



SINGAPORE
ACTUARIAL
SOCIETY

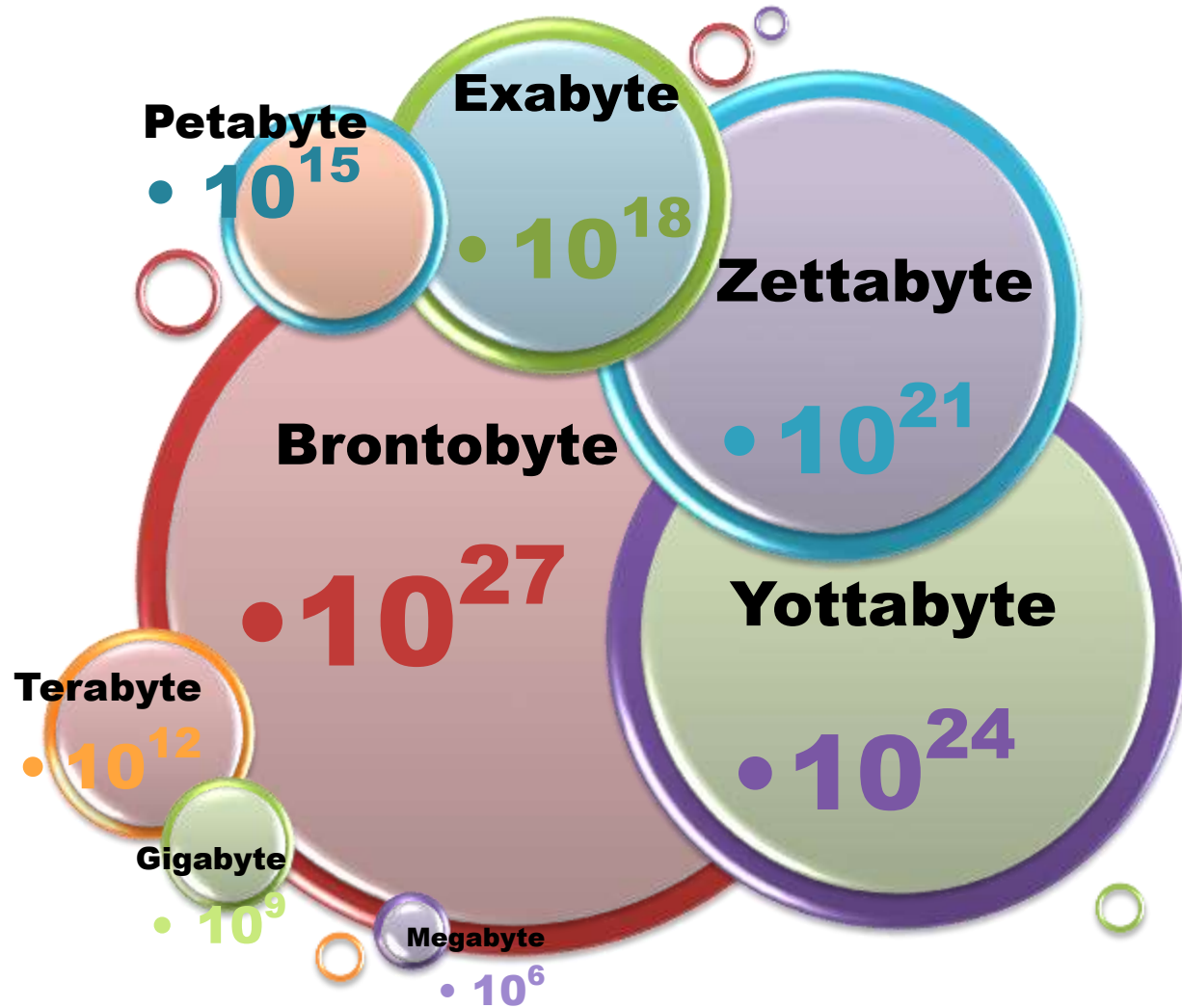
The Bolts and Nuts of building a Big Data infrastructure

Presented to you by eRic Choo

Prepared by JK Wong



So much DATA!!! How to handle?



Source: <http://www.ibm.com/IoT>

Can we (human) still handle them?



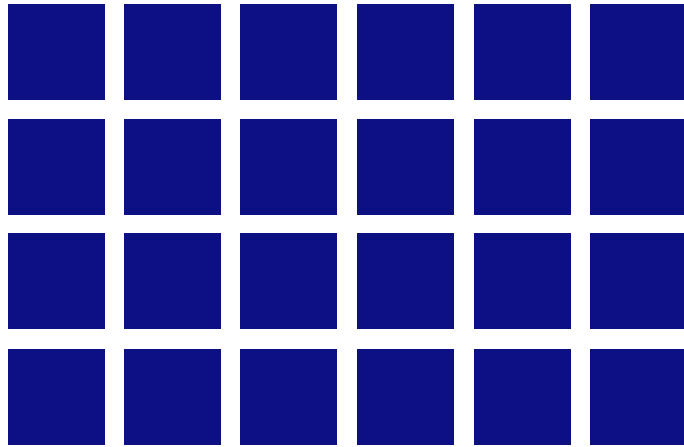
SILOs



of DATA

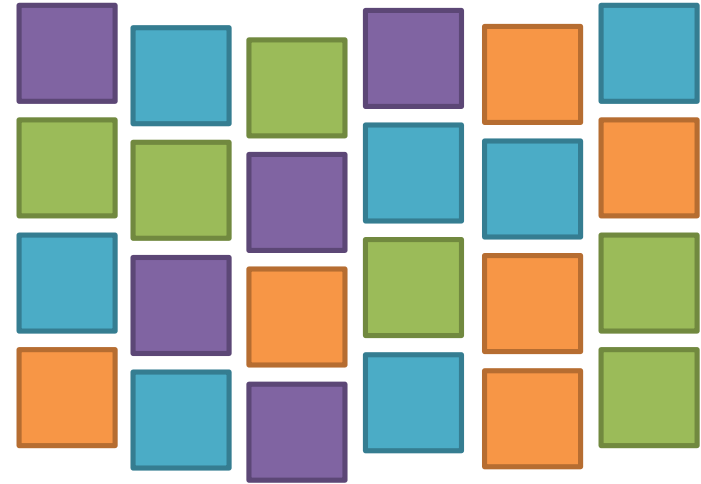


Structured Data



What you find in a DB
(typically)

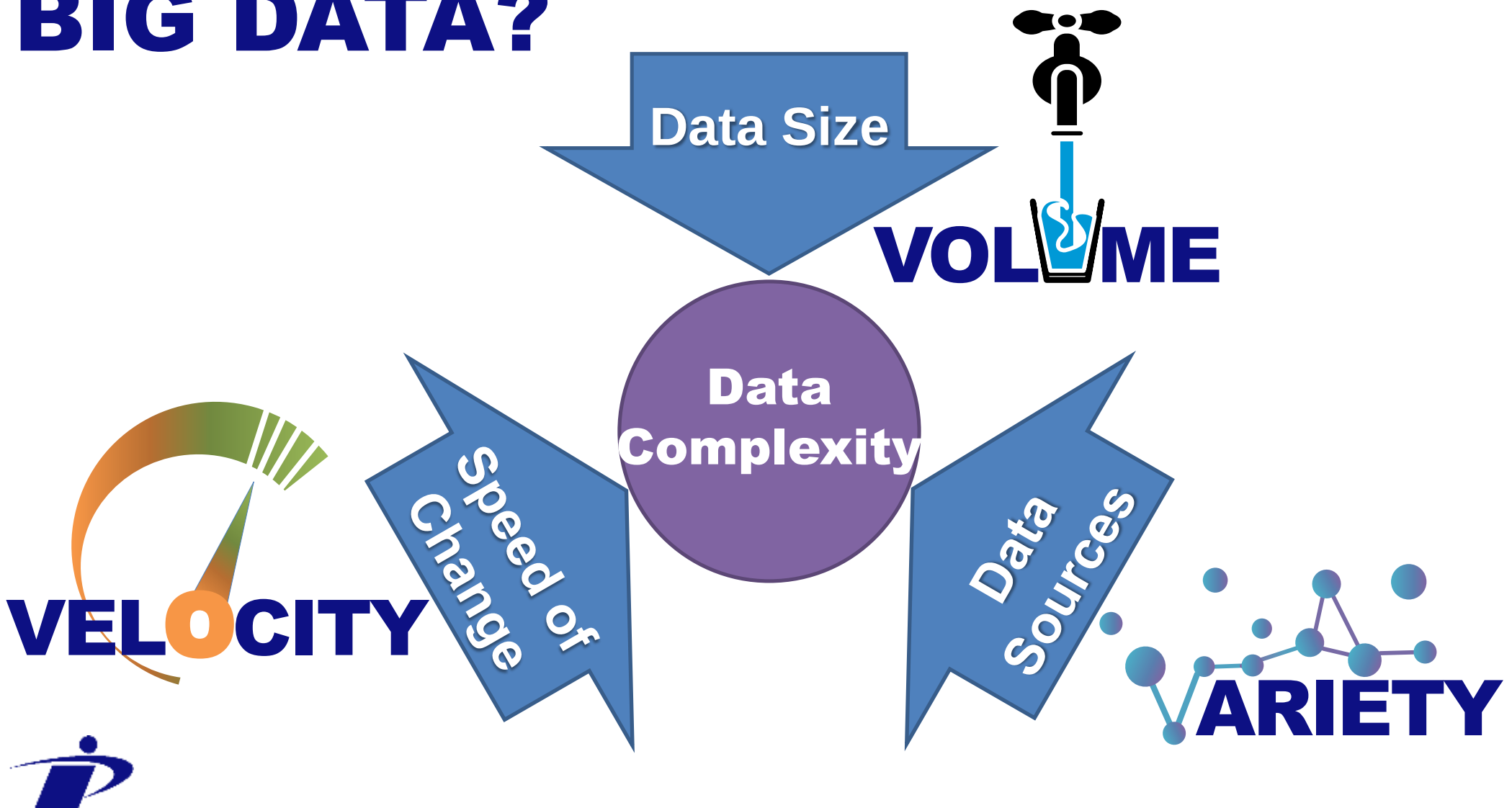
Unstructured Data



What you find in the 'wild'
(text, image, audio, video)



BIG DATA?





We need to
simplify
'simplify
fΛI/
verb



Traditional Processing

- Processor bound
- Complex problem = Larger computer

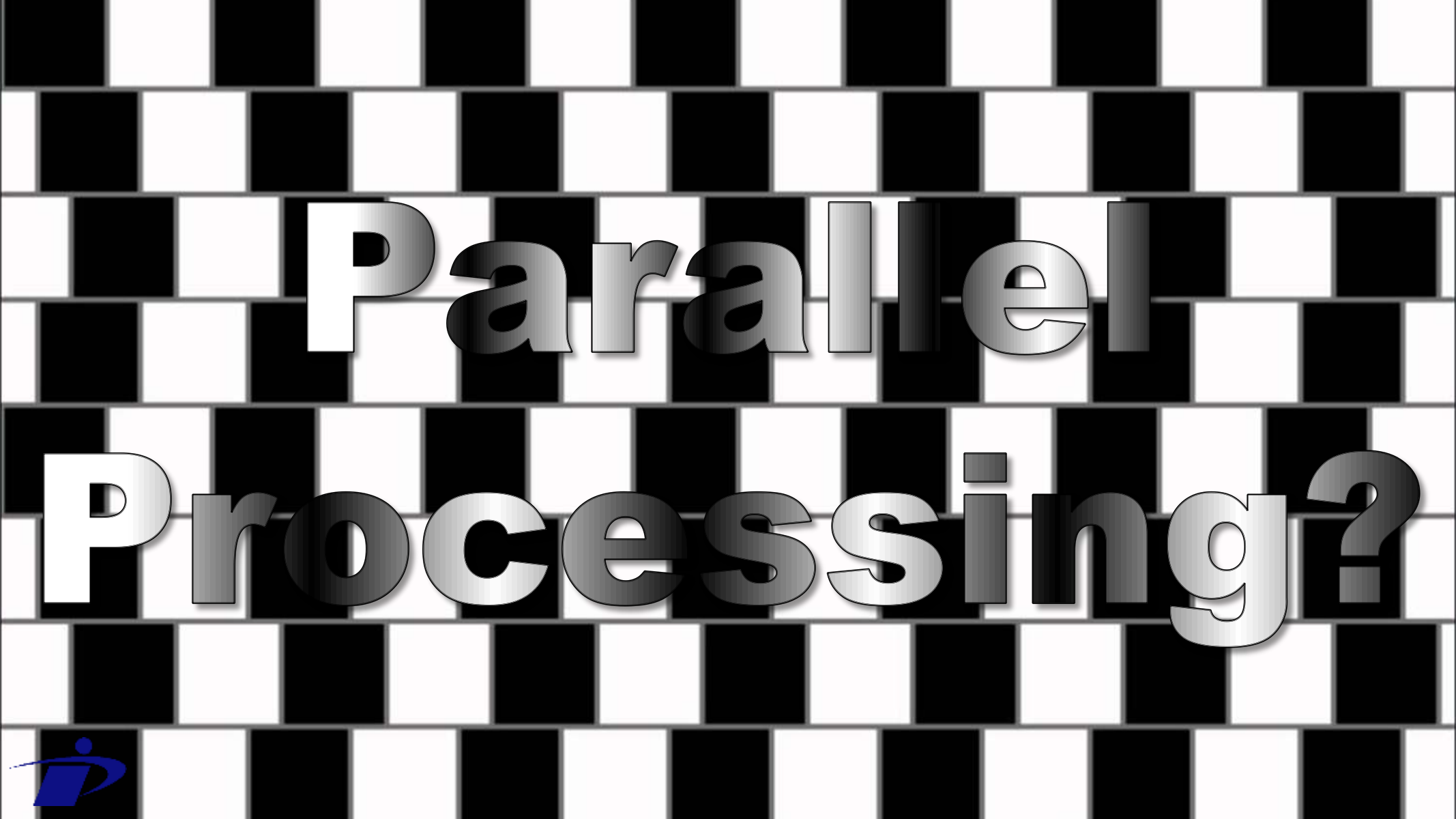


Advancement in Processing

- Parallel Computing
- To solve problem(s) at the same time
- High-end Computers becoming affordable
 - Processor
 - Memory
 - HDD

