Data Engineering mit Apache Kafka

Projektbericht

von

Johannes Weber

Matrikelnummer: 11010021

und

Julian Ruppel

Matrikelnummer: 11010020

09.02.2018

SRH Heidelberg

Fakultät für Information, Medien und Design

Big Data und Business Analytics

Dozent

Frank Schulz

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Aufgabe und Ziel

Im Rahmen dieses Projektes war es die Zielstellung sich mit Data Ingestion, Data Storage sowie Data Retrieval vertraut zu machen.

Data Ingestion ist die Beschaffung der Daten. Dies kann entweder mit Hilfe eines Data Streams erfolgen oder einer statischen Datenquelle - also eine Datei die lokal auf einem Rechner abgelegt wird wie z. B. eine Comma-separated values (CSV) oderJavaScript Object Notation (JSON) Datei.

Unter einem Data Stream versteht man einen kontinuierlichen Datenstrom wie z. B. die Erstellung von immer wieder neuen Twitter Nachrichten. Ein wichtiges Merkmal eines Data Streams ist, dass nicht vorherzusehen ist wann der Datenstom zu Ende ist - er könnte theoretisch unendlich sein. Im Falle des Twitter Datenstroms ist es nicht abzusehen wann die letzte Twitter Nachricht geschrieben wird.

Für unsere Aufgabe ist darauf zu achten, dass der Datenstrom über eine Application Programming Interface (API) öffentlich zugänglich.

Data Storage ist die Speicherung der Daten. Hierbei wurde uns die Anforderung gestellt, das für die Speicherung der Daten die Streaming Plattform Apache Kafka verwende wird. Des Weiteren war es gestattet die Daten zusätzlich in einer relationen Datenbank, NoSQL Datenbank oder mit Spark Streaming zu speichern.

Data Retrieval ist die Beschaffung der Daten aus einer Datenbank mit Structured Query Language (SQL) Abfragen und die abschließende Ausgabe der Ergebnisse in Form von Tabellen oder einfachen Visualisierungen. Es sollten mindestens drei verschieden SQL Abfragen abgesetzt werden mit unterschiedlichen Filter- und Agreggatsfunktionen sowie einer Teilaggregation wie z. B. GROUP BY. Die Visualisierung der Daten sollte in einem vir-

tuellen Notebook erfolgen. Als virtuelles Notebook durften wir uns entscheiden zwischen Apache Zeppelin oder Jupyter.

Unsere Aufgabe ist es eine geeignete Vorgehensweise für die Bewältigung dieser Aufgabe zu finden und umzusetzen.

Wir entschieden uns für unser Szenario Daten von NYC Open Data zu nutzen. NYC Open Data gibt allen „New Yorkern“ und somit auch der ganzen Welt, die Chance, Open Data, also frei zugängliche Daten, einfach zu konsumieren..[1]

NYC Open Data ermöglicht es sowohl einen kontinuierlichen Data Stream als auch eine statische CSV Datei zu konsumieren. Dank diesem Umstand entschieden wir uns im Rahmen dieses Projektes beide Möglichkeiten umzusetzen und zu vergleichen. Auch bei Data Retrieval entschieden wir uns dafür sowohl Apache Zeppelin als auch Jupyter zu nutzen und zu vergleichen.

Kapitel 2 - *Werkzeuge und technische Rahmenbedingungen* beschäftigt sich detailierter mit den verwendeten Tools und Programmiersprachen. Kapitel 3 - *Lösungsansatz* erläutert das gewählte Szenario und beleuchtet die Themen Data Ingestion, Data Storage und Data Retrieval sowie die verwendeten Architektur.

[1] NYC Open Data. *Our mission: open data for all*. 2017. URL: https://opendata.cityofnewyork.us/overview/ (besucht am 03. 02. 2018).

Kapitel 2

Werkzeuge und technische Rahmenbedingungen

Im folgenden Abschnitt werden einige wichtige technische Werkzeuge, die im späteren Verlauf eingesetzt werden, kurz und prägnant erläutert.

Java ist eine objektorientierte Open-Source Programmiersprache die ursprünglich von Sun Microsystems entwickelt wurde und heute zum Oracle Konzern gehört. In den letzten beiden Hauptversionen wurde der Sprachumfang um funktionale und reaktive Aspekte erweitert. Java ist dank der *Java Virtual Maschine* (JVM) als Laufzeitumgebung plattformunabhängig und hat sich vorwiegend in Enterprise Systemen und Web-Backends etabliert. Im Zuge der Verbreitung von BigData Projekten unter dem Dach der Apache Software Foundation, allem voran Hadoop und Spark, werden JVM sprachen wie Java und Scala nun auch im Bereich BigData eingesetzt.

Python ist eine Open-Source Skripsprache, die sich hauptsächlich durch eine gut lesbare und knappe Syntax auszeichnet und unter anderem das objektorientierte und funktionale Programmierparadigma unterstützt. Im Gegensatz zu Java ist Python dynamisch typisiert und wird interpretiert anstatt kompiliert. Dank eines sehr umfangreichen und ausgereiften Ökosystems aus Frameworks und Bibliotheken zur Datenanalyse und maschinelles Lernen[1] ist Python im Bereich BigData und dank der minimalinvasiven Eigenschaften zum Rapid Prototyping beliebt.

[1] z.B. TensorFlow von Google

Apache Kafka ist eine verteilte Data-Streaming Plattform der Apache Software Foundation, die ursprünglich von LinkedIn entworfen wurde. Beliebt ist Kafka im BigData Umfeld wegen seiner Skalierbarkeit und Fehlertoleranz. Zu den Einsatzszenarien zählen vor allem Stream Processing, es kann aber auch als reiner Message Broker oder Speichersystem für Streaming Data verwendet werden. Die wesentlichen Komponenten von Kafka sind ***Producer*** um einen Stream für einen ***Topic*** zu veröffentlichen, ***Kafka Cluster*** um die die Streaming-Daten verteilt pro *Topic* im Dateisystem zu speichern und ***Consumer*** um einen *Topic* zu abonnieren und dessen Nachrichten zu lesen. Zudem können mit ***Kafka Streams*** Nachrichten im Cluster transformiert werden. Mittels ***Kafka Connectors*** kann man per Konfiguration gängige Datenquellen und -senken[2] anschließen und stellen somit eine deklarative alternative zu den imperativen *Producer API* und *Consumer API* dar. 2014 haben sich die verantwortlichen LinkedIn Mitarbeiter vom Mutterkonzern getrennt um sich mit der neu gegründeten Firma Confluent dediziert dem Apache Kafka Ökosystem zu widmen. Entwickelt wurde die quelloffene Software in der JVM-basierten Programmiersprache Scala, welche objektorientierte und funktionale Aspekte vereint.

