Welcome to

Apache Spark

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: Understanding Spark
- 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 11 sessions on Big Data & Spark course. It suffices as an intro to Big Data Tech.

WELCOME - KNOWBIGDATA

- ❑ Expert Instructors
- ❑ CloudxLabs
- ❑ 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	Introduction to Big Data with Apache Spark
II	Downloading Spark and Getting Started
III	Programming with RDDs
IV	Working with Key/Value Pairs
V	Loading and Saving Your Data
VI	Advanced Spark Programming
VII	Running on a Cluster
VIII	Tuning and Debugging Spark
IX	Spark SQL, SparkR
X	Spark Streaming
XI	Machine Learning with MLlib, GraphX

About Instructor?

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

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

DISTRIBUTED COMPUTING

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

DISTRIBUTED COMPUTING

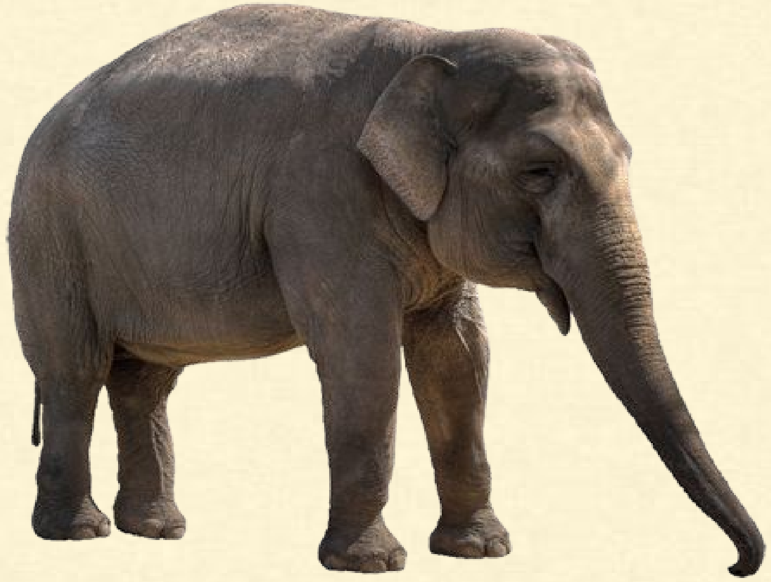
Take the code to the data.

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

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.

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

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.

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

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.

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?

$1.1259x10^{\wedge 15}$

WHY IS IT IMPORTANT NOW?

Smart Phones

Connectivity:
Internet Of Things

Connectivity:
Social Networks

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

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. Processor
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. Processor

