Government Big Data Platform

EGA

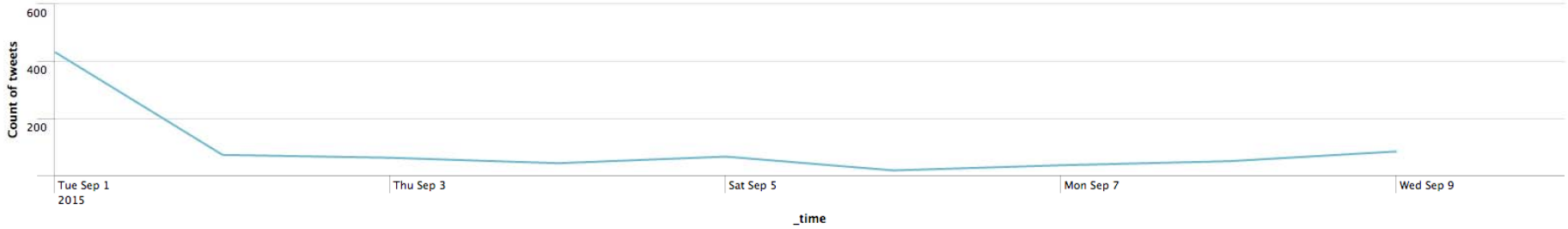
Big Data is not just about handing large datasets.

“It also present a new paradigm shift in way data is managed and analyzed.”

Topics ...

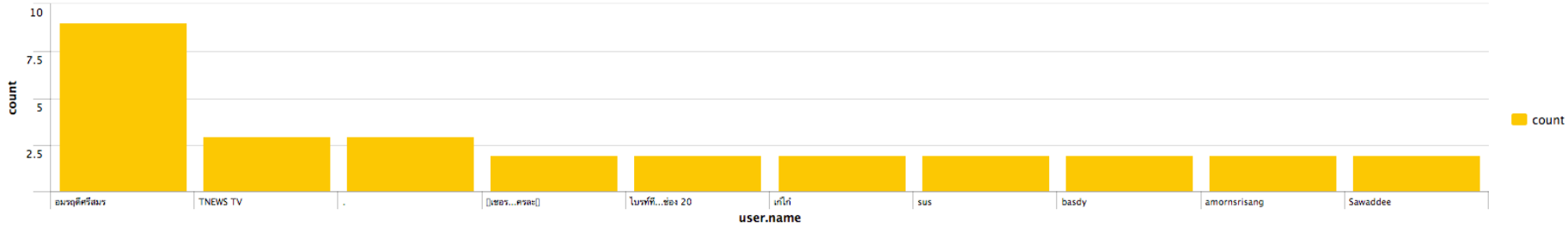
- Traditional BI vs. Big Data
- Big Data Platform: Data Management
- Big Data Platform: Analytics
- Architecture
- Big Data as a Service

