iykra Data Fellowship

IYKRA
Data Fellowship Program

(Cloudera Hadoop)

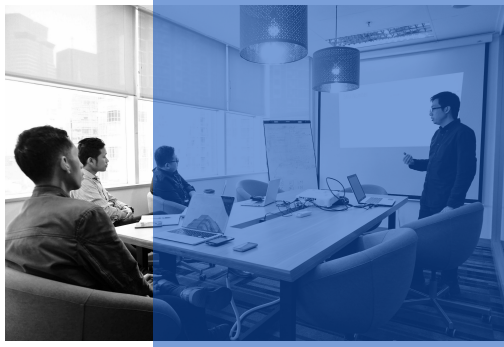
by (Bangun Sasongko)

(date of delivery)

(week1day3)

Trainer Profile

Bangun W. Sasongko
Data Engineer at BRI
Ex. Devops Engineer
Bachelor of Physics, Electrical and Instrumentations
@sasongkobgn
sasongkobgn@gmail.com



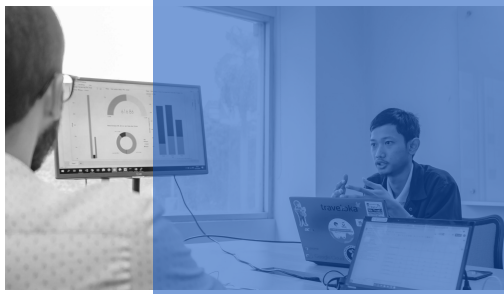
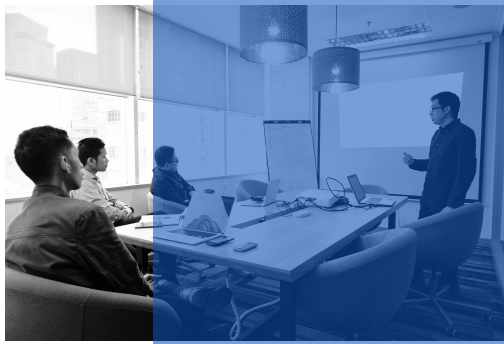
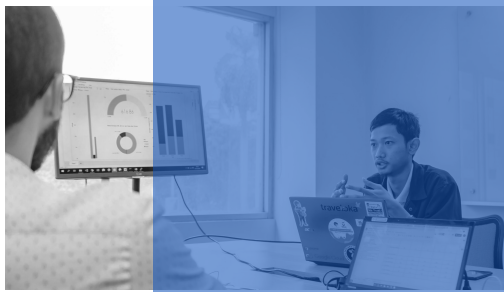


