JOURNAL
COMPUTER SCIENCES INSTITUTE

JCSI 19 (2021) 126–131
Received: 19 April 2021
Accepted: 26 April 2021

The comparative analysis of modern ETL tools

Analiza porównawcza współczesnych narzędzi ETL

Ivan Falchuk, Vitalii Mayuk, Piotr Muryjas*

Department of Computer Science, Lublin University of Technology, Nadbystrzycka 36B, 20-618 Lublin, Poland

Abstract

Each data warehouse requires loading properly processed transactional data. The process that performs this task is known as extract-transform-load (ETL). The efficiency of its implementation affects how quickly the user will have the access to the current analytical data. The paper presents the results of research efficiency of ETL performance of its stage with the use of Azure Synapse (AS) and Azure Data Factory (ADF). The research included selection, sorting and aggregating data, joining tables, and loading data into target tables. To evaluate the efficiency of these operations, the criterion of their execution time has been used. The obtained results indicate that the ADF tool provides a much higher time efficiency of loading transactional data into the data warehouse comparing to AS.

Keywords: Azure Synapse; Azure Data Factory; ETL tools

Streszczenie

Każda hurtownia danych wymaga ładowania odpowiednio przetworzonych danych transakcyjnych. Procesy realizujące to zadanie określane są jako ekstrakcja-transformacja-ładowanie (ETL). Od efektywności ich wykonania zależy jak szybko użytkownik będzie miał dostęp do bieżących danych analitycznych. W artykule przedstawiono istotę procesu ETL oraz wyniki badań efektywności realizacji jego etapów z użyciem Azure Synapse (AS) oraz Azure Data Factory (ADF). Badania obejmowały selekcję, sortowanie i agregację danych, złączenie tabel oraz zapis danych do tabel docelowych. Do oceny efektywności tych operacji zastosowano kryterium czasu ich wykonania. Uzyskane wyniki wskazują, iż narzędzie ADF zapewnia znacznie wyższą efektywność czasową ładowania danych transakcyjnych do hurtowni danych w porównaniu do AS.

Słowa kluczowe: Azure Synapse; Azure Data Factory; ETL tools

*Corresponding author

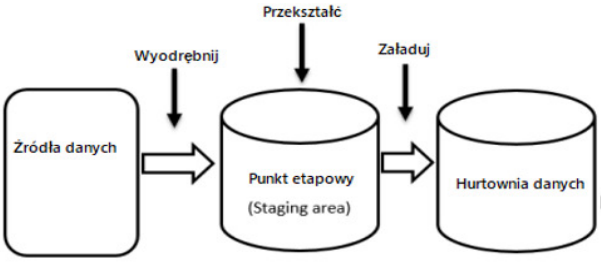
Email address: p.muryjas@pollub.pl (P. Muryjas)

©Published under Creative Common License (CC BY-SA v4.0)

1. Wstęp

Każda współczesna organizacja staje dziś przed wyzwaniem przechowywania dużej ilości danych o zróżnicowanej jakości oraz ich efektywnej analizy. Transakcyjna postać tych danych, pochodzących z systemów ERP, nie jest jednak właściwa dla osób, które podejmują decyzje i muszą posługiwać się danymi o pewnym stopniu agregacji. Z tego względu przedsiębiorstwa coraz częściej korzystają z systemów klasy business intelligence (BI), które znacznie zwiększają efektywność analizy danych oraz zaspokajają potrzeby analityczne osób decyzyjnych. Głównym źródłem danych dla tego rodzaju systemów jest hurtownia danych. Jest to miejsce składowania danych analitycznych, które są używane do dokonywania bardziej trafnych wyborów w procesach decyzyjnych [1].

Podejmowanie decyzji wymaga posiadania informacji o aktualnej sytuacji w organizacji. Dlatego też konieczne jest, by dane w hurtowniach były okresowo uzupełniane i aktualizowane. Zadanie to wykonywane jest w ramach działań określanych jako ekstrakcja, transformacja i ładowanie danych (ETL – Extract, Transform, Load), które opisują sposób pobierania danych ze źródeł, logikę ich transformacji zgodną z wymaganiami biznesowymi oraz miejsce ich docelowego składowania w hurtowni danych. Ideę procesu ETL przedstawiono na Rysunku 1.



Rysunek 1: Idea procesu ETL.

