SAKARYA ÜNİVERSİTESİ

BIG DATA

TOO BIG TO IGNORE

SÜMEYYE KAYNAK

OUTLINE

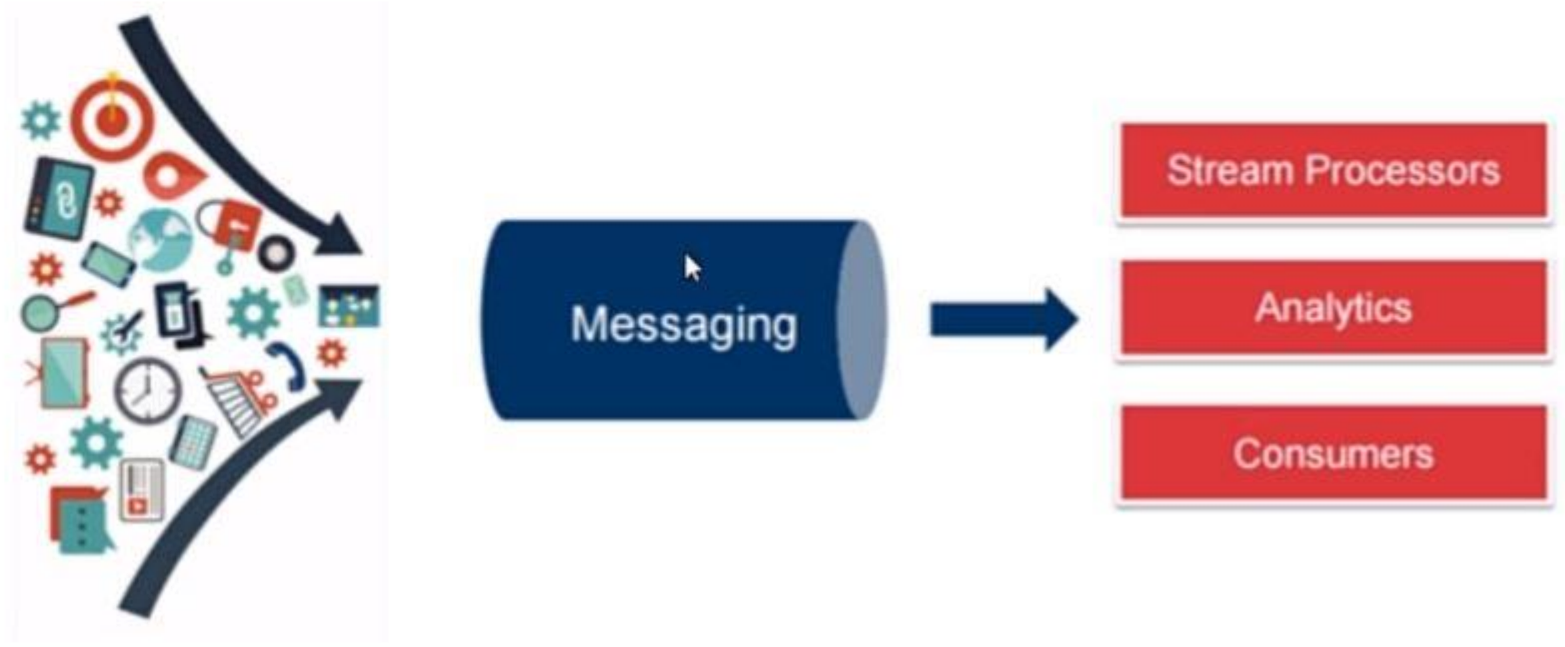
Apache Kafka

Apache Zookeeper

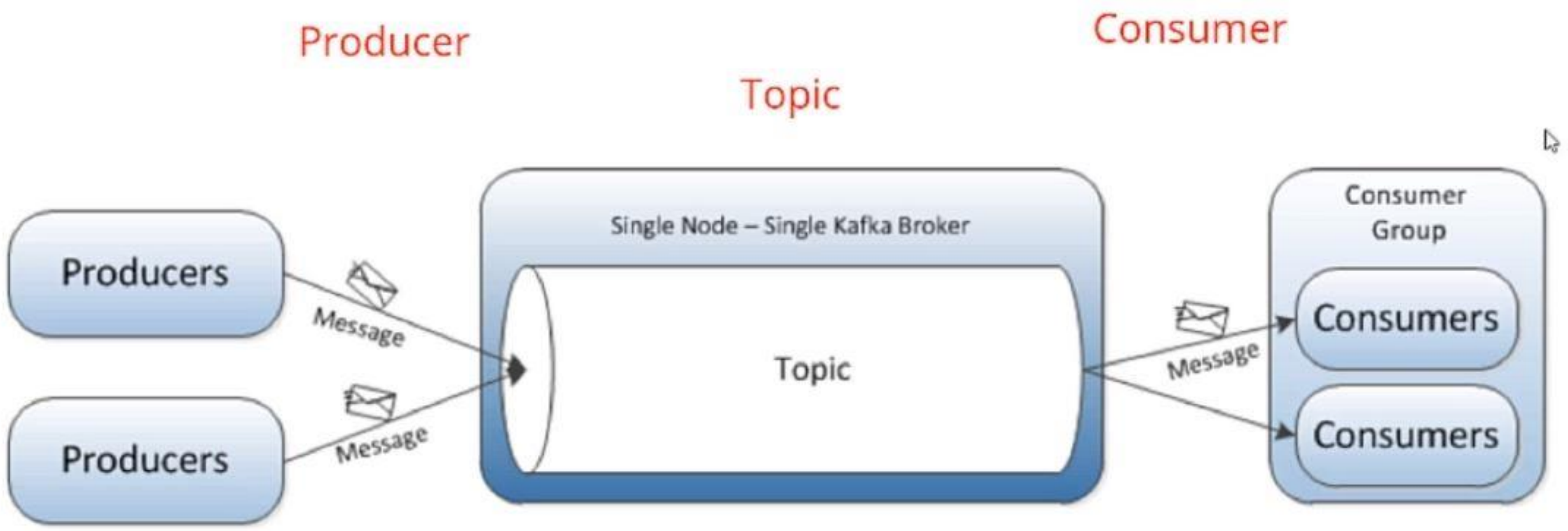
Docker

APACHE KAFKA

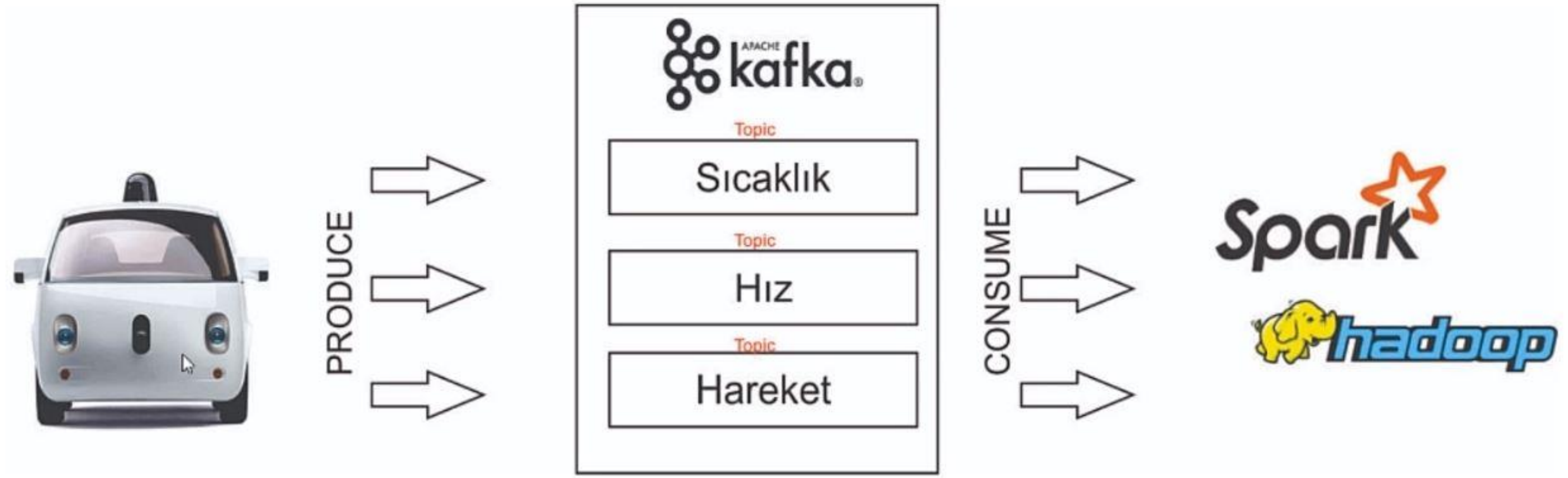
- It is very important to collect and analyze big data quickly and without errors.



