Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Mariana Gonçalves de Oliveira

Modelagens de um Data Lake: De dados brutos a área de negócio

Curitiba
2020

Mariana Gonçalves de Oliveira

Modelagens de um Data Lake: De dados brutos a área de negócio

Monografia apresentada ao Programa de Especialização em Data Science e Big Data da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Eduardo Cunha de Almeida

Curitiba
2020

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Modelagens de um Data Lake: De dados brutos a área de negócio

Mariana Gonçalves de Oliveira[1]
Eduardo Cunha de Almeida[2]

Resumo

A necessidade de maior velocidade na captação de informações bem como a integridade das informações geradas nas organizações são fatores que crescem de maneira latente dentro das empresas. Os obstáculos vêm desde como e quando captar os dados, passando pela melhor forma de armazená-los, e como conseguir modelar essas informações para passar mensagens claras e assertivas que auxiliem as áreas de Negócio em seu processo decisório. Neste trabalho serão exploradas de maneira detalhada cada uma dessas etapas, partindo de um Data Lake estruturado com dados brutos da área de negócio, que se utilizou de um processo de ELT (*Extract, Load and Transform*) para a população inicial dos dados. O resultado desse artigo deve servir de suporte para o entendimento das etapas e desafios existentes ao se integrar grandes volumes de dados que são trazidos de maneira bruta de sistemas robustos OLTP (*On Line Transaction Processing*) e armazenados um *data lake*, comumente utilizados nas empresas de médio ou grande porte.
Palavras-chave: *d*ata lake, big data, ELT, business

Abstract

The need for higher speeds and integrity in data generation on business orgaizations is something that grows exponencially within the companies. The obstacles to do so comes in different forms such as how and when to capture data, the best way to store it and how to model it so it can reflect and comunicate a correct clear message in order to help different business areas in the process of decision making. In this article it will be explored in a detailed matter each and every step of modeling data from a structured Data Lake with raw data from the business areas utilizing an ELT (Extract, Load, Transform) process to begin the process of modeling data. The results of this article should be of support for the understandig of each step of the way and the challenges within the process to connect large volumes of data that are captured in a raw state from robust OLTP ((On Line Transaction Processing) systems that are commonly used by medium/large size companies and stored into a Data Lake.
***Keywrods**: data lake, big data, ELT, business*

[1]Mariana Gonçalves de Oliveira, mgoliveira1990@gmail.com.
[2]Eduardo Cunha de Almeida-DINF/UFPR, eduardo@inf.ufpr.br.

1 Introdução

A crescente demanda por informações nas organizações faz com que as etapas de captação, tratamento e modelagem dos dados esteja cada vez mais flexível e acompanhe as flutuações das demandas da gestão e do negócio como um todo. Uma vez que esses dados são processados e armazenados, um dos desafios que se apresenta é como traduzir dados brutos em informações que cheguem ao usuário final de maneira inteligível e de fácil acesso.

Com esse fluxo bem desenhado é possível que empresas passem cada vez mais a tomar decisões baseadas em dados. Com isso uma empresa não só tem o potencial de maximizar seus resultados como também de analisar de maneira mais acelerada o efeito de seus planos de ação e identificar fatores de oportunidade.

Nesse trabalho os dados da amostra explorada são provenientes de um *Data Lake* que foi implementado em uma empresa de grande porte de varejo. Dentre os principais benefícios dessa estrutura estão: (a) melhorar o processamento do elevado volume de dados de diferentes fontes; (b) flexibilidade no desenvolvimento e geração de novas análises e informações (c) possibilidade de cruzamento de dados entre diferentes áreas de maneira rápida e robusta.

No entanto a escolha do *Data Lake* em detrimento de um modelo de *Data Warehouse* clássico traz algumas dificuldades nas rotinas de manutenção. Para manter um fluxo como esse atualizado, é necessário a revisão frequente dos parâmetros utilizados nos sistemas e modelos de dados uma vez que a dinâmica de negócio costuma mudar com frequência. Para contornar esses desafios a compreensão de alguns conceitos que serão apresentados a seguir pode auxiliar.

2 Conceito de um Data Lake

O *Data Lake* pode ser definido como armazenamento centralizado, consolidado e persistente de dados brutos, não modelados e não transformados de múltiplas fontes, sem um esquema pré-definido explícito e sem metadados definidos externamente [1].

Os dados de um DL podem ser estruturados como por exemplo dados de uma tabela em um modelo relacional, como também dados não estruturados como imagens,

áudio, arquivos de texto e vídeos.

