Getting Started with Hadoop

Raanan Dagan
Paul Tibaldi

cloudera

What is Apache Hadoop?

Hadoop is a platform for data storage and processing that is…

- ✓ Scalable
- ✓ Fault tolerant
- ✓ Open source

CORE HADOOP COMPONENTS

Hadoop Distributed File System (HDFS)

File Sharing & Data Protection Across Physical Servers

+

MapReduce

Distributed Computing Across Physical Servers

Flexibility

- A single repository for storing processing & analyzing any type of data
- Not bound by a single schema

Scalability

- Scale-out architecture divides workloads across multiple nodes
- Flexible file system eliminates ETL bottlenecks

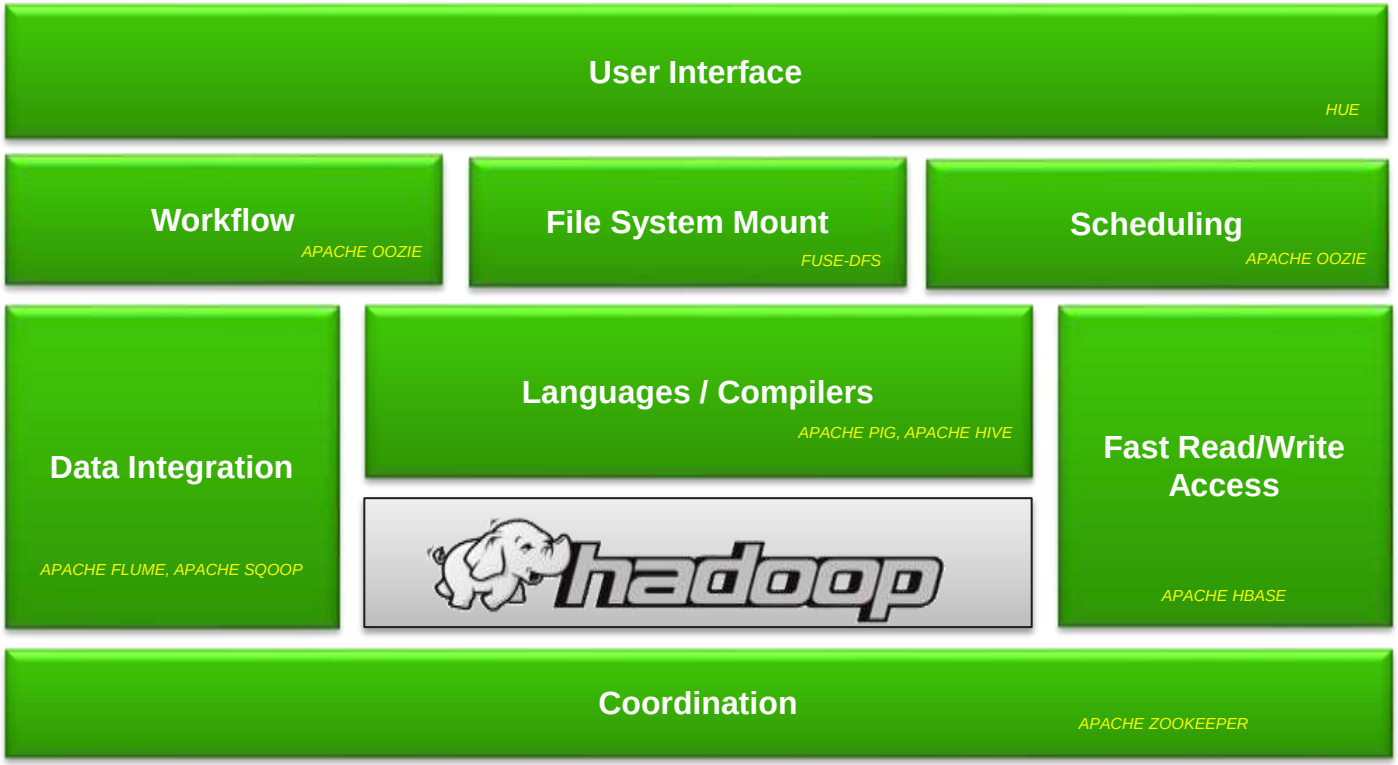
Low Cost

- Can be deployed on commodity hardware
- Open source platform guards against vendor lock

What Makes Hadoop Different?

