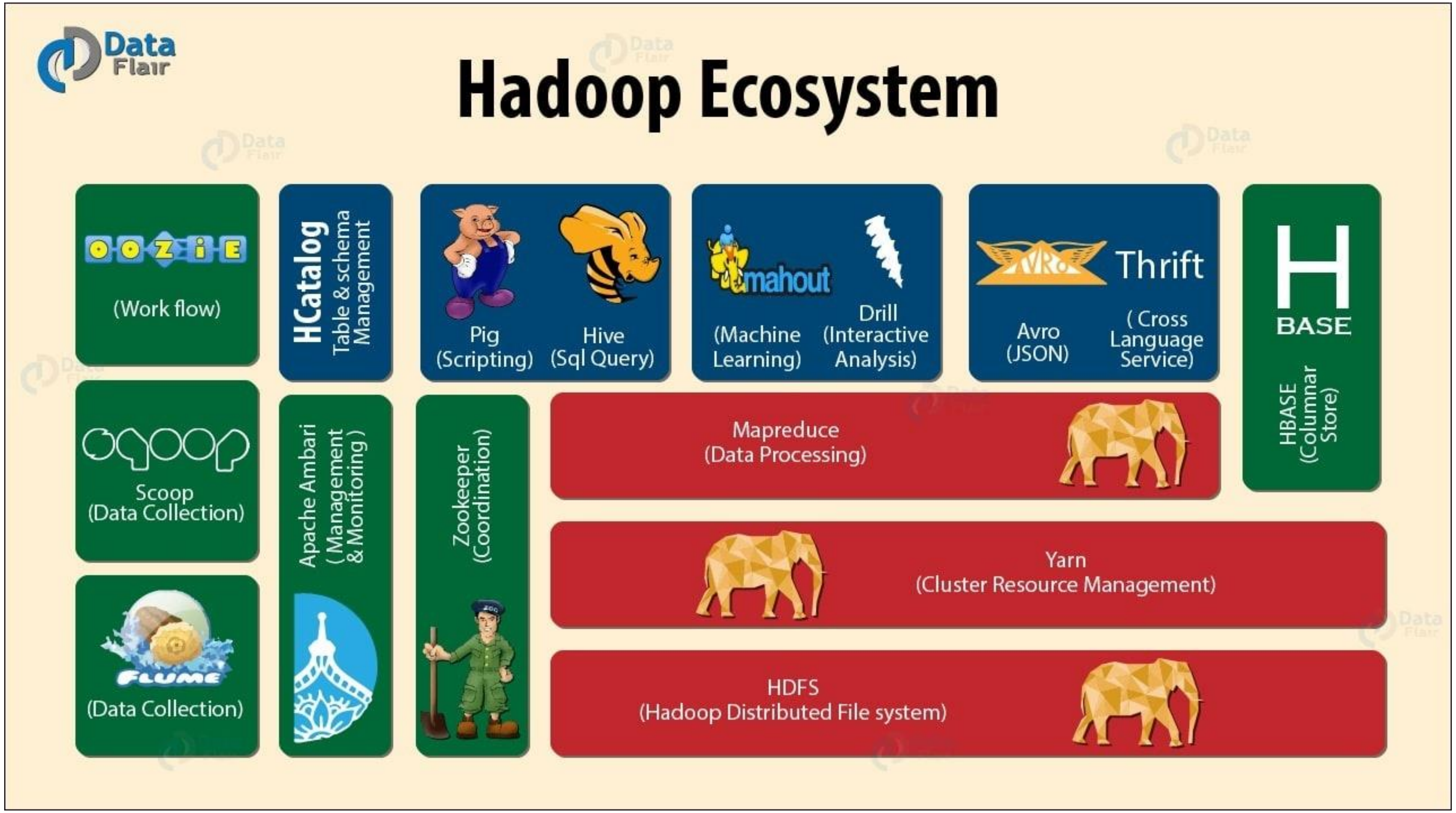
DATA INGRESS AND EGRESS IN HADOOP

Apache Flume and SQOOP

THE CONTEXT

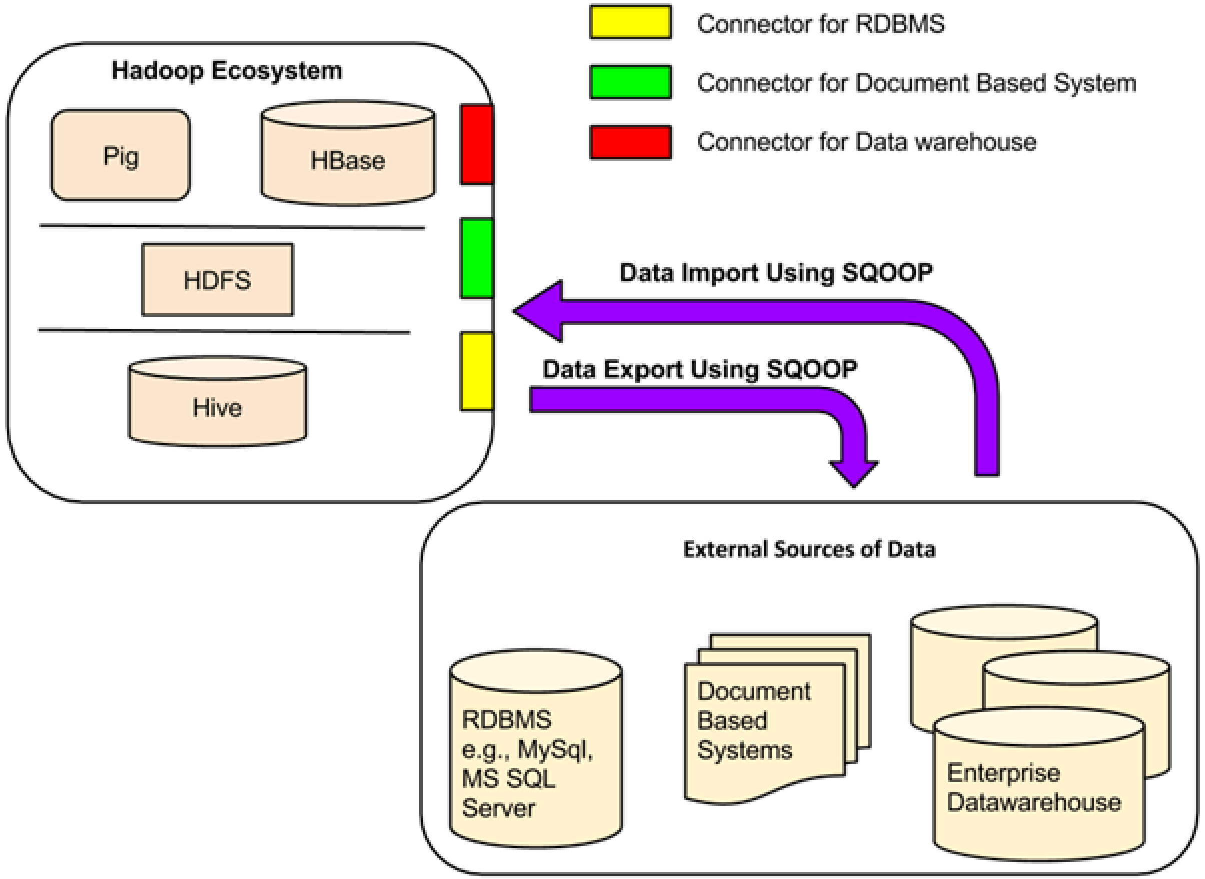
- So far you have been placing data into the Hadoop cluster using the **put** command
- This method is not very efficient or realistic for ingesting large amounts of data and streaming data
- Although methods for data ingress and egress operations in Hadoop are increasing and expanding, there are 2 fundamental methods that data analysts typically use to bring data into Hadoop:
  - Flume: used to import unstructured data, like log files
  - SQOOP (**SQ**L meets Had**oop**): used to bring in structured, typically tabular, sources of data

REFERRING BACK TO THE ECOSYSTEM



- HBASE is a NoSQL Database running on Hadoop and it persists semi and complex structured data in a Big Table/Wide Column format
- Hive is Hadoop's version of a data warehouse. It has a metastore to house tabular schema for HDFS data and it can persist its schemas to HCatalogue - an interface point for other APIs

