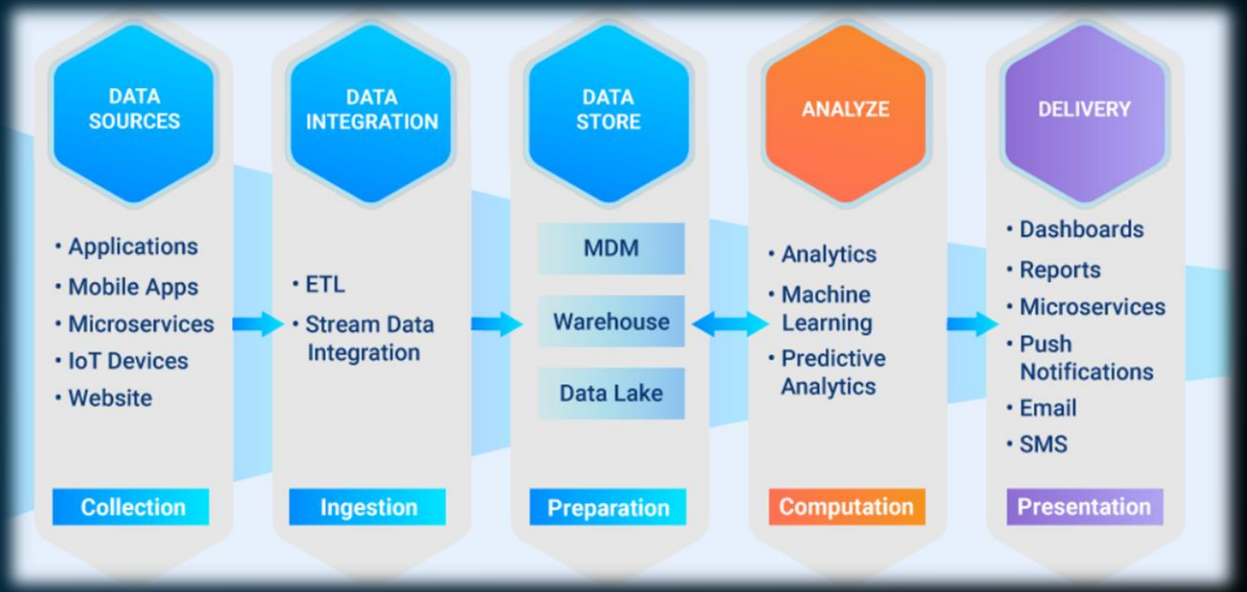
Apache Flume & ZooKeeper

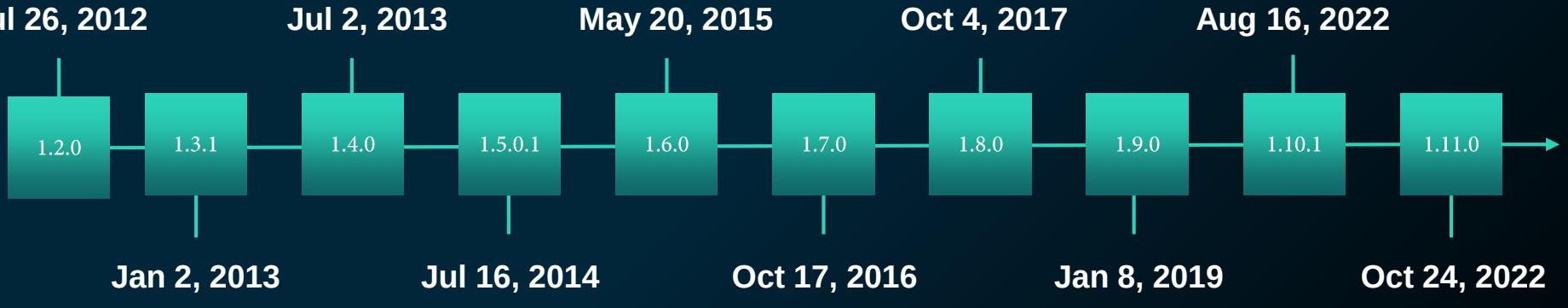
Parham Pishro

Dr. Nasiri

Pipeline of Big Data

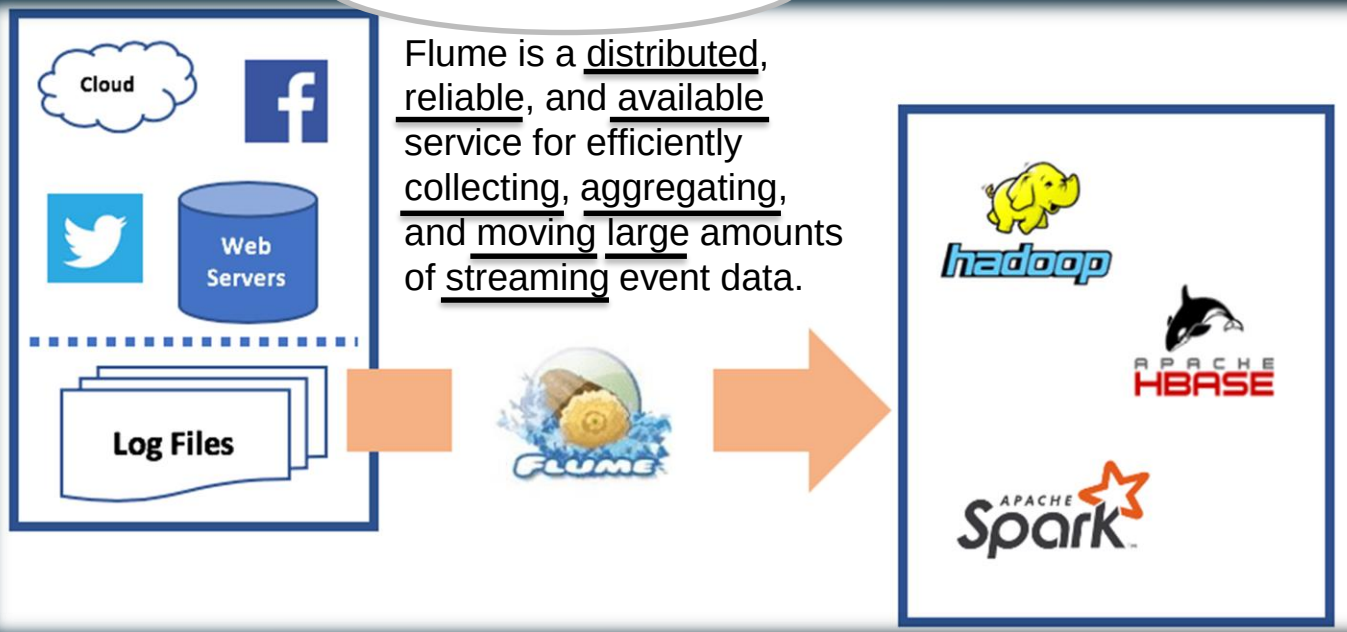


Apache Flume



