Big Data

Organizacyjnie

Prowadzący:

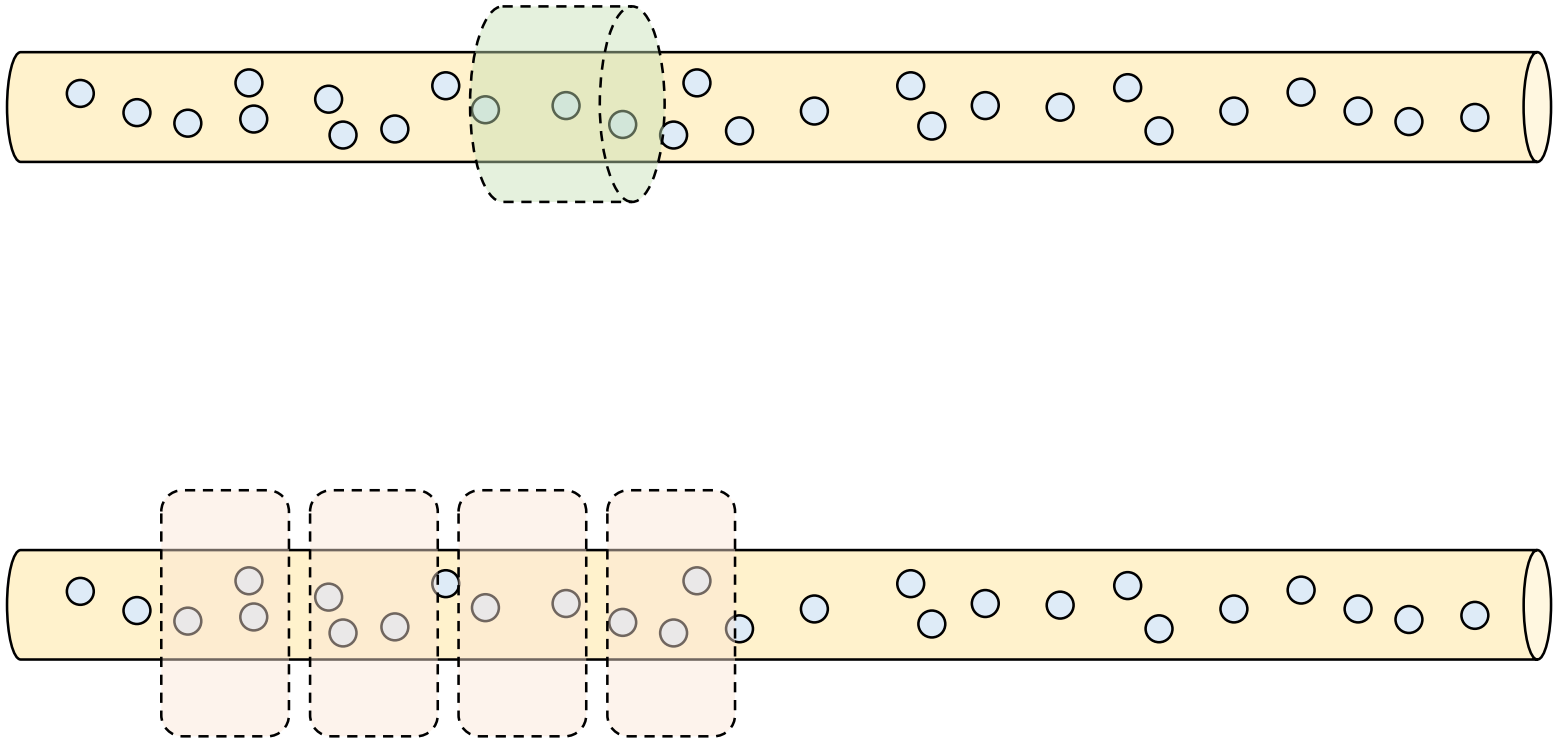
dr Mariusz Rafało

mrafalo@sgh.waw.pl

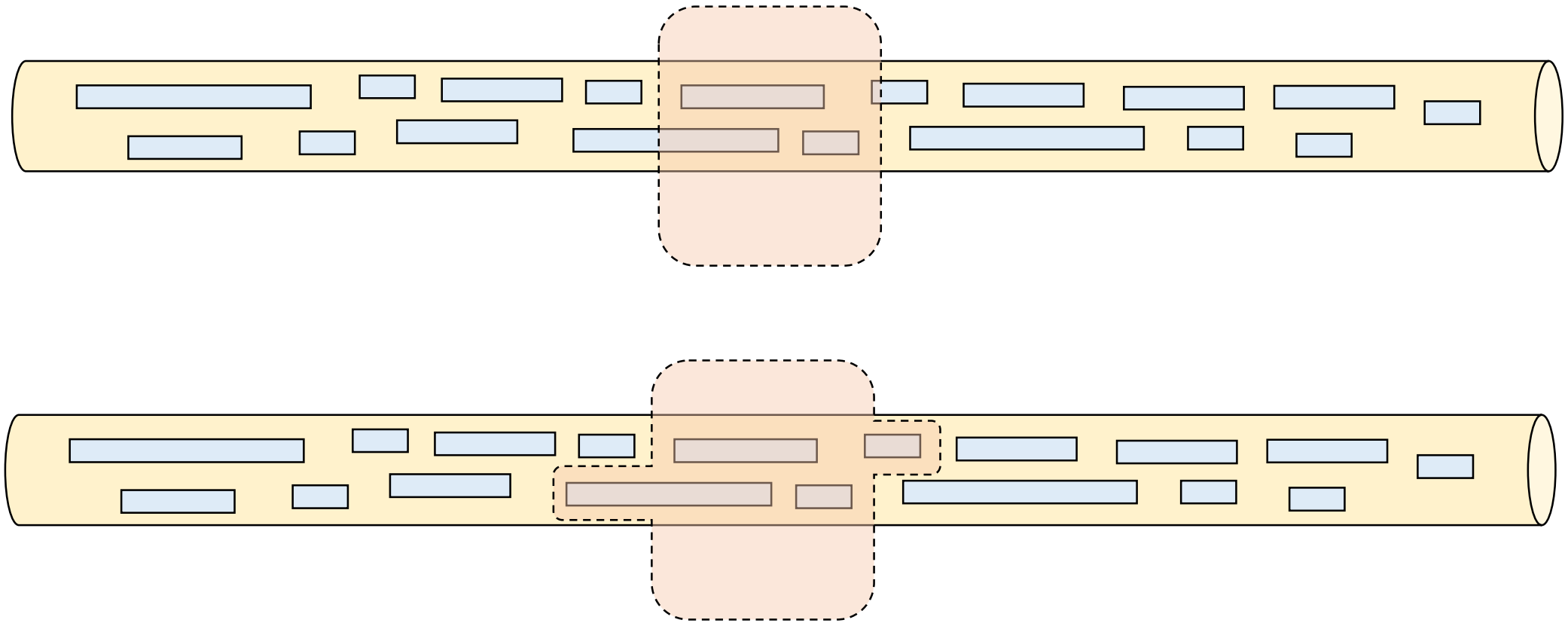
http://mariuszrafalo.pl (hasło: BIG)

