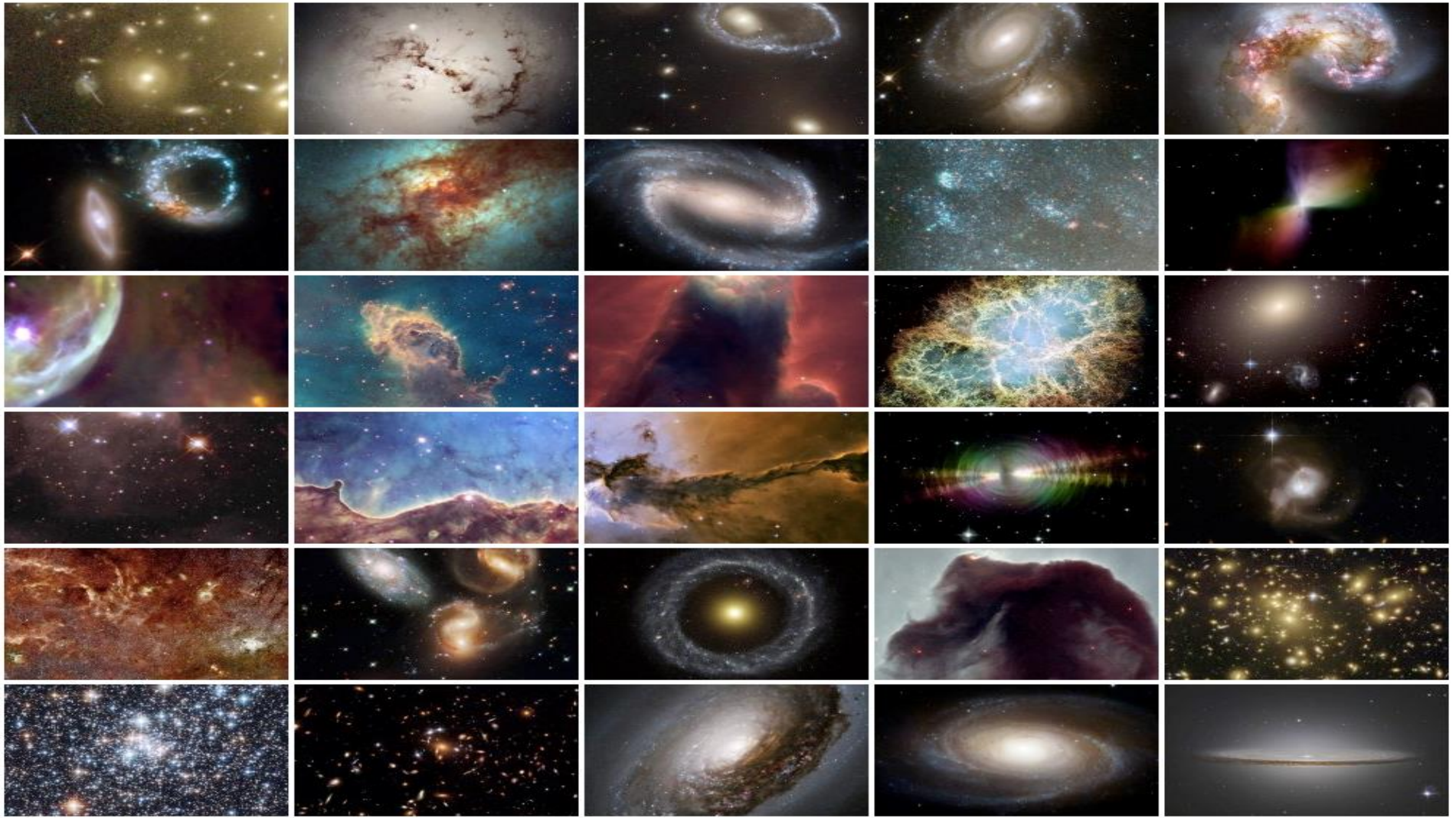
INTRODUCTION TO

BIG DATA MANAGEMENT

Björn Þór Jónsson

CRESS and School of Computer Science,
Reykjavík University


IN 60 SECONDS...
1 NEW DEFINITION IS ADDED ON urbar
1,600+ READS ON Scribd
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
320+ NEW twitter ACCOUNTS
100+ NEW Linked in ACCOUNTS
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!!
6,600+ NEW PICTURES ARE UPLOADED ON flickr
50+ WORDPRESS DOWNLOADS
125+ PLUGIN DOWNLOADS
695,000+ facebook STATUS UPDATES
79,364 WALL POSTS
510,040 COMMENTS
1,700+ Firefox DOWNLOADS
694,445 SEARCH QUERIES
168 MILLION EMAILS ARE SENT
60+ NEW BLOGS
1,500+ BLOG POSTS
70+ DOMAINS REGISTERED
600+ NEW VIDEOS
25+ HOURS TOTAL DURATION
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO! ANSWERS
20,000+ NEW POSTS ON tumblr.
13,000+ iPhone APPLICATIONS DOWNLOADED
THE LARGEST SOCIAL READING PUBLISHING COMPANY
New Craigslist Ads
Google
Google Search
GO-Globe.com



Big Data Analytics: Making Government Data Work

"Big data" comes with many promises, but the data alone is not a silver bullet. True, it holds the potential for extracting business or mission intelligence and improving decision-making, but without the application of expert domain knowledge to give data contextual meaning, big data is nothing but a whole lot of dark figures.


