Karol Śliwa

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

- Apache Spark jest otwarto źródłowym frameworkiem do przetwarzania danych, który umożliwia skalowalne i szybkie analizy danych. Zapewnia wsparcie dla różnych języków programowania, takich jak Scala, Python i Java, oraz oferuje elastyczność w przetwarzaniu dużej ilości danych w klastrach. Można znaleźć dokumentację Apache Spark na oficjalnej stronie projektu: <https://spark.apache.org/documentation.html>

- Apache Hadoop to framework, który umożliwia rozproszone przetwarzanie i przechowywanie dużych zbiorów danych w klastrach komputerowych. Hadoop zapewnia narzędzia takie jak Hadoop Distributed File System (HDFS) do przechowywania danych oraz MapReduce do przetwarzania równoległego. Możesz znaleźć dokumentację Apache Hadoop na oficjalnej stronie projektu: <https://hadoop.apache.org/documentation.html>

- Apache Kafka to rozproszona platforma do przesyłania strumieniowego danych w czasie rzeczywistym. Jest to narzędzie o wysokiej wydajności i skalowalności, które umożliwia przesyłanie, przechowywanie i przetwarzanie strumieni danych między różnymi aplikacjami i systemami. Można znaleźć dokumentację Apache Kafka na oficjalnej stronie projektu: <https://kafka.apache.org/documentation/>

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

Do realizacji PoC (Proof of Concept) na wykrycie anomalii z linii produkcyjnej opartego na technologii IoT w platformie Azure, można skorzystać z następujących narzędzi:

1. Azure IoT Hub: Służy do zbierania, przetwarzania i zarządzania danymi z urządzeń IoT. Umożliwia przesyłanie strumieniowych danych z linii produkcyjnej do chmury Azure.
2. Azure Stream Analytics: Pozwala na przetwarzanie strumieniowych danych w czasie rzeczywistym. Można użyć go do analizy i wykrywania anomalii na podstawie danych zebranych z IoT Hub.
3. Azure Machine Learning: Narzędzie do tworzenia i wdrożenia modeli uczenia maszynowego. Można zastosować je do szkolenia modeli wykrywających anomalie na podstawie zebranych danych i ich implementacji w aplikacji.

Przebieg PoC może obejmować następujące kroki:

1. Konfiguracja i rejestracja urządzeń IoT w Azure IoT Hub.
2. Utworzenie strumienia danych z urządzeń IoT do Azure Stream Analytics.
3. Skonfigurowanie reguł i zapytań w Azure Stream Analytics do wykrywania anomalii na podstawie danych strumieniowych.
4. Szkolenie modelu wykrywającego anomalie za pomocą Azure Machine Learning na podstawie zebranych danych historycznych.
5. Wdrożenie modelu uczenia maszynowego w aplikacji lub w Azure Stream Analytics.
6. Testowanie i monitorowanie systemu w celu wykrywania anomalii na linii produkcyjnej i analizy wyników.

Oczywiście, w zależności od specyfiki przypadku, konkretny PoC może różnić się szczegółami i dodatkowymi narzędziami, ale powyższe wskazówki stanowią ogólny przewodnik.

1. **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…**

Nasze pierwsze zadanie to zbieranie danych z różnych źródeł, takich jak serwery gry, logi zdarzeń oraz dane graczy. Do tego celu wykorzystuję platformę do zbierania danych, na przykład Apache Kafka, aby przesyłać strumieniowe dane do naszego systemu Big Data.

Kolejnym krokiem jest przechowywanie zgromadzonych danych. Wykorzystuję technologie Big Data, takie jak Apache Hadoop lub Apache Spark, aby przechowywać duże zbiory danych. Dzięki temu mogę korzystać z Hadoop Distributed File System (HDFS) do trwałego przechowywania danych oraz Apache Hive lub Apache HBase do analizy i zapytań.

Aby analizować dane w czasie rzeczywistym i wykrywać błędy na bieżąco, wykorzystuję narzędzia do przetwarzania strumieniowego, takie jak Apache Storm, Apache Flink lub Apache Kafka Streams.

Przetwarzanie wsadowe odbywa się cyklicznie lub na żądanie, korzystając z narzędzi takich jak Apache Spark lub Apache Hadoop wraz z Apache Hive lub Apache Pig. To pozwala nam na przeprowadzanie zaawansowanych analiz i modelowania danych.

Wprowadzenie technik uczenia maszynowego i sztucznej inteligencji jest kluczowe.

Wykorzystuję narzędzia do uczenia maszynowego, takie jak TensorFlow, scikit-learn lub Apache Spark MLlib, aby trenować modele na podstawie zgromadzonych danych. Następnie wykorzystuję te modele do wykrywania wzorców, identyfikowania błędów i podejmowania odpowiednich działań.

Ostatecznie, tworzę interaktywne panele i raporty, które pozwalają łatwo analizować i wizualizować wyniki. Do tego celu używam narzędzi takich jak Tableau, Power BI lub Apache Superset.

Dzięki tej architekturze możemy skutecznie przetwarzać dane, wykrywać błędy w grze oraz zwiększać satysfakcję użytkowników poprzez szybką analizę i reagowanie na zgromadzone dane.

**Notatki Dotyczące DP-203 i DP-100**

Notatki dotyczące certyfikacji DP-203 - Data Engineering on Microsoft Azure:

* Egzamin DP-203 obejmuje zagadnienia z zakresu inżynierii danych w środowisku Microsoft Azure.
* Wymagania certyfikacyjne oraz szczegółowe informacje można znaleźć pod adresem: <https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4MbYT>
* W ramach egzaminu należy posiadać umiejętności związane m.in. z przetwarzaniem danych przy użyciu narzędzi i usług Azure, takich jak Azure Stream Analytics, Azure Databricks czy Azure HDInsight, a także z przygotowaniem i ładowaniem danych przy użyciu Azure Data Factory, Azure Synapse Analytics czy Azure Cosmos DB, oraz z przetwarzaniem i przechowywaniem danych przy wykorzystaniu narzędzi takich jak Azure Data Lake Storage, Azure Blob Storage czy Azure SQL Database.

Notatki dotyczące certyfikacji DP-100 - Designing and Implementing a Data Science Solution on Azure:

* Egzamin DP-100 dotyczy projektowania i wdrażania rozwiązań z dziedziny nauki danych w środowisku Azure.
* Wymagania certyfikacyjne oraz szczegółowe informacje można znaleźć pod adresem: <https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE3VEHg>
* W ramach egzaminu należy posiadać umiejętności związane m.in. z tworzeniem środowiska Azure Machine Learning, takich jak tworzenie i konfiguracja eksperymentów, zarządzanie danymi oraz zarządzanie środowiskiem. Ponadto, wymagane jest przygotowanie danych do analizy, włączając w to czyszczenie danych, scalanie danych i wizualizację. Ostatecznie, egzamin obejmuje również implementację modeli uczenia maszynowego.