

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

آشنایی با

Big Data

سید محمدامین ساجدی



بهار ۱۳۹۵



حقایق یا قسمتی از اطلاعات

Foreign Keys

students:

id	name
1	Anna Malli
2	Anders Andersen
3	Pierre Untel
4	Erika Mustermann
5	Juan Pérez
6	Fulano de Tal
⋮	⋮

grades:

student	course	grade
4	MATH201	A-
1	CS413	A
3	CS100	B+
6	B10301	B
1	PHY222	A
2	ARTH213	B
⋮	⋮	⋮

Courses:

id	name
CS100	Intro Comp Sci
MATH201	Calculus
ARTH213	Surrealism
CS413	Purely Functional..
B10301	Anatomy
PHY222	Electromagnetism
⋮	⋮

ظرفیت...





تنوع....



ما همه هوشمندیم....





۹۰ درصد داده های دنیا در ۲ سال اخیر تولید شده اند.

انواع داده

- Feed
- شبکه های اجتماعی
- تصاویر
- ویدئوها
- فایل های متنی
- اسناد



START

فقط حجم؟



VOLUME

VOLUME

- Terabytes
- Records
- Transactions
- Tables, files

3 Vs of Big Data

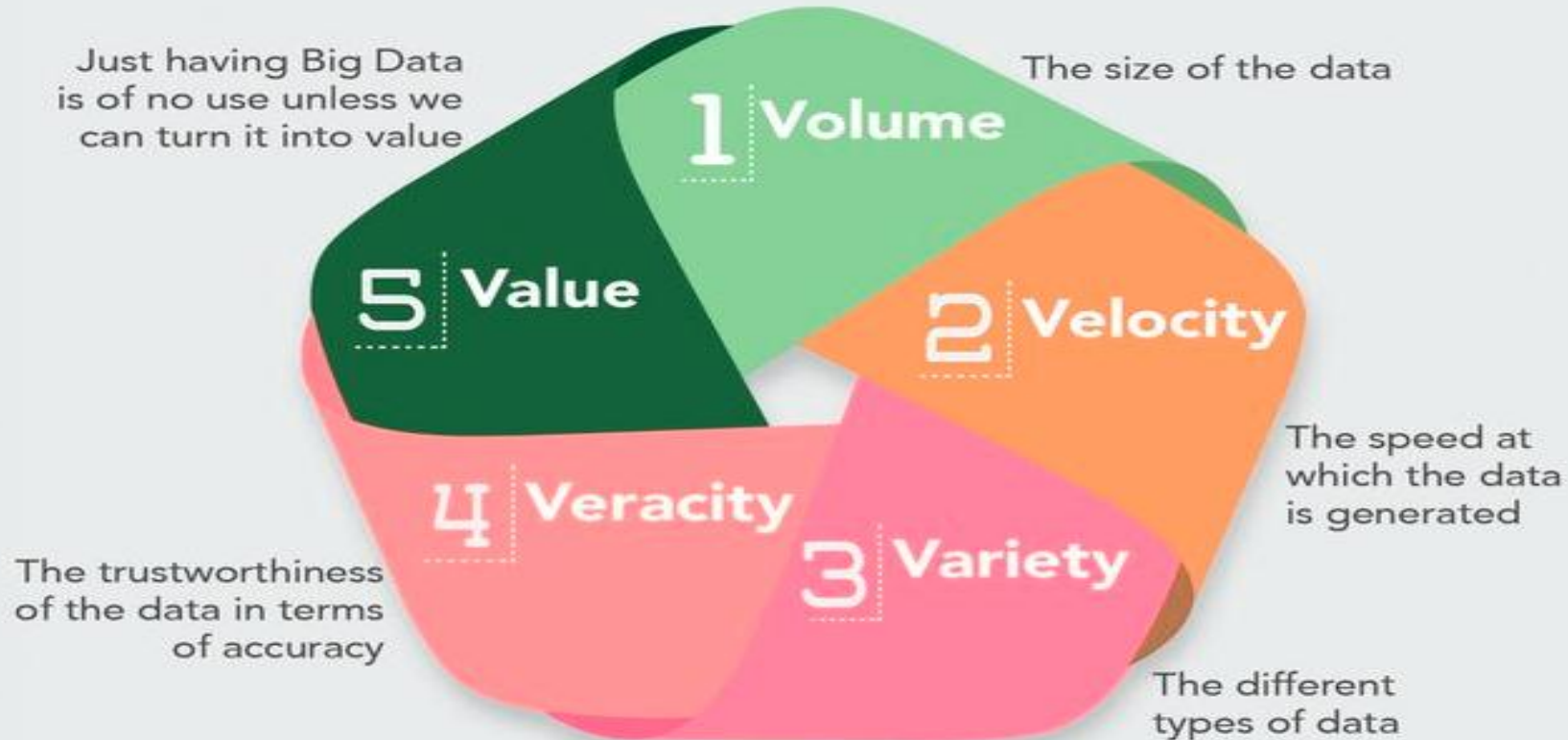
- Batch
- Near time
- Real time
- Streams

