
*Together with* MAPR TECHNOLOGIES

# Realising Value from Data

## Open Source Drives Innovation & Adoption in Big Data

**BCS Open Source SIG** | London | 1 May 2013

# Timings

6:00 - 6:30pm. Register / Refreshments

6:30 - 8:00pm, Presentation Session

8:00 - 8:30pm, Networking

# Objectives

- What is Big Data?
- Evolution of Open Source Hadoop and its influence on the Big Data phenomenon
- The Open Source ecosystem around Big Data
- What is the importance of Open Source for Hadoop?
- What business challenges does Hadoop address?
- What does a Hadoop architecture look like?
- How have MapR built a unique offering on top of Hadoop?

# Introduction

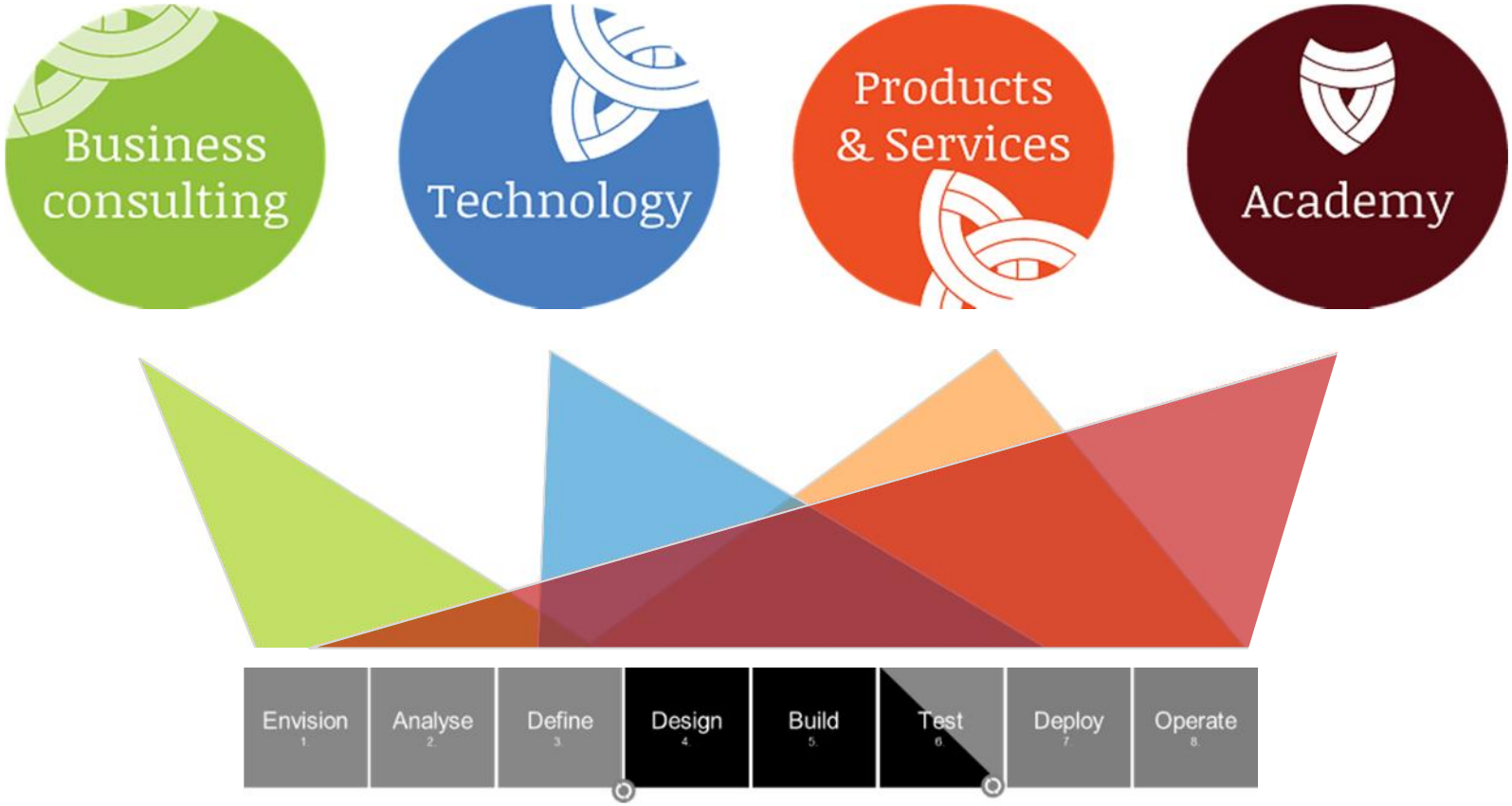
## Big Data & Apache Hadoop

## MapR / Drill Demo

## Summary & Further Learning

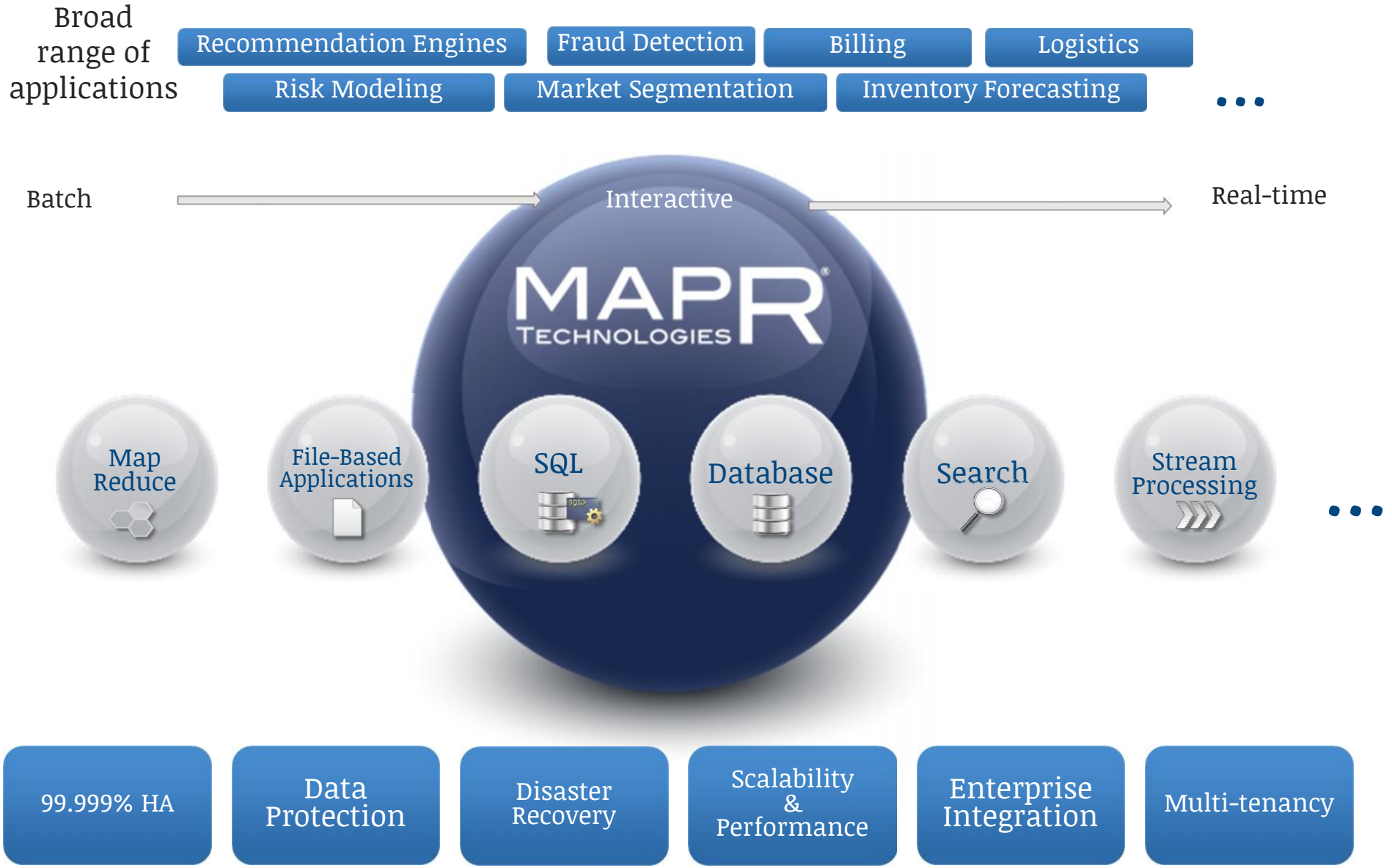
# About Onepoint IQ

## Onepoint IQ empowers individuals and organisations to discover and deliver real value from big (and small) data



# About MapR

## One Platform for Big Data



# Introduction – Onepoint IQ & MapR

## Presentation Team Today

Shashin Shah
*Technology Zen Master | Founding Dir*

Michael Hausenblas
*Chief Data Engineer - EMEA*

Introduction

**Big Data & Apache Hadoop**

MapR / Drill Demo

Summary & Further Learning

# Big Data Definitions

Although there is not a universally accepted definition for 'Big Data', all acknowledge the technology advances that are now available to handle data (big and small).

"Datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse."

McKinsey&Company

"Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it."

O'REILLY radar

"Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis."

IDC
*Analyze the Future*

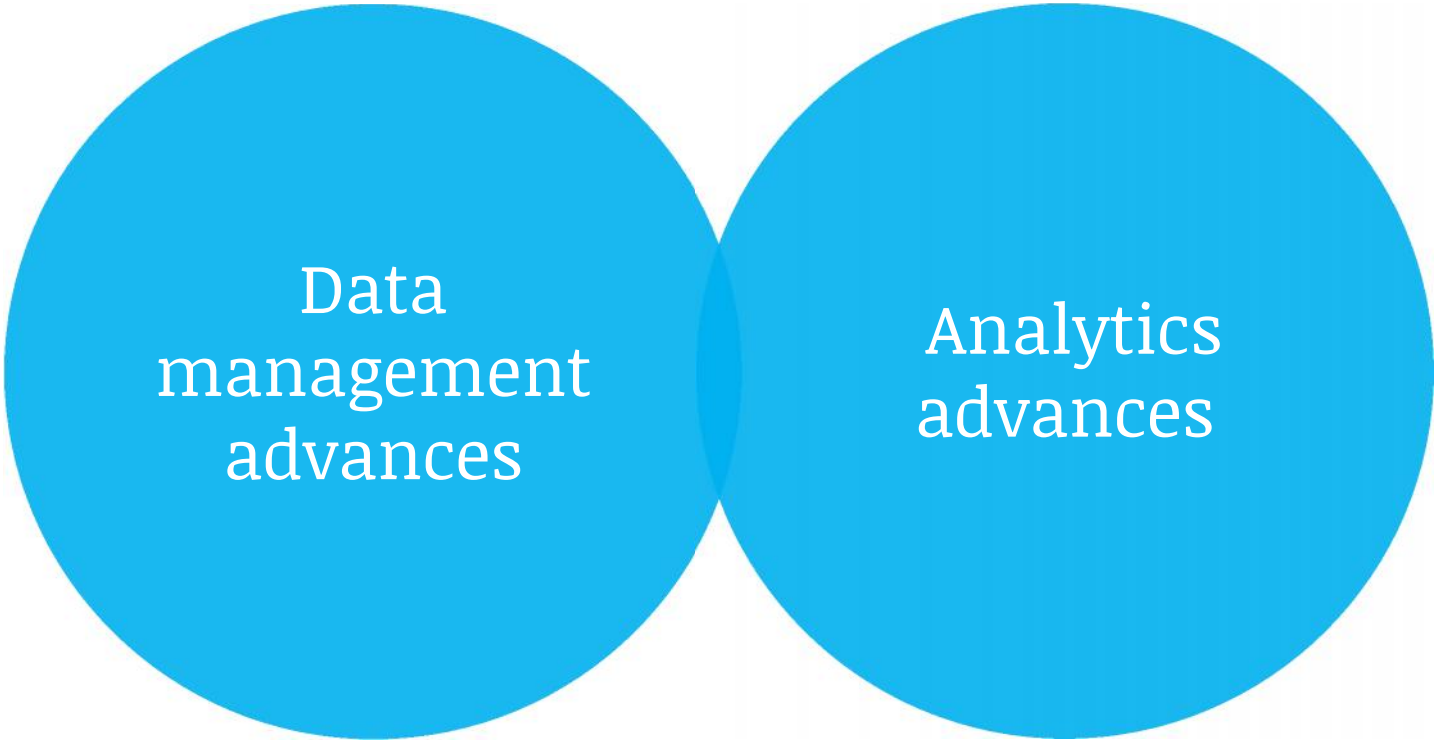
# Business Needs Spurn Innovation

Web search engines were among the first to confront the 'Big Data' problem. Today, social networks, mobile phones, sensors and science contribute to petabytes of data created daily.


