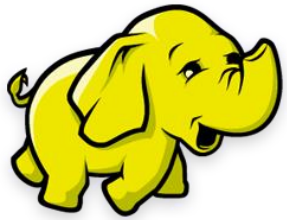


Overview on Big Data and Hadoop

John Jernigan

10/30/2020

And also 

, but just a little bit

- What is Big Data?
- What is Hadoop, and what is it not?
- ~~Let's talk about Spark and H_2O~~

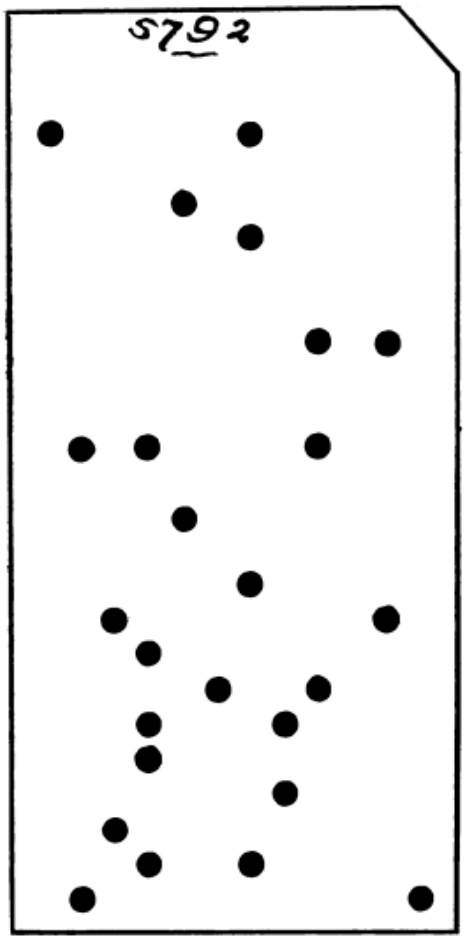
Classic “Big Data” Definition

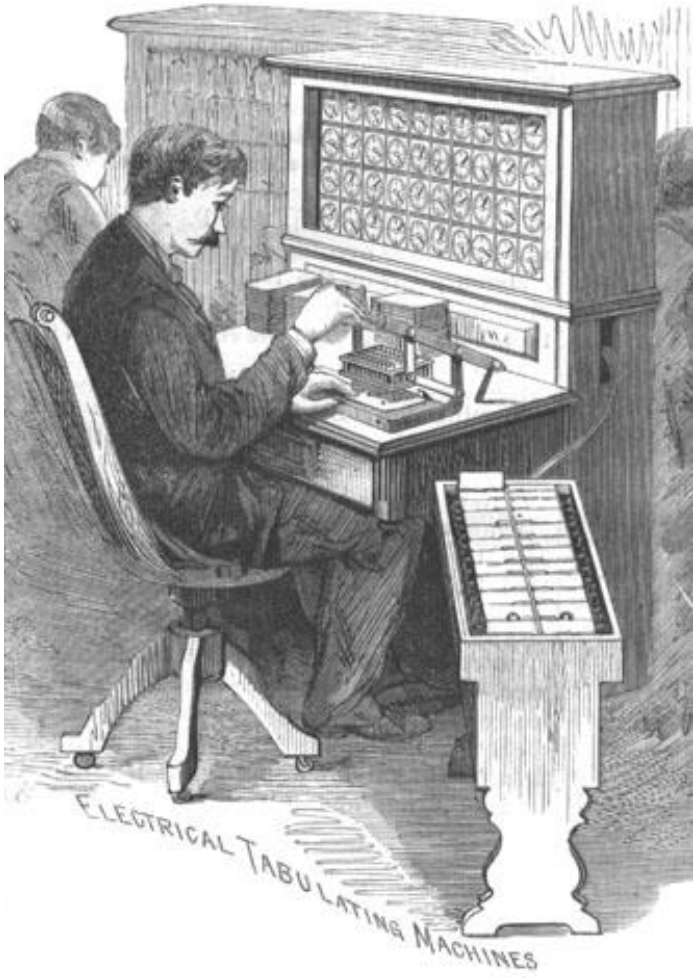
http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

Off on a tangent...

Punch Cards:

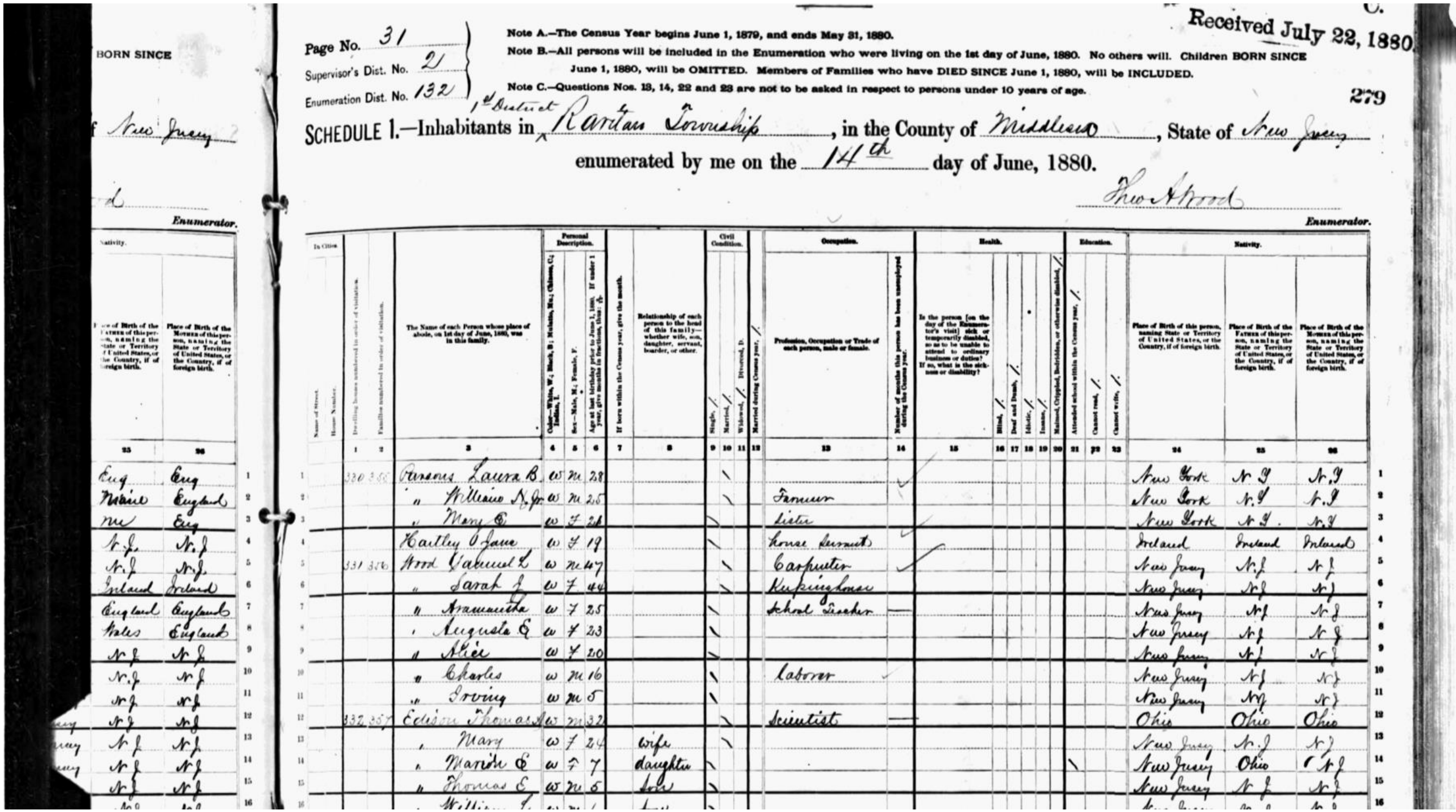
Hadoop of the 1900's?





United States: 1880

1880 Census:
More Data Than Ever



Page No. 31
Supervisor's Dist. No. 2
Enumeration Dist. No. 132

Note A.—The Census Year begins June 1, 1879, and ends May 31, 1880.
Note B.—All persons will be included in the Enumeration who were living on the 1st day of June, 1880. No others will. Children BORN SINCE June 1, 1880, will be OMITTED. Members of Families who have DIED SINCE June 1, 1880, will be INCLUDED.
Note C.—Questions Nos. 13, 14, 22 and 23 are not to be asked in respect to persons under 10 years of age.

Received July 22, 1880

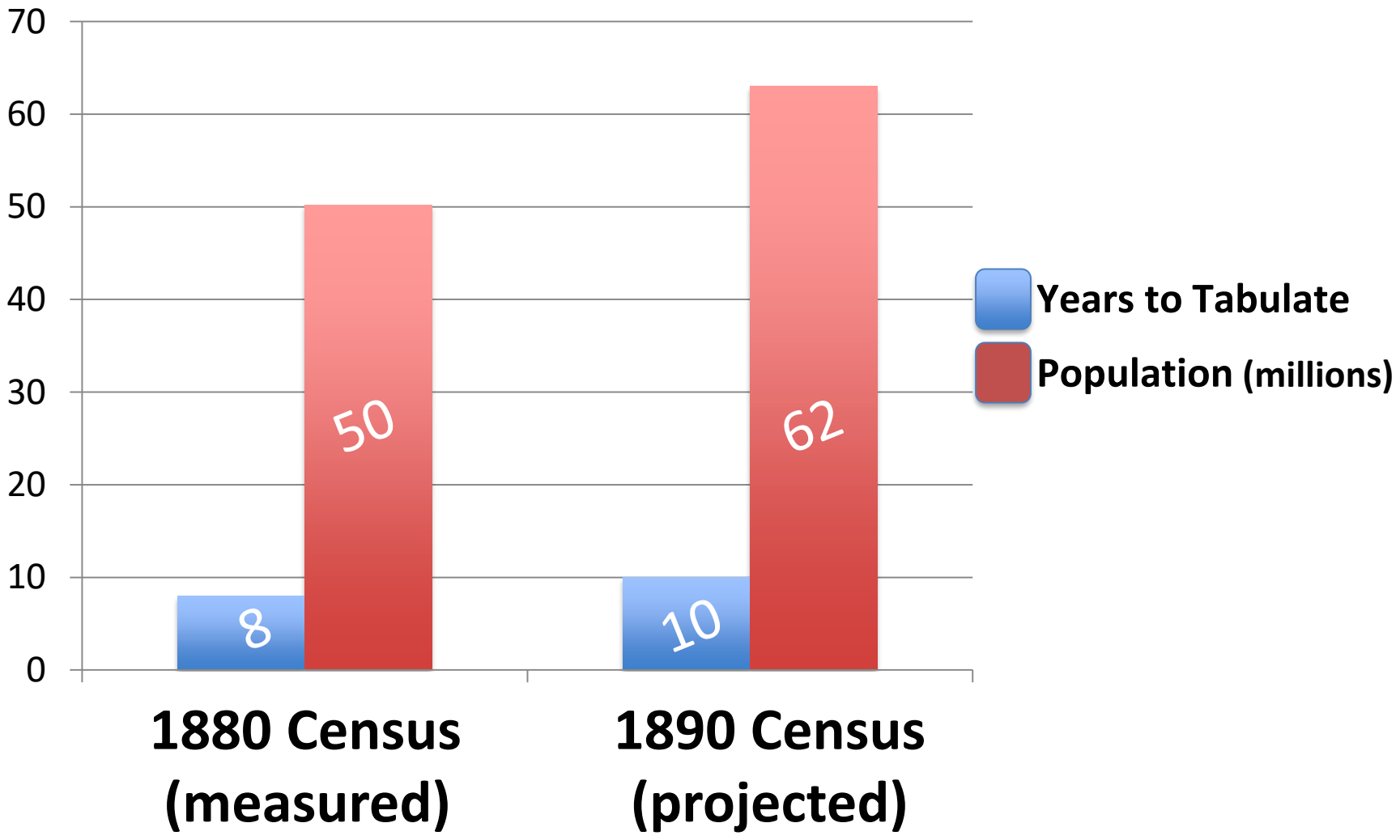
279

SCHEDULE 1.—Inhabitants in 1st District Raritan Township, in the County of Middlesex, State of New Jersey, enumerated by me on the 14th day of June, 1880.

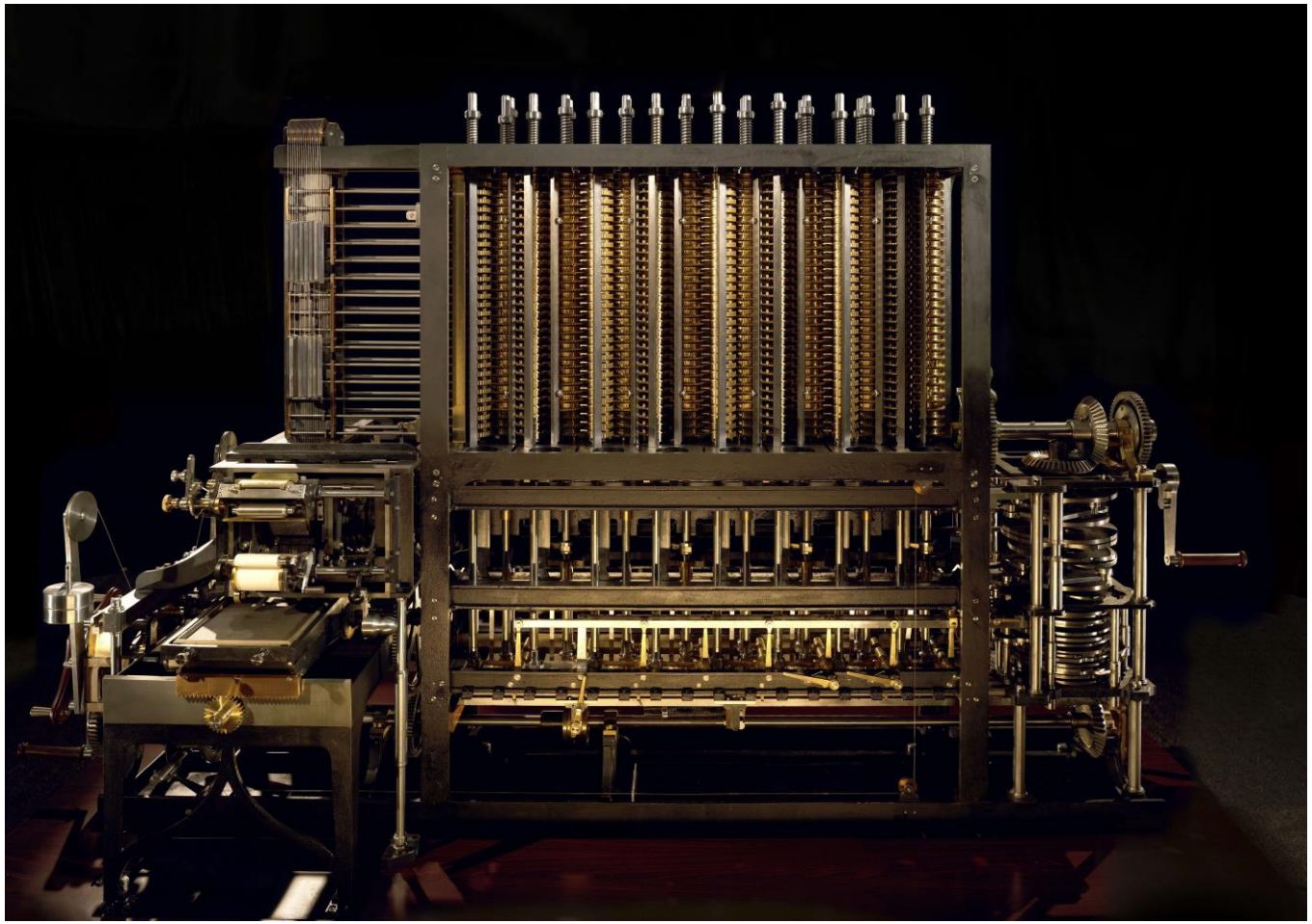
Theo Wood, Enumerator.

