MAILSON RAFAEL FONTES

COMO MELHORAR A QUALIDADE DE DADOS EM BIG DATA COM METODOLOGIA TDQM E USO DE FERRAMENTAS DE QUALIDADE

SÃO PAULO
2017

MAILSON RAFAEL FONTES

COMO MELHORAR A QUALIDADE DE DADOS EM BIG DATA COM METODOLOGIA TDQM E USO DE FERRAMENTAS DA QUALIDADE

Monografia apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do certificado de Especialização em Gestão e Engenharia da Qualidade – MBA/USP.

Orientador: Professor Doutor Adherbal Caminada Netto

SÃO PAULO

2017

AGRADECIMENTOS

Gostaria de agradecer aos meus pais, Carlos e Vani, e a minha namorada, Claudia de Oliveira, por todo o amor e apoio incondicional, pela força e suporte que sempre precisei decorrer desta jornada. Eles foram essências para eu concluir essa etapa.

RESUMO

Big data é um conceito que vem sendo discutido e adotado em todas as grandes organizações, pois se tornou um diferencial competitivo para o mundo corporativo. O mercado brasileiro ainda não está maduro, pois são poucas empresas que investiram nessas soluções e estão mensurando os resultados. O cenário brasileiro ainda é marcado por algumas confusões conceituais no que diz respeito a big data e um dos objetivos deste trabalho é esclarecer os conceitos ainda obscuros sobre este tema. Derivado da *Total Quality Management* (TQM) que consiste em criar consciência de qualidade em todos os processos organizacionais, será apresentado a metodologia TDQM (*Total Data Quality Management*), onde o foco é direcionado para a qualidade dos dados. Como estudo de caso, será utilizado o exemplo de uma grande instituição financeira brasileira. Será apresentado o histórico da empresa, como a empresa vem utilizando o big data, quais as ferramentas adotadas para criar diferencial competitivo de mercado e quais os tratamentos de dados são utilizados para garantir o valor da informação.

Palavras-chave: Big Data, Hadoop, TDQM, *Total Data Quality Management* Qualidade de dados, Governança de dados.

ABSTRACT

Big data is a concept that has been discussed and adopted in all large organizations because it has become a competitive advantage for the enterprise. The Brazilian market is not yet mature, because there are few companies that have invested in these solutions and are measuring the results. The Brazilian situation is still marked by some conceptual confusions with regard to big data and one of the objectives of this study is to clarify the still obscure concepts on this topic. Derived from Total Quality Management (TQM) that is to create quality awareness in all organizational processes, TDQM methodology (Total Data Quality Management) will be displayed, where the focus is directed to the quality of the data. As a case study, we will use the example of a large Brazilian financial institution.

The history of the company will be presented, as the company has been using big data, which the tools used to create competitive advantage and market data which treatments are used to ensure the value of information.

Key words: Big Data, Hadoop, TDQM, Total Data Quality Management, Data Quality, Data Governance.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

SQL:	Structured Query Language
TDQM:	Total Data Quality Management
SGBD:	Sistemas de Gerenciamento de Banco de Dados
CRM:	Customer Relationship Management
RFID:	Radio Frequency Identification
NoSQL:	Not Only SQL
GFS:	Google File System
HDFS:	Hadoop Distributed File System
PDCA:	Plan, Do, Check, Act

SUMÁRIO

1 INTRODUÇÃO

1.1 CONSIDERAÇÕES INICIAIS

O presente trabalho discorre sobre a importância da qualidade de dados no ambiente big data. O conceito big data pode ser definido como um conjunto de bases de dados, sejam elas com dados estruturados e não estruturados (documentos, imagens, vídeos e até mesmo outros bancos de dados relacionais) tão complexa e volumosa que se torna muito difícil fazer algumas operações básicas e descomplicadas (inserção, remoção, ordenação, atualização e sumarização) de forma eficiente utilizando sistemas de gerenciamento de banco de dados (SGBD) tradicionais como *SQL Server*, Oracle, MySQL entre outros SGBDs consolidados no mercado atual. Por causa desse problema, e outros demais, um novo conjunto de plataformas de ferramentas voltadas para big data tem sido proposta, como por exemplo Apache Hadoop [9].

As grandes instituições financeiras, companhias aéreas, operadoras de telefonia, serviços de busca on-line e redes varejistas são apenas alguns dos inúmeros exemplos de empresas que convivem diariamente com grandes volumes de informações, entretanto ter os dados e saber usa-los não basta, é importante garantir a qualidade dos dados. Atualmente a informação é o recurso mais valioso dentro de uma organização, diante disso, a informação deve ser precisa, acessível, consistente e agregar valor para a companhia. Os dados são a matéria prima da informação que por sua vez é o insumo para gerar conhecimento e acrescentar valor para as instituições. Os dados precisam ser tratados e adaptados a um conjunto de regras e técnicas que ajudam manter a credibilidade e proporcione a interpretação do mesmo de maneira eficiente e eficaz. Atualmente inúmeras organizações não implementam as técnicas e ferramentas para manter a qualidade de dados devido ao investimento alto de recursos monetários para aquisição de ferramentas de mercado, contratação de profissionais especializados em qualidade e o tempo que levaria para ajustar os processos que hoje operam diariamente e suprem de maneira dispendiosa as necessidades das organizações.

Manter a qualidade dos dados, ou seja, ter um ambiente big data com os dados confiáveis, disponíveis para consumo imediato e de fácil interpretação poderá proporcionar as organizações vantagens competitivas, uma vez que o conceito big

data ainda não está popular no mercado brasileiro. São variadas as formas de se criar vantagens competitivas com a utilização dos recursos que o big data proporciona. Realizar a exploração dos dados (*data mining*) à procura de padrões consistentes e de estatísticas para procurar correlação entre diferentes dados que permitam adquirir conhecimento benéfico para uma empresa ou indivíduo são algumas das maneiras de agregar valor à organização que se sujeita a realizar tal investimento. Este trabalho almeja orientar os esforços necessários para adquirir a qualidade dos dados. Para chegar a tal objetivo, foi desenvolvido um processo que utiliza um conjunto de especialidades e princípios da TDQM (*Total Data Quality Management*) juntos com outros conceitos e características de administração de dados já solidificadas no mercado atual.

1.2 OBJETIVO GERAL

Este trabalho tem por objetivo geral apresentar a metodologia TDQM (Total Data Quality Management) e como este processo utilizado de forma estratégica combinado com algumas ferramentas poderá garantir a qualidade dos dados agregando valor as corporações.

1.3 OBJETIVOS ESPECÍFICOS

Para alcançar o objetivo e extrair resultados consistentes e significativos neste trabalho, foram definidos o escopo e as metas abaixo:

- Definir o conceito big data e os demais conceitos e as plataformas que compõem a estrutura no ambiente de dados;
- A importância e influência que o big data proporciona para as empresas;
- Pesquisar modelos padrões utilizados atualmente para garantir a qualidade de dados;
- Investigar e explorar novas plataformas solidificadas que realizem a manipulação massiva de dados;
- Avaliar técnicas e ferramentas de tratamento de qualidade preventiva e reativa;

- Utilizar a metodologia TDQM e definir as etapas para a criação de um processo de qualidade de dados.
- Desenvolver o processo; e,
- Aplicar o processo definido pela metodologia TDQM a um exemplo real e posteriormente analisar os resultados alcançados.

1.4 ESCOPO

O escopo desse trabalho visa apresentar o conceito de big data, apresentar de maneira breve a plataforma hadoop e uma visão geral sobre sua estrutura. O trabalho contempla também apresentar o que é TDQM e algumas técnicas de qualidade de dados. Está fora do escopo deste trabalho apresentar métodos e modelos de extração e exploração de dados, modelagem de dados e arquitetura de dados. Este trabalho se limitará a falar exclusivamente a área de CRM da empresa estudada e mostrará os processos da área, o método de correção de qualidade de dados nas comunicações digitais especificadas pela área de negócio e construída pela área de TI.

2 FORMULAÇÃO TEÓRICA

A finalidade deste capítulo é fornecer as informações necessárias para compreender melhor a proposta deste trabalho. Na primeira seção será abordado o conceito de dados, informação e conhecimento, a segunda seção apresenta o conceito de big data e os 5V's que o compõem, na seção seguinte discorre sobre a importância do big data para as instituições e cenário atual, na quarta seção será apresentada a tecnologia Hadoop, utilizada em grandes empresas para realizar o processamento massivo dos dados e a última seção descreve a metodologia *Total Data Quality Management*.

2.1 DADOS, INFORMAÇÃO E CONHECIMENTO

Antes de abordar os conceitos de qualidade de dados, big data e a metodologia *Total Data Quality Management* é de extrema importância passar por alguns conceitos básicos. Diariamente recebemos milhares de dados, informações e conhecimento por diversos meios de comunicação, entretanto sabemos a real diferença entre eles? De acordo com Setzer, dado é definido como uma sequência de símbolos quantificados ou quantificáveis [4]. O dado é a matéria prima que formará a informação, ou seja, dado é a informação não tratada. Os dados de maneira isolada não podem transmitir uma mensagem ou representar conhecimento. Exemplos de dados: Nome do funcionário, número de horas trabalhadas, peças no estoque e etc.

Existem diversas definições para a informação e uma delas parte do conceito de que a informação é o dado configurado de forma adequada ao entendimento e à utilização pelo ser humano [5]. Ou seja, a informação é o resultado dos dados tratados, comparados e classificados utilizado para a tomada de decisão.

O conhecimento facilita reconhecer quais dados e informações são úteis para se atingir os objetivos traçados pela organização. Para LAUDON e LAUDON (1999, p. 10), conhecimento é o conjunto de ferramentas conceituais e categorias usadas pelos seres humanos para criar, colecionar, armazenar e compartilhar a informação [7]. A tabela 01 mostrada abaixo apresenta as caraterísticas de dados, informação e conhecimento:

Dado	Informação	Conhecimento
Simples informações sobre o estado do mundo	**Dados dotados de relavância e propósito**	**Informação valiosa da mente humana. Inclui reflexão, síntese e contexto**
Facilmente estruturado	Requer unidade de Análise	Difícil estruturação
Facilmente obtido através de máquinas	Exige consenso em relação ao significado	Difícil captura por máquinas
Frequentemente quantificado	Exige necessáriamente a medição humana	Difícil de ser medido e quantificado
Facilmente previsível		

Tabela 1: Dado, informação e conhecimento

Uma vez entendido alguns conceitos básicos sobre dados, iremos aprofundar nos demais assuntos.

2.2 BIG DATA E OS 5V'S

Diversas definições de big data podem ser encontradas, porém algumas características são recorrentes possibilitando sua conceituação. Big data é o termo que descreve o imenso volume de dados, estruturados e não estruturados que impactam os negócios no dia a dia. Mas o importante não é a quantidade de dados e sim o que as empresas fazem com os dados que realmente importam [1]. Dados estruturados são aqueles organizados e armazenados em SGBDs, geralmente relacional, que possui como características uma descrição sobre os dados, conhecidos como metadados. Dados não estruturados são dados que não possuem uma estrutura definida, ou seja, são quaisquer documentos, arquivos, imagens, vídeos que não tenham sido codificados, ou de outra forma estruturados em linhas e colunas ou registros.

Big Data pode ser analisado para a obtenção de insights que levam a melhores decisões e direções estratégicas de negócio [1]. Para garantir o bom uso desta tecnologia é essencial conhecer os 5 V's do big data: Volume, Velocidade, Variedade, Veracidade e Valor.

- Volume: O conceito de volume no big data é melhor evidenciado pelos fatos do quotidiano: diariamente, o volume de troca de e-mails, transações bancárias, interações em redes sociais, registro de chamadas e tráfego de

dados em linhas telefônicas. Todos esses servem de ponto de partida para a compreensão do volume de dados presentes no mundo atualmente [2].

- Velocidade: É necessário que a transmissão de informações seja feita a uma velocidade ímpar para gerar bilhões de dados. Por isso, certos informes podem ser considerados ultrapassados, minutos depois de serem veiculados em uma rede social, como o *Facebook* e o *Twitter*, que tem como característica ser altamente dinâmico [2].
- Variedade: os dados vêm de sistemas estruturados (hoje minoria) e não estruturados (a imensa maioria), gerados por e-mails, mídias sociais, documentos eletrônicos, apresentações estilo *PowerPoint*, mensagens instantâneas, sensores, etiquetas RFID (*Radio Frequency Identification*), câmeras de vídeo, etc [3].
- Veracidade: É preciso destacar o que é rico em conteúdo em meio a tanta informação. Ao garantir essa separação, que é possível fazer a partir do big data, o que sobra são conhecimentos importantes para compreender melhor o consumidor e o seu comportamento [2].
- Valor: É necessário geral valor dos resultados que retornam do Big Data. O processo necessita estar focado para a orientação do negócio e assim auxiliar nas tomadas de decisões [2].

O conceito big data pode ser comparado com a descoberta do microscópio, que possibilitou a medicina a ver outras coisas que existiam, mas que não enxergávamos, como as bactérias e vírus. O que o microscópio foi para a medicina e para a sociedade, o big data terá a mesma importância para as empresas e para a sociedade [11]. Ele nos proporcionará enxergar comportamentos, identificar padrões e tendências que hoje com as ferramentas de modelos relacionais não nos proporciona. De forma resumida, o big data é um conjunto de tecnologias, práticas e processos que permitem as organizações analisarem e tomarem decisões baseadas em análise de dados de forma massiva ou até mesmo gerenciar atividades de forma muito mais eficiente.

2.3 A IMPORTANCIA DO BIG DATA

Segundo a empresa SAS, a importância do big data não gira em torno da quantidade de dados que você tem, mas em torno do que você faz com eles. Quando as organizações combinam big data com *analytics*, as mesmas podem realizar tarefas relacionadas ao negócio, como:

- Determinar a causa raiz de falhas, problemas e defeitos em tempo quase real;
- Gerar cupons no ponto de venda com base em hábitos de compra dos clientes;
- Recalcular carteiras de risco inteiras, em questão de minutos;
- Detectar comportamentos fraudulentos antes que eles afetem sua organização[1].

