Big Data

Organizacyjnie

Prowadzący:

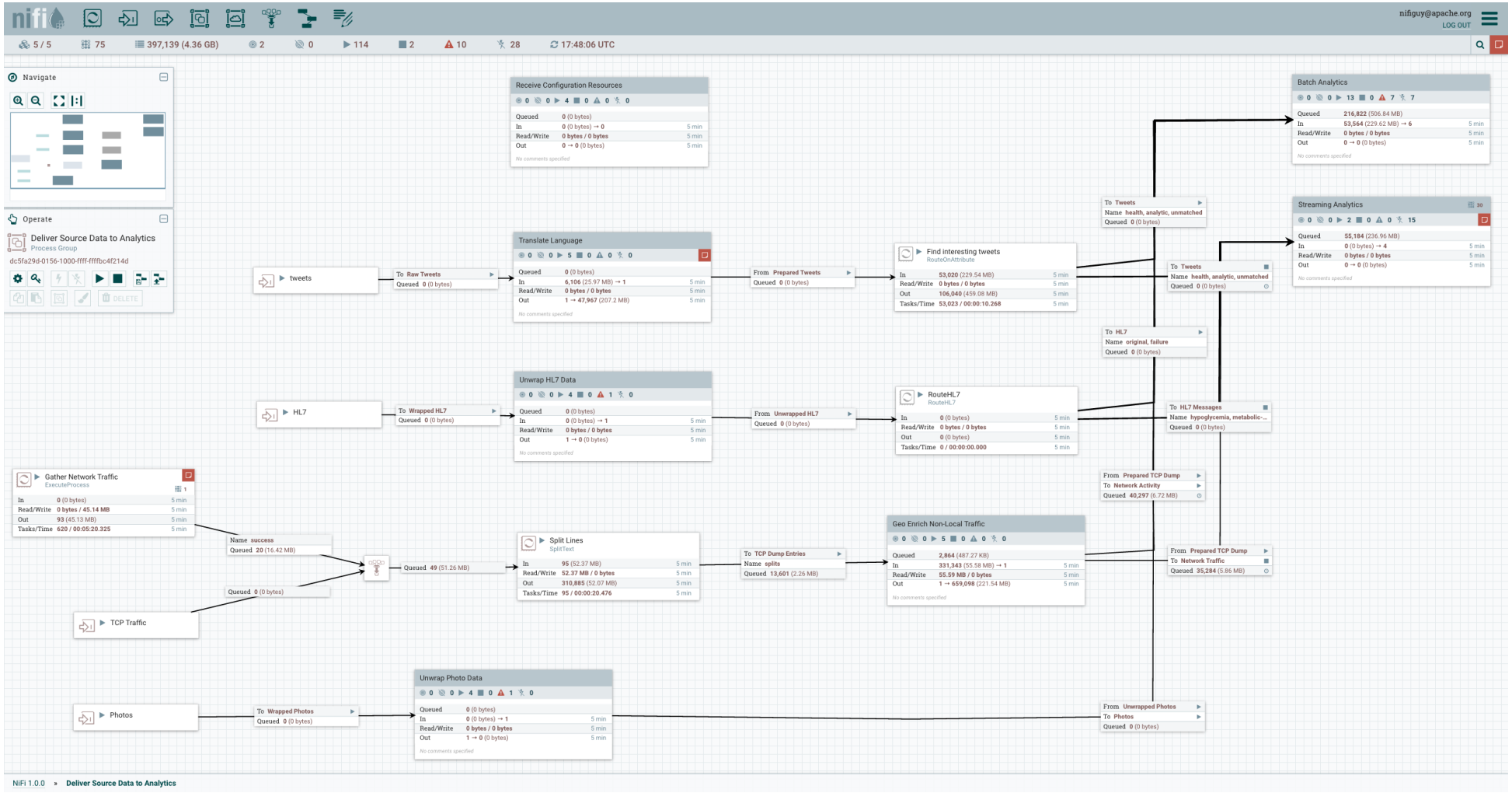
dr Mariusz Rafało

mrafalo@sgh.waw.pl

http://mariuszrafalo.pl (hasło: **BIG**)

