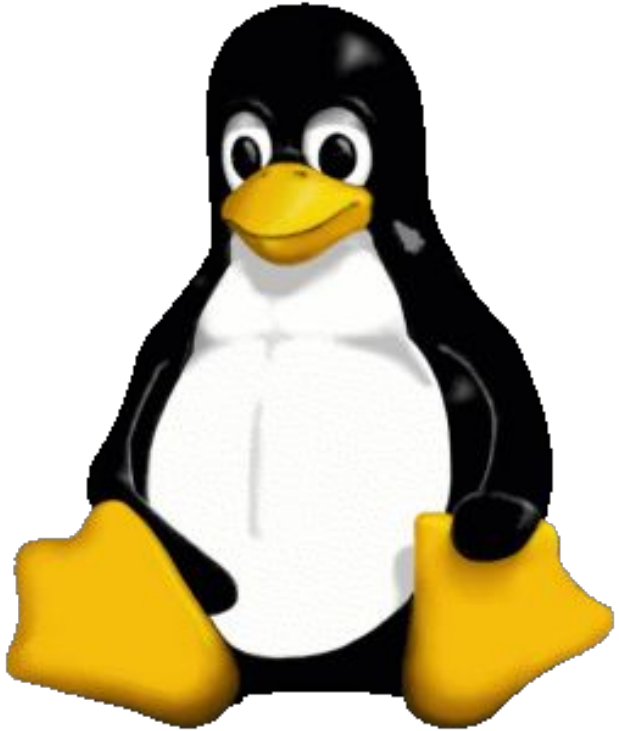


Welcome To CIALUG

A Brief Intro to Big Data

Andrew Denner, Central Iowa Linux User's Group

Welcome

Our website: http://cialug.org

List Server, IRC, and Slack...

Slides will be emailed out afterwards and posted to http://denner.co

Twitter: @adenner

Email denner@gmail.com

What is big data?

Doug Laney of Gartner Research “three V’s”

- Volume
- Velocity
- Variety

Others say:

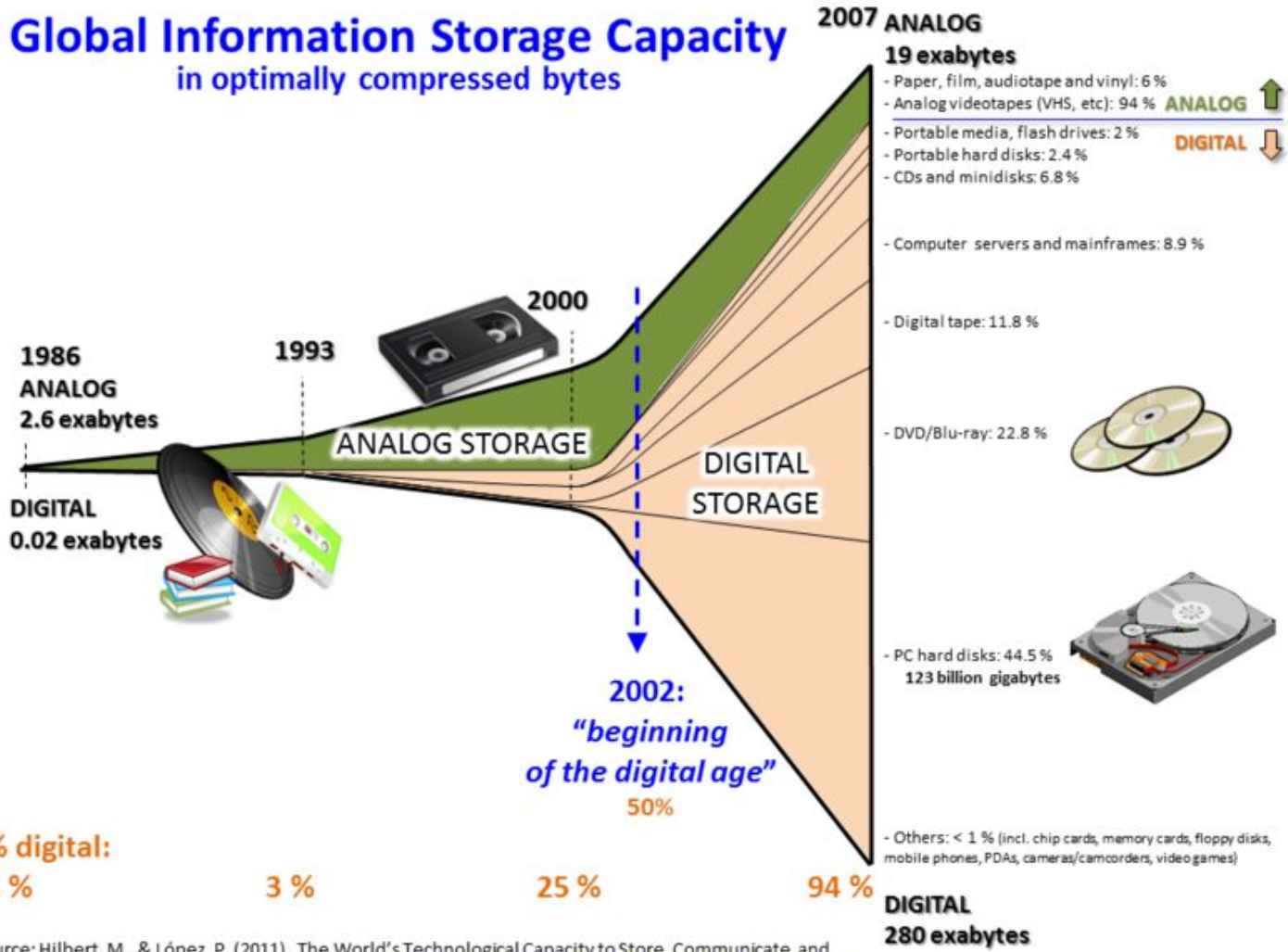
- Volume
- Velocity
- Variety
- Variability
- Veracity
- Visualization
- Value.

Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and information privacy.

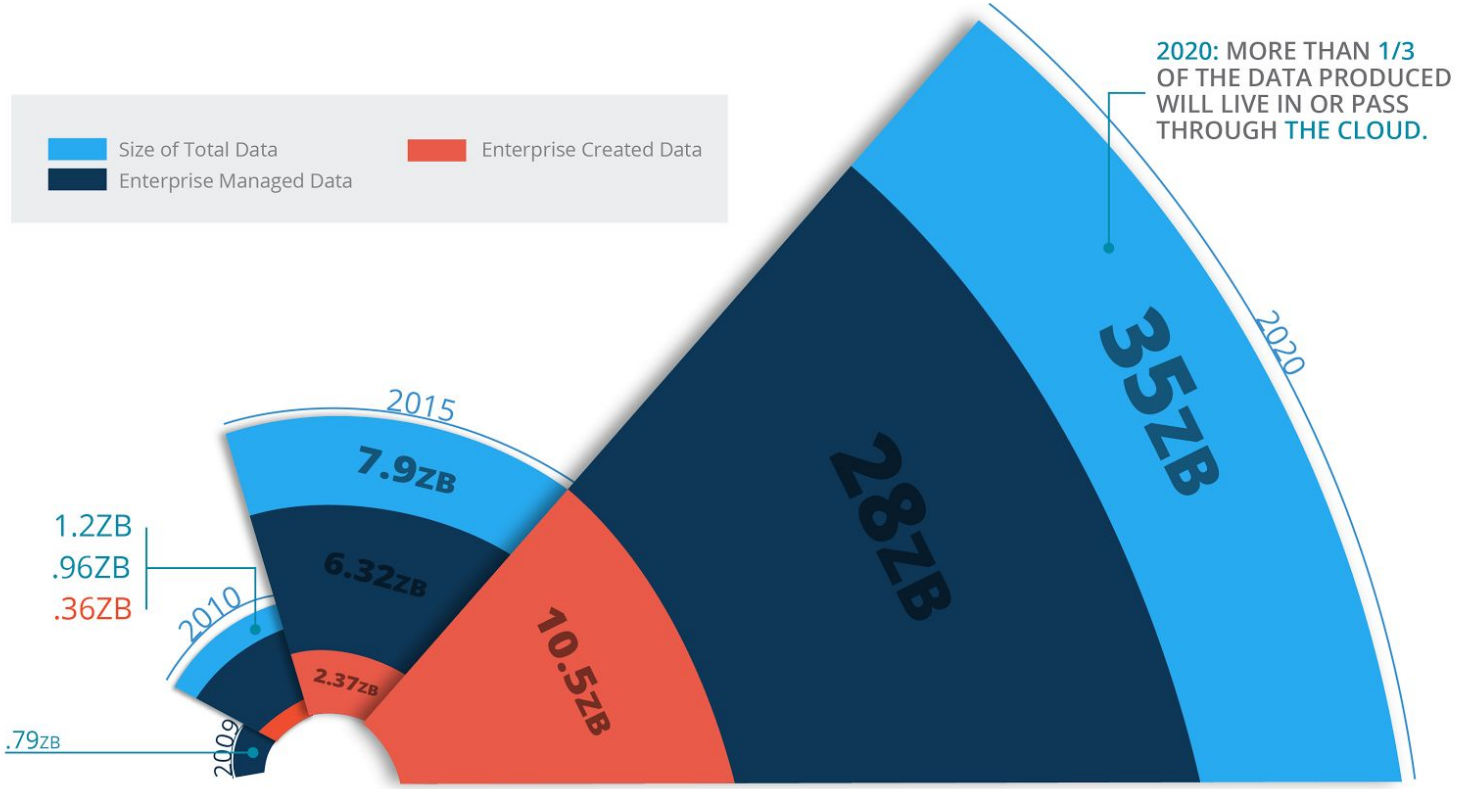
You know big data when you see it. When you can't handle the data using traditional methods.