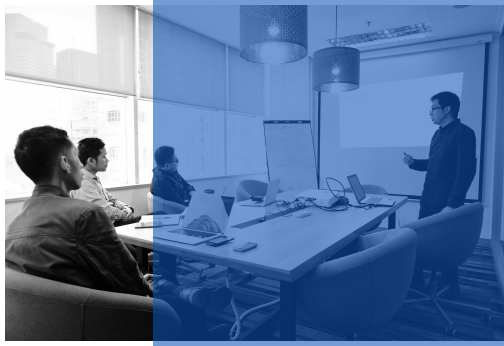
Table of Content

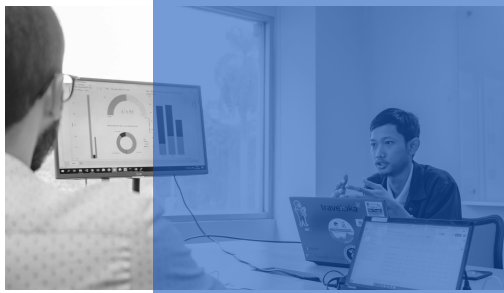




About this Course

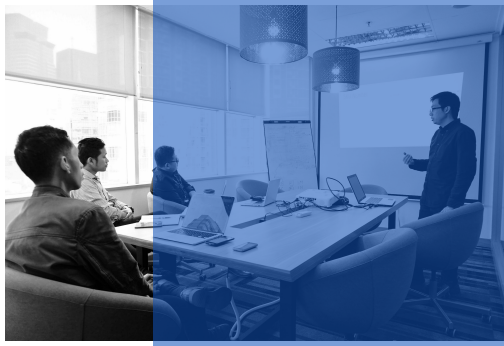
In this course, we will learn about the Hadoop ecosystem and its derivatives, as well as the technologies used in enterprises, especially in the Cloudera Data Platform
Hadoop ecosystem
Mapreduce, Sqoop, Spark
Hbase, Hive

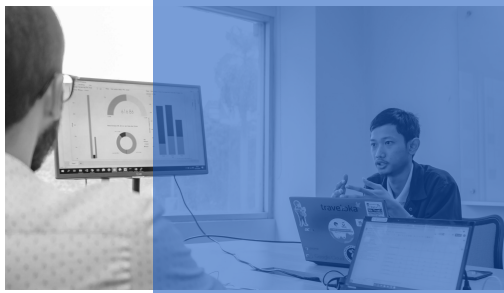




The Objectives

- Hadoop ecosystem concept
- Get insight about processing framework on big data
- Get preliminary concept about analytics database and operational database at hadoop ecosystem





Data Fellowship

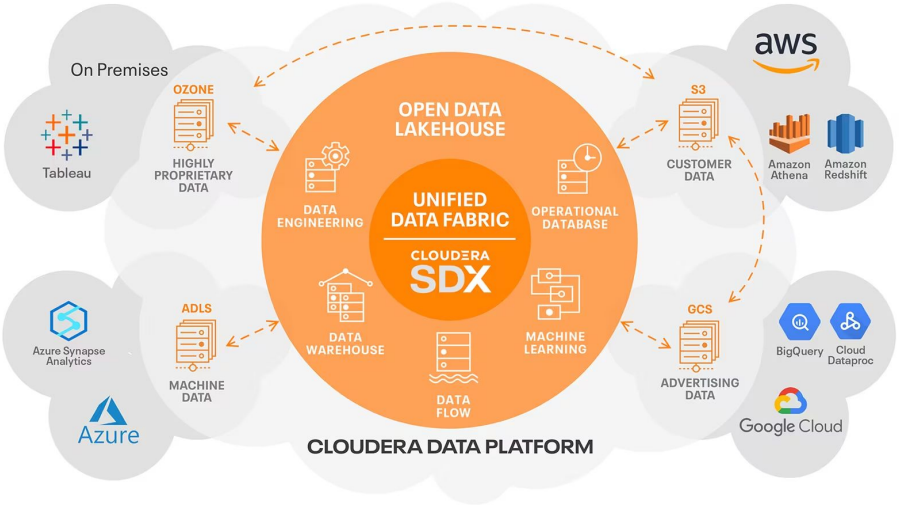
Apache Hadoop Ecosystem

What is Apache Hadoop?

