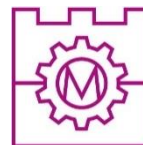




POLITECHNIKA KRAKOWSKA im. T. Kościuszki
Wydział Mechaniczny
Katedra Informatyki Stosowanej



Kierunek studiów: Informatyka Stosowana

STUDIA STACJONARNE

PRACA DYPLOMOWA

INŻYNIERSKA

Krystian Dutka

BUDOWA EKOSYSTEMU BUSINESS INTELLIGENCE
W OPARCIU O ŚRODOWISKA BIG DATA

BUILDING A BUSINESS INTELLIGENCE ECOSYSTEM
BASED ON BIG DATA ENVIRONMENTS

Promotor:
Mgr inż. **Marek Lewiński**

Kraków, rok akad. 2022/2023

OŚWIADCZENIE O SAMODZIELNYM WYKONANIU PRACY DYPLOMOWEJ

Oświadczam, że przedkładana przeze mnie praca dyplomowa inżynierska została napisana przeze mnie samodzielnie. Jednocześnie oświadczam, że ww. praca:

- 1) nie narusza praw autorskich w rozumieniu ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych (Dz.U. z 2021 r. poz. 1062) oraz dóbr osobistych chronionych prawem cywilnym, a także nie zawiera danych i informacji, które uzyskałem/am* w sposób niedozwolony,
- 2) nie była wcześniej podstawą żadnej innej procedury związanej z nadawaniem tytułów zawodowych, stopni lub tytułów naukowych.

Jednocześnie wyrażam zgodę na:

- 1) poddanie mojej pracy kontroli za pomocą systemu Antyplagiat oraz na umieszczenie tekstu pracy w bazie danych uczelni, w celu ochrony go przed nieuprawnionym wykorzystaniem. Oświadczam, że zostałem/~~am~~ poinformowany/~~a~~ i wyrażam zgodę, by system Antyplagiat porównywał tekst mojej pracy z tekstem innych prac znajdujących się w bazie danych uczelni, z tekstami dostępnymi w zasobach światowego Internetu oraz z bazą porównawczą systemu Antyplagiat,
- 2) to, aby moja praca pozostała w bazie danych uczelni przez okres wynikający z przepisów prawa. Oświadczam, że zostałem poinformowany i wyrażam zgodę, że tekst mojej pracy stanie się elementem porównawczej bazy danych uczelni, która będzie wykorzystywana, w tym także udostępniana innym podmiotom, na zasadach określonych przez uczelnię, w celu dokonywania kontroli antyplagiatowej prac dyplomowych/doktorskich, a także innych tekstów, które powstaną w przyszłości.


.....
podpis

- 1) Wyrażam zgodę na udostępnianie mojej pracy dyplomowej w Akademickim Systemie Archiwizacji Prac na PK do celów naukowo-badawczych z poszanowaniem przepisów ustawy o prawie autorskim i prawach pokrewnych (Dz.U. z 2021 r. poz. 1062)..

TAK/NIE*

.....
podpis

Jednocześnie przyjmuję do wiadomości, że w przypadku stwierdzenia popełnienia przeze mnie czynu polegającego na przypisaniu sobie autorstwa istotnego fragmentu lub innych elementów cudzej pracy, lub ustalenia naukowego, Rektor PK stwierdzi nieważność postępowania w sprawie nadania mi tytułu zawodowego (art. 77 ust. 5 ustawy z dnia 18 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce, (Dz.U. z 2021 r., poz. 478, z późn. zm.)).


.....
podpis

Spis treści

1. CEL I ZAKRES PRACY	6
2. WSTĘP	7
3. WPROWADZENIE TEORETYCZNE	8
3.1. Przetwarzanie danych.....	8
3.1.1. Systemy OLTP/OLAP	8
3.1.2. Proces ETL	9
3.1.3. Hurtowanie danych	9
3.2. Big Data.....	11
3.2.1. Zastosowanie systemów Big Data	11
3.2.2. Opis i architektura systemów Big Data.....	13
3.3. Apache Spark.....	16
3.3.1. Architektura Apache Spark	16
3.3.2. Ekosystem Apache Spark.....	18
3.4. Business Intelligence	19
3.4.1. Charakterystyka i zastosowanie Business Intelligence	19
3.4.2. Platforma Tableau	23
3.4.3. Platforma Apache Superset	24
3.4.4. Platforma Power Bi	26
4. CZĘŚĆ PROJEKTOWA	29
4.1. Założenie budowy ekosystemu BI z wykorzystaniem Big Data	29
4.1.1. Uzasadnienie biznesowe dla utworzenia projektu	29
4.1.2. Identyfikacja źródeł danych	30
4.1.3. Identyfikacja wymagań początkowych	33
4.1.4. Identyfikacja wymagań końcowych.....	33
4.2. Architektura z doбором technologii	35
4.2.1 Architektury ekosystemu BI	35
4.2.2 Harmonogramowanie, zarządzanie metadanymi oraz administracja bezpieczeństwem ekosystemu	35
4.2.3 Przetwarzanie danych w ekosystemie	36
4.2.4 Magazyn danych i organizacja danych	37
4.2.5 Udostępnianie oraz analiza danych.....	37
4.2.6 Diagram architektury oraz tabela użytych technologii	38
4.3. Opis implementacji.....	41
4.3.1 Przetwarzanie danych w Apache Spark	41
4.3.2 Analiza danych w Power Bi	45
4.3.3 Podsumowanie	52
5. WNIOSKI	53
LITERATURA	54
SUMMARY	55

1. Cel i zakres pracy

Przedstawiona praca inżynierska ma na celu zaprezentowanie kluczowych cech koncepcji budowania architektury ekosystemu Business Intelligence z wykorzystaniem Big Data. Ekosystem został stworzony do przetwarzania, wizualizacji oraz analizy danych. W celu realizacji projektu architektury zostały wykorzystane koncepcje zaprezentowane w części teoretycznej. Wprowadzenie teoretyczne uwzględnia cztery aspekty:

- sposoby przetwarzania danych,
- budowa architektury Big Data,
- przedstawienie środowiska Apache Spark,
- koncepcja Business Intelligence oraz narzędzia do wizualizacji danych.

W celu zademonstrowania wykorzystania technologii Business Intelligence część praktyczna została podzielona na 3 etapy budowy ekosystemu. Etapy te są niezbędnymi krokami, aby możliwa była budowa prawidłowego ekosystemu. Na potrzeby przetestowania danej architektury została ona zaimplementowana dla przedsiębiorstwa Advenutre Works. Budowa przedstawionej koncepcji w części teoretycznej została podzielona odpowiednio na:

- Stworzenie założeń ekosystemu oraz uzasadnień biznesowych do jego realizacji. Przedstawienie stanu zastanego w przedsiębiorstwie, identyfikacja źródeł danych oraz wymagań do ich przetwarzania.
- Przygotowanie architektury wraz z wykorzystanymi technologiami do jej wykonania tak, aby spełniała określone cele. Oparcie jej o warstwę przetwarzania danych w Apache Sparku oraz warstwę wizualizacji danych w środowisku Power Bi.
- Prezentację implementacji stworzonej architektury do firmy Advenutre Works. Przygotowanie przepływu danych w środowisku Apache Spark oraz budowę interaktywnej aplikacji opartej o cztery pulpity nawigacyjne dla analityków biznesowych na platformie Power Bi. Wizualizację za pomocą wykresów i raportów danych historycznych, których analiza pomogła w optymalizacji przedsiębiorstwa Adventure Works.

2. Wstęp

W dzisiejszym świecie maksymalizacja przychodów oraz optymalizacja procesów produkcyjnych jest niezbędna do prawidłowego rozwoju przedsiębiorstwa. Rządy, korporacje, media próbują znaleźć jak najlepszy wzorzec projektowy, którego zadaniem ma być trafienie do jak największej grupy odbiorców oraz zatrzymanie ich przy sobie. Analityka danych zdobytych na przestrzeni lat z przedsiębiorstwa lub innych źródeł danych ma za zadanie przekształcić dane w wiedzę, która pomóc ma w uzyskaniu oczekiwanych celów.

Wraz ze wzrostem ilości danych oraz ich różnorodności na przestrzeni lat zwiększyła się trudność w ich przetwarzaniu oraz analizie. Technologia Big Data ma za zadanie pomóc w przetwarzaniu danych, z którymi nie były w stanie sobie poradzić tradycyjne metody. Systemy i technologię Big Data poprzez swoją skalowalność sprawiają, że analiza petabajtów danych jest możliwa poprzez wykorzystanie wielu węzłów roboczych. Wykorzystanie omawianych systemów przetwarzania może być problematyczne przez złożoność ich struktury – do rozpoczęcia pracy z technologią wymagana jest duża ilość wiedzy. Dane uzyskane z różnych źródeł danych są gotowe do analizy po przeprowadzeniu procesów ETL/ELT. Posłużyć temu ma Business Intelligence jest to współczesna odpowiedź sektorów informatycznych na dzisiejsze wymagania rynkowe. Narzędzia należące do technologii Business Intelligence służą do raportowania i wizualizacji danych. Procesy te ułatwiają podejmowanie decyzji oraz analizę danych dla analityków biznesowych. Architektura ekosystemu Business Intelligence jest równie skomplikowana co architektura Big Data, jednak jej stworzenie daje wiele korzyści i można ją implementować do wielu różnych przedsiębiorstw. Ekosystemy te są wykorzystywane do zarządzania relacjami z klientami w sektorach gospodarczych takich jak przemysł, handel, usługi, budownictwo, edukacja oraz opieka zdrowotna.

Praca dyplomowa przedstawia koncepcje architektury ekosystemu Business Intelligence w oparciu o środowisko Big Data. W części praktycznej zostało przedstawione na przykładzie rzeczywistego przypadku wykorzystanie ekosystemu w celu zaprezentowania jego możliwości wkładu w analizę rozwoju przedsiębiorstwa.

3. Wprowadzenie teoretyczne

3.1. Przetwarzanie danych

3.1.1. Systemy OLTP/OLAP

Systemy baz danych rozwijają się w błyskawicznym tempie. Zarówno ilość połączeń, jak i ilość danych wzrastały wykładniczo od momentu przyjęcia systemów otwartych i oprogramowania serwerowego baz danych takich jak Oracle. Systemy do przetwarzania danych online można podzielić na OLTP (*ang. Online Transaction Processing*) oraz OLAP (*ang. Online Analytical Processing*) [1]. Bazy danych OLTP zazwyczaj posiadają wysoce znormalizowane tabele, które doświadczają wzrostu poprzez liczne krótkie transakcje. Bazy danych kont bankowych, punktów sprzedaży czy systemu składania zamówień są idealną reprezentacją baz danych OLTP. Systemy przetwarzają wiele danych jednocześnie, są one szyfrowane i zapisywane w celach zbierania informacji oraz tworzenia raportów [1][2]. Niektóre bazy danych OLTP mogą obsługiwać więcej niż jedną aplikację. Dobrym przykładem jest księgowość ze względu na możliwość rozdzielenia modułów [1]. Bazy danych OLAP są wykorzystywane do tworzenia raportów danych i podsumowań za pomocą narzędzi OLAP oraz aplikacji zarządzających jak systemy planowania. Podejście to jest wielowymiarowe do organizacji i analizy danych biznesowych. Przechowywanie danych odbywa się w wysoce zoptymalizowanych strukturach. Korporacje mogą bardzo szybko eksplorować dane i odkrywać ważne spostrzeżenia, które bez wykorzystania tego systemu pozostałyby ukryte [1][3]. Skuteczne rozwiązania OLAP rozwiązują problemy działu technologii informacyjnych (*ang. Information Technology*) oraz klientów biznesowych. OLAP oferuje: szybkościowe analizy, wiarygodne dane i obliczenia, elastyczność, dane wielowymiarowe oraz, co najważniejsze dla użytkowników biznesowych, samoobsługowe raportowanie [3]. Różnicę pomiędzy systemami OLTP i OLAP przedstawia rysunek 3.1.

System OLTP	System OLAP
Umożliwia wykonywanie w czasie rzeczywistym dużej liczby transakcji bazodanowych przez dużą liczbę osób	Zazwyczaj wykonuje zapytania dotyczące wielu rekordów (nawet wszystkich rekordów) w bazie danych do celów analitycznych
Wymaga błyskawicznych czasów reakcji	Nie wymaga błyskawicznych czasów reakcji (reaguje o rząd wielkości wolniej niż system OLTP)
Używa indeksowanych danych w celu skrócenia czasu odpowiedzi	Przechowywanie danych w formacie kolumnowym, aby zapewnić łatwy dostęp do dużej liczby rekordów
Wymaga częstego lub równoczesnego wykonywania kopii zapasowych bazy danych	Wymaga znacznie rzadszego tworzenia kopii zapasowych bazy danych niż system OLTP
Zazwyczaj obsługuje proste zapytania dotyczące tylko jednego lub kilku rekordów	Obsługuje złożone zapytania obejmujące dużą liczbę rekordów
Wymaga stosunkowo niewiele miejsca do przechowywania danych	Zazwyczaj ma duże wymagania dotyczące miejsca do przechowywania danych, ponieważ ma duże ilości danych historycznych

Rys. 3.1. Porównanie systemów OLTP i OLAP

3.1.2. Proces ETL

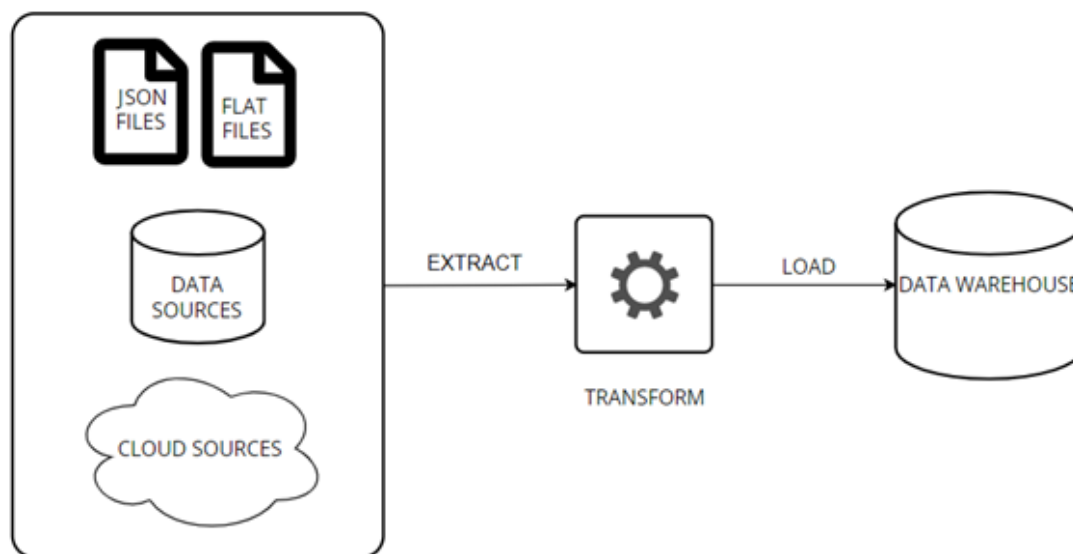
System ETL (ang. *Extract-Transform-Load*) jest podstawą hurtowni danych. Pobiera on dane z systemów źródłowych. Egzekwuje standardy jakości i spójności danych i dostarcza dane w formacie gotowym do prezentacji [4]. ETL ma za zadanie połączyć dane w jedną spójną całość z wielu różnych źródeł danych.

Wyodrębnianie (ang. *Extract*) – ETL wyodrębnia dane ze źródeł i je kopiuje, następnie przetransportowuje dane do magazynu danych. Dane mogą pochodzić z ustrukturyzowanych i nieustrukturyzowanych źródeł danych np. API, dokumentów, baz danych, detektorów.

Transformacje (ang. *Transform*) – dane pochodzą z różnych źródeł, dlatego powinny zostać ujednolicone. Transformacja ma za zadanie usunąć duplikaty oraz walidować dane w celu ich oczyszczenia, aby mogły trafić do miejsca docelowego, czyli hurtowni danych.

Załadowanie (ang. *Load*) – ETL załadowuje dane do postaci docelowej. Dane te mogą zostać zaimportowane w czasie rzeczywistym [5].

Różnica pomiędzy ETL a ELT występuje w momencie transformacji oraz miejscu jej wykonania. Proces ELT to w pierwszej kolejności załadowywanie danych do obszaru tymczasowego, a następnie dokonywanie ich transformacji danych [4]. Schemat budowy architektury przedstawia rysunek 3.2.



Rys. 3.2. Architektura ETL

3.1.3. Hurtowanie danych

Koncepcja hurtowni danych powstała w latach 1980. Oznacza ona: „zorientowany tematycznie, zintegrowany, zmienny w czasie i nieulotny zbiór danych wykorzystywany w strategicznym procesie decyzyjnym”. Hurtownia danych pełni funkcje epicentrum integracji danych. Hurtownia ma przekształcać dane w informacje biznesowe. Jest to rozwiązanie skupiające się na przedsiębiorstwach. Zapewnia ona jednolity widok danych firmy, niezależnie

od tego, jak będą one później wykorzystywane przez użytkowników. Tworzy stabilne źródło historycznych informacji, które jest stałe, spójne i niezawodne dla każdego konsumenta [6].

Pojawienie się hurtowni danych doprowadziło do tego, że przedsiębiorstwa mogą:

- spojrzeć na dane z perspektywy korporacyjnej,
- popatrzeć na dane z perspektywy zintegrowanej,
- zwrócić uwagę na dane z perspektywy czasu,
- mieć jeden fundament, który służył wielu grupom w organizacji.

Korzyścią płynącą z wykorzystywania hurtowni danych jest dostarczanie dużej ilości danych z różnych źródeł, dając korporacjom możliwość ich analizy oraz archiwizacji. Cechy, jakie powinna posiadać hurtownia danych, przedstawia rysunek 3.3 [6][7].

Przedmiotowość	Analiza danych dotyczące konkretnego tematu lub obszaru funkcjonalnego
Integracja	Spójność różnych typów danych pochodzących z różnych źródeł.
Trwałość	Dane znajdujące się w hurtowni danych są stabilne i nie zmieniają się
Zmienność w czasie	Analiza danych w hurtowni uwzględnia zmiany zachodzące w czasie

Rys. 3.3. Cechy charakterystyczne hurtowni danych

Prawidłowo zaprojektowana hurtownia powinna być w stanie dostosować się do wymagań użytkownika, a jakość i szybkość analizy – być dostosowane do poziomu złożoności zapytania konsumenta. Hurtownia danych w ostatnim kroku ma za zadanie dostarczyć do środowiska Business Intelligence dane natomiast użytkownik otrzyma raporty pasywne, proaktywne, analizy silosowe i inne interfejsy [6].

Różnica pomiędzy hurtownią danych a widokiem zapytania do bazy danych jest następująca:

- hurtownia danych istnieją jako trwałe magazyny zamiast być tworzone na żądanie,
- hurtownie zazwyczaj nie są relacyjne, a wielowymiarowe w przeciwieństwie do baz danych,
- widoki nie mogą być indeksowane natomiast hurtownie danych tak celem optymalizacji,
- hurtownie wyciągają zazwyczaj więcej danych niż zawiera jedna baza danych [8].

3.2. Big Data

3.2.1. Zastosowanie systemów Big Data

Big Data to termin, który oznacza ogromne i jednocześnie złożone zbiory danych, które są uzyskiwane z wielu różnych źródeł. Wielkość tych zbiorów danych jest tak duża, że użytkownik nie jest w stanie nimi zarządzać za pomocą tradycyjnych oprogramowań. Analizowane są one w celu odkrycia wzorców dotyczących użytkowników i maszyn, dzięki czemu korporacje są w stanie rozwiązywać problemy biznesowe [9].