Etap ekstrakcji danych polega na ich wyodrębnieniu ze źródeł i zapisaniu w miejscu określanym jako obszar pośredni (staging area). W tym etapie dane nie zmieniają jeszcze swojej pierwotnej, transakcyjnej postaci. W procesie transformacji, który jest realizowany w obszarze pośrednim, dane są oczyszczane, integrowane, grupowane, a także konwertowane do wymaganej postaci, zgodnej ze strukturą przechowywania danych w hurtowni. Na etapie ładowania dane są pobierane z obszaru pośredniego i bezpośrednio umieszczane w tabelach hurtowni danych. Należy zauważyć, że w procesie ETL nie wszystkie dane transakcyjne są brane pod uwagę, ale tylko te, które są nowe lub zostały zmienione [2].

W niniejszej publikacji dokonano analizy wydajności dwóch współczesnych narzędzi ETL firmy Micro-

soft, tj. Azure Data Factory i Azure Synapse. Wykorzystując różne scenariusze procesów ETL i ten sam źródłowy system transakcyjny, przeprowadzono badania, których rezultaty umożliwiły wskazanie narzędzia zapewniającego wyższą efektywność realizacji tych procesów.

2. Charakterystyka narzędzi ETL firmy Microsoft

Azure Data Factory (ADF) to usługa ETL w chmurze platformy Microsoft Azure umożliwiająca skalowalną w poziomie integrację danych i ich transformację bez użycia odrębnego, dedykowanego serwera. Usługa ta dostępna jest poprzez interfejs użytkownika zapewniający intuicyjne tworzenie, monitorowanie i zarządzanie procesem ETL w jednym panelu [3]. Narzędzie ADF zapewnia migrację danych między wieloma lokalnymi i chmurowymi źródłami danych. Lista obsługiwanych platform jest rozbudowana i obejmuje zarówno rozwiązania firmy Microsoft jak i innych dostawców. Jest to zaawansowane narzędzie zapewniające pełną elastyczność przenoszenia ustrukturyzowanych i nieustrukturyzowanych zestawów danych, w tym RDBMS, XML, JSON i różnych magazynów danych typu NoSQL. Jego podstawową zaletą jest duża elastyczność wynikająca z możliwości wykorzystania języka U-SQL lub HiveQL [4-5].

Azure Synapse (AS) jest kolejną usługą wykorzystywaną zarówno do realizacji procesów ETL jak i do analizy danych. Z jej pomocą możliwa jest integracja i transformacja danych transakcyjnych jak również danych typu big data. Narzędzie AS zapewnia wykonywanie zapytań do różnych baz danych bez konieczności użycia zasobów serwerowych. W usłudze tej rozproszone zasoby danych źródłowych są łączone w ujednolicone środowisko umożliwiające pozyskiwanie, eksplorowanie, przygotowywanie i udostępnianie danych oraz zarządzanie nimi na potrzeby analizy biznesowej oraz uczenia maszynowego.

Azure Synapse zapewnia wysoką elastyczność łączenia danych nierelacyjnych oraz relacyjnych. Narzędzie umożliwia łatwy podgląd danych za pomocą wysyłania zapytań do bazy danych Azure SQL. Tworząc przepływy danych przy pomocy predefiniowanych bloków zadaniowych, można zaimplementować zaawansowane scenariusze transformacji i analizy danych bez konieczności pisania kodu źródłowego [6].

Azure Synapse posiada wbudowaną bezserwerową pulę SQL, czyli logiczną przestrzeń, w której udostępniane są w elastyczny sposób różne zasoby umożliwiające efektywną realizację procesów ETL. Każdy obszar roboczy posiada rozbudowaną pulę SQL bez możliwości usunięcia takiej puli. Bezserwerowe pule SQL zapewniają możliwość użycia języka SQL bez konieczności rezerwowania określonej wielkości zasobów pamięci. W przeciwieństwie do dedykowanych pul SQL, rozliczanie dla bezserwerowej puli SQL jest oparte na ilości danych przeskanowanych w celu uruchomienia zapytania, a nie pojemności przydzielonej do puli [7].

3. Metodyka badań

Porównanie wydajności badanych narzędzi ETL wymagało przeprowadzeniu operacji ekstrakcji, transformacji i ładowania danych. Dokonano pomiaru czasu wykonania podstawowych operacji ETL oraz dodatkowo czasu napełniania tabeli faktu w przykładowej hurtowni danych.

Na potrzeby badań, w każdym z narzędzi utworzony został zasób bezserwerowej bazy danych Azure SQL Database o pojemności 5 GB oraz wydajności 100 DTU (Data Transaction Unit). DTU jest to pakiet miary zasobów obliczeniowych (przydzielonych do bazy danych), operacji we/wy i magazynu [8]. W każdej tabeli, stanowiącej źródło danych dla procesów ETL, zostały utworzone klastrowane indeksy kolumnowe w celu zwiększenia wydajności zapytań SQL.

