نقش تحلیل کلان داده ها در بهینه سازی فرآیندهای سازمانی

محمد حیدری
ارشد مهندسی فناوری اطلاعات
دانشکده مهندسی و علوم کامپیوتر
دانشگاه شهید بهشتی

خرداد ماه سال 97

داده در حال تبدیل شدن به مهمترین دارایی برای یک سازمان است
ورنر وگلز

مهمترین منبع دنیا دیگر نفت نیست، دیتاست
مجله اکونومیست

سوال تحقیق

- ویژگی های یک متدولوژی انطباق پذیر جهت پیاده سازی فرآیند تحلیل کلان داده در سازمان به منظور نوسازی فرآیندهای سازمانی

ادبیات تحقیق

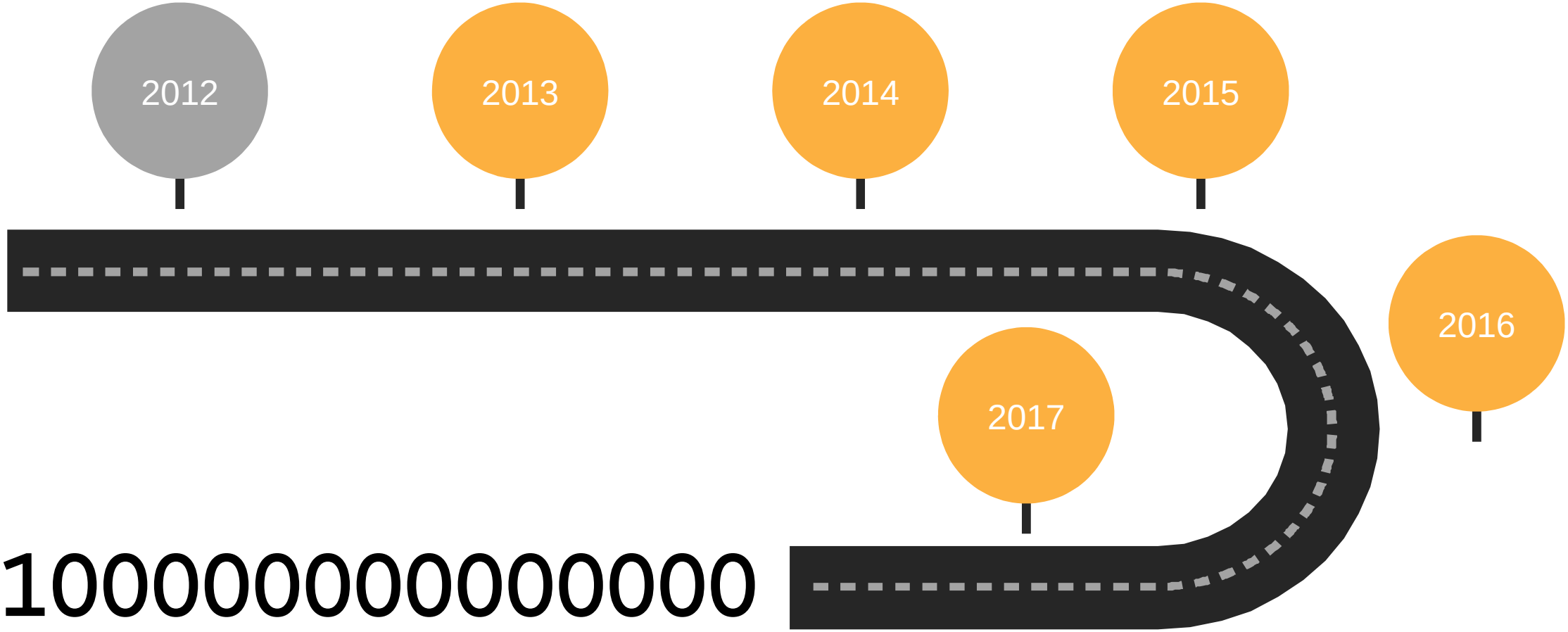
- معرفی کلان داده
- ویژگی های 8 گانه کلان داده
- تکنیک های تحلیل کلان داده
- ارتباط کلان داده با بهبود فرآیندهای درون سازمانی
- بهره مندی از فواید کلان داده ها با پیاده سازی زیرساخت های لازم
- ارائه مدل پیشنهادی به منظور پیاده سازی فرآیند تحلیل کلان داده در سازمان
- انتشار یک پژوهش در راستای تحلیل احساسات و عقاید در کلان داده های سازمانی
- نتیجه گیری


The Economist
Crunch time in France
Ten years on: banking after the crisis
South Korea's unfinished revolution
Biology, but without the cells
MAY 6TH–12TH 2017
The world's most valuable resource
Data and the new rules of competition
Google
amazon
Microsoft
Google
TESLA
TESLA

رشد صعودی تولید داده از سال 2012

از سال 2012 به بعد هر روز 1000 پتابایت داده تولید می شود.

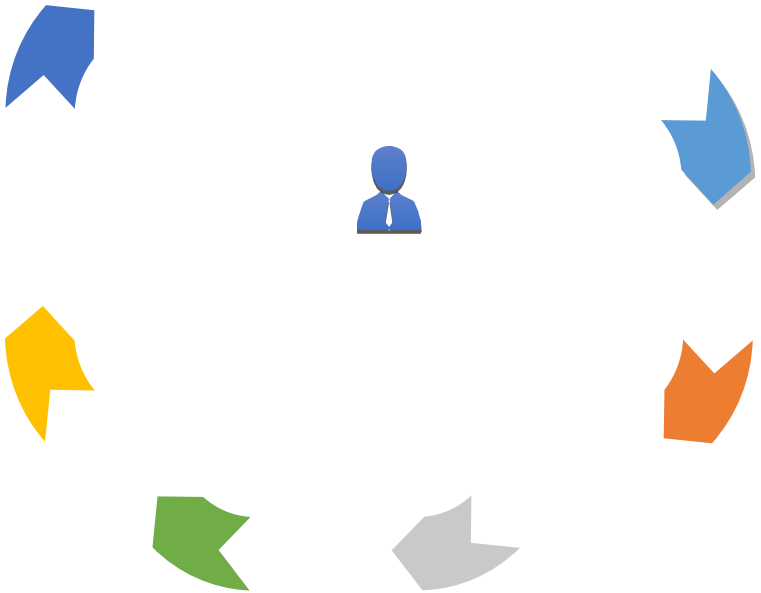
(دو برابر شدن حجم داده های تجاری به صورت سالیانه)



10000000000000000 بایت

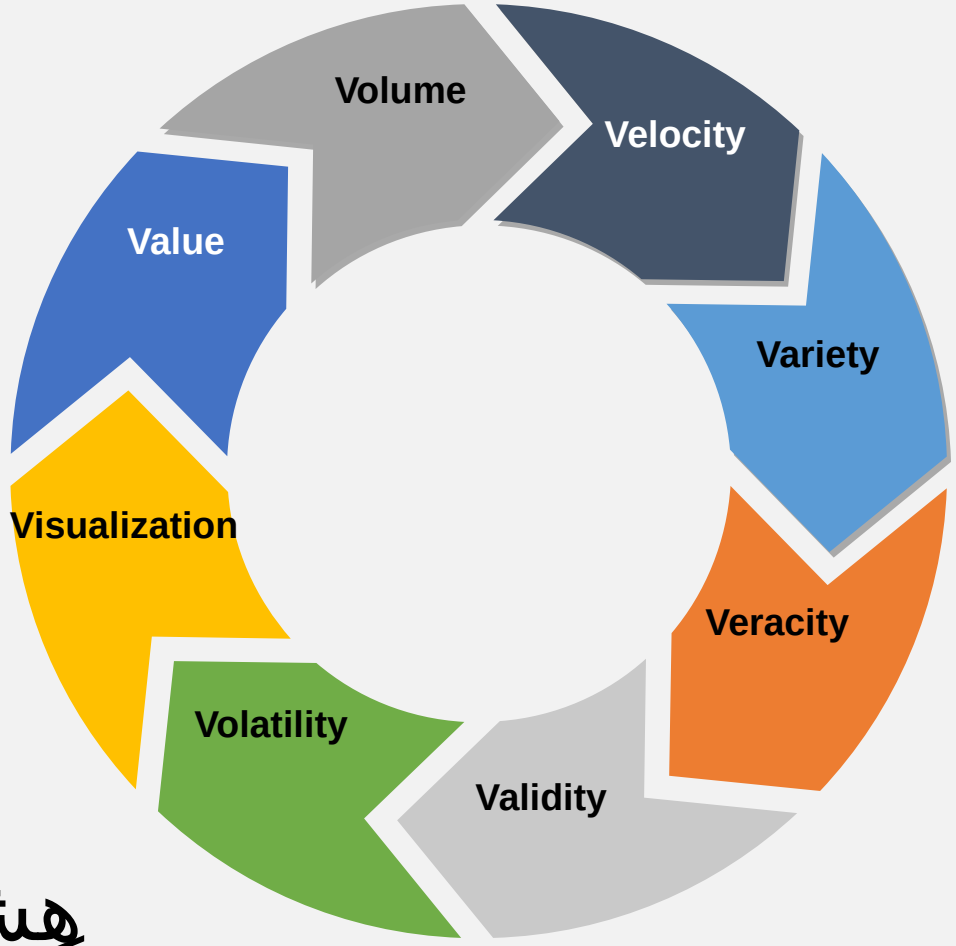
معرفی کلان داده

- کلان داده به مجموعه داده هايي اشاره دارد که با استفاده از روش هاي سنتي فناوري اطلاعات و ابزارهاي سخت افزاري و نرم افزاري موجود در آن نمي توانند در زمان معقولي درك، گردآوري، مديريت و پردازش شوند.
- کلان داده روش ها و فناوري هاي نويني را جهت جمع آوري، ذخيره و آناليز داده هاي غيرساخت يافته به صورت مقياس پذير معرفي مي کند.



