

COMARCH IT AKADEMIA

Analiza danych dla analityków (212 godz.)

Realizowane projekty

W trakcie szkolenia uczestnicy będą realizowali trzy projekty praktyczne:

1. Projekt realizowany w trakcie trwania modułu B.

Celem projektu jest przygotowanie kompletnego modelu analitycznego obejmującego zdefiniowany obszar merytoryczny. Realizacja projektu będzie wymagała:

- Pozyskania danych z różnorodnych źródeł (bazy danych, dane publiczne w Internecie, pliki płaskie, arkusze Excel).
- Utworzenia zautomatyzowanych procedur pobierania i przekształcania danych.
- Zaprojektowania i realizacji tabelarycznego modelu analitycznego.
- Zdefiniowania w ramach modelu szeregu kalkulacji pozwalających na analizę w różnych perspektywach czasowych.

2. Projekt realizowany w trakcie trwania modułu C.

Celem projektu jest przygotowanie wizualizacji danych na modelu utworzonym w ramach pierwszego projektu. Realizacja projektu będzie wymagała:

- Doboru odpowiednich do celów wizualizacji danych.
- Utworzenia raportów prezentujących analizę danych oraz wnioski z niej płynące.
- Utworzenie interaktywnej prezentacji opowiadającej historię zapisaną w modelu (storytelling).

Do realizacji projektu uczestnik będzie mógł wykorzystać Power BI lub SQL Server Reporting Services.

3. Projekt końcowy.

Jest to projekt podsumowujący wiedzę nabytą w trakcie kursu. Punktem wyjścia będzie zestaw pytań, na które, przy pomocy utworzonego modelu oraz analizy wspartej uczeniem maszynowym, będzie musiał odpowiedzieć uczestnik kursu. Realizacja projektu będzie wymagała:

- Pozyskania danych adekwatnych do wymagań projektu.
- Utworzenia modelu analitycznego wraz z zautomatyzowanymi procedurami pobierania i transformacji danych.
- Przeprowadzenia eksploracyjnej analizy danych.
- Wykorzystania algorytmów uczenia maszynowego do budowy modelu, który będzie mógł być użyty w predykcji danych.

- Utworzenia raportów prezentujących wypracowane wnioski.
- Zaprezentowania całości efektów projektu.

Do realizacji projektu uczestnik będzie mógł wykorzystać dowolny zestaw narzędzi poznanych w trakcie trwania kursu.

Moduł A: Organizacja i przebieg procesu analizy danych (2 godz.)

Analiza danych powinna odbywać się w ramach dobrze zorganizowanego i prowadzonego procesu. Pozwala to zwiększyć efektywność, lepiej gospodarować posiadanymi zasobami oraz zmniejszyć ryzyko niepowodzenia.

Metodyka CRISP-DM (2 godziny)

Zrozumienie celów biznesowych
Zrozumienie danych
Przygotowanie danych
Budowa modelu
Ewaluacja
Wdrożenie

Moduł B: Pozyskiwanie danych. Przekształcanie danych. Tworzenie modeli analitycznych (98 godz.)

Celem modułu jest zapoznanie Uczestników z różnymi formami przechowywania danych, które są stosowane w procesach analitycznych. Uczestnicy pozyskają umiejętności pozwalające na dobór odpowiedniej techniki magazynowania i przetwarzania danych, zaprojektowanie, budowę oraz wykorzystanie magazynów danych we własnych rozwiązaniach analitycznych.

Formy magazynowania danych – wady, zalety, korzyści, potencjalne problemy (2 godz.)

