Big Data

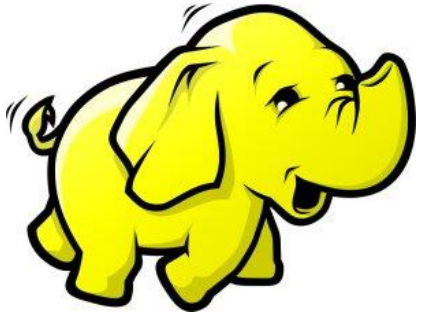
Mohammad Amin Nikbakht

Electrical & Computer Department

Isfahan University Of Technology

Agenda

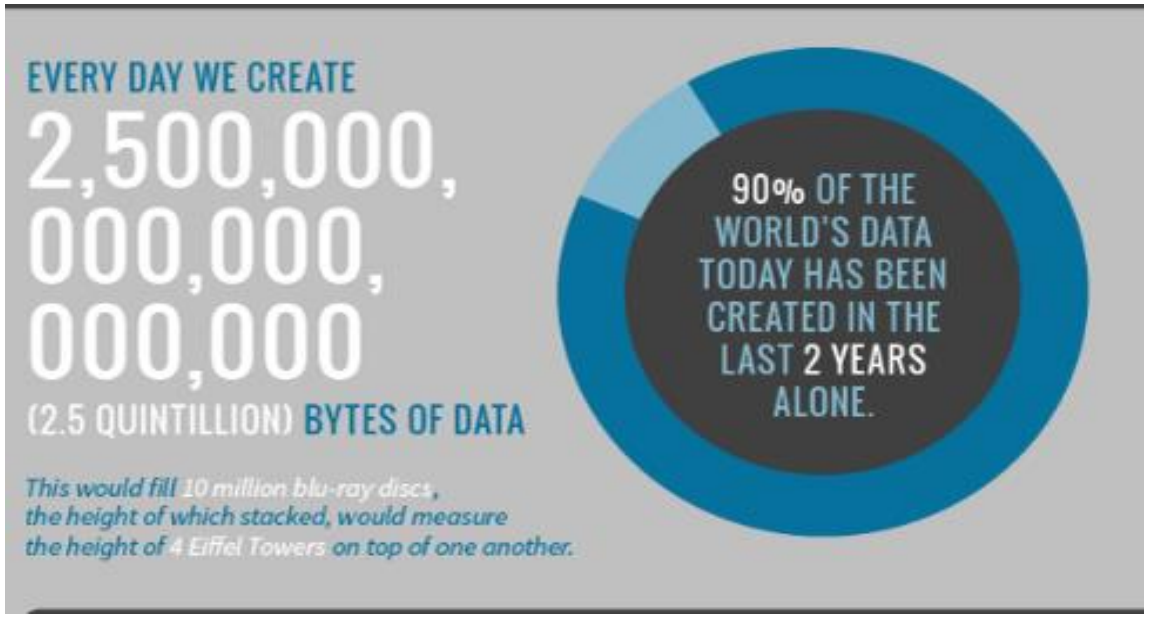
- What is Big Data?
- What is Hadoop?



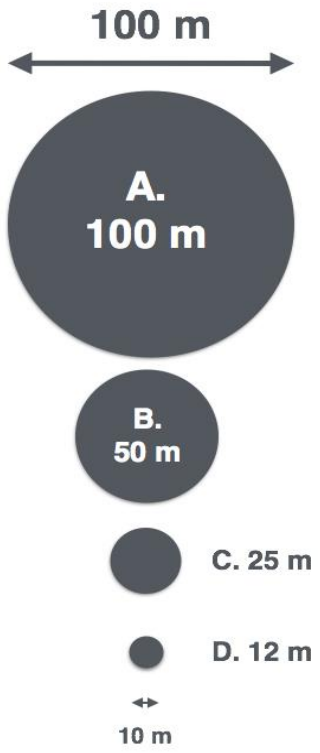
Big Data


EVERYDAY
BIG
DATA
Big data describes the collection of complex and large data sets such that it's difficult to capture, process, store, search and analyze using conventional data base systems. Its uses are shaping the world around us, offering more qualitative insights into our everyday lives.