Tweets along with Time



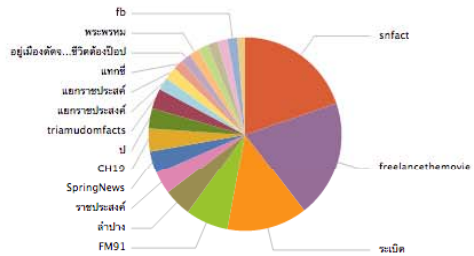
Loading - 0%

Top 10 User



Loading - 0%

Top 20 Retweeted Hashtags



Loading - 0%

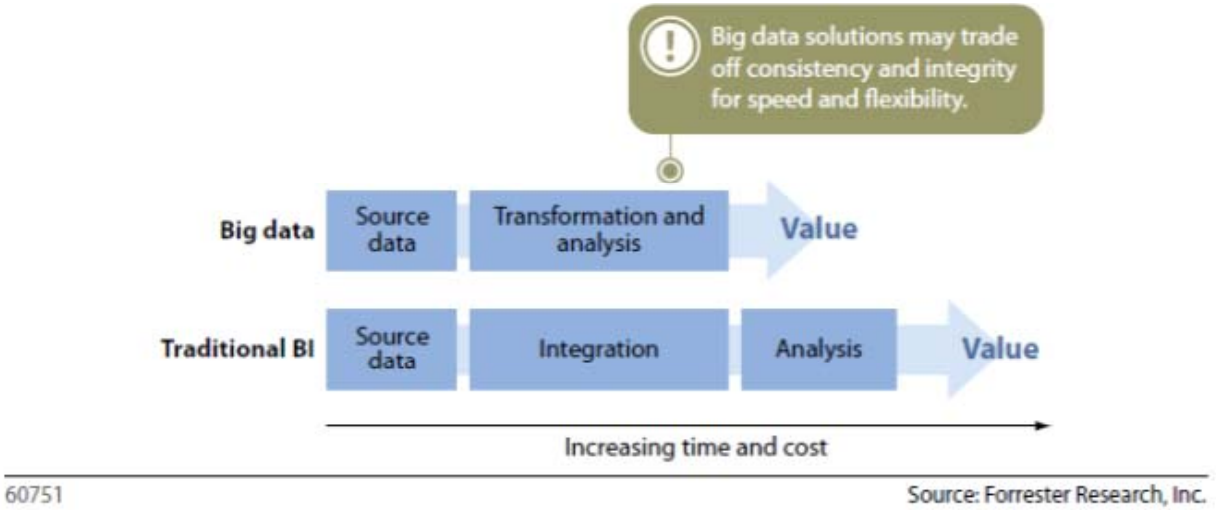
Retweeted Hashtags

retweeted_status.entities.hashtags{}.text	count	percent
snfact	47	37.301587
freelancethemovie	47	37.301587
ระเบิด	40	31.746032
ลำปาง	18	14.285714
FM91	17	13.492063
SpringNews	16	12.698413
CH19	16	12.698413
ราชประสงค์	10	7.936508
ป	9	7.142857
triamudomfacts	9	7.142857

« prev 1 2 next »

Loading - 0%

Traditional BI vs. Big Data [1]



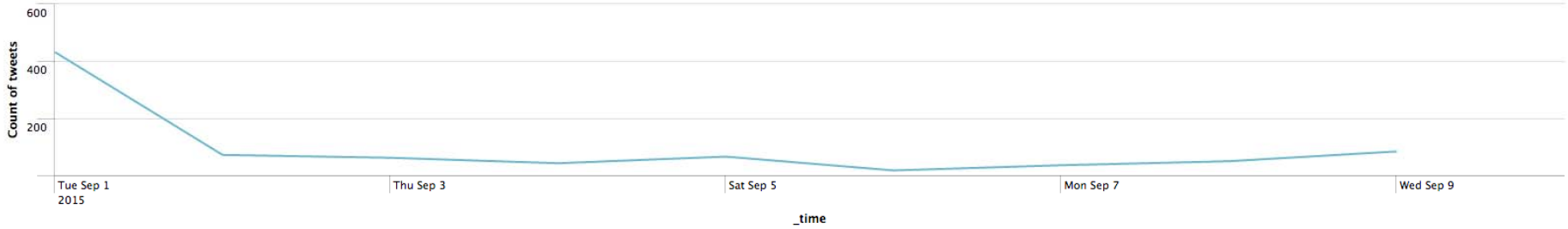
Traditional BI	Big Data
Performed on transformed data from raw sources – Loss of valuable information.	Work with raw data – Yield valuable information that might not been discovered before.
Do not handle unstructured data.	Enable effective analysis of unstructured data.
Consistency and Integrity	Speed and Flexibility

[1] B. Bhagattjee, "Emergence and Taxonomy of Big Data as a Service," Working Paper, Cambridge, 2014

