FGV EMAp

João Pedro Jerônimo

Modelagem Informacional

Revisão para A2

Rio de Janeiro

2025

Conteúdo

Big Data

1.1 Introdução

Na primeira parte do curso, nós vimos uma expansão dos conceitos de Banco de Dados, como eles eram expandidos dependendo da sua finalidade. Vimos a extensão de bancos operacionais para bancos analíticos (Data Warehouses) e sua estruturação e aqui não será diferente! Porém, antes de entender como estruturar novos dados, temos que entender os tipos de dados que vamos armazenar nessa parte do curso

Quando vamos trabalhar com conjuntos de dados, podemos categorizar eles em 3 tipos

- Bancos operacionais
- Data Warehouses
- Big Data

E é exatamente essa terceira que iremos abordar. Big Data são os conjuntos de dados em corporações que tem grande volume e diversificação, além de rápido crescimento. Não são modelados formalmente para consulta e recuperação e não são acompanhados de metadados detalhados

Ou seja, são aqueles tipos de dados que não são bem estruturados. Podemos fazer uma subdivisão também nessa classificação:

- **Não-estruturados**: Não tem metadados detalhados, exemplo: Documentos de Texto
- **Semi-estruturados**: Possuem alguns metadados, mas não o suficiente para descrever completamente o dado num todo, exemplo: Mensagens de texto (Você tem estruturas de destinatário, remetente, horário, etc., mas o texto da mensagem em si não tem uma estruturação)

A gente pode tentar descrever Big Data com o que chamamos de V's. Porém, essa descrição não é algo 100%, já que esse conteúdo e o que é ou não um dado de Big Data vai bastante do contexto e da interpretação

- **Volume**
- **Variedade** (Fontes)
- **Velocidade** (Entrada dos dados)
- **Veracidade** (Qualidade)
- **Variabilidade** (Interpretação)
- **Value**
- **Visualization** (Elaborada e complexa)

Definição 1.1.1 (Big Data): Grandes volumes de conjuntos de dados diversificados e de crescimento rápido que, quando comparados com bancos operacionais ou data warehouses:

- Consideravelmente menos estruturados (Poucos ou nenhum metadado)
- No geral: $+$Volume, $+$Velocidade, $+$Variabilidade
- Mais problemas na qualidade dos dados (Veracidade)
- Maiores as gamas de interpretações (Variabilidade)
- Abordagem mais exploratória e experimental para gerar Valor
- Se beneficia mais com visualizações elaboradas e inovadoras

1.2 MapReduce e Hadoop

Certo, mas se Big Data são dados não-estruturados, o que podemos fazer para lidar com eles? É uma selva sem lei? Na verdade não, existem algumas alternativas! É aí que entram abordagens como **MapReduce**, que se divide em duas etapas:

1. **Map** $\rightarrow$ Mapeia cada registro em um par chave-valor
2. **Reduce** $\rightarrow$ Reúne todos os registros com a mesma chave e gera uma única para cada chave

E o Hadoop é uma implementação OpenSource do MapReduce. O melhor meio de entender o MapReduce é com um exemplo:

> *Exemplo (Contagem de Palavras)*: Vamos imaginar que temos um repositório com milhares de documentos e queremos contar a quantidade de palavras de cada um, será que há um meio de agilizar esse processo? Ou eu vou ter que passar os documentos 1 por 1? Com o MapReduce, esse problema fica computacionalmente viável e eficiente!

