Analytics 2017

An overview of some of the common terminology and usage concerning analytics and business intelligence.

Prepared by Justin Simonds for the IIBA Orange County Chapter

The world is fundamentally changing

Netflix

Amazon

Apple Pay

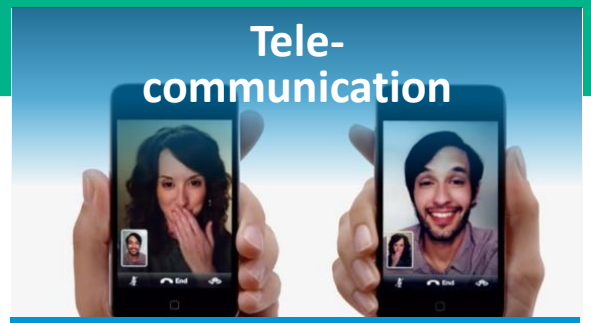
Uber

Market Disruptions

Fitbit

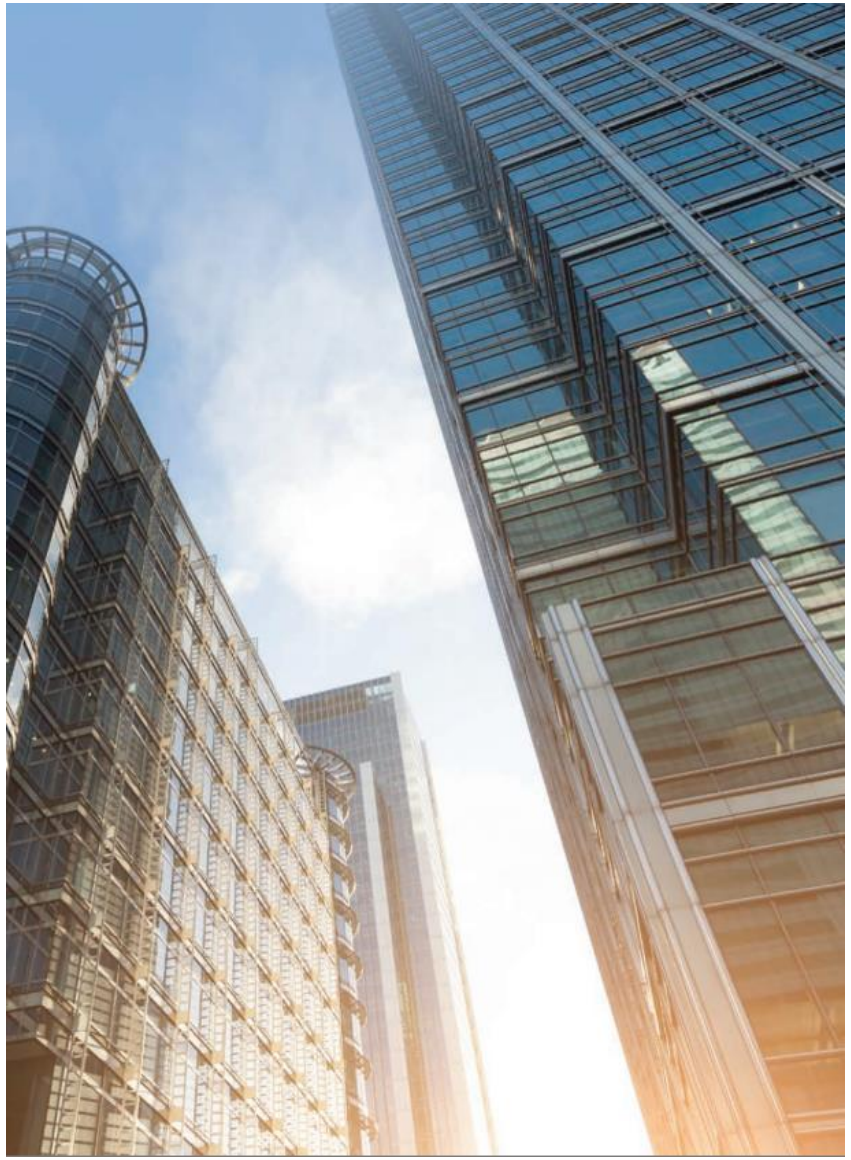
Linkedin

Progressive



Google

The Marketplace:



1/3rd of the top 20 market leaders will be disrupted by competitors by 2018

(IDC Predictions)

68% of businesses view time to market as a critical differentiator

(CSO Magazine)

5-20% increase in business value to organization via fast time to market

(CSO Insights)

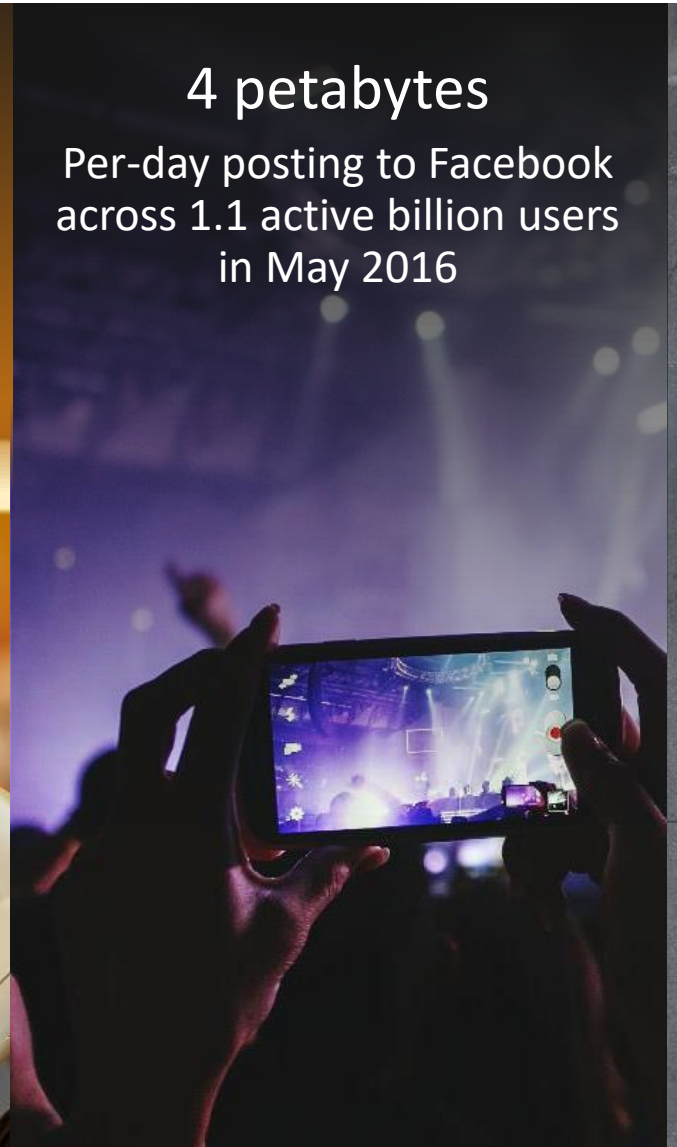
50+% of IT Budget going to LOB leaders vs. the CIO

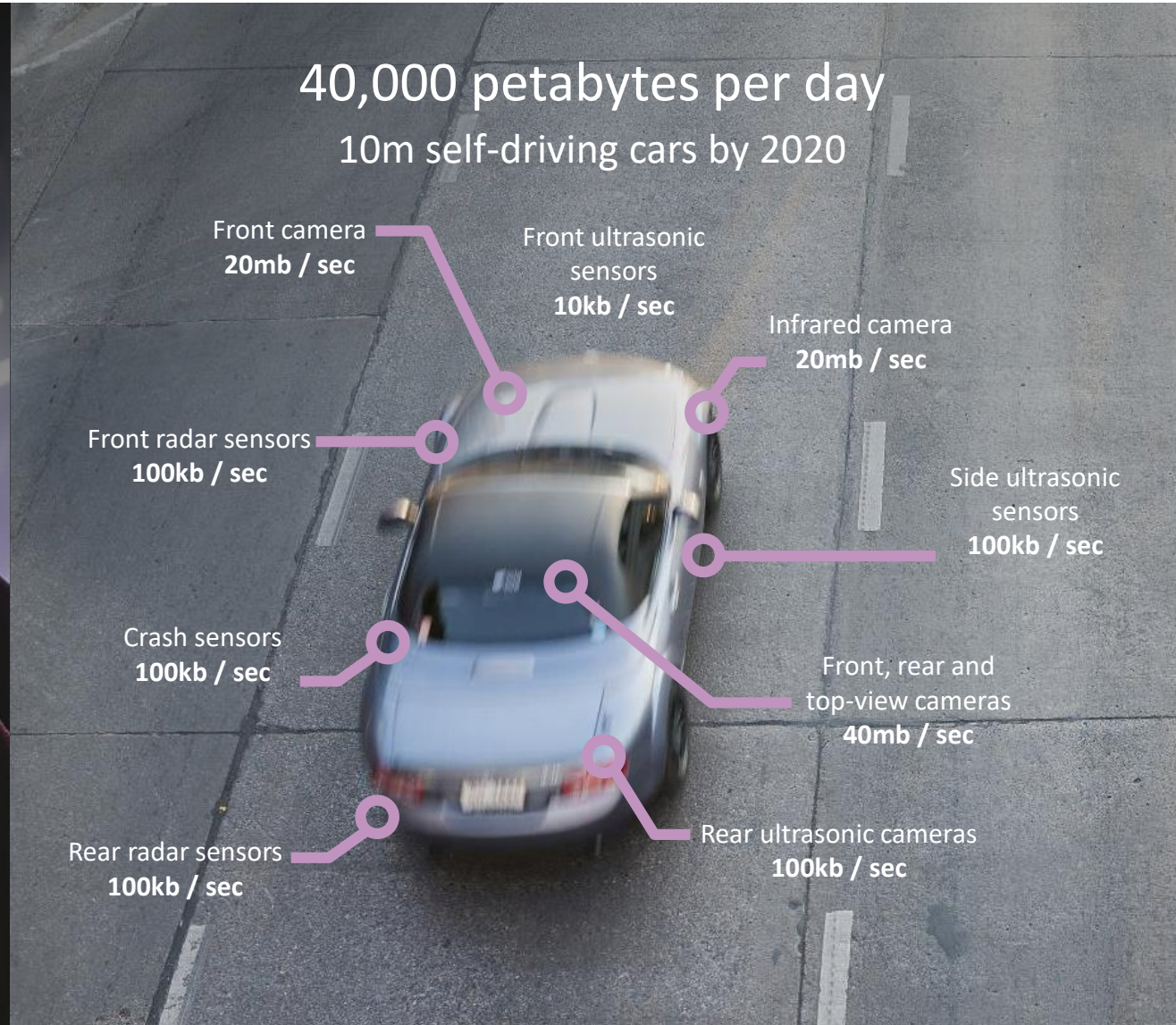
(PWC Study)

1 Trillion sensors deployed globally by 2020

(IBM estimate. IDC Estimate: 17.6 Billion)

