Azure Data Factory: Uma Nova Ferramenta ETL

Fábio João Anastácio, Thiago Francisco Ferreira

Escola Politécnica - Pontifícia Universidade Católica do Paraná (PUCPR)
Curitiba - PR - Brasil

fbianastacio@gmail.com, ferreira_thiago@outlook.com

Resumo. *Este trabalho descreve a construção de um processo ETL que utiliza a tecnologia Azure, tendo como um dos recursos o Data Factory da Microsoft, no contexto de um processo de retenção comercial. Demonstra a forma como são orquestrados, preparados e transformados os dados dentro da ferramenta, de maneira que sirva de estrutura base para um sistema de apoio à tomada de decisão em empresas de negócios.*

Palavras-chave: *Data Factory, Azure, ETL*

Abstract. *This work describes the construction of an ETL process that uses Azure technology, having Microsoft Data Factory as one of its resources, in the context of a commercial retention process. It demonstrates the way how data is orchestrated, prepared and transformed within the tool, in such a way that it serves as a base structure for a decision support system in business companies.*

1. Introdução

Um Data Warehouse (DW) é um conjunto de tecnologias que permitem que a melhor e mais rápida decisão seja tomada. É neste ponto em que os data warehouse se diferem dos bancos de dados operacionais, pois os DWs são orientados ao assunto, integrados e não voláteis. Além disso, são também resumidos e utilizam OLAP - que é a capacidade para manipular e analisar um grande volume de dados sob múltiplas perspectivas [EL-SAPPAGH, 2011].

As ferramentas de Extração-Transformação-Carregamento (ETL: Extraction-Transformation-Loading) existem desde o início dos anos 90 e atualmente tiveram muitos avanços, pois uma boa ferramenta de ETL deve ser capaz de se comunicar com diversos BD e ler diferentes tipos de formatos de arquivos [FERREIRA, 2010].

Neste contexto a seleção de uma ferramenta ETL adequada é uma decisão muito importante a ser tomada pois ela opera o núcleo do DW; e é neste cenário em que entra uma nova ferramenta disponível no mercado: o Azure Data Factory.

Este artigo irá demonstrar na prática como o Azure Data Factory, uma nova ferramenta ETL que foi lançada recentemente no mercado foi utilizada para um projeto de captação de informações para a equipe de retenção comercial em uma organização.

O restante deste trabalho está organizado da seguinte forma: a sessão 2 descreve os sistemas de apoio à tomada de decisão, apresentando as ferramentas ETL e os DW. A sessão 3 apresenta o Azure Data Factory e demonstra como a ferramenta funciona na prática. Por fim, na sessão 4 seguem as conclusões e trabalhos futuros. Este artigo foi apresentado como

requisito parcial à obtenção do grau de Especialista em Gestão de Banco de Dados e Big Data.

2. Os Sistemas de Apoio à Tomada de Decisão

Durante o início da década de 60, existiam na computação, diversas aplicações escritas especialmente em COBOL, não integradas e executadas sobre diversos arquivos mestres. Esses arquivos eram armazenados em fitas magnéticas de alta capacidade, o que também ocasionou uma crescente complexidade na manutenção dos programas existentes e o surgimento de novos sistemas de apoio à tomada de decisão [INMON, 1997].

Os sistemas de apoio à tomada de decisão são definidos por POLLONI (2000) como "aqueles que tratam de assuntos específicos, estatísticas, projeções, comparações de dados referentes ao desempenho de uma organização, estabelecendo parâmetros para novas ações dentro do negócio da empresa. Esses sistemas se caracterizam pela utilização de pacotes interativos para cálculos e/ou simulações".

Nessa circunstância, o volume de dados nas organizações não para de crescer e, ao mesmo tempo, aumenta a necessidade de se obter uma visão consistente de todos os dados que são armazenados Warehouse. A maior parte do esforço exigido durante o desenvolvimento de um Data Warehouse é consumido no processo de ETL e não é incomum que cerca de 80% de todo o esforço seja empregado nesta fase [INMON, 1997 apud ABREU, 2007].

2.1 ETL

Trata-se do processo mais crítico e demorado na construção de um DW, pois é composto por três etapas: Extração, Transformação e Carregamento. Essas três etapas sofrem diretamente influência da regra de negócio e devem ser modeladas de acordo com o que se espera no front-end, ou seja, na saída das informações pois elas influenciam diretamente na tomada de decisão [PAPASTEFANATOS et al. 2012].