- Structured
- Unstructured
- Semistructured
- All the above

VELOCITY

VARIETY

THE 5 Vs OF BIG DATA



چه حجمی؟

داده های با حجم بیش از ۳۰ ترابایت

هوش تجاری / سازمانی

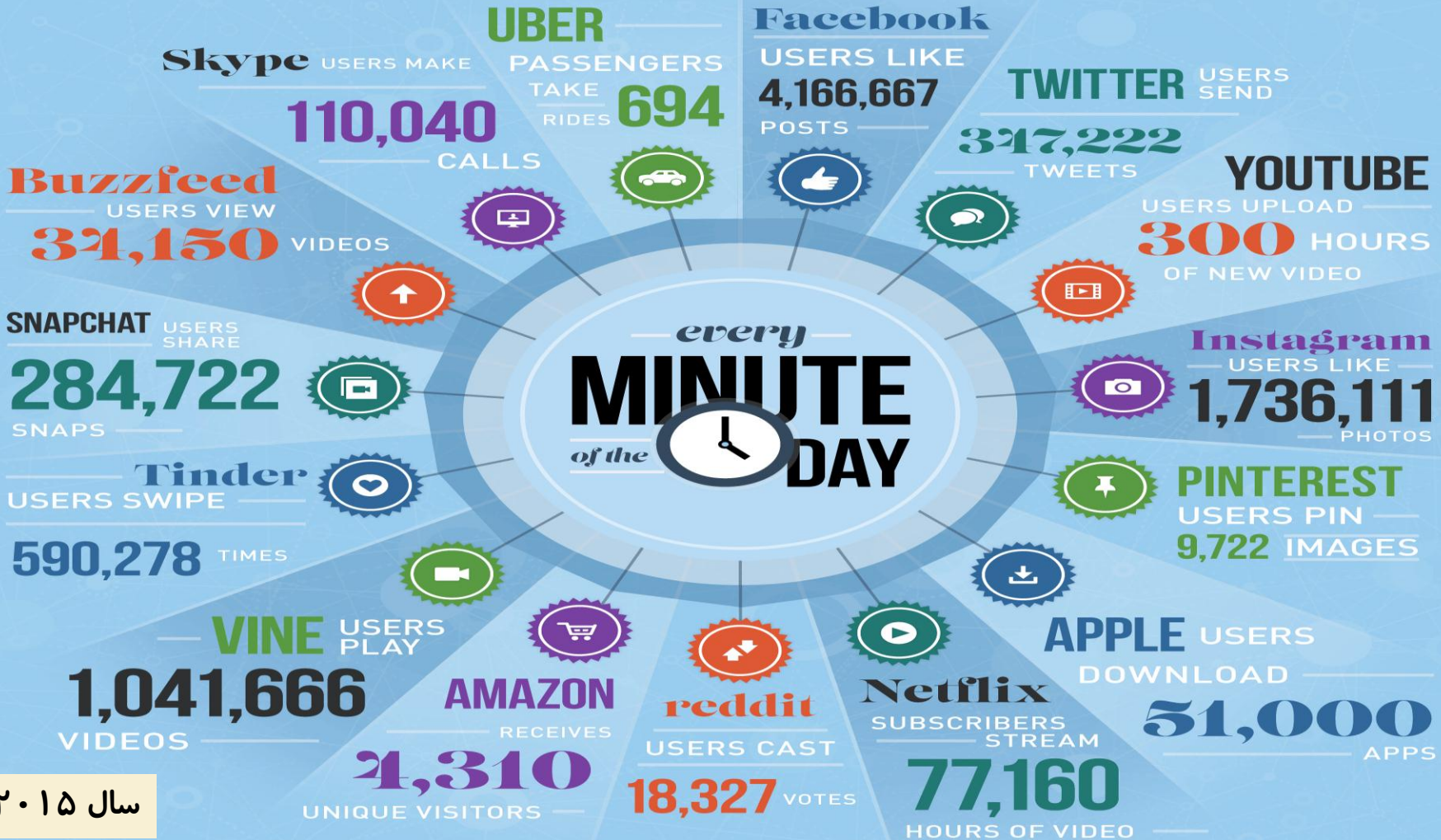
ویژگی های هوش سازمانی / تجاری

- تمام داده ها در یک سرور
- تحلیل آفلاین
- معمولا پایگاه داده های رابطه ای
- در انبار داده انجام می شود

ویژگی های بزرگ داده

- در سرورهای متفاوت توزیع شده اند
- داده ها با ساختار متفاوت
- تحلیل های آنلاین و بلادرنگ
- پردازش بصورت موازی
- توابع پردازشی به سمت داده ها می روند نه برعکس