Porém, mesmo dentro de um universo de dados estruturados em que comumente se atrela o conceito de *Data Warehouse*, a aceitação de um *Data Lake* por dados de diversas fontes, estruturas e modelagens faz com que ele atenda de maneira mais eficiente modelos de negócio maiores principalmente em empresas de grande porte que possuem alta complexidade e variedade em seus ecossistemas de softwares [2].

Para melhorar a organização e gestão dos acessos do *Data Lake* contemplado nesse artigo, foram criadas 4 camadas de dados:

- **Camada 1** - camada de dados crus. Conforme captados nos diferentes sistemas em que são lidos, sem receber nenhum tipo de transformação, é uma camada restrita a desenvolvedores e engenheiros de dados;
- **Camada 2** - camada de negócios. Nessa etapa as regras e restrições de negócio são refletidas. Nessa camada já é possível a criação de *Datamarts* para alimentar *Dashboard* ou análises recorrentes dos usuários;
- **Camada 3** - camada de análise de dados. Nessa camada se inicia a configuração de tabelas mais complexas denominadas "DNA" para construção de modelos preditivos ou prescritivos;
- **Camada 4** - camada usuário. Essa camada seria uma espécie de "sandbox" onde cada usuário tem flexibilidade de montar suas análises mais pontuais ou experimentar novos cruzamentos.

3 Processo ETL

No modelo clássico de inteligência de negócio (*Business Intelligence* ou BI) os dados são extraídas da fonte em seu estado bruto, e anteriormente à carga no *Data Warehouse* tais dados passam uma série de etapas de tratamento, limpeza e aplicação das regras de negócio previamente alinhadas por uma figura geralmente representada por um especialista de negócios no momento da especificação do modelo. O processo de ETL geralmente será exclusivo a uma área de negócio específica, ou a um grupo de indicadores. De acordo com (Kimbal, Ralph 2013) o processo é construído por quatro principais etapas: (1) Selecionar o processo de negócio (2) Definir a granularidade (3) identificar as dimensões (4) identificar os fatos.

4 Processo ELT

[4] Além do crescimento da demanda nas empresas por Big Data um dos motivos que favorecem o modelo ELT também foi a redução de custos para armazenamento dos dados. Ambos fatores contribuíram para que empresas começassem a inverter as etapas do processo clássico

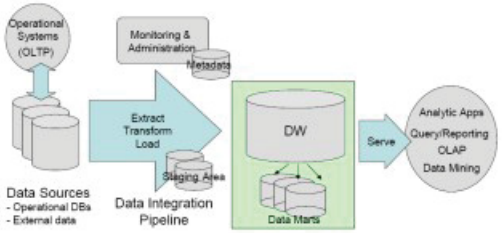


Figure 1. Traditional business intelligence architecture

Figura 1: Ilustração das etapas do Processo ETL iniciando na coleta de dados de um sistema legado ou fonte externa, seguida de cargas de pipeline de dados estruturados com auxílio de áreas lógicas de staging e por fim com a transformação para textitDatamarts que especificam e estruturam o dado para determinada área [3]

de ETL, além deste se tornar defasado devido as restrições que apresentava para soluções de maior flexibilidade nas análises. O processo ELT consiste em extrair o dado bruto e logo em seguido armazená-lo sem que passe por nenhum transformação para na última etapa, já em uma camada diferente do Data Lake, aplicar regras de transformação. Além da flexibilidade em adicionar novas fontes, outras vantagens desse processo é a aceleração do tempo de implementação de uma nova visão, e diferentes possibilidade de agregação sobre os dados brutos.

