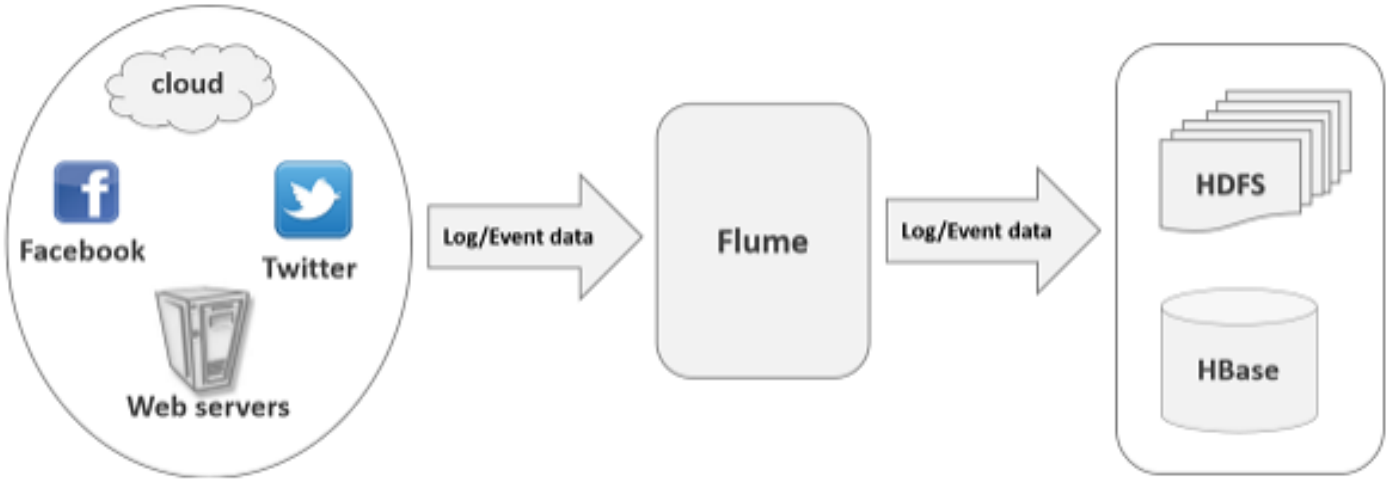
Flume

مقدمه

- Flume یک ابزار استاندارد ساده، قابل انعطاف، قوی و قابل توسعه برای جمع آوری داده از تولید کنندگان مختلف داده (وب سرورها) به هادوپ است.

Flume چیست؟

- Flume یک ابزار برای جمع آوری و انتقال داده های حجیم نظیر log فایل ها، events ها و... از منابع مختلف به یک مرکز داده مرکزی می باشد.



log فایل ها

مرکز داده مرکزی

کاربرد Flume

- تصور کنید یک وب اپلیکیشن که بصورت آنلاین کار تجاری انجام می دهد بخواهد رفتار مشتریها را در یک استان خاص آنالیز کند.
- برای اینکار، باید داده های log فایل ها را برای آنالیز به هادوپ بفرستند.

✓Flume به نجات آنها می آید.

Flume برای انتقال داده های ایجاد شده توسط سرورها به HDFS با سرعت بالا استفاده می شود.

مزایای Flume

- به کمک Flume می توان داده ها را در هر محل ذخیره سازی مرکزی (HBase, HDFS) ذخیره کرد.
- زمانی که حجم داده های ایجاد شده به نسبت میزان داده ای که می تواند در مقصد ذخیره شود بسیار بیشتر است، Flume به عنوان یک واسطه بین آندو عمل می کند و جریان ثابتی از داده را بین آنها فراهم می کند.

مزایای Flume

- ساز و کار Flume بر اساس کانال انجام می شود (بین فرستنده و گیرنده) که تحویل پیام قابل اعتماد را گارانتی می کند.
- Flume قابل اعتماد، تحمل پذیر در برابر خطا، مقیاس پذیر، قابل مدیریت و برنامه ریزی با قالیت سفارشی سازی است.

ویژگیهای Flume

- با استفاده از Flume می توان داده ها را از چندین سرور بلافاصله به هادوپ فرستاد.
- علاوه بر log files، Flume برای انتقال حجم زیادی از داده های ایجاد شده در شبکه های اجتماعی مانند Facebook , Twitter ویا سایتهای تجاری مثل آمازون را داراست.

