

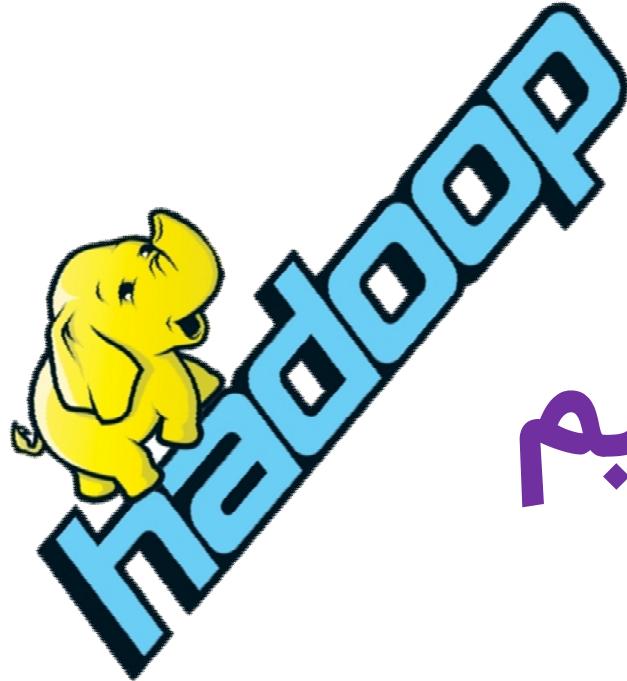
# دوره آموزشی

## مهندسی داده [Data Engineer]

مدرس: مجتبی بنائی



جلسه: سوم



# جلسه سوم

# آشنایی با مفاهیم

# کلانداده

کار عملی با هدوپ

# آنچه خواهیم دید

- جلسه اول : مفاهیم پایه کلان داده / هدوب
- جلسه دوم : آشنایی و کار عملی با Hive و HBase
- جلسه سوم : اکوسیستم هدوب و بنیاد آپاچی



# زمان بندی

- مفاهیم پایه کلان داده
- آشنایی با هدوب
- کارگاه عملی هدوب



# کلان داده



# دلیل اصلی ترافیک کلان شهرها

منابع شهری جو جو عظیم  
و سایر وسائل نقلیه نیست ...  
با سختگویی موجود





# انفجار داده در عصر امروز

- شبکه های اجتماعی
- برنامه های تحت موبایل
- حسگرها / اینترنت اشیاء
- تصاویر دوربینهای نظارتی / تصاویر لاغ های سرورها
- خدمات آنلاین و وب سایتها
- کاربردهای خاص : نجوم، هواشناسی، پزشکی

# بررسی موردی - یک کلیک ساده

فروشگاه اینترنتی دیجی کالا > کالای دیجیتال > لپ تاپ > لپ تاپ و التراپوک

لپ تاپ و التراپوک (نمایش ۱ - ۴۰ محصول از ۹۸۹)

مرتب: سا

لپ تاپ 13 اینچی اپل مدل 2017 ...32

لپ تاپ 14 اینچی ایسوس مدل B - ...

از 16 رای 4.0 ★

از 28 رای 3.1 ★

The screenshot shows a search results page for 'Laptop and tablet' on an e-commerce platform. The results are filtered to show only laptops and tablets. Two products are displayed: an Apple MacBook 13-inch (2017 model) and an ASUS B14 laptop. Below each product image is its name and a rating section. An orange hand cursor icon is overlaid on the second product's image, pointing to the rating area.

# داده های نهفته در یک کلیک

- کاربر چه کالایی را انتخاب کرد ؟
  - ❖ به روزرسانی تعداد بازدید کالا، دسته، روزانه، ماهیانه و ...
  - ❖ شناخت علایق کاربر در نمایش تبلیغات و سایر کالاهای
- زمان و ساعت کلیک
- توالی کلیک
- کالای اصلی صفحه
- زمان ماندن در صفحه قبل از کلیک

# بررسی موردی - جستجو



- بررسی کاربر/ مهمان
- متون عامیانه و زبان گفتاری
- پیشنهاد کلمه همزمان با تایپ
- ذخیره و تحلیل جستجوهای ناموفق
- تحلیل رفتار کاربر بعد از نمایش نتایج