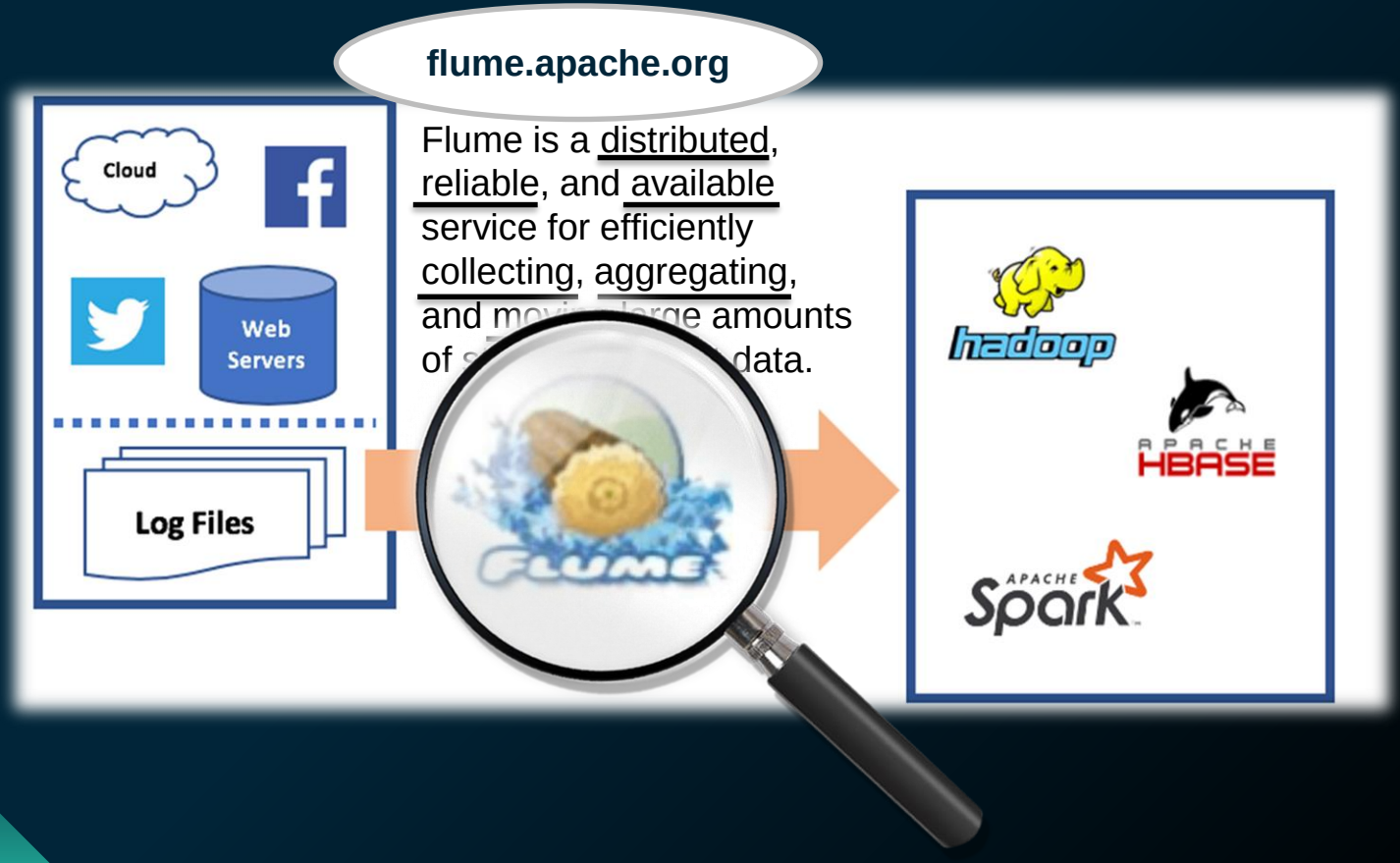
Introduction to Apache Flume

flume.apache.org

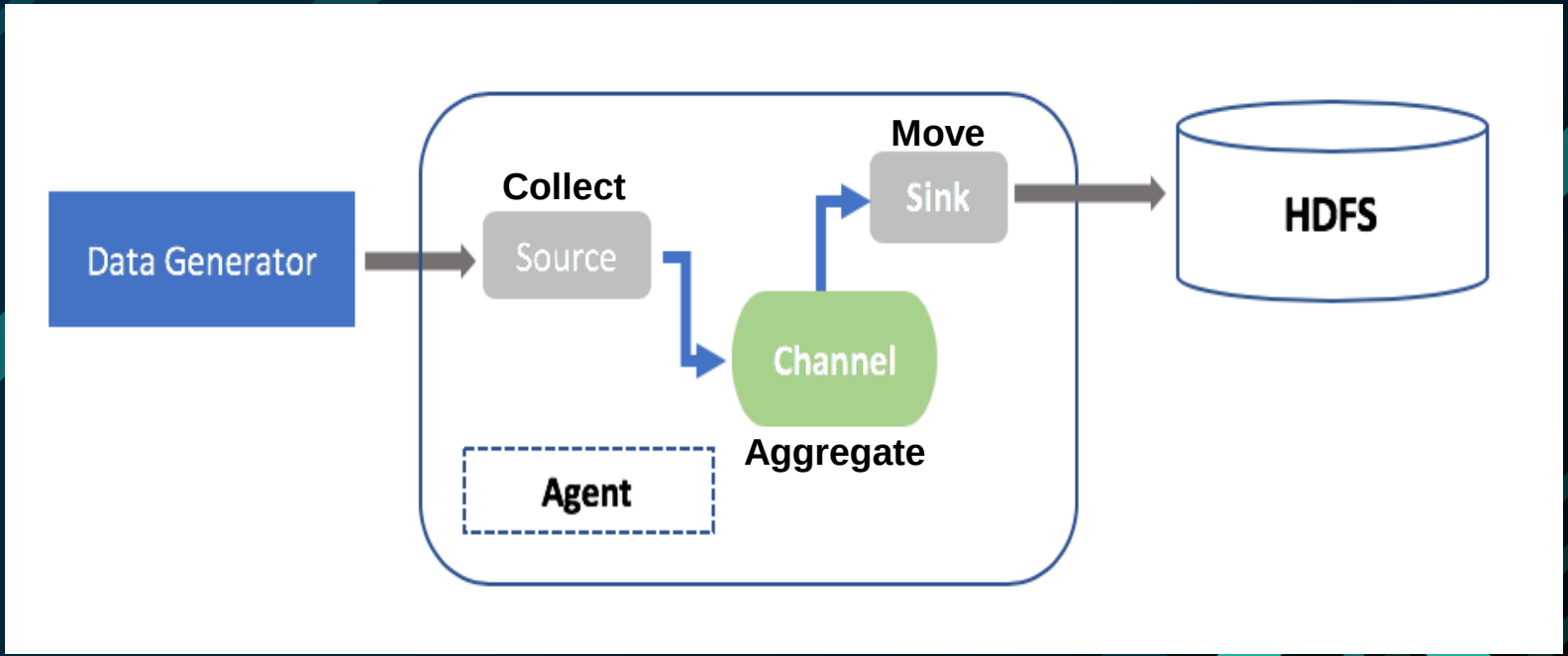


Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming event data.

