

Big Data

# Organizacyjnie

Prowadzący:

dr Mariusz Rafało

[mrafalo@sggw.edu.pl](mailto:mrafalo@sggw.edu.pl)

<http://mariuszrafalo.pl> (hasło: BIG)

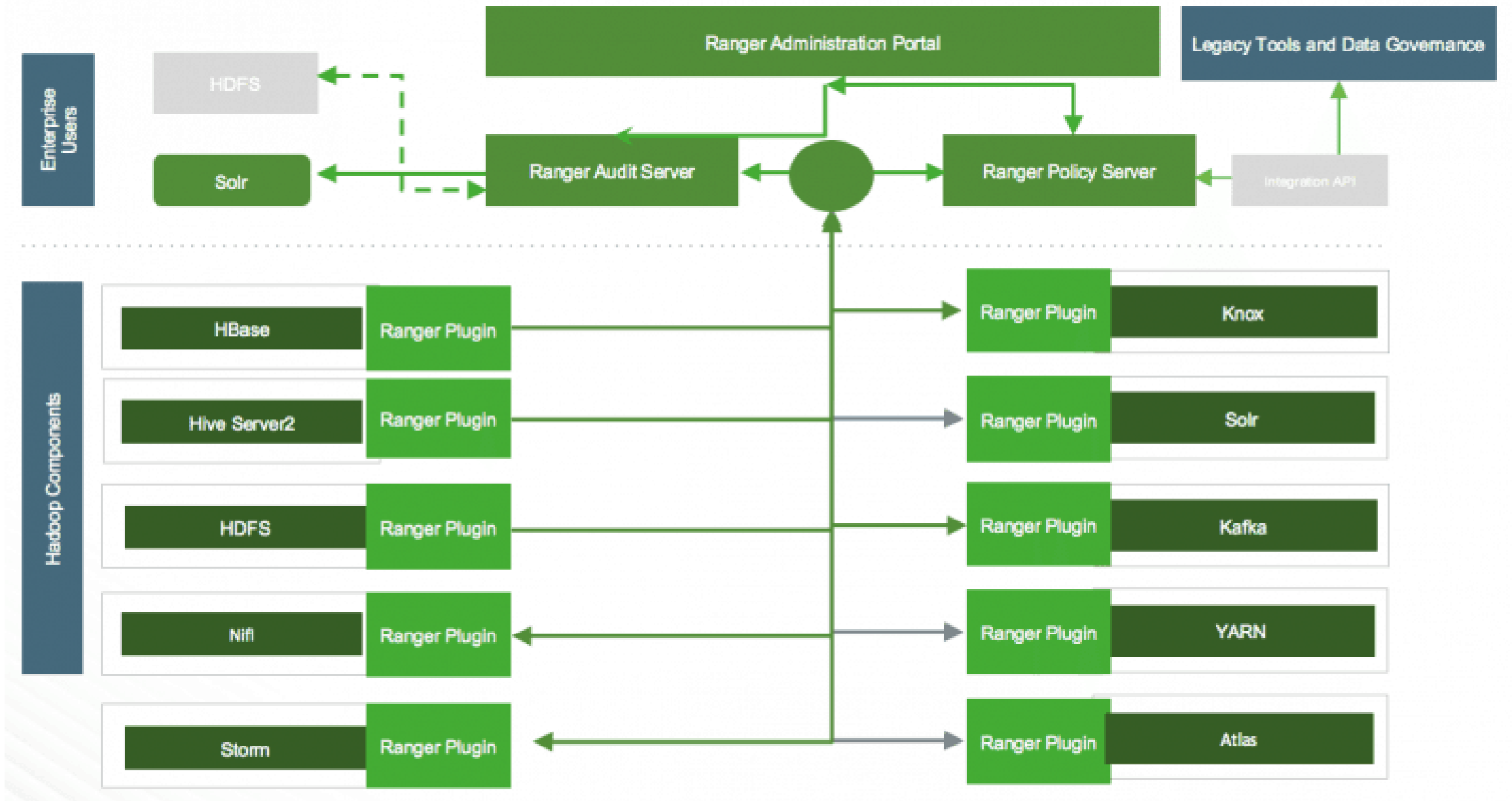
# Bezpieczeństwo

# Wyzwania

- Przetwarzanie danych w oparciu o technologie obejmuje wiele technologii (Spark, Kafka, Storm, Hive, Hbase, itp.), co sprawia że proces przepływu danych jest złożony
- Dane w systemie big data są udostępniane wielu interesariuszom, pojawia się kwestia zapewnienia dostępu na poziomie bazy danych, tabeli, zakresu danych
- System big data przechowuje informacje o wymogu wysokiej dostępności, pojawia się kwestia architektury *disaster recovery*