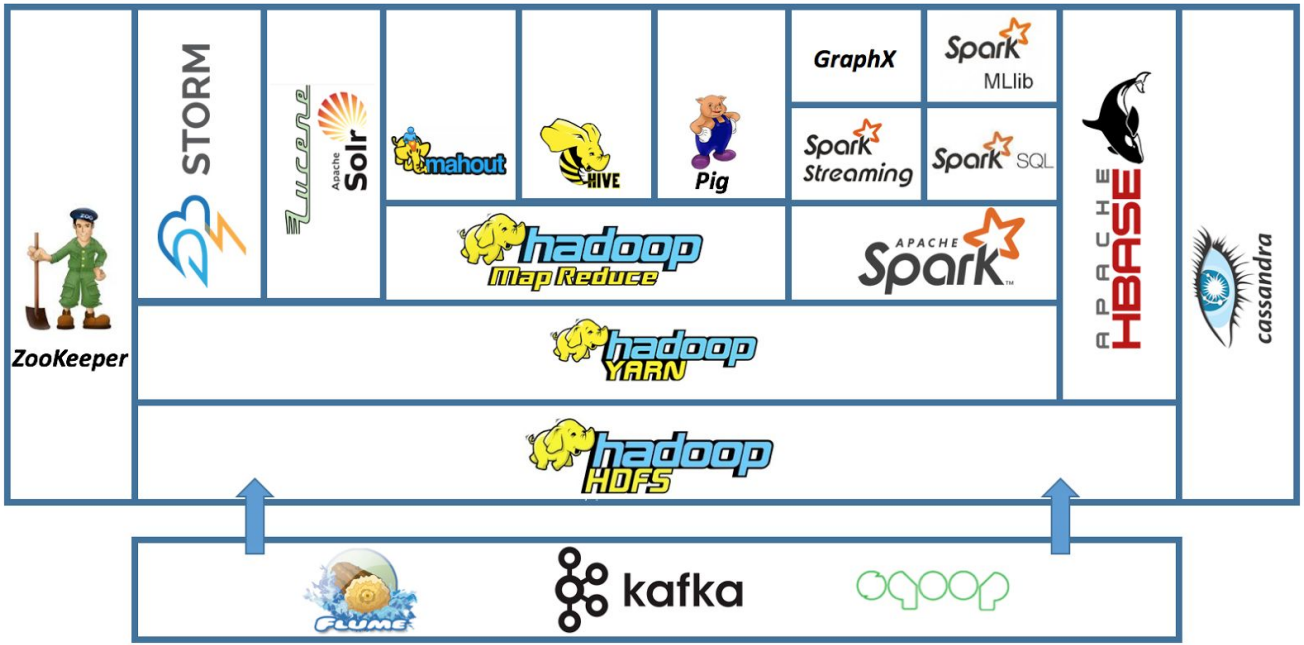
Hadoop is an open-source Java framework for distributed applications and data-intensive management. It allows applications to work with thousands of **nodes** and **petabytes** of data. Hadoop was inspired by Google's MapReduce, GoogleFS and BigTable publications.

What about Cloudera?