Pliki płaskie
Arkusze kalkulacyjne
Dokumenty XML/JSON
Relacyjne bazy danych
Bazy NoSQL

Big Data
Chmura

Microsoft Excel – praca ze zbiorami danych (16 godz.)**Poprawność danych.**

- Narzędzie tabela
- Narzędzie sprawdzania poprawności nazw
- Analiza poprawności formuł i testowanie arkusza
- Konwersja typów liczbowych, dat, typów znakowych oraz stron kodowych
- Łączenie danych z wielu źródeł

Użycie danych pochodzących z różnych źródeł zewnętrznych

- Import danych z plików tekstowych
- Import danych z baz danych
- Import danych z plików XML – omówienie standardu XML
- Dane ze stron internetowych
- Eksportowanie danych

Narzędzia analizy

- Wykorzystanie narzędzi analizy warunkowej
- Używanie różnych typów adresowania
- Wykorzystanie formuł tablicowych
- Analiza warunkowa
- Agregacja danych
- Wybrane funkcje: finansowe, statystyczne, matematyczne, tekstowe

Kontrola danych

- Tworzenie różnego rodzaju sum
- Ustalanie rankingów
- Badanie realizacji celu
- Analiza bazodanowa
- Udziały procentowe
- Wykorzystanie formantów do zmiany parametrów prowadzonej analizy

Relacyjne bazy danych – wykorzystanie języka SQL – Microsoft SQL Server (32 godz.)**Wprowadzenie i przygotowanie środowiska pracy**

- Relacyjne bazy danych – podstawowe pojęcia
- Instalacja i konfiguracja serwera baz danych Microsoft SQL Server
- Narzędzia wykorzystywane w pracy z Microsoft SQL Server – Management Studio, Azure Data Studio

Projektowanie i budowa relacyjnych baz danych

Projektowanie modelu danych na potrzeby bazy relacyjnej

Dostępne typy danych

Tworzenie tabel

Reguły kontrolowania poprawności i integralności danych

SQL – język dostępu do danych

Konstrukcja zapytania w języku SQL

Pobieranie danych – instrukcja SELECT

Filtrowanie danych

Formatowanie wyniku zapytania

Sortowanie

Grupowanie wierszy oraz agregacja danych

Złączenia – pobieranie danych z wielu tabel

Podzapytania

Operacje na zbiorach

Wykorzystanie funkcji standardowych

Transakcje

Polecenia modyfikacji danych – INSERT, UPDATE, DELETE, MERGE

Tworzenie widoków

Wykorzystanie funkcji okienkowych – klauzula OVER

Tworzenie procedur składowanych oraz funkcji użytkownika

Wydajność zapytań – monitorowanie zapytań, definiowanie indeksów

Tworzenie indeksów kolumnowych

Podstawy administracji bazą danych SQL Server

Wydajność zapytań – monitorowanie zapytań, definiowanie indeksów

Tworzenie indeksów kolumnowych

Definiowanie praw dostępu do danych

Zabezpieczenie przed awarią – tworzenie i odtwarzanie kopii zapasowych

Automatyzacja działań z wykorzystaniem SQL Server Agent

Tworzenie wielowymiarowych modeli analitycznych (24 godz.)

Zasady modelowania wielowymiarowego

Analiza wymagań

Projektowanie modeli wielowymiarowych wg podejść Kimball'a oraz Inmon'a.

Wymiary oraz miary

Obsługa zmian w wymiarach (Slowly Changing Dimensions – SCD)

Realizacja analitycznych modeli wielowymiarowych w bazach relacyjnych

Tworzenie modeli tabelarycznych (Excel, Power BI)

Import danych ze źródeł zewnętrznych
Tworzenie relacji pomiędzy tabelami
Rodzaje relacji w modelach tabelarycznych (jeden do wielu, jeden do jeden, wiele do wielu)
Propagacja filtrów
Kolumny sortujące
Definiowanie tabeli dat
Dodatkowe właściwości obiektów w modelu – foldery wyświetlania, synonimy
Tworzenie hierarchii
Obsługa różnych poziomów szczegółowości tabel faktów w modelu

