Microsoft

Microsoft Build

May 7–9, 2018 // Seattle, WA

Microsoft

Real-time data streams with Apache Kafka and Spark

Alena Hall, @lenadroid

THR3504

Data

Ever-increasing

@lenadroid

Data Producers and Consumers

Are data workflows flexible enough?

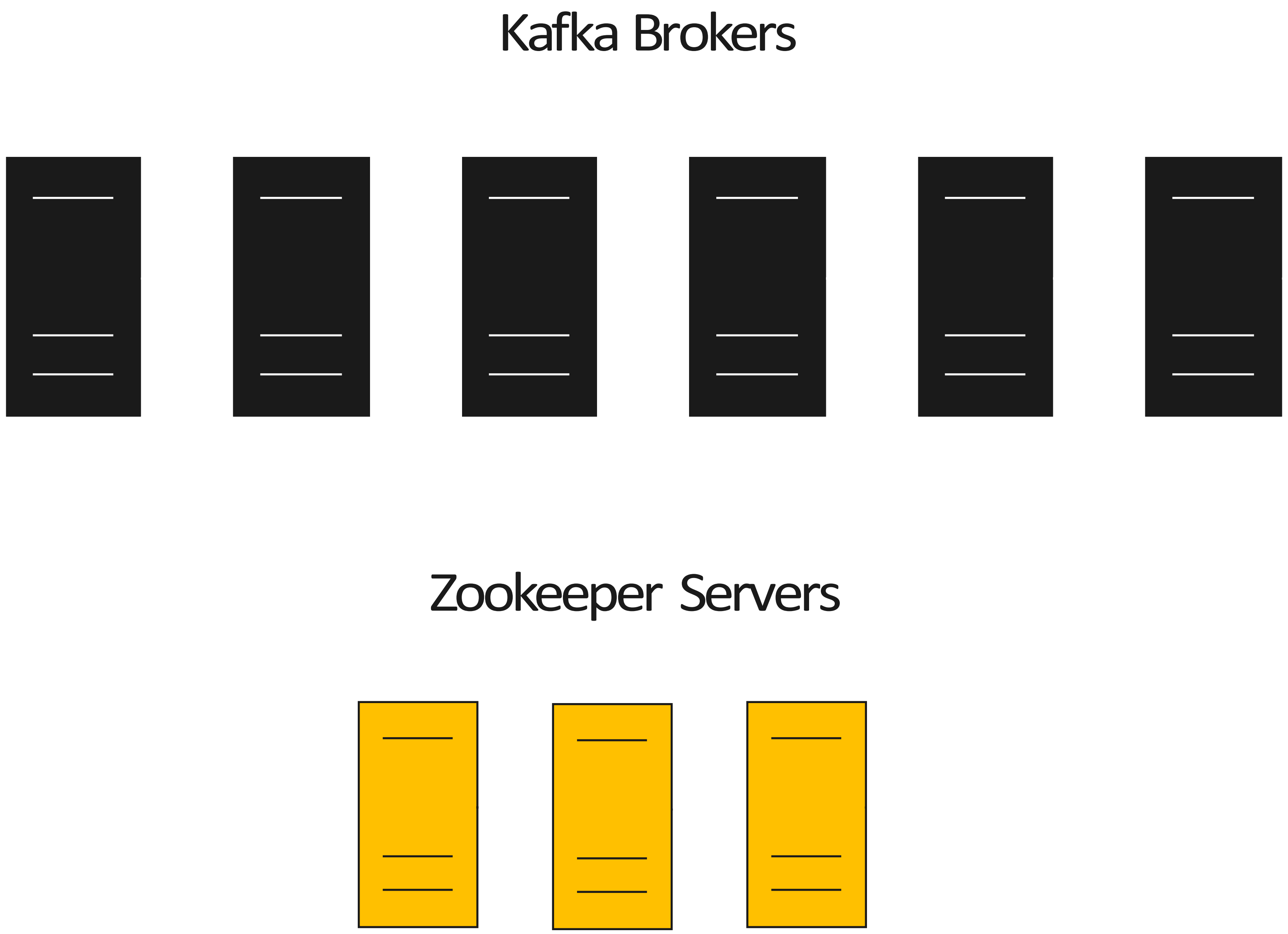
@lenadroid

Challenges:

Simplicity, Scalability, Reliability

Meet Apache Kafka

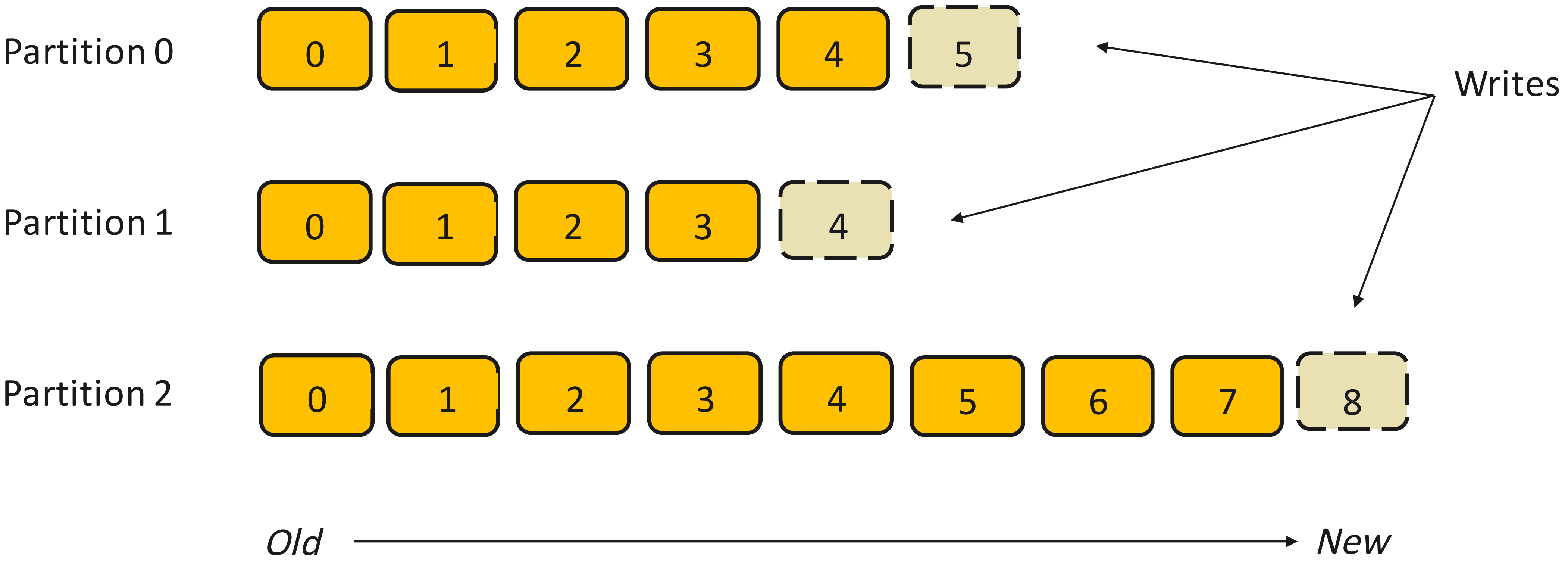
Apache Kafka is an open-source stream-processing software platform developed by the Apache Software Foundation written in Scala and Java.