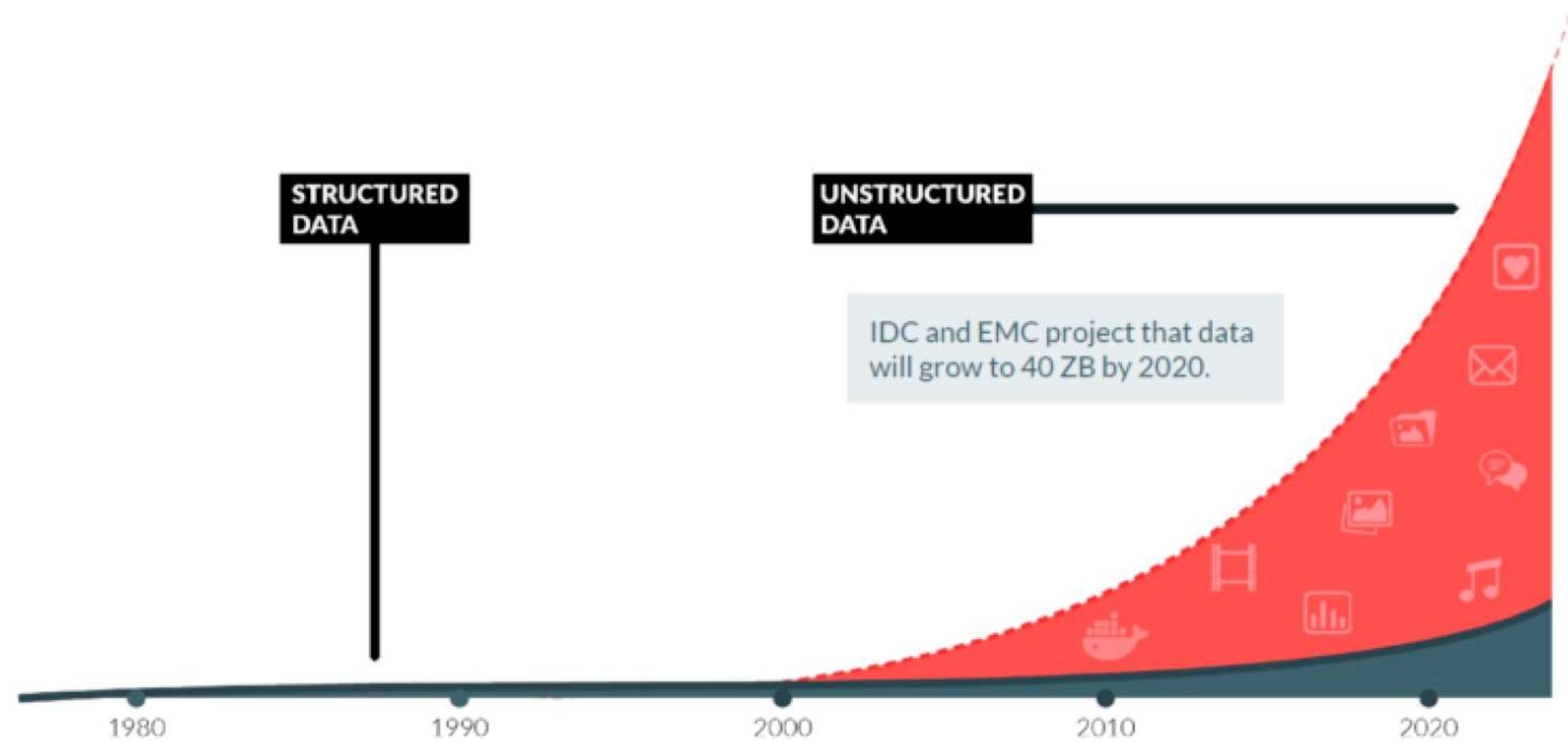
بسیار مهم

# یک دقیقه از یک سایت فروشگاهی

فجوم بی امان  
داده  
عدم پاسخگویی  
وشای سنتی

- انتخاب کالا
- گذاشتن نظر
- خرید کالا
- کنسل کردن سفارش
- جستجو
- امتیازدهی محصول
- .....