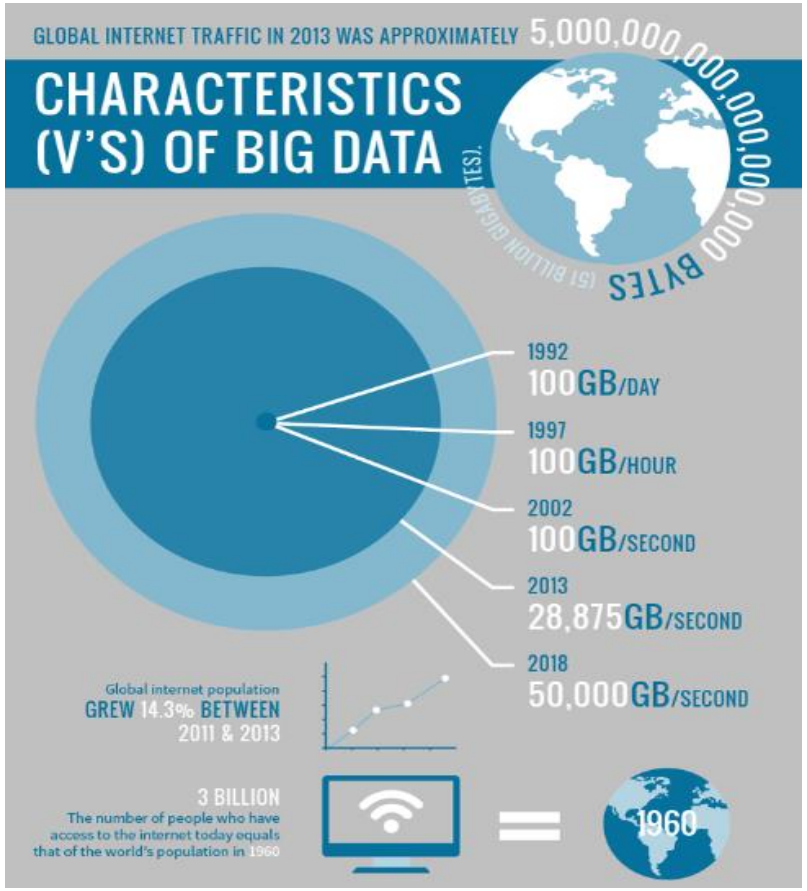
Big Data


EVERY DAY WE CREATE
2,500,000,
000,000,
000,000
(2.5 QUINTILLION) BYTES OF DATA
This would fill 10 million blu-ray discs,
the height of which stacked, would measure
the height of 4 Eiffel Towers on top of one another.
90% OF THE WORLD'S DATA TODAY HAS BEEN CREATED IN THE LAST 2 YEARS ALONE.

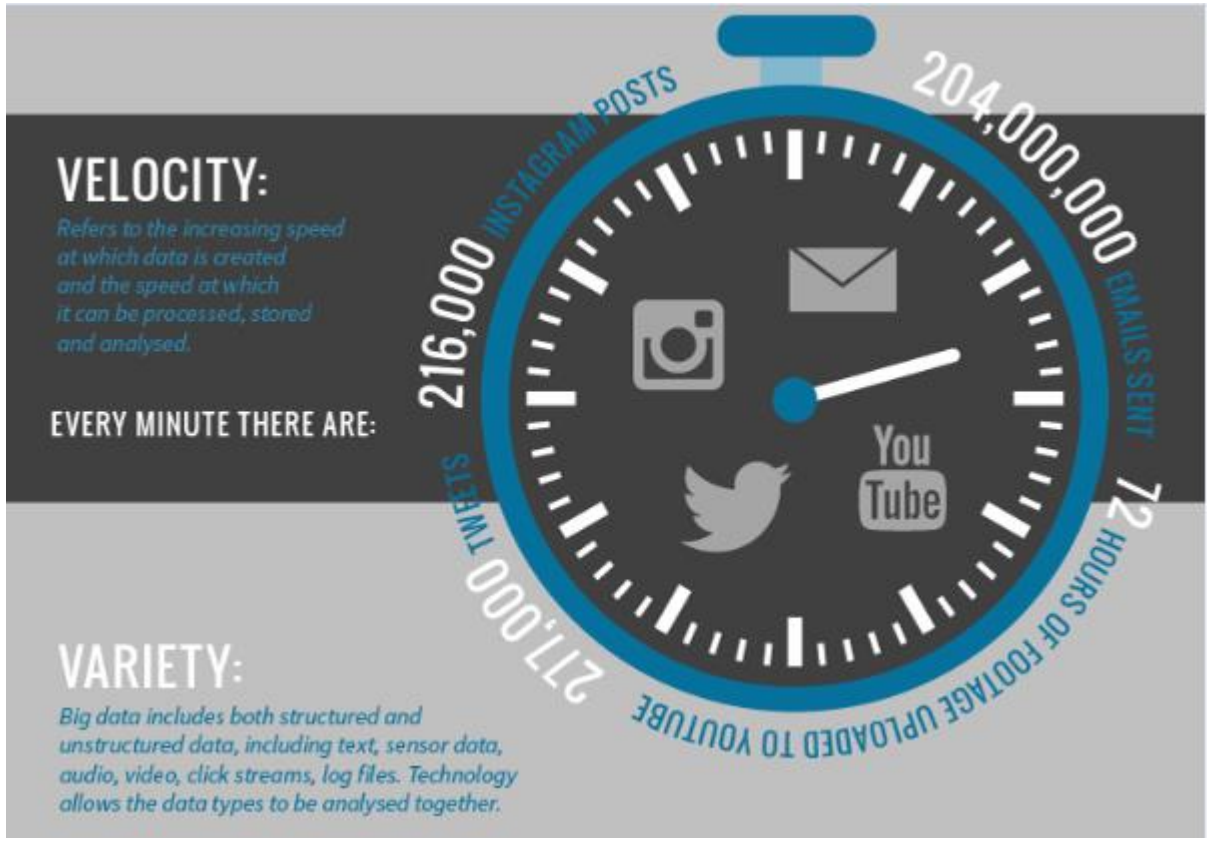
Big Data



Big Data


GLOBAL INTERNET TRAFFIC IN 2013 WAS APPROXIMATELY 5,000,000,000,000,000 BYTES (5 BILLION GIGABYTES).
CHARACTERISTICS (V'S) OF BIG DATA
1992
100GB/DAY
1997
100GB/HOUR
2002
100GB/SECOND
2013
28,875GB/SECOND
2018
50,000GB/SECOND
Global internet population
GREW 14.3% BETWEEN
2011 & 2013
3 BILLION
The number of people who have access to the internet today equals that of the world's population in 1960
1960

Big Data


VELOCITY:
Refers to the increasing speed at which data is created and the speed at which it can be processed, stored and analysed.
EVERY MINUTE THERE ARE:
216,000 INSTAGRAM POSTS
204,000,000 EMAILS SENT
72 HOURS OF FOOTAGE UPLOADED TO YOUTUBE
277,000 TWEETS
You Tube
VARIETY:
Big data includes both structured and unstructured data, including text, sensor data, audio, video, click streams, log files. Technology allows the data types to be analysed together.