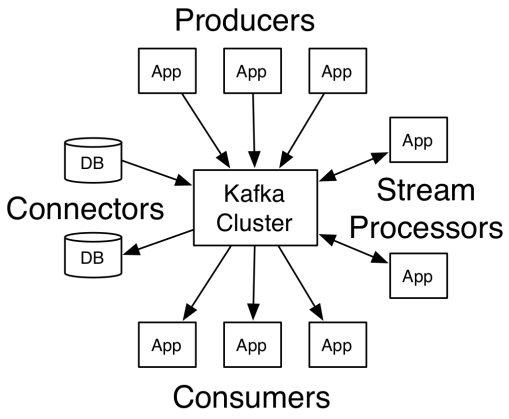

Abbildung 2.1: Apache Kafka Architektur[3]

PostgreSQL ist ein objektrelationales Datenbankmanagementsystem. Das vollständig ACID-konforme und in C geschriebene quelloffene System zeichnet sich durch einen breiten Funktionsumfang, Stabilität, Standardkonformität, hohe Erweiterbarkeit, und als Resultat dessen, eine weite Verbreitung aus. Neben dem traditionellen zeilenorientierten Eigenschaften bietet PostgreSQL zudem Erweiterungen hinsichtlich verteilter, hoch-parallelisierter und spaltenorientierter Datenverarbeitung, ein Geoinformationssystem sowie Volltextsuche. Auch im Bereich NoSQL bietet PostgreSQL eine dokumentenorientierter Speicherung und durch Erweiterungen sogar Graphen und

[2] wie z.B. Twitter oder JDBC

Schlüssel-Werte-Datenstrukturen. Diese Flexibilität eröffnet PostgreSQL vielseitige Einsatzszenarien, darunter sowohl OLTP als auch OLAP.

Apache Zeppelin ist eine Web-basierte Open-Source Software mit der man sog. Notebooks zur datengetriebenen, interaktiven und kollaborativen Analyse erstellen kann. Es werden eine Vielzahl an Speicher- und Analysetechnologien unterstützt, darunter SQL, Scala, Spark, Python und R. Im wesentlichen werden in einem Notebook polyglotte Abfrageskripte ad-hoc gegen die diversen Datenquellen ausgeführt und deren Ergebnisse in einem AngularJS, konfigurierbaren Web-Dashboard visuell und interaktiv dargestellt. Dadurch eignet es sich sowohl zur explorativen Datenanalyse als auch zum veröffentlichen und teilen von Analyseergebnissen.

Jupyter ähnelt in den meisten Aspekten Apache Zeppelin, sodass sich alle oben zu Zeppelin genannten Punkte auch zu Jupyter nennen lassen. Die Unterschiede liegen eher im Detail der einzelnen Funktionen sowie der historisch und organisatorisch bedingten Nähe zu bestimmten Schlüsseltechnologien. Für das hier behandelte Forschungsprojekt spielen die individuellen Stärken und Schwächen der beiden Werkzeuge jedoch keine Rolle, weshalb beide Werkzeuge ebenbürtig eingesetzt werden.

Socrata Open Data (SODA) ist eine quelloffene Open Data Web-Programmierschnittstelle des U.S. Amerikanischen Dienstleisters Socrata. Anhand URLs werden Datasets adressiert und mittels der an SQL angelehnten *Socrata Query Language* (SoQL) per HTTP GET in unterschiedlichen Datenformaten abgefragt. Zudem stehen SDKs für diverse Programmiersprachen zur Verfügung. Im Rahmen des Projektes werden wir die Datensätze der API im JSON Dateiformat abrufen.

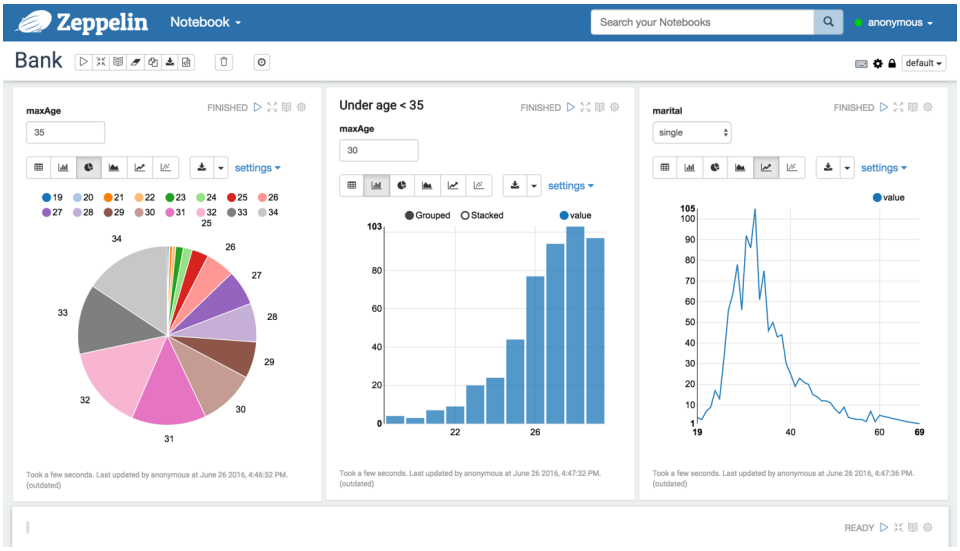

Abbildung 2.2: Beispielhaftes Notebook mit Apache Zeppelin[4]

Kapitel 3

Lösungsansatz

In diesem Kapitel wird ein Konzept und die Herangehensweise zur prototypischen Lösung der in Kapitel 1 - *Einleitung* genannten Problemstellung erörtert.