Kalkulacje w modelu tabelarycznym

Wprowadzenie do języka DAX
Rodzaje kalkulacji w modelu tabelarycznym: kolumny kalkulowane, miary, tabele kalkulowane
Zasady działania kontekstu wiersza i filtra
Przekształcanie kontekstu
Iteratory
Definiowanie kolumn kalkulowanych, funkcje RELATED, RELATEDTABLE
Modyfikacja kontekstu obliczeń – funkcja CALCULATE
Wykrywanie kontekstu obliczeń ustalonego w wizualizacjach
Kalkulacje operujące na okresach czasu

Rozszerzenia modelu tabelarycznego dostępne w Analysis Services

Partycjonowanie danych
Tłumaczenia metadanych
Usługa Analysis Services w Azure

Wdrożenie modelu tabelarycznego

Wdrożenie modelu przygotowanego w Excel – Office 365, migracja modelu do Analysis Services
Wdrożenie modelu przygotowanego w Power BI Service
Wdrożenie na serwerze Analysis Services
Wykorzystanie DAX jako języka zapytań

Bezpieczeństwo danych

Definiowanie ról zabezpieczeń
Testowanie ról
Wdrożenie zabezpieczeń w Power BI Service oraz Analysis Services

Wykorzystanie Power Query do ładowania i transformacji danych (8 godz.)**Wprowadzenie do Power Query**

Power Query w Excel, Power BI, Analysis Services
Język M
Edytor zapytań
Edytor zaawansowany

Źródła danych

Grupy źródeł danych
Definiowanie połączenia z serwerem SQL Server
Pliki płaskie
Arkusze Excel
Łączenie danych z wielu plików – narzędzie Combine
Pobieranie danych ze stron internetowych

Przekształcenia danych

Tworzenie kolumny z przykładów
Zmiana typu danych
Filtrowanie wierszy wg zawartości, wg pozycji, przy pomocy kolumny indeksowej
Modyfikacja zawartości kolumny (narzędzia Zastąp, Wyodrębnij itp.)
Kolumna niestandardowa
Kolumna warunkowa
Przestawienie kolumn

Zaawansowane funkcje Power Query

Parametryzacja zapytań Power Query
Tworzenie własnych funkcji
Wykorzystanie języków R/Python

Tworzenie procesów ładowania i transformacji danych z wykorzystaniem SQL Server Integration Services (16 godz.)**Wprowadzenie**

Integration Services – podstawowe pojęcia, architektura, komponenty
Narzędzia –SQL Server Data Tools, BIDS Helper
Tworzenie projektu

Tworzenie pakietu Integration Services

- Definiowanie połączeń (Connection Managers)
- Dodawanie i konfigurowanie elementów Control Flow
- Wykorzystanie kontroli przebiegu
- Wykonanie pakietu, automatyzacja uruchamiania pakietu
- Zachowanie pakietu jako pliku lub w repozytorium
- Narzędzia dtexec i dtexecui

Definiowanie przepływu danych - Data Flow

- Zasada działania
- Definiowanie źródeł danych oraz miejsc docelowych
- Wykorzystanie jako źródeł baz danych, plików tekstowych, plików Excel
- Wykorzystanie Execute SQL Task do wykonywania operacji na bazach danych

Najczęściej używane transformacje

- Dodawanie kolumn wyliczanych
- Agregacja danych
- Sortowanie danych
- Łączenie danych z różnych źródeł (Lookup)
- Konwersja typów danych
- Łączenie i rozdzielanie przebiegów danych (Union All oraz Conditional Split)

Wzbogacanie pakietów Integration Services

- Obsługa błędów w przepływie danych
- Użycie Send Mail Task do wysyłania wiadomości pocztowych
- Kontener sekwencyjny, kontenery pętli
- Wykorzystanie zmiennych oraz parametrów
- Użycie wyrażeń do dynamicznego ustawiania wartości właściwości elementów pakietu
- Konfiguracja pakietów

