de trabalho não puder ser dividido em partes entregáveis, é possível identificar subtarefas do item. Diferentes membros da equipe podem assumir as diversas subtarefas em paralelo para ajudar o item a avançar.
- Removendo pontos de bloqueio – Muitas vezes, obter novas perspectivas sobre um problema, que um único membro da equipe tem enfrentado, ajuda a encontrar soluções criativas. Seja por meio de técnicas como “discutir com um pato de borracha (rubber ducking)” ou trabalho em par interfuncional para obter novas perspectivas, elas ajudam um item envelhecido a progredir.

Desbloqueio de Trabalho Bloqueado

Por definição, qualquer trabalho que esteja bloqueado ou suspenso não está fluindo. Esses itens de trabalho envelhecem, geralmente devido a dependências internas ou externas. Se a causa for uma dependência interna, precisamos examinar as nossas políticas de processo e buscar melhorias. Se a causa for externa, precisamos descobrir como reduzir a probabilidade de dependências externas para futuros itens ou reduzir o

impacto dessas dependências na idade. Em outras palavras, como podemos nos aproximar da eliminação da dependência ou tornar o tempo de resolução insignificante. Trazer especialistas externos para dentro da equipe, melhorar os relacionamentos com parceiros/fornecedores ou eliminar completamente a dependência são todas opções que podemos avaliar. Seja a dependência interna ou externa, precisamos estabelecer algumas políticas sobre como tratamos o trabalho bloqueado. Existem pelo menos três níveis de bloqueio que precisam ser estabelecidos:

- Quando marcar um item como bloqueado – Quanto tempo precisa se passar antes de marcarmos um item, cujo progresso está interrompido, como bloqueado? Isso está na ordem de horas, dias ou semanas?
- Itens Bloqueados e Limites de WIP — Por quanto tempo um item bloqueado deve contar para nossos limites de WIP e nos impedir de pegar outros trabalhos? Incluir isso no WIP aumenta o nosso foco em resolvê-lo?
- Removendo itens bloqueados do sistema — Em que momento decidimos que o item está bloqueado há tanto tempo que pode não ser mais relevante rastreá-lo? Devemos cancelar o item ou movê-lo de volta para o backlog?

Reexaminando o Dimensionamento Adequado

Leia o livro "Principles of Product Development Flow" de Don Reinertsen e você logo perceberá que um dos maiores prejuízos para o fluxo é trabalhar em itens muito grandes. Em termos de fluxo, isso significa controlar o tamanho do lote. Vimos anteriormente que geralmente quando um item está preso no seu processo é porque ele é muito grande — não foi dimensionado adequadamente.

O dimensionamento adequado é a arte de permitir que o trabalho flua em pequenos lotes de valor em todos os níveis. Isso significa dividir as coisas em pequenos pedaços gerenciáveis.

Para dimensionar, tudo o que você precisa fazer antes de puxar um item para o seu processo, é ter uma rápida conversa sobre se você consegue – com base no que você sabe agora – concluir o item em 12 dias ou menos. Se a resposta for sim, então a conversa acabou, você puxa o item e começa a trabalhar nele. Se a resposta for não, então você discute como pode redefinir o item de trabalho para que ele tenha o tamanho certo. Talvez você precise dividi-lo. Possivelmente você precise ajustar os critérios de aceitação (mais sobre a divisão de itens na próxima seção). Seja qual for o caso, tome a ação necessária antecipadamente e só puxe o item quando estiver 85% confiante de que pode concluí-lo em 12 dias ou menos.

Prateek Singh orientou uma de suas equipes com este guia sobre dimensionamento adequado:

“Temos uma ideia geral de quão grandes foram os itens de trabalho em cada nível no passado. Podemos usar isso para 'dimensionar' os próximos itens. Atualmente, temos uma ideia ótima do dimensionamento no nível de história devido aos dados disponíveis. Também vamos estabelecer diretrizes no nível Épico com base na compreensão histórica do fluxo.

“No gráfico de dispersão do Tempo de Ciclo para nossa equipe, notamos que 85% das histórias nas quais trabalhamos são concluídas em 11 dias ou menos. Isso é um guia para o dimensionamento adequado. Sempre que a equipe pegar a próxima história, eles devem ser capazes de se perguntar: ‘Este é o menor valor possível e pode ser concluído em 11 dias ou menos?’ Se a resposta para essas perguntas for sim, ótimo, não é necessário mais estimativas, comece a trabalhar nela. Se a resposta for não, tentaremos dividir esta história. Esta é a essência do dimensionamento adequado. Cada

equipe vai descobrir o tamanho certo das suas histórias com base nos seus próprios dados.

“Para Épicos – como não temos ótimos dados, mas uma ideia geral razoável a esse respeito, estamos emitindo algumas orientações. Os Épicos devem ter 10 histórias ou menos, 90% do tempo. 10 é um número flexível, é uma intenção. A realidade é que haverá Épicos que se tornarão grandes o suficiente para ter 11/12 histórias. 90% dos nossos Épicos devem ter 10 histórias ou menos. Isso não significa que tentamos forçar todos os Épicos a ficarem próximos de 10 histórias. A parte ‘ou menos’ é importante. Se um Épico puder ser entregue a um cliente em 3 histórias, ótimo, vamos deixá-lo assim. 10 é um limite superior flexível, não uma meta.”

Isso tudo é ótimo, você pode estar pensando, mas como procedemos para dividir os itens, uma vez que reconhecemos que eles não foram dimensionados adequadamente? Fico feliz que você tenha perguntado.

Dividindo Itens de Trabalho

Quando você descobre que um item é muito grande para o seu processo, a sua primeira hipótese deve ser que este grande item de trabalho é provavelmente composto por partes menores, individualmente entregáveis, de valor. Agrupar vários itens em um único item de trabalho pode negar muitos dos benefícios proporcionados pela limitação do Trabalho em Andamento. Por exemplo, se estamos operando com um limite de WIP de 1, mas esse único item poderia potencialmente ter sido 5 itens separados, o nosso WIP, na verdade, é cinco vezes maior do que o que é visível. O WIP muitas vezes está oculto nos itens de trabalho. Para expor o nosso WIP real, devemos quebrar o trabalho em peças individuais 'passíveis de feedback' cedo e frequentemente.

Vamos examinar algumas estratégias simples que podem ser usadas para decompor o trabalho. É uma lista que eu e outros colegas já usamos com sucesso no passado para decompor o trabalho. Essas estratégias são eficazes em todos os níveis de trabalho — histórias, funcionalidades ou iniciativas. Muitas dessas estratégias podem ser aplicadas ao mesmo item de trabalho também. O objetivo final é criar unidades de trabalho menores que possam nos ajudar a obter feedback mais rápido dos nossos clientes.

Critérios de Aceitação

A prática de adicionar Critérios de Aceitação do usuário (CA) nos ajuda a entender como um cliente espera se beneficiar do item de trabalho. Se a equipe trabalha em iniciativas que se desdobram em funcionalidades, que por sua vez são compostas por histórias, cada um desses níveis deve ter critérios de aceitação. Em cada nível, devemos ser capazes de decompor o trabalho, chegando cada vez mais perto de uma proporção de 1:1 entre os critérios de aceitação e o item de trabalho. Isso não significa que cada história, funcionalidade ou iniciativa deva ter apenas um critério de aceitação. Em vez disso, devemos observar que cada item de trabalho tenha o número mínimo de CA que nos ajude a obter feedback.

Por exemplo, considere o seguinte item de trabalho.

Como revisor, quero ver as seções relevantes de um artigo submetido, separadas para que eu possa avaliá-las facilmente.

- CA 1 - Os revisores devem conseguir ver o título do artigo separado do corpo
- CA 2 - Os revisores devem conseguir ver a contagem de palavras da descrição
- CA 3 - Os revisores podem dar notas para cada seção separadamente

- CA 4 - Os revisores podem visualizar a hipótese principal em uma seção separada
- CA 5 - Os revisores podem opcionalmente separar os detalhes do experimento para serem avaliados separadamente

Neste caso, cada um desses CAs pode ser um item de trabalho separado. Todos eles podem ser entregues independentemente aos clientes (internos ou externos) para obtermos feedback. Cada um desses CAs resolve um problema do cliente e entrega valor sem ser impedido de terminar pelos outros.

Conjunções e Conectores

Muitas vezes, o WIP se esconde nos títulos dos itens de trabalho. Sempre que vemos conjunções, especialmente “e” e “ou”, em um título de item de trabalho, é um indicativo de que o item pode ser dividido. Às vezes, essas conjunções são substituídas por vírgulas, travessões e barras também. Essas conjunções e conectores geralmente tentam agrupar várias ações ou atores em um único item de trabalho. Separá-los pode nos ajudar a entregar ações individuais mais cedo.

Por exemplo, o item de trabalho mostrado abaixo pode ser dividido em vários itens:

- Os clientes devem poder avaliar os restaurantes sobre a qualidade da comida, o serviço, o ambiente e a experiência completa.

Este item de trabalho pode facilmente se tornar quatro itens de trabalho diferentes, conforme listado abaixo. A entrega da experiência completa pode ser parte do produto mínimo viável, enquanto os outros podem ser aprimoramentos posteriores.

- Os clientes devem poder avaliar a qualidade da comida dos restaurantes.
- Os clientes devem poder avaliar o serviço dos restaurantes.
- Os clientes devem poder avaliar o ambiente dos restaurantes.
- Os clientes devem poder avaliar a experiência completa nos restaurantes.

Termos Genéricos ou Plurais

Termos genéricos são frequentemente usados nos títulos dos itens de trabalho para representar o que, de outra forma, seriam múltiplos requisitos. Procurar obter mais especificidade pode ajudar a identificar os itens de trabalho menores, que estão escondidos no item maior. Maior especificidade também leva a planos de teste melhores, com menos casos de teste ausentes. De fato, falar sobre como testar um item muitas vezes é uma ótima maneira de passar de itens genéricos para específicos.

- O sistema deve sinalizar itens com cores apropriadas com base na gravidade.

No item acima, os termos genéricos “cores” e “gravidade” podem ser tornados mais específicos para criar itens de trabalho menores que podem então ser priorizados. Esses itens podem ser os seguintes:

- O sistema deve sinalizar eventos com impacto superior a 90% em vermelho.
- O sistema deve sinalizar eventos com impacto entre 50% e 90% em amarelo.
- O sistema deve sinalizar eventos com impacto inferior a 50% em verde.

Otimizar Agora vs. Depois

Esta estratégia se concentra em uma abordagem de incremento mínimo viável em todos os níveis. Qual é a solução mais simples e menos otimizada que podemos entregar para começar a obter feedback dos clientes internos ou externos? Os clientes podem querer muito mais do que o primeiro entregável, mas nos alinhamos cada vez mais e entregamos valor com cada incremento. O exemplo abaixo mostra como isso pode ser feito.

- Fornecer aos clientes a capacidade de pedir pizzas listadas em nosso cardápio online.

O item de trabalho acima pode ser dividido em algumas tarefas menores:

- Fornecer aos clientes um botão para pedir pizza Margherita online.
- Fornecer aos clientes a capacidade de alterar o tamanho da pizza pedida.
- Fornecer aos clientes a capacidade de pagar pela pizza online.
- Fornecer aos clientes a capacidade de selecionar entre uma lista de pizzas online.
- Permitir que os clientes adicionem/removam coberturas dos pedidos de pizza.
- Permitir que os clientes montem sua própria pizza adicionando coberturas a uma pizza básica de queijo.

Existem várias estratégias que podem ser usadas para dividir o trabalho. Essas técnicas podem ser usadas em conjunto umas com as outras para permitir uma entrega de valor mais rápida e frequente. Elas devem ser utilizadas em todos os níveis de definição do trabalho.

A tabela a seguir é um conjunto de dicas que Becky McKneeley montou enquanto trabalhava na Ultimate Software. Ele representa um excelente guia inicial para a divisão de itens de trabalho.

Estratégia	Quando usar	Perguntas a serem feitas
Critério de Aceitação	Critérios de Aceitação individuais adotam os princípios do acrônimo INVEST e podem ser divididos	Algum dos CAs adota os princípios do acrônimo INVEST individualmente? Todos os CAs são necessários para se receber feedback (de alguém)?
Conjunções e Conectores	Procure palavras de conexão (e, ou, se etc.), vírgulas, travessões, barras	Existem conjunções ou conectores no título da história? Existem conjunções ou conectores em algum dos CAs? Essas conjunções/conectores podem ser divididas para recebermos feedback mais cedo?
Termos Genéricos & Plurais	Procure por substantivos comuns ou outras palavras genéricas que poderiam ser substituídas por algo mais específico. Busque por plurais que podem ser divididos (páginas, campos etc.)	Algum dos termos listados podem ser mais claramente identificados? Os plurais usados podem ser divididos e processados separadamente?
Papel do Usuário ou Persona	Mais de um papel poderá ser impactado e a funcionalidade funciona de forma diferente para cada papel	Quais papéis estão envolvidos na história? Papéis diferentes são impactados de forma diferente pela funcionalidade?
Regras de Negócio	Podem ser difíceis de serem descobertas – mas pense em casos de teste. Casos de teste muitas vezes implicam em ou são sinais de regras de negócio que podem ser quebradas em histórias individuais	Quais regras de negócio se aplicam a essa história? Regras mais simples podem ser suficientes para receber algum feedback sobre a funcionalidade? Quais cenários de teste podem ser usados para verificar essa história?
Opções de Plataforma	Oferecer suporte a plataformas diferentes. Por exemplo, celular vs. tablet, iOS vs. Android. Também pode ser usado para compatibilidade entre navegadores.	Quais plataformas e/ou navegadores precisam ser suportados? Todas essas plataformas e/ou navegadores são necessários inicialmente?
Processamento de Exceções	Caminho Feliz vs. Infeliz	Qual é o caminho feliz? Existem exceções e/ou casos extremos identificados na história?
Operações	Baseado nas diferentes operações executadas. Típico no gerenciamento de entidades como clientes etc. (CRUD)	Quais operações estão envolvidas nessa história? É necessário que todas as operações sejam executadas de uma vez para receber algum feedback?
Passos do Fluxo de Trabalho	Identificar os passos no fluxo de trabalho e implementá-los nos estágios do fluxo de trabalho	Quais são os passos do fluxo de trabalho nessa história? Os passos do fluxo de trabalho podem ser divididos e ainda assim permitir algum feedback sobre a funcionalidade?
Otimizar Agora vs. Depois	** Otimização Funcional **. Também conhecido como do Simples ao Complexo. Implementar funcionalidades simples que oferecem algum valor. Funcionalidades mais complexas serão adicionadas mais tarde.	A funcionalidade pode ser simplificada e ainda oferecer valor/feedback? Existem outras formas de lidar com essa funcionalidade que permitem feedback mais cedo?
Grande Esforço	Um esforço significativo é necessário para a primeira história. Novos aspectos técnicos, etc.	Estamos implementando alguma tecnologia nova? Existe alguma parte da história que não sabemos como lidar tecnicamente ou é nova para a EQUIPE?

