Big Data

Henrik W. Andersen
Consulting Manager

Big Data is like teenage sex:

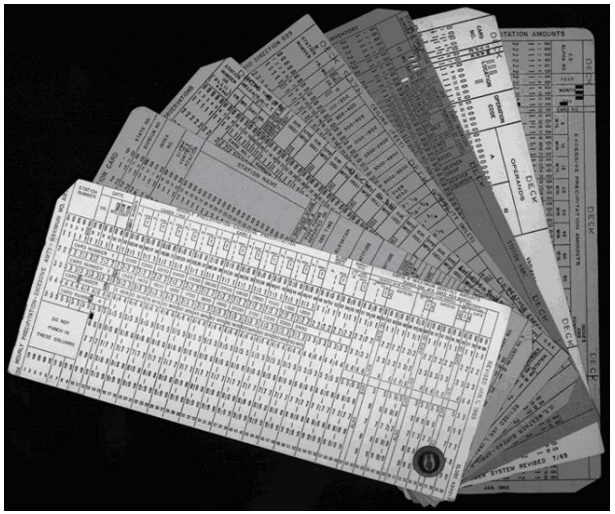
everyone talks about it
nobody really knows how to do it
everyone thinks everyone else is doing it
so everyone claims they are doing it

Affecto

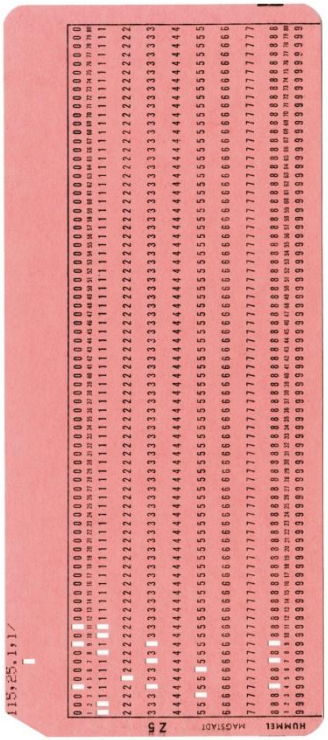
When I was young… ☺

We had Big Data, too!