Source: Hortonworks, Apache Lucene Eurocon, Barcelona, 2011

# Big Data Advances

The technology advances around 'Big Data' can be broadly grouped into two categories.




Big, fast data

Big, better analytics

# Traditional vs. Big Data – Data Management Advances

The advances do not replace existing enterprise data warehousing (EDW), business intelligence (BI), or analytics approaches – they instead enhance and extend them.


| Traditional decision-making environment | Big data innovations and extensions |
|---|---|
| • Integrated data sources | • Virtualised and blended data sources |
| • Structured data | • Unstructured, semi-structured, multi-structured data |
| • Aggregated and granular data (with limitations) | • Large volumes of granular data (without limits) |
| • Relational EDW with at-rest data | • Non-relational EDW with at-rest data |
| • Dimensional cubes / marts with at-rest data | • Streaming systems with in-motion data |
| • One-size-fits-all data management | • Flexible and optimised data management |

# Traditional vs. Big Data – Analytics advances

The advances do not replace existing enterprise data warehousing (EDW), business intelligence (BI), or analytics approaches – they instead enhance and extend them.


| **Traditional decision-making environment** | **Big data innovations and extensions** |
|---|---|
| • Reporting and OLAP | • Advanced analytic functions and predictive models |
| • Dashboards and scorecards | • Sophisticated visualisation of large data sets |
| • Structured navigation (drill down / up; slice / dice) | • Flexible exploration of large data sets |
| • Humans interpret results, patterns and trends | • Sophisticated trend and pattern analysis through machine learning |
| • Manual analyses, decisions, actions | • Model / rules-driven decisions & actions |