Kafka Brokers

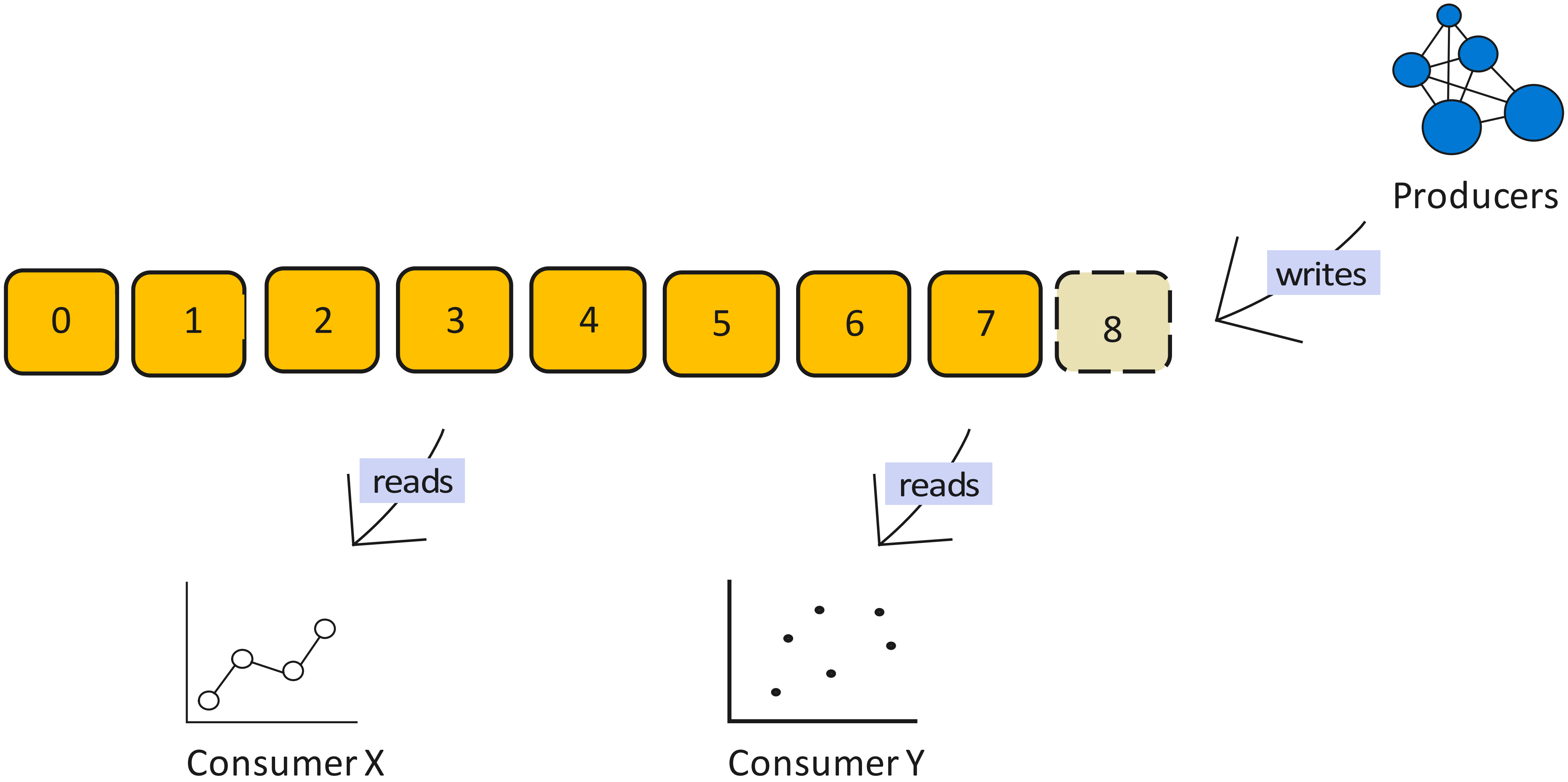
Zookeeper Servers

Inside of a Kafka Topic



@lenadroid

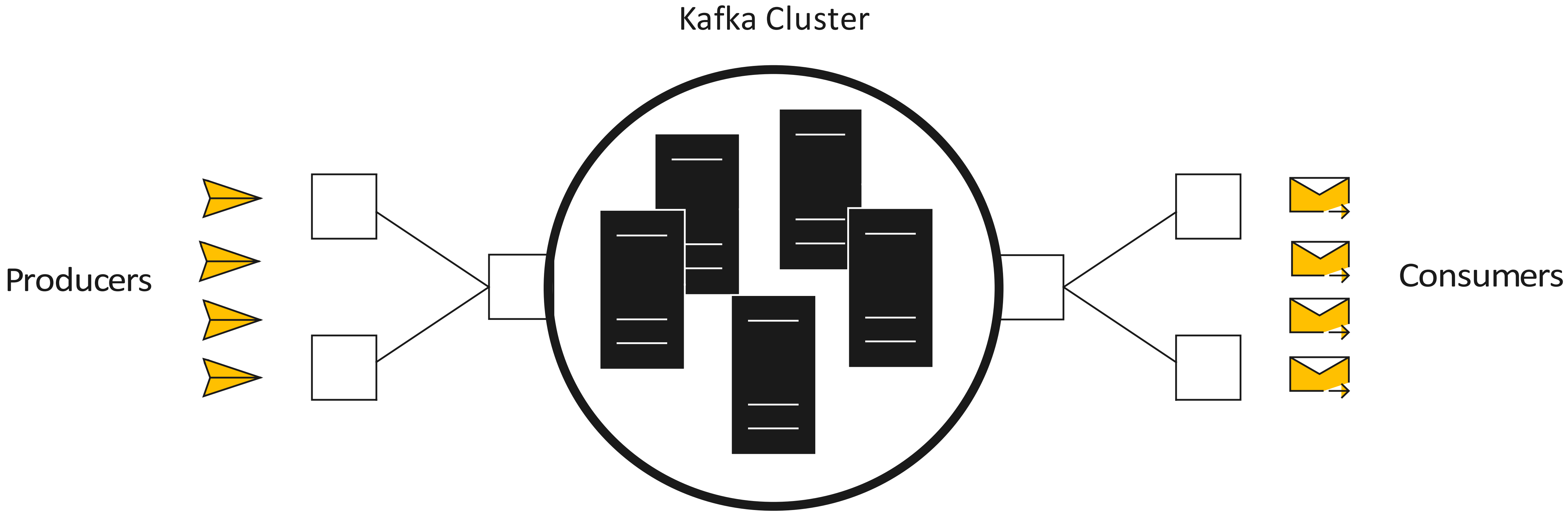
Kafka Topic Partition



@lenadroid

Create a Kafka topic

Kafka Producers and Consumers



@lenadroid

Meet Apache Spark

Apache Spark is a unified analytics engine for large-scale data processing: batch, streaming, machine learning, graph computation. Access data in hundreds of data sources.