Big Data


90%
OF DATA GENERATED IS "UNSTRUCTURED"
This includes tweets, photos, customer purchase history and even customer service call logs.

How does BIG DATA improve our daily lives ?

How does BIG DATA improve our daily lives ?

How does BIG DATA improve our daily lives ?

How does BIG DATA improve our daily lives ?

How does BIG DATA improve our daily lives ?

How does BIG DATA improve our daily lives ?

Hadoop!

Ask bigger questions.

Timeline

- Dec 2004: Dean/Ghemawat (Google) MapReduce paper

- 2005: Doug Cutting and Mike Cafarella (Yahoo) create Hadoop, at first only to extend Nutch

Timeline

- 2006: Yahoo runs Hadoop on 5-20 nodes
- March 2008: Cloudera founded
- July 2008: Hadoop wins TeraByte sort benchmark (1st time a Java program won this competition)
- April 2009: Amazon introduce “Elastic MapReduce” as a service on S3/EC2
- June 2011: Hortonworks founded

Timeline

- 27 dec 2011: Apache Hadoop release 1.0.0
- June 2012: Facebook claim “biggest Hadoop cluster”, totalling more than 100 PetaBytes in HDFS
- 2013: Yahoo runs Hadoop on 42,000 nodes, computing about 500,000 MapReduce jobs per day
- 15 oct 2013: Apache Hadoop release 2.2.0 (YARN)

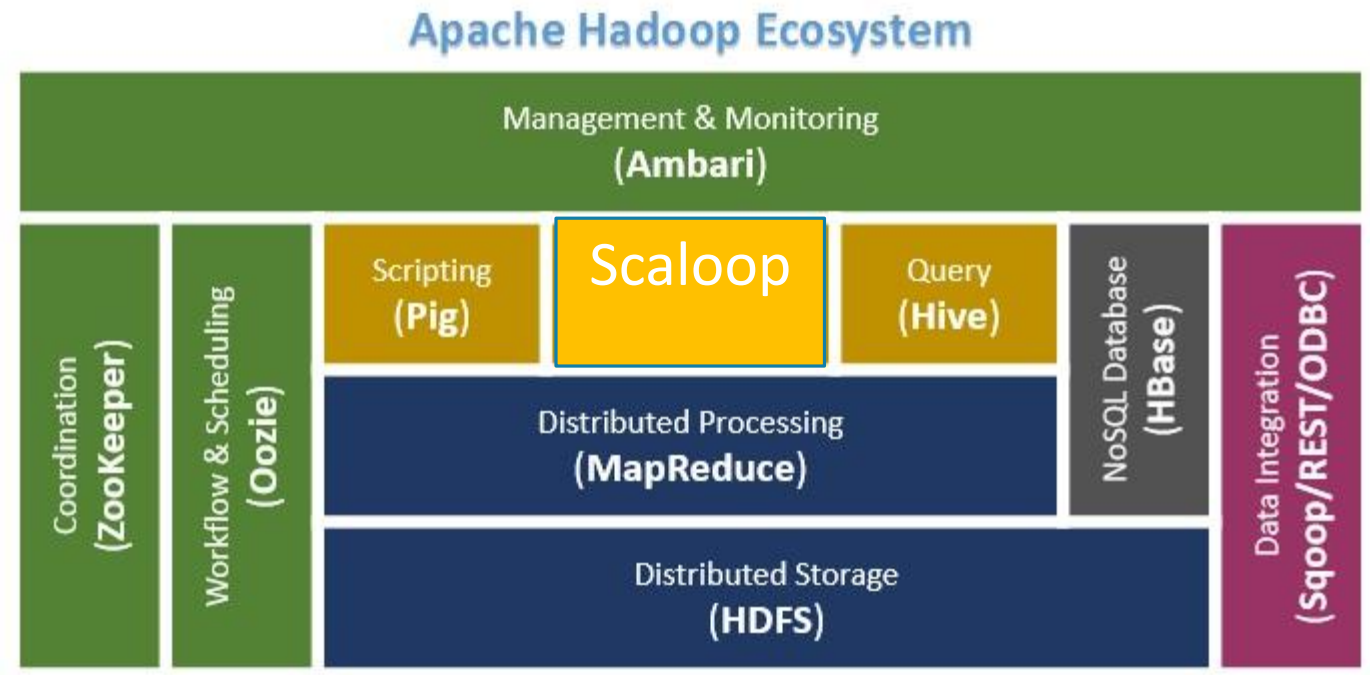
Who uses Hadoop?

What is hadoop?