B. Memory

C. Memory Read Speed

D. Disk Speed

E. Disk Size

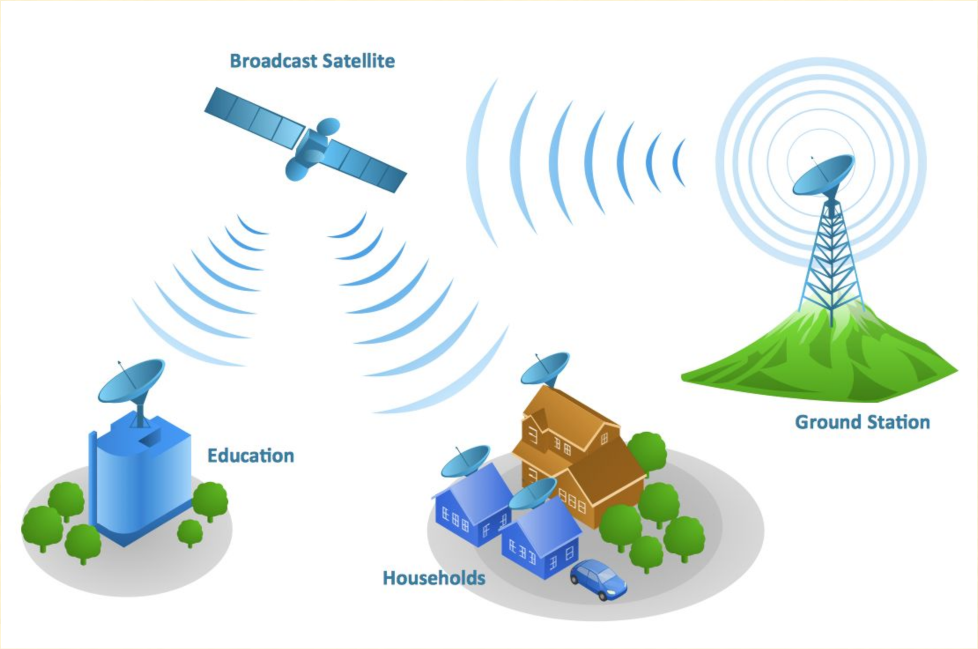
F. Network Speed

✓ G. All of Above

EXAMPLE BIG DATA CUSTOMERS

Web and e-commerce

1. Recommendation Engines
2. Analytics
3. Predicting demand

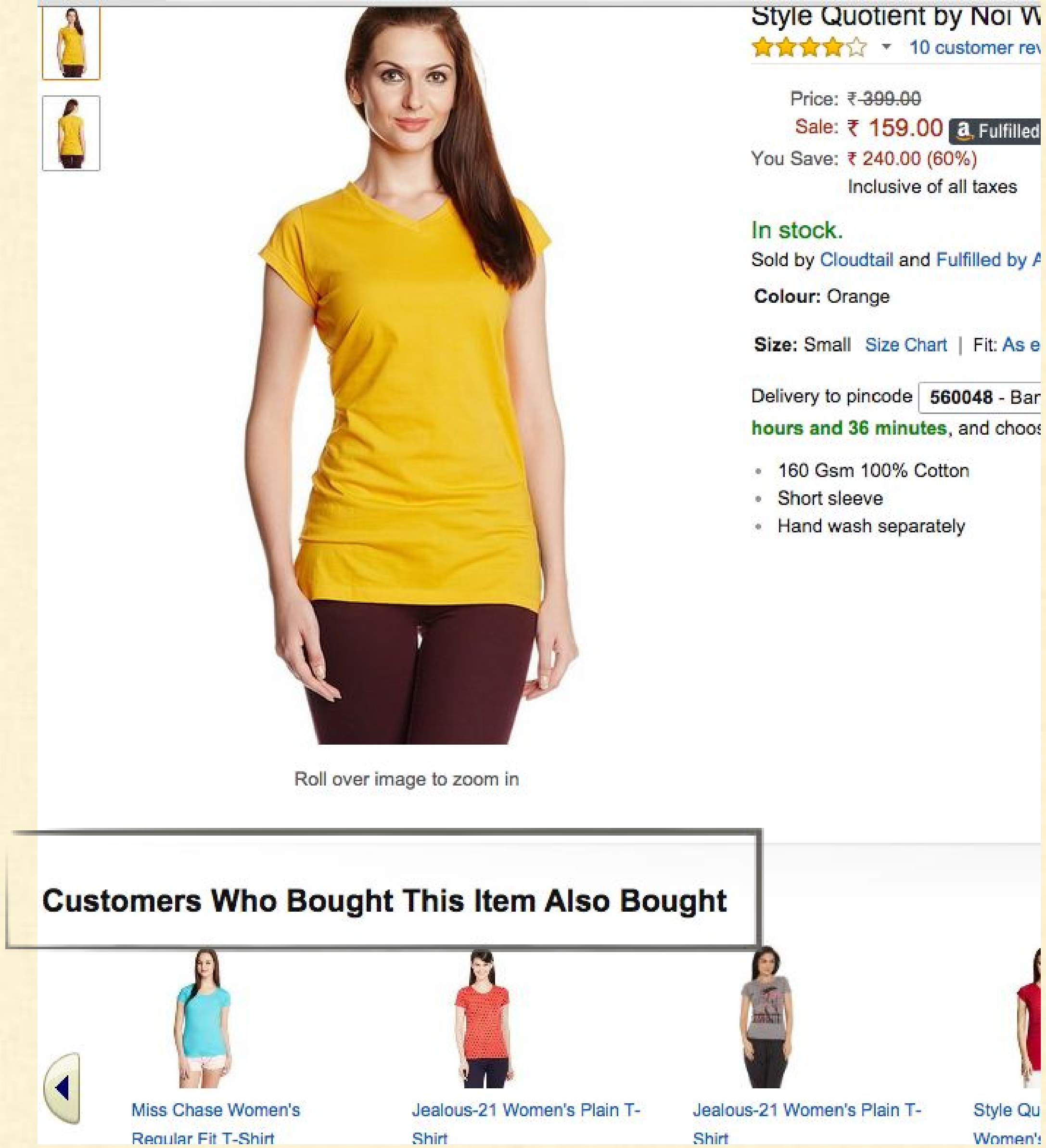


Telecommunications

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

EXAMPLE BIG DATA PROBLEMS

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

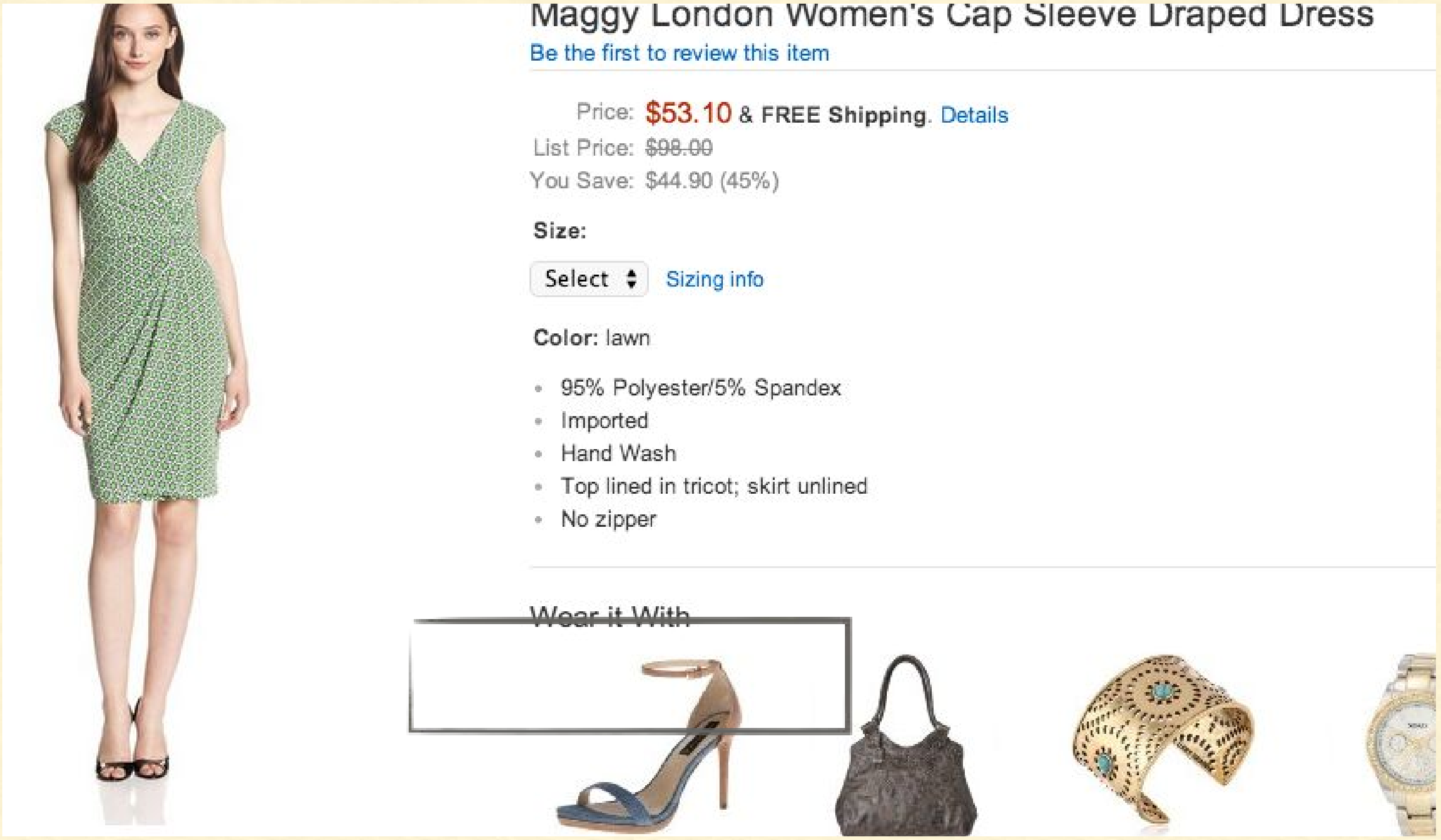
Customers Who Bought This Item Also Bought

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

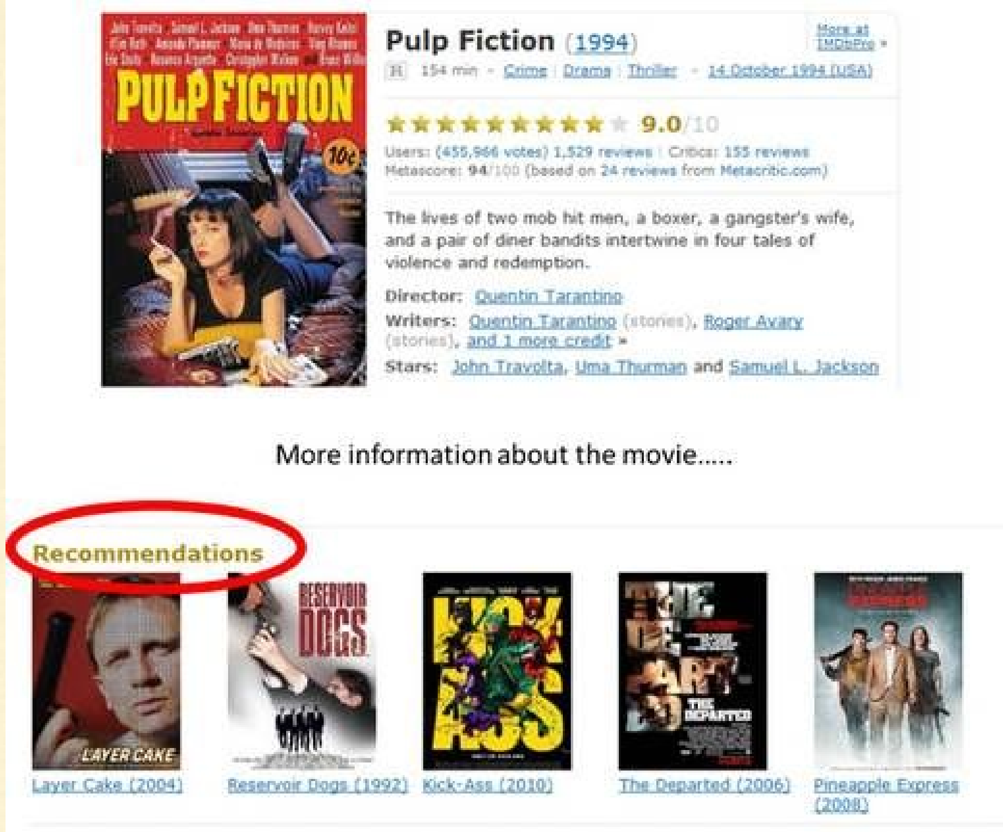
Color: lawn

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

Wear it With

EXAMPLE BIG DATA PROBLEMS

Recommendations



Pulp Fiction (1994)

More at IMDbPro »

R 154 min - Crime | Drama | Thriller - 14 October 1994 (USA)

★★★★★★★★★☆ 9.0/10

Users: (455,966 votes) 1,529 reviews | Critics: 155 reviews

Metascore: 94/100 (based on 24 reviews from Metacritic.com)

The lives of two mob hit men, a boxer, a gangster's wife, and a pair of diner bandits intertwine in four tales of violence and redemption.

Director: Quentin Tarantino

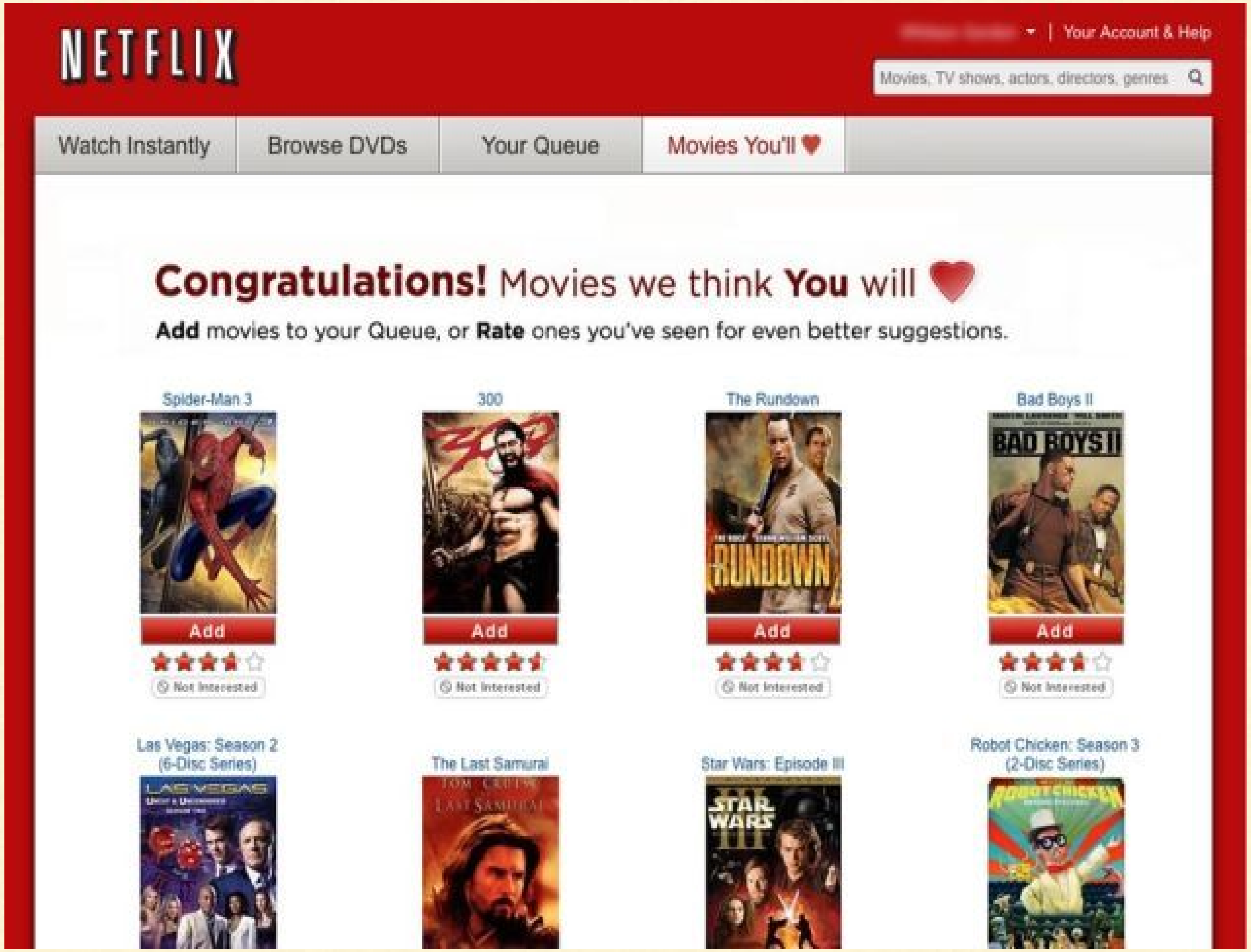
Writers: Quentin Tarantino (stories), Roger Avary (stories), and 1 more credit »

