1. **Fundamentação Teórica**

* Explicar Engenharia de Dados
* Explicar ETL
* Mencionar, sucintamente, algumas (talvez poucas) soluções preliminares já identificadas/encontradas para o problema
  + Mencionar que outras soluções poderão ser escolhidas ao longo do projeto
* Explicar as principais tecnologias por trás das soluções mencionadas
  + Hadoop? Spark? ...
* Aprofundar na discussão do cenário escolhido (Ciências sociais)

Data Warehouse

Surgindo de um conceito acadêmico da década de 1980, o data warehouse , DW, vem suprir a necessidade de organizar e obter dados de bases transacionais, que realizam apenas a transação de inserir, atualizar e excluir nos bancos de dados relacionais, como registros de vendas de mercadorias, e favorecer a análise de dados de negócio ou criação de relatórios de gestão.

Para Raslan e Calazans o data warehouse:

deve ser entendido como uma estratégia de organização de dados voltada para a gestão estratégica das corporações, e não para o controle do negócio. (...) Os DW possuem dados sumarizados, agregados e consolidados que são armazenados por longos períodos de tempo (CALAZANS, 2003, p.7). DW é uma forma de gerir volumes muito grande de dados que se encontram, geralmente, espalhados em diversos sistemas de uma organização. Ele possibilita a análise de grandes volumes de dados coletados dos sistemas transacionais (OLTP). Em outras palavras, os DW das corporações são construídos a partir dos diferentes bancos de dados gerenciais de uma organização.(RASLAN e CALAZANS, 2014, p.31)

Por sua vez, complementando no que se relaciona a análise de dados, o OLAP, *Online Analytical Processing*, favorece a visualização e exploração dos dados. Gouveia e Freitas argumentam que:

As tecnologias OLAP, por sua vez, são projetadas para apoiar consultas não triviais, além de auxiliar a sintetizar informações através de comparações, visões personalizadas e dados históricos. As tecnologias OLAP têm como característica principal permitir visões intuitivas dos dados multidimensionais, por meio de análises em diferentes perspectivas. (GOUVEIA e FREITAS, 2018, p. 1-2)

Os DW são modelados pensando que o sistema está em constante crescimento e que os dados devem estar integrados, são orientados por assunto e posições históricas. Os dados não são atualizados ou excluídos no DW, mas apenas inseridos iterativamente. Os DW podem ser ainda divididos em data mart, que são subconjuntos voltados para um tema, assunto, específico

As abordagens normalizada e dimensional são as mais usadas para o armazenamento. Na perspectiva normalizada as tabelas são agrupadas por “áreas temáticas que refletem categorias de dados gerais”(RASLAN e CALAZANS, 2014, p.34). Já na perspectiva dimensional os dados são divididos em fatos, os dados numéricos da transação, e em dimensões, informações de referência que contextualizam os fatos.