FIAP ON

ENSINO A DISTÂNCIA

BIG DATA & ANALYTICS

INOVANDO COM **BIG DATA**

Anderson Paulucci e Leandro Rubim

CAPÍTULO 2

LISTA DE FIGURAS

SUMÁRIO

2 INOVANDO COM O BIG DATA

2.1 A evolução computacional até o Big Data

Big Data não é uma tecnologia e não se trata apenas de um grande volume de dados. O termo Big Data remete aos conceitos que abordamos até o momento e enfatiza a grande transformação que estamos presenciando nesta nova era baseada em dados.

A seguir, temos 3 definições de Big Data reconhecidas por consultorias e especialistas no mercado (apud TAURION, 2013):

“Intensa utilização de redes sociais online, de dispositivos móveis para conexão à internet, transações e conteúdos digitais e também o crescente uso de computação em nuvem tem gerado quantidades incalculáveis de dados. O termo Big Data refere-se a este conjunto de dados cujo o crescimento é exponencial e cuja a dimensão está além da habilidade das ferramentas típicas de capturar, gerenciar e analisar dados.” *McKinsey Global Institute*

“Big Data é o termo adotado pelo mercado para descrever problemas de gerenciamento e processamento de informações extremas as quais excedem a capacidade das tecnologias tradicionais ao longo de uma ou varias dimensões.” *Gartner*

“A interseção de ferramentas de análise de dados scale-out com armazenamento de dados scale-out.” *Rob Peglar*

Gostaria de chamar a atenção para a terceira definição (de Rob Peglar), por ser uma abordagem mais técnica. E, definitivamente, é a melhor para compreendermos o impacto de uma grande mudança de arquitetura. As soluções criadas para atender Big Data irrefutavelmente possuem estas características nativas.

As soluções tradicionais adotadas no mundo corporativo são arquiteturas de segunda geração em sua grande maioria, baseadas em um modelo cliente-servidor e não foram projetadas para escalabilidade horizontal. Algumas sofreram adaptações para acompanhar a evolução e atendem a computação distribuída com muitas limitações.

Ao longo dos últimos anos tivemos evoluções importantes na computação distribuída, como o uso mais intensivo de soluções em grid e softwares open-source.

O software cliente-servidor foi originalmente desenvolvido para atender a computação de escalabilidade relativamente limitada, por exemplo:

- 1 servidor.
- 1 instância.
- 1 base de dados local.

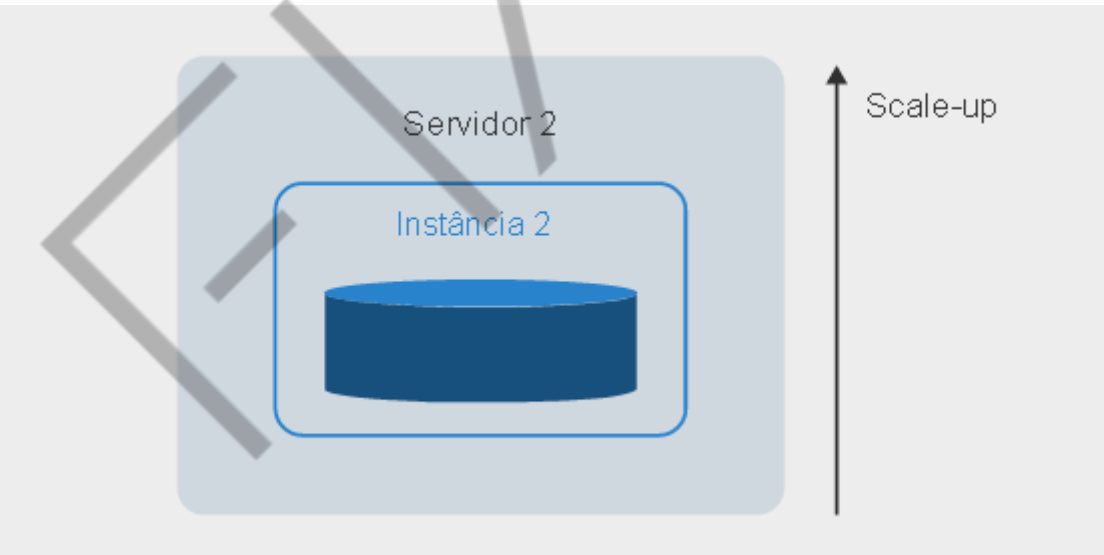

Figura 1 – Solução Cliente Servidor
Fonte: Elaborado pelo autor (2016), adaptado por FIAP (2017).

Este modelo de arquitetura *standalone* contribuiu para:

- Aumento do número de silos de informações.
- Baixa e/ou desigual utilização de recursos.

- O servidor se tornou um ponto único de falha (SPOF) tanto para computação, quanto armazenamento.

Os datacenters evoluíram com o a necessidade de aumentar o volume e garantir a disponibilidade dos dados. As soluções baseadas em Storage remoto ajudaram as empresas com uma melhor utilização (eficiência) do armazenamento e proteção de dados.