Abaixo, segue a figura 1 demonstrando o processo de ETL:

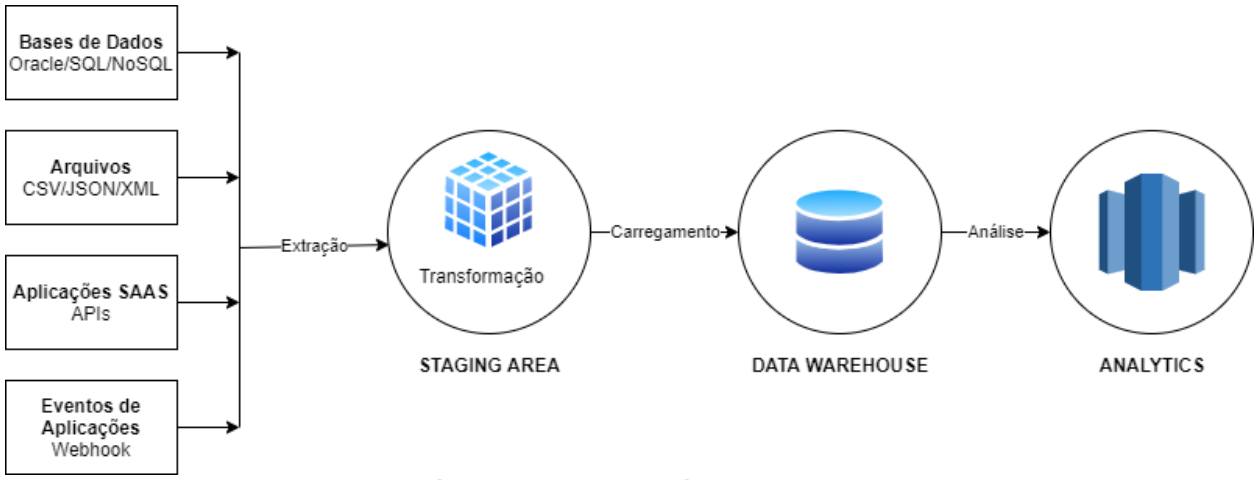

Figura 1 - Processo de ETL

Extração: Nesta etapa os dados são copiados ou exportados dos locais de origem para uma área de preparação, a "Staging Area". Os dados podem vir praticamente de qualquer fonte estruturada ou não-estruturada (servidores SQL ou NOSQL), sistemas ERP e CRM, arquivos de textos, de páginas da web, de documentos, e-mails entre outros.

Transformação: Durante esta etapa os dados brutos são transformados para serem úteis à análise e para se ajustar ao esquema do DW de destino, onde são alimentados por um processamento analítico OLAP (Online Analytical Processing), envolvendo:

- Limpar, filtrar, eliminar as possíveis duplicidades de dados, validações e autenticidade dos dados;
- Execução de cálculos, traduções ou resumos com bases em dados brutos, incluindo desde a alteração de cabeçalhos de linhas e colunas para obter consistência de dados, a conversão de moedas ou unidades de medidas para padronização, a edição de strings de texto, a soma ou média de valores – tudo o que for necessário para se adequar ao modelo de DW específico de acordo com o propósito
- Remover, criptografar, ocultar ou proteger de outra forma os dados regidos por regulamentações governamentais ou setoriais
- Formatação dos dados em tabelas para corresponder ao esquema do armazenamento dos dados no destino

Essas transformações realizadas na "Staging Area" limitam o impacto do desempenho nos sistemas de origem e reduzem a probabilidade de corrupção de dados.

Carregamento: Na última etapa, os dados transformados são movidos da área de preparação para o DW de destino. Geralmente isso envolve um carregamento inicial de todos os dados, seguido pelo carregamento periódico das alterações incrementais de dados (Carga Full e Delta por exemplo) e, com uma menor incidência, atualizações completas para substituir os dados.

Todas essas etapas são necessárias para que possamos ter um DW com todas as características qualitativas e quantitativas necessárias para atender às demandas solicitadas pela área de negócios.

2.2 Data Warehouse

O DW é um repositório proveniente de dados operacionais onde deve ser criado um ambiente homogêneo e padronizado para propiciar as análises de negócio concentradas em um só local. Segundo Kimball (1998), as características mais relevantes para garantir a qualidade dos dados em um DW são:

- **Unicidade:** evitando assim duplicações de dados;
- **Precisão:** os dados não podem perder suas características originais assim que são carregados para o DW;
- **Completude:** não gerando dados parciais de todo o conjunto relevante às análises
- **Consistência:** os fatos devem apresentar coerência com as dimensões que o compõem;

Para que o DW seja considerado bem-sucedido, os interessados ao negócio devem aceitá-lo e ter confiança nas informações fornecidas por ele, possuindo alguns requisitos destacando-se [Kimball, 1998]:

- Permitir fácil acesso à informação com conteúdo compreensível e dados intuitivos, além de retornar resultados às consultas no menor tempo possível;
- Ser adaptável e flexível às mudanças, às necessidades dos usuários, às regras de negócio vigentes, aos dados e à tecnologia empregada;
- Ser seguro pois a maioria das informações armazenadas são confidenciais;

Os DW não necessitam ser construídos de uma vez, pois a complexidade exigida é muito alta, ao invés disso, podem abordar processos de negócio de maneira crescente e serem concebidos aos poucos.