Global Information Storage Capacity

in optimally compressed bytes



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html



https://www.impactradius.com/blog/7-vs-big-data/


Chromogram 1.2b
Choose file: Pearle
View: Timeline Block
Color by: Title
Search:
May 28, 2005
0 50 100 150 200 250 300 350 400 450
28 28 28 28 29 31
Jun 1 1 2 2 2 2 2
3 21 21 22 23 23 24
Jul 1 3 3 3 4 5 6
Aug 20 20 22 23 23 24 24 24 25 25 25 25 26 26 26 28
Sep 4 5 11 15 15
wpdms
uss
his
history
hist...
poli...
rus
battl...
davi...
geor...
james
list
robe...
the
the
vill...
env
universiti...
museum...
buil...
bui
Total 106776
a 131
battle 176
big 108
black 113
buildings 100
community 324
david 156
east 382
fort 182
george 125
grand 110
green 89
history 371
indian 103
james 129
john 332
la 225
lake 387
list 1123
little 101
michael 92
mount 233
music 129
national 286
new 578
north 299
operation 105
pages 344
paul 97
political 168
robert 119
south 98
st 94
st. 132
the 1298
towns 89
u.s. 154
united 194
uss 197
west 114
wikifun 101
wikipedia 123
wikipedians 90
william 156

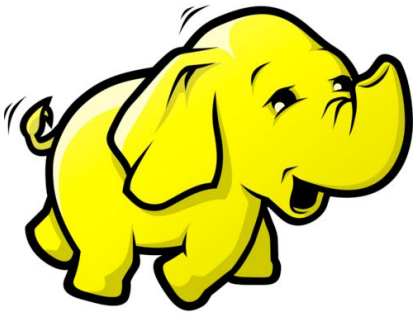
So I have big data, what do I do about it?

There are many ways to get to the same result.