# رشد داده‌های غیرساختیافته



# ادیات جدید در حوزه اندازه داده

Welcome to the new vocabulary

## Geopbyte\*

This will be our digital universe tomorrow...

$10^{30}$

## Yottabyte

This is our digital universe today  
= 250 trillion of DVDs

$10^{27}$

## Brontobyte

A 1BB hard drive would cover the earth 23,000 times

$10^{24}$

## Exabyte

$10^{21}$

## Zettabyte

1.3 ZB of network traffic by 2016

$10^{18}$

## Petabyte

The CERN Large Hadron Collider generates 1PB per second

$10^{15}$

$10^{12}$   
Terabyte

500TB of new data per day are ingested in Facebook databases

$10^9$

Gigabyte

Megabyte

$10^6$

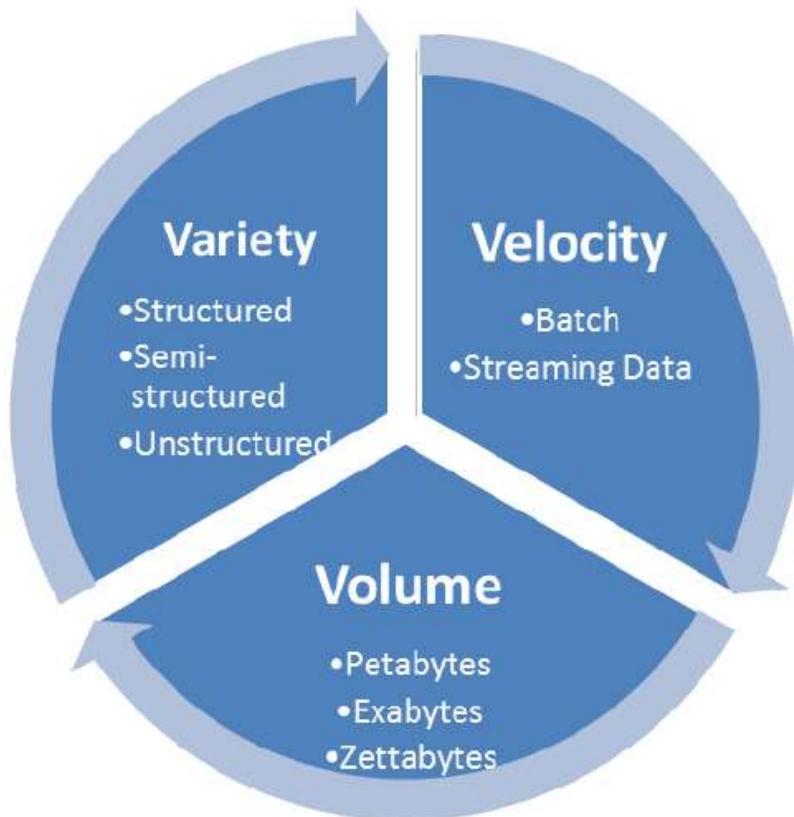
1 EB of data is created on the internet each day = 250 million DVDs worth of information.  
The proposed Square Kilometer Array telescope will generate an EB of data per day