Stars: John Travolta, Uma Thurman and Samuel L. Jackson

More information about the movie.....

Recommendations

Layer Cake (2004) · Reservoir Dogs (1992) · Kick-Ass (2010) · The Departed (2006) · Pineapple Express (2008)



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

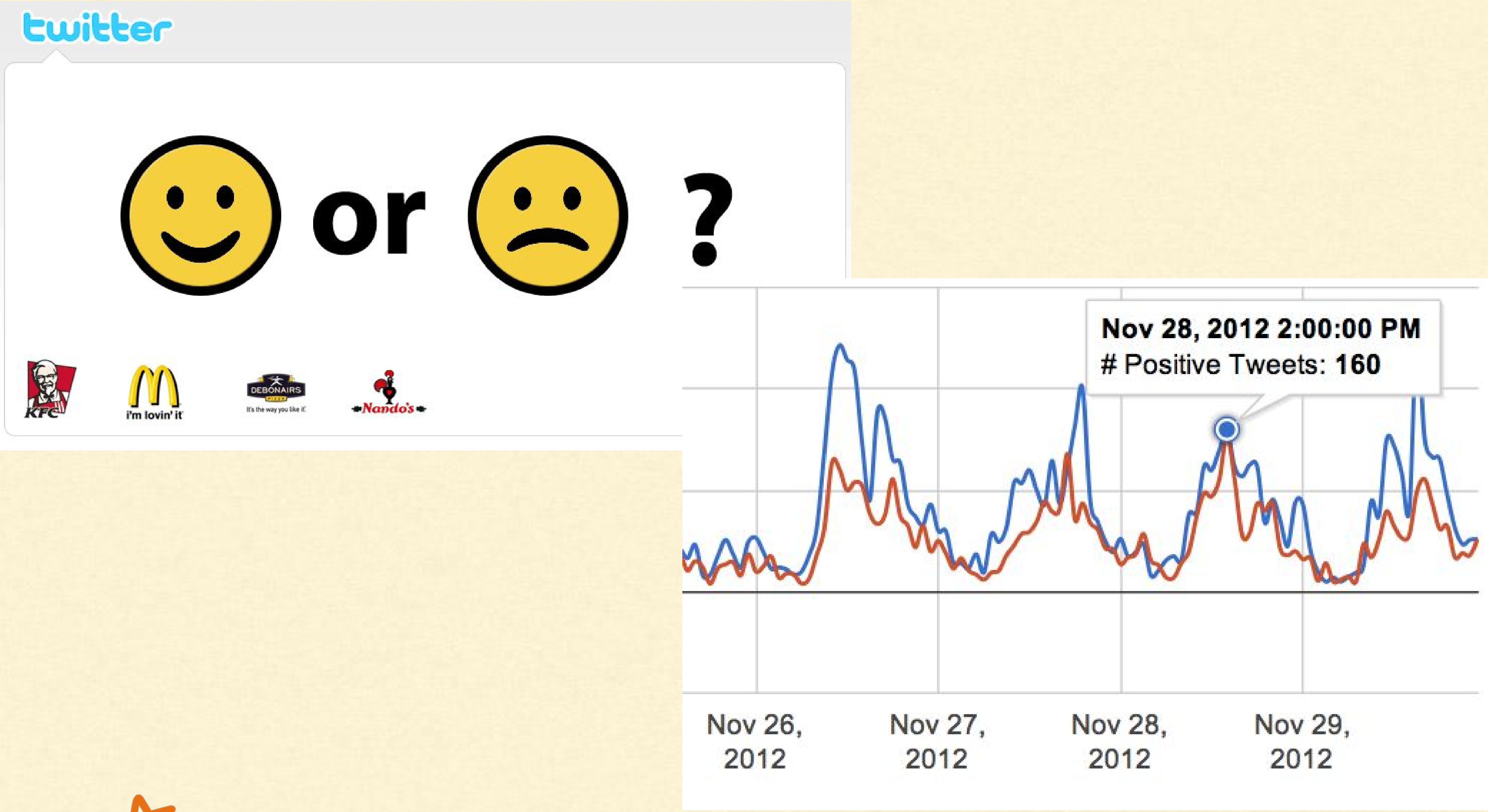
Add movies to your Queue, or **Rate** ones you've seen for even better suggestions.

