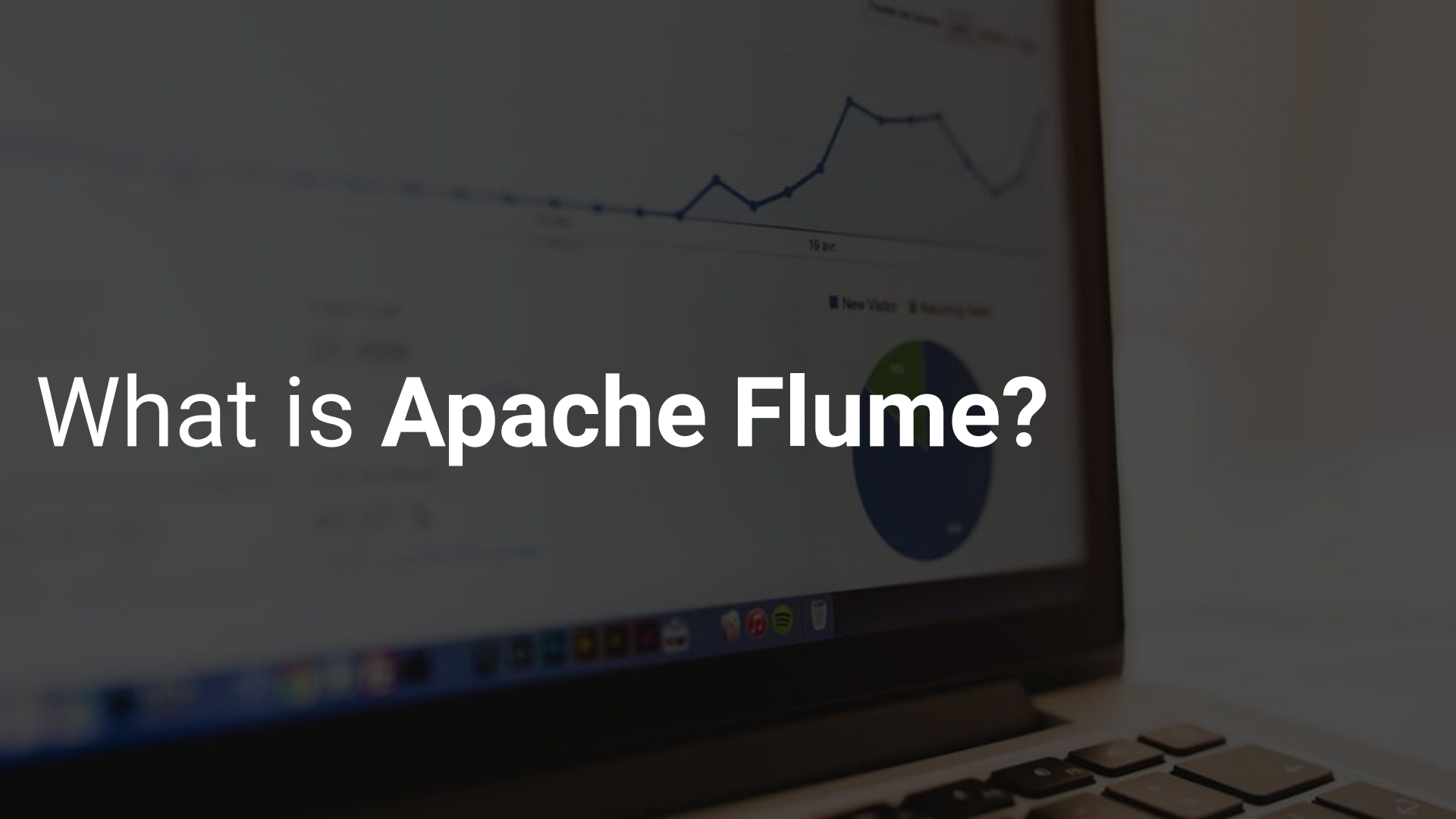
Apache Flume



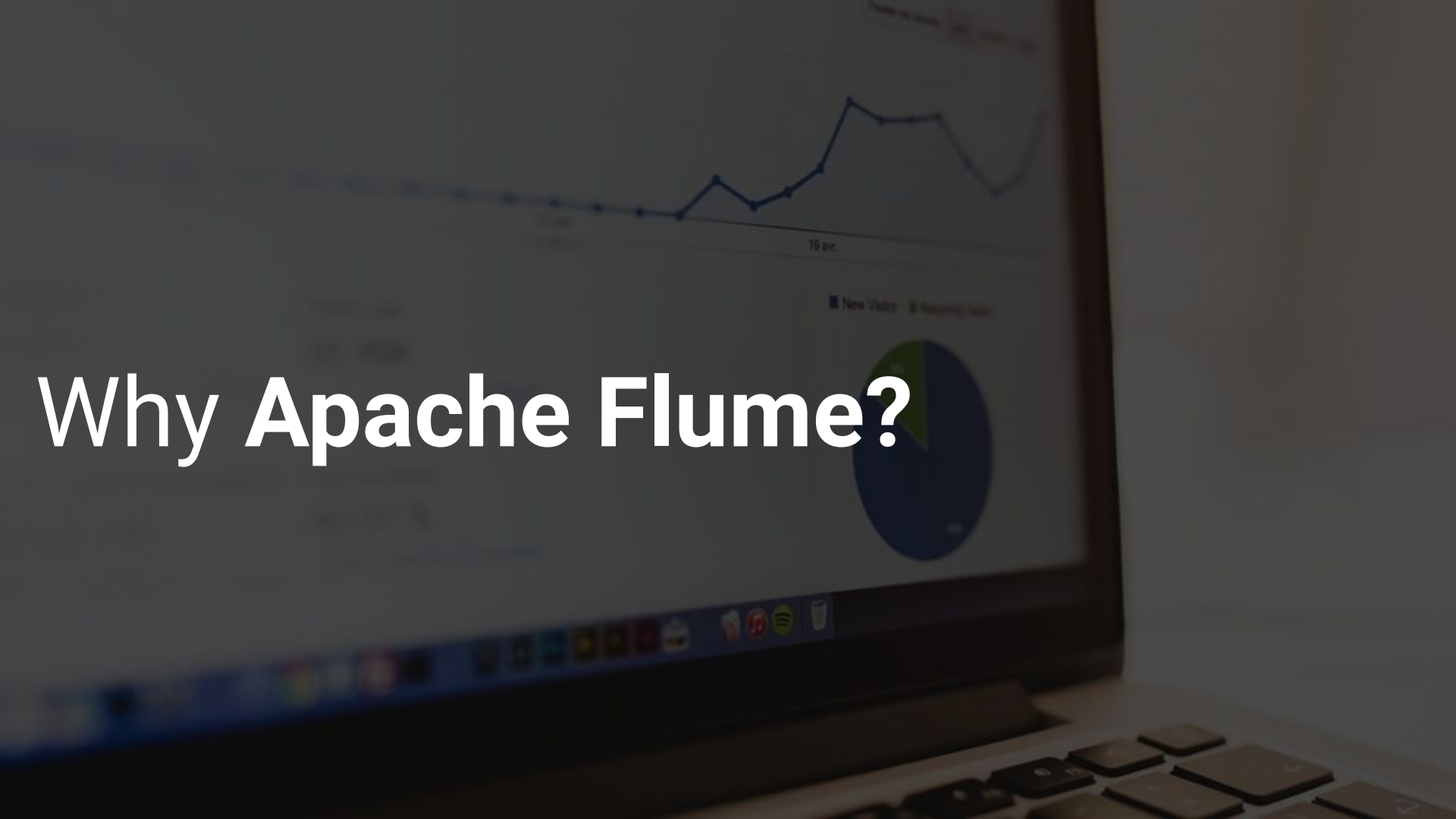
What is **Apache Flume?**

Apache Flume

Flume is an **open-source, distributed, and reliable** system designed for collecting, aggregating, and transferring huge volumes of log data from various different sources to the centralized repository. The centralized repository can be **HDFS**, **HBase**, etc.

