# APAN PS5400: Managing Data

## Week 13: Hadoop-2

Lecturer: François Scharffe

COLUMBIA UNIVERSITY
School of Professional Studies

# Last week

- Introduction to Hadoop Ecosystem
- HDFS
- MapReduce
- Inverted Spectrum as a data model
- YARN

# This week

- Installing Cloudera Hadoop on Google Cloud
- Pig
- Hive
- Flume

# What is Hive?

- Hive is a data processing tool in Hadoop
- Data ingestion tools like Flume, Sqoop, etc., bring data into HDFS files, which may need to be processed.
- We can do data processing with MapReduce
  - But setting up MapReduce for a small data processing task is very tedious. Need to create mappers, reducers, driver code, etc.
  - Not worth doing it just to access a few lines of an HFDS file.
- Hive provides a SQL-like interface (HiveQL) for doing that task
  - A HiveQL instruction is automatically converted as a MapReduce job by its compiler.

# Hive vs RDBMS

- Hive instances do not store data.
  - Hive instances may store meta-data that have pointers (links) to the data
- Unlike RDBMS, Hive is not designed for OLTP (transaction processing). It is designed for OLAP (analytical processing).
  - Hive is suited for data warehouse applications.
- Hive supports queries on structured data
  - Very complex to query unstructured data using HiveQL.

# Hive Architecture

- A Hive query can be submitted to the Hive server using a Web UI, or programmatically using JDBC/ODBC, or Hive CLI (command line interface)
- The Hive server converts it into a MapReduce job.
- The MapReduce job is executed on a Hadoop cluster.

# Managed vs. External Tables

- A Hive database consists of managed tables and external tables.
  - The difference between the two consists in what happens in response to LOAD and DROP TABLE commands.
- Loading a managed table results in the data being moved to a Hive directory
- Dropping a managed table results in the schema as well as the data being deleted.
- Loading an external table results in the schema pointing to the HDFS files.
- Dropping an external table results in only the schema being deleted.

# What is Apache Pig?

- Pig, like Hive, is an abstraction over MapReduce
  - Enables analysts to process data in HDFS files without having to write Java programs for setting up MapReduce jobs.
- With Pig larger data analysis jobs can be performed than with Hive
- Pig programs are written in a high-level scripting language called Pig Latin
- Pig Latin is SQL-like. However, it is a data flow language.
  - A data flow programming language models a program as a directed graph of the data flowing between operations.
- Pig Latin programs are converted into MapReduce jobs by the Pig Engine.

# Apache Pig Applications

- Used for ad-hoc processing and quick proto-typing
- To process huge data sources (e.g., web logs)
- To process time sensitive data loads
  - This is enabled because data can be ingested without creating a schema (schema on read)