— every —
MINUTE
of the
DAY



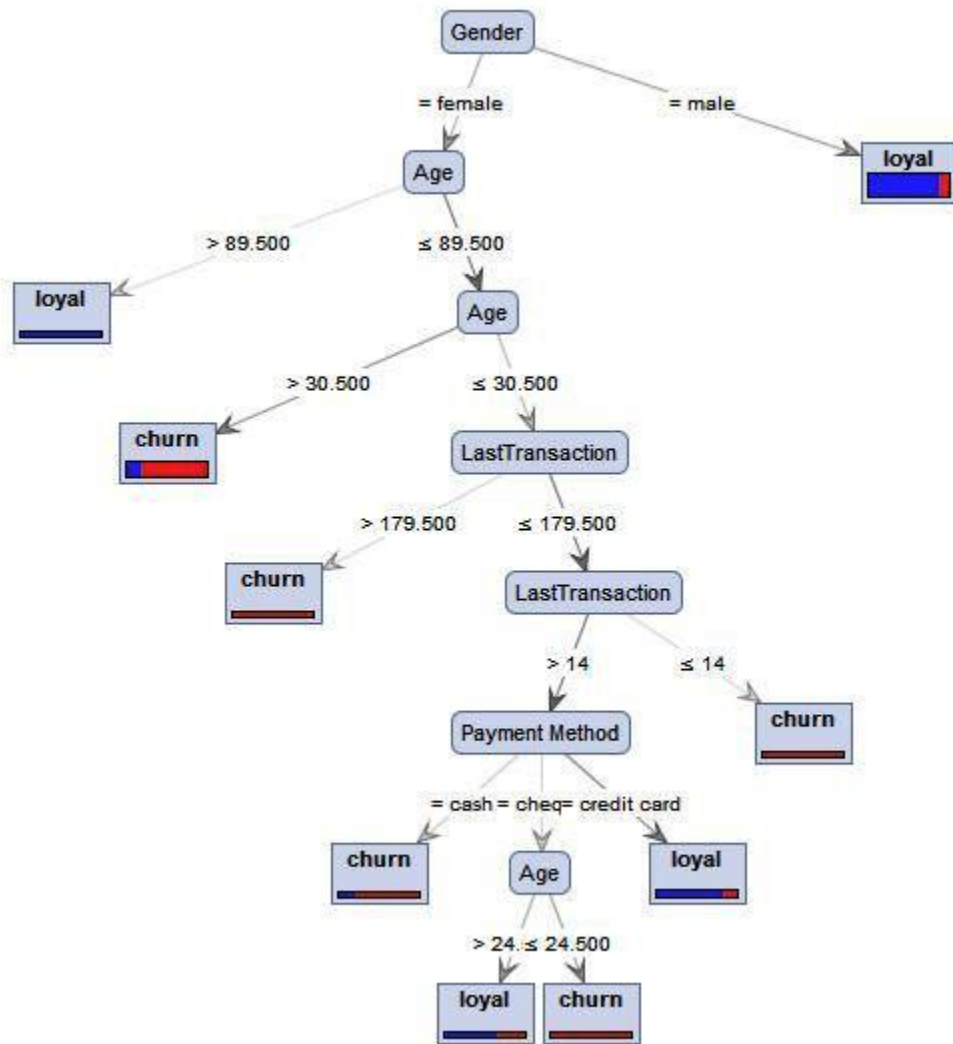
خب با این داده ها چکار میشه کرد؟

استخراج بینش

- رضایت مشتریان
- کسب سود بیشتر
- افزایش وفاداری مشتریان

insight

Row No.	Churn	Gender	Age	Payment Me...	LastTransa...
1	loyal	male	64	credit card	98
2	churn	male	35	cheque	118
3	loyal	female	25	credit card	107
4	?	female	39	credit card	177
5	loyal	male	39	credit card	90
6	churn	female	28	cheque	189
7	loyal	female	21	credit card	102
8	loyal	male	48	credit card	141
9	churn	female	70	credit card	153
10	loyal	male	36	credit card	46



ب عنوان مثال

- شرکت های تامین انرژی
- نمایش تبلیغات تلویزیونی و اینترنتی
- سیستم های تشخیص جرم و خطا
- بهبود سیستم ها
- تحلیل محتوای شبکه های اجتماعی

منابع جدید داده ای

- سنسورها
- لاگ ها
- فعالیت های کلیک کاربران
- توییت ها
- چت ها
- پیامک های موبایلی
- ایمیل ها

تحلیل در داده های بدون ساختار

- شناسایی الگو در متن
- شناسایی الگو در فیلم
- شناسایی الگو در تصویر
- الگوهای تکرار شده در اسناد، ایمیل ها، گفتگوها

چالش های پیش رو

- فهمیدن داده ها و اولویت بندی در ورود به سیستم
- از بین بردن اختلال موجود درون داده ها
- محل ذخیره سازی و نگهداری داده
- مدت زمان نگهداری داده ها
- مهارت کار با ابزارهای بزرگ داده ای
- مدیریت سیستم های بزرگ داده ای

