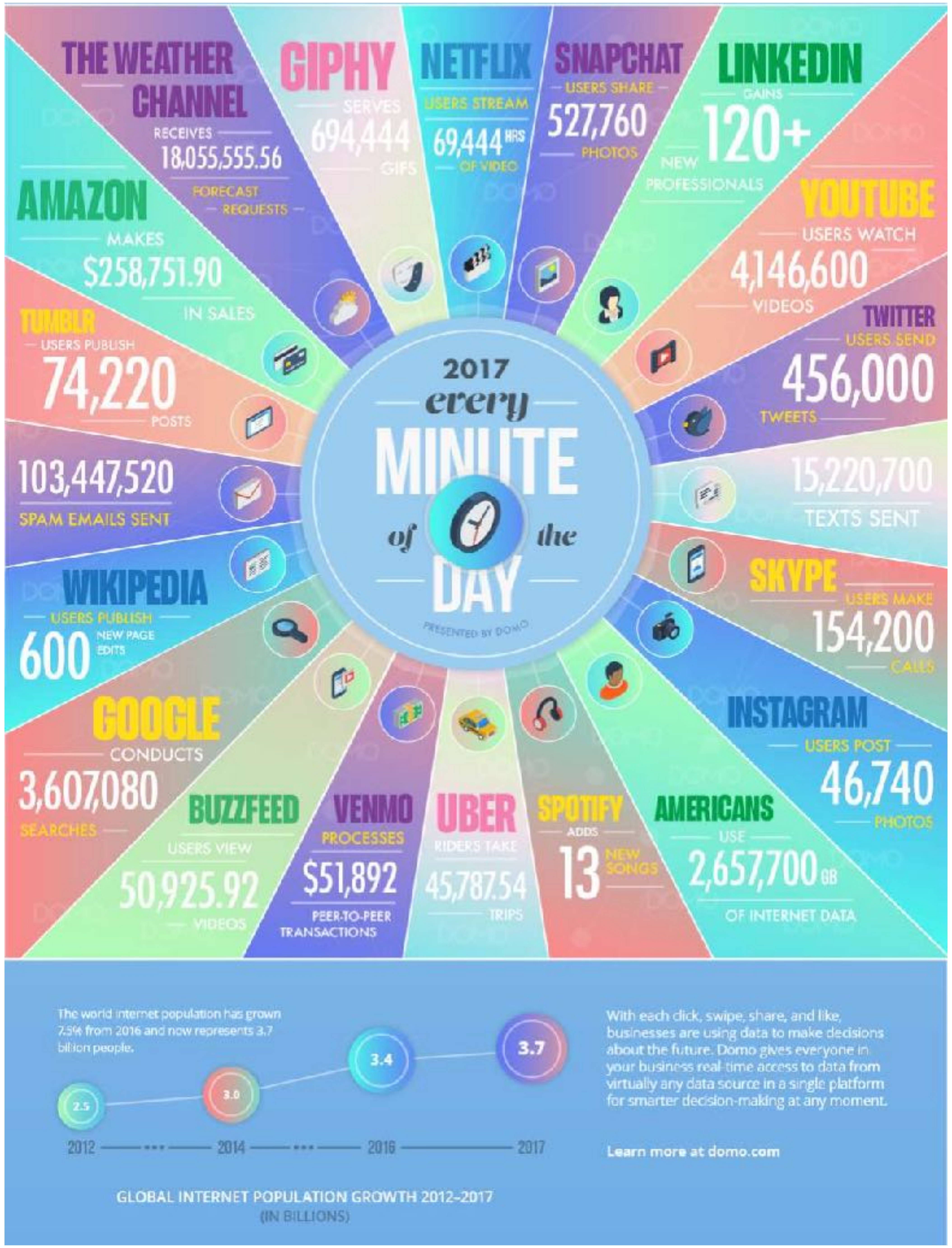
ThoughtWorks®

BIG DATA

ThoughtWorks®

Why is it needed

- Ninety percent of the data in the world today has been created in the last two years alone
- As a whole, the Internet population has grown by 7.5 percent since 2016 and now includes over 3.7 billion humans
- On average, the US alone spits out 2,657,700 gigabytes of Internet data every minute
- Uber taking 45,787 trips each minute, Spotify adding 13 new songs, we tweet 456,000 times, post 46,740 Instagram photos, Google 3.6 million searches, and publish 600 new page edits on Wikipedia each minute. The Internet also copes with 103,447,520 spam emails every minute.


