**Kafka Questions**

Apache Kafka Foundation Course - What is Apache Kafka

<https://www.learningjournal.guru/courses/kafka/kafka-foundation-training/what-is-kafka/>

**delete.topic.enable**

If you want to delete a topic, you can use topic management tool. But by default, deleting a topic is not allowed. You can't remove a topic because the default value for this parameter is false. That is reasonable protection for production environments. But in development or testing environment, you may want to delete topics. So, if you want Kafka to allow deleting a topic, you need to set this parameter to true.

**auto.create.topics.enable**

We have already discussed auto-create topic feature. If a producer starts sending messages to a non-existent topic, Kafka will create the topic automatically and accept the data. This behaviour is suitable for dev environments. But in a production environment, you may want to implement a more controlled approach. You can set this parameter to false, and Kafka will stop creating topics automatically. You can create topics manually using the topic management tool, and no one will be able to send data to a non-existent topic.

**default.replication.factor and num.partitions**

These two parameters are quite straightforward. The default values for both of them is one, and they are effective when you have auto create topics enabled. So, if Kafka is creating your topic automatically, the new topic will have only one partition and a single copy. If you want some other values, you can change the default settings accordingly.

**log.retention.ms and log.retention.bytes**

These two are critical and not obvious. So, whatever data you send to Kafka, it is not retained by Kafka forever. Kafka is not a database. You don't send data to Kafka for storage so that you can query it later. It is a message broker. It should deliver the data to the consumer and then clean it up. There is no reason to retain messages for longer than needed.  
Kafka gives you two options to configure the retention period. The default option is retention by time, and the default retention period is seven days. So, in this case, Kafka will clean up all the messages older than seven days. If you want to change the duration, you can specify your value for log.retention.ms configuration.  
Kafka gives you another option to define this retention period. You can specify it by size. That's where the second parameter log retention bytes is applicable. But this size applies to partition. So, if you set log.retention.bytes = 1 GB, Kafka will trigger a clean-up activity when the partition size reaches to 1 GB. Remember that it is not a topic size. It is partition size.

kafka Producers common Properties configurations interview questions :

Acknowledgements-

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Diagram

Description automatically generated

Text

Description automatically generated with medium confidence

Text

Description automatically generated with medium confidence

[Diagram

Description automatically generated](https://www.learningjournal.guru/_resources/img/jpg-7x/kafka-producer-workflow.jpg)Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated**max.in.flight.requests.per.connection : 5. It means producer don’t want to wait for acknowledgement for each request it he wants only wait on every Fifth request for ack.**

Apache Kafka Foundation Course - What is Apache Kafka

Welcome to Apache Kafka Tutorial at Learning journal. In this session, I will introduce you to Kafka. We will try to understand Kafka in less than 10 minutes. I am assuming that you have at least heard about Kafka and you already know that it is an Open Source project. Kafka was initially developed at LinkedIn and later open sourced in 2011. Since then it has evolved and established itself as a standard tool for building real-time data pipelines. Now it's securing its share in real-time streaming applications as well.  
The Kafka documentation says it is a distributed streaming platform. That's good for definition. But I want to know what it can do for me or what I can do using Kafka.

**Messaging System**

The official documentation says that Apache Kafka is similar to enterprise messaging system. I guess, you already understand a messaging system. In a typical messaging system, there are three components.

Producer or Publisher

Broker

Consumer.

[Diagram

Description automatically generated](https://www.learningjournal.guru/_resources/img/jpg-5x/typical-messaging-system.jpg)Fig.1-Typical Messaging System

The producers are the client applications, and they send some messages.  
The Brokers receive those messages from publishers and store them.

The consumers read the message records from brokers.

**Kafka Use Case**

A messaging system looks very simple. Now let us look at the data integration problem in a large organization. I borrowed the below diagram from Jey Creps blog.

[Diagram

Description automatically generated](https://www.learningjournal.guru/_resources/img/jpg-5x/data-integration-problem.jpg)Fig.1-Data Integration Problem

The above diagram shows the data integration requirement in a large enterprise.  
Does it look like a mess?  
There are many source systems and multiple destination systems. And you are given a task to create data pipelines to move data among those systems. For a growing company, the number of source and destination systems keep getting bigger and bigger. Finally, your data pipeline looks like a mess. I am sure that I don't need to explain that you can't manage and maintain that kind of data pipeline. Some part of your pipeline will keep breaking every day.  
However, if we can use a messaging system for solving that kind of integration problem, the solution may be neater, and cleaner as shown below.

[Diagram

Description automatically generated](https://www.learningjournal.guru/_resources/img/jpg-5x/unified-messaging-system.jpg)Fig.1-Unified Messaging System

That's the idea discovered by the team at LinkedIn. Then they started evaluating existing messaging systems, but none of them meet their criteria to support the desired throughput and scale. Finally, they end up creating Kafka.

