Big Data Trends 2017

Arató Bence

BI Consulting

arato@biconsulting.hu

Introduction

Arató Bence

Consulting and Advisory
BI/DW/Big Data strategy, Architecture planning, vendor and tool selection. Also provides QA and on-the-job mentoring services.

Publications
Editor of the BI.hu portal and the BI Yearbook series.

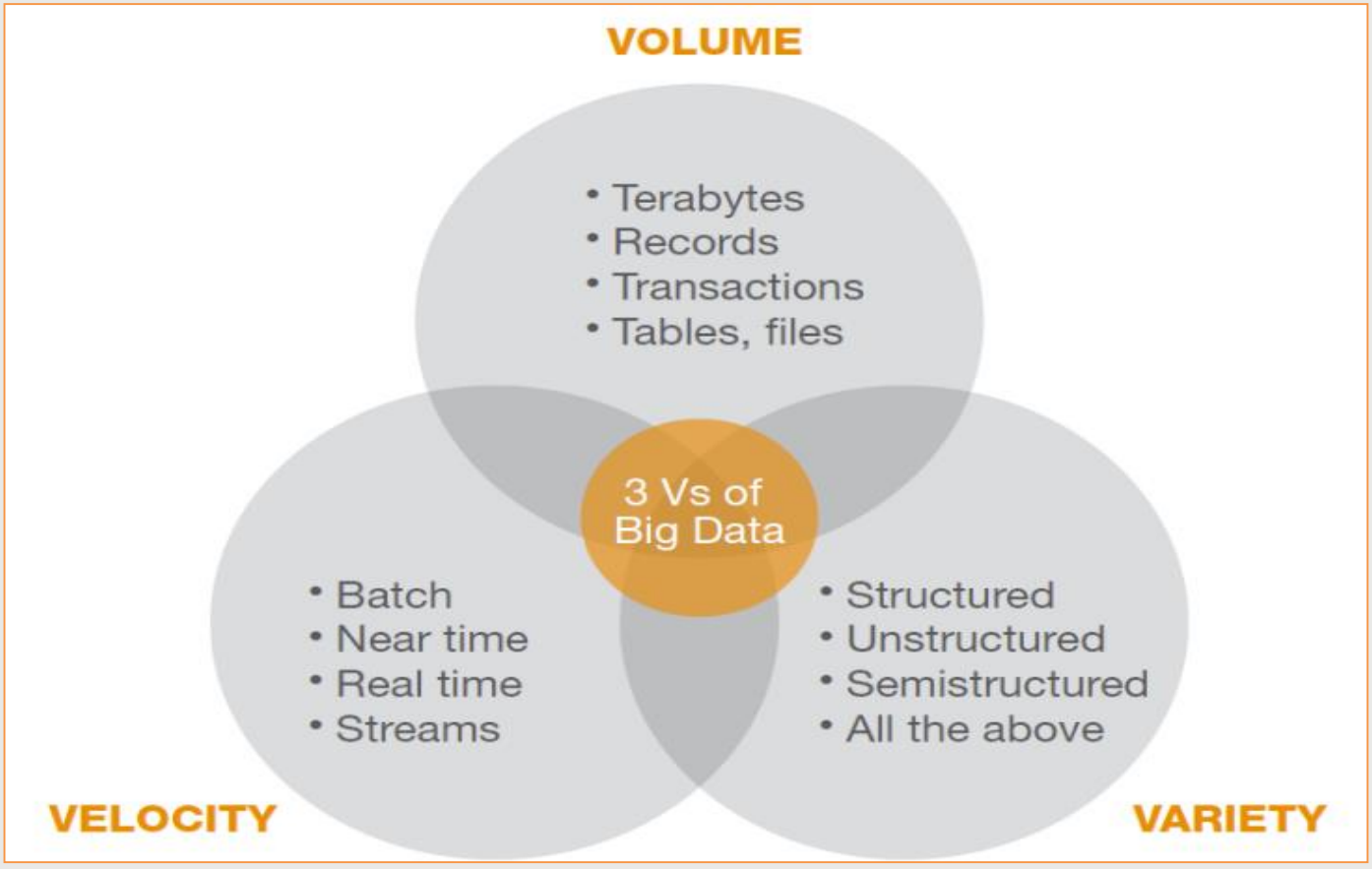
Research
Leader of the Hungary-focused BI-TREK, DW-TREK and NOSQL-TREK surveys.

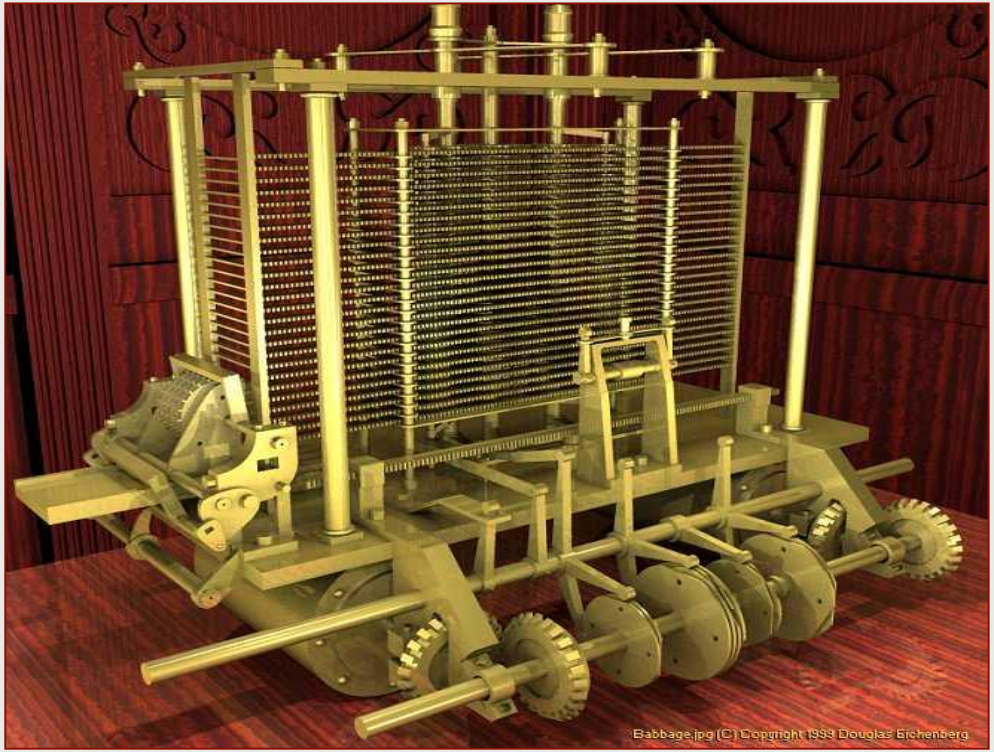
Teaching
Teaching several courses at BI Akadémia from data visualization to Big Data.

Conferences and events
Head organizer of the Budapest Data Forum, Budapest BI Forum and Budapest NOSQL Forum conferences and several data-related meetups.

Big data and the Vs

Big data and the Vs





Big Data in 2016

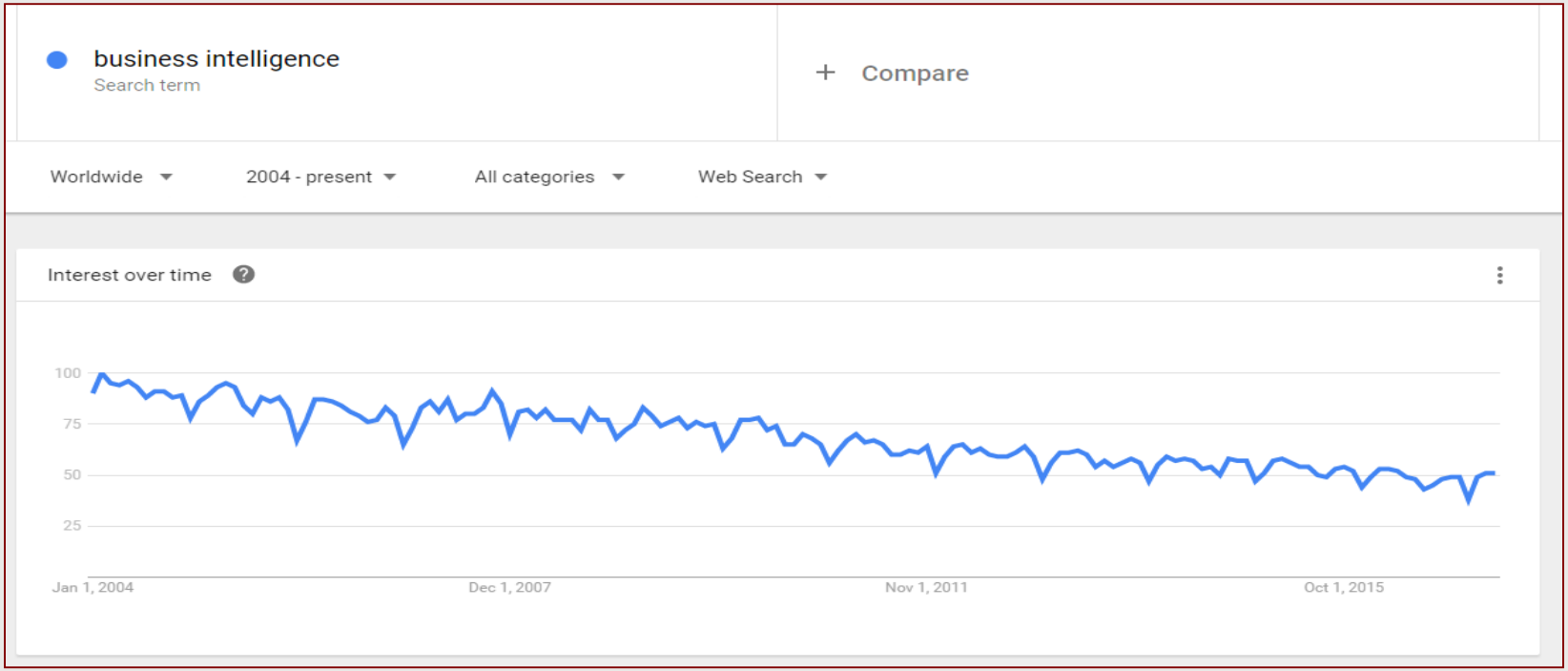
Big Data – hype is over?

„The biggest big data event of 2016 was people ceasing to talk about big data. Big data now 'just is'. „

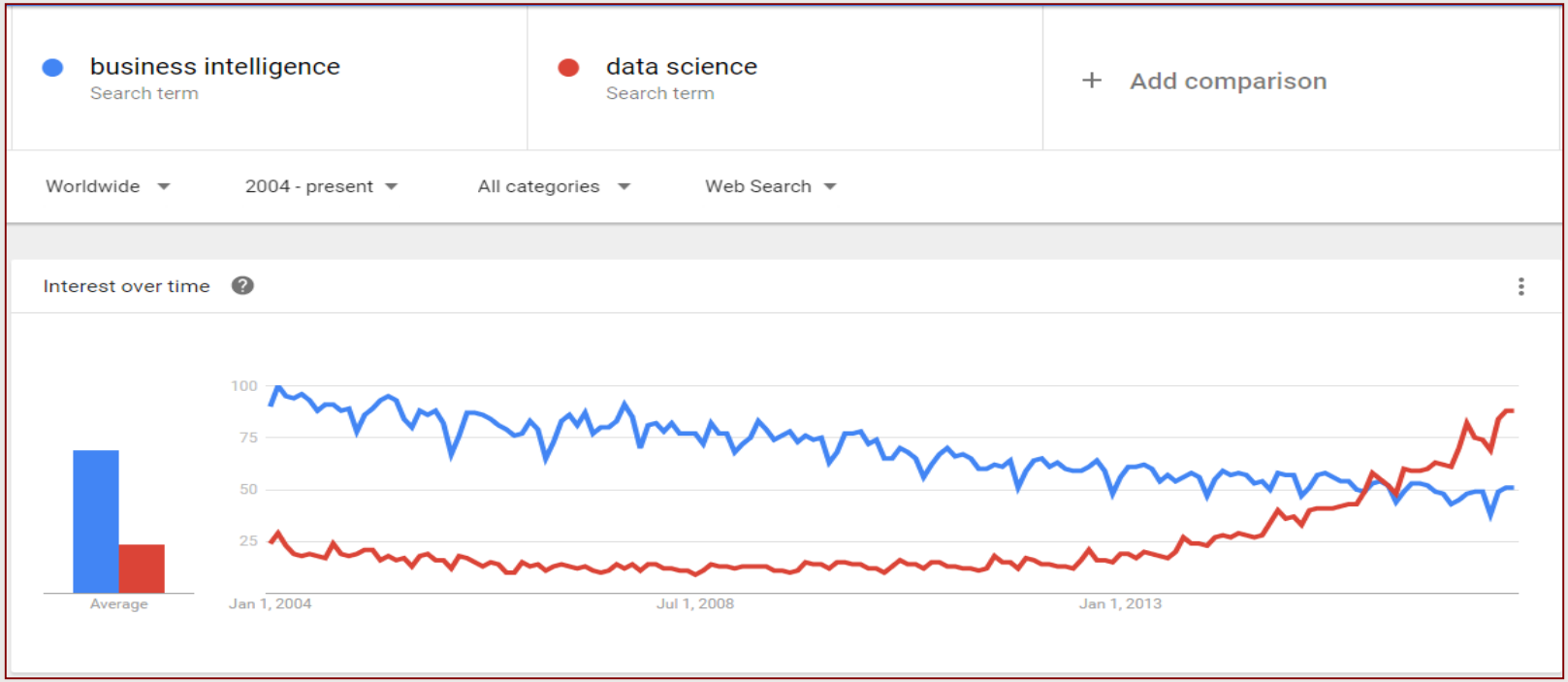
„2016, felt like Big Data was losing the buzz as compared to a few years ago”

2016 was an exciting year for big data, as finally, Big data is no longer a hype or a buzzword.

