Ingesting Data

HORTONWORKS UNIVERSITY

Lesson Objectives

After completing this lesson, students should be able to:

- Describe data ingestion
- Describe Batch/Bulk ingestion options
  - Ambari HDFS Files View
  - CLI & WebHDFS
  - NFS Gateway
  - Sqoop
- Describe streaming framework alternatives
  - Flume
  - Storm
  - Spark Streaming
  - HDF / NiFi

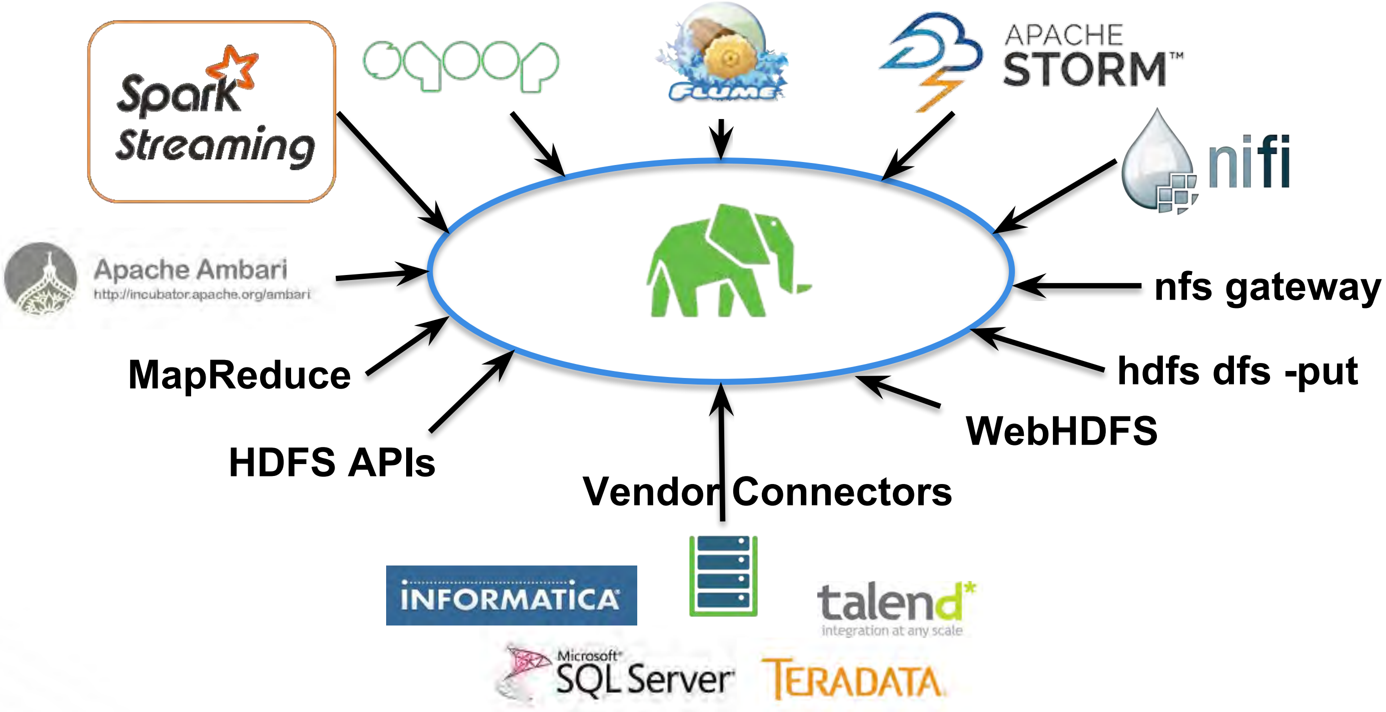
→ **Ingestion Overview**

Batch/Bulk Ingestion

Streaming Alternatives

HORTONWORKS UNIVERSITY

Data Input Options



Real-Time Versus Batch Ingestion Workflows

Real-time and batch processing are very different.

Factors		Real-Time	Batch
Data	**Age**	Real-time – usually less than 15 minutes old	Historical – usually more than 15 minutes old
	Location	Primarily in memory – moved to disk after processing	Primarily on disk – moved to memory for processing
Processing	**Speed**	Sub-second to few seconds	Few seconds to hours
	Frequency	Always running	Sporadic to periodic
Clients	**Who**	Automated systems only	Human & automated systems
	Type	Primarily operational applications	Primarily analytical applications