# Ranger

- Autentykacja użytkowników
- Autoryzacja dostępu do zasobów klastra
- Standaryzacja autoryzacji w ramach poszczególnych komponentów klastra
- Audytowanie dostępu do danych
- Centralizacja uprawnień w jednym repozytorium



# Ranger – dobre praktyki

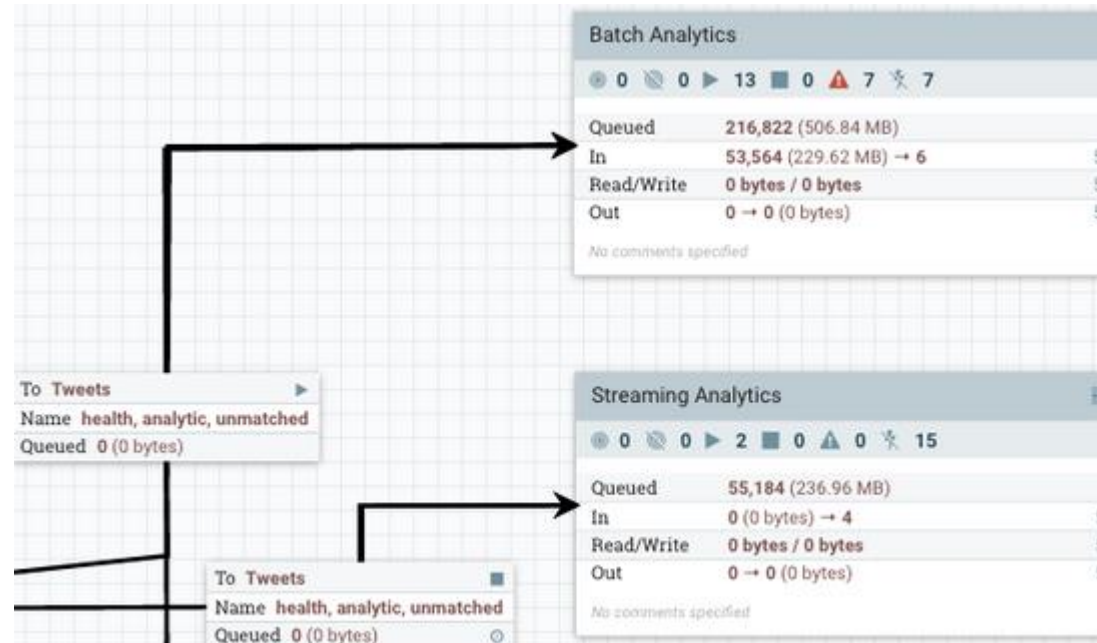
- Odebranie uprawnień dla użytkowników na poziomie OS (*chmod 700*)
- Wszystkie usługi klastra powinny być dostępne wyłącznie poprzez Ranger
- Integracja z LDAP
- Autoryzacja Kerberos: minimalne uprawnienia dla plików z kluczami

