UNIVERSIDADE DO VALE DO RIO DOS SINOS – UNISINOS

MBA EM ADMINISTRAÇÃO DE TECNOLOGIA DA INFORMAÇÃO

FELIPE RENZ

COMO BIG DATA PODE AUXILIAR AS EMPRESAS A GERAR VANTAGEM COMPETITIVA

ESTUDO DE CASO ÚNICO EM EMPRESA DA INDÚSTRIA AGRÍCOLA

SÃO LEOPOLDO

2014

Felipe Renz

COMO BIG DATA PODE AUXILIAR AS EMPRESAS A GERAR VANTAGEM COMPETITIVA

Estudo de caso único em empresa da indústria agrícola

Trabalho de Conclusão de Curso apresentado como requisito parcial para a obtenção do título de Especialista em Administração de Tecnologia da Informação, pelo MBA em Administração de Tecnologia da Informação, da Universidade do Vale do Rio dos Sinos.

Orientador: Prof. Dr. Stanley Loh

São Leopoldo

2014

Dedico este trabalho aos meus filhos Nicolas e Isadora, que apesar de pouco entender o que estava acontecendo, quando o pai chegava tarde porque havia voltado a estudar, foi fazer um MBA que vem mudando a cada dia a maneira de pensar e enxergar o mundo, tanto nos negócios como nas pessoas, além do crescimento que trouxe para minha vida.

AGRADECIMENTOS

Deixo nesta página, os meus sinceros agradecimentos a todos aqueles que, direta ou indiretamente, contribuíram no desenvolvimento deste trabalho, bem como na conquista do título acadêmico.

Ao meu orientador, prof. Dr. Stanly Loh, pela sua disposição em me orientar, pelas recomendações de leitura, por compreender meus interesses e limitações.

Ao prof. Dr. Oscar Rudy Kronmeyer Filho pela aceitação no MBA, e sua didática fantástica demostrada em sala.

Ao prof. Dr. Jerônimo Lima pelos ensinamentos não apenas acadêmicos, mas na maneira de enxergar e pensar.

A minha esposa, Ruti, pelo apoio e compreensão, amor e carinho, dedicação eterna à família.

Aos meus filhos, Isadora e Nicolas, pelos sorrisos puro e espontâneos, uma verdadeira fonte de energia para prosseguir.

Aos meus pais, João e Iene, pelo exemplo de vida.

A AGCO, objeto da pesquisa, pela permissão e acesso às informações.

A todos meus amigos e colegas do MBA, que fizeram parte desta importante etapa da minha vida, e aos meus colegas de empresas, que muitas vezes me ouviram e expressaram; suas opiniões. Não ousarei citar nomes, pois posso cometer injustiça ao esquecer alguém.

ABSTRACT

This work has the objective to perform a case study within an agricultural machinery organization, It's regarding the use of Big Data solution associated with the predictive analysis in order to foresee when an agricultural machine will be damaged, so that a unscheduled stop occurs, avoiding losses to its customers and generating competitive advantages against market competitors. It was used several bibliographical references from many different authors whose studied, developed and applied the matter in a thorough way. It was presented a strong theoretical basis regarding Big Data solution, thoroughly exploring the 5V's solution. Through this work it was proposed an architectural model for Big Data usage, as well as the verification of its organization acceptance. The data survey was made through interviews with managers, by gathering their knowledge and expertise in the system involved and applied in this case study.

Keywords: Big Data. 5V ' s. Predictive Analysis. Competitive Advantage. Architecture. Technology and innovation.

RESUMO

Este trabalho tem por objetivo realizar um estudo de caso em uma organização do ramo de máquinas agrícolas, sobre a utilização de uma solução de Big Data associada a análise preditiva com o intuito de prever quando uma máquina agrícola for estragar, para que uma parada não programada ocorra, evitando prejuízos aos seus clientes, e geração de vantagem competitiva frente à concorrência. Para tanto foram utilizadas várias referências bibliográficas oriundas dos mais diversos autores que de uma forma ou outra estudaram, desenvolveram e aplicaram o assunto de uma maneira aprofundada. Foi apresentado um forte embasamento teórico quanto à solução de Big Data, explorando detalhadamente cada um dos 5V's utilizados na solução. Por meio deste trabalho realiza-se a sugestão de arquitetura para utilização de Big Data, e verificação de sua aderência com a organização. O levantamento dos dados ocorreu através de entrevistas, envolvendo colaboradores e gestores com conhecimentos nos sistemas envolvidos no estudo de caso.

Palavras-chave: Big Data. 5V's. Analise Preditiva. Vantagem Competitiva. Arquitetura. Tecnologia e Inovação.

LISTA DE QUADROS

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

3V’s	Volume, Velocidade, Variedade
4V’s	Volume, Velocidade, Variedade, Veracidade
5V’s	Volume, Velocidade, Variedade, Veracidade, Valor
AGCO	Allis-Gleaner Corporation
BOM	Build of Material
ERP	Enterprise Resource Planning
GPS	Global Positioning System
HDFS	Hadoop Distributed File System
HiveQL	Hive Query Language
HPIL	Hadoop Physical Infrastructure Layer
IDC	International Data Corporation
JDE	JD Edwards
MDM	Master Data Management
NoSQL	Not Only Structured Query Language
ODP	Ordem De Despacho
PDI	Pré-Delivery-Inspection
RFID	Radio-frequency identification
ROP	Registro de Ocorrência de Peças
SQL	Structured Query Language
TI	Tecnologia da Informação
XML	Extensible Markup Language
YARN	Yet Another Resources Negotiator

SUMÁRIO

1 INTRODUÇÃO

Tole (2013) afirma que os cenários no mundo econômico e científico sofreram mudanças nos últimos anos através de complexos métodos que asseguraram sua evolução, no intuito de melhorar a eficiência em produtos e serviços. Para que fosse possível essa evolução uma grande quantidade de dados foi necessária, possibilitando que informações valiosas fossem extraídas.

Devido sua importância dados são hoje considerados o petróleo do século XXI, (HEY, TANSLEY e TOLLE, 2011), porque a partir deles podemos gerar informações relevantes para os negócios e quando transformados em informações e utilizados de maneira eficiente tendem a prover uma melhora nos resultados das organizações.

Nos dias atuais a área de Tecnologia da Informação (TI) está cada vez mais exercendo um papel fundamental nas organizações, estando ela diretamente ligada à sustentabilidade do negócio. Outro fator importante é o aumento na quantidade de dados, podendo ser gerados por pessoas e, ou máquinas que criam um enorme volume de dados e que quando compilados transforma-se em informações. Para este processo de transformação utiliza-se a ciência dos dados também chamada de e-Science (TAURION, 2013).

A TI estando tão próxima ao negócio, tende a prover inovação para que cada vez mais se possam identificar oportunidades que tendem a gerar vantagem competitiva perante seus concorrentes. Para isso deve-se utilizar a maior quantidade de informações possíveis, nesse contexto temos a inserção do Big Data, que auxilia nesse desafio, através dos seus "4Vs" (volume, variedade, velocidade, valor) para Mayer-schönberger & Cukier (2013) ou na ótica de Taurion (2013) os "5Vs" (volume, variedade, velocidade, veracidade gerando valor).

Existe muita variedade na geração de dados o conjunto de multiplicação dos sensores espalhados por diversos dispositivos que iniciam com computadores e se expandem para celulares, tablets, GPS, etiquetas de RFID, somando ao aumento da utilização das Redes Sociais, gera um volume assombroso de dados. Por isso é importante que no meio desse emanharado de dados sejam retirados os "lixos", criando assim veracidade.

Onde Taurion (2013) afirma que para poder lidar com todos esses dados será necessária uma grande velocidade de processamento, processando dados estruturados ou não estruturados, mas isso de nada adianta se não puder gerar valor, por meio da diferenciação dos concorrentes, abertura de novos negócios e ou mercados.

Para Taurion (2013) este valor pode se dar por varias maneiras, novos negócios, novos modelos de negócio, retenção de clientes, melhora da imagem perante o mercado, diferenciando-se da concorrência com maior eficácia e qualidade ou com exclusividade estabelecendo assim vantagens de mercado conforme Kaplan e Norton (2008).

Esta quantidade de dados possibilita trabalhar em varias frentes para aprimorar os negócios, criar cidades inteligentes, auxiliar no controle de doenças, de uma maneira geral tende a ajudar a melhorar o mundo, neste contexto existe a técnica de analise preditiva, que está fundamentada nas correlações entre os dados (MAYER-SCHONBERGER & CUKIER, 2013).

Para que se consiga prover uma solução de Big Data capaz de atender as necessidades de velocidade, variedade e volume, somente com a evolução das tecnologias, fato esse que possibilitará a geração de valor a partir dos dados disponíveis nas organizações (SAWANT e SHAH, 2014).

1.1 Situação problemática e pergunta de pesquisa

Para Taurion (2013) os dados podem ser gerados por máquinas por meio de sensores ou pessoas, através de sistemas, redes sócias, entre outros, para se contextualizar quanto à quantidade de dados que são gerados, citam-se os seguintes exemplos: o facebook que recebe mais de quinhentos terabytes de dados por dia, o Twitter sozinho gera doze terabytes de tuítes, os dados gerados na internet aumentam 90% a cada dois anos.

As empresas, por sua vez geram diariamente uma quantidade assombrosa de dados, que muitas vezes são utilizados de maneira ineficientes sem a exploração total da sua potencialidade. O excesso de dados pode acabar trazendo problemas para as organizações ao invés de trazer soluções, acabam por consumir muitos recursos e gerar pouco ou nenhum resultado.

Uma questão a ser explorada, certamente, é a análise preditiva, onde a questão de padrões e correlações é a fundamentação da análise, ela permite encontrar um padrão de comportamento através das correlações (TAURION, 2013).

Na área agrícola as máquinas podem ser equipadas com sensores que geram uma quantidade significativa de dados, sobre sua localização, tempo de atividade ou inatividade, histórico de utilização, produtividade da área entre outros.

A AGCO, pioneira na utilização de sistema de telemetria em máquinas agrícolas, já se diferencia no mercado agrícola com a transformação dos dados gerados e coletados em informações para seus clientes, porém entende-se que existe a possibilidade de uma maior exploração destes dados.

A AGCO possui muitas fontes de dados, por meio de ERP e seus sistemas complementares, com diferentes bancos de dados, planilhas, oriundos de locais externos, coletados por sensores, que se não utilizados de maneira agregativa ao negócio acabam se tornando um passivo dentro da empresa (TAURION, 2013).

Neste cenário coloca-se a seguinte questão de pesquisa: **Como a AGCO pode obter vantagem competitiva utilizando Big Data e análise preditiva em manutenção de máquinas agrícola?**

1.2 Objetivos

1.2.1 Objetivo geral

Este trabalho tem por objetivo verificar a aderência de uma solução de Big Data associado à análise preditiva na AGCO, analisando a oportunidade de obter vantagem competitiva diante da forte concorrência que existe neste mercado.

1.2.2 Objetivos específicos

- Buscar na literatura as teorias e conceitos sobre Big Data;
- Analisar as fontes de dados disponíveis na AGCO, possíveis de serem utilizadas em uma solução de Big Data;
- Propor uma arquitetura de Big Data para a AGCO;
- Planejar a utilização de análise preditiva, correlacionando os dados disponíveis na empresa que possam auxiliar na identificação de padrões.
- Verificar a aderência do Big Data com o negócio da AGCO;

1.3 Justificativa

Uma vez que a TI está alinhada com as estratégias de negócio da AGCO, suas iniciativas podem possibilitar a melhora e até a transformação do próprio negócio. Taurion (2013) afirma que a utilização de uma solução de Big Data aliada a análise preditiva nos negócios é possível anteceder falhas, reduzir paradas não programadas das operações, aperfeiçoar os ciclos de manutenção e obter um melhor retorno financeiro das máquinas.

Cada máquina que está no campo sofre um desgaste pela sua utilização, que ocorre de maneira não uniforme, devido a diferentes fatores como o solo, o clima, o manuseio, etc. Este desgaste pode ocasionar falhas nas máquinas, que normalmente acarretam em prejuízos ao agricultor, além da perda de confiabilidade do maquinário.

As soluções de Big Data podem auxiliar na visualização destes problemas de maneira preditiva, com a utilização dos dados fornecidos pelos sistemas de telemetria onde são coletados os dados sobre o equipamento e sua produtividade, esforço e eficiência, que são armazenados em um banco de dados, correlacionados a outros dados, que podem estar na própria organização, nas máquinas, na Internet, ou em qualquer outro local, processados com uma velocidade que pode chega a ser quase que em tempo real, pode gerar ganho competitivo para a AGCO. A não exploração destes dados pode acabar se tornando um problema para a empresa, gerando apenas custos com seu armazenamento, e possibilitando que oportunidades importantes passes se ser percebidas.

O presente trabalhado está divido em quatro capítulos.

O Capítulo 1 – Que objetiva, justifica e introduz o assunto a ser abordado.

O Capítulo 2 - Estudo Bibliográfico — apresenta o estado da arte em relação ao tema de estudo: O que é Big Data? Análise preditiva, o que a solução se propõe.

O Capítulo 3 - Traz a preparação, delineamento da pesquisa, definição da unidade de análise dos dados, técnicas de coletas de dados e técnicas de análise dos dados.