Big Data to duże ilości danych, które mogą być różnego rodzaju. Struktura ta charakteryzuje się kilkoma cechami. Najważniejsze z nich zostały wymienione poniżej wraz z ich szczegółowym opisem:

- **Objętość** (ang. *Volume*) jest jedną z głównych cech Big Data i dotyczy rozmiaru danych. Aspekt ten zmienia się wraz ze zwiększeniem się ilości gromadzonych danych. W zależności od rozmiaru przedsiębiorstwa dane są gromadzone w terabajtach lub petabajtach danych.
- **Różnorodność** (ang. *Variety*) opisuje szeroką gamę typów danych, które są przechowywane, przetrzymywane i analizowane. Dane te (np. pliki audio i wideo, zdjęcia, głos, dane z GPS, pliki medyczne), często nieustrukturyzowane lub tylko częściowo ustrukturyzowane, wymagają wstępnego przetwarzania w celu zrozumienia ich znaczenia.
- **Prędkość** (ang. *Velocity*) odnosi się do tempa zapisywania i przetwarzania danych. Duże dane ulegają szybkim zmianom. Prędkość ma za zadanie przetwarzać strumienie danych w celu stworzenia hipotezy lub zauważenia trendu w czasie rzeczywistym lub zbliżonym do rzeczywistego, aby uzyskać wynik, który jest szukany w danym momencie.
- **Wartość** (ang. *Value*) to cecha, która zawiera informacje, jakie dane można uzyskać z przechowywania danych. Konkretnie informacje wzbogacają analizę o większą pewność dokonywanych przewidywań.
- **Prawdziwość** (ang. *Veracity*) pokazuje jakość danych i źródło danych. Jest stopniem dokładności lub prawdziwości zbioru danych. Odnosząc się do Big Data, nie oznacza tylko jakości, a również wiarygodność źródła danych. Usuwa stroniczość, niespójność lub duplikacje w danych. Cecha ta wpływa na poprawę dokładności Big Data.
- **Zmienność** (ang. *Variablity*) odnosi się do tego, jak dane mogą ewoluować w czasie czyli zmieniać swoje znaczenie, strukturę oraz informację, którą niosą [10].

Branże wykorzystujące technologii Big Data:

- Produkcja
- Handel detaliczny
- Opieka zdrowotna
- Telekomunikacja
- Usługi finansowe

Rewolucja cyfrowa zmieniła branżę produkcyjną. Producenci w przemyśle wytwórczym wykorzystują wszystkie wygenerowane przez nich dane, aby zmaksymalizować zyski oraz rozwój przedsiębiorstwa. Duże zbiory danych zawierających informacje o sprzęcie takie jak: rok produkcji, model, producent. Zostaną one wykorzystane do przewidzenia prawdopodobieństwa wystąpienia awarii sprzętu. Informacje jakie można uzyskać dzięki analizie danych, posłużyły również do wyznaczenia czasu sprawności sprzętu lub momentu, w którym należy dokonać konserwacji. Efektywność operacji oraz optymalizacja produkcji to zagadnienia, w których wykorzystanie technologii Big Data jest w stanie poprawić kondycję funkcjonowania firmy. Analiza procesów produkcji ukazała procent rentowności projektów oraz miejsca możliwe do optymalizacji, dzięki czemu korporacje mogły reagować proaktywnie na wymagania klientów. Optymalizacja linii produkcji wpłynęła na koszty oraz efektywność obszarów produkcji.

Handel detaliczny cechuje się konkurencyjnością, więc firma, aby nadal utrzymywać się w czołówce, powinna cały czas się rozwijać. Wykorzystując Big Data, sprzedawcy detaliczni tworzą innowacyjne metody, mające na celu zwiększenie popytu na ich produkty. Firmy mogą maksymalizować jakość swoich interakcji z klientami w sieci za pomocą proaktywnego rozwiązywania problemów lub personalizacji ofert pod konkretnych klientów. Kompleksowa analiza rentowności oraz marż pomaga w zidentyfikowaniu obszarów wymagających interwencji przedsiębiorstwa oraz w poprawie cen tam, gdzie zyski są zagrożone.

Opieka zdrowotna to sektor gospodarki, w którym wykorzystanie Big Data pomoże w ratowaniu życia ludzkiego. Firmy medyczne mogą wykrywać trendy, wzorce oraz zagrożenia u swoich pacjentów. Dzięki analizie źródeł danych będą mogli tworzyć modele predykcyjne. Naukowcy mogą wykorzystać Big Data do badań genomicznych w celu zidentyfikowania genów chorobowych, aby pomóc pacjentom w ich problemach zdrowotnych. Oszustwa dotyczące roszczeń to sektor, w którym również dzięki setkom powiązanych raportów w różnych formatach, firmy medyczne będą mogły wskazać oszustwa mające na celu wyłudzenia odszkodowań.

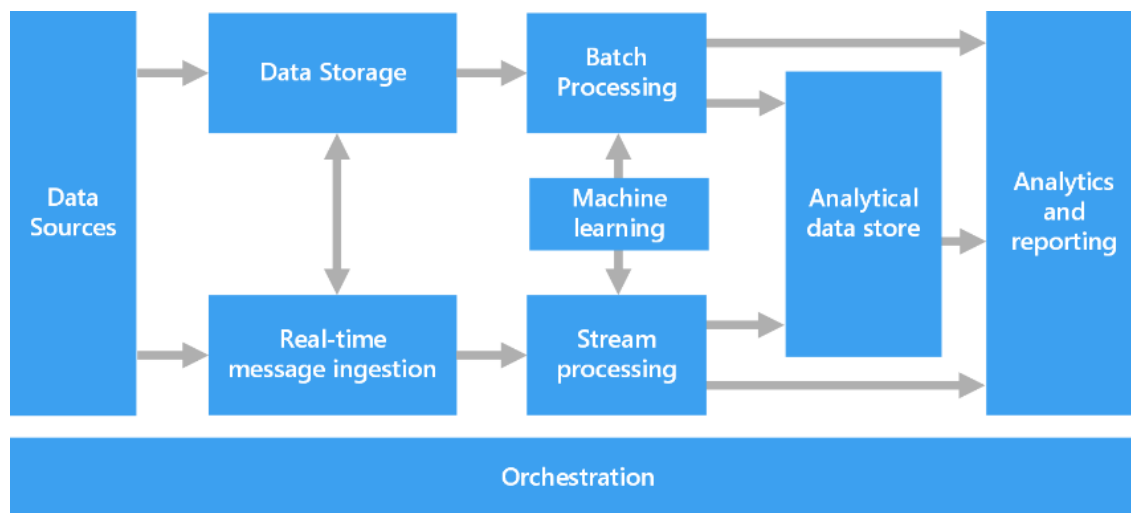
Technologia Big Data nadaje się idealnie do wykorzystania w telekomunikacji. Analiza przepustowości sieci wykaże obszary o nadmiernej eksploatacji sieci w celu ich optymalizacji. Big Data posłuży do planowania inwestycji w infrastrukturę i tworzeniu nowych usług spełniających wymagania klienta. Wskaźnik rezygnacji klientów z usług telekomunikacyjnych

jest obszarem, w którym analiza danych wykaże jakość usług, poziom zadowolenia z nich oraz możliwość przedłużenia oferty. Produktem będzie raport, który będzie można wykorzystać, aby zatrzymać u siebie klienta. Nowa oferta produktów powstanie dzięki wykorzystaniu danych historycznych oraz danych z czujników do tworzenia modeli predykcyjnych na temat zapotrzebowania konsumenckiego.

Banki oraz firmy świadczące usługi z sektora finansowego wykorzystują Big Data do przeciwdziałania oszustwom. Bezpieczeństwo jest niezwykle ważne dla banków ze względu na przepisy wymagające staranności i tworzenia raportów na temat podejrzanych działań. Big Data jest w stanie wykryć wzorce potencjalnych prób oszustwa [9].

3.2.2. Opis i architektura systemów Big Data

Architektura Big Data w przeciwieństwie do prostych systemów bazodanowych ma za zadanie zbierać, przechowywać i analizować złożone zbiory danych o wielkości setek tera-, a nawet petabajtów. Dane mogą być dostarczane w czasie rzeczywistym szybko i stale lub wolniej jako dane historyczne [11]. Rysunek 3.4 przedstawia przykładową architekturę systemów Big Data.

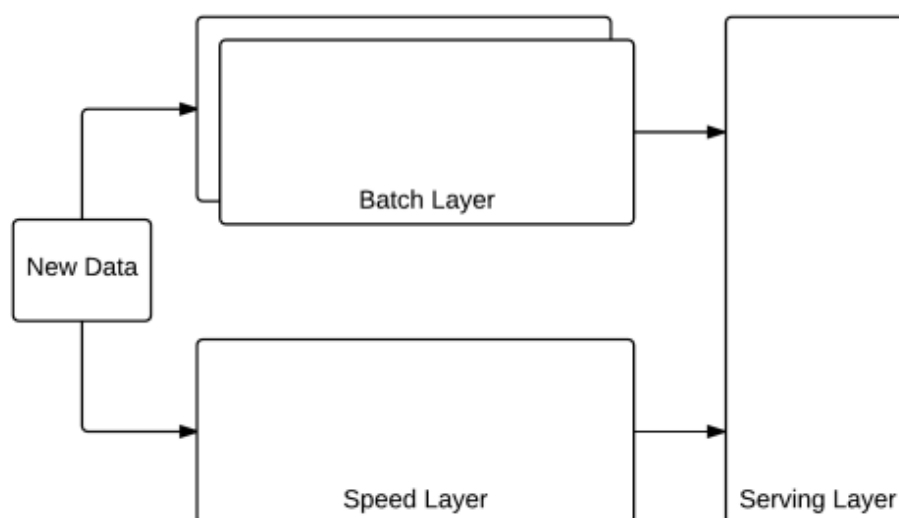


Rys. 3.4. Styl architektury Big Data
<https://learn.microsoft.com/pl-pl/azure/architecture/data-guide/big-data/>

Architektura Big Data musi posiadać co najmniej jedno źródło danych w swojej strukturze. Może nim być np. aplikacja mobilna, bazy OLTP, API (ang. *application programming interface*), relacyjne bazy danych. Magazyn danych jest miejscem, w którym dane są przechowywane w celu późniejszego wykorzystania (np. SQL Server). Przetwarzanie wsadowe jest to metoda wykorzystywana do przetwarzania dużej ilości powtarzających się danych. Podczas przetwarzania wsadowego odbywa się agregacja, filtracja oraz sortowanie danych do analizy w zależności od potrzeb użytkownika. Wprowadzanie komunikatów w czasie

rzeczywistym jest to etap w architekturze, który musi uwzględnić zbieranie danych z komunikatorów zewnętrznych takich jak np. kontrolery, czujniki itp. Dane mogą być przechowywane w zwykłych magazynach danych. Przetwarzanie strumieniowe to proces przechwytywania strumieni wydarzeń w czasie rzeczywistym. Zgromadzone dane są przetwarzane w celu późniejszej analizy. Analityczna hurtownia danych to miejsce, w którym dane są przechowywane w ustrukturyzowanej formie. Narzędzia analityczne mogą zostać wykorzystane w celu analizy biznesowej uzyskanych danych. Analiza i raportowanie danych to finalny produkt rozwiązania Big Data. Służy ona do tworzenia raportów, pulpitów nawigacyjnych, widoków raportowych, hurtowni tematycznych przez użytkownika. Dane można wizualizować w celu lepszego zobrazowania informacji. Harmonogramowanie i monitoring operacyjny opisuje przetwarzanie danych, do jakich miejsc mają być one przetwarzane oraz w jakich odstępach czasowych [11].

Architektura lambda – jest ona niezależna od technologii i definiuje pewne praktyczne i dobrze opanowane zasady obsługi dużych zbiorów danych. Schemat budowy architektury pozwala na obsługę danych historycznych, jak i obsługę danych w czasie rzeczywistym. Proces przepływu danych w architekturze ilustruje rysunek 3.5. Wadą takiego rozwiązania jest złożoność zarządzania oraz duplikacja obliczeń dla obu ścieżek przetwarzania danych [12].

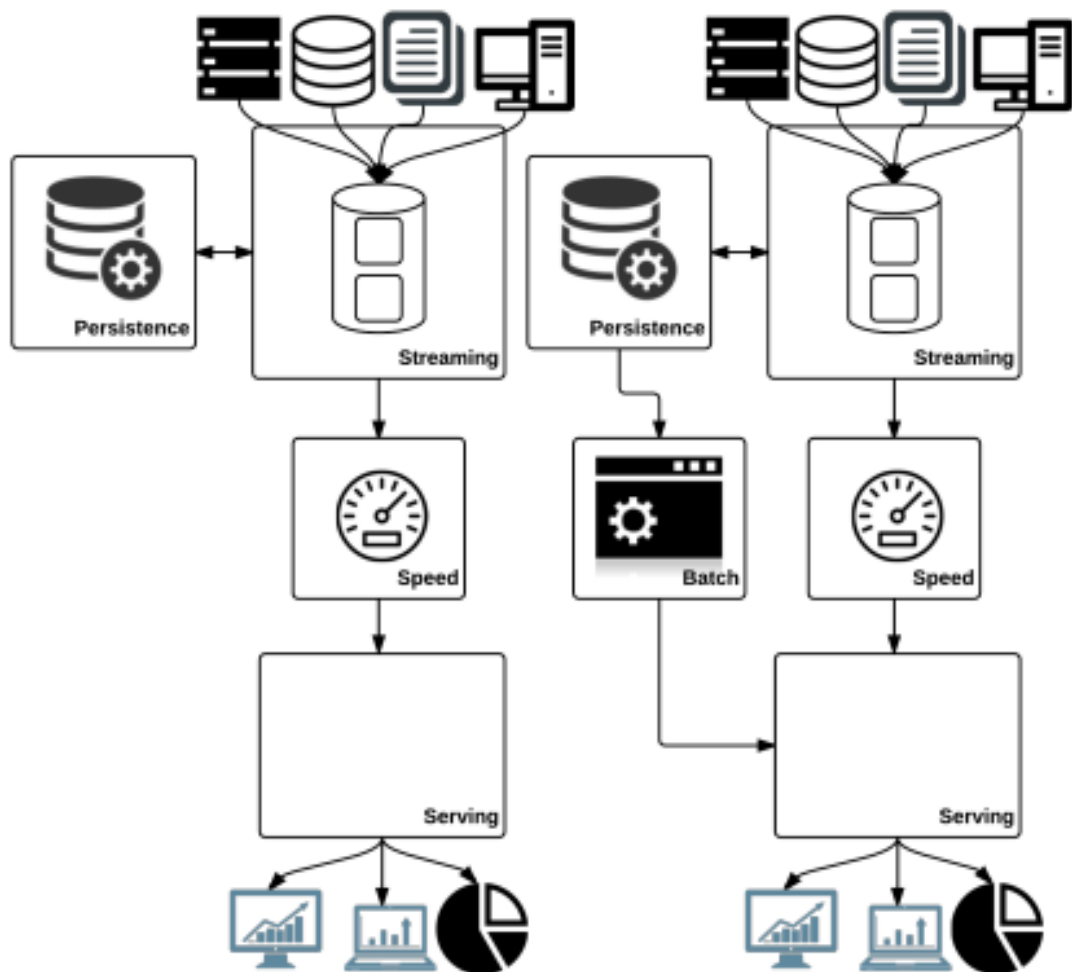


Rys. 3.5. Komponenty architektury lambda
Data Lake for Enterprises Autor: Tomcy John, Pankaj Misra

Nowe dane są dostarczane zarówno do warstwy wsadowej (ang. *Batch layer*) oraz warstwy czasu rzeczywistego (ang. *Speed layer*). Warstwa wsadowa to miejsce, w którym dane są przechowywane w formie nieprzetworzonej. Operacje na danych takie jak aktualizacja oraz usuwanie są niemożliwe. Warstwa ta jest ścieżką zimną (ang. *cold path*), czyli zawiera dane historyczne z opóźnieniami, co pozwala na uzyskanie dokładniejszych wyników. Ścieżka gorąca

(ang. *hot path*) to ścieżka, w której dane trafiają do warstwy czasu rzeczywistego. Spełnia ona wymagania analizy w czasie rzeczywistym. Warstwa wsadowa jest przeznaczona dla dużej ilości danych. Dane potrzebne do analizy w czasie rzeczywistym są inkrementowane do warstwy szybkości z niskim opóźnieniem implementacji. Zapytanie utworzone przez użytkownika trafia zarówno do warstwy wsadowej, jak i do warstwy czasu rzeczywistego, która ostatecznie trafia do warstwy serwującej (ang. *Serving layer*). Warstwa ta, zajmując się tworzeniem widoków zapytań użytkowników oraz systemu, jednocześnie odpowiada za orkiestrację. [12]

Architektura kappa jest alternatywą dla architektury lambda. Różnica pomiędzy tymi dwiema architekturami wynika z usunięcia warstwy wsadowej. Architektura ta chce uniknąć konieczności dwukrotnego wykonywania tej samej logiki. Dane w tej architekturze są również niemodyfikowalne. Cel tej architektury jest identyczny, jednak droga do jego osiągnięcia odbywa się przy użyciu przetwarzania strumieniowego. Różnice w strukturze obu architektur przedstawia rysunek 3.6. [11][12]



Rys. 3.6. Porównanie architektury kappa (z lewej strony) oraz architektury lambda (z prawej strony)
Data Lake for Enterprises Autorzy: Tomcy John, Pankaj Misra

3.3. Apache Spark

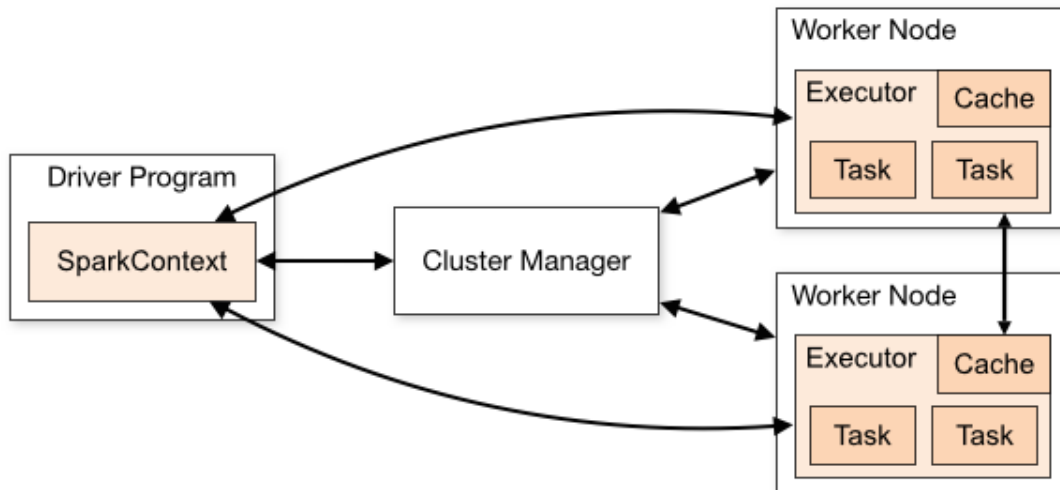
3.3.1. Architektura Apache Spark

Architektura Spark to oprogramowanie obliczeniowe z otwartym źródłem (ang. *open source*) przeznaczone dla Big Data, którego zadaniem jest przetwarzanie nieustrukturyzowanych i ustrukturalizowanych danych do zastosowań analitycznych. Architektura ta składa się z czterech komponentów Spark Driver, Spark Executors, Cluster Manager, węzły robocze (ang. *worker nodes*). Apache Spark jako struktura (ang. *framework*) do obliczeń klastrowych powstał w celu przyspieszenia aplikacji do przetwarzania danych [13].

Cechy Apache Spark:

- Szybkość – podczas przetwarzania dużych wolumenów danych działa nawet 100 razy szybciej niż MapReduce. Równocześnie jest w stanie podzielić dane na części według wymagań użytkownika.
- Wydajne buforowanie – dzięki wykorzystaniu prostej warstwy programistycznej Spark posiada wydajne możliwości buforowania.
- Implementacja – może zostać użyty do wdrożeń.
- Real-Time – zapewnia obliczenia w czasie rzeczywistym.
- Poliglotyczność – możliwość pisania kodu Sparka w wielu językach programowania takich jak: Python, R, Scala, Java.
- Skalowalność – bezproblemowa inkrementacja wielkości systemu.