# Pig Latin script execution flow

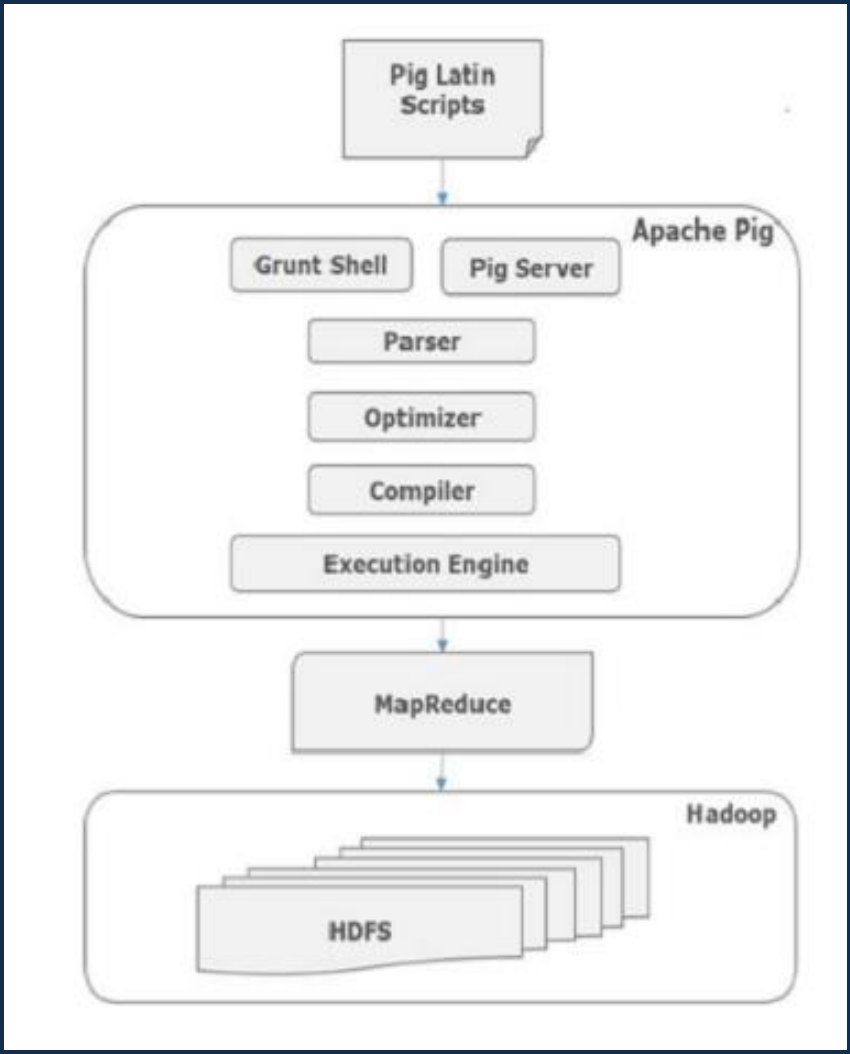



Image from TutorialsPoint.com

# What is Apache Flume?

Flume is a data ingestion tool for collecting and transporting large amounts of data from different sources into a centralized data store.