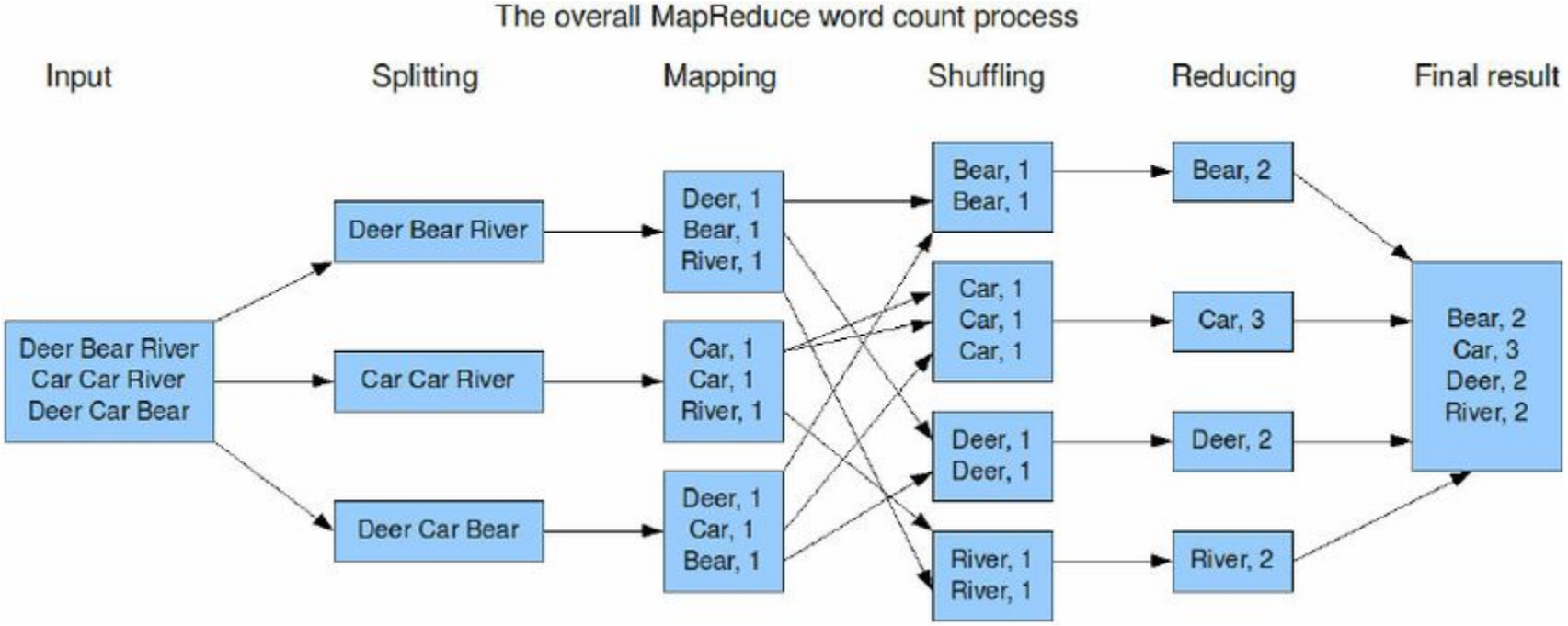
Tonight we are only going to scratch the surface.

Hadoop

- 2003 “Google File System” paper,
- 2004 "MapReduce: Simplified Data Processing on Large Clusters"
- Development implementing these ideas started in the Apache Nutch project
- 2006 moved to the sub project Apache Hadoop named by Doug Cutting at Yahoo for his son’s toy elephant