W skład architektury Apache Spark wchodzi dwie główne warstwy abstrakcyjne. Odporne na awarie rozproszone kolekcje danych (ang. *Resilient Distributed Datasets*, w skrócie RDD) to narzędzie pomocne w ponownym obliczaniu danych po awarii. RDD są elementami każdej aplikacji Apache Spark. Odporny (ang. *Resilient*) odpowiada za zmniejszyć podatność na błędy oraz rekonstrukcje danych utraconych podczas awarii. Rozproszona (ang. *Distributed*) to rozproszenie danych pomiędzy wieloma węzłami w klastrze. Zestawy danych (ang. *Datasets*) to zbiory partycjonowanych danych posiadających wartości. Drugą warstwą jest skierowany graf acykliczny (ang. *Directed Acyclic Graph* w skrócie DAG) jest to sterownik konwertujący program dla każdego zadania. Skierowany graf acykliczny to zbiór wierzchołków i krawędzi, gdzie wierzchołki reprezentują RDD, a krawędzie reprezentują operacje, które mają być zastosowane na RDD. W Spark DAG każda krawędź kieruje od wcześniejszej do późniejszej części sekwencji. Spark DAG jest ścisłym uogólnieniem modelu MapReduce. Ekosystem Apache Spark składa się z wielu komponentów, DAG odpowiada w nim za sekwencje połączeń między węzłami. Architekturę Apache Spark ilustruje rysunek 3.7 [13][16].



Rys. 3.7. Architektura Apache Spark
<https://www.interviewbit.com/blog/apache-Spark-architecture/>

W architekturze Apache Spark, Driver Program to komponent odpowiedzialny za wywołanie programu aplikacji oraz utworzenie SparkContext. SparkContext zawiera w sobie funkcje, natomiast Spark Driver – komponenty DAG. Komponenty te tłumaczą kod stworzony przez użytkownika na zadania (ang. *tasks*), które są wykonywane na klastrze. Cluster Manager jest odpowiedzialny za kontrolowanie oraz zarządzanie wykonywaniem różnych zadań w klastrze. Spark Driver jest w relacji z Cluster Manager. Cluster Manager przydziela dane oraz zasoby do danego zadania. Jedno zadanie jest dzielone na mniejsze fragmenty, a następnie rozdzielane między węzłami roboczymi. Węzeł roboczy będzie wykonywał zadanie mu przydzielone pod kontrolą Spark Driver. Zadania te będą przetwarzane z użyciem RDD stworzonego w Spark Context, natomiast wyniki będą buforowane (ang. *cached*). Zadania można przypisywać większej ilości węzłów w celu przyspieszenia wydajności systemu. Dzięki takiemu rozwiązaniu zadania będą podzielone na bardziej spójne części. Cluster Manager przekazuje informacje o zadaniach do Spark Context, który zapisuje je na węzłach roboczych. W skład węzłów roboczych wchodzi proces wykonawczy (ang. *Executor*). Proces wykonawczy to miejsce w architekturze, które odpowiada za wykonanie tych zadań. [13][14][15]

W poniższych akapitach omówiono główne elementy składowe programu utworzonego z wykorzystaniem platformy Apache Spark:

Spark Driver – sterownik Spark posiada węzeł główny, który odpowiada za nadzór zadań oraz koordynację między węzłami roboczymi. Spark Contexts tworzone są przez sterownik Spark do monitorowania zadań w konkretnym klastrze oraz relacji pomiędzy pozostawionymi klastrami Spark. Cykl życia sesji Spark w Spark Context posiada sterownik składający się z komponentów niezbędnych do wykonywania zadań w innych klastrach.

Spark Executors – sterownik monitoruje procesy wykonawcze podczas ich pracy. Procesy wykonawcze uruchamiają zdania i przetwarzają dane, a po tym procesie zostają usunięte w trybie beczynności. Uruchamiają one aplikacje współbieżnie, działają w procesie Java podczas pracy z danymi. Przydzielane są one dynamicznie, dodawane i usuwane zaraz po wykonaniu swojego zadania.

Cluster Manager – menedżer klastra przechowuje dane w pamięci buforowanej (ang. *cache*) oraz steruje wykonywaniem zadań w klastrach. Wydaje on polecenia węzłom klastra, na których działa.

Węzły robocze (ang. *Workers nodes*) – węzły podrzędne pełnią rolę procesów wykonawczych, przetwarzają one dane podczas zadań (ang. *tasks*), a następnie wysyłają wynik do Spark Context. Węzeł nadrzędny wydaje polecenia do Spark Context, a węzły podrzędne je wykonują. Spark Worker monitoruje wykonywanie zadań przez węzły robocze, do których może być przypisane jedno zadanie Spark. W Spark partycja jest jednostką pracy, do której przypisany jest jeden proces wykonawczy [13][15].

3.3.2. Ekosystem Apache Spark

Ekosystem Sparka składa się z następujących komponentów:

- Spark Core – bazowy silnik do rozproszonego, równoległego i wielkoskalowego przetwarzania danych. Zarządza pamięcią, planowaniem, monitorowaniem, dystrybucją zadań na klastrach oraz interakcjami z systemami pamięci masowej.
- Spark Streaming – komponent Sparka, służący do przetwarzania danych strumieniowych na żywo. API Spark pozwala na skalowalne, wysokiej przepustowości i odporne na błędy przetwarzanie strumieni danych. Zbiorami danych mogą być różne źródła takie jak np. Apache Kafka.
- Spark SQL – API Spark pozwalające na przetwarzanie relacyjne z funkcjonalnym. Spark SQL pozwala na wykonywanie zapytań dla zaimportowanych danych oraz istniejących RDD.
- GraphX – komponent Spark przeznaczony dla obliczeń równoległych oraz grafów. GraphX rozszerza Spark RDD dzięki dodaniu nowej abstrakcji grafu.
- MLlib (ang. Machine learning library) – służy on do uczenia maszynowego w Apache Spark. MLlib zawiera wysokiej jakości algorytmy. Algorytmy te wykorzystują iteracje (w przeciwieństwie do MapReduce, który używa aproksymacji), dzięki czemu uzyskują lepsze wyniki.
- SparkR – pakiet R zawierający rozproszoną implementację ramki danych. Stworzony został do obsługi operacji takich jak filtrowanie, sortowanie, agregacja dla dużych zbiorów danych [14][16].

3.4. Business Intelligence

3.4.1. Charakterystyka i zastosowanie Business Intelligence

W dzisiejszych czasach każda organizacja biznesowa powinna nieustannie monitorować otaczające ją środowisko rynkowe oraz swoje własne wyniki, a następnie umiejętnie zaadaptować się do obecnej sytuacji. Zadaniem Business Intelligence jest monitorowanie branży, konkurencji, klientów oraz dostawców. Dzięki wykorzystaniu kluczowych wskaźników efektywności (ang. *key performance indexes*) w skrócie *KPIs* dyrektorowie firm mogą określać, co chcą obserwować w raportach dotyczących firmy. Kluczowe wskaźniki efektywności określają skuteczność, jaką jest się w stanie osiągnąć dzięki wykonaniu konkretnych działań. Narzędzie to może zostać wykorzystane przez organizacje na wielu obszarach. Business Intelligence to szeroki zestaw rozwiązań informatycznych. Obejmuje on narzędzia do analizy i raportowania użytkownikowi informacji z pozyskanych danych. Analiza danych przedsiębiorstw (np. produktów jakie sprzedają) może pomóc w identyfikacji produktów, które sprzedają się szybciej od innych. Analiza może również okazać się przydatna podczas obserwacji zachowań tendencyjnych wśród klientów. Informacje oraz spostrzeżenia zdobyte po analizie Business Intelligence mogą zostać wykorzystane do opracowania lepszej strategii marketingowej przedsiębiorstwa oraz do utworzenia lepszej jakości produktu lub usług jakie świadczy firma.

Hurtowanie danych mogą, ale nie muszą, być częścią struktury architektury Business Intelligence. Business Intelligence dotyczy wszystkich sektorów przemysłowych oraz funkcji wszystkich pracowników korporacji, klientów i dostawców. Wiedza uzyskana z informacji, które zawierają dane podczas analizy biznesowej zapewnia wzrost wartości przedsiębiorstwa, pokazując sektory, w których firma jest zdolna poprawić proces działania. Bez wywiadu biznesowego menadżerowie korporacji nie są w stanie reagować na bieżącą sytuację dziejącą się w firmie. Analityka biznesowa pozwala na zwiększenie elastyczności oraz terminowości w monitorowaniu zarządzanym przedsiębiorstwem. Korporacje mogą wykorzystywać Business Intelligence do monitorowania wydatków, zamiast czekać do końca kwartału, a także na bieżąco sprawdzać, czy budżet nie zostanie przekroczony. Wykorzystywana efektywnie, Business Intelligence pozwala organizacji na poprawę wydajności. Wyniki biznesowe są mierzone za pomocą wielu wskaźników finansowych (przychody, koszty, marża, rentowność inwestycji itd.). Pozwala ukazać, dlaczego niektóre korporacje działają lepiej od innych. Kluczową oznaką sukcesu Business Intelligence jest poziom, w jaki wpływa na wyniki działalności. [18]

Eksploracja danych (ang. *Data mining*) to proces, którego zadaniem jest wydobywanie informacji zawartych w zbiorach danych. Wzorcem są modele lub projekty, pozwalające ukazać zjawisko korelacji pomiędzy danymi, które początkowo wydają się ze sobą niepowiązane. Mogą działać zgodnie z zasadą Pareto – 20% produktów generuje 80% przychodów.

Wyróżniamy wzorce: czasowe, przestrzenne, funkcjonalne, symetryczne. Świadomość tych wzorców oraz posiadanych przez nie cech ułatwia stworzenie bardziej zrównoważonych produktów, które trafią do jak największej grupy odbiorców. Jednak utrwalone od dawna wzorce mogą okazać się błędne. Zawsze istnieje ryzyko wystąpienia zjawiska czarnego łabędzia, czyli sytuacji, która nie powinna się nigdy wydarzyć lub sytuacji, która powinna się wydarzyć, a się nie wydarzyła. Osoba zajmująca się tworzeniem wzorców projektowych powinna wiedzieć, jakich konkretnie danych potrzebuje, a następnie przetworzyć odpowiednią ilość surowych danych w poszukiwaniu unikalnych użytecznych wzorców.

Wykorzystanie odpowiednich oczyszczonych danych zwiększa szanse na dokonanie właściwych odkryć. Surowe dane (ang. *raw data*) mogą mieć różny poziom jakości, a nawet mogą być sprzeczne. Posiadane dane mogą być lub nie być odpowiednie do znalezienia dobrych wzorców [17][18]. Proces przepływu danych ilustruje rysunek 3.8.



Rys. 3.8. Łańcuch przetwarzania danych
Business Intelligence and data mining, Autor: Maheshwari, Anil

Dane są centralnym punktem u podstaw Business Intelligence, stanowią zasób naturalny. Powyżej zaprezentowana sekwencja to sekwencja kroków, która jest konieczna do czerpania korzyści z danych w sposób systematyczny. Dane są przechowywane w bazie danych, gdzie mogą być modelowane. Następnie odpowiednie dane trafiają do hurtowni danych, gdzie są składowane. Dane te mogą być łączone z innymi źródłami danych. Przedsiębiorstwo, aby uzyskać przewagę nad konkurencją musi analizować spostrzeżenia z uzyskiwanych danych [17].

Dane są wszystkim, co może zostać zarejestrowane. Bazy danych to zbiory danych, relacyjne lub nierelacyjne, do których dostęp jest możliwy na wiele sposobów. Techniki stosowane przy modelowaniu danych określają reguły i ograniczenia w celu zapewniania spójności i integralności danych. Hurtownia danych to zorganizowany magazyn danych zaprojektowany tak, aby pomóc w podejmowaniu decyzji zarządczych. Dane mogą zostać uzyskane z kilku różnych baz danych, aby odpowiedzieć na określony zestaw zapytań. Eksploracja danych to nauka odkrywania użytecznych, innowacyjnych wzorców z danych.

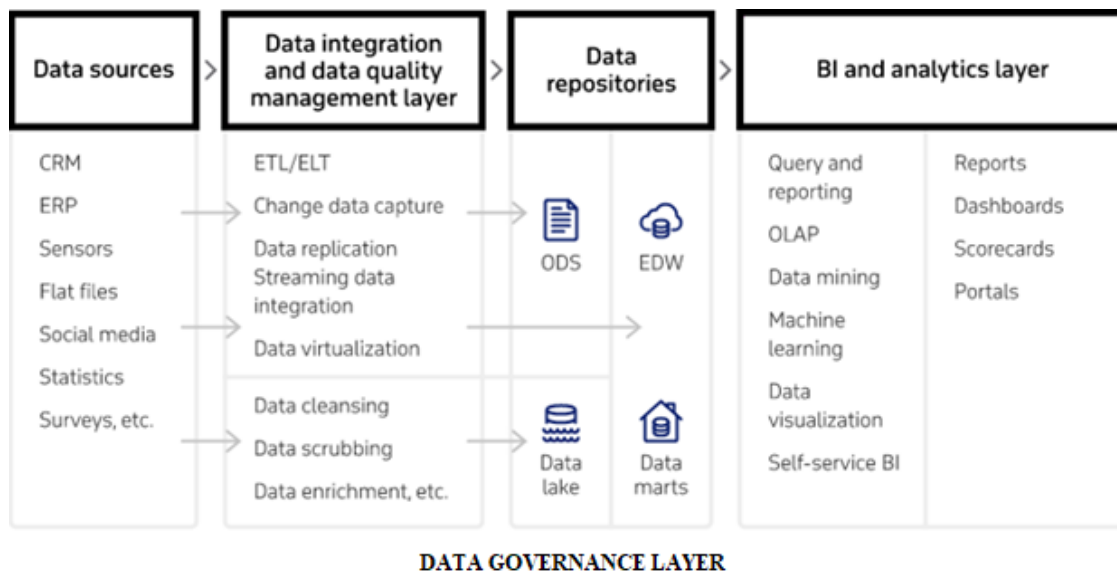
Rozróżnia się różne wzorce danych oraz techniki, które pomagają w ich znalezieniu. Powstała ona do tworzenia priorytetów i zarządzania mniejszą ilością ale kluczowych zmiennych, które odnoszą się bezpośrednio do kluczowych obszarów rezultatów danej roli [17].

Narzędzia Business Intelligence są wymagane w niemal wszystkich branżach. Informacje płynące z danych oraz szybkość ich przetwarzania są niezbędne do prawidłowego funkcjonowania przedsiębiorstwa. Korporacje muszą wdrożyć nowe spostrzeżenia uzyskane z informacji w swoje procesy operacyjne w celu ewoluowania ich działalności w kierunku bardziej efektywnych praktyk. Przykładowe zastosowanie ekosystemów Business Intelligence w poszczególnych branżach:

- Zarządzanie relacjami z klientami – wielkie korporacje zarabiają na spełnianiu potrzeb klientów. Jeżeli klient jest zadowolony z jakości usługi, istnieje duże prawdopodobieństwo, że zostanie stałym klientem. Pozyskiwanie nowych klientów jest zadaniem trudniejszym niż utrzymanie obecnych. Aplikacje Business Intelligence mogą zapewnić przedsiębiorstwu:
 - maksymalizację zwrotu z kampanii marketingowej,
 - utrzymanie stałych klientów i pozyskiwanie nowych,
 - identyfikację najbardziej wartościowej grupy klientów wg zasady Pareto,
 - zarządzanie wizerunkiem marki.
- Opieka zdrowotna – medycyna oparta na danych jest najnowszym trendem w zarządzaniu opieką zdrowotną. Aplikacje Business Intelligence mogą pomóc opiece zdrowotnej w:
 - skuteczności leczenia,
 - diagnozowaniu choroby u pacjentów,
 - kontrolowaniu oszustw i nadużyć.
- Edukacja – szkolnictwo staje się coraz kosztowniejsze i istnieje potrzeba poprawy jakości doświadczeń studentów na różnych poziomach edukacji. Decyzje w oparciu o dane wpłyną na:
 - rekrutację studentów,
 - oferty szkoleń,
 - tworzenie kierunków studiów na podstawie trendów.
- Bankowość – zajmuje się utrzymaniem klientów, którzy posiadają u nich konta bankowe oraz oferowaniu im kredytów. Business Intelligence może wpłynąć w sposób znaczący na poniższe procesy:
 - optymalizacja procesów składania wniosków kredytowych,
 - wykrywanie nieuczciwych transakcji,
 - optymalizacja rezerw gotówkowych oraz ich prognoza.

- Sprzedaż detaliczna – organizacje zajmujące się handlem detalicznym rozwijają się poprzez pozyskiwanie nowych klientów. Zrozumienie potrzeb konsumenckich jest niezbędne do stworzenia nowych wzorców zakupowych klientów. Informacje te pozwolą na:
 - optymalizację zasobów w różnych lokalizacjach,
 - usprawnień układów sklepów,
 - optymalizację logistyczną pod kątem produktów sezonowych.

Ogólną koncepcję architektury Business Intelligence przedstawia rysunek 3.9 [12][17][18].



Rys. 3.9. Architektura Business Intelligence
<https://www.itransition.com/business-intelligence/architecture>

Komponenty budowy ekosystemu Business Intelligence można podzielić na: źródła danych, warstwę integracji danych i zarządzania jej jakością, repozytoria danych, warstwę Business Intelligence i analityki, warstwę zarządzania danymi. Źródła danych możemy podzielić na wewnętrzne oraz zewnętrzne, gdy dane są generowane poza korporacją. Informacje mogą być strukturyzowane, pół strukturyzowane lub nieustrukturyzowane oraz różnią się wielkością. Drugi etap procesu Business Intelligence ma na celu konsolidację zbiorów danych z wielu źródeł w jeden widok najczęściej za pomocą procesu ETL. Następnie wykonywane są takie procesy jak pozbycie się zreplikowanych danych lub danych z brakującymi rekordami. Integracja oraz oczyszczanie danych to dwa procesy, które zachodzą jednocześnie. Dane pobierane z wielu źródeł mogą być niespójne lub zestawy danych mogą zostać powielone. Repozytoria danych (ang. *Data repositories*) to komponent, do którego dane trafiają do analitycznych hurtowni danych lub składka danych (ang. *Data marts*), czyli repozytoriów pamięci masowej dostosowanych do potrzeb analityki i raportowania poszczególnych grup użytkowników. Warstwa Business Intelligence analityki obejmuje dostęp do danych oraz ich wizualizację.

W zależności od potrzeb organizacji z interfejsu użytkownika można sterować analityką opisową oraz diagnostyczną. Ewentualnie użytkownik może wykorzystać kompleksową analizę wspieraną przez uczenie maszynowe lub sztuczną inteligencję. Warstwa zarządzania danymi to element ściśle powiązany z pozostałymi czterema, ponieważ jego głównym celem jest monitorowanie i zarządzanie całym procesem Business Intelligence. Firma może kontrolować kto i jak korzysta z dostępu do informacji oraz czy dane są w odpowiedni sposób zabezpieczone [18][19].

3.4.2. Platforma Tableau

W 1981 roku, kiedy powstał arkusz kalkulacyjny, dziedzina analizy danych rozwinęła się bardzo mocno. Zdolność analityków danych do tworzenia i przechowywania danych wzrosła wykładniczo. Wraz z tym wzrostem powstała dziedzina informatyki Business Intelligence. Tradycyjni liderzy w branży analityki biznesowej stworzyli narzędzia raportowe, które skupiają się na renderowaniu danych z ich własnych produktów. Przeprowadzanie analiz i tworzenie raportów za pomocą tych narzędzi wymaga wiedzy technicznej i czasu. Skala, szybkość i zakres danych wymagają dziś narzędzi do raportowania, które szybko się wdrażają. Muszą one przekształcać dane w informacje oraz być odpowiednie do opanowania przez użytkowników nietechnicznych [21].