Spider-Man 3 · 300 · The Rundown · Bad Boys II

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

Sentiment Analysis



EXAMPLE BIG DATA CUSTOMERS

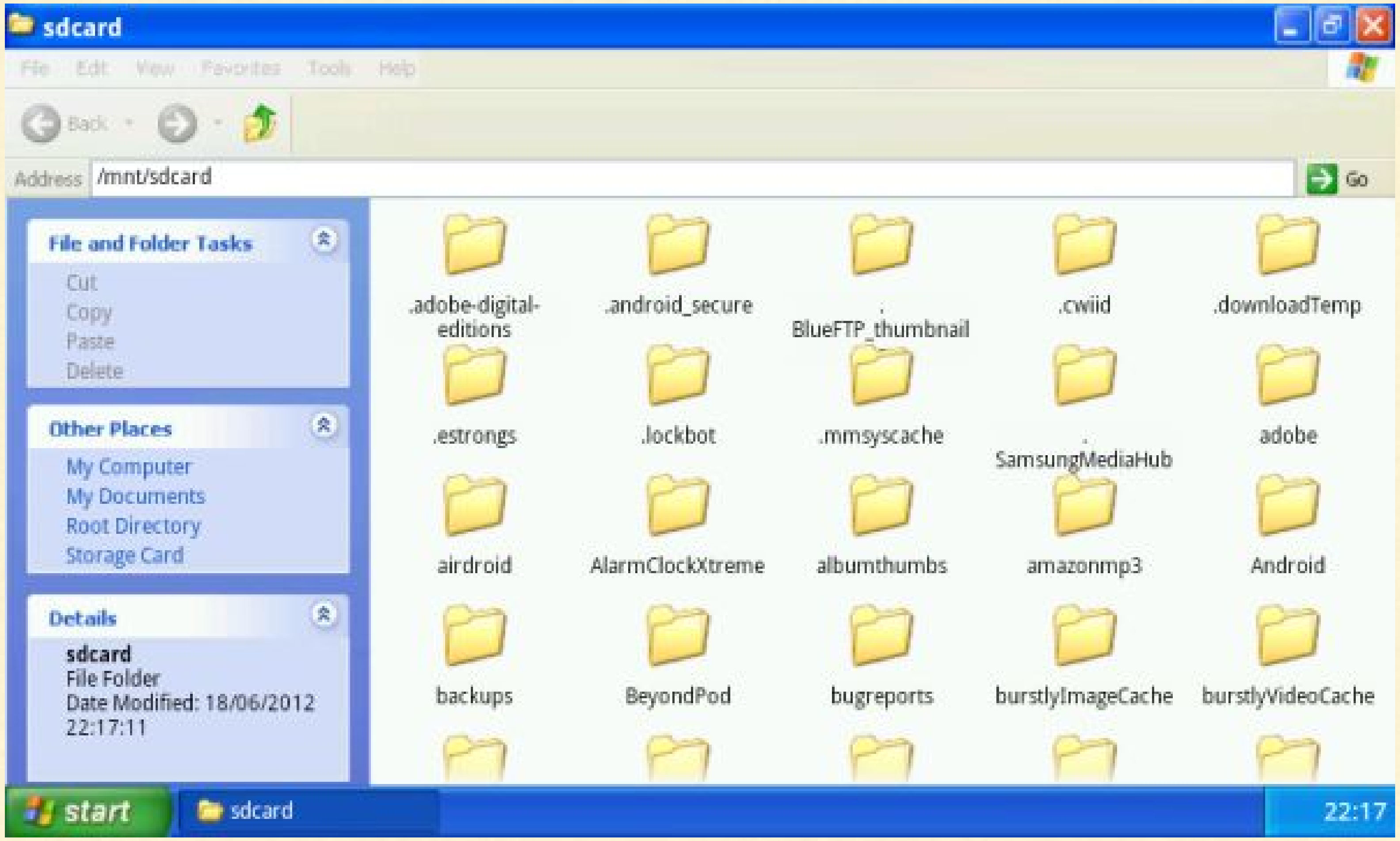
Government

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

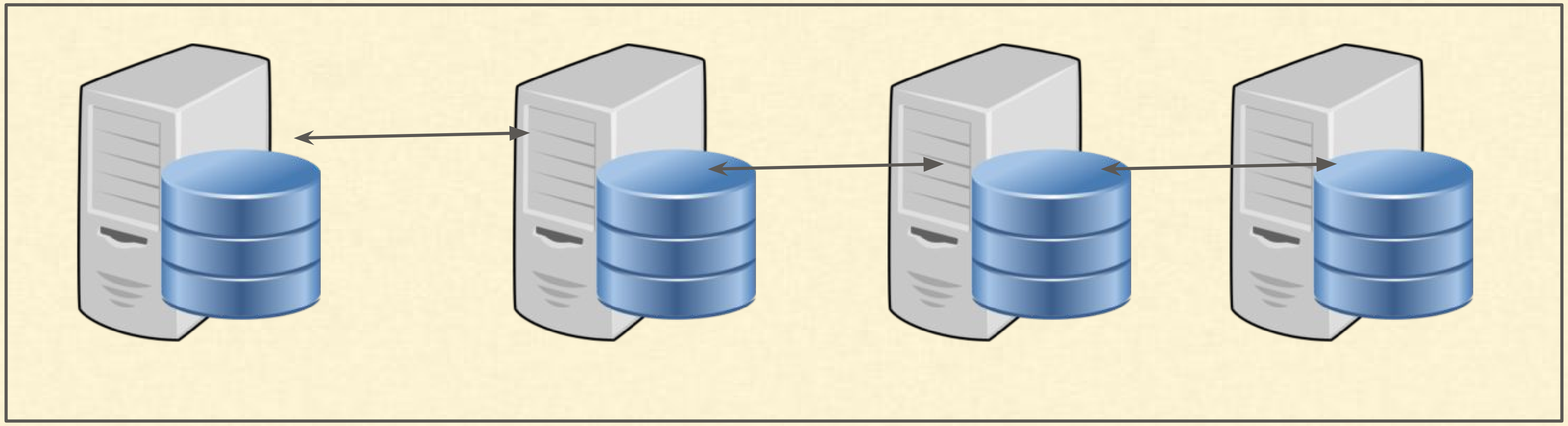
Solving Storage Problem - HDFS



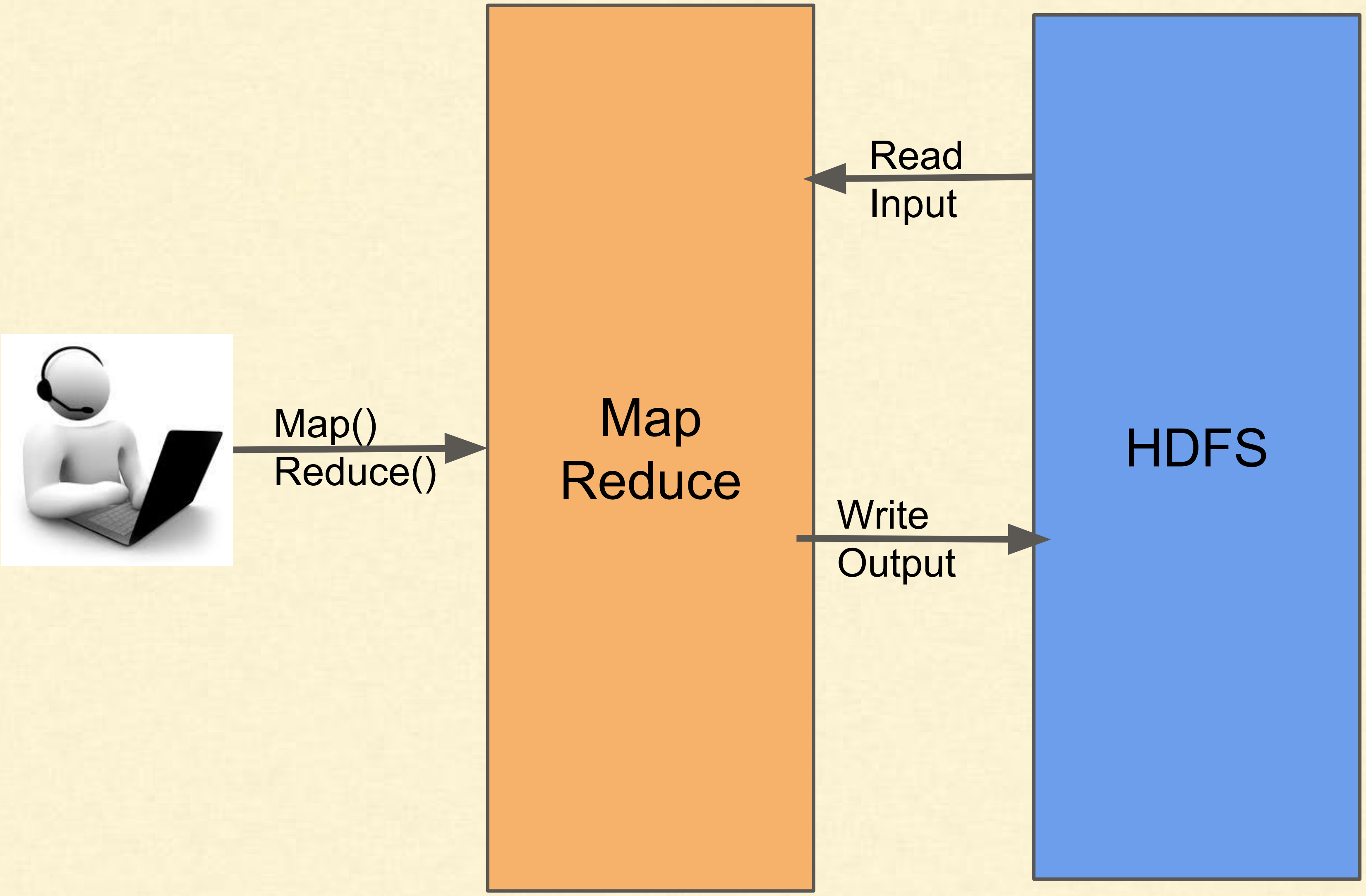
Address /mnt/sdcard

Details
sdcard
File Folder
Date Modified: 18/06/2012 22:17:11

- Uses Many Disks
- Of Many Computers
- Over network
- To Provide
- Scalable
- Fault Tolerant
- Simple Storage

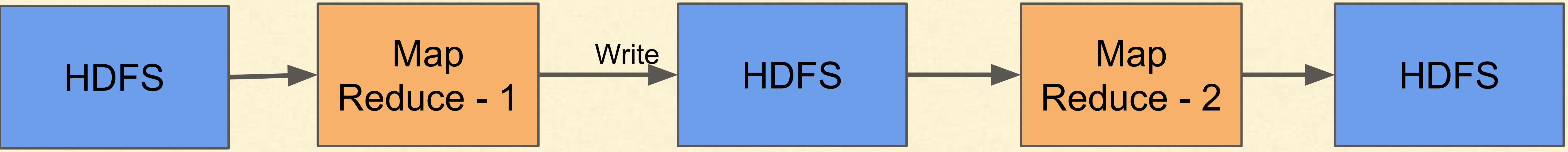


Hadoop Map Reduce



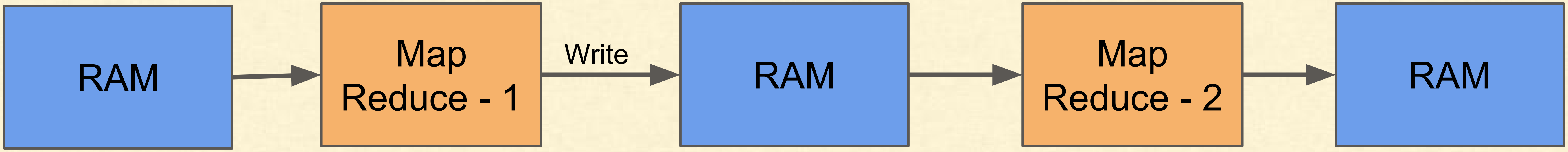
- User Sends Logic
- In form of Map() & Reduces
- Tries to do execute near data
- Saves result to HDFS

Hadoop Map Reduce - Multiple Phases



Shortcoming of Map Reduce

1. Batchwise Design
    a. Every map-reduce cycle reads from and writes to HDFS
    b. Heavy Latency
2. Converting logic to Map-Reduce paradigm is difficult
3. In-memory computing was not possible