CSC

A Briefing from GBC: Industry Insights
November 2012

Making Big Data Work for Government

"Big data" comes with many promises. Not the least of these is access to previously unaccounted information, "dark data," and large volumes of unstructured information. Many estimates place the amount of unstructured data worldwide at 90 percent of all data, and as the average federal agency stores 1.4 petabytes of data,[1] this represents a potentially vast store of previously untapped information for analysts.[2] But big data alone is not a silver bullet. True, it holds the potential for extracting business or mission intelligence and improving decision-making, but without the application of expert domain knowledge to give data contextual meaning, big data will be nothing but a whole lot of dark figures.

Currently, federal agencies cannot make use of all their data because they do not (or cannot afford to) employ enough data scientists – that is, experts who possess domain knowledge and can use big data technologies to ask the right questions and extract business or mission intelligence from vast pools of data. Making use of big data under these circumstances presents a unique challenge.

"Gartner estimates that big data initiatives in 2013 will total $34 billion," and big data will drive $232 billion in spending over the next five years."

Subduing Big Data

The technological challenges of capturing, securing and managing the worldwide explosion of data are not insignificant (the amount of worldwide data currently measures about 2.7 zettabytes and is projected to double every two years[3]). Yet, correlating big data is as much an organizational challenge as a technological one. Industry development of big data technologies has largely bifurcated into two main functions: data discovery and the more traditional business intelligence.[4] Companies have been investing much time and money in the former.[5] Gartner estimates that big data initiatives in 2013 will total $34 billion,[6] and big data will drive $232 billion in spending over the next five years.[7]

Many organizations are trying to address the technology challenge of analyzing big data. The TechAmerica Foundation's Big Data Commission, for example, recently released a report that gives organizations guidelines for an effective big data program. Examination of these steps shows a decided focus not just on technical capabilities, but on organizational policies as well. The Commission recommends that CIOs take a "holistic approach to help guide the agency from an information management perspective and follow these key steps:

- Identify data and content that are critical to its mission
- Identify how, when, where and to whom information should be made available
- Determine appropriate data management, governance and security practices
- Identify and prioritize the information projects that deliver the most value."[8]

CSC

Industry Insights

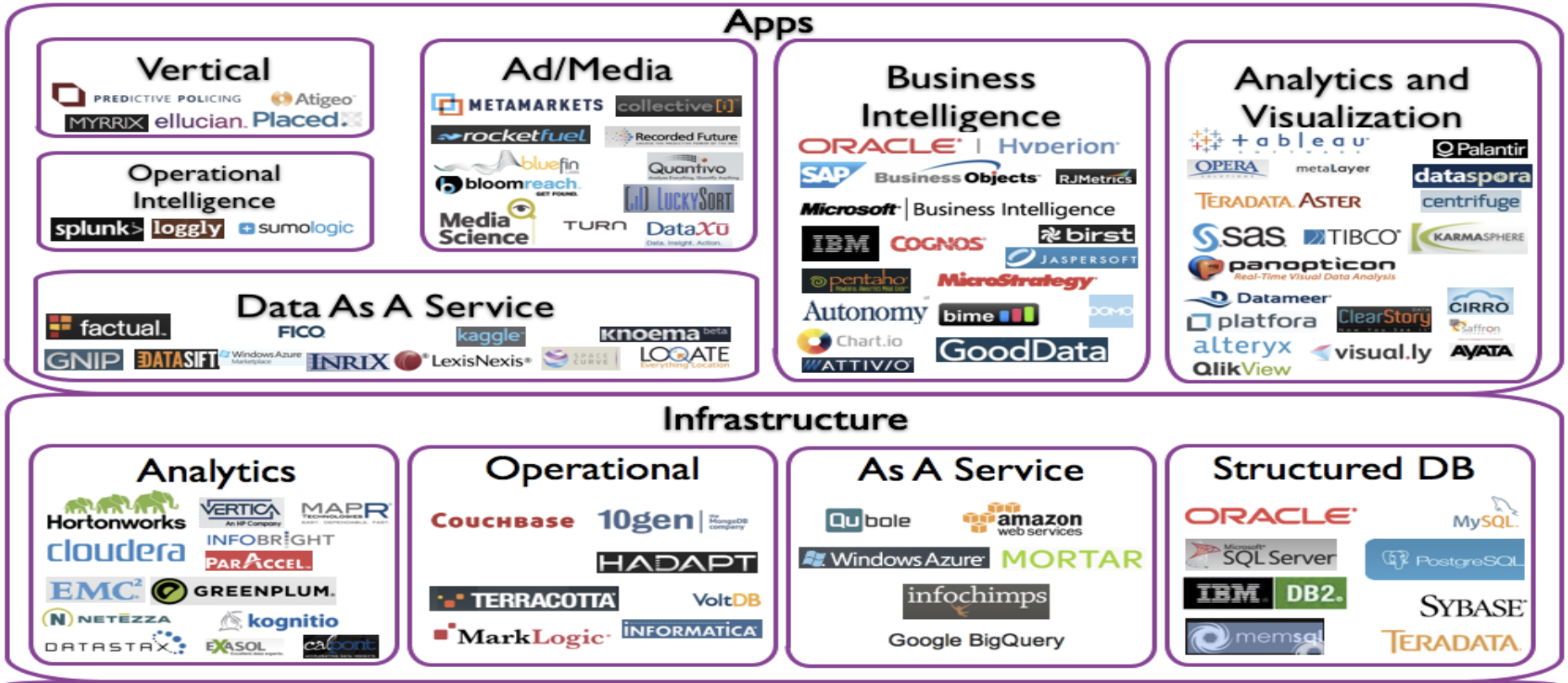
Government Big Data

Currently, federal agencies cannot make use of all their government data because they do not (or cannot afford to) employ enough data scientists—that is, experts who possess domain knowledge and can use government big data analytic technologies to ask the right questions and extract business or mission intelligence from vast pools of data. Making use of big data under these circumstances presents a unique challenge.