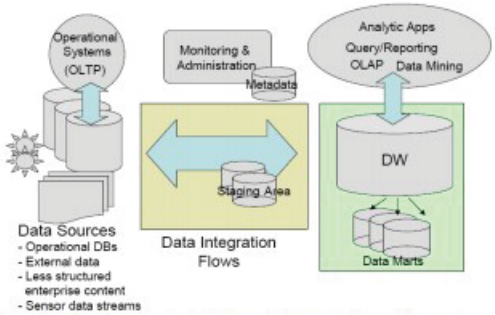


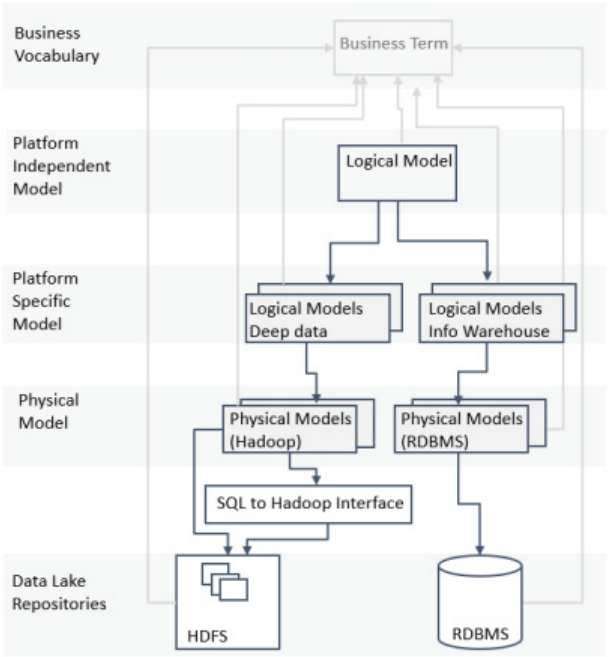
Figure 2. Next generation business intelligence architecture

Figura 2: Ilustração das etapas do Processo ELT, que em distinção ao modelo ETL, recebe dados crus, estruturados ou não, de diferentes fontes e possui fluxo contínuo de retroalimentação entre as mesmas com a criação de Data Marts de dados mais flexívies [3]

Apesar dos ganhos mencionados, um dos principais desafios do modelo ELT é lidar com os logs de erros provenientes de bases cruas e diferentes fontes de dados que não são tratadas, além de traduzir posteriormente um ecossistema extenso de dados para uma necessidade de negócio. Uma solução bastante explorada no mercado para mitigar esses erros é a construção de um modelo conceitual na etapa de "Transformação".

5 Modelo de Dados Conceitual

De acordo [5] o propósito do ciclo de criação de um modelo conceitual para um Data Lake consiste numa série de etapas prévias que irá fornecer estruturas consistentes no momento da implementação. Tais etapas deverão refletir a descrição da realidade para o Negócio e deverá servir de insumo independente da escolha de Modelo Físico podendo ser Relacional ou NoSQL conforme descrito no figura abaixo.



[5]

Figura 3: Descrição IBM das principais etapas e associações consideradas em um processo de Modelagem dos Dados em um Data Lake

Para construção do modelo conceitual existem diferentes opções, entre as mais conhecidas estão o Modelo Entidade Relacionamento ou UML. Para este artigo foi selecionado o modelo Entidade Relacionamento. Em ambos os casos,[6] a construção do modelo fornece os meios de identificação das entidades que serão representadas no *Data Lake* e além disso, como elas se relacionam entre si. Nessa etapa é fundamental o envolvimento do desenvolvedor com a área de Negócios e dada a complexidade de fontes e bases contidas na camada de dados crus de um Data Lake, frequentemente será necessária a consulta de mais usuários para garantir que todas as restrições e regras de negócio sejam incorporadas ao modelo. Nesse momento é importante se atentar para problemas de desenhos dessas bases como redundâncias e dados incompletos. É importante reforçar como mencionado na introdução, a importância da constante manutenção das regras desenhadas inicialmente o que será possível com a comunicação contínua entre área de negócio e desenvolvedores.

Uma vez criado o modelo conceitual o desenvolvedor está habilitado para transformar essas regras em um modelo físico dentro do *Data Lake* na camada de negócios. Com a estruturação dessa nova camada, habilita-se diferentes fins ao uso desses dados, como *Dashboards* de Negócio ou Modelos Preditivos ou Prescritivos.

6 Experimento

Para esse experimento foi coletada uma amostra de um *Data Lake* IBM DB2 com suporte de tecnologia IIAS referente a dados de fornecedores da cadeia de suprimentos de uma empresa de grande porte.

Como mencionado na introdução a estrutura do *Data Lake* confere uma melhora significativa na capacidade de processamento dos dados, permitindo por exemplo a leitura de tabelas de 1 bilhão de registros em poucos segundos. Além disso como falado também, nessa estrutura foi possível cruzar dados para análises pontuais de diferentes áreas e modificá-los de maneira constante e flexível a medida que era necessário para atender novos requisitos da área de negócio.

Para a amostra analisada foram trazidas 8 tabelas, 3 de fato e 5 de dimensões, para caracterizar a atividade de compras de insumos da empresa.

