FACULDADE DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA ELETROTÉCNICA

LICENCIATURA ENGENHARIA INFORMÁTICA

ESTÁGIO PROFISSIONAL

Engenharia de Dados: Proposta de um Pipeline ETL para integração de dados das vendas na CDM

Caso de estudo: **Cervejas de Moçambique**

Autor:

CHAÚQUE, Hélio Carlos

Supervisora:

Eng.ª Leila Omar

Maputo, agosto de 2023

UNIVERSIDADE
EDUARDO
MONDLANE

FACULDADE DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA ELETROTÉCNICA

LICENCIATURA ENGENHARIA INFORMÁTICA

ESTÁGIO PROFISSIONAL

Engenharia de Dados: Proposta de um Pipeline ETL para integração de dados das vendas na CDM

Caso de estudo: Cervejas de Moçambique

Autor:

CHAÚQUE, Hélio Carlos

Supervisora:

Eng.ª Leila Omar

Maputo, agosto de 2023

UNIVERSIDADE EDUARDO MONDLANE

FACULDADE DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA ELECTRÓTECNICA

TERMO DE ENTREGA DE RELATÓRIO DO ESTÁGIO PROFISSIONAL

Declaro que o estudante **Hélio Carlos Chaúque** entregou no dia 04 / 12 / 2023 às 2 cópias do relatório do seu Trabalho da Disciplina de Estágio profissional com a referência: 2023EIEPD204 intitulado: Proposta de um Pipeline ETL para integração de dados das vendas na CDM

Caso de Estudo: Cervejas de Moçambique

Maputo, 04 de dezembro de 2023

A Chefe da Secretaria do DEEL

UNIVERSIDADE EDUARDO MONDLANE

FACULDADE DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA ELECTRÓTECNICA

DECLARAÇÃO DE HONRA

Declaro sob compromisso de honra que o presente trabalho é resultado da minha investigação e que foi concebido para ser submetido apenas para a obtenção do grau de Licenciatura em Engenharia Informática na Faculdade de Engenharia da Universidade Eduardo Mondlane.

Maputo, 04 de Dezembro de 2023.

O Autor

_____

(Hélio Carlos Chaúque)

Dedicatória

Dedico este trabalho aos meus avôs, Hilário Alberto e Salmina Cuna Alberto

Agradecimentos

Em primeiro lugar agradecer a Deus pelo dom da vida, e suporte nesta caminhada e não só (em todas as circunstâncias da vida). Em segunda instância quero agradecer aos meus queridos avós Hilário Benjamim Alberto e Salmina Cuna Alberto já falecidos, eles fizeram a fundação dos meus estudos, foram as primeiras pessoas a me guiarem e mostrar este caminho certo, que acreditaram em mim desde o início até o dia em que partiram, serei eternamente grato e queria poder retribuir um dia. Agradecer também aos meus pais Carlos Chaúque e Isaura Alberto pelo suporte em tudo sem exceção, agradecer também aos meus tios Glória, Otília e Dércio Alberto que fizeram e fazem muito por mim, pelo suporte em todos níveis, agradecer aos meus irmãos e primos que sempre foram parte do motivo para poder ir avante com os estudos e poder servir de inspiração, agradecer a Miralda Simbine por me apoiar nos maus momentos, por motivar sempre e acreditar em mim e no meu potencial, sem esquecer de agradecer aos meus Amigos e colegas do curso o Tomás Mondlane, Pedro Madabula, Gilvaldo Massunguine, Luís Macuvele, Manuel Novela, António Cossa, Cany Mangue, Télvio Sheldon, Sara Tivane, Fátima Massicame e Milton Chiluvane que sempre apoiaram e incentivaram-me em qualquer obstáculo e desânimo para continuar e acreditar, aos Docentes que fizeram parte desta enorme caminhada, pelos ensinamentos, conselhos, experiências, apoio, paciência e boa vontade que me guiaram a melhorar como pessoa e durante o processo de formação profissional.

Agradecer aos amigos da Infância que sempre me acompanharam durante este caminho de forma indirecta, ao Augusto Machapata, Alberto Waite, Dércio Alberto, Armindo Murrombe, Agostinho Macucule, Paulo Cherene, Alfa Mugabe e Dércio Timba.

Agradecer ao Supervisor Nelson Mascarenhas por ter apoiado este projecto, pelas dicas e melhorias durante este período, a Soraya Fernandes minha Directora pela força, motivação e acolhimento, aos restantes colegas do trabalho que ajudaram de forma directa e indirecta.

Epígrafe

“Em Deus nós confiamos, todos os outros tragam dados”.

W. Edwards Deming

Resumo

Actualmente, é irrefutável que o maior ativo nas empresas é o dado. As empresas lidam com a competitividade e para ganhar vantagens da concorrência desenham estratégias e tomam decisão orientados aos dados. Ultimamente o termo *Business intelligence* vem ganhando notoriedade nas organizações para impulsionar os negócios de forma inteligente. As organizações têm gerado milhares de dados diariamente nas suas operações sendo as principais, vendas e interação com clientes. Estes dados são gerados em diversas fontes e formatos o que torna desafiador ter de usá-los para as análises e posterior tomada de decisão. As soluções de Engenharia de dados ajudam a integrar esses dados, com o conceito de *pipeline* de dados pois oferecem mecanismos para centralização e padronização de dados. O objectivo principal do trabalho é propor um *pipeline* ETL para colectar, transformar e disponibilizar dados para gerar relatórios e *dashboards.* Verificado esse cenário e os consequentes constrangimentos foi realizada uma pesquisa bibliográfica sobre as principais soluções de integração de dados para um estudo profundo delas. Para a escolha da solução ideal para o caso em estudo foi imperioso fazer uma análise comparativa das possíveis ferramentas. Com base nesta comparação chegou-se à conclusão de que a solução que melhor se adequa a realidade da empresa CDM é o *SQL Server Integration Services* e prosseguiu-se com a sua implementação. Para testar a solução adotada foi necessário criar um processo ETL em um computador adequado à realidade do negócio.

Palavras-chave: Engenharia de dados, *Business Intelligence*, *Dashboards, SQL Server Integration Services e* Integração de dados*.*

Abstract

Currently, it is irrefutable that the biggest asset in companies is data. Companies deal with competitiveness and to gain competitive advantages, they design strategies and make data-driven decisions. Lately, the term Business intelligence has been gaining notoriety in organizations to drive business intelligently. Organizations have generated thousands of data daily in their operations, the main ones being sales and interaction with customers. This data is generated in different sources and formats, which makes it challenging to use them for analysis and subsequent decision making. Data engineering solutions help to integrate this data, with the concept of data pipeline as they offer mechanisms for centralizing and standardizing data. The main objective of the work is to propose a data pipeline to collect, transform and make data available to generate reports and dashboards. Once this scenario and the consequent constraints were verified, bibliographical research was carried out on the main data integration solutions for an in-depth study of them. To choose the ideal solution for the case under study, it was imperative to carry out a comparative analysis of the possible tools. Based on this comparison, it was concluded that the solution that best suits the CDM company's reality is Azure Data Factory and implementation continued. To test the adopted solution, it was necessary to create a data pipeline reflecting the reality of the business.

Keywords: Data Engineering, Business Intelligence, *Dashboards*, *SQL Server Integration Services* and Data Integration.

Índice

Índice de figuras

Índice de tabelas

Lista de abreviaturas e acrónimos

BI	Business Intelligence
IT	Information Technology
TIC	Tecnologias de Informação e Cominicação
HTTPs	Hypertext Transfer Protocol Secure
SSAS	SQL Server Analysis Service
SSRS	SQL Server Report Service
DW	Data Warehouse
DL	Data Lake
ETL	Extract, Transform and Load
ELT	Extract, Load and Transform
IoT	Internet of Things
CIC	Centro de Interação com Cliente
BDR	Business Development Representative
DS	Distributor Specialist
POC	Point of Consume
WEB	World Wide Web
DBA	Database Administrator
CSV	Comma Separated Values
KPI	Key Performance Indicator
SKU	Stock Keeping Unit

Glossário de Termos

Internet - Maior rede de computadores do mundo que interliga o mundo todo.

Ativo – Qualquer recurso que tenha valor para organização, podendo ser hardware, dado etc.

Data Warehouse - Repositório ou armazém de dados estruturados

Data Lake – Repositório ou lago de dados não estruturados, semiestruturados e estruturados;

Dashboard – Painel de Controle com principais métricas do negócio.

Stakeholder - são indivíduos ou grupos que podem afetar ou serem afetados pelos resultados ou pelo desempenho de uma organização ou projeto.

On-Premise - Diz respeito ao ambiente local.

Data-Driven – Orientado a dados.

OneDrive – OneDrive é um serviço de armazenamento em nuvem da Microsoft, projetado para permitir que os usuários armazenem, acessem e compartilhem arquivos e dados de forma segura pela internet.

Excel - É um software de planilha amplamente utilizado que faz parte do pacote de aplicativos do Microsoft Office. Ele oferece recursos poderosos para criação, manipulação e análise de dados em formato de planilha.

PowerBI – É uma ferramenta da Microsoft de Análise de dados e visualização de dados.

KUJA – Sistema web usado para registar as vendas SellOut (do armazém ao POC).

Syspro – Sistema usado para registar as vendas SellIn (da fábrica ao armazém).

SellIn – Vendas diretas da Fábrica ao Armazém.

SellOut – Vendas indiretas do Armazém ao ponto de Consumo (POC).

Insight - é um termo utilizado para descrever uma compreensão profunda e repentina de um problema, situação, ou aspecto específico.

Stock – Neste contexto, refere-se ao produto mantido ou disponível no armazém.

Big Data – é um termo que se refere a conjuntos de dados extremamente grandes e complexos que não podem ser facilmente processados com ferramentas de processamento de dados tradicionais.

Poster - cartaz que ilustra a disponibilizadade das marcas de cerveja, inclui outras informações como preço.

Distric Manager – Directores regionais, nomeadamente, Maputo, Sul (Gaza e Inhambane), Centro e Norte.

Sales Manager - Gestor de vendas em uma região.

Capítulo I – Introdução

1.1. Contextualização

Na era de tecnologia, quase tudo no mundo gira por volta dela, a tecnologia é aplicada em quase todos os sectores da vida. A cada segundo são gerados grandes volumes de dados, as fotografias que tiramos, os vídeos que assistimos no Youtube, os seriados e filmes assistimos na Netflix, as mensagens que trocamos com amigos e familiares pelo WhatsApp, os sensores das máquinas nas fábricas, as postagens que fazemos nas redes sociais estes são apenas alguns exemplos desse grande volume de dados.

No ramo empresarial são também gerados dados em várias fontes, dados de vendas, relação com clientes, dados de marketing, gestão interna de recursos etc. As Empresas precisam usar estes dados para melhorar a estratégias comerciais, ganhar mais lucros e descobrir novas oportunidades de investimento, neste âmbito, os gestores precisam de basear-se nos dados para tomar decisões assertivas na perspectiva de gerar mais lucros e minimizar as perdas. Na era da tecnologia as decisões são tomadas baseadas em dados e não de forma empírica, a tecnologia suporta a tomada de decisões oferecendo ferramentas poderosas para colectar, transformar, armazenas para análise e visualização de dados.

Os dados nas empresas são gerados por vários sistemas, e estes podem ser internos (dados transacionais) e ou externos (interação com clientes), em diversos formatos e locais, isso remete-nos que estamos diante de dados heterogéneos. Havendo assim a necessidade de colectar, transformar ou consolidar e disponibilizá-los para várias finalidades como análise e aprendizado de máquina. Antes da análise os dados passam por estas etapas a se levar em conta.

Um dado bruto não é suficientemente útil para um tomador de decisões. O presente trabalho consiste em descrever e executar as etapas de engenharia de dados.

1.2. Motivação

Face a morosidade que leva o processo de obtenção de dados para análise, sinto-me motivado em melhorar os processos internos da empresa que direcionam as operações usando ferramentas tecnológicas para flexibilizar as etapas de análise e disponibilização de dados.

Para contribuir na ciência com a aplicação dos conceitos da engenharia para ajudar o negócio a melhorar na camada estratégica e tática a ganhar frente aos concorrentes directos no mercado.

Outra motivação cientifica é facto de poder contribuir na literatura sobre o conteúdo de engenharia de dados, em especifico *pipeline* ETL que é muito escaca.

Outra grande motivação é referente a satisfação individual em colocar em prática os tópicos estudados durante a formação académica de forma adequada na perspectiva de resolver os problemas da sociedade (da Empresa em particular).

1.3. Descrição do problema

Em uma era domina pela tecnologia os dados são classificados como o novo petróleo do século XI. Nos últimos anos tem-se gerado um grande volume de dados e essa tendência vem crescendo cada vez mais. Soluções são criadas na tentativa de acompanhar o rítmo de geração de dados para o processamento e armazenamento em larga escala. Segundo Da Silva *et al.* (2016) afirmam que, a máxima de que a informação é a alma do negócio nunca foi tão actual. A cada dia aumenta a necessidade de as empresas tomarem decisões estratégicas com base nos dados históricos e informações em tempo real.

Actualmente é essencial para uma empresa obter controle total sobre os seus dados, visto que por meio desses dados, é possível identificar os erros e acertos que estão sendo cometidos, além de facilitar na obtenção de *insights* e auxiliar na tomada de decisão (Kondado, 2022).

No dia a dia nas actividades laborais para gerar relatórios, *Dashboards* aos tomadores de decisões, há necessidade de cruzar dados vindo de várias fontes e formatos o que torna a tarefa nada trivial aos analistas, o que leva bastante tempo para consolidar e disponibilizar essa informação, tornando os processos internos lentos para o nível

operacional, sem contar com a necessidade de a informação ser visualizada a tempo real para traçar estratégias instantâneas para ganhar vantagem à concorrência.

Os processos de análise de dados carecem de soluções que possam minimizar o tempo de disponibilizar informação e flexibilidade em gerar os relatórios.

1.4. Pergunta de pesquisa

No final do projecto de pesquisa como resultado espera-se responder a seguinte pergunta:

De que forma a integração de dados pode melhorar eficiência operacional das vendas na CDM?

Problema: São vários problemas identificados, destaca-se o seguinte: Ineficiência operacional (Na ausência de um processo ETL, os processos internos tornam-se mais lentos e suscetíveis a erros, afectando a eficiência operacional das vendas).

1.5. Objetivos

1.5.1. Geral:

- Propor um Pipeline ETL para integração de dados das vendas.

1.5.2. Específicos:

- Conceituar os principais pontos referentes a ETL e a inteligência de negócios;
- Explicar a situação actual dos processos que geram os dados para análise;
- Descrever e comparar as possíveis soluções ETL para a solução;
- Implementar a solução ideal aos processos de negócio para análise de dados.

1.6. Metodologia

Nesta secção vão ser apresentados os métodos utilizados para realizar o trabalho de pesquisa de modo a atingir os objetivos definidos. Mostrando como se vai realizar a colecta, análise de dados.

Para alcançar o primeiro objetivo, conduziremos uma pesquisa exploratória. Este enquadra-se naqueles que buscam descobrir ideias e intuições, na tentativa de adquirir maior familiaridade com o fenômeno pesquisado. Eles possibilitam aumentar o conhecimento do pesquisador sobre os factos (Selltiz et al, 1965) citados por (Oliveira M. F., 2011). Recorrer-se-á a bibliografia para buscar conteúdo sobre os processos de ETL, para melhor entendimento do seu funcionamento e relevância para a CDM.

Para o segundo objectivo, usar-se-á o procedimento de estudo de caso. O estudo de caso que visa dar uma compreensão aprofundada do caso em estudo. Segundo Gil (1991) citado por (Menezes & Da Silva, 2005) estudo de caso é realizado quando a pesquisa envolve o estudo profundo e exaustivo de um ou poucos objectos de maneira que se permita o seu amplo e detalhamento conhecimento. Neste objectivo a ideia é descrever como funciona o processo das vendas, a geração de dados até a fase da análise.

Para o terceiro objectivo, adotaremos uma pesquisa de abordagem qualitativa. Segundo Creswell (2007) a abordagem qualitativa provê ao pesquisador um conhecimento mais profundo de um fenômeno e produz um alto nível de detalhes. Esta abordagem se concentra em compreender as complexidades de um fenômeno, esta é frequentemente descritiva e exploratória. Aqui serão apresentadas e descritas as possíveis soluções e realizar-se-á em seguida uma comparação das mesmas, de acordo com os critérios relevantes para CDM vai se definir a ferramenta ETL ideal. Para o quarto e último objectivo, realizaremos a pesquisa de natureza Aplicada. A pesquisa aplicada visa em gerar conhecimento para aplicação prática, dirigidos a solução de problemas específicos. Envolve verdades e interesses locais (Menezes & Da Silva, 2005).

Será implementada uma solução ETL para a integração de dados que são geradas nas diversas fontes adequada as necessidades da área das vendas para melhor prover informação aos tomadores de decisão.

1.7. Técnicas de coleta de dados

Colecta de dados é um processo que visa reunir dados sobre um determinado assunto de interesse para o uso secundário por meio de técnicas específicas de pesquisa. Para este trabalho irá-se usar duas técnicas, pesquisa bibliográfica e pesquisa documental.

Pesquisa Bibliográfica

Fez-se uma revisão na literatura com a finalidade de reunir conceitos gerais relativas a ETL e *Business Intelligence* para a produção deste trabalho de pesquisa*.* Para este fim efectuou-se uma pesquisa bibliográfica, através da qual pôde-se inteirar sobre o que já foi debatido sobre o assunto em discussão, esta foi possível através da consulta às obras, livros e artigos científicos, variadas páginas da internet, sites e artigos eletrônicos que sejam concedíveis.

Pesquisa Documental

De acordo com Gerhardt e Silveira (2009, p. 69) “Pesquisa documental é aquela realizada a partir de documentos, contemporâneos ou retrospectivos, considerados cientificamente autênticos (não-fraudados); tem sido largamente utilizada nas ciências sociais, na investigação histórica, a fim de descrever/comparar fatos sociais”. Os documentos aqui usados são classificados como os de fonte secundária, nomeadamente, artigos científicos, manuais de procedimentos.

