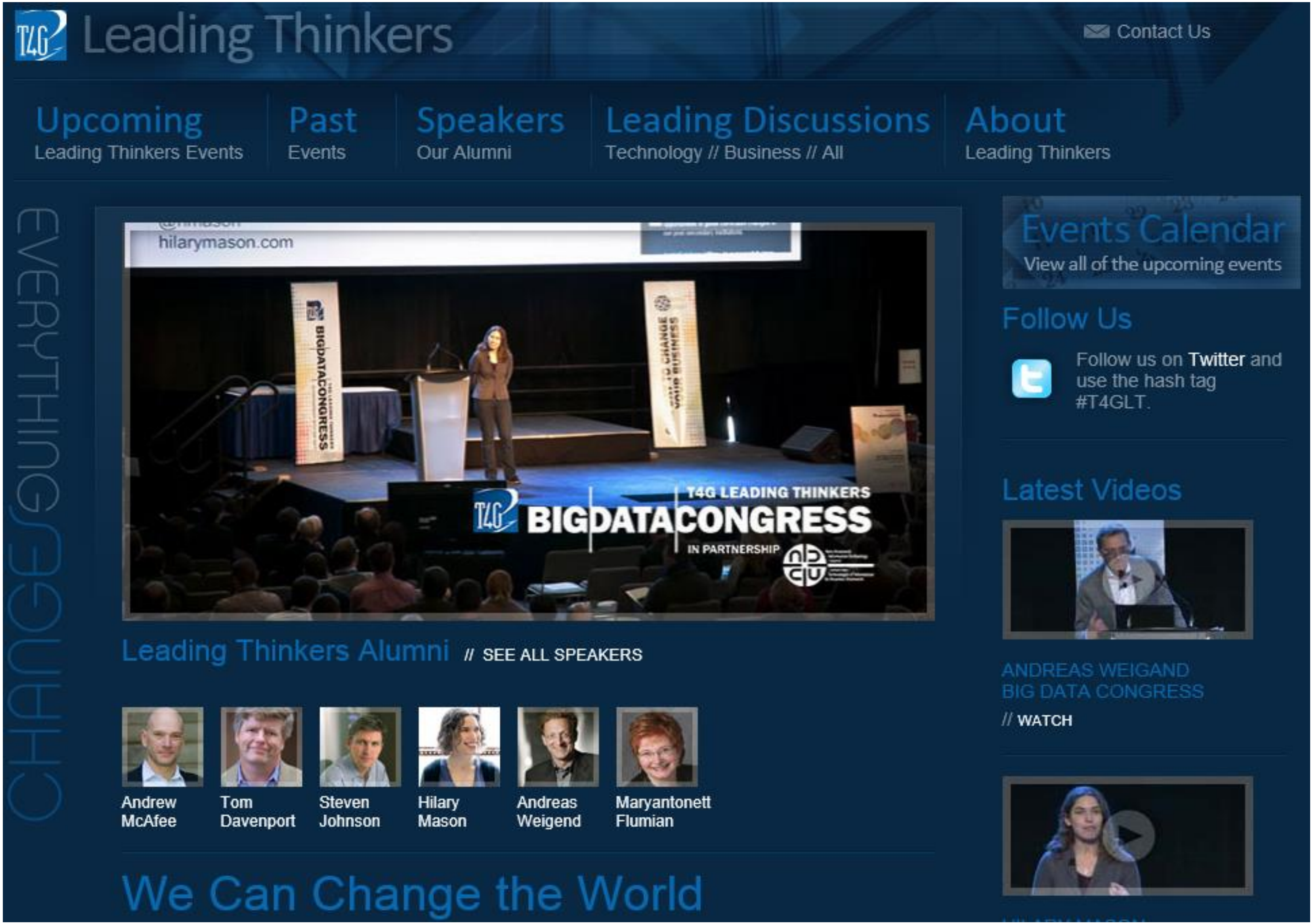
Intro to Big Data

Darryl Dutton

Email: darryl.dutton@t4g.com

GREAT PLACE TO WORK® Best Workplaces 2012 Canada



Leading Thinkers

Contact Us

Upcoming
Leading Thinkers Events

Past
Events

Speakers
Our Alumni

Leading Discussions
Technology // Business // All

About
Leading Thinkers

Events Calendar
View all of the upcoming events

Follow Us

Follow us on Twitter and use the hash tag #T4GLT.

Latest Videos

ANDREAS WEIGAND
BIG DATA CONGRESS
// WATCH

Leading Thinkers Alumni // SEE ALL SPEAKERS

- Andrew McAfee
- Tom Davenport
- Steven Johnson
- Hilary Mason
- Andreas Weigend
- Maryantonett Flumian

We Can Change the World

Agenda and Goal

Continue the conversation to the next level

- Understand the characteristics of Big Data
  - Why is Big Data different
  - Common Scenarios
  - Key Ingredients
- How to address Big Data problems
  - Storage
  - Processing
  - Platform
  - Coding Pattern

