Duże zbiory danych
big data

dr hab. inż. Przemysław Korytkowski, prof. ZUT

Wykład 4

1

- Wczesne wersje Hadoopa zakładały wykorzystanie klastrów HDFS i MapReduce przez grupę współpracujących użytkowników w bezpiecznym środowisku.
- Środki dotyczące ograniczenia dostępu zostały opracowane w celu zapobieżenia przypadkowej utracie danych, a nie w celu zapobieżenia nieautoryzowanemu dostępowi do danych.
- Na przykład, system uprawnień do plików w HDFS zapobiega przed przypadkowym wymazaniem całego systemu plików z powodu błędu w programie, lub przez błędne wpisanie `hadoop fs -rmr /`.
- Nie zapobiega złośliwemu użytkownikowi przejęcia tożsamości roota, aby uzyskać dostęp lub usunąć wszelkie dane w klastrze.

2

2

- Hadoop obsługuje dwa mechanizmy uwierzytelniania:
  - Prosty
  - kerberos.
- Prosty mechanizm, który jest domyślny, wykorzystuje UID klienta w celu określenia nazwy użytkownika, którą przekazuje Hadoopowi. W tym trybie, serwery Hadoop w pełni ufają swoim klientom. Ta domyślna wartość jest wystarczająca dla klastrów, w których każdy użytkownik, który może uzyskać dostęp do klastra, jest w pełni wiarygodny.

3

3

Zagrożenia bezp.

Usługi Hadoop nie uwierzytelniają użytkowników ani innych usług. W związku z tym Hadoop jest narażony na następujące zagrożenia dla bezpieczeństwa.

- Użytkownik ma dostęp do klastra HDFS lub MapReduce jak każdy inny użytkownik. Uniemożliwia to egzekwowanie kontroli dostępu w niepewnym środowisku. Na przykład, kontrola uprawnień do plików w HDFS może być łatwo obchodzona.
- Napastnik może zamaskować się jako usługa Hadoop. Na przykład, kod użytkownika uruchomiony na klastrze MapReduce może się zarejestrować jako nowy TaskTracker.

DataNodes nie egzekwuje żadnej kontroli dostępu do swoich bloków danych. Dzięki temu nieautoryzowany klient może odczytać blok danych tak długo, jak długo może podać swój identyfikator bloku. Każdy może również zapisywać dowolne bloki danych do DataNodes.

4

4

CIA - confidentiality, integrity, availability

Teoria bezpieczeństwa informacji - model CIA:

- **Poufność** – jest zasadą bezpieczeństwa skupiającą się na założeniu, że informacje są dostępne tylko dla zamierzonych odbiorców.
- **Integralność** – oznacza utrzymywanie i zapewnianie dokładności i kompletności danych w całym cyklu życia. Oznacza to, że dane nie mogą być modyfikowane w sposób nieautoryzowany lub niezauważalny.
- **Dostępność** – systemy o wysokiej dostępności mają na celu utrzymanie stałej dostępności, zapobiegając przerwom w świadczeniu usług spowodowanym przerwami w dostawie prądu, awariami sprzętu i modernizacją systemu. Zapewnienie dostępności wiąże się również z zapobieganiem atakom typu denial-of-service.

5

5

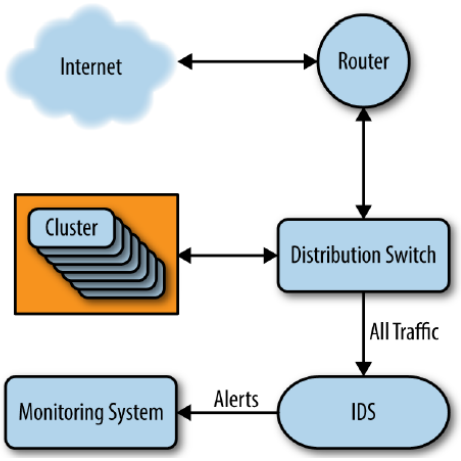
AAA – Authentication, Authorization, Accounting

- **Identyfikacja** to stwierdzenie, kto jest kimś lub czym coś jest. Zazwyczaj zgłoszenie jest w formie nazwy użytkownika.
- **Uwierzytelnienie** jest aktem weryfikacji stwierdzenia tożsamości. Wprowadzając prawidłowe hasło, użytkownik dostarcza dowód, że jest osobą, do której należy jego nazwa użytkownika.
- **Autoryzacja** - po udanej identyfikacji i uwierzytelnieniu osoby, programu lub komputera należy określić, do jakich zasobów informacyjnych mają dostęp i jakie działania będą mogły być wykonywane (uruchamiane, przeglądane, tworzone, usuwane lub zmieniane).
- **Audytowanie** – jest mechanizmem umożliwiającym śledzenie tego, co użytkownicy i usługi robią w klastrze. Jest to krytyczny element systemu bezpieczeństwa, ponieważ bez niego mogą wystąpić naruszenia bezpieczeństwa, których nikt nie zauważy, dostarcza zapis tego, co się stało.

