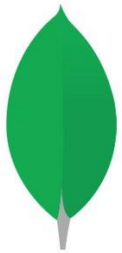
Wide-Column Databases

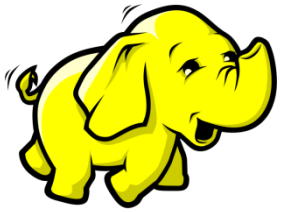
Michael Enudi

Journey through the world of databases and data engineering



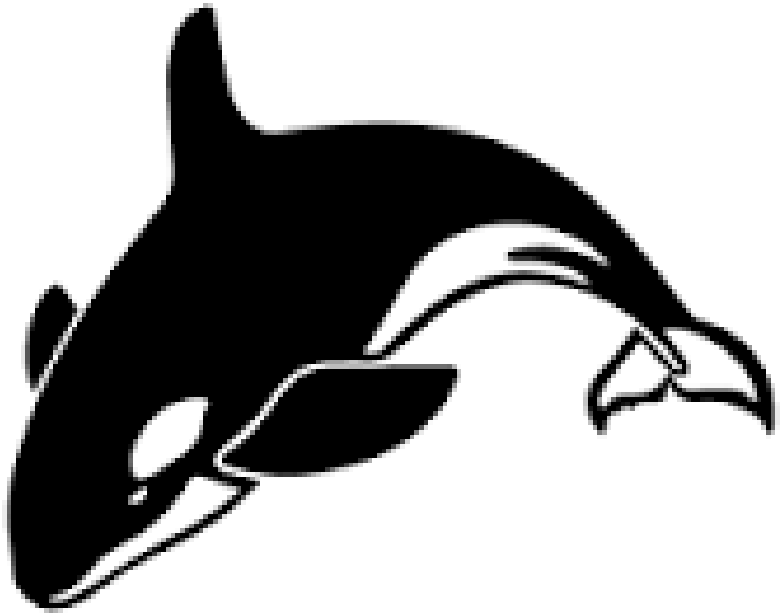




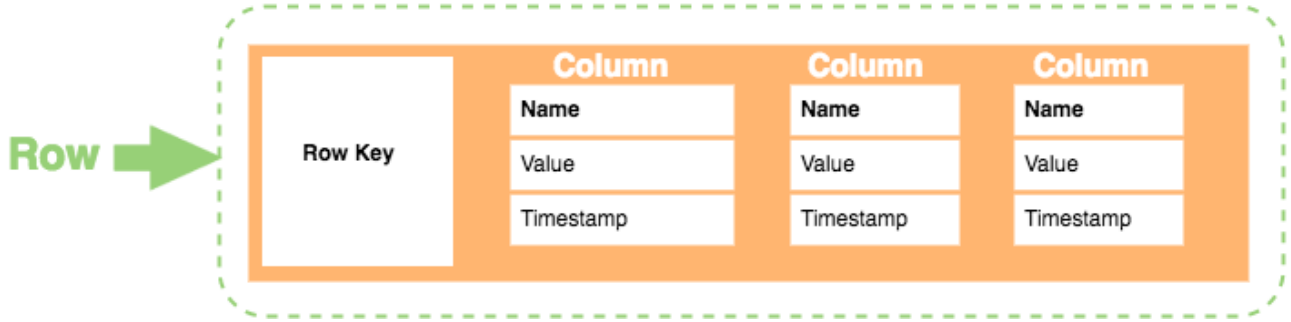




Scope

- Introduction to Columnar databases.
- HBase
- Apache Zookeeper
- Movielens Data in HBase
- Performing CRUD in HBase
- SQL on HBase - Apache Phoenix
- Demo : GeoLife GPS Trajectories
- Wide Column Store: Wrap Up.