Moduł C: Wizualizacja danych, raportowanie, udostępnianie analiz (80 godz.)**Analiza danych z wykorzystaniem arkusza kalkulacyjnego Excel (16 godz.)****Narzędzia analizy**

- Wykorzystanie analizy warunkowej
- Używanie różnych typów adresowania
- Wykorzystanie formuł tablicowych

Analiza warunkowa
Agregacja danych
Wybrane funkcje: finansowe, statystyczne, matematyczne, tekstowe

Kontrola danych

Tworzenie różnego rodzaju sum
Ustalanie rankingów
Badanie realizacji celu
Analiza bazodanowa
Udziały procentowe
Wykorzystanie formantów do zmiany parametrów prowadzonej analizy

Prognozowanie na podstawie danych

Prognozy w oparciu o linie trendu
Prognozy z wykorzystaniem modeli regresji
Prognozy sezonowe
Konsolidacja danych z wielu arkuszy
Konsolidacja danych za pomocą tabeli przestawnej
Konsolidacja wielu źródeł danych (w tym zewnętrznych)

Praca z tabelami i wykresami przestawnymi

Automatyczne obliczanie odpowiednich udziałów procentowych
Obliczanie wartości skumulowanych
Szukanie zmian wartości
Grupowanie danych
Tworzenie wewnętrznych formuł obliczeniowych
Wykres przestawny

Wybór rodzaju wizualizacji

Wybór odpowiedniej formy zestawień
Odpowiedni dobór wykresu
Nakładanie wielu wykresów na jedną oś
Prezentacja danych na wykresie o dużej rozpiętości skali
Formatowanie warunkowe
Tworzenie wykresów niestandardowych

Inne narzędzia prezentacji danych

Sparklines - wykresy przebiegu w czasie
Wykresy na mapach
Mapy jako element wizualizacji

Infografiki

Pulpit menedżerski

- Omówienie koncepcji pulpitu menedżerskiego
- Zasady tworzenia pulpitu menedżerskiego
- Realizacja z wykorzystaniem MS Excel

Publikacja danych

- Publikacja danych w różnych formatach zewnętrznych
- Publikacja danych i wykresów w PowerPoint i Word
- Aktualizacja danych w ramach pakietu Office
- Publikacja danych do formatu PDF
- Publikacja danych do formatów HTML (webowych)
- Publikacja danych na serwerach SQL

Analiza i wizualizacja danych z wykorzystaniem Power BI (16 godz.)

Wprowadzenie

- Komponenty platformy Power BI
- Poziomy licencjonowania
- Najczęstsze scenariusze wykorzystania Platformy Power BI

Wizualizacja danych

- Tworzenie raportu w Power BI desktop
- Zasada działania interakcji wizualizacji w raporcie – filtrowanie danych, wyróżnianie danych
- Filtry w wizualizacjach i raporcie
- Właściwości wizualizacji
- Funkcje analityczne

Tworzenie interaktywnej prezentacji danych

- Definiowanie zakładek
- Ukrywanie i pokazywanie wizualizacji w zależności od kontekstu
- Wizualizacje z dodatkowymi opcjami interaktywnymi

Publikacja i udostępnianie raportów

- Definiowanie obszarów roboczych w usłudze Power BI
- Tworzenie pulpitu nawigacyjnych
- Wykorzystanie zawartości Power BI na urządzeniach mobilnych

Tworzenie i udostępnianie aplikacji

Język R w analizie danych (24 godz.)

