|  |  |
| --- | --- |
| **Cloudera** | **Hortonworks** |
| Cloudera announced its long-term achievement to be an enterprise data hub thus eliminating the need for a Data Warehouse. | Hortonworks looks forward to firmly provide Hadoop distribution partnering with data warehousing company Teradata, just for this purpose |
| Cloudera CDH can run on Windows server | Hortonworks HDP is a native component on Windows Server. A Hadoop based Hadoop cluster can be deployed on Windows Azure through HDInsight service |
| Cloudera has the proprietary management software called the Cloudera Manager, SQL Queries handling interface called the Impala, Cloudera Search to provide real-time and easy access of products | Hortonworks uses Ambari for management, Stinger for handling queries and Apache Solr for data search. Hence there are no proprietary software in its ecosystem. |
| Cloudera with its proprietary software in usage has a commercial license. Cloudera also encourages the use of its open source projects absolutely free but it doesn’t include Cloudera Manager or any other proprietary software in the package | Hortonworks on the other goes by an open source license. |
| Cloudera comes with a 60 days free trial | Hortonworks is completely free, absolutely. |
| Cloudera has been in this market than any other of its counterparts with more than 350 customers. | Hortonworks is catching up the race quite fast and has more innovations in the Hadoop ecosystem than Cloudera in the recent past |
| Cloudera has many enterprise software laid over its open source distributions to help the customers with their unique requirements | Hortonworks provides a framework constituting just the open source projects striving to fulfil all the customer requirements |

Horton works vs cloudera vs databricks vs mapr

**Hortonworks**' distribution is claimed to be 100 percent open source

when IT analyst Gartner released its Magic Quadrant report for Data Science Platforms (see below) in 2017, Cloudera didn’t even hit the qualifiers list.

Cloudera

Pros

Support Spark

More mature libraries and community

More stable

Cloudera Manager – very rich user interface

allows creation of groups of nodes in a Hadoop cluster with varying configuration

Cons

Sells commercial software

components in CDH like Cloudera Manager need to buy a license

Support Linux only

Lock-in

Hortonworks

Management tools: Ambari UI

Pros

No proprietary software associated

Open Source

Support windows and Linux

offers only Apache Foundation certified software

easy to deploy and implement. Work well with 3rd applications like API calls, etc.

Hive is better in performance and ease of use compared to Impala

**Apache Kylin** – Data Cube Management on top of hadoop

Multi-dimensional analysis (OLAP) on Hadoop supporting extremely large datasets.

 OLAP cube meets real-time

Open source Distributed Analytics Engine

Provide SQL interface

Reduce query latency on Hadoop for 10+ billion rows of data

Sub second interactive queries for data Volumes over billions of rows on the fact table.

better than Hive queries

Apache Kylin based on two Hadoop stack: [Apache Hive](http://hive.apache.org/) and [HBase](https://hbase.apache.org/).

First, implement Data Warehouse (DW) on Hive database using star or snow flake schemas.

Then define an OLAP cube on Kylin.

There are keys technologies for Kylin; [Apache Hive](http://hive.apache.org/) and [Apache HBase](https://hbase.apache.org/).   
The Data Warehouse is based on a Start Model stored on Apache Hive.   
Using this model and a definition of a meta-data model, Kylin builds a multidimensional MOLAP Cube in HBase.   
After the cube is builded the users can query it, using an SQL based language with its JDBC driver.

 USE CASE

1. Design a star schema using HiveQL b) Load data into Hadoop using a Hive connector c) Kylin would handle mapping the Hive schema to a cube schema and executing the Map/Reduce through HiveQL d) Output the results of the query to HBase e) Use Kylin for executing SQL statements via Calcite to HBase and returning the results as JSON
2. Define a Kylin’s cube model using Kylin’s GUI with wizard. At this moment, Kylin can generate the MOLAP cube in an automatic process. After cube creation, we can query the OLAP cube using SQL queries or connecting to a BI tool using the available J/ODBC connectors.

<https://www.ebayinc.com/stories/blogs/tech/announcing-kylin-extreme-olap-engine-for-big-data/>

**Cons**

* A cube is pre-constructed, therefore provides access to a data snapshot. It should be re-built in case of updates and built differently in case of changes in dimensions or measures
* Kylin sometimes failed to build a cube or exited during the experiment
* One limitation however is Kylin provides only aggregated results, or in other word, SQL should contain a "group by" clause to yield correct result.
* Kylin only supports the [star schema](https://en.wikipedia.org/wiki/Star_schema). You are limited to a single fact table for each cube.
* 60 dimensions in one cube
* only works with hive tables as an input source

Cloudera

Apache kudo relational database - RDBMS

Cloudera navigator – Data security and lineage

**KYLO –** Self-service data ingestion, data wrangling, data profiling, data validation, data cleansing/standardization, and data discovery

open source project under the Apache 2.0 license

Kylo supports both the Horton Works and Cloudera distributions of Hadoop

Kylo is an open source enterprise-ready data lake management software platform for self-service data ingest and data preparation with integrated metadata management, governance, security and best practices.

Kylo uses Apache Spark for data profiling, data validation, data cleansing, data wrangling, and schema detection

Kylo integrates best practices around metadata capture, security, and data quality.

Kylo has an integrated metadata server currently compatible with databases such as MySQL and PostgreSQL.

Kylo can manage the creation and usage of Nifi RDBMS data source configurations, through a simple [Data Source UI](http://localhost:8400/index.html#!/datasources).