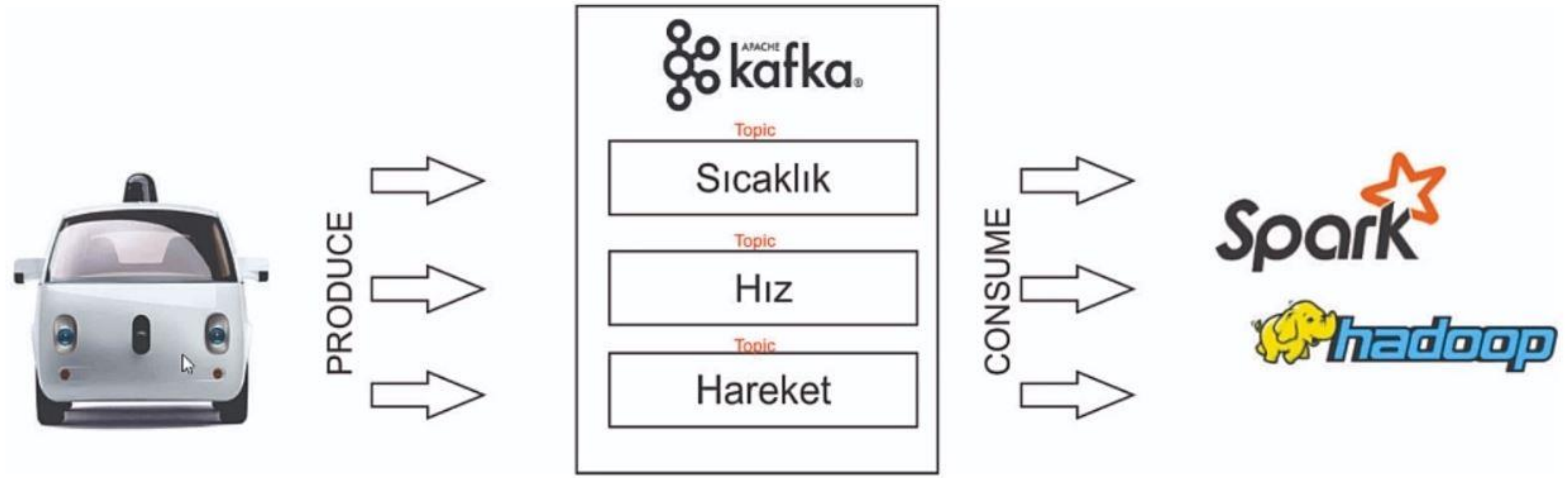
APACHE KAFKA



APACHE KAFKA

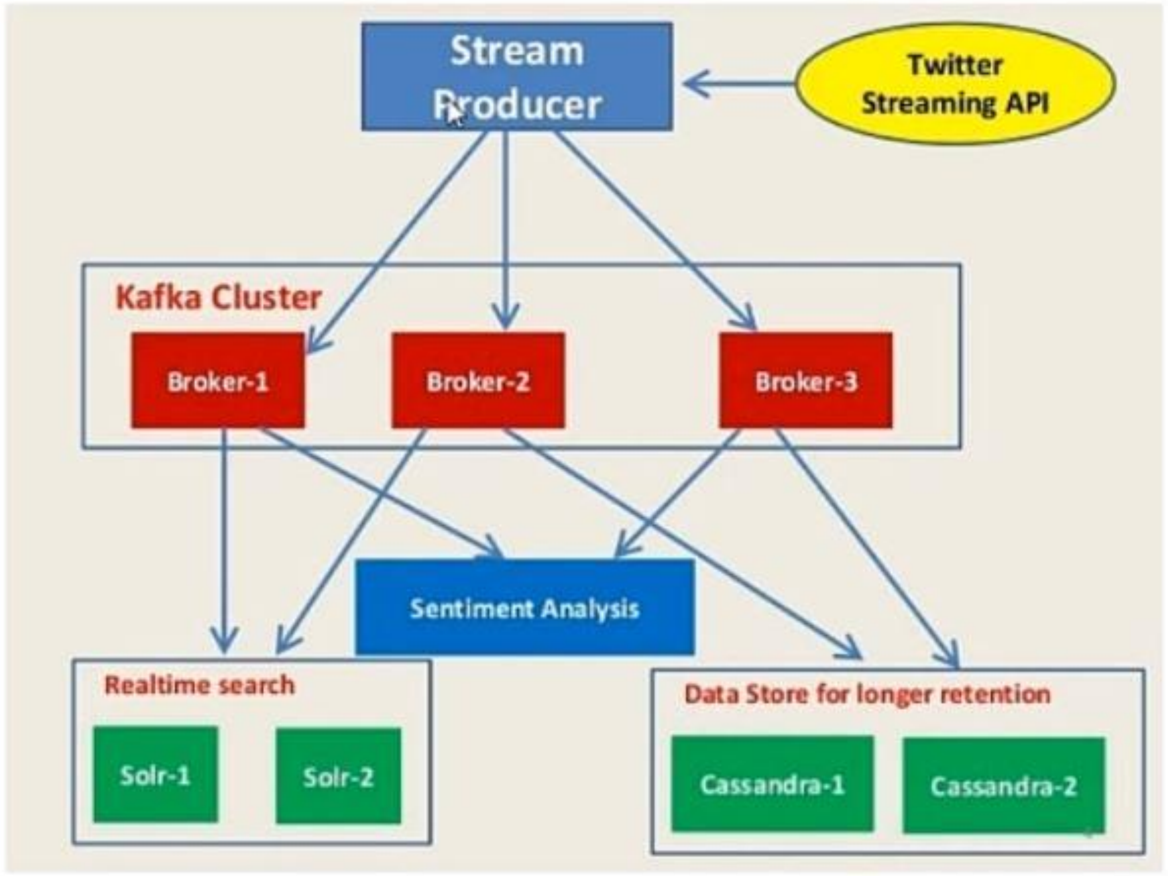


APACHE KAFKA



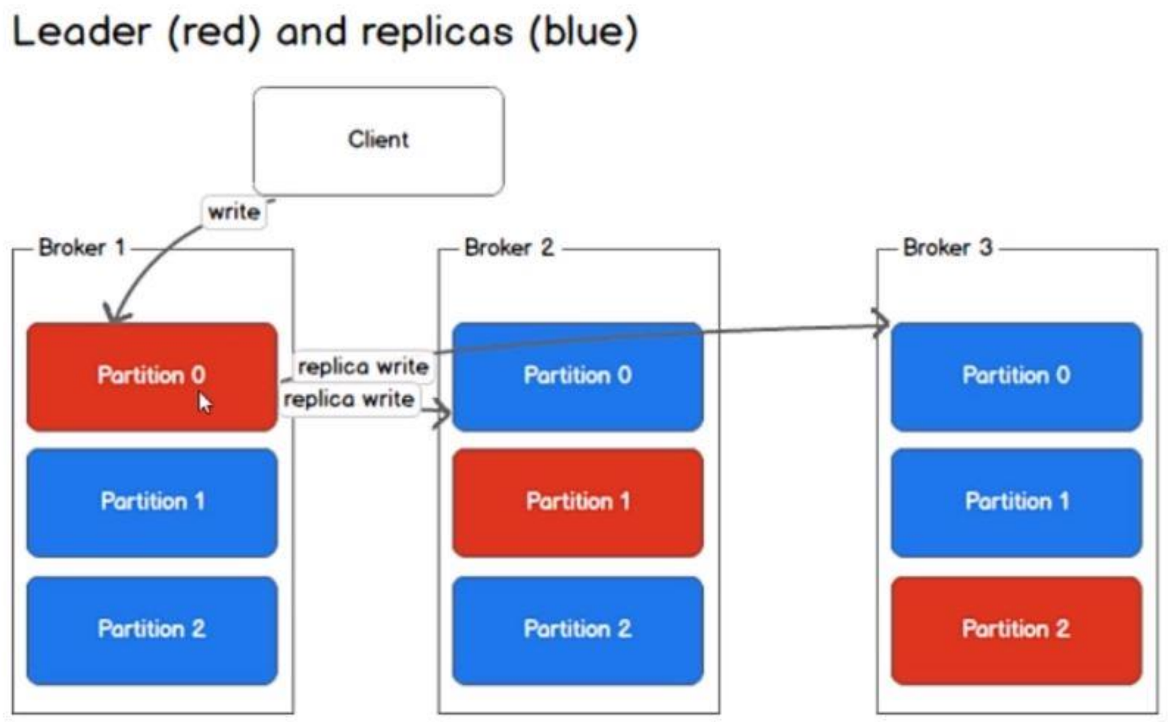
APACHE KAFKA

Distributed data storage and replication are available in Kafka.



APACHE KAFKA

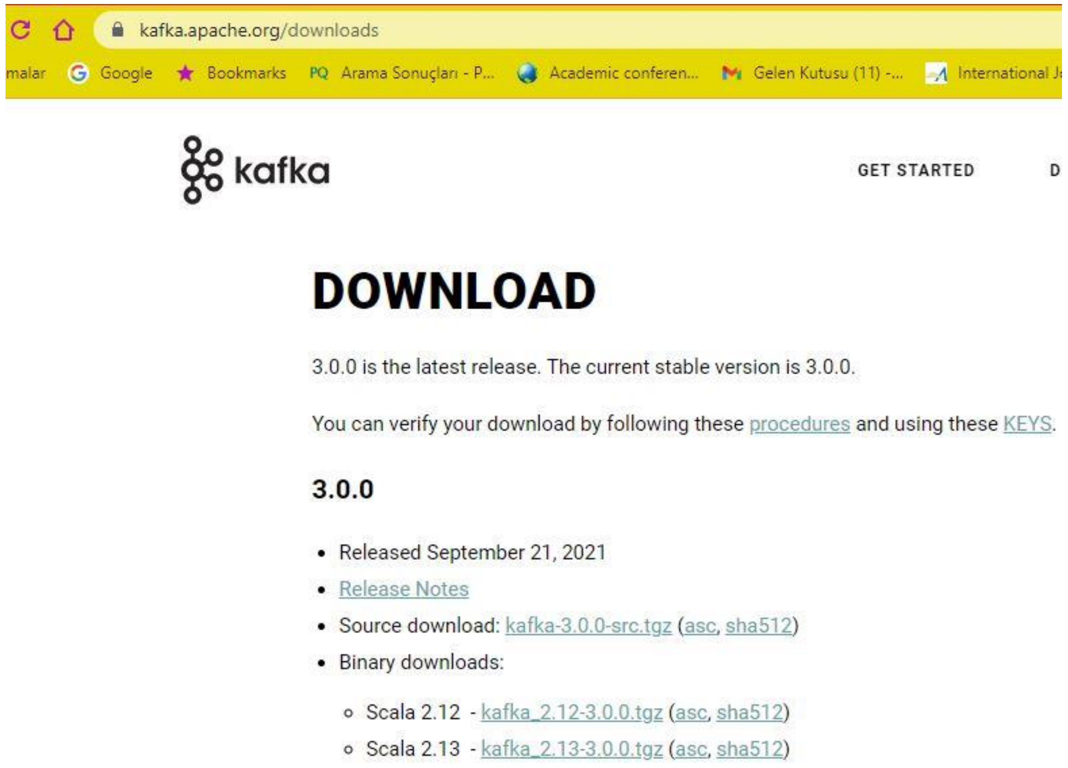
- A file is divided into 3 parts.
- Partition 0 of the file is the leader in broker1.
- Replication factor value is 3.
- The leading part in each machine changes.
- Data loss is prevented.