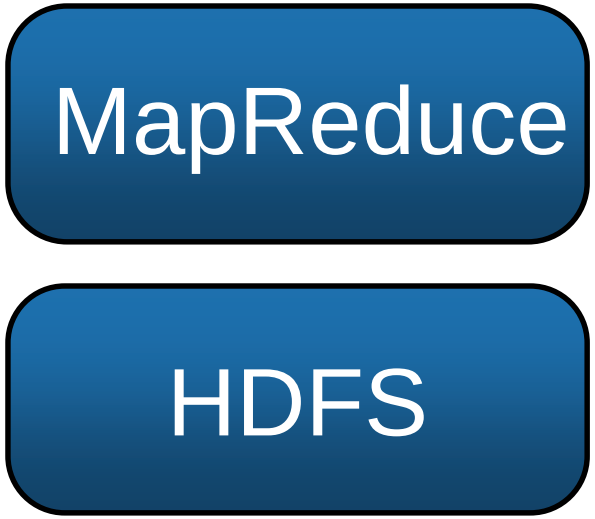
- **Open sourced**, **flexible** and **available** architecture for **large scale** computation and data processing on a network of **commodity hardware**
- Inspired by Google

Hadoop amazing name!

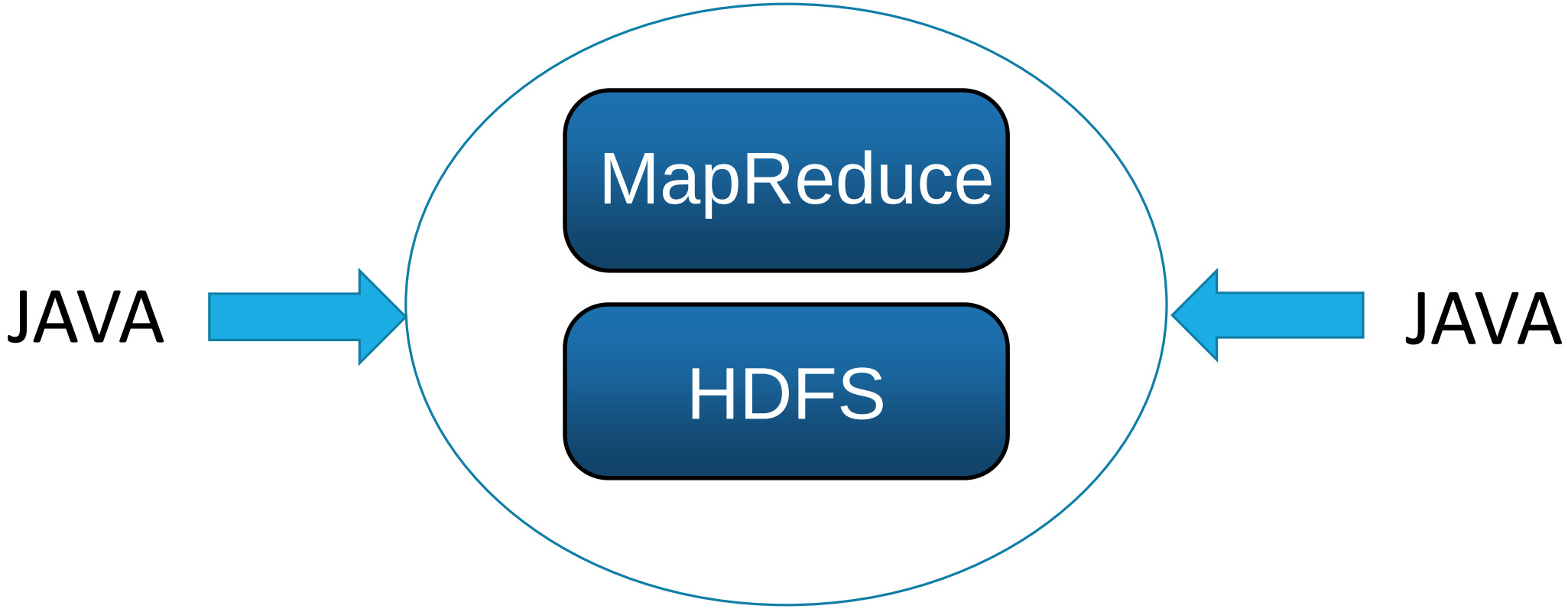
Hadoop Ecosystem!