No início de 2016 a empresa Xerox desenvolveu um estudo na Europa sobre a gestão de dados e a implementação de iniciativas de big data. O estudo foi desenvolvido com base num inquérito a 330 executivos de topo na Europa Ocidental. A análise permite identificar três tendências que salientam o papel do big data na gestão das organizações [9]:

- Big data são fator chave nas decisões: 61% das organizações referiram que as decisões feitas durante 2015 iriam ser provavelmente mais baseadas em informação orientada por dados que em fatores como instinto, opinião ou experiência.
- Dados incorretos são dispendiosos: 70% das organizações ainda encontram dados incorretos nos seus sistemas e 46% acreditam que este fato tem um impacto negativo no negócio, requerendo novo cálculo ou conjuntos de dados totalmente inutilizáveis.
- Segurança e privacidade dos dados: 37% dos inquiridos referiram a segurança e privacidade como um dos maiores desafios quando implementam estratégias de big data.

A utilização do big data é fundamental para acelerar os negócios e nortear a alta administração, seja para detectar e prevenir ocorrências de fraude, para otimizar

processos de gestão ou mesmo para potenciar o valor de um cliente. De acordo com Taurion[11], o big data permite a criação de novos modelos de negócios para diversas áreas baseados no valor da informação. Através das análises preditivas as empresas de diversos setores passam a ter condições de evitar inúmeros tipos de desperdício. Utilizando um exemplo básico e comum para a maioria das pessoas que adquire um carro, a concessionária orienta os compradores a fazerem a manutenção preventiva ao atingirem a quilometragem especifica. Ainda baseado em Taurion [11], se as análises fossem embasadas em algoritmos e base de dados, as manutenções passariam a ser preditivas para cada veículo, pois os desgastes do carro variam conforme o uso e hábitos dos motoristas.

Um setor que tem muito a ganhar com o big data é a educação. Os grandes centros de pesquisas atualmente já estão acostumados a trabalhar com volumes de dados consideráveis. Porém adequar os modelos educacionais e torná-los mais atrativos para os alunos tem sido um grande desafio. Segundo a DataStorm [12], atualmente as avaliações em massa para os alunos é uma falha do modelo tradicional, ou seja, não são consideradas as características próprias de aprendizado e aplicação de conhecimento. Com o auxílio do big data, torna-se possível realizar uma análise de performance dos alunos e com isso focar os pontos fracos de conhecimento e assim potencializar os resultados. Utilizando o big data para análise de dados no ambiente escolar, será possível criar métodos de ensino mais eficazes, tornando o trabalho do docente mais dinâmico e facilitar o aprendizado do estudante [12].

Outro exemplo que pode ser utilizado de big data no setor da educação é sugerir, baseado em análise de padrões de milhares de alunos, quais seriam as profissões mais adequada a cada pessoa. Os conhecidos testes vocacionais combinados com o uso massivo de dados podem ajudar a identificar a melhor combinação carreira/personalidade [11].

2.4 AS INFRAESTRUTURAS QUE SUSTENTAM O BIG DATA

De acordo com Taurion [11], as tecnologias que sustentam o big data podem ser avaliadas de duas formas: as entrelaçadas com o *analytics*, tendo a plataforma *Hadoop* e *Mapreduce* como principais tecnologias de infraestrutura para armazenamento e processamento de grande volume de dados no mercado atual. A

localização, identificação, armazenamento, interpretação e processamento massivo de dados não estruturados começam a demandar novas tecnologias. Neste cenário, o big data impulsiona o uso dos bancos de dados *NoSQL (Not only SQL)*, ou seja, começam a se destacar os bancos de dados não relacionais. Ainda segundo Taurion [11], a resiliência é um ponto que se torna cada vez mais importante, pois a infraestrutura do big data tem que ser desenhada para alta disponibilidade quando o objetivo é realizar o processamento de dados em tempo real. Um bom exemplo que pode ser utilizado é o aplicativo de navegação *Waze* [13].

Waze é um dos maiores aplicativos de transito e navegação atualmente, onde milhares de pessoas vão reportando em tempo real a velocidade em que estão andando, acidentes nas vias, radares e tem como objetivo impedir que os usuários fiquem em congestionamentos e com isso economizem tempo e combustível. Caso a infraestrutura do *Waze* não esteja disponível, o impacto pela falta de informação poderá causar engarrafamento e isso afetar a qualidade de vida de muitos cidadãos, sejam eles usuários ou não do aplicativo.

2.5 HADOOP

Atualmente uma tecnologia que está se destacando no cenário de big data é a plataforma Hadoop. O Apache Hadoop é uma plataforma *open source* para armazenamento e processamento distribuído de grandes conjuntos de dados com alta disponibilidade, escalabilidade, e tolerante a falhas. O Hadoop é utilizado em *hardware commodity*, ou seja, é utilizado em *hardware* simples e mais baratos que os servidores convencionais [14]. O Hadoop foi inspirado na plataforma GFS (*Google File System*) e no paradigma de *MapReduce*, que divide o trabalho nas etapas de *Mapper e Reducer* que manipulam os dados distribuídos em um cluster de servidores [11]. Basicamente, o Hadoop na prática é uma junção de dois projetos: *Hadoop Map Reduce e Hadoop Distributed File System (HDFS).*

De acordo com a *Apache Haddop* [15], o *Hadoop Distributed File System (HDFS)* é um sistema de arquivos distribuídos projetado para rodar em hardware commodity. O HDFS fornece acesso de alta taxa de transferência de dados de aplicativos e é adequada para aplicações que têm grandes conjuntos de dados.

A arquitetura do HDFS é composta por clusters de nós interconectados no local onde os arquivos e diretórios residem. O HDFS é implementado sobre a

arquitetura mestre/escravo, havendo como mestre uma instância do *NameNode* e em cada escravo uma instância do *DataNode*. Um cluster HDFS consiste em um único *NameNode*, um servidor mestre que gere o espaço de nomes do sistema de arquivos e regule o acesso a arquivos por clientes. Os *DataNodes* são responsáveis por servir, ler e escrever pedidos de clientes do sistema de arquivos. Os *DataNodes* também realizam a criação do bloco, exclusão e replicação sob comando da *NameNode* [16].

A figura 01 ilustra a arquitetura HDFS:

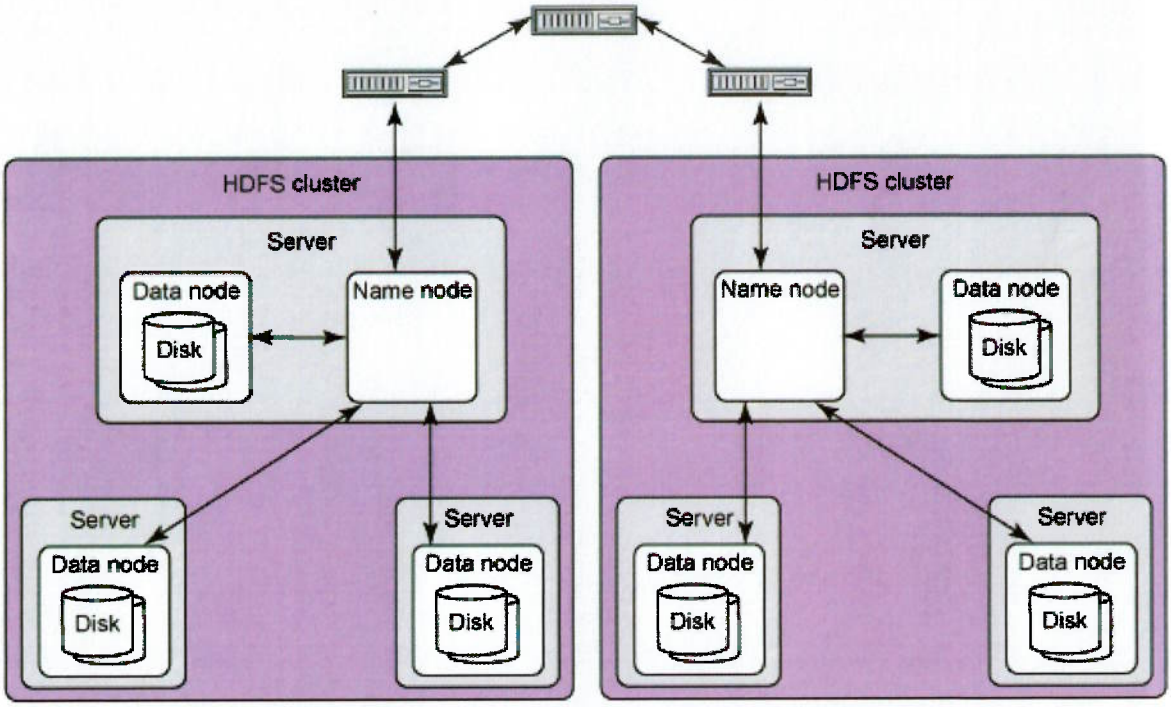

Figura 1: Arquitetura HDFS. Retirada e Adaptada de [16]

O HDFS é projetado para armazenar arquivos muito grandes de forma confiável, onde cada arquivo é armazenado em uma sequência de blocos. Para aumentar a segurança e se manter tolerante a falhas, cada bloco possui três réplicas alocadas em diferentes nós [15]. A figura 03 representa a replicação dos dados no HDFS. Neste exemplo há dois arquivos armazenados, Arquivo A e Arquivo B, de tamanho e formato distintos. Os arquivos são subdivididos em vários blocos antes de serem enviados para os *DataNodes*. O Arquivo A e o Arquivo B foram divididos em 04 blocos, sendo A1, A2, A3 A4 e B1, B2, B3, B4 respectivamente. Para cada bloco dos arquivos foram replicados em 03 unidades, observe que o bloco A1 foi armazenado no *DataNode11, DataNode11 e DataNode22*. Propositalmente, as

réplicas dos blocos são colocadas em *racks* diferentes para garantir a disponibilidade dos dados. A figura abaixo representa a distribuição dos blocos de dados do arquivo A e B divididos nos *racks*:

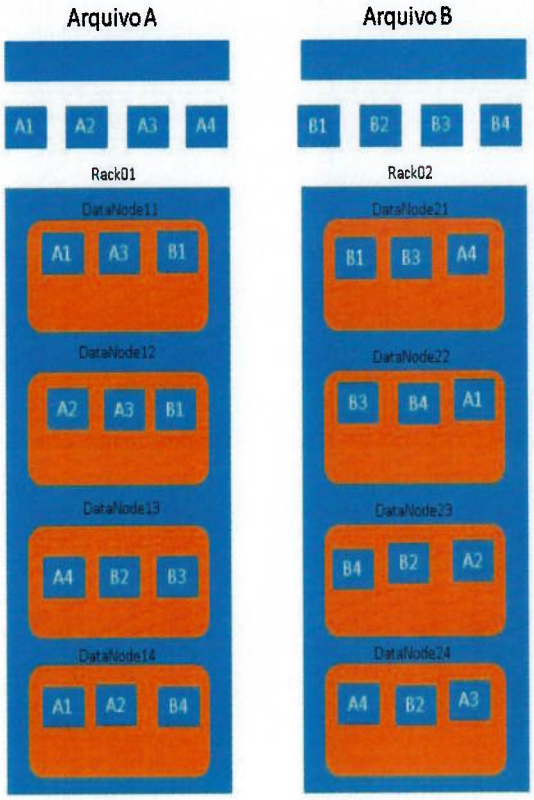

Figura 2: Distribuição dos blocos de dados

Ainda de acordo com a *Apache Hadoop*[15], o *MapReduce* é uma estrutura de software que processa grandes quantidades de dados em paralelo em grandes aglomerados de *hardware commodity*, de forma confiável, tolerante a falhas. Como citado acima, o *MapReduce* possui duas fases: *Map* e *Reduce*. A primeira fase é a fase do mapeamento que acontece o processamento primário dos dados de entrada, ou seja, a função *Map* recebe uma lista como entrada de dados e, aplicando uma função gera uma lista de saída. Um exemplo simples é aplicar a função de multiplicação a uma lista de entrada, dobrando o valor de cada item da lista: *Map* ({2, 4, 6, 8}, (2x)) > {4, 8, 12, 16}

No exemplo de *Map* acima, foi aplicado a função de multiplicação por 2 e gerou a lista de saída {4, 8, 12, 16}. Então os resultados da fase do *Map* são enviados para a função *Reduce* que por sua vez realiza a redução que gerará o

resultado final. O *Reduce* irá receber como entrada o resultado da saída do *Map* e, em geral, irá aplicar uma função para que tenha apenas um único elemento na saída. Exemplo de funções *Reduce* seriam "soma", "media", "max". Aplicando essas funções no exemplo, teríamos o seguinte resultado:

Reduce ({4, 8, 12, 16}, soma) > 40

Reduce ({4, 8, 12, 16}, media) > 10

Reduce ({4, 8, 12, 16}, max) > 16

A figura 03 representa a arquitetura MapReduce:

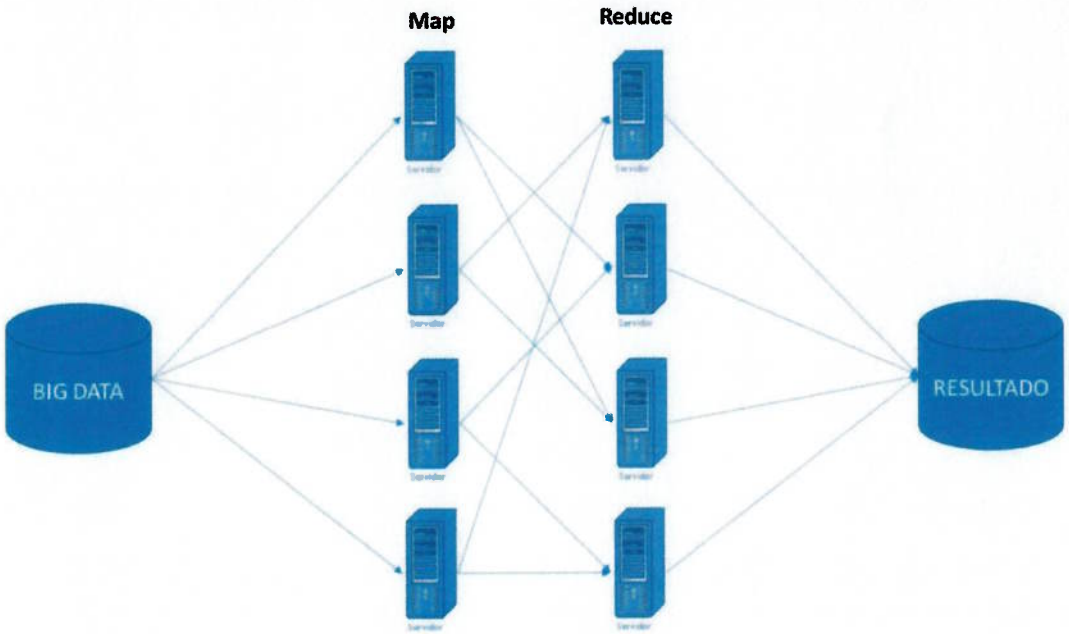

Figura 3: Arquitetura MapReduce

Segundo a *Hortonworks* [14], os benéficos e razões para a organizações utilizarem *hadoop* são:

- Confiabilidade: os dados são replicados em vários nós e caso haja uma falha o processamento é direcionado para o nó ativo.
- Flexiblidade: pode armazenar dados em qualquer formato, incluindo dados estruturados e não estruturados.
- Baixo custo: *Hadoop* é uma estrutura de código aberto e é executado em hardware de baixo custo.

