UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

INSTITUTO DE MATEMÁTICA

CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

DAVID CHRISTIAN ALENCAR GOMES

DESENVOLVIMENTO DE SOLUÇÕES PARA EXPLORAÇÃO DE DADOS: WATSON ANALYTICS E AS CAPACIDADES DE FERRAMENTAS ANALÍTICAS

RIO DE JANEIRO

2020

DAVID CHRISTIAN ALENCAR GOMES

DESENVOLVIMENTO DE SOLUÇÕES PARA EXPLORAÇÃO DE DADOS: WATSON ANALYTICS E AS CAPACIDADES DE FERRAMENTAS ANALÍTICAS

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Ciência da Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Profa. Maria Luiza Machado Campos

RIO DE JANEIRO

2020

CIP - Catalogação na Publicação

Gomes, David Christian Alencar
G633d Desenvolvimento de soluções para exploração de dados: Watson Analytics e as capacidades de ferramentas analíticas / David Christian Alencar Gomes. -- Rio de Janeiro, 2020.
72 f.

Orientador: Maria Luiza Machado Campos.
Trabalho de conclusão de curso (graduação) - Universidade Federal do Rio de Janeiro, Instituto de Matemática, Bacharel em Ciência da Computação, 2020.

1. Análises avançadas. 2. Inteligência de negócio. 3. Exploração e tratamento de dados. 4. IBM Watson Analytics. 5. IBM Cognos Analytics. I. Campos, Maria Luiza Machado, orient. II. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

DAVID CHRISTIAN ALENCAR GOMES

DESENVOLVIMENTO DE SOLUÇÕES PARA EXPLORAÇÃO DE DADOS: WATSON ANALYTICS E AS CAPACIDADES DE FERRAMENTAS ANALÍTICAS

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Ciência da Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em 22 de junho de 2020.

BANCA EXAMINADORA:

Maria Luiza Machado Campos, Ph.D. (UFRJ)

(Participação por videoconferência)

Nome do Professor, Titulação (Instituição)

(Participação por videoconferência)

Nome do Professor, Titulação (Instituição)

Dedicatória: Aos meus pais, por todo esforço e dedicação em me dar a melhor educação, e à minha amada esposa, namorada e melhor amiga Jenny por ser minha razão de tudo.

AGRADECIMENTOS

À minha orientadora, professora Maria Luiza Machado Campos, pela colaboração direta neste trabalho com embasamento e material que me auxiliaram imensamente a traçar um caminho. Além disso, por compreender as minhas limitações e ocupações e por ter sido a melhor professora que eu já tive em toda minha vida, me ajudando sempre em tudo que eu necessitei no que tange a parte acadêmica e sendo um símbolo e exemplo do que é verdadeiramente ser um professor em um meio acadêmico ainda tão carente de empatia com seus alunos.

RESUMO

Em conjunto ao avanço de toda a infraestrutura e capacidade no mundo dos negócios, tem-se o crescimento da necessidade do melhor aproveitamento e melhor gerenciamento dos dados gerados todos os dias em seus diversos ambientes.

Nas últimas décadas, nota-se uma evolução acelerada e impulsionada pela necessidade de respostas cada vez mais rápidas à competitividade numa economia globalizada, que acabou por originar toda uma área da Tecnologia da Informação (TI) dedicada ao estudo da extração, armazenamento e exploração de dados oriundos de diferentes áreas da empresa, a assim chamada de Business Intelligence (BI) ou Inteligência de Negócio.

A partir daí a área continuou a evoluir, de forma a atender as demandas e expectativas das instituições, levando a ciência de dados a um patamar de estudo cada vez mais analítico e preciso, ao mesmo tempo em que se facilita o uso e a exibição de resultados de forma a garantir inclusive um Self-service BI, passando ao usuário do negócio maior autonomia no atendimento de suas necessidades.

Assim, entre os recursos atuais para soluções de apoio à análise de dados, direcionadas à análise exploratória de dados, este trabalho visa analisar as ferramentas analíticas da IBM, através de uma aplicação geradora de insights acerca de um conjunto de dados abertos sobre trabalho escravo, assim como mostrar a trajetória evolutiva no tratamento dos dados frente aos principais tópicos do mundo de negócios que mudaram a visão sobre TI nesse universo corporativo ao longo dos anos.

Por fim, avalia-se também o atual estado da arte da exploração de dados, analisando potencialidades e limitações, utilizando os resultados como uma pequena contribuição para ampliar estudos em um domínio de tanta relevância como o do trabalho escravo, todavia ainda com pouca visibilidade.

Palavras Chave: Análises Avançadas. Inteligência de Negócio. Exploração e Tratamento de Dados. IBM Watson Analytics. IBM Cognos Analytics. SaaS. Self-service BI.

ABSTRACT

Together with the advancement of all infrastructure and capacity in the business world, there has been a growing need to better use and better manage the data generated every day in its various environments.

In the last decades, there has been an accelerated evolution, driven by requirements for increasingly rapid responses to competitiveness in a globalized economy, which has led to an entire Information Technology (IT) area dedicated to the study of the extraction, storage, and exploration of analytical data from different areas of a company, on what has been called Business Intelligence (BI).

From then on, the area continued to evolve to meet the demands and expectations of institutions, taking data science to an ever more analytical and precise level of study, while facilitating the use and the display of results to guarantee even the so called Self-service BI, which gives the business user more autonomy to fulfill his own analytical needs.

Thus, among the current features of Data Analytics, directed to exploratory data analysis, this paper aims to analyze IBM's analytics tools, through an example of insights generation on an open dataset that deals with slave labor, as well as to show the evolutionary trajectory in the treatment of data against the main topics of the business world that have changed the vision of IT in this corporate universe over the years.

Finally, the current state of the art of data exploration is also assessed, analyzing potentialities and limitations, using the results as a small contribution to expanding studies in a domain as relevant as slave labor, yet with little visibility.

Keywords: Advanced Analytics. Business Intelligence. Data Mining and Management. IBM Watson Analytics. IBM Cognos Analytics. SaaS. Self-service BI.

LISTA DE ABREVIATURAS E SIGLAS

ANOVA	Analysis of variance
BA	Business Analytics
BD	Base de Dados
BI	Business intelligence
CPF	Cadastro de Pessoas Físicas
DL	Deep Learning
DM	Data mart
DW	Data warehouse
ETL	Extraction, Transformation and Loading
FOFA	Força, Oportunidade, Fraqueza e Ameaça – Análise FOFA
IA	Inteligência Artificial
IBM	International Business Machines Corporation
IBMWA	IBM Watson Analytics
ML	Machine Learning
OLAP	OnLine Analytical Processing
OLTP	OnLine Transaction Processing
SAAS	Software as a service
TI	Tecnologia da Informação
SPSS	Statistical Package for the Social Sciences
SWOT	Strenght, Weakness, Opportunity and Threats - Análise SWOT

SUMÁRIO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

A inteligência de negócio (*business intelligence* ou simplesmente BI) passou a compor o cotidiano dos negócios, os quais passaram a destinar um orçamento especificamente para o desenvolvimento dessa área internamente, com uma equipe especializada, e não mais com a ideia de que investimento em Tecnologia da Informação (TI), pode se comparar a um investimento em commodity.

Com o crescimento exponencial na quantidade de dados e a necessidade de melhores estratégias para lidar com esse aumento, ocorreu então uma reformulação da posição da TI dentro do escopo estratégico das empresas, passando de um caráter complementar a uma força competitiva atuante no êxito de cada um dos setores do negócio. A atuação da empresa na economia globalizada está correlacionada com sua capacidade na tomada de decisões sobre os dados que possui, muito além de simplesmente produzir e vender.

Segundo Porter (2005), a TI afeta diretamente a forma de competição entre as empresas, reestruturando o setor, adicionando vantagem competitiva e originando novos negócios. Assim, os dados armazenados em bancos ou planilhas processuais passam a ter o potencial, a partir da implementação de um data warehouse (DW), de fornecer informações mais precisas que geram desde economia setorial a melhores decisões de negócios gerenciais.

Na medida em que cresce a necessidade competitiva dos dados, ou seja, da formulação de um ambiente de BI, cresce também a urgência da agilidade na obtenção de informações estratégicas provenientes desses dados. Logo, as ferramentas e aplicativos que os acessam devem ser simples e fáceis de usar e devem retornar os resultados da consulta para o usuário com tempos de espera mínimos (KIMBALL e ROSS, 2013).

Nesse contexto, tem-se o crescimento do *self service* BI, como forma de facilitar e acelerar a obtenção de informações úteis dos dados, todavia, previamente tratados pelos ambientes analíticos. O IBM Watson Analytics é um *software as a service*[1], SaaS, com capacidade cognitiva de aceleração de *insights* que visa trazer um fluxo de interação com o dado a partir de interpretação de linguagem natural, o que motiva o questionamento do papel do cientista de dados e a mensurar em que cenário estamos nessa nova era digital.

[1] Serviço que oferece ao usuário a capacidade de utilizar aplicativos de um provedor com execução em uma infraestrutura direto na nuvem

1.2 OBJETIVOS

O objetivo deste trabalho é realizar um levantamento e discussão sobre as atuais capacidades e limitações das ferramentas de análise de dados , direcionadas à exploração de dados, ou seja, com enfoque na descoberta de informações relevantes acerca de um conjunto de dados. A partir daí realizar uma avaliação sobre as ferramentas analíticas da IBM com intuito de explorar seus recursos de análises exploratórias visuais, levantando *insights* sobre dados abertos já estruturados e tratados, analisando sobre o que se espera, de maneira abrangente, das ferramentas de análises avançadas. Para embasar a discussão sobre o estado da arte desta ferramenta, será realizado um exemplo de aplicação no IBM Cognos Analytics com dados abertos sobre trabalho escravo no Brasil evidenciando os pontos chave das análises avançadas abordadas neste trabalho.

Além disso, de forma complementar, este trabalho visa realizar uma revisão dos termos e nomenclaturas mais utilizadas em BI e análise de dados, realizando uma breve descrição da evolução dessa área da TI correlacionando seu crescimento dentro das empresas e os impactos no ambiente de negócios que levaram ao surgimento das análises avançadas.

1.3 DOMÍNIO DO EXEMPLO DE APLICAÇÃO

Desde 2003, foi estabelecida uma ferramenta governamental de grande importância para fiscalizar e combater o trabalho escravo contemporâneo, a Lista Suja do Trabalho Escravo (SECRETARIA DE TRABALHO, 2019).

A lista suja do trabalho escravo é um cadastro que expõe casos em que houve resgate de pessoas em condições consideradas análogas à escravidão. Antes de serem efetivamente adicionados à lista, os empregadores têm direito de se defenderem.

O enfraquecimentos das leis e fiscalizações de combate ao trabalho escravo ocorrem por grupos de poder que agem na estrutura política para manutenção dessas atividades e impunidade dos envolvidos. É possível, então, relacionar partidos políticos e doações eleitorais de empresas com o envolvimento na lista suja do trabalho escravo.

O foco de nosso exemplo de análises será nos dados das listas sujas do trabalho escravo contemporâneo publicadas desde 2003, incluindo as operações que identificaram trabalho escravo, e dados eleitorais publicados desde 1998 contendo doações exclusivamente diretas para candidatos, assim como os cargos envolvidos de 2002 a 2016 oriundos do trabalho de Verona (2018), que abordou essas relações.

1.4 ESTRUTURA DA MONOGRAFIA

A estrutura deste trabalho visa a compreensão evolutiva dos conceitos para melhor entendimento do estudo de caso realizado. Dessa forma, divide-se nos seguintes capítulos:

- **Capítulo 1:** Introduz o cenário a ser discutido, com os objetivos do trabalho e do estudo de caso desenvolvido.

- **Capítulo 2:** Aborda a linha evolutiva da Inteligência de Negócio, evidenciando seus desafios e tecnologias na exploração de dados, além de definir e correlacionar termos e conceitos.

- **Capítulo 3:** Explora a relação entre a área de ciência de dados com o mundo dos negócios, situando a nova visão de TI dentro dos conceitos corporativos.

- **Capítulo 4:** Retrata a evolução na análise de dados, relacionando conceitos inerentes às principais aplicações com recursos na nuvem, situando a análise avançada de dados com aspectos de Inteligência Artificial (IA).

- **Capítulo 5:** Desenvolve as capacidades e limitações do IBM Watson Analytics frente ao cenário de exploração inteligente e análise avançada de dados com o estado da arte atual das ferramentas de análise de dados.

- **Capítulo 6:** Apresenta o uso da aplicação IBM Cognos Analytics sobre um conjunto de dados abertos já estruturados sobre trabalho escravo, unindo a avaliação tecnológica e conceitual do processo de exploração de informações relevantes dos dados.

- **Capítulo 7:** Denota as conclusões acerca do que foi desenvolvido ao longo do trabalho, situando os desafios para trabalhos posteriores.

2 EVOLUÇÃO DA INTELIGÊNCIA DE NEGÓCIO NA EXPLORAÇÃO DE DADOS

2.1 PANORAMA

Na atual perspectiva da tecnologia, o período de transição evolutivo é muito curto e, dessa forma, poderia se dizer que o passar de uma década, no referencial da TI, não é linearmente comparável ao mesmo período da década passada, sendo inclusive exponencialmente distinto.

A profissão e a prática de sistemas de informação são, certamente, imaturas, porque existem somente desde o início dos anos 60 e teve seu crescimento acelerado sob demanda o que acarretou diversos pontos falhos (INMON, 2002).

Nesse cenário, compreendemos que o surgimento e crescimento dos sistemas de suporte à decisão e dos sistemas analíticos acompanharam essa dinâmica de crescimento. Portanto, não é raro verificar usos incorretos de nomenclaturas sobre arquiteturas e tecnologias, a exemplo de classificar ambientes de Processamento de Transações em Tempo Real, Online Transaction Processing (OLTP), como BI.

Dessa forma, com um avanço exponencial das tecnologias, seguido de uma necessidade de uso que por vezes não permite o amadurecimento necessário delas, tem-se o uso de estruturas modernas, mas ainda não bem definidas, em projetos de BI que, consequentemente, acarretam diversos investimentos perdidos e projetos abandonados pela falta de retorno útil ao usuário.

Essa explosão de desenvolvimento na exploração de dados não poderia ter se iniciado senão a partir da evolução na tecnologia de armazenamento. De fato, a amplitude na capacidade de armazenamento e gerenciamento de dados sobre novos tipos de mídia abriu o caminho para esse desenvolvimento (INMON, 2002). Todavia, todo crescimento exponencial de arquitetura acaba por gerar novos desafios, como a redundância de dados, a credibilidade do próprio dado e o desenvolvimento correlato da complexidade em transformar um montante de dados em informação relevante para agregar conhecimento.

Os dados são o ponto focal de todo BI, de maneira que os processos envolvidos, desde a extração, estão pautados na solução dos problemas relatados. A etapa de análise dos dados, com foco nos negócios, é a mais crítica em um roteiro de construção de um BI (MOSS, 2003).

Assim, pode-se definir um roteiro de BI, no qual o armazenamento de dados é um obstáculo inicial, seguido da análise dos dados como um ponto de atenção crítico, depois temos a verificação de credibilidade, o cruzamento de fontes, a manutenção dos ambientes até finalmente a obtenção de uma informação gerencial de alto nível para tomada de decisão de negócio.

2.2 REVISÃO DE CONCEITO DA ÁREA DE BI

2.2.1 Inteligência de negócio

Esta seção visa uma revisão de literatura dos termos mais utilizados ao longo deste trabalho, de forma que não serão apresentados detalhes de arquitetura e modelos de tudo que compõe o universo do BI, pois o foco está na evolução da área de exploração de dados até o que ela representa hoje em termos de análise e extração de informação de dados em um cenário corporativo.

Em meados de 1958, com a publicação do artigo escrito pelo cientista da computação da IBM, Hans Peter Luhn, o potencial do BI foi reconhecido. Nesse artigo, Luhn (1958) descreveu o BI como um sistema automático, desenvolvido para disseminar informações para os diversos setores de qualquer organização industrial, científica ou governamental.

O BI não é um produto nem um sistema. É uma arquitetura com uma coleção de aplicativos operacionais integrados e de suporte à decisão que fornecem à comunidade corporativa fácil acesso a dados de negócio (MOSS, 2003).

Com isso, o BI é um conjunto orquestrado de softwares, serviços e processos para processar e extrair dados e apresentá-los em diferentes visualizações amigáveis ao usuário, como relatórios, visualizações, painéis etc., visando o auxílio à tomada de decisões.

No que se refere à composição do ambiente de inteligência de negócio, há uma diversidade de definições, segmentações e detalhamentos. Todavia, pode-se então, de maneira mais alto nível, dividir em 4 grandes áreas, conforme mostra a figura 1: origem dos dados, área de transição (*staging)*, *data warehouse* e *data mart (DM)*. É comum encontrar o *data warehouse* englobando todos os processos e inclusive sendo utilizado como sinônimo de BI, em grau conceitual, e, portanto, essas áreas estão dívidas neste trabalho na sua forma prática de uso.

Segundo Moss (2003), iniciativas de apoio à decisão são esforços dispendiosos, pois dados corporativos de diferentes setores devem ser extraídos e mesclados e exigem que sejam consideradas novas tecnologias, tarefas adicionais a serem desempenhadas, transferência de funções e responsabilidades, e ainda, por fim, os aplicativos de suporte à decisão sejam entregues rapidamente em qualidade aceitável. Com isso, incríveis 60% dos projetos de BI terminam em abandono ou falha devido a planejamento inadequado, prazos perdidos, má gestão de projetos ou resultados de baixa qualidade. Os responsáveis pelos projetos precisam, então, saber os prós e contras na implementação do BI baseados em experiência prática confiável.

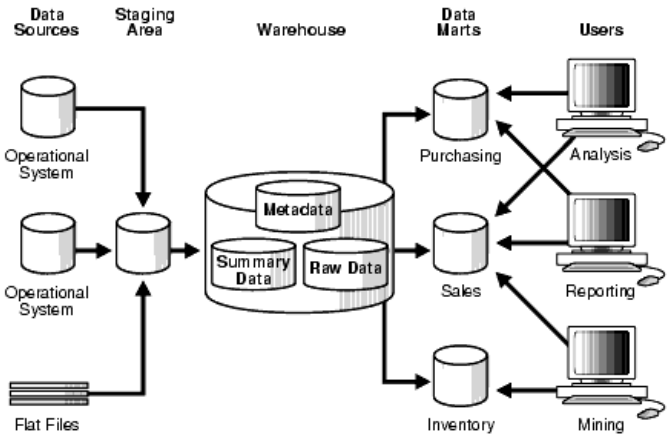

Figura 1 : Arquitetura de um DW com área de *staging* e DM
Fonte: Web Stanford Edu (2003)

Dessa forma, quando um problema de negócio ou oportunidade de negócio é definida, uma solução de BI é proposta. Cada versão do aplicativo de BI é usualmente justificada pelo custo e deve se definir claramente os seus benefícios, seja resolvendo um problema de negócio específico ou visando uma análise futura mais abrangente.

Essa análise necessita de aprofundamento técnico com alinhamento severo a uma política de custos que leve a justificar cada etapa a ser desenvolvida com intuito de atender demandas mais urgentes de uma área ao mesmo tempo em que se planeja uma estrutura adequada para adições futuras.

2.2.2 Origem dos dados

A camada representativa da origem dos dados é composta por diversas infraestruturas relevantes para as análises de negócio. Com isso, define-se o conjunto de ambientes transacionais presentes em uma empresa sob a sigla OLTP, ou seja, tudo que esteja categorizado no processamento de dados em tempo real de solicitação orientado a transações. Segundo Inmon (2002), no ambiente OLTP, o tempo de resposta é quase sempre de missão crítica. O negócio começa a sofrer imediatamente quando o tempo de resposta for alto.

O OLTP, de maneira geral, será composto por múltiplos tipos de transações, como inserir, atualizar e excluir por muitos usuários concomitantes e a camada de origem dos dados pode conter bancos de dados transacionais, arquivos de texto e planilhas, majoritariamente sendo fontes mais específicas de cada unidade de negócio, dando independência de gestão de

dados a cada setor. Os sistemas OLTP tem como função executar transações eficientes, e consultar dados transacionais, através de uma divisão em partes menores, normalizadas, que são, consequentemente, menos redundantes.

Com isso, o OLTP adquire capacidade de processamento de grandes quantidades de transações de forma independente e preserva-se do excesso de carregamento associado a manutenção da integridade dos dados por conta de redundância.

2.2.3 Área de transição

Para Inmon (2002), a área de transição, s*taging*, é um local onde os dados em trânsito são colocados, geralmente provenientes de ambientes legado, antes de entrar na camada de extrair, transformar, carregar, ETL. Dessa forma, a *staging* funciona como um grande repositório inicial, onde os dados de diversos sistemas de origem são extraídos e armazenados.