As amostras foram inseridas na máquida via MySQL 8.0 e o processo de transformação dos dados foi feito no Pentaho Versão 8.2. A escolha do Pentaho Data Integration para a etapa de transformação ocorreu por uma série de requisitos que a ferramenta entrega [7] sendo os principais:

1. Suporte da plataforma e fóruns de usuários
2. Integração com outras ferramentas e serviços
3. Ferramenta open source
4. Diferentes modalidades além da Integração e Transformação de Dados, como o Pentaho Dashboard Designer

Para transformação dos dados de maneira a refletir a necessidade da área de Negócios, foram feitas 4 transformações ao todo, 5 delas sendo as dimensões, e para tabela fato optou-se por uma consolidação dos 3 diferentes fatos provenientes da camada de dados crus.

No processo de criação de tabelas fato em um modelo, muitas vezes é conveniente fazer a consolidação de diferentes processos de negócio em apenas uma tabela fato caso esses processos possuam a mesma granularidade. Tabelas fato consolidadas podem adicionar mais carga para o processo de ELT porém trará ganho na carga das etapas analíticas dos processos. Esses casos deveriam ser considerados para métricas de negócio que são constantemente analisadas no mesmo contexto. [8] A construção do modelo de dados da amostra, foi baseada nesse formato.

No processo de transformação das tabelas um dos principais desafios é o de dicionário de dados para os campos da camada de dados crus. Pelo motivo dos dados serem captados de maneira bruta no estado em que se encontra na origem, muitos deles trazem denominações incompreensíveis. Nessa etapa foi precisa estar

muito próximo a área de negócio e Arquitetura para entender o significado de cada campo e aplicá-lo na transformação do Pentaho.

Dados Dimensionais da Transformação	
Tempo de Processamento	44 seg
Total de Registros Amostra	278 583 registros
Nº Etapas Transformação	42 nós

Figura 4: Dimensões da Transformação feitas no Pentaho Data Integration

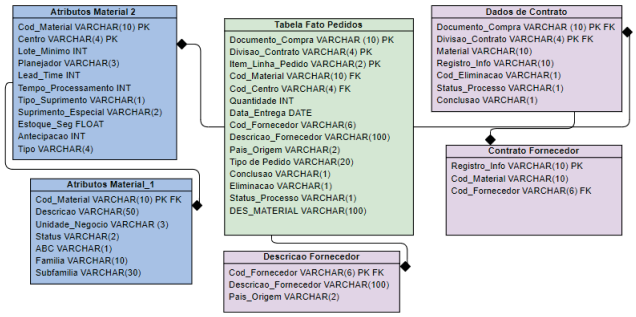


Figura 5: Modelo Conceitual mapeado no Visual Paradigm referente ao processo de planejamento de compras da empresa analisada. O processo descreve a previsão de compras e pedidos firmados para horizonte de 18 meses com seus fornecedores além da caracterização dos produtos planejados e classificação de fornecedores em diferentes grupos

7 Habilitando os Dados para Modelos Descritivos, Preditivos e Preditivos

As últimas etapas do processo de Tratamento de Dados consiste em otimizar e publicar os dados. A otimização pode se aplicar principalmente para a preparação de dados para modelos prescritivos, e pode ocorrer em forma de redução de dados, clusterização ou amostragem. [9]Já para modelos descritivos a parte de publicação e reportes se torna mais relevante principalmente via Dashboards.

Através desse artigo os autores apresentam um estudo de caso que traz relevância a essa etapa focada em entender o público alvo desses dados, e seu nível de expertise para manipulação dos mesmos. Para auxiliar nessa compreensão, e no mapeamento das diferentes estratégias de abordagem eu usuário final, o autor cria uma estrutura de "personas" baseado na função de cada uma, no objetivo que terá ao receber tais dados e com a expertise possui em diferentes ferramentas. Esse princípio foi aplicado também em nosso experimento, auxiliando para que o output dessa solução tenha o maior nível de alcance possível.

Dentro do experimento foram classificados 3 tipos de usuários distintos: Cientista de Dados, Analista Tático, Analista de Negócios e Gestor que serão descritos abaixo.

Cientista de Dados

Escopo: Atua com criação de modelos preditivos e de otimização.Possui conhecimento de programação o suficiente para acessar diretamente o Data Lake e consumirá dados previamente estruturados e agendados. Seu software de acesso para consumo dos dados é o próprio Data Lake.

Principais benefícios no uso do DL: Eliminação de etapas de limpeza e tratamento de dados que podem se tornar bastante custosas no dia a dia de um Cientista de Dados. Maior confiabilidade nos outputs uma vez que não sofrem alteração manual no processo de input.

Analista Tático

Escopo: Atua como consolidador de informações de negócios, gera KPIs de áreas inteiras e faz gestão e crítica de massas grandes de dados. Normalmente não tem conhecimento de programação para acessar diretamente o Data Lake porém possui conhecimento avançado em ferramentas como Excel. Seu acesso será via Excel ou BI com dados na menor granularidade possível para que de flexibilidade as suas análises e publicações.