6

6

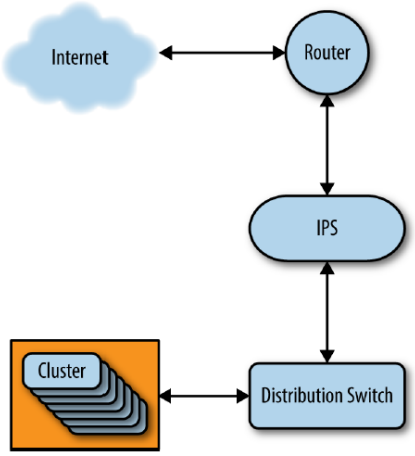
Intrusion Detection System (IDS)



7

7

Intrusion Prevention Systems (IPS)



8

8

Wymagania bezp.

1. Użytkownicy mogą uzyskać dostęp do plików HDFS tylko wtedy, gdy mają do nich uprawnienia.
2. Użytkownicy mogą uzyskać dostęp lub modyfikować tylko swoje zadania MapReduce.
3. Uwierzytelnianie użytkownika i usługi, aby zapobiec nieautoryzowanemu dostępowi do NameNodes, DataNodes, JobTrackers lub TaskTrackers.
4. Usługa do obsługi wzajemnego uwierzytelniania w celu uniemożliwienia nieautoryzowanym usługom dołączania do HDFS lub MapReduce klastra.
5. Nabywanie i korzystanie z uprawnień Kerberosa będzie przejrzyste dla użytkownika i aplikacji, pod warunkiem że system operacyjny nabył Kerberos Ticket Granting Tickets (TGT) dla użytkownika przy logowaniu.
6. Degradacja wydajności GridMix powinna wynosić nie więcej niż 3%.

9

9

Kerberos - the Network Authentication Protocol

- Z mitologii greckiej, wielogłowy pies, który pilnuje wejścia do Hadesu, aby nikt, kto wejdzie, nigdy nie wyszedł.
- Kerberos mechanizm uwierzytelniania opracowany w Massachusetts Institute of Technology (MIT). Kerberos stał się faktycznie standardem silnego uwierzytelniania dla systemów komputerowych dużych i małych.
- Został on zaprojektowany w celu zapewnienia silnego uwierzytelnienia dla aplikacji klienckich/serwerowych poprzez zastosowanie kryptografii tajnych kluczy.

10

10

Kerberos vs. SSL

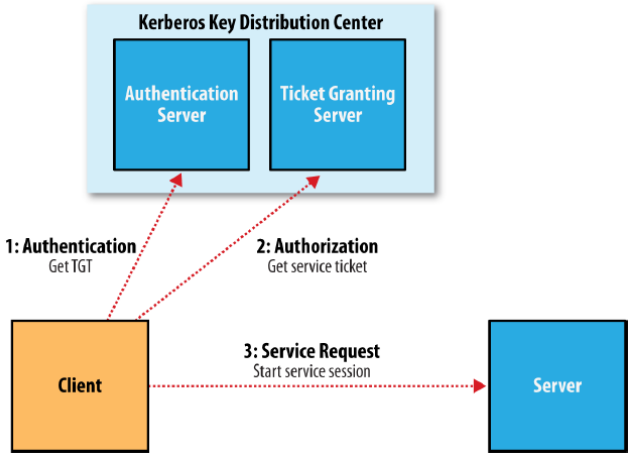
Kerberos wykorzystuje symetryczne klucze, które są o rząd wielkości szybsze niż operacje klucza publicznego używane przez SSL.

Prostsze zarządzanie użytkownikami. Na przykład, cofnięcie uprawnień użytkownika może być dokonane poprzez proste usunięcie użytkownika z centralnie zarządzanego KDC (centrum dystrybucji kluczy) Kerberos. Natomiast w przypadku SSL należy wygenerować nową listę cofnięć certyfikatów i rozesłać ją na wszystkie serwery.

