

Hadoop i Spark

Mariusz Rafało

mrafalo@sgh.waw.pl

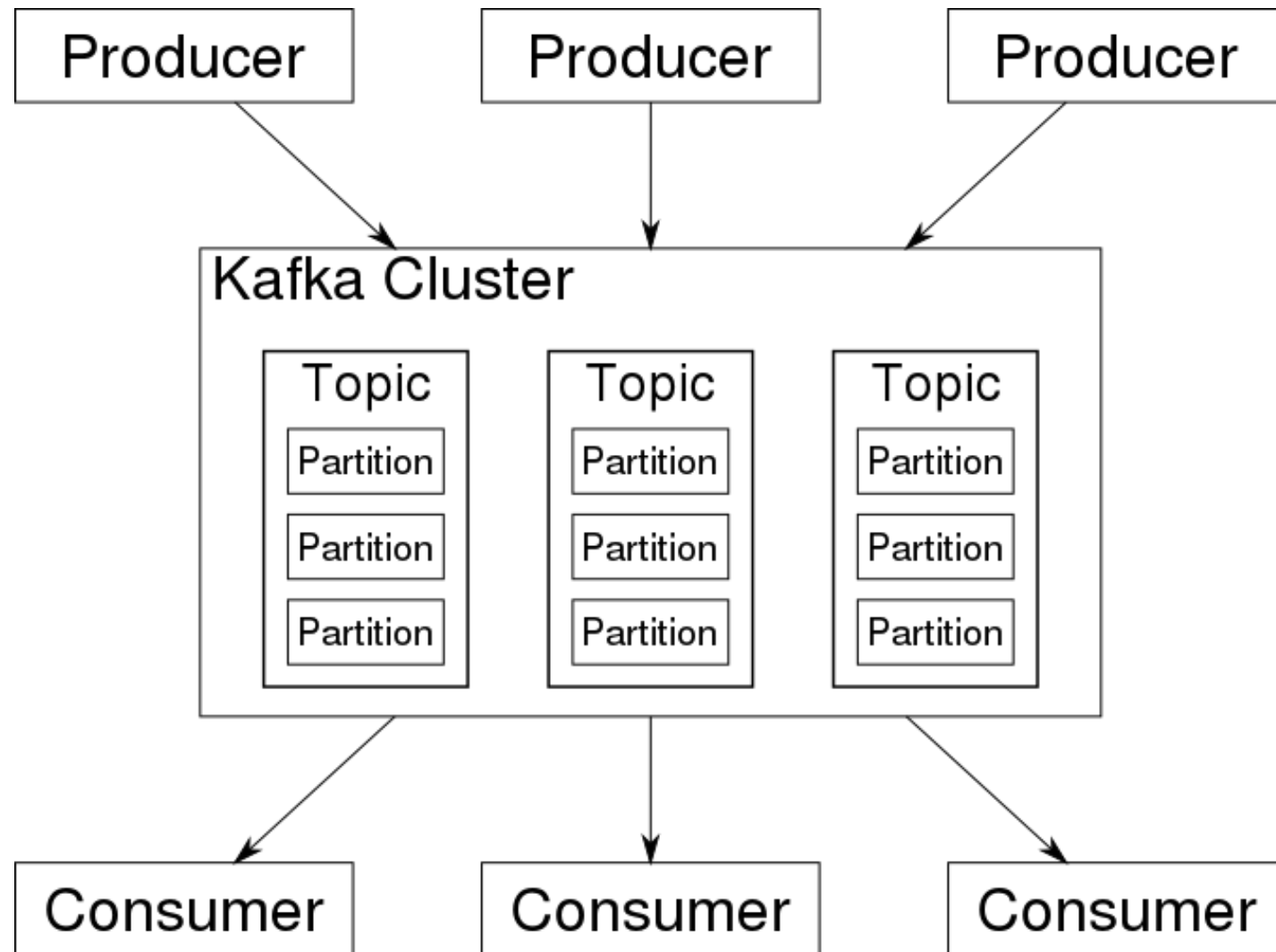
<http://mariuszrafalo.pl>

TECHNOLOGIE REAL-TIME

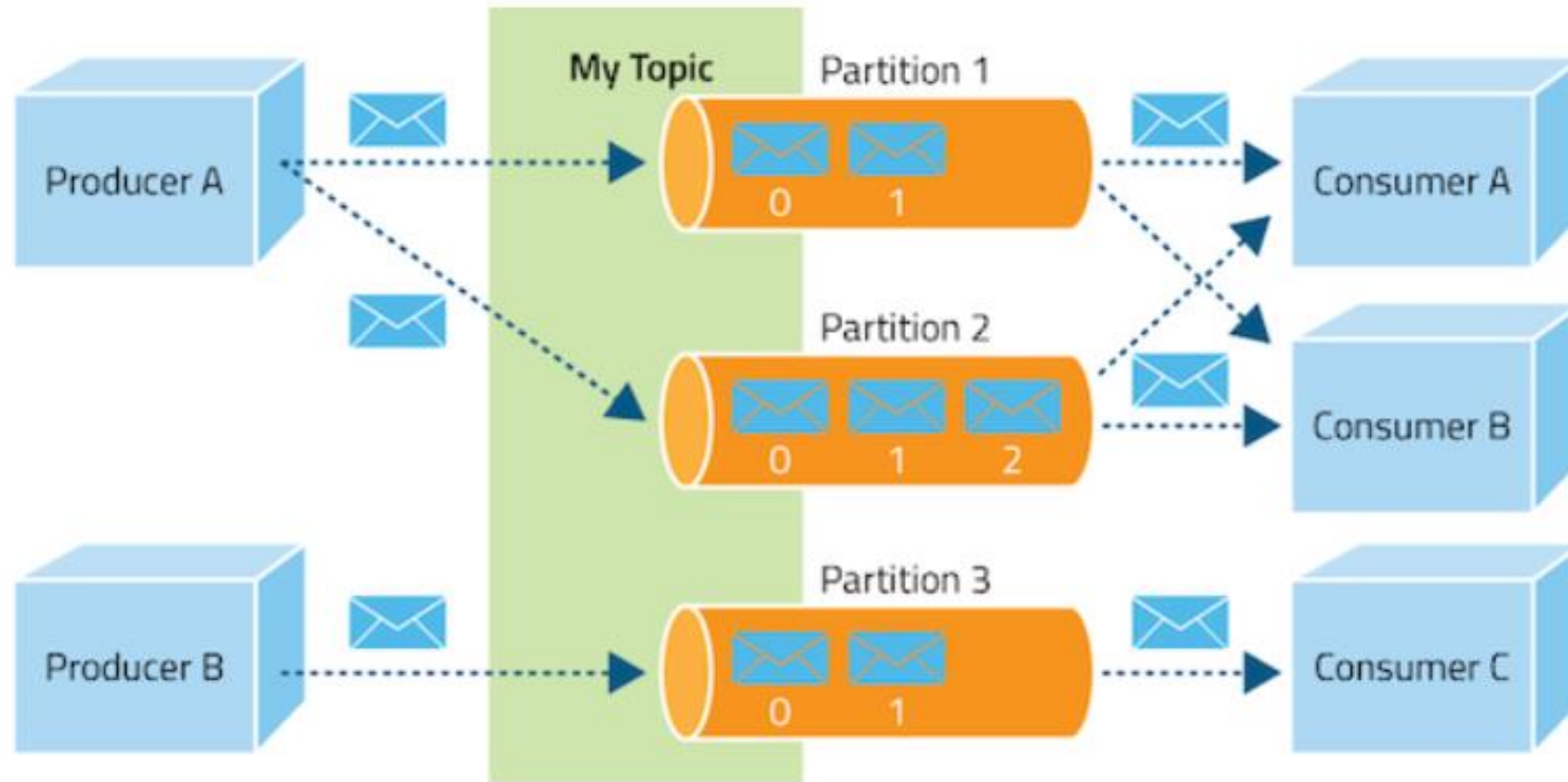
Apache Kafka