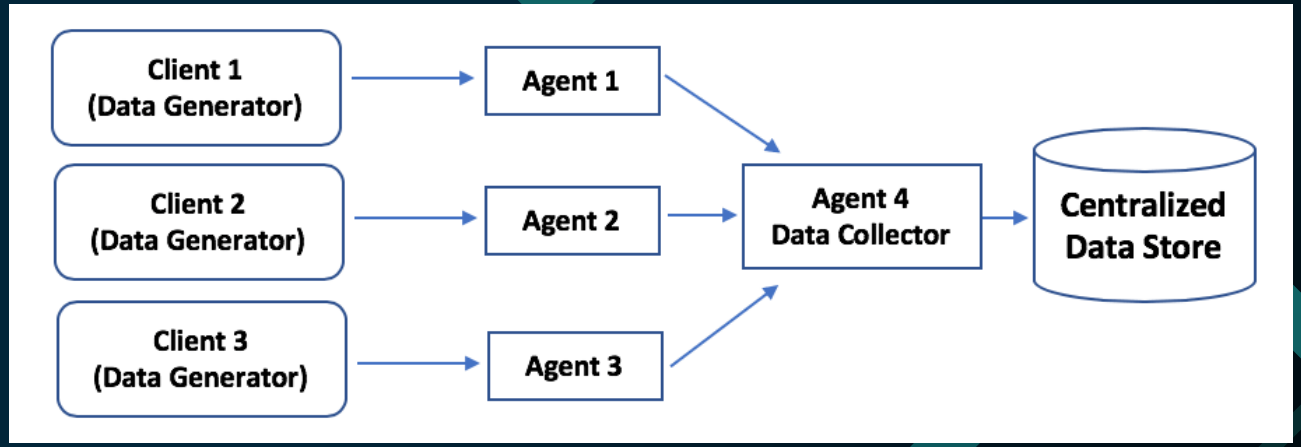
Introduction to Apache Flume

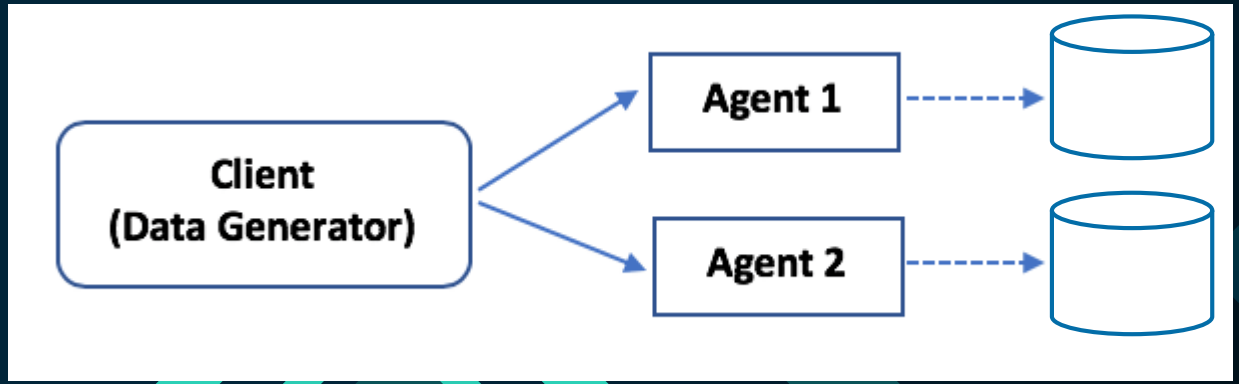


Architecture of Apache Flume

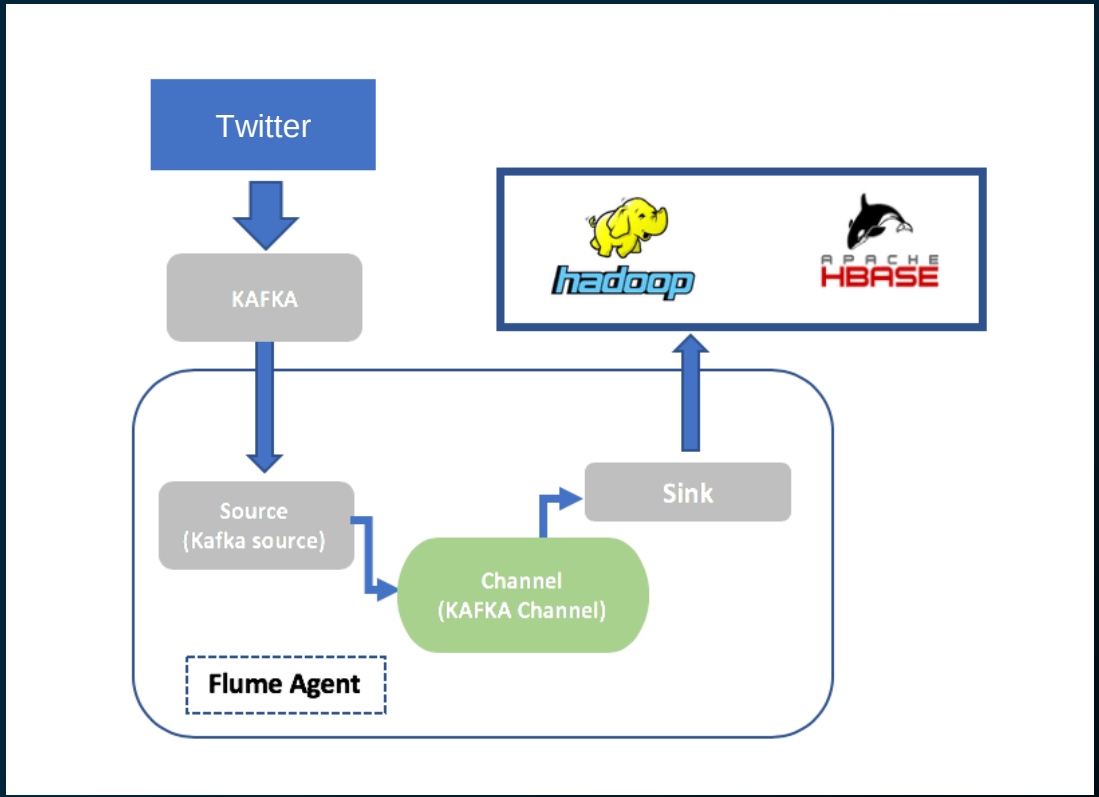


Two Types of Architecture

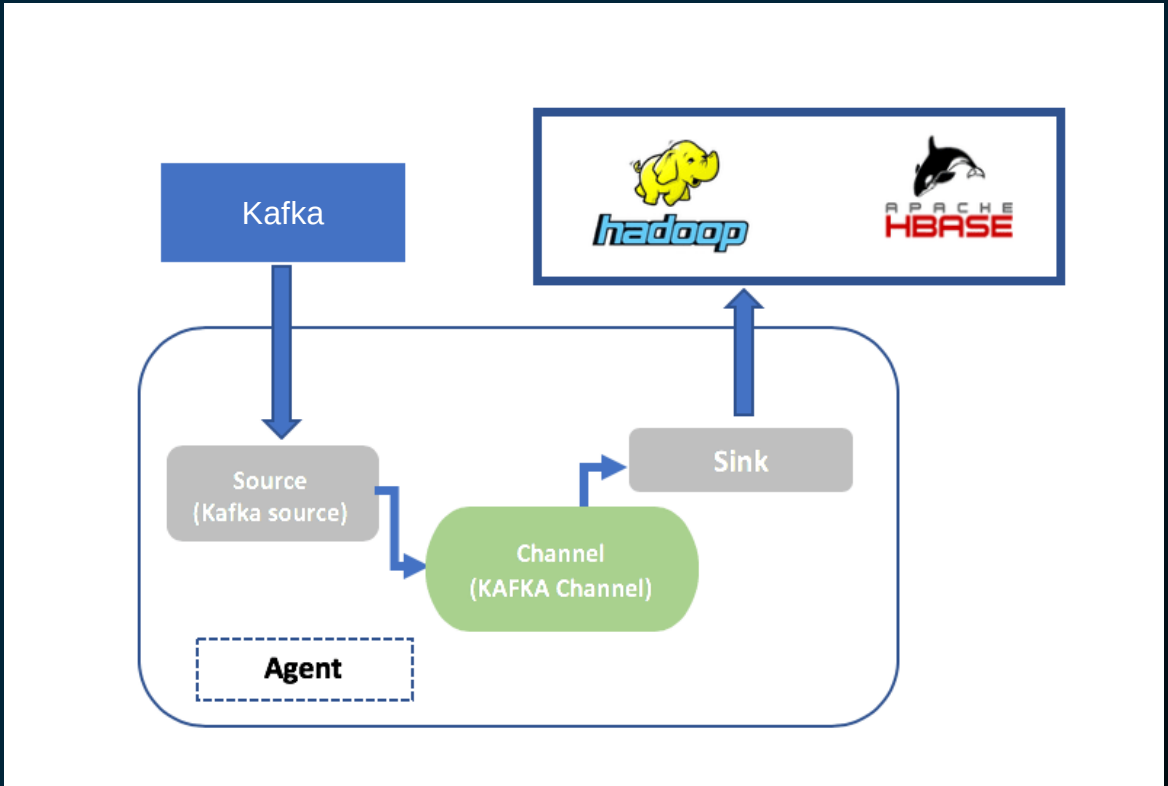




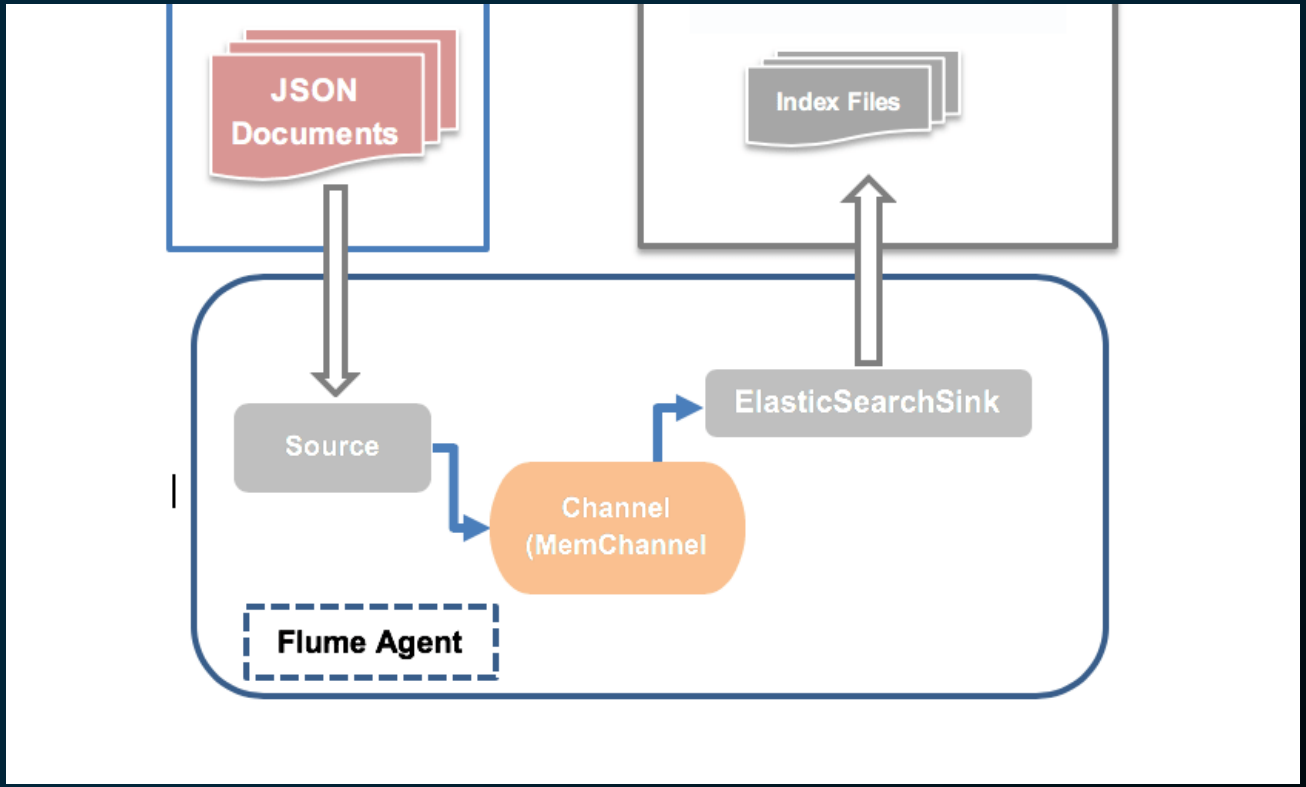
Example
Streaming Log Data to HDFS from Twitter