If we want to transfer social-media-generated data, email messages, log data to **Hadoop**, then we use **Apache Flume**. It is the top-level project at Apache Software Foundation.

The main purpose of designing **Apache Flume** is to copy the streaming data (log data) from different web servers to **HDFS**.



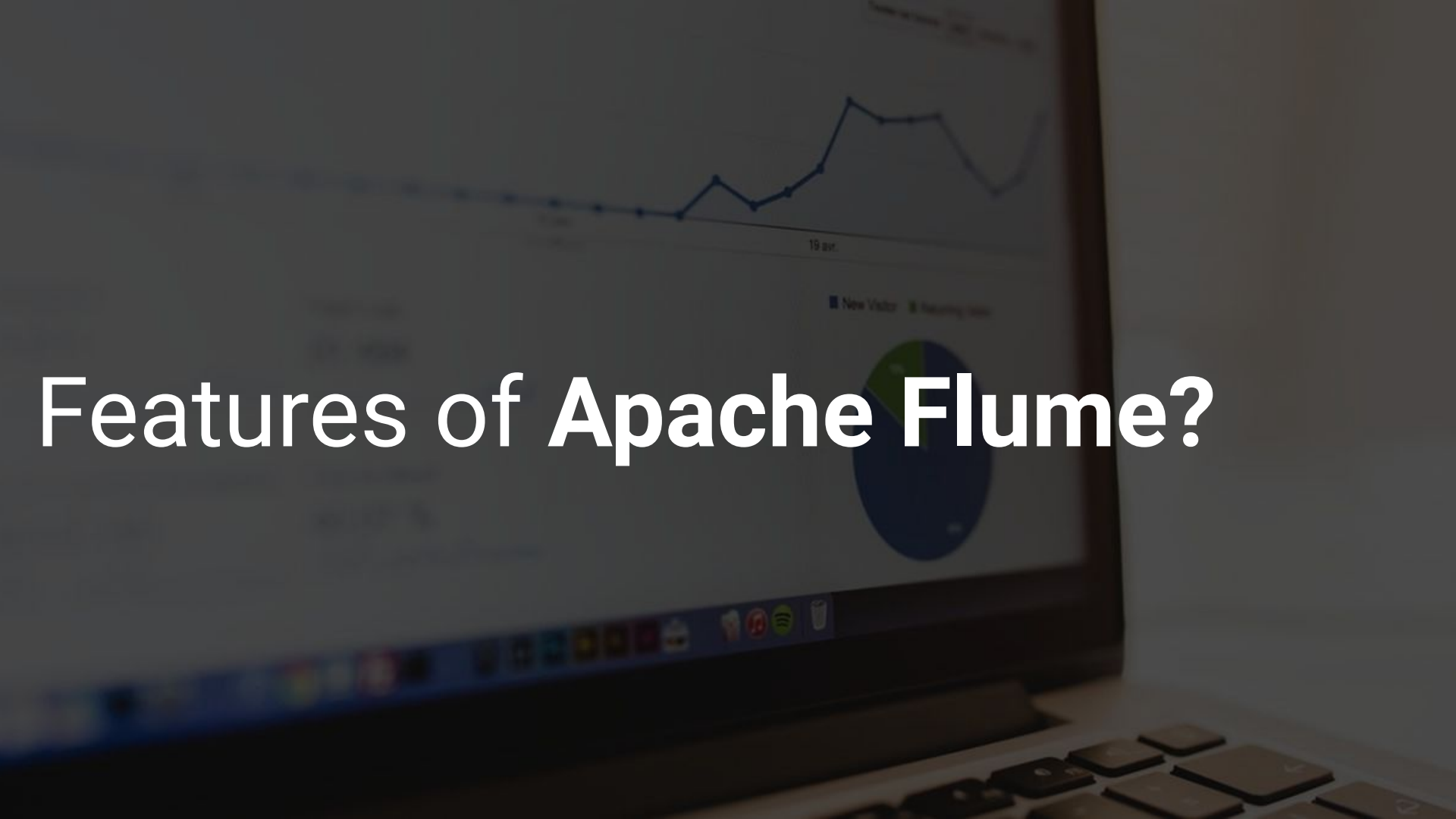
Why **Apache Flume?**

Why using Apache Flume?

