

Project Presentation

Course: Event Processing

Summer term 2022

Tung Nguyen

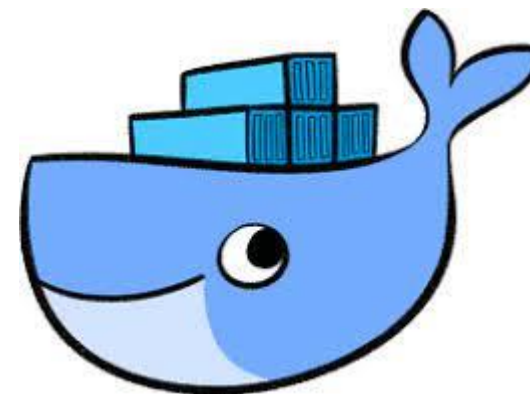
Tools



Kafka + KSQL



Java + Maven



Docker

Setup – Kafka + KSQL

2 brokers

Deployment using Docker

```
broker1:
  image: confluentinc/cp-server:7.1.2
  hostname: broker1
  container_name: broker1
  depends_on:
    - zookeeper
  ports:
    - "29092:29092"
    - "9092:9092"
    - 9101:9101
  environment:
    KAFKA_BROKER_ID: 1
    KAFKA_ZOOKEEPER_CONNECT: 'zookeeper:2181'
    KAFKA_LISTENER_SECURITY_PROTOCOL_MAP: INTERNAL:PLAINTEXT,EXTERNAL:PLAINTEXT
    KAFKA_ADVERTISED_LISTENERS: INTERNAL://broker1:29092,EXTERNAL://localhost:9092
    KAFKA_INTER_BROKER_LISTENER_NAME: INTERNAL
    KAFKA_METRIC_REPORTERS: io.confluent.metrics.reporter.ConfluentMetricsReporter
    KAFKA_OFFSETS_TOPIC_REPLICATION_FACTOR: 2
    KAFKA_GROUP_INITIAL_REBALANCE_DELAY_MS: 0
    KAFKA_CONFLUENT_LICENSE_TOPIC_REPLICATION_FACTOR: 2
    KAFKA_CONFLUENT_BALANCER_TOPIC_REPLICATION_FACTOR: 2
    KAFKA_TRANSACTION_STATE_LOG_MIN_ISR: 2
    KAFKA_TRANSACTION_STATE_LOG_REPLICATION_FACTOR: 2
    KAFKA_JMX_PORT: 9101
    KAFKA_JMX_HOSTNAME: localhost
    KAFKA_CONFLUENT_SCHEMA_REGISTRY_URL: http://schema-registry:8081
    CONFLUENT_METRICS_REPORTER_BOOTSTRAP_SERVERS: broker1:29092
    CONFLUENT_METRICS_REPORTER_TOPIC_REPLICAS: 1
    CONFLUENT_METRICS_ENABLE: 'true'
    CONFLUENT_SUPPORT_CUSTOMER_ID: 'anonymous'
```

connection
ports

```
broker2:
  image: confluentinc/cp-server:7.1.2
  hostname: broker2
  container_name: broker2
  depends_on:
    - zookeeper
  ports:
    - "29093:29093"
    - "9093:9093"
    - "9102:9102"
  environment:
    KAFKA_BROKER_ID: 2
    KAFKA_ZOOKEEPER_CONNECT: 'zookeeper:2181'
    KAFKA_LISTENER_SECURITY_PROTOCOL_MAP: INTERNAL:PLAINTEXT,EXTERNAL:PLAINTEXT
    KAFKA_ADVERTISED_LISTENERS: INTERNAL://broker2:29093,EXTERNAL://localhost:9093
```