80 times faster than disk

See: Latency Numbers Every Programmer Should Know

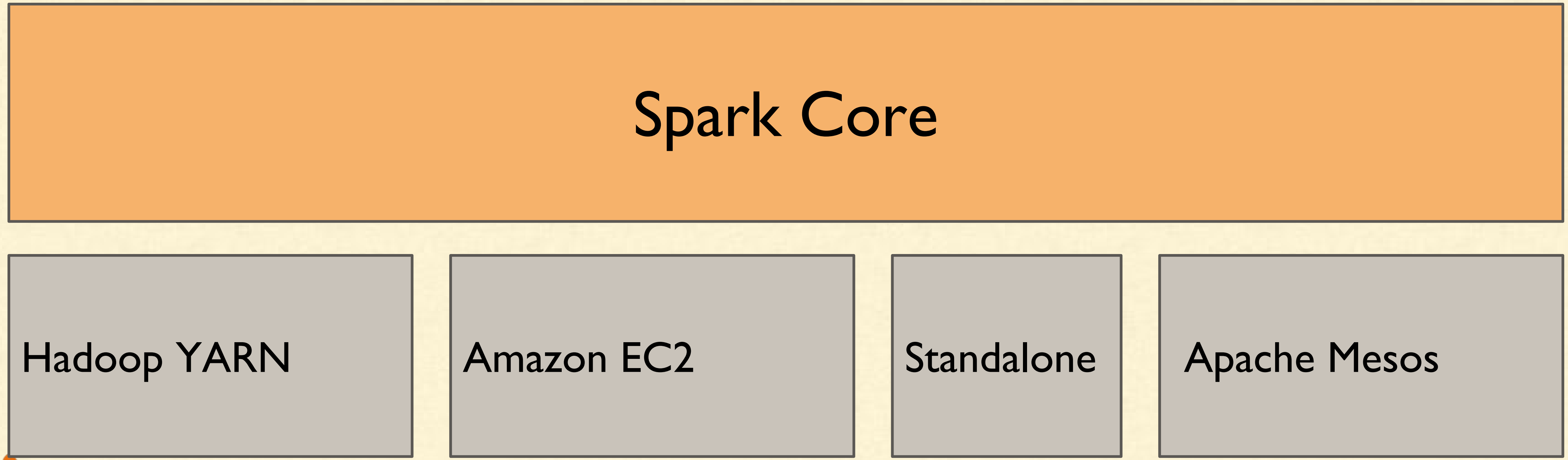
Apache Spark

- 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.

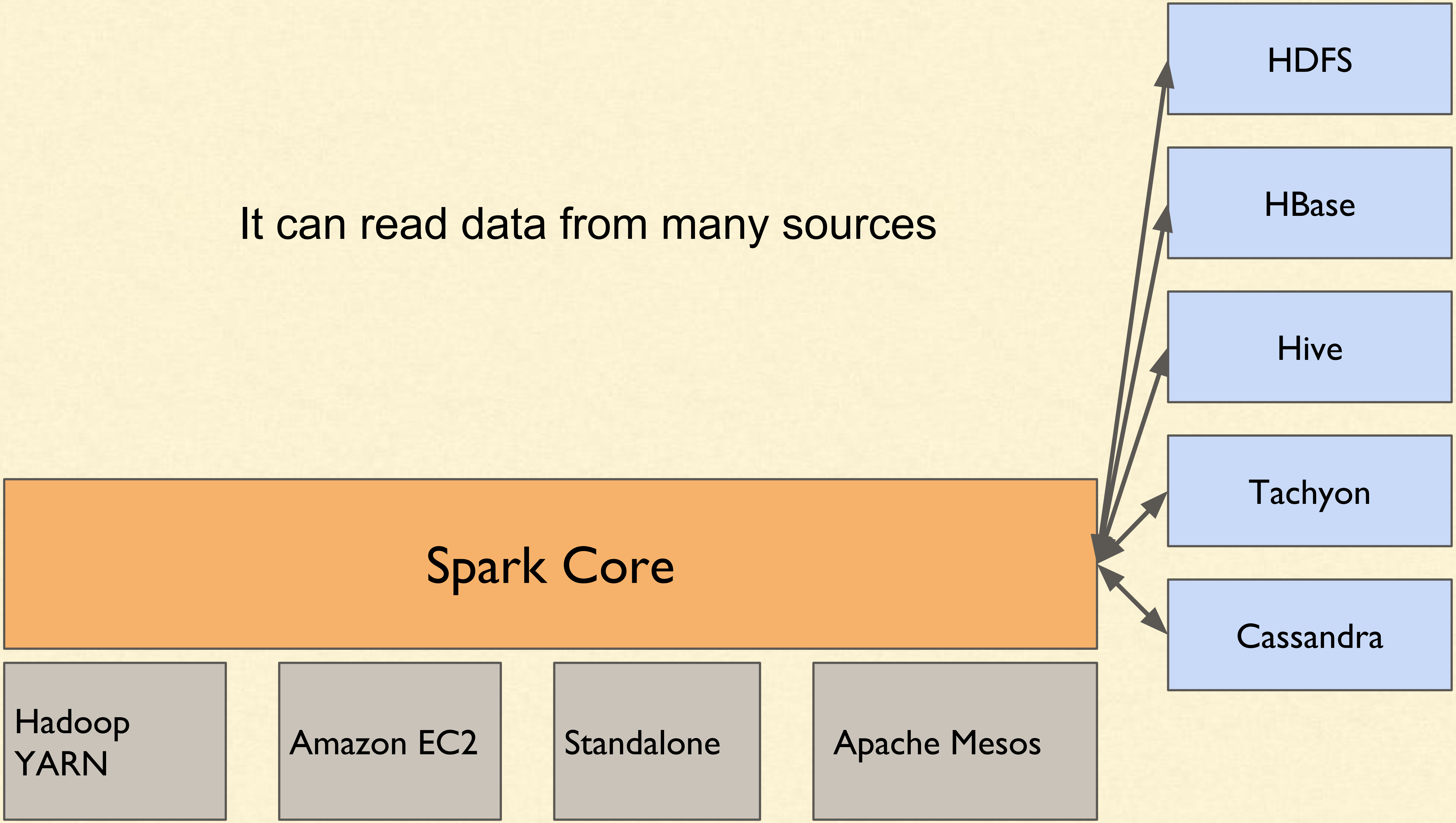
Spark Architecture

It can run on almost all popular cluster resource managers.

