ENEPEX
ENCONTRO DE ENSINO, PESQUISA E EXTENSÃO
8° ENEPE UFGD • 5° EPEX UEMS

BUSINESS INTELLIGENCE COM DADOS EXTRAÍDOS DO FACEBOOK UTILIZANDO A SUÍTE PENTAHO

Francy H. Silva de Almeida[1]; Maycon Henrique Trindade[2]; Everton Castelão Tetila[3]

UFGD/FACET – Caixa Postal 364, 79.804-970 – Dourados – MS, E-mail: helder.dds@gmail.com
[1]Bacharel em Sistemas de Informação pela UFGD. [2]Bacharelando em Sistemas de Informação da UFGD. [3]Orientador, Professor FACET.

RESUMO

A suíte *Pentaho* é uma plataforma *Open Source* utilizada na criação de soluções de *Business Intelligence* (Inteligência Aplicada aos Negócios). Os recursos da *Pentaho* permitem a análise de um grande volume de dados, como é o caso dos dados registrados pelos usuários do Facebook – atualmente, a maior rede social da Internet. Dentre as ferramentas disponibilizadas pela *Pentaho* está o *Schema Workbench*, que facilita o trabalho de criação do cubo OLAP através de uma interface visual. Para realizar a extração de dados é necessário a criação de um aplicativo dentro do Facebook que servirá de porta de entrada para a plataforma de desenvolvimento. Utilizando os recursos da API disponibilizada pelo Facebook, é possível estabelecer uma conexão com essa aplicação e realizar requisições de acesso aos dados públicos dos usuários. A conexão é feita através do protocolo HTTP, sendo possível executar consultas que retornem arquivos do tipo JSON, formato utilizado pelo serviço RESTful do Facebook. Após a coleta dos dados é possível organizá-los para a utilização no OLAP, que possibilita a análise e a identificação de comportamentos e características relacionadas aos usuários do Facebook.

Palavras-chave: *On-line Analytical Processing, Data Warehouse, Web mining.*

INTRODUÇÃO

O Facebook, a rede social mais utilizada atualmente, vem se expandindo dia após dia, conquistando usuários por todo o mundo. De acordo com a agência Reuters, só no Brasil são 76 milhões de usuários — terceiro no ranking — perdendo apenas para a Índia, com 82

milhões, e para os Estados Unidos, com 179 milhões[1]. Pessoas diariamente acessam a rede para atualizar seus perfis, postar fotos, trocar informações, divulgar suas preferências, promovem marcas, e muito mais. O que poucos percebem é que o Facebook vai além do simples conceito de rede social. O Facebook se tornou um ambiente muito atrativo para os negócios, pois empresas podem se beneficiar das informações compartilhadas na rede a fim de viabilizar estratégias comerciais. As informações dos usuários podem ser acessadas por softwares que interagem com a rede social através da API (*Application Programming Interface*) oferecida pelo próprio Facebook, possibilitando assim, uma infinidade de aplicações. Unindo essa tecnologia com os processos já conhecidos de Processamento Analítico On-Line (OLAP) é possível manipular e analisar os dados sob múltiplas perspectivas, e então, extrair conhecimento desse poderoso depósito de dados, abastecido diariamente por mais de 1 bilhão de usuários.

MATERIAIS E MÉTODOS

O desenvolvimento de um ambiente *Business Intelligence* é iniciado pela construção do *Data Warehouse* (DW), o qual pode ser definido como "uma coleção de dados orientada a assunto, integrada, não volátil, variável no tempo e de apoio às decisões da gerência" (INMON et al., 2008). O DW é resultado do processo chamado *Extract/Transform/Load*(ETL), tecnologia que permite retirar dados de ambientes externos e transformá-los em dados corporativos. Para iniciar esse processo, primeiro é necessário selecionar os atributos (campos) que irão compor o cubo OLAP. Para a aplicação proposta, foram utilizados dados simples para detectar o comportamento de usuários na rede ao longo do tempo, usando como métrica a quantidade de *posts* publicados. Os atributos e os tipos utilizados são descritos na Tabela 1.