Download the Big Data brief (PDF, 267KB)

Contact Us

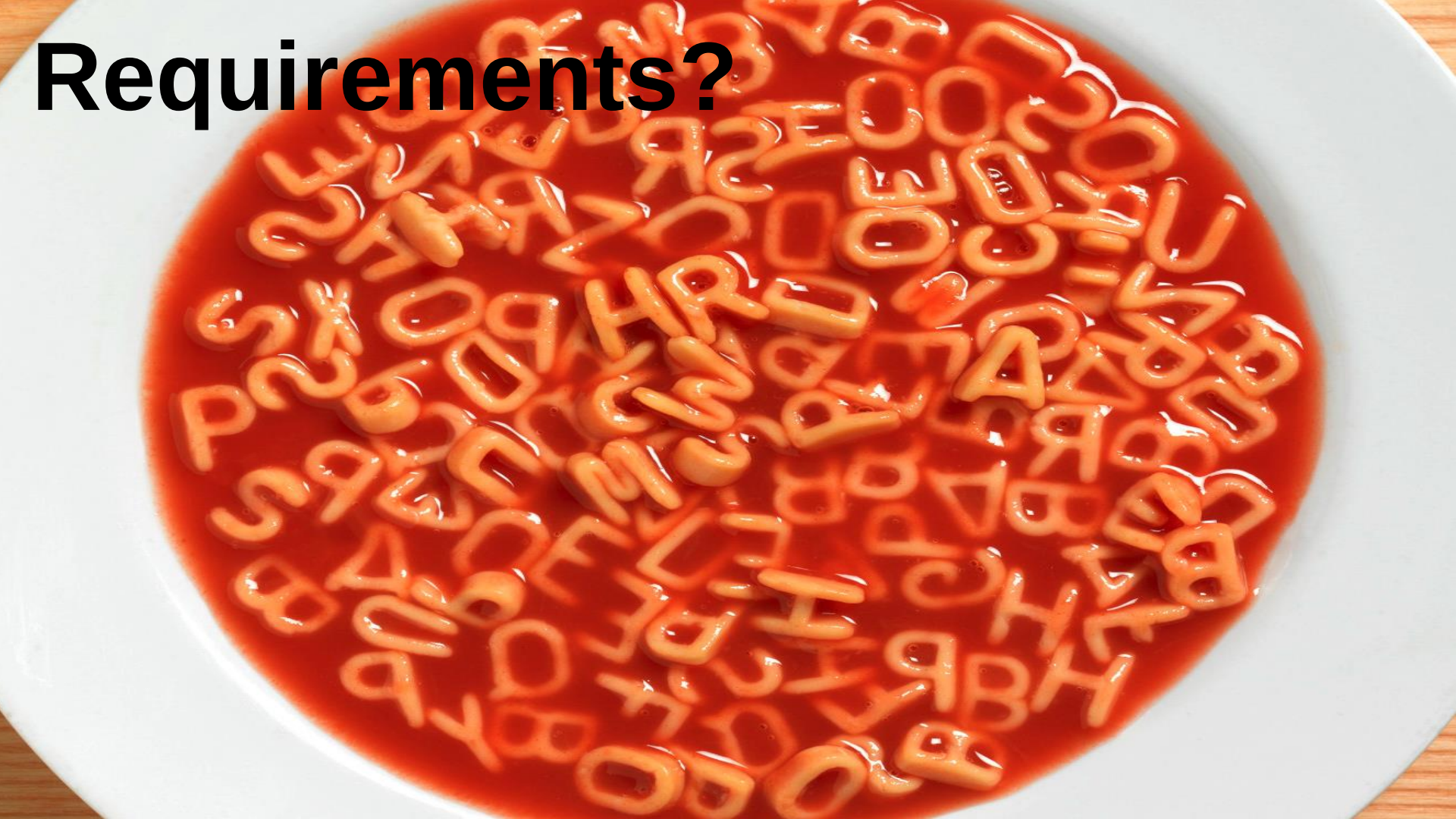
The Big Data Landscape



dave@vcdave.com

bigdatalandcsape.com

Requirements?



The Three “V”s

Volume

Velocity

Variety

Veracity

Validity

Viability

Value

The Five “W”s

Why?

Who?

Where?

When?

What?