Demo: Reduce a transformation and directly work on raw data

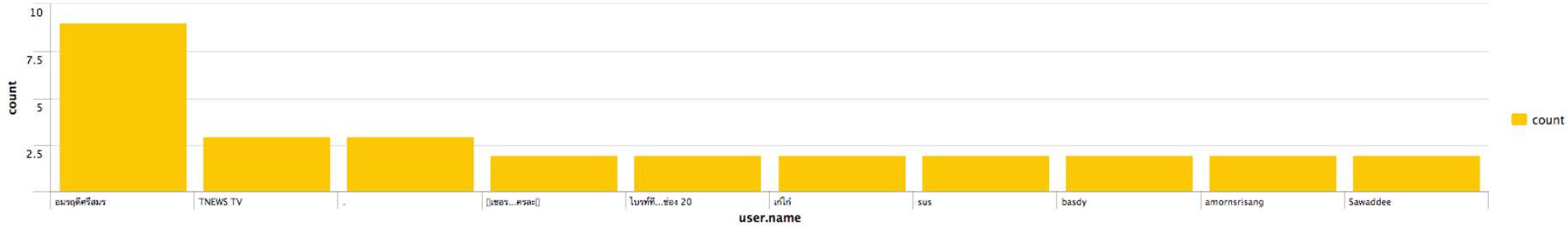
Twitters – Recent Bombing

Tweets along with Time



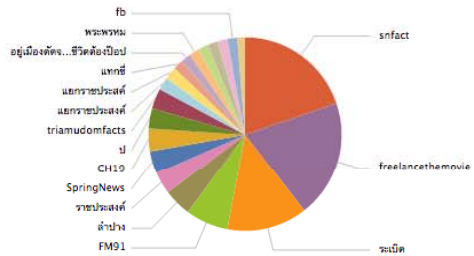
Loading - 0%

Top 10 User



Loading - 0%

Top 20 Retweeted Hashtags



Loading - 0%

Retweeted Hashtags

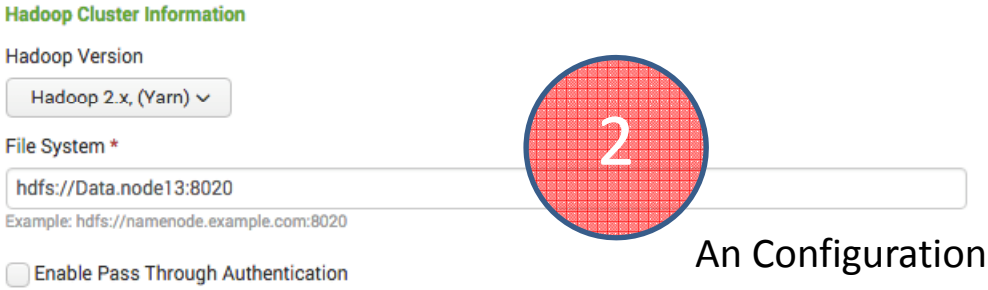
retweeted_status.entities.hashtags{}.text	count	percent
snfact	47	37.301587
freelancethemovie	47	37.301587
ระเบิด	40	31.746032
ลำปาง	18	14.285714
FM91	17	13.492063
SpringNews	16	12.698413
CH19	16	12.698413
ราชประสงค์	10	7.936508
ป	9	7.142857
triamudomfacts	9	7.142857

« prev 1 2 next »

Loading - 0%

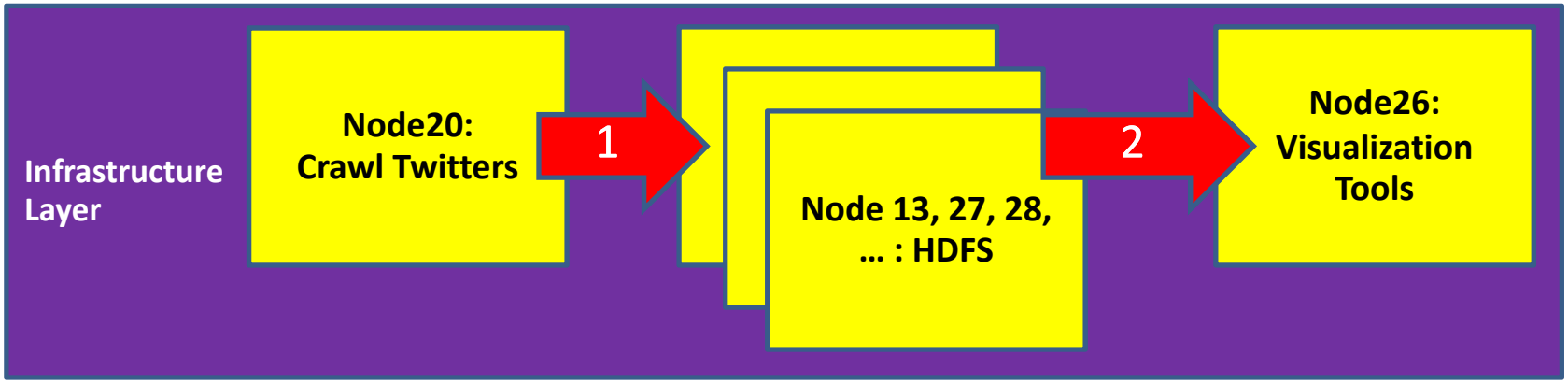
```
[root@Data flume-ng]# tail -f flume-cmf-flume-AGENT-Data.node20.log
2015-09-10 09:47:08,393 INFO org.apache.flume.sink.hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSystem = false
2015-09-10 09:47:08,425 INFO org.apache.flume.sink.hdfs.BucketWriter: Creating hdfs://Data.node13:8020/user/flume/eventtweets/15/09/10/09/47/05/FlumeData.1441853228394
.tmp
2015-09-10 09:47:38,470 INFO org.apache.flume.sink.hdfs.BucketWriter: Closing hdfs://Data.node13:8020/user/flume/eventtweets/15/09/10/09/47/05/FlumeData.1441853228394.
tmp
2015-09-10 09:47:38,484 INFO org.apache.flume.sink.hdfs.BucketWriter: Renaming hdfs://Data.node13:8020/user/flume/eventtweets/15/09/10/09/47/05/FlumeData.1441853228394
.tmp to hdfs://Data.node13:8020/user/flume/eventtweets/15/09/10/09/47/05/FlumeData.1441853228394
2015-09-10 09:47:38,485 INFO org.apache.flume.sink.hdfs.HDFSEventSink: Writer callback called.
2015-09-10 10:25:33,633 INFO org.apache.flume.sink.hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSystem = false
2015-09-10 10:25:33,664 INFO org.apache.flume.sink.hdfs.BucketWriter: Creating hdfs://Data.node13:8020/user/flume/eventtweets/15/09/10/10/25/31/FlumeData.1441855533634
.tmp
2015-09-10 10:26:03,709 INFO org.apache.flume.sink.hdfs.BucketWriter: Closing hdfs://Data.node13:8020/user/flume/eventtweets/15/09/10/10/25/31/FlumeData.1441855533634.
tmp
2015-09-10 10:26:03,720 INFO org.apache.flume.sink.hdfs.BucketWriter: Renaming hdfs://Data.node13:8020/user/flume/eventtweets/15/09/10/10/25/31/FlumeData.1441855533634
.tmp to hdfs://Data.node13:8020/user/flume/eventtweets/15/09/10/10/25/31/FlumeData.1441855533634
2015-09-10 10:26:03,722 INFO org.apache.flume.sink.hdfs.HDFSEventSink: Writer callback called.
```

