# Kafka streams

Data lake is stationery

Stream – flow of data, or data in motion

When to use stream-

1. if there is any need of monitoring the data continuously like traffic and patient then use streams

Here patient data needs to monitor and send alerts to doctors

1. if there is any need to processing continuous flow of data like ETL then also use streams

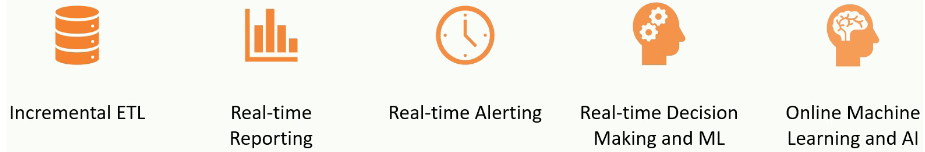
When we are dealing with data which is in motion then it is called stream processing

Definitions: - event is an outcome of an action performed by an actor

Real time use cases:-

1. Flipkart orders - Ordering event , each order in kafka when u place an order and reject an order everything is an event and placed in kafka
2. Stock market trades – real time rates

Stock market updates – means they will immediately send the updates to topic and all users will immediately get the data when all subscribers are subscribed to that topic



1. Advertisement click events- monitors when an user clicked on an add or not , if the user clicked it may go as an event and will be placed in kafka
2. All POS machine UPI transactions – when any person transacted on a pos machine it may put a message to kafka
3. Ordering food in our mobile app
4. Booking a cab- once we book it may send a message to kafka
5. In Facebook reading a post and creating a post which is a continuous flow data

If 10,000 followers are there for that user ( like sundar pichai ) means he will put that post into some topic and whoever follow him means indirectly those consumers are subscribed to that topic then immediately every will get notifications

1. ­Traffic monitoring – traffic updates may come from satellite, here streams are used to monitor this continuously flow of traffic , if high we can send them the alerts
2. Patient health monitoring- sensors will continuously send data here kafka streams will continuously monitor and do some transformations if patient readings are in danger then they will send the alerts
3. Finding fraud transactions- monitor the transaction on bank account and identify if it is fraud or not in real time and stop the transaction if it is fraud
4. Personalized customer experience- flipkart and myntra
5. Real time bidding
6. ETL scenarios- ETL- Extract, transform, load in ETL requirements after extracting it will be like streams and do some transformations and insert data into db

Data coming into my feed, If someone liked by post, commented and tagged me and who commented and liked our post and someone tagged you or if someone tagged ur friend all those events may flow into ur database partition

Misc use cases

1. Infrastructure management –monitor cpu health ram, rom
2. Campaign management- whether recent offers in myntra is driving traffic or not

## Design considerations for data integration solutions

2 applications should be integrated by following points

1. Time sensitivity 2) Decoupling 3)Data format evolution 4) Reliability 5) scalability 6)D
2. Decoupling

Says we should minimize the dependency between source and destination systems within milliseconds we should consume it should be very fast

1. We cant constraint the data format , source system may not send messages to target system in desired formats we should have flexibility to convert the messages

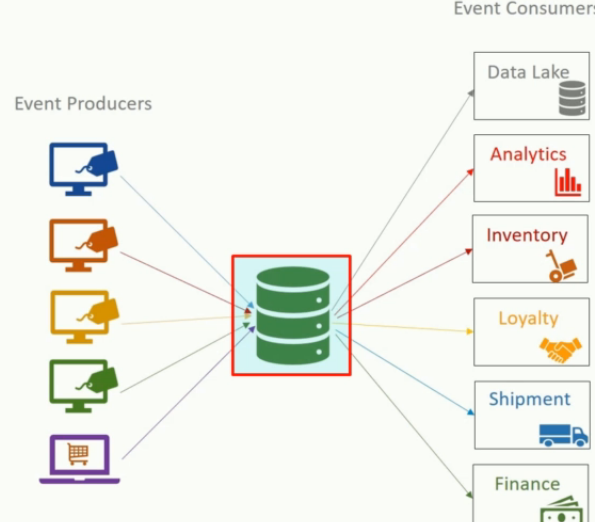
Src may send json and target may need only xml

1. Reliability- even if the consumer went down offline for some times messages should be stored until customer comes online – kafka does that, all messages will be stored in broker partition’s file segment means messages are written to files so those are reliable
2. Scalability – we want horizontal scaling this data integration solution must support at least millions of messages per second