1.8. Técnicas de Análise de dados

Para analisar os dados colectados serão usadas duas técnicas de dados qualitativos nomeadamente análise de conteúdo e de discurso.

- **Análise de conteúdo**

Segundo Bardin (1979, p.42) citado por (Gerhardt & Silveira, 2009, p. 84) “A análise de conteúdo representa um conjunto de técnicas de análise das comunicações que visam a obter, por procedimentos sistemáticos e objetivos de descrição do conteúdo das mensagens, indicadores (quantitativos ou não) que permitam a inferência de conhecimentos relativos às condições de produção e recepção dessas mensagens.”

Esta análise consistiu na descrição do conteúdo que os intervenientes dos debates depuseram, na tentativa de perceber o sentido que manifestaram durante os seus discursos.

- **Análise de discurso**

Na perspectiva de Gerhardt e Silveira (2009, p. 85) “A análise de discurso objectiva realizar uma reflexão sobre as condições de produção e apreensão do significado de textos produzidos em diferentes campos, como, por exemplo, o religioso, o filosófico, o jurídico e o sociopolítico”. Das leituras feitas interpretou-se o conteúdo contido nos textos, na perspectiva de compreender as diferentes ideias que os autores transmitem.

1.9. Estrutura do trabalho

O presente trabalho é composto por seis capítulos, que são devidamente enumerados a seguir, e mais uma secção de referências bibliográficas que não foi enumerada.

- **Capítulo I – Introdução:**

Neste capítulo são apresentados de forma simples e clara a formulação do trabalho de investigação. Compostos pela contextualização, motivação, definição do problema, objectivos e metodologia.

- **Capítulo II – Revisão da Literatura:**

Neste capítulo faz-se uma análise, referente ao trabalho e aos dados da pesquisa, seguindo um roteiro lógico, envolta do tema do trabalho, no que concerne a integração de dados, inteligência de negócios, onde a aplicação das TICs é imprescindível para solucionar o problema apresentado.

- **Capítulo III – Caso de Estudo:**

Neste capítulo, apresenta-se uma descrição atual aprofundada da situação, em seguida ir-se-á propor um processo ETL para a integração de dados, para área das vendas da CDM, e obter conclusões relacionados com o problema apresentado.

- **Capítulo IV – Desenvolvimento da solução proposta:**

Neste capítulo, após ter-se apresentado de forma clara e simples o problema, então propõe-se uma solução ETL ideal na tentativa de resolver o problema anteriormente identificado e apresentado.

- **Capítulo V – Discussão de resultados:**

Neste capítulo, são apresentados as análises das abordagens dos vários autores e os resultados descobertos dos estudos realizados e os possíveis impactos que esta solução irá oferecer.

- **Capítulo VI – Considerações finais:**

Neste capítulo apresentam-se a análise dos dados e a interpretação dos resultados, com maior preocupação na verificação do cumprimento dos objectivos inicialmente propostos.

Secção das Bibliografias

Na secção das bibliografias são incluídas todas as obras, livros, artigos, trabalhos de conclusão de curso, dissertações, documentações que ajudaram na elaboração deste trabalho.

Capítulo II – Revisão da Literatura

Neste capítulo serão abordados os pontos chave do trabalho que passa por conceituar os termos relacionados a ETL e os seus intervenientes, para uma melhor emersão ao problema a solucionar.

2.1. Dado

Segundo Oliveira (2005) citado por Mülbert e Ayres (2011, p.22), “Dado é qualquer elemento identificado em sua forma bruta que, por si só, não conduz a uma compreensão de determinado fato ou situação”. Para Lampert e Badalotti (2015) dado é um caractere no seu estado bruto, sem que haja transformações nela pode ser representado por símbolos, números. Com estes conceitos acimas citados, percebe-se que um dado é apenas um elemento, facto bruto sem ser contextualizado não tem significado.

2.2. Informação

Segundo Stair e Reynolds (2015, p.5), “Informação é uma coleção de factos organizados e processados de modo que tenham valor adicional, que se estende além do valor de fatos individuais”. Para Cezar (2016) O “conceito de informação envolve aspectos que abrangem desde sua colecta na forma bruta (dados), conduzindo ao processamento, que pode ser sob a forma de agrupamentos, cálculos, transformação dos dados, até sua disponibilização para a tomada de decisão”. O mesmo autor afirma ainda que a informação está vinculada à capacidade de relacioná-la ao contexto ao qual pertence, podendo estar associada a uma ação ou regra.

2.3. Conhecimento

Stair e Reynolds (2015, p.6) afirmam que “Conhecimento é a consciência e compreensão de um conjunto de informações e maneiras como essas informações podem ser uteis para apoiar uma tarefa específica ou para chegar a uma decisão”. Para (De Cezar, 2016), conhecimento é um método de transformação que envolve informações, meios (objectos) e pessoas e se desenvolve por aprendizagem com base em experiências anteriores, acúmulo de informações e vivências adquiridas com o tempo. Os autores trazem abordagens diferentes e complementares para este conceito, porém as duas estão certas, em forma de síntese o conhecimento resume-

se em concepção da informação ou aprendizado suficiente para apoiar uma ação a tomar.

2.4. Decisão

De acordo com Lima (2012, p. 28), “Decisão é o processo de análise e escolha entre as alternativas disponíveis de cursos de ação que a pessoa deverá seguir”. Em linhas gerais uma decisão é acto de escolha coerente duma posição dentre várias alternativas em benefício específico.

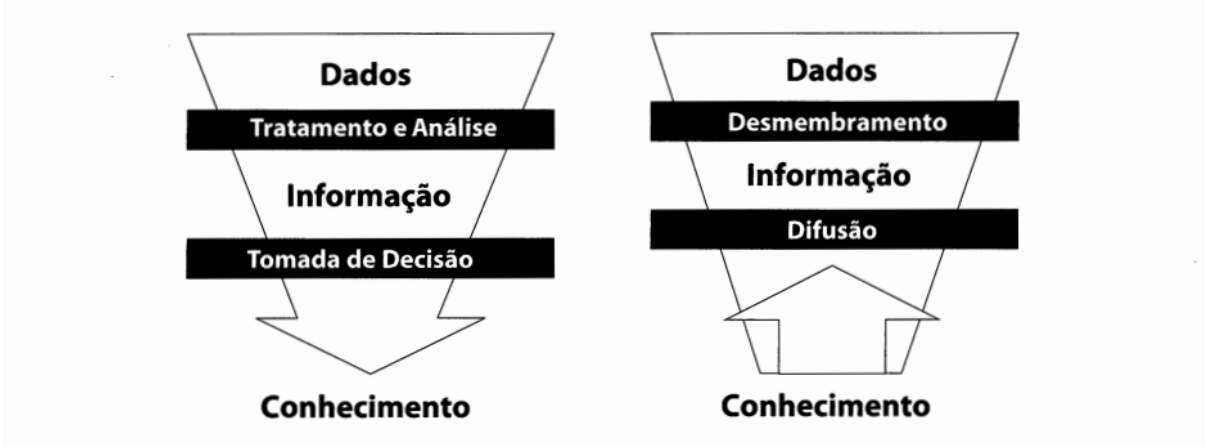

Figura 1: Hierarquia de Dado, Informação e Conhecimento

Fonte: (Favero & Belfiore, 2017)

2.5. Engenharia de dados

Para Reis e Housley (2022, p. 3), “Engenharia de dados é um conjunto de operações que visam a criação de interfaces e mecanismos para o fluxo e acesso à informação”.

Para uma explicação mais clara sobre Engenharia de dados segundos os autores acima citamos, Engenharia de dados é o desenvolvimento, implementação e manutenção de sistemas e processos que recebem dados brutos e produzem informações consistentes e de alta qualidade que suporta casos de uso, como análises e aprendizado de máquina. Ainda os mesmos autores, o Engenheiro de dados é o profissional responsável por manter os dados disponíveis e utilizáveis por terceiros, ou seja, um engenheiro de dados obtém dados, armazena-os e prepara-os para consumo por cientistas de dados, analistas e outros. O processo de engenharia

de dados é a fase inicial para um projecto de análise de dados, modelos de aprendizado de máquina.

2.5.1. Pipeline de dados

Segundo Densmore (2021, p. 1), “Pipelines de dados são conjuntos de processos que movem e transformam dados de várias fontes para um destino onde o novo valor pode ser derivado. Eles são a base de análises, relatórios e aprendizagem de máquina”. Para Munappy *et al* (2020, p. 2), “Pipeline de dados são a cadeia conectada de processos em que a saída de um ou mais processos se torna uma entrada para outro. É um software que remove muitas etapas manuais do fluxo de trabalho e permite um fluxo simplificado e automatizado de dados de um nó para outro”.

O *pipeline* de dados é uma abordagem mais ampla, englobando não apenas a extração, transformação e carregamento de dados, podendo incluir não apenas a integração, mas também outras actividades como, mas também a orquestração, ingestão, transmissão e qualquer outra actividade associada ao gerenciamento de dados.

Uma analogia feita a um *pipeline* de dados é com a tubulação de petróleo, o petróleo é extraído no seu estado bruto, passando por várias etapas de refinação até virar combustível para alimentar os automóveis. É exactamente desta forma como funciona com os dados. Estes são extraídos da origem na sua forma bruta, quase sempre se apresentam inconsistentes, em seguida passam por uma etapa de transformação onde são padronizados, enriquecidos e então armazenados em um repositório central e a partir daí disponíveis para o uso, alimentar *Dashboards*, modelos de aprendizados de máquina e outros.

Um *pipeline* de dados tem importância em um ambiente organizacional, pois este automatiza várias etapas manuais envolvidas na transformação e otimização de carregamento de dados, normalmente um *pipeline* inclui carregar dados em várias origens para uma área intermediária chamada de *Staging Area* onde são armazenados temporariamente, e em seguida são transformados e por fim inseri-los no destino, onde são armazenados de forma permanente, podendo então ser usualmente em um *Data Warehouse* ou *Data Lake*. Há que salientar que um pipeline é um conceito e que pode ser implementado de muitas formas, podendo ser por meio de ferramentas ou programação, cada uma destas soluções apresentam vantagens

e desvantagens podendo se julgar qual melhor forma de implementar adequando-se as necessidades de cada propósito. Um *pipeline* pode ser encadeado, onde a saída de um define a entrada doutro *pipeline*.

2.5.2. Pipeline ETL

Os *Pipelines* ETL são um conjunto de processos usados para transferir dados de uma ou mais fontes para um banco de dados, como um *Data Warehouse*. Extração, transformação e carregamento são três procedimentos interdependentes usados para movimentar dados (Sharma, 2022). Para Fátima (2023) Pipeline ETL é um conjunto de processos que inclui a extração de dados de uma variedade de fontes e sua transformação. Os dados são subsequentemente carregados nos sistemas de destino, como em nuvem, *Data Mart* ou uma BD para análise ou outros propósitos. Um *pipeline* ETL é constituído essencialmente três (3) por elementos: Origem, Processamento e Destino.

a) **Origem**: Considera-se origem de dados todo sistema ou ferramenta que captura e armazena dados. É onde residem os dados no seu estado bruto.

Segundo Reis e Housley (2022), “As fontes produzem dados consumidos por sistemas *Downstream*, incluindo planilhas geradas por humanos, sensores IoT, aplicações web e móveis”. Os dados podem ser gerados por humanos e por máquinas (Marquesone, 2017). Dados gerados por humanos são aqueles em que o conteúdo foi gerado a partir do pensamento de uma pessoa, na qual a propriedade intelectual está integrada ao dado, pode se entender também como sendo dados que refletem a interação das pessoas no mundo digital (Marquesone, 2017). Geralmente grande parte desses dados são oriundas de mídias sociais, *WhatsApp*, *Instagram*, *Snapchat*, *blogs* e outros. Os dados gerados por máquinas, Marquesone (2017) afirma que são dados digitais produzidos por processos de computadores, aplicações e outros mecanismos sem necessitar explicitamente a intervenção humana, sensores IoT, arquivos de log e outros.

b) **Processamento**: De acordo com Ralph e Reynolds (2015), processamento de dados em sistema de informação é converter ou transformar dados brutos em úteis. Esta é a etapa mais complexa e extensa de um *pipeline*, é onde ocorre o processamento, limpeza, transformação, gestão de metadados e ou enriquecimento dos dados.

Suponhamos que em um campo de sexo de uma tabela de registro de clientes os dados registados apareçam com as seguintes variações: “Masculino, Masc, M, 1”, estas variações remetem a informar sobre o sexo masculino apresentado assim não se pode levar esta inconsistência à uma análise e muito menos para construir um modelo de aprendizado de máquina.

c) **Destino**: É onde são armazenados os dados depois de serem processados, limpos e enriquecidos. Geralmente são armazenados em um *Data Warehouse*, *Data Lake*, *Data Store*, em um ambiente local, em nuvem ou mesmo podem ser usados em tempo real.

Nesta etapa os dados estão disponíveis para atender as perguntas do negócio, podendo gerar relatórios, construir *dashboards* e modelos de aprendizado de máquina.

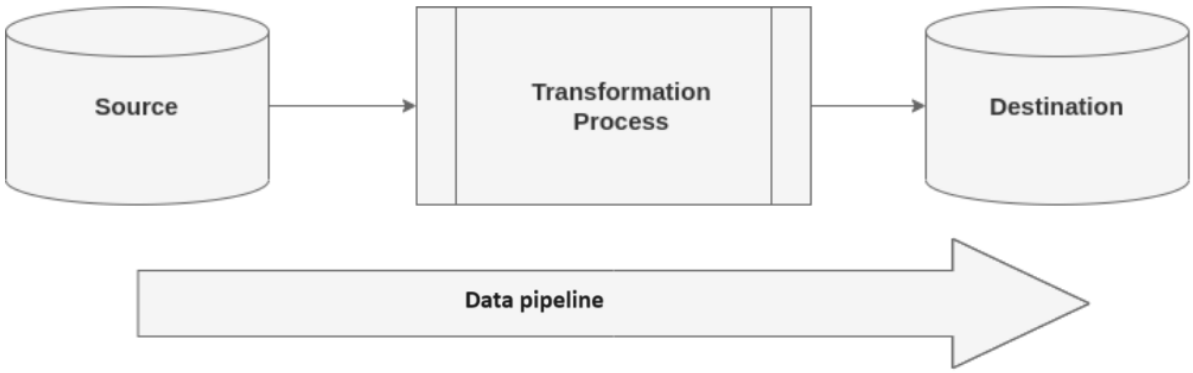

Figura 2: Pipeline de Dados.

Fonte: Adaptado de Suleman (2023).

2.5.3. Pipeline de Dados x Pipeline ETL

Estes dois conceitos deixam uma margem de dúvidas quanto a sua definição, vamos agora descrever as suas diferenças. *Pipeline* ETL é um tipo específico de *pipeline* de dados focado na extração, transformação e carga de dados para propósitos analíticos e de relatórios, já o termo *Pipeline* de Dados é mais abrangente, englobando qualquer fluxo automatizado de dados, independentemente de incluir ou não as fases tradicionais de ETL. Este pode incluir também a orquestração, ingestão, transmissão e qualquer outra actividade associada ao gerenciamento de dados.

2.5.4. Etapas ETL e ELT

Para conceituar essas duas terminologias vamos começar pelo ETL, que significa extrair, transformar e carregar.

Extract (Extrair)

Para Braghitoni (2017, p. 24), “O Extact é o processo de extração periódica dos dados das origens por meio da leitura de uma ou mais fontes de informação”. Densmore (2021, p. 22) afirma que “A etapa de extração reúne dados de várias fontes em preparação para transformar e carregar”. Os conceitos acima descritos trazem uma clara percepção da etapa de extração, onde ocorre a busca e colecta dos dados das diversas fontes.

Transformation (Transformação)

A etapa de transformação é onde os dados brutos de cada sistema de origem são combinados e formatados de tal forma que são úteis para analistas, ferramentas de visualização ou qualquer caso de uso que seu *pipeline* esteja servindo (Densmore, 2021). Depois que os dados são extraídos, colectados e armazenados temporariamente eles passam por esta etapa onde são padronizados, limpos, enriquecidos de modo a garantir qualidade e consistência no destino.

Load (Carga)

De acordo com Densmore (2021, p. 22), “A etapa de carga traz os dados brutos (no caso de ELT) ou dados totalmente transformados (no caso de ETL) para o destino. De qualquer forma o resultado é o carregamento de dados no DW, DL ou noutro destino”. Etapa onde os dados são armazenados no destino.

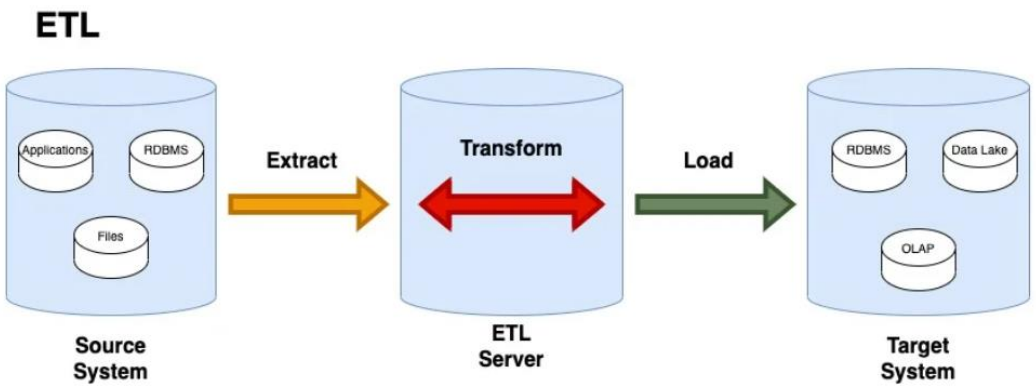

Figura 3: Esquema do Processo ETL

Fonte: (Nicollete, 2022)

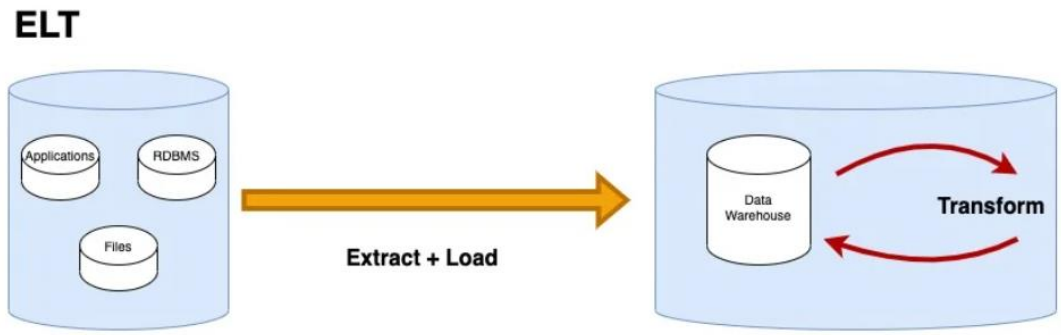

Figura 4: Esquema do processo ELT

Fonte: (Nicollete, 2022)

A diferença entre ETL e ELT está na ordem em que os dados são carregados no destino, no ETL os dados são transformados antes de serem carregadas no destino enquanto no ELT os dados são armazenados antes da transformação. Há casos em que será necessário carregar dados no destino antes de transformá-los, na situação de dados gerados em tempo real e em lote, querendo transformá-los antes faria com que dados não sejam capturados.