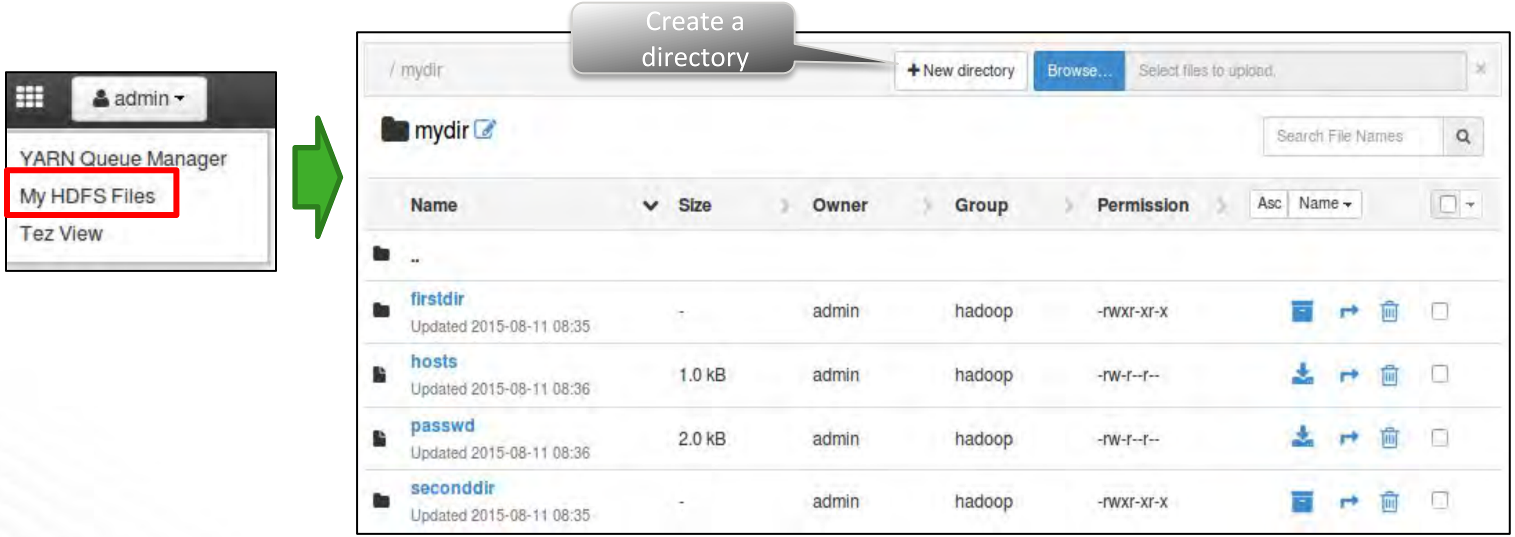
Ingestion Overview

→ **Batch/Bulk Ingestion**

Streaming Alternatives

Ambari Files View

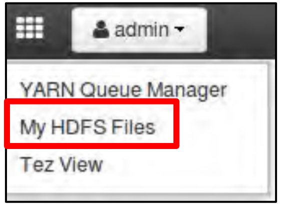
The Files View is an Ambari Web UI plug-in providing a graphical interface to HDFS.

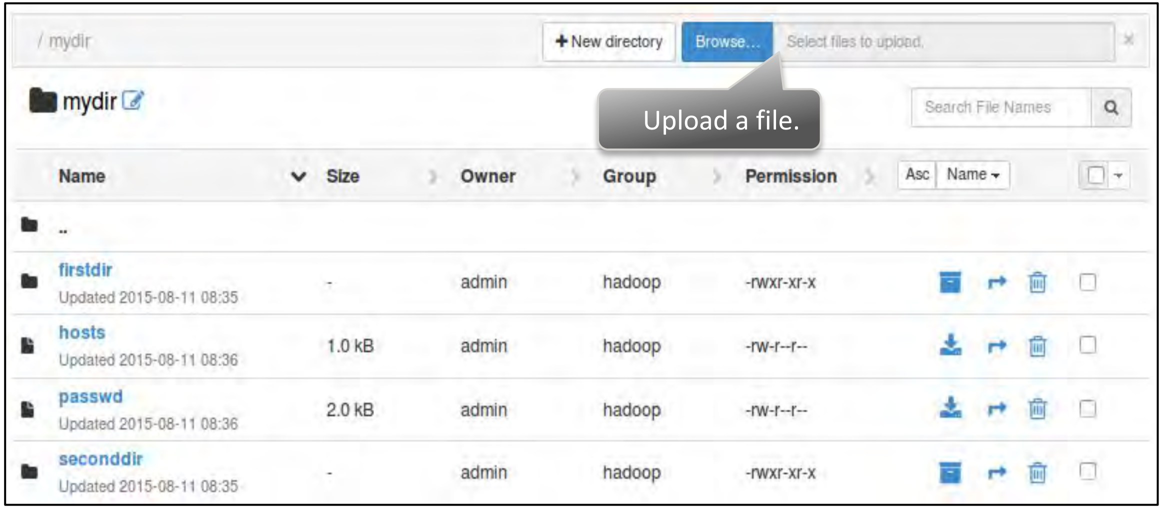


Name	Size	Owner	Group	Permission
..				
firstdir Updated 2015-08-11 08:35	-	admin	hadoop	-rwxr-xr-x
hosts Updated 2015-08-11 08:36	1.0 kB	admin	hadoop	-rw-r--r--
passwd Updated 2015-08-11 08:36	2.0 kB	admin	hadoop	-rw-r--r--
seconddir Updated 2015-08-11 08:35	-	admin	hadoop	-rwxr-xr-x

Ambari Files View

The Files View is an Ambari Web UI plug-in providing a graphical interface to HDFS.



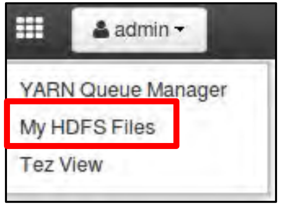


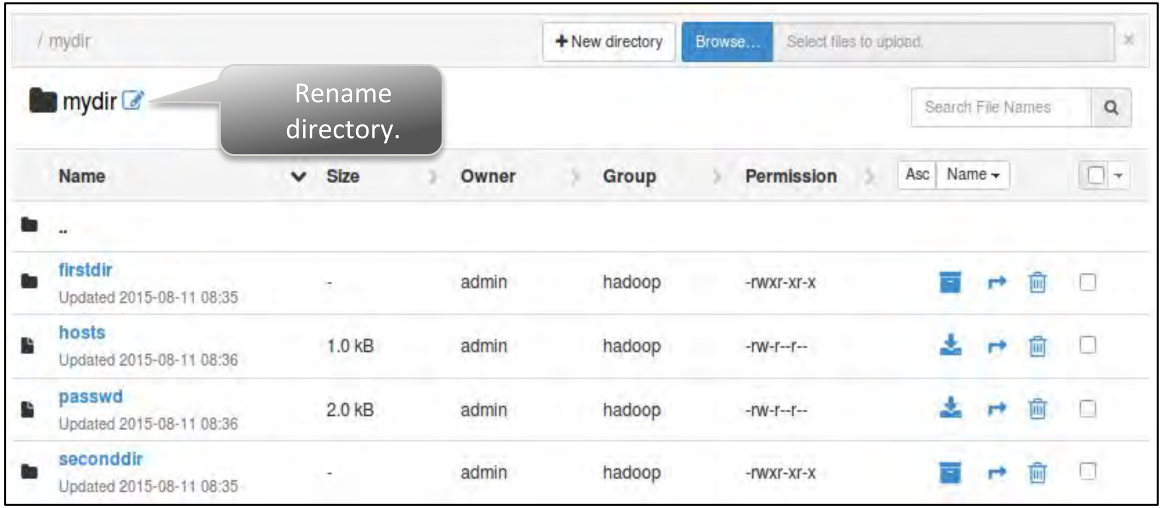
Upload a file.

