

Hadoop ecosystem

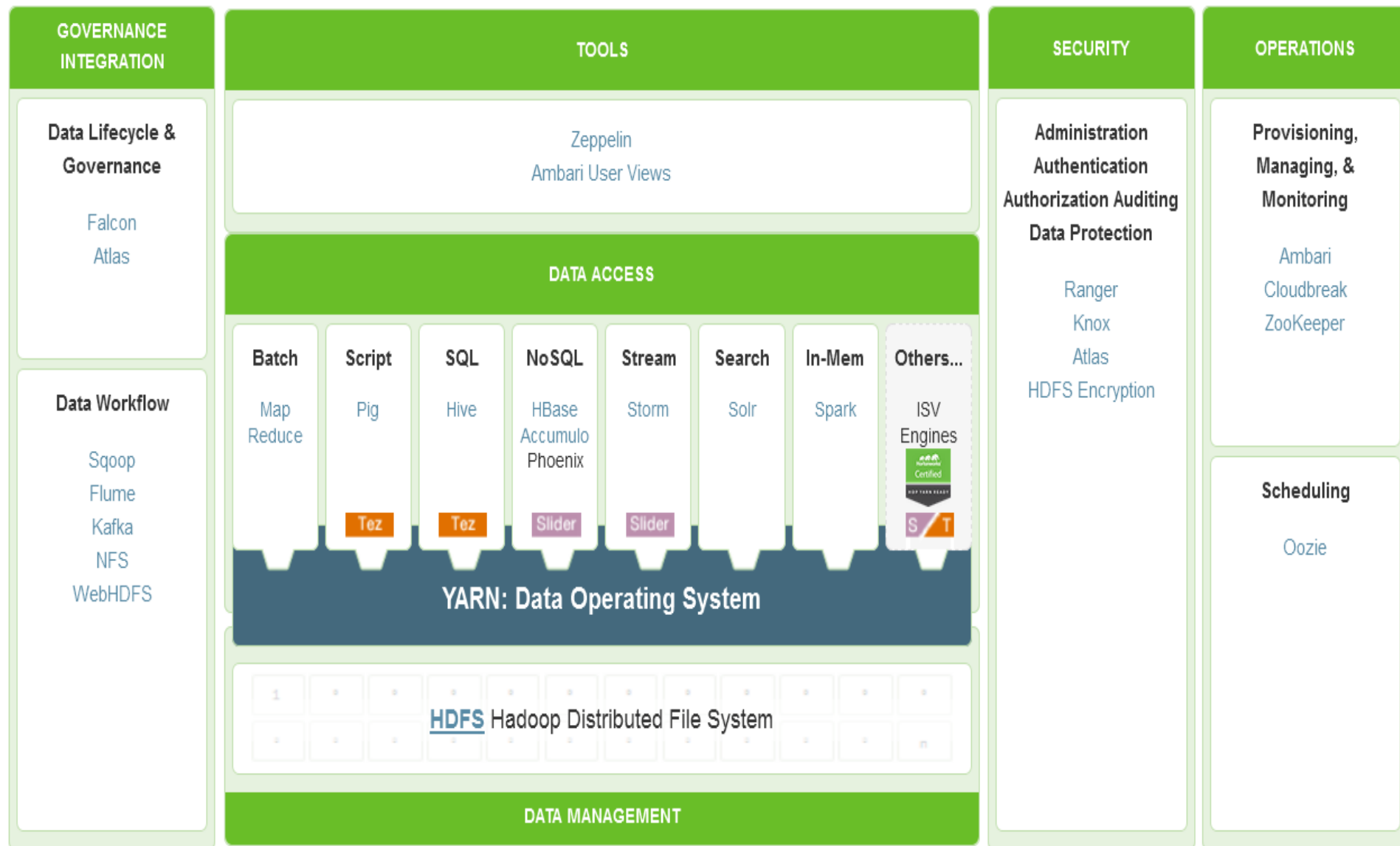
2015/12/21

Scott Miao

@takeshi.miao

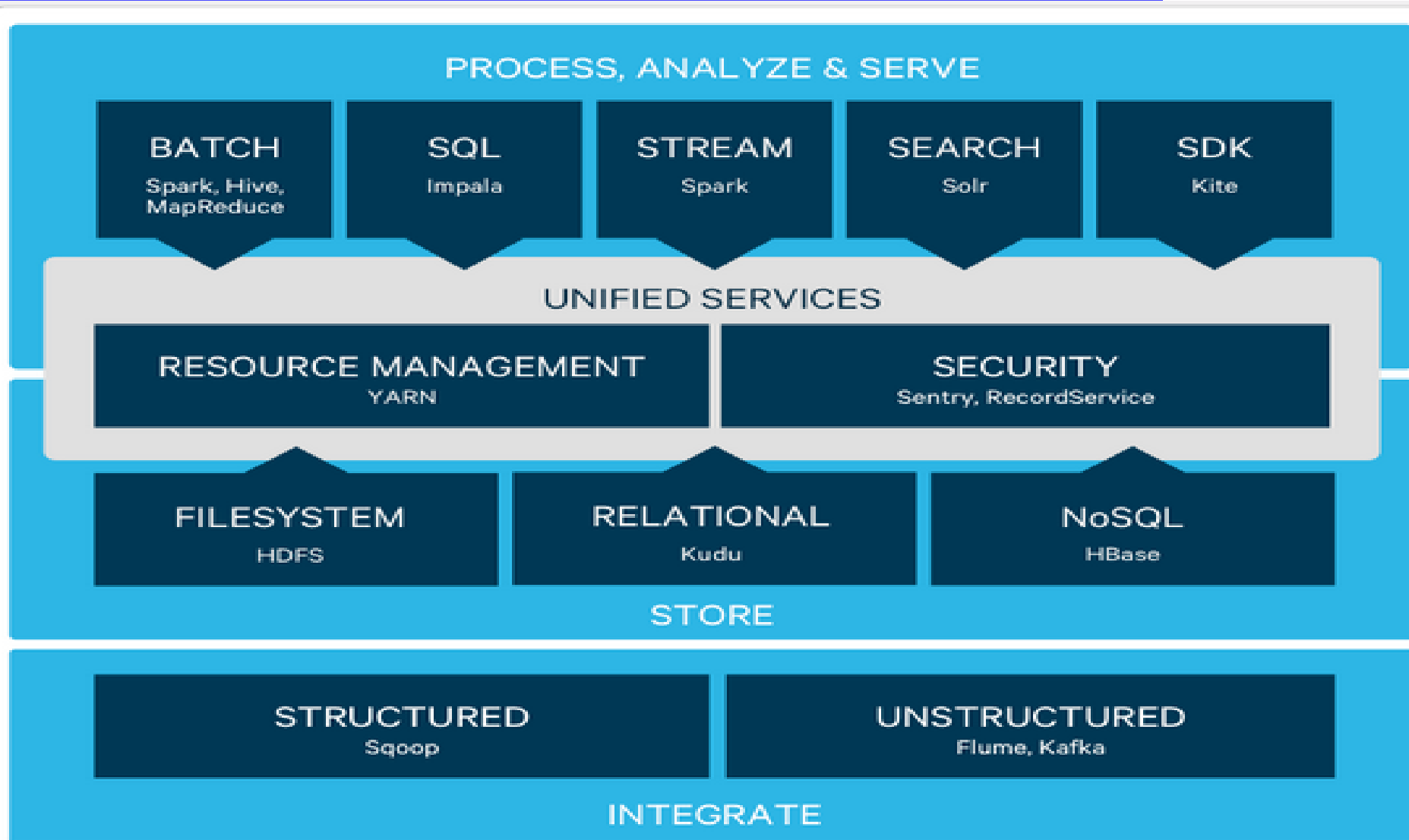
Hadoop ecosystem by Hortonworks

<http://hortonworks.com/hdp/>





Hadoop ecosystem by Cloudera

<http://www.cloudera.com/content/www/en-us/products/apache-hadoop.html>






Key projects should know

Project	Briefing
Apache Hadoop-2.x 	<p>A framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.</p> <ul style="list-style-type: none">• Hadoop Common: The common utilities that support the other Hadoop modules.• Hadoop Distributed File System (HDFS™): A <u>distributed file system</u> that provides high-throughput access to application data.• Hadoop YARN: A framework for <u>job scheduling and cluster resource management</u>.• Hadoop MapReduce: A YARN-based system for <u>parallel processing of large data sets</u>.
Apache Zookeeper 	<p><u>A centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services.</u> All of these kinds of services are used in some form or another by distributed applications.</p> <p>* A de facto Lookup service used in distr. systems</p>