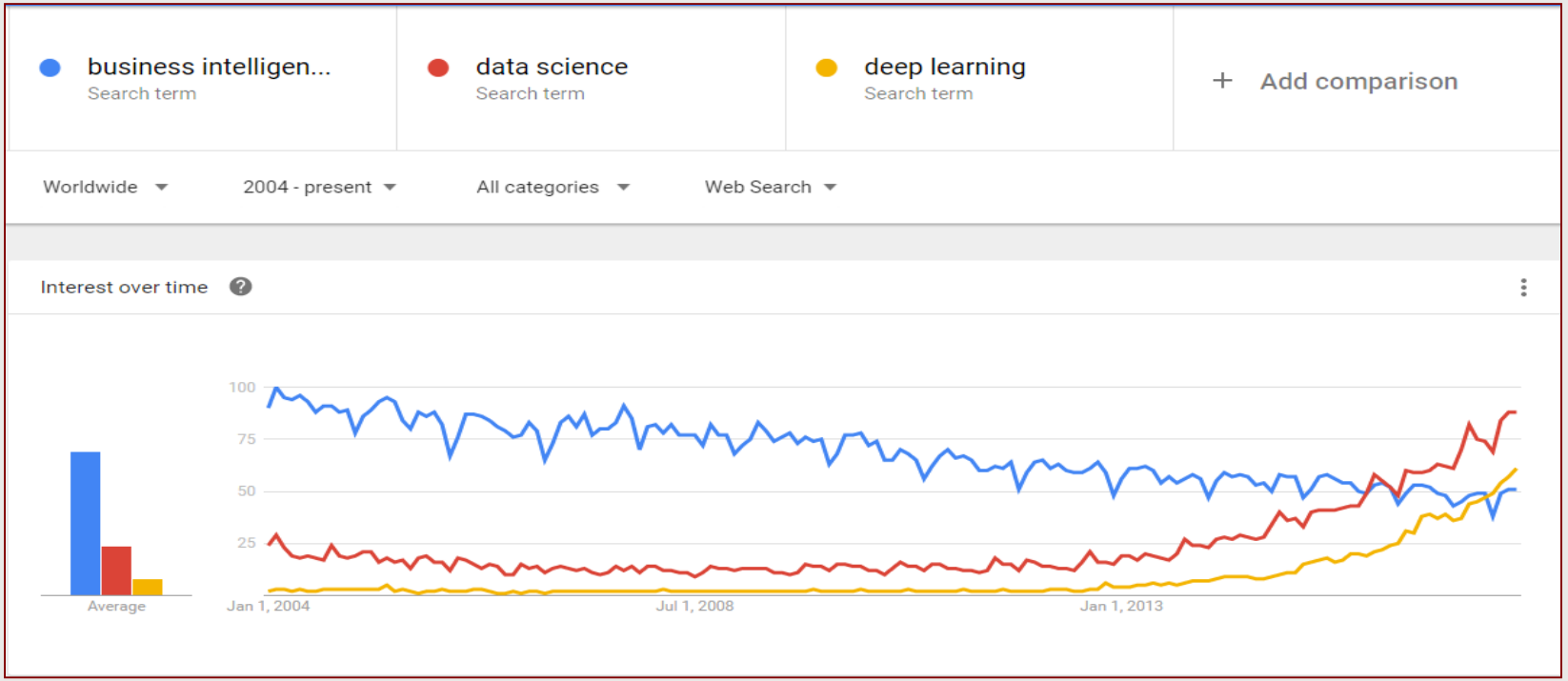
www.kdnuggets.com/2016/12/big-data-main-developments-2016-key-trends-2017.html



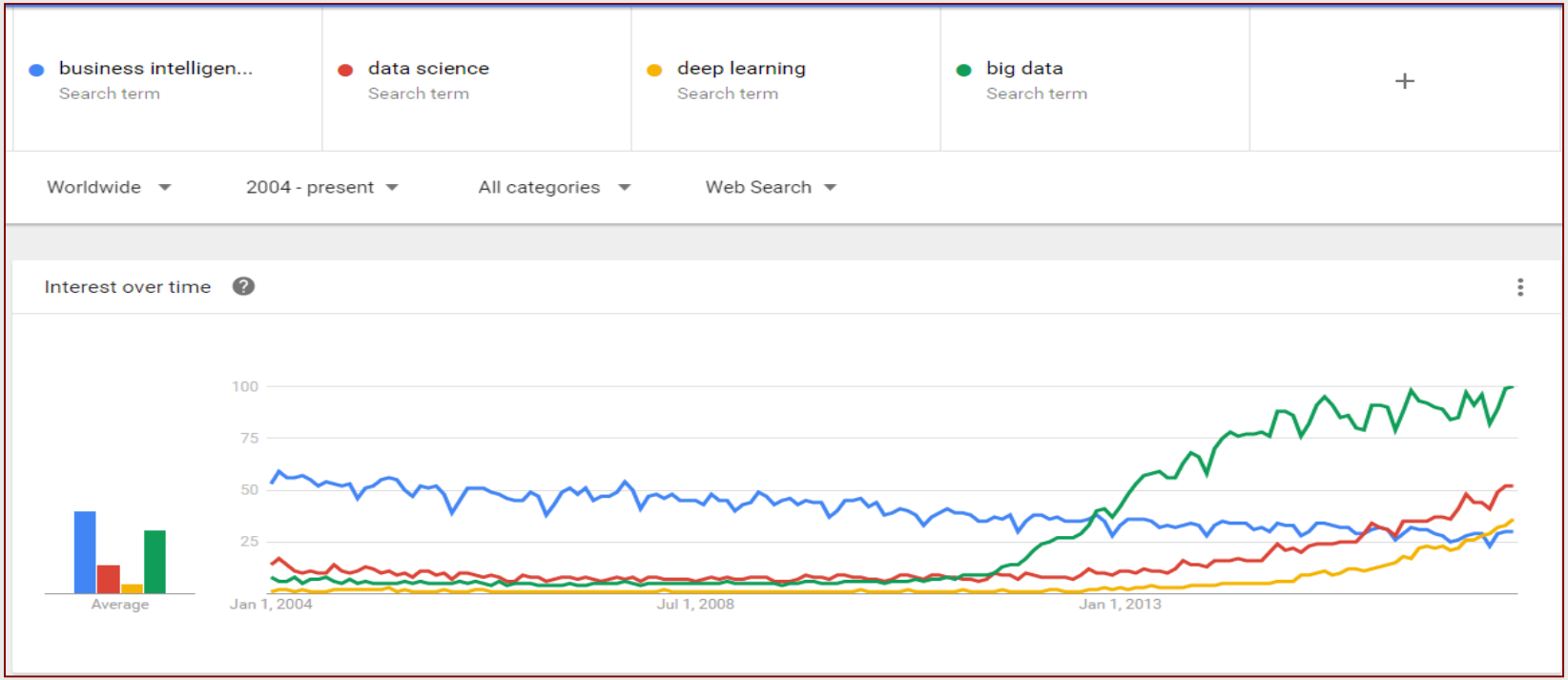
Google Trends



Google Trends



Google Trends



Google Trends

Big Data Market

The Big Data technology and services market will grow at a 27% compound annual growth rate to $32.4 billion through 2017 - or about six times the growth rate of the overall ICT market

IDC Worldwide Big Data Technology and Services 2013-2017 Forecast, Dec 2013

Big Data Market

IDC
Analyze the Future

Search

Homepage | Research | Consulting | For CIOs | Events | Analysts | About IDC | MyIDC

Spending on the Central and Eastern Europe Big Data and Analytics Market Will Reach $4.4 Billion in 2020, According to IDC

Prague, March 10, 2017

Your location: Home » About IDC » Press Center » Spending on the Central and Eastern Europe Big Data and Analytics Market Will Reach $4.4 Billion in 2020, According to IDC

PRAGUE, March 10, 2017 - According to the latest *Worldwide Semiannual Big Data and Analytics Spending Guide* from International Data Corporation (IDC), big data and analytics spending in Central and Eastern Europe (CEE) will exceed $3 billion in 2017 and reach $4.4 billion in 2020, achieving a 12.1% compound annual growth rate (CAGR) over a five-year-forecast period.

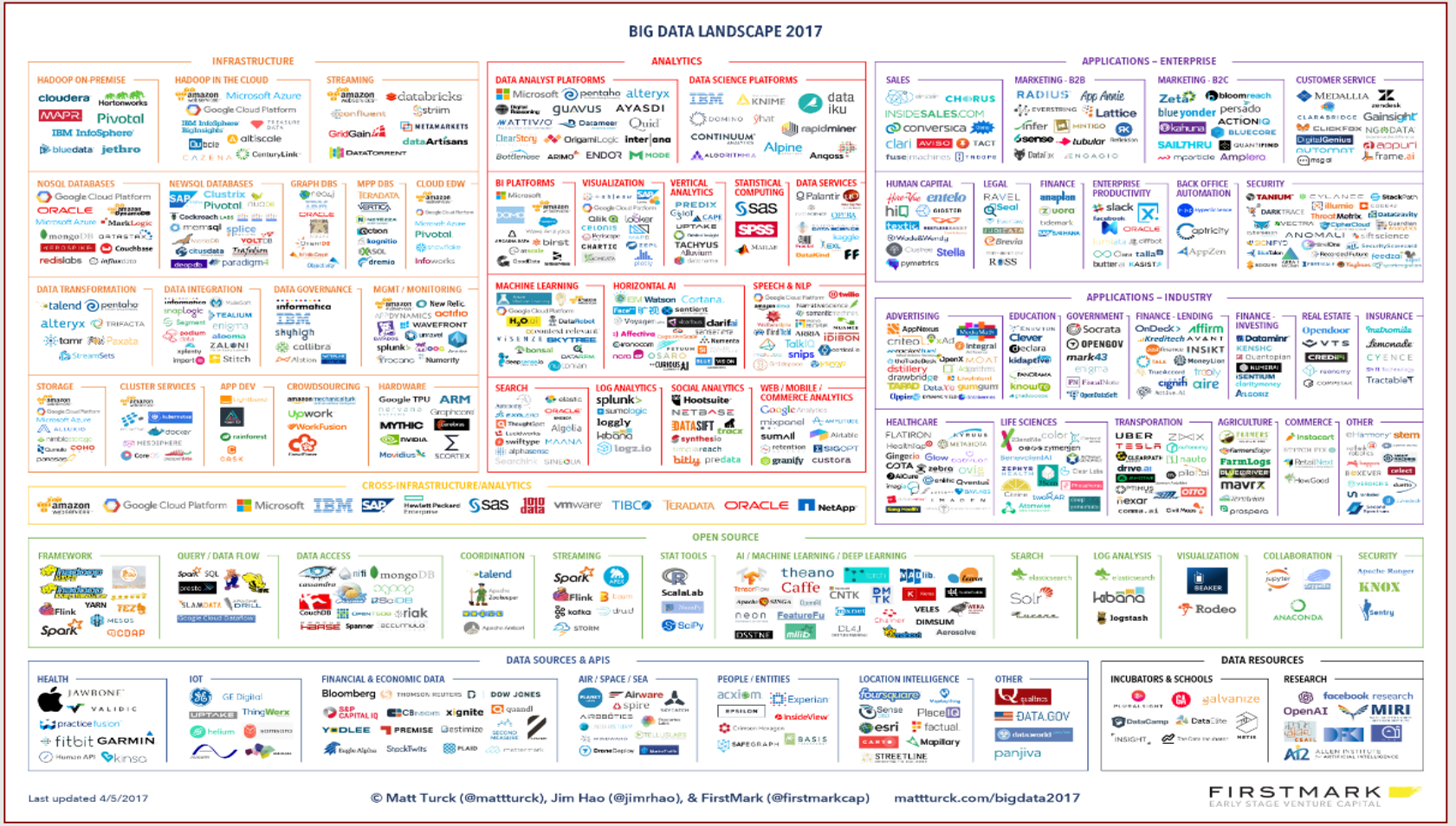
Business and IT services will contribute over 58% to the overall market spending in CEE, with an increasing market share over the forecast period, reaching 63% in 2020. Hardware investments will be the fastest growing submarket, with a 16.4% CAGR, as enterprises invest mainly into storage. In terms of software, cognitive software platforms, content analytics, and search systems will be the fastest growing groups, with CAGRs all exceeding 18%. Investments into end-user query, reporting tools, relational data warehouse management

- Subscribe to IDC Press Releases
- Quotation Guidelines

Print

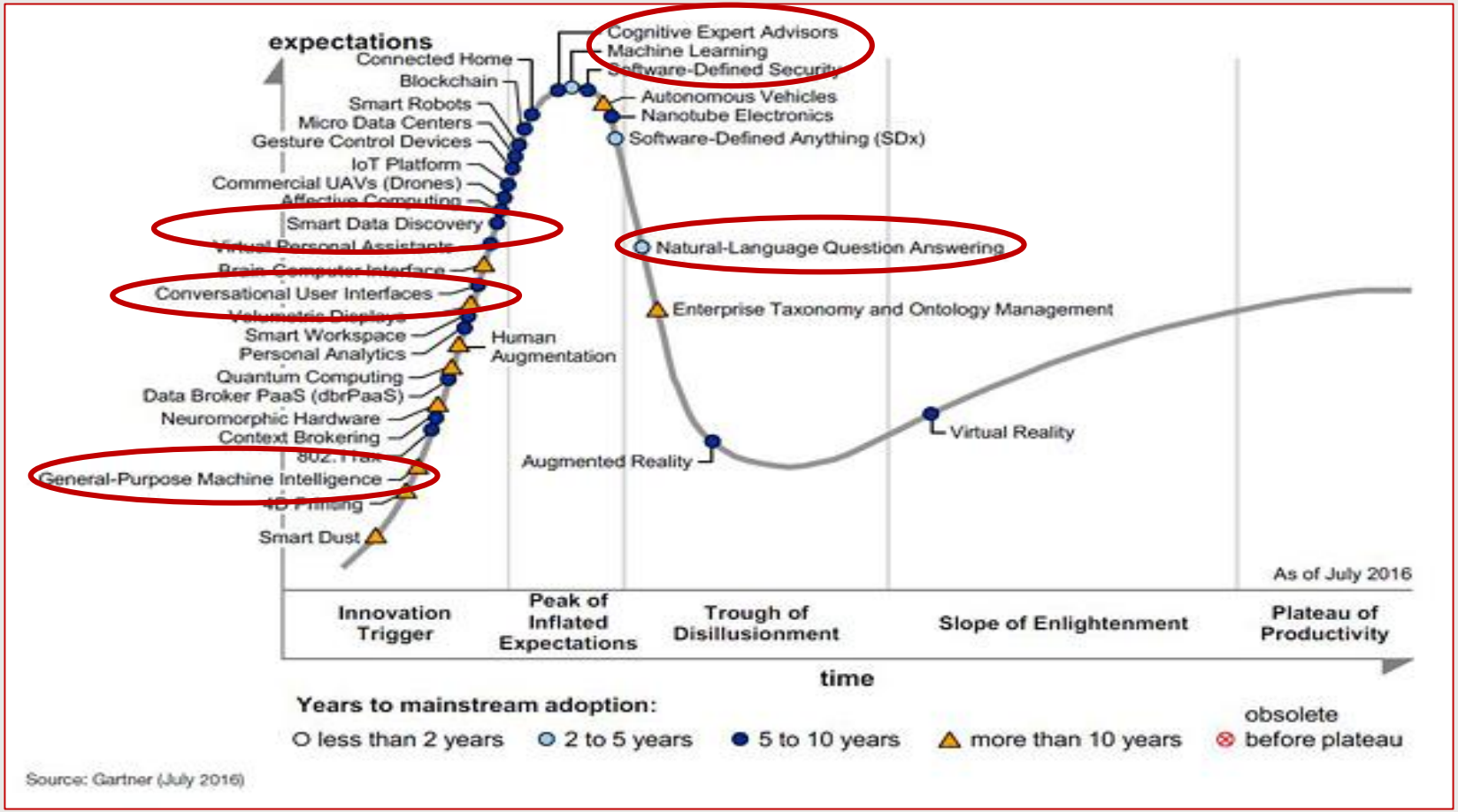
Contact

Big Data Landscape in 2017



mattturck.com/bigdata2017

Hype Cycle



Gartner Hype Cycle for Emerging Technologies, 2016

Hadoop ecosystem

Hadoop ecosystem

„Hadoop declined more rapidly in 2016 from the big-data landscape than I expected. MapReduce, HBase, and even HDFS are less relevant to data scientists than ever.”

www.kdnuggets.com/2016/12/big-data-main-developments-2016-key-trends-2017.html

Hadoop ecosystem

Hadoop is dead, long live Hadoop!

by **Svetlana Sicular** | September 29, 2016 | Comments Off on Hadoop is dead, long live Hadoop!

This year, we are celebrating the 10th anniversary of Hadoop. In 2006, we lived in the dark ages of data: Several databases were old news and a single Gartner analyst was covering them all. Today, a whole team of Gartner database analysts could barely handle all the new options.

In summer 2009, baby-elephant Hadoop made its first steps: a Hadoop distribution was introduced by an unknown startup Cloudera. HDFS and MapReduce became top level Apache projects. Ominously, this very summer of 2009, a CRM icon Tom Siebel was trampled by a charging elephant during an African safari. We didn't know back then that the Era of CRM came to its end, and the Era of Data began.