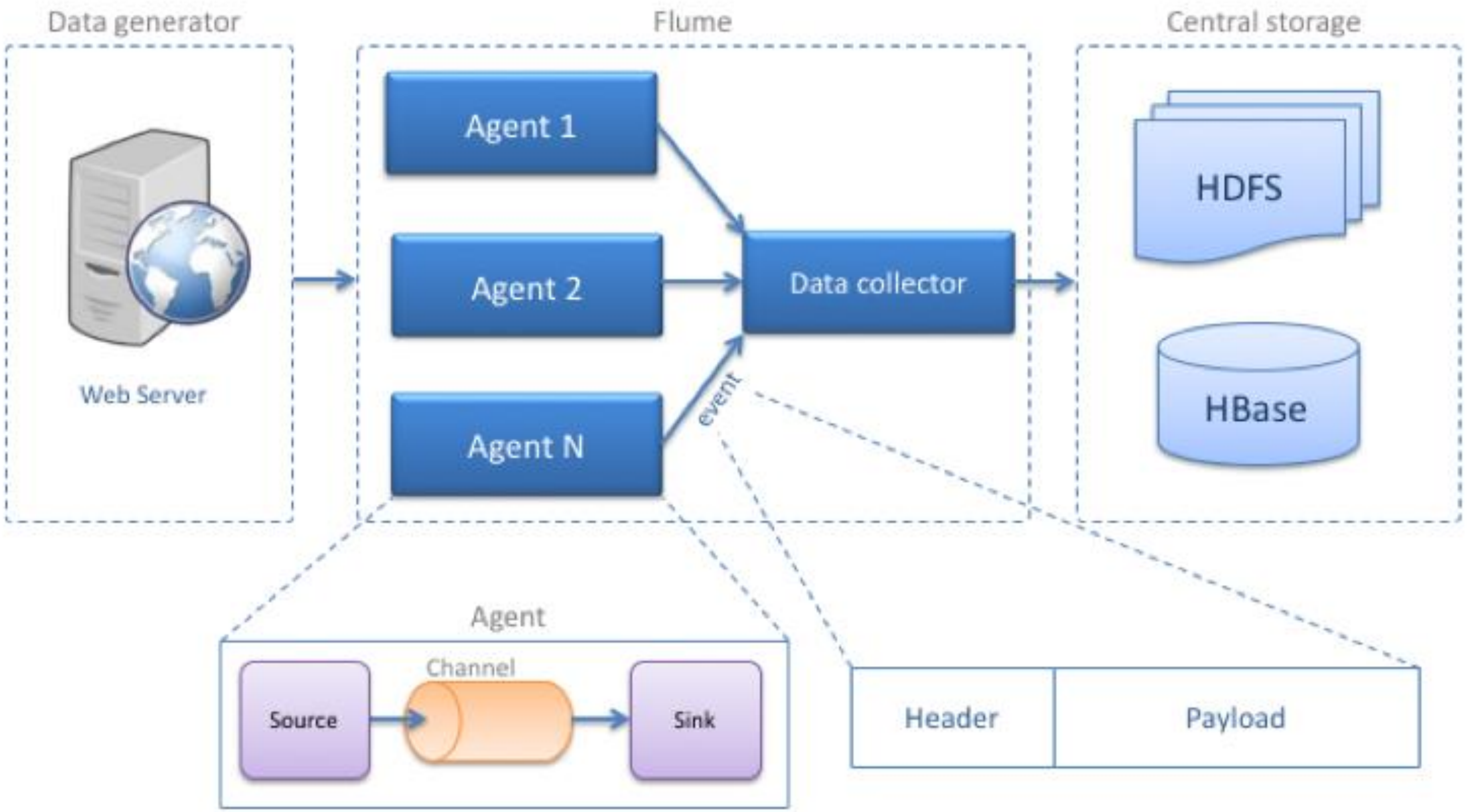
SQOOP INGRESS AND EGRESS



- SQOOP has different connectors to import and export data to and from database sources
- When we import data using SQOOP, its entry points into the Hadoop architecture are in HDFS and Hive

