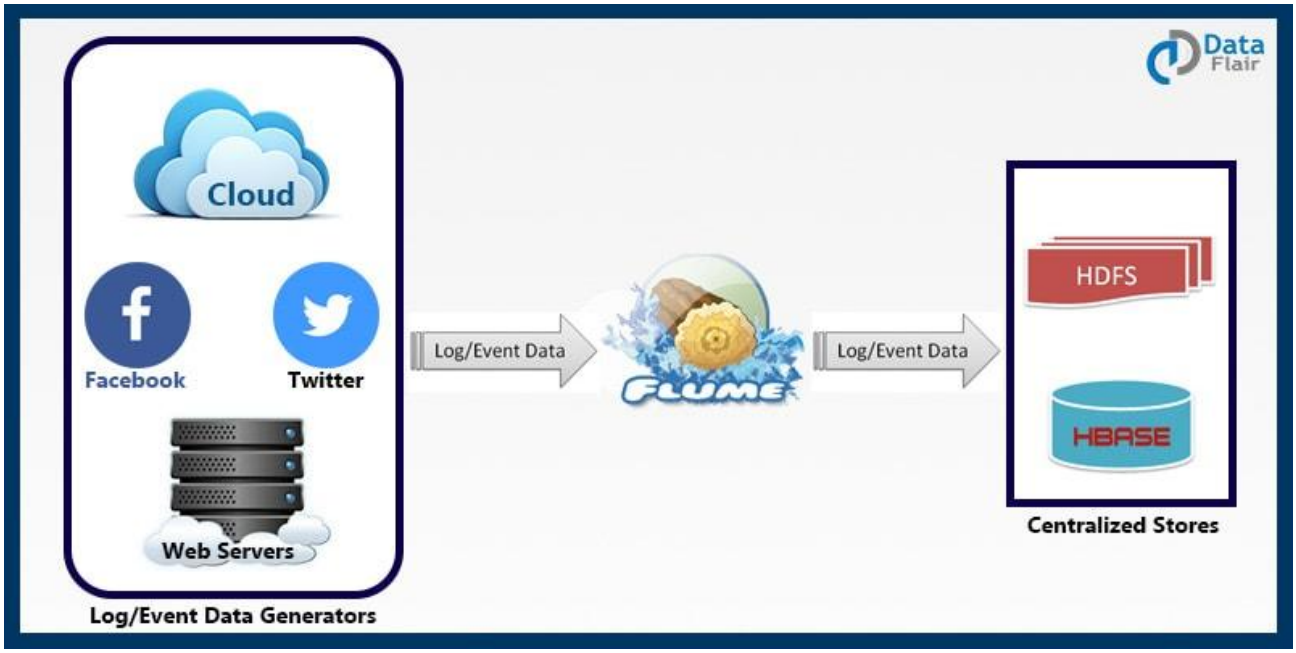
Apache Flume Tutorial – Flume Introduction, Features & Architecture [6]

Contents

1. Objective

Apache flume is an open source data collection service for moving the data from source to destination. In this Apache Flume tutorial, we will discuss what is Apache Flume, what is the need of Flume, features of Apache Flume. This tutorial also covers the architecture of flume, how Flume works and various advantages of Apache Flume.



2. Apache Flume Tutorial – Introduction

Apache Flume is a tool used to collect, aggregate and transports large amounts of streaming data like log files, events, etc., from a number of different sources to a centralized data store (say Hadoop Distributed File System – HDFS).

Flume is a highly distributed, reliable, and configurable tool. Flume was mainly designed in order to collect streaming data (log data) from various web servers to HDFS.

3. Why Apache Flume?

A company has tons of services running on multiple servers. And lots of data (logs) are produced by them, now we need to analyze them altogether. In order process that logs, we need a reliable, scalable, extensible and manageable distributed data collection service which can perform flow of unstructured data (logs) from one location to another where they will be processed (say in HDFS). Apache flume is an open source data collection service for moving the data from source to destination.

Apache Flume is the most reliable, distributed, and available service for systematically collecting, aggregating, and moving large amounts of streaming data (logs) into the Hadoop Distributed File System (HDFS). Based on streaming data flows, it has a simple and flexible architecture. It is highly fault-tolerant and robust and with tunable reliability mechanisms for fail-over and recovery. Flume allows data collection in batch as well as streaming mode.

4. Flume Features

Some of the outstanding features of Flume are as follows:

- From multiple servers, it collects the log data and ingests them into a centralized store (HDFS, HBase) efficiently.
- With the help of Flume we can collect the data from multiple servers in real-time as well as in batch mode.
- Huge volumes of event data generated by social media websites like Facebook and Twitter and various e-commerce websites such as Amazon and Flipkart can also be easily imported and analyzed in real-time.
- Flume can collect data from a large set of sources and move them to multiple destinations.
- Multi-hop flows, fan-in fan-out flows, contextual routing, etc are supported by Flume.
- Flume can be scaled horizontally.

4. Apache Flume Architecture

In this section of Apache Flume tutorial, we will discuss various components of Flume and how this components work-

- **Event:** A single data record entry transported by flume is known as Event.
- **Source:** Source is an active component that listens for events and writes them on one or channels. It is the main component with the help of which data enters into the Flume. It collects the data from a variety of sources, like exec, avro, JMS, spool directory, etc.

- **Sink:** It is that component which removes events from a channel and delivers data to the destination or next hop. There are multiple sinks available that delivers data to a wide range of destinations. Example: HDFS, HBase, logger, null, file, etc.
- **Channel:** It is a conduit between the Source and the Sink that queues event data as transactions. Events are ingested by the sources into the channel and drained by the sinks from the channel.
- **Agent:** An Agent is a Java virtual machine in which Flume runs. It consists of sources, sinks, channels and other important components through which events get transferred from one place to another.
- **Client:** Events are generated by the clients and are sent to one or more agents.

5. Advantages of Flume

Below are some important advantages of using Apache Flume:

- Data into any of the centralized stores can be stored using Apache Flume.
- Flume acts as a mediator between data producers and the centralized stores when the rate of incoming data exceeds the rate at which data can be written to the destination and provides a steady flow of data between them.
- A feature of contextual routing is also provided by Flume.
- Flume guarantees reliable message delivery as in Flume transactions are channel-based where two transactions (1 sender & 1 receiver) are maintained for each message.
- Flume is highly fault-tolerant, reliable, manageable, scalable, and customizable.