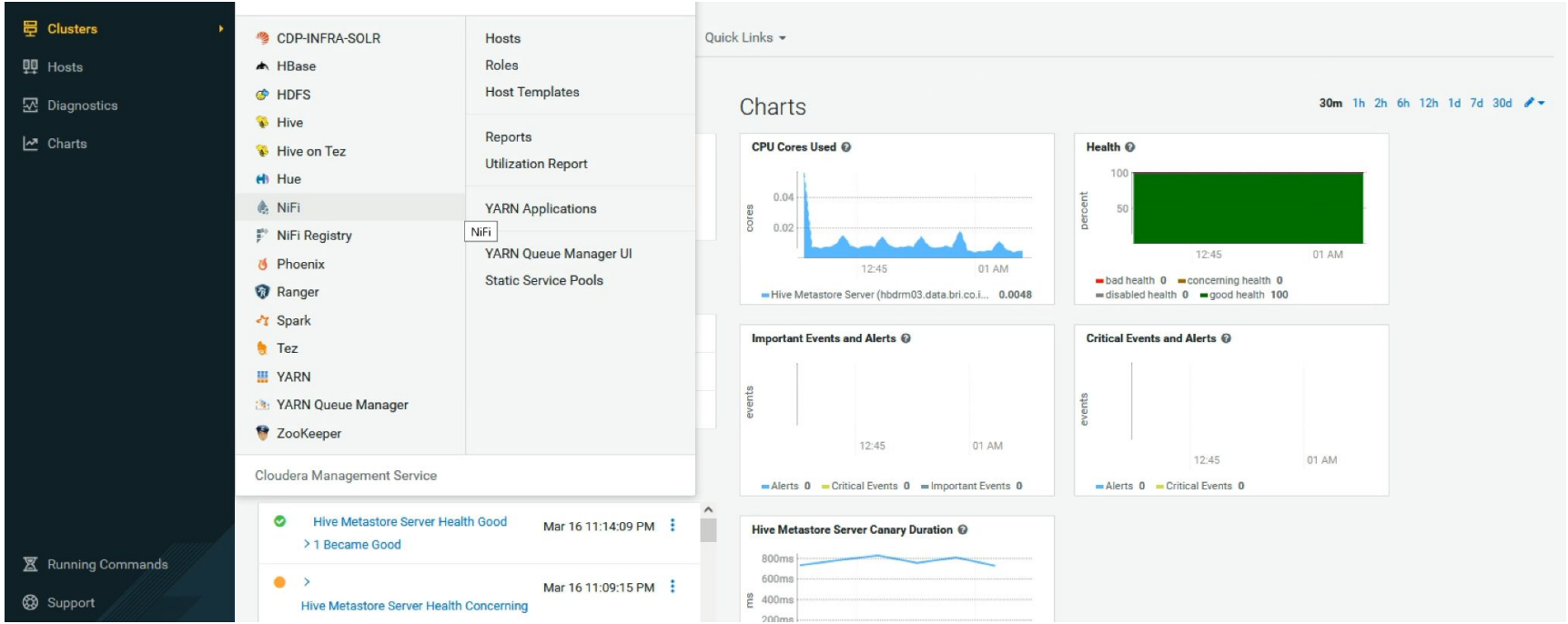
Cloudera was co-founded in 2008 by mathematician Jeff Hammerbach, a former Bear Stearns and Facebook employee. He was in charge of data analysis and developing programs for better advertising targeting. The other co-founders are Christophe Bisciglia, an ex-Google employee. Amr Awadallah, an ex-Yahoo employee who also worked on Hadoop, and Mike Olson, CEO of Cloudera. The chief architect is Doug Cutting, behind the Lucene indexing engine and the Hadoop distributed framework.



