Welcome to

Big Data & Hadoop

An Introductory Session

Please introduce yourselves using Questions Window while others are joining us.

WELCOME TO SESSION 1

- **Session - 3 hours Duration**
    - First Half: Understanding Big Data
    - 10 mins. break
    - Second Half: Hadoop Architecture
- Session is being recorded & Recording & presentation will be shared
- **Asking Questions?**
    - Every one except Instructor is muted
    - Please ask questions by typing in Questions Window
    - Instructor will read out the questions before answering
    - To get better answers, keep your messages short and avoid chat language
- This is Session 1 out of 12 sessions on Big Data & Hadoop course. It suffices as an intro to Big Data Tech.

WELCOME - KNOWBIGDATA

- Expert Instructors
- CloudxLab
- Lifetime access to LMS
  - Presentations
  - Class Recording
  - Assignments + Quizzes
  - Project Work
- Real Life Project
- Course Completion Certificate
- 24x7 support
- KnowBigData - Alumni
  - Jobs
  - Stay Abreast (Updated Content, Complimentary Sessions)
  - Stay Connected

COURSE CONTENT

I	Understanding BigData, Hadoop Architecture
II	Cluster Setup, ETL, Project Environment
III	MapReduce framework
IV	Adv MapReduce & Testing
V	Analytics using Pig
VI	Hive
VII	NoSQL & HBase
VIII	ZooKeeper, Flume
IX	Sqoop, Oozie
X	Spark
XI	Storm, Mahout
XII	Comparisons of No SQLs, Project Assignment

About Instructor?

2015	**CloudxLab**	A big data platform.
2014	**KnowBigData**	Founded
2014	**Amazon**	Built High Throughput Systems for Amazon.com site using in-house NoSql.
2012		
2012	**InMobi**	Built Recommender that churns 200 TB
2011		
	tBits Global	Founded tBits Global Built an enterprise grade Document Management System
2006		
	D.E.Shaw	Built the big data systems before the term was coined
2002		
2002	**IIT Roorkee**	Finished B.Tech.

TODAY'S CLASS

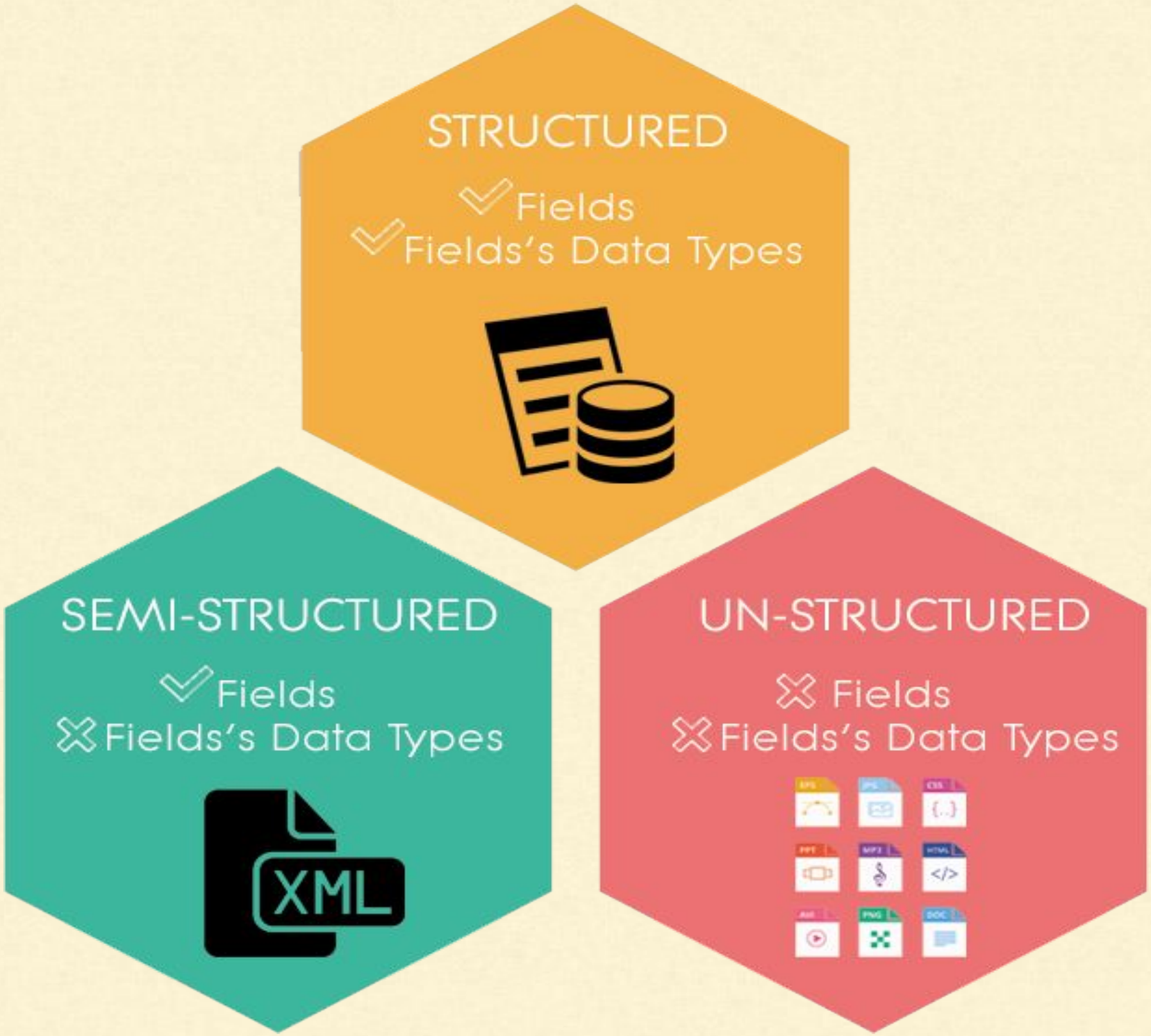
- What/why of Big Data?
- Why Now?
- Examples Customers
- What is Hadoop?
- Components Hadoop
- Further Reading/Assignment

WHAT IS BIG DATA?

WHAT IS BIG DATA?

- Simply: Data of Very Big Size
- Can't process with usual tools
- ***Distributed Architecture Needed***
- Structured / Unstructured

DATA VARIETY



DISTRIBUTED COMPUTING

1. Groups of networked computers
2. Interact with each other
3. To achieve a common goal.

Characterstics of BIG DATA

VOLUME

Data At Rest

Problems related to storage of huge data reliably.
e.g. Storage of Logs of a website, Storage of data by gmail.
FB: 300 PB. 600TB/ day

VELOCITY

Data In Motion

Problems Involving the handling of data coming at fast rate.
e.g. Number of requests being received by Facebook, Youtube streaming, Google Analytics

VARIETY

Data in Many Forms

Problems involving complex data structures
e.g. Maps, Social Graphs, Recommendations

How many bytes in a petabyte?

How many bytes in a petabyte?

1.1259×10^{15}

How many bytes in a petabyte?

1.1259×10^{15}

Kilo	1024	Bytes	1024^1	Bytes
Mega	1024	KB	1024^2	Bytes
Giga	1024	MB	1024^3	Bytes
Tera	1024	GB	1024^4	Bytes
Peta	1024	Tera	1024^5	Bytes
Exa	1024	Peta	1024^6	Bytes
Zeta	1024	Exa	1024^7	Bytes
Yotta	1024	Zeta	1024^8	Bytes

1 byte = 8 bit = can store 256 states

Is 1 PetaByte Big Data?

*If you have to count just vowels in 1 Petabyte data **everyday**, do you need distributed system?*

Is 1 PetaByte Big Data?

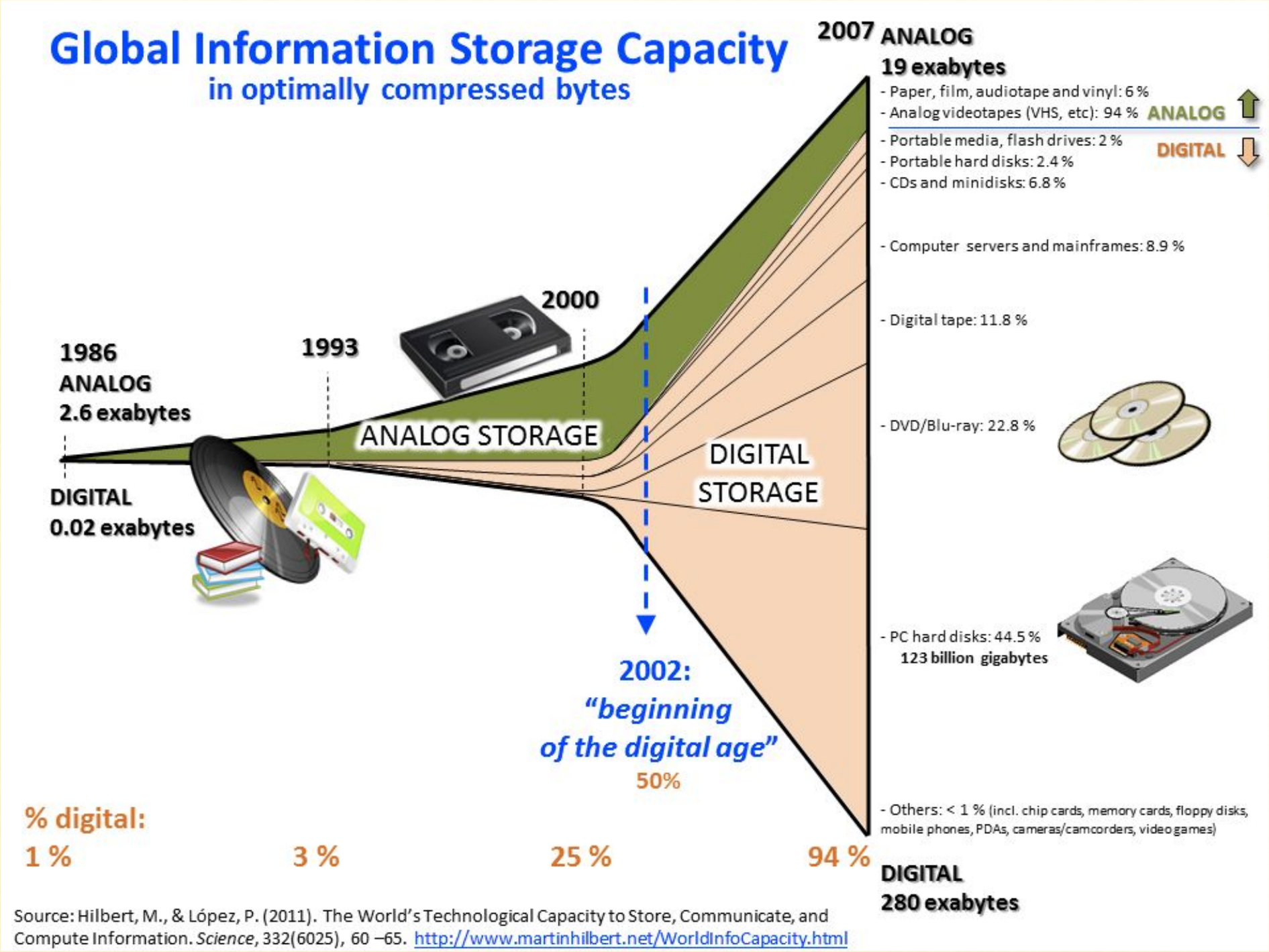
Yes.
Most of the existing systems can't handle it.

