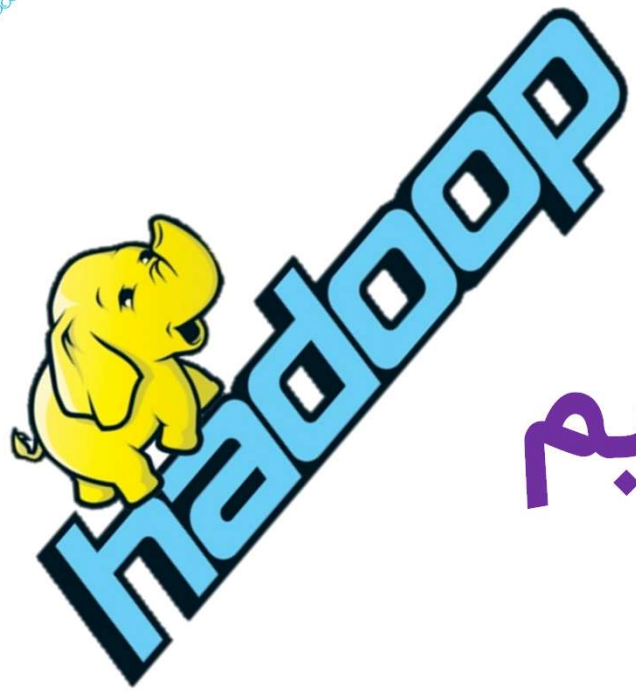


جلسه اول

آشنایی با مفاهیم کلان‌داده

کارعملی با هدوپ

کلان داده

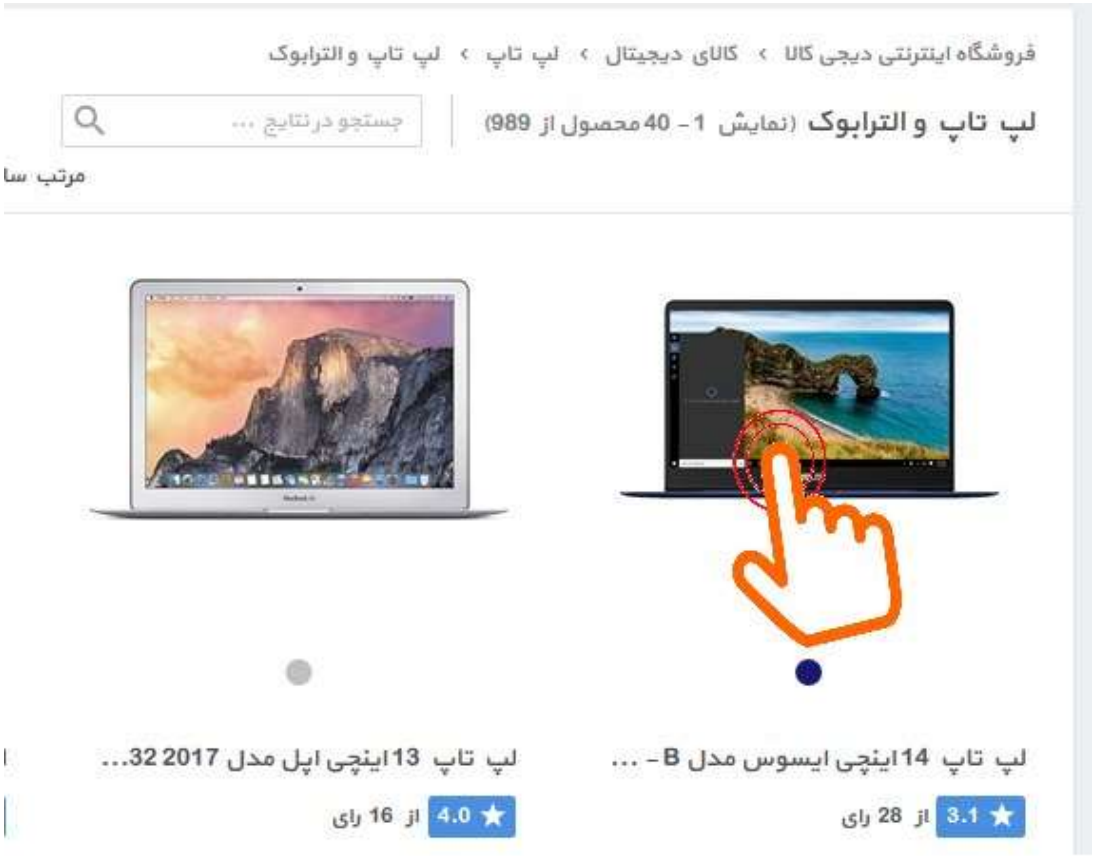
دلیل اصلی ترافیک کلان‌شهرها

منابع شهری موجود
پاسخگوی حجم عظیم
وسایل نقلیه نیست ...