Desde os primeiros DW criados na década dos anos 1970 muita coisa evoluiu, principalmente com o surgimento da computação em nuvem. Constantemente, as empresas têm deixado de investir em suas infraestruturas locais para focar nos serviços em nuvem pois estes oferecem alta performance, simples implementação, fácil administração, alta escalabilidade e disponibilidade a um custo relativamente menor em curto prazo.

Além disso, a integração com as ferramentas de BI com análises em tempo real, a facilidade em trabalhar com linguagens mais universalizadas para consulta, as análises de dados (como por exemplo o SQL) e também a necessidade de não haver nenhuma intervenção diretamente com recursos do servidor (memória, disco, processador por exemplo), por parte do desenvolvedor do DW, tem feito dessas ferramentas uma grande opção quando se deseja atingir os requisitos de desempenho, flexibilidade e segurança exigidos pelas análises de desenvolvimento de grandes volumes de dados.

3 Azure Data Factory

O Azure Data Factory é um serviço de nuvem oferecido pela Microsoft dentro da plataforma Azure que permite a integração de dados de diferentes fontes. O Azure Data Factory é uma solução muito adequada quando se precisa construir pipelines híbridos de extração-transformação-carga (ETL), extração-carga-transformação (ELT) e integração de dados. Além disso, o Data Factory executa de forma simplificada o processamento desses dados, agregando, transformando e carregando esses de forma escalar.

Atualmente, o ADF suporta cerca de 90 tipos de conectores das mais diversas fontes, desde as mais famosas de Big Data como Amazon Redshift, Google BigQuery e HDFS, além de data warehouses corporativos como Oracle Exadata e Teradata e também aplicativos SaaS (Software-as-a-Service) como Salesforce, ServiceNow, Marketo além de todos os serviços de dados já suportados através da nuvem da Azure.

A seguir, são descritos os passos necessários para o projeto e implementação do Data Factory de uma carga de dados de informações de retenção comercial.

3.1 Data Warehouse para a Área de Informações de Retenção Comercial

Por se tratar de informações comerciais provenientes de variados sistemas que serão utilizadas para a criação do DW, utilizaremos uma base fornecida por uma empresa que disponibiliza arquivos EDI (*Eletronic Data Interchange*), que tem o objetivo de facilitar a padronização inicial que deve ser feita nos nomes dos campos, nos nomes dos arquivos, nos produtos e entre outras características e informações importantes.

Para evitar o desabastecimento de produtos dentro dos clientes, existe uma forma de prevenção que auxilia no processo de retenção que é realizado pela empresa através da reposição de estoque contínua e/ou periódica, variando de acordo com o cliente. Ou seja, assim que o produto está para faltar no estoque do cliente, o time de retenção recebe um aviso para já identificar que aquele cliente será um possível comprador em breve.

O objetivo então é desenvolver dentro do Azure Data Factory a ingestão e a orquestração desses dados, atendendo assim a demanda de negócios de retenção comercial. Demonstraremos a seguir, como funciona o processo de ETL dentro do Azure Data Factory.

3.2 Extração, Transformação E Carregamento Na Prática

Inicialmente os arquivos EDI são disponibilizados diariamente em um ambiente de rede compartilhado (file system), onde há a captação de um arquivo em Excel. Este arquivo em Excel já vem previamente padronizado, porém ainda assim, demanda um tratamento.

Abaixo na Figura 2, podemos visualizar as configurações da extração de dados do arquivo, testar a conexão, editar o nome, escolher o tipo de compressão, entre outros.

