数据挖掘和大数据分析



Outline |

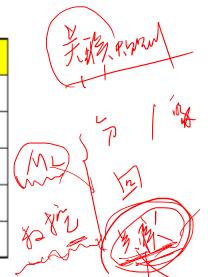
1 Review (Assignments)



Review (Assignments)

9.1 在下列事务数据集中

TID	项集
1	(面包,牛奶)
2	{面包,尿布,啤酒,鸡蛋}
3	{牛奶,尿布,啤酒,可乐}
4	{面包,牛奶,尿布,啤酒}
5	{面包,牛奶,尿布,可乐}



9.2 阅读微信群发布的"关联规则例子.pv",并根据交易单为(T1,

T2, T3, T4, T5, T6, T7, T8, T9), 每笔交易的货物清单为{{I1,

I2, I5},{I2, I4},{I2, I3},{I1, I2, I4},{I1, I3},{I2, I3},{I1, I3},{I1,

I2, I3, I5},{I1, I2, I3}}, 编写代码得到关联规则。

项集{啤酒,尿布,牛奶}的支持度为_____

如果将最小支持度定为3,则数据集中的频繁项集

{{面包}、{牛奶}、{尿布}、{啤酒}、{面包、牛奶}、{面包、尿布}

生奶,尿布》、{尿布、啤

酒}}

规则{牛奶,尿布}→{啤酒}的支持度为_2/5_____,置信度为

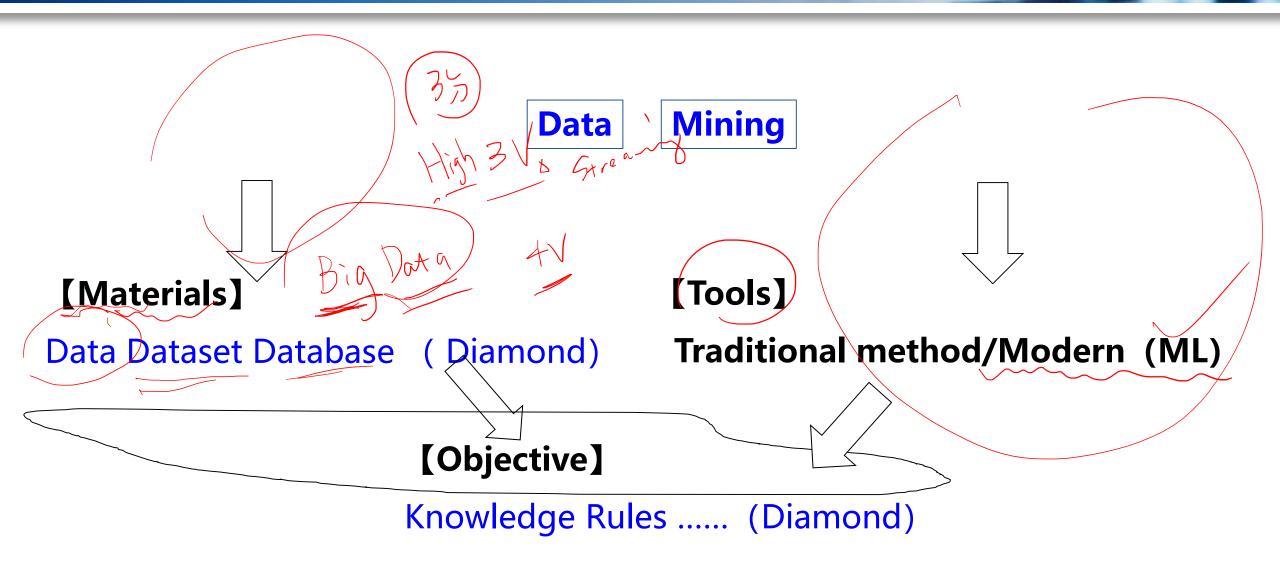








Review (Assignments)



DATA ANALYTICS:

DATA MINING AND BIG DATA



Big Data 1



[Case 1]

Taobao Nuggets Big Data Financial Market

Taobao launched a credit loan service that does not rely on guarantees or mortgages. Based on massive business data, it analyzes the business status and credibility of the company through big data analysis and makes decisions. This financial model has achieved great success!