Principais benefícios no uso do DL: Ganho de tempo, escala e granularidade nos dados extraídos por esse analista, dessa forma é possível o foco em atividades como desenvolver novos reportes, tornar mais visual e dar relevância aos resultados que será gerado.

Analista de Negócios

Escopo: O Analista de Negócios apesar de ser da mesma área que o analista Tático, tem como principal função a gestão de um portfólio específico, que pode ser das mais diversas áreas: Comercial é responsável pelo resultado de contas de clientes, Supply Chain é responsável pelo planejamento de produção nas fábricas e gestão de pedidos dos fornecedores, Marketing é responsável pelo market share dos produtos da empresa. Esse perfil não possui tantas habilidades em gestão de grandes volumes de dados pois se encontra mais na execução. Seu software de acesso é o BI.

Principais benefícios no uso do DL: Analista mais disponível para análises de maior profundidade e priorização dos itens por diferentes critérios de criticidade. Mais facilidade na identificação de causas raiz e reincidência de problemas.

Gestor

Escopo: Por fim, o Gestor de cada área é o responsável por tomar as decisões de negócio para atingimento dos resultados e gestão das pessoas. Como seu perfil é mais voltado pra tomada de decisão mais assertiva e rápida possível, seu software de acesso também deve ser um BI mais customizado e resumido.

Principais benefícios no uso do DL: Com a maior confiabilidade dos dados, e visualizações claras desenvolvidas pelo analista tático, o gestor consegue tomar decisões mais assertivas e baseadas em dados, além de conseguir acompanhar de maneira mais efetiva os resultados a medida que ocorrem para calibrar planos de ação.

Conclusão

Com a automatização do processo de transformação dos dados brutos de fornecedores houve um ganho médio para a área de 2 dias úteis de um supervisor, pela eliminação da atividade de cruzamento de dados, e isso pode ser listado como um ganho tangível. No entanto com esse processo houveram outros ganhos indiretos como habilitação de novas aberturas e insights para área de Negócio devido ao foco que foi dado na análise dos dados e não na construção deles. Outro ganho indireto foi a clusterização dos dados que habilitará que no futuro tenhamos dados históricos suficientes para criação de modelos preditivos ou prescritivos. Por fim com a facilidade da criação dos reportes foram criados novos fóruns de gestão para envolvimento de demais áreas e tomadas de decisão mais conjuntas.

Podemos concluir que o processo de transformação dos dados de um Data Lake tem algumas complexidades como dicionário de dados, formatos de dados na fonte, porém com ajuda de ferramentas de Transformação e um Modelo Conceitual que reflita bem a realidade é possível trazer ganhos significativos as empresas.

Referências

[1] D. Matos, *Data Lake – A Evolução do Armazenamento e Processamento de Dados*, Disponível em: Jun/2019 <https://www.cienciaedados.com/data-lake-a-evolucao-do-armazenamento-e-processamento-de-dados/>

[2] H. Bian *Pixels: Multiversion Wide Table Store for Data Lakes*, (2020), <http://cidrdb.org/cidr2020/gongshow2020/gongshow/abstracts/cidr2020-abstract74.pdf>

[3] D. Umeshwar e M. Castellanos e A. Simitsis e K. Wilkinson, *Data Integration Flows for Business Intelligence* , (2009), <https://doi.org/10.1145/1516360.1516362>

[4] P. Ortega, V. Dmitriyev, M. Abilov, J. Gomez, *ELTA: New Approach in Designing Business Intelligence Solutions in Era of Big Data*, (2014), <doi: 10.1016/j.protcy.2014.10.015>

[5] IBM, *IBM Industry Model Support for a data lake architeture*, (Somers, NY, 2016), Version 1.0.

[6] A. Silberschatz, H. Korth, S. Sudarshan *Database System Concepts, Sixth Edition* (2011),pag. 270.

[7] D.R. Prigol, A.T. Lazaretti, R.M. Bertei *Avaliação da ferramenta Pentaho Comunity Data Integration Através de Estudos de Caso*, (2016), <https://painel.passofundo.ifsul.edu.br/uploads/arq/20161220160913175120358 4.pdf>

[8] R. Kimbal, M. Ross *The Data Warehouse ToolKit 3rd Edition* (2013),pag. 81.

[9] J.M. Hellerstein, J.Heer, S. Kandel *SelfService Data Preparation: Research to Practice*, (2018), <http://sites.computer.org/debull/A18june/p23.pdf>