No cotidiano de um negócio, a camada de *staging* possui duas propriedades fundamentais. A primeira é que esta camada não é acessada pelo usuário final da aplicação, afinal o dado ainda não está limpo, pode estar redundante, entre outros casos. A segunda propriedade é que ela funciona como um espelho com histórico de dados da origem, evitando perda de informação e, assim, podendo se realizar averiguações sobre dados que podem ser oriundos de tabelas transacionais, arquivos, e-mails e todo tipo de fonte de dados que podem não estar mais disponíveis.

Assim, ao invés do ETL realizar requisições dos dados diretamente da fonte de origem do arquivo, do transacional, os processos de transformação acessam os dados da *staging* para iniciarem os tratamentos de implantação.

Uma vez que é comum a existência de múltiplos sistemas origem, acaba sendo recorrente a indisponibilidade de alguns sistemas ou até mesmo a perda de histórico dos dados em bancos transacionais.

Logo, a área de *staging* é fundamental como um repositório de dados brutos, espelho da origem da informação, para uso pela camada de ETL, evitando problemas de performance e garantindo segurança dos dados históricos.

2.2.4 Data warehouse

Como citado no início deste capítulo, definir conceitualmente um *data warehouse* acaba sendo muito próximo de definir um BI, devido a sua importância quantitativa e qualitativa dentro dessa definição.

Dessa forma, a camada de DW compõe uma parte do processo, na qual, tecnicamente definindo, as informações ficam armazenadas após serem efetuados todos os processos de limpeza e transformação do dado. Esses processos de limpeza e transformação são os que agregam valor a esta etapa, pois não haverá somente um repositório de dados, mas sim a qualificação do dado para um estágio em que ele pode ser utilizado, sintetizado e, mais importante, o DW irá manter um registro histórico das inserções e atualizações das informações, pois o DW possui diferentes níveis de agregação e detalhamento, assim como diferentes níveis de idade dos fatos (INMON, 2002).

Segundo Kimball e Ross (2013), a construção de um *data warehouse* é dado pela combinação da necessidade de informações de um negócio com os dados que realmente estão disponíveis para acesso. Essa trajetória dos dados transacionais até se tornarem informação útil, está, efetivamente, na combinação da inteligência do negócio com o potencial da tecnologia do DW.

Nesta camada, também se armazena informações sobre os dados contidos dentro dele, os Metadados, os quais são responsáveis por descrever e detalhar esse dado, tornando-o acessível e interpretável.

Metadados são informações que descrevem várias facetas de um ativo de informação para melhorar a sua usabilidade em todo o seu ciclo de vida. Então, é a definição de metadados que fornece o entendimento do valor dos dados (GARTNER, 2019).

2.2.5 Data mart

O *data mart* é a etapa final no que tange à preparação do dado para consumo pelo usuário em diversas aplicações. Pode se defini-lo como a camada responsável por organizar os dados em uma estrutura multidimensional com dimensões e fatos contendo métricas que podem ser analisadas por diferentes perspectivas.

Ou seja, é nessa camada onde efetivamente sintetiza-se a informação para cada área de negócio dentro da empresa utilizando o dado em caráter multidimensional, seguindo bons modelos de padronização, como o esquema estrela ou o snowflake.

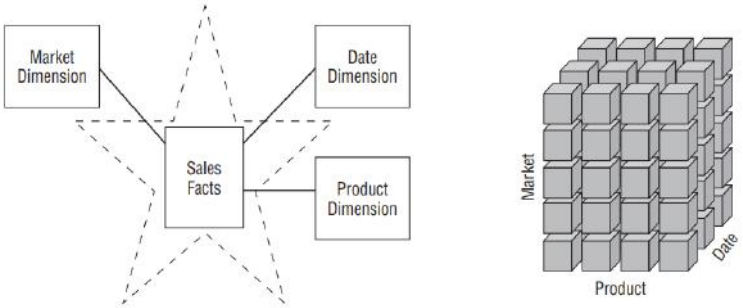

Figura 2 : Esquema Estrela versus Cubo OLAP
Fonte: KIMBALL e ROSS (2013)

Modelos dimensionais implementados em sistemas de gerenciamento relacionais são chamados de "esquema estrela" devido a sua semelhança com a forma de uma estrela. Já modelos dimensionais implementados em ambientes de dados multidimensionais são também chamados de cubos de processamento analítico online (online analytical processing - OLAP), (KIMBALL e ROSS, 2013).

É comum o BI incluir ambas as aplicações, esquemas estrela ou cubos OLAP, conforme mostra a figura 2, pois ele aproveita os conceitos dimensionais, uma vez que ambos possuem um design lógico comum com dimensões reconhecíveis, no entanto com implementações físicas distintas.

Segundo Kimball e Ross (2013), um dos grandes desafios está em como os dados são transferidos do *data warehouse* para o *data mart*, pois os dados no DW são muito granulares, enquanto os dados no DM são muito compactos e resumidos. Assim, periodicamente, os dados devem ser movidos entre eles e esse movimento de dados é análogo ao movimento de dados no data warehouse oriundo do ambiente legado operacional.

Dessa forma, os dados do *data warehouse* devem ser selecionados e, em seguida, reformulados para atender às necessidades do *data mart*, conforme mostra a figura 3. Geralmente, os dados do *data mart* residem em cubos e esses precisam ser processados e muitos cálculos diferentes precisam ser executados nos dados detalhados que residem no *data warehouse*. Em suma, um processo não trivial ocorre quando os dados são transmitidos de um mundo normalizado para um mundo multidimensional. Por fim, é necessário o levantamento

de quantos dados devem ser acessados e com que frequência o processo de atualização deve ser executado (KIMBALL e ROSS, 2013).

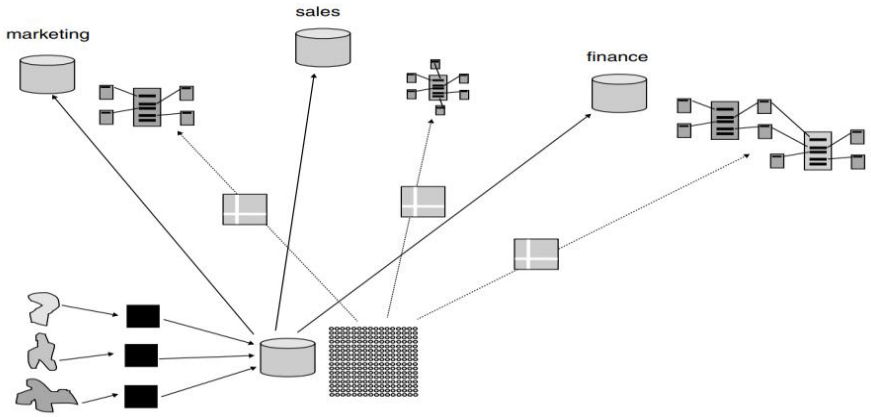

Figura 3 : A relação entre o DW e o DM

Fonte: KIMBALL e ROSS (2013)

Por fim, uma vez que se tem o dado em seu estado mais analítico e gerencial possível, diferenciando-se completamente daquele dado bruto e desnormalizado das origens, o dado está pronto para atender diversas aplicações de consumo, visualização e análise para os usuários.

2.3 PROGRESSO DO BI

O BI representou um grande salto no que se refere a forma como os dados são vistos e tratados. É notável, que ao longo da história, a importância da evolução na capacidade de armazenamento crescia em conjunto com a necessidade das empresas em armazenarem mais informação e de maneira mais segura e estável.

Todavia, todo o conceito na construção de um *data warehouse* acaba indo além de apenas guardar dados. Ele passa a representar uma evolução no conceito de integração dos dados oriundos de diversas infraestruturas dentro de um mesmo negócio.

No passado, os sistemas não eram projetados ou construídos com a integração em mente. Cada sistema tinha um começo e um fim, e cada sistema foi projetado para resolver apenas um problema isolado para um grupo de pessoas de uma mesma linha de negócios. As antigas práticas de desenvolvimento estáticos não são adequadas para iniciativas de BI integradas, devido à ausência de integrações interorganizacionais necessárias para sustentar um ambiente

de suporte a decisões em toda a empresa. Antigamente as atividades interorganizacionais não só eram consideradas desnecessárias, mas também eram percebidas como obstáculos ao progresso, porque retardavam os projetos (MOSS, 2003).

Construções em cima de dados organizacionais setorizados acabavam por causar inúmeras limitações e obstáculos nas análises e tentativas de entendimento do porquê de certos eventos ocorrerem. Entre esses desafios, estão a falta de credibilidade dos dados e a produtividade nas solicitações.

Conforme a figura 4, a estrutura organizacional de forma setorizada causa discrepância de informação entre setores: no exemplo da figura, o departamento A chega ao valor de +10%, enquanto o departamento B chega em -15%.

Segundo Inmon (2002), toda vez que uma nova extração é feita, as probabilidades de uma discrepância surgem devido à temporização ou ao diferencial algorítmico nesse caso setorizado. É comum que uma corporação tenha múltiplos níveis de extração realizados desde o momento em que os dados entram no sistema da empresa até o preparo para o gerenciamento.

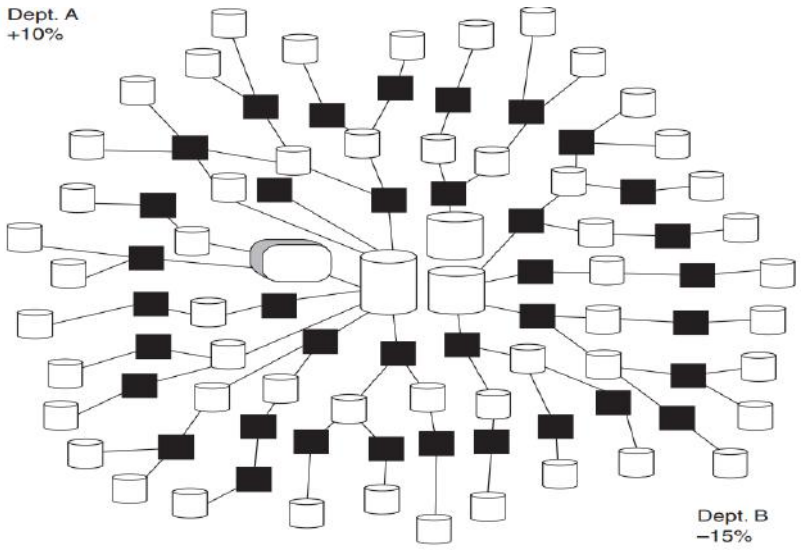

Figura 4 : Falta de credibilidade na evolução da arquitetura

Fonte: INMON (2002)

Há extratos, extratos de extratos e assim por diante, fazendo com que cada novo nível de extração amplie os problemas e riscos causados.

Dadas essas razões, acaba por constituir uma crise de credibilidade em todas as organizações que permitam que seu legado de hardware, software e dados evolua sem uma integração como de um data warehouse.

Outro fator citado é a falta de produtividade nessas solicitações. Objetivamente, solicitações de criação de relatórios gerenciais para sistemas apartados demandam uma enorme quantidade de tempo e acabam ainda contendo falhas, conforme ilustra a figura 5.

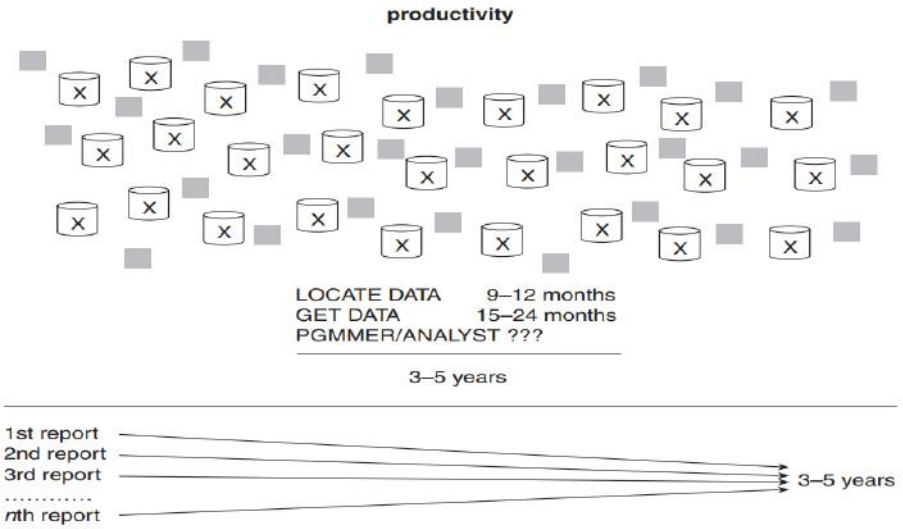

Figura 5 : Falha de geração de relatórios futuros

Fonte: INMON (2002)

Nesse sentido, gerar relatórios envolvendo todas as áreas era um desafio árduo e, principalmente, custoso para as empresas, alocando inúmeros recursos e tendo como resultado um cenário estático, ou seja, para uma nova solicitação todo o esforço teria de ser reempregado.

O estopim da evolução do processamento de informações se dá em meados da década de 1960, com a explosão dos arquivos-mestres e fitas magnéticas que acabaram por produzir enormes quantidades de dados redundantes. A chegada da inteligência de negócio representou a solução efetiva dessas e outras complexidades envolvidas no trabalho sobre dados nas últimas décadas (INMON, 2002).

Assim, o progresso do BI revolucionou toda uma infraestrutura e conceitualização na exploração de dados, uma vez que as suas aplicações apresentam características fundamentais que se fundamentaram nas dificuldades apresentadas ao longo das décadas, segundo Moss (2003):

- As aplicações de BI são principalmente impulsionadas por oportunidades de negócios e não por necessidade de negócios.

- As aplicações de BI implementam uma estratégia de apoio à decisão entre organizações, em vez de silos departamentais de apoio à decisão.
- Os requisitos de suporte à decisão de BI são principalmente requisitos de informações estratégicas, e não requisitos funcionais operacionais.

A análise de projetos de BI enfatiza a análise de negócio em vez da análise do sistema, e a análise é a atividade mais importante no desenvolvimento de um ambiente de suporte a decisões de BI.

3 GESTÃO ESTRATÉGICA DE TI E SEU CARÁTER CORPORATIVO

3.1 CENÁRIO CORPORATIVO DA TI

Termos e conceitos no cenário corporativo possuem uma forte base consolidada, uma vez que estudos econômicos e administrativos possuem longa data de sustentação. Todavia, na era da informação, novos tópicos entram na equação dos elementos competitivos de mercado que podem predizer o sucesso ou o fracasso de uma organização.

Historicamente, a TI possui, economicamente, o caráter de uma *commodity* dentro da empresa. O termo *commodity* é definido como um artigo de comércio, uma unidade permutável de riqueza econômica, especialmente um produto primário ou matéria-prima (Dicionário Collins, 2019). Neste sentido, interpretar a TI dentro da empresa como uma *commodity* é equivalente a compreender somente a questão tecnológica, como os recursos de infraestrutura necessários ao funcionamento de uma empresa.

No entanto, uma típica *commodity* obedece a padrões rígidos e seu preço varia com a demanda, como ocorre com produtos de mercado ou serviços prestados como a distribuição de água. A TI não segue esses padrões amarrados, uma vez que se pode não somente encontrar discrepâncias acentuadas no valor de prestação de consultoria em tecnologia, como também na composição da TI no setor de infraestrutura e no setor estratégico, vide o próprio BI.

"Quanto maior a presença da TI na sociedade, mais recursos financeiros são canalizados para ela e maior a dependência de empresas e pessoas com relação ao seu uso" (LOBATO, 2012, p.117). Dessa forma, a Tecnologia da Informação está presente em todos os setores de uma empresa, através de computadores, notebooks, smartphones, servidores, rede e tudo mais o que se possa imaginar ao que se refere ao caráter operacional e, por conseguinte, a TI está envolvida na maioria das atividades empresariais.

Contudo, houve uma imensa alteração neste cenário com o advento do BI, seguido de todo o conhecimento empregado na exploração de dados. A partir de então, a TI é inserida em um contexto antes regido somente por um caráter administrativo e passa a dominar o caráter gerencial e estratégico a partir do controle dos dados para o suporte à decisão.

Assim, a TI pode afetar diretamente a forma como as empresas competem (PORTER, 2005). Como resultado, há um crescimento nos estudos sobre o retorno financeiro e qualitativo do investimento empregado nos setores tecnológicos das empresas que acabam por denotar

como a TI participa de uma cadeia de valor, agregando valor e transformando a maneira como as atividades podem ser exercidas em todos os cenários (PORTER, 2005).

3.2 CONCEITOS ESTRATÉGICOS DE NEGÓCIO

3.2.1 Análise SWOT

O termo SWOT é um siglônimo dos termos *strengths, weaknesses, opportunities* e *threats* que significam respectivamente: forças, fraquezas, oportunidades e ameaças, também referenciada no Brasil pela sigla FOFA.

O SWOT pode ser aplicado para a totalidade de uma empresa ou para projetos individuais em um único departamento. As análises SWOT são utilizadas a nível organizacional para determinar o quanto uma empresa está alinhada com suas trajetórias de crescimento e referências de sucesso.

Então, uma empresa deve empregá-la como uma ferramenta de autoconhecimento, aprofundando-se a respeito do seu negócio (SEBRAE, 2018). Com isso levantando os pontos elementares conforme o quadro 1.

> "A análise de SWOT e a escolha de grupos estratégicos em que competir têm implicações para a formulação de estratégia. Portanto, estas decisões precisam levar em consideração a comparação entre os pontos fortes e fracos de uma empresa, sejam eles estruturais, sejam em relação à implementação da estratégia selecionada, destacando suas competências distintivas de forma a maximizar as oportunidades e minimizar as ameaças oriundas do ambiente competitivo" (CARVALHO e LAURINDO, 2010, p. 52).

Logo, a análise SWOT, como mostra o quadro 1, situa as características que elucidam as vantagens e desvantagens de uma empresa em uma avaliação ponderada sobre cada aspecto dos setores estratégicos, e é nesse contexto que o setor de TI de uma empresa passa a compor esta avaliação como um diferencial competitivo, seja já como um ponto forte ou como uma oportunidade dentro da estratégia da empresa, principalmente devido ao suporte a tomada de decisão oferecido pelo BI. Mas é a matriz SWOT que complementa efetivamente a partir da adesão de cenários internos e externos que decisões tomar a partir desse levantamento de dados.

Quadro 1 : Análise SWOT e grupos estratégicos

S Pontos fortes	**W Pontos fracos**
Fatores que constroem barreiras de mobilidade. Fatores que reforçam o poder de negociação de seu grupo com compradores e fornecedores. Fatores que isolam seu grupo da rivalidade de outras empresas. Escala maior em relação a seu grupo estratégico. Fatores que permitem custos menores de entrada em seu grupo estratégico. Forte capacidade de implementação de sua estratégia em relação a seus concorrentes. Recursos e habilidades que permitem superar barreiras de mobilidade e penetrar em grupos estratégicos mais interessantes.	Fatores que debilitam barreiras de mobilidade. Fatores que enfraquecem o poder de negociação de seu grupo com compradores e fornecedores. Fatores que expõem seu grupo da rivalidade de outras empresas. Escala menor em relação a seu grupo estratégico. Fatores que causam custos maiores de entrada em seu grupo estratégico. Capacidade menor de implementação de sua estratégia em relação a seus concorrentes. Falta de recursos e habilidades que pudessem permitir superar barreiras de mobilidade e penetrar em grupos estratégicos mais interessantes.
O Oportunidades	**T Ameaças**
Criação de um novo grupo estratégico. Mudança para um grupo estratégico em posição mais favorável. Fortalecimento da posição estrutural do grupo existente ou da posição da empresa no grupo. Mudança para um novo grupo e fortalecimento da posição estrutural deste grupo estratégico.	Outras empresas entrarem em seu grupo estratégico. Fatores que reduzam as barreiras de mobilidade do grupo estratégico, diminuam o poder em relação a compradores ou fornecedores, piorem a posição em relação a produtos substitutos ou exponham a uma maior rivalidade. Investimentos que visam melhorar a posição da empresa aumentando as barreiras de mobilidade. Tentativas de superar barreiras de mobilidade para entrar em grupos estratégicos mais interessantes ou inteiramente novos.

Fonte: CARVALHO e LAURINDO (2010)

3.2.1.1. Matriz SWOT

A matriz SWOT é intrinsecamente relacionada à análise SWOT, pois somente após o levantamento dos pontos da análise SWOT que se pode estabelecer as ações adequadas ao cenário que foi identificado, conforme quadro 2.

O quadro 2 mostra as ações que podem ser tomadas frente aos cenários identificados com uma análise interna e externa à empresa em si. Por exemplo, uma empresa concorrente surge oferecendo o menor preço para serviço de armazenamento na nuvem e a empresa analisada já possui esse setor nos pontos fracos, ou seja, um investimento que já tem dificuldades.