Hadoop core

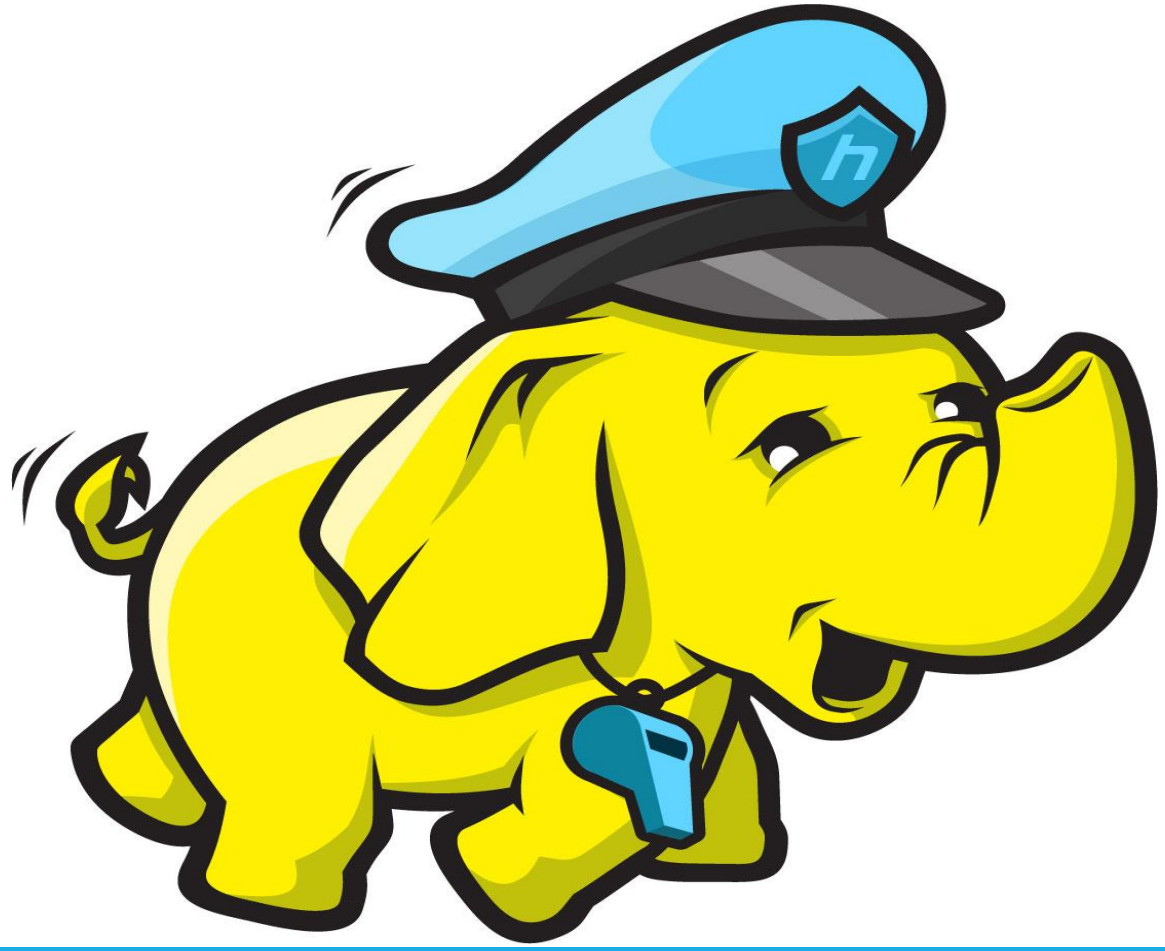


Hadoop core



What is HDFS?

- Hadoop Distributed File System
- Redundancy
- Fault Tolerant
- Scalable
- Self Healing
- Write Once, Read Many Times
- Java API
- Command Line Tool



Map Reduce

- Is a framework for performing high performance distributed data processing using the divide and aggregate programming paradigm.