Wide-Columnar Databases



Google BigTable

Amazon DynamoDB

cassandra

DATA MODEL

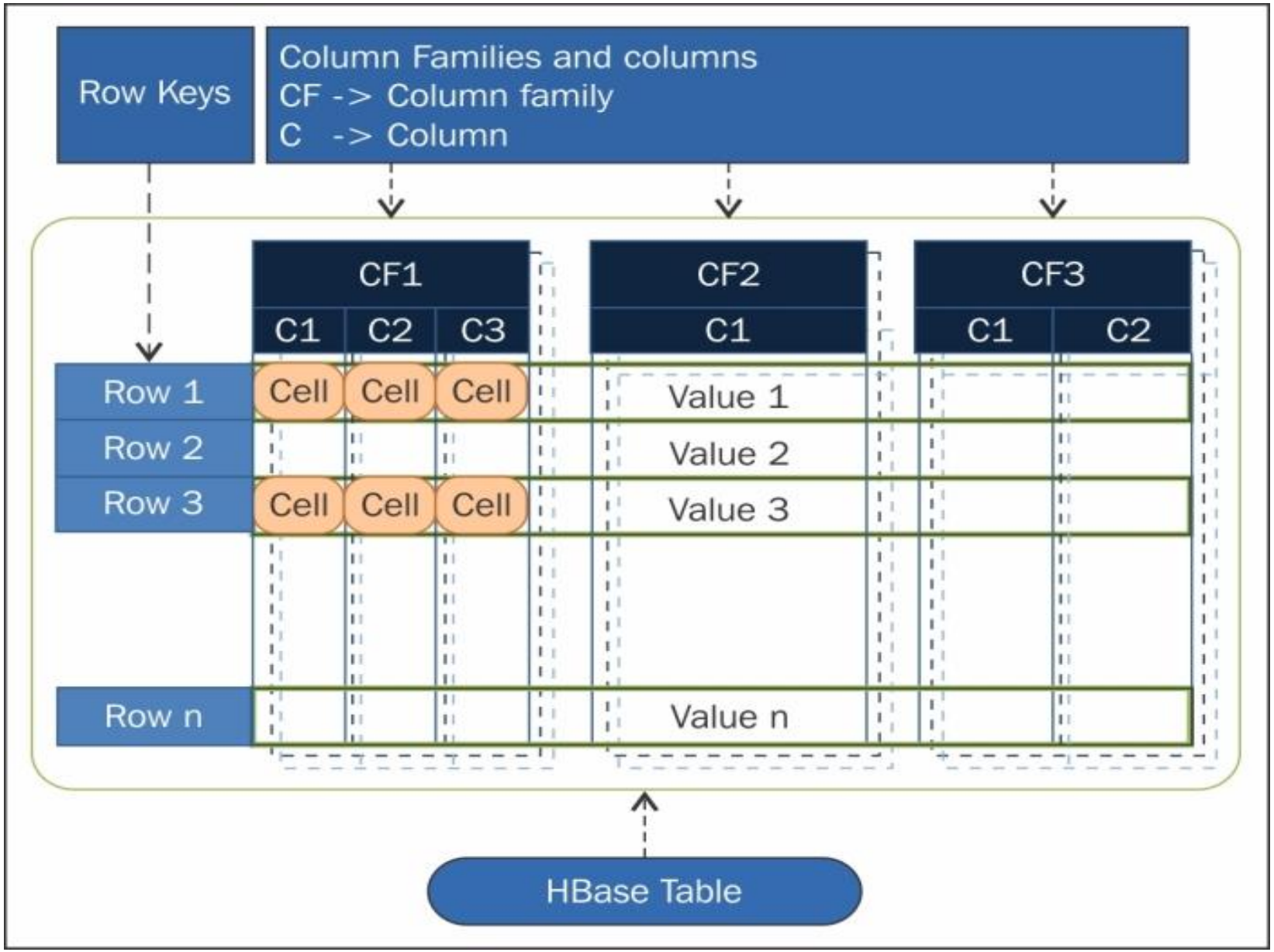
(TABLE, ROWKEY, [COLUMN FAMILY], COLUMN, TIMESTAMP) ⟶ VALUE

SCYLLA.