Time taken to read 1 TB from HDD?

Time taken to read 1 TB from HDD?

Around 6 hours.

WHY BIG DATA


Global Information Storage Capacity
in optimally compressed bytes
2007 ANALOG
19 exabytes
- Paper, film, audiotape and vinyl: 6 %
- Analog videotapes (VHS, etc): 94 %
ANALOG
DIGITAL
- Portable media, flash drives: 2 %
- Portable hard disks: 2.4 %
- CDs and minidisks: 6.8 %
- Computer servers and mainframes: 8.9 %
- Digital tape: 11.8 %
- DVD/Blu-ray: 22.8 %
- PC hard disks: 44.5 %
123 billion gigabytes
- Others: < 1 % (incl. chip cards, memory cards, floppy disks, mobile phones, PDAs, cameras/camcorders, video games)
DIGITAL
280 exabytes
2000
1993
1986
ANALOG
2.6 exabytes
DIGITAL
0.02 exabytes
ANALOG STORAGE
DIGITAL
STORAGE
2002:
"beginning
of the digital age"
50%
% digital:
1 %
3 %
25 %
94 %
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

WHY IS IT IMPORTANT NOW?

Smart Phones

4.6 billion mobile-phones.
1 - 2 billion people accessing the internet.

Connectivity: Internet Of Things

Connectivity: Social Networks

Facebook:1.06 bn monthly active users, 30 billion pieces shared monthly.
~175 million tweets every day

The connectivity improved.
The devices became cheaper, faster and smaller.

Which components impact the speed computing?

A. CPU
B. Memory
C. Memory Read Speed
D. Disk Speed
E. Disk Size
F. Network Speed
G. All of Above

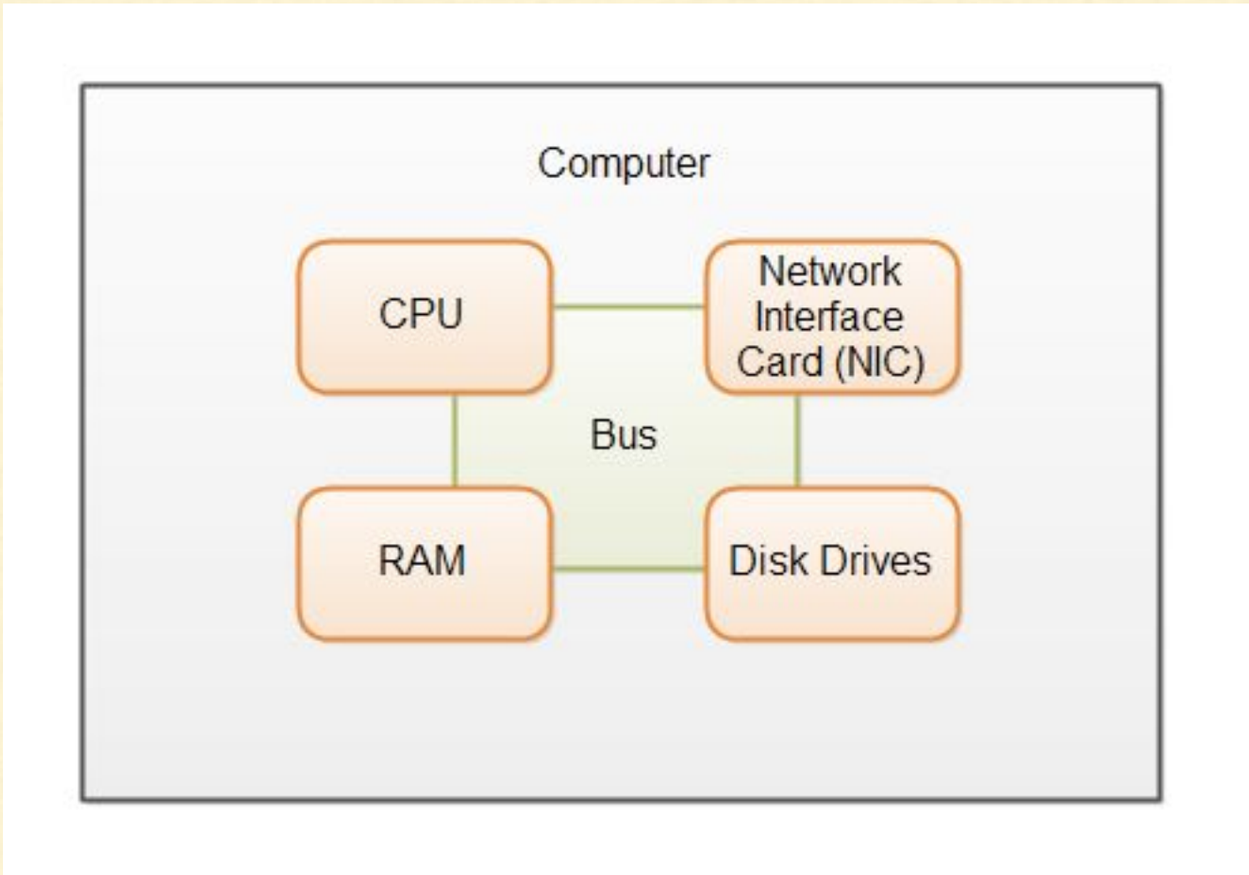
Which components impact the speed computing?

A. CPU
B. Memory
C. Memory Read Speed
D. Disk Speed
E. Disk Size
F. Network Speed
✓ G. All of Above

BIG DATA PROBLEM

To process & store data we need

1. CPU Speed



4. Network

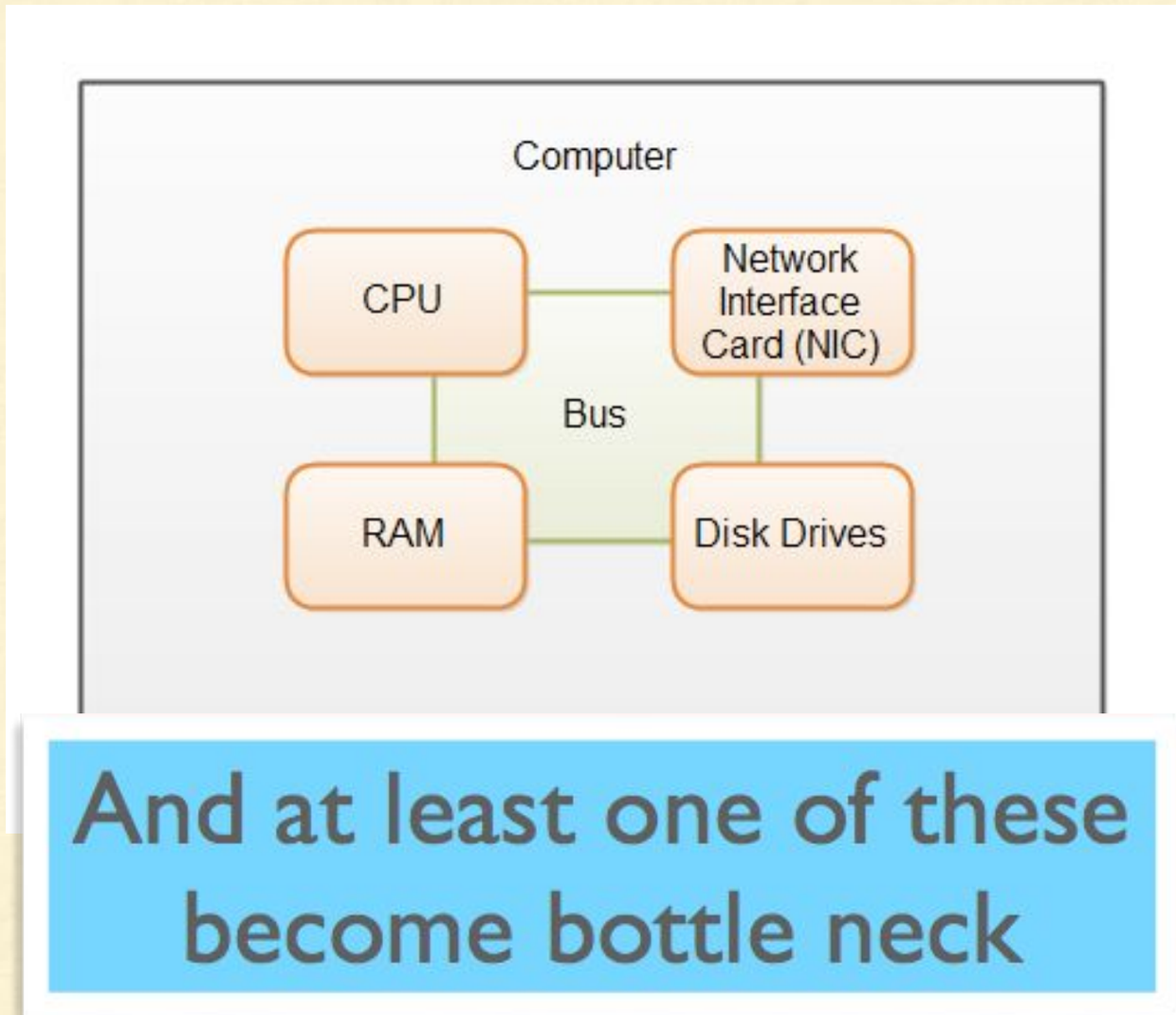
2. RAM - Speed & Size

3. Disk Size + Speed

BIG DATA PROBLEM

1. CPU Speed

To process & store data we need



4. Network

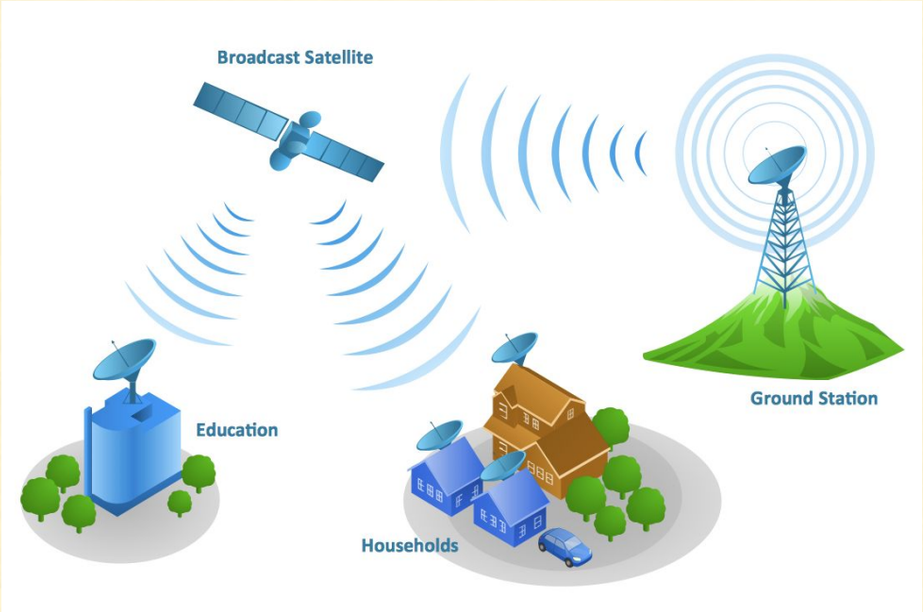
2. RAM - Speed & Size

3. Disk Size + Speed

EXAMPLE BIG DATA CUSTOMERS

Web and e-commerce

1. Recommendation Engines
2. Search Quality
3. Sentiment Analyses
4. A/B testing



Telecommunications

1. Customer Churn Prevention
2. Network Performance Optimization
3. Calling Data Record (CDR) Analysis
4. Analyzing Network to Predict Failure

