Fluxo de Grandes Dados em Ambientes Corporativos

Felipe Manikowski[1], Gustavo C. de P. Santos[1], Luiz A. C. Júnior[1] e Luiz C. G. Junior

[1]Departamento Acadêmico de Informática
Universidade Tecnológica Federal do Paraná (UTFPR)
Avenida Sete de Setembro, 3165, 80.230-901 – Curitiba – Paraná – Brasil

{gustavosantos,felipem,ljunior}@alunos.utfpr.edu.br

gomesjr@dainf.ct.utfpr.edu.br

Resumo. *No presente artigo foram abrangidos os processos de um fluxo de grandes dados em um ambiente corporativo, desde a coleta até a exibição, e como esses processos podem ser realizados utilizando tecnologias atuais. Como exemplo, simulou-se um e-commerce onde os diretores desejaram obter um panorama geral a respeito dos produtos a venda na loja, para identificar oportunidades e problemas. Para isso, recorreram à equipe de Business Intelligence da organização. Ao final dos processos é elaborada uma matriz que classifica os produtos da loja em 4 quadrantes, onde cada um exigirá uma estratégia diferente para aumentar as vendas da loja.*

1. Introdução

De forma recorrente as empresas estão adotando o *Business Intelligence* (BI) como forma de estreitar o relacionamento com o consumidor, na expectativa de gerar lucro. Setores responsáveis exclusivamente pela administração do banco de dados relacionado com negócios começam a aparecer nas empresas, analisando os dados gerados por seus clientes. E agora, com o advento da *Big Data* (grandes dados) na área de BI, há como explorar ainda mais o relacionamento com os consumidores [Chen et al. 2012].

Contudo, embora este crescimento da mineração de dados seja benéfico para os setores de negócios, problemas relativos a infraestrutura podem vir agregados, uma vez que todos estes dados novos, gerados por usuários, sensores e outras fontes de dados onipresentes, têm um elevado impacto estrutural. Assim sendo, na tentativa de conter este problema, gigantes da computação, como o *Google*, que desenvolveu ferramentas como o *Google File System* (GFS) e modelos de programação *MapReduce*, criaram alternativas para encarar esta sobrecarga estrutural. Jim Gray, um pioneiro em *softwares* de banco de dados, percebendo todo este avanço na mineração de dados, nomeou isto em [Hey et al. 2009] como "O Quarto Paradigma". Ele argumenta que a única maneira de lidar com este crescimento seria o desenvolvimento de uma nova geração de ferramentas computacionais para administrar, visualizar e analisar dados massivos [Chen et al. 2014].

Esta problemática descrita por Gray é de interesse dos autores deste artigo, uma vez que passa por áreas da programação, da análise estatística, das práticas de banco de

dados, entre outras. Assim sendo, este trabalho é um desafio, um avanço no estudo de áreas de interesse.

Portanto, ao término do artigo espera-se abranger a melhor forma de tratar dados, criando novas estratégias, identificando erros e acertos e trazendo assim benefícios para o futuro das organizações. Logo, o projeto foca-se na coleta, armazenamento, tratamento e exibição de dados estruturados ou não para o benefício das corporações.

2. Objetivos

2.1. Objetivo Geral

- Abranger, através de um ambiente corporativo simulado, os processos de um fluxo de grandes dados, desde a coleta até a exibição, utilizando ferramentas tecnológicas atuais, como o ambiente de análise do *Google Analytics*, o *software* de gerenciamento de dados *MySQL* e o *software* de visualização de dados *Tableau*.

2.2. Objetivos Específicos

- Explicar de forma geral como é o algoritmo de extração de dados do *Google Analytics*;
- Tratar dados alegóricos e armazená-los de forma simulada em um *Data Warehouse*, fazendo uso do *software MySQL*;
- Manipular os dados de forma teórica e exibi-los, para que facilitem a geração de estratégias e tomadas de decisão por parte dos responsáveis, a partir do *software Tableau*.