Hadoop revealed to the world the power of data. It evangelized the idea of distributed computing and massively-parallel processing. It brought compute to the data, rather than brining data to compute. It was the center of a quickly growing open source community and a vibrant ecosystem.

blogs.gartner.com/svetlana-sicular/hadoop-is-dead-long-live-hadoop

Hadoop ecosystem

Hadoop Is Dead, Long Live Apache (And Open Source)

During the 2 days at Hadoop Summit, I heard about the following big data related Apache projects: **Flink, Drill, Zeppelin, Eagle, Helium, Falcon, Atlas, Kylin, Phoenix, Ranger, Flume, Airavata, Beam** - with some of them being hailed as "the next big thing". While some were designed to overcome some of Hadoop's limitations (batch only, mapreduce bottleneck, security…), others are addressing new areas or challenges. While I did not always understand the exact value proposition of these products, it seems clear that a new wave of innovative frameworks to efficiently work data at scale is coming from the Apache foundation. This raises serious questions for companies that are having trouble seeing clearly through the big data technologies ecosystem and keeping up with rapidly evolving products and standards.

blog.dataiku.com/2016/07/19/trends-observed-from-q2-q3-2016-european-big-data-events

Photo: iStocks

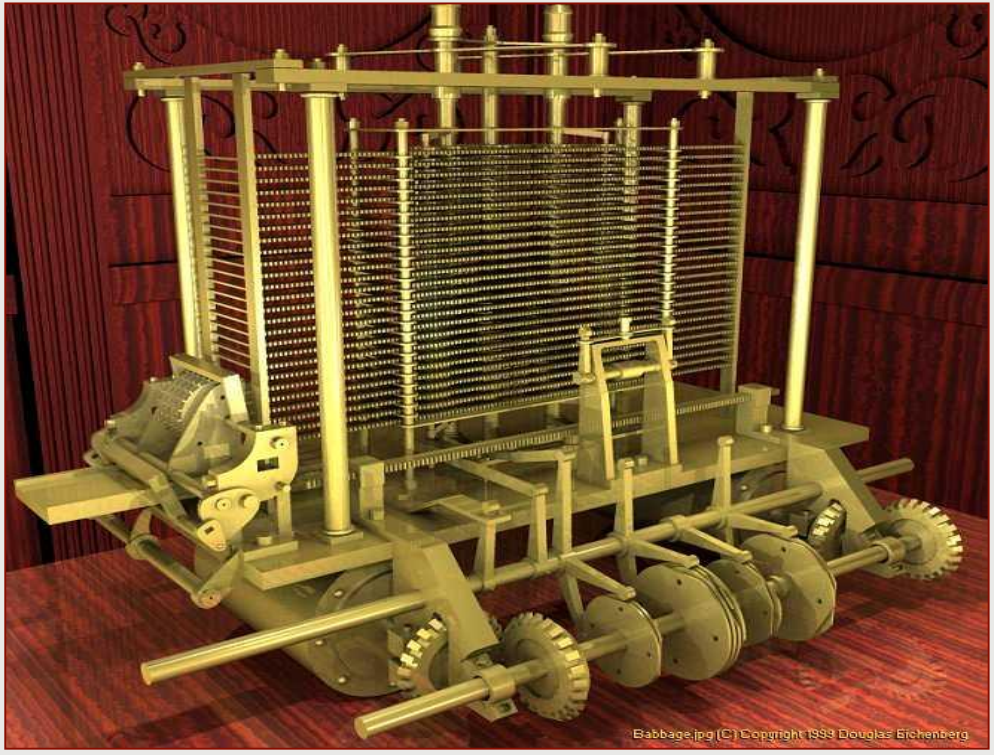
Gartner 2017 Magic Quadrant for Data Management Solutions for Analytics

Gartner 2016 Magic Quadrant for Data Warehouse and Database Management Solutions for Analytics

Gartner 2016 Magic Quadrant for Data Warehouse and Database Management Solutions for Analytics

Gartner 2017 Magic Quadrant for Data Management Solutions for Analytics

Gartner Magic Quadrant for Data Science Platforms, February 2017



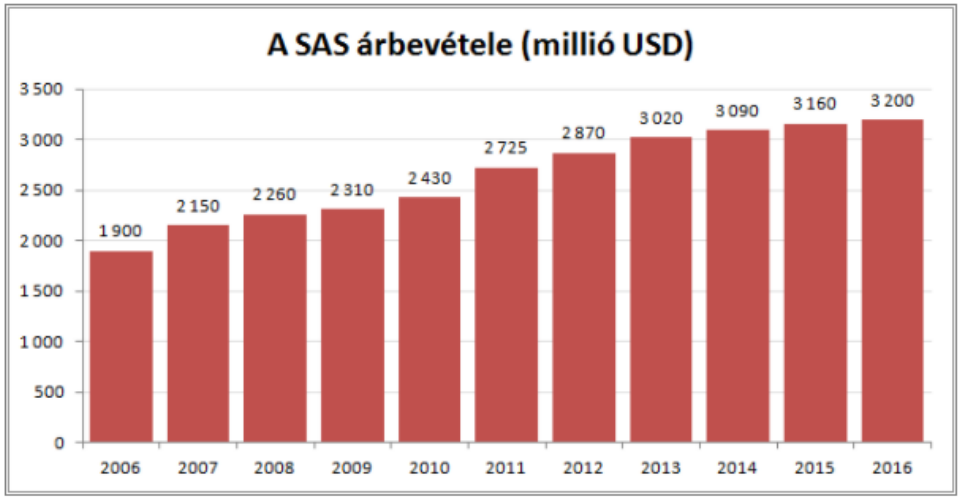
Szállítók

2016-ban is növekedett a SAS

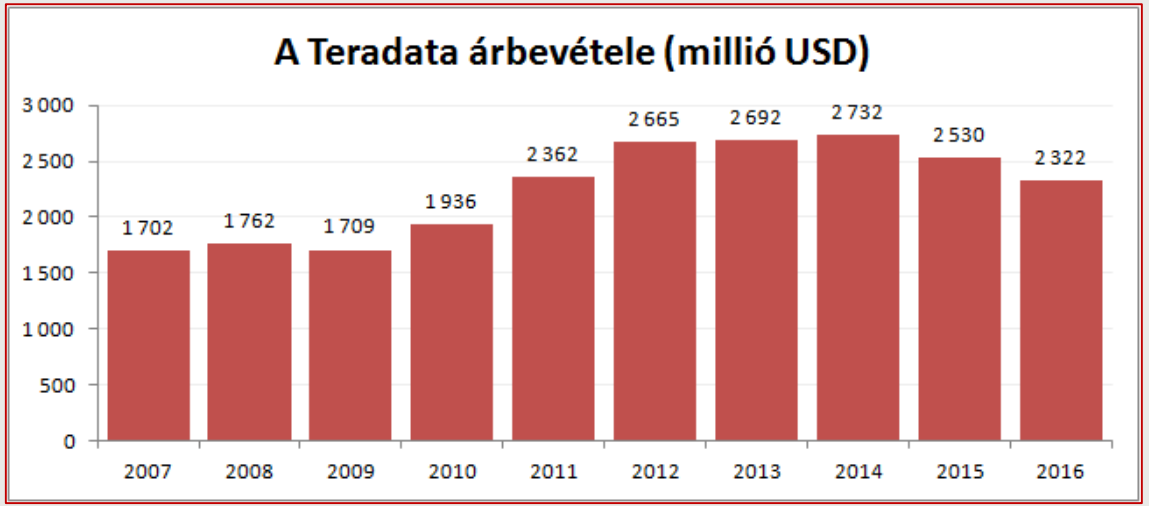
2017. január 27. ■ Üzleti intelligencia · Elemzés · SAS ■ 4459 Edit

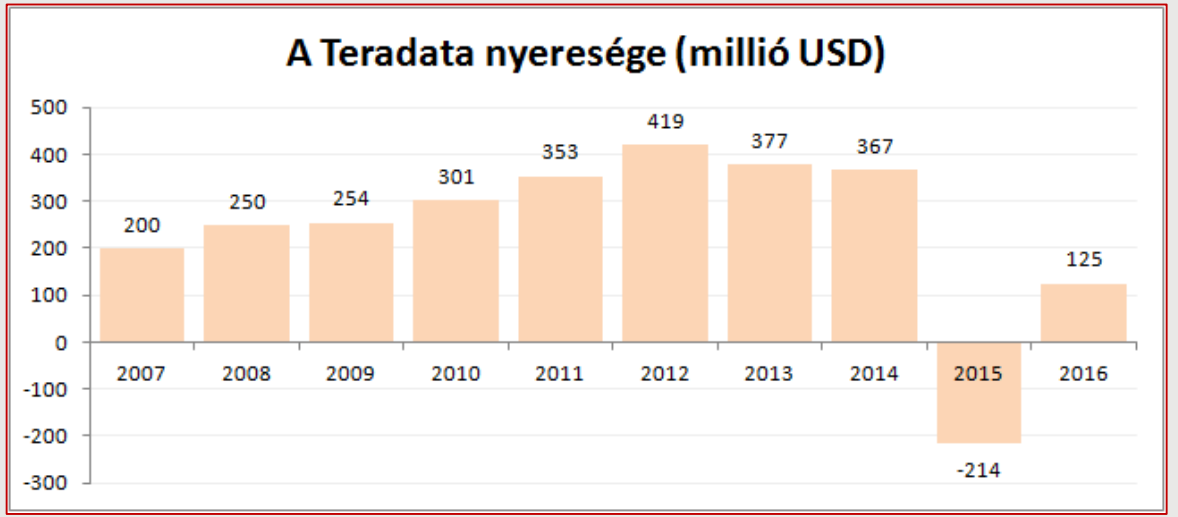
A tavalyi évben is folytatódott a SAS több évtizedes töretlen növekedési sorozata. Az analitikai óriás tíz éve, 2007-ben lépte át a 2 milliárd dolláros árbevételi határt, 2013-ban pedig a hárommilliárdosat.

A 2016-os árbevétel 3,2 milliárd dollárra rúgott, ami 1,3%-os növekedés az előző évhez képest, illetve 4 százalékos, ha a dollár árfolyamváltozásának hatását kiszűrjük.



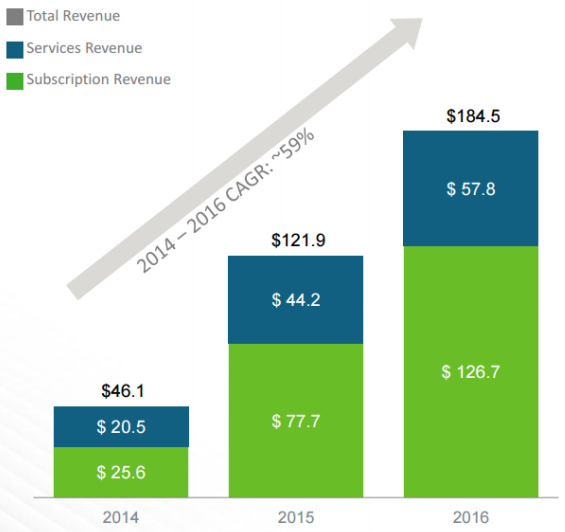
Teradata


A Teradata árbevétele (millió USD)
3 000
2 500
2 000
1 500
1 000
500
0
1 702
1 762
1 709
1 936
2 362
2 665
2 692
2 732
2 530
2 322
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016

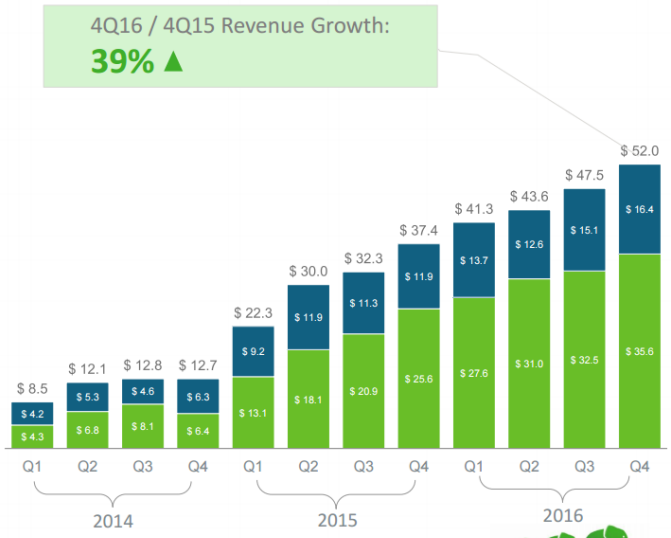

A Teradata nyeresége (millió USD)
500
400
300
200
100
0
-100
-200
-300
200
250
254
301
353
419
377
367
-214
125
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016

Significant Revenue Growth and Momentum *($mil)*

Annual Revenue ($mil)



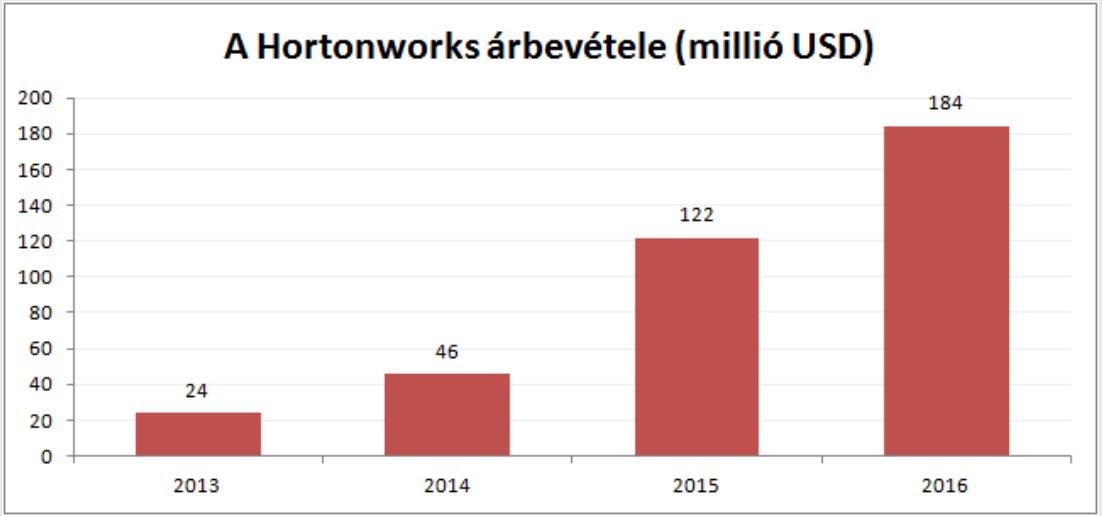
Quarterly Revenue ($mil)