The amount of data we process is about to change — again

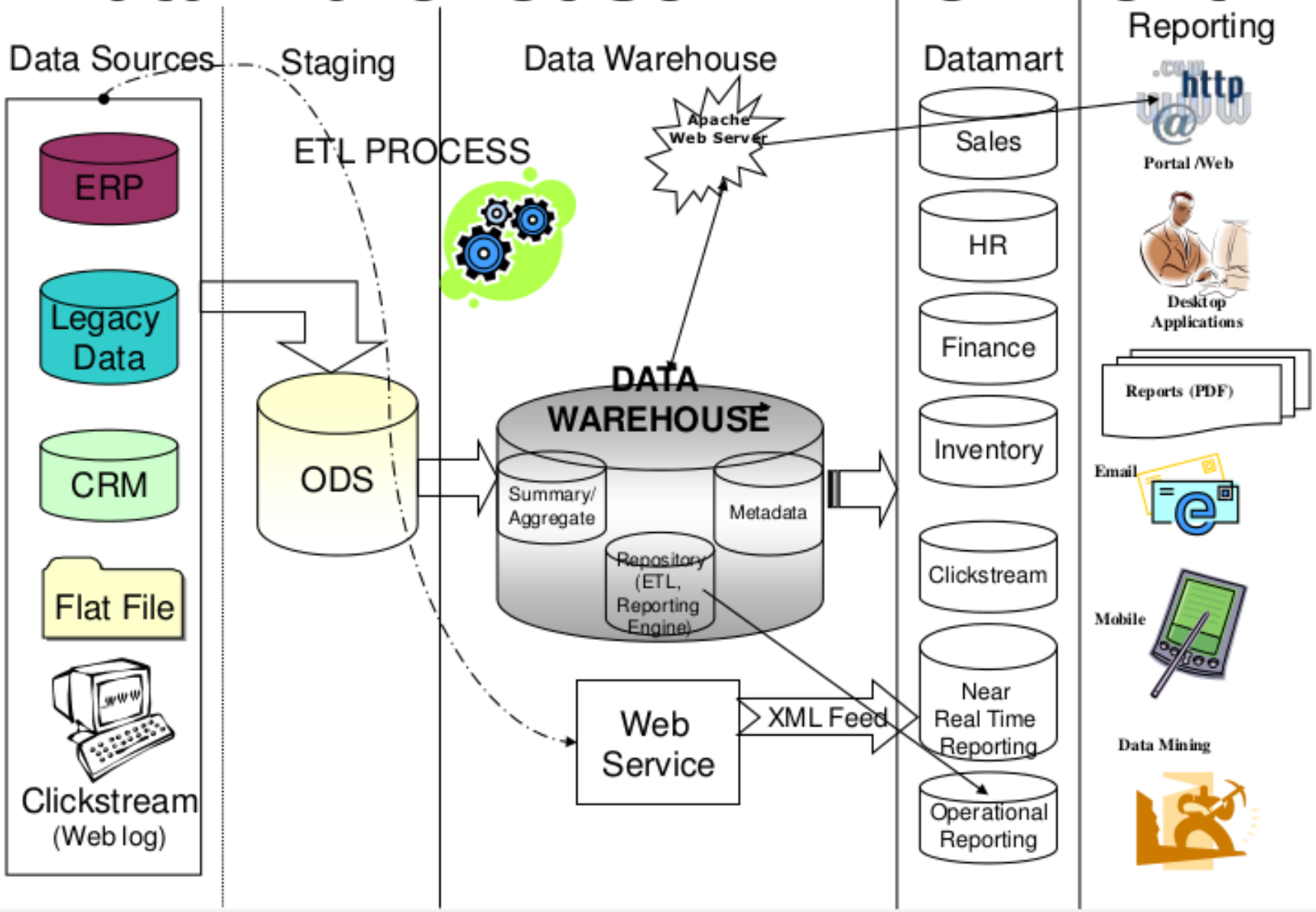




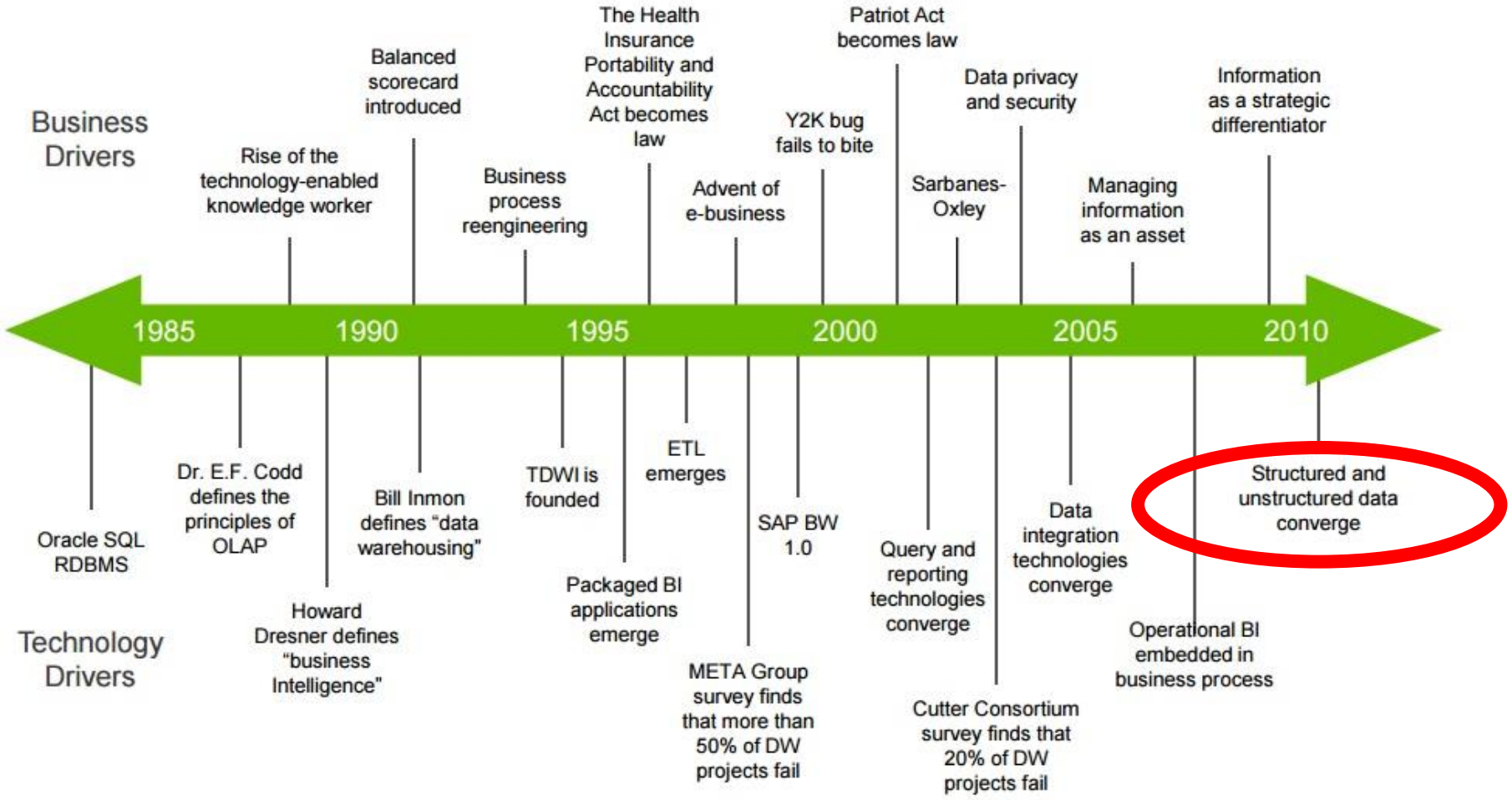
“Often stepping back
you see more,
don’t you?”

- *David Hockney*

Data Warehouse Environment

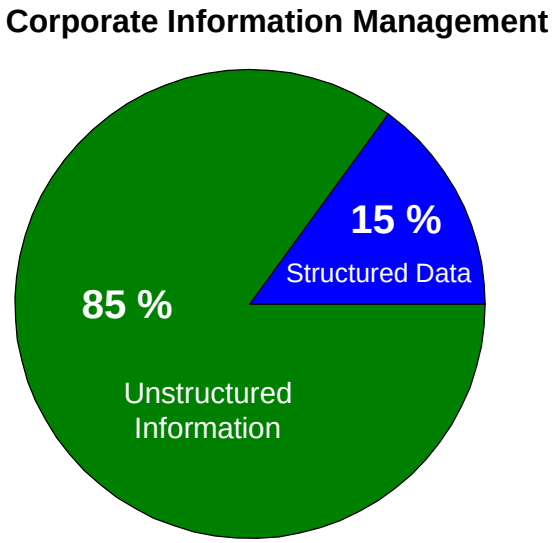


Data Warehousing



The Opportunity – Unstructured Data

Suddenly



- Information and knowledge management is - the management of both structured data (15% of information) and unstructured data (85% of information), according to the Butler Group.
- 80 percent of business is conducted on unstructured information (Gartner Group).
- Over 43 percent of organizations have more than six content stores. (Forrester Research).
- Billions of dollars have been invested - using technologies such as data mining, business intelligence and on-line analytical processing tools to extract value from databases.
- The unstructured data market is a significant opportunity - estimated at least 4 times the size of the structured data market and growing exponentially

Need for external data – structured, semi-structured and unstructured

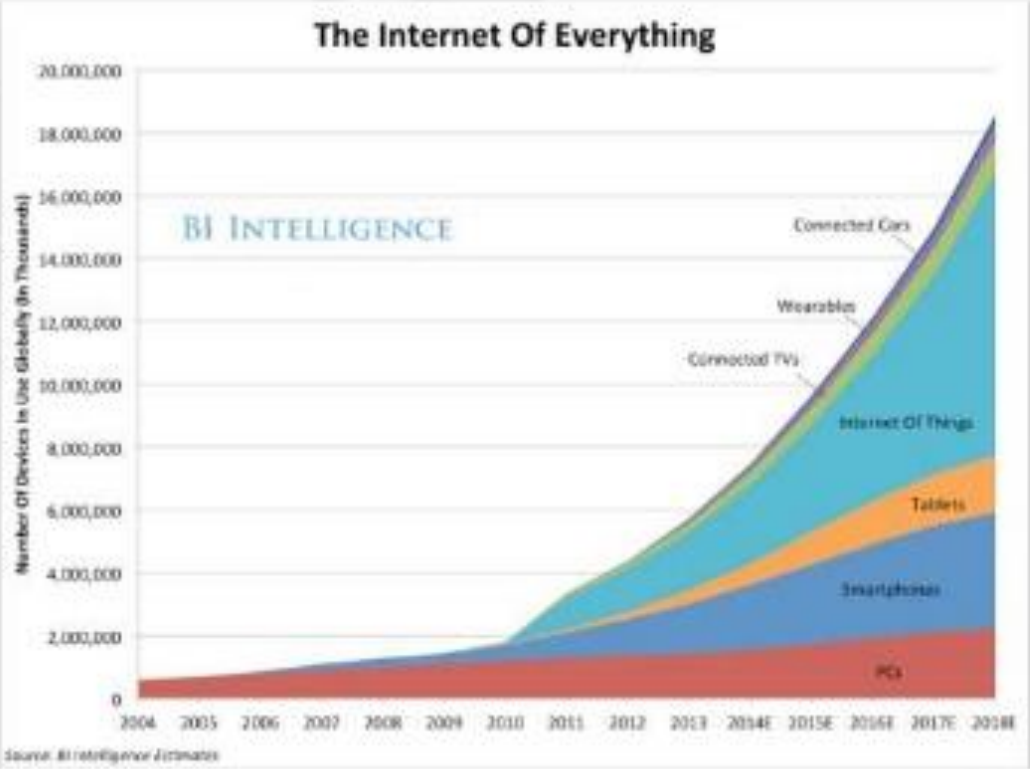
THEN

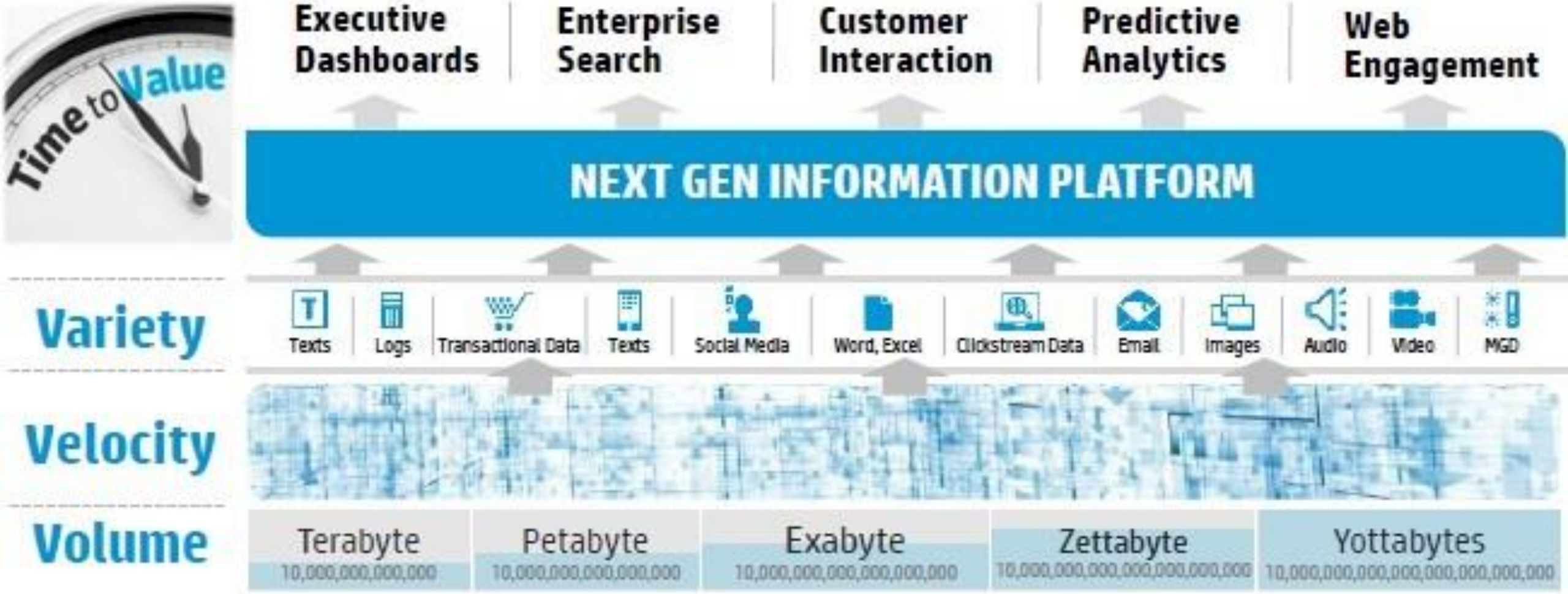
More data structured, hyper-structured and who knows…

THEN

IoT Trend

We're in the midst of a perfect storm with several inflection points contributing to an impending massive growth (e.g. low cost sensors and RF, mobile penetration, strides in cloud and big data analytics.) IoT is where the Web was 15 years ago. This is just the beginning.