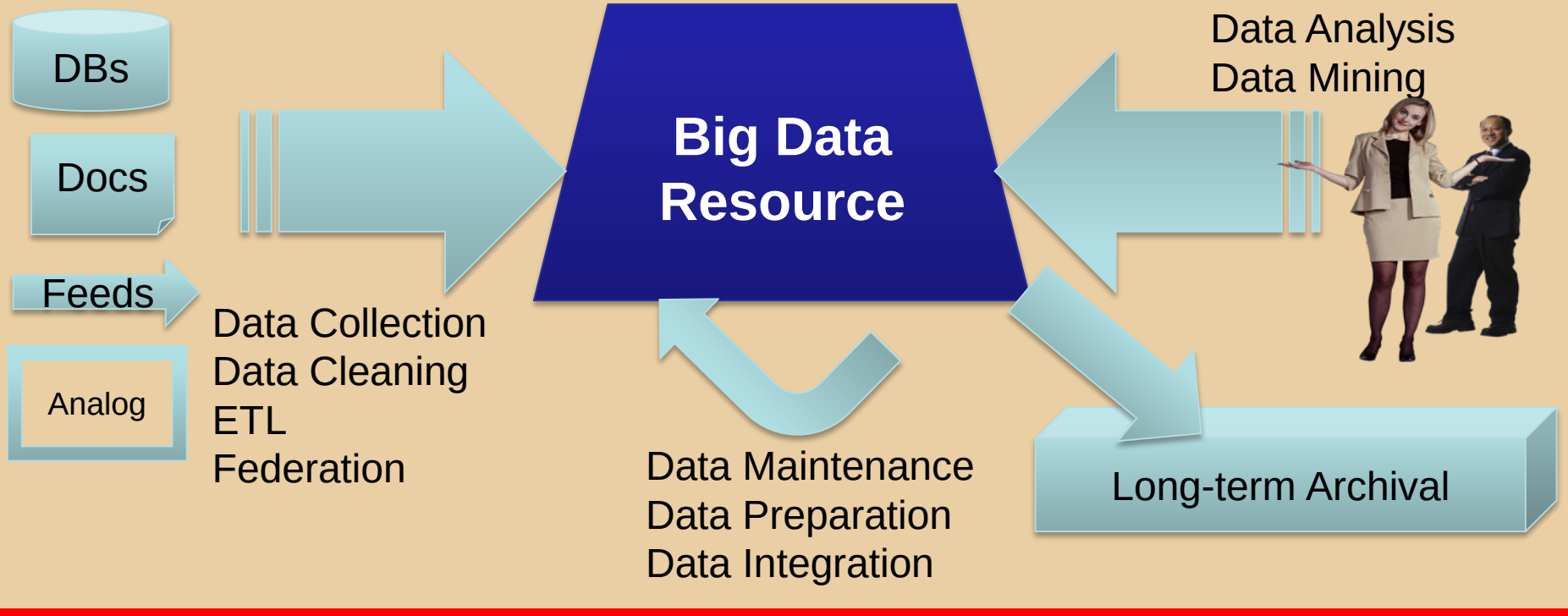
Identification

Introspection

Integration

Immutability

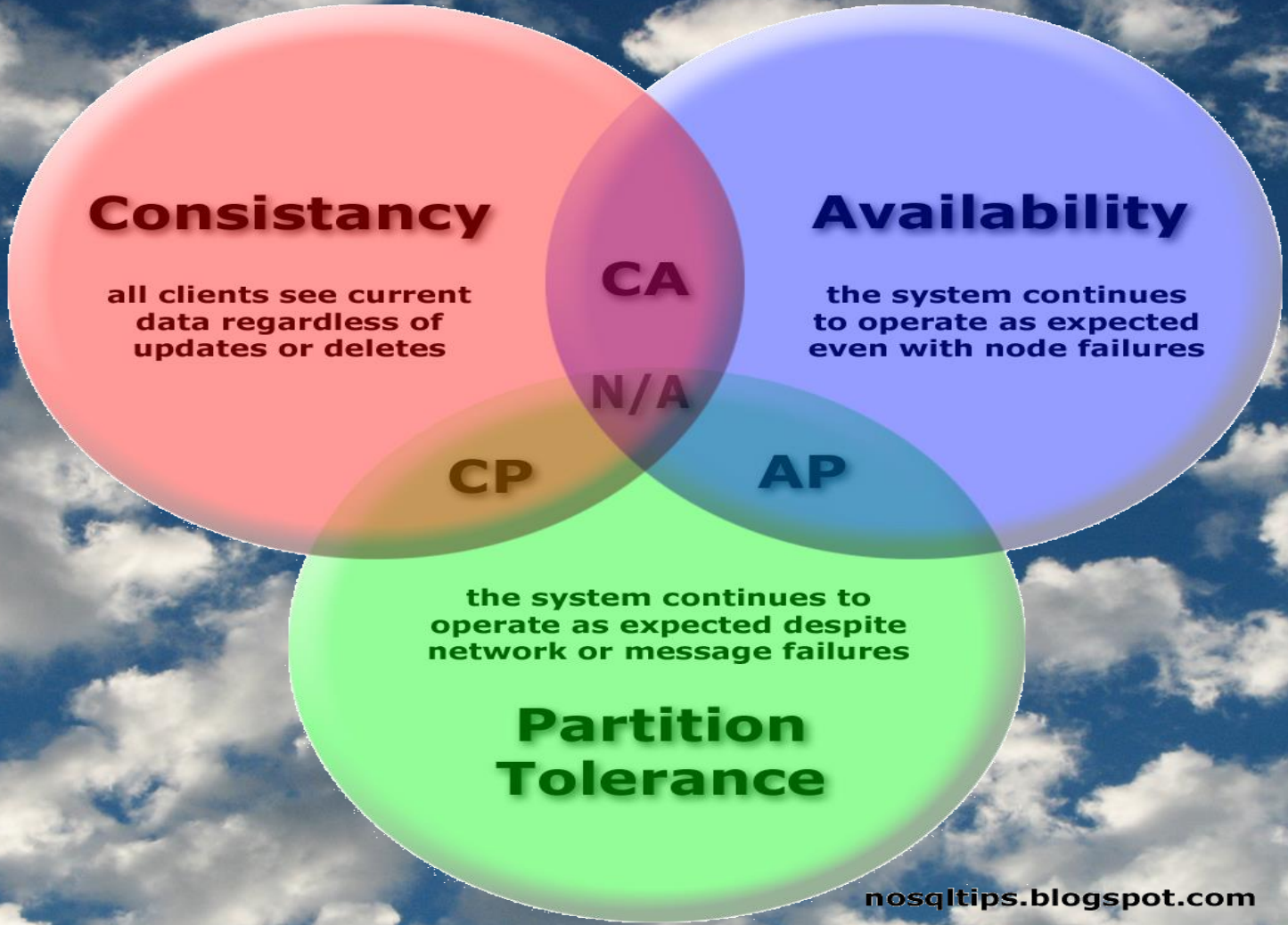
SMALL DATA		BIG DATA
Specific questions	***GOAL***	Broad concerns
One location	***LOCATION***	Many locations
Structured	***STRUCTURE***	Varied, unstructured
Single user	***SOURCE***	Many providers
Transient	***LONGEVITY***	Durable
Focused	***MEASUREMENTS***	Broad
Can be recreated	***REPRODUCIBILITY***	Gone if not captured
Small risk	***STAKES***	Big risk
Simple	***INTROSPECTION***	Metadata is vital
Complete	***ANALYSIS***	Incremental



