



INTRODUCTION TO  
**BIG DATA MANAGEMENT**

**Björn Þór Jónsson**

CRESS and School of Computer Science,  
Reykjavík University



1 NEW DEFINITION IS ADDED ON UPON

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



20,000+ NEW POSTS ON tumblr.

THE LARGEST SOCIAL READING PUBLISHING COMPANY

320+ NEW twitter ACCOUNTS

100+ NEW Linked in ACCOUNTS

13,000+ iPhone APPLICATIONS DOWNLOADED

1 NEW ARTICLE IS PUBLISHED associatedcontent

THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!



QUESTIONS ASKED ON THE INTERNET...

100+ Answers.com 40+ Yahoo! Answers

6,600+ NEW PHOTOS ARE UPLOADED ON flickr



600+ NEW VIDEOS YouTube

50+ WORDPRESS DOWNLOADS

70+ DOMAINS REGISTERED

60+ NEW BLOGS

168 MILLION EMAILS ARE SENT

694,445 SEARCH QUERIES

1,700+ Firefox DOWNLOADS

695,000+ facebook STATUS UPDATES



125+ PLUGIN DOWNLOADS

25+ HOURS TOTAL DURATION

1,500+ BLOG POSTS

79,364 WALL POSTS

510,040 COMMENTS



Google

Google Search











# 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.



The image shows the cover of a report titled "Making Big Data Work for Government" published by CSC. The cover features the CSC logo at the top left and bottom left. The title is prominently displayed in the center. Below the title, there is a brief summary of the report's content, followed by a quote from the report. At the bottom right, there is a logo for the Center for Strategic and Policy Studies.

**CSC**

A Briefing  
from GBC  
Industry Insights  
November 2012

### Making Big Data Work for Government

"Big data" comes with many promises, but the fact of these is access to previously unprocessed information, "dark data," and large volumes of unprocessed information. Many estimates place the amount of unprocessed data worldwide at 10 percent of all data, and as the average federal agency stores 1.4 petabytes of data, this represents a potentially vast store of previously unprocessed information for analysis.<sup>1</sup> 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.

**44**

Center estimates that big data initiatives in 2012 will total \$24 billion,<sup>2</sup> and big data will drive \$200 billion in spending over the next five years.<sup>3</sup>

### Subsiding 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).<sup>4</sup> Yet, capturing big data is as much an organizational challenge as a technological one. Industry observations of big data technologies are largely bifurcated into two main functions: data discovery and the more traditional business intelligence.<sup>5</sup> Companies have been heading south since and moving to the future.<sup>6</sup> Center estimates that big data initiatives in 2012 will total \$24 billion,<sup>7</sup> and big data will drive \$200 billion in spending over the next five years.<sup>8</sup>

Many organizations are trying to address the technology challenge of analyzing big data. The Technomic Foundation's Big Data Commission, for example, recently released a report that gives organizations guidelines for an effective big data program. Examination of these areas 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 vital to its mission
- Specify 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.<sup>9</sup>

**CSC**

Center for Strategic and Policy Studies  
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

