Welcome to

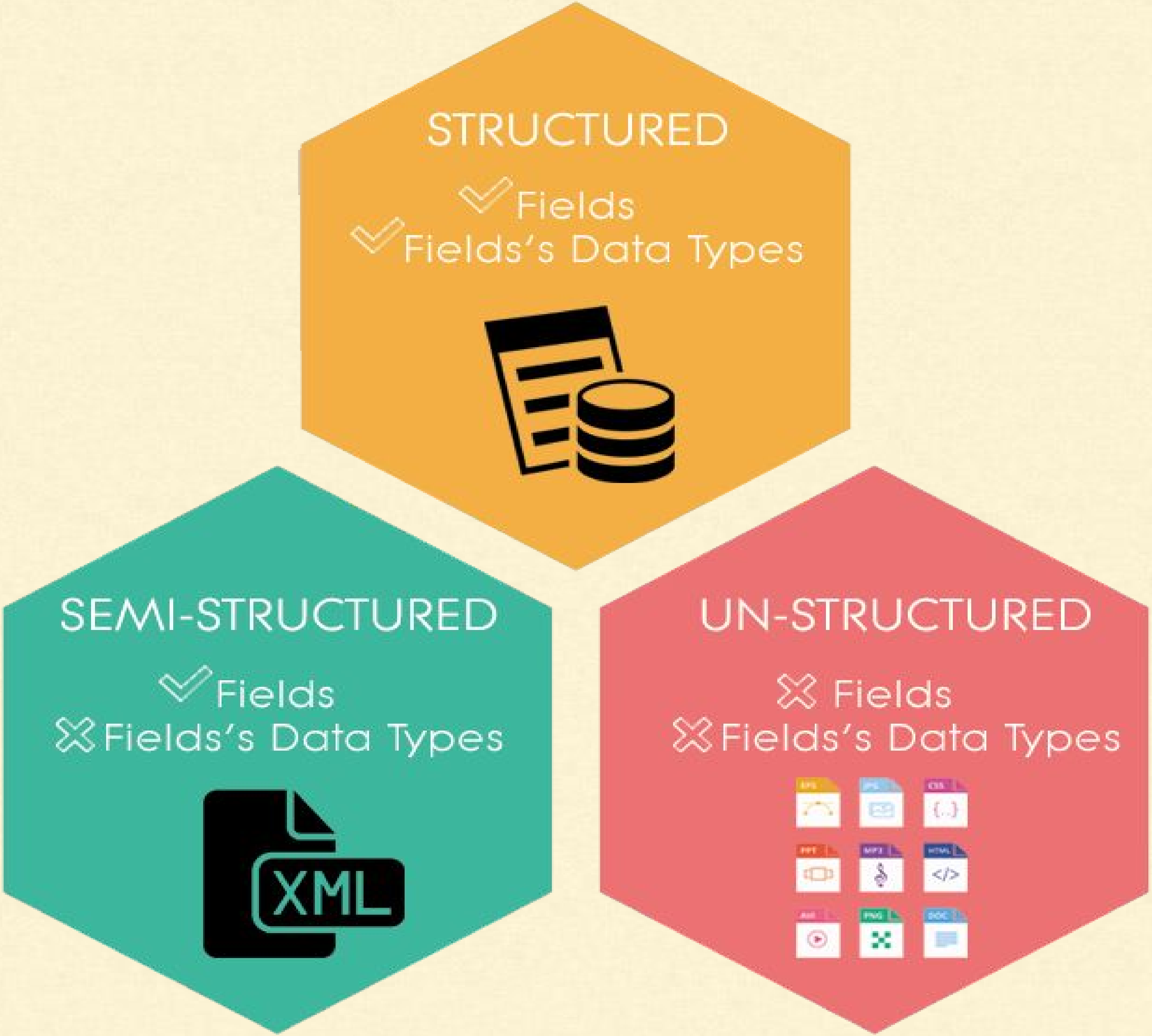
Big Data

with

Hadoop & Spark

Introduction

Data Variety



STRUCTURED

✓ Fields

✓ Fields's Data Types

SEMI-STRUCTURED

✓ Fields

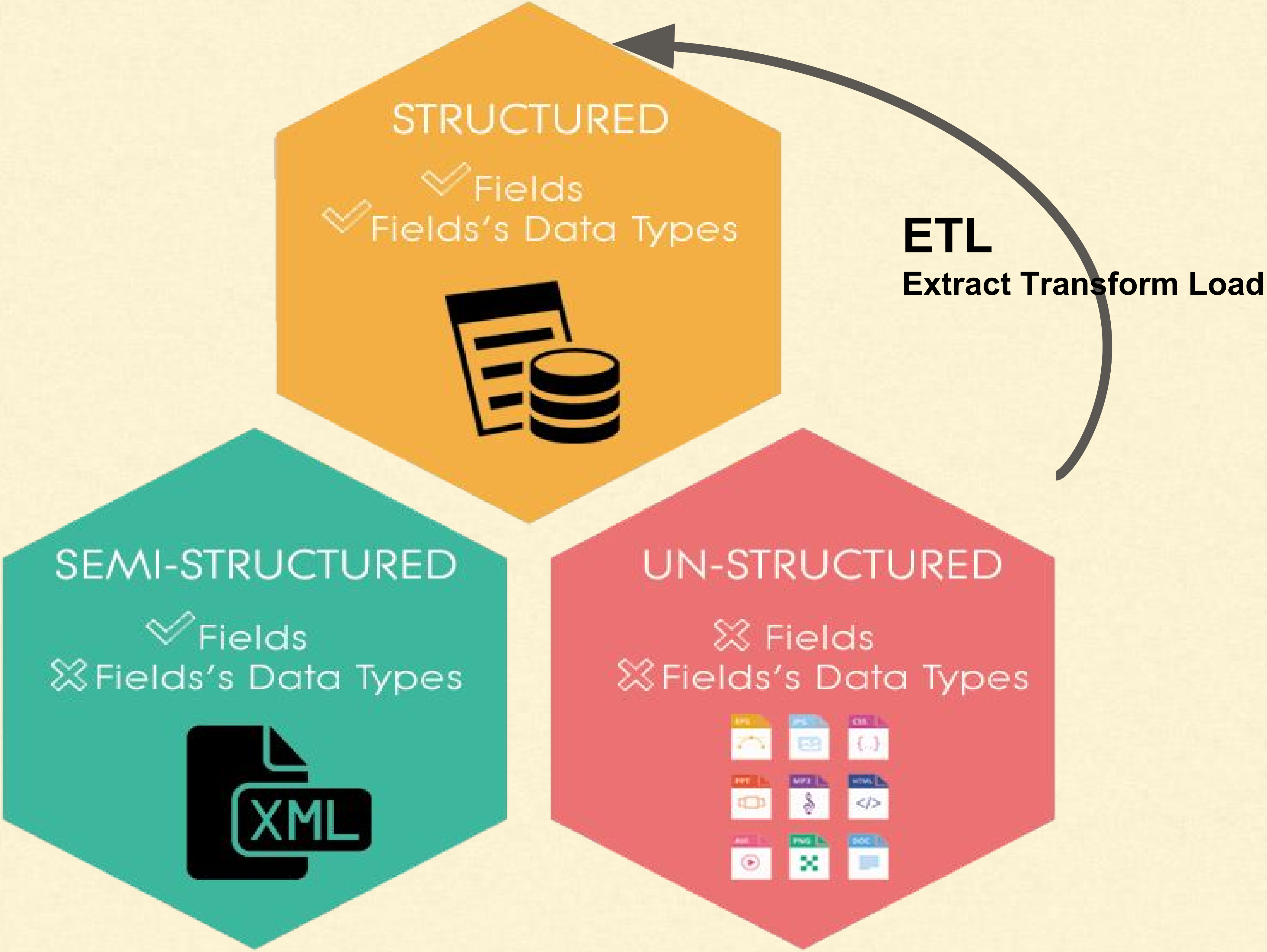
✗ Fields's Data Types

UN-STRUCTURED

✗ Fields

✗ Fields's Data Types

Data Variety



STRUCTURED

✓ Fields

✓ Fields's Data Types

SEMI-STRUCTURED

✓ Fields

✗ Fields's Data Types

UN-STRUCTURED

✗ Fields

✗ Fields's Data Types

ETL

Extract Transform Load

Distributed Systems

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

Question

How Many Bytes in One Petabyte?

$1.1259\text{x}10^{\wedge 15}$

Question

How Much Data Facebook Stores in One Day?

