FIAP ON
ENSINO A DISTÂNCIA

BIG DATA & ANALYTICS

ECOSSISTEMA E **PROFISSÕES EM BIG DATA**

Anderson Paulucci e Leandro Rubim

CAPÍTULO 5

LISTA DE FIGURAS

LISTA DE QUADROS

SUMÁRIO

5 ECOSSISTEMA E PROFISSÕES EM BIG DATA

5.1 Plataformas e Soluções de Dados

5.1.1 Data Lake - Uma nova abordagem para o DW

Ralph Kimball – um dos precursores dos conceitos de Data Warehouse e da técnica de modelagem dimensional – menciona na terceira edição do famoso *best-seller* "*The Data Warehouse Toolkit*":

Figura 5.1 – Ralph Kimball
Fonte: Banco de Imagens Shutterstock

- Estes termos recorrentes (se referindo a grande parte dos conceitos de DW/BI) já existem há mais de três décadas.
- Recolhemos toneladas de dados, mas não podemos acessá-los, precisamos "slice e dice" os dados em todas as direções (quebrar um conjunto de informações em partes menores permitindo examiná-lo a partir de diferentes pontos de vista para que possamos compreendê-lo melhor).
- Pessoas de negócios precisam obter os dados facilmente.
- Mostrar apenas o que é importante.

- Desperdiçamos reuniões inteiras discutindo sobre quem tem os números corretos ao invés de tomar decisões.
- Queremos usar a informação para apoiar a tomada de decisão baseada em fatos.

Os primeiros estudos sobre a abordagem Data Warehouse surgiram na década de 1980 como conceitos acadêmicos. Naturalmente, a sistematização das empresas com ERP, CRM e os demais sistemas transacionais ajudaram a criar as fontes de dados corporativas que alavancaram a necessidade da implementação de uma base de dados também conhecida como armazém de dados (o DW), com o objetivo de consolidar uma visão estratégica dos dados, organizados na linha do tempo de acordo com as necessidades de negócio.

Consequentemente, os dados evoluíram, as fontes de dados da era digital são bastante expansivas e possuem uma variedade de dados muito maior.

Não proponho discutirmos a importância e a necessidade de trabalharmos com DW, pois isso é evidente. Big Data não anula o Data Warehouse que foi projetado para atender soluções de Business Intelligence, o mercado não terá uma resposta rápida e definitiva para qualquer tipo de comparações neste sentido. Mas realmente precisamos considerar o apoio das plataformas de Big Data para compor uma arquitetura de Data Warehouse tradicional, portanto, podemos afirmar que os domínios são complementares.

A arquitetura do Data Warehouse não está preparada para atender ao Big Data, por várias razões:

- Toda e qualquer nova fonte de dados deverá passar por um processo burocrático de estruturação (modelagem, ETL, ajustes etc.) e quanto maior o volume e abrangência dos dados, maior será o tempo necessário para armazenamento e processamento.
- As tecnologias de armazenamento e processamento tradicionais não atendem as capacidades das demandas de negócio. E os 3Vs de Big Data excedem seus limites facilmente.

- O custo de soluções tradicionais de DW, considerando licenciamento de software, armazenamento, processamento e segurança, dificultam e até inviabilizam os projetos de evoluções.

James Dixon (2015), Diretor de Tecnologia na Pentaho mencionou em seu blog “Union of the State – A Data Lake Use Case” que para facilitar o pensamento sobre Data Mart, basta pensar como uma reserva de água engarrafadas, higienizadas, embaladas e estruturada para fácil consumo, fazendo uma analogia com o objetivo de preparar os dados para o fácil consumo.

Figura 5.2 – Data Marts
Fonte: Banco de Imagens Shutterstock

Os Data Marts são parte de um domínio de dados, orientados aos departamentos ou assuntos da empresa por exemplo.

Bill Inmon é considerado o pai do Data Warehouse e assim como Kimball possui dezenas de livros sobre o assunto. Um ponto de divergência curioso e conhecido, entre os dois, está relacionado aos Data Marts.

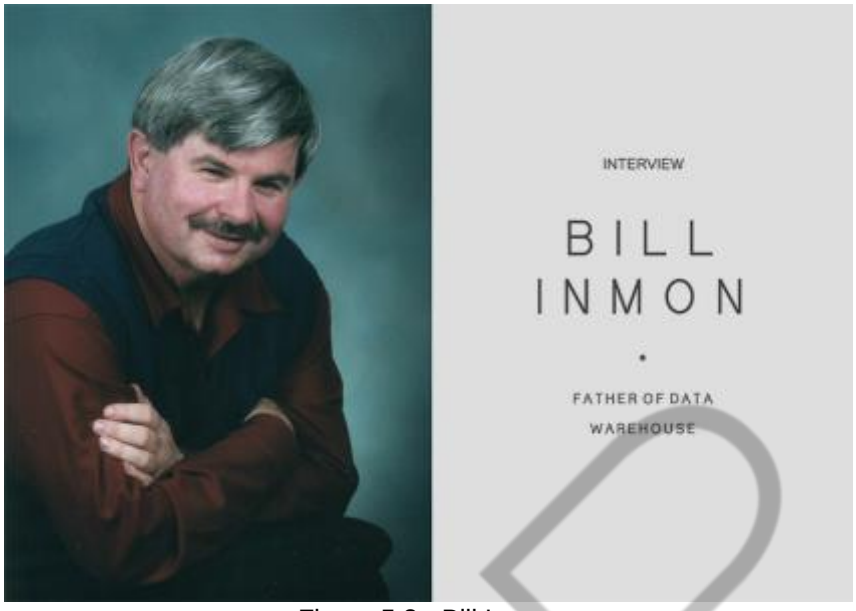

Figura 5.3 - Bill Inmon
Fonte: Banco de imagens Shutterstock (2017)

Kimball & Ross (2013) defende iniciar a construção dos Data Marts e integrá-los posteriormente, definindo o conceito que apelidou de Data Warehouse Bus Architecture.