ویژگی های کلان داده ها

هشت ویژگی مهم کلان داده به گزارش گروه Gartner



01 حجم داده

02 نرخ تولید

03 تنوع

04 صحت

05 اعتبار

06 نوسان

07 نمایش

08 ارزش


حجم داده
01
نرخ تولید
02
03
تنوع
04
صحت
05
اعتبار
06
نوسان
07
نمایش
08
ارزش

01 حجم داده

02 نرخ تولید

03 تنوع

04 صحت

05 اعتبار

06 نوسان

07 نمایش

08 ارزش

01 حجم داده

02 نرخ تولید

03 تنوع

04 صحت

05 اعتبار

06 نوسان

07 نمایش

08 ارزش

01 حجم داده

02 نرخ تولید

03 تنوع

04 صحت

05 اعتبار

06 نوسان

07 نمایش

08 ارزش

01 حجم داده

02 نرخ تولید

03 تنوع

04 صحت

05 اعتبار

06 نوسان

07 نمایش

08 ارزش

01 حجم داده

02 نرخ تولید

03 تنوع

04 صحت

05 اعتبار

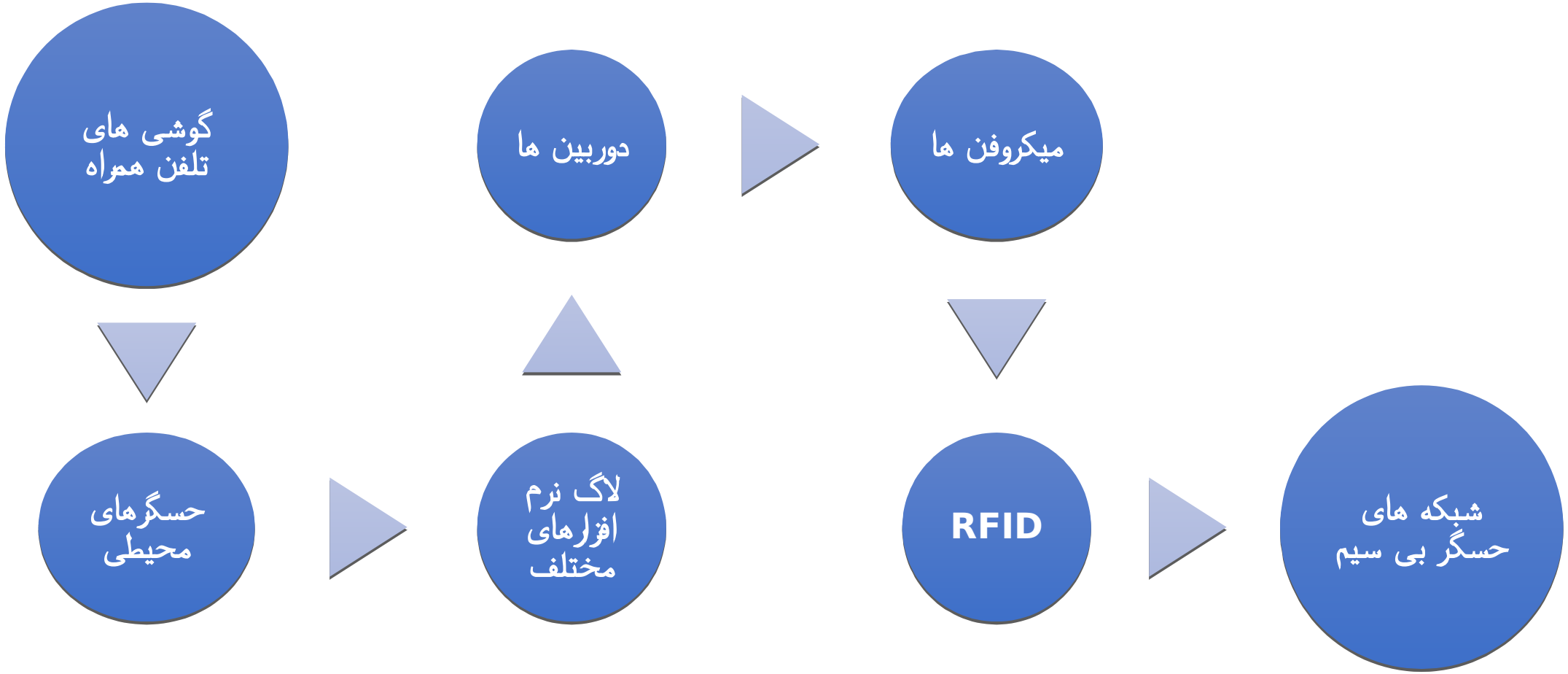
06 نوسان

07 نمایش

08 ارزش


01
حجم داده
02
نرخ تولید
03
تنوع
04
صحت
05
اعتبار
06
نوسان
07
نمایش
08
ارزش

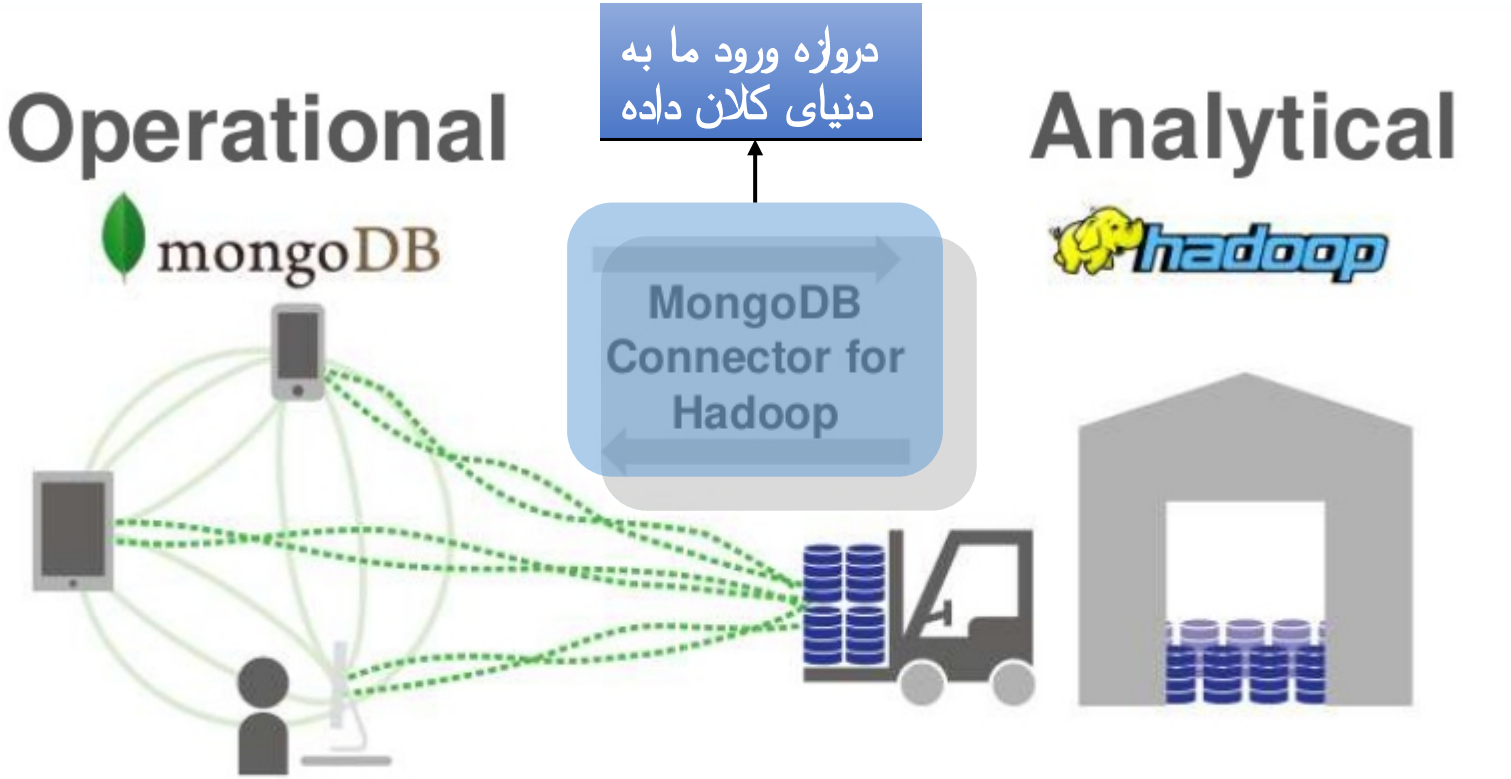
کلان داده ها چگونه تولید و جمع آوری می شوند ؟



مدل هاي داده مورد استفاده براي داده هاي عظيم

NoSQL (Column, Document, Key-value, Graph)	XML	رابطه ای
پشتيباني از ارتباط ها	انعطاف پذير اما امنيت پايين	ساخت يافته
نسبتاً مقياس پذير	ساختار درختی	ضعف در پشتيباني از ارتباط ها
در اغلب كاربردهای فعلي داده هاي عظيم نظير شبكه هاي اجتماعي		عدم مقياس پذيري

اما چرا پایگاه داده های غیر رابطه ای Nosqlمهم هستند ؟



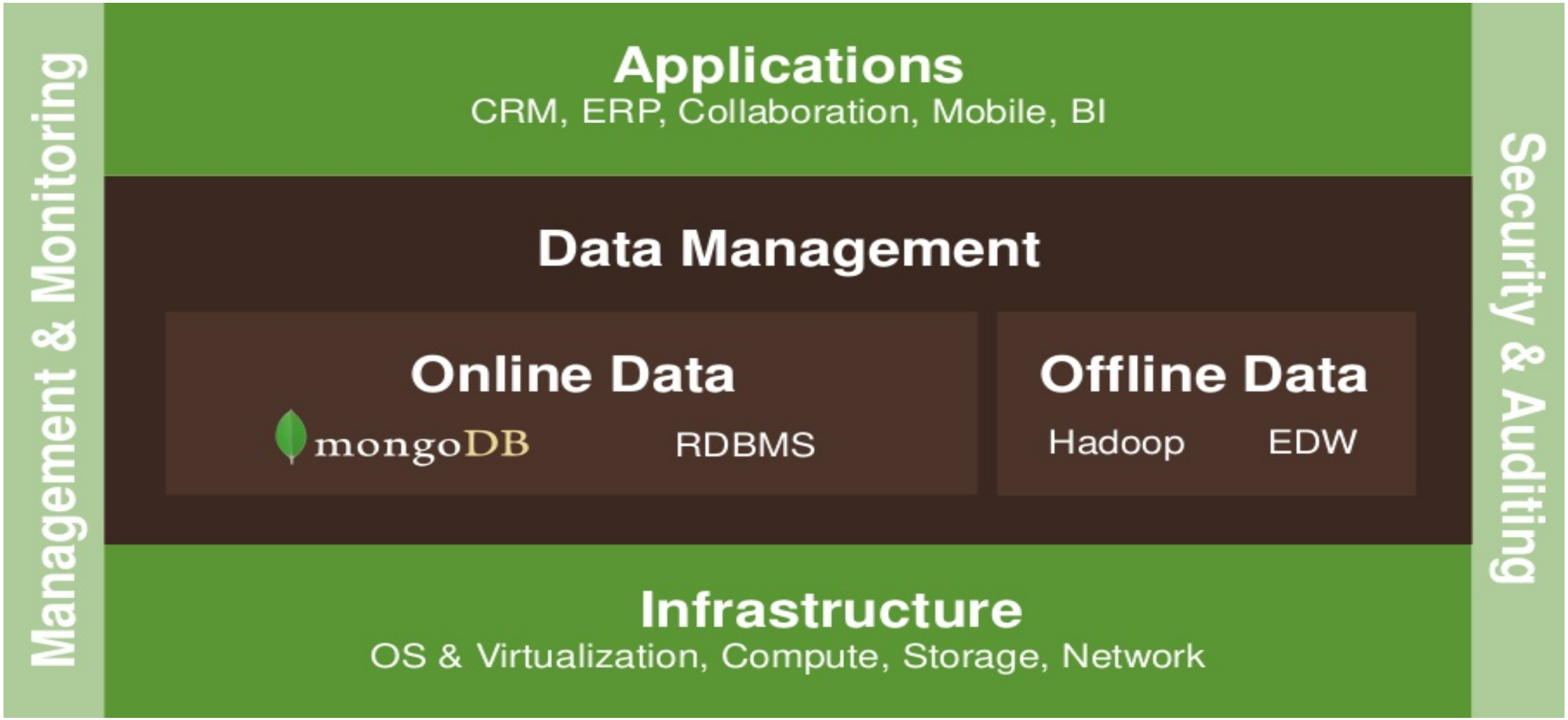
- Online, Real-time
- High concurrency & HA
- Live analytics