600 TB

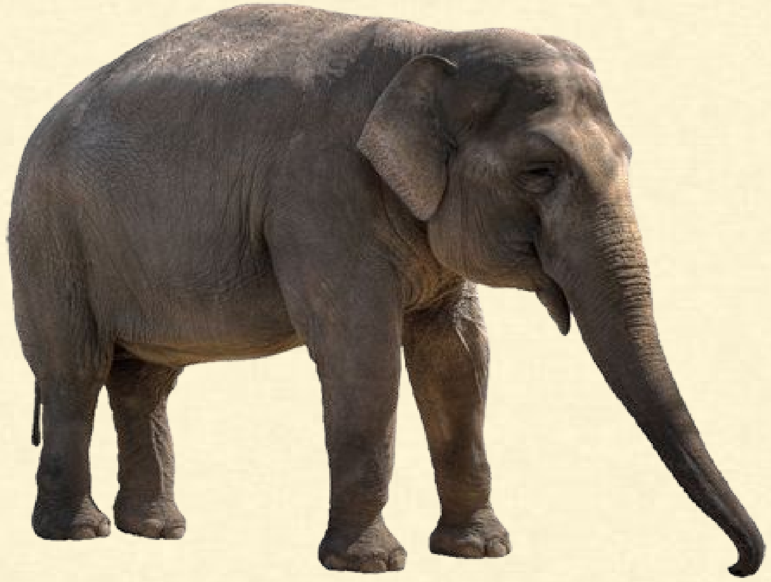
What is Big Data?

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

Characteristics 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

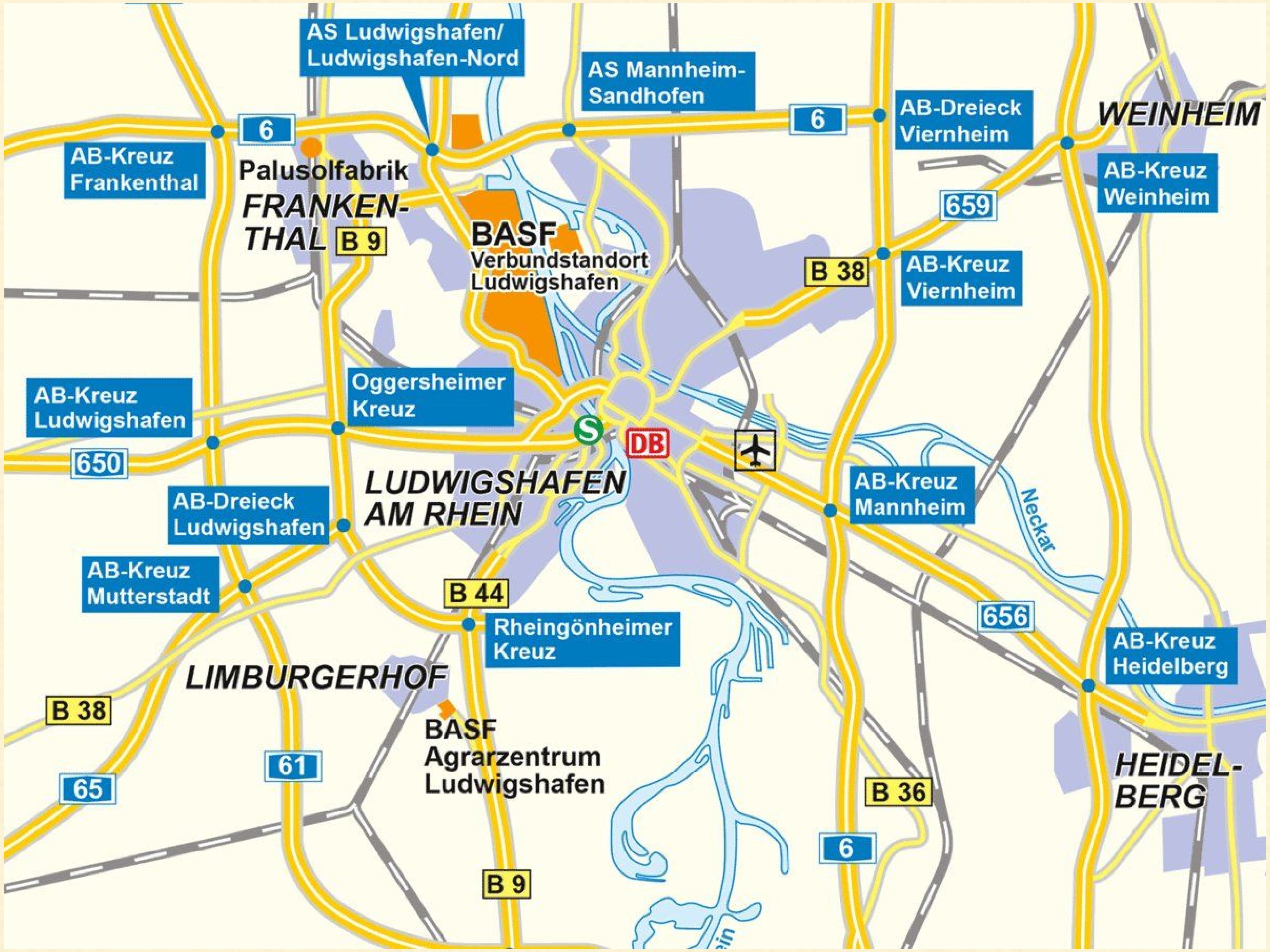
Problems Involving the handling of data coming at fast rate.
e.g. Number of requests being received by Facebook, Youtube streaming, Google Analytics

VARIETY

Data in Many Forms

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

Characteristics of Big Data - Variety



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

Question

Time taken to read 1 TB from HDD?

Around 6 hours

Is One PetaByte Big Data?