Na avaliação de Inmon (1995) é recomendável criar um Data Warehouse com um único modelo de dados corporativo e posteriormente derivar, construindo os Data Marts por assuntos ou departamentos, propondo o conceito do CIF – *Corporate Information Factory.*

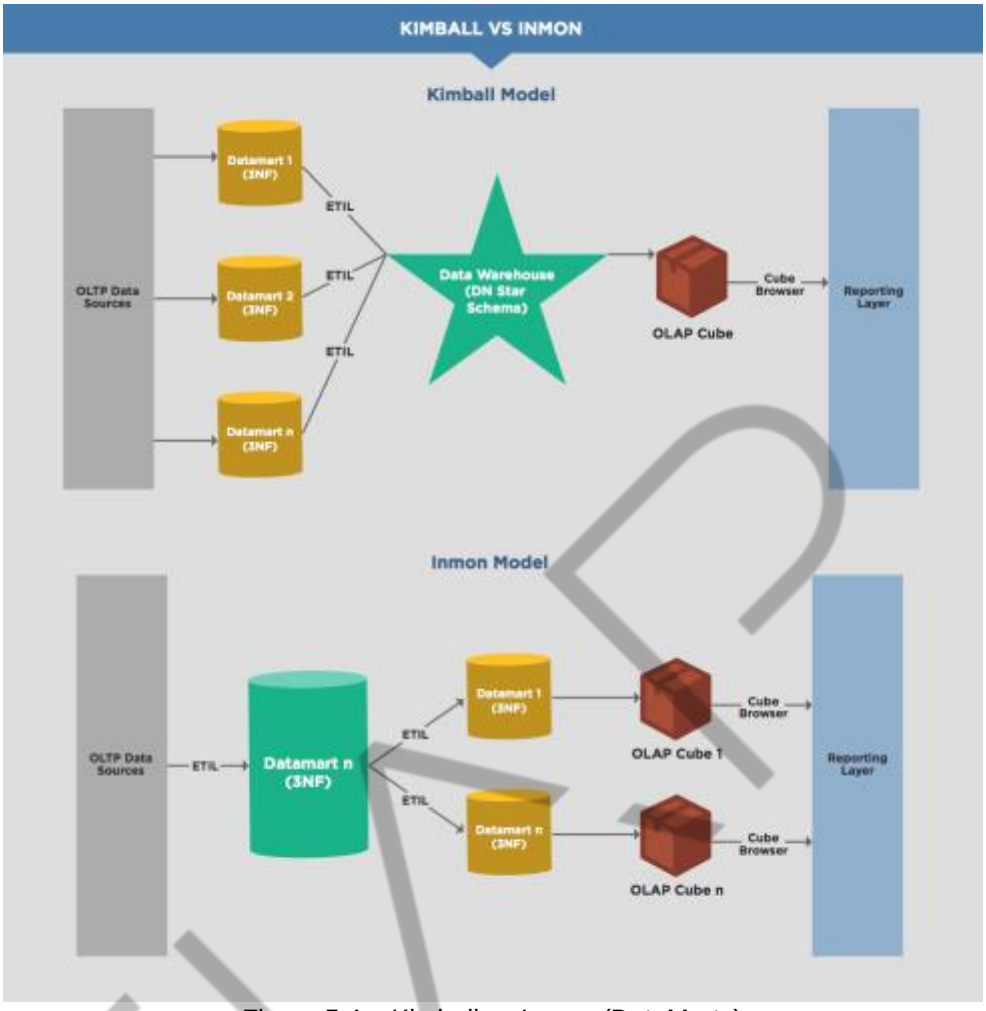

Figura 5.4 – Kimball vs Inmon (DataMarts)
Fonte: https://bennyaustin.wordpress.com/2010/05/02/kimball-and-inmon-dw-models/, adaptado por FIAP (2017)

Independentemente da opção sobre as melhores práticas de implementação dos Data Marts, James Dixon propôs em Dixon (2017) o termo Data Lake para contrastar a abordagem dos Data Marts. De acordo com o autor, existem duas grandes limitações na abordagem de Data Marts:

- Apenas um subconjunto dos atributos é analisado, para que apenas perguntas predeterminadas possam ser respondidas.
- Os dados são agregados e perdemos visibilidade para os níveis mais baixos.

Estes problemas são muitas vezes definidos como silos de informações e acabam criando dificuldades para a evolução de modelos analíticos.

A EMC está criando um “Data Lake” em conjunto com a White House Climate Data Initiative, com o objetivo de compreender os dados climáticos e suas implicações quanto ao aquecimento da Terra entre outros possíveis fatores. Para este “Data Lake” a premissa é que quanto maior o número de pessoas adicionando mais dados, maior será o impacto.

Foi realizado um estudo inicial com a seguinte questão: para onde foram 1 milhão de Andorinhas e qual é o motivo?

Figura 5.5 – Para onde foram as andorinhas?
Fonte: Banco de imagens Shutterstock (2017)

Para entender como as mudanças climáticas afetam até mesmo uma ave durante o ano, comece multiplicando os milhões de aves por milhares de locais, todos os dias, durante 365 dias.

Essa visualização de big data revelou tudo isso em apenas alguns segundos e pode fazer mais, muito mais. Os cientistas conseguem ver rapidamente as mudanças na migração, observando os dados de vários anos. A migração está aumentando? E a temperatura? E as chuvas? Há correlações que apontam para as mudanças climáticas?

Data Lake é um repositório de armazenamento e *engine* para processamento de Big Data. Fornece armazenamento massivo para qualquer tipo de dados, tem enorme poder de processamento e capacidade de lidar com tarefas e jobs simultâneos, praticamente ilimitadas.

Um conceito chamado Data Lake e não uma tecnologia, isso significa que pode ir além da solução *Hadoop*. Porém o Ecossistema Hadoop é a tecnologia que melhor atende as necessidades definidas pelo conceito do Data Lake, seu custo-benefício acaba sendo decisivo para que as empresas iniciem uma prova de valor (PoV) e evoluam com a implementação.