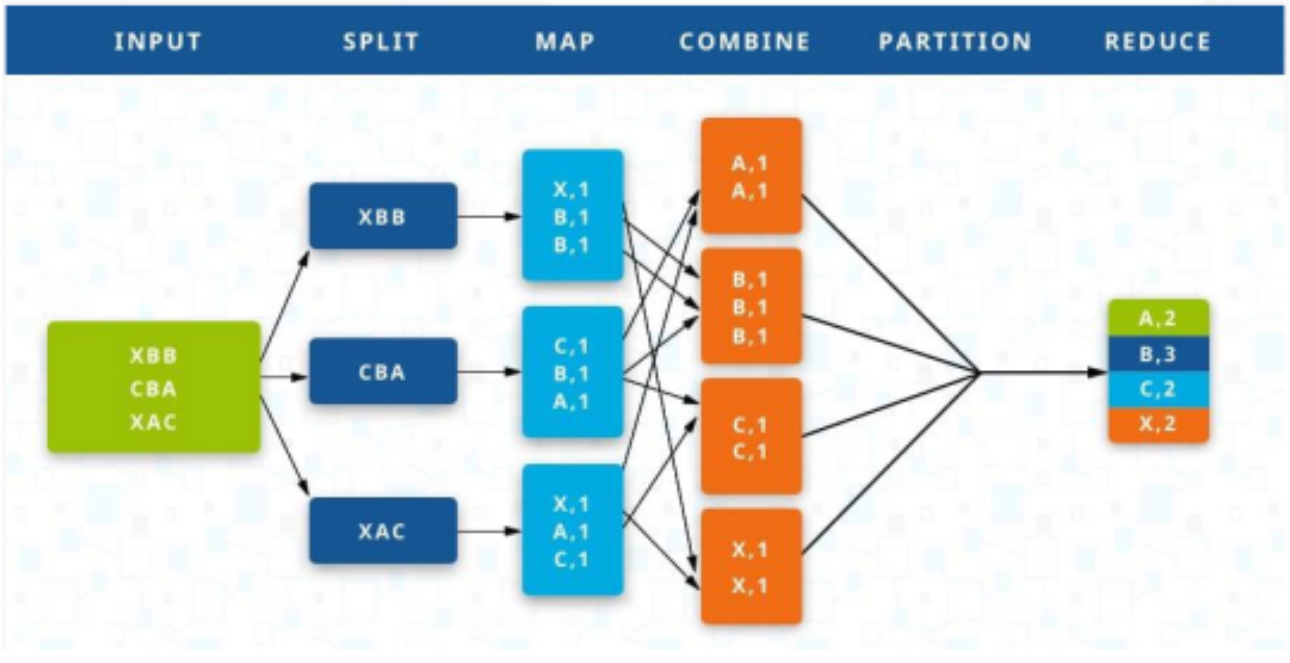

Figura 1: MapReduce no contexto de contagem de palavras

> Nós passamos um ou mais documentos para nosso MapReduce, então ele divide (Seja por linha, parágrafo, etc., na nossa imagem de exemplificação, está separando por linha), e cada divisão é enviada para um node (Fase Split), onde cada node está fazendo o mapeamento das palavras em pares de chave e valor independentemente (Como se cada node fosse um computador separado), então pegamos aqueles valores que possuem chaves iguais (No nosso caso, todas as contagens de cada palavra), e então fazemos o processo de combinar todas em um único par chave-valor de acordo com o contexto. No nosso caso, vamos somar todos os valores para obter todas as quantidades de repetição daquela palavra, obtendo no final, a contagem total de todas as palavras

1.3 Data Lake

Certo, então como que deve ser o processo de tratamento de um problema envolvendo Big Data? Algo muito errado, mas comum, que acontece, é tratar o problema de Big Data como algo separado e distinto, mas ele ta incluso em todo um contexto de problema de dados, na verdade, o mesmo vale para os bancos operacionais e data warehouses! A empresa sempre deve analisar quais tipos de dados são adequados para cada conjunto de dados e fazer sua estruturação e planejamento tendo isso em mente

Certo, dito isso, uma dúvida vem na mente. Os dados operacionais tem seus bancos de dados próprios, os analíticos são extraídos dos data warehouses, e o big data, onde fica?