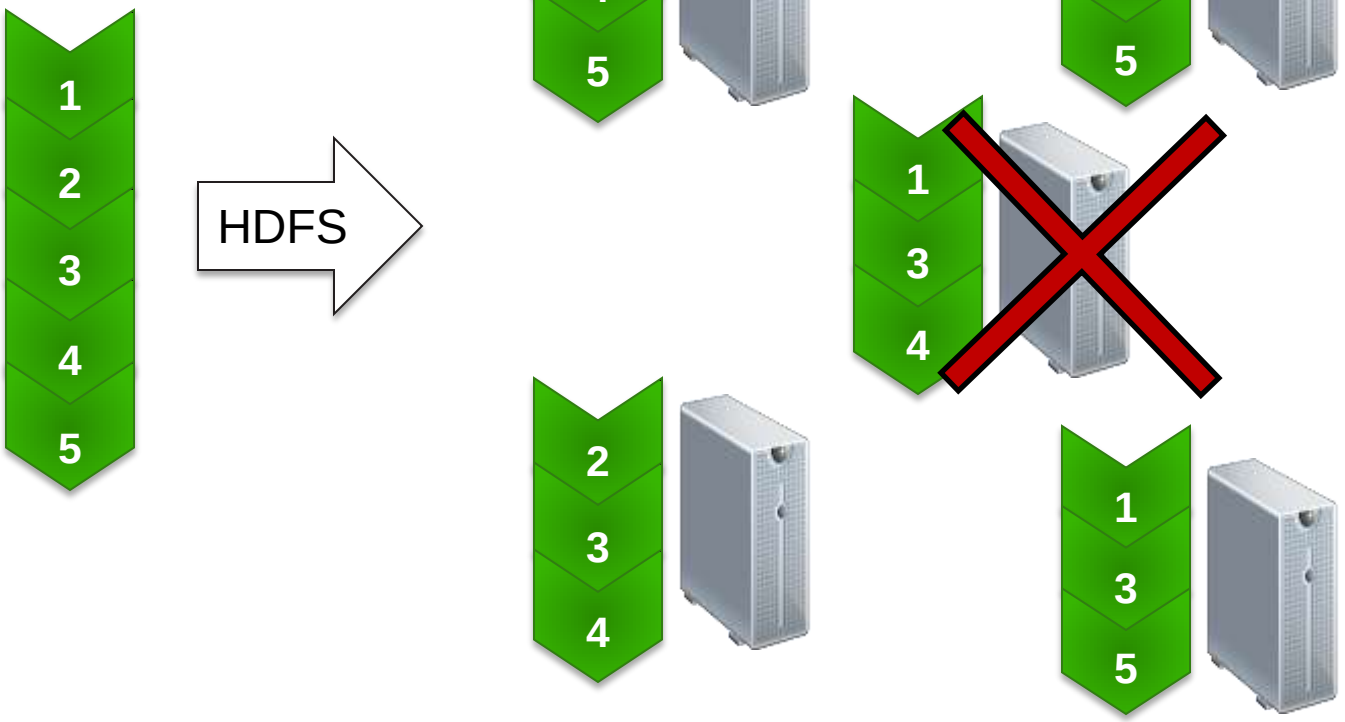
- Ability to scale out to Petabytes in size using commodity hardware
- Processing (MapReduce) jobs are sent to the data versus shipping the data to be processed
- Hadoop doesn’t impose a single data format so it can easily handle structure, semi-structure and unstructured data