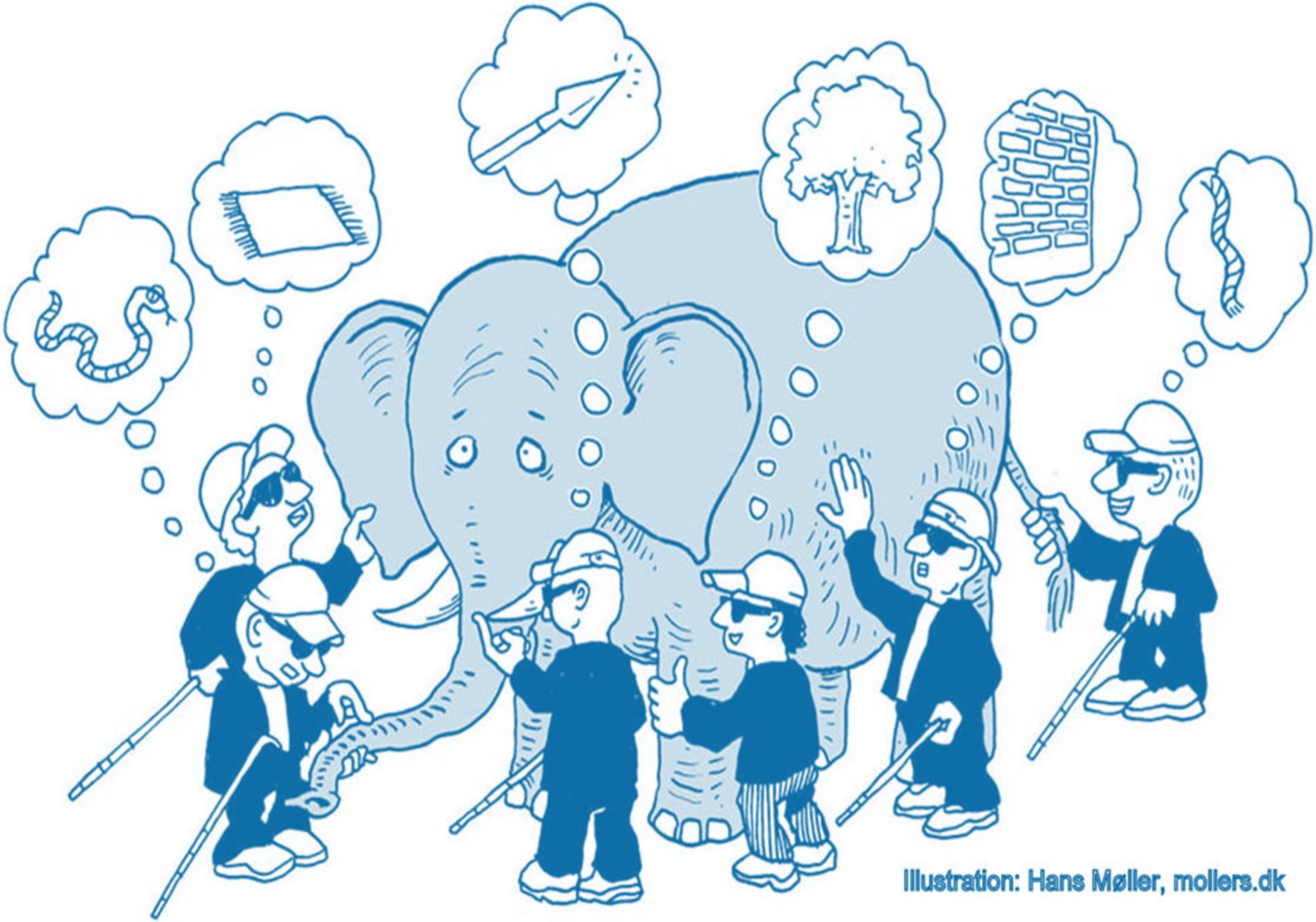
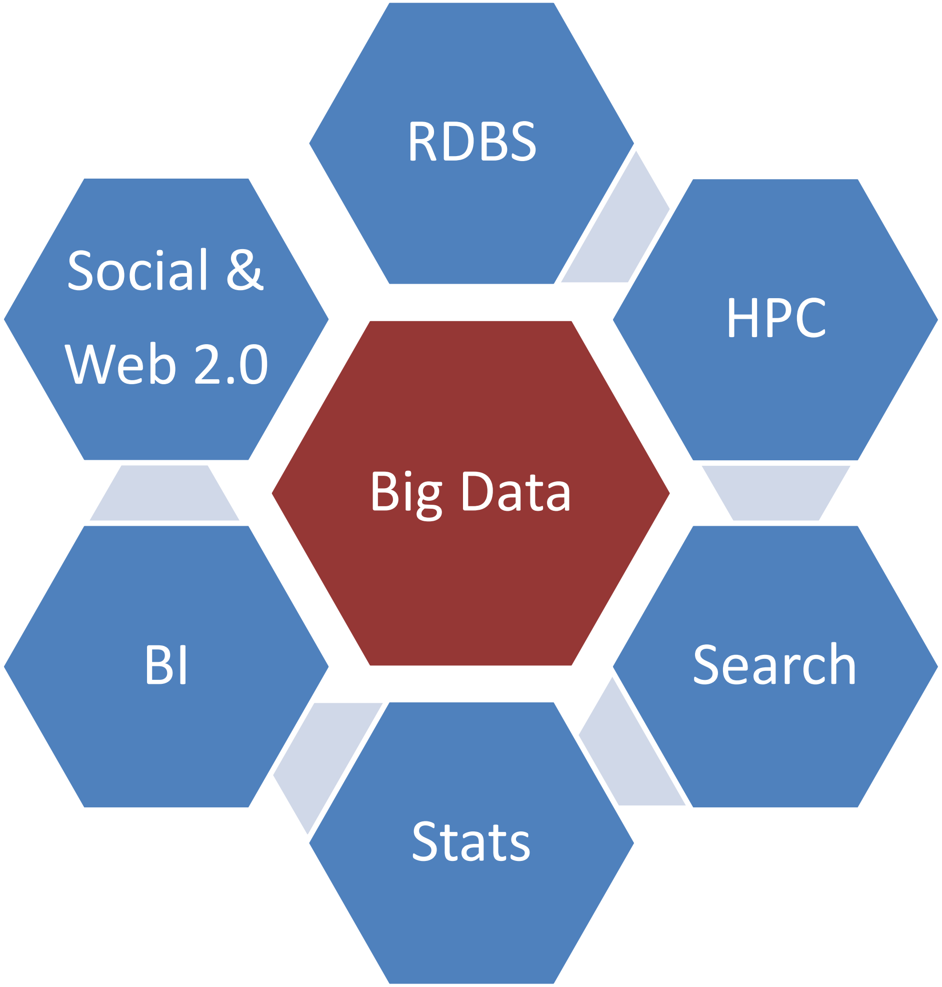


Illustration: Hans Møller, mollers.dk

Six Perspectives



Why is Big Data different?

How big is big....

- Users will generate over 300Gb per year
- Predicate to have about 8 ZETTABYTES data by 2015
  - 8,000,000,000,000,000,000,000 - 10^{21}

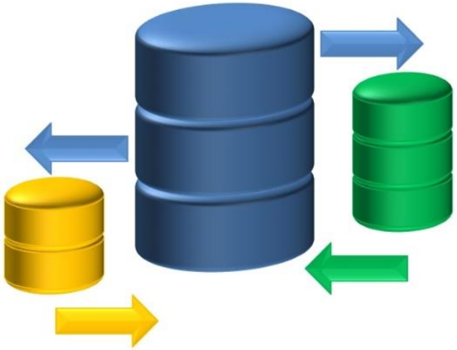
Why is Big Data different?

The allure of finding useful information…

Why is Big Data different?

I have my favorite hummer…

- The tools and approaches we know and love become awkward in this scale of data
- It is difficult to capture store, search, share and analyze at a certain point
- Most of this data is unstructured and it comes in oceans and streams
- More exploratory in nature instead of structured



Why is Big Data different?