Line	Dwelling	Family	Name	Color	Sex	Age	Relationship	Occupation	Place of Birth	Father's Birthplace	Mother's Birthplace
1	330	355	Parsons Laura B.	W	M	28			New York	N.Y.	N.Y.
2			" William N. Jr.	W	M	25		Farmer	New York	N.Y.	N.Y.
3			" Mary E.	W	F	21		Sister	New York	N.Y.	N.Y.
4			Hartley Jane	W	F	19		House Servant	Ireland	Ireland	Ireland
5	331	356	Wood Samuel L.	W	M	47		Carpenter	New Jersey	N.J.	N.J.
6			" Sarah J.	W	F	44		Keeping house	New Jersey	N.J.	N.J.
7			" Araminta	W	F	25		School Teacher	New Jersey	N.J.	N.J.
8			" Augusta E.	W	F	23			New Jersey	N.J.	N.J.
9			" Alice	W	F	20			New Jersey	N.J.	N.J.
10			" Charles	W	M	16		Laborer	New Jersey	N.J.	N.J.
11			" Irving	W	M	5			New Jersey	N.J.	N.J.
12	332	357	Edison Thomas A.	W	M	32		Scientist	Ohio	Ohio	Ohio
13			" Mary	W	F	24	wife		New Jersey	N.J.	N.J.
14			" Marion E.	W	F	7	daughter		New Jersey	Ohio	N.J.
15			" Thomas E.	W	M	5	son		New Jersey	N.J.	N.J.
16			" William L.	W	M	[illegible]	son		New Jersey	[illegible]	N.J.

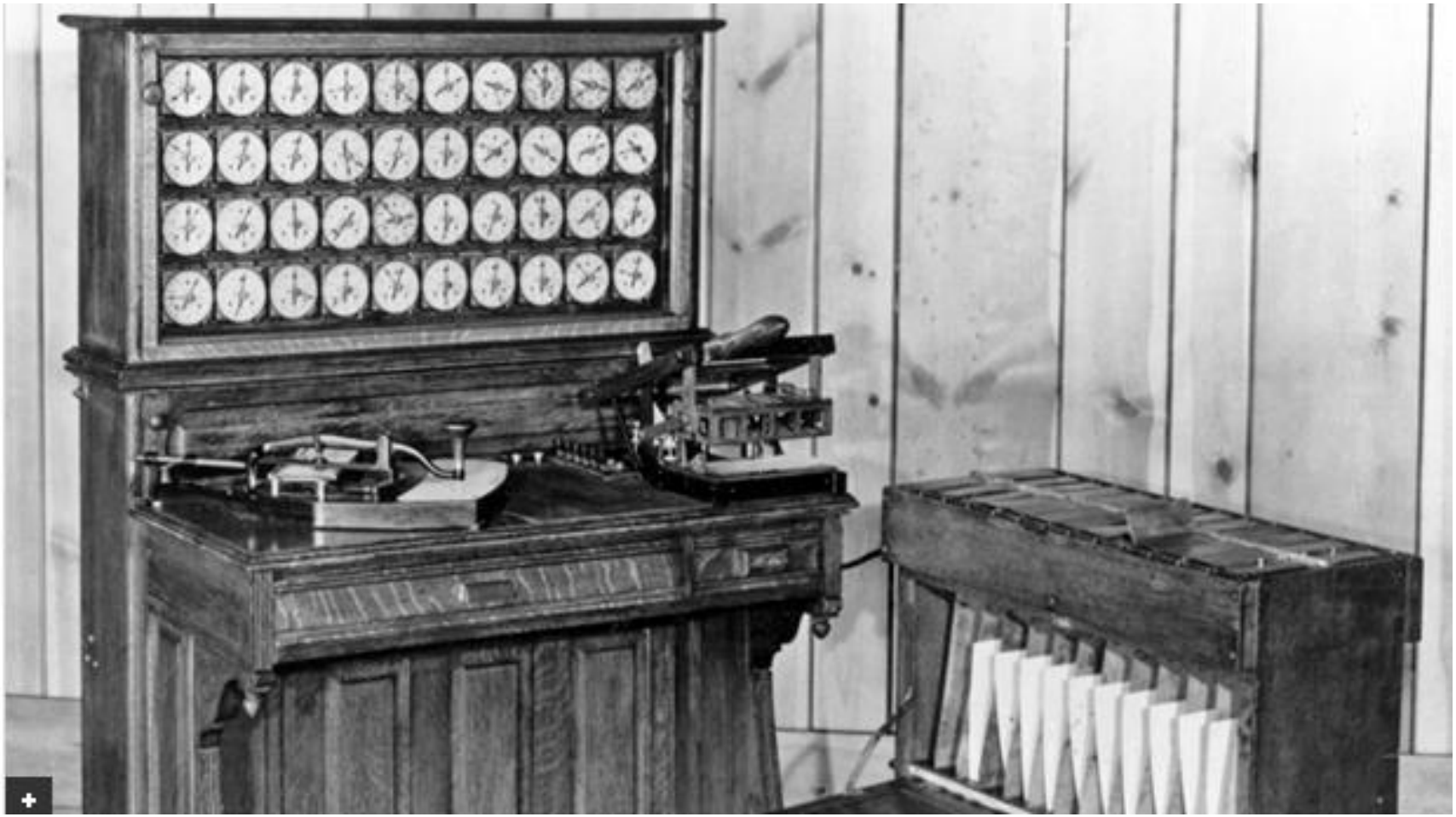
Taking Too Long to Process Data



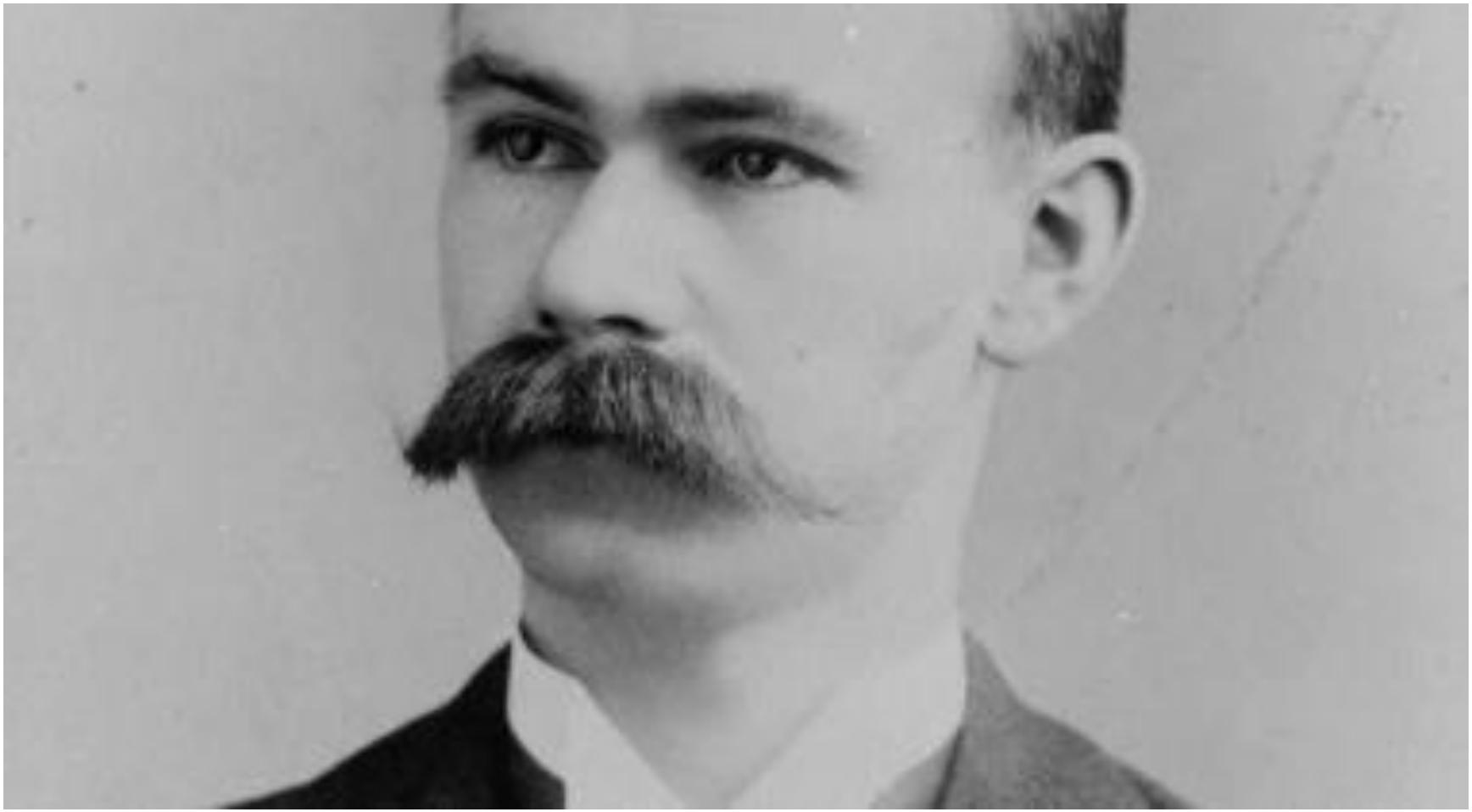
“Second Industrial Revolution” Mentality: Machines Will Solve All Problems