Components of Hadoop



Hadoop Distributed File System

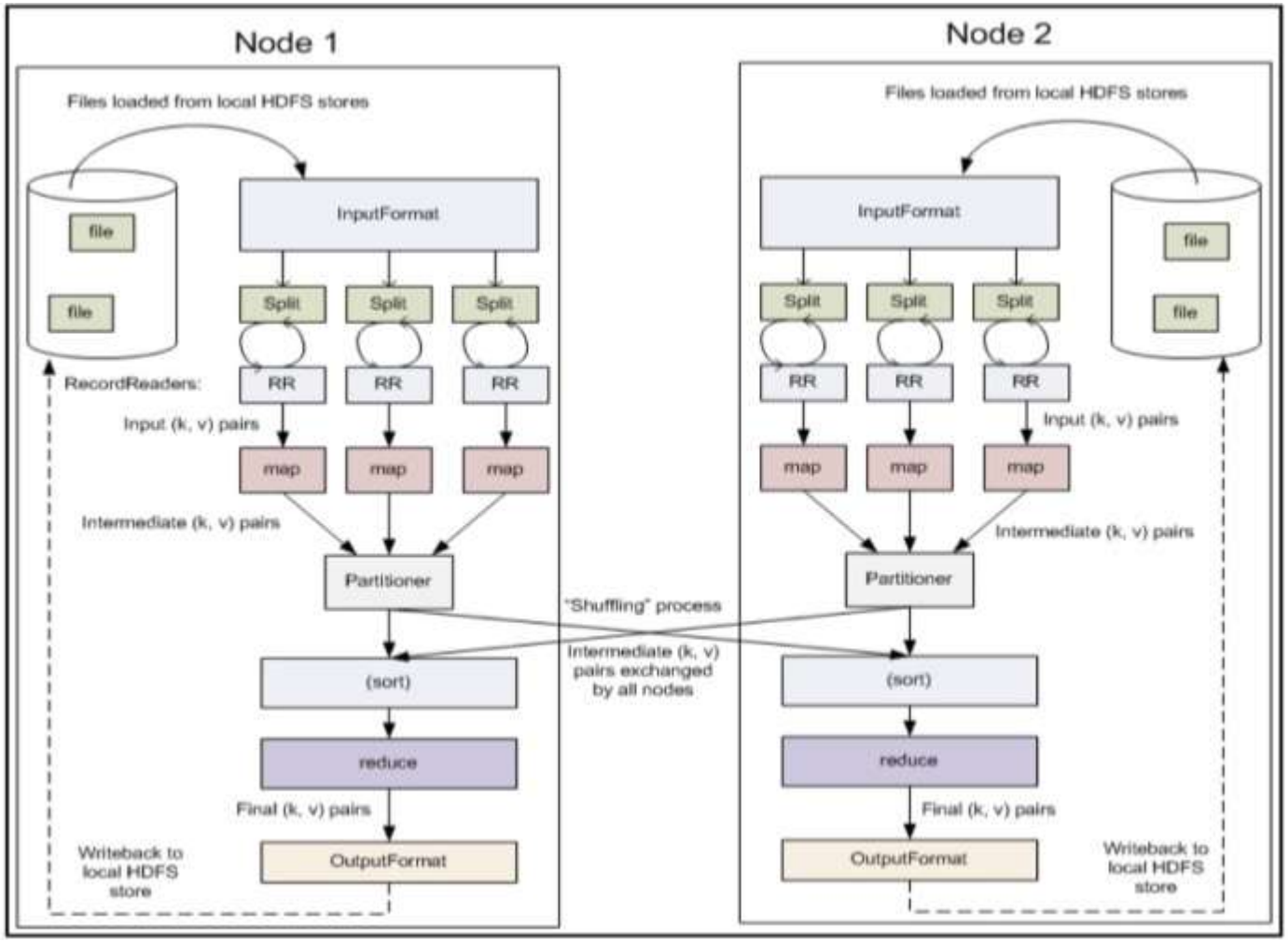
Block Size = 64MB
Replication Factor = 3



Five Hadoop Deamons

- Master Nodes (very few nodes)
  - NameNode (Holds the metadata)
  - Secondary NameNode (housekeeping)
  - JobTracker (Manages MapReduce jobs)
- Slave Nodes (Majority of the nodes)
  - DataNode (Stores actual HDFS data blocks)
  - TaskTracker (Map and Reduce tasks)