Jako szczegółowe kryteria oceny wydajności narzędzi w każdym scenariuszu badawczym przyjęto:

- czas uruchamiania klastra – czas potrzebny na uruchomienie izolowanego klastra Apache Spark, który jest używany podczas uruchamiania przepływów danych [9];
- czas ekstrakcji danych ze źródła – czas potrzebny na odczyt danych z tabeli źródłowej do pierwszego bloku zadaniowego w przepływie danych;
- czas transformacji danych – czas potrzebny do wykonania operacji przekształcenia oraz czyszczenia danych w przepływie;
- czas zapisywania danych w zasobie docelowym - czas potrzebny na załadowanie danych do tabeli docelowej (np. tabela wymiaru lub faktu w hurtowni danych).

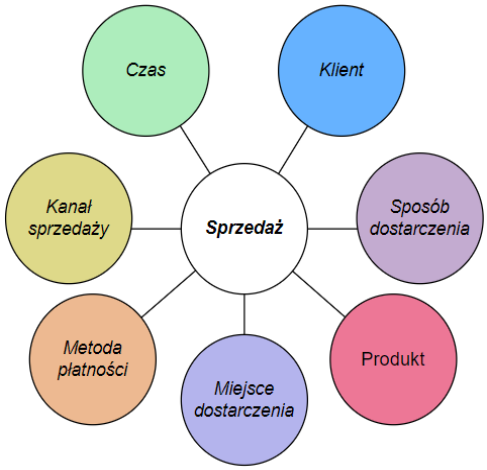
W każdym z narzędzi zostało utworzonych 9 przepływów danych opisujących wykonanie podstawowych operacji przekształcenia danych w procesie ETL. Wykaz tych operacji oraz ich szczegółowy opis biznesowy zostały przedstawione w Tabeli 1.

Tabela 1: Operacje ETL w scenariuszach badawczych

Lp.	Operacja	Opis
1	Selekcja wszystkich kolumn	Selekcja wszystkich kolumn z tabeli zawierającej dane o klientach.
2	Selekcja wybranych kolumn	Selekcja wybranych kolumn z tabeli zawierającej dane o klientach.
3	Sortowanie	Sortowanie danych w tabeli zawierającej dane o produktach.
4	Zliczanie	Zliczanie faktur wszystkich klientów.
5	Zliczanie unikalnych wartości	Zliczanie liczby sprzedanych produktów.
6	Sumowanie	Sumowanie wartości sprzedanych produktów.
7	Złączenie	Złączenie danych o nagłówkach i pozycjach faktur sprzedaży.
8	Złączenie + Sumowanie	Złączenie danych o produktach i pozycjach faktur oraz sumowanie wartości sprzedaży według grupy zapotrzebowania.

9	Filtrowanie	Wyszukiwanie faktur korygujących z podanym kodem przyczyny korekty.

Dodatkowo, w ramach przeprowadzonych badań, zbudowano 6 przepływów danych realizujących zadanie napełnienia tabel w hurtowni danych. Pięć przepływów danych służyło do napełnienia tabel wymiarów oraz jeden – do napełniania tabeli faktu opisującego sprzedaż. Schemat konceptualny wykorzystanej hurtowni danych został przedstawiony na Rysunku 2.



Rysunek 2: Schemat konceptualny hurtowni danych wykorzystanej w badaniach.

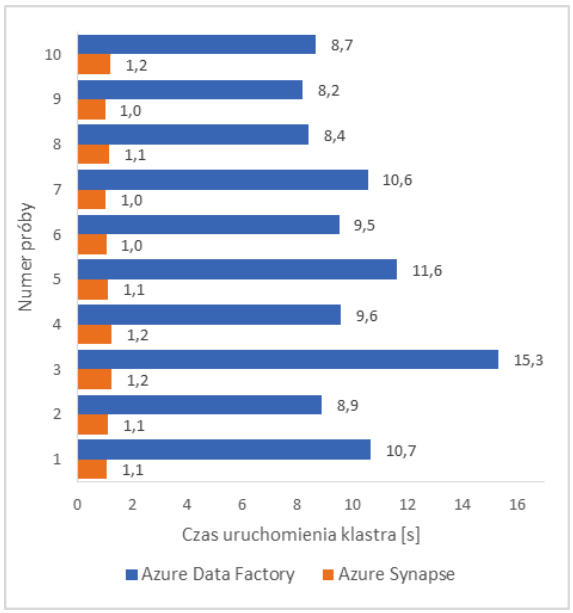
Każdy z przepływów danych, tj. 9 do wykonania podstawowych operacji ETL i 1 do napełniania tabeli faktu, został uruchomiony 10 razy. Wyniki przeprowadzonych badań zostały zapisane w plikach *csv*. Czas wykonania badanych operacji był mierzony przy pomocy wbudowanego w każde z narzędzi środowiska monitorowania, które podaje szczegółowe informacje o wykonanych przepływach danych [10].