Since we all know that millions of services of a company are running on multiple servers. These servers produce lots of logs. With the advent of Big Data technology, that is, Apache Hadoop, companies want to analyze these logs to generate insights. Businesses want to analyze these log data to understand their customer behavior.

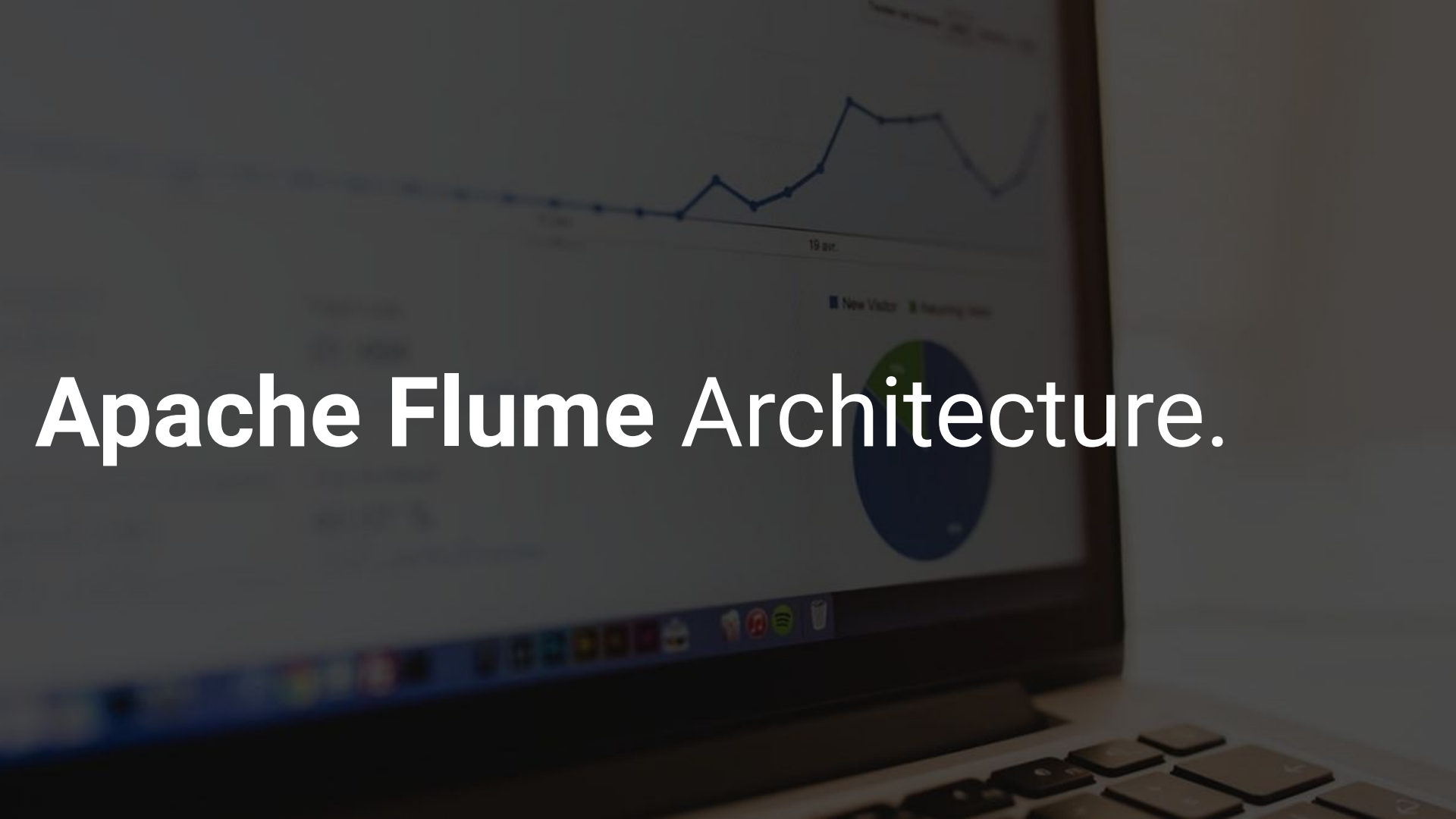
So for processing logs, they require a scalable and reliable distributed data collection service. This service must be capable of transferring logs from their web servers to the system, which can store and process these logs (such as HDFS).

