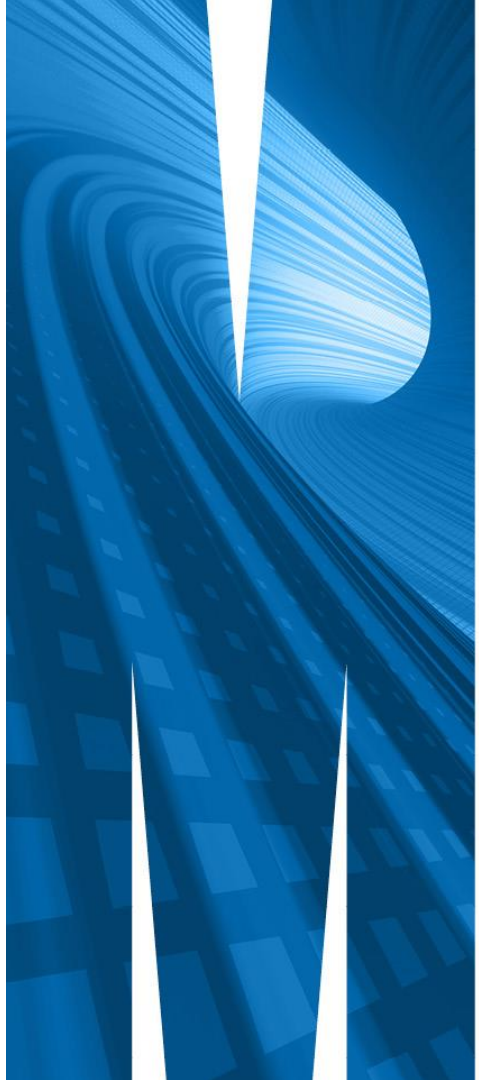
MONASH University

Session 09

FIT5202 Big Data Processing

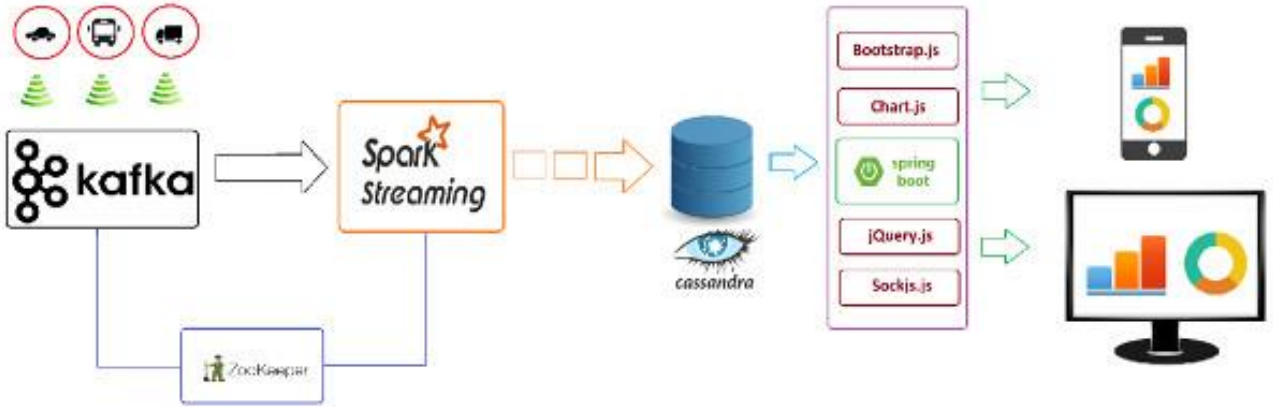
Data Streaming using Apache Kafka and Spark



Session 09 Agenda

- Session 08 Review
  - Implicit vs Explicit Data
  - Matrix Factorization
  - Collaborative Filtering with ALS
- Streaming using Apache Kafka
  - Kafka Producer
  - Kafka Consumer
    - Visualizing in real-time
  - **Use case : Click stream visualization**
- Spark Streaming Basics
  - Demo : word count
  - **Lab Task : Click Stream Analysis and Visualization**