2.6. Business Intelligence

Business intelligence ou simplesmente BI é um termo cunhado por Howard Dresner em 1989, para descrever um conjunto de conceito e métodos para melhorar o processo de tomada de decisão das empresas, utilizando-se de sistemas fundamentados em factos e dimensões (Braghittoni, 2017). O mesmo autor afirma que o BI se baseia em agrupar informações de diversas fontes e apresentá-los de forma unificada e sob uma métrica comum, a fim de indicadores aparentemente distantes possam fazer sentido entre si.

Segundo Howard Dresner Citado por Braghittoni (2017), “BI é uma metodologia pela qual se estabelecem ferramentas para obter, organizar e prover acesso às informações necessárias aos tomadores de decisão das empresas analisarem os fenômenos acerca dos seus negócios”.

Da definição acima citada é interessante observar que, BI é uma metodologia, não uma ferramenta. Isso significa que se pode implementar BI com praticamente

qualquer ferramenta de controle de dados, ou com o conjunto de quaisquer ferramentas próprias de BI, (Braghittoni, 2017).

BI serve para analisar os fenómenos acerca do negócio, isso significa que o BI precisa ser uma plataforma capaz não só de aglutinar as informações transacionais, mas também de exibi-las, fazendo com que fenómenos escondidos se tornem visíveis, (Braghittoni, 2017).

2.6.1. Data Source

Segundo Kimball e Ross (2002, p. 7), "*Data Source* são os sistemas operacionais de registro que capturam as transações do negócio". As fontes de dados são onde os dados residem, que geralmente em um ambiente corporativo em base de dados, onde são armazenadas as transações diárias do negócio, podendo ser também em ficheiro Excel e outros.

2.6.2. Data Staging

Segundo Kimball e Ross (2002), O *Data Staging* é uma área de armazenamento e um conjunto de processos chamados *ETL*, está entre os sistemas de origem e área de apresentação de dados, ou seja, repositório de dados.

O *Data Staging* é a área onde os dados são armazenados temporariamente antes de serem transformados. Os dados após serem extraídos da origem eles precisam ser armazenados em algum lugar para que se siga o processo de limpeza e transformação.

2.6.3. Data Warehouse

Os dados após passarem pelo processo de limpeza e transformação são armazenados permanentemente em um repositório central o dito *Data Warehouse* para finalidades de relatórios e análises. Um DW armazena dados estruturados e é criado com o propósito de facilitar acesso à informação para o negócio. A partir do DW os dados são conectados ao destino, para o alimentar *Dashboards*, modelos de aprendizado de máquina, relatórios e outros.

De acordo com Watson e Haley (1998) citados por (Oliveira J. V., 2009), "Sistemas de *Data Warehouse* possibilitam que as organizações tenham acesso à informação

de gestão que é determinante para obterem ganhos significativos, nomeadamente, aumento de vendas, redução nos custos, oferta de novos serviços e produtos".

> Um *Data Warehouse* é um repositório de dados corporativos integrados. Um armazém de dados é usado especificamente para suporte à decisão, ou seja, existe (normalmente ou idealmente) apenas um *Data Warehouse* em uma empresa. Um *Data Warehouse* normalmente contém dados colectados de um grande número de fontes dentro e às vezes também fora da empresa (Jensen, Pedersen, & Thomsen, 2010, p. 3).

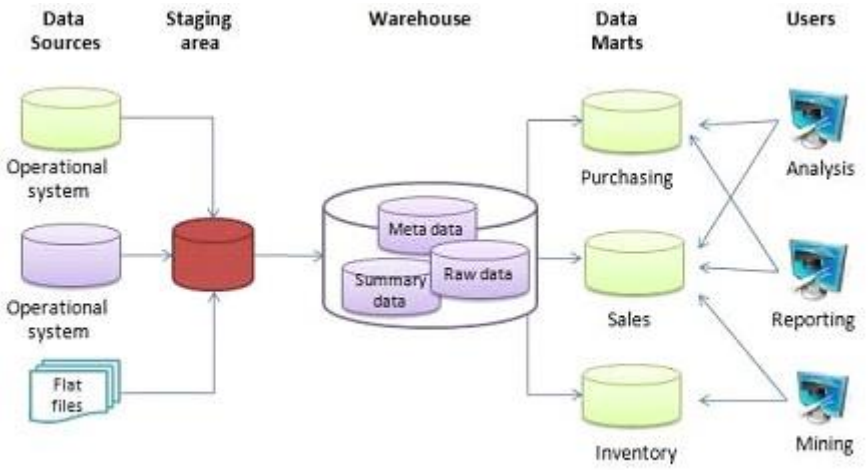

Figura 5: Arquitectura de um Data Warehouse

Fonte: https://commons.wikimedia.org/wiki/File:Data_warehouse_architecture.jpg

Bill Inmon em 1989 descreveu *Data Warehouse* como "uma coleção de dados orientado por assunto, integrado, não volátil e variante com o tempo para apoiar a tomada de decisões".

- **Orientado por assunto**

Os dados quando armazenados são segmentados por assuntos diferentes. Por exemplo vendas, stocks, pagamentos em pequenos blocos chamados *Data Mart*.

- **Integrado**

Os dados provêm de várias fontes e geralmente apresentam inconsistências, precisam ser integrados, ou seja, padronizados.

- **Não volátil**

Após dados serem carregados num DW não mais podem ser alterados, podendo apenas ser consultados.

- **Variante com o tempo**

Significa que um *Data Warehouse* armazena dados históricos, ou seja, de anos anteriores.

Os principias benefícios de um Data Warehouse são:

a) **Dados Integrados** – Integrar dados de diferentes sistemas;
b) **Consultas mais rápidas** – O DW é projetado para lidar com consultas grandes, executa consultas mais rápida que uma base de dados convencional;
c) **Qualidade de dados melhorada** – Os dados passam por processo de limpeza, enriquecimento e validação antes do armazenamento, garantindo a precisão, integridade e consistência.
d) **Visão Histórica de dados** – são armazenados dados históricos com detalhes do negócio, podendo analisar a qualquer momento.

2.6.4. Data Mart

Segundo Jensen *et al*. (2010, p. 16), “Um *Data Mart* é geralmente considerado um subconjunto de um *Data Warehouse*. Enquanto um *Data Warehouse* pode conter dados sobre diferentes assuntos, uma Data *Mart* contém dados sobre um único assunto, por exemplo, vendas”. Uma *Data Mart* pode ser vista como uma pequena *Data Warehouse* segmentado por assunto em uma organização.

2.6.5. Data Lake

É um grande lago de dados onde são armazenados todos os tipos de dados, sendo eles estruturados, semi-estruturados e não estruturados.

2.7. Modelo Dimensional

Para Machado (2013) “A modelagem dimensional é a técnica estruturada desenvolvida para obtenção de modelos de simples entendimento e alta performance de acesso aos dados”.

Um DW é formado por dois tipos de entidade, a facto e a dimensão. O modelo dimensional é um tipo de modelagem de dados virado ao negócio. Este é projectado para atender as visões do negócio.

Tabela Facto

Kimball e Ross (2002), afirmam que a “Facto é a principal tabela em um modelo dimensional onde as medidas numéricas de desempenho do negócio são armazenadas”. Essas medidas são médias, quantidades, valores, preços, etc.

Tabela Dimensão

Para Kimball e Ross (2002, p. 19), “As tabelas de dimensões são acompanhantes integrantes de uma tabela de factos. As tabelas de dimensão contêm os descritores textuais do negócio, em um modelo dimensional bem projectado, as tabelas de dimensão têm muitas colunas ou atributos”.

2.7.1. Modelo Estrela e Snow flake

De acordo com Braghittoni (2017), existem algumas regras para implementar um *Data Warehouse*, a representação do esquema dum DW será baseada em um desenho dimensional (factos e dimensões), este desenho pode ter dois esquemas.

Modelo estrela

“Um esquema em estrela possui uma tabela de dimensão para cada dimensão. Esta tabela possui uma coluna chave e uma coluna para cada nível da dimensão ligada unicamente a uma tabela facto” (Jensen, Pedersen, & Thomsen, 2010).

As tabelas dimensão estão directamente relacionadas com a tabela facto.

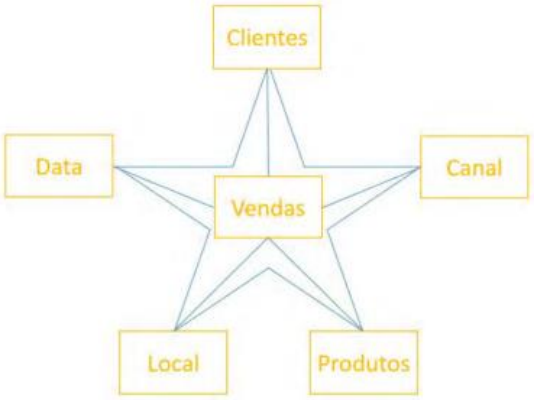

Figura 6: Modelo em Estrela

Fonte: Braghittoni *(2017, p. 63)*

Modelo Snow Flake

Para (Jensen, Pedersen, & Thomsen, 2010), “Um esquema *Snow Flake*, em português floco de neve possui uma tabela de facto, assim como um esquema em estrela. Os esquemas floco de neve, no entanto, contêm várias tabelas de dimensão para cada dimensão, nomeadamente uma tabela para cada nível”. As tabelas dimensão podem não estar directamente relacionadas a facto, podendo haver assim um relacionamento entre as tabelas dimensões.

Figura 7: Modelo Snow Flake

Fonte: Braghittoni *(2017, p. 63)*

III – Caso de Estudo

3.1. Cervejas de Moçambique

A Empresa CDM (Cervejas de Moçambique) é do ramo industrial fundada em agosto de 1995, resultando da privatização das fábricas 2M, (abreviatura de Mac-Mahon em homenagem ao antigo presidente francês Marie Esme Patrice Maurice, conde de Mac-Mahon fundada em 1962 na capital Maputo) e da fábrica Manica fundada em 1959 na Beira. No mesmo ano foi vendida ao grupo Sul-Africano SABMiller.

Conta actualmente com quatro (4) fábricas em todo país, duas (2) em Maputo no bairro do Jardim e no distrito de Marracuene a mais recente e moderna fábrica, uma na Beira e a quarta em Nampula. Conta com mais de 1000 colaboradores em todo país.

A CDM anualmente produz acima de 3 milhões de hectolitros anuais e distribui marcas locais e internacionais para todo país e além de fronteiras, para África do sul, Portugal e Inglaterra. Actualmente detém uma quota de 94% do controle do mercado Moçambicano de cervejas.

A CDM é Subsidiária da AB InBev a partir de outubro de 2016, o seu principal acionista com 51% das ações. Faz parte da Zona África do Business Unit (BU) da ABInBev com a Tanzânia, Uganda, Zâmbia, Gana e Botswana.

De acordo com Site oficial da CDM a AB InBev é a maior cervejeira do mundo, multinacional belga-brasileira criada em 2004 pela união de duas cervejeiras a Belga *Interbrew* e Brasileira *Ambev*. Conta com mais de 400 marcas de bebidas, entre as quais se destacam a Mexicana Corona, a Belga Stella Artois e a Norte Americana Budweiser.

Patrocina grandes eventos dos mais variados tipos, torneiros de futebol, ações de causa social. Conta com variedades de marcas nacionais a 2M, Impala, Laurentina, Manica, Dourada e internacionais Castle Lite, Budweiser, Flying Fish, Smirnoff, Brutal Fruit, Stella Artois e Corona.

No ano de 2011, a CDM introduziu a cerveja feita na base da mandioca a Impala, primeira cerveja no mundo produzida industrialmente com base na mandioca. Com o propósito de apoiar a agricultura do país, reduzir necessidade das importações e de trazer para o mercado uma cerveja mais barata. Para a produção da Impala

Mandioca, De acordo com Site oficial da CDM, os ingredientes são locais e eram obtidos em pequenos proprietários de plantações regionais, estes não conseguiam vender as suas plantações por conta de abundância. Em dezembro de 2017, a Impala lançou a variante de milho, produzida na base do milho.

A Empresa Cervejas de Moçambique divide-se em oito (8) departamentos, nomeadamente, Comercial, *Supply*, Logística, *People, IT & Solutions*, Finanças, Assuntos Legais e Corporativos e *Marketing*.

Abaixo segue-se a tabela da segmentação das marcas pelos tamanhos e tipo de embalagem, o termo "*ml*" é uma medida de capacidade que significa mililitros.

Marca	Tamanho da Embalagem	Tipo de Embalagem
2M	550ml	Garrafa Retornável
	250ml	Garrafa não retornável
	330ml	Lata
	750ml	Garrafa Retornável
2M Flow	250ml	Garrafa não retornável
	330ml	Lata
	490ml	Garrafa Retornável
Laurentina Preta	250ml	Garrafa não retornável
	550ml	Garrafa Retornável
	330ml	Lata
Manica	550ml	Garrafa Retornável
	330ml	Lata
Impala Mandioca	550ml	Garrafa Retornável
	330ml	Garrafa Retornável
Impala Milho	500ml	Garrafa Retornável
	330ml	Garrafa Retornável
	330ml	Lata
Dourada	500ml	Garrafa retornável
	330ml	Lata
Castle Lite	500ml	Garrafa Retornável
	340ml	Garrafa não retornável
	330ml	Lata
	250ml	Garrafa não retornável
	500ml	Garrafa Retornável
Brutal Fruit	275ml	Garrafa não retornável
	500ml	Lata
Smirnoff	300ml	Garrafa não retornável
Budweiser	330ml	Garrafa não retornável
Stella Artois	250ml	Garrafa não retornável
	330ml	Garrafa não retornável
Corona	210ml	Garrafa não retornável

	355ml	Garrafa não retornável
2M	50Lt	Barril
	30Lt	Barril
Laurentina Preta	30Lt	Barril
Caste Lite	30Lt	Barril

Tabela 1: Marca dos produtos, Tamanho e tipo de garrafa.

Fonte: Elaborado pelo Autor.

3.1.1. Marcos Históricos na CDM

A seguir são descritos os marcos importantes da CDM na cronologia do tempo de acordo com o site oficial.

- 2020 - É inaugurada a Cervejeira de Marracuene, a maior e mais moderna fábrica do país;
- 2019 - As cervejas 2M, Manica, Laurentina Preta e Impala Mandioca foram distinguidas pelo prémio de qualidade de “Ouro” e a Impala Milho, pelo prémio de qualidade de “Prata” pela Monde Selection;
- 2018 - Cerimónia de Lançamento da Primeira Pedra da Nova Fábrica da CDM em Marracuene;
- 2017 - Início da produção da Castle Lite em Moçambique;
- 2017 - Lançamento da Impala Milho, mundialmente, a primeira cerveja clara produzida usando o milho como matéria-prima principal;
- 2015 - A Laurentina Preta foi considerada a melhor cerveja preta de África, nos African Beer Awards;
- 2013 - A Laurentina Preta foi considerada a melhor cerveja preta de África, nos African Beer Awards;
- 2010 - Chibuku, marca de cerveja opaca, foi lançado em Moçambique;
- 2010 - A Cervejas de Moçambique inaugura a fábrica de Nampula, construída de raíz;
- 2009 - A Laurentina Premium ganhou a medalha Grand Gold pela sua qualidade no concurso internacional Monde Selection, em Bruxelas;
- 2008 - A Laurentina Preta foi reconhecida com Medalha de Ouro no concurso internacional de qualidade Monde Selection;
- 2008 - Foi lançada a Laurentina Premium, uma cerveja especial, moderna e sofisticada, feita de 100% malte;

- 2002 - A CDM adquire a Laurentina Cervejas e passa a produzir as marcas Laurentina Clara e Lautentina Preta;
- 2001- A CDM torna-se a primeira empresa moçambicana cotada na Bolsa de Valores de Moçambique;
- 1995 - Nasce a Cervejas de Moçambique, SARL, resultado da privatização das fábricas de Cerveja Mac-Mahon e Manica, localizadas, respectivamente, em Maputo e na Beira;
- 1980 - O Estado Moçambicano nacionalizou a Sogere;
- 1955 - Foi instalada a Fábrica Manica, na Beira;
- 1950 - Foi inaugurada a fábrica Mac Mahon em Maputo, cujo nome deu origem também ao nome da mais popular cerveja de Moçambique, a 2M;
- 1932 - A Laurentina Clara, primeira cerveja de Moçambique, foi lançada por um imigrante grego chamado Cretikos, que fundou a fábrica Vitória;

3.1.2. Estrutura Organizacional

O departamento comercial é composto pelas seguintes áreas: Vendas, RTM *(Route to Market)*, *Trade Market* e *Revenue*. A seguir é ilustrada a estrutura organizacional.

