BITS Pilani
Pilani Campus

BITS Pilani presentation

Dr. Vivek V. Jog
Dept. Of Computer Engineering

BITS Pilani

Pilani Campus

Big Data Systems (S1-24_CCZG522)

Lecture No.3

Big Data Analytics

Tools Taxonomy

The Five V's of Big Data

Scale of Data

This refers to the sheer volume of data being generated every second.

6 Billion People have cell phones

40 Zettabytes of data will be created by 2020 and increase of 300 times from 2005

Most companies in the U.S. have at least **100 Terabytes** of data stored.

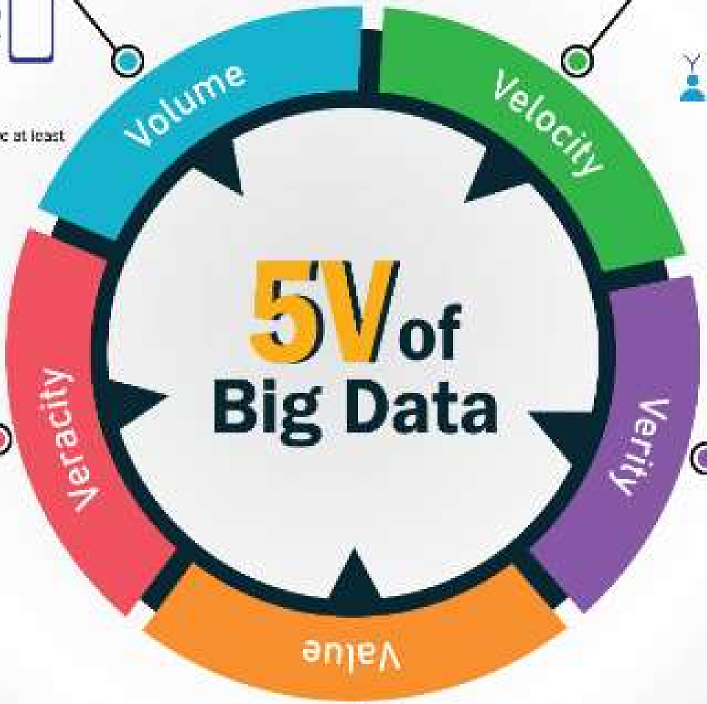
The New York Stock Exchange capture **1 TB of Trade Information**

Analysis of Streaming Data

Denotes the speed at which data is emanating and changes are occurring between the diverse data sets.

By 2016 it is projected there will be **18.9 Billion** network connections

Modern cars have close to **100 Sensors**



4 Billion+ hours of video are watched on You Tube each month

30 Billion pieces of content are shared on facebook every month

400 Million tweets are sent per day by about 200 million monthly active users

1 in 3 Business leaders don't trust the information they use to make decisions

Uncertainty Of Data

This refers to the discrepancies found in the data.

Poor data quality costs the US economy around **$ 3.1 Trillion a year**

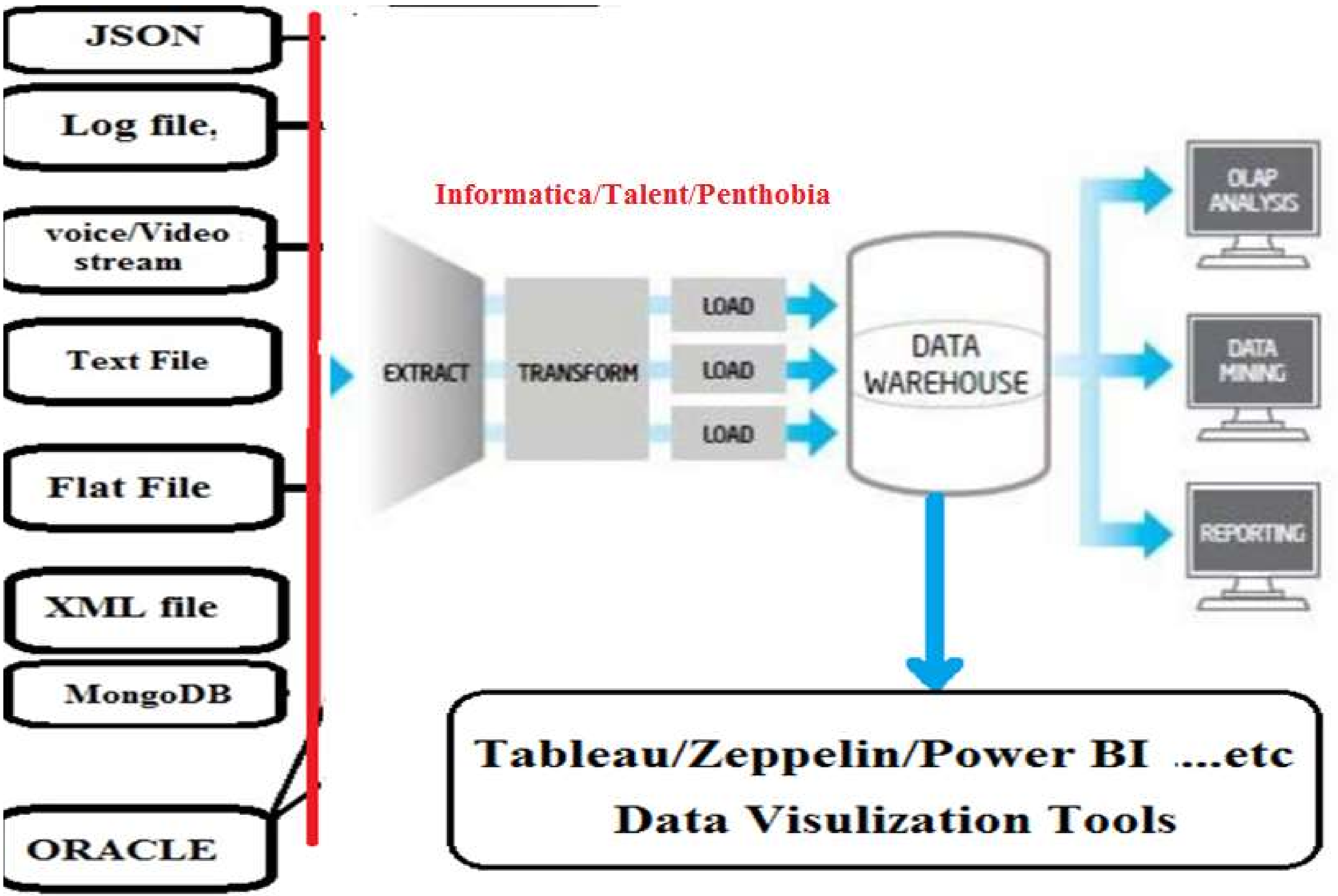
Diffrent forms of data

As more and more data is being digitized.

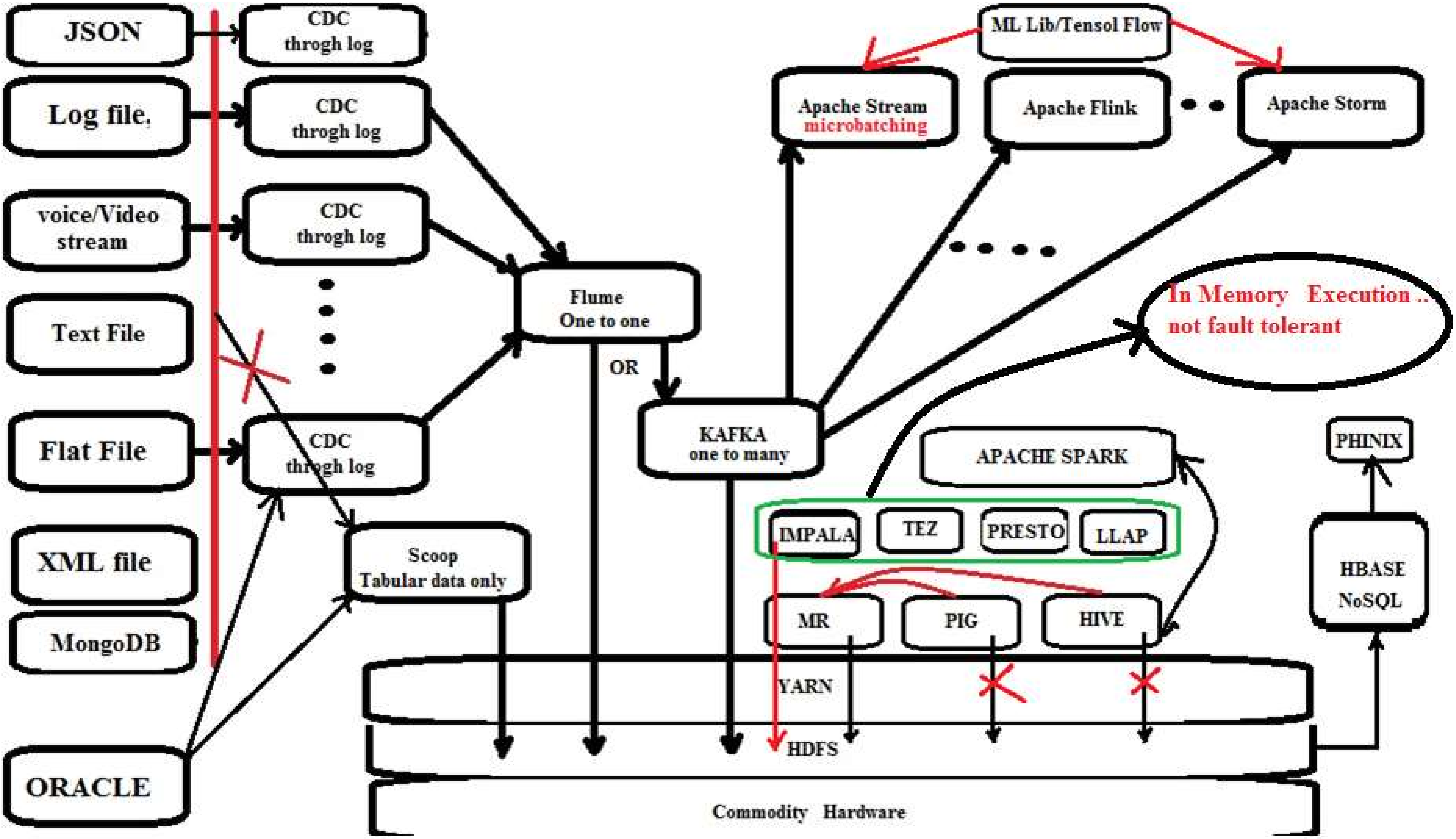
Value Of Data

Having access to big data is all well and good but that's only useful if we can turn it into a value.