Aplicações distribuídas passaram a adotar este modelo baseado em um subsistema de armazenamento compartilhado e centralizado:

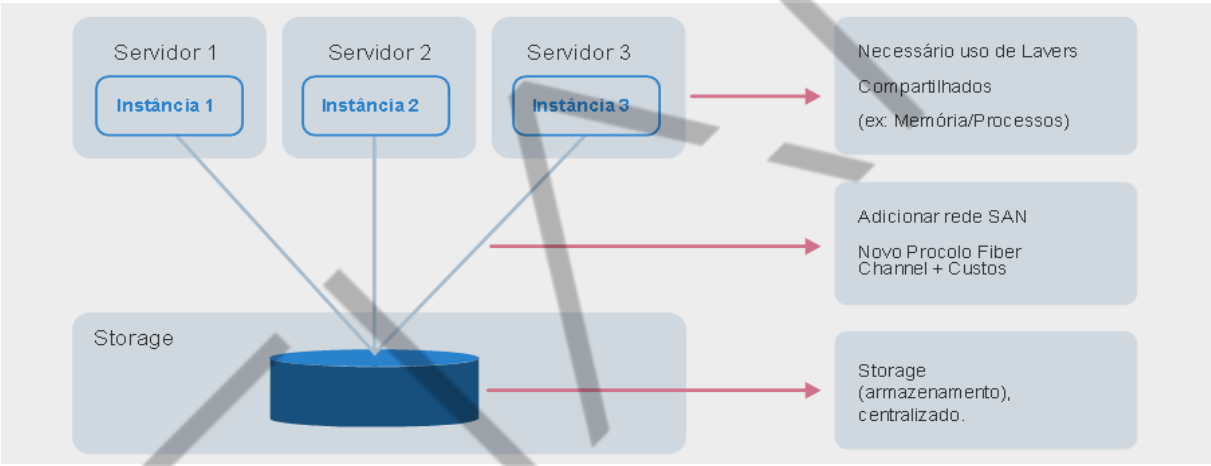

Figura 2 – Solução Distribuída (Storage Centralizado)
Fonte: Elaborado pelo autor (2016), adaptado por FIAP (2017).

Na era digital, com o crescimento acelerado e desordenado dos datacenters, o *storage* incluindo rede SAN, passou a ser um grande desafio, principalmente devido ao custo e dificuldades de evoluir seu *roadmap*.

Apesar da grande maioria das aplicações cliente-servidor ter sofrido alguma adaptação para atender a cenários de computação distribuída, as decisões para adoção de escalabilidade das aplicações evitam a computação distribuída, adotado como primeira opção a escalabilidade vertical.

Qual é o motivo?

- Complexidade.
- Complexidade.

- Complexidade (Isso não foi um erro de digitação, cenários de computação distribuídas com storage centralizados e aplicações da era cliente-servidor são muito complexos).
- Custos.
- Instabilidade da solução.
- Escalabilidade não linear.
- Limitações de funcionalidades.
- Adaptações de códigos legados.

Estas são algumas das motivações. Vamos entender um pouco mais no detalhe o motivo principal.

Qualquer adaptação com esta finalidade de adicionar mais capacidade para a computação distribuída, afeta diretamente a estrutura principal de uma solução que não foi originalmente concebida para escalar, causando grandes impactos.

É como a construção de uma nova cidade, se podemos planejar uma nova infraestrutura e arquitetura sobre uma área ainda não ocupada. Usaremos o que há de mais moderno para atender aos padrões de arquitetura/engenharia atuais. Do contrário, se o plano é aumentarmos a capacidade de um grande centro já existente, mantendo a arquitetura estrutural, teremos prazos e custos maiores, e no final iremos conviver com os desafios das limitações do legado.

Desde os primórdios da computação, aprendemos que existe uma relação, chamada Lei de Moore, conceito criado em 1965 pelo então presidente da Intel Gordon E. Moore. Com base no crescimento do número de transistores nos chips teríamos uma capacidade de dobrar o poder computacional a cada 24 meses, aproximadamente. E isso realmente nunca parou.