2017 every MINUTE of the DAY
PRESENTED BY DOMO
THE WEATHER CHANNEL RECEIVES 18,055,555.56 FORECAST REQUESTS
GIPHY SERVES 694,444 GIFS
NETFLIX USERS STREAM 69,444 HRS OF VIDEO
SNAPCHAT USERS SHARE 527,760 PHOTOS
LINKEDIN GAINS 120+ NEW PROFESSIONALS
AMAZON MAKES $258,751.90 IN SALES
YOUTUBE USERS WATCH 4,146,600 VIDEOS
TUMBLR USERS PUBLISH 74,220 POSTS
TWITTER USERS SEND 456,000 TWEETS
103,447,520 SPAM EMAILS SENT
15,220,700 TEXTS SENT
WIKIPEDIA USERS PUBLISH 600 NEW PAGE EDITS
SKYPE USERS MAKE 154,200 CALLS
GOOGLE CONDUCTS 3,607,080 SEARCHES
INSTAGRAM USERS POST 46,740 PHOTOS
BUZZFEED USERS VIEW 50,925.92 VIDEOS
VENMO PROCESSES $51,892 PEER-TO-PEER TRANSACTIONS
UBER RIDERS TAKE 45,787.54 TRIPS
SPOTIFY ADDS 13 NEW SONGS
AMERICANS USE 2,657,700 GB OF INTERNET DATA
2.5
3.0
3.4
3.7
2012
2014
2016
2017
GLOBAL INTERNET POPULATION GROWTH 2012–2017 (IN BILLIONS)
Learn more at domo.com

The world internet population has grown 7.5% from 2016 and now represents 3.7 billion people.

With each click, swipe, share, and like, businesses are using data to make decisions about the future. Domo gives everyone in your business real-time access to data from virtually any data source in a single platform for smarter decision-making at any moment.

Courtesy-: http://www.iflscience.com/technology/how-much-data-does-the-world-generate-every-minute/

Why is it needed?

- Need to process lot of data in batch/live
- Rate of growth of processing speed and memory has slowed down
- Horizontally scaled systems are cheaper to use
- Computationally complex queries - companies no longer do just the old and simple show and tell

Let’s take an example

Why the sudden interest - what were people doing before?

Simple web analytics application - want to list the number of page views per customer and the top 100 URLs by number of page views?

Schema

id: Int

user_id: Int

url: varchar(255)

pageviews: bigInt

Why is it needed?

You hit Black Friday - page views are off the charts, database is the contention, how do you fix?

1.Batched writes through queues

2.Sharding

3.Tackling fault tolerance

Increments for unavailable shards on a pending queue

Replication - Read Slave

4.Corruption/ Resilience

Cyber Monday begins - and you are doomed!

Why is it needed?

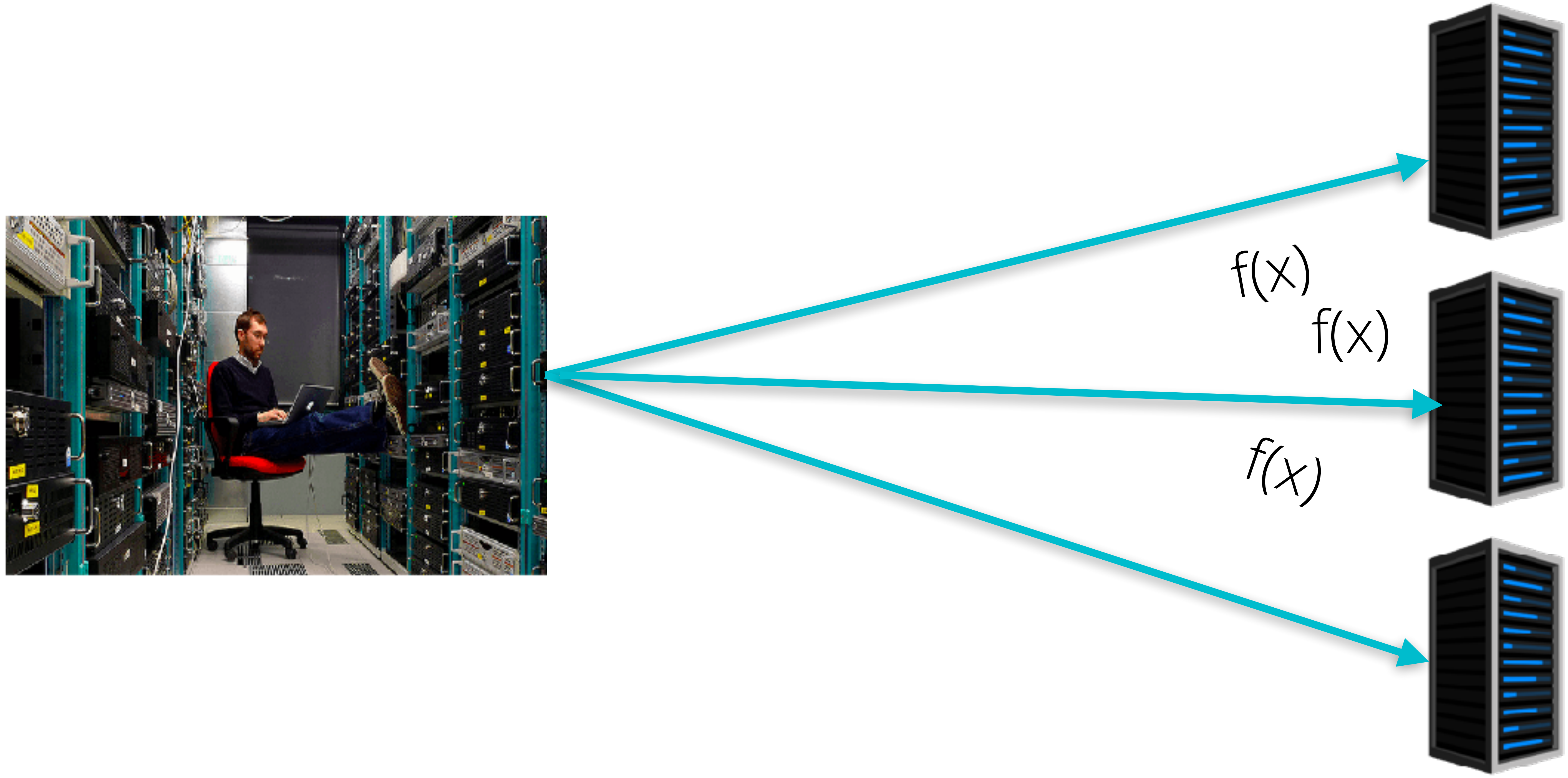
Does big data then just only apply to problems where the quantity of data is huge?