Wprowadzenie

Instalacja i konfiguracja niezbędnych narzędzi: dystrybucja języka R, RStudio
Dystrybucja Anaconda, notatniki Jupyter
Repozytoria CRAN, MRAN
Wyszukiwanie oraz instalacja bibliotek dodatkowych

Podstawowy język R, typy danych

Praca z konsolą i edytorem w RStudio
Praca z wykorzystaniem notatników Jupyter
Operacje na plikach
Wektory, macierze, tablice, listy
Faktor
Działania na danych – wektoryzacja
Podstawowe konstrukcje języka R

Pobieranie i przekształcanie danych

Pobieranie danych z plików płaskich oraz Excel
Pobieranie danych z baz relacyjnych
Pobieranie danych z internetu
Praca z ramkami danych - data.frame
Przekształcanie danych z wykorzystaniem biblioteki dplyr

Wizualizacja danych

Wizualizacja z wykorzystaniem funkcji wbudowanych w język R
Zasada pracy z biblioteką ggplot2
Wizualizacja zmiennych ciągłych i kategoriowych
Wykresy jednej zmiennej, dwóch oraz wielu zmiennych
Dodatkowe elementy na wykresie
Wykorzystanie języka R do wizualizacji danych w Excel oraz Power BI

Publikacja danych

Tworzenie publikacji oraz prezentacji wyników analiz z wykorzystaniem biblioteki knitr
Budowa interaktywnych stron internetowych z użyciem pakietu Shiny

Analiza i wizualizacja danych w Tableau (8 godz.)**Wprowadzenie**

- Komponenty platformy Tableau
- Tableau Public
- Zasady licencjonowania

Pobieranie danych

- Tworzenie połączeń
- Współdzielenie połączeń
- Zmiany w danych źródłowych

Tworzenie prostych wizualizacji

- Elementy wizualizacji
- Formatowanie widoku
- Filtrowanie i sortowanie danych
- Używanie grup, hierarchii, zestawów i pól złożonych

Wykorzystanie wymiaru czasu

- Przetwarzanie dat
- Tworzenie filtrów opartych na datach

Wizualizacja danych przy pomocy wykresów

- Prezentacja wielu miar na wykresie
- Wykresy złożone
- Typy wykresów i ich zastosowania
- Mapy temperaturowe
- Wykresy kołowe, tree maps
- Wykorzystanie map

Tworzenie kalkulacji

- Typy kalkulacji
- Tworzenie pól obliczeniowych
- Użycie instrukcji warunkowych
- Konwersje typów danych
- Operowanie na datach i czasie
- Agregacje danych

Interaktywność

- Parametry widoku
- Tworzenie kokpitów analitycznych
- Akcje i drażnienie danych

Współdzielenie prac

- Eksport widoku
- Eksport danych
- Inne opcje współdzielenia

Raportowanie z wykorzystaniem SQL Server Reporting Services (16 godz.)**Wprowadzenie**

- Cykl życia raportu
- Architektura i podstawowe komponenty Reporting Services
- Narzędzia wykorzystywane do budowy raportów
- SQL Server Data Tools
- Report Builder
- Tworzenie i zarządzanie projektem w SSDT
- Publikacja raportu na serwerze Reporting Services

Definiowanie źródła danych raportu

- Określanie uprawnień dostępu do źródła danych
- Definiowanie zestawu danych. Budowa zapytania
- Korzystanie z baz relacyjnych (SQL) i wielowymiarowych (OLAP – Analysis Services)

Budowa raportu

- Prezentacja danych przy pomocy elementu tablix
- Formatowanie tabeli
- Właściwości pól tekstowych
- Sortowanie, filtrowanie i grupowanie danych w tabelach. Dodawanie podsumowań
- Definiowanie wyrażeń w raportach. Formatowanie warunkowe
- Wizualizacja danych przy pomocy wykresów i wskaźników
- Zagnieżdżanie obiektów
- Wykorzystanie elementu Rectangle do grupowania obiektów
- Wykorzystanie map

Właściwości i formatowanie raportu

- Podziały stron

Parametryzacja raportu
Parametryzacja zestawów danych

Opcje interaktywne w raportach

Mapa dokumentu
Zakładki oraz akcje
Interaktywne sortowanie
Przełączanie widoczności elementów

Zarządzanie raportami

Automatyzacja generowania i dystrybucji raportów przy pomocy subskrypcji
Definiowanie praw dostępu do raportów
Tworzenie, zarządzanie i wykorzystanie Web Parts do szybkiej budowy raportów

Moduł D: Analiza zaawansowana – uczenie maszynowe (24 godz.)