Name	Size	Owner	Group	Permission
..				
firstdir Updated 2015-08-11 08:35	-	admin	hadoop	-rwxr-xr-x
hosts Updated 2015-08-11 08:36	1.0 kB	admin	hadoop	-rw-r--r--
passwd Updated 2015-08-11 08:36	2.0 kB	admin	hadoop	-rw-r--r--
seconddir Updated 2015-08-11 08:35	-	admin	hadoop	-rwxr-xr-x

Ambari Files View

The Files View is an Ambari Web UI plug-in providing a graphical interface to HDFS.



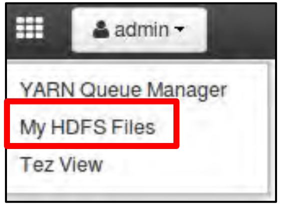


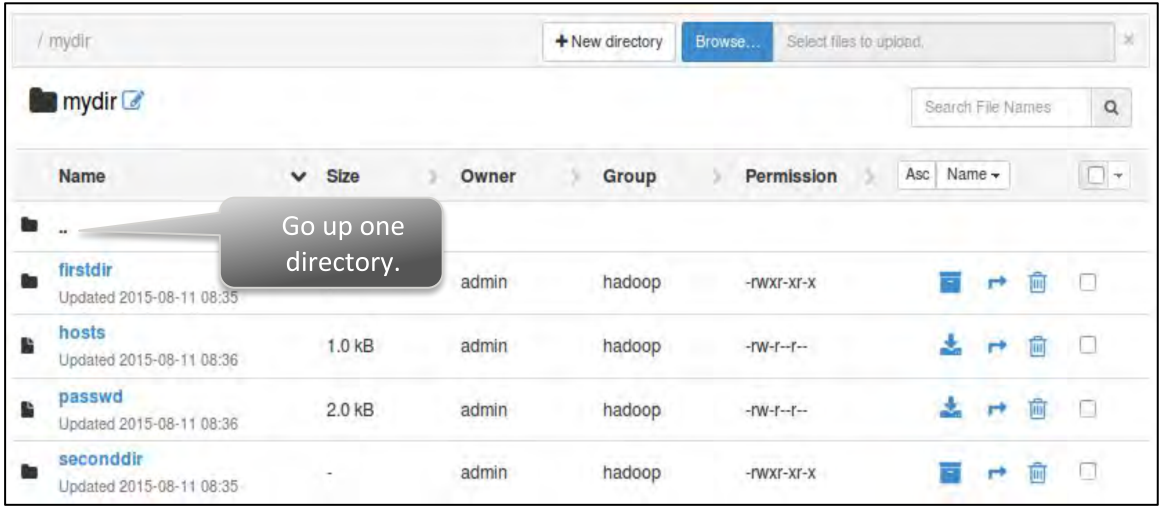
Rename directory.

Name	Size	Owner	Group	Permission
..				
firstdir Updated 2015-08-11 08:35	-	admin	hadoop	-rwxr-xr-x
hosts Updated 2015-08-11 08:36	1.0 kB	admin	hadoop	-rw-r--r--
passwd Updated 2015-08-11 08:36	2.0 kB	admin	hadoop	-rw-r--r--
seconddir Updated 2015-08-11 08:35	-	admin	hadoop	-rwxr-xr-x

Ambari Files View

The Files View is an Ambari Web UI plug-in providing a graphical interface to HDFS.



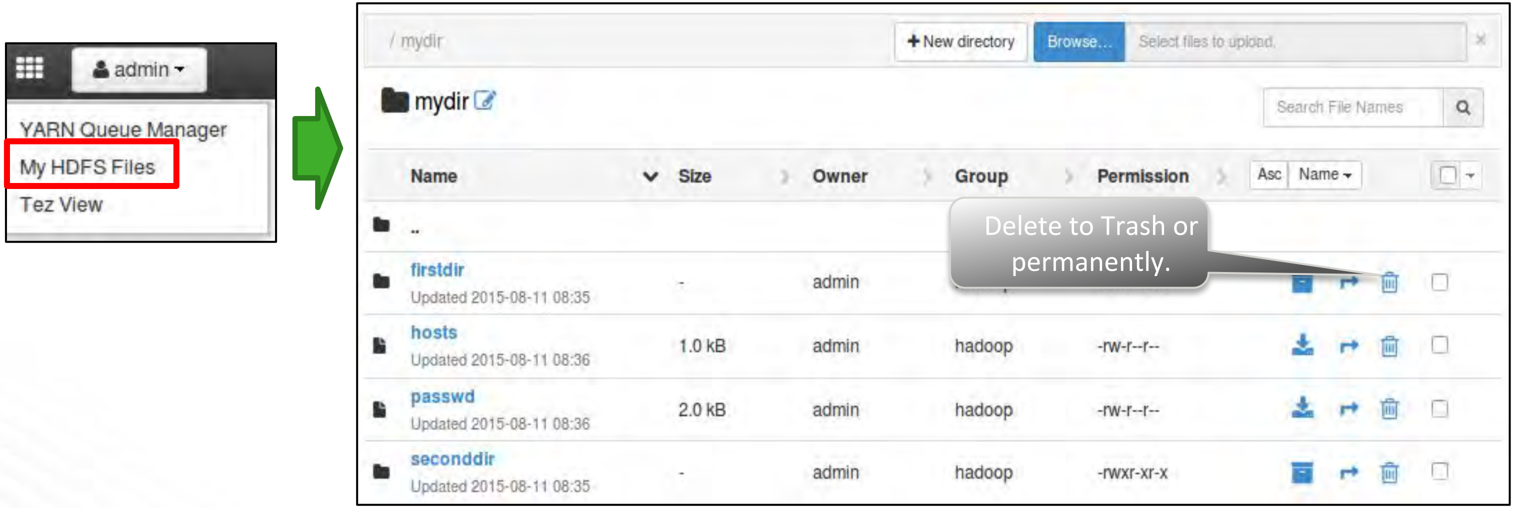


Name	Size	Owner	Group	Permission
..				
firstdir Updated 2015-08-11 08:35		admin	hadoop	-rwxr-xr-x
hosts Updated 2015-08-11 08:36	1.0 kB	admin	hadoop	-rw-r--r--
passwd Updated 2015-08-11 08:36	2.0 kB	admin	hadoop	-rw-r--r--
seconddir Updated 2015-08-11 08:35	-	admin	hadoop	-rwxr-xr-x

Go up one directory.