Ways of communication

1. Shared database
2. File transfer – here will transfer file by file
3. Messaging
4. RPC remote procedure call – Its is by rest calls – worst for scalable application

### **1) Shared database**

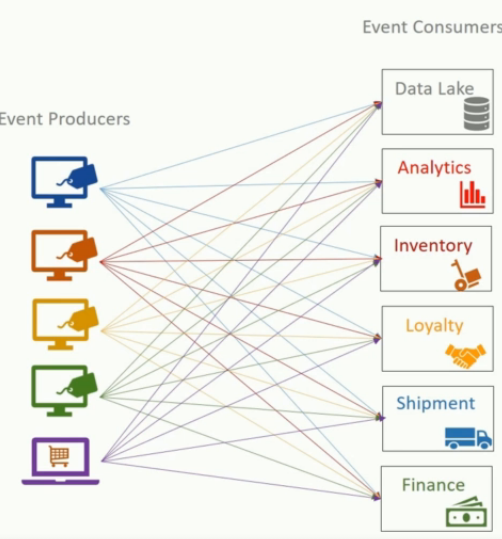
1) This is a good approach for storing now and query later not good to deal with the data in motion and not good for realtime and to deal in milliseconds

2) reg decoupling – all consumers may not support RDBMS and 3) not good for data format evolution because producer may send json and consumer may need xml

5) and this is not horizontally scalable not but the problem is   
**RMI- RPC**

Fails on decoupling (data exchange formats) and scalability , for million messages should we do million rest calls

### **Why we need Kafka Messaging than REST service calls**



**1) consumer offline :-** remote method invocation this is like REST service call, we can’t send millions of data over REST, main issue – what if the consumer is offline so is chain broken

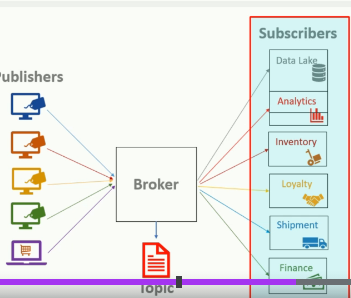
, in kafka all messages will be stored in topic-folder-files whenever he is online he can consume all the messages from the beginning

1. **Resend same message to all consumers –** un necessary REST calls if we want to send millions of same messages to all 5-6 consumers its terrible to make 6 REST calls to 6 services if one is down how can the producer send the same data again

**Ex:**- a) sending invoice message to all 5 consumers, if kafka topics are not used then we have to do same rest call to all 5 consumers so 4 times waste REST svc call

b) That to there is no guarantee that all 5 consumers will be online

Hence if same message is needed by multiple consumers then use kafka

   
Here all publishers will send message to the broker, consumer will poll for the messages

Kafka Streams API vs Kafka Client API

If u have kafka transformation logics then u might need kafka streams API, else for ordinary consumption and storing u can happily use kafka client API

Features of Kafka streams

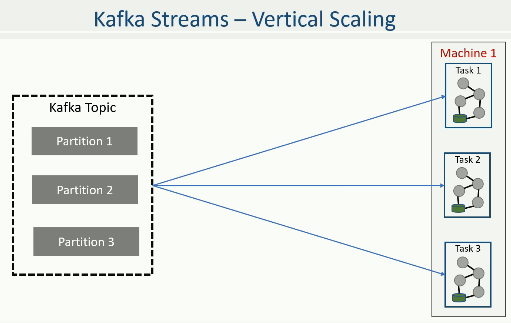
These are not like java streams, java streams once we have used we cannot re use the same stream

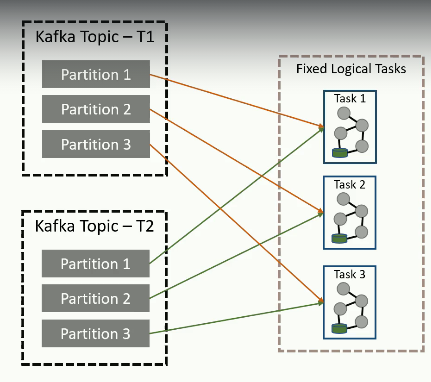
Whereas kafka streams We can reuse them.