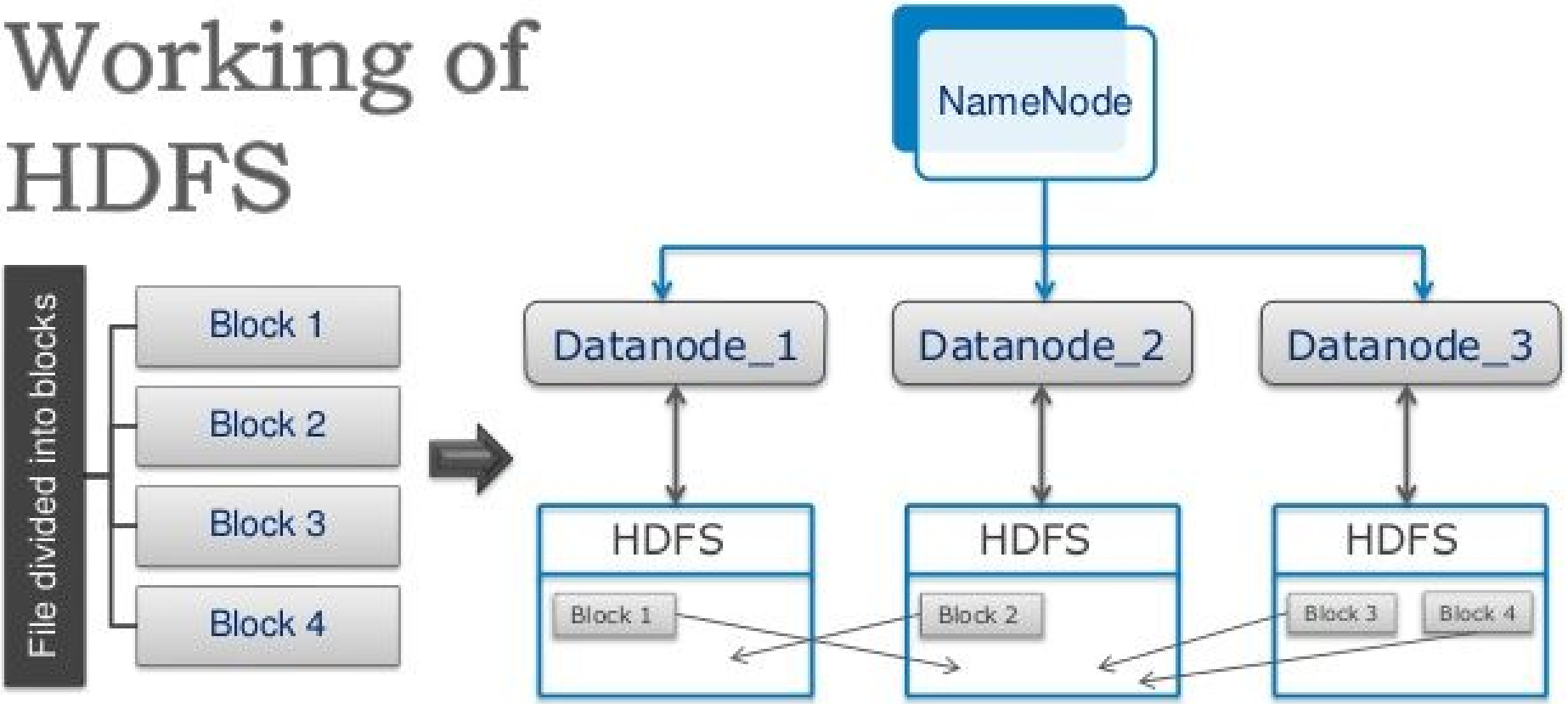


What is map reduce?