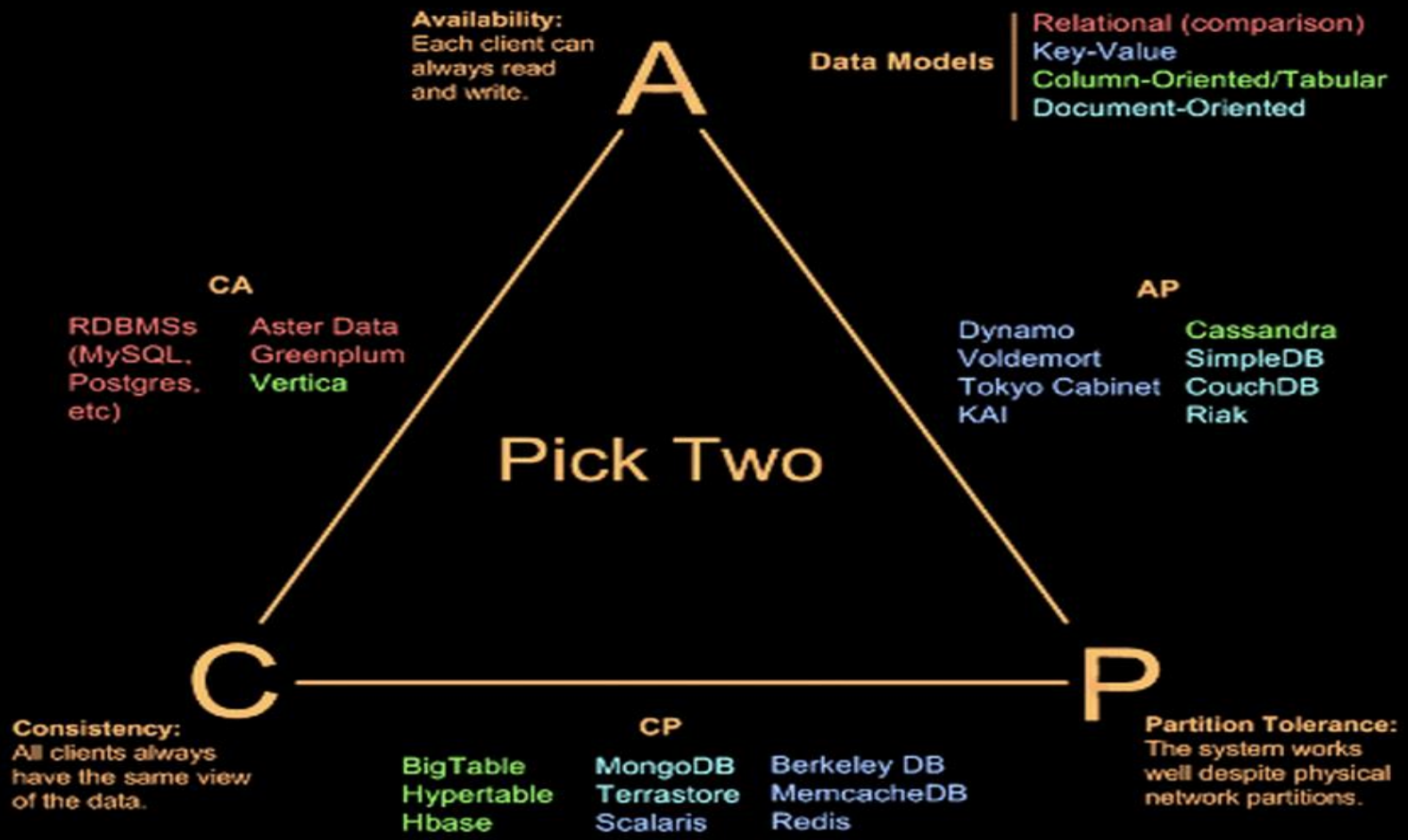
Big data is not a product, but a collection of **processes**

LOTSA

SMALL DATA		BIG DATA
Specific questions	***GOAL***	Broad concerns
One location	***LOCATION***	Many locations
Structured	***STRUCTURE***	Varied, unstructured
Single user	***SOURCE***	Many providers
Transient	***LONGEVITY***	Durable
Focused	***MEASUREMENTS***	Broad
Can be recreated	***REPRODUCIBILITY***	Gone if not captured
Small risk	***STAKES***	Big risk
Simple	***INTROSPECTION***	Metadata is vital
Complete	***ANALYSIS***	Incremental