Here Apache Flume came into the picture. It is an open-source distributed data collection service that we can use for data transferring from source to destination.

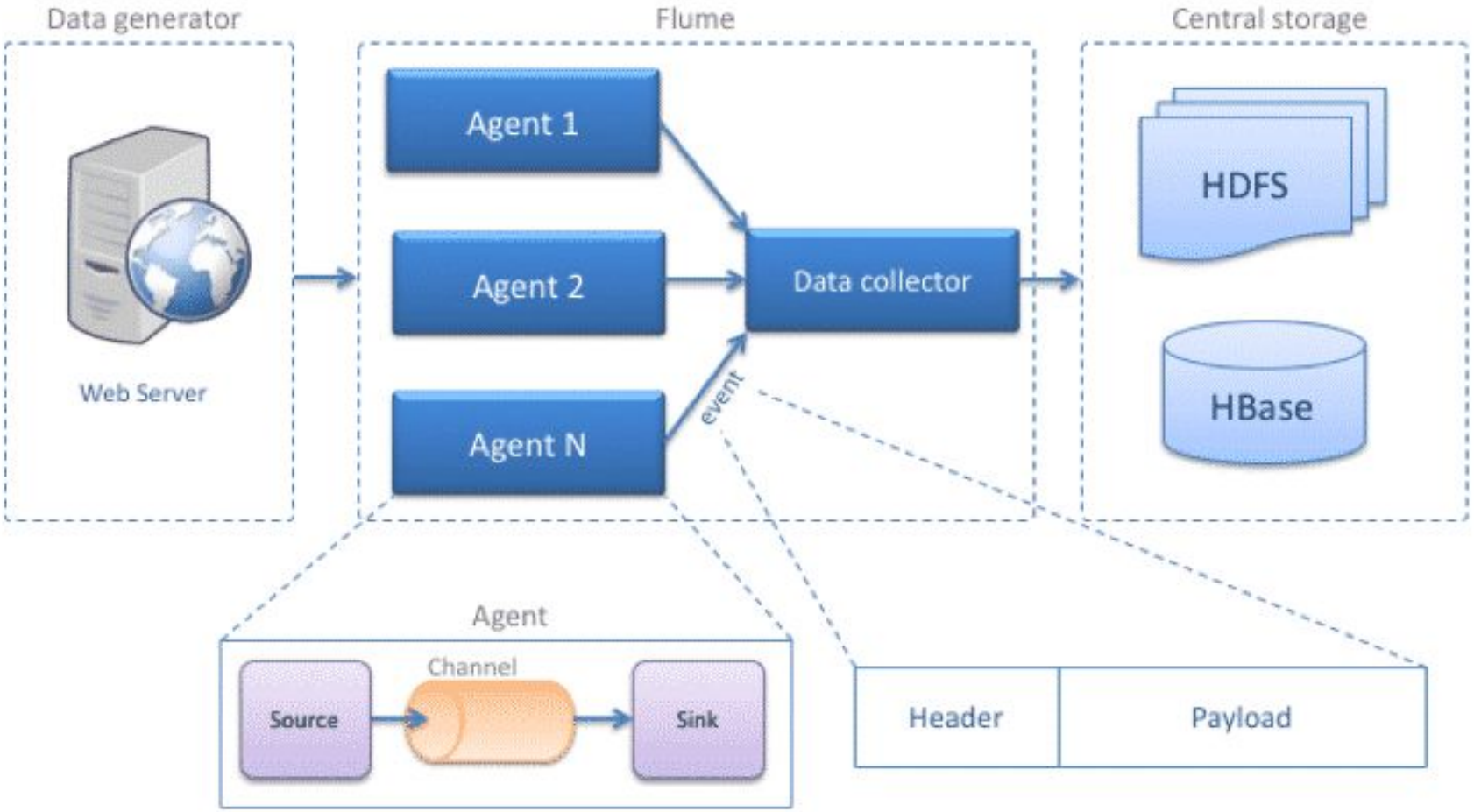


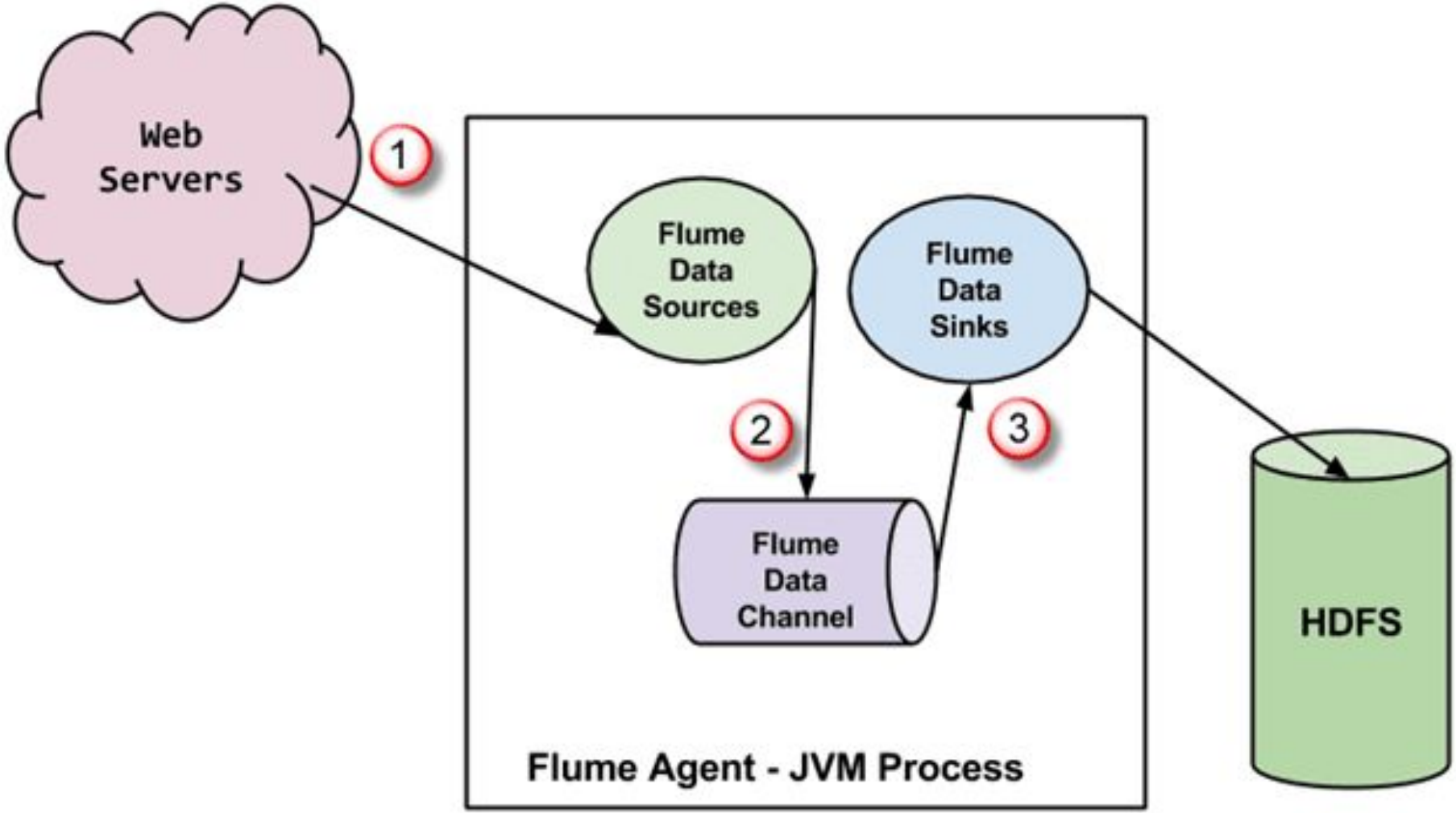
Features of **Apache Flume?**

1. It is an open-source framework.
2. It is a highly available, robust, and fault-tolerant service.
3. Apache Flume has tunable reliability mechanisms for failover and recovery.
4. It provides support for the complex data flows such as fan-in flows, multi-hop flows, fan-out flows. It also provides support for Contextual routing as well as backup routes.
5. Flume is horizontally scalable.
6. Flume supports large sets of channels, sources, and sinks.
7. We can use Apache Flume for efficiently ingesting log data from other servers into the centralized data store.