DANE W CZASIE RZECZYWISTYM

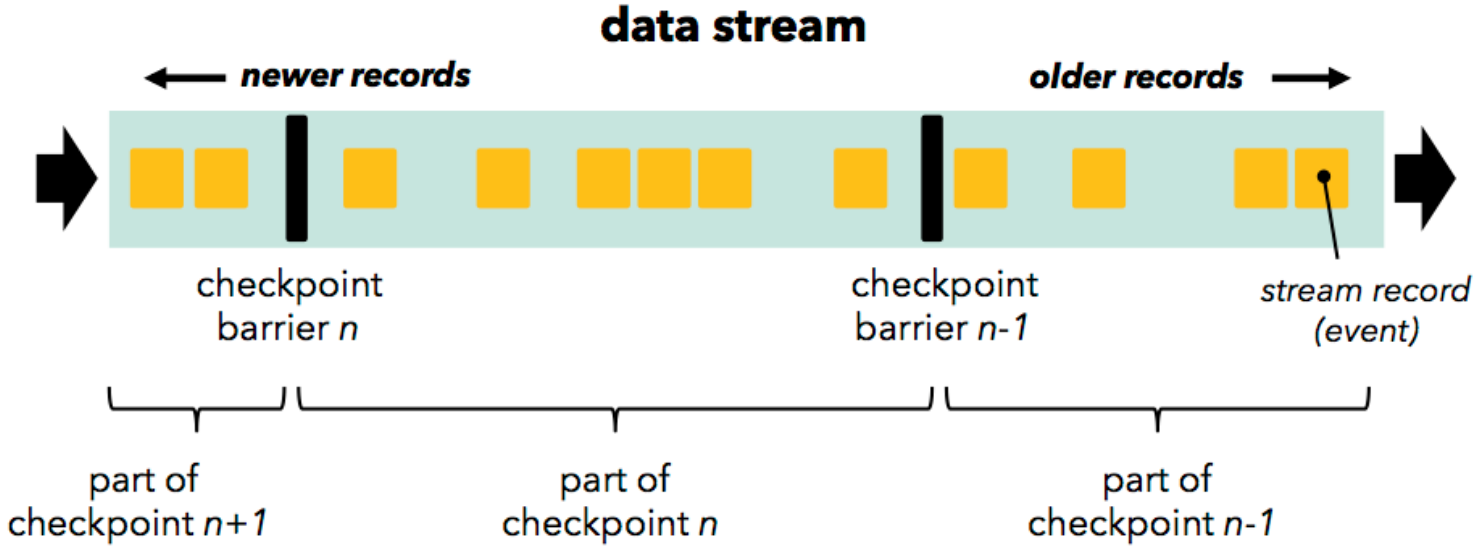
Tryb analizowania danych

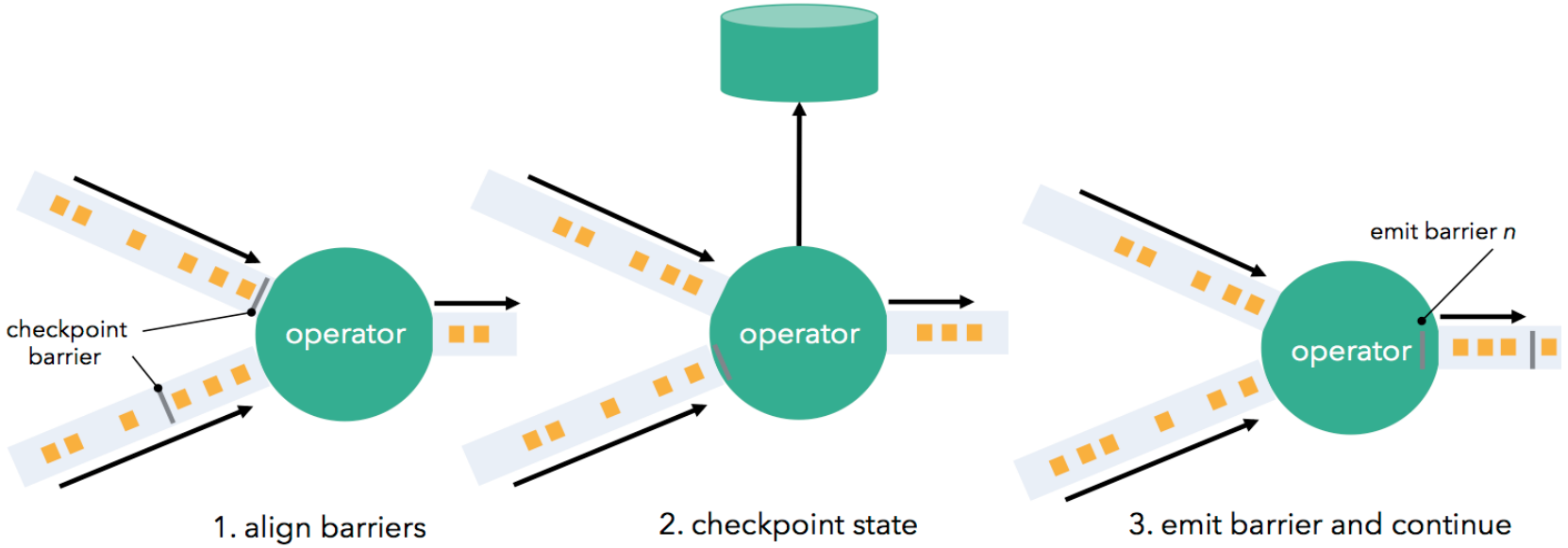


Okno analizowania

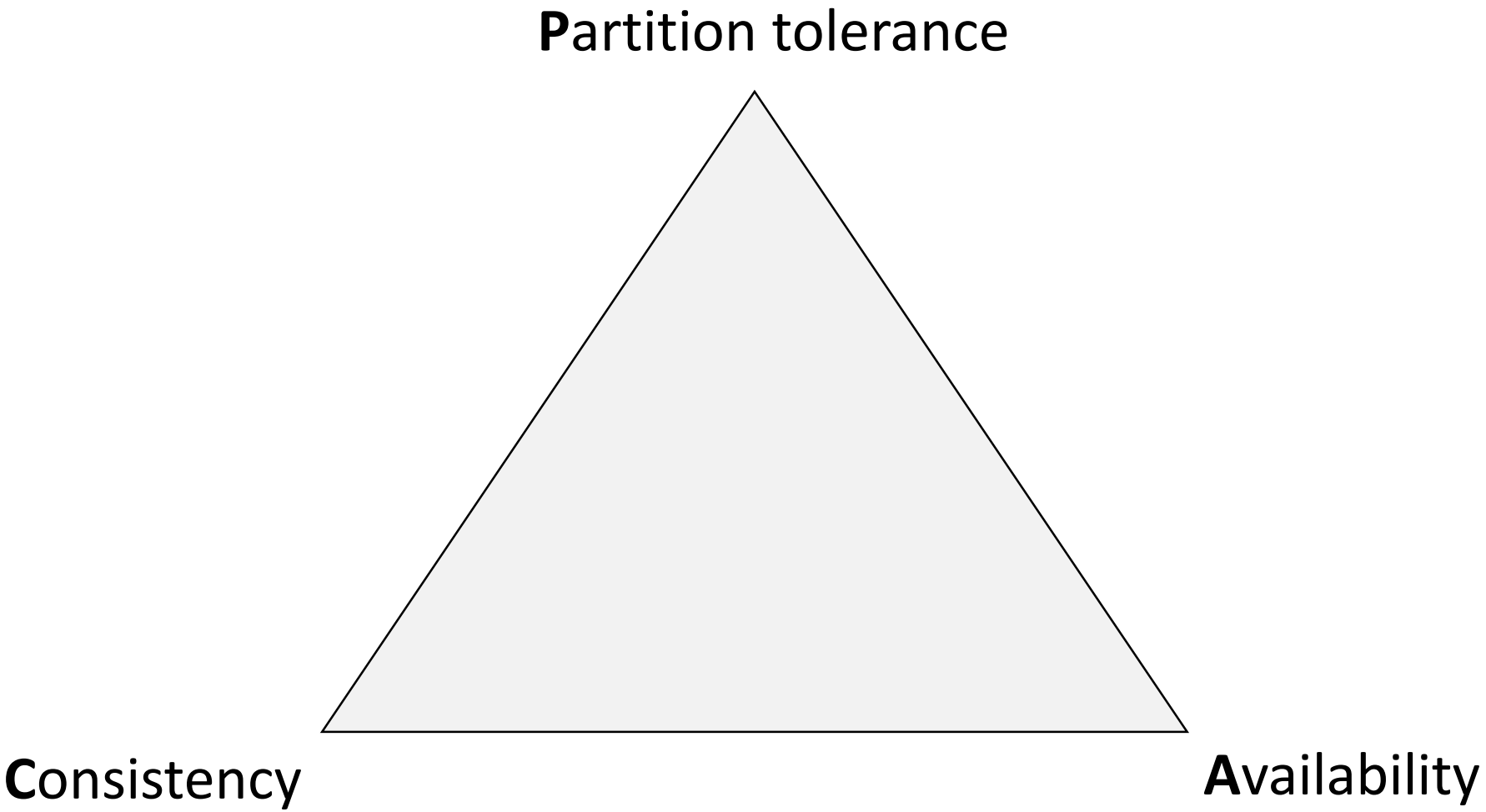


Real-time: Checkpointing





Teoria CAP

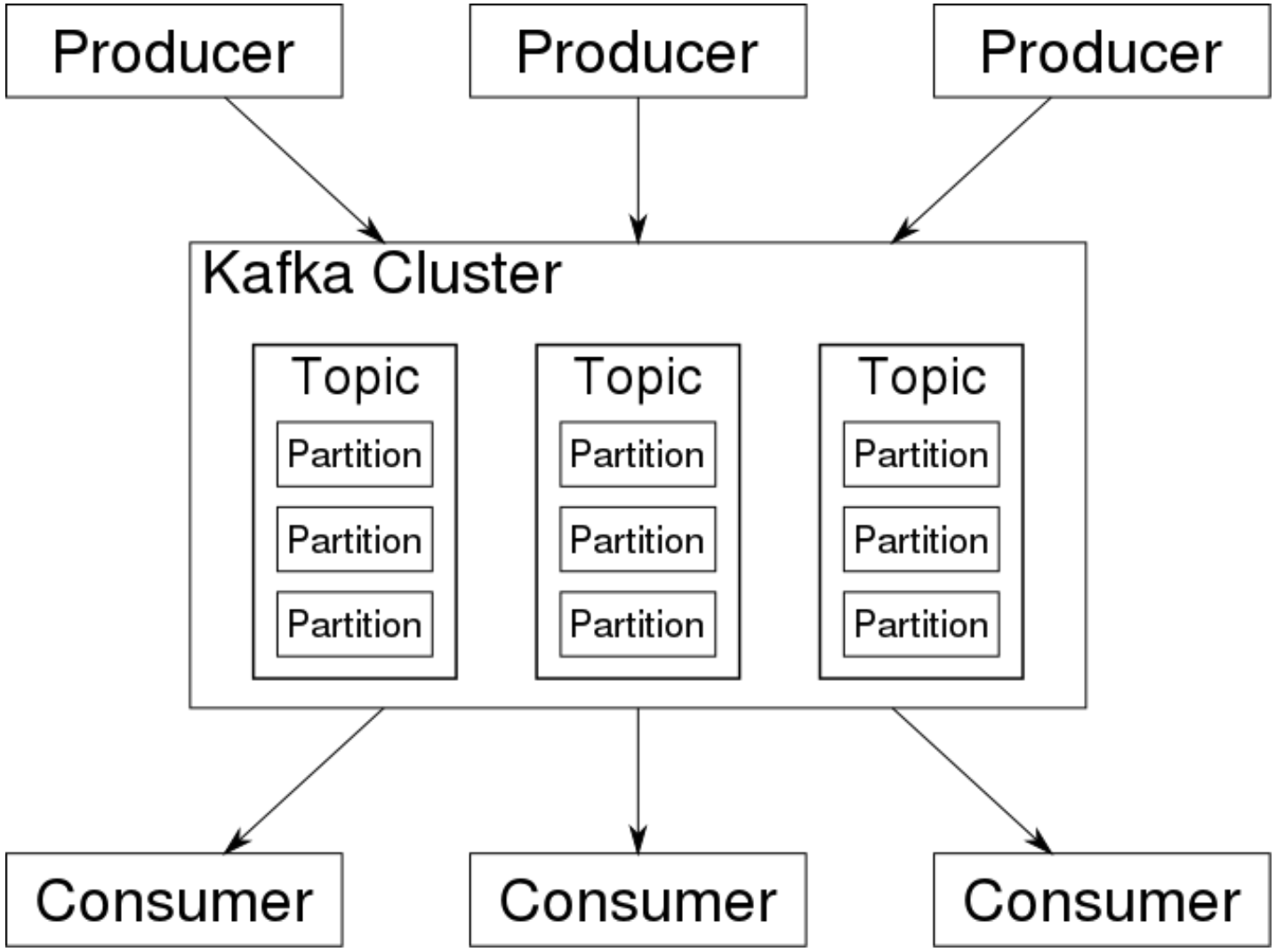


TECHNOLOGIE REAL-TIME

Apache Kafka