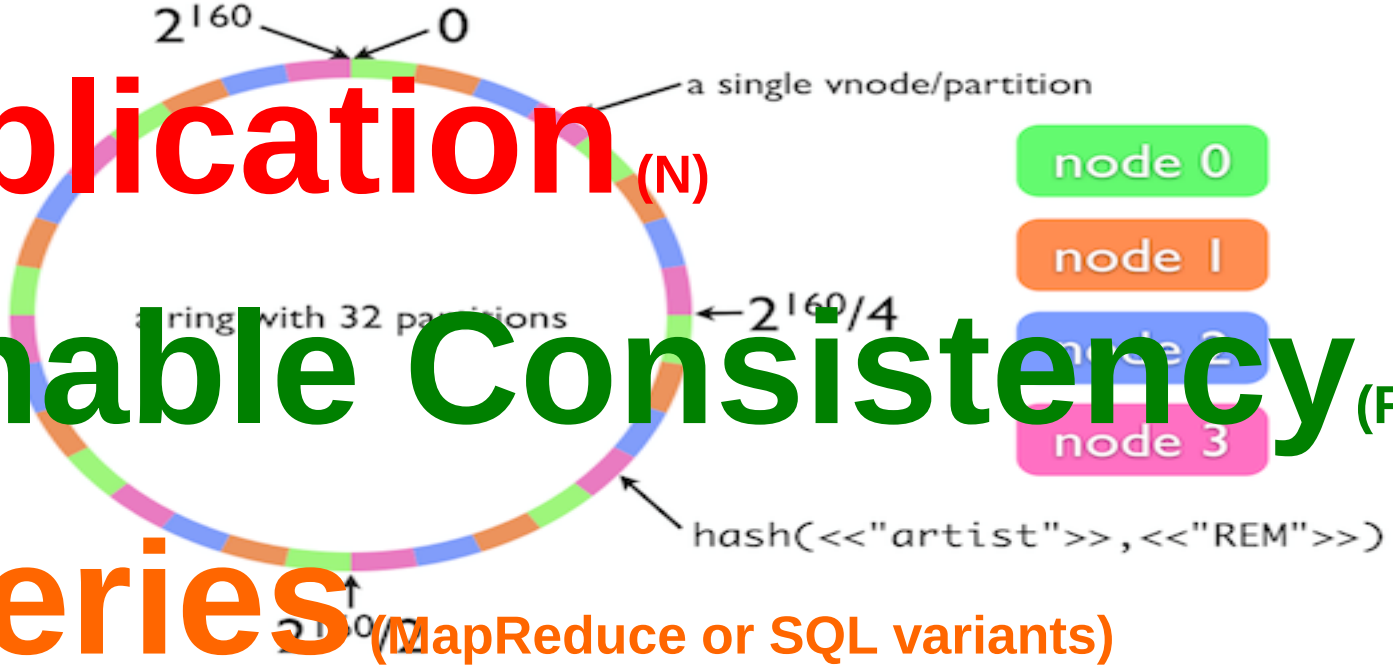

Consistancy
all clients see current data regardless of updates or deletes
Availability
the system continues to operate as expected even with node failures
CA
N/A
CP
AP
the system continues to operate as expected despite network or message failures
Partition Tolerance
nosqltips.blogspot.com

Visual Guide to NoSQL Systems



Consistent Hashing

Replication (N)

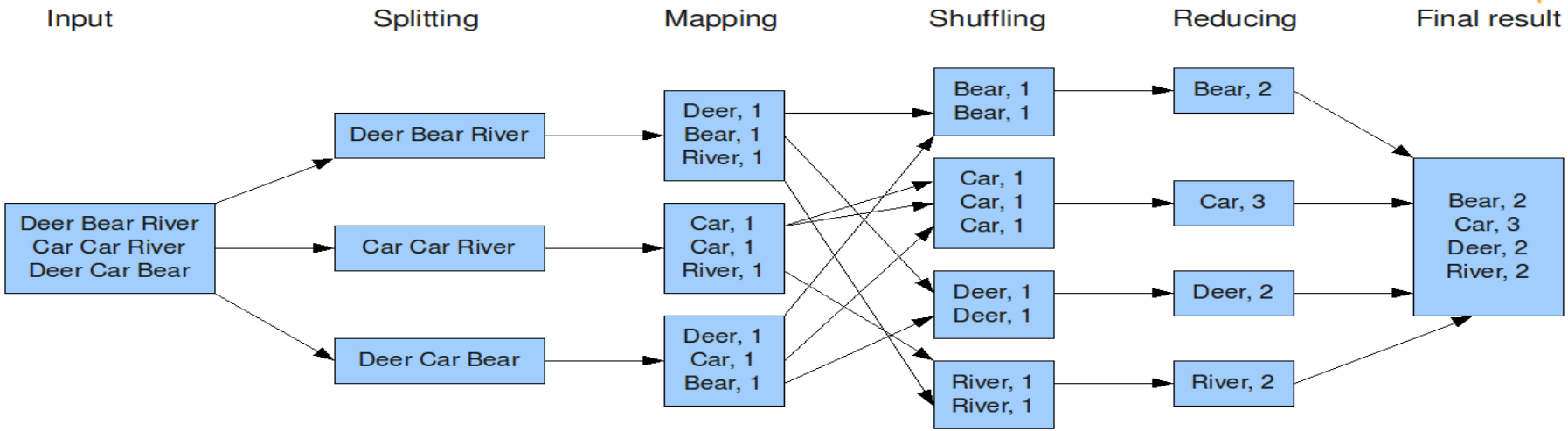


Tunable Consistency (R + W < N)