Data were produced, stored and transported using punch cards



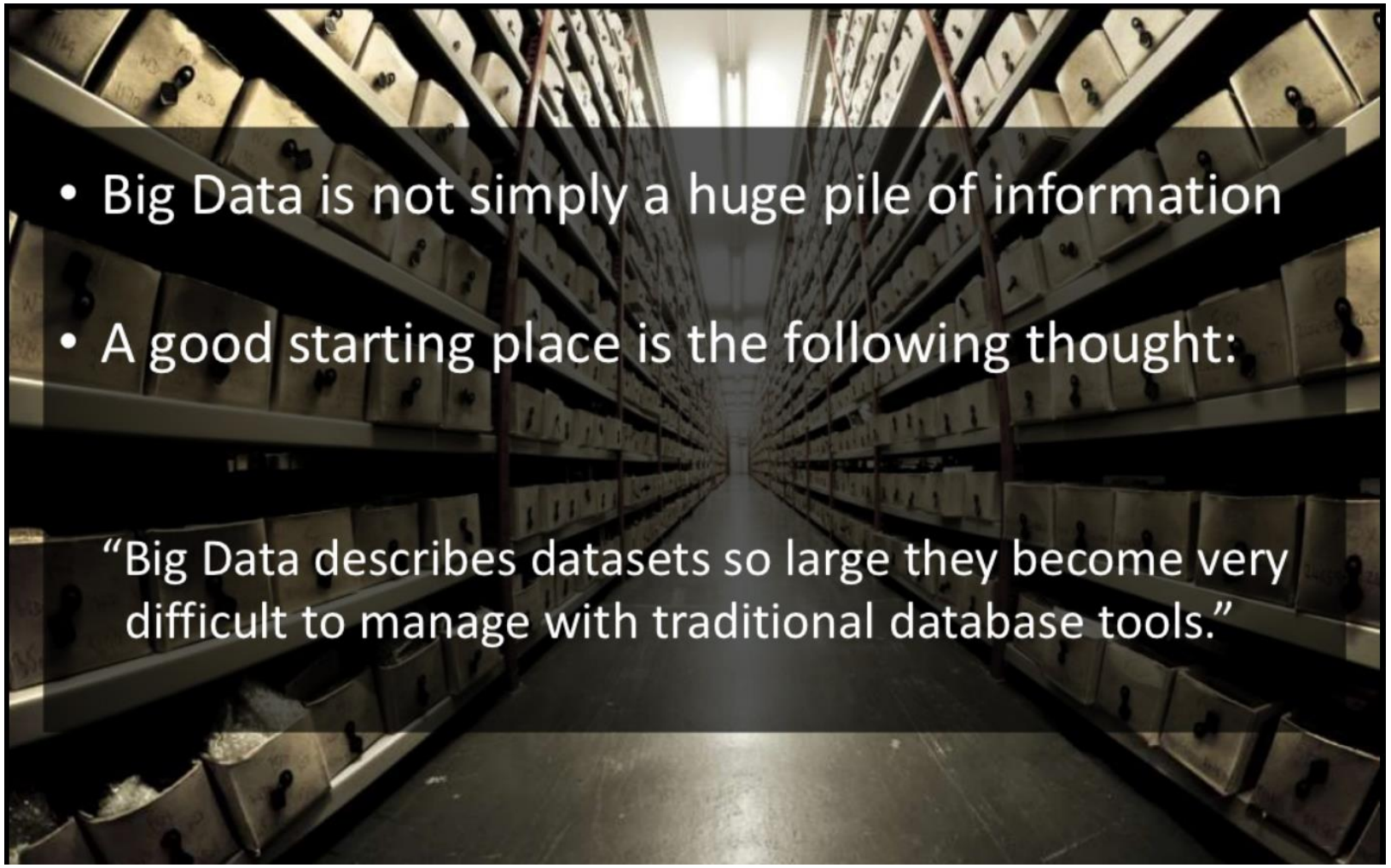
A punch card contained 80 characters and had a size of 19 x 8 x 0.02 cm.



A standard moving box could then contain up to 2.5 kB of data.

Affecto

What *is* Big Data?



Affecto

How big is big?

While experts agree that Big Data is big, exactly how big is a matter of debate.