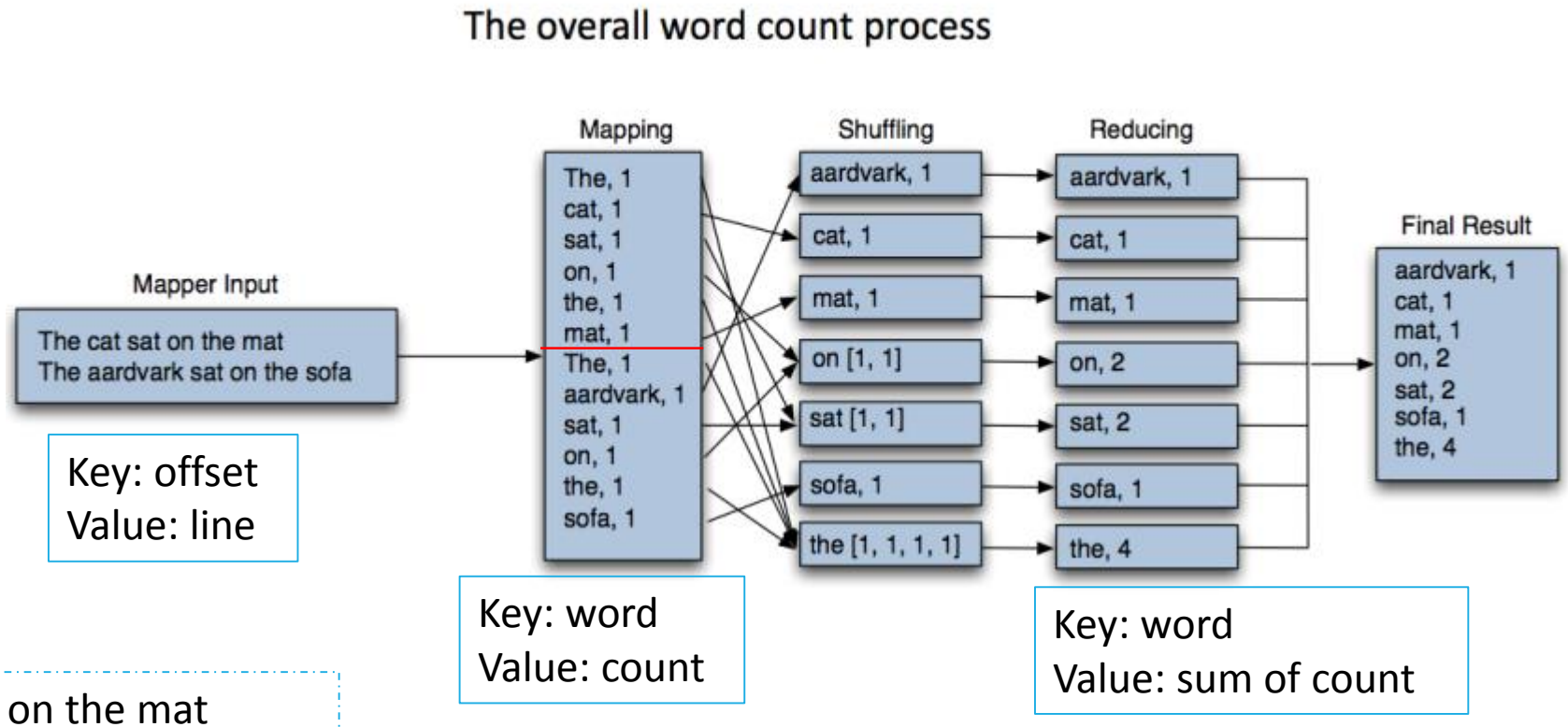
Simple example

- Goal: count the number of books in the library.
- Map:
  You count up shelf #1, I count up shelf #2

  (The more people we get, the faster this part goes)
- Reduce:
  We all get together and add up our individual counts.

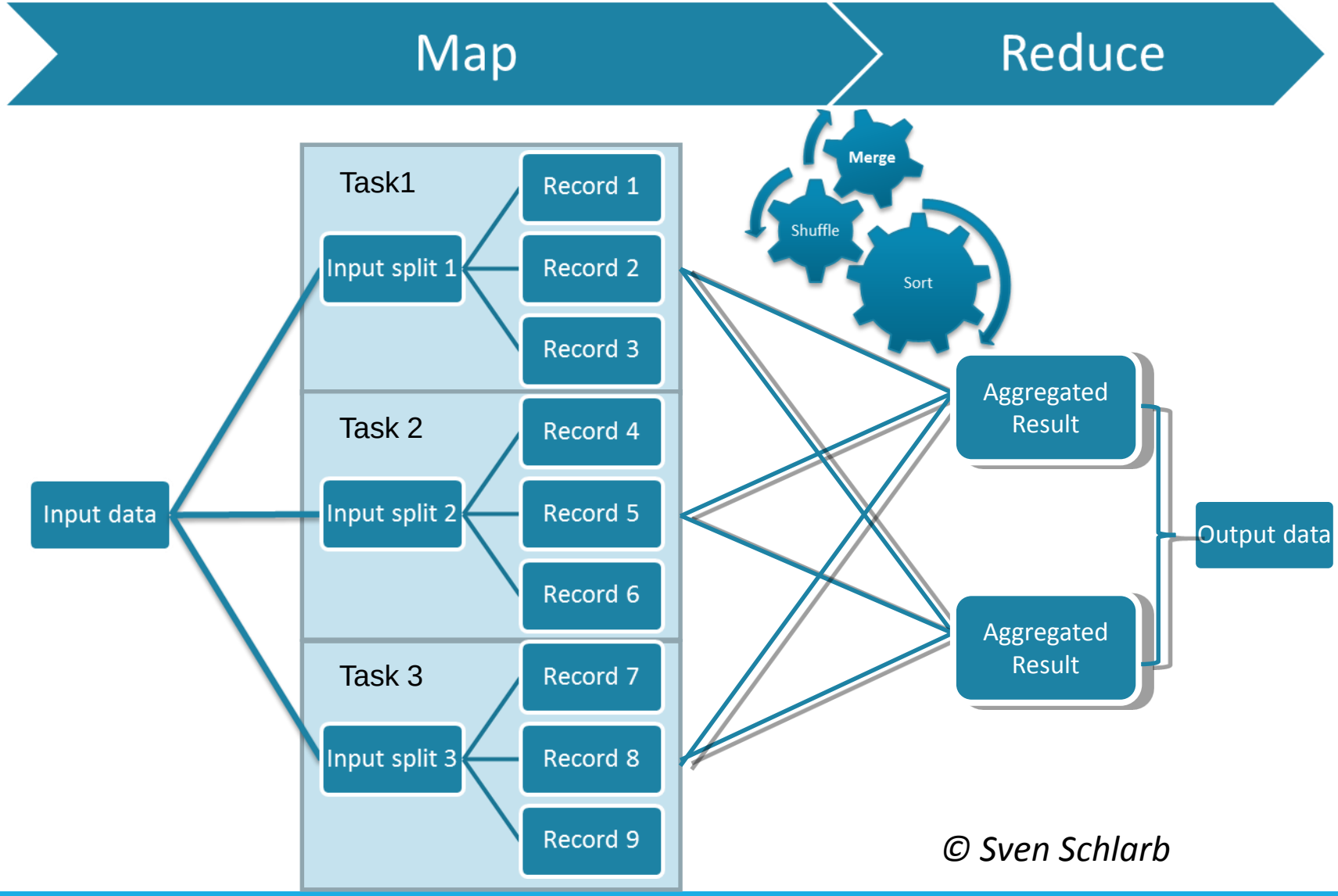
(Cf. http://www.chrisstucchio.com/blog/2011/mapreduce_explained.html)