- Multi-source analytics
- Interactive & Batch
- Data lake

کاربردهای تحلیل کلان داده در انواع سازمان ها

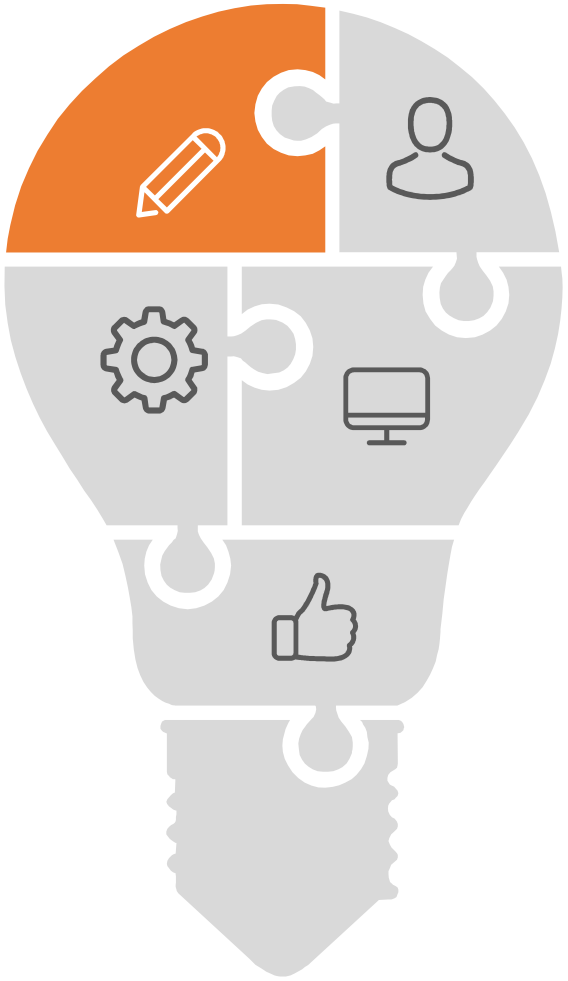
- نرم افزارهای امنیتی: مثلا نرم افزاری مانیتورینگ شبکه
- سیستم های مدیریتی: مثلا مدیریت ارتباط با مشتریان یا CRM
- علوم اجتماعی و سیاسی: مثلا پیش بینی یا تحلیل نتایج انتخابات
- مالی و اقتصادی: مثلا پیش بینی قیمت یک یا چند سهام یا شاخص
- سیستم های بانکی: مثلا تخصیص اعتبار به مشتریان و طبقه بندی آن ها
- علوم پزشکی: مثلا پیش بینی خطرات احتمالی ناشی از یک عمل جراحی خاص
- برنامه ریزی و مکان یابی: مثلا چیدن داخلی فروشگاه های بزرگ و یا تخصیص امکانات شهری

Enterprise Big Data Stack



تکنیک های تحلیل کلان داده ها

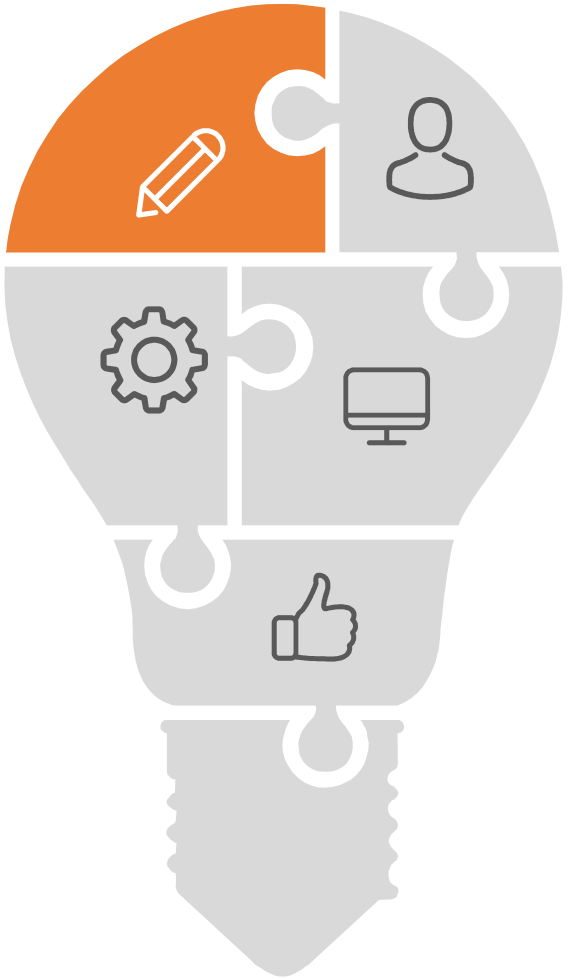
Machine Learning - ML



یادگیری ماشین

- یادگیری تحت نظارت
  - در دسترس بودن دیتاست
  - داده ها برچسب گذاری شده اند
- یادگیری بدون نظارت
  - فقدان دیتاست
  - داده ها برچسب گذاری نشده اند
  - سیستم باید خودش یاد بگیرد

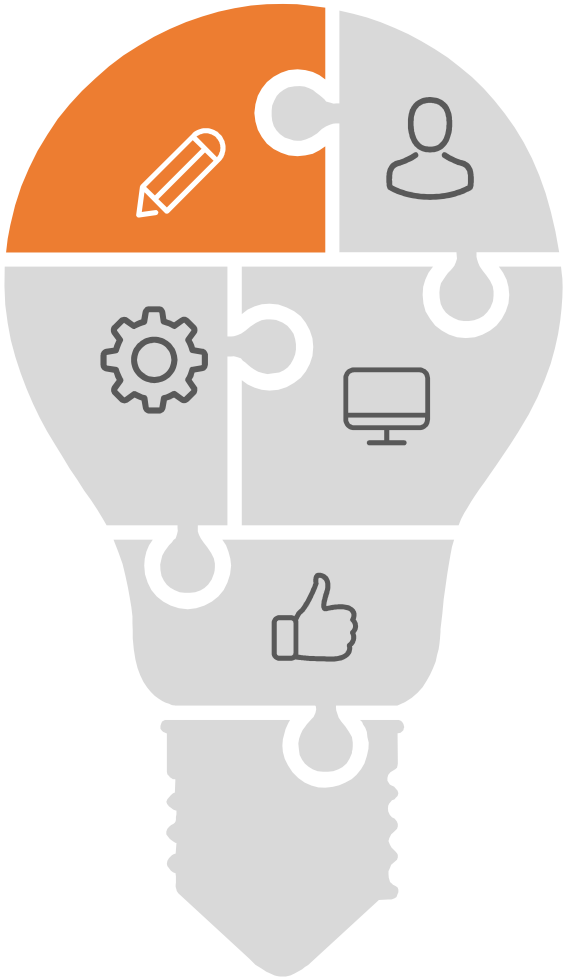
Machine Learning - ML



یادگیری ماشین

- طبقه بندی – Classification
  - پیشنهاد خواندن کتاب آشنایی با کلان داده
- یعنی شما را در کلاس افراد علاقه مند به این مبحث طبقه بندی کرده است.

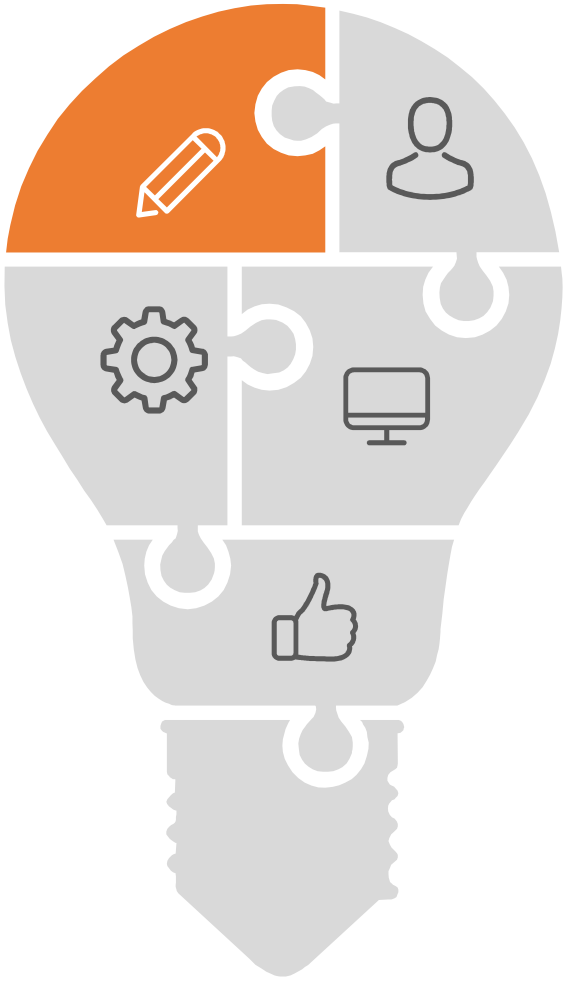
Machine Learning - ML



یادگیری ماشین

- خوشه بندی Clustering
  - مساله از قبل حل نشده است.
- تقسیم بندی فرآیندهای سازمانی در 3 خوشه :
  - حیاتی، مهم، معمولی
- **فرآیندهایی که در یک خوشه قرار می گیرند بیشترین شباهت را به هم دارند و میان خوشه 1 و 3 بیشترین تفاوت ممکن وجود دارد.**

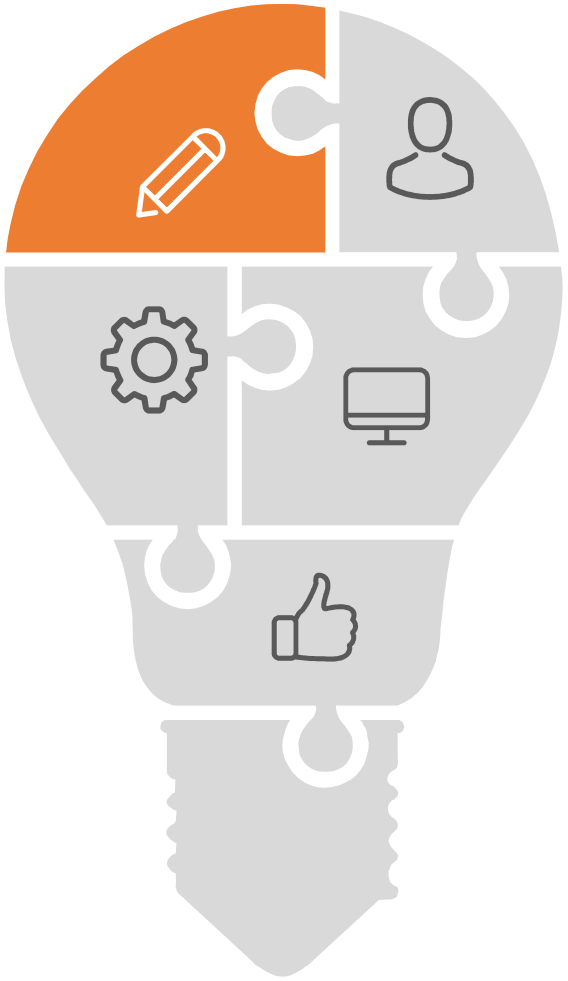
Machine Learning - ML



یادگیری ماشین

- رگرسیون Regression
- هدف از حل مساله رسیدن به یک رابطه ریاضی برای توصیف یک پدیده است.
- رابطه میان ساعت مراجعه به یک سایت، محل سکونت، سن، سوویس ایمیل مورد استفاده و مقدار سفارش انجام شده توسط یک کاربر.

