



Big Data Platform

Assoc. Prof. Dr. Thanachart Numnonda

Executive Director, IMC Institute

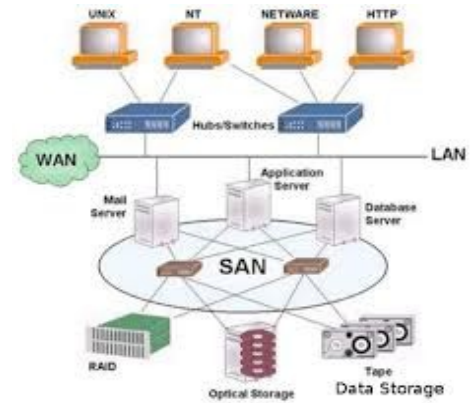
Aj. Araya Florence

Lecturer - Ubon Ratchathani University

11 September 2015



Data Sources



Technology



Analytics

**“By 2015, 20% of Global 1000 organizations
Will have established a strategic focus on
information infrastructure”**

Gartner

The Need for new technology

- Size of data is too big
- RDBMS are the square peg for all round holes
- Not all data requires ACID
- Many data are unstructure

Big Data Technology !!



ORACLE®

PostgreSQL



SYBASE®



Cassandra

mongoDB



membase



MAPR™
TECHNOLOGIES



Hortonworks

Big Data Technology



Storage

Analytics

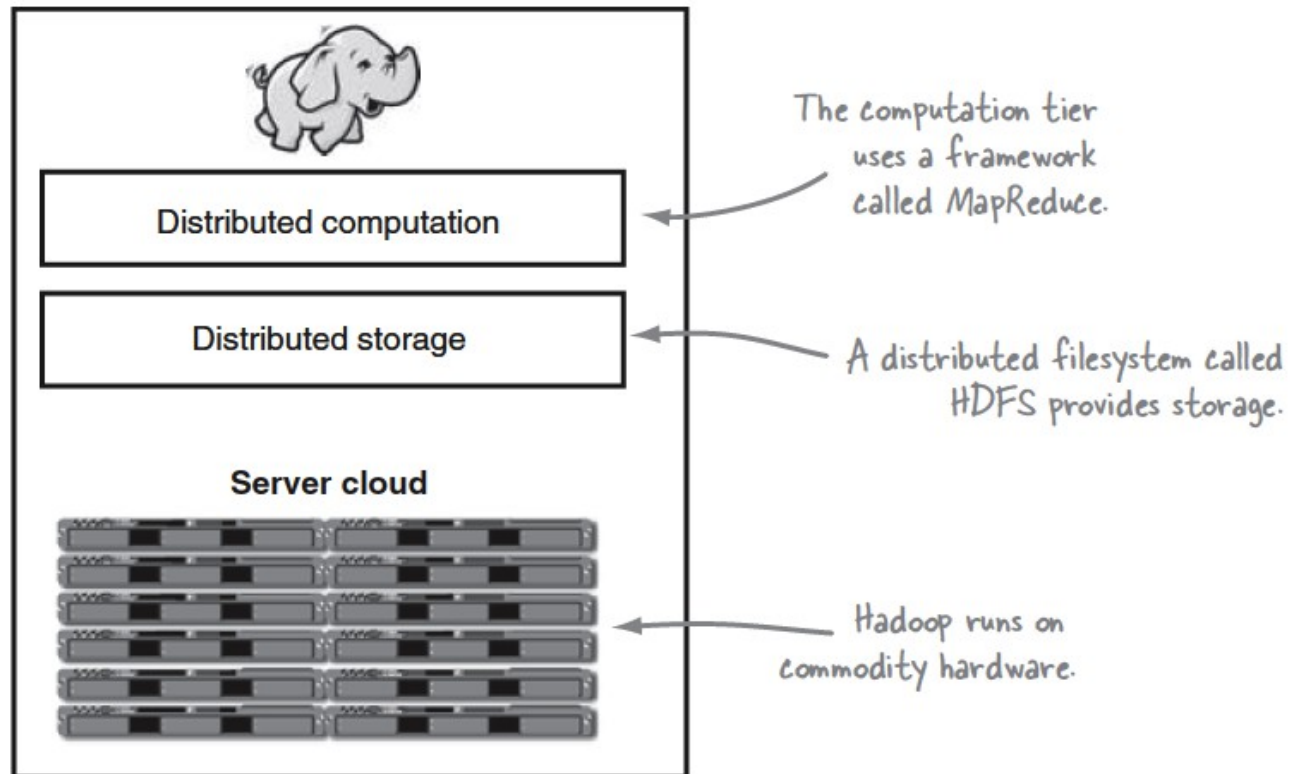
What is Hadoop?