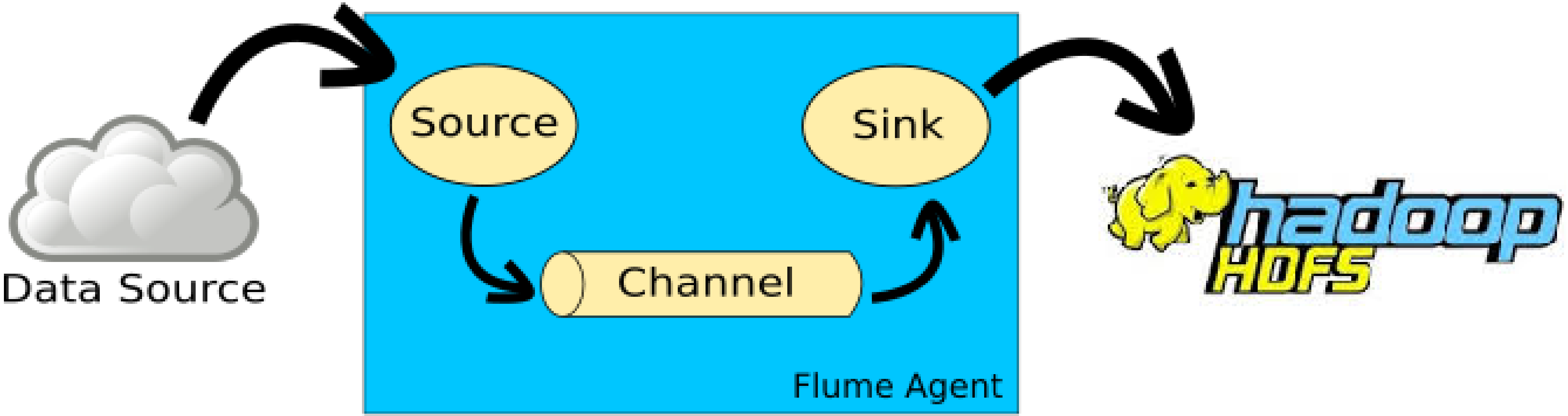
FLUME AND DATA INGRESS

- Flume uses configured agents to transfer unstructured or semi-structured data from external sources to HDFS and HBase



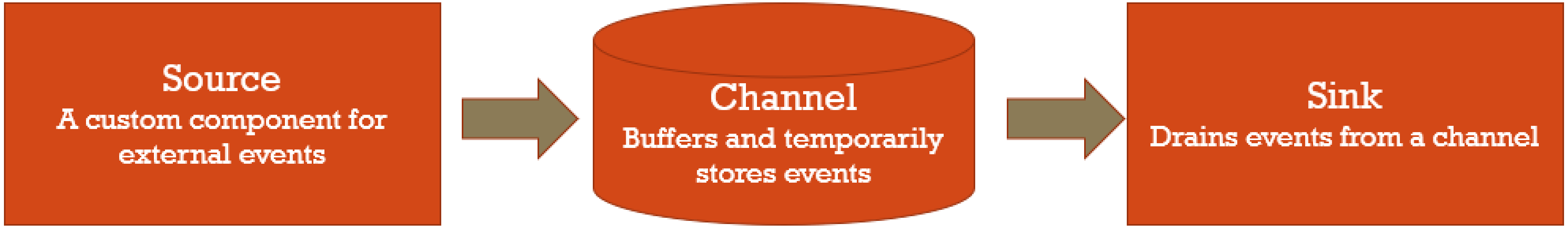
FLUME COMPONENTS AND ARCHITECTURE

- In Flume we have a data source and a destination (Hadoop). In between them lies the Flume Agent, which is a Java Process that links the source to its destination