Traditional approach

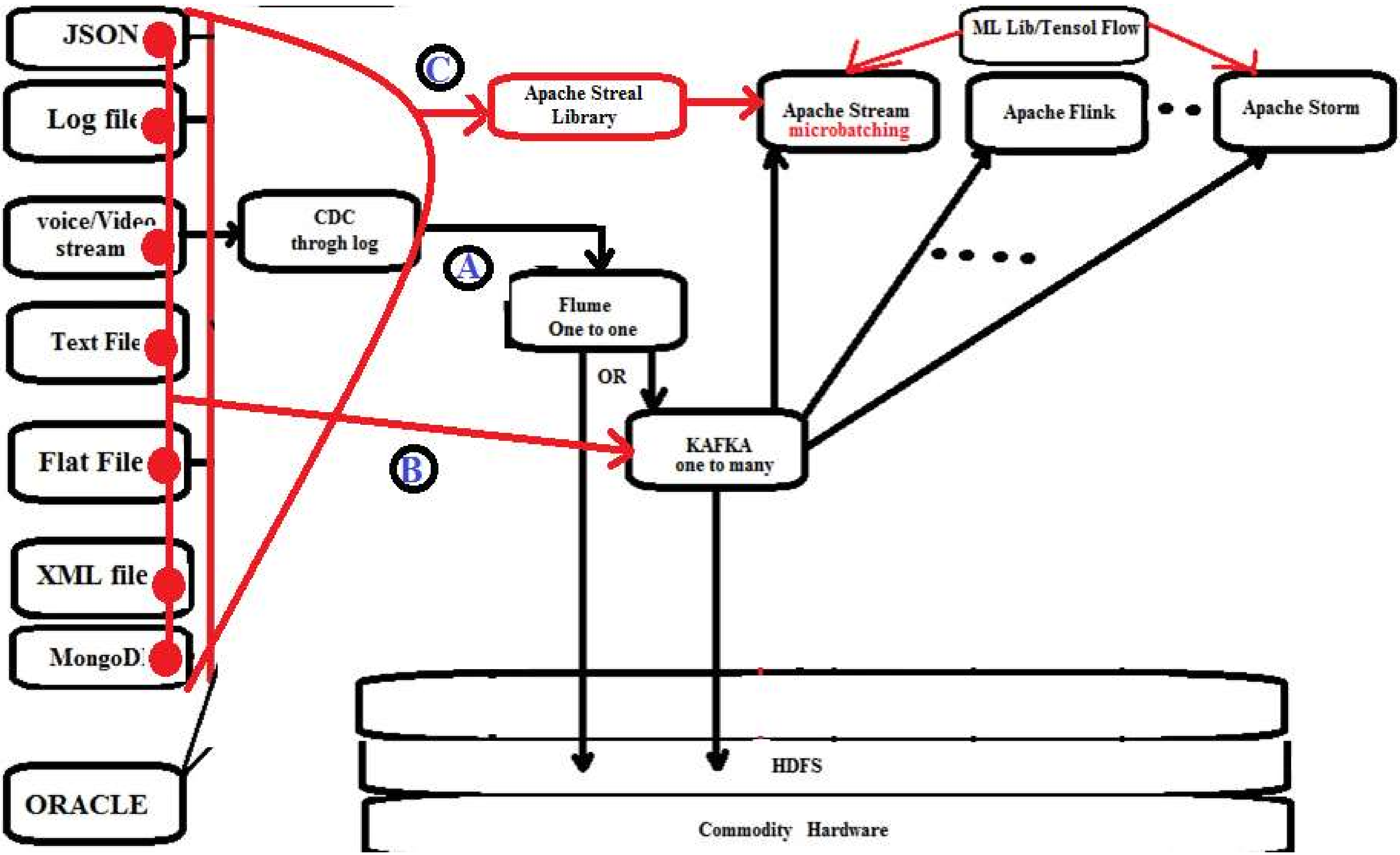


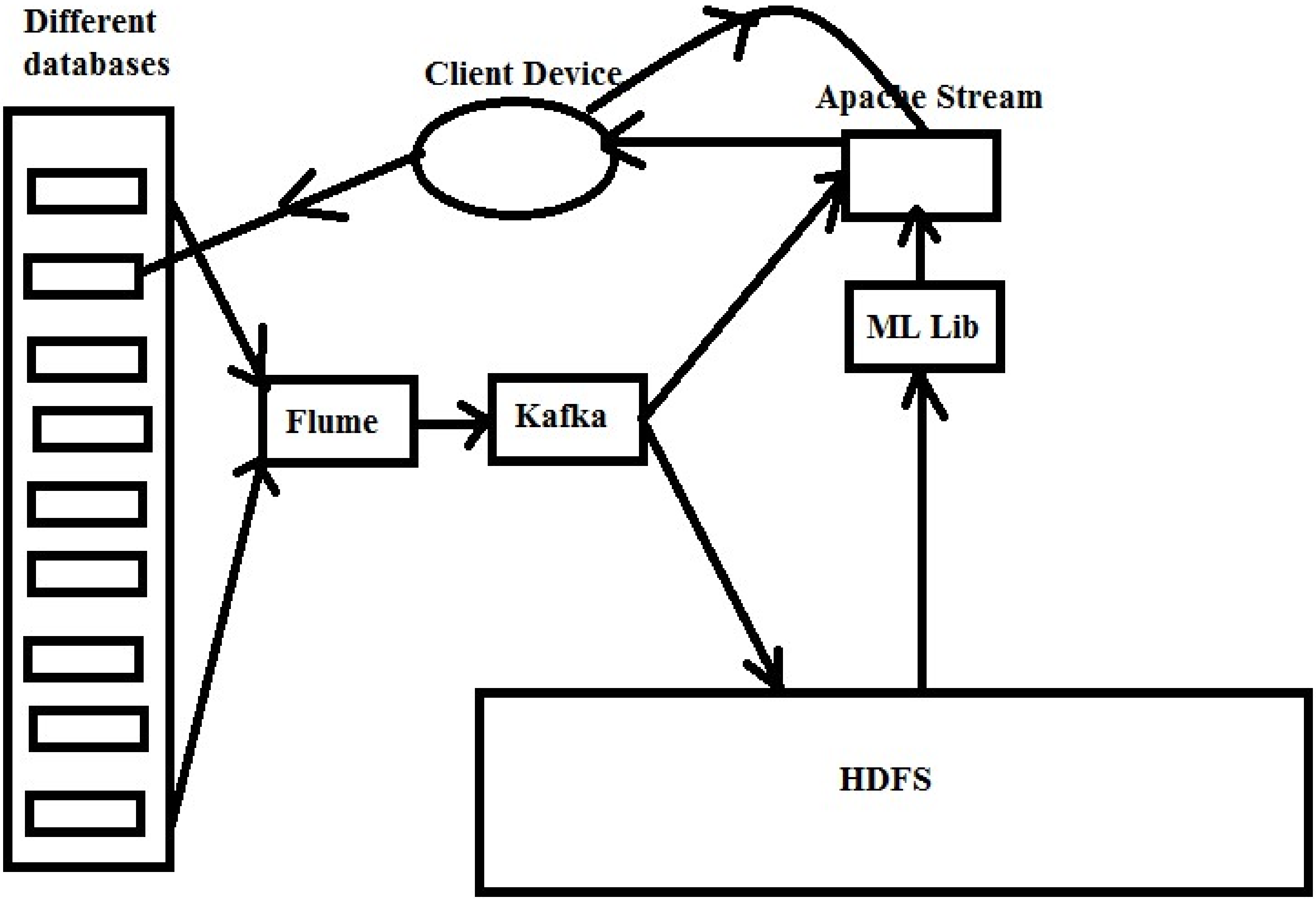
Hadoop Approach -Bird Eye View

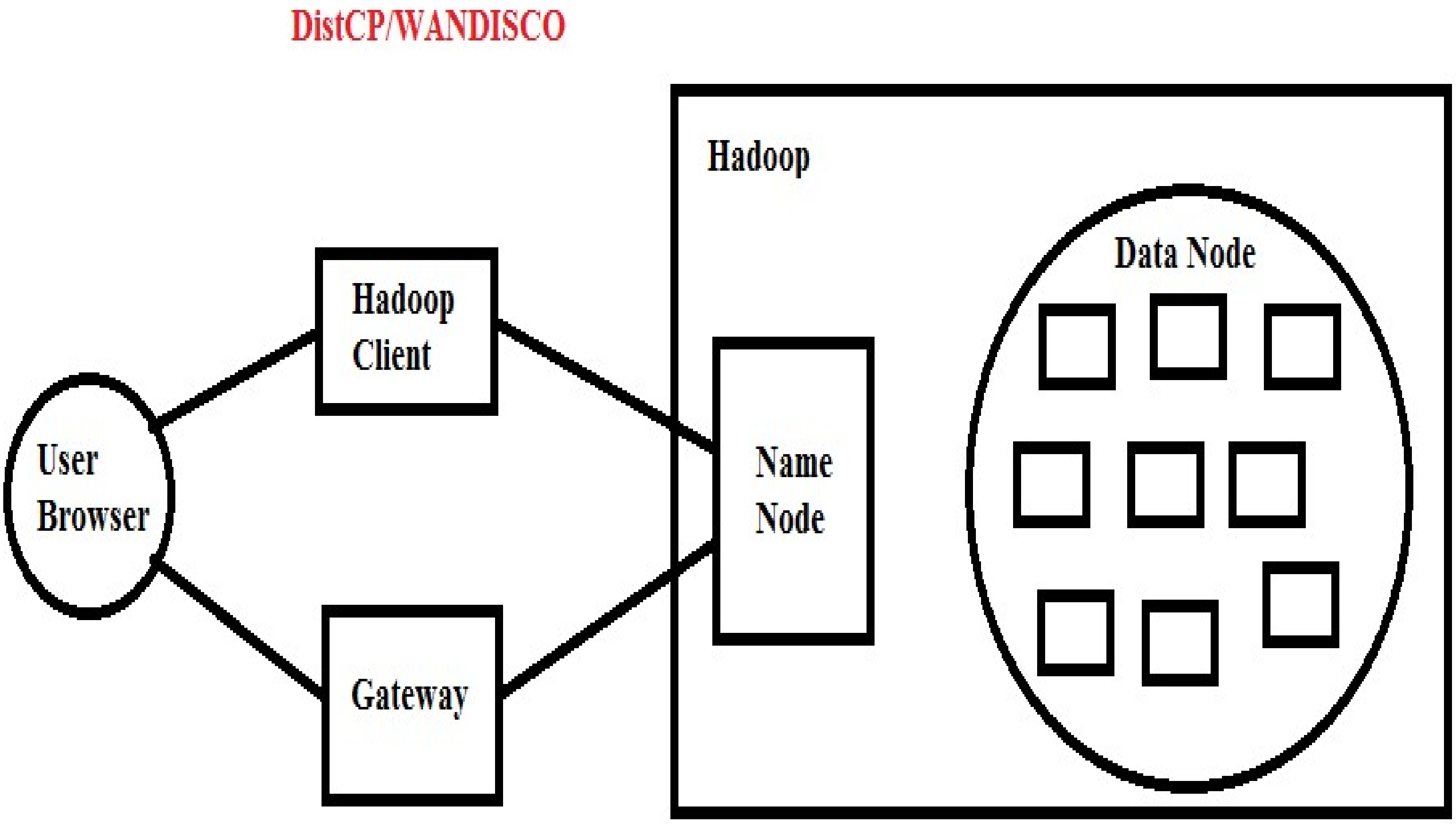


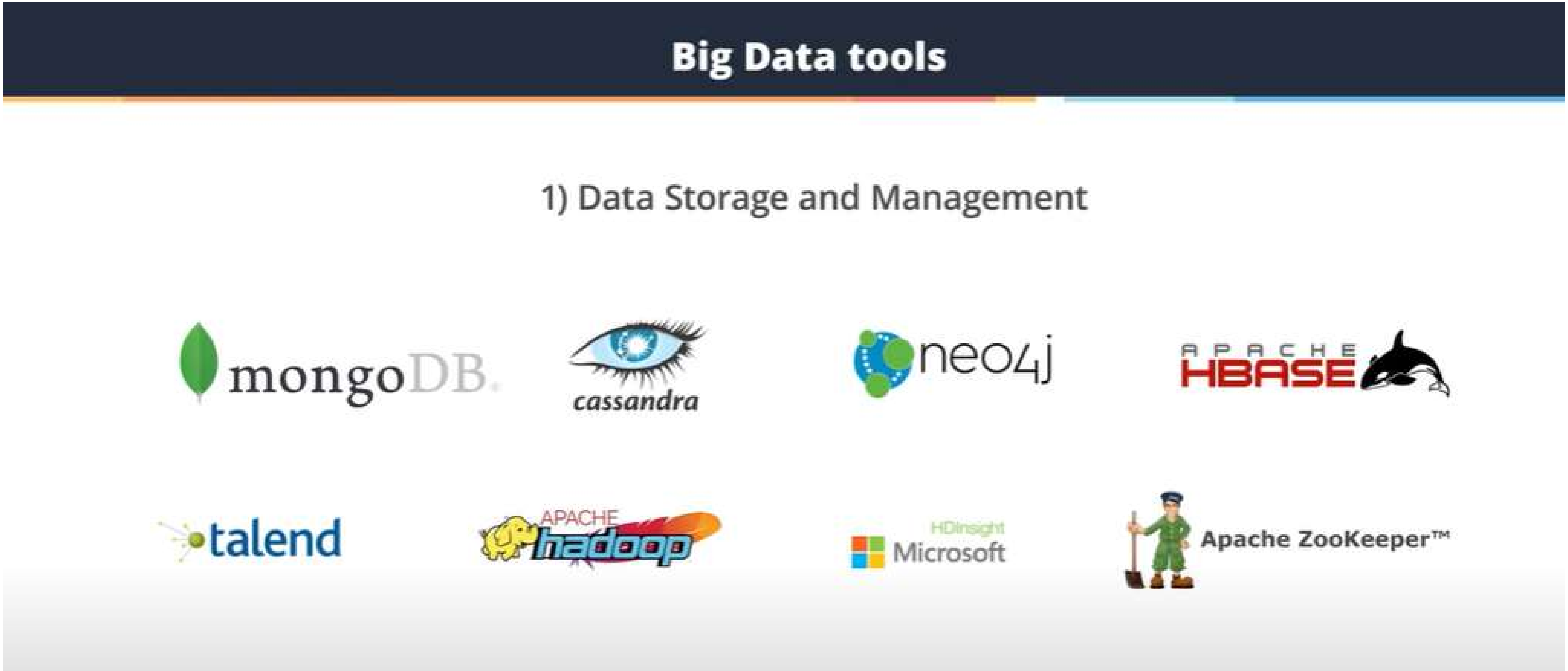
ELT APPROACH

Different Approaches – Know your requirements




Different databases
Client Device
Apache Stream
ML Lib
Flume
Kafka
HDFS


DistCP/WANDISCO
Hadoop
Data Node
Hadoop Client
User Browser
Name Node
Gateway



Big Data tools

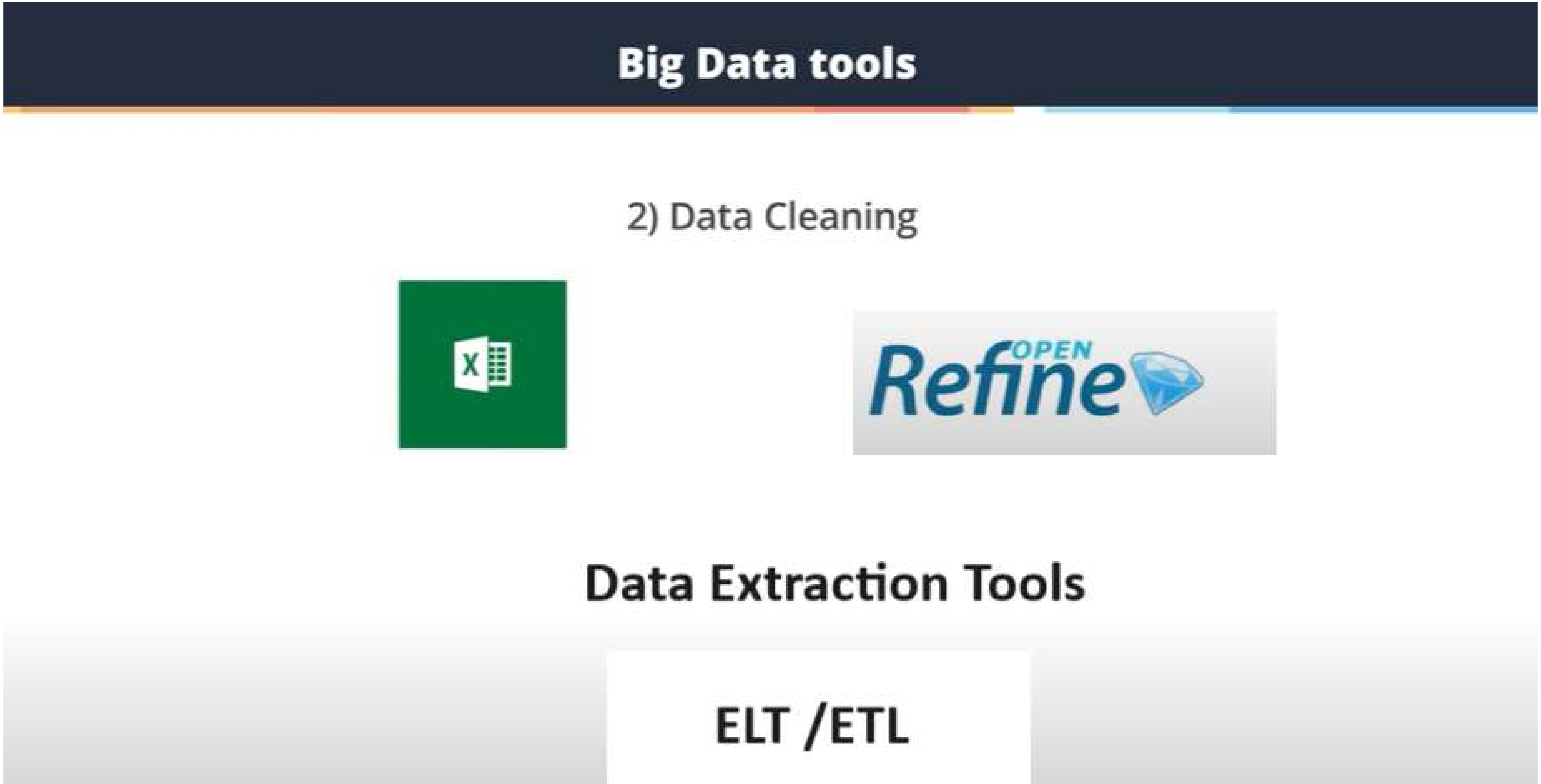