3.1 Architektur

Grundsätzlich lassen sich zwei unterschiedliche Architekturansätze im Bereich Data Streaming unterscheiden:

Lambda Architektur: Benannt nach dem griechischen λ zeichnet sich dieser Ansatz dadurch aus, dass die Streaming Daten zweigleisig verarbeitet werden. Zum einen wird der Datenstrom direkt in einen häufig auf In-Memory Technologie basierenden *Speed Layer* geleitet, der diese in Echtzeit verarbeitet und dem *Serving Layer* zur Verfügung stellt. Da Streams per Definition unendlich und der Speed Layer teuer und physikalisch endlich ist, wird der Stream von *Ingestion Layer* parallel in den *Batch Layer* geleitet. Dieser speichert zunächst die Daten persistent und startet nach einem fest definierten Intervall einen Batch-Job um die bis dahin angelaufenen Daten zu verarbeiten und zum Serving Layer zu übertragen. Es ist also die nicht zu unterschätzende Verantwortung des Serving Layers, die aggregierten Bestandsdaten aus dem Batch Layer mit den Echtzeitdaten des Speed Layers, die noch nicht vom Batch Layer verarbeitet worden sind, abzumischen.

Kappa Architektur: Dieser von Confluent Mitgründer und CEO Jay Kreps entworfene Ansatz verzichtet auf eine Batch-Verarbeitung und kommt somit mit lediglich mit *Ingestion-*, *Speed-* und *Serving-Layer* aus. Damit spart man sich Entwicklung und Betrieb von zwei separaten Schichten und das aufwändige Abmischen von Batch-Daten mit Live-Daten im *Serving-Layer*. Voraussetzung ist allerdings, dass der *Ingestion-Layer* nicht nur Daten volatil durchreicht sondern vielmehr als **Puffer** die Rohdaten persistent im *Master-Dataset* vorhält, um im Falle einer neuen, noch nicht vorberech-

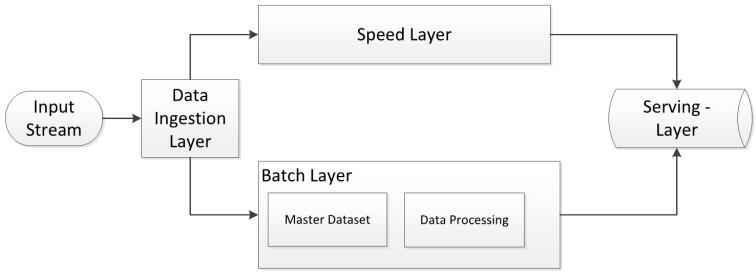

Abbildung 3.1: Schema der λ Architektur[1]

neten Anfrage oder Änderung im *Speed-Layer* die Rohdaten erneut bereit zu stellen. Um ein korrektes Replay der Nachrichten sicherzustellen beruht der Puffer auf einem kanonisches Log, in dem lediglich Nachrichten unverändert hinzugefügt, aber bereits gespeicherte Nachrichten nicht mehr verändert oder in ihrer Reihenfolge verschoben werden können. Um gleichzeitig Nähe als auch Abgrenzung zur λ Architektur zu veranschaulichen wurde dieser Ansatz nach dem griechischen κ benannt.

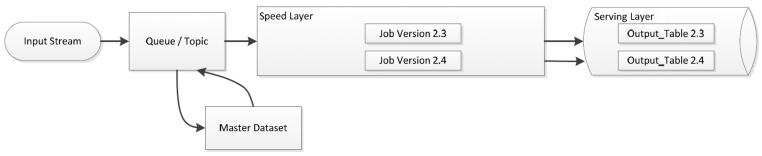

Abbildung 3.2: Schema der κ Architektur[2]

Der erste Lösungsentwurf für unser Problem orientiert sich an der Lambda Architektur. Weil die von uns gewählte Datenquelle kein Stream sondern ein statisches, online abrufbares Datenset ist, sieht unser Lösungsansatz vor, mittels eines Kafka Producers **(1)** die Online-Daten zu laden und sukzessive in einen Kafka Topic zu schreiben, um somit eine Art Pseudo-Stream zu imitieren. Da das Web-API ohnehin *Paging* über das Datenset erlaubt sollte die Implementierung nicht sonderlich kompliziert sein.

In Schritt **(2)** sammelt ein Consumer zeitgesteuert die Nachrichten des Kafka Topics ein und schreibt die Daten per SQL `INSERT` [3] in eine relationale Datenbanktabelle. Auf dieser Tabelle horcht ein `AFTER INSERT TRIGGER` **(3)**, der, sobald Daten in die Tabelle geschrieben wird, eine `FUNCTION` bzw. `STORED PROCEDURE` aufruft um Aggregate über die neuen Datensätze zu berechnen und in einer separaten Tabelle abzuspeichern bzw. mit bereits bestehenden Aggregaten zu verrechnen. Danach kann die Tabelle mit den Rohdaten theoretisch geleert werden, spätestens jedoch sobald der verfügbare Speicherplatz zu neige geht[4].

[3] Um die IO-Last der Datenbankverbindung gering zu halten sollte die Daten per BULK-INSERT erfolgen

[4] Alternativ könnte man einen `INSTEAD OF TRIGGER` benutzen und ausschließlich Aggregate permanent zu speichern

Parallel dazu konsumiert ein in Zeppelin- oder Jupyter-Notebook eingebetteter Consumer[5] die Rohdaten aus dem gleichen Topic **(4)**.