Ambari Files View

The Files View is an Ambari Web UI plug-in providing a graphical interface to HDFS.

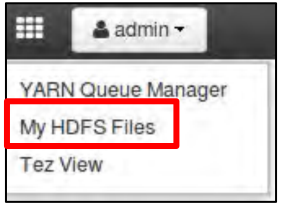


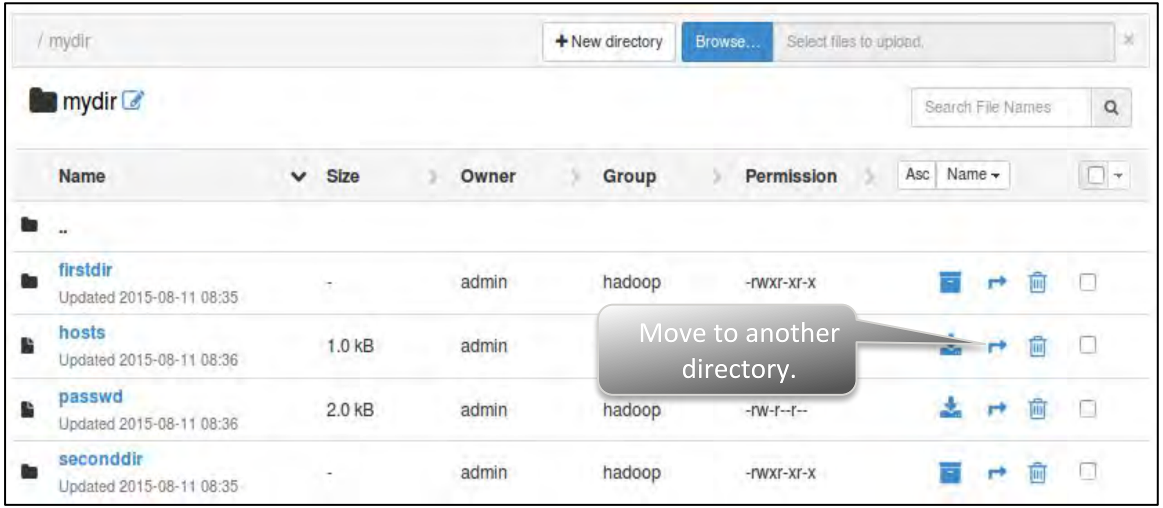
Delete to Trash or permanently.

Name	Size	Owner	Group	Permission
..				
firstdir Updated 2015-08-11 08:35	-	admin		
hosts Updated 2015-08-11 08:36	1.0 kB	admin	hadoop	-rw-r--r--
passwd Updated 2015-08-11 08:36	2.0 kB	admin	hadoop	-rw-r--r--
seconddir Updated 2015-08-11 08:35	-	admin	hadoop	-rwxr-xr-x

Ambari Files View

The Files View is an Ambari Web UI plug-in providing a graphical interface to HDFS.



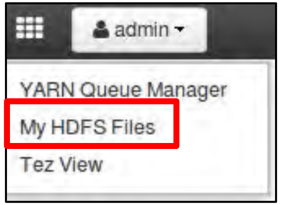


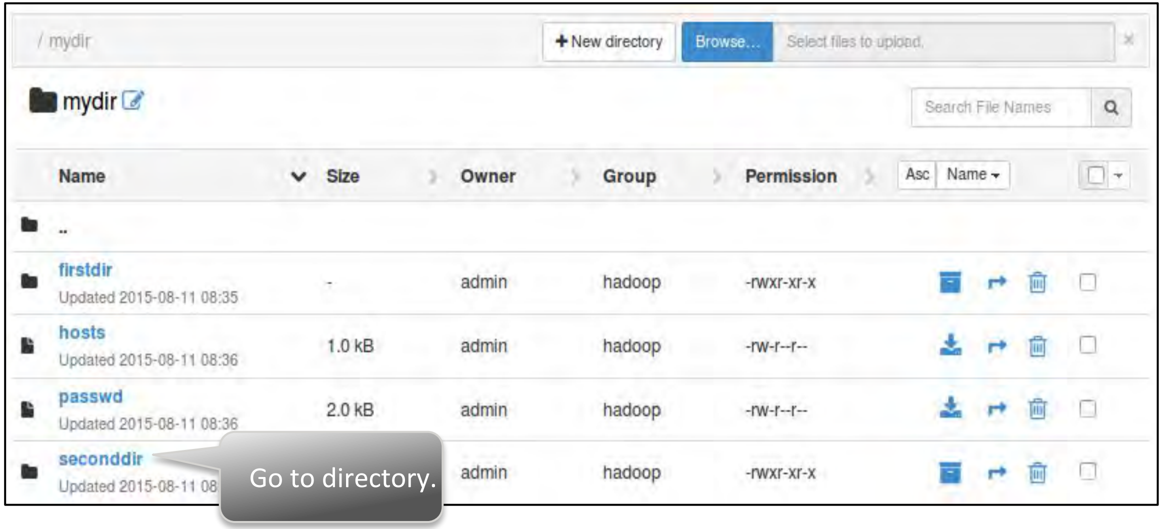
Name	Size	Owner	Group	Permission
..				
firstdir Updated 2015-08-11 08:35	-	admin	hadoop	-rwxr-xr-x
hosts Updated 2015-08-11 08:36	1.0 kB	admin		
passwd Updated 2015-08-11 08:36	2.0 kB	admin	hadoop	-rw-r--r--
seconddir Updated 2015-08-11 08:35	-	admin	hadoop	-rwxr-xr-x

Move to another directory.

Ambari Files View

The Files View is an Ambari Web UI plug-in providing a graphical interface to HDFS.