ZOOKEEPER

- Zookeeper coordinates resource management in distributed server architectures.
- Zookeeper is generally used for configuration and keeps configuration files.

DOWNLOAD KAFKA



DOWNLOAD

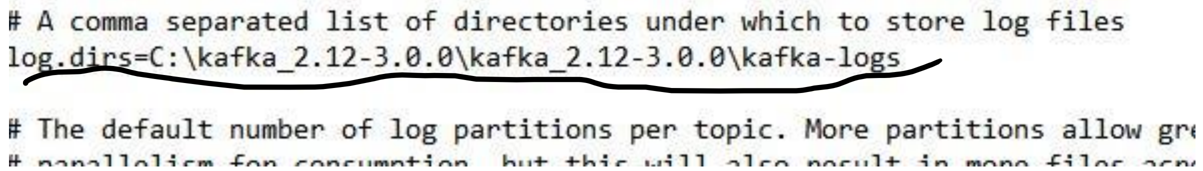
3.0.0 is the latest release. The current stable version is 3.0.0.

You can verify your download by following these procedures and using these KEYS.

3.0.0

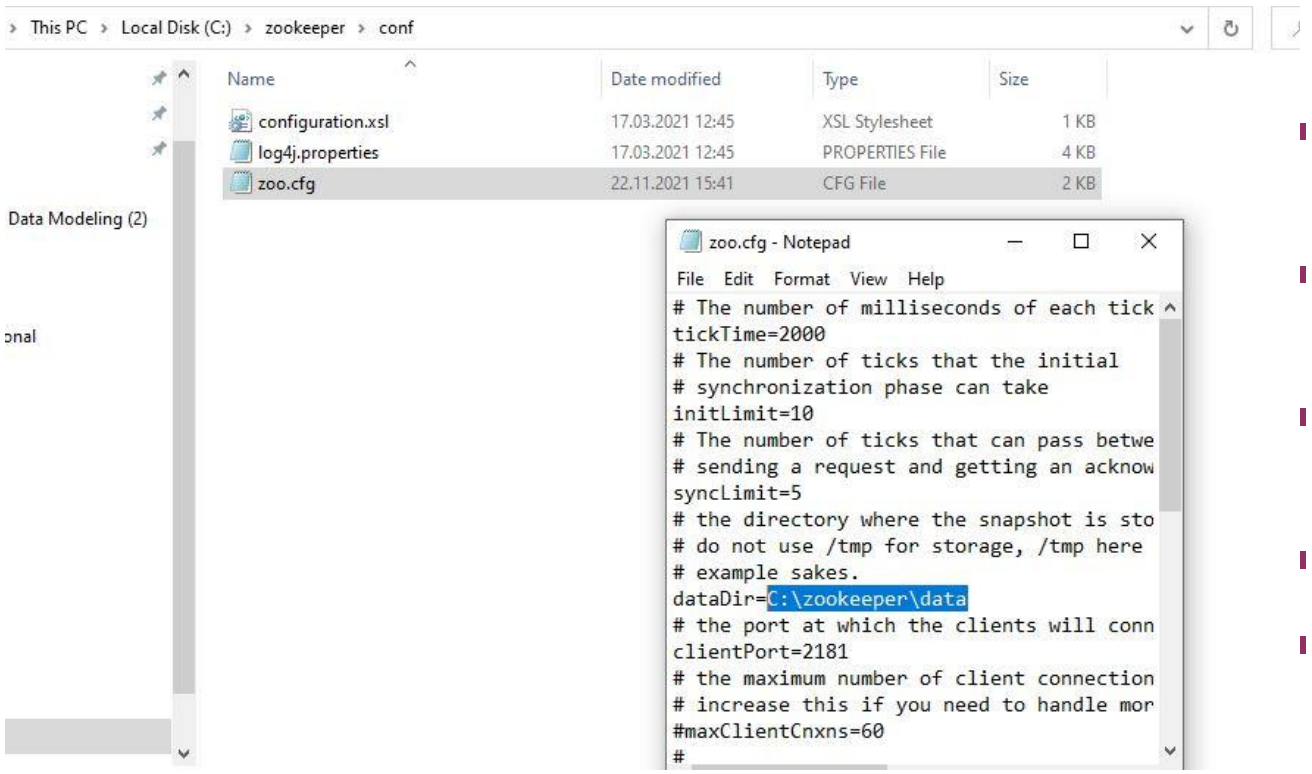
- Released September 21, 2021
- Release Notes
- Source download: kafka-3.0.0-src.tgz (asc, sha512)
- Binary downloads:
  - Scala 2.12 - kafka_2.12-3.0.0.tgz (asc, sha512)
  - Scala 2.13 - kafka_2.13-3.0.0.tgz (asc, sha512)

- Download the appropriate version from the Kafka site.
- The downloaded zip file is extracted to the C: directory.
- Update the server.properties file in the config folder.



```
# A comma separated list of directories under which to store log files
log.dirs=C:\kafka_2.12-3.0.0\kafka_2.12-3.0.0\kafka-logs

# The default number of log partitions per topic. More partitions allow gre
# parallelism for consumption, but this will also result in more files acro
```

DOWNLOAD ZOOKEEPER



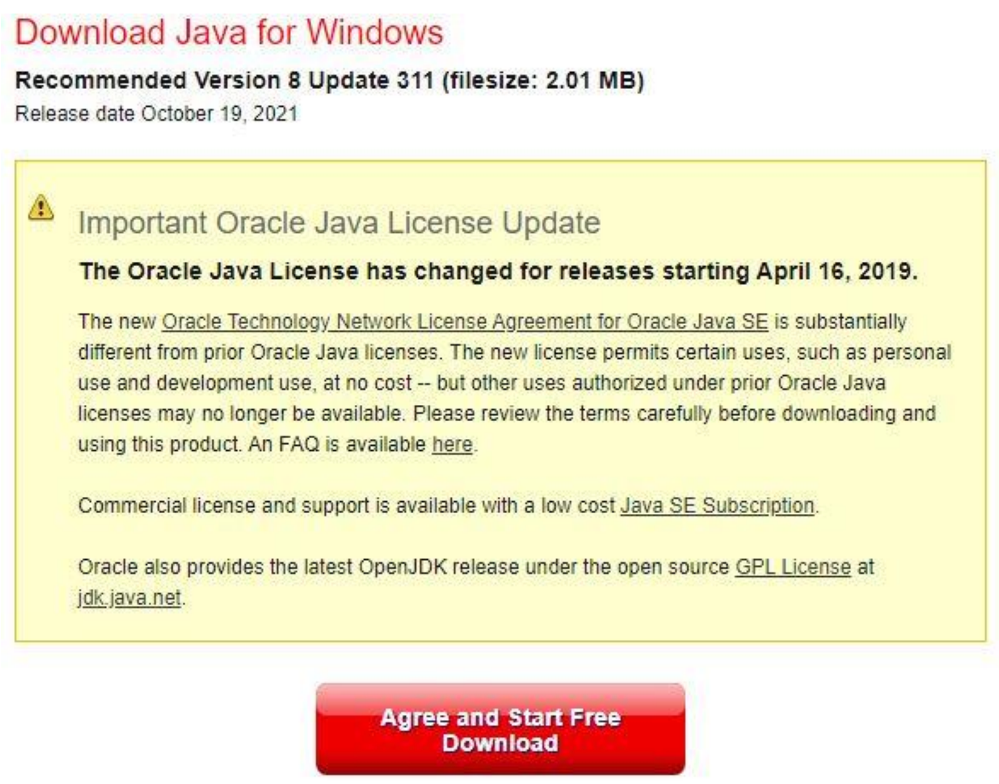
This PC › Local Disk (C:) › zookeeper › conf
Data Modeling (2)
onal
zoo.cfg - Notepad
File Edit Format View Help

Name	Date modified	Type	Size
configuration.xsl	17.03.2021 12:45	XSL Stylesheet	1 KB
log4j.properties	17.03.2021 12:45	PROPERTIES File	4 KB
zoo.cfg	22.11.2021 15:41	CFG File	2 KB

```
# The number of milliseconds of each tick
tickTime=2000
# The number of ticks that the initial
# synchronization phase can take
initLimit=10
# The number of ticks that can pass betwe
# sending a request and getting an acknow
syncLimit=5
# the directory where the snapshot is sto
# do not use /tmp for storage, /tmp here
# example sakes.
dataDir=C:\zookeeper\data
# the port at which the clients will conn
clientPort=2181
# the maximum number of client connection
# increase this if you need to handle mor
#maxClientCnxns=60
#
```

- Download the appropriate version from the Zookeeper site.
- The downloaded zip file is extracted to the C: directory.
- Update the zoo_sample.cfg’s file name in the config folder and named zoo.cfg.
- Update the zoo.cfg
- Create folders named data.