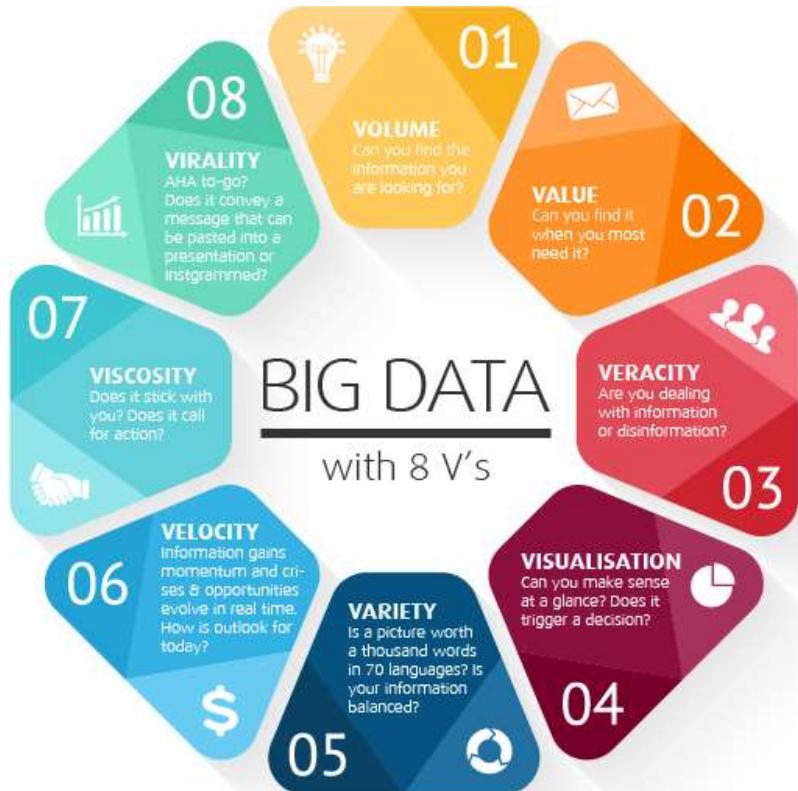
# تعریف کلان داده

Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

# تعریف کلانداده - 3V



# تعزیف کلان داده – ۵V , ۸V



# تعریف کلان داده



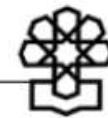
<https://bigdatawg.nist.gov/>

## 3.1 BIG DATA DEFINITIONS

Big Data refers to the need to parallelize the data handling in data-intensive applications. The characteristics of Big Data that force new architectures are as follows:

- *Volume* (i.e., the size of the dataset);
- *Velocity* (i.e., rate of flow);
- *Variety* (i.e., data from multiple repositories, domains, or types); and
- *Variability* (i.e., the change in velocity or structure).

# کلان داده در ایران



فناوری داده‌های عظیم و الزامات قانونی آن

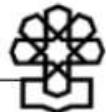
## مرکز پژوهش‌های مجلس - خرداد ۹۴

چکیده

در سال‌های اخیر صنعت فناوری اطلاعات، با ظهر حجم گسترده‌ای از داده‌ها و اطلاعات مواجه بوده که این داده‌ها عمدتاً از منابع متفاوتی از جمله ابزارهای علمی، ماهواره‌ها، رسانه‌های دیجیتالی،

دانشگاه‌ها، اینترنت، موسسه‌های علمی و تحقیقاتی، شرکت‌های فناوری، و اینترنت اشیاء

## کلان داده در ایران



### شکل ۳. چارچوب حقوقی برای داده‌های عظیم

- سطح اول: زیرساخت‌های فنی

سطح دوم: معماری اطلاعات

سطح سوم: حقوق مالکیت فکری داده‌ها

سطح چهارم: الزامات قراردادی داده‌ها

سطح پنجم: حفاظت از داده‌ها و تنظیم

سطح ششم: مدیریت امنیت اطلاعات

## فناوری داده‌های عظیم و الزامات قانونی آن

۹۴ - خرداد مجلس پژوهش‌های مرکز

چکیدہ

در سال‌های اخیر صنعت فناوری اطلاعات، با ظهر حجم گستردگی از داده‌ها و اطلاعات که این داده‌ها عمدتاً از منابع متفاوتی از جمله ابزارهای علمی، ماهواره‌ها، رسانه‌های