Setup – Kafka + KSQL

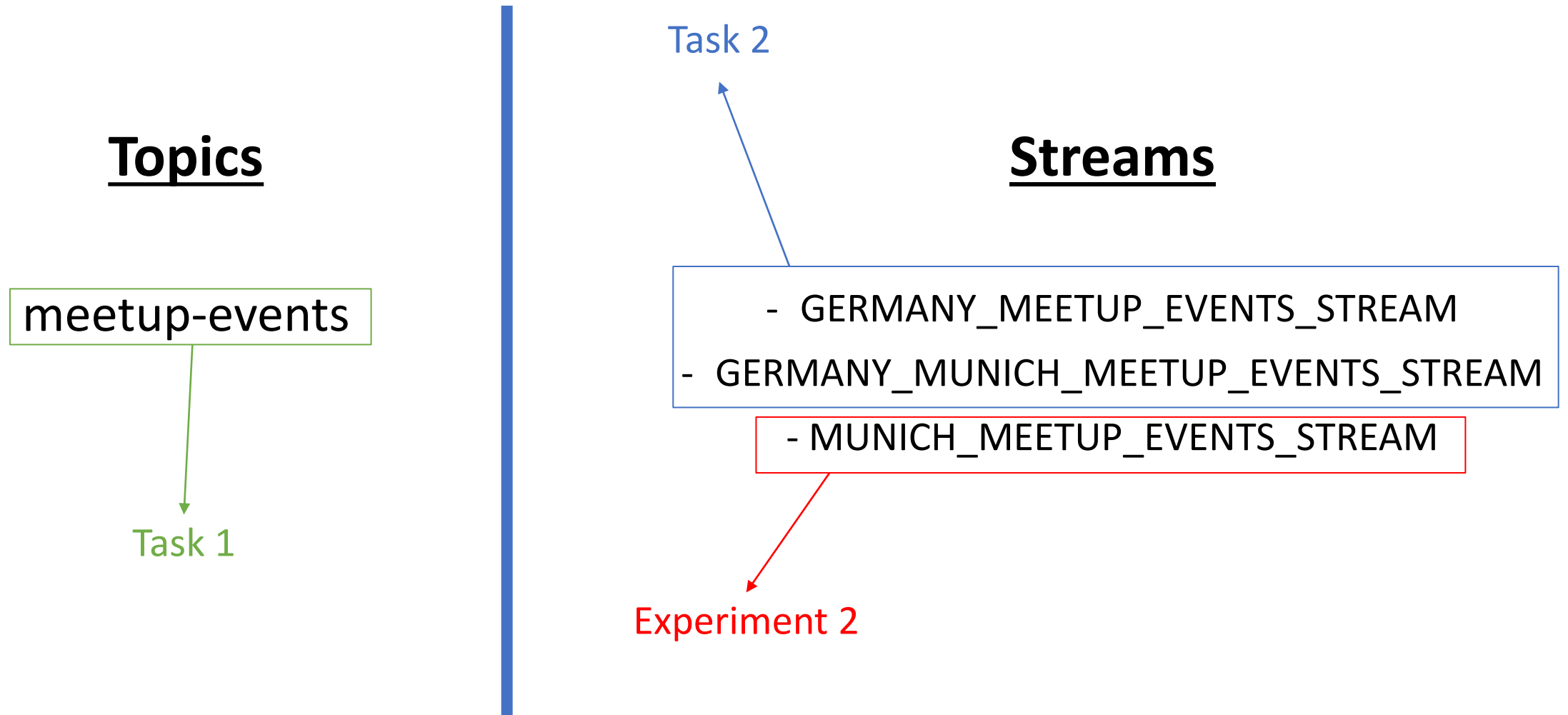
Topics

meetup-events

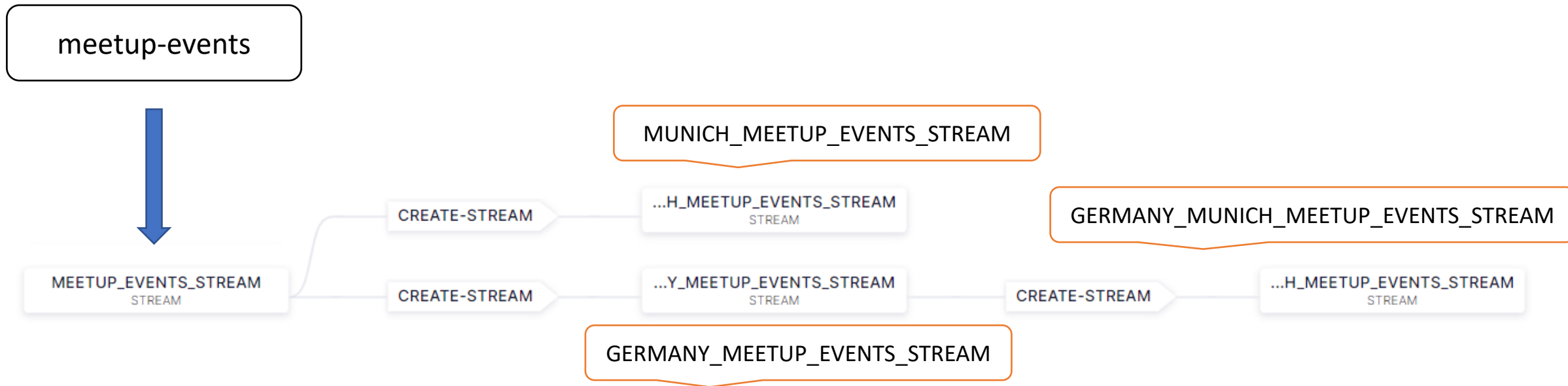
Streams

- GERMANY_MEETUP_EVENTS_STREAM
- GERMANY_MUNICH_MEETUP_EVENTS_STREAM
- MUNICH_MEETUP_EVENTS_STREAM

Setup – Kafka + KSQL

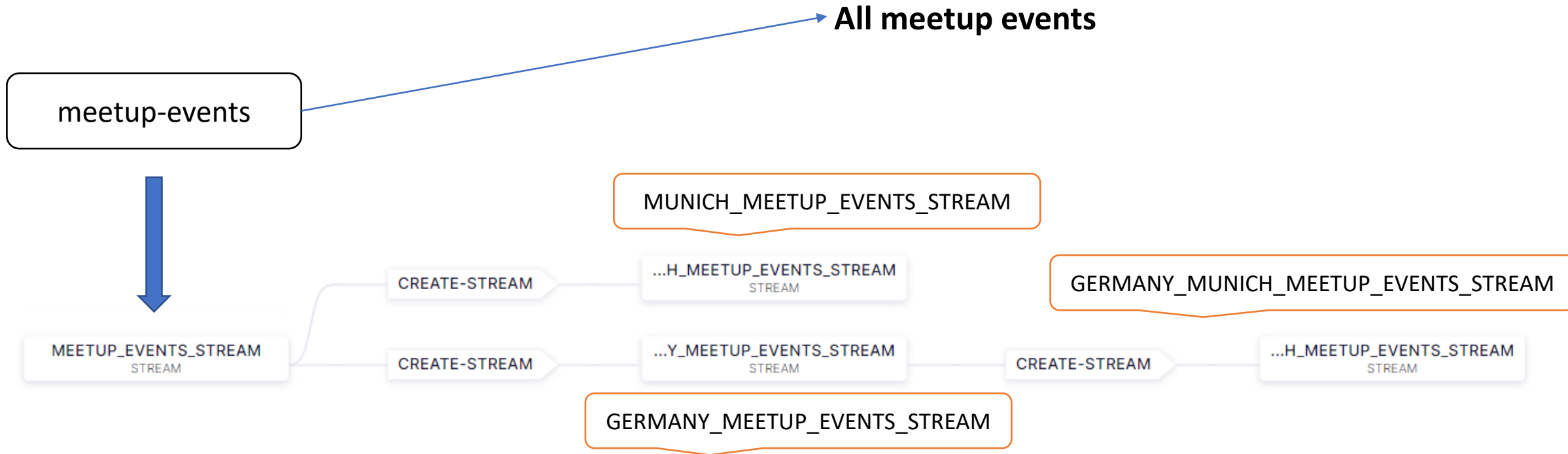


Setup – Kafka + KSQL



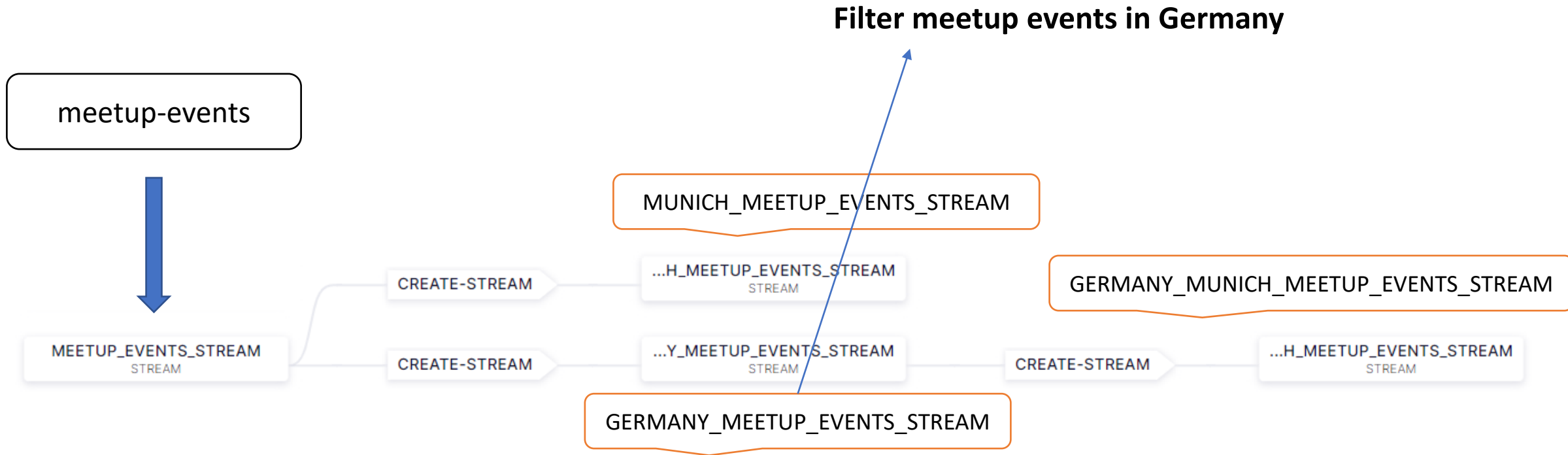
Overview of Stream Flow Topology

Setup – Kafka + KSQL



Overview of Stream Flow Topology

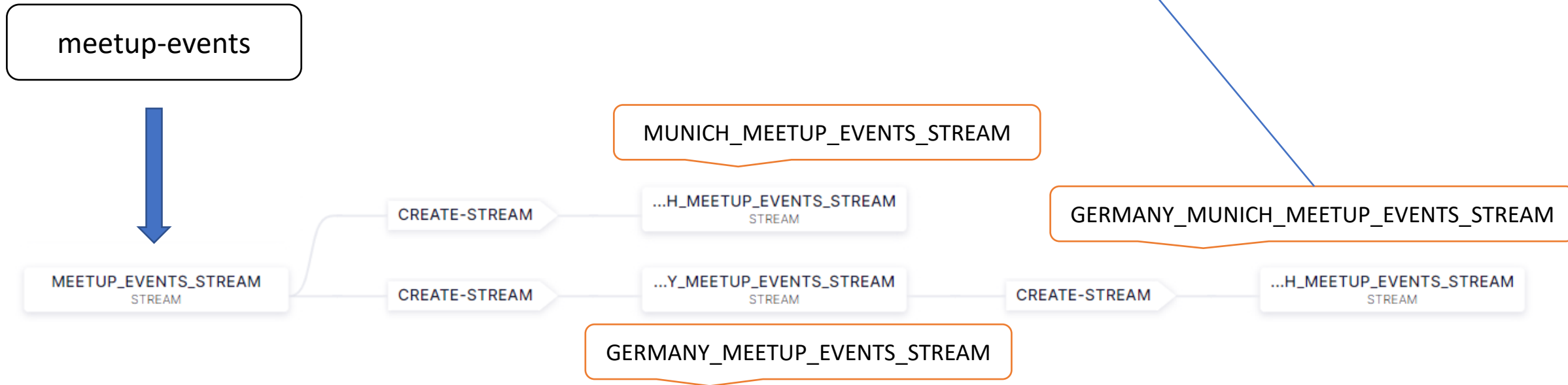
Setup – Kafka + KSQL



Overview of Stream Flow Topology

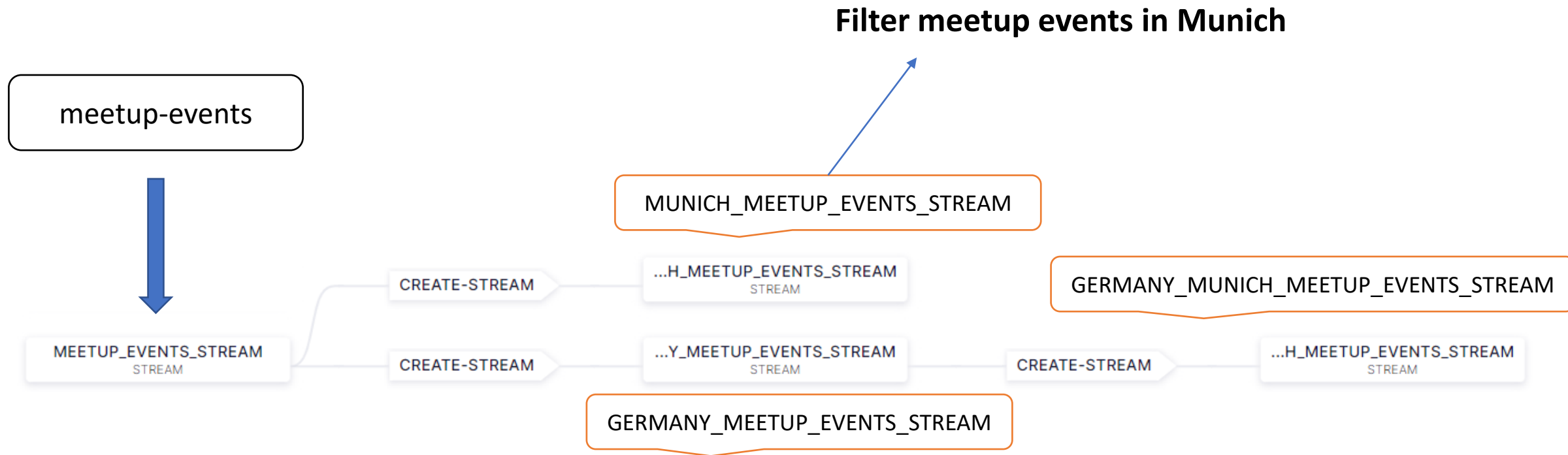
Setup – Kafka + KSQL

Filter meetup events in Munich - Germany



Overview of Stream Flow Topology

Setup – Kafka + KSQL



Overview of Stream Flow Topology

Setup – Producer + Consumer

Producer and Consumer connect to brokers on port **9092, 9093**

```
public Producer() {
    // create Producer properties
    Properties properties = new Properties();
    properties.setProperty(ProducerConfig.BOOTSTRAP_SERVERS_CONFIG, "127.0.0.1:9092,127.0.0.1:9093");