1) Data Storage and Management

Flume V/s Kafka

- F –point to point
- K - Multi
- F- Can pull data without disturbing client
- K- Need Kafka producer services on client box

Apache STORM

– Better real time processing system then Flink and Apache streaming

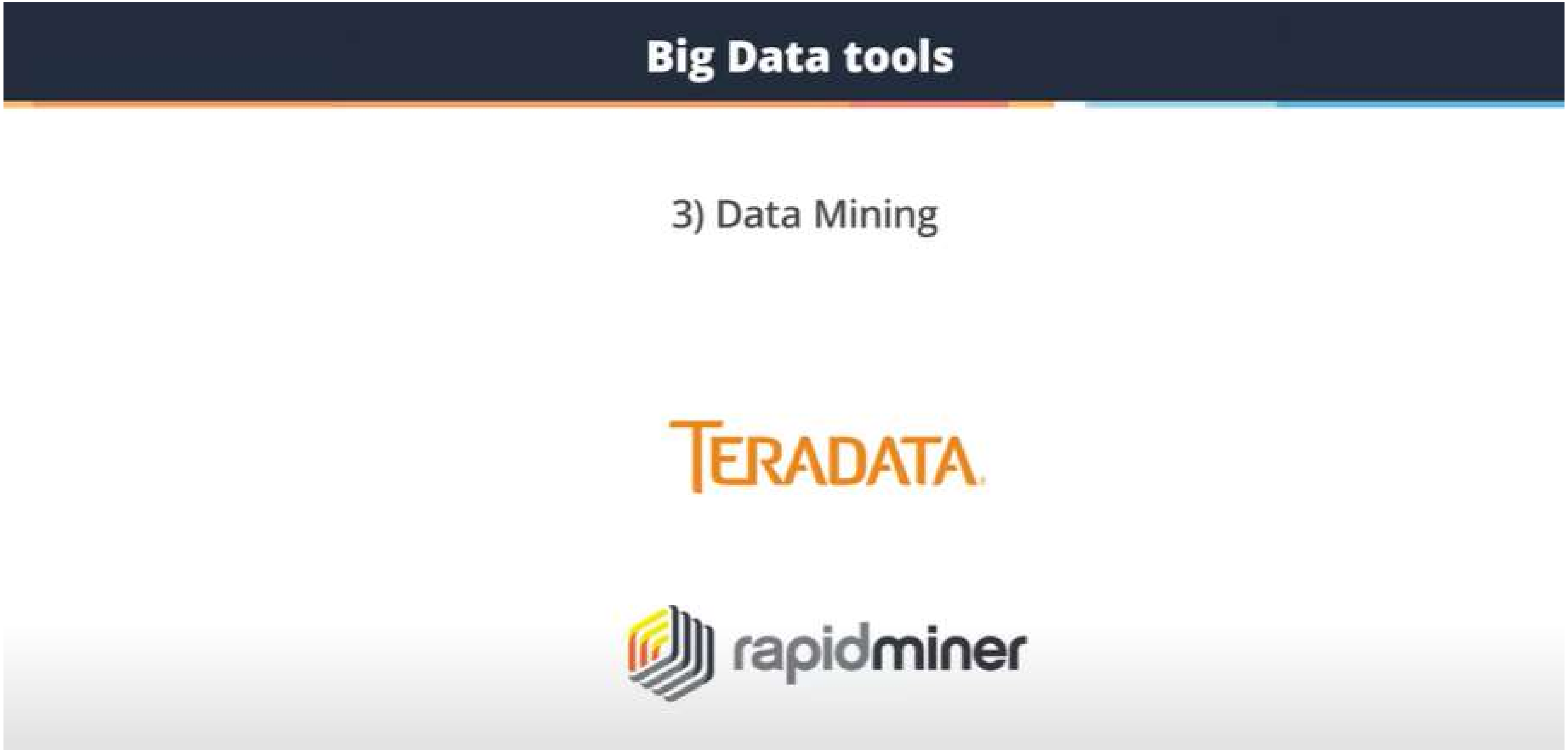


Big Data tools

2) Data Cleaning

Data Extraction Tools

ELT /ETL



Big Data tools

3) Data Mining

TERADATA

rapidminer

Big Data tools

4) Data Visualization

tableau

IBM Watson Analytics

IBM.