Example
Streaming Log Data from Kafka to HDFS



Example
Streaming Log Data to Elasticsearch



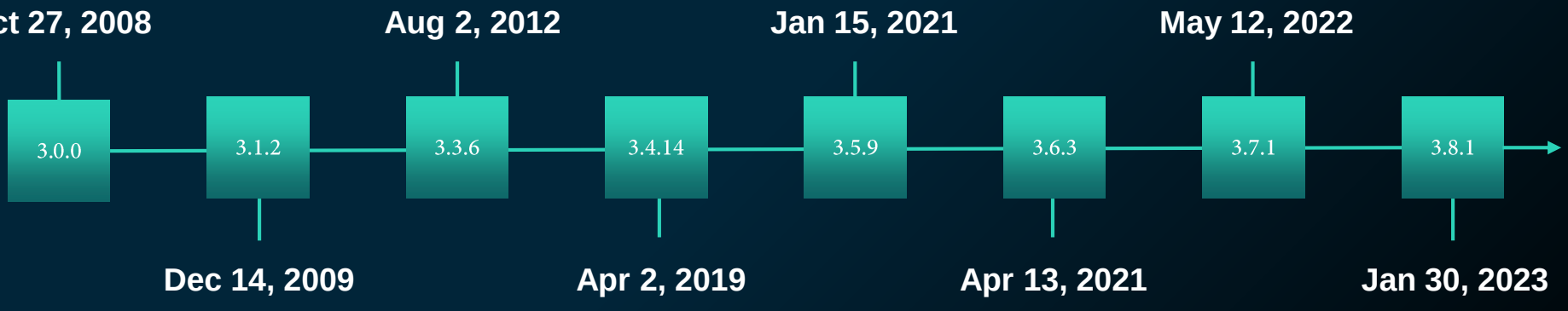
The Limitations of Apache Flume

- Complexity and difficulty of architecture to manage and maintain (From multiple sources to multiple destinations)

- Streaming is not 100% real-time

- Weakness to identify duplicate data

Apache ZooKeeper

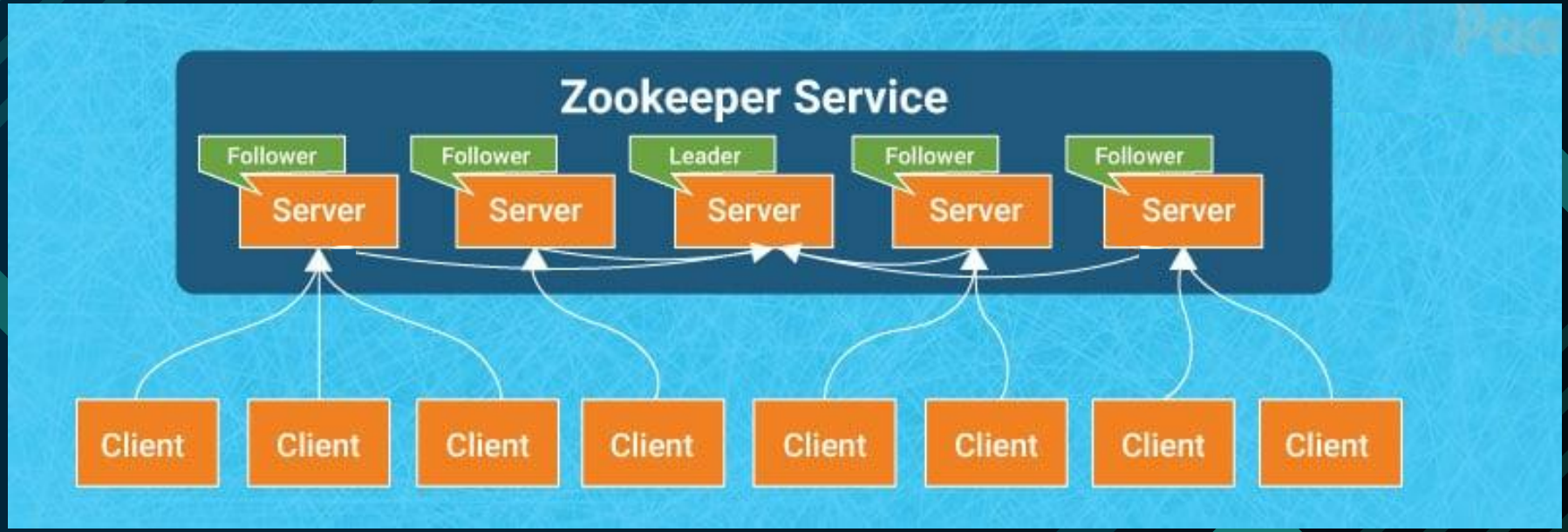


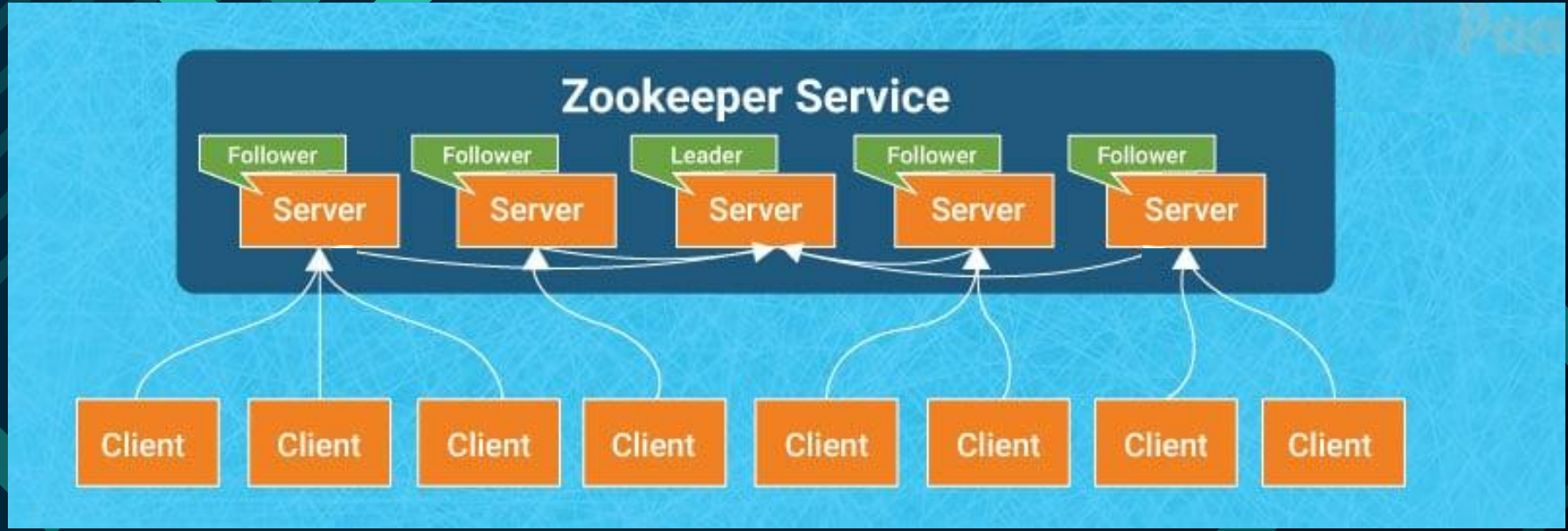
Why Do We Need Apache ZooKeeper?

Coordinating Different Services in Hadoop Ecosystem
(Systems Configuration with Data Synchronization)

Why Do We Need Apache ZooKeeper?