The data sets and collection process is different

- Look outward to leverage external data
- “Collecting data first and ask questions later”
- Applying schema and context later in the process

The Characteristics of Big Data

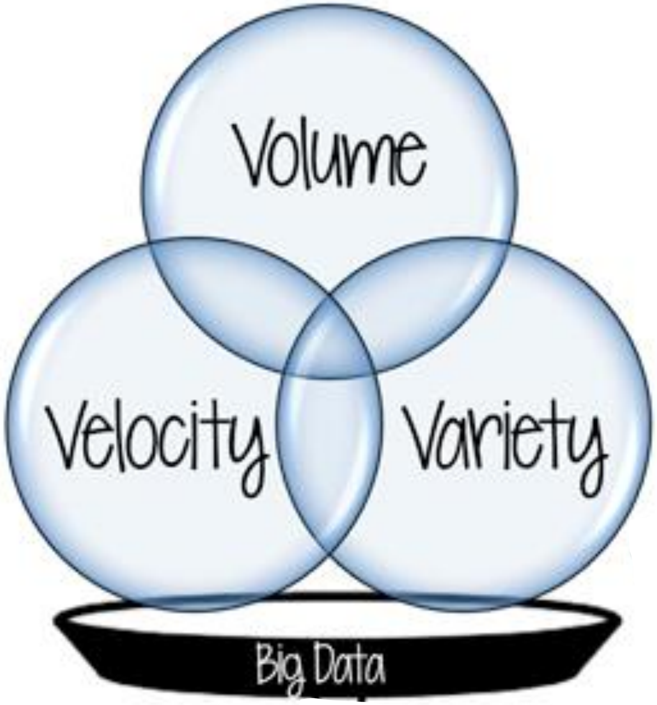
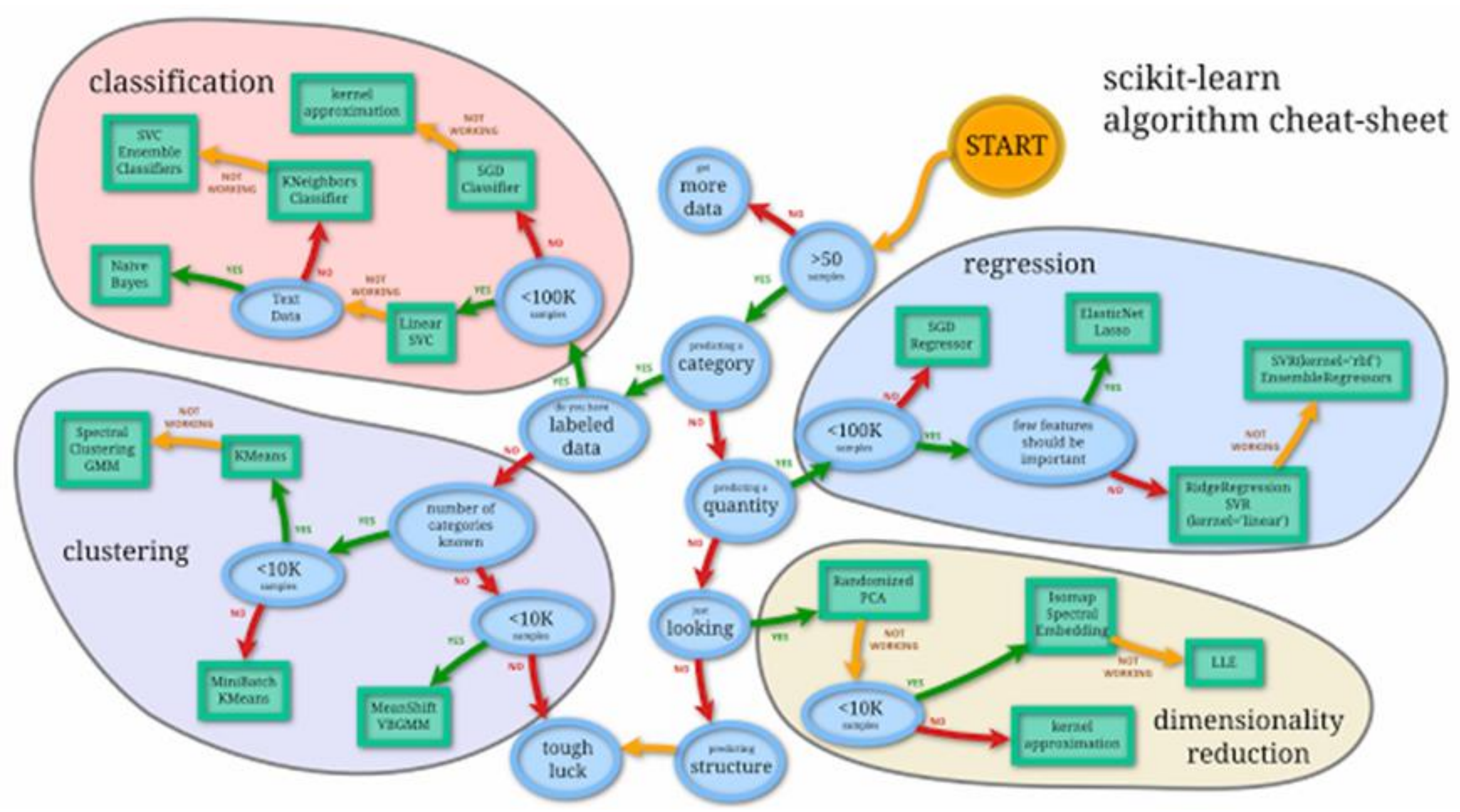


Image source: DamFoundation.org

Some of this is teaching old dog new tricks….

Source: http://www.thebigdatainsightgroup.com

Big Data is exploratory by nature….

