Hadoop i Spark

Mariusz Rafało
mrafalo@sgh.waw.pl
http://mariuszrafalo.pl

WPROWADZENIE DO EKOSYSTEMU APACHE HADOOP

Czym jest Hadoop

- Platforma służąca przetwarzaniu rozproszonemu dużych zbiorów danych
- Jest to system open-source, na licencji Apache Software Foundation

- Główne cechy Apache Hadoop:
  - Zarządzanie danymi w różnych formatach
  - Skalowalność
  - Tolerancja na awarie

- Wspierane języki programowania
  - Java/Python/Scala/R
  - SQL

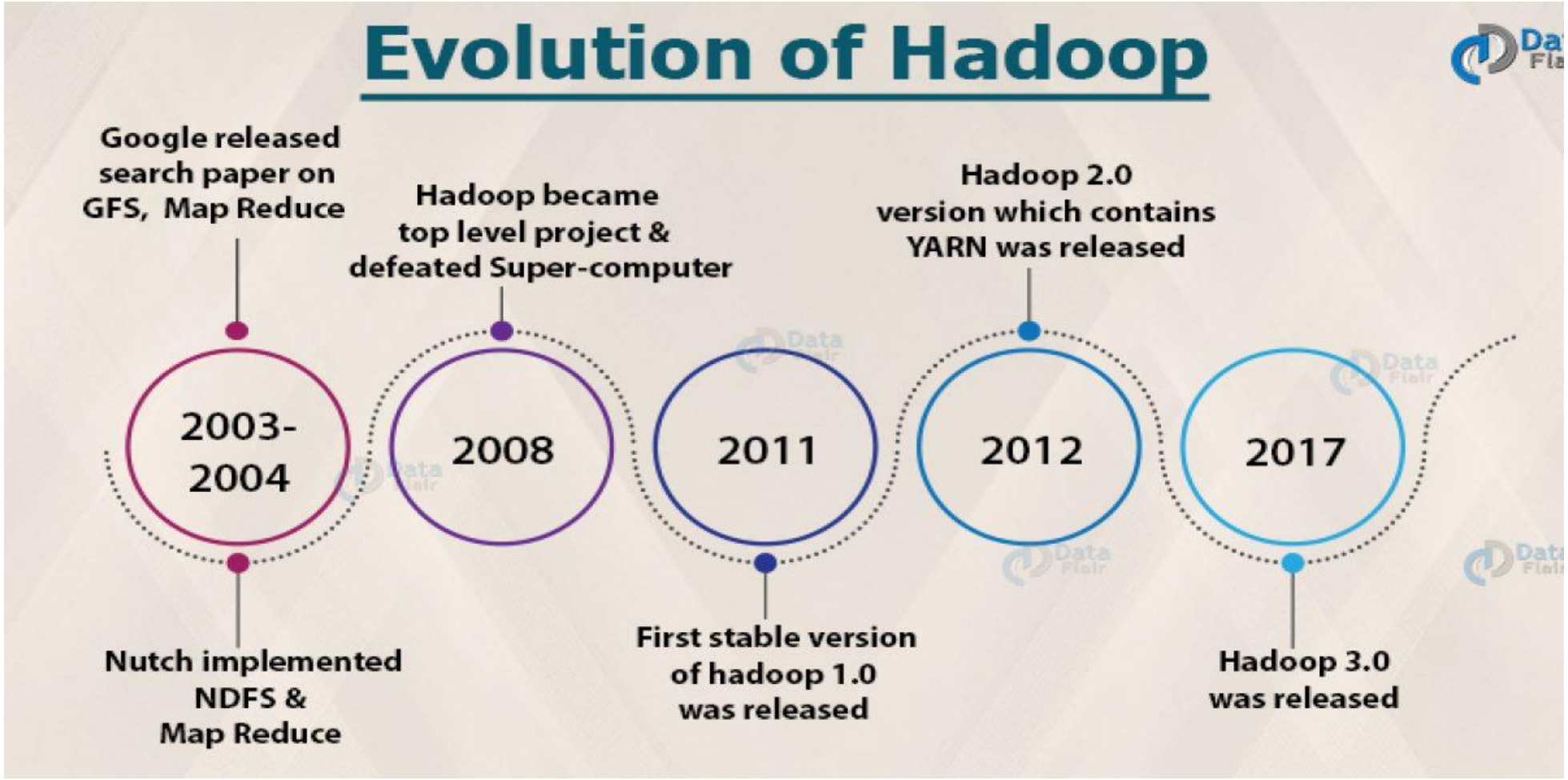
HDFS

- HDFS (*Hadoop Distributed File System*) to rozproszony system plików, umieszczony na wielu serwerach (węzłach – *nodes*)

- HDFS cechuje się wysokim poziomem tolerancji na awarie sprzętowe (*fault tolerant*)

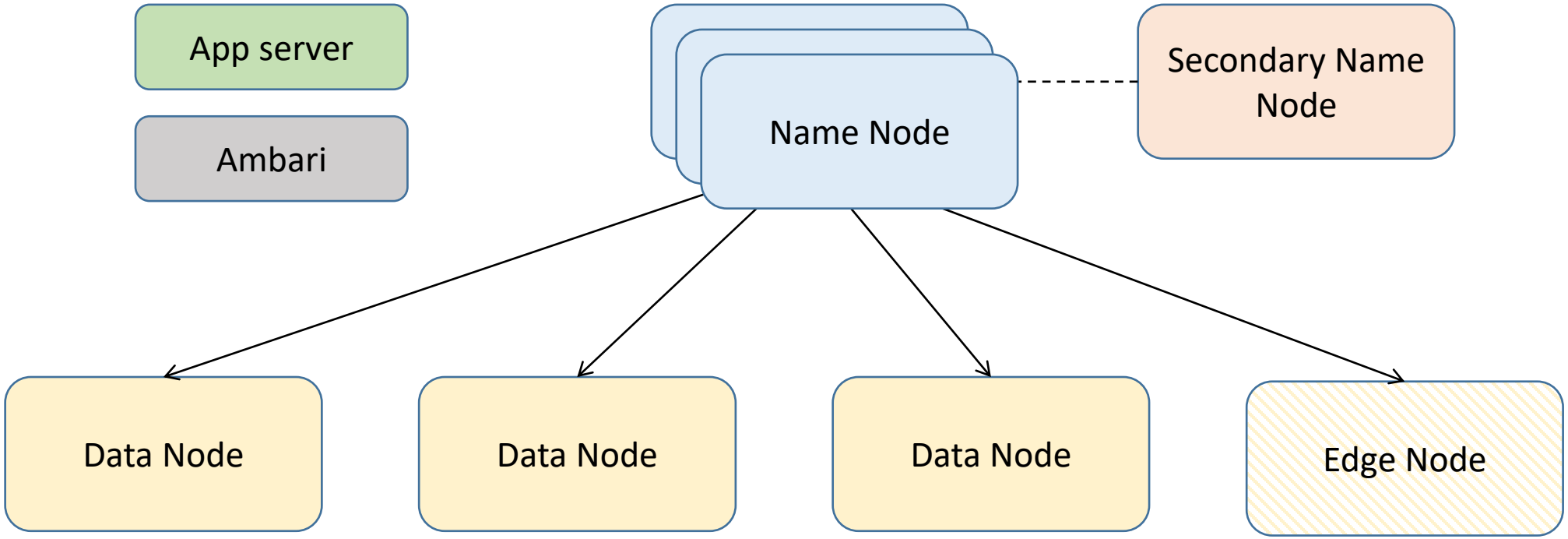
- HDFS powstał na potrzeby projektu wyszukiwarki Nutch, dla firmy Yahoo (Doug Cutting/Mike Cafarella in 2005)