Kylo retrieves the list of database connections from Apache NiFi

Feed creation wizard UI allows end-users to configure cleansing and standardization functions to manipulate data into conventional

**Automatic Data Profiling and Search-based Data Discovery**

Kylo automatically generates profile statistics such as minimum, maximum, mean, standard deviation, variance, aggregates (count & sum), occurrence of null values, occurrence of uniqueness, occurrence of missing values, occurrence of duplicates, occurrence of top values, and occurrence of valid & invalid values.

Work with HORTONWORKS and is on top of Hadoop, NiFi

Banking Tips:

Customer Spending

Congratulating on birthdays

Data Collection

Account transactional history

Personal information

To determine what products or services would be most useful.

Notes:

Python and Pyspark libraries

Spark SQL is much faster than HQL

Hive DB

temp - - not useful

#### [test\_user\_db](javascript:void(0)) - - not useful

-check date format

**cxttbls.ct\_trxn\_date**

**cxttbls.ct\_post\_date\**

**baf** - - **empty**

#### [transformation](javascript:void(0))

#### No Use

**trn\_ac\_inactive\_reset**

trn\_account\_class

**trn\_baf\_cao**

trn\_baf\_cao\_hist

trn\_banker\_cheque

trn\_branch

**trn\_consolidate\_asst\_liab**

 trn\_currency

IMPORTANT

transformation.trn\_account

transformation.trn\_account\_closed

transformation.trn\_currency & trn\_currency\_hist

transformation.trn\_customer

transformation.trn\_md\_deal

#### [bafl\_dw](javascript:void(0))

atm\_locations

bafl\_complaints

bafl\_complaints\_view

cardpro\_analysis\_view

cardprocustomers

cardprocustomers1

NOT COMPLETED

#### 

#### [phoenix](javascript:void(0))

#### 

#### Important

#### Atm

#### atm\_location

#### atm\_location1

#### branch\_list ----------- For Grouping

#### Staging

#### stg\_account\_closed

#### stg\_customer

#### stg\_loan\_deposit

#### 

#### cardpro

#### cardpro\_analysis

#### cp\_crdtrx

#### cp\_csttbl

#### dup

USE CASE

[AI-driven chatbot](https://www.wellsfargo.com/about/press/2017/chat-bot_0418.content) through the Facebook Messenger platform to communicate with users and provide assistance with passwords and accounts.

Recommendation Engine of financial products to customer

Fixed deposits, credit cards, home loans Automation

Uses past information about borrowers’ default rates to predict the likelihood of default for future borrowers.

Incorporating the predictive models into a real-time loan approval process is easy, allowing businesses to scale up and expand their loan portfolios.

Investment Predictions for customer

Identifying a risk score of a customer based on his/ her nationality, occupation, salary range, experience, industry he/she works for, credit history et.

**Hadoop Ecosystem:**

HDFS – Distributed File System

Sqoop – Data Exchange

Hive – SQL Queries

Hbase – Column Based Data Storage, data is stored in key value pairs. No SQL

MapReduce - processing and generating big data sets with a parallel, distributed algorithm on a cluster

Pig - Scripting

Flume – Log Collector – fetching, aggregating, moving large amount of log data

Apache Hadoop YARN – YET ANOTHER RESOURCE NEGOTIATOR

-- The resource management and job scheduling tool

Flink - open source stream processing framework. Best choice for streaming applications

**Data Security**

Apache Ranger:

 A framework to enable, monitor and manage comprehensive data security across the Hadoop platform

**Data Governance:**

**Apache Falcon**: A data governance engine that defines, schedules, and monitors data management policies

**Data Lineage:**

**Apache Atlas**:

**Data Lineage:**Captures lineage across Hadoop components at platform level

To view Data Lineage

### [Data Governance and Metadata framework for Hadoop](https://atlas.apache.org/)

Data Ingestion Tools:

Sqoop

Flume

Kafka

Data Access Tools:

HIVE

PIG

Hbase

Storm

Monitoring tools

Ambari

Data governance tools

Apache Atlas

Support for data classification

Support for centralized auditing

Proving a search and lineage function

Building a security and policy engine

**Batch layer**

MapReduce

Pig

**Serving layer**

Hbase

MongoDB

For interactive queries and full CRUD operations on the datasets

**Speed layer**

Apache Storm

Apache Spark Streaming

**Authorization / Authentication**

Kerberos

**DATA MANAGEMENT**HADOOP / YARN **DATA ACCESS**PIG  
HIVE  
TEZ  
SOLR  
SPARK  
SLIDER  
Hbase  
PHEONIX  
ACCUMULO  
STORM

**GOVERNANCE AND INTEGRATION**FALCON  
ATLAS  
SQOOP  
FLUME  
KAFKA

**OPERATIONS**AMBARI  
CLOUDBREAK  
ZOOKEEEPER  
OOZIE

**SECURITY**KNOX  
RANGERS

|  |  |
| --- | --- |
| **Data Lake Tier Name** | **Technology Used** |
| Ingestion Tier | Apache Flume  HDFS Copy  Apache Sqoop |
| Storage Tier | HDFS |
| Insights Tier | Apache Zeppelin, Hive QL |
| Operations Tier | Apache Ranger  HDFS Permissions |

**OOZIE**

This tools provide better administration capabilities, as well as improved logging, auditing, and retry functionality.

**OBJECTIVES:**

**360-degree view of a customer**

information about a customer's attitude, behavior, preferences and static data such as date of birth, and presents it as a single integrated view

**Churn detection**

**DATA SOURCES**

DBMS

Transactional System

POS

Web Logs

it\_finance

phoenix

cardpro

bafl\_dwh

bafl\_dw

**Columns**

**LIST OF IMPORTANT FEATURES TO EXTRACT**

id

first\_name

last\_name

email

country

ip\_address

gender

userid

currency

job title

company name

creditcardnumber

domain name

latitude

longitude