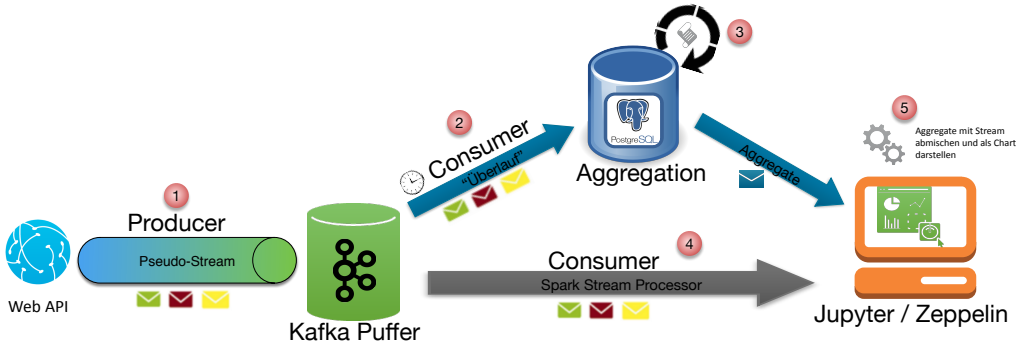

Abbildung 3.3: Lösungsentwurf nach λ

Somit wären im Auswertungs-Dashboard **(5)** sowohl Charting auf Live-Daten als auch über aggregierte Bestandsdaten aus PostgreSQL, die ggf. aus Speicherplatzgründen gar nicht mehr in Kafka vorgehalten werden können, möglich. Der Serving Layer würde in diesem Fall in die Auswertungskomponente fallen.

Da allerdings mit diesem Ansatz die bereits erörterten Nachteile der Lambda Architektur einhergehen und wir in unserem Beispiel keinen unendlichen Stream sondern eine endliche Datenmenge haben, die ganzheitlich in den Kafka-Puffer passt, haben wir den Lösungsentwurf überarbeitet und an die einfachere Kappa-Architektur angeglichen.

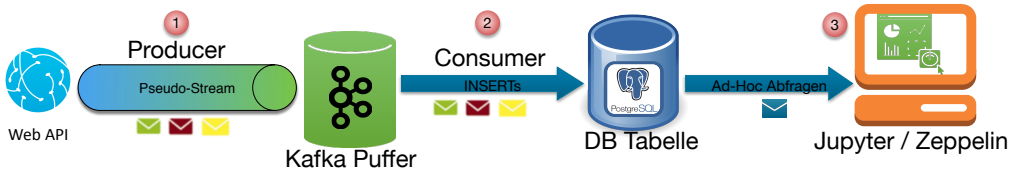

Abbildung 3.4: Finaler Lösungsentwurf nach κ

Schritt **(1)** bleibt unverändert, wohingegen in Schritt **(2)** nur noch ein einziger Consumer die Nachrichten des Kafka Topics subskribiert und direkt in eine Datenbanktabelle weiter leitet. Der zweite Consumer sowie Voraggregationen in PostgreSQL entfallen. Somit muss in den Zeppelin- bzw. Jupyter-Notebooks **(3)** auf nur eine Datenquelle zugegriffen werden, um die Daten auszuwerten und zu visualisieren.

[5] Apache Spark liefert bereits eine Kafka-Consumer Bibliothek

3.2 Data Ingestion

Wie in Kapitel 1 - *Einleitung* erläutert wird für die Erstellung des Prototyps NYC Open Data als Datenquelle genutzt. NYC Open Data bietet verschiedene Datenätze wie z. B. der Standort von Wi-Fi Hotspopts oder offene Stellenauschreibungen.[6]

Innerhalb dieses Projektes wird der Datensatz mit dem Kürzel *fhrw-4uyv* verwendet. Er beinhält alle Service Request die seit 2011 von den Einwohner von New York City abgesetzt worden sind. Beispielsweise kann mit diesem Datensatz herausgefunden werden wo welche Straßenlaternen in New York City ausgefallen sind oder in welchem Haus zu welcher Uhrzeit gefiert wurde da sich jemand über den Lärm beschwert hat.

In unserem Projekt beschaffen wir die Daten über eine CSV Datei die man sich bei NYC Open Data herunter laden kann und direkt über die SODA API die einen eigenen Endpunkt anbietet um die Service Requests abzufragen.

Die Data Ingestion per CSV Datei setzt Julian Ruppel mit der Programmiersprache Java um und den kontinuierlichen Datenstrom über die SODA API wurde von Johannes Weber mit der Programmiersprache Python ausgelesen.

Beide Ansätze werden in diesem Kapitel beschrieben und miteinander verglichen.

3.2.1 Data Ingestion via SODA Schnittstelle

Folgende Python Bibliotheken wurden genutzt um die Implementierung des Producers zu unterstützen:

- `sodapy`
- `kafka-python`

Im Kern basiert `sodapy` auf dem `Request` Paket von Python und vereinfacht das Absenden von Anfragen an die SODA API.[7]

kafka-python ist ein offiziell unterstützter Klient für Apache Kafka in der Programmiersprache Python und basiert lose auf der offiziellen Java Implementierung des Kafka Klients.[8] Gegenüber dem offiziellen Paket `confluent-kafka-python` ist `kafka-python` komplett in Python geschrieben und somit entfällt die Installation des Pakets `librdkafka` , das bei dem offiziellen Klient von Confluent noch installiert werden muss.[9]

[6] NYC Open Data. *List of most popular datasets*. 2017. URL: https://data.cityofnewyork.us/browse?provenance=official&sortBy=most_accessed&utf8=%E2%9C%93 (besucht am 06. 02. 2018).
[7] Cristina. *sodapy*. 2017. URL: https://github.com/xmunoz/sodapy (besucht am 05. 02. 2018).
[8] Dana Powers. *kafka-python*. 2017. URL: http://kafka-python.readthedocs.io/en/master/ (besucht am 05. 02. 2018).
[9] Confluent. *Python*. 2017. URL: https://cwiki.apache.org/confluence/display/KAFKA/Clients#Clients-Python (besucht am 05. 02. 2018).