- Platforma służąca do tworzenia i zarządzania strumieniami danych
- Technologia oparta jest na tzw. brokerach: pojedynczy broker może obsługiwać setki megabajtów danych zapisywanych i odczytywanych w ciągu sekundy
- Dane w strumieniu mogą być partycjonowane; wówczas różne maszyny obsługują różne elementy strumienia
- Wszystkie komunikaty są składowane na dysku, przez co platforma jest odporna na zatrzymania pracy i inne awarie

Apache Kafka

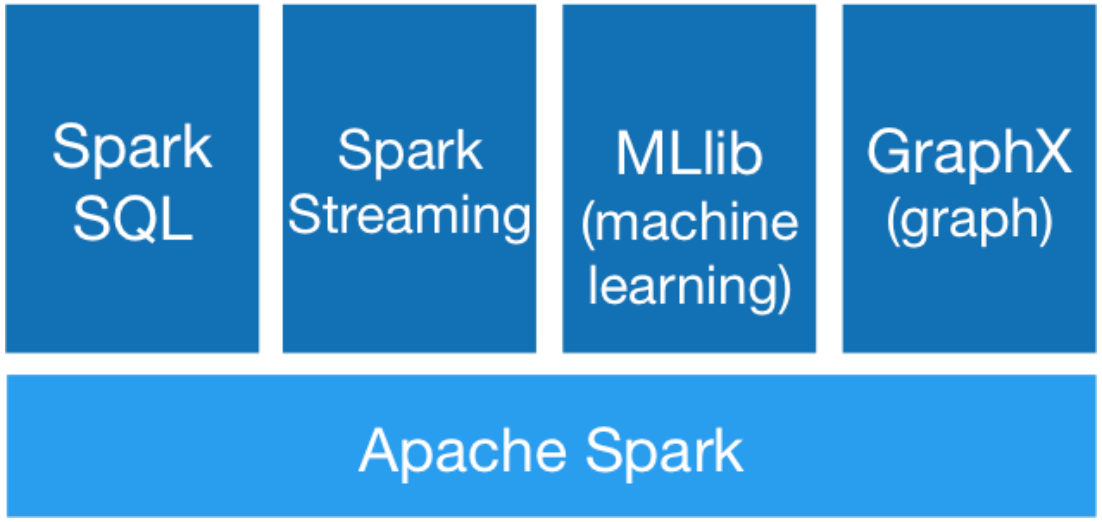


Apache Storm

- Platforma do strumieniowego przetwarzania danych
- Może służyć do analityki w czasie rzeczywistym, wykorzystania algorytmów statystycznych a także jako narzędzie ETL
- Platforma zintegrowana z narzędziem Apache Kafka