4. Wyniki badań

Wyniki przeprowadzonych badan wydajnościowych zostaną przedstawione dla każdego kryterium czasowego wymienionego w rozdziale 3. Dodatkowo, na wstępie rozdziału zaprezentowano rezultaty badań dotyczących szybkości uruchomienia klastra Apache Spark, w którym zlokalizowane są dane źródłowe. Czas wykonania tej operacji ma istotny wpływ na całkowity czas realizacji procesu ETL. Dlatego też zasadne jest, aby określić jego wielkość dla poszczególnych narzędzi.

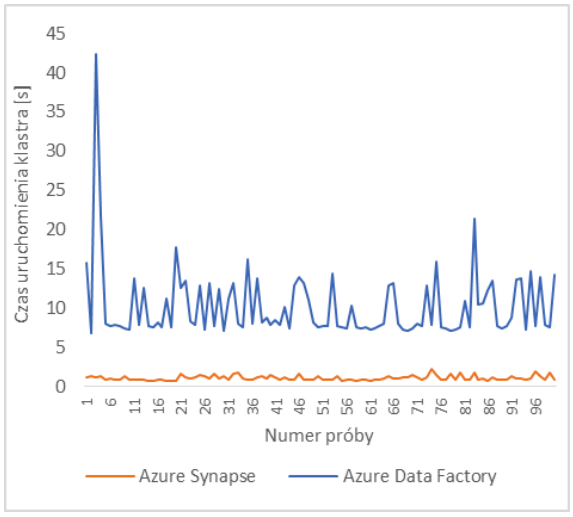
4.1. Uruchomienie klastra

Na Rysunku 3 przedstawiono wyniki badań dotyczących szybkości uruchomienia klastra Apache Spark przez ten sam proces ETL wykonywany 10-cio krotnie w każdym z badanych narzędzi. Można zauważyć, że czas uruchomienia klastra w narzędziu Azure Synapse jest średnio 9 razy krótszy niż w środowisku Azure Data Factory.



Rysunek 3: Czas uruchomienia klastra Apache Spark.

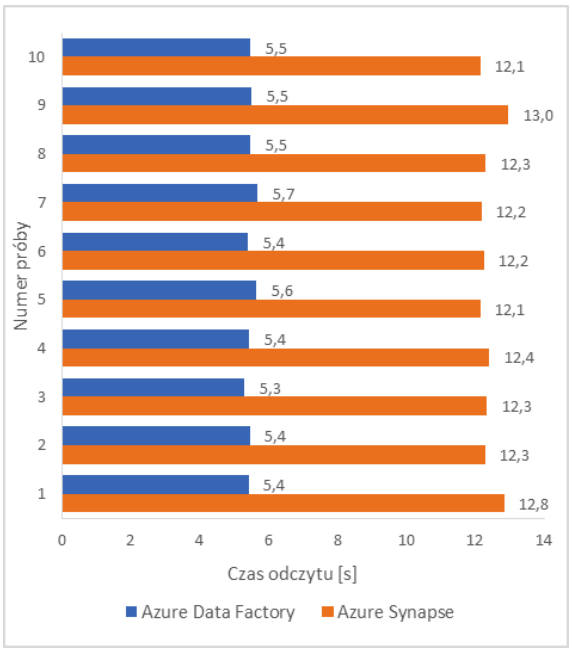
Ponadto, na podstawie 10-krotnego powtórzenia tego badania dla wszystkich 10-ciu scenariuszy można stwierdzić, że klaster używany w narzędziu Azure Synapse jest bardziej stabilny niż w narzędziu Azure Data Factory (Rysunek 4).



Rysunek 4: Czasy wielokrotnego uruchomienia klastra dla wszystkich scenariuszy badawczych.