**A scalable fault-tolerant distributed system
for data storage and processing**

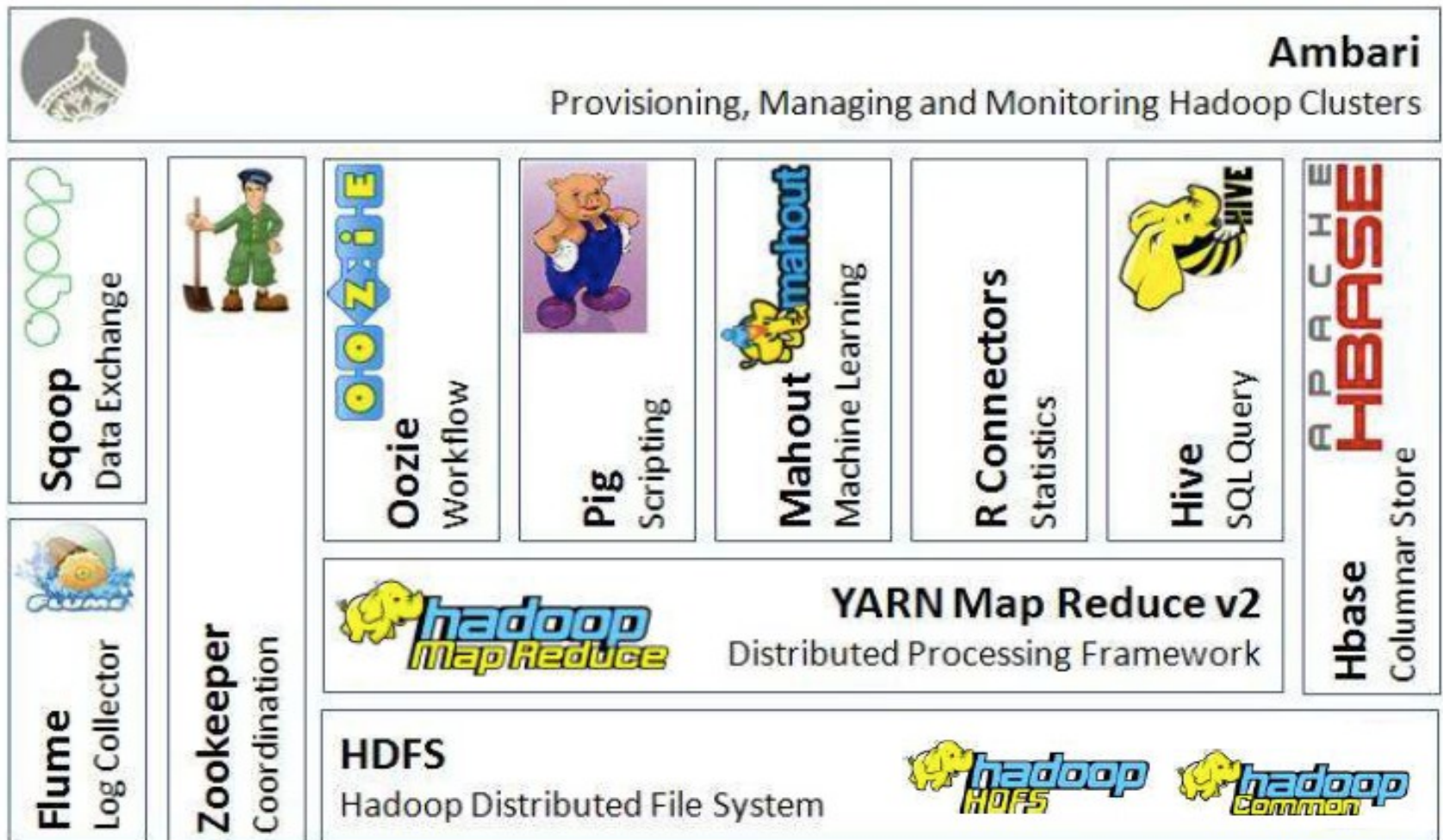
**Completely written in java
Open source & distributed under Apache license**