Log file

- در حالت کلی، یک Log file فتیلی است که اتفاقات/عملیات رخ داده در یک سیستم عامل را لیست می کند.
- با استفاده از Log file می توان:
  - اطلاعاتی درباره کارایی اپلیکیشن بدست آورد
  - محل اشکالات سخت افزاری و نرم افزاری را تشخیص داد
  - اطلاعاتی درباره فعالیتهای یک کاربر بدست آورد که می تواند برای تجارت مهم باشد

HDFS put Command

- شل هادوپ فرمان هایی برای وارد کردن یا خواندن داده ها از هادوپ را فراهم می کند.
- برای مثال دستور put :

```
$ Hadoop fs -put /path of the required file  /path in HDFS where to save the file
```

HDFS put Command

- مشکلات دستور put :
  - در هر زمان فقط امکان انتقال یک فایل وجو د دارد.
  - در حالیکه داده ها ی تولید شده با سرعت بسیار زیادی تولید می شوند، در نتیجه نتایج تحلیل روی داده های قدیمی است و در نتیجه نمی تواند صحیح باشد.
  - دستور put باید داده ها را بسته بندی کرده تا برای آپلود اماده شوند. از انجایی که وب سرورها داده ها را بصورت دائم تولید می کنند، کار را بسیار سخت می کند.

راه حل های موجود

- Facebook’s Scribe :
  - Scribe یک ابزار محبوب است که برا ی جمع آوری جریان داده ها استفاده می شود.
- Apache Kafka :
  - Kafka توسط شرکت آپاچی توسعه داده شده است.
  - Kafka یک واسط متن باز است
- Apache Flume :

معماری Flume

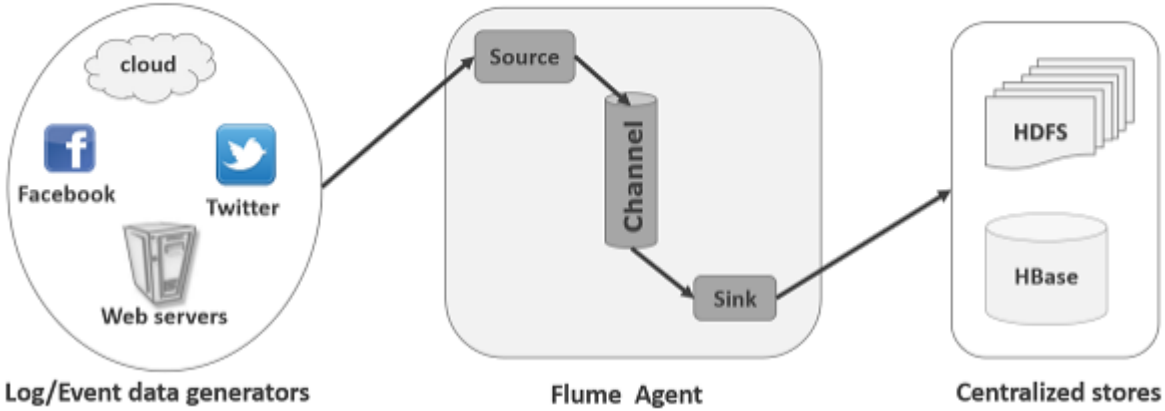
Flume Event

- یک event واحد اصلی داده منتقل شده درون Flume است.
- در قسمت payload داده هایی که باید به مقصد منتقل شوند قرار دارند.
- قسمت header انتخابی است.

Flume event

Flume Agent

- همانطور که در شکل زیر مشاهده می شود یک agent یک پروسه مستقل در flumeاست.
- داده ها را (eventها) را از کاربر دریافت کرده و به مقصد بعدی هدایت می کند