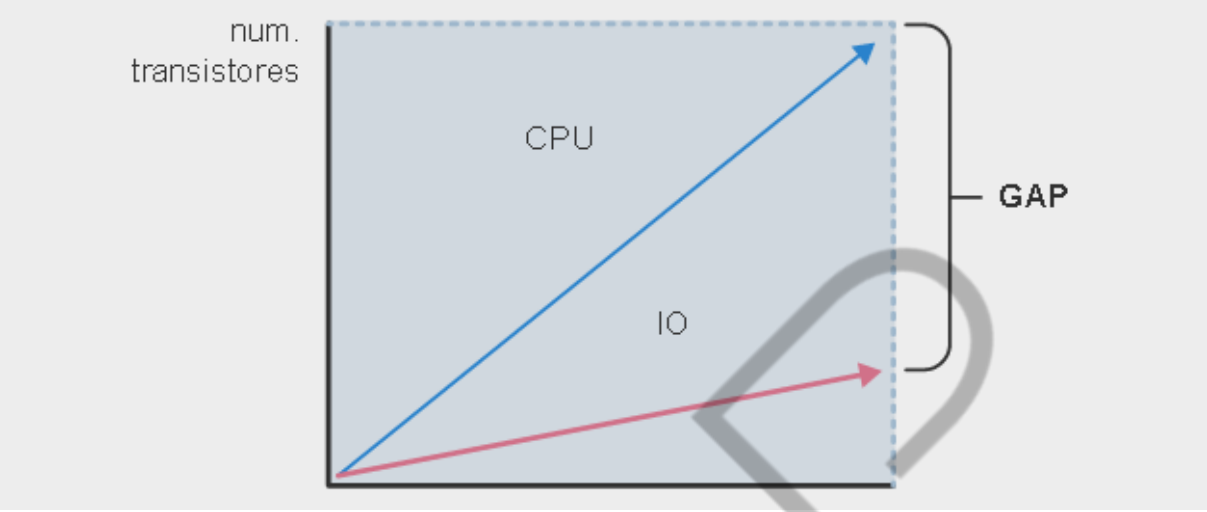

Figura 3 – Evolução Processamento vs I/O
Fonte: Elaborado pelo autor (2016)

Com relação à outra capacidade computacional, o I/O (entrada e saída no disco), a evolução não aconteceu da mesma forma, apesar de aumentarmos muito o volume e as demandas por velocidade no acesso aos dados.

Os *datacenters* de segunda geração foram projetados para trabalhar com armazenamento remoto usando equipamentos robustos conhecidos como *storage*, dessa forma, centralizamos um grande volume de dados e crescemos vertiginosamente durante duas décadas com esta arquitetura de armazenamento.

Os softwares e as tecnologias de gerenciamento de dados evoluíram bastante com os *storages*, porém a capacidade de I/O não acompanhou a evolução de CPU, e criamos um gap enorme para resolver problemas de I/O.

Permanecemos durante anos sem grandes evoluções, com padrões de conectividade SAN (Storage Area Network) entre a computação e o storage baseado em Fibre Channel, usando protocolos FC, inicialmente de alto Throughput, 4Gb/s e 8Gb/s. Os custos de uma rede SAN continuam altos, e mantêm os roadmaps de datacenters limitados a evolução.

Os discos usados para o armazenamento dos dados também não acompanharam os requisitos de performance nos últimos anos, e os datacenters permaneceram com padrões de HDs magnéticos, saturados fisicamente devido ao

superaquecimento do hardware, mantemos grande parte dos dados armazenados em discos SATA e SAS. A opção de armazenamento com discos SSDs está se tornando viável financeiramente e poderá ajudar a reduzir este impacto.

Figura 4 – SSDs
Fonte: Banco de imagens Shutterstock (2017).

Concluímos que existe um grande gap e o I/O é o principal ofensor. Imagine uma cidade como Alphaville no ano de 2000 ainda em fase de crescimento inicial, com estruturas bem definidas para condomínios residencial, comercial e empresarial. Ruas largas, atendendo satisfatoriamente a todo o tráfego de veículos, mesmo em horários de pico. As pessoas também transitando nos restaurantes e comércios com grande facilidade. Em uma década, o cenário já havia mudado bastante, o crescimento expressivo, rapidamente saturou a capacidade de acompanhar o roadmap de infraestrutura necessária para manter o padrão. O colapso nas vias de locomoções foi inevitável (comparando com redes SANs), apesar da infraestrutura de condomínios continuarem crescendo e evoluindo.

Desta forma, podemos usar esta analogia e entender como nossos datacenters cresceram nos últimos anos (exponencialmente) e a deficiência criada com a complexidade e custos de redes SANs com uso de storage remoto, dificultando o roadmap evolutivo, resultou na saturação das capacidades computacionais, mesmo com a garantia da lei de Moore.

Para que possamos resolver este gap, uma proposta é a necessidade de aproximar a computação dos dados e usar tudo o que aprendemos nos últimos anos com os gigantes da internet para quebrar este paradigma.