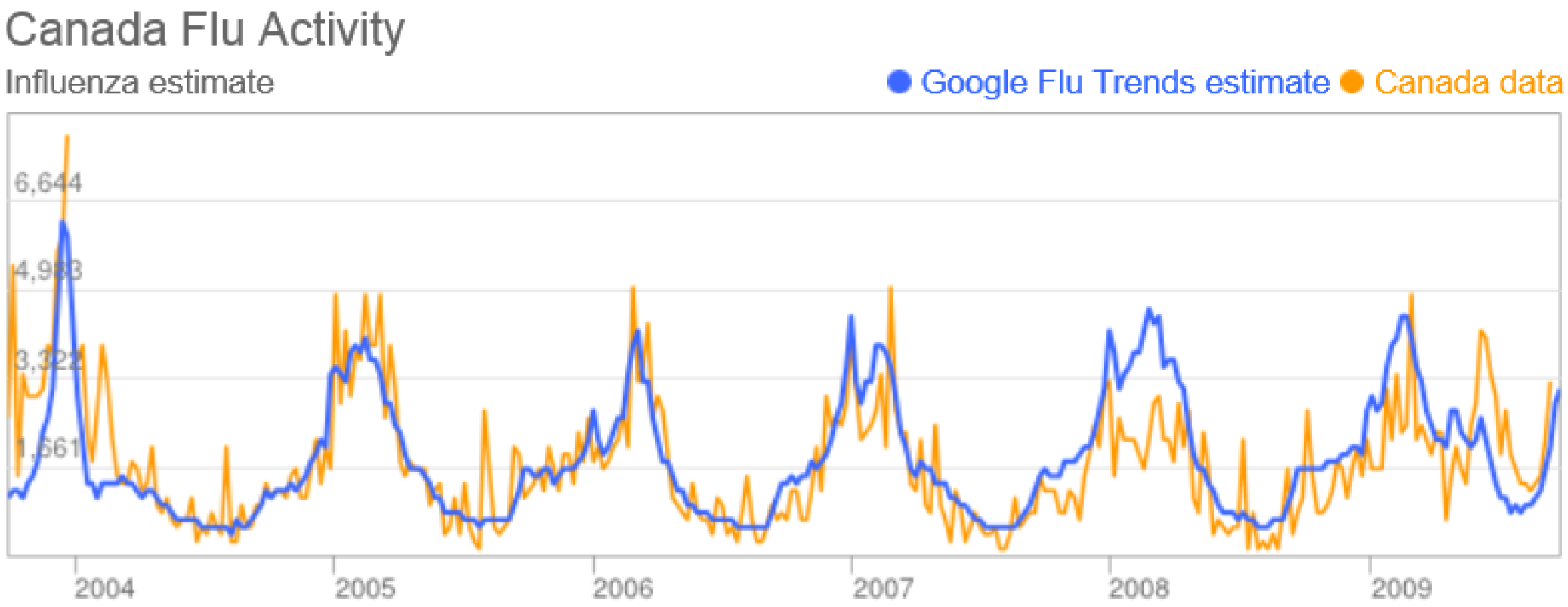


Source: Nishant Chandra http://n-chandra.blogspot.ca/

Common Scenarios

- Behavioral Analysis
  - customer churn
  - trending
- Sentiment Analysis
  - social media analysis
  - multi channel analysis
- Recommendation Engines
  - cross-sell
  - up-sell
- Fraud Detection
  - clickstream analysis
  - mining
- Risk Mitigation
  - asset portfolio analysis
  - transactional log analysis
- Root Cause
  - network log
  - sensor data analysis
- Marketing Effectiveness
  - campaign ROI
  - Ad targeting