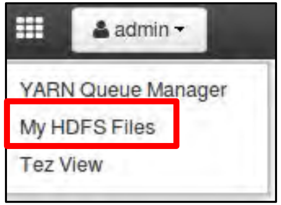


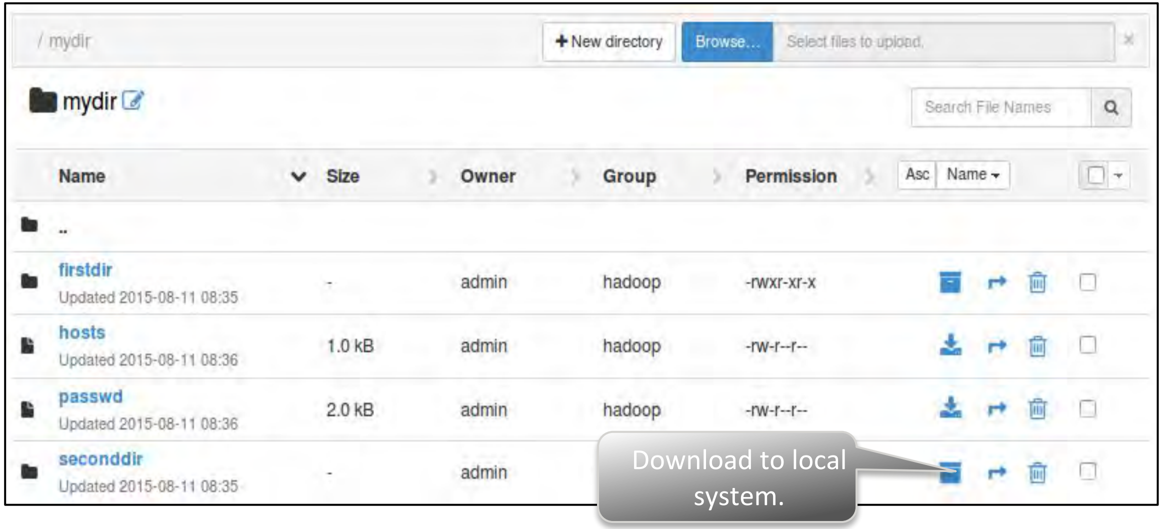
Name	Size	Owner	Group	Permission
..				
firstdir Updated 2015-08-11 08:35	-	admin	hadoop	-rwxr-xr-x
hosts Updated 2015-08-11 08:36	1.0 kB	admin	hadoop	-rw-r--r--
passwd Updated 2015-08-11 08:36	2.0 kB	admin	hadoop	-rw-r--r--
seconddir Updated 2015-08-11 08		admin	hadoop	-rwxr-xr-x

Go to directory.

Ambari Files View

The Files View is an Ambari Web UI plug-in providing a graphical interface to HDFS.





Name	Size	Owner	Group	Permission
..				
firstdir Updated 2015-08-11 08:35	-	admin	hadoop	-rwxr-xr-x
hosts Updated 2015-08-11 08:36	1.0 kB	admin	hadoop	-rw-r--r--
passwd Updated 2015-08-11 08:36	2.0 kB	admin	hadoop	-rw-r--r--
seconddir Updated 2015-08-11 08:35	-	admin		

Download to local system.

The Hadoop Client

- The `put` command to uploading data to HDFS
- Perfect for inputting local files into HDFS
- Useful in batch scripts
- Usage:

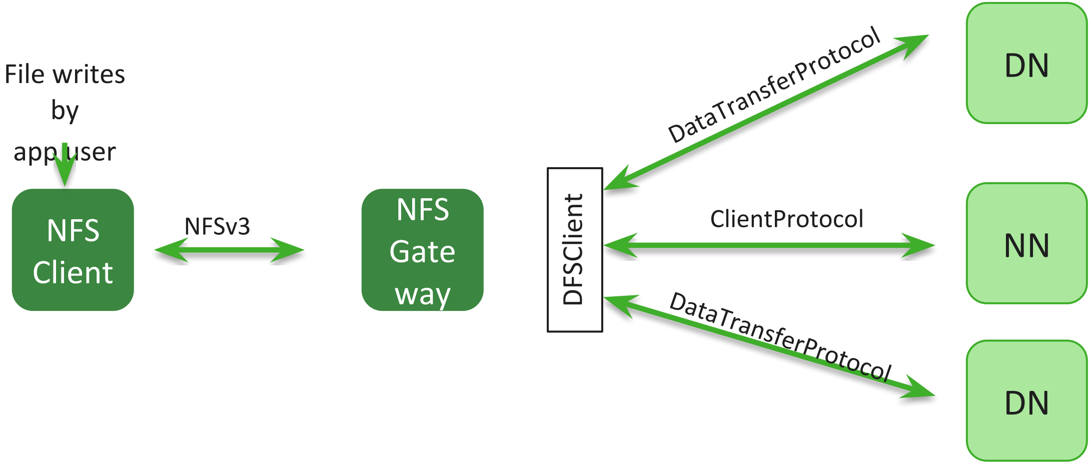
```
hdfs dfs -put mylocalfile /some/hdfs/path
```