Figura 3.1 Exemplos de Estratégias para a Quebra de Itens de Trabalho

Conclusão

O mundo vai conspirar contra você para tornar o seu processo o mais imprevisível possível. É necessário diligência para identificar quando os itens estão demorando demais, mas, mais importante ainda, é necessária diligência para agir prontamente. O mundo lá fora é de fato um lugar desagradável, e você deveria agradecer sua divindade favorita por ter o Kanban ao seu lado.

Não seria bom se o envelhecimento por si só nos permitisse ver todos os males do nosso processo? Mas, não é o caso. Infelizmente, o envelhecimento em si depende de como definimos nosso fluxo de trabalho, e assim também são as nuances de um fluxo de trabalho definido que discutiremos a seguir.

Notas

1. W. Edwards Deming, "Out of the Crisis" (The MIT Press, 2000)

Capítulo 4 - Definindo e Visualizando um Fluxo de Trabalho

Definição de Kanban pré-pandemia de COVID-19: “Nós temos *Post-Its* em um quadro branco”.

Definição de Kanban pós-pandemia de COVID-19: “Nós temos um quadro configurado no Jira”.

Gostaria de saber de onde vem o equívoco de que o Kanban se trata apenas de visualizar o trabalho. Dizer que o Kanban é apenas sobre visualização é como dizer que a Escócia se resume apenas a uísque ou que Nadal é apenas um jogador de quadra de saibro. Se você pensa que a primeira prática do Kanban “Definir e Visualizar um Fluxo de Trabalho” significa “ter um quadro no Jira”, então você está profundamente enganado. Esperamos que, até o final deste capítulo, possamos convencê-lo do contrário.

O propósito fundamental de um quadro Kanban não é apenas visualizar o trabalho, mas visualizar um conceito muito mais amplo conhecido como Definição de Fluxo de Trabalho (DoW). Um DoW tem o objetivo de abranger todos os aspectos de como o valor potencial flui pelo seu sistema para ser convertido em valor tangível quando entregue aos seus clientes. Especificamente, no mínimo, um DoW deve incluir:

- Uma definição das unidades individuais de valor que estão se movendo através do fluxo de trabalho. Essas unidades de valor são referidas como itens de trabalho (ou itens).
- Uma definição para quando os itens de trabalho são iniciados e concluídos dentro do fluxo de trabalho. Seu fluxo de trabalho pode ter mais de um ponto de início ou conclusão, dependendo do item de trabalho.
- Um ou mais estados definidos pelos quais os itens de trabalho passam de iniciados para concluídos. Qualquer item de trabalho entre um ponto de início e um ponto de conclusão é considerado Trabalho em Andamento (WIP).
- Uma definição de como o WIP será controlado entre iniciado e concluído.
- Políticas explícitas sobre como os itens de trabalho podem fluir por cada estado, entre iniciado e concluído.
- Uma expectativa de nível de serviço (SLE), que é uma previsão de quanto tempo deve levar para que um item de trabalho flua desde iniciado até concluído.

Uma breve observação sobre valor: neste livro, sempre que digo “valor”, provavelmente estou me referindo a “valor potencial”. Na maioria, mas não em todos os casos, o estado ideal da sua implementação do Kanban é que você entregue itens aos seus clientes, valide que o que foi entregue é valioso e, em seguida, faça quaisquer melhorias no seu processo com base no que foi ou não validado. Seria um pouco difícil para mim substituir “valor” por “valor potencial” em todas as instâncias, então por favor entenda que, a menos que estejamos falando de um contexto em que o valor tenha sido explicitamente validado, o que realmente queremos dizer é valor potencial.

Itens de Trabalho

Como o objetivo principal do Kanban é otimizar a entrega de valor, é razoável começarmos nossa discussão sobre como esse valor é encapsulado em nosso sistema. A maioria das implementações Lean-Agile já possui alguma noção de como o valor é capturado para ser trabalhado. Ferramentas como histórias de usuário, épicos, funcionalidades, etc., têm o propósito de tornar explícito o que é uma unidade individual de valor em nosso contexto. O Kanban não se importa com qual ferramenta você usa para definir valor em seu sistema, tudo o que importa no Kanban é que você tenha alguma maneira de descrever valor em unidades discretas. Essas unidades individuais de valor que fluem pelo seu sistema são chamadas de itens de trabalho (ou simplesmente itens).

Não há nenhum outro requisito sobre itens de trabalho no Kanban além de sua existência em algum lugar do seu sistema. O Kanban não se importa se você usa histórias, ou épicos, ou funcionalidades, ou qualquer outra coisa. O Kanban não se importa se eles foram definidos por um product owner, ou ordenados em algum backlog, ou refinados por uma equipe (embora você possa escolher fazer qualquer uma dessas coisas e mais se desejar). Do ponto de vista do Kanban, tudo o que você precisa é de uma maneira de encapsular o que você considera valor em um item que pode fluir através do seu processo, ser entregue ao seu cliente e ser validado por ele.

Fluxo de Trabalho

Falando em fluxo, uma vez que os itens tenham sido definidos, eles precisam de um conjunto de etapas de trabalho que agregam valor, pelas quais possam passar para se tornarem algo tangível.

Iniciado e Concluído

O primeiro passo para identificar seu fluxo de trabalho é desenhar limites ao redor dele. Isso significa ter pelo menos um ponto claramente definido em que você considera o trabalho iniciado e ter pelo menos um ponto claramente definido em que você considera o trabalho concluído. Então, digamos que você tenha um fluxo de trabalho que seja composto por Opções → Descoberta → Construção → Validação → Concluído. Podemos definir nosso ponto de início quando um item passa para a fase de Descoberta e podemos definir o ponto de conclusão quando um item passa para a etapa de Concluído. Ou podemos definir o início como Construção e a conclusão como Validação. Os pontos de início e conclusão corretos dependem totalmente do seu contexto e o Kanban não se importa como você os define, contanto que o faça.

Isso, na verdade, é um pouco mais complicado porque no Kanban é perfeitamente permitido ter mais de um ponto de início e mais de um ponto de conclusão. No exemplo acima, talvez queiramos medir o início tanto da Descoberta quanto da Construção, e a conclusão tanto da Validação quanto do Concluído. Porque podemos querer fazer isso dependerá das perguntas que você deseja responder com base nas métricas que coleta. Por exemplo, talvez queiramos saber quanto tempo leva para concluir itens desde quando começamos a trabalhar neles (Descoberta até Concluído), ou então quanto tempo leva nosso passo de Construção (da Construção até a Validação).

Incentivamos a experimentação com diferentes pontos de início e conclusão em seu processo para entender melhor seu contexto. A única coisa a lembrar é que o Kanban exige que você tenha a definição de pelo menos um de cada. O restante dos elementos do DoW, bem como as métricas básicas de fluxo que você acompanhará, serão definidos em termos de sua decisão sobre começado/concluído (uma discussão detalhada das métricas de fluxo será abordada em um capítulo posterior).

Estados do Fluxo de Trabalho

Uma vez que você tenha decidido seus pontos de início/conclusão, o próximo passo é mapear os estados discretos, que agregam valor, entre esses dois pontos. O número de estados que você escolher será específico do contexto, mas você deve ter pelo menos um entre iniciado e concluído. Contrariando a crença popular, um fluxo Fazer → Fazendo → Feito (onde iniciamos no estado Fazendo e concluímos no estado Feito) é um fluxo de trabalho perfeitamente válido no Kanban.

Exatamente como selecionar quais estados devem existir em seu fluxo de trabalho está muito além do escopo deste livro. No entanto, é preciso ter em mente alguns pontos importantes.

Em primeiro lugar, existem algumas diretrizes muito básicas sobre como identificar diferentes estados do fluxo de trabalho (mas tenha em mente que não há regras rígidas e rápidas):

1. Qualquer atividade que agregue valor, como “Descoberta”.
2. Quando você deseja modelar uma transferência explícita entre duas pessoas ou grupos, como separar um estado de “Design” em "Design" e “Revisão de Design” (assumindo que as pessoas que fazem a revisão são diferentes das pessoas que fizeram o design).
3. Quando você deseja rastrear métricas para uma atividade, como no exemplo acima, talvez a revisão não seja feita por um grupo separado, mas você deseja saber quanto tempo leva para que uma revisão aconteça.
4. Quando você geralmente deseja trazer mais transparência para uma atividade específica.

É uma prática típica (embora não seja um requisito) que cada estado escolhido para o seu fluxo de trabalho se torne uma coluna em seu quadro Kanban. Mais sobre isso abaixo.

Em segundo lugar, sou um grande defensor do princípio KISS (Keep It Simple, Stupid — Mantenha Simples, Estúpido). Especialmente quando engenheiros estão envolvidos, é muito comum projetar um fluxo de trabalho excessivamente complicado. Não há uma regra geral em torno do número exato de estados que você deve ter em seu fluxo de trabalho, mas gostaria que você desconfiasse sempre que alguém quiser adicionar um estado ou coluna. Mais estados dificultam o entendimento do que está realmente acontecendo em um fluxo de trabalho, e ter mais estados geralmente são uma desculpa para aumentar o Trabalho em Andamento (a propósito, isso é uma coisa ruim).

Terceiro, como você nomeia suas colunas pode ajudar ou prejudicar a colaboração. Geralmente, é uma prática ruim nomear estados ou colunas se baseando em funções específicas na sua organização — “Análise”, por exemplo. Uma razão para isso é que, se uma coluna representa uma função, pode haver uma tendência de as pessoas se isolarem em uma parte específica do fluxo de trabalho e relutarem em ajudar em outras áreas. Quando se trata de nomear colunas especificamente, o que você rapidamente perceberá é que o nome em si importa muito menos do que o entendimento compartilhado da organização sobre qual atividade está sendo realizada nessa etapa. Uma das melhores maneiras de facilitar esse entendimento é tornar as políticas explícitas para cada estágio do fluxo de trabalho. O que nos leva ao próximo elemento do DoW

Políticas Explícitas

As políticas são as regras pelas quais você joga o seu jogo de processo. Pense em como o jogo de beisebol seria drasticamente diferente se cada

rebatedor pudesse ter apenas 1 “strike” em vez de 3, ou se os jogos terminassem em três “innings” em vez de nove. Para vocês fãs de críquete por aí, pensem em como uma partida de teste é diferente de uma partida de um dia, que é diferente de uma partida de twenty20. O que leva a essas diferenças massivas e a uma mudança nos resultados são nossas políticas ou regras de processo.

Alguns exemplos de políticas que existem dentro do seu processo (quer você perceba ou não):

1. Regras sobre como lidar com itens bloqueados.
2. A ordem na qual você move os itens através do seu quadro.
3. O que significa para um item ser concluído em uma coluna específica.
4. Quando, ou se é aconselhável, dividir um item em itens menores.
5. Se os itens podem fluir para trás ou não (a propósito, não há regra no Kanban que impeça os itens de fluir para trás – mas o fluxo para trás pode não ser ideal em muitos contextos).

Tornar essas políticas explícitas pode ajudar bastante a eliminar a confusão sobre como o trabalho deve fluir em um determinado sistema. Na verdade, se você acertar essas políticas, coisas que você pensou que poderiam ser importantes – como nomes de colunas, número de colunas, ou qualquer outra coisa – podem se tornar irrelevantes.

Existe uma política que deixamos de mencionar, mas que pode ser a mais importante de todas. Aqueles acompanhando de casa, sabem que obviamente estamos nos referindo à política em torno de como o WIP é controlado.

Controle de WIP

Como você chamaria uma rodovia que está operando a 100% de sua capacidade? De que forma você chamaria uma rede que está operando a 100% de sua capacidade? De que maneira você chamaria uma fila do Starbucks que está operando com mais de 100% de sua capacidade?

Experimentamos problemas com o fluxo todos os dias de nossas vidas e esses problemas geralmente são diretamente atribuíveis a não controlar efetivamente o WIP. O trabalho do conhecimento não é diferente. Em geral, uma pessoa, equipe ou organização é muito mais eficiente ao trabalhar em uma ou duas coisas de cada vez do que ao trabalhar em 100 ou 200 coisas ao mesmo tempo. Em outras palavras, se você deseja habilitar o fluxo, em algum momento terá que considerar o controle do WIP.

Assim como em cada uma dessas seções, os detalhes completos sobre como controlar o WIP estão muito além do escopo deste livro, mas temos espaço para alguns pontos-chave.

Primeiro, o Kanban não se importa como você controla o WIP. Ele apenas se importa que você o faça. Não deixe ninguém lhe dizer que o Kanban requer que cada coluna em seu quadro tenha um limite de WIP. Isso simplesmente não é verdade. Você pode controlar o WIP da maneira que desejar. Você pode ter um único limite para o quadro inteiro, pode ter um limite por pessoa, pode agrupar limites em várias colunas e/ou pode ter um limite em cada coluna. E isso, apenas para citar alguns exemplos. Seu DoW requer que você tenha uma política explícita sobre como o WIP deve ser controlado, mas a implementação desse controle é completamente sua.

Segundo, controlar o WIP é apenas o primeiro passo. Você também precisa decidir como vai responder quando estiver acima ou abaixo do seu limite de

WIP. Você precisará decidir coisas como “os itens bloqueados devem contar no cálculo do limite de WIP?”.