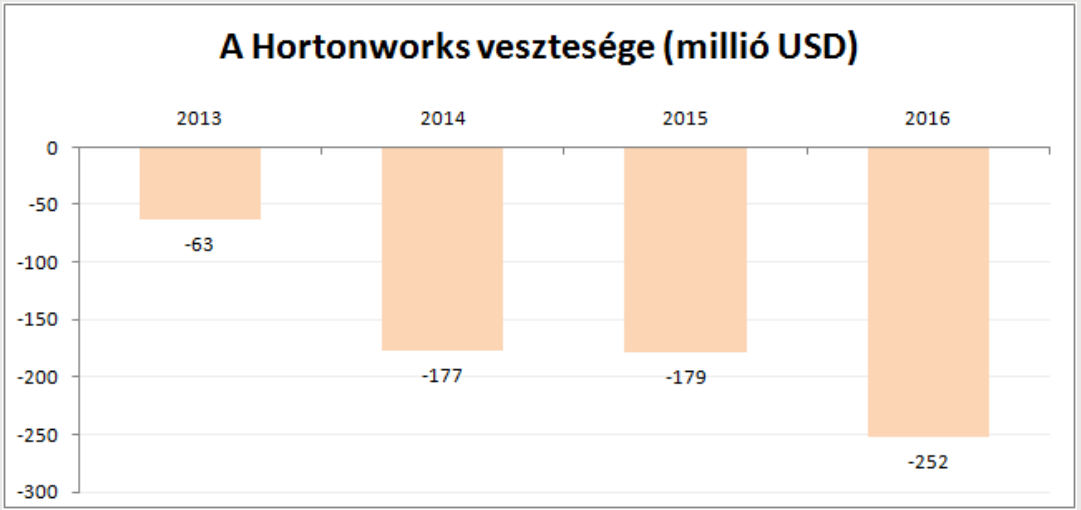


HORTONWORKS

© Hortonworks Inc. 2011 – 2017. All Rights Reserved

Hortonworks



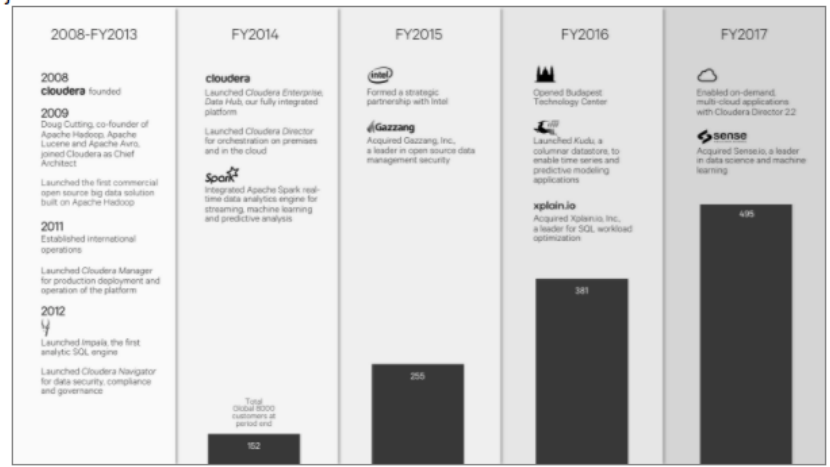


Tőzsdére megy a Cloudera is

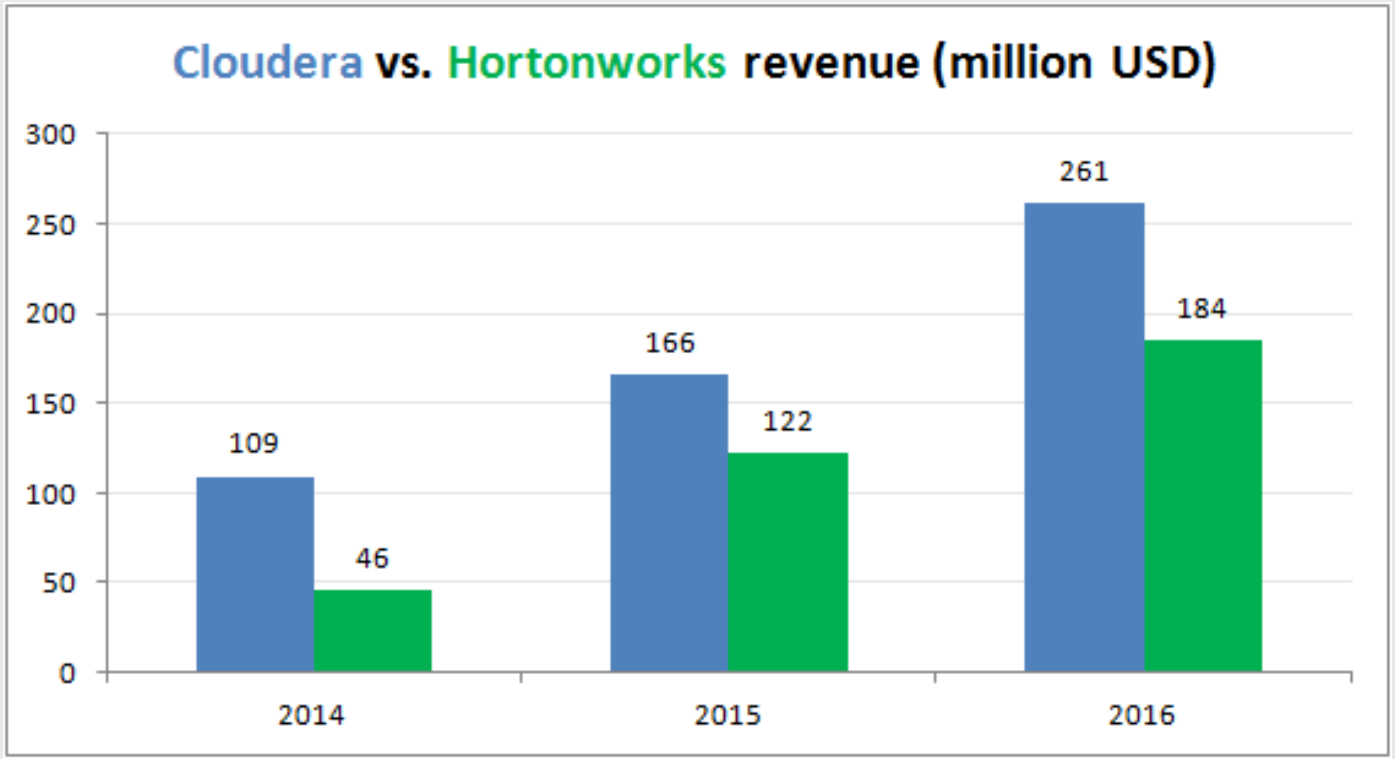
2017. április 19. ■ Üzleti intelligencia · BigData · Cloudera · Elemzés · Hortonworks ■ 4515
Edit

Hosszú várakozás után bejelentette a Cloudera, hogy megjelenik a tőzsdén. A részvénykibocsátás keretében körülbelül 200 millió dollár tőkét terveznek bevonni.

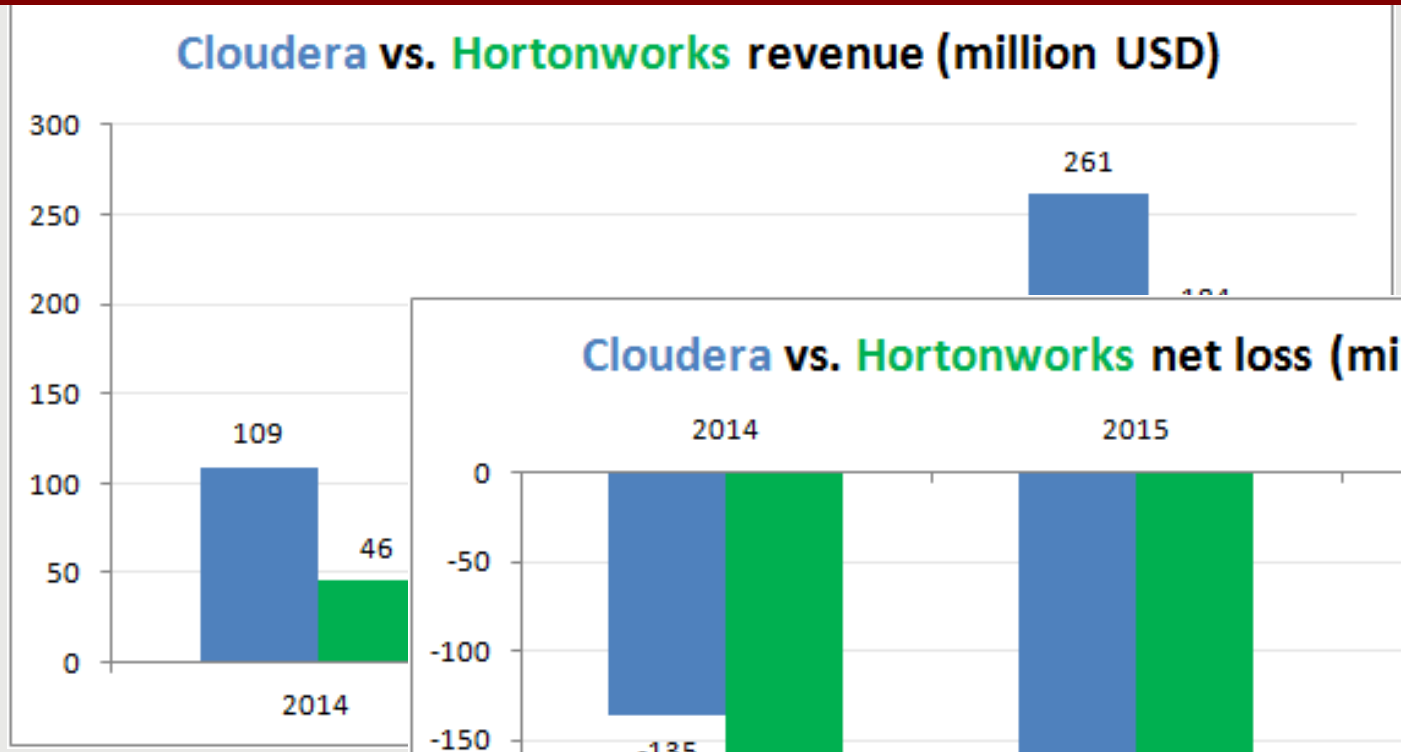
A 2008-ban alapított Cloudera történetének legfontosabb mérföldköveit az alábbi ábra mutatja be.

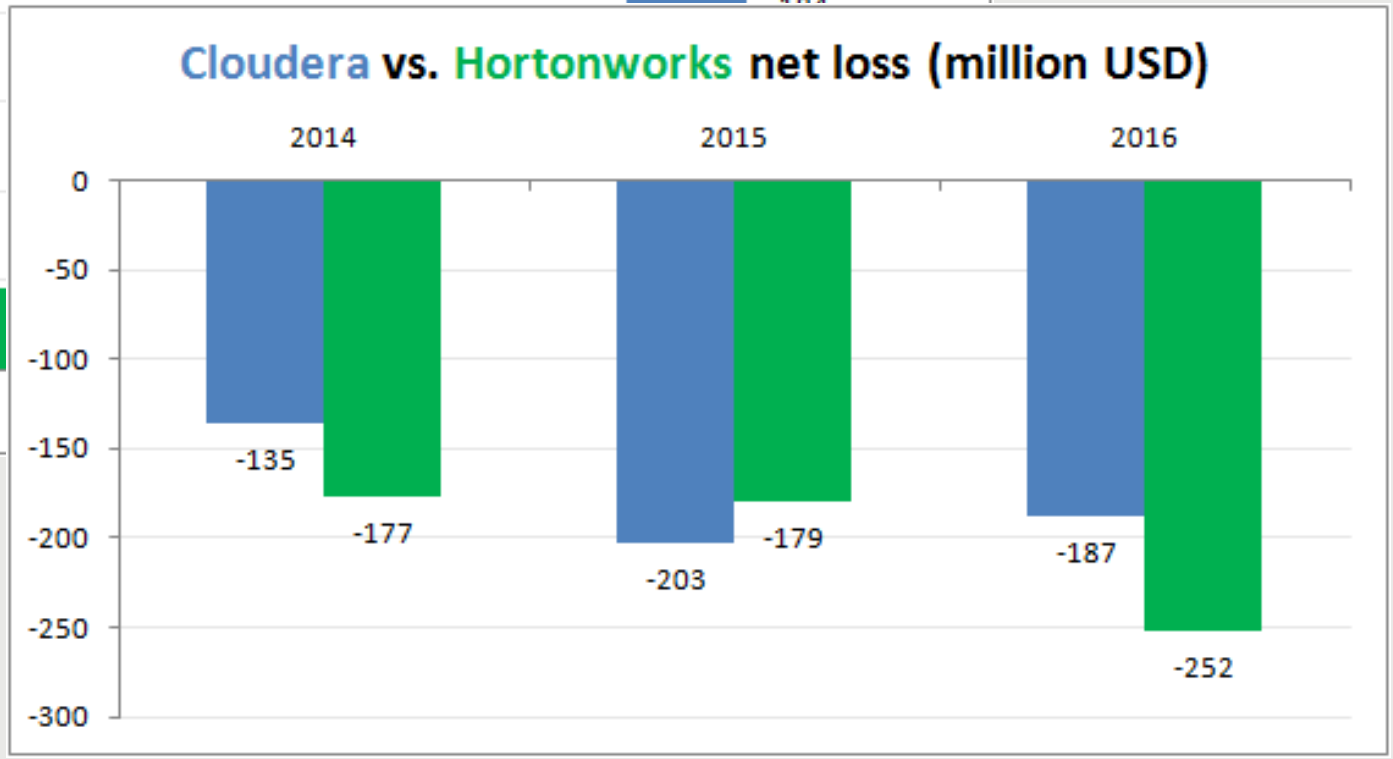


Hortonworks



Hortonworks





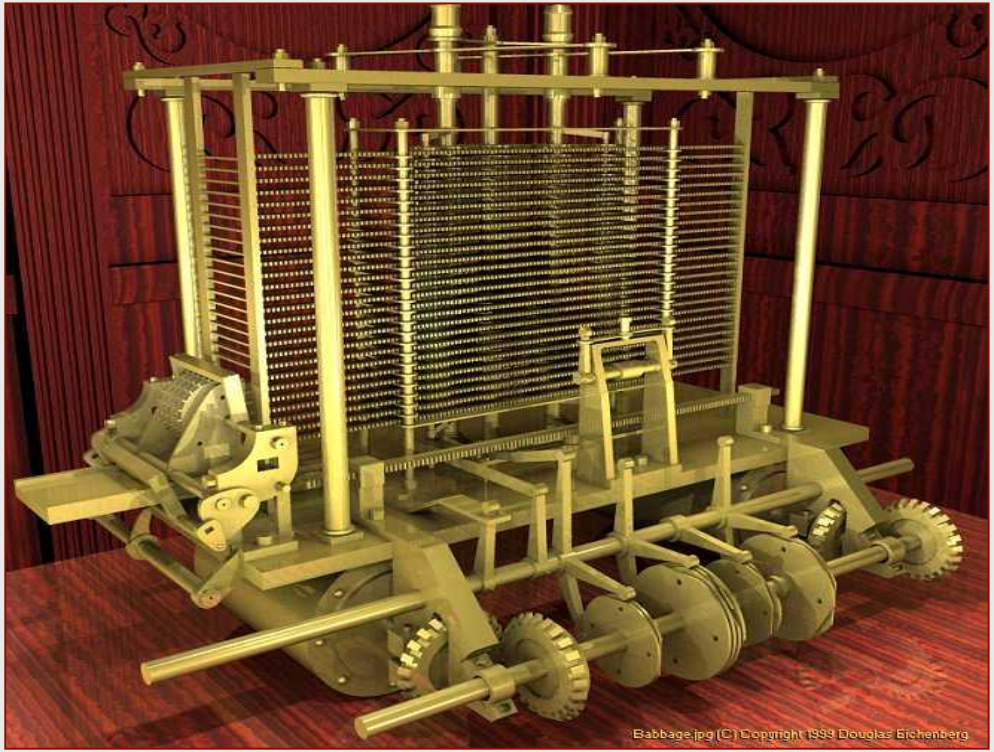
Felvásárolja a Google a Kaggle-t

2017. március 9. ■ Adatbányászat ■ 4507 Edit

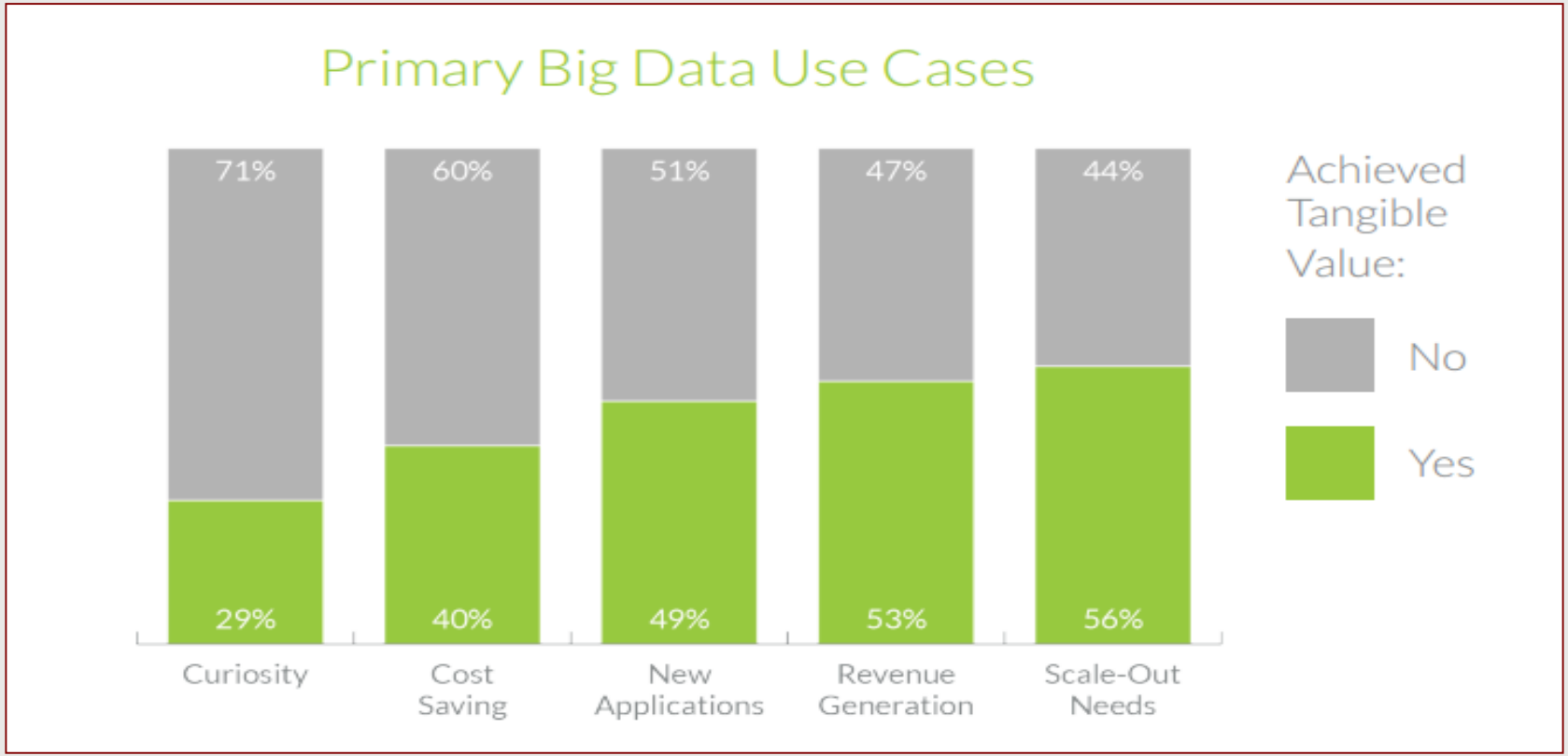
A most zajló Cloud Next konferencián jelentette be a Google, hogy felvásárolja az adatbányászati versenyek szervezésről ismert Kaggle-t.