DOWNLOAD ZOOKEEPER



Download Java for Windows

Recommended Version 8 Update 311 (filesize: 2.01 MB)

Release date October 19, 2021

Important Oracle Java License Update

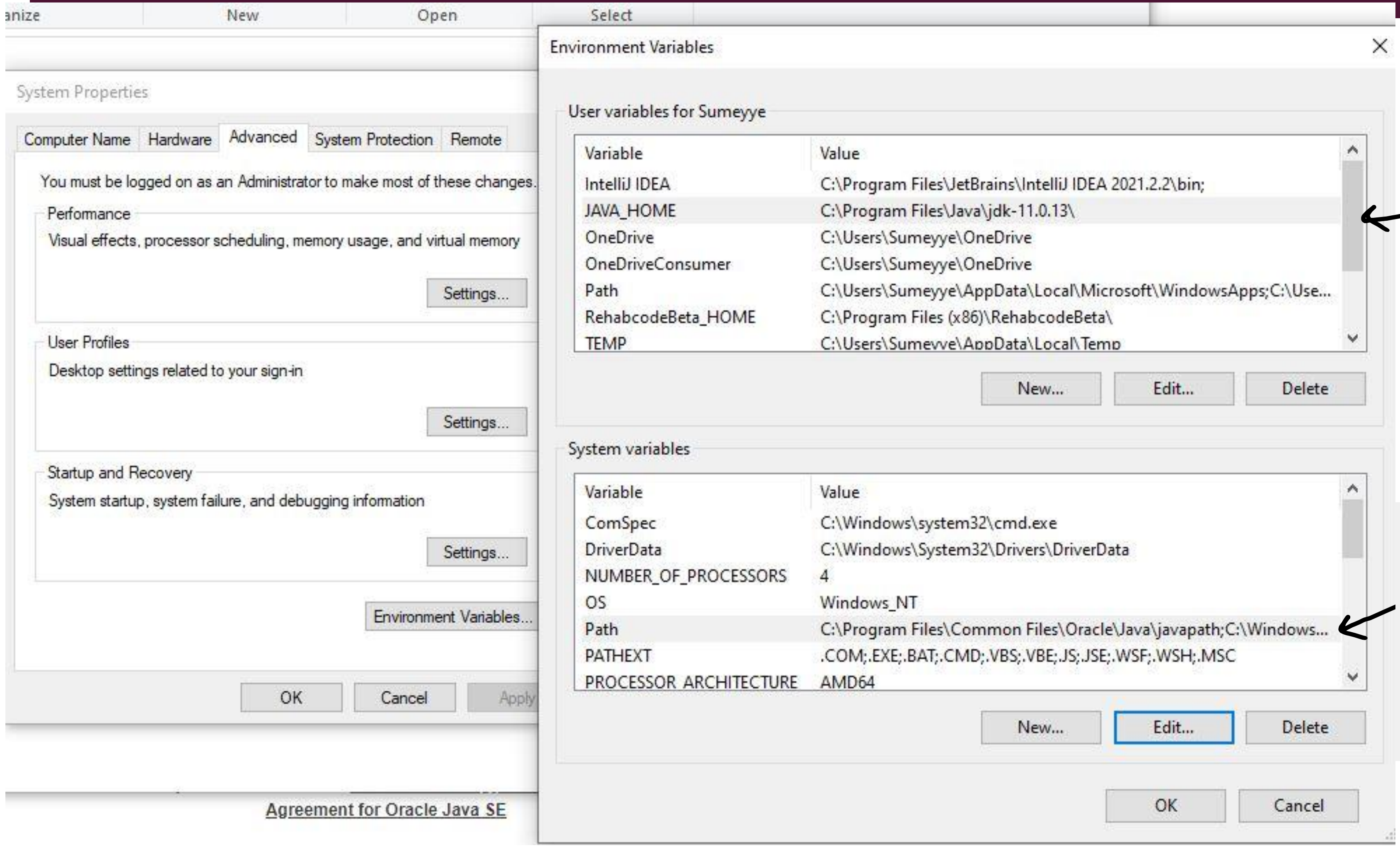
The Oracle Java License has changed for releases starting April 16, 2019.

The new Oracle Technology Network License Agreement for Oracle Java SE is substantially different from prior Oracle Java licenses. The new license permits certain uses, such as personal use and development use, at no cost -- but other uses authorized under prior Oracle Java licenses may no longer be available. Please review the terms carefully before downloading and using this product. An FAQ is available here.

Commercial license and support is available with a low cost Java SE Subscription.

Oracle also provides the latest OpenJDK release under the open source GPL License at jdk.java.net.

JAVA_HOME IS NOT SET



anize New Open Select

System Properties

Computer Name | Hardware | Advanced | System Protection | Remote

You must be logged on as an Administrator to make most of these changes.

Performance
Visual effects, processor scheduling, memory usage, and virtual memory
Settings...

User Profiles
Desktop settings related to your sign-in
Settings...

Startup and Recovery
System startup, system failure, and debugging information
Settings...

Environment Variables...

OK Cancel Apply

Agreement for Oracle Java SE

Environment Variables

User variables for Sumeyye

Variable	Value
IntelliJ IDEA	C:\Program Files\JetBrains\IntelliJ IDEA 2021.2.2\bin;
JAVA_HOME	C:\Program Files\Java\jdk-11.0.13\
OneDrive	C:\Users\Sumeyye\OneDrive
OneDriveConsumer	C:\Users\Sumeyye\OneDrive
Path	C:\Users\Sumeyye\AppData\Local\Microsoft\WindowsApps;C:\Use...
RehabcodeBeta_HOME	C:\Program Files (x86)\RehabcodeBeta\
TEMP	C:\Users\Sumeyye\AppData\Local\Temp

New... Edit... Delete

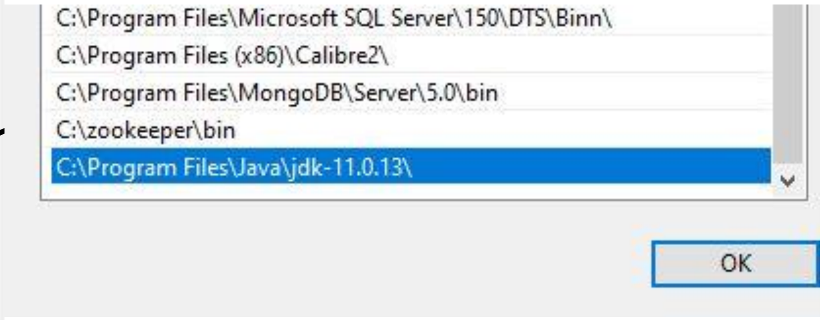
System variables

Variable	Value
ComSpec	C:\Windows\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS	4
OS	Windows_NT
Path	C:\Program Files\Common Files\Oracle\Java\javapath;C:\Windows...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE	AMD64

New... Edit... Delete

OK Cancel

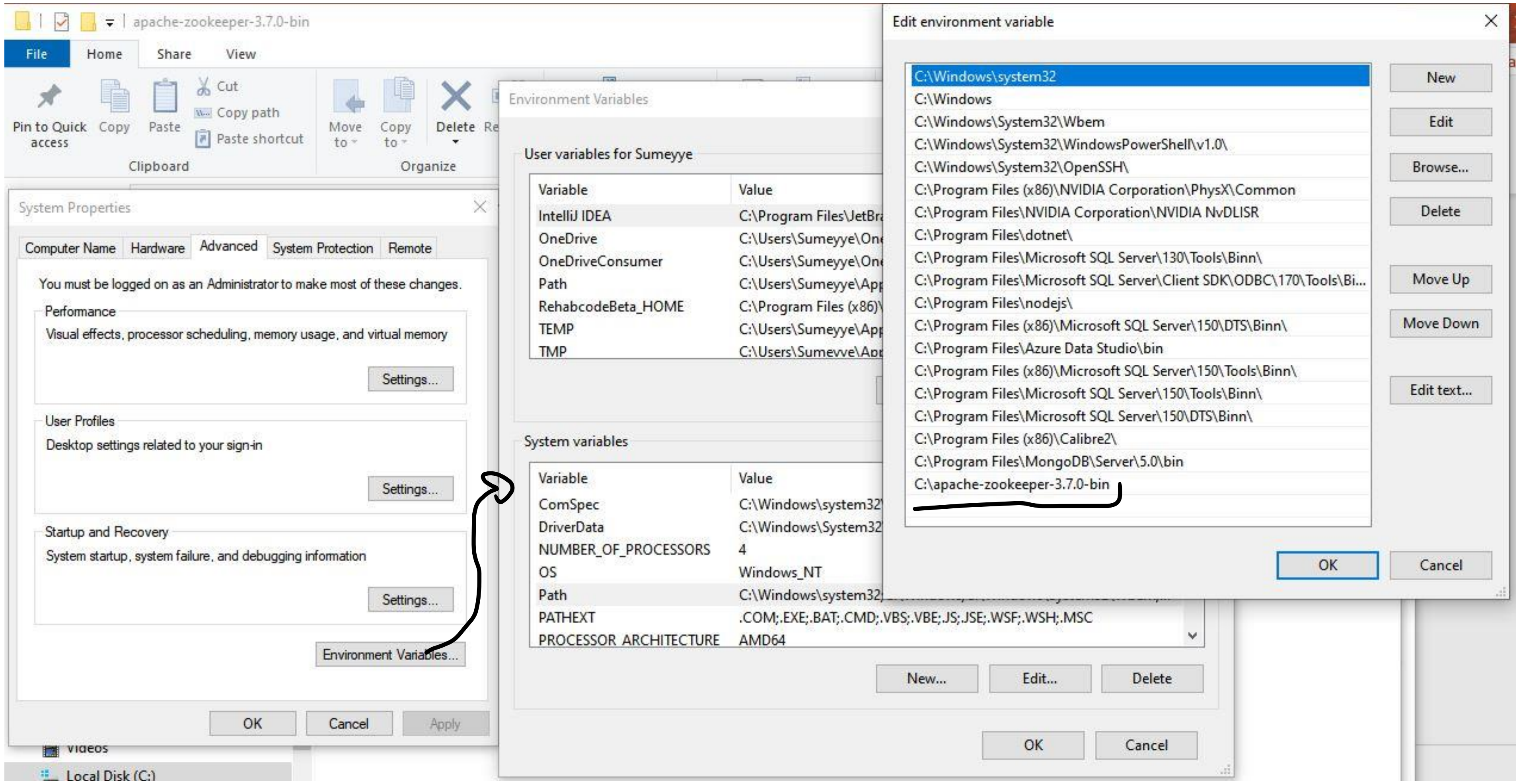
create new env. variable



C:\Program Files\Microsoft SQL Server\150\DTS\Binn\
C:\Program Files (x86)\Calibre2\
C:\Program Files\MongoDB\Server\5.0\bin
C:\zookeeper\bin
C:\Program Files\Java\jdk-11.0.13\

OK

DOWNLOAD ZOOKEEPER



You must be logged on as an Administrator to make most of these changes.