Terceiro, no início, tenha em mente que o controle de WIP que você escolheu é muito mais um gatilho para uma conversa, do que uma regra rígida. Não se preocupe muito se você está acima ou abaixo de um limite, mas preste atenção se esses limites nos incentivam a fazer as perguntas certas mais cedo. Se suas conversas estão acontecendo tardiamente ou muito cedo, esses são bons indicadores de que um limite de WIP possivelmente precisa mudar.

A Expectativa de Nível de Serviço

Você deve ter lido tudo sobre SLEs no Capítulo 2, então não há mais nada a dizer aqui além de que os SLEs são um membro de primeira classe do seu DoW. Se você não está usando SLEs, então você não está fazendo Kanban.

O Quadro Kanban

A implementação visual coletiva de cada elemento do seu DoW é chamada de Quadro Kanban. Mesma música, diferente estrofe: como você escolhe visualizar seu DoW é uma escolha completamente sua. Você pode usar um quadro branco, pode usar uma ferramenta virtual, pode usar uma planilha – o Kanban não se importa. Também só está limitado pela sua imaginação na sua escolha de como visualizar seu DoW. Você pode mostrar o fluxo da esquerda para a direita, da direita para a esquerda, de cima para baixo, de baixo para cima. Não há nada que diga que você precise usar colunas para os estados. Você pode usar círculos ou espirais, ou o que quiser.

Ultimamente, tenho considerado um tópico muito interessante: como tornar os quadros Kanban mais inclusivos e acessíveis para aqueles de nós que têm dificuldades visuais? Achamos que técnicas de visualização mais “tradicionais” precisam ser questionadas ou mesmo descartadas por completo. Mas também pensamos que qualquer solução que envolva mais de um dos nossos sentidos para resolver um problema estaria conforme o espírito do Kanban. Esse sentido não precisa se limitar apenas à visão.

Conclusão

Assim que finalizar a definição do seu fluxo de trabalho e criar seu quadro, é hora de usá-lo. Sim, você me ouviu corretamente, você realmente precisa usar o quadro que acabou de criar. Reveja o último capítulo para obter mais informações sobre como fazer isso.

Mas, se lembra daquela exigência irritante da definição de Kanban de que realmente temos que “otimizar” o fluxo de valor através do nosso processo? A otimização não vem de ficar parado. Ela vem de pegar o sistema original que projetamos e realizar experimentos para descobrir como fazê-lo funcionar melhor. Em resumo, para otimizar nosso processo, vamos precisar melhorá-lo continuamente...

Capítulo 5 - Melhorando um Fluxo de Trabalho para Otimizar o Fluxo

No Guia Kanban, demos preferência à palavra “otimizar” em vez da palavra “maximizar”, de propósito. Isso ocorre porque “maximizar” valor é um sentimento muito perigoso. É bem simples maximizar o valor a curto prazo. Basta esgotar suas pessoas, tomar decisões onde você favorece o risco de longo prazo em detrimento do ganho de curto prazo (em vez do risco de curto prazo para o ganho de longo prazo), ignorar custos, etc. O truque é como você se posiciona para entregar valor aos seus clientes ano após ano. Por exemplo, a Toyota vem fazendo esse tal de “lean” há mais de 80 anos e meu palpite é que eles diriam que ainda não terminaram.

Otimizar também reconhece que você está operando dentro de certas restrições. Tentar maximizar uma coisa significa que você pode minimizar outras. A otimização implica encontrar o equilíbrio certo entre todas as restrições organizacionais. Acreditamos que o Guia Kanban diz isso de maneira bem clara:

“Ao contrário, a otimização de valor significa esforçar-se para encontrar o equilíbrio certo de eficácia, eficiência e previsibilidade na forma como o trabalho é feito:

- Um fluxo de trabalho eficaz é aquele que fornece o que os clientes querem quando eles o querem.

- Um fluxo de trabalho eficiente aloca os recursos econômicos disponíveis da forma mais otimizada possível para gerar valor.
- Um fluxo de trabalho mais previsível significa ser capaz de prever com precisão a entrega de valor dentro de um grau aceitável de incerteza.

Opções para Melhoria

A terceira prática do Kanban é melhorar o fluxo de trabalho. Qualquer elemento da sua Definição de Fluxo de Trabalho (DoW) é um candidato para melhoria experimental. Dizemos “experimental” porque nem todas as mudanças de processo funcionarão. Se o experimento melhorar as coisas, ótimo. Se piorar as coisas, volte ao que estava fazendo antes. O importante é tentar.

Itens de Trabalho

As atualizações nos Itens de Trabalho podem variar desde quais tipos rastrear, quais informações são registradas no item de trabalho, até como o item de trabalho é dimensionado. Por exemplo, digamos que Histórias de Usuário sejam um tipo de item de trabalho que você rastreia em seu quadro Kanban. Talvez uma melhoria seja (como mencionado no Capítulo 3) estabelecer uma política de que uma determinada história só pode ter 1 ou 2 critérios de aceitação. Possivelmente você queira introduzir cores para segmentar os diferentes tipos de itens de trabalho. Talvez você queira eliminar o campo de responsável dos seus itens de trabalho (o nome do responsável provavelmente não seja mesmo uma boa informação a ser rastreada). Porventura você queira parar de estimar itens de trabalho em pontos (com certeza, a melhor coisa que você pode fazer para um item de trabalho). Seja qual for o caso, o ponto é que você tem muitas opções

quando se trata de experimentos com itens de trabalho em seu fluxo de trabalho. O que nos leva a...

Fluxo de Trabalho

Da mesma forma, você deve experimentar diferentes aspectos do próprio fluxo de trabalho. Talvez adicionar ou remover colunas, talvez renomear colunas, talvez remover raias (melhor) ou adicionar raias (pior). Possivelmente você queira começar a visualizar etapas no *upstream* e no *downstream* de seus pontos de partida e de chegada para obter uma visão mais abrangente do seu processo (mais sobre isso em breve). A questão é que, se você estiver usando uma ferramenta virtual de Kanban, ajustar seu fluxo de trabalho pode ser algo limitado. Eu não escondi meu caso de amor com quadros físicos, mas entendemos que em nosso mundo cada vez mais remoto/distribuído, os quadros físicos podem não ser práticos (não impossíveis, mas também não práticos). A vantagem de um quadro físico é que você só está limitado pela sua imaginação em termos de como deseja visualizar seu fluxo de trabalho. Com uma ferramenta eletrônica, você fica naturalmente limitado pela funcionalidade que a ferramenta oferece. No entanto, isso não é algo muito grave. Uma vez trabalhei com uma equipe que usava uma ferramenta eletrônica que não suportava controles de WIP nativamente. Isso não os impediu, pois eles colocaram *Post-its* com números para representar os limites de WIP no topo do monitor que transmitia o quadro na sala da equipe. O que nos leva a...

Controles de WIP

Em 2016, Steve Reid, Prateek Singh e Daniel Vacanti publicaram um estudo de caso da Ultimate Software detalhando nosso sucesso com o Kanban naquele momento[1]. Nesse estudo de caso, exploramos em detalhe alguns de nossos esforços com um grupo conhecido como a “Equipe Aces”:

“Após observar o gráfico de dispersão, a equipe começou a investigar os motivos pelos quais as histórias estavam demorando tanto para serem concluídas. O que descobriram foi que a maioria das histórias de longa duração estava parada na coluna “Pronto para QA” por períodos prolongados. Isso era um problema porque “Pronto para QA” é uma coluna de fila onde as histórias simplesmente ficam paradas e não são trabalhadas ativamente. Essas colunas de ‘espera’ são os pontos mais óbvios de melhorias no processo, e foi o “Pronto para QA” que a equipe decidiu atacar primeiro, colocando um limite de WIP de 5 nessa coluna. Essa decisão significava que os desenvolvedores não podiam pegar novos trabalhos se houvesse 5 ou mais coisas esperando pelo QA. Eles teriam que ajudar nos testes do produto. Essa implicação foi discutida e aceita pela equipe como o comportamento apropriado para garantir o fluxo de trabalho.”

O que não é mencionado neste estudo de caso é que os Aces revisaram e ajustaram continuamente o Limite de WIP em sua coluna “Pronto para QA” e, por fim, conseguiram reduzi-lo para 2. Você pode ler no estudo de caso que eles tinham evidências de que essa mudança funcionou, pois o que viram foi uma diminuição proporcional no Tempo de Ciclo e um aumento na Vazão em seu processo.

Você notará que escondido neste estudo de caso está um ajuste de política que a equipe Aces implementou. Sempre que a coluna “Pronto para QA” ficava cheia, adotaram uma política de que os membros da equipe deveriam ajudar em outro item em andamento em algum lugar no quadro. A política não era começar algo novo, ou não ir para casa, era procurar oportunidades de trabalho em par ou em conjunto (swarm). O que nos leva a…

Políticas

O ambiente mais propenso a melhorias, quando se trata de políticas, está provavelmente nas suas políticas de processo. Você tem políticas que talvez nem saiba que são políticas.

Por exemplo, como você lida com bloqueios? O que significa que algo está bloqueado? Os bloqueios devem contar no cálculo do Limite de WIP? Qual é a ordem em que você puxa itens através do seu processo? O que significa algo estar concluído em uma etapa específica do fluxo de trabalho? Devemos trabalhar em par por padrão? Se você der um passo para trás e pensar sobre isso, provavelmente existem dezenas de políticas — implícitas ou explícitas — que afetam sua capacidade de realizar o trabalho todos os dias. Uma das maiores alavancas que você pode acionar para mudar seu processo é alterar essas políticas.

Fazer uma mudança de política mudará, por definição, como seu processo se comporta. O que nos leva a...

A Expectativa de Nível de Serviço

No exemplo dos Aces mencionado anteriormente, o Tempo de Ciclo do processo deles era de 33 dias no percentil 85 (antes dos controles de WIP), e foi para 14 dias no percentil 85 (após os controles de WIP). Se, quando começaram com o Kanban, eles estabeleceram um SLE de 33 dias ou menos no percentil 85, então certamente esse SLE não era mais válido após semanas ou meses de melhoria.

Uma pergunta comum que recebemos é “como eu sei quando devo ajustar meu SLE?” Como quase tudo relacionado ao fluxo, não há uma resposta direta. Podemos dizer o seguinte: uma mudança no SLE provavelmente não

é justificada a menos que tenha havido alguma outra mudança no processo (ou seja, uma mudança em um dos elementos DoW discutidos até agora). Quando você faz uma mudança — especialmente o que pode ser considerado uma mudança importante (por exemplo, alterar limites de WIP, remover colunas, alterar a política de bloqueio, etc.) – o que geralmente precisa fazer é esvaziar o quadro de quaisquer itens que estavam em andamento quando você fez a mudança e começar a rastrear o Tempo de Ciclo para novos itens que entram no processo após a mudança. Provavelmente, se a mudança for grande o suficiente, quando a mudança aconteceu ficará óbvio no seu gráfico de dispersão, e você poderá ajustar os dados históricos que você usa para calcular seu SLE de acordo. Se não for tão óbvio no seu gráfico de dispersão, então, como regra geral, assim que tiver cerca de 10 “novos” itens concluídos após a mudança, você deve ter dados suficientes para calcular um novo SLE. Note que nem todas as mudanças de processo resultarão em mudanças de SLE. É aqui que entra o discernimento da equipe. Use sua compreensão do contexto para decidir se uma mudança no SLE é apropriada ou não.

Uma última coisa sobre os SLEs: você pode estar se perguntando o que fazer se não tiver dados históricos suficientes para calcular um SLE. Embora este seja um caso muito incomum (geralmente o problema é ter dados demais), se você se encontrar nessa situação, a resposta é fácil. Adivinhe. Sério, adivinhe. Considere o que pode ser um SLE apropriado e então (você está vendo um padrão aqui ainda?) ajuste o SLE assim que tiver dados.

Um caso em que você pode não ter dados suficientes para calcular um SLE acontece quando tiver alterado os pontos de início e fim do seu processo. O que nos leva a...

Pontos de Início e Fim

Digamos que você tenha examinado todos os exemplos aqui e mais alguns e tenha ficado sem ideias de quais mudanças de processo fazer para melhorar. Para mim, isso é uma indicação bastante clara de que pode ser hora de expandir os limites do seu sistema Kanban. Talvez você queira expandir mais no *upstream* para cobrir o trabalho que acontece antes que os itens entrem no seu processo. Ou talvez você queira expandir mais no *downstream* para cobrir o trabalho que acontece depois que os itens saem do seu processo. De qualquer forma, uma vez que você expanda seus limites, quase sempre terá todo tipo de perguntas de melhoria para responder: nossos limites de WIP estão definidos corretamente? Que políticas temos em vigor para lidar com o trabalho no *upstream*/*downstream*? Qual deve ser o nosso novo SLE? Ao expandir continuamente os limites do seu processo, mais cedo ou mais tarde você deverá chegar àquela utopia conhecida como verdadeira agilidade de negócios de ponta a ponta. Ainda não vimos isso acontecer na prática, mas acreditamos que existe (ou pelo menos pode existir).

A propósito, para os céticos de plantão que podem dizer algo como “nós fazemos implantação/entrega contínua, então não há espaço para nós expandirmos no *downstream*”, deixe-me fazer uma pergunta: uma vez que você entregou ao seu cliente, como você está validando que o que você entregou é valioso? Vamos até dar o benefício da dúvida e dizer que você está validando o que é/não é valioso, então como você está acompanhando quais mudanças está fazendo com base nesse feedback? Meu ponto é que, se você for criativo o suficiente, quase certamente pensará em oportunidades para expandir o escopo do seu fluxo de trabalho geral.

Conclusão

Não há como enumerar neste livro todas as possíveis maneiras de melhorar seu sistema Kanban. O melhor que podemos fazer é dar alguns exemplos de melhoria e deixar que você experimente em seu próprio processo a partir daí.

Além disso, se você não fez uma mudança significativa em seu processo em meses, então você não está praticando o Kanban. A essência do profissionalismo é comparecer ao trabalho todos os dias acreditando que você pode melhorar. Começamos este capítulo falando sobre a Toyota. A Toyota vê a melhoria como uma jornada e não um destino. Você deveria fazer o mesmo.