Apache Hadoop



https://data-flair.training/blogs/hadoop-history/

Architektura logiczna klastra Hadoop



Name node

- Przechowuje metadane plików; także dane dotyczące lokalizacji poszczególnych plików składowanych na HDFS
- *name node* jest kluczowym elementem architektury fizycznej – w klastrze zawsze jest jeden *name node*
- Zarządzania rozkładem plików podczas przyłączania nowych *data node* oraz w przypadku wystąpienia awarii

Secondary name node

- Przechowuje logi replikowane w określonym czasie z *name node*
- Zadaniem *secondary name node* jest redukcja czasu zarządzania metadanymi klastra oraz czasu restartu klastra
- *secondary name node stanowi* zapisuje stany danych(*checkpoint*) w systemie HDFS; służy to wsparciu wydajności pracy *name node*
- *secondary name node* nie służy zapewnieniu wysokiej dostępności klastra (HA)

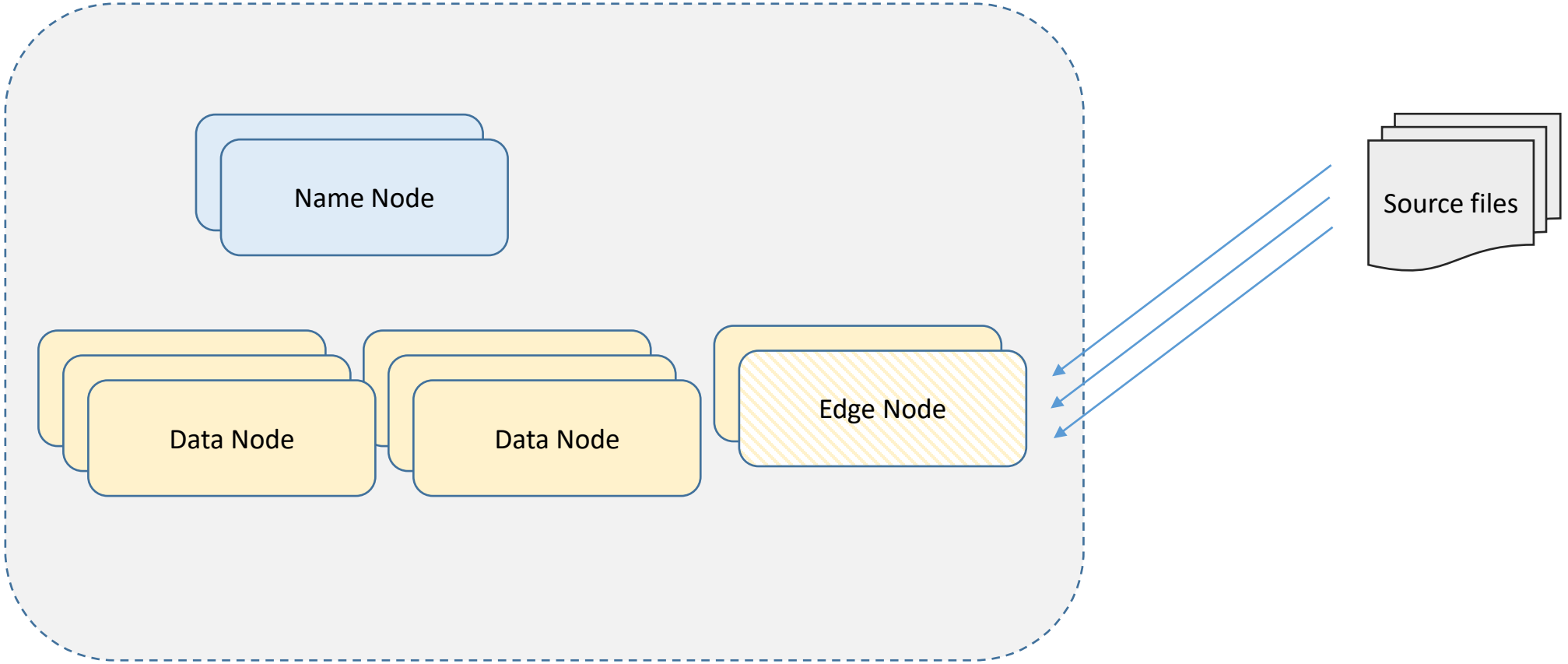
Data node

- Składuje dane na systemie HDFS
- Przekazuje informacje od *name node*, dotyczące swojego statusu oraz posiadanych plików
- Może pracować w trybie replikacji danych (także RAID)
- Wykonuje zadania obliczeniowe zlecane poprzez MapReduce lub Yarn

Edge node

- Służy do wymiany informacji z otoczeniem klastra
- Są na nim instalowane aplikacje
- Posiada dedykowaną konfigurację sieciową, pozwalającą na wymianę danych z innymi systemami

