W ekosystemie Big Data i AI mamy kilka kluczowych kategorii narzędzi:

1. **Przechowywanie danych** – Tu dominują rozwiązania typu data lake, takie jak Amazon S3, Azure Data Lake czy Google Cloud Storage. Służą do przechowywania surowych danych z różnych źródeł: logów, plików CSV, strumieni danych itp.
2. **Przetwarzanie danych (ETL/ELT)** – Wykorzystuje się narzędzia takie jak Apache Spark (często przez Databricks), AWS Glue czy Azure Data Factory. Umożliwiają one oczyszczanie, transformację i ładowanie danych do hurtowni lub innych systemów analitycznych.
3. **Strumieniowanie danych** – Do przetwarzania danych w czasie rzeczywistym używa się Apache Kafka, AWS Kinesis, Azure Event Hubs lub Google Pub/Sub. Pozwalają na szybkie reagowanie na zdarzenia, np. wykrycie błędu w grze lub awarii maszyny.
4. **Hurtownie danych** – Dla analiz ad hoc i raportowania wykorzystuje się hurtownie takie jak Amazon Redshift, Azure Synapse czy Google BigQuery. Umożliwiają one szybkie zapytania SQL nad dużymi zbiorami danych.
5. **Uczenie maszynowe i AI** – Każda chmura oferuje własne platformy: Amazon SageMaker (AWS), Azure Machine Learning, Vertex AI (GCP). Pozwalają na trenowanie modeli ML/AI, ich ocenę, deployment i monitoring.
6. **Dashboardy i wizualizacja** – Dane są prezentowane w narzędziach takich jak Power BI, Tableau, Looker lub Amazon QuickSight. Dzięki nim można tworzyć raporty, śledzić KPI i analizować trendy.
7. **Monitoring i alerty** – Systemy takie jak AWS CloudWatch, Azure Monitor czy Prometheus pozwalają na automatyczne śledzenie stanu systemów i reagowanie na błędy lub anomalie.

Przykład PoC – Wykrywanie anomalii na linii produkcyjnej IoT w Azure

Na hali produkcyjnej zainstalowane są czujniki zbierające dane o wibracjach, temperaturze i ciśnieniu maszyn. Celem jest wykrycie anomalii, zanim dojdzie do awarii.

W Azure zaczynam od podłączenia urządzeń do **Azure IoT Hub**, który służy do przyjmowania danych z urządzeń IoT. Dane te w czasie rzeczywistym trafiają do **Azure Stream Analytics**, gdzie można w locie analizować ich zmienność, wykrywać skoki lub odstępstwa od norm.

Surowe dane archiwizuje się w **Azure Data Lake**, natomiast dane przetworzone kieruje się do **Azure Databricks**, gdzie trenowany jest model detekcji anomalii, np. model oparte na Isolation Forest lub Autoencoderze. Gotowy model wdrożony zostanie do **Azure Machine Learning**, skąd jest wykorzystywany w pipeline'ie do przewidywania odchyleń.

Na końcu wszystkie dane i alerty są wizualizowane w **Power BI**, gdzie użytkownicy mogą przeglądać trendy, analizować zdarzenia i podejmować działania operacyjne.

Architektura Big Data dla Rockstar Games (na AWS)

Rockstar Games potrzebuje architektury, która pozwala na analizę ogromnych ilości danych generowanych przez graczy na całym świecie. Te dane obejmują działania w grze, logi błędów, dane czatów, wyniki rozgrywek i interakcje sieciowe.

Gromadzenie danych

Dane z aplikacji i serwerów gier trafiają najpierw do **Kafka (w AWS MSK)**. Pozwala to na przetwarzanie ich w czasie rzeczywistym i przesyłanie dalej do różnych systemów.

Przechowywanie

Surowe dane są składowane w **Amazon S3** jako Data Lake. Dane przetworzone trafiają do **Amazon Redshift** – hurtowni danych, która umożliwia szybkie zapytania i raportowanie.

Przetwarzanie

ETL jest wykonywany za pomocą **AWS Glue** (serverless) lub **Amazon EMR z Apache Spark**, co pozwala na równoległe przetwarzanie ogromnych zbiorów danych, np. w celu wykrywania typowych błędów lub schematów zachowań graczy.

Sztuczna inteligencja

Do predykcji porzuceń gry (churn), rekomendacji treści (np. misje dopasowane do stylu gracza), detekcji nadużyć (cheatów, exploitów), używa się **Amazon SageMaker**, gdzie trenowane są modele ML/AI. Model można wdrożyć jako endpoint, który działa w czasie rzeczywistym.

Monitoring

Wszystkie komponenty są monitorowane przez **Amazon CloudWatch** – logi, metryki i alerty (np. spike błędów w jednym regionie gry) są automatycznie wykrywane.

Prezentacja danych

Dane z Redshift są używane przez **Amazon QuickSight** lub Tableau, które prezentują dashboardy dla analityków, programistów czy zespołu operacyjnego.