Tabela 1. Dados dos usuários utilizados no cubo OLAP.

Atributo	Tipo
Nome	String
Idade	Integer
Sexo	String
Posts	String
Data_post	Date

1 http://br.reuters.com/article/businessNews/idBRSPE97D03020130814

Nesse trabalho, foi implementado uma rotina de computador, intitulada Extrator, capaz de realizar todas as etapas do processo ETL. Esse código realiza a extração dos dados utilizando o pacote Facebook SDK (*Software Development Kit*), o qual "permite interagir com a plataforma Facebook e viabiliza o acesso aos dados" (RUSSELL, 2013). Executando essa primeira etapa, é possível obter os resultados em formato JSON, os quais devem ser transformados - de acordo com o processo de ETL - para, depois, realizar a carga no *Data Warehouse*. As Figuras 1 e 2 apresentam, respectivamente, o Modelo Multidimensional que representa o projeto lógico do *Data Warehouse* da aplicação proposta e parte do arquivo JSON extraído.

Figura 1. Modelo Multidimensional *Star Schema*.


```
"data": [
  {
    "gender": "female",
    "name": "Patricia [redacted]",
    "birthday": "05/10",
    "id": "530706343"
  },
  {
    "gender": "female",
    "name": "Ruth [redacted]",
    "birthday": "06/08/1989",
    "id": "547474115",
    "posts": {
      "data": [
        {
          "created_time": "2013-12-26T14:01:53+0000",
          "id": "547474115_10152102974039116"
        }
      ],
      "paging": {
        "previous": "https://graph.facebook.com/547474115/posts?limit=40&fields=created_time&since=1388066513",
        "next": "https://graph.facebook.com/547474115/posts?limit=40&fields=created_time&until=1388066512"
      }
    }
```


Figura 2. Arquivo JSON contendo os dados dos usuários.

O Extrator permite tratar o arquivo JSON e organizar os dados no modelo *Star Schema*. Este modelo consiste em uma tabela central de fatos relacionada com tabelas de dimensões, formando assim, o esquema multidimensional do armazém de dados, ou *Data Warehouse* (TURBAN et al., 2010). O procedimento realizado pelo Extrator, é responsável por realizar o cruzamento dos dados armazenados nas duas tabelas de dimensão (dimensão_usuário e dimensão_tempo) e inseri-los na tabela fato. Esse processo de

cruzamento e carregamento de dados, caso realizado manualmente, despenderia muito tempo e seria considerado inviável de acordo com o tamanho da base de dados.

Neste trabalho, foi analisado a frequência com que os usuários utilizam o Facebook, no intuito de identificar algum padrão que responda à perguntas como, “Quais dias ou meses os usuários do Facebook são mais interativos?” ou “Quais usuários são mais interativos de modo geral?”. O modelo da Figura 1 viabiliza essas respostas.

Após utilizar o Extrator para gerar e carregar a base de dados, podemos iniciar a comunicação entre a plataforma *Pentaho* e o *Data Warehouse.* A comunicação é realizada através do arquivo XML que pode ser criado com a ferramenta *Workbench* da *Pentaho*, como mostra a Figura 3.