EXAMPLE BIG DATA CUSTOMERS

Government

1. Fraud Detection
2. Cyber Security Welfare
3. Justice

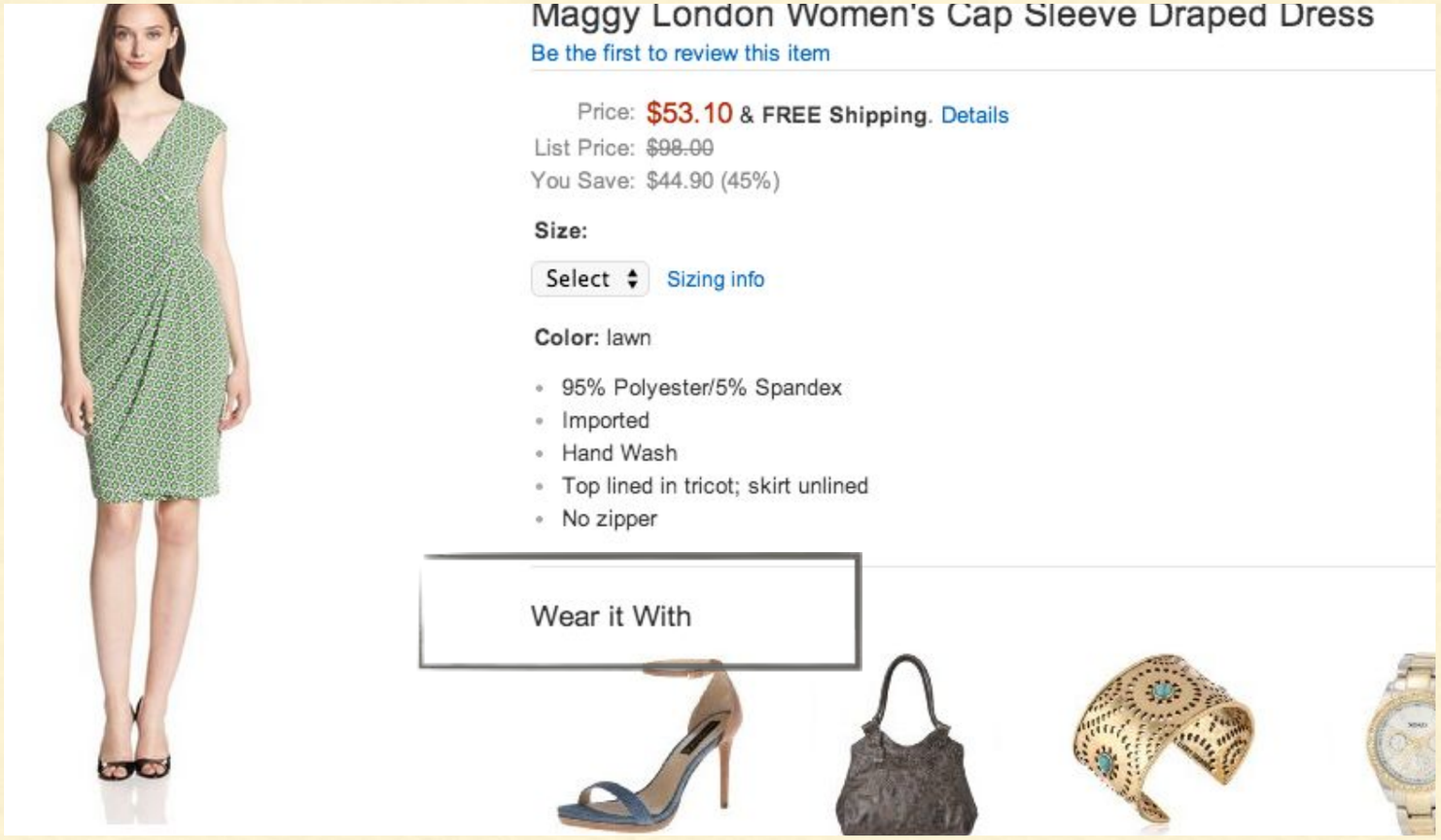
Healthcare & Life Sciences

1. Health information exchange
2. Gene sequencing
3. Healthcare improvements
4. Drug Safety

Break of 10 mins. back by 9:42pm IST.

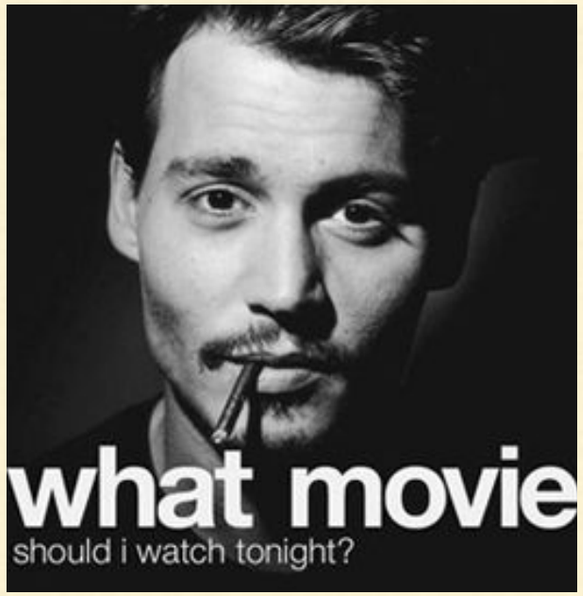
EXAMPLE BIG DATA PROBLEMS

Recommendations



EXAMPLE BIG DATA PROBLEMS

Recommendations



EXAMPLE BIG DATA PROBLEMS

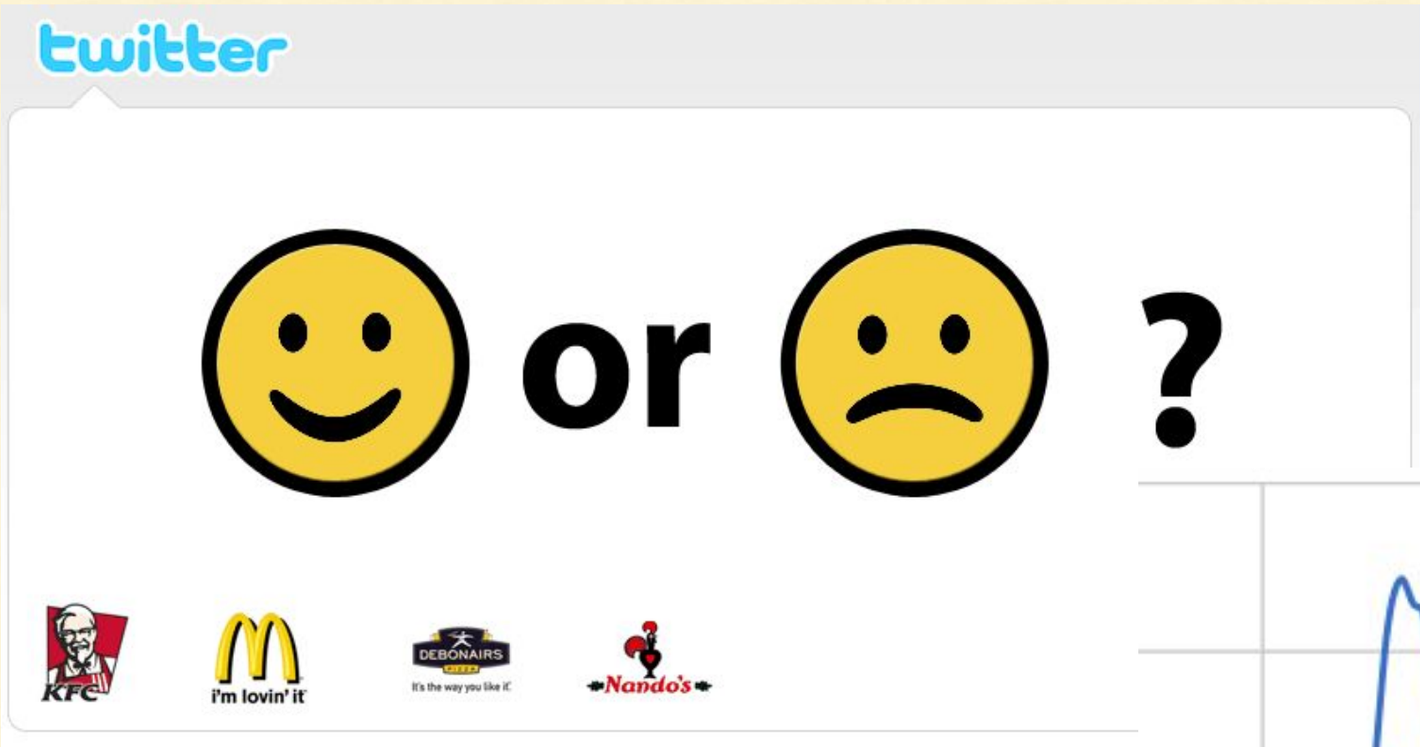
Recommendations - How?