Herman Hollerith's Tabulating Machine

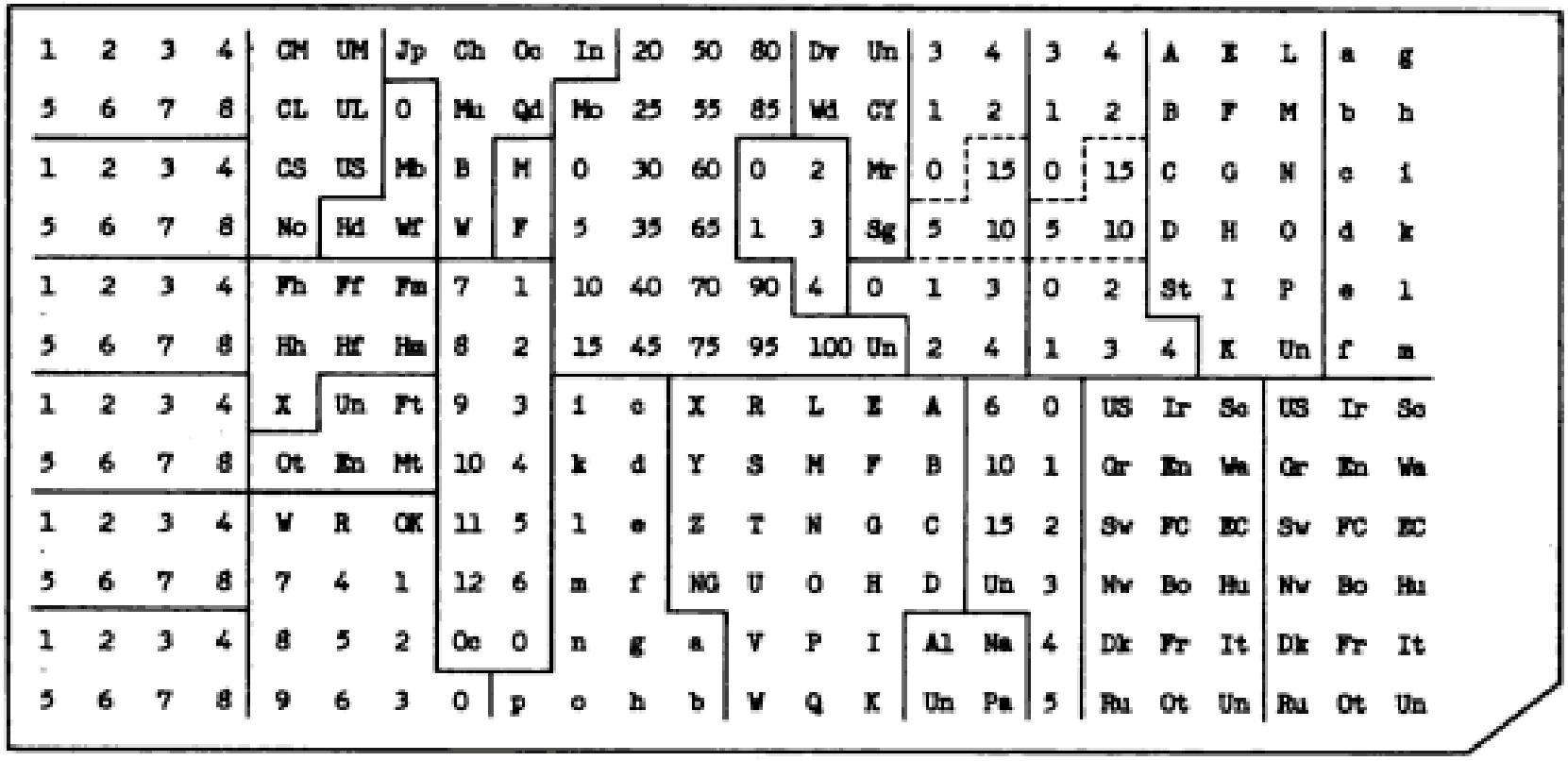


Herman Hollerith



Punch Card Transcription

Punch Card for 1890 Census



1	2	3	4	CM	UM	Jp	Ch	Oc	In	20	50	80	Dv	Un	3	4	3	4	A	E	L	a	g
5	6	7	8	CL	UL	0	Mu	Qd	Mo	25	55	85	Wd	CY	1	2	1	2	B	F	M	b	h
1	2	3	4	CS	US	Mb	B	M	0	30	60	0	2	Mr	0	15	0	15	C	G	N	c	i
5	6	7	8	No	Nd	Wf	W	F	5	35	65	1	3	Sg	5	10	5	10	D	H	O	d	k
1	2	3	4	Fh	Ff	Fm	7	1	10	40	70	90	4	0	1	3	0	2	St	I	P	e	l
5	6	7	8	Hh	Hf	Hm	8	2	15	45	75	95	100	Un	2	4	1	3	4	K	Un	f	m
1	2	3	4	X	Un	Ft	9	3	i	o	X	R	L	E	A	6	0	US	Ir	So	US	Ir	So
5	6	7	8	Ot	En	Mt	10	4	k	d	Y	S	M	F	B	10	1	Or	En	Wa	Or	En	Wa
1	2	3	4	W	R	OK	11	5	l	e	Z	T	N	G	C	15	2	Sw	FC	EC	Sw	FC	EC
5	6	7	8	7	4	1	12	6	m	f	NG	U	O	H	D	Un	3	Nw	Bo	Hu	Nw	Bo	Hu
1	2	3	4	8	5	2	Oc	0	n	g	a	V	P	I	Al	Na	4	Dk	Fr	It	Dk	Fr	It
5	6	7	8	9	6	3	0	p	o	h	b	W	Q	K	Un	Pa	5	Ru	Ot	Un	Ru	Ot	Un

FIGURE 16.—Content of 1890 card (from reading board)

Punch Card For 1890 Census

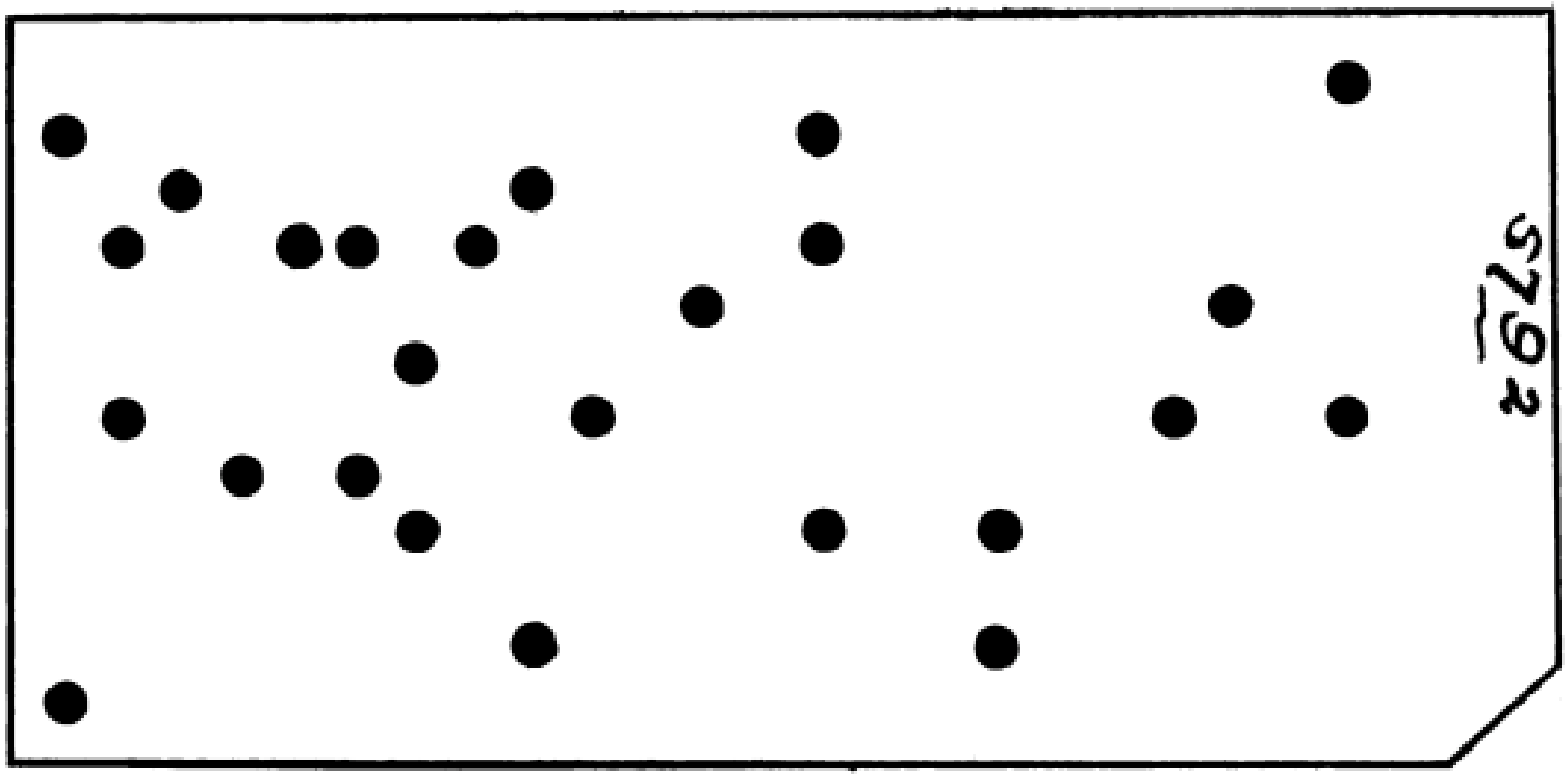
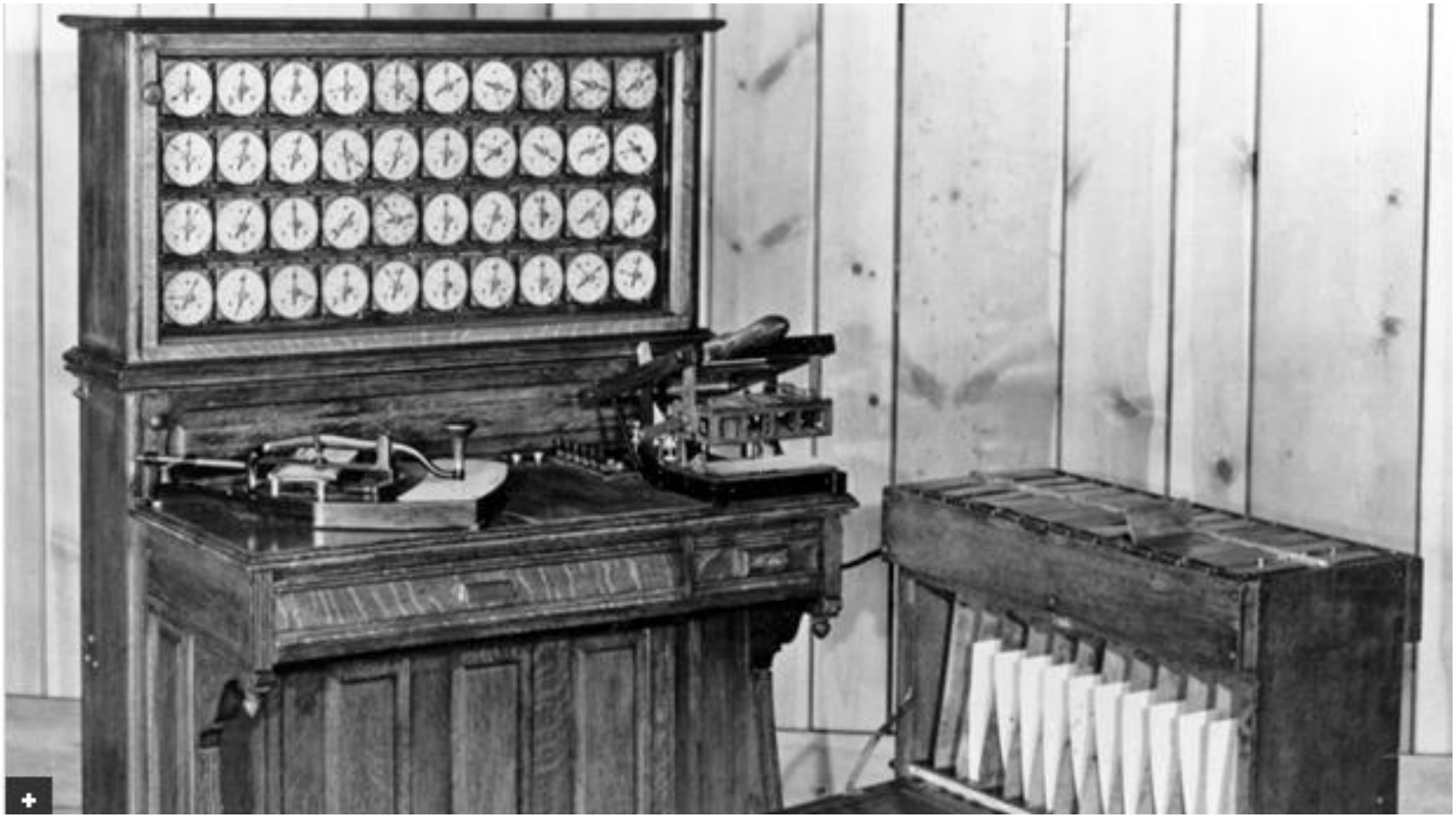


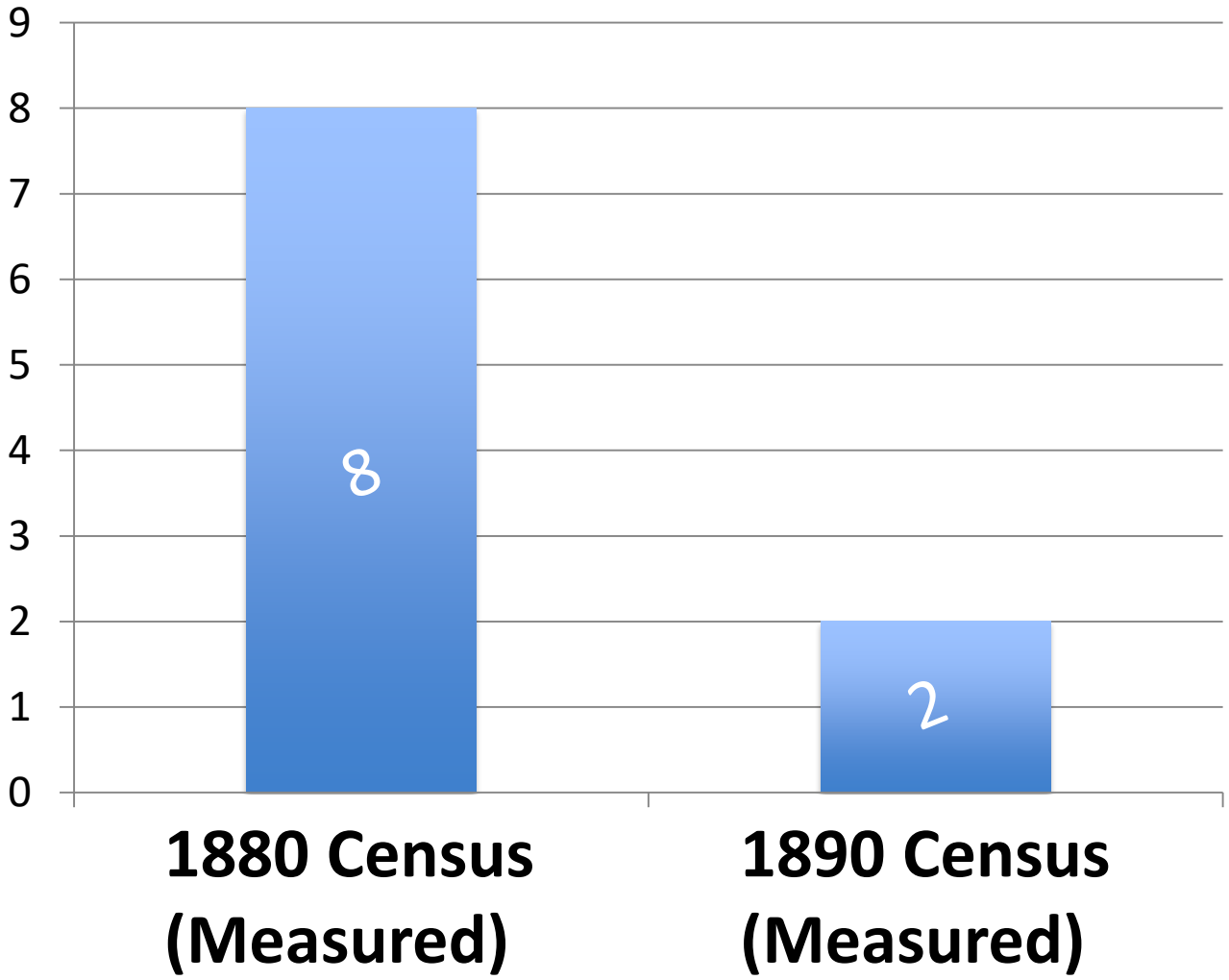
FIGURE 15.—Complete card as punched for 1890