11

11

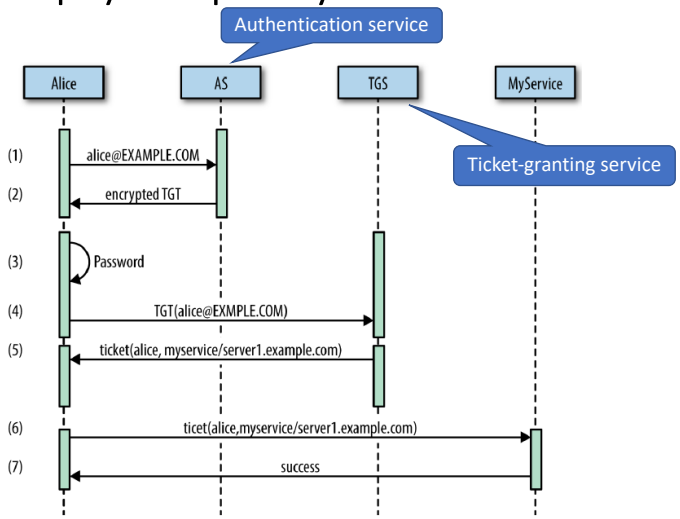
Protokół wymiany kluczy Kerberos'a



12

12

Przykład przepływu pracy Kerberosa



13

13

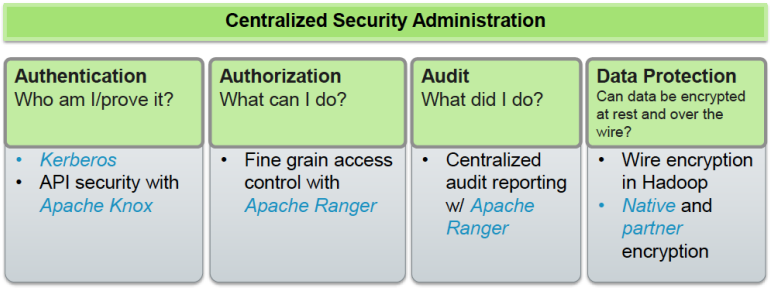
Założenia bezp.

- Hadoop nie wydaje poświadczeń użytkownika ani nie tworzy kont dla użytkowników. Hadoop zależy od danych logowania użytkownika zewnętrznego (np. login systemu operacyjnego, dane logowania Kerberos itp.). Oczekuje się, że użytkownicy uzyskają te poświadczenia od Kerberos podczas logowania do systemu operacyjnego. Usługi Hadoop powinny również być skonfigurowane przy użyciu odpowiednich poświadczeń, w zależności od konfiguracji klastra, w celu wzajemnego uwierzytelnienia.
- Każdy klaster jest konfigurowany niezależnie. Aby uzyskać dostęp do wielu klastrów, klient musi uwierzytelnić się w każdym klastrze osobno. Jednak pojedyncze logowanie, które uzyskuje token Kerberos, będzie działać na wszystkich klastrach.
- Użytkownicy nie będą mieli dostępu do kont root w klastrze lub na komputerach używanych do uruchamiania zadań.
- Komunikacja HDFS i MapReduce nie będzie działać w niezaufanych sieciach.
- Zadanie Hadoop będzie działać nie dłużej niż 7 dni (konfigurowalne) w klastrze MapReduce lub dostęp do HDFS z zadania zakończy się niepowodzeniem.
- Bilety Kerberos nie będą przechowywane w zadaniach MapReduce i nie będą dostępne dla zadań zadania. Dostęp do HDFS będzie autoryzowany za pomocą tokenów.

14

14

Technologie bezpieczeństwa w Hadoop



HORTONWORKS

15

15

Apache Ranger

- Kompleksowo zapewnienia bezpieczeństwo w całym ekosystemie Apache Hadoop.
- Firmy mogą wykonywać wiele zadań, w środowisku wielu najemców.
- Bezpieczeństwo danych w ramach Hadoop musi obsługiwać przypadki wielokrotnego wykorzystania dostępu do danych, jednocześnie zapewniając ramy dla centralnej administracji politykami bezpieczeństwa i monitorowania dostępu użytkowników.