plotly

Big Data tools

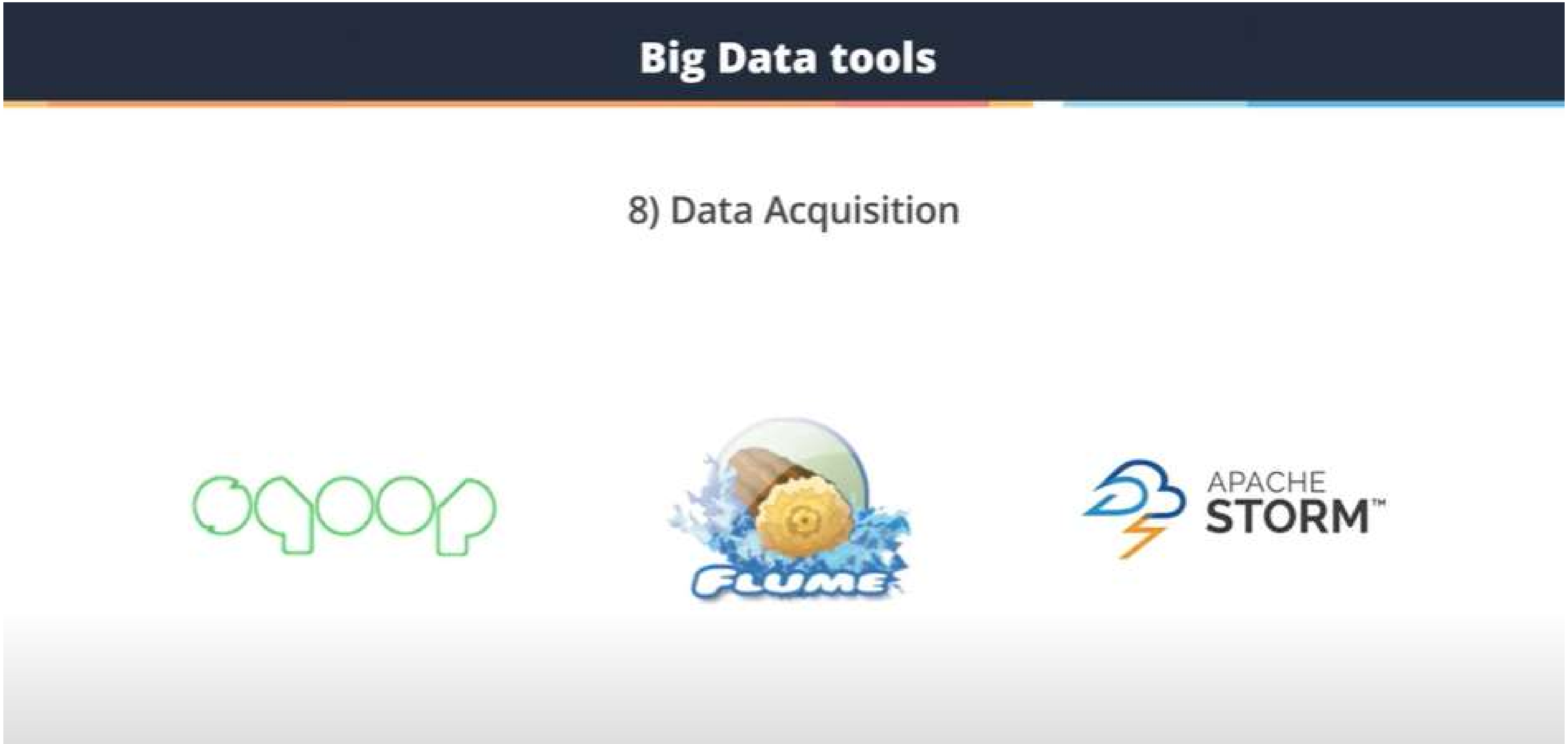
5) Data reporting

Big Data tools

7) Data Analysis

hadoop
MapReduce

APACHE
Spark
TM



Big Data tools

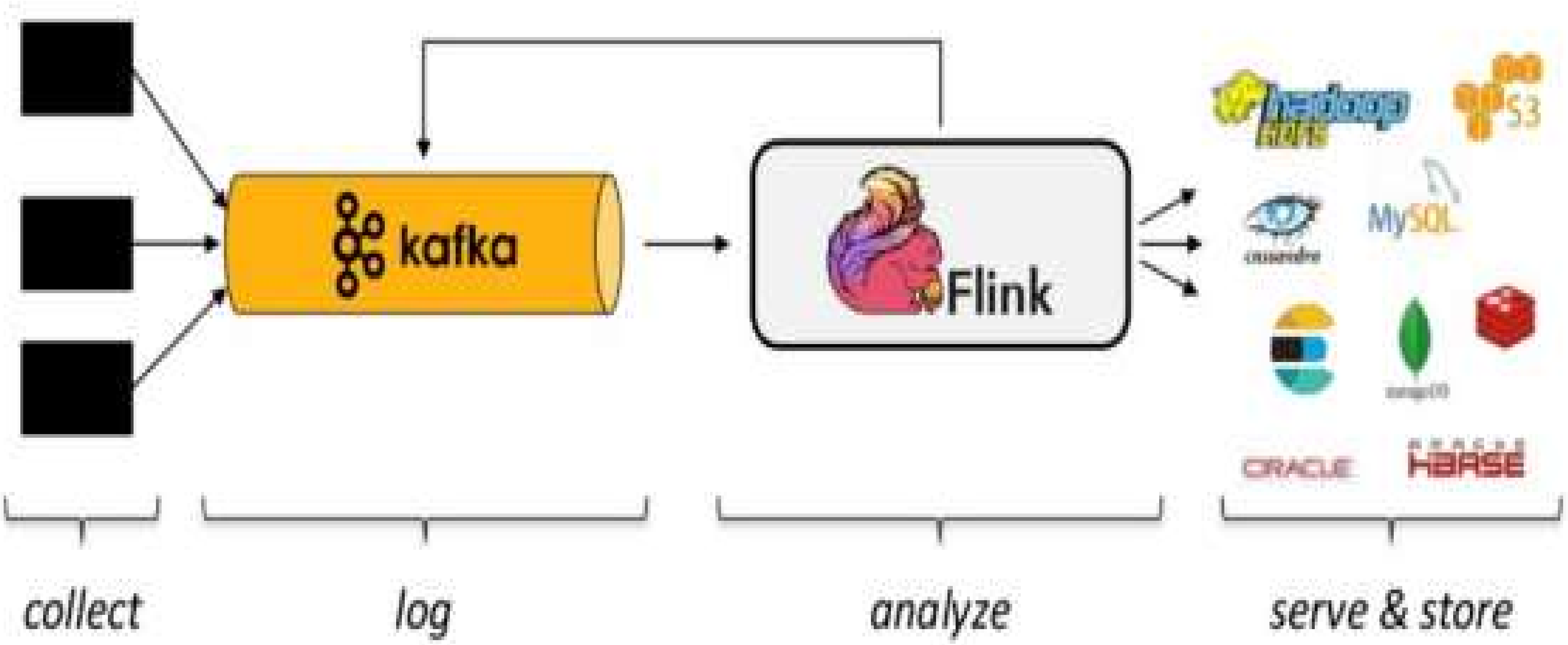
8) Data Acquisition

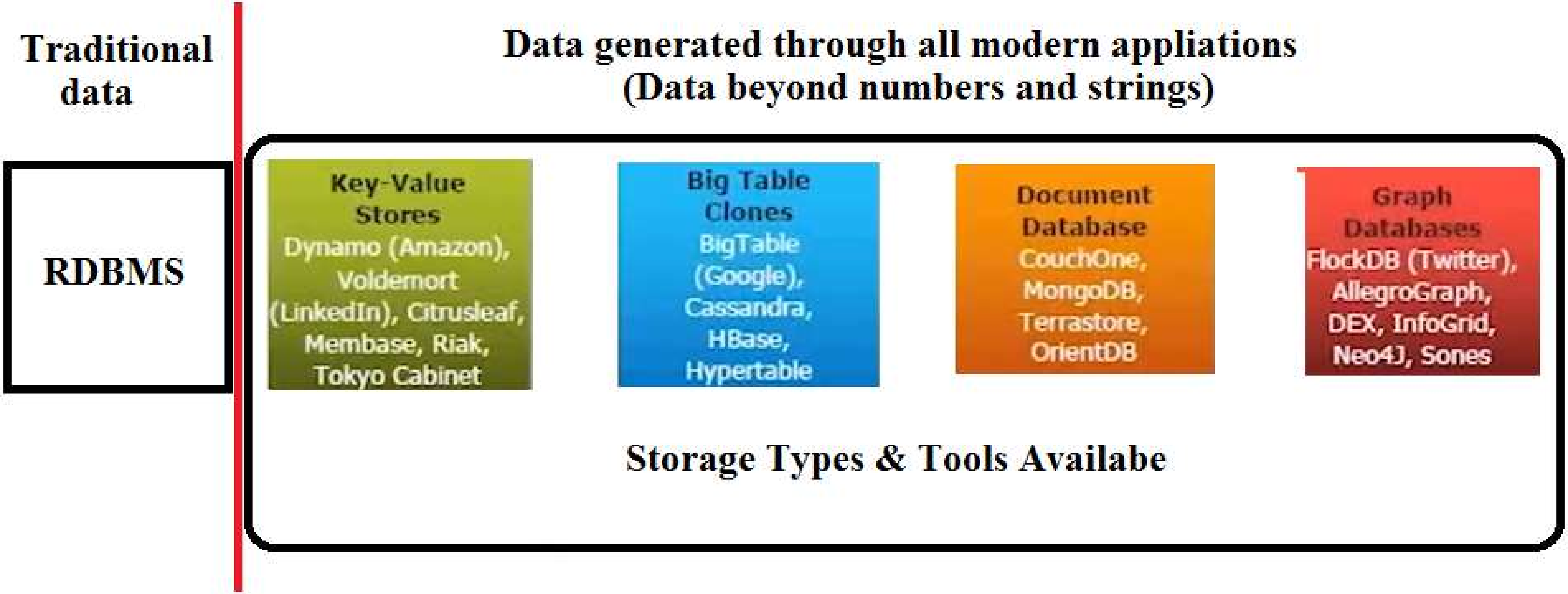
FLINK over Apache Streaming

1)Consistent Data movement

2)more realistic data streaming

3)Window based over micro batching





Traditional data

RDBMS

Data generated through all modern appliations
(Data beyond numbers and strings)

Key-Value Stores
Dynamo (Amazon), Voldemort (LinkedIn), Citrusleaf, Membase, Riak, Tokyo Cabinet