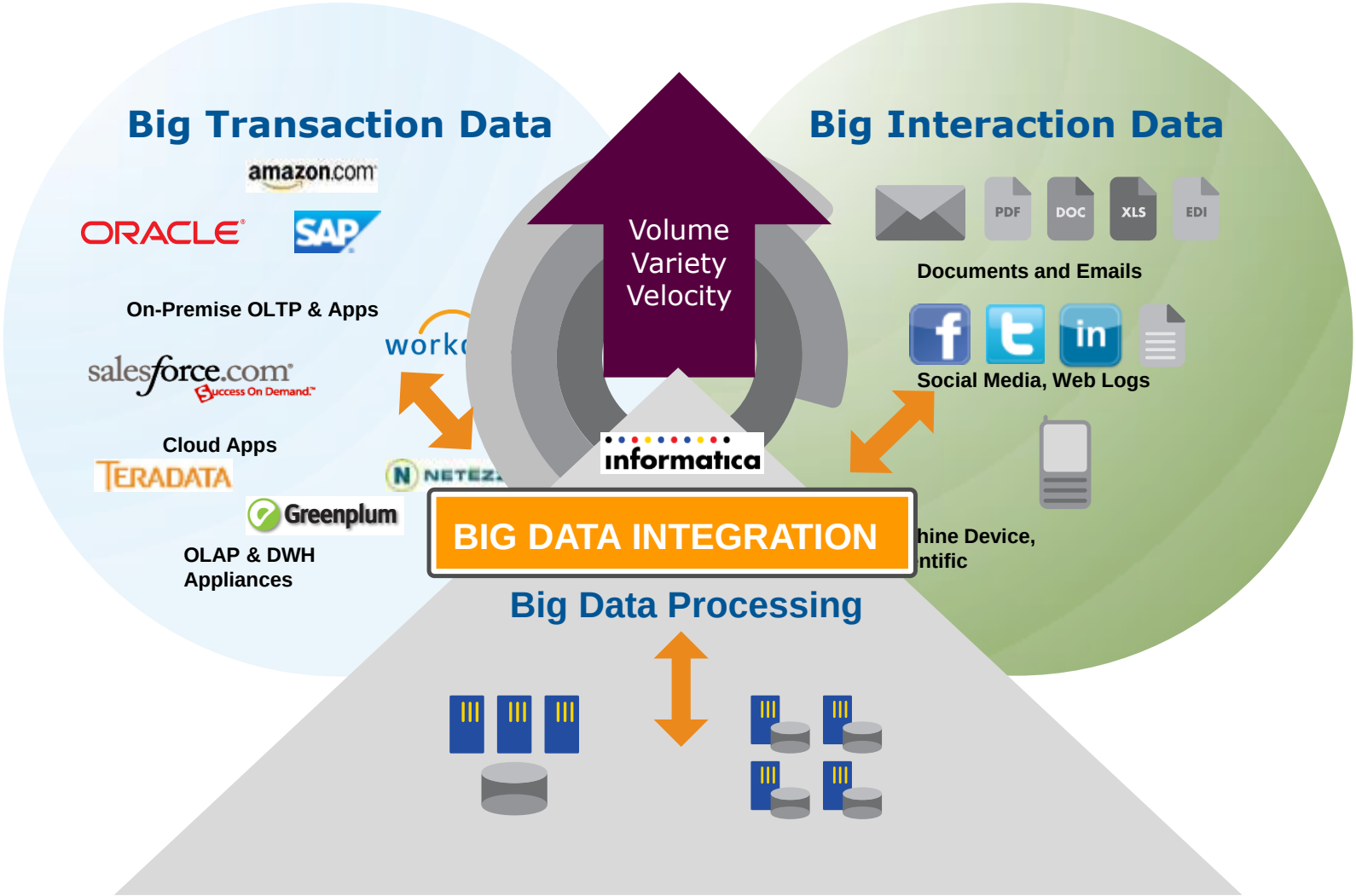
- IDC forecasts:
  - a roughly 50 percent annual growth rate for what it calls the world's "digital universe,"
  - more than 70 percent of which IDC estimates is generated by consumers
  - over 20 percent by enterprises.
  - Between 2009 and 2020, increasing by a factor of 44 to 35 zettabytes, or 35 million petabytes.
- Computer scientists at the University of California at S iego estimates:
  - The world's enterprise servers processed 9.57 ZB of t counting 3.6 ZB it
- an
- e and
- ıy

Bytes?

1 Kilobytes	1.000 bytes
1 Megabytes	1.000.000 bytes
1 Gigabytes	1.000.000.000 bytes
1 Terabytes	1.000.000.000.000 bytes
1 Petabytes	1.000.000.000.000.000 bytes
1 Exabytes	1.000.000.000.000.000.000 bytes
1 Zettabytes	1.000.000.000.000.000.000.000 bytes