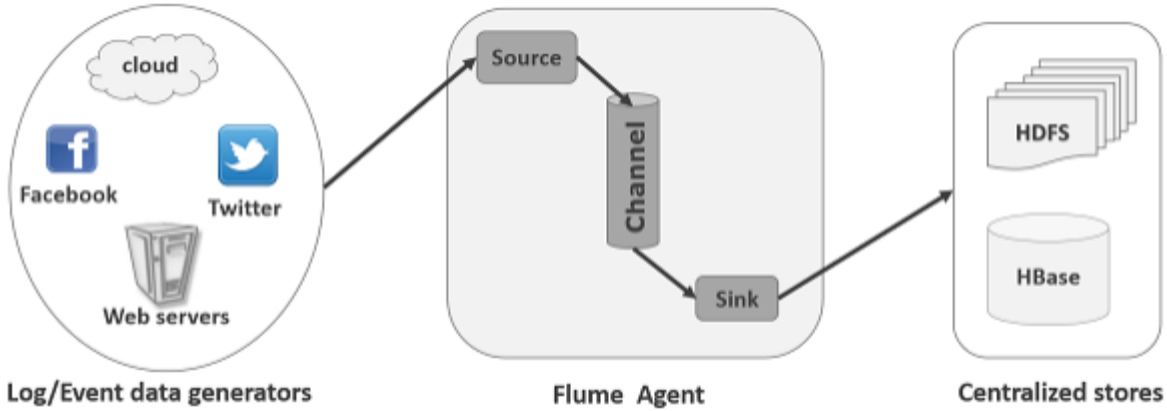


- یک agent شامل سه مولفه اصلی است:

source, channel, sink

Flume Agent

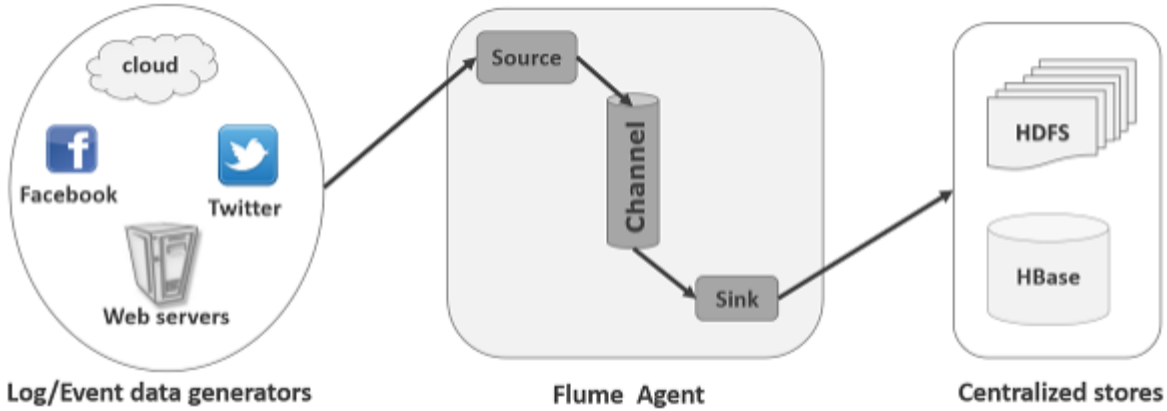
- Source :
- یک source داده ها را از یک تولید کننده log/event دریافت می کند (مانند Facebook, Twitter) و آنها را به یک کانال به شکل Flume events منتقل می کند.
- داده ها ممکن است به یک یا چندین کانال منتقل شوند
- انواع داده های مختلفی که توسط سرورهای مختلف تولید می شوند، توسط source پشتیبانی می شوند



Flume Agent

: Channel

- یک Channel یک محل ذخیره سازی موقت است که events ها را از source دریافت کرده و آنها را تا زمانی که توسط یک sink مصرف شوند بافر می کند
- Channel به عنوان یک پل بین sources و sinks عمل می کند.
- می تواند با هر تعداد sources , sinks کار کند.



Flume Agent

: Sink

- در نهایت، یک Sink داده ها را در یک محل ذخیره سازی مانند Hbase و HDFS ذخیره می کند.

- یک Sink داده ها (events) را از channels گرفته و تحویل مقصد می دهد.
- مقصد می تواند یک agent دیگر باشد

```
agent.sinks.k1.type=hdfs
agent.sinks.k1.hdfs.path=/path/in/hdfs
```

Additional Components of Flume Agent

What we have discussed above are the primitive components of the agent. In addition to this, we have a few more components that play a vital role in transferring the events from the data generator to the centralized stores.

Interceptors

Interceptors are used to alter/inspect flume events which are transferred between source and channel.

Channel Selectors

These are used to determine which channel is to be opted to transfer the data in case of multiple channels. There are two types of channel selectors:

- **Default channel selectors**: These are also known as replicating channel selectors they replicates all the events in each channel.

