Odejścia pracowników i ich czynniki

Spis treści

[1. Cel i opis projektu 1](#_Toc128773784)

[2. Kluczowe wymagania klienta 1](#_Toc128773785)

[3. Obecna architektura klienta 2](#_Toc128773786)

[4. Data dictionary (słownik danych) 2](#_Toc128773787)

[5. Decision logs (dzienniki decyzji) – uzgodnione z klientem 3](#_Toc128773788)

[6. Zaproponowana architektura (logiczna) 5](#_Toc128773789)

# Cel i opis projektu

Celem projektu jest wielowymiarowa analiza czynników i motywacji pracowników odchodzących z firmy klienta, jak również zestawienie tych danych z publicznie dostępnymi informacjami w tej tematyce w celu zrozumienia faktorów na to wpływających, jak również zmniejszeniu ilości odejść tychże pracowników. Jako pracowników odchodzących, klient kwalifikuje:

1. Pracowników, którzy odeszli z pracy przed końcem umowy
2. Pracowników, którzy odeszli z pracy po zakończeniu umowy (pomimo otrzymania propozycji przedłużenia)

Zgodnie z informacjami od klienta, analizie **nie będą** podlegać pracownicy, którzy zostali zwolnieni, jak również pracownicy którzy byli na stażu.

# Kluczowe wymagania klienta

Klient jest elastyczny pod względem metryk oraz wniosków, które będą wynikiem analizy, aczkolwiek MVP (Minimum Viable Product) jest:

1. Analiza odejść pracowników na podstawie wieku pracownika
2. Analiza odejść pracowników na podstawie departamentu pracownika
3. Analiza odejść pracowników na podstawie sumarycznego stażu pracy
4. Analiza odejść pracowników na podstawie stażu pracy u klienta
5. Analiza odejść pracowników na podstawie płci

Jako narzędzie do wizualizacji danych, klient wybrał narzędzie **Power BI**. Klient nie posiada subskrypcji do tego narzędzia, dlatego też ich koszt zostanie uwzględniony w części kosztowej

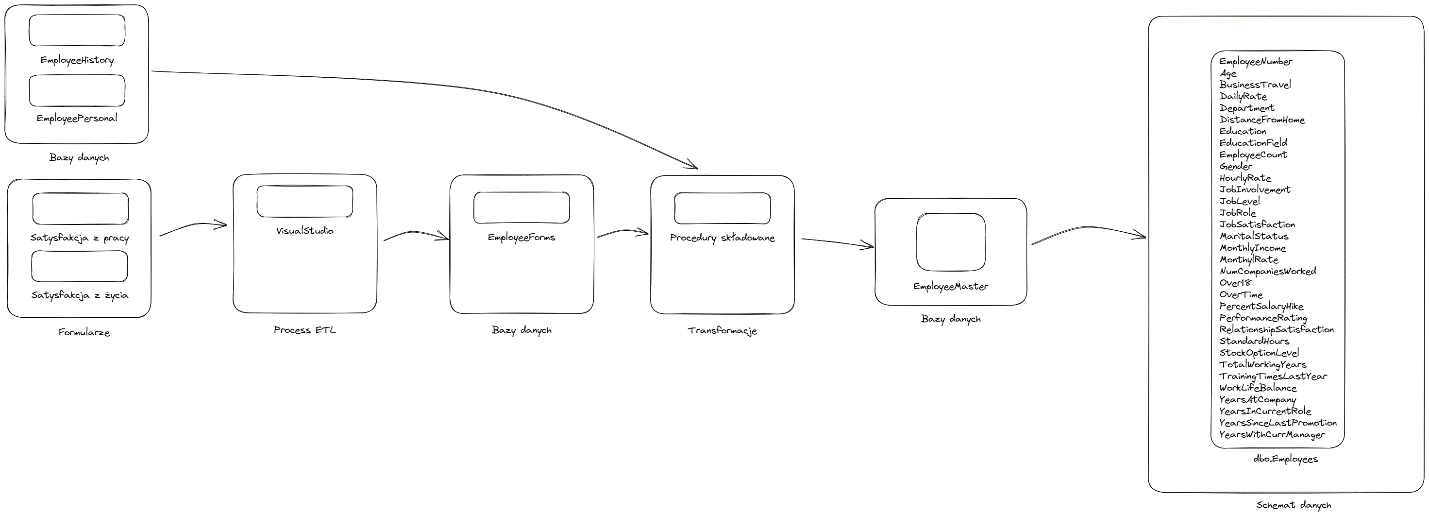
Do celów integracji danych, przepływu danych, transformacji, przechowywania danych oraz zarządzania danymi (m.in. Data Governance), klient zdecydował się na środowisko Azure. Klient otrzymał wstępne wyceny, poznał mocne i słabe strony rozwiązań chmurowych od różnych dostawców oraz rozwiązań on-premise, stąd też nie będą proponowane inne rozwiązania. Szczegółowe informacje znajdą swoje miejsce w decision logach w dalszej części dokumentu.