## Apps

### Vertical



### Operational Intelligence



### Ad/Media



### Data As A Service



### Business Intelligence



### Analytics and Visualization



## Infrastructure

### Analytics



### Operational



### As A Service



### Structured DB



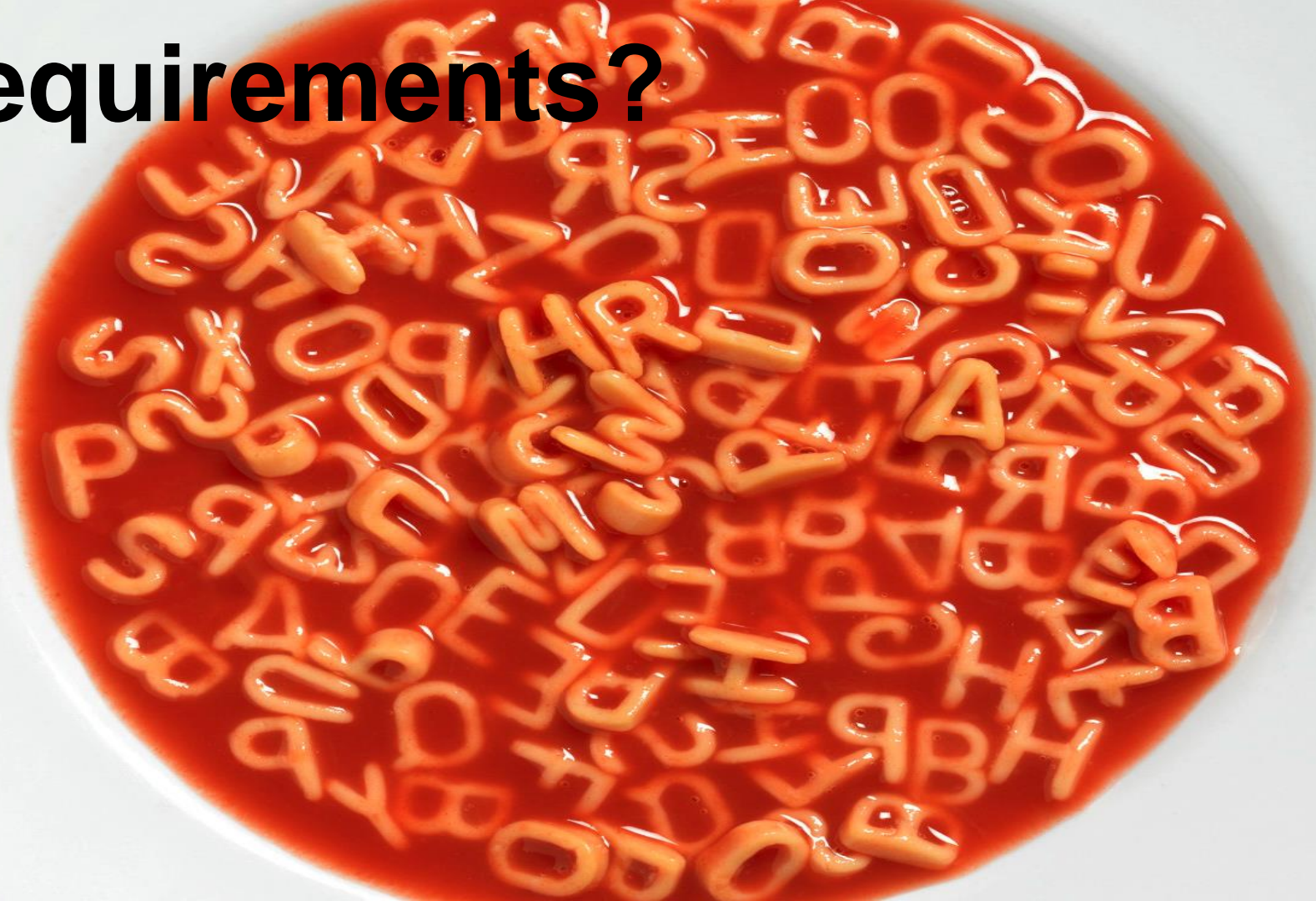
Technologies



APACHE HBASE



**Requirements?**





# The Three “V”s

Volume

Velocity

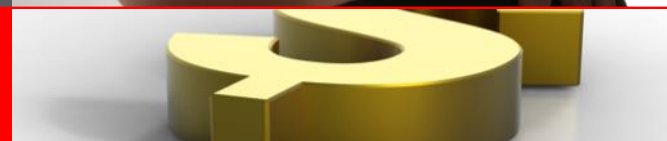
Variety

Veracity

Validity

Viability

Value



# The Five “W”s



Why?

Who?

Where?