- Obsługuje okna czasowe oparte na datach biznesowych

Apache Spark

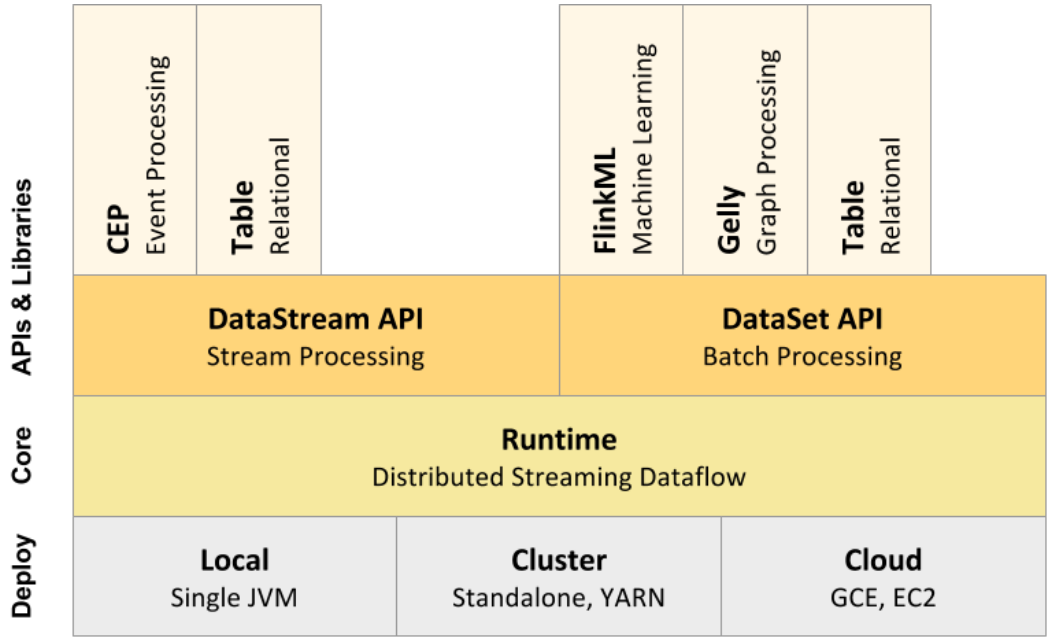
- Platforma do przetwarzania danych w dużej skali
- Obsługuje języki programowania: Java, Scala, Python, R
- Może pracować w trybie *batch* lub *stream*
- Posiada wiele wbudowanych bibliotek:
  - SQL
  - Data Frame
  - Mlib
  - GraphX