THE FLUME AGENT

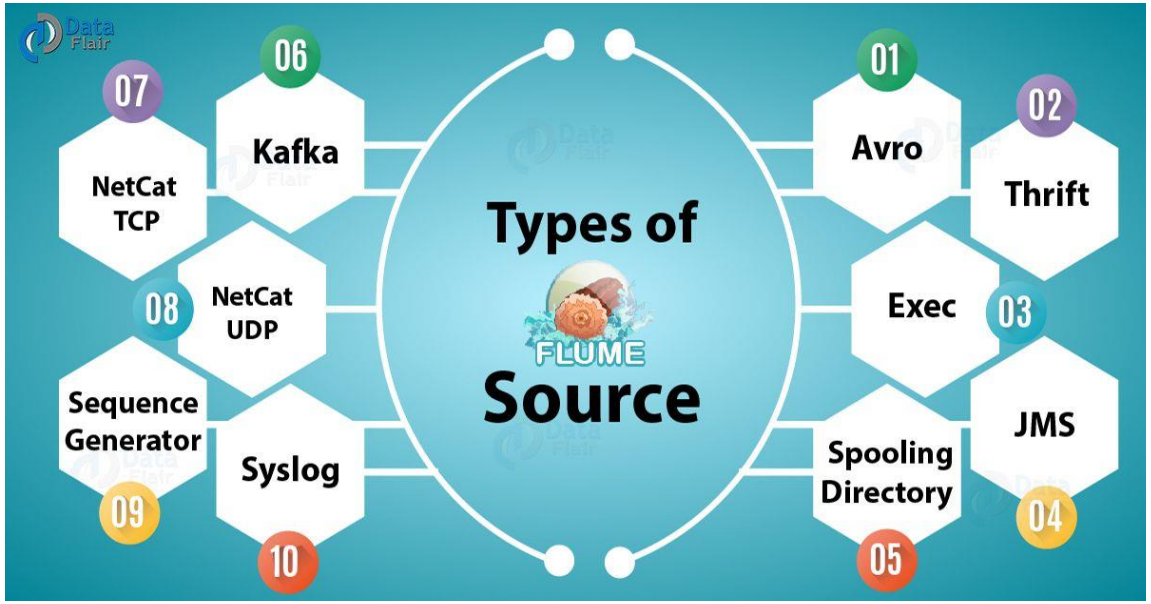
- Broadly speaking, the Flume Agent has 3 sections of Java property configurations:
  - **Source** properties to describe the origin
  - **Sink** properties to describe the destination
  - **Channel** properties, which join the source object to the destination object

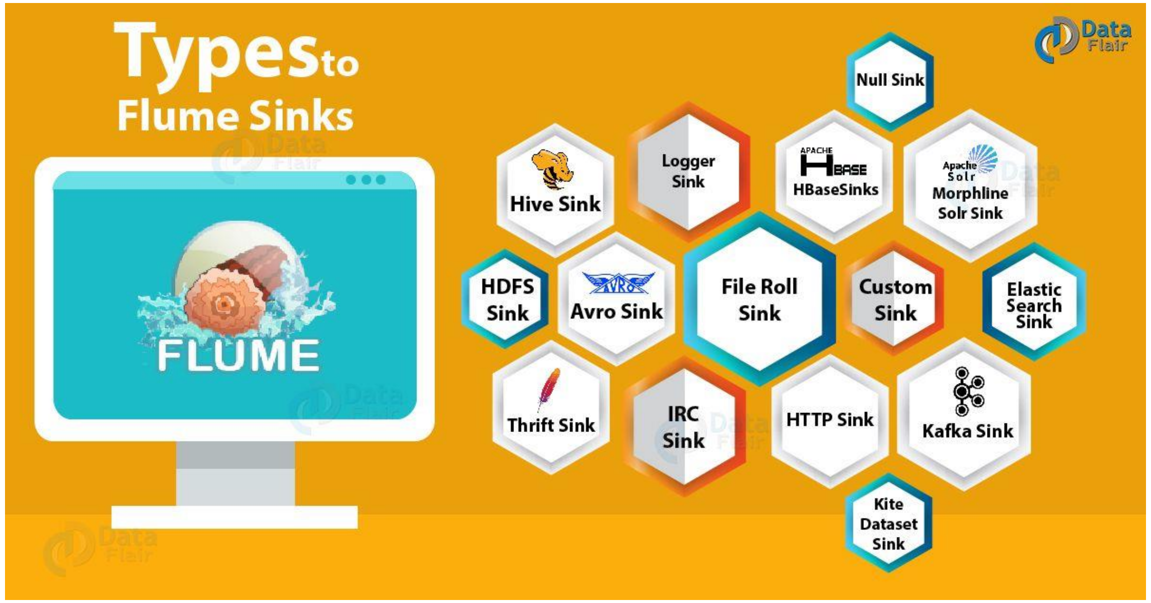


Source
A custom component for external events
Channel
Buffers and temporarily stores events
Sink
Drains events from a channel

TYPES OF FLUME SOURCES, CHANNELS, AND SINKS

- The following chart lists many of the current kinds of sources, channels, and sinks