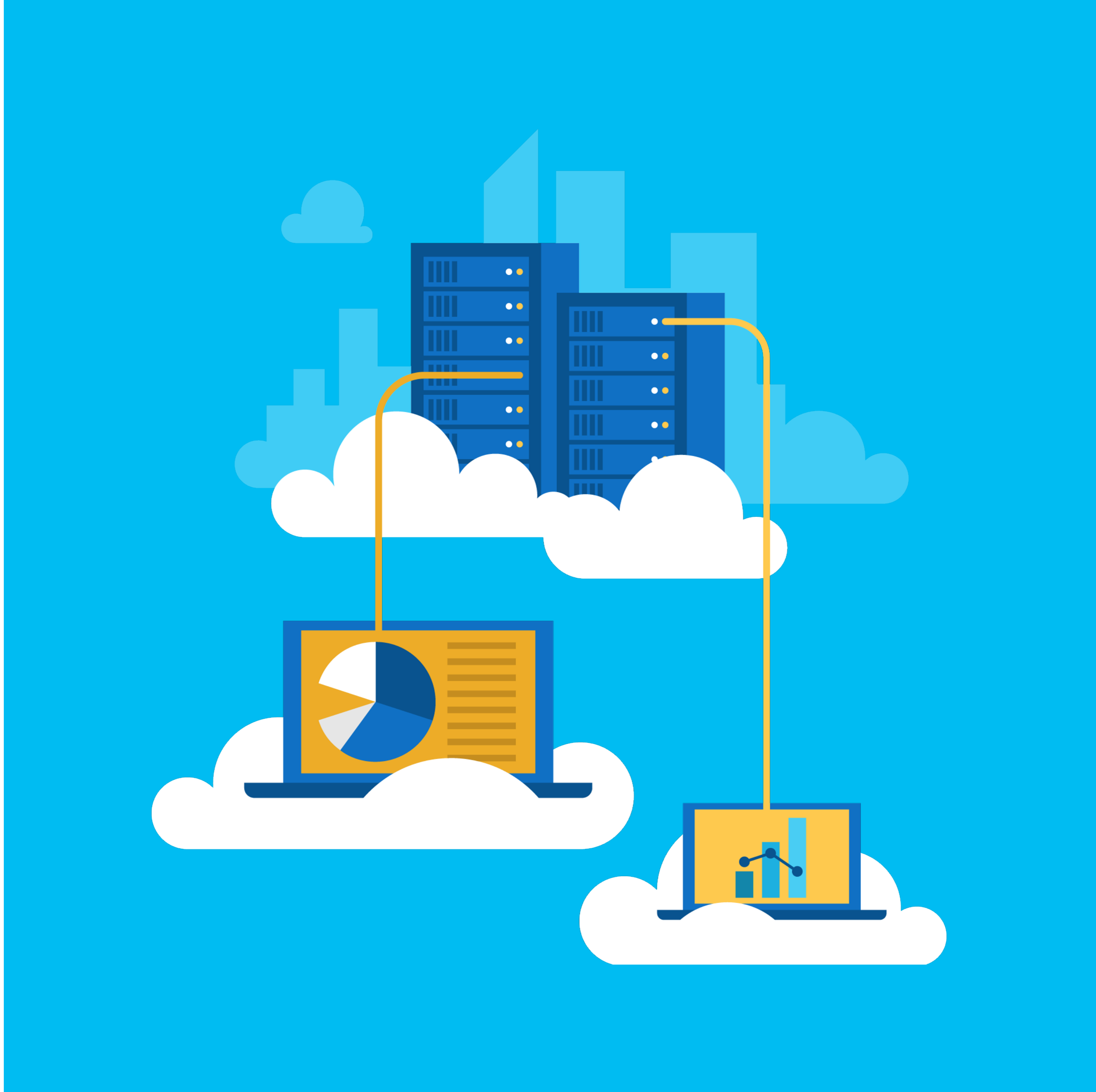
What Apache Spark can do

- Spark SQL and batch processing
- Stream processing with Spark Streaming and Structured Streaming
- Machine Learning with Mllib
- Graph computations with GraphX