↓ **Prices**



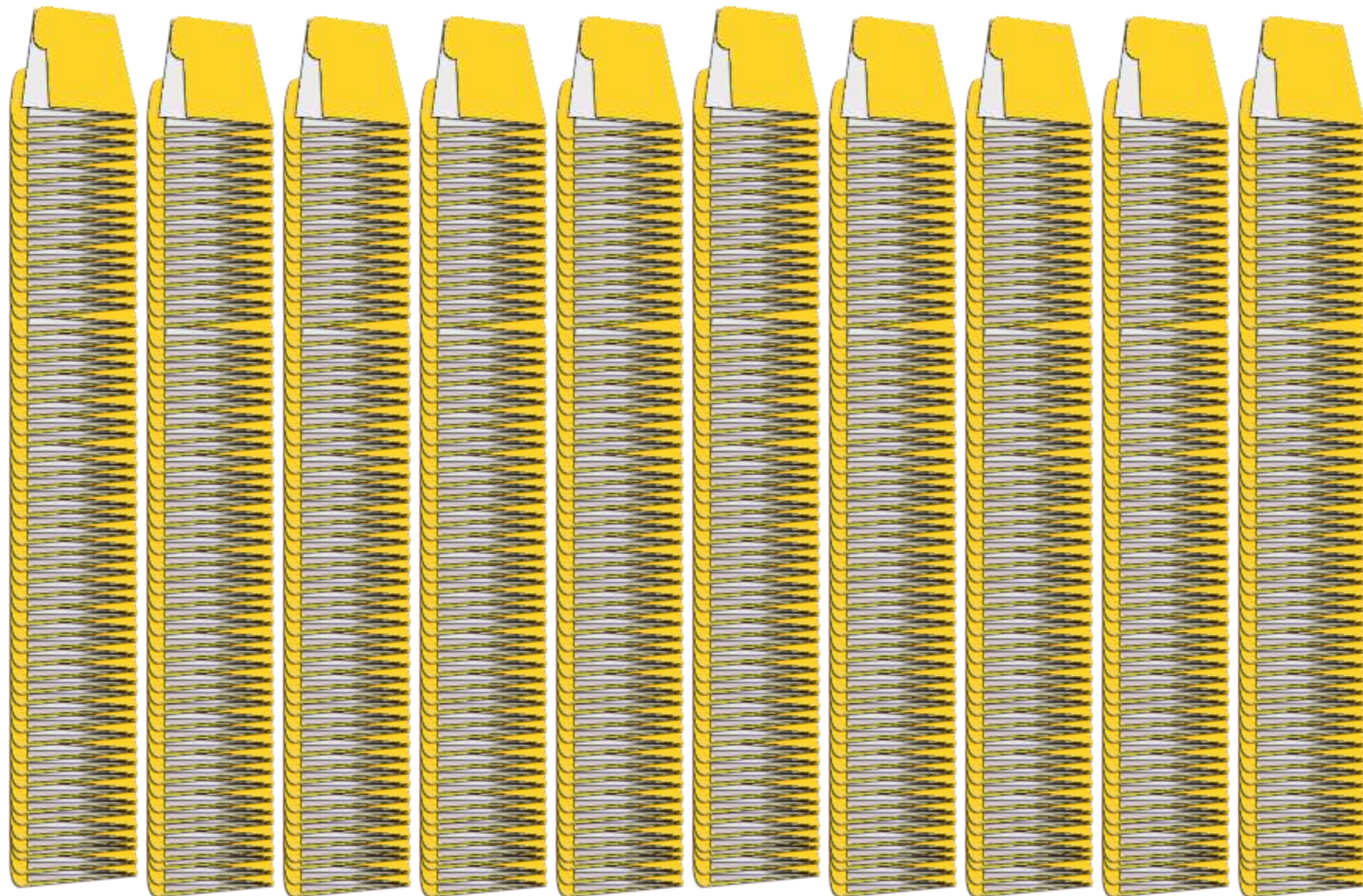


Parallel Processing?

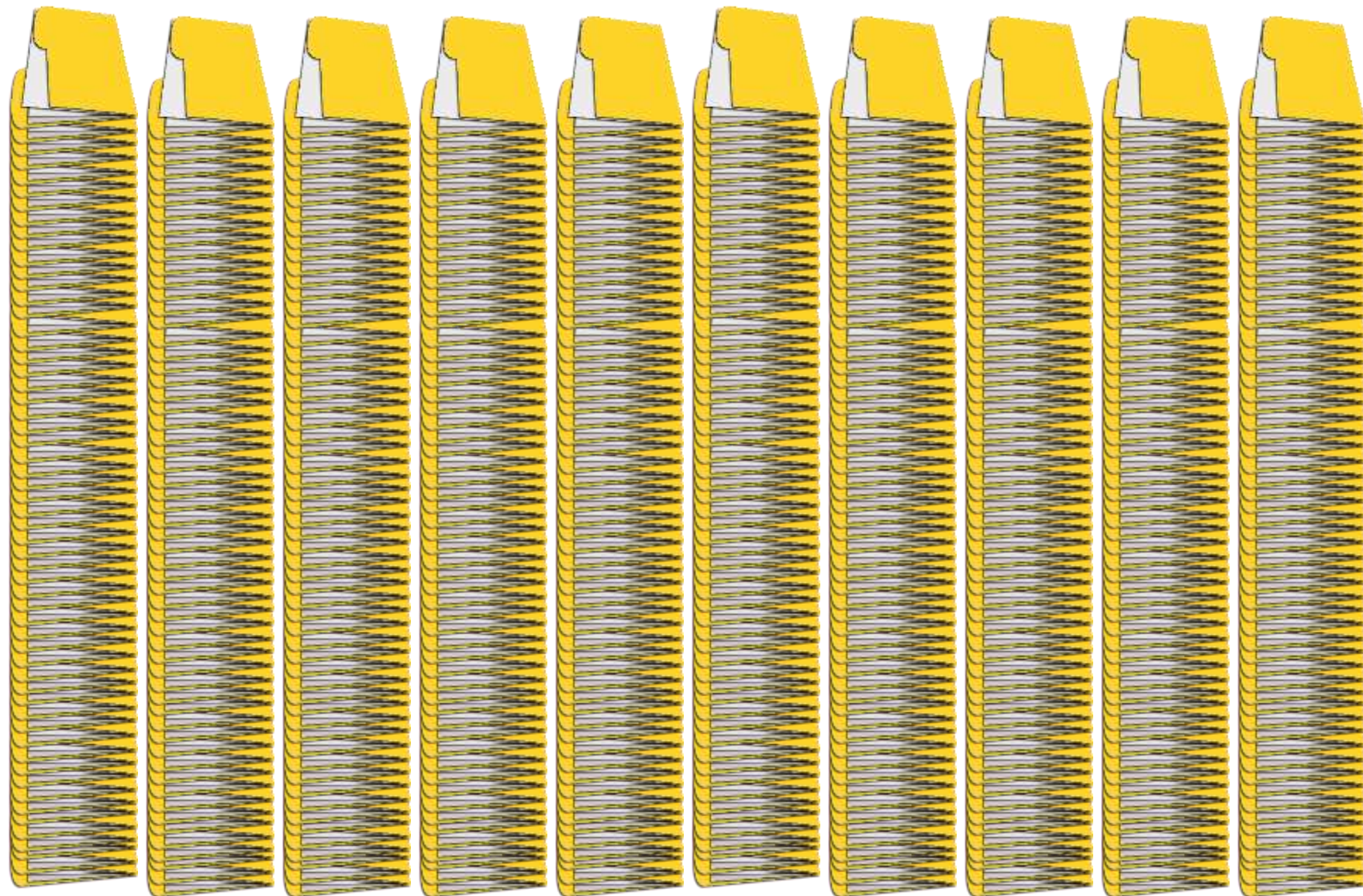






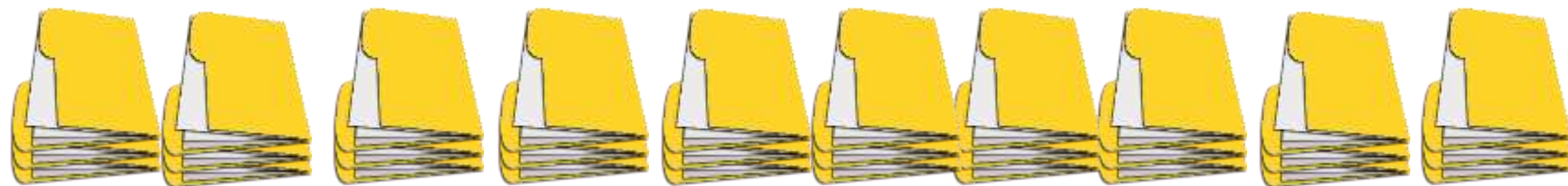
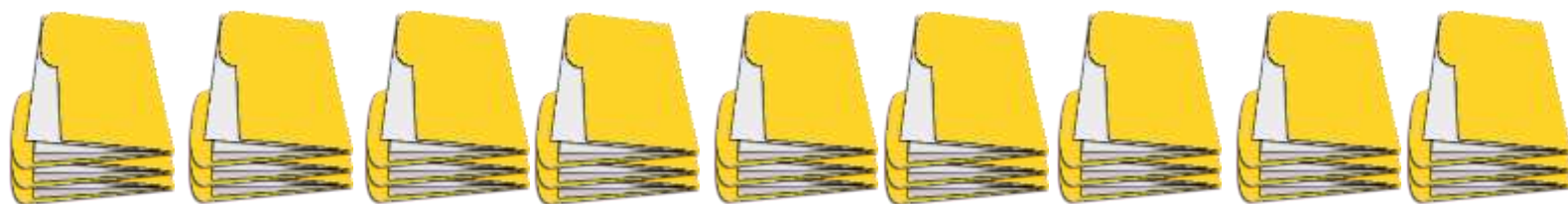


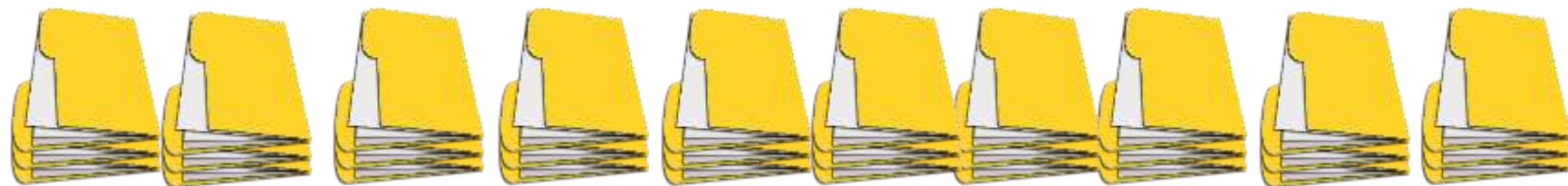
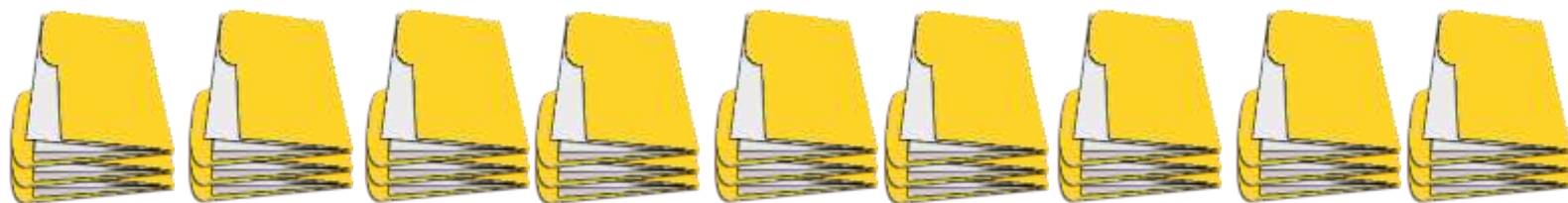


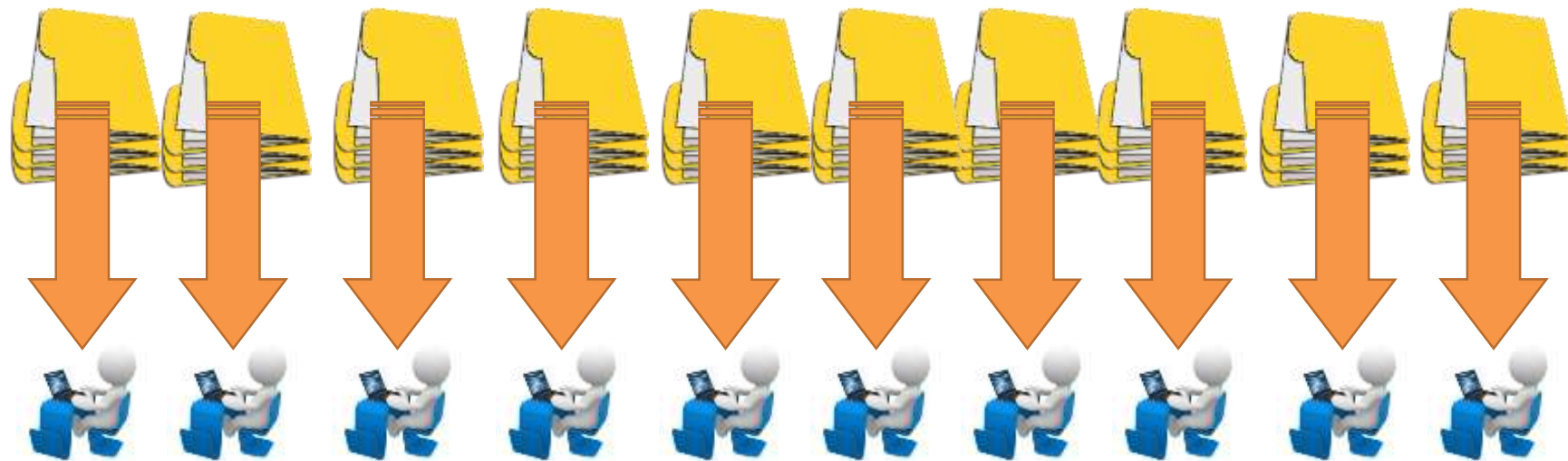
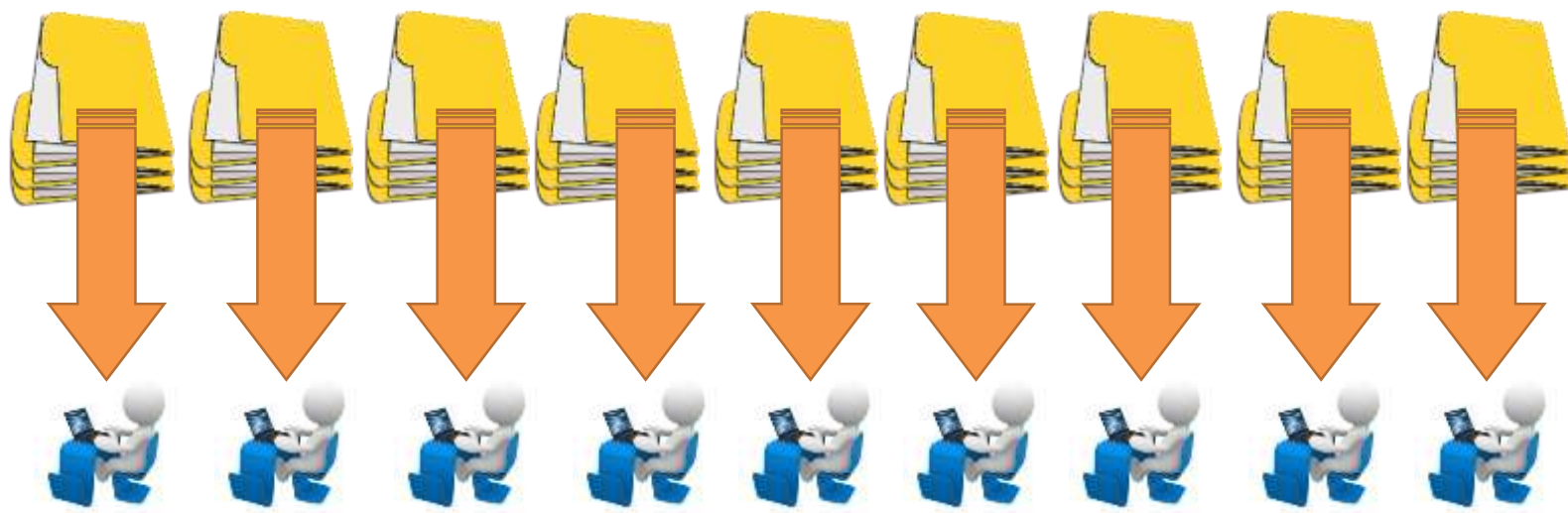