**What is Kafka?**

At the core, Kafka is a highly scalable and fault tolerant enterprise messaging system. Take a look at the [Apache Kafka diagram](https://kafka.apache.org/intro.html#_blank) from official documentation. I hope you understand the producer, consumer and the broker that the figure shows. At the top of the diagram, the Producer applications are sending messages to Kafka cluster. The Kafka cluster is nothing but a bunch of brokers running in a group of computers. They take message records from producers and store it in Kafka message log.  
At the bottom of the picture, there are consumer applications. They read messages from Kafka cluster, processes it and do whatever they want to do. They may want to send them to Hadoop, Cassandra, HBase or may be pushing it back again into Kafka for someone else to read these modified or transformed records.

**Kafka Streams**

**What is a stream?**  
Well, I will say continuous flow of data. or you can define it as a constant stream of messages.  
Kafka, as a messaging system is so powerful regarding throughput and scalability that it allows you to handle a continuous stream of messages. If you can just plug in some stream processing framework to Kafka, it could be your backbone infrastructure to create a real-time stream processing application. And that is what right side of the diagram is trying to explain. Those are some stream processing applications. They read a continuous stream of data from Kafka, process them and then either store them back in Kafka or send them directly to other systems. Kafka provides some stream processing APIs as well. So you can do a lot of things using Kafka stream processing APIs, or you can use other stream processing frameworks like Spark streaming or Storm.

**Kafka Connect**

The next thing is Kafka connector. These are the most compelling features. They are ready to use connectors to import data from databases into Kafka or export data from Kafka to databases. These are not just out of the box connectors but also a framework to build specialized connectors for any other application.

Summary

**Let us summarize all that we learned in this session.**

Kafka is a distributed streaming platform. You can use it as an enterprise messaging system. That doesn't mean just a traditional messaging system. You can use it to simplify complex data pipelines that are made up of a vast number of consumers and producers.

You can use it as a stream processing platform. There are two parts of stream processing. Stream and a Processing framework. Kafka gives you a stream, and you can plug in a processing framework.

Kafka also provides connectors to export and import bulk data from databases and other systems.

But implementing these things is not that simple. There is no plug and play component. You need to use APIs and write a bunch of code. You need to understand some configuration parameters and tune or customize Kafka behavior according to your requirement and use case.  
We will cover all these things in this training. So, keep watching.

**basic concepts associated with Kafka**

The objective of this article is to introduce you to the main terminologies and build a foundation to understand and grasp rest of the training.

Producer

Consumer

Broker

Cluster

Topic

Partitions

Offset

**Consumer Groups**

We will be using these terms extensively during our discussion of Apache Kafka. It is crucial that we both, myself and you have the same understanding of these concepts. So let me explain my understanding of these terms and some other related concepts associated with these keywords.

**Kafka Producer**

The first item is the producer. So, what is a producer?  
The producer is an application that sends data. Some people call it data, but we will call it a message or a message record. These messages can be anything ranging from a simple string to a complex object. Ultimately it is small to the medium-size piece of data. The message may have different meaning or schema for us. But for Kafka, it is a simple array of bytes.  
For example, if I want to send a file to Kafka, I will create a producer application and push each line of the file as a message. In this case, a message is one line of text. But for Kafka, it is just an array of bytes. Similarly, If I want to send all the records from a table, I will submit each row as a message, or if I want to send the result of a query. I will create a producer application, fire a query against my database, collect the result and start throwing each row as a message. So, while working with Kafka, if you want to send some data, you have to create a producer application. It is unlikely that you get a readymade producer that fits your purpose.

**Kafka Consumer**

The next thing is the consumer. The consumer is again an application that receives data. If producers are sending data, they must be sending it to someone. Right? The consumers are the recipients. But remember that the producers don't send data to a recipient address. They just send it to Kafka server. And anyone who is interested in that data can come forward and take it from Kafka server. So, any application that requests data from a Kafka server is a consumer, and they can ask for data send by any producer provided they have permissions to read it.  
So just continuing the file example, If I want to read the file sent by a producer, I will create a consumer application, then I will request Kafka for the data. The Kafka server will send me some messages. I think you remember that each message is a line of text in this example.  
So, the client application will receive some lines from Kafka server, it will process them and again request for some more messages. The client keeps demanding data, and Kafka server will keep giving message records as long as new messages are coming from the producer.

**Kafka Broker**

Now, let’s move on and try to understand a Broker. The broker is Kafka server. It is just a meaningful name given to Kafka server. And this title makes sense as well because all that Kafka does is act as a message broker between producer and consumer. The producer and consumer don not interact directly. They use Kafka server as an agent or a broker to exchange messages.