1

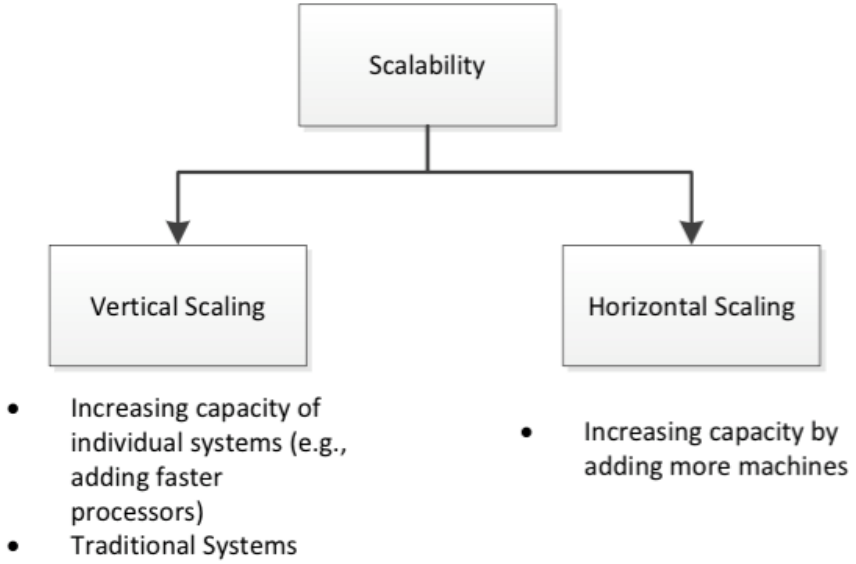


2

An Configuration of Visualization Tool



Data Management: Scalability [1]

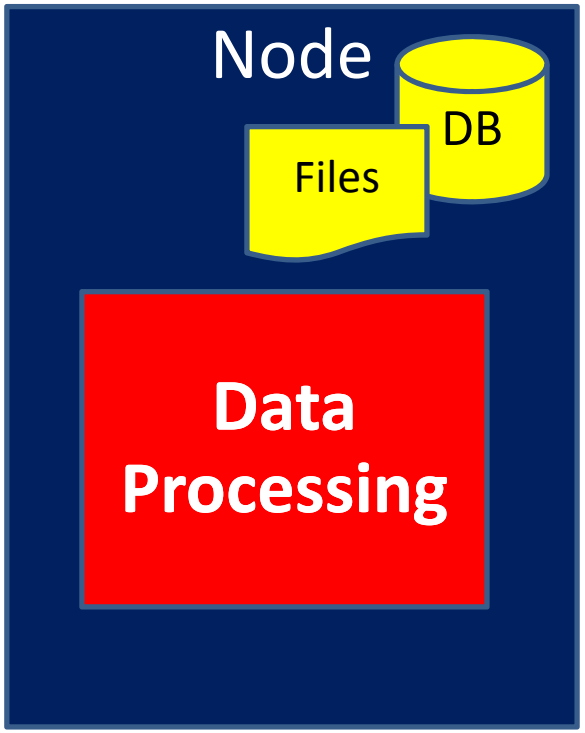


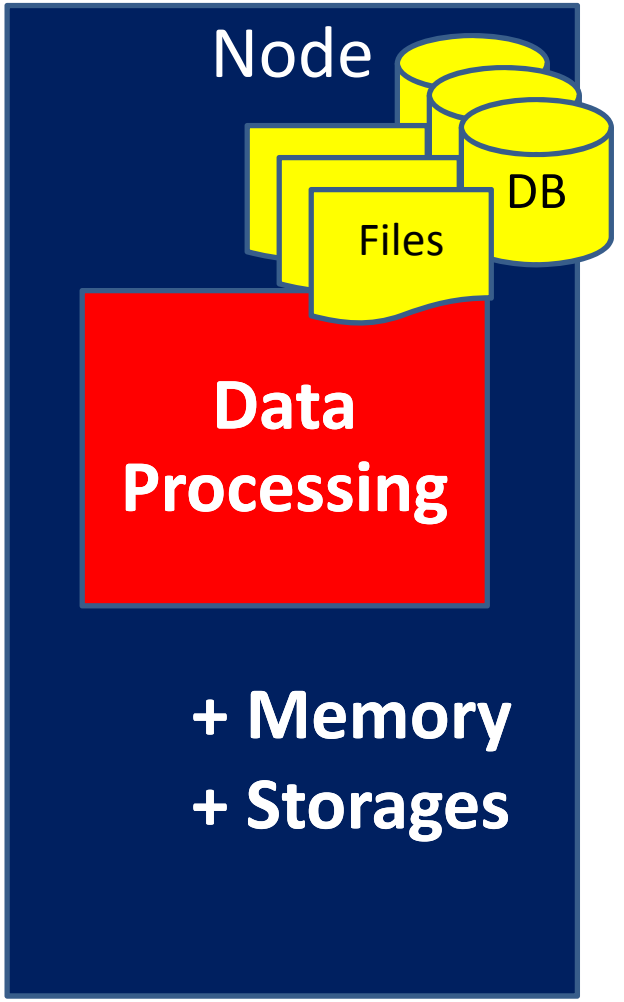
- Distributed computing architecture.
- Horizontally scalable to handle large datasets.
- For example, using Hadoop, users can add more ‘nodes’ or computer systems to the existing architecture to scale up.