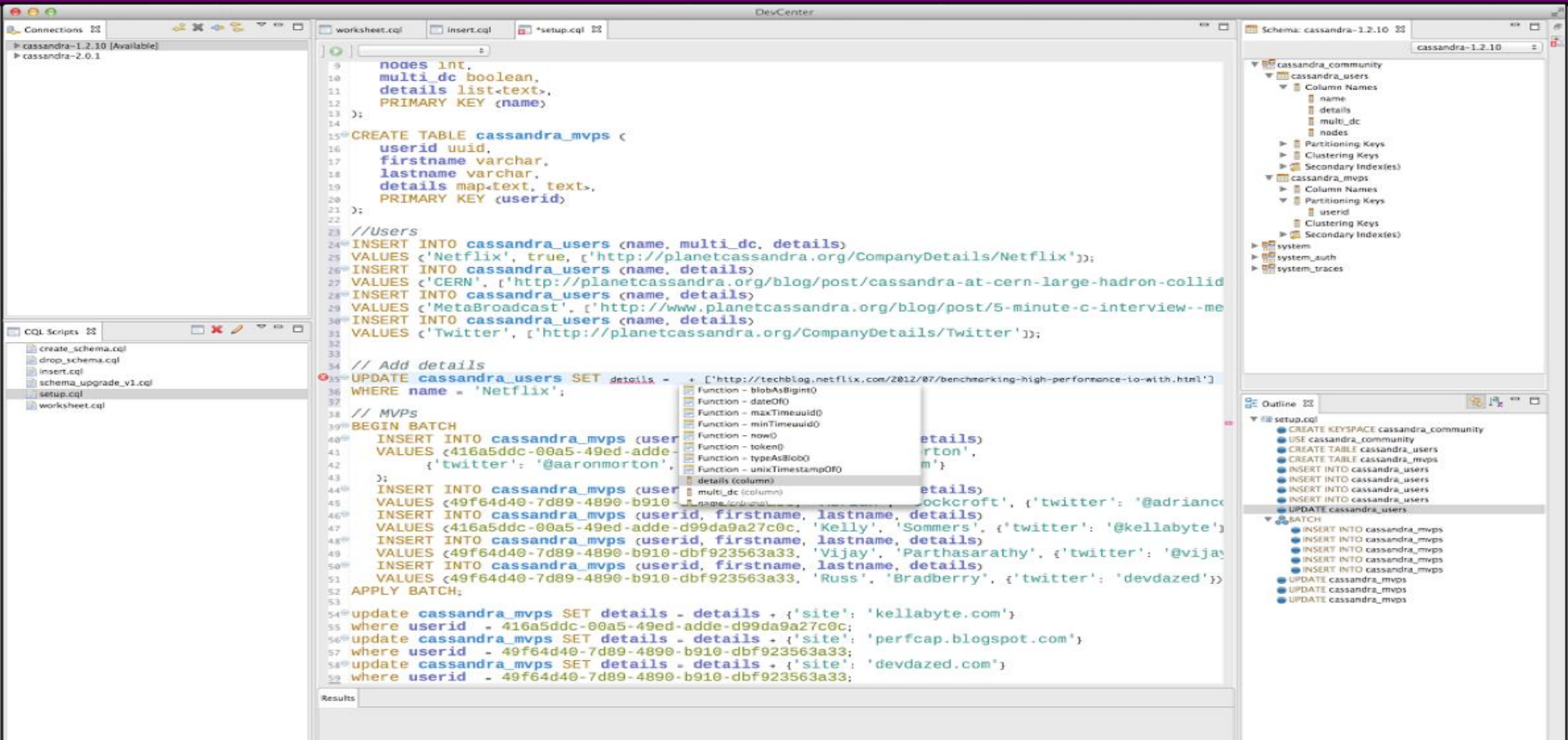
Queries (MapReduce or SQL variants)

Components

The overall MapReduce word count process



SQL variants



Solving Big Data Challenges for Enterprise Application Performance Management

Tilmann Rabl
Middleware Systems Research Group
University of Toronto, Canada
tilmann@msrg.utoronto.ca

Mohammad Sadoghi
Middleware Systems Research Group
University of Toronto, Canada
mo@msrg.utoronto.ca

Hans-Arno Jacobsen
Middleware Systems Research Group
University of Toronto, Canada
arno@msrg.utoronto.ca

Sergio Gómez-Villamor
DAMA-UPC
Universitat Politècnica de Catalunya, Spain
sgomez@ac.upc.edu