Electronic Punch Card Reader

Herman Hollerith's Tabulating Machine

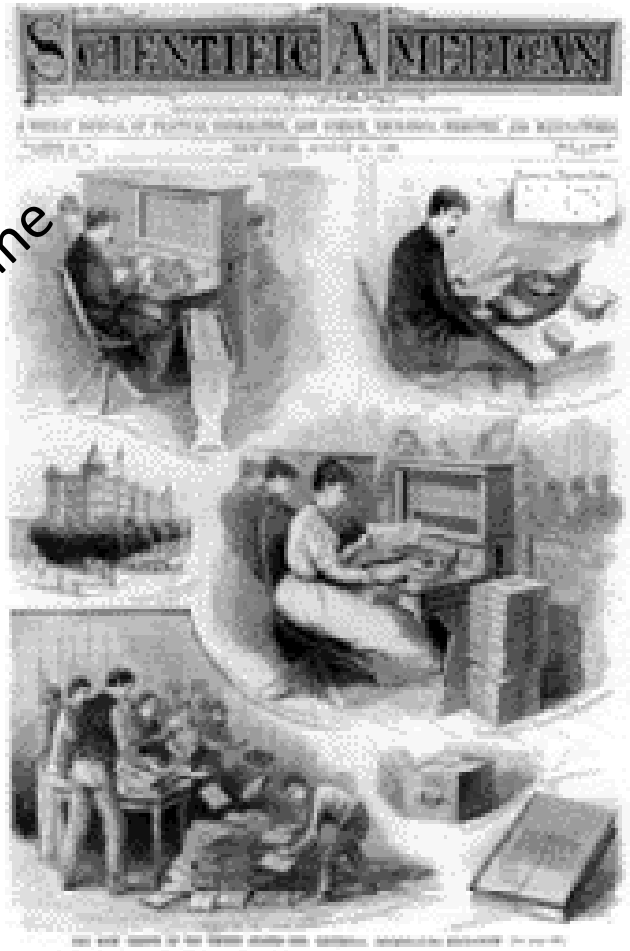


Years to Tabulate Census



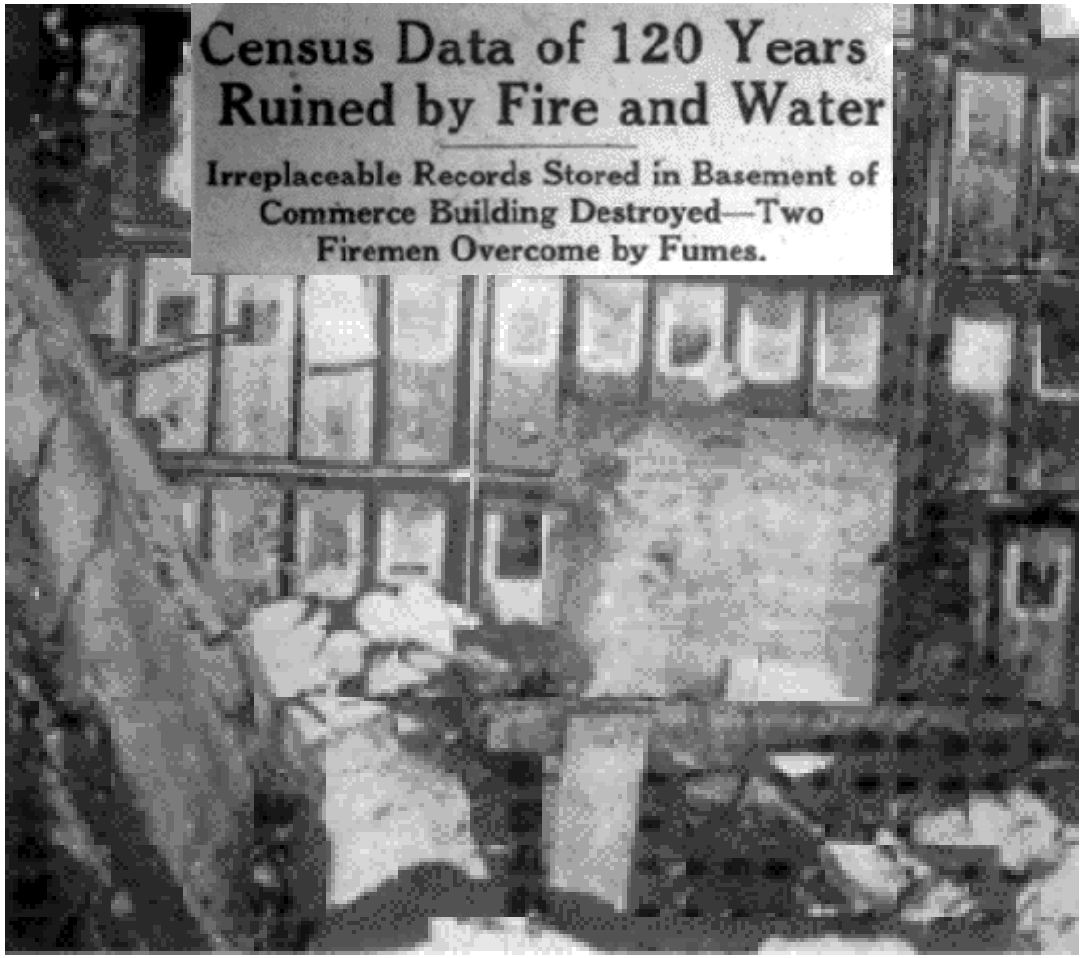
Tabulating Machines A Success!

Cover of Scientific American Magazine August 30, 1890



Punch card machines in action

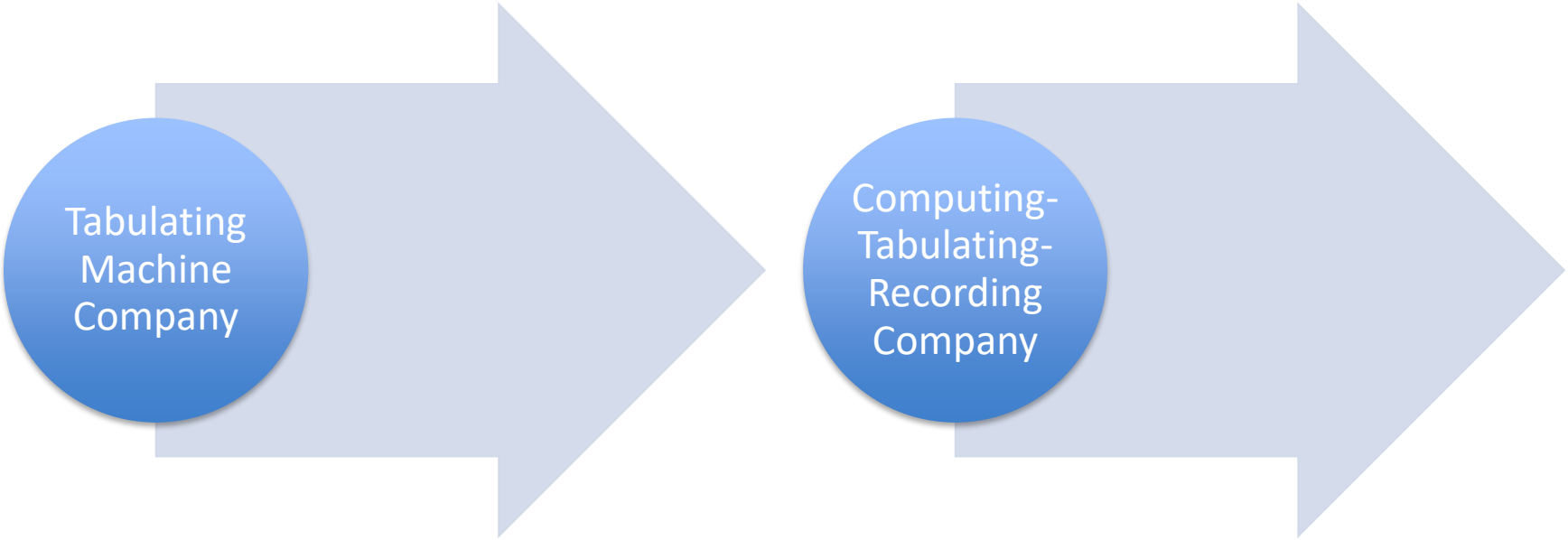
Side Note:
1890 Census Was Lost to Fire : (



Tabulating Machines Took Off

Countries Tabulating Census With Hollerith Machine by 1911

Hollerith Sold Company in 1911



Eventually Changing its Name To:

(In 2005, IBM sold its computer line to Lenovo)

Thoughts and Tie-In

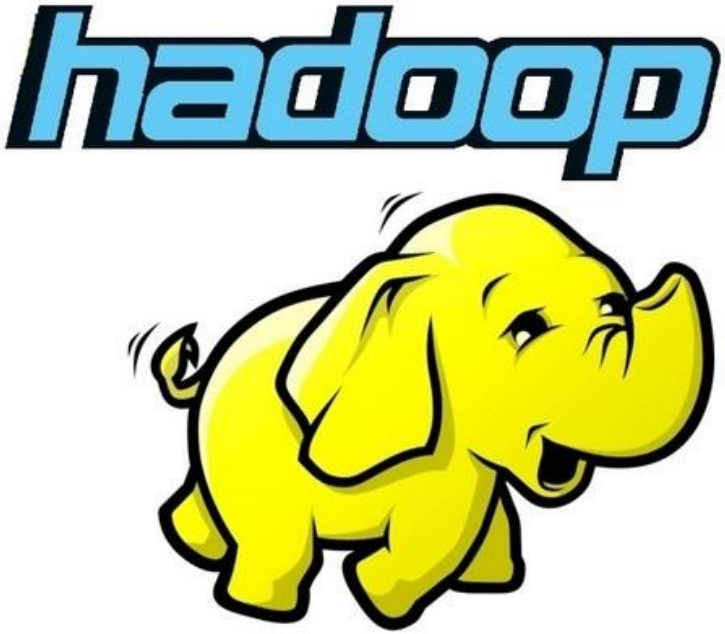
Now you know how  got its start

solving a big data challenge by developing tabulation machines that used punch cards to tabulate the 1890 U.S. Census.