Segundo Taurion [11], a comunidade Apache mantém inúmeros projetos relacionados ao *Hadoop*, como o *Hbase*, que é um banco de dados *NoSQL* que trabalha em cima do HDFS. Atualmente o facebook utiliza o *Hbase* para suportar seus serviços de mensagens instantâneas e informações analíticas. Há também o subprojeto *Hive* que é uma camada de *data warehouse* que executa em cima do *hadoop*. O *hive* facilita a leitura, escrita e gerenciamento de grandes conjuntos de dados que residem em armazenamento distribuído utilizando a linguagem SQL [15].

A tabela 01 apresenta a camada funcional e os subprojetos hadoop:

Camada funcional do Hadoop	Subprojetos
Modelagem e desenvolvimento	MapReduce, Pig, Mahout
Armazenamento e gestão de dados	HDFS, Hbase, Cassandra
Data warehousing e queries	Hive, Scoop
Coleta, agregação e análise de dados	Chukwa, Flume
Metadados, tabela e esquemas	Hcatalog
Cluster Management, job scheduling e workflow	Zookeeper, Oozie, Ambari
Serialização de dados	Avro

Tabela 2: Subprojetos Hadoop

2.6 GOVERNANÇA DE DADOS

De acordo com o BERGSON [17], podemos definir Governança de Dados da seguinte forma:

> *"Governança de dador é o exercício de autoridade e controle (planejamento, monitoramento e engajamento) sobre o gerenciamento de ativos de dados. A função de Governança de Dados guia como todas as outras funções da Gestão de Dados são realizadas. Governança de Dados é de alto nível, ou seja, é gestão estratégica de dados na esfera executiva"*

Hoje em dia ainda há empresas que muitas vezes enxergam os dados como insumos operacionais para realizarem um trabalho pontual, como um projeto novo que possa gerar algum lucro ou até mesmo matéria prima para resolver um problema corriqueiro do sistema. Entretanto deveriam enxergar os dados como componentes preciosos de seu patrimônio e teriam que receber os mesmos cuidados que quaisquer outros bens que tenham alta relevância para a organização. A Governança de dados irá garantir um gerenciamento apropriado e destinado,

assim como processos, monitoramento continuo e toda a arquitetura necessária para que os dados atendam as legítimas necessidades da área de negócio. Segundo o site da Assesso [21], Governança de Dados é uma gestão conjunta de políticas organizacionais, onde serão definidas as regras para realizar a melhor gestão dos dados. São necessários processos eficazes para garantir a agilidade na execução e padronização da produção. Pessoas bem treinadas e capacitadas para executarem os afazeres e alcançar os objetivos estratégicos definidos na política da organização e por último e não menos importante, as tecnologias, que auxiliarão em todas as etapas da governança de dados.

Os quatros itens (Politicas, Processos, Pessoas e Tecnologia) visam estruturar e administrar os ativos de informação, com o objetivo de suportar a tomada de decisão, aprimorar a eficiência operacional e promover a rentabilidade do negócio.

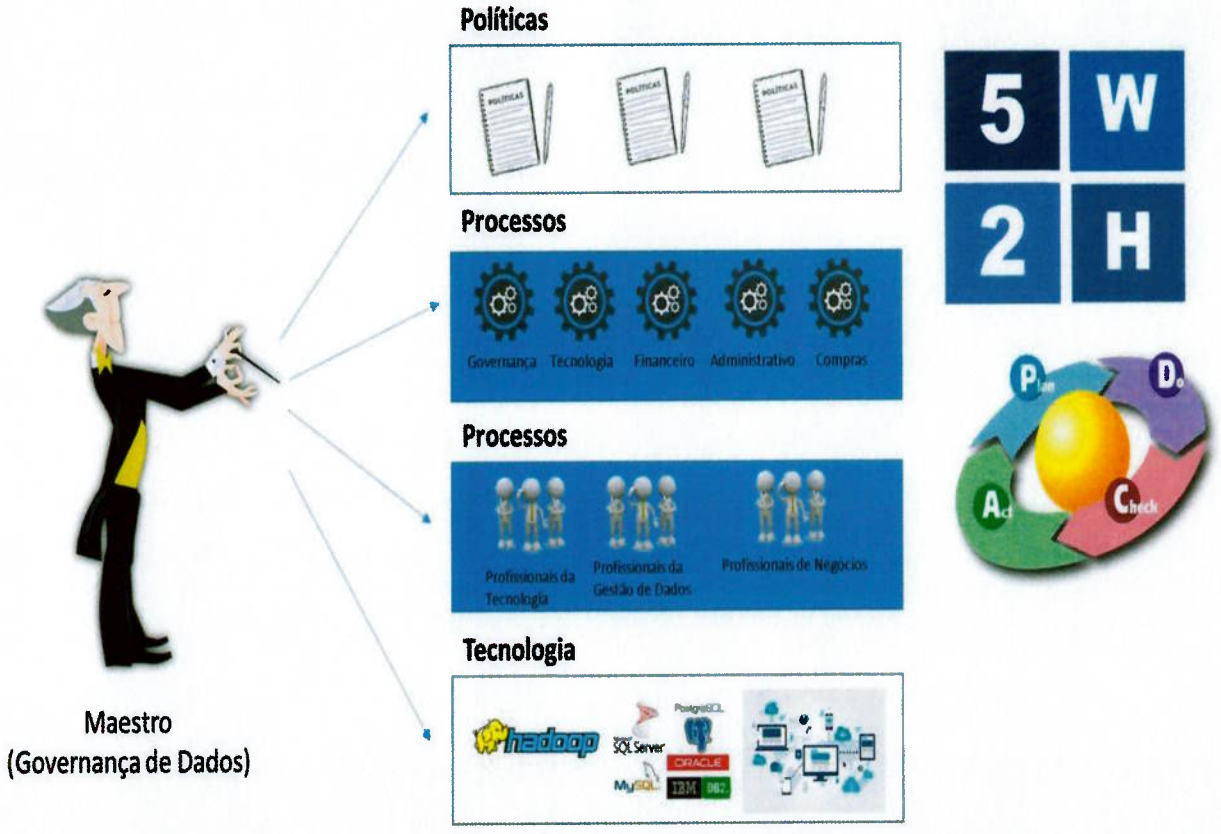

Figura 4: Governança de Dados. Retirado e adaptado de [20].

A ferramenta 5W2H pode ser utilizada na governança de dados para determinar as atividades que necessitam ser desenvolvidas e gerenciadas com o máximo de clareza no ambiente de dados da empresa. Segundo Gustavo Periard [22], através desta ferramenta é possível realizar o mapeamento das atividades e determinar quem será o responsável por cada atividade, o tempo que essa atividade levará para ser finalizada, por qual setor da companhia essa atividade será feita e o motivo desta atividade ser responsabilidade daquele setor. Mas o que vem a ser essa ferramenta 5W2H? A ferramenta foi denominada 5W2H, pois as primeiras letras juntas formam o a sigla das diretrizes do processo.

- WHAT – O QUE SERÁ FEITO (ETAPAS)
- WHY – POR QUE SERÁ FEITO (JUSTIFICATIVA)
- WHERE – ONDE SERÁ FEITO (LOCAL)
- WHEN – QUANDO SERÁ FEITO (TEMPO)
- WHO – POR QUEM SERÁ FEITO (RESPONSABILIDADE)
- HOW – COMO SERÁ FEITO (MÉTODO)
- HOW MUCH – QUANTO CUSTARÁ FAZER (CUSTO)

A ferramenta 5W2H é muito utilizada para resolução de problemas dentro das empresas, porém a mesma é de grande valia quando se deseja planejar atividades, seja algo simples do dia a dia ou até mesmo um projeto de grande porte. Felizmente a ferramenta pode ser aplicada em diferentes cenários e com isso auxiliar em todo processo. No caso da governança de dados, essa ferramenta auxiliará a responder perguntas como:

- O que significa esse conceito de dados?
- Porque esse dado está sendo capturado desta maneira?
- Onde é utilizado esse dado?
- Quando esse dado poderá ser descartado?
- Quem é o gestor de um determinado dado?
- Como consigo acessar esse dado?
- Quanto custa a geração deste dado?

Outra ferramenta que podemos abordar que auxiliará nos processos da governança de dados é o PDCA. Também conhecido como ciclo de Deming, o

PDCA é um processo de melhoria contínua que não possui intervalos e nem interrupções, deste modo, o ciclo de Deming visa melhorar os processos e produtos de forma contínua. A ferramenta PDCA pode ser aplicada em qualquer empresa independente da área ou departamento, e com isso, garantir o sucesso esperado pela alta administração. Abaixo será mostrado o significado da sigla PDCA através da figura 5:

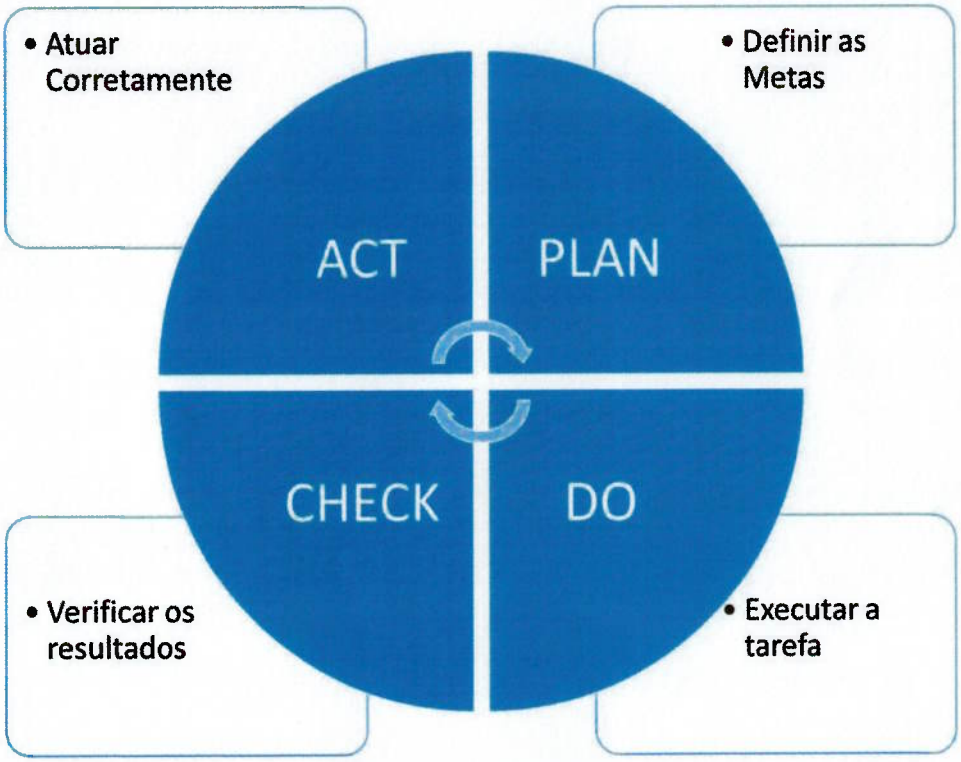

Figura 5: Ciclo PDCA

De acordo com o BERGSON [17], a ferramenta PDCA pode ser aplicada em todo o processo de governança de dados e *Data Quality*.

2.7 QUALIDADE DE DADOS

Atualmente o diferencial competitivo está sendo a capacidade de produzir inovações dos produtos e serviços, gerar e reter conhecimento. Com isso, o dado e a informação estão entrando em evidência sendo considerados recursos estratégico valiosíssimos. Garantir a qualidade dos dados é essencial para que a organização tenha uma visão clara do mundo real que aquele dado representa.

A qualidade dos dados é sempre definida sobre a visão de negócio da empresa, ou seja, os dados possuem qualidade quando eles satisfazem as necessidades para que foram criados [17]. A qualidade de dados necessita da tecnologia para gerar valor para as organizações a partir do armazenamento dos dados, também são necessários processos e pessoas interagindo de forma integrada.

De acordo com Bergson[17], a qualidade de dados pode ser dividida em três pilares:

- Qualidade de metadados.
- Qualidade de conteúdo de dados.
- Qualidade dos processos de gestão de dados.

A figura abaixo ilustra os três pilares da qualidade de dados:

Figura 6: Pilares da qualidade de dados

Metadados são definidos como informações úteis para identificar, localizar, entender e gerenciar os dados, em outras palavras, metadados são os "dados que descrevem os dados" [18]. A qualidade de metadados é executada tanto pelos

técnicos tanto pelo negócio, onde ambos atuam no processo em que os metadados são definidos, checados e manipulados [17]. Segundo Bergson, a qualidade do conteúdo é aplicada em cima de dados produtivos, ou seja, dados que já estão sendo consumidos no ambiente de produção durante o dia a dia. A qualidade do conteúdo é efetivada através de técnicas que como foco identificar dados de baixa qualidade e corrigi-los. A qualidade dos processos monitora tanto os processos de qualidade de dados quanto os de gestão de dados, e utiliza de métricas e indicadores dos processos controlados para tomar ações de melhoria continua [17].