File Storage

Working of HDFS



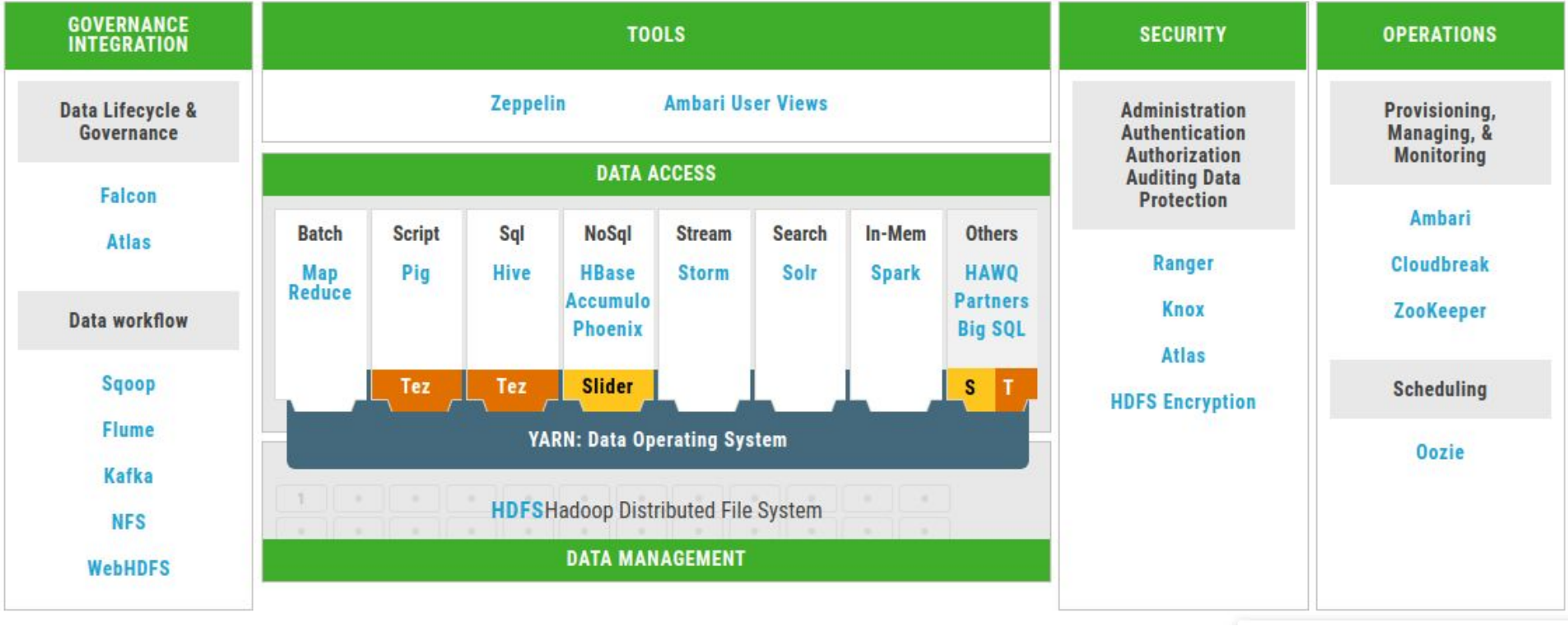
Storage & Replication of Blocks in HDFS

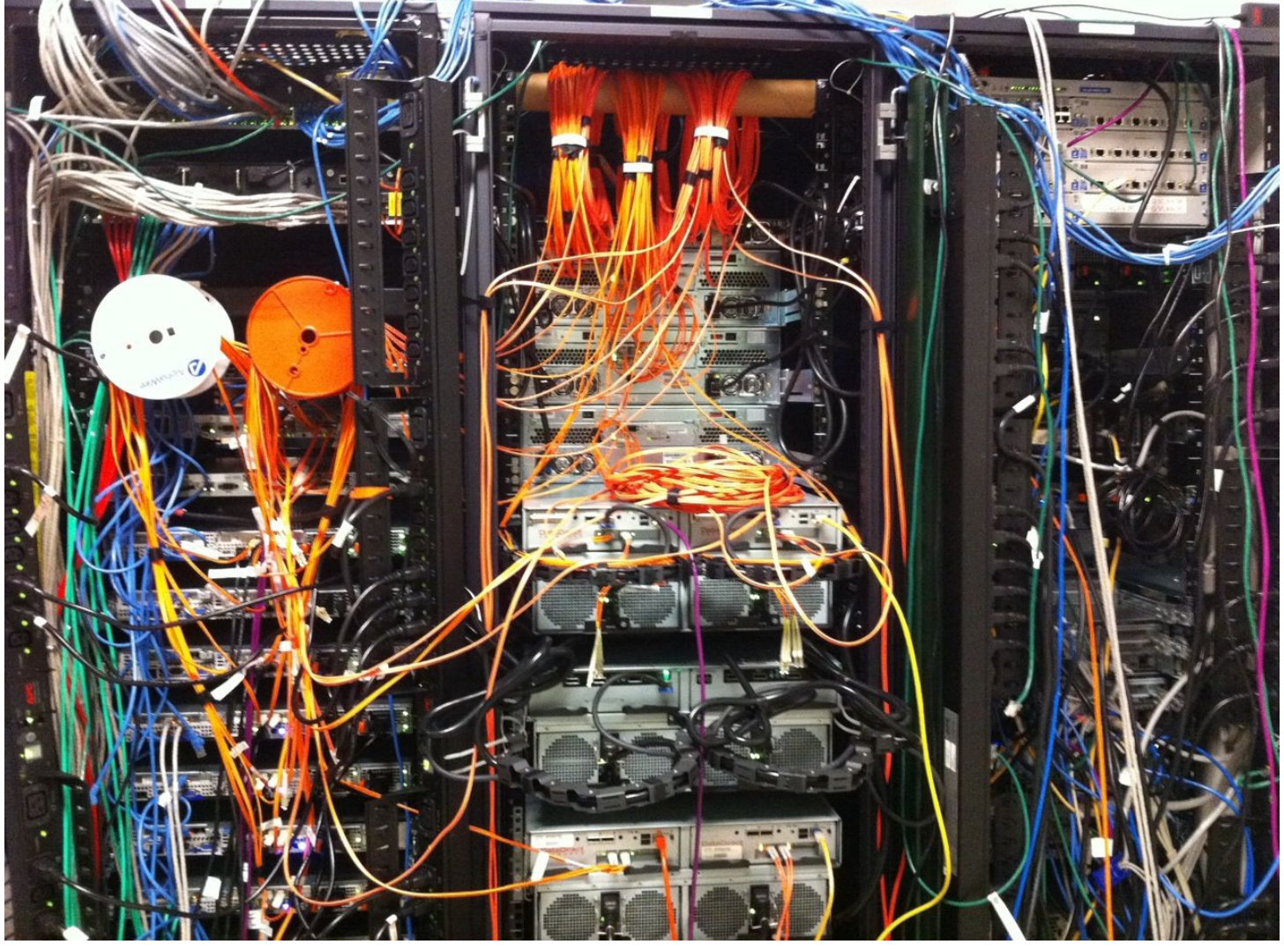
Other storage

Amazon S3 /Azure Blob Storage

- Cheeper
- Resilient and redundant
- Loose data locality
- Eventually consistant, not a “real file system”
- Move is not move… but copy and delete
- Size is limited to your bank account

The rest of the ecosystem





Cloud Resources

Visit the sandbox

- You don’t need to use the cloud if you don’t want to for development/play.
- Hortonworks Sandbox
- One machine VM in Virtual Box, VMWare, Docker

Lets Play with some data...

Demo Time...