# Our View

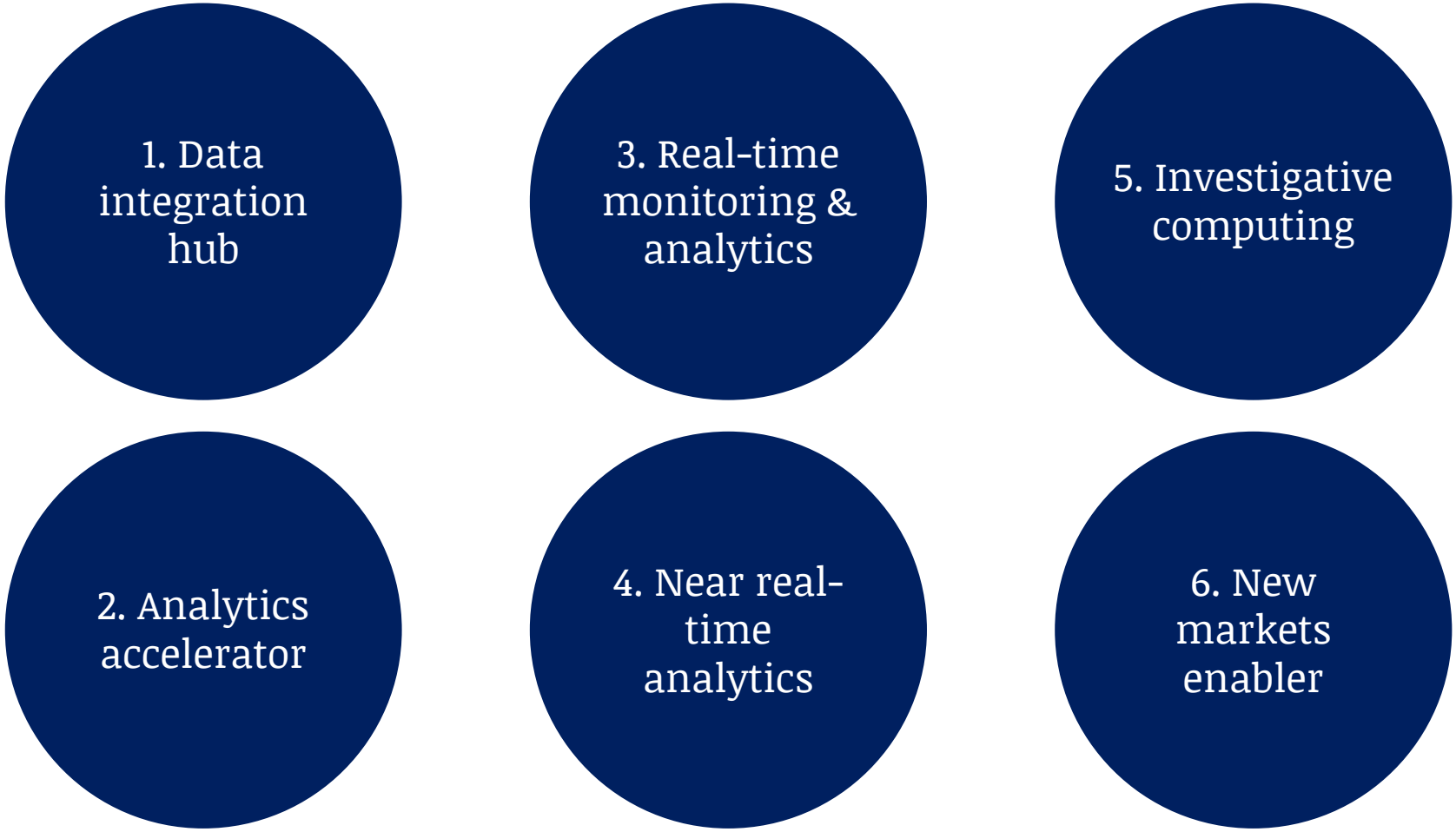
'Big Data' is data that becomes large enough that it cannot be processed using conventional methods. A fantastic array of advances aim to address this challenge.

"True business value comes from:

- *properly* choosing and applying
- *appropriate* 'Big Data' technology advances to
- *specific* data challenges (or opportunities) of an organisation
- *whether* big or small."

# Use Case Clusters (or Deployment Patterns)

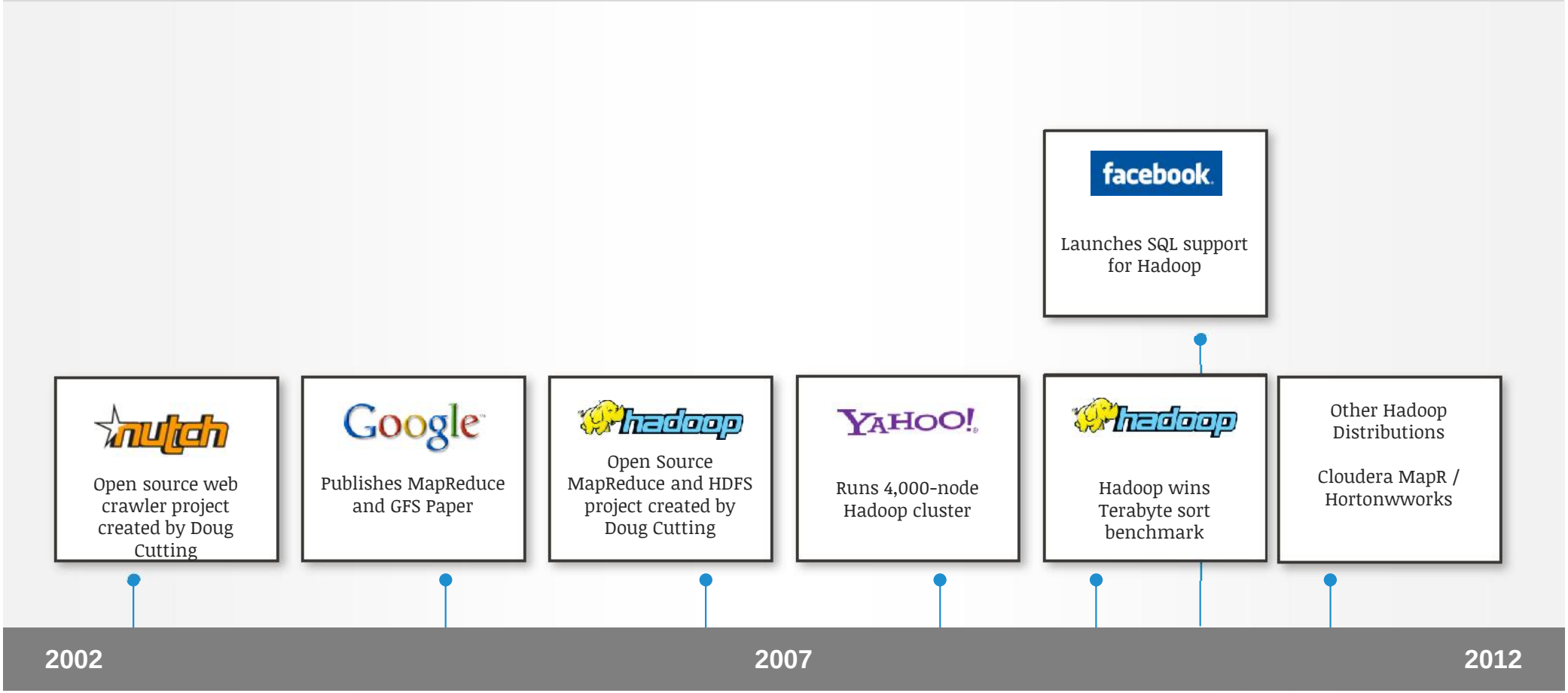
To speed up the discovery of benefits and value from Big (and small) Data, we group the nearly endless permutations of Big Data use cases into a few key clusters.



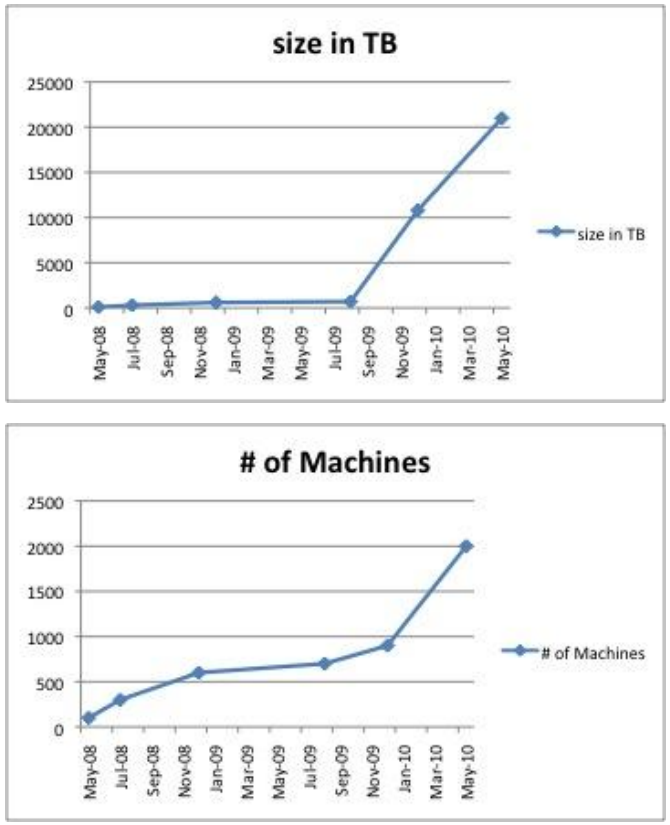

These clusters are not mutually exclusive. A given client use case is likely to be a combination of these clusters. For example, a data integration hub (1) may be a pre-requisite to investigative computing (5).