Automatyzacja

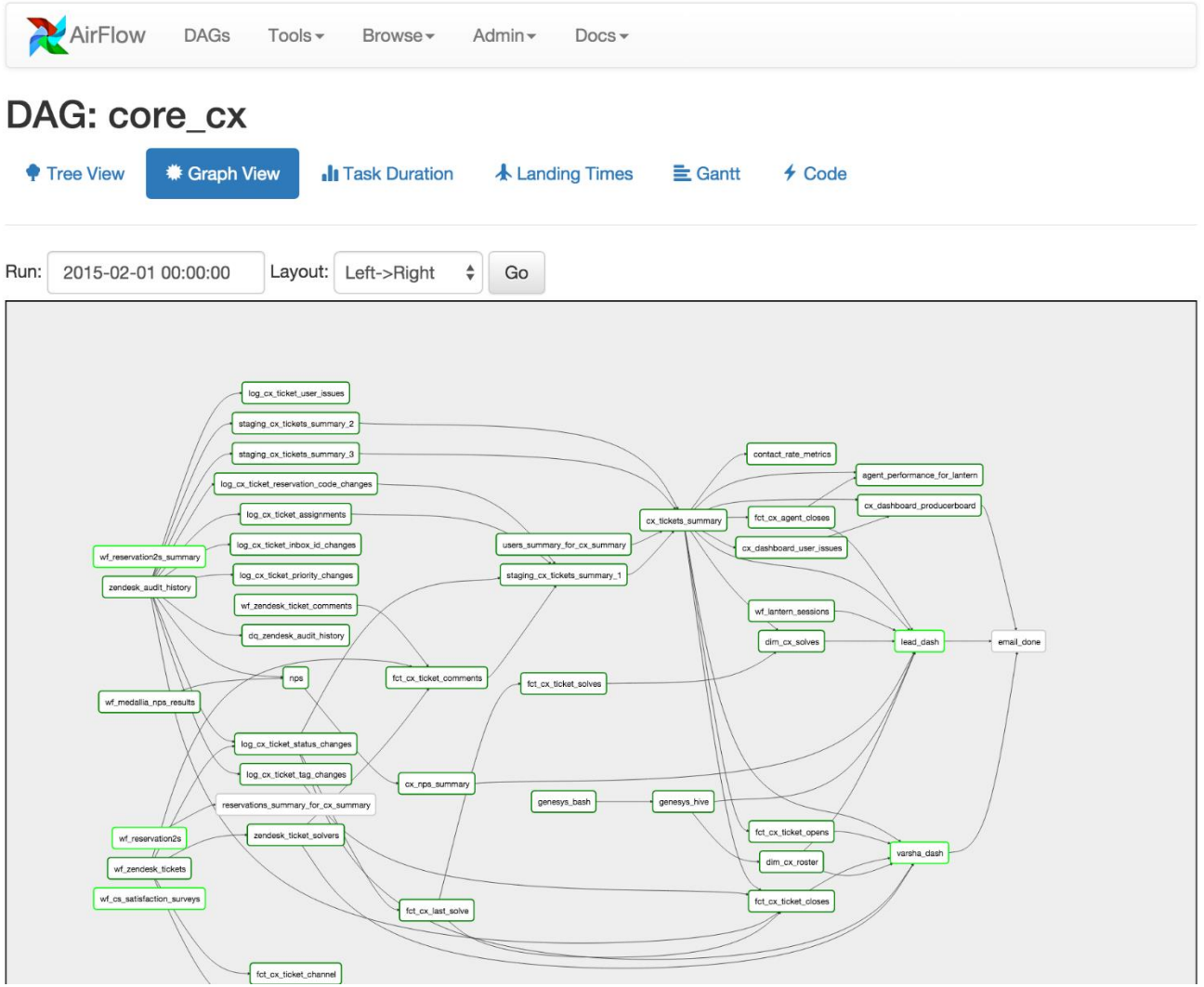
Automatyzacja przetwarzania: Apache NiFi



Źródło: nifi.apache.org

Automatyzacja przetwarzania: ETL

- Oozie
- Airflow
- Falcon
- SLJM
- CRON(sic!)