# Obecna architektura klienta

Klient posiada swoje dane w architekturze on-premise w środowisku bazodanowym SQL Server 2016. Klient korzysta z następujących baz danych:

* EmployeeHistory
* EmployeePersonal
* EmployeeMaster

Dane są pobierane z baz danych EmployeeHistory i EmployeePersonal oraz z formularzy: Satysfakcja z pracy oraz Satysfkacja z życia.



Rysunek 1 - Obecna architektura klienta (większa rozdzielczość w pliku obecna\_architektura.png)

Zgodnie z ustaleniami, klient nie chce udostępniać dostępu do baz danych EmployeePersonal oraz EmployeeHistory z uwagi na wysoce wrażliwe dane (np. numer PESEL, numer dowodu osobistego), dlatego też dane będą wyciągane bezpośrednio z bazy EmployeeMaster. Klient w przeszłości próbował używać tabeli dbo.Employees w rozwiązaniach Big Data, dlatego też tabela ma postać tabeli zdenormalizowanej.

# Data dictionary (słownik danych)

Na spotkaniu przedwdrożeniowym klient udostępnił słownik pojęć, który definiuje nazwę kolumny oraz opis kolumny.

|  |  |
| --- | --- |
| Nazwa kolumny | Opis |
| Age | Wiek |
| Attrition | Flaga czy pracownik jest zwolniony (0 - nie, 1 - tak) |
| BusinessTravel | Jak często pracownik podróżuje w ramach pracy |
| DailyRate | Stawka dzienna |
| Department | Oddział |
| DistanceFromHome | Odległość od domu do miejsca pracy |
| Education | Stopień edukacji (1-5) |
| EducationField | Ścieżka edukacji |
| EmployeeCount | Ilość pracowników, używane do sprawdzania duplikatów |
| EmployeeNumber | Numer pracownika |
| EnvironmentSatisfaction | Satysfakcja z miejsca pracy (1-5) |
| Gender | Płeć |
| HourlyRate | Stawka godzinowa |
| JobInvolvement | Zaangażowanie w pracę (1-5) |
| JobLevel | Poziom pracownika (1-5) |
| JobRole | Nazwa roli |
| JobSatisfaction | Satysfakcja z pracy (1-5) |
| MaritalStatus | Stan cywilny |
| MonthlyIncome | Miesięczny dochód (po uwzględnieniu podatków) |
| MonthlyRate | Stawka miesięczna (bez uwzględnienia podatków) |
| NumCompaniesWorked | Ilość miejsc pracy, w których pracownik pracował wcześniej |
| Over18 | Czy pracownik jest pełnoletni (0 - nie, 1 - tak) |
| OverTime | Czy pracownik bierze nadgodziny (No - nie, Yes - tak) |
| PercentSalaryHike | Procentowa podwyżka wynagrodzenia |
| PerformanceRating | Wydajność pracownika (1-5) |
| RelationshipSatisfaction | Satysfakcja w związku (1-5) |
| StandardHours | Średnia ilość godzin pracy |
| StockOptionLevel | Poziom posiadania akcji firmy (0-5) |
| TotalWorkingYears | Staż pracownika w latach |
| TrainingTimesLastYear | Ilość szkoleń w ostatnim roku |
| WorkLifeBalance | Poziom równowagi między pracą, a życiem prywatnym (1-5) |
| YearsAtCompany | Staż pracownika w obecnej firmie w latach |
| YearsInCurrentRole | Ilość lat na obecnym stanowisku |
| YearsSinceLastPromotion | Ilość lat od ostatniego awansu |
| YearsWithCurrManager | Ilość lat z obecnym menadżerem |

