Anna Kaniowska

**Popularne narzędzia (3) analityczne/big data na platformach Azure, AWS, GCP, IBM, Alibaba:**

1. **Apache Spark** – Apache Spark to silnik przetwarzania rozproszonego, który umożliwia wydajne przetwarzanie dużych zbiorów danych. Spark udostępnia wiele narzędzi, takich jak Spark SQL, Spark Streaming i MLlib do analizowania, przetwarzania i uczenia maszynowego dużych zbiorów danych.

Dokumentacja:

• https://spark.apache.org/

• https://docs.microsoft.com/en-us/azure/databricks/spark/

2. **Apache Hadoop** - Apache Hadoop to projekt typu open source, na który składa się wiele narzędzi do przetwarzania dużych ilości danych. Głównymi narzędziami są Hadoop Distributed File System „HDFS”, który umożliwia przechowywanie dużych zbiorów danych w klastrze, oraz MapReduce, który umożliwia równoległe przetwarzanie tych danych.

dokumentacja:

• https://hadoop.apache.org/

• https://docs.microsoft.com/en-us/azure/hdinsight/hadoop/

3. **Apache Kafka** – Apache Kafka to platforma streamingowa, która może przetwarzać i przesyłać dane w czasie rzeczywistym. Kafka umożliwia zbieranie danych z wielu źródeł i wysyłanie ich do wielu miejsc docelowych, takich jak bazy danych, analizy i aplikacje czasu rzeczywistego.

dokumentacja:

• https://kafka.apache.org/

• https://docs.microsoft.com/en-us/azure/hdinsight/kafka/

**Notatki do DP-203:**

• Egzamin DP-203: Data Engineering on Microsoft Azure

• Wymagania certyfikacyjne: https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4MbYT

• Co trzeba umieć:

o Przetwarzanie danych przy użyciu narzędzi i usług Azure (np. Azure Stream Analytics, Azure Databricks, Azure HDInsight)

o Przygotowanie i ładowanie danych (np. Azure Data Factory, Azure Synapse Analytics, Azure Cosmos DB)

o Przetwarzanie i przechowywanie danych przy użyciu narzędzi Azure (np. Azure Data Lake Storage, Azure Blob Storage, Azure SQL Database)

**Notatki do DP-100:**

• Egzamin DP-100: Designing and Implementing a Data Science Solution on Azure

• Wymagania certyfikacyjne: https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE3VEHg

• Co trzeba umieć:

o Tworzenie środowiska Azure Machine Learning (np. tworzenie i konfiguracja eksperymentów, zarządzanie danymi, zarządzanie środowiskiem)

o Przygotowanie danych do analizy (np. czyszczenie danych, scalanie danych, wizualizacja danych)

o Implementacja modeli uczenia maszynowego

Dodatkowo: **Przykład**

**Robisz PoC na wykrycie anomalii z linii produkcyjnej (jakieś IoT). Jakich narzędzi użyjesz np. w Azure. Jak będzie wyglądało PoC.**

Do wykrywania anomalii na linii produkcyjnej można wykorzystać wiele narzędzi dostępnych na Azure, takich jak:

o Azure Stream Analytics — narzędzie do przetwarzania strumieni, które umożliwia przetwarzanie, analizę i wizualizację danych w czasie rzeczywistym. Może służyć do analizy danych z czujników IoT i wykrywania anomalii w czasie rzeczywistym.

o Azure Functions — platforma do budowania mikrousług, które mogą wykonywać kod w chmurze w odpowiedzi na zdarzenia. Może służyć do przetwarzania i analizowania danych z czujników IoT oraz wykrywania anomalii.

o Azure Machine Learning — platforma uczenia maszynowego, która umożliwia tworzenie i trenowanie modeli uczenia maszynowego, których można używać do wykrywania anomalii. Przykładowy PoC wygląda tak:

1. Skonfiguruj czujniki IoT do łączenia się z usługą Azure IoT Hub.

2. Skonfiguruj strumienie danych i przetwarzaj dane z czujników za pomocą Azure Stream Analytics.

3. Użyj Azure Functions do przetwarzania danych przesyłanych strumieniowo i wykrywania anomalii.

4. Użyj usługi Azure Machine Learning, aby wyszkolić model uczenia maszynowego w celu wykrywania anomalii.

5. Zintegruj swój model z usługą Azure Functions, aby automatycznie wykrywać anomalie i reagować na nie.

**Zaproponuj architekturę Big Data dla Rockstar Games przetwarzanie danych w celu wykrycia i zapobiegania błędom w grze, satysfakcji użytkowników. Jakie narzędzia, gdzie będzie compute ect…**

Propozycja architektury Big Data dla Rockstar Games mogłaby składać się z następujących kroków i narzędzi:

1. Zbieranie danych z gry: Zastosowanie mechanizmów telemetrii gry, które umożliwią zbieranie i przekazywanie danych o działaniu gry do chmury.

Narzędzia: Azure Stream Analytics, Azure Event Hubs

1. Składowanie danych: Dane zebrane z gry muszą być składowane w sposób umożliwiający ich szybki i łatwy dostęp dla analizy.

Narzędzia: Azure Blob Storage, Azure Data Lake Storage

1. Przetwarzanie danych: Dane muszą być przetwarzane w celu identyfikacji błędów i wykrywania trendów w zachowaniu użytkowników.

Narzędzia: Azure Databricks, Apache Spark

1. Analiza danych: Wykorzystanie narzędzi do analizy danych umożliwi wykrycie błędów oraz określenie trendów w zachowaniu użytkowników.

Narzędzia: Azure Machine Learning, Power BI

1. Reagowanie na dane: Automatyczne lub ręczne podejmowanie działań w odpowiedzi na wykryte błędy lub niezwykłe zachowanie użytkowników.

Narzędzia: Azure Functions, Logic Apps

1. Monitorowanie systemu: Monitoring systemu umożliwi wczesne wykrywanie problemów i reagowanie na nie zanim wpłyną one na doświadczenie użytkowników.

Narzędzia: Azure Monitor, Application Insights

1. Compute: Do wykonania powyższych kroków potrzebna będzie odpowiednia infrastruktura obliczeniowa. W zależności od specyfikacji i wymagań systemu, możliwe będzie użycie wirtualnych maszyn, kontenerów lub funkcji Azure Functions.

Narzędzia: Azure Virtual Machines, Azure Kubernetes Service, Azure Functions.