Dessa forma, a melhor alternativa seria desativar essa área e focar em outros pontos de oportunidade.

Quadro 2 : Matriz SWOT

Fonte: Adaptação própria do conceito de matriz SWOT

Logo, uma empresa deve enxergar o ambiente externo e o interno em conjunto para definir as estratégias para seus setores identificados na análise SWOT.

3.2.2 As 5 forças de Porter

Segundo Porter (2005), em seu modelo das cinco forças, criado em 1979, as forças competitivas são divididas em cinco fatores, Força dos concorrentes diretos, Força dos compradores, Força dos fornecedores, Força dos novos entrantes e Força dos substitutos. O foco das cinco forças de Porter é examinar o ambiente em que a empresa está inserida e, a partir de levantamentos da estratégia, avaliar um posicionamento frente aos concorrentes, dando uma maior visibilidade do negócio.

O modelo de Porter tem como premissa avaliar, em um determinado setor da economia, o nível de atividade, levantando os elementos que impactam a sua competitividade, e, simultaneamente, prover uma perspectiva global das empresas concorrentes e, também, do negócio ao qual pertence. Nesse aspecto, é necessário aprofundar as informações sobre o que rege as forças competitivas.

Porter (2005), define o modelo das 5 forças como o microambiente da organização, de forma que o uso desse modelo amplia as capacidades da empresa e otimiza suas estratégias. É

relevante considerar que, dependendo do cenário em que se encontra inserido, alguma das vertentes terá maior relevância e impacto no sucesso da empresa.

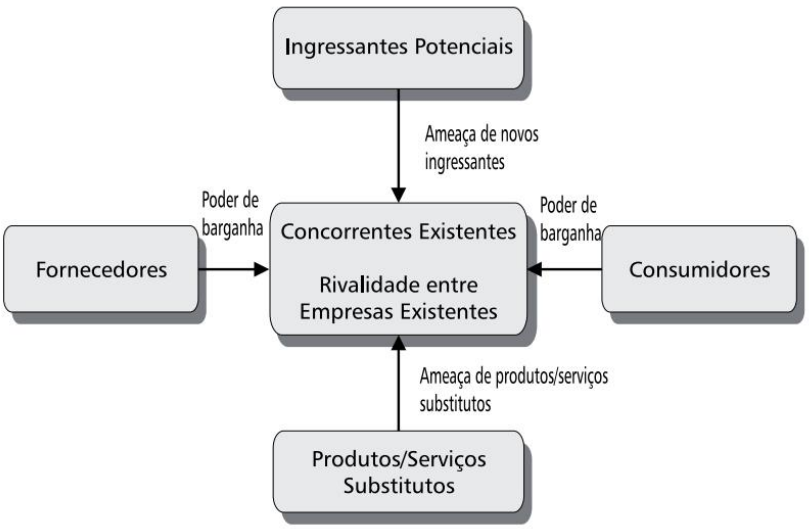

Figura 6 : Análise estrutural da indústria - as cinco forças competitivas
Fonte: LAURINDO e CARVALHO (2010)

Com isso, temos que as 5 forças, representadas na figura 6, exercem pressão simultaneamente e a intensidade dessa pressão e a forma como ela está distribuída entre as 5 forças define o caráter do mercado. Um mercado com uma enorme concorrência interna tende a ter lucros menores, enquanto um mercado que possui a força de entrantes muito enfraquecida pela dificuldade natural de adentrar um determinado ramo acaba por dominar com estabilidade uma fatia de consumidores, aumentando os lucros.

3.3 APLICAÇÃO DOS CONCEITOS E POSIÇÃO DO BI

A partir dos termos trabalhados na seção 3.2, a discussão acerca do setor de TI, em especial o setor de BI dentro das empresas, pode ser aprofundada. Uma vez que se entende que a análise SWOT e a matriz SWOT definem, em suma, os riscos e vantagens nos cenários externos e internos, enquanto as 5 forças de Porter definem o caráter competitivo no qual a empresa está inserida, têm-se a correlação com as características do BI nesse mesmo universo, uma vez que tomar decisões melhores, melhorar as eficiências operacionais, aumentar as

receitas e aumentar a vantagem competitiva são os quatro principais objetivos das organizações de BI atualmente (DRESNER, 2018).

Com papel fundamental dentro das decisões empresariais, o BI altera profundamente o papel da TI, antes visto como *commodity*, e passa a agregar valor à nível estratégico dentro dos conceitos apresentados, uma vez que a TI afeta diretamente a forma de competição entre as empresas, reestruturando o setor, adicionando vantagem competitiva e originando novos negócios (PORTER, 2005).

A figura 7 representa a inserção do BI dentro dos principais ramos corporativos, assim Dresner (2018), em seu estudo de Market, avalia que as indústrias verticais de seguros, tecnologia e serviços de negócios têm a maior taxa de adoção do BI, sendo o setor de seguros o líder, seguido pelo setor de tecnologia, com 40% das organizações.

Setores como o de Seguro estão entre os que mais necessitam de análises precisas para melhor tomada de decisão. Logo, é natural que a inteligência de negócio faça cada vez mais parte desses setores. Assim, pode-se destacar que empresas de seguro estão procurando agregar cada vez mais o BI dentro do seu negócio, pois ele foi considerado um diferencial competitivo dentro de análises como a SWOT e a de Porter no âmbito corporativo.

A figura 7 também apresenta outros setores essenciais que tiveram alta inserção do BI como parte integrante da tomada de decisões de negócio, tais como a área de educação, saúde, finanças e indústrias.

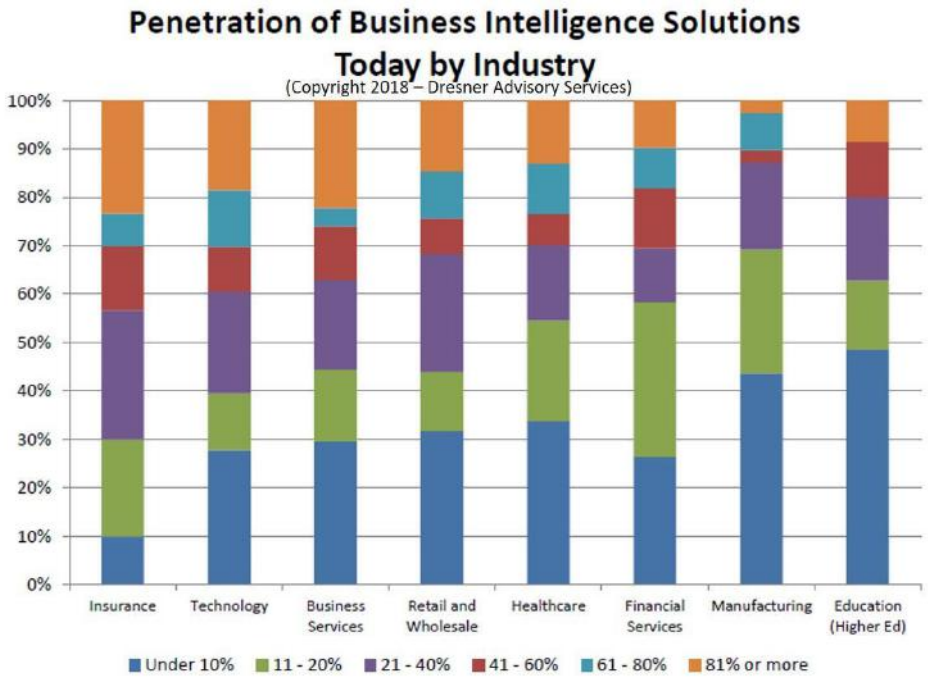

Figura 7 : Participação do BI por tipo de indústria

Fonte: DRESNER (2018)

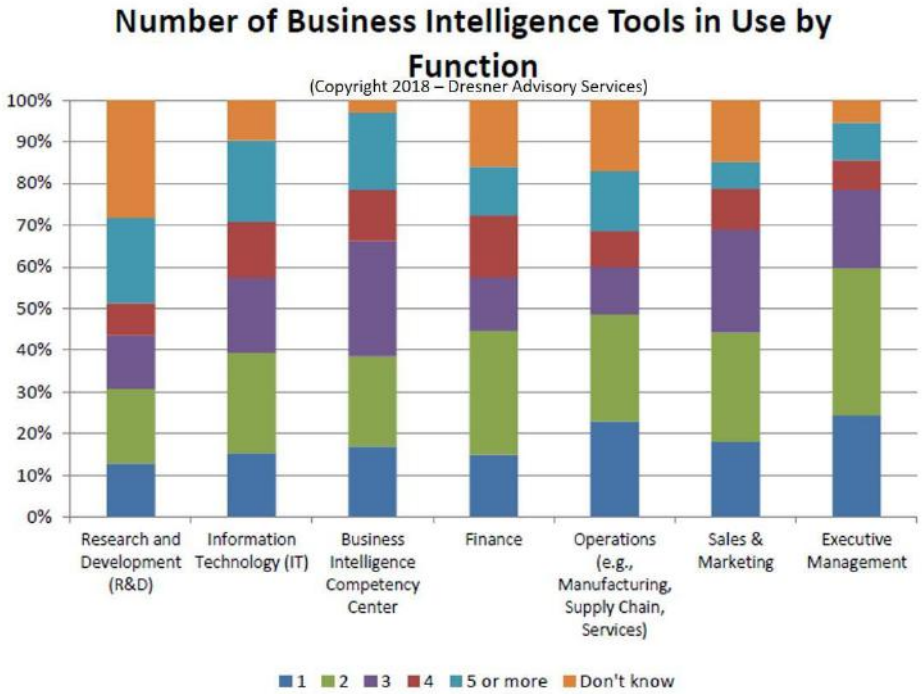

Figura 8 : Número de ferramentas de BI em uso por ferramenta

Fonte: DRESNER (2018)

Também segundo Dresner (2018), os Centros de Competência de inteligência de negócio relatam consistentemente um número maior de ferramentas de BI em uso do que outras áreas funcionais, dado seu forte envolvimento em todas as fases de análise e execução de projetos de BI. TI, Vendas e Marketing e Finanças provavelmente terão mais ferramentas de BI em uso do que as Operações, conforme mostra a figura 8.

Dessa forma, temos inúmeros *data warehouse* já implantados em setores estratégicos de cada empresa, não somente otimizando decisões, mas atuando em conjunto com ferramentas de análise e visualização de dados para redefinir estratégias consolidadas de negócio, como as citadas anteriormente. O setor de TI dentro de uma empresa não está mais atrelado unicamente à infraestrutura e *commodity*, ele passou a ser ponderado em todos os setores de uma organização, em diversos níveis de complexidade que incorporam diversos setores fundamentais ao suporte, gestão e manutenção desse ambiente analítico.

4 A EVOLUÇÃO DA ANÁLISE DE DADOS

4.1 INTRODUÇÃO

A inteligência de negócio tem como ponto focal o aprimoramento da tomada de decisões no meio corporativo. Uma vez que o *data warehouse* foi construído, tem-se uma poderosa gama de dados consolidados e tratados que retratam o presente e o passado da organização.

A partir disso, é necessário extrair dos dados, agora uniformes e bem consolidados, informação útil. Essa análise de dados é realizada através de relatórios e dashboards em cima do BI estruturado, de maneira que o ramo gerencial da empresa seja capaz de tomar decisões em cima disso.

Esse tipo de análise é chamada de análise descritiva, que, segundo o Gartner IT Glossary, é definido como um exame de dados, geralmente feito manualmente, para responder às perguntas como "O que aconteceu?" ou “O que está acontecendo?, caracterizado pelo BI tradicional com visualizações como gráficos de pizza, gráficos de barras, gráficos de linhas, tabelas ou narrativas geradas.

Dessa forma, é comum que grandes empresas necessitem de uma equipe de cientista de dados para extrair informação mais valiosa ou complexa de um conjunto de dados, como por exemplo tentar responder “O que vai acontecer?”. E foi nessa lógica que a análise de dados vem evoluindo de forma a tentar tornar esse acesso à informação útil mais dinâmico, independente e acessível.

Assim, surgiram múltiplas ferramentas que tornam a consulta mais dinâmica, com um simples arrastar dos campos desejados e a informação é atualizada em tempo real. Além de já utilizar os recursos desenvolvidos de IA e mineração de dados para consolidar análises exploratórias avançadas, classificadas como análises avançadas que elevam a análise ao nível preditivo.

Por fim, conceder múltiplos *insights* dinâmicos em tempo de análise e disponibilizar todas essas opções citadas acessíveis de qualquer lugar com acesso à internet através dos SaaS, que possuem a interoperabilidade necessárias ao acesso por múltiplos sistemas distintos, possibilitando a troca de informação eficaz para enriquecer cada análise.

4.2 INTELIGÊNCIA DE NEGÓCIO EM TEMPO REAL (REAL-TIME BI)

Em um *data warehouse* de um BI tradicional, armazena-se os dados tratados em um grande repositório, consumindo informação da origem mensalmente, semanalmente ou até diariamente para então apresentar ao usuário final informações não suscetíveis as instabilidades de um ambiente transacional.

Historicamente, para esses dados, eram então desenvolvidas visualizações, como *dashboards*, pré-determinadas e fixas que atendiam a uma determinada necessidade de negócio. Dessa forma, a cada nova necessidade de negócio, mesmo acerca do mesmo conjunto de informação, como uma troca de consolidação do período da informação, levava então os especialistas a terem de criar uma visão fixa, enquanto tentavam corresponder a necessidade e expectativa de um gerente de negócios.

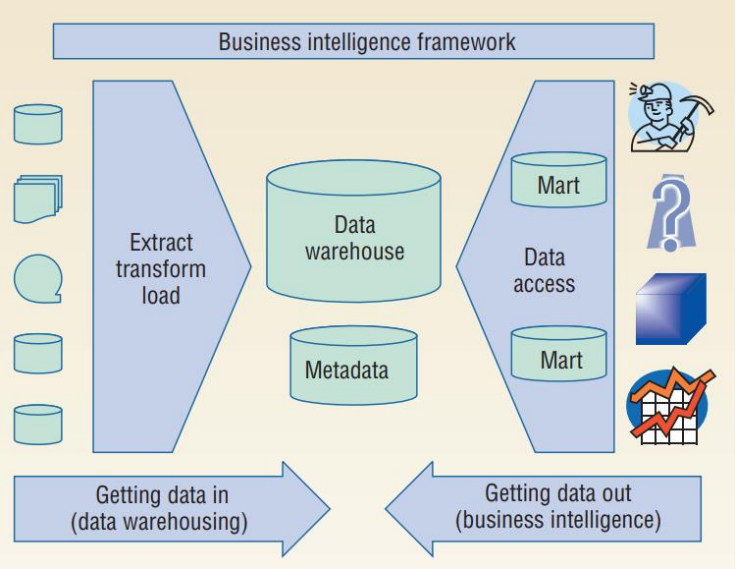

Figura 9 : Business intelligence framework

Fonte: WATSON e WIXON (2007)

Conforme a figura 9, a inteligência de negócio está pautada no retorno de informação útil e é natural avaliar que o crescimento da necessidade de informação competitiva cada vez mais rápida não era compatível com o modelo de entrega de dados sob demanda de alteração por analistas.

Portanto, ferramentas atuais de análise de dados estão conectadas aos DWs de uma empresa e permitem a personalização completa da análise, garantindo ao gerente de negócio a

capacidade de experimentar múltiplos cenários com um simples arrastar de campos, ou seja, garantindo o completo *self-service*, em tempo real, em termos de visualizações sobre os dados.

Contudo, segundo MOSS (2003), embora alguns possam achar que o *data warehouse* deva ser completamente *ad hoc*, ou seja, um ambiente de consulta de autoatendimento com a capacidade de alterar os parâmetros de um relatório para criar sua própria versão personalizada, não faz sentido fazer com que todos os usuários comecem do zero.

A construção de um conjunto de aplicativos de BI estabelece uma estrutura analítica consistente para a organização, logo é necessário evitar que cada planilha conte uma versão ligeiramente distinta da mesma história.

4.3 ANÁLISES AVANÇADAS (ADVANCED ANALYTICS)

Segundo o Gartner, o termo análises avançadas é a exploração de dados automática ou semiautomática utilizando técnicas e ferramentas sofisticadas, além daquelas da inteligência de negócios, BI tradicional, para descobrir *insights* mais profundos, fazer previsões ou gerar recomendações. As técnicas analíticas avançadas incluem: mineração de dados / texto, aprendizado de máquina, correspondência de padrões, previsão, visualização, análise semântica, análise de sentimento, análise de rede e *cluster*, estatística multivariada, análise de gráficos, simulação, processamento de eventos complexos e redes neurais.

Desta forma, com o auxílio das técnicas analíticas avançadas, com múltiplos algoritmos de identificação de padrões e relevância do dado, baseado inclusive em matemática probabilística e estatística, tudo isso dentro de uma ferramenta de análises avançadas, é possível revelar ao analista menos experiente *insights* mais profundos e padrões menos previsíveis, que necessitariam de um cientista de dados mais experiente, assim como também pode auxiliar no trabalho de um cientista de dados sênior na exploração mais eficiente de uma grande gama de dados.

Uma ferramenta de análises avançadas vai além de uma ferramenta de auto atendimento de BI, pois além de gerar visualizações personalizadas pelo usuário acerca dos dados carregados, seja de um DW ou de qualquer outra fonte, ele deve ser capaz de sugerir uma análise que traga alguma informação relevante ou a melhor visualização para uma determinada análise, inclusive, em um conjunto melhor detalhado, ser capaz de fazer predições.

Com isso, as técnicas de análise de dados são divididas em 2 grandes grupos, as análises tradicionais (*traditional analytics)* e as análises avançadas (*advanced analytics)*. A análise tradicional é composta essencialmente de análise descritiva, que responde perguntas sobre o que ocorreu no passado, e análises diagnósticas, que oferece insights sobre o porquê dos fatos ocorridos no passado.

Já a análise avançada é composta por análises preditivas, prescritivas e cognitivas. A análise preditiva está embasada em uma avaliação acerca dos dados atuais e passados para prover insights sobre o que pode ocorre no futuro. A análise prescritiva já implementa o uso de técnicas de dados sofisticadas, como simulação e aprendizado de máquina, m*achine learning*, para sugerir ações visando melhorias de resultados. Por fim, a análise cognitiva utiliza tecnologias de inteligência artificial, *machine learning e deep learning*, para automatizar uma tomada de decisão ou dar suporte complementar a decisões manuais no processo (INTEL, 2015).

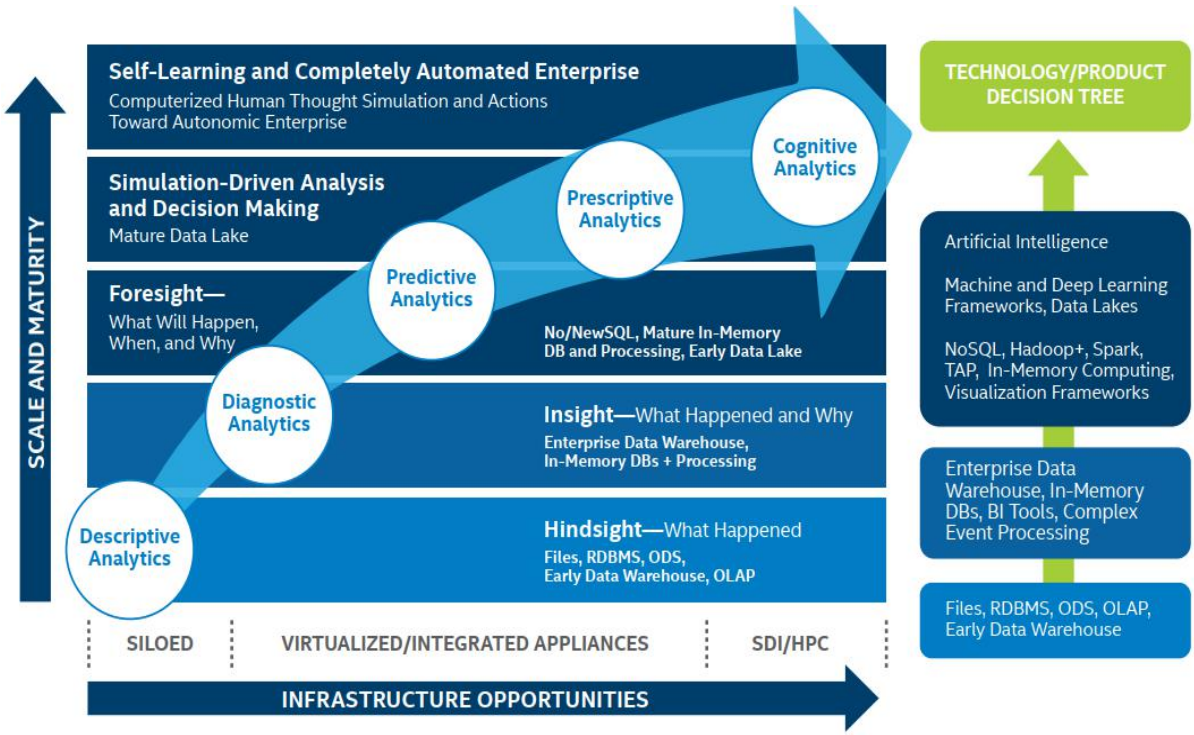

Figura 10 : Amadurecimento do Advanced Analytics

Fonte: INTEL (2015)

Assim sendo, conforme a figura 10, as análises acerca dos dados têm amadurecido e tomado perspectivas mais dinâmicas que acompanham a necessidade de respostas em tempo real exigida nas empresas, com todos os tipos de análise que abrangem os mais diversos tipos de pergunta.