# کلان داده در ایران



Open Community of Cloud Computing  
جامعه آزاد رایانش ابری ایران

کارگروه تاکسونومی و استانداردسازی

OCCC.IR

نقشه راه توسعه کلان داده در کشور



نسخه ۴ - دیماه ۹۵

وضعیت کلان داده در دنیا و ایران

چالش های کلان داده در ایران

برناهه ریزی در حوزه کلان داده

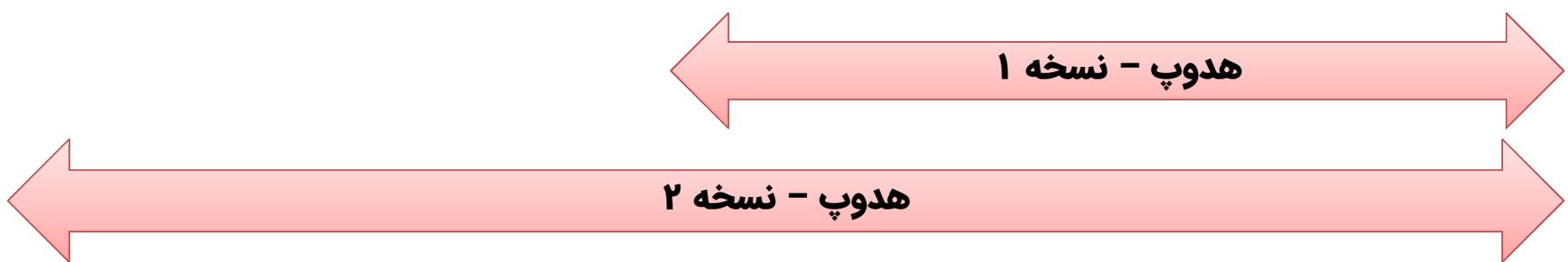
نقشه راه

# سه چالش اصلی در کار با کلان داده

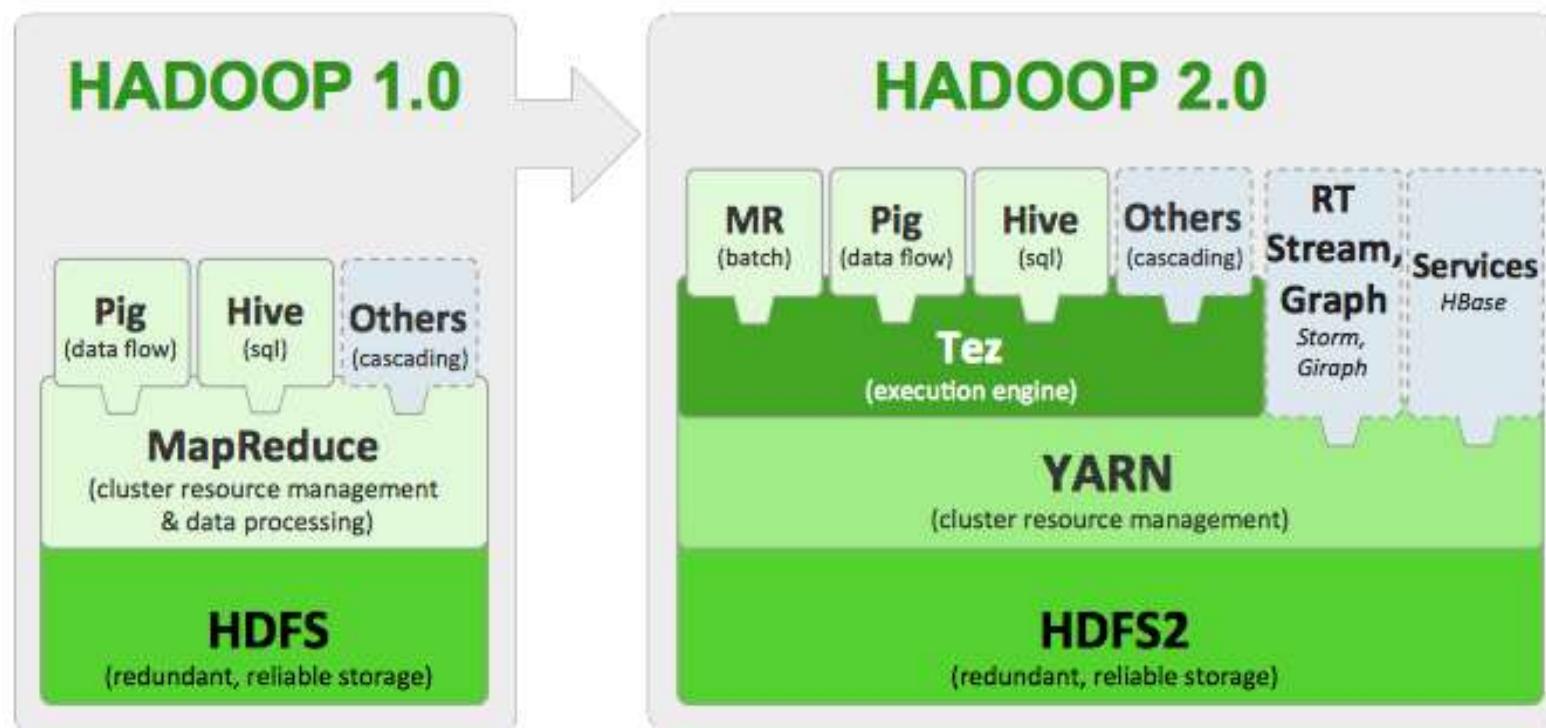
- مدیریت منابع**
- اختصاص منابع
  - مدیریت کلاستر
  - تعادل بار