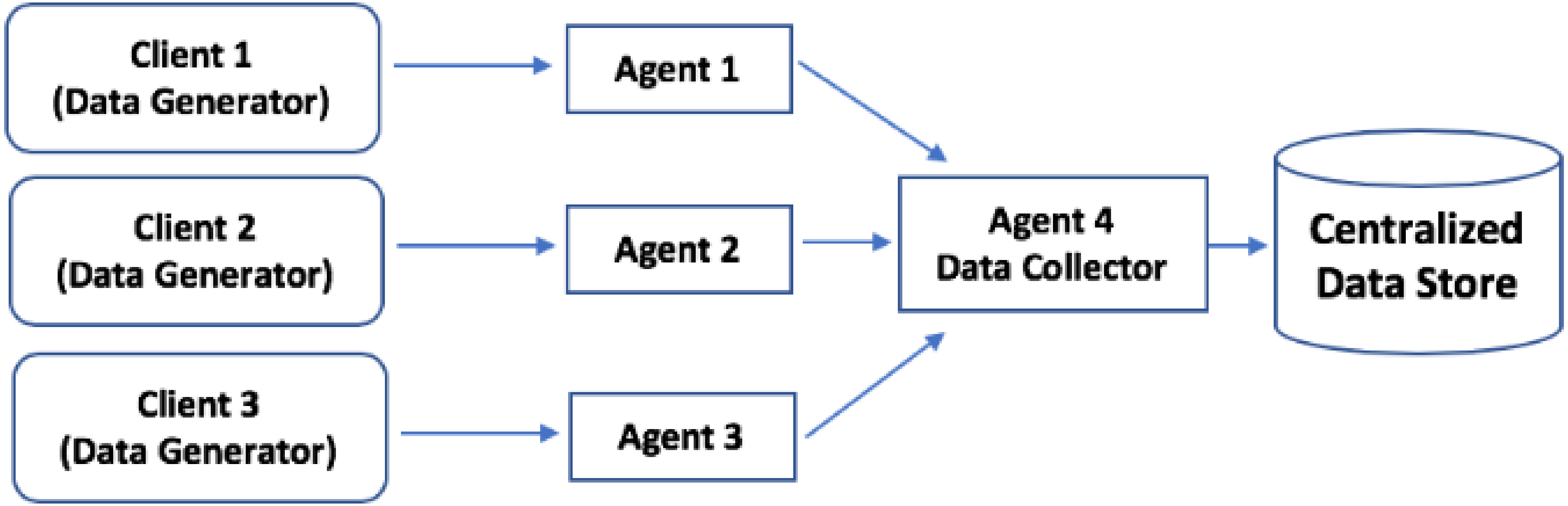




DEMONSTRATION TIME!

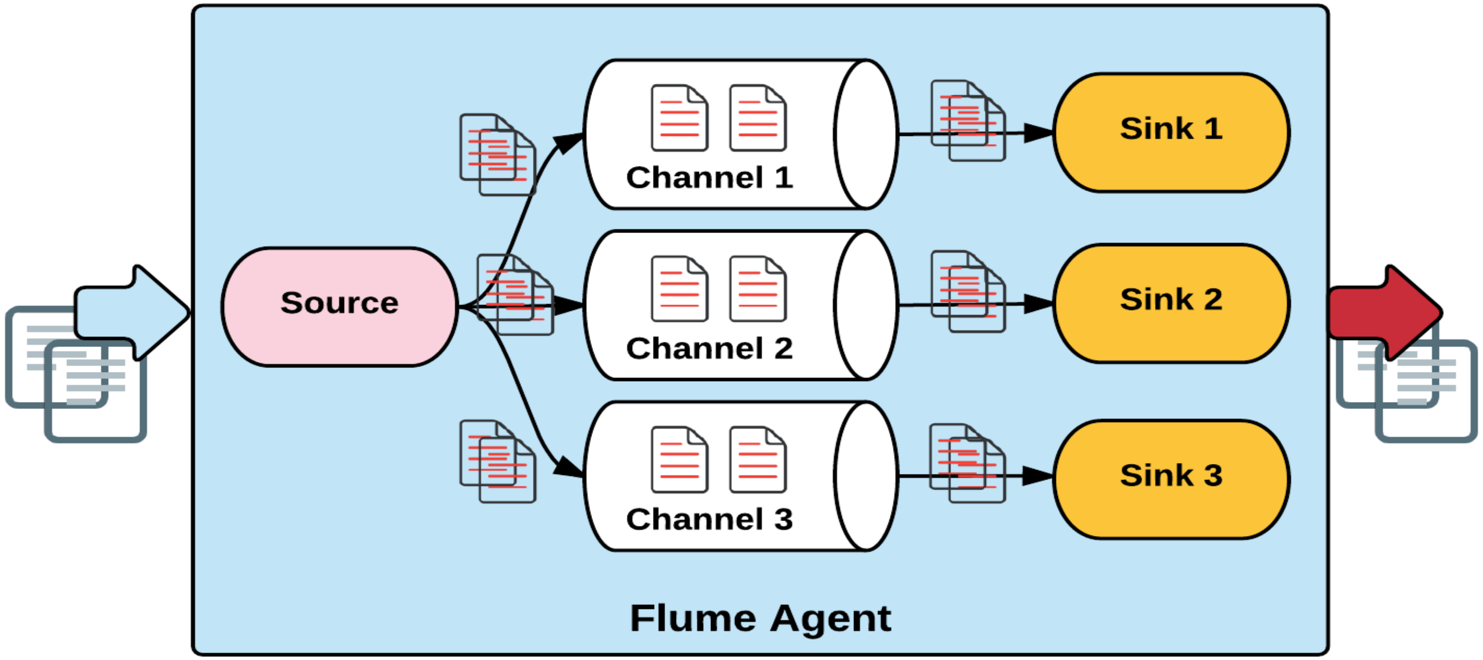
- Installation Process
- Flume using:
  - NetCat TCP source to HDFS sink

