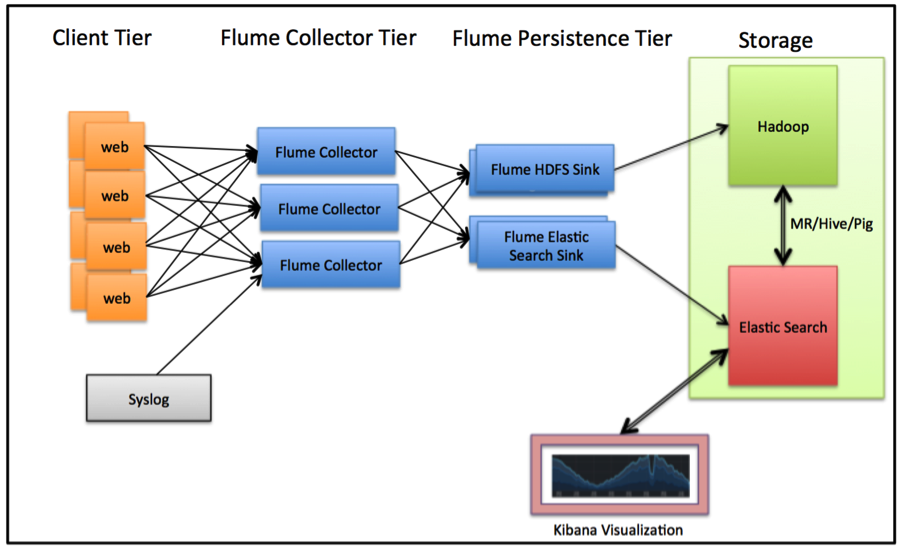
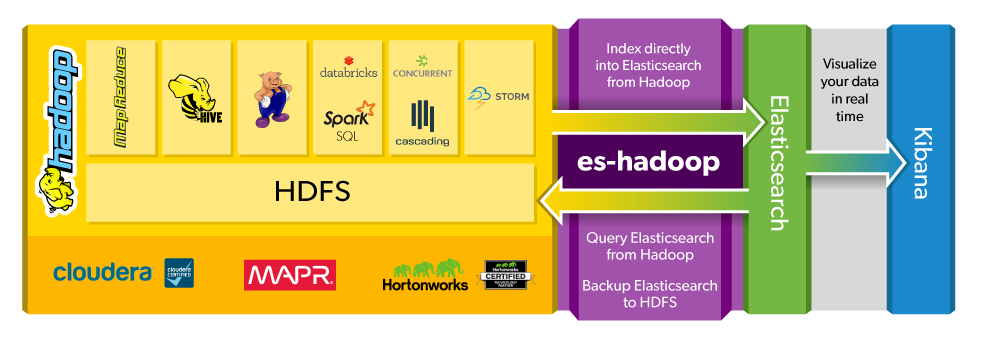
ElasticSearch（ES）简介和使用

ElasticSearch是基于Lucene的搜索服务器。它提供了一个分布式多用户能力的全文搜索引擎，基于RESTful web接口，Elasticsearch是用Java开发的，能够达到实时搜索，文档，可靠，快速及安装使用方便。Elasticsearch引擎结合YARN，可以提供Hadoop中的数据的实时访问和搜索能力。使用框架如下图所示：