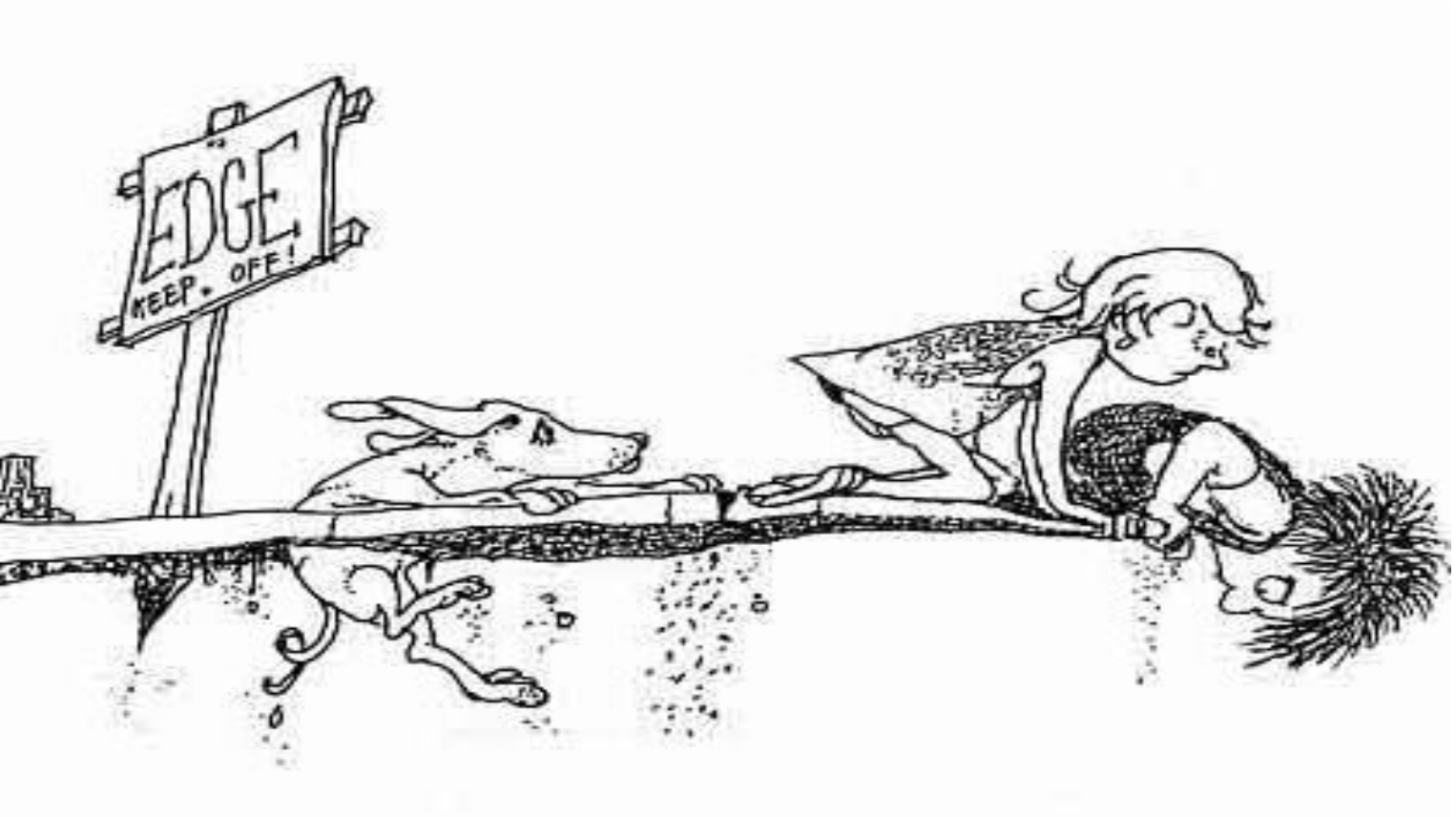



Time to Value
Executive Dashboards
Enterprise Search
Customer Interaction
Predictive Analytics
Web Engagement
NEXT GEN INFORMATION PLATFORM
Variety
Texts
Logs
Transactional Data
Texts
Social Media
Word, Excel
Clickstream Data
Email
Images
Audio
Video
MGD
Velocity
Volume
Terabyte
10,000,000,000,000
Petabyte
10,000,000,000,000,000
Exabyte
10,000,000,000,000,000,000
Zettabyte
10,000,000,000,000,000,000,000
Yottabytes
10,000,000,000,000,000,000,000,000

Houston we have a problem…

- Traditional RDMS/EDW cannot handle
  - New data types (**Variety**)
  - Extended analytical and stream processing (**Value**)
  - TB/hour loading with immediate Query access (**Velocity**)
- RDBS/EDW storage constraints (**Volume**)

Where the Data Warehouse ends…



But not the end of the Data Warehouse!!!

hadoop vs EDW
And

- Hadoop is an open source distributed storage and processing framework
- Architectural differences

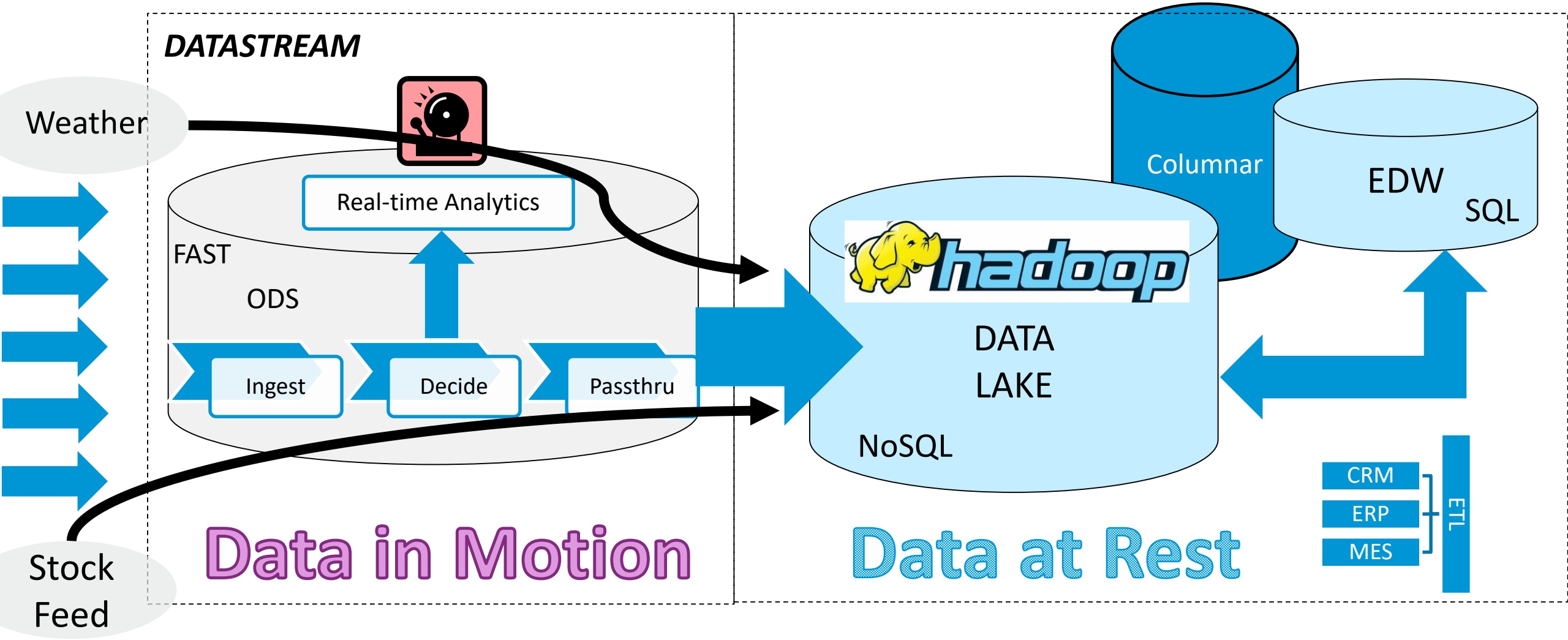
Database - Tables	**STORAGE**	**HDFS Files (raw to highly structured)**
System Tables	**METADATA**	Hcatalog (multiple possible descriptions)
SQL Query Engine	**QUERY**	Multiple Engines (SQL, NoSQL and non SQL)

All three layers provided as proprietary bundle

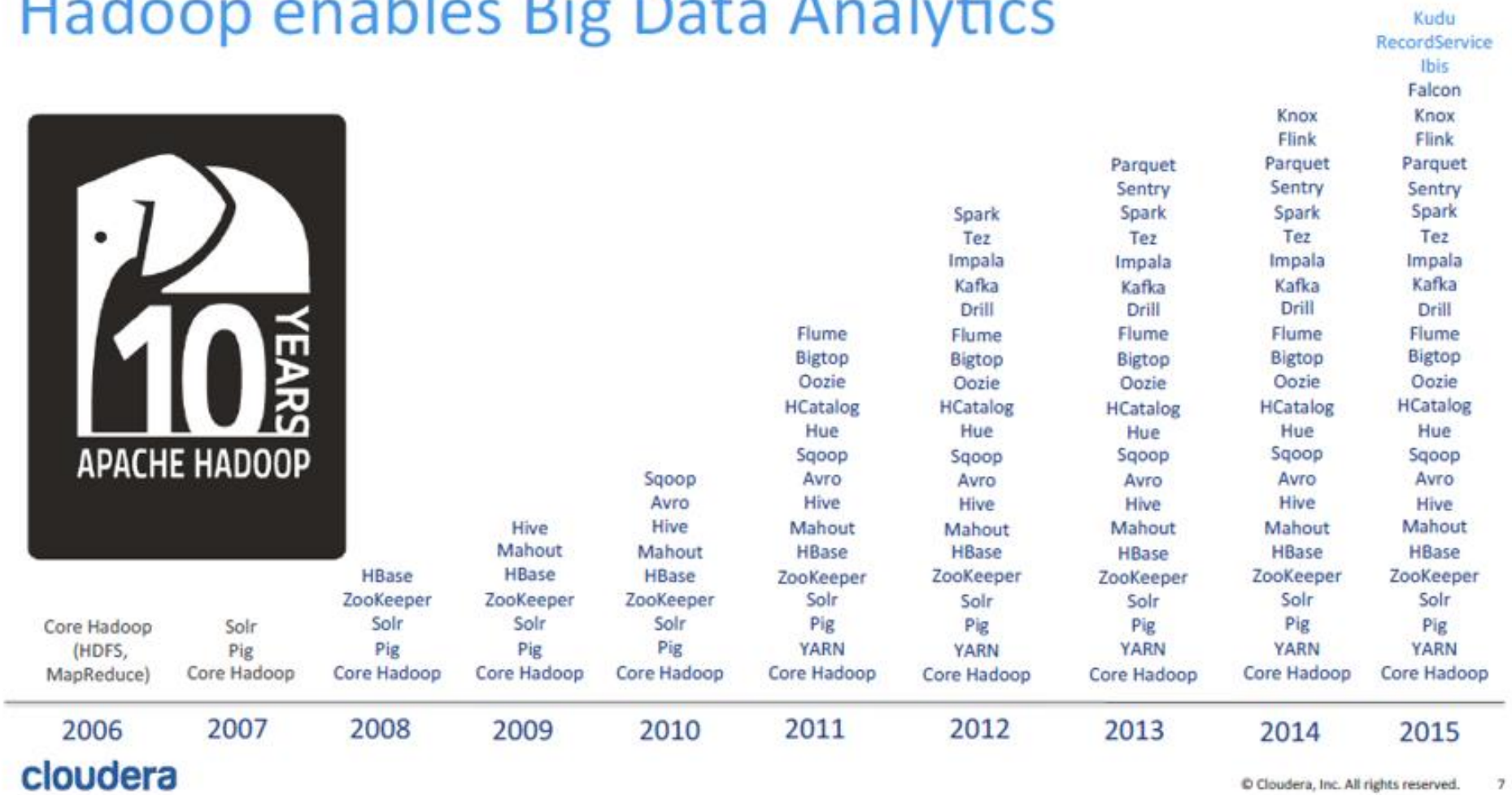
All three layers separate and independent allowing multiple access