[1] B. Bhagattjee, “Emergence and Taxonomy of Big Data as a Service,” Working Paper, Cambridge, 2014

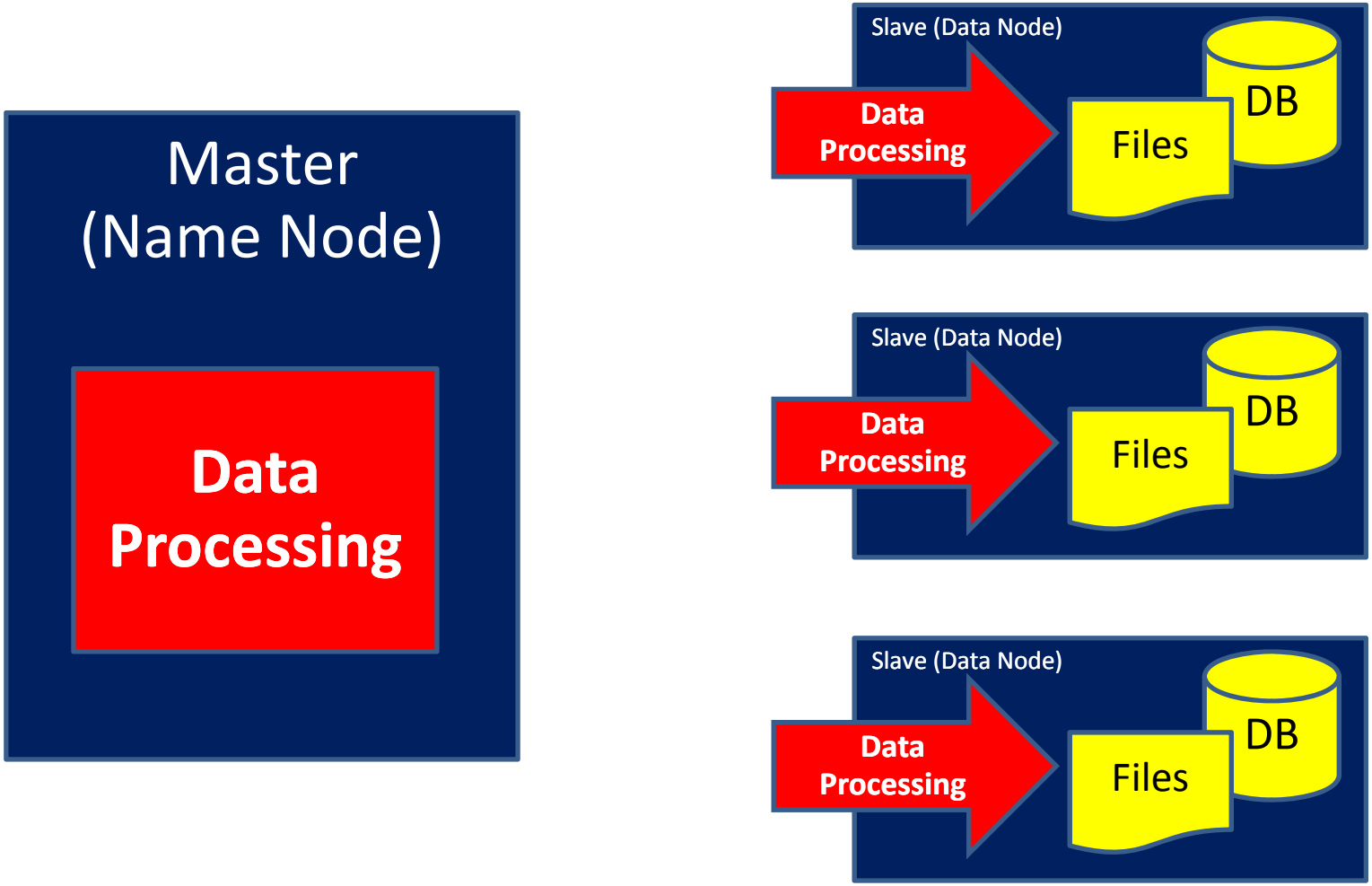
G-Cloud
Government Cloud Service

General Computing





Distributed Computing



Data Management: MapReduce [1]