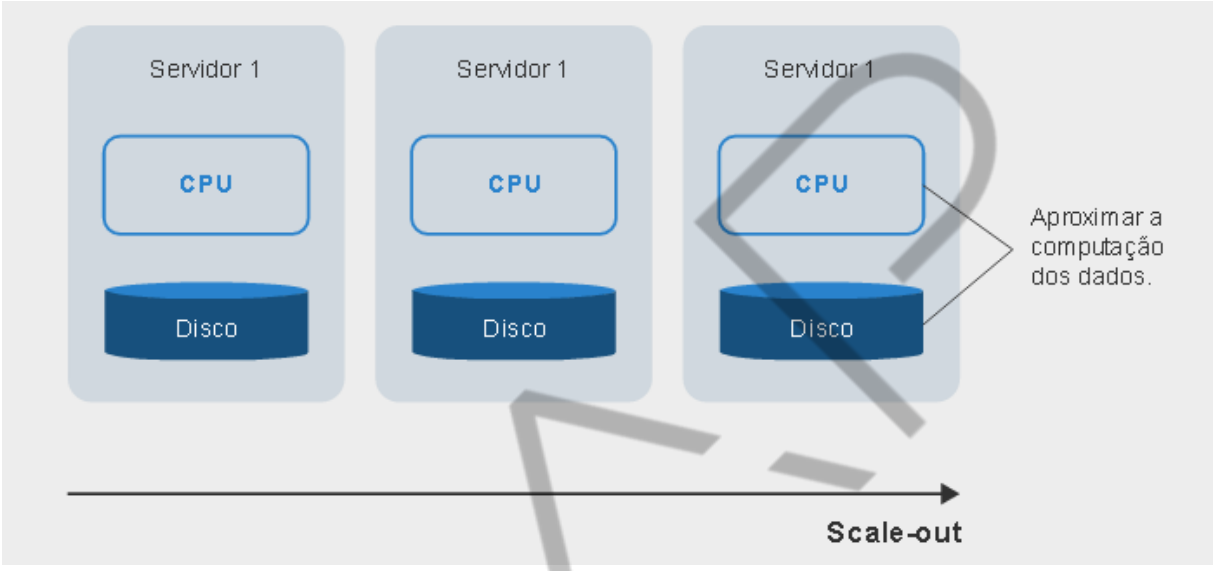

Figura 5 – Solução scale-out
Fonte: Elaborado pelo autor (2016), adaptado por FIAP (2017).

Soluções de Big Data que requerem alta escalabilidade, como Hadoop, Spark e databases NoSQL, do qual você verá mais a frente, foram concebidas com este modelo de arquitetura. E isso faz muita diferença, pois não há “puxadinhos” ou “remendos” na arquitetura para resolver a computação distribuída. Efetivamente foram construídas para atender Big Data.

Estas plataformas possuem as seguintes características:

- Escalabilidade Linear.
- Baixa complexidade, considerando a possibilidade de trabalhar com centenas e milhares de nós em cluster, gerenciando Peta Bytes de dados.
- Hardware commodities.

- Baixo Custo.
- Software Open Source.

SCALE-UP: Escalabilidade Vertical (Modelo da arquitetura de computação da geração cliente-servidor):

- Adicionar mais recursos no mesmo nó de computação (ex.: memória, discos).
- Arquitetura criada para suportar cenários de baixa escalabilidade.
- Não atendem efetivamente as soluções de cloud e big data.

SCALE-OUT: Escalabilidade Horizontal (Modelo da arquitetura de computação para geração de cloud computing):

- Adicionar mais nós ao cluster (sistema).
- Arquitetura para suportar cenários de alta escalabilidade.

A arquitetura de terceira geração deve ser baseada em scale-out e atender as exigências cada vez maiores de escalabilidade on-demand.

O termo **Web-Scale IT** é baseado no modelo de arquitetura scale-out e define novos padrões de arquitetura – referência adotada pelas gigantes da Internet para entregar seus serviços online praticamente sem interrupções, o que permite que essas empresas ganhem tempo (agilidade) e reduzam seus custos (Qual empresa não tem estes objetivos?).

De fato, as empresas tradicionais, diferentes das empresas digitais, aprenderam a evoluir com outros padrões de arquitetura, e para implementar esse conceito, é preciso adotar uma visão aberta em relação ao hardware e software, agilidade nos processos, uma cultura de colaboração forte e aceitar correr riscos em função da inovação.

2.2 Inovando com Big Data

Agora que aprendemos sobre os 3 Vs, vamos entender como acontece a disrupção de tecnologia, comparando um cenário semelhante.

O livro “The Innovator's Dilemma” de Clayton Christensen, professor de Harvard, descreve que as empresas que tiveram sucesso em uma geração de inovação, inevitavelmente, serão paralisadas por seu próprio sucesso e, portanto, condenadas a perder na próxima onda de inovação.

Já vimos isso acontecer com grandes empresas, e um exemplo clássico é a Kodak e a fotografia. Os nativos digitais não vivenciaram essa época, mas há pouco tempo, a fotografia era limitada a uma tecnologia analógica. As máquinas dependiam de um filme, material responsável em armazenar a imagem, limitado a 12/24/32 posições (o filme está para o cartão de memória assim como o a fita cassete está para o MP3). O processo de reproduzir a imagem no papel dependia de uma alta latência, após todas as posições do filme serem ocupadas por cada foto, era necessário ir até um laboratório e submeter as imagens do filme a um processo químico com prazos mínimos de 24h (D-1) para revelação (execução).

Figura 6 – Filmes para câmeras
Fonte: Google Images (2016)

Dizem que a Kodak criou a primeira câmera digital, e a gigante da fotografia não acreditou que a disrupção seria tão rápida e devastadora. Não precisamos dizer o que aconteceu com a Kodak, restaram apenas algumas patentes sobre tecnologias de câmeras digitais.

Após o ano 2000, já na era digital, o mundo não tinha mais espaço para fotografias analógicas, das quais mantemos apenas o charme sobre as lembranças daquele processo burocrático de produzir a fotografia.

Com as câmeras digitais, podemos produzir imagens com altíssima qualidade e em grande quantidade, um cartão de memória de celular é capaz de armazenar dezenas de milhares delas, com variedades que incluem vídeos e animações. É possível visualizar e compartilhar as fotos on-line em segundos, com pessoas em qualquer continente. A produção de vídeos e transmissão em real-time eleva a comunicação para um modelo cada vez mais interativo e sem fronteiras para a informação.