FANNING IN DESIGN



- In this design, Flume is used to pull data from several servers into one sink

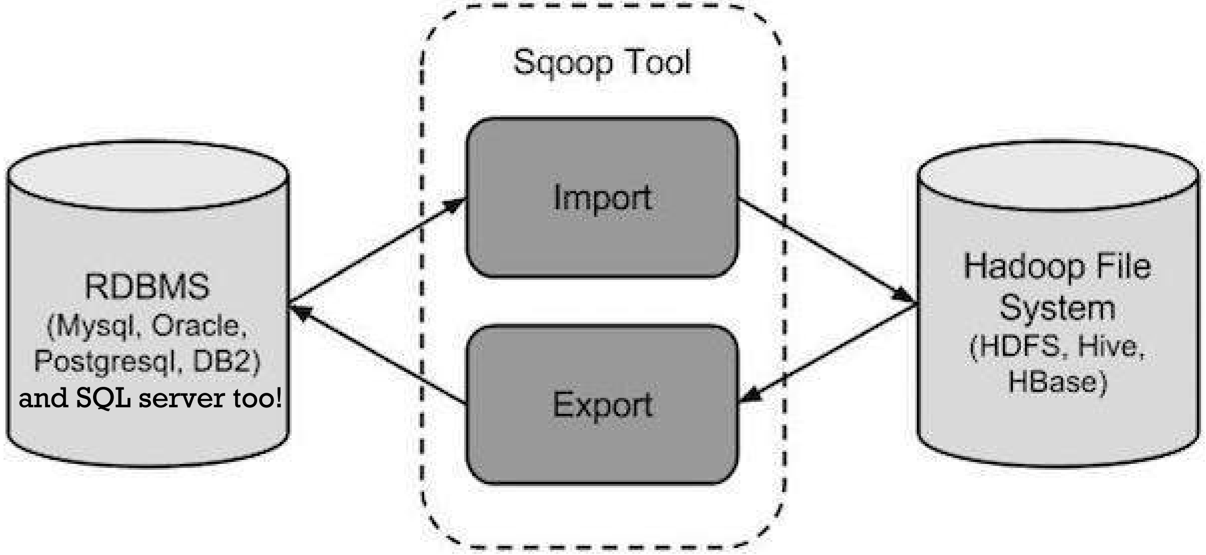
FANNING OUT DESIGN



- In this design, we can take one source but send it's data to many sinks

SQOOP

- Implementing Sqoop is a bit more straight-forward than working with Flume
- It has import and export tools as part of its driver that allow you to perform data ingress and egress operations
    - Use **import** and **export** commands to move data to and from Hadoop
- Sqoop comes with different connectors to "talk" to various data sources. We will use the **jdbc:sqlserver** connector



LAB TIME

- In this lab you will start to play with independently figuring out how to perform installation and testing tasks within the Hadoop ecosystem
- You will use the SQOOP driver in this assessment
- Let’s go over a few hints before you get started