




Introduction to Kafka Stream Processing

Travis Deason and Michael Landrum

<https://github.com/tdeason416/esports>
DS7330 April 25th 2018



What is Kafka

- Kafka is a message system
 - Acts like a database where the only indices for the data are sequential integers
- Messages are sent to Kafka via one or many producers
- Messages are read from Kafka via one or many consumers

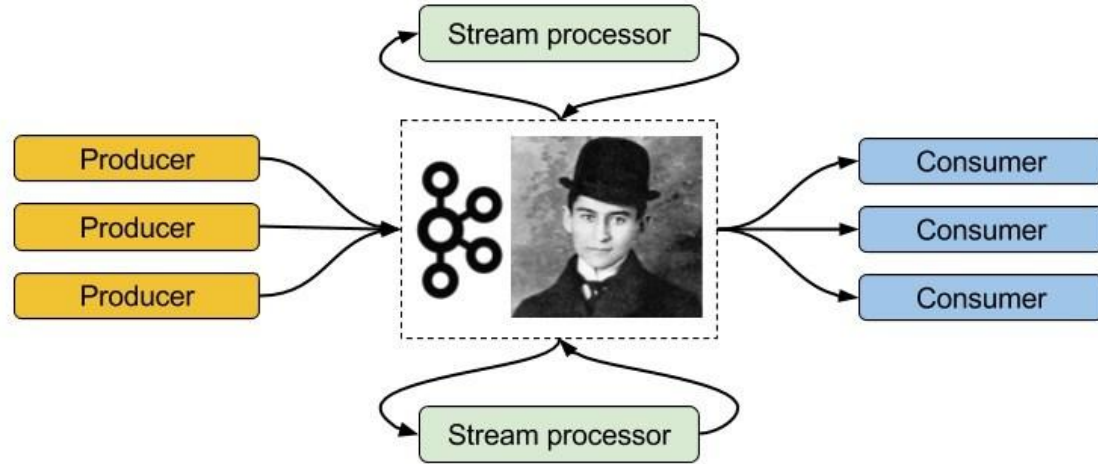
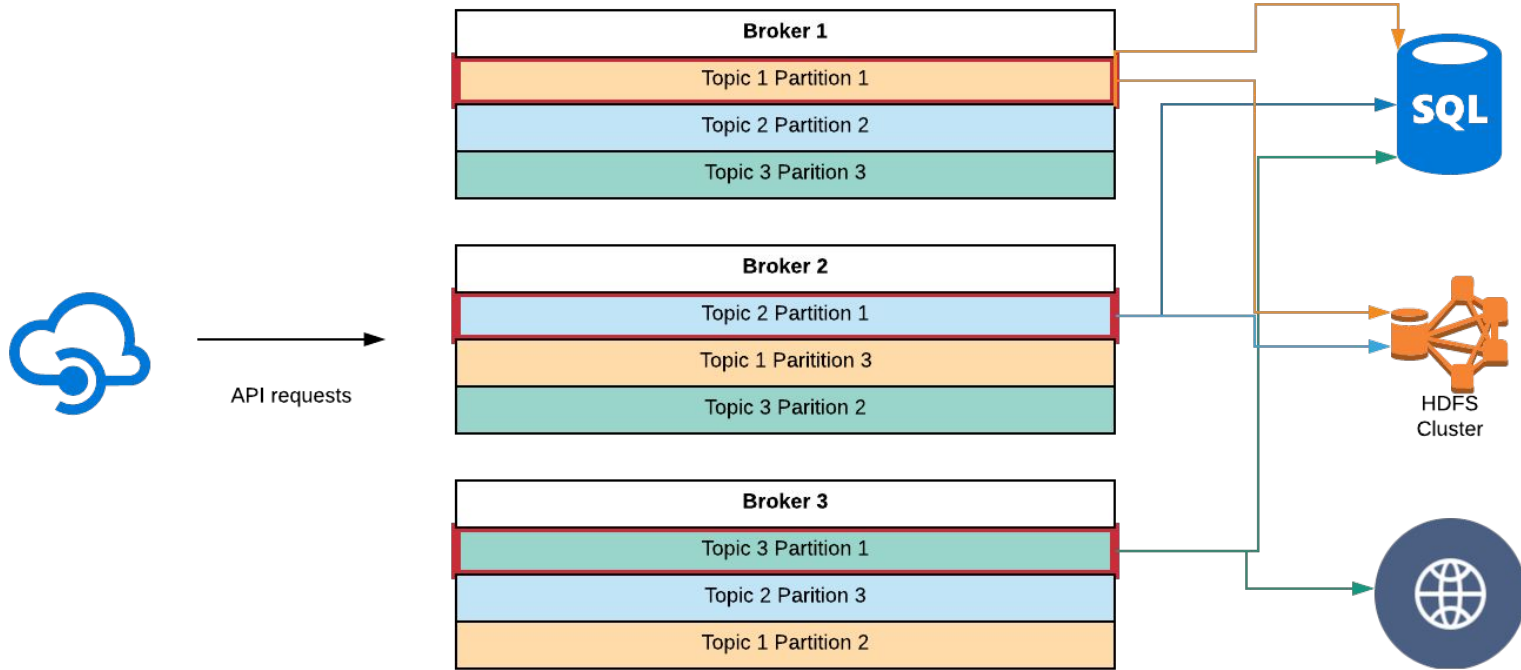