Realizar experimentos em seu processo é ótimo, mas como saber se esses experimentos fizeram alguma diferença? A peça final que falta – e última peça do nosso quebra-cabeça do Kanban – será um conjunto de métricas básicas de fluxo que nos darão perspectivas adicionais sobre a saúde e o desempenho do nosso processo. Vamos discutir isso a seguir.

Notas

1. Steve Reid, “Ultimate Kanban: Scaling Agile without Frameworks at Ultimate Software” https://www.infoq.com/articles/kanban-scaling-agile-ultimate/

Capítulo 6 - As Métricas Básicas do Fluxo

As quatro medidas de fluxo a serem rastreadas no Kanban são:

- WIP (Trabalho em Andamento): O número de itens de trabalho iniciados, mas não concluídos.
- Templo de Ciclo (Cycle Time): A quantidade de tempo decorrido entre quando um item de trabalho foi iniciado e quando um item de trabalho foi concluído.
- Idade do Item de Trabalho (Work Item Age): A quantidade de tempo decorrido entre quando um item de trabalho foi iniciado e o momento atual.
- Vazão (Throughput): O número de itens de trabalho concluídos por unidade de tempo. Note que a medição de vazão é a contagem exata de itens de trabalho.

No Capítulo 4, falamos sobre a importância de ter um ponto bem definido onde o trabalho é iniciado e um ponto bem definido onde o trabalho é concluído. Para nossos propósitos aqui, a razão para essa importância é que todas as métricas básicas de fluxo são definidas em termos desses pontos de início e fim. Para uma discussão mais completa desse conceito, eu o direcionarei para o Capítulo 4. Apenas saiba que, para o restante deste capítulo, presumimos que você tem um processo com limites bem definidos.

Por si só, as métricas acima são irrelevantes a menos que possam informar uma ou mais das três práticas do Kanban. Além disso, essas quatro medidas de fluxo representam apenas o mínimo necessário para o funcionamento de

um sistema Kanban. Equipes e organizações podem e muitas vezes devem usar medidas adicionais, específicas do contexto, que auxiliam na tomada de decisões baseadas em dados[1].

Trabalho em Andamento

WIP: Todas as unidades discretas de valor potencial para o cliente que entraram em um determinado processo, mas ainda não saíram.

Para calcular o WIP, você simplesmente conta o número discreto de itens de trabalho dentro dos limites do seu processo conforme definido acima. É isso: apenas conte.

Sua objeção natural pode ser: “Isso não significa que você tem que fazer todos os seus itens de trabalho do mesmo tamanho?” Afinal, os itens de trabalho que passam pelo seu processo têm durações diferentes, são de complexidades díspares e podem exigir uma ampla variedade de recursos para trabalhar neles. Como você pode possivelmente considerar toda essa variabilidade e criar um sistema previsível apenas contando itens de trabalho? Embora essa seja uma pergunta razoável, não é algo para se preocupar.

Como vimos nos Capítulos 2 e 3, quando se trata de WIP, não há necessidade de que todos os seus itens de trabalho sejam do mesmo tamanho. Não haverá necessidade de adicionar qualquer complexidade extra ao cálculo do WIP, como estimar em Pontos de História ou atribuir horas ideais a cada item de trabalho. Esse conceito provavelmente é muito desconfortável para aqueles que estão acostumados a pensar no trabalho em termos de complexidade relativa ou nível de esforço, mas não há requisito para fazer qualquer tipo de estimativa inicial ao praticar o fluxo.

Também não há restrição no nível em que você rastreia o WIP do item de trabalho. Você pode acompanhar o WIP no nível de portfólio, projeto, equipe, indivíduo, apenas para citar alguns. Todas essas decisões são completamente suas e devem ser tomadas considerando as restrições do seu contexto.

Também se note que há uma diferença entre WIP e limites de WIP. Você não pode calcular o WIP simplesmente somando todos os limites de WIP em seu quadro. Deveria funcionar assim, mas na realidade não funciona. Este resultado deveria ser óbvio, já que a maioria dos quadros Kanban nem sempre tem colunas ou quadros que estão no limite máximo de WIP. Uma situação mais comum é ter um quadro Kanban com violações de limite de WIP em várias colunas — ou em todo o quadro. Em qualquer um desses casos, simplesmente somar os limites de WIP não lhe dará um cálculo preciso de WIP. A triste verdade é que não há como escapar de contar fisicamente o número de itens em progresso para chegar ao seu WIP total.

Em suma, se você quer usar o Kanban, mas atualmente não está rastreando o WIP, então você deve começar. Quanto antes, melhor.

Tempo de Ciclo

Na seção anterior, afirmamos que um processo tem fronteiras de chegada e partida específicas e qualquer item de valor para o cliente entre essas duas fronteiras pode ser logicamente contado como WIP. Uma vez que sua equipe determine os pontos de fronteira que definem o Trabalho em Andamento, a definição de Tempo de Ciclo se torna muito fácil:

Tempo de Ciclo: A quantidade de tempo decorrido que um item de trabalho passa como Trabalho em Andamento.

Essa definição é baseada na definição dada por Hopp e Spearman em seu livro “Factory Physics”, e acreditamos que se sustenta bem na maioria dos contextos de trabalho do conhecimento. Definir o Tempo de Ciclo em termos de WIP remove grande parte — senão toda — a arbitrariedade de algumas das outras explicações de Tempo de Ciclo que você pode ter visto (e ficado confuso) e nos dá uma definição mais precisa para começar a medir essa métrica. A moral dessa história é: você essencialmente tem controle sobre quando algo é contado como Trabalho em Andamento em seu processo. Dedique um tempo para definir essas políticas sobre o que significa para um item ser “Trabalho em Andamento” em seu sistema e comece e pare o relógio do Tempo de Ciclo conforme necessário.

Você também não deve subestimar a ênfase no “tempo decorrido”.

O uso do tempo decorrido é provavelmente muito diferente da orientação que você pode ter recebido anteriormente. A maioria das outras metodologias pede que você meça apenas a quantidade real de tempo gasto trabalhando ativamente em um determinado item (quando elas pedem para medir o tempo). Nós achamos que essa orientação está errada. Temos algumas razões para isso.

Primeiro, e mais importante, seus clientes provavelmente pensam no mundo em termos de tempo decorrido. Por exemplo, digamos que em 1º de março você comunique aos seus clientes que algo será feito em 30 dias. Suponho que a expectativa do seu cliente seja receber o item em, ou antes, do dia 31 de março. No entanto, se você quiser dizer 30 “dias úteis”, então sua expectativa é que o cliente receba algo por volta do meio de abril. Tenho certeza de que você consegue ver como essa diferença de expectativas pode ser um problema.

Segundo, se você medir apenas o tempo ativo, ignorará grande parte do seu problema de fluxo. É o tempo que um item passa esperando ou atrasado (ou seja, não está sendo trabalhado ativamente) que é onde geralmente está a maioria da sua imprevisibilidade. É exatamente nessa área que vamos procurar as melhorias mais substanciais em termos de previsibilidade. Lembre-se, o atraso é o inimigo do fluxo!

Ainda há uma razão mais importante para entender o Tempo de Ciclo. O Tempo de Ciclo representa a quantidade de tempo que leva para obter feedback do cliente. Este feedback é de vital importância em nosso mundo de trabalho do conhecimento. O valor em si é determinado pelo cliente, o que significa que sua equipe vai querer ter certeza de obter esse feedback de valor o mais rápido possível. A última coisa que você quer é desenvolver algo de que o cliente não precisa — especialmente se isso levar uma eternidade para fazer. Reduzir o Tempo de Ciclo vai encurtar o ciclo de feedback do cliente. E para reduzir o Tempo de Ciclo, você primeira precisa medi-lo.

Idade do Item de Trabalho

A Idade é de longe a métrica mais importante de todas as métricas de fluxo a serem acompanhadas, e por essa razão ela foi coberta em grande detalhe no Capítulo 1 deste livro (o que pode ser o motivo pelo qual você começou a ler este capítulo em primeiro lugar).

A definição de Idade de Item de Trabalho é:

Idade do Item de Trabalho: o tempo total que passou desde que um item entrou em um fluxo de trabalho.

A Idade do Item de Trabalho é uma medida do tempo atual em progresso que se aplica para todo o seu trabalho em andamento atual, por definição, e por esta razão, ela se aplica apenas a itens que entraram, mas não saíram do fluxo de trabalho. Uma vez que um item sai do fluxo de trabalho, todo o tempo acumulado até aquele momento imediatamente se converte em Tempo de Ciclo.

Vazão

Deixei a métrica mais fácil de definir por último. Simplificando, Vazão é definida como:

Vazão: a quantidade de WIP (número de itens de trabalho) concluídos por unidade de tempo.

De forma um pouco diferente, a Vazão é uma medida de quão rápido os itens saem de um processo. A unidade de tempo que sua equipe escolhe para sua medição de Vazão é uma escolha completamente sua. Sua equipe pode optar por medir o número de itens que conclui por dia, por semana, por iteração, etc. Por exemplo, você pode declarar a Vazão do seu sistema como “três histórias por dia” (para um determinado dia) ou “cinco funcionalidades por mês” (para um determinado mês).

Uma coisa adicional a saber sobre a Vazão é que muitos coaches e consultores ágeis usam as palavras “Velocidade” e “Vazão” de forma intercambiável. Embora a Velocidade possa ser definida em termos que são os mesmos que a Vazão, na maioria das vezes quando um coach diz “Velocidade” ele ou ela quer dizer “pontos de história por sprint”. Quando

definido em termos de pontos de história, você deve saber que Vazão e Velocidade estão longe de ser sinônimos.

Se a Vazão é a velocidade com que os itens saem de um processo, então a Taxa de Chegada é a velocidade com que os itens chegam a um processo. Mencionamos este fato aqui porque, dependendo do seu ponto de vista, a Taxa de Chegada pode ser considerada como um análogo à Vazão. Por exemplo, digamos que a etapa “Desenvolvimento” e a etapa “Teste” sejam adjacentes no seu fluxo de trabalho. Então, a Vazão da etapa “Desenvolvimento” também poderia ser pensada como a Taxa de Chegada para a etapa “Teste”.

Ainda mais importante, porém, comparar a Taxa de Chegada de uma etapa em seu processo com a Vazão em outra etapa, diferente, pode lhe fornecer algumas informações muito necessárias sobre problemas de previsibilidade. Vamos entrar em muito mais detalhes sobre essa comparação nos próximos capítulos. No entanto, minha razão mais imediata para discutir a Taxa de Chegada é simplesmente para apontar que quão rápido as coisas chegam ao seu processo pode ser tão importante quanto quão rápido as coisas saem.

A métrica de Vazão responde a uma pergunta muito importante: “Quantas funcionalidades vou ter no próximo lançamento?” Em algum momento, você precisará responder a essa pergunta, então rastreie a Vazão e esteja preparado.

Conclusão

O que apresentamos aqui são as métricas fundamentais do fluxo para iniciar sua jornada: WIP, Tempo de Ciclo, Idade do WIP e Vazão. Embora existam, sem dúvida, outras métricas que possam ser relevantes para seu ambiente

específico, essas métricas formam a base comum a todas as implementações de fluxo. Se seus objetivos incluem melhoria e previsibilidade, então essas são as métricas que devem estar no centro de seus esforços de rastreamento.

Notas

1. John Coleman and Daniel Vacanti, “The Kanban Guide” https://kanbanguides.org/

Capítulo 7 - Liberando o Verdadeiro Poder do Kanban

Tendo chegado a este ponto do livro, é natural pensar que o Kanban só é aplicável a equipes de desenvolvimento. No entanto, acreditamos que o verdadeiro poder do Kanban só será liberado quando começarmos a escalar nossa implementação ao longo das dimensões descritas neste capítulo. Embora a implementação e a otimização de um sistema Kanban no nível da equipe ajudem você a caminhar, a escalada dessas práticas ajudará você a correr. A boa notícia é que escalar o Kanban não requer nenhum conjunto novo de princípios ou práticas. Kanban em escala é simplesmente a aplicação das mesmas práticas sobre as quais você já leu neste livro, em novas dimensões.

Dimensões de Escala

Existem múltiplas dimensões de escalabilidade ao longo das quais podemos aplicar as práticas do Kanban. Veremos abaixo algumas maneiras pelas quais o Kanban pode ser aplicado para alcançar ótimos resultados além dos limites de uma única equipe de desenvolvimento.

Várias Equipes Trabalhando no Mesmo Produto

Os produtos geralmente são grandes demais para serem desenvolvidos e mantidos por uma única equipe. As equipes frequentemente selecionam funcionalidades do produto para desenvolver por conta própria. Logo, percebemos que existem múltiplas dependências entre essas equipes, pois elas contribuem para o mesmo produto. Isso leva ao bloqueio do trabalho e, muitas vezes, ao início de mais funcionalidades. Nos vemos em uma situação em que as equipes individuais podem ser capazes de manter seus WIPs de itens de trabalho, mas o WIP geral de funcionalidades para esse conjunto de equipes continua crescendo. Um padrão muito comum que leva a este problema são equipes divididas por especialização, por exemplo, equipes focadas em Front End, Back End, Banco de Dados, QA, etc. Este problema também pode existir quando temos um grupo de equipes multidisciplinares trabalhando no mesmo produto.

O leitor atento provavelmente já terá adivinhado a resposta dos autores a este problema, com base nos capítulos anteriores. Esta é a oportunidade perfeita para olhar para essas equipes como um sistema único. Podemos fazer isso usando um único modelo representativo para o trabalho no produto, provavelmente um quadro Kanban contendo as funcionalidades do produto. O quadro Kanban força duas práticas muito importantes neste contexto — A limitação do WIP, ao nível das funcionalidades entre estas equipes e o estabelecimento de um SLE para estas funcionalidades.

