**Big Data tools in AZURE**

Azure Data Lake Storage gen2 – miejsce do przechowywania danych. Ustrukturyzowanych jak i nie, w formatach natywnych np. plikach binarnych. Początkowe położenie dla danych skąd trafiają do dalszej przeróbki. Zcentralizowane repozytorium

Azure Data Factory – służy do automatyzacji przetwarzania danych. Jest to serwis ETL (Extract Transform Load). Dzięki temu narzędziu możemy wyciągnąć dane z Data Lake Storage, następnie Transformować i przekazać dalej, do hurtowni. Świetne narzędzie do automatyzacji procesów np. aby przetwarzać raporty z gier.

Azure databricks – serwis bazujący na apache spark, przeznaczony głównie do big data oraz machine learning. Można tam korzystać ze scali, pythona, java, sql, R oraz z bibliotek do AI/ML TensorFlow, Pytorch. Nastawienie na Sparka.

Azure HDInsight – Zawiera więcej narzędzi do Big Data, generalnie Hadoop tools, jednak spark jest bardziej wydajny na databricks, ponadto jest mniej user friendly. Do naprawdę Big Data, narzędzie może już trochę przestarzałe.

Azure Stream Analytics – platforma przeznaczona do obsługi danych napływających strumieniowo. Zorientowana na JavaScript i SQL. Lepiej się sprawdzi, kiedy nie mamy stałych danych tylko napływające w czasie rzeczywistym np. z czujników IoT.

Azure Machine Learning – narzędzie dedykowane do ML. Zapewnia świetne środowisko do tworzenia różnych modeli. Pozwala także wykonywać testy, aby zdecydować się, który model wybrać, dzięki możliwości ich ewaluacji i wizualizacji pewnych metryk.

Azure Synapse Analytics – platforma od Azure, data warehouse. Wspiera sparka, ale przede wszystkim jest przeznaczona do dużej ilości zapytań T-SQL. Można tam też tworzyć pipeline’y oraz podłączyć takie usługi jak powerBI. Nastawienie na SQL.

Data Lake Analytics – alternatywa dla Azure Synapses. W przeciwieństwie do konkurenta nie pobiera on do siebie danych tylko bezpośrednio zasięga od Data Lake Storage. Dużo lepiej sprawdzi się w przypadku, kiedy chcemy wykonać pewną jednostkową analizę, ponieważ nie trzeba stawiać całej infrastruktury.

Azure Analysis services – kolejna usługa do analizy danych. Tutaj zaletą jest to, że nie wspiera ona tylko SQL. Tak więc, kiedy byłoby 500 zapytań SQL i stać w kolejce. Do wybrania, jeśli potrzebujemy wykonać coś na szybko w BI.

Azure Monitor – pozwala kontrolować przepływ danych oraz udostępnia statystyki na temat ich przetwarzania. Możemy korzystać z różnych metryk oraz je wizualizować. Przydatne do obserwacji infrastruktury.

Azure role-based acces control – narzędzie do kontrolowania uprawnień do infrastruktury. Bazując na rolach możemy przydzielać dostęp do poszczególnych części użytkownikom. Kolejne przydatne narzędzie w potoku przetwarzania danych.

Image on the next page.

Obraz zawierający diagram

Opis wygenerowany automatycznie