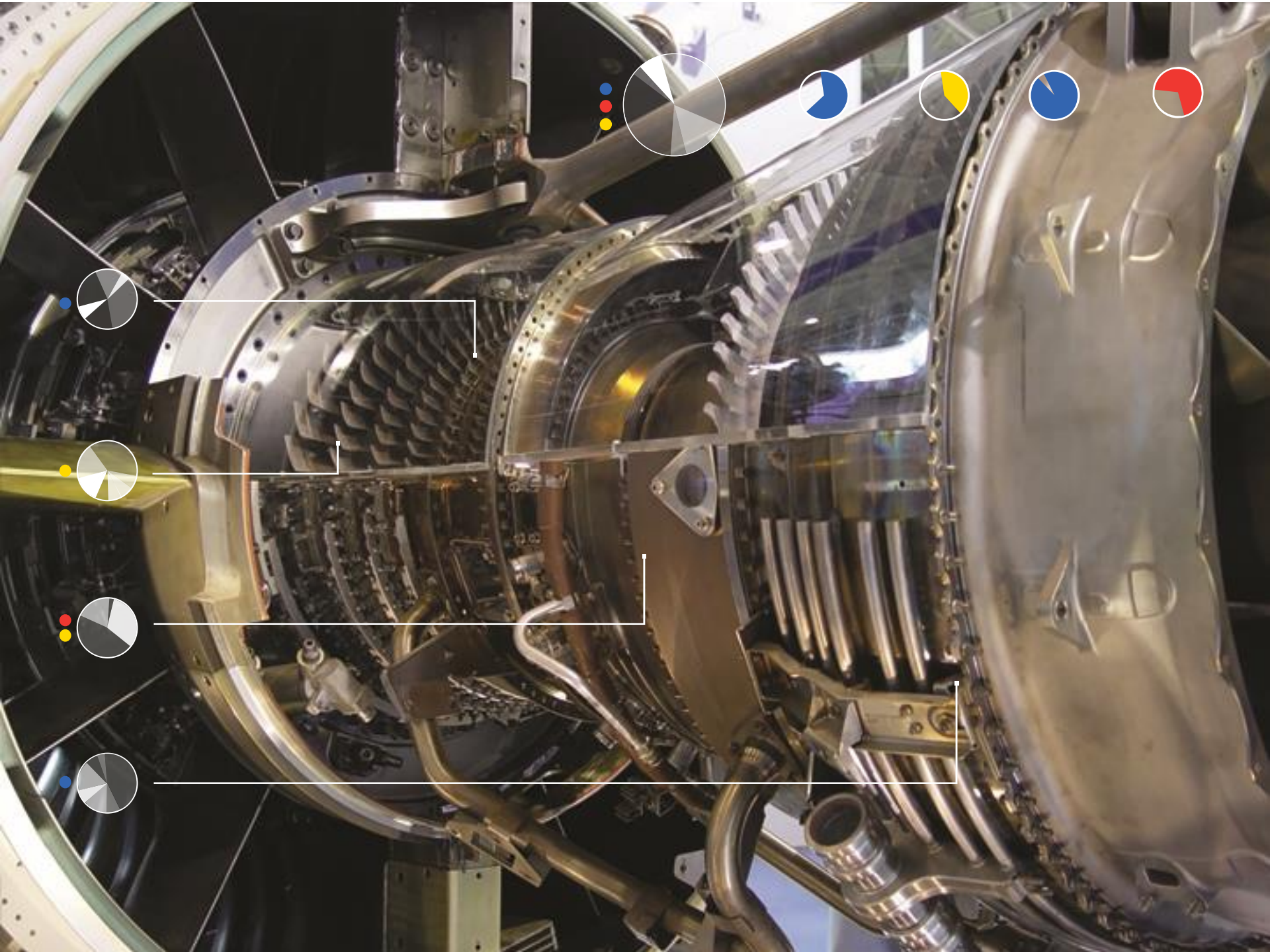
Affecto

What *is* Big Data?



Affecto


6
Cold
Cold




1372
1557
2423
2332
1671
1941
2232
2101
1832
2106

The next wave?

2014

2011

2007

Devices
& Machines

ommunities

BUSINESS

Real-Time
Optimization

TE

TECHNOLOG

cosm

ioBridge
Connect things.