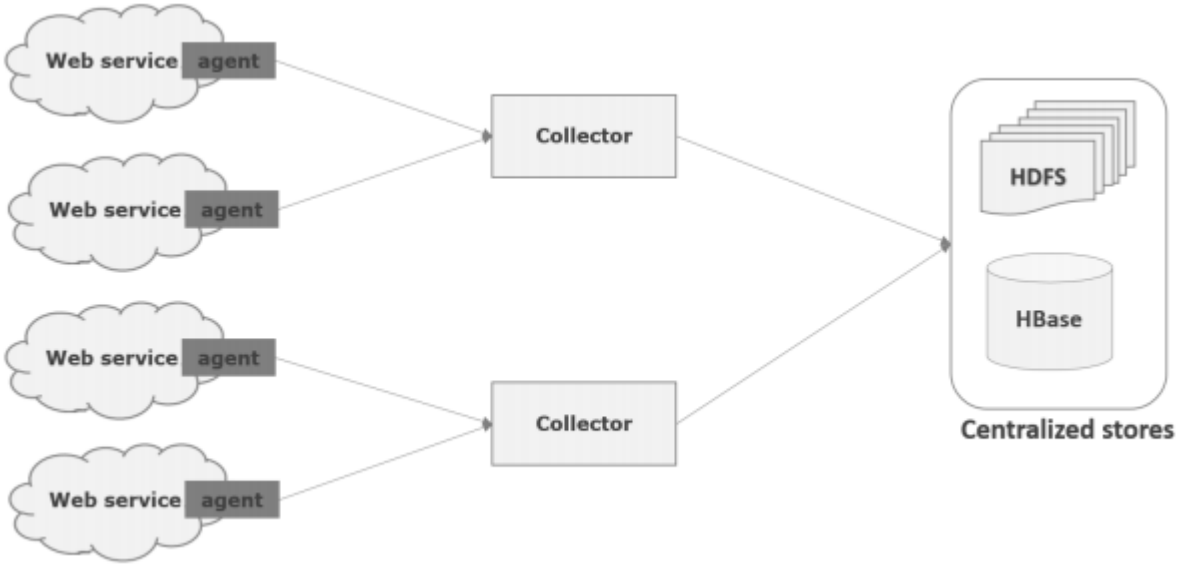
- **Multiplexing channel selectors**: These decides the channel to send an event based on the address in the header of that event.

Sink Processors

These are used to invoke a particular sink from the selected group of sinks. These are used

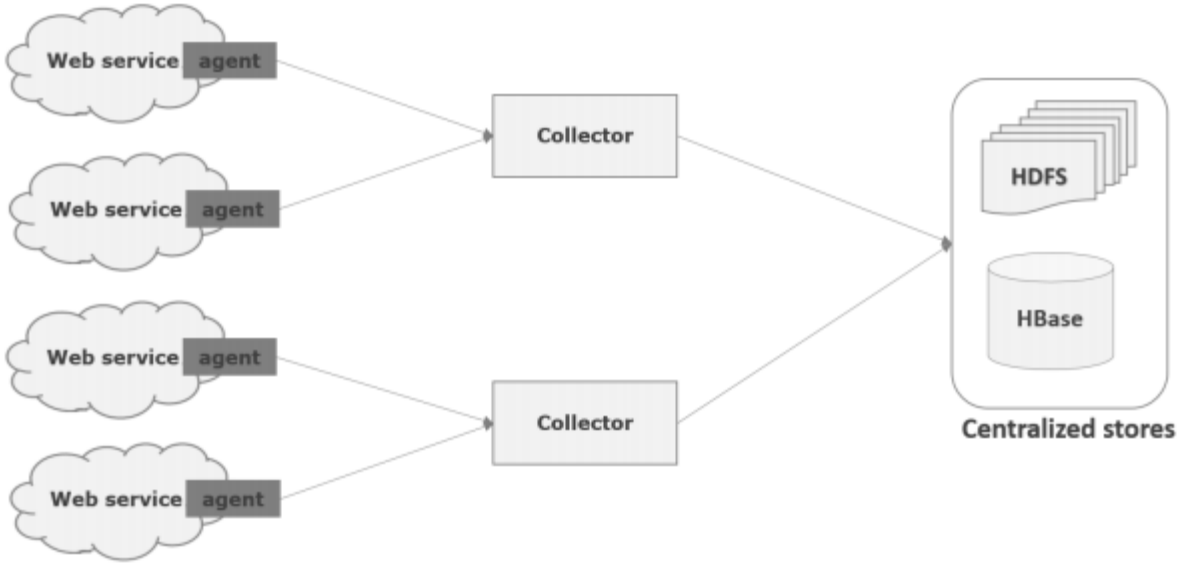
FLUME – DATA FLOW

- در حالت کلی، event ها توسط سرورهایی تولید می شوند که Flume agents هایی را روی خود در حال اجرا دارند.
- این agentها داده ها را از تولید کننده داده، دریافت می کنند.