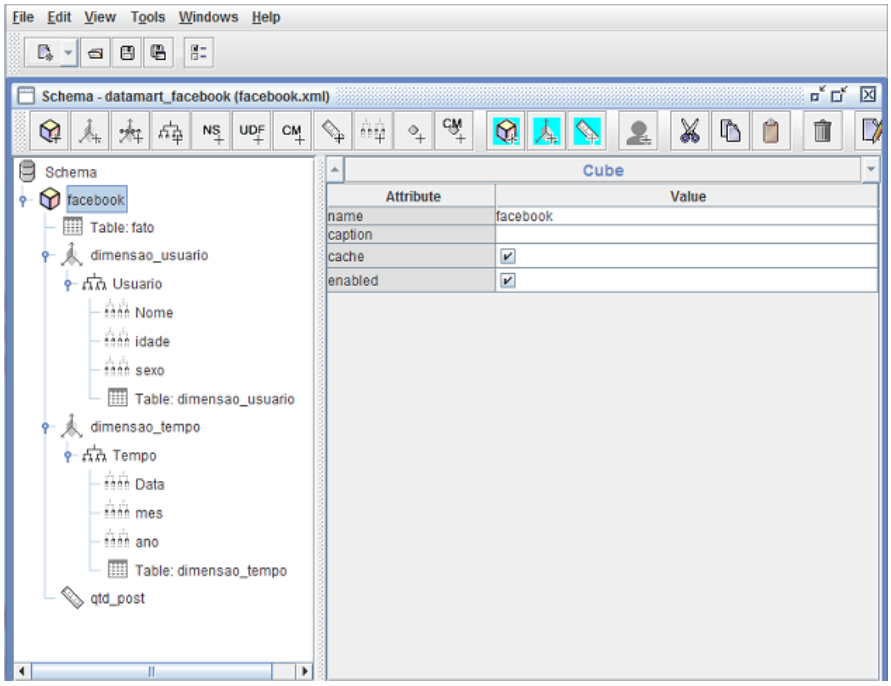

Figura 3. Interface Gráfica da ferramenta *Schema Workbench.*

RESULTADOS E DISCUSSÃO

Esta seção apresenta as visões analíticas executadas na plataforma *Pentaho Analisys Services*. A Figura 4 ilustra o ambiente visual da plataforma, com as dimensões e métrica definidas no modelo multidimensional.

Para facilitar a compreensão, foi utilizado os dados extraídos de apenas 100 usuários interativos. Suponha o seguinte caso: deseja-se identificar qual usuário possui o maior número de *posts* publicados. A Figura 5 apresenta o gráfico resultante dessa pesquisa. Observe que o

usuário 34 é o mais ativo, de acordo com a quantidade de *posts* publicados. É possível ainda identificar qual data possui mais *posts* publicados. Essa informação pode ser obtida na Figura 6, onde é visível o aumento de *posts* no final do ano. Finalmente, o gráfico da figura 7 apresenta qual usuário possui a maior quantidade de *posts* em relação ao tempo.

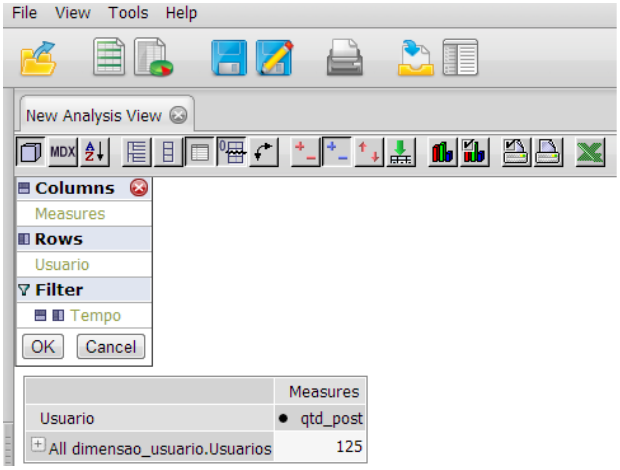

Figura 4. Ambiente visual da *Pentaho Analisys Services*.