3. Referencial Teórico

3.1. *Business Intelligence*

Business Intelligence são diretivas e *softwares* que visam o melhoramento das tomadas de decisão no ambiente corporativo. Como o texto [Davenport 2006] expõe: "O termo BI apareceu no final da década de 80, e abrange uma cadeia de processos e ferramentas que são utilizados para coletar, analisar e disseminar dados, sempre com objetivo de melhorar as tomadas de decisão".

Também compõe o *Business Intelligence* as tecnologias, sistemas e praticas que analisam os dados referentes aos negócios para darem uma melhor compreensão da corporação e do mercado. Os principais objetivos da BI são: Possibilitar a interação e o fácil acesso para dados diversos, possibilitar a manipulação e transformação desses dados e prover para os gestores e analistas de negócios as habilidades de conduzirem análises apropriadas e executarem tomadas de decisão [Lim et al. 2013].

3.2. *Data Warehouse*

Data Warehouse é um conceito fundamental do *Business Intelligence* [Lim et al. 2013].

Segundo a [Oracle], "é uma base de dados relacional desenvolvida para consulta e análise, além do processamento de transações (operações). Geralmente ela possui dados derivados de suas transações, mas também dados de outras fontes são armazenados". Seu objetivo é fornecer informações que auxiliem no processo de tomada de decisão, descrevendo o comportamento da organização com dados históricos relacionados de forma significativa para a análise gerencial e estratégica [ZIULKOSKI 2003].

O processo de desenvolvimento de um DW, o *Data Warehousing*, de acordo com [Machado 2000], [Kimball 1998], [J. Han 2001] e [W. Inmon 2001] deve incluir as seguintes atividades:

- Modelagem de dados;
- Extração, tratamento e limpeza dos dados originais;
- Criação do OLAP (*OnLine Analytical Processing* (análise de dados em tempo real));
- Análise das informações;

3.3. *Big Data*

Apesar dos estudiosos sobre o assunto ainda não chegarem a um consenso quanto a sua definição, este conceito pode ter significados mais aprofundados do que simplesmente "dados massivos". "De forma geral, *Big Data* significa as bases de dados que não poderiam ser percebidas, coletadas, gerenciadas e processadas pela tecnologia da informação tradicional em um tempo tolerável." [Chen et al. 2014].

A *International Data Corporation* (IDC), uma das líderes de maior influência em *Big Data* e em pesquisas na área, segundo [Chen et al. 2014], expôs a seguinte frase em [Gantz and Reinsel 2011]: "*Big Data* descreve uma nova geração de tecnologias e arquiteturas, designadas a extrair valores econômicos de volumes muito grandes de uma variedade abrangente de dados, utilizando uma alta velocidade de captura, descoberta e/ou análise". Com isto os grandes dados podem ser classificados em quatro Vs – volume, variedade, velocidade e valor (alto valor, mas baixa densidade). Para a IDC, *Big Data* significa "explorar os valores ocultos" [Chen et al. 2014].

4. Metodologia

A metodologia será baseada no estudo de caso do fluxo de grandes dados em um ambiente corporativo simplificado, onde será exposto como é possível alcançar dados da conversão (valor) de um produto em um *e-commerce*:

- A partir da extração dos dados da quantidade de visualizações e sessões pelo *Google Analytics* e inserção destes dados em um *Data Warehouse*, calcular qual o percentual de conversão de cada produto a venda, ou seja, qual o percentual de pessoas que compram o produto após acessar sua página. As informações geradas podem também expor indiretamente que o produto está com problemas de venda.

- Exibir estas informações utilizando a ferramenta de visualização *Tableau*, facilitando às tomadas de decisão por parte dos responsáveis.