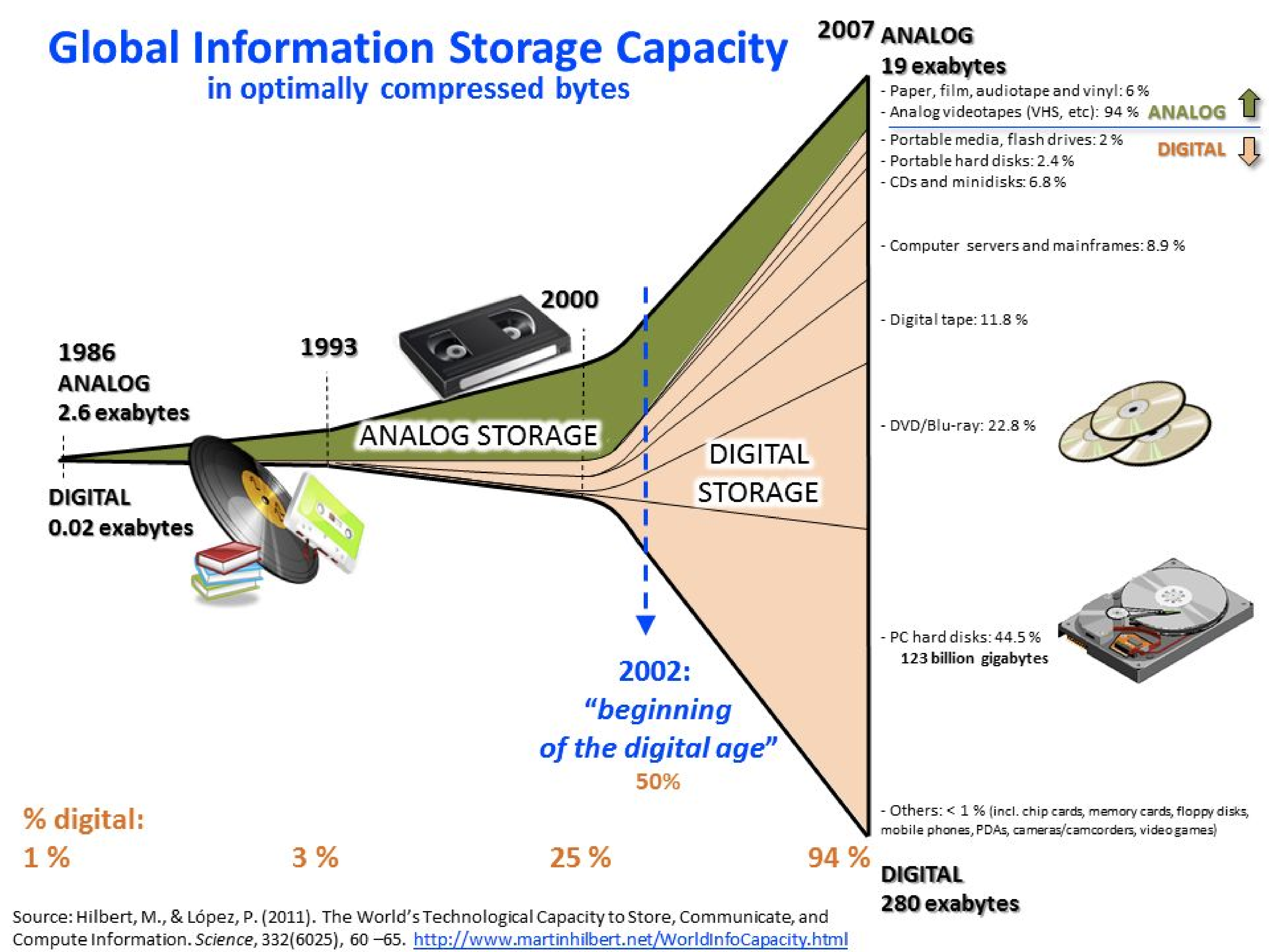
*If you have to count just vowels in 1 Petabyte data **everyday**, do you need distributed system?*

Is One PetaByte Big Data?

Yes.

Most of the existing systems can’t handle it.

Why Big Data?



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

Why is It Important Now?

X

=>

Devices:

Smart Phones

4.6 billion mobile-phones.

1 - 2 billion people accessing the internet.

Connectivity

Wifi, 4G, NFC, GPS

Application

Social Networks

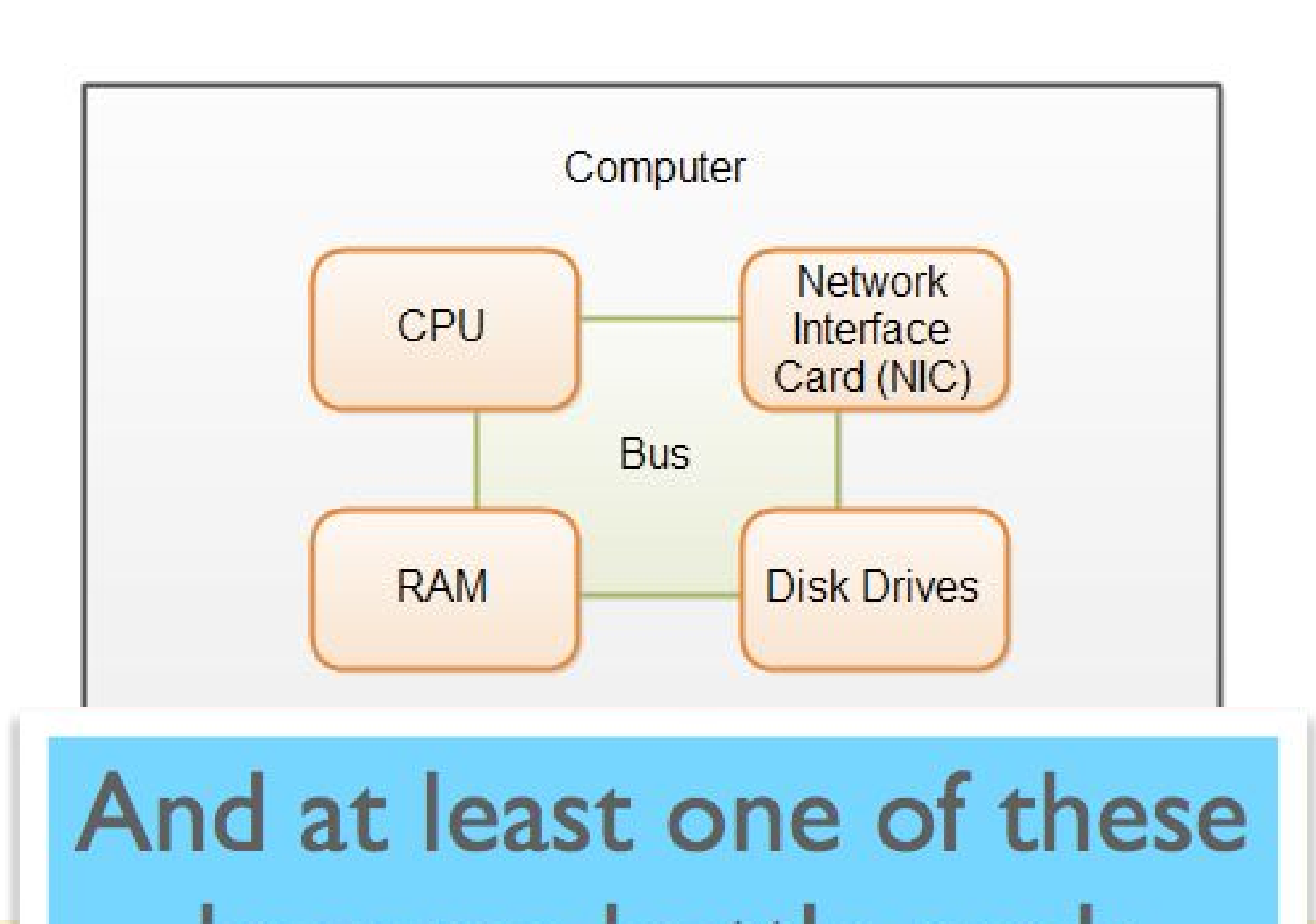
Internet of Things

The devices became cheaper, faster and smaller.
The connectivity improved. **Result: Many Applications**

Computing Components

1. CPU Speed

To process & store data we need



And at least one of these become bottle neck

4. Network

2. RAM - Speed & Size

3. HDD or SSD
Disk Size + Speed

Which Components Impact the Speed of Computing?

