Rozwiązania
Obiektowe HCI

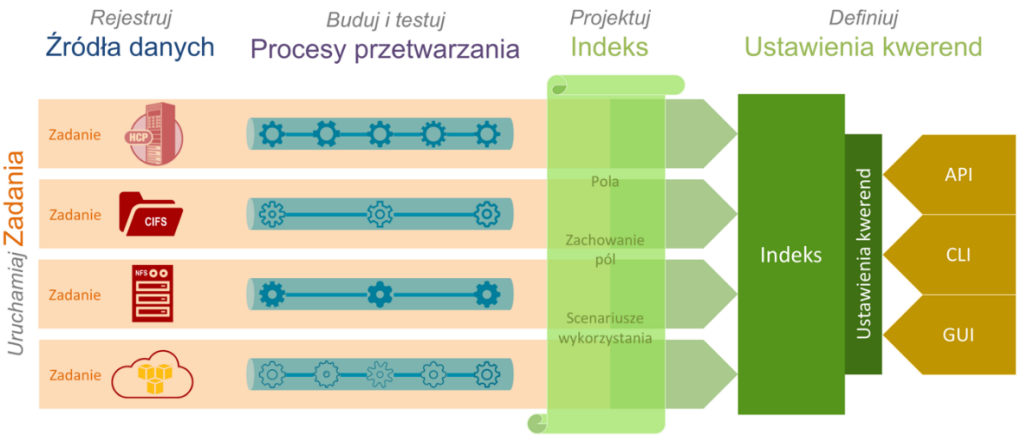
Hitachi
Content Intelligence

Agregacja danych oraz ich kontekstowa świadomość. Korelacja oraz wzbogacanie danych (w tym również danych niestrukturalnych), w taki sposób, aby były one gotowe do wykorzystania w analizie biznesowej. I w końcu akcje podejmowanie w inteligentny sposób w oparciu o rezultaty takiej analizy. To tylko niektóre wartości dostarczane przez Hitachi Content Intelligence (HCI).

HCI jest to rozwiązanie, wykorzystywane do indeksowania i wyszukiwania danych oraz metadanych. Oprócz tego jest ono jednak również wyposażone w potężne mechanizmy ETL (Extract, Transform, Load). I to te mechanizmy pozwalają na odpowiednie przygotowanie danych, zanim zostaną one zaindeksowane. HCI pozwala dzięki temu na lepsze zrozumienie danych poprzez ich centralizacje oraz dzięki analizie treści i możliwości eksploracji różnych źródeł danych.

A wszystko to, aby dostarczyć wiedzy na temat przechowywanych danych, zwiększyć produktywność (m.in. produktywność wyszukiwania w źródłach niestrukturalnych) oraz obniżać ryzyko biznesowe (dzięki świadomości tego co znajduje się w plikach oraz innych źródłach danych).

> **"HCI jest to rozwiązanie, wykorzystywane do indeksowania i wyszukiwania danych oraz metadanych."**



Logiczna struktura działania Hitachi Content Intelligence.

Projektowanie zadań w HCI rozpoczyna się od poznania wymagań operatorów, którzy będą używali tej aplikacji.

Znając potrzeby operatorów oraz wiedząc, jakich informacji będą oni poszukiwać, można zaprojektować kwerendy, ich ustawienia oraz sposób wyświetlania rezultatów wyszukiwania. Te mogą być dostarczane albo w standardowej aplikacji do wyszukiwania dostępnej w HCI, albo w dowolnej aplikacji biznesowej operatora, która zostanie zintegrowana z HCI poprzez dostępny interfejs programistyczny API.