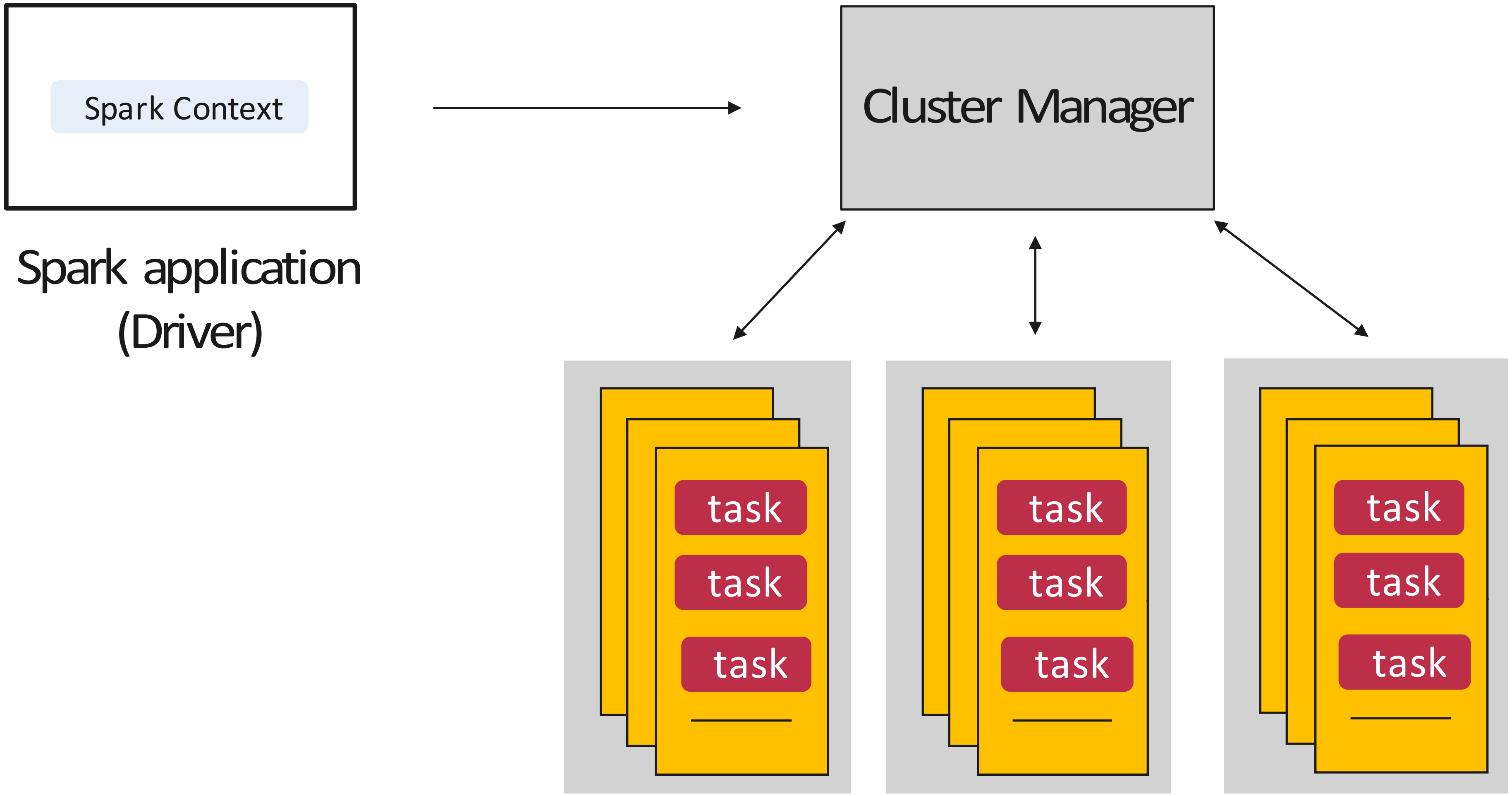
Spark program

```
val textFile = sc.textFile("hdfs://...")

val counts =
     textFile
     .flatMap(line => line.split(" "))
     .map(word => (word, 1))
     .reduceByKey(_ + _)

counts.saveAsTextFile("hdfs://...")
```

How does Spark work?



@lenadroid



Spark Context

Spark application (Driver)