A Kaggle hét évvel ezelőtt indult, az Eurovision dalverseny szavazási mintáit vizsgáló első kiírásban 21 csapat küzdött az 1.000 dolláros díjért. Azóta a platformnak több százezer felhasználója van, és rengeteg különböző data mining versenyt rendezett üzleti, tudományos és toborzási célokból.

A Kaggle blogposztja szerint továbbra is önállóságukat és saját arculatukat megőrizve fognak működni a Google berkein belül is.

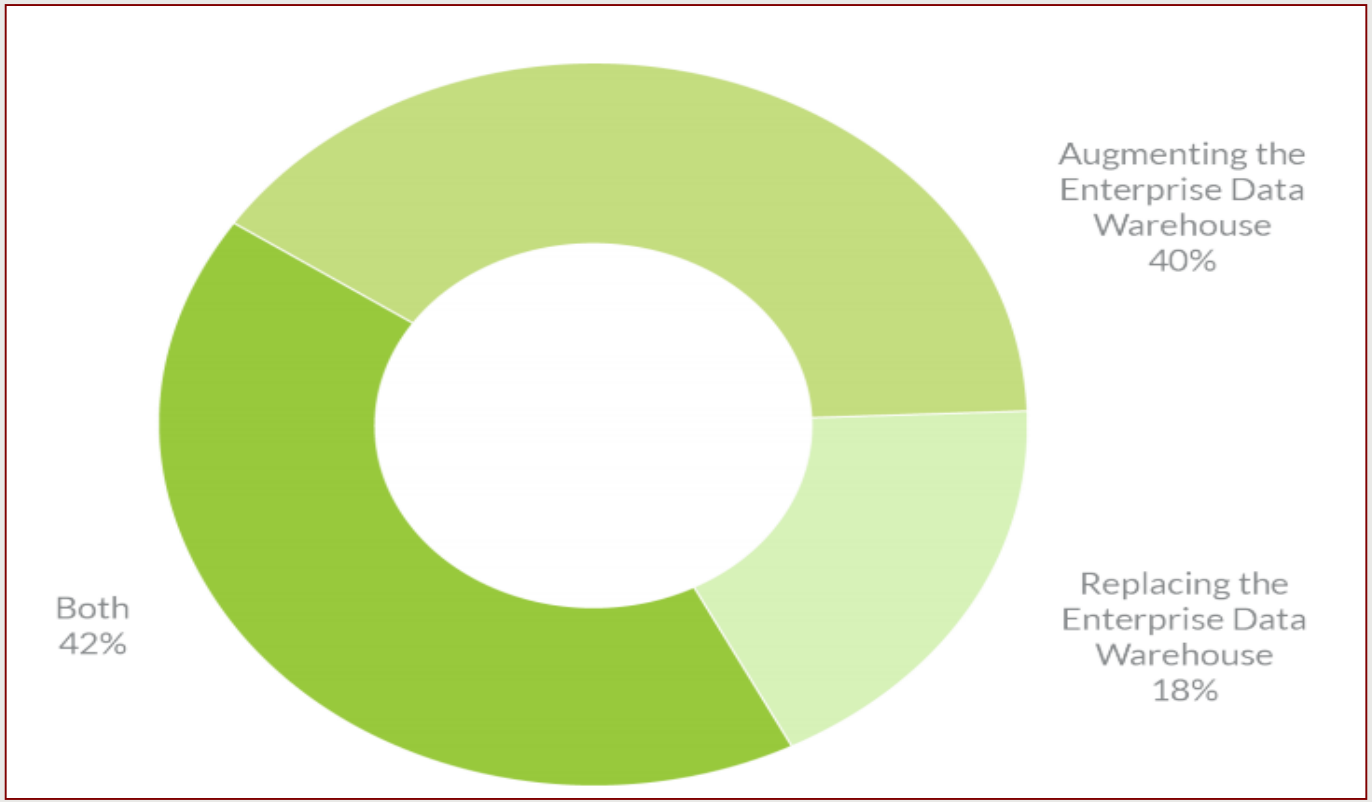


Big Data Trends



Atscale Big Data Maturity Survey 2016

Big Data vs. DW



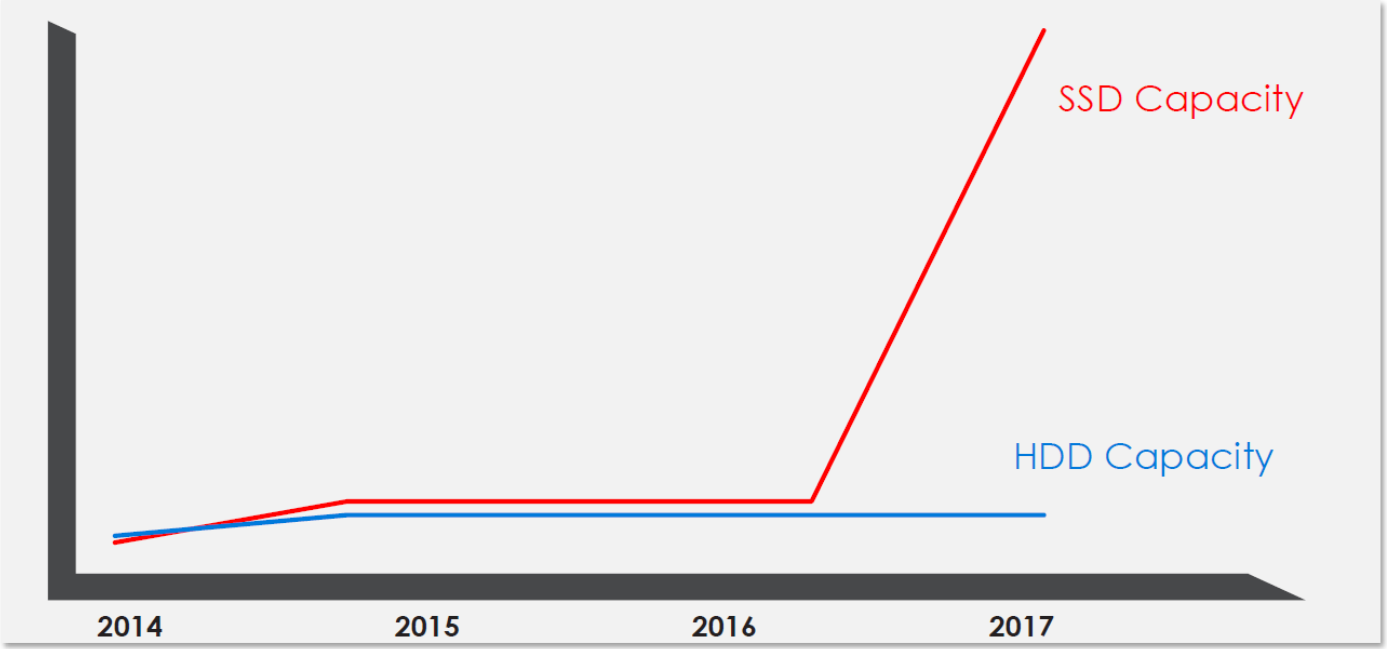
Atscale Big Data Maturity Survey 2016

Big Data & Cloud

Atscale Big Data Maturity Survey 2016

HDD vs. SSD

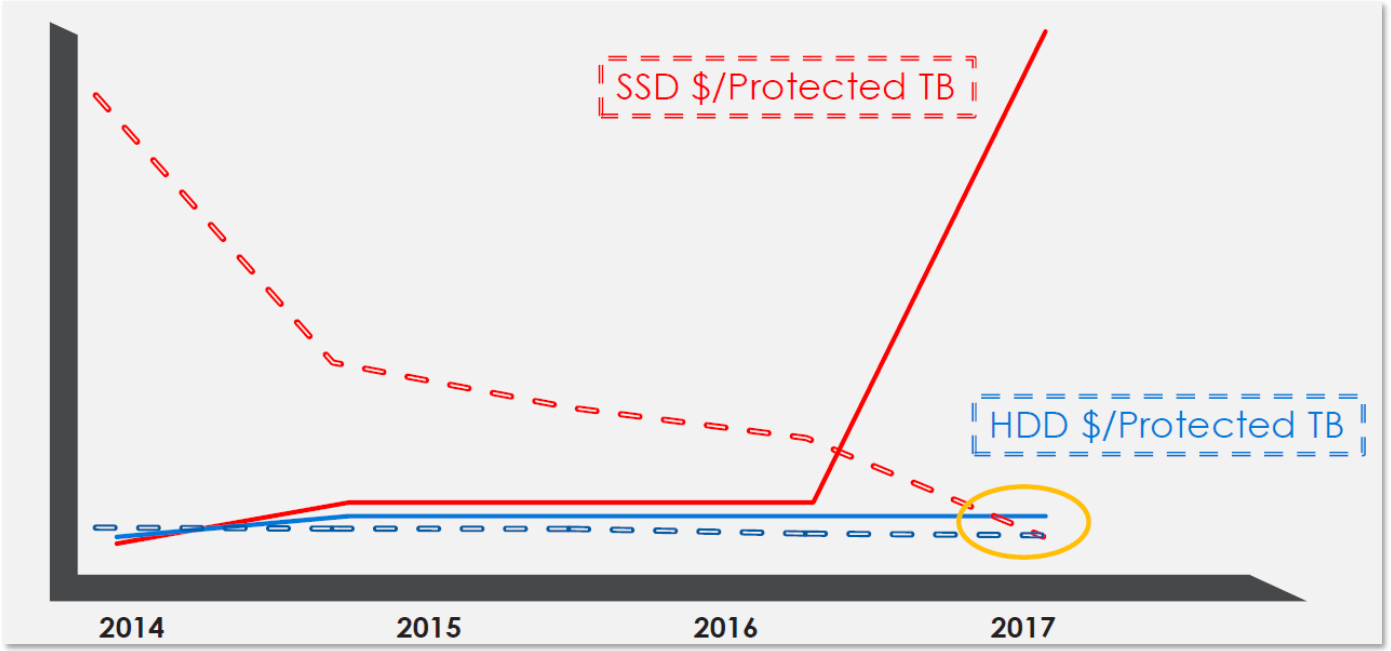
We've Reached the Cross-Over Point



TERADATA.

HDD vs. SSD

We've Reached the Cross-Over Point



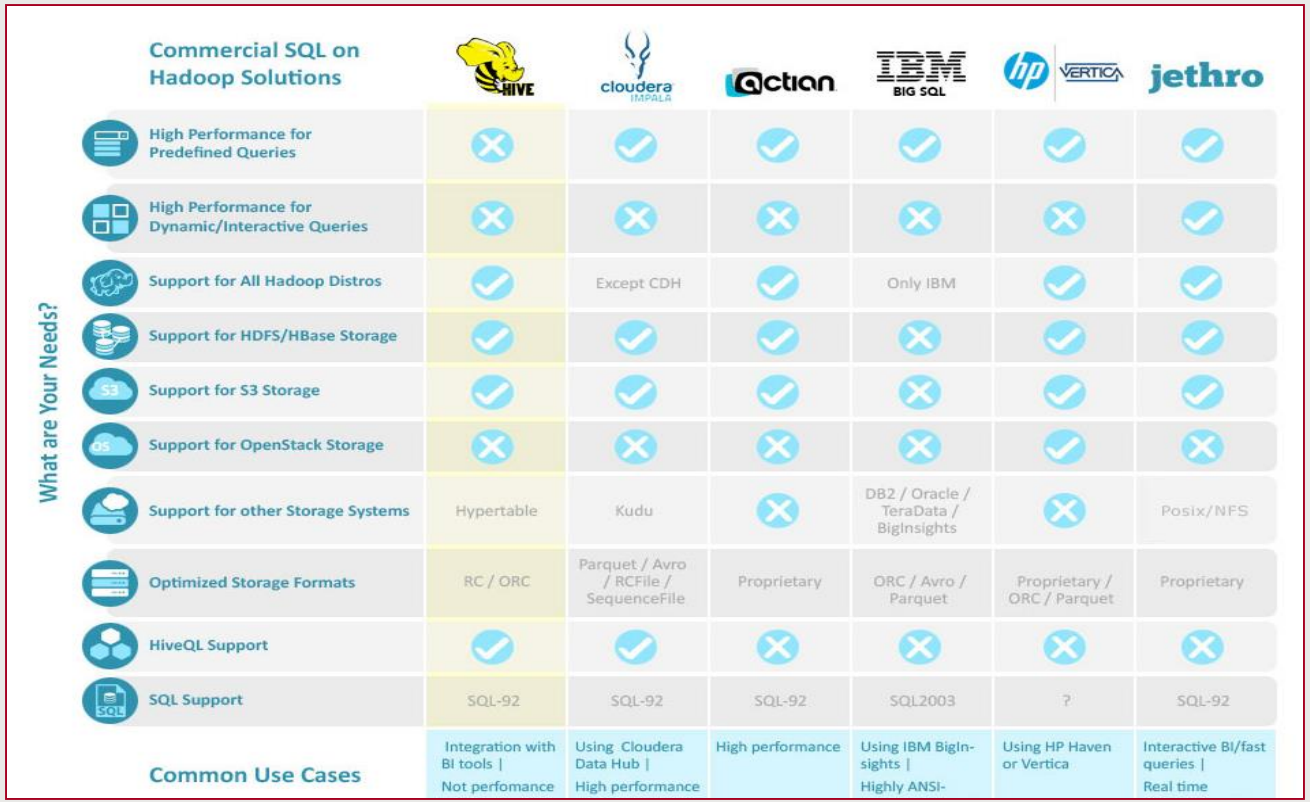
TERADATA.