ThingWorx

Texas Instruments

CISCO

Honeywell

TERNET
OF THINGS

10 9

10 11

Affecto

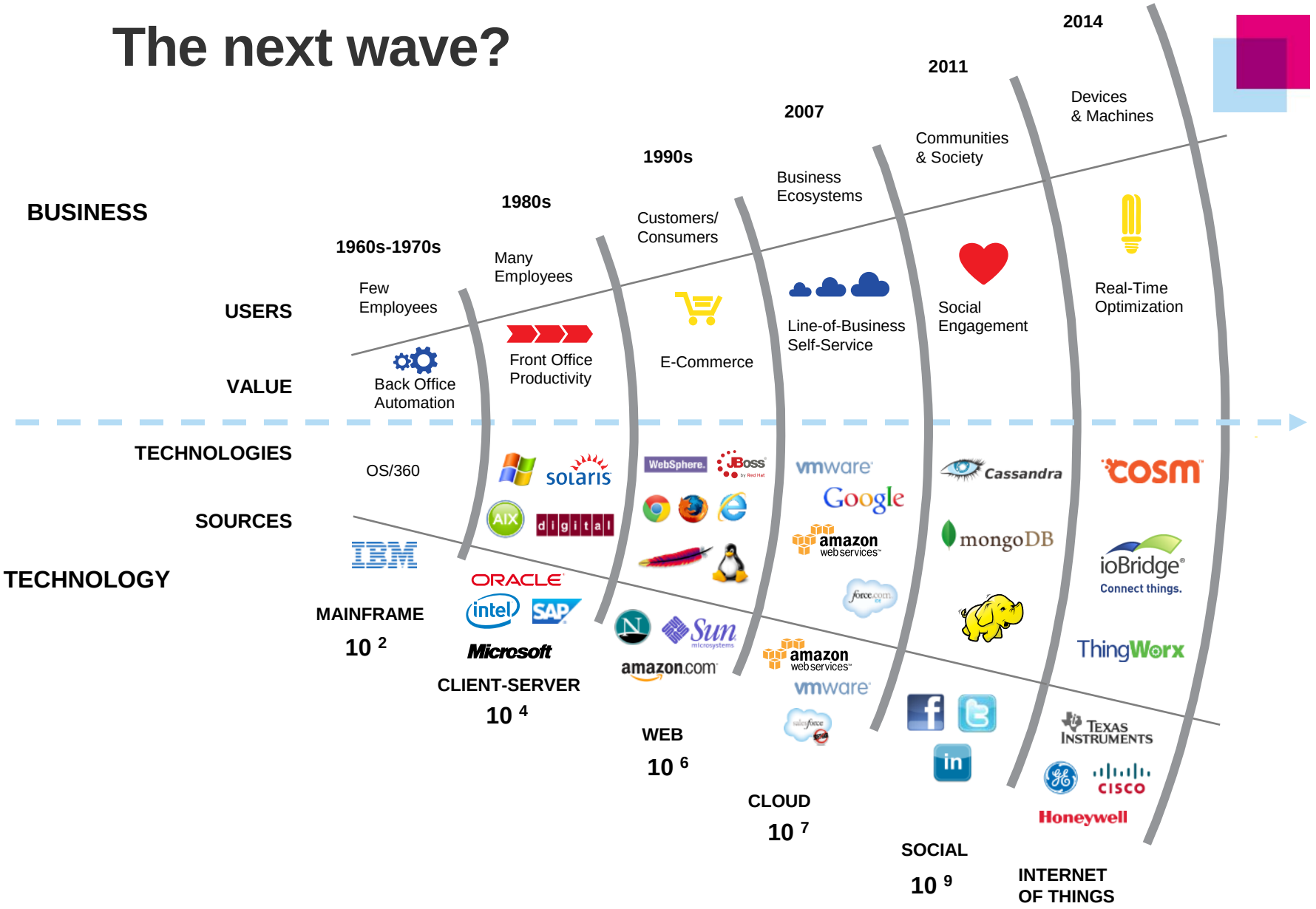
Google Acquires Smart Thermostat Maker Nest For $3.2 Billion

+ Comment Now + Follow Comments

Three years after redefining what thermostats are capable of, Palo Alto-based Nest is being bought by Google GOOG -3.74% for $3.2 billion.

"Google will help us fully realize our vision of the conscious home and allow us to change the world faster than we ever could if we continued to go it alone. We've had great momentum, but this is a rocket ship," said CEO and cofounder Tony Fadell in a blog post.

The next wave?



