Aby dobrze zrozumieć, jak stworzyć architekturę Big Data najpierw trzeba opanować proces, który jest podstawą w tego typu architekturach, czyli ETL (Extract, Transform, Load). Proces ten idealnie nadaję się do infrastruktury Big Data, ponieważ pozwala on na skuteczne i efektywne przenoszenie danych między różnymi systemami i bazami danych. W ujęciu Big Data, jak sama nazwa wskazuję mamy do czynienia z ogromnymi ilościami danych, a więc zależy nam na jak najszybszym i jak najbardziej skutecznym przenoszeniu ich.

Dodatkowo proces ten pozwala na transformację i łączenie danych w bardziej użyteczne i zrozumiałe dla użytkownika formy.

Jest on również elastyczny, co powoduje, że wykorzystywany jest w wielu różnych dziedzinach, takich jak biznes, nauka, medycyna, finanse i wiele innych. Pomaga zarządzać danymi, a co za tymi idzie wpływa pomocnie na podejmowanie decyzji i analizowanie dużych zbiorów danych.

Myślę, że z roku na rok coraz więcej firm i dziedzin będzie stosować ten proces, ponieważ mamy coraz częściej do czynienia z dużymi zbiorami danych.

Zatem, przechodząc po kolejnych krokach procesu ETL.

1. Extract (wyodrębnienie) – pierwszy krok ETL, polega na wyodrębnieniu danych z różnych źródeł. Mogą to być dane z baz danych, plików tekstowych, arkuszy kalkulacyjnych lub innych aplikacji. Często źródła te różnią, się formatem, strukturą, a nawet sposobem przechowywania danych. W tym kroku należy wyodrębnić tylko te dane, które są potrzebne do dalszej obróbki, czyli spełniają określone kryteria wyboru.
2. Transform (przekształcenie) – po wyodrębnieniu danych należy dokonać ich transformacji do docelowej postaci. W tym kroku przeprowadzane są różne operacje na danych, takie jak usuwanie duplikatów, filtrowanie, sortowanie, scalanie lub konwersja formatu danych. Wszystkie te operacje mają na celu przygotowanie danych do dalszej analizy i łatwej ich integracji z innymi źródłami danych.
3. Load (ładowanie) – ostatnim krokiem ETL jest ładowanie danych do docelowej lokalizacji, takiej jak hurtowania danych lub baza danych, a następnie są indeksowane i oznaczane odpowiednimi etykietami, aby ułatwić ich późniejsze wyszukiwanie i analizę. Wszelkie błędy lub nieprawidłowości w danych są usuwane lub oznaczane, aby uniknąć wprowadzania błędnych informacji do systemu.

Oprócz wymienionych kroków, często w połączeniu z ETL stosuje się również krok

- Compute – przetwarzanie danych, w którym dane są analizowane, przetwarzane i agregowane, aby wygenerować pożądane wyniki. W tym etapie można wykorzystać różne narzędzia i technologie, takie jak Hadoop, Spark, Scala lub inne, aby przeprowadzić analizę danych w dużych ilościach. W tym etapie można również wykorzystać algorytmy uczenia maszynowego, aby wyciągnąć wnioski z danych.

- Visualize – etap wizualizacji danych, w którym wyniki przetwarzania danych są przedstawiane w formie wykresów, tabel lub innych wizualizacji, które pomagają w zrozumieniu danych i ich analizie.

Oba etapy, są zwykle realizowane po etapie Transform, kiedy dane zostały oczyszczone i przekształcone do odpowiedniego formatu.

**Postanowiłem stworzyć architekturę dla Valve w celu zapobiegania błędów w grze Counter Strike i poprawienia satysfakcji użytkowników**

Counter Strike jest grą z ponad milionową dzienną liczbą graczy, posiadać będziemy zatem ogromną ilość danych zbieranych każdego dnia.

1. Extract: wyodrębnienie danych z serwisów społeczniościowych, opinii użytkowników na portalach gamingowych, a także opinii o produkcie na steam. Dodatkowo najważniejszymi wyodrębnionymi danymi będą logi (dziennik zdarzeń), zobaczymy tym sposobem skąd pochodzą gracze, o jakich godzinach grają i ile czasu. Dane będą pochodzić z różnych źródeł, a co za tym idzie będą posiadać różne formaty. W celu wyodrębnienia danych z różnych źródeł wykorzystam Azure Data Lake Storage Gen2, który umożliwia przechowywanie i przetwarzanie dużych zbiorów danych z różnych źródeł w jednym miejscu, co ułatwia ich dalsze przetwarzanie i analizę.
2. Transform: teraz dane należy poddać transformacji, tak aby można je było wykorzystać w dalszym przetwarzaniu, a docelowo zanalizować. W tym kroku skorzystam z Azure Databricks, który oferuje łatwe tworzenie i konfigurowanie klastrów przetwarzania danych w celu przetwarzania dużych zbiorów danych w czasie rzeczywistym.
3. Compute: po transformacji dane należy przetworzyć w celu uzyskania konkretnych wyników i analizy. W tym kroku również można skorzystać z wielu narzędzi od Azure, takich jak np. Azure HDInsight, które umożliwia skalowanie poziomu przetwarzania dla dużych zbiorów danych.
4. Load: w tym kroku dane są ładowane do hurtowni danych lub bazy danych, które umożliwią łątwe wyszukiwanie, analizę i wizualizację wyników. Tutaj skorzystamy z Azure Synapse Analytics, które skaluje i monitoruje wydajność Warehouse’u, udostępnia narzędzia związane z uczeniem maszynowym, służące do budowania modeli predykcyjnych, aby zwiększyć skuteczność zapobiegania błędom w grze, a także wykryć trendy w negatywnych jak i pozytywnych opiniach graczy, aby wiedzieć co należy ulepszyć, a co zostawić.
5. Visualise: przetworzone dane zostaną zwizualizowane, aby ułatwić analizę i zrozumienie danych. W tym celu można skorzystać z Power BI.

**3 narzędzia analityczne/Big Data** **Azure, które są popularne i często wykorzystywane**

1. Azure Databricks – platforma do przetwarzania dużych zbiorów danych, która łączy w sobie silnik przetwarzania danych Apache Spark z platformą chmurową Microsoft Azure. Platforma ta oferuje narzędzia do przetwarzania danych, analizy, eksploracji i wizualizacji danych. Pozwala ona również na tworzenie modeli uczenia maszynowego.

Link do dokumentacji: <https://azure.microsoft.com/en-us/products/databricks/>

1. Azure Synapse Analytics – usługa ta pozwala zarządzać hurtowniami danych i całymi systemami Big Data. Oferuje wiele narzęzi i funkcji, takich jak przetwarzanie wielu kroków (ETL), integracja z usługami Azure, w tym Azure Data Factory i Power BI oraz możliwość przetwarzania danych w czasie rzeczywistym.

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

1. Azure HDInsight – usługa chmurowa, która umożliwia wdrażanie i zarządzanie klastrami Big Data, które wykorzystują popularne frameworki, akie jak Hadoop, Spark, Hive, HBase i wiele innych.

Link do dokumentacji: https://azure.microsoft.com/en-us/products/hdinsight/