- No but thats how it all began

BIG DATA PARADIGM

- Some rules and design ideas that limit how you process your data
- In return you get
  - Computational Reasoning
  - Efficiency dealing with large/unstructured/streaming data sets
  - Better operations and maintainability
  - Fault Tolerance
  - Resilience
  - Debugging ability

BIG DATA PARADIGM



THE 4 V'S OF BIG DATA

- Volume -> Petabytes of data - Queries/Statistical Analysis/Batch Data - Analyze billions of taxi rides
- Velocity -> Streaming Applications - Twitter/Facebook trends every minute/Swarm health/Real time trading
- Variety -> Different data formats - Videos/Tweets/Facebook comments
- Veracity/Quality -> Reliability/life time of data/Context/Data cleaning

THE 4 V'S OF BIG DATA

- 2016 US Elections - Big data and social media used heavily

Some interesting stats-:

https://medium.com/google-cloud/big-data-and-the-elections-2016-5bd53dda2315

And why it failed?

http://www.techrepublic.com/article/election-tech-lies-damned-lies-and-statistics/ - Data veracity

- Tesla

https://evannex.com/blogs/news/ai-and-autopilot-how-tesla-is-winning-the-race-to-achieve-vehicle-autonomy-infographic

https://www.inc.com/kevin-j-ryan/how-tesla-is-using-ai-to-make-self-driving-cars-smarter.html

- Google Maps

https://www.ncta.com/platform/broadband-internet/how-google-tracks-traffic/

- Trend Analysis
  Facebook + Twitter + Youtube
  https://trends.google.com/trends/

- Predicting Crop Yields

https://www.theverge.com/2016/8/4/12369494/descartes-artificial-intelligence-crop-predictions-usda