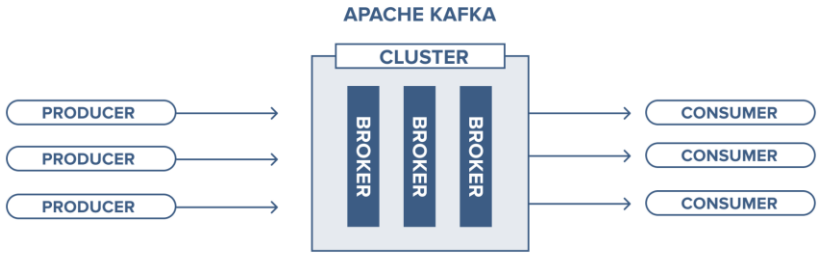
Kafka Use Case (Traffic Data Monitoring)



https://www.infoq.com/articles/traffic-data-monitoring-iot-kafka-and-spark-streaming/

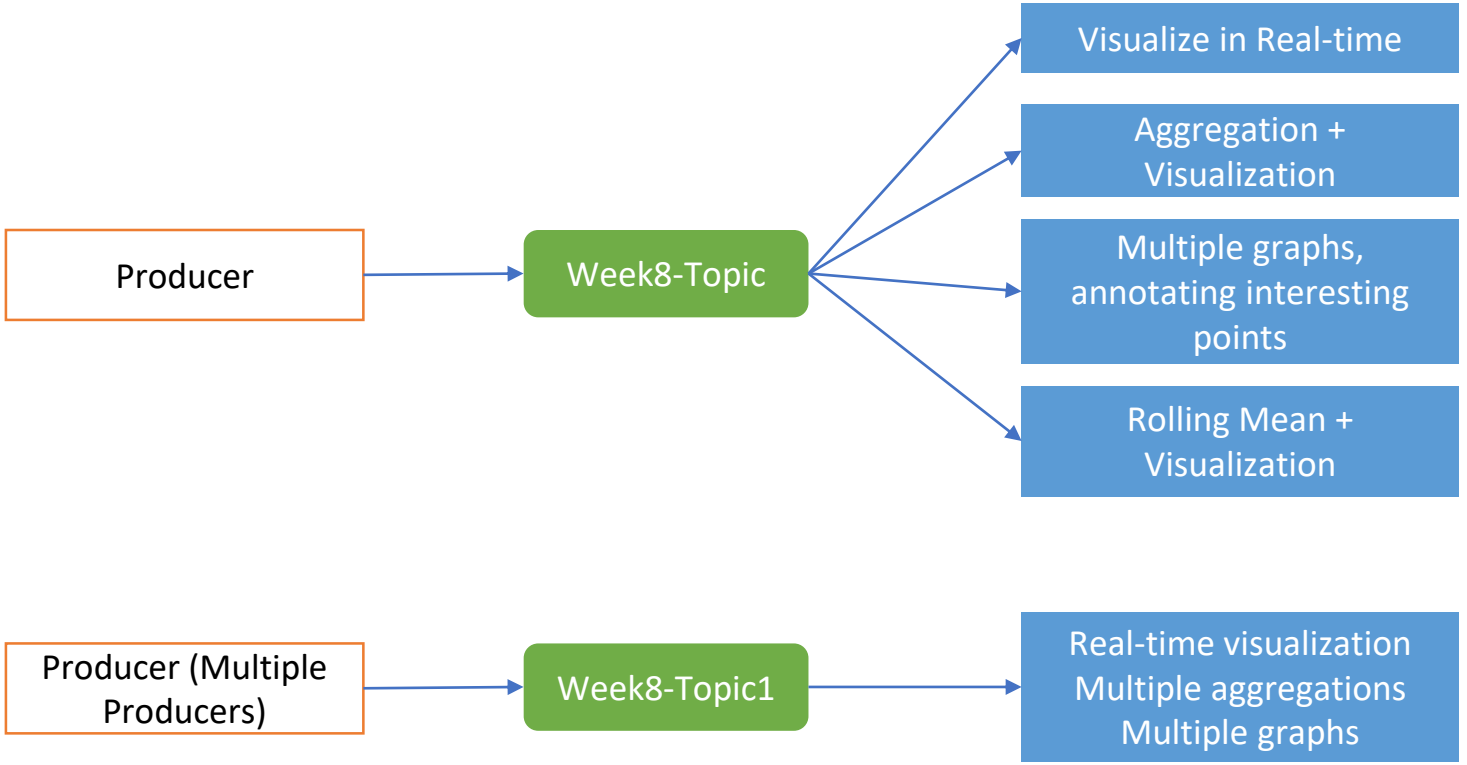
What is Apache Kafka?