Tableau to oprogramowanie do inżynierii danych typu „przeciągnij i upuść” (ang. *drag-and-drop*), które pozwala na podłączenie się do różnych źródeł danych i przygotowanie ich do użycia w Tableau Desktop lub innych programach. Tableau Public to darmowa wersja Tableau Desktop, która pozwala na połączenie z ograniczoną liczbą źródeł danych takich jak Microsoft Excel i tworzyć publiczne analizy. Tableau Public jest darmowe, jednak posiada ono wadę – stworzona praca zapisuje się w sieci publicznej, gdzie każdy ma do niej dostęp, co czyni Tableau narzędziem niewłaściwym do wykorzystania dla danych zastrzeżonych. Excel oraz Tableau są narzędziami do analizy danych, jednak posiadają dwa różne podejścia do eksploracji danych. Excel pracuje z wierszami i kolumnami w arkuszach kalkulacyjnych. Tableau formatuje dane w postaci wykresów i obrazów, które są zdecydowanie bardziej klarowne niż te utworzone w Excelu. Różnica jest najbardziej widoczna w obszarach takich jak interaktywne wizualizacje danych oraz możliwości pracy z danymi na dużą skalę. Tableau Server jest zaprojektowany w sposób umożliwiający połączenie wielu warstw danych. Może łączyć klientów z komputerów stacjonarnych, mobilnych i internetowych. System ten jest wieloużytkownikowy, wieloprocesowy i wielowątkowy [20]. Możliwości oprogramowania Tableau ilustruje rysunek 3.10.



Rys. 3.10. Przykładowy dashboard w programie Tableau

<https://www.tableau.com>

Przekształcanie danych w informacje za pomocą analizy wizualnej musi zostać zrobione dokładnie. Dane nie mogą być zbyt szczegółowe, ponieważ ich interpretacja byłaby utrudniona, a przy tym istnieje ryzyko, że dane utracą swoją zdolność do przedstawiania informacji. Narzędzie do analizy i raportowania powinno posiadać następujące cechy:

- Prostota – łatwe do opanowania przez użytkowników nietechnicznych.
- Łączność – możliwość łączenia z wieloma różnymi źródłami danych.
- Dostępność – ułatwienie dzielenia się spostrzeżeniami.
- Skalowalność – obsługa dużych zbiorów danych.

Tableau Software zostało zaprojektowane w celu zaspokojenia tych wymagań [21].

3.4.3. Platforma Apache Superset

Apache Superset to nowoczesne narzędzie open-source typu Business Intelligence, które służy użytkownikowi do analizy, wizualizacji oraz eksploracji danych. Superset jest elastyczny i łatwy w obsłudze ze względu na dynamiczne elementy pulpitu nawigacyjnego. Apache Superset jest wyjątkowo przydatny dla firm i naukowców, którzy chcą rozpoznawać kluczowe trendy i wzorce z nieprzetworzonych danych, aby pomóc w procesie podejmowania decyzji [22].

Głównym celem Apache Superset jest pomoc w wizualizacji danych oraz ich analiza. Cechy jakie posiada ta platforma to:

- Umożliwia tworzenie własnych wizualizacji i rozszerzenie ich możliwości.
- Pozwala na wykonywanie zapytań SQL, aby móc badać dane.

- Posiada skalowalne rozwiązanie do pobierania danych, które funkcjonuje z istniejącą infrastrukturą danych bez konieczności stosowania oddzielnej warstwy pobierania.
- Wykorzystując bazowe warstwy semantycznej, użytkownik kontroluje, jak źródła danych są wyświetlane i przetwarzane. [23]

Przykład wizualizacji danych w środowisku Apache Superset przedstawia rysunek 3.11.



Rys. 3.11. Przykładowy dashboard w programie Apache Superset
<https://superset.apache.org>

Pulpity nawigacyjne (ang. *Dashboard*) to interfejs użytkownika, który pozwala na przeglądanie różnych wykresów i danych. W świecie danych głównym problemem jest bezpieczeństwo. Apache Superset pozwala nadać różnym użytkownikom różny poziom dostępu. Superset zapewnia funkcjonalność łączenia się z wieloma bazami danych i narzędziami. Bezproblemowo łączy się z prawie wszystkimi głównymi bazami danych. Ułatwia to wizualizację i analizę danych, czyniąc rozwój modelu prostszym. Superset jest kompatybilny z Apache Spark SQL, Google Sheets, Amazon Redshift, Azure MS SQL oraz PostgreSQL. Apache Superset udostępnia szeroką gamę wykresów, tabel i układów. Główną zaletą korzystania ze środowiska jest swoboda, jaką daje użytkownikowi. Zapewnia ona bezpieczeństwo poprzez całkowitą kontrolę nad dostępnością danych w oprogramowaniu. Pozwala na tworzenie użytkowników w bazie danych, dawania im uprawnień oraz śledzenia ich zachowań. W narzędziu tym użytkownik tworzy interaktywne zapytanie dotyczące konkretnych widoków, tabel albo schematów [22][23].

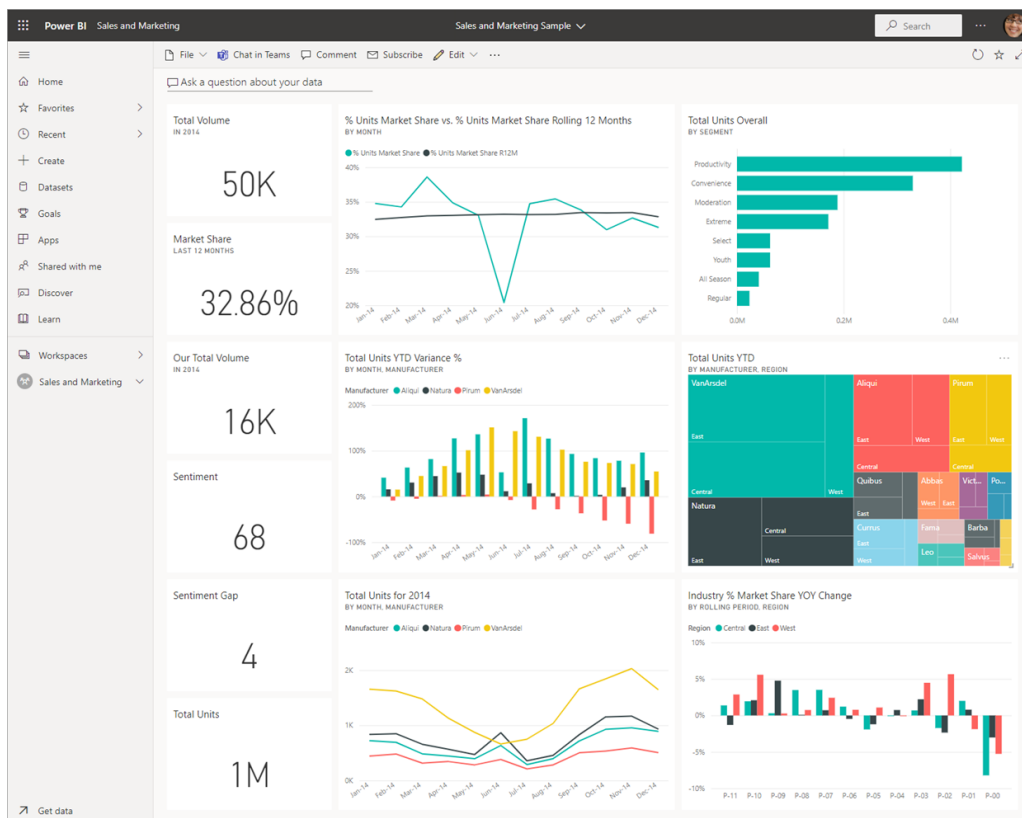
Wizualizacja danych pełni kluczową rolę w cyklu uczenia maszynowego. Ułatwia ona przetwarzanie dużych danych, ponieważ zmniejsza wymagane obciążenie poznawcze. Szybkie znalezienie schematów w dużych zbiorach danych może być szczególnie przydatne do interpretacji złożonych systemów. Wizualizacja danych zawsze była integralną częścią statystyki, ale jest również wykorzystywana w innych profesjach, takich jak informatyka, ekonomia, socjologia, biologia i Business Intelligence [22].

3.4.4. Platforma Power Bi

Power Bi to narzędzie utworzone przez firmę Microsoft. Aplikacja ta to nowoczesne podejście do samoobsługowej analizy Business Intelligence. Opiera się ono na umożliwieniu użytkownikowi korzystania z prostych, ale potężnych narzędzi do przetwarzania danych w ilościach przemysłowych. Power Bi jest to chmurowa usługa, która pozwala na przechowywanie i udostępnianie istotnych danych biznesowych w formie pulpitów i raportów. Do tworzenia pulpitów nawigacyjnych wykorzystuje się Power BI Desktop. Narzędzie to jest całkowicie bezpłatne. Służy do wyszukiwania, oczyszczania i łączenia danych. Przygotowane tabele, wykresy i mapy w formie raportów, użytkownik może udostępnić wybranym odbiorcom w chmurze Microsoft Azure Cloud. Proces przetwarzania danych na platformie Power Bi odbywa się w trzech następujących po sobie krokach: importowanie danych, modelowanie danych, tworzenie raportów [24].

Importowanie danych z różnych źródeł to pierwszy krok w wizualizacji danych za pomocą narzędzia Power Bi. Power BI Desktop pozwala na import danych z różnych miejsc wewnętrznych, jak dane korporacyjne, oraz zewnętrznych, jak media społecznościowe. Scala i integruje dane w jedną spójną strukturę. Oczyszcza dane, aby były wiarygodne i łatwe w użyciu oraz rozbija dane na kolumny i rekordy, które odpowiadają wymaganiom. Modelowanie danych umożliwia rozszerzanie i rozwijanie oczyszczonych danych. Pozwala na tworzenie modelu danych poprzez łączenie tabel w celu uzyskania spójnej struktury danych z wielu oddzielnych źródeł. Ten wzorzec danych jest używany następnie w wizualizacji danych. Modelowanie danych wzbogaca aplikację interaktywną o spójne typy danych i nazwy zmiennych. Oczyszczanie danych nie jest procesem koniecznym podczas ich wizualizacji może zostać pominięty w sytuacji, której dane źródłowe przed oczyszczaniem będą dobrej jakości. Tworzenie wizualizacji danych (ang. *dashboard*) oraz raportów to finalny proces analizy biznesowej.

Analiza Business Intelligence za pomocą narzędzia Power Bi pozwala na przedstawienie informacji składających się z: tabel, macierzy, wykresów, map, wskaźników itd. Dane można filtrować i dzielić w zależności od potrzeb użytkownika [25]. Wizualizację danych w środowisku Power Bi przedstawia rysunek 3.12.



Rys. 3.12. Przykładowy dashboard w programie Power Bi
<https://learn.microsoft.com/en-us/power-bi>

Oprogramowanie Power Bi Desktop umożliwi tworzenie zbiorów stron w jednym pliku. Power Bi Desktop to narzędzie całkowicie niezależne. Oznacza to, że pliki zawierają potrzebne elementy do wizualizacji na każdej ze stron. Służy on do wizualizacji ogromnych zbiorów danych oraz kompresuje dane, które ładowane są w sposób efektywny. Dzięki temu pliki Power BI Desktop zajmują często tylko ułamek miejsca, które zajmowałyby, gdyby zawierały tylko nieprzetworzone dane źródłowe. Kompresja ta dotyczy również danych, które Power BI Desktop wykorzystuje podczas modelowania danych i tworzenia wizualizacji [24].

Tableau, Apache Superset i Power Bi to narzędzia stworzone do wizualizacji źródeł danych wykorzystywanych w branży Business Intelligence. Rysunek 3.13 przedstawia porównanie kluczowych cech, w których każde z oprogramowań się różni. Oprogramowania zostały stworzone do wykonywania konkretnych zadań. Posiadają one swoje indywidualne zalety oraz wady.

Apache Superset	Tableau	Power Bi
Podstawowym celem Supersetu jest uproszczenie procesu przetwarzania, oczyszczania i przeglądania danych. Daje on użytkownikowi możliwość przeprowadzania analiz z dużą prędkością.	Tableau posiada funkcję przeciągnij i upuść, które pozwala użytkownikom na łatwe budowanie interaktywne wizualizacje. Za pomocą zaledwie kilku kliknięć można tworzyć również interaktywne pulpity.	Zapewnia on prosty interfejs typu przeciągnij i upuść oraz możliwość powielania wszystkich formatowanie w porównywalnych wizualizacji.
Ograniczone możliwości dostosowywania	Szczegółowe dostosowanie	Szczegółowe dostosowanie
Wsparcie dla baz danych SQL	Zarówno SQL jak i NoSQL wsparcie dla baz danych	Obsługa baz danych SQL, bez Bezpośrednie wsparcie dla baz danych NoSQL
Dla dużych bazach danych mogą wystąpić opóźnienia podczas przetwarzania	Dobre do obsługi dużych bazą danych	Dla dużych bazach danych mogą wystąpić opóźnienia podczas przetwarzania
Łatwy w użyciu oraz zrozumieniu	Może być trudny dla początkujących	Łatwy w użyciu

Rys. 3.13. Porównanie Apache Superset z Tableau i Power BI

4. Część projektowa

4.1. Założenie budowy ekosystemu BI z wykorzystaniem Big Data

W ramach projektu zrealizowane zostały poniżej wypisane założenia:

- uzasadnienie biznesowe dla utworzenia projektu,
- identyfikacja źródeł danych,
- identyfikacja wymagań początkowych,
- identyfikacja wymagań końcowych,
- architektura z doбором technologii,
- opis implementacji danej architektury.

Poniżej w punktach zostało opisane wykonywanie poszczególnych kroków.

4.1.1. Uzasadnienie biznesowe dla utworzenia projektu

Celem projektu jest utworzenie środowiska, w którym użytkownik będzie mógł analizować dane firmy w celu polepszenia jej funkcjonowania. Stworzony ekosystem zostanie wykorzystany do analizy bazy danych AdventureWorks.

Baza danych AdventureWorks jest przykładową bazą danych wykorzystującą technologię OLTP (Online Transaction Processing) wyprodukowaną przez firmę Microsoft. Poniżej zostały wypisane najważniejsze obiekty, jakie ona zawiera.

- tabele,
- widoki (wyznaczone w oparciu o tabele),
- indeksy,
- procedury składowe,
- funkcje zaprogramowane przez użytkownika,
- typy danych zdefiniowane przez użytkownika,
- schematy,
- katalog tekstowy.

Adventure Works to międzynarodowa firma produkcyjna, która dane firmy magazynuje w bazie danych AdventureWorks. Firma ta sprzedaje akcesoria, rowery, sprzęty rowerowe i ubrania. Baza danych zawiera dane dotyczące różnych scenariuszy biznesowych, takich jak sprzedaż, produkcja i finanse. Posiada ona informacje o około dwudziestu tysiącach klientów, pięciuset produktach oraz ponad siedemdziesięciu tysiącach zamówień. Struktura budowy bazy danych została omówiona szczegółowo w następnej części projektu.

Stan zastany w firmie Adventure Works to relacyjna baza danych, która zawiera istotne informacje dla firmy z następujących tabel pogrupowane na poszczególne grupy danych dotyczące: Zamówienia, Produkcja, Zasoby ludzkie, Klienci, Sprzedaż itd. Firma Adventure

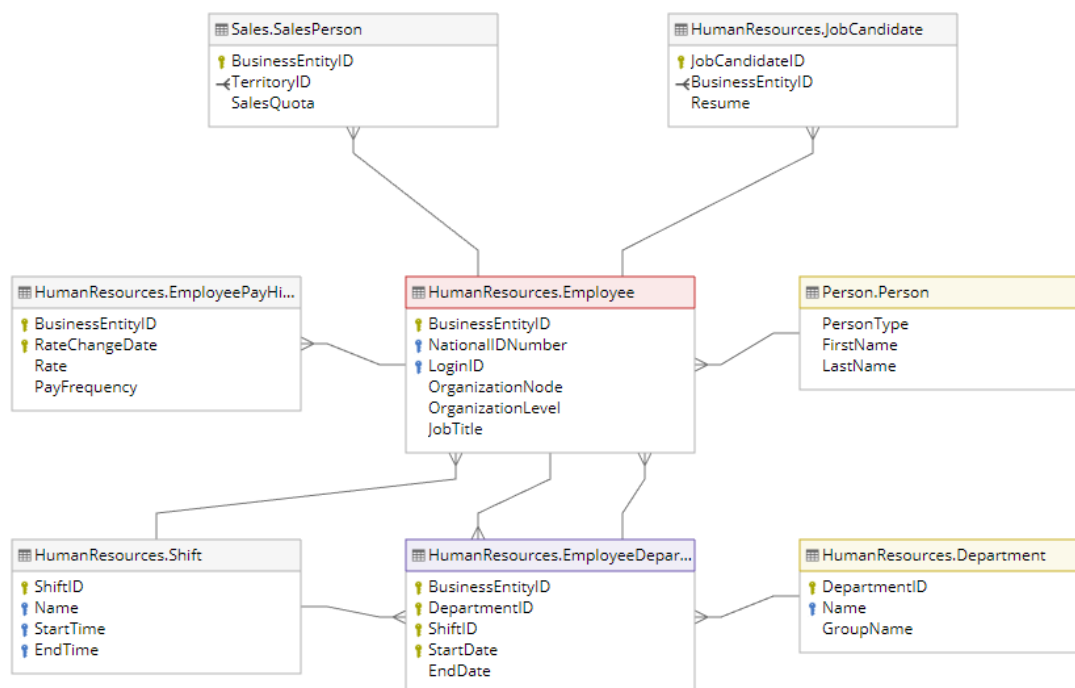
Works chciałaby posiadać platformę, w której mogłaby analizować dane dostarczane w czasie rzeczywistym oraz analizować dane historyczne, aby móc zoptymalizować zarządzanie firmą. Stworzony ekosystem Business Intelligence będzie przekształcał dane w informacje, a informacje w wiedzę, która będzie przedstawiana w przejrzysty sposób za pomocą grafów i wykresów wygenerowanych na platformie Power BI.

4.1.2. Identyfikacja źródeł danych

Źródła danych definiują sposób połączenia z bazą danych. Zawierają również informacje takie jak parametry oraz referencje. Identyfikacja źródeł danych jest krokiem niezbędnym przed wykonaniem analizy, aby można było zdecydować, jakie dane warto przeanalizować. Poniżej wypisane zostały schematy bazy danych AdventureWorks.

Zasoby ludzkie (ang. *Human resources*) zawierają informacje dotyczące kandydatów do pracy, pracowników, ich zarobków oraz lokalizacji pracy. Zarządzanie zasobami ludzkimi (ang. *Human resources management*) odnosi się do zadań, jakie musi wykonać dział kadr w celu zbierania informacji, jakie osoby pracują w firmie oraz jakie prace wykonują. Są oni pośrednikami między pracownikami a zarządem firmy. Tabele oraz relacje w schemacie Zasoby Ludzkie przedstawia rysunek 4.1.

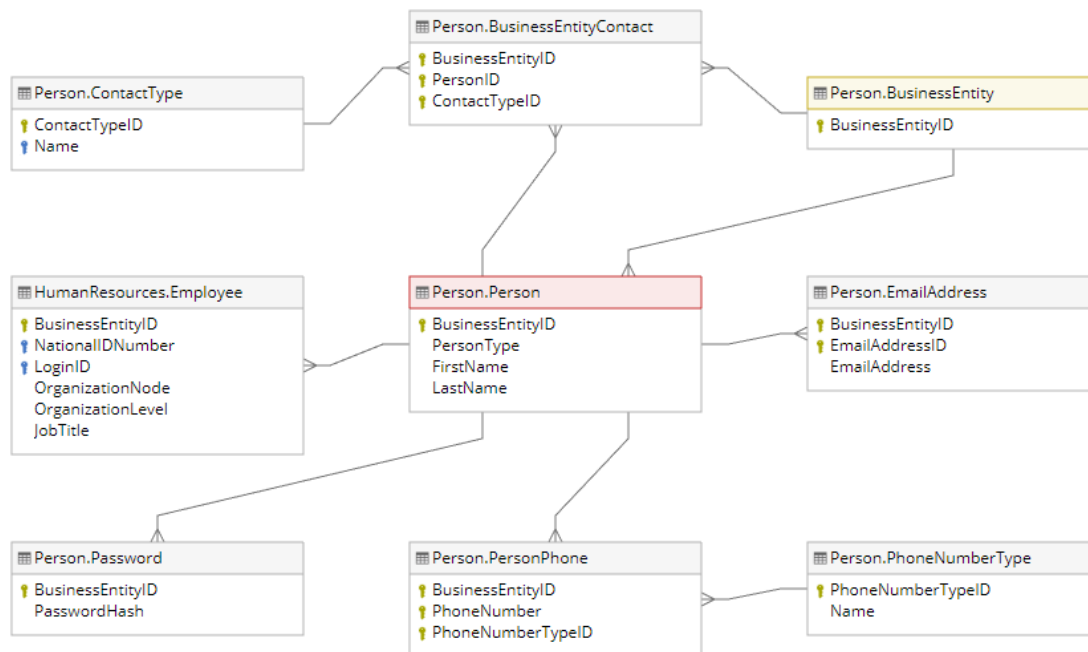
Human Resources



Rys. 4.1. Pracownicy Adventure Works
<https://dataedo.com/samples/html/AdventureWorks>

Schemat Osoby to źródło danych zawierających informacje takie jak nazwy i adresy klientów indywidualnych, pracowników oraz sprzedawców. Zawiera w sobie osiem tabel oraz dwa widoki *Person.vAdditionalContactInfo* (wyświetla nazwę kontaktu i zawartość z każdego elementu w kolumnie xml dla tej osoby), a także widok *Person.vStateProvinceCountryRegion* (połączenie tabeli *CountryRegion* z tabelą *StateProvince*). Rysunek 4.2 przedstawia relacje pomiędzy zbiorami danych oraz informacje jakie dane magazynowane są w konkretnej tabeli.