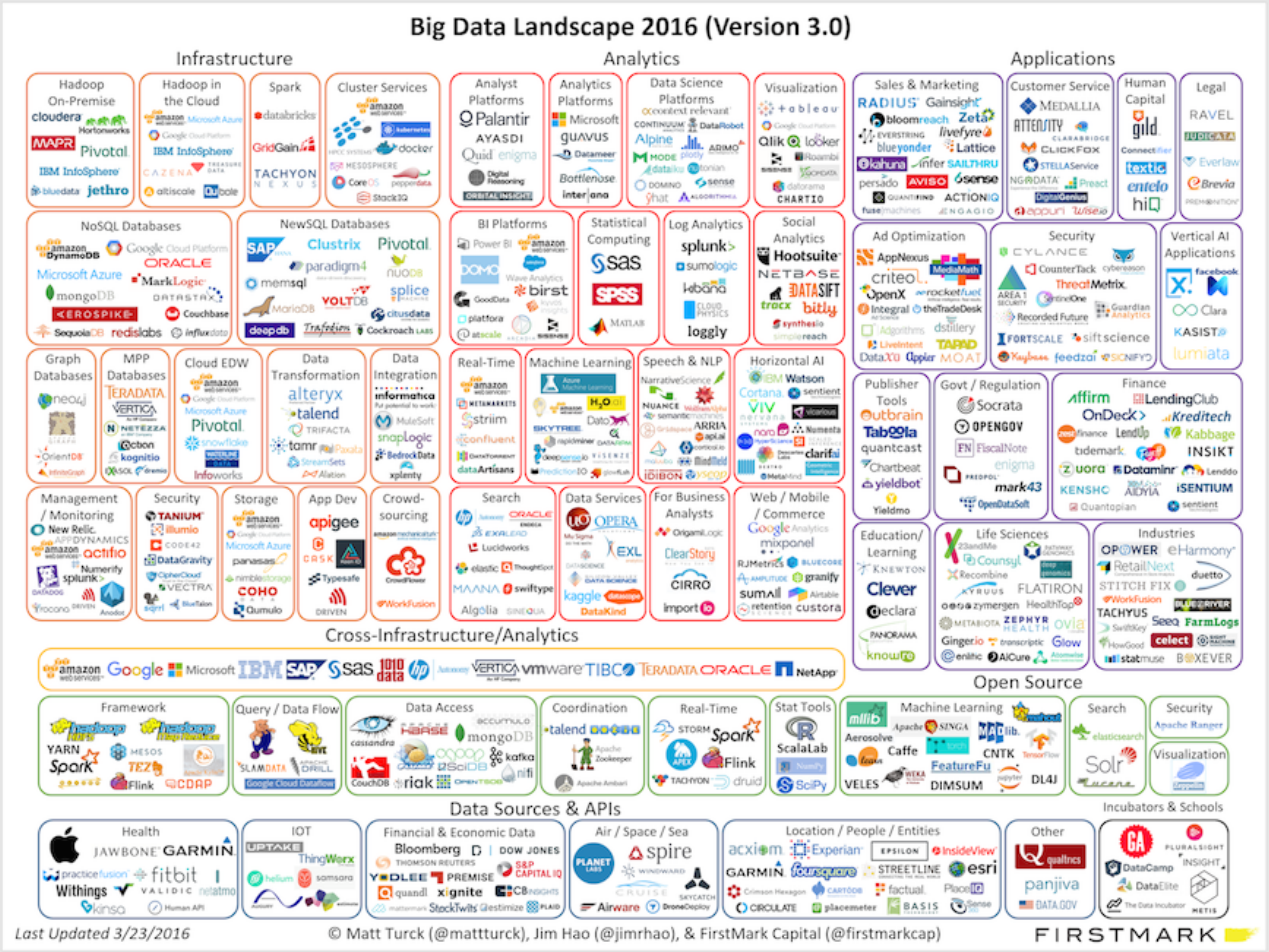
Why is it needed?

- Huge amounts/sources/formats of data being generated.
- Modern hardware hasn't been able to keep up with this huge burst
- Trend towards horizontal scaling rather than vertical scaling.Traditional Databases can't keep up.
- There has also been a trend of moving from simple web apps to more complicated and interactive websites
- Metadata has evolved over the years and is now almost of equal importance as the data itself in some cases
- The rise of Data Analytics and Machine Learning over the years

