Instituto Politécnico do Cávado e do Ave

Integração de Sistemas de Informação

***Big Basket* – Processos ETL**

Licenciatura em Engenharia de Sistemas Informáticos

Nathaniel Joshua Armando da Silva Lloyd Jones

Nº 21116

Barcelos, Portugal

15/novembro/2022

Resumo

**Conteúdo**

# Lista de Figuras

# Introdução

Neste capítulo será elabora os conteúdos deste relatório, dando o leitor a perceber no que consiste este projeto, a contextualização e os objetivos do mesmo, tal como também a estrutura do documento.

# Contextualização

Maior parte dos sistemas de informação exigem uma troca de dados, e com esta exigência podem aparecer vários problemas tais como uma sequência incorreta de informação, inconsistência dos dados ou informação errada ou até um formato de dados errado. Para evitar estes problemas surgiu a área de integração de dados, passando pela análise das regras e processos de negócio e, a partir dai, construir uma solução que permita a comunicação entre dois ou mais sistemas. Um dos processos de integração de sistemas passa pela Extração, Transformação e Carregamento (sigla ETL em inglês) dos dados.

A importância do ETL numa organização é diretamente proporcional quanto à organização do armazenamento de dados. As ferramentas de ETL coletam, leem e migram grandes volumes de dados brutos de várias fontes de dados e entre plataformas diferentes, carregando esses dados num único banco de dados ou um armazenamento de dados para facilitar o acesso. Após isso processam os dados para torná-los significativos com operações como classificação, junção, reformatação, filtragem, incorporação e agregação. Por fim, eles incluem interfaces gráficas para resultados mais rápidos e fáceis do que os métodos tradicionais de movimentação de dados por meio de pipelines de dados codificados manualmente.

As ferramentas modernas de ETL incluem inteligência integrada que deteta e reflete continuamente as alterações nos objetos de origem e destino para garantir a consistência dos dados – a força vital da estratégia e da experiência do cliente.

# Objetivos

O principal propósito deste projeto é a transformação e extração de dados de um ficheiro *.csv*, chamado *Big Basket*, proveniente do website *Kaggle*.

Para a extração destes dados, deve resultar treze ficheiros *Extensible Markup Language (.xml)* e onze ficheiros *JavaScript Object Notation (.json)*.

# Estrutura do documento

O documento encontra-se organizado em .. capítulos, detalhados de seguida.

# Estado de Arte

Qualquer supermercado, tanto de grandes dimensões como de pequenas, tem sempre uma vasta lista de artigos e quantidades dos mesmos, entre outras informações. Para a melhor gestão deste estabelecimento é crucial uma base de dados e uma tecnologia que manipule eficazmente o mesmo consoante as necessidades dos utilizadores.

Sem esta organização um funcionário terá de ver artigo a artigo o que tem em stock, sem mencionar os outros dados que tem de ser tomados em conta, tal como o prazo de validade. Com uma base de dados o trabalho deste funcionário será extremamente simplificado e com uma margem de erros no armazenamento dos dados quase inexistente.

Já existem programas que conseguem resolver estes problemas e foi ao estudar alguns deles que eu consegui desenvolver o meu projeto.

# *Extract, Transform and Load (ETL) Tools*

As ferramentas *ETL* tratam de um processo de três passos de gestão de dados sendo o primeiro a extração dos dados de uma fonte de dados estruturada ou não estruturada, transformando-os num formato que satisfaça os requisitos funcionais e analíticos da empresa e, por fim, carrega os dados no alvo destino.

Um sistema de *ETL* bem desenvolvido extrai os dados, impõem padrões de qualidade dos dados, preenche os dados num formato padronizado para que varias fontes possam ser usadas juntas e entrega os dados prontos para que os desenvolvedores da aplicação a possam desenvolver e os donos do produto possam tomar decisões.

# Conteúdo do ficheiro *.csv*

O ficheiro *.csv* tem os seguintes campos:

* *ProductName* do tipo *String* que guarda o nome do produto;
* *Brand* do tipo *String* que guarda a marca do produto;
* *Price* do tipo *Number* com formato “#.#” que guarda o preço do produto;
* *DiscountPrice* do tipo *Number* com formato “#.#” que guarda o preço do produto em desconto;
* *Quantity* do tipo *String* que guarda a quantidade do artigo em peças, litros ou kilos;
* *Category* do tipo *String* que guarda a categoria do produto;
* *SubCategory* do tipo *String* que guarda a subcategoria do produto;