Com este exemplo, podemos ver claramente os 3 Vs de Big Data, que há pouco mais de uma década transformou a fotografia com o conceito de disrupção tecnológica. E agora estamos nos preparando para ver isso acontecer com as aplicações e principalmente com o analytics.

Empresas que não entenderem rapidamente este modelo disruptivo correm o mesmo risco que a Kodak, não terão tempo de reverter um cenário devastador.

2.2.1 Big Data é a próxima fronteira de Inovação

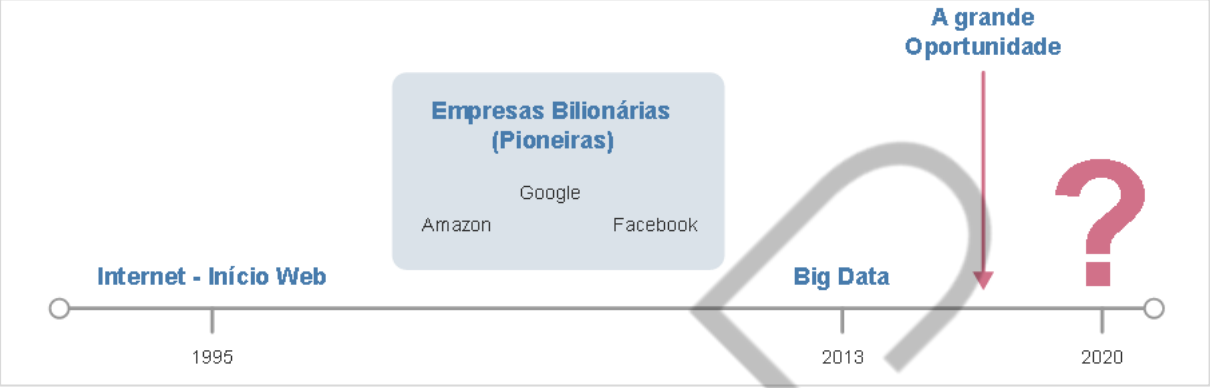

Figura 7 – Fronteira de Inovação
Fonte: FIAP (2017).

Podemos afirmar que estamos na idade das pedras de Big Data. Segundo o Trends do Google, o termo começou a ser usado efetivamente em 2012.

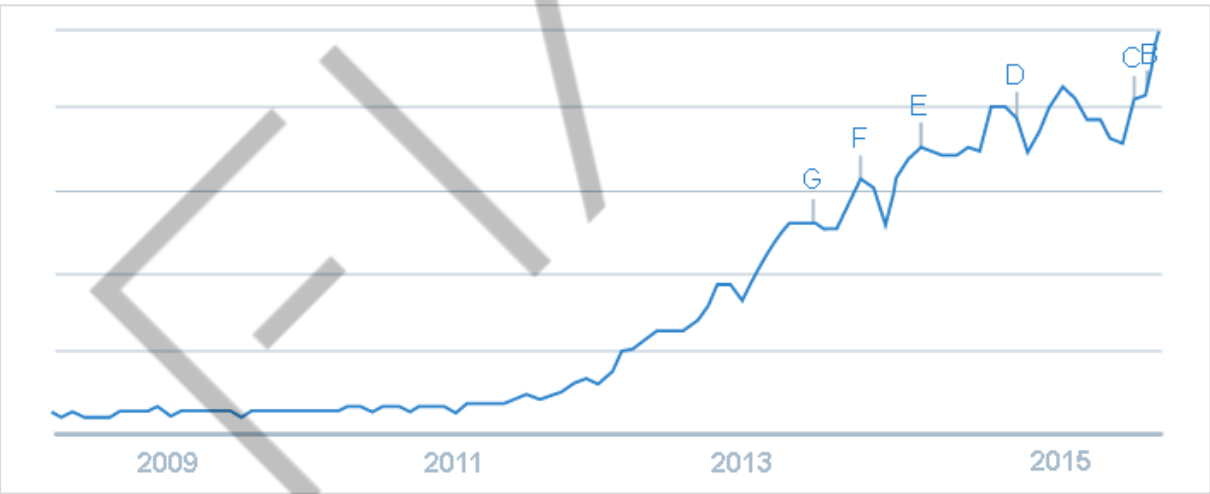

Figura 8 – Google Trends – Big Data
Fonte: Google Images (2017).

Não há dúvidas que será necessário avançarmos com a maturidade das soluções para suportarmos Big Data nas empresas. Mas já temos grandes referências que usam tecnologias de big data há mais de dez anos, as gigantes da internet Amazon, Google, Facebook entre outras, já estão trabalhando em um estágio mais avançado e preparando os próximos passos.

Sem dúvidas, as tecnologias são importantes para ultrapassarmos as barreiras da inovação. Mas outra questão relevante é o perfil das pessoas que desenvolvem e administram estas tecnologias. É preciso aprender com as empresas digitais e desenvolver um DNA de inovação.