Źródło: airflow.apache.org

Technologie składowania danych

Hive

- Baza danych oparta na systemie plików HDFS
- Oprogramowanie powalające na zadawanie zapytań do rozproszonego systemu HDFS
- Oferuje język zapytań HQL; jest to język programowania o podobnej składni do SQL (HiveSQL)
- Dodatkowo, podobnie jak Pig, Hive udostępnia komendy pozwalające na stosowanie algorytmów MapReduce

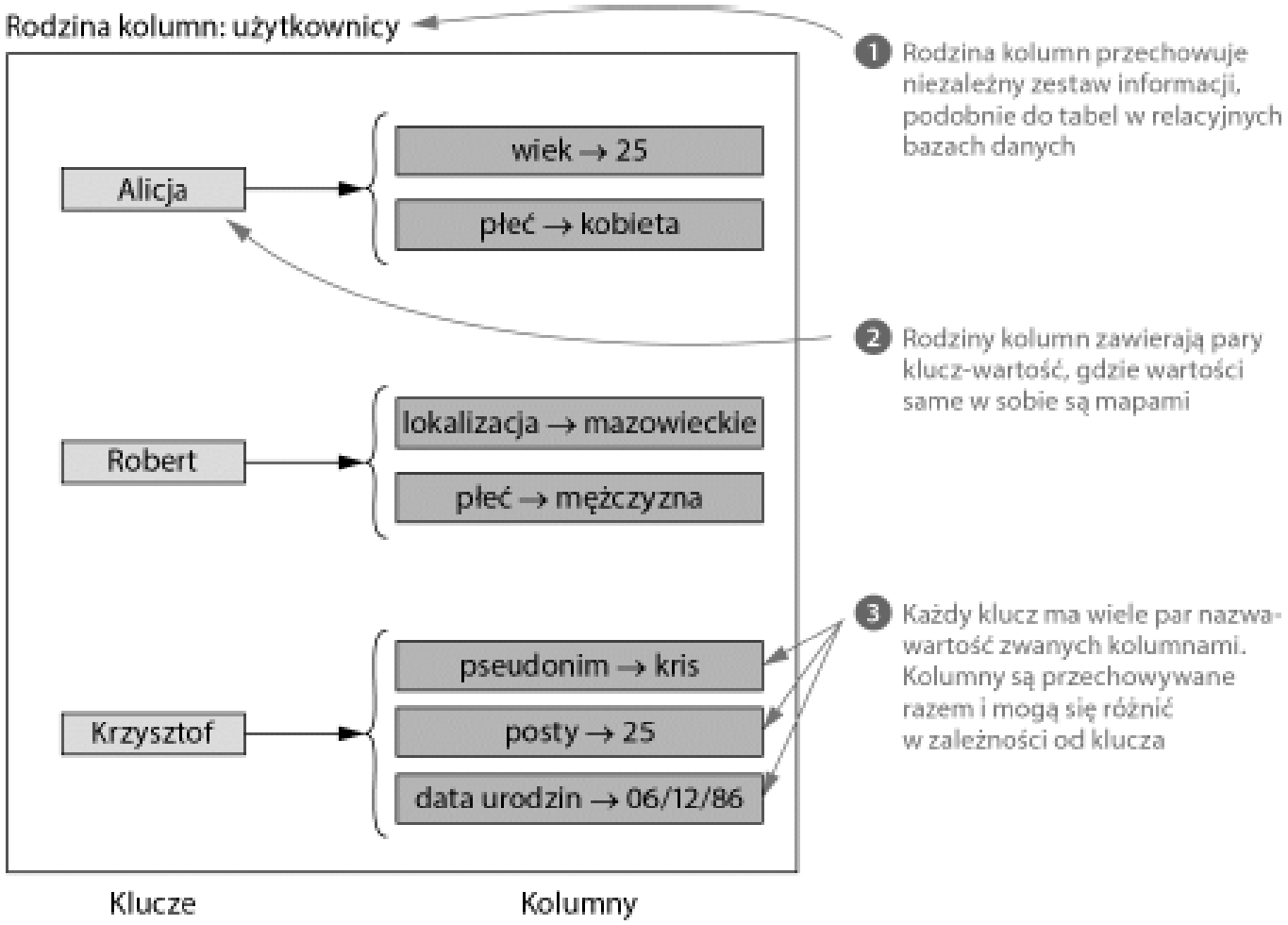
```
SELECT a.year, a.player_id, a.runs from batting a
JOIN (SELECT year, max(runs) runs FROM batting GROUP BY year ) b
ON (a.year = b.year AND a.runs = b.runs);
```