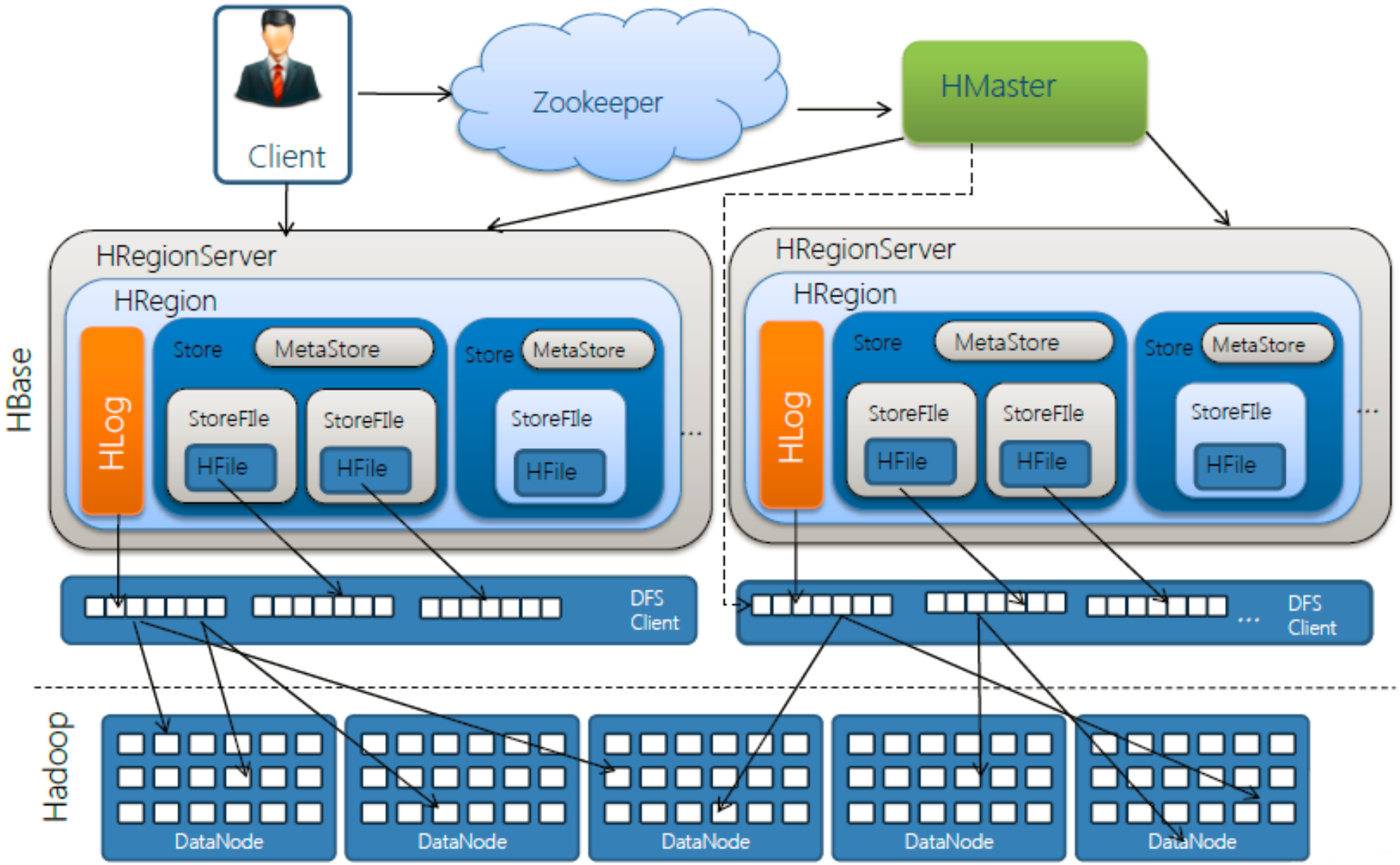
accumulo™

- Modeled after Google's BigTable. provides super fast lookup and range query over billions of records
- stores data on HDFS.
- is linearly scalable.
- Consistency + Partition Tolerance
- benefits from HDFS features such as data reliability, scalability, and durability.
- does not support SQL natively
- offers Java, thrift and Rest API for client access.
- can be integrated with other tools in the Hadoop ecosystem for processing or analytical workloads.
- written with Java.
- deployed by a list of high-profile enterprises that include Netflix, Pinterest, Salesforce, Spotify, Yahoo, Adobem Bloomberg, Alibaba, Amadeus, Mozilla and more.