5. Cronograma

- Etapa 1: Estruturar o cenário e as ferramentas que serão estudadas. (todos os membros)
- Etapa 2: Após a estruturação do cenário e as tecnologias escolhidas, pesquisar teorias de todos os processos do fluxo [coleta, armazenamento, tratamento e exibição]. (todos os membros)
- Etapa 3: Desenvolver um texto e uma apresentação de qualificação do projeto. (todos os membros)
- Etapa 4: Apresentar a qualificação aos professores e ao orientador do projeto. (todos os membros)
- Etapa 5: Implementar, por meio de uma documentação do projeto, a fundamentação teórica pesquisada no cenário escolhido. (todos os membros)
- Etapa 6: Apresentar o projeto aos professores e ao orientador do projeto. (todos os membros)

	Março		Abril		Maio		Junho	
	1ª Parte	2ª Parte	1ª Parte	2ª Parte	1ª Parte	2ª Parte	1ª Parte	2ª Parte
Etapa 1	X	X						
Etapa 2		X	X					
Etapa 3			X	X				
Etapa 4				X				
Etapa 5					X	X	X	X
Etapa 6								X

Tabela 1. Cronograma

6. Desenvolvimento

6.1. Cenário

O cenário escolhido para demonstração de exemplos é um comércio *online* que vende diversos tipos de produtos, o endereço de cada produto seguirá o modelo "http://www.lojasimulada.com.br/nome-produto-id.html", facilitando a inserção e a localização das informações de cada página no *Google Analytics* (GA), já que o mesmo traz dados relacionando com a *Uniform Resource Locator* (URL), ou no português, Localizador Padrão de Recursos, de cada item.

A lojasimulada.com.br vende cerca de 100.000 produtos, divididos em 100 categorias e recebe cerca de 1.000.000 de visitas por dia, contando todas as páginas. Não é

possível comprar um produto sem acessar a sua página. Essas visitas possuem diversas origens, como acessos diretos, acessos provenientes de páginas de busca, empresas de *marketing* que trabalham com divulgações, redes sociais, etc.

Os diretores e gerentes demandaram que a equipe de *Business Intelligence* os ajude a terem um melhor controle da loja, buscando informações sobre quais são os produtos que mais vendem, quais produtos estão com problemas e quais os motivos do sucesso e do fracasso de cada produto. Esses dados devem ser exibidos de forma clara e resumida, facilitando o entendimento por pessoas de todas as áreas de empresa, através de relatórios, *dashboards* (painéis de informações) e análises estatísticas.

Para isso, a equipe de BI buscará informações no GA da lojasimulada.com.br a fim de ter o conhecimento da quantidade de visualizações das páginas, rejeições e sessões dos usuários.

De forma simplificada, *Google Analytics* é uma ferramenta que auxilia os setores corporativos responsáveis pela mineração de dados dos clientes de uma dada corporação. Dados estes que podem estar relacionados com identidade e interesse daqueles.

As informações coletadas pelo GA são:

- Quantidade de pessoas que visitaram as páginas da empresa por dia, qual o sistema operacional utilizado por elas e até mesmo a localização territorial das mesmas.
- Origem do tráfego nas páginas, quais as palavras chaves utilizadas para chegar nelas e se foi usado um buscador para acessá-las [Chan 2011].

6.2. Extração

Com o acesso aos relatórios do GA, a necessidade é extrair estes dados provenientes da ferramenta por meio de um algoritmo desenvolvido pela equipe de programação a partir da *Application Program Interface* (API) do *Google Analytics*.

De maneira geral, uma API é um "conjunto de rotinas e padrões de programação para o acesso a um aplicativo de *software* ou plataforma baseada na *Web*. Uma API é criada quando uma empresa tem a intenção de que outros desenvolvedores criem produtos associados ao seu serviço" [Canaltech].