A TI sobrevive ao caos operacional, passa o tempo todo apagando incêndios. A complexidade criada na segunda geração da computação contribui para um modelo pouco evolutivo. Falta tempo e dinheiro para a inovação.

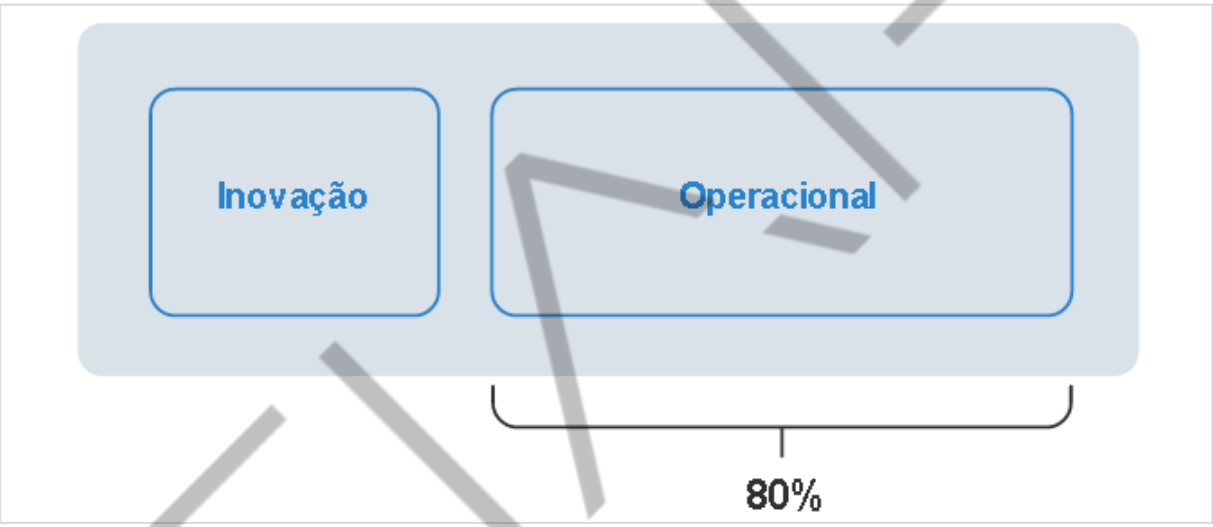

Figura 9 – IT Operacional
Fonte: Elaborado pelo autor (2016), adaptado por FIAP (2017).

As empresas da era digital nasceram com o objetivo de reverter este cenário e transformar a TI. Para estas empresas, a TI é o motor principal (alimentado por dados) que movimenta o negócio. Para se manter competitiva, está em constante evolução e dedica grande parte dos esforços em inovação. A computação da terceira geração baseada na arquitetura em nuvem é o caminho.

O profissional da era digital deve ser capaz de inovar, colaborar com a evolução dos softwares e, consequentemente, dos negócios, aplicando mais inteligência aos processos e quando falamos em Big Data, buscando sempre o valor dos dados. Os profissionais com perfil operacional que aprenderam a resolver problemas recorrentes e “se acomodaram” a desenvolver uma carreira aplicada em uma única solução, terão cada vez menos espaço no mercado.

Um dos grandes desafios desta transição, para o modelo de digitalização das empresas é manter o legado e trabalhar com a visão de inovação, buscando-se evolução na nova geração da computação.

O Gartner usa o termo TI Bimodal para descrever esta transição, e para isso, algumas regras de evolução de roadmap devem ser analisadas, afinal existe a necessidade de manter o avião na rota enquanto trocamos a turbina e os motores. É preciso tirar dinheiro do legado e investir no novo. É preciso manter o avião no ar, mas não se limite a evitar grandes turbulências, é necessário chacoalhar as equipes, rever as culturas enraizadas, mudar pessoas e processos. Antes mesmo de mudar tecnologias. As empresas e os profissionais que entenderem esta transformação, estarão melhor posicionados para esta transição entre a segunda e terceira geração da computação.

Esta “missão impossível” é o grande desafio das empresas que precisam manter o legado, enquanto as gigantes da internet que já nasceram digitalizadas estão expandindo rapidamente, e startups do dia para a noite passam a valer milhões de dólares conquistando os seus clientes.

2.3 Big Data e IoT (Internet of Things)

No ano de 2015, o mercado estimava um número aproximado de sete bilhões de chips de celular no mundo, praticamente um por habitante. A estimativa é que até o ano 2020 este número aumentará sete vezes.

A Internet das Coisas atingiu seu hype em 2015, segundo o Gartner, e está pronta para inundar o mundo com dados. Nossas geladeiras, carros, televisores, despertadores, tênis, relógios e, assim, todas as coisas poderão ter um endereço IP e estarão produzindo dados a cada minuto.

Estamos nos primeiros passos, é uma grande oportunidade para profissionais que desejam se especializar em dados, será um mercado ainda mais abundante.

Imagine uma consulta ao médico, que poderá considerar informações coletadas pelo seu tênis ou sapato com métricas de quantos passos você se movimentou durante os últimos dias, por hora, por dia ou por minuto, seguindo uma time-line semelhante a uma rede social. Agora, considere a possibilidade de o médico correlacionar estas informações com dados coletados pela sua geladeira sobre alimentos consumidos nos últimos dias, considerando as propriedades de sódio, gordura saturada, ou qualquer outra informação relevante sobre os alimentos consumidos.