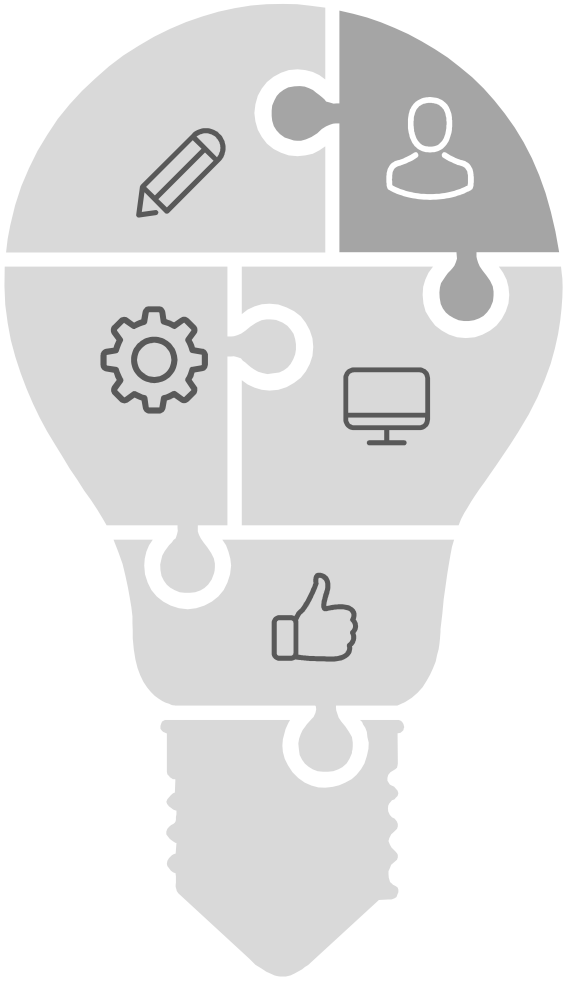
Machine Learning - ML



یادگیری ماشین

- کاوش قواعد وابستگی Association Rule Mining
  - مثال یک قاعده :
  - اگر فردی لپ تاپ و موس خریداری کند حتما ماوس پد را هم تهیه خواهد کرد.
- **پس سعی می شود لوازم 3 گانه فوق حتی المقدور در کنار هم در ویترین نمایش یابند.**

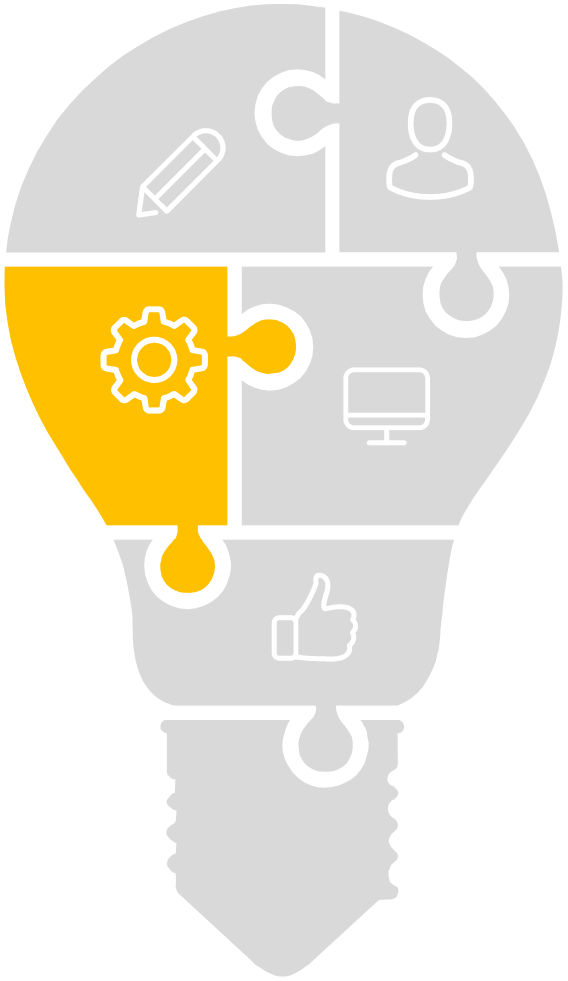
Artificial Neural Network - ANN



شبکه های عصبی مصنوعی

- تشخیص الگو
  - Pattern Recognition
- تجزیه و تحلیل تصاویر
  - Computer Vision

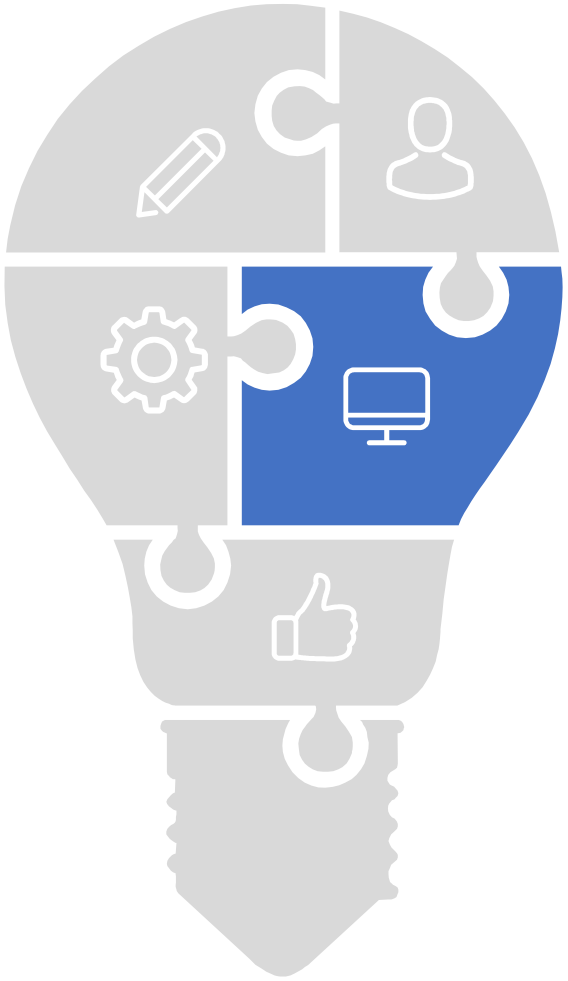
Natural Language Processing - NLP



پردازش طبیعی متن

- متن کاوی
  - پیش بینی بازار سهام
  - کشف تقلب اسناد مدیریتی
- تحلیل احساسات پنهان در متن
  - سطح سند
  - سطح جمله
  - سطح بُعد : نظر مشتری پیرامون ویژگی های مختلف یه موبایل

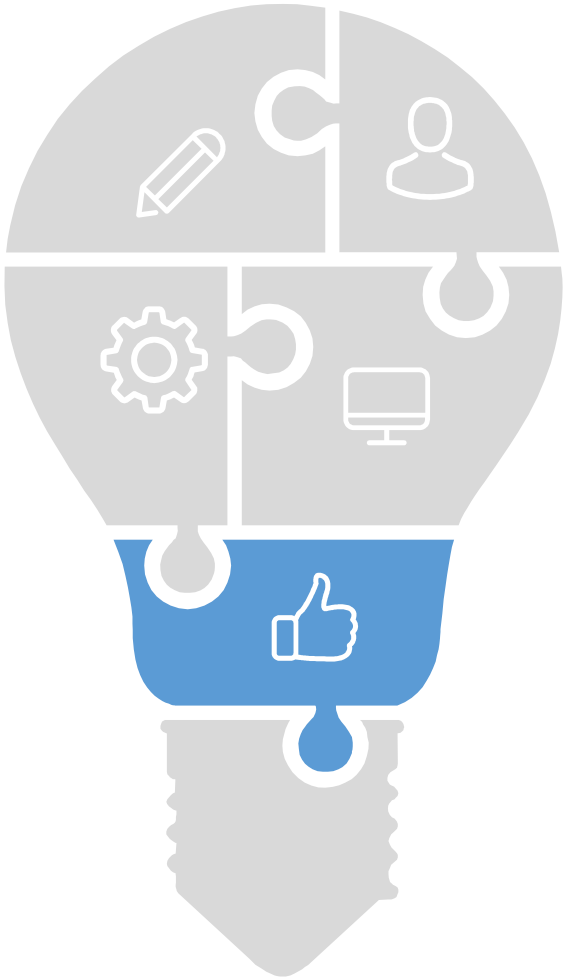
Visualization



مصورسازی

- **Google Chart**
- **Tableau**
- **Microsoft Power BI**
- **Oracle Visual Analyzer**

Data Mining



داده کاوی

استخراج اطلاعات و دانش و کشف الگوهای پنهان از پایگاه داده‌های بسیار بزرگ با بهره مندی از انباره داده

کلان داده چطور سازمان را تحت تاثیر قرار می دهد؟

برقراری ارتباط نزدیک تر با مشتریان

بهینه سازی فعالیت های کاری کارکنان

تصمیم گیری کارکنان سازمان با بینش وسیع تر

سوق دهی مدیران فناوری اطلاعات به استفاده از ارزش های پنهان در داده ها

متمرکزسازی سرمایه های فکری و فنی کارکنان سازمان بر روی منبع جدیدی از سود سرشار پنهان در داده ها