Automatyzacja



# Zarządzanie procesem przepływu danych

- Oozie
- Apache NIFI
- Apache Beam

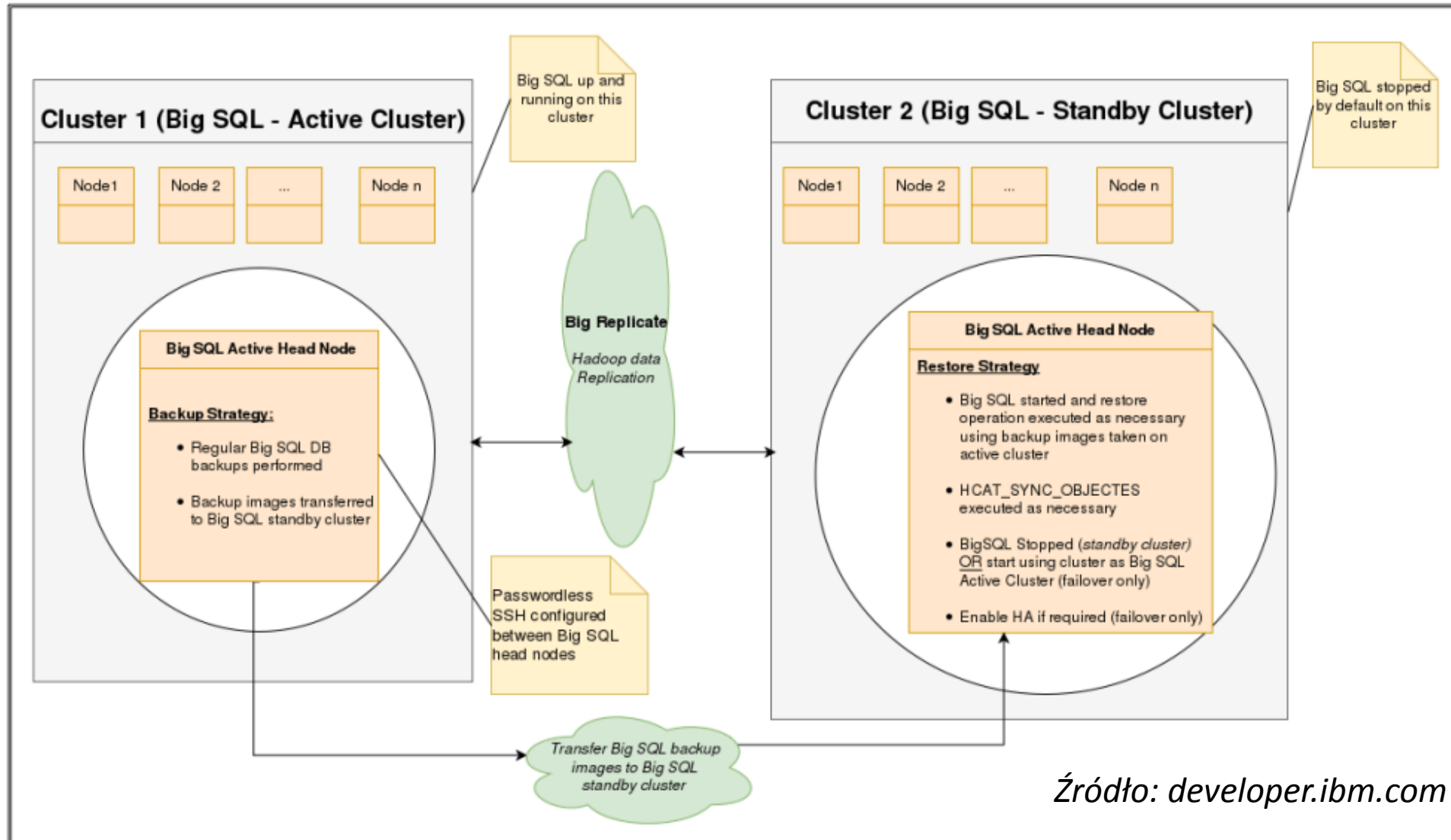


# Zarządzanie dostępem do danych

- Integracja z SSO
- Integracja z Kerberos
- Sterowanie dostępem do danych (Hive, HDFS, Hbase, itp..) na poziomie:
  - Bazy danych
  - Kolumny
  - Wiersza
- Zastosowane technologie
  - Knox
  - Ranger

# Konceptcja środowiska DR

## IBM Big Replicate



ETYKA

# Etyka



# RODO

- Zakres regulacji
- Jakie dane znajdują się w repozytorium danych?
- Jak anonimizować dane?
- Jak chronić dane?
- Przetwarzanie danych osobowych a udzielone zgody
- Zastosowanie zaawansowanej analityki predykcyjnej

Dziękuję za uwagę