- پردازش**
- توزیع شده
  - سخت افزار ارزان
  - عدم نیاز به بکاپ

- ذخیره**
- توزیع شده
  - تحمل خطا
  - مقیاس پذیر
  - تحمل خطا
  - مقیاس پذیر



# هدوپ : چارچوب پردازشی کلاسیک کلان داده



# تعریف رسمی هدوب

**The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.**

- تعریف رسمی بنیاد آپاچی در صفحه رسمی این پروژه -

## Modules

The project includes these modules:

- **Hadoop Common:** The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™):** A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN:** A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce:** A YARN-based system for parallel processing of large data sets.
- **Hadoop Ozone:** An object store for Hadoop.

# سه چالش اصلی / سه مولفه اصلی

مدیریت منابع

پردازش

ذخیره



**YARN**

**Map/Reduce**

**HDFS**



# تاریخچه هدوب

- Nutch Web Crawler - to index 1 billion web page

2002

- The Google File System - Google Article

2003

- Nutch DFS : NameNode and DataNode
- MapReduce - Google Article

2004

- Nutch Joind Apache Incubator
- MapReduce in Nutch

2005

# تاریخچه هadoop

- Hadoop Subproject Created from Nutch : NDFS+MR
- Hadoop 0.1.0 Released
- Doug Cutting Joined Yahoo as Hadoop team Leader
- Hadoop sorts 1.8 TB on 188 Nodes in 48 Hours
- **Yahoo** deploys 300 Machine Hadoop cluster - 600 by end of the year

2006

- First release of Hadoop including Hbase
- Only 3 Companies listed in “Powered By Hadoop” section