Inicialmente, a API implementada pela equipe de desenvolvimento da lojasimulada.com.br faz os *links* com o banco de dados realizando autenticações e atualizações neste. Em seguida um laço de repetição é ativado para a coleta de dados novos, referentes à quantidade de visualizações de cada página (*pageviews*), à quantidade de abandonos de cada página (*bounces*) e à quantidade de sessões abertas pelo usuário (*sessions*) – "A sessão é um grupo de interações que ocorrem em seu *website* em um determinado período. Por exemplo, uma única sessão pode conter várias exibições de página ou tela, eventos, interações sociais e transações de comércio eletrônico" [Google] –. Além da coleta, o algoritmo também organiza estas informações para que sejam inseridas no banco de dados no formato ID do produto (cada ID contém visualizações, rejeições e sessões). Todos estes dados são postos em seguida no *Data Warehouse*.

Diariamente este código será executado e os produtos que tiveram algum dos dados descritos acima selecionados, os quais foram gerados no dia anterior, serão inseridos no banco. Logo, todo dia serão inseridas na tabela de dados algo entre 0 a 100.000 novas linhas.

6.3. Armazenamento

A estrutura escolhida para o armazenamento dos dados que serão analisados será um *Data Warehouse* (DW), que é um banco de dados que guarda informações de forma íntegra e facilita a realização de análises em diversas dimensões [Favaretto 2007], integrando dados de vários ambientes da empresa – banco de dados de vendas, banco de dados do sistema ERP, etc. – O DW possibilitará que os dados provenientes do *Google Analytics* possam ser analisados junto com dados de todos os ambientes da empresa através de consultas.

As duas primeiras atividades, coleta e armazenamento, já são tratadas no próprio algoritmo de extração de dados. As duas últimas atividades serão melhor explanadas na próxima seção.

Neste artigo será utilizado como Sistema Gerenciador de Banco de Dados a ferramenta *MySQL*, por conta de sua flexibilidade e escalabilidade.

O DW terá o seguinte esquema para a tabela global de dados:

```
dw_venda (_id_pedido_, _id_produto_, datahora, status,
quantidade, frete, valor_absoluto, _id_cliente_);

dw_cliente (_id_cliente_, nome, endereco,
endereco_entrega, cep, cpf, celular);

dw_produto (_id_produto_, nome, marca,
categoria, preco_custo, preco_venda);

dw_analytics (_id_produto_, data, visualizacoes,
rejeicoes, sessoes);
```

A tabela `dw_venda` possui informações referentes a pedidos, seus produtos e id do cliente que comprou, a tabela `dw_cliente` possui as informações de todos os clientes cadastrados, assim como a `dw_produto` possui todos os dados de cada produto a venda, e por fim a tabela que é populada com as informações extraídas do GA, que é denominada `dw_analytics`.

6.4. Análise e Exibição

Com o DW já estruturado e populado, agora a equipe de BI necessita fazer as análises e pensar na forma de exibir essas informações de forma clara e objetiva. A seguinte métrica foi estipulada para ser analisada e exibida:

- Taxa de Conversão (resultado do cálculo: quantidade de pedidos de um determinado produto / total de visualizações deste produto) - Esta métrica vai mostrar quais produtos estão bem posicionados em relação ao mercado, ou seja, se suas descrições estão suficientes, afinal se o cliente visitou a página e comprou quer dizer, na maioria dos casos, que ele encontrou o que necessitava a um preço adequado.

Como essas análises podem ser feitas a nível de loja, categoria e produto, é possível indicar desde problemas gerais, quanto a problemas muito pontuais.

Porém, para mostrar esses dados ferramentas tradicionais não são viáveis, pois além de demandarem um certo tempo de processamento, a exibição das informações acaba ficando inelegível e difícil de ser filtrada, caracterizando assim um problema de *Big Data*. Então a equipe utilizou o *software Tableau*, uma ferramenta de exibição de dados voltada para BI e *Big Data*, onde é possível adicionar filtros direto nas exibições dos dados e criar painéis interativos a partir dos dados extraídos do *Data Warehouse*. Tudo isso feito de forma *online* e pelo próprio navegador de *internet*. O *Tableau* ainda permite o *download* da exibição em formato de planilha eletrônica.

