**Lab6 – Grzegorz Cichy – 405708**

**Przykładowa architektura Big Data dla przemysłu do wykrywania anomalii i zapobiegania błędom na produkcji.**

Cały system jest oparty o proces ETL.

**Extract** – Zbierane są dane zarówno historyczne czy transakcyjne (są one przechowywane w produkcyjnych bazach danych) jak i real-time data z logów produkcyjnych. Mają one różne struktury, np.: json, csv, log etc.

**Transform** – Zbierane dane ulegają transformacji: są pre-procesowane – następuje czyszczenie i ujednolicanie struktury plików, aby później dało się je składować w Data Warehouse. Do Batch Data stosowana jest usługa Azure Databricks, natomiast do danych czasu rzeczywistego usługa Azure Stream Analytics.

**Load –** Kolejny krok to załadowanie przeprocesowanych danych do hurtowni danych. Do jej obsługi użyta jest platforma Azure Synapse Analytics, która skaluje i monitoruje wydajność Warehouse’u, udostępnia narzędzia związane z ML, służące do budowania modeli predykcyjnych, aby zwiększyć skuteczność zapobiegania anomaliom. Dodatkowo jest zintegrowana z narzędziami do wizualizacji, tworzenia raportów jak Power BI.

**Narzędzia analityczne/Big Data w Azure:**

**Azure Data Lake Storage Gen2 –** repozytorium umożliwiające składowanie danych wszelkiego rodzaju, ustrukturyzowanych i nieustrukturyzowanych. Nie ma ścisłej struktury jak chociażby baza danych, można tam składować pliki nie licząc z się z ograniczeniami, a transformować je w dalszym etapie procesu.

<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-introduction>

**Azure Data Factory –** usługa chmurowa, umożliwia integrację danych między różnymi systemami. Za jej pośrednictwem możliwe jest podłączenie się do systemu źródłowego, pobranie i transformacja danych oraz załadownie ich do innego systemu.

<https://learn.microsoft.com/en-us/azure/data-factory/>

**Azure Databricks –** narzędzie chmurowe oparte na silniku Apache Spark, służące do transformacji dużych zbiorów danych oraz ich integrację z innymi usługami.

<https://learn.microsoft.com/en-us/azure/databricks/introduction/>

**Azure Synapse Analytics –** jest to narzędzie analityczne pozwalające zarządzać hurtowniami danych i całymi systemami Big Data. Dostarcza różne technologie Big Data takie jak Apache Spark, SQL, narzędzia do przetwarzania danych czasu rzeczywistego itp.

<https://azure.microsoft.com/en-us/products/synapse-analytics>