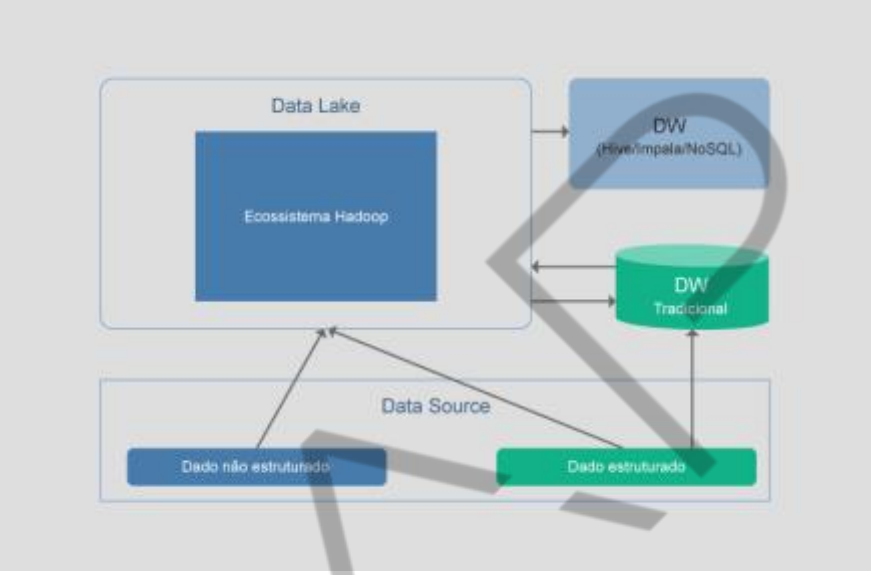

Figura 5.6 – Data Lake
Fonte: do Autor (2015), adaptado por FIAP (2017)

Esta arquitetura possibilita manter um grande repositório de dados “brutos”, preservando o princípio de imutabilidade, garantindo maior capacidade de retenção de dados com custo consideravelmente reduzido.

Um modelo tradicional de ETL (*Extract/Transform/Load*) ou Extração/ Transformação/Carga, exige um esforço grande para modelagem e desenvolvimento de rotinas para preparar os dados. Aplicando a abordagem de Data Lake, o processo de carga poderá ser priorizado, com a possibilidade de abrir mão da estruturação dos dados inicialmente, assim podemos iniciar o processo de ingestão (carga) de dados no Data Lake e já submetermos os dados aos processos de análises.

Todavia, nem todos os dados devem ser tratados de forma bruta, ainda continuaremos dependendo de processos pesados de transformação dos dados para grande parte dos dados, envolvendo técnicas de estruturação, enriquecimento e qualidade dos dados. Fazendo uso da grande capacidade do cluster de

armazenamento, podemos otimizar o processo com um fluxo invertido, conhecido como ELT.

Aplicando o ELT, é possível usar a capacidade do *cluster* Hadoop para executar processamentos massivos, aumentando a velocidade e agilidade de todo o processo.

Mantendo o princípio de imutabilidade (não alterando os dados no Data Lake) será possível corrigir erros de regras de negócios e falhas de cadeias de transformação sem a necessidade de grandes manobras e esforços adicionais para movimentações de dados.

O Data Lake armazena um grande volume de dados independentemente do seu formato ou estrutura, e não define um esquema, mantendo o dado no seu formato nativo até que os dados sejam efetivamente demandados.

O modelo propõe despejar todos os dados no lago, seguindo a abordagem "*free-for-all*". Com o tempo, não será tão simples navegar por grandes lagos escuros com pouca visibilidade, passando por brejos improdutivos que limitaria o potencial dos dados.

Figura 5.7 – Analogia a navegar por grandes lagos escuros
Fonte: Banco de imagens Shutterstock (2017)

Precisamos considerar técnicas importantes como a curadoria, para manter as águas claras e passíveis de navegação, afinal nosso maior objetivo é extrair o valor dos dados com agilidade.

A **Curadoria** dos dados envolve a captura de metadados e gestão da linhagem dos dados, incorporando as informações no catálogo de metadados. Está intimamente ligado à governança dos dados e será importante para sustentar os processos de análise, estruturação, limpeza e transformação dos dados.

Com toda a confusão e volume de informações que estamos produzindo, é cada vez mais necessária a adoção da curadoria como premissa para um projeto de Big Data, fator decisivo para a sustentação do Data Lake.

Um exemplo de curadoria usado em redes sociais é o *Moments* do Twitter (https://twitter.com/i/moments).