Se gerenciarmos ativamente o trabalho neste quadro, podemos garantir que nenhuma das funcionalidades envelheça por muito tempo. Também podemos facilitar que as equipes se ajudem mutuamente, desbloqueando o trabalho ou até mesmo trabalhando em algumas funcionalidades juntas. Isso geralmente leva a um único backlog de produto, em vez de um backlog para cada equipe. Como resultado, novos padrões de trabalho e colaboração começam a surgir. As equipes podem escolher trabalhos inicialmente programados para outra equipe, permitindo maiores graus de

flexibilidade e transferência de conhecimento. As fronteiras entre as equipes começam a ficar nebulosas. Podemos chegar à conclusão de que seria benéfico ao fluxo geral a combinação de algumas equipes, ou criar equipes maiores que possam aproveitar melhor a nova flexibilidade. Ter um quadro Kanban de nível Funcionalidade ou Épico entre equipes que trabalham no mesmo produto pode garantir o sucesso de todas as equipes e, consequentemente, do produto.

Várias Equipes e Vários Produtos

É fácil perceber os benefícios de um único quadro para equipes que trabalham no mesmo produto. O trabalho deles está provavelmente relacionado e eles conseguem se ajudar mutualmente para concluir os itens de trabalho. E se tivermos várias equipes, mas agrupamentos delas estiverem focados em vários produtos? Seria aconselhável implementar o Kanban em um nível superior? A resposta ainda é ‘Sim’.

Um quadro Kanban ao nível de portfólio pode garantir que nossa execução esteja alinhada com nossa estratégia atual. Se expormos todas as iniciativas/funcionalidades/épicos (independentemente do nível que faça sentido aqui) em que a organização está trabalhando atualmente, que informações poderemos obter? Descobriremos muito rapidamente qual é o nosso WIP atual nesses itens de alto nível em geral. Também saberemos se o foco das equipes corresponde à estratégia da organização. Quando cada produto tem seu próprio backlog, é muito provável que o item de maior prioridade no backlog de um produto não seja apenas de menor prioridade para a organização, mas também contraria a estratégia geral. Além disso, se a equipe que trabalha neste produto for muito eficiente, a organização produzirá mais daquilo que não deseja do que aquilo que considera estrategicamente importante.

Ter uma implementação Kanban ao nível portfólio nos ajuda a nos tornarmos eficientes e eficazes no nível organizacional. Seremos capazes de tornar o fluxo de trabalho eficiente monitorando as métricas de fluxo e melhorando-as ao longo do tempo. Também conseguiremos produzir mais coisas certas, garantindo que a distribuição desse trabalho ativo e futuro esteja alinhada com a nossa estratégia global. A frequência com que um quadro neste nível é gerenciado pode não ser a mesma de um quadro de equipe. Pode ser com menos frequência, mas com frequência suficiente para podermos fazer ajustes a tempo.

Níveis Superiores da Organização

Da mesma forma que várias equipes podem ficar fora de sincronia com a direção estratégica geral, departamentos inteiros podem cair na mesma armadilha. Os objetivos estratégicos e as prioridades da organização podem muitas vezes ser incompatíveis com as prioridades de departamentos individuais. Escalar o Kanban para o nível da organização, onde os objetivos de alto nível são representados em um quadro, pode ajudar a evitar isso. Um quadro para cada nível pode ajudar a vincular os objetivos da organização de alto nível aos itens de trabalho no nível do departamento, que por sua vez podem ser vinculados às funcionalidades dos produtos e das equipes e, consequentemente, aos itens de trabalho nos quadros das equipes. O primeiro e imediato resultado disto é compreender se as nossas atividades atuais estão alinhadas com a nossa estratégia e objetivos organizacionais.

Um quadro Kanban não está completo sem uma Expectativa de Nível de Serviço e políticas explícitas. É aqui que começam a surgir os verdadeiros benefícios de ter quadros em níveis mais altos. Um problema comum com os objetivos organizacionais é quão demorados e vagos eles podem ser. A aplicação de critérios de saída explícitos e a identificação de uma fase «concluída» podem ajudar a garantir que os objetivos globais não sejam

vagos. Um SLE também pode garantir que não tenhamos objetivos de longa duração que estejam desatualizados, mas ainda ativos. O gerenciamento ativo de itens em um quadro em qualquer nível da organização exige a mesma disciplina que os quadros ao nível de equipe. Quanto mais elevado for o nível do quadro, maior impacto terá na eficiência, eficácia e previsibilidade da organização.

O gerenciamento ativo de itens de trabalho em níveis superiores pode estender os benefícios do Kanban a todos os níveis. O envelhecimento dos itens de trabalho nesses quadros pode revelar problemas de WIP, dimensionamento ou dependência. Nos níveis mais elevados da organização, muitas vezes temos as pessoas certas envolvidas para resolver estas questões. Seremos capazes de reunir equipes e departamentos para trabalhar em nossas prioridades coletivas e garantir a entrega de valor aos clientes.

Estendendo *Upstream* e *Downstream*

Até agora falamos sobre a escala para incluir mais equipes, produtos e departamentos em nosso sistema Kanban. Outra dimensão é a escala para incluir mais atividades. Muitas equipes iniciam frequentemente a implementação do Kanban começando do ponto onde se inicia a implementação de uma solução até o ponto em que a solução está pronta para ser entregue. Esta é uma ótima forma de iniciar o sistema, mas se deixado sozinho, este sistema pode acabar ficando abaixo do ideal. É provável que criemos um subsistema altamente otimizado que produza resultados mais rápido do que podemos entregá-los aos clientes. Isso também pode encorajar os processos *upstream* a tentar abastecer o sistema e criar desperdício.

Quando tivermos estabilizado o sistema Kanban inicial, é hora de expandir o sistema para incluir atividades *upstream* e *downstream*. Precisamos então

redefinir os pontos inicial e final dentro dos quais nosso sistema funcionará. Uma vez otimizado o processo de criação da solução para os problemas do cliente, devemos incluir os processos de compreensão do problema (*upstream*) e entrega da solução (*downstream*), bem como obter feedback sobre a solução entregue (*downstream* que informa o *upstream*). À medida que movemos as linhas de início e de chegada, aplicamos as mesmas práticas descritas neste livro — identificar itens de trabalho, definir estágios, limitar WIP, políticas explícitas e ter um SLE. Com o tempo, seremos capazes de estender nossa prática Kanban desde o ponto em que qualquer pessoa na organização começa a entender o problema até o ponto em que recebemos feedback dos clientes de que o problema foi ou não resolvido.

Outros Departamentos da Organização

Muitas vezes, em uma organização, apenas um departamento começa com Kanban. Outro modo de escalar o Kanban é polinizar o aprendizado em outros departamentos. Se o departamento de TI obteve grandes benefícios com a implementação de sistemas Kanban, alguns dos departamentos com os quais ele interage regularmente podem ser ótimos lugares para espalhar os benefícios do Kanban. Vimos em primeira mão departamentos como suporte ao cliente, ativações de clientes e estratégia de produtos utilizando os métodos adotados inicialmente pelas equipes de desenvolvimento. Essas adoções levaram a uma maior eficiência, eficácia e previsibilidade para estes departamentos.

Escalar desta maneira muitas vezes coincide com as outras dimensões de escala já mencionadas. Ter um sistema Kanban no nível da organização pode atuar como um veículo para espalhar os benefícios do Kanban para os vários departamentos.

Conclusão

Quando se trata de escalar o Kanban, há várias dimensões a serem consideradas. Descrevemos algumas delas aqui. Podemos liberar o verdadeiro poder do Kanban ampliando essas dimensões. Frequentemente, uma dessas dimensões leva à necessidade de trabalhar em outra diferente. Recomendamos que a organização escolha aquela que parece mais fácil de progredir e observe algum sucesso antes de embarcar numa dimensão diferente. Kanban é uma estratégia aplicável a vários tipos de equipes e níveis de uma organização. Não tenha medo de jogar fora das linhas do campo.

Um exemplo de escala é o estudo de caso Ultimate Software Scaling Kanban. Este trecho descreve como a organização combinou algumas das abordagens de expansão descritas aqui:

“A adoção de técnicas ágeis proporcionou os benefícios de maior produtividade e previsibilidade. Porém, em uma perspectiva geral, a Ultimate Software está em um sanduíche de Cascata. A organização de desenvolvimento ágil fica no meio das organizações tradicionais de vendas e suporte e das organizações tradicionais de implantação e ativação. Como parte da próxima evolução do pensamento ágil e baseado em fluxo na Ultimate Software, estamos expandindo para organizações que flanqueiam o desenvolvimento. A cultura da Ultimate, que incentiva gerentes e funcionários a experimentar e tomar as decisões certas para a Ultimate, ajudou muito na difusão dos princípios fora do desenvolvimento central. Os departamentos da Ultimate Software começaram a recorrer aos serviços dos instrutores de agilidade no desenvolvimento para ajudá-los com os mesmos princípios.

“O nosso envolvimento mais próximo com a Estratégia de Produto e a capacidade de lhes proporcionar mais previsibilidade melhoraram

enormemente a capacidade do Desenvolvimento de ajudar com questões de suporte sem interromper o trabalho ativo. O suporte Nível 3 também adotou práticas Kanban para melhorar sua capacidade de oferecer suporte aos nossos clientes. A Estratégia de Produto consegue utilizar a previsibilidade e os ganhos de produtividade do Desenvolvimento para fornecer melhor orientação a Vendas sobre os próximos produtos e funcionalidades. À medida que continuamos a melhorar a previsibilidade que podemos fornecer a Vendas, podemos começar a criar solicitações de funcionalidades e prioridades em conjunto com Vendas. As funcionalidades podem então ser puxadas por todo o fluxo de valor e o rastreamento do Tempo de Ciclo e da Vazão nos permitirá assumir e manter compromissos mais precisos com nossos clientes.

“Embora a expansão *upstream* nos ajude a melhorar a criação de valor, a expansão *downstream* para implantação e ativações é onde podemos melhorar a entrega de valor aos nossos clientes. Assim que a Ultimate Software começa a trabalhar em novos produtos, transferimos as atividades de implantação para as equipes. Para nossos produtos mais antigos, sempre fizemos uma transferência para nosso grupo de implantação SaaS. Quebramos a mentalidade do ‘por cima do muro’ ao incorporar engenheiros de implantação nas equipes de desenvolvimento de novos produtos e ajudá-los a educar o restante da equipe sobre como manter seus próprios pipelines de implantação. As equipes inicialmente ficaram preocupadas em assumir essa responsabilidade adicional. Esses receios diminuíram à medida que as equipes se deram conta do apoio que estava disponível pelo resto da organização. Essa prática também reduziu bastante as ocorrências de surpresas no ambiente de produção. Como as equipes ajudam a construir os ambientes nos quais implantam o código, o código não se comporta de forma inesperada quando enviado para produção. Essas equipes são apoiadas por três grupos fora da Engenharia de Produto. Os grupos que gerenciam a infraestrutura de Build e Deployment dos produtos em desenvolvimento também adotaram os princípios Kanban e começaram a medir os Tempos de Ciclo para disponibilizar a infraestrutura às equipes.

Eles estabeleceram SLAs para diferentes tipos de solicitações e tornaram-se previsíveis com essas métricas.

“Agora conseguimos ver uma funcionalidade percorrendo todo o caminho desde uma solicitação gerada em Vendas até a Estratégia de Produto, passando pelo Desenvolvimento e finalmente pela Produção. Ao conseguirmos acompanhar o progresso de uma funcionalidade dessa maneira, podemos começar a identificar oportunidades de melhoria no ciclo do início à entrega. A organização na totalidade pode identificar onde as funcionalidades ficam impedidas e aplicar nosso entendimento de fluxo para eliminar o tempo que as funcionalidades têm de esperar em filas em toda a organização.

Outro aspecto posterior ao desenvolvimento e até mesmo ao grupo de implantação são as ativações. Ativações é o grupo que ajuda um novo cliente a lançar os produtos da Ultimate Software. O processo de ativação pode levar até um ano e envolver várias equipes. A cada dia que um cliente gasta em fase de ativação, a Ultimate Software investe tempo, mas não recebe receita integral. Esta é uma área que pode aproveitar os benefícios que a Organização de Desenvolvimento obteve com o fluxo e as práticas ágeis. O Desenvolvimento começou a trabalhar com Ativações para compartilhar os princípios e práticas que fizeram uma diferença positiva na previsibilidade e na velocidade de conclusão dos resultados.

“Mover o Kanban para fora das linhas do campo é o próximo grande passo para a Ultimate Software. Já começamos a avançar nessa direção através do nosso trabalho com equipes de suporte e implantação. A Ultimate continua a expandir sua implementação Ágil sem usar nenhum framework específico. Configurar os canais certos de comunicação e visualizar nosso trabalho de uma maneira que seja facilmente compreendida por todos é o ponto crucial de como a Ultimate conseguiu adotar e evoluir com sucesso o Ágil em escala.”

Capítulo 8 - Reflexões Sobre Como Começar

Ao ler este livro, você provavelmente já está formando uma boa ideia de como seria um sistema Kanban em seu contexto. A parte mais difícil muitas vezes é apenas começar. Aqui descrevemos uma abordagem para ajudá-lo a começar. Existem várias maneiras de começar com Kanban, a descrita abaixo é a que consideramos mais bem-sucedida. Elementos desta abordagem foram descritos em capítulos anteriores (especialmente no capítulo 4). Indicamos os capítulos que se aprofundam no conceito ao longo de cada etapa. Consulte esses capítulos para uma discussão mais profunda dessas etapas.

Etapas iniciais

Definir representações de valor (Capítulo 4)

O primeiro passo é entender quais são os itens de trabalho que representam os pedaços individuais de valor que fluem pelo seu sistema. São itens que entregamos e validamos com os clientes.

Defina os pontos inicial e final (Capítulo 4)

Em seguida, precisamos determinar quando consideramos que esses itens foram iniciados e quando “terminamos” esses itens. Estes formam os limites

dentro dos quais iremos gerenciar o sistema Kanban. Esses limites certamente mudarão ao longo do tempo, mas precisamos sempre estar cientes deles para podermos observar e melhorar efetivamente o sistema.