People








Rys. 4.2. Moduł People w bazie danych AdventureWorks
<https://dataedo.com/samples/html/AdventureWorks>

Zamówienia to równie ważny schemat co poprzednie, ponieważ zawiera informacje o sprzedawcach, od których kupowane są produkty oraz części. Fragment bazy danych jest w relacji ze schematami takimi jak: Osoby, Zasoby Ludzkie, Produkty. Zawiera pięć tabel, dwa widoki oraz jedną funkcję.

Schemat Produkty to źródło danych zawierających informacje o produktach produkowanych i sprzedawanych przez firmę Adventure Works. Struktura schematu produkty została przedstawiona na rysunku 4.3. Rysunek 4.4 przedstawia relacje pomiędzy tabelami w schemacie Produkty.

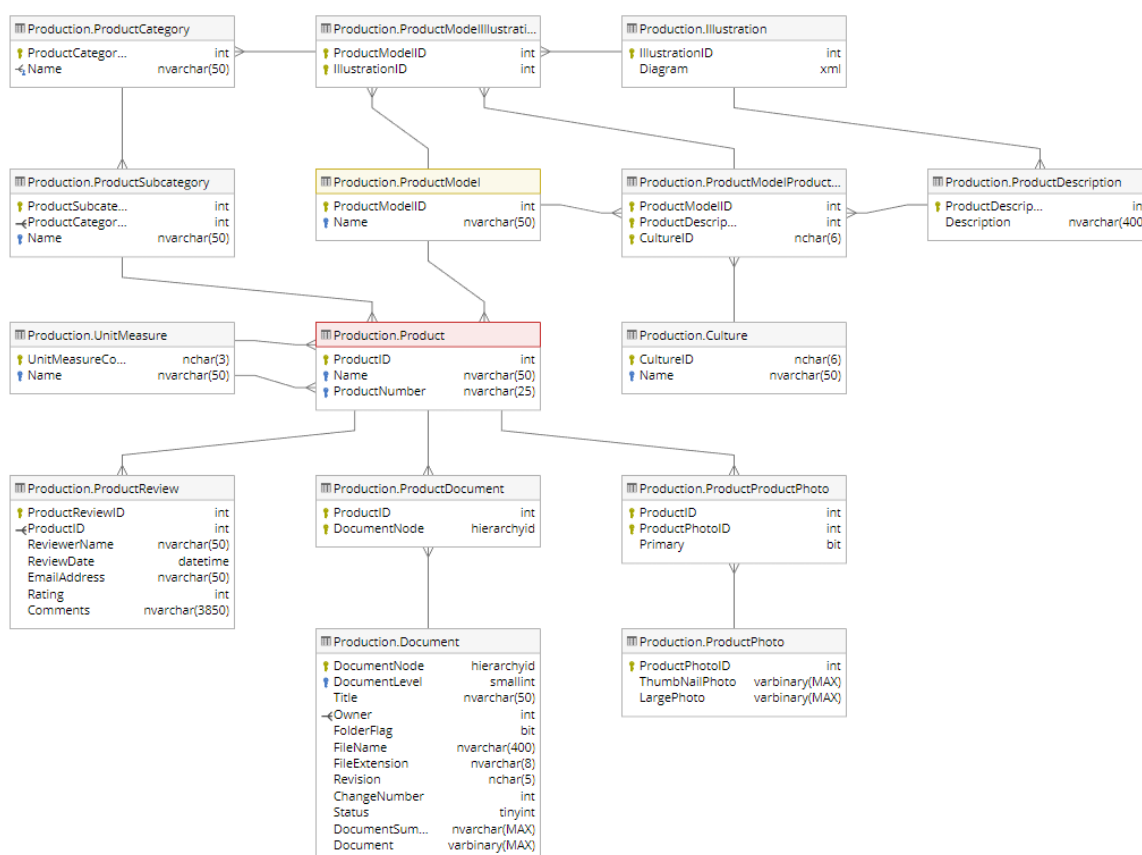
Name		
1		Production.Culture
2		Production.Document
3		Production.Illustration
4		Production.Product
5		Production.ProductCategory
6		Production.ProductDescription
7		Production.ProductDocument
8		Production.ProductModel
9		Production.ProductModelIllustration
10		Production.ProductModelProductDescriptionCulture
11		Production.ProductPhoto
12		Production.ProductProductPhoto
13		Production.ProductReview
14		Production.ProductSubcategory
15		Production.UnitMeasure

Name	
1	 Production.vProductAndDescription
2	 Production.vProductModelCatalogDescription
3	 Production.vProductModelInstructions

		Name
1		dbo.uspGetBillOfMaterials
2		dbo.uspGetWhereUsedProductID

	Name
1	<i>fx</i> dbo.ufnGetDocumentStatusText
2	<i>fx</i> dbo.ufnGetProductDealerPrice
3	<i>fx</i> dbo.ufnGetProductListPrice
4	<i>fx</i> dbo.ufnGetProductStandardCost
5	<i>fx</i> dbo.ufnGetStock

<https://dataedo.com/samples/html/AdventureWorks>



Rys. 4.4. Moduł Produkty w bazie danych AdventureWorks
<https://dataedo.com/samples/html/AdventureWorks>

4.1.3. Identyfikacja wymagań początkowych

Wymagania postawione przez firmy wskazują na zainteresowanie wykorzystaniem danych historycznych w celu uzyskania informacji mających za zadanie maksymalizację przychodów przedsiębiorstwa. Dane pochodzą z wielu źródeł jak: zasoby ludzkie, produkcja, sprzedaż, magazyn, itd. Analiza tych danych jest konieczna w celu optymalizacji firmy. Przedsiębiorstwo Adventure works posiada fabryki zlokalizowane na całym świecie, które importują dane dotyczące produkcji oraz składowania produktów w magazynach. Firma zainwestowała w modernizację lin produkcyjnych w celu optymalizacji procesów produkcji oraz redukcji kosztów np. podczas awarii maszyn.

Architektura Business Intelligence zostanie zaimplementowana dla przedsiębiorstwa w celu polepszenia tego obszaru firmy. Ścieżka danych jest trasą, przez którą dane mogą być przesyłane w ramach systemu komputerowego lub sieci. Odnosi się zarówno do fizycznych, jak i logicznych połączeń. Obejmuje różne elementy jak np. kable, złącza, przełączniki i routery. Ścieżki danych w przedsiębiorstwie Adventure works zostały wykorzystywane do przesyłania różnych typów danych takich jak tekst, obrazy, wideo i audio. Ze względu na posiadane źródła danych będą one dostarczane ścieżką zimną (ang. *cold path*), ponieważ jest ona używana w kontekście systemów komputerowych i sieci. W przeciwieństwie do gorącej ścieżki (ang. *hot path*) dane wykorzystane w ekosystemie nie będą dotyczyć danych dostarczanych w czasie rzeczywistym, a danych historycznych, aby stworzyć optymalny model w odniesieniu do dłuższej perspektywy zarządzania firmą. Zimne ścieżki mogą być przydatne w optymalizacji wydajności systemu i alokacji zasobów, ponieważ są używane do obsługi rzadziej dostępnych danych lub zadań.

Architektura systemu przetwarzania danych będzie zbierać dane z różnych źródeł danych oraz je identyfikować. Dane te będą przepływać do magazynów, gdzie będą modelowane oraz organizowane. Następnym procesem w architekturze będzie zarządzanie metadanymi w zależności od potrzeb operacyjnych oraz kontroli bezpieczeństwa przedsiębiorstwa. Przetworzone dane będą udostępniane do kolejnych narzędzi, gdzie ostatnim etapem będzie analiza utworzonych raportów i wizualizacja kluczowych dla przedsiębiorstwa danych.

4.1.4. Identyfikacja wymagań końcowych

Przypadek użycia (ang. *use case*) jest opisem określonego zestawu działań lub zdarzeń, które są wykonywane w celu osiągnięcia określonego celu lub zadania. W kontekście systemów komputerowych i oprogramowania przypadek użycia opisuje kroki, przez które przechodzi użytkownik w celu wykonania określonego zadania lub procesu. Zostanie on wykorzystany w celu rozwoju projektowanego ekosystemu przedsiębiorstwa Adventure works. Pomoże również w identyfikacji wymagań i funkcji, które są niezbędne do wspierania pożądaných zadań

i procesów użytkownika. Dla przedsiębiorstwa zostanie on w głównej mierze wykorzystany do optymalizacji procesów produkcyjnych oraz zarządzania zasobami ludzkim a także w celu testowania i walidacji funkcjonowania architektury.

Istnieje wiele różnych sposobów prezentacji i wykorzystania danych, w zależności od konkretnych celów i kontekstu danych. Użyte w projekcie sposoby prezentacji danych obejmują:

- Tabele dane mogą być prezentowane w formacie tabelarycznym, z wierszami i kolumnami używanymi do organizacji i wyświetlania danych.
- Diagramy oraz wykresy dane można wizualizować za pomocą wykresów i diagramów, które mogą pomóc w łatwiejszym zrozumieniu i interpretacji danych.
- Mapy termiczne (ang. *Heatmap*) dane (często geograficzne) można przedstawić za pomocą map, które pomogą w pokazaniu przestrzennego rozmieszczenia danych.

Wykorzystanie danych obejmuje:

- Analizę danych – obejmuje to analizę statystyczną, uczenie maszynowe lub inne techniki.
- Wizualizację danych – dotyczy to tworzenia wykresów, diagramów lub innych rodzajów wizualizacji.
- Raportowanie danych – dane mogą być wykorzystywane do tworzenia raportów, które podsumowują i prezentują kluczowe punkty w danych.
- Przechowywanie i wyszukiwanie danych – obejmuje to wykorzystanie baz danych lub innych rodzajów systemów przechowywania danych.

Finalnie stworzona architektura będzie ekosystemem, w którym użytkownik będzie mógł: pozyskiwać i przechowywać dane, integrować dane z różnych źródeł do odczytu, modelować i organizować dane, udostępniać dane, administrować dostęp do informacji, harmonogramować, wizualizować dane, generować raporty oraz wykresy, zarządzać metadanymi, analizować uzyskane informacje z danych.

4.2. Architektura z doborem technologii

4.2.1 Architektury ekosystemu BI

Architektura ekosystemu Business Intelligence zależy od wielu czynników takich jak: cel systemu, wymagania dotyczące wydajności i skalowalności, a także ograniczenia budżetowe. Dla przedsiębiorstwa Adventure works została stworzona architektura, której struktura uwzględnia konkretne potrzeby firmy. Wymagania te zostały opisane szczegółowo podczas identyfikacji wymagań. Kształt architektury powstał na podstawie elementów budowy struktury firmy Adventure Works. Zostały w nim uwzględnione takie elementy jak: źródła danych, magazyny danych, warstwa przetwarzania oraz integracji danych, narzędzia analityczne, warstwa wizualizacji, zabezpieczenia. Projekt ekosystem Business Intelligence to zestaw narzędzi i rozwiązań, które służą do zbierania, przetwarzania, analizowania i prezentowania danych biznesowych dla podanej wyżej firmy.

4.2.2 Harmonogramowanie, zarządzanie metadanymi oraz administracja bezpieczeństwem ekosystemu

Apache Airflow jest platformą do definiowania, planowania i monitorowania przepływów pracy. Został on wykorzystany w ekosystemie ze względu na jego zdolności do przetwarzania danych oraz współpracy z ETL.

Airflow jest narzędziem do orkiestracji przepływów pracy, dlatego odpowiedzialny był w zarządzaniu przedsiębiorstwem za koordynację wykonywanymi zadaniami. Odpowiadał również za zależności między nimi, zapewniając, że zadania zostaną uruchamiane w odpowiedniej kolejności i we właściwym czasie. Platforma wykorzystuje graf skierowany acykliczny DAG do reprezentowania przepływów pracy, gdzie każdy węzeł w grafie reprezentuje zadanie, a krawędzie między węzłami zależność [26].

Airflow dostarcza szeregu funkcji do zarządzania i monitorowania przepływów pracy, które są niezbędne podczas kontrolowania prawidłowego funkcjonowania przedsiębiorstwa:

- Interfejs użytkownika do definiowania i harmonogramowania przepływów pracy.
- Harmonogram do wykonywania zadań regularnie albo nieregularnie na żądanie administratora.
- Oprogramowanie zapewniające wsparcie dla uruchamiania w różnych środowiskach, w tym w systemach lokalnych, zdalnych i opartych na chmurze.
- Powiadomienia oraz alerty do śledzenia statusu zadań i przepływów pracy [26].

Apache Ranger to framework bezpieczeństwa umożliwiający scentralizowaną administrację bezpieczeństwem i kontrolę dostępu dla Apache Hadoop. W ekosystemie odpowiadał za pomoc w organizacji zabezpieczeń oraz zarządzał dostępem do zasobów danych

przechowywanych w Hadoop. Apache Ranger w projekcie został użyty do zapewnienia precyzyjnej kontroli dostępu do danych i zasobów używanych przez narzędzia Business Intelligence. Ranger odpowiadał za definiowanie i egzekwowanie polityki dostępu dla użytkowników i grup oraz kontrolował dostęp do określonych zbiorów danych lub tabel. Jednocześnie monitorował i audytował dostęp do zasobów danych. [27].

Metadane to dane, które przekazują informacje o innych danych i często są wykorzystywane do zapewnienia dodatkowego kontekstu i znaczenia zbiorów danych. Opisują cechy zbiorów danych, takie jak ich format, zawartość i struktura. Metadane są często wykorzystywane w celu ułatwienia organizacji, zarządzania i zrozumienia zasobów danych. Business Intelligence wykorzystuje metadane, aby pomóc przedsiębiorstwom zrozumieć i nadać sens posiadanym danym. Ekosystem będzie wykorzystywał metadane w czterech głównych sektorach: odkrywanie danych, ich pochodzenie, jakość danych oraz zarządzanie nimi. Tworzenie, utrzymywanie i wykorzystywanie metadanych w celu wspierania bieżących działań organizacji zawierają się w kontekście celów operacyjnych zarządzania nim. W ramach budowy architektury ekosystemu wykorzystana została platforma Data Hub, która przechowuje i zarządza danymi z wielu źródeł. Środowisko Data Hub w kontekście ekosystemu Business Intelligence zostało użyte do: integracji danych z wielu źródeł, przekształcania i oczyszczania danych, eksploracji i wstępnej analizy danych [29].

4.2.3 Przetwarzanie danych w ekosystemie

Scenariusz przedsiębiorstwa wymaga wykorzystania danych historycznych do tworzenia raportów oraz wykresów. Oznacza to, że do budowy architektury ekosystemu została wybrana architektura Lambda ze względu na warstwę wsadowa. Przetwarzanie odbędzie się kosztem czasu natomiast architektura oferuje wysoką jakość i pełną historię danych. Dane gromadzone przez Adventure works należą do dwóch grup: dane ustrukturyzowane oraz dane półustrukturyzowane. Architektura systemu nie zakłada analizy danych ze ścieżki gorącej (ang. *hot path*), ponieważ jej założeniem jest rozwinięcie przedsiębiorstwa poprzez spojrzenie na nią z dłuższej perspektywy. Dlatego nie powstanie model klucz-wartość dla danych z czujników sensorowych umieszczonych na maszynach, a model relacyjny dla danych w firmie dotyczących: zasobów ludzkich, magazynów, produkcji oraz sprzedaży. Apache Spark to narzędzie wykorzystane przy budowie architektury ekosystemu ze względu na przetwarzanie potoku danych z systemów wyżej wymienionych. Apache Spark, jako wydajny i szybki silnik do przetwarzania danych, został systemem ELT. Zadaniem Apache Spark jest przetwarzanie danych za pomocą filtrowania, grupowania i agregowania z różnych źródeł, a następnie przenoszenie ich do magazynu danych [28].

4.2.4 Magazyn danych i organizacja danych

W architekturze ekosystemu został wykorzystany Hadoop do utworzenia magazynu danych. Umożliwia on przetwarzanie dużych zbiorów danych, co dla przedsiębiorstwa Adventure works jest ważne w kontekście Business Intelligence w celu analizy danych historycznych. Struktura Hadoop składa się z : HDFS (ang. *Hadoop Distributed File System*), Map Reduce oraz YARN (ang. *Yet Another Resource Negotiator*).

Hadoop Distributed File System (HDFS) to dystrybuowana struktura plików, która została specjalnie zaprojektowana do przechowywania dużych ilości danych na zestawie komputerów. MapReduce to narzędzie do przetwarzania danych w sposób równoległy na klastrach. Pozwala ono na szybkie przetwarzanie dużych zbiorów danych. YARN to system do zarządzania zasobami klastra.

Zadaniem HDFS w ekosystemie jest bycie magazynem danych, które są dostarczane przez Apache Spark po ich przetworzeniu. Środowisko HDFS zostało wykorzystane dla zastosowań Business Intelligence ze względu na możliwości jakie oferuje czyli skalowalność, elastyczność oraz integracja z innymi narzędziami Business Intelligence jak Tableau czy Power BI [30].

4.2.5 Udostępnianie oraz analiza danych

Za udostępnianie danych w ekosystemie odpowiedzialny był Apache Druid narzędzie do wykonywania szybkich zapytań analitycznych. Druid jest szczególnie przydany dla Business Intelligence ze względu na przyspieszenie operacji w Apache Spark. Również procesy przetwarzania danych ETL mogą zostać wykonane za pomocą Apache Druid. Dane przetworzone można zapisać w formatach JSON, CSV, itd. Następnie połączyć Druida z innymi narzędziami do wizualizacji danych jak Power Bi, Apache Superset lub inne. Integracja Apache Druid z Power BI umożliwia szybkie wizualizowanie danych analitycznych przechowywanych w Druidzie za pomocą wykresów, tabel itp. Power BI Desktop łączy się z Apache Druid poprzez adres URL Druidowej instancji. Wybranie odpowiednich baz danych oraz tabel sprawi że Power Bi wyświetli dostępne dane w postaci listy pól. Odpowiednie zapytania do Apache Druid pozwalają na skonfigurowanie go tak, aby pobierał odpowiednie dane do wizualizacji, które zostaną przekazane do narzędzia Power Bi. Środowisko Power BI może być wykorzystane do przetwarzania, wizualizowania i raportowania danych z Adventure Works. Stworzyło ono różne wizualizacje, takie jak wykresy słupkowe, kołowe, liniowe i inne, które przedstawiają dane z Adventure Works w czytelny i przejrzysty sposób [28].

4.2.6 Diagram architektury oraz tabela użytych technologii

W proponowanej architekturze ekosystemu użytkownik lub analityk biznesowy posiada cztery punkty dostępu do danych:

- widok warstwy wsadowej w architekturze Lambda,
- dostęp do danych surowych w HDFS,
- przetwarzania wizualizacji oraz udostępniania danych w środowisku Apache Druid,
- raporty oraz wykresy danych w środowisku Power Bi.