De acordo com a MJV Tecnologia & Inovação [24], diariamente nos deparamos com problemas relacionados a dado de baixa qualidade. Por exemplo, uma correspondência que chegou na residência, porém é de uma pessoa desconhecida, um contato telefônico feito por uma central de atendimento procurando por uma pessoa que antigamente era a proprietária do número telefônico, e-mails com propagandas de produtos ou serviços ao qual você já possui, todos esses exemplos evidenciam a baixa qualidade dos dados que as empresas possuem. Entretanto esses problemas podem estar relacionados a diversas causas: erros sistêmicos, erros de digitação no cadastro, erros na transformação dos dados entre muitos outros motivos. A seguir é apresentado uma tabela com os impactos da baixa qualidade de dados:

BAIXA QUALIDADE DE DADOS	
CATEGORIA DO IMPACO	Exemplos de aspectos para revisão
FINANCEIRO	Perda de oportunidade nos custos • Identificação de clientes com elevado patrimônio líquido • Tempo e custo para a limpeza de dados e o processamento de correções • Avaliação imprecisa do desempenho de funcionários • Incapacidade para identificar fornecedores para análise de gastos
CONFIANÇA	• Maior facilidade de uso por parte da equipe (vendas, central telefônica, etc.) • Maior facilidade de interação para os clientes • Incapacidade de oferecer fatura unificada aos clientes • Decisão prejudicial no estabelecimento de preços
PRODUTIVIDADE	• Menor capacidade de processamento direto mediante utilização de serviços automatizados
RISCO	• A incapacidade de avaliar plenamente o histórico de crédito leva a uma avalição incorreta dos riscos • Dados incompletos levam a uma avaliação incorreta do risco de crédito • Violações da obrigatoriedade de adequação a regulamentos • Violações de privacidade

Tabela 3: Baixa qualidade de Dados. Retirada e adaptada [24]

Como mostrado no quadro acima, a baixa qualidade dos dados pode acarretar em diversos problemas, para BERGSON [17] as principais causas de dados de baixa qualidade são divididas em quatro partes:

Parte 1 - Humanos:

- Falta de treinamento;
- Erros na entrada de dados, seja no momento do cadastro ou da revisão do mesmo de forma não intencional;
- Erros intencionais e maldosos;
- Várias portas de entradas para o mesmo dado;

Parte 2 - Técnico

- Processos mal definidos ou a inexistências dos processos;
- Falta de ferramenta para aferir a qualidade dos dados;

Parte 3 - Organizacional

- Falta de comprometimento da alta direção em conscientizar os funcionários sobre a importância da qualidade de dados;
- Base de dados antigas sem documentações históricas e sem controle;
- Base de dados não integradas;

Parte 4 - Negligência Corporativa

- Não perceber que a qualidade de dados é necessária para todos os setores da organização;
- Não definir quem é o responsável pela qualidade de dados;
- Desconhecimento do valor em melhorar a qualidade de dados;

Nos dias atuais com o crescimento grandioso do número de informações que são armazenadas, há vários processos e soluções de Qualidade de Dados que as empresas podem adotar para tratar os dados. Entretanto para BERGSON [17], há seis processos que são considerados obrigatórios conforme mostra a figura 7:

Figura 7: Processo de Qualidade de dados. Retirado e Adaptado de [17]

Esse processo pode ser feito através de treinamentos, e-mail corporativo e workshops com especialistas da área. Esta fase é de extrema importância para conseguir patrocinadores e *stakeholders* para garantir que tudo corra bem nos próximos processos[17]. De acordo com BERGSON, o processo definir requisitos e a gestão da qualidade irão possibilitar a organização gerar indicadores sobre a qualidade dos dados existentes e com isso estabelecer metas mensuráveis. Segundo BERGSON, há alguns requisitos comuns para a Qualidade de dados:

- **Acurácia:** determina se a entidade real está representada corretamente nos dados.
- **Completude:** determina se os dados estão completos de acordo com as informações exigidas na execução dos processos de negócio.
- **Consistência:** determina a integridade e consistência dos dados.
- **Atualidade:** determina se os dados representam o real estado das informações.
- **Precisão:** determina se os dados representam o grau de precisão necessário. Podemos usar como exemplo o tipo do dado (*double, int, char*) e casas decimais.
- **Privacidade:** indica que o dado é conhecido e está disponível para as pessoas que tem a permissão de acesso.
- **Razoabilidade:** considera como relevante a volumetria esperada de dados para o contexto em que o dado será utilizado.
- **Validade:** o valor dos dados tem que estar em conformidade com os atributos associados aos elementos de dado: tipo do dado, precisão, valores predefinidos e etc.

De acordo com a DataSource [26], O processo de perfilar dados é o processo de conhecer o conteúdo dos dados. Neste ponto é necessário analisar os dados e não apenas confiar nas documentações, modelos de dados ou no perito responsável pela origem dos dados. O processo de perfilar dados, ou também conhecido como *data profiling*, é um diagnóstico da qualidade dos dados. O *Data Profiling* irá mostrar as informações sobre os dados e não as informações a partir dos dados. Segue abaixo alguns exemplos:

- Na coluna NUM_CPF da tabela de clientes não é adicionado o dígito verificador do documento.
- Todos os valores nulos da coluna VALOR_COMPRA receberão o valor padrão "0".
- 40% das entradas de dados na coluna QTD_COMPRA estão com o caractere especial "&".

Segundo BERGSON, o processo de Analisar Dados avalia os dados através das regras de negócios e de métricas. As métricas devem ser quantificáveis e estar muito bem definida e conjunta aos objetivos do negócio. Deve ser determinado como medir e analisar os dados e planos de ações caso seja necessário em determinado ponto dos processos. Ainda de acordo com BERGSON, é na fase de analisar dados que poderá ser decidido qual as prioridades de correções caso haja baixa qualidade dos dados.

O processo de Limpar e Corrigir dados, também conhecido como *Data Cleansing* envolve a detecção e correção de registros incorretos em uma base de dados [17]. O processo de limpar e corrigir dados pode remover redundâncias, remover dados corrompidos e manter os dados consistentes e confiáveis [27]. Segundo BERGSON [17], a limpeza dos dados pode ser feita tanto de forma manual quanto automática. Existem algumas ferramentas e técnicas para fazer o processo de limpeza de forma automática. Uma das técnicas mais conhecidas é a deduplicação de dados, que consiste em identificar os registros duplicados dentro de uma base de dados.

Por fim, o processo de Garantia da Qualidade dos Dados é realizado através de registros das atividades do dia a dia. Neste ponto, são avaliados os indicadores a fim de obter melhorias nos processos [17]. Segundo BERGSON, pode se criar vários indicadores como Qualidade dos modelos de dados, Índices de chamados abertos pelos usuários, Índices de correções de dados, entre inúmeros outros.

Monitorar a qualidade de dados, tomar ações para corrigir e aprimorar os processos, os processos de gestão de dados e a própria qualidade de dados se referem ao conceito melhoria contínua [17].

2.8 METODOLOGIA TOTAL QUALITY MANAGEMENT

Nos últimos anos, a maioria das empresas, tantos grandes quanto as pequenas, iniciaram programas com objetivos que incluem total satisfação para os clientes e sem defeitos de produtos. Estas empresas passaram a enxergar que definir, analisar, medir e melhorar continuamente é essencial para assegurar um alto nível de qualidade dos produtos e serviços, isso não apenas em setores isolados da empresa e sim permeando toda a organização. O *Total Quality Management* é uma técnica formada por um conjunto de programas, ferramentas e procedimentos, aplicados no controle do processo de produção das organizações, para melhorar a qualidade dos produtos e serviços, reduzindo os custos e atendendo as exigências de mercado e provocando o encantamento do cliente [19]. Todavia, direcionando o foco para a qualidade de dados, temos a metodologia *Total Data Quality Management* em 1995. Segundo Batini [23], a TDQM pode ser vista com uma extensão da qualidade total, que foi originalmente proposto para fabricação de produtos.

Segundo o MIT [20], o TQDM foi desenvolvido devido a necessidade das indústrias em ter dados de alta qualidade. O objetivo TDQM é estabelecer sólidos fundamentos teóricos no campo de qualidade de dados, conceber métodos práticos para empresas e indústrias para melhorar a qualidade dos dados que produzem e consomem. De acordo com o MIT [20], os esforços contínuos do grupo de pesquisa do TDQM têm buscado aperfeiçoar a metodologia e especializá-la, isto, para solidificar a disciplina de qualidade de dados, divulgar o impacto de pesquisas e promover colaborações entre universidades, indústria, governo e organizações.

Segundo o site do MIT (*Massachusetts Institute of Technology*), há quatro pilares principais para o programa TDQM:

- Definição de qualidade de dados: este pilar é responsável por definir a qualidade dos dados.
- Medir: tem a responsabilidade de medir a qualidade nos sistemas fornecedores, mantenedores e consumidores da informação
- Análise: tem o papel de identificar o componente e calcular os impactos de dados de má qualidade e os benefícios de dados com alta qualidade para o negócio.

- Melhoria: envolve implementar novas tecnologias e remodelar as práticas de negócio, a fim de melhorar a qualidade dos dados.

Entretanto para Batini [23] os pilares são definidos da seguinte maneira:

- Definição: esta fase inclui a identificação das dimensões de qualidade de dados e requisitos relacionados.
- Medição: produz métricas de qualidade que fornecem feedback do gerenciamento da qualidade dos dados.
- Análise: identifica as causas raízes dos problemas relacionados a qualidade de dados e estuda suas relações.
- Melhoria: estuda e implanta atividades de melhoria de qualidade de dados.

Figura 8: Ciclo TDQM. Retirado e Adaptado de [25]

3 CARACTERIZAÇÃO DA ORGANIZAÇÃO

3.1 CARACTERÍSTICAS DO NEGÓCIO

A empresa em questão, batizada com o nome fictício de Orange é líder no segmento bancário no território nacional. A Orange é um banco brasileiro fundado em 04 de novembro de 2008 mediante a fusão de duas das maiores instituições financeiras do país. A Orange tem mais de 105 mil funcionários atuando em centros administrativos e agências bancárias, espalhados pelo Brasil e em outros 19 países onde todos estão conectados por um mesmo jeito de se relacionar com os públicos do banco e guiados pelo compromisso de fomentar negócios que permitam que todos cresçam. A sustentável visão da Orange é ser um banco líder em performance e em satisfação dos clientes. A empresa personaliza soluções para cada um e promove educação financeira, contribuindo para que as pessoas e empresas tenham melhores relações com o dinheiro. A estratégia de sustentabilidade da Orange está fundada em três pilares: Educação Financeira, Riscos e Oportunidades Socioambientais e Diálogo e Transparência.

Esses temas foram definidos a partir da visão e da cultura corporativa chamada de "Nosso Jeito". A cultura corporativa denominada de "Nosso Jeito" é composta por sete pilares:

- Só é bom para a gente, se for bom para o cliente: Somos pessoas servindo pessoas, com paixão e excelência. Trabalhamos com o cliente e para o cliente, porque ele é a razão maior de tudo o que fazemos.
- Fanáticos por performance: A geração de resultados sustentáveis está no nosso DNA. O desafio constante de buscar a liderança em performance nos trouxe até aqui e continuará guiando a nossa empresa em direção aos nossos objetivos.
- Gente é tudo para a gente: Tudo o que realizamos é por meio de gente. Gente de talento, que gosta de trabalhar em um ambiente de colaboração, meritocracia e alta performance.
- O melhor argumento é o que vale: Cultivamos um ambiente desafiador, aberto ao questionamento e ao debate construtivo. Para nós, a hierarquia que conta é a da melhor ideia.

- Simples. Sempre: Acreditamos que a simplicidade é o melhor caminho para a eficiência. Por isso, lutamos para que a profundidade não se confunda com complexidade e para que a simplicidade não se transforme em simplismo.
- Pensamos e agimos como donos: Pensamos sempre como donos da empresa, liderando pelo exemplo e colocando os objetivos coletivos acima de ambição pessoal.
- Ética é inegociável: Fazemos o que é certo, sem jeitinho, sem atalhos. Exercemos nossa liderança de forma transparente e responsável, totalmente comprometidos com a sociedade e com as melhores práticas de governança e gestão.

A figura abaixo representa a hierarquia da área proposta no estudo da empresa Orange. O nível mais baixo da hierarquia é a posição onde está atualmente o responsável pelo estudo do caso da empresa em questão.

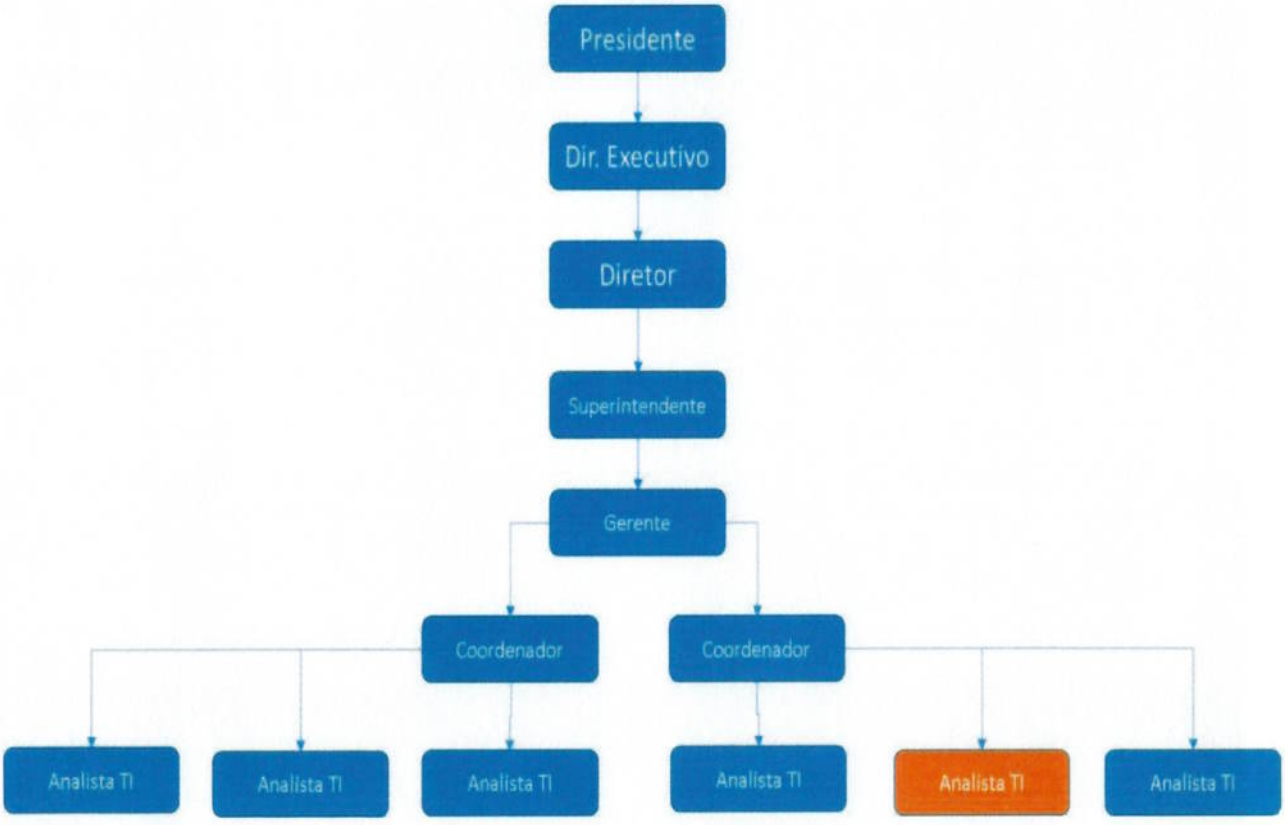

Figura 9: Hierarquia da Orange

3.2 CARACTERISTICA DA OPERAÇÃO

Neste trabalho, o foco será dado para a área de CRM (*Customer Relationship Management*) da empresa Orange. O termo CRM se refere a um conjunto de práticas, estratégias de negócio e tecnologias focadas no cliente que podem utilizar para gerenciar e analisar as interações com seus clientes, antecipar suas necessidades e desejos, otimizar a rentabilidade e aumentar as vendas e a assertividade de suas campanhas de captação de novos clientes. O CRM armazena informações de clientes atuais e potenciais e suas atividades e pontos de contato com a empresa, incluindo visitas a sites, ligações telefônicas, e-mails, entre outras interações.