Streams to Lake

BIG DATA

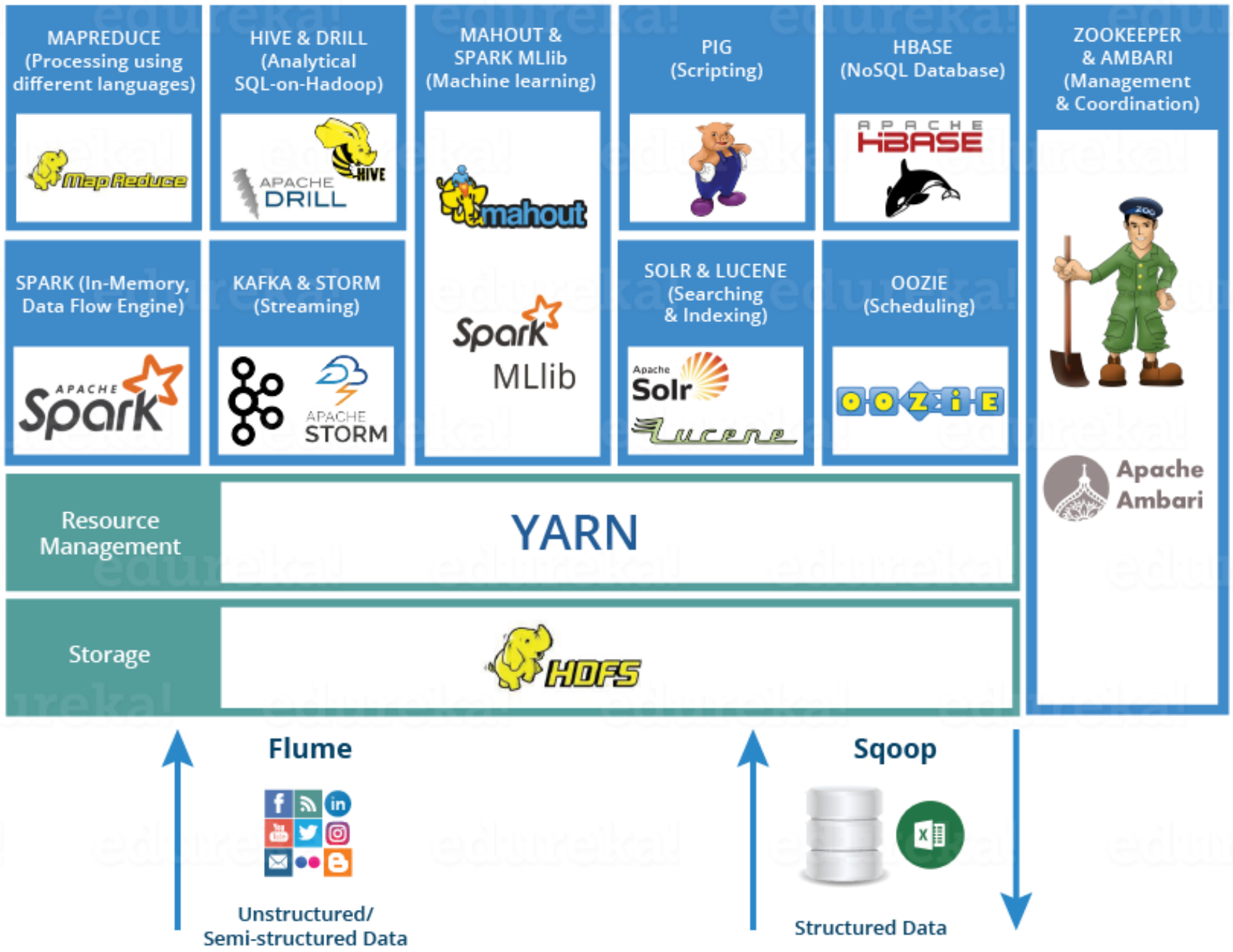


Hadoop enables Big Data Analytics

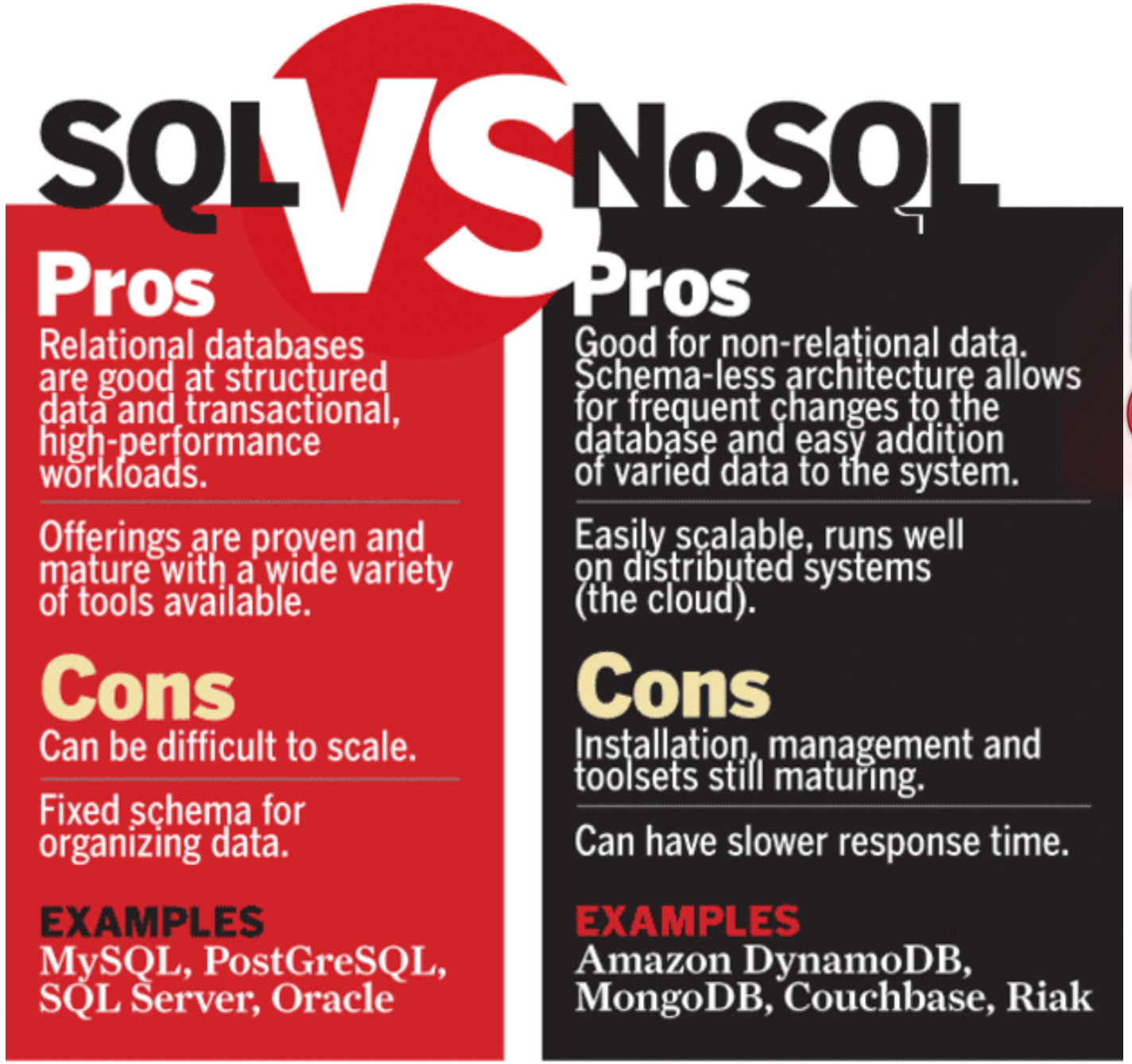


cloudera

 7


MAPREDUCE (Processing using different languages)
HIVE & DRILL (Analytical SQL-on-Hadoop)
MAHOUT & SPARK MLlib (Machine learning)
PIG (Scripting)
HBASE (NoSQL Database)
ZOOKEEPER & AMBARI (Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
KAFKA & STORM (Streaming)
SOLR & LUCENE (Searching & Indexing)
OOZIE (Scheduling)
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/ Semi-structured Data
Sqoop
Structured Data

ACID
- Atomicity
- Consistency
- Isolation
- Durability

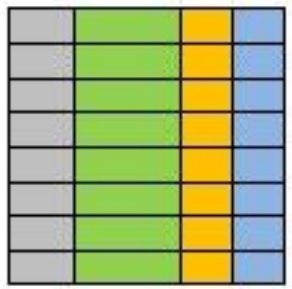


Not Only SQL

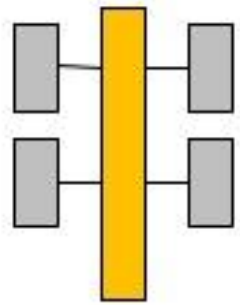
BaSE
- Basic availability
- Soft state
- Eventually consistent

NoSQL Database Patterns

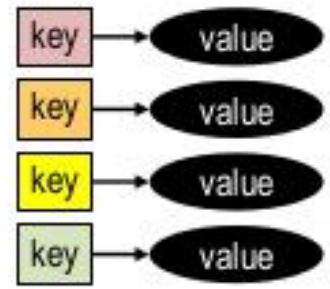
Relational



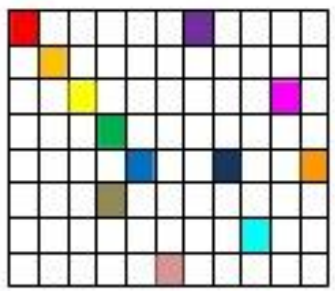
Analytical (OLAP)