Cloud Computing

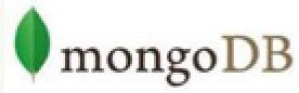
پردازش ابری



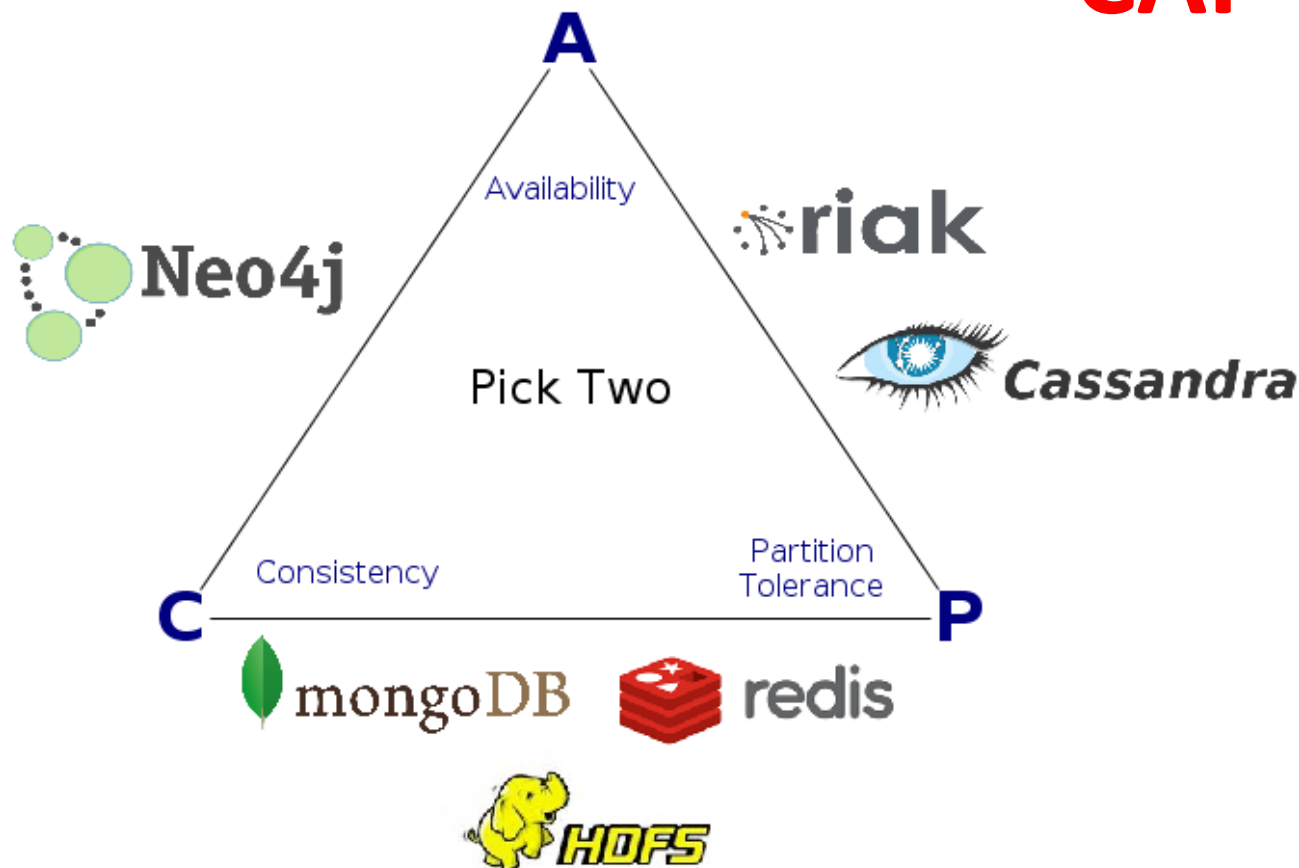
لایه ذخیره سازی

- توزیع شده است.
- به کاربران قول تحمل خطا می دهد.
- با موازی سازی امکان اجرای الگوریتم های پردازشی توزیع شده را می دهد.

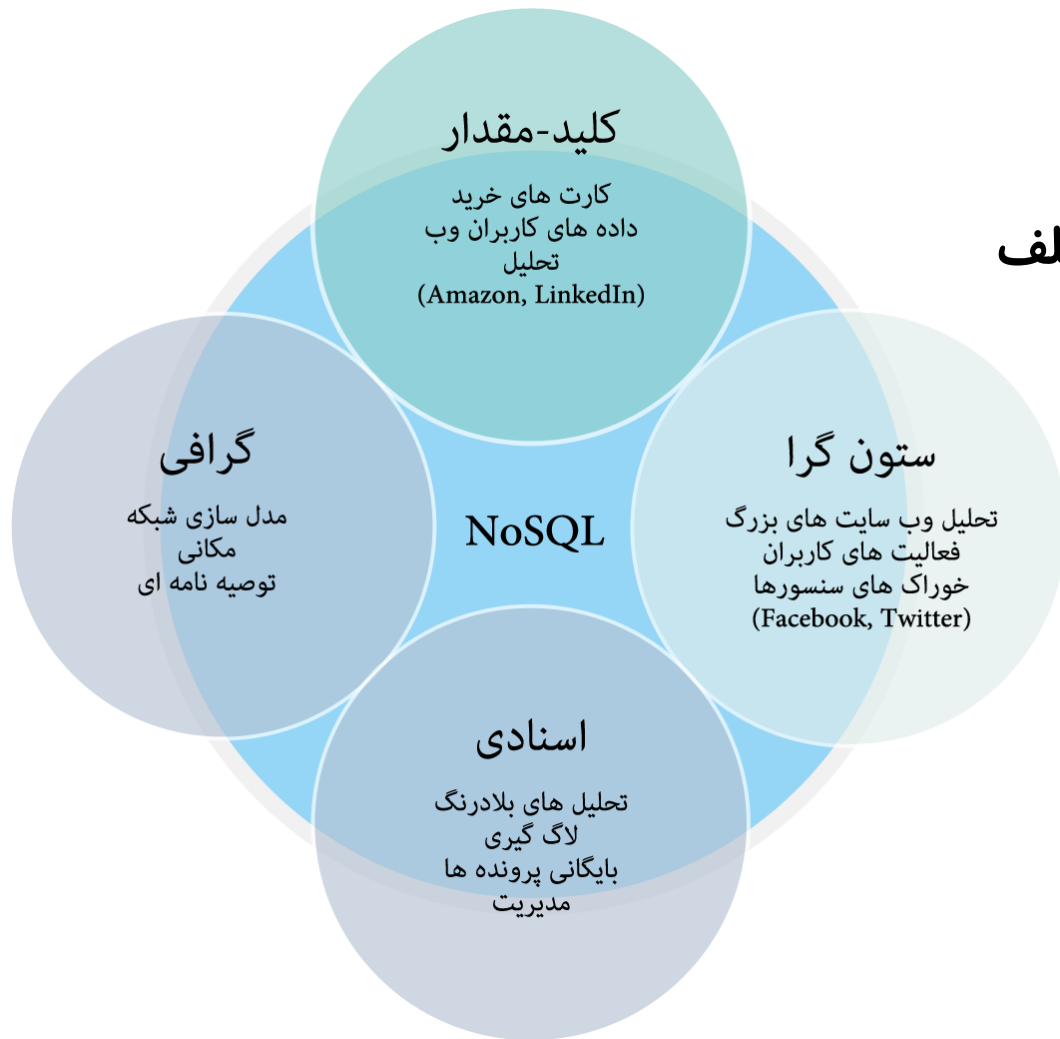
پایگاه داده های NOSQL

پایگاه داده های کلید-مقدار	پایگاه داده های ستون گرا	پایگاه داده های اسنادی	پایگاه داده های گرافی
   	     	    	    

نظريه CAP



کاربردهای تجاری رایج مختلف



هadoop



یک چارچوب متن باز است که به ما اجازه ذخیره سازی حجم بالای داده بر روی ماشین های ارزان قیمت را می دهد.

هadoop



- از بلاک ها برای ذخیره سازی فایل یا قسمتی از فایل استفاده می کند.
- رویکرد یکبار نوشتن و چندین بار خواندن
- باید از سیستم های توزیع شده جدید و غیررابطه ای برای دسترسی به داده های روی آن استفاده کرد.