انفجار داده در عصر امروز

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

بررسی موردی – یک کلیک ساده



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 کالاها
- 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

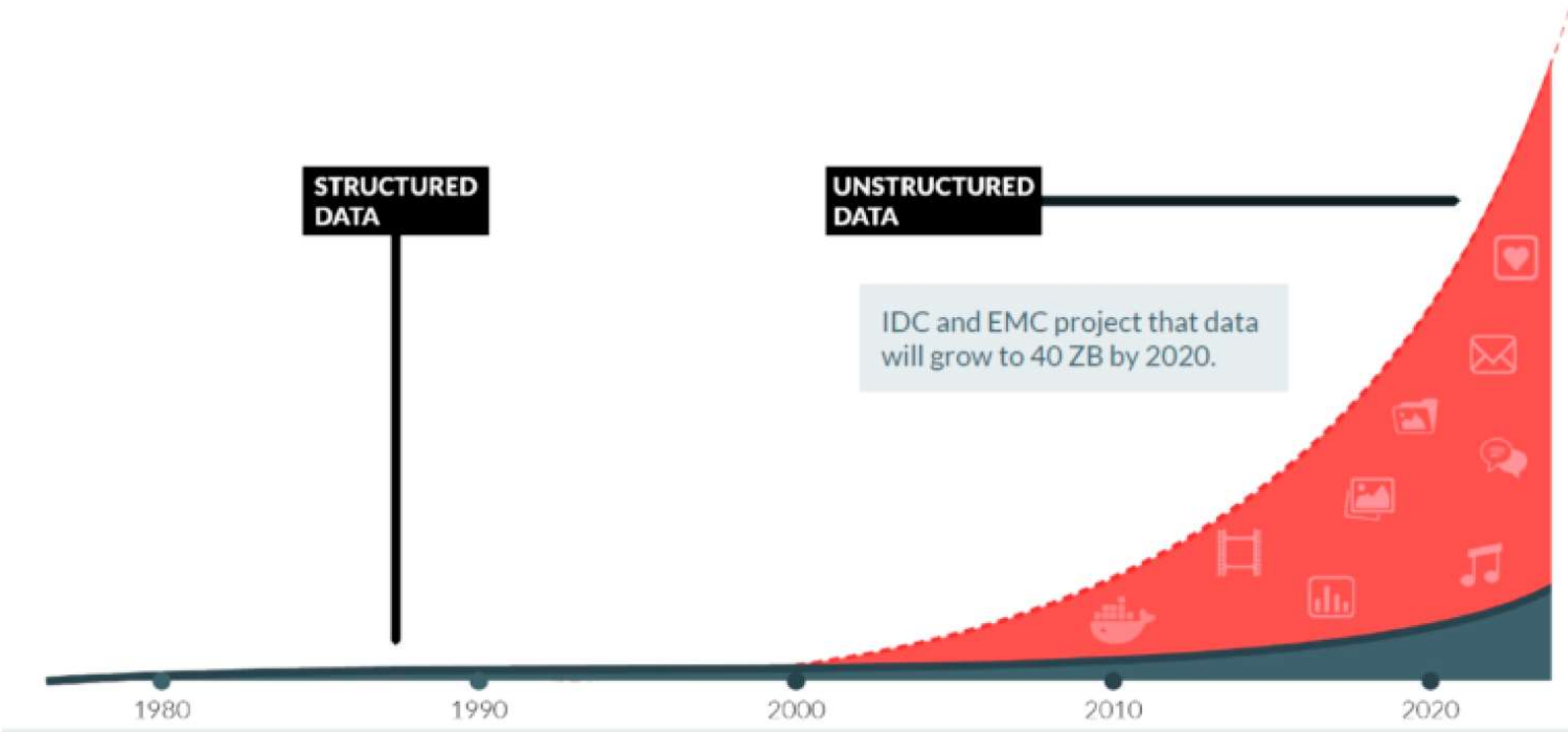
یک دقیقه از یک سایت فروشگاهی

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

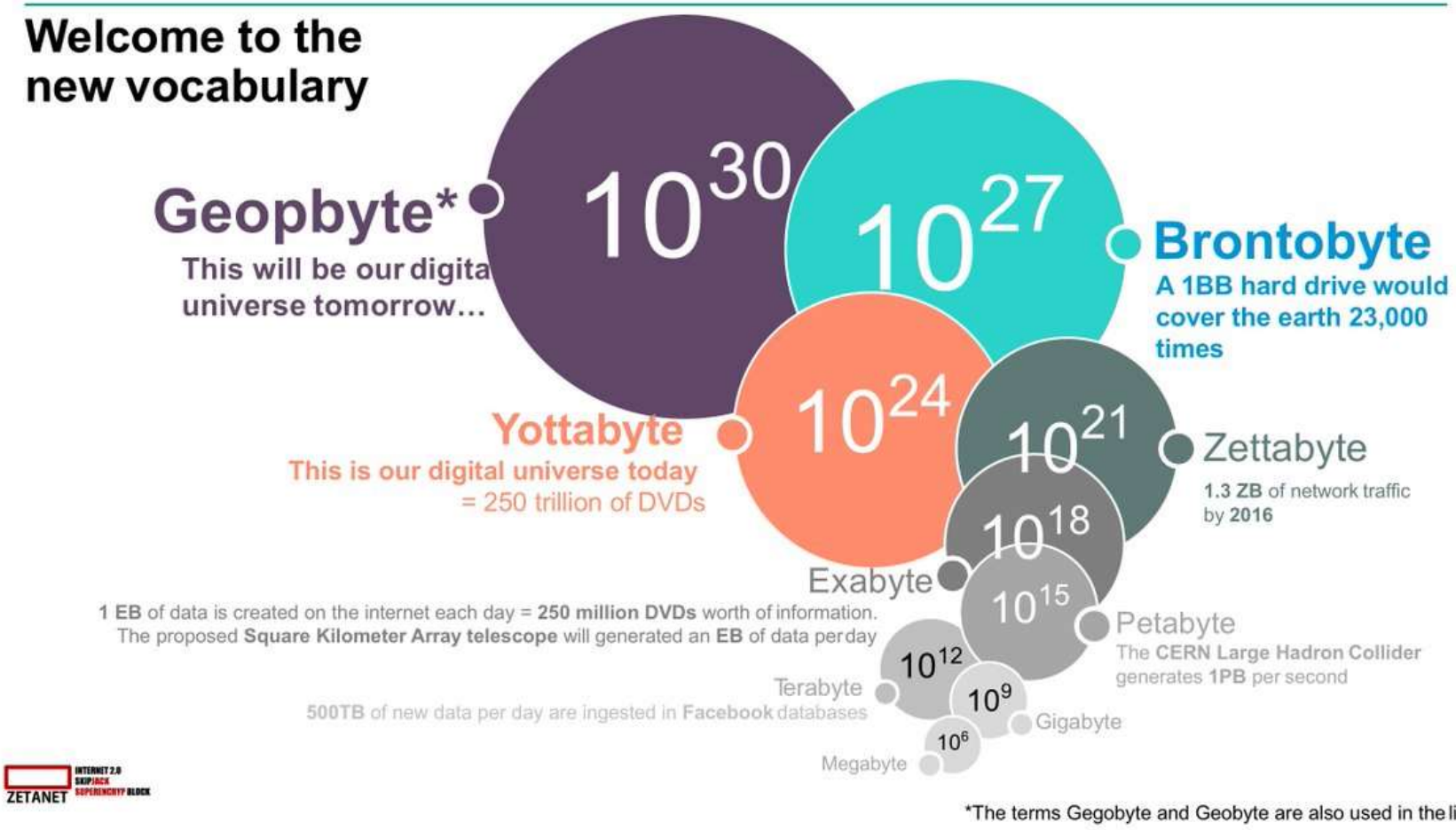
هجوم بی امان داده

و عدم پاسخگویی روشهای سنتی

رشد داده‌های غیرساختیافته