- Publish-subscribe messaging system
- Enables distributed applications
- Brokers utilize Apache ZooKeeper for management and coordination of the cluster
- Each broker instance is capable of handling read and write quantities reaching to the hundreds of thousands each second (and terabytes of messages) without any impact on performance.



https://www.cloudkarafka.com/blog/2016-11-30-part1-kafka-for-beginners-what-is-apache-kafka.html
https://www.instaclustr.com/apache-kafka-architecture/

DEMO Kafka Implementation Scenarios for Lab



Kafka Producer and Consumer Properties

- **KafkaProducer**
  - Bootstrap_servers
  - Value_serializer
  - Api_version

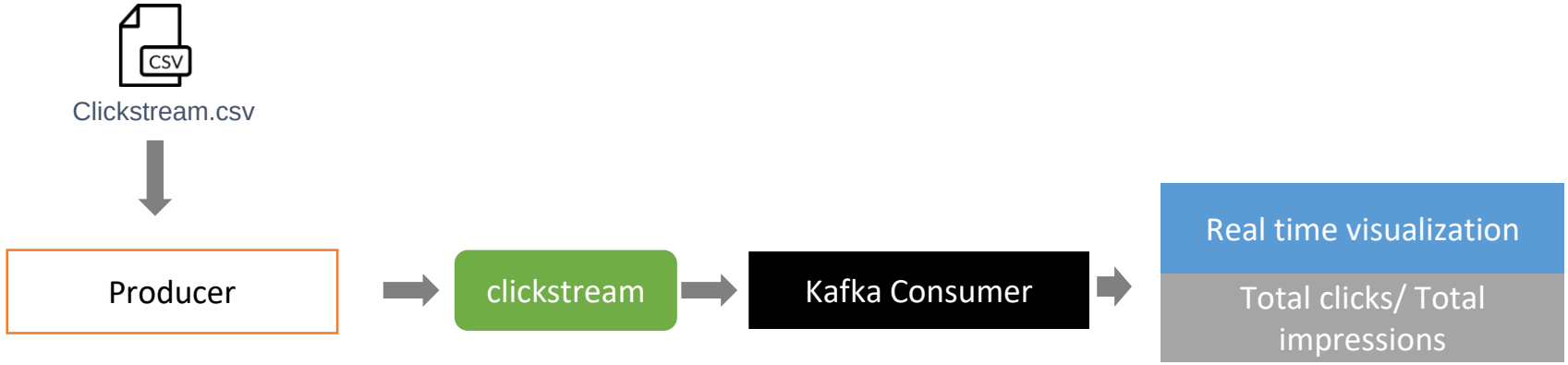
- **KafkaConsumer**
  - Consumer_timeout_ms
  - Auto_offset_reset
  - Bootstrap_servers
  - Value_deserializer
  - Api_version

```
_producer = KafkaProducer(bootstrap_servers=['localhost:9092'],
                          value_serializer=lambda x: dumps(x).encode('ascii'),
                          api_version=(0, 10))
```

```
_consumer = KafkaConsumer(topic,
                          consumer_timeout_ms=10000, # stop iteration if no message aft
                          auto_offset_reset='latest', # comment this if you don't want
                          bootstrap_servers=['localhost:9092'],
                          value_deserializer=lambda x: loads(x.decode('ascii')),
                          api_version=(0, 10))
```