Para atender essas requisições, é necessária uma solução em análise de dados com tecnologias que complementem umas às outras com flexibilidade e capacidade de disponibilizar análises tradicionais e avançadas. Essa solução é dividida em 4 grandes camadas conforme a figura 11.

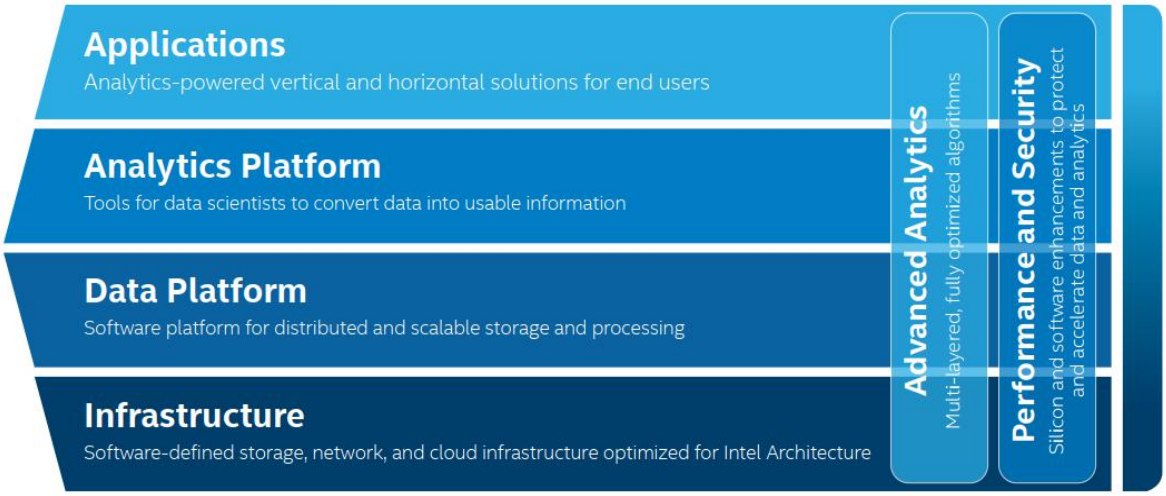

Figura 11 : Solução de Analytics

Fonte: INTEL (2015)

Similar ao conceito de ambiente estruturado para um BI, é necessário ter um cenário adequado à adaptação de análises para o nível de análises avançadas em tempo real. Dessa forma, as 4 camadas principais: Camada de Infraestrutura, Camada de Dados, Camada de Análises e Camada de Aplicações, formam o ambiente ideal para execuções de análises avançadas, acompanhada do gerenciamento de performance e segurança.

A camada de infraestrutura compõe a base tecnológica que permite o armazenamento dos dados com segurança. Ela usualmente alimenta combinações de processos distribuídos de códigos abertos, bancos de dados analíticos não relacionais e aplicativos analíticos.

A camada de dados é o repositório dos dados a serem analisados. No nível de maturidade analítica tradicional, a camada de dados consistia, principalmente, de bancos de dados relacionais. Agora, com a necessidade de armazenar e analisar *streamings*, comum em consumo de aplicações de internet das coisas, e dados não estruturados, a camada é complementada com tecnologias modernas como o Hadoop e NoSQL.

A camada de análise fornece o insumo para o usuário final, as aplicações analíticas. Ela contém múltiplos sistemas conhecidos de análise de dados, como os da Microsoft, IBM, Tableau, entre outros.

A camada de aplicações é a camada superior das soluções analíticas, incluindo as análises prontas para os aplicativos setorizados (INTEL, 2015).

Desse modo, de ponta a ponta do processo de análise de dados, começando desde ambientes transacionais e DWs corporativos até camadas de aplicações de análises avançadas preditivas, prescritivas e cognitivas, reflete-se o amadurecimento contínuo mediante a valorização da inteligência de negócio na tomada de decisão em setores estratégicos das empresas.

Cientistas de dados, desenvolvedores e pesquisadores utilizam ferramentas de IA, como o aprendizado de máquina, *machine learning*, para definir algoritmos e modelos matemáticos que consigam "aprender" a partir dos dados para, então, obter *insights* invisíveis às análises comuns. Soluções para análises de dados facilitam e aceleram o processo de aprendizagem de máquina com multicamadas e algoritmos otimizados que permitem insights mesmo em dados complexos e não estruturados sem uma diretiva explícita.

O *deep learning*, como um ramo do *machine learning,* utiliza redes neurais artificiais que aprendem com experiências massivas e repetitivas, a partir de gráficos profundos com múltiplas camadas para, então, formar modelos que podem inferir *insights* de novos dados. Entre os avanços notáveis nesse ramo estão o reconhecimento de fala e o processamento de linguagem natural (INTEL, 2015).

Por fim, essas abordagens combinadas viabilizam uma aplicação real de análises avançadas, que agregam valor de negócio e transformam, dinamicamente, os setores corporativos.

4.4 SOFTWARE AS A SERVICE SAAS

Nessa perspectiva de avanço na simplificação de utilização de recursos para análise de dados, ampliou-se também a necessidade de integração entre os múltiplos sistemas que participam dos processos, de forma que se tornou imprescindível que o recurso de análise seja um recurso multiplataforma.

Conforme o quadro 3, sobre o crescimento e previsão de crescimento do uso dos mais diversos tipos de serviços e aplicações na nuvem pelo Gartner, é notável a tendência constante de ampliação na utilização da nuvem de forma globalizada.

Quadro 3 : Previsão de receita de serviços na nuvem em bilhões de dólares

	2017	2018	2019	2020	2021
Cloud Business Process Services (BPaaS)	42.2	46.6	50.3	54.1	58.1
Cloud Application Infrastructure Services (PaaS)	11.9	15.2	18.8	23.0	27.7
Cloud Application Services (SaaS)	58.8	72.2	85.1	98.9	113.1
Cloud Management and Security Services	8.7	10.7	12.5	14.4	16.3
Cloud System Infrastructure Services (IaaS)	23.6	31.0	39.5	49.9	63.0
Total Market	**145.3**	**175.8**	**206.2**	**240.3**	**278.3**

Fonte: GARTNER (2019)

Miyachi, em sua publicação na IEE Cloud Computing em 2018, definiu os termos agora bem conhecidos IaaS, PaaS e SaaS:

A Infraestrutura como serviço (IaaS) oferece ao usuário capacidade de processamento, armazenamento, redes e outros recursos fundamentais de computação através de virtualização, onde é possível implantar e executar programas, incluindo sistemas operacionais e aplicativos. O usuário não gerencia ou controla a infraestrutura de nuvem, mas tem controle sobre os sistemas operacionais, o armazenamento e os aplicativos implantados além de controle limitado dos componentes da rede, por exemplo *firewalls* de *hosts*. Esse modelo viabiliza que as empresas utilizem recursos sob demanda, evitando assim de ter de mantê-los fisicamente.

Plataforma como serviço, PaaS, oferece ao usuário capacidade de criar aplicações com linguagens de programação, bibliotecas, serviços e ferramentas suportadas pelo ambiente, o provedor. O usuário não gerencia ou controla a infraestrutura de nuvem, incluindo rede, servidores, sistemas operacionais ou armazenamento, mas tem controle sobre os aplicativos implantados e sobre as configurações para o ambiente da hospedagem dos aplicativos.

Software como serviço, SaaS, oferece ao usuário a capacidade de utilizar aplicativos do provedor com execução em uma infraestrutura direto na nuvem. Os aplicativos são acessíveis a partir de vários dispositivos do cliente como um navegador, por exemplo. O usuário não gerencia ou controla a infraestrutura de nuvem, incluindo rede, servidores, sistemas operacionais, armazenamento, nem mesmo recursos de aplicativos individuais.

Assim, um exemplo de modelo moderno de aplicações e serviço na nuvem encontra-se representado no quadro 4:

Quadro 4 : Um modelo de computação em nuvem baseado no modelo de den Hann

.

6	SaaS		Applications		*End User*
5	App Services	Apps in the Cloud	Communications and Social Media Apps	Data as a Service	*Any User*
4	Built-up PaaS	Business as a Process	Social Media PaaS	Data Analytics	*Rapid Developers*
3.5		Serverless Computing			*Speed Developers*
3	PaaS				*Developers*
2	Foundational PaaS	Application Containers	Routing, Messaging, Orchestration	Object Storage	*DevOps*
1	Software Defined	Virtual Machines	Software Defined Networks (SDN)	Software Defined Storage	*Infrastructure Engineers*
0	Hardware	Servicers	Switches, Routers	Storage	
		Compute	***Communicate***	***Store***	

Fonte: MIYACHI (2018)

Tratando-se de aplicações de análise de dados, o SaaS é o serviço mais comumente utilizado para acesso aos diversos programas multiplataforma de grandes empresas como a IBM, Google e Microsoft.

Isso deve-se ao fato de que aplicações SaaS tipicamente dispensam instalações físicas, que são uma das grandes causas de incompatibilidade, pois, dessa forma, atualizações do serviço seguido da manutenção do funcionamento adequado ficam a cargo somente da máquina principal que hospeda o serviço, sendo necessária somente a manutenção sobre o acesso online do serviço.

A estrutura de pagamento pré-pago ou por usuário é um dos aspectos inestimáveis do modelo SaaS. Para usuários e empresas que compram o software, isso proporciona um investimento inicial menor e evita o pagamento de um pacote fechado para toda a empresa, em vista que apenas uma equipe específica possivelmente irá utilizar.

Precisamente o cenário de maior benefício do SaaS, que é a utilização de um recurso sempre atualizado e de qualquer máquina com acesso à internet, também acaba por ser uma de suas desvantagens. Um recurso 100% online costuma sofrer com instabilidade de conexão, assim como lentidão ao tratar arquivos pesados devido a necessidade de carregar na nuvem tudo o que se for trabalhar, reduzindo então a produtividade.

No entanto a disponibilidade e compatibilidade do serviço prestado em SaaS torna esse modelo altamente atrativo e certas aplicações também dispõem das aplicações em sua versão desktop como alternativa a ausência de conexão ou necessidade de sigilo de informação. Assim, recursos de análise avançada de dados são majoritariamente SaaS e o IBM Watson Analytics, foco nesse estudo, também utiliza essa estrutura.

5 FERRAMENTAS DE ANÁLISE E CONTRIBUIÇÕES DO IBM WATSON ANALYTICS

5.1 INTRODUÇÃO

O cenário atual é de grande volume de dados, que tende a crescer ainda mais. Dessa forma, migram-se cada vez mais aplicações, bancos e infraestruturas para a nuvem a fim de suportar a magnitude dessa expansão na quantidade de dados.

Com essa explosão de dados, necessita-se de ferramentas mais modernas e eficientes na análise de dados. Mas, em conformidade com a quantidade, amplia-se a dificuldade na análise de informação, uma vez que, tradicionalmente, a análise de dados exige um conhecimento avançado em estatística e ciência da computação.

Mediante esse cenário, surgem múltiplas ferramentas de análises de dados que visam realizar análises sobre grandes volumes, de forma intuitiva, para atender desde o usuário com menor conhecimento até aquele de conhecimento avançado. Assim, nesse universo de análise avançada de dados, temos algumas ferramentas que se destacam, como o Microsoft Power BI, o Tableau e, ferramenta chave desse estudo, o IBM Watson Analytics.

Essas ferramentas cresceram ao longo dos anos, captando avaliações dos usuários e otimizando seus processos para entregar uma visão dos dados com nível estratégico de informação até mesmo para usuários sem conhecimento na área.

Esse capítulo visa analisar numa abordagem teórica, a partir de exemplos de manuais de algumas das ferramentas de análises avançadas, em especial o IBMWA, quais as vantagens e limitações na exploração de dados a partir dessas ferramentas.

5.2 ESTADO DA ARTE NO MERCADO DE ANÁLISE DE DADOS

A análise de dados terá um desempenho melhor quanto mais limpos forem os dados, e estruturada e organizada for a sua base de dados de entrada para análise, o que se resume em qualidade do dado mediante ao maior pré-processamento (HOYT, 2016). Nesse sentido, o desafio das ferramentas nesse mercado é o de gerar *insights* e visualizações amigáveis que

entreguem efetivamente informação acerca de um montante cada vez maior de dados, não estruturados e complexos, de maneira intuitiva ao usuário final.

Visando mensurar isso, a cada 1 ou 2 anos, o grupo Gartner monta um Quadrante Mágico, parte de uma série de pesquisas de mercado da empresa, para classificar as ferramentas analíticas e de inteligência de negócio disponíveis no mercado:

Figura 12 : Quadrante mágico para plataformas de Analytics e BI

Fonte: GARTNER (2019)

Conforme a figura 12, o Tableau e o Power BI estão isolados na liderança efetiva nesse segmento e por isso são importantes contribuições ao estado da arte dessa categoria de ferramenta.

O Tableau é o mais antigo dentre os citados na área de visualização de dados. Fundada em 2003, a Tableau entregou uma ferramenta de compartilhamento de visualização de dados equipada com um gama de visualizações interativas a partir de uma intuitiva exploração de dados.

Em nível de negócio, conforme descreve Porter (2005), a empresa pode construir um ambiente sustentável frente a suas concorrentes situando-se com a diferenciação do seu produto ou o custo. Neste caso, a Tableau teve seu foco na diferenciação, construindo uma grande

comunidade para atender as demandas de seus usuários de forma a criar um produto dinâmico que evolui conforme as necessidades emergentes e tem uma implantação flexível.

Seguindo outra frente no conceito de Porter, o Power BI se apresenta com um grande foco em custo, disponibilizando seu serviço integrado ao Office 365, acessando múltiplos serviços de nuvem com a portabilidade conhecida de produtos Microsoft.

Lançado uma década depois da Tableau, em 2013, o Power BI ganhou expressividade no mercado, precisamente por oferecer uma adoção sem riscos para o usuário, ou seja, um investimento inicial baixo para empresas de pequeno porte que podem não possuir necessariamente um BI.

Todavia, é importante ressaltar que isso está limitado frente ao serviço SaaS oferecido para esse caso, e, como citado dentre as desvantagens desse modelo, isso torna complexo trabalhar com um volume de dados muito grande. Para atender setores maiores, outras funcionalidades são necessárias o que, efetivamente, muda o custo dessa implementação. Dessa forma, é comum que empresas de grande porte que efetivamente necessitam de uma margem segura de consultoria e implantação do serviço de maneira integrada a um *data warehouse* previamente existente acabam optando por ferramentas com uma infraestrutura mais robusta.

Assim, existem múltiplas ferramentas que visam auxiliar na visualização de dados, conforme evidenciado na figura 12. Dentre essas ferramentas, um quantitativo menor é capaz de efetivamente explorar dados, funcionando como uma ferramenta inteligente de análises avançadas. Logo, o Quadrante Mágico funciona como um balizador, pois ele não analisa somente a aplicação em si, mas também o conjunto da solução, envolvendo empresas e parceiras, no qual o principal fator não reside unicamente no nível técnico e funcionalidades da ferramenta, mas, por muitas vezes, nas regras de negócio de cada empresa que podem melhor se adequar a cada modelo oferecido.

O relatório da Frost & Sullivan (2018), *Latin American Big Data and Analytics* , BDA, Market, aponta uma projeção de receita de US$ 8.5 bilhões até 2023 para o mercado, alinhado ao crescimento constante de adaptação à nuvem que acompanha esse crescimento. Segundo a IDC Brasil (2019), o segmento de nuvem pública no Brasil pode chegar a US$ 2,6 bilhões em 2019, crescendo 35% anualmente até atingir US$ 6,5 bilhões em 2022.

Assim, o mercado de inteligência de negócio e análises avançadas continuará a ganhar cada vez mais espaço e essas aplicações continuarão a se desenvolver e integrar cada vez mais funcionalidades a fim de atender a demanda crescente de empresas migrando seus serviços para a nuvem a medida que continua crescendo a quantidade de dados a serem analisados para melhor tomada de decisão.

Quadro 5 : Resumo do cenário evolutivo da exploração de dados

DÉCADAS DE 80 e 90	DÉCADA DE 2000	DÉCADA DE 2010	PRÓXIMOS PASSOS
BI Tradicional	**self-service BI**	**Advanced Analytics**	**“Auto Analytics”**
Relatórios e Dashboards estáticos com alterações solicitadas pelo usuário e realizadas por técnicos de TI	**Dashboards e Scorecards dinâmicos atualizados sem necessidade de interferência técnica**	**Ferramentas de exploração de dados integradas com aplicação de IA, Machine Learning e Deep Learning**	**Ambiente analítico integrado com todas as aplicações executando como uma mesma interface**
Responde às perguntas: “O que aconteceu?” e “Por que isso aconteceu?”	**Responde à pergunta: “O que está acontecendo?”**	**Responde à pergunta: “O que pode acontecer?”**	**Recomenda análise de cenários relevantes através de relatórios automáticos**
Tem foco em gerar relatórios com dados históricos para entender o passado	**Tem foco em monitorar e gerar visualizações para entender o presente com atualizações constantes**	**Tem foco em gerar insights acerca do conjunto de dados de entrada, mostrando relacionamentos e pontos relevantes**	**Tem foco em explorar um fluxo constante de entrada de dados e levantar insights relevantes de cenários a serem analisados**
A nível de negócio, TI era visto como commodity, ou seja, apenas infraestrutura tecnológica essencial a análises longas	**A nível de negócio, o BI passou a integrar como diferencial competitivo para tomada de decisões em menos tempo**	**A nível de negócio, a análise de dados se consolida como um grande conjunto de práticas essenciais a toda empresa de médio e grande porte e passa a ser considerado na estratégia de pequenas empresas entrantes. A empresa que não realiza análise sobre seus dados tem um enorme déficit competitivo**	**A nível de negócio, a análise exploratória de dados torna-se meio fundamental, com baixo custo e baixa complexidade, de apoio a tomada de decisão gerencial e evidencia uma relação direta entre investimento em gestão e exploração de dados e retorno financeiro das melhores decisões**
Altamente Adhoc	**mais bem integrado, mas ainda muito Adhoc**	**Com as aplicações em IA, ML e DL, torna-se menos Adhoc**	**O ambiente analítico é integrado ao ambiente de negócio como um todo e não mais a um cenário específico. Deixa de ser Adhoc para abrir caminhos a automação.**

Conforme o quadro 5, a exploração de dados para tomada de decisão evoluiu ao longo dos anos de maneira a atender necessidades de negócio, que efetivamente impulsionam a evolução dessa área, principalmente em pontos como velocidade de atendimento, capacidade analítica e integração.

Na coluna “próximos passos”, situo minha análise do que seriam os próximos passos necessários para generalizar a exploração de dados, desconectando-se do atual cenário ainda muito *ad hoc* e migrando para um cenário mais abrangente no qual a exploração de dados é alinhada às técnicas avançadas de inteligência artificial, aprendizado de máquina e aprendizagem profunda, para alcançar a maior capacidade de automação possível, fundamental ao crescimento continuo do quantitativo de dados.

Nesse cenário, o IBM Watson Analytics, sendo uma ferramenta da IBM, atualmente referência em IA, possui potencial de integração com diversas aplicações com intuito de consolidar esse ambiente analítico em uma mesma interface. Por isso foi a escolha de ferramenta foco desse trabalho devido ao seu potencial de consolidar-se junto às técnicas que efetivamente definem as análises como avançadas.

5.3 IBM WATSON ANALYTICS

5.3.1 SPSS

Em 2015, a International Business Machines Corporation, IBM, lançou o IBM Watson Analytics, IBMWA, que adentrava o mercado de ferramentas de análises avançadas tendo como base o Pacote Estatístico para Ciências Sociais ou SPSS (IBM, 2018).

Conforme apresentado no quadro 6 acerca das principais características das funcionalidades de análise do SPSS, o software base de exploração do Watson Analytics reúne um grande compilado de aplicações de mineração, estatística e demais componentes de análise para exploração de dados que visam auxiliar na tomada de decisão, tendo sua primeira versão em 1968 e adquirido pela IBM em 2009 (IBM, 2018).