Hadoop ecosystem heuristic map



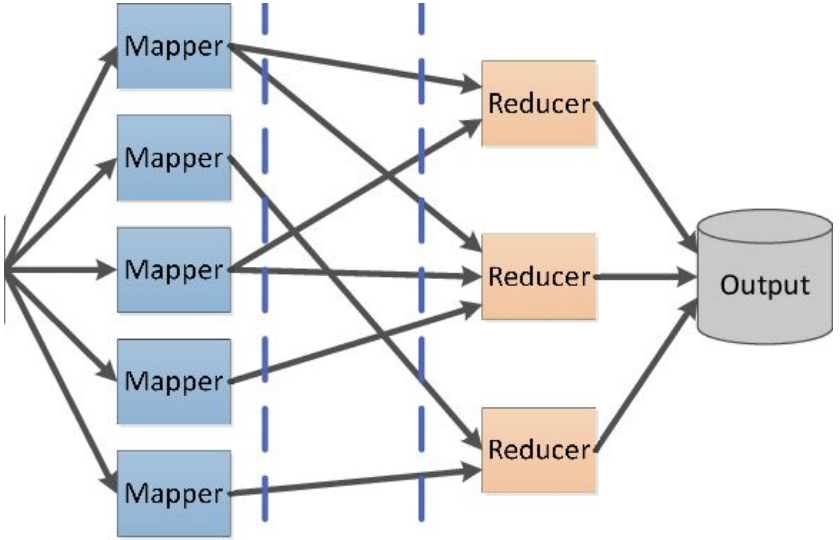
Cloudera Manager



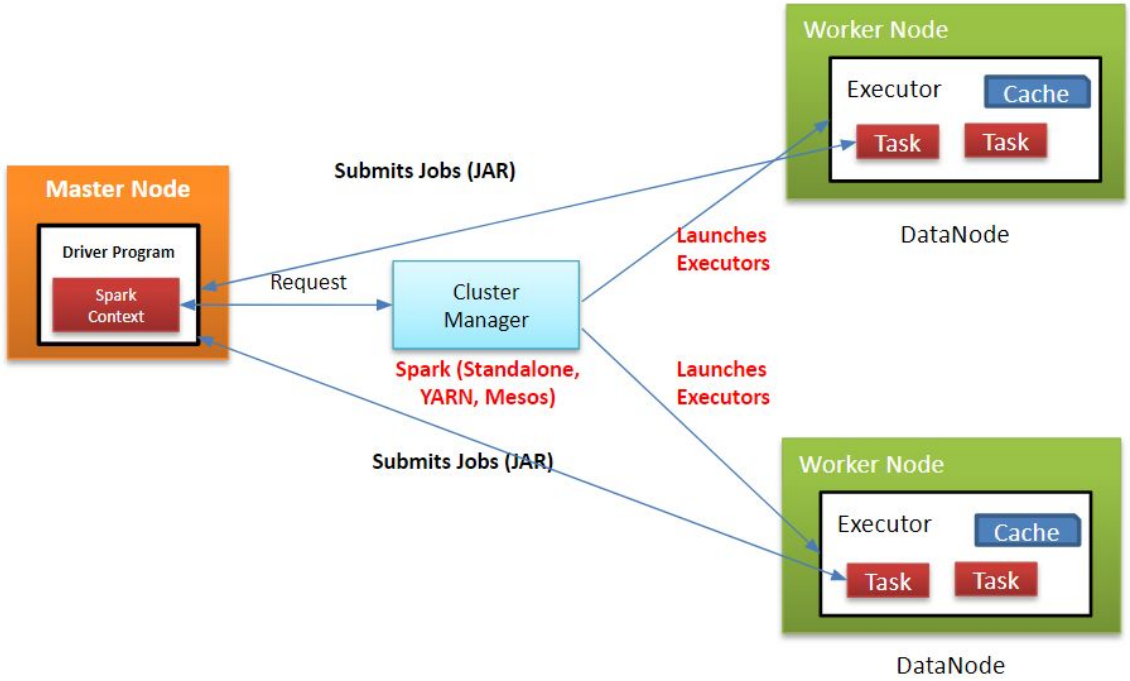
Apache Spark

- Data is loaded into memory, performed much faster since accessing data from memory is orders of magnitude quicker than fetching it from disk