Python Quellcode Programmablauf

Der Producer besteht insgesamt aus zwei Python Skripten:

- `SodaHelper.py`
- `producer.py`

Das SodaHelper Skript ist ein separater Wrapper um die `sodapy` Bibliothek um die Verbindung zu der API herzustellen und die Daten zu holen und der Producer ist für die Weiterleitung der empfangenen Daten an Apache Kafka zuständig. Somit wird eine klare Aufgabentrennung erreicht.

Da es sich hierbei nicht um ein Live Stream handelt wie z. B. bei der Twitter API, haben wir uns dazu entschieden einen „Fake Stream“ zu erstellen, indem nicht alle Daten sofort an Apache Kafka weitergeleitet werden sondern immer ein gewisser Abstand zwischen dem Senden der einzelnen Datensätze erzwungen wird.

Der ausgewählter Datensatz ist sehr groß. Wenn z. B. der Nutzer Daten von einem Monat abrufen will kann es vorkommen, dass in diesem Zeitraum mehr als 100.000 Datensätze geladen werden. Da der Abruf von einer solch großen Menge an Daten über die API eine sehr hohe Rechenleistung erfordert und diese im Rahmen des Projektes nicht zur Verfügung steht wurde ein Paging entwickelt.

Durch die feste Angabe eines Limits von 10.000 wird sichergestellt, dass ein Request an die SODA API nur maximal 10.000 Datensätze liefern kann und der Algorithmus führt so lange Requests an die API aus bis der gewünschte Zeitraum des Users komplett empfangen und die Request an Apache Kafka gesendet wurden.

Vorgegeben durch das Paket *kafka-python* können die Einträge eines Topics nur als byte String abgelegt werden. Aus diesem Grund wird der empfangende Datensatz zuerst in ein byte string umgewandelt und dann an Apache Kafka gesendet.

Folgendes Code Snippet zeigt den Algorithmus um das Paging zu realisieren. Mit Hilfe des SodaHelpers werden zunächst die Datensätze von der API - unter Berücksichtigung des Limits, des Anfangs- und Enddatums geholt. Die Variablen *from_date* und *to_date* werden beim Starten des Programms von dem User gesetzt.

Der Algorithmus beruht auf der Annahme, dass wenn der empfangene Datensatz genau die Länge des Limits hat es immer noch weitere Datensätze gibt die von der API abgerufen werden müssen. Wenn also das Limit erreicht wurde wird das Datum des letzten Datensatzes als neues Enddatum festgelegt und der Prozess beginnt von vorne. Dies geschieht solange die Anzahl der empfanenen Datensätze nicht dem Limit entsprechen oder die verwendeten Anfangs- und Enddatumswerte identisch sind.

```
1     def fetch_data(self, from_date, to_date, limit):
```

```
        requests = self.soda.get_data(dataset_identifier=config.
            SOCRATA_DATASET,
                                      from_date=from_date,
                                      to_date=to_date,
                                      limit=limit)
        length = len(requests)
        if length > 0:
            date = requests[-1]["created_date"]
        else:
            date = self.to_date
        for request in requests:
            json_string = json.dumps(request)
            json_byte = b"" + json_string
            print("Sending to Kafka...")
            self.producer.send(self.topic, json_byte)
            # time.sleep(random.randint(0, 50) * 0.1)

        return len(requests), date

    def run(self):
        number_of_entries, to_date = self.fetch_data(from_date=self.
            from_date,
                                                     to_date=self.to_date,
                                                     limit=self.limit)
        while number_of_entries == self.limit and to_date != self.
            from_date:
            print("Next Request...")
            number_of_entries, to_date = self.fetch_data(from_date=
                self.from_date,
                                                         to_date=to_date,
                                                         limit=self.limit)
        print("End.")
```

3.2.2 Data Ingestion via CSV Datei

Als offline-fähige Alternative zur SODA API gibt es einen weiteren Kafka Producer in Java. Dieser liest die Nachrichten zeilenweise aus einer lokalen CSV Datei ein und publiziert die Datensätze als JSON einzeln an einen Kafka Topic. Die CSV Datei wurde vorher aus dem NYC OpenData Portal runtergeladen.

Der Programmablauf lässt sich in wenigen Stichpunkten beschreiben:

1. Verbindung des `KafkaProducer<Long, String>` zum Kafka Cluster konfigurieren.Dazu zählen hauptsächlich Host und Port sowie die Datentypen, um Schlüssel und Wert der Nachrichten zu serialisieren.
2. Zeilenweises einlesen der CSV Datei und dabei jeweils den Datensatz in einen JSON-String transformieren, diesen `String` als Wert in einen `ProducerRecord<Long, String>` setzen und an den bestimmten Kafka Topic senden. Um einen realen Stream zu simulieren wartet der Producer-Thread pro verarbeiteten Datensatz eine zufällige Wartezeit zwischen 0 bis 2 Sekunden.

Listing 3.1 zeigt einen Überblick über die relevanten Methoden.

```
1
2   public static void main(String[] args) throws Exception {
3
4     Reader in = new FileReader("/data/311
          _Service_Requests_from_2010_to_Present.csv");
5     Iterable<CSVRecord> records = CSVFormat.RFC4180.
          withFirstRecordAsHeader().parse(in);
6
7     new MyProducer("").runProducer(records);
8
9   }
10
11  public void runProducer(final Iterable<CSVRecord> records) throws
        Exception {
12
13    try (final Producer<Long, String> producer = createProducer()) {
14
15      for (final CSVRecord csvRecord : records) {
16        String json = objectMapper.writeValueAsString(csvRecord.toMap
              ());
17        final ProducerRecord<Long, String> record = new ProducerRecord
              <Long, String>(KafkaCommons.loadProperties().getProperty("
              TOPIC", KafkaCommons.TOPIC), Long.parseLong(csvRecord.get(
              "Unique Key")), json);
18        Thread.sleep(new Random().nextInt(2000));
19        RecordMetadata metadata = producer.send(record).get();
20        System.out.printf("sent record(key=%s value=%s) " + "meta(
              partition=%d, offset=%d)\n", record.key(), record.value(),
               metadata.partition(), metadata.offset());
21
22      }
23
```

```
24      }
25    }
26
27    private Producer<Long, String> createProducer() throws IOException{
28      Properties props = new Properties();
29      props.put(ProducerConfig.BOOTSTRAP_SERVERS_CONFIG, KafkaCommons.
            loadProperties().getProperty("BOOTSTRAP_SERVERS",
            KafkaCommons.BOOTSTRAP_SERVERS));
30      props.put(ProducerConfig.CLIENT_ID_CONFIG, "KafkaCSVProducer");
31      props.put(ProducerConfig.KEY_SERIALIZER_CLASS_CONFIG,
            LongSerializer.class.getName());
32      props.put(ProducerConfig.VALUE_SERIALIZER_CLASS_CONFIG,
            StringSerializer.class.getName());
33      return new KafkaProducer<Long, String>(props);
```