8. Apache Flume allows us to collect data from web servers in real-time as well as in the batch mode.
9. We can easily move social networking sites generated data and e-commerce site data into the Hadoop Distributed File System through Flume.
10. Flume offers steady data flow between the read and the write operations.
11. It offers a high throughput and lower latency.
12. Flume is an in-expensive distributed system.



Apache Flume Architecture.


Data generator
Flume
Central storage
Agent 1
Agent 2
Agent N
Data collector
HDFS
HBase
Web Server
event
Agent
Channel
Source
Sink
Header
Payload


Web Servers
1
Flume Data Sources
Flume Data Sinks
2
3
Flume Data Channel
HDFS
Flume Agent - JVM Process

Flume Event:

It is a unit of data that is to be transferred from source to destination.

Flume Agent:

It is an independent JVM process (JVM) in Flume. Flume agents receive events from the clients or the other Flume agents. It passes these events to another flume agent or to the centralized store.

Flume Agent basically contains three main components. Let’s explore each of them in detail.

Flume Source:

It is a component of a flume agent that receives the data from data generators. Source transfers the data received from data generators to one or more flume channels in the form of events.

There are several types of sources supported by Flume.

Ex: Thrift source, Exec source, Avro source, twitter 1% source, etc.

Flume Channel:

It is a component of a flume agent that receives events from source and buffers them until the flume sinks consume them.