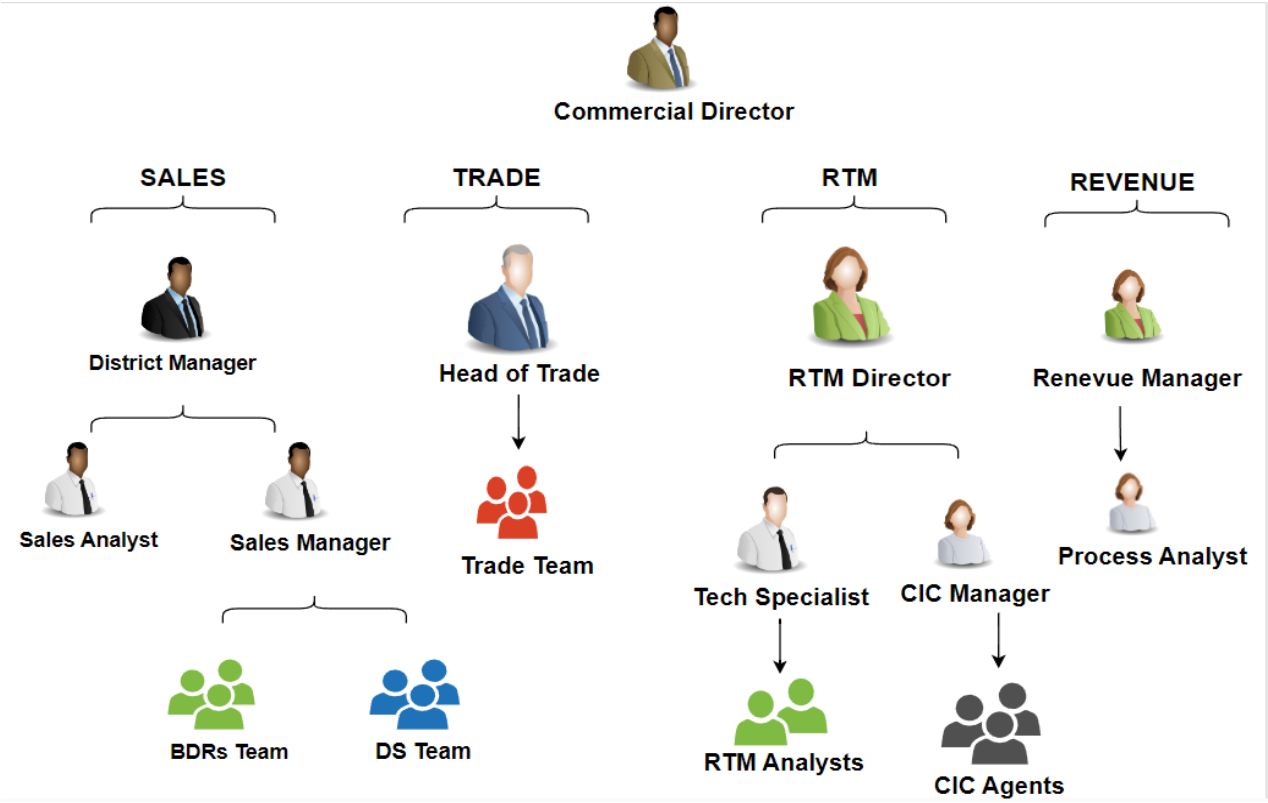

Figura 8: Estrutura Hierárquica do Departamento Comercial.

Fonte: Elaborado pelo autor.

3.1.3. Descrição da Situação atual

Neste trabalho vamos nos focar no Departamento comercial em particular área de vendas. Em uma sala enorme está uma equipe com cerca de trinta (30) pessoas designadas CIC (Centro de Interação com Cliente), dividida em dois (2) grupos, uma está virada ao *SellIn* e o outro no *SellOut*. A equipe de *SellIn* é responsável por processar as encomendas dos pedidos feitos pelos distribuidores, ou seja, os agentes CIC da equipe do *SellIn* recebem encomendas vindo dos armazenistas e registram no sistema designado *Syspro*. Por outro lado, a equipe de *SellOut* dedica-se nas vendas directas aos POCs (Ponto de Consumo), ou seja, ao cliente final, que pode ser uma barraca, discoteca, restaurante e outros. esta equipe com o auxílio dos BDRs (os representantes da CDM no mercado), efectua 30 chamadas telefônicas diárias em média por cada agente com a finalidade de gerar encomendas nos POCs. Cada agente da equipe do *SellOut* tem uma base de clientes os quais liga diariamente para 30 clientes agendados por dia de semana com o intensão de influenciar os POCs a comprar no armazém mais próximo.

Esses POCs são visitados uma vez por semana pela equipe dos BDRs, esses são responsáveis por criar a ligação entre os agentes CIC e os POCs. Essas visitas incluem como outras atribuições dos BDRs a colagem dos *posters*, fornecer materiais (geleiras, mesas, panos, sombreiros, copos, abre-garrafas, bandejas), garantir o cumprimento do preço recomendado, verificar a disponibilidade dos produtos e outras dificuldades que o cliente possa reportar.

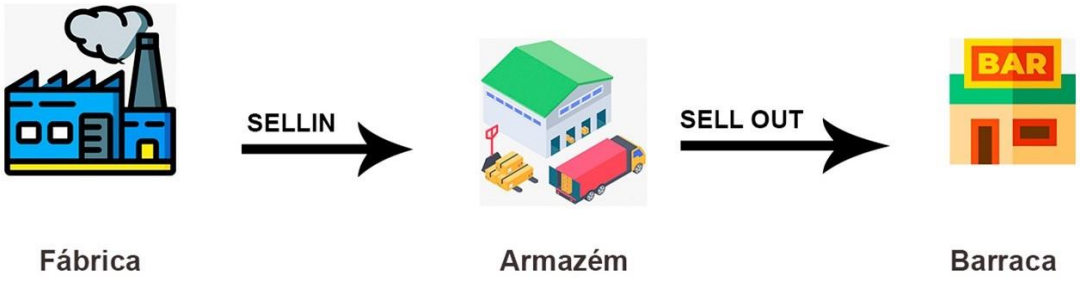

Figura 9: Processo de Sellin e Sellout

Fonte: Elaborado pelo autor.

Para a assistência aos armazéns existe uma equipe chamada DS (*Distributor Specialist*) em português Especialistas de Distribuição, assim como há uma cooperação entre a equipe do *SellOut* e os BDRs também há uma cooperação entre os DS e a equipe do *SellIn*, essa equipe negoceia e projecta as encomendas que os armazéns fazem, são também atribuições dos DS garantir que o armazém tenha um stock suficiente para cobrir maior número de dias, gerenciar os pagamentos das encomendas e reportar os níveis de *stocks* no armazém.

Nos armazéns parceiros da CDM, para gestão de vendas foi instalado um sistema designado KUJA pela equipe de soluções e IT da CDM, para o registro das vendas que os armazéns realizam aos POCs, neste caso o *SellOut*, são também registrados os níveis de *stock* do armazém.

A equipe das vendas trabalha orientado a metas, são definidas metas mensais aos armazéns, aos BDRs, aos DS, aos agentes CIC, aos *Sales Managers*, aos *District Managers*.

As metas mensais são definidas tanto para *SellIn* e para *SellOut* para cada uma das equipes, assim como regionais (metas que cada região deve atingir) e nacional (meta geral do país), são também definidos por marca e *SKU*. Um Mapeamento de definição e atribuição das metas é elaborado mensalmente.

Da situação descrita acima, podemos perceber que as vendas são registadas em vários sistemas, no caso de *Sellin* usa o *Syspro*, para *SellOut* registadas no KUJA e as metas mensais são definidas em planilhas de Excel de forma isolada.

3.1.4. Constrangimentos Identificados

Para fazer análises das vendas é necessário usar dados provenientes dos diferentes sistemas de vendas anteriormente descritos, cruzar esses dados para gerar *insights* do negócio. Neste cenário verificam-se os seguintes constrangimentos:

- Dados são provenientes de diferentes sistemas;
- Os dados das vendas não estão centralizados;
- Leva-se bastante tempo para a responder as perguntas do negócio, ou seja, as perguntas do negócio não são respondidas a tempo real;
- Tarefas Manuais e repetitivas para obtenção e consolidação de dados para análise;

Capítulo IV – Proposta de Solução

4.1. Descrição das Soluções de ETL

Para avaliar as possíveis alternativas de solução ir-se-à realizar uma análise comparativa das ferramentas de ETL. Já referenciado ETL é um conceito e que a sua implementação pode ser mediante várias ferramentas, destas são listadas quatro para a comparação, são eles o Oracle Data Integrator, Talend Open Studio, Microsoft SQL Server Integration Services e IBM DataStage. A seleção da solução ideal para o caso em estudo será mediante a ferramenta que conseguir maior pontuação durante a análise comparativa.

4.1.1. Oracle Data Integrator

O *Oracle Data Integrator* é uma ferramenta *on-premise* da Oracle. Permite a integração de dados, oferece recursos avançados de ETL de dados permitindo que as organizações movam dados entre sistemas diferentes. Florea *et al.* (2015) afirmam que uma das principais características do ODI é poder realizar transformações complexas tanto do lado da origem quanto no lado do destino, e a maioria dessas transformações ocorre em lote quando não há consultas do usuário a serem executadas pelo servidor.

Existem cinco principais componentes da plataforma ODI: Repositório, ODI Studio, Agente, Console e *Oracle Enterprise Manager*.

i. Repositório

É o elemento central da arquitectura do ODI e representa o principal local de armazenamento onde é gravada toda informação que o ODI executa nomeadamente, metadados, detalhes de conexão, regras e cenários de transformação, registo de actividade e estatísticas e o código gerado.

Os repositórios são divididos em dois grupos: Um repositório mestre onde hospeda dados confidenciais (informações de segurança e topologia) e repositórios de trabalho onde residem dados relativos ao projecto.

ii. ODI Studio

O ODI Studio representa a interface gráfica da plataforma e oferece acesso aos repositórios para que os utiliza, geralmente por administradores, desenvolvedores.

Segundo Florea *et al.* (2015) o ODI Studio é usado para administrar infraestrutura, fazer a engenharia reversa dos metadados, desenvolver projectos, programar e monitorar execuções.

Ainda os mesmos autores o Studio está organizado em quatro navegadores diferentes que são utilizados por vários usuários de acordo com as funções e perfis de segurança. São descritos em seguida:

Navegador de segurança – usado por administradores de sistemas e DBAs para gerenciar funções e privilégios dos usuários regulares.

Navegador de Topologia – normalmente usados pelos administradores de sistemas e DBAs para definir as conexões e credenciais necessárias para que o ODI se conecte aos sistemas origem e destino.

Navegador de Design – é a parte central do mundo do desenvolvedor, sendo utilizado para definir as transformações necessárias em interfaces.

Navegador de Operador – é o componente do Studio que garante a gestão e acompanhamento das actividades.

iii. Agente

O Agente é o componente responsável pela execução física dos processos de integração de dados. Ele e instalado em servidores remotos ou em máquinas locais executando as tarefas a mando do ODI Studio. Os agentes são responsáveis na movimentação dos dados durante as operações de ETL.

iv. Console

O Console é uma interface web que oferece a possibilidade de monitoramento e administração para os processos de integração de dados. Permite que os administradores gerenciem operações administrativas e monitorem as tarefas e o desempenho das operações.

v. Oracle Interprise Manager

O *Interprise Manager* ou *Topology Manager* é utilizado para configurar e gerenciar as conexões e topologias necessárias para o ODI. Permitindo definir com detalhes as

conexões com as fontes e destino, assim como configurar o ambiente para garantir uma eficiente execução dos processos de integração.

Estes componentes trabalham em conjunto para proporcionar uma solução robusta e escalável no ODI, ajudando as organizações a movimentarem dados de maneira a garantir qualidade e consistência de dados.

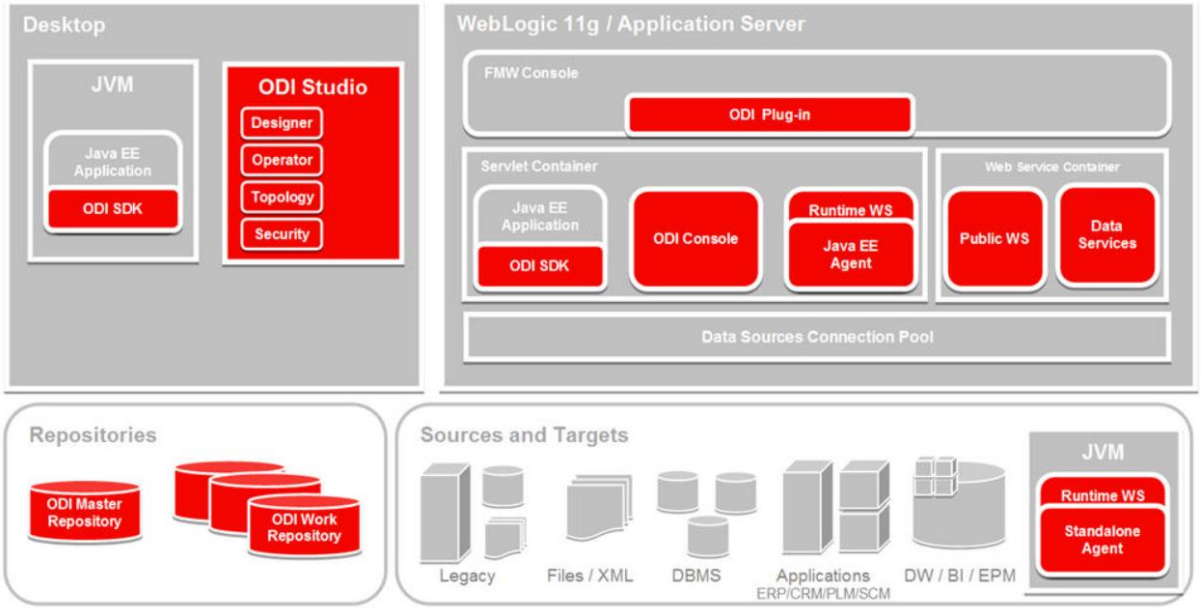

Figura 10: Arquitetura do Oracle Data Integrator

Fonte: https://docs.oracle.com/cd/E17904_01/integrate.1111/e12641/overview.htm#ODIGS421

4.1.2. Talend Open Studio

Kumar *et al.* (2019) afirmam que o Talend OS e o primeiro software de integração de dados de código aberto lançado em 2006 pela Talend. Ele é baseado no Eclipse que oferece suporte a implementação de ETL para implantação *on-premise* e modelos de software como serviço.

O Talend Open Studio é utilizado para integração de sistemas transacionais, assim como uma ferramenta de ETL para processamento de dados, *Business Intelligence*, *Data Warehouse* e movimentar dados.

A seguir são descritas as características do Talend Open Studio segundo Kumar *et al.* (2019):

- Implementado com sucesso à aplicação no mundo real;
- Sincroniza dados entre fontes heterogêneas e destinos;
- Fácil de usar, apresenta uma interface intuitiva rica em ferramentas de modelagem;
- O ambiente de desenvolvimento e amigável e abrangente;
- Apresenta mais de 450 conectores de dados incluindo em nuvem;
- Pode gerar código a partir dos pacotes desenvolvidos;
- O código gerado pode ser modificado para alcançar as necessidades de maior controle e flexibilidade;
- Para a integração de dados o Talend Open Studio é gratuito para baixar e usar;

Na versão comunitária desta ferramenta há certas limitações, pois ele foi desenvolvido para o uso individual e por conta disso não é possível ter mais de um usuário (não em simultâneo, mas um usuário por sistema). Deste modo cria sérios problemas de implementação pois pode ser necessário que vários usuários usem o mesmo computador em diferentes momentos. A versão gratuita não suporta a automação de tarefas como agendamento, roteamento de dados etc. Carece ainda de suporte comercial.

A arquitectura funcional do Talend OS é dividida em cinco blocos funcionais: bloco de clientes, bloco de servidor, bloco de repositórios, base de dados e execução de servidores. Esta é comparada ao canivete suíço por conta da sua disposição.

A seguir é ilustrada a arquitectura funcional de integração dos dados Talend.

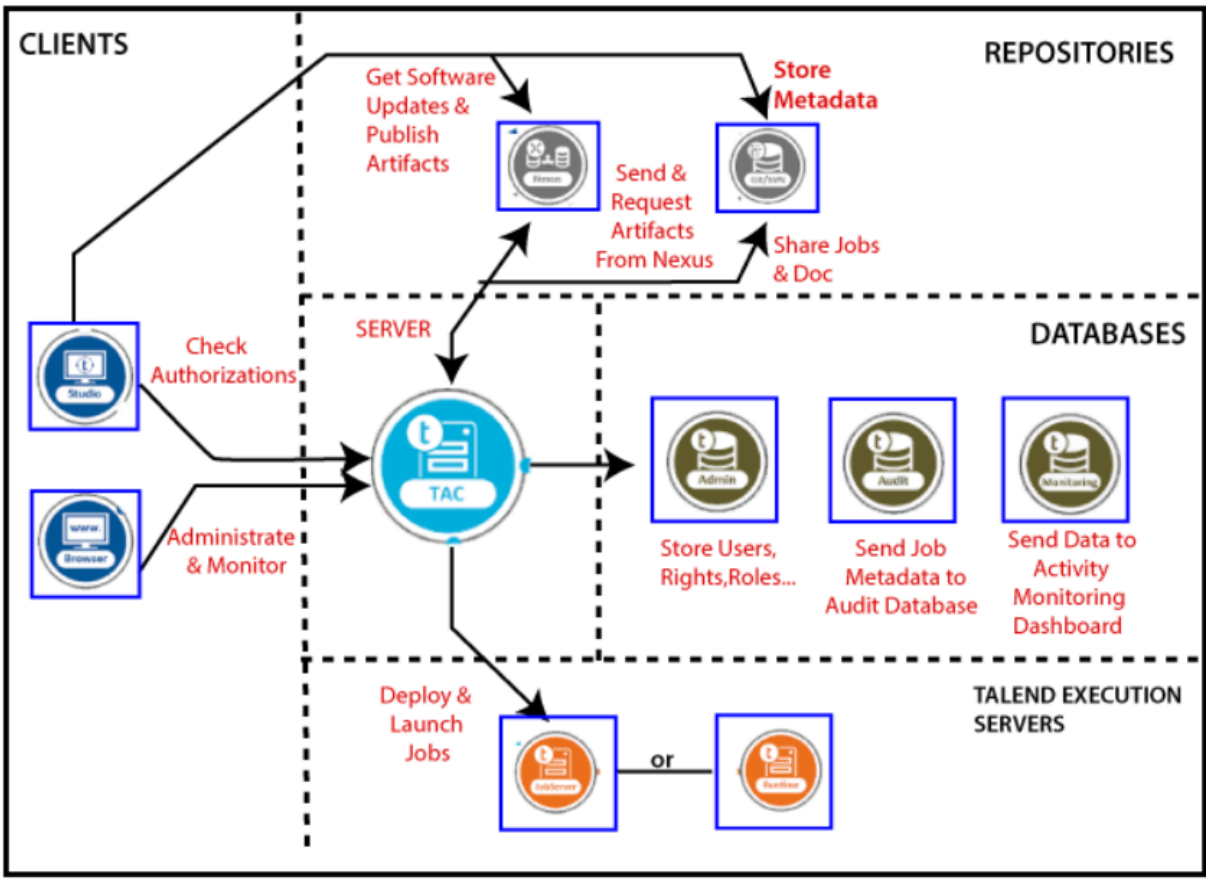

Figura 11: Arquitetura funcional do Talend

Fonte: Kumar *et al.* (2019)

Bloco Clientes

Este bloco inclui navegadores web que podem estar na mesma máquina ou em máquinas diferentes. O Talend Open Studio permite que você trabalhe em qualquer projeto apenas para processos autorizados. No navegador, você precisa se conectar remotamente ao Centro de Administração Talend de forma remota por meio do protocolo HTTPs (Kumar *et al*. 2019).