HBASE MODELLING



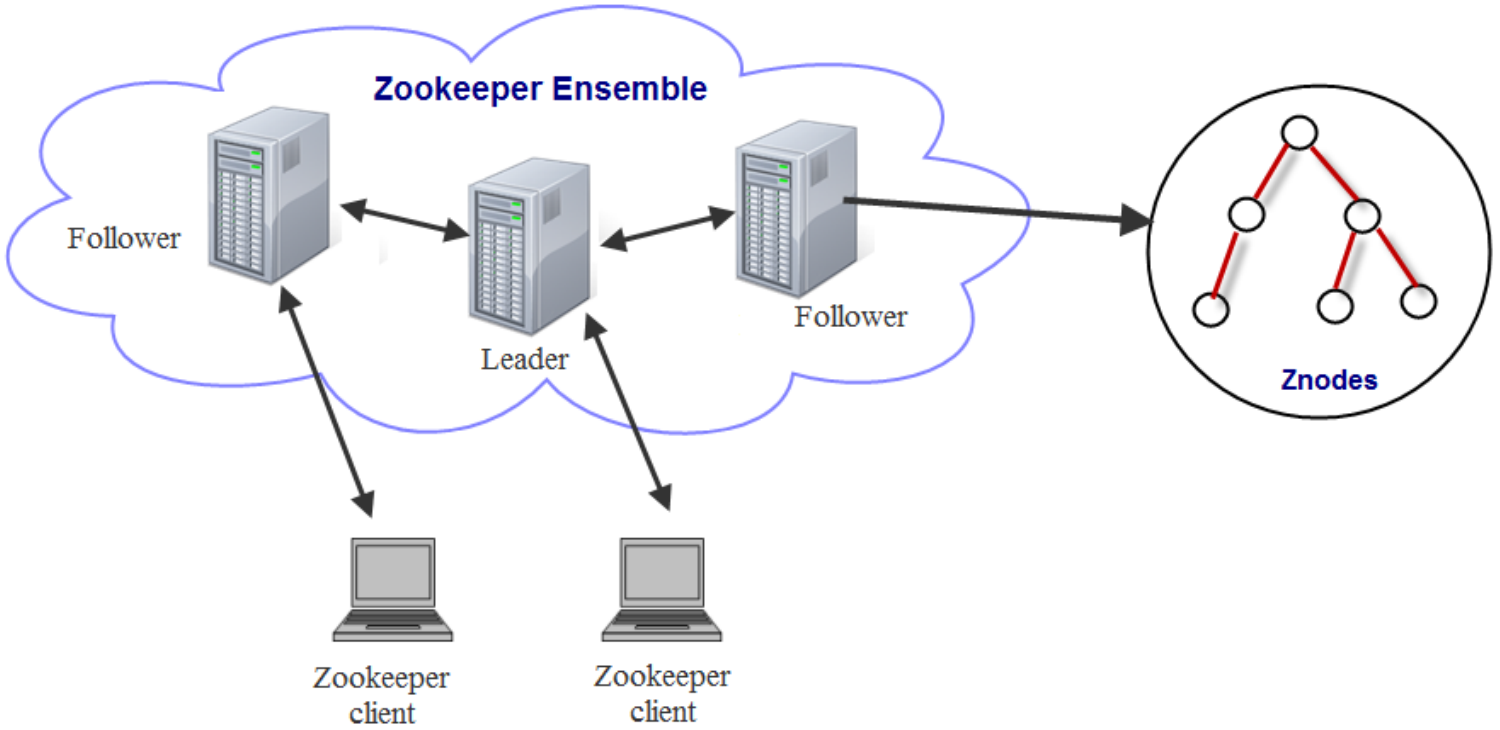
HBASE ARCHITECTURE



APACHE
ZooKeeper™

ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. All of these kinds of services are used in some form or another by distributed applications.