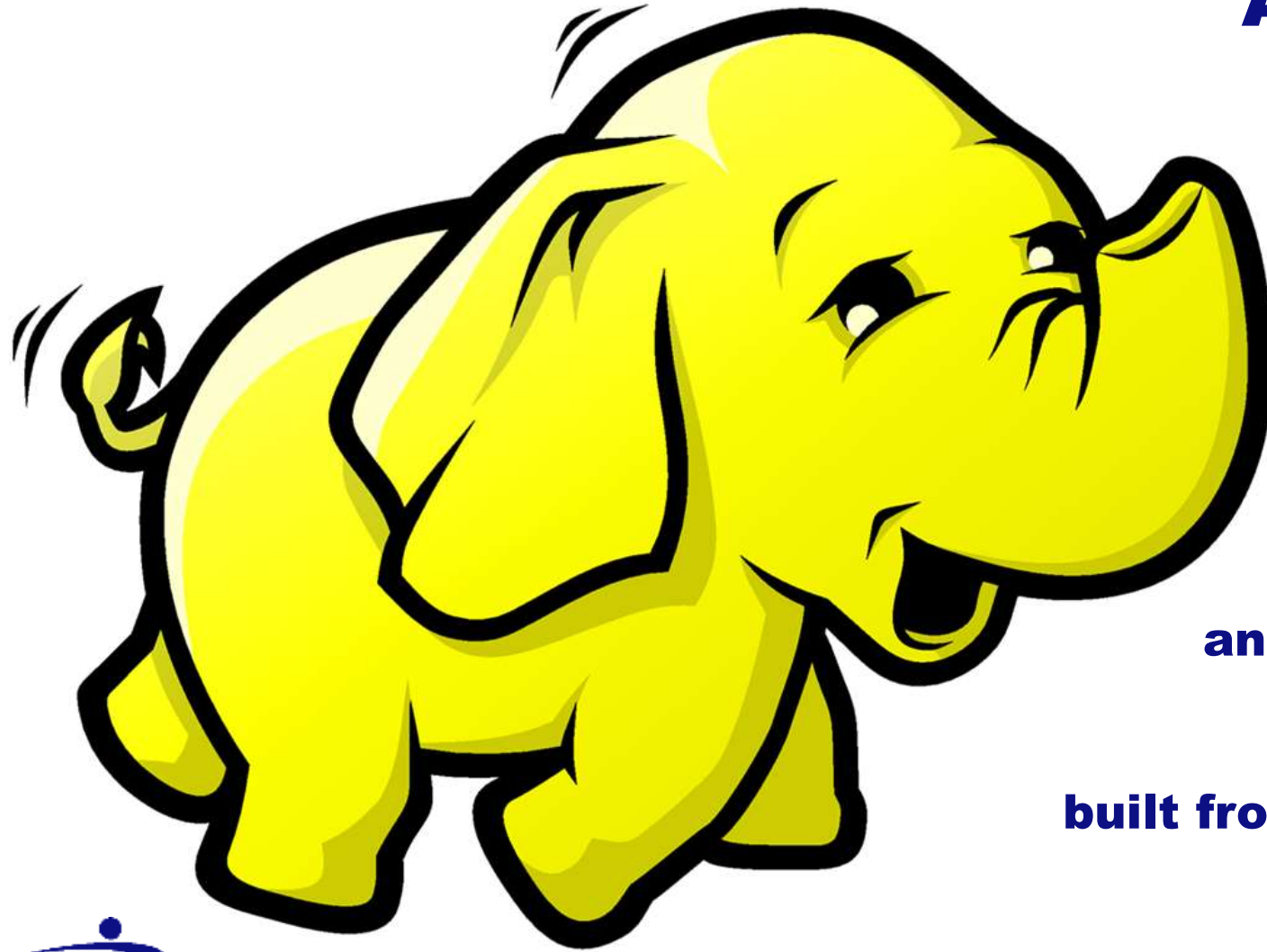






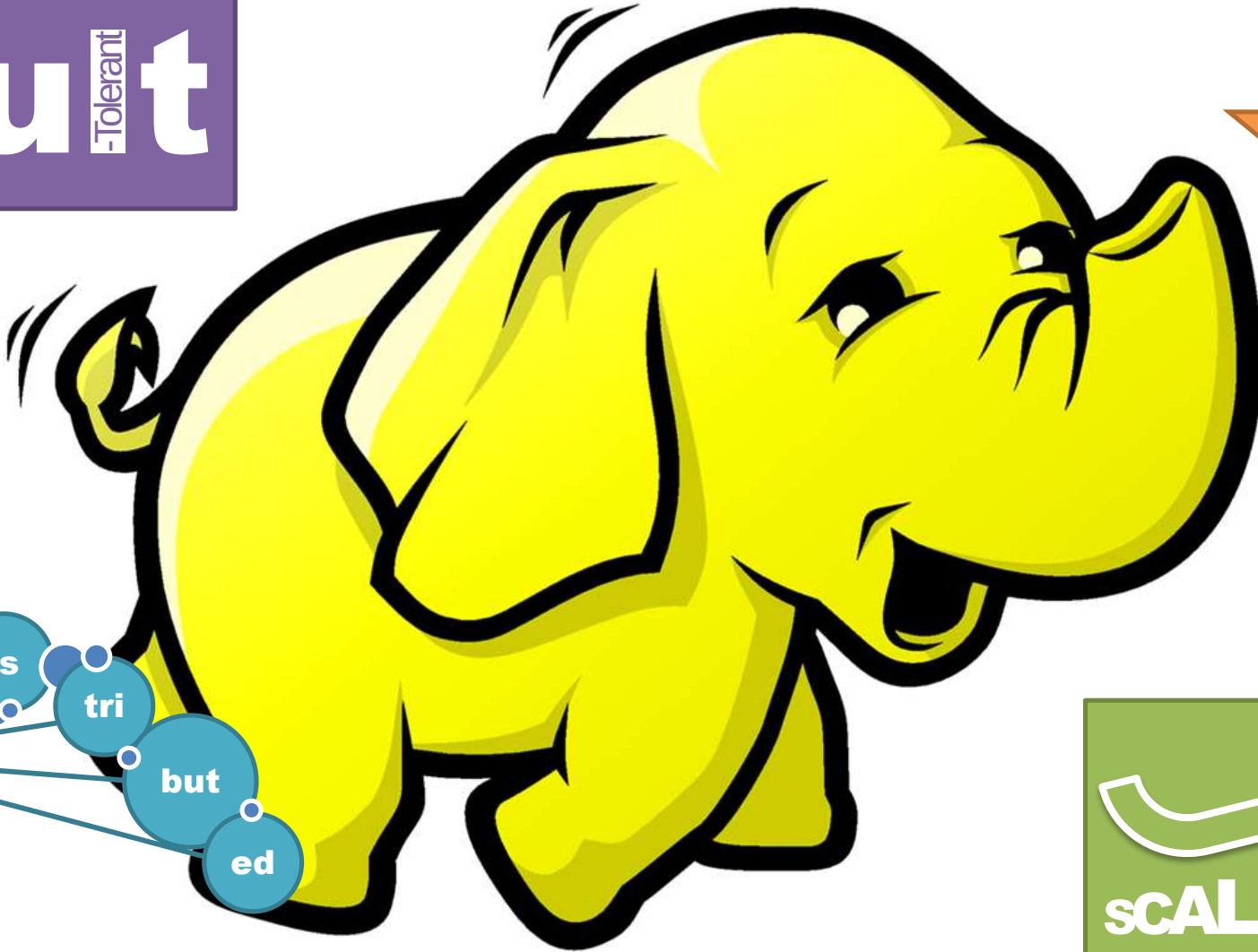
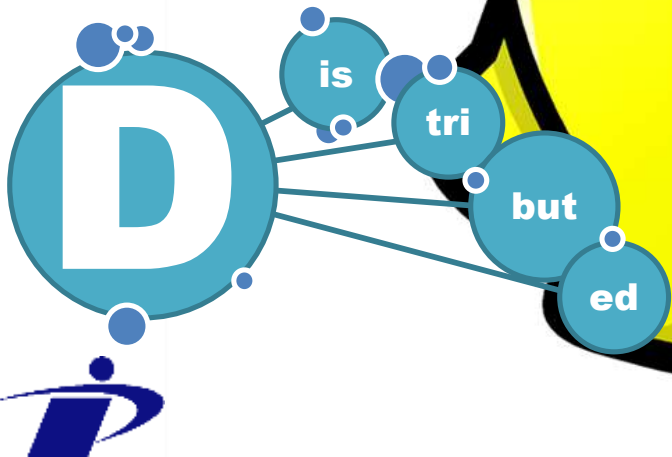
Apache Hadoop **(pronunciation:** **/hə'du:p/)**

**is an open-source
software framework
for distributed storage
and distributed processing
of very large data sets
on computer clusters
built from commodity hardware.**



Fault
-Tolerant

cost



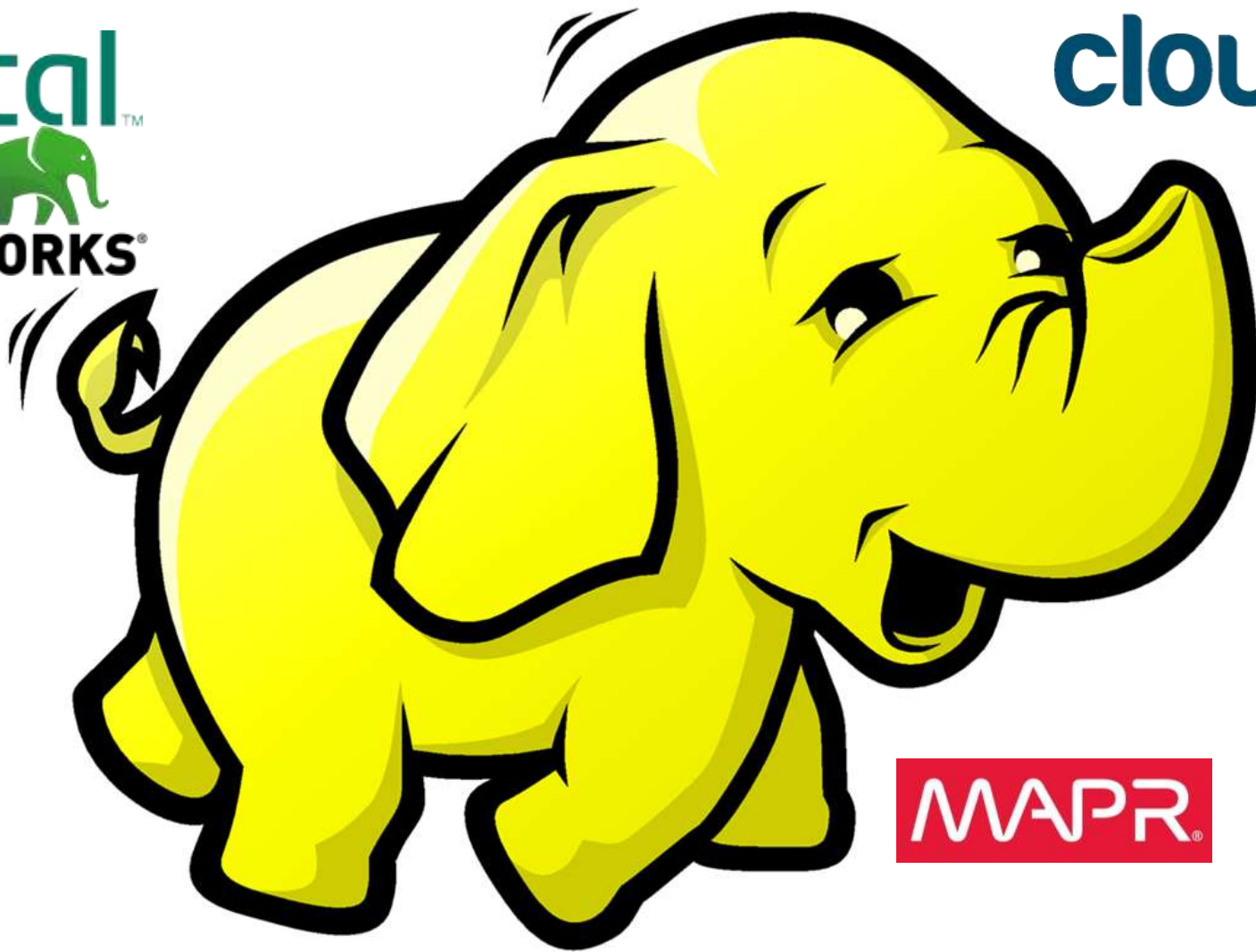
scalable

Pivotal™



HORTONWORKS®

cloudera



MAPR®



Hadoop Layers

WORKLOAD MANAGEMENT

PROCESSING

STORAGE



Core Hadoop Components

**WORKLOAD
MANAGEMENT**

YARN

PROCESSING

MapReduce

Spark

STORAGE

HDFS



**BATCH
PROCESSING**
(MapReduce,
Hive, Pig)

**ANALYTICS
SQL**
(Impala)

**SEARCH
ENGINE**

**MACHINE
LEARNING**
(Spark,
MapReduce,
Mahout)

**STREAM
PROCESSING**
(Spark)

**3RD PARTY
APPS**

WORKLOAD MANAGEMENT
(YARN)