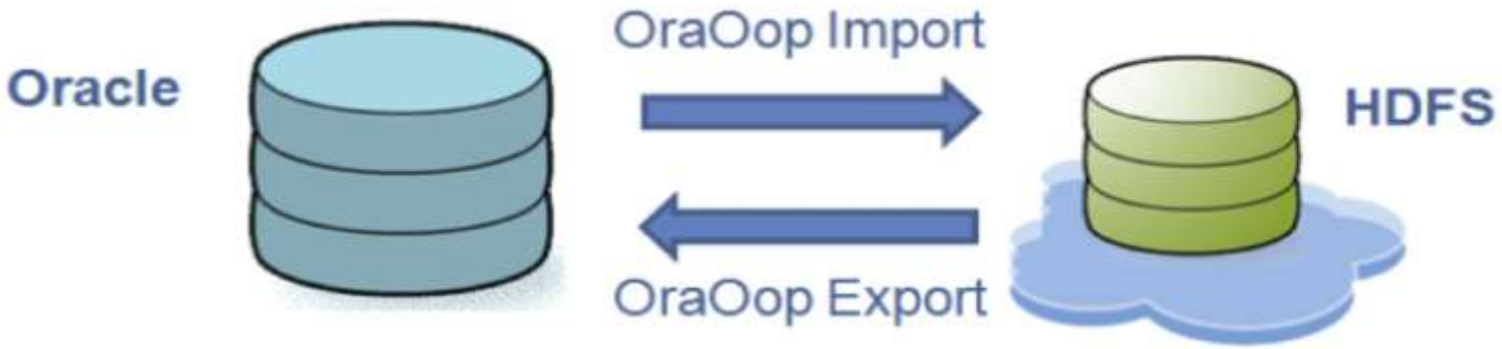
MapReduce


Node 1
Files loaded from local HDFS stores
file
file
InputFormat
Split
Split
Split
RecordReaders:
RR
RR
RR
Input (k, v) pairs
map
map
map
Intermediate (k, v) pairs
Partitioner
"Shuffling" process
Intermediate (k, v) pairs exchanged by all nodes
(sort)
reduce
Final (k, v) pairs
OutputFormat
Writeback to local HDFS store
Node 2
Files loaded from local HDFS stores
InputFormat
file
file
Split
Split
Split
RR
RR
RR
Input (k, v) pairs
map
map
map
Intermediate (k, v) pairs
Partitioner
(sort)
reduce
Final (k, v) pairs
OutputFormat
Writeback to local HDFS store