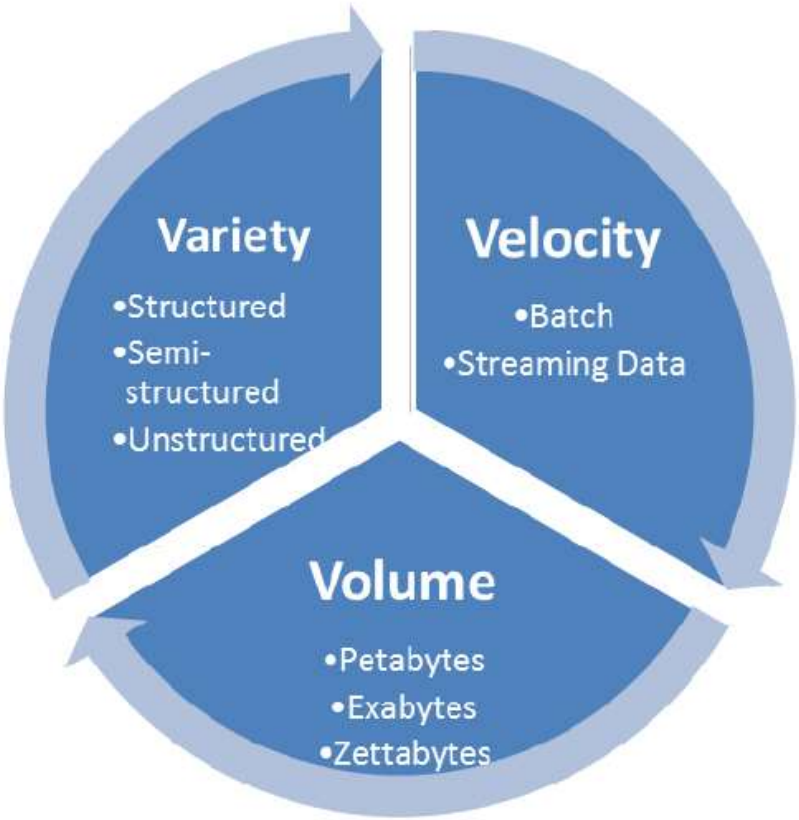
ادبیات جدید در حوزه اندازه داده



تعریف کلان‌داده

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



تعریف کلان‌داده – 5V , 8V

تعریف کلان‌داده

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

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 ظهور حجم گسترده‌ای از داده‌ها و اطلاعات مواجه بوده که این داده‌ها عمدتاً از منابع متفاوتی ازجمله ابزارهای علمی، ماهواره‌ها، رسانه‌های دیجیتالی،

کلان داده در ایران

شکل ۳. چارچوب حقوقی برای داده‌های عظیم

فناوری داده‌های عظیم و الزامات قانونی آن

مرکز پژوهش‌های مجلس - خرداد ۹۴

چکیده

در سال‌های اخیر صنعت فناوری اطلاعات، با ظهور حجم گسترده‌ای از داده‌ها و اطلاعات که این داده‌ها عمدتاً از منابع متفاوتی ازجمله ابزارهای علمی، ماهواره‌ها، رسانه‌های

کلان داده در ایران

نقشه راه توسعه کلان داده در کشور

نسخه ۴ – دیماه ۹۵

Open Community of Cloud Computing
جامعه آزاد رایانش ابری ایران
کارگروه تاکسونومی و استانداردسازی