Architecture of Apache ZooKeeper





Server

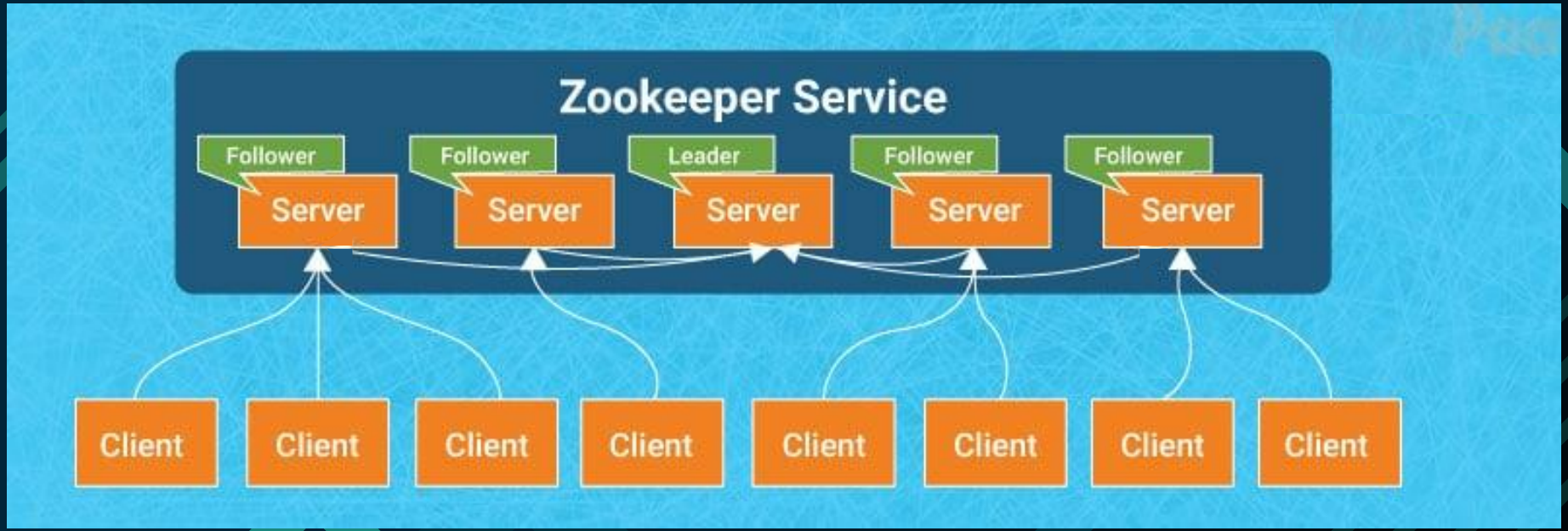
Leader

Follower

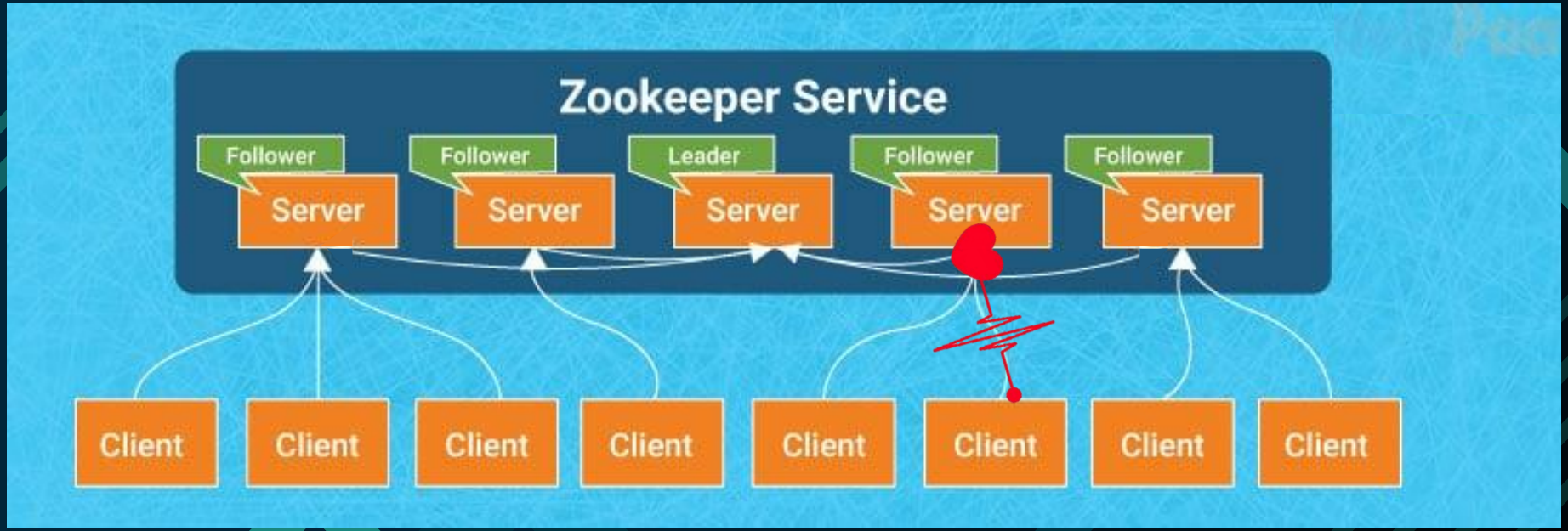
Client

Working of Apache Zookeeper

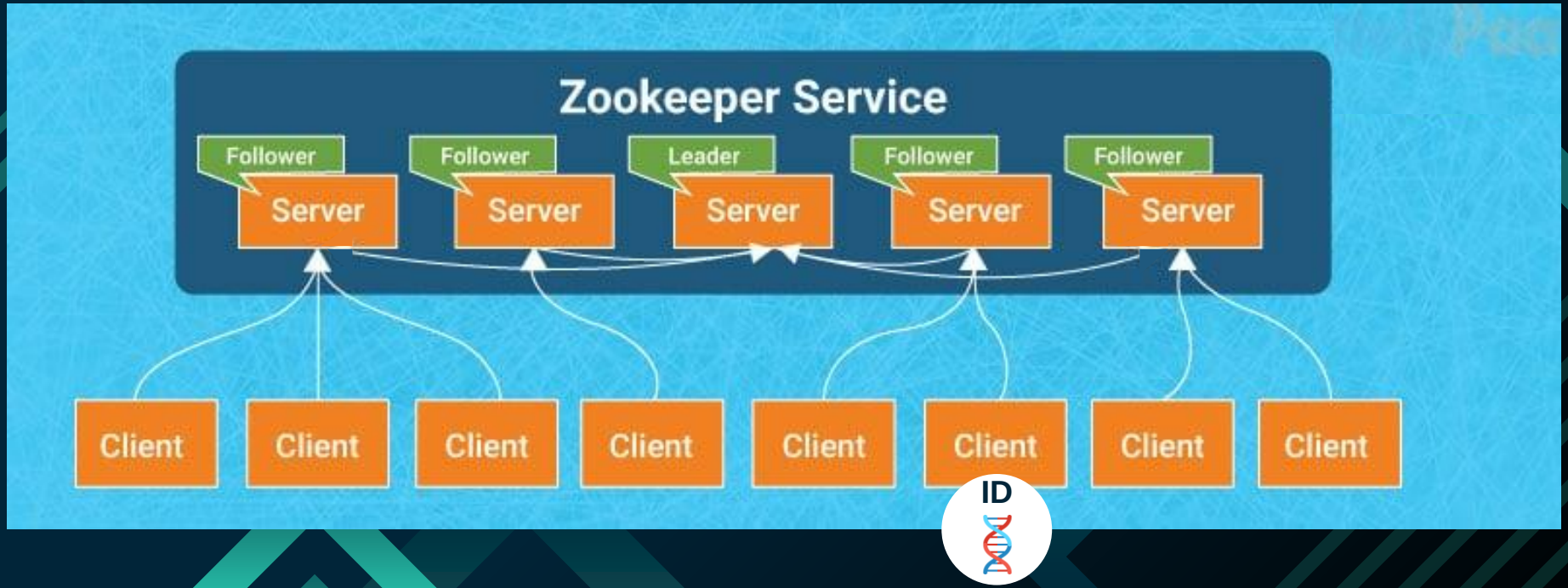
Working of Apache Zookeeper



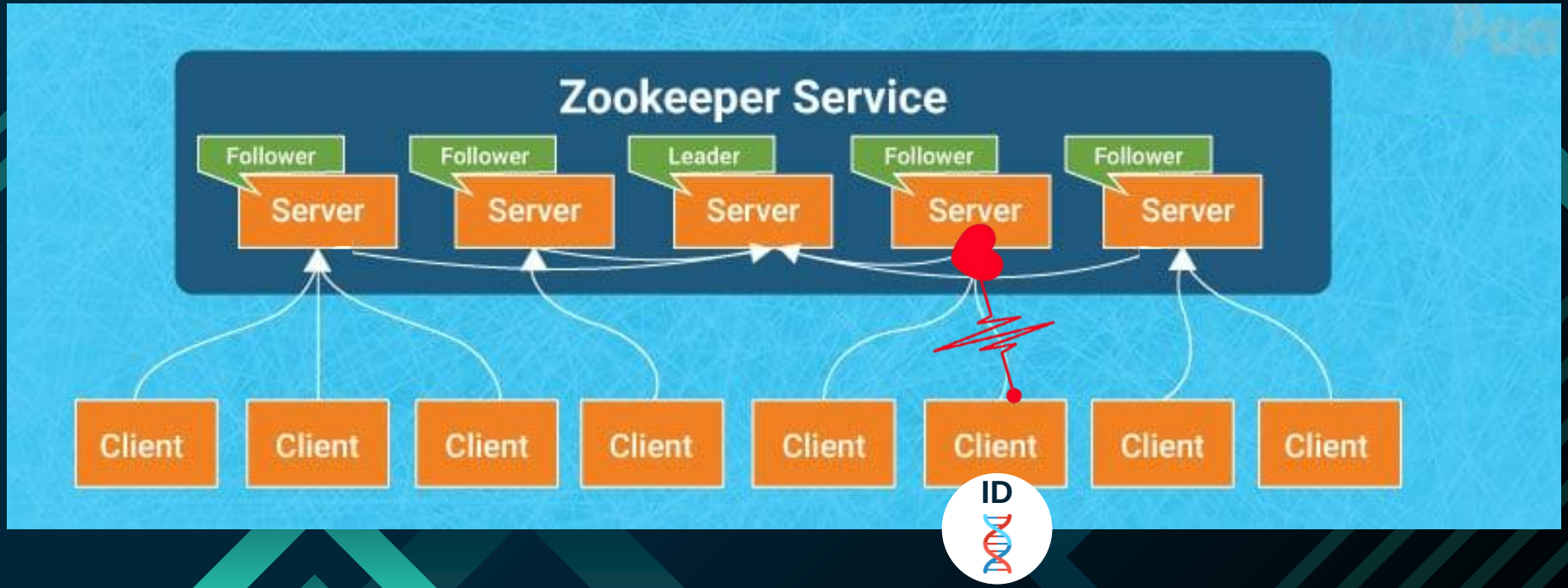
Working of Apache Zookeeper



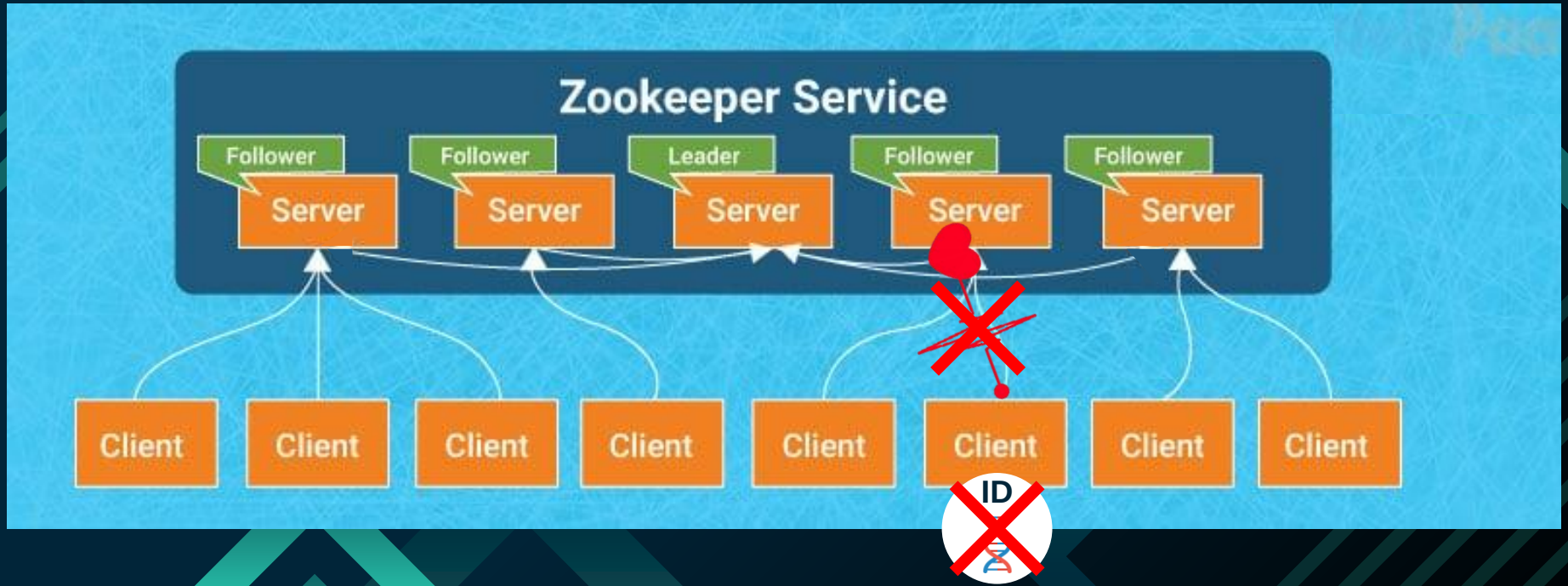
Working of Apache Zookeeper