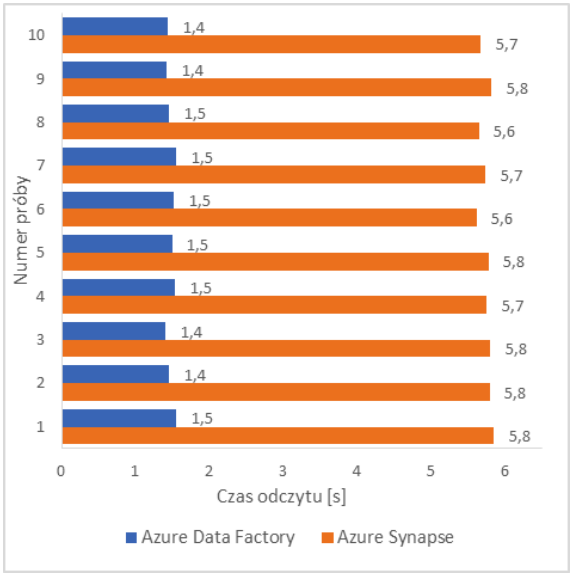
4.2. Ekstrakcja danych ze źródła

Kolejne badanie miało na celu identyfikację wpływu liczności kolumn w zapytaniu do tabeli źródłowej na długość czasu jego wykonania. Operacje odczytu wykonano z użyciem tabeli zawierającej 49 383 rekordy opisujące klientów. W pierwszej kolejności zmierzono czas odczytu danych ze wszystkich szesnastu kolumn tej tabeli (Rysunek 5), a następnie z pięciu wybranych (Rysunek 6).



Rysunek 5: Czasy odczytów danych ze wszystkich kolumn tabeli *Klienci*.

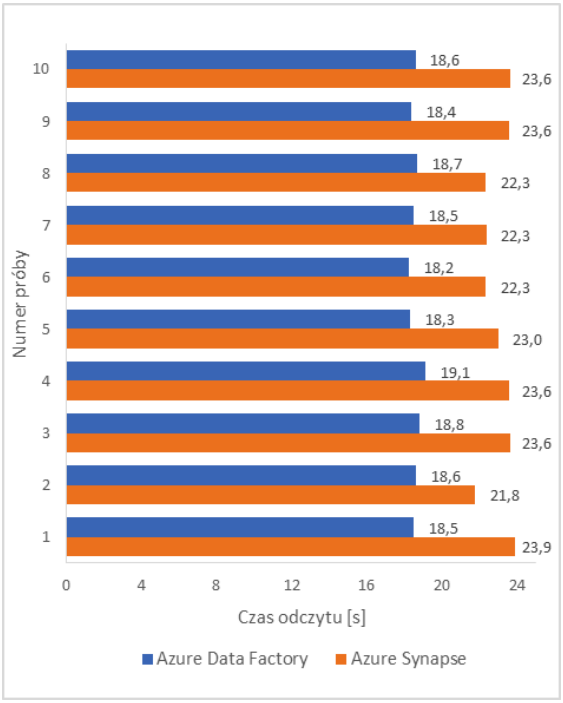
Na podstawie otrzymanych wyników badań można stwierdzić, że czas odczytu danych ze wszystkich kolumn tabeli źródłowej jest ponad 2 razy dłuższy w środowisku Azure Synapse niż w Azure Data Factory.



Rysunek 6: Czasy odczytów z wybranych 5 kolumn tabeli *Klienci*.

Natomiast zapytanie zawierające 5 wybranych kolumn tabeli źródłowej, przy tej samej liczbie rekordów, wykonuje się niemal 4 razy szybciej w narzędziu Azure Data Factory niż w Azure Synapse.

Na podstawie przeprowadzonych eksperymentów stwierdzono, że zwiększenie liczby odczytywanych rekordów i kolumn z tabeli źródłowej prowadzi do zmniejszenia różnicy w czasie wykonania zapytań w badanych środowiskach. W przypadku ekstrakcji danych z tabeli opisującej produkty, zawierającej 921 855 rekordy i 41 kolumn, zaobserwowano, iż zapytanie w środowisku Azure Data Factory wykonywało się już tylko 1,3 razy szybciej niż w narzędziu Azure Synapse.



Rysunek 7: Czasy odczytu wszystkich kolumn z tabeli *Produkty*.

