Notatka – narzędzia analityczne, Big Data i AI w chmurze (Azure, AWS, GCP)

Współczesna analiza danych i rozwój AI opierają się na usługach chmurowych, które zapewniają skalowalność, elastyczność i łatwe zarządzanie zasobami obliczeniowymi. Trzy główne platformy chmurowe – Microsoft Azure, Amazon Web Services (AWS) i Google Cloud Platform (GCP) – oferują bogate zestawy narzędzi wspierających cały cykl życia danych: od ich zbierania, przez przetwarzanie, analizę, aż po wdrażanie modeli uczenia maszynowego.

Do najważniejszych narzędzi i usług należą:

- Przetwarzanie danych (ETL/ELT): Azure Data Factory, AWS Glue, Google Cloud Dataflow
- Przechowywanie danych: Azure Data Lake, Amazon S3, Google Cloud Storage
- Bazy danych: Azure Synapse Analytics, Amazon Redshift, BigQuery
- Uczenie maszynowe: Azure Machine Learning, Amazon SageMaker, Vertex AI
- Wizualizacja i raportowanie: Power BI, Amazon QuickSight, Looker
- Zarządzanie pipeline'ami i orkiestracja: Azure Data Factory, AWS Step Functions, Cloud Composer
- Zbieranie danych z urządzeń IoT: Azure IoT Hub, AWS IoT Core, Cloud IoT Core (GCP)
- Bezpieczeństwo i monitorowanie: Azure Monitor, AWS CloudWatch, Cloud Logging

W każdej chmurze możliwe jest przygotowanie pełnej architektury od surowych danych po gotowe insighty i modele predykcyjne wdrażane w środowisku produkcyjnym.

Przykład 1 – Wykrywanie nieprawidłowości w transakcjach kartą (AWS)

- Dane zbierane są z systemów płatniczych w czasie rzeczywistym i trafiają do Amazon Kinesis.
- Przetwarzanie odbywa się w Amazon Glue, gdzie dane są czyszczone i przygotowywane.
- Dane są przechowywane w Amazon S3 jako warstwa Data Lake.
- W Amazon SageMaker trenowany jest model klasyfikujący anomalie.
- Model jest wdrażany jako endpoint w SageMaker i zintegrowany z aplikacją analizującą transakcje w czasie rzeczywistym.
- Wyniki są przesyłane do Amazon QuickSight do dalszej analizy i wizualizacji.

Przykład 2 – Architektura analityczna dla firmy streamingowej (GCP)

- Dane o użytkownikach, ich aktywnościach, lokalizacji i czasie spędzonym w aplikacji trafiają do BigQuery jako centralnego magazynu danych.
- GCP Dataflow służy do przetwarzania danych strumieniowych i batchowych.
- Na danych działa Vertex AI, który tworzy modele przewidujące churn lub rekomendacje treści.
- Trening modeli oparty jest o dane historyczne z BigQuery.
- Wyniki są publikowane do Looker Studio i dostępne dla zespołów produktowych i marketingowych.

Przykład 3 – Monitorowanie zużycia energii w smart city (Azure)

- Czujniki IoT wysyłają dane do Azure IoT Hub.
- Dane trafiają do Azure Stream Analytics, gdzie są przetwarzane w czasie rzeczywistym.
- Następnie zapisywane są w Azure Data Lake i agregowane w Azure Synapse Analytics.
- W Azure Machine Learning trenowany jest model, który przewiduje szczytowe zużycie energii i potencjalne awarie.
- Dashboard w Power BI umożliwia władzom miejskim monitorowanie wskaźników w czasie rzeczywistym.
- Całość jest zabezpieczona za pomocą Azure Monitor i Azure Security Center.