WARNING: BIG TECH AHEAD

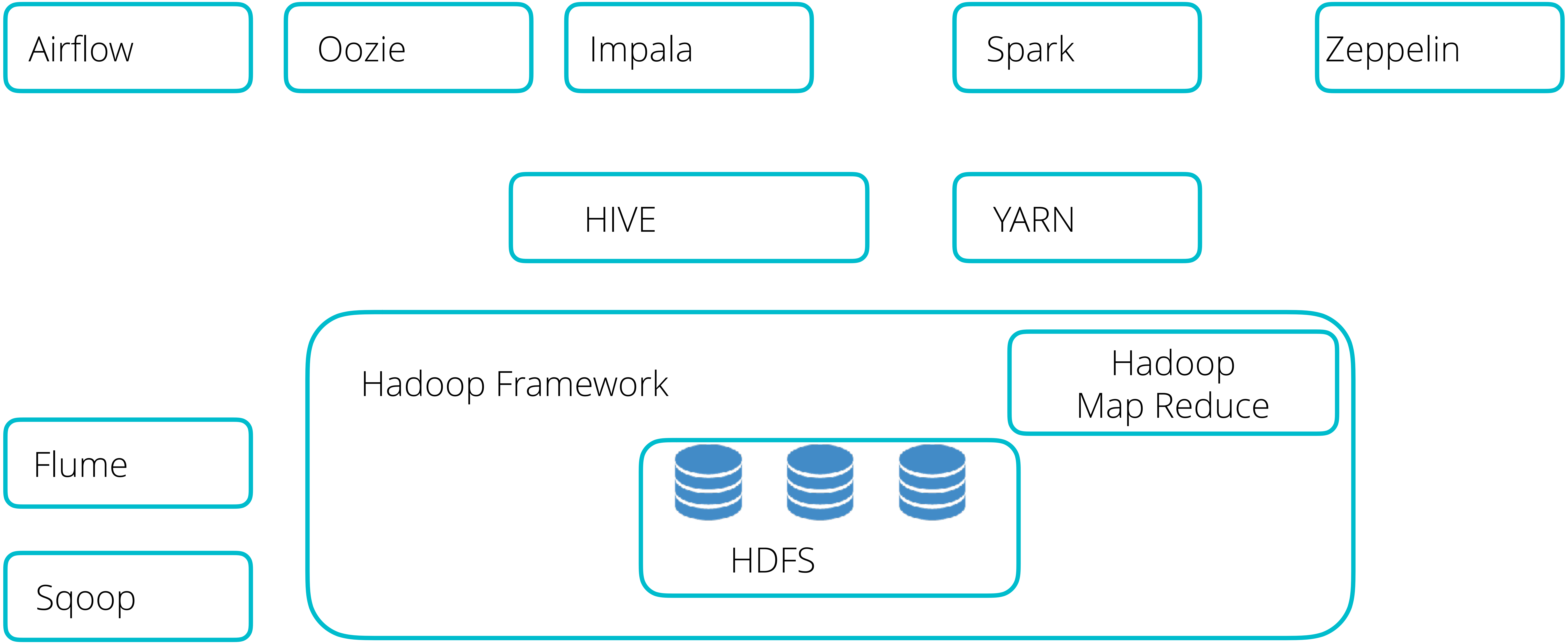
BREAK ADVISORY

BIG DATA ECOSYSTEM



BIG DATA ECOSYSTEM - OUR VERSION

Apache Hadoop V. 3.0



Airflow

Oozie

Impala

Spark

Zeppelin

HIVE

YARN

Hadoop Framework

Hadoop
Map Reduce

Flume

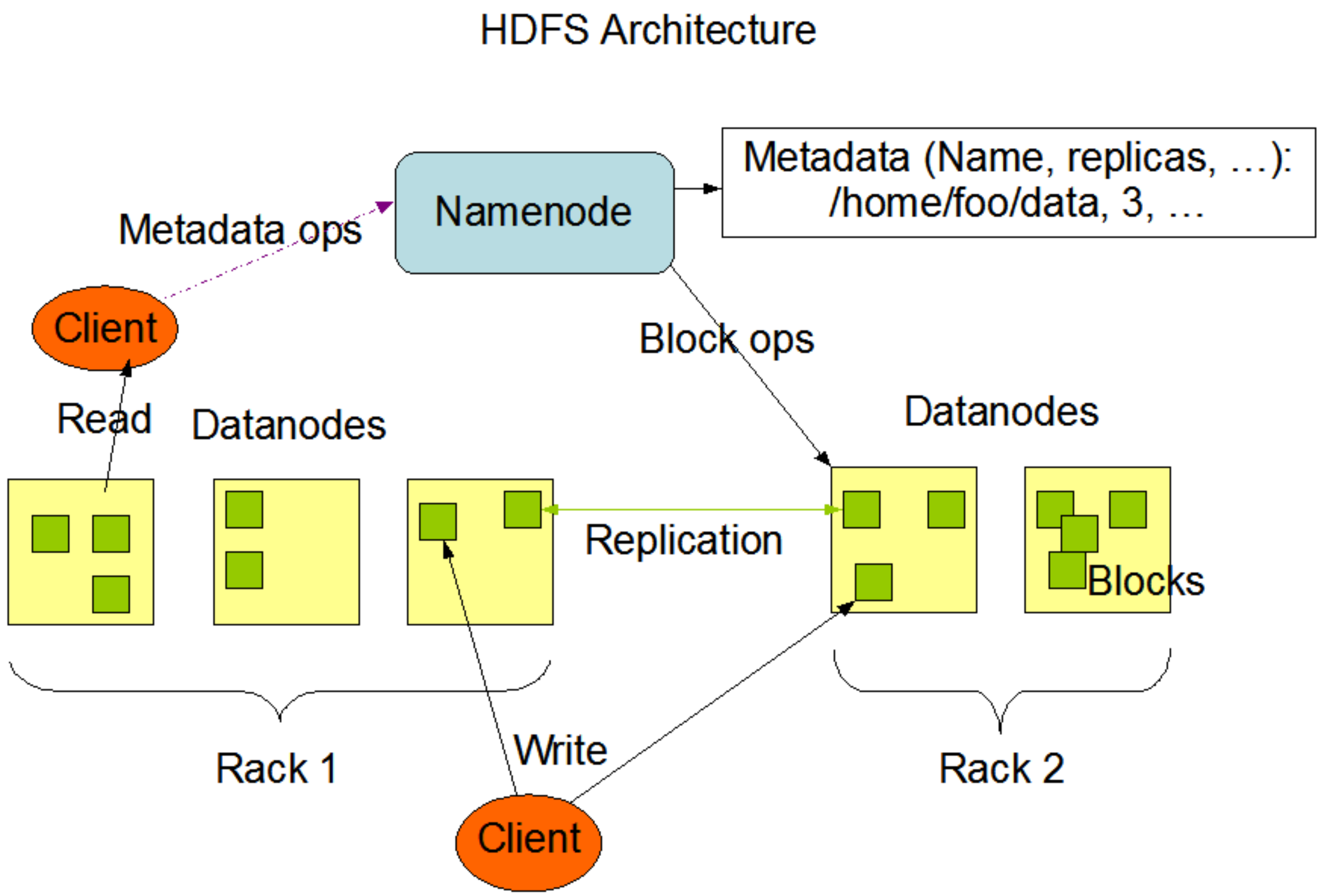
HDFS

Sqoop

HDFS

- Distributed FS based on Google File System (GFS)