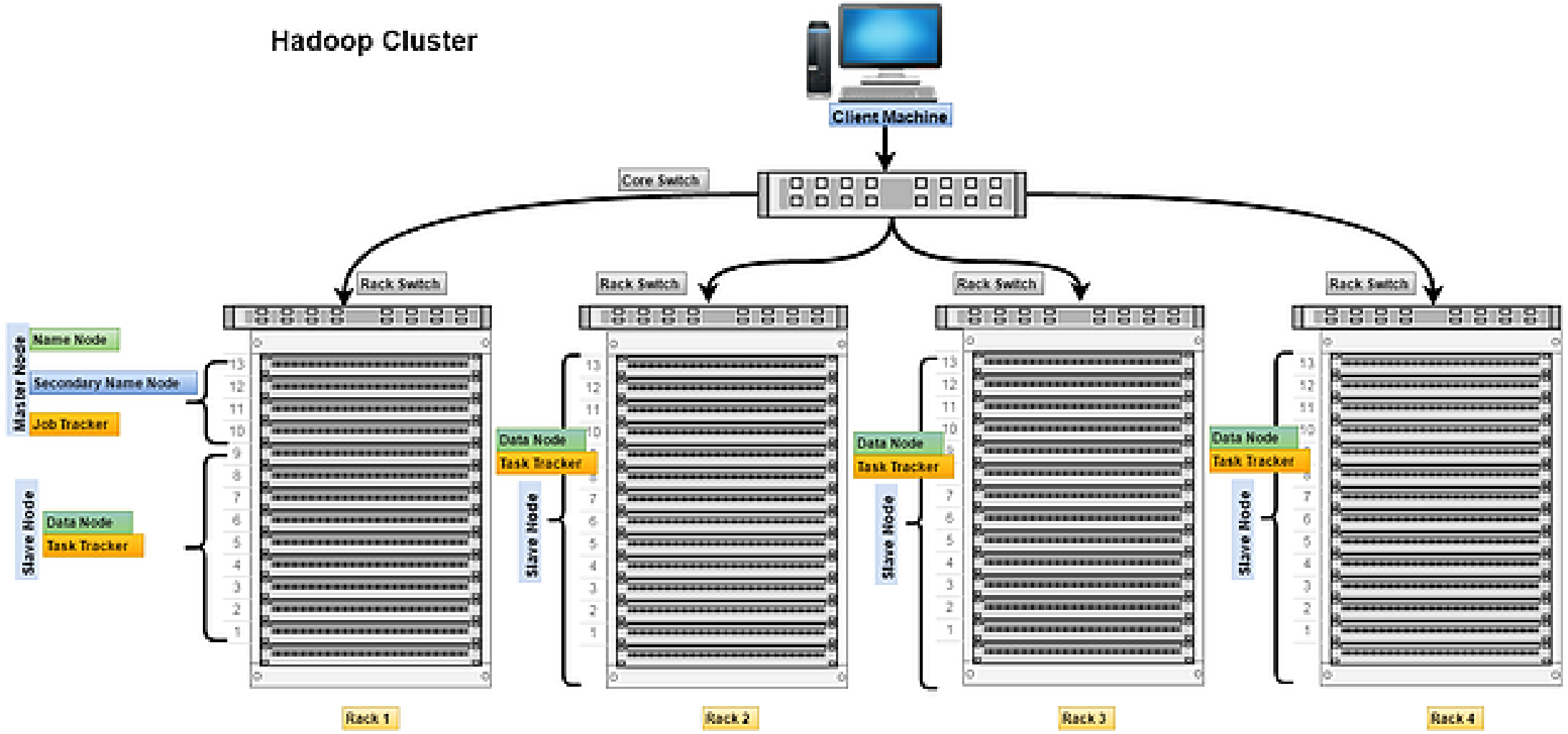
Edge node



Serwer aplikacyjny

- Hostuje aplikacje realizujące różne funkcjonalności
- Aplikacji nie instaluje się na węzłach klastra (*name node*/*data node*)
- Dobrą praktyką jest wirtualizacja poszczególnych aplikacji (np. ESXi)
- Przykładowe usługi:
  - Ambari
  - NiFi
  - Zeppelin/Jupyter
  - Oozie
  - Ambari