Common Scenarios

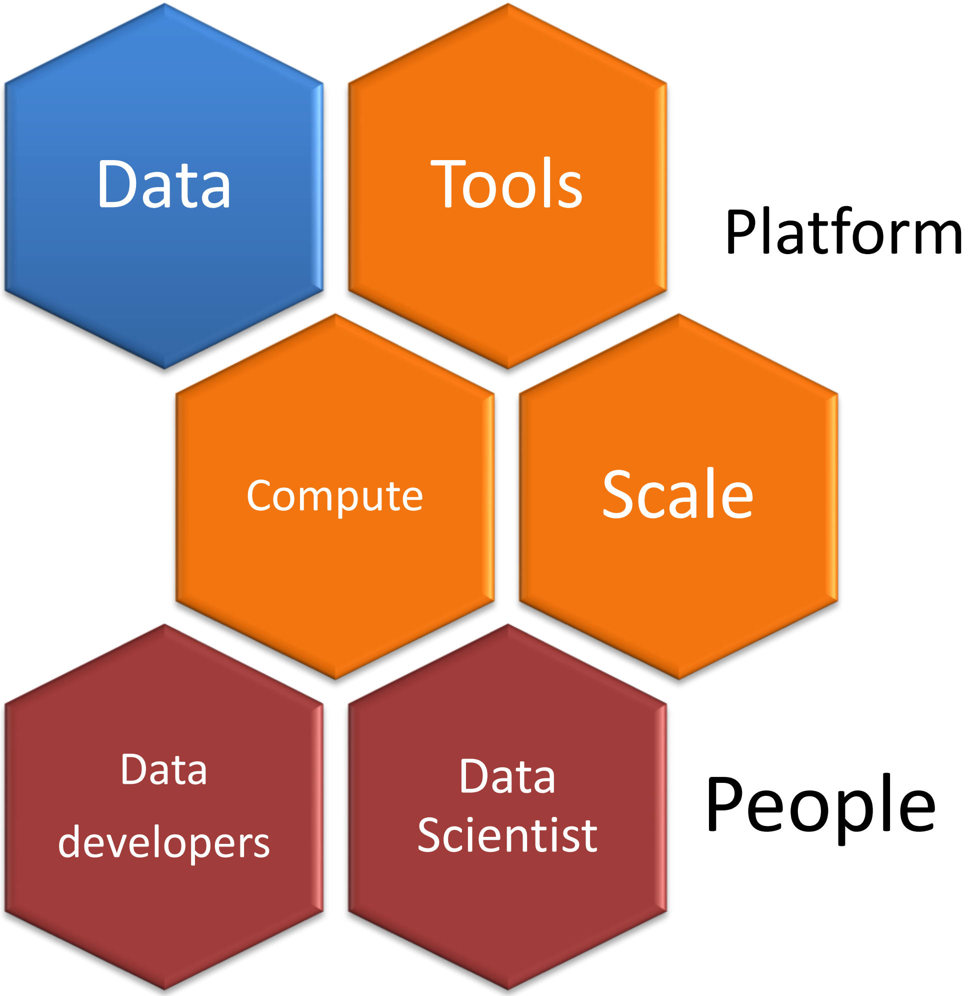


Canada Flu Activity

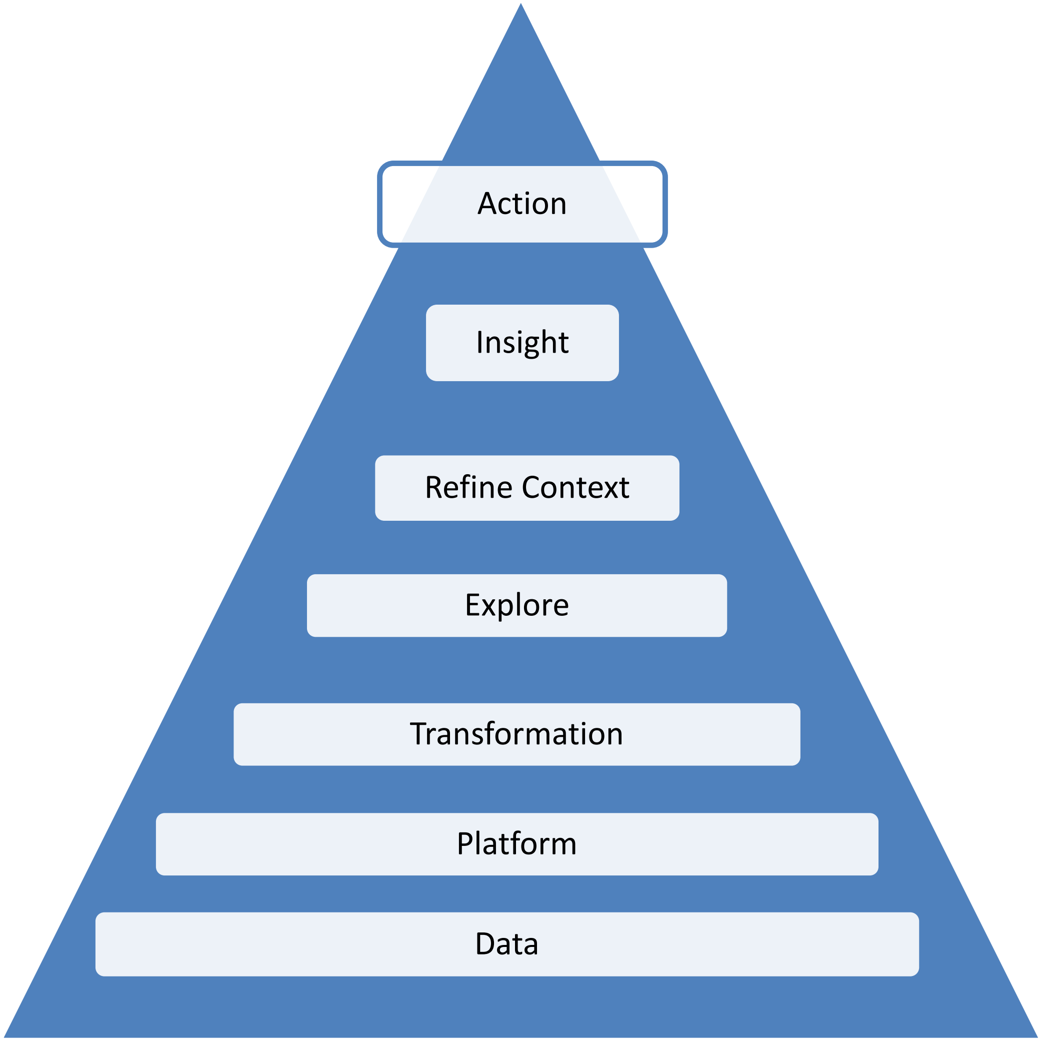
Influenza estimate

Canada: Influenza-like illness (ILI) data provided publicly by the Public Health Agency of Canada.