Visual effects, processor scheduling, memory usage, and virtual memory

Desktop settings related to your sign-in

System startup, system failure, and debugging information

User variables for Sumeyye

Variable	Value
IntelliJ IDEA	C:\Program Files\JetBra
OneDrive	C:\Users\Sumeyye\One
OneDriveConsumer	C:\Users\Sumeyye\One
Path	C:\Users\Sumeyye\App
RehabcodeBeta_HOME	C:\Program Files (x86)\
TEMP	C:\Users\Sumeyye\App
TMP	C:\Users\Sumeyye\App

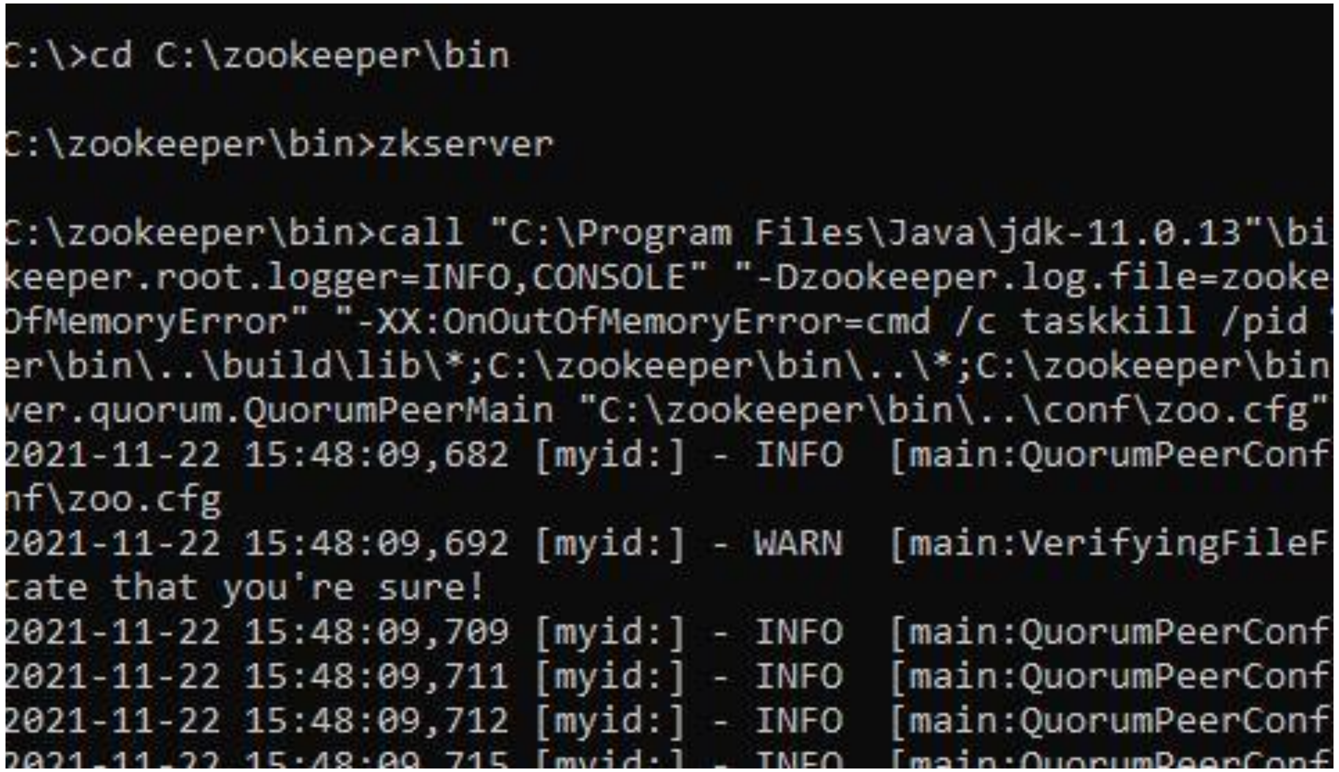
System variables

Variable	Value
ComSpec	C:\Windows\system32\
DriverData	C:\Windows\System32
NUMBER_OF_PROCESSORS	4
OS	Windows_NT
Path	C:\Windows\system32;
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE	AMD64

Edit environment variable

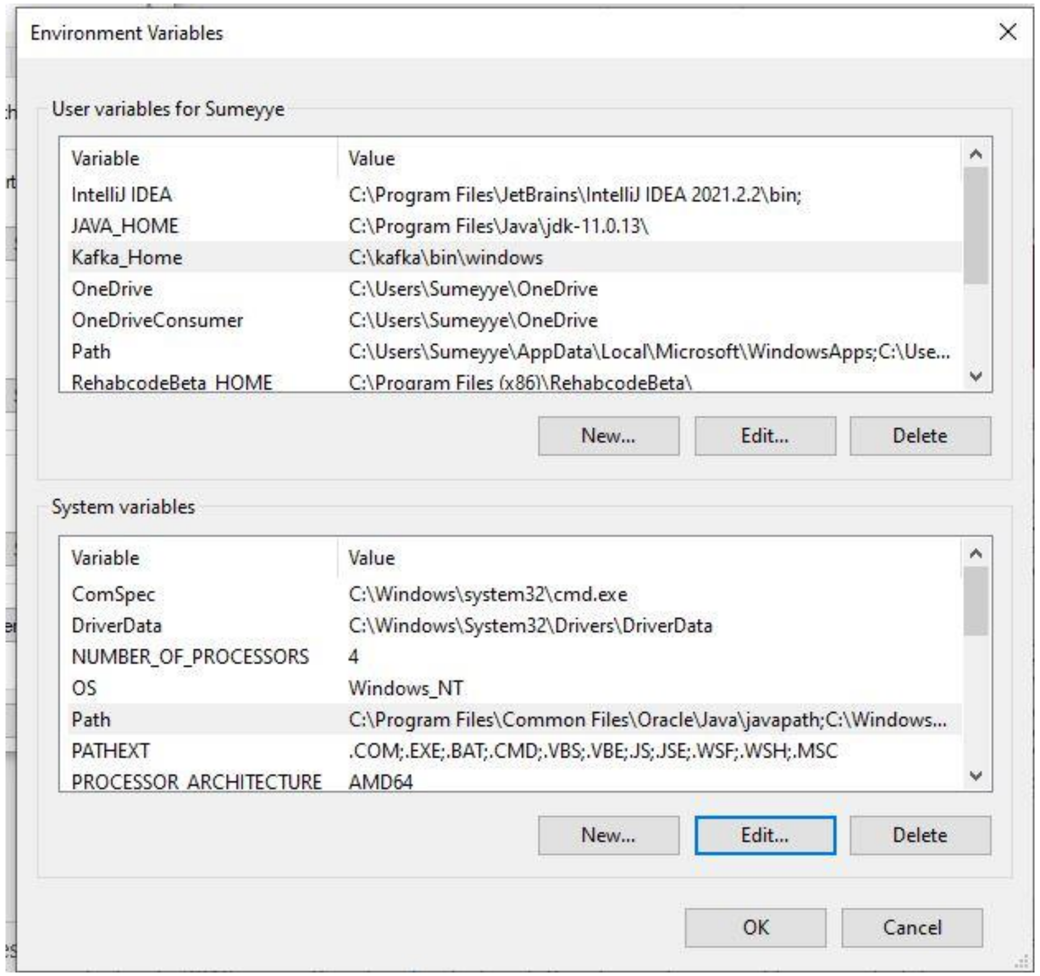
```
C:\Windows\system32
C:\Windows
C:\Windows\System32\Wbem
C:\Windows\System32\WindowsPowerShell\v1.0\
C:\Windows\System32\OpenSSH\
C:\Program Files (x86)\NVIDIA Corporation\PhysX\Common
C:\Program Files\NVIDIA Corporation\NVIDIA NvDLISR
C:\Program Files\dotnet\
C:\Program Files\Microsoft SQL Server\130\Tools\Binn\
C:\Program Files\Microsoft SQL Server\Client SDK\ODBC\170\Tools\Bi...
C:\Program Files\nodejs\
C:\Program Files (x86)\Microsoft SQL Server\150\DTS\Binn\
C:\Program Files\Azure Data Studio\bin
C:\Program Files (x86)\Microsoft SQL Server\150\Tools\Binn\
C:\Program Files\Microsoft SQL Server\150\Tools\Binn\
C:\Program Files\Microsoft SQL Server\150\DTS\Binn\
C:\Program Files (x86)\Calibre2\
C:\Program Files\MongoDB\Server\5.0\bin
C:\apache-zookeeper-3.7.0-bin
```

RUNNING ZOOKEEPER



```
C:\>cd C:\zookeeper\bin

C:\zookeeper\bin>zkserver

C:\zookeeper\bin>call "C:\Program Files\Java\jdk-11.0.13"\bi
keeper.root.logger=INFO,CONSOLE" "-Dzookeeper.log.file=zooke
OfMemoryError" "-XX:OnOutOfMemoryError=cmd /c taskkill /pid
er\bin\..\build\lib\*;C:\zookeeper\bin\..\*;C:\zookeeper\bin
ver.quorum.QuorumPeerMain "C:\zookeeper\bin\..\conf\zoo.cfg"
2021-11-22 15:48:09,682 [myid:] - INFO  [main:QuorumPeerConf
nf\zoo.cfg
2021-11-22 15:48:09,692 [myid:] - WARN  [main:VerifyingFileF
cate that you're sure!
2021-11-22 15:48:09,709 [myid:] - INFO  [main:QuorumPeerConf
2021-11-22 15:48:09,711 [myid:] - INFO  [main:QuorumPeerConf
2021-11-22 15:48:09,712 [myid:] - INFO  [main:QuorumPeerConf
2021-11-22 15:48:09,715 [myid:] - INFO  [main:QuorumPeerConf
```