Bloco do Servidor

Neste bloco e onde se encontram o servidor de aplicações web e o centro de administração Talend, que permitem a administração e gerenciamento de todos os projetos.

Base de dados

A base de dados é usada para armazenar os metadados e configuração de informação como: auditoria, monitoria e administração de contas de usuários.

Bloco de Repositórios

Composto pelo servidor SVN (Sistema de versionamento) e o repositório de dados. O servidor SVN é usado para organizar todos os itens de projeto, como trabalhos e modelos de negócio compartilhado entre diferentes usuários (Kumar *et al*. 2019). O repositório é usado para armazenar dados.

Bloco de Execução de servidores

Segundo Kumar *et al.* (2019) os blocos de execução de servidores incluem um ou mais servidores de execução que são configurados dentro do sistema de informação.

4.1.3. IBM DataStage

O DataStage é uma plataforma de integração de dados da IBM projetada para simplificar o processo de extração, transformação e carga de dados (ETL) entre sistemas diferentes. A principal função é permitir que as organizações transformem e movimentem dados de forma confiável e eficiente, suportando o processamento em grande escala.

O IBM DataStage é muito utilizado em ambientes corporativos para as tarefas de integração, ETL e movimentação entre sistemas heterogêneos. É caracterizado por sua escalabilidade, flexibilidade e robustez na manipulação de dados mais complexos e em ambientes diferentes de armazenamento.

Os principais componentes do IBM DataStage são: Design Center, Repositório, Director, DataStage Administrator e Palette.

I. **Design Center**

Consiste em um ambiente gráfico que possibilita aos programadores projectarem, criarem e definirem os processos de ETL. Usa uma abordagem de desenvolvimento visual, permitindo que os desenvolvedores criem fluxos de dados usando uma interface gráfica.

II. Repositório

Armazena metadados do projeto, tarefas e objectos do *DataStage*. Os metadados contém informações sobre as fontes, transformações e destino de dados e outros elementos integrantes usados no processo de ETL.

III. Director

É o elemento responsável por executar as tarefas iniciadas no Design center. Gerencia as execuções simultâneas de ETL e fornece recursos para monitoramento das tarefas em tempo real.

IV. DataStage Administrator

Proporciona funções administrativas para configurar e gerenciar todos os recursos de um ambiente *DataStage*, ele é o motor de todo funcionamento. Desde a gestão de conexões a base de dados, configuração de servidores e gerenciamento de usuários.

V. Palette

É um conjunto de operadores e funções criadas no *DataStage* que possibilita aos desenvolvedores poderem arrastar e soltar no Design Center para construir os fluxos de trabalho. Esta funcionalidade facilita a criação rápida e eficiente de transformações de dados muito complexas.

Segundo Cetax (2022) o *DataStage* fornece os seguintes recursos e benefícios:

Plataforma poderosa e escalável – Suporta a colecta, integração e transformação de grandes volumes de dados, com estruturas de dados variando de simples a complexas.

Suporte a Big Data e Hadoop – Permite o acesso direto ao Big Data em um sistema de arquivos distribuídos.

Integração de dados a tempo real - Permite uma conexão entre as origens dos dados e aplicativos dos usuários.

Gerenciamento de regras de carga de trabalho e negócio – Ajuda a optimizar a utilização de recursos de hardware e priorizar as tarefas mais essenciais.

Facilidade de uso – Ajuda a melhorar a velocidade, flexibilidade e efetividade para implementação, atualização e gestão da infraestrutura de integração de dados.

4.1.4. SQL Server Integration Service - SSIS

O SSIS é uma plataforma de extração, transformação e carga (ETL) da Microsoft lançada em 2005 em conjunto com o SQL Server nesta versão. Permite que os desenvolvedores criem e executem processos de integração de dados, transformando e movendo dados de uma origem a um destino. O SSIS suporta a integração com outras ferramentas da Microsoft nomeadamente *SQL Server Analysis Services (SSAS)* e *SQL Server Reporting Services (SSRS)* que são muitas das vezes usadas como destinos de um processo ETL para soluções de *Business Intelligence*.

Permite que você consiga se conectar a uma variedade de fontes de dados incluindo, base de dados SQL Server, Excel, CSV, entre outros.

Segundo Lisboa (2023) com o SSIS, é possível automatizar tarefas repetitivas e demoradas com o ETL. Isso pode economizar tempo e recursos, permitindo que os profissionais se foquem mais em actividades táticas e estratégicas.

São principais componentes do SISS os seguintes: *Data Flow*, *Package Explorer, Control Flow*, *Connection Managers*, *Variables* e *Event Handlers*.

A. Data Flow

Knight *et al.* (2008) afirmam que a maior parte do tempo levado no SSIS é gasto na guia de fluxo de dados. Quando uma tarefa é criada no *control flow* subsequentemente é criado também um fluxo de dados neste guia.

O Data Flow é o componente central para definir as operações ETL, permite a extração, transformação e carga de dados entre fontes de destinos.

B. Control Flow

Este componente controla o fluxo e a lógica de execução dos pacotes SSIS, incluindo tarefas condicionais, loops e execução paralela. Com o *control flow* pode se definir restrições de precedências, em um caso de sequenciamento de tarefas permite definir se a próxima tarefa execute mesmo que anterior falhe, podendo assim garantir a confiabilidade dos dados.

C. Package Explorer

Fornece uma visão única de todos painéis e componentes de um pacote SSIS.

D. Variables

São variáveis, Segundo Knight *et al.* (2008) variáveis são uma peça poderosa da arquitectura do SSIS. Eles permitem que você controle dinamicamente o pacote em tempo de execução, existem dois tipos de variáveis: de sistema e usuário.

- **Variáveis de Sistema** - são aquelas incorporadas ao SSIS como nome de um pacote.
- **Variáveis de Usuário -** são aquelas criadas pelo desenvolvedor do SSIS. As variáveis podem ter escopos variados, nomeadamente escopo padrão e inteiro.

E. Connection Manager

O gerenciador de conexões contém uma lista de conexões que as tarefas de controle de fluxo e fluxo de dados podem usar. Quer seja uma conexão com um endereço FTP ou uma conexão com um servidor do *Analysis Services* (Knight *et al*, 2008).

Basicamente o gerenciador configura a conexão com fontes e destino de dados.

F. Event Handlers

O Manipulador de eventos permite criar fluxos de trabalhos para lidar com erros, avisos ou finalização de tarefas e pacotes. Permite responder à eventos específicos durante a execução de um pacote.

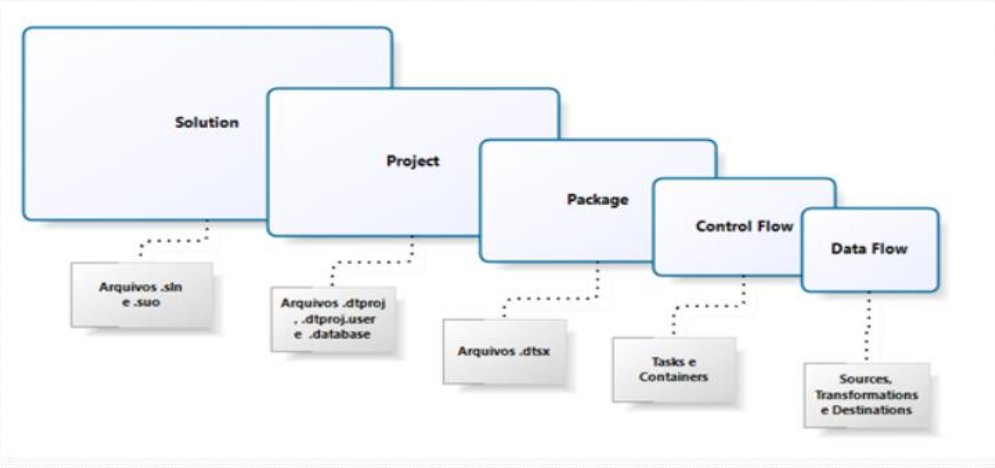

Figura 12: Hierarquia dos Componentes SSIS

Fonte: https://www.devmedia.com.br/microsoft-etl-arquitetura-ssis-sql-server-integration-services/30862

Da figura acima pode se notar que o recipiente solução é o elemento macro da hierarquia, nela são gerenciadas um ou mais projetos, tem a finalidade de abarcar todos os projetos para a visão do negócio. O recipiente projeto é o local onde efetivamente se desenvolvem os pacotes, definem-se aqui as fontes e destinos dos dados. O recipiente pacotes é onde se desenvolvem os fluxos, sendo de controle ou de dados. Os Fluxos de controle são os principais componentes de um pacote, pois é onde se definem os pacotes e as tarefas.

Já definido anteriormente o recipiente fluxo de dados é o componente central, define o processo de ETL.

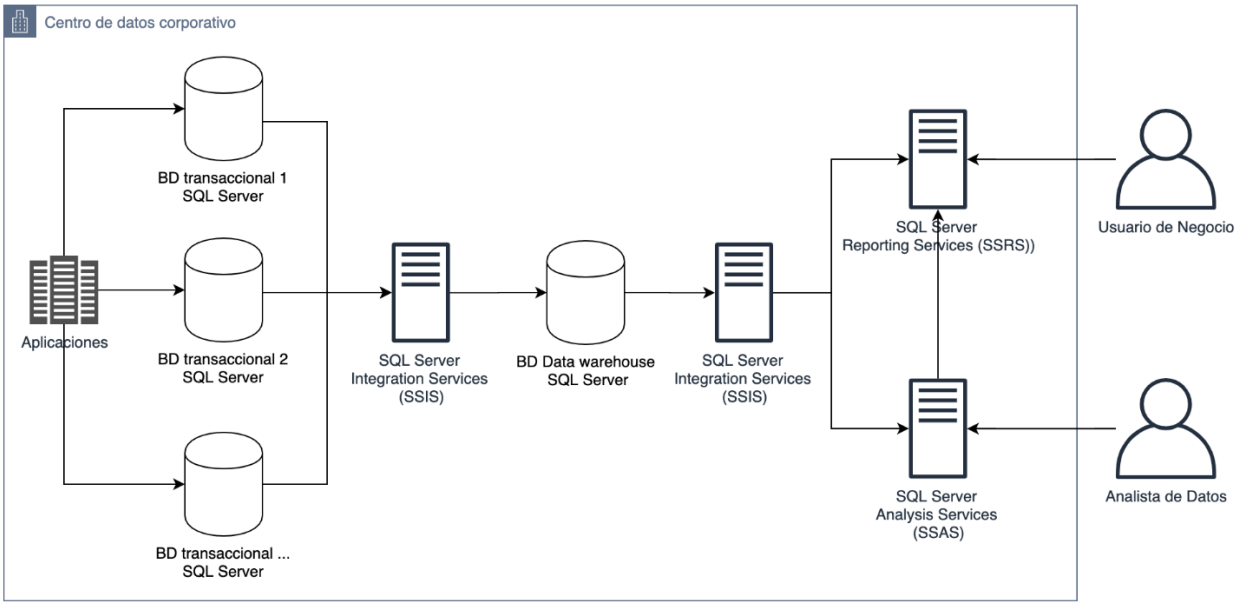

Figura 13: Arquitetura de *Business Intelligence* com SSIS como um elemento central

Fonte: https://aws.amazon.com/pt/blogs/aws-brasil/executando-workloads-de-microsoft-business-intelligence-na-nuvem-aws/

O SSIS é essencialmente desenvolvido para finalidade de impulsionar o negócio em uma organização, na figura 12 é um ilustrada uma arquitectura completa de *Business Intelligence* onde o SSIS é o elemento central pois ele é responsável por extrair os dados nas distintas fontes, as várias aplicações na organização geram dados que são armazenados nas bases de dados e outras formas de armazenamento o SSIS faz a extração, transformação e carrega os dados para um *Data Warehouse*. Daqui os dados podem ser usados pelas ferramentas que consomem dados, o SSRS e SSAS para gerar relatórios e análises pelos usuários de negócio e analistas respectivamente.

4.2. Análise comparativa das ferramentas de Solução

Para analisar as soluções ETL, far-se-à uma comparação das principais ferramentas ETL considerando os critérios relevantes para alcançar o objectivo: Implementação, Infraestrutura, Escalabilidade, Flexibilidade, Código aberto, Custo, Segurança, Facilidade de uso e Desempenho.

A seguir irá-se descrever cada um dos critérios para que não restem dúvidas que do estamos avaliando.

Implementação – Este critério diz respeito a modo de implementação. Podendo assumir valores com GUI (interface gráfica) ou desenvolvendo código (programação).

Infraestrutura – Define onde a solução desenvolvida é disponibilizada, assume dois valores possíveis, nuvem e *on-premise* (local).

Escalabilidade - Avalia a capacidade da ferramenta de escalar de acordo com seus volumes de dados e crescimento de negócios. Pode assumir valores como baixo, médio e alto.

Flexibilidade - Garante que a ferramenta ETL permita flexibilidade e personalização para atender aos seus requisitos de negócios específicos. Pode assumir valores como baixo, médio e alto.

Código aberto - Avalia a atribuição da licença de uso, podendo então ser pago ou de código aberto significa que você pode acessar, modificar e distribuir o código-fonte. Assume apenas dois valores, sim ou não.

Custo - Avalia o custo geral de propriedade, incluindo taxas de licenciamento, manutenção e quaisquer despesas adicionais. Assume dois valores, pago ou gratuito.

Segurança - A segurança dos dados é fundamental. Avalia se a ferramenta ETL fornece recursos de segurança robustos para proteger informações confidenciais durante os processos de extração, transformação e carregamento. Pode assumir três valores nomeadamente, baixo, médio e alto.

Facilidade de uso - Avalia se a ferramenta apresenta uma interface de usuário intuitiva e baixa curva de aprendizado. Uma ferramenta ETL fácil de usar pode aumentar a produtividade e reduzir o tempo necessário para desenvolvimento.

Desempenho - Avalia o desempenho da ferramenta em termos de velocidade e eficiência. Deve se considerar a rapidez com que ele pode processar e transferir dados, especialmente ao lidar com grandes volumes de dados. Este critério pode assumir três valores, baixo, médio e alto.

Processamento paralelo – é a capacidade de poder executar várias tarefas em simultâneo. Este critério pode assumir três valores, que são, baixo, médio e alto.

Critérios	Oracle Data Integrator	Talend Open Studio	IBM DataStage	SSIS
Implementação	GUI	GUI	GUI	GUI
Infraestrutura	Nuvem/Local	Local/Nuvem	Nuvem/Local	Nuvem/Local
Escalabilidade	Média	Média	Média	Média
Flexibilidade	Alta	Alta	Alta	Alta
Código aberto	Não	Sim	Não	Não
Custo	Pago	Gratuito	Pago	Pago
Segurança	Alta	Média	Alta	Alta
Facilidade de uso	Média	Alta	Alta	Alta
Desempenho	Alto	Médio	Alto	Alto
Processamento Paralelo	Alto	Alto	Alto	Alto

Tabela 2: Análise comparativa das soluções ETL.

Fonte: Elaborado pelo Autor.

Com a comparação feita, agora irá-se avaliar de acordo com os critérios a solução ideal para alcançar o objectivo. Agora vai se apurar os resultados de acordo com as cores na tabela 3. A cor verde corresponde a Bom, amarela a normal e o vermelho Mau.

Ferramentas ETL	Mau	Normal	Bom
ODI	2	2	6
Talend OS	0	3	7
IBM DS	2	1	7
SSIS	2	1	7

Tabela 3: Resumo da análise comparativa.

Fonte: Adaptado de Massunguine (2022).

Para decidir a solução ideal, vamos considerar a coluna “BOM” na tabela 4. Verifica-se um empate em três ferramentas, para desempatar vamos reconsiderar a avaliação de dois critérios, segurança e custo. A segurança é o elemento crucial quando lhe damos com dados, sendo assim vamos desqualificar o Talend Open Studio por apresentar um nível de segurança “Médio” e ficamos apenas com o IBM DataStage e o SSIS para o desempate dessas ferramentas vamos reconsiderar o critério custo, ambas soluções são pagas, porém o SSIS aparece como oferta quando é paga a licença do SQL Server e pelo ecossistema da CDM que é da Microsoft, bases de dados são em SQL Server, Excel, *PowerBI*. Porque a CDM usa produtos da *Microsoft* e sendo assim não será necessário pagar poder usar o SSIS. De acordo com o desempate estabelecido a ferramenta ideal para o processo de ETL para integrar os dados das vendas na CDM é o *SQL Server Integration Services* (SSIS).