Key ingredients for a Big Data project



Big Picture, it is about getting to action…



Action

Insight

Refine Context

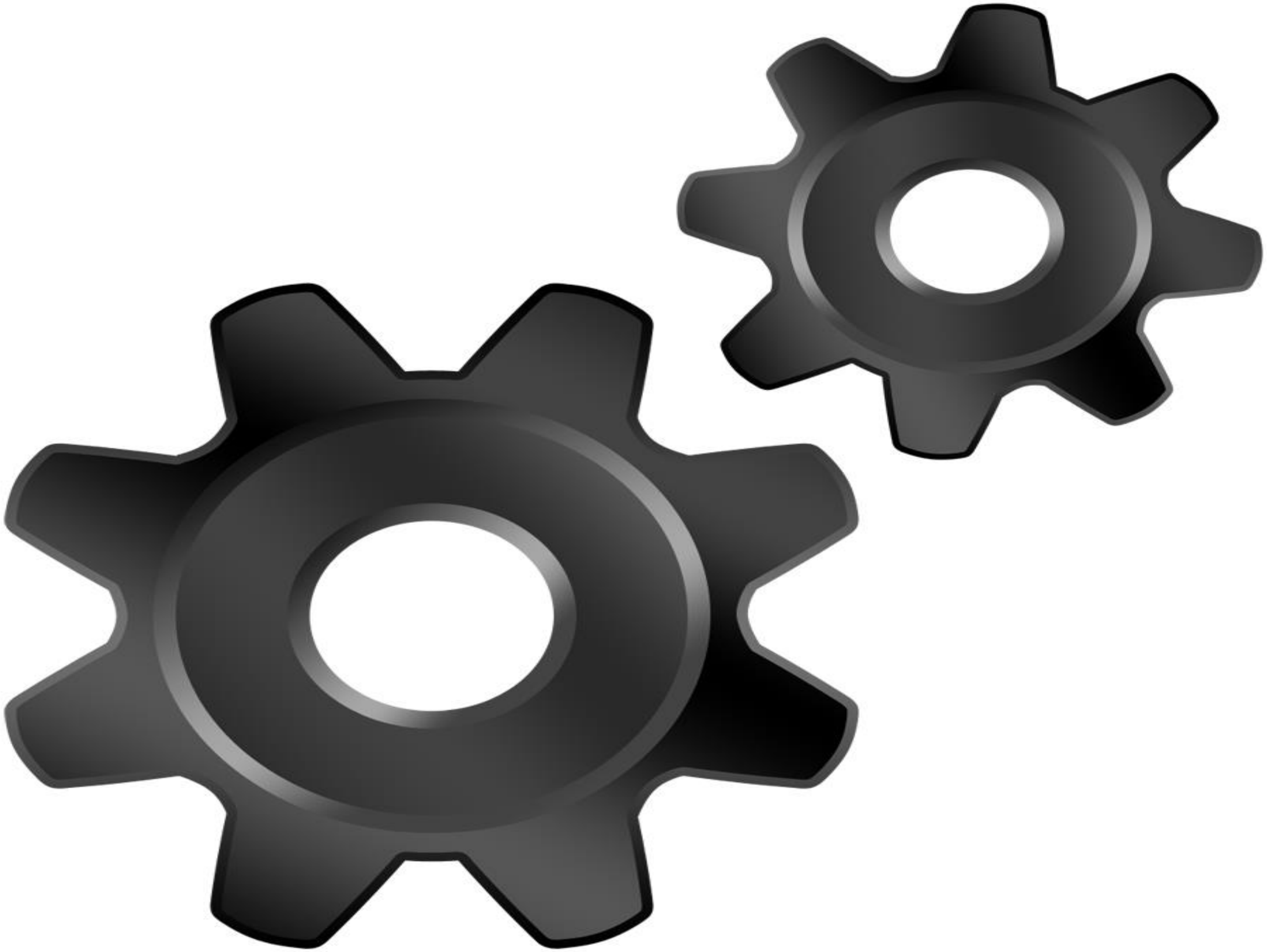
Explore

Transformation

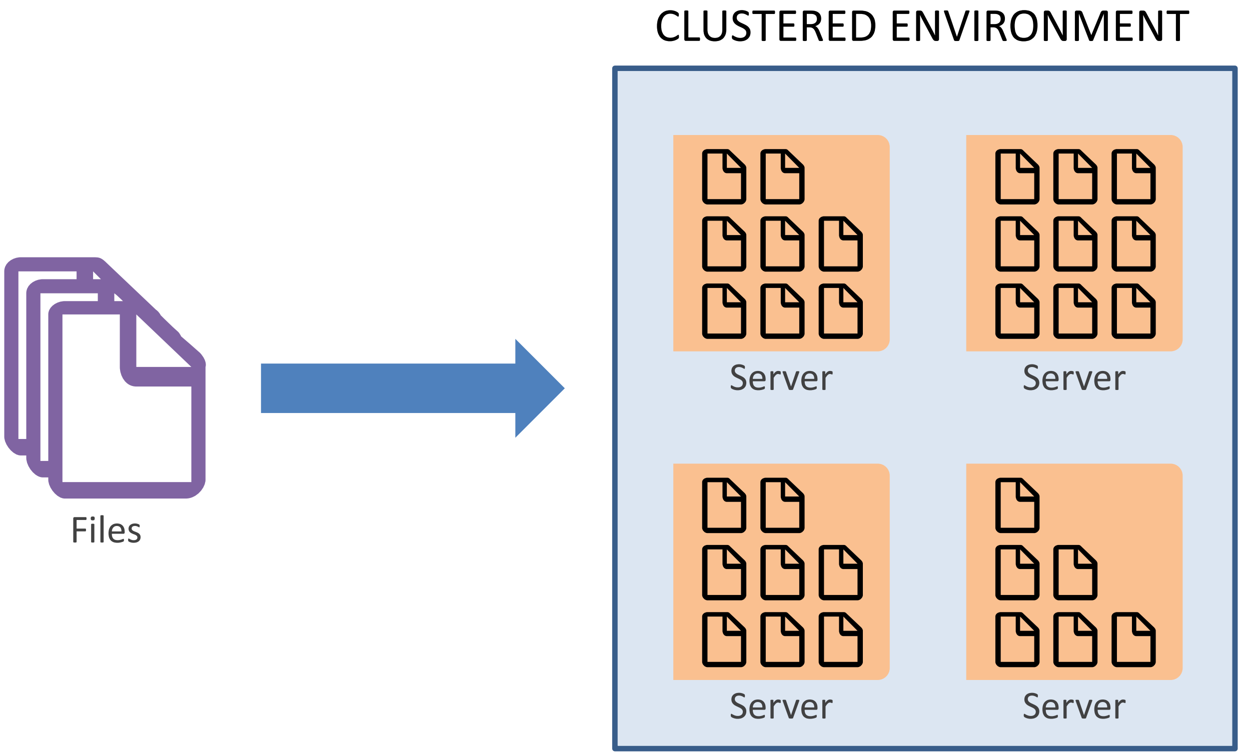
Platform

Data

Let’s switch gears and focus on the ‘how’



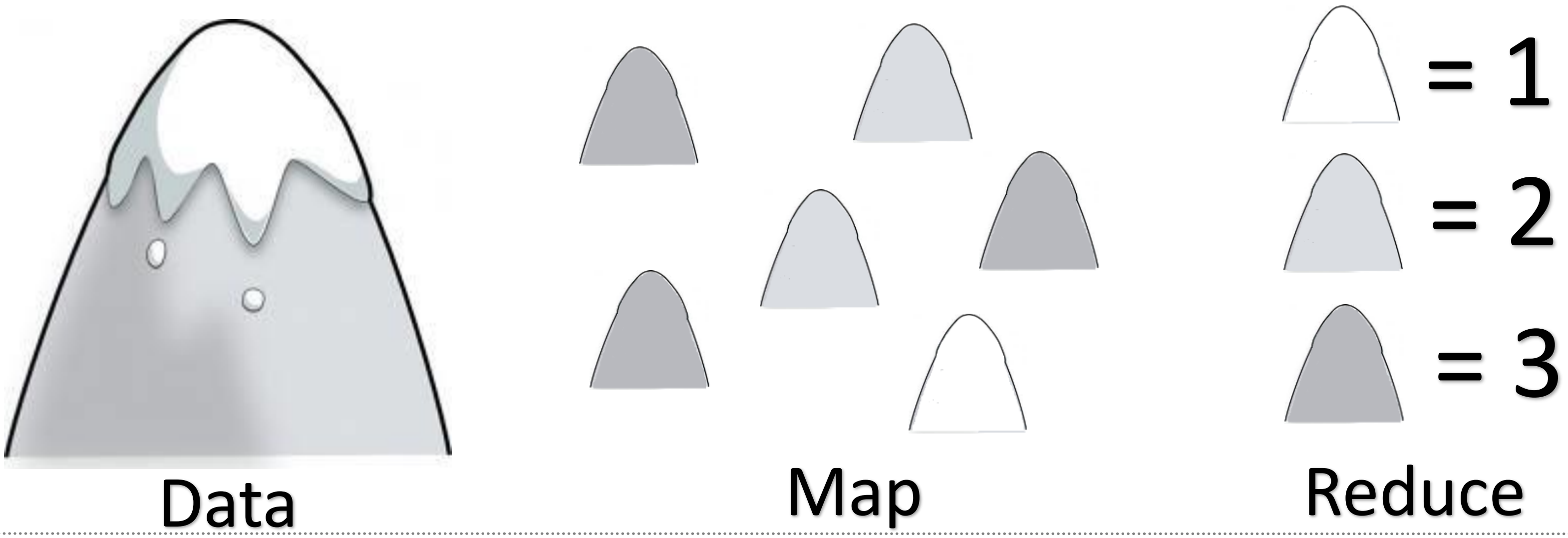
How do we store big data?