Classic “Big Data” Definition

http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

Brief Hadoop Origin Story…



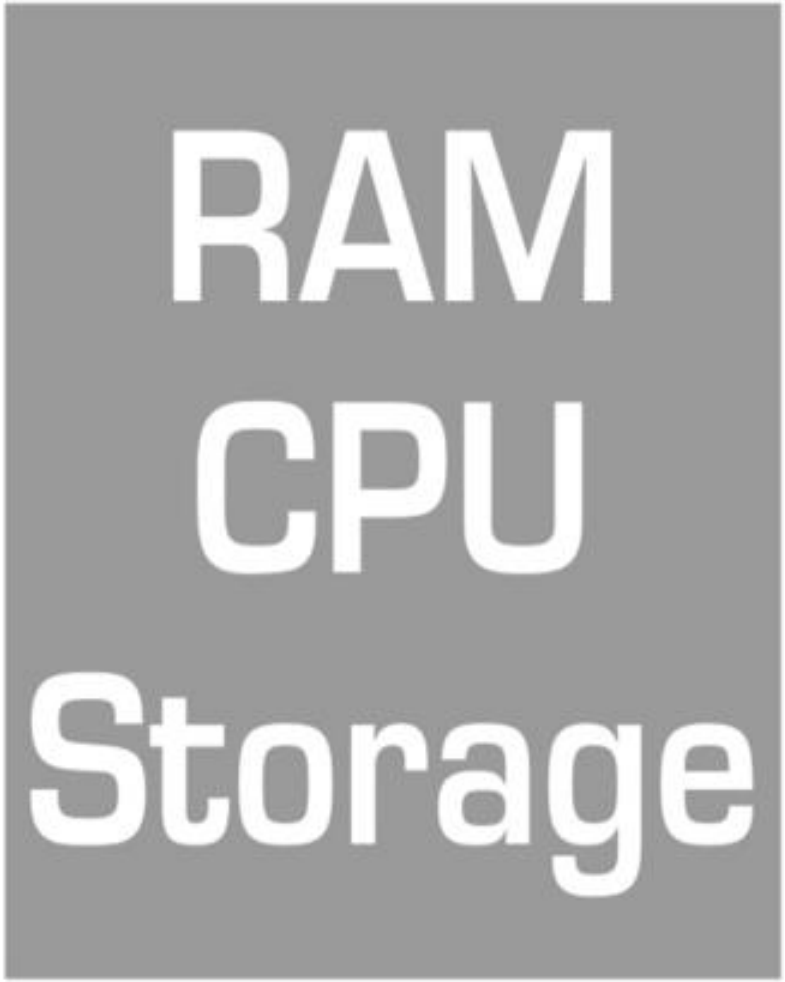
Vertical Scaling

http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

Vertical Scaling

http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

Vertical Scaling



http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

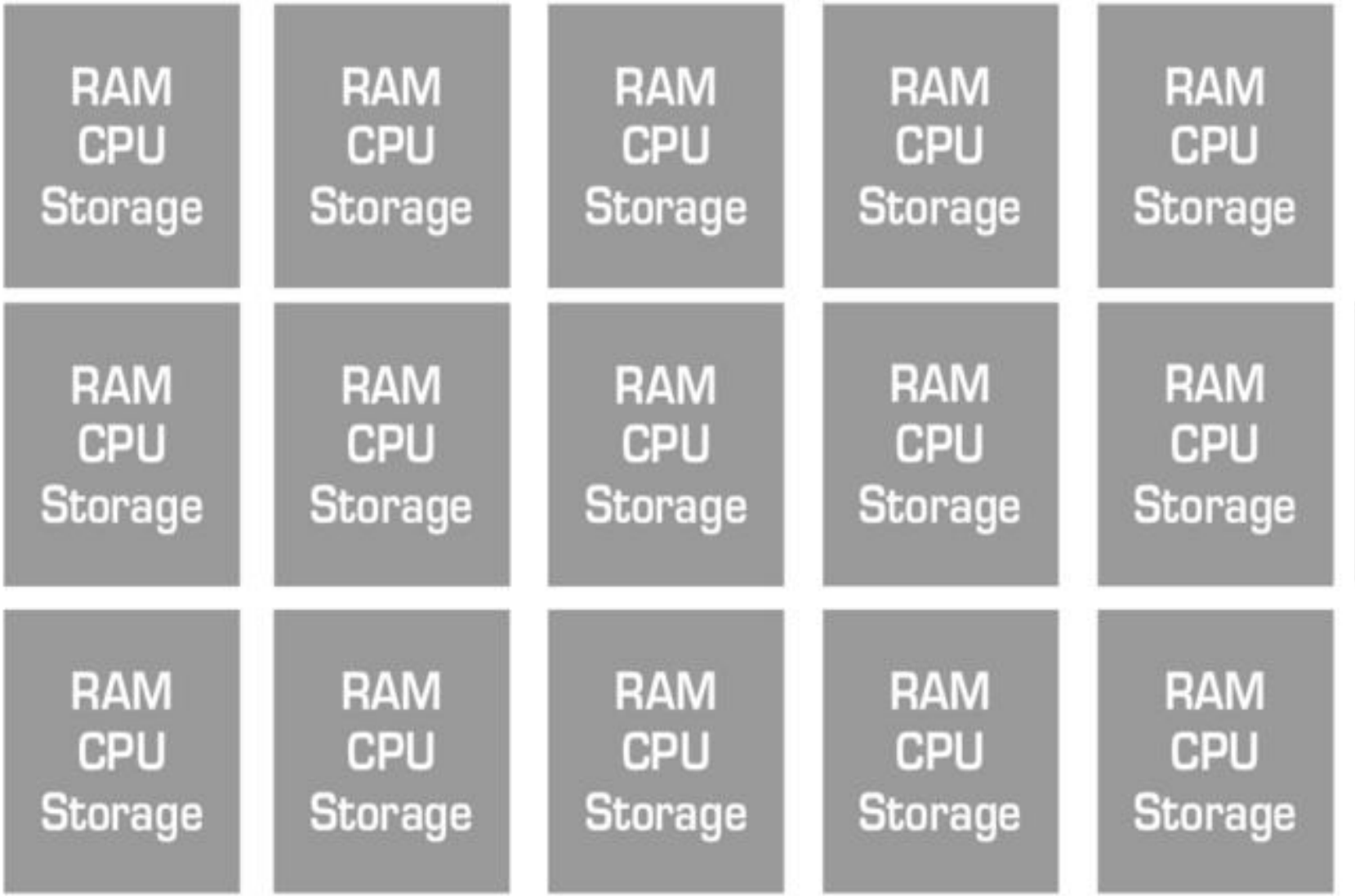
Horizontal Scaling

http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

Horizontal Scaling

http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

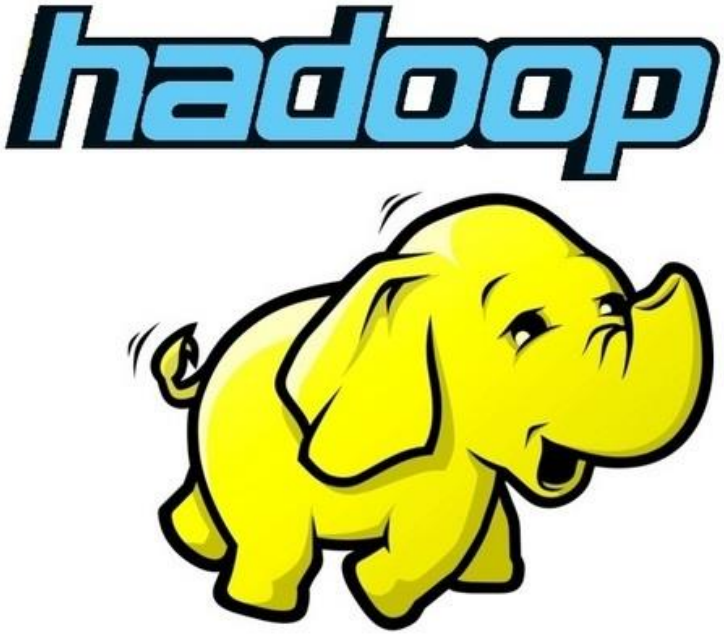
Horizontal Scaling



http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

But...

HADoop



HaDOOP

Hadoop is a *framework*

http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

Hadoop is a *framework*

http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

Hadoop is a *framework*

http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

Hadoop is meant to be:

- Massively parallel (horizontally scalable)
- Reliable (fault tolerant)
- Run on ordinary computers (cheap)

No one said anything about high performance!

The Hadoop App Store

http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

Data Storage

OK, first things first!

I want to store all of my <<Big Data>>

http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

Data Storage

There is an app for that!

http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

Data Processing

Data stored, check!

Now I want to create insights from my data!

http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

Data Processing

There is an app for that!

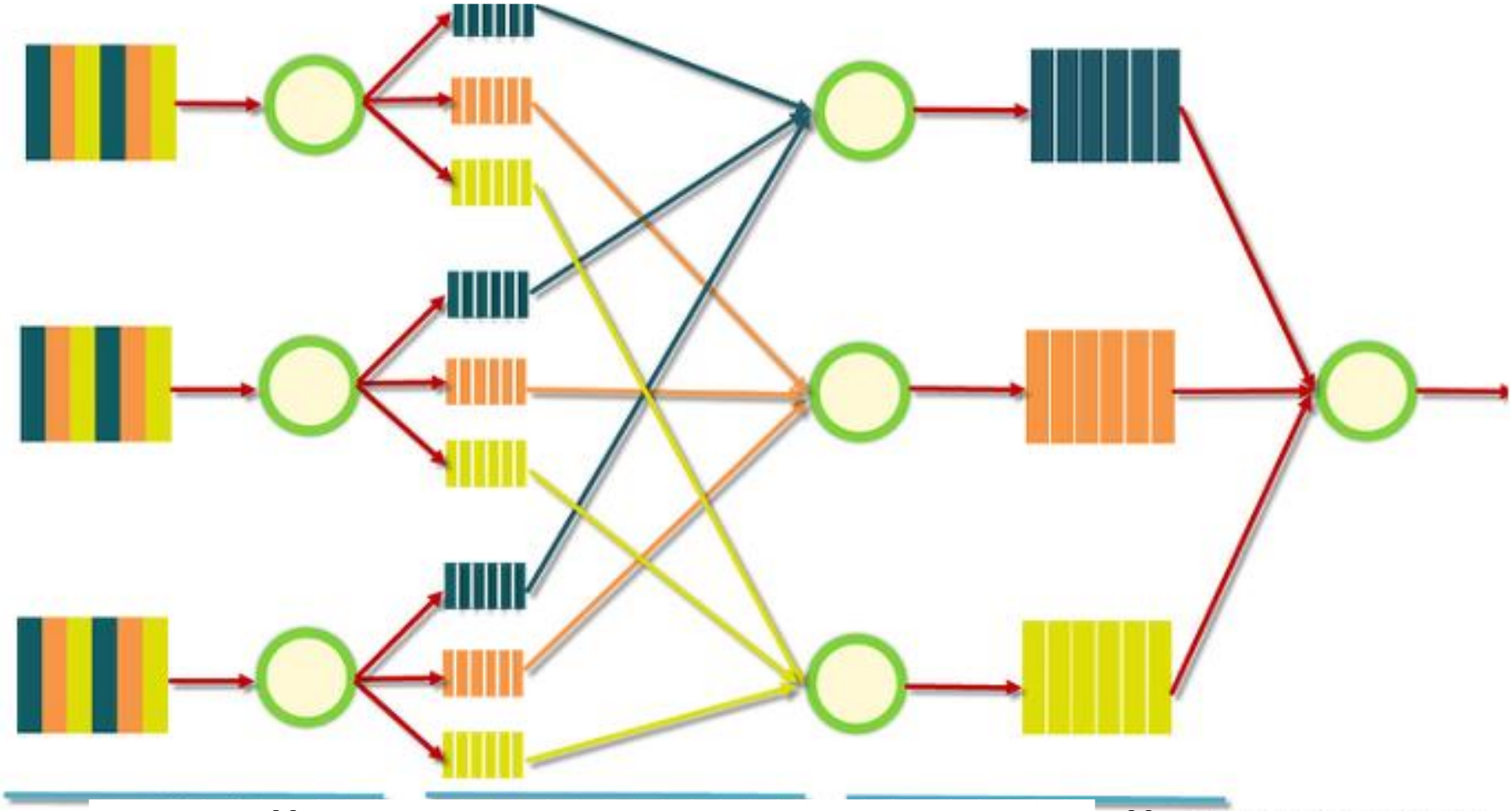
http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

MapReduce

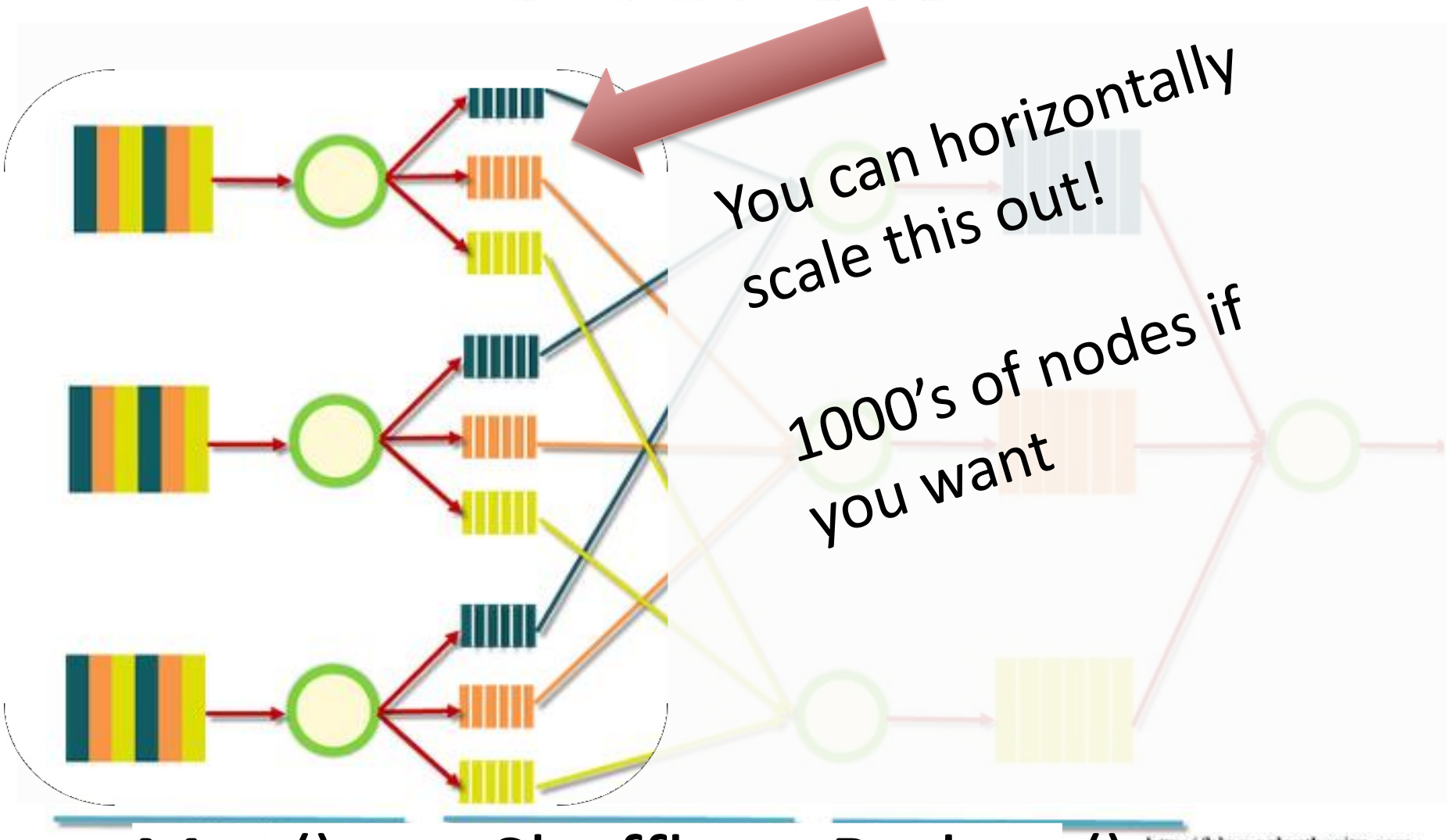
- **Programming model for distributed computations at a massive scale**
- **Execution framework for organizing and performing such computations**