- An architectural model for parallel processing of tasks on a distributed computing.
- Allows splitting of a single computation task to multiple nodes.
  - Number of nodes determines the processing power of the system.
- For starting a job, like copying codes and scheduling, is another problem that prevents it from executing interactive jobs and near real-time queries. [2]
- An implementation of MapReduce is the Apache Hadoop

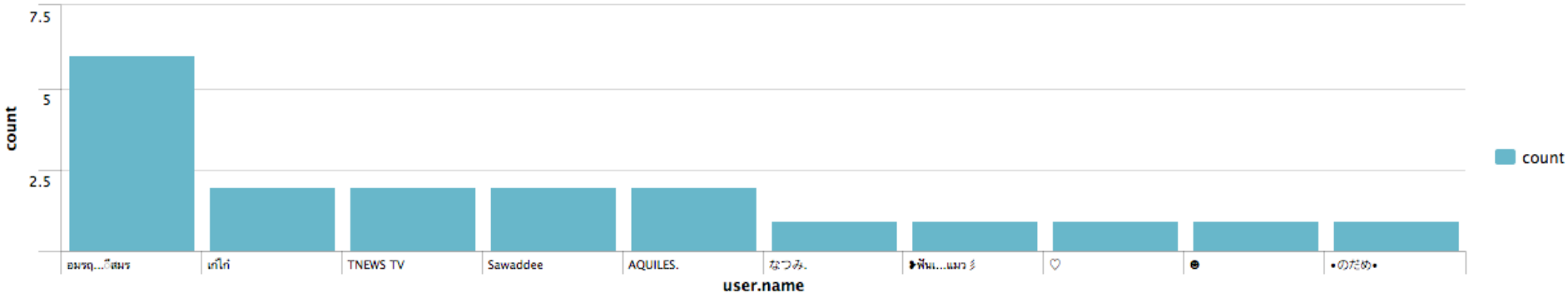
[1] B. Bhagattjee, “Emergence and Taxonomy of Big Data as a Service,” Working Paper, Cambridge, 2014

[2] S. Shahrivari, “Beyond Batch Processing: Towards Real-Time and Streaming Big Data,” Computers 2014, 3, 117-129.

Demo: Move processing to data

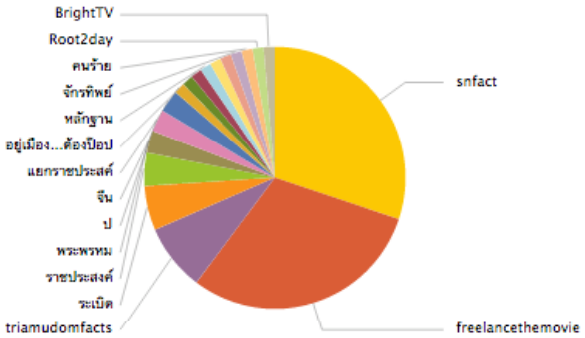
& MapReduce Jobs

Top 10 User



Loading - 0%

Top 20 Retweeted Hashtags



Loading - 0%

Retweeted Hashtags

retweeted_status.entities.hashtags{}.text	count	percent
snfact	30	60.000000
freelancethemovie	30	60.000000
ระเบิด	8	16.000000
triamudomfacts	7	14.000000
ราชประสงค์	5	10.000000
พระพรหม	4	8.000000
ป	4	8.000000
จีน	4	8.000000
แยกราชประสงค์	3	6.000000
แทกซี่	2	4.000000

« prev 1 2 next »

Loading - 0%

MapReduce Job job_1441553098625_0011

Logged in as: dr.who

- Cluster
- Application
- Job
  - Overview
  - Counters
  - Configuration
  - Map tasks
  - Reduce tasks
  - AM Logs
- Tools

Job Overview	
Job Name:	SPLK_Data.node26_user1__user1__search__search5_1441852537.2000_0
State:	RUNNING
Uberized:	false
Started:	Thu Sep 10 09:37:28 ICT 2015
Elapsed:	3mins, 35sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Thu Sep 10 09:37:22 ICT 2015	Data.node28:8042	logs

Task Type	Progress	Total	Pending	Running	Complete
Map		10000	9893	10	97
Reduce		0	0	0	0

Attempt Type	New	Running	Failed	Killed	Successful
Maps	9893	10	0	0	97
Reduces	0	0	0	0	0

Logged in as: dr.who

Map Tasks for job_1441553098625_0011

- Cluster
- Application
- Job
  - Overview
  - Counters
  - Configuration
  - Map tasks
  - Reduce tasks
  - AM Logs
- Tools

Show 20 entries Search:

Task	Progress	Status	State	Start Time	Finish Time	Elapsed Time
task_1441553098625_0011_m_000000		map	SUCCEEDED	Thu Sep 10 09:37:37 +0700 2015	Thu Sep 10 09:38:14 +0700 2015	37sec
task_1441553098625_0011_m_000123		map	SUCCEEDED	Thu Sep 10 09:41:27 +0700 2015	Thu Sep 10 09:41:44 +0700 2015	17sec
task_1441553098625_0011_m_000002		map	SUCCEEDED	Thu Sep 10 09:37:37 +0700 2015	Thu Sep 10 09:38:21 +0700 2015	44sec
task_1441553098625_0011_m_000003		map	SUCCEEDED	Thu Sep 10 09:37:37 +0700 2015	Thu Sep 10 09:38:06 +0700 2015	29sec

Show 20 entries Search:

Attempt	Progress	State	Status	Node	Logs	Started	Finished	Elapsed	Note
attempt_1441553098625_0011_m_000000_0	100.00	SUCCEEDED	map	Data.node16:8042	logs	Thu Sep 10 09:37:37 +0700 2015	Thu Sep 10 09:38:14 +0700 2015	37sec	Container killed by the ApplicationMaster. Container killed on request. Exit code is 143 Container exited with a non-zero exit code 143

Showing 1 to 1 of 1 entries First Previous 1 Next Last

Show 20 entries Search:

Attempt	Progress	State	Status	Node	Logs	Started	Finished	Elapsed	Note
attempt_1441553098625_0011_m_000123_0	100.00	SUCCEEDED	map	Data.node28:8042	logs	Thu Sep 10 09:41:27 +0700 2015	Thu Sep 10 09:41:44 +0700 2015	17sec	Container killed by the ApplicationMaster. Container killed on request. Exit code is 143 Container exited with a non-zero exit code 143

Showing 1 to 1 of 1 entries First Previous 1 Next Last

Show 20 entries Search:

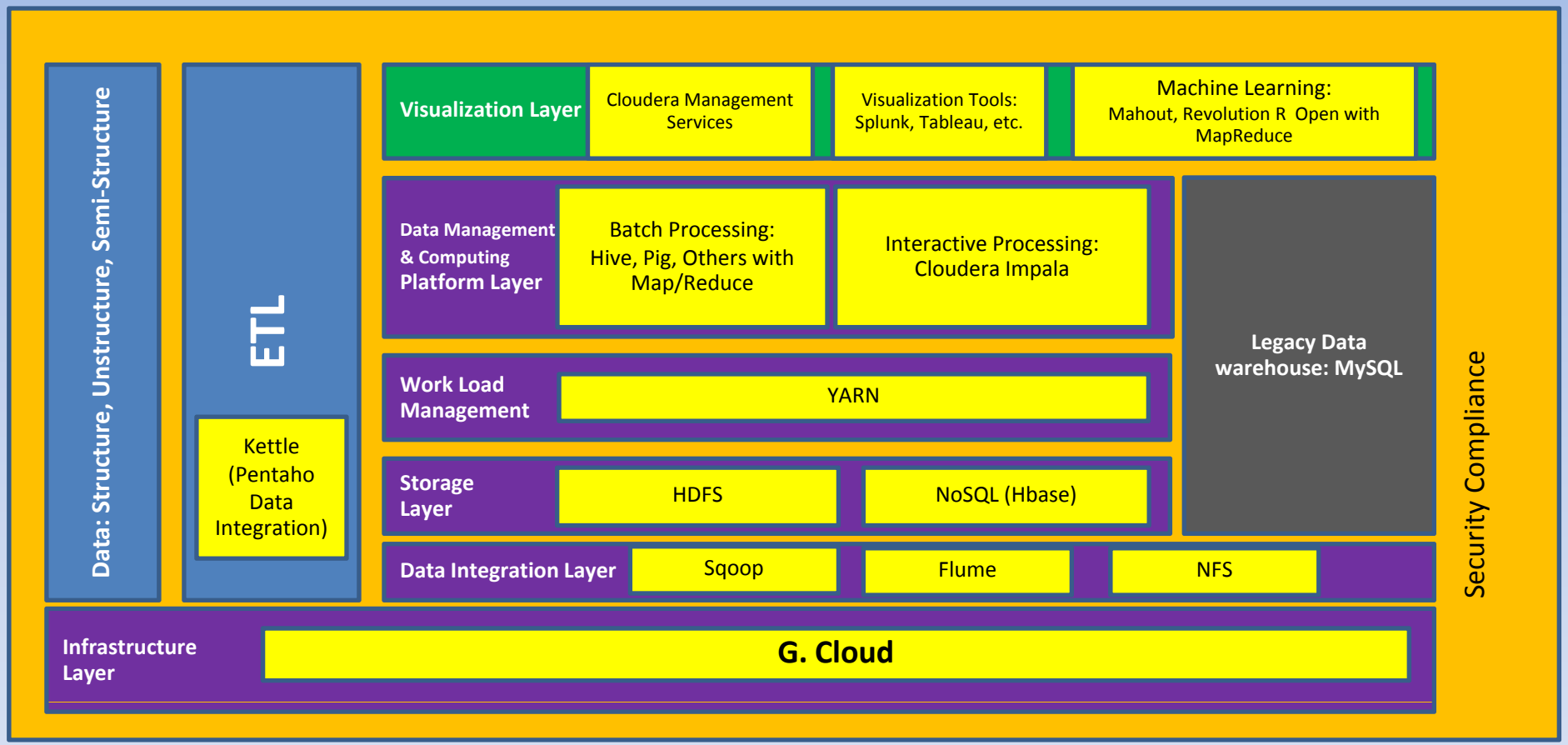
Attempt	Progress	State	Status	Node	Logs	Started	Finished	Elapsed	Note
attempt_1441553098625_0011_m_000002_0	100.00	SUCCEEDED	map	Data.node27:8042	logs	Thu Sep 10 09:37:37 +0700 2015	Thu Sep 10 09:38:21 +0700 2015	44sec	Container killed by the ApplicationMaster. Container killed on request. Exit code is 143 Container exited with a non-zero exit code 143