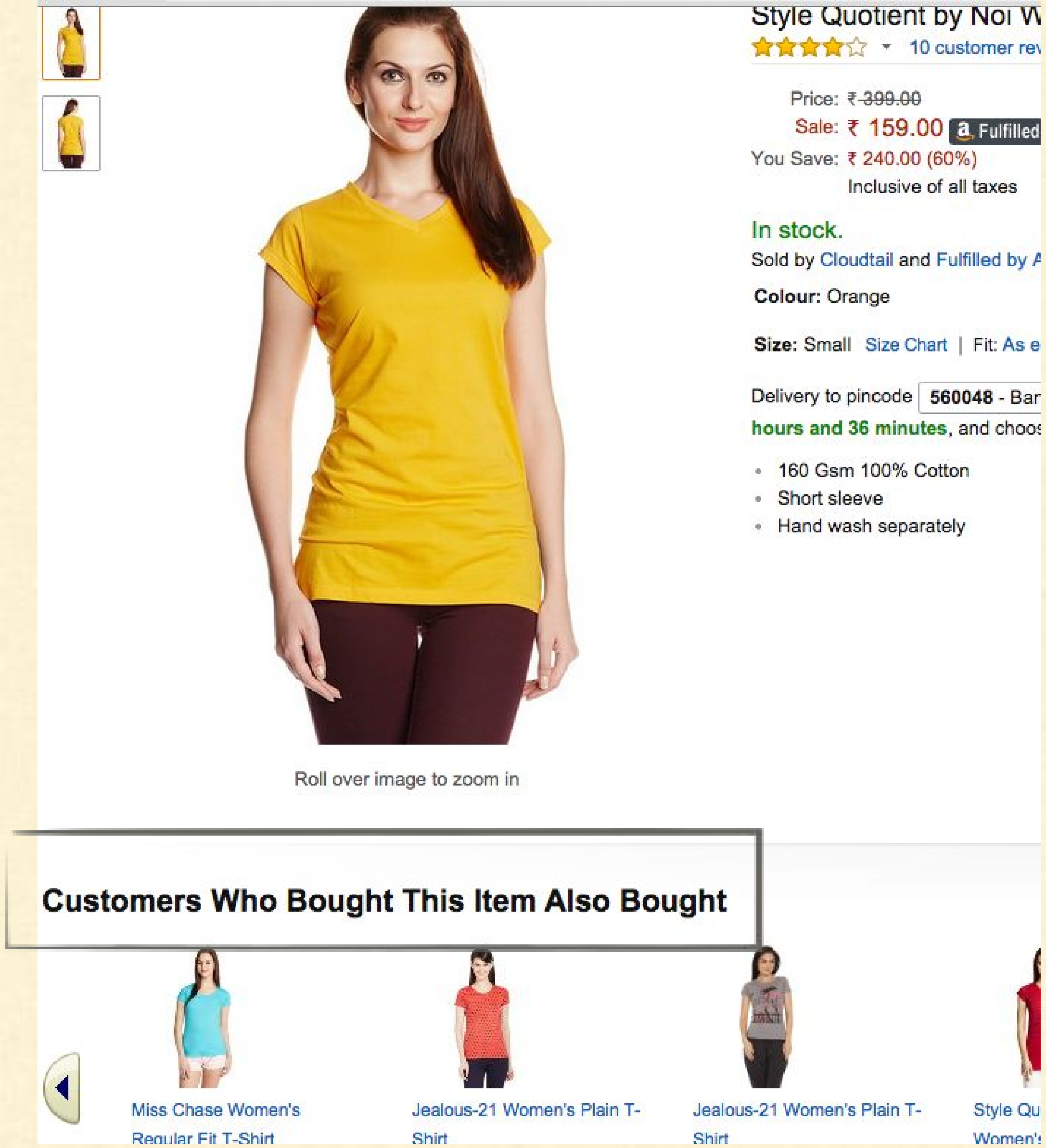
A. CPU
B. Memory Size
C. Memory Read Speed
D. Disk Speed
E. Disk Size
F. Network Speed
G. All of Above

Which Components Impact the Speed of Computing?

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

Example Big Data Customers

1. Ecommerce - Recommendations



Style Quotient by Noi W
★★★★☆ ▾ 10 customer rev

Price: ~~₹ 399.00~~
Sale: ₹ 159.00 a Fulfilled
You Save: ₹ 240.00 (60%)
Inclusive of all taxes