SQL on Big Data

What are Your Needs? — Open Source SQL on Hadoop Solutions	Hive	Hive + Spark	Hive + Tez	Spark SQL	Presto	Apache Drill	HAWQ
High Performance for Predefined Queries	✗	✓	✓	✓	✓	✓	✓
High Performance for Dynamic/Interactive Queries	✗	✗	✗	✗	✗	✗	✗
Support for All Hadoop Distros	✓	✓	✓	✓	✓	✓	✓
Support for HDFS/HBase Storage	✓	✓	✓	✓	✓	✓	✓
Support for S3 Storage	✗	✓	✓	✓	✓	✓	✗
Support for WASB Storage	✗	✗	✓	✗	✗	✗	✗
Support for other Storage Systems	Hypertable	Cassandra	✗	Cassandra	RDBs / Redis	Kudu / GoogleCS / MongoDB	✗
Optimized Storage Formats	RC / ORC	SEQ / Parquet	ORC	SEQ / Parquet	ORC	JSON / Parquet	Parquet
HiveQL Support	✓	✓	✓	✗	✗	✗	✗
SQL Support	SQL-92	SQL-92	SQL-92	SQL-92 Subset	SQL-92	ANSI	SQL-99
Common Use Cases	Integration with BI tools | Not performance sensitive	Need high performanc | Multiple data streams	High performance | Complex distributed processing	High performance | Multiple data streams	Very large data scale | Interactive analytics	Both schema-declared and schema-on-read | Arbitrary formats	Greenplum user | Advanced analytics with MADlib

jethro.io/hadoop-hive

SQL on Big Data



What are Your Needs?	Commercial SQL on Hadoop Solutions	HIVE	cloudera IMPALA	Actian	IBM BIG SQL	hp VERTICA	jethro
	High Performance for Predefined Queries	✗	✓	✓	✓	✓	✓
	High Performance for Dynamic/Interactive Queries	✗	✗	✗	✗	✗	✓
	Support for All Hadoop Distros	✓	Except CDH	✓	Only IBM	✓	✓
	Support for HDFS/HBase Storage	✓	✓	✓	✗	✓	✓
S3	Support for S3 Storage	✓	✓	✓	✗	✓	✓
OS	Support for OpenStack Storage	✗	✗	✗	✗	✓	✗
	Support for other Storage Systems	Hypertable	Kudu	✗	DB2 / Oracle / TeraData / BigInsights	✗	Posix/NFS
	Optimized Storage Formats	RC / ORC	Parquet / Avro / RCFile / SequenceFile	Proprietary	ORC / Avro / Parquet	Proprietary / ORC / Parquet	Proprietary
	HiveQL Support	✓	✓	✗	✗	✗	✗
SQL	SQL Support	SQL-92	SQL-92	SQL-92	SQL2003	?	SQL-92
	Common Use Cases	Integration with BI tools | Not perfomance	Using Cloudera Data Hub | High performance	High performance	Using IBM BigInsights | Highly ANSI-	Using HP Haven or Vertica	Interactive BI/fast queries | Real time

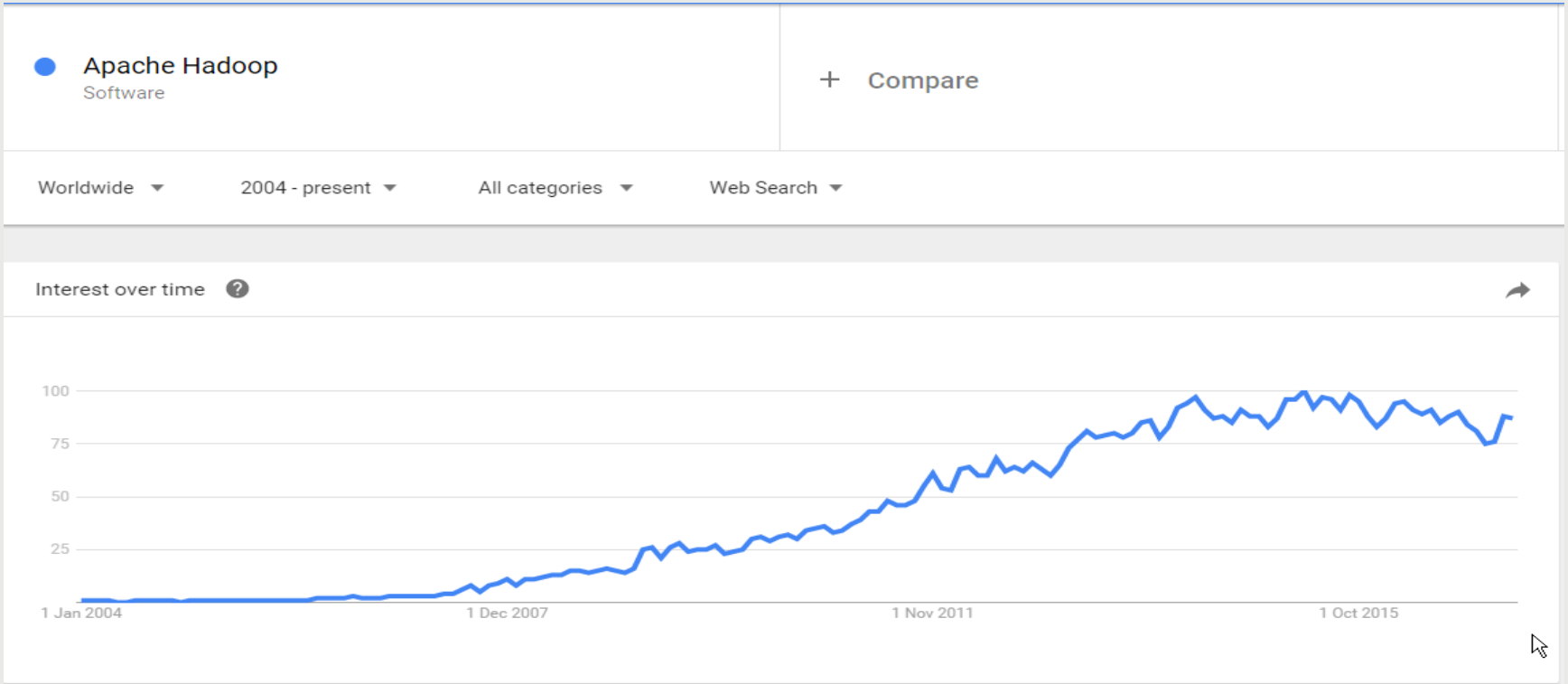
jethro.io/hadoop-hive

SQL on Big Data

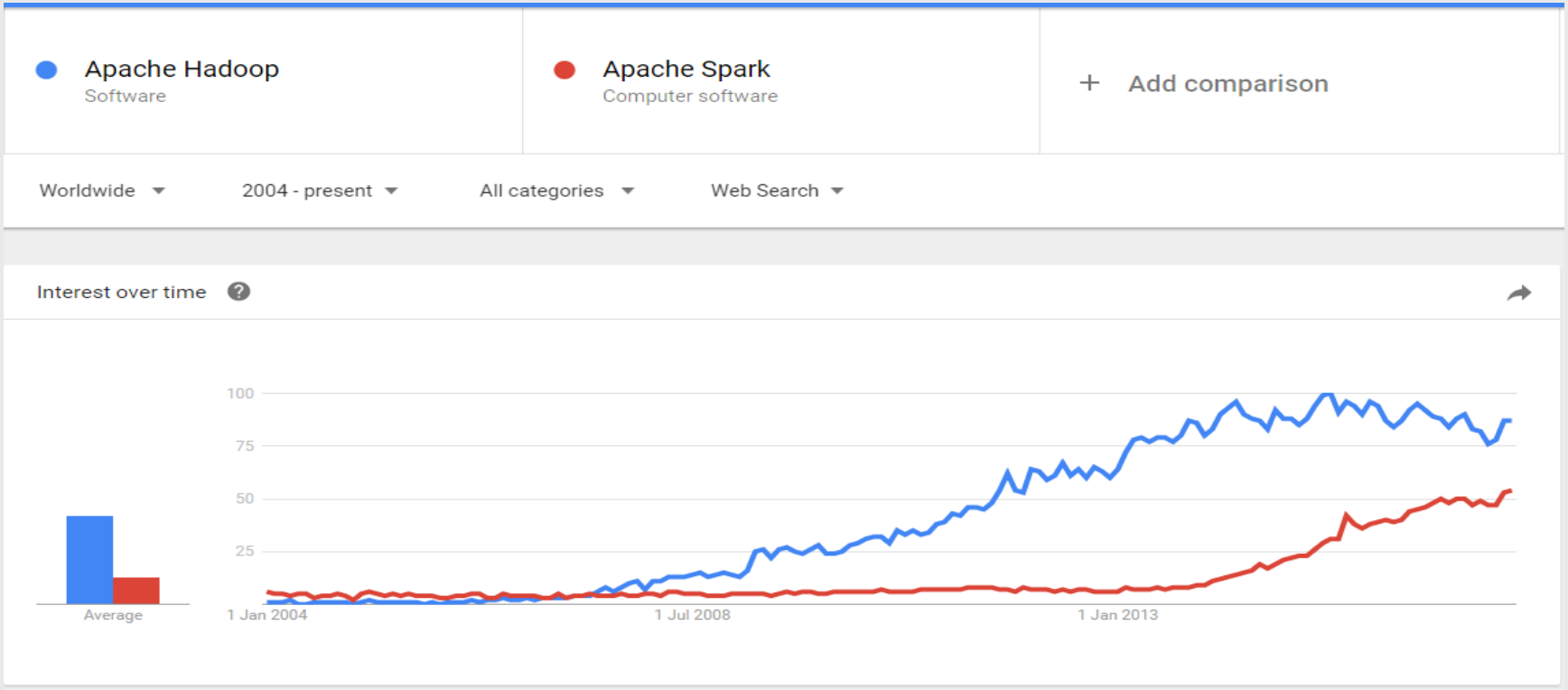
Table 3: Benchmark Query Results for Large Tables

Query	Query Execution Time (in seconds)			
	Impala 2.6	Spark 2.0	Hive 2.1 (LLAP)	Presto 0.152
Q1.1	5.6	4.8	10.5	10.3
Q1.2	5.0	3.8	9.1	8.7
Q1.3	5.6	3.3	9.3	7.9
Q2.1	8.0	11.8	10.0	12.4
Q2.2	6.2	11.0	9.3	10.4
Q2.3	6.0	10.6	9.3	9.8
Q3.1	12.6	15.5	37.5	35.5
Q3.2	15.5	15.6	38.1	39.9
Q3.3	8.0	9.2	28.1	24.9
Q3.4	7.4	6.8	28.7	15.7
Q4.1	97.3	64.9	137.6	134.4
Q4.2	49.1	30.3	131.9	123.8
Q4.3	26.8	13.2	137.9	135.5

Atscale BI on Hadoop Benchmark Q4-2016



Google Trends



Google Trends

Integrált stack

NEW

Hortonworks Data Platform 2.6 Key Highlights

Interactive Query in Seconds: Hive with LLAP and ACID Merge

Enterprise Spark at Scale: Spark 2.1 and Zeppelin 0.7

Real-Time Applications: HBase/Phoenix and Druid Support

Streamlined Operations: Apache Ambari 2.5

© Hortonworks Inc. 2011 – 2017. All Rights Reserved

HORTONWORKS

Enterprise adoption

Customers Speaking in Enterprise Adoption Track

BMW Group – "Driving the Elephant from the Lab to Global Production

ING – "ING CoreIntel – Collect and Process Network Logs Across Data Centers in Real Time"

Danske Bank– "Depositing Vale from Transactional Data at Danske Bank"

Worldpay – ""Delivering Multi-Tenancy Applications in a Secure Operational Platform"

Klarna – "Data Journey at Klarna"

Natixis – "How Hadoop Makes the Natixis Pack More Efficient"

Groupe Renault– "Renault Group's Journey to a Data-Driven Business"

Societe Generale – "Big Data & Data Science @ Societe Generale"

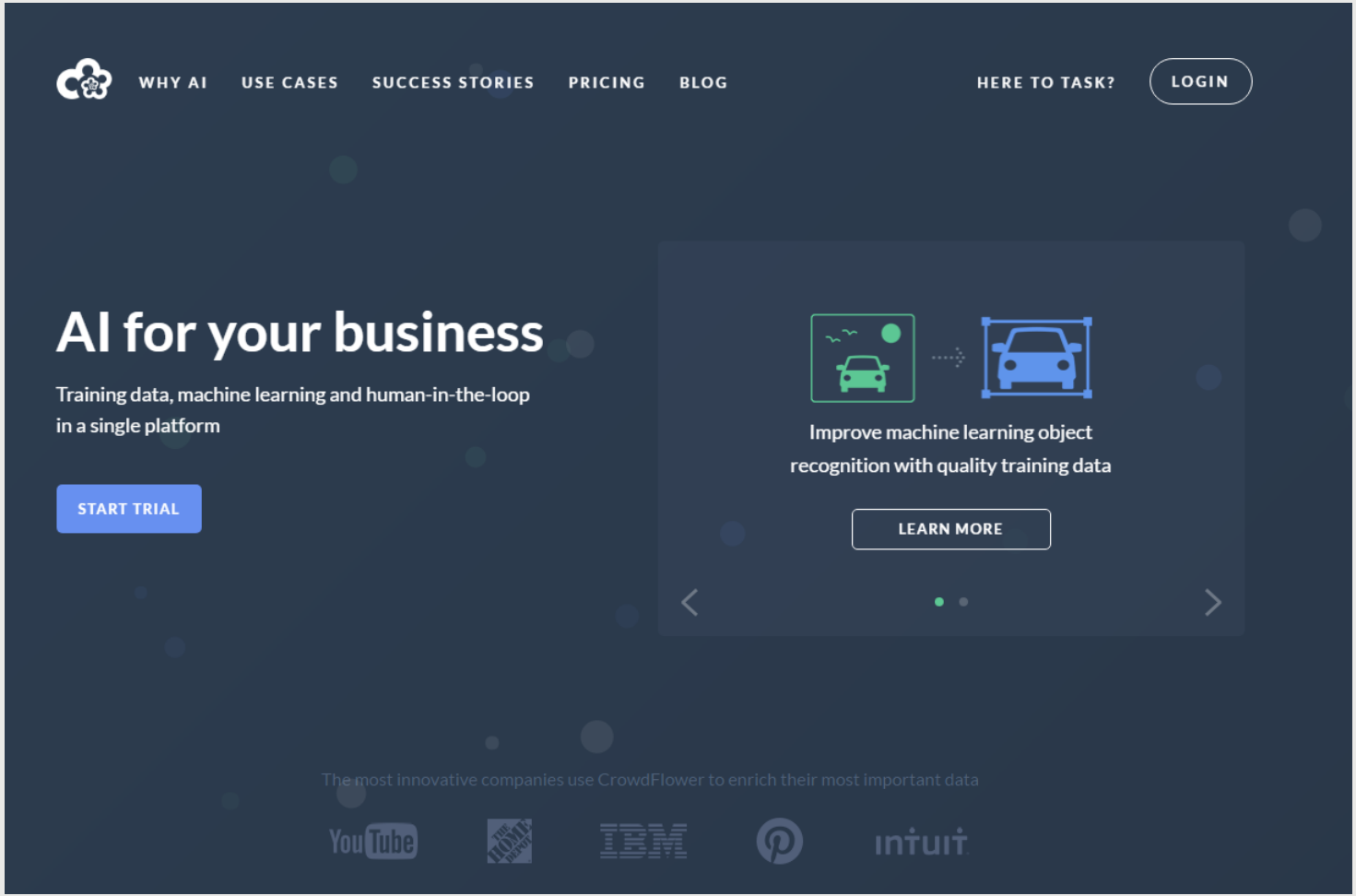
Munich Re – "Munich Re's Journey to the Big Data & Analytics Self-Service Platform Including a Central Data Lake"

Open Energi – "Machine Learning in the IoT with Apache NiFi"