In java, we will create a stream from the collection , similarly here we will create a stream from a topic, wherever u go we will create a stream on some data source

Streams vertical scaling

Vertical scaling means, scaling on same machine/same hardware, in single machine multiple threads



the framework will see how many partitions are then it will create those many logical task

no of logical task=no of partitions

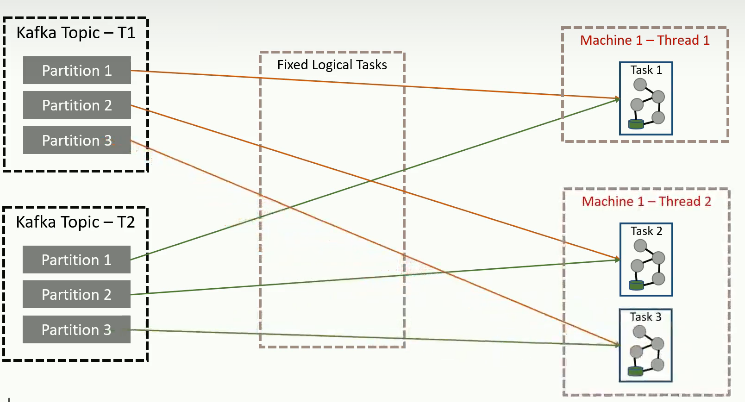
if u have only 1 thread all tasks will be performed by single thread

if u have 3 threads then each thread will perform 1 task

To configure vertical scaling we have to set a property

Props.put(Streamsconfig.NUM\_STREAM\_THREADS\_CONFIG,3);

Once u build the topology, logical task is equal to number to partitions, each logical task Is performed by separate thread , in the above if u have 3 available threads all 3 tasks will be started by 3 separate threads , if 3 threads are not available then if only 2 are available ,1st thread will handle 1 task and second thread will only handle 2 task as per below image



Since only 2 threads are available they shared the tasks among themelves

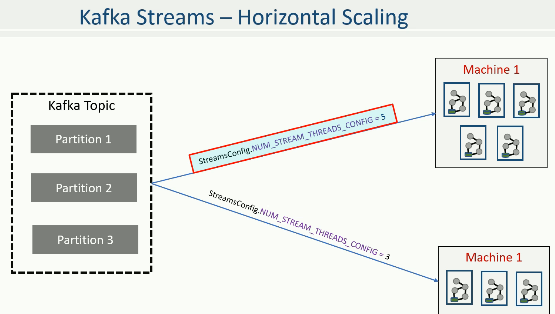
Streams horizontal scaling

It is like launching same app in multiple computers

You can scale application horizontally by starting your streams application in different computers

Same like we deploy our code in multiple servers in production / like deploying your application

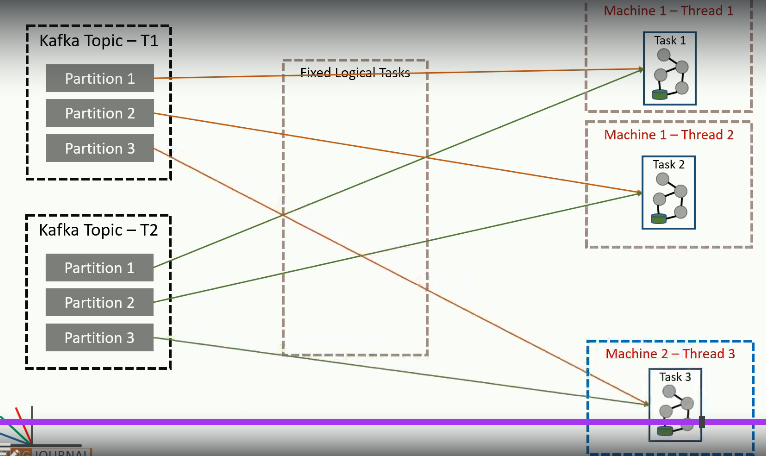
In servers like oxmore/shoreview ,its like we create our streams application and generally we deploy our app in prod in 3-4 servers



Horizontal and vertical scaling

Simple first do vertical scaling by configuring 5 threads in your application and later to implement horizontal scaling just deploy the same application in multiple servers lets say we have deployed in 4 servers

Therefore in each server 5 threads and total 4 servers =5\*4 =20 threads are there in all vms