余额宝手机端界面

[Case 2]

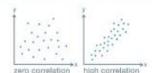
McKinsey: How to use big data to improve manufacturing

Identify initial patterns (using moving averages, distribution histograms, standard deviations, and clustering) to prioritize data collection and analysis Identify core determinants of process performance and form an initial hypothesis about root causes of yield drop and variability

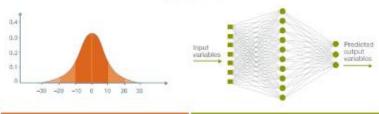
Using data visualizations

Using correlation analyses









Using significance testing

Using artificial neural networks

Fest initial hypothesis of root causes of whodel complex processes to quantify the impact of and optimal ranges for the interestination.

In the impact of and optimal ranges for the identified parameters.

3 main areas where big data helps improve manufacturing performance

better forecast product demand and adjust production capacity

(46%); understand factory performance across multiple

indicators (45%); provide faster service and support to

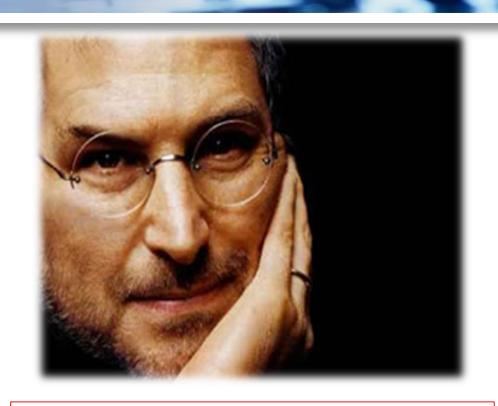
consumers (39%).



(Case 3)

Jobs' cancer treatment

Apple 's legendary president Steve Jobs took a different approach in the fight against cancer, becoming the first person in the world to sort all his DNA and tumor DNA. After Steve Jobs became ill, the treatment team also called big data to match Steve Jobs' personal situation with massive amounts of diagnosis and treatment data to get a better treatment plan, so his life was extended for several years.



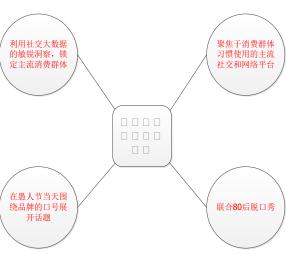
Jobs joked: "I am either the first person to defeat cancer in this way, or the last person to die of cancer in this way.

(Case 4)

Rely on big data to play April Fools' Day

This marketing activity on April Fool 's Day Quduoduo created more than 600 million page views and affected nearly 15 million unique users. The number of times the brand was mentioned increased by 270%. It can be said that this is a successful brand marketing campaign, a wide range of voices, so that a lot of Quduoduo's users attention has been greatly improved, the humorous brand genes are more deeply into the level of user awareness.

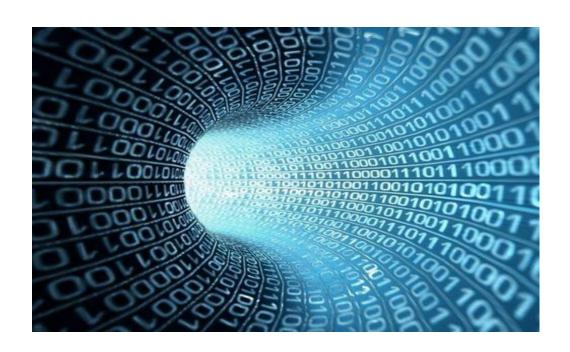




McKinsey, a world-renowned consulting firm, first proposed the era of "big data".

McKinsey said: Data has penetrated into every industry and business function in every field today, and has become an important production factor!







Data management

- Structure based: key value integration Your text in here
- Content based: spatial-temporal linkage
- Knowledge: ontology guided data fusion, Common sense integration, Domain knowledge

Data Fusion Knowledge development

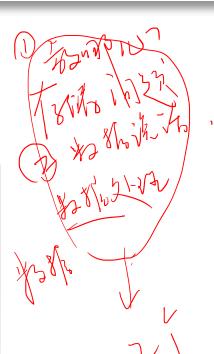
- Data mining: association, clustering, time-series and classification analytics
- Graph mining
- Ranking everything