Źródło: spark.apache.org

Apache Flink

- Platforma do obsługi danych w trybie strumieniowym; pracuje w środowisku rozproszonym z wysokim poziomem tolerancji na awarie
- Flink składa się z kilku narzędzi:
  - DataStream API – służące do strumieni, które nie posiadają ograniczeń
  - DataSet API – służące do obsługi strumieni statycznych
  - Table API – pozwalające na stosowanie składni SQL
  - Biblioteka CEP (*Complex Event Processing*)
  - Biblioteka Machine Learning



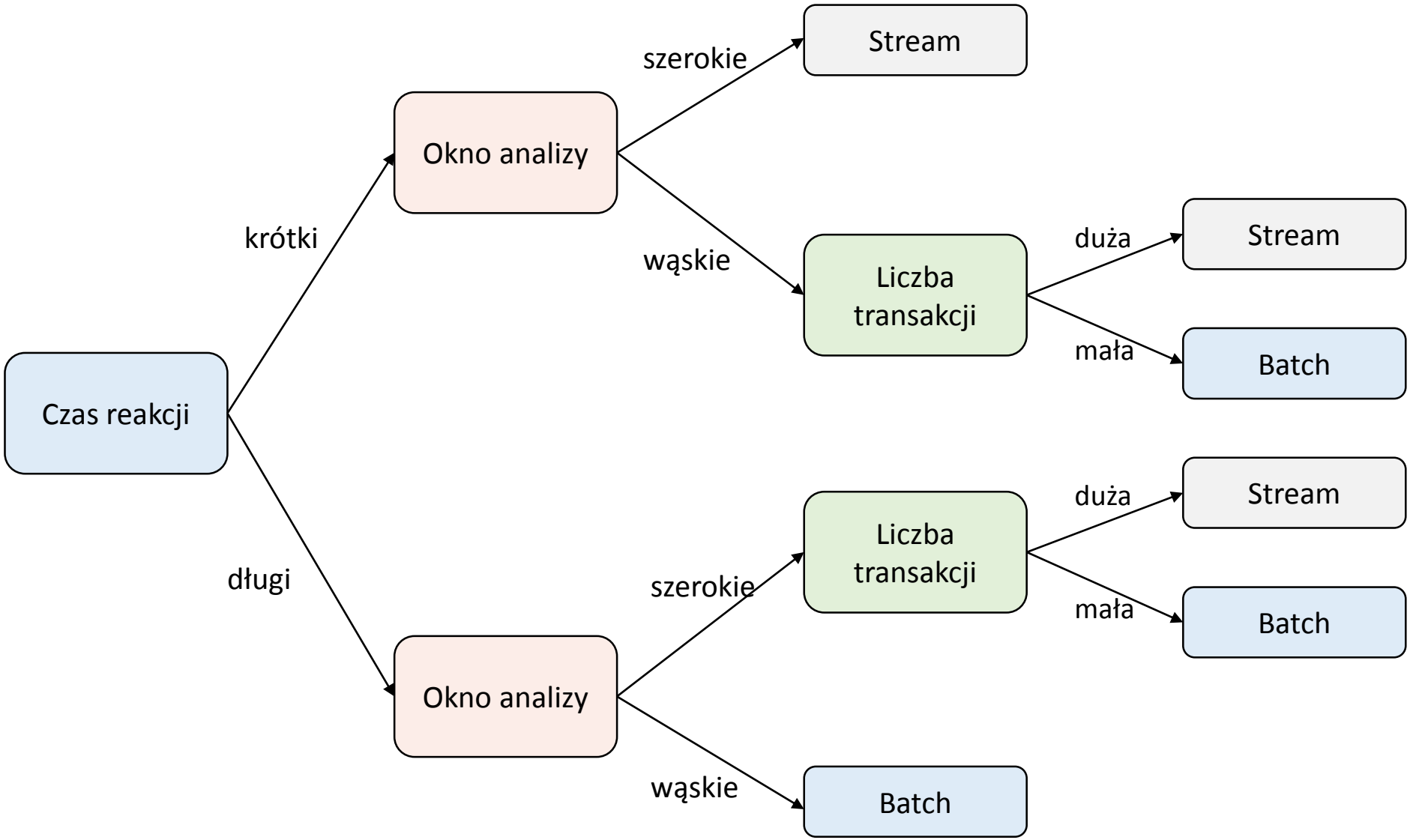
Źródło: flink.apache.org

Apache Samoa

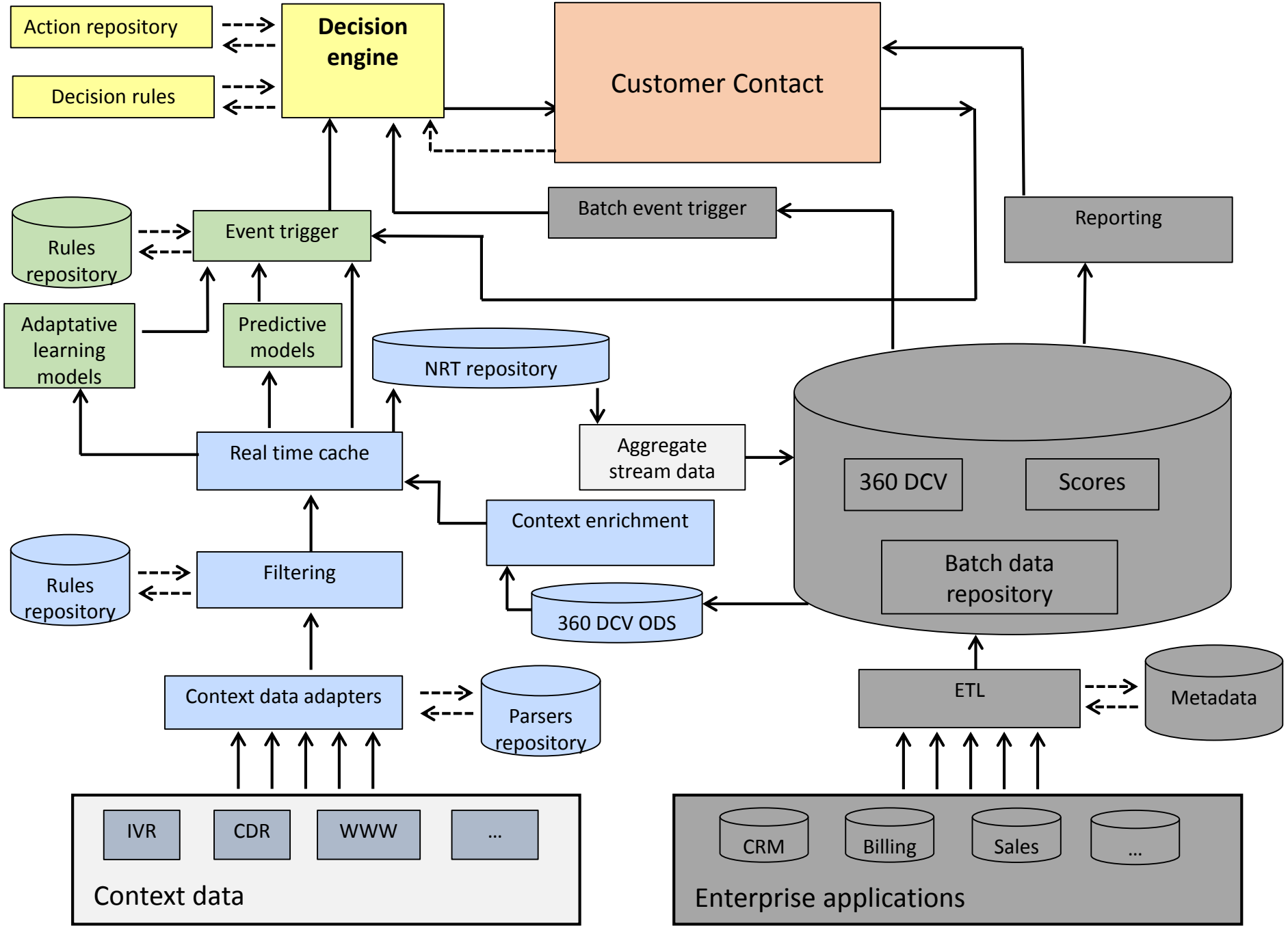
- Platforma służąca rozproszonemu wykonywaniu zadań związanych ze statystyczną obróbką danych
- Obsługuje najpopularniejsze algorytmy Machine Learning, jako biblioteki
- Obsługuje automatycznie strumienie danych pochodzące z różnych źródeł (Kafka, Storm, Samza, itp.)
- Pozwala na tworzenie i wykorzystanie w aplikacjach własnych bibliotek