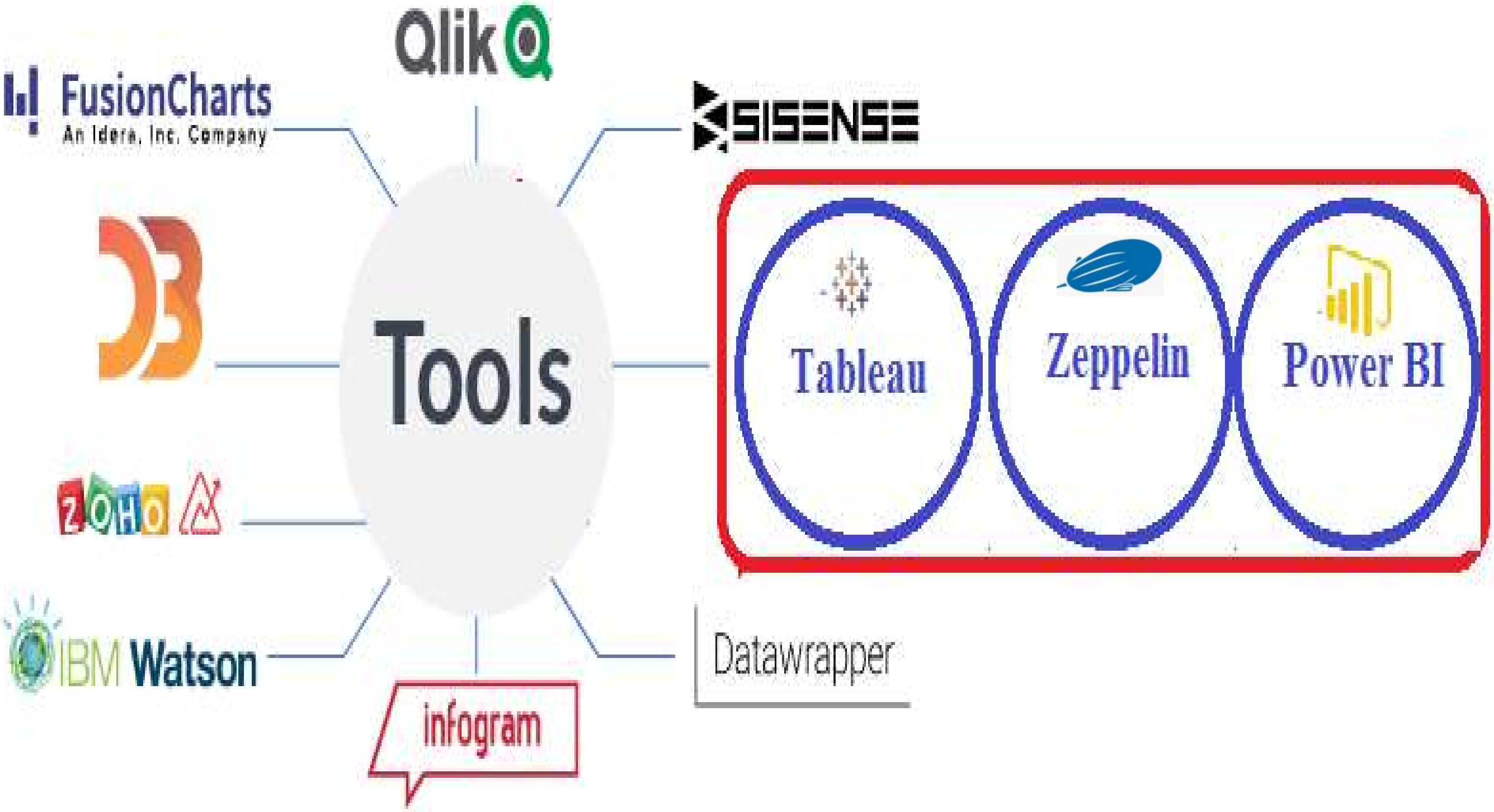
Big Table Clones
BigTable (Google), Cassandra, HBase, Hypertable

Document Database
CouchOne, MongoDB, Terrastore, OrientDB

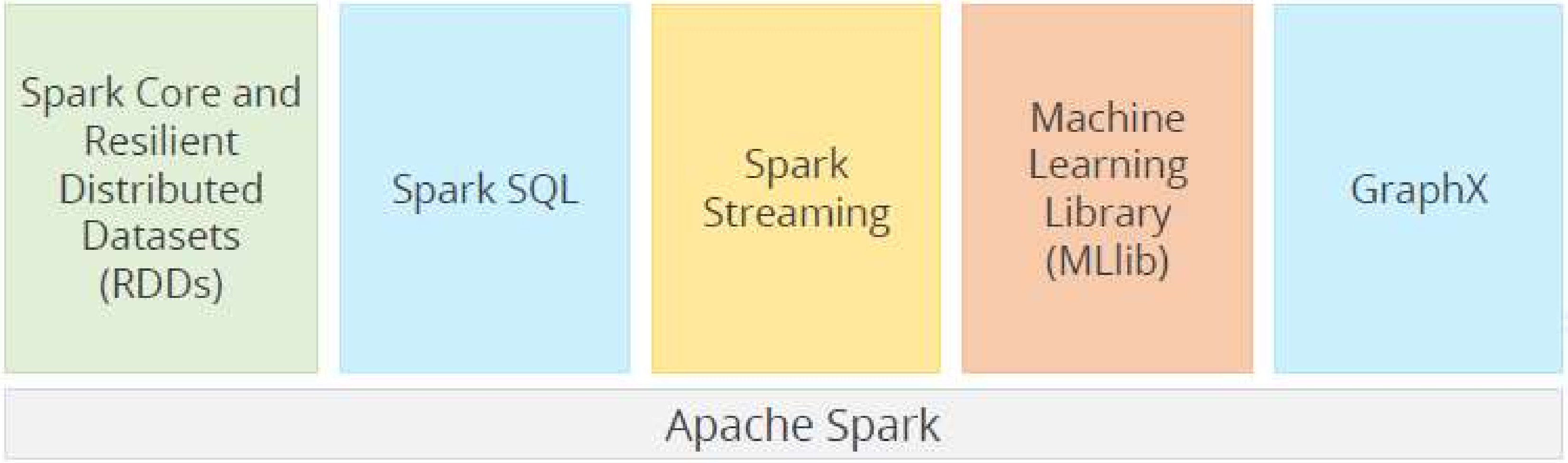
Graph Databases
FlockDB (Twitter), AllegroGraph, DEX, InfoGrid, Neo4J, Sones

Storage Types & Tools Availabe

Visualization: How you want to see!!



SPARK Components



Spark Core and Resilient Distributed Datasets (RDDs)