Listing 3.1: Auszug aus `com.srh.bdba.dataengineering.MyProducer`

Zur Implementierung wurden folgende Bibliotheken über Maven eingebunden:

- `org.apache.kafka:kafka-clients`
- `org.apache.commons:commons-csv`
- `com.fasterxml.jackson.core:jackson-core`
- `com.fasterxml.jackson.core:jackson-databind`

3.3 Data Storage

Apache Kafka wird dazu verwendet um den kompletten Datensatz eines Service Requests abzuspeichern. Ein Consumer liest die Daten von Apache Kafka aus und speichert die relevanten Attribute eines Datensatzes in der Datenbank. Als Datenksystem haben wir in unserem Projekt für PostgreSQL entschieden.

Zunächst erstellten wir eine Tabelle mit dem Namen ***service_request*** um die Service Requests von New York zu speichern. Eine stichprobenartige Analyse des Datensatzes hat ergeben, dass einzelne Felder des original Datensatzes sehr häufig mit `NULL` Werten belegt sind. In der Datenbanktabelle ***service_request*** wurden diese Felder außer Acht gelassen.

Die Datenbanktabelle besitzt folgende Attribute

1. unique_key
2. created_date
3. agency_name

4. complaint_type
5. descriptor
6. longitude
7. latitude
8. agency
9. location_type
10. incident_zip
11. incident_address
12. street_name
13. cross_street_1
14. cross_street_2
15. address_type
16. city
17. status
18. due_date
19. borough
20. resolution_description

In der offiziellen Dokumentation des Datensatzes werden die einzelnen Attribute eines Service Requests genau beschrieben und deren technische Bezeichner aufgelistet. [10]

Die Attribute der Tabelle ***service_request*** sind identisch zu den Bezeichnern des original SODA Datensatzes.

Um das Handling mit Aoache Kafka zu vereinfachen wurden zwei weitere Python Bibliotheken installiert. Diese sind:

- kafka-python
- sqlalchemy
- psycopg2 (wird von sqlalchemy benötigt)

Genau wie bei dem Producer abstrahiert das Paket `kafka-python` die Verbindung zu unserem Apache Kafka Server und erleichtert den Zugriff auf das Topic. Durch den Einsatz von `sqlalchemy` ist es möglich auf die Datenbank und deren Tabelle(n) in einer objektorien-

[10] https://dev.socrata.com/foundry/data.cityofnewyork.us/fhrw-4uyv

tierten Weise zuzugreifen und die Ausführung von SQL Statements wird vereinfacht bzw. durch Klassenmethoden abstrahiert.

Python Quellcode Programmablauf

Neben den verwendeten Bibliotheken besteht derConsumer aus zwei Python Skripten.

- DBHelper.py
- consumer.py

Während sich das `consumer` Skript um das Auslesen eines Kafka Topics kümmert ist die `DBHelper` Klasse für die Verbindung zu der Datenbank zuständig, liest Tabellenspalten aus und speichert einen Datensatz in der Tabelle.

Nachfolgend die Code Snippet der Consumer Klasse.

```
1         self.db_helper = DBHelper(config.DATABASE_NAME)
2         self.db_columns = self.db_helper.get_table_column_names(self.
              db_table)
3
4     def run(self):
5         requests = KafkaConsumer(config.KAFKA_TOPIC,
6                                  bootstrap_servers=config.KAFKA_SERVER)
                                      # auto_offset_reset='earliest'
7
8         for message in requests:
9             db_entry = {}
10            message_json = json.loads(message.value)
11            for column in self.db_columns:
12                if column in message_json: # check if key is available
                       in request object
13                    db_entry[column] = message_json[column]
14            print("Saving in DB...")
15            self.db_helper.insert(db_entry, self.db_table)
```

Sobald das Consumer Skript aufgerufen wird, wird eine Verbindung mit der Datenbank aufgebaut und der `KafkaConsumer` stellt eine Verbindung mit dem Apache Kafka Server bzw. dem Topic SServiceRequest"her. (Zeile 1 - 9)

Sobald von seitens des Producers ein neuer Datensatz in das Topic SServiceRequests"geschrieben wird, wird der Datensatz ausgelesen in ein JSON umgewandelt, die benötigten Attribute aus dem JSON gelesen und in ein temporäres Dictionary geschrieben. Dieses Dictionary wird abschließend mit Hilfe des `DBHelper` Skripts in die Datenbanktabelle geladen. (Zeile 11 - 18)

Wie eingangs erwähnt bezeichneten wir die Attribute der Datenbanktabelle und des SODA Datensatzes identisch.

Mit diesem kleinen „Kniff“ kann mit Hilfe der Namen der Tabellenspalten auf die Keys des JSON zugegriffen, den zugehörigen Wert ausgelesen und als neuen Wert für das temporäre Dictionary genutzt werden. (Zeile 14 - 16)

Java Quellcode Programmablauf

Auch für den Consumer gibt es eine alternative Implementierung in Java. Diese findet sich in `com.srh.bdba.dataengineering.MyConsumer`. Der Programmablauf ist ähnlich zu der Python Implementierung, weshalb an dieser Stelle auf Codelistings im Anhang TODO verwiesen wird. Kurz zusammengefasst wird ein `KafkaConsumer<Long, String>` konfiguriert und pro 100 Millisekunden am Kafka-Cluster nachgefragt, ob es neue `ConsumerRecord <Long, String>` für das entsprechende Topic gibt. Falls dem so ist wird die JSON Nachricht aus dem `ConsumerRecord<Long, String>` ausgepackt und per `PreparedStatement` über `JDBC` in die PostgreSQL Tabelle eingefügt.