# Big Data – A Brief History

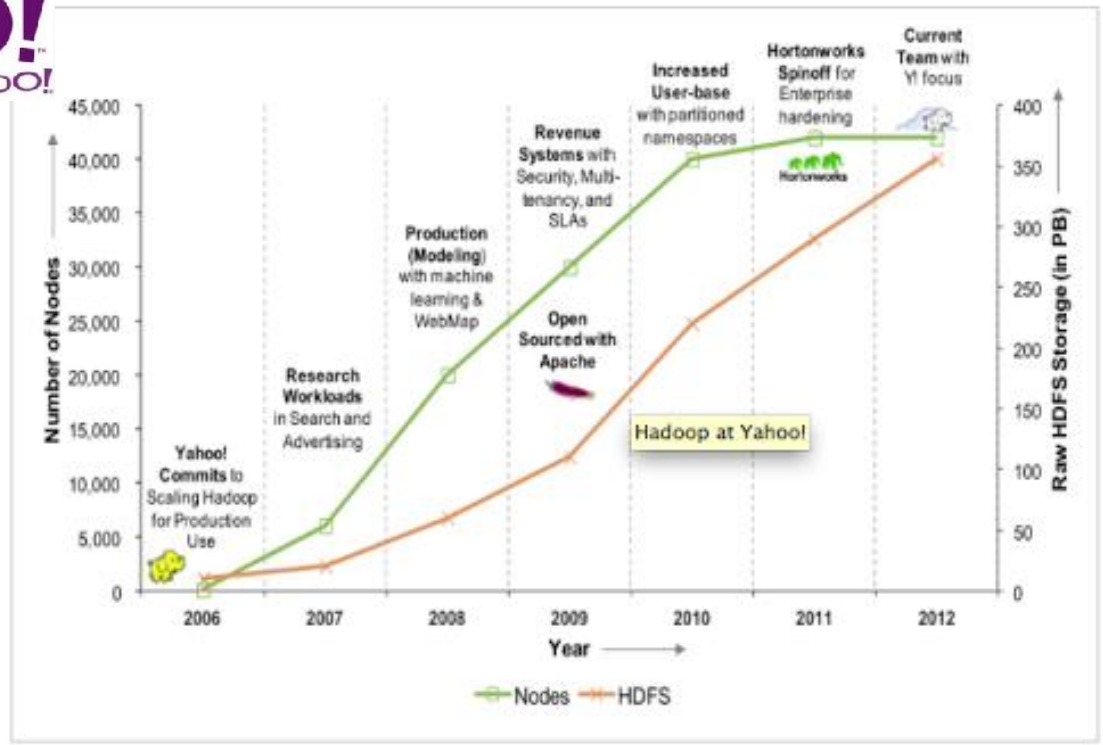
## Big Data technologies have developed over time to meet increasing need to process large data volumes



# Growth of Hadoop – Examples




Source : http://hadoopblog.blogspot.co.uk/2010/05/facebook-has-worlds-largest-hadoop.html




Source: http://nosql.mypopescu.com/post/42441081929/hadoop-at-yahoo-2013-update

**2009: 10 & 28 node clusters**
**2010: Hundreds node cluster / multi PB**
**2011: Thousands node cluster(s) / 10s PB**

Source: eBay's Hadoop Stack: Evolution and Revolution, Juhan Lee, ebay

"There's this wonderful technology at Google. I would love to be able to use it but I can't because I don't work at Google. There are probably a lot of other people who feel that same way, and open source is a great way to get technology to everyone.".

*I've always loved open source because it's such a tremendous lever.*

What I look for is a way to find the smallest thing I can do, with the least amount of work that will have the most impact. Where is the leverage point?

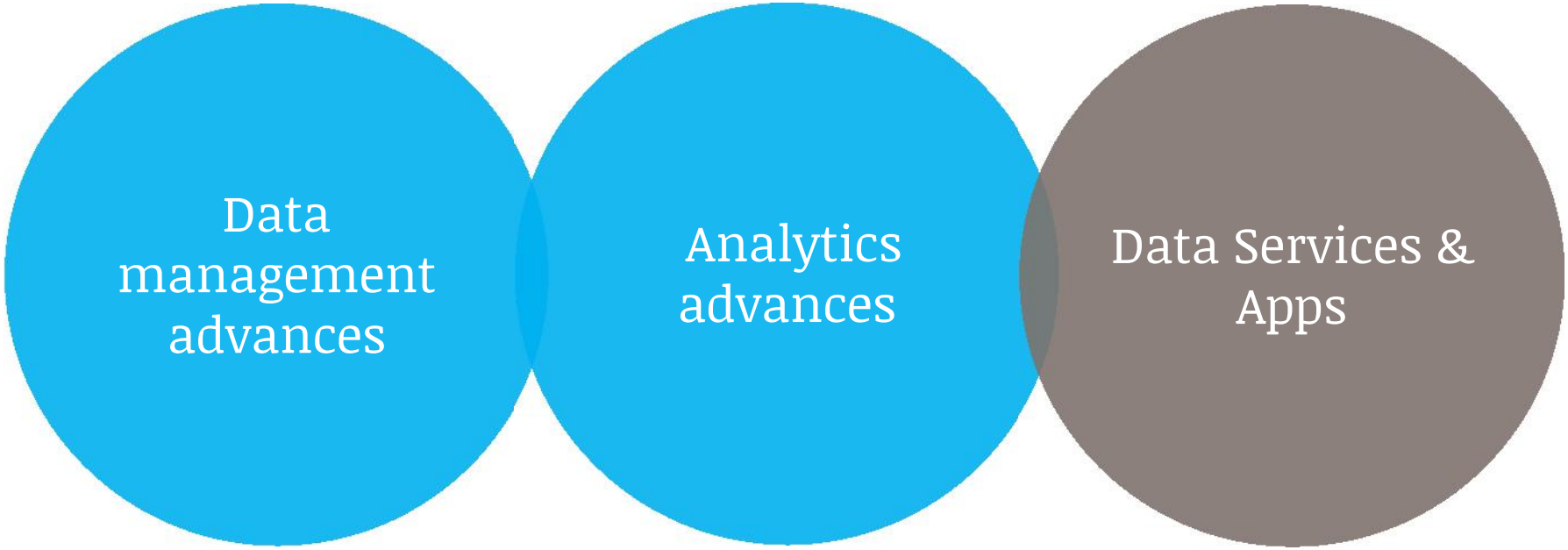
Hadoop came out of that. We needed to do some vast computing, but I also saw a lot of other workloads that could benefit from this.

**-Doug Cutting, Director Apache Software Foundation (2006)**

Source : Forbes http://www.forbes.com/sites/netapp/2013/01/16/big-data-hadoop-doug-cutting/

# Big Data Advances

Real value comes from taking advantage of these advances to develop data products and services.