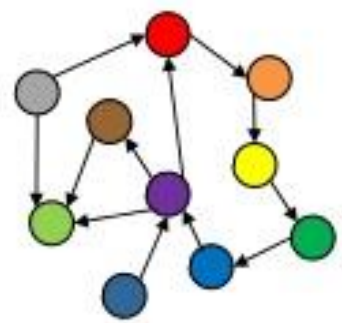
Key-Value



Column-Family

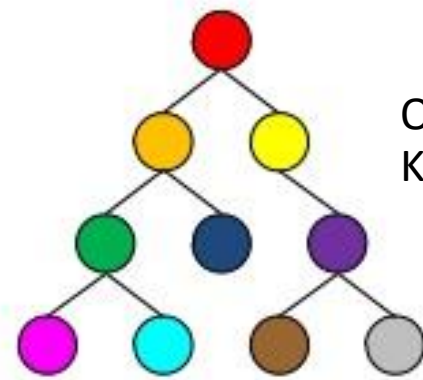


Graph



Node/relationship

Document

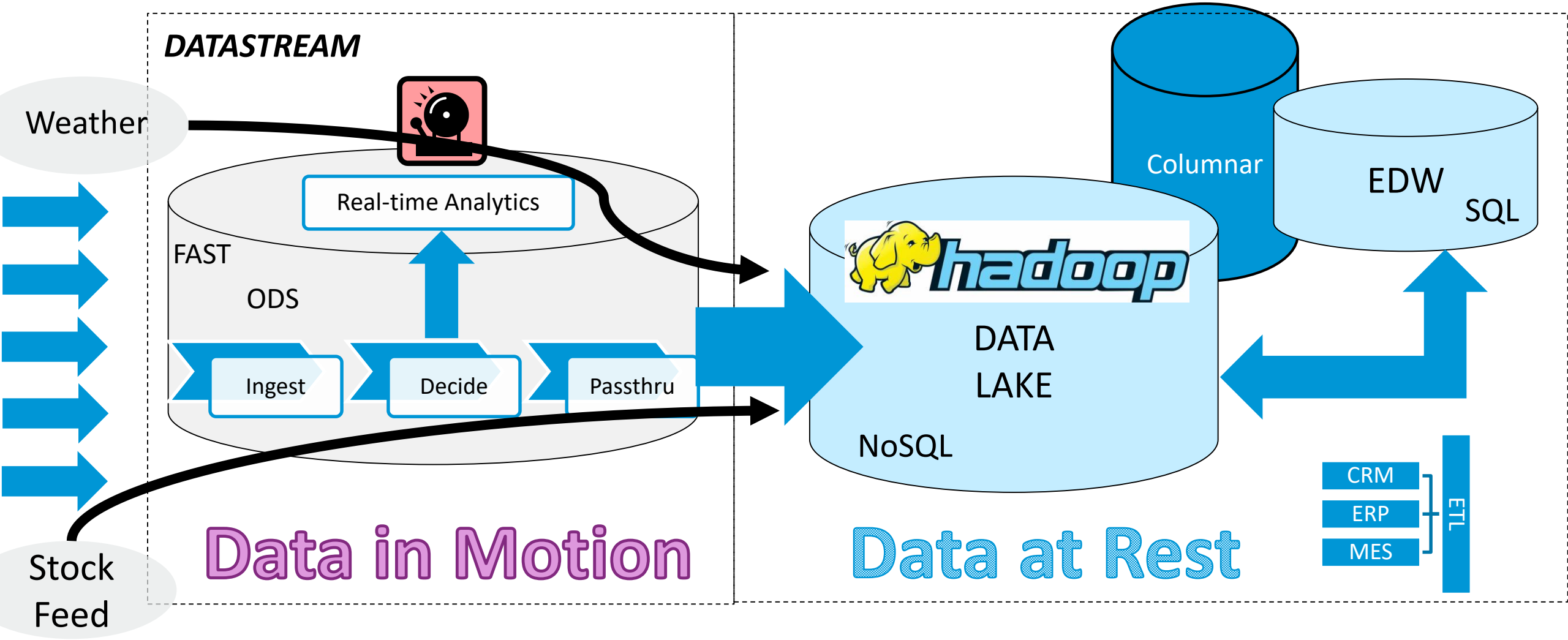


Organized
Key-Value

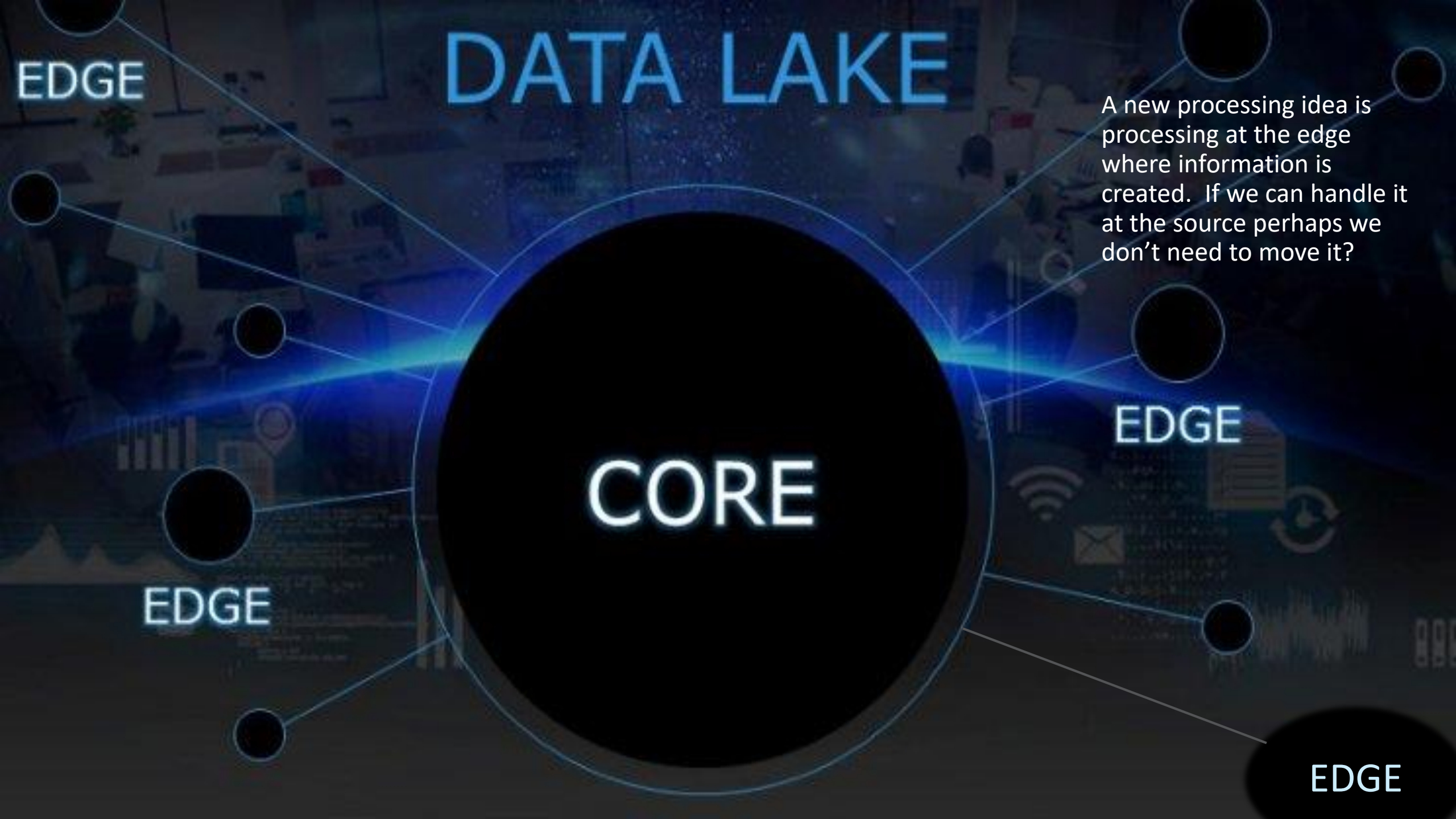
Type	Example
Key-Value Store	redis, riak
Wide Column Store	HBASE, cassandra
Document Store	mongoDB, CouchDB relax
Graph Store	Neo4j, InfiniteGraph The Distributed Graph Database

Streams to Lake

BIG DATA



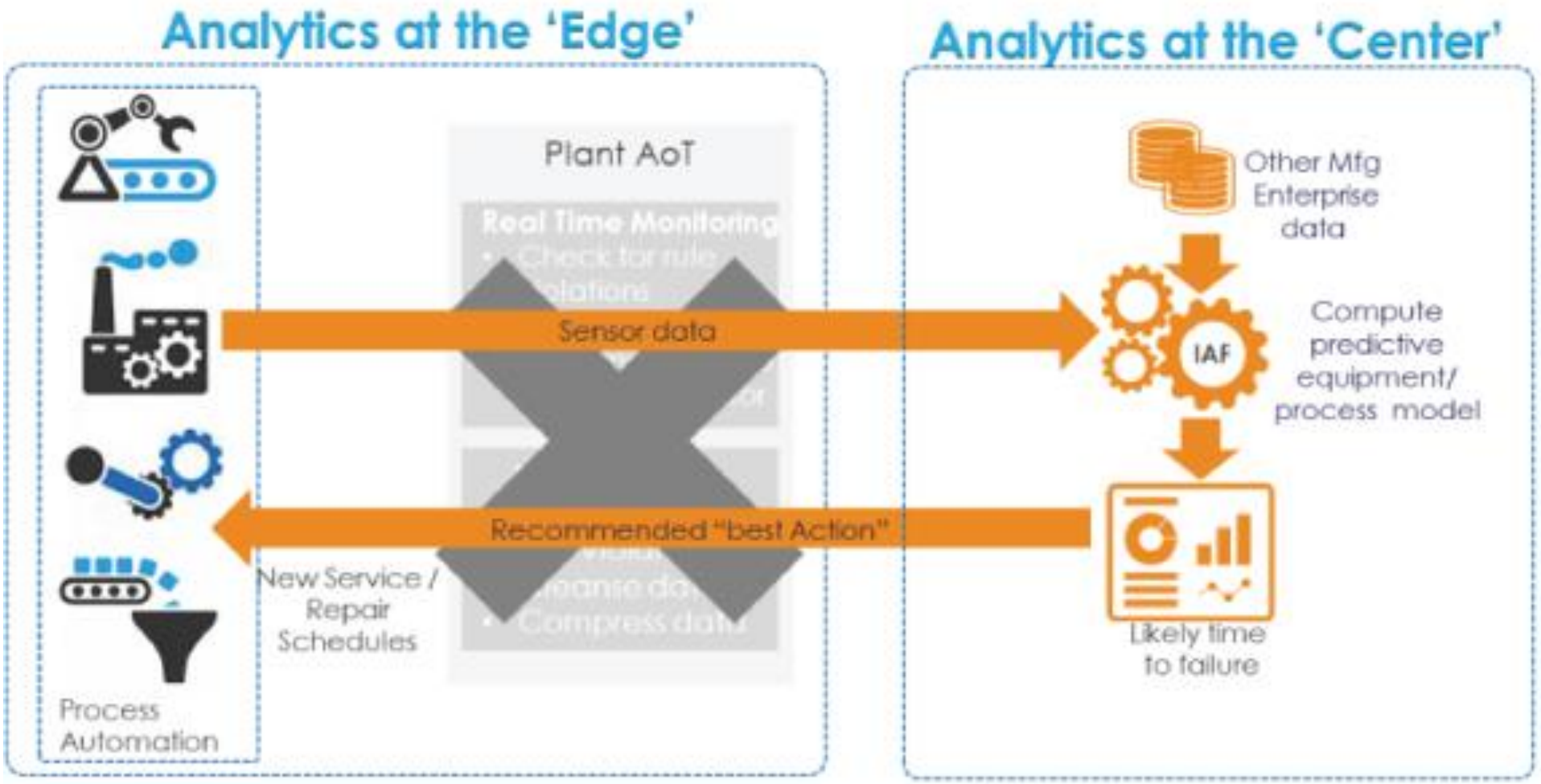
Can we continue to move all this data?



A new processing idea is processing at the edge where information is created. If we can handle it at the source perhaps we don’t need to move it?

The need for speed

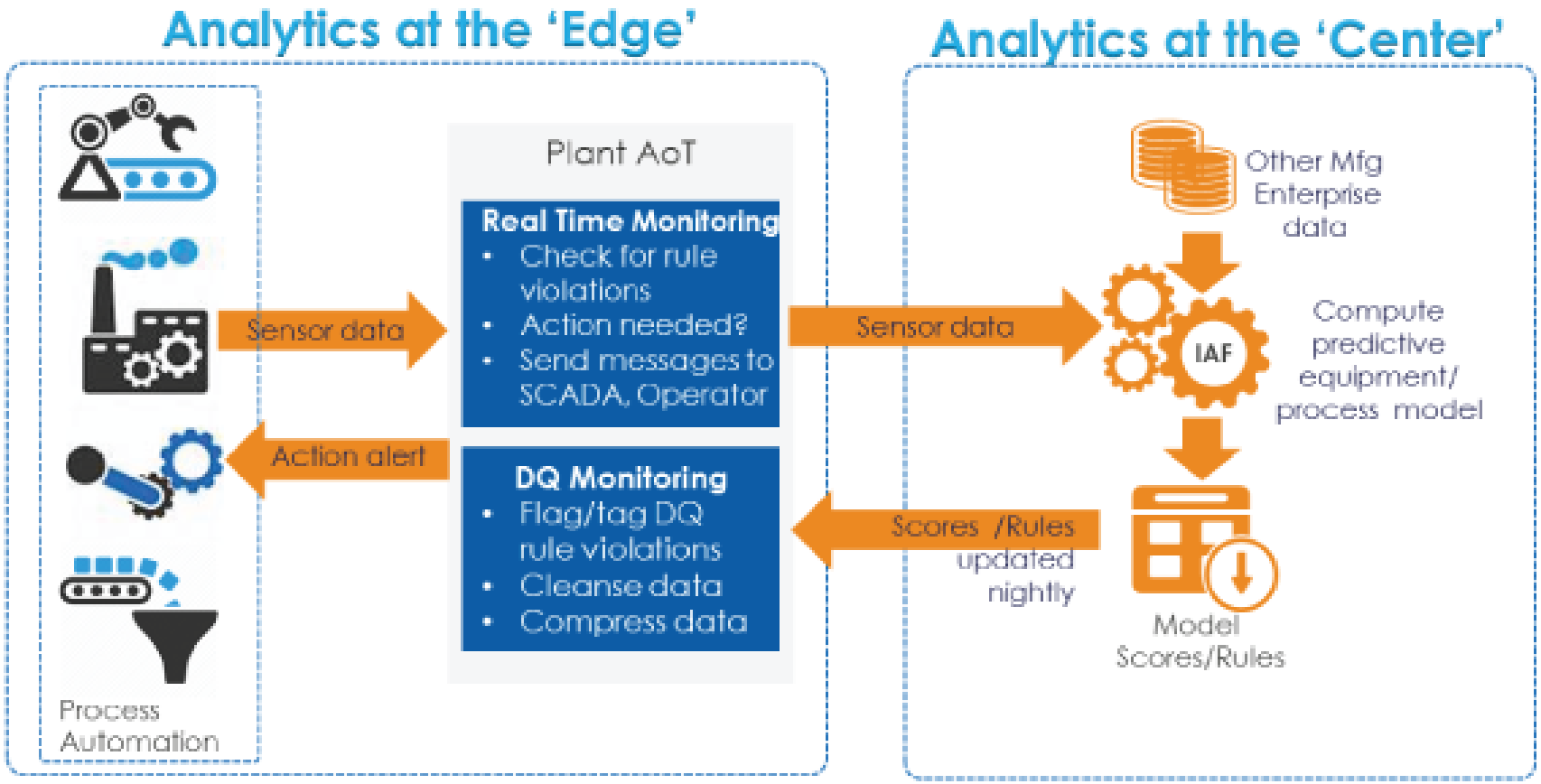
So initially we may require all data from shop floor/supply chain systems to be moved to the data lake where analytics are run against it. The analytics might uncover patterns indicating the requirement for maintenance or possible outage.