- 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 Kafka

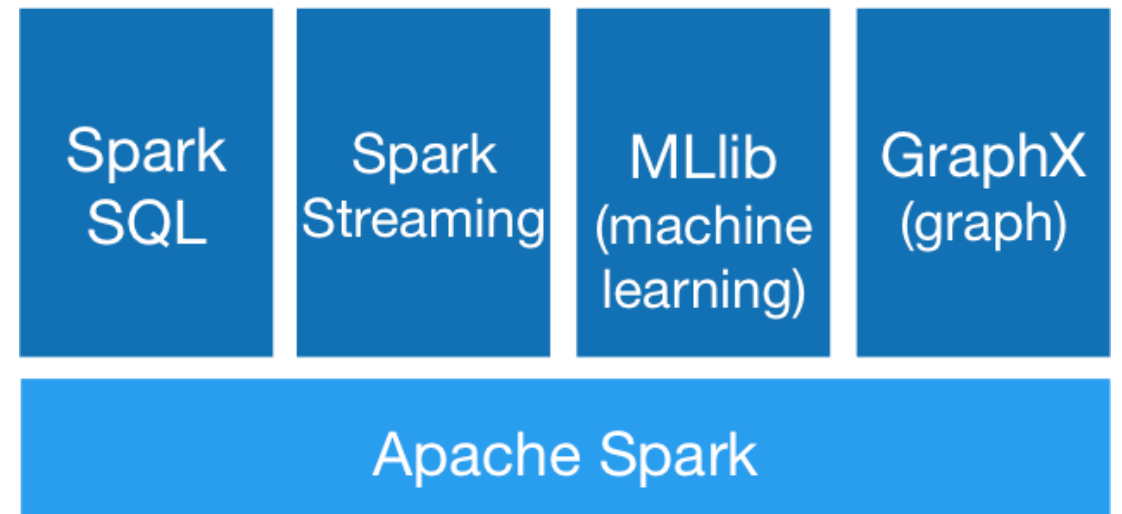


Apache Storm

- 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 danych biznesowych

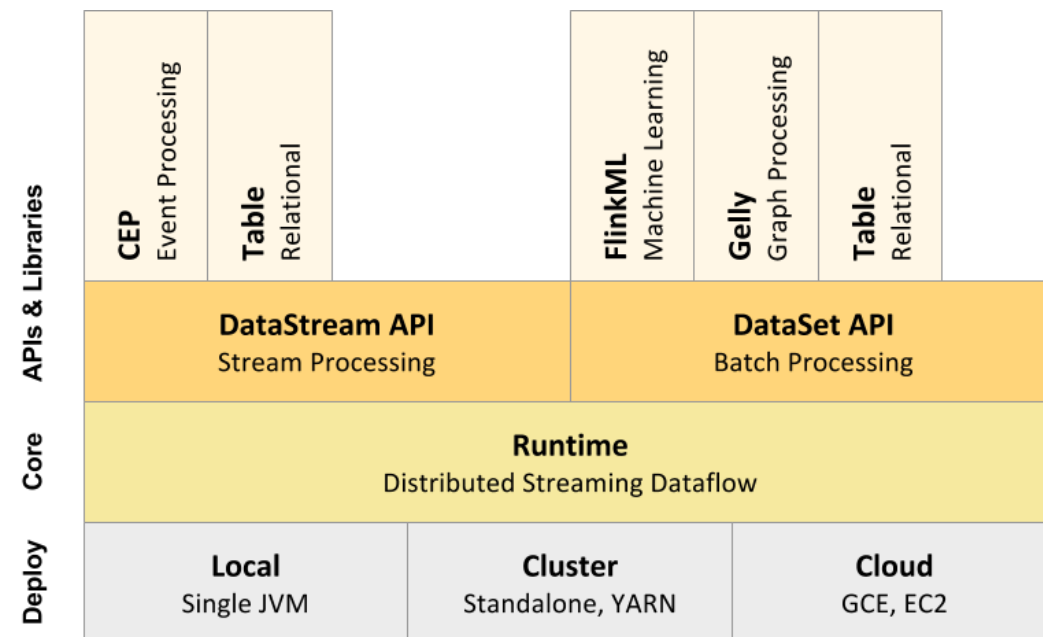
Apache Spark

- 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



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



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

Dziękuję za uwagę