In stock.
Sold by Cloudtail and Fulfilled by A

Colour: Orange

Size: Small Size Chart | Fit: As e

Delivery to pincode 560048 - Bar
hours and 36 minutes, and choos

- 160 Gsm 100% Cotton
- Short sleeve
- Hand wash separately

Roll over image to zoom in

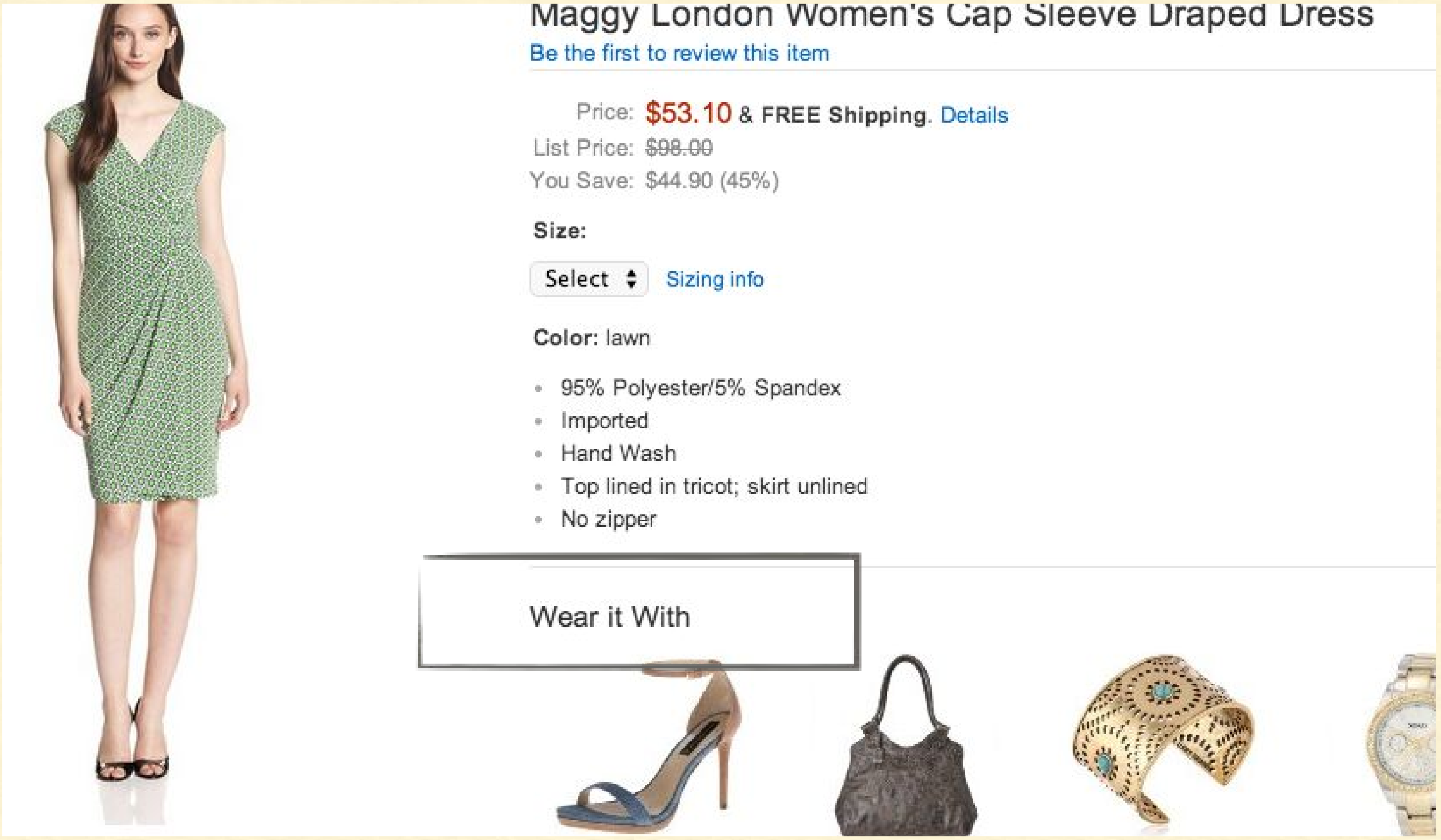
Customers Who Bought This Item Also Bought

Miss Chase Women's Regular Fit T-Shirt

Jealous-21 Women's Plain T-Shirt

Jealous-21 Women's Plain T-Shirt

Style Qu Women's



Maggy London Women's Cap Sleeve Draped Dress
Be the first to review this item

Price: $53.10 & **FREE Shipping**. Details
List Price: ~~$98.00~~
You Save: $44.90 (45%)

Size:

Select ⬍ Sizing info

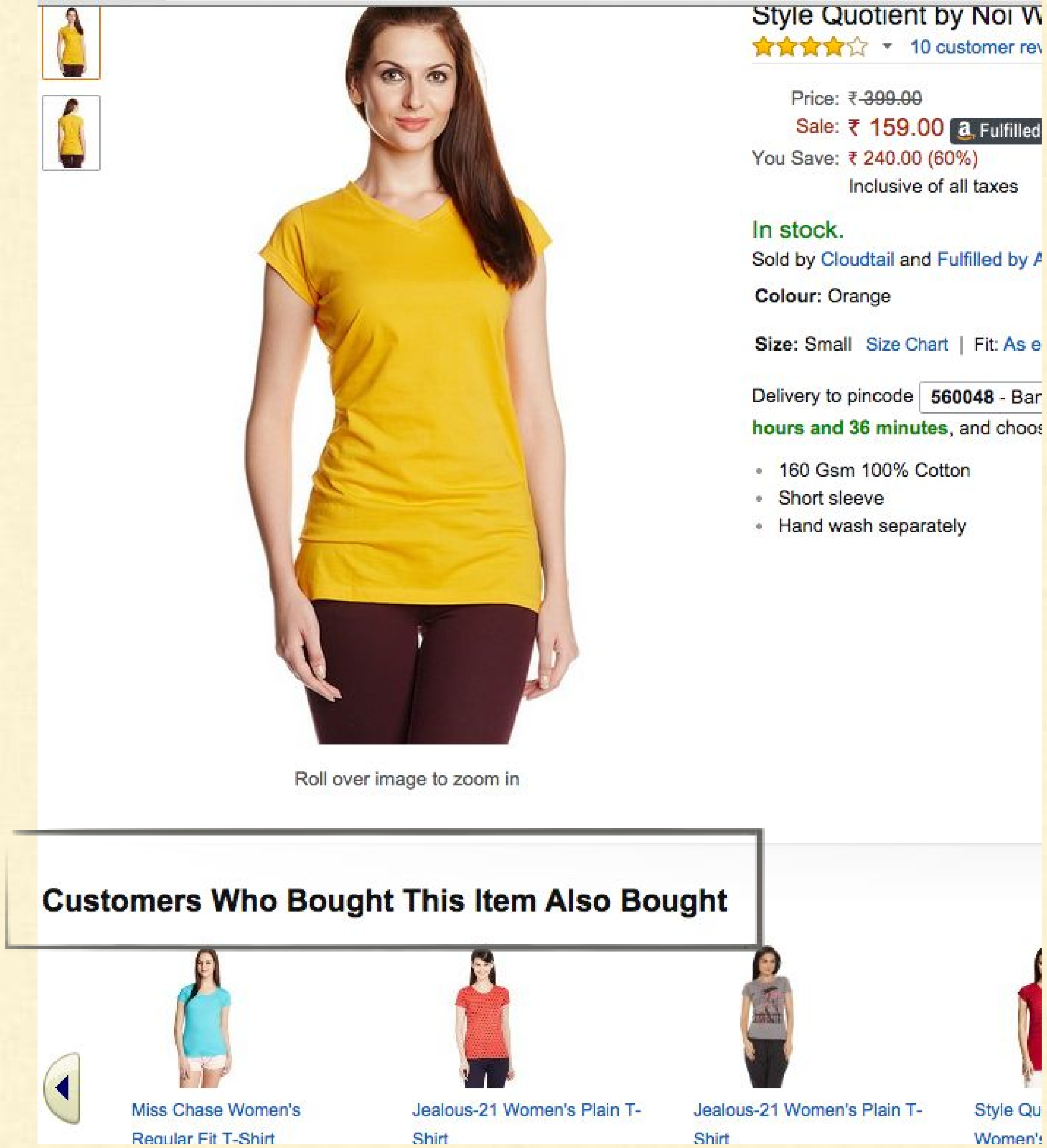
Color: lawn

- 95% Polyester/5% Spandex
- Imported
- Hand Wash
- Top lined in tricot; skirt unlined
- No zipper

Wear it With

Example Big Data Customers

1. Ecommerce - Recommendations



Style Quotient by Noi W

★★★★☆ 10 customer rev

Price: ~~₹ 399.00~~

Sale: ₹ 159.00 a Fulfilled

You Save: ₹ 240.00 (60%)

Inclusive of all taxes

In stock.