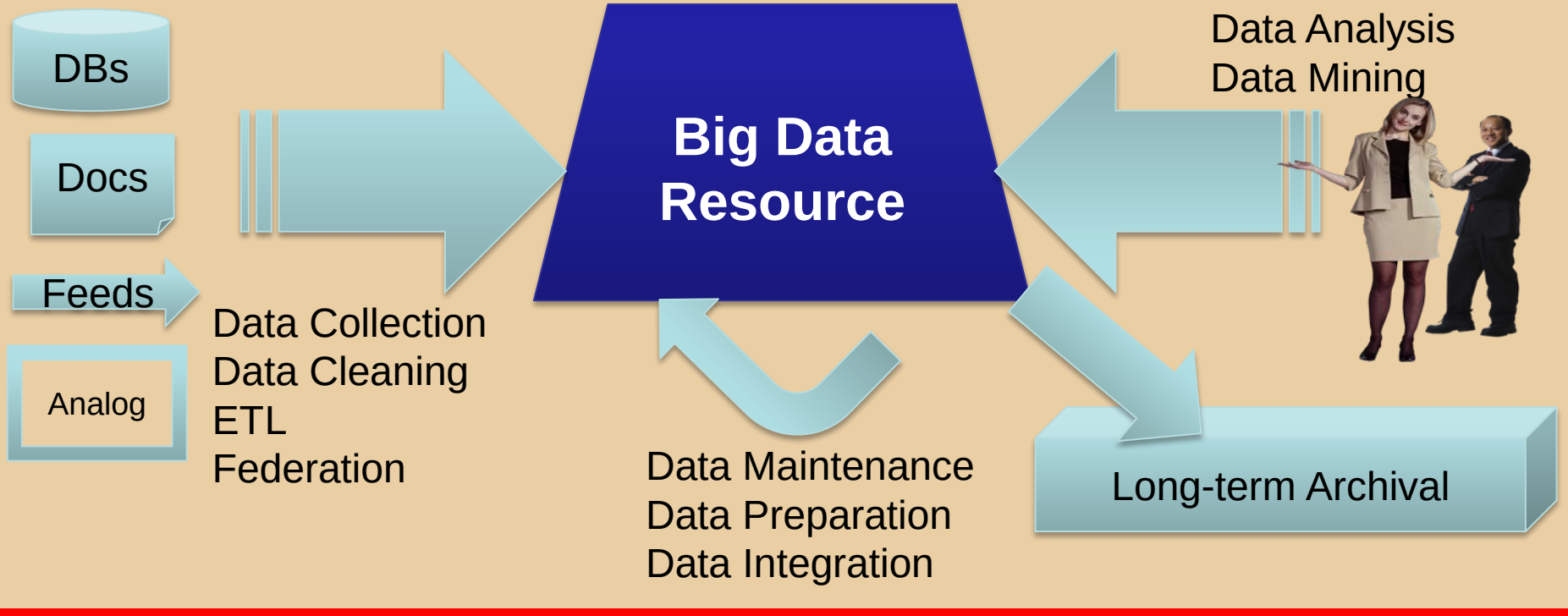
Victor Muntés-Mulero
CA Labs Europe
Barcelona, Spain
victor.muntes

Serge Mankovskii

ABSTRACT

As the complexity of enterprise systems increases, the need for monitoring and analyzing such systems also grows. A number of companies have built sophisticated monitoring tools that go far beyond simple resource utilization reports. For example, based on instrumentation and specialized APIs, it is now possible to monitor single method invocations and trace individual transactions across geographically distributed systems. This high-level of detail enables more precise forms of analysis and prediction but comes at the price of high data rates (i.e., big data). To maximize the benefit of data monitoring, the data has to be stored for an extended period of time for ulterior analysis. This new wave of big data analytics imposes new challenges especially for the application performance monitoring systems. The monitoring data has to be stored in a system that can sustain the high data rates and at the same time enable an up-to-date view of the underlying infrastructure. With the advent of modern key-value stores, a variety of data storage systems

...lete data ... heteroge- ... es their admin- istrat... administrators an on-line view of the system ... monitoring frameworks have been developed. Common examples are Ganglia [20] and Nagios [12]. These are widely used in open-source projects and academia (e.g., Wikipedia[1]). However, in industry settings, in presence of stringent response time and availability requirements, a more thorough view of the monitored system is needed. Application Performance Management (APM) tools, such as Dynatrace[2], Quest PerformaSure[3], ... and CA APM[5] ... or combinations of services, as well as about failure rates, resource utilization, etc. Different monitoring targets such as the response



Big data is not a product, but a collection of **processes**

Sources

- http://jobs.aol.com/articles/2011/08/10/data-scientist-the-hottest-job-you-havent-heard-of/
- http://en.wikipedia.org/wiki/Data_science, http://en.wikipedia.org/wiki/MapReduce, http://en.wikipedia.org/wiki/Big_data, http://en.wikipedia.org/wiki/List_of_countries_by_population
- http://www.delphianalytics.net/wp-content/uploads/2013/04/GrowthOfDataVsDataAnalysis.png
- http://media.economist.com/images/20100227/20100227SRC696.gif
- http://www.datasciencecentral.com/profiles/blogs/structured-vs-unstructured-data-the-rise-of-data-anarchy
- http://www.zerohedge.com/sites/default/files/images/user5/imageroot/2012/10-2/Food%20For%20Thoughts.jpg
- http://www.theguardian.com/news/datablog/2012/mar/09/Big-data-theory
- http://blogs-images.forbes.com/daveteinleb/files/2012/07/Big-Data-Trends_0031.png
- http://www.slideshare.net/4Neha/big-data-15681560
- http://www.mimul.com/pebble/default/images/blog/cloud/nosql_cap04.png
- http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg
- http://reflectionsblog.emc.com/2012/06/scientific-big-data/
- http://go.nutanix.com/rs/nutanix/images/CAP_Diagram_dist-copy.jpg
- http://www.paperplanes.de/2011/12/9/the-magic-of-consistent-hashing.html
- Jules J. Berman, Principles of Big Data, Morgan Kaufmann, 2013
- Research papers, Wikipedia, …