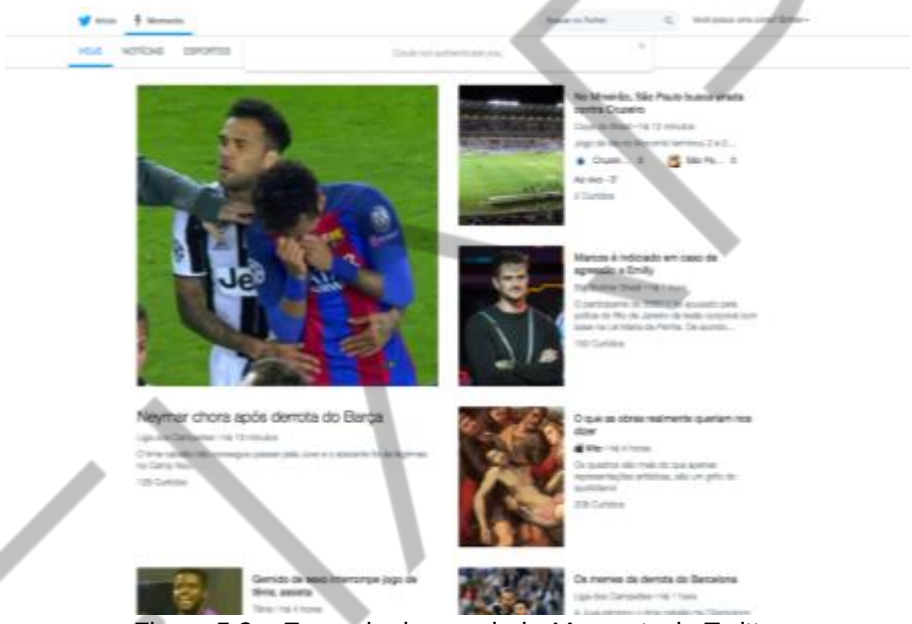

Figura 5.8 – Exemplo de curadoria *Moments* do Twitter
Fonte: Twitter

Com um grande volume e *tweets* gerados a cada segundo *versus* vários assuntos, não é simples construirmos um roteiro sobre os fatos e qualificarmos os conteúdos com base em colaborações. Afinal, não temos possibilidade de seguir todas as pessoas e acompanhar um grande número de publicações a todo o momento.

Figura 5.9 – Usuário twittando
Fonte: Banco de imagens Shutterstock (2017)

O Twitter aposta em parcerias com empresas especializadas em conteúdo para fazer o trabalho de curadoria e apresentar aos usuários os *moments* (ou histórias) que serão complementadas com os *tweets* de usuários que representem relevância para os assuntos discutidos.

5.1.2 Arquitetura LAMBDA

Processamento de grandes volumes e análises com baixíssima latência (*real time*), são grandes desafios inclusive para a arquitetura de Big Data.

Nathan Marz propôs o termo Arquitetura Lambda (LA) para o *design* genérico de processamento de dados escalável e tolerante a falhas, após experiências com processamentos distribuídos na Backtype e Twitter.

A arquitetura Lambda atende as características de Big Data e Fast Data, sendo uma base de referência para implementação dos mais variados *use cases.*

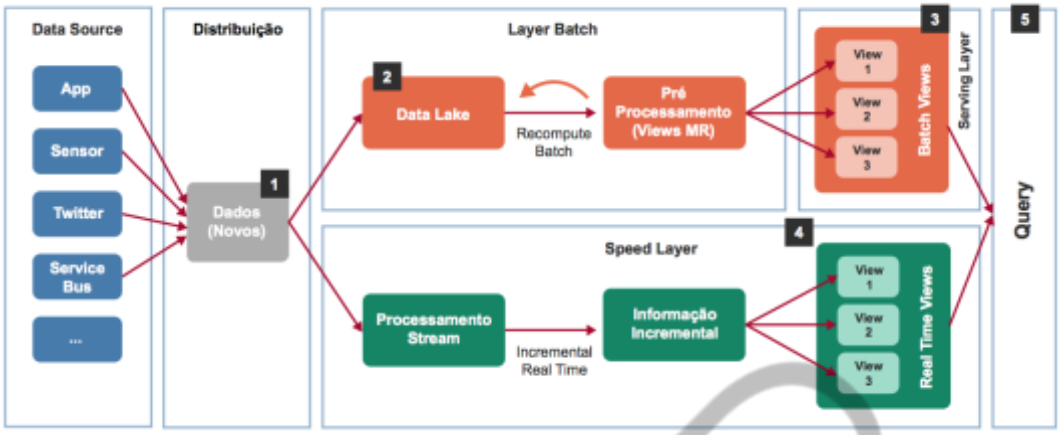

Figura 5.10 – Arquitetura Lambda
Fonte: Nathan Marz (2015), adaptado por FIAP

1) Todos os dados que entram no sistema são enviados tanto para a camada de processo *batch* e *real time*.

2) A camada batch tem duas principais funções:

    a. Administrar o conjunto de dados *master* (Data Lake), considerando o princípio da imutabilidade dos dados, ou seja, preservar os dados de entrada sem alterações para que eles possam ser recomputados e gerar visões pontuais (particionadas) a qualquer momento, evitando a necessidade de refazer uma cadeia completa de processamentos.

    b. Pré-calcular (processar) pontos de visões *batch* para servir a camada de acesso de consultas e análises.

3) A camada *serving* indexa os pontos de visões *batch* para que possam ser consultados com baixa latência, por exemplo: exploração *ad-hoc*.

4) A camada *speed* compensa a alta latência de atualizações da camada *serving* e lida apenas com os dados mais recentes (o processo de preparação dos dados *batch* pode levar minutos ou horas, e apesar de ser muito mais ágil, comparado com os fluxos do BI (D-1) tradicional, ainda assim pode representar limitações quando o objetivo é trabalhar com Fast Data, que demanda segundos e microssegundos de latência em toda a cadeia do processo de análise).

5) Todas as consultas podem ser respondidas através da junção das visões *batch* e *real time*.

A arquitetura Lambda possibilita entendermos a abrangência e complexidade de soluções necessárias para a implementação de um projeto de Big Data. São necessários vários componentes para atender os requisitos de dos 3Vs de Big Data.

5.1.3 Armazenamento e Processamento Analítico

As tecnologias de Big Data estão evoluindo rapidamente e o *roadmap* de um ano pode representar uma grande distância entre as soluções. O termo Big Data foi adotado efetivamente pelo mercado em 2012, e em 2015, o mercado corporativo iniciou discussões sobre a segunda geração de Big Data, baseada em Fast Data.

O desafio da arquitetura de TI para acompanhar esta evolução exige uma atenção para a adoção de cada componente. Uma engrenagem errada pode travar a evolução da plataforma. Às vezes, a definição da solução deve aguardar o momento certo, afinal, muitas tecnologias ainda não foram efetivamente experimentadas para que possamos obter confiança na sua adoção.

O Ecossistema Hadoop é, sem dúvidas, a principal plataforma para a implementação de um projeto de Big Data, podendo ser considerado também um *framework*.