Os recursos do SPSS são um compilado históricos de múltiplas funções de pacotes de ferramentas e funções estatísticas que foram compiladas em um único *software*, no qual é possível realizar regressões, tabelas e criar relacionamentos entre variáveis dependentes e independentes para análise histórica de eventos e dados de duração.

Quadro 6 : SPSS principais recursos de análise

Estatísticas descritivas	Relatórios
	Livro de códigos
	Frequências
	Descritivos
	Explorar
	Crosstabs
	Estatísticas descritivas da relação
Comparar meios	Medida de análise com Eta e Eta2
	Teste de linearidade com R e R2
	Estatísticas independentes da amostra
	Estatísticas de amostra emparelhadas
	Estatísticas: intervalos de confiança
	Teste t de uma amostra
ANOVA unidirecional	Contrastes: linear, quadrático, cúbico, de ordem superior
	Testes de alcance
	Testes Post hoc
	Estatísticas ANOVA
	Teste de homogeneidade de variância
	Estatísticas descritivas do grupo
Correlação e distâncias	Bivariado
	Parcial
	Calcular proximidades entre casos ou variáveis
	Medidas de dissimilaridade e semelhança
Outros	Modelagem linear automática (MLA)
	Regressão - regressão linear
	Regressão ordinal - PLUM
	Modelagem de simulação
	Análise geoespacial
	Estimar Curvas
	Testes não paramétricos
	Resposta múltiplas
	Redução de dados
	Classificar
	Cluster
	Cluster rápido
	Análise de vizinhos mais próximos
	Discriminante
	Dimensionamento
	Operações matriciais

Fonte: Baseado em SPSS Features em IBM (2018)

Dessa forma, o SPSS surge como um integrado de algoritmos estatísticos e depois passou a agregar técnicas de mineração de dados. Posteriormente, o SPSS passa a integrar o Watson Analytics, aliando-se a métodos de inteligência artificial como aprendizado de máquina, análises de dados não estruturados e análise preditiva. A figura 13 apresenta um resumo da evolução do SPSS ao longo das últimas décadas.

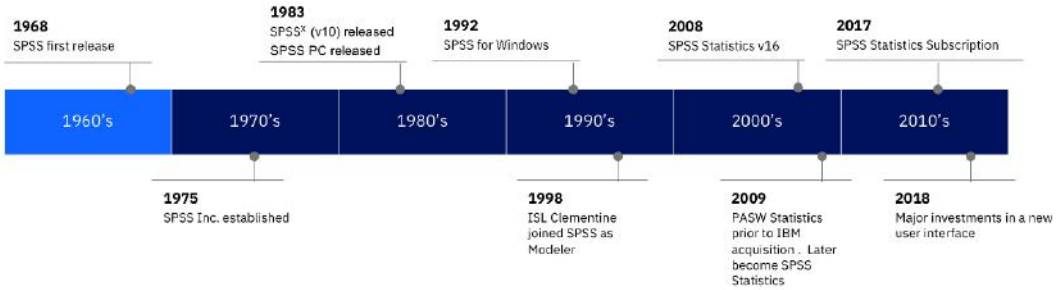

Figura 13 : 50 anos de SPSS

Fonte: STAUBER (2018)

5.3.2 Funcionalidades e características do IBMWA

O IBMWA pode ser definido como um SaaS, s*oftware as a service*, com capacidade cognitiva, cujo foco está na aceleração de *insights* através de linguagem natural, na automação de análises preditivas e na criação de infográficos detalhados com mínimo esforço (IBM, 2018).

De uma maneira geral, o Watson funciona em um modelo *self service* que guia o usuário pelo processo de descoberta de *insights* dentro do seu conjunto de dados, permitindo uma interação direta com essa base, como uma "conversa", em linguagem natural, na qual você pode realizar algumas perguntas ao seu dado, a partir de alguns termos chave, e extrair respostas de informações estruturadas e não estruturadas com maior facilidade.

Além disso, após definir um escopo, o IBMWA pode auxiliar a identificar os fatores que mais influenciam os resultados dessa consulta e apresentar isso de maneira visual e dinâmica com diversos modelos, dashboards ou infográficos, sugeridos de forma automática a melhor realçar essa visualização.

Conforme a figura 14, o Watson Analytics é associado a um determinado *dataset* que pode ser continuamente alimentado de maneira a enriquecer os retornos oferecidos pela ferramenta, nas dimensões como usuários ou visão geográfica por exemplo.

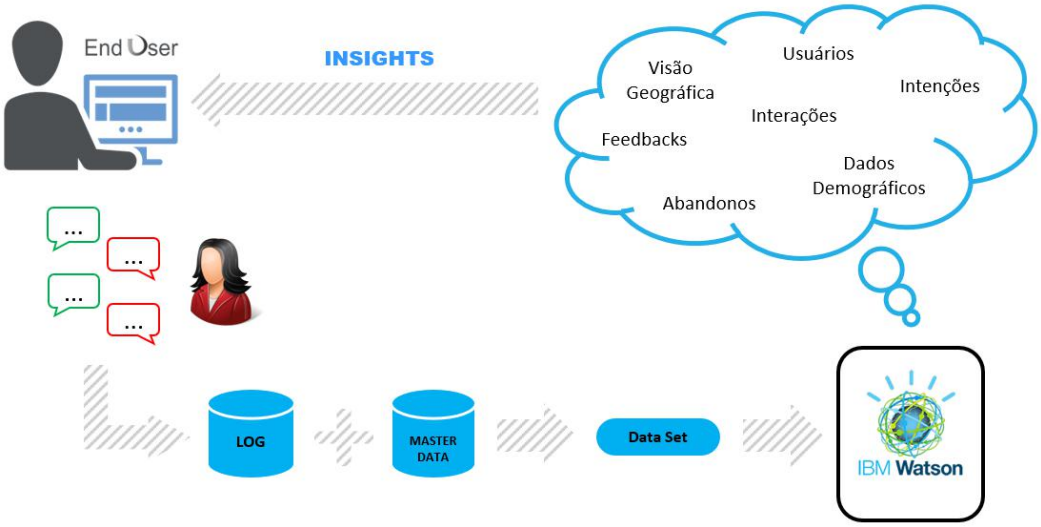

Figura 14 : Ciclo do Watson Analytics

Fonte: IBM (2018)

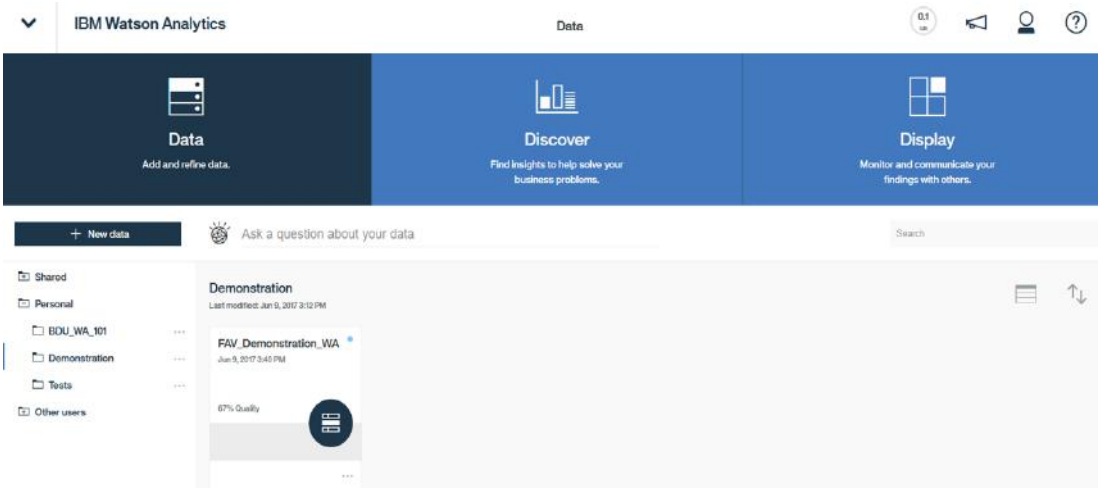

Figura 15 : Watson Analytics tela inicial

A ferramenta possui três funcionalidades abrangentes principais, “Data”, “Discover” e “Display”, disponíveis na tela inicial da aplicação que compõem o ciclo de vida da exploração, que se resume em carregar e refinar os dados, explorar os dados e apresentar os dados conforme mostra a figura 15.

5.3.2.1. Entrada de dados IBMWA

A seção "Data" é o ponto de partida lógico para examinar qualquer conjunto de dados (JMIR Public Health Surveill, 2016). Como ponto de partida, deve-se adicionar a fonte de dados que se deseja trabalhar, que se resume em três tipos: importar, conexão e arquivo local.

As fontes externas são atualizadas frequentemente com intuito de ampliar a gama de opções de importação. A figura 16 mostra os diferentes fontes possíveis de importação.

Figura 16 : Watson Analytics fontes de importação de dados

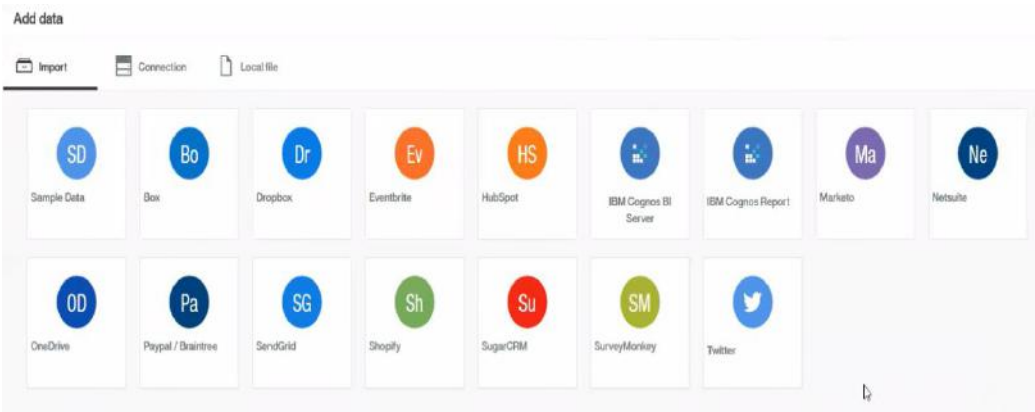

Figura 17 : Watson Analytics fontes de conexão

A aba de conexão possibilita a conexão com bancos de dados previamente configurados no login de conta do usuário na aba de criar conexões de dados. Assim como o *import*, o IBMWA possui diversas fontes possíveis de BDs, conforme a figura 17.

Por fim, a aba de arquivo local possibilita o carregamento de planilhas e demais arquivos direto da máquina local de acesso para a nuvem do IBMWA.

Vale ressaltar que apesar de múltiplas possibilidades de entrada de tipos de arquivo, o IBMWA trabalha com o modelo tabular.

Uma vez definido o conjunto de dados a ser carregado para o IBMWA, ele não somente será carregado como também já será processado, qualificando as colunas, atribuindo uma nota de qualidade para cada coluna e depois atribuindo uma nota de qualidade para o conjunto completo. Essa nota de qualidade tem como objetivo dar um parâmetro de quão acurada pode ser a análise realizada em cima desses dados pelo próprio IBMWA.

Esse valor está relacionado a possíveis valores faltantes nas colunas, a distribuição de valores dentro da coluna, linha de títulos bem definido, limpeza do dado e o quanto é possível relacionar essa coluna com demais campos, mas, no final, sem um detalhamento dos critérios.

O IBMWA classifica alguns campos como o *driver* de uma predição, ou seja, que possui um bom relacionamento com os demais campos e pode calcular outras informações a partir dele.

Antes de começar a exploração desse dado, é recomendado entrar no modo de refinar o dado, onde é possível renomear atributos, incorporar cálculos, e colocar os dados em grupos ou hierarquias para análise de subgrupos. Além disso, é possível visualizar a qualidade do dado atribuída a cada coluna e a distribuição de dados na coluna.

A distribuição de uma variável é o padrão apresentado por um conjunto de dados resultantes da observação da variabilidade de uma variável, e o padrão desta variabilidade é a distribuição da variável (ROSSMAN, 2001).

Terminado essa etapa, deixando o conjunto de dados o mais otimizado possível, como, por exemplo, eliminando duplicatas e registros nulos que podem reduzir a qualidade do dado, pode-se iniciar a etapa de descoberta.

5.3.2.2. Descoberta e Apresentação de Dados IBMWA

Ao iniciar a etapa de exploração de um determinado conjunto de dados, o IBMWA, que já processou previamente esse conjunto, já é capaz de sugerir alguns pontos de partida para análise, classificados como *insights* de pontos de partida, realizado pela IA presente na própria ferramenta.

Nesse ponto é relevante observar que ele não somente propõe como entrada uma sugestão de *insight* acerca do dado que o IBMWA considerou relevante ao processo, mas também sugere um modelo de visualização que melhor se adequa à apresentação desse ponto de partida, conforme mostra a figura 18.

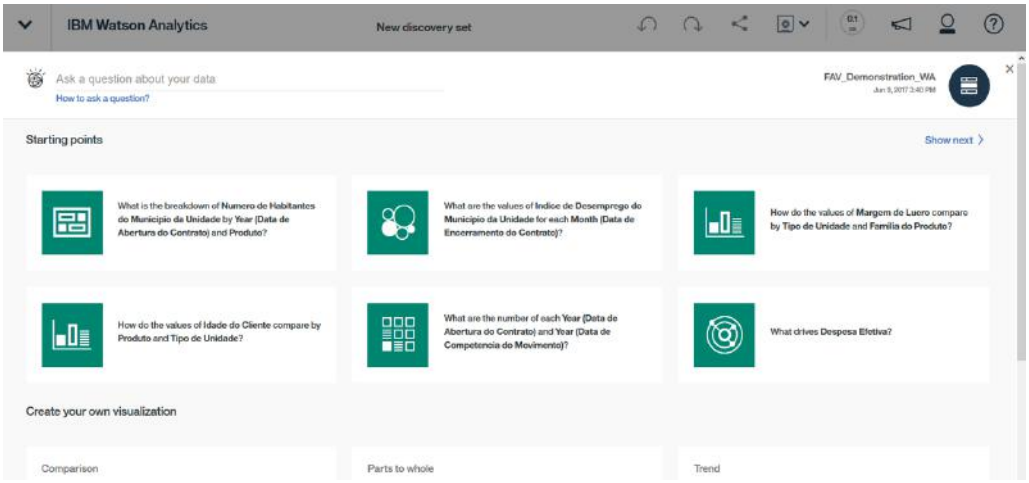

Figura 18 : Watson Analytics pontos de início

O ponto, todavia, mais relevante do IBMWA é o recurso de pergunta em linguagem natural, disponível em inglês e espanhol, que tenta simular uma conversa com o *dataset*.

Nesse caso, o conhecimento do *dataset* é fundamental, pois conforme a figura 18, com os exemplos gerados automaticamente, é possível usar palavras chave como *compare*, *values* e *drives* para relacionar os campos do seu escopo de dados a fim de gerar informações expressivas e visuais anteriormente ocultas.

O IBMWA gera novos pontos de partida relacionados à pergunta em linguagem natural, de maneira a encadear os *insights* do IBMWA com os seus *insights* para então chegar mais longe do que se poderia ir sozinho (IBM, 2018).

Esse é um dos fatores de grande destaque da ferramenta, pois ela torna o processo de navegar dentro dos dados muito mais intuitivo e vai encadeando as análises sugerindo novas análises e tendências.

Como um comparativo, temos que a conversa com o dado nas demais ferramentas como o Power BI, por exemplo, é um recurso a mais sem maiores desenvolvimentos, conforme mostra a figura 19, na qual a janela de conversação com o dado apresentado à direita possui a interface simplificada, com apenas sugestões de campos a serem visualizados mas sem qualquer prévia dessa visualização ou sugestão de visualizações.

Nesse sentido, o IBM Watson Analytics, que veio originalmente de uma IA voltada precisamente para o diálogo com o usuário e sugestão de recomendações para os mais diversos fins através do supercomputador Watson, traz o recurso de navegação através das perguntas como um ponto focal que facilita ao usuário realizar descobertas dentro dos dados em análise.

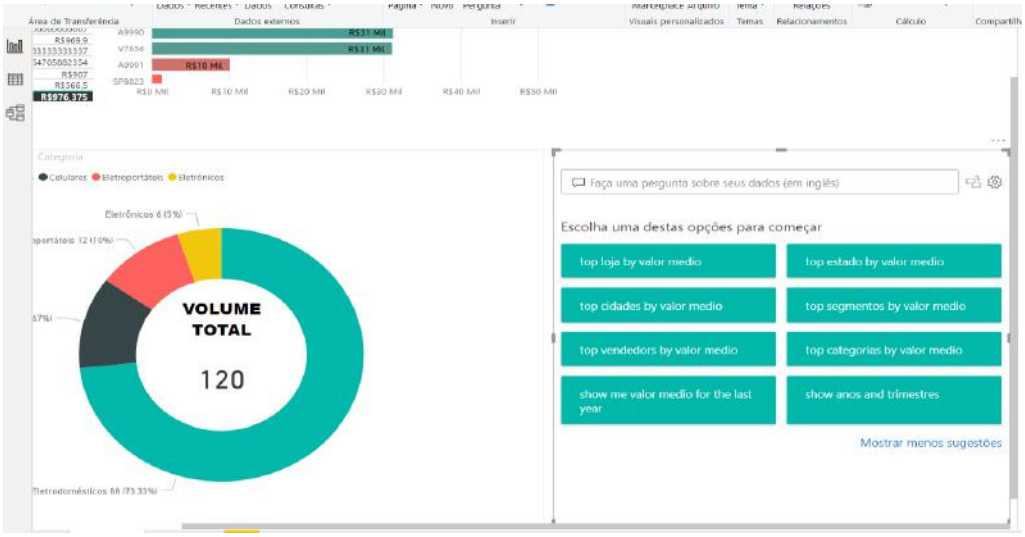

Figura 19 : Power BI recurso de conversa com o dataset

Dessa forma, o IBMWA se auto incrementa com novas soluções a partir das capacidades de outros recursos incorporados ao seu funcionamento, conforme citado no quadro 5 no conceito de ambiente analítico.

Neste capítulo teórico, com intuito apenas de comentar as funcionalidades do IBMWA, foi utilizado um dataset fictício com viés bancário, no qual não se levará em consideração os resultados de informação sobre os dados em si.

Na figura 20, após a escolha de uma das recomendações, o IBMWA gera a visualização recomendada e incrementa com sugestões de descobertas relacionadas a visualização. Neste exemplo, utilizando apenas como referência um *dataset* fictício com viés bancário, analisou-se a margem de lucro por cada gerente e segmento com dados fictícios de segmentos de ouro, platina etc.

O modelo escolhido traz consigo os recursos disponíveis de visualização que são os vetores desse gráfico. Neste caso, gerente é a quantidade de bolhas, segmento é a cor das bolhas e margem de lucro é o tamanho das bolhas. Além disso ainda seria possível utilizar o slot de mapa de calor e multiplicador.

Com essa visualização já é possível ter alguns *insights* acerca desse conjunto de dados sem utilizar funções de quantificar, somar ou qualquer outro elemento de agregação comum das análises tradicionais.

Alguns exemplos fictícios são:

- Poucos gerentes comercializam platina, representada pela cor amarela, todavia possuem uma margem de lucro maior que diversos outros gerentes de outros setores.

- Há muitos gerentes comercializando prata com uma margem de lucro pequena.