WebHDFS

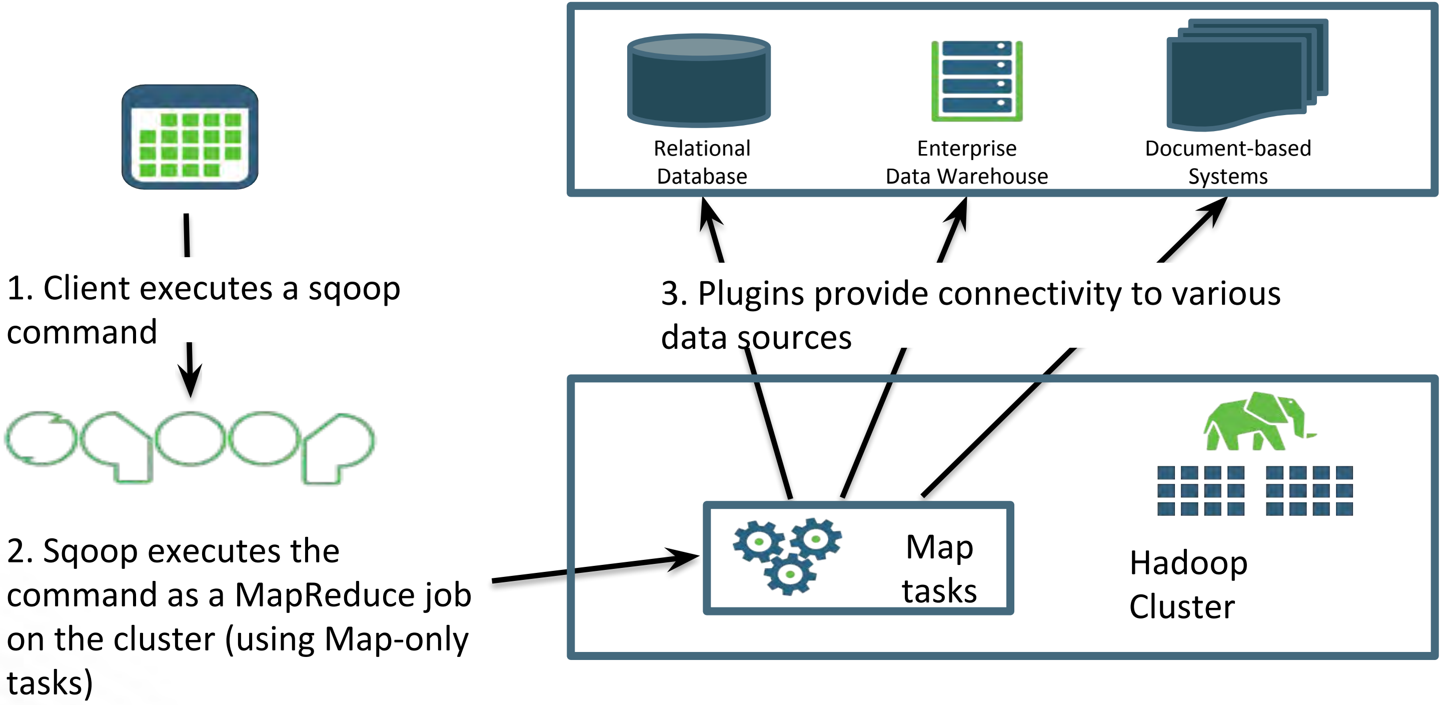
- REST API for accessing all of the HDFS file system interfaces:
  - `http://host:port/webhdfs/v1/test/mydata.txt?op=OPEN`
  - `http://host:port/webhdfs/v1/user/train/data?op=MKDIRS`
  - `http://host:port/webhdfs/v1/test/mydata.txt?op=APPEND`

NFS Gateway

- Uses NFS standard and supports all HDFS commands
- No random writes



Sqoop: Database Import/Export



1. Client executes a sqoop command

2. Sqoop executes the command as a MapReduce job on the cluster (using Map-only tasks)

3. Plugins provide connectivity to various data sources

The Sqoop Import Tool

The **import** command has the following requirements:

- Must specify a connect string using the **--connect** argument
- Credentials can be included in the connect string, so use the **--username** and **--password** arguments
- Must specify either a table to import using **--table** or the result of an SQL query using **--query**

Importing a Table

```
sqoop import
--connect jdbc:mysql://host/nyse
--table StockPrices
--target-dir /data/stockprice/
--as-textfile
```

Importing Specific Columns

```
sqoop import
--connect jdbc:mysql://host/nyse
--table StockPrices
--columns StockSymbol,Volume, High,ClosingPrice
--target-dir /data/dailyhighs/
--as-textfile
--split-by StockSymbol
-m 10
```

Importing from a Query

```
sqoop import
--connect jdbc:mysql://host/nyse
--query "SELECT * FROM StockPrices s
WHERE s.Volume >= 1000000
AND \$CONDITIONS"
--target-dir /data/highvolume/
--as-textfile
--split-by StockSymbol
```

The Sqoop Export Tool

- The export command transfers data from HDFS to a database:
  - Use **--table** to specify the database table
  - Use **--export-dir** to specify the data to export
- Rows are appended to the table by default
- If you define **--update-key**, existing rows will be updated with the new data
- Use **--call** to invoke a stored procedure (instead of specifying the **--table** argument)

Exporting to a Table

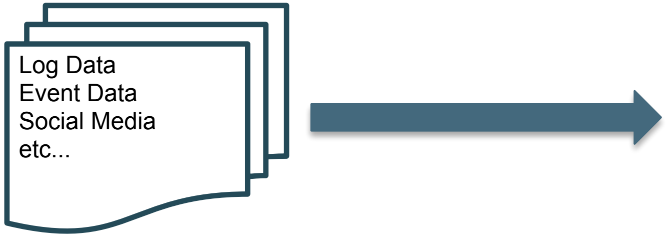
```
sqoop export
--connect jdbc:mysql://host/mylogs
--table LogData
--export-dir /data/logfiles/
--input-fields-terminated-by "\t"
```