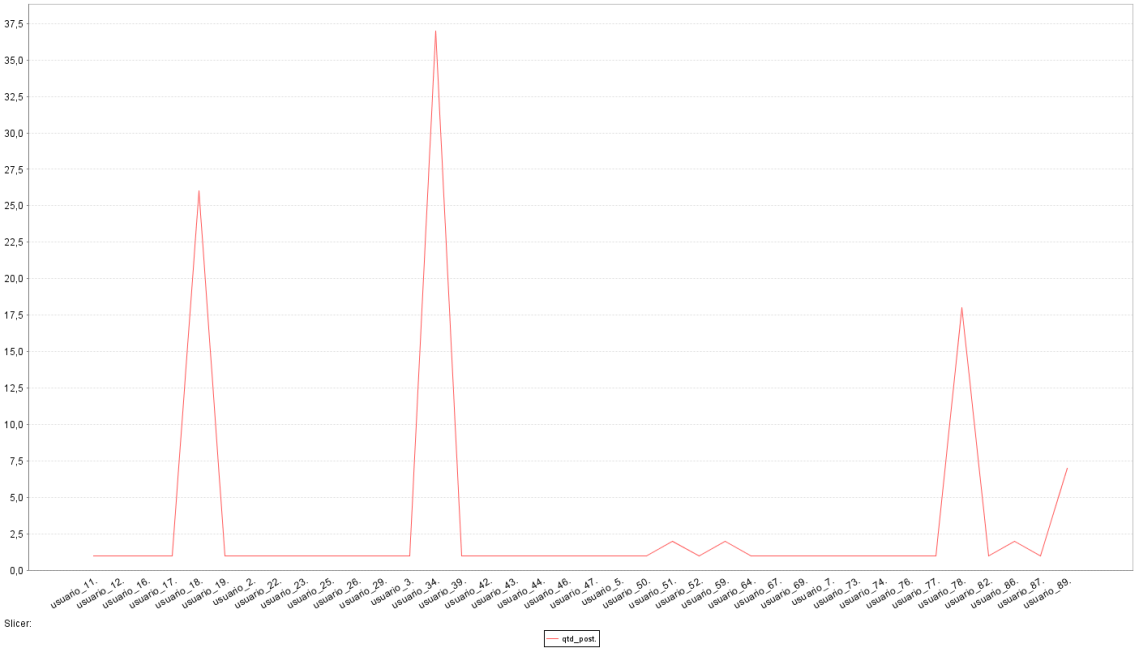

Figura 5. Relatório gráfico mostrando o usuário mais ativo.

Por meio deste estudo é possível observar que os gráficos gerados não apresentam informação satisfatória para datas distantes. Segundo o artigo de Marcelo Brito, publicado no site DevMedia, isso se deve ao fato da plataforma Facebook armazenar apenas os dados recentes em memória e por um pequeno período de tempo (BRITO, 2014). Portanto, em aplicações reais o processo de ETL deve ser feito periodicamente. Mais uma vez, podemos

perceber a importância de automatizar o processo de ETL para reduzir o tempo despendido, utilizando códigos personalizados como o Extrator desenvolvido neste trabalho.

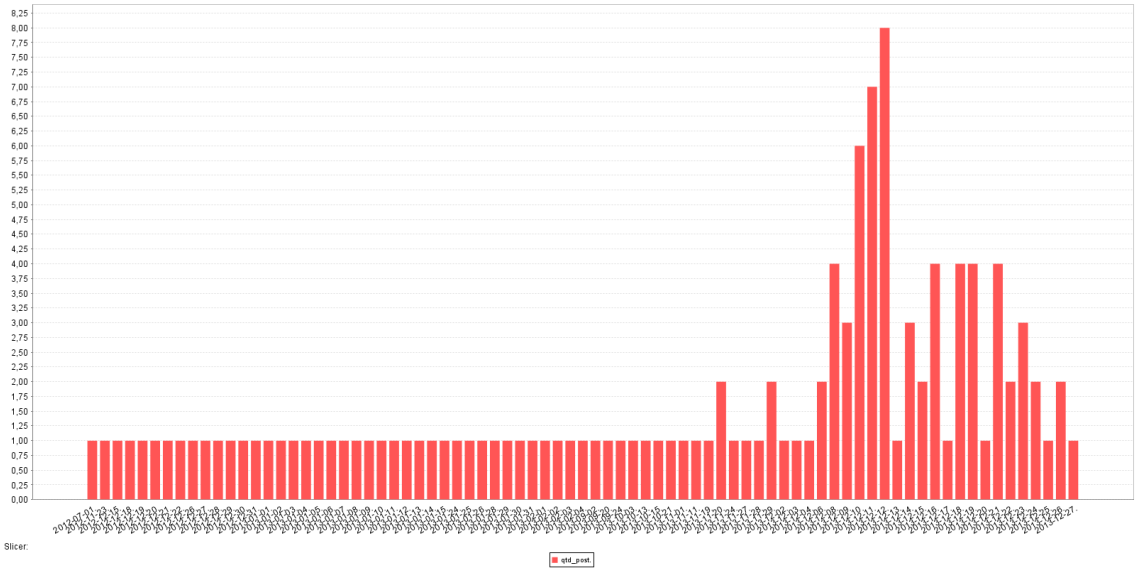

Figura 6. Quantidade de *posts* em relação ao tempo.

