Unit - V

Flume

Data Ingestion Tool

Flume – Data Ingestion Tool - HDFS

- Flume is a popular data ingestion tool for Hadoop.

It is used to

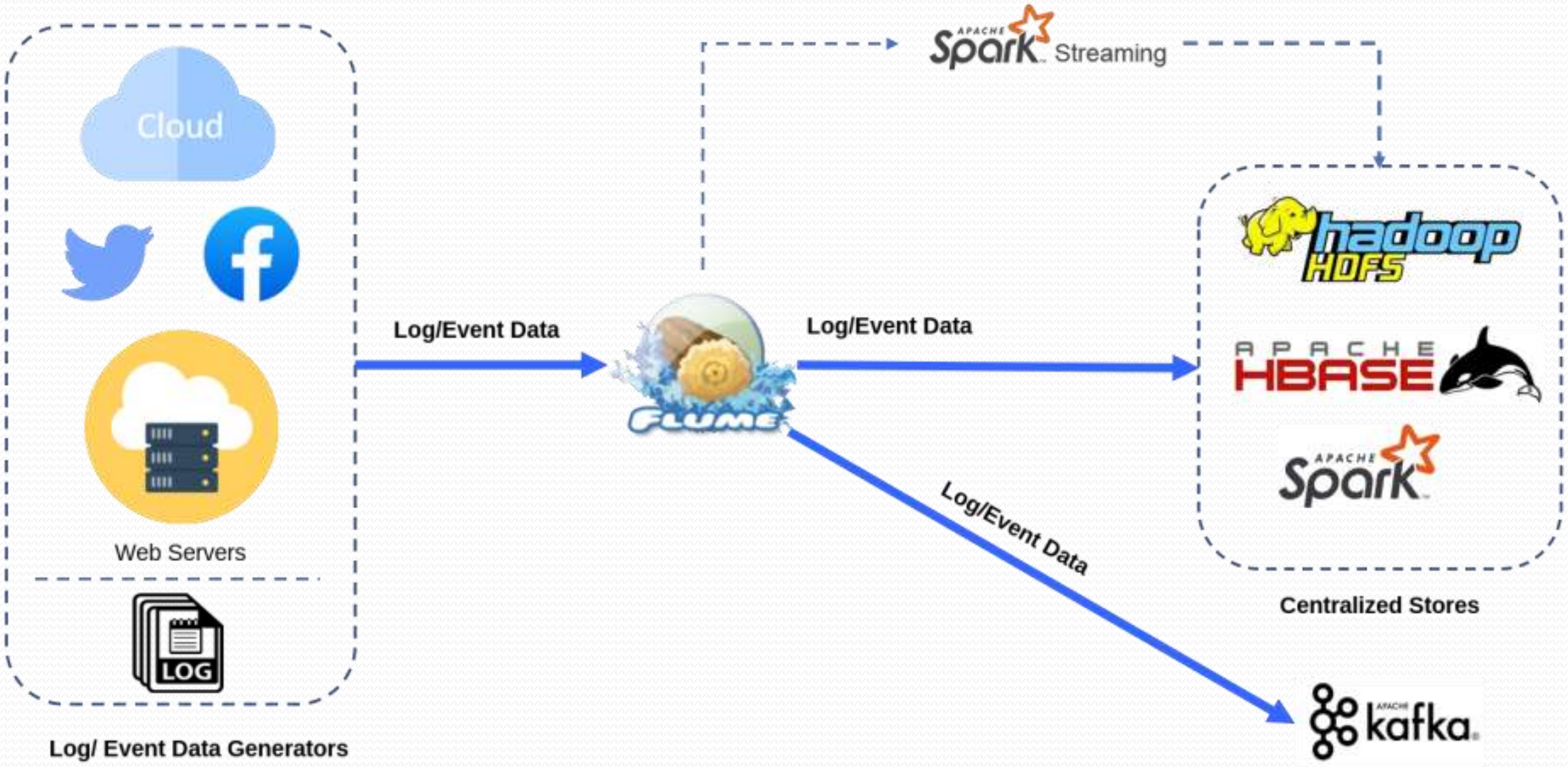
- ✓ Collect,
- ✓ Aggregate
- ✓ Move Large Amounts Of Streaming Data From A Variety Of Sources To Hadoop Distributed File System (HDFS).

- Flume is highly scalable and reliable, making it ideal for handling high-volume data streams.

- Flume is used by a wide range of companies, including

- ✓ Twitter
- ✓ LinkedIn
- ✓ Sales force - to collect and process data from a variety of sources.