Cluster Manager

task

task

task

task

task

task

task

task

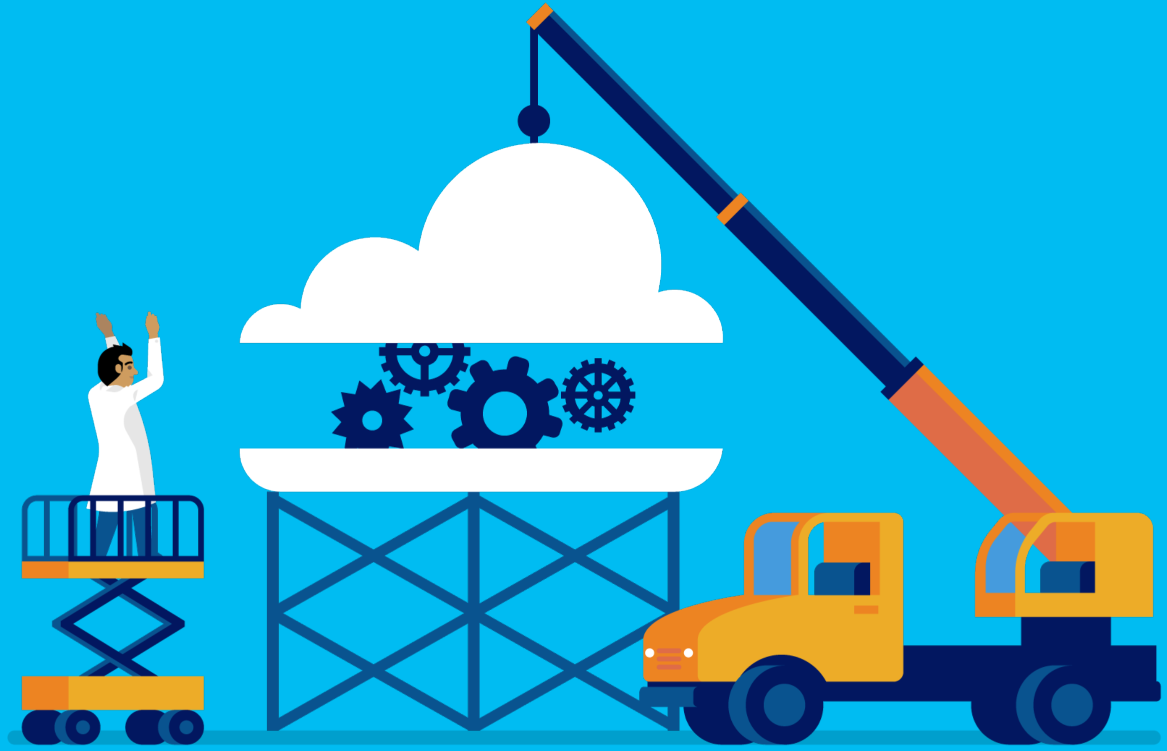
task

Spark workers have executors of tasks

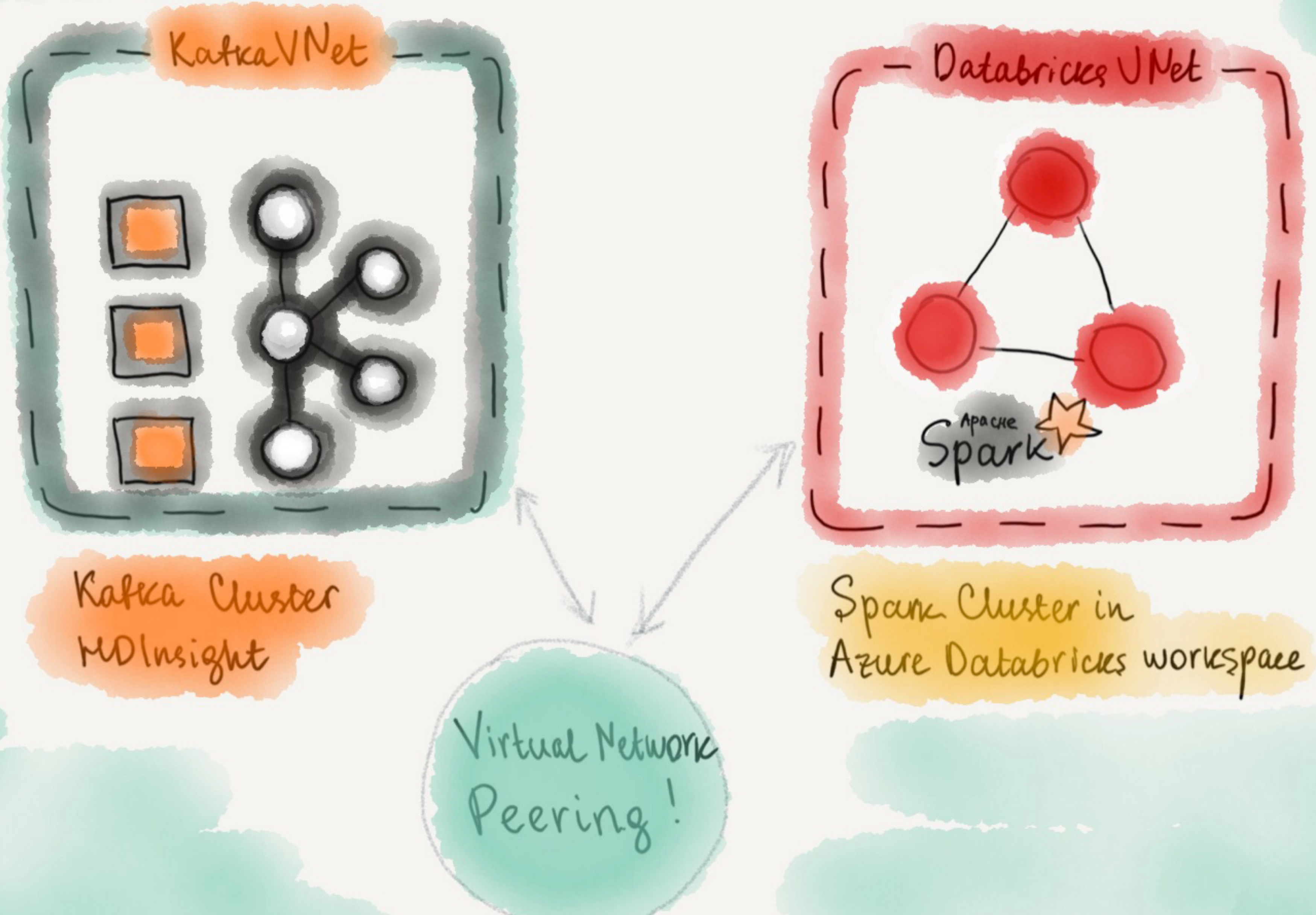
@lenadroid

How to use Apache Kafka and Spark on Azure?

HDInsight and Azure Databricks



@lenadroid


KafkaVNet
DatabricksVNet
Apache Spark
Kafka Cluster HDInsight
Spark Cluster in Azure Databricks workspace
Virtual Network Peering !

Existing infrastructure and resources

- HDInsight Kafka cluster
- Azure Databricks workspace with a Spark cluster
- Kafka and Spark Virtual Networks peered together
- Used sources of data:
  - Public dataset files saved on Azure storage account
  - Twitter data