Definição 1.3.1 (Data Lake): Grande pool de dados não-estruturados (Até o momento da consulta). Dados brutos em seu formato nativo até necessário

- Esquema sob-demanda
- Usuários devem transformar os dados antes da análise

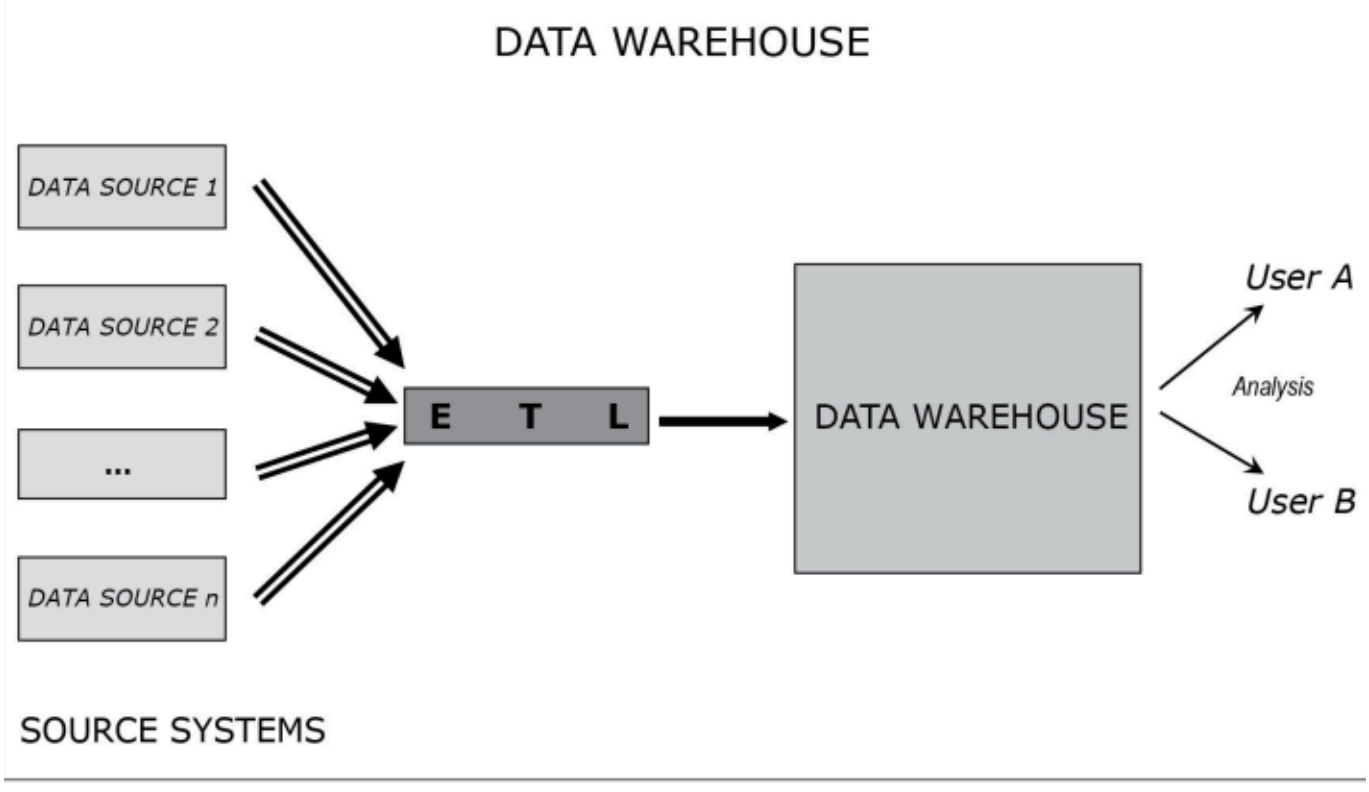

Figura 2: Estrutura e fluxo dos dados em um ecossistema Data Warehouse