Photo ref:

<https://i1.wp.com/codeblog.dotsandbrackets.com/wp-content/uploads/2016/11/kafka-1.jpg?resize=745%2C374>

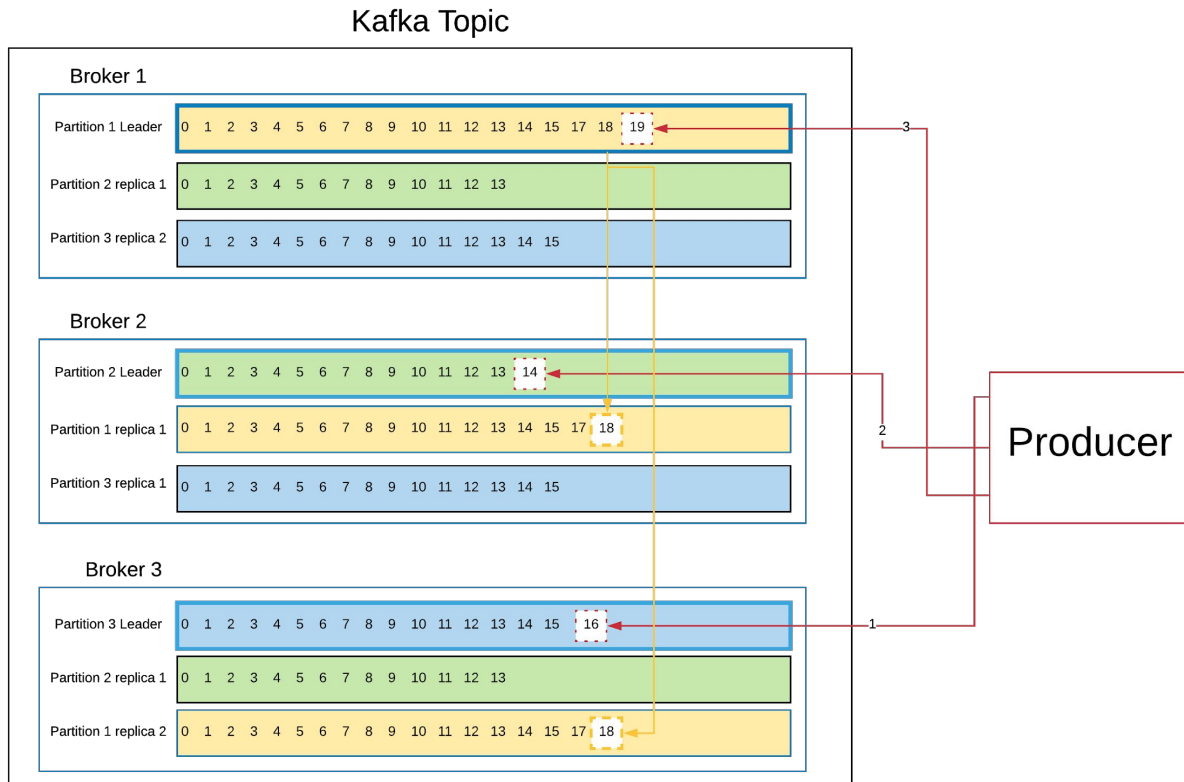
Shows a basic concept of what Kafka is used for, and also has a photo of its German philosopher namesake, Franz Kafka