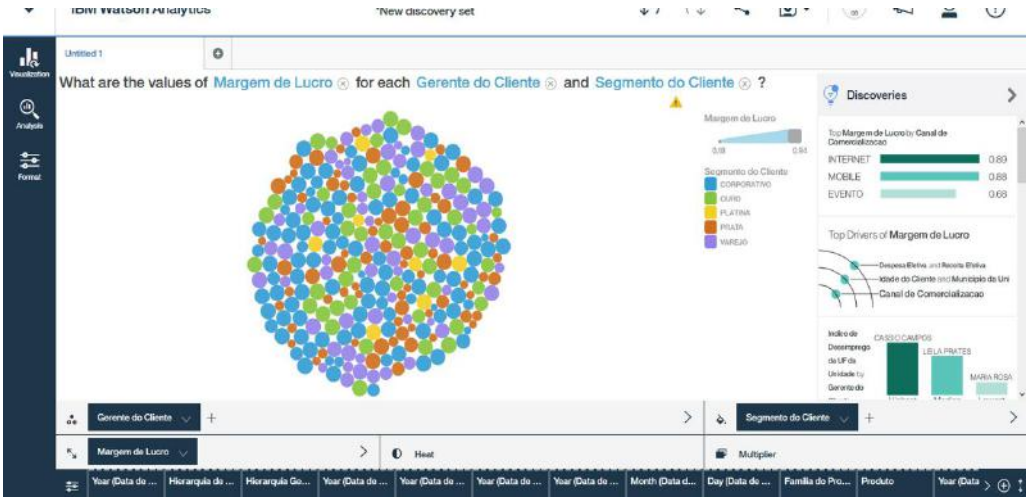

Figura 20 : Watson Analytics modelo de visualização em bolhas

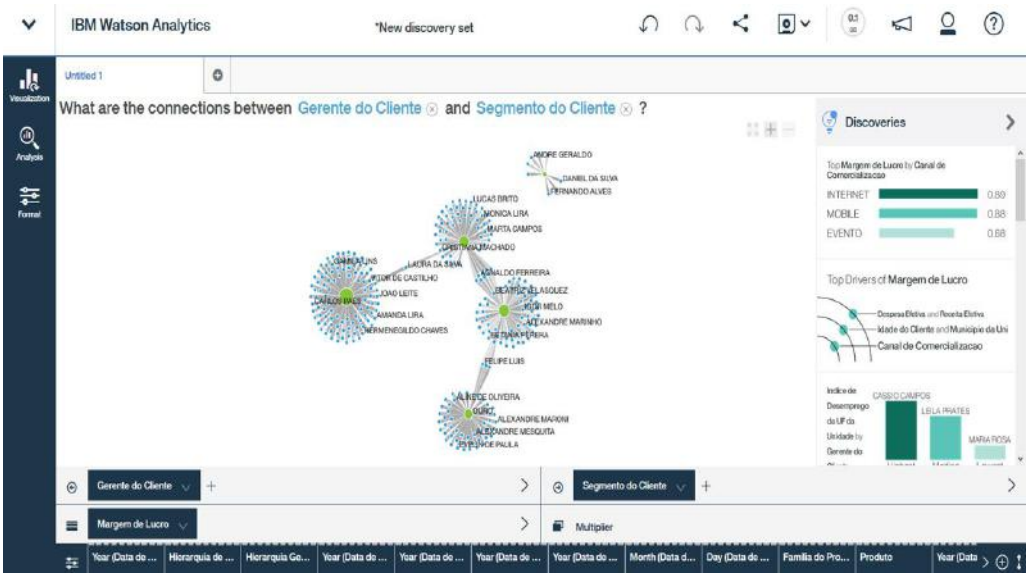

Figura 21 : Watson Analytics modelo de visualização de conexões

Com esse ponto de partida inicial, já é possível formular novas perguntas e considerações inclusive acerca do mesmo conjunto já exibido e, portanto, vale modificar a forma de visualização apresentada para tirar outras conclusões mais estratégicas.

Na aba *visualization*, é possível verificar outras visualizações recomendadas pelo próprio IBMWA para essa relação específica, assim como todas as demais visualizações possíveis. Dentre as recomendações, a figura 21 representa a escolha do modelo de *network*.

Neste novo modelo, cada bola central é um segmento e quanto maior ela for maior o número de gerentes comercializando esse segmento e quanto mais denso o traço que liga o gerente ao segmento, maior é a margem de lucro.

Agora, novos insights são possíveis e complementam os apresentados anteriormente e enriquecem com novas informações fictícias visuais, tais como:

- Os gerentes que comercializam Platina, um dos segmentos menos comercializados, não comercializam nenhum outro segmento, logo são especializados nesse setor.
- É possível visualizar que a gerente Laura da Silva é a única com conhecimento dos dois segmentos com maior quantitativo de comercializações, logo ela poderia ser um ponto focal de compartilhamento de conhecimento ou ponto estratégico de administração simultânea das duas áreas.

5.3.2.3. Modelo preditivo de dados históricos IBMWA

O IBMWA possui internamente o pacote SPSS, que ele utiliza para montar modelos preditivos sobre dados históricos que é capaz de responder algumas perguntas tais como: O que influencia tal métrica? Qual o valor previsto dado uma determinada configuração?

Dito isto, o IBMWA não se enquadra como uma ferramenta de predição de dados, pois ele não monta um modelo que poderia ser exportado e aplicado a um outro *dataset* que não possui esses mesmos registro históricos, ou seja, ele não realiza o *scoring* dos dados.

O *Scoring* é tratado como uma transformação dos dados, definido como um conjunto de transformações numéricas a serem aplicadas a um determinado conjunto de variáveis - os preditores especificados no modelo - para obter um determinado resultado. Nesse sentido, o processo de *Scoring* de um determinado modelo é inerentemente o mesmo que a aplicação de qualquer função, como uma função de raiz quadrada, a um conjunto de dados (IBM, 2018).

Então, o IBMWA é capaz de gerar *insights* de predições acerca do conjunto que se está trabalhando de maneira que um cientista de dados poderia efetivamente gerar as predições com os modelos fornecidos. Já para um usuário comum, é possível enxergar de forma mais clara qual o conjunto de influenciadores, *drivers* do modelo de dados, que possuem mais força de predição para uma determinada métrica.

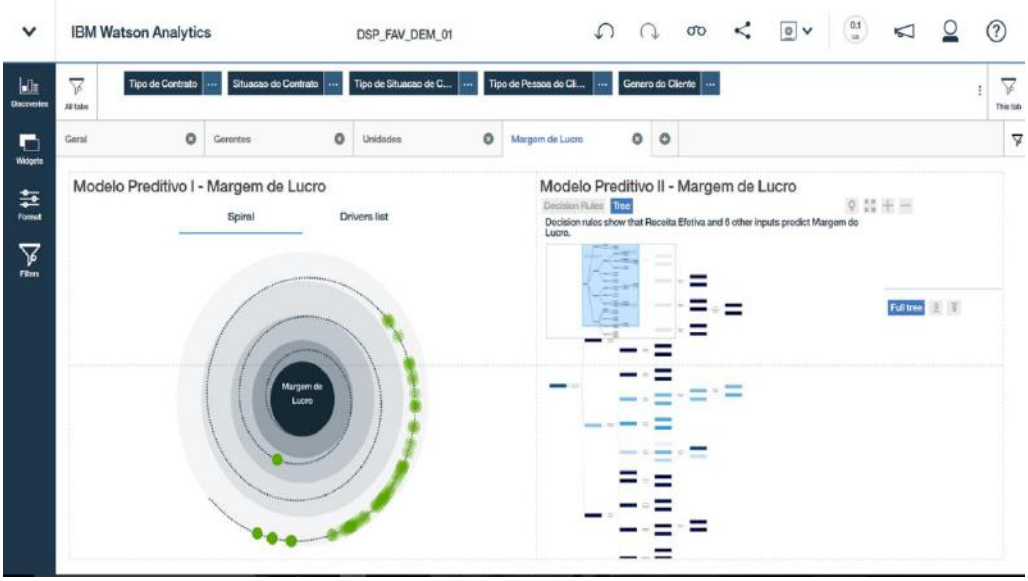

Figura 22 : Watson Analytics visualização dos Drivers

Na figura 22, o IBMWA mostra qual é o *driver*, ou seja, o fator que mais influencia determinada métrica, da margem de lucro. O modelo escolhido pela ferramenta é o espiral, patenteado pela IBM, no qual quanto mais próximo do centro da espiral maior é o poder de influência sobre esse fator. Dessa maneira, temos de forma visual e numérica que o canal de comercialização é o fator que mais influencia a margem de lucro.

Além disso, à direita, na figura 22, é apresentado o modelo de árvore de decisão que você pode navegar entre as previsões de lucro em cima desse modelo histórico específico.

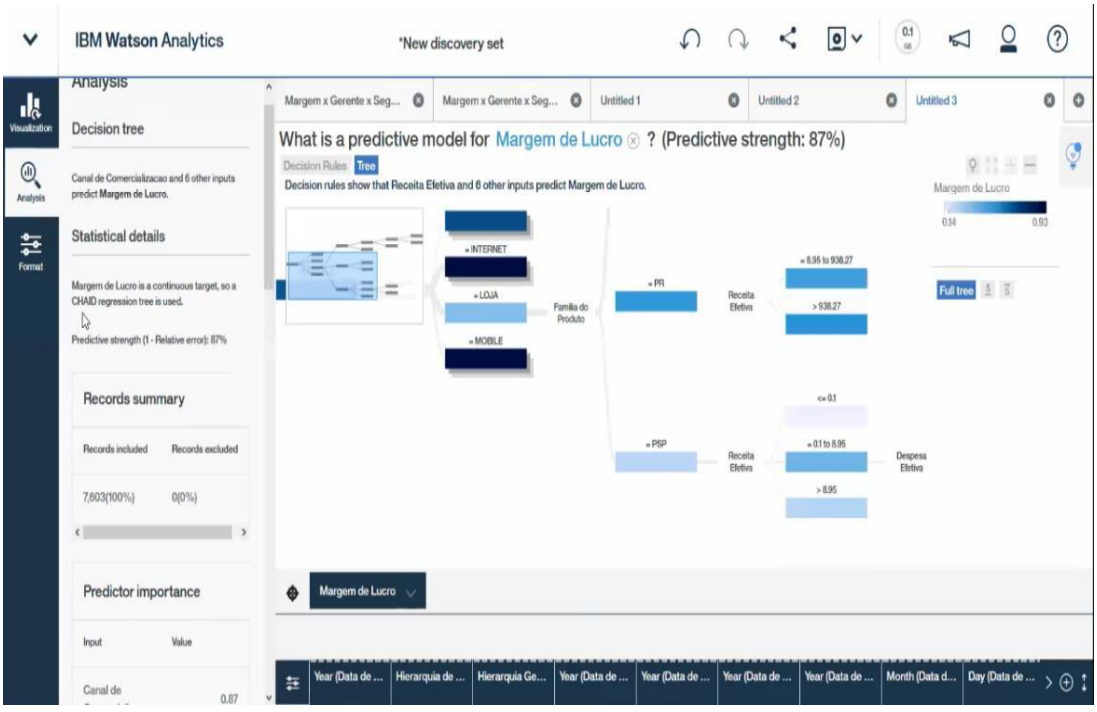

Figura 23 : Watson Analytics visualização dos Drivers

Dentro de todos esses percursos, navegando entre os *insights*, o IBMWA não fornece um detalhamento de passos e processos utilizados para ele chegar nesse resultado.

Ele possui para alguns insights específicos, por exemplo para aqueles que utilizaram o modelo preditivo histórico, algumas métricas utilizadas pelo software para chegar nesses resultados. Todavia, esse escopo é pouco transparente acerca de quais recursos estatísticos que foram necessários para afirmar um insight específico.

A figura 23 apresenta esse caso, em que o IBMWA utilizou uma árvore de regressão Chaid junto com outros elementos para montar a árvore de decisão. Todavia, não é possível exportar nenhum modelo do IBMWA para uso em outras aplicações ou mesclar regras.

Essa é uma limitação recorrente das ferramentas de análise avançada, uma vez que elas não disponibilizam opções para utilizar os resultados das análises. É necessário, então, exportar todo o conjunto e recomeçar as análises com os dados em uma ferramenta complementar para depois retornar com o conjunto de volta para a ferramenta.

6 APLICAÇÃO DO IBM COGNOS ANALYTICS SOBRE DADOS ABERTOS

6.1 INTRODUÇÃO

O IBM Watson Analytics seguiu a tendência principal descrita no quadro 5 deste trabalho, em que os próximos passos de ferramentas de análise são se integrarem para formarem verdadeiros ambientes analíticos.

Dessa forma, ele foi descontinuado para fazer parte de outra ferramenta de mercado da IBM que é o IBM Cognos Analytics. Uma vez que dentro do Cognos Analytics foram mantidas as capacidades e propriedades exibidas neste trabalho, então foi possível realizar as mesmas funções aqui relatadas sem qualquer prejuízo em cima dos dados abertos.

Nesta seção, utilizamos as funcionalidades do Cognos Analytics sobre um conjunto de dados abertos já estruturados sobre trabalho escravo no Brasil. Assim, uniu-se a avaliação tecnológica e conceitual do processo de exploração de informações relevantes dos dados com uma pequena contribuição para ampliar estudos em um domínio de tanta relevância como o do trabalho escravo, de modo que, através das facilidades do IBMWA utilizadas neste trabalho, não sejam necessários conhecimentos mais avançados dentro da área de ciência de dados.

Desse modo, exploramos essas funcionalidades para tentar levantar informações relevantes acerca dos dados e comentar sobre os diferenciais do IBM Cognos Analytics e sobre as capacidades e limitações no cenário geral de análises avançadas, em uma abordagem mais prática, com exemplos de uso criados e adaptados ao escopo de dados.

6.2 ESCOPO

Conforme descrito no item 1.3 deste trabalho, utilizamos dados das listas sujas do trabalho escravo contemporâneo, publicadas desde 2003, incluindo as operações que identificaram trabalho escravo, e dados eleitorais publicados desde 1998 contendo doações exclusivamente diretas para candidatos, assim como os cargos envolvidos de 2002 a 2016 (SECRETARIA DE TRABALHO,2019).

Com isso, a ferramenta IBM Cognos Analytics foi aplicada a esse conjunto de dados abertos já tratados sobre o trabalho escravo para obter informações e *insights* relevantes com recursos visuais e de IA sobre dados de trabalho escravo que necessitam ainda de maior

visibilidade e estudo no Brasil, para enfim abordar os conceitos e mecanismos tratados até este ponto.

6.3 ANÁLISES EXPLORATÓRIAS SOBRE OS DADOS ABERTOS

Os arquivos base referentes à lista de envolvidos e operações realizadas, além de doações recebidas por candidato, foram disponibilizados em formato CSV e importados para dentro do ambiente analítico do SaaS IBM Cognos Analytics.

O Cognos Analytics, semelhante a outras plataformas citadas neste trabalho, interpreta os dados importados como fontes externas semelhantes a tabelas de dados em que podemos criar relacionamentos.

Dessa forma, a fim de correlacionar todos os dados importados, os relacionamentos entre eles foram realizados através dos campos chave comuns às fontes semelhantes à estrutura do modelo estrela utilizado em *data warehouse*, mas desvinculado do conceito geral de fatos e dimensões.

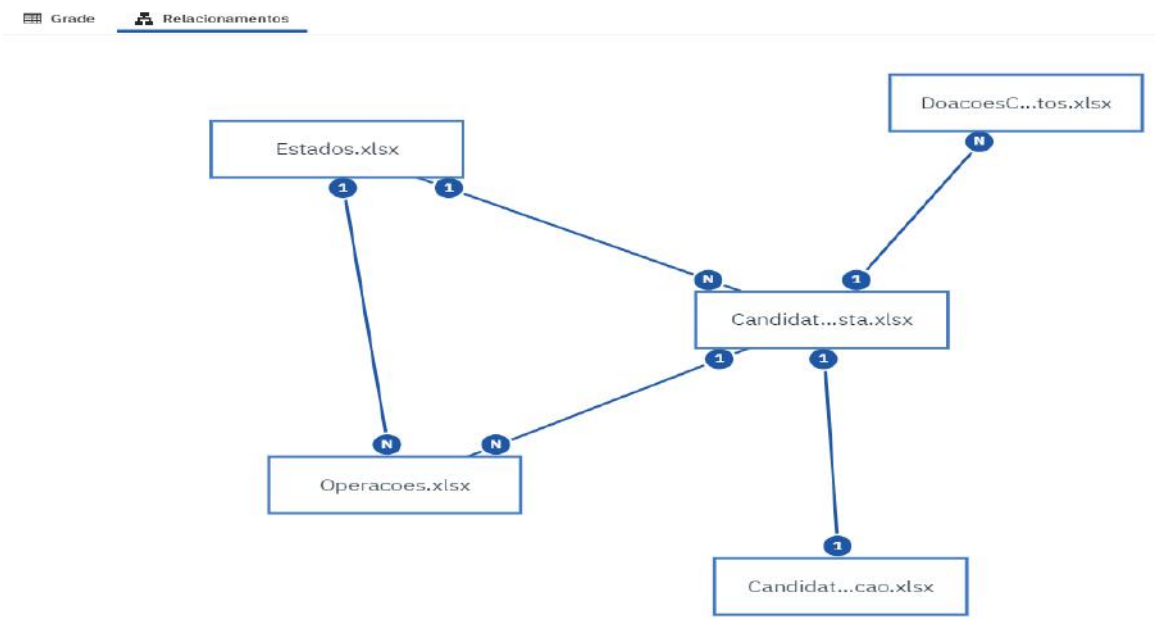

Figura 24 : Cognos Analytics relacionamentos dos fontes

A figura 24 mostra o relacionamento dos CSVs fontes, que são: Candidatos sujos lista, centralizado, que é a lista com dados cadastrais, CPF, Data de Nascimento etc. dos candidatos eleitos entre 1998 e 2016 relacionados; Candidatos sujos eleição, mais inferior, com a lista dos candidatos contidos na lista suja de trabalho escravo; Estados, com a lista dos estados do Brasil;

Operações, com a lista das operações contra o trabalho escravo de fiscalização com CPF dos envolvidos e a quantidade de trabalhadores envolvidos; Doações candidatos, mais superior, com a relação de doações recebidas por candidatos entre 2002 e 2016.

Como citado, existe a centralização do arquivo fonte relacionado a lista de candidatos envolvidos, uma vez que este contém a chave primária única, CPF, compartilhada pela maioria das tabelas, e possui uma única linha por CPF, evitando duplicidade de informação.

A partir deste ponto, as análises foram criadas navegando do menor ao maior grau de complexidade de aprofundamento dentro dos dados, realizando uma navegação através dos *insights* sugeridos pela ferramenta alinhado a *insights* próprios que visavam desde uma visualização mais clara de um determinado fator, até análises de projeção de valores futuros com auxílio de recursos de IA presentes na ferramenta.

6.3.1 Análise dos cargos envolvidos na lista suja

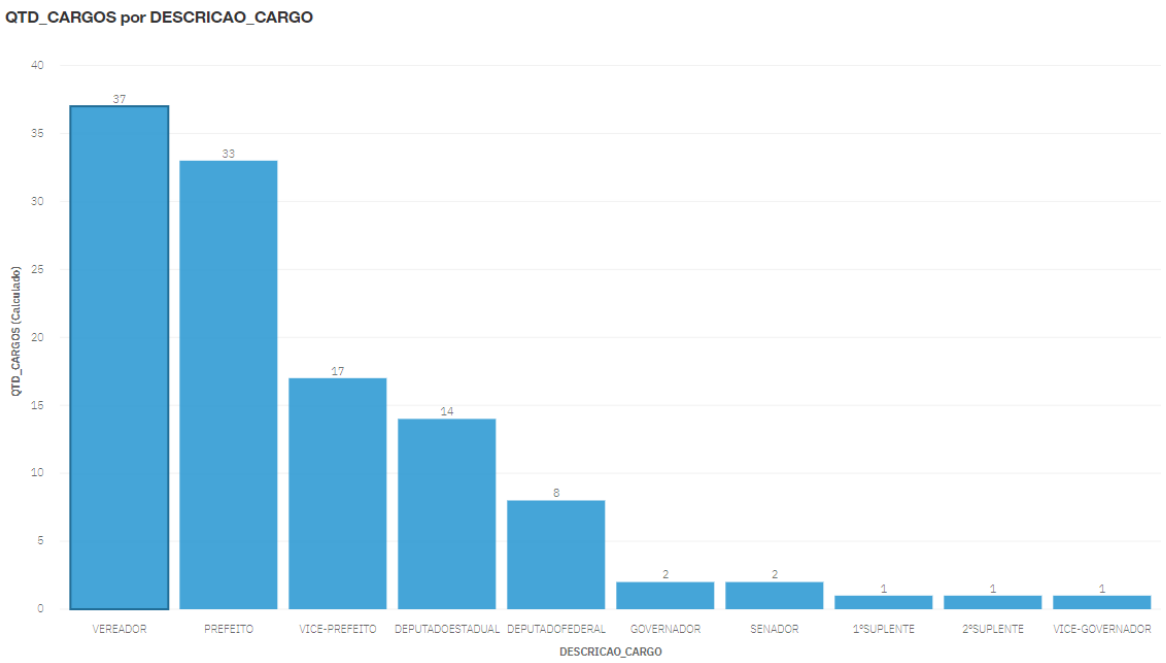

Figura 25 : Cognos Analytics cargos envolvidos

Na análise representada na figura 25, foi possível observar de forma descendente os principais cargos envolvidos na lista suja de candidatos. Para isso, primeiramente, foi criado no modelo um *calculation*, que corresponderia a uma métrica do BI.

A métrica criada corresponde a um agrupamento dos cargos envolvidos, disponibilizando para o Cognos Analytics um valor numérico que ele pudesse correlacionar com a tabela de cargos para montar a visão correspondente.

Com isso, observa-se uma informação relevante, pois apesar da quantidade de vereadores ser significativamente maior, em termos de quantidade de cargos nacionais, que o número de prefeitos, ainda assim o cargo de prefeito ficou muito próximo do de vereadores.