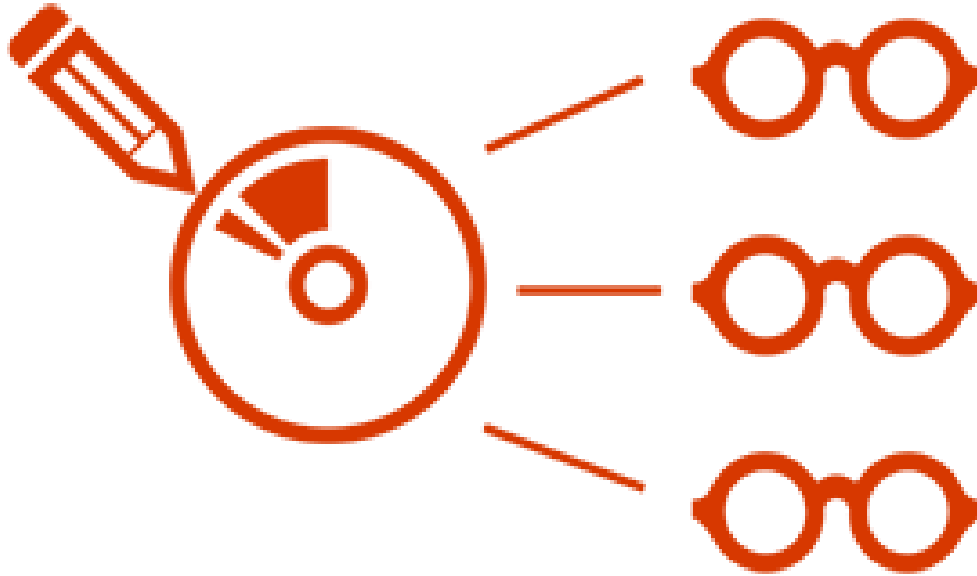
Hadoop = HDFS + MapReduce

HDFS

- مخفف Hadoop Distributed File System
- برای اجرا بر روی سخت افزارهای ارزان قیمت
- مناسب نرم افزارهای با مجموعه داده های بزرگ
- شامل صدها یا هزاران ماشین مختلف برای ذخیره سازی
- برای پردازش های دسته ای توسط کاربران طراحی شده است
- برای کاهش تاخیر زمان دستیابی به داده ها طراحی شده است

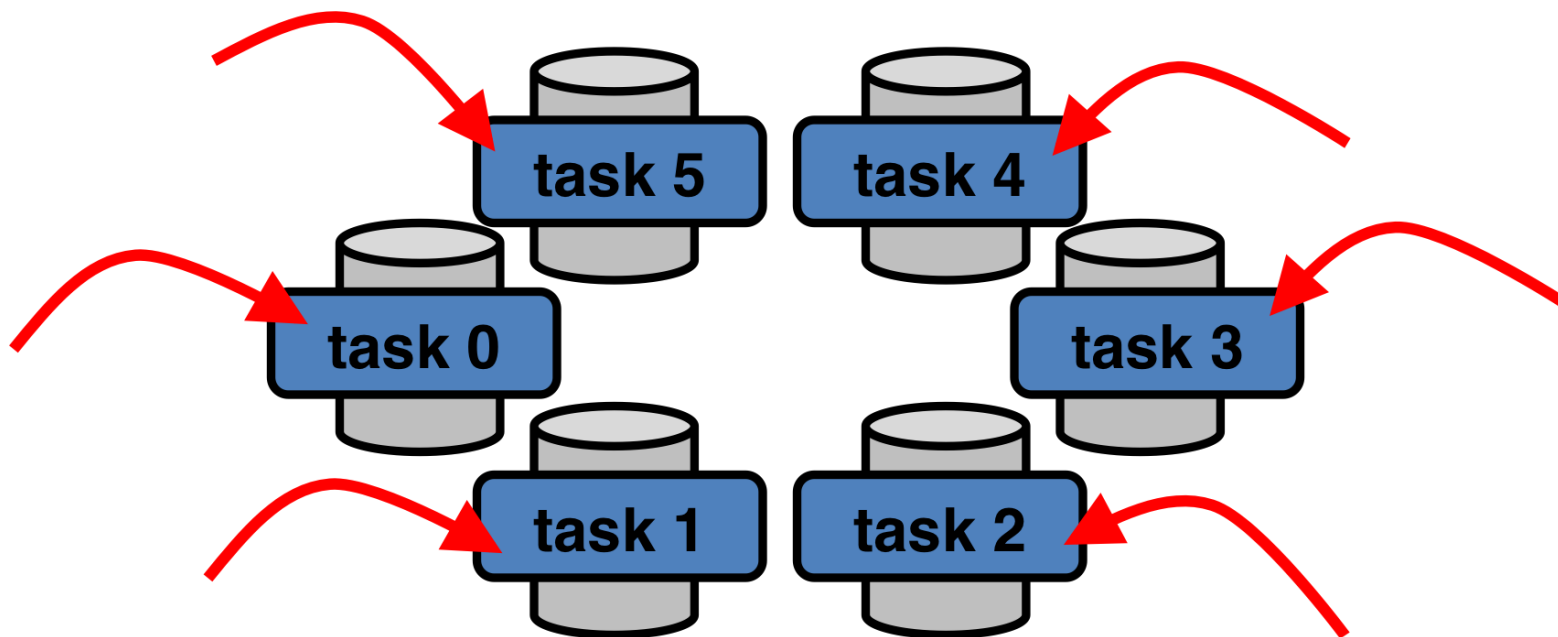
HDFS

Write once read many •



HDFS

- محاسبات متحرک بهتر از داده های متحرک!