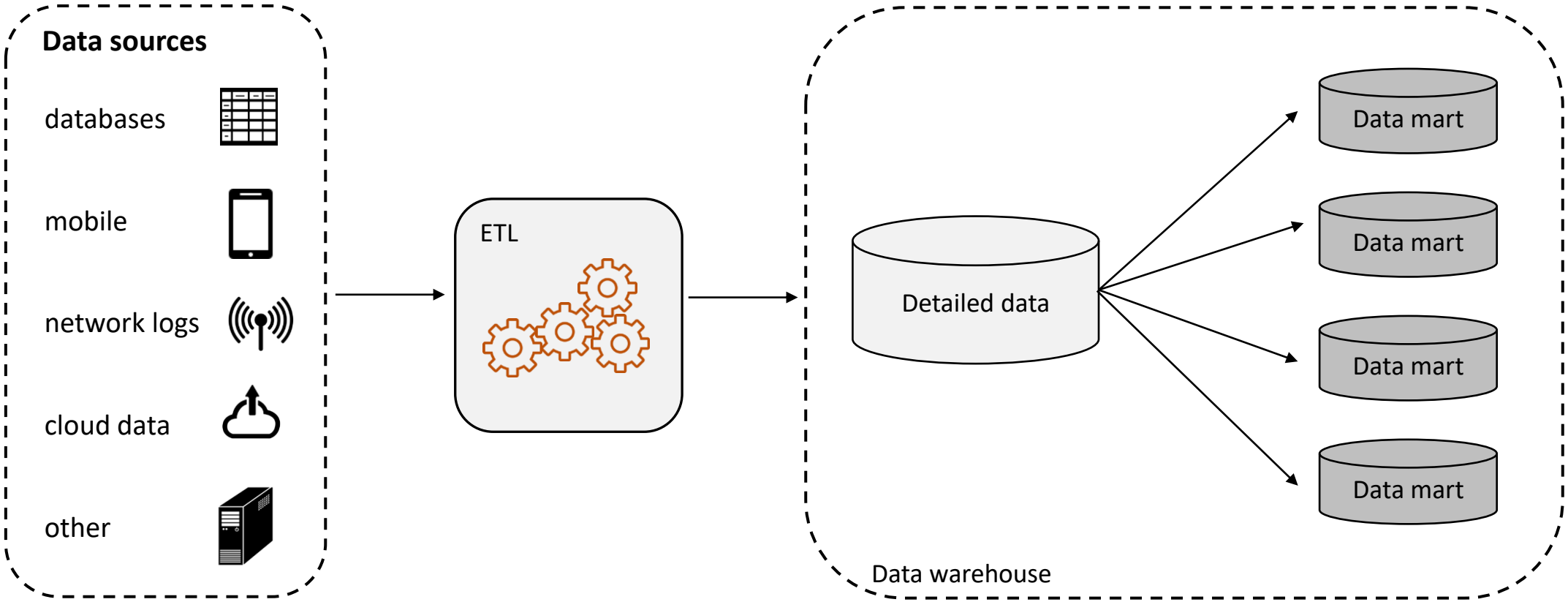
Architektura fizyczna



Źródło: pacificbigdata.com

ARCHITEKTURY BIG DATA

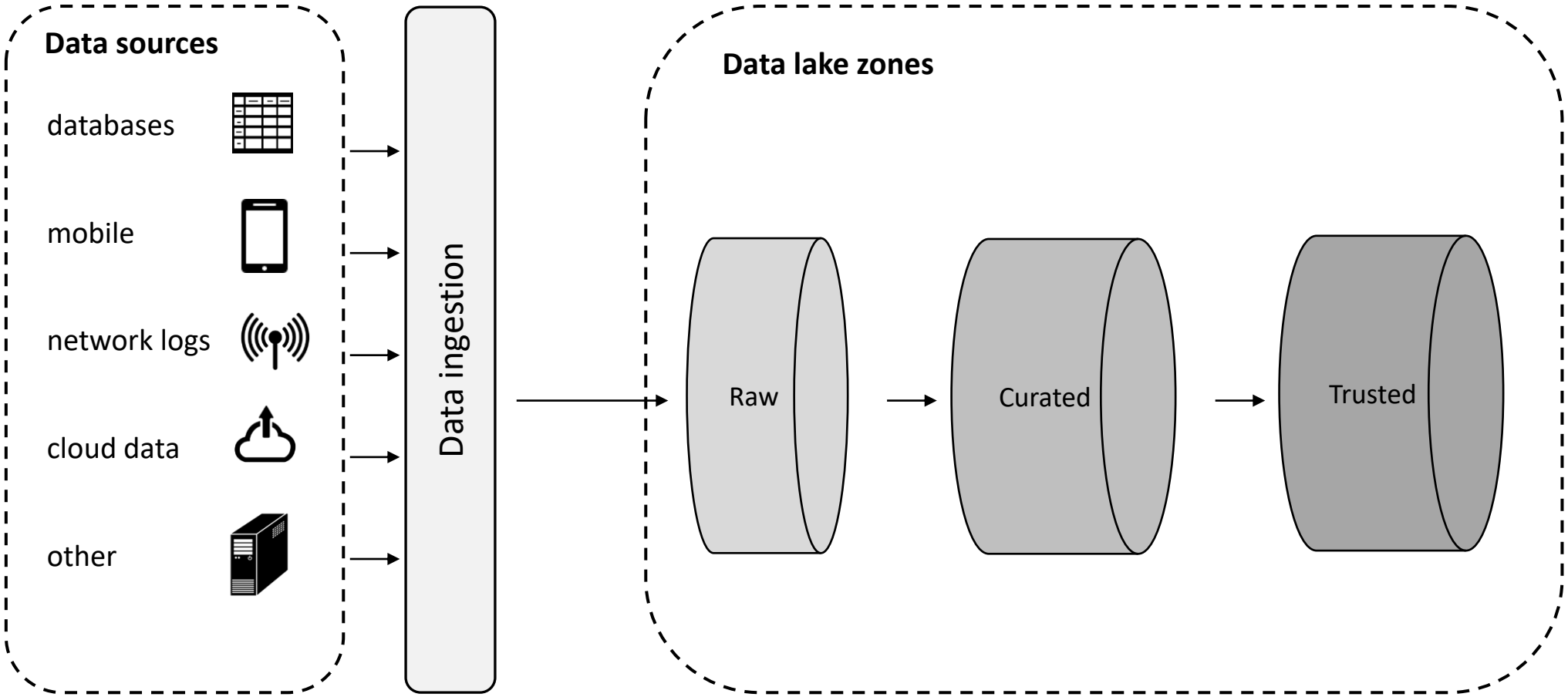
Data Warehouse

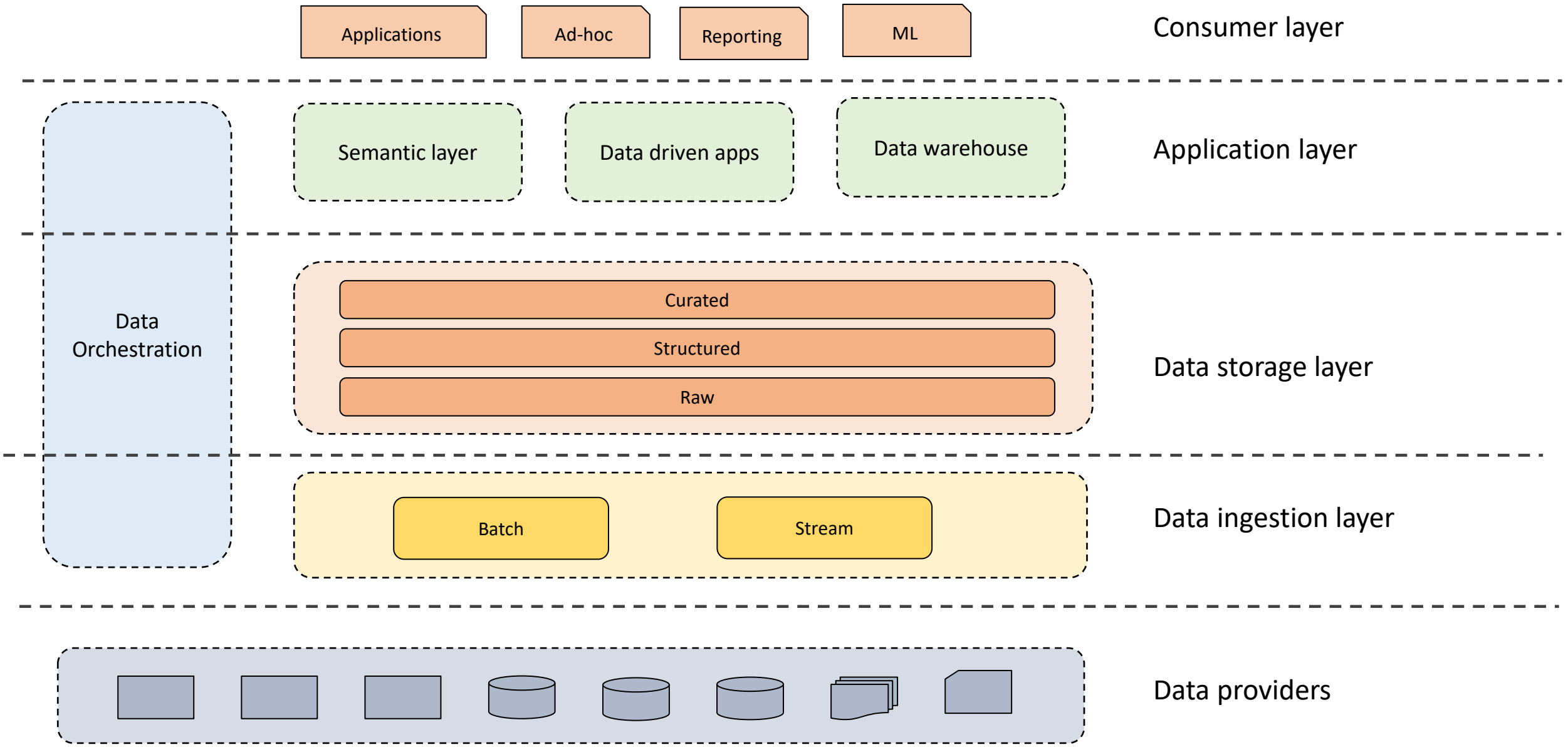


Data Lake

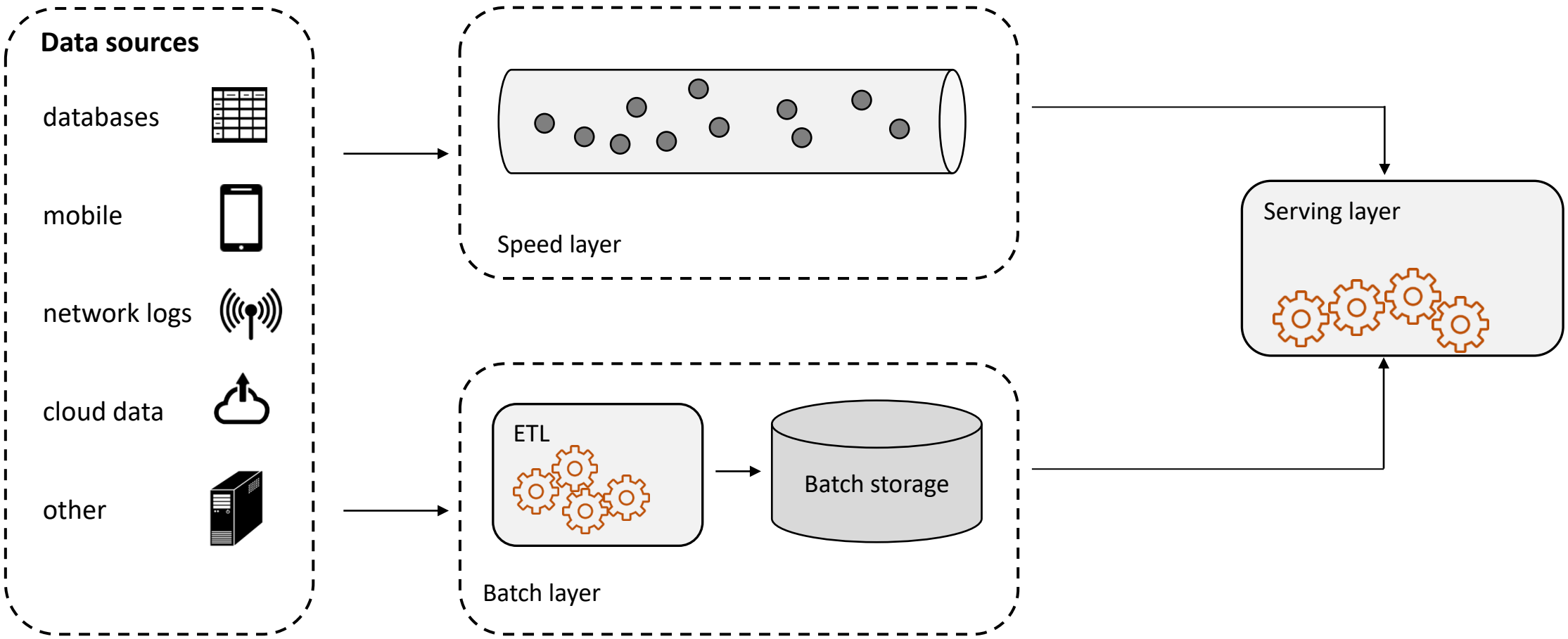
- Repozytorium służące składowaniu i przetwarzaniu danych o bardzo dużej skali i zróżnicowaniu
- Możliwość podłączania zróżnicowanych źródeł danych, zarówno posiadających strukturę jak i pozbawionych struktury; danych wsadowych oraz strumieni
- Dane nie są składowane w sposób uporządkowany jak w przypadku hurtowni danych czy data martów. Jest to często federacja technologii, baz danych i strumieni danych
- Architektura powstała jako odpowiedź na wady „klasycznych” hurtowni danych:
  - HD odpowiadają tylko na pytania, które były znane wcześniej
  - Hurtownie danych i data marty posiadają dane o określonej szczegółowości. Nie można jej zwiększyć
  - HD opierają się na zdefiniowanych źródłach danych