When?

What?

Four Eyes



**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



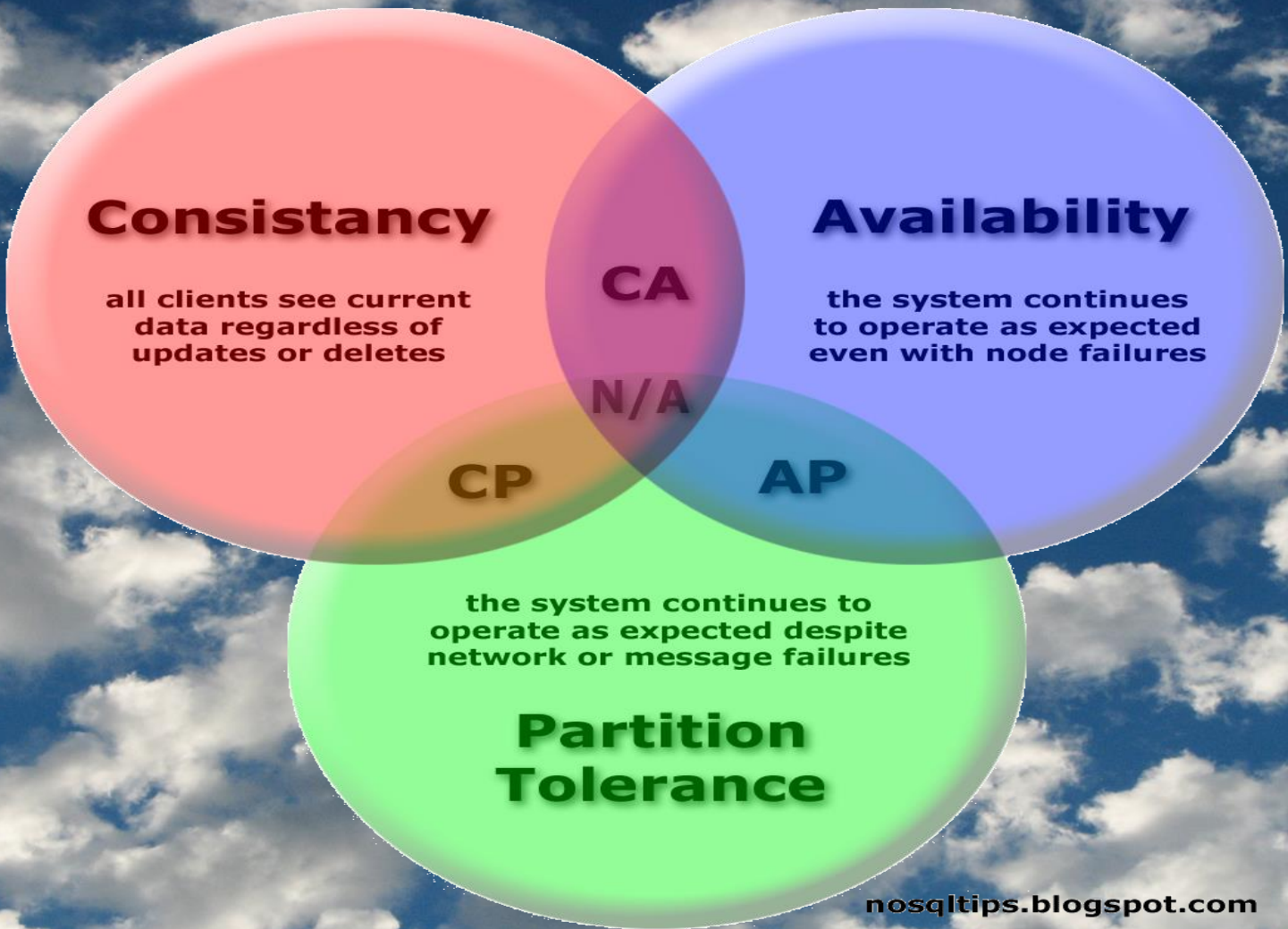
**Big data is not a product, but a collection of processes**