2007

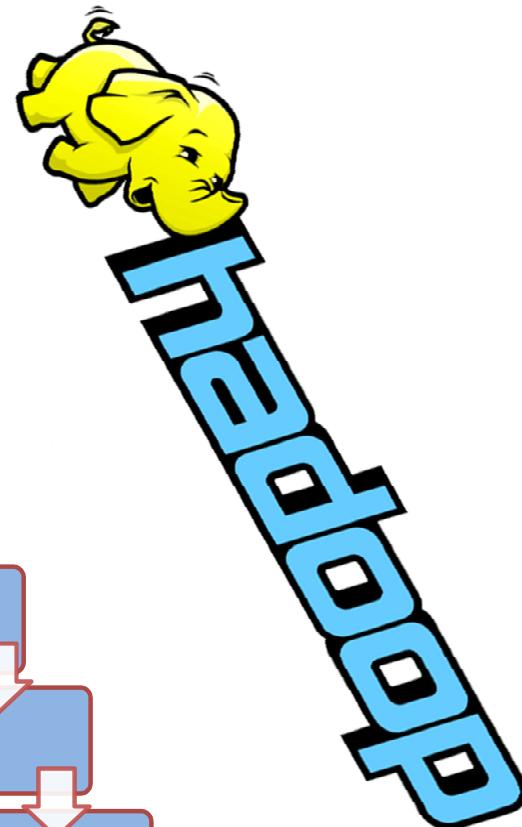
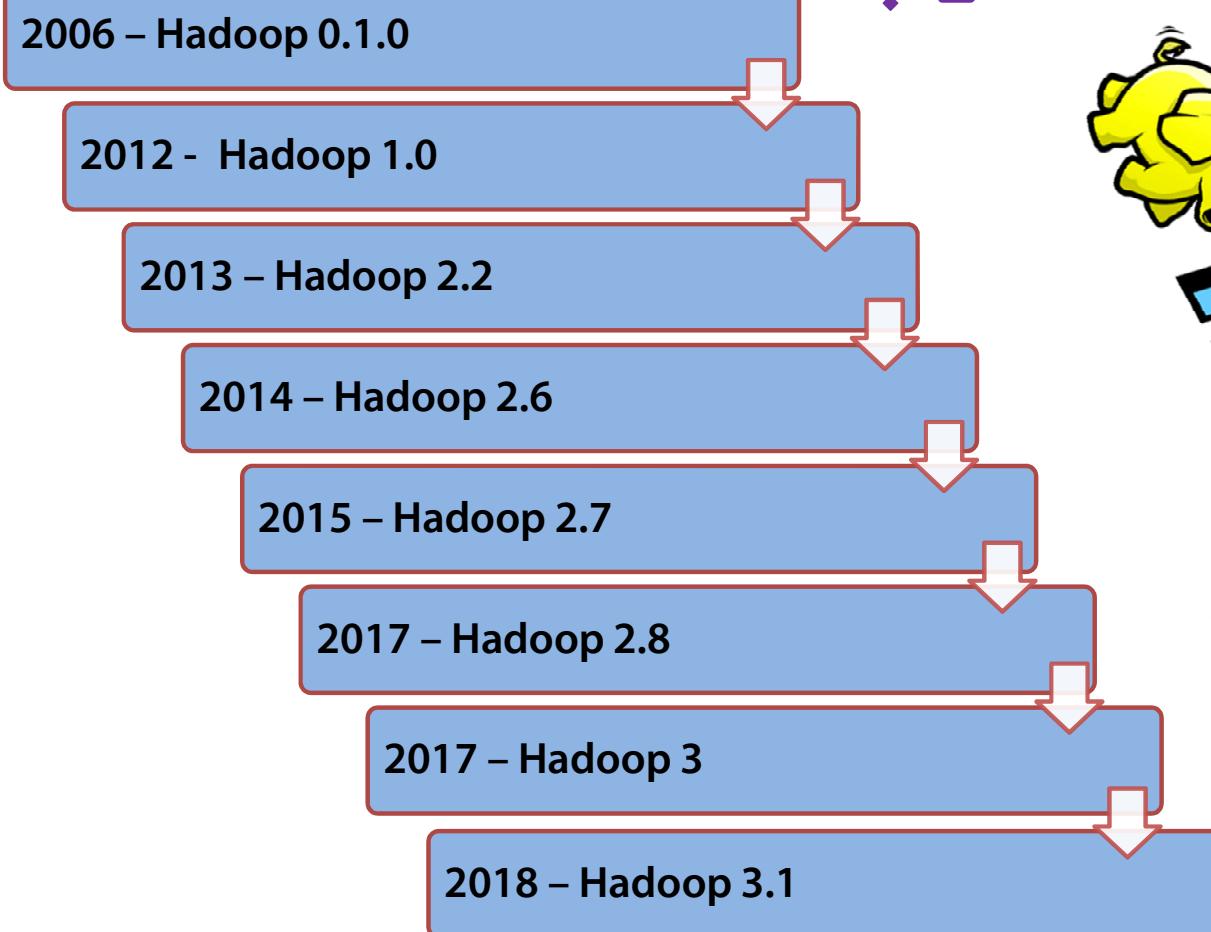
- Hadoop : Apache Top-Level Project
- Hadoop wins TeraByte Sort Benchmark - 1TB / 910 node / 209 Sec

2008

- Hadoop Core renamed to Hadoop Common
- MapReduce / HDFS are Separate subproject
- Google sorted 1 TB in 68Sec / Yahoo sorted in 62Sec