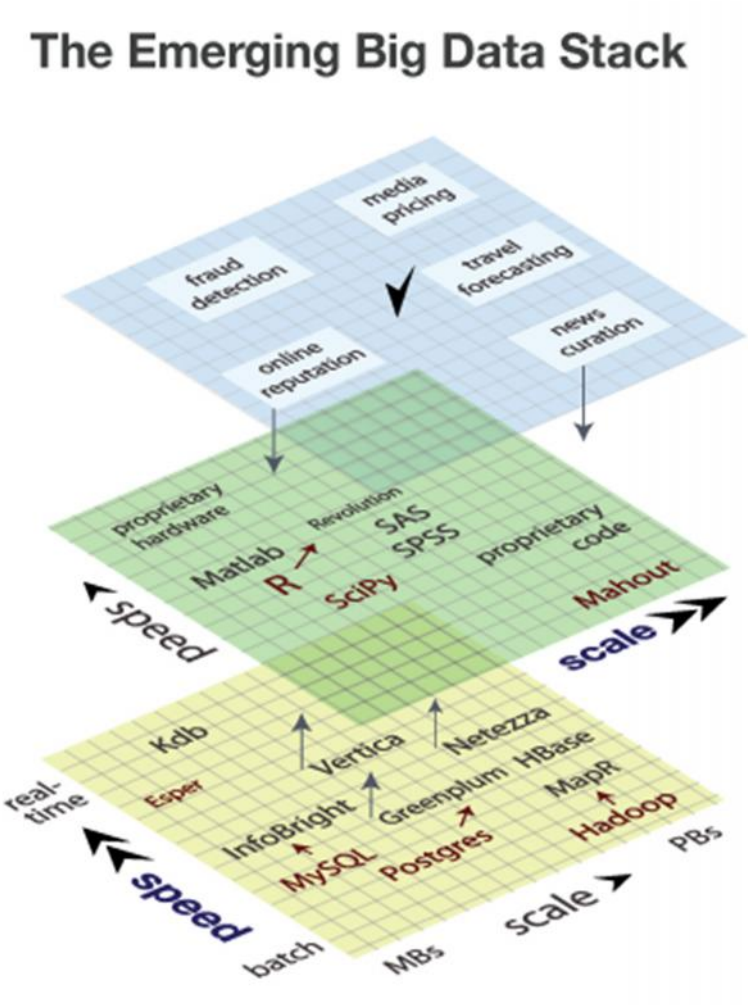

Big, fast data

Big, better analytics

Real business value

# Emerging Big Data Reference Stack

- As the **foundational layer** in the big data stack, the cloud provides the **scalable persistence and compute power** needed to manufacture data products.
- At **the middle layer of the big data stack is analytics**, where features are extracted from data, and fed into classification and prediction algorithms.
- Finally, **at the top of the stack are services and applications**. This is the level at which consumers experience a data product, whether it be a music recommendation or a traffic route prediction.




