Czym jest Big Data?

Big Data odnosi się do ogromnych, złożonych zbiorów danych, które mogą być analizowane w celu uzyskania wartościowych informacji wspierających proces podejmowania decyzji. Ich rozmiar i różnorodność przekraczają możliwości tradycyjnych narzędzi bazodanowych, jeśli chodzi o ich pozyskiwanie, przechowywanie, zarządzanie i analizę.

Big Data opiera się na tzw. **trzech filarach 3V**:

- **Objętość (volume)** – ilość generowanych danych jest olbrzymia.
- **Prędkość (velocity)** – dane napływają w szybkim tempie, często w czasie rzeczywistym.
- **Różnorodność (variety)** – dane pochodzą z różnych źródeł i mają różne formaty (np. tekst, obrazy, nagrania).

Kluczowe technologie i narzędzia Big Data

- **NoSQL** – nierelacyjne bazy danych, które nie wymagają sztywnego schematu danych.
- **Data Lake** – scentralizowane repozytorium pozwalające przechowywać dane w ich pierwotnej formie (strukturalne, półstrukturalne i niestrukturalne).
- **Apache Hadoop** – otwarta platforma do rozproszonego przetwarzania dużych zbiorów danych w klastrach.
- **Apache Hive** – narzędzie do analizy danych oparte na Hadoop, umożliwia zapytania SQL.
- **Apache Spark** – platforma do szybkiego przetwarzania danych w rozproszonej architekturze.
- **Azure Synapse Analytics** – usługa analityczna integrująca dane hurtowniane i Big Data.
- **Power BI** – narzędzie do budowy interaktywnych pulpitów (dashboardów) dla użytkowników biznesowych.

Przykładowa architektura Big Data – Rockstar Games

1. Pozyskiwanie danych (Data Ingestion)

Źródła danych to zarówno strumieniowe, jak i wsadowe:

- Logi z gier (np. błędy, zawieszenia, spadki FPS)
- Telemetria graczy (czas rozgrywki, interakcje)
- Media społecznościowe i zgłoszenia do supportu
- Dane głosowe i wideo – opcjonalnie do analizy emocji

Wykorzystywane narzędzia:

- *Azure Event Hubs* – do odbierania danych w czasie rzeczywistym
- *Azure Data Factory* – do zbierania danych w partiach (np. kopie zapasowe)
- *Azure IoT Hub* – jeśli urządzenia gracza zbierają dane czujnikowe

2. Składowanie danych (Data Storage)

Dane są przetwarzane warstwowo:

- **Warstwa surowa (Raw)** – dane w niezmienionej formie
- **Warstwa oczyszczona (Cleansed)** – dane sformatowane (np. Parquet)
- **Warstwa finalna (Curated)** – dane gotowe do analizy i modelowania

Narzędzia:

- *Azure Data Lake Storage Gen2* – centralny magazyn danych
- *Azure Synapse Analytics* – analiza i przetwarzanie danych

3. Przetwarzanie danych (Compute Layer)

Obejmuje zarówno przetwarzanie strumieniowe, jak i wsadowe:

- *Azure Stream Analytics* – analiza danych w czasie rzeczywistym, np. alerty
- *Apache Spark w Azure Synapse / Databricks* – zaawansowane przetwarzanie i analiza
- *Azure Functions* – lekkie akcje wywoływane przez zdarzenia (np. wysyłanie powiadomień)

4. Sztuczna inteligencja i uczenie maszynowe (ML/AI)

Zastosowania:

- Predykcja awarii i błędów na podstawie logów
- Analiza nastrojów graczy (komentarze, Discord, zgłoszenia)
- Przewidywanie odejść graczy (churn prediction)
- Wykrywanie anomalii i nietypowych zachowań

Technologie:

- *Azure Machine Learning* – zarządzanie eksperymentami i modelami
- *Cognitive Services* – przetwarzanie języka naturalnego, analiza emocji
- *AutoML, MLflow* – szybkie prototypowanie modeli
- *ONNX, AKS (Azure Kubernetes Service)* – wdrażanie modeli jako API

5. Wizualizacja i analizy biznesowe

Narzędzia:

- *Power BI + Azure Synapse* – interaktywne dashboardy dla zespołów
- *Azure Monitor + Application Insights* – narzędzia dla zespołów IT i testerów

6. Zarządzanie i bezpieczeństwo

Rozwiązania:

- *Azure Purview* – zarządzanie danymi, śledzenie ich pochodzenia
- *Azure Key Vault* – bezpieczne przechowywanie kluczy i haseł
- *RBAC* i *Managed Identity* – kontrola dostępu do danych
- *CI/CD z GitHub Actions lub Azure DevOps* – automatyzacja procesów wdrażania