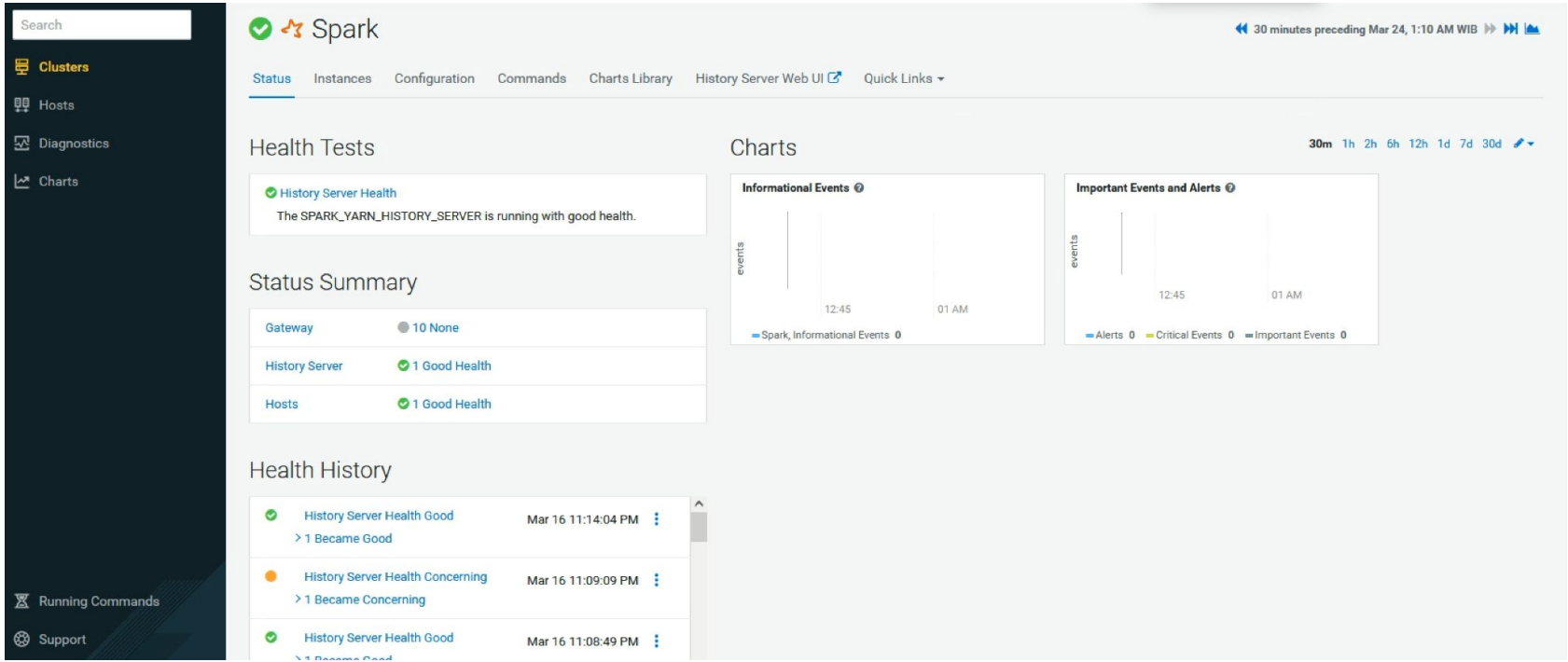
- Compared to MapReduce, which works in batch mode, Spark's computation model works in interactive mode, i.e., assembles the data in memory before processing it and is therefore very suitable for Machine Learning processing.



Apache Spark



Apache Spark at cloudera



Apache Hive

Hive is a computing infrastructure similar to the Data Warehouse that provides query and aggregation services for very large volumes of data stored on a distributed file system such as HDFS

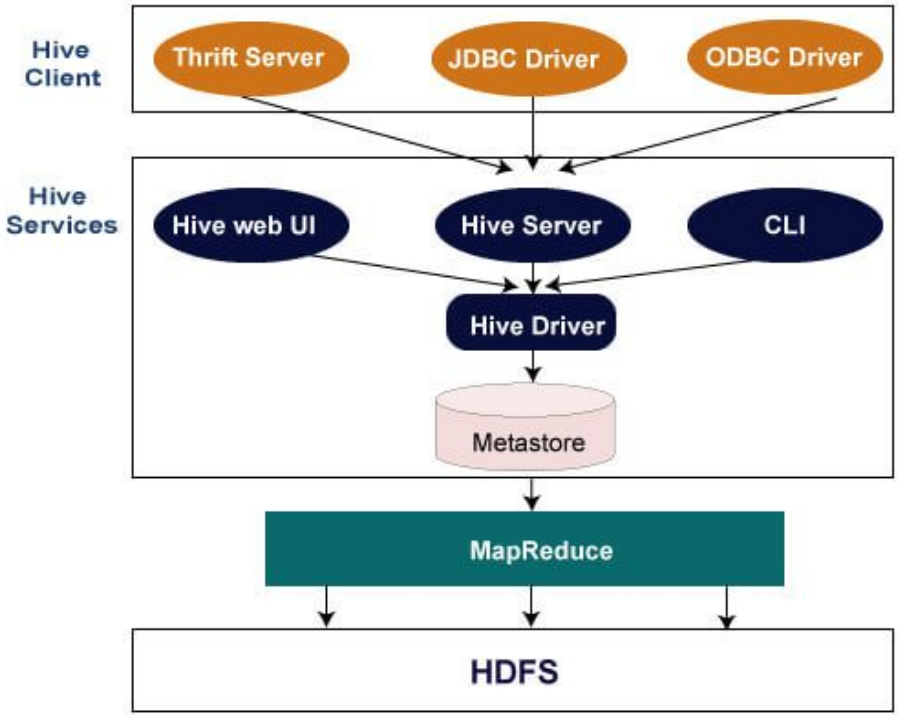
Hive offers a SQL-based (ANSI-92 standard) query language called HiveQL (Hive Query Language)