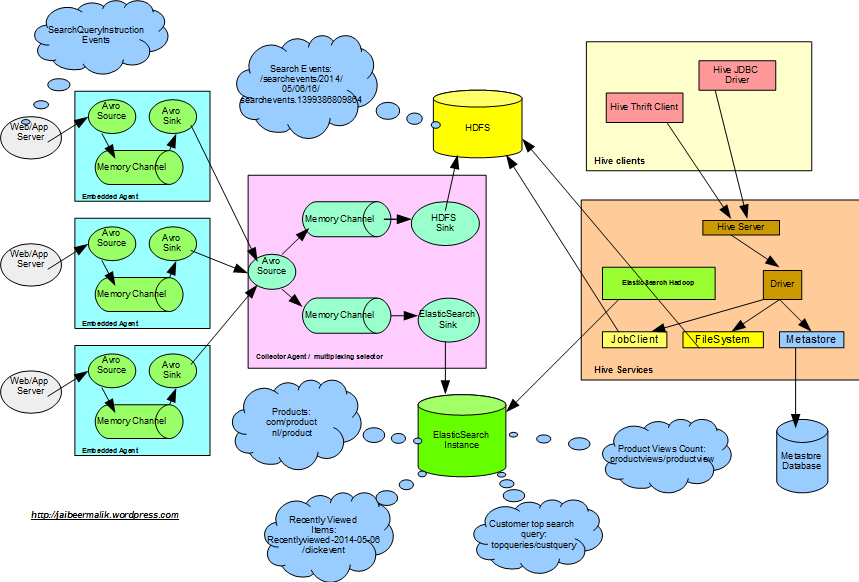


上图中将日志文件以近实时的方式在Elasticsearch中建立索引，并保存到Hadoo中，用于长实时批处理分析。

ElasticSearch-Hadoop，集合Hadoop和ElasticSearch的项目，也是ES官方维护的一个子项目，通过实现Hadoop和ES之间的输入和输出，可以在Hadoop里面对ES集群的数据进行读取和写入，充分发挥Map-Reduce并行处理的优势，为Hadoop数据代理实时搜索的可能。框架图如下所示：



使用Flume从客户端系统中收集数据，并将数据存储到Hadoop HDFS和ElasticSearch中，可以使用Hive分析相同数据，并产生统计数据。



Elasticsearch-hadoop-hive，使用Hive访问ElasticSearch，从Hive Table中抽取数据，读取并将数据索引到ElasticSearch，并在website中进行展示。

示例如下：

1. 从Hive table中搜索数据，表如下所示：

# search.search\_productviews : id, productid, viewcount

61, 61, 15

48, 48, 8

16, 16, 40

85, 85, 7

1. 建立索引

创建Hive External Table: search\_productviews\_to\_es，用于保存索引数据

Use search;

DROP TABLE IF EXISTS search\_productviews\_to\_es;

CREATE EXTERNAL TABLE search\_productviews\_to\_es (id STRING, productid BIGINT, viewcount INT) STORED BY 'org.elasticsearch.hadoop.hive.EsStorageHandler' TBLPROPERTIES('es.resource' = 'productviews/productview', 'es.nodes' = 'localhost', 'es.port' = '9210', 'es.input.json' = 'false', 'es.write.operation' = 'index', 'es.mapping.id' = 'id', 'es.index.auto.create' = 'yes');

INSERT OVERWRITE TABLE search\_productviews\_to\_es SELECT qcust.id, qcust.productid, qcust.viewcount FROM search\_productviews qcust;

该外表用于对应ES Instance，索引Productviews和Document Type用于索引数据。

1. produce views的数据构建索引，使用HiveSearchClicksServiceImpl.java

Collection<HiveScript> scripts = new ArrayList<>();

HiveScript script = new HiveScript(new ClassPathResource("hive/load-search\_productviews\_to\_es.q"));

scripts.add(script);

hiveRunner.setScripts(scripts);

hiveRunner.call();

1. productviews索引的样例数据

{id=48, productid=48, viewcount=10}

{id=49, productid=49, viewcount=20}

{id=5, productid=5, viewcount=18}

{id=6, productid=6, viewcount=9}

1. Hive Data的查询

# search.search\_customerquery : id, querystring, count, customerid

61\_queryString59, queryString59, 5, 61

298\_queryString48, queryString48, 3, 298

440\_queryString16, queryString16, 1, 440

47\_queryString85, queryString85, 1, 47

参考文献：

ES-Hadoop github源码：<https://github.com/elasticsearch/elasticsearch-hadoop#readme>

ES-Hadoop: <http://www.elasticsearch.org/overview/hadoop/>

ElasticSearch中文发行版：https://github.com/medcl/elasticsearch-rtf

ElasticSearch-Hadoop : <https://www.elastic.co/products/hadoop>

http://www.javacodegeeks.com/2014/05/elasticsearch-hadoop-indexing-product-views-count-and-customer-top-search-query-from-hadoop-to-elasticsearch.html