O Capítulo 4 - Analisa como a AGCO utiliza seus dados e verifica se existe a oportunidade de gerar ganho competitivo, verifica a aderência com o negócio da AGCO.

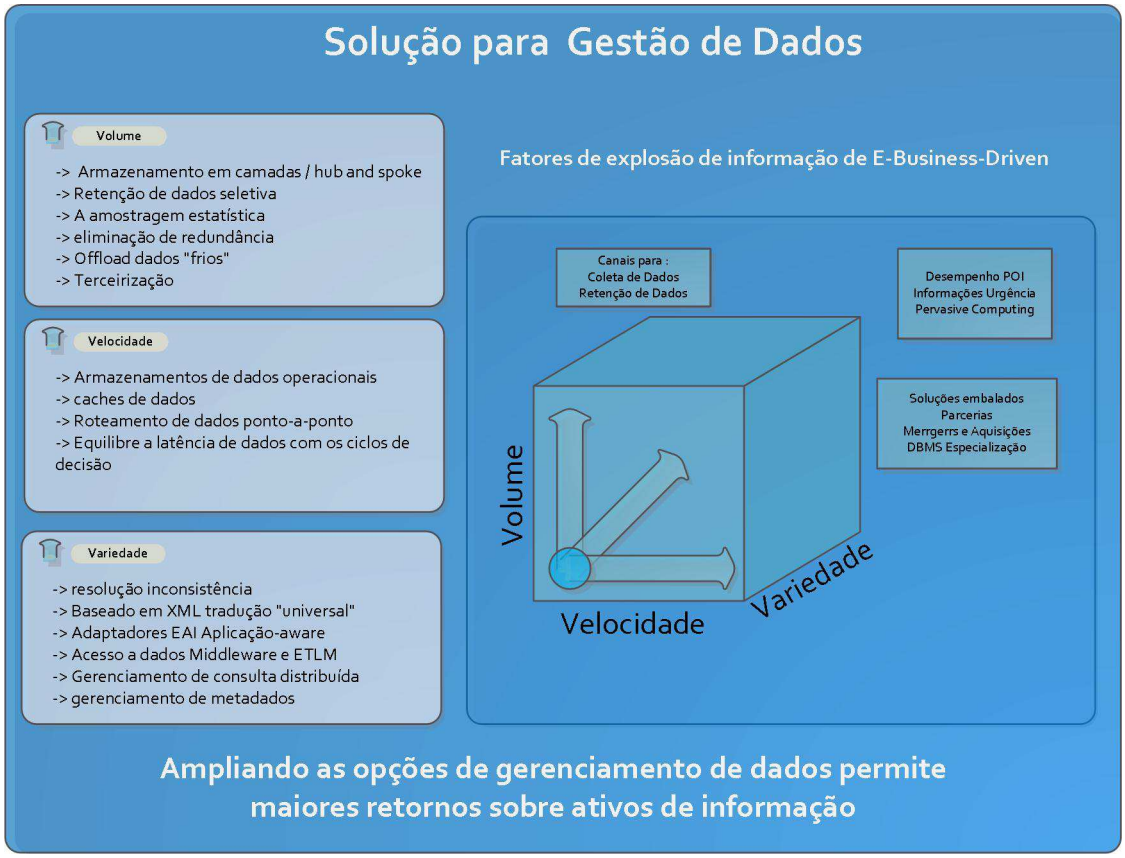
2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo será apresentado os conceitos fundamenteis do objeto de estudo o Big Data, analisando sua relação na atualidade, seu surgimento, e as soluções que ele se propõe. O estudo pretende embasar a problemática em questão, a capacidade de prever falha, em uma máquina agrícola.

2.1 Conceitos de big data

Big Data teve sua previsão feita por Laney (2001) que em um artigo cientifico onde a forma de como lidamos com os dados mudaria de maneira substancial, pois os negócios dependeriam da interpretação de novos e maiores volume, velocidade e variedade de dados. Na Figura 1 uma representação da solução de dados apresentada.

Figura 1 – Solução para gestão dos dados



Fonte: Figura adaptada (LANEY, 2001, p.4).

Para Mayer-schönberger & Cukier (2013), Big Data é a capacidade de uma sociedade em obter informações de maneiras novas a fim de gerar ideias e bens e serviços de valor significativo.

> Big Data se refere a trabalhos em grande escala que não podem ser feitos em escala menor, para extrair novas ideias e criar novas formas de valor de maneira que alterem os mercados, as organizações, a relação entre cidadãos e governos, etc. (Mayer-schönberger & Cukier, 2013, p. 4).

Ainda sob a ótica desses autores percebe-se que Big Data está em constate atualização, pois sempre haverá desafios na maneira em que se vive e interage com o mundo, as mudanças que ainda estão por vir serão de grande magnitude. Relaciona-se Big Data com o microscópio, que permitiu a compreensão dos germes, da mesma formas as novas técnicas de coleta e análise e grandes quantidades de dados ajudarão a compreender o mundo de uma maneira diferente. Porém um fator relevante neste tema é como os dados serão utilizados, está é a verdadeira revolução. O conceito inicialmente introduzido nas Ciências de Astronomia e Genômica, agora está sendo migrado para os demais campos do conhecimento humano. (MAYER-SCHÖNBERGER & CUKIER, 2013).

Para Taurion (2013) uma maneira simples para conceituar Big Data é a utilização da formula: “Big Data = volume + variedade + velocidade + veracidade, tudo agregando + valor” (pagina 61). Além de toda a complexidade envolvida nas soluções de Big Data o exemplo utilizado, onde o microscópio para descrever o seu importância no nosso mundo, o microscópio foi para a medicina e a sociedade, o Big Data também o será para as empresas e própria sociedade.

> Big Data não é apenas um produto de software ou hardware, mas sim um conjunto de tecnologias, processos e praticas que permitem às empresas analisarem dados a que antes não tinham acesso e tomar decisões ou mesmo gerenciar atividades de forma muito mais eficiente. (TAURION, 2013, p. 30)

2.2 Os 5vs do big data

Laney (2001) traz três características fundamentais ao Big Data Figura 1, que iniciam pela letra “V”, são elas: volume, variedade e velocidade , estas também são referenciadas por

diversos autores, alguns, adicionam ainda outros dois Vs são eles : veracidade e valor. Os itens abaixo aprofundam a concepção de cada um desses “5V”s envolvidos na técnica de Big Data.

2.2.1 Volume

Segundo estudos de Mayer-schönberger & Cukier (2013) em 2007 existiam cerca de 300 exabytes de dados armazenados no mundo, sendo que 7% eram analógicos, já em 2013 a quantidade é estimada é de 1.200 exabytes sendo menos de 2% a representação dos dados analógicos. Está claro que o volume de dados gerados é avassalador, explorando um pouco os números que esse volume gera, o Google recebe três bilhões de consultas diariamente, todas são armazenadas em seu banco de dados, além de receber mais de uma hora de vídeo por segundo no YouTube, através de seus 800 milhões de usuários. Mayer-schönberger & Cukier (2013) afirma que o crescimento no volume de informações armazenadas cresce quatro vezes mais rápido que a economia mundial.

“O Facebook recebe mais de 2,7 bilhões de comentários por dia e mais de 300 milhões de fotos por hora, a opção curtir é clicada mais de três bilhões de vezes por dia, gerando mais de quinhentos Terabytes de dados todos os dias, o Twitter cresce a uma taxa aproximada de 200% ao ano, em 2012 ultrapassou os 400 milhões de tweets por dia. (Taurion 2013, pag. 26).

Sintetizando Tole (2013) afirma que o volume refere-se à quantidade de dados que são manipulados e analisados para que se obtenha resultados desejados, o que representa um desafio, pois, para manipular e analisar uma grande quantidade de dados há necessidade de grande quantidade de recursos computacional para processa-los e por fim exibi-los.

2.2.2 Variedade

Para Taurion (2013) assim, como o volume de dados cresce de forma a dobrar a quantidade a cada dezoito meses a variedade também. Os dados são oriundos de sistemas estruturados que hoje já são a minoria e não estruturados de onde vem a avalanche de dados através de e-mails, mídias sociais (Facebbok, LinkedIn, etc), documentos eletrônicos, mensagens instantâneas (whatsapp, viber, etc), câmeras de vídeos, etiquetas RFID, sensores, etc.

A variedade possui papel importante no Big Data, pois a partir de fontes de dados incialmente sem relação, pode-se derivar informações que são capazes de se mostrar extremamente eficientes, um exemplo citado por Taurion (2013) foi à correlação de dados meteorológicos com padrões de compra de clientes, assim é possível planejar quais tipos de produtos deverão estar em determinada loja quando detectado que haverá um período de temperatura elevada em determinado período.

Segundo PRAJAPAT, (2013) o conceito de variedade é reforçado, colocando os diferentes tipos de dados que podem existir, por exemplo, texto, áudio, vídeo e fotos, sendo estes estruturados ou não estruturados.

Para HURWITZ (2013) uma quantidade crescente de dados proveniente de uma imensa variedade de fontes, incluindo os que vêm de máquinas ou sensores, fonte público e fontes privada, fornecidas por outras empresas, pode fazer parte das soluções de Big Data.

Conforme TOLE (2013) a variedade representa o tipo de dados que é armazenado, analisado e utilizado. Os tipos de dados podem consistir em coordenadas de localização (GPS), arquivos de vídeo, dados enviados de navegadores, simulações entre muitos outros. O grande desafio está em como classificar todos esses dados de maneira que eles possam ser "lidos" por todos os usuários e sistemas que o acessam sem criar resultados ambíguos ou confusos.

2.2.3 Velocidade

PRAJAPAT (2013) refere-se a velocidade com baixa latência, velocidade em tempo real em que as análises precisam ser aplicadas. Um exemplo disto está na realização de análises em um fluxo contínuo de dados provenientes de rede social, onde a velocidade de processamento é fundamental para identificação do evento o quanto antes.

Taurion (2013) coloca a velocidade em evidência, pois muitas vezes são necessárias execuções de analises muito velozes. Dependendo do caso imagina-se a utilização em tempo real dessas analises. Sua importância aumenta quando se verifica a crescente velocidade com que os negócios precisam reagir às mudanças. No caso de sensores utilizados no controle de transito, os dados só tem significado se analisados em tempo real, para que um congestionamento possa ser evitado.

Conforme Mayer-schönberger & Cukier (2013) o Google processa diariamente mais de 24 petabytes de dados por meio de seus servidores espalhados ao redor mundo, sem uma velocidade adequada isto seria realmente impossível.

TOLE (2013) diz que a velocidade é a maneira como os dados se deslocam de um ponto “X”, para um ponto “Y”. Está é uma questão importante devido ao grande numero de transmissões que ocorrem em diversos dispositivos (Notebook, Smartphones, tablets...).

2.2.4 Veracidade

Taurion (2013) adiciona o “V” de veracidade, e afirma que não são todos os dados que podem ser analisados, eles precisam fazer sentidos e ter autenticidade, segundo estudos, em 2013 apenas 20% de todos os dados poderiam ser considerados validos para serem analisados, sendo que atualmente somente 5% são realmente utilizados nas analises.

Para HURWITZ (2013) a veracidade é mostrada como validade, certamente todos gostariam de obter resultados precisos, porém quando se inicia a coleta e análise de petabytes de dados, não é necessário preocupar-se com a veracidade de todos os dados, inicialmente a coleta deve ser realmente suja, após é que os mesmos serão analisados e limpos nas próximas etapas. O mais importante é saber se existe alguma relação entre os dados coletados do que assegurar que todos os dados são realmente válidos.

2.2.5 Valor

Taurion (2013) afirma que para uma organização implementar um projeto de Big Data é imprescindível que se obtenha o retorno sobre o investimento, podemos considerar que o resultado da aplicação dos conceitos de Big Data é o valor gerado para os negócios. A missão do Big Data enfim é criar valor descobrindo padrões e relacionamentos entre dados que antes estavam espalhados e perdidos em diferente locais.

Conforme Michael Porter (1990), a vantagem competitiva se obtém por dois caminhos, baseada nos custos ou na diferenciação.

“As empresas que conseguem descobrir uma tecnologia melhor para executar uma atividade do que os seus concorrentes ganham, portanto, vantagem competitiva.” (Porter, 1990 p.158).

Outro importante ponto de vista do valor do Big Data é o apoio para a tomada de decisão, para Gomes (2007) o processo de decisão é uma de escolha de forma direta ou indireta, a partir de diferentes alternativas que podem resolver um problema. O Big Data pode levar embasamento suficiente para que o tomador decisão possa realizar a escolha de maneira assertiva.

Neste sentido Mohanty; Jagadeesh; Srivatsa (2013) afirmam que sensores de coleta de dados sobre o solo e equipamentos agrícolas tornaram-se essenciais no sucesso da agricultura de precisão, para a gestão de dados agrícolas, na qual existe grande quantidade de dados oriundos desses sensores, pode-se gerir eficientemente de forma a auxiliar na tomada de decisão.

Para Tole (2013) a quantidade de dados coletados, armazenados e devidamente gerenciados oferece como resultado insights para um desenvolvimento de produtos e serviços.

2.3 Análise preditiva

Conforme TURBAN, (2010, p 464) a Análise Preditiva "é uma ferramenta que ajuda a determinar o provável resultado futuro de um evento ou a probabilidade de uma situação ocorrer". A Análise Preditiva tem potencial para a identificação de relacionamentos e padrões.