How Does Kafka Work



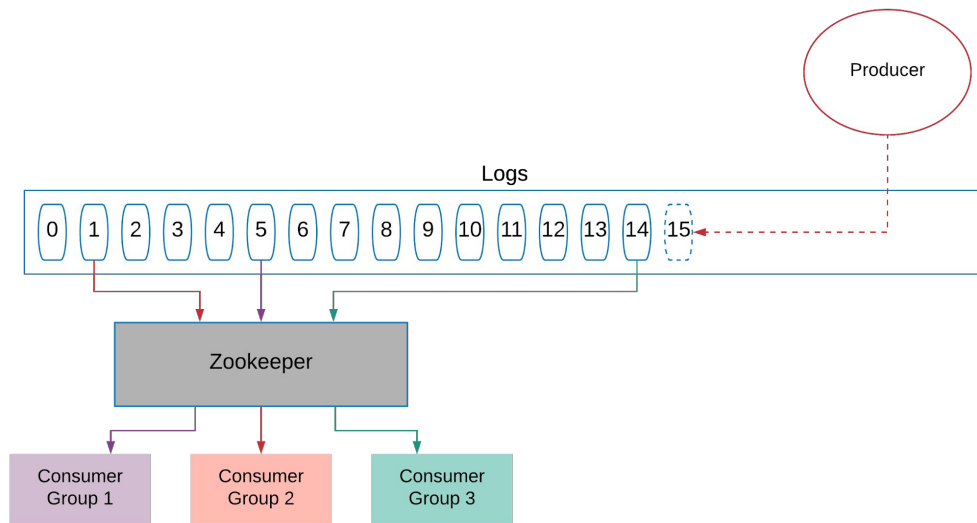
Topics Partitions Offsets and Replicas

- A Kafka topic can be split into multiple partitions
 - Each partition will contain data unique to that partition
 - Each piece of data in the partition can be identified using an integer value called an offset
- Partitions can be duplicated by the broker for durability



Logs

- Kafka Zookeeper keeps a log of which consumers have consumed which offsets
 - If a consumer is multiple machines, these can be grouped into a consumer group
- Logs allow each consumer to read each message exactly once
 - So long as each offset is consumed by only one consumer within a consumer group



Baseball Streaming Project

- To demonstrate a practical application of Kafka, we built a streaming application based which taps into the Major League baseball streaming API available on Sportradar
- Source code for this project is at
 - <https://github.com/tdeason416/esports>