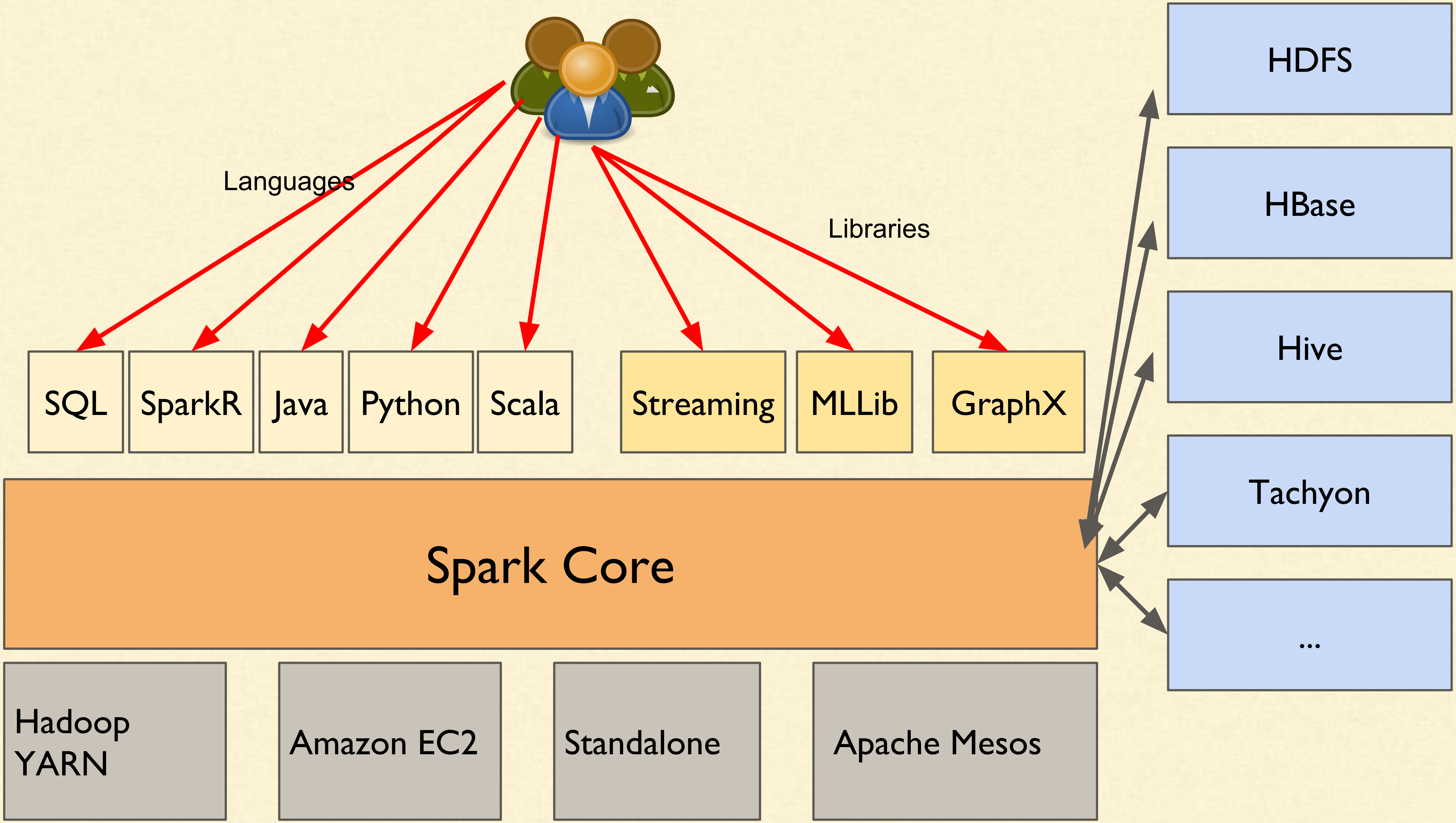


Spark Architecture

It can read data from many sources

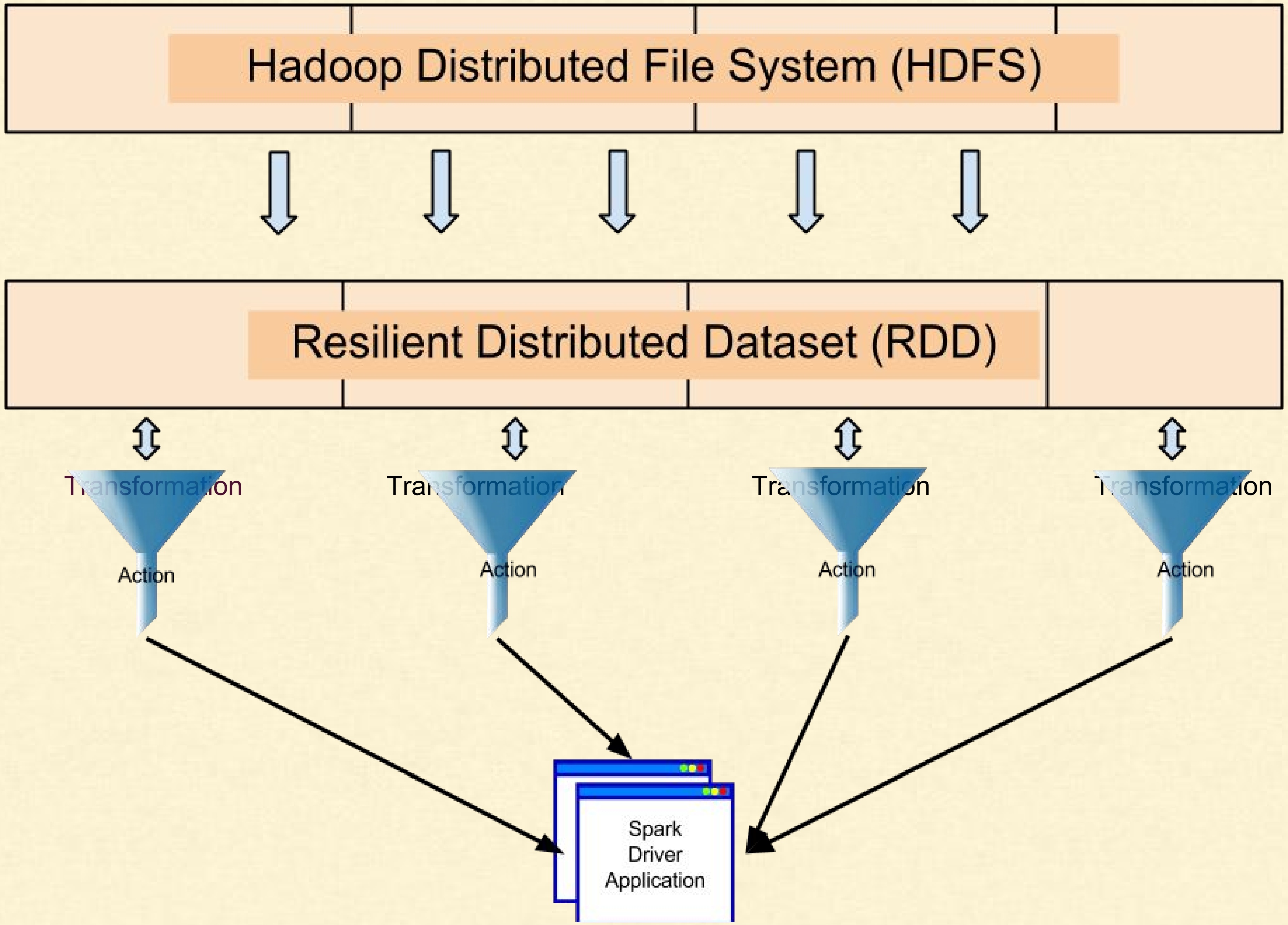


Spark Architecture



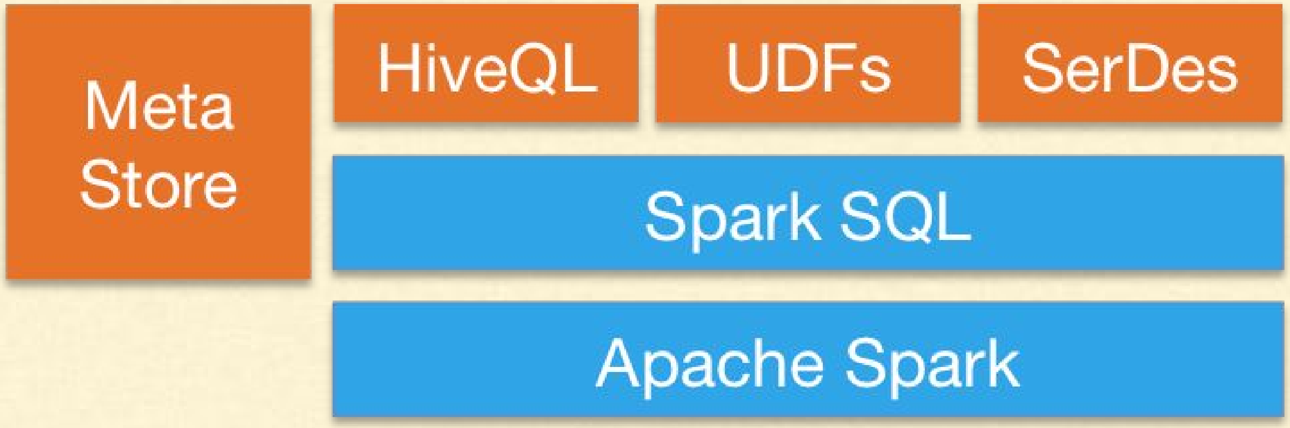
Spark Architecture - Core

RDD is a distributed data set on which either we can run actions or transformations.



Spark SQL

- Hive Compatibility
- Standard Connectivity
  - JDBC / ODBC
- Performance & Scalability