Sensor data reveals wear-out trend

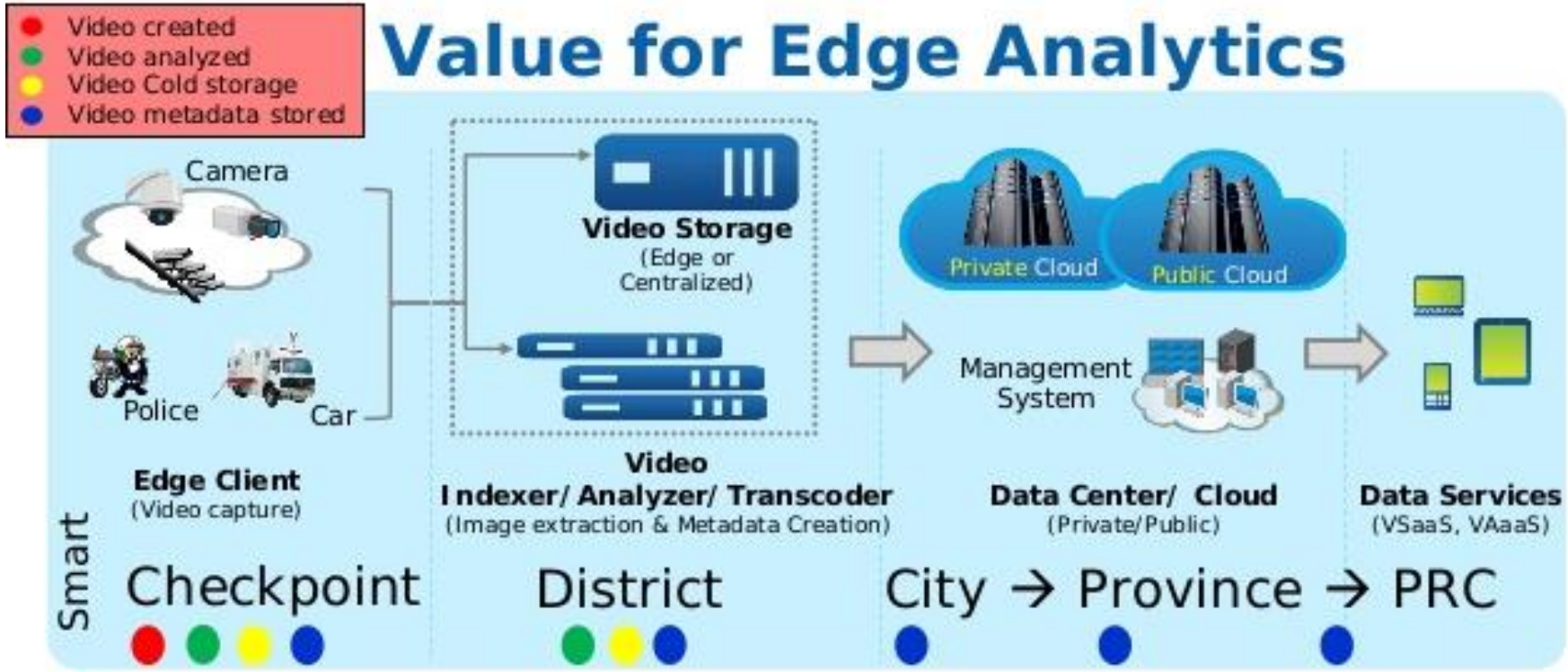
Reducing Big Data movement and latency

Those patterns can be exported to local, smaller systems and the information processed locally without the need to move all the data just the information that cannot be processed locally.



A local sensor detects anomaly

Value for Edge Analytics



1 **By end of 2017 457 PB raw video for traffic generated per day**

Edge VA's Value

- Real-time intelligence (into metadata) 1/6 of video data
- Reduce the footage to be 1/8 ~ 1/12 of its original size
- Resolve the bandwidth issue and backend storage capacity constrain

By end of 2017

76 PB Metadata Per Day

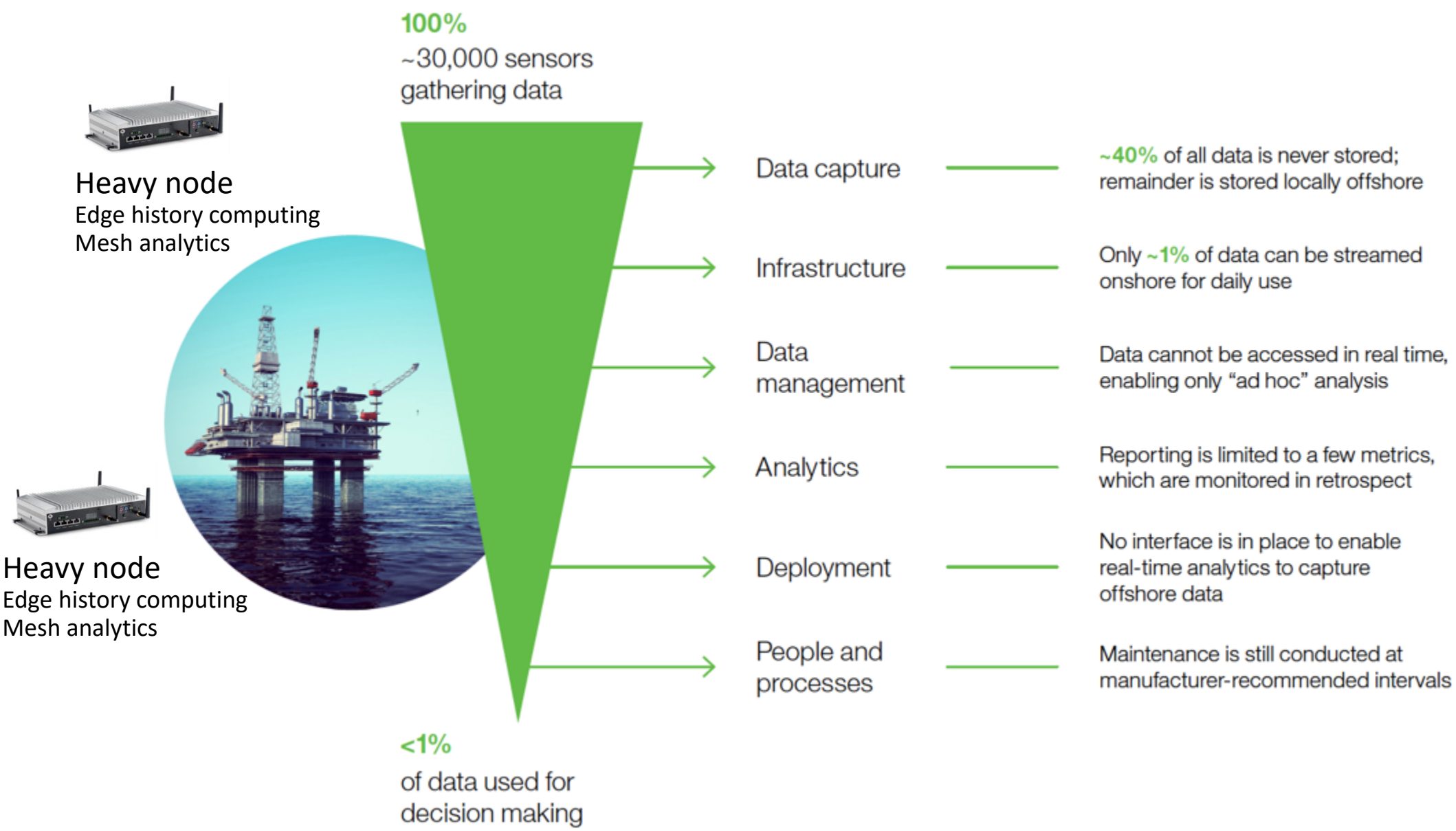
People, cars, license plates

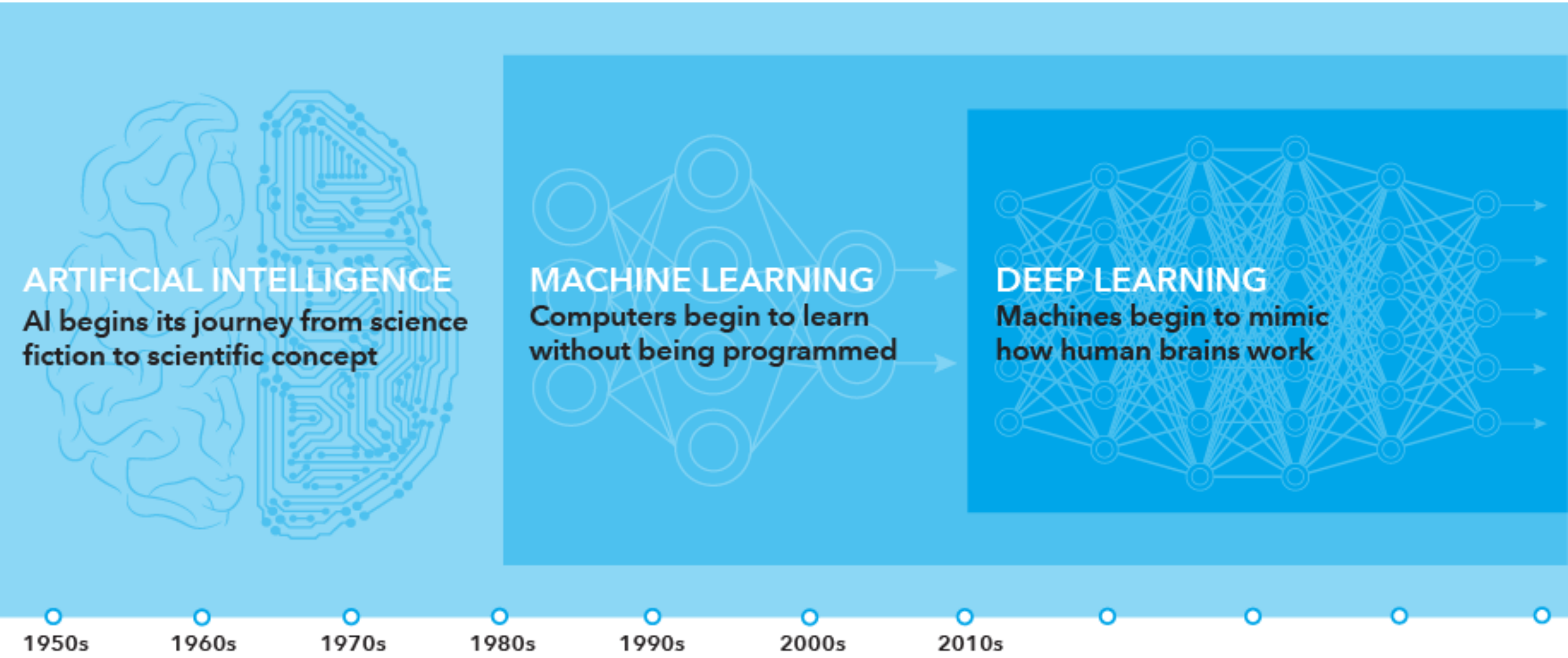
1 Source: Internal Team Analysis based on IHS* Research Report

Enhanced NVR (Network Video Recorder) Key Features for Intelligent Transportation System (ITS)

闯红灯/ / Run the red light		车牌颜色识别 // Plate colour recognition
逆行 / / Retrogradation		车身颜色识别 // Vehicle colour recognition
车牌识别 / / Plate Recognition		交通拥堵 // Traffic jam
车流统计/ / Vehicle counting		行使缓慢 // Run slowly
禁停/ / No parking		行使超速 // Speeding
禁左禁右 // No turn right or left		行人横穿//Jaywalk
占道(不按规定车道行使)		车标识别//Auto logo recognition
变道 / / Lane Change		机动车抓拍 / / Abnormality quick shot
压线 / / Line crossing		超速 / / Speeding

Vehicle features recognition


100%
~30,000 sensors gathering data
Heavy node
Edge history computing
Mesh analytics
Heavy node
Edge history computing
Mesh analytics
Data capture
~40% of all data is never stored; remainder is stored locally offshore
Infrastructure
Only ~1% of data can be streamed onshore for daily use
Data management
Data cannot be accessed in real time, enabling only "ad hoc" analysis
Analytics
Reporting is limited to a few metrics, which are monitored in retrospect
Deployment
No interface is in place to enable real-time analytics to capture offshore data
People and processes
Maintenance is still conducted at manufacturer-recommended intervals
<1%
of data used for decision making


ARTIFICIAL INTELLIGENCE
AI begins its journey from science fiction to scientific concept
MACHINE LEARNING
Computers begin to learn without being programmed
DEEP LEARNING
Machines begin to mimic how human brains work
1950s
1960s
1970s
1980s
1990s
2000s
2010s

“Machine learning is basically a way for a computer to find the nuggets of information that a human can’t,” explains Fausto Ibarra, director of global product management for Google Cloud Platform. “Once you have your data and train and deploy your models, the machine can go through terabytes of data and get smarter and smarter—basically train itself—and ultimately make predictions for you.”