Sold by Cloudtail and Fulfilled by A

Colour: Orange

Size: Small Size Chart | Fit: As e

Delivery to pincode 560048 - Bar

hours and 36 minutes, and choos

- 160 Gsm 100% Cotton
- Short sleeve
- Hand wash separately

Roll over image to zoom in

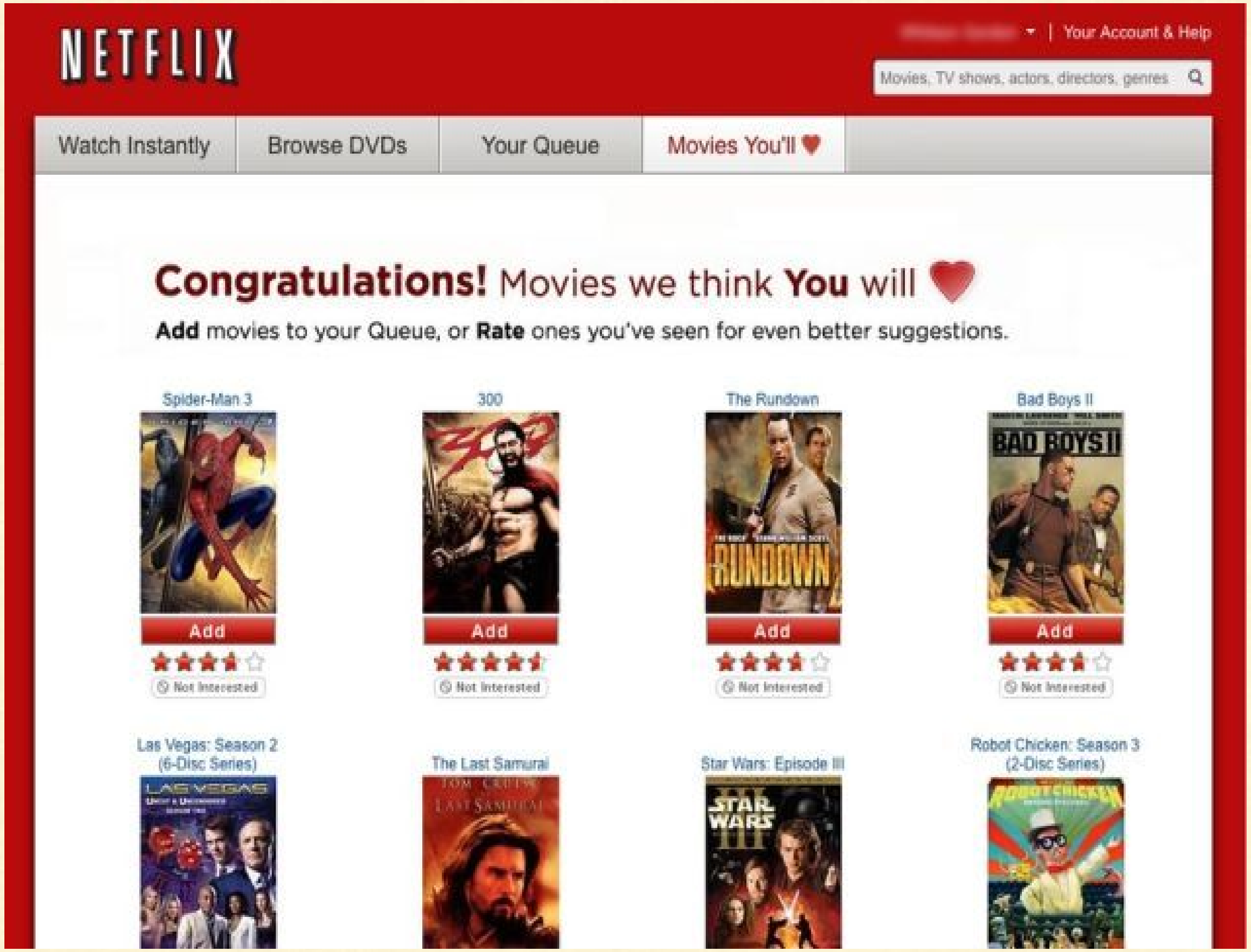
Customers Who Bought This Item Also Bought

Miss Chase Women's Regular Fit T-Shirt

Jealous-21 Women's Plain T-Shirt

Jealous-21 Women's Plain T-Shirt

Style Qu Women's



NETFLIX

[redacted] | Your Account & Help

Movies, TV shows, actors, directors, genres

Congratulations! Movies we think **You** will ♥

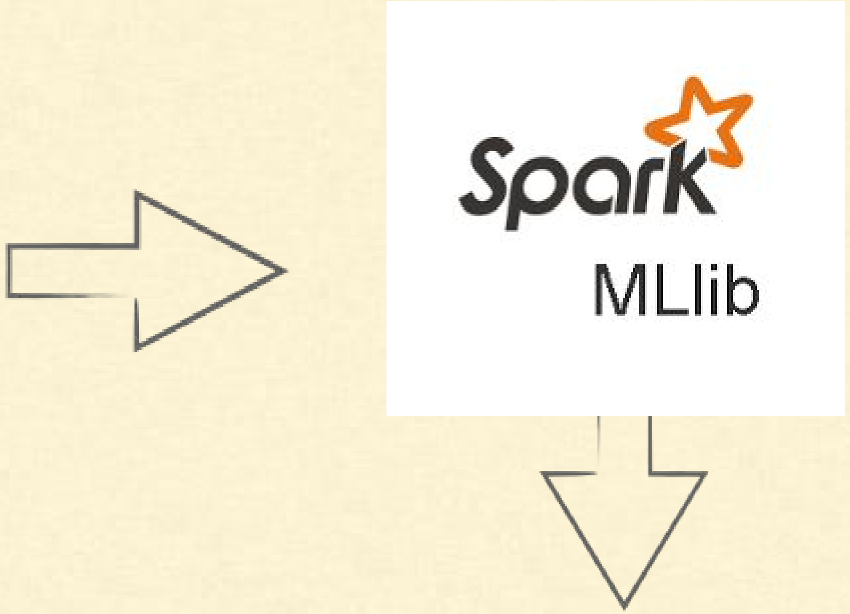
Add movies to your Queue, or **Rate** ones you've seen for even better suggestions.

Spider-Man 3	300	The Rundown	Bad Boys II
Add	Add	Add	Add
Not Interested	Not Interested	Not Interested	Not Interested
Las Vegas: Season 2 (6-Disc Series)	The Last Samurai	Star Wars: Episode III	Robot Chicken: Season 3 (2-Disc Series)

Example Big Data Problems

Recommendations - How?