HiveQL also allows advanced users/developers to integrate Map and Reduce functions directly into their queries to cover a wider range of data management problems

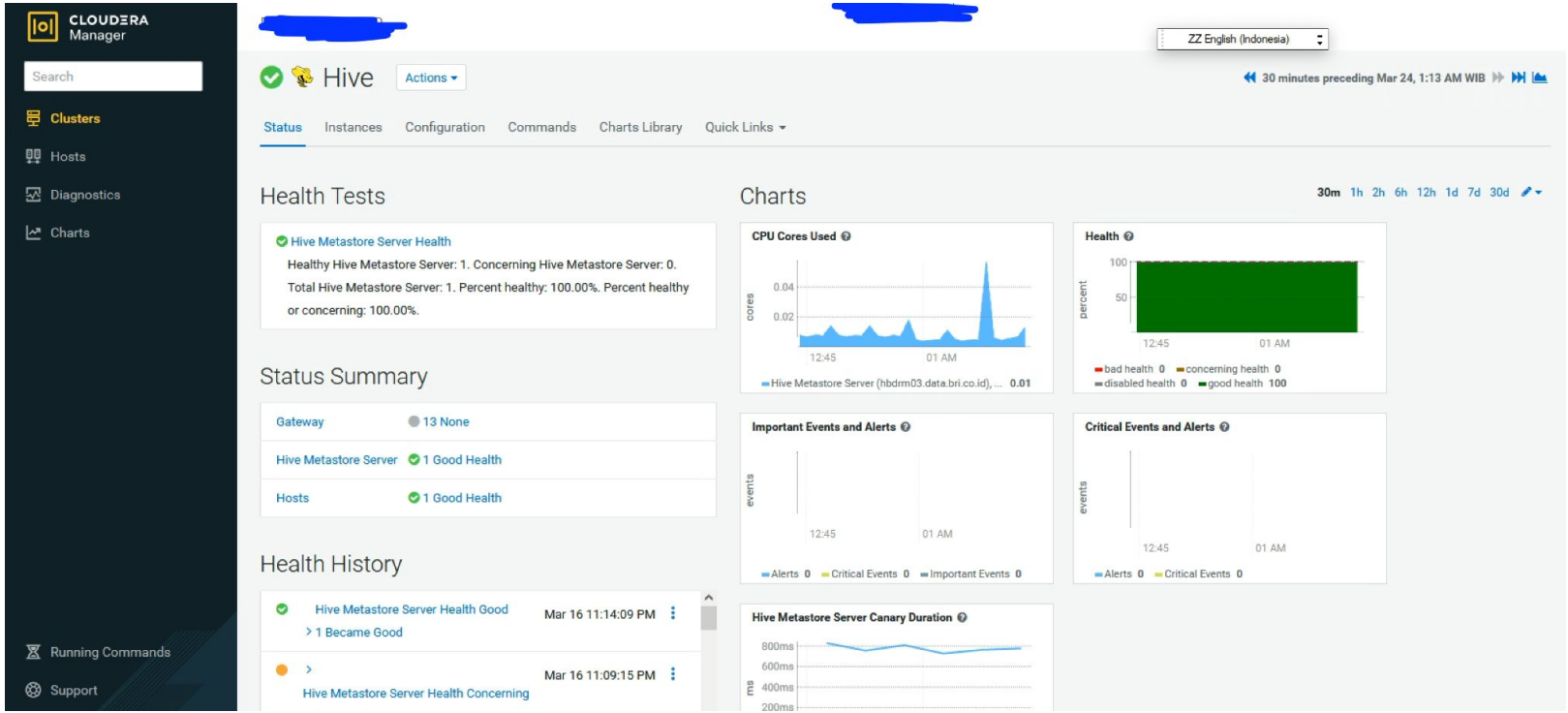
On board, give explanations simple architecture on hadoop how hive is used