* Added by Scott Miao

Key projects should know

Project	Briefing
Apache Pig 	<p>A platform for analyzing large data sets that consists of a <u>high-level language</u> for expressing data analysis programs, coupled with infrastructure for evaluating these programs. At the present time, Pig's infrastructure layer consists of a <u>compiler that produces sequences of Map-Reduce programs</u>, for which large-scale parallel implementations already exist.</p> <p>* Script, usually used for ETL</p>
Apache HBase 	<p>An <u>open-source, distributed, versioned, non-relational database</u> modeled after Google's Bigtable: A Distributed Storage System for Structured Data by Chang et al.</p> <p>* NoSQL DB usually bundled with Hadoop</p>
Apache Phoenix 	<p>A relational database layer over HBase delivered as a client-embedded JDBC driver targeting low latency queries over HBase data.</p> <p>* SQL on HBase</p>




* Added by Scott Miao

Key projects should know

Project	Briefing
Apache Hive 	A data warehouse software facilitates <u>querying and managing large datasets residing</u> in distributed storage. * SQL on Hadoop
Apache Impala 	The open source, native analytic database for Apache Hadoop. * SQL on Hadoop
Apache Storm 	A free and open source distributed realtime computation system. Storm makes it easy to reliably process unbounded streams of data, doing for realtime processing what Hadoop did for batch processing. * Streaming data digest





* Added by Scott Miao

Key projects should know

Project	Briefing
Apache Spark 	<p>A fast and general engine for large-scale data processing.</p> <ul style="list-style-type: none">* There are different libs for different usecases<ul style="list-style-type: none">* Spark SQL* Spark Streaming* MLLib (Machine Learning)* GraphX (Graph processing)
Apache Flume 	<p>A distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data.</p> <ul style="list-style-type: none">* Unstructured log collector
fluentd 	<p>An open source data collector for unified logging layer.</p> <ul style="list-style-type: none">* Unstructured log collector