Determine todas as atividades que ajudam a criar valor (Capítulo 4)

À medida que um item se move do ponto inicial até o ponto final de um sistema, realizamos ações nele para transformá-lo em valor entregável. Precisamos determinar quais são essas atividades. Para um software, podem ser: compreensão do problema, design técnico, criação do código, revisão de código, teste de unidade, teste de integração, construção e implantação, validação do cliente. Os sistemas terão um número variável de atividades que contribuem para que um item de trabalho se torne uma parte do valor entregável. Ao listá-los, temos uma ideia do fluxo do sistema.

Mapeie atividades e itens de trabalho determinados para etapas do processo (Capítulo 4)

Até o momento temos itens de trabalho, pontos de início e término, e atividades. Estamos começando a entender o fluxo através do sistema. A próxima etapa é descobrir quais são os principais estágios entre os pontos inicial e final pelos quais os itens de trabalho fluem. Podemos mapear as atividades definidas nas etapas anteriores para essas etapas. Essas atividades fazem parte do nosso conjunto de políticas para o sistema. Elas se tornam os critérios de saída para cada estágio – Um item de trabalho não pode passar de um estágio a menos que as atividades mapeadas tenham sido concluídas. Também podemos mapear os itens atualmente ativos no sistema para esses estágios usando os critérios de saída. Isso nos dá uma ideia do estado atual do sistema.

Decidir sobre um método para limitar o WIP (Capítulos 1 e 4)

O tema mais importante deste livro é: todas as práticas Kanban podem ser derivadas da motivação singular de não querer que os itens envelheçam desnecessariamente. O ponto de partida para controlar a idade e, consequentemente, o Tempo de Ciclo é limitar o WIP. Existem várias maneiras de limitar o WIP – Adicionar limites de WIP a cada estágio, adicionar um limite geral de WIP a todo o sistema, limitar o WIP com base no número de pessoas na equipe, etc. Neste passo precisamos escolher um desses métodos e garantir que temos uma boa compreensão de como limitaremos o WIP visando evitar o envelhecimento.

Selecione um SLE inicial (Capítulo 2)

Em seguida, precisamos criar um entendimento comum sobre “Quanto tempo leva para concluir o trabalho?”, por meio de um SLE. Podemos fazer isso analisando a distribuição do Tempo de Ciclo dos itens que já foram concluídos. Com base nesta distribuição, podemos escolher um SLE que representa a quantidade de tempo que a maioria (70%, 85%, 95%, etc.) dos itens leva para ser concluída. Este SLE se torna o parâmetro para itens que estão ativos no sistema.

Definir expectativas sobre a inspeção e adaptação de políticas (Capítulo 5)

Todo o trabalho de definição e configuração de um sistema Kanban muitas vezes pode dar às equipes a falsa confiança de que descobriram a maneira perfeita de executar e monitorar o trabalho. Mas é apenas o começo da jornada e precisamos ser explícitos sobre isso. Os líderes e as equipes devem esperar que à medida que aprendem mais sobre o seu trabalho, e

fluxo de trabalho, as fases do trabalho e as políticas mudem. À medida que evoluímos o sistema, ele se ajustará melhor ao nosso contexto específico.

Gerenciar ativamente o trabalho para melhorar o sistema (Capítulo 3)

Com o sistema já configurado, precisamos nos concentrar na operação do fluxo de trabalho. Diariamente, temos que gerenciar os itens de trabalho visando não os deixar envelhecer desnecessariamente. Esse gerenciamento ativo pode ocorrer de várias formas – Trabalho em Par, “Swarming”, Remoção de Bloqueios, Dimensionamento Correto e outros não discutidos neste livro. Ao fazermos isso diariamente, criamos um ciclo de feedback no processo. Esse ciclo de feedback ajuda a ajustar nossas políticas e nosso fluxo de trabalho para melhorar o sistema na totalidade.

Conclusão

Há muitas maneiras de começar a usar o Kanban. Este capítulo fornece um modelo da abordagem que vimos funcionar (ex post – como observado na prática). A maioria dos trabalhos iniciais envolve a compreensão de suas atividades atuais e a definição de seu fluxo de trabalho e políticas. Assim que tivermos um entendimento comum sobre isso, podemos começar a observar os motivos pelos quais os itens envelhecem em nosso sistema e fazer os ajustes apropriados. Damos o passo revolucionário, mas simples, de definir um sistema com WIP limitado para podermos dar vários passos evolutivos para melhorar o sistema todo.

Epílogo - Profissionalismo e ProKanban.org

Como será o futuro do Kanban?

Você teve uma ótima visão de como o Kanban começou e como deveria ser, então o que vem a seguir para o Kanban?

Não sei quanto tempo durarão as mudanças que a pandemia trouxe ao mundo do trabalho, mas há uma que espero que permaneça: a incansável priorização do autocuidado. Trabalhar em casa e em diferentes fusos horários tornou facilitou trabalhar do anoitecer ao amanhecer, sem pausas e nunca vendo a luz do sol, o que também significa que estamos exaustos. A vida doméstica e a vida profissional estão tão inextricavelmente misturadas que não têm mais início e fim. E é aqui que o Kanban brilha.

Precisamos de um mecanismo (sejamos uma equipe de um ou de vinte) para nos organizar em torno do trabalho a ser feito. Para ver nossas prioridades, ver o que está travado e concentrar nosso tempo em fazer as coisas. Um novo termo para este défice diário é “pobreza de tempo” e tem provocado um efeito profundo na nossa saúde física, bem-estar e produtividade.

Ouço desenvolvedores falarem sobre como é fazer parte da “fábrica de funcionalidades” (feature factory), como Melissa Perri descreve, onde produzimos código sem ter tempo para fazer uma pausa e entender qual problema estamos tentando resolver. Quase sempre os ouço dizer que nunca têm tempo para pensar profundamente. Tempo realmente focado

para resolver problemas de forma criativa. Tempo de inatividade para lidar com a crescente dívida técnica. Oportunidade de aprender novas tecnologias ou orientar outras pessoas. E o resultado disso é que as pessoas desistem e seguem em frente na esperança de encontrar um ritmo que não as esgote. Um ritmo que raramente encontram.

Mas, é possível? Um lugar onde há 100 coisas no backlog e não estamos bebendo da mangueira de incêndio tentando acompanhar? Acredito que sim. São lugares onde se usa o Kanban.

Sempre haverá novos métodos (e novos nomes para métodos antigos) de como entregar trabalho de conhecimento, mas é difícil argumentar contra a simples necessidade de foco. Passamos muito tempo na comunidade ágil falando sobre a saúde e o moral da equipe, mas nunca vi uma equipe encontrar melhor equilíbrio e ter melhor controle sobre o volume de trabalho que tem pela frente todos os dias do que quando tem o controle para criar políticas que lhes permitam desligar no final do dia e retornar no dia seguinte com uma compreensão clara do que fazer a seguir.

As equipes que têm controle sobre como trabalham são mais felizes e produtivas. Elas sabem como contribuir quando têm capacidade. Elas sabem como ajudar quando o trabalho está bloqueado. E elas têm autonomia para ajustar as coisas que não estão funcionando à medida que avançam. É assim que envolvemos os funcionários e evitamos o esgotamento – com o Kanban.

Alguns anos antes da pandemia, comecei a trabalhar com uma grande organização de segurança para implementar o Kanban. O que aprendi rapidamente é que o espaço de segurança tem um dos maiores níveis de rotatividade porque está constantemente combatendo incêndios. Tudo é urgente. O esgotamento é alto. Raramente há tempo de inatividade. E, em última análise, as pessoas não duram muito nesse ritmo.

As equipes que treinei queriam o Kanban como uma forma de equilibrar sua carga de trabalho e criar um SLE previsível para entrega aos seus “clientes”, mas ficaram agradavelmente surpresos quando descobriram que isso também deixava suas equipes mais felizes. A rotatividade da equipe diminuiu e permitiu que a equipe equilibrasse projetos internos de longa duração com os requisitos diários de uma organização de segurança e continuaram assim indefinidamente. Não foi um sprint bem-sucedido ou uma semana lenta após uma semana de insanidade – foi um ritmo que criou equilíbrio em um ambiente normalizado por tempestades insanas de urgências.

O que é importante aqui é que esta “pobreza de tempo” também se espalha por tudo o que fazemos fora do trabalho. Estamos sem energia, sem paciência e sem a centelha que nos torna excelentes no que fazemos. A resposta raramente é dizer não – em muitos desses casos isso pode nem ser possível. Trata—se de foco, fluxo e eficiência — e de ter controle sobre o processo para tornar isso possível.

Espero que o que venha em seguida para a nossa indústria e, mais importante, para as pessoas que nela trabalham, seja exigirmos uma forma diferente de trabalhar. Uma forma que nos permite trazer o nosso melhor, e o nosso melhor trabalho para o que fazemos – e acho que isso é melhor quando feito com Kanban.

Kanban nos dá a capacidade de sermos melhores naquilo que fazemos, permitindo—nos focar totalmente. Kanban nos dá a capacidade de sermos melhores membros na equipe, organizando colaborativamente o trabalho a ser realizado. O Kanban nos dá a oportunidade de sermos melhores profissionais, buscando melhorias incessantes em nossos sistemas e na forma como entregamos valor. E tudo isso nos torna humanos melhores quando desligamos nosso laptop.

Apêndice A — Uma Introdução à Lei de Little

Nota: este apêndice apareceu originalmente como capítulo 3 do livro de Daniel Vacanti "Actionable Agile Metrics for Predictability". Ele foi editado a partir da forma original para torná-lo mais alinhado com o objetivo deste livro.

O Capítulo 6 tratou das métricas básicas de fluxo: WIP, Tempo de Ciclo e Vazão. No que pode ser considerado um dos resultados mais milagrosos da história da análise de processos, estas três métricas estão intrinsecamente ligadas por uma relação muito simples e poderosa conhecida como Lei de Little:

Tempo de Ciclo Médio = Trabalho em Andamento Médio / Vazão Média

Se você já viu a Lei de Little antes, provavelmente já a viu na forma da equação acima. O que poucos profissionais do Ágil percebem, entretanto, é que a Lei de Little foi originalmente enunciada de uma forma ligeiramente diferente:

Média de Itens na Fila = Taxa Média de Chegada * Tempo Médio de Espera

Este fato é importante porque diferentes pressupostos precisam de ser satisfeitos dependendo da forma da lei que se utiliza. E compreender os pressupostos por trás da equação é a chave para compreender a própria lei. Após compreender as suposições, você poderá usá-las como um guia para

algumas políticas de processo que podem ser implementadas para ajudar na previsibilidade.

A matemática da Lei de Little é simples. Mas neste apêndice não nos importamos com a matemática. O que nos importa – e não posso enfatizar o suficiente este ponto, se quisermos obter uma maior apreciação da aplicabilidade da lei ao nosso mundo – é olhar muito além da elegância da equação para obter uma compreensão mais profunda das suposições básicas necessárias para fazer a lei funcionar. É aí que as coisas ficam mais complicadas, mas é também onde encontramos os maiores benefícios. Uma compreensão completa de porque a Lei de Little funciona dessa maneira será a base para a compreensão de como as métricas básicas de fluxo podem se tornar previsivelmente acionáveis.

Precisamos de um Pouco de Ajuda

Primeiro, algumas informações básicas.

O Dr. John Little passou grande parte do início de sua carreira estudando sistemas de filas. Na verdade, uma das melhores definições desse sistema de filas vem do próprio Dr. Little:

“Um sistema de filas consiste em objetos discretos que chamaremos de itens, que chegam ao sistema em certa taxa. Os itens podem ser carros em um pedágio, pessoas em uma fila de refeitório, aeronaves em uma linha de produção ou instruções esperando para serem executadas em um computador. O fluxo de chegadas entra no sistema, junta—se a uma ou mais filas e eventualmente recebe atendimento, e sai em um fluxo de saídas. O serviço pode ser uma corrida de táxi (viajantes), uma tigela de sopa (trabalhadores famintos) ou conserto de automóveis (proprietários de automóveis). Geralmente o atendimento é o gargalo que cria a fila, e por

isso, normalmente, temos uma operação de atendimento com tempo de atendimento, mas isso não é obrigatório. Nesse caso, presumimos que ainda existe um tempo de espera. Às vezes é feita uma distinção entre o número na fila e o número total na fila mais o serviço, sendo este último chamado de número no sistema."

A diversidade de domínios que ele menciona aqui é extraordinária. Embora ele não mencione especificamente o desenvolvimento de software ou o trabalho de conhecimento em geral, vou sugerir que essas áreas também podem ser facilmente modeladas dessa forma.

Em 1961, o Dr. Little decidiu provar o que parecia ser uma propriedade muito geral, e muito comum, exibida por todos os sistemas de filas. O resultado que ele estava pesquisando era uma conexão entre a taxa média de chegada de uma fila, o número médio de itens na fila e o tempo médio que um item passava na fila (para os fins deste apêndice, quando digo "média" estou realmente falando de "média aritmética"). Matematicamente, a relação entre essas três métricas é semelhante a:

Equação (1): $L = \lambda * W$

Onde:

L = o número médio de itens no sistema de filas.

λ = o número médio de itens que chegam por unidade de tempo.

W = tempo médio de espera no sistema por um item.

Observe que a Equação (1) é expressa estritamente em termos da Taxa de Chegada de um sistema de filas. Este ponto será de especial interesse um pouco mais adiante neste capítulo.

Observe também que – se ainda não for óbvio – a Lei de Little é uma relação de médias. A maioria das aplicações de trabalho de conhecimento e discussões sobre a lei negligenciam esse detalhe muito importante. O fato da Lei de Little se basear em médias não é necessariamente bom ou mau. Só é mau quando as pessoas tentam aplicar a lei para usos que nunca foram pretendidos.

O Dr. Little foi o primeiro a fornecer uma prova rigorosa para a Equação (1) e, como tal, esta relação é desde então conhecida como Lei de Little. Segundo ele, uma das razões pelas quais a lei é tão importante é que (a ênfase é minha): “L, λ e W são três medidas bastante diferentes e importantes de eficácia do desempenho do sistema, e a Lei de Little insiste que elas devem obedecer à 'lei'… *A Lei de Little reúne as três medidas de uma forma única e consistente para qualquer sistema em que se aplica. A Lei de Little não dirá aos gestores como lidar com os compromissos ou fornecer inovações para melhorar as medidas escolhidas, mas estabelece uma relação necessária*. Como tal, fornece estrutura para pensar sobre qualquer operação que possa ser lançada como uma fila e sugere quais dados podem ser valiosos para serem coletados.”