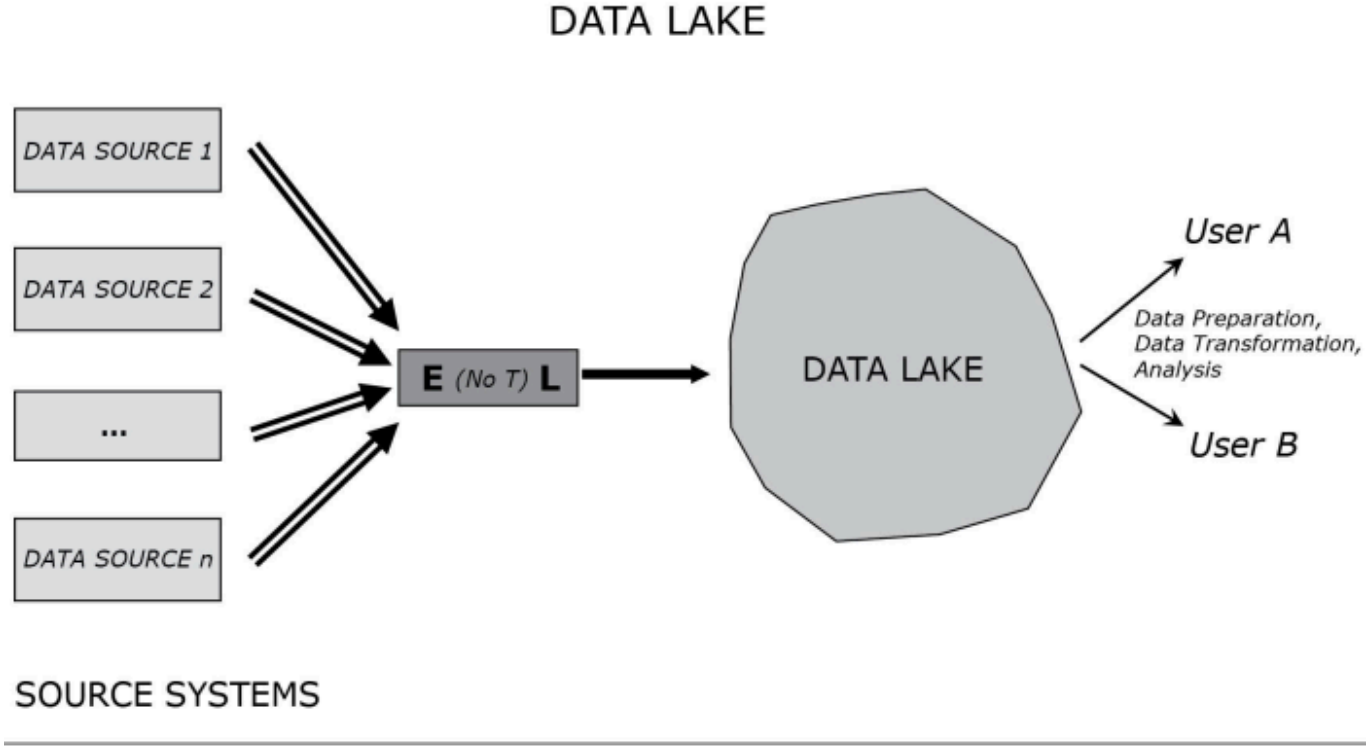

Figura 3: Estrutura e fluxo dos dados em um ecossistema Data Lake

Como falamos, Big Data não é estruturado, então não há transformação dos dados ao serem colocados no Data Lake

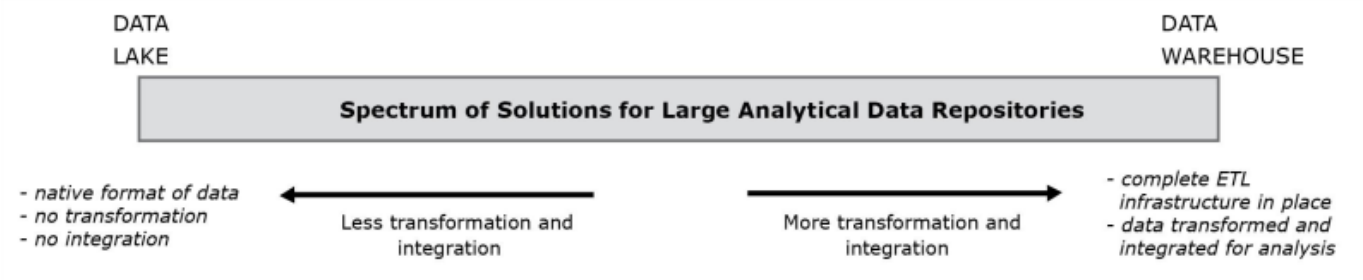

Figura 4: Descrição da melhor solução de repositório de dados para seu problema baseado em características do mesmo