Data Services & Apps

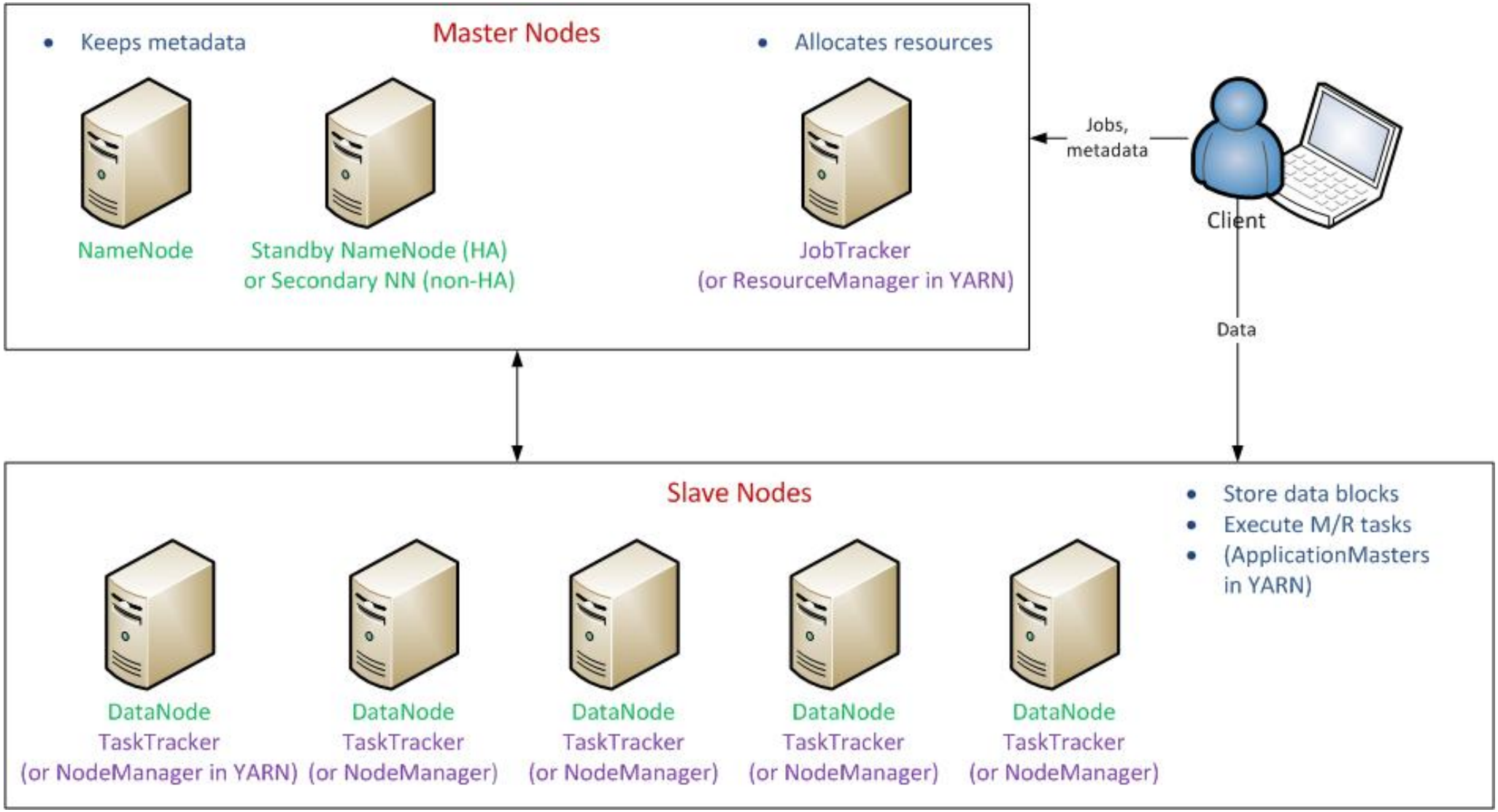
Big, better analytics

Big, fast data

Source: O'Reilly Strata

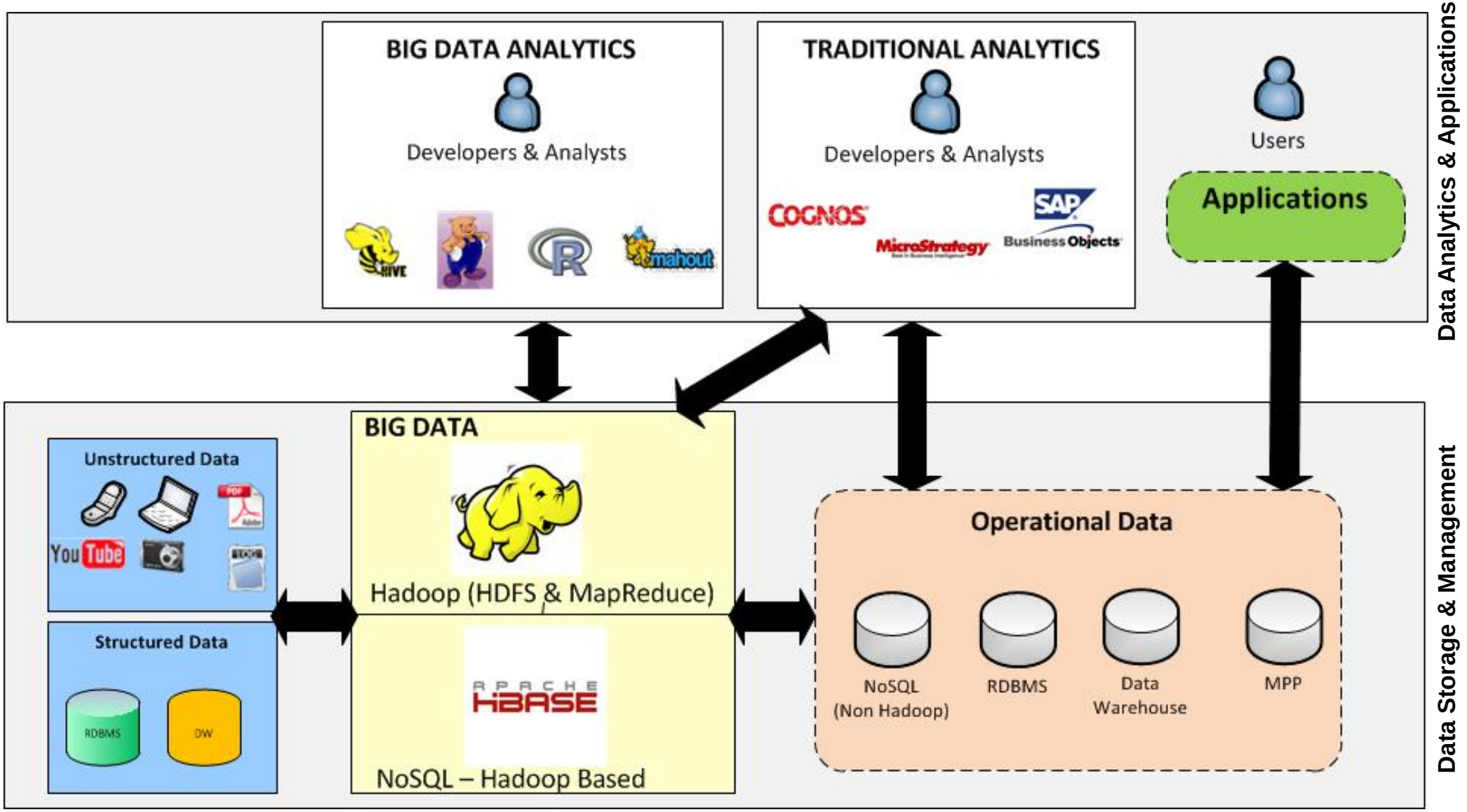
# Core Hadoop Architecture

Highly scalable, distributed, fault tolerant Architecture running on off the shelf hardware; two core components – **HDFS** & MapReduce



# Hadoop within existing Enterprise Architecture

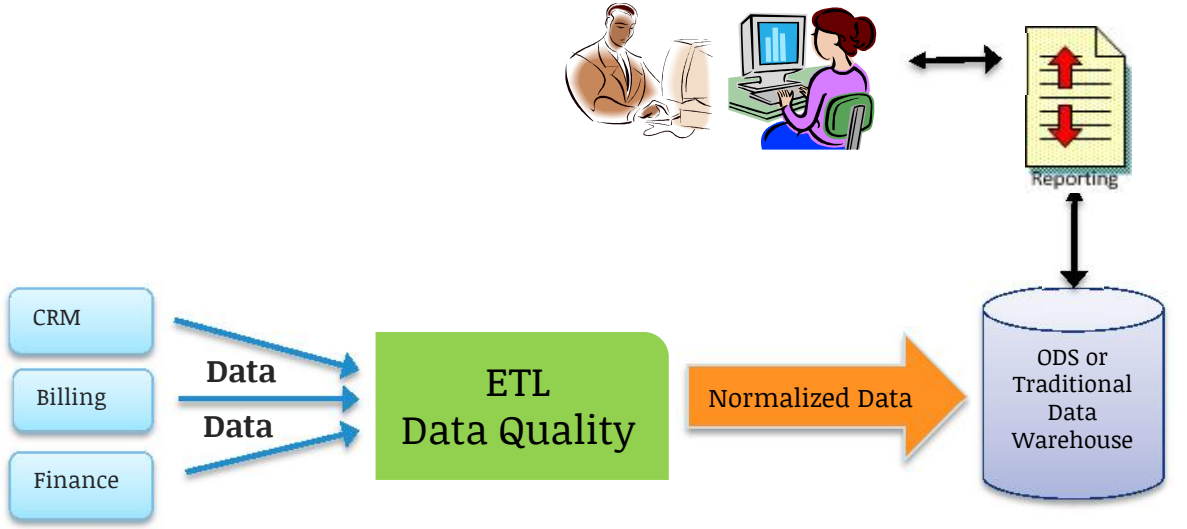
We see Hadoop as complimentary to current Enterprise Architecture to extract value from newer data sources both structured and unstructured



# A Simplified Comparison

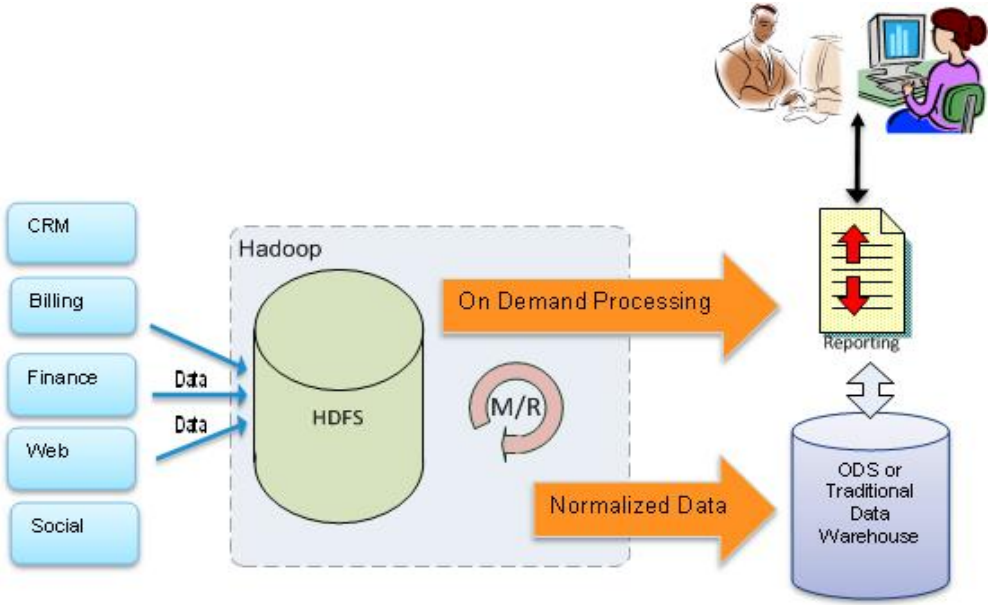
**The 'Old' way:**

- Store ***some*** of the data.
- Process and analyze ***some*** of the data.
- Setup ***specific*** schemas and queries.
- ***Huge effort*** when schemas have to change.