FLUME – DATA FLOW

- این داده ها توسط نودهایی میانی به نام Collector جمع اوری می شوند.
- همانند agents ، در Flume ممکن است چندین Collector وجود داشته باشد

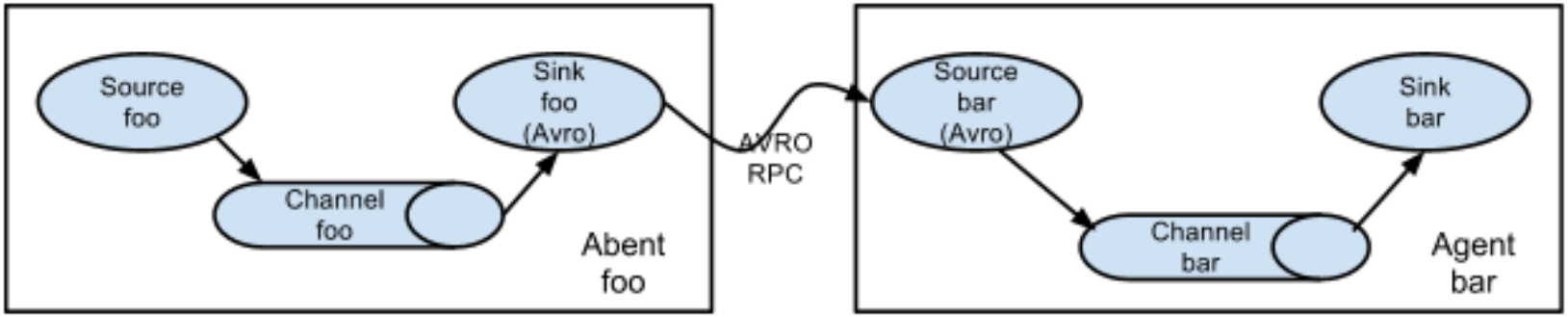


Multi-hop Flow

- قبل از رسیدن به مقصد نهایی، ممکن است چندین agents وجود داشته باشد.
- یک event ممکن است از چندین agent عبور کند که به عنوان multi-hop flow شناخته می شود.

Multi-hop Flow

- قبل از رسیدن به مقصد نهایی، ممکن است چندین agents وجود داشته باشد.
- یک event ممکن است از چندین agent عبور کند که به عنوان multi-hop flow شناخته می شود.