http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

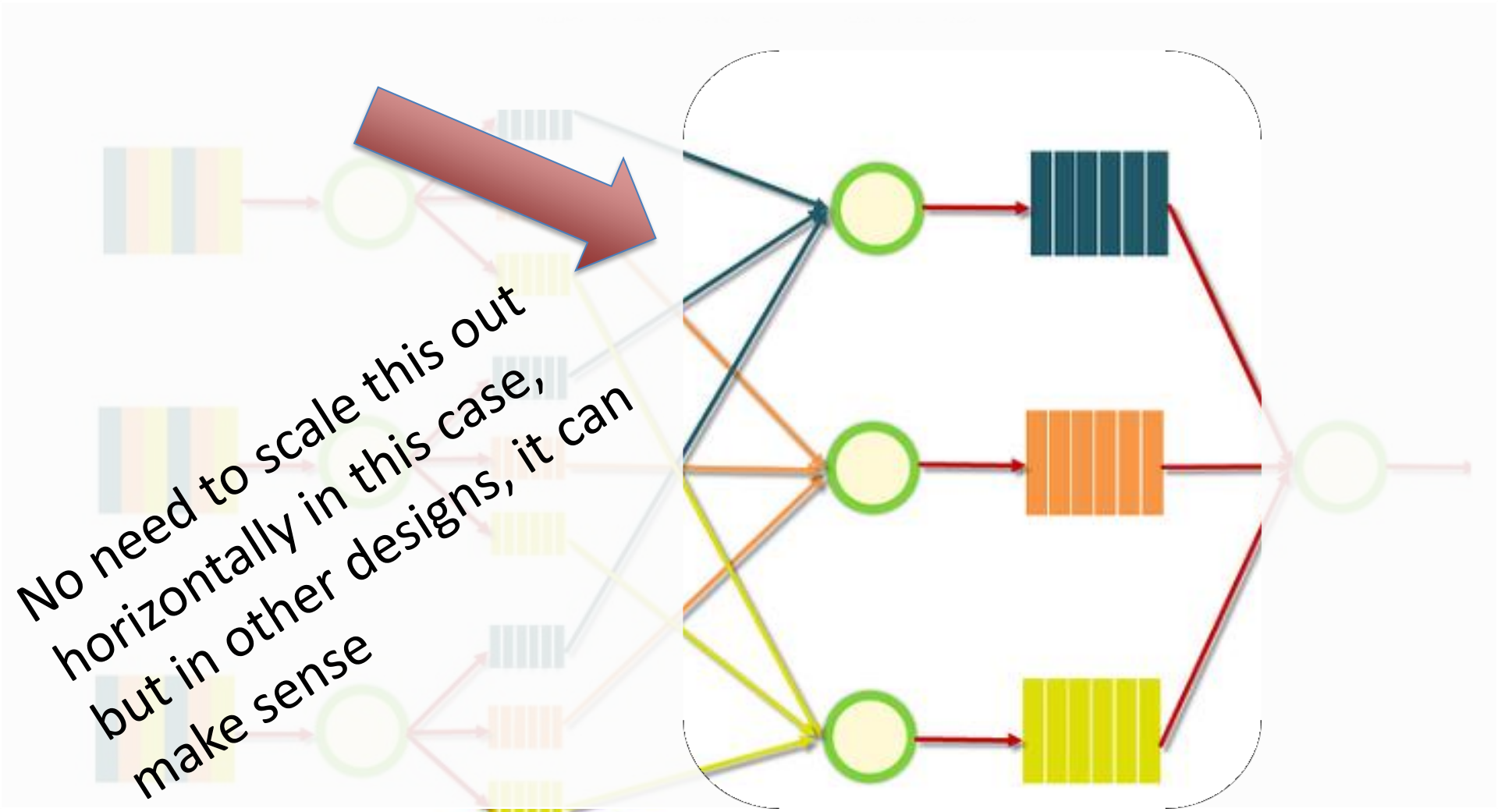
How MapReduce Works (simplified)



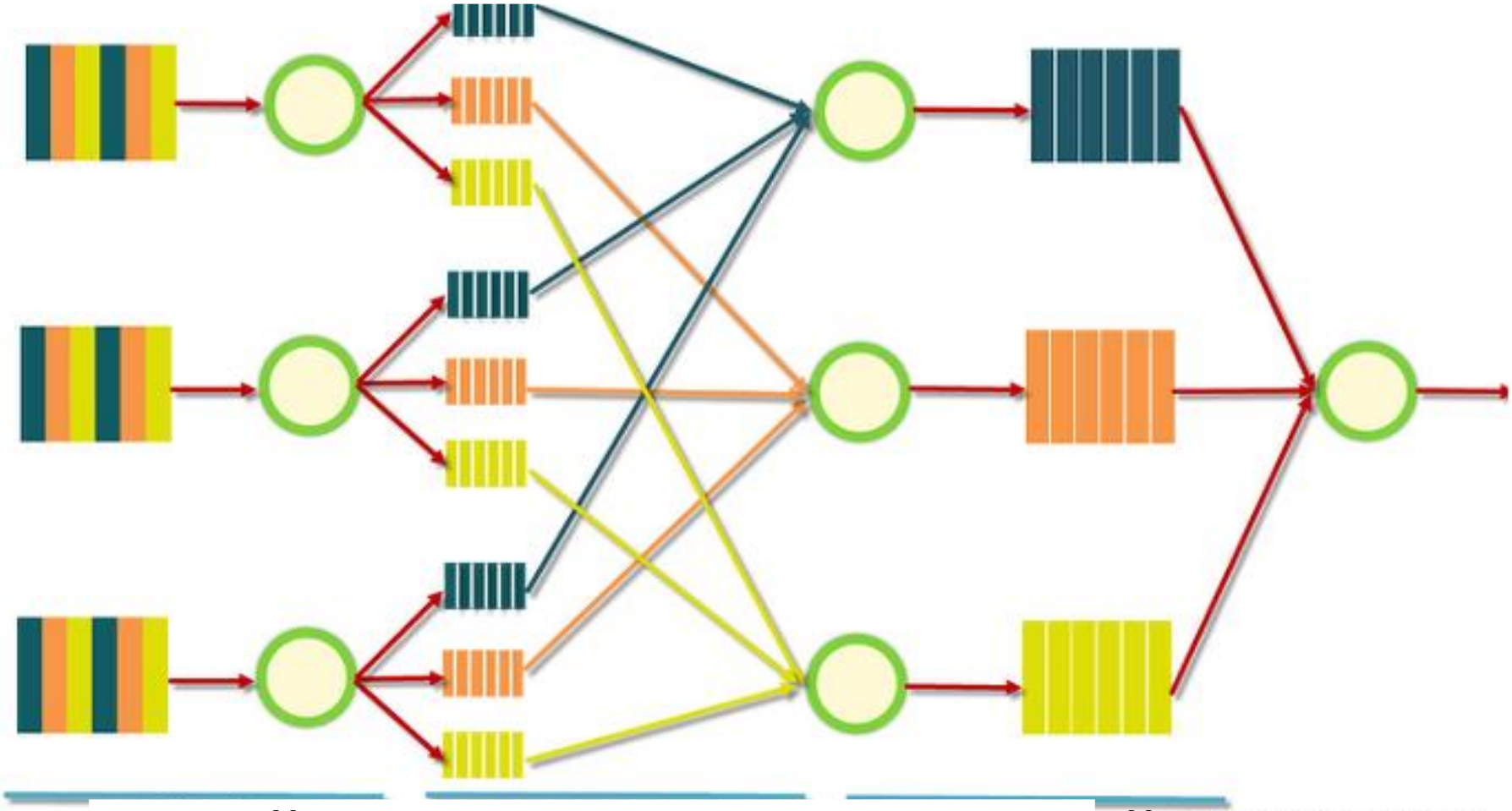
How MapReduce Works (simplified)



How MapReduce Works (simplified)



How MapReduce Works (simplified)



Scripting for Hadoop

Java for MapReduce?

I dunno,

I'm more of a

scripting guy...

http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

Scripting for Hadoop

There is an app for that!

Apache Pig

- **High-level data flow language**
- **Made of two components:**
  - **Data processing language Pig Latin**
  - **Compiler to translate Pig Latin to MapReduce**

Abstracts you from specific details and allows you to focus on data processing

Pig in the Hadoop ecosystem

http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

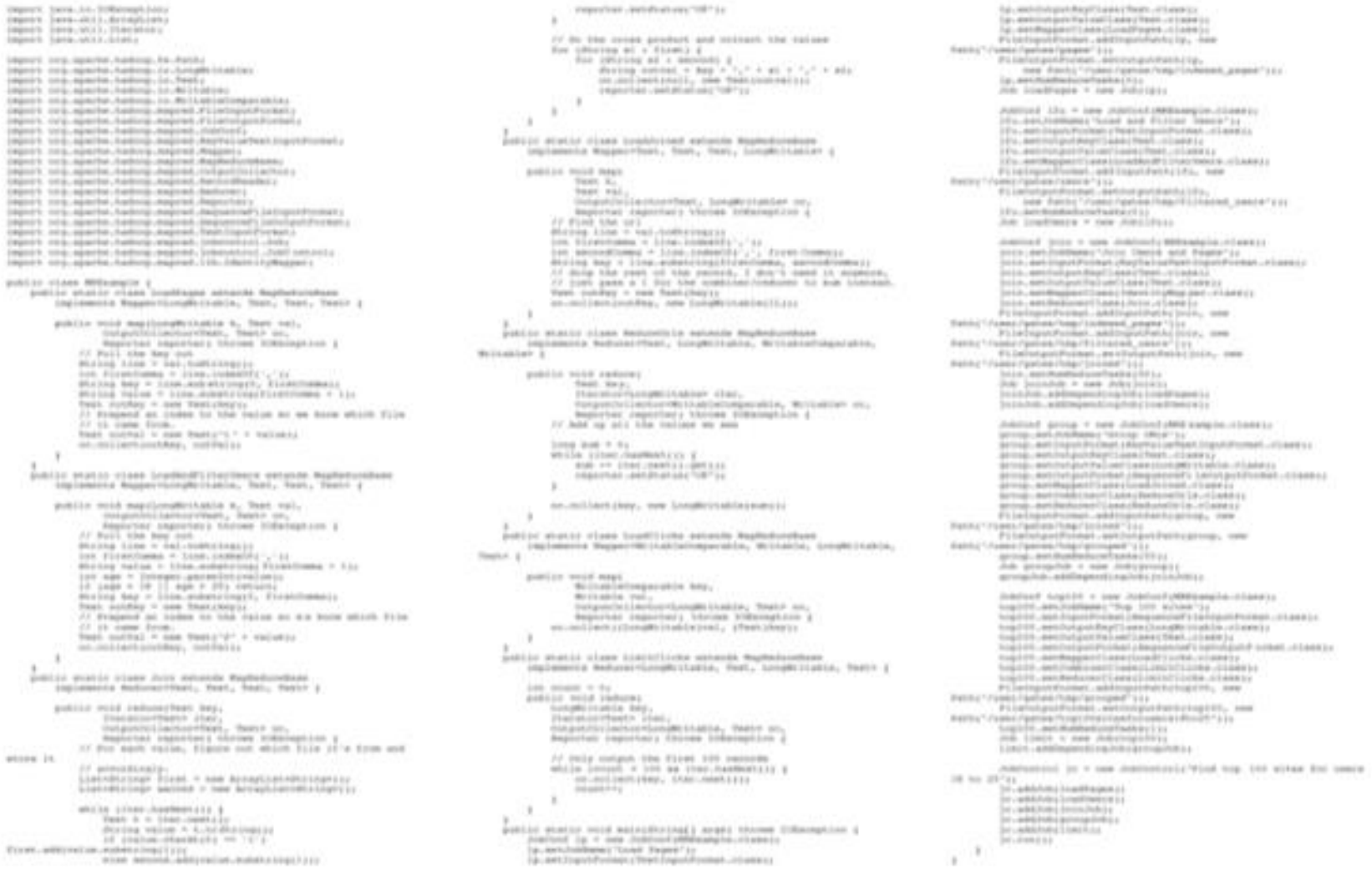
Pig Latin

```
users = LOAD 'users.txt' USING PigStorage(',') AS (name,
  age);
pages = LOAD 'pages.txt' USING PigStorage(',') AS (user,
  url);

filteredUsers = FILTER users BY age >= 18 and age <=50;
joinResult = JOIN filteredUsers BY name, pages by user;
grouped = GROUP joinResult BY url;
summed = FOREACH grouped GENERATE group,
  COUNT(joinResult) as clicks;
sorted = ORDER summed BY clicks desc;
top10 = LIMIT sorted 10;

STORE top10 INTO 'top10sites';
```

Try that with Java...



http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

SQL for Hadoop

OK, Pig seems quite useful...

But I'm more of a SQL person...

SQL for Hadoop

There is an app for that!

http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

Apache Hive

- **Data Warehousing Layer on top of Hadoop**
- **Allows analysis and queries using a SQL-like language**

Hive is best for data analysts familiar with SQL who need to do dynamic queries, summarization and data analysis

http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

Hive in the Hadoop ecosystem

http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

Hive Example

```
CREATE TABLE users(name STRING, age INT);
CREATE TABLE pages(user STRING, url STRING);

LOAD DATA INPATH '/user/sandbox/users.txt' INTO
TABLE 'users';
LOAD DATA INPATH '/user/sandbox/pages.txt' INTO
TABLE 'pages';

SELECT pages.url, count(*) AS clicks FROM users JOIN
pages ON (users.name = pages.user)
WHERE users.age >= 18 AND users.age <= 50
GROUP BY pages.url
SORT BY clicks DESC
LIMIT 10;
```

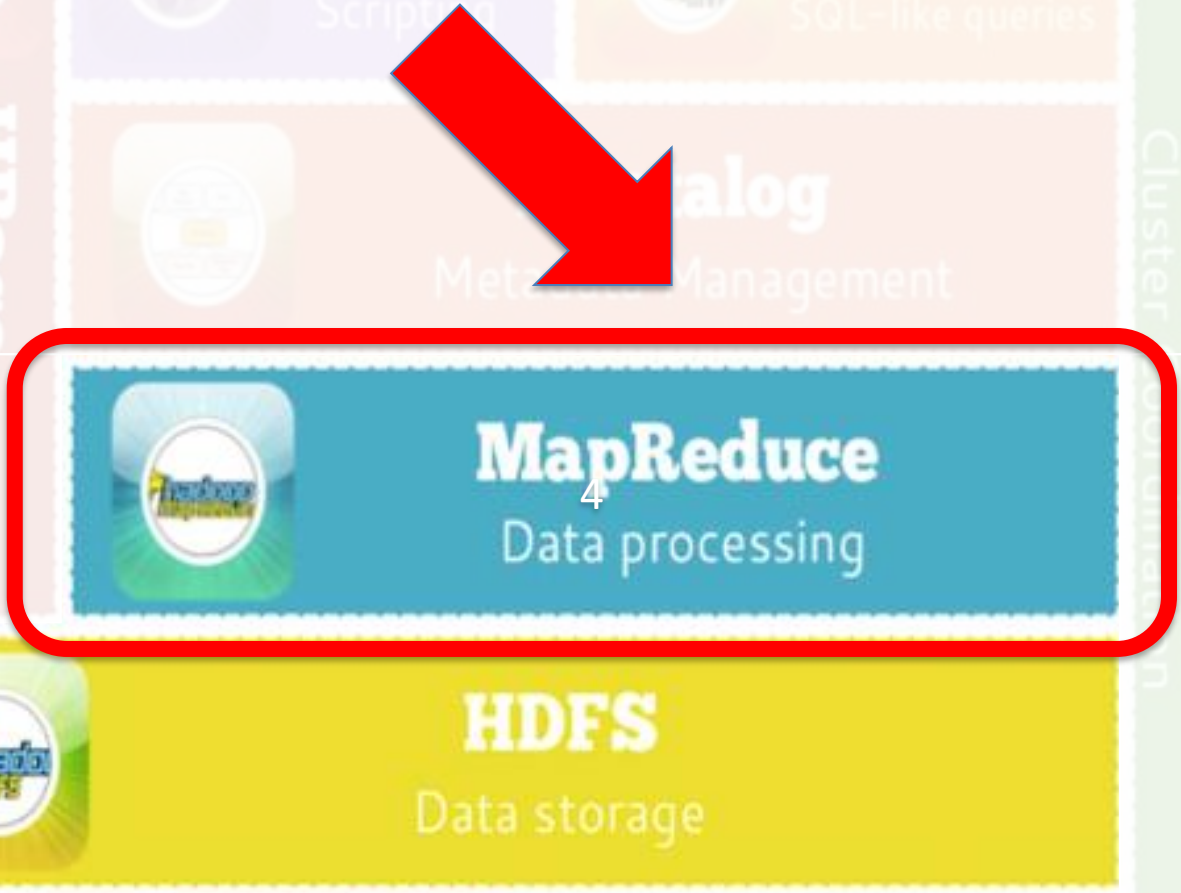
http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-25557364

Mahout
Machine Learning
Pig
Scripting
Hive
SQL-like queries
HBase
NoSQL Database
HCatalog
Metadata Management
MapReduce
Data processing
HDFS
Data storage
Scoop
Import & Export of
Flume
Import & Export of
ZooKeeper
Cluster Coordination
Ambari
Cluster installation & management
Oozie
Workflow automatization


Mahout
Machine Learning
Pig
Scripting
Hive
SQL-like queries
HBase
NoSQL Database
HCatalog
Metadata Management
MapReduce
Data processing
HDFS
Data storage
ZooKeeper
Cluster Coordination
Ambari
Cluster installation & management
Oozie
Workflow automation
Scoop
Import & Export of
Flume
Import & Export of

Let’s Take a Step Back…

Have you noticed that MapReduce is the ONLY data processing algorithm? What a limitation!