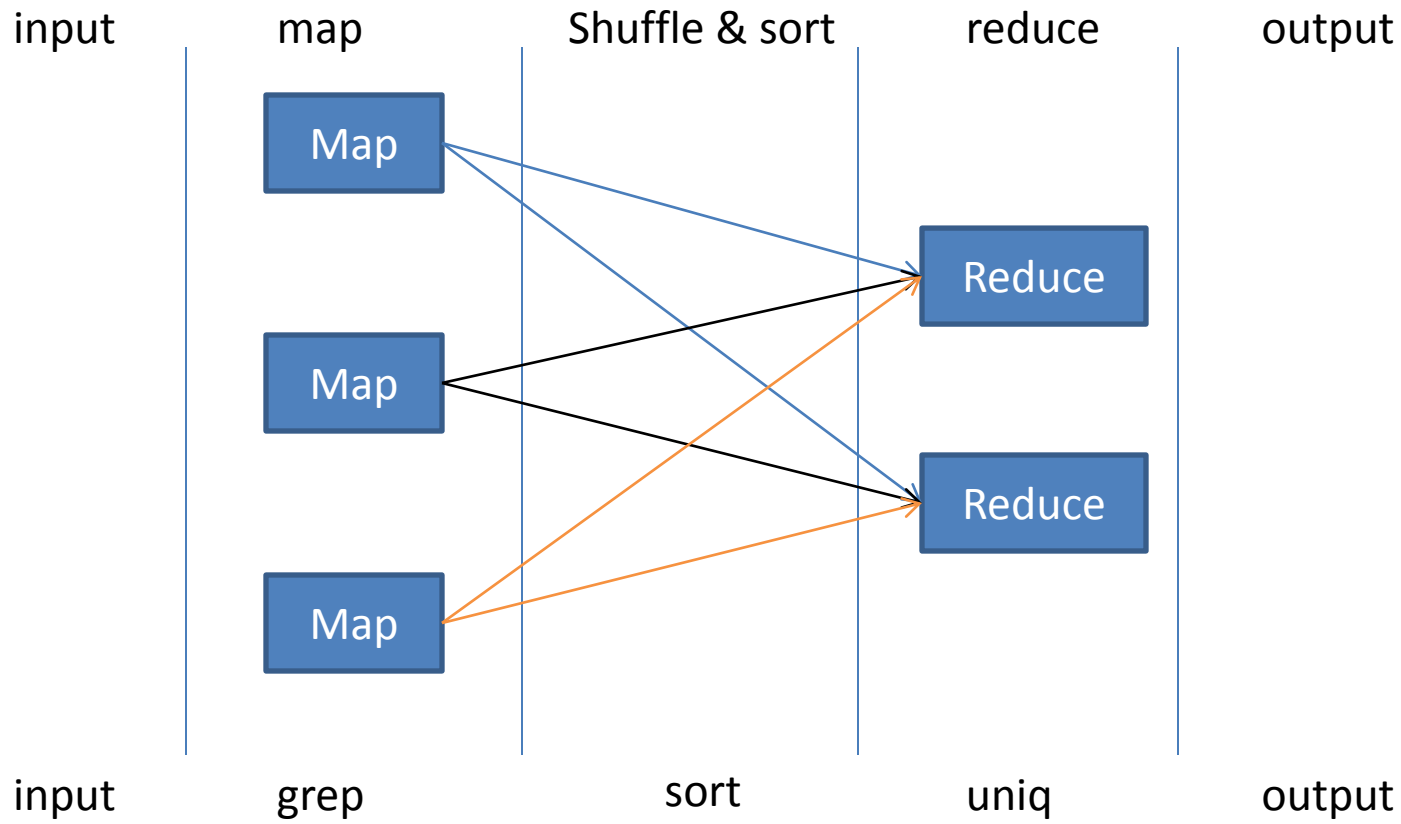


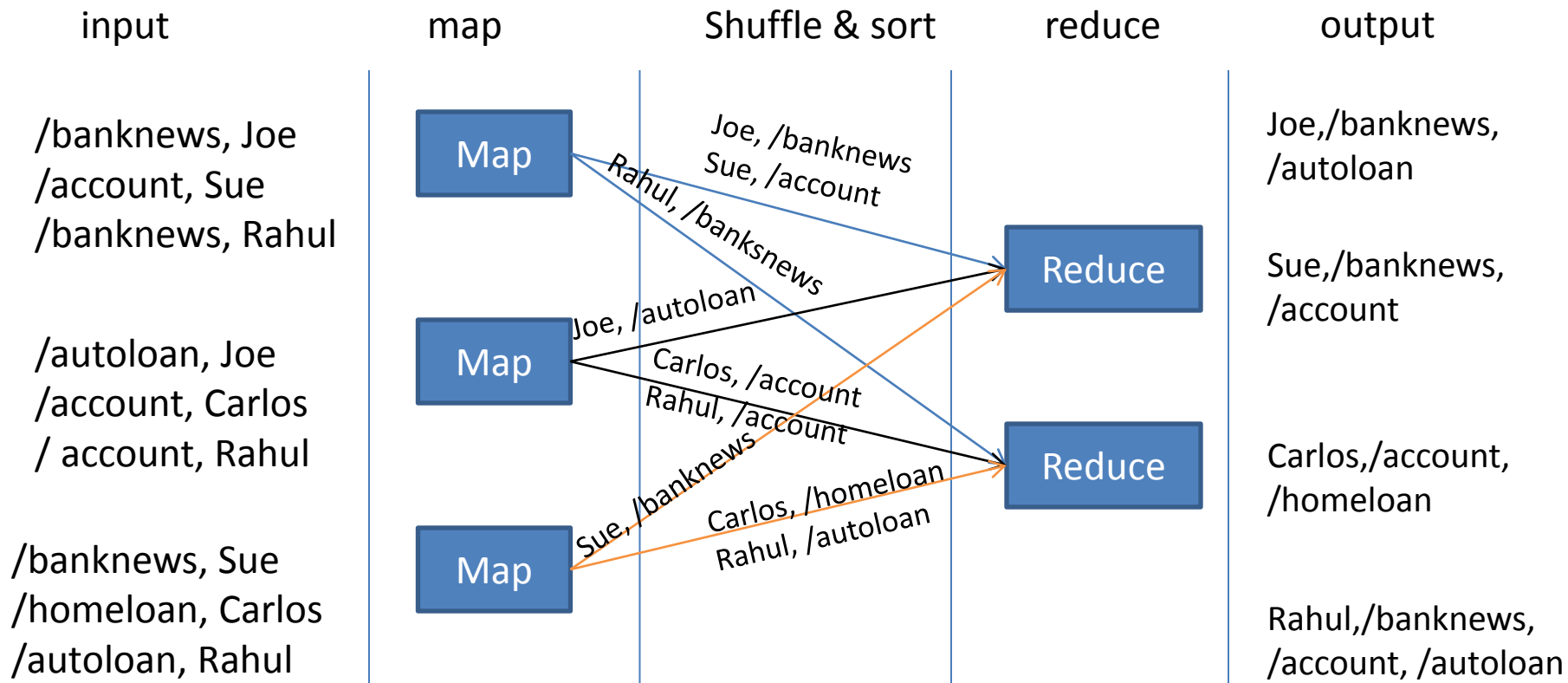
HDFS

- قابل حمل بودن بر روی سیستم ها و پلت فرم های مختلف
- معماری Master-Slave
- یک گره بنام NameNode و چندین گره بنام DataNode

MapReduce

- برای اجرای مجموعه ای از توابع روی حجم داده بالا ارائه شد.
- Map تعداد زیادی task را روی سیستم توزیع می کند.
- پس از انجام محاسبات توزیع شده، reduce تمام عناصر را باهم ترکیب و یک نتیجه تولید می کند.