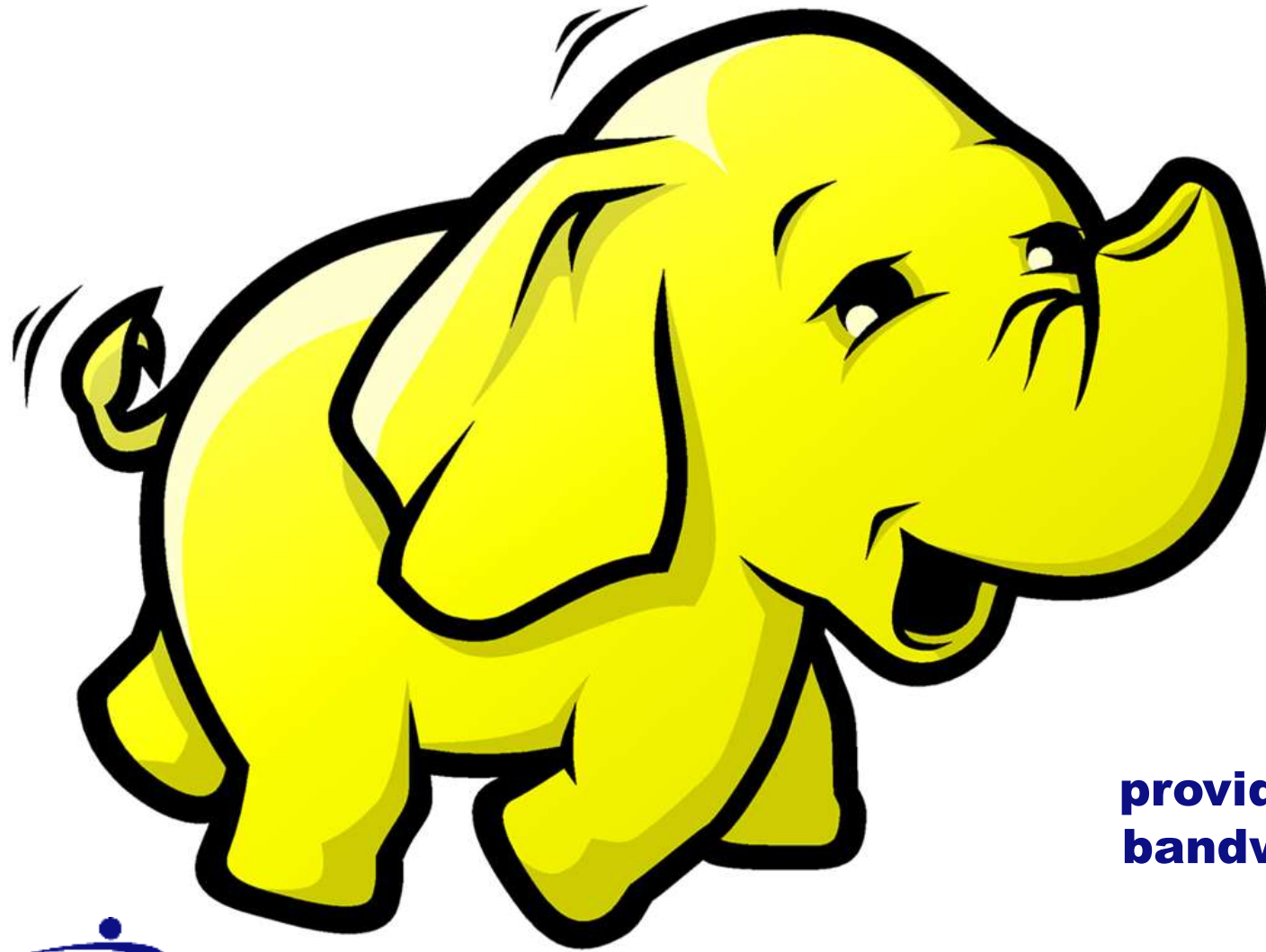
STORAGE FOR ANY TYPE OF DATA
UNIFIED, ELASTIC, RESILIENT, SECURE

Filesystem
(HDFS)

Online NoSQL
(HBASE)

DATA INTEGRATION





Hadoop HDFS (Hadoop Distributed File System)

**is a
distributed file-system
that stores data on
commodity machines,
providing very high aggregate
bandwidth across the cluster.**