بهره مندی از کلان داده ها

نصب، راه‌اندازی و تنظیمات زیرساخت

طراحی و مدیریت انواع پلیگاه‌های داده‌ای

پردازش جریان داده‌ها

انجام محاسبات و یادگیری ماشین

به‌کارگیری انواع واسط SQL

انتقال داده‌ها

پیام رسانی و مدیریت صف

طراحی و مدیریت جست و جو و شاخص گذاری

مدیریت Log Files

نصب، راه‌اندازی و تنظیمات زیرساخت‌های بیگ دیتا

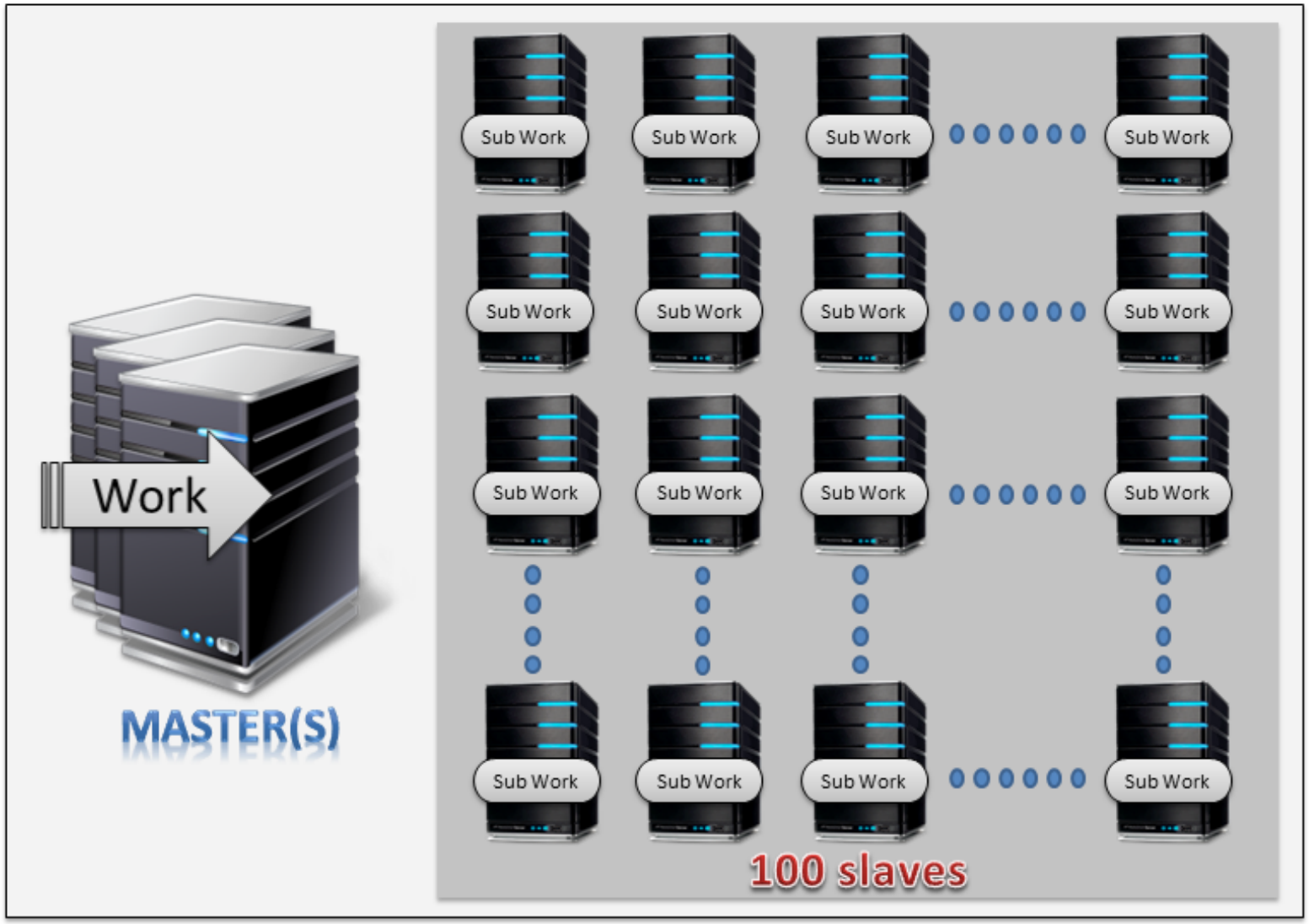
Apache Hadoop یک چهارچوب برنامه نویسی جاوا، جهت توسعه و ذخیره سازی داده هایی با حجم بسیار بالا توسط کمپانی Apache. ساخته شده است

معمولاً پردازش های را به صورت توزیع شده یعنی بر روی چند کامپیوتر مختلف انجام می دهد و نتایج را به کامپیوتر مقصد برمیگرداند.

در این نرم افزار داده ها در یک سیستم فایلی خاص به نام HDFS ذخیره می شوند.

از مدل برنامه نویسی Map Reduce استفاده می کند.


USER


Work
MASTER(S)
Sub Work
100 slaves

معماری Apache Hadoop

- هدوپ به سبک Master/Slave کار می‌کند.
- در کلاستر هدوپ یک گرهMaster وجود دارد و تعداد زیادی گره Slave
- گره Master، گره‌هایSlave را مدیریت، حفظ و کنترل می‌کند
- در حالی که گره‌های Slave، مولفه‌های واقعی انجام کار هستند.
- گره Master، فقط فراداده‌ها (داده‌های درباره داده) را ذخیره می‌کند
- در حالی که Slaveها گره‌هایی هستند که داده‌ها را ذخیره می‌کنند

Map Reduce

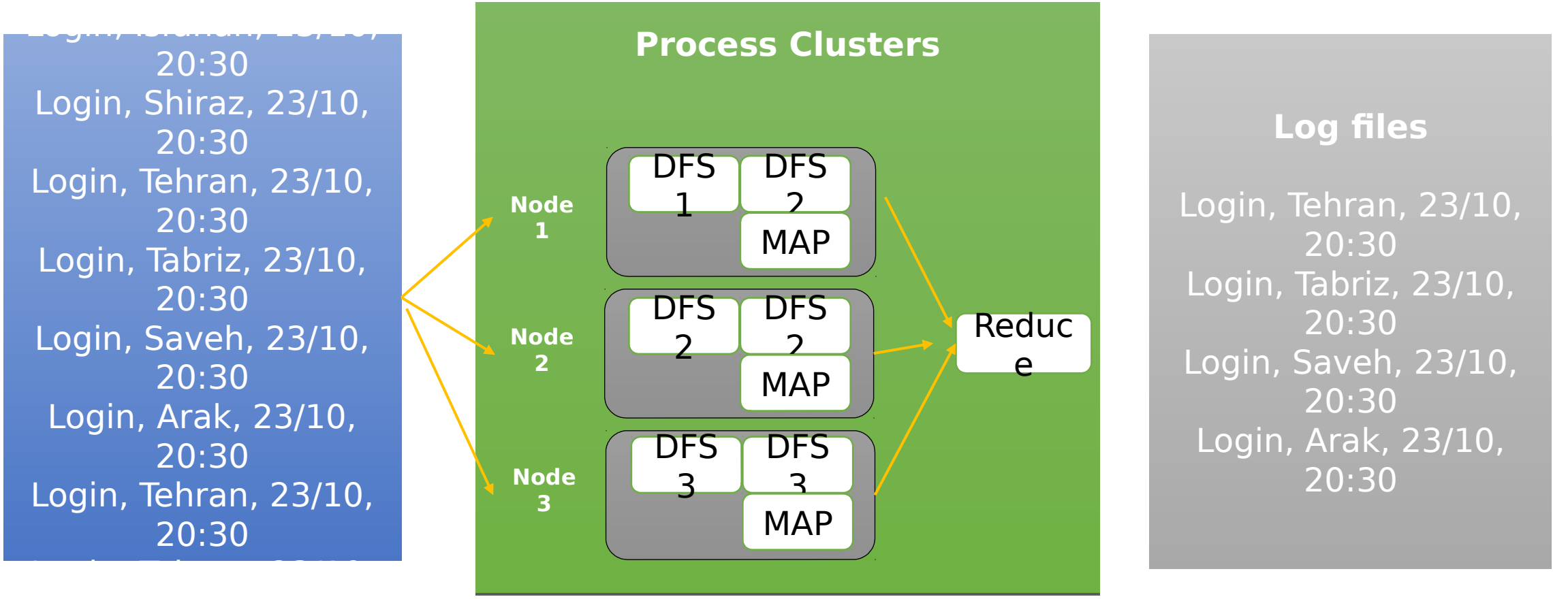
این تابع داده ها را از ورودی دریافت می کند و آنها را برروی پردازشگرهای مختلف توزیع می کند. سپس هر پردازشگر داده های گرفته شده را به صورت مستقل محاسبه می کند و آنها را بصورت **key/value** ذخیره می کند.

Map (k1, v1) -> list (k2, v2)

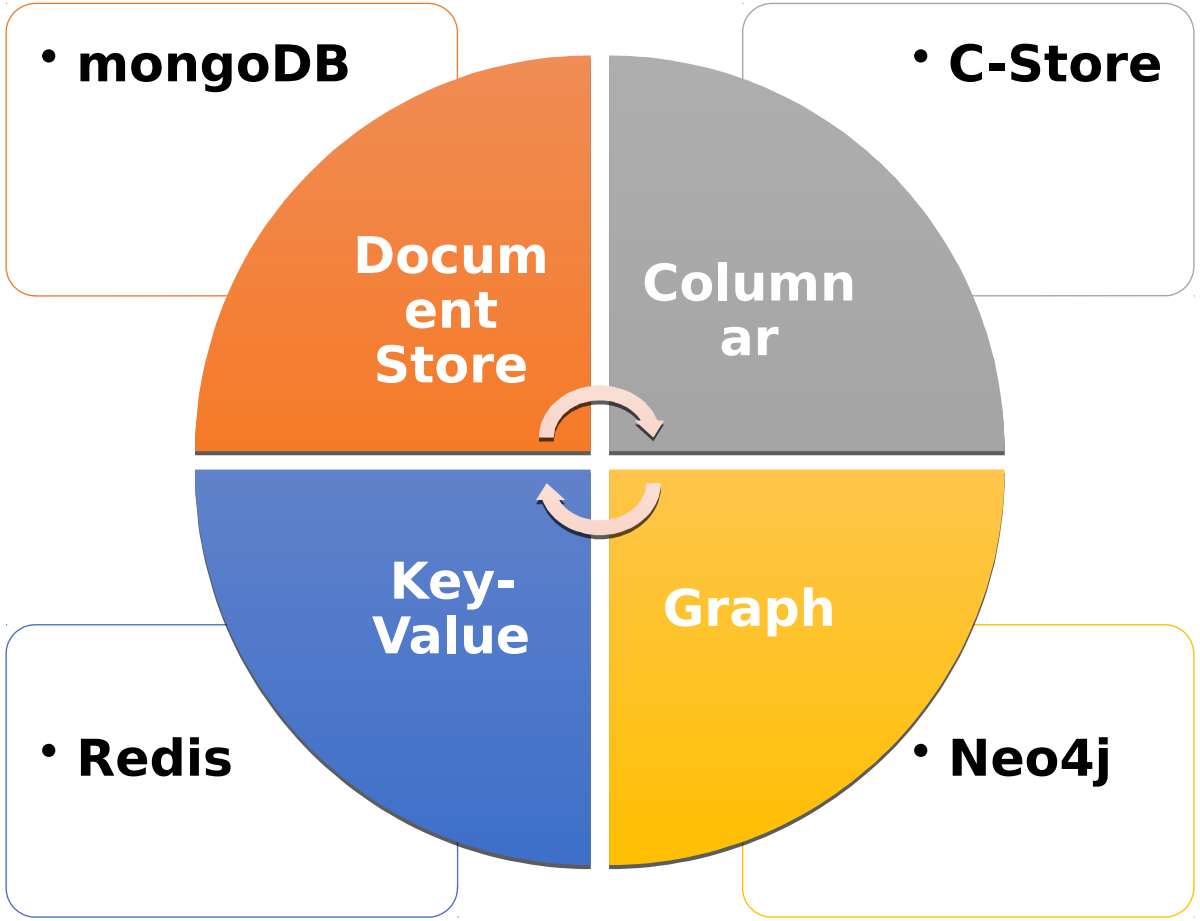
Reduce (k2, list (v2)) -> list(v3)

همه داده ها باید بر روی یک پردازش واحد، جمع آوری و شاخص گذاری شوند. به همین دلیل از تابع **Reduce** استفاده می کنیم. در تابع مذکور هر**Key** مجموعه ای از محاسبات در پردازشگرهای مختلف دارد که آنها را با یکدیگر ترکیب می کند و به صورت**Out_value** در خروجی نشان می دهد.