Para Fogarty (2004) a análise preditiva utiliza poderosos e sofisticados algoritmos para que se possa identificar padrões de comportamento que surgem.

A correlação destaca-se no "mundo Big Data", pela possibilidade de obtermos ideias de forma mais fácil, rápida e clara. O princípio da correlação está na quantificação da estatística entre dois pontos, em que se um dado se altera o outro provavelmente se alterará também.

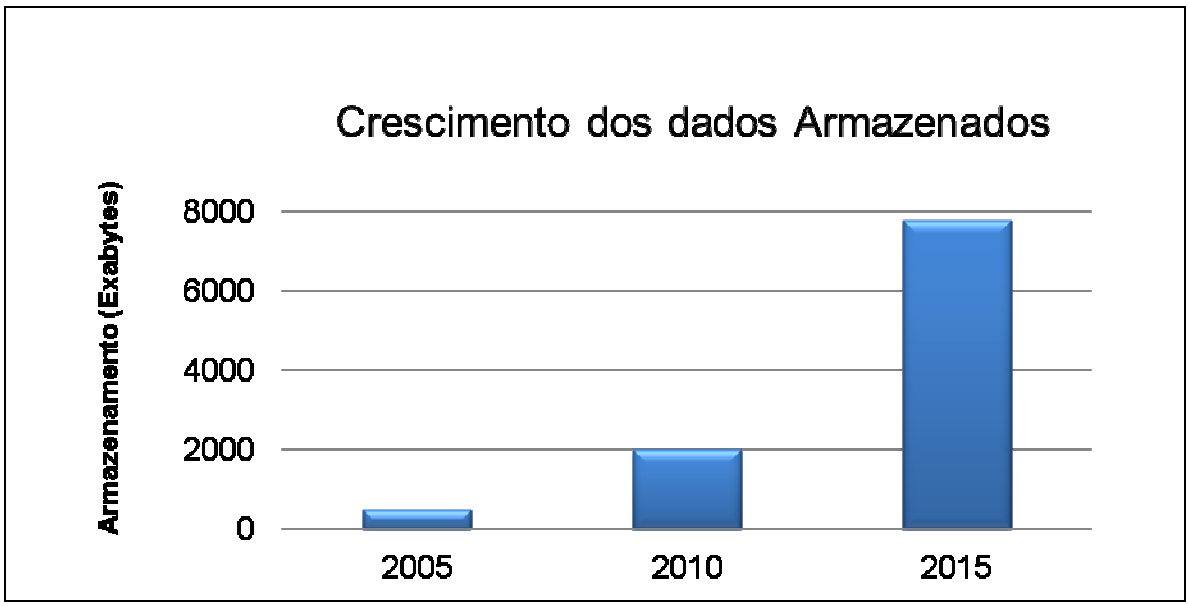
A correlação de "análise de previsão" ou análise preditiva está na finalidade de prever eventos. Esta técnica vem sendo utilizada na previsão de grandes falhas estruturais e mecânicas através da instalação de sensores em máquinas, motores e infraestrutura para que dados como calor, vibração e estresse possam ser analisados a fim de identificar uma possível falha mais adiante. O principal objetivo é identificar e observar uma "ponte" para prever eventos futuros. As previsões com base em correlações estão na essência do Big Data. É importante entender que a análise preditiva não tem como objetivo identificar o porquê determinado evento acontece, mas mostrar o quê acontece e geralmente isto basta, ela pode informar que um motor está superaquecendo, mas talvez não informe que o superaquecimento

se deve a uma correia, parafuso ou polia. Graças a quantidade de dados disponíveis e melhores instrumentos as correlações surgem cada vez com maior rapidez e menor custo, preciso porém ter cuidado com estas correlações, pois conforme a quantidade de dados aumenta em magnitude a possibilidade de encontrar correlações não verdadeiras também aumenta (MAYER-SCHONBEGER & CUKIER, 2013).

2.4 **O que a técnica de big data se propõe a resolver**

Como a geração e armazenamento de dados estão em crescimento e acabam gerando uma curva em formato exponencial, as medidas de volume já mudaram de gigabytes, terabytes para petabytes 10 ˜18 e ate zettabytes 10 ˜21. segundo (Gantz e Reinsel, 2010, Gantz e Reinsel, 2011) a previsão é que em 2015 haverá perto de 8. Zettabytes, um aumento de 300 % se comprado a 2010 quando havia cerca de 2 zettabytes.

Figura 2 - Aumento dos dados armazenados estimados pela IDC



Fonte: IDC's Digital Universe Study, patrocinado pela EMC, Junho de 2011

Deste modo as técnicas utilizadas para coletar, armazenar e analisar não serão mais capazes de lidar de maneira satisfatória com a quantidade, até pouco tempo inimaginável de dados. Muitas são as razões que justificam tamanho crescimento. Usuários estão permanentemente interconectados criando bilhões de conexões que se transformam em fontes

de dados empresas estão monitorando informações de seus clientes, fornecedores e suas operações comerciais, existem milhões de sensores; celulares, medidores eletrônicos, dispositivos portáteis, automóveis sensoriam, criam e trocam dados remotamente na “Internet das coisas” (THE ECONOMIS, 2010).

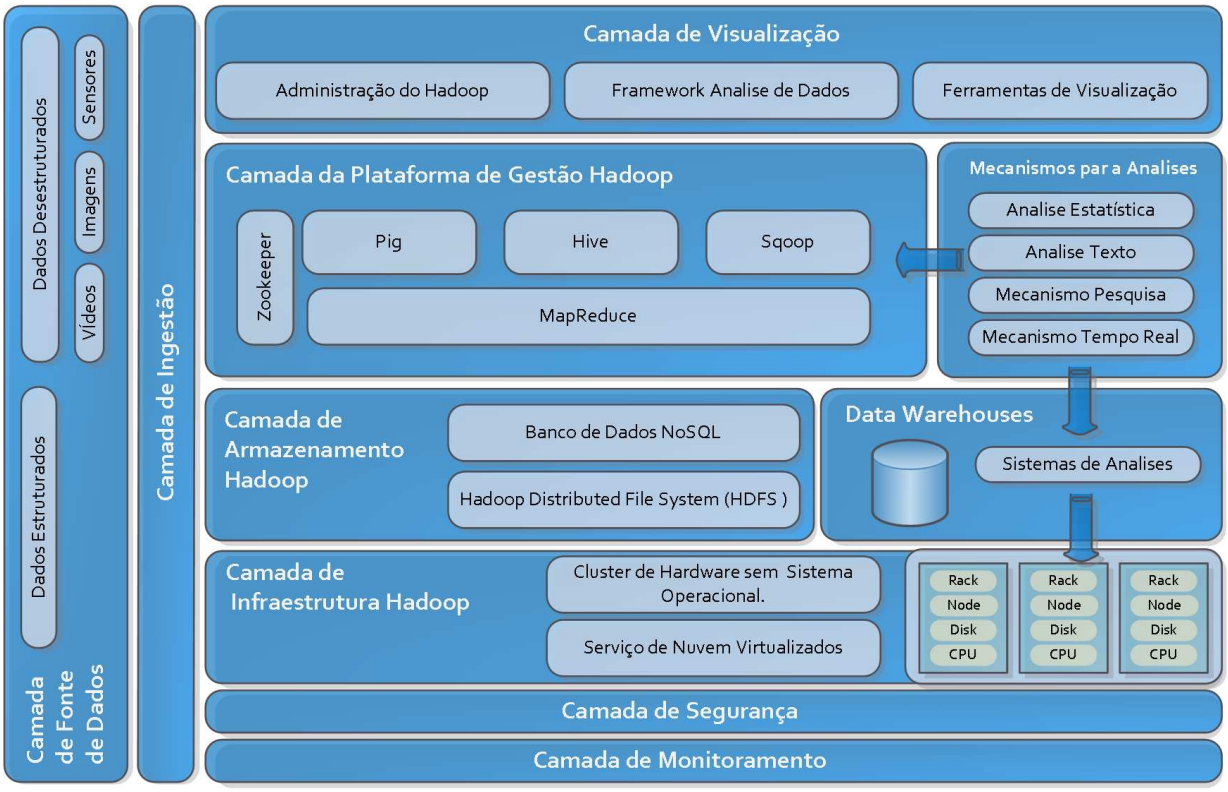
Para que se possa pensar em lidar com essa quantidade de dados entendeu-se que haveria necessidade de uma nova visão para à coleta, armazenamento e analise, o Big Data vêm com a audaciosa missão de trabalhar com esse mar de dados e transforma-los em valiosas informações. (COSTA, 2012).

2.5 Arquitetura do big data

Conforme Sawant e Shah (2014), com o crescimento exponencial dos dados, não só no mundo dos negócios, mas nas organizações de pesquisas, educacionais e governo, os bancos de dados relacionais não estão preparados para atender a este crescente aumento de maneira satisfatória, pois há necessidade de analisar cenários e modelos complexos envolvendo imagens, vídeos e dados textuais. Aliado a isso estão às novas fontes de dados internas e externas incluindo mídia social, dispositivos móveis, sensores e outros dados gerados por máquina que os cientistas de dados precisam analisar para encontrar a “agulha num palheiro”. Para que isso seja possível o modo como os dados são gerenciados, armazenados e analisados precisar mudar e por fim trazer valor para organizações.

Como se pode observar na Figura 3, adaptada a partir SAWANT e SHAH, (2014), uma solução de Big Data requer grande quantidade de componentes para que se tenha uma arquitetura capaz de produzir resultados satisfatórios. Na arquitetura do Big Data conforme visualizado na Figura 3 estão sendo utilizados frameworks de código aberto, porém existem produtos licenciados, embalados capazes de tirar o grande proveito das funcionalidades dos vários componentes de uma solução de Big Data.

Figura 3 – Arquitetura para solução de Big Data



FONTE: Figura adaptada (Sawant e Shah, 2014, p.10).

Para um melhor entendimento todas as camadas da arquitetura são aprofundadas para demostrar sua função na solução de Big Data, conforme propõe SAWANT e SHAH, (2014).

2.5.1 Camada de Fonte de Dados

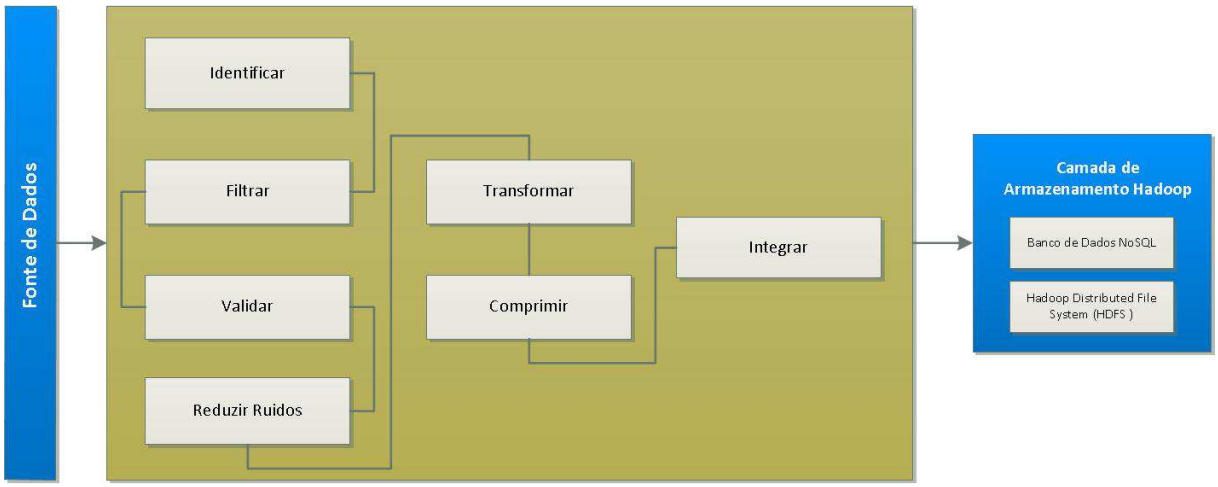
Na camada de fonte de dados está à origem dos dados, que são classificados como dados estruturados, aqueles oriundos de banco de dados relacional, por exemplo, Oracle, SQL Server, MySQL, PostgreSQL, ou dados desestruturados originados por imagens, textos, vídeos, sensores, redes sociais e que geram enorme quantidade de dados.

Um dos problemas para definição e implementação de uma solução de Big Data está na camada de fontes de dados, onde existe grande quantidade de fontes de dados, a velocidade e variedade competem entre si para ser utilizada na analise dos dados elegíveis da arquitetura do Big Data.

2.5.2 Camada de Ingestão

A camada de ingestão tem como responsabilidade separar os dados com ruídos dos dados relevantes, ela tem que ser eficiente o bastante para lidar com o grande volume de dados, precisa alta velocidade de processamento e capacidade para interpretar a imensa variedade de dados. Deve ser capaz de validar, limpar, transformar, reduzir e integrar os dados com as demais tecnologias da solução de Big Data. A camada de ingestão carrega a informação final relevante, sem ruídos, para a camada de armazenamento Hadoop (HDFS) as aplicações responsáveis por este trabalho são Apache Flume e Storm. Na Figura 4 temos a representação de como funciona a camada de ingestão.

Figura 4 – Funcionamento da camada de Ingestão



FONTE: Adaptada de (Sawant e Shah, 2014, p.31).

A responsabilidade de cada uma das necessidades envolvidas nesta camada será explicada de maneira detalhada.