Fazendo uma analogia com o *framework* PMBOK para gerenciamento de projetos, no qual temos 47 processos (PMBOK 5ª edição) que poderão ajudar com gerenciamento de escopo, riscos, qualidade, prazos etc., você não precisa aplicar todos os processos para gerenciar um projeto, mas definitivamente é o *framework* mais completo.

O Ecossistema Hadoop é composto da integração de vários componentes com o objetivo de atender praticamente qualquer cenário de aplicação de Big Data. Ele é baseado em *software open source* e escalabilidade horizontal, projetado para atender Peta Bytes de dados e processar esses grandes volumes com paralelismo.

O Hadoop foi criado em 2006 e é amplamente adotado pelas empresas digitais. A necessidade de evolução com mais *performance* nos levou à segunda

geração de Big Data (Fast Data) e novas tecnologias estão surgindo com características *in-memory,* ou seja, baixar a latência para atender análises *near real-time.*

Uma nova plataforma eminente é o Spark, criada em 2009, que vem ganhando força e promete ser a solução *core* para o Fast Data. Trata-se de projeto *open source* criado na Universidade de Berkeley, disponibilizado para a Fundação Apache. Com o objetivo de evoluir com a comunidade, rapidamente foi acoplado ao Hadoop, sendo complementar neste caminho evolutivo.

Uma solução de Big Data possui vários componentes integrados, criando um ecossistema conforme abaixo.

Obs.: o diagrama abaixo não resume todas as soluções de Big Data, apenas ilustra a complexidade e abrangência da arquitetura analítica.

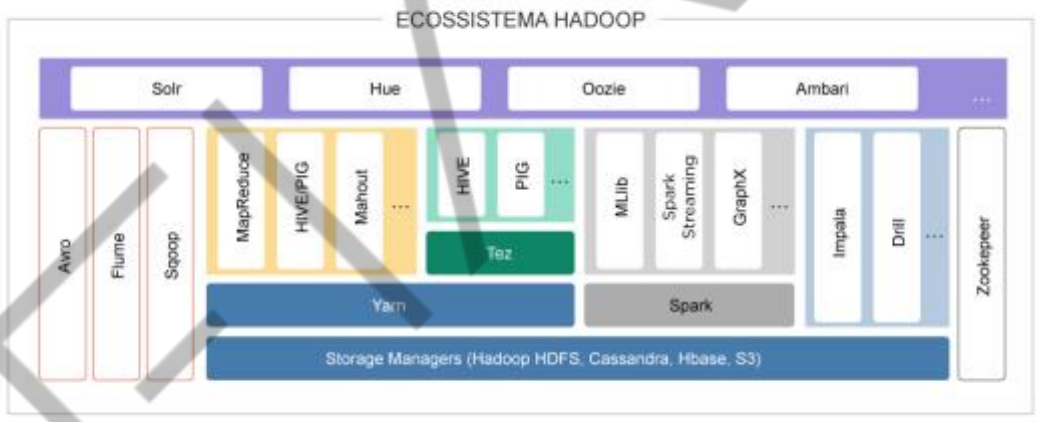

Figura 5.11 – Ecossistema Hadoop
Fonte: Elaborado pelo Autor (2015), adaptado por FIAP (2017)

Todos os componentes são *open sources,* e a cada dia, a comunidade se depara com grandes desafios para evoluir o *roadmap* de maneira integrada e consistente.

Assim como o Spark, na segunda geração de Big Data surgiu mais uma grande promessa, criada em Berlim na Alemanha, a solução Apache Flink é completa, compreende processamento *batch* e *streaming* e pode se integrar a todo o ecossistema.

A adoção de soluções *open-source* ainda é um tabu para a maioria das empresas, devemos entender que realmente existe um risco associado a qualquer dependência de uma solução de código aberto. Portanto, as empresas devem mitigar questões importantes como:

Figura 5.12 – Questões importantes na adoção de soluções *open-source*.
Fonte: Elaborado pelo Autor (2017), adaptado por FIAP (2017)

Uma empresa não deve correr o risco de depender da comunidade para aguardar a solução de correção de um *bug*, por exemplo, por mais inovadora que ela seja os riscos podem ser consideráveis.

Aprendemos a usar *softwares* abertos com o avanço do sistema operacional Linux. Empresas que fazem a intermediação do *software open-source* para o cenário corporativo, foram criadas para garantir a segurança e confiança na adoção da solução. Um bom exemplo disso é a RedHat que se tornou referência na distribuição do sistema operacional Linux. Um *software* de código aberto não pode ser comercializado com a venda de licenças, o modelo é baseado em subscrição e propõe um contrato de manutenção e suporte do produto.

A Cloudera é a maior empresa criada para distribuição do Hadoop, cujo CTO é ninguém menos que o pai do Hadoop, Doug Cutting. Temos opções e cada uma com seus diferenciais:

- Hortonworks.
- MapR.
- IBM Big Insights for Apache Hadoop.
- MS HD Insight.
- AWS EMR.

As empresas precisam decidir sobre a implementação de uma solução *on-premises* (local) ou optar por um serviço em nuvem.

Provedores de serviços em nuvem como Google, Amazon e Microsoft desenvolveram um nível alto de maturidade em suas soluções de Big Data, e podem prover tecnologias como um serviço, garantindo a abstração de toda a complexidade para que os clientes possam focar em evoluir os algoritmos, afinal, eles já são experientes e estão direcionando a evolução das tecnologias.

5.1.4 *Softwares* para *Analytics*

Após a primeira onda de Big Data, praticamente todos os *softwares* analíticos tradicionais (Tableau, SAS, Microstrategy etc.) foram adaptados para integração com as plataformas de Big Data, usando conectores customizados ou genéricos. E já podem fazer uso de toda a robustez das plataformas de maneira relativamente simples, limitada para análises de *front-end*, que requerem uma estrutura bem definida.

Vamos destacar um *software* analítico que está evoluindo na velocidade de Big Data: o R.