Big data analysis

User interaction

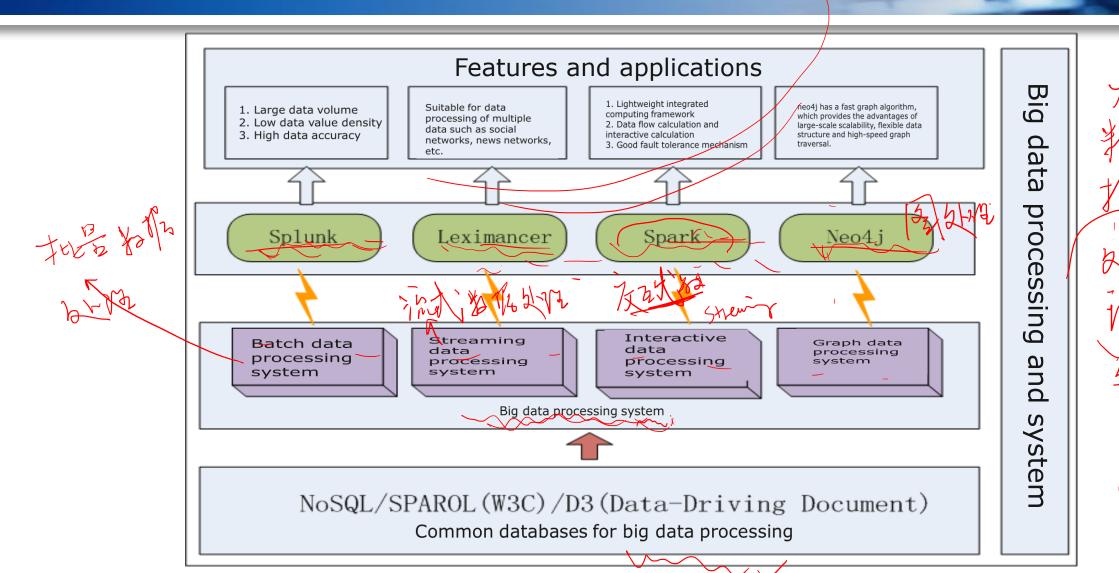
- Observe relationships between every data items
- Reveal the trends Create new concepts
- Show the categories and hierarchies