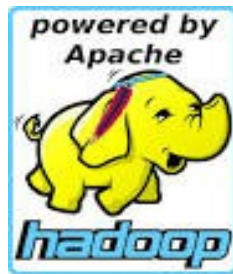
Hadoop Environment



Hadoop 2



Hadoop Distribution



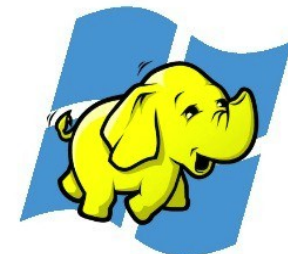
cloudera



Pivotal™

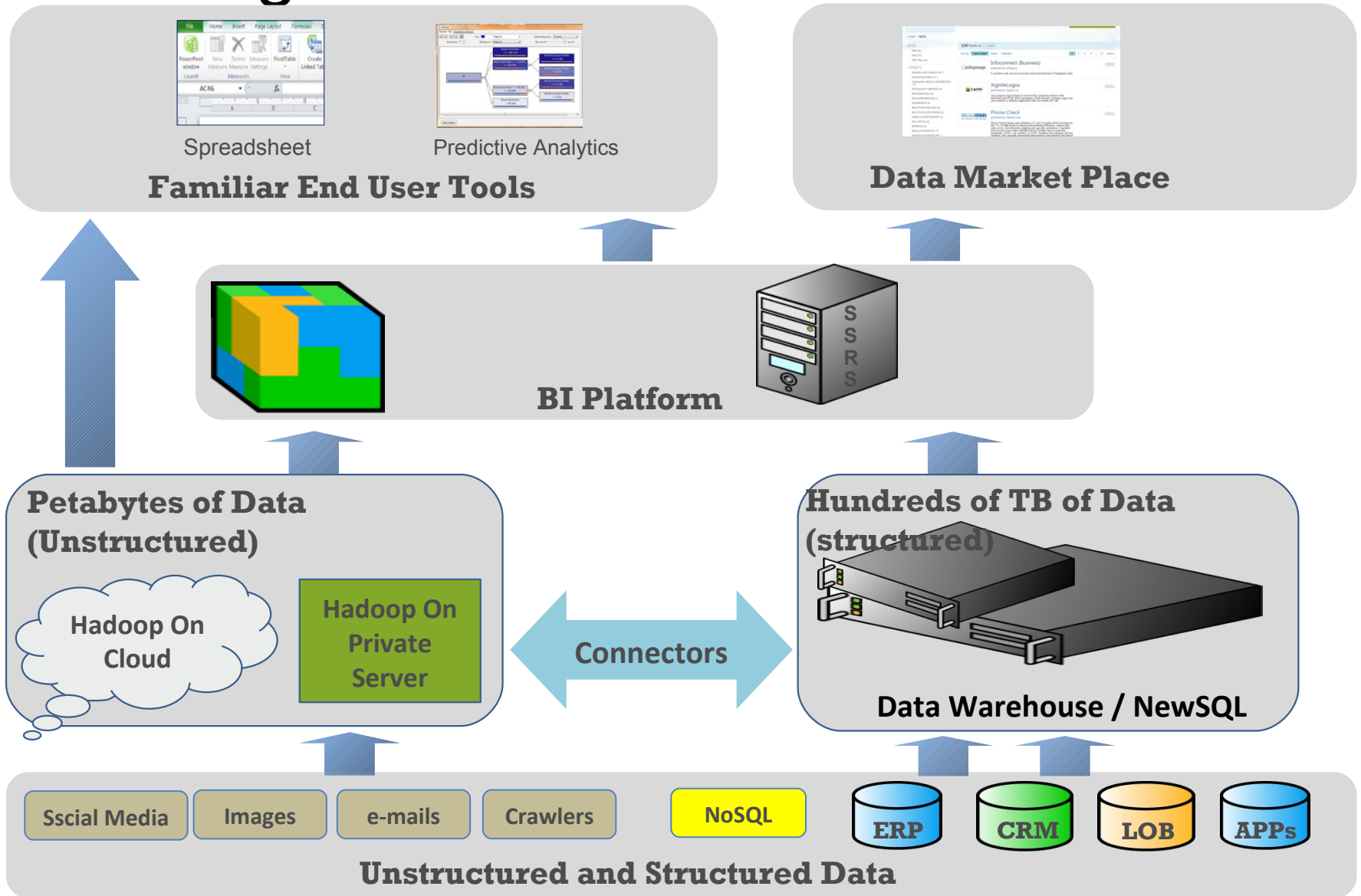


TERADATA®



Microsoft Azure

Big Data Future Architecture



ebay Data Infrastructure*

ORACLE®

- Thousands of nodes
- > 2K sharded logical host
- > 16K tables
- > 27K indexes
- > 140 billion SQLs/day
- > 5 PB provisioned



- Hundreds of nodes
- Persistent & in-memory
- > 40 billion SQLs/day



- 10+ clusters
- 100+ nodes
- > 250 TB provisioned (local HDD + shared SSD)
- > 9 billion writes/day
- > 5 billion reads/day

Not replacing RDMBS but
complementing!