** Added by Scott Miao*

Key projects should know

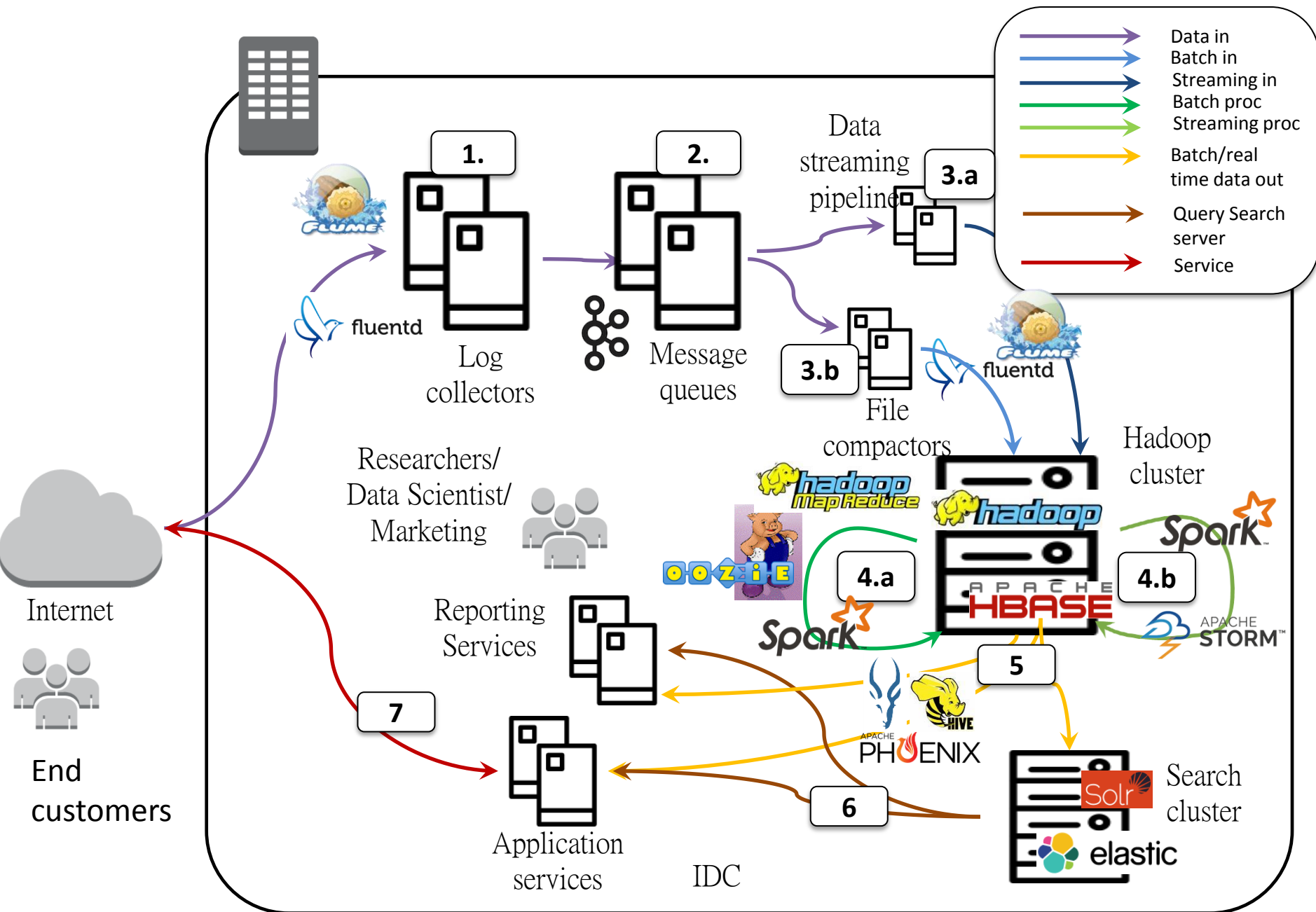
Project	Briefing
Apache Sqoop 	A tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases. * Structured data collector
Apache Kafka 	is publish-subscribe messaging rethought as a distributed commit log. * A distributed and easy scale out message queue
Apache Oozie 	A workflow scheduler system to manage Apache Hadoop jobs
Apache Ambari  Apache Ambari	Is aimed at making Hadoop management simpler by developing software for provisioning, managing, and monitoring Apache Hadoop clusters * Hadoop clusters manager