Outro ponto importante nessa análise é que se mostrou ainda mais fundamental a questão do tratamento prévio dos dados. Nesse modelo de dados, havia dados parcialmente duplicados, ou seja, não era a linha inteira que era duplicada, mas se tratava do mesmo candidato, pois possuía o mesmo CPF.

Dessa forma, o CPF não funciona como uma chave única, afinal esse modelo de dados não compõe verdadeiramente uma tabela física, mas apenas aparenta uma tabela nos moldes de uso dentro da ferramenta. Assim, as primeiras análises deram resultados divergentes do resultado apresentado na figura 25 e foi necessário realizar previamente um tratamento de limpeza de duplicidade baseado em CPFs distintos para então garantir a unicidade.

Esse caso não é exclusivo do IBM Cognos Analytics, e se mostrou presente em outras ferramentas analíticas, como o Power BI, por exemplo, o que se mostra como uma limitação intrínseca desses processos em serem altamente dependentes de processos de limpeza e tratamento previamente à análise exploratória.

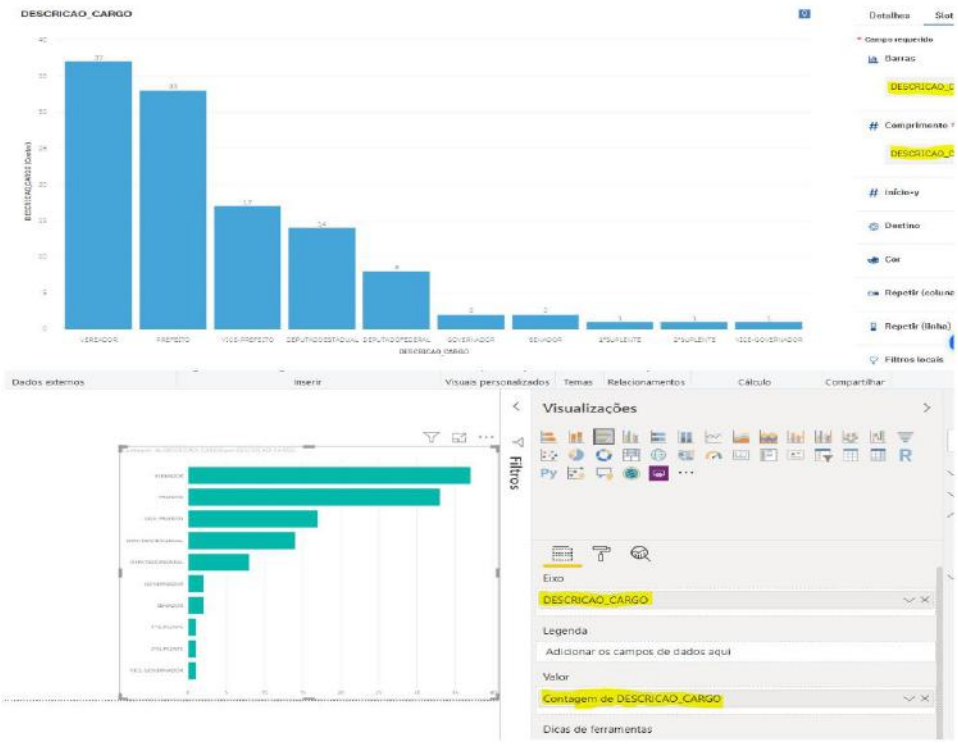

Figura 26 : Cognos Analytics e Power BI uso do mesmo campo

Ademais, aproveitando dos recursos de IA presentes nas ferramentas de análise, foi possível tanto no IBM Cognos Analytics, parte superior da figura 26, quanto no Power BI, parte inferior da figura 26, não criar a métrica de quantidade, de maneira que quando se utilizou o mesmo campo como eixo e valor simultaneamente na mesma análise, ele foi capaz de realizar uma contagem do campo informado comparativamente de forma automática.

Isso mostra que as ferramentas de análise avançada, de maneira geral, realizam o processo de entendimento da intenção do usuário em calcular uma determinada agregação de dados sem a necessidade de criação dessa métrica.

6.3.2 Análise dos estados de ocorrência do trabalho escravo

Na análise apresentada na figura 27, utiliza-se o mapa de calor para identificar os estados com o maior número de citações na lista suja. Neste caso, além da métrica quantitativa do estado, também se utilizou o serviço de refinamento do dado, em que classificamos a coluna relacionada aos estados como uma coluna do tipo geoespacial.

Com isso, o Cognos Analytics foi capaz de identificar diretamente no mapa as localidades citadas e, a partir do mapa de calor, evidenciar as regiões com pequeno ou grande envolvimento na lista suja. Neste caso, o estado do Pará é o mais citado.

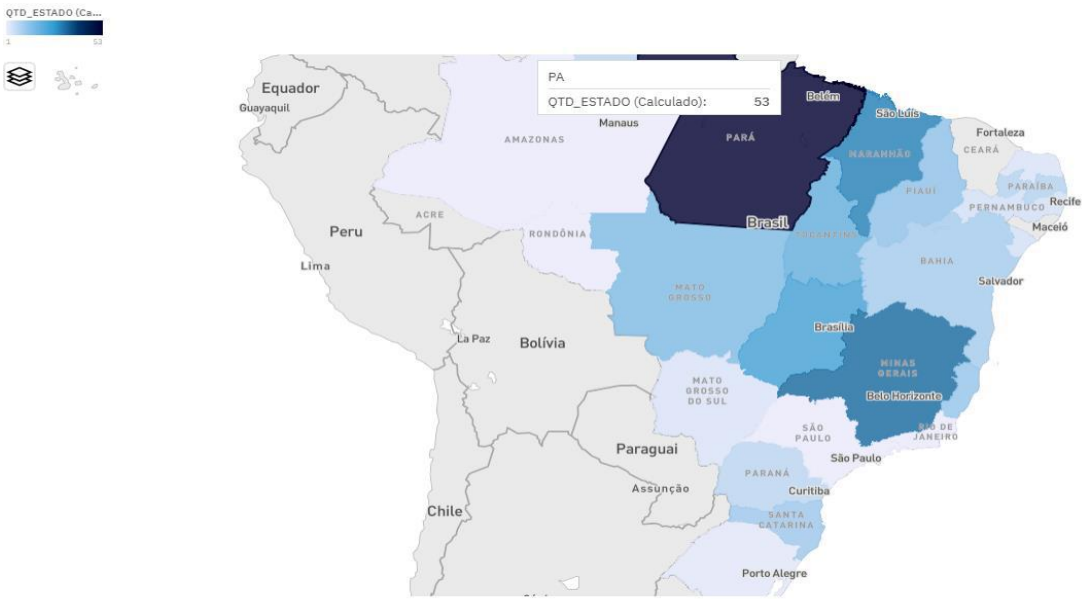

Figura 27 : Cognos Analytics estados citados

Com um pouco mais de complexidade, devido à necessidade de implementação da variação de cor pela intensidade, também é possível implementar o mesmo caso no Power BI, conforme figura 28, podendo selecionar inclusive o modelo de mapa para aéreo.

Assim, é natural que ferramentas de mercado quando identificam novas funcionalidades em ferramentas concorrentes, tentem implementar o mesmo recurso para entregar um resultado semelhante, todavia ocorrem, na maioria das vezes, diferenças de dificuldade na utilização ou qualidade do resultado.

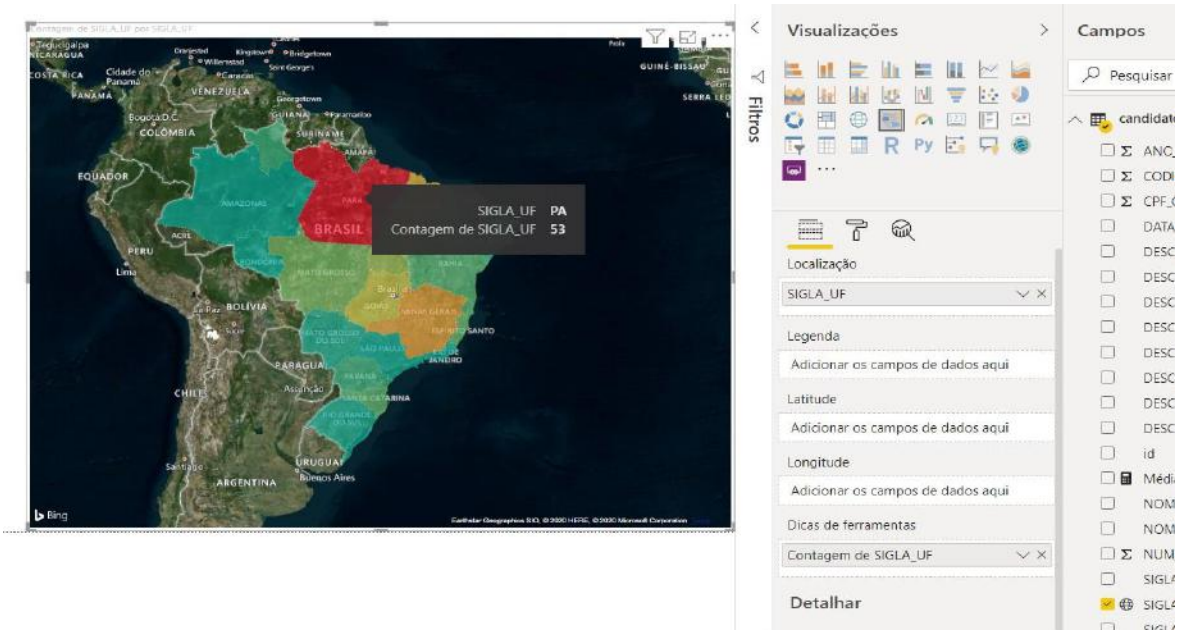

Figura 28 : Power BI estados citados

6.3.3 Análise comparativa dos partidos envolvidos e doações recebidas

Na análise apresentada na figura 29, utilizou-se a ferramenta de comparativo a fim de evidenciar disparidades ou proporcionalidades entre fatores dentro do conjunto de dados.

Desta forma, agrupando o valor de doações por partido, no gráfico à esquerda, e o quantitativo de trabalhadores encontrados em situação de trabalho escravo associados a partidos, no gráfico à direita, foi possível constatar uma relação quase que direta entre esses fatores.

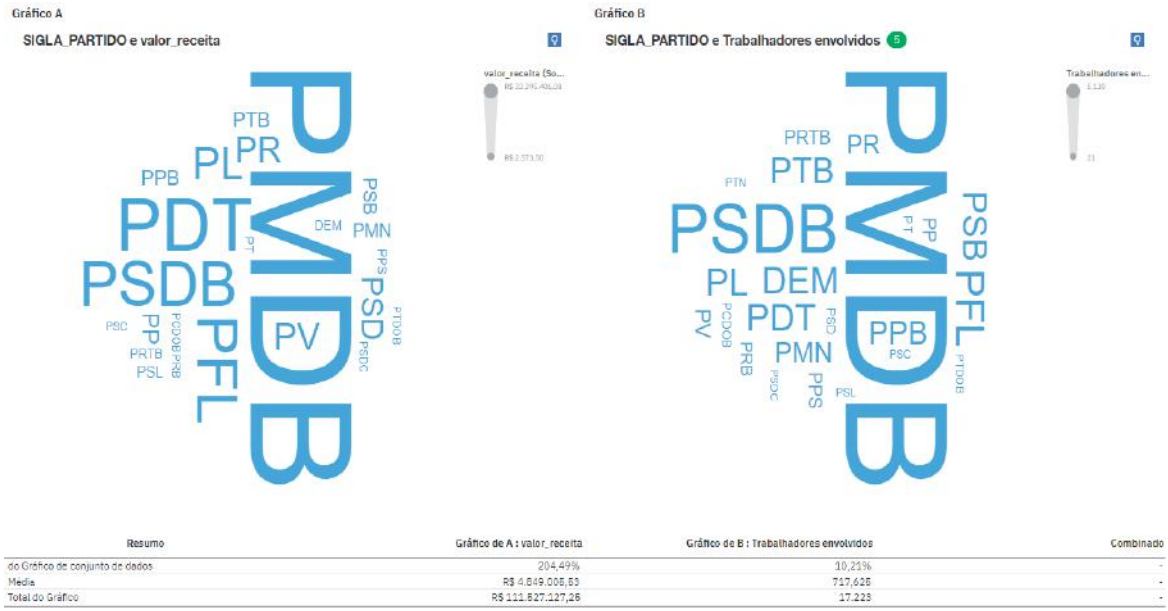

Figura 29 : Cognos Analytics comparativo de doações e envolvimento

Assim, os partidos que mais possuem doações são, na maioria dos casos, os com maior envolvimento em operações de trabalho escravo envolvendo proporcionalmente uma grande quantidade de trabalhadores em situação semelhante à escravidão.

6.3.4 Análise dos trabalhadores envolvidos por atividade e estado

Na análise apresentada na figura 30, foi realizado um comparativo de 3 fatores simultâneos e a própria ferramenta sugere a melhor visualização conforme a quantidade de campos envolvidos.

Neste caso, são relacionados a quantidade de trabalhadores envolvidos, limitada por uma escala dos 6 maiores estados envolvidos e o tipo de atividade exercida. Além disso, o Cognos Analytics permite a visualização não somente pelo tamanho de cada fatia do gráfico de pizza, mas com uma identificação percentual para cada atividade e, inclusive, o quantitativo em *popup*.

Neste caso, a atividade em azul é a sem valor informado e acaba sendo a com maior quantitativo, uma vez que é comum encontrar-se registros incompletos. Todavia, ainda é possível quantificar as demais atividades que representam o perfil desse tipo de trabalho em cada estado.

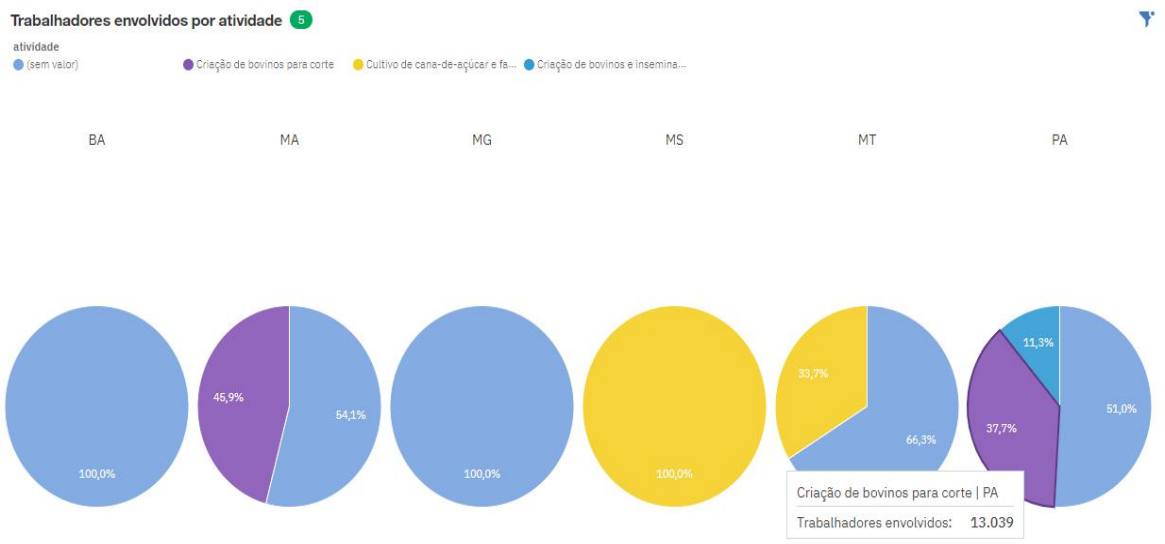

Figura 30 : Cognos Analytics trabalhadores envolvidos por atividade e estado

Essa análise é relevante, pois mostra a capacidade analítica da ferramenta, pois quando tentamos relacionar muitos fatores, a visualização de forma clara e objetiva fica majoritariamente comprometida se não for escolhido o modelo ideal que se adeque ao escopo dessa análise.

Assim, ferramentas de análise avançada, como um todo, possuem recursos de IA para renderização gráfica mediante o número de variáveis de entrada, adequando para uma visualização clara do usuário dos fatores envolvidos, com intuito de facilitar os *insights* sobre essa análise.

6.3.5 Análise de doação por ano e partido com projeção futura

Por fim, como última análise deste trabalho, a figura 31 apresenta a evolução das doações agrupadas por partido ao longo dos anos. Nesta análise, aparecem os *insights* disponibilizados pelo IBM Cognos Analytics durante toda investigação, à direita na aba detalhes.

O Cognos Analytics faz alguns apontamentos sobre o escopo em análise que revelam algumas informações úteis que podem direcionar qual caminho de descoberta tomar. Neste exemplo, ele mostra a média ponderada de todas as doações ao longo do gráfico e informa que o PMDB é o partido em que seus candidatos mais receberam doações diretas, totalizando mais

de 33 milhões em doações que correspondem a aproximadamente 30% do total de doações realizadas ao longo dos 8 anos da análise.

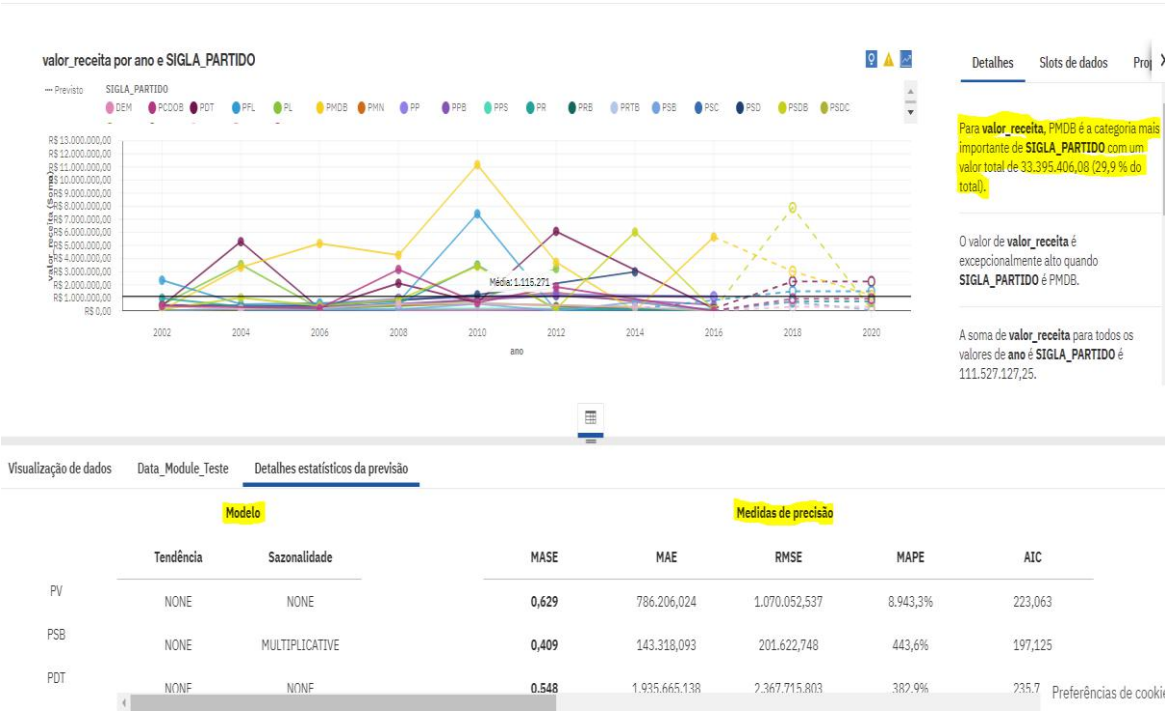

Figura 31 : Cognos Analytics doação por ano e partido com projeção

Além disso, foi acionado o recurso de previsão, dentro das capacidades preditivas do IBM Cognos Analytics, equivalentes aos do IBM Watson Analytics tratados no item 5.3.2.3 deste trabalho. O recurso de previsão gerou automaticamente uma visão do gráfico para os valores de doações agregadas dos partidos para os anos de 2018 e 2020, com base no modelo histórico de dados bienais.

Com uma melhoria com relação ao Watson Analytics, o IBM Cognos Analytics possui um detalhamento mais rico de informação acerca dos detalhes estatísticos da IA utilizados pela ferramenta para gerar a previsão, conforme a parte inferior da figura 31 em que aparecem o modelo e as medidas de precisão para essa previsão.

Os detalhes da previsão são gerados desde que os pontos no tempo sejam espaçados igualmente e as informações do modelo especifiquem o tipo de tendência e sazonalidade selecionados para estimar os dados da série temporal (IBM, 2019).