Hadoop Cluster

Server 1

Server 2

Server 3

Server 4



Hadoop Cluster

Customer.pdf

Server 1

Server 2

Server 3

Server 4



Hadoop Cluster

Customer.pdf

Server 1

Server 2

Server 3

Server 4



Hadoop Cluster



Server 1

Server 2

Server 3

Server 4



Hadoop Cluster

x3

Block 1

Block 2

Block 3

Block 4

Server 1

Server 2

Server 3

Server 4



Hadoop Cluster

x3

Block 1

Block 1

Block 1

Block 2

Block 2

Block 2

Block 3

Block 3

Block 3

Block 4

Block 4

Block 4

Server 1

Server 2

Server 3

Server 4



Hadoop Cluster

Block 2

Block 2

Block 2

Block 3

Block 3

Block 3

Block 4

Block 4

Block 4

Server 1

Block 1

Server 2

Block 1

Server 3

Block 1

Server 4



Hadoop Cluster

Server 1

Block 1

Block 2

Server 2

Block 1

Block 2

Server 3

Block 1

Server 4

Block 2

Block 3

Block 3

Block 3

Block 4

Block 4

Block 4



Hadoop Cluster

Server 1

Block 1

Block 2

Server 2

Block 1

Block 2

Block 3

Server 3

Block 1

Block 3

Server 4

Block 2

Block 3

Block 4

Block 4

Block 4



Hadoop Cluster

Server 1

Block 1

Block 2

Block 4

Server 2

Block 1

Block 2

Block 3

Server 3

Block 1

Block 3

Block 4

Server 4

Block 2

Block 3

Block 4



Hadoop Cluster

Server 1

Block 1

Block 2

Block 4

Server 2

Block 1

Block 2

Block 3

Server 3

Block 1

Block 3

Block 4

Server 4

Block 2

Block 3

Block 4



Hadoop Cluster

Server 1

Block 1

Block 2

Block 4

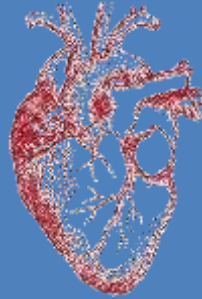
Server 2

Block 1

Block 2

Block 3

Master Node



Server 3

Block 1

Block 3

Block 4

Server 4

Block 2

Block 3

Block 4



Hadoop Cluster

Server 1

Block 1

Block 2

Block 4

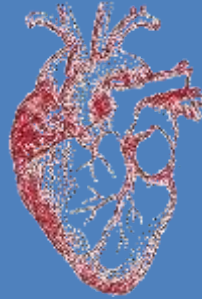
Server 2

Block 1

Block 2

Block 3

Master Node



Server 3

Block 1

Block 3

Block 4

Server 4

Block 2

Block 3

Block 4



Hadoop Cluster

Server 1

Block 1

Block 2

Block 4

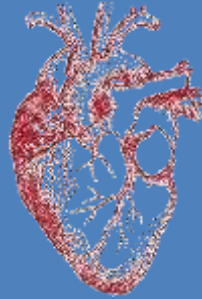
Server 2

Block 1

Block 2

Block 3

Master Node



Server 3

Block 1

Block 3

Block 4

Server 4

Block 2

Block 3

Block 4



Hadoop Cluster

Server 1

Block 1

Block 4

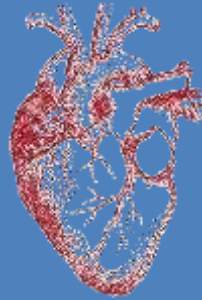
Server 2

Block 1

Block 2

Block 3

Master Node



Server 3

Block 1

Block 3

Block 4

Server 4

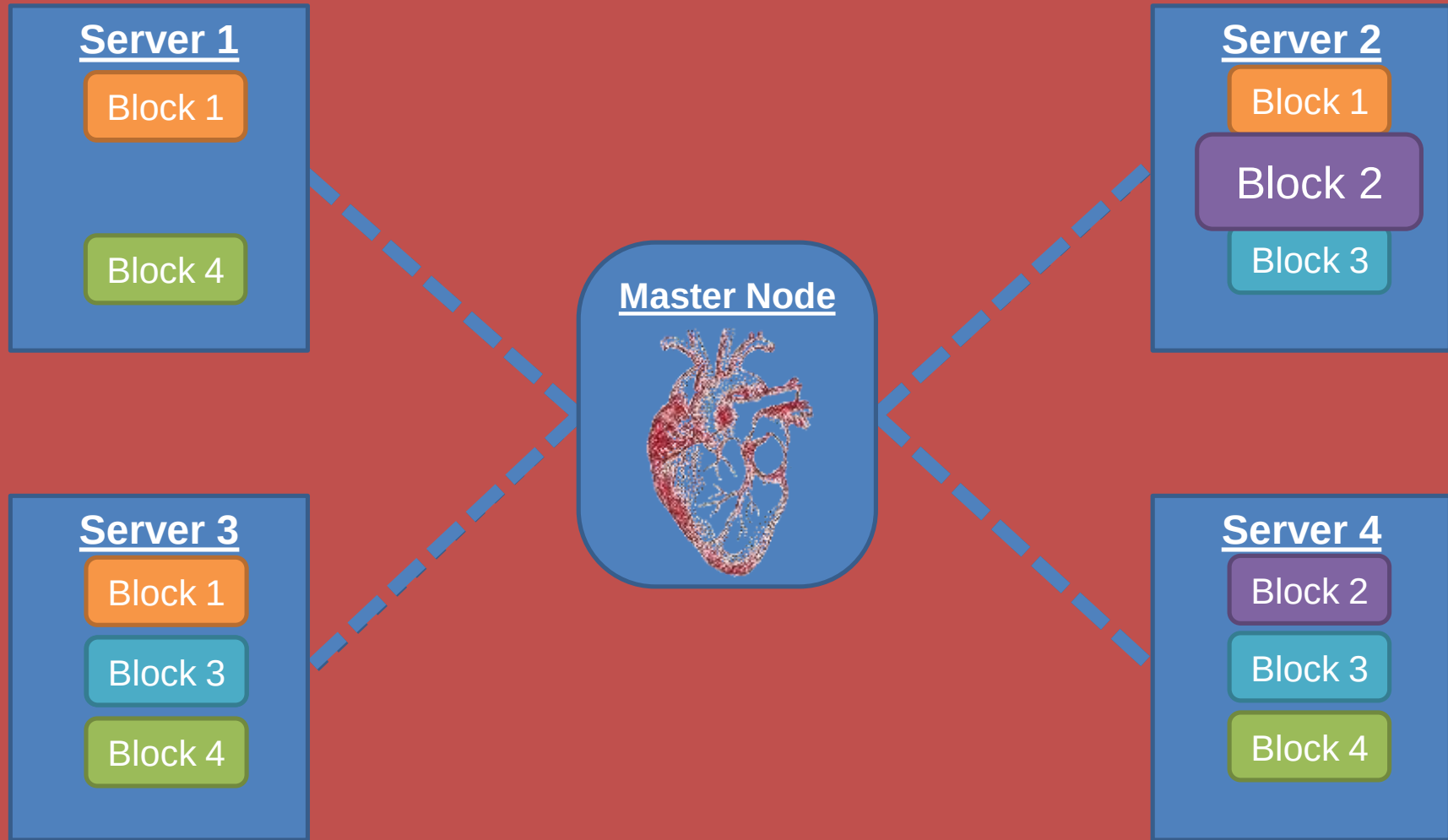
Block 2

Block 3

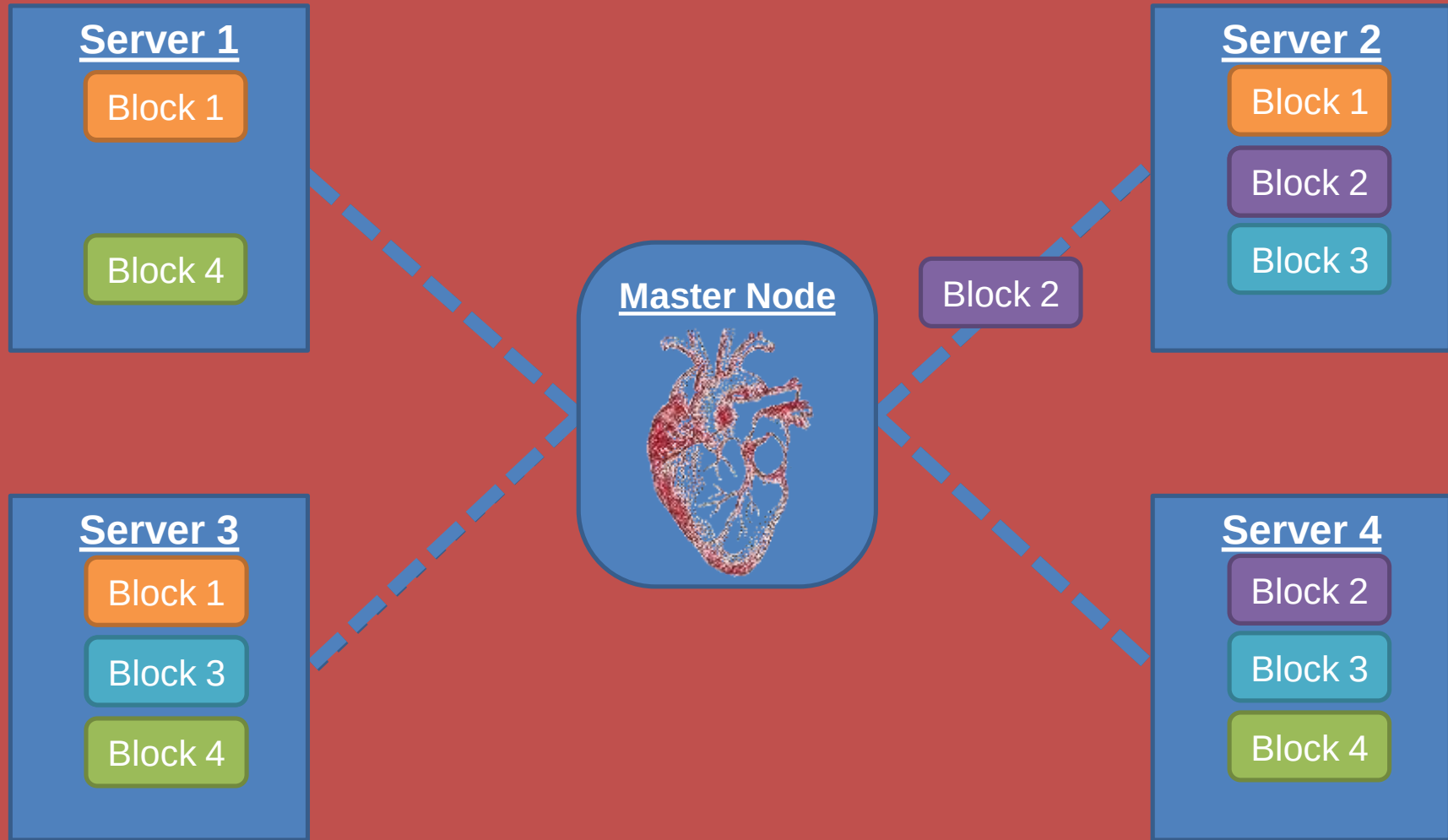
Block 4



Hadoop Cluster



Hadoop Cluster



Hadoop Cluster

Server 1

Block 1

Block 4

Server 2

Block 1

Block 2

Block 3

Master Node

Block 2

Server 3

Block 1

Block 3

Block 4

Server 4

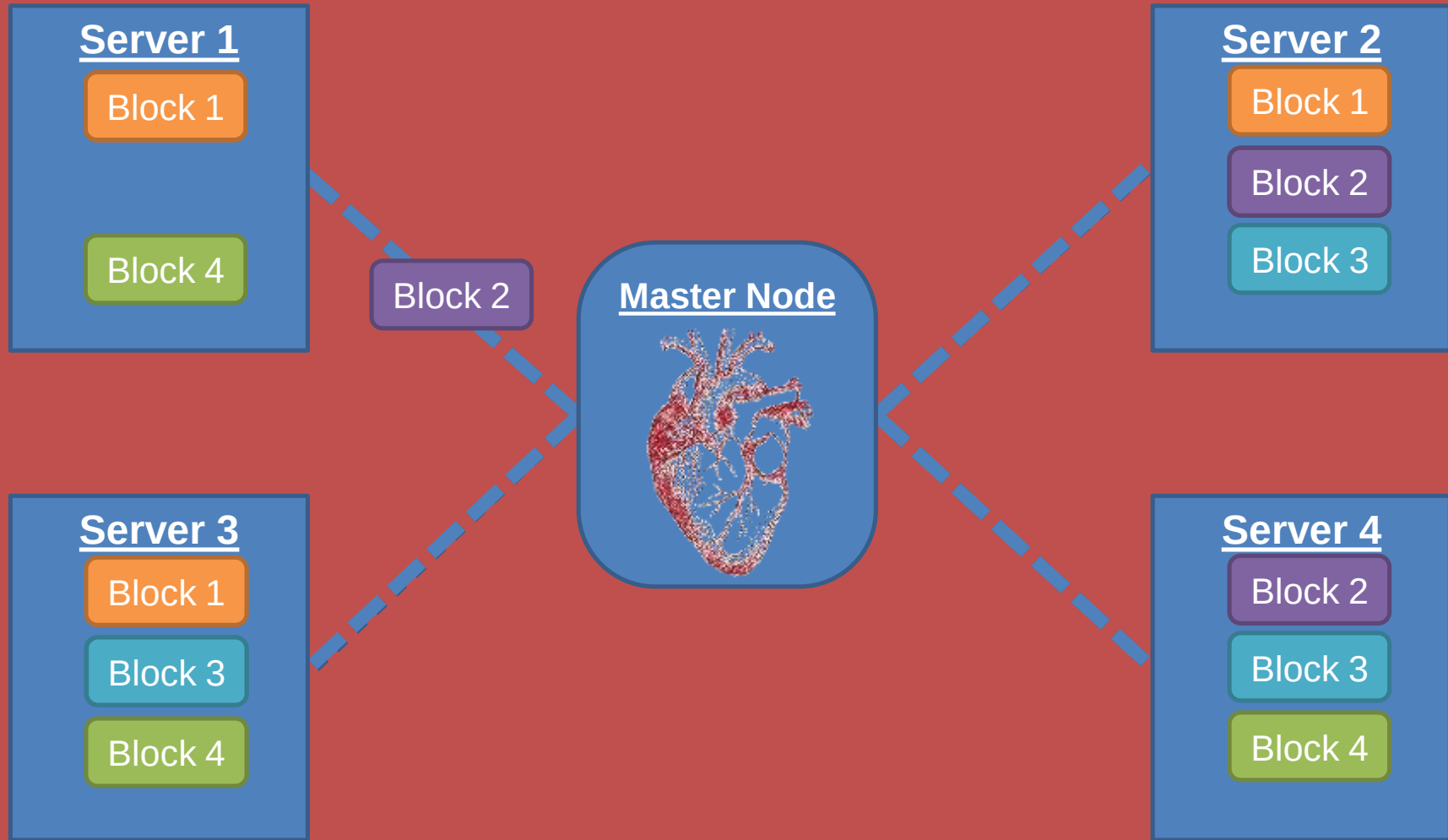
Block 2

Block 3

Block 4



Hadoop Cluster



Hadoop Cluster

Server 1

Block 1

Block 2

Block 4

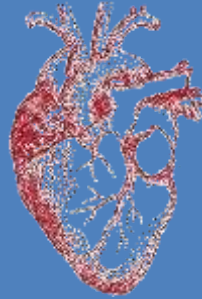
Server 2

Block 1

Block 2

Block 3

Master Node



Server 3