Meta Store

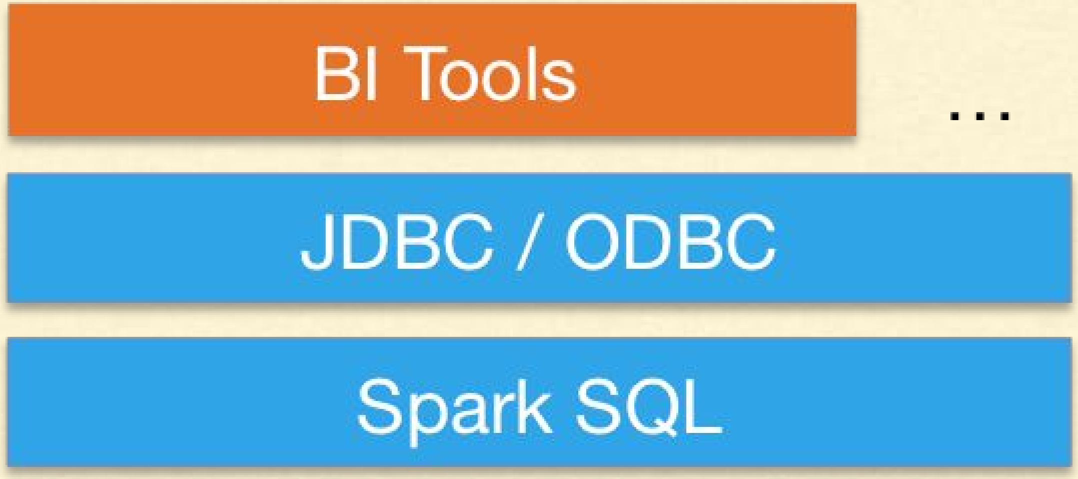
HiveQL

UDFs

SerDes

Spark SQL

Apache Spark



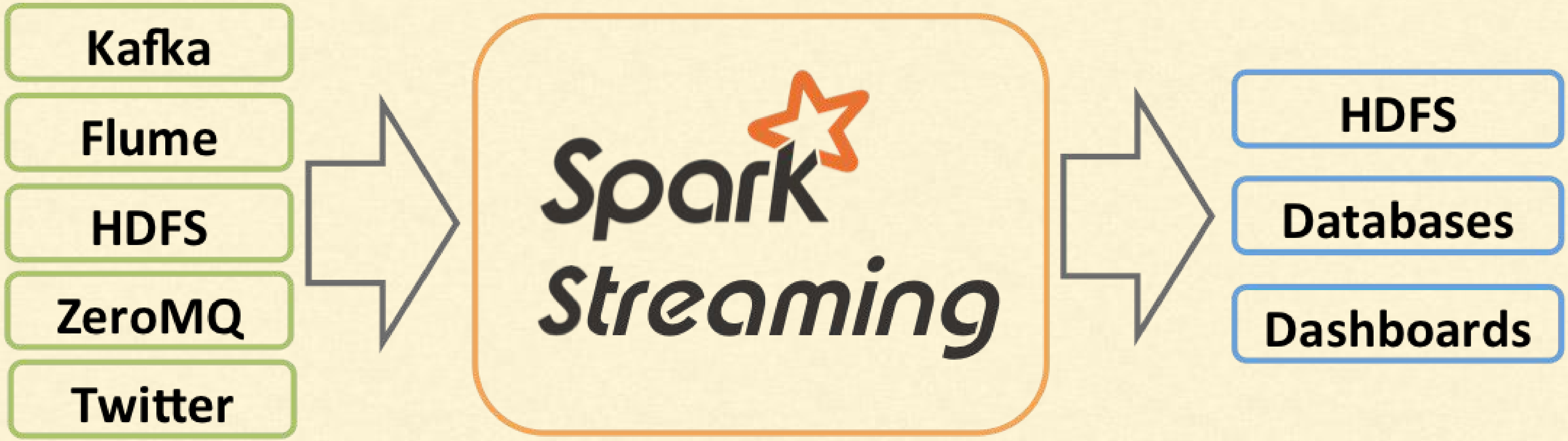
BI Tools

…

JDBC / ODBC

Spark SQL

Spark Streaming



Example: **Show the sentiment on twitter in realtime.**

MLLib - What is Machine Learning?

“Programming Computers to optimize a Performance using Example Data or Past Experience”

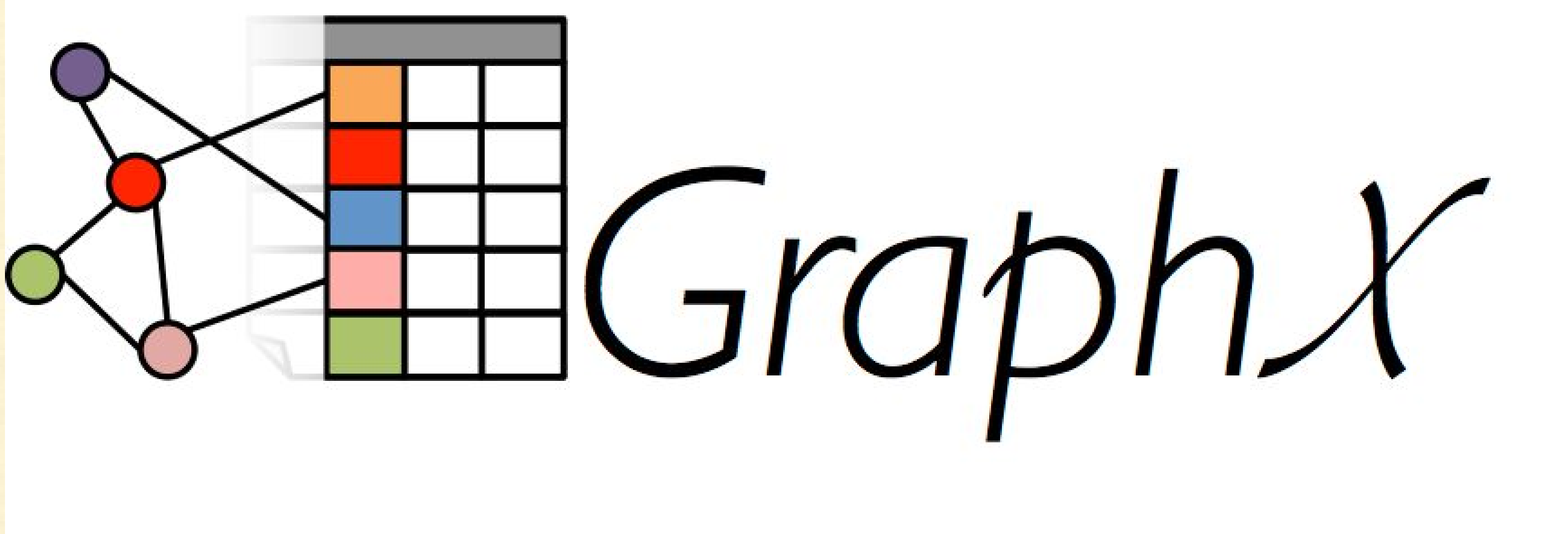
- Branch of Artificial Intelligence
- Design and Development of Algorithms
- Computers Evolve Behaviour based on Empirical Data

MLLib - Machine Learning Applications

- Recommend Friends, Dates, Products to end-user.
- Classify content into pre-defined groups.
- Find Similar content based on Object Properties.
- Identify key topics in large Collections of Text.
- Detect Anomalies within given data.
- Ranking Search Results with User Feedback Learning.
- Classifying DNA sequences.
- Sentiment Analysis/ Opinion Mining
- Computer Vision.
- Natural Language Processing,
- BioInformatics.
- Speech and HandWriting Recognition.

MLLib - Scalable machine learning library

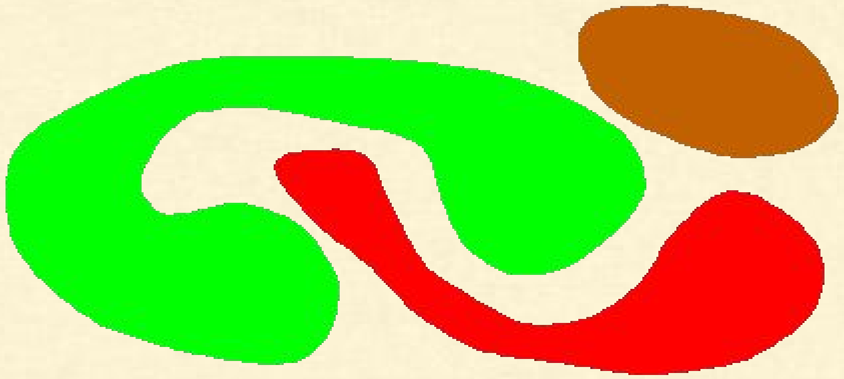
- Ease of Use
  - Usable in Java, Scala and Python.
- Performance
- High-quality algorithms
- High Level APIs for ML Pipelines



- Flexibility
    - Seamlessly work with both graphs and collections.
- Speed
    - Comparable performance to the fastest specialized graph processing systems.
- Algorithms
    - Choose from a growing library of graph algorithms.
- Community

GraphX - Algorithms

- PageRank
- Connected components
- Label propagation
- SVD++
- Strongly connected components
- Triangle count



SparkR - R on Spark

1. Provides dataframe like structure
2. Lets you import data from R
3. Have rich operations such as group by, filter by etc.
4. Overcomes the memory limitations of R
5. Run SQL Queries on R Dataframe
6. SparkR allows using existing R packages

FULL COURSE - Big Data with Spark

www.KnowBigData.com

1. Upcoming Sessions
   - 17 Oct, 8:30pm-11:30pm IST, SAT-SUN
2. 33 hrs - 3 hr x 11 classes
3. ₹24999 (25% off) (Incl. Taxes) - $399
4. Includes CloudLabs + Support + LMS
5. Every class is recorded.

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

reachus@KnowBigData.com

Apache Spark

Thank you.