INSTALLATION OF KAFKA



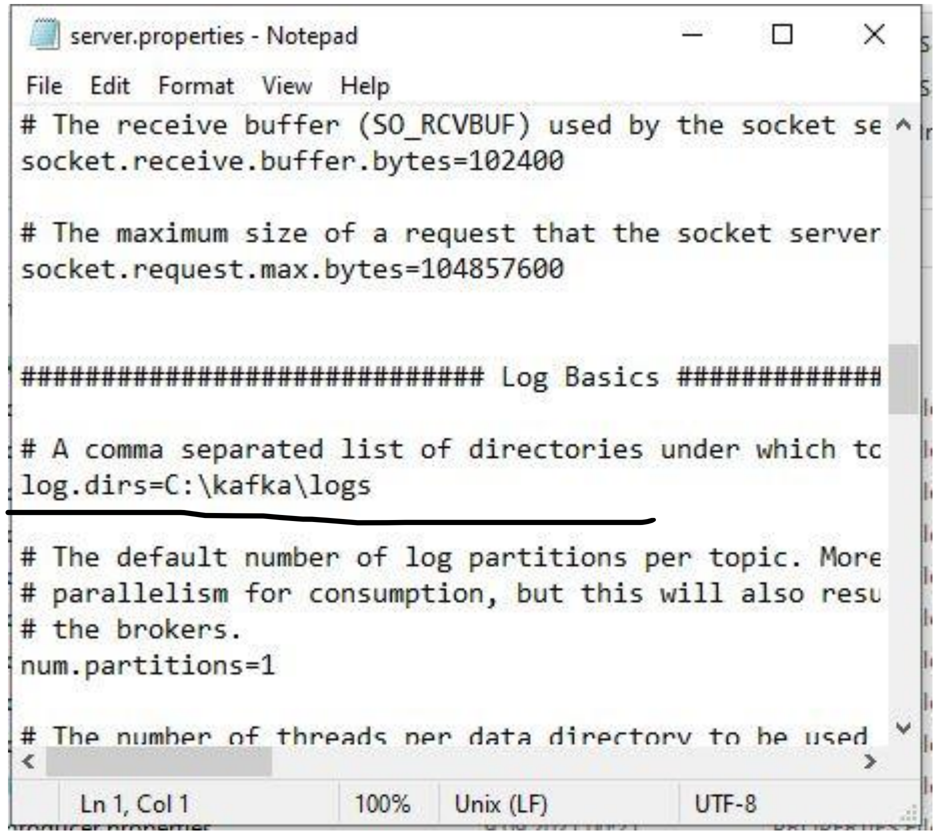
Environment Variables

User variables for Sumeyye

Variable	Value
IntelliJ IDEA	C:\Program Files\JetBrains\IntelliJ IDEA 2021.2.2\bin;
JAVA_HOME	C:\Program Files\Java\jdk-11.0.13\
Kafka_Home	C:\kafka\bin\windows
OneDrive	C:\Users\Sumeyye\OneDrive
OneDriveConsumer	C:\Users\Sumeyye\OneDrive
Path	C:\Users\Sumeyye\AppData\Local\Microsoft\WindowsApps;C:\Use...
RehabcodeBeta_HOME	C:\Program Files (x86)\RehabcodeBeta\

System variables

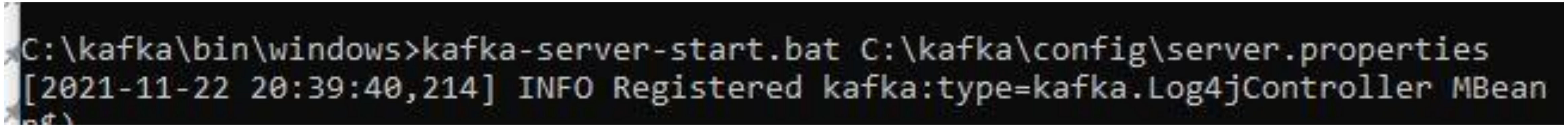
Variable	Value
ComSpec	C:\Windows\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS	4
OS	Windows_NT
Path	C:\Program Files\Common Files\Oracle\Java\javapath;C:\Windows...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE	AMD64



server.properties - Notepad

```
# The receive buffer (SO_RCVBUF) used by the socket se
socket.receive.buffer.bytes=102400

# The maximum size of a request that the socket server
socket.request.max.bytes=104857600


############################# Log Basics ############

# A comma separated list of directories under which tc
log.dirs=C:\kafka\logs

# The default number of log partitions per topic. More
# parallelism for consumption, but this will also resu
# the brokers.
num.partitions=1

# The number of threads per data directory to be used
```

RUNNING OF KAFKA



```
C:\kafka\bin\windows>kafka-server-start.bat C:\kafka\config\server.properties
[2021-11-22 20:39:40,214] INFO Registered kafka:type=kafka.Log4jController MBean
```

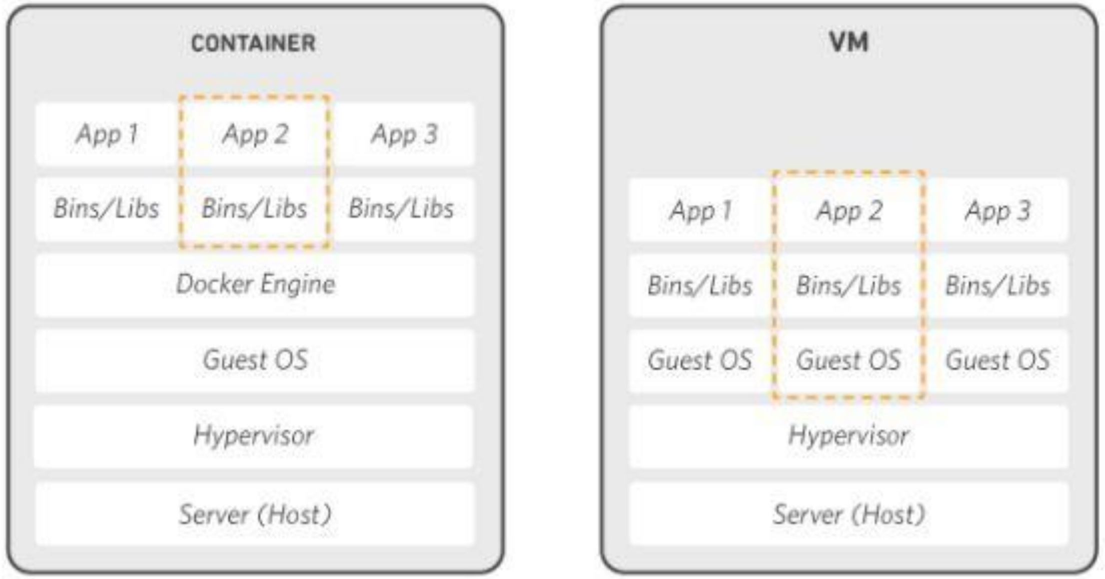
DOCKER

- Docker is a technology that provides virtualization thanks to hundreds or even thousands of isolated and independent containers on the same operating system.
- It is an open source 'container' technology.
- Docker is a new technology that allows development teams to build, manage and secure applications anywhere.

CONTAINER

- It is the name given to each of the processes that are run in isolation from each other in the Linux kernel by the Docker Daemon.
- A container is a special type of transaction that is isolated from other processes. Containers are assigned resources that no other process can access.

CONTAINER-VIRTUAL MACHINE



VM (Virtual Machine)
•OS : **Tam işletim sistemi**
•İzolasyon : **Yüksek**
•Çalışır hale gelmesi : **Dakikalar**
•Versiyonlama : **Yok**
•Kolay paylaşılabilirlik : **Düşük**

Docker
•OS : **Küçültülmüş işletim sistemi imajı**
•İzolasyon : **Daha düşük**
•Çalışır hale gelmesi : **Saniyeler**
•Versiyonlama : **Yüksek**
•Kolay paylaşılabilirlik : **Yüksek**

HYPERVISOR

- Hypervisor is a piece of code that allows multiple operating systems to run on the same hardware.
- This piece of code works directly on the hardware, allowing us to create multiple guest operating systems on our physical server.

DOCKER IMAGE-DOCKER REGISTERY

- It is a structure that contains your application to run and the necessary operating system libraries running on your application's infrastructure.
- Images are kept in Docker Registries.