OCCC.IR

وضعیت کلان داده در دنیا و ایران

چالش های کلان داده در ایران

برنامه ریزی در حوزه کلان داده

نقشه راه

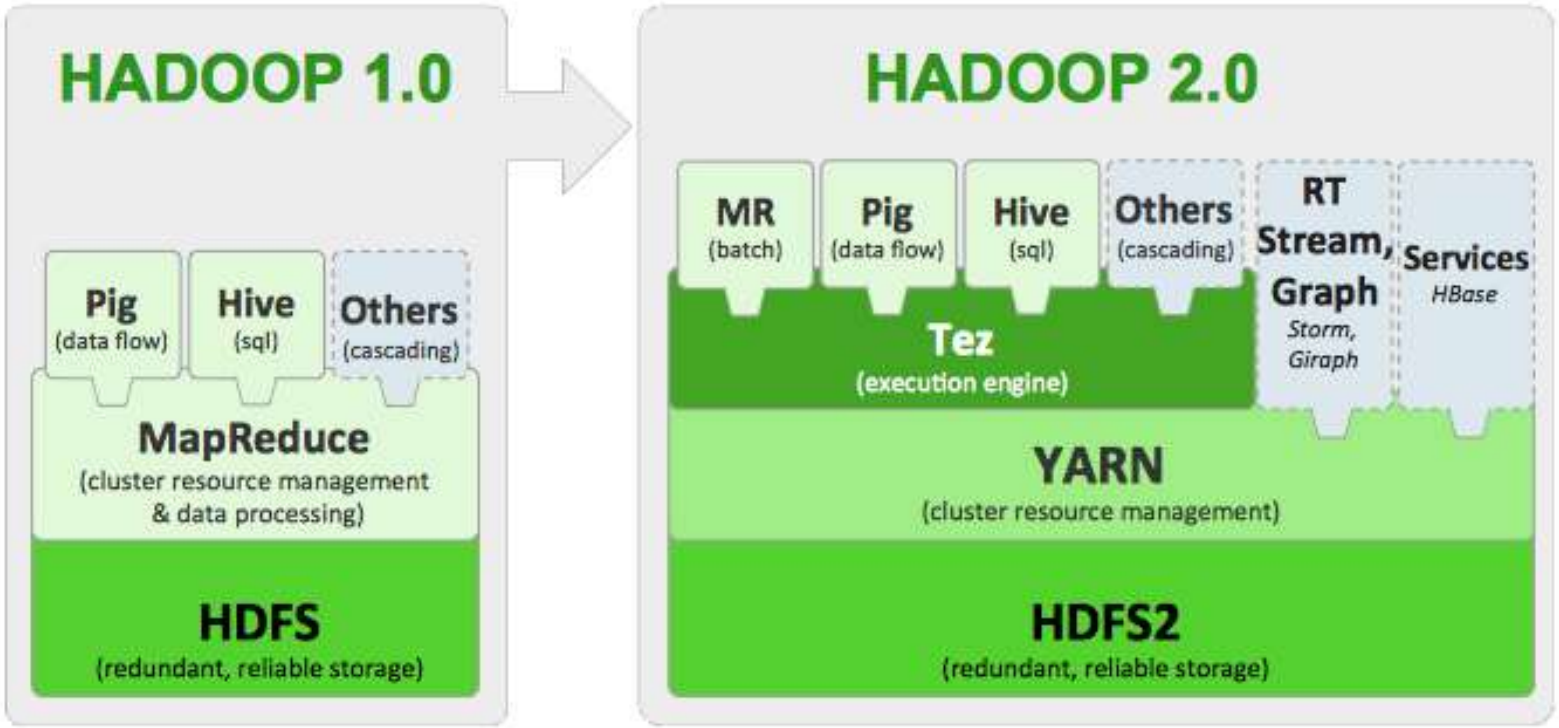
سه چالش اصلی در کار با کلان داده

ذخیره
- توزیع شده
- سخت افزار ارزان
- عدم نیاز به بکاپ
- تحمل خطا
- مقیاس پذیر

پردازش
- توزیع شده
- تحمل خطا
- مقیاس پذیر

مدیریت منابع
- اختصاص منابع
- مدیریت کلاستر
- تعادل بار

هدوپ : چارچوب پردازشی کلاسیک کلان داده



تعریف رسمی هدوپ

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.

