Opis Big Data

Big Data to zbiór dużych, złożonych zbiorów danych, które można analizować w celu wydobycia istotnych informacji pomocnych w podejmowaniu decyzji. Ich rozmiar wykracza poza możliwości typowych narzędzi oprogramowania typu baza danych w zakresie przechwytywania, przechowywania, zarządzania i analizowania. Jest rozwinięciem 3 elementów składowych (3V): Ilość (volume), Szybkość (velocity) i różnorodność (variety).

Najważniejsze narzędzia Big Data:

- Bazy danych NoSQL - nierelacyjne, dane mogą być przechowywane bez wcześniejszego zdefiniowania schematu.
- Data Lakes - scentralizowane repozytorium, które pozyskuje i przechowuje duże ilości danych w oryginalnej postaci. Dane mogą być następnie przetwarzane i używane jako podstawa dla różnych potrzeb analitycznych. Ze względu na otwartą, skalowalną architekturę magazyn typu data lake może obsłużyć wszystkie typy danych z dowolnego źródła, od ustrukturyzowanych (tabel bazy danych, arkuszy programu Excel) po częściowo ustrukturyzowane (pliki XML, strony internetowe) do takich bez struktury (obrazy, pliki dźwiękowe, tweety), a wszystko to bez poświęcania wierności.
- Apache Hadoop - oryginalna struktura typu „open source" do przetwarzania rozproszonego i analizy zestawów danych big data w klastrach
- Apache Hive - projekt oprogramowania hurtowni danych. Jest zbudowany na bazie Apache Hadoop w celu zapewnienia zapytań i analiz danych.
- Apache Spark - platforma do przetwarzania danych w klastrze komputerów, która jest używana do analizy big data. Jest to oprogramowanie open source, które zapewnia równoległe przetwarzanie danych.
- Azure Synapse Analytics - usługa do analizy przedsiębiorstwa, która przyspiesza czas wglądu w magazyny danych i systemy danych big data. Usługa Azure Synapse łączy najlepsze technologie SQL używane w magazynowaniu danych przedsiębiorstwa
- Power BI - technologia do tworzenia interaktywnych dashboardów, które można pokazać klientowi.

Architektura Big Data dla Rockstar Games

1. Dane źródłowe (Data Ingestion)

Strumieniowe i wsadowe źródła danych:

- Logi serwera gry (np. błędy, crash logs, FPS drop, itd.)

- Telemetry danych gracza (np. czas sesji, ruchy, interakcje)
- Social Media i support tickets – do analizy nastroju
- Video/Voice data – opcjonalnie do NLP/audio emotion detection

Narzędzia:

- Azure Event Hubs – dla danych w czasie rzeczywistym
- Azure Data Factory – do danych wsadowych (np. backupy logów)
- Azure IoT Hub (jeśli urządzenia klienta mają sensor-like telemetrię)

2. Przechowywanie danych (Data Storage)

Warstwy przetwarzania danych:

- Raw Layer (landing zone) – surowe dane
- Cleansed Layer – przetworzone dane, np. Parquet, Delta
- Curated Layer – dane gotowe do analityki i modelowania

Narzędzia:

- Azure Data Lake Gen2 (ADLS) – centralny magazyn danych
- Azure Synapse Analytics – do analizy i przetwarzania danych

3. Przetwarzanie danych (Compute Layer)

Batch i Stream Processing:

- Azure Stream Analytics – reguły alertowe na podstawie real-time danych
- Apache Spark on Azure Synapse / Azure Databricks – przetwarzanie wsadowe, featury ML, analizy
- Azure Functions / Durable Functions – do lekkich event-driven akcji, np. wysyłanie alertów

4. Machine Learning i AI

Przykłady zastosowania:

- Wykrywanie błędów i crash prediction – klasyfikacja na podstawie logów
- Analiza sentymentu graczy – NLP na komentarzach / Discord / support
- Churn Prediction – przewidywanie utraty graczy
- Anomaly Detection – wykrywanie nietypowych zachowań

Narzędzia:

- Azure Machine Learning (DP-100) – zarządzanie eksperymentami, pipeline'y, rejestr modeli
- Cognitive Services (AI-102) – Text Analytics, Sentiment Analysis, Language
- AutoML / MLflow – do szybkiego testowania modeli
- ONNX / Azure Kubernetes Service – deployment modeli do real-time API

5. Wizualizacja i analiza

Narzędzia:

- Power BI + Synapse – dashboardy do analizy błędów, stabilności gry
- Azure Monitor + Application Insights – dla devopsów i testerów

6. Zarządzanie i bezpieczeństwo

Narzędzia:

- Azure Purview – data governance, lineage
- Azure Key Vault – zarządzanie sekretami i kluczami
- RBAC + Managed Identity – kontrola dostępu
- CI/CD z GitHub Actions lub Azure DevOps – do pipeline'ów danych i modeli