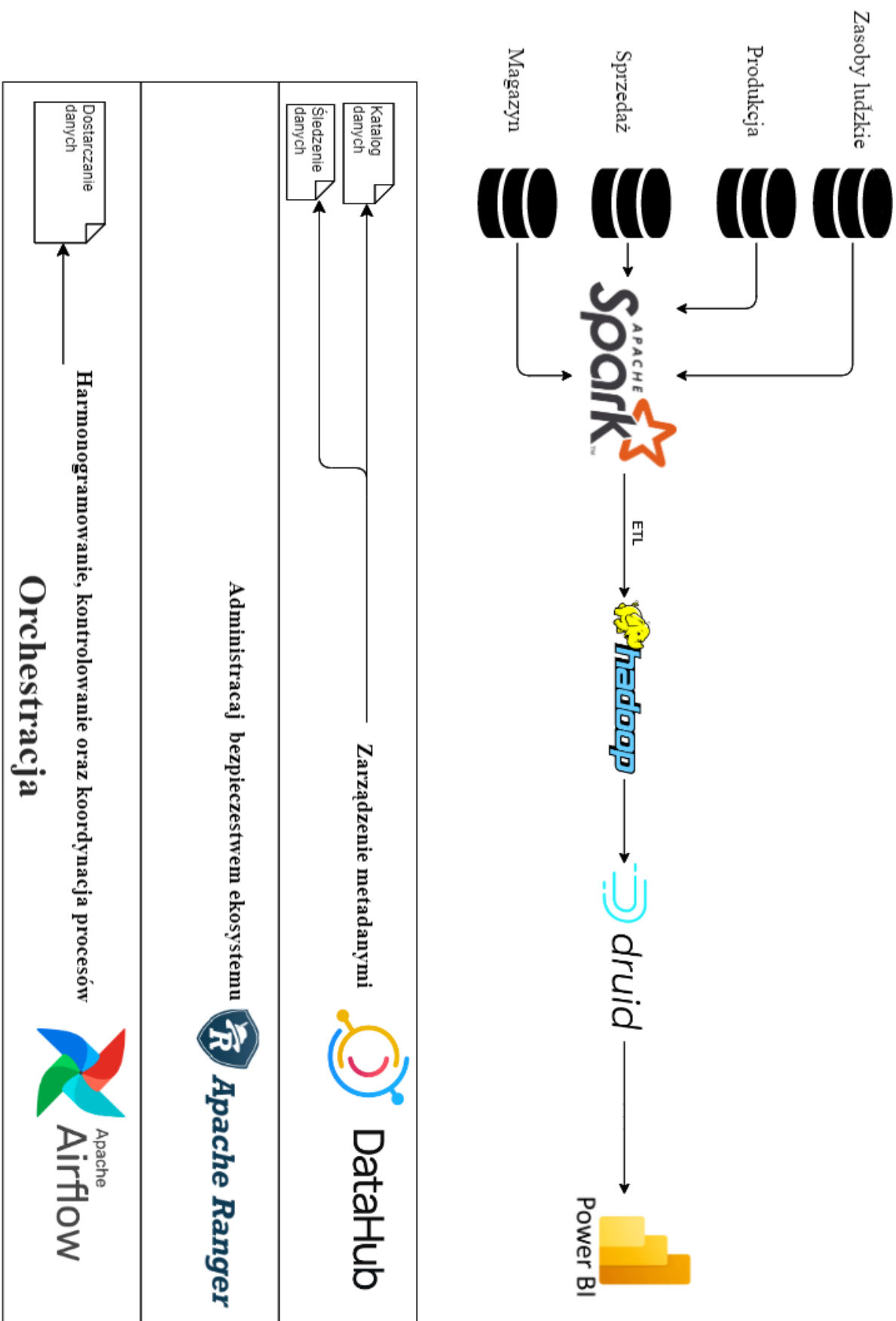
Przedstawiony ekosystem spełnia założenia przedsiębiorstwa Adventure works w odniesieniu do analizowanych scenariuszy. Ekosystem umożliwia analizę danych historycznych uzyskanych z danych wsadowych generowanych w bazach danych. System przetwarzania danych umożliwia tworzenie raportów zawierających wiedzę w zgromadzonych danych oraz udostępnia te zbiory danych do analizy Business Intelligence analitykom i użytkownikom biznesowym danego przedsiębiorstwa. Oferuje również kompleksową, wielomodelową warstwę integracji danych. W systemie jest również warstwa administracji bezpieczeństwa oraz warstwa zarządzania metadanymi, które odpowiadają za nadawanie uprawnień określonym użytkownikom oraz kontrolowanie jakości danych i katalogowanie metadanych. Orchestracja odpowiada w ekosystemie za samoobsługowe dostarczanie danych poprzez koronacje i harmonogramowanie pracy.

Przedstawiona architektura może zostać przyszłościowo rozwinięta poprzez stworzenie ścieżki gorącej (ang. *hot path*), która odpowiadałaby za analizę danych pochodzących z widoku czasu rzeczywistego. Architektura lambda pozwala na analizę danych pochodzących z sensorów zamieszczonych w fabrykach przedsiębiorstwa Adventure works.

Rysunek 4.5 przedstawia wybrane technologie oraz podejścia architektoniczne, które zostały wybrane do stworzenia architektury ekosystemu Business Intelligence w oparciu o środowisko Big Data. Diagram architektury proponowanego ekosystemu zawierający opis połączeń oraz relacji pomiędzy poszczególnymi warstwami został zaprezentowany na rysunku 4.6.

Warstwa	Rozwiązanie architektoniczne
Orchestracja	Apache Airflow, dostarczanie danych
Administracja i bezpieczeństwo	Apache Ranger
Zarządzanie metadanymi	Data Hub, katalog oraz śledzenie danych
Warstwa dostarczania i integracji	Poziomy dostęp: <ul style="list-style-type: none"> • widok warstwy wsadowej w architekturze Lambda, • dostęp do danych surowych w HDFS, • przetwarzania wizualizacji oraz udostępniania danych w środowisku Apache Druid, • raporty oraz wykresy danych w środowisku Power Bi.
Przetwarzanie danych i przepływ danych	Apache spark, architektura Lambda, Apache druid
Modelowanie danych	<ul style="list-style-type: none"> • Dane surowe • Model relacyjne
Magazyn danych	Hadoop Distributed File System
Udostępnianie danych	Apache Druid
Wizualizacja i analiza danych	Power Bi

Rys. 4.5. Wybrane technologie do budowy ekosystemu



Rys. 4.6. Diagram architektury ekosystemu


```
scala> val df = spark.read.csv("D:/inzynierka/AdventureWorks_Products.csv");
df: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 9 more fields]
```

```
scala> df.show(35);
```

	_c0	_c1	_c2	_c3	_c4	_c5	_c6	_c7	_c8	_c9	_c10
	ProductKey	ProductSubcategory	ProductSKU	Product Name	ModelName	ProductDescription	ProductColor	ProductSize	ProductStyle	ProductCost	ProductPrice
214		31	HL-U509-R	Sport-100 Helmet,...	Sport-100	Universal fit, we...	Red	0	0	13.0863	34.99
215		31	HL-U509	Sport-100 Helmet,...	Sport-100	Universal fit, we...	Black	0	0	12.0278	33.6442
218		23	SO-B909-M	Mountain Bike Soc...	Mountain Bike Socks	Combination of na...	White	M	U	3.3963	9.5
219		23	SO-B909-L	Mountain Bike Soc...	Mountain Bike Socks	Combination of na...	White	L	U	3.3963	9.5
220		31	HL-U509-B	Sport-100 Helmet,...	Sport-100	Universal fit, we...	Blue	0	0	12.0278	33.6442
223		19	CA-1098	AWC Logo Cap	Cycling Cap	Traditional style...	Multi	0	U	5.7052	8.6442
226		21	LJ-0192-S	Long-Sleeve Logo ...	Long-Sleeve Logo ...	Unisex long-sleev...	Multi	S	U	31.7244	48.0673
229		21	LJ-0192-M	Long-Sleeve Logo ...	Long-Sleeve Logo ...	Unisex long-sleev...	Multi	M	U	31.7244	48.0673
232		21	LJ-0192-L	Long-Sleeve Logo ...	Long-Sleeve Logo ...	Unisex long-sleev...	Multi	L	U	31.7244	48.0673
235		21	LJ-0192-X	Long-Sleeve Logo ...	Long-Sleeve Logo ...	Unisex long-sleev...	Multi	XL	U	31.7244	48.0673
238		14	FR-R92R-62	HL Road Frame - R...	HL Road Frame	Our lightest and ...	Red	62	U	747.9682	1263.4598
241		14	FR-R92R-44	HL Road Frame - R...	HL Road Frame	Our lightest and ...	Red	44	U	747.9682	1263.4598
244		14	FR-R92R-48	HL Road Frame - R...	HL Road Frame	Our lightest and ...	Red	48	U	747.9682	1263.4598
247		14	FR-R92R-52	HL Road Frame - R...	HL Road Frame	Our lightest and ...	Red	52	U	747.9682	1263.4598
250		14	FR-R92R-56	HL Road Frame - R...	HL Road Frame	Our lightest and ...	Red	56	U	747.9682	1263.4598
253		14	FR-R38B-58	LL Road Frame - B...	LL Road Frame	The LL Frame prov...	Black	58	U	176.1997	297.6346
256		14	FR-R38B-60	LL Road Frame - B...	LL Road Frame	The LL Frame prov...	Black	60	U	176.1997	297.6346
259		14	FR-R38B-62	LL Road Frame - B...	LL Road Frame	The LL Frame prov...	Black	62	U	176.1997	297.6346
262		14	FR-R38R-44	LL Road Frame - R...	LL Road Frame	The LL Frame prov...	Red	44	U	181.4857	306.5636
264		14	FR-R38R-48	LL Road Frame - R...	LL Road Frame	The LL Frame prov...	Red	48	U	181.4857	306.5636
266		14	FR-R38R-52	LL Road Frame - R...	LL Road Frame	The LL Frame prov...	Red	52	U	181.4857	306.5636
268		14	FR-R38R-58	LL Road Frame - R...	LL Road Frame	The LL Frame prov...	Red	58	U	181.4857	306.5636
270		14	FR-R38R-60	LL Road Frame - R...	LL Road Frame	The LL Frame prov...	Red	60	U	181.4857	306.5636
272		14	FR-R38R-62	LL Road Frame - R...	LL Road Frame	The LL Frame prov...	Red	62	U	181.4857	306.5636
274		14	FR-R72R-44	ML Road Frame - R...	ML Road Frame	Made from the sam...	Red	44	U	352.1394	594.83
275		14	FR-R72R-48	ML Road Frame - R...	ML Road Frame	Made from the sam...	Red	48	U	352.1394	594.83
276		14	FR-R72R-52	ML Road Frame - R...	ML Road Frame	Made from the sam...	Red	52	U	352.1394	594.83
277		14	FR-R72R-58	ML Road Frame - R...	ML Road Frame	Made from the sam...	Red	58	U	352.1394	594.83
278		14	FR-R72R-60	ML Road Frame - R...	ML Road Frame	Made from the sam...	Red	60	U	352.1394	594.83
279		14	FR-R38B-44	LL Road Frame - B...	LL Road Frame	The LL Frame prov...	Black	44	U	176.1997	297.6346
282		14	FR-R38B-48	LL Road Frame - B...	LL Road Frame	The LL Frame prov...	Black	48	U	176.1997	297.6346
285		14	FR-R38B-52	LL Road Frame - B...	LL Road Frame	The LL Frame prov...	Black	52	U	176.1997	297.6346
288		12	FR-M94S-42	HL Mountain Frame...	HL Mountain Frame	Each frame is han...	Silver	42	U	623.8403	1204.3248
291		12	FR-M94S-44	HL Mountain Frame...	HL Mountain Frame	Each frame is han...	Silver	44	U	706.811	1364.5

only showing top 35 rows

Rys. 4.8. Przypisanie do zmiennej bazy danych Adventure works Products

```
scala> val df = spark.read.option("header", true).csv("D:/inzynierka/AdventureWorks_Products.csv");
df: org.apache.spark.sql.DataFrame = [ProductKey: string, ProductSubcategoryKey: string ... 9 more fields]
```

```
scala> df.show(35);
```

	ProductKey	ProductSubcategoryKey	ProductSKU	Product Name	ModelName	ProductDescription	ProductColor	ProductSize	ProductStyle	ProductCost	ProductPrice
214		31	HL-U509-R	Sport-100 Helmet,...	Sport-100	Universal fit, we...	Red	0	0	13.0863	34.99
215		31	HL-U509	Sport-100 Helmet,...	Sport-100	Universal fit, we...	Black	0	0	12.0278	33.6442
218		23	SO-B909-M	Mountain Bike Soc...	Mountain Bike Socks	Combination of na...	White	M	U	3.3963	9.5
219		23	SO-B909-L	Mountain Bike Soc...	Mountain Bike Socks	Combination of na...	White	L	U	3.3963	9.5
220		31	HL-U509-B	Sport-100 Helmet,...	Sport-100	Universal fit, we...	Blue	0	0	12.0278	33.6442
223		19	CA-1098	AWC Logo Cap	Cycling Cap	Traditional style...	Multi	0	U	5.7052	8.6442
226		21	LJ-0192-S	Long-Sleeve Logo ...	Long-Sleeve Logo ...	Unisex long-sleev...	Multi	S	U	31.7244	48.0673
229		21	LJ-0192-M	Long-Sleeve Logo ...	Long-Sleeve Logo ...	Unisex long-sleev...	Multi	M	U	31.7244	48.0673
232		21	LJ-0192-L	Long-Sleeve Logo ...	Long-Sleeve Logo ...	Unisex long-sleev...	Multi	L	U	31.7244	48.0673
235		21	LJ-0192-X	Long-Sleeve Logo ...	Long-Sleeve Logo ...	Unisex long-sleev...	Multi	XL	U	31.7244	48.0673
238		14	FR-R92R-62	HL Road Frame - R...	HL Road Frame	Our lightest and ...	Red	62	U	747.9682	1263.4598
241		14	FR-R92R-44	HL Road Frame - R...	HL Road Frame	Our lightest and ...	Red	44	U	747.9682	1263.4598
244		14	FR-R92R-48	HL Road Frame - R...	HL Road Frame	Our lightest and ...	Red	48	U	747.9682	1263.4598
247		14	FR-R92R-52	HL Road Frame - R...	HL Road Frame	Our lightest and ...	Red	52	U	747.9682	1263.4598
250		14	FR-R92R-56	HL Road Frame - R...	HL Road Frame	Our lightest and ...	Red	56	U	747.9682	1263.4598
253		14	FR-R38B-58	LL Road Frame - B...	LL Road Frame	The LL Frame prov...	Black	58	U	176.1997	297.6346
256		14	FR-R38B-60	LL Road Frame - B...	LL Road Frame	The LL Frame prov...	Black	60	U	176.1997	297.6346
259		14	FR-R38B-62	LL Road Frame - B...	LL Road Frame	The LL Frame prov...	Black	62	U	176.1997	297.6346
262		14	FR-R38R-44	LL Road Frame - R...	LL Road Frame	The LL Frame prov...	Red	44	U	181.4857	306.5636
264		14	FR-R38R-48	LL Road Frame - R...	LL Road Frame	The LL Frame prov...	Red	48	U	181.4857	306.5636
266		14	FR-R38R-52	LL Road Frame - R...	LL Road Frame	The LL Frame prov...	Red	52	U	181.4857	306.5636
268		14	FR-R38R-58	LL Road Frame - R...	LL Road Frame	The LL Frame prov...	Red	58	U	181.4857	306.5636
270		14	FR-R38R-60	LL Road Frame - R...	LL Road Frame	The LL Frame prov...	Red	60	U	181.4857	306.5636
272		14	FR-R38R-62	LL Road Frame - R...	LL Road Frame	The LL Frame prov...	Red	62	U	181.4857	306.5636
274		14	FR-R72R-44	ML Road Frame - R...	ML Road Frame	Made from the sam...	Red	44	U	352.1394	594.83
275		14	FR-R72R-48	ML Road Frame - R...	ML Road Frame	Made from the sam...	Red	48	U	352.1394	594.83
276		14	FR-R72R-52	ML Road Frame - R...	ML Road Frame	Made from the sam...	Red	52	U	352.1394	594.83
277		14	FR-R72R-58	ML Road Frame - R...	ML Road Frame	Made from the sam...	Red	58	U	352.1394	594.83
278		14	FR-R72R-60	ML Road Frame - R...	ML Road Frame	Made from the sam...	Red	60	U	352.1394	594.83
279		14	FR-R38B-44	LL Road Frame - B...	LL Road Frame	The LL Frame prov...	Black	44	U	176.1997	297.6346
282		14	FR-R38B-48	LL Road Frame - B...	LL Road Frame	The LL Frame prov...	Black	48	U	176.1997	297.6346
285		14	FR-R38B-52	LL Road Frame - B...	LL Road Frame	The LL Frame prov...	Black	52	U	176.1997	297.6346
288		12	FR-M94S-42	HL Mountain Frame...	HL Mountain Frame	Each frame is han...	Silver	42	U	623.8403	1204.3248
291		12	FR-M94S-44	HL Mountain Frame...	HL Mountain Frame	Each frame is han...	Silver	44	U	706.811	1364.5
292		12	FR-M94S-52	HL Mountain Frame...	HL Mountain Frame	Each frame is han...	Silver	48	U	706.811	1364.5

only showing top 35 rows

Rys. 4.9. Usunięcie pierwszej kolumny w danych Adventure works – Products

Rysunek 4.8 ilustruje nieprzetworzone dane posiadające błędnie zdefiniowany nagłówek. Rysunek 4.9 przedstawia nagłówek poprawiony poprzez usunięcie pierwszego wiersza w pliku CSV bazy danych. Proces poprawiania błędnie zdefiniowanej struktury przedstawia rysunek 4.10.

```
scala> val df = spark.read.option("header", true).option("infernaSchema", true).csv("D:/inzynierka/AdventureWorks_Products.csv");
df: org.apache.spark.sql.DataFrame = [ProductKey: string, ProductSubcategoryKey: string ... 9 more fields]

scala> df.printSchema
root
|-- ProductKey: string (nullable = true)
|-- ProductSubcategoryKey: string (nullable = true)
|-- ProductSKU: string (nullable = true)
|-- ProductName: string (nullable = true)
|-- ModelName: string (nullable = true)
|-- ProductDescription: string (nullable = true)
|-- ProductColor: string (nullable = true)
|-- ProductSize: string (nullable = true)
|-- ProductStyle: string (nullable = true)
|-- ProductCost: string (nullable = true)
|-- ProductPrice: string (nullable = true)

scala> import org.apache.spark.sql.types.{StructType, StructField, StringType, IntegerType};
import org.apache.spark.sql.types.{StructType, StructField, StringType, IntegerType}

scala> val mySchema = StructType(Array(
  | StructField("ProductKey", IntegerType, true),
  | StructField("ProductSubcategoryKey", IntegerType, true),
  | StructField("ProductSKU", StringType, true),
  | StructField("ProductName", StringType, true),
  | StructField("ModelName", StringType, true),
  | StructField("ProductDescription", StringType, true),
  | StructField("ProductColor", StringType, true),
  | StructField("ProductStyle", StringType, true),
  | StructField("ProductSize", StringType, true),
  | StructField("ProductCost", IntegerType, true),
  | StructField("ProductPrice", IntegerType, true)));
mySchema: org.apache.spark.sql.types.StructType = StructType(StructField(ProductKey,IntegerType,true),StructField(ProductSubcategoryKey,IntegerType,true),StructField(ProductSKU,StringType,true),StructField(ProductName,StringType,true),StructField(ModelName,StringType,true),StructField(ProductDescription,StringType,true),StructField(ProductColor,StringType,true),StructField(ProductStyle,StringType,true),StructField(ProductSize,StringType,true),StructField(ProductCost,IntegerType,true),StructField(ProductPrice,IntegerType,true))

scala> val df = spark.read.option("header", true).schema(mySchema).csv("D:/inzynierka/AdventureWorks_Products.csv");
df: org.apache.spark.sql.DataFrame = [ProductKey: int, ProductSubcategoryKey: int ... 9 more fields]

scala> df.printSchema
root
|-- ProductKey: integer (nullable = true)
|-- ProductSubcategoryKey: integer (nullable = true)
|-- ProductSKU: string (nullable = true)
|-- ProductName: string (nullable = true)
|-- ModelName: string (nullable = true)
|-- ProductDescription: string (nullable = true)
|-- ProductColor: string (nullable = true)
|-- ProductStyle: string (nullable = true)
|-- ProductSize: string (nullable = true)
|-- ProductCost: integer (nullable = true)
|-- ProductPrice: integer (nullable = true)
```