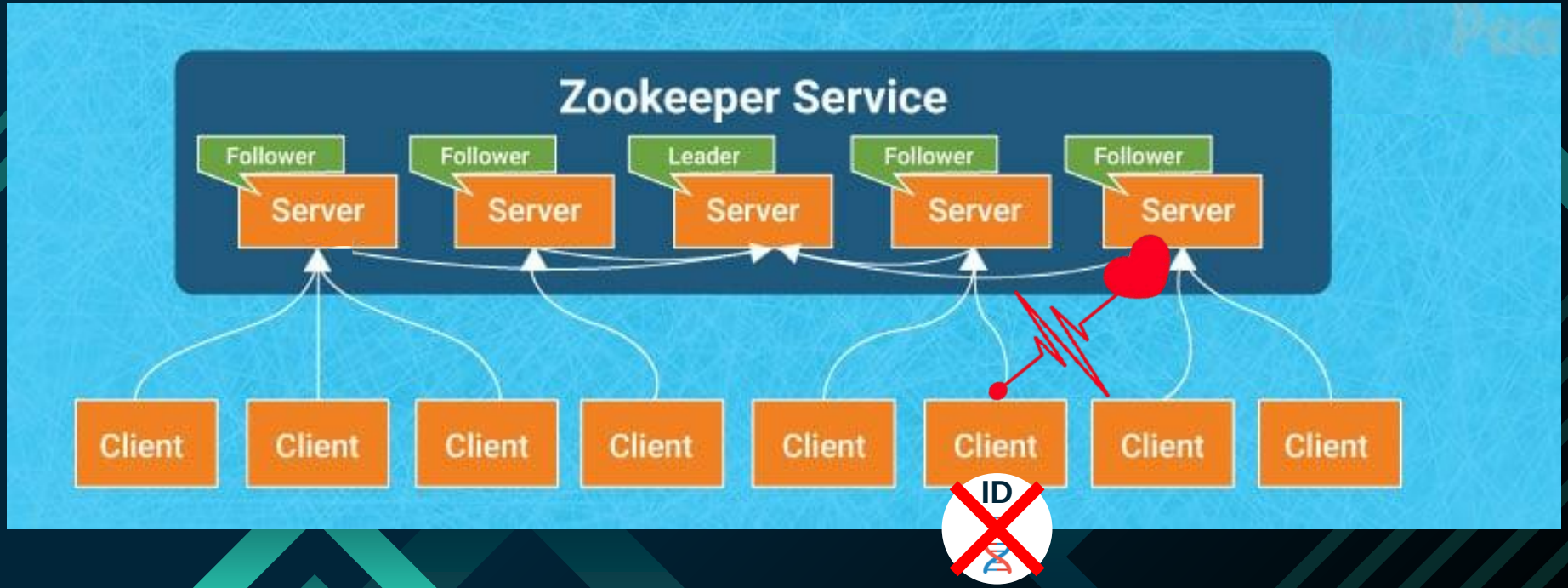
Working of Apache Zookeeper



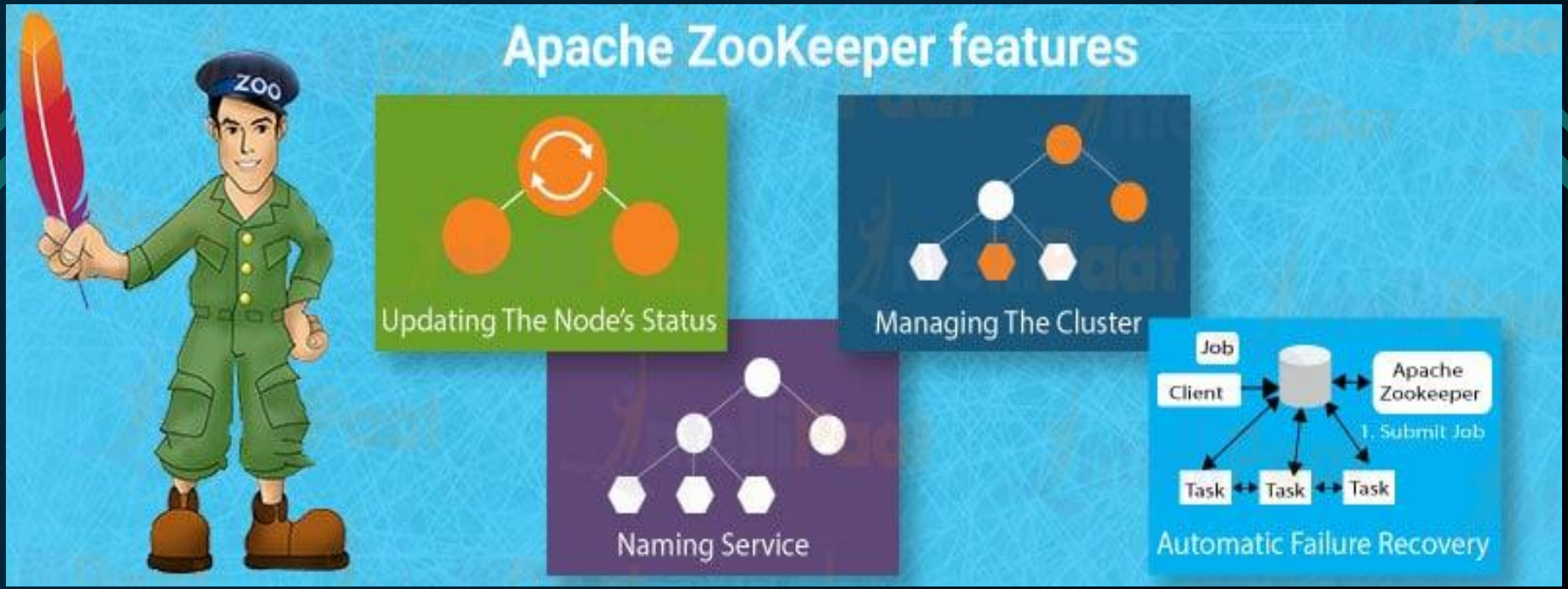
Working of Apache Zookeeper



Working of Apache Zookeeper



Features of Apache ZooKeeper



Benefits of Apache ZooKeeper

Simplicity: Coordination is done

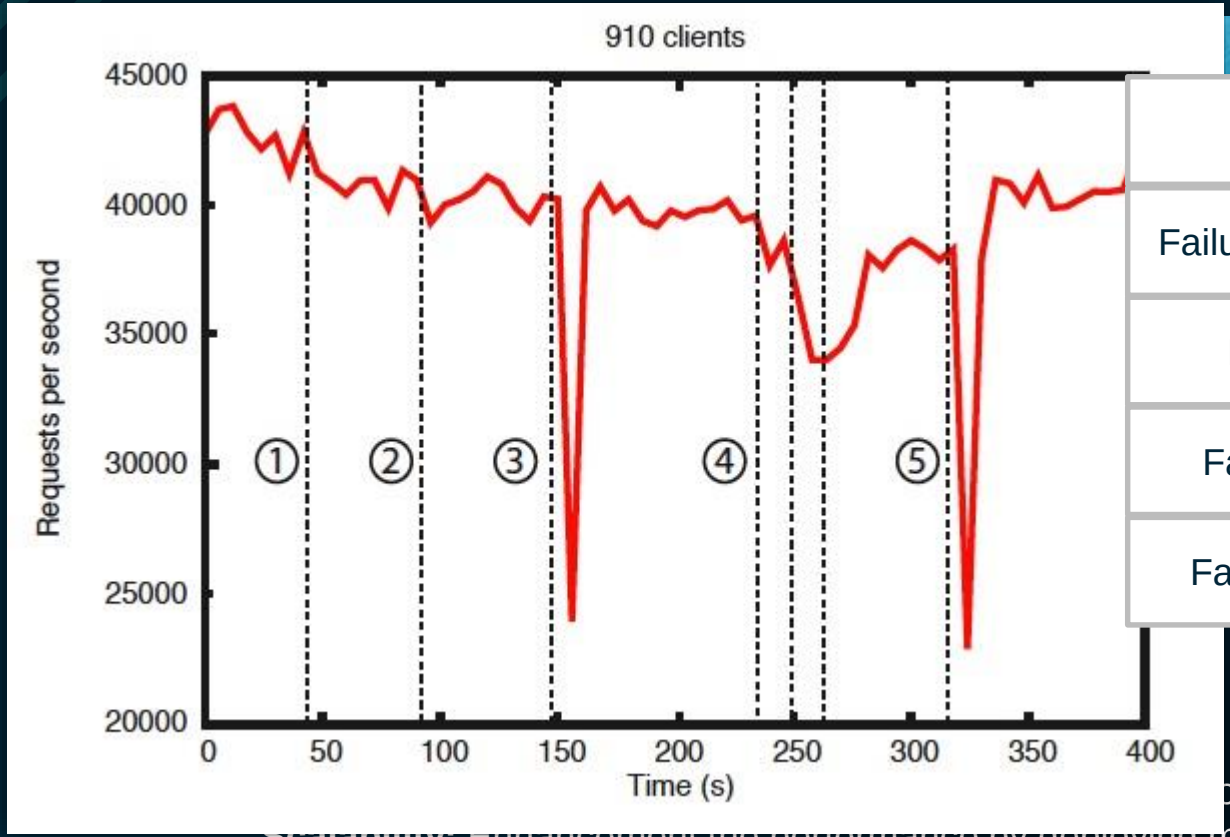
Reliability: $\left[\frac{n}{2}\right] + 1$

Order: Stamping each update with a number