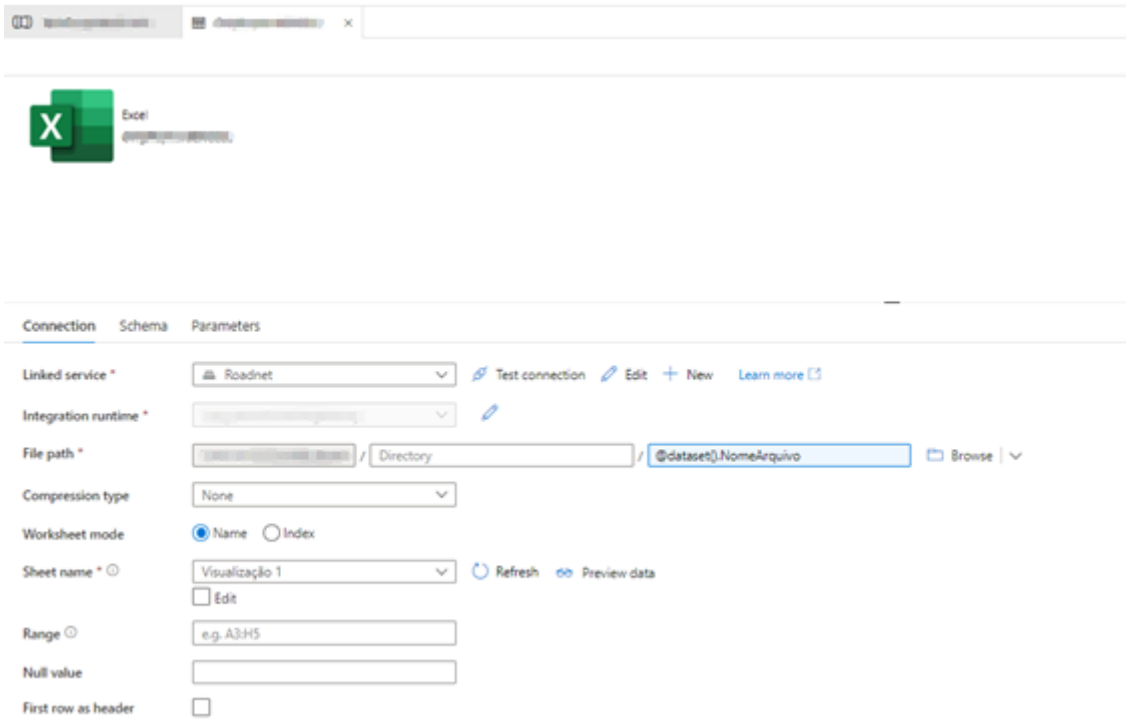

Figura 2 - Configuração da Extração

Após configurarmos a captura dos dados, devemos configurar a saída do arquivo que é a etapa de carregamento, onde deverá compor uma base de dados Azure SQL, diretamente no banco na nuvem. Nessa carga definimos a tabela intermediária “RAW” para onde os dados serão enviados, conforme pode ser visto na Figura 3:

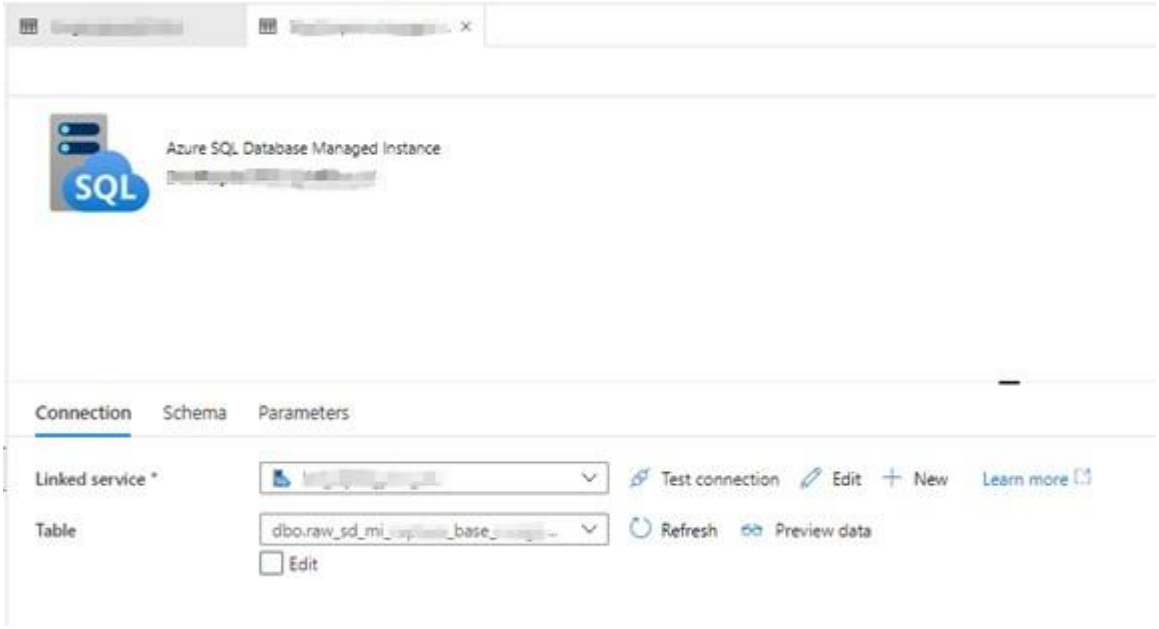

Figura 3 - Configuração da saída para o banco (Extração)