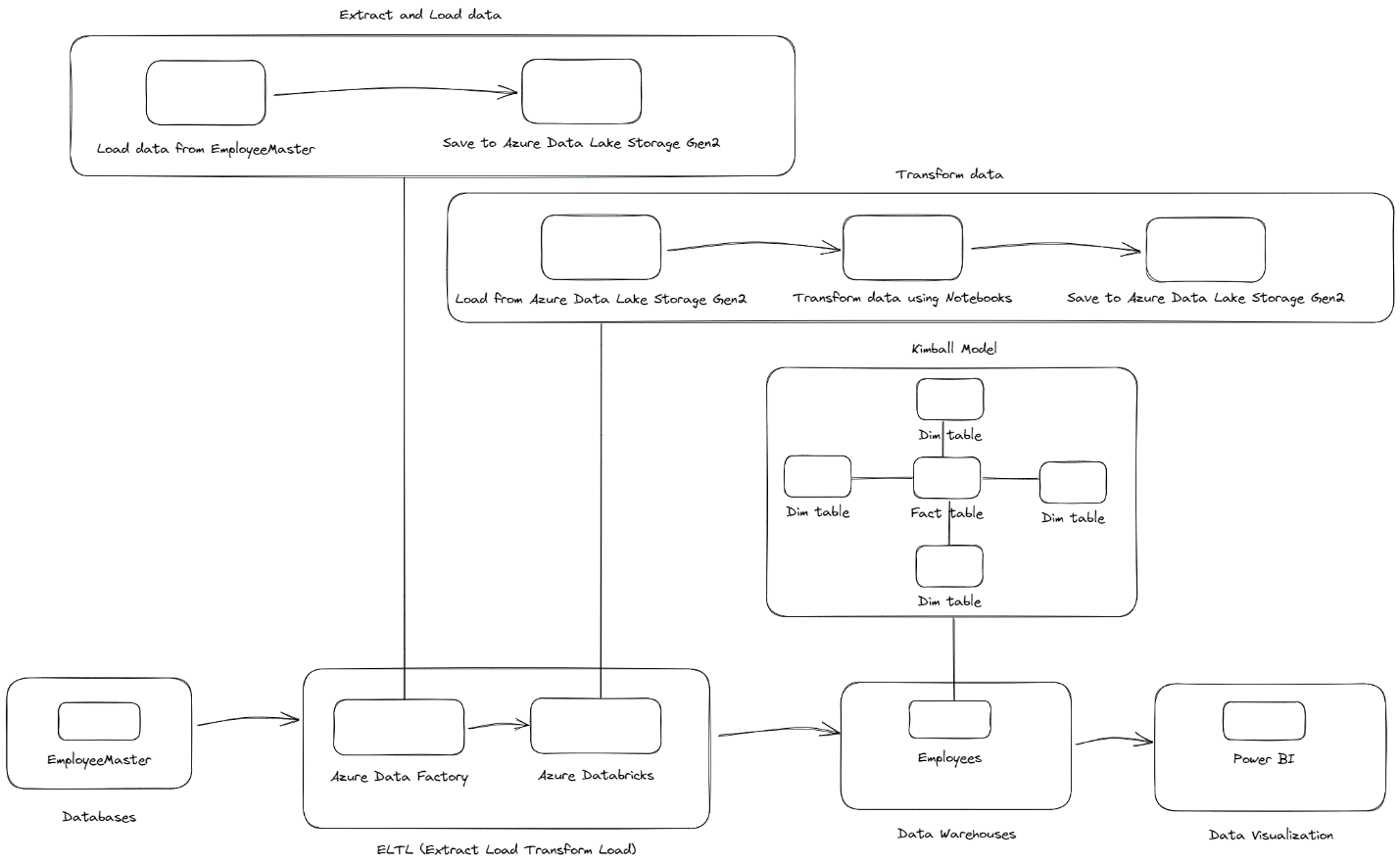
# Decision logs (dzienniki decyzji) – uzgodnione z klientem