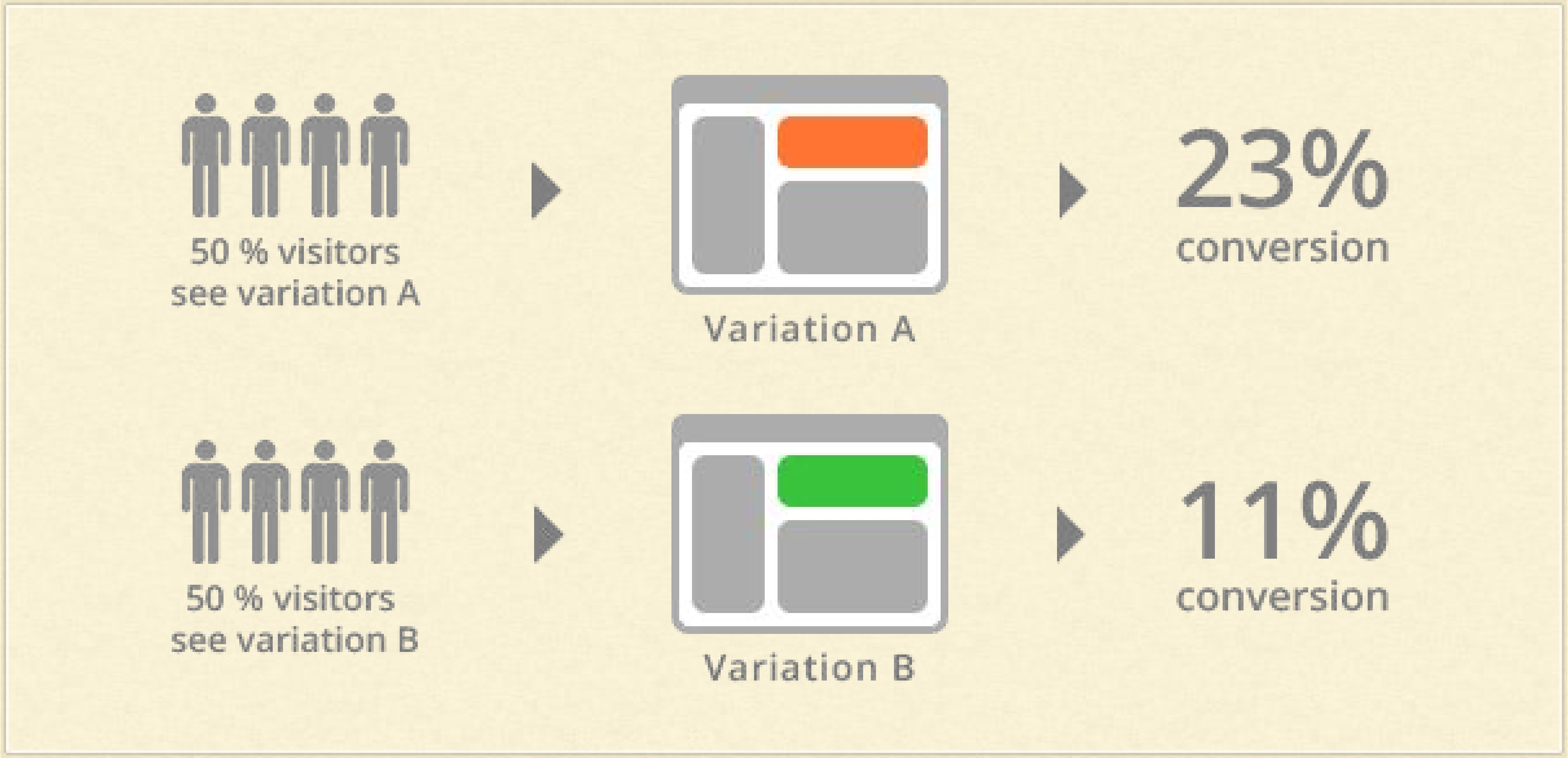
USER ID	MOVIE ID	RATING
KUMAR	matrix	4.0
KUMAR	Ice age	3.5
GIRI	apocalypse now	3.6
GIRI	Ice age	3.5



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

Example Big Data Customers

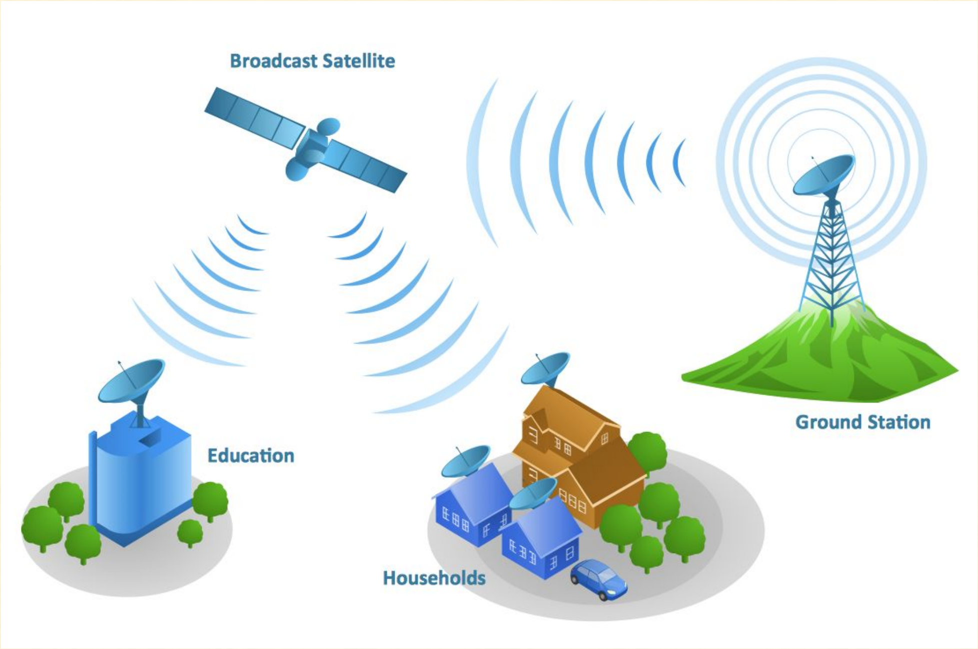
2. Ecommerce - A/B Testing



Big Data Customers

Government

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



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

Healthcare & Life Sciences

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

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)