Então, criamos a tarefa que efetua a extração do arquivo de origem para o arquivo de destino, efetuando uma classificação inicial dos dados oriundos do Excel, ignorando a primeira linha e a primeira coluna, e redefinindo os nomes das linhas de acordo com a tabela intermediária que será o primeiro destino dos dados. Ou seja, a tarefa de extração altera o nome das colunas que no arquivo em Excel vem como “números”, conforme vemos na Figura 4:

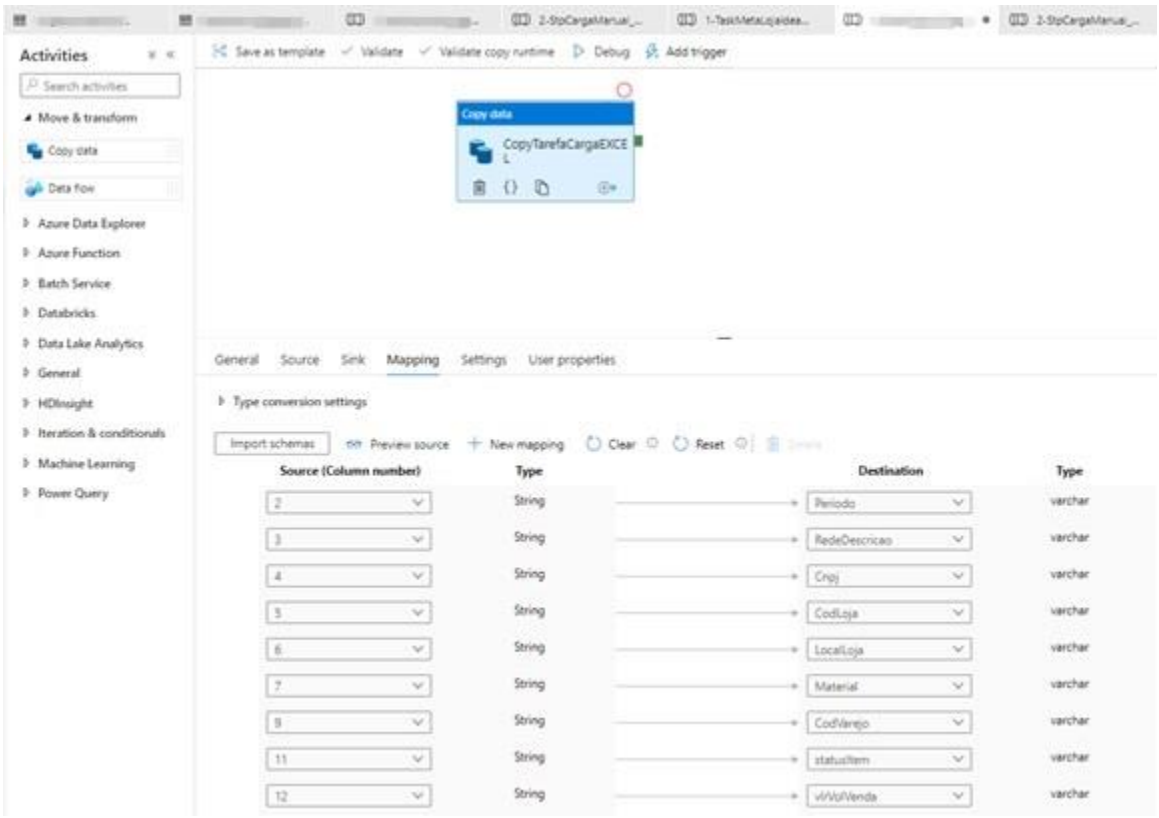

Figura 4 - Configuração da Extração

Como podemos observar, a coluna 2 se torna “Periodo”, a coluna 3 se torna “RedeDescricao” e assim sucessivamente. O nome desse processo é chamado de Mapping dentro do ADF.

Por fim, conforme na figura 5, efetuamos a última tarefa onde configuramos a carga final para a geração do dataset que irá para o DW. Ali, ela ainda irá sofrer uma transformação durante a etapa de carregamento que foi desenvolvida dentro da procedure para fazer a transformação a partir da tabela intermediária, tratando os dados que foram inseridos nela na etapa de extração.