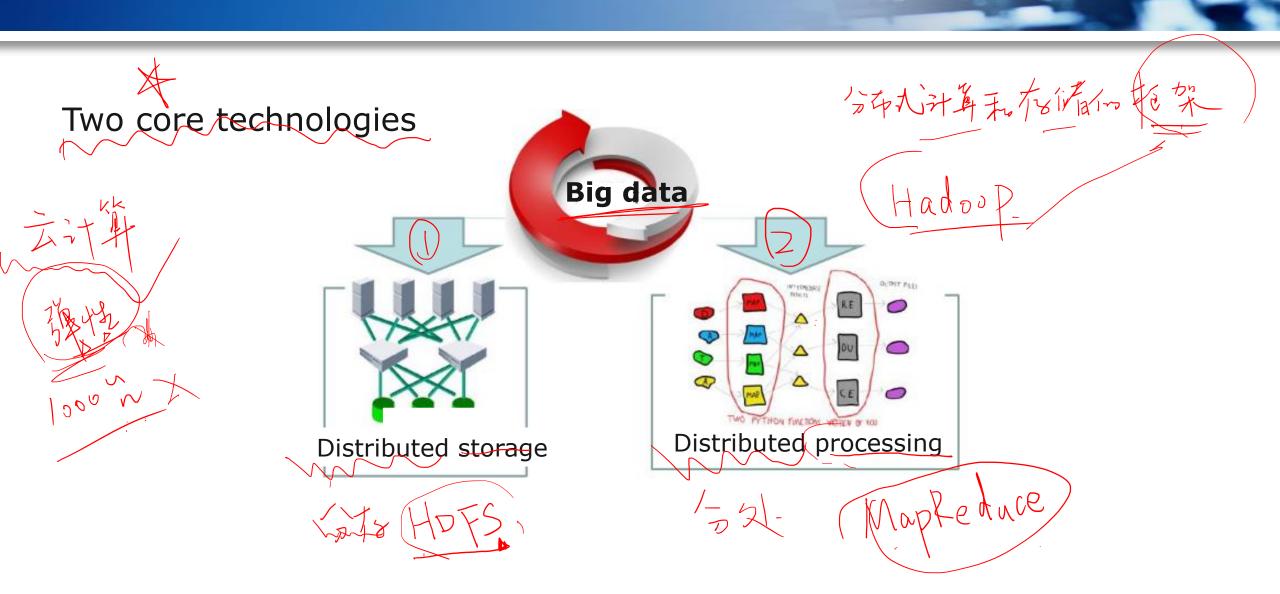
Visualiza tion

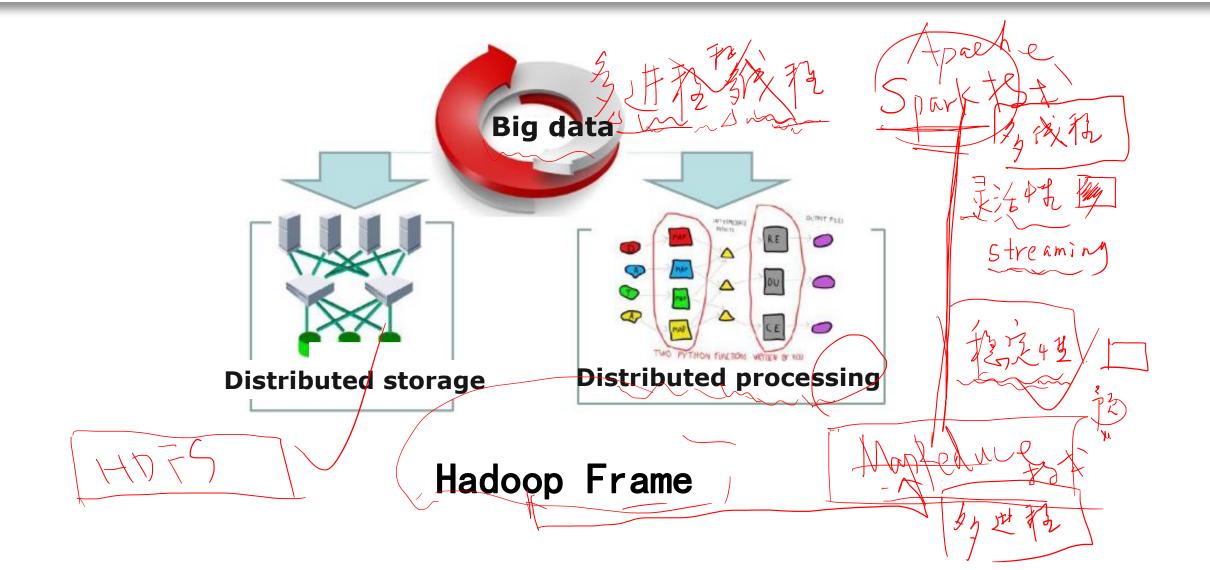


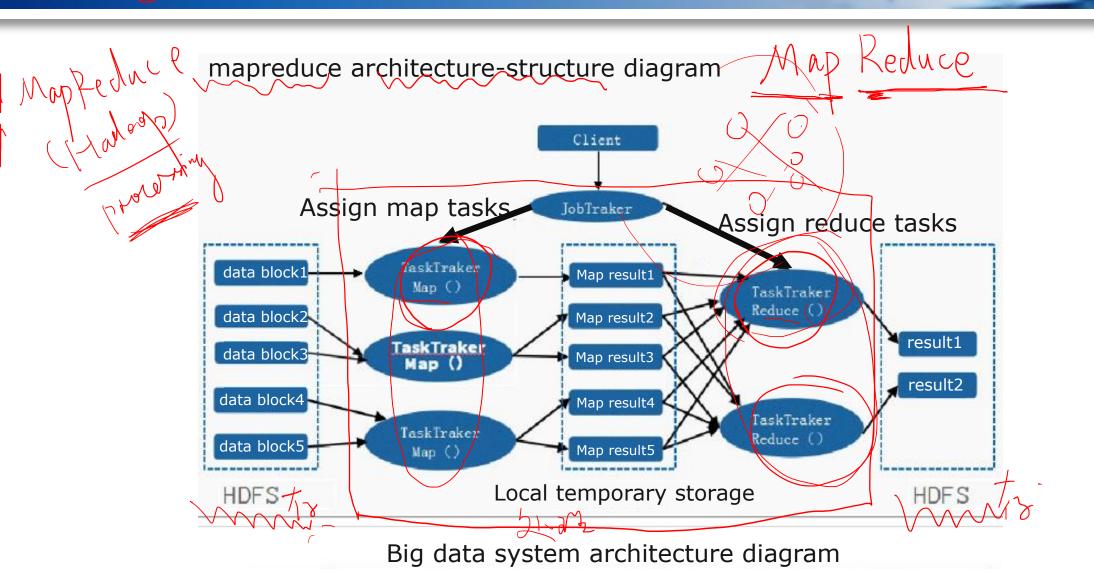
25 May 2

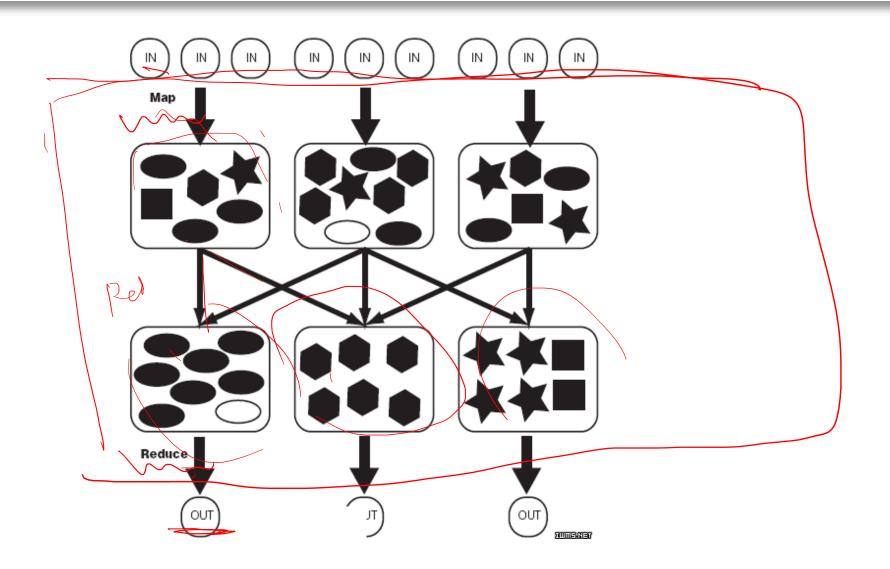
