Aufgabe 3: Kafka mit Partitionen und Consumer-Gruppen

(a) Tankerkönig

Tankerkönig stellt dankenwerterweise auch historische Daten via Git zur Verfügung. Das Datenformat ist auf der Azure-Webseite des Projekts dokumentiert.
Für die Vorlesung wurde das Preisformat sinnvoll reduziert und um die Angabe der Postleitzahl der Tankstelle erweitert. Unt derm Topic `tankerkoenig` sind historische Daten von Anfang 2022 abrufbar.

- **Machen Sie sich mit dem Datenformat vertraut und lesen Sie es via Konsolen-Tool (kcat, …) aus.**
  Beispiel:
  {"date":"2022-01-02T21:01:07.000+00:00","station":"89c931b5-2ddd-4a90-969e-ae250143ddf3","postCode":"15517","pDiesel":1.529,"pE5":1.759,"pE10":1.699}

  Daten werden über Kafka im JSON-Format ausgegeben. Dabei erhält jede Station eine eindeutige ID „station“.

- **Wie sind die Daten organisiert? Achten Sie dabei auf die Einteilung der Partitionen.**
  Jede Partition hält die Daten für einen Postleitzahlenbereich 0-9…
  Befehl mit kafkacat, um Daten einer Topic zu konsumieren:
  `kcat -C -b 10.50.15.52:9092 -t tankerkoenig -p <partition>`

- **Wie werden Partitionen hier verwendet?**
  Die Partitionen werden verwendet, um Daten in sinnvolle Kategorien (Postleitzahlenbereiche) einzuteilen und aufzusplitten. Jede Partition hält dabei die Daten für einen Postleitzahlenbereich.

- **Nützt Ihnen die Einteilung der Daten in den Partitionen für die nächste Aufgabe?**
  Durch die Aufteilung in verschiedene Partitionen, kann ein parallelisiertes Programm entwickelt werden, das für jeden Postleitzahlenbereich einen eigenen Consumer bereitstellt, der die Daten liest und aggregiert.

(b) Datenverarbeitung

Programmieren Sie einen Consumer, der die Daten von Tankerkönig liest und pro Postleizsahlenbereich (und pro Preiskategorie) eine Aggregation aller Daten macht. Dies bedeutet, dass die Preisdaten pro Postleitzahlenbereich (0-9) und Preiskategorie gemittelt und dann auf beispielsweise stündliche Werte zusammengefasst werden sollen.
Diese Daten sollen dann wieder in Grafana visualisiert werden, sodass man die Entwicklung der Preise über die Monate unter den Postleitzahlenbereichen vergleichen kann.

Da die Datenmenge pro Partitoin recht hoch ist, empfiehlt es sich, Ihren Consumer auf der Epyc-Maschine laufen zu lassen.

- **Schreiben Sie ein Programm, das die Partitionen parallel verarbeiten kann (also pro Parition ein Thread). Sie können selbstverständlich auch gerne Kafka Streams dafür verwenden, oder die normale Consumer-API. Hierzu können Sie entweder die zufällige Zuordnung innerhalb einer Consumer-Gruppe nutzen, oder besser die jeweiligen Consumer fet an eine Partition binden.**
  siehe Quelldateien

- **Visualisieren Sie die Daten mit Grafana, sowohl einzeln pro Postleitzahlenbereich, als auch in einer Grafik pro Preiskategorie, die dafür jeweils all PLZ-Bereiche vergleichend anzeigen soll. Dies sind dann in Summe 13 Screenshots (10 PLZ-Kategorien + 3 Preiskategorien).**
  siehe Abgabgecontainer

- **Geben Sie Ihren Code und die Screenshots (mit Gruppen-Id, bzw. Matrikelnummern im Titel des Graphen) im Abgabecontainer ab.**
  siehe Abgabecontainert