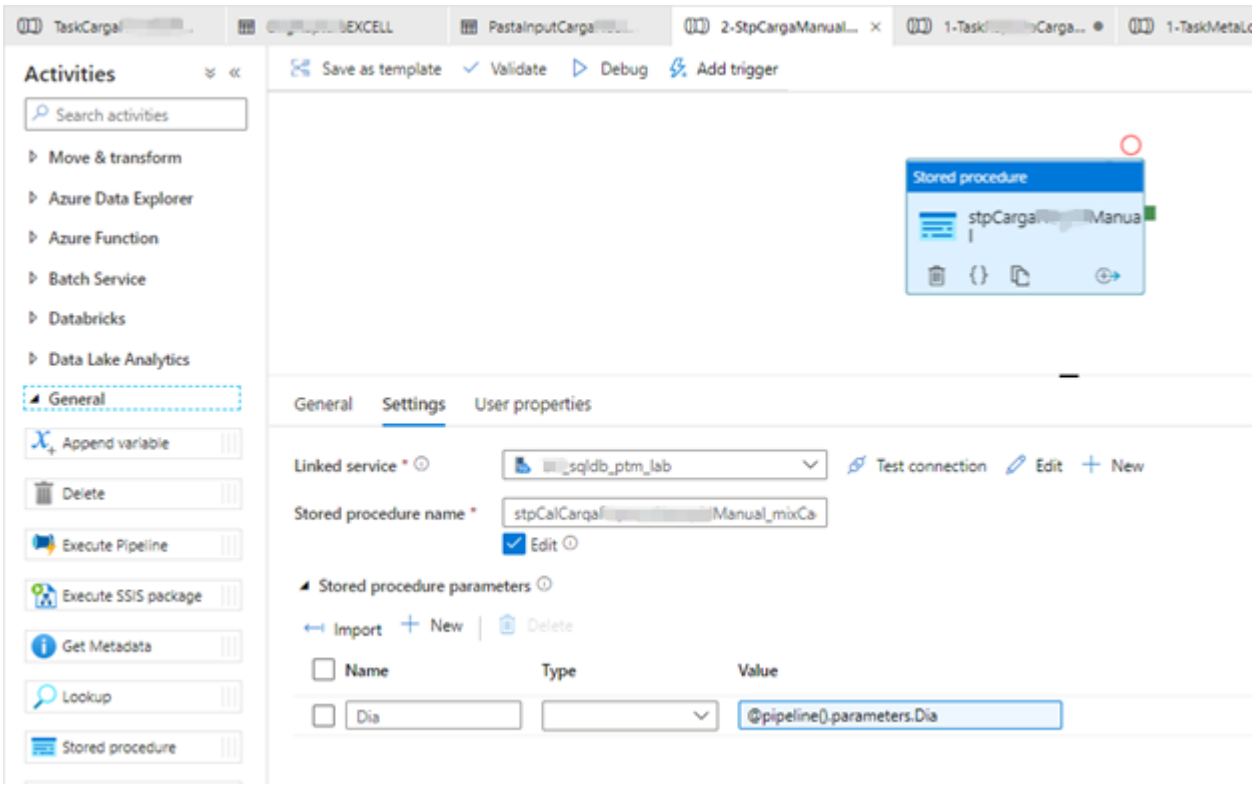

Figura 5 - Configuração da Transformação

A procedure efetua os tratamentos de dados, reportando os problemas que podem acontecer, como por exemplo um período inexistente, um código de loja ou de produto inválido ou inexistente, entre outros, e isso é alertado ao ADF através de uma mensagem de erro que avisa o usuário em uma tela de monitor de tarefas, conforme podemos ver na Figura 6 abaixo:

Showing 1 - 38 items

Pipeline name	Run start	Run end	Duration	Triggered by	Status	Run
stpCargaVendaRealizada	3/17/21, 12:10:14 PM	3/17/21, 12:17:48 PM	00:07:33	b9214c85-bc2b-4ce2-ad	Succeeded	Original
stpCargaMaterialEstrutu...	3/17/21, 12:10:05 PM	3/17/21, 12:10:13 PM	00:00:07	cdfec35b-1ba8-4495-a7b	Succeeded	Original
SpcCargaLogisitcia	3/17/21, 12:01:28 PM	3/17/21, 12:10:04 PM	00:08:36	737192ce-ea13-451f-974	Succeeded	Original
SpcCargaGeral	3/17/21, 12:00:03 PM	3/17/21, 12:01:27 PM	00:01:23	99e1041c-40b2-4638-bd	Succeeded	Original
Gerenciador Cargas	3/17/21, 12:00:02 PM	3/17/21, 12:30:26 PM	00:30:23	Manual trigger	Succeeded	Original
TaskRetorno _App	3/17/21, 12:00:00 PM	3/17/21, 12:00:38 PM	00:00:37	Schedule Retorno App	Succeeded	Original
stpCargaVendaRealizada	3/17/21, 11:47:45 AM	3/17/21, 11:54:08 AM	00:06:23	Manual trigger	Succeeded	Rerun (Latest)
TaskRetorno _App	3/17/21, 11:00:00 AM	3/17/21, 11:02:00 AM	00:01:59	Schedule Retorno App	Succeeded	Original
stpCalculaDMD	3/17/21, 6:00:01 AM	3/17/21, 6:01:03 AM	00:01:02	Schedule Demanda Medi	Failed	Original
stpCargaMaterialEstrutu...	3/17/21, 5:55:00 AM	3/17/21, 5:55:08 AM	00:00:07	14f2830b-1ec3-435c-84c	Succeeded	Original
SpcCargaLogisitcia	3/17/21, 5:47:18 AM	3/17/21, 5:54:59 AM	00:07:41	c379a331-af00-4fad-9ef	Succeeded	Original
SpcCargaGeral	3/17/21, 5:46:00 AM	3/17/21, 5:47:17 AM	00:01:16	670fc989-9b6e-4d61-94	Succeeded	Original
Gerenciador Cargas	3/17/21, 5:46:00 AM	3/17/21, 6:01:07 AM	00:15:07	Schedule_Master	Failed	Original