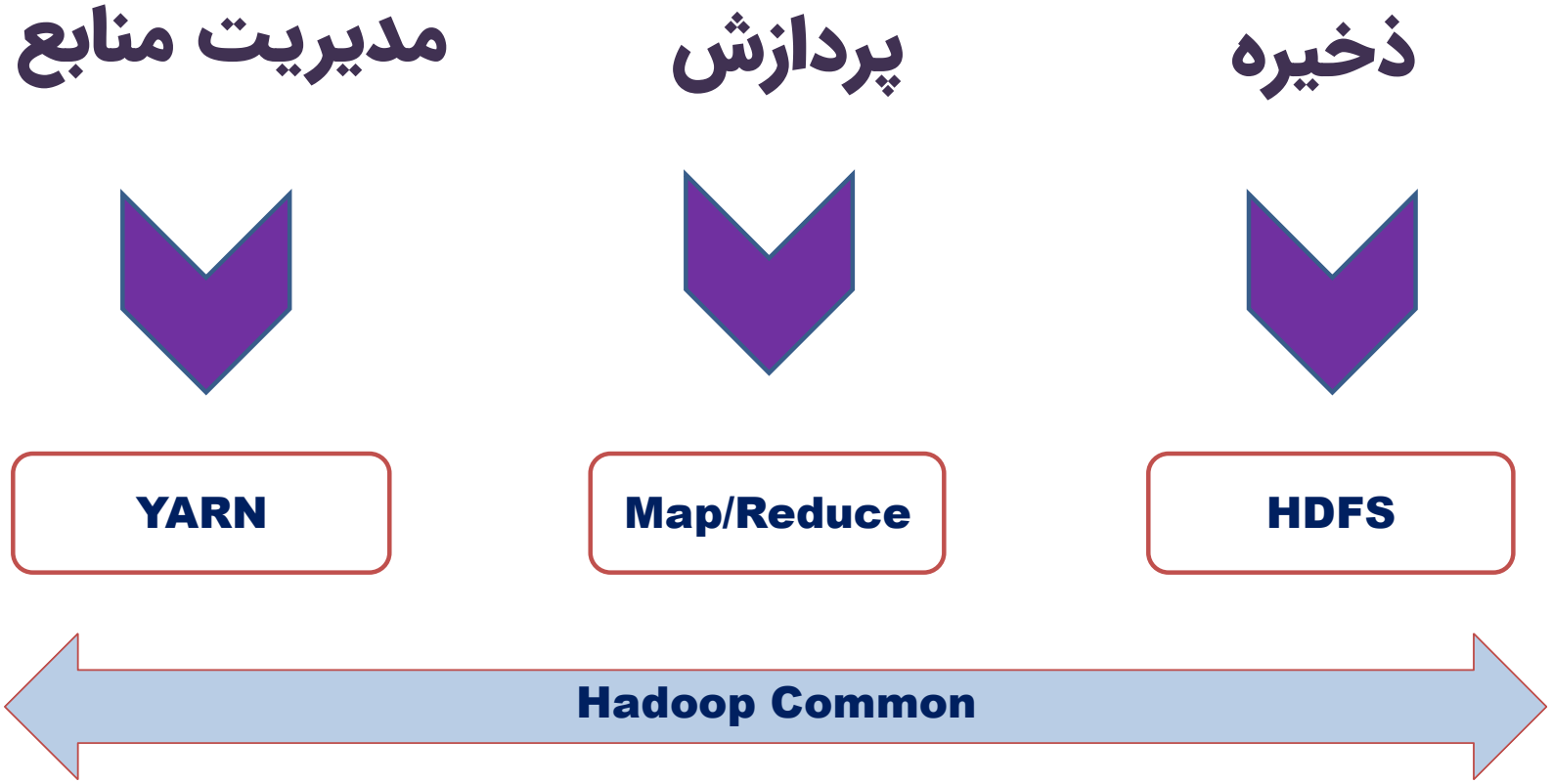
- تعریف رسمی بنیاد آپاچی در صفحه رسمی این پروژه

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



تاریخچه هدوپ

- 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 هدوپ

- 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

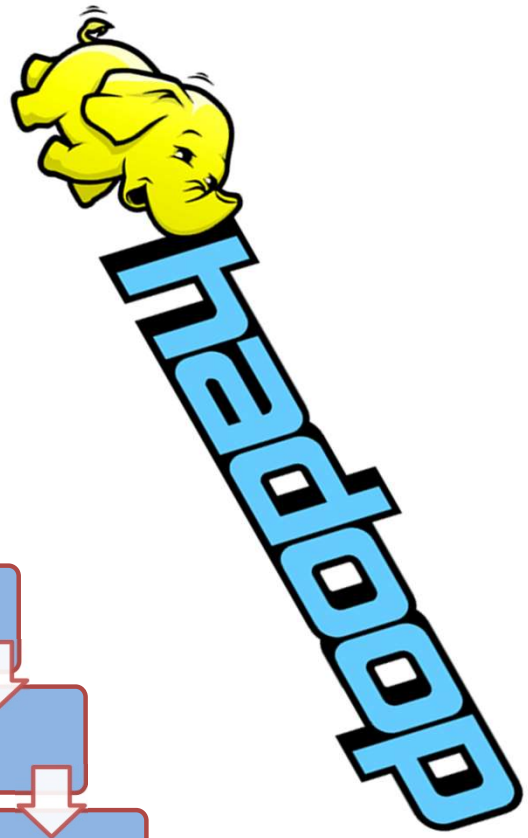
2008

- **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

- 2006 – Hadoop 0.1.0
- 2012 - Hadoop 1.0
- 2013 – Hadoop 2.2
- 2014 – Hadoop 2.6
- 2015 – Hadoop 2.7
- 2017 – Hadoop 2.8
- 2017 – Hadoop 3
- 2018 – Hadoop 3.1