O processo que será estudado realiza as criações e correções das comunicações digitais para serem disparadas para os clientes, sejam eles correntistas (pessoas físicas ou pessoas jurídicas que possuem conta corrente na instituição), não correntistas (pessoas que não possuem conta corrente na instituição), "cartonistas", ou seja, pessoas que possuem os cartões oferecidos pela instituição, "não cartonistas", são o público que não possui cartões da instituição, mas que tem um vínculo com a mesma, pessoa física ou pessoa jurídica. Os contatos com os clientes, que aqui iremos denominar de campanhas de comunicações digitais, podem ser para ofertar um produto ou um serviço que está sendo oferecido pela instituição ou até mesmo enviar algum comunicado importante para os clientes. As comunicações digitais podem ser disparadas por diferentes canais da instituição, como por e-mail, SMS, *Push*, Mala Direta, *Popup* na tela inicial do *Internet banking*, via aplicativo *mobile* entre outros canais.

Antes de iniciar a explanação sobre o processo proposto, há algumas definições que serão úteis para o bom entendimento de todo o conteúdo a ser apresentado. Será muito utilizado o termo "Variável". Segue abaixo as definições deste termo:

Variável de colagem: algumas *tags* utilizadas apenas para "colar" uma informação dentro de uma campanha, seja ela por e-mail, SMS ou qualquer outro canal. Conforme imagem abaixo, a variável de colagem é *<DESCODPR>* e nela poderão ser "coladas" quaisquer informações do cliente, que neste caso seria a informação sobre o cartão a ser desbloqueado:

Figura 10: Variável de colagem

Variável de seleção: nesta variável ficará armazenado o público ao qual receberá as comunicações cadastradas em uma campanha de Marketing criada pela área de CRM.

Atualmente o processo para criação de uma campanha de CRM se inicia através de um pedido do gestor de negócios. O gestor de negócio é responsável por documentar as solicitações que serão feitas para as demais áreas, os pedidos serão submetidos ao comitê de aprovação e assim que as solicitações forem aprovadas, as mesmas entram no fluxo de trabalho dos times responsáveis para que as campanhas sejam criadas com sucesso. O Time 1 (análise) é responsável por analisar e documentar tecnicamente os pedidos do gestor de negócio. O Time 1 é multidisciplinar, pois conta com analistas da área de negócios, analistas de CRM,

analistas de TI e o envolvimento da equipe de administradores de dados e carga/captura dos dados. O time 1 é responsável por entregar as especificações das variáveis que serão utilizadas nas comunicações digitais e o modelo lógico de dados para as criações das tabelas que irão armazenar as informações dos clientes que serão acionados através das comunicações digitais. É válido ressaltar que o time 1 tem o período de 04 semanas para realizar as atividades e assim dar insumos para o time 2 continuar a evolução da campanha. Segue abaixo as entregas do time 1:

- Modelo lógico de Dados (MLD): Compreende uma descrição das estruturas que serão armazenadas no banco de dados e resulta numa representação gráfica dos dados de uma maneira lógica, inclusive nomeando os componentes e ações que exercem uns sobre os outros. Segue abaixo um exemplo de MLD:

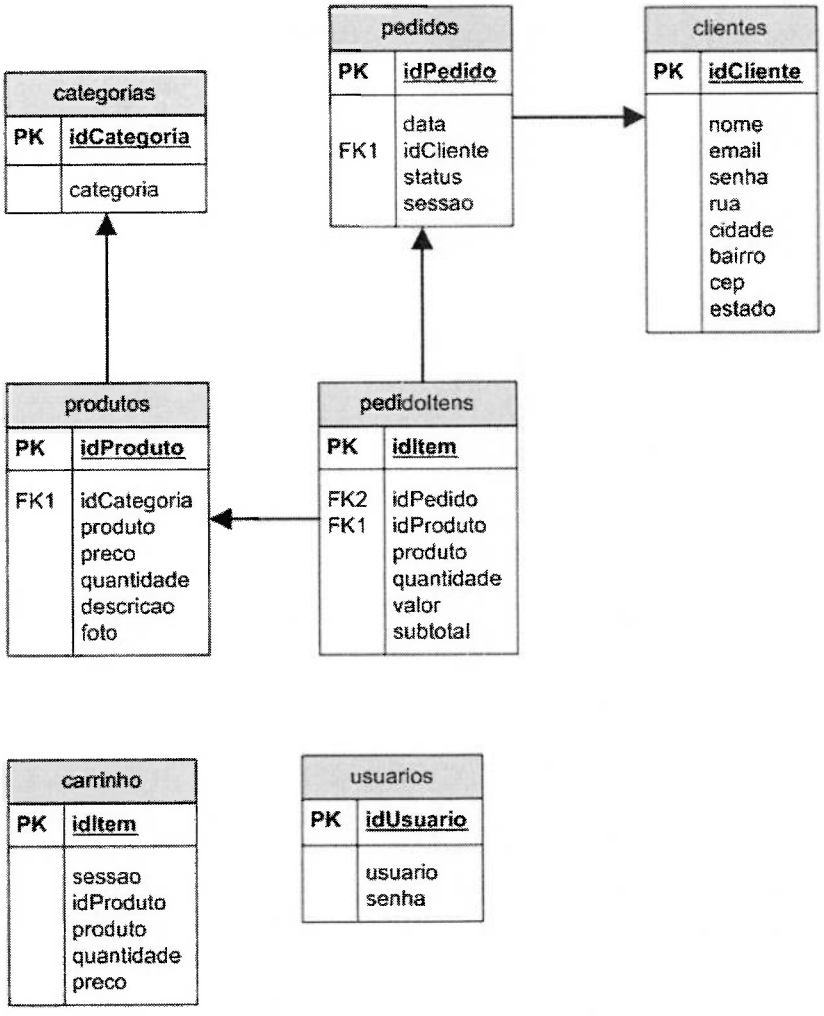

Figura 11: Modelo Lógico de Dados

- Especificações das variáveis: É o documento em que se encontram as regras de negócios para criar as variáveis, as informações sobre as origens dos dados utilizados para criar os *scripts hive*, os campos a serem utilizados nos *scripts hive*, o responsável pela especificação da variável, o nome definitivo da variável e de qual arquivo de origem é o dado que será utilizado.

O time 2 (desenvolvimento) é o responsável pela construção de todos os artefatos que compõem a comunicação digital. Da mesma forma que o time 1, o time 2 também tem 04 semanas para concluir as atividades que são de sua responsabilidade. Essa construção inclui:

- *Scripts hive ou Querys*: são os programas escrito em linguagem *hive* para executarem na plataforma *hadoop*. Estes programas contém os requisitos solicitados pela área demandante que foram refinados e especificados pelo time 1. É através destes scripts que é selecionado o público que será atingido pelas campanhas.
- Rotinas: é um modo de execução que não tem nenhuma interação com o usuário e os *scripts hive* são enfileirados e executados através de um estimulo. Este estimulo pode ser por uma data e horário programado ou até ser estimulado por outra rotina.
- Modelo Físico de Dados: Descreve o modo como os dados são salvos em meios de armazenamentos, como discos e fitas, sendo exigido a definição tanto dos dispositivos de armazenamento físico como dos métodos de acesso (físico) necessários para se chegar aos dados nesses dispositivos, o que o torna dependente tanto de software como de hardware.
- Configuração na ferramenta SAS CI: são as configurações necessárias na ferramenta que é utilizada pela área solicitante para visualização e utilização das variáveis após estar todo o processo concluído.

Por sua vez, o time 3 (CRM) é responsável por homologar todas as variáveis que foram especificadas e refinadas pelo time 1, construídas pelo time 2 em conjunto com a equipe de Sustentação. O time 3 avalia minunciosamente cada variável que foi entregue pelos times anteriores. É analisado se o que foi entregue está de acordo com o que foi especificado pelo time 1, se os volumes de dados

esperados são compatíveis com o que está sendo carregado na tabela, se o tipo do dado que está sendo entregue está de acordo com o esperado pela área solicitante entre outras validações. É nesta etapa que são encontrados os problemas relacionados a qualidade de dados e é neste ponto que o time da Célula de Qualidade entra em ação.

A Célula de Qualidade é a garantia da entrega do que foi construído e implantado pelo time 2, ou seja, qualquer artefato que foi solicitado pela área demandante e que após implantado apresentou alguma não conformidade é de reponsabilidade da Célula de Qualidade tratar este problema.

A célula de qualidade utiliza o PDCA como diretriz para suas atividades:

Figura 12: Ciclo de trabalho da Célula de Qualidade

Caso haja algum erro ou alguma não conformidade com a variável que está sendo entregue, a célula de qualidade irá analisar o problema, realizar as correções necessárias, testar os scripts e configurações ou qualquer outro artefato que foi alterado e realizar a implantação das correções. Após implantado, a célula de qualidade irá checar se o problema foi corrigido com sucesso e posteriormente irá enviar a variável novamente para o time 3 realizar a homologação. Após o time 3 confirmar que entrega foi efetuada com sucesso, a célula de qualidade ficará

responsável por aqueles artefatos por mais 30 dias e após o término deste prazo, todos os artefatos entregues ficam sob a responsabilidade do time de Sustentação.

A equipe de sustentação tem papel primordial em todo o processo de criação de uma variável de campanha, pois é esse time o guardião de todo o ambiente produtivo. Após passar o prazo do acordo de nível de serviço de 30 dias das entregas efetuadas pelos times (time1, time2 e célula de qualidade), fica a cargo da sustentação corrigir e implementar melhorias em todos os sistemas e processos existentes no ambiente produtivo.

A figura a seguir irá exemplificar o modelo de trabalho que é adotado na Orange para criar as campanhas de comunicações:

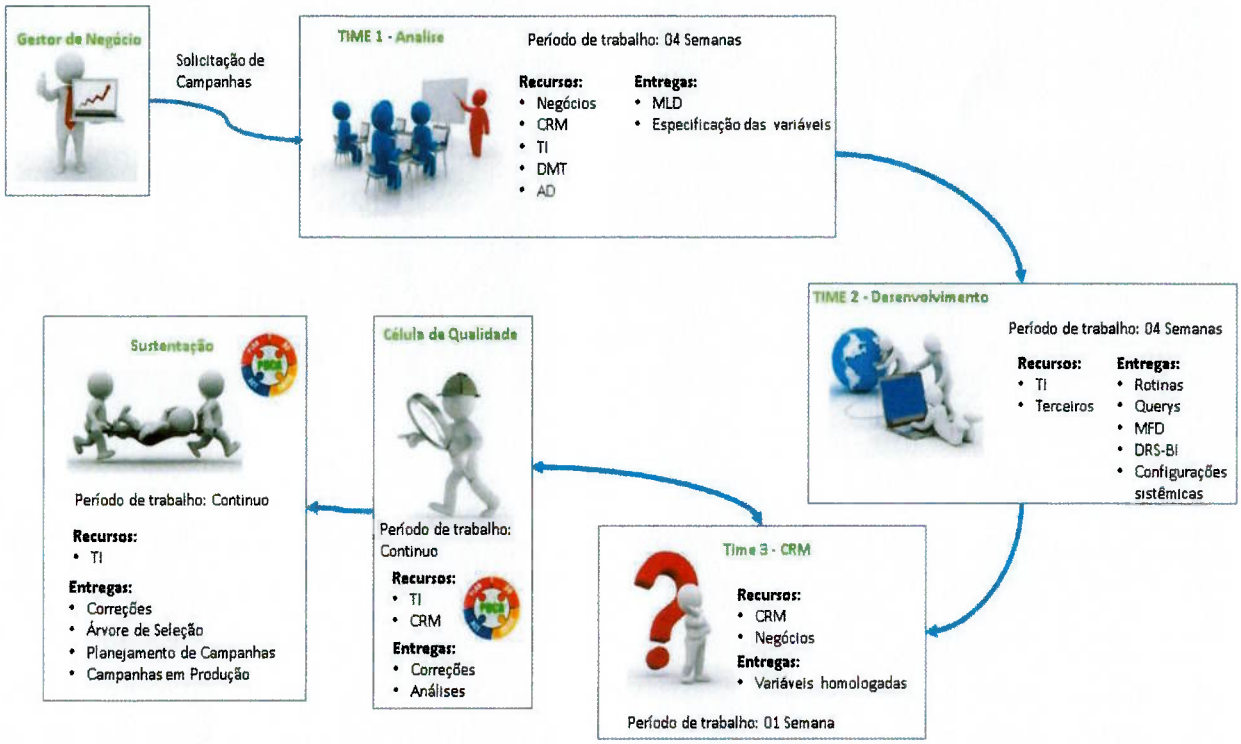

Figura 13: Fluxo de trabalho dos times

De forma simplificada será apresentado como é a estrutura do sistema, desde a origem dos dados até o momento em que a informação chega ao cliente.