    properties.setProperty(ProducerConfig.KEY_SERIALIZER_CLASS_CONFIG, StringSerializer.class.getName());
    properties.setProperty(ProducerConfig.VALUE_SERIALIZER_CLASS_CONFIG, StringSerializer.class.getName());
    // create the producer
    producer = new KafkaProducer<>(properties);

    final Thread mainThread = Thread.currentThread();
    // adding the shutdown hook
    Runtime.getRuntime().addShutdownHook(new Thread(() -> {
        logger.info("Close the producer by calling producer.close()...");
        producer.close();
        // join the main thread to allow the execution of the code in the main thread
        try {
            mainThread.join();
        } catch (InterruptedException e) {
            logger.error("Error occurs while waiting for all threads to be shutdown");
        }
    }));
}
```

```
public Consumer(String groupId) {
    // create consumer configs
    Properties properties = new Properties();
    properties.setProperty(ConsumerConfig.BOOTSTRAP_SERVERS_CONFIG, "127.0.0.1:9092,127.0.0.1:9093");
    properties.setProperty(ConsumerConfig.KEY_DESERIALIZER_CLASS_CONFIG, StringDeserializer.class.getName());
    properties.setProperty(ConsumerConfig.VALUE_DESERIALIZER_CLASS_CONFIG, StringDeserializer.class.getName());
    properties.setProperty(ConsumerConfig.GROUP_ID_CONFIG, groupId);
    properties.setProperty(ConsumerConfig.AUTO_OFFSET_RESET_CONFIG, "earliest");
    properties.put("schema.registry.url", "127.0.0.1:8081");
    // create consumer
    consumer = new KafkaConsumer<>(properties);
    // get a reference to the current thread
    final Thread mainThread = Thread.currentThread();
    // adding the shutdown hook
    Runtime.getRuntime().addShutdownHook(new Thread(() -> {
        logger.info("Close the consumer safely by calling consumer.wakeup()...");
        consumer.wakeup();
        // join the main thread to allow the execution of the code in the main thread
        try {
            mainThread.join();
        } catch (InterruptedException e) {
            logger.error("Error occurs while waiting for all threads to be shutdown");
        }
    }));
}
```

Setup – Producer + Consumer

Producer

- Read events from source **events.json.gz**
- Publish events to topic **meetup-events** every **10ms**

```
private static Callable<Void> startProducer() {  
    return () -> {  
        try(InputStream fileStream = new FileInputStream( name: "events.json.gz");  
            InputStream gzipStream = new GZIPInputStream(fileStream);  
            LineIterator iter = IOUtils.lineIterator(gzipStream, charsetName: "UTF-8");  
        ) {  
            Producer producer = new Producer();  
            while (iter.hasNext()) {  
                String line = iter.nextLine();  
                // publish the message every 10 ms  
                producer.publish( topic: "meetup-events", line);  
                Thread.sleep( millis: 10);  
            }  
        } catch (InterruptedException e) {  
            logger.info("Thread is interrupted");  
        }  
        return null;  
    };  
}
```