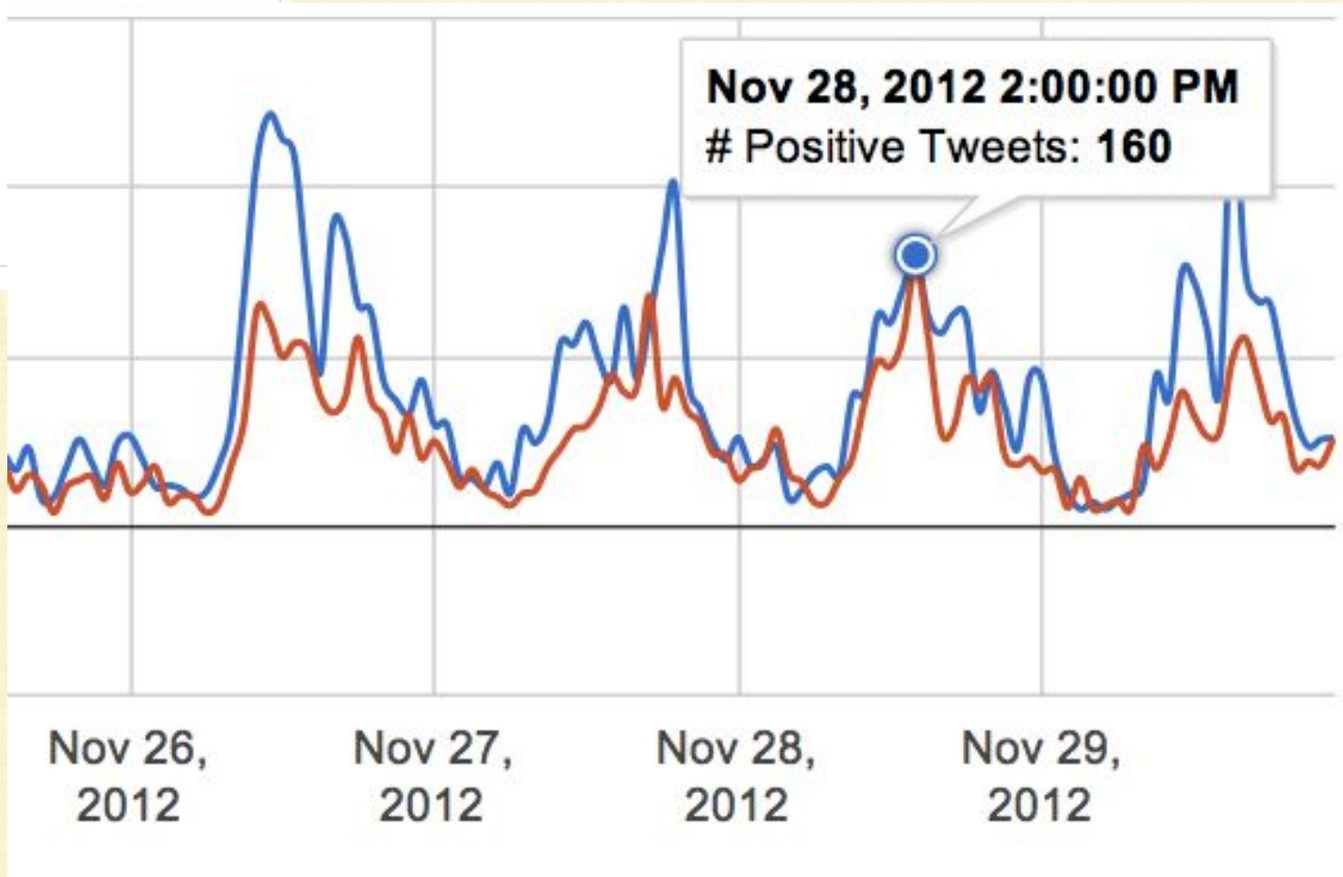
USER ID	MOVIE ID	RATING
KUMAR	matrix	4.0
KUMAR	Ice age	3.5
GIRI	apocalypse now	3.6
GIRI	Ice age	3.5

USER ID	MOVIE ID	RATING
KUMAR	apocalypse now	3.6
GIRI	matrix	4.0

EXAMPLE BIG DATA PROBLEMS

Sentiment Analysis



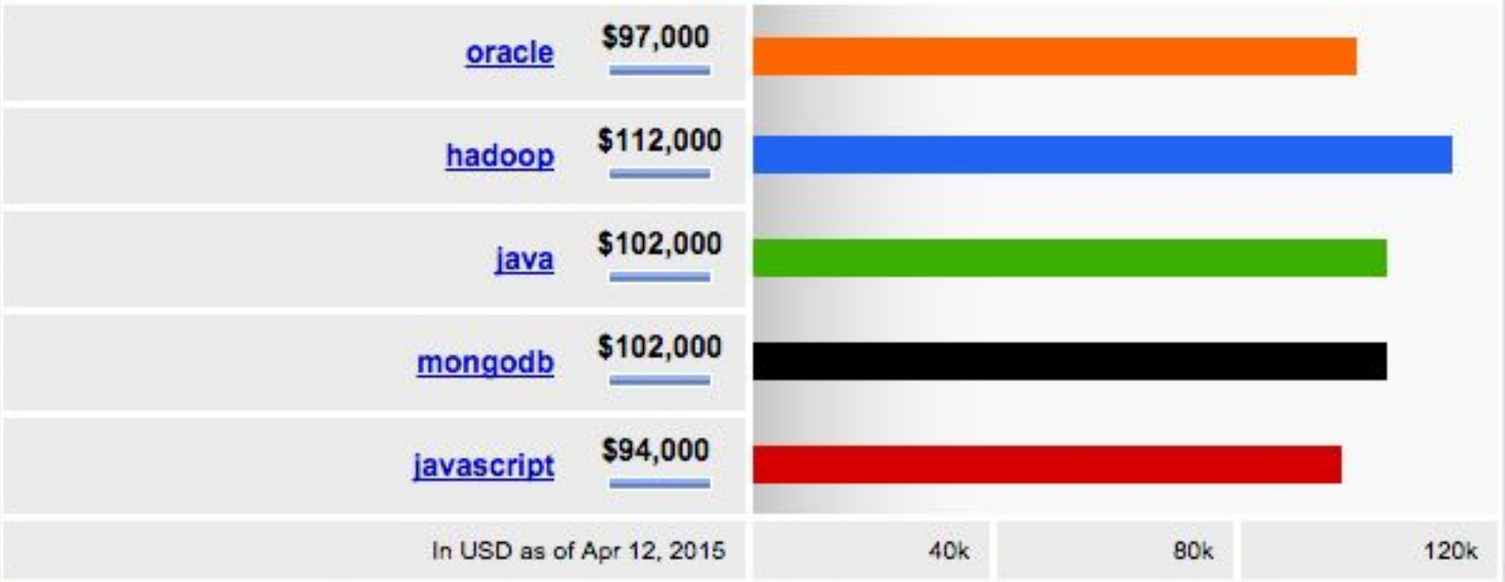


11 COMMON MYTHS

BIG DATA

1. Always means data above or in range of TB
2. Is always about social media. Doesn't apply to me.
3. Will replace EDW [Enterprise data warehouse]
4. Is just a buzzword. No Practical Applications
5. Is New Concept
6. Will be future.
7. Is Expensive
8. Is only for data scientists. Or is magic.
9. We have enough hardware. Don't need any more.
10. We will build it when we need it.
11. Big Data is about Hadoop.