4.3. Desenvolvimento da solução proposta.

4.3.1. Descrição do cenário proposto para implementação da solução

Para colocar em prática a solução ETL (SSIS) para integração de dados das vendas na CDM, projectou-se a arquitectura da figura 13, que melhor descreve o fluxo de dados desde a sua extração até o usuário final.

Para este efeito será necessário um servidor, onde será instalado a ferramenta Visual Studio para criar e gerenciar os projetos SSIS, definir uma pasta no ambiente em nuvem da Microsoft, o *OneDrive* onde os *stakeholders* tem acesso uma pasta, onde serão armazenados os dados das vendas consolidados.

Na arquitetura proposta temos 4 estágios, o primeiro ilustra as fontes que serão usadas para extrair dados na sua forma bruta, no segundo temos a ferramenta Visual Studio que vai criar e gerenciar os processos de ETL, o terceiro estágio define o local onde os dados são armazenados após o processo de extração e ou transformação, no último estágio verifica-se o uso dos dados, na geração de relatórios e criação de *Dashboards.*

As fontes de dados usadas para este processo são das vendas, os metadados das bases de dados e do arquivo Excel não serão revelados por motivos confidenciais.

Nas bases de dados são armazenados os dados das vendas designadas *Sellin*, vendas directas da fábrica ao distribuidor (armazém) e no arquivo Excel temos

armazenados as vendas designadas *SellOut*, vendas a partir do armazém para uma barraca.

Na ferramenta Visual Studio são criados e gerenciados fluxos de dados e controle de fluxos. Esta ferramenta é configurada de forma a extrair dados periodicamente nas fontes, extraindo dados de *SellIn* da base de dados transacional para um arquivo excel disponivel na pasta do *OndeDrive*, e fazer o carregamento diário dos dados do *SellOut* para a base de dados onde de imediato são extraídos para um arquivo Excel na mesma pasta onde são gravados os dados de *SellIn.* Um servidor FTP baixa dados diários referentes ao dia anterior de *SellOut* vindos do KUJA no formato ".csv" para uma pasta gerenciada pelos Administradores de Aplicações, um *Job* será configurado no SQL Server para carregar os dados pelo SSIS, armazená-los na base de dados.

Para o armazenamento dos dados, será definida uma pasta no *OneDrive* da rede da área de vendas. Consta desta pasta dois (2) arquivos Excel, um com dados provenientes da base de dados de *SellIn* e outro arquivo Excel com dados de *SellOu*t provenintes da base de dados. A partir deste ponto os dados são úteis para disponibilizar aos DS, DBRs para poder ter a visibilidade diária das suas vendas e performance de outros KPIs, para os analistas criarem relatórios e Dashboards, possibilitando assim que Gestores e Directores tomem decisões baseadas em dados.

A escolha na forma de armazenamento em um arquivo Excel é motivada pelo facto de que os, DS, BDRs, Gestores e Directores tem maior familiaridade com a ferramenta. Foi escolhida também por ser uma ferramenta usada para armazenar dados, criar planilhas, realizar cálculos, criar gráficos e análise de dados, além de ser a fonte nativa da moderna ferramenta de análise de dados usada na CDM, o *PowerBI*.

Como forma de garantir a confiabilidade e segurança dos dados armazenados, foram definidas permissões de acesso nos arquivos Excel na pasta destino, as permissões de editar e apagar foram removidas para todos os usuários com acesso ao ficheiro, ficando apenas as permissões para visualizar e baixar a planilha.

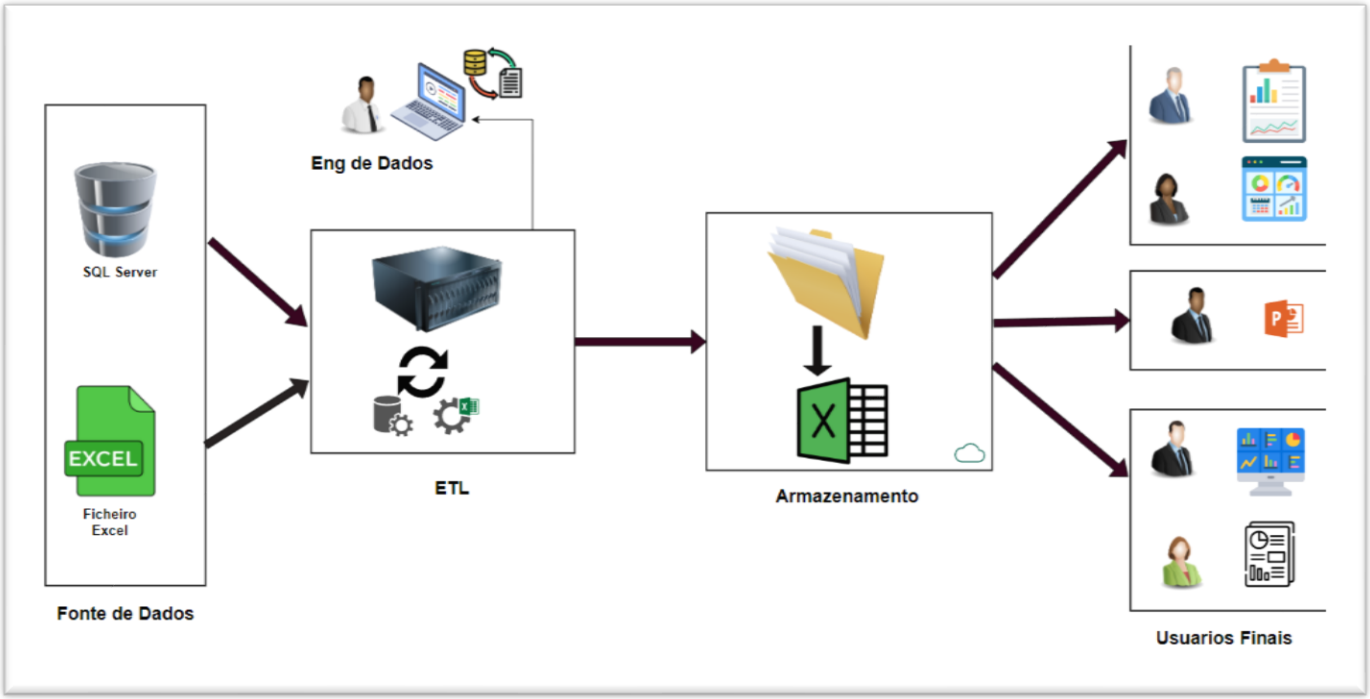

Figura 14: Arquitetura do cenário proposto para implementação da Solução

Fonte: Elaborado pelo Autor.

Capítulo V – Discussão de Resultados

5.1. Revisão de Literatura

As empresas orientam seus negócios com base nos dados, a tomada de decisões é apoiada sobre os dados gerados diariamente pelos diferentes sistemas transacionais de uma empresa. Para empresas que já trabalham com o aprendizado de máquina já tem uma noção de quão crucial é um dado confiável. Um processo que extrai, transforma e centraliza esses dados gerados pelos diferentes sistemas é indispensável para a geração de relatórios e criação de relatórios para o acampanhamento diário do desempenho das actividades duma empresa para conduzir ações do nível operacional. Dados errados, conduzem à tomada de decisões erradas então é de especial atenção a presença de profissionais qualificados para criar e gerenciar os processos de ETL.

Densmore traz a seguinte abordagem sobre *pipeline* de dados, "*Pipelines* de dados são conjuntos de processos que se movem e transformam dados de várias fontes para um destino onde o novo valor pode ser derivado. Eles são a base de análises, relatórios e aprendizagem de máquina". Para Munappy e outros afirmam que, "*Pipelines* de dados são a cadeia conectada de processos em que a saída de um ou mais processos se torna uma entrada para outro. É um software que remove muitas etapas manuais do fluxo de trabalho e permite um fluxo simplificado e automatizado de dados de um nó para outro". Estas duas abordagens se complementam, pois, um *pipeline* de dados é projectado para mover dados de um ponto de origem a outro, destino, passando ou não por transformações, o objectivo é de simplificar o fluxo de trabalho deixando tarefas automáticas e mais eficientes.

O termo "*Business Intelligence*" vem se tornando mais popular nas empresas, porém este conceito existe a bastante tempo. O intuíto da implementação do BI é de poder impulsionar os negócios, melhorando os processos de tomada decisões usando das ferramentas computacionais.

O processo de ETL existe de duas abordagens, o habitual ETL, que segue as etapas convencionais de extração, transformação e carga de dados, a outra abordagem é o ELT que tem mesmo propósito, porém esta após a extração armazena os dados permanentemente e só depois os manipula, ou seja, transforma-os. Após o tratamento dos dados eles são armazenados em um repositório central, o *Data*

Warehouse. A afirmação do Oliveria baseada na ideia do Watson e Haley, “Sistemas de *Data Warehouse* possibilitam que as organizações tenham acesso à informação de gestão que é determinante para obterem ganhos significativos, nomeadamente, aumento de vendas, redução nos custos, oferta de novos serviços e produtos”. Mas o Jensen e Thomsen, defendem que um *Data Warehouse* é um repositório de dados corporativos integrados, usado especificamente para suporte à decisão em uma empresa. Um *Data Warehouse* normalmente contém dados colectados de muitas fontes dentro e às vezes também fora da empresa. As duas contribuições são válidas e consumam-se. Um DW é projectado para atender as perguntas do negócio e dirige à tomada de decisões, com objectivo de maximizar lucros, minimizar os custos de produção e visão de novas oportunidades.

5.2. Caso de estudo

As Cervejas de Moçambique é uma empresa de produção de cerveja, conta com quatro (4) fabricas em todo país, duas (2) em Maputo, uma (1) na Beira e uma (1) em Nampula. Conta com várias marcas de cerveja, distribui a nível nacional e internacional, comercializa marcas estrangeiras. E formado por vários departamentos, Comercial, IT, Logística, Assuntos Legais e Corporativos, *People*, *Marketing* e *Supply*. O foco deste trabalho é no departamento comercial para área das vendas em específico, que é constituída por quatro (4) *Distric Managers*, oito (8) *Sales Managers,* três (3) equipes de vendas, divididas em: mais de oitenta (80) representantes de área (BDRs), mais de vinte (20) agentes de vendas (CIC) e desaseis (16) Especialistas de Distribuição (DS). As vendas são divididas em dois (2) tipos: *SellIn* e *SellOut*, o tipo *SellIn* é venda directa da fábrica até o armazém onde os especialistas e os clientes (armazenistas) são os intervenientes e o *SellOut* é a venda indireta que é feita do armazém até o ponto de consumo (POC). Os intervenientes do *SellOut* são os representantes de área local, agentes CIC e o armazenista. Essas equipes trabalham em conjunto para o atingimento das metas, os dados das vendas estão disponíveis em várias fontes, KUJA, *Syspro*. Há uma necessidade de extrair esses dados para as análises e relatórios. Pelo que, se verificam os constrangimentos: Dados desorganizados, inconsistências nos dados, custos elevados de manutenção (especialmente em termos de tempo e recursos necessários para gerenciar manualmente os dados) e falta da actualização em tempo real.

5.3. Proposta de solução

Verificou-se neste capítulo a importância de um processo de ETL para as organizações. Um processo de ETL é indispensável para a análise de dados como verificou-se na CDM.

A identificação da problema foi por meio das actividades de análise de dados no campo de estudo (CDM), onde foi possível colectar informações suficientes que dirigiram a proposta de uma solução de integração de dados.

Após obter dados do campo e relacionar com as informações conseguidas no capítulo II, foi possível propor a implementação de um processo ETL usando SSIS na CDM para o departamento comercial na área de Vendas para a integração de dados. Deste modo, como base nas ferramentas que as TICs oferecem, foi exequível desenhar uma arquitectura funcional que é usado para ilustrar a solução proposta.

Capítulo VI – Considerações Finais

6.1. Conclusão

Nos dias que correm, as soluções ETL se mostram importantes para umas empresas com a cultura de *Data-Driven*, ou seja, orientada a dados, pois eles assumem a responsabilidade de garantir a qualidade, integridade e disponibilidade de dados para análises e relatórios. A integração de dados envolve a combinação de dados de diferentes fontes, formatos e proporciona uma visão holística, fornecendo informações abrangentes para apoiar a tomada de decisões mais informadas e estratégicas.

O trabalho de pesquisa teve como objectivo geral a proposta de um Pipeline ETL para integração de dados das vendas na CDM, para chegar à solução foram definidos quatro objectivos específicos.

O primeiro consistiu na contextualização dos pontos cruciais do trabalho no que diz respeito a *Pipeline* ETL e seus intervenientes. Este objectivo foi cumprido, visto que foram aqui trazidos conceitos, arquiteturas e funcionamento de Pipelines ETL e abordagem do *Business intelligence*, estes pontos foram abordados de forma detalhada e clara.

O segundo objectivo foi também concretizado com êxito, onde foi possível descrever a situação realística dos processos de vendas e seus constrangimentos na geração de dados para análise e relatórios, graças a uma interação profunda durante este processo de fluxo de dados, desde a extração até a análise por parte dos tomadores de decisão.

O terceiro objectivo visava em apresentar quatro possíveis soluções nomeadamente, (Oracle Data Integration, IBM DataStage, Talend open Studio e SQL Server Integration Services) e realizar uma análise comparativa segundo critérios relevantes para a CDM, foi possível através de um quadro comparativo avaliar cada solução e por fim a classificação e escolha da ferramenta ideal que foi o *SQL Server Integration Services* (SSIS) a escolha desta ferramenta foi graças aos critérios com maior impacto e relevância, sendo eles a segurança, custo e por ser um produto da Microsoft.

Finalmente, o quarto objectivo, consistia na implementação de uma solução ETL ideal escolhia para integração de dados das vendas. Este objectivo foi cumprido, porém teve de se usar o computador pessoal e não um servidor que seria a melhor opção para simular um pipeline ETL que seja adequado com o ambiente corporativa da CDM e que atenda à questões das vendas (negócio).

Em suma, pode se afirmar que o principal objectivo deste trabalho de pesquisa foi alcançado.

Em relação à pergunta de pesquisa feita no início, na introdução:

De que forma a integração de dados pode melhorar eficiência operacional das vendas na CDM?

Terminado o trabalho, chegou-se a conclusão de que a integração de dados pode melhorar significativamente o acesso rápido e centralizado dos dados para as análises e relatórios, redução e ou eliminação de trabalhos manuais e garantir qualidade de dados.

6.2. Recomendações

Recomenda-se que as Cervejas de Moçambique (CDM), implemente a Solução ETL que é proposta neste trabalho o SSIS para melhorar a eficiência operacional, sobretudo para permitir tomada de decisões mais informadas por parte dos tomadores de decisões.

Para os futuros trabalhos de pesquisa recomenda-se:

- Explorar mais sobre Engenharia de dados para apoiar a tomada de decisões baseada em dados;
- Implementação do *Business Intelligence* para impulsionar o negócio em outras áreas.

6.3. Constrangimentos

Durante a realização do presente trabalho, verificou-se os seguintes constrangimentos;

- Dificuldade em conciliar o tempo para realizar a pesquisa e com as actividades do trabalho simultaneamente;

- Não permissão na divulgação dos metadados das bases de dados e ficheiro Excel;
- Domínio da Empresa é limitado apenas ao material disponibilizado pela Empresa, obrigou a usar um computador pessoal para ter mais liberdade na pesquisa e para testar e simular o pipeline ETL;
- Entre outros constrangimentos.

7. Referências Bibliográficas

Bibliografia

[1].Menezes, E. M., & Da Silva, E. L. (2005). *Metodologia da Pesquisa e Elaboração de Dissertação.*

[2].AWS Glue. (2023). p. 1559.

[3].Braghittoni, R. (2017). *Business Intelligente.* Brasil: Casa do Codigo. Obtido em 2023

[4].Cetax. (26 de Janeiro de 2022). *Ferramenta ETL: Data Stage IBM*. Obtido em 14 de Novembro de 2023, de Cetax: https://cetax.com.br/datastage-ibm-ferramenta-de-etl/

[5].Creswell, J. W. (2007). *PROJETO DE PESQUISA* (2 ed.). (L. O. Rocha, Trad.) California, USA: Bookman.

[6].Densmore, J. (2021). *Data Pipelines Pocket Reference.* USA: O'REILLY.

[7].Fátima, N. (31 de Outubro de 2023). *Pipeline de dados versus pipeline de ETL: qual e a diferenca?* Obtido em 31 de Novembro de 2023, de Astera: https://www.astera.com/pt/type/blog/etl-pipeline-vs-data-pipeline/

[8].Favero, L. P., & Belfiore, P. (2017). *MANUAL DE ANALISE DE DADOS.* Sao Paulo, Brasil: ELSEVIER.

[9].Florea, A. M., Diaconita, V., & Bologa, R. (2015). Data integration approaches using ETL. *Database System Journal, 6*, 19. Obtido em 14 de Novembro de 2023

[9].Fuentes, A. (17 de Novembro de 2022). Obtido em 11 de Novembro de 2023, de AQUARELA: https://www.aquare.la/apache-airflow-o-que-e-e-como-funciona/

[10].Gerhardt, T. E., & Silveira, D. T. (2009). *Métodos de pesquisa.* Universidade Federal do Rio Grande do Sul. Porto Alegre: UFRGS EDITORA.

[11].Gomes, C. F., & Da Silva, R. A. (Abril de 2016). O USO DO BUSINESS INTELLIGENCE (BI) EM SISTEMA DE APOIO À TOMA- DA DE DECISÃO ESTRATÉGICA USING BUSINESS INTELLIGENCE (BI) IN MAKING SUPPORT SISTMA STRA- TEGIC DECISION. p. 20.

[11].Harenslak, B., & Ruiter, J. d. (2021). *Data Pipelines with Apache Airflow.* USA: MANNING.

[12].Jensen, C. S., Pedersen, T. B., & Thomsen, C. (2010). *Multidimensional databases and Data Wrehousing.* Morgan & Claypool.

[13].Kijoma, G. K., Campus, L. N., & Molina, M. F. (2022). *ESTUDO COMPARATIVO ENTRE FERRAMENTAS DE EXTRAÇÃO, TRANSFORMAÇÃO E CARGA DE DADOS.* São Paulo.