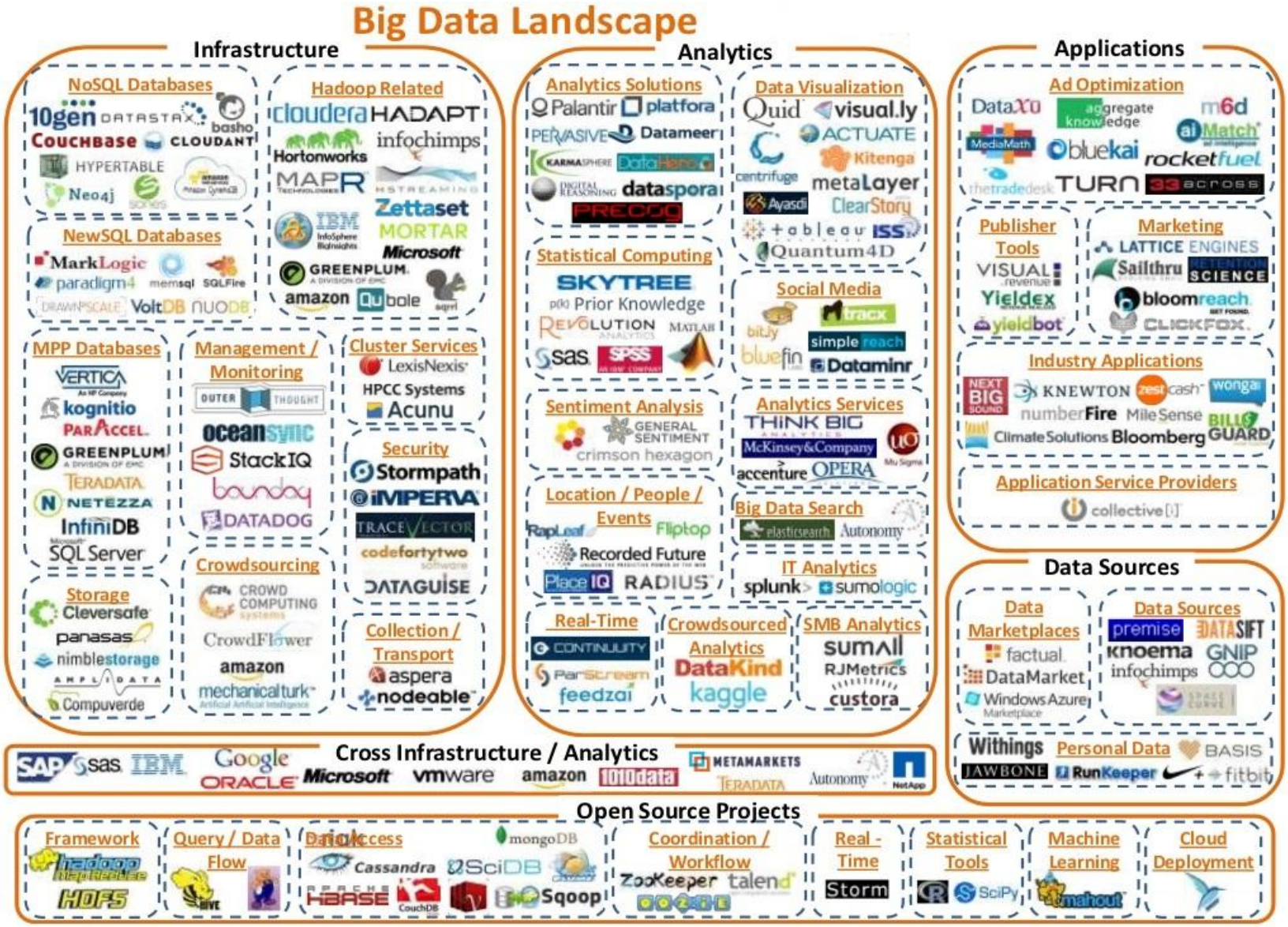



**The 'Big Data' way:**

- Store ***all*** the data you want.
- Process and Analyze ***all*** Your Data.
- Ask ***new*** Questions for further analysis.
- Ask ***more*** Questions
- Get Answers ***faster***
- Get ***clearer*** Insight
- ***Make better business decisions***



# Evolving Big Data Landscape




Source: Bloomberg Ventures (Matt Truck @matttruck & Shivon Zillis @shivonz)

# Ecosystem

## Our practical experience with an ever-evolving ecosystem and our vendor independence help us to quickly define a tailored solution that is 'right' for our clients.

| | |
|---|---|
| Hadoop infrastructure | hadoop, cloudera, Hortonworks, MAPR TECHNOLOGIES |
| Big data management | Flume, HIVE, OOZIE, Spark, sqoop, talend |
| NoSQL & other databases | APACHE HBASE, Cassandra, CouchDB, mongoDB |
| Big data search | Lucene, Apache Solr |
| Analytic discovery | HIVE, mahout, R, Shark, vectorwise |
| Analytic visualisation | tableau, JASPERSOFT, pentaho |
| Cloud platforms | amazon web services, ElasticHosts, rackspace |
| Systems integrators | OSSCube The Open Source People |
| Business consultancies | northhighland. WORLDWIDE CONSULTING, radbourne |
| Sector & point specialists | adqura, Lifecode |

(This is a highly simplified characterisation of the landscape, as solution stacks often cut across multiple categories.)

Introduction

Big Data & Apache Hadoop

**MapR / Drill Demo**

Summary & Further Learning

Introduction

Big Data & Apache Hadoop

MapR / Drill Demo

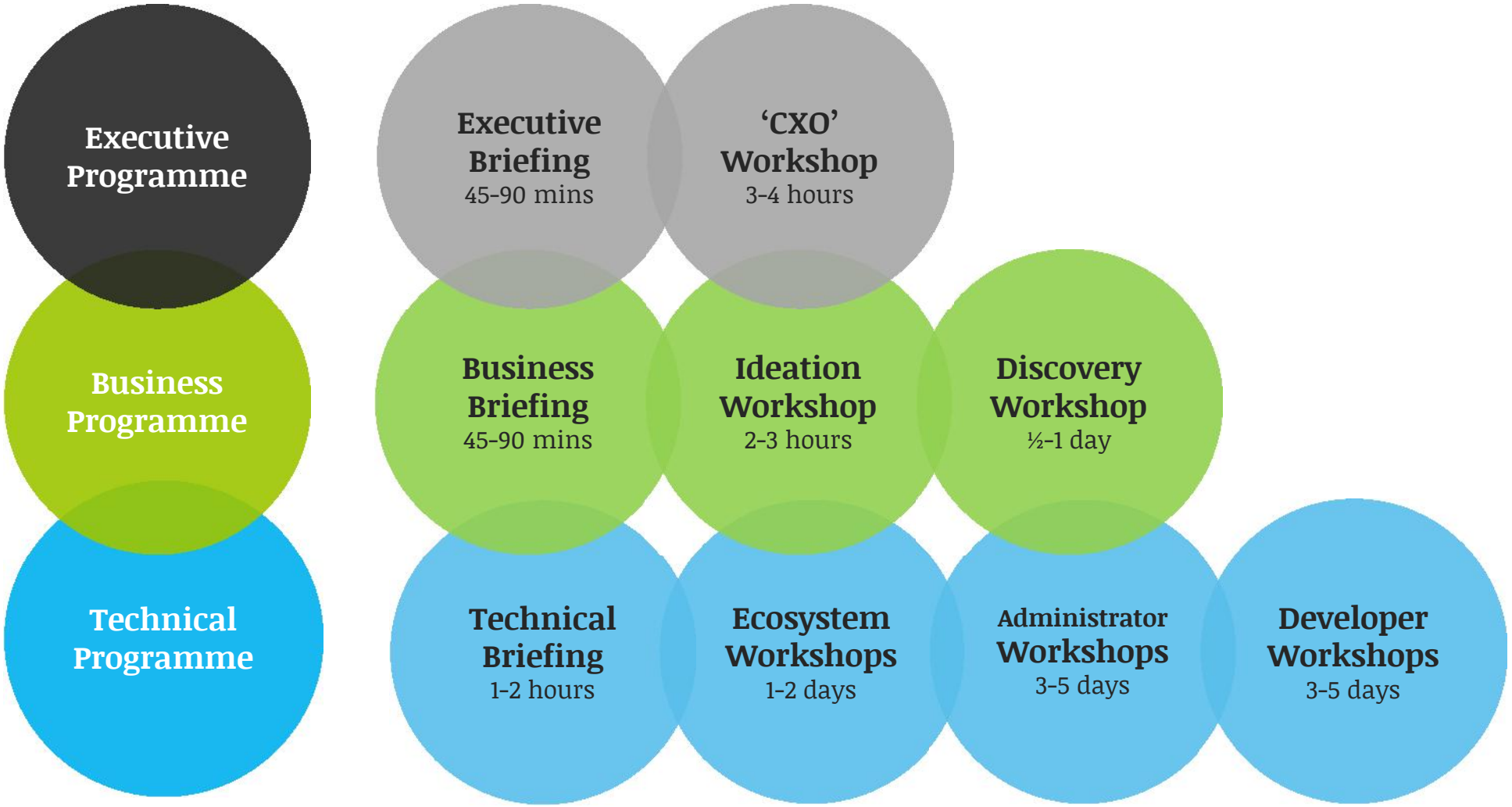
**Summary & Further Learning**

# Summary / Wrap Up

- There is more to Big Data than the hype
- Many of the advances are powered by Open Source
- The ecosystem around Big Data is evolving rapidly
- Most organisations can experimentation with Big (and small) Data
- MapR have built a unique offering on top of Open Source

# Further Learning

Learning can take many shapes, depending on organisational and individual needs.