HBase

- Rozproszona, kolumnowa baza danych
- Dobrze sprawdza się do obsługi szybkich zapytań na tabelach zawierających miliardy rekordów i tysiące (nawet miliony) kolumn
- Jest to baza danych nie-relacyjna, obsługująca zapytania w pamięci operacyjnej
- Mniej radzi sobie z zapytaniami analitycznymi, które wymagają agregowania danych (np. zapytania o dane detaliczne będą mniej wydajne)

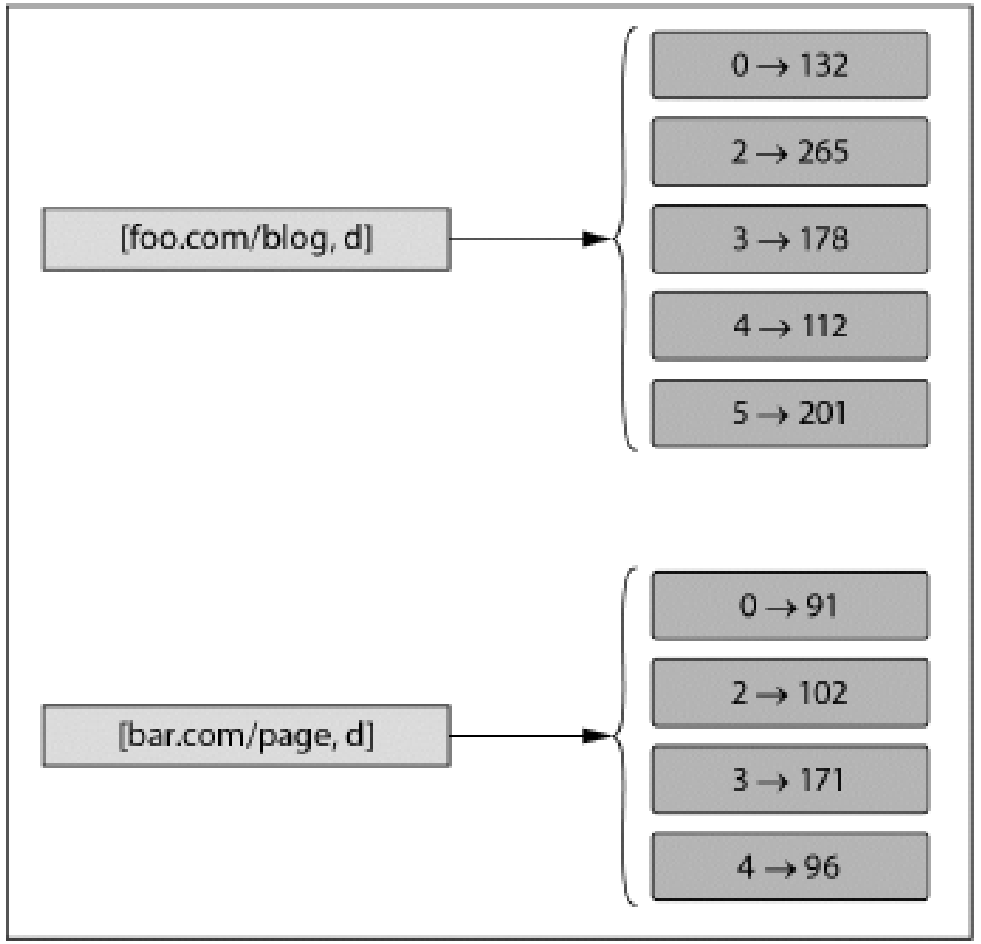
Cassandra

- Kolumnowa baza danych; cechuje się krótkimi czasami odpowiedzi
- Posiada cechy bazy klucz-wartość
- Operuje na rodzinach kolumn, kluczach i kolumnach



Źródło: Big Data. Najlepsze praktyki budowy skalowalnych systemów obsługi danych w czasie rzeczywistym, Nathan Marz, James Warren, Helion, 2016