- Well suited for commodity hardware
- Built around the idea of “Write Once, Read multiple times”
- Reliability through replication

HDFS

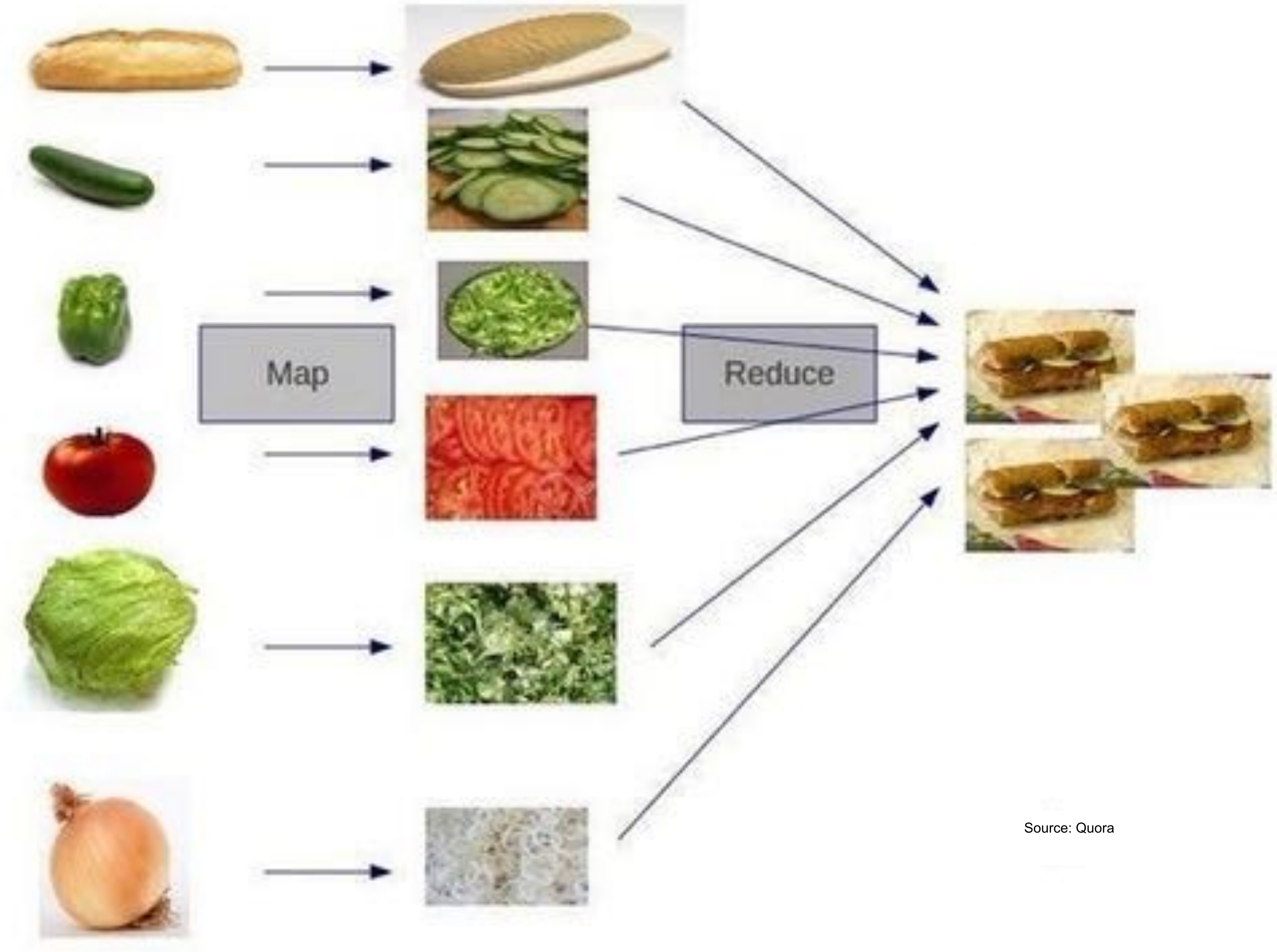

HDFS Architecture
Namenode
Metadata (Name, replicas, …):
/home/foo/data, 3, …
Metadata ops
Client
Block ops
Read
Datanodes
Datanodes
Replication
Blocks
Rack 1
Write
Rack 2
Client