- Identificar os vários formatos de dados conhecidos ou cessão de formatos padrão para dados não estruturados.
- Filtrar as informações de entrada pertinente para a organização, com base no repositório de MDM.
- Validar e analisar os dados de forma contínua, sempre atualizado com novos metadados do MDM .

- Reduzir o ruído envolve a limpeza de dados, removendo os ruídos e minimizando discrepâncias.
- Transformar envolve a divisão, convergindo, desnormalizar ou resumir dados.
- Comprimir envolve a redução do tamanho dos dados, mas sem perder a relevância dos dados no processo. Não pode afetar os resultados das análises após os dados serem comprimidos.
- Integrar os dados com a camada de armazenamento Hadoop, ou seja, tem a responsabilidade de integrar os dados com os Hadoop sistema de arquivos distribuídos (HDFS) e os bancos de dados NoSQL.

2.5.3 Camada de Monitoramento

Devido ao grande número de *clusters* utilizados na solução de Big Data para o armazenamento dos dados, é de extrema importância que se possua um monitoramento sobre a solução tecnológica adotada a fim de garantir que os acordos de nível de serviço estejam sendo atendidos, garantindo assim o mínimo de inatividade da solução.

Os sistemas que terão a responsabilidade de monitorar devem ser capazes de lidar com aglomerado distribuído de servidores, e ainda trabalhar com diferentes sistemas operacionais, e tipos de hardware, sendo capaz de se comunicar com os protocolos de alto nível como o XML.

Além do desempenho e paralelismo que são fundamentais para monitorar a sobrecarga a ferramenta deve prover o armazenamento e visualização de dados, as ferramentas de código aberto largamente utilizadas para esta monitoração são Ganglia e Nagios.

2.5.4 Camada de Segurança

Camada de extrema importância e grande preocupação, devido à sensibilidade dos dados, por exemplo: hábitos de compra de clientes, histórico médico de pacientes, dados demográficos de doenças genéticas e muitos outros que precisam ser fortemente protegidos tanto para atendimento de requisitos quanto a privacidade de indivíduos.

Para que se possa garantir a segurança nos projetos de Big Data, mantendo performance, escalabilidade e funcionalidade, de maneira simples, utiliza-se alguns pré-requisitos de segurança principalmente na utilização de servidores distribuídos, como por exemplo:

- Autenticar utilizando protocolos seguros como o Kerberos,
- Permitir a criptografia de arquivos por camadas,
- Garantir que toda a comunicação entre os servidores distribuídos seja segura,
- Registrar a comunicação entre o conjuto de servidores dsistribuidos, utilizando mecanismos de registro distribuídos capaz de rastrear anomalias entres as camadas.

2.5.5 Camada de Infraestrutura do Hadoop.

A camada de infraestrutura do Hadoop tem como responsabilidade suportar a camada de storage, para tal uma infraestrutura física robusta, barata e confiável é fundamental para garantir a operação e escalabilidade da arquitetura de dados.

Para que seja capaz de atender a demanda de volume imprevisto ou imprevisível, velocidade e variedade de dados, a infraestrutura física para soluções de Big Data tem que ser diferente do que para os dados tradicionais.

A camada de infraestrutura física Hadoop (HPIL) é baseada em um modelo de computação distribuída. Isto significa que os dados podem estar fisicamente armazenados em muitos locais diferentes e interligados entre si através de redes e sistemas de arquivo distribuído.

2.5.6 Camada de Armazenamento do Hadoop

Com a utilização de massiva de dados o armazenamento e processamento distribuído causa uma grande mudança na forma como as empresas lidam com o Big Data. O sistema de armazenamento distribuído tem como referência a tolerância a falhas, e permitindo a paralelização de algoritmos para que o processamento seja distribuído provendo alta velocidade em larga escala. O sistema de arquivos distribuído Hadoop (HDFS) é a pedra fundamental da camada de armazenamento de dados.

O framework Hadoop, open source, que permite o armazenamento de grande volume de dados distribuído através de máquinas de baixo custo. Prove a dissociação entre a engenharia de software em computação distribuída e a logica da aplicação. O Hadoop permite que você interaja com o conjunto logico de processamento ao invés do conjunto físico, os dois principais componentes são o sistema distribuído de arquivo que pode suportar petabytes de dados e um mapa altamente escalável capaz de reduzir o tempo que o motor que calcula os resultados em lote.

O HDFS exige programas complexos de leitura de arquivos/gravação escrito por desenvolvedores qualificados. A sua estrutura não permite que os dados sejam manipulados de maneira fácil, assim é necessário utilizar distribuição de banco de dados não relacional capaz de trabalhar com uma grande quantidade de dados, incluindo chave-valor, documentos, gráficos, colunas, e banco de dados geoespaciais. Estes bancos de dados são referidos como NoSQL, ou Not Only SQL.

2.5.7 Camada da plataforma de gestão do Hadoop

Na camada da plataforma de gestão do Hadoop são fornecidas as ferramentas e linguagens de consulta para acessar os bancos de dados NoSQL, que utilizam o sistema de armazenamento de arquivos HDFS sobre a camada de infraestrutura física Hadoop.

A camada de gerenciamento da plataforma de gestão Hadoop acessa os dados, executa consultas e gerencia as camadas inferiores, utilizando linguagens, como Pig e Hive. O Hadoop e MapReduce são as novas tecnologias que permitem as empresas armazenar, acessar e analisar grandes quantidades de dados em tempo quase que real.

Estas tecnologias abordam um dos principais problemas, a capacidade de processar gigantescas quantidades de dados de maneira eficiente, com custo-benefício apropriado, e dentro de um tempo satisfatório.

2.5.8 Mecanismos para Análise

As organizações precisam adotar diferentes abordagens para resolver diferentes problemas com Big Data, em algumas análise os data warehouse tradicionais conseguem atender, enquanto outras análises precisará também da utilização de Big Data, assim como métodos tradicionais de Bussines Inteligencie.

As análises podem ocorrer em banco de dados relacionais tradicionais ou em grandes volumes de dados com o processamento distribuído. Data warehouses continuará gerenciando os dados transacionais baseados em banco de dados relacional com o ambiente centralizado. Para o gerenciamento dos dados não estruturados e ou distribuídos ficará a cargo das soluções Hadoop.

As análises de dados entre o Data warehouse e Big Data como o Hive e HBase,é o ponto importante de troca de informações que acontece em qualquer direção com a utilização da ferramenta de integração Apache Sqoop. Com os bancos Cassandra, Vertica, Hbase devido sua baixa latência é possível analisar dados em tempo real, como os dados produzidos em rede sociais. As ferramentas Madlib e Mahout ambas Open Source permitem a utilização de complexos algoritmos e de fácil acesso para que seja executada a análise pelos cientistas de dados.

2.5.9 Camada de Visualização

Devido ao grande volume de dados do Big Data, se a camada de visualização não for pensada no momento da concepção da arquitetura da solução poderá ter sobrecarga deixando analistas e cientistas de dados sem a possibilidade de obter *insights* rápidos e diminuir a capacidade de olhar para diferentes dados em diversos modos visuais.

Os gráficos de pizza e barras são ao mais utilizados para a interpretação dos dados, por si só, já mudou a partir de uma representação de dados de amostra para uma contemplando grande volume e variedade de dados residentes nas organizações, bem como o sentimento no mercado com a análise dos dados de mídia social.

2.6 O Cientista de Dados

O cientista de dados, papel que não existia até pouco tempo, vem sendo muito falado e requisitado no mercado, Taurion (2013) define o cientista de dados como profissional de alto nível de formação com muita curiosidade no mudo de Big Data, com capacidade de criar algo novo. Este Cientista vai trabalhar diretamente na ciência dos dados, onde a tecnologia não será impeditiva, mas as pessoas sim. Quanto ao conhecimento o cientista de dados precisa saber sobre estatística, matemática impreterivelmente conhecer o negócio e ter familiaridade com as tecnologias envolvidas no mundo Big Data como Haddop e Pig.

Segundo Sawant e Shah (2014), o cientista de dados tem como missão encontrar a agulha no palheiro, para isso precisa obter valiosos *insights* por meio dos dados, para dessa forma aumentar a probabilidade de um projeto Big Data ser bem sucedido é fundamental que o cientista de dados participe desde a definição do escopo do projeto. Ele precisa estar sempre atualizado, pois a cada pouco tempo surgem novas tecnologias no complexo mundo do Big Data.

Para Mohanty; Jagadeesh e Srivatsa, (2013) o cientista de dados tem de ser capaz de programar, preferencialmente em diferentes linguagens de programação como Python, R, Java, Ruby, Clojure, Matlab, Pig ou SQL. Precisa necessariamente compreender tecnologias de Big Data como Hadoop, Hive e MapReduce. Outros conhecimentos importantes para o cientista de dados são:

- Linguagem Natural: Interações entre computadores e humanos
- Aprendizagem de máquina: melhorar o uso de computadores, bem como desenvolver algoritmos eficientes.
- Modelagem conceitual: ser capaz de compartilhar e articular modelagem
- Análise estatística: para entender e solucionar possíveis limitações em modelos estatísticos.
- Modelagem preditiva: a grande maioria dos problemas de dados está em capaz de prever os resultados futuros.
- Testes de hipóteses: ser capaz de desenvolver hipóteses e testar com cuidadosos experimentos.

3 MÉTODOS E PROCEDIMENTOS

Este capítulo delineia o método de pesquisa, a unidade-caso e indica as técnicas de coleta e de analise de dados utilizadas. Por fim descreve as limitações do estudo.

3.1 Delineamento da pesquisa

Para que o objetivo fosse alcançado utilizou-se o método de pesquisa estudo de caso descritivo que segundo Yin (2005, p.32), é “uma investigação empírica que analisa um fenômeno contemporâneo dentro de seu contexto real, quando os limites entre o fenômeno e o contexto não estão claramente definidos”. Não havendo controle sobre os eventos comportamentais por parte do pesquisador, fazendo com que a analise de dados apresente suas características especificas.

Nesta mesma linha está GOODE & HATT (1969) onde em sua percepção o estudo de caso é uma forma de organizar os dados preservando a forma original do objeto estudado.

Algumas as principais características do estudo de caso segundo: Trauth & O'Connor (1991), são:

- Fenômeno observado em seu ambiente natural;
- Dados coletados por diversos meios;
- Uma ou mais entidades (pessoa, grupo, organização) são examinadas;
- A complexidade da unidade é estudada intensamente;
- Pesquisa dirigida aos estágios de exploração, classificação e desenvolvimento
- de hipóteses do processo de construção do conhecimento;
- O pesquisador não precisa especificar previamente o conjunto de variáveis dependentes e independentes;
- Os resultados dependem fortemente do poder de integração do pesquisador;
- Podem ser feitas mudanças na seleção do caso ou dos métodos de coleta de dados à medida que o pesquisador desenvolve novas hipóteses;
- Pesquisa envolvida com questões "como" e "por que" ao invés de frequências ou incidências;
- Enfoque em eventos contemporâneos.

Nas pesquisas bibliográficas para fundamentação teórica onde temas Big Data, análise preditiva são abordados, sendo esta vital para a elaboração deste trabalho, a pesquisa realizada em livros, artigos publicados, e diversos referencias teóricos que de alguma maneira já foram analisadas.

3.2 Definição da unidade de análise

Esta pesquisa foi realizada na AGCO. A escolha da empresa ocorreu pelo fato do pesquisador atualmente ser funcionário da mesma.

A empresa AGCO é composta por múltiplas empresas, como a Massey Ferguson, Fendt, Challenger, Valtra e GSI, com este portfólio de bandeiras , é uma das principais industrias no seguimento, com variedade de produtos incluindo, tratores, colheitadeiras, implementos, pulverizadores, equipamentos de proteínas para alimentação animal, silos para grãos e colhedoras de cana-de-açúcar.

No ano de 2013 teve um faturamento na casa de 13 bilhões de dólares. A região da América do Sul representou no ano de 2013 cerca de 21% do faturamento total, mostrando sua força perante o restante da companhia.

Quanto à visão, missão e valores da AGCO:

a) **MISSÃO** – Crescimento sustentável através do atendimento ao cliente, inovação, qualidade e comprometimento superiores.
b) **VISÃO** – Soluções de alta tecnologia para produtores rurais que alimentam o mundo.
c) **VALORES** – A AGCO acredita nos seguintes valores:

- **Foco no Cliente**

  Criamos soluções excelentes para nossos clientes ouvindo atentamente suas necessidades e excedendo suas expectativas.

- **Foco na Concessionária e Distribuidores**

  Percebemos que a lucratividade da concessionária é o meio para o nosso sucesso e esperamos ser o fornecedor preferido.

- **Dimensões Humanas**

  Valorizamos nossos funcionários.

  Queremos ser o empregador preferido no nosso segmento.

Queremos desenvolver funcionários altamente motivados que sejam os mais instruídos e melhor treinados no segmento.

Desenvolvemos as habilidades e qualificações dos funcionários.

Queremos que nossos líderes sejam pró-ativos e indiquem a direção.

Queremos que nossos líderes influenciem e estabeleçam as regras.

Queremos alcançar vantagem competitiva através da agilidade, qualidade e comportamento inovador.

- **Número Um na Qualidade Percebida pelo Cliente**

Mais do que entregar alta qualidade em produtos e serviços, queremos ser reconhecidos por isso.