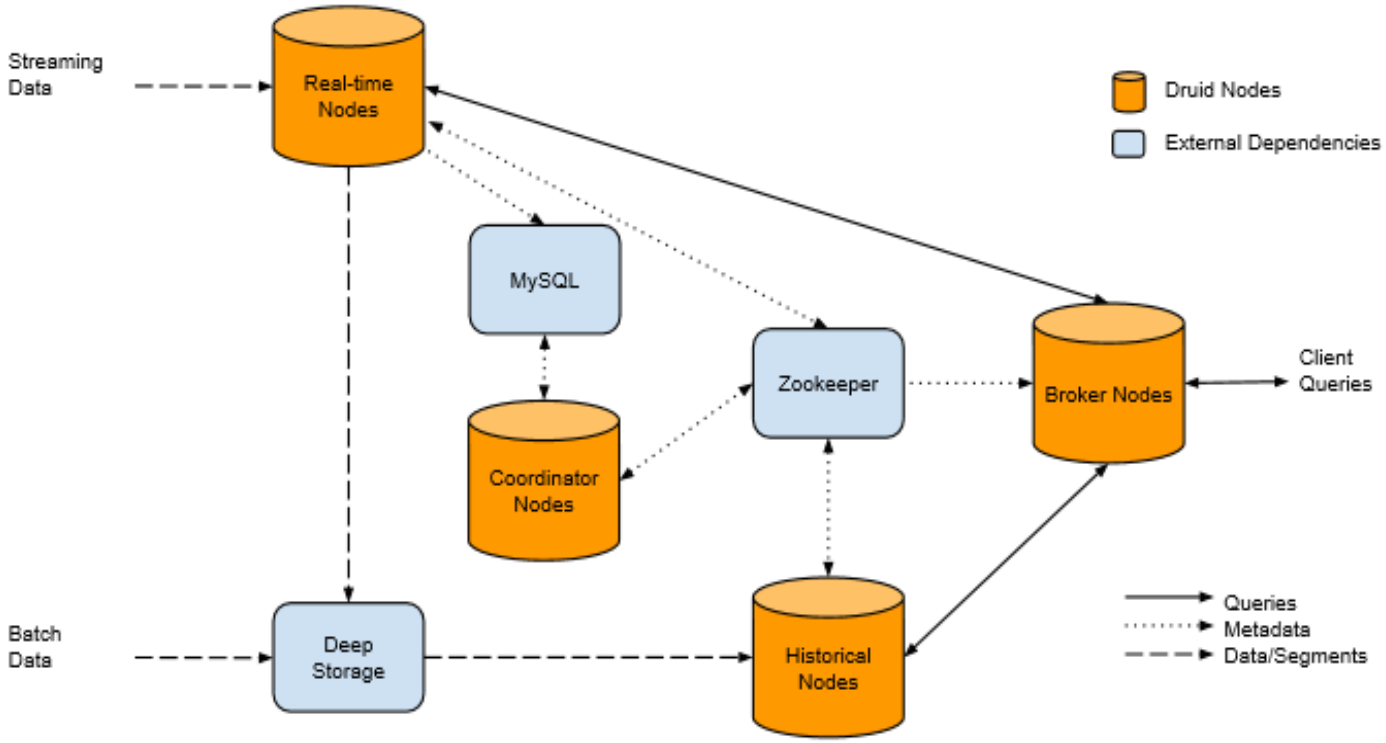
Cassandra(przykład)



Źródło: Big Data. Najlepsze praktyki budowy skalowalnych systemów obsługi danych w czasie rzeczywistym, Nathan Marz, James Warren, Helion, 2016

Druid

- Kolumnowa baza danych wspierająca analizy wielowymiarowe (OLAP)
- Możliwość ładowania danych za pomocą ETL lub poprzez strumień danych (Kafka)
- Rozproszona, skalowalna architektura



