# Take Notes

Kafka Connect and Kafka Streams are two built-in libraries to the base Kafka packaging

If you want to put data into Hadoop, Confluent maintains an [HDFS Connect plugin](https://docs.confluent.io/current/connect/connect-hdfs/docs/hdfs_connector.html#quickstart). If you want to get data "from APIs", that's a little vague, but you could write Kafka Connect tasks yourself that are "source connectors", otherwise tools such as Apache NiFi or Streamsets integrate well with many products (and you may not even need Kafka using these two, other than it being a persistent message bus)

If you did want to program something on a larger scale, you can look into Spark, or Flink.

If you want to use any format, Spark, Flink, NiFi, or Streamsets have more flexibility for handling that out of the box

Another solution includes using NiFi or Streamsets instead of Flume

NiFi and Streamsets don't offer support for setting the record timestamps (yet).

First of all, Spark doesn't have a filesystem. That's why primarily why Hadoop is nice, in my book. Sure, you can use S3, or many other cloud storages, or bare metal data stores like Ceph, or GlusterFS, but from what I've researched, HDFS is by far the fastest when processing data.

Maybe you're not familiar with the concept of rack locality that YARN offers. If you use Spark Standalone mode with any file system not mounted under the Spark executors, then all your data requests will need to be pulled over a network connection, therefore saturating the network, and causing a bottleneck, regardless of memory. Compare that to the Spark executors running on the YARN NodeManagers, HDFS datanodes are ideally also NodeManagers.

A similar problem - people say Hive is slow, SparkSQL is faster. Well, that's true if you run Hive with MapReduce instead of Tez or Spark execution modes.

Now, if you're wanting streaming and real-time events rather than the batch world commonly associated with Hadoop. You might want to research the SMACK stack.

You can use NiFi; nothing is stopping you. It would run closer to real-time than Spark micro batches. And it is a good tool to pair with Kafka.

It's up to some BigData architect in your company to come up with a solution. You'll find that vendor support from Confluent is mostly for Kafka. I haven't seen them talking about Samza much. Hortonworks will support Storm, Nifi, and Spark, but they aren't running the latest version of Kafka if you want fancy features like KSQL. Streamsets is a similar company offering a tool competing with NiFi which consists of employees with backgrounds in other batch/streaming Apache projects.

Storm and Samza are two ways to do the same thing, as far as I know. I think Flink is more programmer friendly than Storm. I don't have experience with Samza, though I work closely with people who primarily are using Kafka Streams rather than it. And Kafka Streams isn't DAG based - it's just a high level Kafka library, embeddable in any JVM application.

<https://stackoverflow.com/questions/48104475/relevance-of-hadoop-streaming-solutions-when-spark-exists/48105619#48105619>

If you want just a FS.a ==> FS.b "real-time" pipe, the best options I know of are either Apache NiFi or StreamSets because there is no coding required.

KAFKA to StreamSets issue:

<https://stackoverflow.com/questions/48167694/kafka-cannot-retrieve-metadata-for-topic-when-changing-active-controller>

<https://stackoverflow.com/questions/44995605/how-to-parse-multiple-lines-record-of-log-file-using-streamsets?rq=1>

<https://github.com/streamsets/tutorials/tree/master/tutorial-origin>

Additional Driver

I had to add the ojdbc jar under the repository

/opt/streamsets-extra/streamsets-datacollector-jdbc-lib/lib

Your pipeline is reasonable, though I might suggest using Apache NiFi (which is in the Hortonworks HDF distribution), or Streamsets, which is installable in CDH easily, from what I understand.

# **Scheduling the JDBC consumer job in Stream Sets** ?

There is no built-in scheduler in SDC, but you could use cron and the [StreamSets CLI](https://streamsets.com/documentation/datacollector/latest/help/index.html#Administration/Administration_title.html#concept_ywx_d5x_pt) to start the pipeline.

## https://stackoverflow.com/questions/36899612/difference-between-apache-nifi-and-streamsets?rq=1

## Email:

BigDecimal

Avro, Parquet formats

Username, password encryption for data sources