Flume is used to ingest to unstructured data from to Hadoop.

Collects data efficiently from different sources to a centralized store.

Simple & flexible architecture

**Architecture**

Source is used to connect to a source storage & sends to channel & Sink is used to data to HDFS

Channel types

1. Memory channel – faster, no fault tolerance
2. File channel – backed by local file system, has fault tolerance

We need to mention which channel to used in script

Sink

Source, channel & sink together are called flume agent

**Pig**

Pig converts scripts into m/R

Syntax is similar to SQL

Can deal with csv files

Doesn’t have control structures, loops

Not suitable for all tasks for tasks involved with audios, videos

**Architecture**

Pig script is parsed, compiled, optimized &planned then converted to mapreduce jobs -> YARN 🡪 HDFS

**Modes**

1. **Local mode: job runs in single JVM process.**

**It js for testing**

**Picks & stores data from local linux path**

**2.**

**3,**

Pig Latin – language

Pig Engine – execution environment for pig latin

Pig supports shell commands 🡪 should mention fs followed by shell command