SALARY TRENDS



Source:Indeed.com

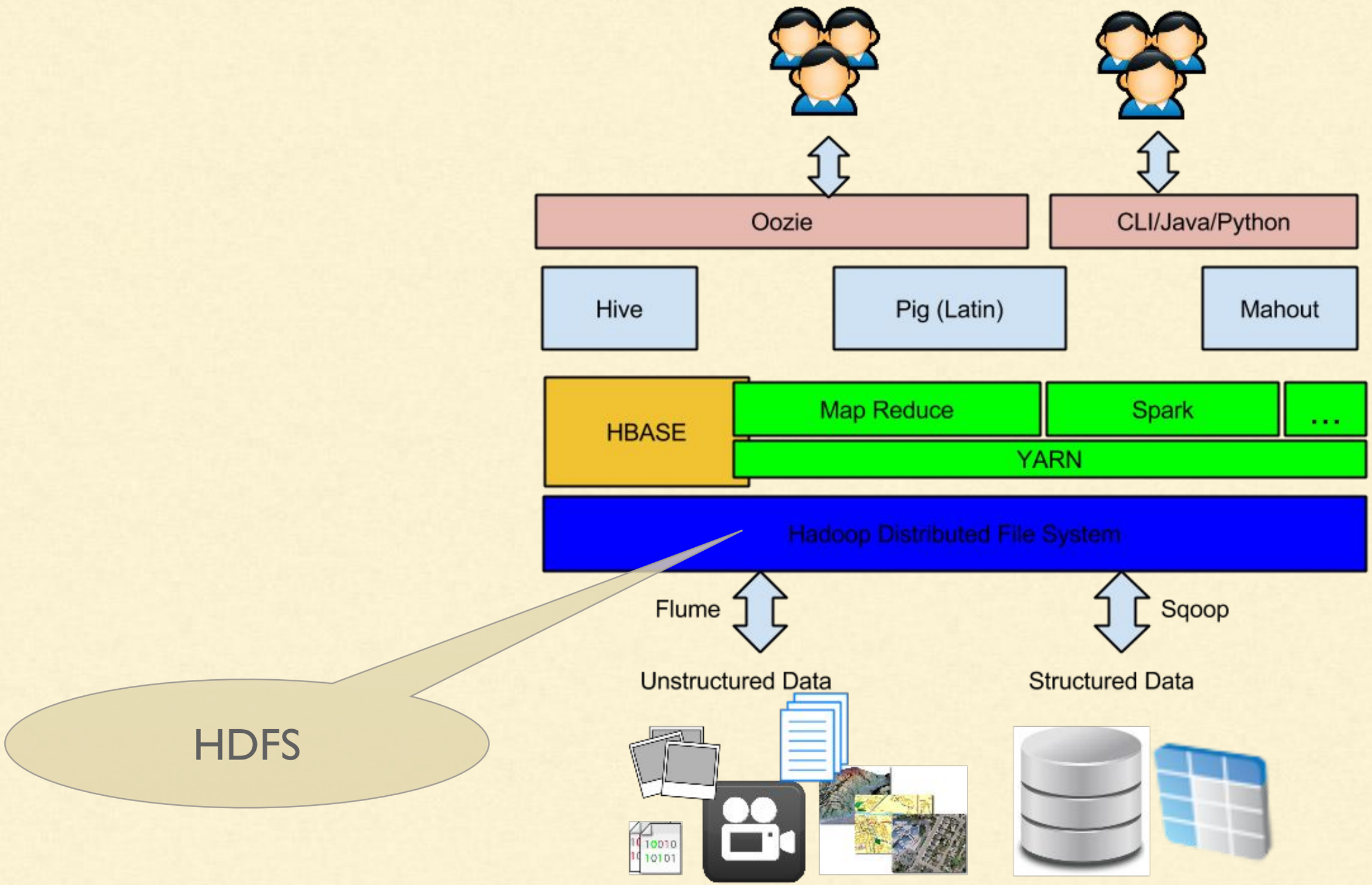
BIG DATA SOLUTIONS

1. Apache Hadoop
   - Apache Spark
2. Cassandra
3. MongoDB
4. Google Compute Engine

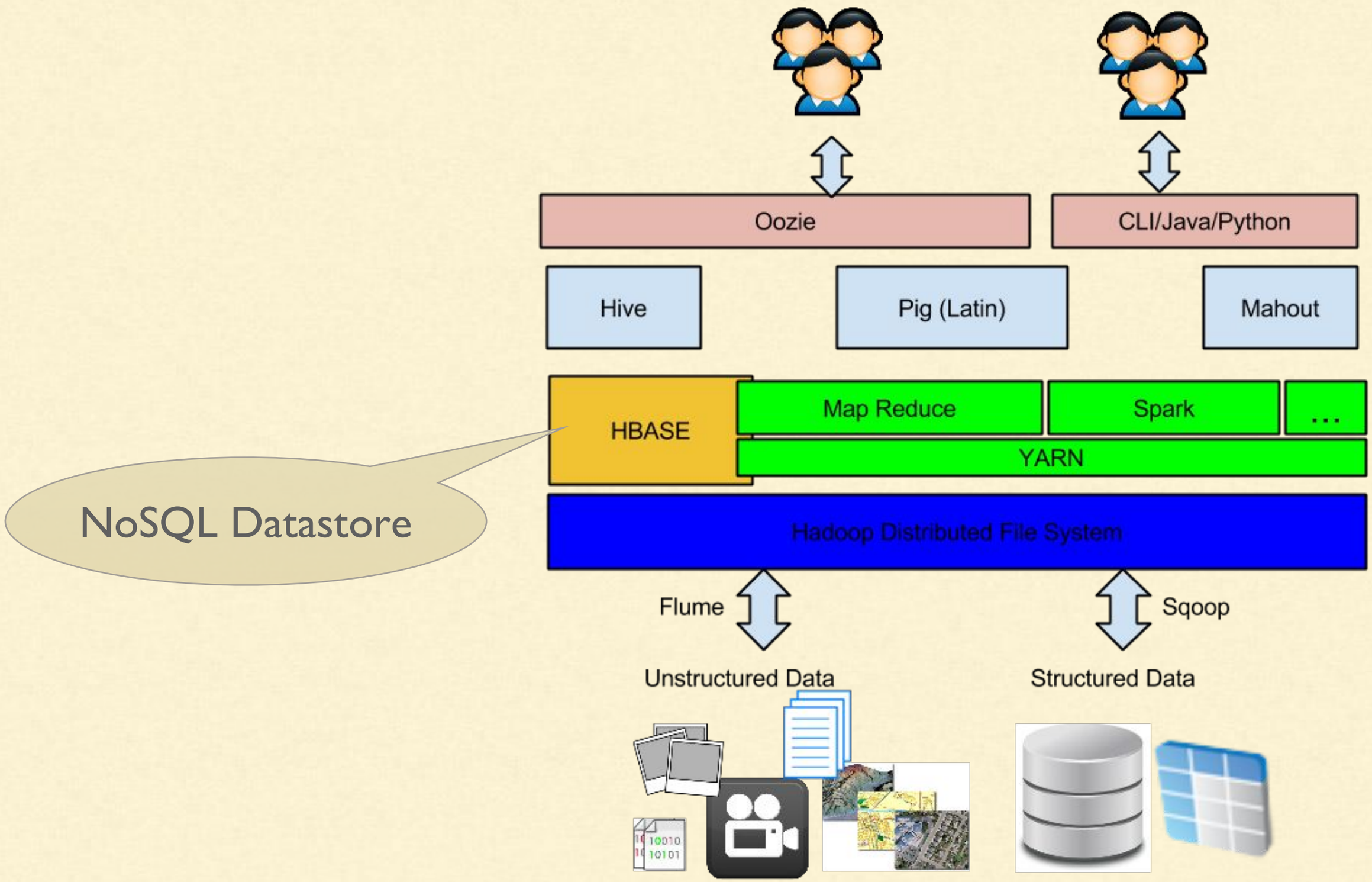
WHAT IS HADOOP?

A. Created by Doug Cutting (of Yahoo) and Mike Cafarella
B. Based on GFS, GMR & Google Big Table
C. Built for Nutch search engine project
D. Named after Toy Elephant
E. Open Source - Apache
F. Power, Popular & Supported
G. Framework to handle Big Data
H. For reliable, scalable, distributed computing
I. Written in Java

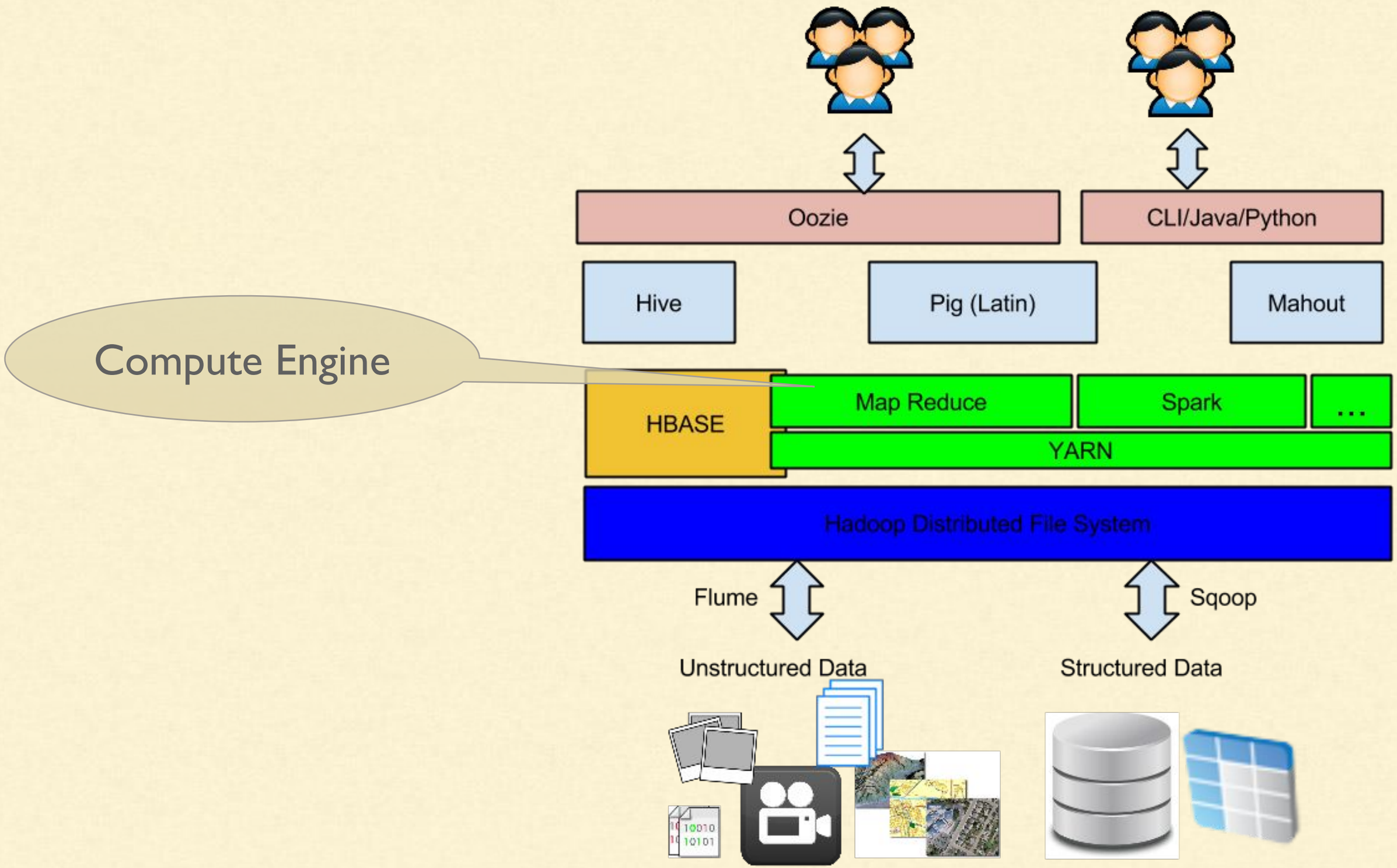
COMPONENTS - HDFS



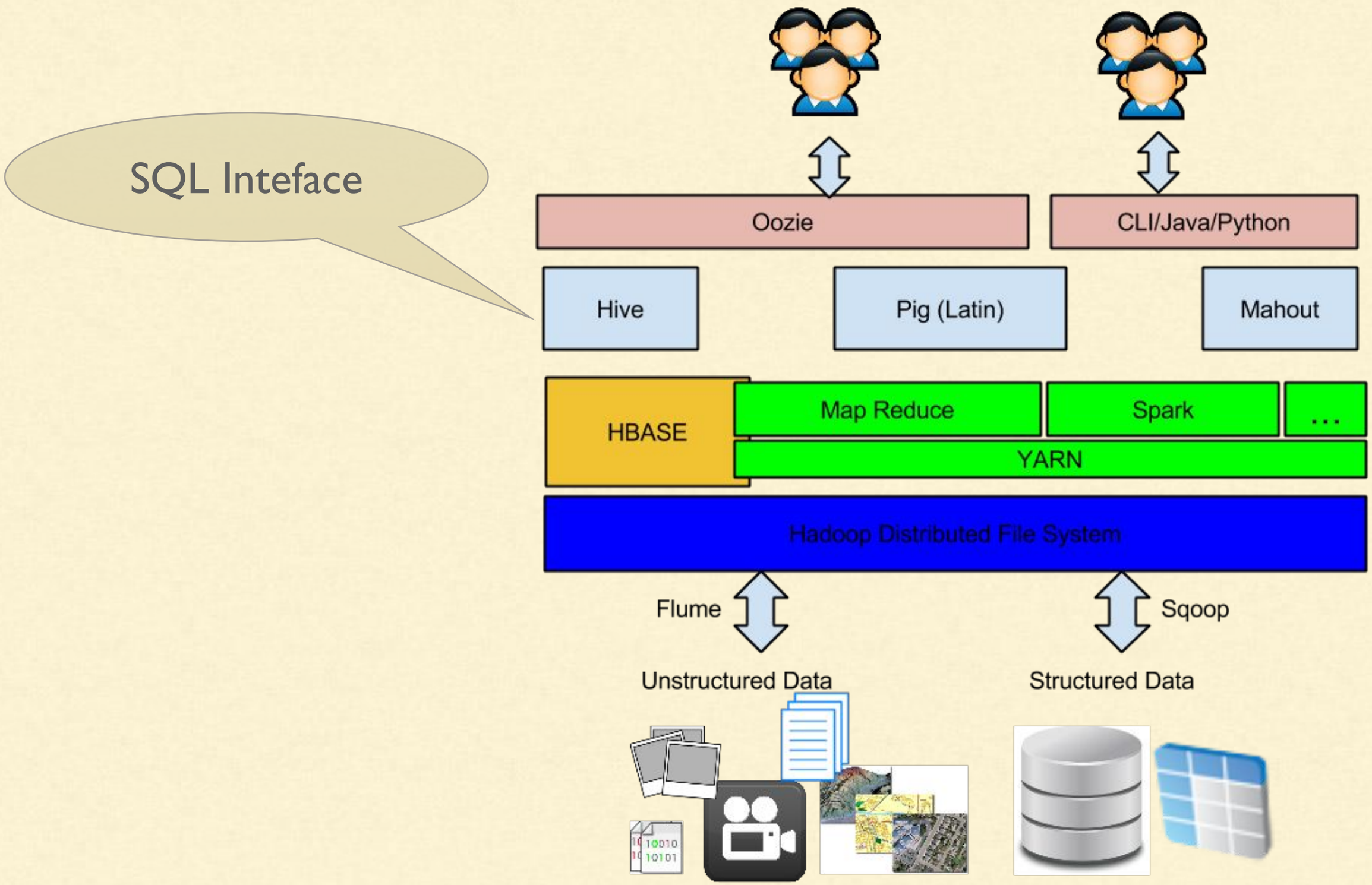
COMPONENTS


Oozie
CLI/Java/Python
Hive
Pig (Latin)
Mahout
HBASE
Map Reduce
Spark
...
YARN
Hadoop Distributed File System
Flume
Sqoop
Unstructured Data
Structured Data
NoSQL Datastore

