



# TECHNIQUES OFFERED BY BIG DATA AVAILABLE FOR USE IN THE BANKING SECTOR

RADOSŁAW KITA

Partner



Sponsor



Media Partners



## Before we embark on the journey

Why me?

- For over 12 years I was responsible for reporting systems in Onet.pl. I was responsible for the choice of data analysis tools. In this the transition from commercial to open source solutions.
- I took an active part in one of the most interesting (and somewhat controversial) Big Data project at Alior Bank.
- I worked on the implementation of the recommendation system in TVN content.
- Expert Data Scientist at Allegro.
- Actively dealing with the analysis of large data sets and analytical tools selection from the beginning of this century 😊

## **Before we embark on the journey**

What I do not want to talk

I do not want to talk about what I read on the Internet just about something that I did in practice.

## Before we embark on the journey

Big Data is like...

“Teen sex:

everyone talks about it,  
nobody really knows how to do it,  
everyone thinks everyone else is doing it,  
so everyone claims they are doing it...”

*Dan Ariely*

If I were to be malicious:

Children of a certain age do not believe that their parents have sex,  
despite fairly obvious proof.

## **Before we embark on the journey**

When you are responsible for the choice of technology

“I do not really have anything against so that our competitors has issued 3,000,000 for technology that they will not help.”

## **Before we embark on the journey**

When you are responsible for the choice of technology

“The truth is out there”

But there is a world where no one has ever lied.

# Before we embark on the journey

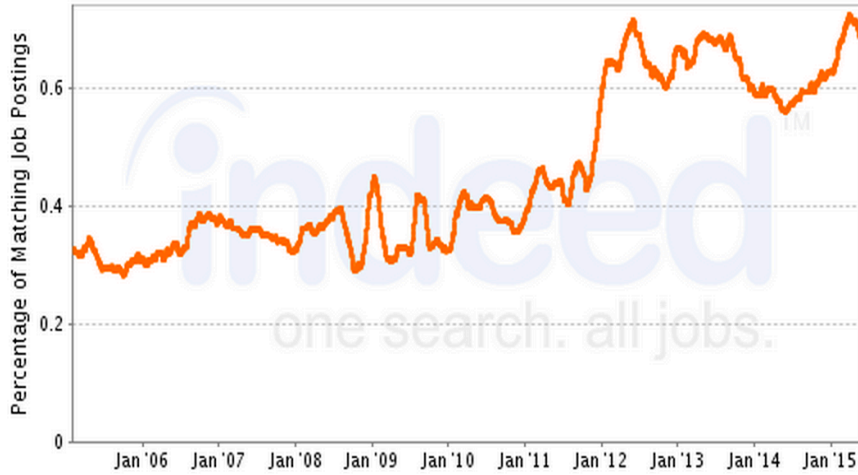
World where no one has ever lied



Scale: **Absolute** - [Relative](#)

Job Trends from Indeed.com

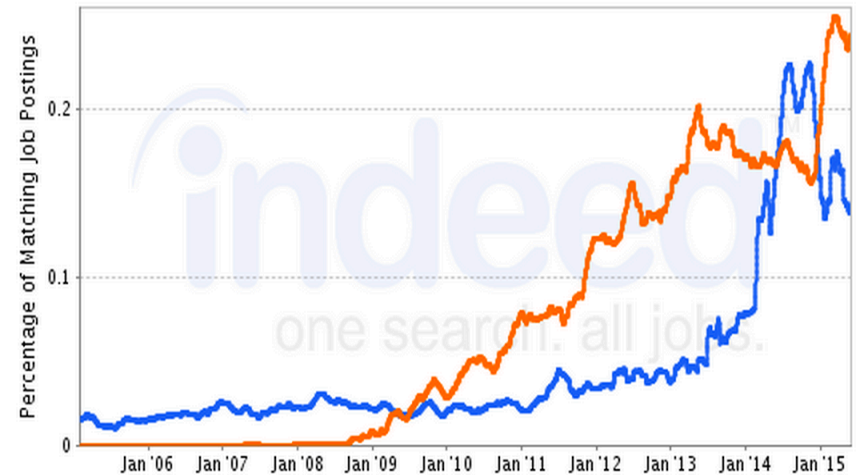
— big data



Scale: **Absolute** - [Relative](#)

Job Trends from Indeed.com

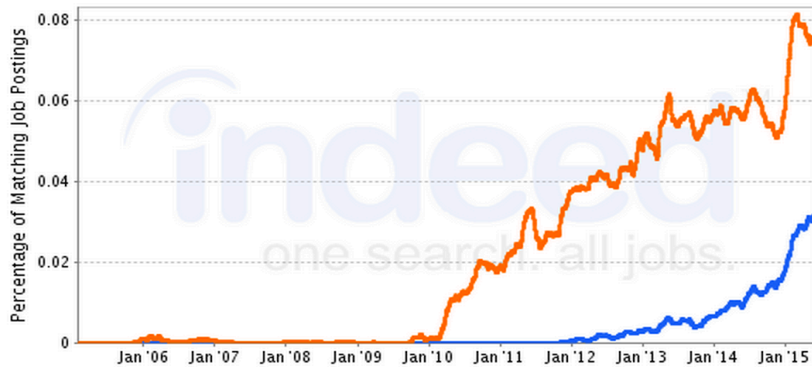
— Hadoop — Spark



Scale: **Absolute** - [Relative](#)

Job Trends from Indeed.com

— cassandra — elasticsearch



▶

▶

Tc

## All that jazz is about

in my humble opinion

The ability to create a coherent picture of customer needs.

