moa

Mining Big Data in Real Time

Albert Bifet

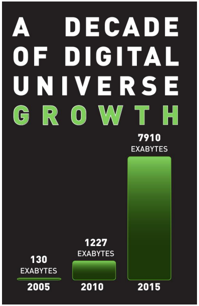
YAHOO! RESEARCH

Turing/SLAIS 2012 Conference

BIG DATA

Measure and React

Motivation



Source: IDC's Digital Universe Study (EMC), June 2011

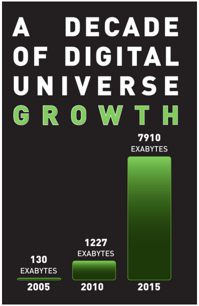
Data is growing

Motivation

Memory unit	Size	Binary size
kilobyte (kB/KB)	10^3	2^{10}
megabyte (MB)	10^6	2^{20}
gigabyte (GB)	10^9	2^{30}
terabyte (TB)	10^{12}	2^{40}
petabyte (PB)	10^{15}	2^{50}
exabyte (EB)	10^{18}	2^{60}
zettabyte (ZB)	10^{21}	2^{70}
yottabyte (YB)	10^{24}	2^{80}

Data is growing

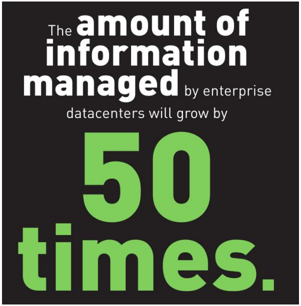
Motivation



Source: IDC's Digital Universe Study (EMC), June 2011

Data is growing

Motivation

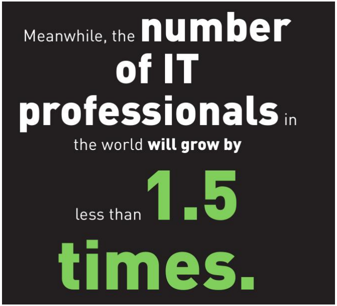


The amount of information managed by enterprise datacenters will grow by 50 times.

Source: IDC’s Digital Universe Study (EMC), June 2011

Data is growing

Motivation



Meanwhile, the **number of IT professionals** in the world **will grow by** less than **1.5 times.**

Source: IDC's Digital Universe Study (EMC), June 2011

Data is growing

Streaming Data

Big Data & Real Time

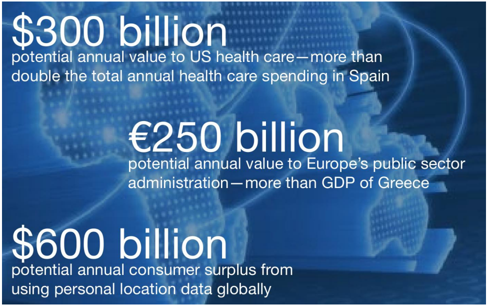
Big Data

Big data—a growing torrent
$600 to buy a disk drive that can store all of the world's music
5 billion mobile phones in use in 2010

McKinsey Global Institute (MGI) Report on Big Data, 2011.

Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.

Big Data



$300 billion
potential annual value to US health care—more than double the total annual health care spending in Spain

€250 billion
potential annual value to Europe's public sector administration—more than GDP of Greece

$600 billion
potential annual consumer surplus from using personal location data globally

McKinsey Global Institute (MGI) Report on Big Data, 2011.

Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.

BIG Data

- Volume
- Variety
- Velocity

3 Vs

Methodology

Sampling and distributed systems

Methodology

Paolo Boldi

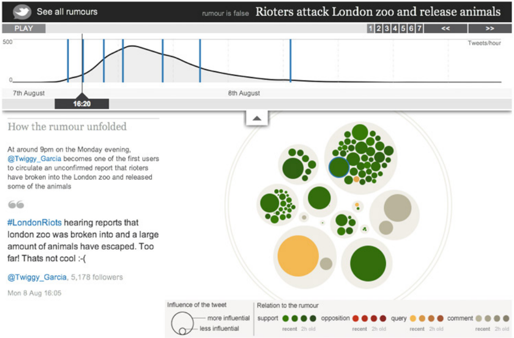
Facebook Four degrees of separation

Big Data does not need big machines,
it needs big **intelligence**

Real time analytics

We want to analyze what is happening **now**.