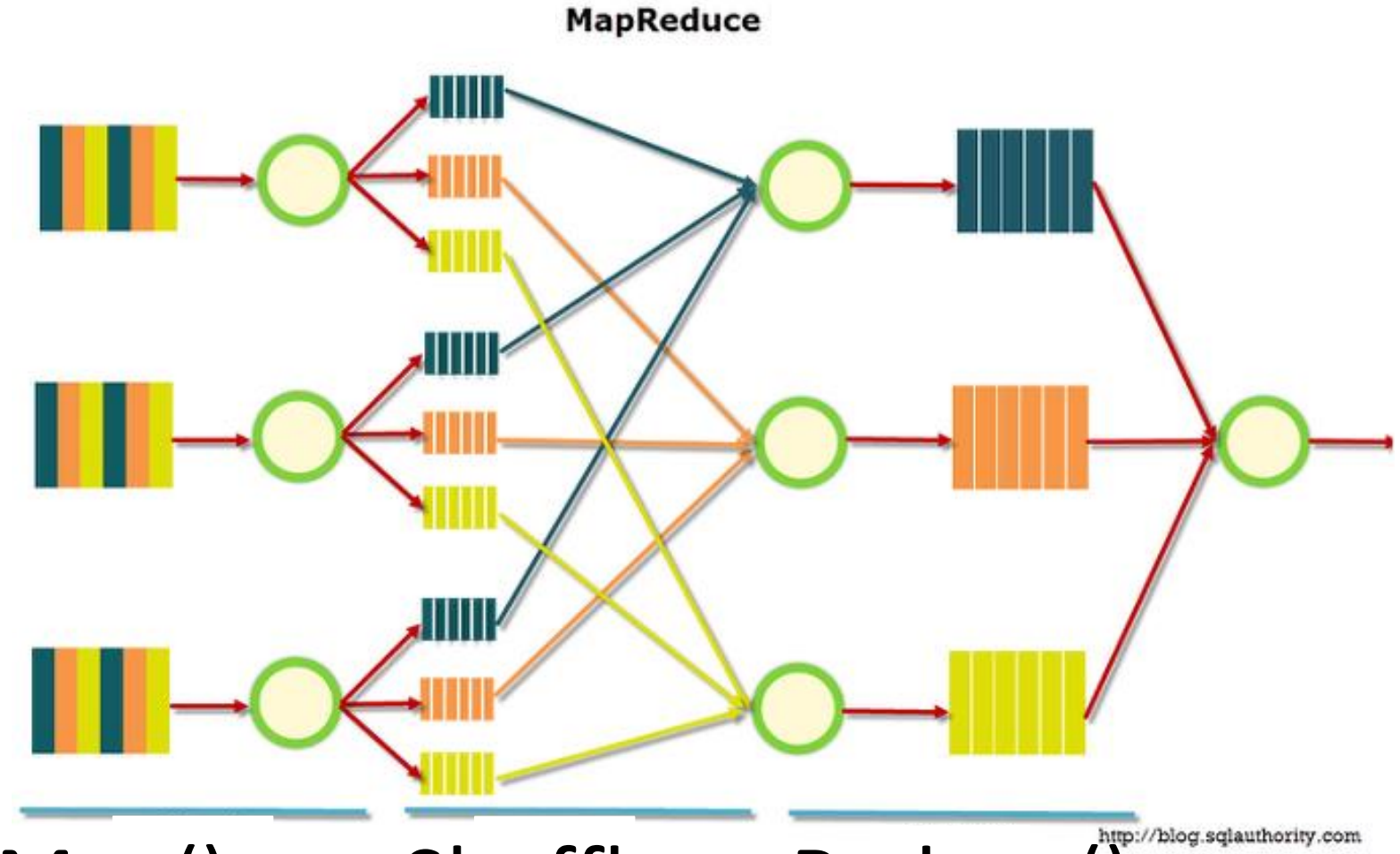


Don’t Use Hadoop - Your Data Isn’t That Big

(Blog post by Chris Stucchio)

https://www.chrisstucchio.com/blog/2013/hadoop_hatred.html

Is MapReduce Our ONLY Choice?



Map() Shuffle Reduce()

https://www.chrisstucchio.com/blog/2013/hadoop_hatred.html

MapReduce Ties Our Hands

https://www.chrisstucchio.com/blog/2013/hadoop_hatred.html

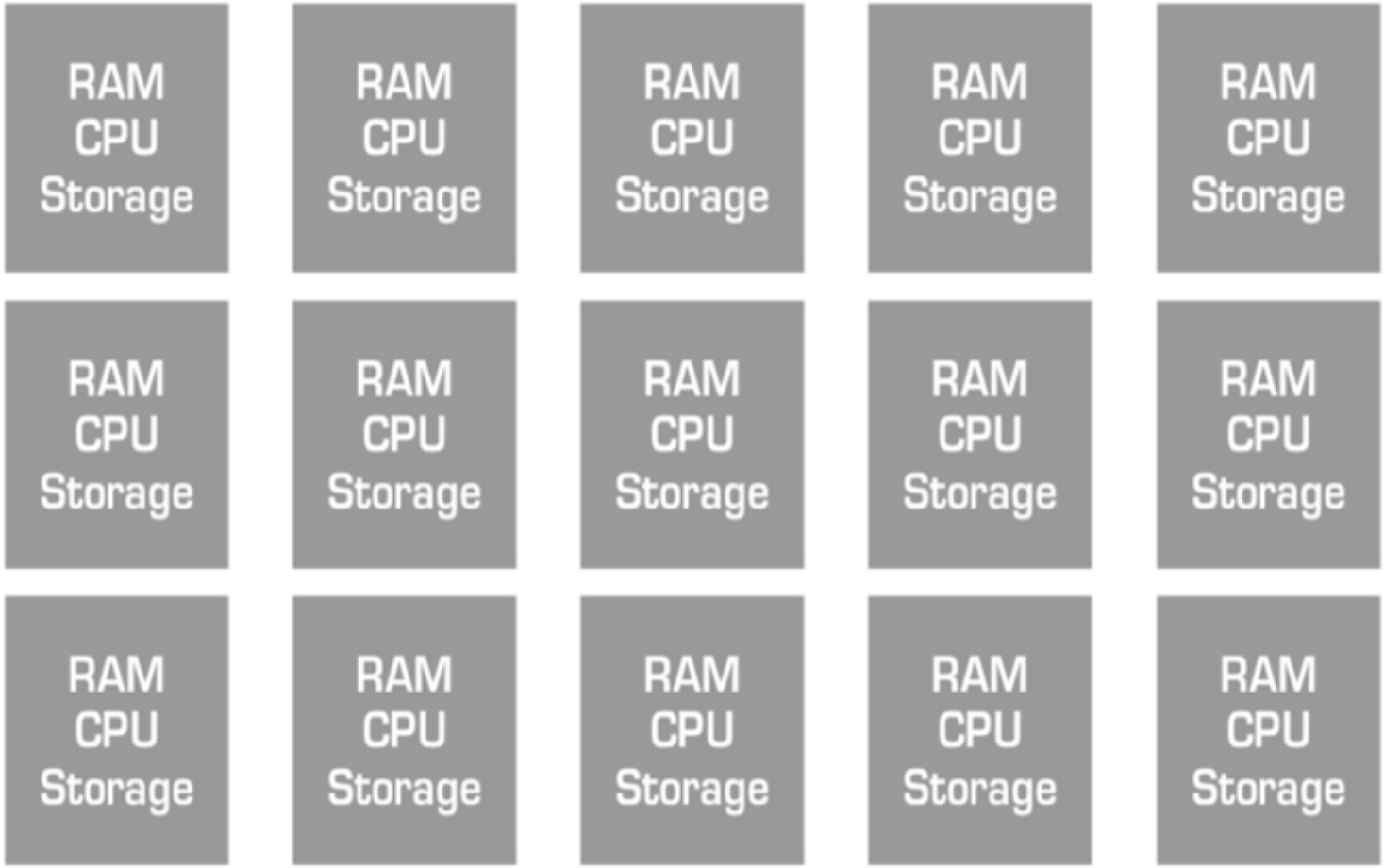
MapReduce Only Makes Sense If You NEED Horizontal Scaling

https://www.chrisstucchio.com/blog/2013/hadoop_hatred.html

MapReduce Only Makes Sense If You NEED Horizontal Scaling

https://www.chrisstucchio.com/blog/2013/hadoop_hatred.html

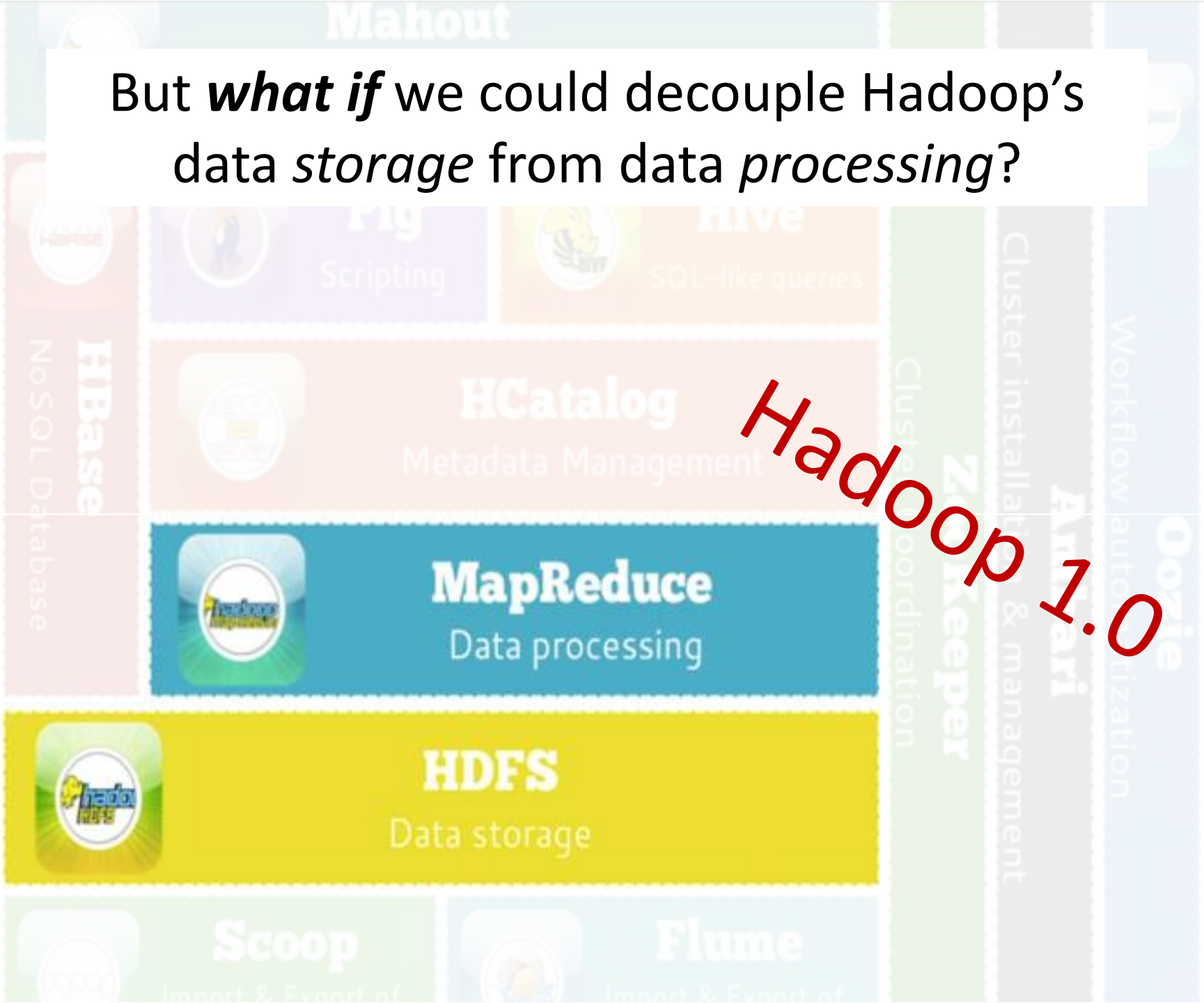
MapReduce Only Makes Sense If You NEED Horizontal Scaling