# Cheatsheet & Mythbuster

onepointiq

Academy

## The Big Data Cheatsheet

We at Onepoint IQ like to educate and empower our clients to cut through the noise around 'Big Data'. We hope you find this primer a useful starting point.

**What is Big Data?**

A simple definition of Big Data is a dataset that is so large that it cannot be managed with traditional information technology tools. However, the term is more broadly used to describe a data-driven economy or society. This is fuelled by an array of technology advances that can handle data in new ways and deliver business insights from this data.

**Where did the term come from?**

The term 'Big Data' has been used in academia and in some technology circles for at least two decades. Some attribute the term to John Mashey, chief scientist at Silicon Graphics in the 1990s. More recently in 2011, McKinsey Global Institute, the research arm of McKinsey & Co, released a report titled "Big Data: The Next Frontier for Innovation, Competition and Productivity." This caught the imagination of businesses and media, widely popularising the term.

**How big is big?**

When people speak of Big Data in terms of volume, they are usually talking about data in the petabyte (1 million gigabyte) to exabyte (1 billion gigabyte) range. What constitutes 'big' varies by perspective and will certainly change over time. This is why at Onepoint IQ, we stay focused on data—big or small.

Big Data is commonly defined in terms of the three V's: volume (how much data), velocity (how fast the data is generated or processed) and variety (the different types or formats of data). We believe this is often an over-simplification. A small volume of complex data, a huge volume of simple data or sophisticated analytics and predictions from any of your data can still benefit from the technology advancements.

**Why is there so much excitement about Big Data?**

The technology advances surrounding Big Data, which were designed to handle large, diverse, complex and growing datasets, can be applied to all sorts of data management challenges (use cases) by all sorts of organisations. One exciting outcome is having the right business information (or insights) available at the right time. In some cases, the insights would not have been possible before due to the challenges of dealing with different types of data.

**What is unstructured data?**

It is data with unpredictable structures, such as documents, e-mails, blogs, social media chatter, digital images, videos, and satellite imagery. Contrast this with data stored in 'traditional' databases: they have predefined formats (or structures), such as customer records and point of sale data. The key point is this: in 'traditional' data stores, the data format (or schema) is defined before the data is stored. With unstructured data, this luxury is not available.

**What is Hadoop?**

Hadoop is essentially software that Google, Yahoo!, Facebook and others use to process vast amounts of data. The name comes from one the creators, whose son's elephant toy was called Hadoop.

Hadoop originated from search technology invented by Google. It is now an open source software project run by a non-profit foundation called Apache.

In technical terms, Hadoop is a framework for processing and querying vast amounts of disparate data. This unstructured data is stored on large collections (or 'clusters') of commodity hardware.

In the past, only large corporations, academia and government organisations could afford the computing capability to process large and complex data. With Hadoop, smaller organisations and even individuals have access to the power of distributed computing.

Because Hadoop software can detect and handle failures at the application layer, expensive high-end hardware is not needed to give resilience for the data.

**What is HDFS?**

Hadoop Distributed File System or HDFS is a versatile way to automatically break data into smaller pieces and store these across multiple nodes (servers). When this data needs to be processed, multiple computations can be run simultaneously (asynchronously), on the smaller pieces.

**What is MapReduce?**

MapReduce is a software programming model for distributed computing. A software developer can use this model to write code—typically in the Java programming language—to process massive amounts of multi-structured data in parallel across hundreds or even thousands of nodes (servers), such as a Hadoop cluster. MapReduce was designed by Google as a way of efficiently executing a set of functions against a large amount of data in batch mode.

The 'map' part distributes the programming problem or tasks across the servers. After the distributed computation is completed, another function called 'reduce' aggregates all the 'answers' back together to provide a result.

The model manages most of the complexities of distributed computing, which makes it easier for the developer to focus on getting 'answers' or insights from your data, rather than getting bogged down with structures (schema), assembling, disassembling data, managing workloads of each server and recovery from failures.

For the fuller, up-to-date Big Data Cheatsheet, visit academy.onepointiq.com, where you can also sign-up for our programme of noise-removal sessions.

onepointiq

Academy

## The Big Data Mythbuster

A client called us the Big Data Mythbusters. We liked that and thought we'd compile this starter list of common myths.

**Big Data is a technology**

Big Data is not one technology. The term is used as a 'catch-all' for an array of technology advances, including new ways of acquiring, manipulating, storing, making sense, making predictions from and visualising data.

**Value creation from Big Data only applies to large organisations with lots of data**

Many of the 'Big Data' technology advances were invented and applied by organisations with huge quantities of data such as Google, Yahoo! and Facebook. Yet, they are relevant to all types of organisations.

**Big Data is all hype**

Big data is a relatively new term for one of information technology's oldest trends: the exponential growth of business data. Business data has grown dramatically for the past 40+ years.

The graph shows the growth of disk drive average capacities – from 1MB in 1980 to 1TB in 2010 (that's 10,000,000MB). Data volumes have grown to take up these capacities and then some.


**Big Data is all about volumes of data**

Often Big Data is described in terms of the three V's: volume, velocity and variety. Yes, there is huge volume, as above. This volume is created very quickly (velocity) through a hugely diverse data sources and stored in various formats (variety). These sources include video, audio, all sorts of sensors, social media and enterprise operational data.

**There is a silver bullet solution for Big Data**

Big Data is not a product or solution. It is a series of complex and diverse business (and technical) problems that are addressed with numerous tools and architectures. A single vendor is not able to provide an 'end-to-end' solution, let alone give an independent perspective.

This is why we created a trusted partner and associate ecosystem at Onepoint IQ, allowing us to pull together the multi-disciplinary expertise needed—including business consultants, sector experts, business and technology architects, developers, data scientists, business intelligence experts and data visualisation experts—to harness Big Data advances for the greatest business impact.

**Big Data is open source**

Many 'Big Data' advancements were first invented by commercial organisations. They released it to the open source community and are now advanced by non-commercial entities like the Apache Foundation and an array of companies, both large ones and startups.

**Big Data solutions are a replacement for traditional databases**

More often than not, Big Data advances are complementary to traditional relational database management systems (RDBMS) like those by IBM, Oracle and others. Most organisations have huge investments in traditional databases and these systems are well integrated to operational processes.

Big Data solutions can sit alongside these and help address specific data challenges and opportunities. We see Big Data advancements as particularly well suited for six clusters of use cases (or implementation patterns):

| | | |
|---|---|---|
| 1. Data integration hub | 3. Real-time monitoring & analytics | 5. Investigative computing |
| 2. Analytics accelerator | 4. Near real-time analytics | 6. New markets enabler |

Onepoint IQ Academy's courses go into more details and give plenty of examples to illustrate each cluster.

**Big Data solutions are free**

Some Big Data advancements are available as open source downloads that can be freely downloaded. These often have commercial variants (distributions) with support and enhancements which come with commercial fee models.

See more Big Data Mythbusters at academy.onepointiq.com, where you can also sign-up for our programme of mythbusting lessons.

# Thank you

**Shashin Shah** | Technology Zen Master | Founder Director | shashin@onepointiq.com
**Christy Kulasingam** | Chief Stratnologist | Founder Director | christy@onepointiq.com
**Alexander (Sasha) Polev** | (The Data Professor) Chief Architect | sasha@onepointiq.com
**Michael Hausenblas** | Chief Data Engineer EMEA at MapR Technologies | mhausenblas@maprtech.com