Hive architecture

When you write a query in HiveQL, that query is transformed into a MapReduce job and submitted to the JobTracker for execution by Hive.



Cloudera Apache Hive



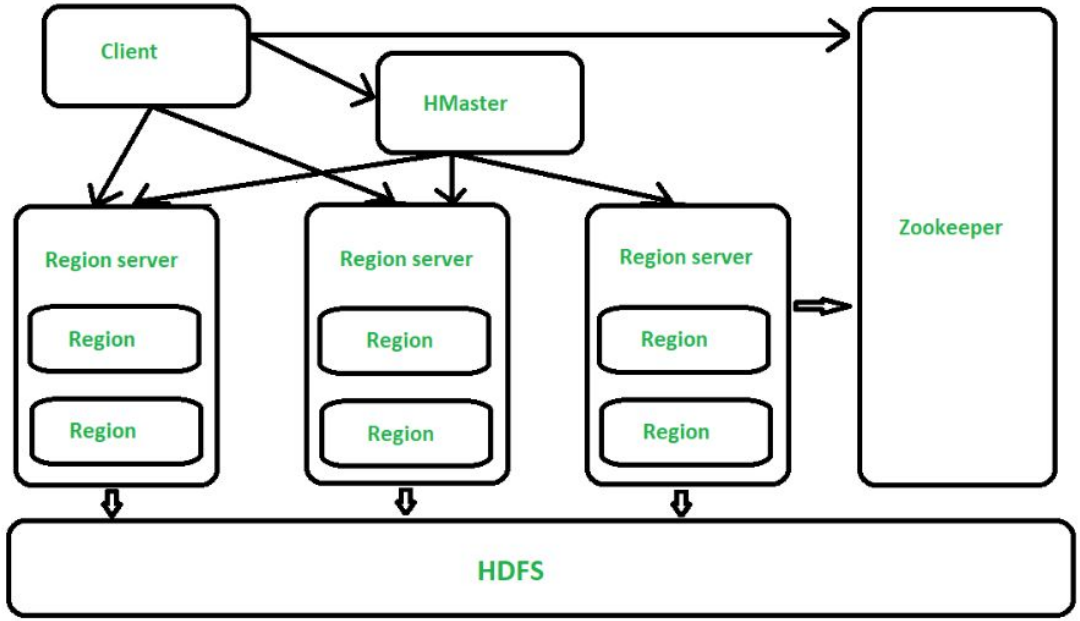
Apache Hbase

At Cloudera, HBase is referred to as operational databases, which is combined with Apache Phoenix as a translation layer."

- Distributed and Scalable
- Column-Family Data Model
- Built for Big Data
- Hadoop Ecosystem Integration
- Highly Available and Fault-Tolerant
- Use Cases

On board, give explanations simple architecture on hadoop how **hbase** is used

Hbase architecture



Hadoop ecosystem: Summary

What have we learned?

- Apache Hadoop
- Apache Hive
- Apache Hbase

(impala, storm, oozie, zookeeper, sqoop) define later

References

1. Hadoop definitive guide: storage and analyst at internet scale
2. https://www.ibm.com/blog/hadoop-vs-spark/
3.

iykra Data Fellowship

Thank you!