Example: Processing a stream of events from Twitter using Apache Kafka and Spark

Part 1: Kafka Producer

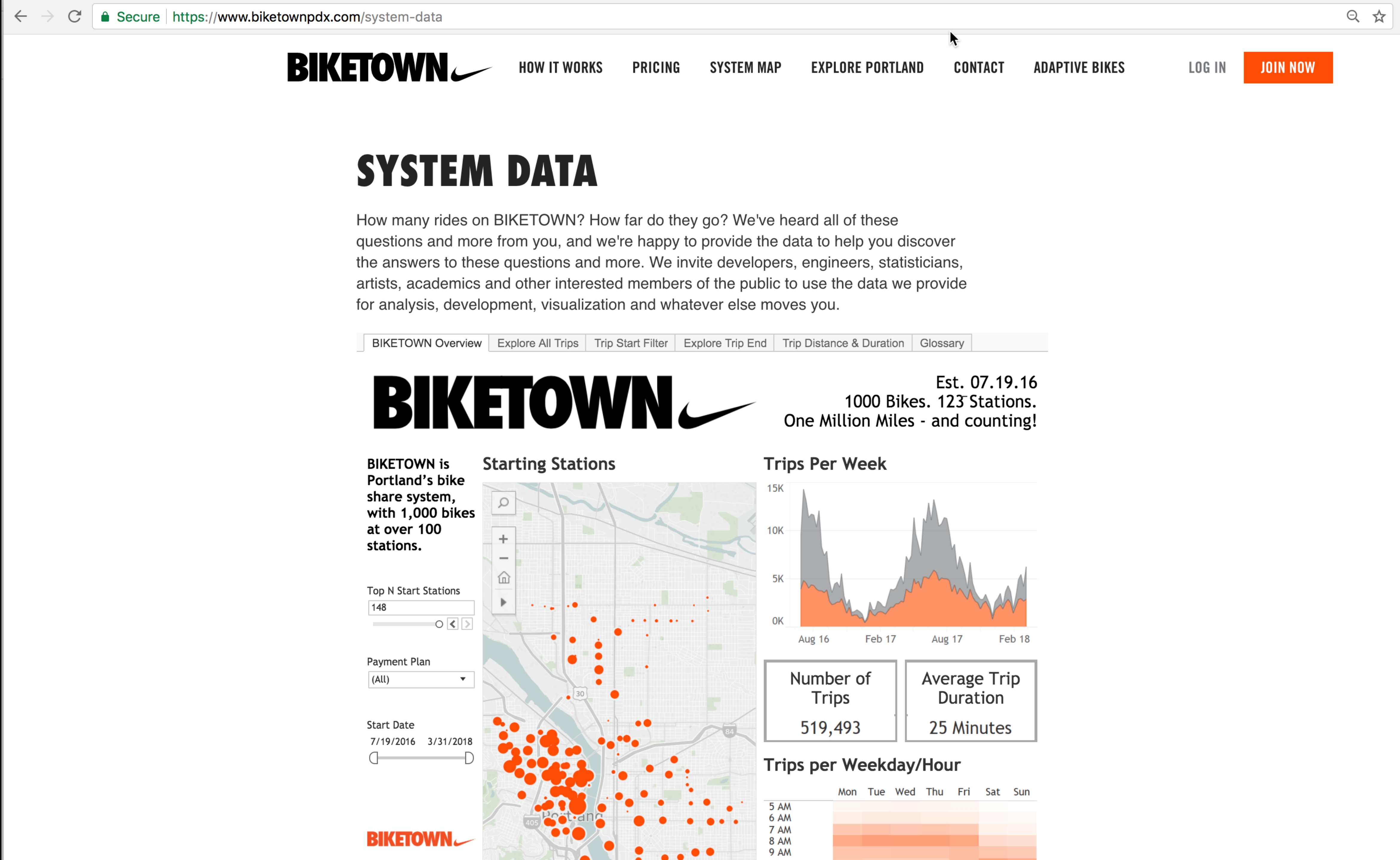
Part 2: Spark Consumer

Kafka + Spark

=

Reliable, scalable event ingestion and real-time stream processing

Example: Analyzing a public dataset using Apache Kafka and Spark



Secure https://www.biketownpdx.com/system-data

BIKETOWN HOW IT WORKS PRICING SYSTEM MAP EXPLORE PORTLAND CONTACT ADAPTIVE BIKES LOG IN JOIN NOW

SYSTEM DATA

How many rides on BIKETOWN? How far do they go? We've heard all of these questions and more from you, and we're happy to provide the data to help you discover the answers to these questions and more. We invite developers, engineers, statisticians, artists, academics and other interested members of the public to use the data we provide for analysis, development, visualization and whatever else moves you.

BIKETOWN Overview | Explore All Trips | Trip Start Filter | Explore Trip End | Trip Distance & Duration | Glossary

BIKETOWN

Est. 07.19.16
1000 Bikes. 123 Stations.
One Million Miles - and counting!

BIKETOWN is Portland's bike share system, with 1,000 bikes at over 100 stations.

Top N Start Stations: 148

Payment Plan: (All)

Start Date: 7/19/2016 – 3/31/2018

Starting Stations

Trips Per Week

Number of Trips: 519,493

Average Trip Duration: 25 Minutes