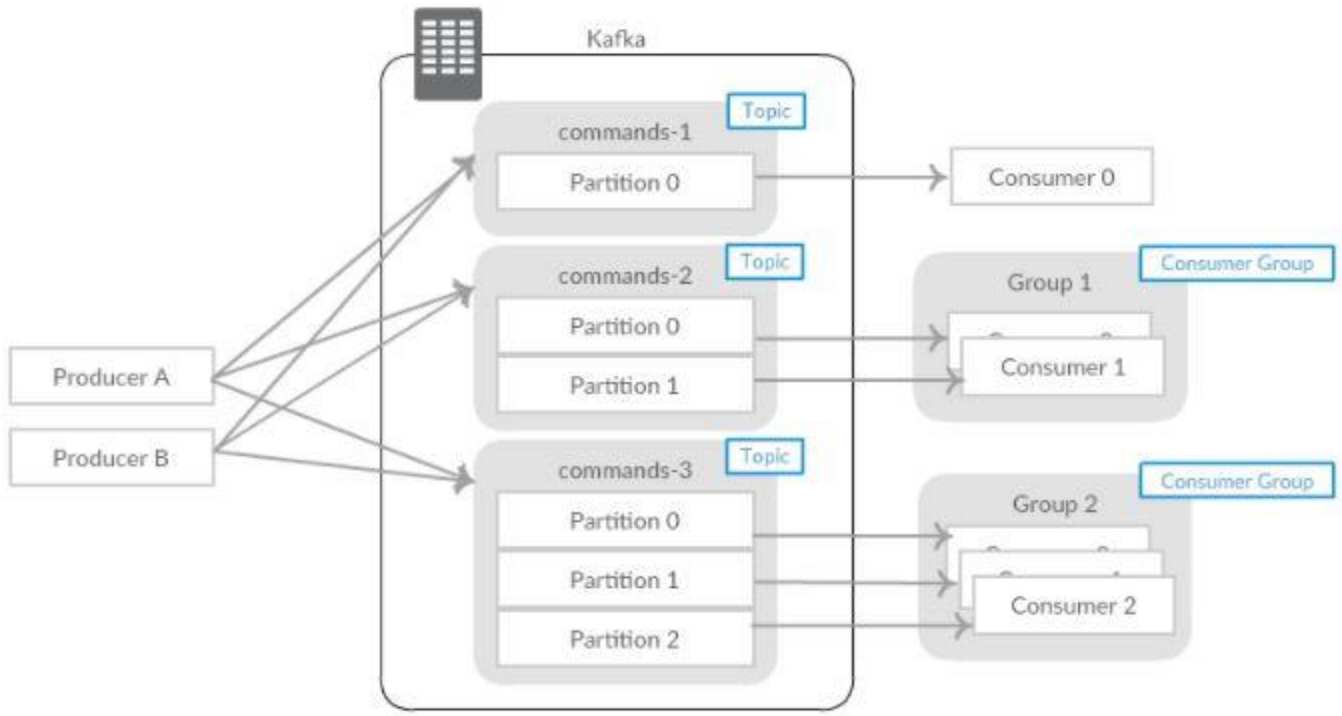
DOCKER COMPOSE

- With Docker Compose, you can define more than one container in a single file, you can run the application by standing up all the requirements your application needs with a single command.

KAFKA WITH DOCKER

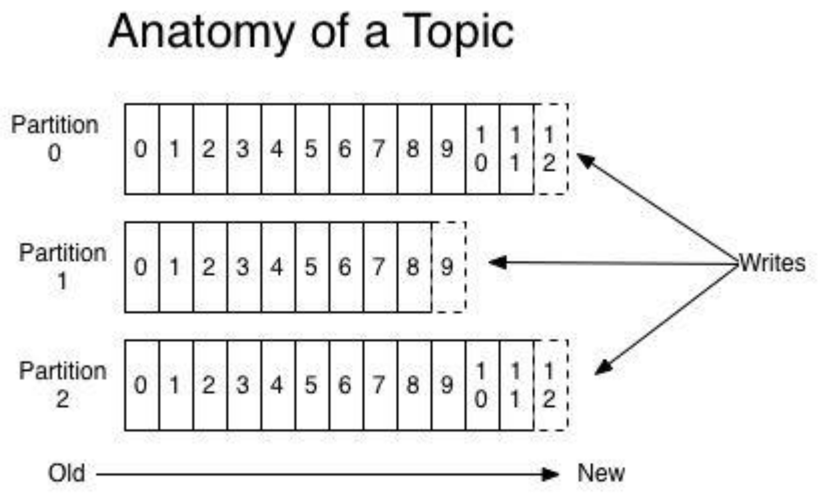
- Apache Kafka is a distributed publish-subscribe messaging system that is designed to be fast, scalable, and durable.
- Kafka stores streams of records (messages) in topics. Each record consists of a key, a value, and a timestamp.
- Producers write data to topics and consumers read from topics.

KAFKA WITH DOCKER



TOPICS AND LOGS (KAFKA)

- A topic is a category to which records are published. It can have zero, one, or many consumers that subscribe to the data written to it.
- For each topic, the Kafka cluster maintains a partitioned log. Since Kafka is a distributed system, topics are partitioned and replicated across multiple nodes.



PRODUCERS AND CONSUMERS

- Producers publish data to the topics of their choice. It is responsible for choosing which record to assign to which partition within the topic.
- Consumer groups can subscribe to one or more topics. Each one of these groups can be configured with multiple consumers.
- Every message in a topic is delivered to one of the consumer instances inside the group subscribed to that topic. All messages with the same key arrive at the same consumer.

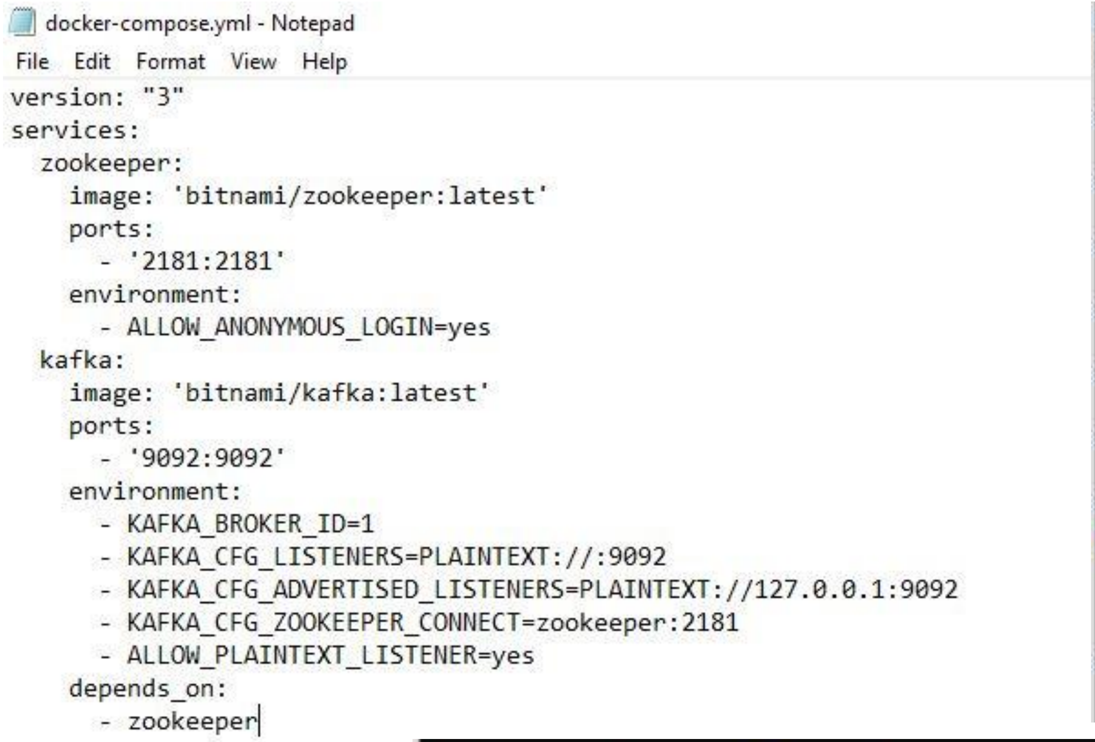
MESSAGES

- Kafka is message-based, the main element that is processed is messages.
- Messages consist of a simple key-value pair. Both key and message contents can be anything that can be serialized.
- Examples of messages are page clicks, comments, and orders.

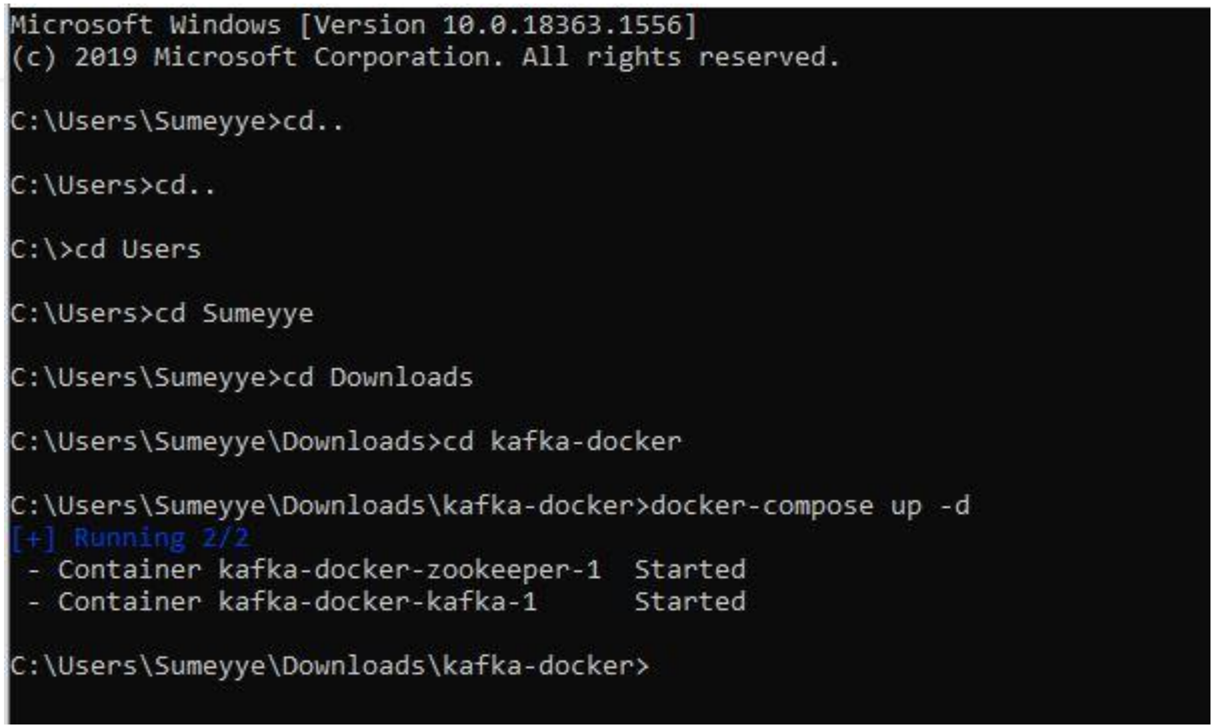
PARTITION

- Kafka runs as multiple nodes on different physical/virtual machines for load distribution and fault tolerance.
- Topics are also physically divided into different partitions on these machines.
- Each partition can also be copied to other nodes as master/slave.