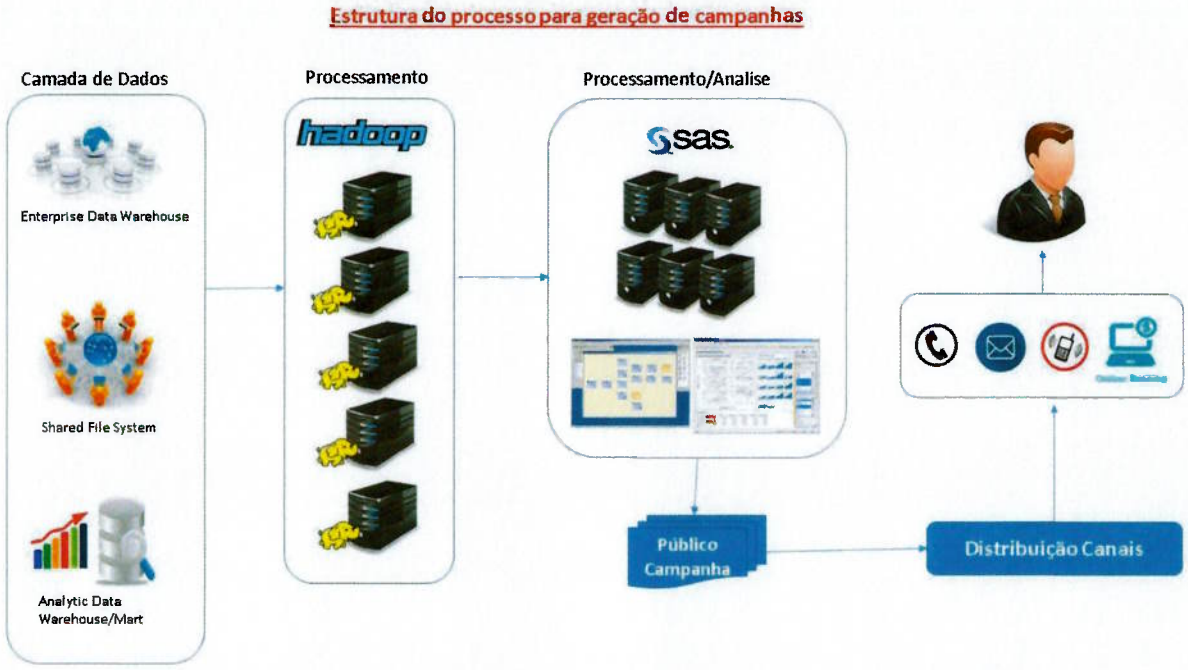

Figura 14: Estrutura do processo para geração de campanhas

O objetivo da camada de dados é prover as informações de toda a instituição para que as mesmas possam ser usadas como insumo para as campanhas planejadas pela equipe de CRM. Os dados são capturados de diversos sistemas de diferentes tecnologias existentes dentro da corporação. Sistemas origens são os responsáveis por gerar e disponibilizar as informações para o CRM efetuar o processamento através do hadoop. Os sistemas origens são os donos das informações geradas através dos canais eletrônicos da instituição. A camada de processamento é onde o dado bruto é manipulado e se transforma em informação para que a área de negócio possa traçar o perfil de cada cliente e direcionar com mais assertividade as campanhas que serão construídas. Nesta etapa do processo, é utilizada a tecnologia *hadoop* para realizar o processamento do dado bruto. Após o *hadoop* efetuar o processamento dos dados disponibilizados pelos sistemas origens, as informações são enviadas para a ferramenta SAS para que a informação seja estudada e para que as campanhas sejam construídas de acordo com as regras estipuladas. Em seguida é gerado o público ao qual se quer atingir com determinada campanha, como dito anteriormente, essas campanhas podem ser disparadas por diferentes canais como SMS, E-mail, Mala direta, *Internet Banking* e etc.

4 ESTUDO DOS CASOS

O presente trabalho tem como objetivo o estudo de alguns casos relacionados a problemas de qualidade de dados encontrados nos processos de criação e manutenção de variáveis de campanhas de CRM que serão disparadas para os clientes. Para os casos que serão estudados, as informações já foram capturadas e processadas na camada de processamento via *hadoop*, ou seja, as variáveis foram processadas e disponibilizadas para o Time 3 – CRM e a equipe detectou algum problema. As variáveis serão analisadas e corrigidas exclusivamente pela equipe da Célula de Qualidade.

Como já citado acima, o time da Célula de Qualidade utiliza o PDCA como diretriz para as atividades, entretanto é valido ressaltar que o diagrama de Ishikawa também é utilizado para auxílio na resolução dos problemas de qualidade de dados encontrados no decorrer do processo. Segue abaixo a figura de como o diagrama de Ishikawa é utilizado nos casos que serão descritos:

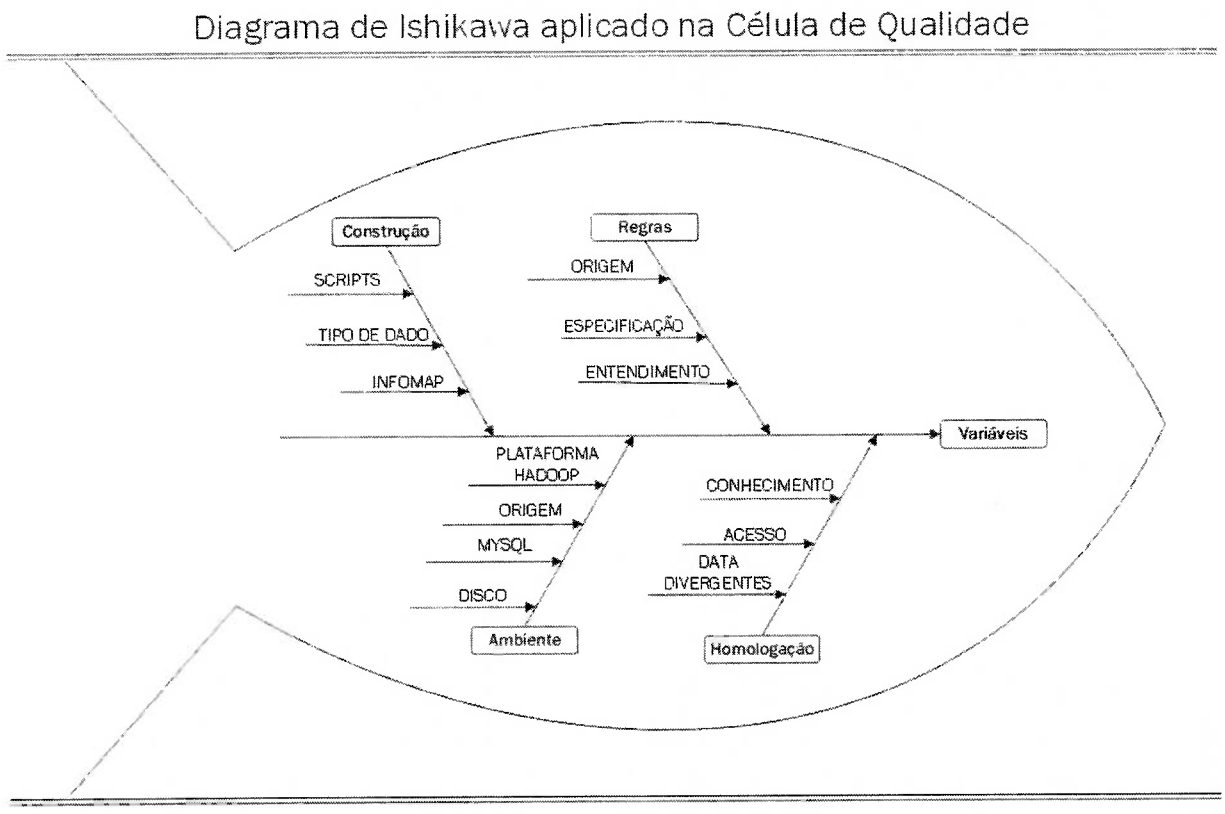

Figura 15: Diagrama de Ishikawa

Como base para início das análises são utilizados 4 possíveis causadores dos problemas:

- Regras: neste item é importante verificar se a regra para a variável está bem definida e entendida por todos os envolvidos. Se a origem dos dados atende fielmente a regra definida pela área de negócio.
- Construção: é verificado se não há erro na codificação das regras descritas pela área de negócio e documentada pelo time1. É verificado também se o dado é o mais atual e se o volume de dados está condizente com o que o demandante espera receber.
- Ambiente: neste ponto é verificado se houve a comunicação correta entre os servidores, se não ocorreu perda de informação durante a disponibilização dos dados para o solicitante e se o dado está sendo disponibilizado de acordo com o planejado pelo Time 2.
- Homologação: este ponto é crítico, pois nem sempre o usuário que solicitou a construção da variável é o mesmo que irá realizar a homologação. Com isso é necessário que todos os envolvidos tenham conhecimento do que vai ser entregue e o que terá que ser validado. É nesta etapa do processo que ocorrem as maiores reclamações a respeito da qualidade do dado, pois a equipe que está verificando a informação não sabe qual o valor exato que a variável deveria ter.

A maioria dos casos que serão descritos a seguir estão fortemente atrelados a pessoas e processos. Atualmente o maior índice causador dos problemas relacionados a qualidade de dados no processo proposto está ligado a pessoas. Os problemas são ocasionados por erros na definição do pedido para criação das campanhas, erro ao documentar o que o gestor de negócios está solicitando, falta de atenção do desenvolvedor do Time 2 para entender e codificar as regras criadas pelo time1 entre inúmeros outros erros.

O PDCA foi a ferramenta escolhida para ser aplicada por ser fácil de utilizar e bastante intuitiva. O PDCA pode ser aplicado praticamente qualquer tipo de projeto, dos mais simples aos mais complexos, já que ajuda a direcionar a equipe para o desenvolvimento de melhorias contínuas, aguça os sentidos para a identificação de falhas e oportunidades de aprimoramento e ainda contribui para que

todos os envolvidos visualizem as mudanças realizadas. O intuito de utilizar o PDCA é aumentar a eficiência dos processos e a produtividade por parte dos times envolvidos (Time 1, Time 2, Time 3, Célula de Qualidade e Sustentação). O PDCA foi considerado de extrema importância neste estudo, pois a ferramenta pode garantir um diagnóstico apurado sobre os processos e tratar os problemas encontrados referentes a qualidade de dados. O PDCA contribuiu de forma positiva para os casos apresentados neste trabalho, pois proporcionou medir o que estava sendo entregue e com que qualidade os artefatos estavam sendo disponibilizados para o negócio. O PDCA também proporcionou estruturar as atividades do time da Célula de Qualidade dentro de suas quatro fases (*Plan, Do, Check, Act*) e com isso aumentaram as chances de atingir o objetivo proposto e melhorar a qualidade dos dados continuamente.

O diagrama de Ishikawa foi escolhido para ser utilizada neste estudo, pois é uma ferramenta que permite agrupar e visualizar várias causas que estão na origem de qualquer problema ou do resultado que se deseja alcançar. O diagrama de Ishikawa por ser uma ferramenta gráfica, fica fácil visualizar em que ponto estão os possíveis erros que contribuem para a baixa qualidade dos dados. Para estudos de casos apresentados abaixo, o diagrama de Ishikawa ajudou a medir e identificar em que ponto do processo eram encontrados os maiores problemas no desenvolvimento das variáveis de CRM, com isso foi possível contabilizar os maiores ofensores dos problemas encontrados. A seguir está a tabela que classificam os processos ofensores:

Classificação do Problemas							
Construção		Regras		Ambiente		Homologação	
Erros na construção dos Scripts	23	Erro na definição da Origem	12	Dado não disponível	2	Falta de conheicmento do Usuário	8
Tipo de dado	0	Especificação errada	10	Plataforma não disponível	2	Falta de Acesso as tabelas	0
infomap errado	0	Entendimento errado	18	Erros MySQL	0	Bases com datas divergentes	5
Total:	**23**	**Total:**	**40**	**Total:**	**4**	**Total:**	**13**

Tabela 4: Classificação dos problemas encontrados pela Célula de Qualidade

Ao todo foram encontrados 80 problemas nas variáveis que estavam passando por homologação no decorrer de 45 dias. Com a ajuda do Ishikawa, foi possível realizar a classificação dos problemas encontrados no processo de criação

de variáveis de campanhas que ocorriam com frequência. Como evidenciado na tabela 4, a coluna Regras é a que possui o maior número de problemas e estes são ocasionados por pessoas, seja por falta de conhecimento do que se está querendo implantar, por erro do analista no momento da especificação da variável ou até mesmo por definição de uma origem de dados pelo analista que está especificando o pedido.

4.1 CASO 1

Neste primeiro caso iremos abordar a variável "Valor Total Tomado Cliente". A origem dos dados que compões essa variável é oriunda do Banco Central, ou seja, as informações são enviadas pelo Bacen para a empresa Orange. Neste arquivo há informações referente a empréstimos que os clientes têm tanto na empresa Orange quanto nas demais instituições financeiras do mercado. Ou seja, o arquivo que está sendo enviado pelo Bancen contém informações de empréstimos na visão Clientes Orange e também a visão dos Clientes Mercado. Este arquivo contém todas as informações referentes a empréstimos dos clientes, ou seja, há informações dos empréstimos vencidos e a vencer.

A variável "Valor Total Tomado Cliente" foi calculada para selecionar clientes que não contrataram um determinado tipo de produto especifico oferecido pela instituição Orange há mais de 06 meses e se o cliente tem esse produto especifico contratado em outras instituições financeiras. Caso o cliente tenha esse produto contratado nas instituições financeiras de mercado e na Orange, é verificado se o valor do produto que o cliente tem contratado no mercado é maior que o valor que o mesmo tem contratado na Orange. Se o cliente estiver dentro das características descritas acima e tem limites disponíveis na conta, esse cliente será abordado através de uma campanha via canais eletrônicos.

O problema referente a qualidade de dados para essa variável foi identificado pelo Time 3 – CRM no momento da validação da variável. O Time 3 identificou que a informação do valor de empréstimos contidos no campo referente a Orange era idêntica para o Mercado, ou seja, a informação estava apresentando que o cliente possuía o valor de empréstimo tomado no Mercado e na Orange completamente iguais.

Cliente	Orange	Mercado	Valor tomado
Mailson	R$ 10.850,30	R$ 10.850,30	R$ 21.700,60
José	R$ 5.000,00	R$ 5.000,00	R$ 10.000,00
Patricia	R$ 554,00	R$ 554,00	R$ 1.108,00

Tabela 5: Exemplo Valores de empréstimos iguais

A célula de qualidade com auxílio do diagrama de Ishikawa, conseguiu identificar que havia alguns erros na construção do script que calculava os valores de empréstimos para os clientes. Na figura 17 está representado através do diagrama de Ishikawa o ponto em que foi identificado o problema de qualidade de dado para a variável em questão.