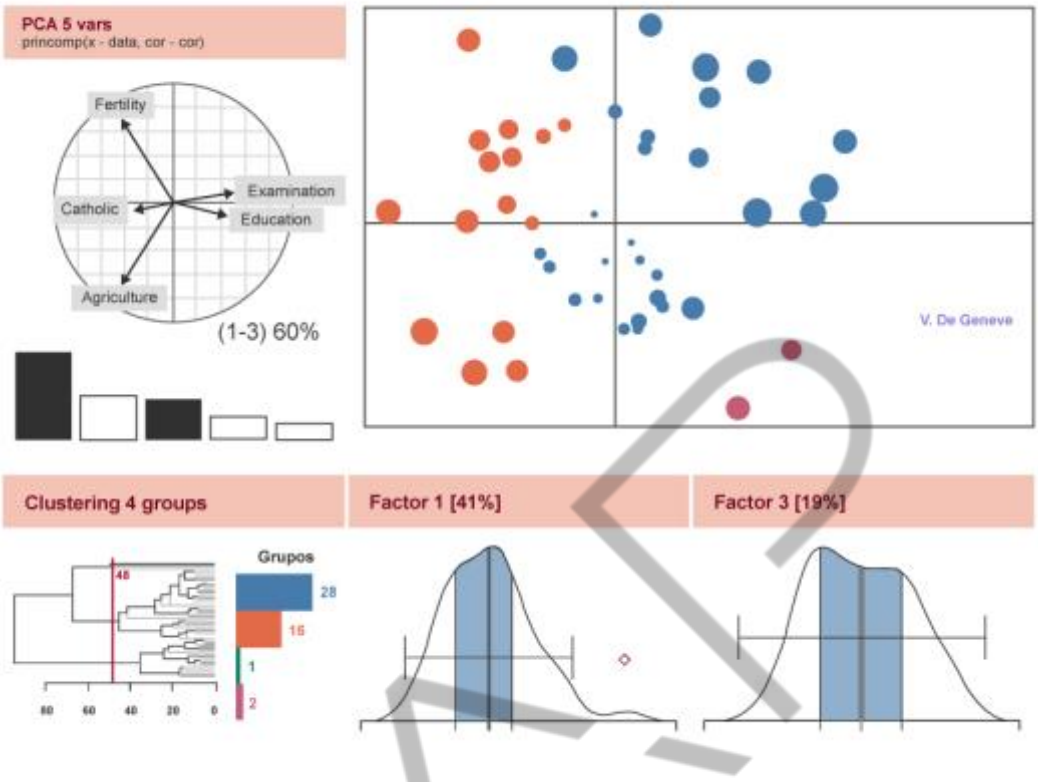

Figura 5.13 – Linguagem R
Fonte: Robert Gentleman e Ross Ihaka (2001), adaptado por FIAP (2017)

- R é um *software* de análise de dados.
- R é uma linguagem de programação, você faz a análise de dados em R gravando *scripts* e funções na linguagem de programação R.
- R é uma linguagem orientada a objeto completa e interativa, projetada por estatísticos para estatísticos.
- R é um ambiente para análise estatística, a maioria das pesquisas de ponta em estatísticas e modelagem preditiva é desenvolvida em R.
- R é um projeto de *software open-source*, seu potencial é estendido por meio de pacotes criados por usuários, que permitem técnicas estatísticas especializadas, dispositivos gráficos, capacidades de importação / exportação, ferramentas de relatórios etc.

5.1.5 Armazenamento e processamento operacional

Apesar da grande ênfase de Big Data para as soluções analíticas, é importante ressaltar que o conceito Big Data também deve ser aplicado para atender as capacidades operacionais. Empresas estão operando modelos de negócios que demandam arquiteturas de terceira geração, robustas, geograficamente distribuídas e assim como evoluíram as grandes plataformas analíticas, notamos um avanço das aplicações operacionais em busca de maior volume, velocidade e variedade.

Como vimos nos capítulos anteriores, a arquitetura cliente-servidor limita soluções tradicionais como o SGDBR (sistema gerenciador de banco de dados relacional) para atender as necessidades de escalabilidade e novas semânticas de dados.

As gigantes da internet (Amazon, Google, Facebook) se depararam com limitações operacionais e idealizaram novos modelos de persistências para resolverem os problemas de armazenamento e processamento distribuído, que demandam arquiteturas com alta escalabilidade.

O termo *databases* NoSQL (Not Only SQL), por exemplo, foi utilizado inicialmente em 2009 para nomear um evento sobre *databases* não relacionais.

Soluções alternativas para o modelo de persistência tradicional (SQL) começaram a ganhar força nesta era da informação, e as empresas digitais alavancaram grandes *cases* com a adoção destas soluções. É o caso da Netflix, que possui um *cluster* de *database* Cassandra com milhares de instâncias, suportando milhões de usuários pelo mundo.

Figura 5.14 – Cluster de database Cassandra do Netflix.
Fonte: Techblog – Netflix (2017)

Conceitualmente, os *databases* NoSQL são baseados em estruturas de chave-valor e podem ser organizados em diferentes modelos.

Figura 5.15 – Modelos NoSQL
Fonte: do Autor (2015), adaptado por FIAP (2017)

Os *databases* NoSQL possuem o DNA da terceira geração da computação (*cloud*):

- Arquitetura *scale-out* (escalabilidade horizontal).
- Escalabilidade linear.
- Capacidade de aproximar a computação dos dados.
- *Software open-source.*
- *Hardware commodities.*
- Arquitetura geograficamente distribuída.
- Suporte para dados não estruturados.

5.2 Profissionais para Big Data

A democratização da tecnologia com adoção de soluções mais abertas e abrangentes demandam capacitação constante. A velocidade com que as tecnologias evoluem está impactando e causando um *gap* enorme de mão de obra qualificada na TI.

A computação em nuvem ajudará a TI das empresas a saírem do caos operacional e manter um foco maior no negócio. Este é um caminho importante para que os profissionais possam dedicar mais esforços no que realmente agrega valor para a empresa.