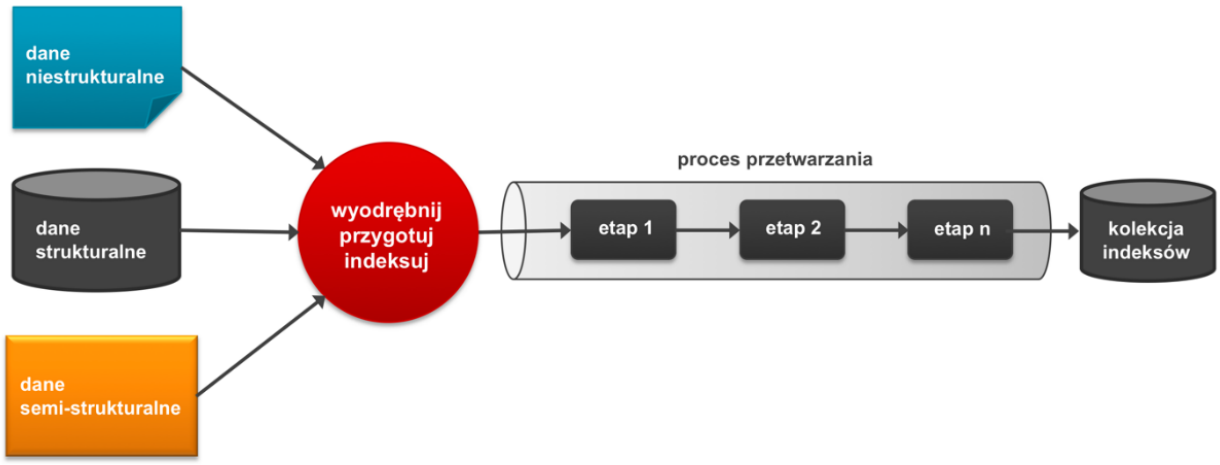
W ten sposób użytkownik biznesowy może realizować wyszukiwania, nie wychodząc ze swojej aplikacji. Gdy już wiemy co i w jaki sposób należy wyświetlić w rezultacie wyszukiwania, wówczas możemy zacząć projektować indeks. Aby zwiększyć efektywność działania rozwiązania dobrze jest wybrać do zaindeksowania tylko potrzebne pola (dane i metadane). Wcześniej natomiast bardzo często dane oraz ich metadane trzeba w odpowiedni sposób przygotować, wzbogacić, sklasyfikować, przefiltrować i wyodrębnić. I temu służy proces przetwarzania (tzw. pipeline) w ramach którego administrator (osoba, która przygotowuje dane do zaindeksowania) może korzystać z wbudowanych w HCI etapów przetwarzania oraz tworzyć swoje własne etapy w oparciu o gotowe biblioteki SDK oraz interfejs programistyczny API.

> **„Aby zwiększyć efektywność działania rozwiązania dobrze jest wybrać do zaindeksowania tylko potrzebne pola (dane i metadane).”**

Aby jednak można było zacząć przetwarzać dane, należy przede wszystkim podłączyć HCI do źródła danych, które chcemy, aby było analizowane i indeksowane. W tym celu wykorzystywane są albo gotowe konektory do takich źródeł jak HCP, HCP Anywhere, HCP Anywhere Edge, CIFS, NFS, S3, Kafka, HDFS, JDBS, albo biblioteki SDK i API do budowania konektorów do innych źródeł.

Praca z Hitachi Content Intelligence polega na budowaniu procesów przepływu (tzw. workflow), które składają się z trzech logicznych komponentów:

- **Konektorów do źródeł danych** – gotowe lub własne zaprojektowane w oparciu o biblioteki SDK i interfejs API.
- **Procesów przetwarzania danych (pipeline)** – proces taki składa się z etapów przetwarzania danych, których celem może być: normalizacja, filtrowanie, kategoryzowanie, analiza, ekstrakcja i wiele innych typów transformacji przeprowadzanych na danych i metadanych, które mają być zaindeksowane i możliwe do wyszukiwania. Gotowe lub własne zaprojektowane w oparciu o biblioteki SDK i interfejs API.
- **Indeksów** – zaprojektowanych w taki sposób, aby uwzględnione zostały w nich wszystkie pola (otrzymane w wyniku transformacji przeprowadzonych w procesie przetwarzania – pipeline), które będą wyszukiwane przez użytkowników końcowych. Tutaj również projektowany jest wygląd strony dla użytkowników końcowych (wyszukiwarka) oraz definiowane są kwerendy zapytań.



Proces przepływu danych (workflow) w HCI.