B. Built for Nutch search engine project

C. Joined by Mike Cafarella

D. Based on GFS, GMR & Google Big Table

E. Named after Toy Elephant

F. Open Source - Apache

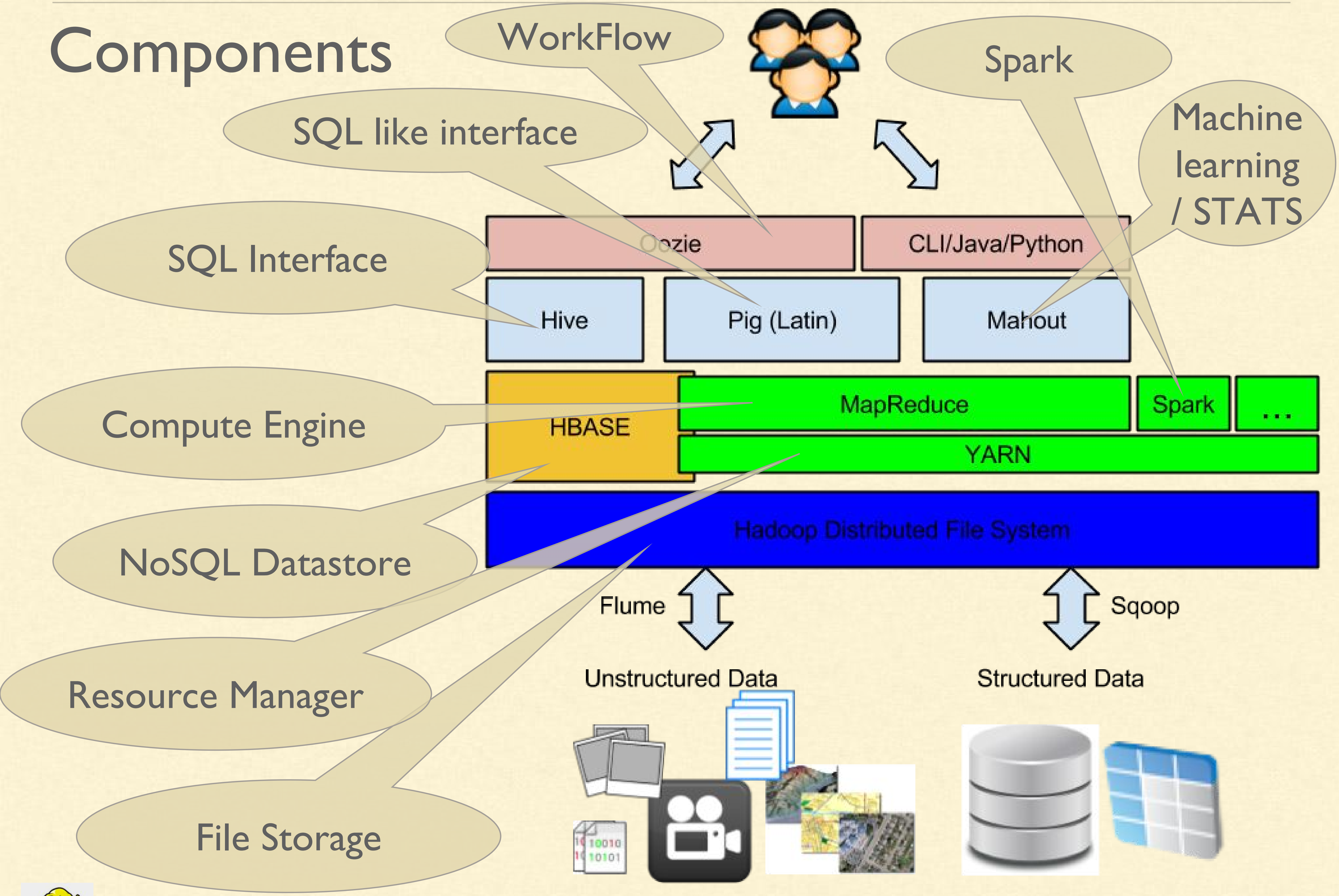
G. Powerful, Popular & Supported

H. Framework to handle Big Data

I. For distributed, scalable and reliable computing

J. Written in Java

Components

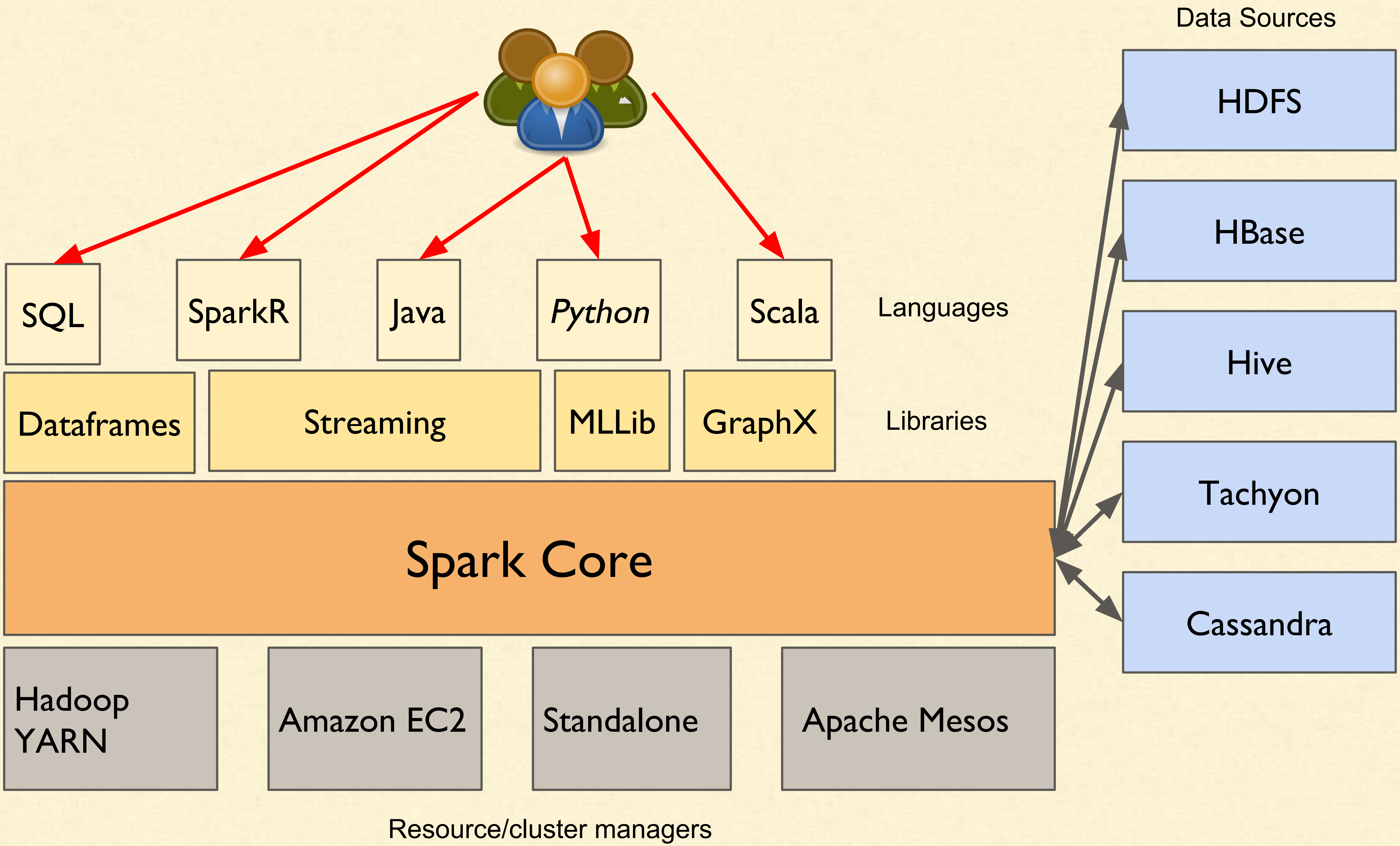

WorkFlow
SQL like interface
SQL Interface
Compute Engine
NoSQL Datastore
Resource Manager
File Storage
Spark
Machine learning / STATS
Oozie
CLI/Java/Python
Hive
Pig (Latin)
Mahout
HBASE
MapReduce
Spark
...
YARN
Hadoop Distributed File System
Flume
Sqoop
Unstructured Data
Structured Data
10010
10101

Apache Spark

- Really fast MapReduce
  - 100x faster than Hadoop MapReduce in memory,
  - 10x faster on disk.
- Builds on similar paradigms as MapReduce
- Integrated with Hadoop

Spark Core - A fast and general engine for large-scale data processing.

Spark Architecture



Data Sources

HDFS

HBase

Hive

Tachyon

Cassandra

SQL

SparkR

Java

Python

Scala

Languages

Dataframes

Streaming

MLLib

GraphX

Libraries

Spark Core

Hadoop YARN

Amazon EC2

Standalone

Apache Mesos

Resource/cluster managers