INOVANDO COM O BIG DATA

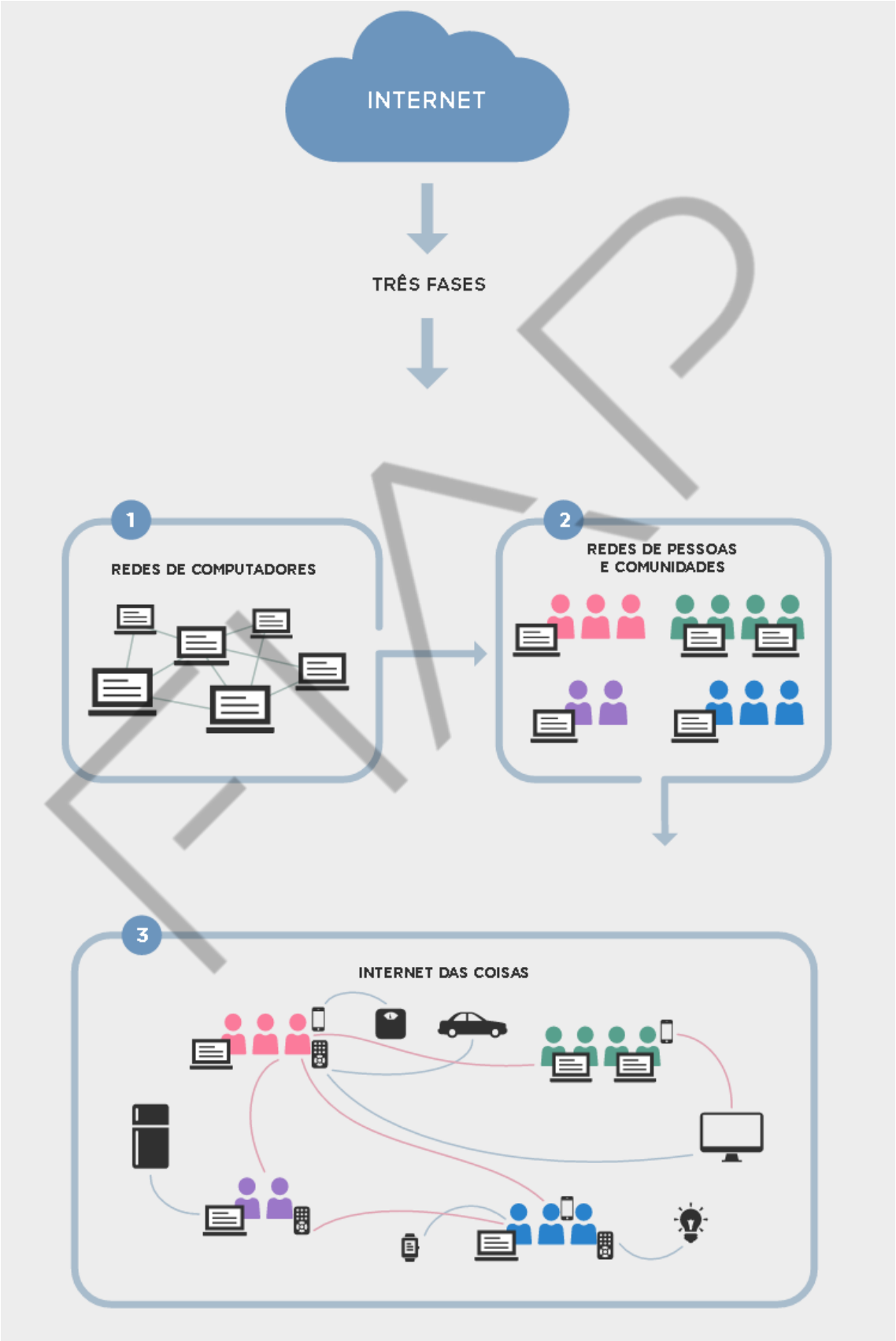

Figura 10 – Internet das Coisas
Fonte: Google Images (2016), adaptado por FIAP (2017).

O custo acessível de mapeamento do seu genoma e as bases dos sistemas de saúde mais integradas aumentam as possibilidades de uma medicina cada vez mais dependente de Big Data.

Nos últimos anos, vivenciamos problemas críticos de falta de água nas grandes cidades. Os estudos apontam que grande parte do volume de água dos reservatórios são desperdiçados antes mesmo de chegarem nas torneiras, como podemos observar no relatório da organização Trata Brasil (http://www.tratabrasil.org.br/desperdicio-na-distribuicao-de-agua), as perdas chegam a 45% em algumas regiões brasileiras. Vazamentos que permanecem durante anos pelo simples fato de não estarem visíveis, dificultando a identificação e manutenção.

E assim, vários problemas que não tratamos diretamente como problemas de dados serão considerados problemas de Big Data. Suponha que a rede de distribuição de água possua sensores a cada 500 m de distância e coletem os dados a cada minuto, com a possibilidade de identificar os desvios em tempo real e atuar com preditividade.