16

16

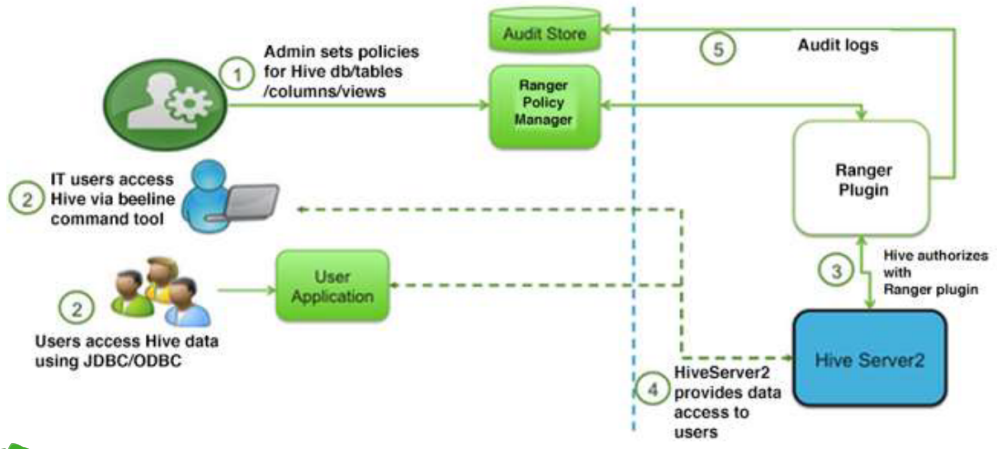
Apache Ranger

- Scentralizowana administracja bezpieczeństwa w celu zarządzania wszystkimi zadaniami związanymi z bezpieczeństwem w centralnym UI lub przy użyciu interfejsów REST API.
- Drobiazgowe uprawnienia do wykonywania określonych działań i/lub operacji za pomocą komponentu/narzędzia Hadoop i zarządzanie nimi za pomocą centralnego narzędzia administracyjnego
- Jedna metoda autoryzacji we wszystkich komponentach Hadoopa.
- Wsparcie dla różnych metod autoryzacji - kontrola dostępu oparta na rolach, kontrola dostępu oparta na atrybutach itp.
- Centralizacja kontroli dostępu użytkowników i działań administracyjnych (związanych z bezpieczeństwem) we wszystkich komponentach Hadoopa.