Block 1

Block 3

Block 4

Server 4

Block 2

Block 3

Block 4



Hadoop Cluster

Server 1

Block 1

Block 2

Block 4

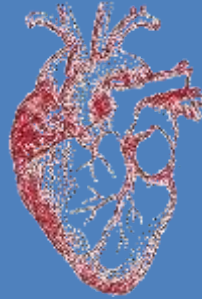
Server 2

Block 1

Block 2

Block 3

Master Node



Server 3

Block 1

Block 3

Block 4

Server 4

Block 2

Block 3

Block 4



Hadoop Cluster

Server 1

Block 1

Block 2

Block 4

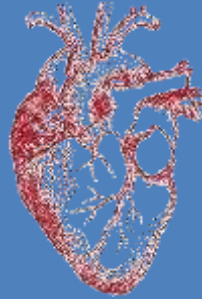
Server 2

Block 1

Block 2

Block 3

Master Node



Server 3

Block 1

Block 3

Block 4

Server 4

Block 2

Block 3

Block 4



Hadoop Cluster

Server 1

Block 1

Block 2

Block 4

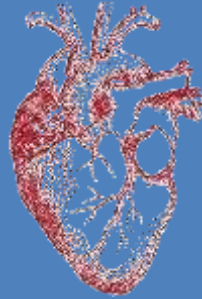
Server 2

Block 1

Block 2

Block 3

Master Node

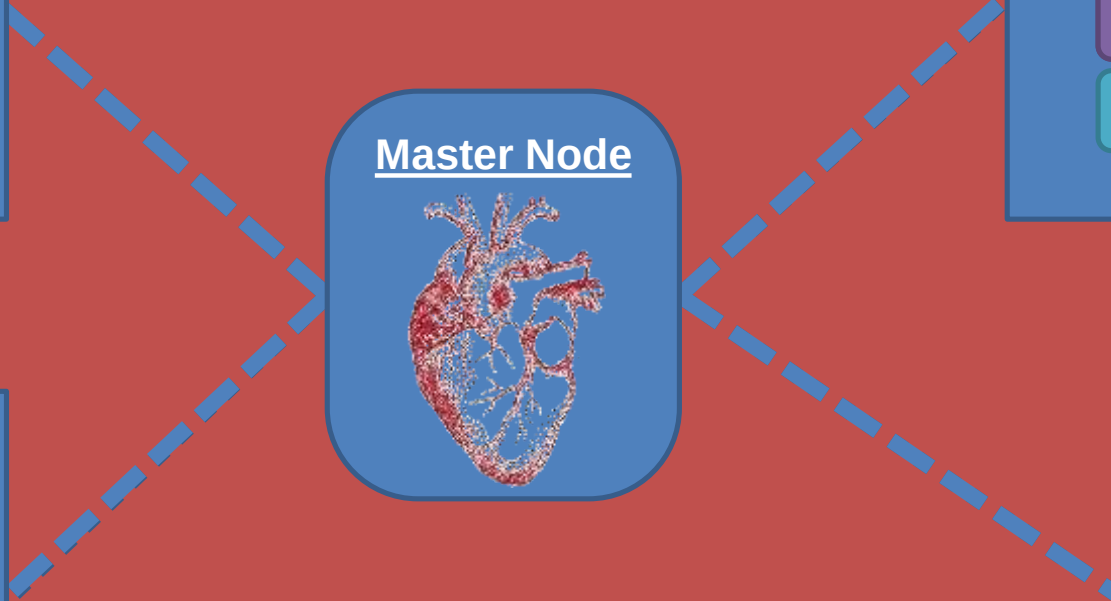


Server 3

Block 1

Block 3

Block 4



Hadoop Cluster

Server 1

Block 1

Block 2

Block 4

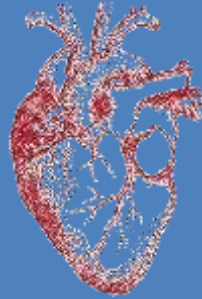
Server 2

Block 1

Block 2

Block 3

Master Node

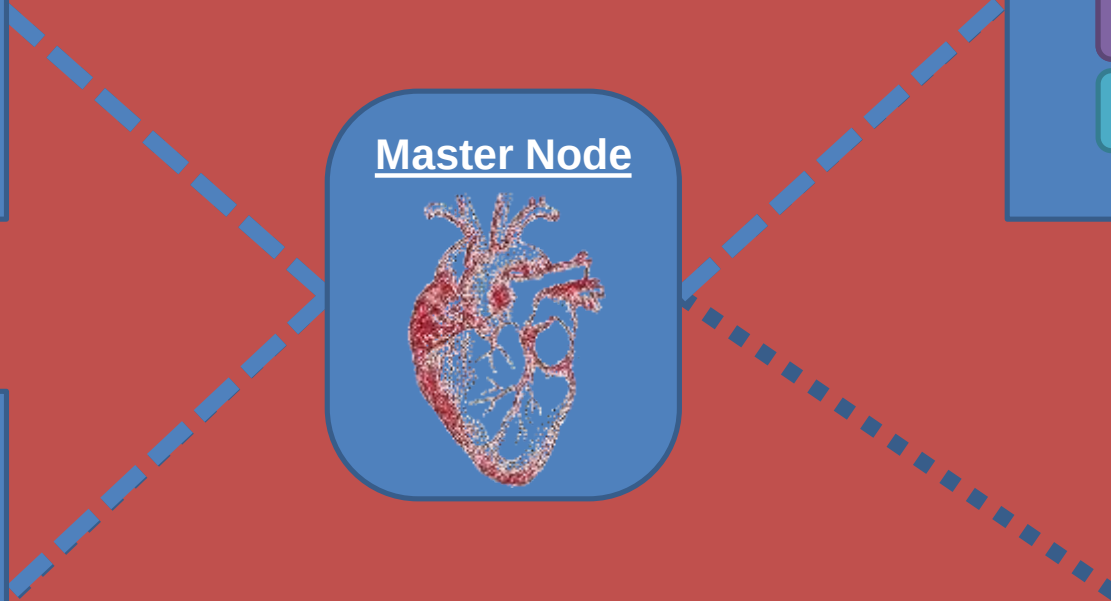


Server 3

Block 1

Block 3

Block 4



Hadoop Cluster

Server 1

Block 1

Block 2

Block 4

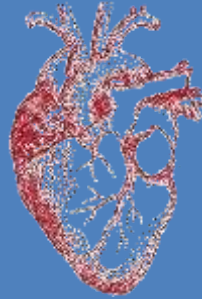
Server 2

Block 1

Block 2

Block 3

Master Node



Server 3

Block 1

Block 3

Block 4



Hadoop Cluster

Server 1

Block 1

Block 2

Block 4

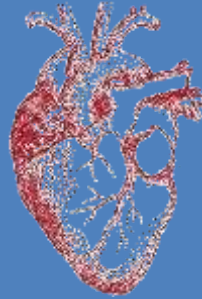
Server 2

Block 1

Block 2

Block 3

Master Node



Server 3

Block 1

Block 3

Block 4

Server 4



Hadoop Cluster

Server 1

Block 1

Block 2

Block 4

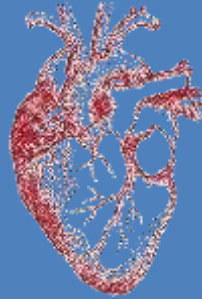
Server 2

Block 1

Block 2

Block 3

Master Node



Server 3

Block 1

Block 3

Block 4

Server 4



Hadoop Cluster

Server 1

Block 1

Block 2

Block 4

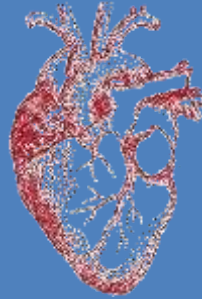
Server 2

Block 1

Block 2

Block 3

Master Node



Server 3

Block 1

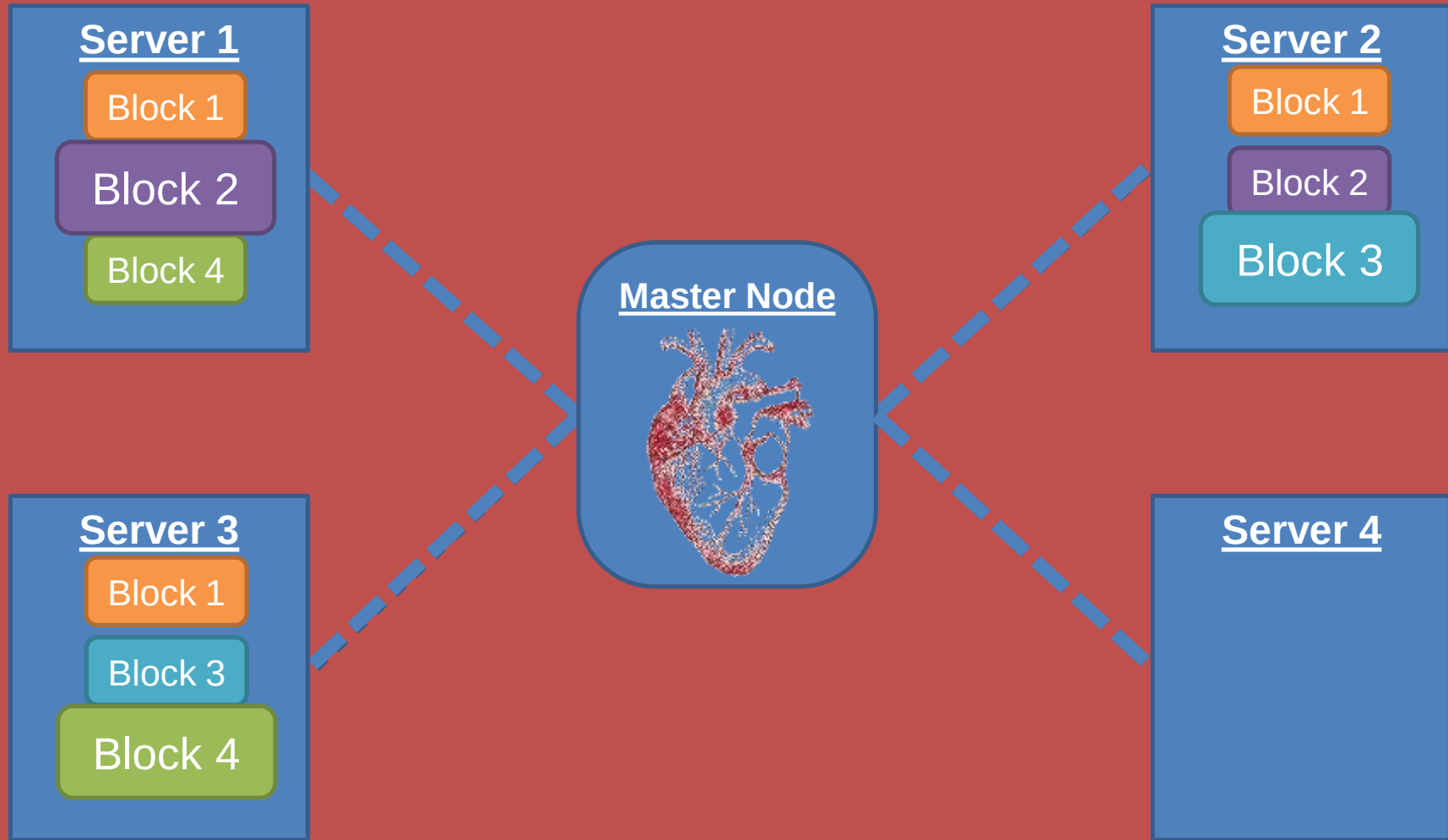
Block 3

Block 4

Server 4



Hadoop Cluster



Hadoop Cluster

Server 1

Block 1

Block 2

Block 4

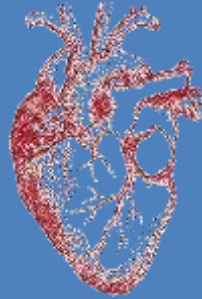
Server 2

Block 1

Block 2

Block 3

Master Node



Server 3

Block 1

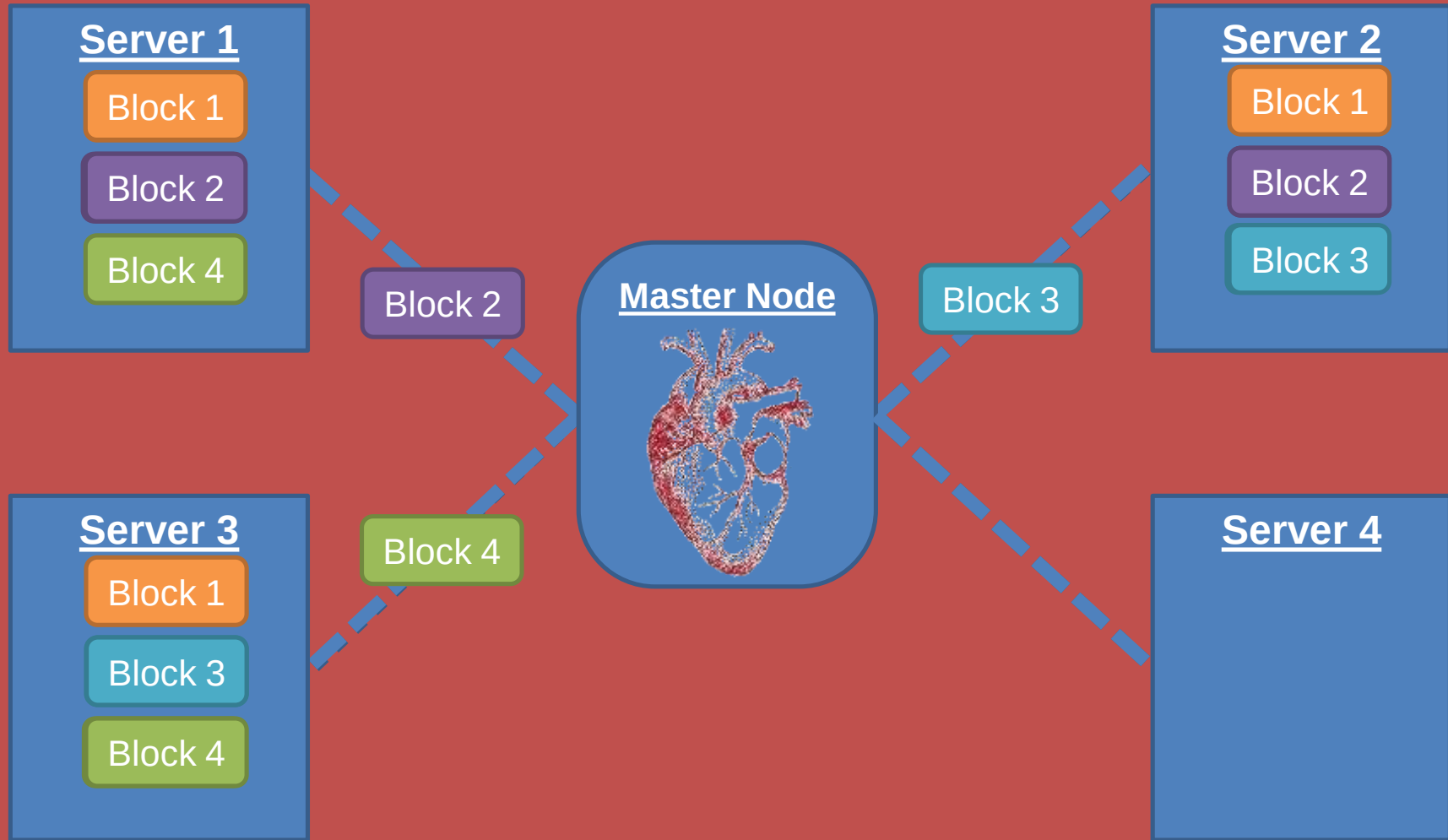
Block 3

Block 4

Server 4



Hadoop Cluster



Hadoop Cluster

Server 1

Block 1

Block 2

Block 4

Server 2

Block 1

Block 2

Block 3

Master Node

Block 2

Block 3

Block 4

Server 3

Block 1

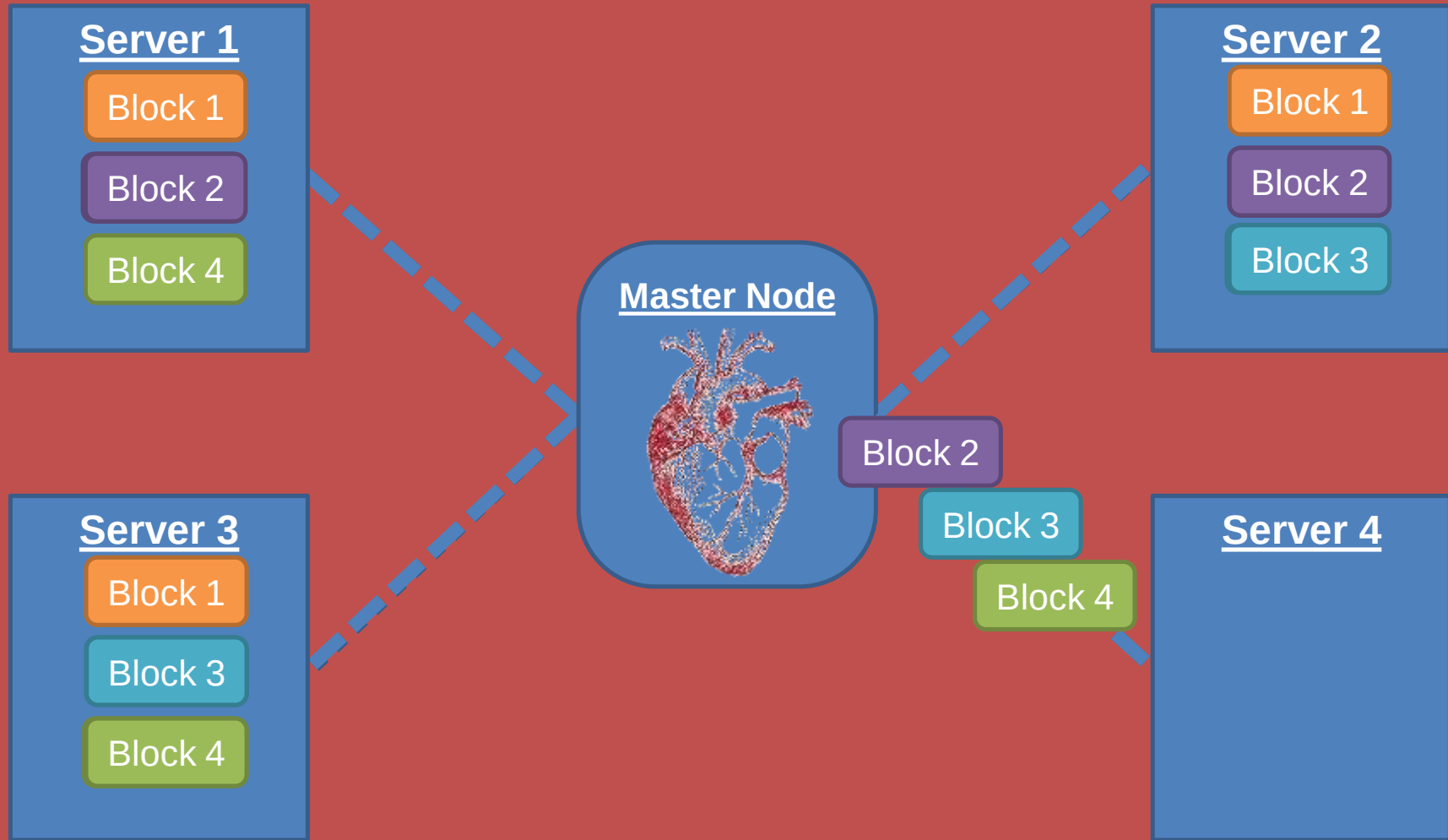
Block 3

Block 4

Server 4



Hadoop Cluster



Hadoop Cluster

Server 1

Block 1

Block 2

Block 4

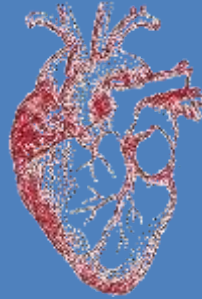
Server 2

Block 1

Block 2

Block 3

Master Node



Server 3

Block 1

Block 3

Block 4

Server 4

Block 2

Block 3

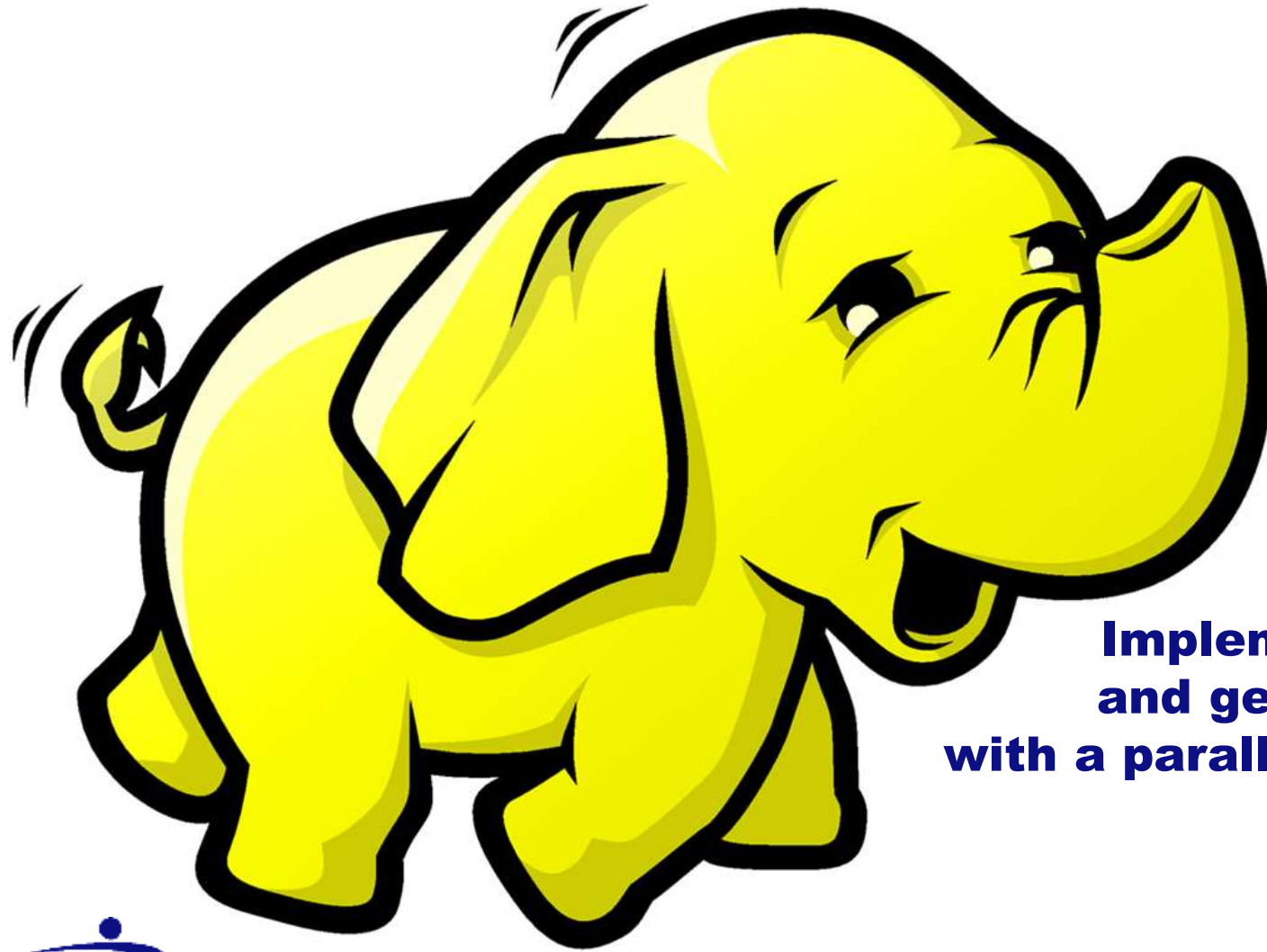
Block 4



MapReduce

“Map” step
“Shuffle” step
“Reduce” step

is a programming model
and an associated
Implementation for processing
and generating large data sets
with a parallel, distributed algorithm
on a cluster.