Trips per Weekday/Hour

BIKETOWN

Exploratory data analysis

With Azure Databricks and Spark

@lenadroid

Processing real-time streams of trip data and making decisions

With Kafka and Spark

@lenadroid

Part 1: Kafka Producer

Part 2: Spark Consumer

Kafka + Spark
=
Reliable, scalable event ingestion and real-time stream processing

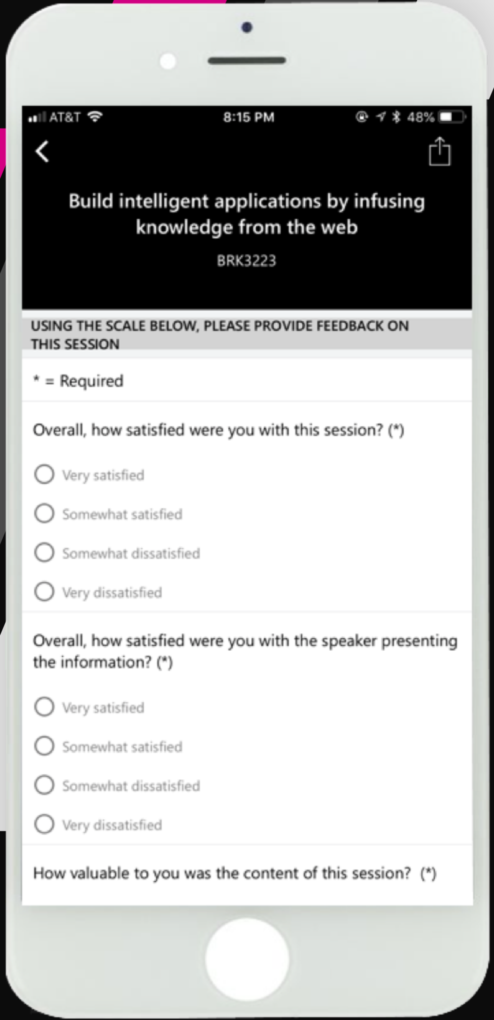
Thank you!

- Apache Kafka: aka.ms/apache-kafka
- Apache Spark: aka.ms/apache-spark
- Event stream processing architecture on Azure with Apache Kafka and Spark: aka.ms/kafka-spark-azure
- Create HDInsight Kafka cluster using ARM: aka.ms/hdi-kafka-arm
- Create Kafka topics in HDInsight: aka.ms/hdi-kafka-topic

- Lena on Twitter: twitter.com/lenadroid
- Lena on Github: github.com/lenadroid

Please evaluate this session

Your feedback is important to us



Build intelligent applications by infusing knowledge from the web

BRK3223

USING THE SCALE BELOW, PLEASE PROVIDE FEEDBACK ON THIS SESSION

* = Required

Overall, how satisfied were you with this session? (*)

○ Very satisfied
○ Somewhat satisfied
○ Somewhat dissatisfied
○ Very dissatisfied

Overall, how satisfied were you with the speaker presenting the information? (*)

○ Very satisfied
○ Somewhat satisfied
○ Somewhat dissatisfied
○ Very dissatisfied

How valuable to you was the content of this session? (*)

Microsoft