* Added by Scott Miao

Key projects should know

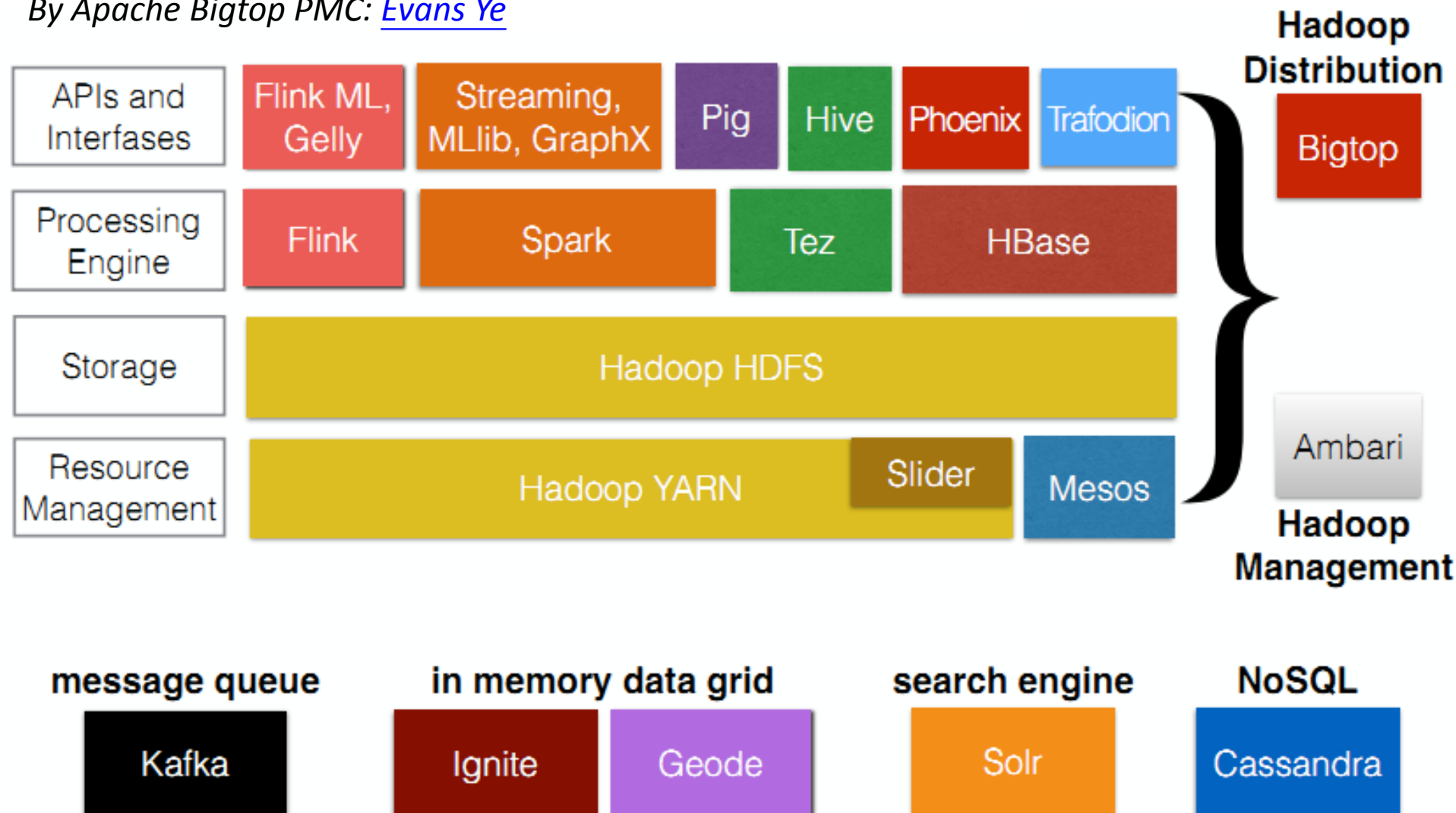
Project	Briefing
<p>Apache Lucene</p>  	<p>A open-source search software</p> <p>Lucene Core: our flagship sub-project, provides Java-based indexing and search technology</p> <p>Solr™: is a high performance search server built using Lucene Core</p> <p>Open Relevance Project: is a subproject with the aim of collecting and distributing free materials for relevance testing and performance.</p> <p>PyLucene: is a Python port of the Core project.</p> <p>* Search engine + Server</p>
<p>Elastic Search</p> 	<p>Elastic believes getting immediate, actionable insight from data matters. As the company behind the three open source projects — Elasticsearch, Logstash, and Kibana — designed to take data from any source and search, analyze, and visualize it in real time, Elastic is helping people make sense of data.</p> <p>* Search engine + server, log ETL, Visualization UI</p>

* Added by Scott Miao



Hadoop ecosystem supported by Apache Bigtop

By Apache Bigtop PMC: [Evans Ye](#)



Where are we all?

- [Home](#)
- [Committer Map](#)
- [Calendar](#)
- [Event Pictures](#)
- [Gallery](#)
- [Projects](#)
- [Resources](#)
- [Weblogs](#)