3.4 Data Retrieval

Die Aufgabe im Bereich Data Retrieval war es die zum einen die Daten aus der Datenbank auszulesen und diese mit einem virtuellen Notebook zu visualisieren.

Die Beschaffung und Auswertung der Daten mit Apache Zeppelin übernahm Julian Ruppel. Johannes Weber bereitete die Daten mit Python auf und visualisierte sie in einem Jupyter Notebook.

3.4.1 Data Retrieval mit Apache Zeppelin

URL öffnen

Interpreters konfigurieren

Notebook laden

SQL Queries ausführen und Diagramme auswerten

3.4.2 Data Retrieval mit Jupyter

Das Jupyter Notebook bietet die Möglichkeit mit *Magic-Commands* direkt aus dem Notebook heraus eine Verbindung mit der Datenbank aufzubauen und SQL Statements abzusetzen. Einfache Tabellen werden direkt in Jupyter visualisiert wohingegen komplexere

Visualisierungen wie z. B. Balkendiagramme oder Geo Plots mit Python Bibliotheken dargestellt werden müssen. Für die Darstellung in Jupyter werden sog. *Widgets* installiert und aktiviert.

Folgende Python Bibliotheken wurden installiert um SQL Statements in Jupyter ausführen und visualisieren zu können.

- ipython-sql
- sqlalchemy (wird von ipython-sql benötigt)
- bokeh
- gmaps

Mit `ipython-sql` werden SQL *Magic-Command* in Jupyter aktiviert. `ipython-sql` nutzt `sqlalchemy` um sich mit der Datenbank zu verbinden. Ein abgesetztes SQL Statement lässt sich entweder direkt in Jupyter ausgeben oder einer beliebigen Variable zuordnen die dann weiterverarbeitet werden kann.

`bokeh` ist eine mächtige Python Bibliothek um viele Arten der Visualisierung umzusetzen wie z. B. Balkendiagramme, Scatter Plots, Geo Maps oder Zeitreihen.

Die Visualisierung von Geo Daten erfolgt mit der Bibliothek `gmaps`. Diese greift auf die Karten von Google Maps zu erlaubt es die Geo Daten in einem Layer über einen beliebigen Kartenausschnitt zu legen. Der Vorteil von `gmaps` gegenüber `bokeh` ist zum einen die Nutzung des Kartenmaterials von Google aber auch die interaktive Nutzung des Kartenauschnitts mit z. B. StreetView.

Folgende Fragestellungen wurden im Rahmen von Data Retrieval beantwortet.

1. Zeige alle Beschwerdetypen die häufiger als 400 aber seltener als 8000 Mal gemeldet wurden?
2. Wie lautet die Beschreibung der häufig vorkommenden Service Requests?
3. An welchen Orten von New York City wurden Service Request vom Typ 'Noise - Residential' abgesetzt?
4. Wieviele Service Requests sind im Jahr 2017 eingegangen? Gruppiert nach Tag und Sortiert nach dem Erstellungsdatum.

Nachfolgender Abschnitt listet die SQL Statements zu jeder Fragestellung sowie ein dazugehöriges Beispiel.

zu 1.

```
SELECT complaint_type, COUNT(complaint_type)FROM service_request GROUP
BY complaint_type HAVING COUNT(complaint_type)> 400 AND COUNT(complaint_type
)< 8000
```

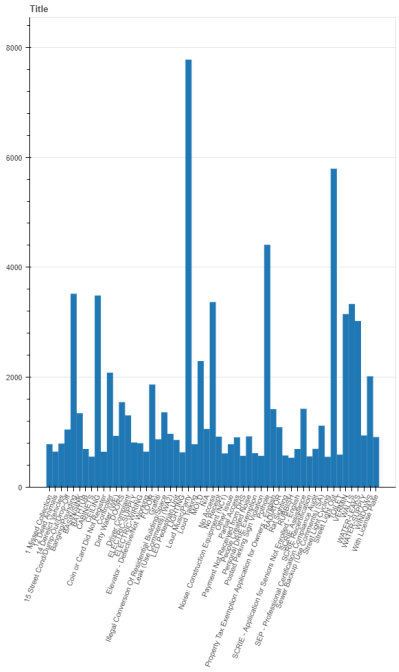

Abbildung 3.5: Tabelle mit allen Beschwerdetypen[11]

zu 2.

```
SELECT descriptor, COUNT(descriptor)FROM service_request WHERE descriptor
 IS NOT NULL GROUP BY descriptor ORDER BY count DESC
```

descriptor	count
HEAT	33615
Loud Music/Party	7771
Street Light Out	5788
Pothole	4403
Banging/Pounding	3512
CEILING	3481
No Access	3361
WALLS	3326
VERMIN	3144
WATER-LEAKS	3018

Abbildung 3.6: Auszug aus den meisten Service Requests[12]

zu 3.

```
SELECT longitude, latitude FROM service_request WHERE complaint_type =
```

[11] eigene Darstellung
[12] eigene Darstellung

```
'Noise - Residential'and latitude IS NOT NULL and longitude IS NOT NULL
```