Spark SQL

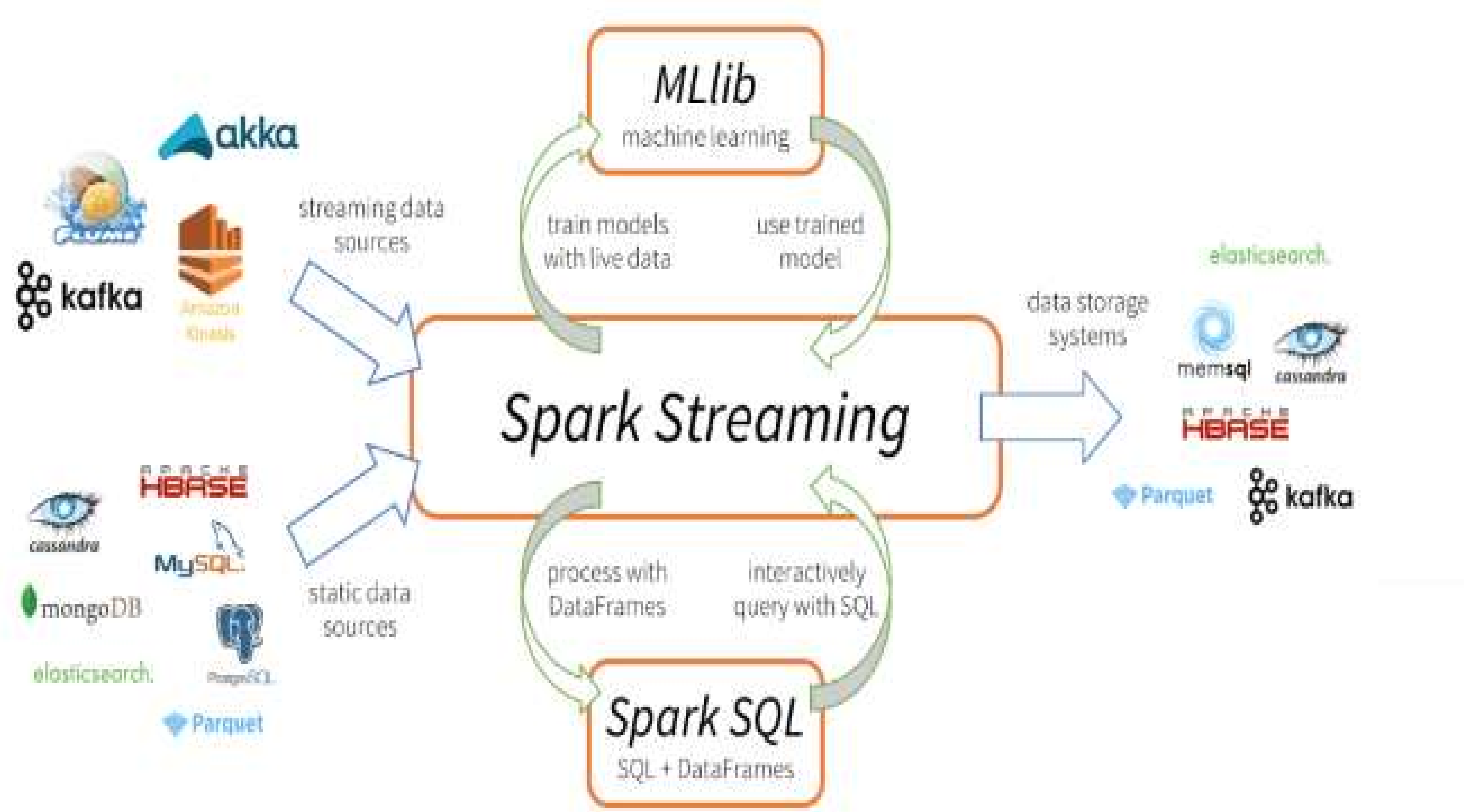
Spark Streaming

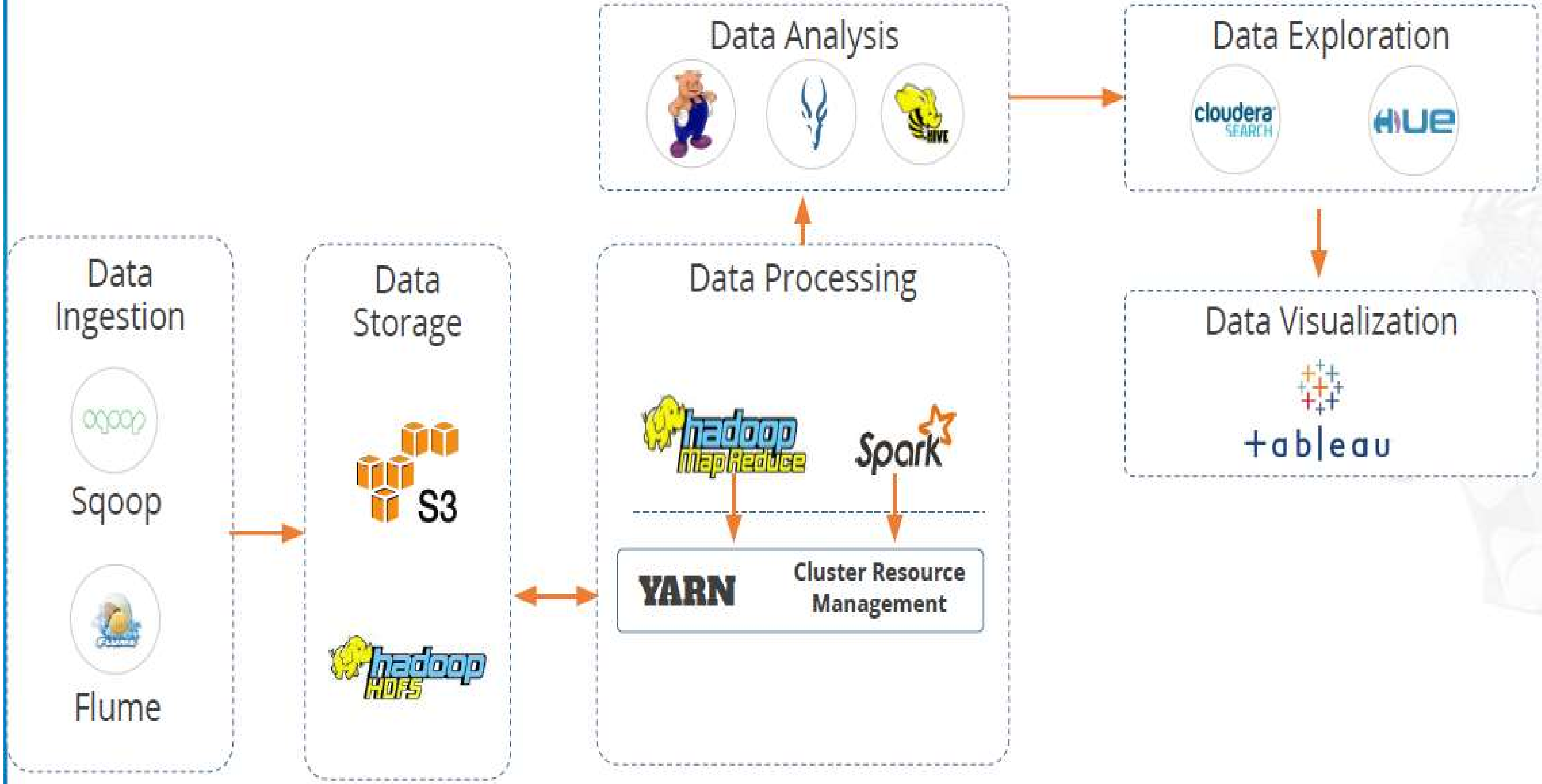
Machine Learning Library (MLlib)

GraphX

Apache Spark

Spark Streaming





Data Ingestion

Sqoop

Flume

Data Storage

S3

Data Processing

YARN

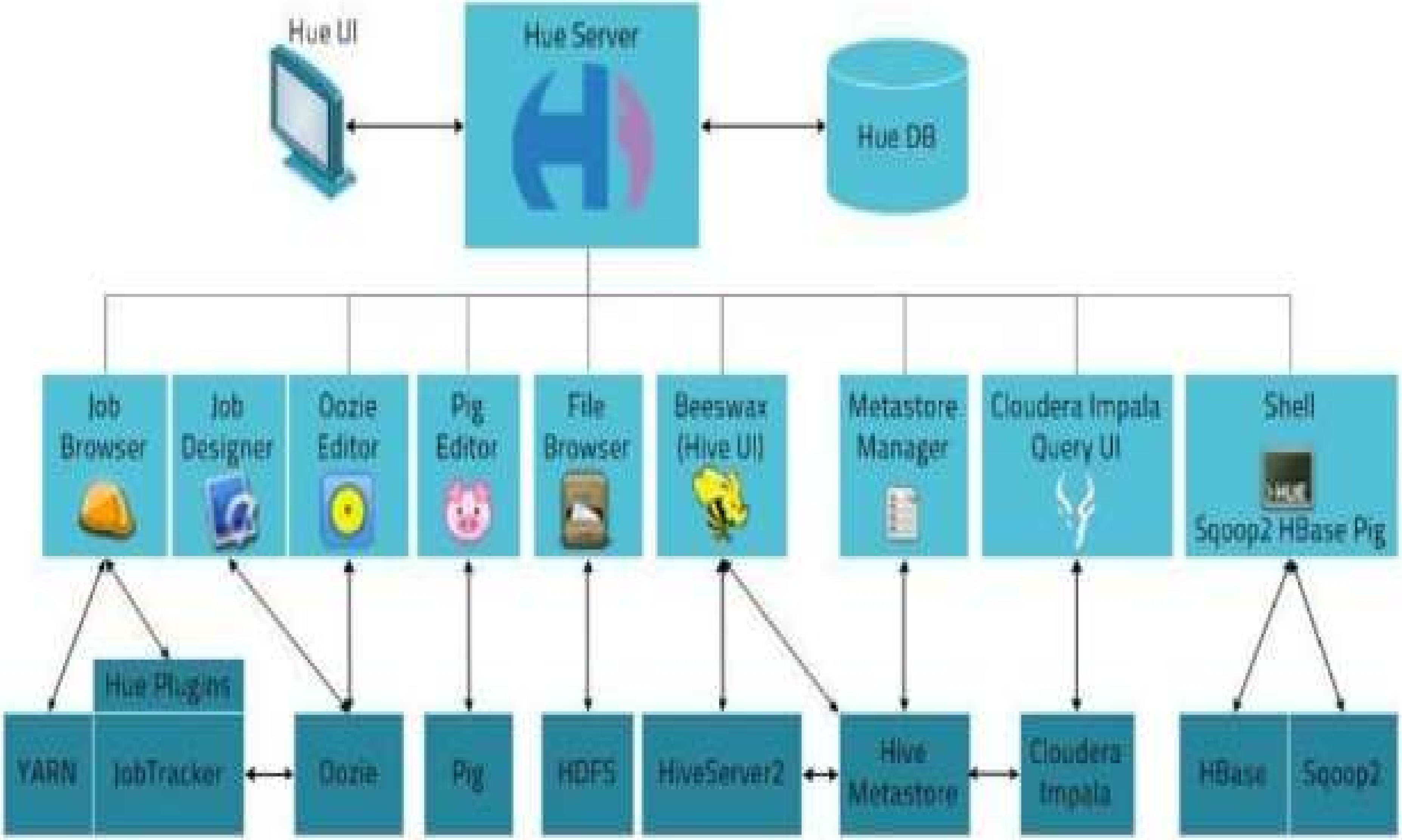
Cluster Resource Management

Data Analysis

Data Exploration

Data Visualization

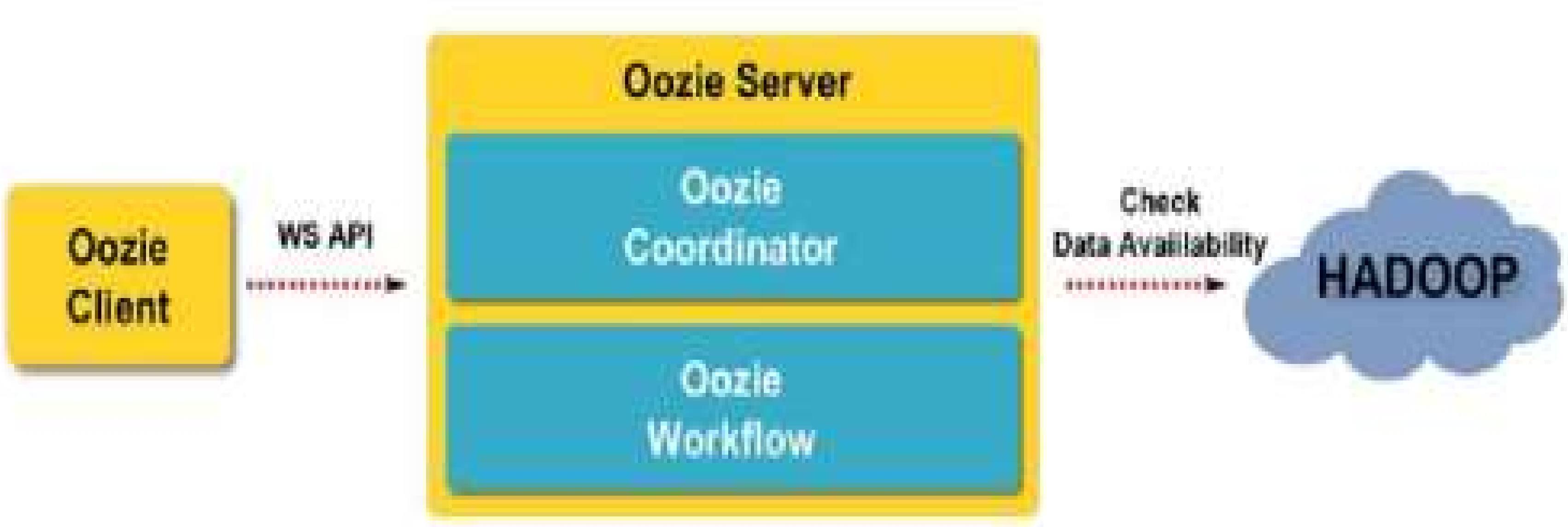
Interface to Hadoop : HUE



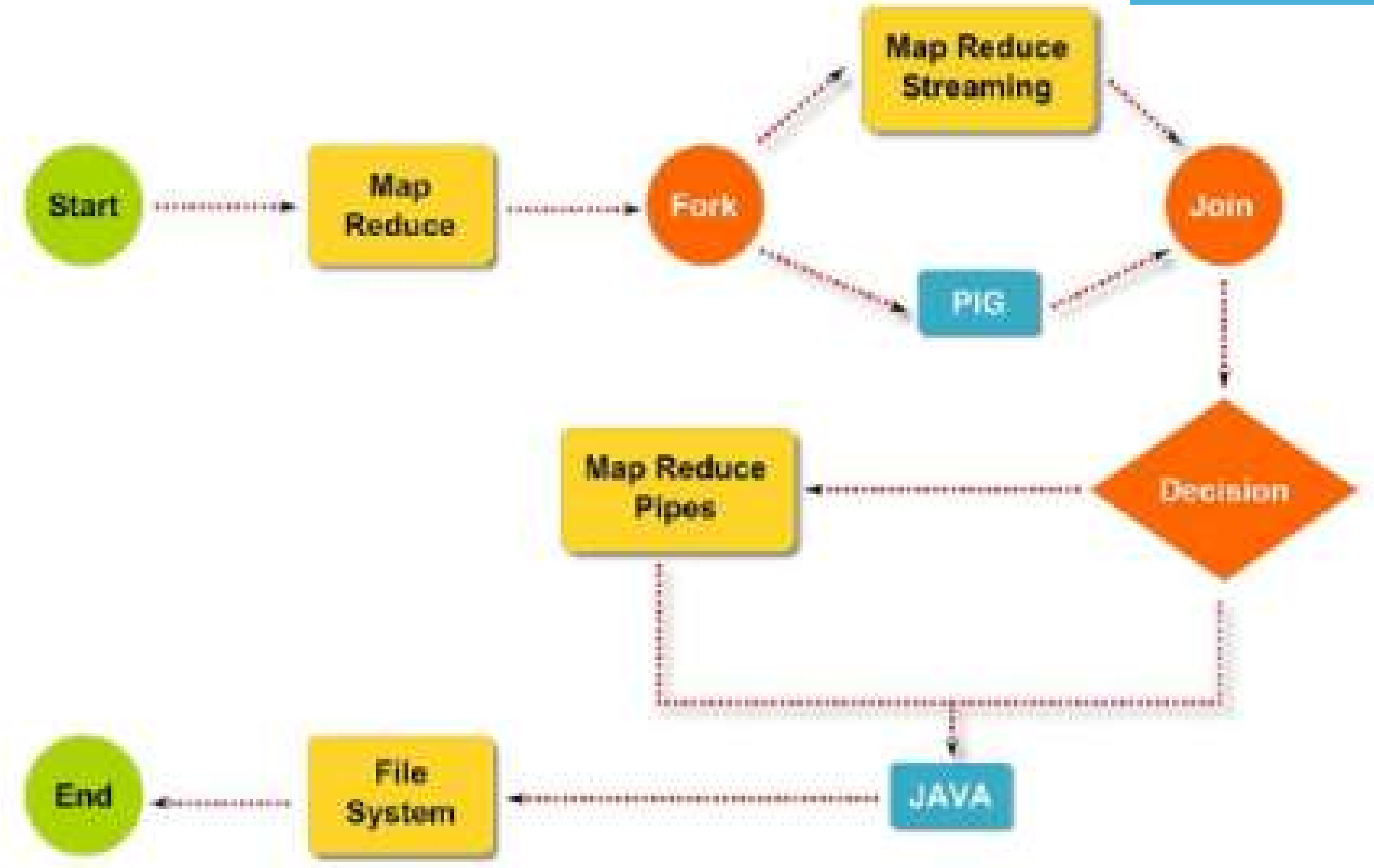
Hadoop Task Scheduler : OOZIE
Scheduling Batch Jobs