Setup – Producer + Consumer

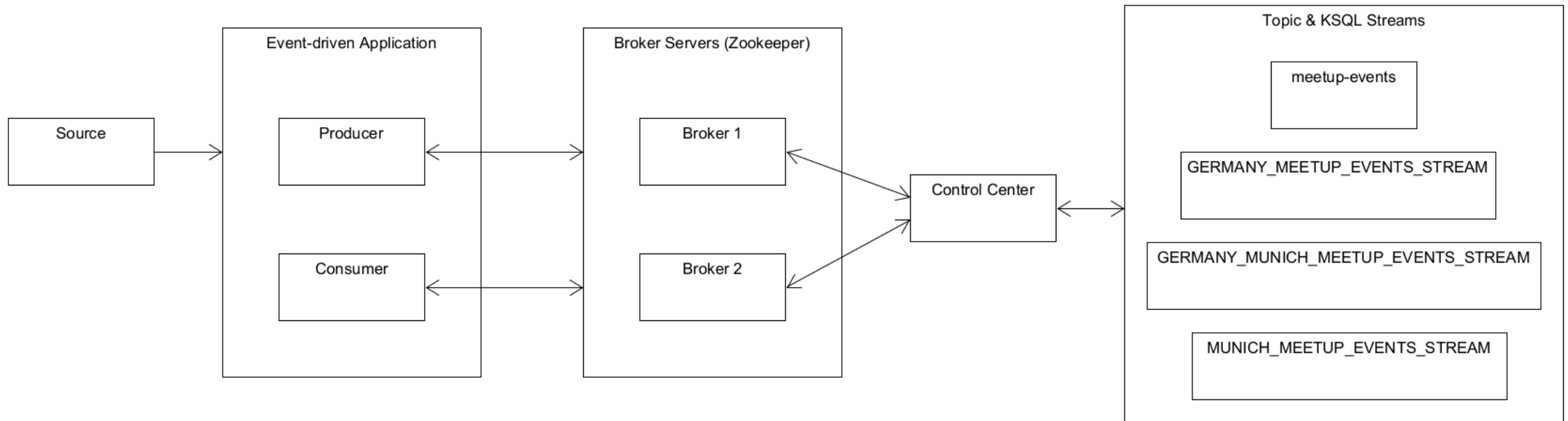
Consumer

- Subscribe the listed topic and streams
- Poll events every 100ms

```
private static Callable<Void> startConsumer() {  
    return () -> {  
        Consumer consumer = new Consumer( groupId: "meetup-events-consumer");  
        consumer.subscribe(Arrays.asList(  
            "meetup-events",  
            "GERMANY_MEETUP_EVENTS_STREAM",  
            "GERMANY_MUNICH_MEETUP_EVENTS_STREAM",  
            "MUNICH_MEETUP_EVENTS_STREAM"  
        ));  
        return null;  
    };  
}
```

```
public void subscribe(List<String> topics) {  
    try {  
        // subscribe consumer to our topic(s)  
        consumer.subscribe(topics);  
        // poll for new data  
        while (true) {  
            ConsumerRecords<String, String> records = consumer.poll(Duration.ofMillis(100));  
            for (ConsumerRecord<String, String> record: records) {  
                long currentTimestamp = System.currentTimeMillis();  
            }  
        }  
    }  
}
```

Setup – Overview



Kafka Topic – meetup-events

▼ {"utc_offset":0,"venue":{"country":"ie","city":"Dublin 6","address_1":"183 Rathmines Road Lower","name":"Holistic...
Partition: 1 Offset: 7335 Timestamp: 1658648673743

▼ {"utc_offset":0,"venue":{"country":"no","city":"Tromso","address_1":"Skattørvegen 27, 9018","name":"Skattøra Mari...
Partition: 1 Offset: 7334 Timestamp: 1658648673712

▼ {"utc_offset":0,"venue":{"country":"ie","city":"Dublin","address_1":"Unit A Gandon Mews Lucan Village","name":"Se...
Partition: 0 Offset: 7678 Timestamp: 1658648683293