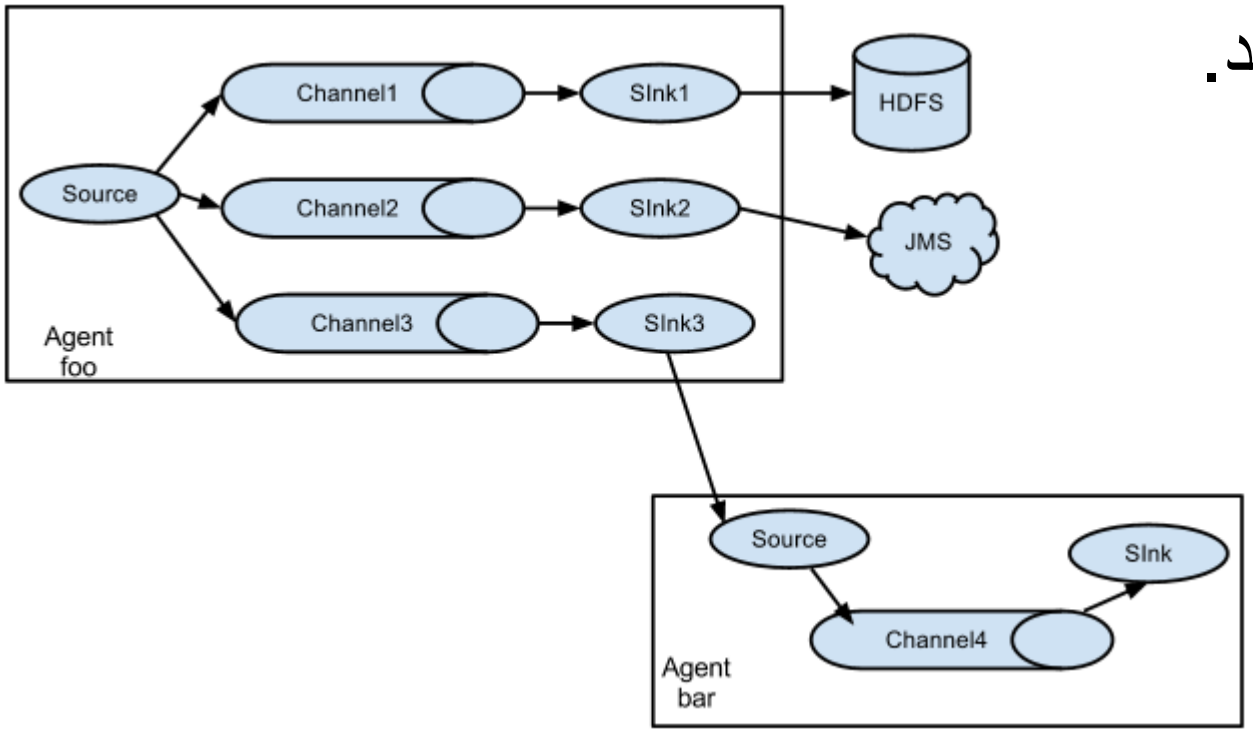


Fan-out Flow

- جریان داده از یک منبع به چندین channels به fan-out flow معروف است که دو نوع دارد:
- Replicating : جریان داده ای که داده باید در همه کانال ها تکرار شود

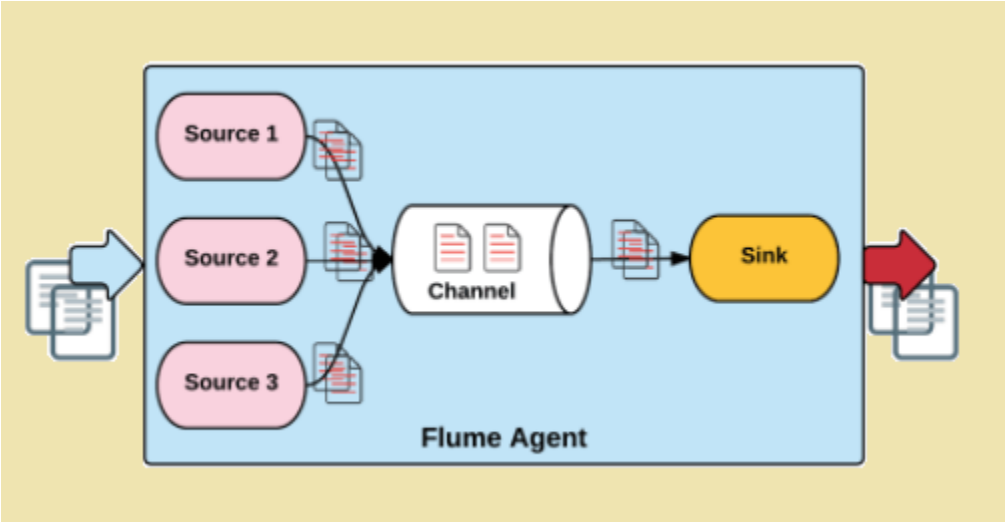
Fan-out Flow

- جریان داده از یک منبع به چندین channels به fan-out flow معروف است که دو نوع دارد:
- Multiplexing : جریان داده ای که داده باید به کانال های انتخابی که در قسمت header یک event مشخص شده فرستاده شوند.



Fan-in Flow

- جریان داده ای که داده از چندین منبع به یک کانال منتقل می شوند Fan-in Flow هستند



Failure Handling

- در Flume ، برای هر event دو تراکنش رقم می خورد: یک در سمت فرستنده و دیگری در سمت گیرنده. یک فرستنده که یک event را برای گیرنده می فرستد منتظر دریافت پیام تایید از سمت گیرنده می ماند. فرستنده پایان تراکنش را اعلام نمی کند مگر اینکه پیام تایید را دریافت کرده باشد.