Oozie executes workflow based on

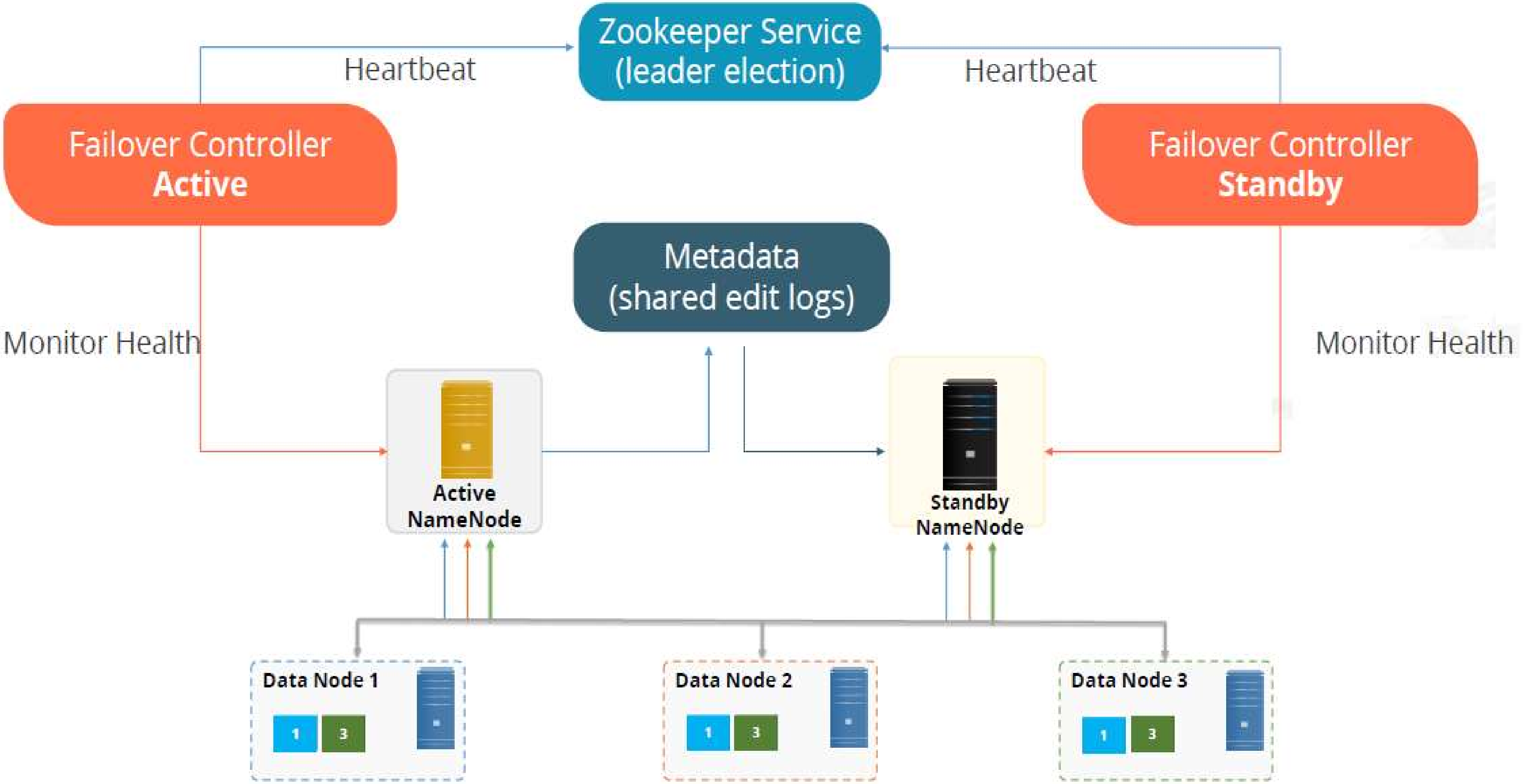
- Time Dependency (Frequency)
- Data Dependency