17

17

Apache Ranger

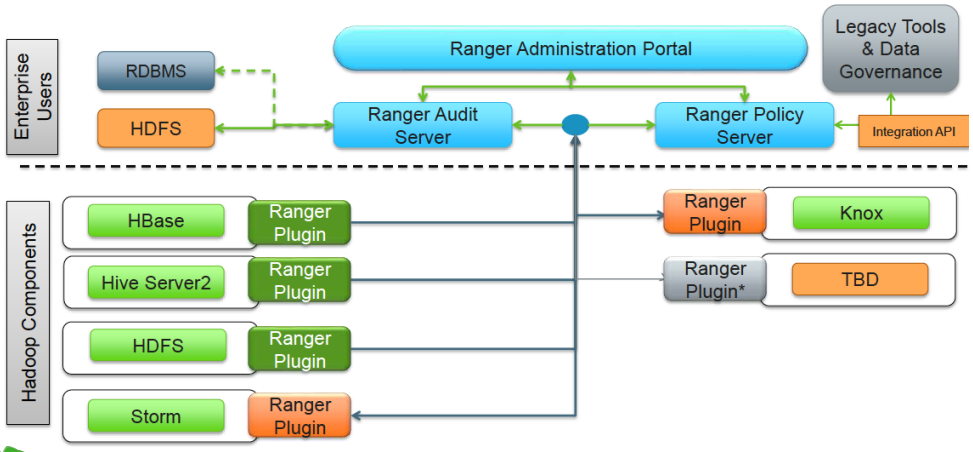


HORTONWORKS

18

18

Apache Ranger

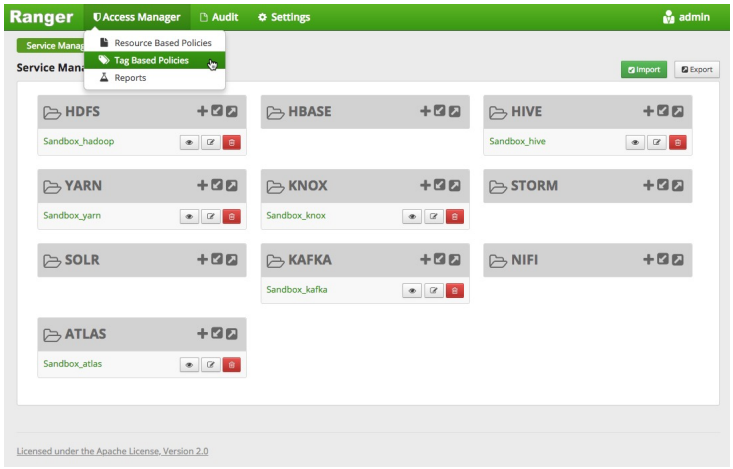


HORTONWORKS

19

19

Apache Ranger



HORTONWORKS

20

20

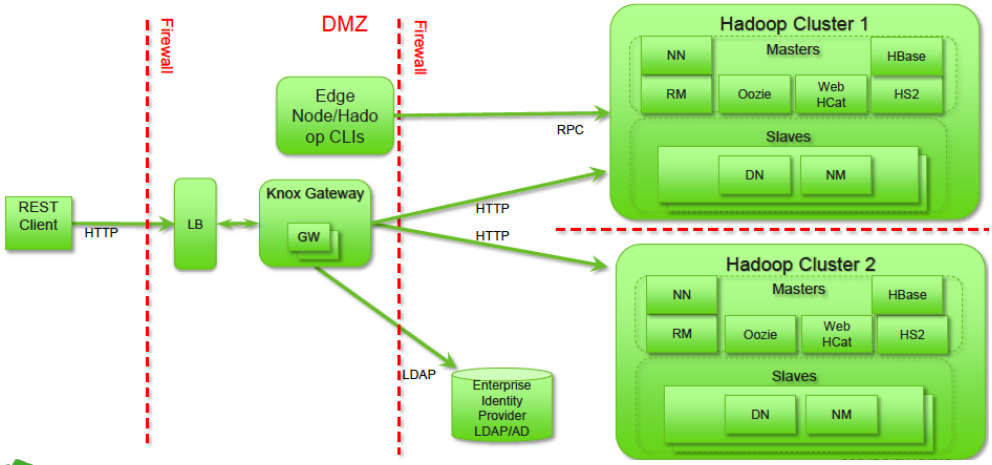
APACHE KNOX

- Bramka API Knox została zaprojektowana jako odwrotny serwer pośredniczący (proxy) na potrzeby zapewnienia bezpieczeństwa.
- Egzekwowanie polityki sięga od uwierzytelniania/federacji, autoryzacji, audytu, rozsyłki, mapowania.
- Klaster jest zdefiniowana w ramach deskryptora topologii i pozwala bramkom Knox na routing i translację pomiędzy adresami URL użytkowników i wewnętrznymi zasobami klastra.
- Każdy klaster Apache Hadoop, który jest chroniony przez Knoxa, ma swój zestaw REST API reprezentowany przez specyficzną ścieżkę kontekstową aplikacji. Pozwala to Knox zarówno chronić wiele klastrów, jak i prezentować konsumentowi REST API z jednym punktem dostępu do wszystkich wymaganych usług, w obrębie wielu klastrów.