[14].Kimball, R., & Ross, M. (2002). *The data Warehouse Toolkit* (2 ed.). Robert Elliot.

[15].Knight, B., Veerman, E., Dickinson, G., Duoglas, H., & Herbold, D. (2008). *Professional Microsoft SQL Server 2008 Integration Services* (1 ed.). WROX.

[16].Kondado. (29 de Novembro de 2022). *O que significa Pipeline de Dados?* Obtido de Kondado: https://kondado.com.br/blog/blog/2022/11/29/o-que-significa-pipeline-de-dados/

[17].Kumar, V. A., Anandhi, M. D., Gopika, T. K., Devi, V. S., & Thenmozhi, S. (Fevereiro de 2019). Reliable Data Integration using Talend. *International Journal of Research in Engineering, Science and Management, 2*, 4.

[18].Lampert, E. d., & Badaloti, G. M. (2015). *Sistema de Informacao.* Uniasselvi.

[19].L'Esteve, R. (2022). *The Azure Lakehouse Toolkit.* Chicago, USA: Apress.

[20].Lima , J. A. (2012). *LIDERANÇA E TOMADA DE DECISÃO NA ORGANIZAÇÃO.*

[21].Lisboa, P. (18 de Agosto de 2023). *Tutorial SSIS: Aprenda Passo a Passo em Portugues*. Obtido em 15 de Novembro de 2023, de awari: https://awari.com.br/tutorial-ssis-aprenda-passo-a-passo-em-portugues-2/

[22].Machado, F. R. (2013). *Tecnologia e Projecto de Data warehouse* (6 ed.).

[23].Marquesone, R. (2017). *Big Data.* Sao Paulo, Brasil: Casa do Codigo.

[24].Massunguine, G. P. (2022). *Segurança Cibernética: Proposta de Implementação de uma Plataforma SIEM.* Relatorio de Estagio Profissional, Faculdade de Engenharia da UEM, DEEL, Maputo. Obtido em 16 de Novembro de 2023

[25].Microsoft. (2023). *Documentação do Azure Data Factory.* Obtido em 08 de Novembro de 2023, de Microsoft: https://learn.microsoft.com/pt-pt/azure/data-factory

[26].Mulbert, A. L., & Ayres, N. M. (2011). *Gestão da Informação* (3 ed.). (UnisulVirtual, Ed.) Brasil.

[27].Munappy, A. R., Bosch, J., & Olsson, H. (2020). Data Pipeline Management in Practice:Challenges and Opportunities. p. 17.

[29].Nicollete, D. (20 de Maio de 2022). Obtido em 01 de Novembro de 2023, de StreamSets: https://streamsets.com/blog/data-pipeline-architecture-principles/

[30].Oliveira, J. V. (2009). *Meodologia de Sistemas de Data Warehouse.* Tese de Doutoramento, Universidade do Minho, Sistemas de Informacao. Obtido em 19 de Novembro de 2023

[31].Oliveira, M. F. (2011). *METODOLOGIA CIENTÍFICA: um manual para a realização de pesquisas em administração.* UNIVERSIDADE FEDERAL DE GOIÁS, Goias.

[32].Reis, J., & Housley, M. (2022). *Fundamentals of Data Engineering* (1 ed.). USA: Wilkey.

[33].Savjani, P. (10 de Setembro de 2019). *Design your Data pipelines in Azure Data Factory to load data into Azure Database for MySQL*. Obtido em 13 de Novembro de 2023, de Microsoft: https://techcommunity.microsoft.com/t5/azure-database-for-mysql-blog/design-your-data-pipelines-in-azure-data-factory-to-load-data/ba-p/847871

[34].Sharma, V. (4 de Agosto de 2022). *A Complete Guide on Building an ETL Pipeline for Beginners*. Obtido em 27 de Novembro de 2023, de Analytics Vidhya: https://www.analyticsvidhya.com/blog/2022/06/a-complete-guide-on-building-an-etl-pipeline-for-beginners/

[35].Stair, R. M., & Reynolds, G. W. (2015). *Principles of Information Sytems* (11 ed.). (C. Learning, Ed.) USA.

[36].Suleman, E. (2023). Obtido em 13 de Novembro de 2023, de Traindex: https://www.traindex.io/blog/introduction-to-data-pipelines-26o7/

ANEXOS

Anexo 1: Guia de Entrevista

UNIVERSIDADE EDUARDO MONDLANE

FACULDADE DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA ELECTROTÉCNICA

Guia de Entrevista

1. **Quando foi fundada a CDM?**

   R: A CDM foi fundada em 1995.

2. **Quantos Departamentos tem a CDM?**

   R: Oito Departamentos (8) nomeadamente, Comercial, Recuros Humanos, Assuntos Legais e Corporativos, Finanças, Produção, Markenting, IT e Logística.

3. **Quando (Ano) é que a CDM passou a fazer parte de ABInBev?**

   R: outubro de 2016.

4. **Quais países fazem parte da Zona Business Unit south com Mocambique?**

   R: Tanzânia, Uganda, Zâmbia, Gana e Botswana.

5. **Qual foi o melhor ano em termos de vendas da CDM?**

   R: Foi o ano de 2021 com cerca de 3,3milhões de hectolitros.

6. **Como esta dividida a equipe de vendas?**

   R: Equipe de BDRs, DS ambas lideradas pelos *Sales Manager* em todo o país, e Agentes CIC que efectuam as vendas via chamada telefônica.

7. **Quais são os maoires concorrentes da CDM?**

   R: A Heineken Moçambique.

8. **Qual é a presenca (%) da CDM no Mercado Moçambicano de Cerveja?**

   R: De acordo com os últimos estudos de pesquisa realizados, a CDM detém cerca de 94% de *share* no mercado.

9. **Quantos colaboradores tem a CDM?**

   R: Conta com mais de 1.000 colaboradores.

Anexo 2: Especificaçôes do Computador a usar

Sistema Operativo	Windows 11 Home 22H2
Arquitectura	64 bits
Processador	12th Gen intel(R) Core™ i7-12650H 3.30GHz
Memoria RAM	16 GB
Armazenamento (HDD)	1TB

Tabela 4: Especificações do Computador a usar.

Anexo 3: Download e Instalação da ferramenta Visual Studio

Download do Visaual Studio

O visual Studio é um ambiente de desenvolvimento integrado (IDE) criado pela Microsoft. Ele oferece um conjuto abragente de ferramentas e serviços para desenvolvedores criar software para diversas plataformas. Esta será a ferramenta usada para criar os projectos SSIS, para obter a ferramenta vai se acessar o site da Microsoft, eis o link para baixar o VS: https://visualstudio.microsoft.com/downloads/. Vai se optar pela versão comunity.

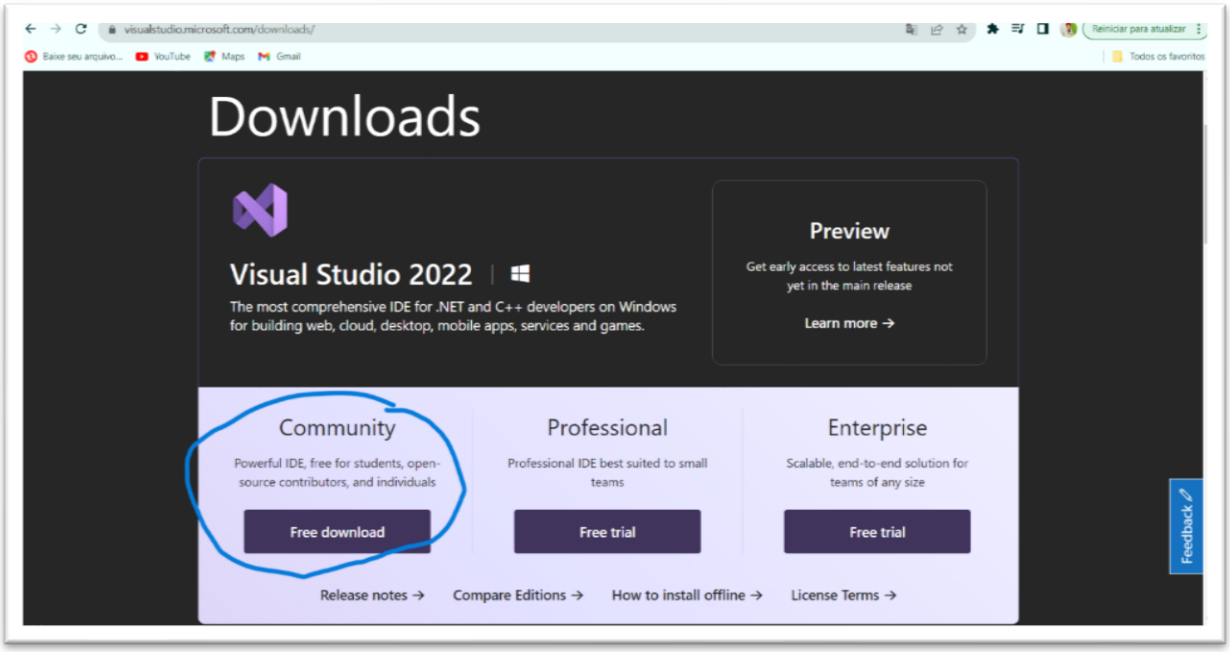

Figura 15: Escolha da versão comunity para baixar

Após baixar o VS, segue-se o processo de instalação.

Executando o Setup temos:

Passo 1: Instalando o Visual Studio

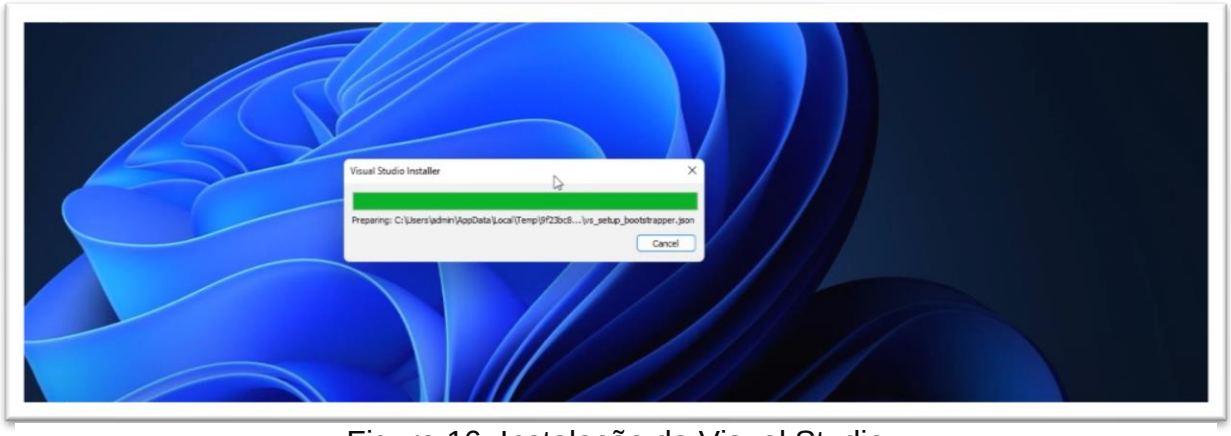

Figura 16: Instalação do Visual Studio

Passo 2: Clicar em continuar

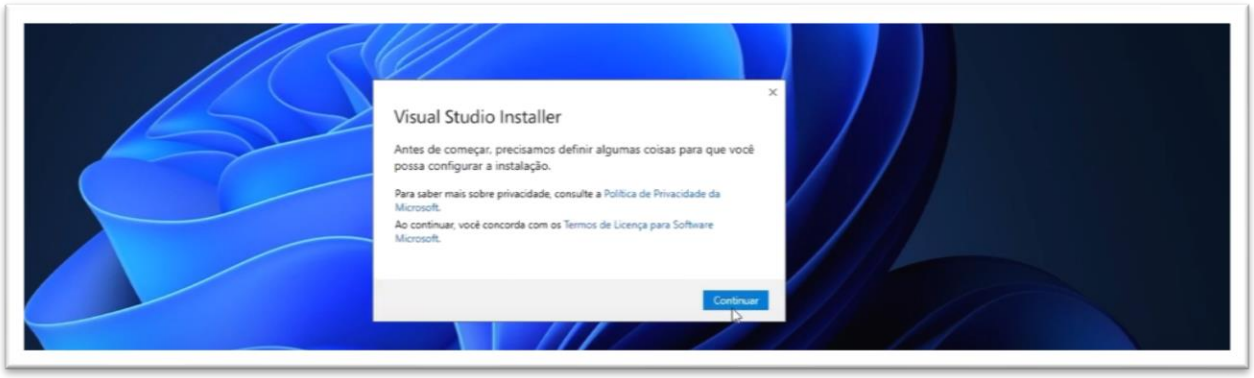

Figura 17:Obtendo o Instalador

Passo 3: Obtendo o Instalador

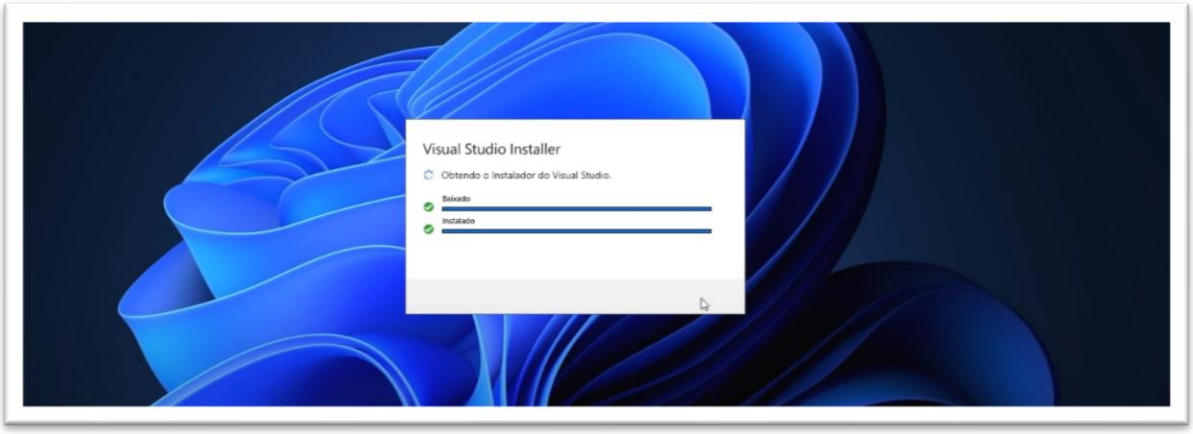

Figura 18: Instalador do Visual Studio

Passo 4: Escolhendo Módulo de Processamento armazenamento de dados

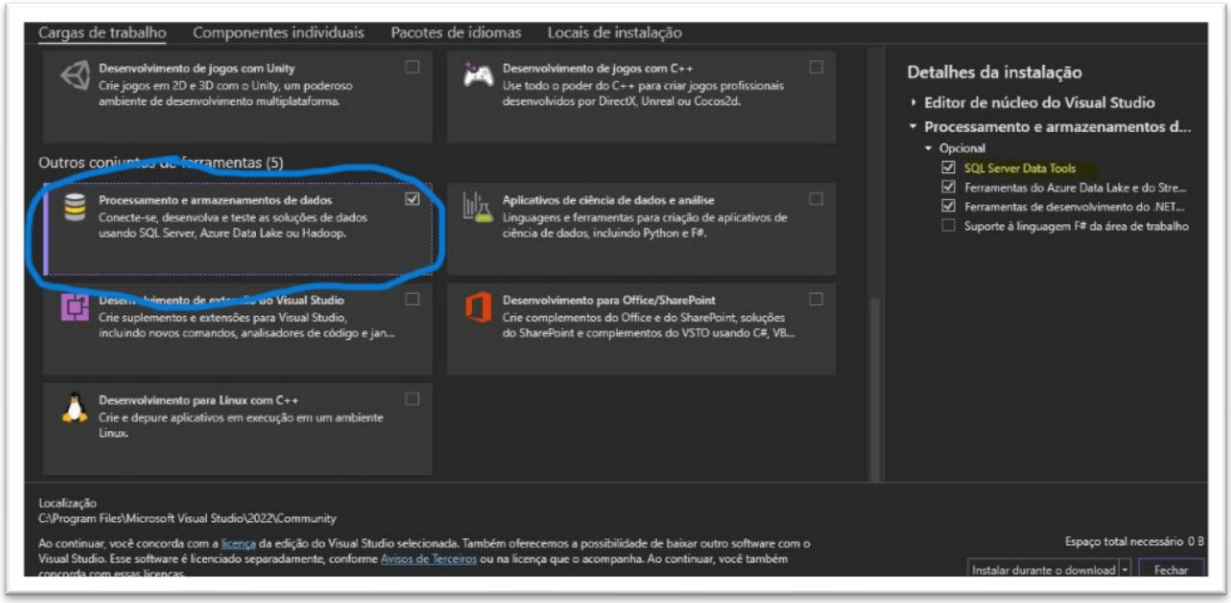

Figura 19: Escolha do Módulo de Processamento e armazenamento de dados

Passo 5: Terminada instalação agora vamos abrir o Visual Studio

Figura 20:Abrindo o Visual Studio

Depois de abrir o Visual Studio, a janela abaixo será exibida, de seguida clicar em “Continuar sem código”

Figura 21: Painel Inicial do Visual Studio

Após em “Continuar sem códigos”, uma janela em branco será exibida em seguida vamos instalar o SSIS, clicando em extensões.