Sqoop: RDBMS Integration

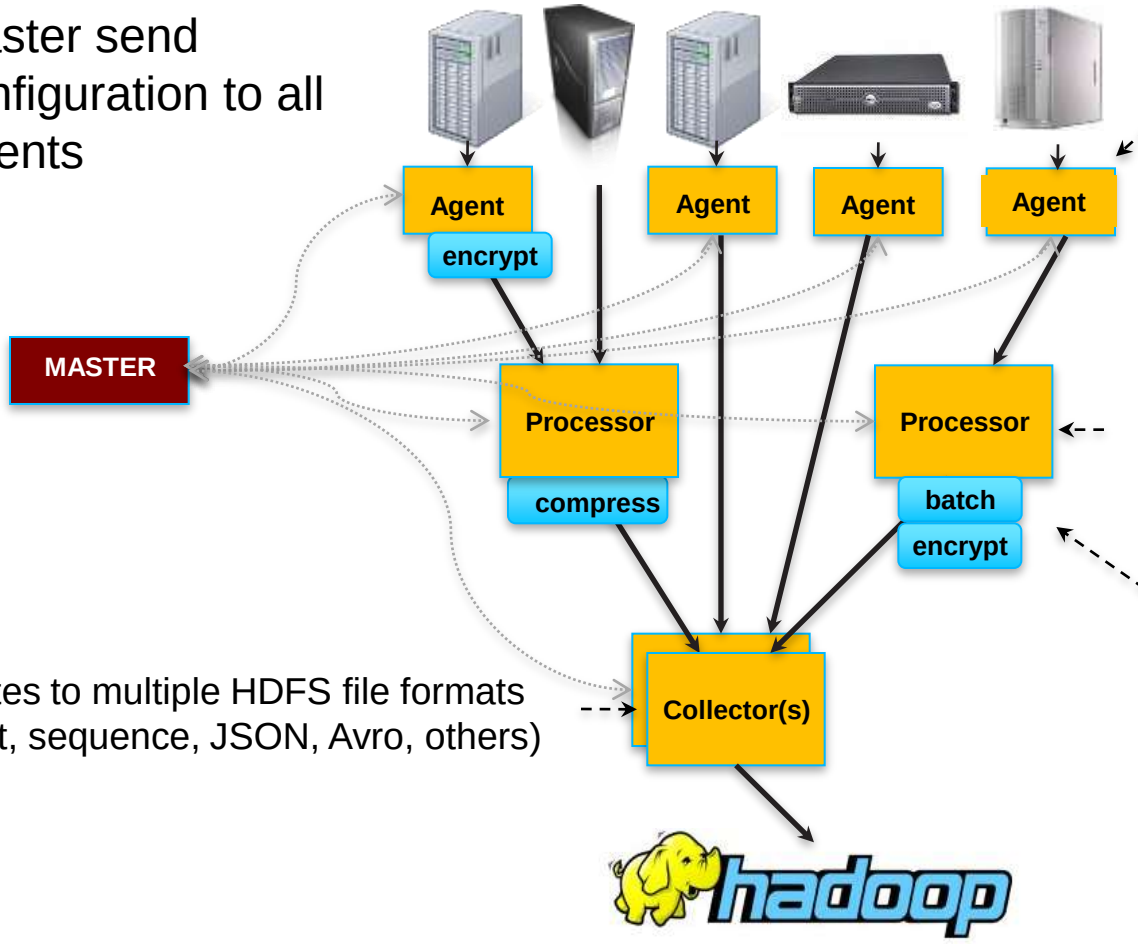
SQL to Hadoop

- Tool to import/export any JDBC-supported database into Hadoop
- Transfer data between Hadoop and external databases or EDW
- High performance connectors for some RDBMS



Flume: Log file collector