- Official Website.



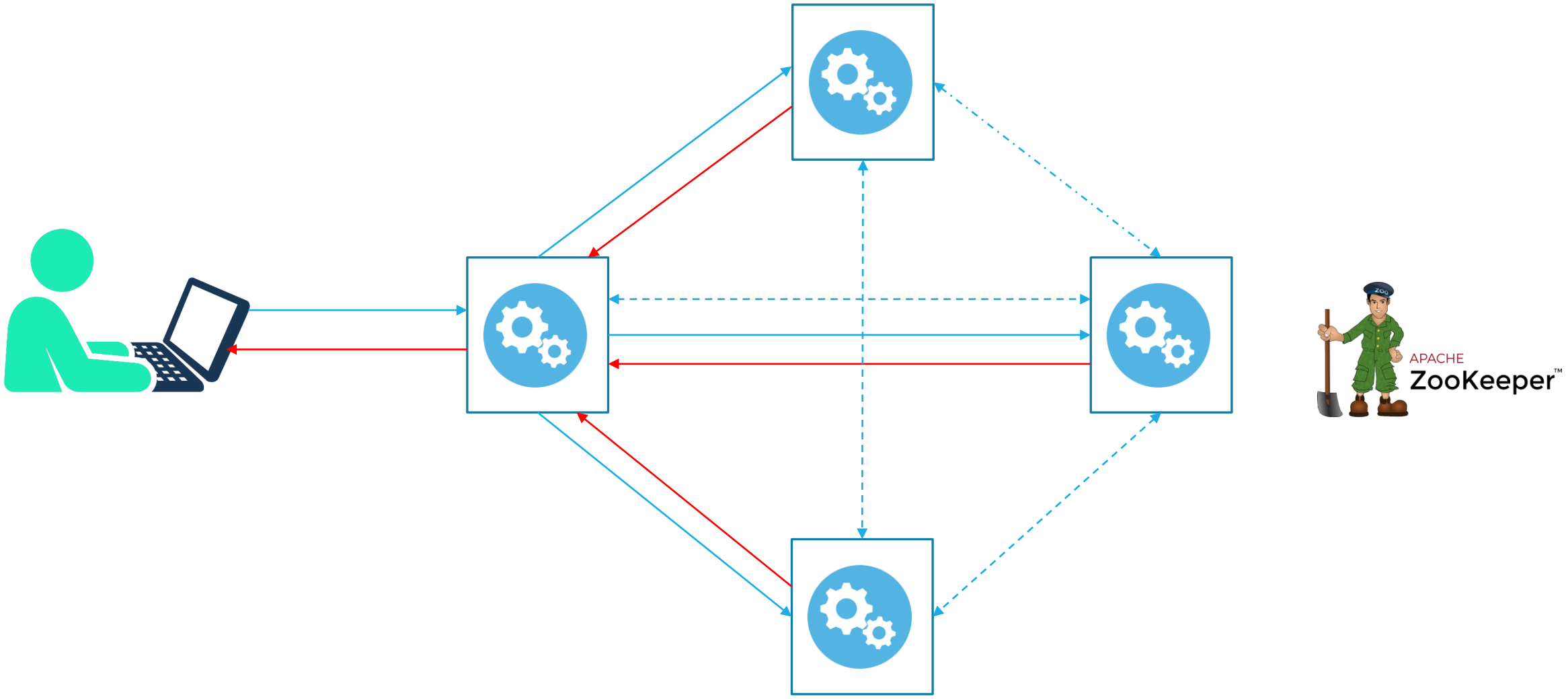
Used By :-

hadoop HDFS

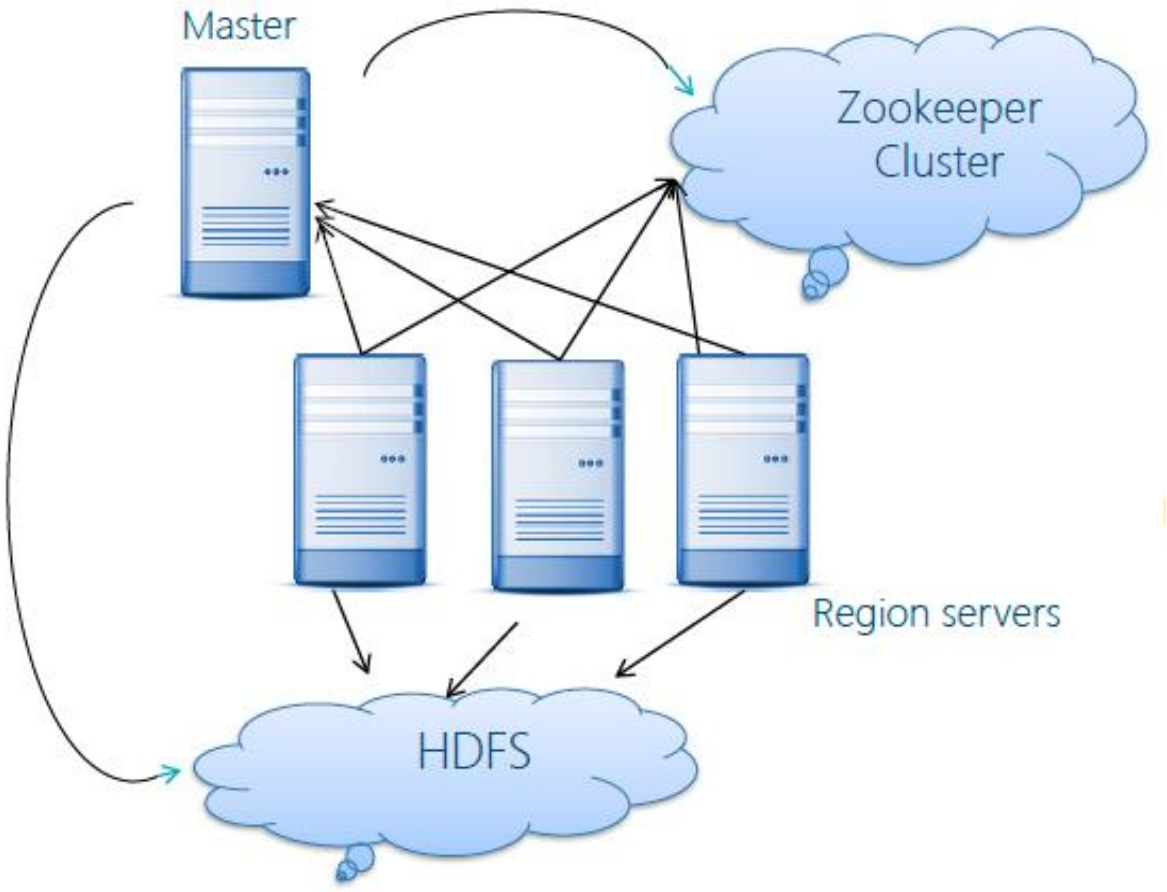
APACHE HBASE

APACHE kafka®

DISTRIBUTED SYSTEMS & ZOOKEEPER



HBASE AND ZOOKEEPER



- HBase uses Zookeeper extensively for
  - region assignment
  - tracking active and healthy servers
  - electing a new active master in case of the failure of the active master.
  - ensuring that only one master is active at a time
  - storing shared state or variables across cluster
- HBase can manage Zookeeper daemons for you (in development) or you can install/manage them separately (in production)