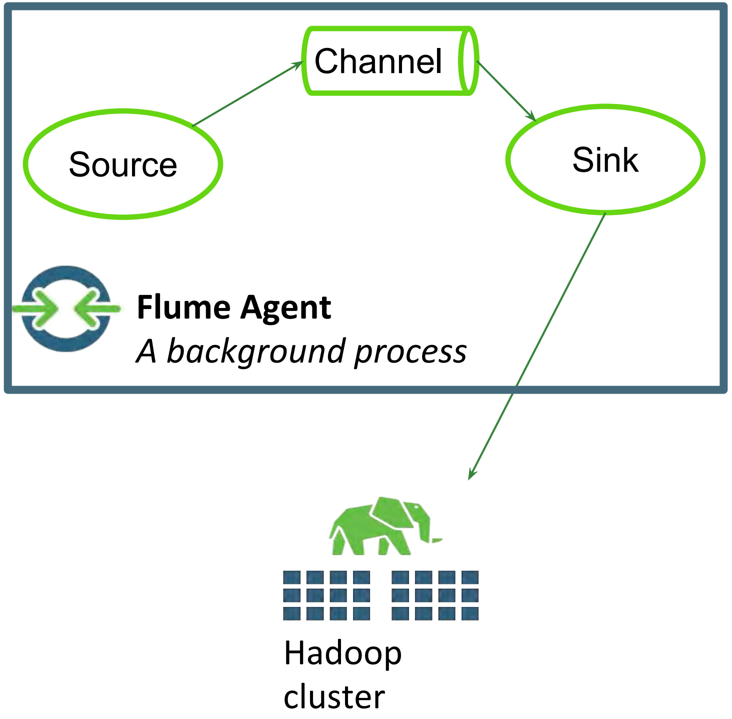
Ingestion Overview

Batch/Bulk Ingestion

Streaming Alternatives

Flume: Data Streaming

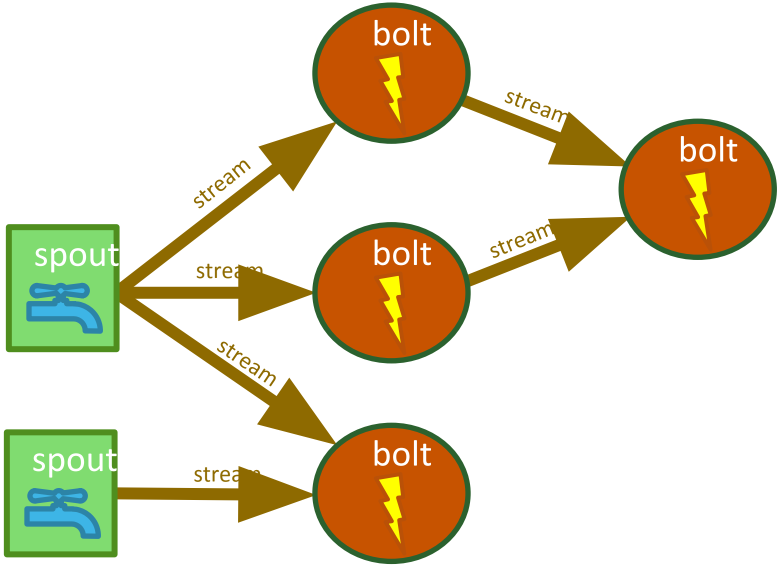




Flume uses a **Channel** between the **Source** and **Sink** to decouple the processing of **events** from the storing of events.

Storm Topology Overview

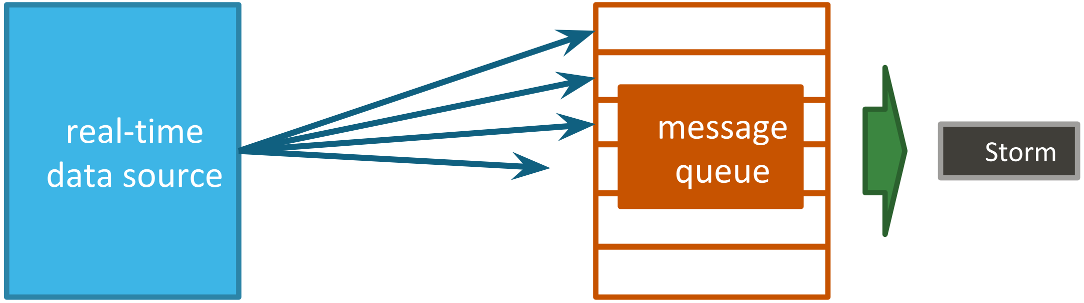
- Storm data processing occurs in a topology.
- A topology consists of spout and bolt components.
- Spouts bring data into the topology
- Bolts can (not required) persist data including to HDFS



Storm topology

Message Queues

Various types of message queues are often the source of the data processed by real-time processing engines like Storm



operating systems, services and applications, sensors

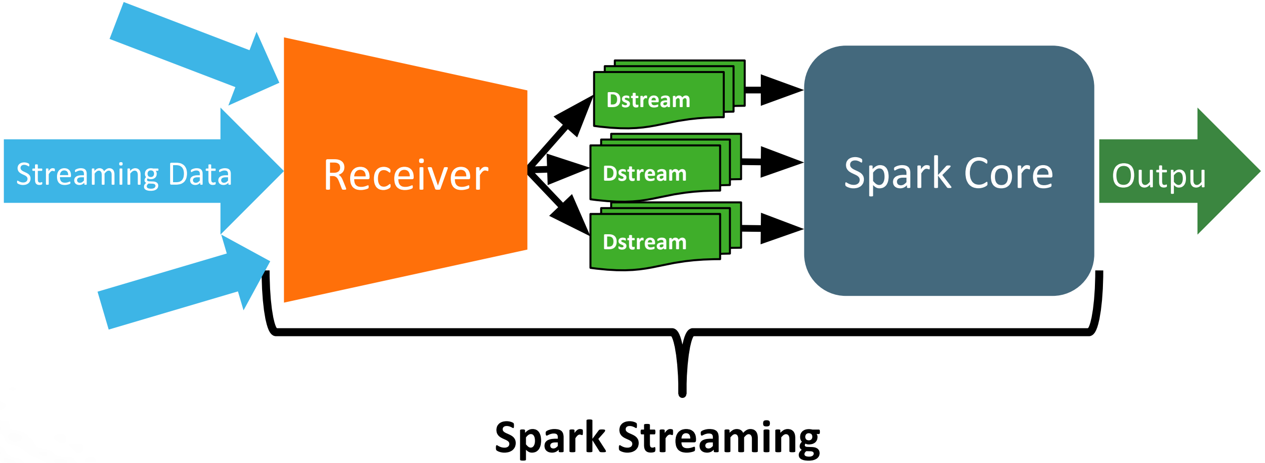
log entries, events, errors, status messages, etc.

Kestrel, RabbitMQ, AMQP, Kafka, JMS, others…

data from queue is read by Storm

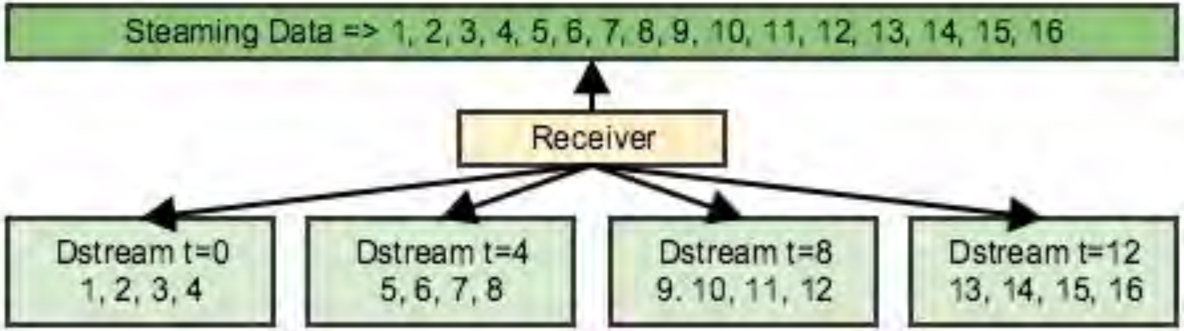
Spark Streaming

- Streaming Applications consist of the same components as a Core application, but add the concept of a receiver
- The receiver is a process running on an executor