- Hundreds of nodes
- > 50 TB
- > 2 billion ops/day



- Thousands of nodes
- The world largest cluster with 2K+ nodes

*by Jay Patel, Cassandra Summit June 2013 San Francisco

13/01/2014

Cassandra Introduction & Key Features by Philipp Potisk

18

Hadoop Led the Way

Big Data Platform: Demo

Hadoop Cluster

cloudera manager Home Clusters ▾ Hosts Diagnostics ▾ Audits Charts ▾ Backup ▾ Administration ▾ 5 Search (Hotkey: /) × Support ▾ admin ▾

30 minutes preceding September 11, 2015, 2:53 AM UTC

Home Status All Health Issues Configuration **✕ 3** ▾ All Recent Commands Add Cluster

Cluster 1 (CDH 5.4.5, Parcels) ▾

Hosts	✕ 3
Flume	▾
HBase	▾
HDFS	▾
Hive	▾
Hue	▾
Impala	▾
Key-Value Store...	▾
Oozie	▾
Solr	▾
Spark	▾
Sqoop 2	▾
YARN (MR2 Incl...	▴
ZooKeeper	▴

Charts

30m 1h 2h 6h 12h 1d 7d 30d 📄

Cluster CPU

Cluster 1, Host CPU Usage Across Hosts **0.82%**

Cluster Disk IO

Total Disk Byt... **137b/s** Total Disk Byt... **300K/s**

Cluster Network IO

Total Bytes R... **33.3K/s** Total Bytes Tr... **34.7K/s**

HDFS IO

Total Bytes Read Ac... **0** Total Bytes Wri... **1.9b/s**

Completed Impala Queries

Cloudera Management Service

Storage: HDFS

HUE [Home](#) [Query Editors](#) [Data Browsers](#) [Workflows](#) [Search](#) [File Browser](#) [Job Browser](#) [cloudera](#) [?](#) [🚩](#) [🔑](#)

File Browser

Search for file name [⚙️ Actions](#) [✖ Move to trash](#) [📁 Upload](#) [➕ New](#)

[🏠 Home](#) / [user](#) / [cloudera](#) [✎](#) [▼ History](#) [🗑 Trash](#)

☐	⬆ ⬆ Name	⬆ Size	⬆ User	⬆ Group	Permissions	⬆ Date
☐	📁 ↑		hdfs	supergroup	drwxr-xr-x	September 10, 2015 03:43 AM
☐	📁 .		cloudera	cloudera	drwxr-xr-x	September 10, 2015 05:18 AM
☐	📁 .Trash		cloudera	cloudera	drwxr-xr-x	September 10, 2015 05:10 AM
☐	📁 .staging		cloudera	cloudera	drwx-----	September 10, 2015 04:54 AM
☐	📁 2008.csv		cloudera	cloudera	drwxr-xr-x	September 06, 2015 04:47 AM
☐	📁 airline		cloudera	cloudera	drwxr-xr-x	September 09, 2015 08:59 PM
☐	📁 ml-1m		cloudera	cloudera	drwxr-xr-x	September 09, 2015 09:19 PM
☐	📄 movies.dat	489.1 KB	cloudera	cloudera	-rw-rw-rw-	September 10, 2015 04:39 AM
☐	📁 rating.dat		cloudera	cloudera	drwxr-xr-x	September 10, 2015 04:49 AM
☐	📄 ratings.dat	224.2 MB	cloudera	cloudera	-rw-rw-rw-	September 10, 2015 04:39 AM
☐	📄 tmovies.dat	159.7 KB	cloudera	cloudera	-rw-r--r--	September 10, 2015 05:18 AM

Hadoop: Analytics





hue ▾

Configuration

Check for misconfiguration

Server Logs

Ambari enabled successfully

[Leave Feedback](#)