Data Lake

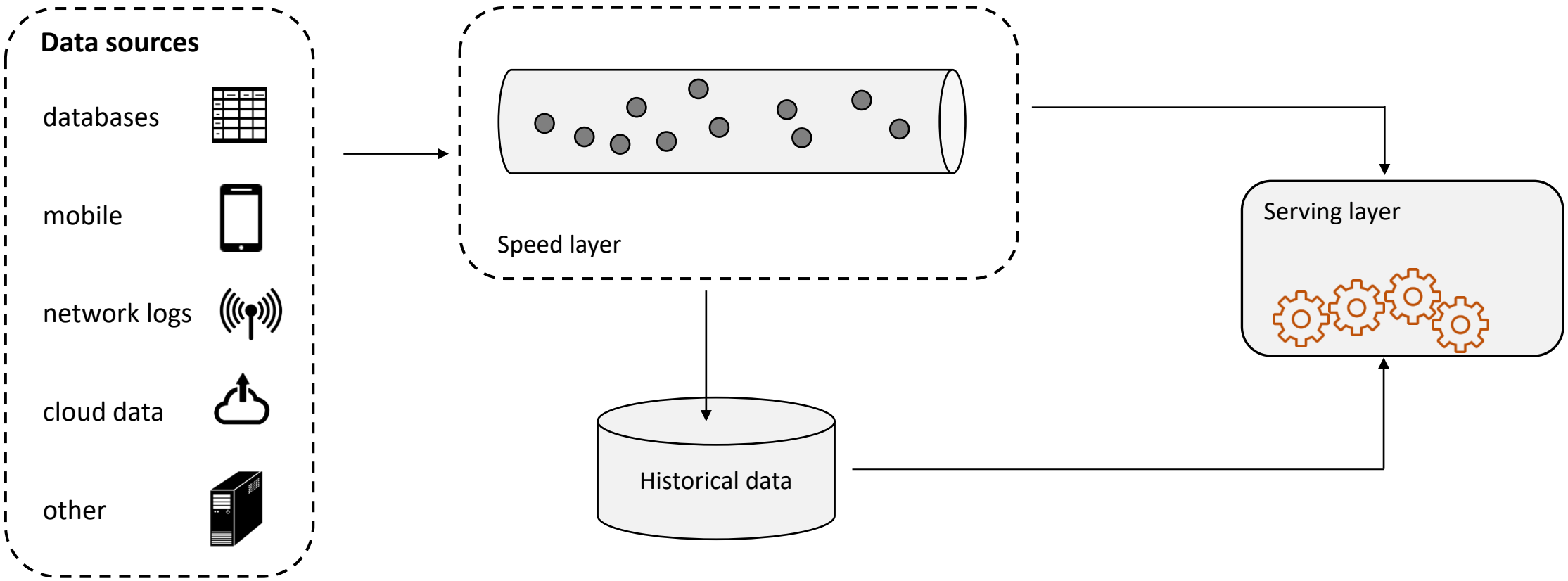

Data sources
databases
mobile
network logs
cloud data
other
Data ingestion
Data lake zones
Raw
Curated
Trusted


Applications
Ad-hoc
Reporting
ML
Consumer layer
Semantic layer
Data driven apps
Data warehouse
Application layer
Data
Orchestration
Curated
Structured
Raw
Data storage layer
Batch
Stream
Data ingestion layer
Data providers

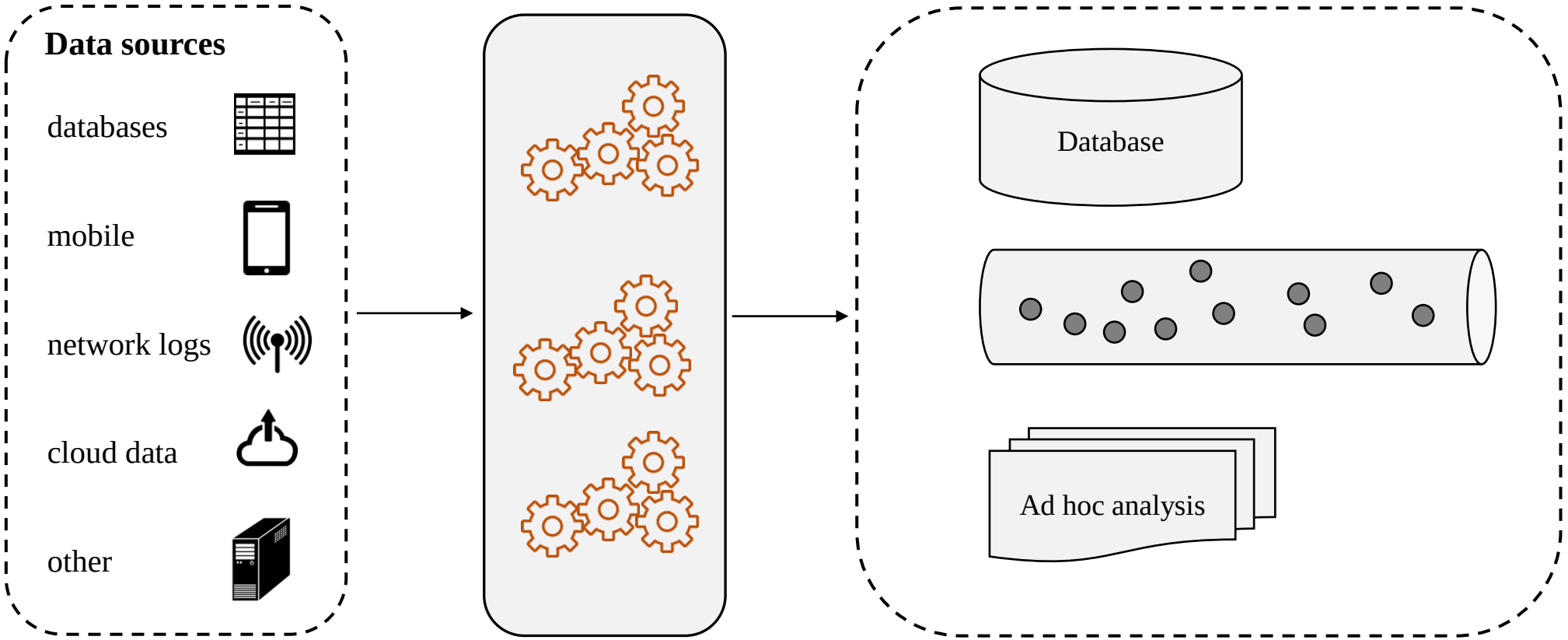
Lambda


Data sources
databases
mobile
network logs
cloud data
other
Speed layer
ETL
Batch storage
Batch layer
Serving layer