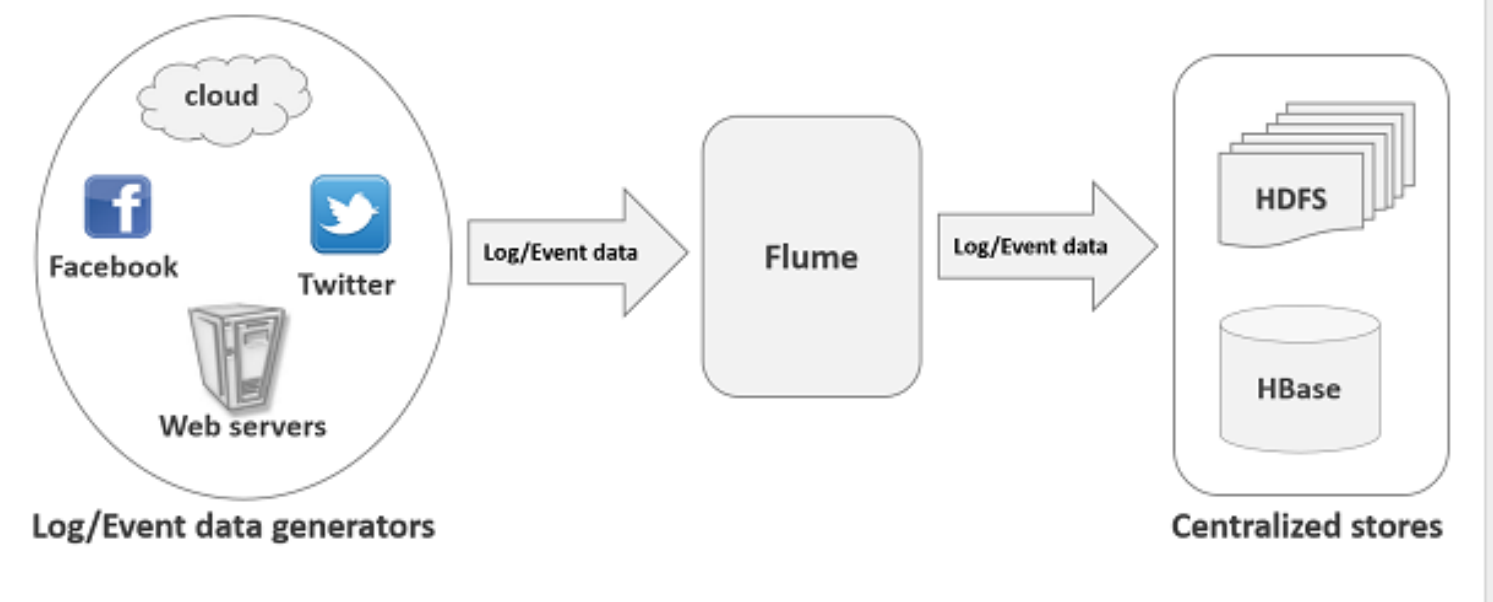



Image from TutorialsPoint.com

# Features of Flume

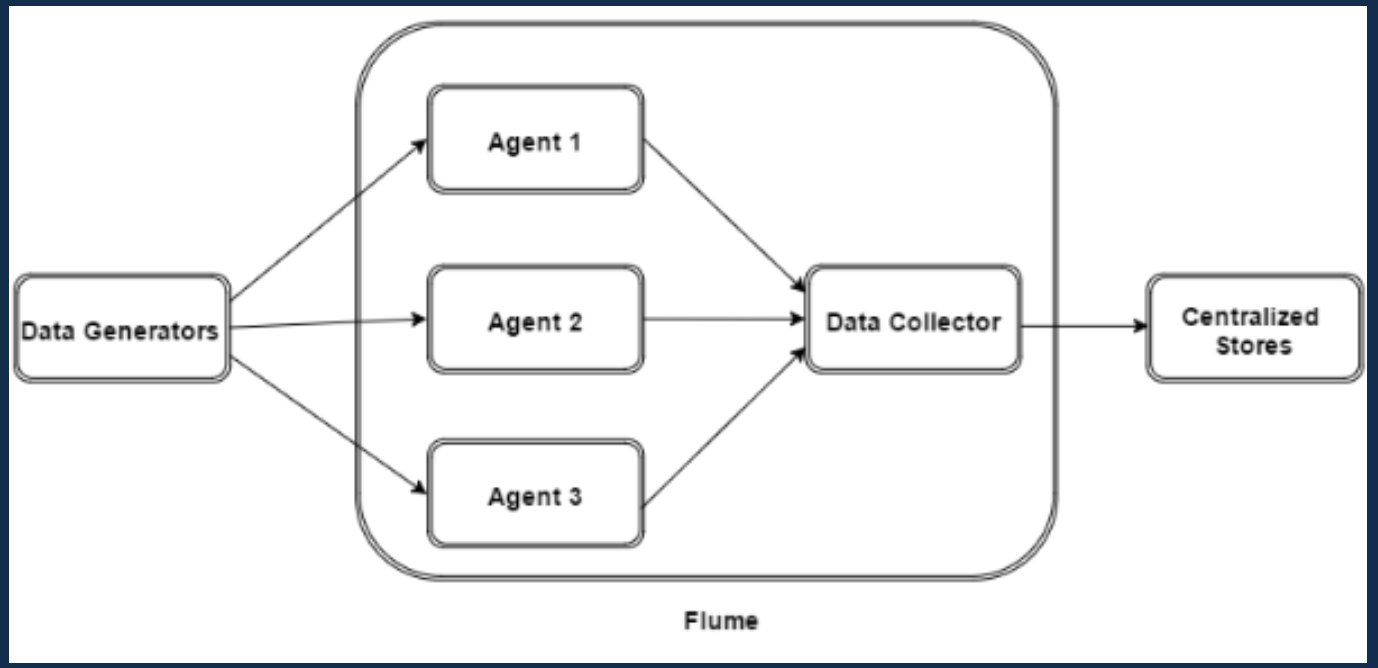
- Apache Flume can be used to store data in to any of the centralized stores (HBase, HDFS)
- Flume permits the data to be buffered when incoming rate exceeds write rate of the centralized store.
- Flume is reliable, fault tolerant, horizontally scalable, manageable, and customizable.
- Based on the concept of a channel connecting sender to receiver
- Flume can be used to get data (log files, event files) from multiple servers into centralized stores.

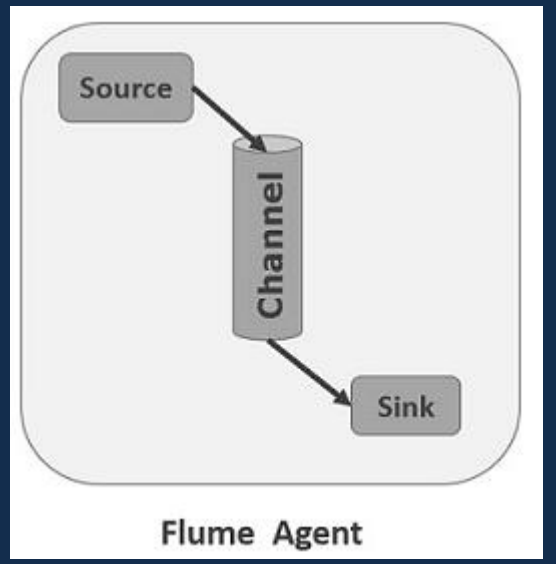
# Why a service like Flume is needed

- In Hadoop the traditional way of bringing data into HDFS is through the **put** command.
- The **put** command permits transfer of **only one file at a time** even if data generators generate data at a much higher rate.
- The **put** command service requires the data to be packaged and ready for upload. This is not possible if the data is generated continuously, as in the case of web servers.
- If the network is interrupted during the writing of a file using standard put-service, then the data written in the file could be lost (if the file was not closed before network interruption).
- So, we need a service that can overcome these problems. Flume comes to the rescue!

Data from generators like social media sites, etc., go to one or more agents. An agent contains a channel, which can provide transient storage (buffer), which is used to transmit data from the source (which can be another agent or a data generator) to the sink, which can be HDFS or HBase or another Agent.

A Data Collector is itself an Agent.




Images from TutorialsPoint.com

In the rest of this lecture we will go over Project 3.

# Recap of this week

- Pig
- Hive
- Flume
- Installing Cloudera Hadoop on Google Cloud
- Discussion of Project 3