# LOTS ~~SA~~ SMALL DATA

# BIG DATA

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





## Consistency

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

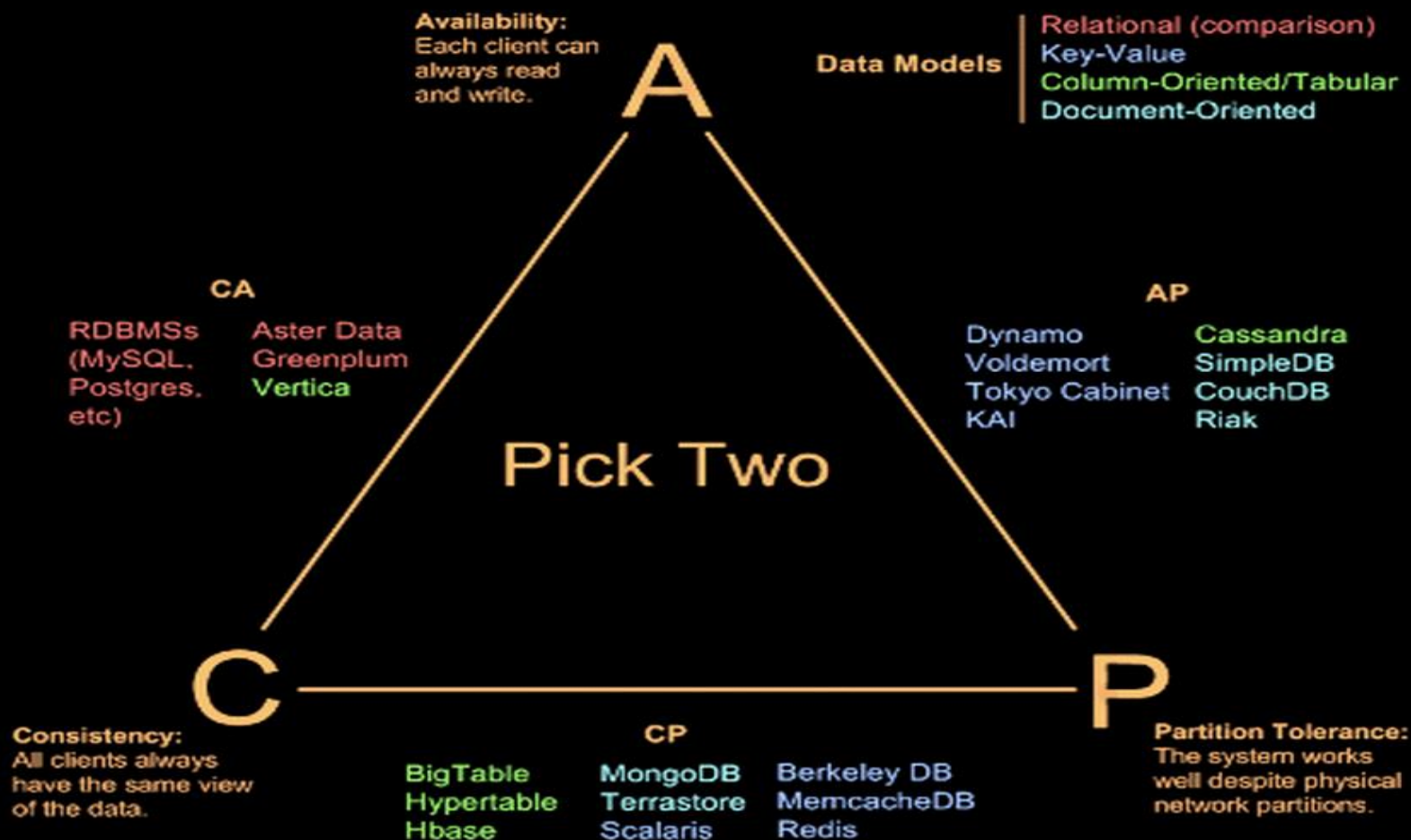
CP

AP

the system continues to operate as expected despite network or message failures

## Partition Tolerance

# Visual Guide to NoSQL Systems



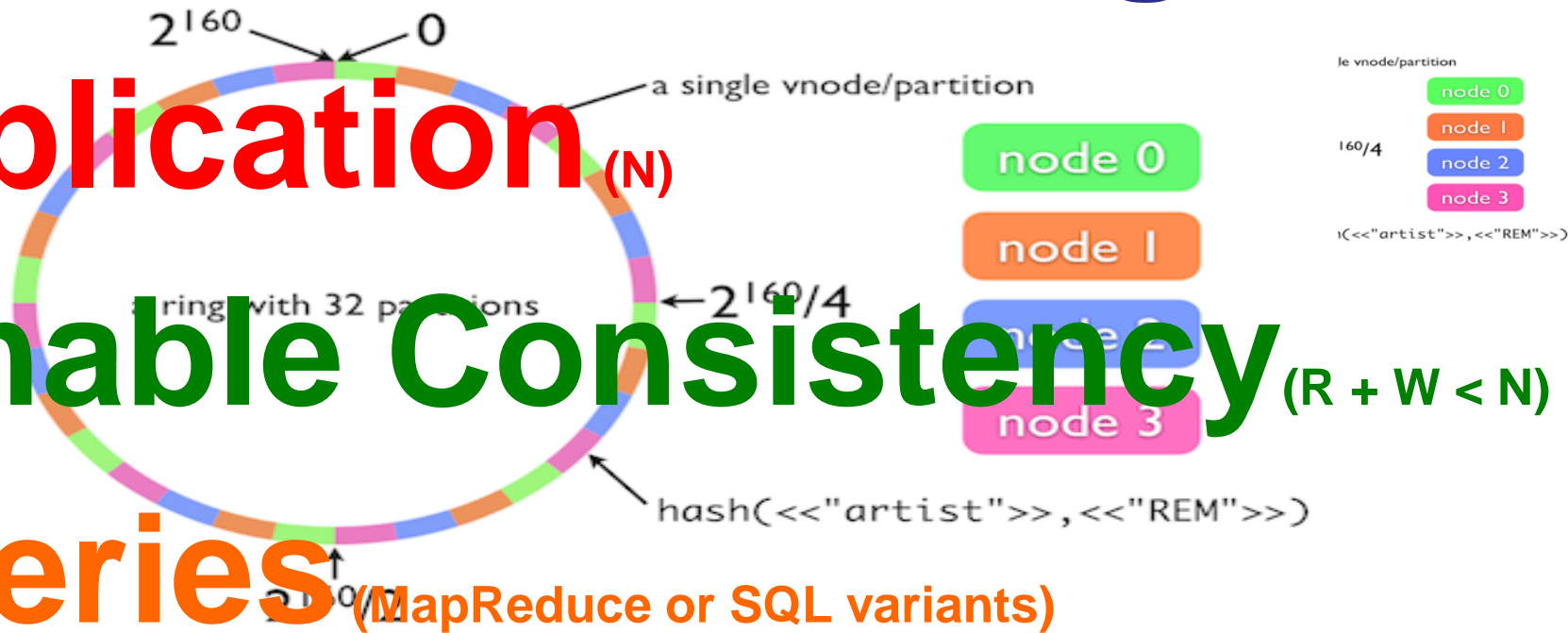
# Consistent Hashing

**Replication** (N)

**Tunable Consistency** ( $R + W < N$ )

**Queries** (MapReduce or SQL variants)

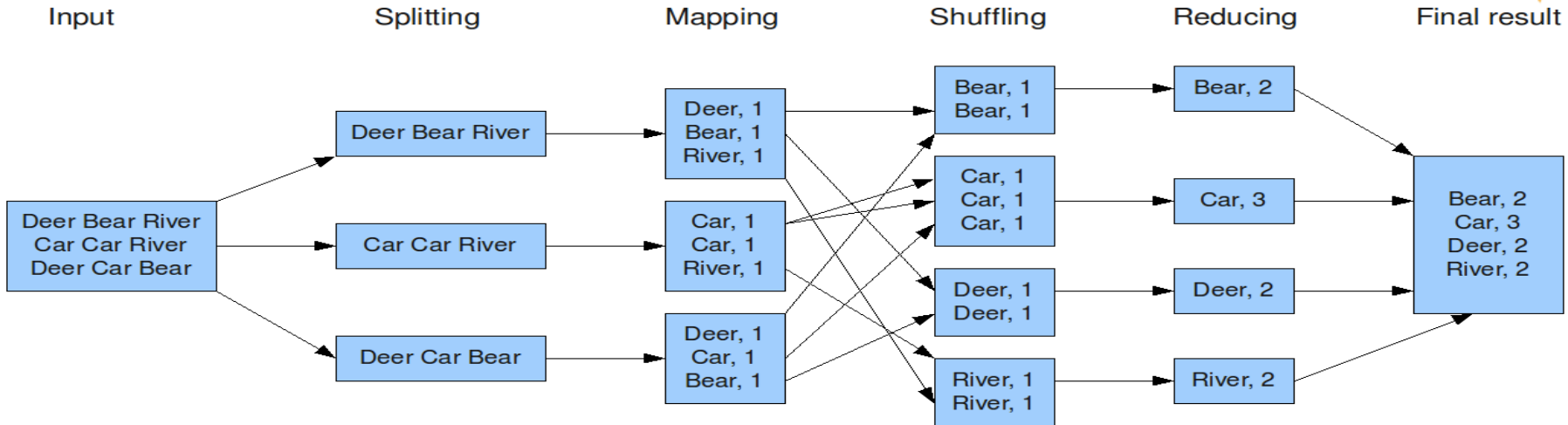
**Components**







The overall MapReduce word count process



# SQL variants

The screenshot displays a SQL IDE interface with a script for creating and populating Cassandra tables. The script includes the following SQL statements:

```
9 nodes int,
10 multi_dc boolean,
11 details list<text>,
12 PRIMARY KEY (name)
13 );
14
15 CREATE TABLE cassandra_mvps (
16 userid uuid,
17 firstname varchar,
18 lastname varchar,
19 details map<text, text>,
20 PRIMARY KEY (userid)
21 );
22
23 //Users
24 INSERT INTO cassandra_users (name, multi_dc, details)
25 VALUES ('Netflix', true, ['http://planetcassandra.org/CompanyDetails/Netflix']);
26 INSERT INTO cassandra_users (name, details)
27 VALUES ('CERN', ['http://planetcassandra.org/blog/post/cassandra-at-cern-large-hadron-collid
28 INSERT INTO cassandra_users (name, details)
29 VALUES ('MetaBroadcast', ['http://www.planetcassandra.org/blog/post/5-minute-c-interview--me