Speed: Runs with a ratio of 10:1 (when ‘reads’ are more common)

Scalability: Enhancement the performance by deploying more machines

Benefits of Apache ZooKeeper



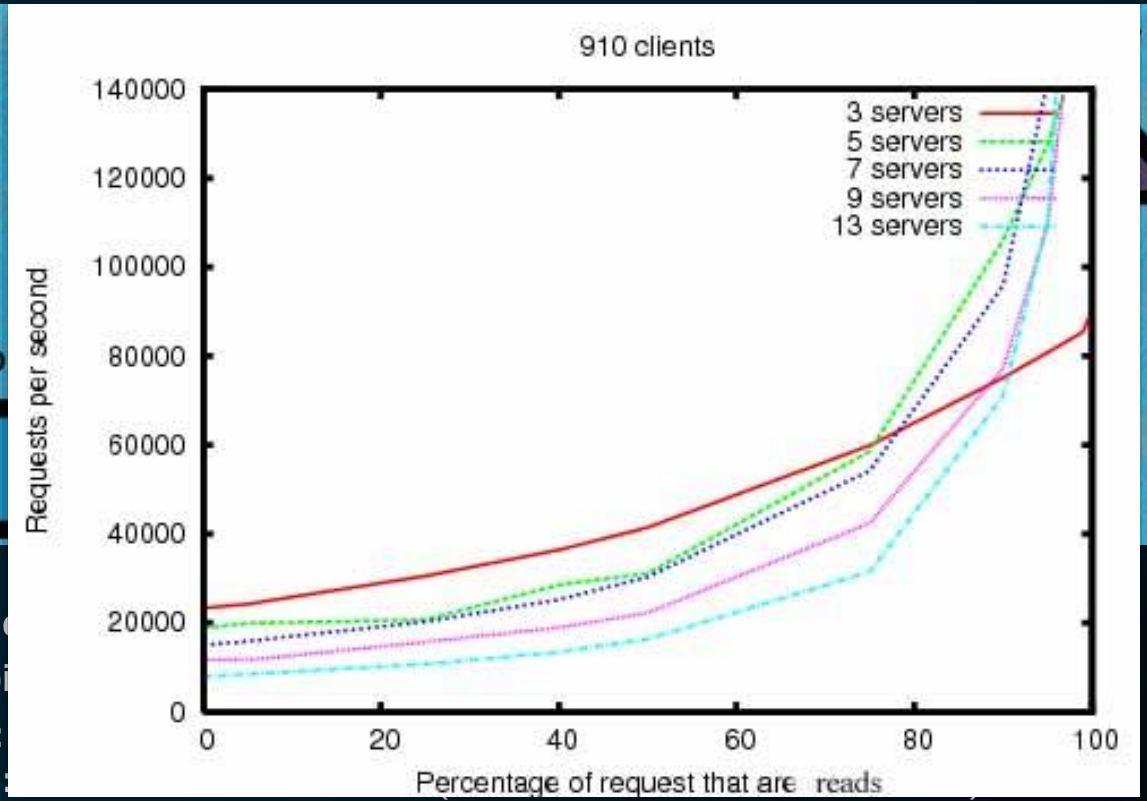
Simplicity

- Failure and recovery of a follower
- Failure and recovery of a different follower
- Failure and recovery of the leader
- Failure and recovery of two followers
- Failure and recovery of another leader

ommon)