Rys. 4.10. Formatowanie błędnych typów danych w bazie danych Adventure Works – Products

Kluczowym procesem, który musiał zostać wykonany, było usunięcie pierwszego rzędu z pliku CSV bazy danych Adventure Works Products. Proces ten był wymagany ze względu na poprawienie typu danych wartości takich jak: *ProductKey*, *ProductCost*, *ProductPrice*. Te dane posiadały niepoprawny format danych tekstowych (ang. *string*), co uniemożliwiłoby późniejszą analizę w narzędziach analitycznych. Typ danych został zmieniony na typ całkowity (ang. *integer*) czyli prawidłowy dla danych numerycznych jak np. liczby całkowite. Oprócz tego z bazy zostały również usunięte dane niekompletne. Proces ten był konieczny do wykonania również w pozostałych bazach danych, które zostały przeanalizowane w narzędziu Power Bi.

```
scala> val expensiveProducts = df.filter("ProductPrice > 3000");
expensiveProducts: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [ProductKey: string, ProductSubcategoryKey: string ... 9 more fields]

scala> expensiveProducts.orderBy("ProductPrice").show(20);
```

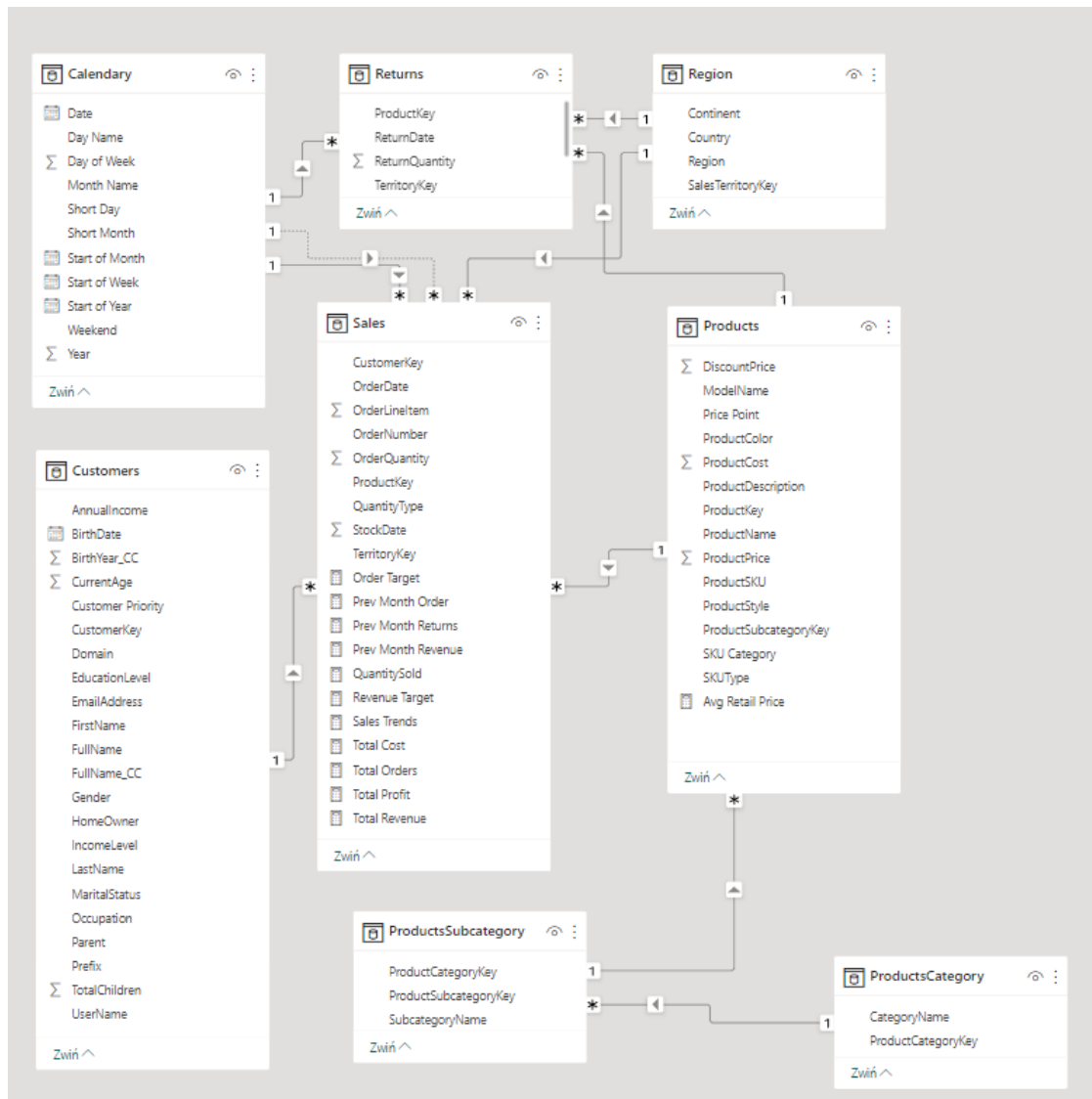
	ProductKey	ProductSubcategoryKey	ProductSKU	ProductName	ModelName	ProductDescription	ProductColor	ProductSize	ProductStyle	ProductCost	ProductPrice
351	1	BK-M82B-48	Mountain-100 Blac...	Mountain-100	Top-of-the-line c...	Black	48	U	1898.0944	3374.99	
348	1	BK-M82B-38	Mountain-100 Blac...	Mountain-100	Top-of-the-line c...	Black	38	U	1898.0944	3374.99	
349	1	BK-M82B-42	Mountain-100 Blac...	Mountain-100	Top-of-the-line c...	Black	42	U	1898.0944	3374.99	
350	1	BK-M82B-44	Mountain-100 Blac...	Mountain-100	Top-of-the-line c...	Black	44	U	1898.0944	3374.99	
347	1	BK-M82S-48	Mountain-100 Silv...	Mountain-100	Top-of-the-line c...	Silver	48	U	1912.1544	3399.99	
346	1	BK-M82S-44	Mountain-100 Silv...	Mountain-100	Top-of-the-line c...	Silver	44	U	1912.1544	3399.99	
344	1	BK-M82S-38	Mountain-100 Silv...	Mountain-100	Top-of-the-line c...	Silver	38	U	1912.1544	3399.99	
345	1	BK-M82S-42	Mountain-100 Silv...	Mountain-100	Top-of-the-line c...	Silver	42	U	1912.1544	3399.99	
310	2	BK-R93R-62	Road-150 Red, 62	Road-150	This bike is ridd...	Red	62	U	2171.2942	3578.27	
311	2	BK-R93R-44	Road-150 Red, 44	Road-150	This bike is ridd...	Red	44	U	2171.2942	3578.27	
313	2	BK-R93R-52	Road-150 Red, 52	Road-150	This bike is ridd...	Red	52	U	2171.2942	3578.27	
312	2	BK-R93R-48	Road-150 Red, 48	Road-150	This bike is ridd...	Red	48	U	2171.2942	3578.27	
314	2	BK-R93R-56	Road-150 Red, 56	Road-150	This bike is ridd...	Red	56	U	2171.2942	3578.27	

Rys. 4.11. Przetwarzanie danych poprzez filtrację bazy Adventure Works – Products

Agregacja w Apache Spark polega na przetwarzaniu danych w taki sposób, aby uzyskać określone statystyki lub podsumowania danych. Możliwe jest wykonywanie różnego rodzaju agregacji, takich jak: *count*, *sum*, *avg*, *min*, *max*, itd. Rysunek 4.11 przedstawia wykonaną filtrację na danych, które spełniają warunek, że cena danego produktu musi być większa niż 3000 dolarów. Następnie za pomocą *orderBy* są one sortowane od najtańszych do najdroższych. Wstępne przetworzenie danych było krokiem, który należało wykonać, ponieważ baza danych Adventure works jest ogromna. Dzięki pozbywaniu się z baz danych rekordów, które nie spełniały określonych warunków, baza danych zmniejszyła swoją objętość. Proces ten był wymagany w celu uzyskania informacji dotyczących konkretnego wydziału przedsiębiorstwa. Pozostałe bazy danych struktury Adventure Works zostały również agregowane i filtrowane w procesach przetwarzania danych w celu pozyskania niezbędnych, danych których analiza odbyła się w środowisku Power Bi. Dane z Apache Sparka do Hadoop, zostały przeniesione za pomocą metody *saveAsHadoopFile*, który pozwolił na zapis danych w formacie używanym przez Hadoop.

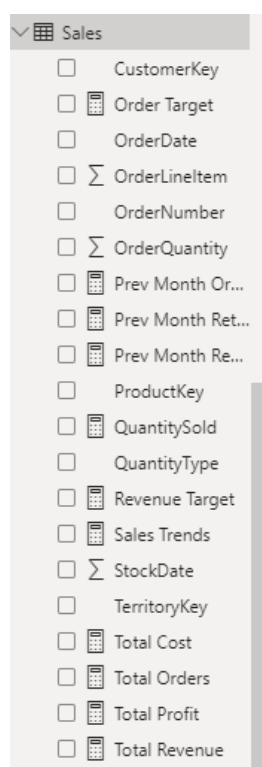
4.3.2 Analiza danych w Power Bi

Dane z Hadoop do Apache Druid zostały przeniesione za pomocą narzędzia Hadoop *fs*. Skopiował on dane do lokalnego systemu plików, a następnie użycie narzędzia *curl* posłużyło do przesłania danych do Apache Druida. Po przetworzeniu danych również w Apache Druid zostały one przeniesione do narzędzia do wizualizacji danych Power Bi. Import danych odbył się za pomocą plików CSV utworzonych w interfejsie API HTTP Apache Druid. Na ostatnim etapie dane były agregowane i wizualizowane w środowiskach raportowych. W celu tworzenia pulpitów nawigacyjnych najpierw musiały zostać stworzone relacje pomiędzy odpowiednimi tabelami. Rysunek 4.12 przedstawia bazy danych Adventure works, która została przeanalizowana w ekosystemie BI.



Rys. 4.12. Relacje pomiędzy przetworzonymi bazami danych Adventure works

W Power Bi stworzone zostały miary, czyli formuły, które posłużyły do obliczania wartości w modelu danych. Miary zdefiniowane zostały za pomocą języka DAX (ang.. *Data Analysis Expression*) i użyte do wykonywania obliczeń w locie, podczas interakcji z danymi w raportach. Zostały one utworzone w celu bardziej szczegółowej analizy danych w Power BI Desktop. Miary utworzone samodzielnie pojawiają się na liście Fields z ikoną kalkulatora, co można zauważyć na rysunku 4.13. Posłużyły one do odpowiedzenia na pytania biznesowe jak np. „Ile wynosiła całkowita sprzedaż produktów?”. Miary wykraczają poza podstawową agregację danych jednocześnie adaptują się do filtrowania, przez co są bardzo wydajne w eksploracji i analizie danych.



Rys. 4.13. Przykładowe utworzone miary

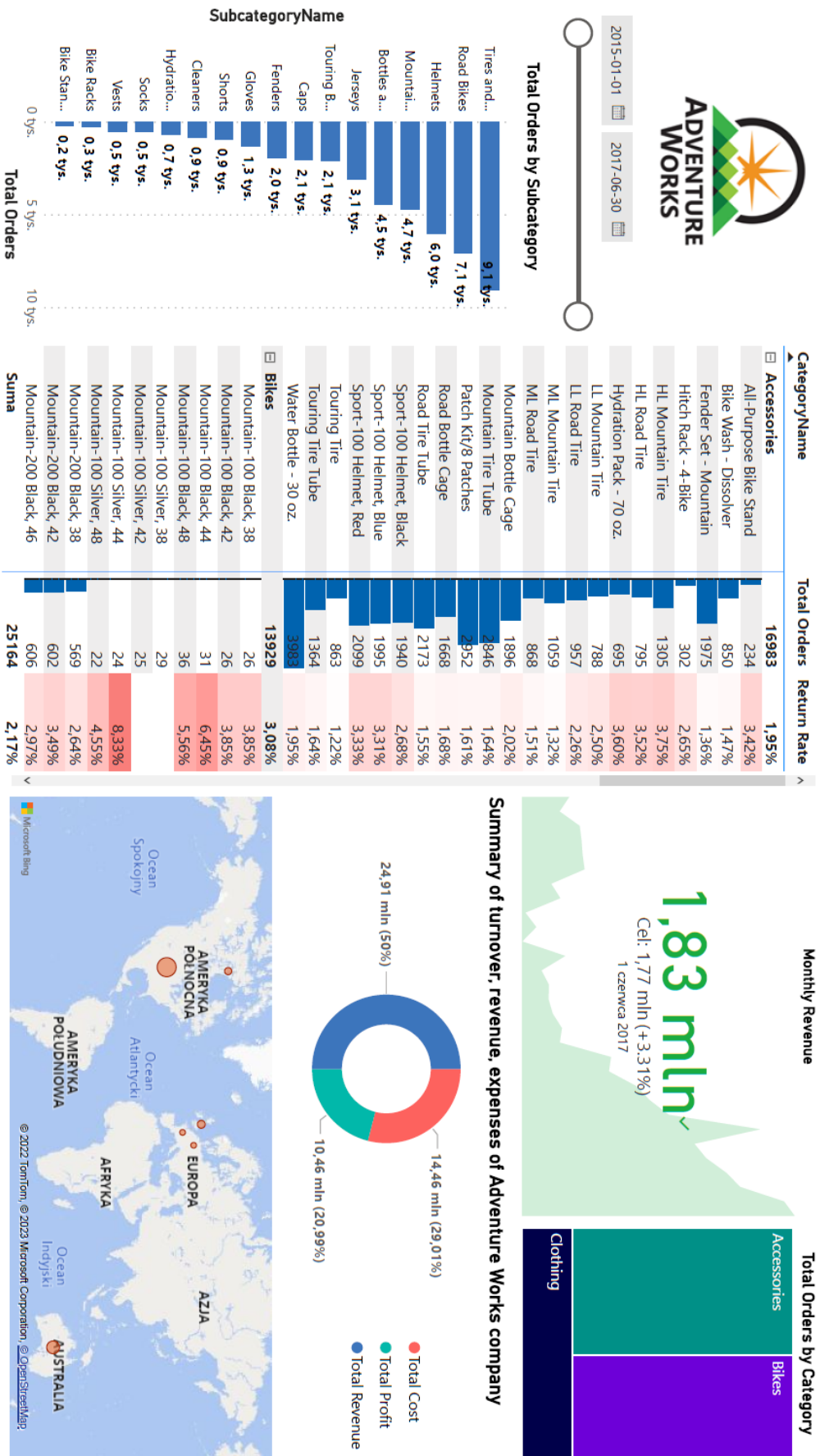
Rysunek 4.14 przedstawia kod niezbędny do utworzenia miar, które zostały później wykorzystane podczas tworzenia pulpitów nawigacyjnych. Miary te wraz z innymi były użyte do tworzenia zaawansowanych wizualizacji z uwzględnieniem ram czasowych.

```
Total Cost = sumX(Sales,Sales[OrderQuantity]*related(Products[ProductCost]))
Total Orders = DISTINCTCOUNT(Sales[OrderNumber])
Total Profit = [Total Revenue]-[Total Cost]
Total Revenue = sumX(Sales,Sales[OrderQuantity]*related(Products[ProductPrice]))
Prev Month Order = CALCULATE([Total Orders], DATEADD(Calendar[Date],-1,month))
Prev Month Returns = CALCULATE([Total Returns], DATEADD(Calendar[Date],-1,month))
Prev Month Revenue = CALCULATE([Total Revenue], DATEADD(Calendar[Date],-1,month))
QuantitySold = SUM(Sales[OrderQuantity])
Revenue Target = [Prev Month Revenue]*1.1
```

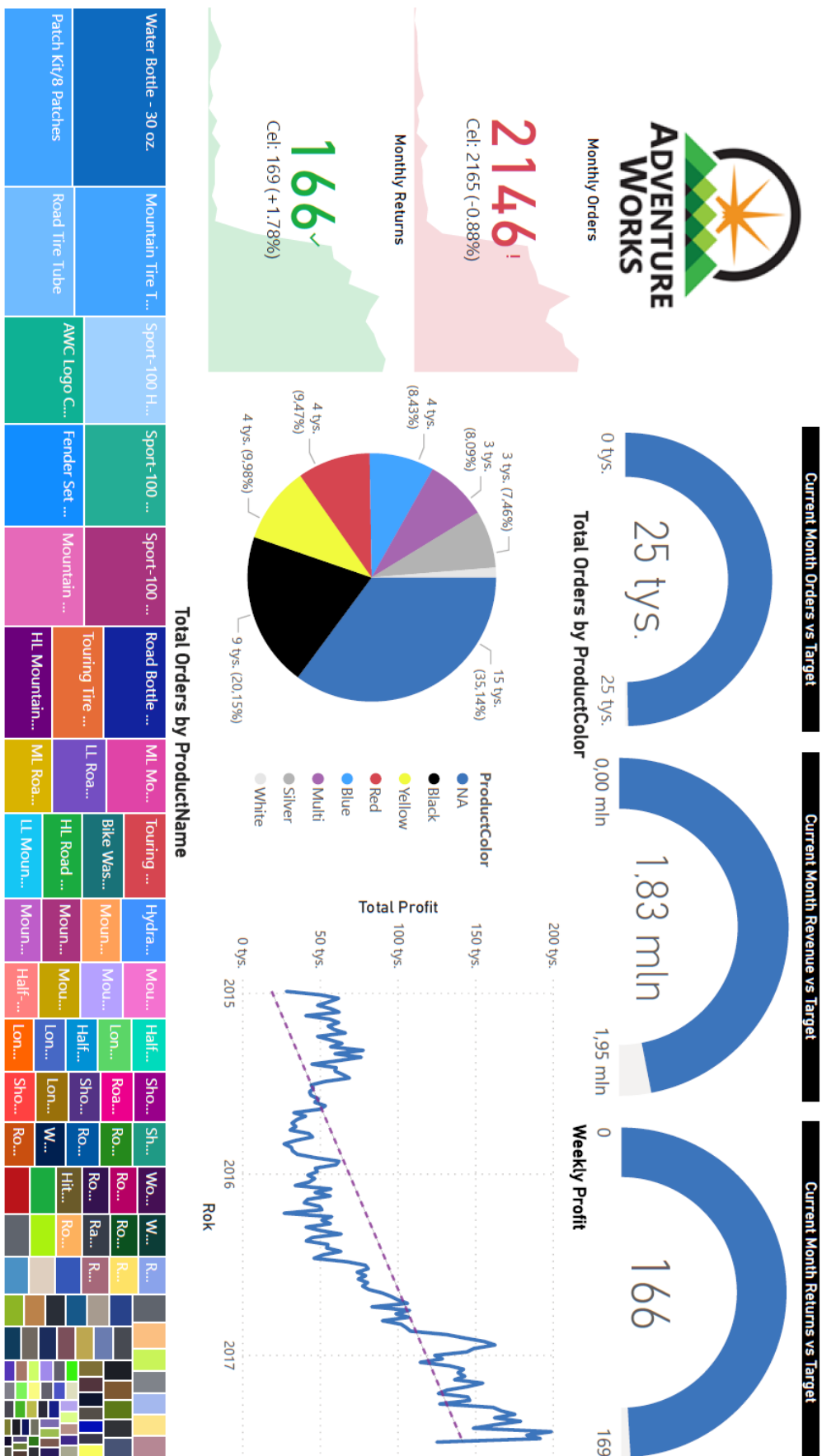
Rys. 4.14. Kod definiujący przykładowe miary

W programie Power BI została utworzona interaktywna aplikacja dla użytkowników oraz analityków biznesowych do analizy danych przedsiębiorstwa Adventure Works. Analizuje ona dane historyczne od 1 stycznia 2015 roku do 30 czerwca 2017 roku. Wykresy adaptują się w zależności do fragmentu czasu, do którego mają się odnosić. Ekosystem udostępnia 4 pulpity nawigacyjne, w których analityk biznesowy może analizować dane.