30 INSERT INTO cassandra_users (name, details)
31 VALUES ('Twitter', ['http://planetcassandra.org/CompanyDetails/Twitter']);
32
33
34 // Add details
35 UPDATE cassandra_users SET details = ['http://techblog.netflix.com/2012/07/benchmarking-high-performance-10-with.html']
36 WHERE name = 'Netflix';
37
38 // MVPS
39 BEGIN BATCH
40 INSERT INTO cassandra_mvps (userid, firstname, lastname, details)
41 VALUES ('416a5ddc-00a5-49ed-adde-d99da9a27c0c', 'Kelly', 'Sommers', {'twitter': '@kellybyte'});
42 INSERT INTO cassandra_mvps (userid, firstname, lastname, details)
43 VALUES ('49f64d40-7d89-4890-b910-dbf923563a33', 'Vijay', 'Parthasarathy', {'twitter': '@vijay
44 INSERT INTO cassandra_mvps (userid, firstname, lastname, details)
45 VALUES ('49f64d40-7d89-4890-b910-dbf923563a33', 'Russ', 'Bradberry', {'twitter': 'devdazed'});
46 INSERT INTO cassandra_mvps (userid, firstname, lastname, details)
47 VALUES ('416a5ddc-00a5-49ed-adde-d99da9a27c0c', 'Kelly', 'Sommers', {'twitter': '@kellybyte'});
48 INSERT INTO cassandra_mvps (userid, firstname, lastname, details)
49 VALUES ('49f64d40-7d89-4890-b910-dbf923563a33', 'Vijay', 'Parthasarathy', {'twitter': '@vijay
50 INSERT INTO cassandra_mvps (userid, firstname, lastname, details)
51 VALUES ('49f64d40-7d89-4890-b910-dbf923563a33', 'Russ', 'Bradberry', {'twitter': 'devdazed'});
52 APPLY BATCH;
53
54 update cassandra_mvps SET details = details + {'site': 'kellybyte.com'};
55 where userid = 416a5ddc-00a5-49ed-adde-d99da9a27c0c;