یکی از بخش های پرفرمدار

بصری سازی

حجم بالای داده ها فهم آن ها را مشکل می کند در صورتیکه در حالت دیداری می توان بهتر این حجم داده آنالیز شده را درک نمود.

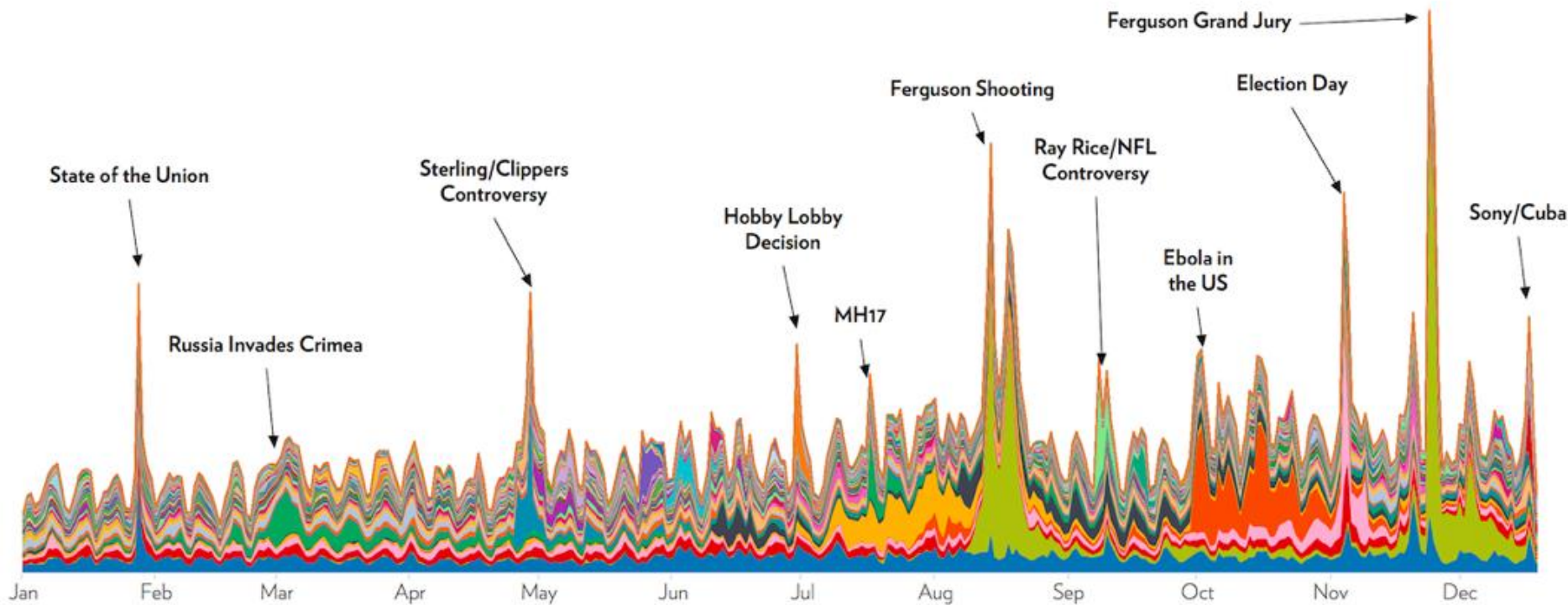
VISUALIZATION



THE YEAR IN NEWS

from ECHELON INSIGHTS

What America talked about in 2014, as viewed through 184.5 million Twitter mentions.