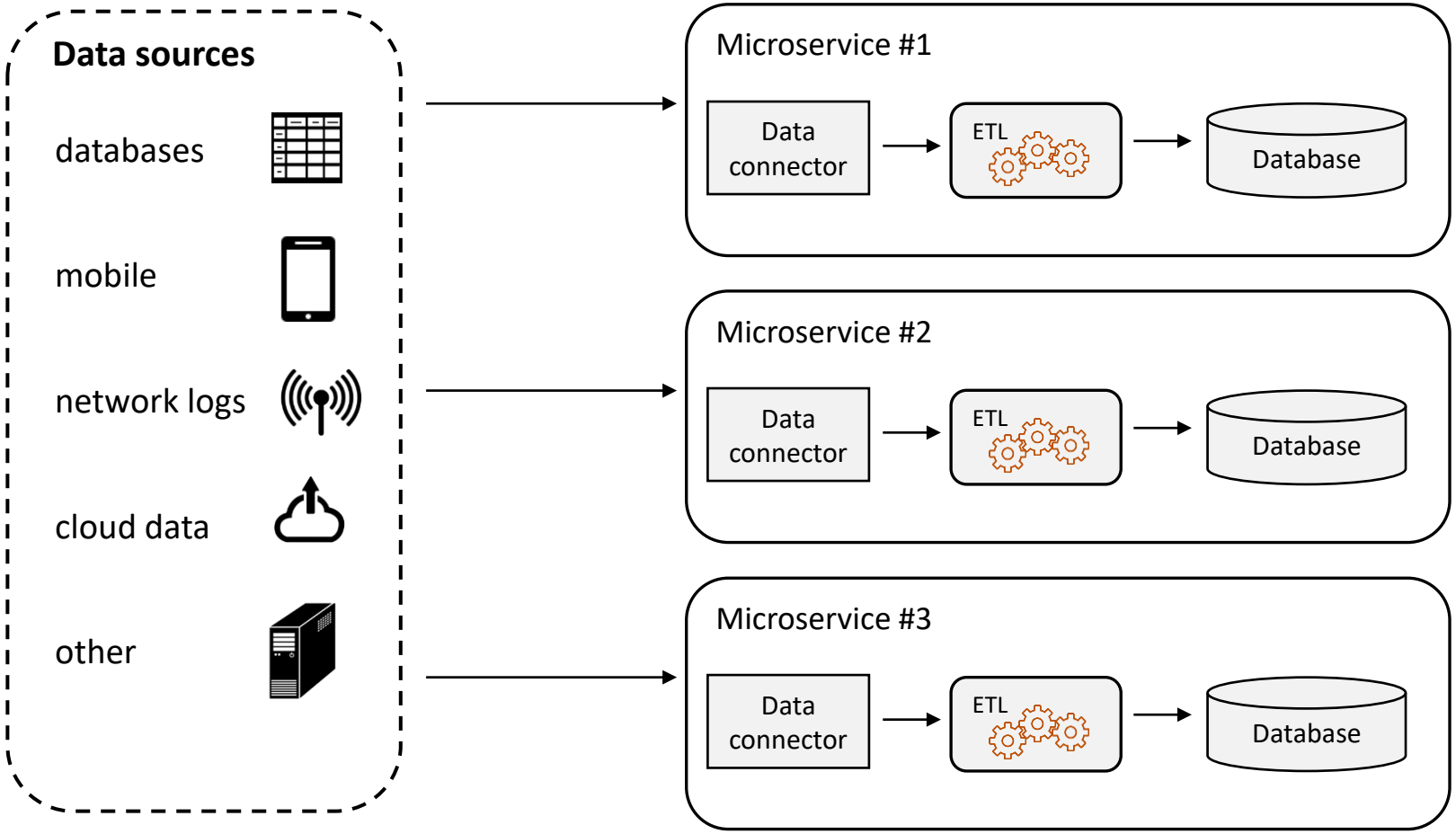
Kappa


Data sources
databases
mobile
network logs
cloud data
other
Speed layer
Historical data
Serving layer

Lakehouse


Data sources
databases
mobile
network logs
cloud data
other
Database
Ad hoc analysis

Microservices



APACHE HADOOP

Komponenty technologiczne - klasyfikacja

- Składowanie danych
- Przetwarzanie
- Automatyzacja przetwarzania
- Administrowanie klastrem
- Bezpieczeństwo
- Przetwarzanie strumieniowe
- Machine learning

Dystrybucje Apache Hadoop

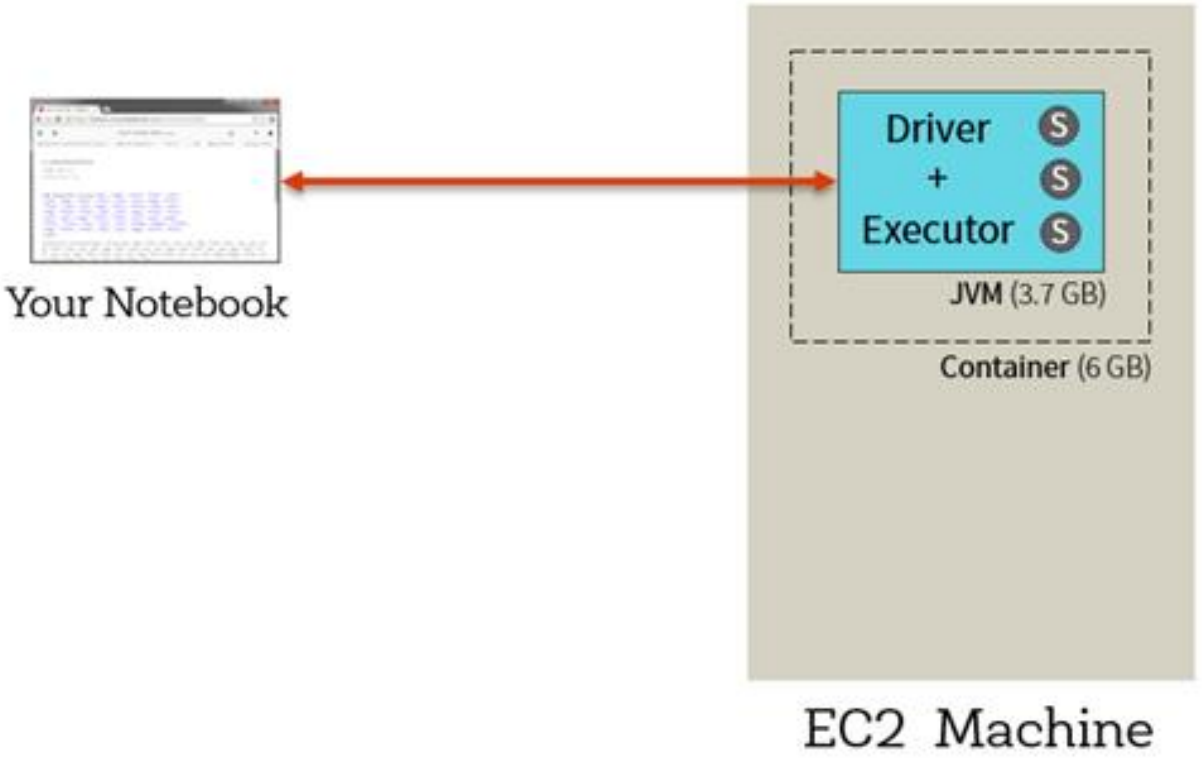
- Alibaba Cloud E-MapReduce
- Amazon EMR
- Azure HDInsight
- Google Cloud Dataproc
- Cloudera/Hortonworks