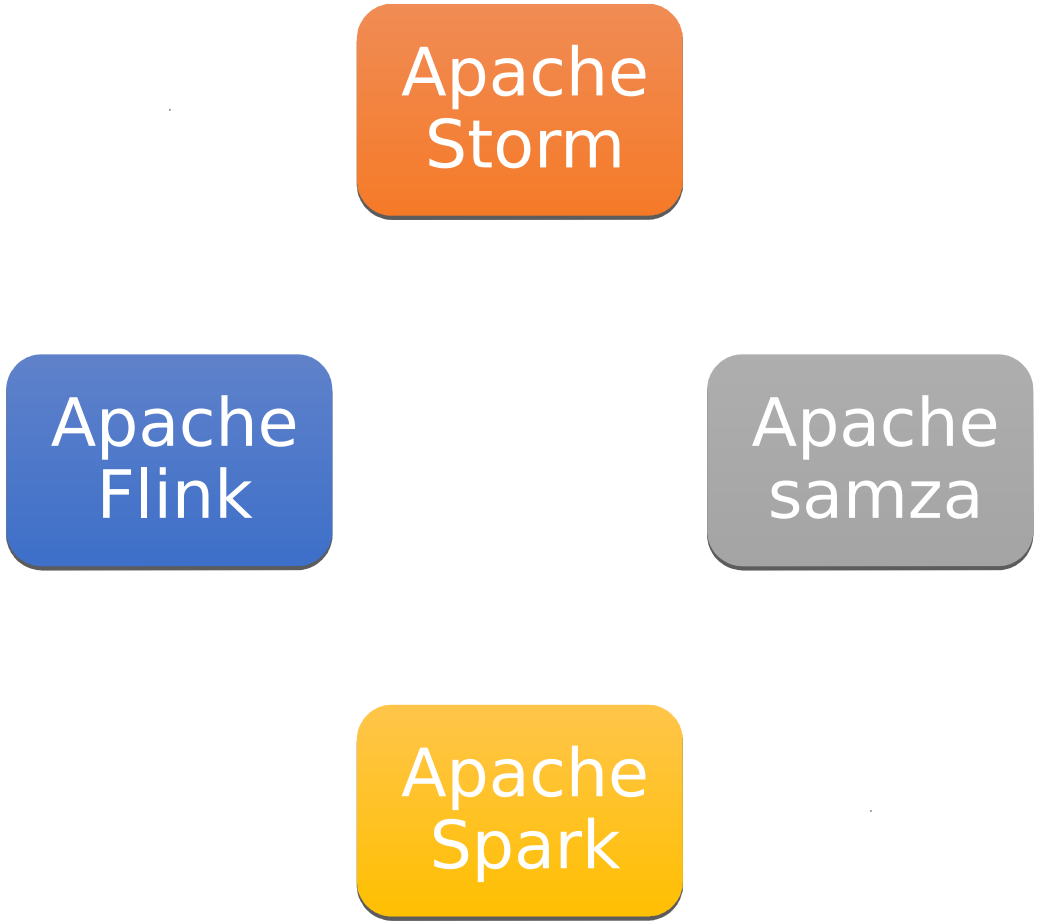
یک مثال از کارکرد Hadoop



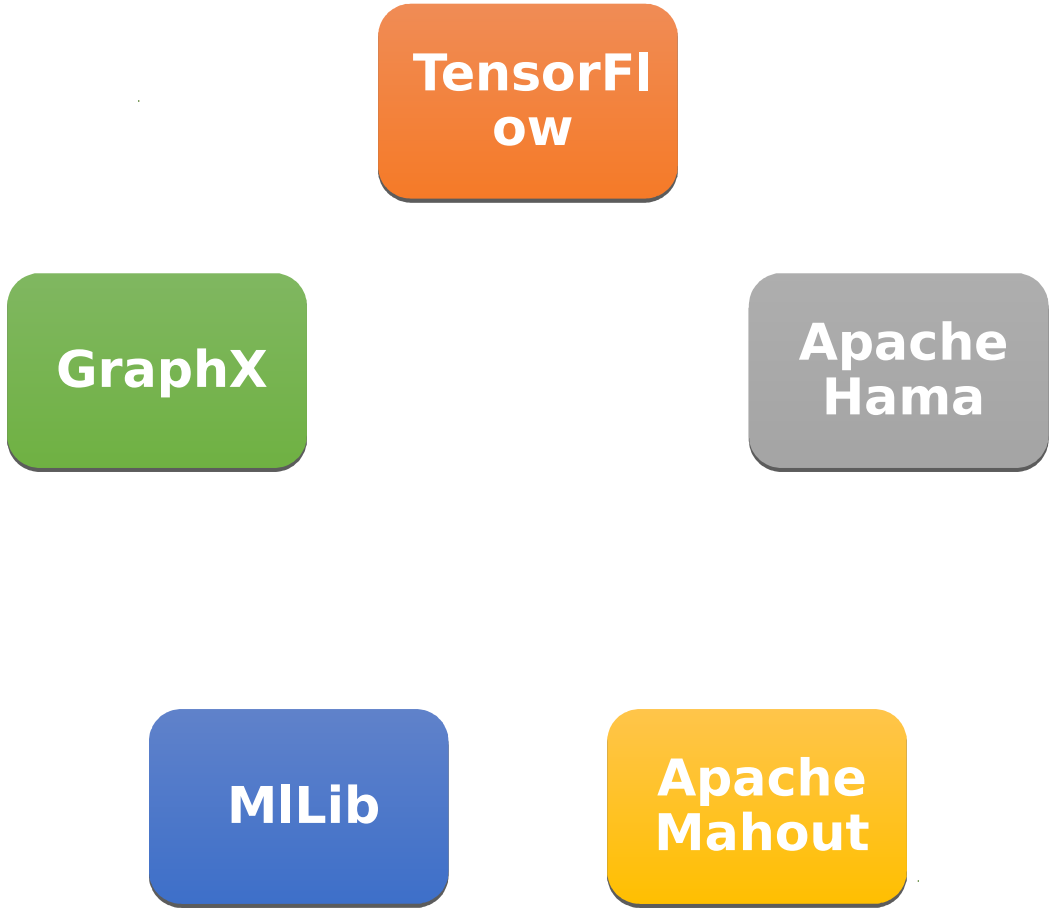
طراحی و مدیریت انواع پایگاه های داده



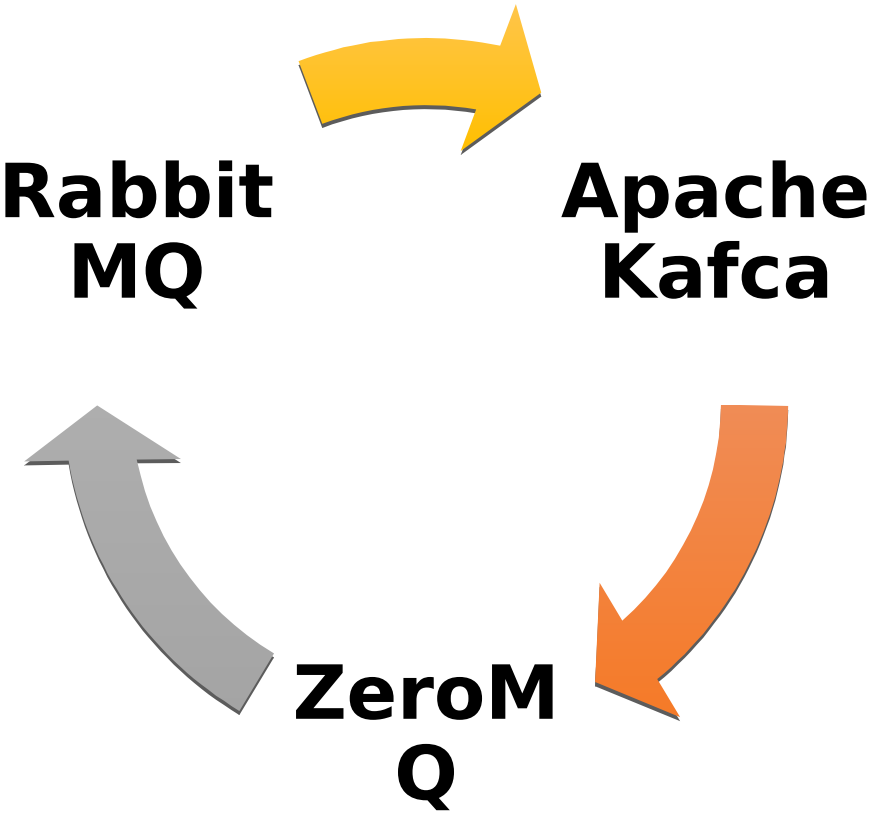
پردازش جریان داده



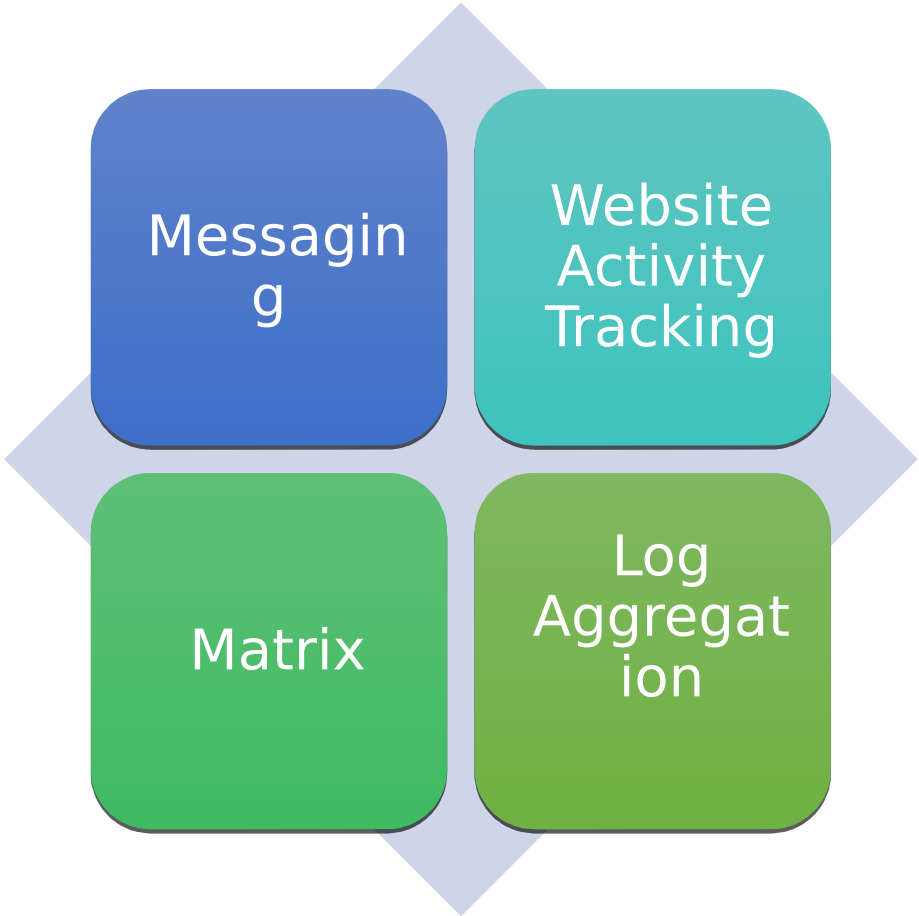
انجام محاسبات و یادگیری ماشین



پیام رسانی و مدیریت صف



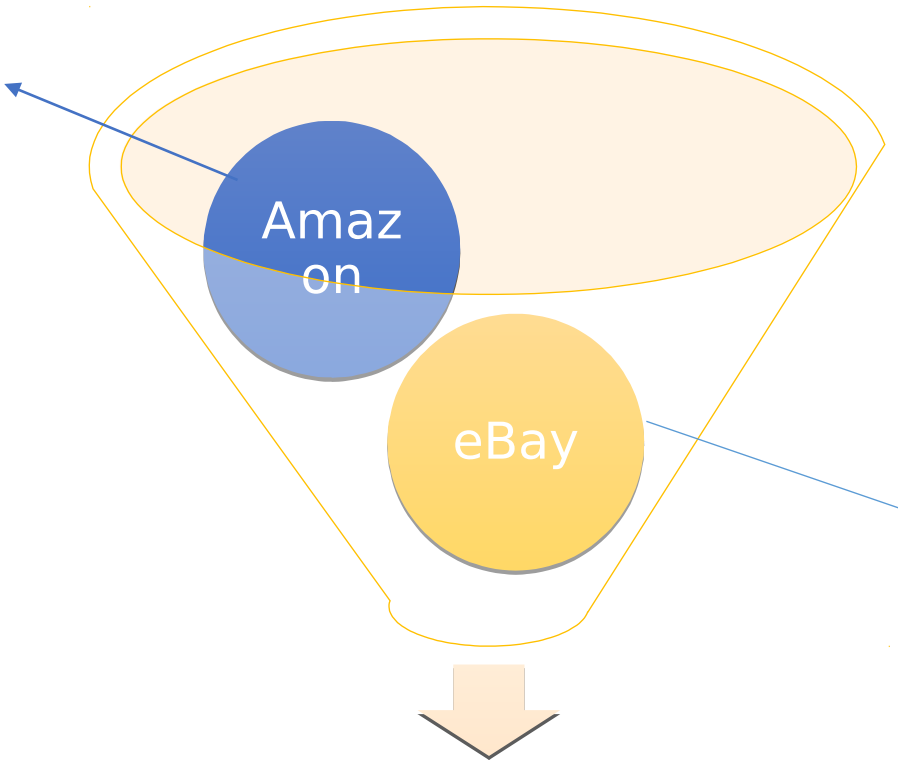
پیام رسانی و مدیریت صف




amazon