Showing 1 to 1 of 1 entries First Previous 1 Next Last

Data Management: Data Processing

- Batch Processing
- Interactive Processing
- Real-time Processing

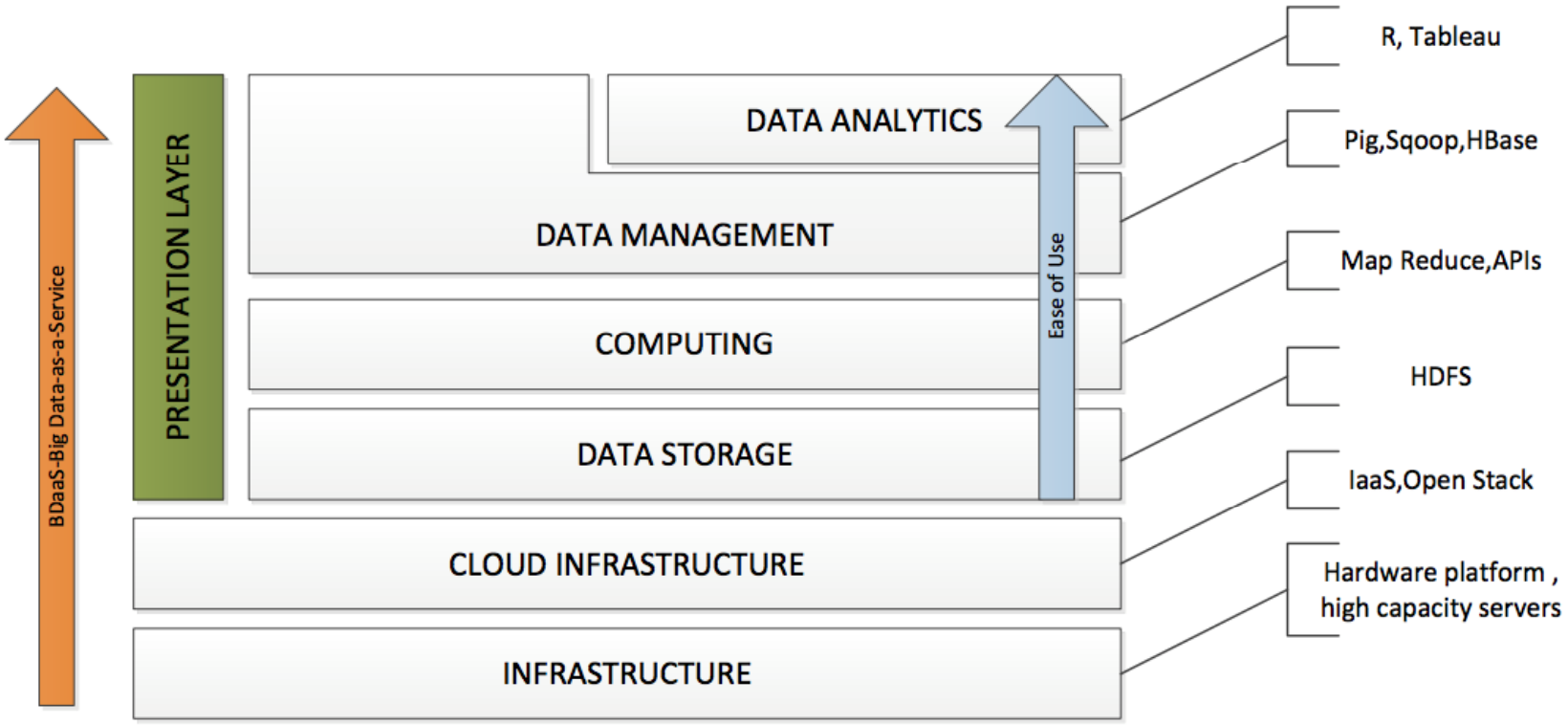
Architecture of Big Data Hadoop on Cloud Computing



Powered by

G-Cloud
Government Cloud Service

Framework of Big Data as a Service



Adapted from http://web.mit.edu/smadnick/www/wp/2014-06.pdf