Inserindo a seguinte *query* (consulta) no *Tableau*:

```
SELECT
v.id_produto,
a.data,
COUNT(DISTINCT v.id_pedido) as pedidos,
SUM(a.visualizacoes) as visualicacoes,
COUNT(DISTINCT v.id_pedido)/SUM(a.visualizacoes) as taxa_de_conversao
FROM
dw_analytics a
LEFT JOIN dw_venda v ON v.id_produto = a.id_produto
AND DATE_FORMAT(v.datahora, '%d-%m-%Y') = a.data
WHERE a.data > SUBDATE(NOW(),30)
GROUP BY v.id_produto, a.data;
```

Teremos como resultado uma tabela que trará dados referentes à quantidade de visualizações dos produtos que tiveram *pageviews* dentre uma determinada faixa de dias. A quantidade de linhas a serem retornadas vai ser X, onde X é quantidade de produtos que tiveram alguma visualização, ou seja, $X \leq 100.000 \times 30$ (que foi o período de dias escolhido). Estamos falando de uma grade quantidade de dados, o que seria impossível de tratar em ferramentas comuns de análises de dados, como planilhas, por exemplo.

Com os dados no *Tableau* conseguimos montar a seguinte exibição:

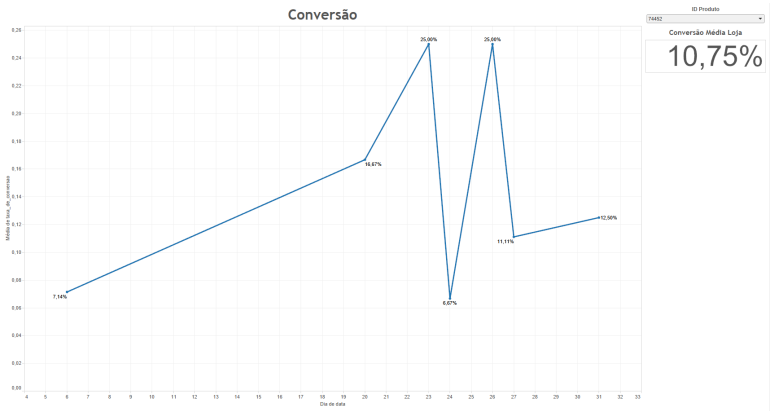

Figura 1. Gráfico de evolução da conversão de um produto no *Tableau* facilmente comparada com a média geral da loja

7. Resultados

Com as informações presentes no *Tableau* a equipe de BI chegou a conclusão que a melhor forma de exibir estes dados da taxa de conversão por produto é numa matriz de conversão onde os eixos são pedidos e visualizações. A partir das quantidades destas variáveis por produto relacionadas com a média de pedidos e visualizações de toda a loja, os itens a venda serão classificados em:

- Produtos exemplares - São os produtos "carro-chefe" da loja, possuem muitas visualizações e muitos pedidos. (Categoria A)
- Produto com algum problema perante ao mercado - São os produtos que recebem uma grande quantidade de visualizações, mas vendem pouco. (Categoria B)
- Produtos com potencial - São produtos que apesar do baixo número de visualizações, possuem um alto número de pedidos. (Categoria C)
- Produtos ruins - São produtos que vendem pouco e recebem poucas visualizações. São muito pouco influentes para a loja. (Categoria D)

Estrutura da matriz é:

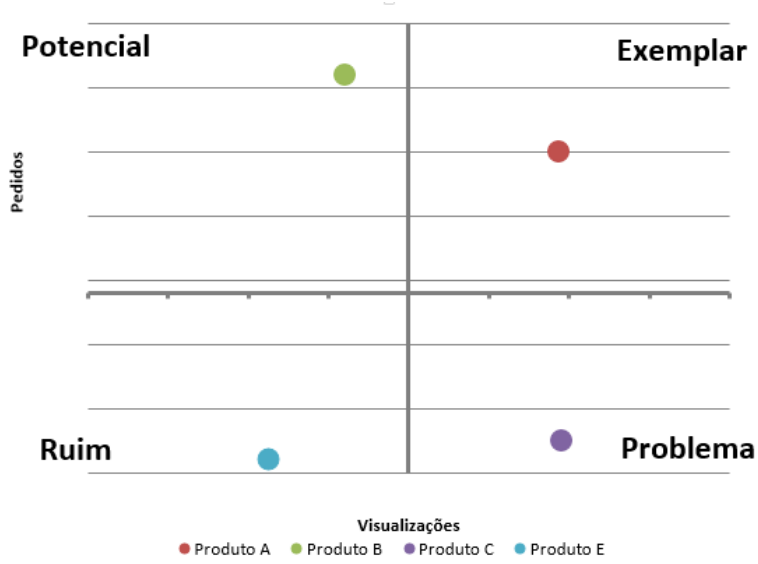

Figura 2. Matriz dos produtos de acordo com suas métricas relacionadas à média da loja

Para facilitar a visualização dos dados dos produtos, a equipe de BI irá exibir os quadrantes da seguinte forma para os tomadores de decisão:

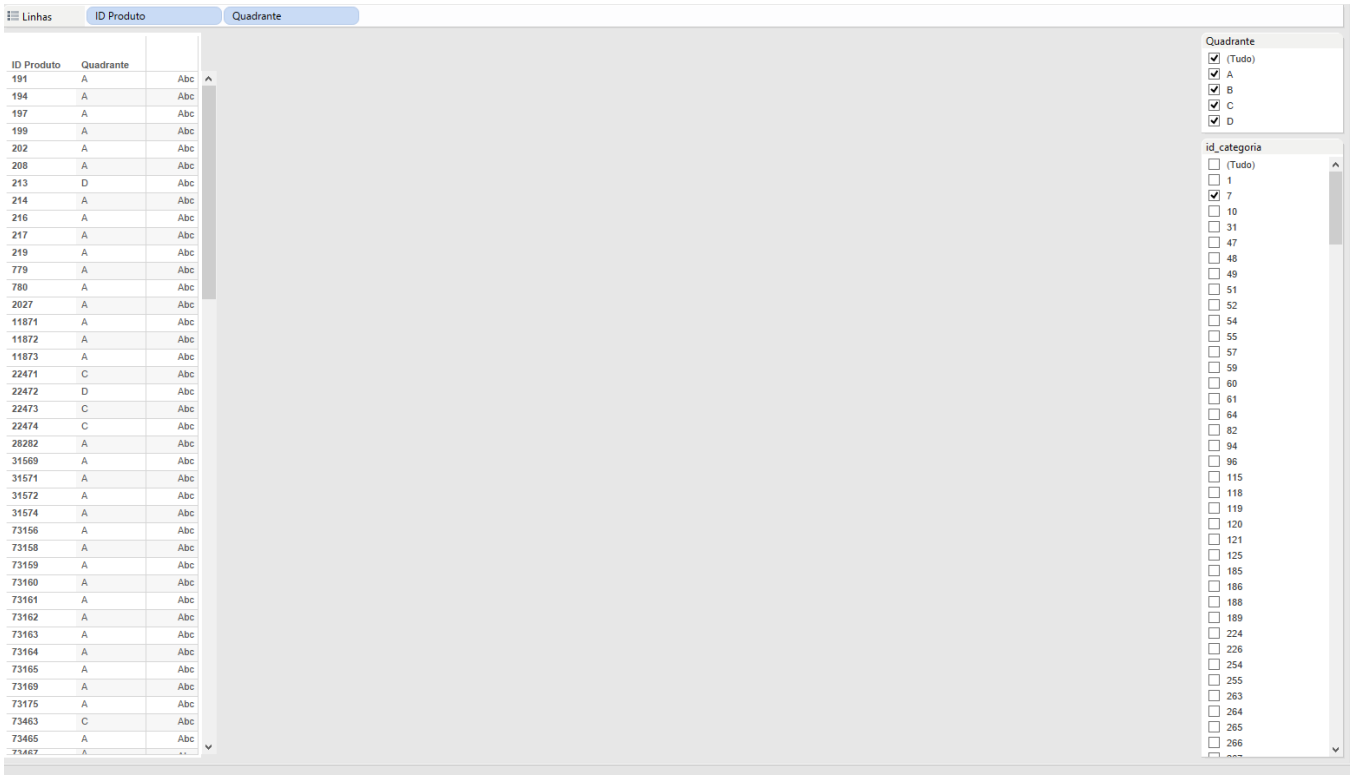

Figura 3. Tabela que exibe os produtos de acordo com seus quadrantes e/ou de acordo com suas categorias, dependendo da aplicação dos filtros

A ferramenta, *Tableau*, permite que todas as exibições estejam disponíveis em tempo real e de forma simplificada para os diretores e gerentes, sendo possível filtrar produtos por quadrantes, categorias, entre outros.

8. Conclusões

Com a estruturação dos processos de obtenção de dados provenientes do *Google Analytics*, e a junção desses dados com outros já comuns relacionados às vendas de uma loja *online*, como a quantidade de pedidos de determinado produto, foi possível atender a demanda dos gerentes e diretores de saber como está a situação da loja a nível dos seus produtos a venda. Tal demanda foi completada por meio da elaboração da matriz de pedidos X visualizações, dividida por quadrantes, aliada a uma visão no *Tableau* que permite saber qual a situação de cada produto de acordo com seu quadrante no período desejado e em tempo real.

A princípio foi feita uma análise apenas a respeito das informações relacionadas a visualizações das páginas dos produtos, porém, futuramente, com os dados de sessões e rejeições já presentes no DW, outros tipos de verificações podem vir a serem feitas sem maiores dificuldades.

Referências

[Canaltech] Canaltech. O que é api?

[Chan 2011] Chan, M. (2011). Google analytics. *Spring*.

[Chen et al. 2012] Chen, H., Chiang, R. H., and Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly*, 36(4):1165–1188.

[Chen et al. 2014] Chen, M., Mao, S., and Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, 19(2):171–209.

[Davenport 2006] Davenport, T. H. (2006). Competing on analytics. *harvard business review*, 84(1):98.

[Favaretto 2007] Favaretto, F. (2007). Melhoria da qualidade da informação no controle da produção: estudo explanatório utilizando data warehouse. *Produção*, 17:343–353.

[Gantz and Reinsel 2011] Gantz, J. and Reinsel, D. (2011). Extracting value from chaos. *IDC iview*, 1142:1–12.

[Google] Google. Como uma sessão é definida no analytics.

[Hey et al. 2009] Hey, A. J., Tansley, S., Tolle, K. M., et al. (2009). *The fourth paradigm: data-intensive scientific discovery*, volume 1. Microsoft research Redmond, WA.

[J. Han 2001] J. Han, M. K. (2001). Data mining. *New York: Morgan Kaufmann Publishers*.

[Kimball 1998] Kimball, R. (1998). *Data Warehouse tool kit: técnicas para a construção de data warehouses dimensionais.* Makron Books.

[Lim et al. 2013] Lim, E.-P., Chen, H., and Chen, G. (2013). Business intelligence and analytics: Research directions. *ACM Transactions on Management Information Systems (TMIS)*, 3(4):17.

[Machado 2000] Machado, F. N. R. (2000). *Projeto de Data Warehouse: uma visão multidimensional*. Editora Érica.

[Oracle] Oracle. Data warehousing concepts.

[W. Inmon 2001] W. Inmon, R. Terderman, C. I. (2001). *Data Warehousing: como transformar informações em oportunidades de negócios.* Editora Berkely.

[ZIULKOSKI 2003] ZIULKOSKI, L. C. C. (2003). Coleta de requisitos e modelagem de dados para data warehouse: um estudo de caso utilizando técnicas de aquisição de conhecimento. *Bacharelado em Ciência da Computação pelo Instituto de Informática da Universidade Federal do Rio Grande do Sul*.