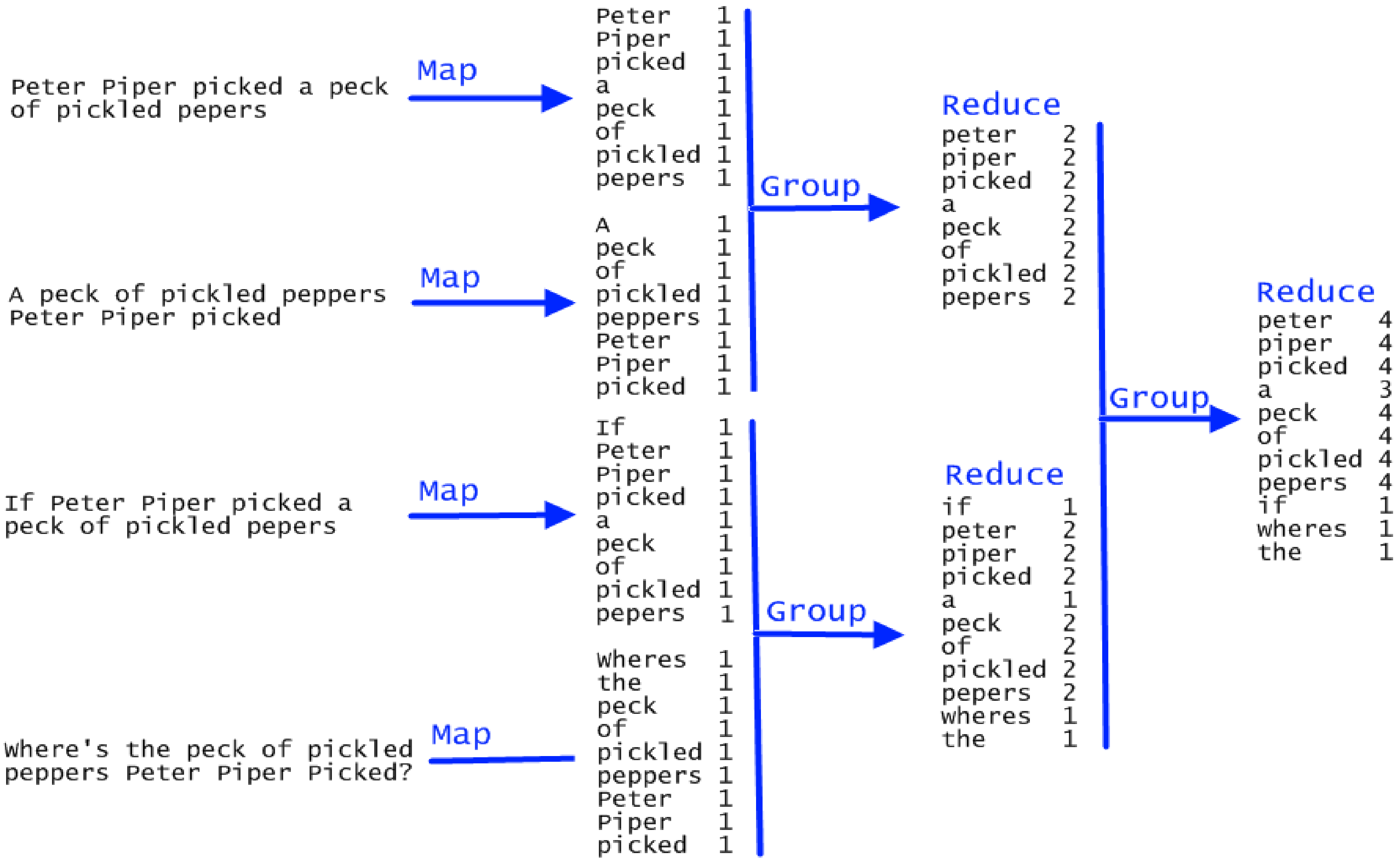
How do we process big data?

Map and Reduce - Simple ‘divide and conquer’

- Help simplify the complexities of analyzing data
- Data structure is based on <key , value> pairs
- Map applies a function to each element in the list
- Reduce applies a function to elements with the same key to aggregate back to a smaller list



Simple MapReduce example – How many words?



Peter Piper picked a peck
of pickled pepers

Map

Peter	1
Piper	1
picked	1
a	1
peck	1
of	1
pickled	1
pepers	1

A peck of pickled peppers
Peter Piper picked

Map

A	1
peck	1
of	1
pickled	1
peppers	1
Peter	1
Piper	1
picked	1

If Peter Piper picked a
peck of pickled pepers

Map

If	1
Peter	1
Piper	1
picked	1
a	1
peck	1
of	1
pickled	1
pepers	1

Where's the peck of pickled
peppers Peter Piper Picked?

Map

Wheres	1
the	1
peck	1
of	1
pickled	1
peppers	1
Peter	1
Piper	1
picked	1

Group

Reduce

peter	2
piper	2
picked	2
a	2
peck	2
of	2
pickled	2
pepers	2

Group

Reduce

if	1
peter	2
piper	2
picked	2
a	1
peck	2
of	2
pickled	2
pepers	2
wheres	1
the	1

Group

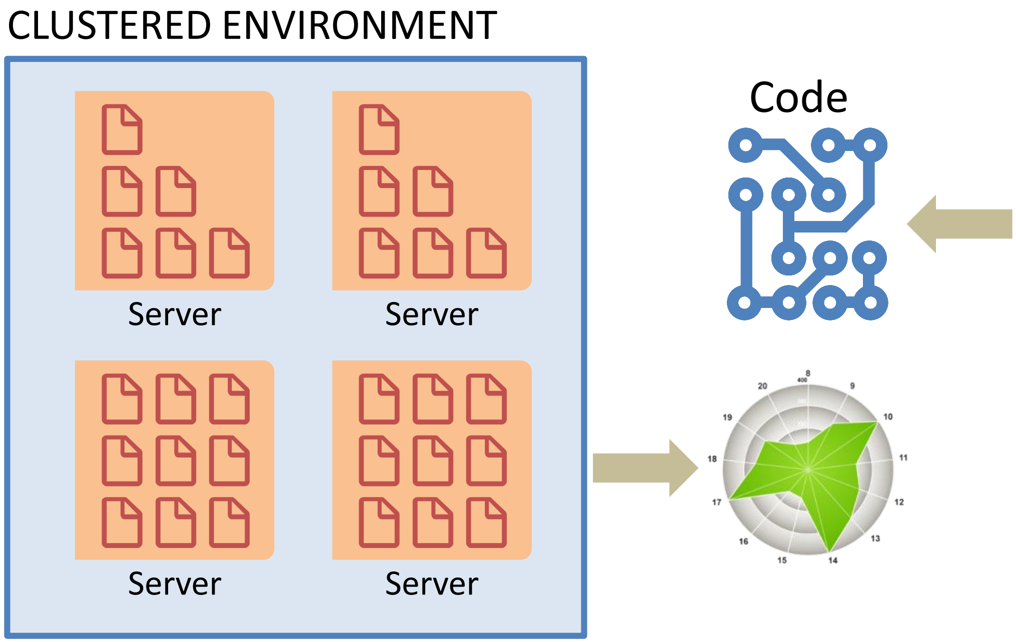
Reduce

peter	4
piper	4
picked	4
a	3
peck	4
of	4
pickled	4
pepers	4
if	1
wheres	1
the	1

Source: http://tarnbarford.net/journal/mapreduce-on-mongo

How do we process big data?

Running our code….