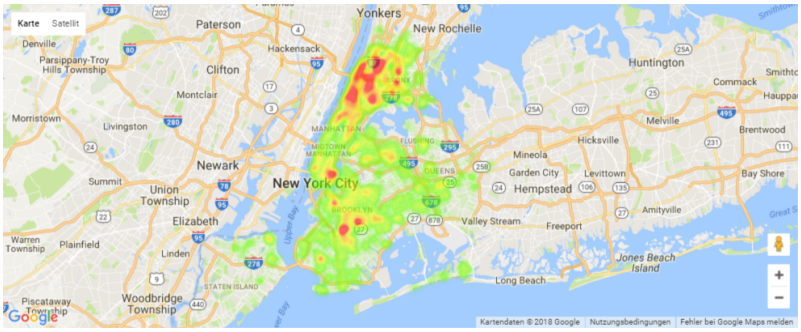

Abbildung 3.7: Heatmap der Lärmquellen in New York[13]

zu 4.

```
SELECT date_trunc('day', created_date)AS dd, COUNT(created_date)as daily_sum
 FROM service_request where EXTRACT(year from created_date)= '2017'GROUP
 BY dd ORDER BY date_trunc('day', created_date)
```

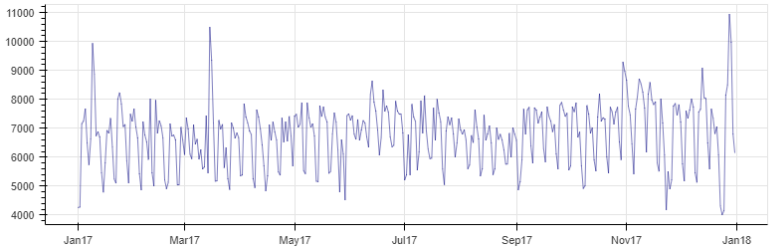

Abbildung 3.8: Timeline aller Service Requests in 2017[14]

[13] eigene Darstellung
[14] eigene Darstellung

Kapitel 4

Inbetriebnahme

In diesem Kapitel werden die einzelnen Schritte beschrieben, welche durchgeführt werden müssen um den Prototyp ausführen zu können.

Hinweis:
Hierbei handelt es sich lediglich um eine grob-granulare Beschreibung der Inbetriebnahme des Prototypen. Weitere Details zur Installation und Nutzung können den Quellen entnommen werden.

4.1 Voraussetzung & Infrastruktur

4.1.1 Installation

Folgende Komponente müssen installiert und lauffähig sein:

- Apache Kafka und Apache Zookeeper
- PostgreSQL

Je nachdem welche Technologie man für die Consumer, Producer und Auswertung verwenden möchte ergeben sich folgende zwei alternative Technologiestacks:

- Python 2
- pip
- Jupyter
- JDK 1.8
- Maven
- Apache Zeppelin

4.2 Infrastruktur Setup & Konfiguration

1. Projekt aus Git Repository[1] auschecken
2. Apache Zookeeper starten und initialisieren
3. Apache Kafka starten und initialisieren
4. Kafka Topic anlegen
5. PostgreSQL Datenbankschema initialisieren[2]

4.3 Setup und Starten des Python Prototypen

1. Python Bibliotheken installieren
   - `$pip install sodapy`
   - `$pip install kafka-python`
   - `$pip install psycopg2`
   - `$pip install sqlalchemy`
   - `$pip install jupyter`
   - `$pip install bokeh`
   - `$pip install ipython-sql`
   - `$pip install gmaps`
2. `gmaps` Widget für Jupyter aktivieren
   - `$jupyter nbextension enable --py --sys-prefix widgetsnbextension`
   - `$jupyter nbextension enable --py --sys-prefix gmaps`
3. Google Maps API Key erstellen um die Google Maps Visualisierungen nutzen zu können[3]
4. Applikation Token für die SODA API beziehen[4]
5. Python Konfigurationsdatei `/python/config.py` pflegen

[1] https://github.com/johannesweber/BDBA_DataEngineering.git
[2] DDL Skripte um das Datenbankschema in PostgreSQL aufzubauen befinden sich im Ordner /db
[3] Eine Anleitung zum Erzeugen eines API Keys gibt es http://jupyter-gmaps.readthedocs.io/en/latest/authentication.html
[4] Socrata. *Obtaining an Application Token*. 2017. URL: https://dev.socrata.com/docs/app-tokens.html (besucht am 05. 02. 2018).

6. Python Consumer und Producer starten via `/python/main.py`
7. Jupyter starten und vorkonfiguriertes Notebook BDBA/DataEngineering.ipynb laden

4.4 Setup und Starten des Java Prototypen

1. CSV Datei von https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9 runterladen und lokal speichern
2. Java Projekt mit Maven bauen
3. Konfiguration `/java/src/main/resources/kafka_config.properties` pflegen
4. Python Consumer und Producer starten via `com.srh.bdba.dataengineering.Main`. Beim Aufruf der `Main()` muss man den Pfad zur CSV Datei angeben.
5. Apache Zeppelin starten und Datenbankverbindung sowie JDBC Treiber im PSQL Interpreter pflegen
6. Vorkonfiguriertes Notebook BDBA/TODO laden

Kapitel 5

Fazit

fazit in die loesung packen? Fazit = Fazit zum Vergleich der Tools und der eingesetzetn Architektur

Abkürzungsverzeichnis

JSON JavaScript Object Notation

CSV Comma-separated values

API Application Programming Interface

SODA Socrata Open Data

SQL Structured Query Language

Abbildungsverzeichnis

Literatur

Confluent. *Python*. 2017. URL: https://cwiki.apache.org/confluence/display/KAFKA/Clients#Clients-Python (besucht am 05. 02. 2018).

Cristina. *sodapy*. 2017. URL: https://github.com/xmunoz/sodapy (besucht am 05. 02. 2018).

Data, NYC Open. *List of most popular datasets*. 2017. URL: https://data.cityofnewyork.us/browse?provenance=official&sortBy=most_accessed&utf8=%E2%9C%93 (besucht am 06. 02. 2018).

– *Our mission: open data for all*. 2017. URL: https://opendata.cityofnewyork.us/overview/ (besucht am 03. 02. 2018).

Powers, Dana. *kafka-python*. 2017. URL: http://kafka-python.readthedocs.io/en/master/ (besucht am 05. 02. 2018).

Socrata. *Obtaining an Application Token*. 2017. URL: https://dev.socrata.com/docs/app-tokens.html (besucht am 05. 02. 2018).