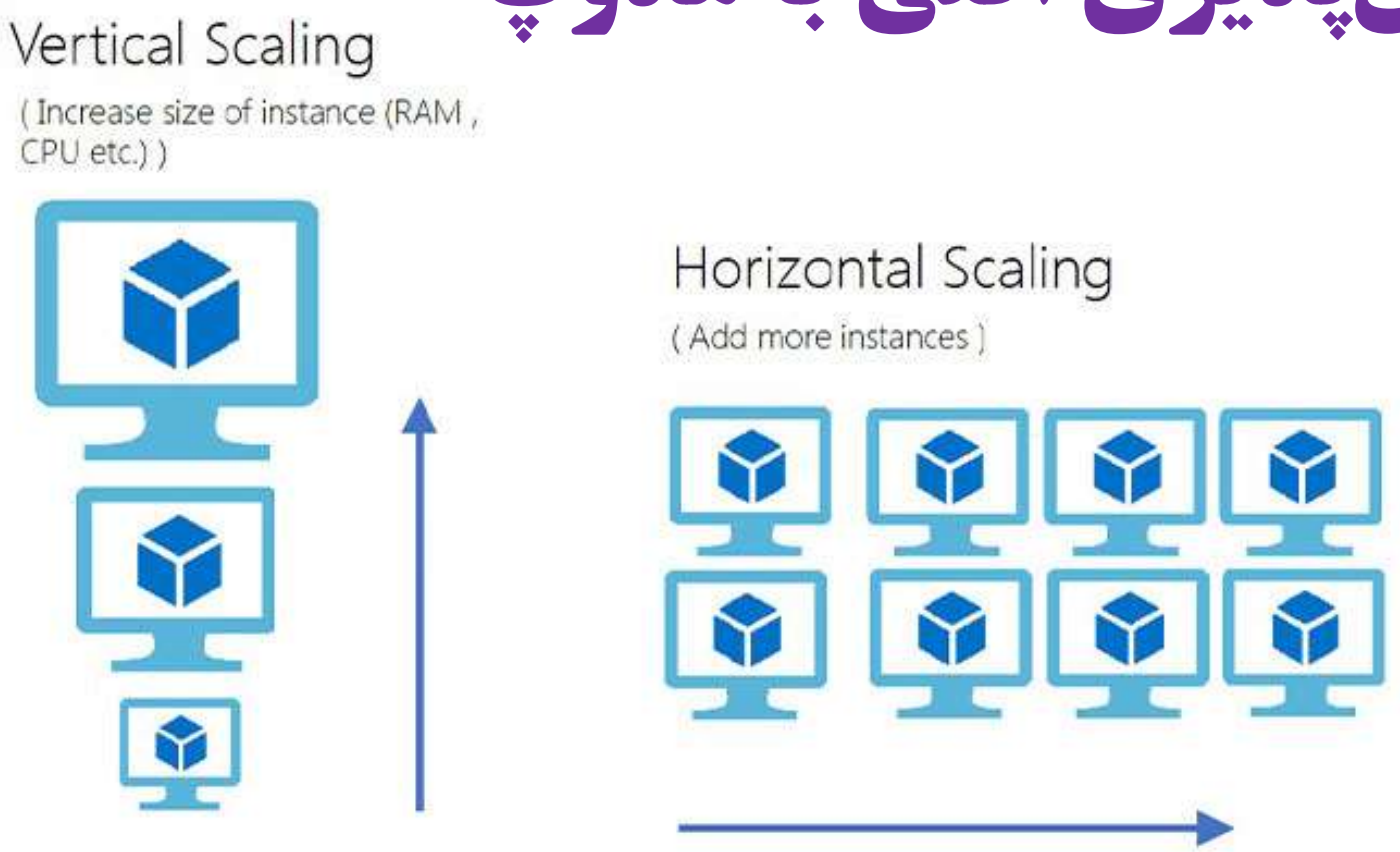
نام هدوپ و نماد فیل : عروسک فرزند Doug Cutting

چرا هدوپ ؟ چرا توزیع فایل ها و پردازش ؟

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 هدوپ



منابع و مطالعات بیشتر