MOVIELENS MODELLING

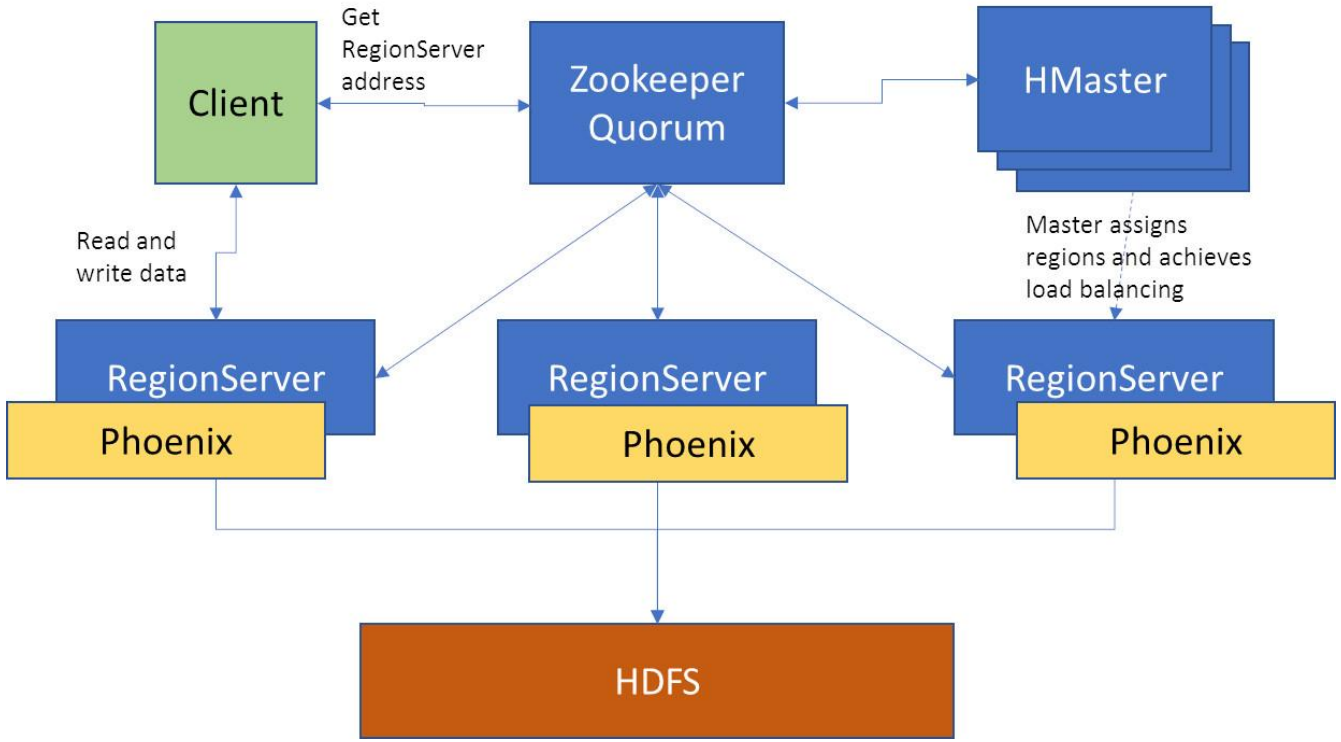
DATA MANIPULATION

- **HBase API**
  - Get(row)
  - Put(row, <column,value>)
  - Scan(key range, filter)
  - Increment(row, columns)
  - Check and Put, delete etc
- **HBase Interface**
  - Java
  - Thrift(Ruby, Php, Python, Perl, C++,..)
  - HBase Shell

APACHE PHOENIX

Apache Phoenix enables OLTP and operational analytics in Hadoop for low latency applications by combining the best of both worlds:

- the power of standard SQL and JDBC APIs with full ACID transaction capabilities and
- the flexibility of late-bound, schema-on-read capabilities from the NoSQL world by leveraging HBase as its backing store



WIDE-COLUMNAR DATABASE WRAP-UP