- **Padrões Éticos**

Agiremos de maneira ética como bons cidadãos corporativos em todas as comunidades nas quais a Companhia atua.

Cuidamos do meio ambiente.

Queremos proteger o meio ambiente de influências nocivas, conservar os recursos naturais e promover a consciência ambiental.

- **Valores da Marca**

Reconhecemos a tradição e o valor das marcas, a lealdade de nossos clientes e a identificação de nossas concessionárias e distribuidores com as marcas. A estratégia multi-marcas da AGCO mantém o valor de cada marca.

- **Agregar valor para o acionista**

Queremos alcançar crescimento lucrativo.

A AGCO administrará o negócio para oferecer retorno superior a seus acionistas

Neste estudo será analisada a aderência da organização com uma solução de Big Data com foco em Analise preditiva, onde o principal ponto de abordagem será prever quando uma máquina agrícola apresentará algum problema para o seu proprietário. Para que está máquina faça parte do estudo à mesma deverá estar equipada com o sistema de telemetria, este estudo acontecerá através do cruzamento de dados por meio de diversos sistemas da organização como, sistema de garantia, ERP, sistema de venda de peças de reposição, e os dados gerados pelos sensores oriundos do sistema de telemetria.

3.3 Técnicas de coleta de dados

Na coleta de dados para este trabalho, utilizaram-se as seguintes técnicas: Pesquisa bibliográfica: seleção de autores que publicam obras relacionadas ao assunto: Big Data, vantagem competitiva, entre outros. As pesquisas foram realizadas por meio físico e eletrônico, buscando por livros e trabalhos acadêmicos com o intuito de obter melhor entendimento do assunto.

Entrevistas Dinâmicas: Realizadas entrevistas de maneira dinâmica com diversos colaboradores de diferentes setores da companhia, conforme propôs (Yin, 2010) “uma das fontes mais importantes de informação para o estudo de caso”. As entrevistas são conversas guiadas, diferente de investigações estruturadas.

3.4 Técnicas de análise de dados

A técnica de analise de dados utilizada para apreciação do estudo de caso foi a qualitativa, para analisar as respostas com suas questões nesta técnica é possível analisar o conteúdo de inúmeras fontes tais como livros, revistas, jornais, discursos, entre outras, inclusive documentos pessoais MARCONI; LAKATOS, (1999).

O estudo de caso foi realizado com a compilação da luz da teoria previamente estudada e analisada em profundidade na fundamentação teórica, a qual forneceu forte embasamento teórico.

3.5 Limitações do método

O estudo foi realizado em uma indústria multinacional de máquinas agrícolas, porém os dados coletados e analisados foram de pessoas e equipamentos produzidos na AGCO América do Sul.

As máquinas que participaram da análise foram restritas a um conjunto que possuíam o dispositivo de telemetria instalado, sendo este um dos principais pontos para captura de dados para posterior analise.

O cruzamento dos dados ocorreu entre sistemas da organização através dos seus diversos bancos de dados utilizados, podendo ainda contar com a utilização de dados não estruturados, os valores reais dos dados não serão expostos no decorrer do estudo de caso.

4 APRESENTAÇÃO E ANÁLISE DOS DADOS

Neste capítulo será apresentada a proposta de utilização de arquitetura para solução de Big Data, contemplando as fontes de dados, arquitetura para Big Data e a analise dos dados para verificar se existe aderência do Big Data com o negócio da AGCO.

4.1 Fontes de Dados

A AGCO possui inúmeras fontes de dados, diversos sistemas compõe o conjunto de dados da organização, o possível correlacionando entre os conjuntos de dados poderá fornecer valiosas informações para o negócio, estas fontes de dados serão exploradas mais profundamente neste capítulo.

Para saber quais sistemas podem fornecer dados a fim de serem utilizados na solução de Big Data, e verificar a correlação para aplicação de uma analise preditiva, foi consultada a área de arquitetura de sistemas do departamento de (TI) da AGCO. que apontou alguns sistemas que podem auxiliar no estudo de caso, os principais sistemas informados foram: o ERP da organização (JD Edwards), o sistema de Garantias, o sistema AGCOOLINE e o sistema AGCOMMAND responsável pela Telemetria (Tecnologia embarcada nas máquinas).

Após compreender quais os sistemas são elegíveis para o estudo de caso, direcionou-se entrevistas dinâmicas com diversos gestores e colaboradores que trabalham diretamente com estes sistemas ou que participaram de sua concepção e ou implantação, outras informações Institucionais sobre os sistemas foram solicitadas aos colaboradores da área de TI da organização.

4.1.1 ERP – JD Edwards

Para o levantamento das informações sobre o sistema de ERP da empresa, foi realizada entrevista com um profissional que está há mais de 20 anos na empresa. Participou da implantação do sistema de ERP atual. De acordo com Oracle (2014):

> O JD Edwards Enterprise One da Oracle, um pacote de aplicações integradas com software abrangente de planejamento de recursos Empresariais. Ele combina valor comercial, tecnologia com base em padrões e experiência aprofundada no setor em uma solução de negócios com custo total de propriedade reduzido.

O sistema é formado por módulos e atualmente está suportando as áreas de Finanças, Vendas, Compras, Planejamento e Materiais, Manufaturado das às áreas de negócios e unidades da AGCO no Brasil e na Argentina.

Segundo a gestora o sistema do JDE armazena todas as operações da organização de compras, vendas, produção e financeiras (Contas a Pagar e a Receber, Custos, fiscais, informações sobre a linha de montagem, etc). As informações mais relevantes para o estudo de caso estão ligadas ao processo de produção e venda, a seguir os itens indicados para ser analisados.

- Configuração do produto e materiais utilizados na sua produção, quem vem do termo "Build of Material" (BOM);
- Ordem de Fabricação
- Data de fabricação do produto;
- Lista de materiais utilizados na ordem de fabricação de produto em questão;
- Lotes/séries dos itens rastreáveis e suas respectivas compras/fornecedores
- Faturamento de máquinas;
- Faturamentos de peças para atendimento de máquina parada/garantia
- Cadastro da peça (caso não exista)
- Pedido de vendas de peças (garantia ou venda normal)
- Ordem de despacho (ODP)
- Inúmeros relatórios (vendas, faturamento, entrega, etc.)

Quanto à tomada de decisão, por tratar-se de um ERP, todas as decisões de caráter tático-operacional podem ser tomadas a partir de informações deste sistema. No tema em questão (previsão de quebra de máquinas), o faturamento de itens em garantia e de peças para máquinas paradas são essenciais para o processo de identificação e recuperação de garantias e poderiam ser utilizados para suportar uma analise preditiva, cruzando-se as ordens de produção e respectivas informações.

Especificações técnicas do sistema JD Edwards EnterpriseOne (JDE):

- Banco de dados: Oracle
- Tamanho do banco de dados: 3 Terabytes

- Quantidade de usuários: 1.100
- Tempo de operação: 16 anos

4.1.2 AGCOONLINE

No sistema AGCOONLINE foi entrevistado o gestor que concebeu o sistema em 2002, o sistema contempla um portal para realizar transações entre a AGCO e suas redes de concessionárias (Massey Ferguson, Valtra, AGCO Allis e Challenger) na América do Sul, Central e Caribe.

As principais transações do sistema AGCOONLINE ocorrem entre cinco departamentos da AGCO e suas concessionárias, contemplando Vendas, Marketing, Peças, Serviço e Finanças. Neste sistema encontram-se as informações de Pedido de tratores, Pedido de Peças, Registro de revisões, Consulta de disponibilidade de Peças, Consulta de informações técnicas, Preenchimento do Balanço Patrimonial, entre outros. Adicionalmente, o AGCO Online executa single sign-on com outros sistemas corporativos.

Na Figura 5 visualizamos de maneira gráfica as diversas áreas que estão envolvidas no sistema AGCOONLINE.

Figura 5 – Integrações do Sistema AGCOONLINE

FONTE: Elaborada pelo autor

Na opinião do Gestor algumas informações importantes estão disponíveis no sistema que podem auxiliar na descoberta de padrões

- PDI – Pré-Delivery-Inspection – formulário com check list de verificação antes da entrega do produto para o cliente final

- Processos de Garantia
- Pedido de Peças na modalidade Garantia
- ROP – Registro de Ocorrência de Peças

Especificações técnicas do sistema AGCOONLINE:

- Banco de dados: Microsoft SQL Server
- Tamanho do banco de dados: 200 Gb
- Quantidade de usuários: 300
- Tempo de operação: 16 anos

4.1.3 Sistema de Garantias

Segundo Gestor da área, o sistema de Garantias possui informações referente às solicitações de garantia dos produtos faturados pelas companhias Massey Ferguson e Valtra do Brasil. No sistema de Garantias são validadas todas as informações registradas nos processos que as concessionárias enviam para recebe-las.

No sistema informações consideradas importantes pelo gestor são enviadas e recebidas da rede de concessionaria, tais como faturamento de produtos, processos de Garantia, certificado de entrega, revisões, entre outros.

Na opinião do Gestor muitos dados que estão no sistema de Garantia são importantes, para auxiliar na identificação de padrões, o mesmo elenca alguns informações que acredita auxiliar:

- Data de faturamento do produto,
- Serie do monobloco,
- Serie do motor (quando tiver),
- Serie do eixo dianteiro tracionado (quando tiver),
- Serie da transmissão,
- Data de Entrega do Produto,
- Data de abertura e fechamento da Ordem de Serviço,
- Horímetro (mostra o tempo que máquina está operando)
- Dados do cliente que está com o produto,

- Peça causadora da falha,
- Código de falha e defeito,
- Descrição da falha, ação tomada,
- Tipo de operação e cultura,
- Peças que foram trocadas na falha (pode ter processo sem peças)

O Gestor explica que os processos de garantia possuem dois status após as analises, podendo ser procedente nesse caso a fábrica que realiza os reparos necessários, ou improcedentes, onde o dono do produto fica responsável pelo reparo. Atualmente são realizadas analises dos códigos de falhas e peça causadora, porém ele considera que existe oportunidades para exploração dos dados que estão disponíveis no sistema.

Especificações técnicas do sistema de Garantias:

- Banco de dados: Oracle
- Tamanho do banco de dados: 10 Gb
- Quantidade de usuários: 20
- Tempo de operação: 17 anos

4.1.4 AGCOMMAND

O Gestor responsável pelo sistema AGCOMMAND explica que o sistema fornece aos seus usuários uma solução de registro de dados e de controle de gerenciamento de frota de fácil manuseio, com um custo acessível, e que pode ser instalado em todos os seus equipamentos, independente da marca.

Com ele o usuário poderá monitorar detalhadamente a localização exata da máquina em um determinado período, e saberá como está o seu desempenho em campo. O AGCOMMAND ainda informa o histórico de operação e relatórios da frota, incluindo eficiência operacional e eficiência em campo.

Atualmente existem duas versões para o sistema, o AGCOMMAND Standard, que está disponível para tratores, colheitadeiras e pulverizadores e o AGCOMMAND Advanced disponível somente para colheitadeiras (MF 32, e Axiais), no Quadro 1 um detalhamento e comparativo entre as duas versões disponíveis.

Quadro 1 – Informações do Sistema AGCOMMAND

AGCOMMAND		AGCOMMAND ADVANCED
» Posição, data, hora		» Posição, data, hora
» Horas do Motor		» Horas do Motor
» Status da Operação	← Dados →	» Status da Operação
» Velocidade de Deslocamento		» Velocidade de Deslocamento
» RPM do Motor		» Até 25 Mensagens rede CAN da Máquina

AGCOMMAND		AGCOMMAND ADVANCED
A cada 60 segundos	← Registra →	A cada 10 segundos
A cada 15 minutos	← Envia →	A cada 10 Minutos

FONTE: Guia de Informações do Sistema AGCOMMAND

O AGCOMMAND fornece informações como a localização de máquinas em tempo real, e ainda pode indicar o status atual da máquina (em operação, em manobra, em transporte ou parada).

É possível mapear um histórico de funcionamento, e localização da máquina, área trabalhada, status de operação, etc. Podendo comparar diretamente o desempenho e eficiência de até cinco máquinas. Controlar operações em tempo real e informar se estão sendo executadas conforme o planejado.

Algumas de suas aplicabilidades são as identificações dos gargalos de produção nas operações atuais e melhorar processos, para estas os serviços disponibilizados possuem os seguintes recursos:

Cerca Limitadora: É um limite geográfico definido por um raio, podendo abranger toda a propriedade. Um alarme pode ser atribuído a este limite, caso a máquina o ultrapasse a área demarcada.

Geo-Forma: É um limite geográfico que pode ser moldado conforme a forma de uma área específica. Os dados operacionais e de eficiência podem ser individualmente analisados para estas áreas específicas

Segundo Gestor a aplicação pode fornecer dados importantes com os relatórios disponíveis no total são oito os tipos de relatórios que estão disponíveis no sistema

AGCOMMAND, e somente poderão ser gerados no website do usuário. Alguns deles, como o Relatório de Custos e o Relatório de Comparação, estarão disponíveis na versão Advanced.