Real time analytics



Rioters attack London zoo and release animals

How the rumour unfolded

At around 9pm on the Monday evening, @Twiggy_Garcia becomes one of the first users to circulate an unconfirmed report that rioters have broken into the London zoo and released some of the animals

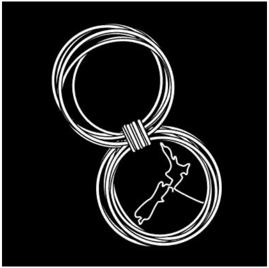
#LondonRiots hearing reports that london zoo was broken into and a large amount of animals have escaped. Too far! Thats not cool :-(

@Twiggy_Garcia, 5,178 followers

Mon 8 Aug 16:05

We want to analyze what is happening **now**.

Time and Memory



Number 8 Wire Mentality

Time and memory are the resource dimensions of the process.

Time and Memory

Time and memory are the resource dimensions of the process.

Algorithms

Classification, Regression, Clustering, Frequent Pattern Mining.

Applications

- sensor data: industry, cities
- telecomm data
- social networks: twitter, facebook, yahoo
- marketing: sales business

Data may come from: humans, sensors, or machines.

New applications: social networks

Twitter: A Massive Data Stream

- Micro-blogging service
- Built to discover what is happening at any moment in time, anywhere in the world.
- 3 billion requests a day via its API.

`MOA-TweetReader`: a real-time system to

- read tweets in real time
- detect changes
- find the terms whose frequency changed

Sentiment Analysis on Twitter

Sentiment analysis

Classifying messages into two categories depending on whether they convey positive or negative feelings

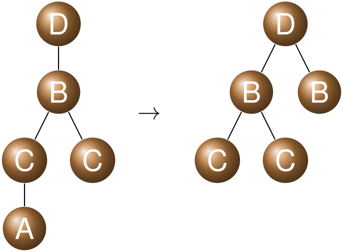
Emoticons are visual cues associated with emotional states, which can be used to define class labels for sentiment classification

Positive Emoticons	Negative Emoticons
:)	:(
:-)	:-(
:)	: (
:D	
=)	

Table : List of positive and negative emoticons.

New problem: structured classification

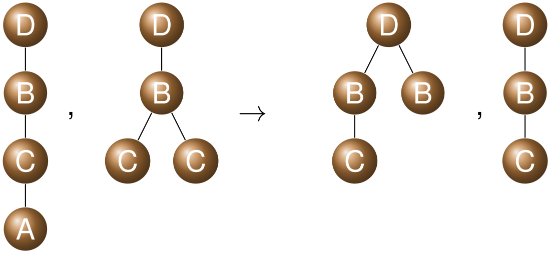
New methods for structured classification



- sequences, trees, graphs

New problem: structured classification

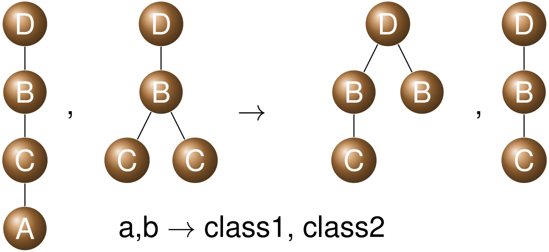
New methods for structured classification



- sequences, trees, graphs
- frequent pattern mining techniques

New problem: structured classification

New methods for structured classification



a,b → class1, class2

- sequences, trees, graphs
- frequent pattern mining techniques
- multi-label data mining
    - Example: Lord of the Rings → Action, Adventure, Fantasy

New Techniques: Distributed Systems

Google

YAHOO!