4.3. Transformacja danych

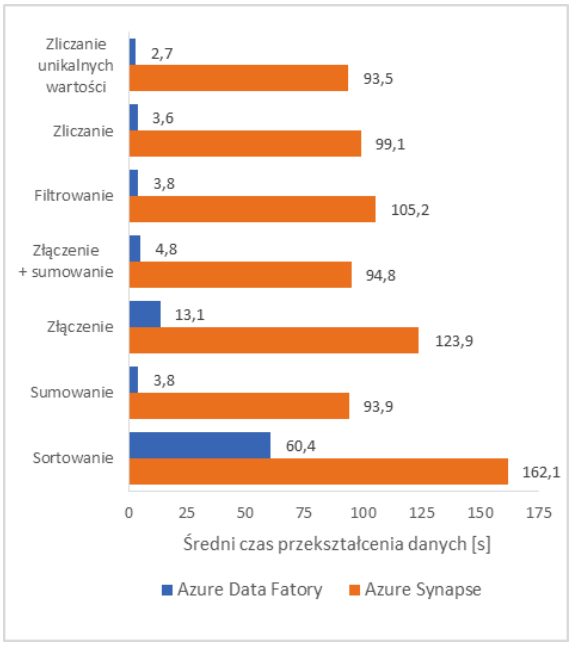
W następnym etapie badań dokonano porównania średniego czasu przekształcenia danych w poszczególnych środowiskach. Pod uwagę wzięto następujące operacje:

- sortowanie,
- zliczanie,
- zliczanie unikalnych wartości,
- sumowanie,
- złączenie,
- złączenie + sumowanie,
- filtrowanie.

Sortowanie danych zostało wykonane z użyciem danych z tabeli *Produkty*, zawierającej 921 855 rekordów. Źródłem danych dla operacji agregacji, tj. zliczania rekordów, zliczania unikalnych wartości i sumowania, oraz złączenia danych była tabela opisująca nagłówek faktury sprzedażowej oraz tabela opisująca poszczególne pozycje tych faktur. Zawierały one odpowiednio 44 570 i 66 995 rekordów. W operacji filtrowania danych została ponownie użyta tabela nagłówków faktur sprzedaży oraz tabela kodów przyczyny zwrotu produktu, zawierająca 5 018 rekordów. Operację złączenia i agregacji danych przeprowadzono z wykorzystaniem tabeli *Produkty* i tabeli z pozycjami faktur sprzedaży.

Uzyskane czasy 10-ciokrotnego wykonania poszczególnych operacji zostały uśrednione i przedstawione na Rysunku 8. Na podstawie przeprowadzonych badań można stwierdzić, że wszystkie działania przekształcenia danych wykonują się znacznie szybciej w środowisku Azure Data Factory niż w Azure Synapse, przy czym najmniejsza różnica średnich czasów

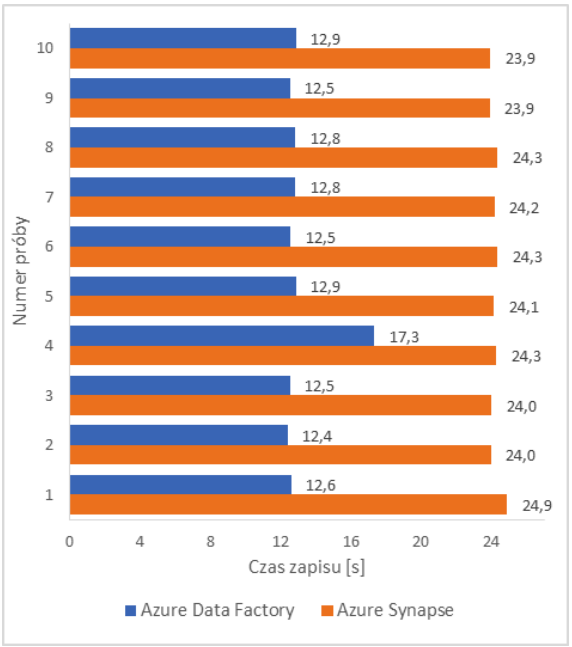
wykonania dotyczy operacji sortowania rekordów (ADF jest niemal 3-krotnie szybszy niż AS), natomiast największa różnica występuje w przypadku operacji zliczania unikalnych wartości (ADF jest 35 razy szybszy od AS).



Rysunek 8: Średni czas transformacji danych dla różnych operacji.