- On the basis of a number of different data sources.
- Suitable for practical use.
- Within a reasonable period of time.
- At a reasonable cost.

Why?

“Data driven company”



## All that jazz is about

more detailed

We want to merge data about the behavior of our customers:

- Internal transaction systems
- What are they doing on our website
- What they write in social media
- What kind of content are looking for on the Internet
- In what case would call for call centers
- How they use a mobile phones...

Definition Big Data (Gartner, 2001): Volume, Velocity, Variety  
+ 2012: Variability and Complexity

## Translating technical data

We have:

Client id	day	Target account	amount	Client id (cookie)	day	url	useragent	IP
1	2009-01-01	0812345345678990	315 zł				Mozilla/5.0 (Windows NT 5.1; rv:26.0) Gecko/20100101 Firefox/26.0	217.67.211.123
1	2009-01-01	0967899045543215	124 zł	23456781	2009-01-01	www.wp.pl		
...	...	...	...	...	...	...	...	...

Event date	event
2015-10-11 12:01:14	{ <pre>"PageView": {"value": "start", "popup": {   "menuitem": [     {"value": "Open", "onclick": "OpenDoc()"},     {"value": "Close", "onclick": "CloseDoc()"}   ] }</pre>
...	...

we need:

Client id	Value od the gas bill M-6	Value od the gas bill M-5	Gas bill # days before/ after M-6	Airplane tickets M-6	Gasoline M-5	Food shoping high-end shops M-5	Time spend on business web pages M-6	Expenditure on education of the child M-5	Class of a held phone	...
1	315 zł	320 zł	10	0 zł	300 zł	20 zł	3:40	140 zł	High	...
2	420 zł	350 zł	5	800 zł	500 zł	550 zł	0:15	0 zł	Medium	...
...	...	...	...	...	...	...	...	...	...	...

## Technical issue

First and last

- We need to merge data coming from our web pages with data from our transactional system
- We do not have an united user id
- We can merge user id's base on IP address and userAgent parameter coexistence close together in time

Computing power!!!!

## Big data

### Strong points

can collect and process vast amounts of data at a reasonable cost



<https://hadoop.apache.org/>



<http://tez.apache.org/>



<http://mesos.apache.org/>

## Big data

### Strong points

realtime computation system - a million tuples processed per second per node



<http://spark.apache.org/>



<https://storm.apache.org/>



<https://flink.apache.org/>



<https://www.elastic.co/>

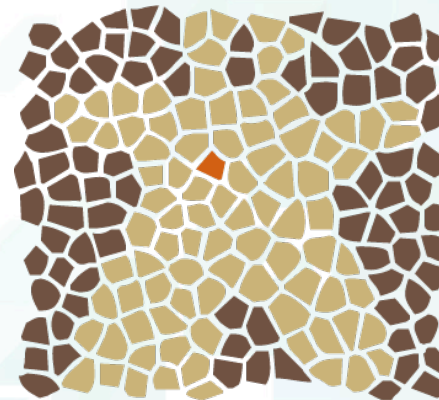
# Big data

Strong points

Graph databases



<http://neo4j.com/>



A P A C H E  
G I R A P H

<http://giraph.apache.org/>

## Big data

Strong points

Linear scalability



# ***Cassandra***

<http://cassandra.apache.org/>

# Big data

Strong points

Schema-Free



<https://drill.apache.org/>



**mongoDB**<sup>®</sup>  
FOR **GIANT** IDEAS

<https://www.mongodb.org/>



# Big data

## Weak points

### Data Governance

- Just on the beginning of this road

# Big data

## Weak points

## Security

- Definitely backend (not frontend) systems

# Big data

## Weak points

surprising gaps

table a

JOIN table b

ON a.index=b.index

table a

JOIN table b

ON a.day <= b.day

# Big data

## Weak points

we are living in a clusters world

# Big data

## Weak points



classical systems  
well organized cities



Big Data systems  
markets

# Big data

## Weak points



<https://hive.apache.org/>



<https://zookeeper.apache.org/>



<http://kafka.apache.org/>



<https://pig.apache.org/>



<http://sqoop.apache.org/>



<http://falcon.apache.org/>



<http://atlas.incubator.apache.org/>

# Big data vs small data

possibilities of cooperation

- Big Data: efficient quality data control.
- Big Data: the ability to integrate data from different sources.
- Big Data: fast access to data from any long period of time and for any size group of customers (data lakes concept).
- Advanced techniques of text analysis.
- Analysis of the data graph.
- Cheap and efficient power.
- And all this to change Big Data in Small Data.



## The Third International Workshop

**Thank you!**  
**Advanced Analytics and Data Science**  
**[radoslaw.kita@allegrogroup.com](mailto:radoslaw.kita@allegrogroup.com)**

Partner



Sponsor



Media Partners