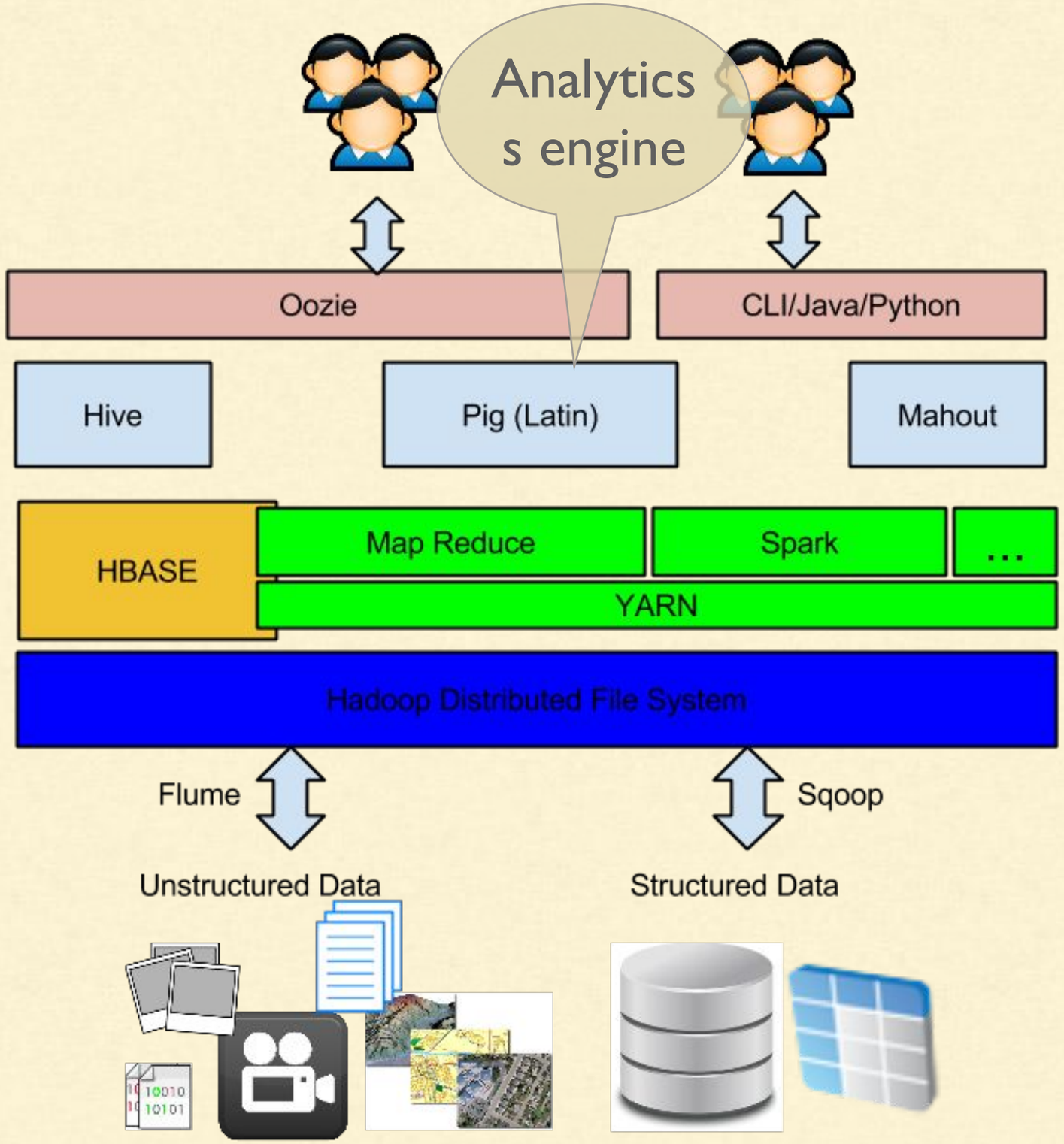
COMPONENTS


Compute Engine
Oozie
CLI/Java/Python
Hive
Pig (Latin)
Mahout
HBASE
Map Reduce
Spark
...
YARN
Hadoop Distributed File System
Flume
Sqoop
Unstructured Data
Structured Data

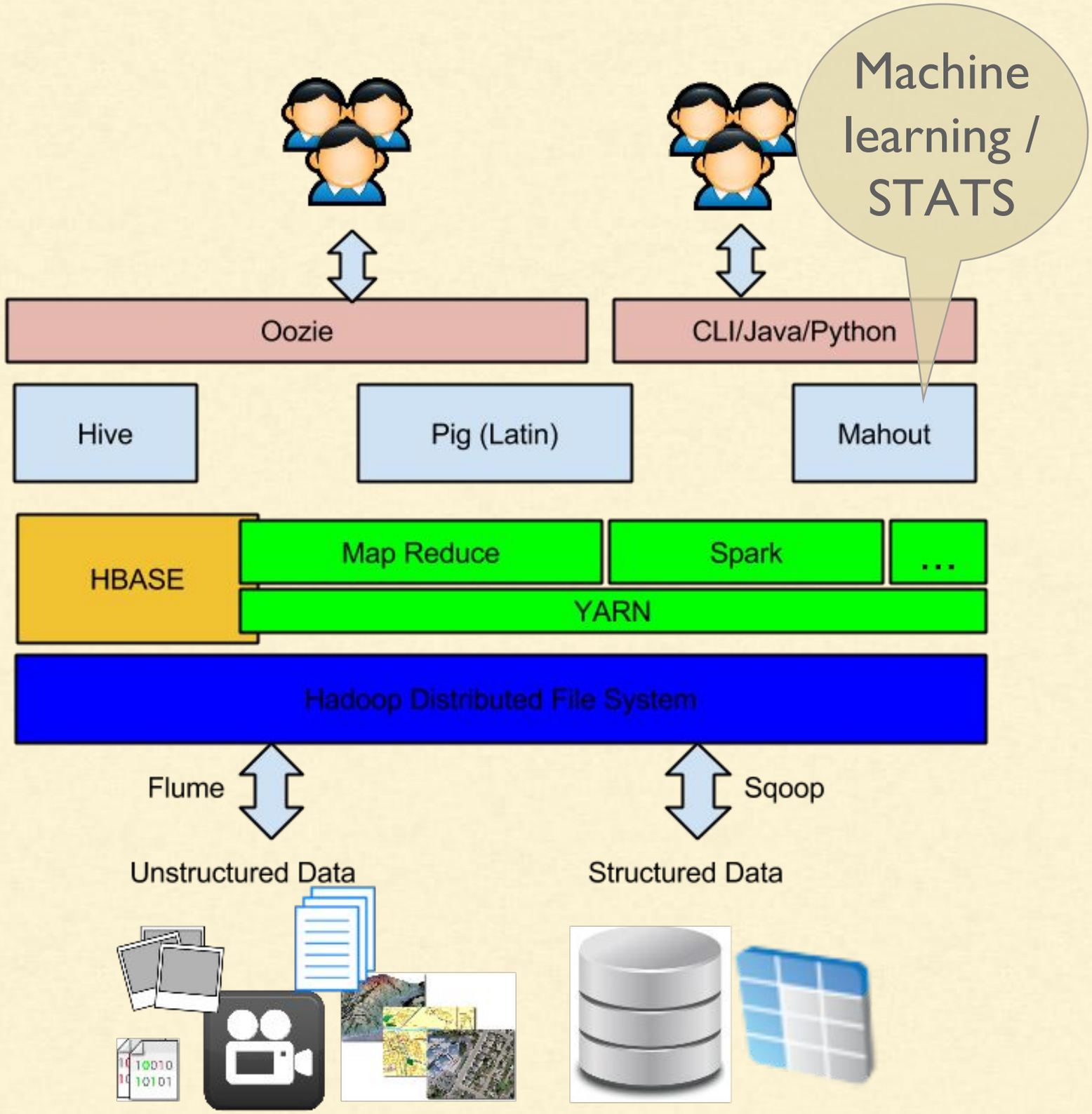
COMPONENTS


SQL Inteface
Oozie
CLI/Java/Python
Hive
Pig (Latin)
Mahout
HBASE
Map Reduce
Spark
...
YARN
Hadoop Distributed File System
Flume
Sqoop
Unstructured Data
Structured Data

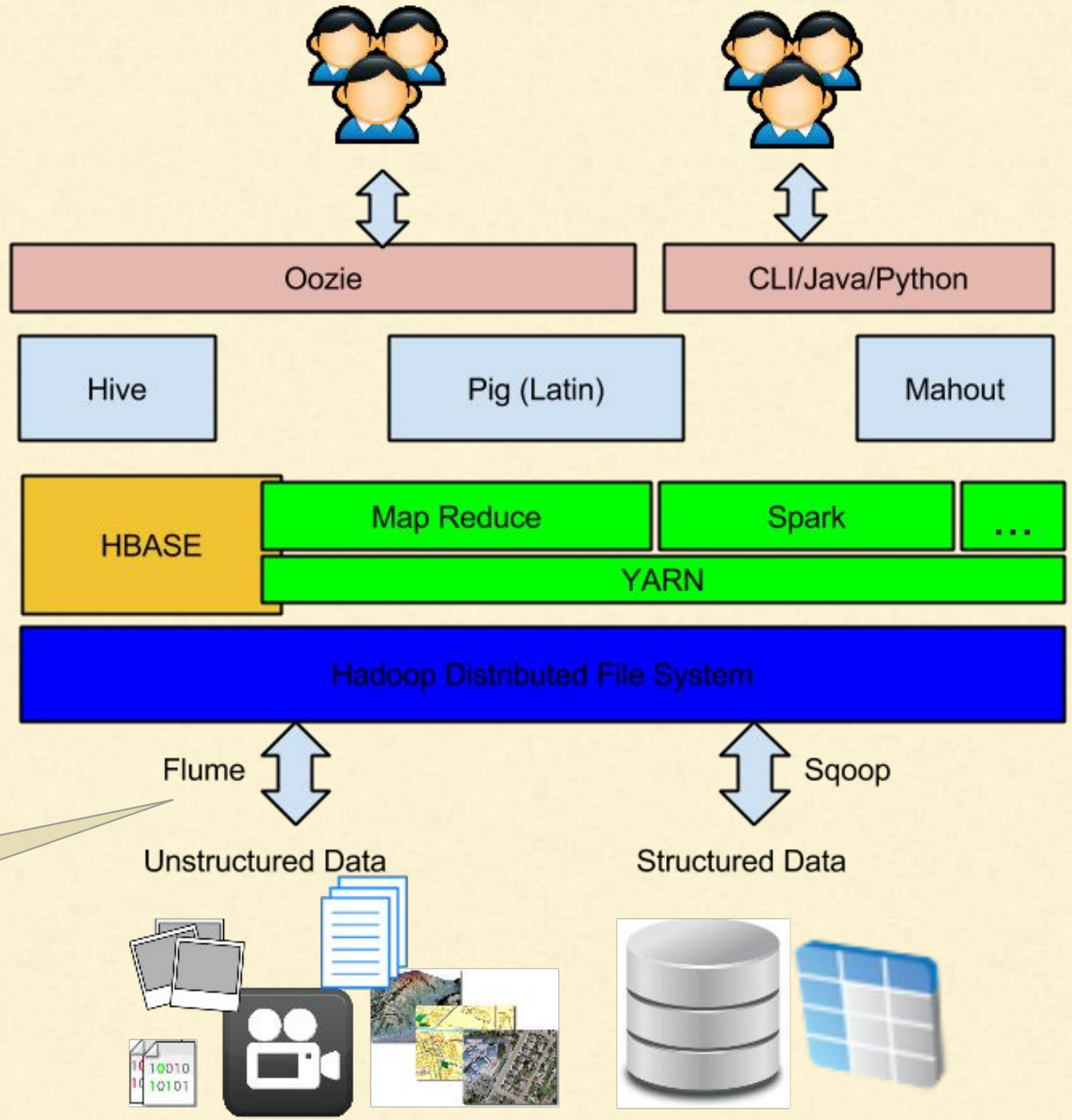
COMPONENTS


Analytics
s engine
Oozie
CLI/Java/Python
Hive
Pig (Latin)
Mahout
HBASE
Map Reduce
Spark
...
YARN
Hadoop Distributed File System
Flume
Sqoop
Unstructured Data
Structured Data