▼ {"utc_offset":-28800000,"venue":{"zip":"92131","country":"us","city":"San Diego","phone":"8585300990","address_1"...
Partition: 0 Offset: 7677 Timestamp: 1658648683262

```
{
  "utc_offset": 0,
  "venue": {
    "country": "ie",
    "city": "Dublin 6",
    "address_1": "183 Rathmines Road Lower",
    "name": "Holistic Centre of Ireland",
    "lon": -6.265949,
    "lat": 53.32288
  },
  "rsvp_limit": 0,
  "venue_visibility": "public",
  "visibility": "public",
  "maybe_rsvp_count": 0,
  "description": "<p>Join me for a day of informative and fun DIY beauty workshop Would you like to make
beauty products that are non toxic, organic and paraben free?<br><br></p>\n<p>These products are all
natural and very nourishing for your skin and body<br>You will learn how to make many products and can
give them as Christmas gifts or can use them to nurture yourself.</p>\n<p>All oils are do Terra- CPTG
essential oils, they are high vibrational and the only oils I use. We will also infuse some of the
products with crystals.<br>Space for this event is limited</p>\n<p>Cost: €65 Early Bird by 22nd October
or €75 thereafter</p>\n<p>6 products, tea/coffee and snacks plus manual with all recipes
included</p>\n<br>",
  "mtime": 1478610276660,
  "event_url": "http://www.meetup.com/Holistic-Courses-Workshops-Dublin/events/235437858/",
  "yes_rsvp_count": 1,
  "duration": 14400000,
  "payment_required": "0",
  "name": "DIY Beauty Workshop - With Mary Henry",
  "id": "235437858",
  "time": 1479034800000,
  "group": {
    "join_mode": "open",
    "country": "ie",
    "city": "Dublin",
    "name": "Holistic Courses & Workshops Dublin",
    "group_lon": -6.25,
```

KSQL Streams – Filter Processors

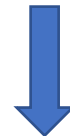
All events stream

```
CREATE STREAM MEETUP_EVENTS_STREAM (  
  key VARCHAR KEY,  
  utc_offset BIGINT,  
  venue STRUCT<country VARCHAR, city VARCHAR>  
) WITH (KAFKA_TOPIC='meetup-events', VALUE_FORMAT='JSON');
```



Filter events held in Germany

```
CREATE STREAM GERMANY_MEETUP_EVENTS_STREAM  
AS SELECT * FROM MEETUP_EVENTS_STREAM WHERE venue->country = 'de'  
EMIT CHANGES;
```



Filter events held in Munich-Germany

```
CREATE STREAM GERMANY_MUNICH_MEETUP_EVENTS_STREAM  
AS SELECT * FROM GERMANY_MEETUP_EVENTS_STREAM  
WHERE venue->city LIKE '%Munich%' or venue->city LIKE '%München%'  
EMIT CHANGES;
```


KSQL Streams – Filter Processors

GERMANY_MEETUP_EVENTS_STREAM

▼ {"UTC_OFFSET":3600000,"VENUE":{"COUNTRY":"de","CITY":"Hamburg"}}
Partition: 0 Offset: 19 Timestamp: 1658648209235

▼ {"UTC_OFFSET":3600000,"VENUE":{"COUNTRY":"de","CITY":"Hamburg"}}
Partition: 1 Offset: 24 Timestamp: 1658648209307

▼ {"UTC_OFFSET":3600000,"VENUE":{"COUNTRY":"de","CITY":"Hamburg"}}
Partition: 1 Offset: 23 Timestamp: 1658648209173

▼ {"UTC_OFFSET":3600000,"VENUE":{"COUNTRY":"de","CITY":"76133 Karlsruhe"}}
Partition: 0 Offset: 18 Timestamp: 1658648207487

GERMANY_MUNICH_MEETUP_EVENTS_STREAM

▼ {"UTC_OFFSET":3600000,"VENUE":{"COUNTRY":"de","CITY":"München"}}
Partition: 0 Offset: 2 Timestamp: 1658648284903

▼ {"UTC_OFFSET":3600000,"VENUE":{"COUNTRY":"de","CITY":"München"}}
Partition: 1 Offset: 2 Timestamp: 1658648261289

▼ {"UTC_OFFSET":3600000,"VENUE":{"COUNTRY":"de","CITY":"München"}}
Partition: 0 Offset: 4 Timestamp: 1658648340897

▼ {"UTC_OFFSET":3600000,"VENUE":{"COUNTRY":"de","CITY":"München"}}
Partition: 0 Offset: 3 Timestamp: 1658648340280

Experiment 1 - Selectivity

$$\text{Selectivity} = \frac{\#(\text{in events})}{\#(\text{out events})}$$

Selectivity of Germany filter processor



$$\frac{\#(\text{germany meetup events})}{\#(\text{meetup events})}$$

Selectivity of Germany-Munich filter processor



$$\frac{\#(\text{munich meetup events})}{\#(\text{germany meetup events})}$$

Experiment 1 - Selectivity

```
[main] INFO tum.de.App - Statistics:  
# of producing events: 75264,  
# of consuming events: 75264,  
# of germany meetup events: 765,  
# of germany-munich meetup events: 75,  
# of munich meetup events: 75
```

```
[main] INFO tum.de.App - Selectivity:  
Germany filter: 0.01016422193877551,  
Germany-Munich filter: 0.09803921568627451,
```