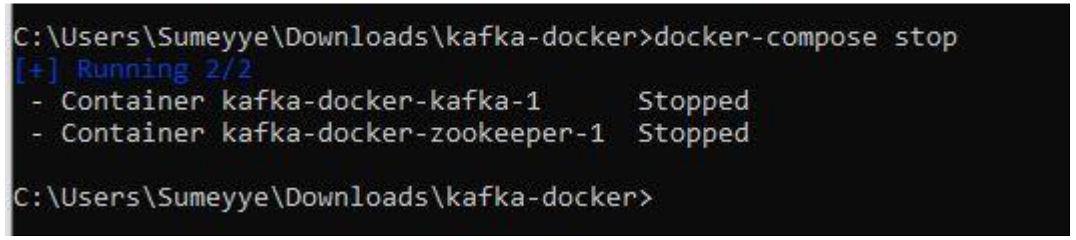
DOCKER-COMPOSE.YML- RUNNING DOCKER-COMPOSE



```
version: "3"
services:
  zookeeper:
    image: 'bitnami/zookeeper:latest'
    ports:
      - '2181:2181'
    environment:
      - ALLOW_ANONYMOUS_LOGIN=yes
  kafka:
    image: 'bitnami/kafka:latest'
    ports:
      - '9092:9092'
    environment:
      - KAFKA_BROKER_ID=1
      - KAFKA_CFG_LISTENERS=PLAINTEXT://:9092
      - KAFKA_CFG_ADVERTISED_LISTENERS=PLAINTEXT://127.0.0.1:9092
      - KAFKA_CFG_ZOOKEEPER_CONNECT=zookeeper:2181
      - ALLOW_PLAINTEXT_LISTENER=yes
    depends_on:
      - zookeeper
```

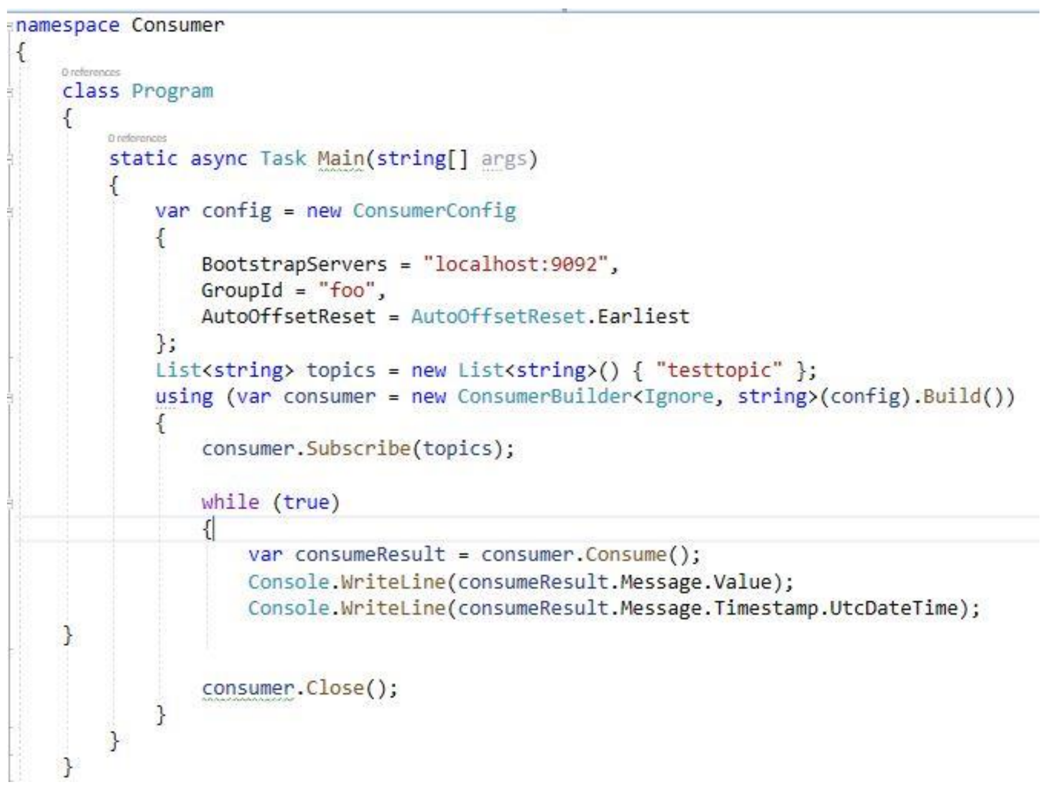


```
Microsoft Windows [Version 10.0.18363.1556]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\Sumeyye>cd..

C:\Users>cd..

C:\>cd Users

C:\Users>cd Sumeyye

C:\Users\Sumeyye>cd Downloads

C:\Users\Sumeyye\Downloads>cd kafka-docker

C:\Users\Sumeyye\Downloads\kafka-docker>docker-compose up -d
[+] Running 2/2
 - Container kafka-docker-zookeeper-1  Started
 - Container kafka-docker-kafka-1      Started

C:\Users\Sumeyye\Downloads\kafka-docker>
```

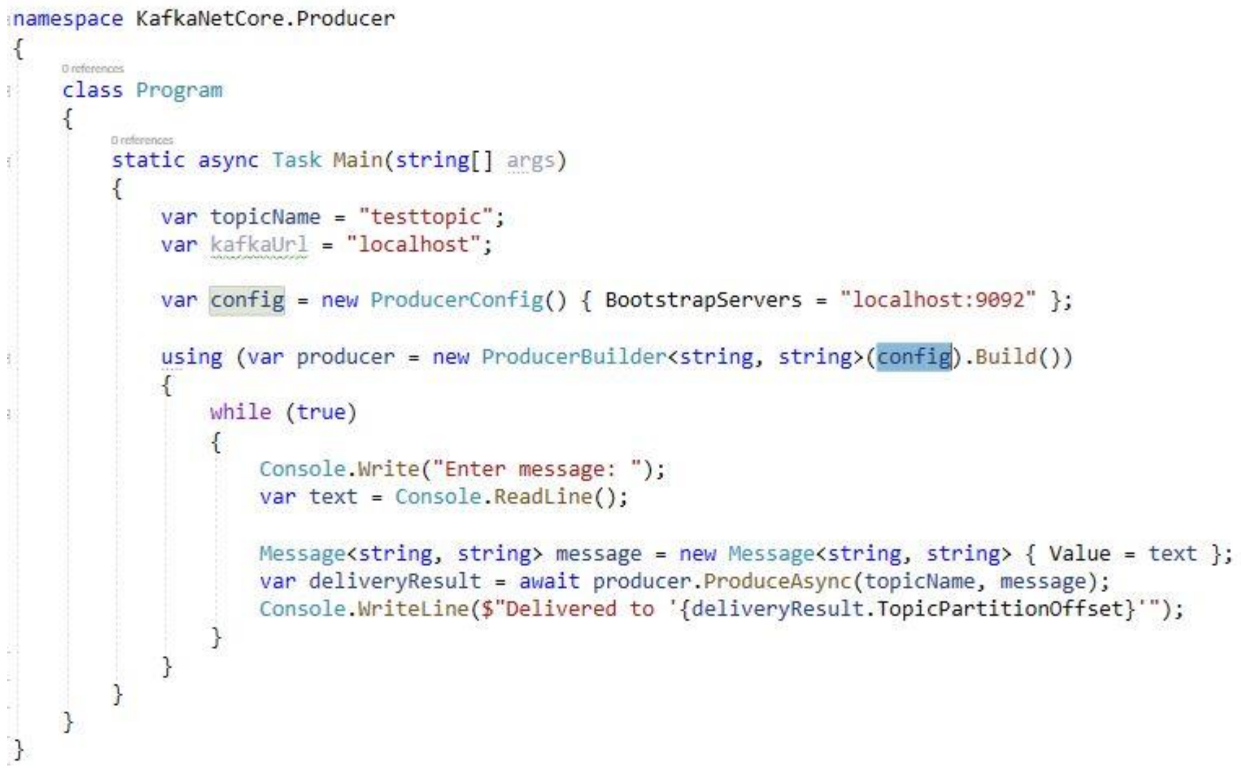


```
C:\Users\Sumeyye\Downloads\kafka-docker>docker-compose stop
[+] Running 2/2
 - Container kafka-docker-kafka-1      Stopped
 - Container kafka-docker-zookeeper-1  Stopped

C:\Users\Sumeyye\Downloads\kafka-docker>
```

VISUAL STUDIO



```
namespace Consumer
{
    0 references
    class Program
    {
        0 references
        static async Task Main(string[] args)
        {
            var config = new ConsumerConfig
            {
                BootstrapServers = "localhost:9092",
                GroupId = "foo",
                AutoOffsetReset = AutoOffsetReset.Earliest
            };
            List<string> topics = new List<string>() { "testtopic" };
            using (var consumer = new ConsumerBuilder<Ignore, string>(config).Build())
            {
                consumer.Subscribe(topics);

                while (true)
                {
                    var consumeResult = consumer.Consume();
                    Console.WriteLine(consumeResult.Message.Value);
                    Console.WriteLine(consumeResult.Message.Timestamp.UtcDateTime);
    }

                consumer.Close();
            }
        }
    }
```



```
namespace KafkaNetCore.Producer
{
    0 references
    class Program
    {
        0 references
        static async Task Main(string[] args)
        {
            var topicName = "testtopic";
            var kafkaUrl = "localhost";

            var config = new ProducerConfig() { BootstrapServers = "localhost:9092" };

            using (var producer = new ProducerBuilder<string, string>(config).Build())
            {
                while (true)
                {
                    Console.Write("Enter message: ");
                    var text = Console.ReadLine();

                    Message<string, string> message = new Message<string, string> { Value = text };
                    var deliveryResult = await producer.ProduceAsync(topicName, message);
                    Console.WriteLine($"Delivered to '{deliveryResult.TopicPartitionOffset}'");
                }
            }
        }
    }
}
```