A grande vantagem da Lei de Little é a simplicidade geral do seu cálculo. Especificamente, se alguém tiver duas das três estatísticas acima, poderá calcular facilmente a terceira. Este resultado é extremamente útil, pois há muitas situações em diversos domínios onde a medição de todas as três métricas de interesse é difícil, cara ou mesmo impossível. A Lei de Little mostra—nos que se pudermos medir quaisquer dois atributos, obteremos automaticamente o terceiro.

Para ilustrar esse ponto, o Dr. Little usou o exemplo muito simples de uma prateleira para vinhos. Digamos que você tenha uma adega que, em média, sempre contém 100 garrafas. Digamos ainda que você reabastece a prateleira a uma taxa média de duas garrafas por semana. Conhecer apenas esses dois números (e nada mais!) nos permite determinar quanto tempo,

em média, uma determinada garrafa fica parada na prateleira. Aplicando a Equação (1), temos L igual a 100 e λ igual a 2. Inserir esses números na fórmula nos diz que uma garrafa de vinho passa, em média, 50 semanas na prateleira.

Antes de prosseguirmos, vale a pena explorar quais são as condições contextuais necessárias para a lei ser válida. Quando expresso na forma da Equação (1), a única suposição necessária é que o sistema em consideração tenha alguma garantia de estar em um regime estacionário. É isso. Realmente, é isso. Para ilustrar o que não precisamos, observe que podemos chegar ao resultado da prateleira de vinhos sem rastrear as datas específicas de chegada ou partida de cada garrafa individual. Também não precisamos saber a ordem específica em que as garrafas foram colocadas na prateleira, ou a ordem específica em que as garrafas foram retiradas da prateleira. Não precisamos entender nada sofisticado como as distribuições de probabilidade subjacentes das taxas de chegada e partida. Curiosamente, nem precisamos rastrear o tamanho das garrafas na prateleira. Poderíamos ter algumas garrafas pequenas de 20cl ou algumas garrafas grandes de 2 litros, além das garrafas mais convencionais de 750ml. A variação no tamanho não tem impacto no resultado básico. (Devo dizer que, para manter o rigor, estou no meio do processo de verificar sozinho o resultado desta prateleira de vinhos. Tenha certeza de que nenhum detalhe foi esquecido na pesquisa deste livro.)

Por mais notável que tudo isto possa ser, a matemática não é realmente o que importa para os nossos propósitos aqui. O que é importante é que reconheçamos que a relação fundamental existe. Compreender a ligação inextricável entre essas métricas é uma das ferramentas mais poderosas à nossa disposição em termos de design de processos previsíveis.

Mas antes de podermos entrar na forma como a Lei de Little pode nos ajudar com a previsibilidade, é provavelmente útil expor primeiro a relação em termos mais familiares.

A Lei de Little de uma Perspectiva Diferente

No final da década de 1980 (ou início da década de 1990, dependendo de quem você pergunta), a Lei de Little foi usurpada pela comunidade de Gestão de Operações (GO) e alterada para enfatizar o foco da GO na Vazão. O povo da GO mudou assim os termos da Lei de Little para refletir a sua perspectiva diferente, conforme mostrado pela Equação (2):

Equação (2): Tempo de Ciclo (CT) = Trabalho em Andamento (WIP) / Vazão (TH)

Onde:

1. Tempo de Ciclo (CT) = tempo médio que leva para um item fluir pelo sistema.
2. Trabalho em andamento (WIP) = o estoque total médio no sistema.
3. Vazão (TH) = vazão média do sistema.

No interesse de sermos completos, não há problema em realizar a álgebra da Lei de Little para que ela assuma as formas diferentes, mas ainda válidas:

Equação (3): TH = WIP / CT

e

Equação (4): WIP = CT * TH

Onde CT, WIP e TH são definidos da mesma forma que na Equação (2).

Devido às suas raízes na Gestão de Operações, a comunidade de trabalho do conhecimento Lean e Kanban adotou esta forma de “Vazão” da Lei de Little como sua. Se você já viu a Lei de Little antes, é quase certo que já a viu na forma da Equação (2) – embora a Equação (2) não represente o formato original da lei.

O resultado da Lei de Little é que, em geral, quanto mais coisas você trabalhar em um determinado momento (em média), mais tempo levará para cada uma dessas coisas terminar (em média), ceteris paribus (mantidas inalteradas todas as outras coisas). Como exemplo, os gestores que desconhecem esta lei entram em pânico quando percebem que os seus Tempos de Ciclo são demasiado longos e realizam a intervenção exatamente oposta ao que deveriam fazer: iniciam mais trabalho. Afinal, eles raciocinam, se as coisas demoram tanto, então eles precisam iniciar novos itens o mais rápido possível para que esses itens terminem no prazo – independentemente do que esteja em andamento no momento. O resultado é que os itens demoram cada vez mais para serem concluídos. Assim, os gestores sentem cada vez mais pressão para começar as coisas cada vez mais cedo. Você pode ver como esse ciclo vicioso começa e se perpetua. Após estudar a Lei de Little, você deve perceber que se os tempos de ciclo forem muito longos, a primeira coisa que você deve considerar é reduzir o WIP. É desconfortável, mas é verdade.

O que o Dr. Little demonstrou é que as três métricas de fluxo são essencialmente três lados da mesma moeda (se uma moeda pudesse ter três lados). Ao alterar um deles, você afetará um ou os outros dois. Em outras palavras, a Lei de Little revela quais alavancas podemos usar ao realizar a melhoria de processos. Além disso, como veremos em breve, a Lei de Little irá sugerir as intervenções específicas que devemos explorar quando o nosso processo não funciona da forma que desejamos.

Correndo o risco de me repetir, o que estou falando aqui é um fato matemático simples e incontestável. Uma mudança em uma métrica resulta

em uma mudança nas outras. A maioria das empresas com quem converso, que se queixam de baixa previsibilidade, quase sempre ignoram as implicações negativas que um excesso de WIP provoca no Tempo de Ciclo ou na Vazão. Ignore essa correlação por sua própria conta e risco.

É Tudo uma Questão de Suposições

Tudo isso parece bastante simples, certo? Bem, infelizmente, não é. Lembre-se de que eu disse no início que a Lei de Little é ilusoriamente simples? É aqui que as coisas ficam mais complicadas.

É fácil ver, de uma perspectiva puramente matemática, que a Equação (1) é logicamente equivalente à Equação (2). Mas é mais importante focar na diferença entre as duas. Como mencionei anteriormente, a Equação (1) é expressamente declarada em termos da *Taxa de Chegada* ao sistema, enquanto a Equação (2) é expressamente declarada em termos da *Taxa de Saída* do sistema. Esta ênfase na Vazão na Equação (2) provavelmente nos parece mais confortável, pois reflete a perspectiva usual de um processo de trabalho do conhecimento. Normalmente, no nosso contexto, nós (e, mais importante, os nossos clientes) nos preocupamos com o ritmo que estamos terminando o nosso trabalho (embora, como veremos em breve, devamos nos preocupar igualmente com o ritmo que começamos a trabalhar). O que é bom saber é que a Lei de Little pode se transformar para corresponder a essa perspectiva necessária.

À primeira vista, esta mudança pode não parecer tão significativa. No entanto, esta transformação da perspectiva das chegadas para a perspectiva das partidas tem um impacto profundo em termos de como pensamos e aplicamos a lei. Quando enunciamos a Lei de Little em termos da Vazão de um sistema, devemos também considerar imediatamente quais

os pressupostos subjacentes que devem existir para que a lei orientada para a partida seja válida.

Anteriormente, quando apresentei a Equação (1), afirmei que havia realmente apenas uma suposição que precisava ser implementada para ela funcionar. Bem, no interesse da integridade, tecnicamente existem três. Para a Equação (1) precisamos:

1. Um regime estacionário (ou seja, em que os processos estocásticos subjacentes sejam estáveis)
2. Um período arbitrariamente longo sob observação (para garantir regime estacionário dos processos estocásticos subjacentes)
3. Que o cálculo seja realizado utilizando unidades consistentes (por exemplo, se o tempo de espera for indicado em dias, então a Taxa de Chegada também deve ser indicada em dias).

A propósito, o objetivo aqui não é lhe dar um diploma avançado em estatística ou teoria das filas. Não se preocupe se não souber o que significa “estocástico” ou “estacionário”. Você não precisa. Como acabei de dizer, menciono essas coisas apenas por completude.

Contudo, quando mudamos a perspectiva para olhar para a Lei de Little a partir da perspectiva da Vazão e não da perspectiva da Taxa de Chegada, precisamos também de alterar os pressupostos necessários para que a lei seja válida. Este ponto é tão importante que quero colocá-lo em destaque:

Olhar para a Lei de Little da perspectiva da Vazão e não da perspectiva da Taxa de Chegada exige uma mudança nos pressupostos necessários para que a lei seja válida.

Em sistemas onde o WIP nunca chega a zero (consulte AAMFP para uma discussão mais completa do caso em que o WIP pode chegar a zero), então as suposições sobre o nosso processo, necessárias para fazer a Lei de Little (na forma da Equação (2)) funcionar são:

1. A entrada média ou Taxa de Chegada (λ) deve ser igual à saída média ou Taxa de Partida (Vazão).
2. Todo o trabalho iniciado será por fim concluído e sairá do sistema.
3. A quantidade de WIP deve ser aproximadamente a mesma no início e no final do intervalo de tempo escolhido para o cálculo.
4. A idade média do WIP não está nem aumentando, nem diminuindo.
5. O Tempo de Ciclo, o WIP e a Vazão devem ser medidos usando unidades consistentes.

As duas primeiras suposições (#1 e #2) compreendem uma noção conhecida como Conservação do Fluxo. As duas segundas suposições (3 e 4) referem-se à noção de estabilidade do sistema.

A última suposição (#5) é necessária para que a matemática (e qualquer análise correspondente) resulte corretamente (você notará que esta é a mesma suposição necessária ao estabelecer a lei em termos de chegadas). A necessidade de usar unidades consistentes ao realizar um cálculo da Lei de Little deveria ser intuitivamente óbvia, mas é bastante fácil tropeçar nisso. Quando dizemos unidades “consistentes”, o que realmente estamos dizendo é, por exemplo, se estamos medindo o Tempo de Ciclo médio usando a unidade de tempo “dia”, então a Vazão média deve estar na forma do número de itens por essa mesma unidade de tempo (dia), e o WIP médio deve ser a quantidade média de itens para uma unidade de tempo (dia). Em outro exemplo, se você deseja medir a Vazão média em termos de itens por semana (ou seja, a unidade de tempo aqui é “semana”), o Tempo de Ciclo médio deve ser declarado em termos de semanas, e o WIP médio deve ser a média para cada semana.

Você pode pensar que estou perdendo seu tempo mencionando isso, mas ficaria surpreso com quantas equipes não entendem esse ponto (me lembro de quando a NASA colidiu um orbitador na superfície de Marte porque uma equipe usou unidades métricas enquanto outra equipe usou unidades imperiais – moral da história: não faça isso). Por exemplo, vi uma equipe Scrum que estava medindo sua velocidade em termos de pontos de história por sprint (como as equipes Scrum costumam fazer, o que é uma pena). Para o cálculo da Lei de Little, eles inseriram seu número de velocidade para Vazão, seu número WIP como o número total de histórias de usuários (histórias reais – não pontos de história) concluídas no sprint e esperavam obter um número de Tempo de Ciclo em dias. Você pode imaginar a surpresa deles quando os números não saíram exatamente como eles esperavam.

Suposições como Políticas de Processo

Compreender essas suposições fundamentais é de uma importância monumental. Apesar do que muitas pessoas lhe dirão, o verdadeiro poder da Lei de Little não está em realizar cálculos matemáticos inserindo números em sua fórmula. Embora eu já tenha passado tanto tempo nisso, reitero que você deve esquecer a aritmética. Na verdade, a maioria de nós nunca terá necessidade de calcular a Lei de Little. Como mencionei no Capítulo 6, os dados das quatro métricas de fluxo são tão fáceis de capturar que você nunca deveria ter que computar nenhuma delas – basta olhar os dados!

Em vez disso, o verdadeiro poder da Lei de Little reside, em primeiro lugar, na compreensão dos pressupostos necessários para que a lei funcione. Se há três coisas que eu quero que você entenda desta conversa sobre a Lei de Little, são elas:

1. É tudo uma questão de suposições.
2. É tudo uma questão de suposições.
3. É tudo uma questão de suposições.

Cada vez que você violar uma suposição da Lei de Little, seu processo se tornará menos previsível. Toda vez. Essa maior imprevisibilidade pode se manifestar como Tempos de Ciclo mais longos ou mais variabilidade de processo, ou menos Vazão, ou WIP mais alto, ou todas as alternativas acima. Pior ainda, essas violações podem nem aparecer imediatamente nos seus dados. Durante todo o tempo em que você viola a Lei de Little, seus dados podem mostrar uma imagem do mundo mais otimista do que realmente ocorre. O perigo aqui é que você possa estar baseando algumas previsões nesta visão excessivamente otimista – apenas para descobrir que as coisas são muito piores do que pareciam.

É claro que vivemos no mundo real e haverá momentos em que violar estes pressupostos será inevitável ou mesmo necessário. Mas é exatamente por isso que é ainda mais importante compreender as implicações quando estas violações ocorrem. Sempre acontecem coisas conosco que estão fora do nosso controle. No entanto, a última coisa que queremos é agravar esses acontecimentos incontroláveis, permitindo que aconteçam coisas ruins que estavam sob o nosso controle e que poderiam ter sido facilmente evitadas. Controle o que você pode controlar e tente eliminar ou mitigar o que não pode.