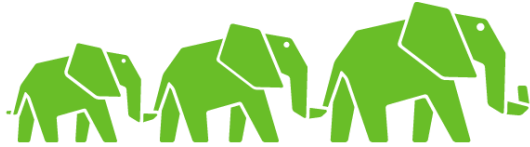
تا حالا فقط آشنا شدیم!

شروع به کار

شروع به کار

cloudera

IBM BigInsights •



Hortonworks

Hortonworks •

MAPR

MapR •



Microsoft
Azure

Microsoft Windows Azure

کاربرد در علوم مختلف



مالی و اقتصادی

- جلوگیری از سوء استفاده های مالی
- تحلیل رفتار مشتری
- تحلیل چگونگی تخصیص وام
- سیستم های پولی نوین
- تحلیل دقیق سود و زیان



زمین شناسی

- پیش بینی زلزله
- مطالعه سنگ ها و نمونه های یافت شده گذشته
- تحلیل رفتار آتش فشان ها و تغییرات زمینی



نجوم و علوم فضایی

- دستیابی به دقت های بالاتر در تصویربرداری از سیاره ها و کرات دیگر
- کشف منظومه های جدید
- تحلیل مدل های برخورد شهاب سنگ ها



انرژی پاک

- تحلیل داده های سنسور جوی و محیطی برای توربین های بادی
- تحلیل تابش های فراخورشیدی در فصول مختلف
- موقعیت یابی نصب سلول های خورشیدی در نقاط مختلف



مدیریت کشوری و سیاسی و نظامی

- پیش بینی نتایج انتخابات، جمع آوری و شمارش سریع آراء
- تحلیل و جمع آوری داده های آرایش نظامی و رفتار سربازان در جهت دفاع از مرزهای کشور
- جمع آوری داده های رفتار شهروندان و کنترل عبور و مرور شهری



ارتباطات و مخابرات

- ایجاد بسترها و شبکه های با امنیت و قابلیت اعتماد بالا
- شبکه های داده گرا و داده محور
- تحلیل شبکه های اجتماعی و سایت های اینترنتی



رسانه و خبر

- دسته بندی و نمایش خبرها و برنامه های مورد علاقه هر بیننده و تحلیل علاقه مندی او بر طبق برنامه ها و خبرهای مشاهده شده



پزشکی و داروسازی

- تجویز دقیقتر دارو بر طبق مدل زندگی، تغذیه و بیماری های ارثی
- تحلیل عواقب جانبی داروها
- کشف سرطان
- نقشه برداری ژنوم انسان و شناخت ساختار مغز



فروش محصول

- شخصی سازی فروشگاه براساس علاقه مندی ها و کالاهای مشاهده شده و پیشنهاد کالاهای مرتبط



نفت، گاز و پتروشیمی

- تحلیل داده های سنسورهای لرزه ای برای کشف چاه های نفتی
- افزایش دقت در حفر چاه های نفتی
- تحلیل داده های جغرافیایی و زمین شناسی در جهت کشف گاز



هواشناسی

- تحلیل وضعیت آب و هوا به صورت دقیق تر
- پیش بینی احتمال وقوع سیل
- پیش بینی تغییرات جوی در سال های آینده



دوست دارید بیشتر بدونید؟

گوگل کنید!

- Apache Hcatalog
- Apache Flume
- Apache Hive
- Apache PIG
- SPARK
- Mahout
- Apache Avro
- Apache Storm
- HBase
- Kafka
- R
- Zookeeper

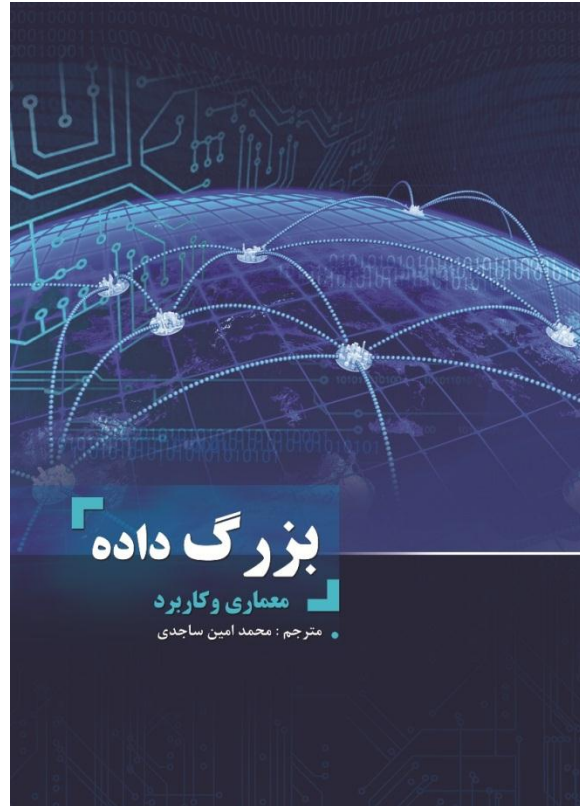
معرفی کتاب

عنوان: بزرگ داده، معماری و کاربرد

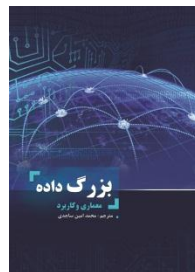
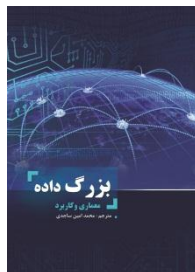
نویسندگان: Nitin Sawant, Himanshu Shah

مترجم: محمد امین ساجدی

سال چاپ: ۱۳۹۴



۲ سوال با ۳ هدیه!



جهت سفارش کتاب : <http://aminsajedi.ir/bigdata>

کد تخفیف: **scu-bigdata** فقط تا ۷ روز آینده



پرسش و پاسخ!

متشکرم!

<http://aminsajedi.ir>