Sample Work Flow: OOZIE

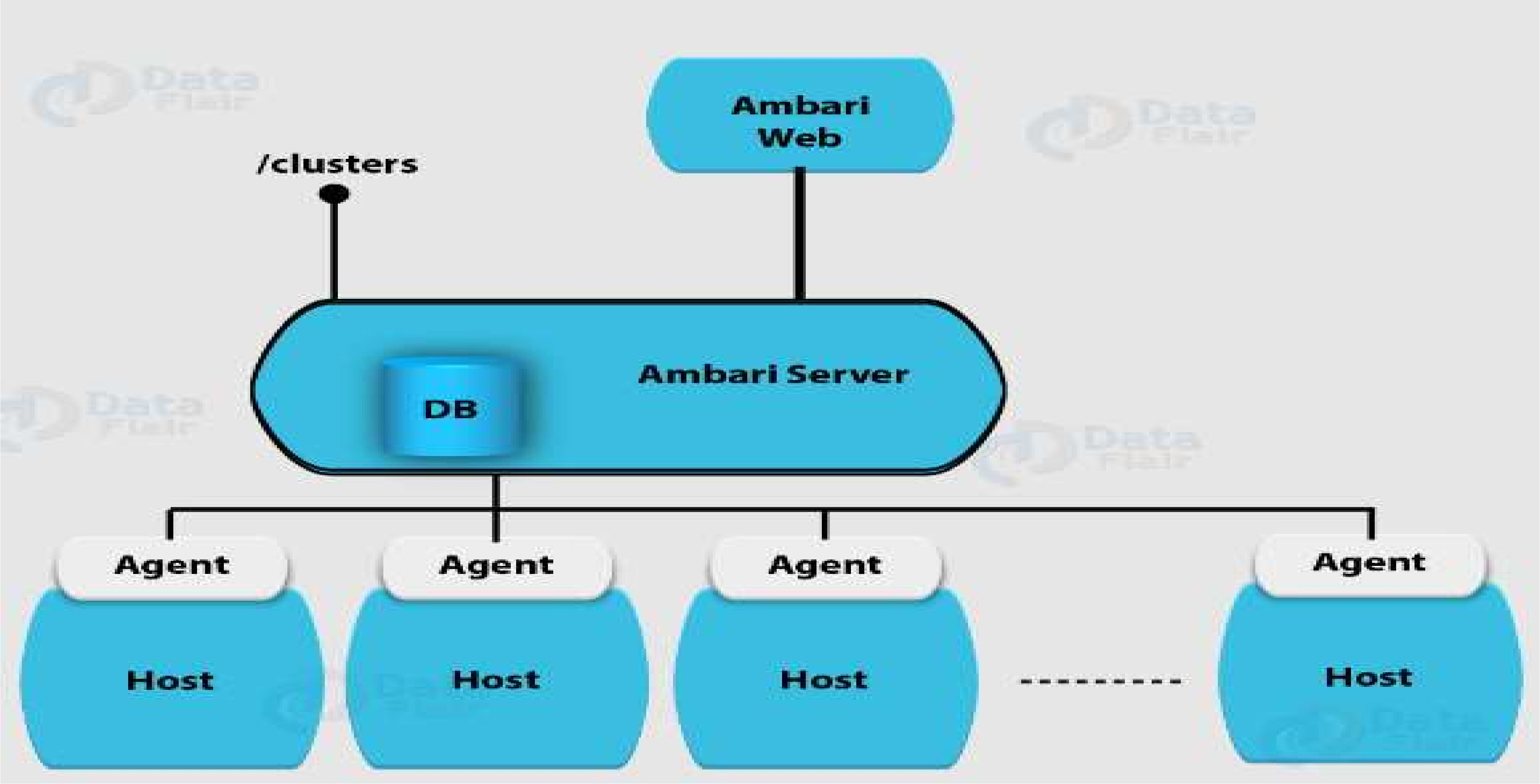


Hadoop Availability – Zookeeper Works with HBase



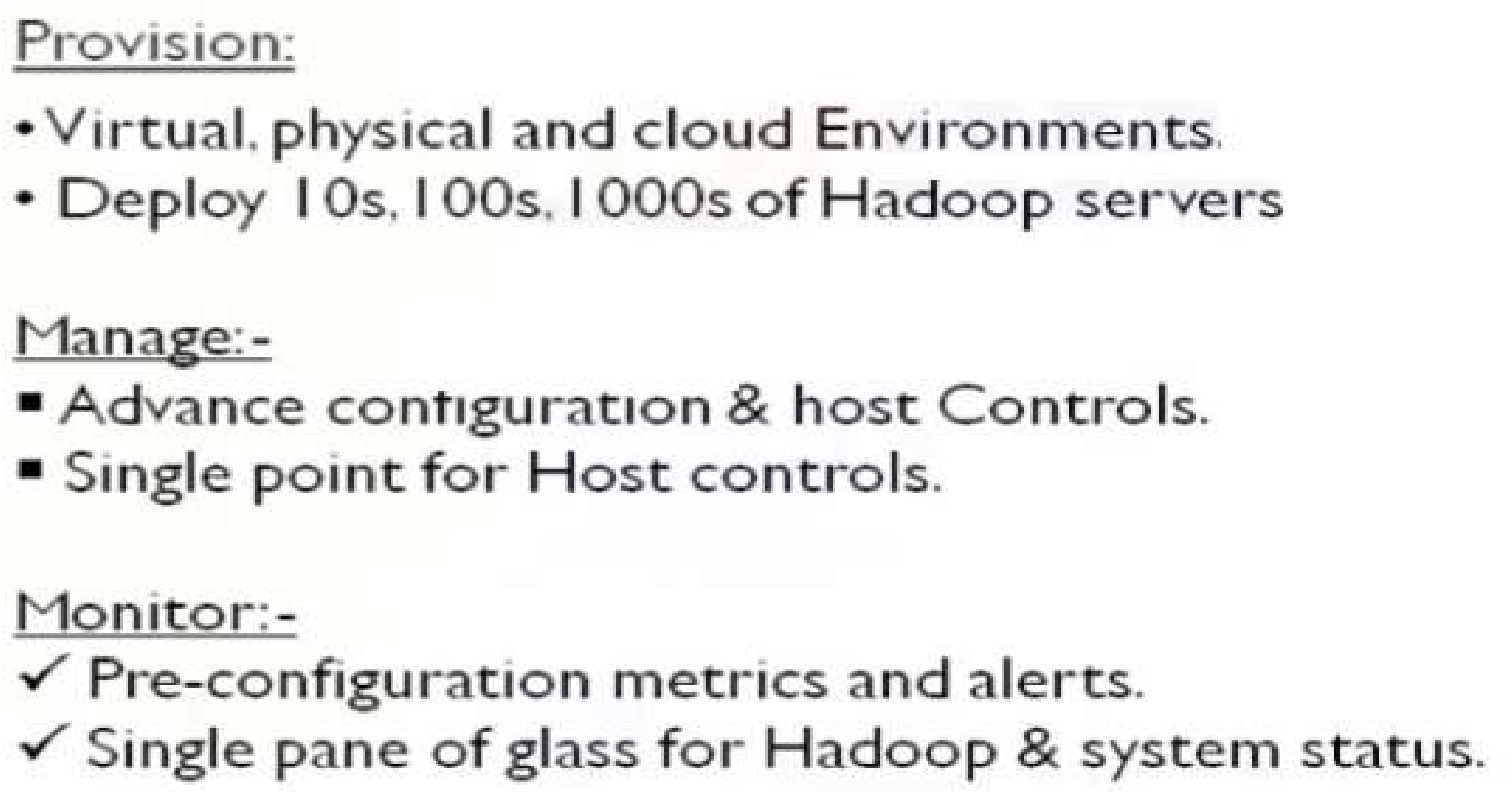
Hadoop Administration : Apache Ambari

Used for management of Apache Hadoop clusters using a web UI. It also integrates with other existing applications using Ambari REST APIs.



Hadoop Administration : Apache Ambari

Ambari



Provision:

- Virtual, physical and cloud Environments.
- Deploy 10s, 100s, 1000s of Hadoop servers

Manage:-

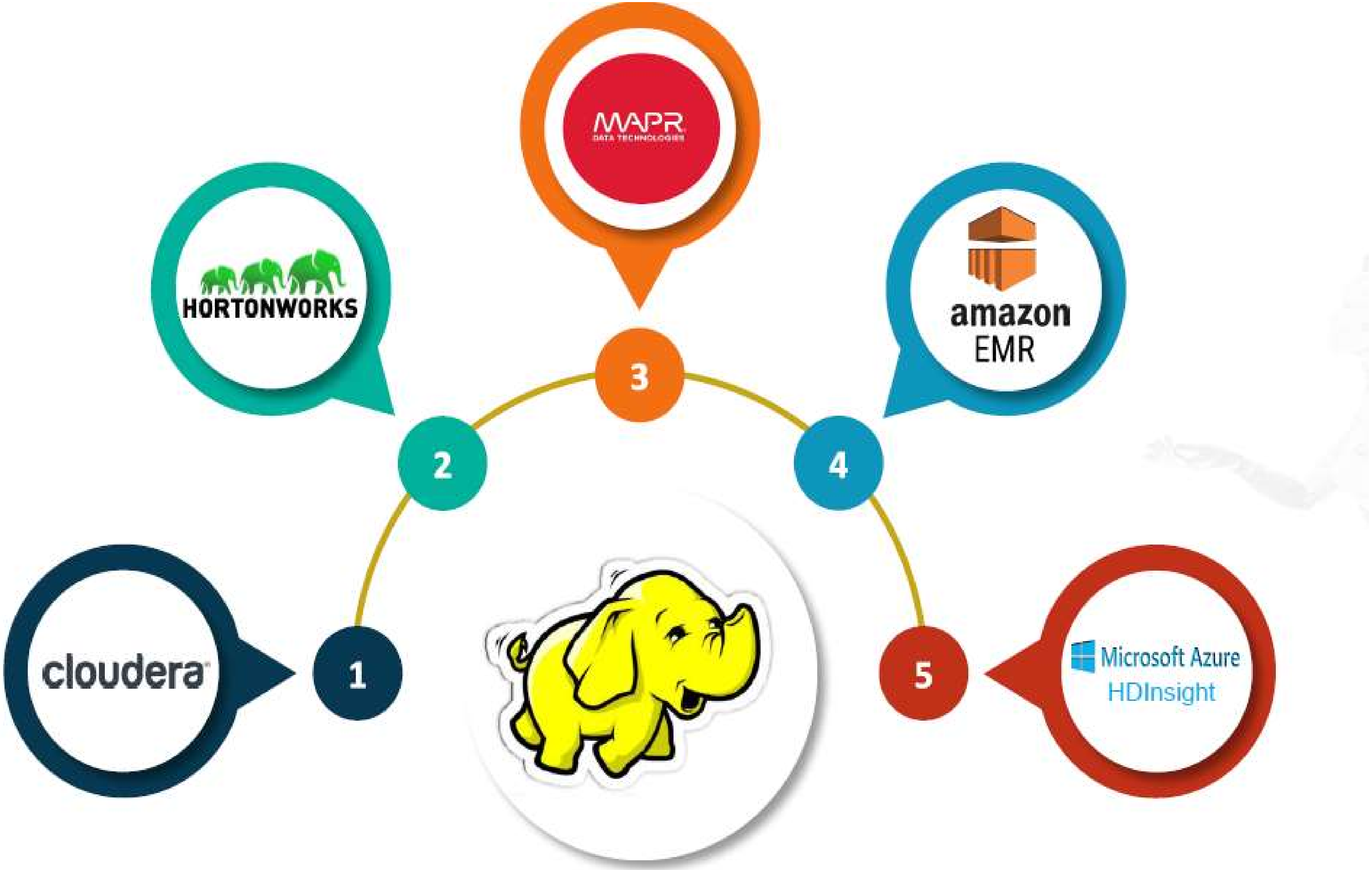
- ▪ Advance configuration & host Controls.
- ▪ Single point for Host controls.

Monitor:-

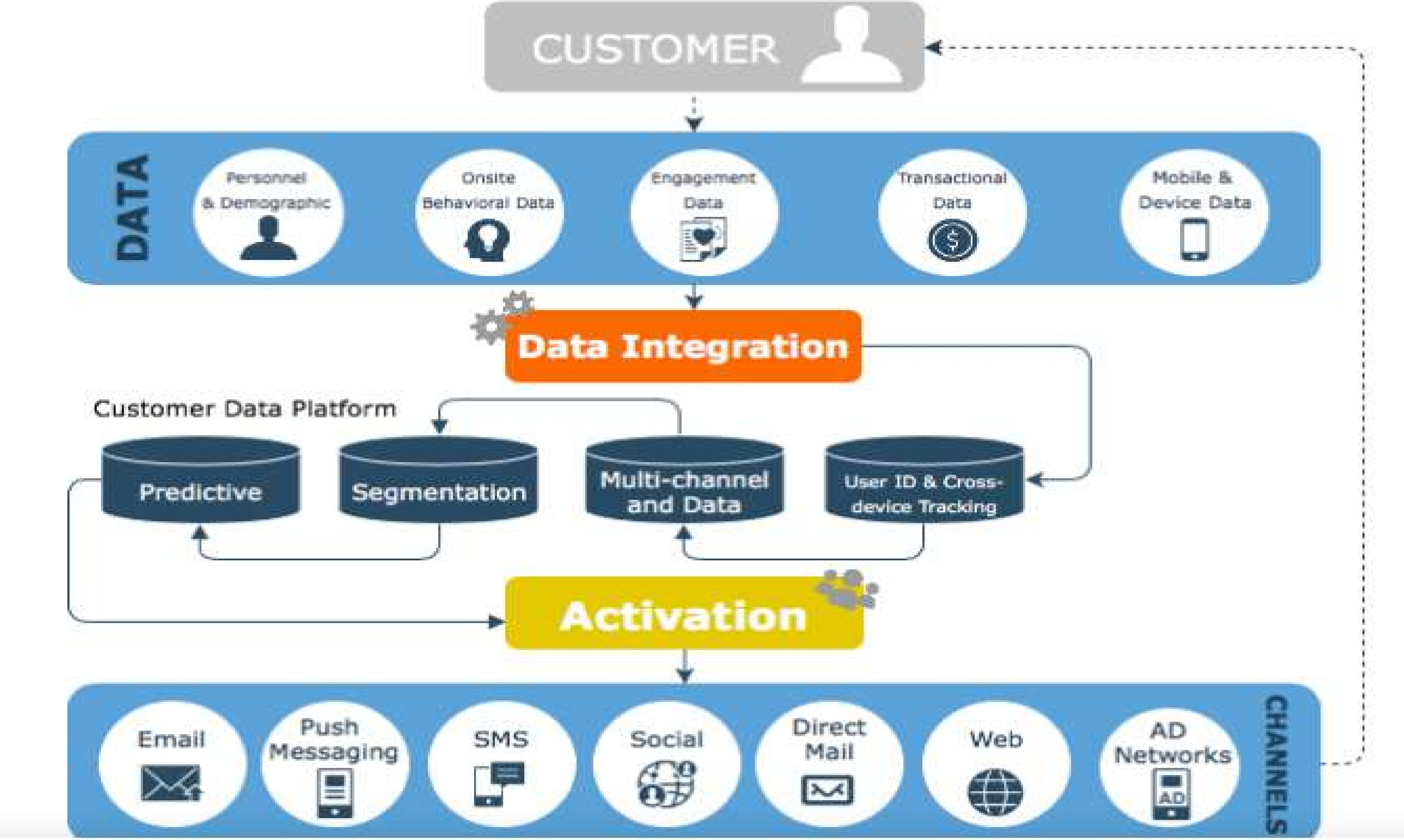
✓ Pre-configuration metrics and alerts.

✓ Single pane of glass for Hadoop & system status.

Market Players



Trending - CDC



Quick Summarization

- Exact need
- Form in which data is available
- Data type : perishable/Nonperishable

Based on the answers obtained ..Find out

- What are the tools available
- How to use those tools
- Do you need to be a programming expert
- Organization protocols/ Infra Prerequisite
- paid or open source

STREAMLINE YOUR OPERATION

- Are you planning to have your own setup ? ..**Bigger Question**
- In what form data is available ?
- What is the speed at which data arrives ?
- Direct access to the data source is available ?
- Do you need to send data to multiple processing tools as well as storage device ?

Data Ingestion

- Are you going to store data using Hadoop native component or proprietary tools ? **Data storage**
- Do you need real time processing ?
- Do you need to take immediate action using data thresholds ?

Data processing

- Do you need to monitor data for decision making?

Data visualization

References

- https://hadoop.apache.org/ -- Apache Foundation
- https://www.ibm.com/analytics/hadoop/big-data-analytics ---IBM
- https://azure.microsoft.com/en-in/solutions/big-data/ - AZURE
- Great Learning – Raghu Raman

This is the Beginning!!

SO

Big thank You