Hitachi Content Intelligence działa w oparciu o technologie kontenerów i mikro-usług

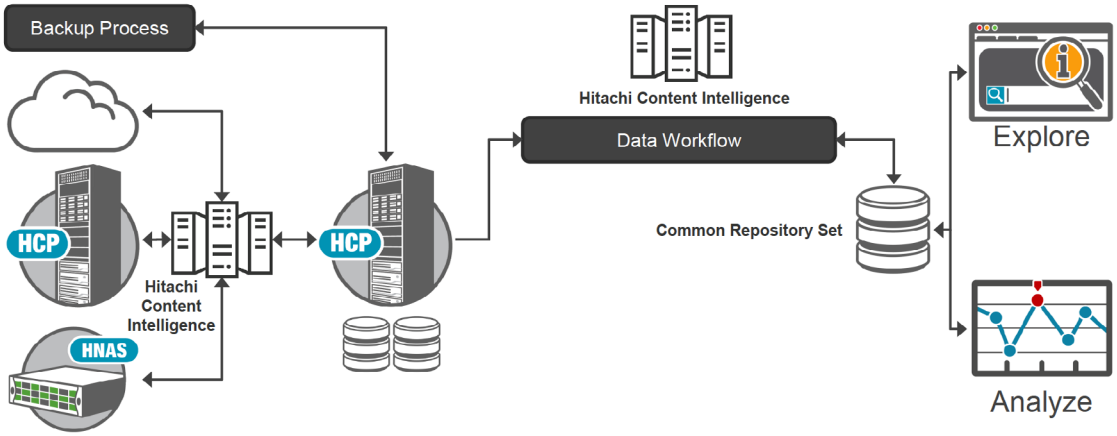
HCI może być zainstalowany na serwerach fizycznych, w wirtualnych maszynach lub w chmurze publicznej. Instancje HCI tworzą klaster HA i są instalowane w systemie operacyjnym Linux z zainstalowanym Dockerem. Klaster HCI może być skalowany (scale-out) w sposób dynamiczny poprzez dodawanie kolejnych instancji, a jego wielkość (min. 4 instancje w konfiguracji HA, maksimum 10 tys. instancji) będzie zależała od:

- ilości indeksowanych plików,
- wymaganego czasu zaindeksowania istniejących i nowych plików,
- wymaganego czasu wyszukiwania plików przez użytkowników końcowych.

„Scenariuszy wykorzystania HCI w przedsiębiorstwie jest bardzo wiele. Niektóre z nich sklasyfikowane w trzech różnych kategoriach przedstawiono poniżej.”

Wyszukiwanie i klasyfikacja treści

W tych przypadkach użycia wyszukujemy dane i metadane w posiadanych źródłach (przestrzeniach dyskowych, systemach, bazach danych), identyfikujemy je i przypisujemy im określoną wartość. Dane są indeksowane i często kategoryzowane.



Przykład wyszukiwania i klasyfikacji treści.

Przykłady:

1. **GDPR/RODO – HCI** pozwala m.in. na wyszukiwanie danych i metadanych w różnych źródłach, zaindeksowanie ich oraz skategoryzowanie i oznaczenie odpowiednią flagą. Dzięki tym mechanizmom możliwe jest przeszukanie istniejących zbiorów danych w poszukiwaniu plików, rekordów, logów, kolejek asynchronicznych, itp., które zawierają dane osobowe klientów przedsiębiorstwa. Przeszukiwane są przy tym nie tylko metadane, ale również dane na przykład wewnątrz pliku (umowy, regulaminy lub inne dokumenty). Przeszukiwanie realizowane jest w oparciu o zdefiniowane wzorce w postaci wyrażeń regularnych, plików JSON i plików XML.