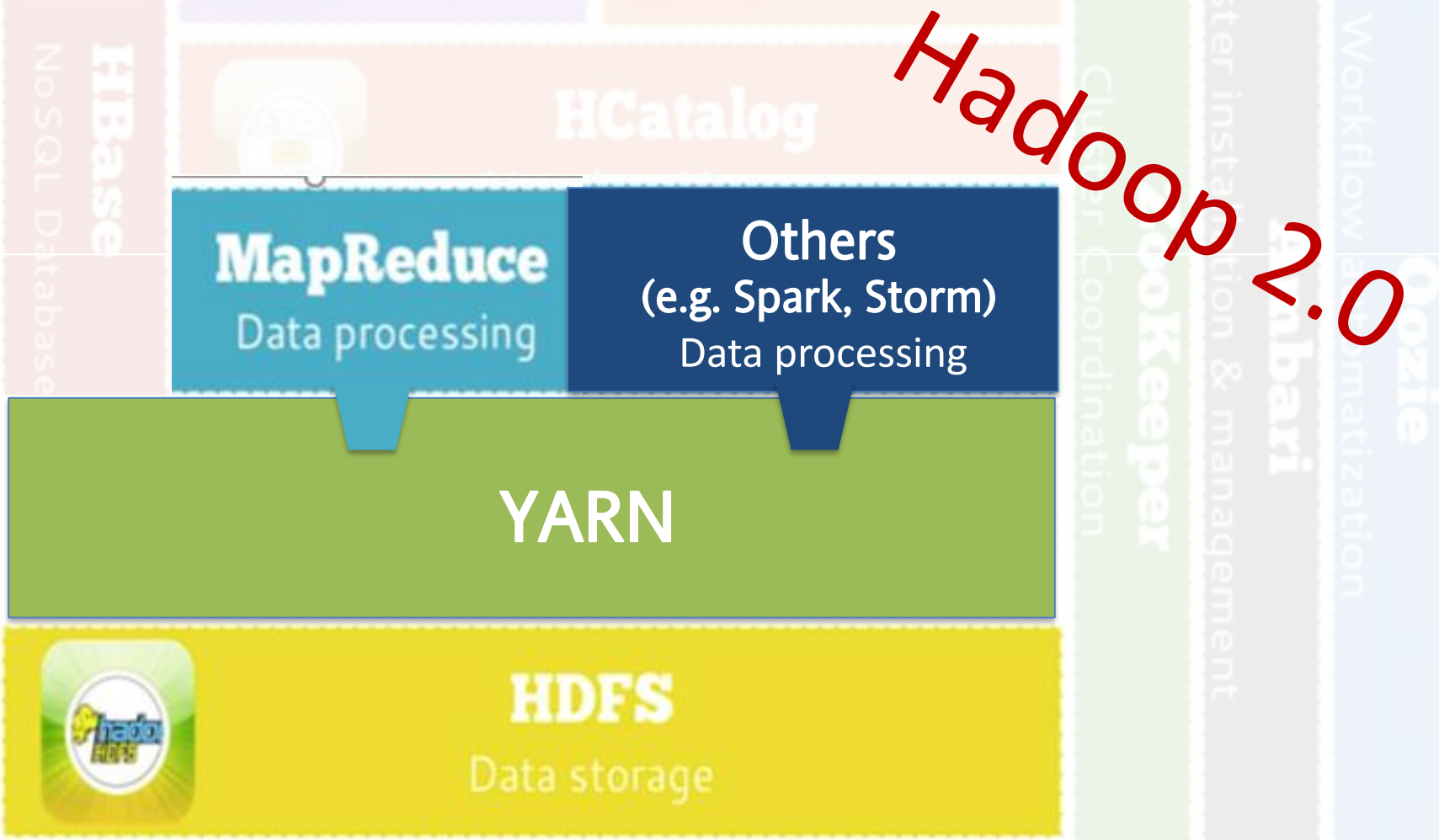
https://www.chrisstucchio.com/blog/2013/hadoop_hatred.html

https://www.chrisstucchio.com/blog/2013/hadoop_hatred.html

But ***what if*** we could decouple Hadoop's data *storage* from data *processing*?



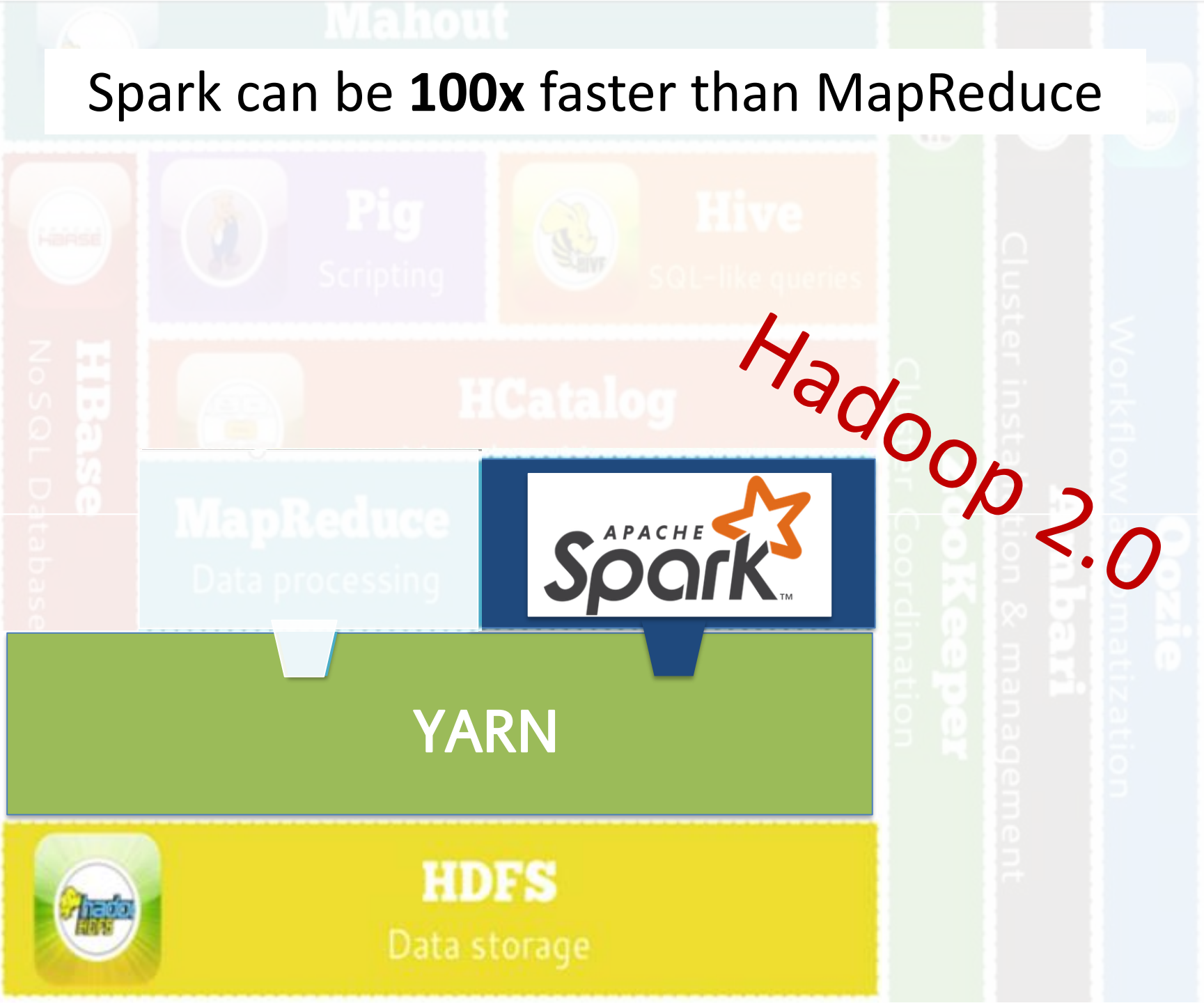
Hadoop “2.0”: YARN (modular data processing engines!)



“Lightning-fast cluster computing”

https://spark.apache.org/

Spark can be **100x** faster than MapReduce



You *could* run Spark on your laptop (rather than a grid of computers)

e.g. via R https://spark.rstudio.com

e.g. via Python https://stackoverflow.com/a/49587560/10485230

Spark Machine Learning: MLlib

MLlib is Apache Spark's scalable machine learning library.

Ease of Use

Usable in Java, Scala, Python, and R.

Performance

High-quality algorithms, 100x faster than MapReduce.

```
data = spark.read.format("libsvm")\
  .load("hdfs://...")

model = KMeans(k=10).fit(data)
```

Calling MLlib in Python

https://spark.apache.org/mllib/

This is a Big Deal (to employers)

This is a Big Deal (to employers)

Prototype analytical model in R/Python…

Using Spark MLlib methods…

This is a Big Deal (to employers)

Prototype analytical model in R/Python…

Using Spark MLlib methods…

Test your code in Spark “standalone” mode (run it just on 1 computer)

This is a Big Deal (to employers)

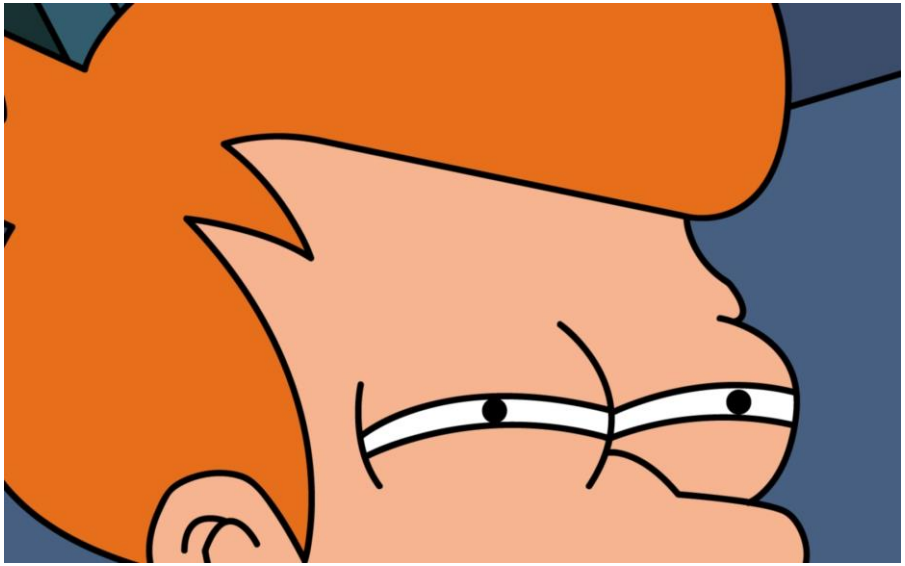
Prototype analytical model in R/Python…

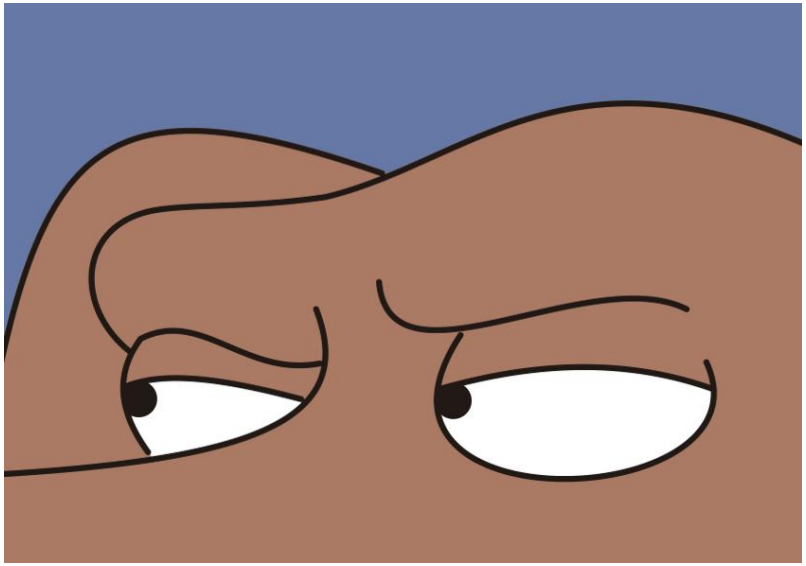
Using Spark MLlib methods…

Test your code in Spark “standalone” mode (run it just on 1 computer)

Scale it out easily over 100’s of nodes in AWS or Azure clouds!

What's the catch?





Why not use Spark (via R/Python) always?

Why not use Spark (via R/Python) always?

- This is frontier territory – tools are just coming into existence

Why not use Spark (via R/Python) always?

- This is frontier territory – tools are just coming into existence
- Not all algorithms / tools are available
  - (modeling may be easier than data cleaning)

Why not use Spark (via R/Python) always?

- This is frontier territory – tools are just coming into existence
- Not all algorithms / tools are available
  - (modeling may be easier than data cleaning)
- Running Spark on Windows can be … tricky

Want to Speed Up R/Python in the Practicum? (with Apache Spark)

- Try implementing code with Spark in single-machine mode on Practicum server so you'll use ALL processing power available.
  - https://spark.rstudio.com/
  - pip install pyspark
    - Note: https://stackoverflow.com/questions/46286436/running-pyspark-after-pip-install-pyspark
  - You'd be implementing methods from Spark ML lib
    - https://spark.apache.org/mllib/

Want to Speed Up R/Python in the Practicum? (with H2O.ai)

- Try implementing code with H20.ai in single-machine mode on Practicum server so you'll use ALL processing power available.
  - Intro to H2O in R
    - https://github.com/h2oai/h2o-tutorials/blob/master/h2o-open-tour-2016/chicago/intro-to-h2o.R
  - Intro to H2O in Python
    - https://github.com/h2oai/h2o-tutorials/blob/master/h2o-open-tour-2016/chicago/intro-to-h2o.ipynb
  - You'd be implementing methods from H2O.ai
    - http://docs.h2o.ai/h2o/latest-stable/h2o-py/docs/index.html
    - http://docs.h2o.ai/h2o/latest-stable/h2o-r/docs/index.html

Questions

- Acknowledgment to Uwe Seiler's excellent slides at: http://www.slideshare.net/uweseiler/introduction-to-the-hadoop-ecosystem-itstammtisch-darmstadt-edition