Scalability: Enhancement the performance by deploying more machines

Benefits of Apache ZooKeeper



Simpli

Reliabi

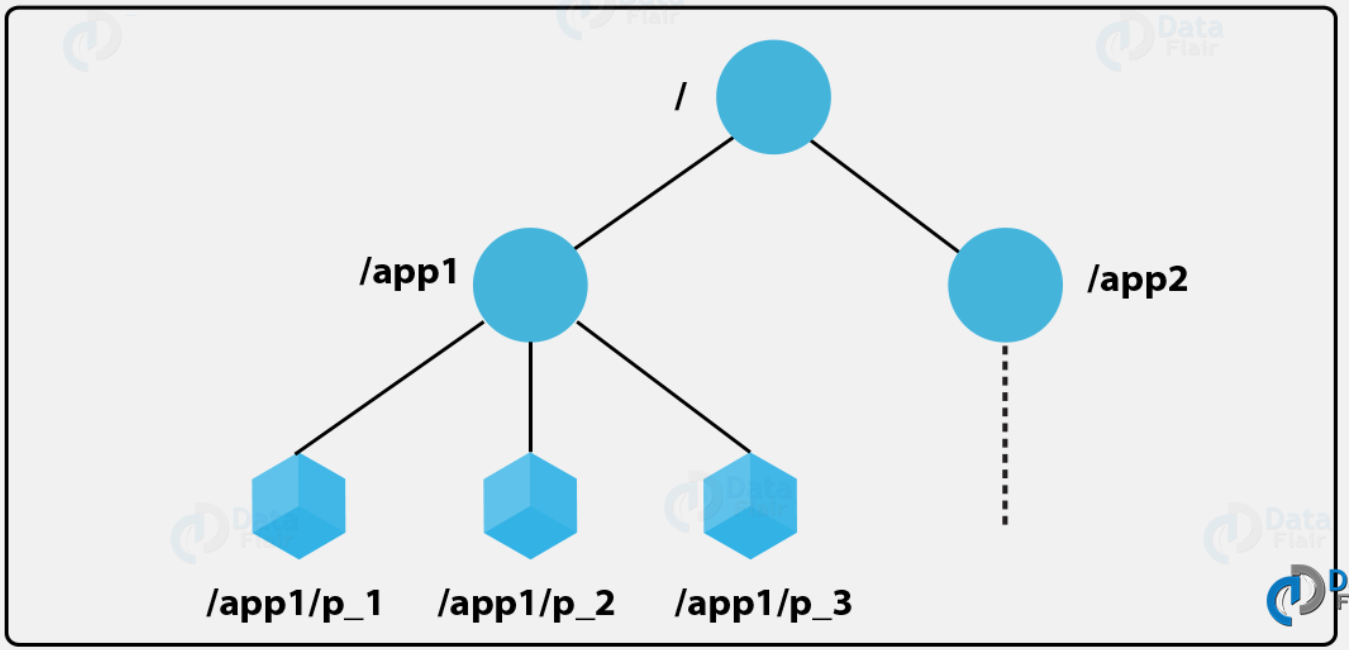
Order:

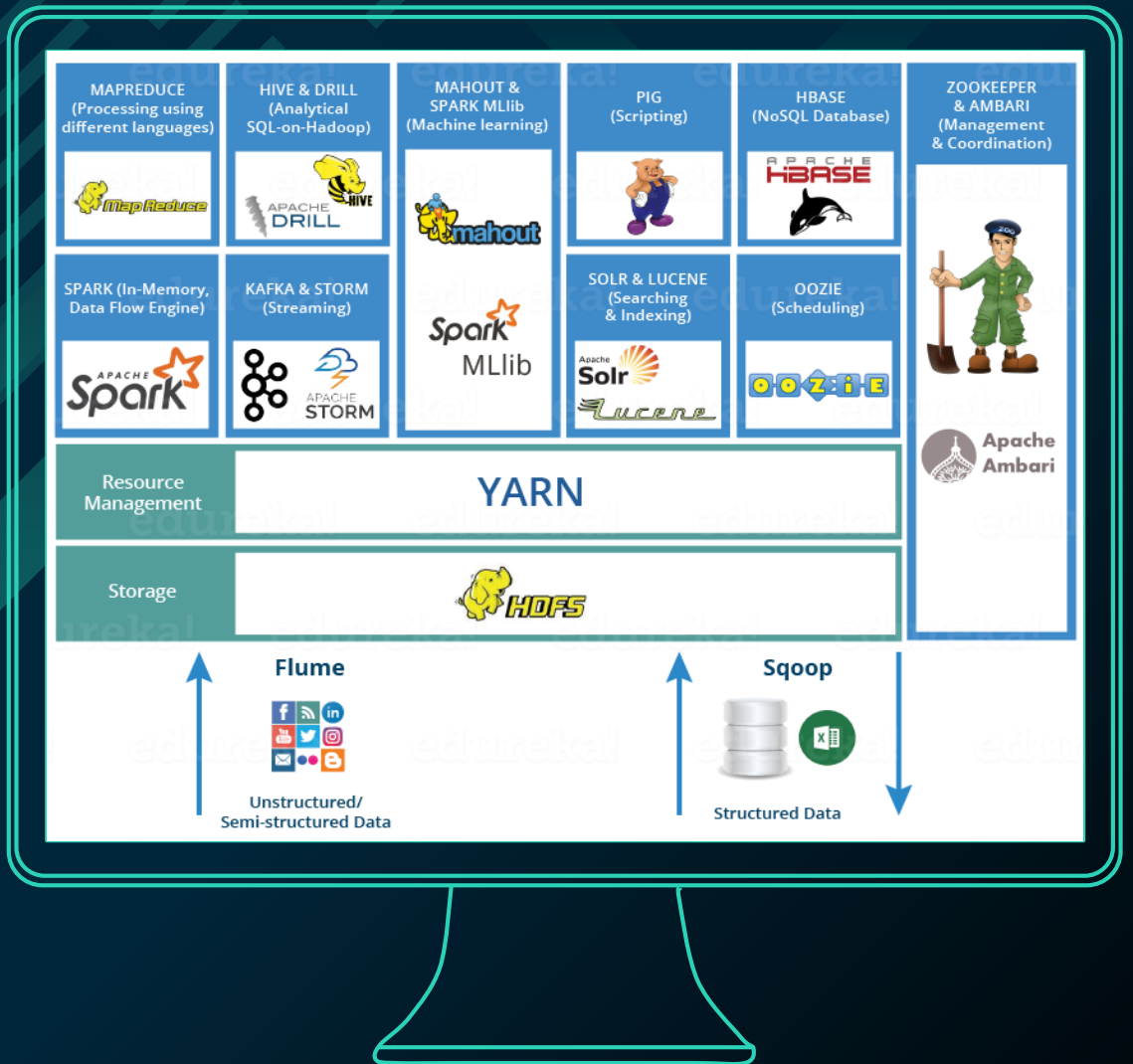
Speed:

Scalability: Enhancement the performance by deploying more machines

Data Model and The Hierarchical Namespace

Name = Sequence of path elements separated by a slash (/)
Every node is identified by a path.




MAPREDUCE (Processing using different languages)
HIVE & DRILL (Analytical SQL-on-Hadoop)
MAHOUT & SPARK MLlib (Machine learning)
PIG (Scripting)
HBASE (NoSQL Database)
ZOOKEEPER & AMBARI (Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
KAFKA & STORM (Streaming)
SOLR & LUCENE (Searching & Indexing)
OOZIE (Scheduling)
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/ Semi-structured Data
Sqoop
Structured Data

Thanks for Your Attention