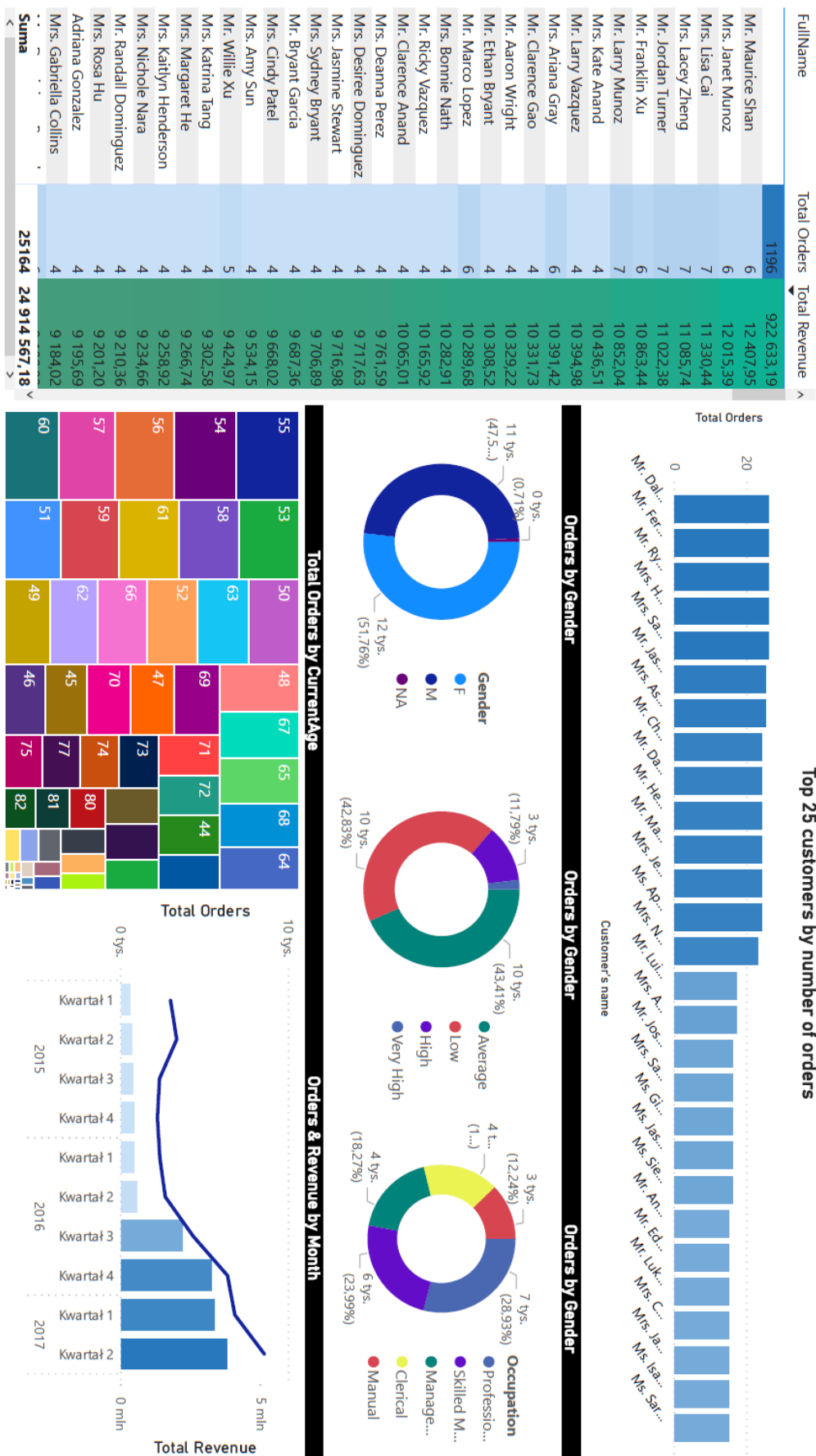
Rysunek 4.15 przedstawia pierwszy pulpit nawigacyjny interaktywnej aplikacji. Prezentuje on rozwój przedsiębiorstwa na przestrzeni czasu. Analityk może w nim analizować takie dane jak: ilość zamówień z podziałem na kategorie oraz podkategorie, ilość zwrotów zamówionych produktów, ilość sprzedanych produktów na różnych kontynentach oraz w konkretnych państwach. Osoba mająca dostęp do aplikacji może również sprawdzić obroty, koszty i zarobki przedsiębiorstwa na przestrzeni czasu. Drugi pulpit nawigacyjny został przedstawiony na rysunku 4.16. Wizualizacje te, w przeciwieństwie do pierwszego pulpitu nawigacyjnego, skupiają się na krótkich odstępach czasu, tygodnia lub miesiąca. Dodatkowo pokazały realizację określonych celów: zarobkowych, sprzedażowych oraz zmniejszania reklamacji produktów. Równie ważne wizualizacje to zysk na przestrzeni tygodni wraz z prognozą na przyszłość oraz najlepiej sprzedające się produkty ze względu na nazwę oraz kolor produktu. Rysunek 4.17 przedstawia trzeci pulpit nawigacyjny, który odnosi się do klientów przedsiębiorstwa Adventure Works. Ilustruje on raport zestawiający wszystkich klientów z ilością złożonych przez nich zamówień oraz wydanych pieniędzy podczas zakupu produktów przedsiębiorstwa. Prezentuje również dwudziestu pięciu najlepszych klientów pod względem zamówień. Wstępna analiza tych danych potwierdza, że w przedsiębiorstwie widoczna jest przedstawiona w części teoretycznej zasada Pareto. Jednocześnie została wykonana wizualizacja ilości zamówień ze względu na wiek, płeć oraz wykształcenie klientów. Ostatni pulpit nawigacyjny został przedstawiony na Rysunku 4.18. Przedstawia on najlepsze produkty w zależności od ilości sprzedaży i zysków w danych regionach oraz raport, który na podstawie analizy danych historycznych sprzedaży produktów ustala trend sprzedażowy na przyszłość poszczególnych produktów.



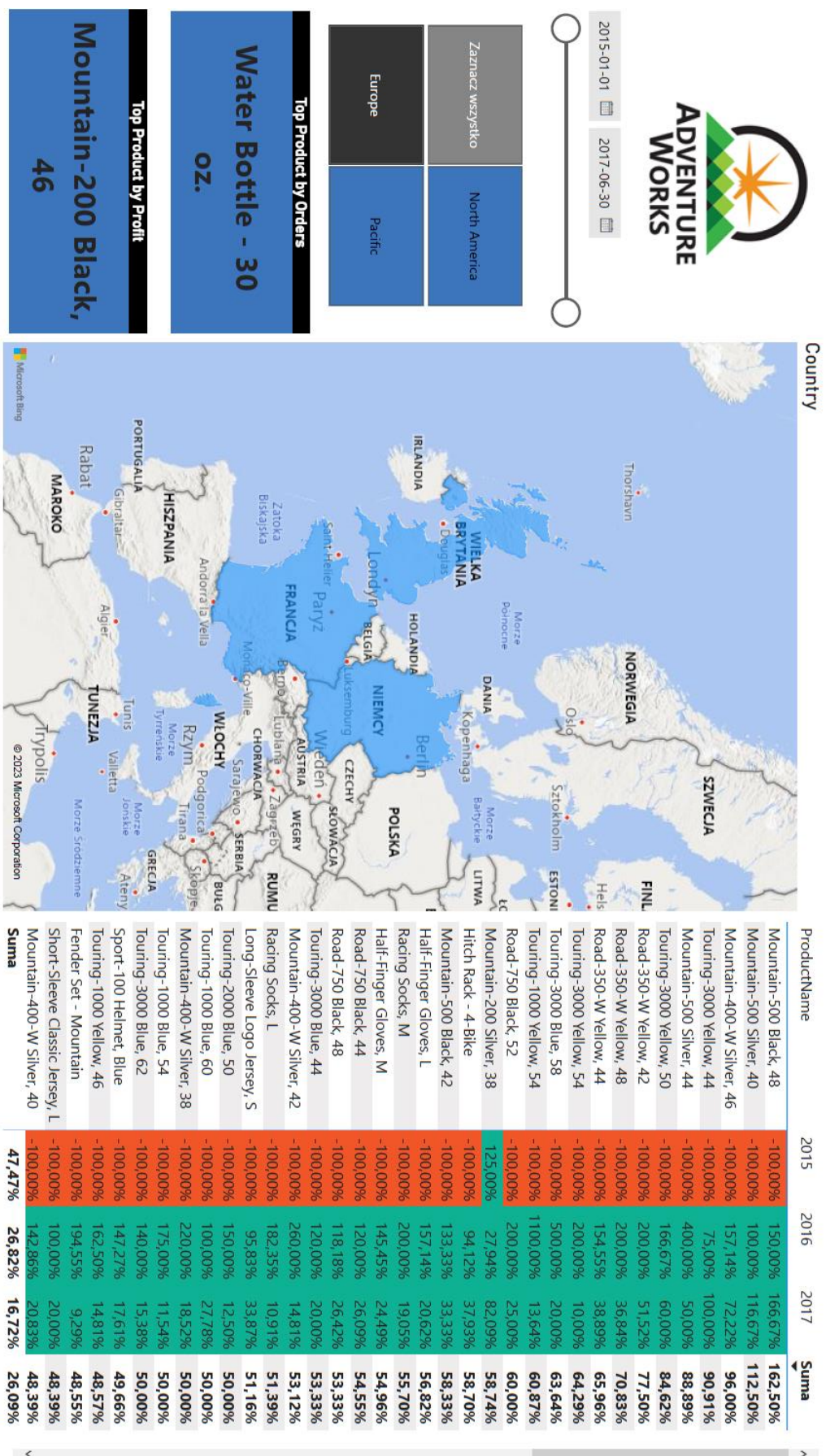
Rys. 4.15. Pierwszy pulpit nawigacyjny interaktywnej aplikacji – Adventure Works



Rys. 4.16 Drugi pulpit nawigacyjny interaktywnej aplikacji – Adventure Works



Rys. 4.17 Trzeci pulpit nawigacyjny interaktywnej aplikacji – Adventure Works



Rys. 4.18 Czwarty pulpit nawigacyjny interaktywnej aplikacji – Adventure Works

4.3.3 Podsumowanie

Cześć praktyczna pracy dyplomowej przedstawia budowę architektury ekosystemu Business Intelligence z wykorzystaniem Big Data. Praca zawiera również zaprezentowanie tematów zawartych w części teoretycznej na przykładzie rzeczywistego problemu. Prezentowana architektura z doбором technologii posiada swoje uzasadnienie biznesowe oraz studium przypadków wyjaśniających wybranie konkretnych elementów podczas budowy jej struktury. Architektura została oparta o środowisko Apache Spark do przetwarzania danych, Hadoop jako magazyn danych, Apache Druid jako środowisko do udostępniania danych oraz platformę Microsoft Power Bi jako narzędzie analityczne do wizualizacji danych. Zbudowany ekosystem został zaimplementowany oraz przetestowany dla przedsiębiorstwa Advenutre Works.

Zaprezentowano możliwości technologii Business Intelligence w środowisko Power Bi poprzez stworzenie interaktywnej aplikacji analitycznej posiadającej cztery pulpity nawigacyjne, w których użytkownik może analizować dane przedsiębiorstwa. Wykorzystane źródła danych nie są zbiorami Big Data, ale prezentują stworzoną koncepcję architektury oraz jej możliwości. Metody oraz rozwiązania technologiczne mogą zostać wykorzystane podczas analizy zbiorów danych stanowiących integralną część zbiorów Big Data dzięki skalowalności środowiska Apache Spark.

Stworzony projekt ekosystemu można rozbudować o utworzenie warstwy przetwarzania w czasie rzeczywistym by odpowiadała ona za analizę danych pochodzących bezpośrednio z czujników.

5. Wnioski

Część teoretyczna pracy zawiera opis przetwarzania danych, koncepcje systemów Big Data i jego narzędzia oraz zastosowanie Business Intelligence. W części tej została opisana szczegółowa budowa architektury systemów Big Data i Business Intelligence oraz analiza działania platform takich jak Apache Spark, Apache Superset, Tableau, Power Bi. Koncepcje te stanowią fundamentalną wiedzę niezbędną podczas projektowania architektury systemów przetwarzania danych oraz ekosystemów Business Intelligence.

Po szczegółowej analizie zagadnienia w części teoretycznej została utworzona część praktyczna, w której na przykładzie została zaprezentowana implementacja ekosystemu dla przedsiębiorstwa. Narzędziem, które posłużyło do masowego przetwarzania danych było Apache Spark. Proces przetwarzania danych nie został wykonany w całości ze względu na złożoność danych. W architekturze została przedstawiona ścieżka przetwarzania danych warstwy wsadowej odpowiadająca za analizę danych historycznych. Narzędzia oraz technologie potrzebne do stworzenia architektury ekosystemu zostały dobrane prawidłowo i spełniały postawione przed nimi zadanie. Pole do przemyśleń to wybór platformy do tworzenia warstwy wizualizacji oraz raportów. Środowisko Power Bi wykonało swój cel prawidłowo w pełnym zakresie oraz po dłuższej pracy w nim staje się bardzo przejrzysty i intuicyjne, przez co tworzenie nawet zaawansowanych pulpitów nawigacyjnych nie jest tak czasochłonne jak w innych środowiskach. Zdecydowanie mniej wartą uwagi jest platforma Tableau, która wymaga więcej czasu do zrozumienia jej działania oraz subiektywnie tworzy mniej intuicyjne raporty oraz potrzeba więcej czasu do stworzenia klarownych wizualizacji. Wartym rozważania zostaje narzędzie Apache Superset ze względu na swoją naturalną integralność z pozostałymi platformami Apache takim jak Apache Spark oraz Apache Druid. Oferuje on więcej od produktu Microsoft Power Bi natomiast również, jak w przypadku Tableau, wymaga zdecydowanie więcej czasu na zrozumienie funkcjonowania środowiska. Jednak w przeciwieństwie do Tableau, Apache Superset jest wart poświęconego czasu.

Założony cel pracy dyplomowej został w pełnym zakresie osiągnięty. Została stworzona architektura ekosystemu Business Intelligence z wykorzystaniem Big Data, która może zostać wykorzystana jako koncepcja architektury dla architektów oraz projektantów odpowiadających za ich budowę. Równocześnie może pomóc w wykonywaniu pracy analitykom biznesowym poprzez zrozumienie budowy ekosystemów, które wykorzystują podczas analizy.

Literatura

- [1] James Morle: *Scaling Oracle8i(TM): Building Highly Scalable OLTP System Architectures*, Addison-Wesley Professional, Maaloev, 2000
- [2] Oracle, <https://www.oracle.com/pl/database/what-is-oltp/> (dostęp 23.10.2022)
- [3] Michael Schrader ; Dan Vlamiš ; Mike Nader, *Oracle Essbase & Oracle OLAP*, McGraw-Hill Osborne Media, Nowy Jork, 2009
- [4] Ralph Kimball, *The Data Warehouse ETL Toolkit : Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*, Wiley Publishing, Nowy Jork, 2004
- [5] Oracle, <https://www.oracle.com/integration/what-is-etl/> (dostęp 28.10.2022)
- [6] Claudia Imhoff, Nicholas Gallemmo, Jonathan G. Geiger, *Mastering Data Warehouse Design: Relational and Dimensional Techniques*, Wiley, Hoboken, 2003
- [7] Oracle, <https://www.oracle.com/pl/database/what-is-a-data-warehouse/> (dostęp 03.11.2022)
- [8] Ramez Elmasri, Shamkant B. Navathe, *Fundamentals of Database Systems*, Pearson Education, London, 2003
- [9] Oracle, <https://www.oracle.com/big-data/> (dostęp 12.11.2022)
- [10] Motivaction, <https://www.motivaction.nl/en/news/blog> (dostęp 20.11.2022)
- [11] <https://learn.microsoft.com/pl-pl/azure/architecture/data-guide/big-data/> (dostęp 28.11.2022)
- [12] Tomcy John, Pankaj Misra, *Data Lake for Enterprises: Leveraging Lambda Architecture for building Enterprise Data Lake*, Packt Publishing, Birmingham, 2017
- [13] InterviewBit, <https://www.interviewbit.com/blog/apache-spark-architecture> (dostęp 08.12.2022)
- [14] Apache Foundation, *Spark Documentation*, <https://Spark.apache.org> (dostęp 13.12.2022)
- [15] Microsoft, <https://learn.microsoft.com/pl-pl/dotnet/Spark/what-is-Spark> (dostęp 15.12.2022)
- [16] Edureka!, <https://www.edureka.co/blog/Spark-architecture> (dostęp 16.12.2022)
- [17] Maheshwari, Anil, *Business Intelligence and data mining*, Business Expert Press, New Jersey, 2015
- [18] Howston C., *Successful Business Intelligence: Secrets to Make BI a Killer App*, 2008
- [19] Itransition, <https://www.itransition.com/business-intelligence/architecture>, (dostęp 18.12.2022)
- [20] Ryan Sleeper, *Tableau Desktop Pocket Reference*, O'Reilly Media, Sebastopol, 2021
- [21] Dan Murray, *Tableau Your Data!: Fast and Easy Visual Analysis with Tableau Software*, Wiley, Hoboken, 2014
- [22] Apache Foundation, *Superset Documentation*, <https://superset.apache.org> (dostęp 29.12.2022)
- [23] Cenius, <https://cenius.ai/blogs/apache-superset-review> (dostęp 30.12.2022)
- [24] Microsoft, <https://learn.microsoft.com/en-us/power-bi/> (dostęp 30.12.2022)
- [25] Adam Aspin , *Pro Power BI Desktop*, Apress, London, 2016
- [26] Apache Foundation, *Airflow Documentation*, <https://airflow.apache.org/docs/> (dostęp 01.01.2023)
- [27] Apache Foundation, *Ranger Documentation*, <https://ranger.apache.org> (dostęp 02.01.2023)
- [28] Apache Foundation, *Druid Documentation*, <https://Druid.apache.org> (dostęp 02.01.2023)
- [29] DataHub, <https://datahub.io/docs> (dostęp 03.01.2023)
- [30] Apache Foundation, *Hadoop Documentation* <https://hadoop.apache.org> (dostęp 03.01.2023)

Summary

The objectives of the thesis were to present ways to process data and present the application and characteristics of Big Data systems and Business Intelligence solutions. The thesis includes a detailed theoretical description and a practical part in which the described ecosystem architecture was created and implemented in the company Advenutre Works. The ecosystem architecture includes:

- Apache Spark as a mass data processing engine,
- Hadoop Distributed File System for data storage,
- Apache Druid for data sharing,
- Power Bi as a platform for creating visualizations and data reporting.
- An interactive application for users and business analysts to analyze Advenutre Works enterprise data has been created in Power BI.

The ecosystem provides 4 dashboards in which the business analyst can analyze:

- the number of orders by categories and subcategories,
- the number of returns of ordered products,
- the turnover, costs and earnings of the company over time,
- the amount of products sold in different continents and specific countries,
- realization of earnings targets, sales targets and reduction of product complaints,
- achievement of targets in specific months,
- weekly profit with a forecast for the future,
- type of color of products sold,
- customers in relation to the number of orders and the money they leave at the company when buying products,
- the influence of gender, age, etc. on the number of orders that a customer makes,
- the best products in relation to the amount of sales and profits they generate for the company,
- trends in product sales over the years.

The created ecosystem design can be expanded to create a real-time processing layer, which would be responsible for analyzing data coming directly from sensors. The architecture of the ecosystem was created with the concept that in the future it can be expanded to include the previously mentioned layer.

The stated goal of the thesis was fully achieved. An architecture of a Business Intelligence ecosystem using Big Data has been created, which can be used as a point of presentation of architecture concepts for architects and designers responsible for their construction.