Dessa forma, conforme a figura 31, o Cognos Analytics utiliza uma série de métricas de precisão, parâmetros de sazonalidade e diagnósticos, com base nos modelos estatísticos e de IA presentes no ambiente analítico, oriundos do SPSS, conforme abordado no item 5.3.1 deste trabalho.

Não abordando minuciosamente os modelos estatísticos por não ser foco deste trabalho, as principais métricas utilizadas foram: o erro absoluto médio, MAE, o erro médio absoluto em escala, MASE, o erro quadrático médio da raiz, RMSE e o erro médio percentual absoluto da raiz, MAPE.

O Cognos utiliza modelos de suavização exponencial para gerar suas previsões. Múltiplos tipos de suavização são utilizados para criar modelos candidatos para cada série temporal em uma previsão. Os modelos utilizam equações de nivelamento, equações de suavização de tendência, equações sazonais de suavização e o método de um passo à frente, *k-step ahead* (IBM 2018). Esses modelos possuem a mesma complexidade em comum que é a dificuldade em determinar o valor mais apropriado da constante de suavização (USP IME 2019).

Todas as medidas de precisão são baseadas exclusivamente nos dados históricos. Logo, como explicado no item 5.3.2.3 deste trabalho, as métricas de precisão podem ser utilizadas como um indicador da precisão da previsão, todavia elas não podem ser transferidas para valores futuros.

Nessa análise o IBM Cognos Analytics mostra seus principais diferenciais, em vista que outras tecnologias de análise, como o Power BI não detalham os recursos estatísticos utilizados para uma determinada previsão, assim como não possuem os *insights* colaborativos durante as análises.

As ferramentas de análise avançada tendem, então, a não explicitar ou a reduzir o acesso a como determinado resultado foi obtido, o que acaba por limitar o rastreamento para validação de confiabilidade da informação.

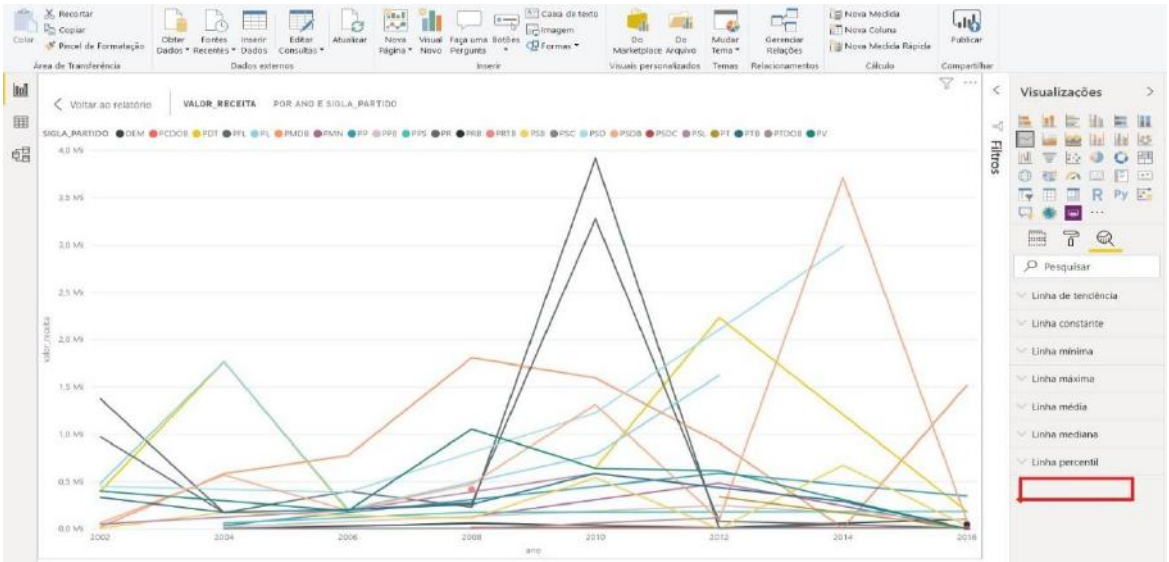

Figura 32: Power BI doação por ano e partido com projeção desabilitada

Na figura 32, está destacado que quando é gerada a mesma visualização de doações por ano e sigla do partido, o Power BI não habilita o recurso de previsão, pois ele só consegue trabalhar a previsão limitada a duas variáveis, enquanto o Cognos Analytics foi capaz de estimar valores futuros por partido, totalizando 3 métricas simultâneas.

No entanto, isso também acarretou o fato da margem de limite superior e inferior da previsão ser muito grande no Cognos Analytics, o que não garante grandes ganhos em informação e confiabilidade da informação.

No exemplo da figura 33, utilizamos somente duas variáveis, doações por ano, e foi possível realizar a previsão pelo Power BI.

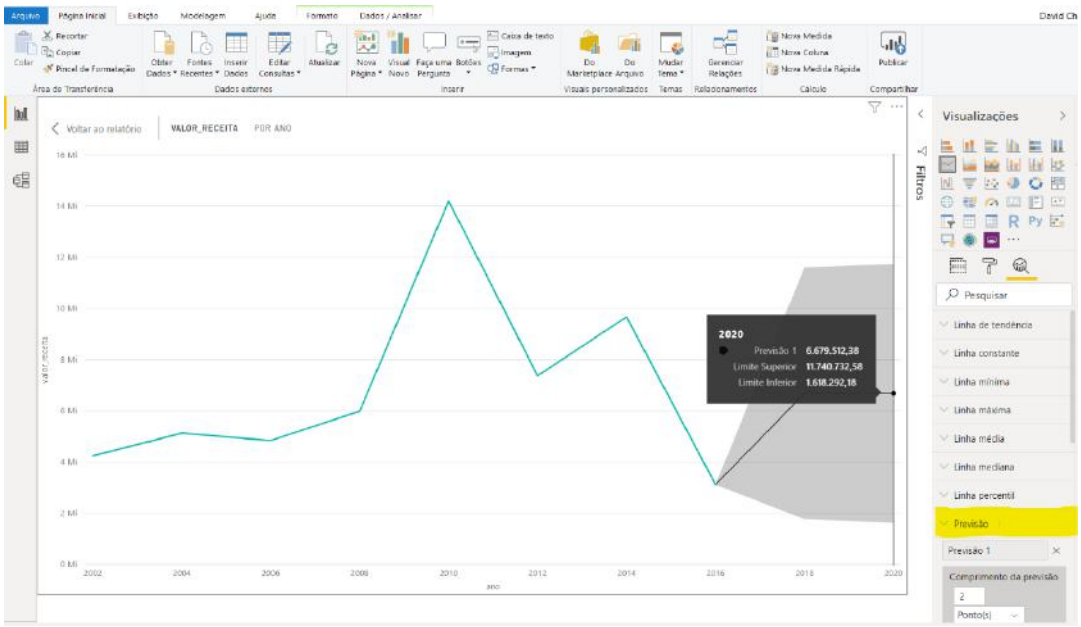

Figura 33 : Power BI doação por ano com projeção habilitada

O Power BI já exibe em sombra os limites inferiores e superiores da análise, que se apresentaram menores e mais confiáveis do que os do Cognos Analytics, todavia sem a possibilidade de adicionar novas variáveis dentro da previsão.

6.4 CONSIDERAÇÕES SOBRE OS RESULTADOS

Ao longo das explorações, foram trabalhadas as potencialidades e limitações das ferramentas de análise, de maneira a situar um cenário geral de seu estado atual, comparativamente aos diferenciais do IBM Cognos Analytics.

Nessas análises foram utilizados modelos de visualização comuns às ferramentas de análise, tais como:

- Gráficos de barra clusterizados (figuras 25 e 26): colunas em cluster utilizadas para representar valores discretos para mais de um item que compartilham de uma mesma categoria (IBM, 2018);
- Mapas coropléticos (figuras 27 e 28): utilizam o tipo de símbolo do mapeamento inteligente, contagens e quantidades, cor, para mostrar dados normalizados como áreas, linhas ou pontos sombreados (Arcgis, 2020);
- Nuvem de palavras (figura 29): possui sua visualização baseada no texto de uma coluna, onde a altura do texto representa a escala (IBM, 2018);
- Gráficos de pizza ou torta (figura 30): em que se coloca proporções são colocadas em destaque, e cada fatia mostra a relação percentual de cada parte para o todo (IBM, 2018);
- Gráficos de linha (figuras 31, 32 e 33): permite comparar as tendências e ciclos, inferir relacionamentos entre as variáveis, ou mostrar como uma única variável está evoluindo ao longo do tempo (IBM, 2018).

Dessa forma, análises exploratórias com recursos visuais elaborados facilitam os *insights* do analista que por sua vez elabora análises mais complexas por compreender mais profundamente o seu dataset.

É notável que todas compartilham da mesma limitação intrínseca de serem altamente dependentes de processos de limpeza e tratamento, todavia o IBM Cognos Analytics destaca-se, após ultrapassada essa barreira, em facilitar o acesso a informações relevantes que necessitariam de experiência em agregação de dados e cruzamento de informações.

Além disso, a ferramenta, ao longo de todo o processo de exploração, sugere novas análises que poderiam ser de interesse do analista associados a análise corrente, assim como sugere a melhor visualização para o relacionamento daqueles campos, uma vez que a quantidade de métricas e campos utilizados torna a visualização do dado muito mais complexa.

Resumidamente, as melhores ferramentas entregam resultados semelhantes em termos de possibilidades, apenas com diferenças na complexidade e qualidade na obtenção deles. Como mostrou o comparativo do Cognos Analytics com outra ferramenta de análise, o Power BI, que não detalhava os recursos estatísticos utilizados para uma determinada previsão, assim

como não possuía os insights colaborativos durante as análises e não habilitava o recurso de previsão para mais de duas variáveis.

No entanto, a aplicação da IA e outros elementos de automatização para gerar sugestões, pontos de partida de análise, visualizações mais adequadas e, inclusive, apresentar informações relevantes acerca do *dataset* durante toda a análise são o que efetivamente torna o IBM Cognos Analytics promissor e a inovação desses recursos, possíveis apenas com a formação de ambiente analíticos, determinará o próximo cenário na exploração de dados.

7 CONCLUSÃO

Esta monografia abordou a evolução da área de exploração de dados como fator fundamental para mudança do cenário de TI no meio corporativo, saindo do conceito generalizado de área de suporte para um conceito de área gerencial indispensável para melhor tomada de decisões.

Além disso, discutiu como a área de análise de dados continuará a crescer pelos próximos anos, mas que ainda possui muitas barreiras a serem rompidas a fim de atender as demandas exponencialmente crescentes de todos os setores empresariais, pequenos e grandes, que necessitam de informações mais rápidas, mais assertivas e sobre uma quantidade cada vez maior de dados.

Através de uma revisão do campo da inteligência de negócio alinhada à apresentação de conceitos de Porter e cenários do ramo de negócios, foi possível visualizar o impulso ao desenvolvimento da análise exploratória de dados que levaram ao surgimento e avanço das análises avançadas.

Este estudo, então, tratou sobre o atual estado da arte de ferramentas de exploração de dados, utilizando as ferramentas analíticas da IBM sobre um exemplo de dados abertos de trabalho escravo no Brasil. Verificou-se que as ferramentas de análise ainda necessitam de dados com altos índices de tratamento e sob modelos muito restritos, o que conflita com a realidade de dados cada vez mais diversificados e não estruturados, todavia que denota o fato de que muitas dessas ferramentas serão utilizadas em cima de DWs que possam garantir o dado no melhor molde para consumo.

Outro ponto observado neste trabalho é que por mais que se busquem métodos de facilitação para que profissionais que não sejam da área de TI possam criar suas análises, as ferramentas de análises avançadas ainda são pouco transparentes com relação às técnicas utilizadas para gerar cada visualização. Demandam ainda conhecimento e experiência para manipular modelos que possam fazer predições adequadas ou até mesmo gerar efetivamente decisões de negócio.

Deste modo, conforme o quadro 5, fica claro que as empresas ainda necessitam construir um ambiente analítico, pois ferramentas de análises avançadas são pouco efetivas sem o dado previamente tratado e são necessárias múltiplas ferramentas com atuação de um time de TI para conseguir enfim extrair informação relevante em larga escala.

Ademais, com o crescimento em larga escala tanto das análises quanto dos ambientes analíticos necessários ao avanço da análise de dados, mais imprescindível se torna uma gestão de dados eficiente e igualmente dinâmica e fluida para evitar problemas semelhantes aos ocorridos no passado de falta de credibilidade e resultados distintos de um mesmo conjunto de dados, assim como o gasto de recursos em soluções repetidas dentro de uma mesma empresa. Resumidamente, manter o dado atualizado, acessível, reduzido em tamanho e seguro.

Por fim, para evoluir ainda mais esse estudo em trabalho futuros, poderia se realizar um comparativo de método a método de exploração com as técnicas de análises avançadas aqui descrita a fim de analisar mais profundamente a capacidade das principais ferramentas do mercado atualmente.

Outro ponto relevante seria realizar um estudo dos modelos estatísticos utilizados pelas ferramentas e das técnicas de mineração e aprendizado de máquina para compreender mais a fundo como as ferramentas funcionam internamente a fim de situar mais tecnicamente suas limitações.

Com intuito de manter o cunho de pequena contribuição, para ampliar estudos em um domínio de tanta relevância como o do trabalho escravo abordados neste trabalho, as análises futuras deverão ser realizadas sobre dados abertos mais recentes que abordem as relações de trabalho escravo e funcionalismo público, tornando ao final as análises e dados puros disponíveis e acessíveis para a população.

REFERÊNCIAS

ARCGIS. **Criar e utilizar um mapa coroplético**. Disponível em: https://doc.arcgis.com/pt-br/insights/latest/create/choropleth-maps.htm. Acesso em: 05 Mar. 2020.

BRESLIN, M. B. **Data Warehousing Battle of the Giants**: Comparing the Basics of the Kimball and Inmon Models. 2004. Disponível em: https://pdfs.semanticscholar.org/c80f/8aaea5bf58846b0125b460401fed8230c2d2.pdf?_ga=2.181811360.669057503.1592348442-1125676588.1592348442. Acesso em: 05 mar. 2020.

CARVALHO, Marly Monteiro de; LAURINDO, Fernando José Barbin. **Estratégia competitiva**: dos conceitos à implementação. 2. ed. São Paulo: Atlas, 2010.

COLLINS ENGLISH DICTIONARY. **Commodity definition and meaning.** Disponível em: https://www.collinsdictionary.com/dictionary/english/commodity. Acesso em: 09 jan. 2019.

DRESNER, Howard. **Wisdom of Crowds business intelligence Market Study**. [S.l.: s.n.], 2018.

FERREIRA, Luciene Braz; RAMOS, Anatália. Tecnologia da Informação: commodity ou ferramenta estratégica? **Journal of Information Systems and Technology Management**, São Paulo, v. 2, n. 1, p. 69-79, 2005.

GARTNER. **Gartner Forecasts Worldwide Public Cloud Revenue to Grow 17.3 Percent in 2019**. Disponível em: https://www.gartner.com/en/newsroom/press-releases/2018-09-12-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-17-percent-in-2019 Acesso em: 21 jun. 2019.

_____. **Gartner Information Technology**: Metadata. Disponível em: https://www.gartner.com/en/information-technology/glossary/metadata. Acesso em: 02 abr. 2019.

_____. **Gartner IT Glossary**: Advanced Analytics. Disponível em: https://www.gartner.com/it-glossary/advanced-analytics. Acesso em: 02 abr. 2019.

_____. **Gartner IT Glossary**: Descriptive Analytics. Disponível em: https://www.gartner.com/it-glossary/descriptive-analytics. Acesso em: 02 abr. 2019.

_____. **Gartner IT Glossary**: Predictive Analytics. Disponível em: https://www.gartner.com/it-glossary/predictive-analytics. Acesso em: 02 abr. 2019.

_____. **Gartner Magic Quadrant for Analytics and Business Intelligence Platforms**. Disponível em: https://cadran-analytics.nl/wp-content/uploads/2019/02/2019-Gartner-Magic-Quadrant-for-Analytics-and-Business-Intelligence-Platforms.pdf . Acesso em: 02 jun. 2019.

HOYT, Robert Eugene et al. IBM Watson Analytics: Automating Visualization, Descriptive, and Predictive Statistics. **JMIR Public Health And Surveillance**, [S.l.], v. 2, n. 2, p.157-167, 2016.

IBM. **IBM Cognos Analytics**. Disponível em: https://www.ibm.com/br-pt/products/cognos-analytics. Acesso em: 30 dez. 2019.

_____. **IBM Knowledge Center**. Disponível em: https://www.ibm.com/support/knowledge center. Acesso em: 30 dez. 2018.

IDC BRASIL. **Cloud Pública mantém ritmo de crescimento forte até 2022**. Disponível em: http://br.idclatin.com/releases/news.aspx?id=2462. Acesso em: 16 nov. 2019.

INMON, W. H. **Building the Data Warehouse**. 3. ed. Indianápolis: Wiley, 2002.

INTEL. **Achieving intel transformation through IT innovation**: 2014-2015 Intel It Business Review - Anual Edition. Disponível em: https://www.intel.com/content/dam/www/public/us/en/documents/best-practices/intel-it-annual-performance-report-2014-15-paper.pdf. Acesso em: 16 nov. 2019.

KIMBALL, Ralph; ROSS, Margy. **The Data Warehouse Toolkit**: the definitive guide to dimensional modeling. 3. ed. Indianápolis: Wiley, 2013.

KOTLER, Philip. **Administração de Marketing.** 10. ed. São Paulo: Prentice Hall, 2000.

LOBATO, David Menezes et al. **Gestão estratégica.** Rio de Janeiro: FGV, 2012.

LUHN, H. A business intelligence System. **IBM Journal of Research and Development**, Armonk, v. 2, n. 4, p. 314-319, 1958.

MIYACHI, Christine. What is “Cloud”? It is time to update the NIST definition? **IEEE Cloud Computing**, [S.l.], v. 5, n. 3, p. 6-11, maio 2018.

MOSS, Larissa T.; ATRE, Shaku. **Business Intelligence Roadmap**: The complete project lifecycle for decision support applications. Boston: Addison Wesley, 2003.

PORTER, Michael. **Estratégia competitiva**: técnicas para análise de indústrias e da concorrência. Rio de Janeiro: Campus, 2005.

ROSSMAN, Allan J; CHANCE, Beth L. **Workshop statistics**: discovery with data. 4. ed. Emeryville: Key College Pub, 2001.

ROZENFELD, Henrique et al. **Gestão de Desenvolvimento de Produtos**: uma referência para a melhoria do processo. São Paulo: Saraiva, 2006.

SEBRAE. **Ferramenta**: Análise SWOT clássico. Disponível em: http://www.sebrae.com.br/Sebrae/Portal%20Sebrae/Anexos/ME_Analise-Swot.PDF. Acesso em: 10 ago. 2018.

SECRETARIA DE TRABALHO. **Combate ao Trabalho Escravo**. Disponível em: http://trabalho.gov.br/fiscalizacao-combate-trabalho-escravo. Acesso em: 30 dez. 2019.

STAUBER, Doug. **SPSS**: 50 Years of Innovation. 2018. Disponível em: https://developer.ibm.com/predictiveanalytics/2018/04/05/spss-50-years-innovation/. Acesso em: 25 ago. 2019.

SUCOLOTTI, Ângelo Alfredo et al. **Modelando um Data Warehouse Dimensional a partir de um Modelo Hierárquico**. [S.l.: s.n.], 2019.

TAVARES, Mauro Calixta. **Gestão Estratégica**. São Paulo: Atlas, 2000.

TI INSIDE. **Big Data e Analytics pode chegar a US$ 8,5 bilhões em 2023 na AL, segundo Frost & Sullivan**. Disponível em: https://tiinside.com.br/tiinside/23/04/2018/o-bilionario-mercado-brasileiro-de-big-data-e-analytics-segundo-a-frost-sullivan. Acesso em: 01 out. 2019.

USP IME. **Modelos de suavização exponencial**. Disponível em: https://www.ime.usp.br/~chang/home/mae325/MAE325-An%E1lise%20de%20S%E9ries%20Temporais/aulas/mae0325-aula08-2019.pdf. Acesso em: 30 dez. 2019.

VERONA, Letícia. **Métricas para análise de poder em redes**: uma abordagem com aporte da teoria de Manuel Castells. 2018. 113 f. Dissertação (Mestrado em Informática) - Instituto de Matemática, Instituto Tércio Pacciti de Aplicações e Pesquisas Computacionais, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2018.

WATSON, Hugh J.; WIXOM, Barbara H. The Current State of Business Intelligence. **Computer**, Long Beach, v. 40, n. 9, p. 96-99, set. 2007.

WEB STANDFORD EDU. **Data Warehousing Concepts**. 2003. Disponível em: https://web.stanford.edu/dept/itss/docs/oracle/10gR2/server.102/b14223/concept.htm. Acesso em: 06 dez. 2018.