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

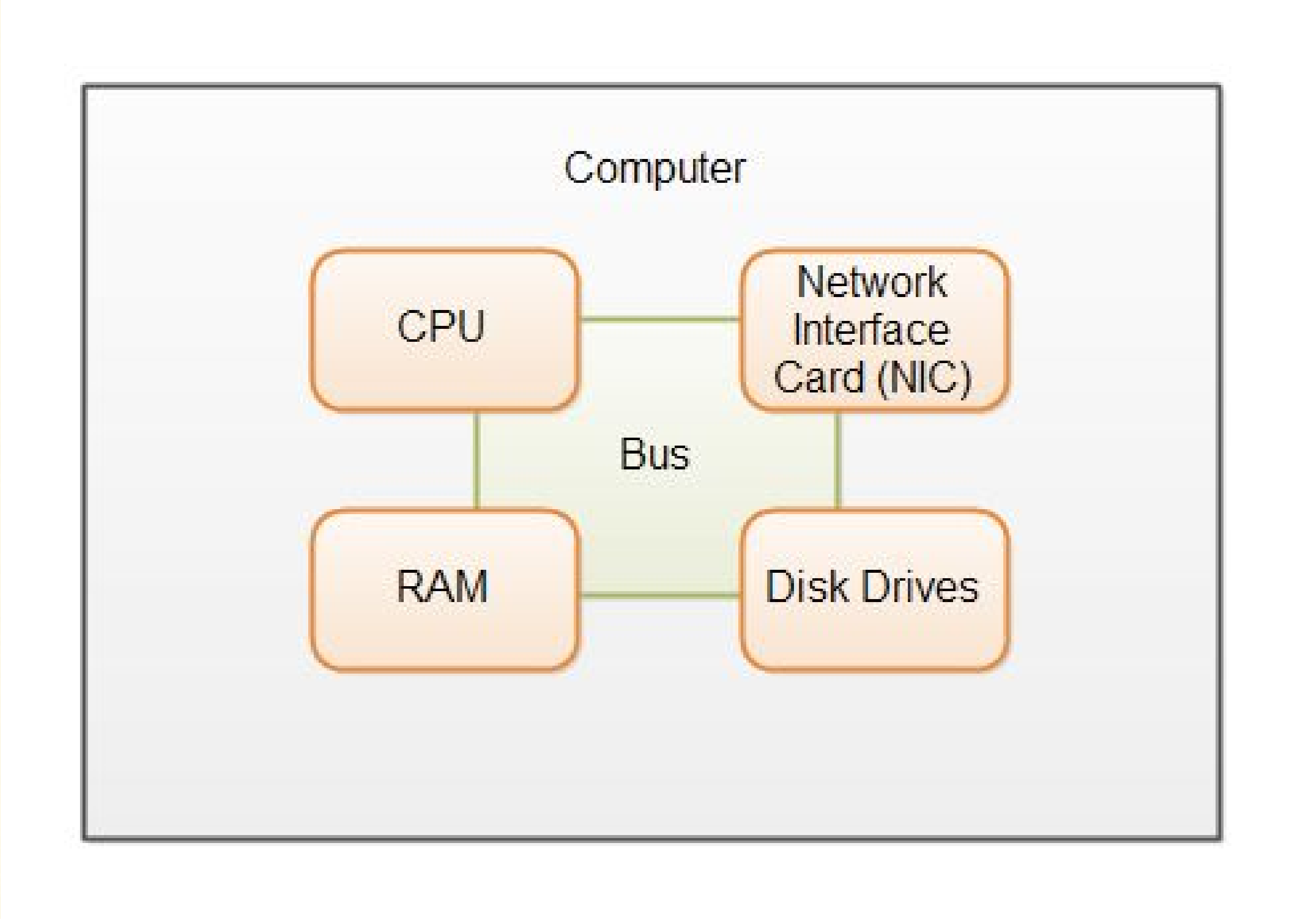
reachus@knowbigdata.com

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

BIG DATA PROBLEM

1. CPU Speed

To process & store data we need



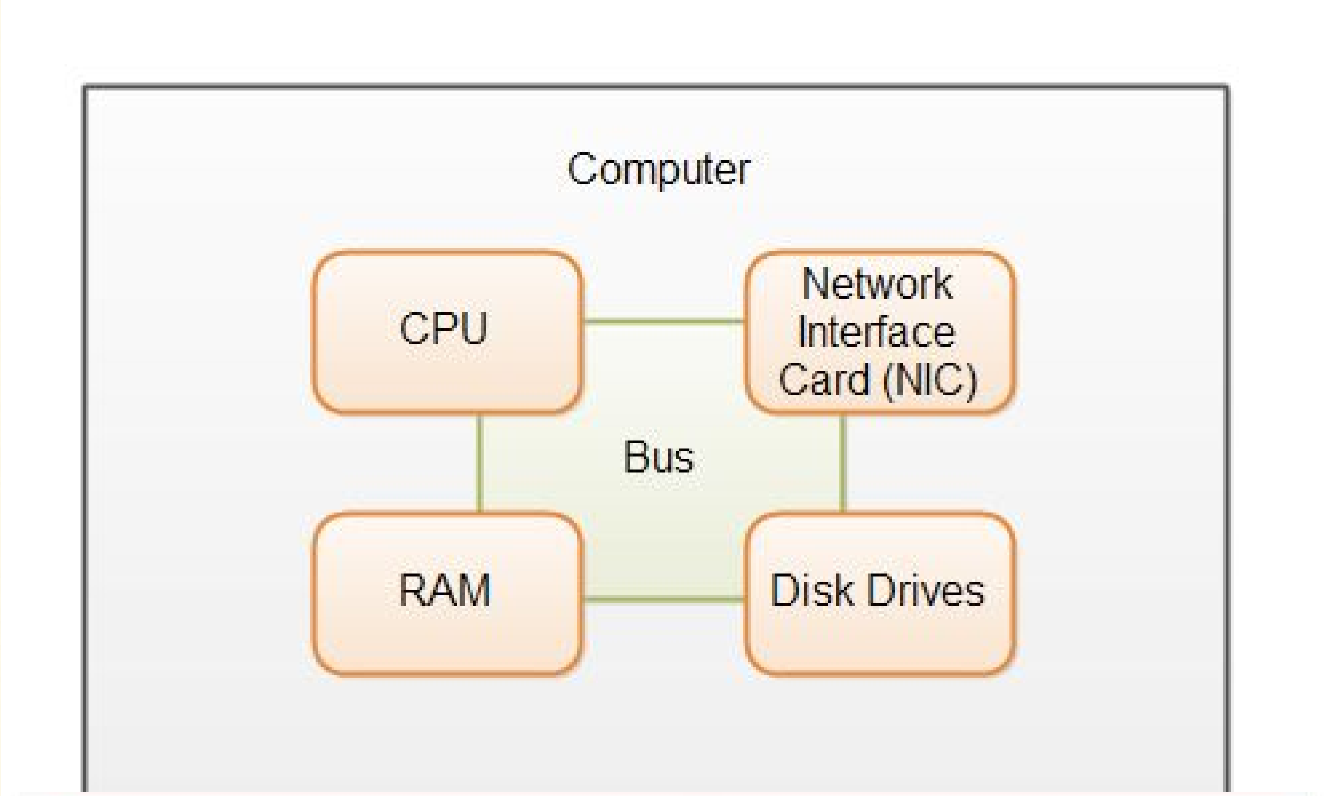
4. Network

2. RAM - Speed & Size

3. Disk Size + Speed

BIG DATA PROBLEM

To process & store data we need



1. CPU Speed

4. Network

And at least one of these become bottle neck

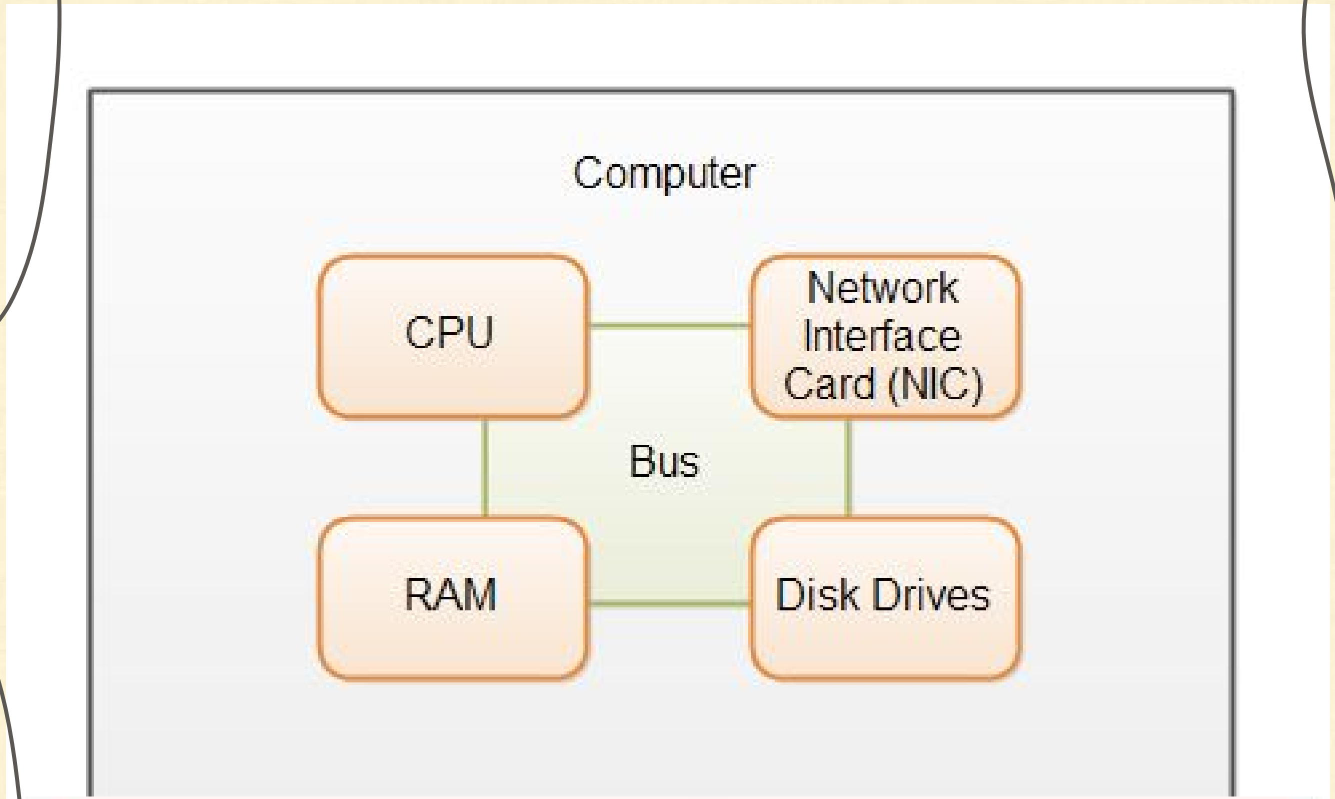
2. RAM - Speed & Size

3. Disk Size + Speed

BIG DATA PROBLEM

To process & store data we need

1. CPU Speed



Computer

CPU

Network Interface Card (NIC)

Bus

RAM

Disk Drives

And at least one of these become bottle neck

4. Network

2. RAM - Speed & Size

3. Disk Size + Speed