- Relatório de Dados
- Relatório de Serviço
- Relatório de Tempo do Motor (Horimetro)
- Relatórios de Eficiência
- Relatório de Campo
- Relatório de Tendência de Custos (somente versão Advanced)
- Relatório de Comparação (somente versão Advanced)
- Relatório de Custo (somente versão Advanced)
- Velocidade média que máquina opera;

Atualmente as principais utilizações do sistema AGCOMMAND são:

- Controlar as operações em tempo real e avaliar se estão sendo executadas conforme o planejado.
- Identificar os gargalos de produção nas operações atuais e melhorar processos, tomando decisões em cima dos históricos e relatórios gerados pelo sistema.
- Recursos de Cerca Limitadora e Geo-forma que permitem atribuir alarmes e relatórios às áreas específicas, o que propicia o controle e a análise de forma mais focada e eficiente.
- Avaliar e classificar o desempenho dos operadores para identificar necessidades e decisões.

Especificações técnicas do sistema de Telemetria:

- Banco de dados: PostgreSQL
- Tamanho do banco de dados: 2 TeraBytes
- Taxa de crescimento anual: 400 Gigabytes
- Quantidade de usuários: 864
- Tempo de operação: 5 Anos
- Quantidade de máquinas monitorada: 1800 +

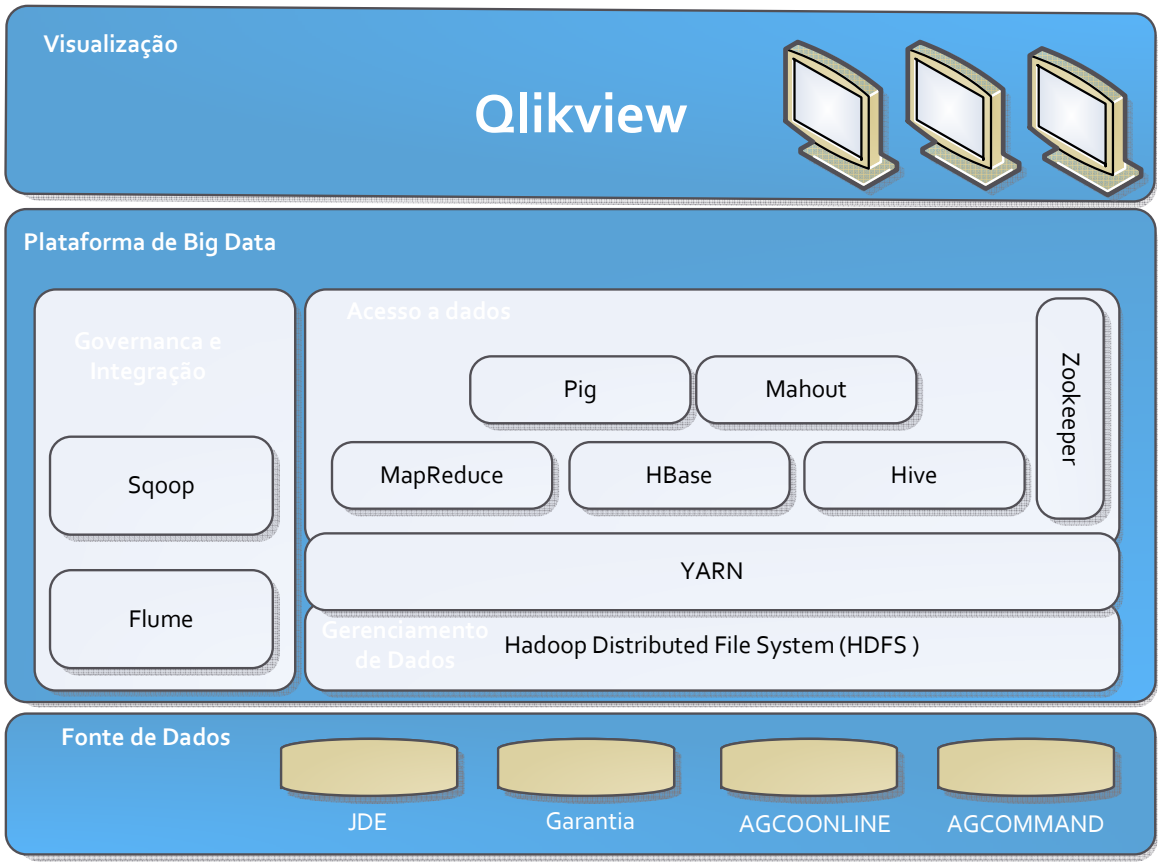
A taxa de crescimento anual aumenta a cada nova ativação do sistema de telemetria em uma nova máquina, crescendo cerca de 200 Megabytes por ano de dados por sistema adicionado, porém depende muito das configurações utilizadas.

4.2 SUGESTÃO DE ARQUITETURA PARA SOLUÇÃO DE BIG DATA NA AGCO

Com base nas informações levantadas, a seguir é apresentada uma proposta de uma solução de Big Data, onde a mesma deverá ser capaz de atender as necessidades do problema em questão, para facilitar o entendimento foi elaborado o desenho da arquitetura, que está representada na Figura 6.

A arquitetura foi desenhada de modo simplificado a fim de prover maior facilidade para o seu entendimento, foi divida em apenas três camadas, uma referenciada como fonte de dados responsável por prover os dados que serão utilizados, outra camada denominada Plataforma Big Data, que possui os recursos de Big data conforme estudado na fundamentação teórica que fica responsável pelo “processamento” e por fim porém não menos importante, está a camada de visualização, sendo está responsável por disponibilizar as informações resultantes do “processamento” e analise executadas pela plataforma de Big Data.

Figura 6 – Proposta de Arquitetura para Big Data



FONTE: Elaborado pelo autor

4.2.1 Fonte de Dados

A camada de fonte de dados possui como responsabilidade prover os dados para serem processados, analisados e posteriormente disponibilizados, os dados serão providos pelos sistemas JDE, AGCOONLINE, Garantia e AGCOMMAND conforme Figura 6.

4.2.2 Plataforma de Big Data

Para a camada denominada plataforma de Big Data, foi adaptado a partir da Figura 3, de acordo com as necessidades necessárias para resolver o problema em questão, na Figura 6, estão às ferramentas necessárias para o estudo.

Para facilitar o entendimento e a responsabilidade de cada camada e suas respectivas ferramentas, abaixo estas foram exploradas conforme o autor LUBLINSKY (2013)

- HDFS – Sistema de arquivo distribuído do Hadoop, solução de armazenamento que tem como responsabilidade gerir o sistema de arquivos, servindo como base para as demais ferramentas.
- Sqoop – Tem como responsabilidade fornecer a conectividade para mover dados entre bancos de dados relacionais, bando de dados de data warehouses e Hadoop.

- Zookepper – Tem como responsabilidade a coordenação Flume – Tem como finalidade validar, limpar, transformar, reduzir, sendo capaz de trabalhar com um grande volume de dados oriundo das mais diversas fontes e move-las para dentro do Hadoop da forma mais eficiente possível
- YARN – YARN significa "Yet Another Resources Negotiator" Monash, (2012). Sua responsabilidade é prover a redução da dependência do MapReduce e outras ferramentas do Hadoop. Criando uma camada de abstração para adicionar ou retirar componentes, exemplo interfaces de programação.
- MapReduce – Modelo de programação para sistemas distribuídos, com processamento paralelo, o processamento é divido em duas etapas, uma chamada Map, que consiste no mapeamento e validação dos dados e a outra chamada Reduce que recebe os dados da fase do Map e para gerar o resultado final.
- HBase – Banco de dados NoSQL orientada a coluna construído sobre o HDFS, o HBase tem como responsabilidade prover um acesso rápido para leitura / gravação com grandes volume de dados entre diversas ferramentas.
- Hive- Uma linguagem de alto nível SQL-like usado para executar consultas sobre os dados armazenados no Hadoop, o Hive permite que desenvolvedores não familiarizados com a forma de escrever em MapReduce possam escrever consultas de dados que são traduzidos em trabalhos de MapReduce no Hadoop. Assim como Pig, Hive foi desenvolvido para ser uma camada de abstração, orientada para os analistas de banco de dados familiarizados com as linguagens SQL e Java.
- Pig – Uma abstração sobre a complexidade da programação do MapReduce, a plataforma de desenvolvimento Pig possui ambiente de execução e uma linguagem de script (Pig Latin) seu compilador traduz Pig Latin em seqüências de programas para o MapReduce.

- Mahout – Biblioteca para aprendizado de máquina e mineração de dados que fornece implementações de MapReduce, para algoritmos populares incluindo algoritmos de analise preditiva, testes de regressão, e modelagem estatística do serviço distribuída do Hadoop. O Zookepper é utilizado para coordenar diversos componentes, exemplos: o Hbase, Hive, Pig, Mahout.

4.2.3 Visualização

A visualização dos resultados deve prover ao usuário fácil interpretação dos dados para isso a ferramenta de visualização que será utilizada na exposição dos resultados disponibilizados pela plataforma de Big Data, será o Qlikview. A AGCO já possui as licenças necessárias para tal utilização, esta ferramenta é capaz de trabalhar de maneira eficiente com a plataforma de Big Data proposta.

4.2.4 O funcionamento da solução de Big Data

A solução de Big Data proposta funcionará da seguinte forma: as fontes de dados proverão os dados dos sistemas: JDE; Garantia; AGCOONLINE e AGCOMMAND, estas fontes serão acessadas pelo Apache Sqoop, e posteriormente gravados no banco de dados HBase. Antes da gravação e do acesso, Apache Flume fará o tratamento dos dados retirando o ruído, para que isso aconteça será necessário a criação das regras na ferramenta.

O HBase que está suportado pelo sistema de arquivos HDFS, e interfaceado a este pelo YARN. Um MapReduce job será desenvolvido com a utilização Pig Latim (linguagem fornecida pelo Apache Pig) . Este realizará o trabalho de processamento utilizando algoritmos de predição criados no Apache Mahout. Os algoritmos criados no Mahout são responsáveis pela predição. Com tal objetivo, utilizarão conceitos de aprendizado de máquinas, onde o algoritmo se alimenta e se reprocessa a cada execução, aumentando sua assertividade quanto à predição.

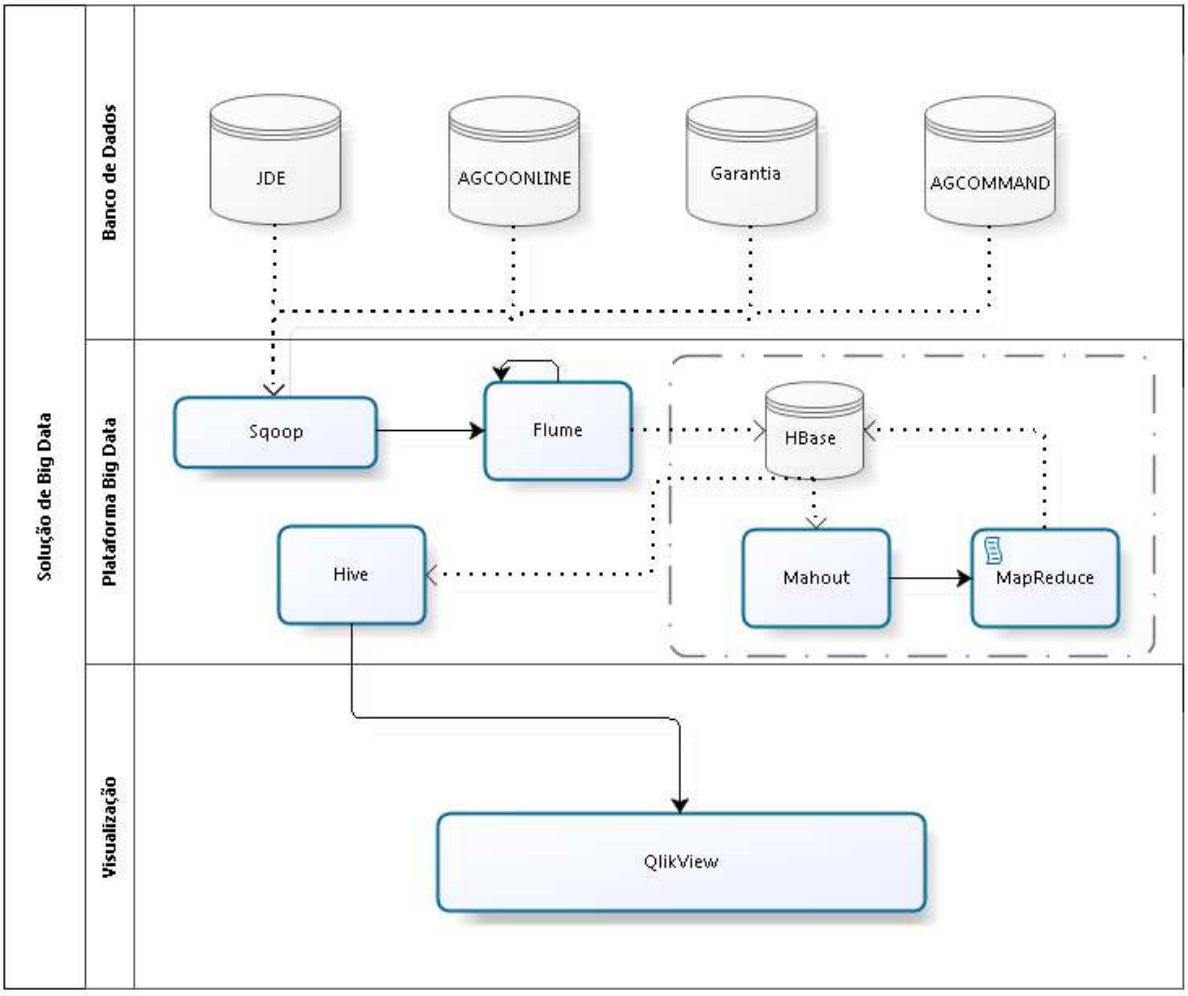
Após a conclusão do processamento dos algoritmos, os resultados serão gravados novamente no HBase. Posteriormente, Apache Hive fornecerá, através de seu HiveQL, uma

interação dos resultados gravados no banco de dados distribuído. Todos os processos descritos acima possuem suas configurações coordenadas pelo Apache ZooKeeper.