Figura 5.16 – Computação em nuvem
Fonte: Banco de imagens Shutterstock (2017)

As empresas precisam se preparar para uma estrutura organizacional que priorize tomadas de decisões baseadas em dados. Afinal, quem é o responsável por investir e manter uma estratégia para os dados? No modelo tradicional, o CIO (*Chief Information Officer*) está sufocado com tantos problemas e necessidades de investimentos para manter a TI, e faltará folego para essa transformação, portanto, deve-se abrir uma nova cadeira executiva para o CDO (*Chief Data Officer*). E os profissionais de TI devem ser preparados com uma nova visão, um novo *mindset* para a era digital, o DNA de inovação deverá prevalecer e funções importantes como o cientista de dados são fundamentais para compor esta nova estrutura.

Figura 5.17 – Analogia a cientista de dados
Fonte: Banco de imagens Shutterstock (2017)

As estruturas organizacionais dominantes nas empresas digitais são baseadas em modelos exponenciais. Segundo Ismail et al. (2014) no livro "*Exponential Organizations*", um dos fatores de sucesso das empresas da Era Digital é manter uma equipe reduzida de funcionários mesmo com crescimentos exponenciais dos negócios. E este "milagre" da gestão só é possível com a ajuda de tecnologias avançadas, principalmente baseadas em Big Data. É preciso reduzir a burocracia e tomar decisões guiadas por dados.

Em 2008, DJ Patil e Jeff Hammerbacher usaram o termo "*Data Scientist*" (Cientista de Dados) para definir suas funções no LinkedIn e Facebook, respectivamente.

Abaixo, a Figura 5.18 mostra um diagrama que representa uma estrutura organizacional genérica e prioriza estratégia de dados e no Quadro 5.1 os cargos em BigData.

Figura 5.18 – Estrutura Organizacional Genérica
Fonte: Elaborado pelo Autor (2015), adaptado por FIAP (2017)

Quadro 5.1 – Cargos em BigData (continua)

CDO	A TI está sufocada com as demandas de infraestrutura, desenvolvimento e projetos que consomem seus valiosos recursos de tempo e dinheiro. Isso impossibilita o investimento adequado nos dados. Tratar dados como ativo requer um foco maior e uma estratégia bem definida. A função do CDO (*Chief Data Officer*) é desenvolver o planejamento adequado da estratégia de dados, alinhado aos propósitos corporativos e garantir a visão executiva necessária para engajar as iniciativas de toda a empresa.
Cientista de Dados	"A profissão mais sexy do século XXI" esta frase foi usada por Hal Varian, economista do Google (2012) para definir o potencial deste profissional. Todas as empresas estão atrás dos cientistas de dados. São especialistas analíticos que possuem habilidades para resolver problemas de alta complexidade, alguns *skills* são essenciais para este perfil profissional. Quantitativo: deve ter uma base sólida de conhecimentos em matemática e estatística. Desenvolvimento: necessário conhecimento em ciências da computação (no mínimo programação) para desenvolver algoritmos e propor soluções para desbloquear o valor dos dados. As linguagens R e Python, por exemplo, são ferramentas indispensáveis para os cientistas de dados.
Governança de Dados	Os profissionais responsáveis em garantir a governança dos dados estão contidos na estrutura da TI atual e serão peças-chave para a estratégia de Big Data nas empresas. Abaixo, algumas funções para garantir melhor governança: Administrador de dados: seu trabalho se torna mais abrangente e vai além do *database* relacional. Com ajuda do desenvolvedor e do cientista de dados, terá um grande desafio no controle dos metadados e estruturas para as novas semânticas; Curador de dados: Responsável em manter os índices de qualidade, veracidade e confiabilidade dos dados aceitáveis. Trabalhando com dados não estruturados e um *time-to-market* agressivo será seu grande desafio.

Continuação do Quadro 5.1

CIO	O CIO já tem muitos problemas com a TI tradicional e não disporá de tempo e orçamento para investir e priorizar a estratégia e evolução de Big Data. Porém, para muitas empresas, não será possível iniciar uma mudança organizacional com mais um cargo executivo (CDO), e o CIO deverá assumir a função de preparar a empresa para este novo modelo de gestão baseada em *Data Driven*. Isso possibilitará a aproximação do CIO da estratégia de negócios e a TI passará a assumir um papel importante como agente de inovação da empresa.
Arquitetura de TI	Os arquitetos de tecnologia/infraestrutura, dados e aplicações responsáveis pelo *roadmap* da TI, serão cada vez mais exigidos. A arquitetura de soluções para suportar um projeto de Big Data envolve muitos componentes e a tecnologia avança em ritmo acelerado. A estratégia dos dados corporativos deverá ser definida com apoio do time de Arquitetura de TI, alinhado com os casos de usos das empresas.
Desenvolvimento	Big Data demanda muito desenvolvimento e, normalmente, este profissional também pode ser enquadrado como engenheiro de dados. Técnicas de computação com grandes volumes são implementadas com o uso de frameworks de processamento paralelo massivo, como o *Map Reduce*, e exigem conhecimentos de métodos de otimização e distribuição dos dados. Os ecossistemas mais comuns de Big Data são implementados em Java e esta, sem dúvida, é a linguagem-padrão, porém linguagens como Python e Scala estão em ascensão com a evolução da chamada segunda geração de Big Data. Os desenvolvedores de Big Data atuam desde o processo de carga com a ingestão de dados nos clusters, transformação, enriquecimento, qualidade e análises. Projetos e Big Data demandam fabricas de *softwares* com muitos profissionais de desenvolvimento e, geralmente é o maior número de pessoas para implementação e evolução de Big Data.