- Both filter processors have high selectivity
- 100 events -> 1 event held in Germany
- 10 events in Germany -> 1 event held in Munich

```
[main] INFO tum.de.App - Statistics:  
# of producing events: 37599,  
# of consuming events: 38003,  
# of germany meetup events: 479,  
# of germany-munich meetup events: 43,  
# of munich meetup events: 43  
[main] INFO tum.de.App - Selectivity:  
Germany filter: 0.012739700523950105,  
Germany-Munich filter: 0.08977035490605428,
```

```
[main] INFO tum.de.App - Statistics:  
# of producing events: 56226,  
# of consuming events: 56236,  
# of germany meetup events: 652,  
# of germany-munich meetup events: 70,  
# of munich meetup events: 70  
[main] INFO tum.de.App - Selectivity:  
Germany filter: 0.011596058762849927,  
Germany-Munich filter: 0.10736196319018405,
```

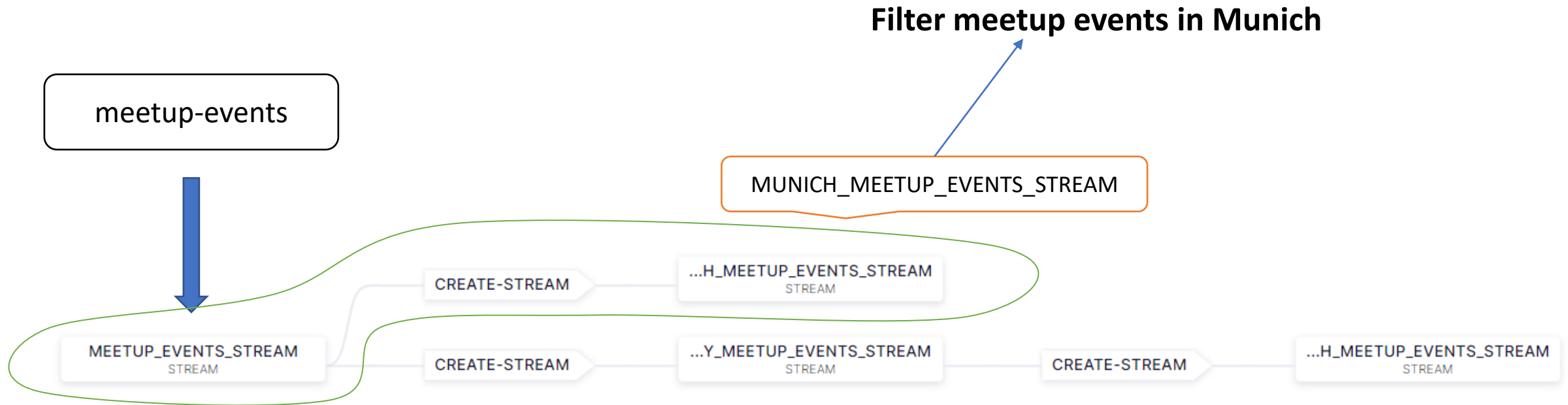
Experiment 2 – Optimization

Fact: Munich is a city in Germany



„Direct“ filter processor

Experiment 2 – Optimization



Experiment 2 – Optimization

Metric: The average latency over events (AvgLat)



$$\text{AvgLat} = \frac{\sum [T(\text{event is consumed}) - T(\text{event is produced})]}{\#(\text{consumed events})}$$

Experiment 2 – Optimization

```
[main] INFO tum.de.App - Statistics:  
# of producing events: 56226,  
# of consuming events: 56236,  
# of germany meetup events: 652,  
# of germany-munich meetup events: 70,  
# of munich meetup events: 70  
[main] INFO tum.de.App - Selectivity:  
Germany filter: 0.011596058762849927,  
Germany-Munich filter: 0.10736196319018405,  
Munich filter: 0.0012449756340483051  
[main] INFO tum.de.App - Average Latency:  
Germany-Munich filter: 199.24285714285713 ms,  
Munich filter: 104.71428571428571 ms
```

```
[main] INFO tum.de.App - Statistics:  
# of producing events: 75264,  
# of consuming events: 75264,  
# of germany meetup events: 765,  
# of germany-munich meetup events: 75,  
# of munich meetup events: 75
```

```
[main] INFO tum.de.App - Selectivity:  
Germany filter: 0.01016422193877551,  
Germany-Munich filter: 0.09803921568627451,  
Munich filter: 9.964923469387755E-4  
[main] INFO tum.de.App - Average Latency:  
Germany-Munich filter: 204.29333333333332 ms,  
Munich filter: 103.90666666666667 ms
```

Twice better than old filter processors setup ✓