https://kafka-python.readthedocs.io/en/master/apidoc/KafkaConsumer.html

Lab Task for Kafka

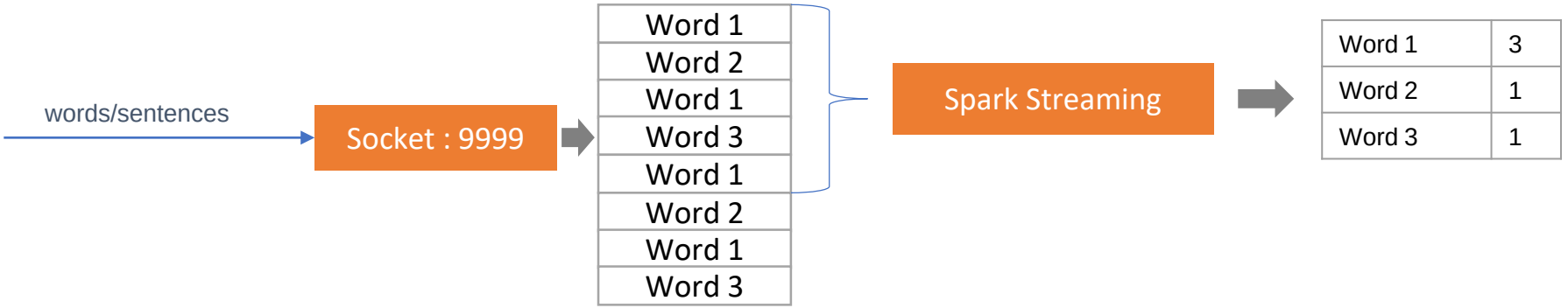


Spark Structured Streaming

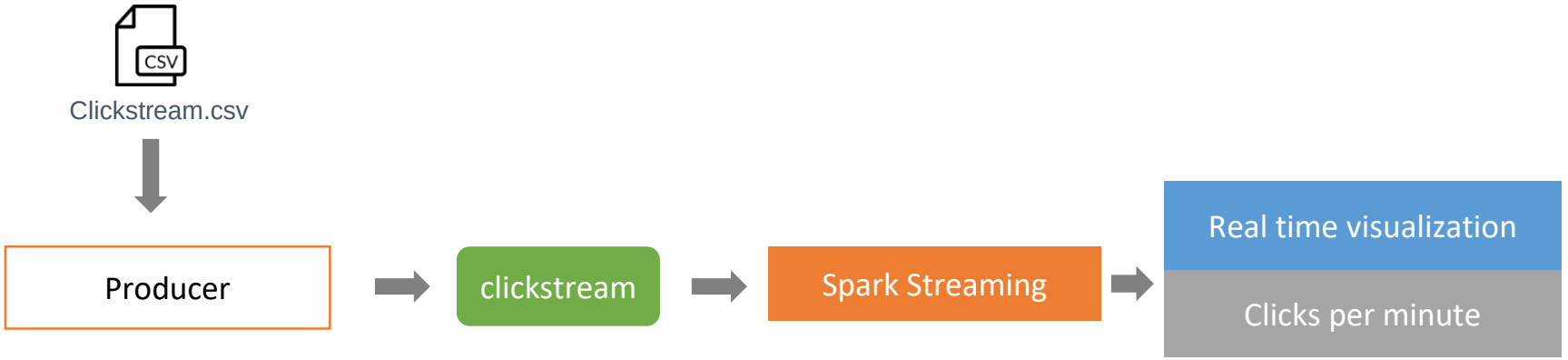
To be covered in Session 10 Lecture

DEMO Spark Structured Streaming

Word Count Demo



Lab Task for Spark Structured Streaming



Thank You!

See you next week.