Figura 6 - Monitor de Tarefas do ADF

Assim, na Figura 7 a seguir, podemos ver todo o processo de ETL dentro do Azure Data Factory:

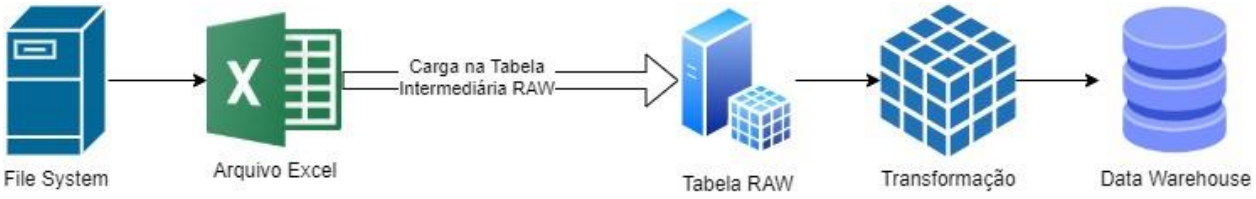

Figura 7 - Processo de ETL de Retenção Comercial

Após todas as configurações de Extração, Transformação e Carregamento, definimos para que a ferramenta envie uma notificação por e-mail toda vez que o processo for executado, confirmando sucesso ou insucesso no processo, conforme a Figura 8:

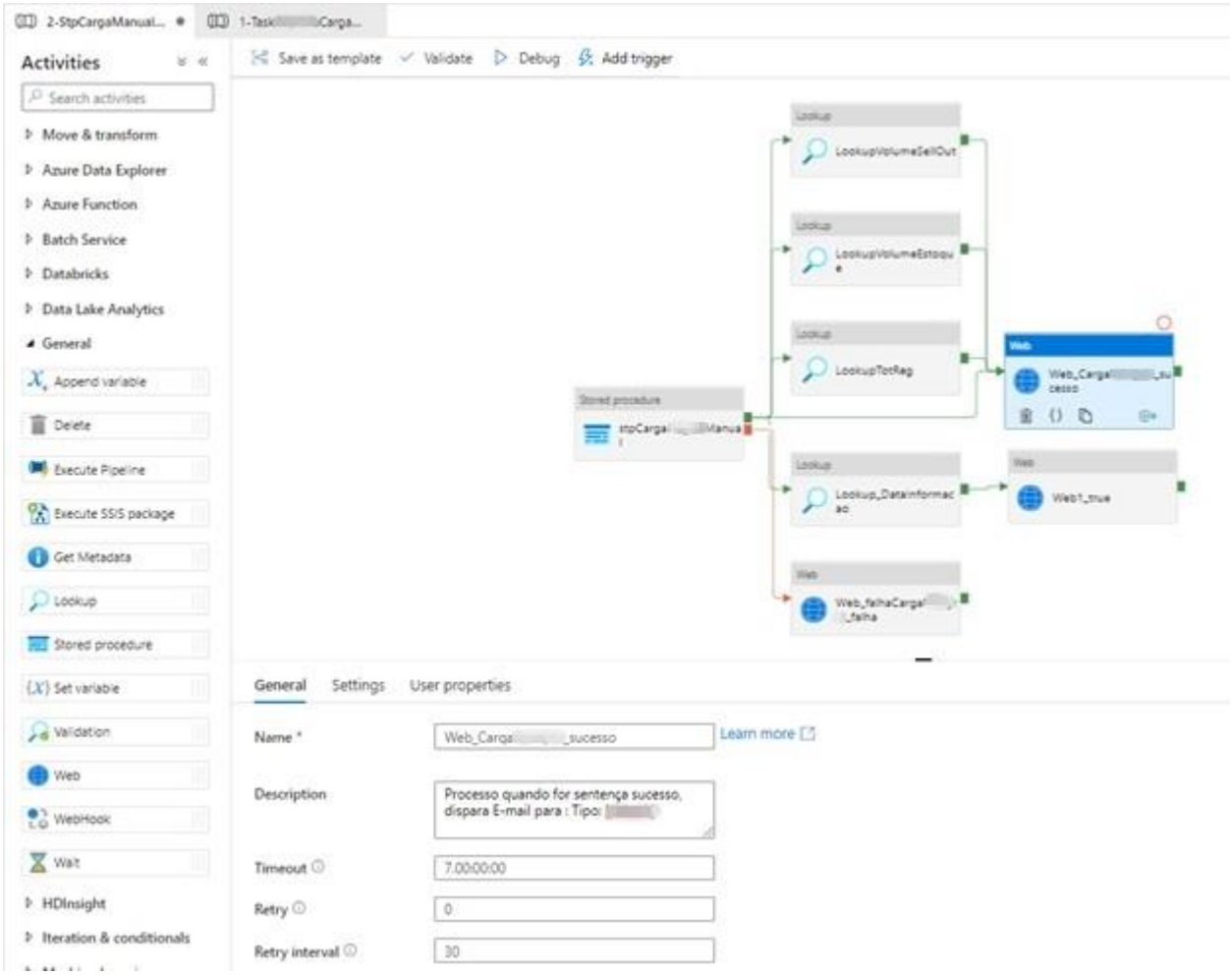

Figura 8 - Configuração da Tarefa de E-mail

4. Conclusão

O Data Warehouse tem como objetivo o auxílio na tomada de decisão de qualquer organização em que seja utilizado, e para isso, um processo de ETL adequadamente preparado é fundamental. Contudo, alguns aspectos devem ser considerados para minimizar as dificuldades inerentes à construção dessas aplicações e o Azure Data Factory é uma nova ferramenta disponível no mercado justamente com essa proposta.

O Azure Data Factory demonstrou ser uma ferramenta que permite uma fácil configuração, efetuando inúmeras integrações e conexões com as mais diversas fontes e as principais origens de dados utilizadas atualmente no mercado. A interface do ADF, além de ser amigável é tudo o que já esperamos de um produto da Microsoft: integrada, robusta, organizada e de fácil interação, não demandando muito esforço com programação e configuração, onde é possível incluir um novo processo apenas arrastando um objeto.

O presente artigo serve de exemplo para apoiar o desenvolvimento de outros ambientes para o suporte à tomada de decisão, que incluam a utilização de ferramentas ETL

que ainda são muito úteis, principalmente com o surgimento de novos produtos oferecidos totalmente em nuvem como o Azure Data Factory.

Em trabalhos futuros pretende-se analisar a performance e agilidade da ferramenta comparada com outras correlatas no mercado. Além disso, a implementação do ADF em processos de outras áreas da organização, como logística e suprimentos também está planejada.

Referências Bibliográficas

ABREU, Fábio Silva Gomes da Gama. Estudo de usabilidade do software Talend Open Studio como ferramenta padrão para ETL dos sistemas-clientes da aplicação PostGeoOlap. *Monografia de Graduação em Sistemas de Informação–Faculdade Salesiana Maria Auxiliadora*, Macaé, 2007.

CASTELLS, Manuel.A sociedade em rede. São Paulo: Paz e Terra, *Wagner Junqueira Prado*, 2003.

EL-SAPPAGH, Shaker H. Ali; HENDAWI, Abdeltawab M. Ahmed; EL BASTAWISSY, Ali Hamed. A proposed model for data warehouse ETL processes. *Journal of King Saud University-Computer and Information Sciences*, v. 23, n. 2, p. 91-104, 2011.

FERREIRA, João et al. O processo etl em sistemas data warehouse. *In: INForum*. 2010. p. 757-765.

KIMBALL, R. et al. The Data Warehouse *Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses*. Wiley, 1998

INMON, William H.; ZACHMAN, John A.; GEIGER, Jonathan G. Data stores, data warehousing and the Zachman framework: managing enterprise knowledge. McGraw-Hill, Inc., 1997.

JACKSON, Lacey. 4 Ways to Improve Your Data Management. *Disponível em: https://www.formstack.com/resources/blog-data-management-tips*. Acesso em: 12 mar. 2021.

MICROSOFT. Documentação do Azure. Disponível em: *https://docs.microsoft.com/pt-br/azure/?product=featured*. Acesso em: 02 mar. 2021.

PANOPLY. Data Warehouse Architecture: Traditional vs. Cloud. Disponível em: *https://panoply.io/data-warehouse-guide/data-warehouse-architecture-traditional-vs-cloud/*. Acesso em: 10 mar. 2021.

PAPASTEFANATOS, George et al. Metrics for the prediction of evolution impact in etl ecosystems: A case study. *Journal on Data Semantics*, v. 1, n. 2, p. 75-97, 2012.

POLLONI, Enrico Giulio Franco. Administrando sistemas de informação: estudo de viabilidade. *Futura*, 2000.