Źródło: druid.io

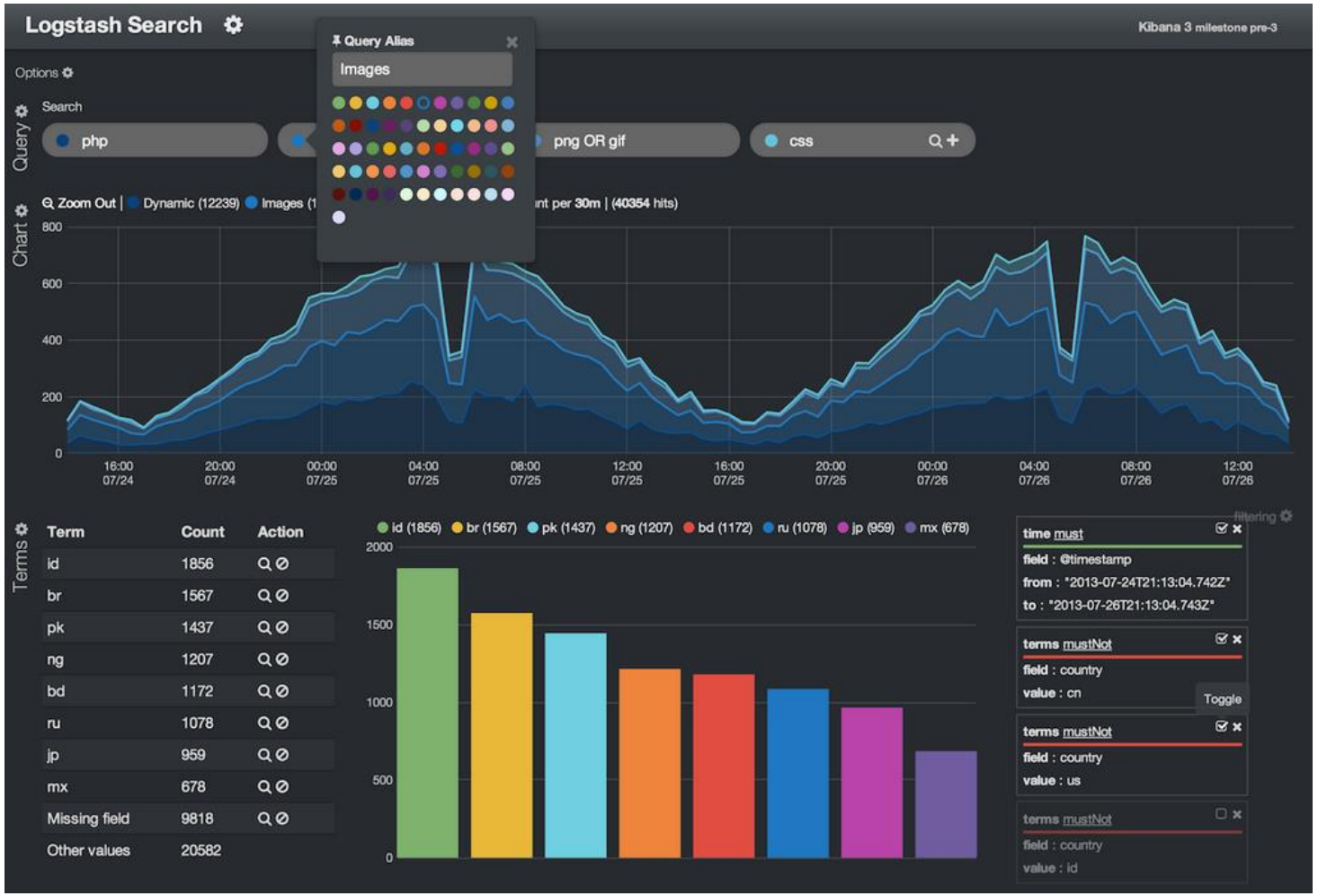
MongoDB

- Nierelacyjna baza danych, napisana w języku C++
- Platforma cechuje się dużą skalowalnością, wydajnością oraz brakiem ściśle zdefiniowanej struktury obsługiwanych danych
- Dane składowane są jako pliki w formacje JSON, co umożliwia aplikacjom bardziej naturalne ich przetwarzanie, przy zachowaniu możliwości tworzenia hierarchii oraz indeksowania
- Posiada możliwość składowania danych w pamięci operacyjnej (*WiredTiger engine*)
- Wybrane narzędzia i komponenty:
  - MongoDB Compass – narzędzie do wizualnej analizy danych
  - MongoDB Spark Connector – integracja z Apache Spark
  - MongoDB Atlas – database as a service

Redis

- Baza danych klasy in-memory, składująca dane w pamięci operacyjnej
- Oferuje bardzo szybki odczyt i zapis danych
- Oferuje wsparcie dla wielu typów danych jak np. string, hash, list, set
- Pozwala na stosowanie indeksów
- Posiada specjalizowane struktury służące obsłudze danych geograficznych