facebook

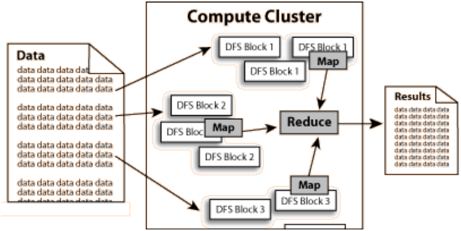
twitter

Hadoop, S4 and Storm

Hadoop

Hadoop

Hadoop

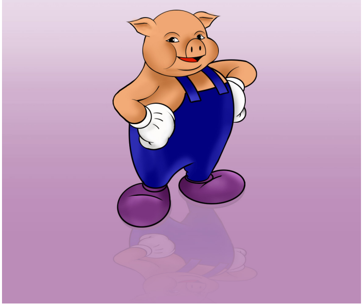


Hadoop architecture

Apache Mahout

Mahout: open source framework

Pig



Pig: Similar to SQL

Pig

- `A = LOAD ’data’ USING PigStorage() AS (f1:int, f2:int, f3:int);`
- `B = GROUP A BY f1;`
- `C = FOREACH B GENERATE COUNT ($0);`
- `DUMP C;`

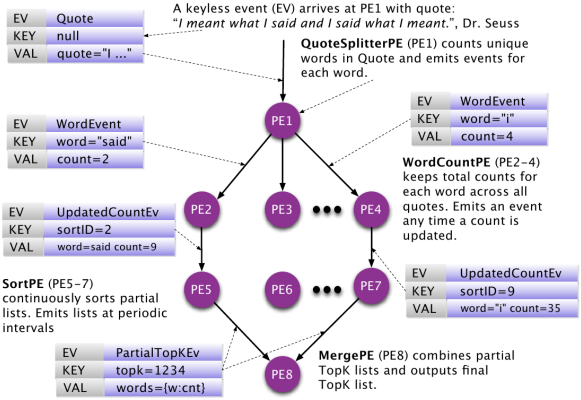
Pig: Similar to SQL

Apache S4

S4 distributed stream computing platform

Apache S4

Apache S4



A keyless event (EV) arrives at PE1 with quote:
"*I meant what I said and I said what I meant.*", Dr. Seuss

EV	Quote
KEY	null
VAL	quote="I ..."

QuoteSplitterPE (PE1) counts unique words in Quote and emits events for each word.

EV	WordEvent
KEY	word="said"
VAL	count=2

EV	WordEvent
KEY	word="i"
VAL	count=4

WordCountPE (PE2–4) keeps total counts for each word across all quotes. Emits an event any time a count is updated.

EV	UpdatedCountEv
KEY	sortID=2
VAL	word=said count=9

EV	UpdatedCountEv
KEY	sortID=9
VAL	word="i" count=35

SortPE (PE5–7) continuously sorts partial lists. Emits lists at periodic intervals

EV	PartialTopKEv
KEY	topk=1234
VAL	words={w:cnt}

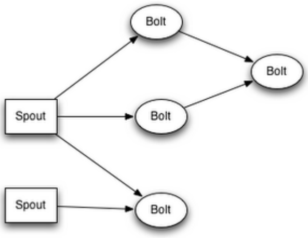
MergePE (PE8) combines partial TopK lists and outputs final TopK list.

PE ID	PE Name	Key Tuple
PE1	QuoteSplitterPE	null
PE2	WordCountPE	word="said"
PE4	WordCountPE	word="i"
PE5	SortPE	sortID=2
PE7	SortPE	sortID=9
PE8	MergePE	topK=1234

Storm

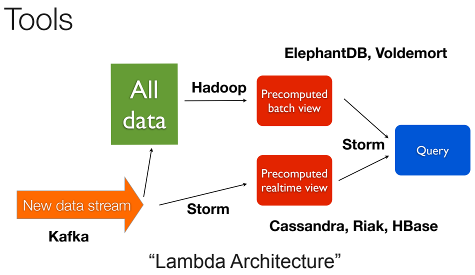
Storm from Twitter

Storm



Stream, Spout, Bolt, Topology

Storm



Tools

ElephantDB, Voldemort

All data

Hadoop

Precomputed batch view

Storm

Query

Precomputed realtime view

New data stream

Storm

Cassandra, Riak, HBase

Kafka

“Lambda Architecture”

Runaway complexity in Big Data
Nathan Marz, 2012

Data Streams

Big Data & Real Time

Data Streams

Thanks!