Por último o Qlikview disponibilizará os resultados para os usuários, visando fácil visualização através de seus painéis intuitos e dinâmicos.

Uma abordagem de forma visual é feita na Figura 7 onde é apresentado o fluxo dos dados da solução de Big Data proposta. O caminho que os dados seguem passando pelo processamento e por fim as informações são disponibilizadas.

Figura 7 – Fluxo de Dados da Solução Big Data



FONTE: Elaborado pelo autor

4.3 ANÁLISE DOS DADOS

Com base nos dados coletados dos sistemas utilizados para a solução de Big Data e para que se consiga criar uma análise preditiva precisamos olhar o passado e identificar padrões capazes de auxiliar na predição buscando assim as principais informações de cada sistema com base nos dados coletados.

No sistema JDE pode-se utilizar os dados referentes às configurações dos produtos e materiais utilizados, ordem e data da fabricação, lotes/séries dos itens rastreáveis, pedidos e faturamento de peças para atendimento de máquinas paradas ou em garantia.

Para o sistema AGCOONLINE utiliza-se os dados do PDI – *Pré-Delivery-Inspection* – formulário com *check-list* de verificação antes da entrega do produto para o cliente final dos processos de garantia e pedido de peças na modalidade Garantia.

O sistema de Garantia certamente irá prover a maior quantidade de dados para a descoberta de padrões, os principais são a data de faturamento e data de entrega do produto, data de abertura e fechamento da ordem de serviço, séries de monobloco, motor, eixos e transmissão, dados referente às falhas como peça causadora e código de falha e defeito, descrição e ação tomada, bem como as peças que foram trocadas, dados do cliente que está com a máquina naquele momento, em que cultura a máquina está trabalhando e o horímetro da máquina.

O sistema de telemetria AGCOMMAND proverá os dados praticamente em tempo real para a análise, utilizam-se dados de tempo do motor (horímetro), referentes a serviços e campo, dados geográficos fornecidos pelo GPS, eficiência da máquina, e quando a analise for realizada com colheitadeiras categorizadas na versão *Advanced* poderá contar com os dados referentes à tendência de custos além da comparação entre máquinas.

Os dados que foram selecionados em cada sistema são copiados de suas tabelas de origem para dentro da solução de Big Data no banco de dados HBase. Para os sistemas JDE, AGCOONLINE e Garantia a cópia dos dados pode ser realizada uma vez ao dia, porém para o sistema AGCOMMAND, que recebe os dados a cada 10 ou 15 minutos, a cópia deve ser de realizada de hora em hora, e a cada cópia um novo processamento do algoritmo de predição deve ser executado, isso para que os dados mantenham-se sempre atualizados.

4.3.1 Identificando padrões

De maneira simples e fácil compreensão Loh (2014) explica de onde vem e como surgem os padrões:

> A identificação de padrões é parte da nossa vida. A descoberta de padrões iniciou há milhares de anos atrás. Nossos antepassados conseguiam prever as variações do tempo, as estações, os ciclos das plantações, as fases lunares e eclipses, e até mesmo o surgimento de reis. E hoje em dia não é diferente. Quem não dá palpites sobre como será o tempo, se vai chover, fazer sol, calor, observando as nuvens? Ou se o próximo inverno será mais frio ou menos frio do que o ano anterior, pelo que viu no outono? Se um local público vai lotar ou não para um evento, observando o movimento das pessoas chegando? Ou quantas pessoas há num concerto ao ar livre num parque público, lembrando o último evento que ocorreu ali? Mesmo algumas superstições são exemplos de padrões, que acreditamos que irão se repetir. Numa entrevista de negócios, usar a mesma roupa de um acontecimento bom. Sentar no mesmo lugar do último título para torcer por seu time. Não quebrar espelho, pois quando isto ocorreu, um evento de má sorte também ocorreu junto. (LOH, 2014, p.17).

Segundo LEDOLTER (2013) os dados das redes sociais contêm informações sobre a presença de links entre os milhares ou milhões de indivíduos, incluindo características demográficas dos indivíduos (tais como sexo, idade, renda, raça e educação) que podem conectar um individuo a outro, ou não. O Google possui vasta informação sobre os 100 milhões de usuários e o Facebook possui muito mais informações em seu banco de dados. Os sistemas de recomendação desenvolvidos por empresas como a Amazon e Netflix utilizam as informações demográficas disponíveis e as modalidades de compra e aluguel referente a milhões de clientes, para pesquisar e encontrar padrões de consumo, podendo assim sugerir novas compras ou alugueis para seus clientes.

Em uma simulação onde analisa-se os dados coletados a partir dos sistemas pertencentes ao estudo caso, identificou-se que máquinas do tipo colheitadeira modelo Massey Ferguson MF32 SR fabricadas nos meses de Janeiro a setembro de 2013, utilizadas principalmente nas culturas de trigo e soja, na região sul e centro oeste, tem como peça causadora de falhas a bomba d’agua, que vaza e acaba gerando a parada da máquina causando grandes prejuízos ao produtor. Este problema pode ocorrer tanto nas cem primeiras horas de trabalho quanto ao longo do tempo. Os principais sintomas identificados nestes casos são aquecimento e perda de performance do motor, podendo a máquina parar em um dois ou três dias.

Com o padrão identificado, a solução de Big Data deve ser capaz de analisar todos estes dados e identificar a probabilidade que as próximas colheitadeiras terão de apresentar este problema, quando encontrada a máquina ou o conjunto de máquinas que se encaixam neste perfil, o sistema mostra a localização das máquinas e probabilidade que cada uma tem de estragar através do padrão identificado. Deve desconsiderar as que já estragaram e quando encontrar alguma outra relação entre elas automaticamente a probabilidade aumenta na Figura 8 um exemplo desta visualização.

Figura 8 – Visualização de máquinas com problema

FONTE: Elaborado pelo autor

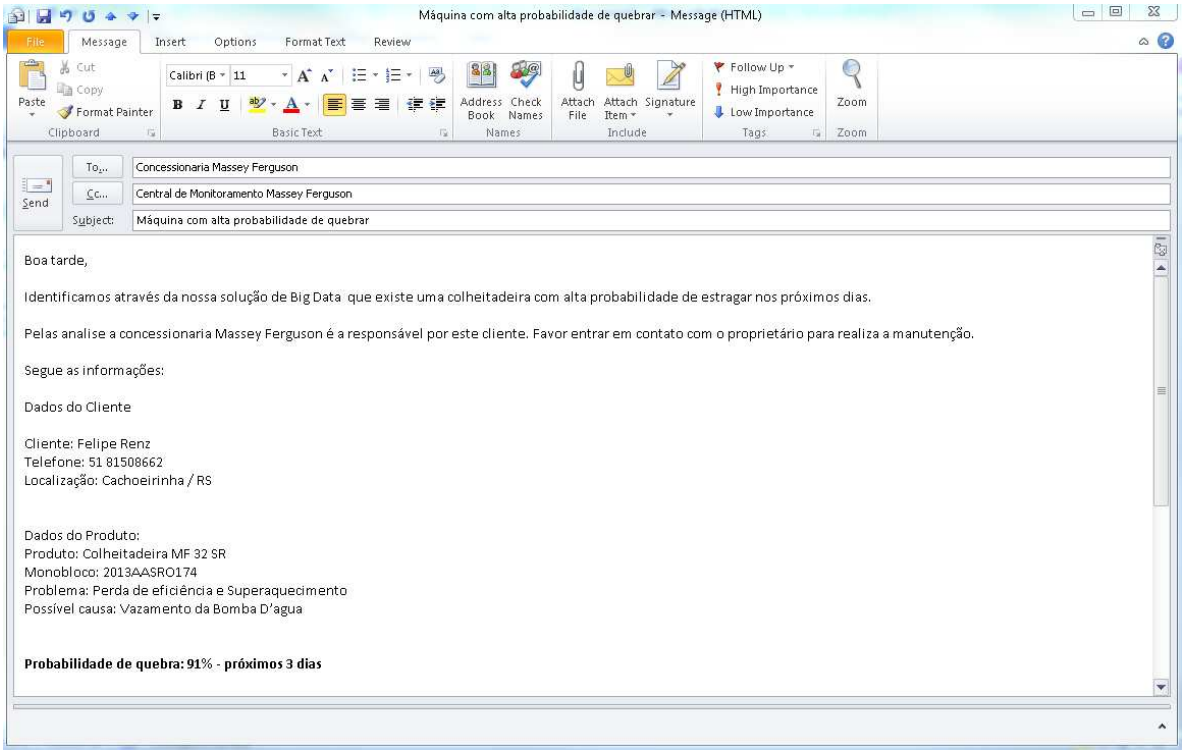
As colheitadeiras em vermelho possuem grande probabilidade de quebrar, as que estão sinalizadas com amarelo, estão com algum sintoma, que se não for tratado na logo devem mudar a cor para vermelho, e todas que estão na cor verde, não apresentam nenhuma anomalia identificada.

Com os dados coletados pelo sistema de telemetria identifica-se que determinada colheitadeira está perdendo eficiência e / ou aquecendo, e assim que os dados são coletados, processados e identificado o padrão de ocorrência, a anomalia é informada. Isso quase que em tempo real, podendo ainda, encontrar a causa do problema se já houver um padrão que

identificou está e assim estimar uma previsão para a quebra da colheitadeira baseado em correlação com eventos passados.

Existem varias maneiras que pode ser utilizada para alertar que uma máquina está a perigo de quebra, através da solução de Big Data proposta que armazena muitas informações de diferentes lugares, a mesma pode identificar a concessionária responsável pelo cliente e enviar um e-mail de alerta conforme visualizado na Figura 9.

Figura 9 - Exemplo de e-mail de alerta enviado pela Solução Big Data



FONTE: Elaborado pelo autor

Na Figura 8 pode-se utilizar o botão identificado como *Create Report* para que seja gerado relatório com o status de cada colheitadeira, podendo enviar o e-mail de alerta demostrado na Figura 9 para cada máquina com potencial de quebra, (sinalizadas com a cor vermelha), e antecipando-se, a possíveis problemas enviar um alertas de perigo para as máquinas que estão sinalizadas em amarelo.

Pode também funcionar de maneira interativa, onde o usuário clica sobre a máquina que apresentou anomalia, demostrada no mapa como na Figura 8, diversas informações do produto que foram capturadas pela solução Big Data ficam disponíveis para visualização imediata, visualizada na Figura 10.

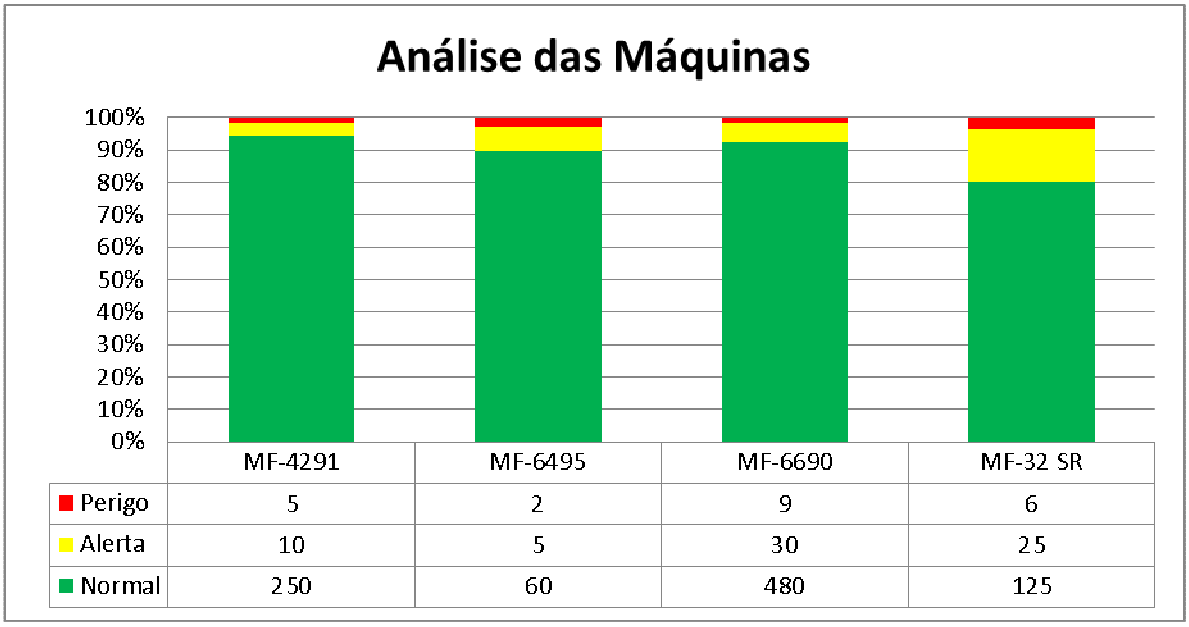
Figura 10 – Detalhamento de uma máquina com problema

FONTE: Sistema AGCOMMAND

Na Figura 10 algumas informações possíveis de serem extraídas, como o nível de combustível, óleo, velocidade, rotações por minuto, temperatura tração, força horas de trabalhos e umidade. Juntos a estes dados pode ser acrescido também informações referente ao cliente, histórico da máquina, entre outros.