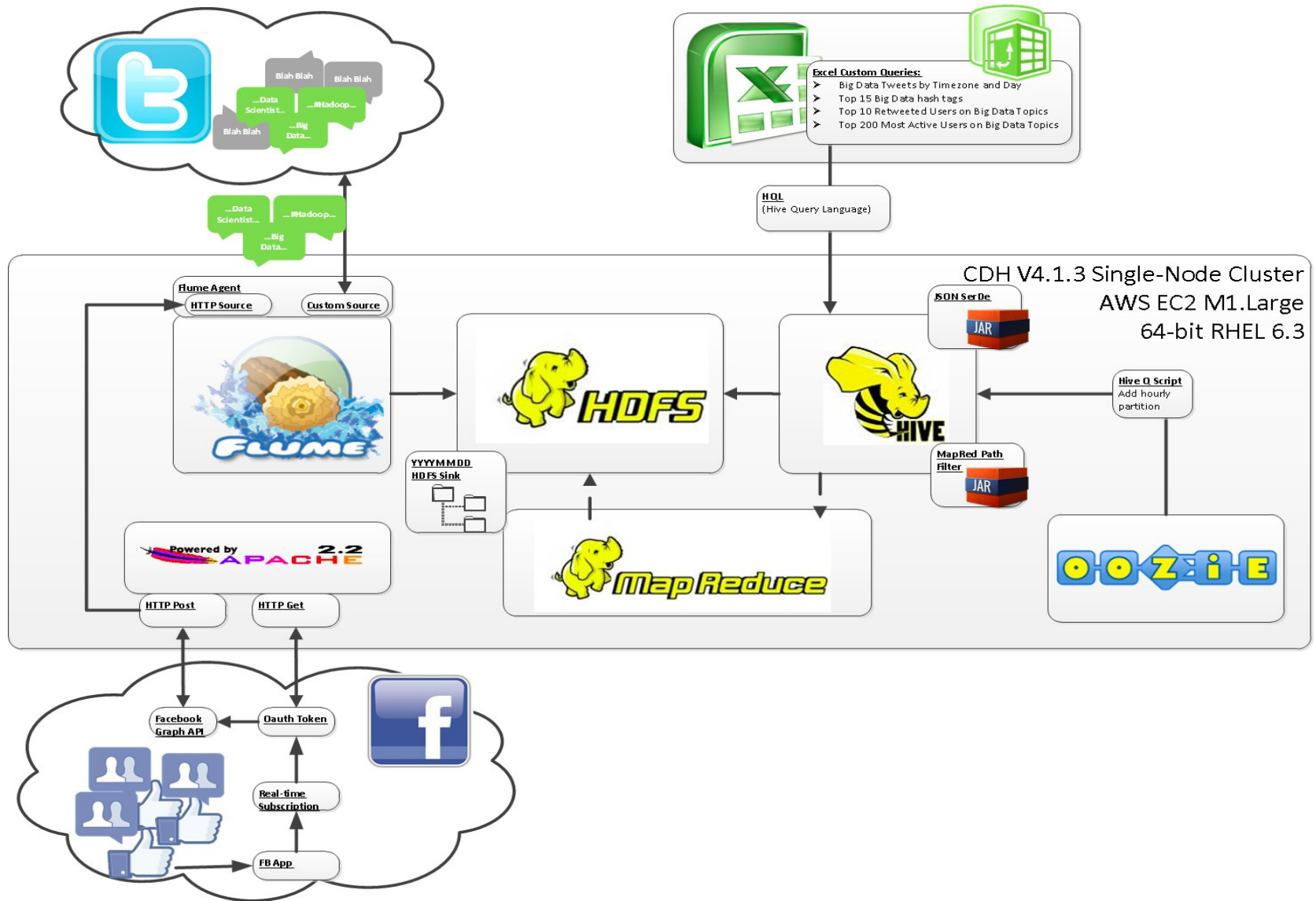
Component	Version
Hue	2.6.1-2
HDP	2.2.4
Hadoop	2.6.0
Pig	0.14.0
Hive-Hcatalog	0.14.0
Oozie	4.1.0
Ambari	2.0-151
HBase	0.98.4
Knox	0.5.0
Storm	0.9.3
Falcon	0.6.0

Disable

Copyright © 2013 The Apache Software Foundation.

Apache Hadoop, Hadoop, HDFS, HBase, Hive, Mahout, Pig, Zookeeper are trademarks of the Apache Software Foundation.

Hue and the Hue logo are trademarks of Cloudera, Inc. and licensed under the Apache 2 license. For more information: gethue.com





www.facebook.com/imcinstitute

Thank you

thanachart@imcinstitute.com

www.facebook.com/imcinstitute

www.slideshare.net/imcinstitute

www.thanachart.org