Podstawy uczenia maszynowego (4 godz.)

Dobór zmiennych, redukcja zmiennych
Przekształcanie danych wejściowych na potrzeby modeli uczenia maszynowego
Podział danych na treningowe, testowe
Techniki walidacji modeli

Tworzenie i wykorzystanie modeli uczenia maszynowego w języku R (16 godz.)

Klasyfikacja – k-NN, drzewa decyzyjne, sieci neuronowe, maszyna wektorów nośnych
Regresja
Grupowanie
Asocjacja

Funkcje uczenia maszynowego w Power BI (4 godz.)

Wykorzystanie modeli uczenia maszynowego tworzonych w języku R
Wizualizacje Power BI z funkcjami uczenia maszynowego (analiza kluczowych czynników wpływu, grupowanie, prognozowanie)

Moduł E: Konsultacje z doradcą HR (4 godz.)

Konsultacje z doradcą HR (4 godz.)

Statystyka rynku pracy w zawodzie analityka
Główne zasady rekrutacji pracowników
Zasady rekrutacji w Comarch SA
Klucz do dobrego CV
Przygotowanie profilu zawodowego na znanych portalach (Linkedin, Goldenline)
Rozmowa rekrutacyjna
 Przygotowanie do rozmowy
 Zasady prowadzenia rozmowy
 Pytania od i do rekrutera
 Stres – geneza i antidotum
Pytania i odpowiedzi

Umiejętności

Po ukończeniu kursu uczestnik/czka będzie potrafił/a:

- Zastosować metodyki analizy danych.
- Przygotować dane do analizy za pomocą różnych narzędzi (Excel, Power BI, SQL Server).
- Wykorzystać język R podczas przygotowania danych do analizy.
- Wykonać analizę za pomocą narzędzi Excel, PowerPivot, PowerBI, Analysis Services.
- Wykonać wizualizację danych za pomocą Excel, Tableau, PowerPivot, Power BI i SQL Server (Reporting Services).
- Zastosować język R w analizie danych.
- Zaprojektować i wdrożyć procesy analityczne w przedsiębiorstwie.
- Przeprowadzić ewaluację uzyskanych wyników analizy.

Bibliografia

Alberto Ferrari, Marco Russo: Kompletny przewodnik po DAX. Analiza biznesowa przy użyciu Microsoft Excel, SQL Server Analysis Services i Power BI.

Wydawnictwo Promise, 2016.

Michael Alexander, Jared Decker, Bernard Wehbe: Analizy Business Intelligence. Zaawansowane wykorzystanie Excela.

Wydawnictwo Helion.

Marcin Szeliga: Data Science i uczenie maszynowe.

Wydawnictwo Naukowe PWN.

Danuta Mendrala, Marcin Szeliga: Praktyczny kurs SQL. Wydanie III
Wydawnictwo Helion.

Katarzyna Żmuda: SQL. Jak osiągnąć mistrzostwo w konstruowaniu zapytań.
Wydawnictwo Helion.

Stacia Misner: Microsoft SQL Server 2012 Reporting Services. (2 tomy).
Wydawnictwo Promise.

Itzik Ben-Gan: Podstawy języka T-SQL Microsoft SQL Server 2016 i Azure SQL Database.
Wydawnictwo Promise.

Alberto Ferrari, Marco Russo: Microsoft SQL Server 2016 Analysis Services: Modelowanie tabelaryczne.
Wydawnictwo Promise.

Wee Hyong Tok i inni: Microsoft SQL Server 2012. Integration Services.
Wydawnictwo Promise.

Ralph Kimball, Margy Ross: The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling,
3rd Edition.
Wydawnictwo Wiley.