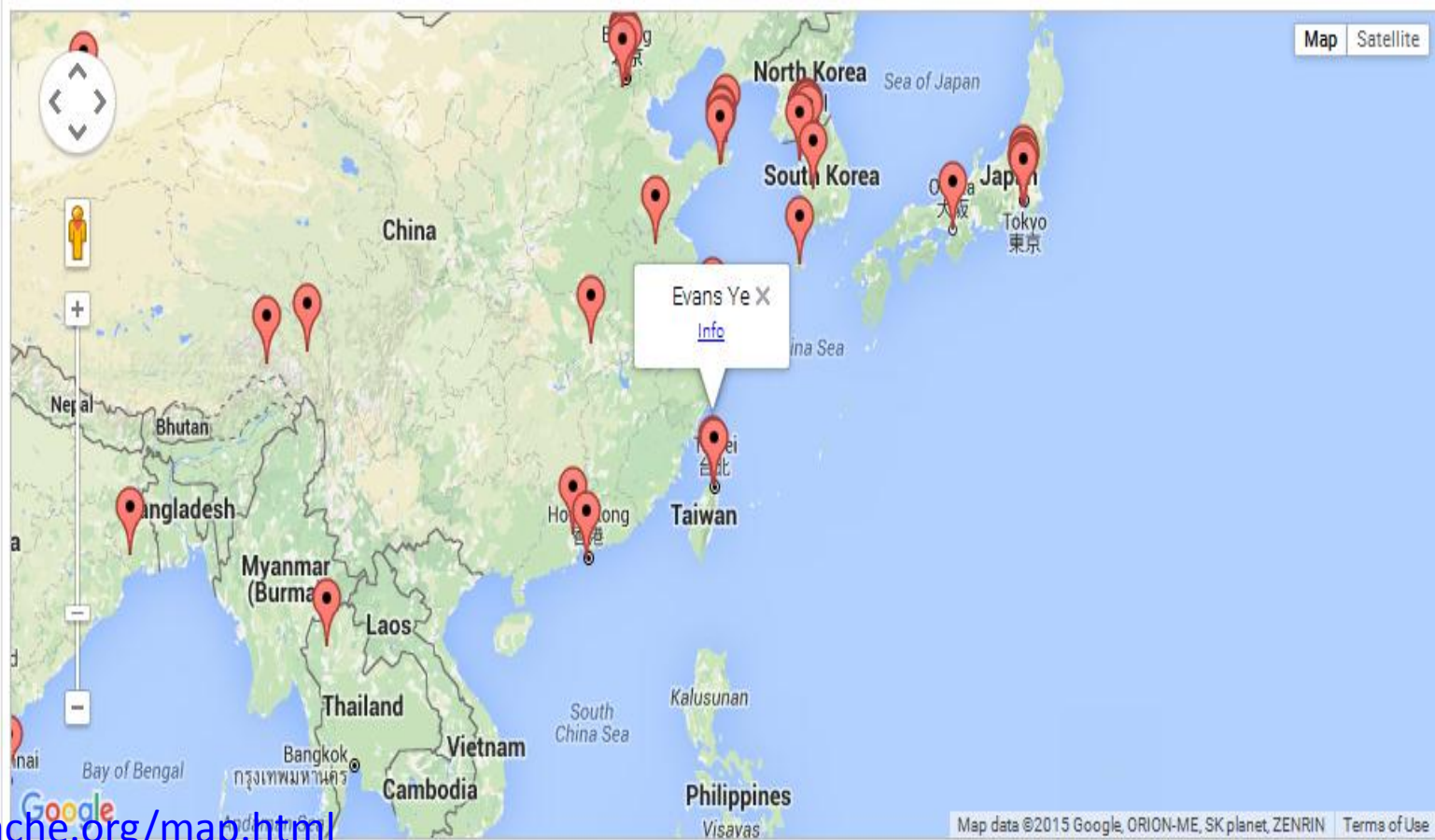
[Managing Your Details](#)

- [Committer Index](#)
[A](#)[B](#)[C](#)[D](#)[E](#)[F](#)[G](#)[H](#)[I](#)[J](#)[K](#)[L](#)
[M](#)[N](#)[O](#)[P](#)[Q](#)[R](#)[S](#)[T](#)[U](#)[V](#)[W](#)
[X](#)[Y](#)[Z](#)

- [Committers by login id](#)
[Committers by auth group](#)

This map shows locations of those committers within the ASF that have added their locations.

Predefined views: [Worldwide](#) [Europe](#) [North America](#)



You can study them all online !

- Hortonworks tech blog
 - <http://hortonworks.com/blog/>
- Cloudera tech blog
 - <http://blog.cloudera.com/>
- Hadoop summit on Youtube
 - <https://www.youtube.com/user/HadoopSummit>
- * usergroup
 - E.g. [Spark usergroup](#)