COMPONENTS

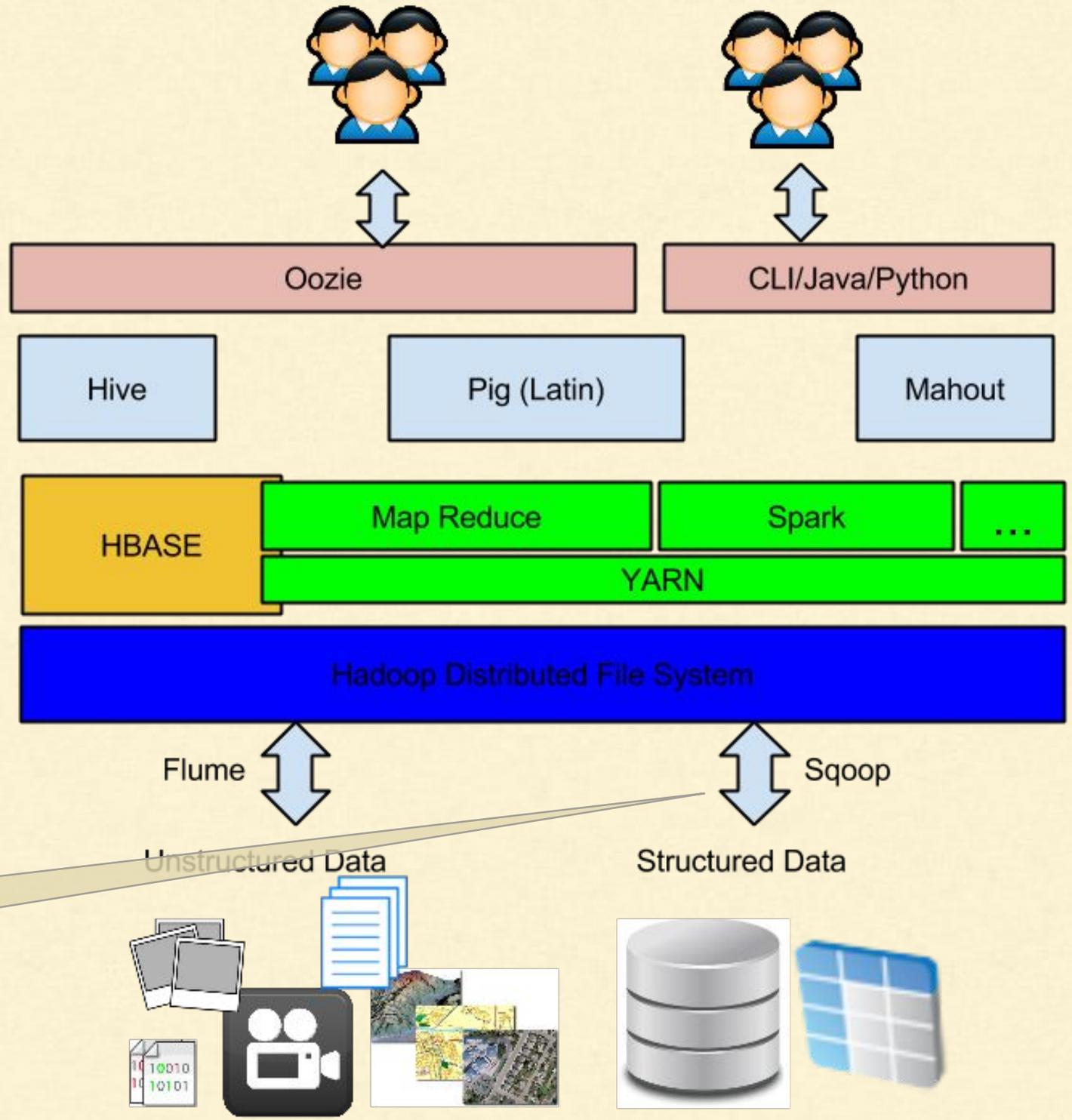

Machine learning / STATS
Oozie
CLI/Java/Python
Hive
Pig (Latin)
Mahout
HBASE
Map Reduce
Spark
...
YARN
Hadoop Distributed File System
Flume
Sqoop
Unstructured Data
Structured Data

COMPONENTS



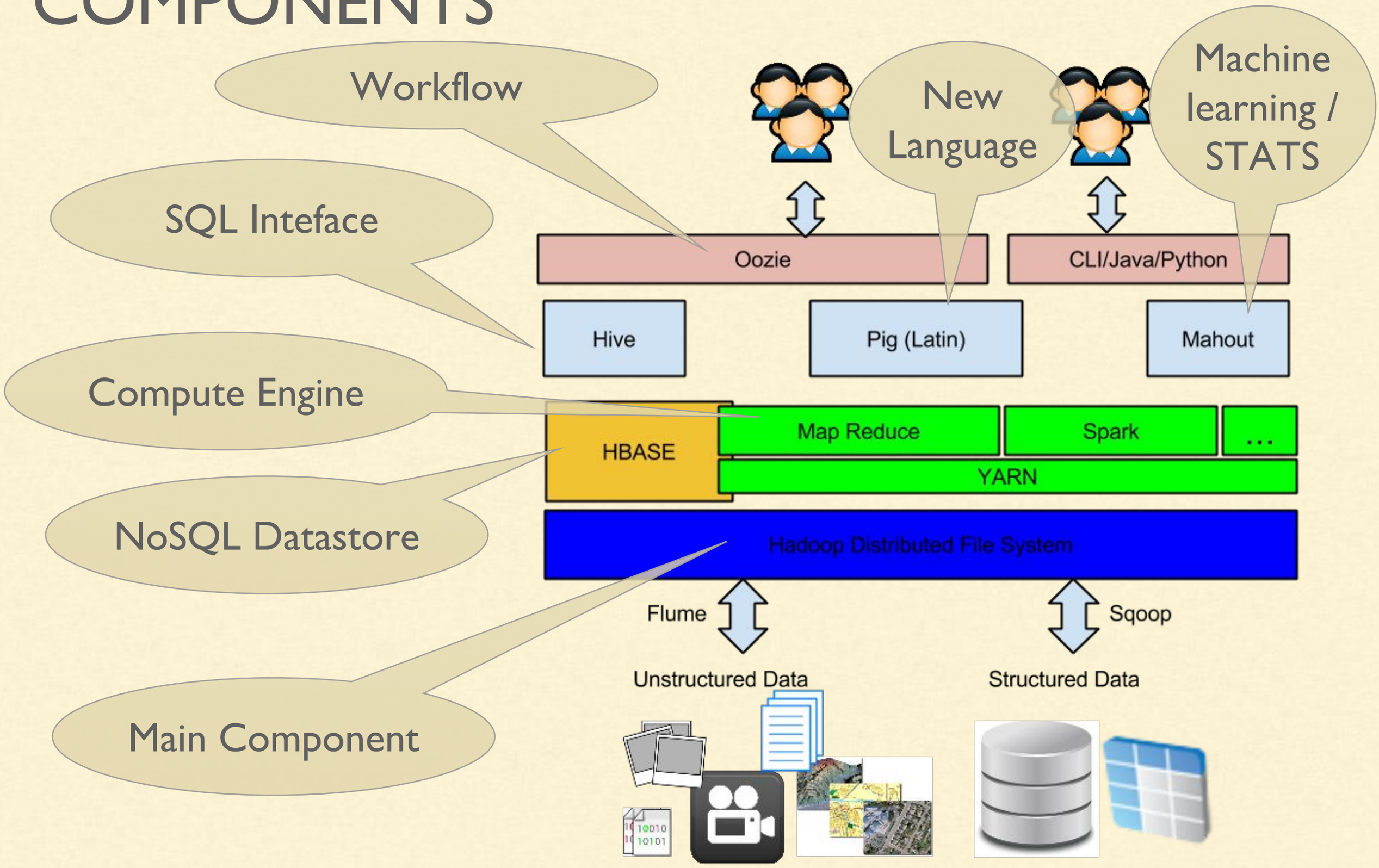
Import unstructed data

COMPONENTS



Import Structured Data

COMPONENTS


Workflow
SQL Inteface
Compute Engine
NoSQL Datastore
Main Component
New Language
Machine learning / STATS
Oozie
CLI/Java/Python
Hive
Pig (Latin)
Mahout
HBASE
Map Reduce
Spark
...
YARN
Hadoop Distributed File System
Flume
Sqoop
Unstructured Data
Structured Data

WHAT IS CLOUDXLABS™?

- For Real Life Experience
- An online cluster of servers
- With all required tools installed
- Accessible globally
- Do not require high end configuration

ASSIGNMENT / PRE-WORK

1. Go through Cloud Labs:
   - Admin Console (Ambari) http://hadoop1.knowbigdata.com:8080
   - Hue http://hadoop1.knowbigdata.com:8000
   - SSH
2. Go through LMS: http://www.knowbigdata.com/my-courses
3. Setup Hadoop (optional) - Environment based on the VM
4. Finish the quiz from LMS
5. See Assignment section on LMS

FURTHER READING

http://en.wikipedia.org/wiki/Apache_Hadoop

http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html

FULL COURSE

www.KnowBigData.com

1. Second Session onwards
   - 10 Apr - 8pm IST
2. Sat-Sun - 3 hours
3. 33 hrs - 3 hr x 12 classes
4. Includes CloudLabs + Support + LMS
5. Every class is recorded.

+1 419 665 3276 (US)
+91 803 959 1464 (IN)

reachus@KnowBigData.com

Big Data & Hadoop

Thank you.