Master send configuration to all Agents



Configurable levels of reliability

Guarantee delivery in event of failure

Deployable, centrally administered

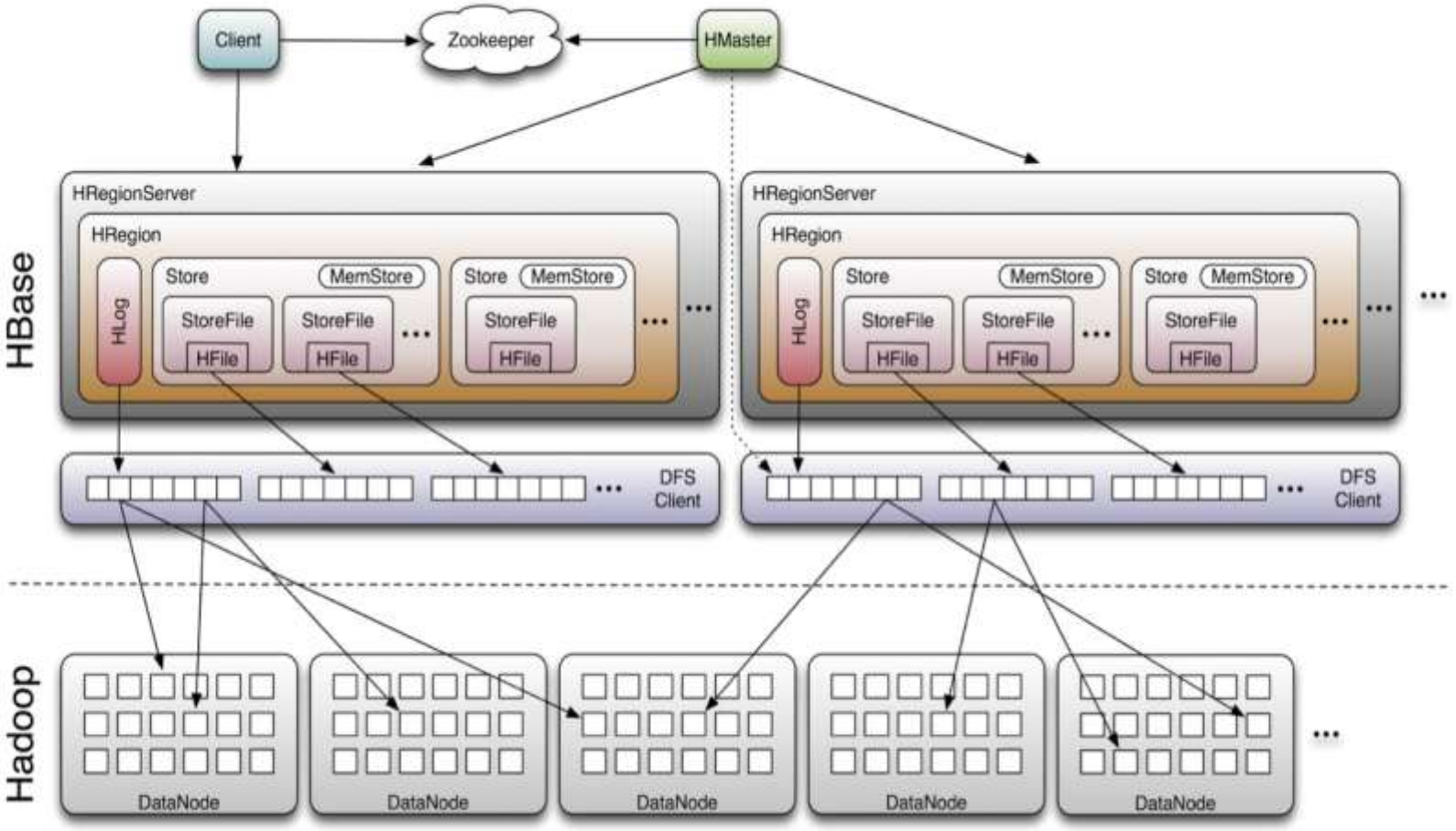
Optionally pre-process incoming data: perform transformations, suppressions, metadata enrichment