کلان داده ها در بخش خصوصی

سه دیتابیس بزرگ جهان مبتنی بر لینوکس را از آن خود کرده است.



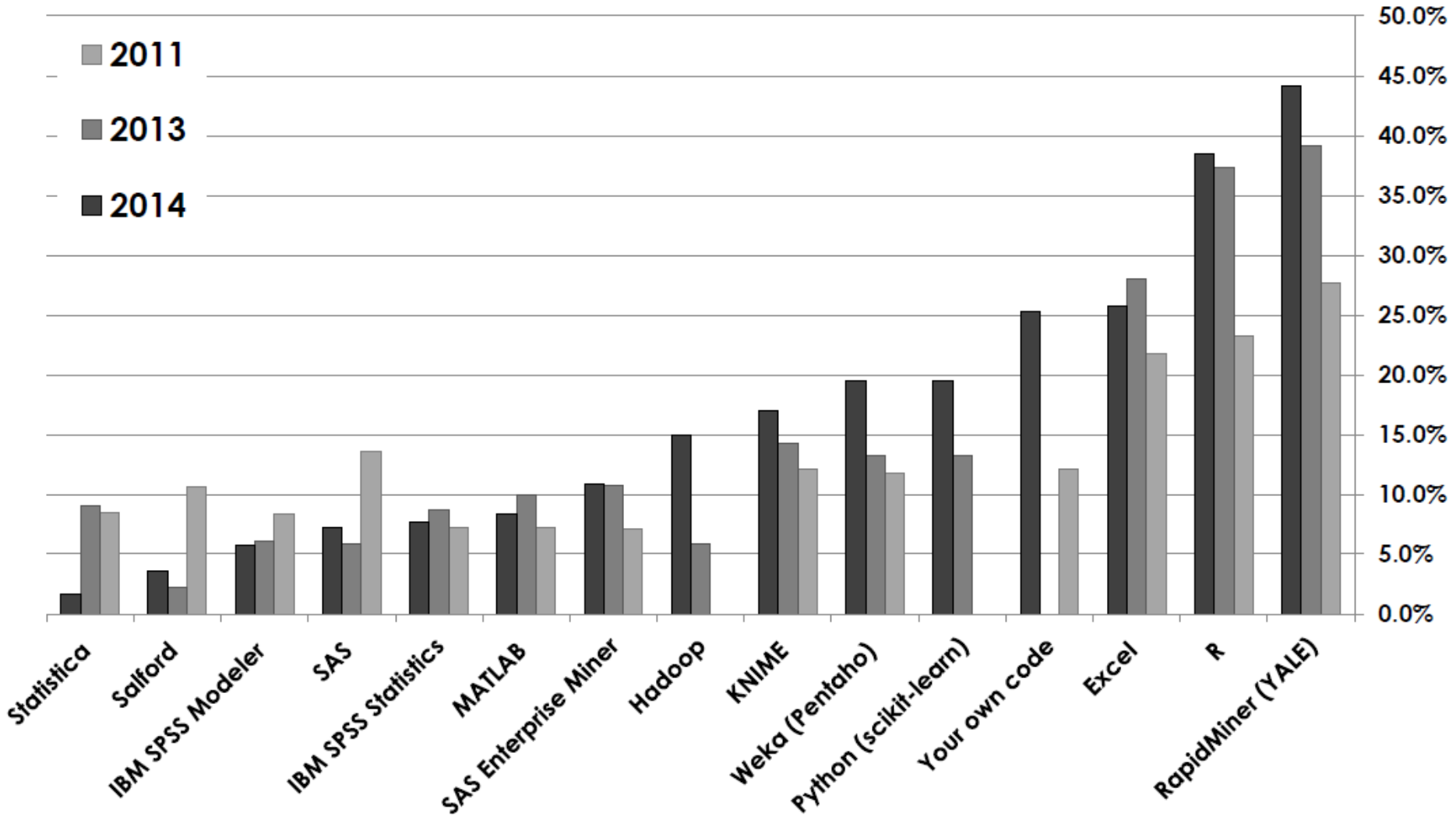
بهره مندی از سیستم های توصیه گر
با تکیه بر دو انباره داده 7.5 پتاباتی
به همراه دو انباره 40 پتابایتی

استخراج گنج های مخفی از معدن دیتاسنترهای عظیم بر بستر **Hadoop**

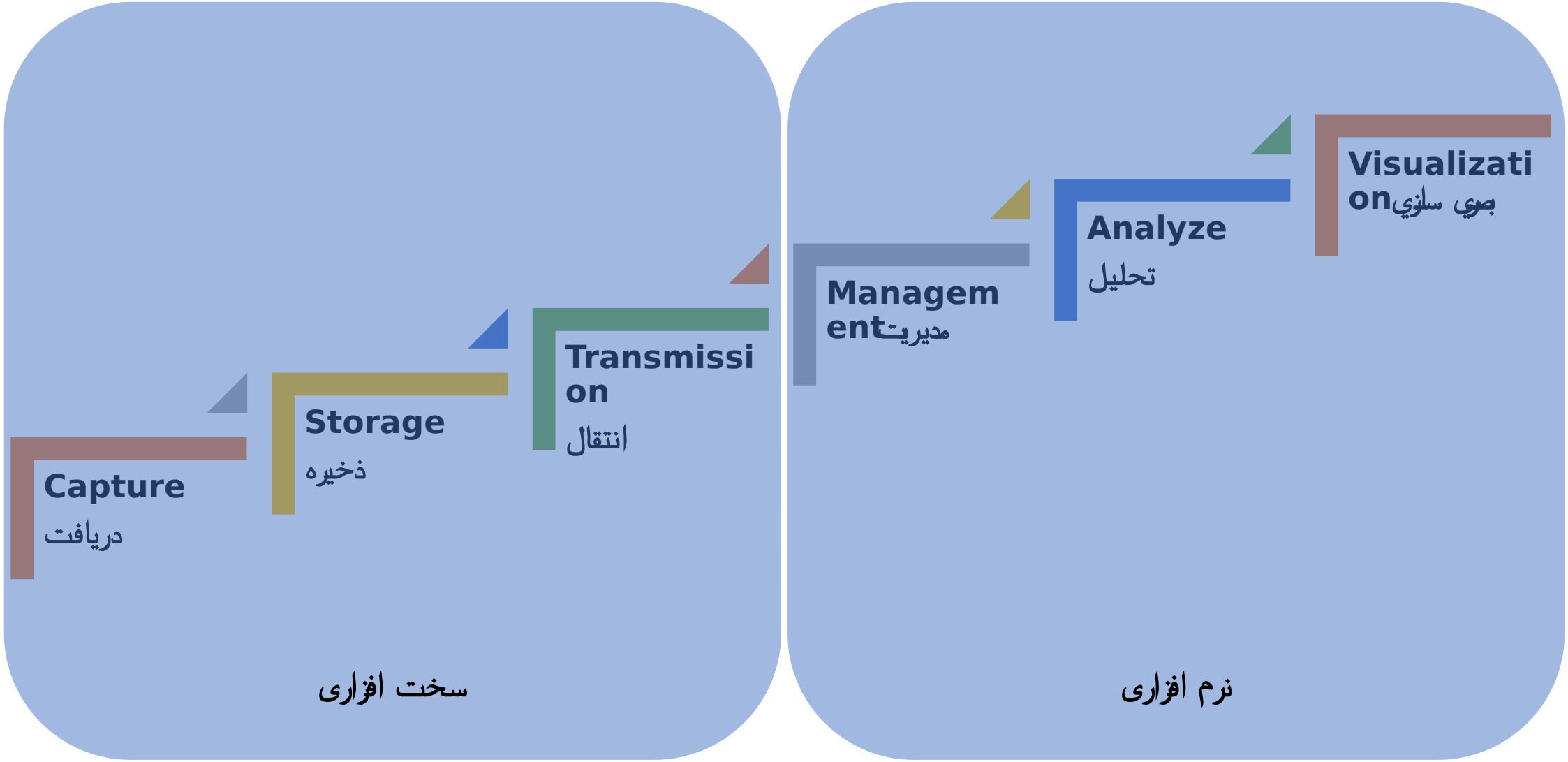
کشف الگو از فرآیند خرید در سیستم های توصیه گر