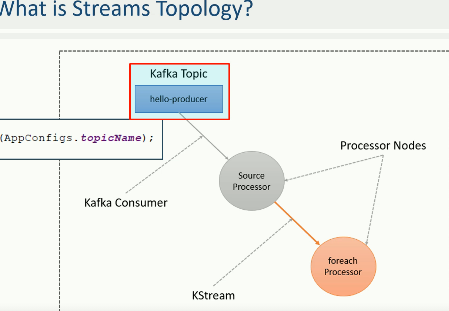


Here we did horizontal scaling and vertical scaling , each machine we have separate threads , example 1st machine have 2 threads and we did horizontal scaling we deployed the same application in second machine also which have 1 thread so they have share the logical tasks among themselves

Ideally there should be 6 threads, each machine should have 3 threads , but in 1st machine as we don’t have enough 3 as we have only 2 it used 2 threads and in 2nd machine also we should have 3 threads but only1 is available so used that

Kafka Terminologies

Source processor vs sink processor

Source processor (it will process from source) will internally uses a kafka consumer to read data from topic

This Source processor (kafka consumer) can consume data from multiple topics

Sink processor will internally uses a kafka producer to write to a topic

These processors can read & write the data to the topic  
For Source processor input is topic and this source processor will internally create a consumer and create a stream on that topic

`

Methods of kstream

Filter(),Map(),flatMap() ,mapValues(),flatMapValues()– these are intermediate operations and these methods will create another streams

forEach(), to() – these methods are terminal operations,

to() method is a sink processor , it will take the stream and write the data to that topic

#### Remaining misc classes

Produced Object- used on KStream.to(String topicName,Produced p) this is a sink processor which will write the data to the topic, this object is used to hold

Consumed object

Sample program

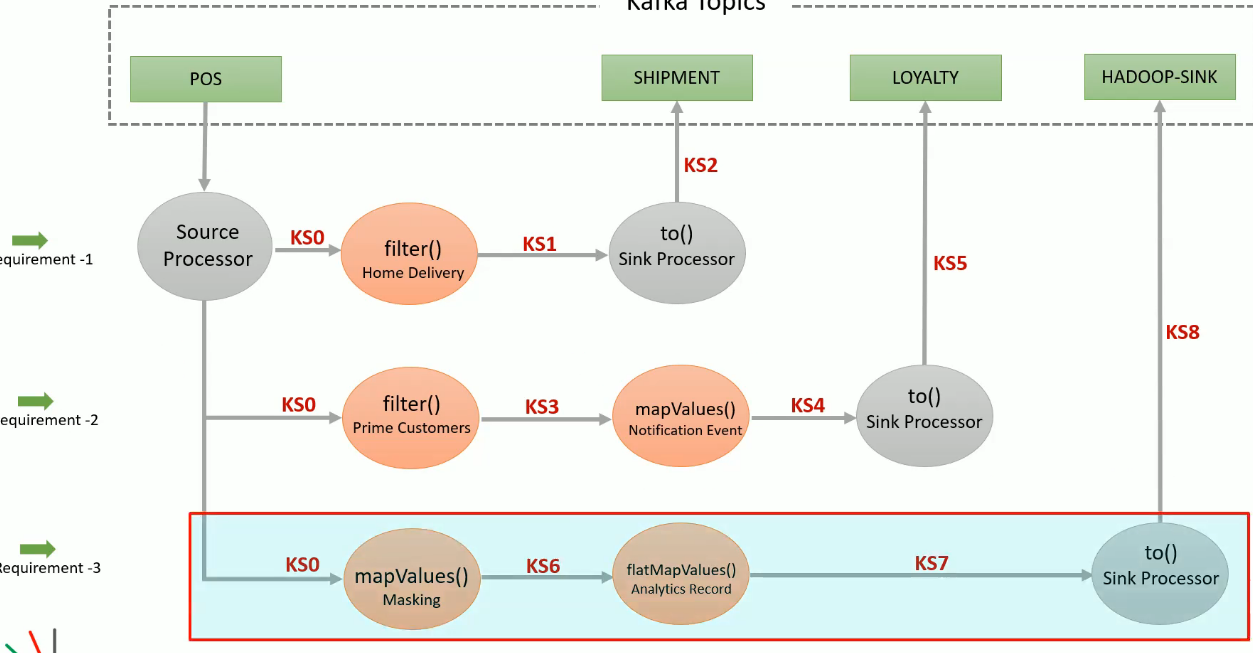
Requirement

1) Select Invoices where DeliveryType = "HOME-DELIVERY" and push them to the shipment service queue.

2. Select Invoices where CustomerType = "PRIME" and create a notification event for the Loyalty Management Service. The format for the new notification event is given here.