Routine example



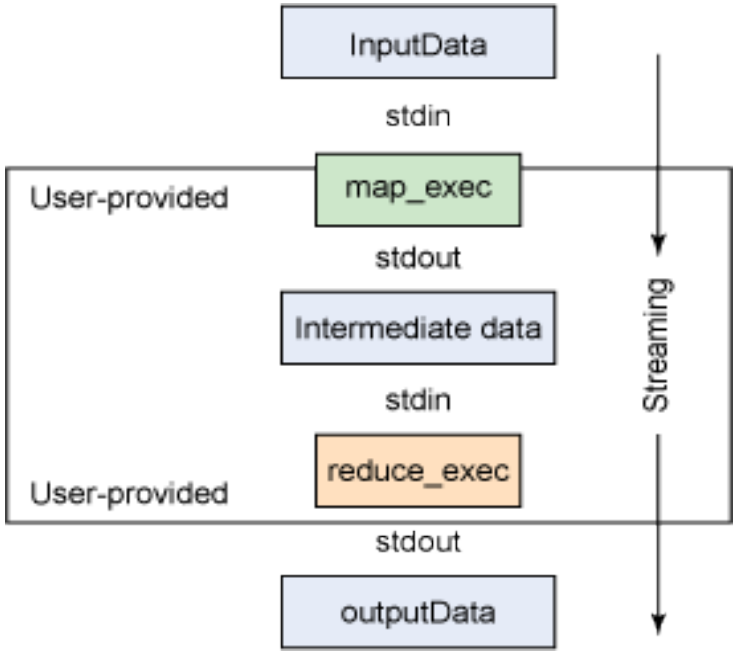
0:The cat sat on the mat
1:The aardvark sat on the sofa