- آمازون یکی از قوی ترین سیستم های توصیه گر را در اختیار دارد.
- افرادی که تلفن همراه خریداری می کنند در کنار آن قاب و برچسب ضد ضربه نیز سفارش می دهند یک الگو یا Pattern است که در گذر زمان سیستم توصیه گر آمازون آن را شناسایی و کشف کرده است.
- همین مفهوم را می توان در سازمان نهادینه کرد. به شکلی که همجواری ارائه خدمات به سوی مشتری با تفکری مناسب پیاده سازی گردد

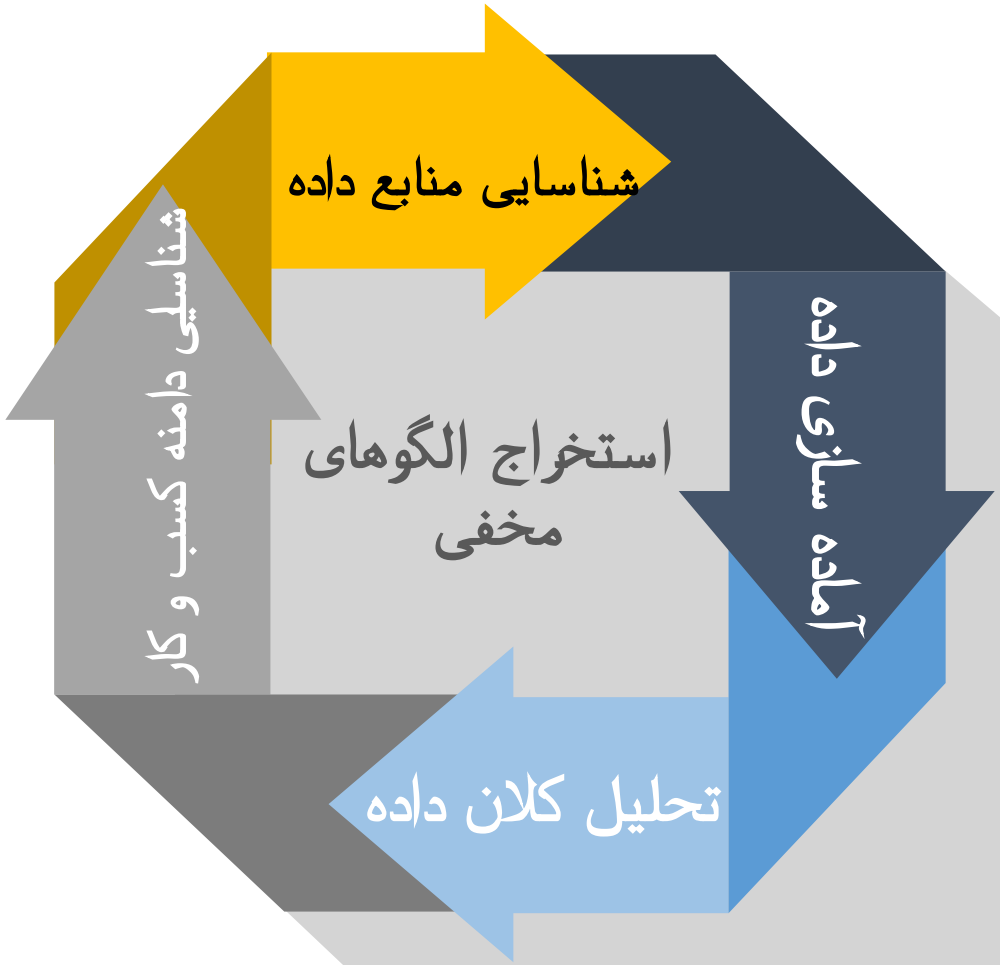
مهم ترین ابزارهای تحلیل کلان داده (خاص، داده کاوی)



چالش های مدیریت کلان داده

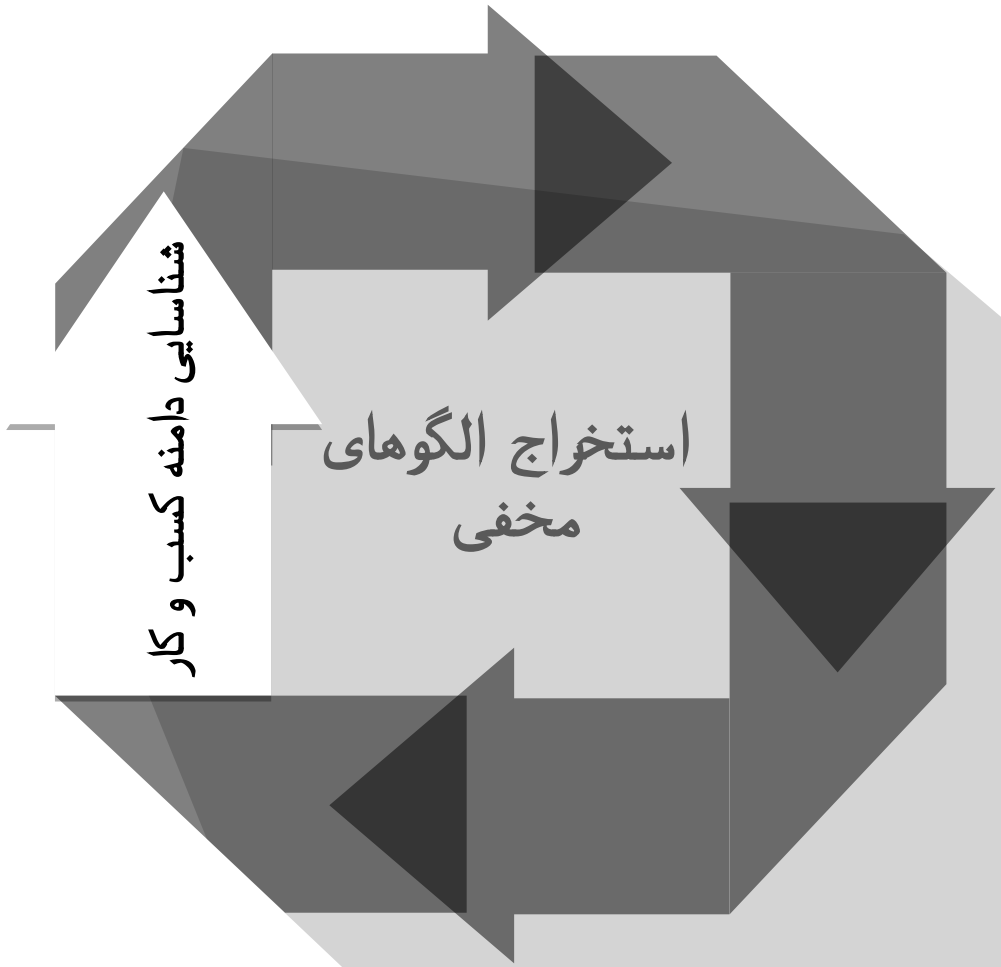


ارائه یک مدل پیشنهادی به منظور تحلیل کلان داده در سازمان



شناسایی دامنه کسب و کار

- انتخاب حوزه های مهم سازمان
- تشخیص اهداف به کار گیری یا کاربردهای مفید داده کاوی در سازمان و کسب و کار موردنظر



شناسایی منابع داده

- شناسایی ویژگی های منابع داده
  - فیلدها
  - متغیرها
- نمونه برداری حجم محدود داده از میان داده های سازمان (در یک بازه زمانی محدود)

آماده سازی داده

- پیش پردازش و آماده سازی داده
  - پاک سازی داده
  - تبدیل و کدینگ داده
  - استخراج ویژگی های مهم از داده
- انتخاب بهترین روش از میان الگوریتم های موجود
- یا
- توسعه یک الگوریتم جدید مطابق با داده های فعلی
- تنظیم پارامترهای روش انتخابی

تحلیل کلان داده

- انجام فرآیند اصلی تحلیل و ساخت مدل یا استخراج الگوهای مخفی از میان داده ها
- ارزیابی خودکار توسط معیارهای مشخص شده
- اجرای فرآیند بر روی داده های واقعی
- استفاده از دانش بدست آمده در کسب و کار سازمان
- تبدیل شدن دانش به حکمت
- اثربخشی هرچه بیشتر مدیران سازمان

پژوهش انجام شده در حوزه چالش های تحلیل عقاید در کلان داده ها در زبان فارسی

Classification of Sentimental Analysis Challenges in Persian Langauge

Mohammad Heydari
Department of Computer Science and Engineering
Shahid Behehsti University
Tehran, Iran
moh.heydari@mail.sbu.ac.ir

نتیجه گیری

در این پژوهش ضمن ارائه یک مدل مبسوط به منظور پیاده سازی تحلیل کلان داده در سازمان موردنظر،تکنیک های مهم تحلیل کلان داده به همراه چالش های پیش روی سازمان ها در مواجهه با این پدیده مهم بررسی شد.

مراجع

- James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers, Big data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute, 2012.
- Mark A. Beyer and Douglas Laney. “The Importance of 'Big Data': A Definition”. Gartner, 2012
- Bill Franks. “Taming the big data tidal wave”. Wiley, 2012
- David R. Hardoon and Galit Shmueli. “Getting started with business analytics – insightful decision making”. Talor & Francis Group.2013
- Foster Provost and Tom Fawcett. “Data science for business”. O’Relly, 2013
- Thomas H. Davenport and D.J. Patil . “Data Scientist: The Sexiest Job of the 21st Century”, Harvard Business Review, 2012
- Villars, R. L., Olofson, C. W., & Eastwood, M. (2011). Big data: What it is and why you should care. White Paper, IDC.

مراجع

- Combining big data analytics with business process using reengineering, 2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)
- Thomas H. Davenport, “ Analytics 3.0”, Harvard Business Review, December, 2013 Issue.
- EY, Building a better working world, “ Big Data Changing the way businesses compete and operate”, April 2014
- C.L. Philip Chen, C.-Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Inform. Sci.(2014)
- Muhammad Sahimi, Hossein Hamzehpour, Efficient computational strategies for solving global optimization problems, Comput. Sci. Eng. 12 (4)(2010) 74–83.
- MongoDB Integration with Hadoop, https://www.slideshare.net/spf13/mongodb-and-Hadoop, 2012
- Introduction to Data mining Presentation, Ehsan Asgarian.

Any Questior

Feel free to ask me face to face

:-)