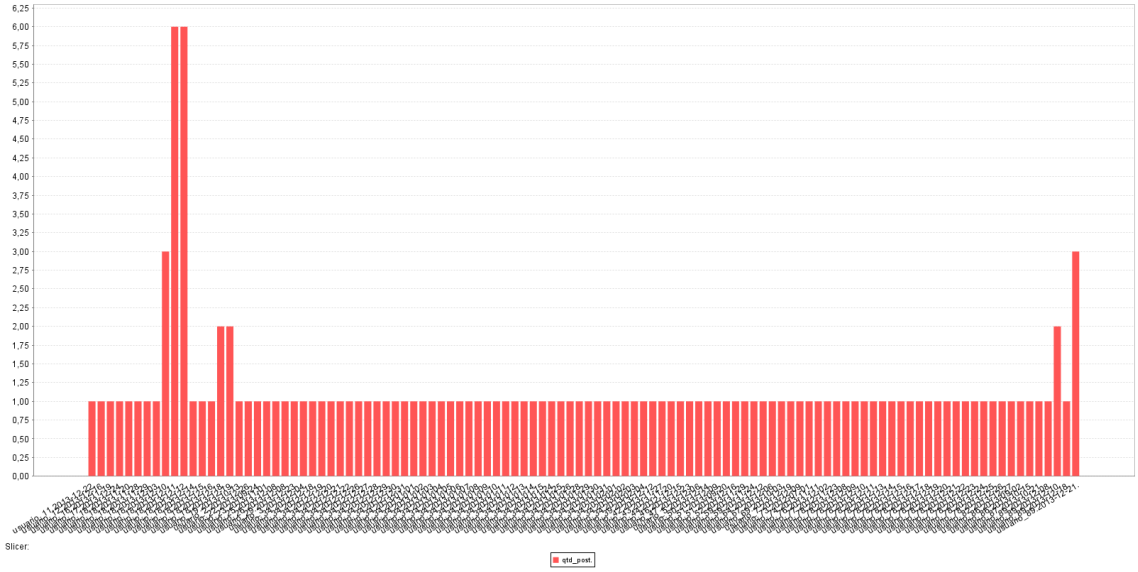

Figura 7. Quantidade de publicação de *posts* por Usuário e Tempo.

CONCLUSÃO

Este artigo apresentou um exemplo simples da possibilidade de aplicar técnicas de *Business Intelligence* em dados gerados pelas redes sociais, em especial, a plataforma Facebook. Com base neste estudo, muitas soluções podem ser desenvolvidas a fim de explorar ao máximo o potencial das técnicas apresentadas. Várias respostas podem ser obtidas utilizando essa abordagem, dependendo do tipo de conhecimento que se deseja adquirir, sendo este trabalho uma base para o desenvolvimento de novas aplicações.

REFERÊNCIAS

BRITO, M. Extração de dados do Facebook com a suíte Pentaho. Disponível em: www.devmedia.com.br/extracao-de-dados-do-facebook-com-a-suite-pentaho/25523. Acesso em: 10 de ago. 2014.

EQUIPE FACEBOOK DE DESENVOLVEMENTO. The Graph API. Disponível em: developers.facebook.com/docs/graph-api. Acesso em: 22 de ago. 2014.

INMON, W. at al. *DW 2.0: The Architecture for the Next Generation of Data Warehousing.* Burlington, USA: Elsevier, 2008. 371p.

RUSSELL, M. A. *Mining the Social Web: data mining facebook, twitter, linkedin, google+, githut, and more.* 2ª Edição. Sebastopol, CA: O'Reilly Media, 2013. 421p.

TURBAN, E. et al. *Business Intelligence*: *a managerial approach.* 2ª Edição. New Jersey: Prentice Hall, 2010. 312p.