MAP REDUCE

- Massively Parallel programming paradigm
- Fault Tolerant
- Designed to run on clusters of commodity hardware
- Provides high level of abstraction to developers

COOL READ-:
https://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf

MAP REDUCE

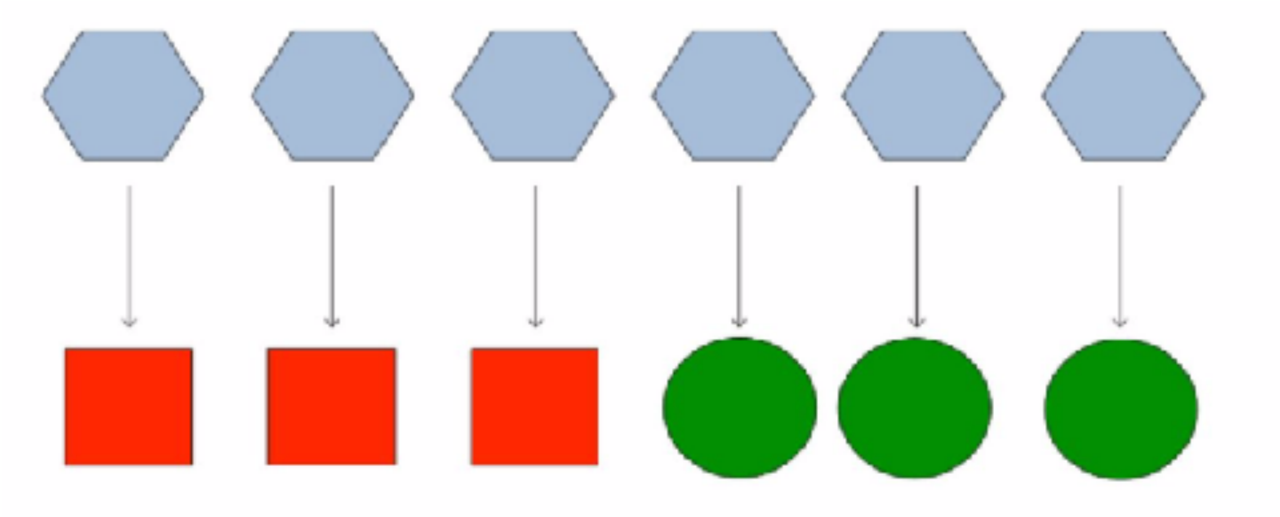


Source: Quora

MAP REDUCE

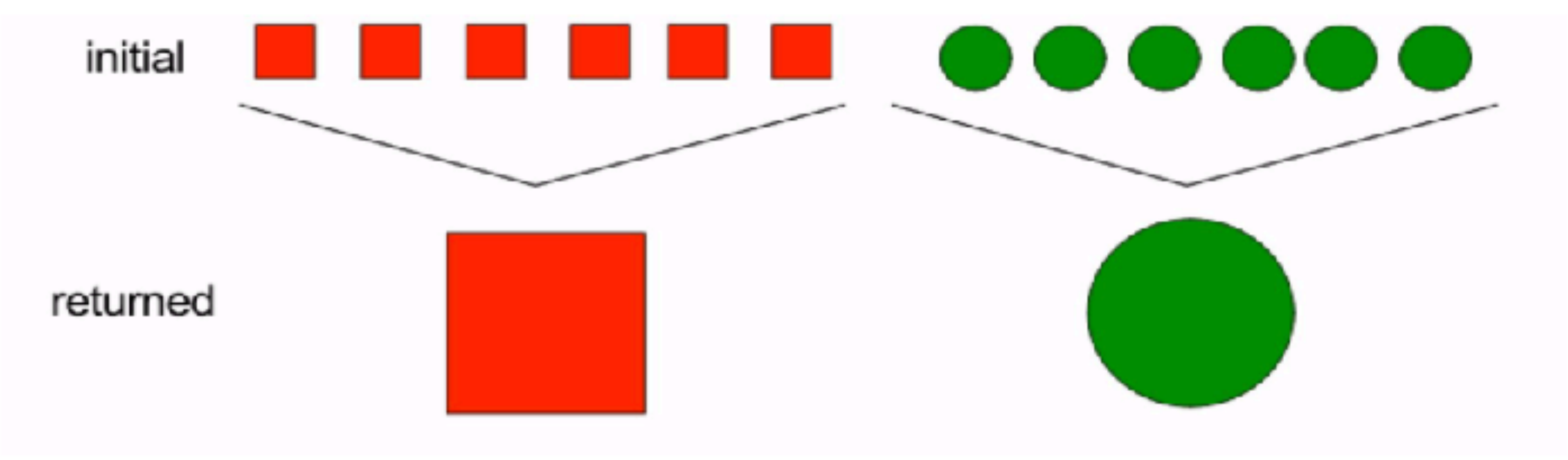
Map

map(K1, V1) -> (K2, V2)