4.4. Zapis danych w zasobie docelowym

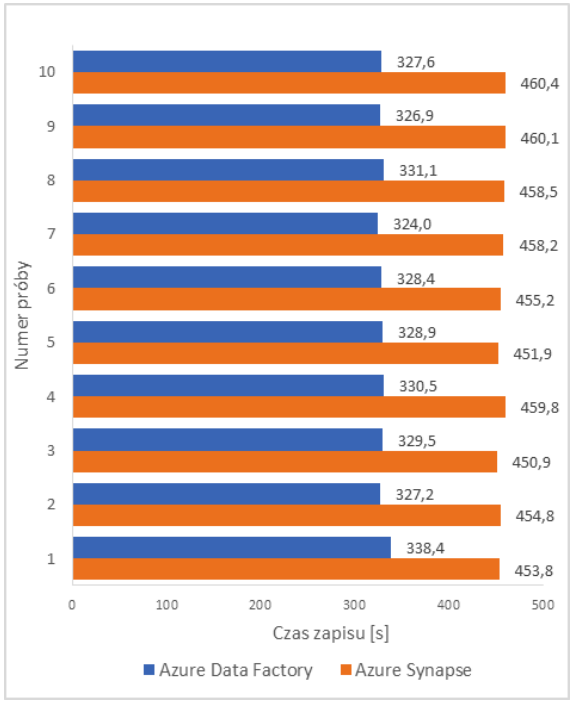
Kolejnym etapem badań była identyfikacja długości czasu ładowania danych do docelowych tabel wymiarów i faktu przy użyciu analizowanych narzędzi. W wyniku transformacji transakcyjnych danych o klientach, do tabeli wymiaru *Klient* załadowano 49 383 rekordy. Czasy wykonania serii 10-ciu prób tej operacji zostały przedstawione na Rysunku 9.



Rysunek 9: Czasy zapisu danych do tabeli wymiaru *Klient*.

Na podstawie otrzymanych rezultatów badań można stwierdzić, że ładowanie danych z użyciem narzędzia Azure Data Factory odbywa się dwukrotnie szybciej niż przy wykorzystaniu środowiska Azure Synapse.

W celu zweryfikowania wpływu liczby rekordów na czas wykonania ładowania danych, przeprowadzono eksperyment polegający na wstawieniu 921 855 rekordów do tabeli wymiaru *Produkt*. Wyniki serii 10-ciu prób tej operacji zobrazowano na Rysunku 10.



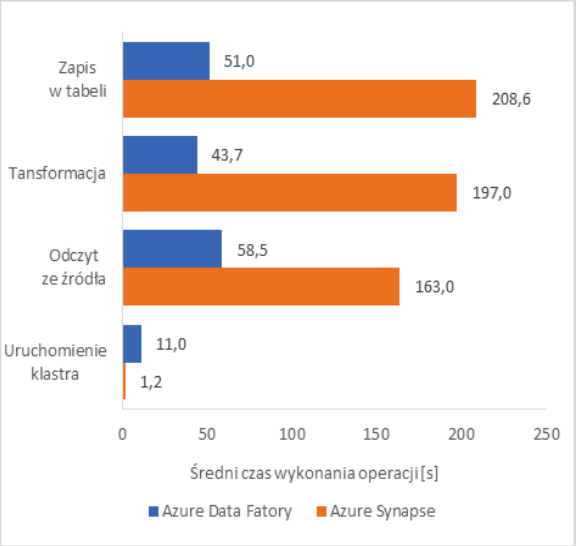
Rysunek 10: Czasy zapisu danych do tabeli wymiaru *Produkt*.

Otrzymane wyniki badania potwierdzają przewagę Azure Data Factory na Azure Synapse również w przypadku ładowania bardzo dużych zbiorów danych. Jednak warto zauważyć znaczne zmniejszenie się różnicy pomiędzy czasami wykonania tej operacji z użyciem wspomnianych narzędzi w miarę wzrostu liczby ładowanych rekordów. Obecnie różnica ta wynosi 39%, podczas gdy dla mniejszej ilości danych była ona równa 91%.

W ostatnim etapie prac badawczych dokonano analizy czasów trwania pełnego cyklu ładowania danych do tabeli faktu *Sprzedaz* z użyciem prezentowanych narzędzi ETL. Uwzględniono tutaj wszystkie operacje, począwszy od pobrania danych z tabel transakcyjnych, opisujących faktury i ich pozycje, poprzez ich transformację do postaci analitycznej, a skończywszy na załadowaniu zagregowanych danych do tabeli faktu. Dodatkowo, w badaniu tym wykorzystano uprzednio utworzone tabele wymiarów *Klient* i *Produkt*. Cykl ładowania został powtórzony 10-cio krotnie w każdym środowisku. W kolejnych powtórzeniach tego procesu zmierzono czasy trwania każdej operacji, a następnie dokonano ich uśrednienia. Otrzymane wyniki zostały zaprezentowane na Rysunku 11.