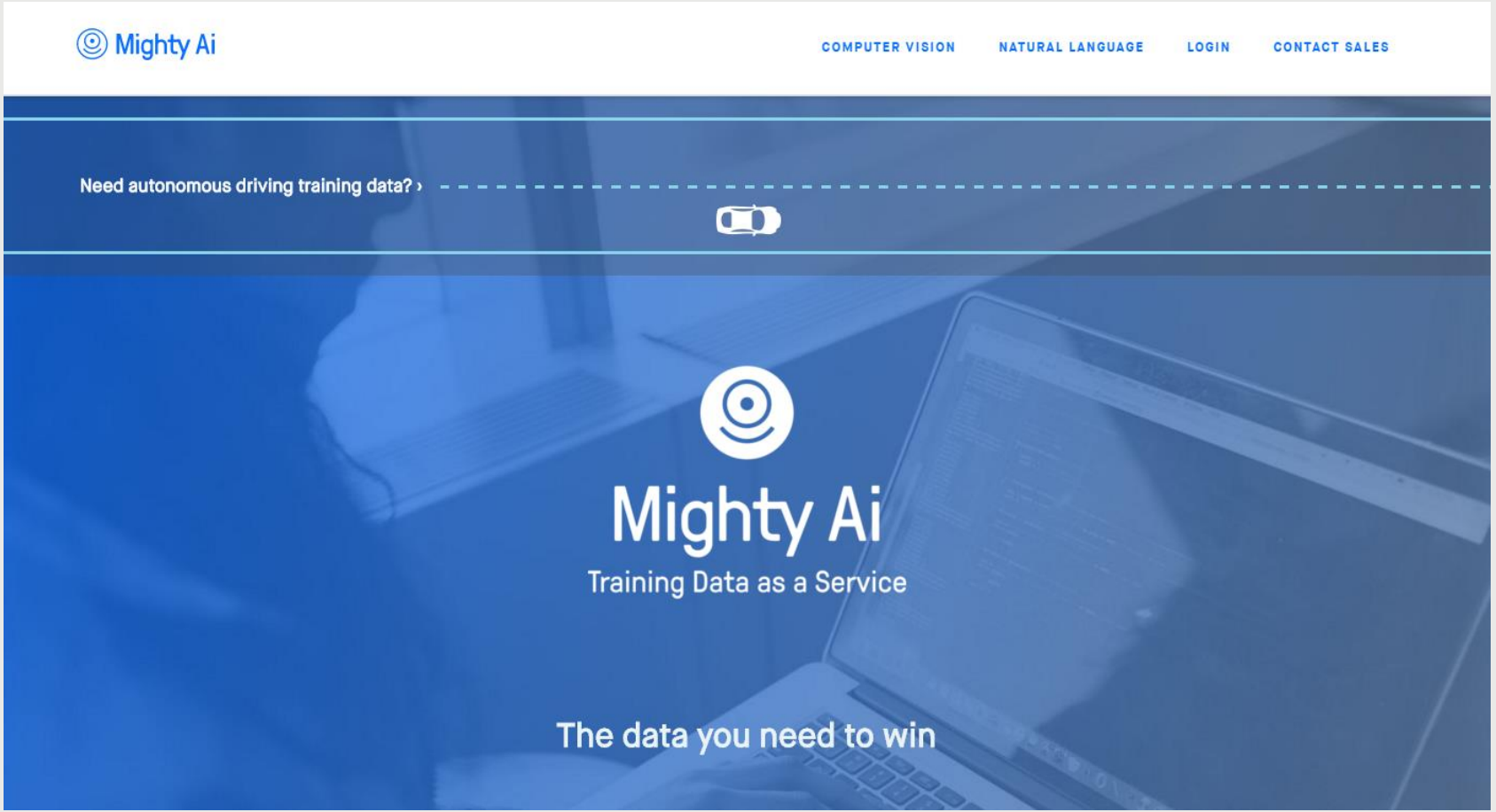
Walgreens Boots Alliance– "How Data is Driving Walgreens Boots Alliance's Retail Space & Location Strategies "

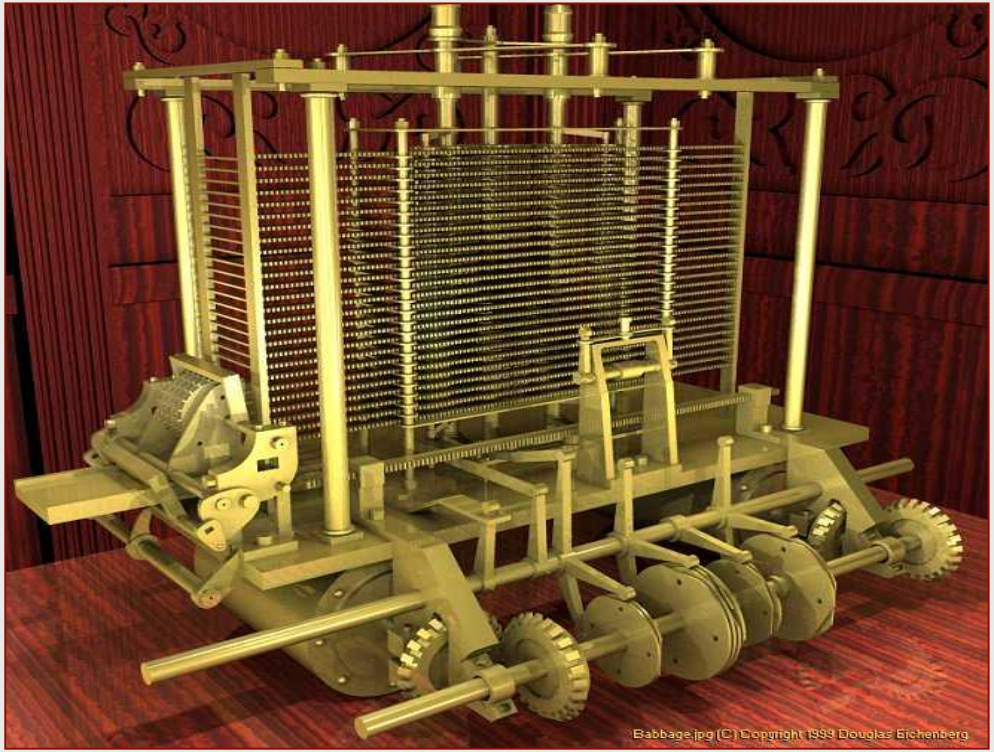
© Hortonworks Inc. 2011 – 2017. All Rights Reserved

Test Data as a service



Test Data as a service





Data Science

Data Science is dead.

Published on June 9, 2017

Cloud ML API

By 2020, the market for machine learning applications will reach $40 billion, IDC, a market research firm, estimates. And 60 percent of those applications, the firm predicts, will run on the platform software of four companies — Amazon, Google, IBM and Microsoft.

Cloud ML API

By 2020, the market for machine learning applications will reach $40 billion, IDC, a market research firm, estimates. And 60 percent of those applications, the firm predicts, will run on the platform software of four companies — Amazon, Google, IBM and Microsoft.

60% of ML applications will run on Amazon, Google, IBM and Microsoft

Apple Announces Core ML: Machine Learning Capabilities on Apple Devices

Like | by **Roland Meertens** on Jun 09, 2017. Estimated reading time: 1 minute | Discuss

Share | Reading List | Read later

At WWDC 2017 Apple announced ways it uses machine learning, and ways for developers to add machine learning to their own applications.

Their machine learning API, called Core ML, allows developers to integrate machine learning models into apps running on Apple devices with iOS, macOS, watchOS, and tvOS. Models reside on the device itself, so data never leaves the device.

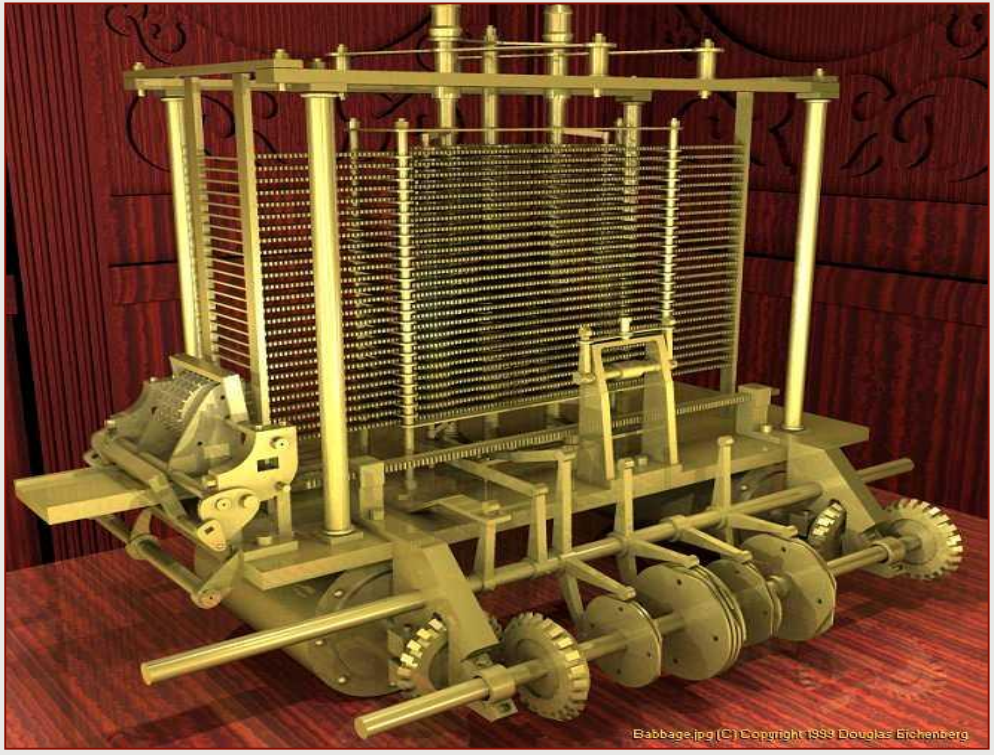
Multiple API calls are already available that application developers can use without having to add any additional models to their app. Examples of such computer vision algorithms are face detection and tracking, landmark detection, and event detection. Developers are also able to analyze natural language, for example in emails, text, and web pages. Natural language

RELATED CONTENT

ARKit Sets the Foundations for Augmented Reality on Apple's Platform Jun 07, 2017

The InfoQ eMag: Introduction to Machine Learning Apr 19, 2017

Machine Learning at Scale Apr 19, 2017



A GPU korszak

github.com/luanfujun/deep-photo-styletransfer

github.com/luanfujun/deep-photo-styletransfer

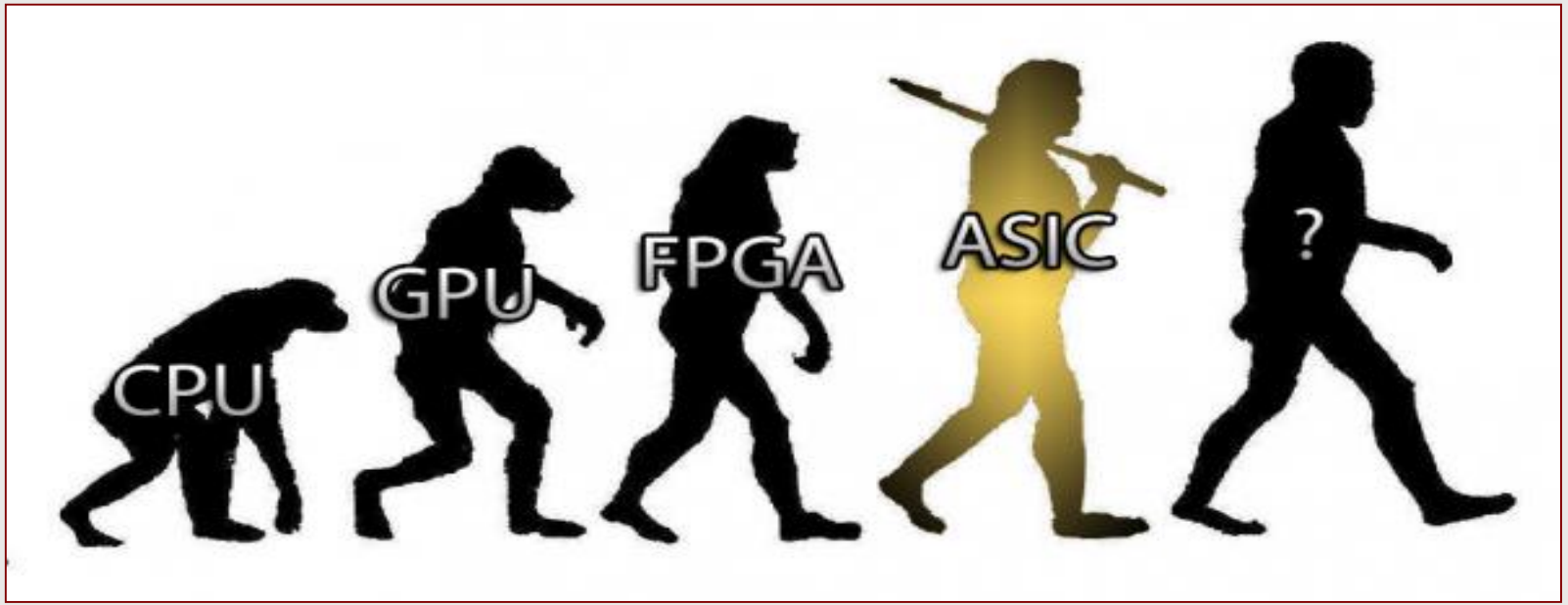
github.com/luanfujun/deep-photo-styletransfer

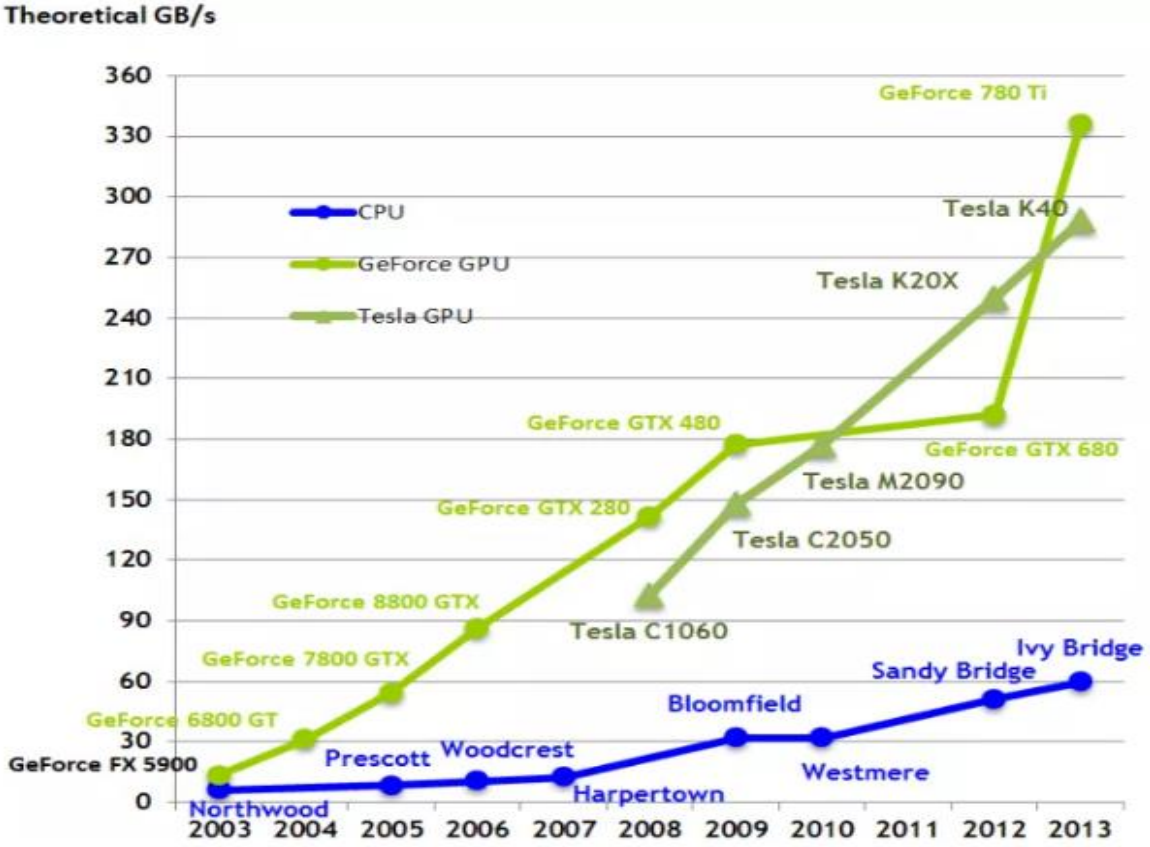
Using human-rated side-by-side comparison as a metric, the GNMT system produces translations that are vastly improved compared to the previous phrase-based production system. GNMT reduces translation errors by more than 55%-85% on several major language pairs measured on sampled sentences from Wikipedia and news websites with the help of bilingual human raters.