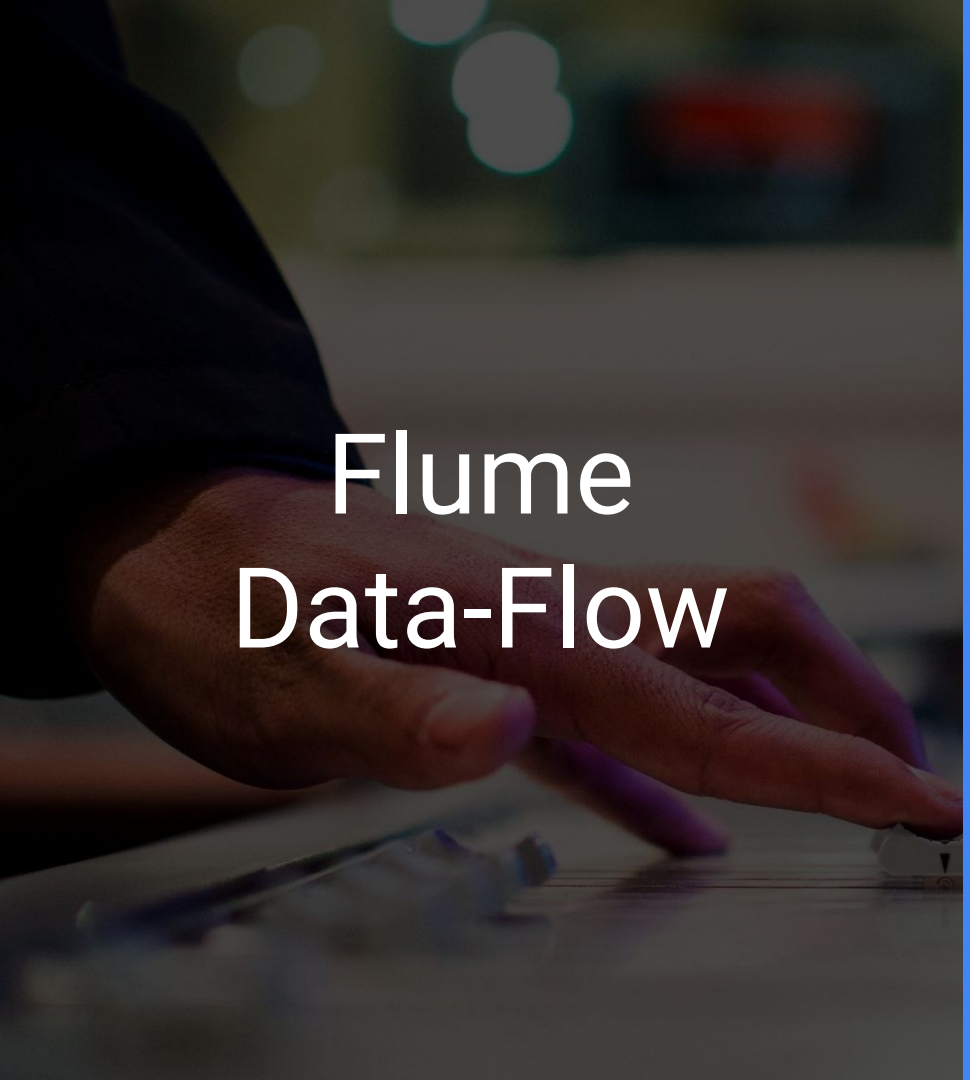
There are several types of channels supported by Flume.

Ex: JDBC channel, Memory channel, File system channel, etc.

Flume Sink:

It is a component of a flume agent that consumes the data from the flume channel and stores them into the next destination, which can be a centralized store or the other flume agents.

Ex: HDFS sink.



Flume Data-Flow

Flume provides support for the complex data flow. The three types of data flow in Flume are:

1. Multi-hop Flow

In multi-hop flow, before reaching the final destination, the event goes through two or more flume agents.

2. Fan-out Flow

In fan-out flow, an event will flow from one source to multiple channels. It is of two types – replicating and multiplexing.

3. Fan-in Flow

In fan-in flow, an event is transferred from many sources to one channel.

Flume Advantages

Flume permits us to store streaming data into centralized repositories (HBase, HDFS).

It offers steady data flow during read/write between producer and consumer.

Flume supports contextual routing.

It guarantees reliable message delivery.

Flume is open source, reliable, fault-tolerant, scalable, extensible, customizable, and manageable.

Flume Disadvantages

It provides weaker ordering guarantees.

Flume doesn’t guarantee about the uniqueness of the messages, that is, the messages reaching are 100% unique.

Flume has a complex topology. Reconfiguration is challenging.

Sometimes Apache Flume suffers from scalability and reliability issues.

Flume Aplications

E-commerce companies use flume for analyzing customer behavior from the particular region.

Flume is useful for dumping large datasets produced by application servers into HDFS at a higher speed.

Flume is useful for detecting frauds.

It is useful in IoT applications.

We can use it for collecting and aggregating machine and sensor-generated data.

We can use Flume in the alerting or SIEM.

Summary

In short, Apache Flume is an open-source system for collecting and moving data from multiple servers to HDFS or HBase.
It can transfer data in real-time as well as in batch mode.
Flume is highly robust, scalable, and fault-tolerant.
It supports complex data flow as well as contextual routing.
The flume agent consists of a flume source, channel, and sink.
Larger set of channels, sources, and sink are supported by Apache Flume.
E-commerce companies use Flume to move their server data to HDFS and then process these data to understand customer behavior.