**Kafka Cluster**

Let's come to the next term. The cluster. This one is simple. If you have any background in distributed systems, you already know that a Cluster is a group of computers acting together for a common purpose. Since Kafka is a distributed system, so the cluster has the same meaning for Kafka. It is merely a group of computers, each executing one instance of Kafka broker.

**Kafka Topic**

Next item is the topic. We learned that producer sends data to Kafka broker. Then a consumer can ask for data from the Kafka broker. But the question is, Which data?  
Let's try to understand this by a simple conversation between Broker and the consumer.  
  
**Broker -** I am collecting data from multiple producers, which one do you want?  
**Consumer** - Give the data sent by producer ABC.  
**Broker -** Oh Man, producer ABC is pushing three different types of records. Which one do you want?  
**Consumer -** Well, send me the sales data.  
**Broker -** Ok, so you are looking for sales data. Two more producers are sending sales data.  
**Consumer -** Gosh, we need to have some identification mechanism.  
  
There comes the notion of the Kafka Topic. So, the topic is an arbitrary name given to a data set. We better say that it is a unique name for a data stream.  
For example, we create a topic called Global Orders, and every point of sales may have a producer. They send their order details as a message to the single Topic named Global Orders. And a subscriber interested in Orders can subscribe to the same Topic.

**Kafka Partitions**

By now, you learned that the broker would store data for a topic. This data can be enormous. It may be larger than the storage capacity of a single computer. In that case, the broker may have a challenge in storing that data. One of the obvious solutions is to break it into two or more parts and distribute it to multiple computers. Kafka is a distributed system that runs on a cluster of machines. So, it is self-evident that Kafka can break a topic into partitions and store one partition on one computer. And that's what the Partition means.  
You may be wondering that how Kafka will decide on the number of partitions. I mean, some topics may be large, but others may be relatively small. So how Kafka knows that it should create 100 partitions or just ten partitions could be enough?  
The answer is simple. Kafka doesn't take that decision. We, as a developer make that decision. When we create a topic, we make that decision, and Kafka broker will create that many partitions for our Topic. But remember that every Partition sits on a single machine. You can't break it again. So, do some estimation and simple math to calculate the number of partitions.

**Offsets**

Let's talk about offset. The offset is simple. It is a sequence number of a message in a partition. This number is assigned as the messages arrive in a partition. And these numbers, once assigned, they never change. They are immutable. This sequencing means that Kafka stores messages in the order of arrival within a partition. The first message gets an offset zero. The next message receives an offset one and so on. But remember that there is no global offset across partitions. Offsets are local to the partition. So, if you want to locate a message, you should know three things.  
Topic name, Partition number, and an offset number. If you have these three things, you can directly locate a message.

**Kafka Consumer Groups**

Now we are left with the last thing. The consumer groups. We already understand the Consumer. What is Consumer Group?  
It is a group of consumers. Several Consumers form a group to share the work. You can think of it like there is one large task and you want to divide it among multiple people, so you create a group, and members of the same group share the work. Let me give you an example.

A Kafka Example

Let's assume that we have a retail chain. In every store, there are few billing counters. You want to bring all the invoices from every billing counter to your data centre. Since you learned Kafka and you find Kafka as an excellent solution to transport data from billing locations to the data centre. You decided to implement it. The first thing you might want to do is to create a producer at every billing site. These Producers will send bills as a message to a Kafka Topic. The next thing you might want to do is to create a consumer. The Consumer will read data from Kafka Topic and write them into your data centre. It sounds like a perfect solution. Right? But there is a small problem. Think of the scale. You have hundreds of producers pushing data into a single topic. How will you handle that volume and velocity?

[Diagram

Description automatically generated](https://www.learningjournal.guru/_resources/img/jpg-7x/apache-kafka-example.jpg)Fig.1-Apache Kafka Example Use Case

You learned Kafka exceptionally well. So, you decided to create large Kafka cluster and partition your Topic. Correct? So, your Topic is partitioned and distributed across the Cluster. Now several brokers are sharing the workload to receive and store data. From the source side, you have many producer and several Brokers to share the workload. What about the destination side? You have a single unfortunate consumer.  
There comes the Consumer group. You create a Consumer group and start executing many Consumers and tell them to divide the work.  
So far so good. But how do we split the work? That's not a difficult question. I have 600 partitions. And I am starting 100 consumers. So why don't each of the consumer take six partitions? We will see, if they can't handle six partitions, we will start some more Consumers in the same group. We can go up to 600 Consumers, so each consumer will have just one partition to read.  
If you followed this example correctly, you understand that partitioning and consumer group is a tool for scalability. And notice that the maximum number of Consumers in a group is the total number of partitions you have on a topic. Kafka doesn't allow more than one Consumer to read from the same partition simultaneously. This restriction is necessary to avoid double reading of records.  
Great. I hope you learned core concepts of Kafka. Now you are familiar with the essential terminology that we will be using throughout the Kafka tutorials.