Databricks

Databricks community edition

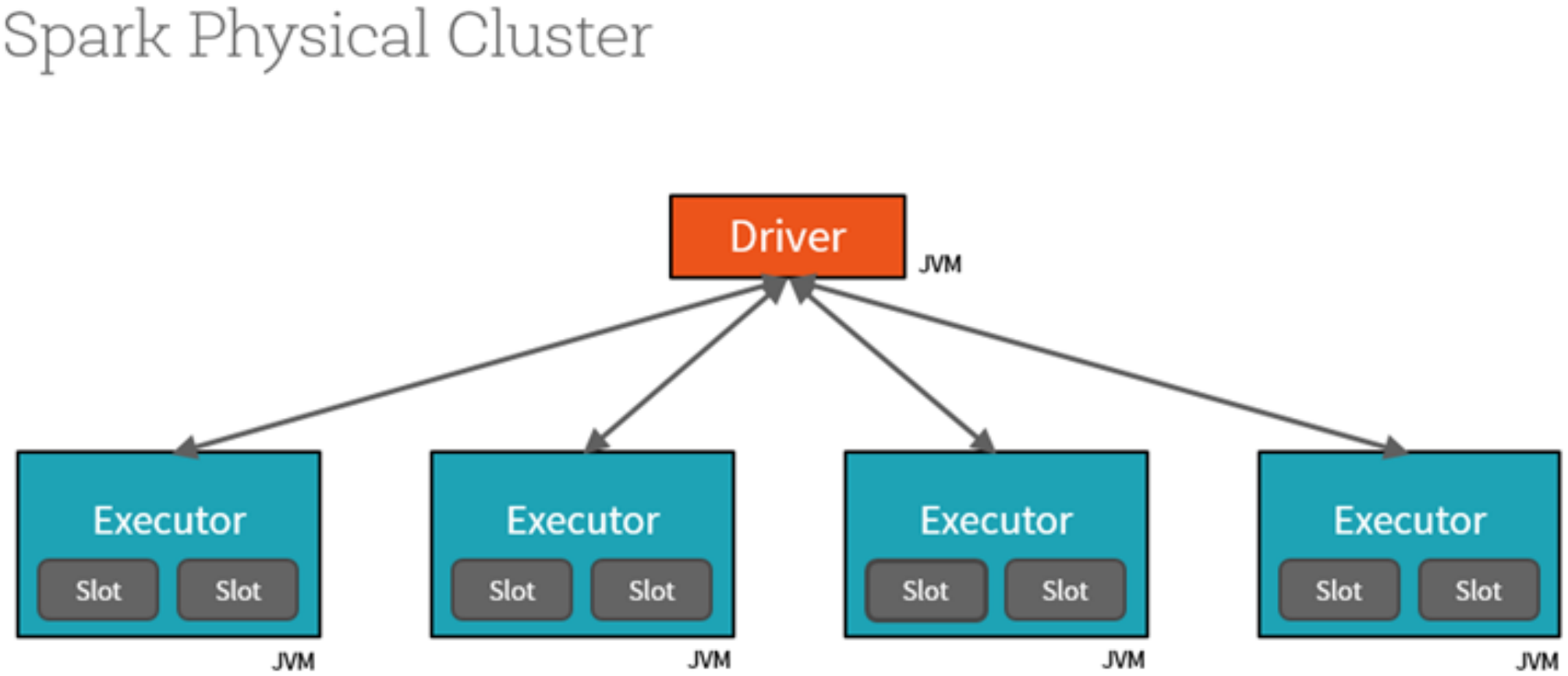
https://community.cloud.databricks.com

Databricks platform



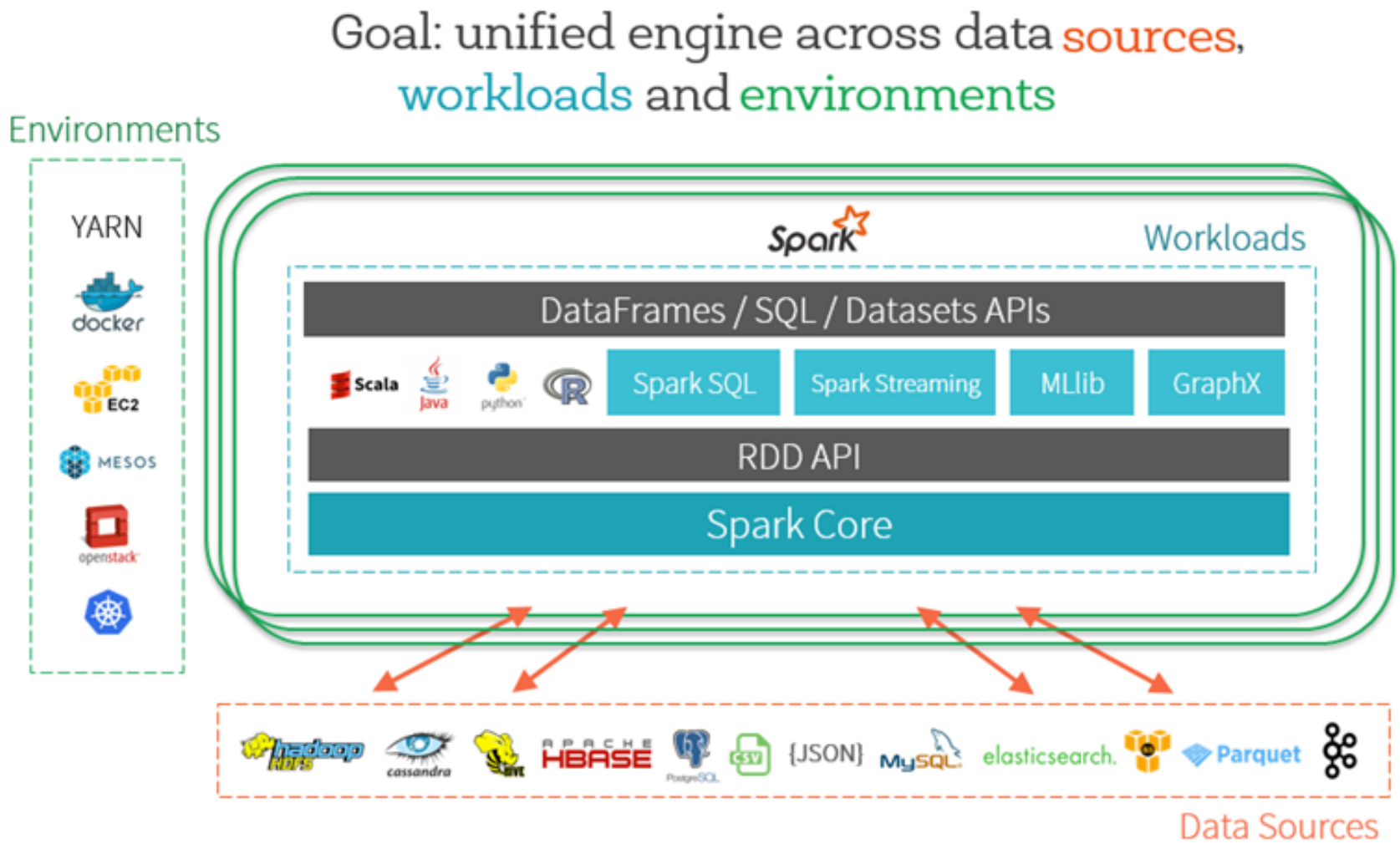
Źródło: databricks.org

Databricks platform



Źródło: databricks.org

Databricks platform



Źródło: databricks.org

Dziękuję za uwagę