Performance Measurements - round 2

Setup:

Here are some sample measurements taken with a **single agent.**

Cluster Config: 20-node Hadoop cluster (1 name node and 19 data nodes).

Machine Config: 24 cores – Xeon E5-2640 v2 @ 2.00GHz, 164 GB RAM, 7200 rpm Hard Drive.

1. File channel with HDFS Sink (Sequence File):

Flume version: 1.4

Source: 4 x Exec Source, 100k batchSize

HDFS Sink Batch size: 500,000

Event Size: 500 byte events.

Channel: File

	Events/Sec					
Sinks	**1 data dirs**	**2 data dirs**	**4 data dirs**	**6 data dirs**	**8 data dirs**	**10 data dirs**
1	14.3k (7 MB/s)					
2	21.9k					
4		35.8k				
8			72.5k	77k	**78.6 (37 MB/s)**	76.6k
10			58k			
12			49.3	49k		

Measurements were taken to get an idea around the configuration that yields best performance. So took measurements only for all data points in the grid that made sense. For example it was not necessary to take measurements for multiple dataDirs at single sink, as it was evident multiple HDFS sink would better than single sink config.

2. HDFS Sink:

Flume version: 1.4

Channel: Memory

Event Size: 500 byte events.

# of HDFS Sinks	**Snappy BatchSz:1.2mill**	**Snappy BatchSz:1.4mill**	**Sequence File BatchSz:1.2mill**
1	34.3 k (17 MB/s)	33 k	33 k
2	71 k	75 k	69 k
4	141 k	145 k	141 k
8	271 k	273 k	251 k
12	382 k	380 k	370 k
16	478 k	**538 k (240 MB/s)**	**486 k (232 MB/s)**

Some simple observations:

- increasing number of dataDirs helps FC perf even on single disk systems
- Increasing number of sinks helps

3. Hive Sink:

Flume version: 1.5 & 1.6

Channel: Memory

BatchSz:1million

Event Size: 500 byte events.

	Flume 1.5		Flume 1.6	
	Event/s	**MBps**	**Event/s**	**MBps**
	1 Sink			
DELIMITED Text	36,885	18	138,461	66
JSON	12,735	6		
	16 sinks (agent maxed out)			
DELIMITED Text	209,600	100	348,214	166
JSON	25,751	12	31,135	14

Observation: Feeding JSON data to Hive sink is much slower, potentially due to higher parsing overhead of JSON in part.

4. HBase Sink:

Flume version: 1.5

Channel: Memory

Serializer: RegexHbaseEventSerializer

Total Sinks: 1

Event Size (bytes)	**BatchSz: 1**	**BatchSz: 100**	**BatchSz: 1000**	**BatchSz: 10,000**
500		11 MB/s		11 MB/s
1000	0.5 MB/s	14 MB/s	22 MB/s	27 MB/s

5. ASync HBase Sink:

Flume version: 1.5

Channel: Memory

Serializer: SimpleAsyncHbaseEventSerializer

Total Sinks: 1

Event Size (bytes)	BatchSz: 1	BatchSz: 100	BatchSz: 1000
500		0.4 MB/s	0.5 MB/s
1000	0.8 MB/s	0.8 MB/s	0.9 MB/s

6. Kafka Source:

Flume version: 1.6

Channel: Memory

Sink: Null Sink

Event Size: 1000 bytes

Total Sinks: 1

Batch Size (bytes)	MB/s
1,000	62
10,000	112
20,000	125
40,000	147
80,000	153