3. Select all Invoices, mask the personal information, and create records for Trend Analytics. When the records are ready, persist them to Hadoop storage for batch analytics. The format for the new Hadoop record is also given.

DAG (Direct acyclic graph)



Sample codes

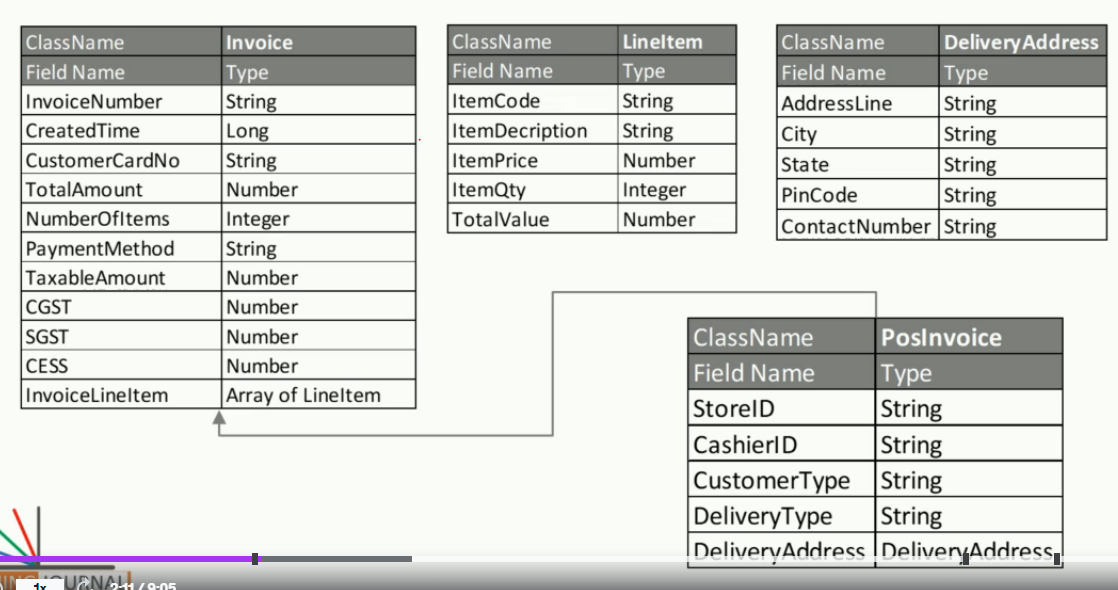
Convert JSON to POJO class

There is a jar called json schema to pojo use that jar and convert

<groupId>org.jsonschema2pojo</groupId>

<artifactId>jsonschema2pojo-maven-plugin</artifactId>

<version>1.1.3</version>



I still didn’t understand why to create a json object that to in below format

Instead directly we can create java classes instead of creating the json objects and use some bloody plugin generate java code

1. First prepare the Example json object

{

"type": "object",

"javaType": "guru.learningjournal.kafka.examples.types.Invoice",

"properties": {

"InvoiceNumber": {"type": "string"},

"CreatedTime": {"type": "object", "javaType": "java.lang.Long"},

"CustomerCardNo": {"type": "string"},

}

First we have to prepare this kind of json object, my suggestion instead of this better we can create java class directly after preparing the json we have to add the below plugin in pom.xml

1. Add 2 maven plugins- json plugin and maven compiler plugin in pom.xml and add related Jackson jars

<plugins>

<plugin>

<groupId>org.jsonschema2pojo</groupId>

<artifactId>jsonschema2pojo-maven-plugin</artifactId>

<version>1.1.3</version>

<plugin>  
 <groupId>org.apache.maven.plugins</groupId>  
 <artifactId>maven-compiler-plugin</artifactId>  
 <version>3.8.0</version>

Refer this page for full information<https://github.com/joelittlejohn/jsonschema2pojo/wiki/Getting-Started>

1. Then run the maven task called “compile” when u run this task then …. That will create the java classes
2. Refer project “09-json-to-pojo” project to and see full code

Creating serdes

If u are not satisfied with existing serdes write a class extending org.apache.kafka.common.serialization.Serdes

Avro to pojo

As of now avro wont support extending classes

If u want avro then let all properties be in single class , why this many conversions why cant u take json directly

1. Create avro files .asc same like json files where we will create pojos for the avro
2. Add maven compiler plugin, maven avro plugin (for that group id of that plugin will be avro
3. Add avro jars (org.apache.avro- avro-1.8.1)