Figura 22: Obtebdo o Data Tools

Após clicar em na aba “Extensões”aparecer a janela da figura abaixo, em seguida pesquisar por “Integration Services” e baixar

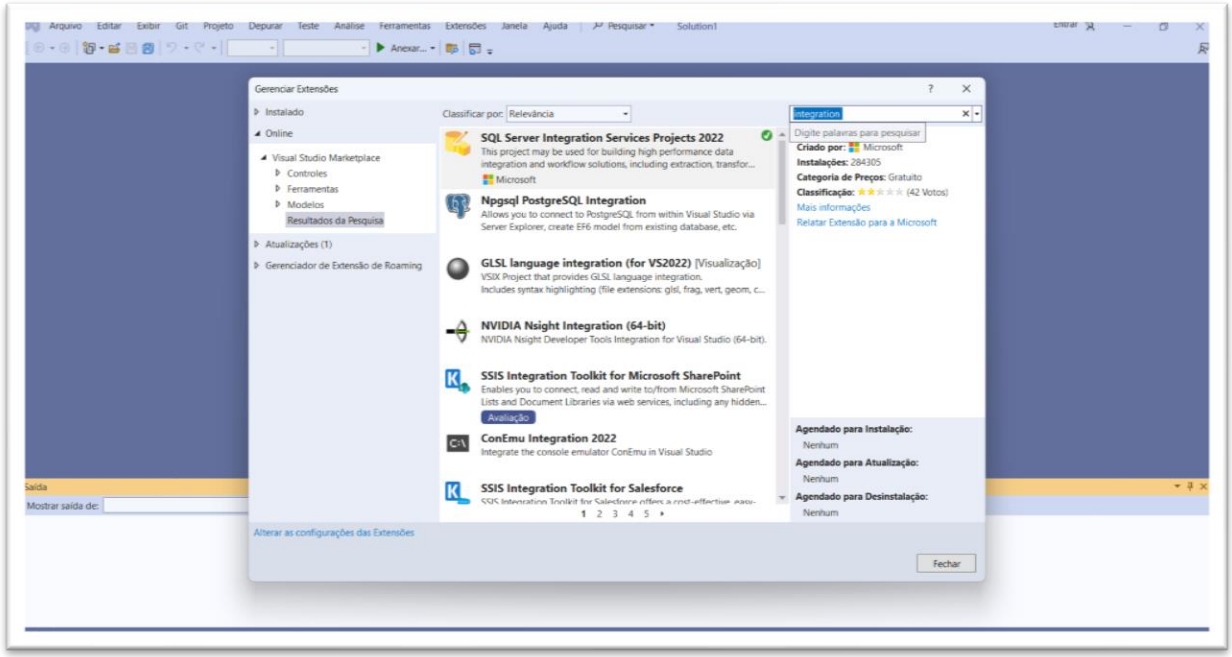

Figura 23: Instalando o SSIS

Depois que o SSIS é instalado, já pode mos criar um projecto.

Figura 24: Criando um Projecto

Pesquisamos pelo tipo de projecto, neste caso o “Integration Services” e clicamos em “próximo”.

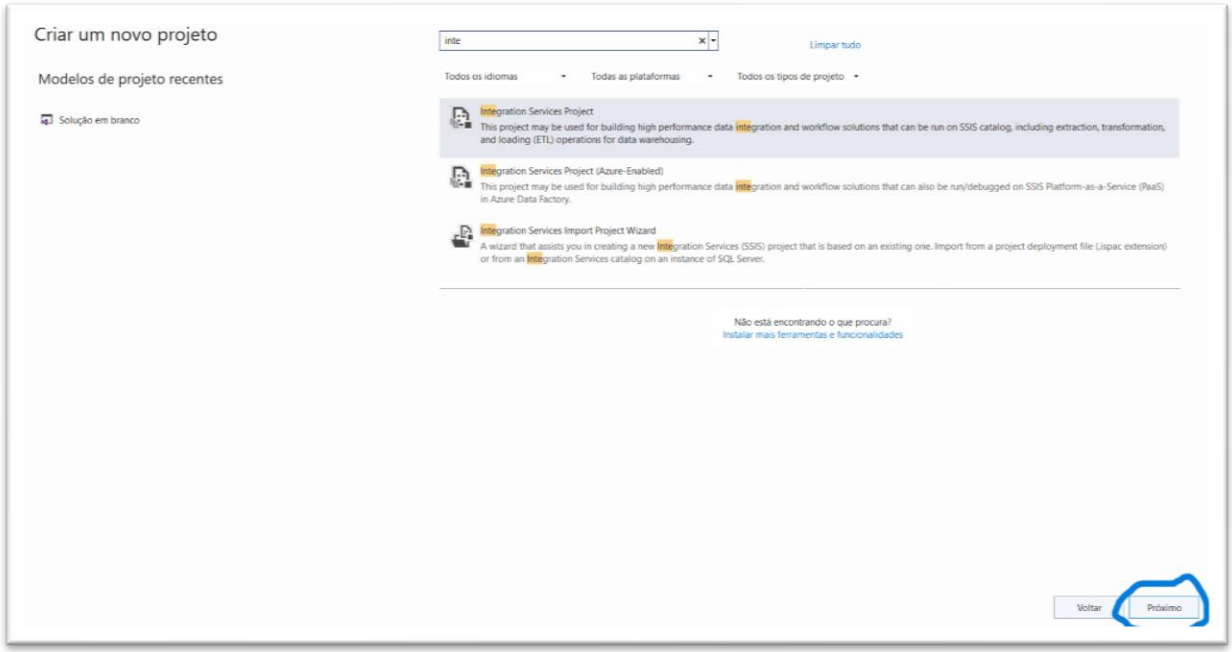

Figura 25: Escolhendo um Projecto de Integração

Anexo 4: Criando um Projecto SSIS

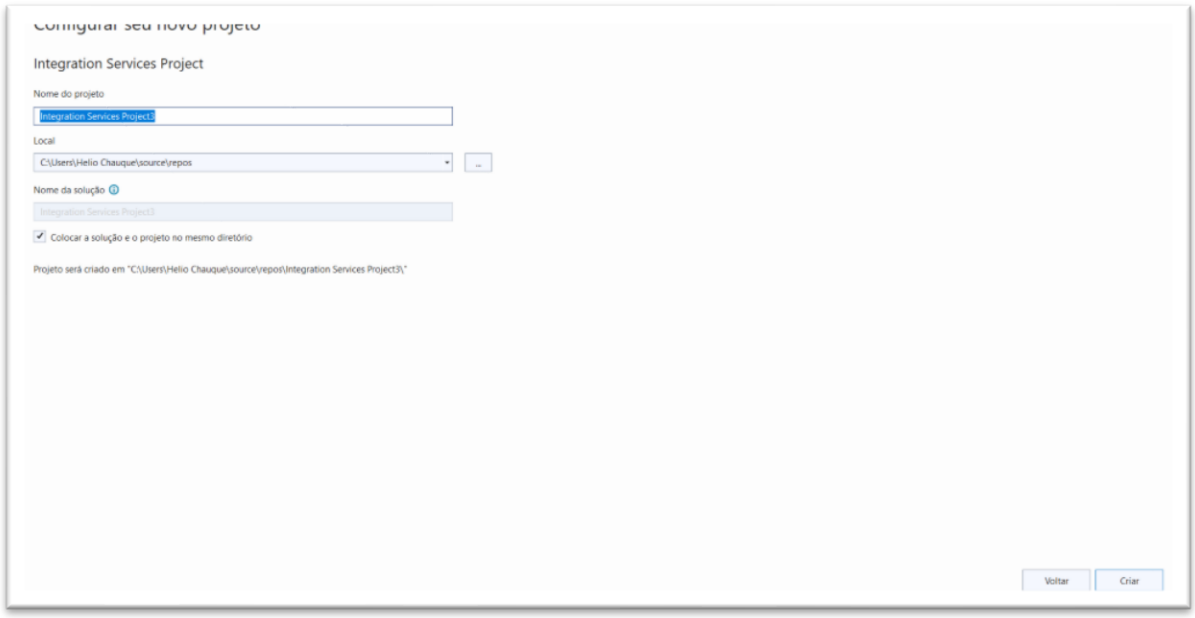

Figura 26: Dando um nome ao Projecto

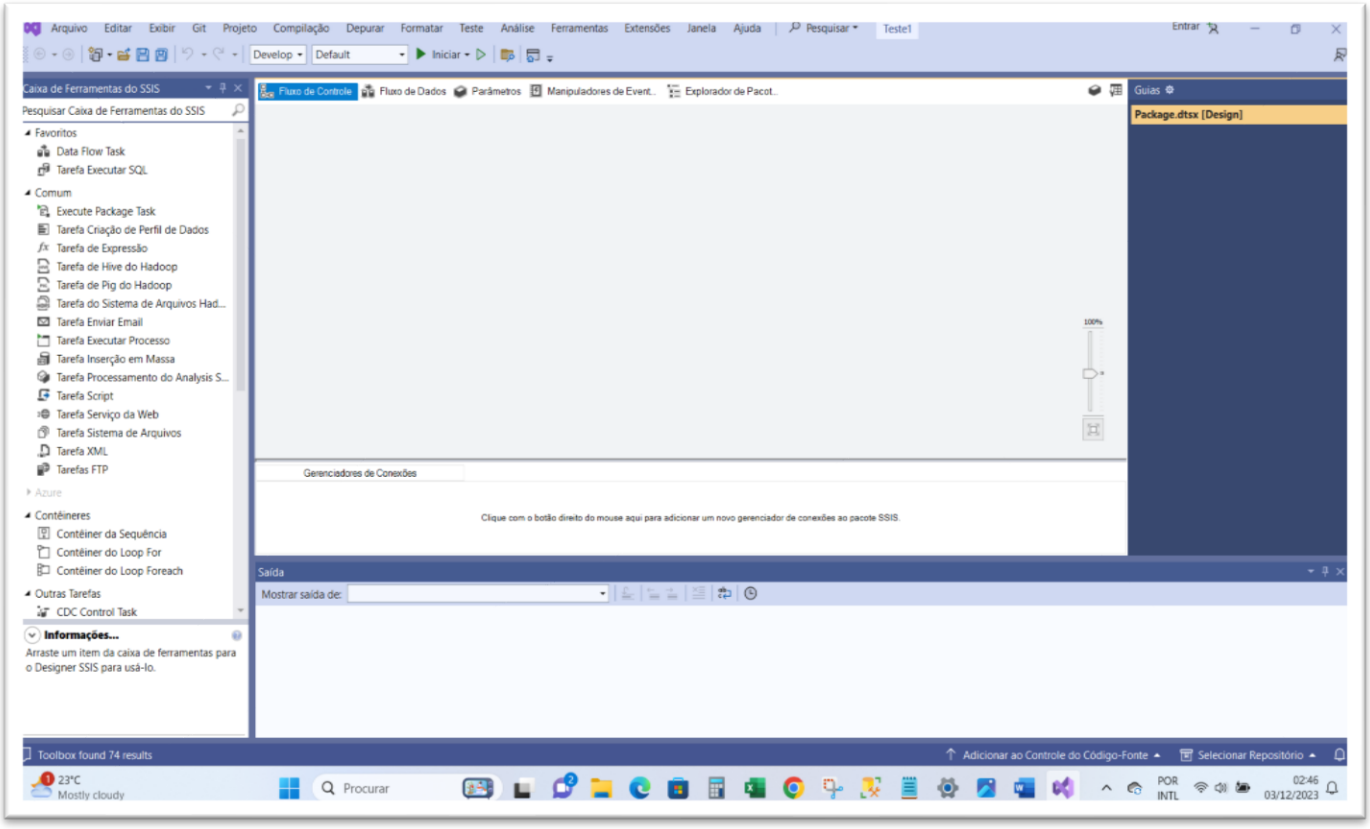

Figura 27: Ambiente do SSIS

O primeiro passo em um projecto SSIS é criar as conexões com as fontes e destino.

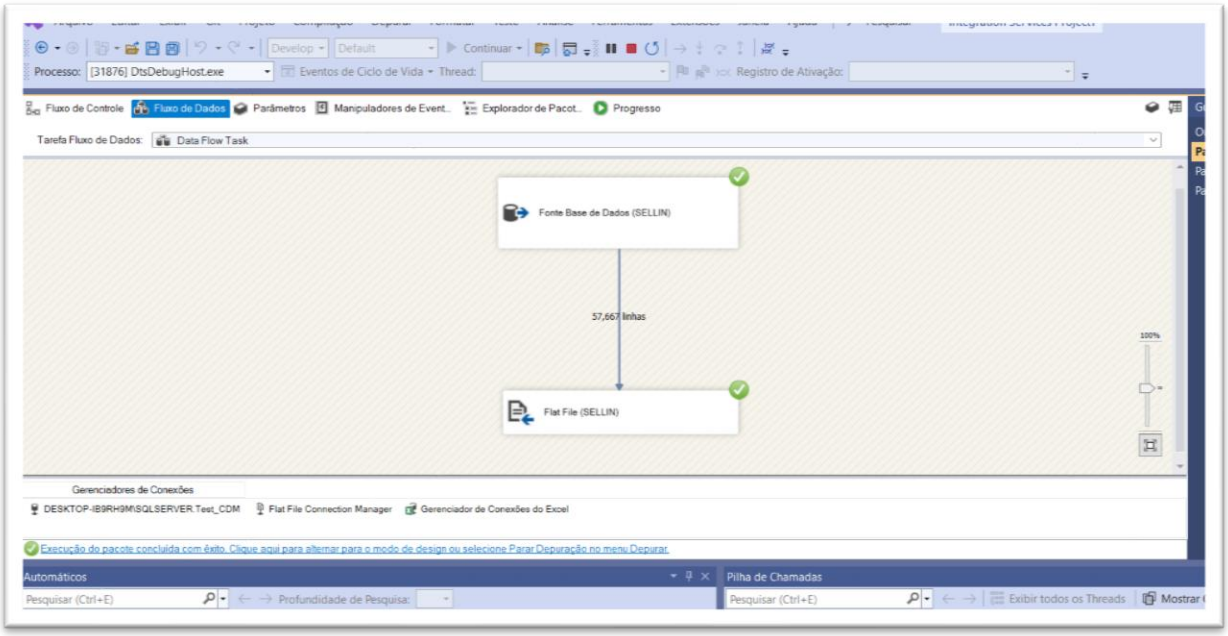

Figura 28: : ETL que Movimenta os dados da Base de Dados transacional para o Excel na pasta destino.

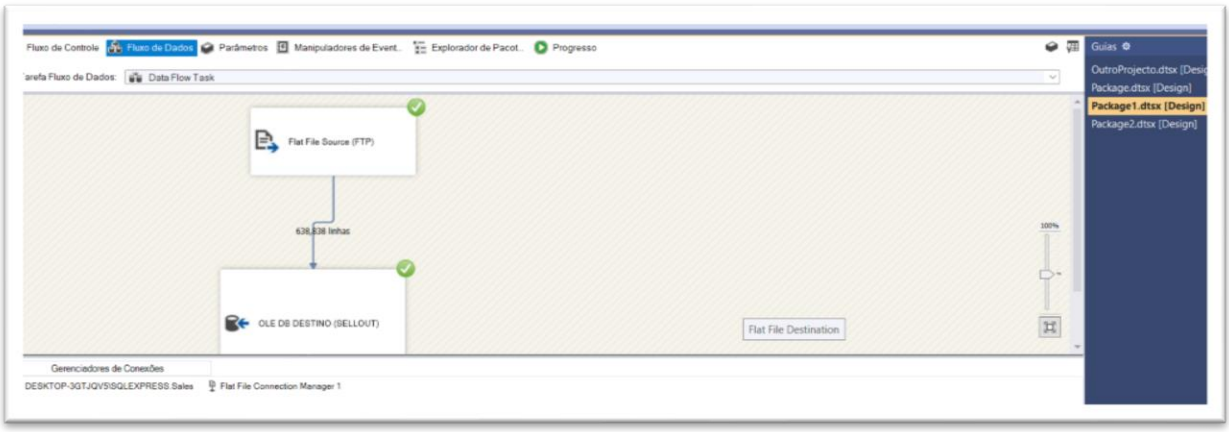

Figura 29: Extraindo os dados de Sellout em um ficheiro Excel para Base de Dados

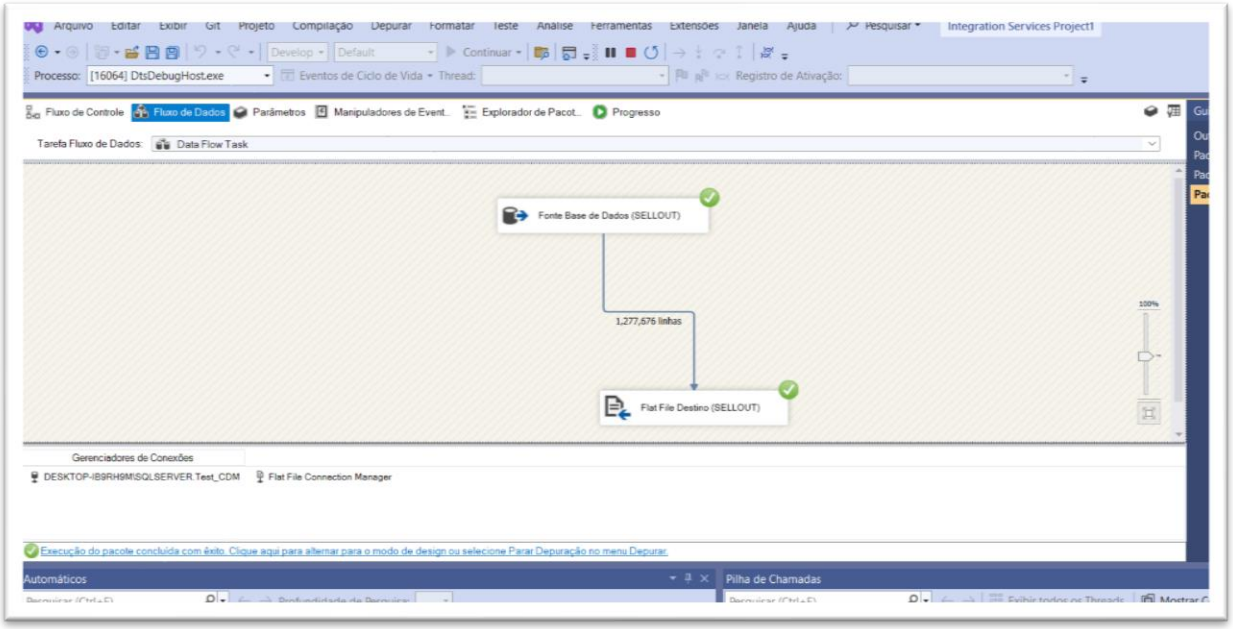

Figura 30: Carregando Dados de SellOut da Base de Dados para o arquivo Excel na pasta destino.