2009

# نسخه های مختلف هدوپ



# نام هدوپ و نماد فیل : عروسک فرزند Doug Cutting



# چرا هدوپ؟ چرا توزیع فایل‌ها و پردازش؟

<http://www.fabak.ir>

**هدف: پردازش ۲۰۰ گیگابایت**

اگر در هر ثانیه بتوانیم ۵۰ مگابایت را بخوانیم

$$\begin{array}{ccc} 200 & = & 200000 \\ \text{Gigabyte} & \downarrow & \text{Megabyte} \\ & & \end{array}$$

$$200,000 \text{ ( MB) } / 50 \text{ (MB) } = 4,000 \text{ ( Second)}$$

۱ سکره پردازش  $4,000 \text{ ( Second) } / 1 \text{ (node) } \sim 1 \text{ Hour.}$

اگر لازم باشد این حجم داده در زمانی کمتر از ۵ دقیقه خوانده شود چه کار باید گردید؟

۱۰۰ سکره پردازش  $4,000 \text{ ( Second) } / 100 \text{ ( Node) } = 40 \text{ Second } < 1 \text{ Min.}$

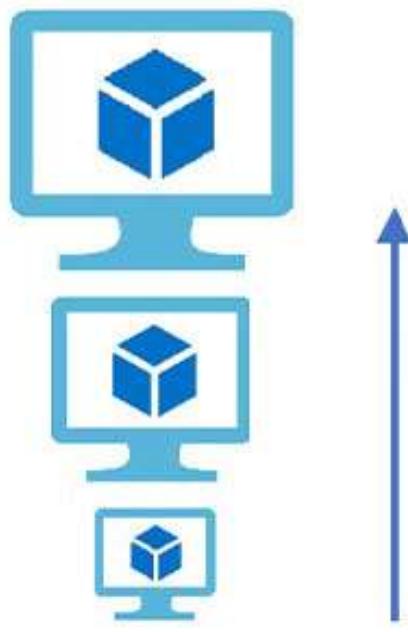
## راز محبوبیت هدوپ

- تمرکز بر پردازش داده و نه نحوه ذخیره و اجرای آن
- وجود خطا در یک گره امری ممکن و همیشگی است
- سخت افزارهای معمولی هم می توانند بخشی از پردازش را انجام دهند.
- وجود مجموعه داده های بزرگ در دنیای واقعی

# مقیاس‌پذیری افقی با هدوب

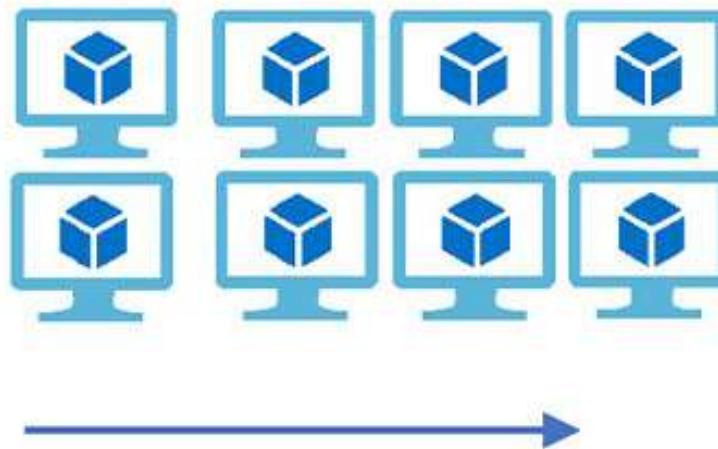
## Vertical Scaling

( Increase size of instance (RAM ,  
CPU etc.) )



## Horizontal Scaling

( Add more instances )



# کارگاه عملی

- دانلود مخزن کد درس

[https://gitlab.com/nikamooz\\_bigdata/de](https://gitlab.com/nikamooz_bigdata/de)

- دستورات این بخش در پوشه Section3 قرار گرفته‌اند
- نرم افزار Typora را حتماً نصب کنید
- در این کارگاه نیاز به اینترنت و نصب داکر بر روی سیستم خود دارید.

# منابع و مطالعات بیشتر