ELK



Źródło: elastic.co

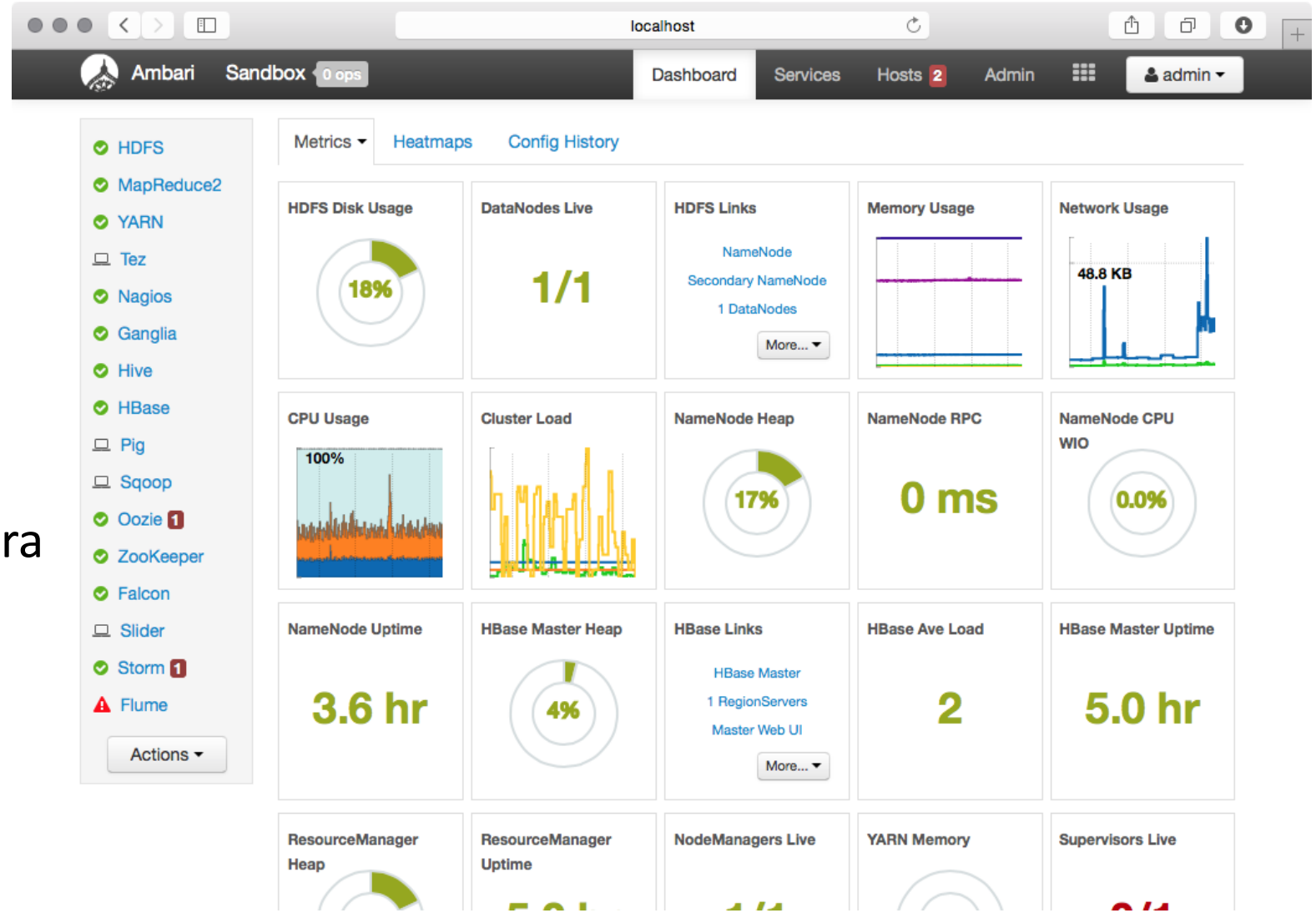
Zarządzanie klastrem

Zookeeper

- Scentralizowana usługa utrzymująca konfigurację klastra
- Zarządza metadanymi i konfiguracją klastra
- Zarządza konfiguracją rozproszoną, pozwala na automatyczne konfigurowanie nowego węzła (*data node*) lub nowej usługi na wielu węzłach

Ambari

- Konsola do zarządzania klastrem oraz wszystkimi usługami
- Monitoruje zarówno zasoby klastra jak i poszczególne usługi



Zarządzanie klastrem: zagadnienia

- Infrastruktura sprzętowa (jak monitorować?)
- Warstwa aplikacji (jak dbać o dostępność?)
- Przetwarzanie danych (jak ładować dane klaster?)
- Zarządzanie dostępem (bezpieczeństwo danych)
- Śledzenie i diagnostyka błędów w danych
- Wydajność

Dziękuję za uwagę