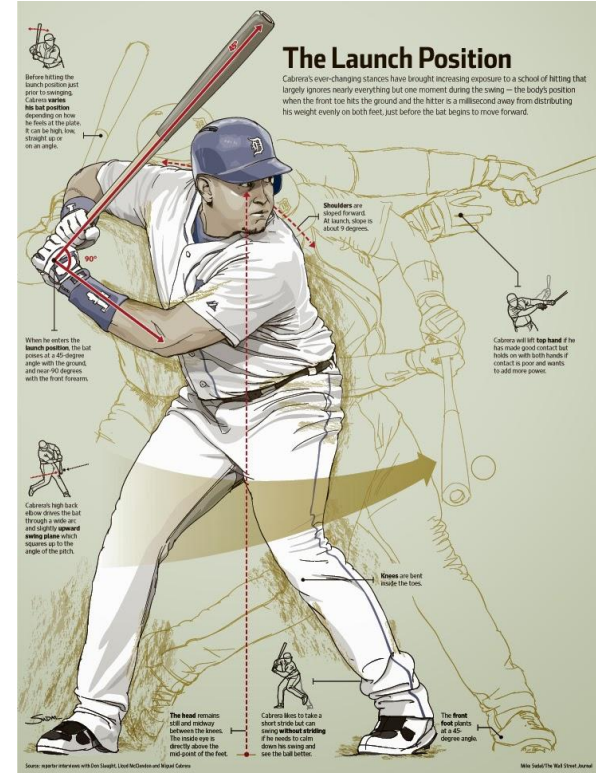


Photo ref:

<https://www.pinterest.com/pin/42291683975309936/>

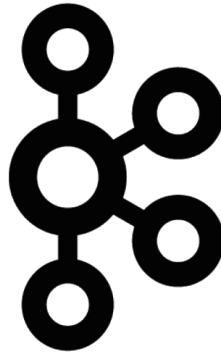
Project Technology



Sportradar API



Airflow
Managed
Python
Producer



Kafka



Airflow
Managed
Python
Consumer



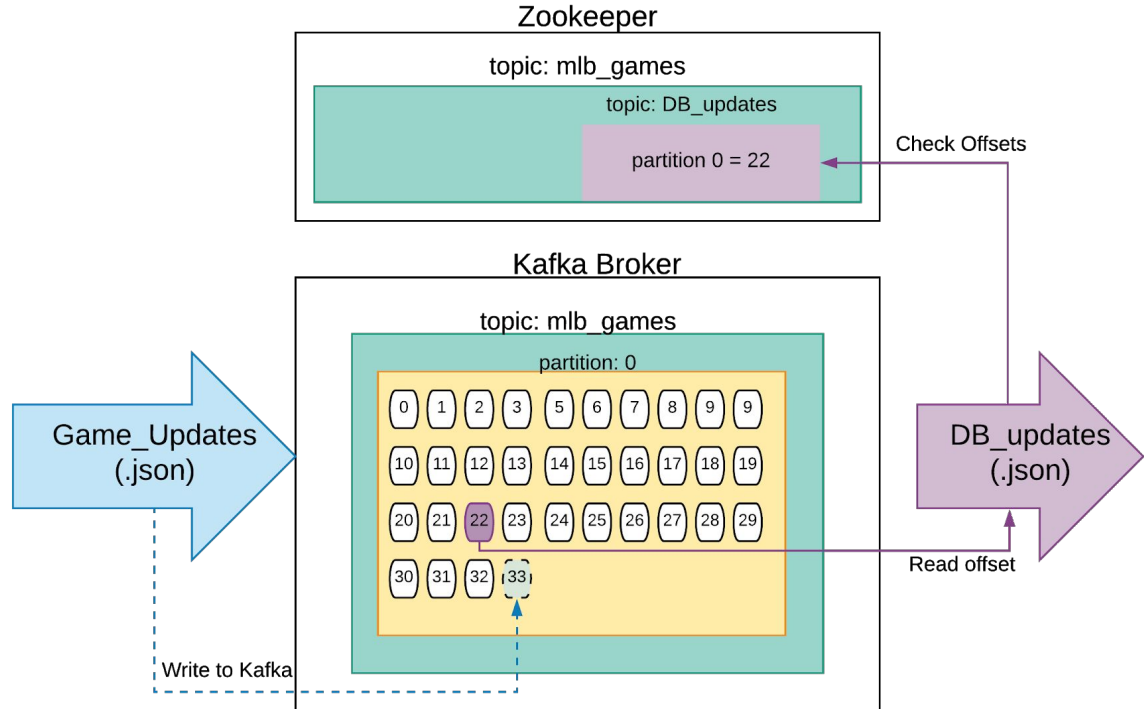
mongoDB

Mongo
Datastore

MLB Kafka Architecture

The Kafka cluster consists of a single node which runs both Zookeeper and the Broker

- Mlb_games is the only topic
- There is only one partition



Questions