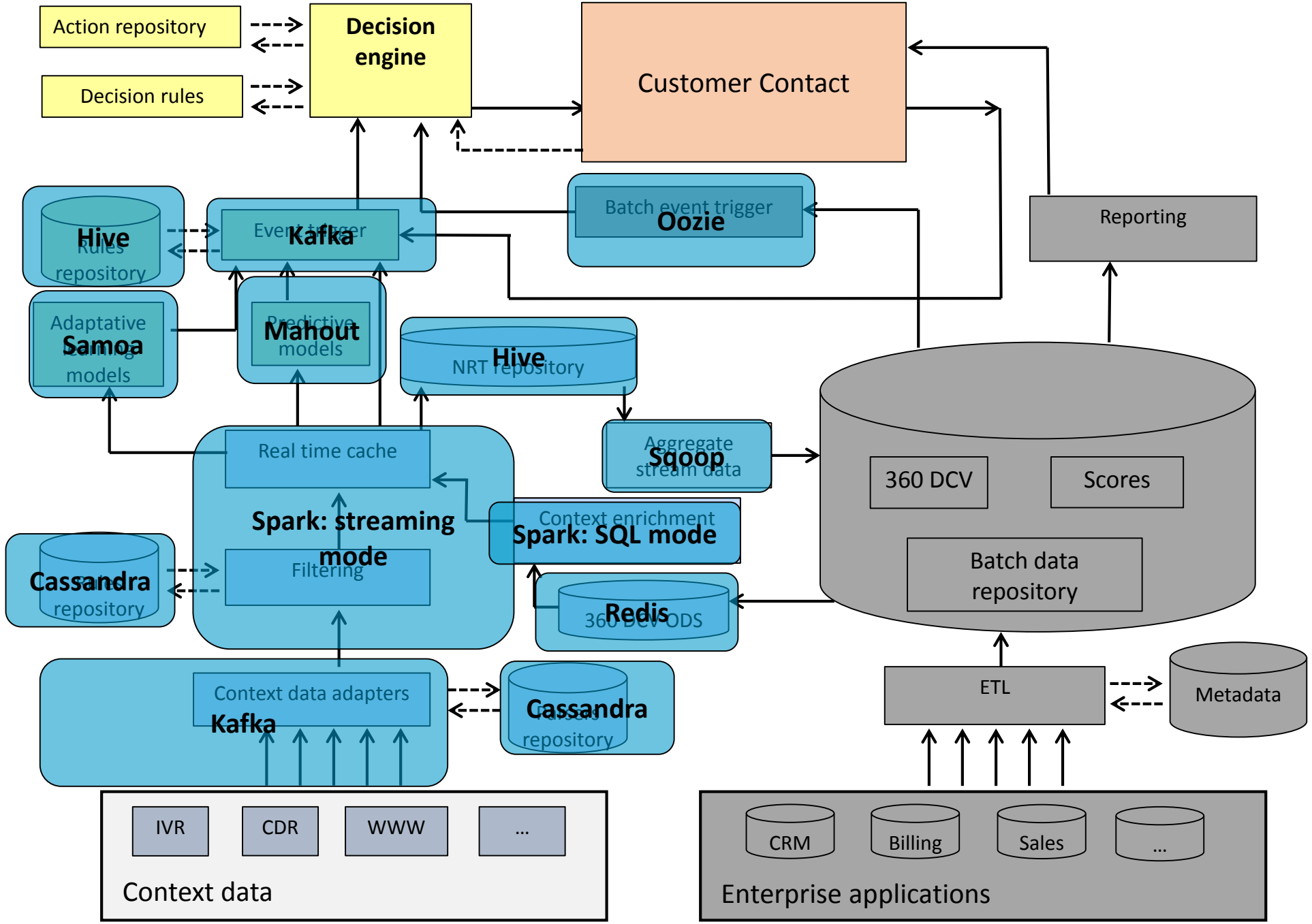
BATCH VS REAL-TIME

Podejście wsadowe vs real-time



WYKORZYSTANIE NARZĘDZI


Action repository
Decision rules
Decision engine
Customer Contact
Batch event trigger
Reporting
Rules repository
Event trigger
Adaptative learning models
Predictive models
NRT repository
Real time cache
Aggregate stream data
360 DCV
Scores
Batch data repository
Context enrichment
Rules repository
Filtering
360 DCV ODS
Context data adapters
Parsers repository
ETL
Metadata
IVR
CDR
WWW
...
Context data
CRM
Billing
Sales
...
Enterprise applications


Action repository
Decision rules
Decision engine
Customer Contact
Hive
Rules repository
Kafka
Batch event trigger
Oozie
Reporting
Adaptative learning models
Samoa
Predictive models
Mahout
NRT repository
Hive
Aggregate stream data
Sqoop
Real time cache
Spark: streaming mode
Filtering
Context enrichment
Spark: SQL mode
360 DCV
Scores
Batch data repository
Cassandra
Redis
360 DCV ODS
Context data adapters
Kafka
Cassandra
ETL
Metadata
IVR
CDR
WWW
...
Context data
CRM
Billing
Sales
...
Enterprise applications

Dziękuję za uwagę