Pensar em utilizar uma central de monitoramento para as máquinas que e que a cada processamento realizado pela da solução Big Data gera informações atualizadas do estado de funcionamento, exemplo de relatório na Figura 11.

Figura 11 – Relatório das máquinas com potencial de quebra



	MF-4291	MF-6495	MF-6690	MF-32 SR
Perigo	5	2	9	6
Alerta	10	5	30	25
Normal	250	60	480	125

FONTE: Elaborada pelo autor

No relatório demostrado por meio gráfico na Figura 11, utiliza-se quatro modelos de máquinas, divida em três estados, demostrando a quantidade de máquina em cada um dos estados, divida em cores, em verde estão sinalizadas as máquinas no estado normal, em amarela as máquinas no estado de alerta e por fim as sinalizadas em vermelho estão apresentando perigo de quebra.

4.4 ADERÊNCIA DA SOLUÇÃO DE BIG DATA NA AGCO

As soluções de Big Data vieram para ficar nas companhias, trazendo consigo uma gama de tecnologias e conceitos conforme explorados na fundamentação teórica. Com essas diretrizes a AGCO pode desenvolver estratégias que representam ganhos substancias em relação aos seus concorrentes.

Durante as entrevistas com os gestores identificou-se um forte alinhamento do Big Data com as diretrizes da organização, a possibilidade de se obter vantagem competitiva, a partir da solução de Big Data, para o processamento de grandes quantidades de dados, com a

capacidade de predizer um evento futuro, através da analise probabilística identificando a ocorrência de defeito na máquina em operação, gerando uma maior proximidade com seus clientes.

O estreitamento do relacionamento da rede de concessionarias da AGCO espalhadas pela América do Sul gerada a partir da solução de Big Data, com seus clientes, tende a proporcionar ganhos para a companhia. Considerando a seguinte situação:

Um cliente que no momento de realizar a colheita, com duração média de trinta dias, onde as colheitadeiras são extremamente exigidas e tem como premissa não parar, pois perder o período correto da colheita, pode gerar grandes prejuízos, no vigésimo dia de colheita, a concessionaria responsável pelo atendimento daquele cliente, liga para o cliente, solicitando um agendamento para a manutenção de uma de suas colheitadeiras.

Está ação será possível de ser tomada com os relatórios disponíveis na solução de Big Data, identificando a relação daquelas máquinas que apresentaram um problema similar no passado, com os dados de telemetria foi possível identificar que está colheitadeira estava com alta probabilidade de ocorrer o mesmo problema que as demais apresentaram. Uma ação proporcionada a partir da solução de Big Data com a utilização da analise preditiva.

O funcionário da concessionaria realiza a troca das peças necessárias que foram incialmente indicadas pelo Big Data, que senão fosse trocada naquele momento poderia ocasionar uma parada que poderia levar dias para resolver o problema, ou ainda gerar problemas maiores, pois os estragos na máquina provavelmente seriam muito maiores o que acabaria por gerar prejuízos para o cliente, e em pouco tempo a colheitadeira volta a trabalhar, estando apta a completar com sucesso sua atividade de colheita dentro do tempo esperado.

Nesta situação onde foi possível identificar uma probabilidade e predizer uma falha, é possível aumentar substancialmente a qualidade percebida pelos clientes, criando uma nova maneira de relacionamento entre ele, a concessionaria e a fábrica, isso coloraria a rede de concessionarias em um excelente posicionamento diante dos clientes trazendo consigo o fortalecimento da marca no mercado e gerar por consequência uma satisfação aguda aos clientes.

Através desse exemplo consegue-se identificar inúmeros ganhos que a utilização de uma solução de Big Data através de um modelo preditivo, com o intuito de prever uma falha nas máquinas pode trazer para a AGCO, mostrando uma aderência incomum com o negócio e principalmente suas estratégias de negócio.

5 CONSIDERAÇÕES FINAIS

O Big Data a partir de Taurion (2013) que cita os 5V’s como velocidade, volume, variedade, veracidade gerando valor, como o poder de transformar as organizações, a exploração feita no estudo de caso, se explorada pode revolucionar a maneira como a AGCO se relaciona com seus concessionários e clientes, colocando-a em um novo patamar, muito a frente de seus concorrentes, aumentando de maneira substancial a qualidade percebida pelos clientes.

A questão proposta para o estudo de caso. **Como a AGCO pode obter vantagem competitiva utilizando Big Data e análise preditiva em manutenção de máquinas agrícola?** Foi amplamente explorada no desenvolvimento do trabalho e respondida com propriedade, demonstrando com exemplos a possibilidade de obtenção de vantagem competitiva.

A proposta dos cinco objetivos específicos foram claramente alancados no decorrer do trabalho. No primeiro objetivo especifico, referente a busca da teoria sobre a solução Big Data, exaustivamente explorada e analisada, por meio de livros, artigos científicos tanto no meio eletrônico quanto no meio físico, a identificação do que a solução de Big Data se propõe a resolver. A importância e complexidade que a implantação da arquitetura de uma solução de Big Data exige.

A exploração dos 5V’s foi desafiadora entender como cada um deles se encaixa na solução, e fundamentalmente a responsabilidade do cientista de dados, que tem e terá cada vez mais um papel de importantíssimo nas organizações pois ele poderá descobrir novas maneiras e ou negócios através da exploração eficiente dos dados.

No segundo objetivo especifico que propõe a analise das fontes de dados da AGCO, foi possível entender quais os principais sistemas poderiam fazer parte da solução de Big Data, a relevância de cada um dos sistemas foi explorada através das entrevistas dinâmicas com gestores e colaboradores da organização.

Através da identificação dos sistemas que poderiam compor a solução junto com a exploração realizada na fundamentação teórica foi possível alcançar o terceiro objetivo especifico onde foi proposta a arquitetura para utilização de uma solução de Big Data, da fundamentação teórica entendeu-se o que é necessário para que se tenha uma arquitetura robusta de Big Data,.

Proposta a arquitetura, tendo definido os requisitos necessários com as camadas de fonte de dados, plataforma de Big Data e a camada de visualização, e criado o fluxo dos dados, chega a hora do quarto objetivo especifico, planejar a utilização dos dados, para análise preditiva, previamente explorada na fundamentação teórica deste trabalho, e a identificação de padrões, com estas informações é possível prever em níveis probabilísticos a possibilidade de novas máquinas agrícolas apresentarem defeitos, caso já se tenha ocorrências destes eventos no passado.

Por fim, porém não menos importante o quinto objetivo especifico: A aderência de uma solução de Big Data com a AGCO, foi analisada e constatada por meio de simulação, que existe sim uma forte aderência com a organização, que pode colocar a AGCO em outro patamar, diante seus concorrentes e criar um forte relacionamento com seus clientes e concessionários, tendo assim um grande diferencial competitivo.

A solução de Big Data tem forte potencial para gerar valor para as organizações, dentre os diversos livros e artigos pesquisados, uma simples analogia, porém muito forte, chamou minha atenção, onde Hey, Tansley e Tolle, (2011) afirmam que os dados são hoje considerados o petróleo do século XXI. Em uma breve analise, vemos que se o petróleo não for manufaturado de nada serve a não ser para poluir o meio ambiente, analisando os dados se eles não forem “manufaturados” estarão apenas poluindo as organizações, mas se explorados de maneira eficiente estarão gerando riquezas para as empresas por meio de novos negócios e ou redução de custo, diferenciação, o mesmo acontece com o petróleo quando ele é manufaturado, gera riqueza para as companhias, através de seus inúmeros subprodutos utilizado pelas mais diversas empresas.

Concluída a identificação dos objetivos do projeto, executada as pesquisas para fundamentação teórica e que serviram de base para o desenvolvimento do trabalho, aplicando as metodologias e procedimento de estudos, para a aplicação no estudo de caso, certificada a da aderência com o negócio da AGCO, o projeto será levado para analise da área de negócio, afim de que seja aprovado para posterior implantação.

REFERÊNCIAS

GANTZ, J. e REINSEL, D. **The Digital Universe Decade – Are You Ready?,** 2010. Disponível em: http://www.emc.com/collateral/analyst-reports/idc-digital-universe-are-you-ready.pdf Acesso em: 5 jul 2014.

_____. **Extracting value from chãos**, 2011. Disponível em: http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf Acesso em: 5 jul 2014.

GIL, A. C. **Métodos e técnicas de pesquisa social**. São Paulo: Atlas, 2008.

GOMES, L. F. A. M. **Teoria da decisão**. São Paulo: Thomson, 2007.

GOODE, W. J. & HATT, P. K. **Métodos em Pesquisa Social**. 3ªed., São Paulo: Cia Editora Nacional, 1969.

HEY, T.; TANSLEY, S.; TOLLE, K. **O Quarto Paradigma: Descobertas Científicas na era da eScience**. São Paulo: C, 2011.

HURWITZ, J.; KAUFMAN, M.; HALPER, F.; KIRSCH, D. **Big Data For Dummies**. New York: Wiley, 2013.

KAPLAN, R. S.; NORTON, D. P. **A execução premium a obtenção de vantagem competitiva através do vinculo da estratégia com as operações do negocio**. Rio de Janeiro: Elsevier Campus, 2009.

LANEY, D. **3D Data Management: Controlling Data Volume, Veocity, and Variety,** 2001.

LEDOLTER, J. **Data Mining and Business Analytics with R**. Hoboken, N.J.: Wiley, 2013.

LOH, S. **BI na Era do Big Data para Cientistas de Dados : Indo além de cubos e dashboards na busca pelos porquês, explicações e padrões**. Porto Alegre, 2014.

LUBLINSKY, B. **Professional Hadoop solutions**. 1st edition ed. Indianapolis, IN: John Wiley and Sons, 2013.

Luis Henrique M. K. Costa1, Marcelo D. de Amorim2,Miguel Elias M. Campista1, Marcelo G. Rubinstein3,Patricia Florissi4e Otto Carlos M. B. Duarte(2012). Disponível em: http://www.gta.ufrj.br/ensino/cpe728/CAC12.pdf Acesso em: 2 jul 2014.

MARCONI, M. DE A.; LAKATOS, E. M. **Técnicas de pesquisa: planejamento e execução de pesquisas ; amostragens e técnicas de pesquisa ; elaboração, análise e interpretação de dados**. São Paulo: Atlas, 1999.

MAYER-SCHÖNBERGER, V.; CUKIER, K. **Big Data - Como Extrair Volume, Variedade, Velocidade e Valor da Avalanche de Informação Cotidiana**. 1. ed. Campus, 2013.

MOHANTY, S.; JAGADEESH, M.; SRIVATSA, H. **Big Data imperatives: enterprise big data warehouse, BI implementations and analytics,** 2013.

Monash, C., 2012. Hadoop YARN- Beyond MapReduce. [Online] Disponível em: http://www.dbms2.com/2012/07/23/hadoop-yarn-beyond-mapreduce Acesso em: 20 Jul 2014.

ORACLE. JD Edwards EnterpriseOne. Disponível em: http://www.oracle.com/br/products/applications/jd-edwards-enterpriseone/overview/index.html Acesso em: 15 jul. 2014

PEETS, S.; MOUAZEN, A. M.; BLACKBURN, K.; KUANG, B.; WIEBENSOHN, J. Methods and procedures for automatic collection and management of data acquired from on-the-go sensors with application to on-the-go soil sensors. **Computers and Electronics in Agriculture**, v. 81, p. 104–112, 2012. Acesso em: 5 ago 2014.

PORTER, Michael E. **. Vantagem Competitiva: Criando e Sustentando um Desempenho Superior**. Editora Campus, Rio de Janeiro, 1990.

PRAJAPATI, V. **Big Data analytics with R and Hadoop set up an integrated infrastructure of R and Hadoop to turn your data analytics into Big Data analytics**. Birmingham: Packt Publishing, 2013.

SAWANT, N. E SHAH, H. **Big Data Application Architecture Q&A A Problem - Solution Approach.** Dordrecht: Springer, 2014.

TAURION, C. **Big Data**. Brasport, 2013.

The Economist. Data, data everywhere, 2010. Disponível em: http://www.emc.com/collateral/analyst-reports/ar-the-economist-data-data-everywhere.pdf Acesso em: 7 jul 2014

TOLE, A. A. Big Data Challenges. **Database Systems Journal**, v. 4, n. 3, p. 31–40, 2013.

Trauth, E.M.; O'Connor, B. **A study of the interaction between information, technology and society: an illustration of combined qualitative research methods.** In: Nissen, H.E., Klein, H.K., Hirschheim R. Information Systems Research: Contemporary Approaches & Emergent Traditions. Amsterdam, 1991.

TURBAN, E. **Tecnologia da informação para gestão transformando os negócios na economia digital**. Porto Alegre: Bookman, 2010.

YIN, R. K. **Estudo de caso: planejamento e métodos**. Porto Alegre (RS): Bookman, 2010.

YIN, R. K.; GRASSI, D., TRAD. **Estudo de caso**. Porto Alegre: Bookman, 2005.