Flume Data Flow



- Flume supports a variety of aggregation mechanisms, such as

✓ batching

✓ buffering

✓ Filtering

Here are some examples of how Flume can be used in Hadoop:-

✓ To collect and store **Web Server Logs** in HDFS for analysis.

✓ To collect and store **Social Media Data** in HDFS for analysis.

✓ To collect and store **IOT Data** in HDFS for analysis.

✓ To **Build A Real-time Data Pipeline** that ingests data from a variety of sources and stores it in HDFS for processing and analysis.

Key Features Of Apache Flume:-

- ✓ **Distributed**: Flume is designed to be distributed, meaning that it can be deployed across multiple servers to handle large amounts of data.
- ✓ **Reliable**: Flume is designed to be highly reliable, with features such as fault tolerance and data replay.
- ✓ **Scalable**: Flume is designed to be scalable, meaning that it can handle large amounts of data and can be scaled up or down to meet the needs of your application.
- ✓ **Flexible**: Flume is designed to be flexible, with a wide range of supported sources and destinations.
- ✓ **Extensible**: Flume is designed to be extensible, with the ability to add custom sources, destinations, and channel processors.

Use Cases For Apache Flume:-

- ✓ **Log Aggregation**: Flume can be used to collect and aggregate log data from various sources, such as web servers, application servers, and database servers. This data can then be analyzed for insights into system performance and security.

- ✓ **Social Media Data Collection**: Flume can be used to collect data from social media platforms, such as Twitter and Facebook. This data can then be analyzed for insights into customer sentiment, trends, and marketing campaigns.

- ✓ **Real-Time Analytics**: Flume can be used to collect and process data in real time, enabling real-time analytics. This can be used for a variety of purposes, such as fraud detection, anomaly detection, and personalized recommendations.

If you are looking for a reliable and scalable solution for streaming data collection and aggregation, Apache Flume is a good option to consider.

Data Sources

Apache Flume Supports A Wide Range Of Data Sources:-

- **Log files**: Flume can collect log files from various sources, such as web servers, application servers, and database servers.
- **Network Streams:** Flume can collect data from network streams, such as TCP sockets and UDP sockets.
- **Messaging Systems**: Flume can collect data from messaging systems, such as Kafka, RabbitMQ, and ActiveMQ.
- **Databases**: Flume can collect data from databases, such as MySQL, Postgre SQL, and Oracle.
- **Cloud Storage**: Flume can collect data from cloud storage, such as Amazon S3, Google Cloud Storage, and Microsoft Azure Blob Storage.
- **Other Sources**: Flume can also collect data from other sources, such as sensors, IoT devices, and social media platforms.

Examples Of Specific Data Sources With Flume:-

- ✓ **Syslog**: Flume can collect **Syslog Messages** from a variety of devices, such as routers, switches, and firewalls.

- ✓ **Netcat**: Flume can collect data from **Netcat Connections**. This can be useful for collecting data from custom applications or devices.

- ✓ **Kafka**: Flume can collect data from **Kafka Topics**. This is a common use case for Flume, as Kafka is a popular streaming data platform.

- ✓ **HDFS**: Flume can collect data from **HDFS Directories**. This can be useful for collecting data that has been processed by other Hadoop applications.

- ✓ **Twitter**: Flume can collect data from the **Twitter API**. This can be useful for collecting real-time data about customer sentiment, trends, and marketing campaigns.

- ✓ Flume also supports **Custom Sources**. This means that you can develop your own source to collect data from any source that you need.

- ✓ Which data source you use will depend on your specific needs. If you are unsure which data source to use, you can consult the Flume documentation or contact the Flume community for assistance.

Flume Architecture

- Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data.
- It has a simple and flexible architecture based on streaming data flows.

Flume Architecture:-

The Flume architecture is based on the following components:

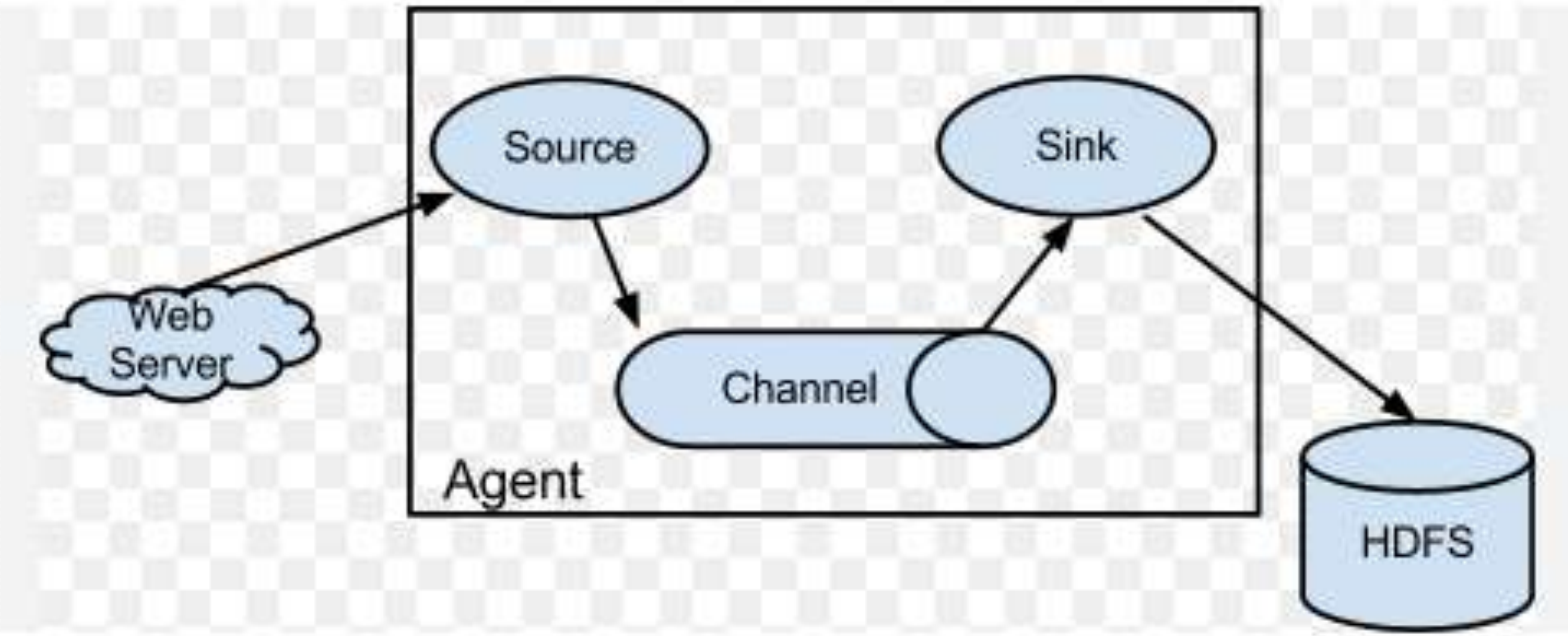
Events: The **Basic Unit Of Data In Flume** is an event.

✓ An event is a collection of key-value pairs, where the key is a string and the value is a byte array.

Sources: Sources are **Components That Generate Events**.

✓ Sources can be local or remote, and they can be configured to generate events at a specified frequency.

Flume Architecture



Channels: Channels are buffers that store events until they are processed by sinks.

- ✓ Channels can be configured to store events in memory, on disk, or in a distributed file system.

Sinks: Sinks are components that consume events from channels and write them to a destination.

- ✓ Sinks can be configured to write events to a variety of destinations, including HDFS, HBase, Kafka, and Elastic search.

Agents: Agents are daemon processes that **Manage Sources, Channels, And Sinks**.

- ✓ Agents can be configured to collect data from multiple sources, aggregate the data, and then write it to one or more sinks.

Flume Work Flow:-

- ✓ **Data Generators** generate events and send them to **Flume Agents.**
- ✓ Flume agents receive events from the data generators and **Store Them In Channels**.
- ✓ Flume agents then process the events in the channels and **Write Them To Sinks**.
- ✓ The **Collector** receives events from the Flume agents and **Aggregates Them**.
- ✓ The collector then writes the aggregated events to a data store, such as **HDFS or HBase**.

Benefits Of Flume:-

- ✓ **Scalability**: Flume is horizontally scalable, meaning that it can be scaled up by adding more agents.
- ✓ **Reliability**: Flume is designed to be reliable and fault-tolerant. It uses a variety of mechanisms to ensure that data is not lost, even if there are failures in the system.
- ✓ **Flexibility**: Flume is a flexible system that can be used to collect and move data from a variety of sources to a variety of destinations.

Use Cases For Flume:-

- ✓ **Log Collection**: Flume can be used to collect logs from servers and applications and store them in a centralized location for analysis.
- ✓ **Data Integration**: Flume can be used to integrate data from different sources into a single data lake.
- ✓ **Real-Time Analytics**: Flume can be used to collect and process streaming data in real time for analytics.