As suposições acima (especialmente as quatro primeiras) vão nos ajudar a fazer exatamente isso. Podemos usar essas suposições como base para algumas políticas simples que regerão a operação do nosso processo. Estas políticas servirão para controlar as coisas que podemos controlar. Estas políticas servirão para tornar o nosso processo mais previsível.

Com base nas suposições acima, algumas políticas de processo podem incluir (mas certamente não estariam limitadas a):

- Só iniciaremos novos trabalhos aproximadamente na mesma proporção em que terminamos trabalhos antigos.
- Faremos todos os esforços razoáveis para terminar todo o trabalho iniciado e minimizar o esforço desperdiçado devido a itens de trabalho descartados (isso exigirá alguma noção de "compromisso" tardiamente vinculado).
- Se o trabalho ficar bloqueado, faremos tudo o que pudermos para desbloqueá-lo o mais rapidamente possível.
- Monitoraremos de perto nossas políticas em relação à ordem em que puxamos os itens para o nosso sistema, de forma que os itens de trabalho não fiquem parados e envelheçam desnecessariamente.

O desenho do seu processo é, na verdade, apenas a soma de todas as políticas que você implementou. O desempenho bom ou ruim do seu sistema é diretamente atribuível a essas políticas e ao quão bem você adere ou não a elas. Quando falo sobre projetar para previsibilidade, estou falando sobre fornecer algumas perspectivas sobre políticas apropriadas que você pode incorporar na operação diária do seu processo. Estas políticas servirão para normalizar e estabilizar o seu sistema, a fim de dar ao seu processo a previsibilidade que você procura. É somente a partir desta base estável que podemos esperar implementar melhorias reais e duradouras nos processos.

Como Frank Vega gosta de dizer com frequência, “as suas políticas moldam os seus dados e os seus dados moldam as suas políticas”. As políticas que mencionei aqui influenciarão na maioria os dados que você coleta em seu processo. A propósito, isso é uma coisa boa. É boa porque esses dados por si só irão potencialmente sugerir ainda mais onde as nossas políticas de processo são deficientes. É deste ciclo virtuoso que estou falando quando digo “métricas acionáveis para previsibilidade”.

Sistemas Kanban

Do ponto de vista do WIP, pode parecer que a execução de um sistema Kanban garante que as suposições da Lei de Little sejam atendidas. Existem vários motivos pelos quais isso pode não ser o caso:

1. É possível que a alteração dos limites de WIP não influencie o WIP médio total (por exemplo, diminuir ou aumentar um limite de WIP após um claro gargalo sistêmico). Esta pode ser uma das razões pelas quais você não obtém o comportamento “previsto” que esperaria da Lei de Little.
2. Definir um limite de WIP não é necessariamente o mesmo que limitar o Trabalho em Andamento. Não sei dizer quantas equipes encontrei que estabelecem limites de WIP, mas depois os violam rotineira e flagrantemente.
3. O WIP médio durante um período depende muito das políticas de puxar em vigor. Por exemplo, o maior número possível de itens é puxado para satisfazer os limites de WIP em todos os momentos?

A questão aqui é que se você estiver usando um sistema Kanban, você não pode simplesmente somar todos os limites de WIP em seu quadro e pensar que calculou o WIP para seu processo. Você terá que rastrear o WIP físico, real.

Por último, a maioria das pessoas pensa que a Lei de Little é o maior motivo para implementar um processo ágil no estilo Kanban. Embora eu não discorde estritamente dessa afirmação, ofereceria uma maneira melhor de afirmá-la. Eu diria que a Lei de Little é a maior razão para migrar para um processo de fluxo mais controlado por WIP e baseado em um sistema

puxado. A questão é que, uma vez feito isso, poderemos começar a usar a Lei de Little como nosso guia para a previsibilidade do processo.

Tamanho Não Importa

Tenho um último tópico que quero abordar antes de encerrar. Observe como nas suposições da Lei de Little não mencionei a exigência de que todos os itens de trabalho fossem do mesmo tamanho. Isso ocorre porque tal exigência não existe. A maioria das pessoas presume que uma aplicação específica da Lei de Little – e uma limitação do WIP, em geral – exige que todos os itens de trabalho sejam do mesmo tamanho. Isso simplesmente não é verdade (ver Capítulo 3).

A primeira razão pela qual o tamanho dos itens de trabalho não importa é porque com a Lei de Little estamos lidando com relações entre médias. Não nos preocupamos necessariamente com cada item individualmente, nos preocupamos com a aparência média de todos os itens.

Em segundo lugar, e mais importante, a variabilidade no tamanho do item de trabalho provavelmente não é a variabilidade que está acabando com a sua previsibilidade. Seus maiores problemas de previsibilidade são geralmente o excesso de WIP, a frequência com que você viola as suposições da Lei de Little, etc.

Geralmente, esses são problemas mais fáceis de resolver do que tentar arbitrariamente tornar todos os itens de trabalho do mesmo tamanho. Mesmo se você estivesse em um contexto em que o tamanho importasse, seria mais uma questão de dimensionar corretamente o seu trabalho e não de dimensioná-los todos no mesmo tamanho (Capítulo 3).

Previsão

Poder ser que você estivesse esperando todo esse tempo que eu dissesse que, após entender a Lei de Little, tudo o que você precisa fazer é inserir os números e surgirá o resultado da previsão que você está procurando (à la F = ma de Newton ou E = mc2 de Einstein). No entanto, nada poderia estar mais longe da verdade.

A este respeito, é importante saber que a Lei de Little se preocupa apenas em olhar para trás, para um período que já passou. Não se trata de olhar para frente; isto é, não se destina a ser usada para fazer previsões determinísticas. Como o próprio Dr. Little diz sobre a lei: “Isso não é de todo ruim. Diz apenas que estamos no negócio de medição, não no negócio de previsões”.

Este ponto requer um pouco mais de discussão, pois geralmente é onde as pessoas ficam presas. A parte “lei” da Lei de Little especifica uma relação exata entre WIP médio, Tempo de Ciclo médio e Vazão média, e esta “lei” só se aplica quando você está analisando dados históricos. A lei não se destina – e nunca foi concebida para isso – a fazer previsões determinísticas sobre o futuro. Por exemplo, vamos supor uma equipe que historicamente teve um WIP médio de 20 itens de trabalho, um Tempo de Ciclo médio de 5 dias e uma Vazão média de 4 itens por dia. Você não pode dizer que aumentará o WIP médio para 40, manterá o Tempo de Ciclo médio constante em 5 dias e magicamente a Vazão aumentará para 8 itens por dia — mesmo se você adicionar pessoal para manter a proporção de WIP por pessoa igual nas duas instâncias. Você não pode presumir que a Lei de Little fará essa previsão. Não vai. Tudo o que a Lei de Little dirá é que um aumento no WIP médio resultará em uma mudança em um ou ambos, no Tempo de Ciclo médio e na Vazão média. Dirá ainda que essas mudanças se manifestarão de formas tais que a relação entre as três métricas ainda obedecerá a essa lei. Mas o que a lei não diz é que se pode prever de forma

determinística quais serão essas mudanças. É preciso esperar até o final do intervalo de tempo de seu interesse e olhar para trás para aplicar a lei.

Mas essa restrição não é grave. A aplicação adequada da Lei de Little no nosso mundo é compreender os pressupostos da lei e desenvolver políticas de processo que correspondam a esses pressupostos. Se o processo que operamos estiver em conformidade – ou praticamente em conformidade – com todos os pressupostos da lei, então chegaremos a um mundo onde poderemos começar a confiar nos dados que recolhemos do nosso sistema. É neste ponto que nosso processo é probabilisticamente previsível. Uma vez lá, podemos começar a usar algo como a simulação de Monte Carlo nos nossos dados históricos para fazer previsões e, mais importante, podemos ter alguma confiança nos resultados que obtemos usando esse método.

Existem outras razões mais fundamentais pelas quais não se deve usar a Lei de Little para fazer previsões. Por um lado, espero que já tenha internalizado que a Lei de Little é uma relação de médias. Menciono isto novamente porque mesmo que você pudesse usar a Lei de Little como uma ferramenta de previsão (o que não é possível), você não iria querer fazê-lo, pois estaria produzindo uma previsão baseada em médias. Existem vários motivos pelos quais você não deve fazer previsões com base em médias – muitos para serem abordados aqui (veja o livro do Dr. Savage, “The Flaw of Averages”). Acontece que temos algo melhor que a média. Ao coletar dados de métricas, teremos ferramentas muito melhores à nossa disposição quando estivermos prontos para fazer previsões.

Dito tudo isto, porém, não há razão para que não se possa usar a lei para estimativas rápidas e simples sobre o futuro. Claro que você pode fazer isso. No entanto, eu não assumiria quaisquer compromissos, decisões de contratação ou demissão de pessoal, nem cálculos de custos de projetos baseados apenas neste tipo de estimativa. Eu diria ainda que é negligência alguém sugerir fazê-lo. Mas este cálculo simples pode ser útil como uma

verificação rápida para decidir se algo como um projeto vale a pena ser explorado mais a fundo.

Lembre-se de que ser previsível não significa apenas fazer previsões. Uma boa parte da previsibilidade é operar um sistema que se comporta da maneira que esperamos. Ao projetar e operar um sistema que segue os pressupostos estabelecidos pela Lei de Little, obteremos exatamente isso: um processo que se comporta da maneira que esperamos. Isso significa que teremos controlado aquilo que podemos controlar e que as intervenções que realizarmos para melhorar as coisas resultarão em resultados mais alinhados com as nossas expectativas.

Conclusão

Sei que já disse isso antes, mas preciso repetir: a Lei de Little não trata da compreensão da matemática da teoria das filas. Trata-se de compreender os pressupostos que precisam estar em vigor para que a lei funcione. Podemos usar essas suposições como guia, esquema ou modelo para nossas próprias políticas de processo. Sempre que suas políticas de processo violam os pressupostos da Lei de Little, você sabe que pelo menos diminuiu – ou possivelmente eliminou – sua chance de ser previsível.

À medida que você opera seu processo, pense nos momentos e nas razões pelas quais o trabalho entra em um ritmo mais rápido do que o trabalho sai. Pense por que os itens envelhecem desnecessariamente devido a bloqueios ou políticas de puxar inadequadas. Considere por que o trabalho é abandonado quando está apenas parcialmente concluído (e como você explica esse abandono). Pondere como essas ocorrências violam as suposições da Lei de Little e como, em última análise, afetam sua capacidade de ser previsível. Mas o mais importante é pensar em como a sua compreensão da Lei de Little deve resultar em mudanças de

comportamento para você e sua equipe. Quando ocorrem violações da Lei de Little, geralmente é devido a algo que você fez ou escolheu (intencionalmente ou não) não fazer.

Lembre-se de que você tem muito mais controle sobre seu processo do que imagina.

Bibliografia

Bertsimas, D., D. Nakazato. The distributional Little's Law and its applications. Operations Research. 43(2) 298–310, 1995.

Brumelle, S. On the relation between customer and time averages in queues. J. Appl. Probab. 8 508–520, 1971.

Coleman, John and Vacanti, Daniel S. "The Kanban Guide" Kanban-Guides.org, 2020.

Deming, W. Edwards. The New Economics. 2nd Ed. The MIT Press, 1994.

Deming, W. Edwards. Out of the Crisis. The MIT Press, 2000.

Glynn, P. W., W. Whitt. Extensions of the queuing relations L = λ W and H = λ G.

Operations Research. 37(4) 634–644, 1989. Goldratt, Eliyahu M., and Jeff Cox. The Goal. 2nd Rev. Ed. North River Press, 1992.

Heyman, D. P., S. Stidham Jr. The relation between customer and time averages in queues. Oper. Res. 28(4) 983–994, 1980.

Hopp, Wallace J., and Mark L. Spearman. Factory Physics. Irwin/McGraw-Hill, 2007.

Little, J. D. C. A proof for the queuing formula: L = λ W. Operations Research. 9(3) 383–387, 1961.

Little, J. D. C., and S. C. Graves. “Little’s Law.” D. Chhajed, T. J. Lowe, eds. Building Intuition: Insights from Basic Operations Management Models and Principles. Springer Science + Business Media LLC, New York, 2008.

Ripley, Ryan and Miller, Todd. Fixing Your Scrum. The Pragmatic Programmers LLC, 2020.

Reid, Steve and Singh, Prateek and Vacanti, Daniel “Ultimate Kanban: Scaling Agile without Frameworks at Ultimate Software” Infoq.com, 2016.

Reinertsen, Donald G. Managing the Design Factory. Free Press, 1997.

Reinertsen, Donald G. The Principles of Product Development Flow. Celeritas Publishing, 2009.

Savage, Sam L. The Flaw of Averages. John Wiley & Sons, Inc., 2009.

Schwaber, Ken and Sutherland, Jeff “The Scrum Guide” Scrumguides.org, 2020.

Shewhart, W. A. Economic Control of Quality of Manufactured Product, 1931.

Shewhart, W. A. Statistical Method from the Viewpoint of Quality Control, 1939.

Stidham, S., Jr. L = λ W: A discounted analogue and a new proof. Operations Research. 20(6) 1115–1126, 1972.

Stidham, S., Jr. A last word on L= λ W. Operations Research. 22(2) 417–421, 1974.

Vacanti, Daniel S. “Actionable Agile Metrics for Predictability”. ActionableAgile Press. 2014.

Vacanti, Daniel S. When Will It Be Done? ActionableAgile Press, 2017.

Vacanti, Daniel S. and Bennet Vallet. “Actionable Metrics at Siemens Health Services”. AgileAlliance.com. 1 Aug 2014.

Vallet, Bennet. “Kanban at Scale: A Siemens Success Story.” In- foq.com. 28 Feb 2014.

Vega, Frank. “Are You Just an Average CFD User?” Vissinc.com. 21 Feb 2014.

Vega, Frank. “The Basics of Reading Cumulative Flow Diagrams”. Vissinc.com. 29 Sep 2011.

Wheeler, Donald J., and David S. Chambers. Understanding Statis- tical Process Control. 2nd Ed. SPC Press, 1992.

Wikipedia “Monte Carlo method.” Wikipedia.com 01 Aug 2014.