Big data processing theory

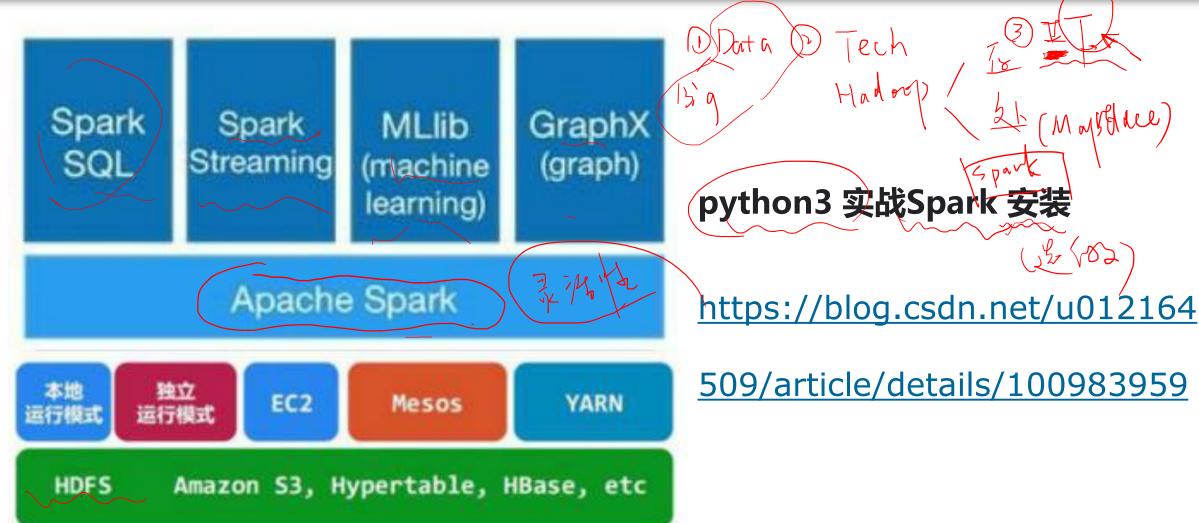


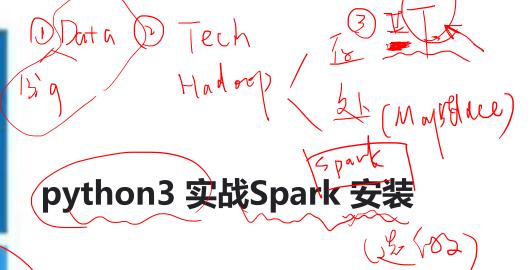












509/article/details/100983959





贵在坚持!