Zajęcia 5

Zadanie 1

**Cel zadania – przygotować się do interview**,

Notatka z krótkim opisem najważniejszych narzędzi analitycznych/Big Data/AI. Móc opisać **wybrane** przez siebie zagadnienie/narzędzia analityczne **w wybranej** chmurze Azure/AWS/GCP/IBM/Alibaba.

Przykład

1. Robisz PoC na wykrycie anomalii z linii produkcyjnej (jakieś IoT). Jakich narzędzi użyjesz np. w Azure. Jak będzie wyglądało PoC.
2. 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…

Podpowiedź - przejrzyj co jest wymagane do certyfikacji np. DP-203, i DP-100, AI-102

<https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE3VEHg>

<https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4MbYT>

<https://learn.microsoft.com/pl-pl/credentials/certifications/resources/study-guides/ai-102>

Apache Spark: Framework do przetwarzania danych w czasie rzeczywistym i przetwarzania wsadowego. Wykorzystuje pamięć podręczną i możliwość obsługi różnych rodzajów danych, takich jak strumienie, dane tekstowe, grafy i dane tabelaryczne.

Apache Hadoop: Framework do przetwarzania i przechowywania dużych zbiorów danych w rozproszonym środowisku. Hadoop składa się z Hadoop Distributed File System (HDFS) do przechowywania danych oraz MapReduce do przetwarzania danych równolegle.

Hive: Narzędzie zbudowane na bazie Hadoopa, które umożliwia przetwarzanie danych w stylu SQL na danych przechowywanych w HDFS. Hive tłumaczy zapytania SQL na zadania MapReduce.

TensorFlow: Biblioteka open-source do uczenia maszynowego i głębokiego uczenia. TensorFlow oferuje narzędzia do budowy, treningu i wdrażania modeli uczenia maszynowego, w tym sieci neuronowych.

PyTorch: Biblioteka open-source do uczenia maszynowego i głębokiego uczenia. PyTorch jest znany ze swojej elastyczności i wydajności, a także umożliwia programowanie dynamiczne.

Google Cloud Platform (GCP) BigQuery: Usługa analityczna w chmurze GCP, która umożliwia wykonywanie zapytań SQL na dużych zbiorach danych bez potrzeby zarządzania infrastrukturą. BigQuery oferuje skalowalność i wydajność przetwarzania danych.

Microsoft Azure Machine Learning: Usługa w chmurze Azure do budowy, treningu i wdrażania modeli uczenia maszynowego. Azure Machine Learning oferuje narzędzia do eksperymentowania z różnymi algorytmami uczenia maszynowego i automatyzacji procesu budowy modeli.

Aby przeprowadzić PoC na wykrycie anomalii na linii produkcyjnej żarówek w chmurze Azure, możemy wykorzystać kilka usług, w tym:

1. **Azure IoT Hub**: Usługa do przesyłania, przetwarzania i zarządzania danymi z urządzeń IoT w chmurze. Azure IoT Hub może być używany do zbierania danych z czujników monitorujących produkcję żarówek na linii produkcyjnej.
2. **Azure Stream Analytics**: Usługa do przetwarzania strumieniowego danych w czasie rzeczywistym. Może być używana do analizy danych z IoT Hub w celu wykrycia anomalii w produkcji żarówek na podstawie różnych czynników, takich jak temperatura, ciśnienie, prąd itp.
3. **Azure Machine Learning**: Platforma do budowy, treningu i wdrażania modeli uczenia maszynowego. Może być wykorzystana do budowy modeli detekcji anomalii na podstawie danych historycznych z produkcji żarówek. Te modele mogą być następnie wdrożone do Azure Stream Analytics w celu analizy danych w czasie rzeczywistym.
4. **Azure Functions**: Usługa do tworzenia i wdrażania mikrousług w chmurze. Może być używana do uruchamiania niestandardowych skryptów lub algorytmów detekcji anomalii na podstawie danych przetwarzanych przez Azure Stream Analytics.

PoC na wykrycie anomalii na linii produkcyjnej żarówek przy użyciu tych narzędzi:

1. **Zbieranie danych z czujników**: Konfiguracja Azure IoT Hub do zbierania danych z czujników na linii produkcyjnej żarówek. Dane te będą obejmować informacje o temperaturze, ciśnieniu, prądzie itp.
2. **Przetwarzanie strumieniowe danych**: Konfiguracja Azure Stream Analytics do przetwarzania strumieniowych danych z Azure IoT Hub. Zdefiniowanie zapytania SQL w Azure Stream Analytics do analizy danych w czasie rzeczywistym i wykrywania potencjalnych anomalii.
3. **Budowa modeli uczenia maszynowego**: Przygotowanie danych historycznych z produkcji żarówek i użycie ich do treningu modeli detekcji anomalii w Azure Machine Learning. Wybranie odpowiednich algorytmów i ocena jego skuteczności na danych testowych.
4. **Integracja modeli z Azure Stream Analytics**: Wdrożenie modeli uczenia maszynowego do Azure Stream Analytics lub Azure Functions w celu wykrywania anomalii na podstawie danych przetwarzanych w czasie rzeczywistym.
5. **Testowanie i optymalizacja**: Przeprowadzenie testów PoC, analiza wyników i dostosowanie algorytmów lub parametrów w celu poprawy skuteczności wykrywania anomalii.
6. **Raportowanie i monitorowanie**: Konfiguracja mechanizmów raportowania i monitorowania, aby śledzić wyniki analizy danych i reagować na wykryte anomalie w produkcji żarówek.

Poprzez wykorzystanie tych narzędzi i przeprowadzenie PoC, można zweryfikować skuteczność wykrywania anomalii na linii produkcyjnej żarówek oraz ocenić potencjalne korzyści biznesowe z wdrożenia pełnego rozwiązania.