Decision logs przedstawiają decyzje, które zostały przygotowane i opracowane przez Grzegorza Borkowskiego, a zostały podjęte przez osobę decyzyjną ze strony klienta (Krzysztofa Drużkowskiego).

|  |  |  |  |
| --- | --- | --- | --- |
| Nazwa | Opis | Proponowane opcje | Podjęta decyzja |
| Hurtownia danych | Zdefiniowanie lokalizacji hurtowni danych, w której będą przechowywane dane używane do wizualizacji | * On-premise (SQL Server) * Azure (Azure SQL Database) * Azure (Azure Synapse) | Azure (Azure SQL Database) |
| Narzędzie ETL | Narzędzie używane do integracji, przenoszenia danych pomiędzy środowiskiem klientem a hurtownią danych oraz do transformacji danych | * On-premise (Apache Airflow) * Azure (Azure Data Factory) + Azure Databricks * Azure (Azure Databricks) * Azure (Azure Synapse) | Azure (Azure Data Factory) + Azure Databricks |
| Wizualizacja danych | Narzędzie do wizualizacji danych | * On-premise (Power BI Report Server) * Microsoft Power BI * QlikView * Tableau | Microsoft Power BI |
| Pośrednia warstwa przechowywania danych | Pośrednia warstwa przechowywania danych jest używana jako tymczasowe miejsce, w którym dane będą przechowywane w celu zapewnienia wysoko wydajnego storage’a do dalszej transformacji danych | * On-premise (dyski NAS/NTFS) * Azure (Azure Data Lake Storage Gen2) * Azure (Azure File Shares) * Azure (Azure Blob Storage) | Azure (Azure Data Lake Storage Gen2) |
| Retencja danych | Ilość dni, po których dane mają zostać usunięte lub przeniesione do innej warstwy | N/A | Dane mają zostać:   * Przeniesione z warstwy hot do warstwy cold po 30 dniach * Przeniesione z warstwy cold do warstwy archive po 180 dniach * Usunięte po 300 dniach od daty utworzenia |
| Historia wersji danych | Historia wersji danych jest istotnym mechanizmem przy przechowywaniu danych. Dzięki niej możemy porównywać dane z danymi z przeszłości celem zestawienia oraz dalszej analizy, np. aby sprawdzić czy zmniejsza się ilość pracowników odchodzących z pracy) | N/A | Będziemy tworzyć snapshot danych codziennie, natomiast ostatniego dnia miesiąca będziemy zachowywać tylko dane z pierwszego dnia miesiąca (reszta będzie usuwana). Klient będzie chciał porównać maksymalnie 6 miesięcy |
| Data Governance | Data Governance jest bardzo szerokim obszarem, którego celem jest łatwiejsze zarządzanie danymi, klasyfikacja danych, zarządzanie uprawnieniami do danych, sprawdzanie jakości danych i inne. | * On-premise (IBM InfoSphere) * Azure (Azure Purview) * Azure (Databricks Unity Catalog) | Azure (Databricks Unity Catalog) |

# Zaproponowana logiczna architektura

Mając na uwadze posiadane dane przez klienta, jak również oczekiwania i wymagania, Grzegorz Borkowski przygotował i opracował architekturę danych. Architektura została omówiona z Krzysztofem Drużkowski i jego zespołem i została zaakceptowana.



Rysunek 2 - Zaproponowana logiczna architektura (większa rozdzielczość dostępna w pliku zaproponowana\_logiczna\_architektura.png)