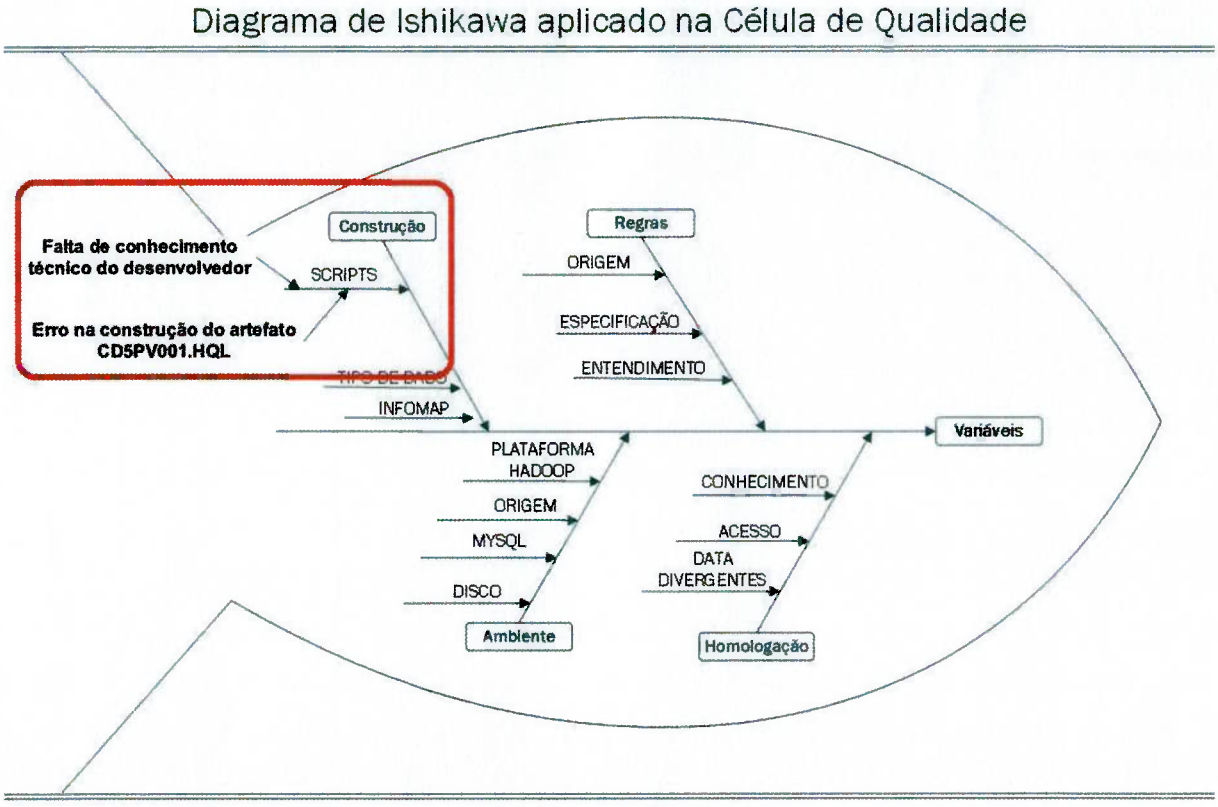

Figura 16: Idenficação do problema

A seguir está a tabela de como o usuário obteve a informação após a correção:

Cliente	Orange	Mercado	Valor tomado
Mailson	R$ 10.850,30		R$ 10.850,30
José		R$ 5.000,00	R$ 5.000,00
Patricia	R$ 554,00		R$ 554,00

Tabela 6: Correção do Valor Tomado Mercado

Observe que este caso, o problema da qualidade de dados ocorreu devido a um erro técnico no momento de codificação do script para gerar o resultado final. Esses erros são menos comuns dentro do processo de criação das variáveis para disparo das campanhas. Após descoberto o problema, a Célula de Qualidade informou o usuário que o problema já havia sido encontrado e que a correção já estava em andamento. Para efetuar a correção foi utilizado o PDCA para auxiliar em todo o processo burocrático para se fazer uma nova publicação no ambiente produtivo. A figura a seguir irá exemplificar como o PDCA utilizado nesta atividade de correção do problema:

Figura 17: Plano de correção da variável Valor Total Tomado Cliente

4.2 CASO 2

Neste segundo caso iremos abordar a variável "Indicador Posse Capitalização Cartão Crédito". Está variável foi construída para selecionar o público que tenha títulos de capitalização contratados ou não, e se não há títulos de capitalização cancelados nos últimos 06 meses. Esse limite de tempo de 06 meses foi incrementado no filtro de seleção de público, para não abordar e aborrecer os clientes que já contrataram ou cancelaram os títulos. O Objetivo desta variável é ofertar campanhas de título de capitalização para clientes que ainda não possuem esse produto atualmente.

O problema de qualidade de dados encontrado nessa variável foi procedente do pedido do gestor de negócio, ou seja, o pedido da criação desta variável foi feito de maneira errada. Quando a variável começou a ser especificada pelo Time1 a área de negócio informou apenas uma fonte de informação que iria conter todo o público desta variável, entretanto eram necessárias duas origens de dados para compor o público esperado. Devido ter sido capturada em uma única fonte de dados, quando a variável chegou para o Time 3 realizar a homologação, o mesmo informou que a volumetria do público esperado estava baixa. Após um longo período de análise pela equipe da Célula de Qualidade, foi identificado que seria necessário mais uma origem para compor toda a informação esperada.

A seguir será apresentado a figura do diagrama de Ishikawa com destaque no item que em foi identificado o problema:

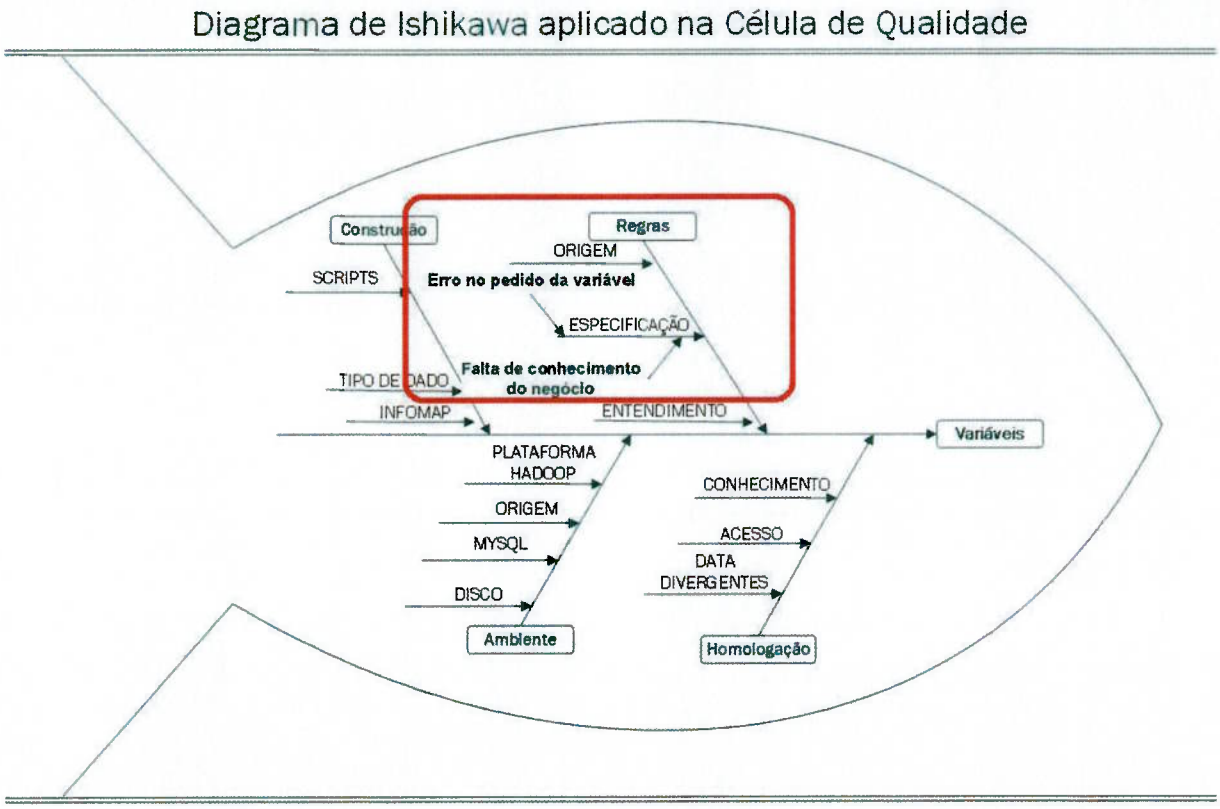

Figura 18: Erro no pedido da criação da variável

O erro de qualidade de dados foi ocasionado por um erro na definição do que seria necessário para construir aquele artefato. Quando ocorre erro de definição da variável ou problemas na especificação da mesma, o solicitante da variável é informado que a solicitação para criação da variável "Indicador Posse Capitalização

Cartão Crédito" deverá retornar para o Time 1, pois será necessário refinar a especificação.

4.3 CASO 3

Neste terceiro caso iremos abordar a variável "Código Situação Cadastro Biometria". Esta variável é calculada para selecionar todo o público que possui e não possui o cadastramento da biometria como dispositivo de segurança. Está variável será utilizada para abordar os clientes que ainda não possuem a biometria de segurança e com isso disparar campanhas de marketing através dos canais eletrônicos incentivando e mostrando as vantagens de se ter a biometria cadastrada. O Time 3 relatou que os clientes que estavam selecionados não estavam corretos, pois haviam muitos clientes que já tinham a biometria cadastrada e estavam sendo indicados como não cadastrada pelo cálculo feito na variável. Após um período de análise da equipe da Célula de Qualidade, foi constatado que a origem dos dados que o Time 1 havia indicado para serem utilizados na construção da variável não eram a melhor fonte de informação sobre o cadastro de biometria do cliente. Neste caso foi possível identificar dois erros que ocasionaram o problema na qualidade de dados.

- Primeiro: a falta de conhecimento das origens dos dados dos demais sistemas da instituição financeira fez com que o artefato especificado pelo Time 1 fosse entregue para o cliente com algumas não conformidades.
- Segundo: a instituição financeira ter duas fontes de informação para o mesmo tipo de dado.

A figura a seguir ilustra através do diagrama de Ishikawa onde os problemas referentes a qualidades de dados foram encontrados para o Caso 3.

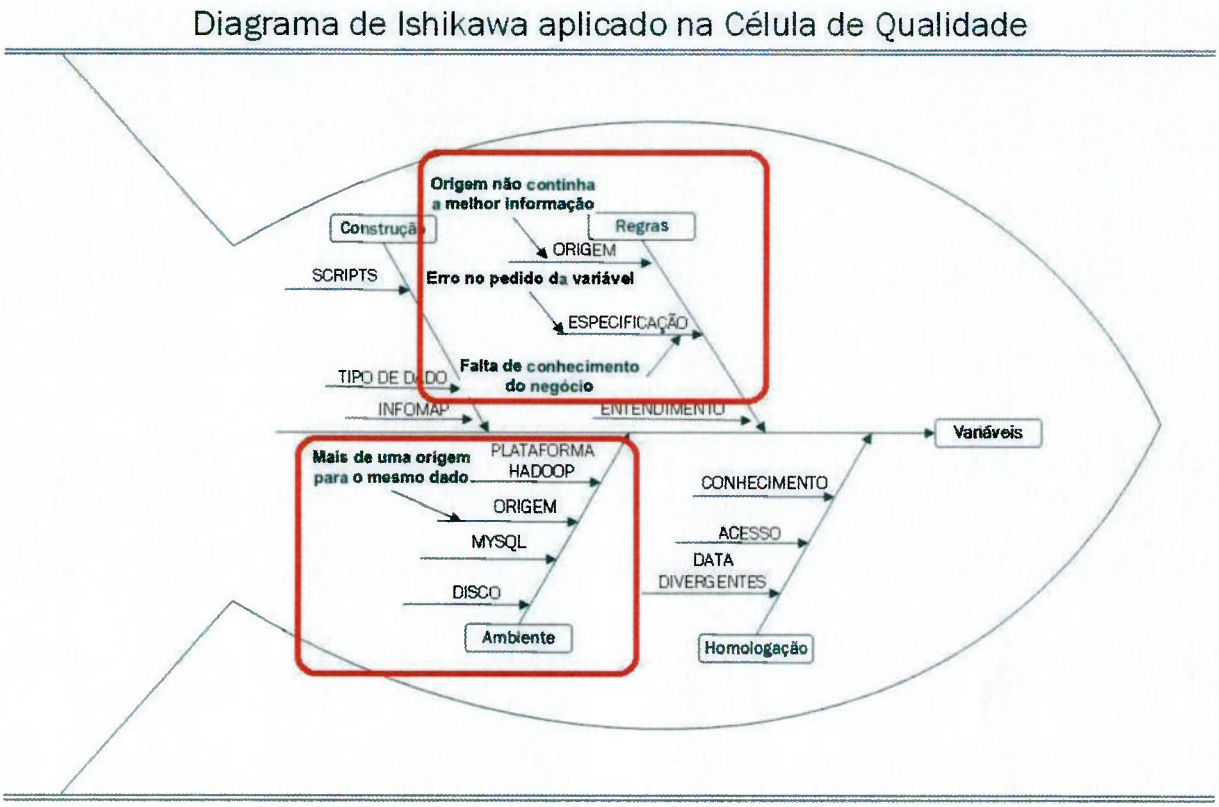

Figura 19: Caso 3 com dois problemas identificados

A equipe da Célula de Qualidade com auxílio do PDCA corrigiu o problema de qualidade de dados realizando a troca da origem da informação e documentando o ocorrido. Para realizar a troca da origem da informação, é necessário realizar uma série de atividades para chegar ao ponto de disponibilizar o artefato para consumo do usuário. A figura a seguir irá exemplificar como o PDCA utilizado nesta atividade de correção do problema:

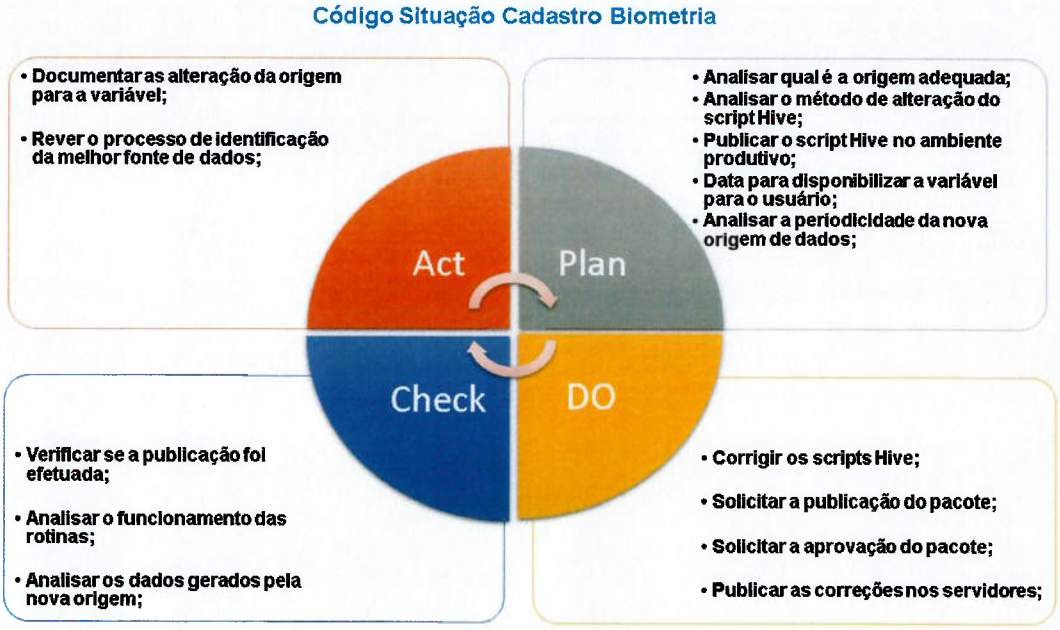

Figura 20: Plano de correção da variável Códico Situação Cadastro Biometria

4.4 CASO 4

Neste quatro e último caso é abordado a variável "Percentual Pagamento Fatura Cartão Crédito Último Mês". O objetivo dessa variável era atingir o público cartonista, ou seja, os clientes que utilizam o cartão de crédito da instituição. A seguir será descrita a regra para selecionar o público desta variável:

Verificar se o cliente pagou mais de 75% do total da fatura de todos os cartões de créditos somados. Caso o cliente tenha pago menos de 75% e tenha limite de crédito disponível maior que o valor que sobrou para quitar o restante da fatura, o cliente será abordado por uma campanha de marketing via canais eletrônicos da instituição. Os clientes que pagaram mais de 75% não serão abordados. Essa variável foi construída e apresentou na etapa de homologação inúmeros erros de qualidade de dados como:

- Origens incorretas;
- Falta de entendimento do conceito de utilização da variável;
- Falta de conhecimento para homologar a variável;

A figura a seguir ilustra através do diagrama de Ishikawa onde os problemas referentes a qualidades de dados foram encontrados para este caso:

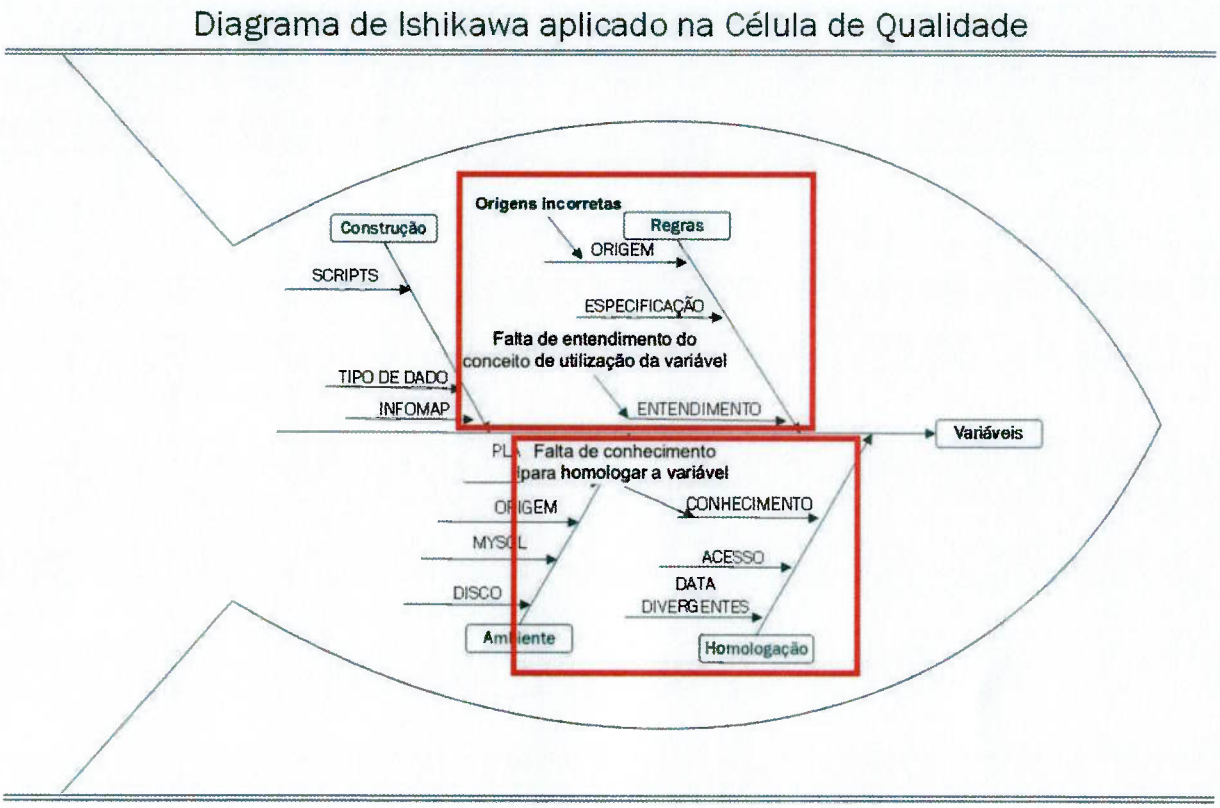

Figura 21: Problema conceitual da variável

O problema desta variável ocorreu na sua concepção, pois o conceito deste artefato estava errado. O conceito com a qual a variável foi solicitada e construída gerava como resultado final uma visão geral do cliente com a soma de todos os cartões de crédito, ou seja, olhamos a porcentagem de pagamento da fatura do cliente de forma unificada. O correto neste caso seria analisar os pagamentos de cada cartão do cliente de forma individual, ou seja, para determinado cliente seria necessário identificar qual cartão não foi pago mais de 75% do valor da fatura e a partir deste ponto realizar uma abordagem mais assertiva e personalizada.

Novamente enfrentamos problemas de qualidade de dados devido a uma falha humana na definição do pedido do componente. O gestor de negócio solicitou a correção da variável e a mesma entrou no início do fluxo de criação.

5 CONSIDERAÇÕES FINAIS

Este capítulo descreve as considerações finais do trabalho, assim como apresenta a lista de melhorias a serem realizadas no processo estudado. As conclusões deste trabalho foram obtidas a partir do confronto das falhas encontradas no processo de criação de variáveis com a aplicação das ferramentas de qualidade descritas nesta monografia.

5.1 CONCLUSÃO

Historicamente, as instituições financeiras sempre foram reconhecidas por serem detentoras de muita informação e por diversas formar de utilizar esse volume de informação em diversos processos, como ofertar produtos via campanha de marketing, para decidir para quem deverá ser aumentado o limite de crédito e até mesmo identificar como está a saúde financeira de cada cliente. Atualmente a Orange atua em diversos países e possuem mais de 50,5 milhões de clientes que transacionam diariamente, seja para fazer um saque nos caixas eletrônicos, pagando contas na internet, aplicando em investimento e até mesmo ligando nas centrais de atendimento. Todas essas transações geram um enorme volume de dados diversificados que podem ser utilizados para diferentes fins.

O Big Data tornou-se fundamental para as instituições financeiras, pois a quantidade de dados disponíveis nessas instituições possibilita a elaboração de novos modelos de negócio e trilhar com mais exatidão o plano de estratégia competitiva no mercado, que está cada vez mais acirrado. Ao contrário do que acontecia no passado, agora na era do Big Data os dados podem ser acessados a qualquer momento e são armazenados em *commodityes*, o que torna o custo do armazenamento administrável. O Big data está apoiado em três pilares que são Tecnologia, Pessoas e Processos. Como demonstrado nos casos acima, o maior desafio encontrado está no pilar de Pessoas. Os colaboradores necessitam de mais qualificações e fazer parte de times que sejam multidisciplinares para cumprir os desafios demandados neste novo cenário.

Todo projeto que envolva dados deve estabelecer alguns compromissos de qualidade para que sua implantação se torne um caso de sucesso. Para isso, deve-se utilizar metodologias como o TDQM que incrementem um nível de qualidade no

gerenciamento e na entrega dos dados, mas também, a definição e acompanhamento da aplicação de técnicas de qualidade durante a continuidade do sistema alvo. O uso da metodologia TDQM foi essencial ao ajuste dos dados de baixa qualidade encontrados nos processos de homologação e correção. A forma simples e estruturada de utilizar o PDCA permitiu padronizar os processos de correção para que as boas práticas encontradas na criação de variável se repetissem constantemente.

A ferramenta de qualidade Ishikawa proporcionou enxergar os pontos mais ofensores de todo o processo de criação das variáveis de campanha, com isso foi possível dar maior atenção e recapacitar os envolvidos para que os erros fossem mitigados em cada iteração do ciclo de vida de criação dos artefatos. Através o Ishikawa foi possível criar métricas mensais de todos os ofensores do processo e essas métricas definiu os processos prioritário em uma ação de melhoria.

5.2 LISTA DE MELHORIAS A SEREM REALIZADAS NO PROCESSO

O uso do TDQM junto com as ferramentas Ishikawa e PDCA em um exemplo real apresentado nesta monografia teve seu foco na aplicação de qualidade de dados em um sistema já existente. Com base nos resultados obtidos e no desempenho apresentado pelo processo, pode-se listar algumas melhorias a serem implantadas no fluxo de criação de variáveis:

- Refinar o conceito da variável antes de solicitar a criação da mesma;
- Especificar detalhadamente a variável a ponto de qualquer pessoa entender o que está sendo solicitado;
- Envolver uma equipe de qualidade para garantir a entrega das variáveis;
- Realizar teste minuciosos antes de efetuar a implantação dos scripts em ambiente produtivo;
- Realizar novos testes após a implantação e antes da variável ser disponibilizada para o usuário (Time 3);
- Definir metas de qualidade para o processo;

REFERÊNCIAS BIBLIOGRÁFICAS

[1] **SAS What is big data.** Disponível em: http://www.sas.com/pt_br/insights/big-data/what-is-big-data.html. Acesso: 17/05/2016

[2] **Sabe o que é Big Data? Já ouviu falar em 5v big data?.** Disponível em: <http://datastorm.com.br/5v-big-data-estrutura/>. Acesso: 17/05/2016

[3] **Você realmente sabe o que é Big Data?.** Disponível em: https://www.ibm.com/developerworks/community/blogs/ctaurion/entry/voce_realmente_sabe_o_que_e_big_data?lang=en. Acesso: 17/05/2016

[4] SETZER, V. W. **Dado, Informação, Conhecimento e Competência.** Disponível em: <http://www.ime.usp.br/~vwsetzer/dado-info.html>. Acesso em: 17/05/2016.

[5] DAVENORT, T. H.; PRUSAK, L. **Conhecimento empresarial: como as organizações gerenciam o seu capital intelectual.** Rio de Janeiro: Campus, 1998

[6] ROSINI, Alessandro Marco, PALMISANO, Angelo. **Administração de Sistemas de Informação e a Gestão do Conhecimento.** Ed. (2003).

[7] LAUDON, Kenneth C.; LAUDON, Jane Price. **Sistemas de informação.** 4. ed. LTC: Rio de Janeiro,1999.

[8] **Quais os gurus da qualidade e suas ferramentas.** Disponível em: http://www.totalqualidade.com.br/2012/09/quais-sao-os-gurus-da-qualidade-e-suas.html. Acesso: 17/05/2016

[9] **What is Apache Hadoop.** Disponível em: http://hadoop.apache.org. Acesso: 30/05/ 2016

[10] **Estudo da Xerox destaca a importância do big data para o sucesso das empresas.** Disponível em: http://www.xerox.com/news/news-archive/2016/prt-estudo-da-xerox-destaca-a-importncia-do-big-data-para-o-sucesso-das-empresas-/ptmz.html. Acesso em: 17/05/2016

[11] TAURION, Cesar. **Big data.** Editora Brasport. Rio de janeiro 2013.

[12] **Big Data na educação: como é usado nas escolas ao redor do mundo.** Disponível em: http://datastorm.com.br/como-o-big-data-esta-sendo-usado-pelas-escolas-ao-redor-do-mundo/. Acesso em 13/07/2016

[13] **Waze.** Disponível em: https://www.waze.com/pt-BR. Acesso em 13/06/2016

[14] **Big data em código aberto: um ecossistema de projetos.** Disponível em: https://br.hortonworks.com/apache/. Acesso em 13/07/2016

[15] **HDFS Architecture Guide.** Disponível em: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html. Acesso em: 17/07/2016

[16] **Uma Introdução ao Hadoop Distributed File System**. Disponível em: http://www.ibm.com/developerworks/br/library/wa-introhdfs/. Acesso em: 19/07/2016

[17] BERGSON, L.R **Gestão e Governança de Dados: Promovendo dados com ativo de valor nas empresas.** Editora Brasport. Rio de Janeiro 2013.

[18] **Banco de Metadados.** Disponível em: http://www.metadados.ibge.gov.br/consulta/default.aspx. Acesso em: 28/07/2016

[19] DEMING, E. W. **Out of the Crisis.** 1. ed. Massachussets: MIT, 1986.

[20] **MIT'S Total Data Quality Management (TDQM).** Disponível em: http://web.mit.edu/tdqm/www/. Acesso: 11/05/2016

[21] **Governança de Dados.** Disponível em: http://www.assesso.com.br/governanca-de-dados. Acesso: 11/10/2016

[22] **O que é 5W2H e como ele é utilizado.** Disponível em: http://www.sobreadministracao.com/o-que-e-o-5w2h-e-como-ele-e-utilizado/. Acesso: 12/10/2016

[23] Batini, Carlo; Scannapieco, Monica. **Data and Information Quality. Dimensions Principles and Techniques.** Ed. Springer, 2016.

[24] **Data Quality Aplicado a Business Inteligence.** Disponível em: http://www.mjv.com.br/wp-content/uploads/2013/10/Apresentacao-Data-Quality.pdf. Acesso em: 18/10/2016

[25] **TDQM - Total Data Quality Management.** Disponível em: <http://www.datasetting.com.br/produtos/tdqm-total-data-quality-management> Acesso em: 27/10/2016

[26] **Data Profiling: What, Why and How?**. Disponível em: < http://ds.datasourceconsulting.com/blog/data-profiling/>. Acesso em: 06/11/2016

[27] **Data cleansing made easy.** Disponível em: <https://www.improvemydata.com/data-cleansing>. Acesso em: 06/11/2016