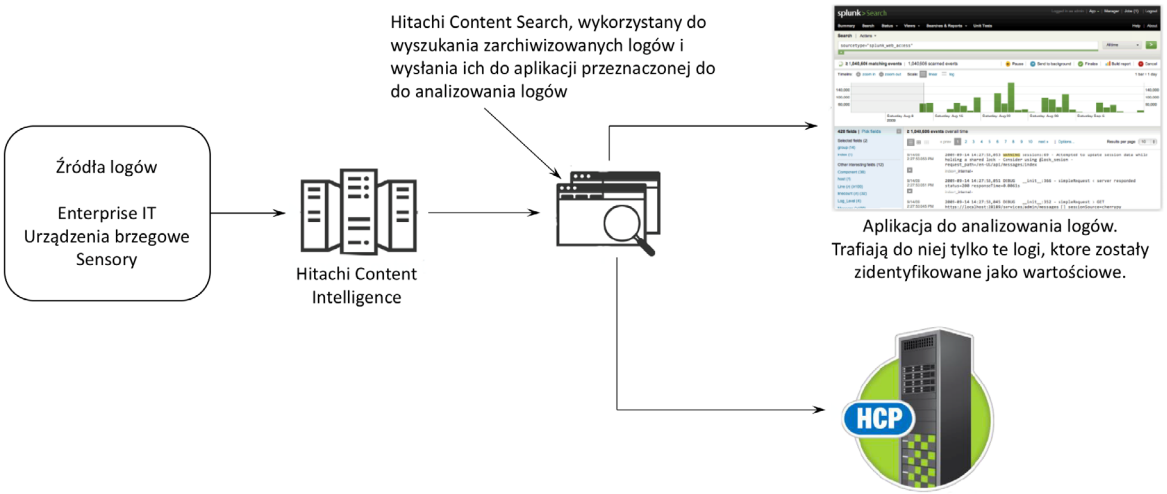
   Pliki zawierające dane osobowe są indeksowane i dodatkowo mogą być oznaczone flagą (np. RODO=1) dopisaną do ich metadanych. HCI pozwala również na dopisanie znalezionych wewnątrz pliku wartości danych osobowych do metadanych tego pliku (w postaci nowych pól XML). Plik wraz z uzupełnionymi metadanymi może być zapisany i przechowywany w obiektowym magazynie danych z gwarancją ich niezmienności i nieusuwalności (np. Hitachi Content Platform). W ten sposób możliwe jest przygotowanie się na jedno z podstawowych wymagań Rozporządzenia, mianowicie wyszukania w systemach komputerowych przedsiębiorstwa danych osobowych obywatela na jego żądanie, w wymaganym przez to Rozporządzenie czasie.

2. **MIFID2 i inne regulacje prawne – HCI** umożliwia przeszukiwanie źródeł danych w oparciu o określone wzorce. Po zdefiniowaniu wzorca, który będzie zawierał informacje o tym, jakie wrażliwe dane (różne dla różnych regulacji prawnych) powinny być wyszukiwane wewnątrz podłączonego do HCI źródła, możliwe jest uruchomienie procesu, który albo w sposób ciągły, albo zgodnie z określonym harmonogramem będzie poszukiwał nowych plików lub rekordów, które potencjalnie mogą zawierać określone przez zdefiniowany wzorzec dane wrażliwe, indeksował je, oznaczał i ewentualnie zapisywał w obiektowym magazynie danych (np. Hitachi Content Platform).

3. **Możliwość odczytywania i zapisywania wiadomości w kolejce Apache Kafka, które są wymieniane pomiędzy aplikacjami – HCI** pozwala na czytanie i zapisywanie danych do strumienia kolejki z wiadomościami, które są wymieniane pomiędzy aplikacjami i systemami.

Analiza

Celem tego typu scenariuszy może być oddzielenie wartościowych informacji z logów różnych systemów i aplikacji od wszystkich pozostałych (od tzw. „szumów"). Logi spełniające określone kryteria mogą być wysłane do aplikacji, która będzie wykonywała ich analizę. Pozostałe logi, te które nie zawierają żadnych wartościowych informacji są odkładane w repozytorium WORM z gwarancją ich niezmienności (np. Hitachi Content Platform).



Przykład analizy treści.

Przykłady:

1. **Filtrowanie logów i wysyłanie tylko tych wartościowych do aplikacji, która zajmuje się ich analizą – HCI** pozwala na uruchomienie zautomatyzowanego procesu, który w oparciu o określone wzorce, metadane lub konkretną zawartość logów będzie interpretował ich wartość. Dzięki temu możliwe jest ograniczenie zakresu i ilości danych, które należy przeanalizować, a to może wpływać na redukcję kosztów licencji oprogramowania do analizy logów.

```
172.18.10.106,admin,[20/Jul/2016:11:30:40],"POST /auth/oauth/ HTTP/1.1",200
172.18.10.106,admin,[20/Jul/2016:11:30:40],"GET /api/admin/setup HTTP/1.1",200
172.18.10.106,admin,[20/Jul/2016:11:30:40],"GET /api/admin/alerts HTTP/1.1",200
172.18.10.106,admin,[20/Jul/2016:11:30:53],"POST /api/admin/pipelines HTTP/1.1",201
172.18.10.106,jsmith,[20/Jul/2016:11:30:53],"GET /api/admin/pipelines/d7dc655f-e42b-4e2e-a48d-72d49beda939
HTTP/1.1",403
```