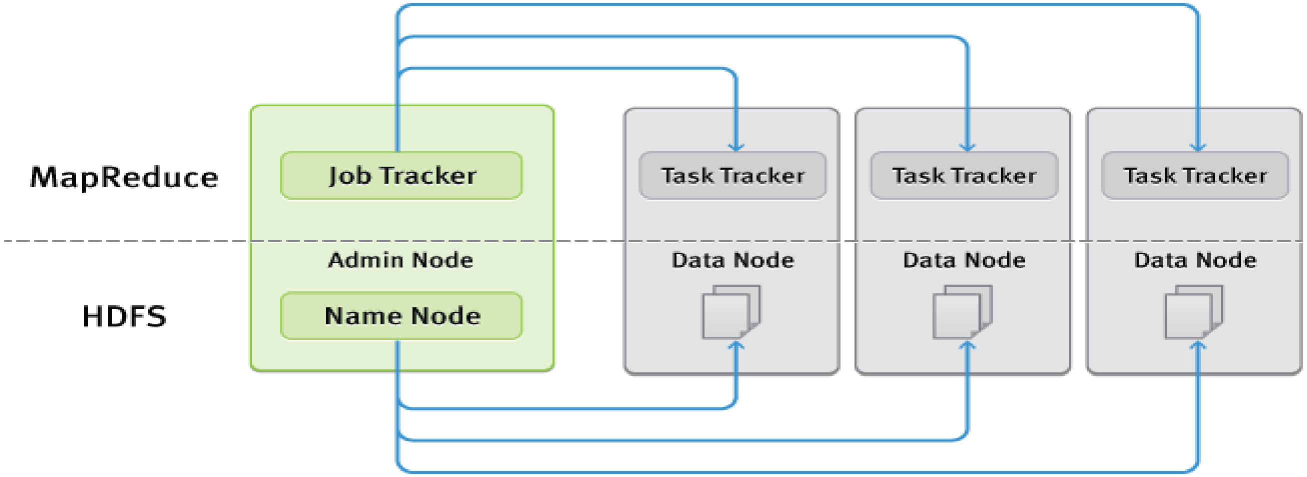
```
// Map Reduce function in JavaScript

var map = function (key, value, context) {
var words = value.split(/[^a-zA-Z]/);
for (var i = 0; i < words.length; i++) {
           if (words[i] !== "")
context.write(words[i].toLowerCase(),
1);}
}};

var reduce = function (key, values,
context) {
var sum = 0;
while (values.hasNext()) {
sum += parseInt(values.next());
      }
context.write(key, sum);
};
```

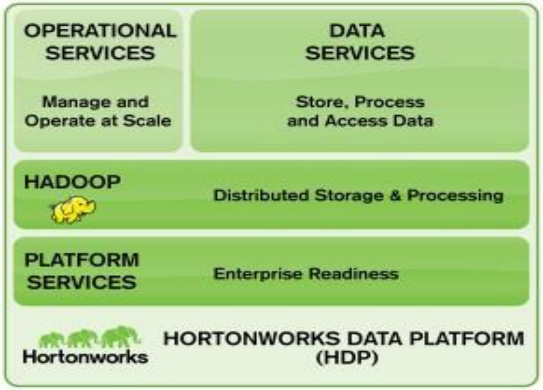
What is one of frameworks we will use?

- Open source software framework that facilitates big data management and analysis
- Parallelizes data processing across many nodes (computers) in a compute cluster using ‘batch’ style jobs
- Parallelizing Map and Reduce functions (MapReduce)



Our Big Data Partners in Hadoop…

Hortonworks



OPERATIONAL SERVICES

Manage and Operate at Scale

DATA SERVICES

Store, Process and Access Data

HADOOP

Distributed Storage & Processing

PLATFORM SERVICES

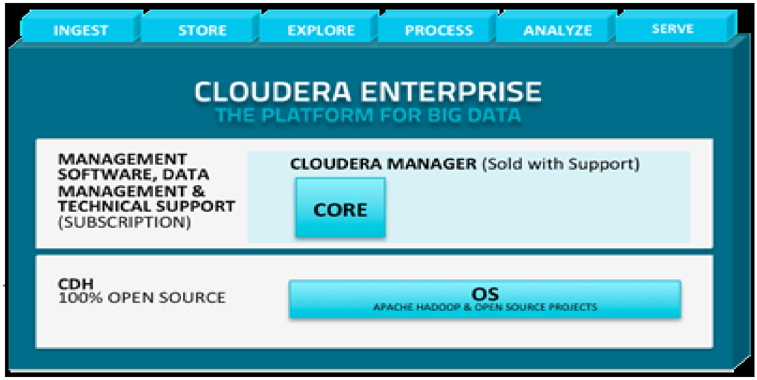
Enterprise Readiness

Hortonworks

HORTONWORKS DATA PLATFORM (HDP)

cloudera®

Ask Bigger Questions



INGEST | STORE | EXPLORE | PROCESS | ANALYZE | SERVE

CLOUDERA ENTERPRISE

THE PLATFORM FOR BIG DATA

MANAGEMENT SOFTWARE, DATA MANAGEMENT & TECHNICAL SUPPORT (SUBSCRIPTION)

CLOUDERA MANAGER (Sold with Support)

CORE

CDH 100% OPEN SOURCE

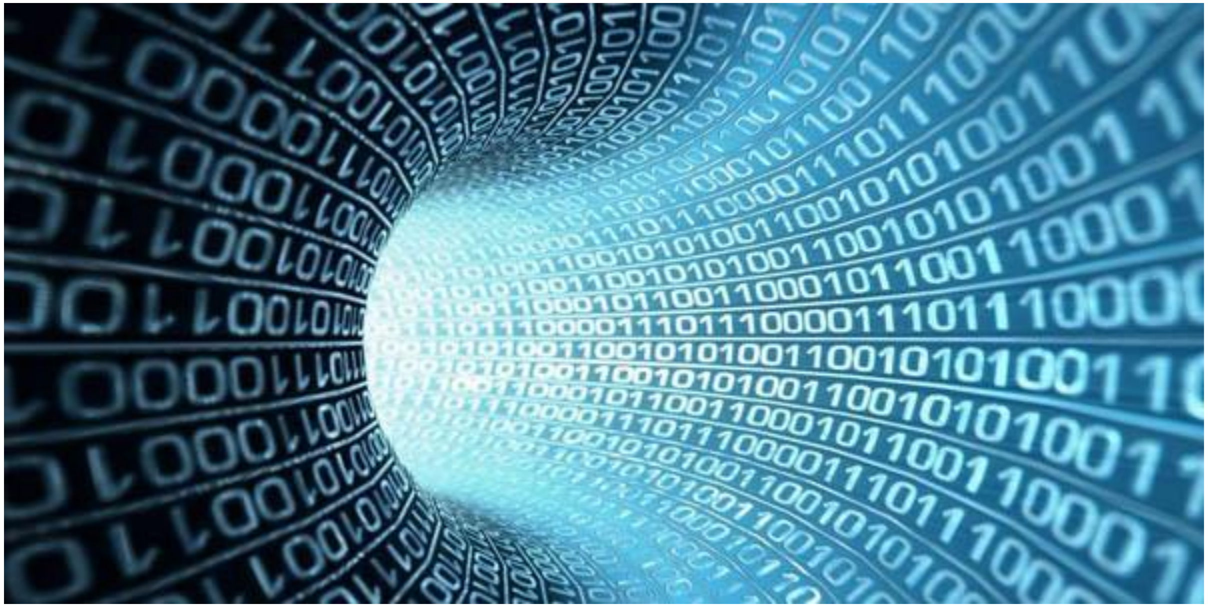
OS

APACHE HADOOP & OPEN SOURCE PROJECTS

In Summary

The Big Data conversation will continue..

- Looked at **'why'** Big Data is different
- Briefly looked at **'how'** Big Data problems are addressed
- The drive to insight and action!



Thank You…

Darryl.Dutton@T4G.com

GREAT PLACE TO WORK® Best Workplaces 2012 Canada