Na podstawie otrzymanych rezultatów badań można stwierdzić, że proces ETL, odpowiedzialny za ładowa-

nie danych do tabeli faktu, jest realizowany znacznie efektywniej z wykorzystaniem narzędzia Azure Data Factory niż z użyciem Azure Synapse.



Rysunek 11: Średni czas realizacji etapów procesu ETL podczas ładowania danych do tabeli faktu *Sprzedaz*.

Przewaga Azure Synapse jest dostrzegalna tylko w fazie uruchamiania klastra z danymi źródłowymi. Jednak wszystkie dalsze etapy procesu ETL są wykonywane 3-4 krotnie szybciej w środowisku Azure Data Factory.

5. Wnioski

Celem niniejszego artykułu była analiza porównawcza wydajności narzędzi ETL Azure Synapse oraz Azure Data Factory oferowanych przez firmę Microsoft. Na potrzeby badań utworzono analityczną bazę danych, w której alokowano dane przetworzone zgodnie z przyjętymi scenariuszami badawczymi, uwzględniającymi wszystkie typowe operacje realizowane w ramach działań ETL.

Uzyskane rezultaty badań pozwalają jednoznacznie stwierdzić, iż w porównaniu do Azure Synapse, narzędzie Azure Data Factory zapewnia znacznie wyższą efektywność czasową ładowania danych z systemów transakcyjnych do hurtowni danych. Przewaga ADF nad AS jest widoczna zarówno podczas wykonywania zasilania tabel wymiarów jak i tabel faktów. Szczegółowa analiza czasów wykonania poszczególnych etapów pełnego cyklu ETL pozwala dostrzec znaczne różnice w ich wielkości na korzyść ADF. Wprawdzie Azure Synapse zapewnia krótszy czas uruchomienia klastra z danymi źródłowymi, to jednak proces ten jest realizowany tylko raz, na początku procesu ETL, w celu zapewnienia dostępu do danych transakcyjnych. Z punktu widzenia całości przetwarzania danych w ramach ETL, najistotniejsza jest efektywność czasowa wykonania transformacji danych i ich ładowania do tabel docelowych, a w tym obszarze Azure Data Factory dominuje nad Azure Synapse.

Warto jednak zauważyć, że przewaga ADF nad AS maleje wraz ze wzrostem liczby przetwarzanych rekordów. Oznacza to, że wybór właściwego narzędzia do realizacji zadań ETL, które zapewni wysoką efektywność tego działania, powinien być poprzedzony głęboką analizą specyfiki źródłowych tabel relacyjnych.

Literatura

[1] Ł. Bielak, P. Muryjas, Integracja Big Data i Business Intelligence jako innowacyjne rozwiązanie wspomagające funkcjonowanie nowoczesnych organizacji, Journal of Computer Sciences Institute 1 (2016) 6–13.

[2] А. С. Черняев, ETL: обзор инструментов, Молодой ученый, 1 (2019) 23–26, https://moluch.ru/archive/239/55368/, [16.04.2021].

[3] Azure Data Factory documentation, https://docs.microsoft.com/en-us/azure/data-factory/, [16.04.2021].

[4] R. Sudhir, A. Narain, Understanding Azure Data Factory: Operationalizing Big Data and Advanced Analytics Solutions, Apress, Berkeley, 2019.

[5] A. Leonard, K. Bradshaw, SQL Server Data Automation Through Frameworks. Building Metadata-Driven Frameworks with T-SQL, SSIS, and Azure Data Factory, Apress, Berkeley, 2020.

[6] Dokumentacja narzędzia Azure Synapse Analytics, https://azure.microsoft.com/pl-pl/services/synapse-analytics/, [16.04.2021].

[7] Architektura dedykowanej puli SQL (dawniej SQL DW) w usłudze Azure Synapse Analytics, https://docs.microsoft.com/pl-pl/azure/synapse-analytics/sql-data-warehouse/massively-parallel-processing-mpp-architecture, [16.04.2021].

[8] Wybór między modelami zakupów rdzeń wirtualny i DTU — Azure SQL Database i wystąpienie zarządzane SQL, https://docs.microsoft.com/pl-pl/azure/azure-sql/database/purchasing-models#dtu-based-purchasing-model, [16.04.2021].

[9] Przewodnik dotyczący wydajności i dostrajania przepływu danych, https://docs.microsoft.com/pl-pl/azure/data-factory/concepts-data-flow-performance, [16.04.2021].

[10] Monitorowanie przepływów danych, https://docs.microsoft.com/pl-pl/azure/data-factory/concepts-data-flow-monitoring, [16.04.2021].