+1 419 665 3276 (US)
+91 803 959 1464 (IN)

reachus@knowbigdata.com

Subscribe to our Youtube channel for latest videos - https://www.youtube.com/channel/UCxugRFe5wETYA7nMH6VGyEA

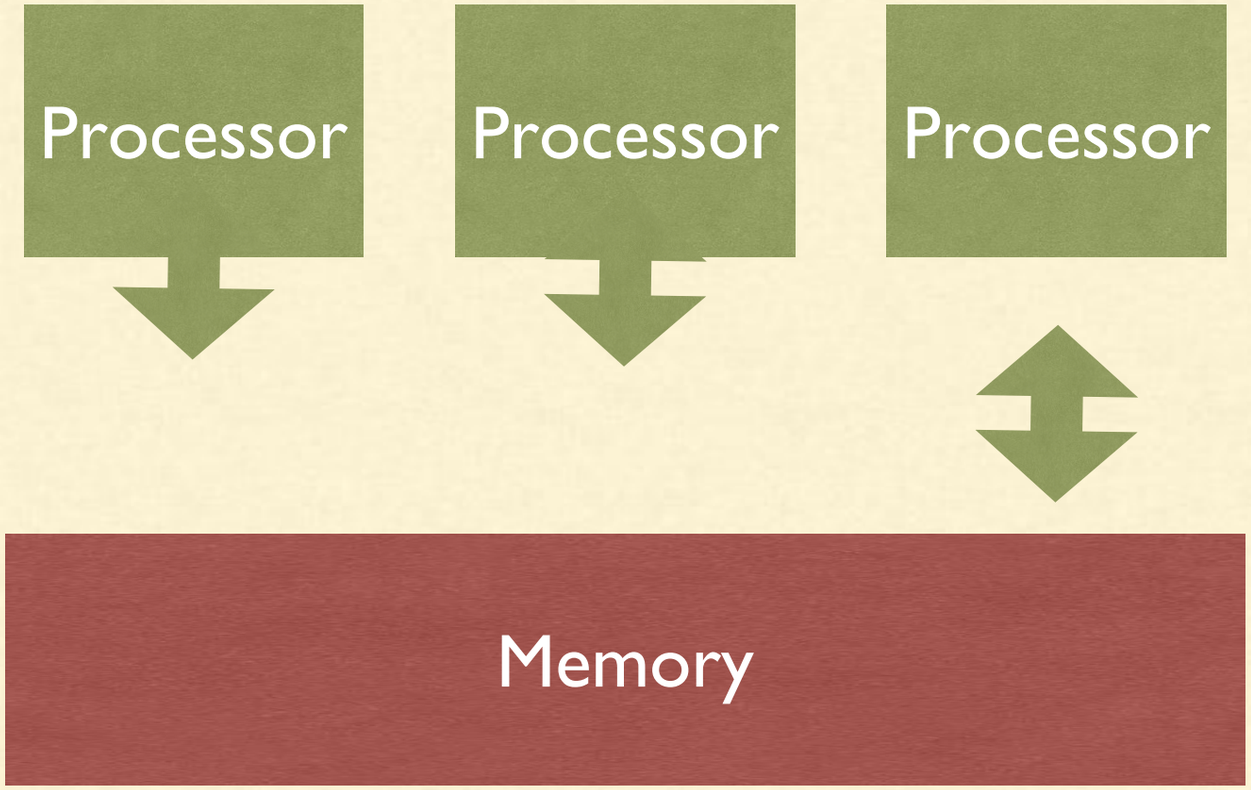
About Instructor?

TYPES OF COMPUTING

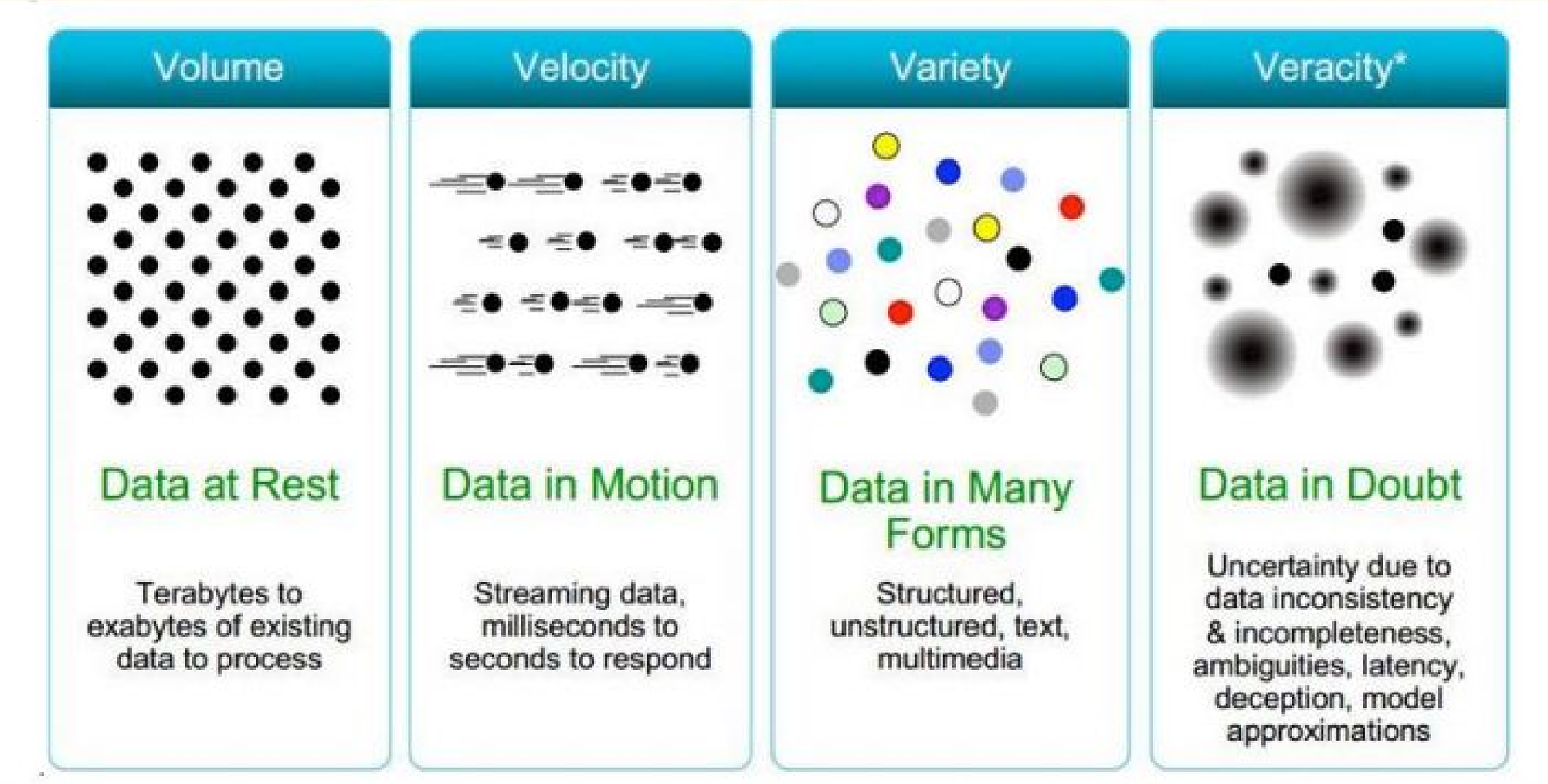
DISTRIBUTED

1. Groups of networked computers
2. Interact with each other
3. To achieve a common goal.

PARALLEL



WHAT IS BIG DATA?

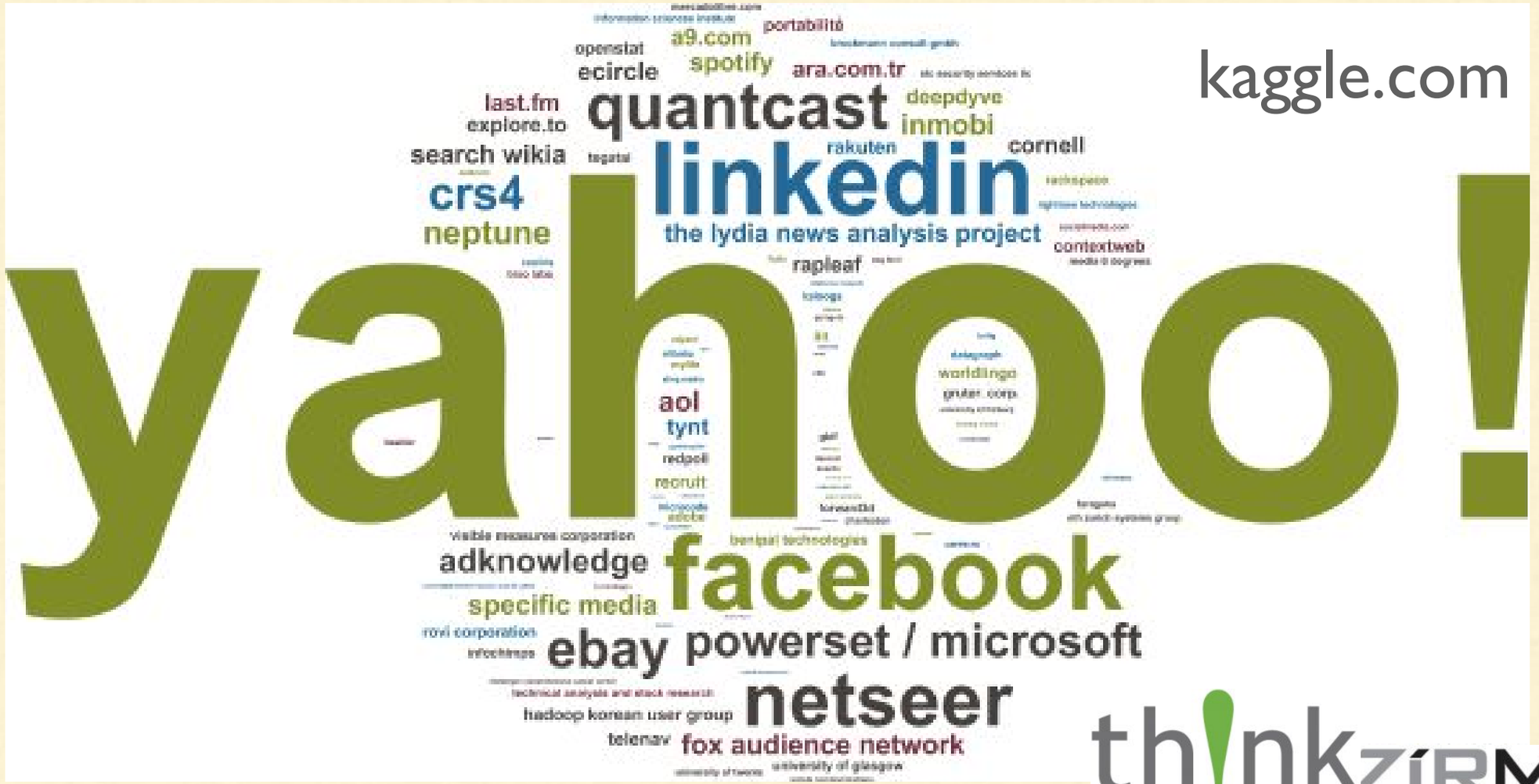


Facebook: 500TB /day
Boeing737: 240 TB / flight

. Clickstreams:
~ 1m events / sec

Geospatial data
3D data
audio & video
Unstructured text

AND MANY MORE…



kaggle.com

DISTRIBUTED COMPUTING

Take the code to the data.

Not data to the code.
Data is very big as compared to size of code.