Spark



- Originally developed in 2009 in UC Berkeley's AMPLab
- Fully open sourced in 2010 – now a Top Level Project at the Apache Software Foundation
- Fast Growing Community

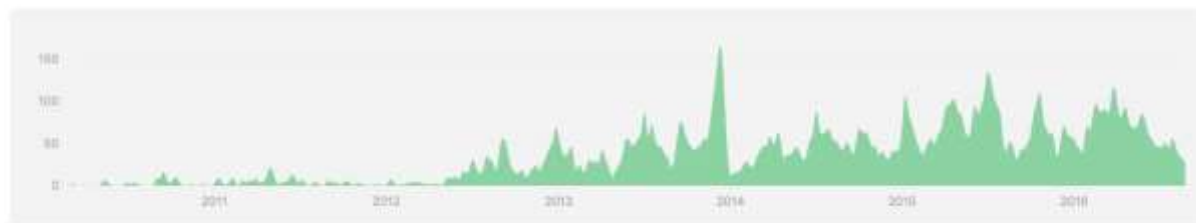
The screenshot shows the Apache Spark website with the following elements:

- Header:** Apache Spark logo with the tagline "Lightning fast cluster computing". Navigation links include Download, Libraries, Documentation, Examples, Community, and FAQ. The Apache Software Foundation logo is on the right.
- Main Text:** "Apache Spark™ is a fast and general engine for large-scale data processing."
- Speed Section:**
 - Text: "Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk." and "Apache Spark has an advanced DAG execution engine that supports cyclic data flow and in-memory computing."
 - Bar chart titled "Logistic regression in Hadoop and Spark" comparing running times. Hadoop is 110 seconds, and Spark is 0.9 seconds.
- Ease of Use Section:**
 - Text: "Write applications quickly in Java, Scala, Python, R." and "Spark offers over 10 high-level constructs that make it easy to build parallel programs. All you can Java interactively from the Scala, Python and R shells."
- Code Snippets:**

```
text_file = spark.textFile("hdfs://...")
text_file.flatMap(lambda line: line.split())
.map(lambda word: (word, 1))
.reduceByKey(lambda a, b: a+b)
```

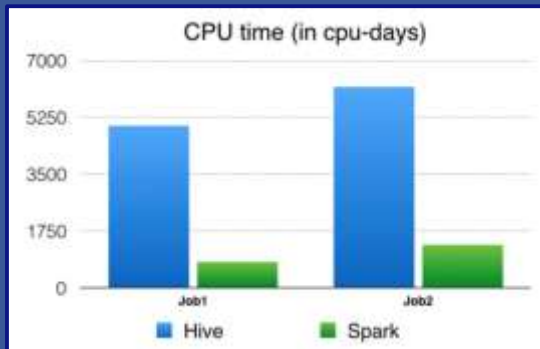
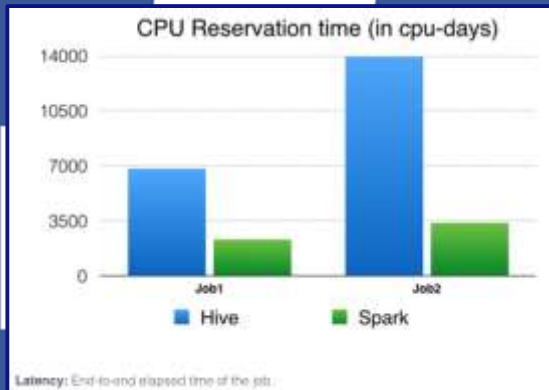
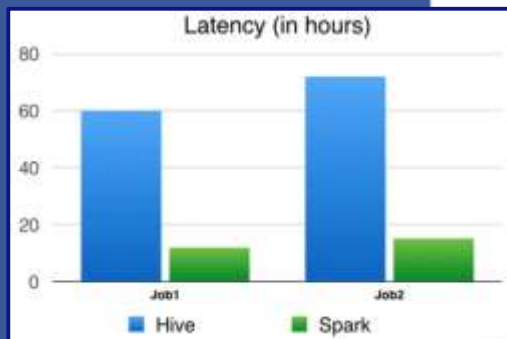
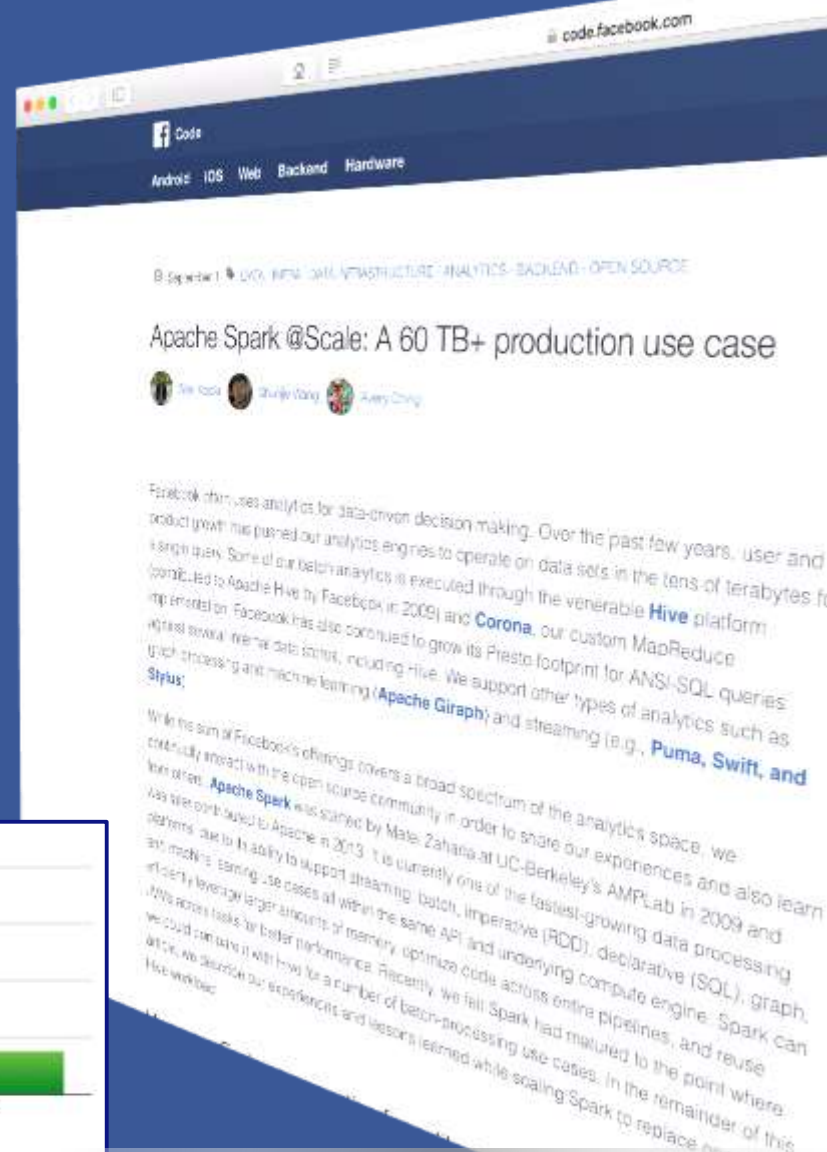
Word count in Spark's Python API
- Latest News:**
 - Spark 2.0.0 released (Jul 26, 2016)
 - Spark 1.6.2 released (Jun 25, 2016)
 - Call for Presentations for Spark Summit EU is Open (Jun 16, 2016)
 - Preview release of Spark 2.0 (May 2016)
- Download Spark:** A large green button.
- Built-in Libraries:**
 - SQL and DataFrames
 - Spark Streaming
 - MLlib (machine learning)
 - GraphX (graph)
 - Third-Party Packages
- Footer:** Logos for Spark, MLlib, and GraphX.

Mar 28, 2010 – Sep 21, 2016
Contributions to master, excluding merge commits



Source: spark.apache.org
github.com/apache/spark





Benefits?

Runs on
commodities
servers

Ecosystem
and tools

Large Scale
Processing
and
Analytics

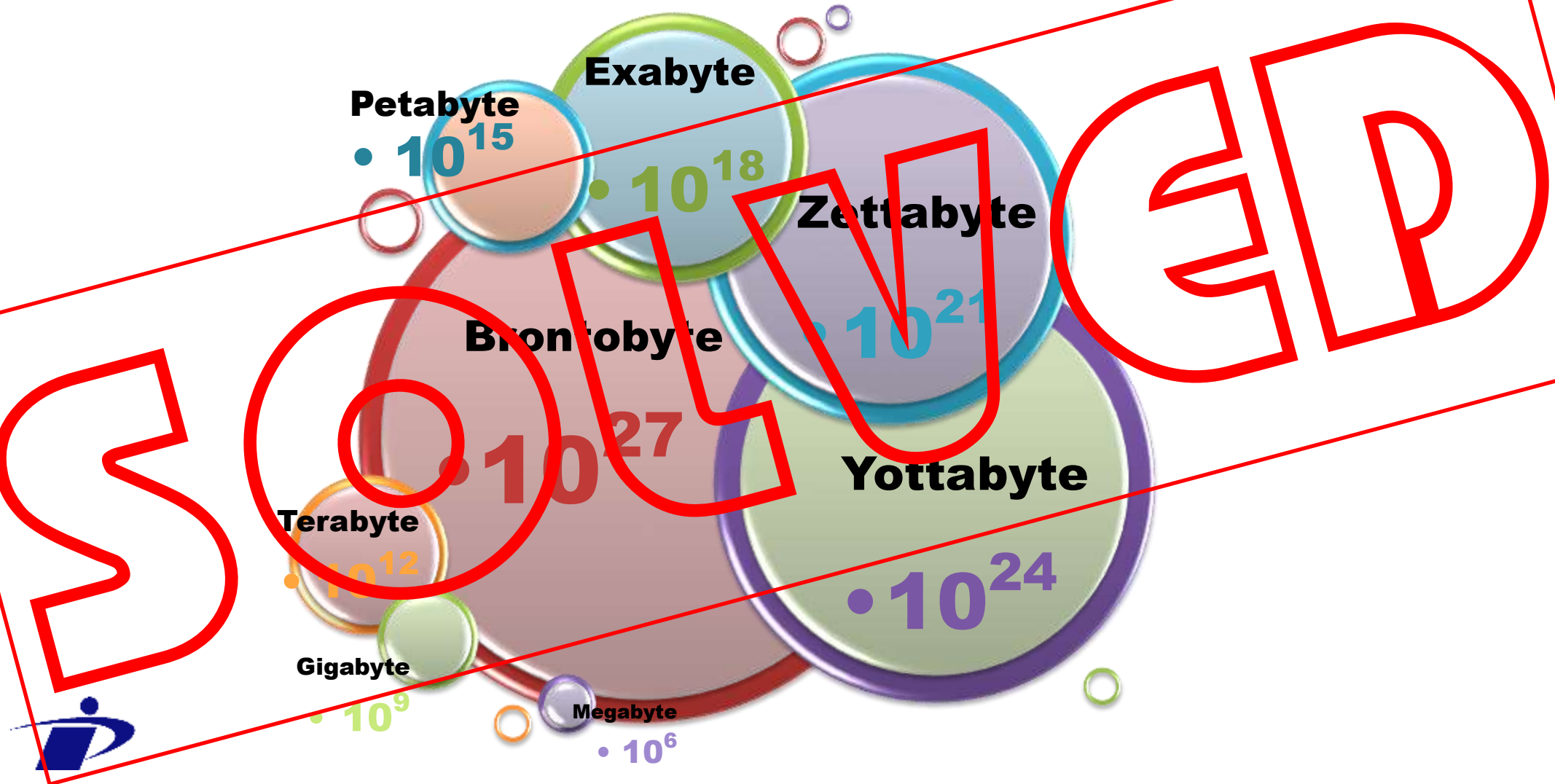
Real-
time/Batch
processing

Parallel
processing

Storage &
Processing
Platform



So much DATA!!! How to handle?



Can we (human) still handle them?