Machine Learning – Deep Learning

- IBM Project Deep Blue
- Chess playing computer system
- May 1997 defeats Garry Kasparov – current world champion

Machine Learning – Deep Learning

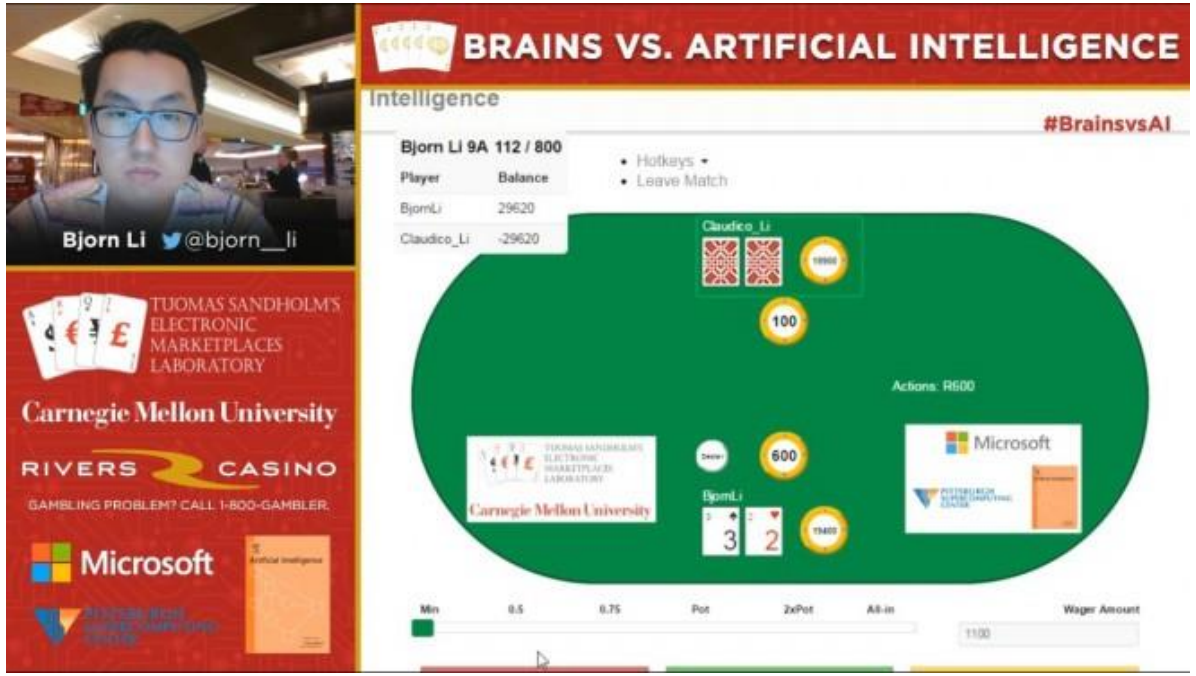
- IBM Project Deep QA aka “Watson”
- Developed NLP (Natural Language Processing) capabilities
- 2011 defeats the top Jeopardy champions Brad Rutter and Ken Jennings

Machine Learning – Deep Learning

- Google DeepMind aka “AlphaGo”
- Developed to play Go, more variations than Chess
- March 2016 AlphaGo defeats Lee Sedol a 9th Dan Go champion
- Jan 2017 AlphaGo defeats Ke Jie the #1 ranked Go player in the world in 3 straight games

Machine Learning – Deep Learning

- Carnegie Mellon University “Libratus”
- No-limit Texas Hold’em Poker
- January 2017 beats Jason Les, Dong Kim, Daniel McAulay and Jimmy Chou



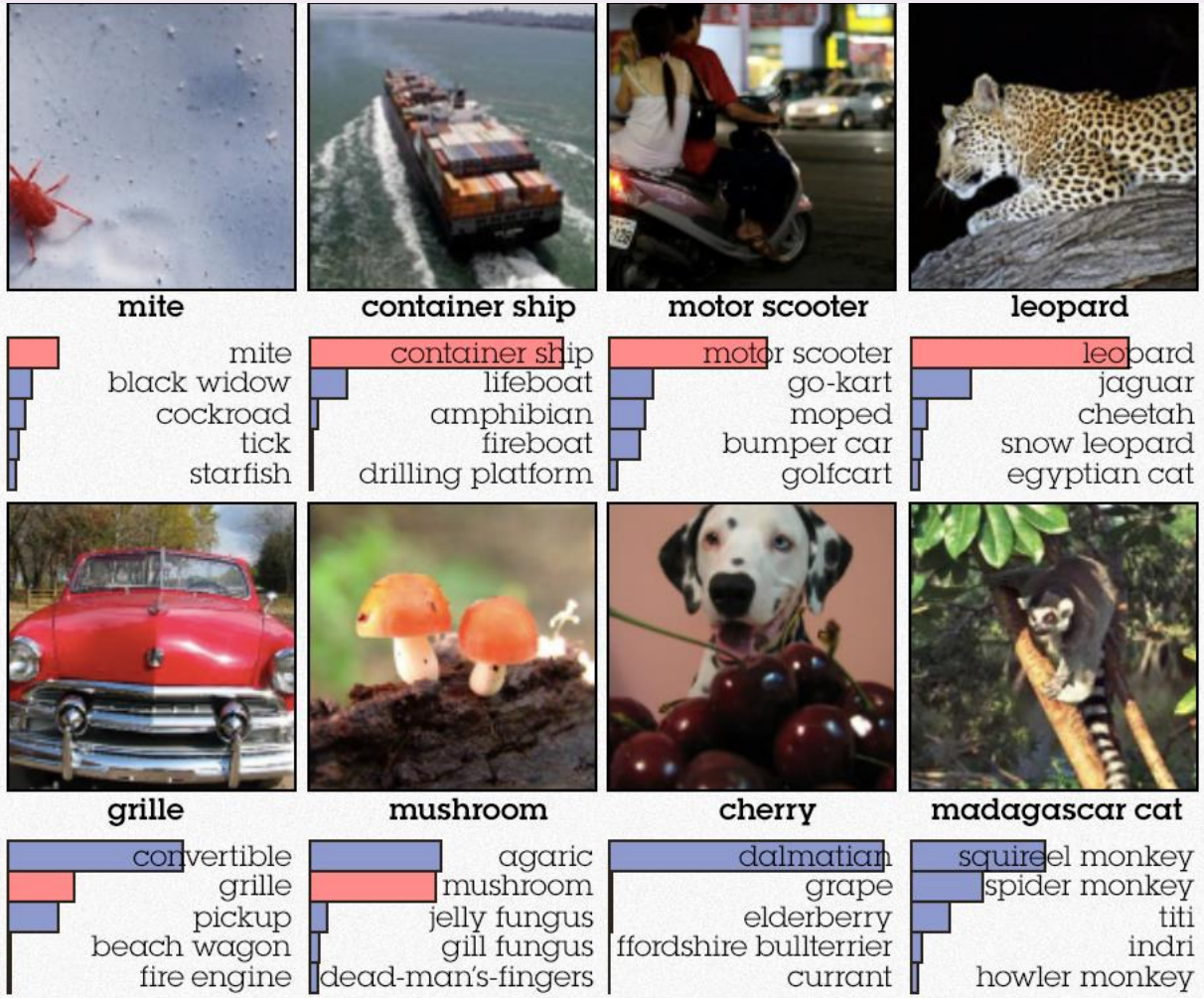
Placeholder for Watson Video

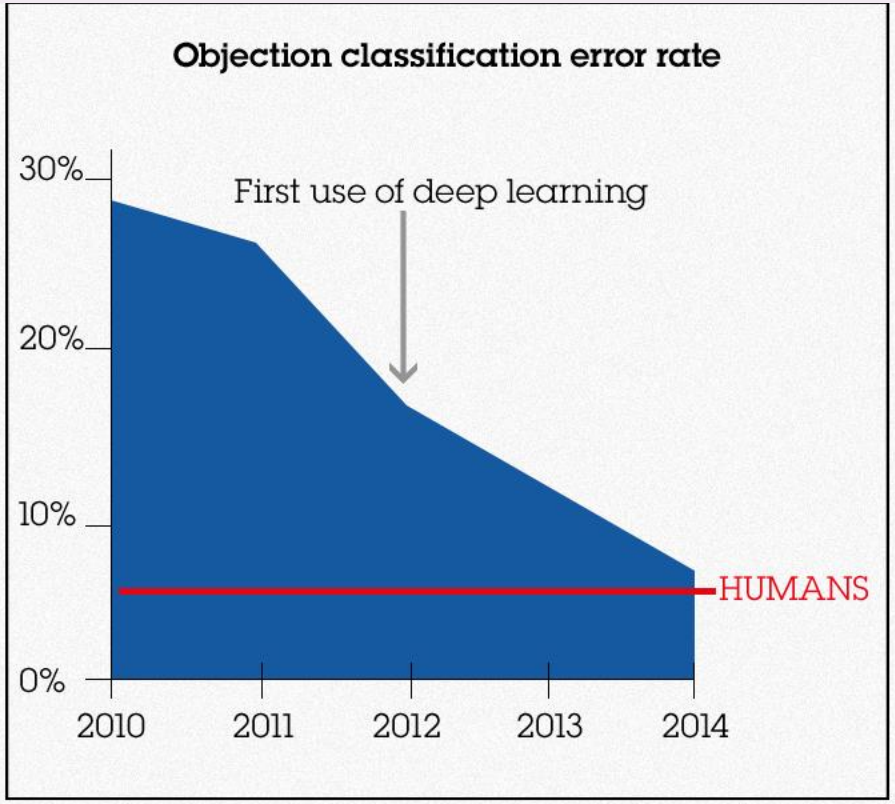
https://www.youtube.com/watch?v=HkEOJnn_zlg

"There's a 60-year-old woman in Tokyo. She was at the University of Tokyo. She had been diagnosed with leukemia six years ago. She was living, but not healthy. So the University of Tokyo ran her genomic sequence through Watson and it was able to ascertain that they were off by one thing. Actually, she had two strains of leukemia. They did treat her and she is healthy."

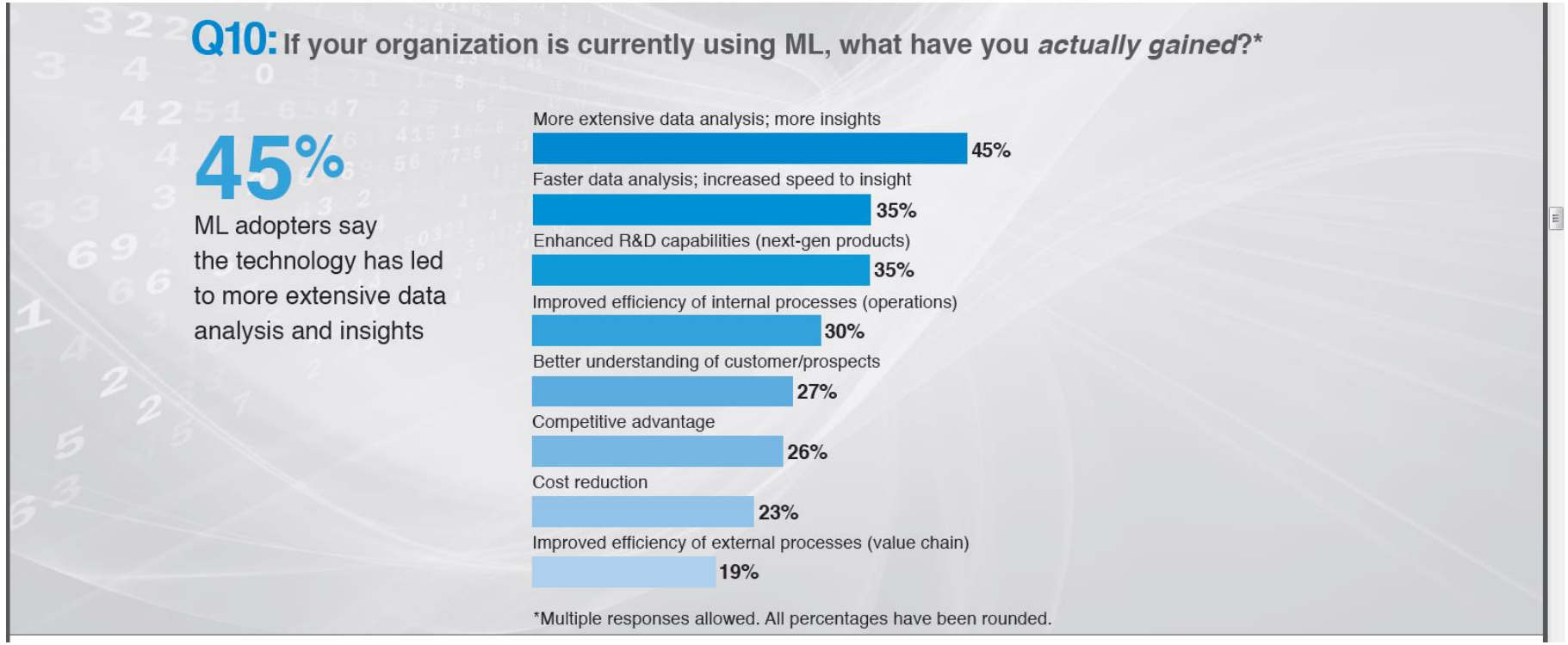
David Kenny, General Manager of IBM Watson 12/7/2016