Writes to multiple HDFS file formats (text, sequence, JSON, Avro, others)

Flexibly deploy decorators at any step to improve performance, reliability or security

Hbase: The Hadoop Database

- Holds extremely large datasets (multi-TB)
- Column – Family Data Store



Hive: SQL-Like Map Reduce

SQL-based data warehousing application

- Language is SQL-like
- Supports SELECT, JOIN, GROUP BY, etc.
- Features for analyzing very large data sets
  - Partition columns, Sampling, Buckets

- Example:

```
SELECT s.word, s.freq, k.freq FROM shakespeares
JOIN ON (s.word= k.word) WHERE s.freq >= 5;
```

Pig: Procedural Map Reduce

Data-flow oriented language – “Pig latin”

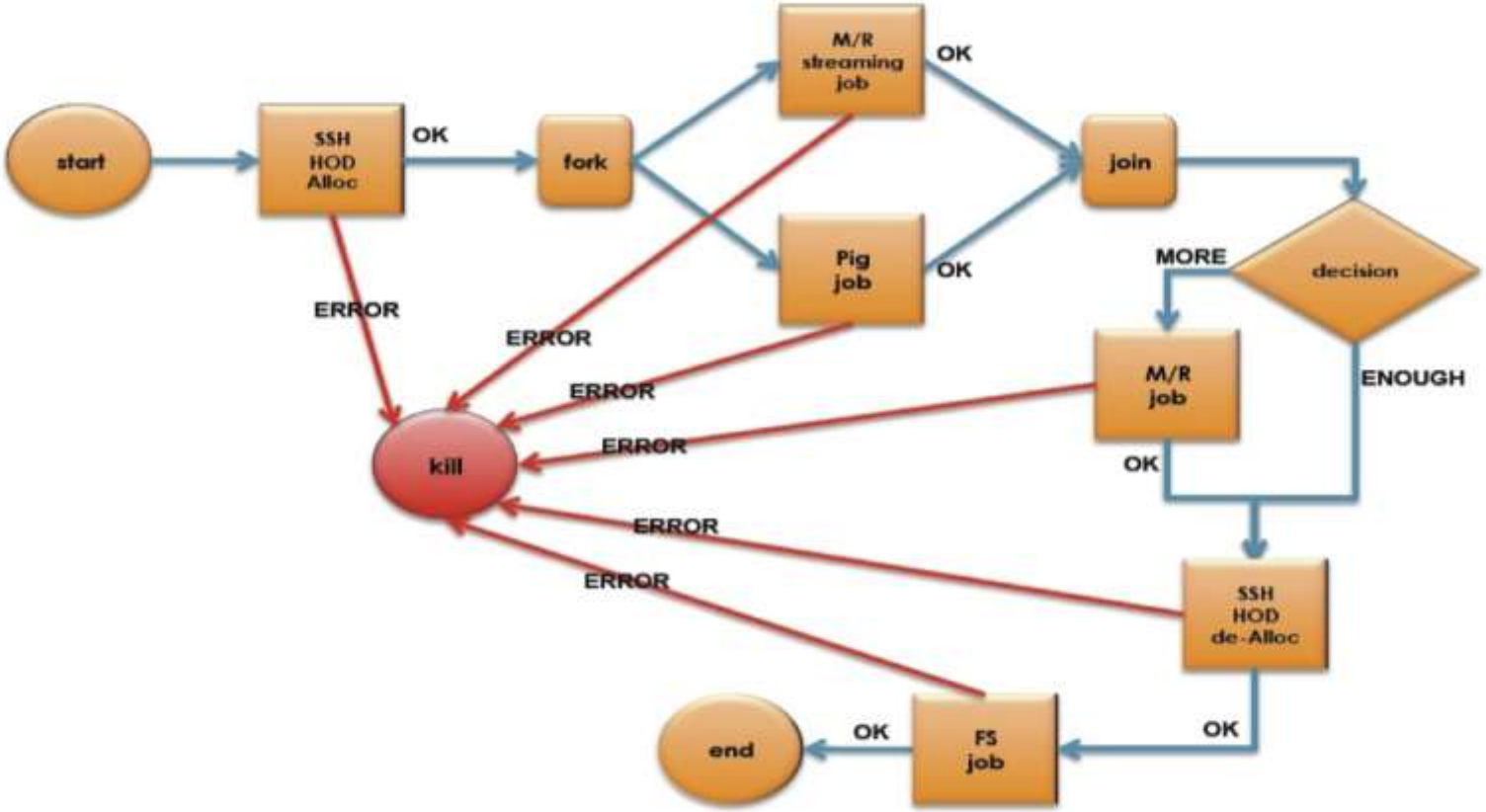
- Datatypes include sets, associative arrays, tuples
- High-level language for routing data, allows easy integration of Java for complex tasks

- Example:

```
emps=LOAD 'people.txt’ AS(id,name,salary);
rich = FILTER emps BY salary > 100000; srtd =
ORDER rich BY salary DESC; STORE srtd INTO ’
rich_people.txt';
```

Oozie: Workflow

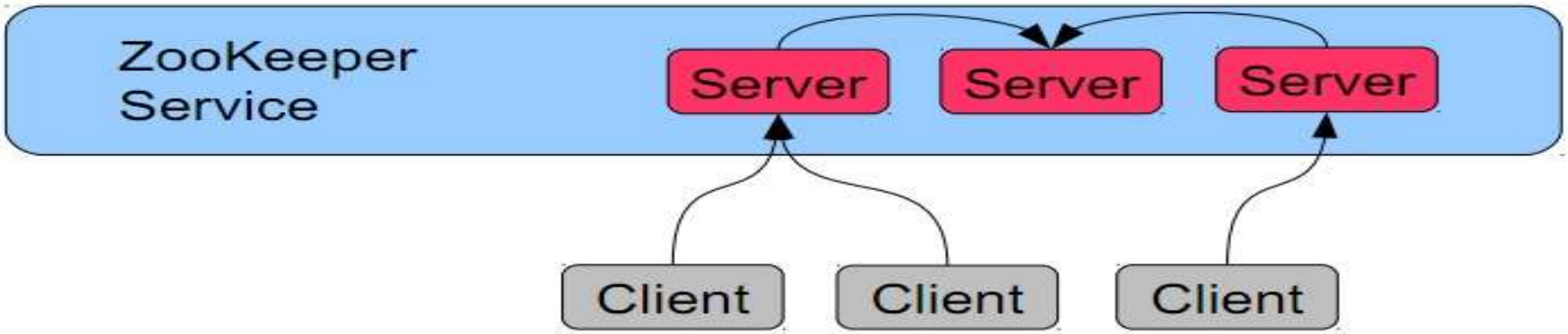
Oozie is a workflow/coordination service to manage data processing jobs for Hadoop