SOLVED





SILLOS



of DATA

We need to

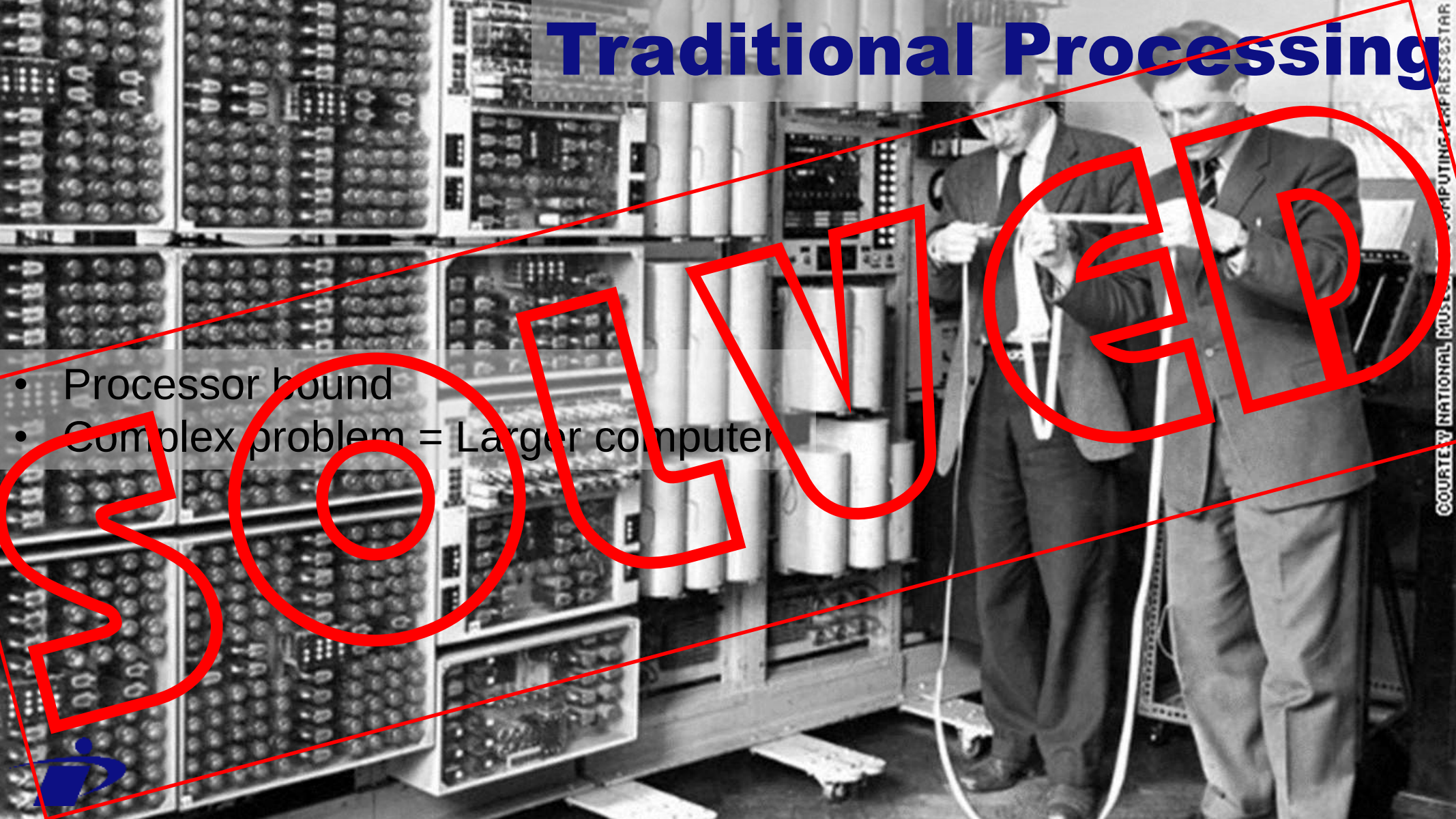
simplify
'simplify'
verb

SOLVED



Traditional Processing

- Processor bound
- Complex problem = Larger computer

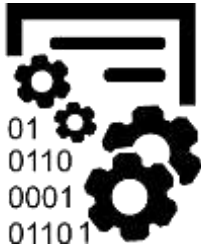


Source: Intro to DataRobot
[www.youtube.comhttps://m.youtube.com/watch?v=L3V_UJugtJ0](https://www.youtube.com/watch?v=L3V_UJugtJ0)

The Future of Machine Learning



faster



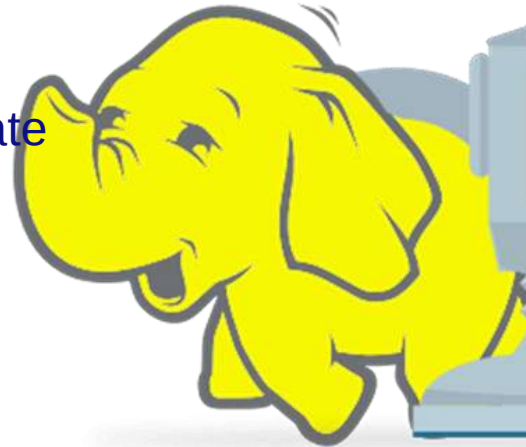
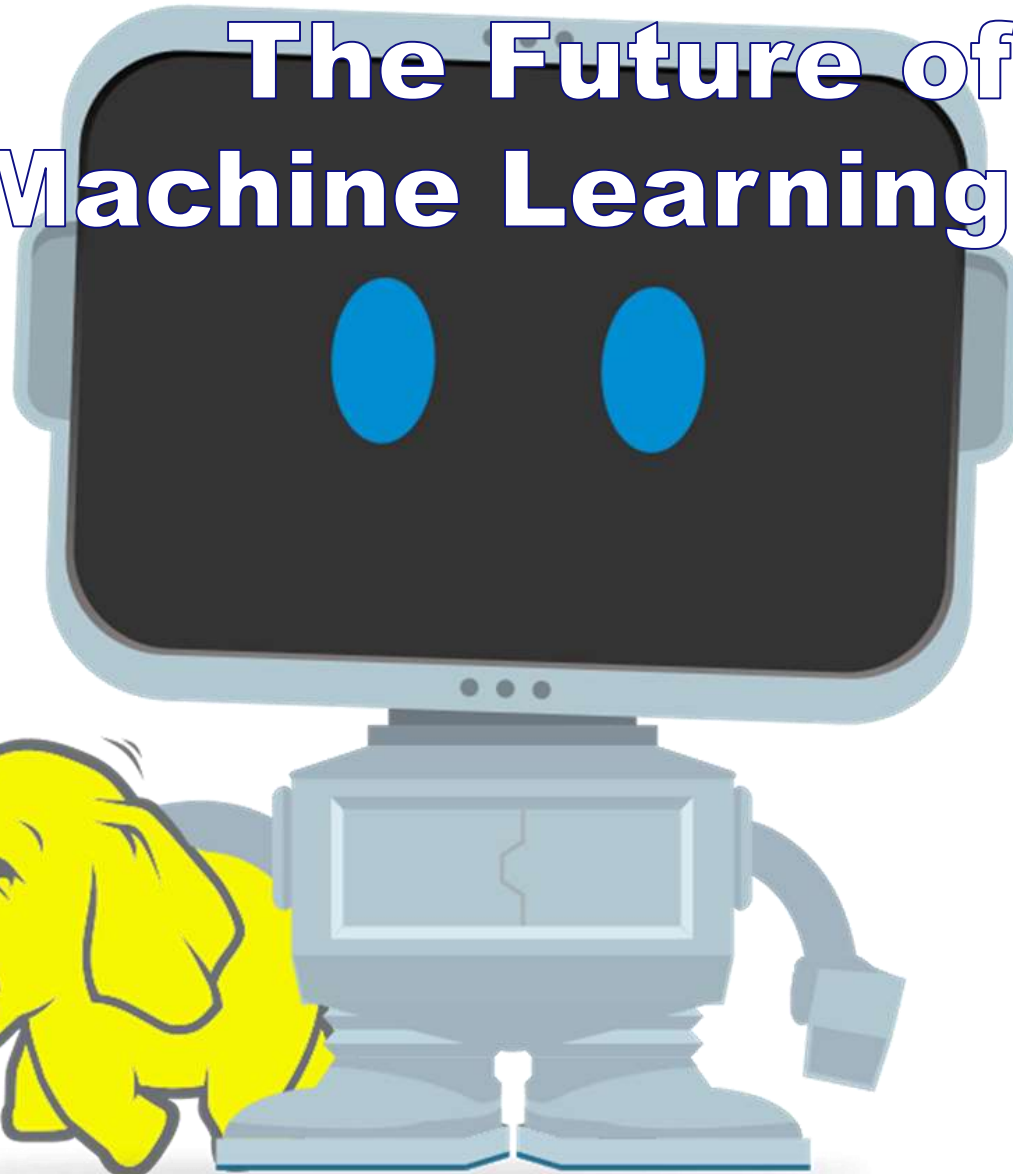
> data



> accurate



< cost



Hadoop+ DataRobot = ?



faster



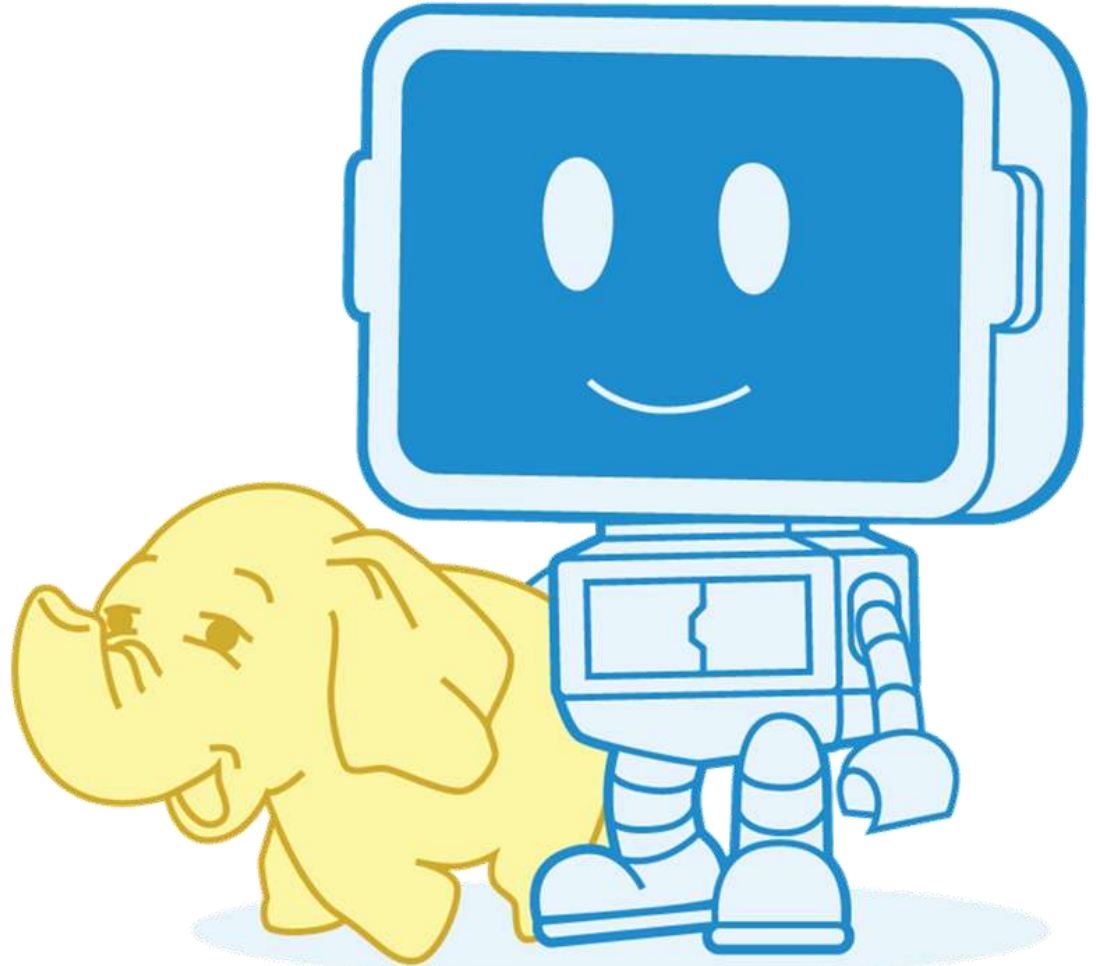
> data



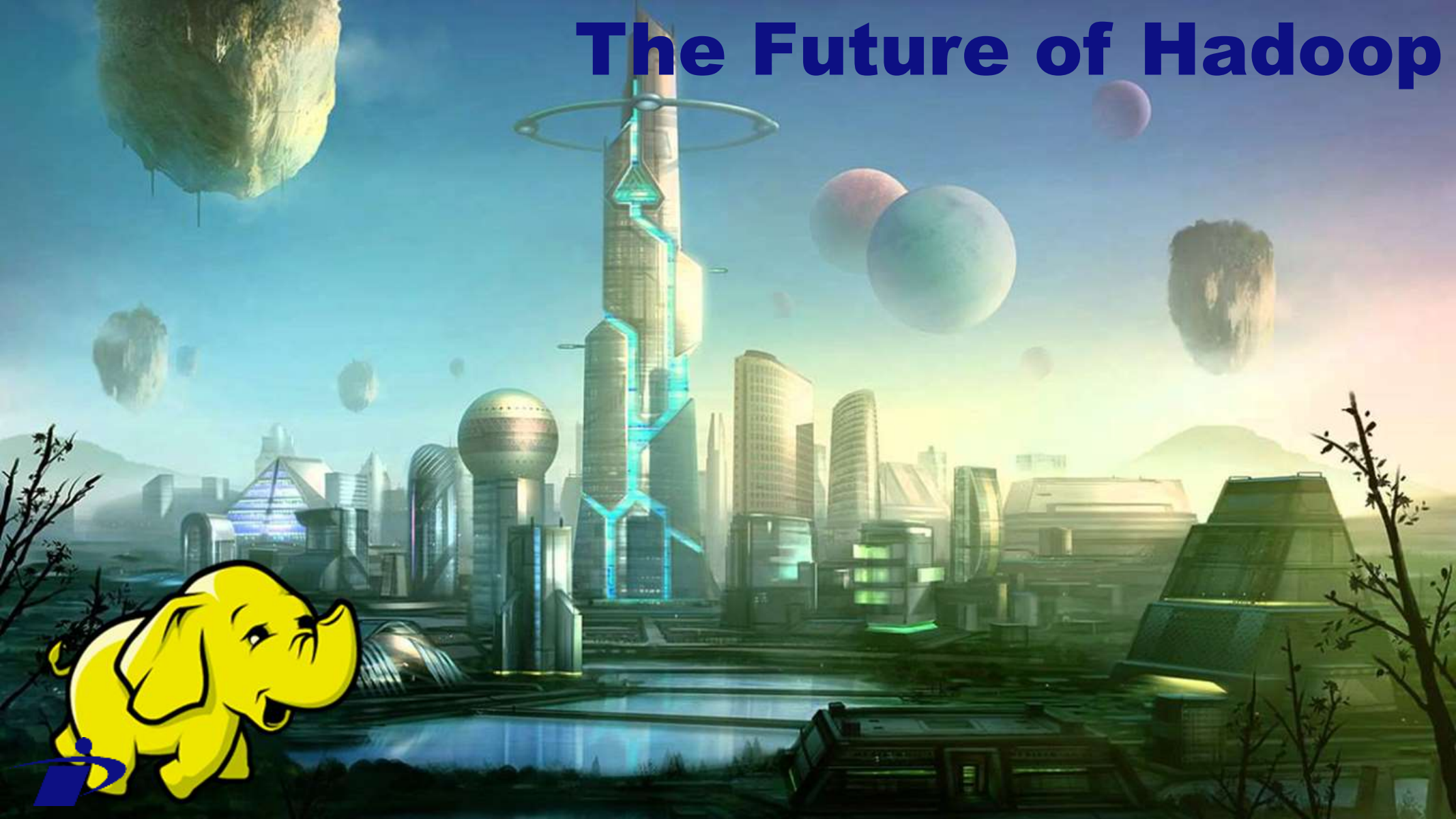
> accurate



< cost



The Future of Hadoop



ec

eRic.Choo@ptcsys.com

JK

Jk.wong@ptcsys.com