56 update cassandra_mvps SET details = details + {'site': 'perfcap.blogspot.com'};
57 where userid = 49f64d40-7d89-4890-b910-dbf923563a33;
58 update cassandra_mvps SET details = details + {'site': 'devdazed.com'};
59 where userid = 49f64d40-7d89-4890-b910-dbf923563a33;
```

The IDE interface includes a 'Connections' panel on the left showing 'cassandra-1.2.10 [Available]' and 'cassandra-2.0.1'. A 'CQL Scripts' panel at the bottom left lists files like 'create\_schema.cql', 'drop\_schema.cql', 'insert.cql', 'schema\_upgrade\_v1.cql', 'setup.cql', and 'worksheet.cql'. The main editor shows the script with line numbers and a tooltip for the 'details' column type. The right sidebar shows a 'Schema: cassandra-1.2.10' tree view with tables 'cassandra\_community', 'cassandra\_users', and 'cassandra\_mvps'. An 'Outline' panel at the bottom right shows a list of SQL statements executed in the session.

# Solving Big Data Challenges for Enterprise Application Performance Management

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

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

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

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


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

Sergey Manokovskii

## 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, provide an up-to-date view of the underlying infrastructure. With the advent of modern key-value stores, a variety of data storage systems

complete data heterogeneity, and the need for 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<sup>1</sup>). 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<sup>2</sup>, Quest PerformanceSure<sup>3</sup>, AppDynamics<sup>4</sup>, and CA APM<sup>5</sup> provide sophisticated views on the monitored system. These tools instrument the applications to monitor the infrastructure, the response time of each service or combinations of services, as well as about failure rates, resource utilization, etc. Different monitoring targets such as the response



cassandra



**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/Data_science), <http://en.wikipedia.org/wiki/MapReduce>,  
[http://en.wikipedia.org/wiki/Big\\_data](http://en.wikipedia.org/wiki/Big_data), [http://en.wikipedia.org/wiki/List\\_of\\_countries\\_by\\_population](http://en.wikipedia.org/wiki/List_of_countries_by_population)
- <http://www.delphianalytics.net/wp-content/uploads/2013/04/GrowthOfDataVsDataAnalysts.png>
- <http://media.economist.com/images/20100227/201009SRC696.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/davefeinleib/files/2012/07/Big-Data-Trends.0031.png>
- <http://www.slideshare.net/4Neba/big-data-15681560>
- [http://www.mimul.com/pebble/default/images/blog/cloud/nosql\\_cap04.png](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://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://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, ...