Zookeeper: Coordination

Zookeeper is a distributed consensus engine

- Provides well-defined concurrent access semantics:
  - Leader election
  - Service discovery
  - Distributed locking / mutual exclusion
  - Message board / mailboxes



Mahout: machine learning

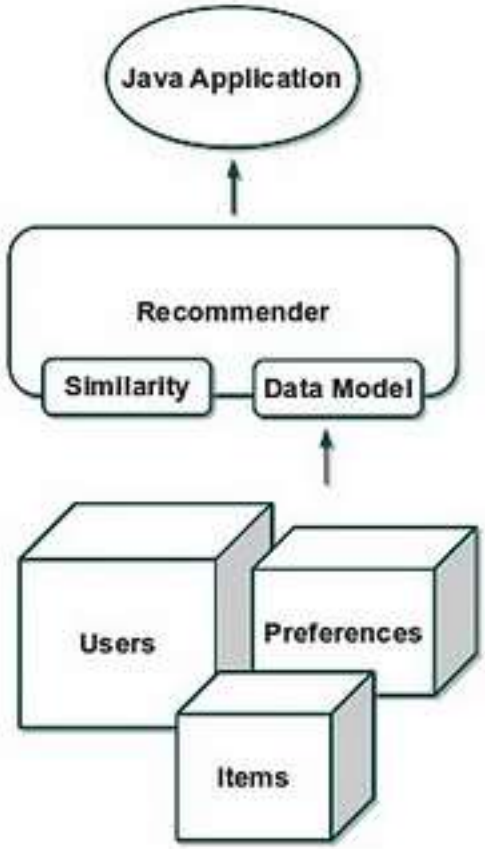
Mahout use cases:

* Recommendation based on users' behavior.

* Clustering groups related documents.

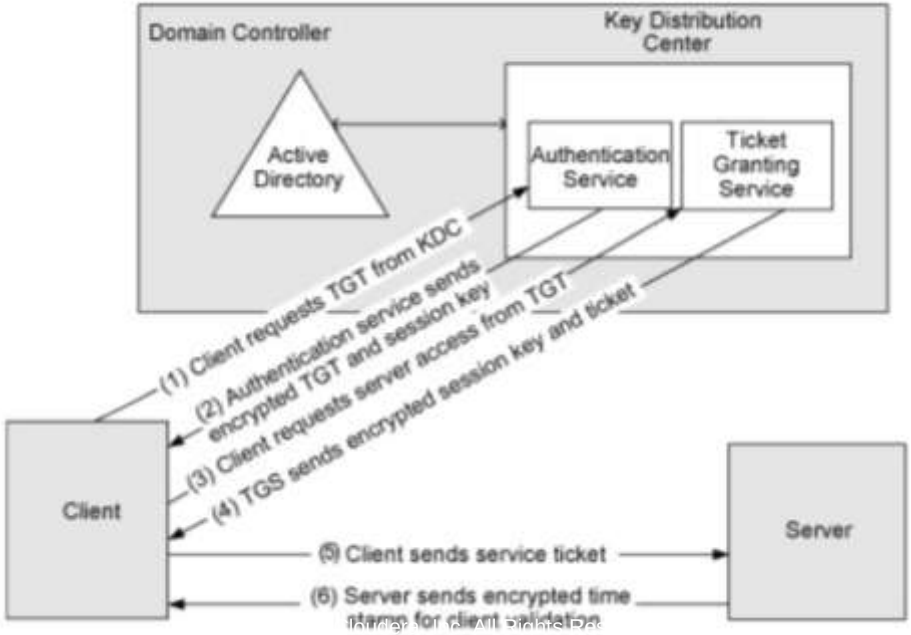
* Classification from existing categorized.

* Frequent item-set mining (shopping cart content).



Hadoop Security

- Authentication is secured by Kerberos v5 and integrated with LDAP
- Hadoop server can ensure that users and groups are who they say they are
- Job Control includes Access Control Lists, which means Jobs can specify who can view logs, counters, configurations and who can modify a job
- Tasks now run as the user who launched the job



Get Hadoop

Cloudera helps you profit from all your data.

cloudera