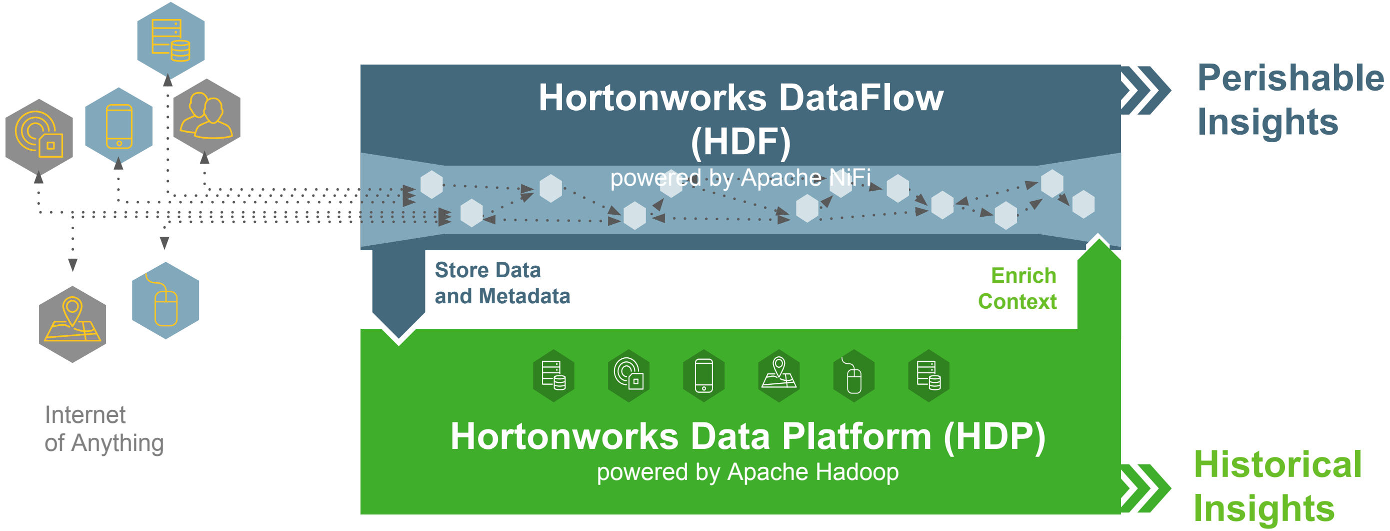


Spark Streaming’s Micro-Batch Approach

- Micro-batches are created at regular time intervals
  - Receiver takes the data and starts filling up a batch
  - After the batch duration completes, data is shipped off
  - Each batch forms a collection of data entities that are processed together



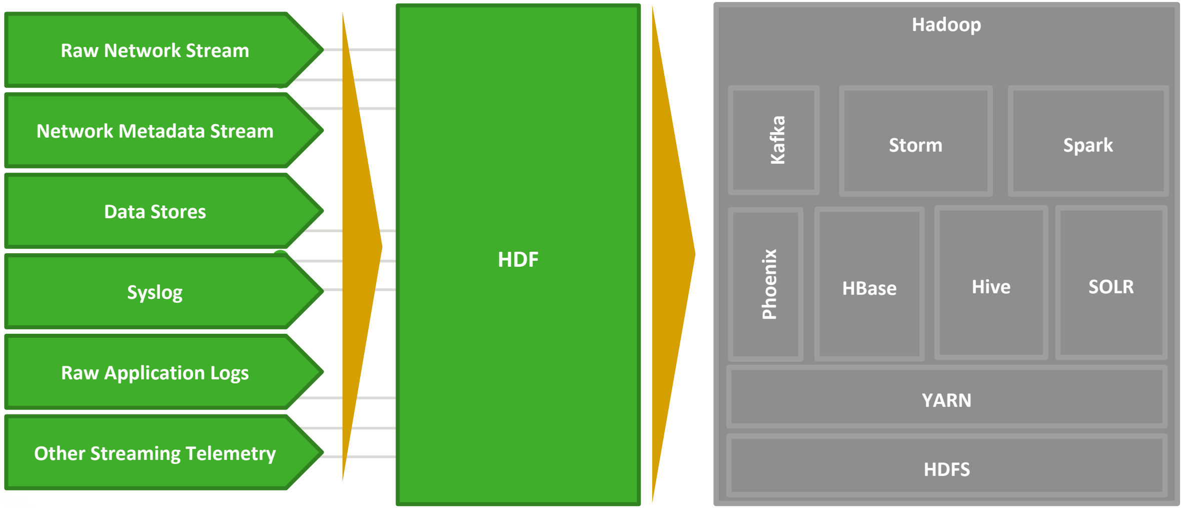
HDF with HDP – A Complete Big Data Solution



Hortonworks DataFlow and the Hortonworks Data Platform deliver the industry’s most complete Big Data solution

Big Data Ingestion with HDF

HDF workflows and Storm/Spark streaming workflows can be coupled



Knowledge Check

Questions

1. What tool is used for importing data from a RDBMS?

Questions

1. What tool is used for importing data from a RDBMS?
2. List two ways to easily script moving files into HDFS.

Questions

1. What tool is used for importing data from a RDBMS?
2. List two ways to easily script moving files into HDFS.
3. True/False? Storm operates on micro-batches.

Questions

1. What tool is used for importing data from a RDBMS?
2. List two ways to easily script moving files into HDFS.
3. True/False? Storm operates on micro-batches.
4. Name the popular open-source messaging component that is bundled with HDP.

Summary

Summary

- There are many different ways to ingest data including customer solutions written via HDFS APIs as well as vendor connectors
- Streaming and batch workflows can work together in a holistic system
- The NFS Gateway may help some legacy systems populate data into HDFS
- Sqoop’s configurable number of database connection can overload an RDBMS
- The following are streaming frameworks:
  - Flume
  - Storm
  - Spark Streaming
  - HDF / NiFi