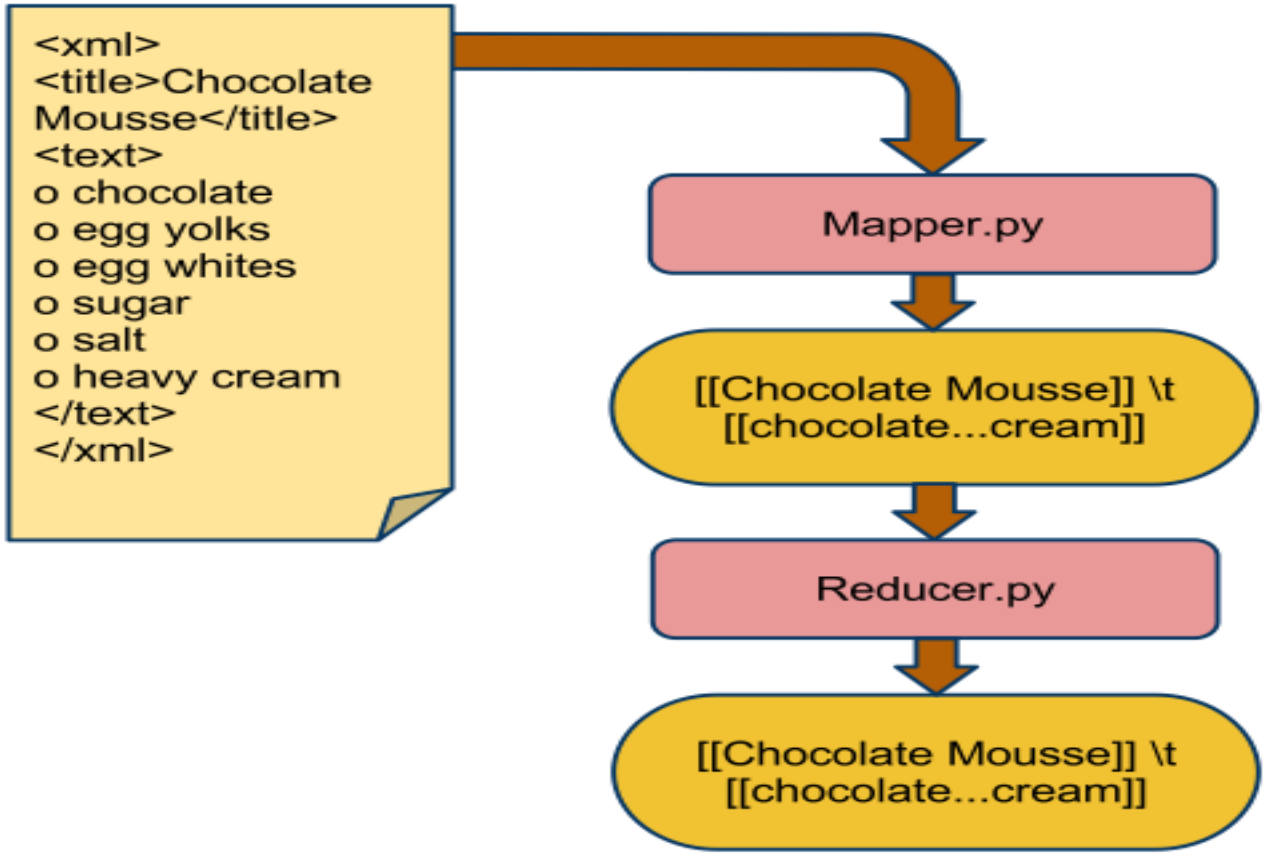
MapReduce in a nutshell



© Sven Schlarb

Hadoop streaming





Exercise

Goal : find total sales of each city!

```
1  2012-01-01  09:00  San Jose  Men's Clothing  214.05  Amex
2  2012-01-01  09:00  Fort Worth  Women's Clothing  153.57  Visa
3  2012-01-01  09:00  San Diego  Music  66.08  Cash
4  2012-01-01  09:00  Pittsburgh  Pet Supplies  493.51  Discover
5  2012-01-01  09:00  Omaha  Children's Clothing 235.63  MasterCard
6  2012-01-01  09:00  Stockton  Men's Clothing  247.18  MasterCard
7  2012-01-01  09:00  Austin  Cameras 379.6  Visa
8  2012-01-01  09:00  New York  Consumer Electronics  296.8  Cash
9  2012-01-01  09:00  Corpus Christi  Toys  25.38  Discover
10 2012-01-01  09:00  Fort Worth  Toys  213.88  Visa
11 2012-01-01  09:00  Las Vegas  Video Games 53.26  Visa
12 2012-01-01  09:00  Newark  Video Games 39.75  Cash
13 2012-01-01  09:00  Austin  Cameras 469.63  MasterCard
14 2012-01-01  09:00  Greensboro  DVDs  290.82  MasterCard
15 2012-01-01  09:00  San Francisco  Music  260.65  Discover
16 2012-01-01  09:00  Lincoln Garden  136.9  Visa
17 2012-01-01  09:00  Buffalo Women's Clothing  483.82  Visa
18 2012-01-01  09:00  San Jose  Women's Clothing  215.82  Cash
```

Mapper

```
1  #!/usr/bin/env python
2  import sys
3
4  def mapper():
5      for line in sys.stdin:
6          data = line.strip().split("\t")
7              if len(data)==6:
8              date,time,store,item,cost,payment = data
9              print "{0}\t{1}".format(store,cost)
10
11
12 mapper()
13
```

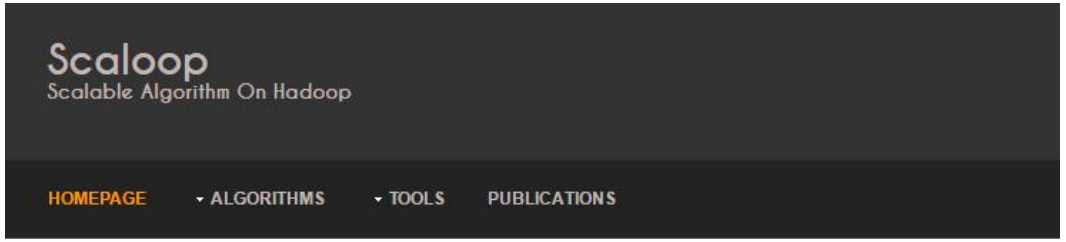
Reducer

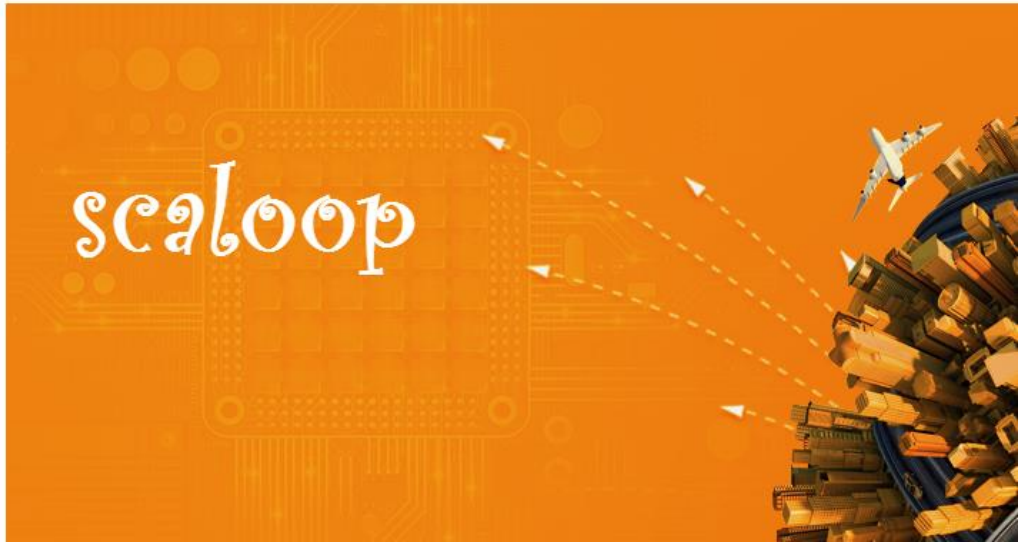
```
#!/usr/bin/env python
import sys

def reducer():
    salesTotal =0
        oldKey = None
    for line in sys.stdin:
            data = line.strip().split("\t")
        if len(data) != 2:
            continue
        thisKey,thisSale =data
        if oldKey and oldKey !=thisKey:
             print "{0}\t{1}".format(oldKey,salesTotal)
             salesTotal =0
        oldKey = thisKey
        salesTotal +=float(thisSale)


reducer()
```

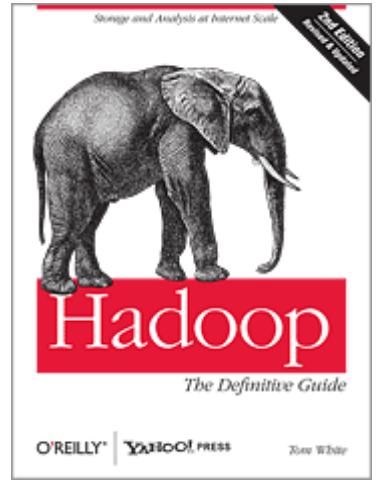
Scaloop Story





Resources

- Tom White: Hadoop. The Definitive Guide
- Udacity Big Data course
- www.vcloudnews.com



Questions?