Continuação do Quadro 5.1

Administração de TI	Plataformas de Big Data não são de simples administração, e requerem habilidades que serão incorporadas por profissionais da TI: DBA: Com o grande conhecimento em administração de banco de dados, estes profissionais são importantes para a evolução de Big Data, pois conhecem muito de dados e soluções de gerenciamento que independentemente da estrutura dos dados e arquitetura serão base para as soluções de Big Data. Infraestrutura: Soluções de Big Data demandam grandes *clusters* (plataformas) de TI, o gerenciamento de sistema operacional, *network*, capacidade de armazenamento/ processamento e são responsabilidades dos especialistas de infraestrutura que devem se preparar para suportar este grande “elefante”.
BI/Analytics	Os profissionais de BI tradicional são peças fundamentais para implementar um projeto de Big Data, afinal, eles possuem um histórico de como os dados são coletados, quais são as regras de transformação e necessidades de análises que não são efetivamente atendidas no modelo atual. Estes profissionais serão beneficiados com a adoção da nova plataforma e devem participar ativamente na construção de Big Data. É importante ressaltar a necessidade da mudança de *mindset* para os especialistas em BI, muitos projetos vão “por água a baixo”, quando tentamos usar modelos e processos tradicionais de BI em uma plataforma de Big Data.
Estatísticos/Matemáticos	Poucas empresas demandam mão de obra qualificada em estatística e matemática. Raramente uma empresa abre uma oportunidade com o requisito de formação “mestrado em matemática”. Durante muitos anos estes profissionais permaneceram mais próximo do lado acadêmico. Empresas financeiras contratam matemáticos e estatísticos com maior frequência, para desenvolver justamente uma análise mais precisa de mercado. Com a possibilidade de armazenar grandes volumes de dados, as empresas irão demandar profissionais quantitativos para apoiar os especialistas de negócios em análises complexas. Estes profissionais são grandes candidatos a evoluírem a carreira como cientista de dados, para isso devem aprender a programar em linguagens como Python, R, Scala e Java.

Conclusão do Quadro 5.1

Especialistas de Negócios	Profissionais que atuam nas áreas de negócio da empresa (ex.: Marketing, Vendas, Gestão de Riscos etc.). À medida que as empresas ganham maturidade nos processos guiados por dados, muitos profissionais de negócio, principalmente funções operacionais, podem perder espaço. Imagine um sistema de recomendações baseado em aprendizado de máquina (*machine learning*), que alavanca vendas com abordagem precisa de marketing, eliminando custos com campanhas e processos burocráticos de CRM convencionais. Os especialistas de negócios deverão ser os maiores aliados da TI, na identificação dos *use cases* para a evolução dos projetos de Big Data. Afinal, ninguém conhece melhor as oportunidade e necessidades da empresa que eles.

Fonte: Elaborado pelo Autor (2015).

REFERÊNCIAS BIBLIOGRÁFICAS

BARTON, Dominic; COURT, David. **Making Advanced Analytics Work for You.** 2012. Disponível em: <https://hbr.org/2012/10/making-advanced-analytics-work-for-you>. Acesso em: 2 dez. 2015.

BASSO, Guilherme Mastrichi. **Terceirização e o mundo globalizado:** o encadeamento produtivo e a complementaridade de serviços como potencializadores da formação de contratos. s/d. Disponível em: https://juslaboris.tst.jus.br/handle/1939/5388. Acesso em: 9 out. 2014.

DAVENPORT, Thomas H; PATIL, D.J. **Data Scientist: The sexiest job of the 21st century**. Harvard Business Review. Publicado em: Out. 2012 Disponível em: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century. Acesso em: 8 dez. 2015.

DIXON, James. **Blog "Union of the State – A Data Lake Use Case.** Disponível em <https://jamesdixon.wordpress.com/>. Acesso em: 19 abr. 2017.

GENTLEMAN, Robert; IHAKA, Ross. R: **A language for data analysis and graphics.** Americam Statistical Association, Instititute of Mathemathical Statistics and Interface Foundations of North America. Journal of Computational and Graphical Statistics. Volume 5, number 3, pages 299-314. Disponível em: http://biostat.mc.vanderbilt.edu/wiki/pub/Main/JeffreyHorner/JCGSR.pdf. Acesso em: 19 abr. 2017.

INMON, William. H. **Building the data warehouse**. Wiley: New York, 1995.

ISMAIL, Salim; MALONE, Michael S.; GESST, Yuri van. DIAMANDIS, Peter H. **Exponential Organizations**. Diversions Books, 2014.

KIMBALL, Ralph; ROSS, Margy. **DATA WAREHOUSE TOOLKIT**. 3 ed. Editora John Wiley & Sons, Inc. Indianapolis, 2013.

MARZ, Nathan; WARREN, James. **Big Data - Principles and best practices of scalable realtime data systems**. Publicado em : 27 abr. 2007. Disponível em: <http://projecteuclid.org/download/pdf_1/euclid.aoms/1177704711. Acesso em: 16 jan. 2015.

_____. **Lambda Architecture**. Disponível em: <http://lambda-architecture.net/>. Acesso em: 9 out. 2014.

TECHBLOG. **Netflix.** Disponível em: <http://techblog.netflix.com/2011/11/benchmarking-cassandra-scalability-on.html>. Acesso em 20/04/2017.

MITCHELL, Tom. **Departamento de Machine Learning**. s/d. Disponível em: <http://www.ml.cmu.edu/>. Acesso em: 16 jan. 2015.

NIXON, James. **Union of the State - A Data Lake Use Case.** Publicado em 22 jan.2015. Disponível em: http://www.pentaho.com/blog/2015/01/22/union-of-the-state-a-data-lake-use-case. Acesso em: 8 dez. 2015.

PAULUCCI, Anderson. **Data Lake - Uma nova abordagem para o DW**. Publicado em 28 de janeiro de 2017. Disponível em: https://pt.linkedin.com/pulse/data-lake-uma-nova-abordagem-para-o-dw-anderson-paulucci.

SATHI, Arvind. **Big Data Analytics**. IBM Corporation: MC Press Online, 2012.

THE R FOUNDATION. **What is R?** Disponível em: https://www.r-project.org/ Acesso em: 8 dez. 2015.