21

21

APACHE KNOX

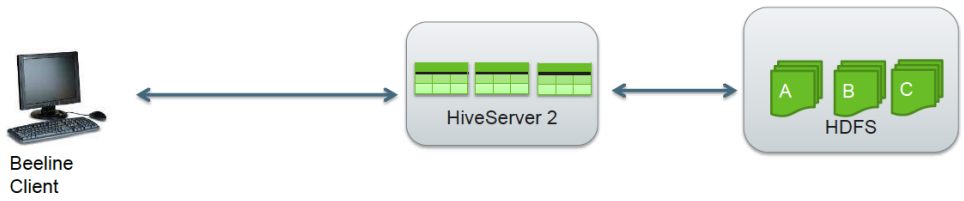


HORTONWORKS

22

22

Dostęp do Hive z CLI Beeline

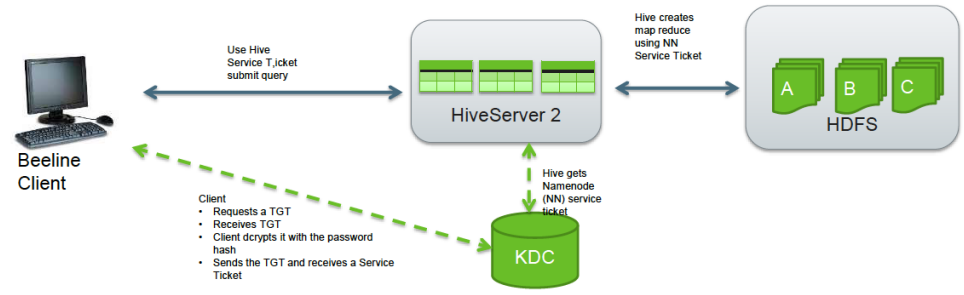


HORTONWORKS

23

23

Autoryzacja z Kerberos

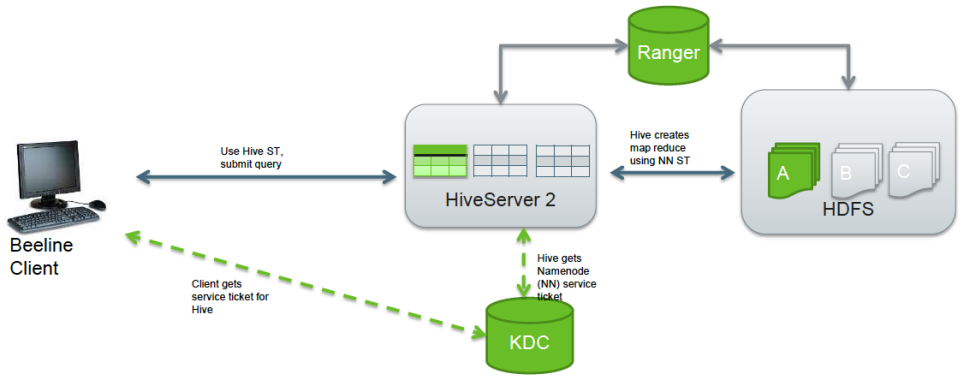


HORTONWORKS

24

24

Autoryzajca z Ranger

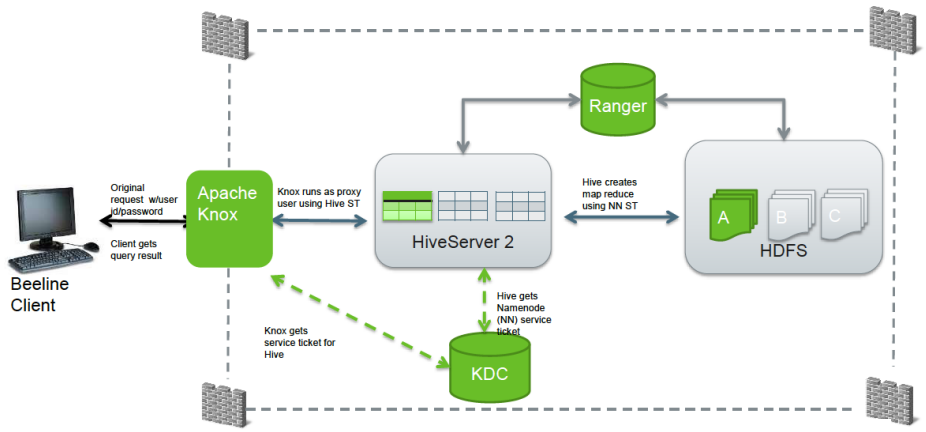


HORTONWORKS

25

25

Dostęp przez Knox



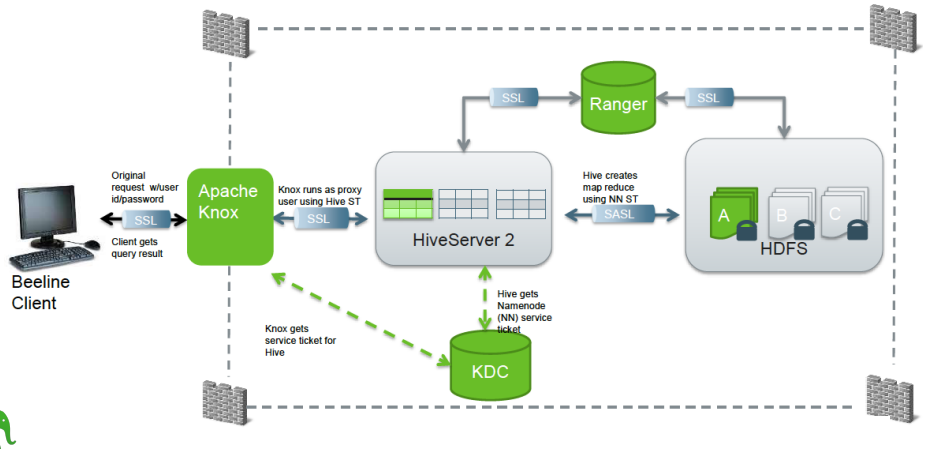
HORTONWORKS

26

26

+ SSL + szyfrowanie danych

Bezpieczeństwo danych statycznych i w ruchu.



HORTONWORKS

27

27

Ekosystem

28

28

Dziękuję za uwagę!

29