Outro exemplo baseado em casas inteligentes: considere que o seu despertador está integrado com a sua agenda e deverá te alertar sobre os seus compromissos, após levantar-se, um sensor identifica o movimento e ascende a luz, a cafeteira inicia o preparo do café enquanto o chuveiro aquece e a TV seleciona as principais notícias com base no seu perfil. O seu elevador irá contabilizar mais uma viagem de descida e a empresa que fornece os serviços do elevador irá computar a métrica para garantir as manutenções preventivas, precisas baseadas em números de viagens e pesos por operação, por exemplo, com informações coletadas a cada minuto, em milhares de condomínios pelo mundo, seguindo um modelo “as a service”, será monetizado conforme o uso, e isso facilitará os ajustes de manutenções.

E por fim, quando entrar no automóvel da Google, Apple, Uber ou Tesla, será conduzido ao seu compromisso guiado por software orientado por sensores e estatísticas avançadas.

Tudo isso já existe e é uma questão de tempo para a disrupção efetivamente acontecer, e com ajuda da evolução de IoT este processo poderá ser acelerado.

Em 2015, o presidente da empresa de tecnologia Cisco, John Chambers, que no início da era digital já foi uma das maiores marcas do mundo, anunciou o seu desligamento do comando da empresa após 20 anos. E uma de suas previsões, que menciona em palestras pelo mundo, ele diz: “Nos próximos 10 anos, 40% das empresas que existem hoje deixarão de existir”.

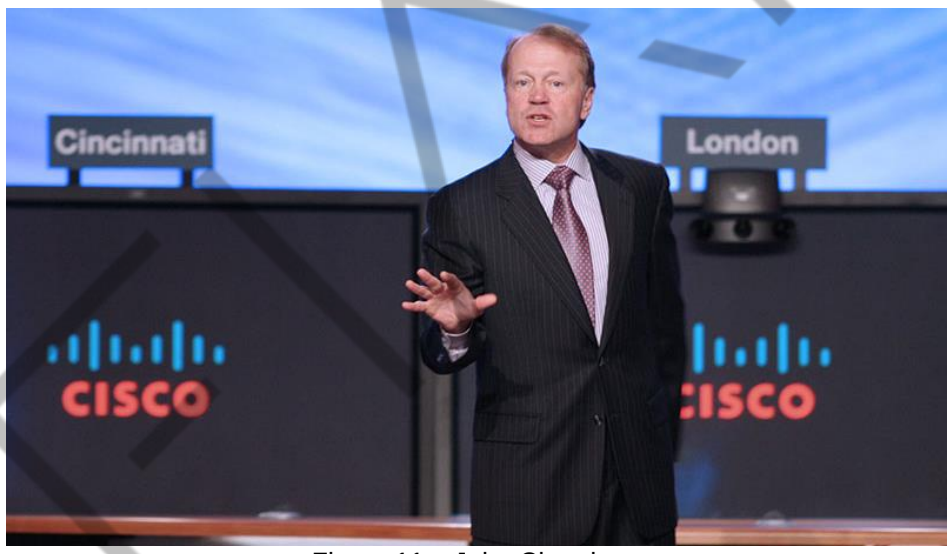

Figura 11 – John Chambers
Fonte: Google Images (2017).

Como podemos notar, o impacto da era digital é devastador, impulsionado pelo avanço da internet das coisas.

REFERÊNCIAS

CHRISTENSEN, Clayton M. **The Innovator's Dilemma.** Nova York: Harper Business Essencials, 2000.

SATHI, Dr. Arvind **Big Data Analytics**. IBM Corporation. MC Press Online, 2012. Disponível em: <ftp://public.dhe.ibm.com/software/pdf/at/SWP10/Big_Data_ Analytics.pdf>. Acesso em: 8 dez. 2015.

TAURION, Cesar. **Big Data.** São Paulo: Brasport, 2013.

GLOSSÁRIO

Grid	É um modulo computacional para distribuir tarefas entre diversos computadores.
Arquitetura *Standalone*	É uma arquitetura capaz de executar seus processamentos isoladamente.
Computação Distribuída:	É permitir que diversos computadores possam, interligados em rede, compartilhar a execução de tarefas de um ou mais sistemas, ou seja, possibilitar “sistemas distribuídos”.
SAS (Serial Attached SCSI)	É um protocolo serial (ponto a ponto) que transfere dados (de e para) dispositivos de armazenamento de computador (Wikipédia).
SATA (Serial AT Attachment)	É uma tecnologia para transferência de dados entre computadores e dispositivos de armazenamento em massa com unidades em disco (Wikipédia).
SSD (Solid State Drive)	Dispositivos de armazenamento de rápido acesso do qual não possui partes móveis. É um evolução do HD.
Storage	São dispositivos projetados especificamente para armazenamento de dados.
Storage Area Network	É uma rede responsável por conectar servidores e dispositivos de armazenamento.
Scale-out	Significa o poder de aumentar (escalar) a quantidade de computadores para um rápido crescimento, que pode ser chamado de “escalabilidade horizontal”.