Input sentence:	*Translation (PBMT):*	*Translation (GNMT):*	*Translation (human):*
李克強此行將啟動中加總理年度對話機制，與加拿大總理杜魯多舉行兩國總理首次年度對話。	Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held its first annual session.	Li Keqiang will start the annual dialogue mechanism with Prime Minister Trudeau of Canada and hold the first annual dialogue between the two premiers.	Li Keqiang will initiate the annual dialogue mechanism between premiers of China and Canada during this visit, and hold the first annual dialogue with Premier Trudeau of Canada.

www.nextplatform.com/2017/03/20/google-team-refines-gpu-powered-neural-machine-translation

Hardware



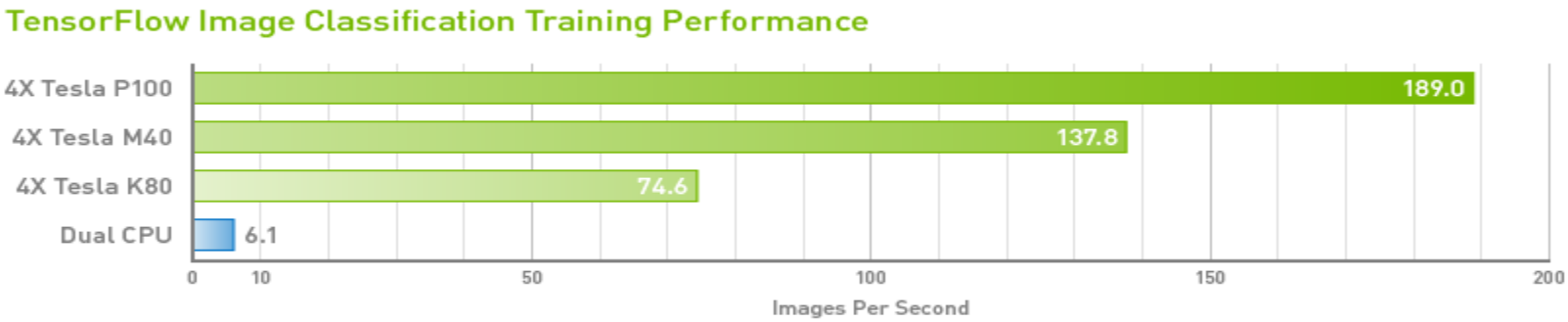


Comparison of bandwidth for CPUs and GPUs over time: Bandwidth is one of the main reasons why GPUs are faster for computing than CPUs are.

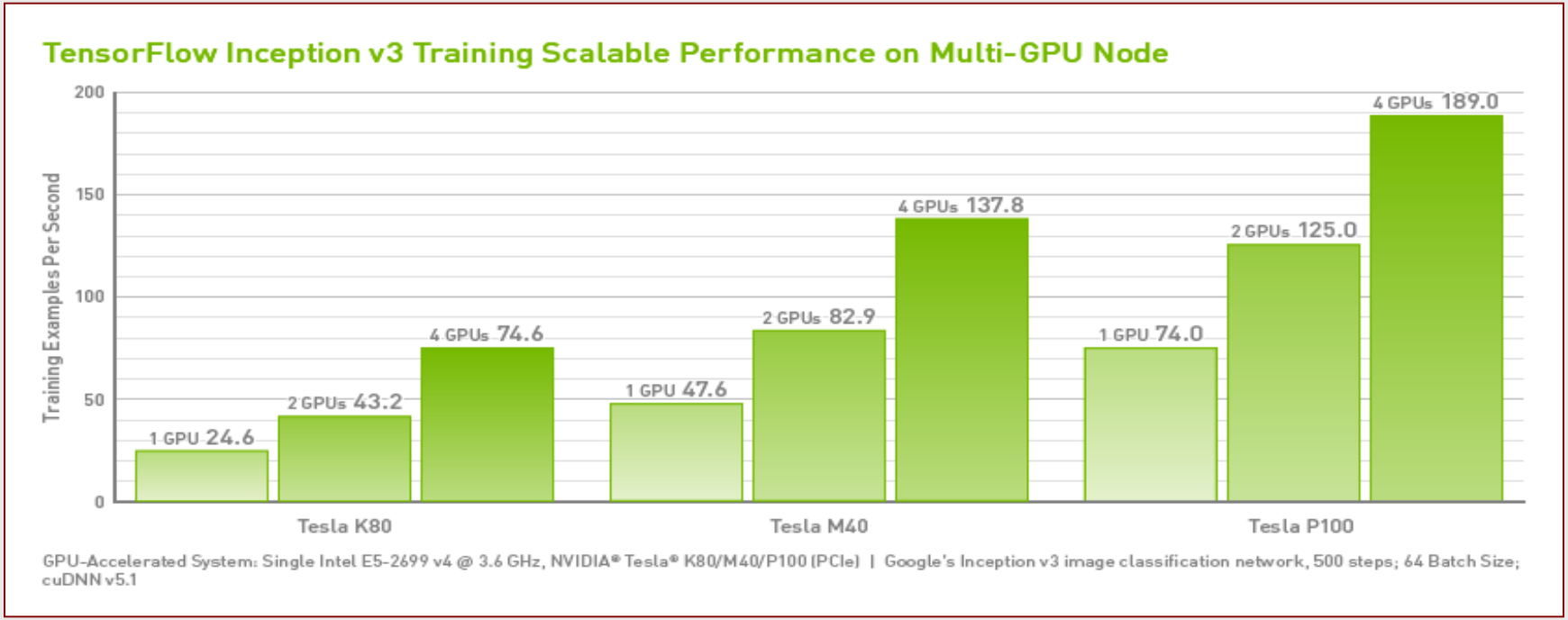
*One shortcoming of current NMT architectures is the amount of compute required to train them. Training on real-world datasets of several million examples typically requires **dozens of GPUs** and convergence time is on the order of **days to weeks**.*

*... an effort that required more than **250,000 GPU hours** on their in-house cluster, which is based on Nvidia Tesla K40m and Tesla K80 GPUs*

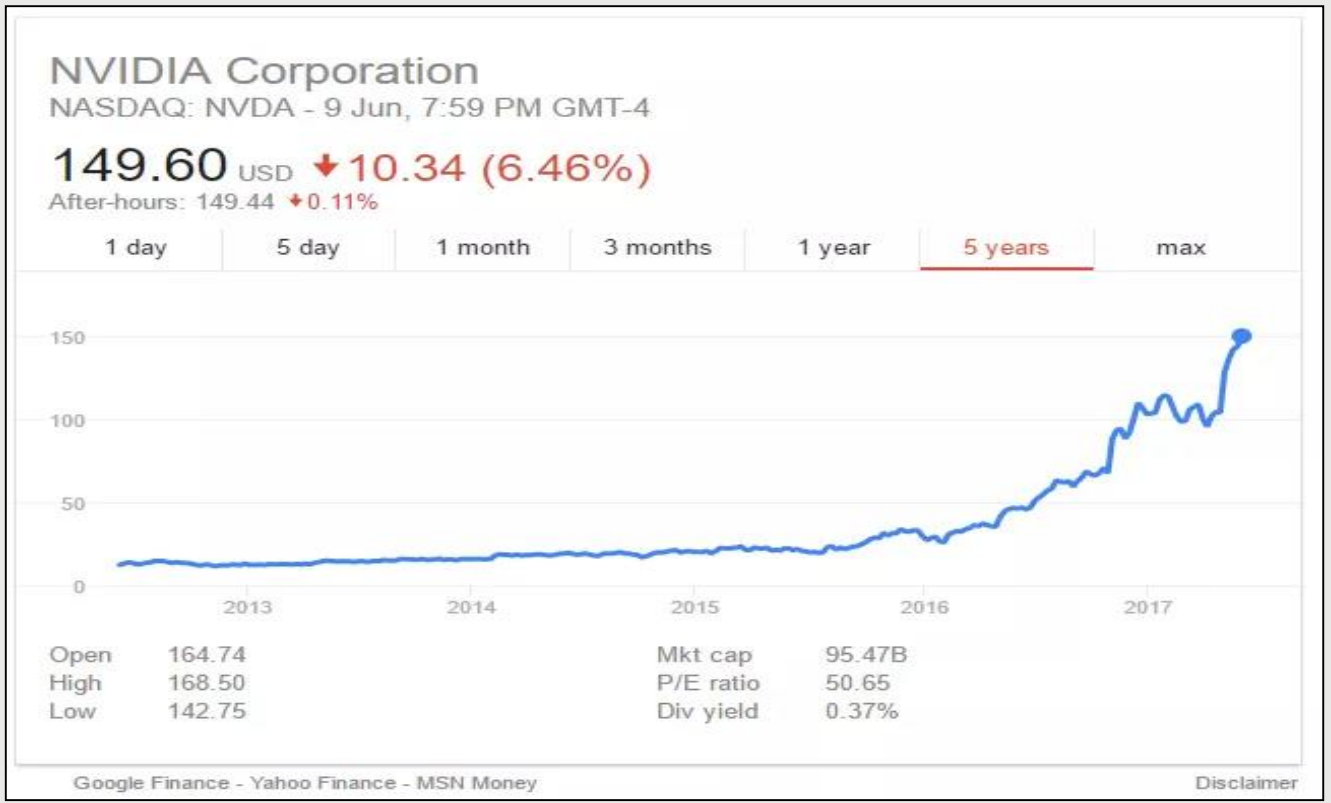
www.nextplatform.com/2017/03/20/google-team-refines-gpu-powered-neural-machine-translation



www.nvidia.com/object/gpu-accelerated-applications-tensorflow-benchmarks.html



www.nvidia.com/object/gpu-accelerated-applications-tensorflow-benchmarks.html


NVIDIA Corporation
NASDAQ: NVDA - 9 Jun, 7:59 PM GMT-4
149.60 USD ↓10.34 (6.46%)
After-hours: 149.44 ↓0.11%
1 day
5 day
1 month
3 months
1 year
5 years
max
150
100
50
0
2013
2014
2015
2016
2017
Open 164.74
High 168.50
Low 142.75
Mkt cap 95.47B
P/E ratio 50.65
Div yield 0.37%
Google Finance - Yahoo Finance - MSN Money
Disclaimer

GOOGLE CLOUD

MAY 17, 2017

Build and train machine learning models on our new Google Cloud TPUs

WRITTEN BY

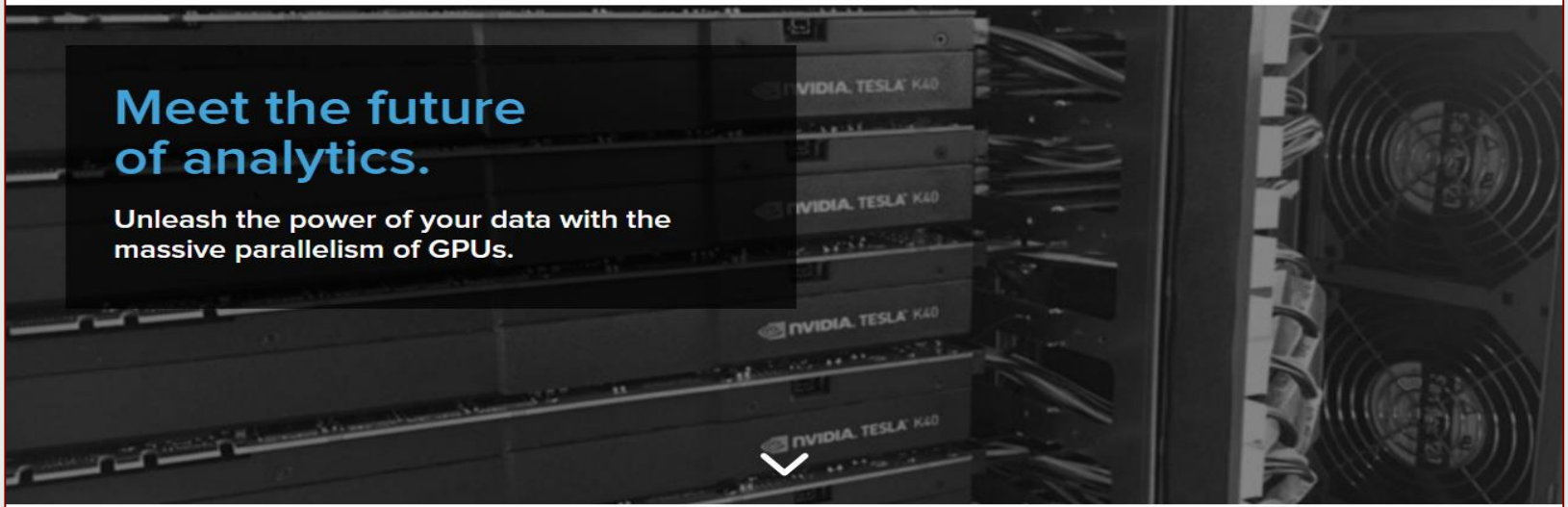
Jeff Dean
GOOGLE SENIOR FELLOW

Urs Hölzle
SENIOR VICE PRESIDENT, GOOGLE CLOUD INFRASTRUCTURE

We're excited to announce that our second-generation Tensor Processing Units (TPUs) are coming to Google Cloud to accelerate a wide range of machine learning workloads, including both training and inference. We call them Cloud TPUs, and they will initially be available via Google Compute Engine.

MAPD

Product Demos Company Resources Blog Get the Whitepaper



UNPARALLELED PERFORMANCE

Supercomputing in a standard server, for your most urgent business questions

MapD's available 4U server fits up to eight Nvidia K80 GPUs with 192GB of ultra-fast video RAM, allowing SQL queries to be processed at rates approaching 3TB/second by almost 40,000 cores.

Huge SQL speedups without the hassle

The MapD database intelligently partitions, compresses and caches your data across all GPUs, meaning users just see up to 100X faster SQL, no indexing or optimizations required.

GPU Open Analytics

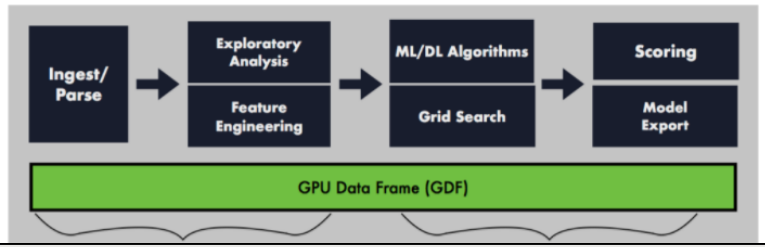
Home | Github | Google Groups

GPU Open Analytics Initiative

Continuum Analytics, H2O.ai, and MapD Technologies have announced the formation of the GPU Open Analytics Initiative (GOAI) to create common data frameworks enabling developers and statistical researchers to accelerate data science on GPUs. GOAI will foster the development of a data science ecosystem on GPUs by allowing resident applications to interchange data seamlessly and efficiently.

Github | Google Groups

goai
GPU OPEN ANALYTICS INITIATIVE



GPU Data Frame

Architecture

Our first project: an open source GPU Data Frame with a corresponding Python API. The GPU Data Frame is a common API that enables efficient interchange of data between processes running on the GPU. End-to-end computation on the GPU avoids transfers back

GPU for Deep Learning

TL;DR advice

Best GPU overall: Titan X Pascal and GTX 1080 Ti

Cost efficient but expensive: GTX 1080 Ti, GTX 1070

Cost efficient and cheap: GTX 1060

I work with data sets > 250GB: Regular GTX Titan X or Titan X Pascal

I have little money: GTX 1060

I have almost no money: GTX 1050 Ti

I do Kaggle: GTX 1060 for any "normal" competition, or GTX 1080 Ti for "deep learning competitions"

I am a competitive computer vision researcher: Titan X Pascal or regular GTX Titan X

I am a researcher: GTX 1080 Ti. In some cases, like natural language processing, a GTX 1070 might also be a solid choice — check the memory requirements of your current models

I want to build a GPU cluster: This is really complicated, you can get some ideas here

I started deep learning and I am serious about it: Start with a GTX 1060. Depending of what area you choose next (startup, Kaggle, research, applied deep learning) sell your GTX 1060 and buy something more appropriate

timdettmers.com/2017/03/19/which-gpu-for-deep-learning

GPU for Deep Learning

TL;DR advice

Best GPU overall: Titan X Pascal and GTX 1080 Ti

Cost efficient but expensive: GTX 1080 Ti, GTX 1070

Cost efficient and cheap: GTX 1060

80.000 Ft

I work with data sets > 250GB: Regular GTX Titan X or Titan X Pascal

I have little money: GTX 1060

I have almost no money: GTX 1050 Ti

50.000 Ft

I do Kaggle: GTX 1080 for any "normal" competition, or GTX 1080 Ti for "deep learning competitions"

I am a competitive computer vision researcher: Titan X Pascal or regular GTX Titan X

I am a researcher: GTX 1080 Ti. In some cases, like natural language processing, a GTX 1070 might also be a solid choice — check the memory requirements of your current models

I want to build a GPU cluster: This is really complicated, you can get some ideas here

I started deep learning and I am serious about it: Start with a GTX 1060. Depending of what area you choose next (startup, Kaggle, research, applied deep learning) sell your GTX 1060 and buy something more appropriate

timdettmers.com/2017/03/19/which-gpu-for-deep-learning