Reduce

reduce(K2, list(V2) -> list(K3, V3)



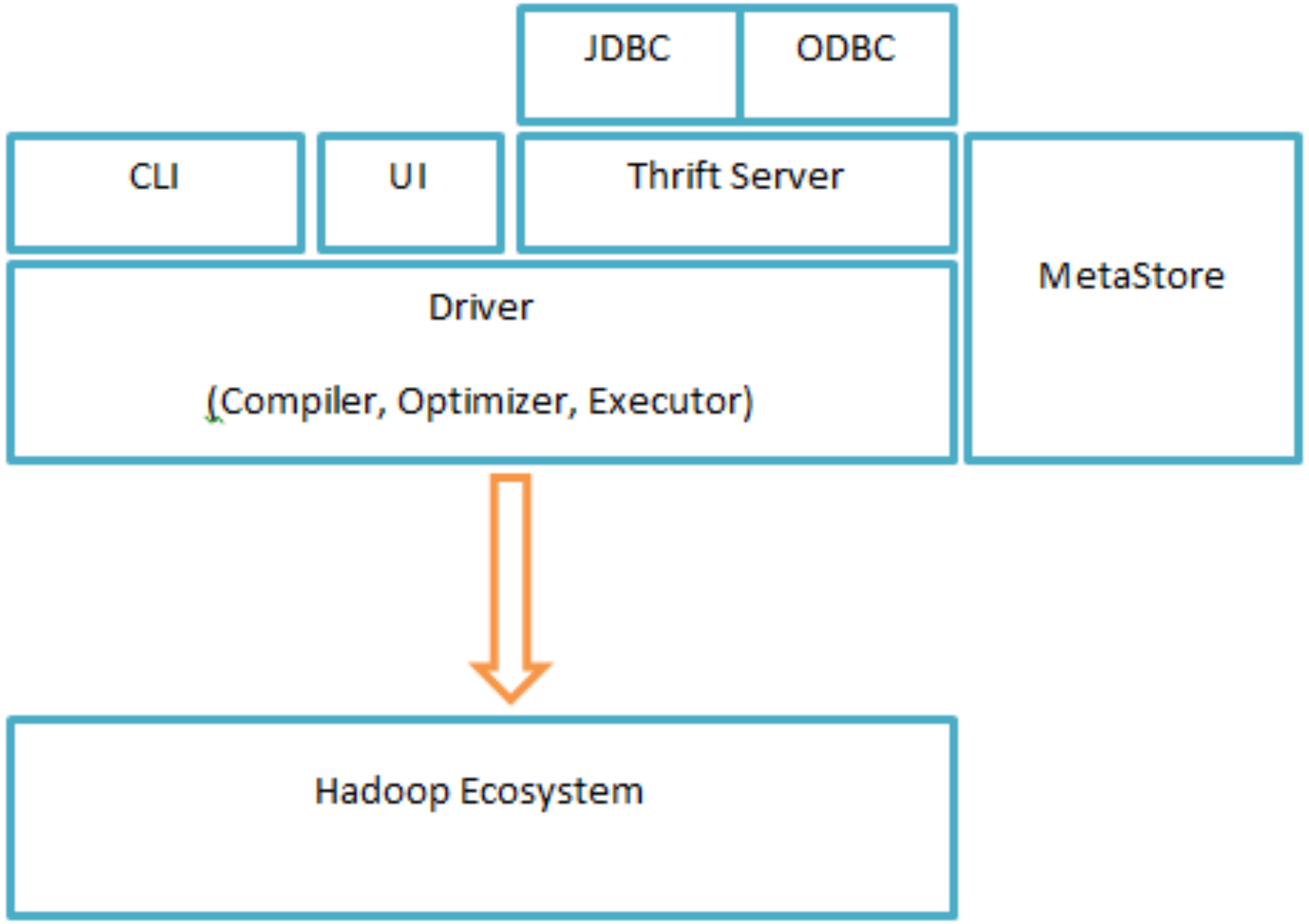
HADOOP

- Open source framework
- Used for storage and large scale processing of data sets
- Mainly consists of the following two modules
  - HDFS
  - Map Reduce

HIVE

- SQL like abstraction over map reduce
- Works for structured data types
- Most commonly used QL engine
- In built optimizations for queries
- Uses beeline and hive client - also available over ODBC
- WORM - Write Once Read Many (NO UPDATES ALLOWED!!!!!)

HIVE



JDBC

ODBC

CLI

UI

Thrift Server

MetaStore

Driver

(Compiler, Optimizer, Executor)

Hadoop Ecosystem

THANK YOU

ThoughtWorks®