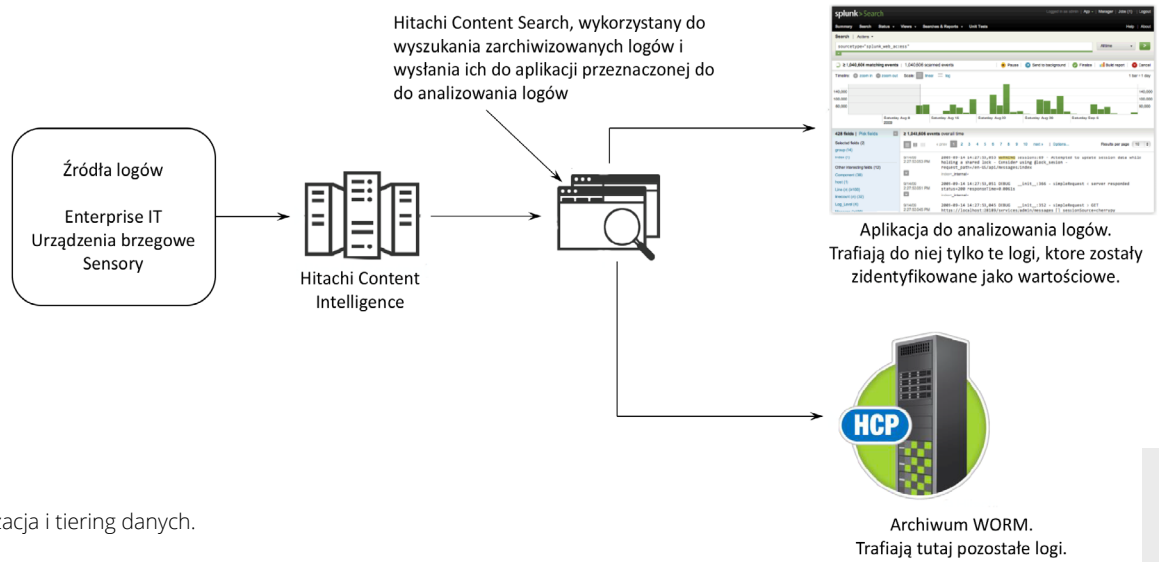
2. **Analizowanie zawartości logów – HCI** pozwala na wykorzystanie procesów wewnętrznych (Field Parser oraz Read Lines) do ekstrakcji pól z plików CSV lub z logów, w których każda linia reprezentuje inne zdarzenie.

   Mając na przykład log jak powyżej *(zob. grafika)* możliwa jest ekstrakcja każdej informacji oddzielonej przecinkiem, potraktowanie jej jako oddzielnego pola w dokumencie oraz zaindeksowanie. Dzięki temu administrator może np. szukać informacji zadając następujące zapytania:

   - pokaż wszystkie akcje, wykonane przez użytkownika admin w okresie od czerwiec 2016 do lipiec 2016
   - pokaż wszystkie zdarzenia, w których dany użytkownik otrzymał odmowę dostępu

Optymalizacja IT i tiering danych

W przypadku tych scenariuszy przesuwamy dane (tiering) pomiędzy różnymi klasami pamięci masowej (różnymi źródłami). Celem jest optymalizacja kosztów, zwiększenie wydajności systemów, poprawienie jakości dostępu do danych przy jednoczesnym zapewnieniu zgodności z regulacjami oraz wymaganiami biznesu dot. ciągłości pracy, retencji danych oraz innych zdefiniowanych SLO.



Optymalizacja i tiering danych.

Przykłady:

1. **Archiwizacja danych z klastra Hadoop do obiektowego magazynu danych Hitachi Content Platform.** Strumienie danych, które nie są już wykorzystywane w procesach MapReduce mogą zostać zapisane do pliku i przeniesione do obiektowego magazynu danych. Zasoby klastra Hadoop mogą być w ten sposób ponownie wykorzystane do realizacji bieżących procesów.
2. **Identyfikacja danych**, które mogą być przeniesione do chmury publicznej, oznaczenie tych danych, zaszyfrowanie ich (jeżeli jest to wymagane) oraz ich migracja do chmury publicznej (MS Azure, AWS, GCP).

CHCESZ WIEDZIEĆ WIĘCEJ? ZAPRASZAM DO KONTAKTU

S4E S.A.

ul. Samuela Lindego 1 C,

30-148 Kraków

www.s4e.pl

HITACHI
Inspire the Next

S4E