Cataloging pictures

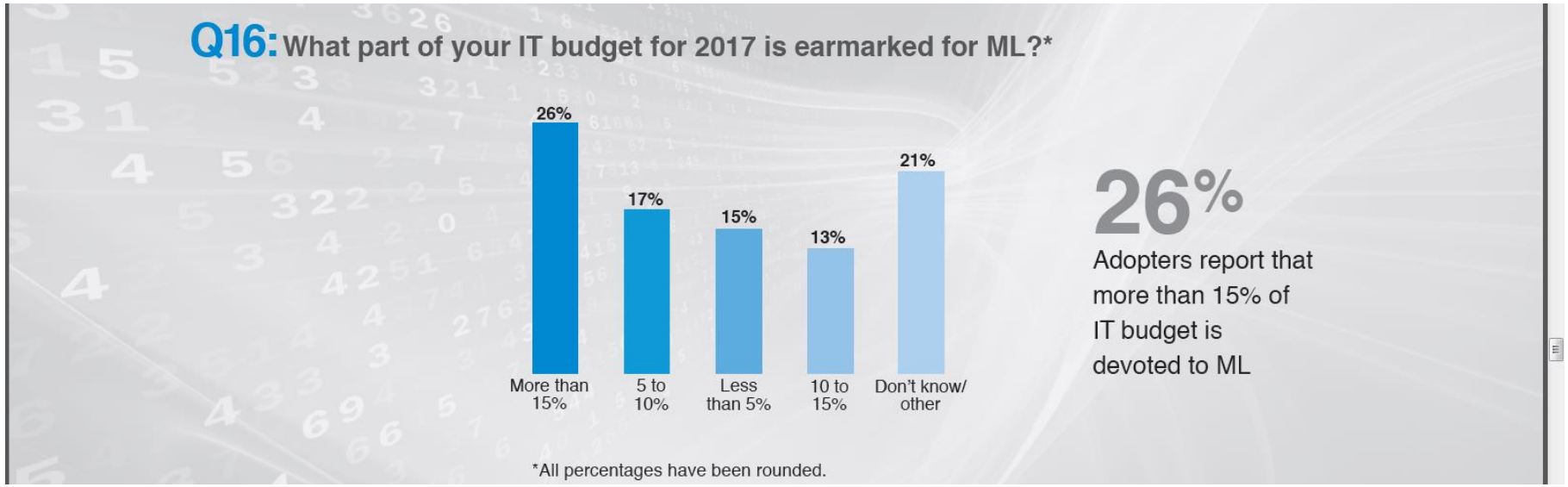




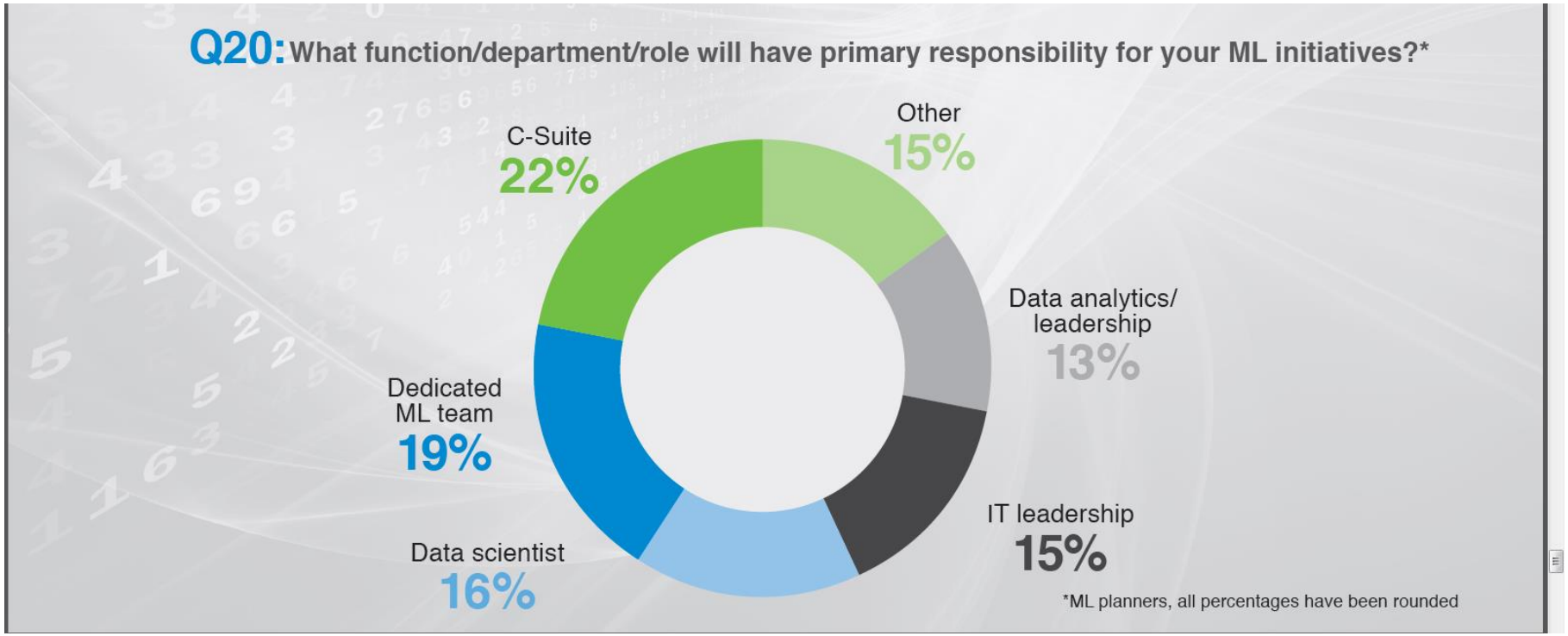
Benefits


Q10: If your organization is currently using ML, what have you actually gained?*
45%
ML adopters say the technology has led to more extensive data analysis and insights
More extensive data analysis; more insights
45%
Faster data analysis; increased speed to insight
35%
Enhanced R&D capabilities (next-gen products)
35%
Improved efficiency of internal processes (operations)
30%
Better understanding of customer/prospects
27%
Competitive advantage
26%
Cost reduction
23%
Improved efficiency of external processes (value chain)
19%
*Multiple responses allowed. All percentages have been rounded.

Budget


Q16: What part of your IT budget for 2017 is earmarked for ML?*
26%
17%
15%
13%
21%
More than 15%
5 to 10%
Less than 5%
10 to 15%
Don't know/ other
*All percentages have been rounded.
26%
Adopters report that more than 15% of IT budget is devoted to ML

C-Suite involved

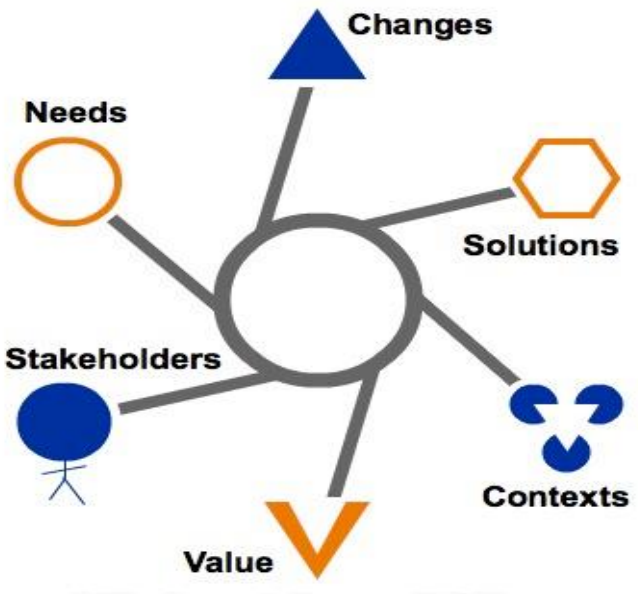


Business Analysts and Business Intelligence

- Analytics have become strategic to the business (Competing on Analytics by Tom Davenport)
- BI transitioning from IT-led system of record reporting to Business-led self-service analytics
  - Accessibility (democratize data)
  - Agility
  - Deeper analytical insights
- Diverse array of data sources
- Expanded range of users

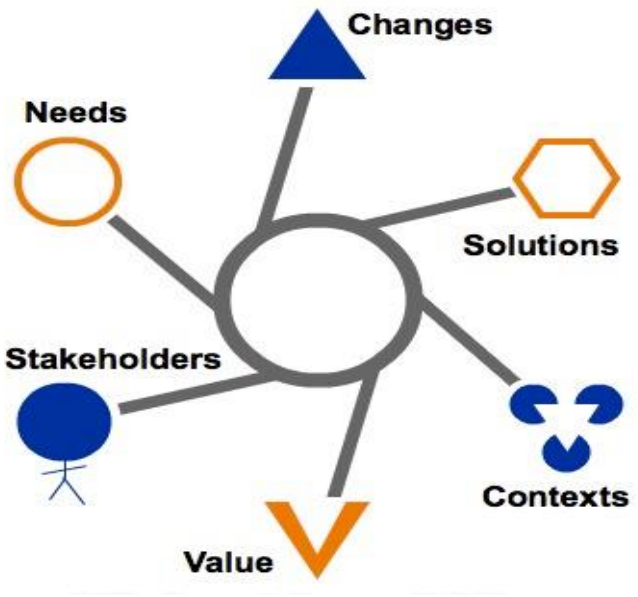
Seven areas where business analysts can help drive BI&A

- BI&A Platform modernization
  - Non-technical users to autonomously create content
- Business value
  - Driving efficiencies and innovation across business domains
- Tool Selection
  - Necessary to maximize adoption of trusted analytics that impact the business
- Organizational change requirements
  - Redefining roles and responsibilities of IT and business users



Seven areas where business analysts can help drive BI&A

- Develop user competence
  - New methods for training and support
  - New rules for governing, managing, promoting and sharing content
- Agile development model
  - Pace of business and exploratory nature of analytics requires agile methodology
- Key technologies and market trends
  - The data lake and the EDW
  - Cloud deployment options
  - Benefits and compromises of open source
  - Large, diverse and real-time streaming data
  - Data Discovery
  - Machine learning
  - Deploying self-service at scale

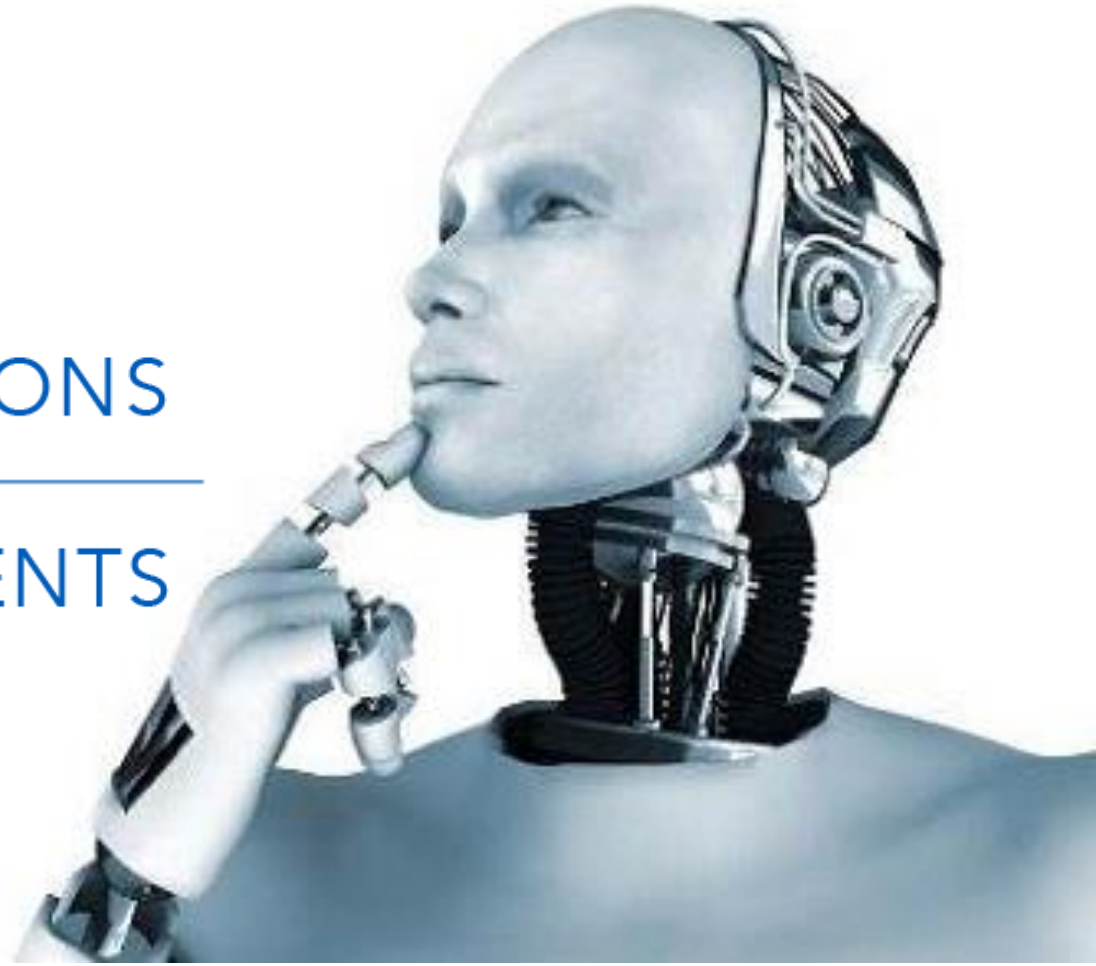


“What the industry is seeing is that more than 50 percent of big data projects fail, and fail miserably,” Poutonnet says. “Businesses are spending a lot of time with the data-gathering and data-preparation phases, as well as trying to figure out the data architecture, and they don’t have time to play with the data itself.”

Philippe Poutonnet, global product marketing lead for Google Cloud Platform

QUESTIONS

COMMENTS



If there’s no questions (unlikely) another Watson Video

https://www.youtube.com/watch?v=ymUFadN_MO4