Affecto

Industrial in the IoT

Wirelessly connected LED lighting & energy management

Flow & pressure sensors

Affecto

Automotive in the IoT

Navigation, Bluetooth hands free & audio, Wi-Fi

Keyless entry, interior lighting, mirror control, sensors

Affecto

Homes in the IoT

Security & safety system, sensors

Smart home energy gateway, thermostats, sensors

Affecto

Fitness and healthcare in the IoT

Informed workouts with activity & performance measurement

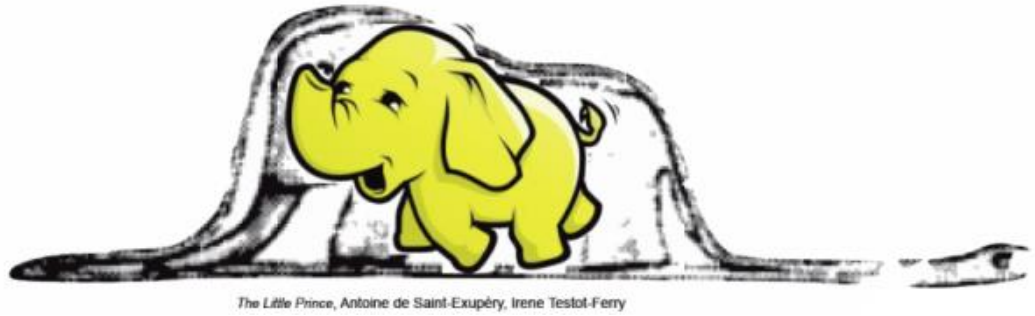
Safe independent living with fall detection, medication monitoring, etc.

Affecto


hadoop

What is Hadoop?

Apache Hadoop is an
open-source system
to reliably store and process
gobs of information
across many commodity computers.



The Little Prince, Antoine de Saint-Exupéry, Irene Testot-Ferry

Scalable

Low Cost

No Rules

It enables organizations to cost-effectively store and process Petabyte of data in a reasonable amount of time

Affecto

What is Hadoop?

HDFS & Map/Reduce

HDFS breaks incoming files into blocks and stores them redundantly across the cluster

Map/Reduce process large jobs in parallel across many nodes and combine the results

Map() procedure that performs filtering and sorting

Reduce() procedure that performs a summary operation

Affecto

```
public class JoinStationMapper extends MapReduceBase implements
Mapper<LongWritable, Text, TextPair, Text> {
     private NcdcStationMetadataParser parser = new
NcdcStationMetadataParser();

     public void map(LongWritable key, Text value,
OutputCollector<TextPair, Text> output, Reporter reporter)
throws IOException {
          if (parser.parse(value)) {
           output.collect(new TextPair(parser.getStationId(),
"0"), new Text(parser.getStationName()));
          }
     }
}

public class JoinRecordMapper extends MapReduceBase implements
Mapper<LongWritable, Text, TextPair, Text> {
     private NcdcRecordParser parser = new NcdcRecordParser();
     public void map(LongWritable key, Text value,
OutputCollector<TextPair, Text> output, Reporter reporter)
throws IOException {
     parser.parse(value); output.collect(new
TextPair(parser.getStationId(), "1"), value);
     }
}

public class JoinReducer extends MapReduceBase implements
Reducer<TextPair, Text, Text, Text> {
     public void reduce(TextPair key, Iterator<Text> values,
OutputCollector<Text, Text> output, Reporter reporter) throws
IOException {
          Text stationName = new Text(values.next());
          while (values.hasNext()) {
               Text record = values.next();
               Text outValue = new Text(static
+ "\t" + record.toString());
               output.collect(key.getFirst(),
          }
     }
}

public class JoinRecordWithStationName extends Configured
implements Tool {
     public static class KeyPartitioner implements
Partitioner<TextPair, Text> {
          @Override
          public void configure(JobConf job) {}
```

```
     @Override
     public int getPartition(TextPair ke
numPartitions) {
          return (key.getFirst().hashCo
Integer.MAX_VALUE) % numPartitions;
          }
     }
     @Override
     public int run(String[] args) throw
          if (args.length != 3) {
               JobBuilder.printUsage(this, "<ncdc input>
<station input> <output>");
               return -1;
          }
          JobConf conf = new JobConf(getConf(), getClass());
          conf.setJobName("Join record with station name");
          Path ncdcInputPath = new Path(args[0]);
          Path stationInputPath = new Path(args[1]);
          Path outputPath = new Path(args[2]);
          MultipleInputs.addInputPath(conf, ncdcInputPath,
TextInputFormat.class, JoinRecordMapper.class);
          MultipleInputs.addInputPath(conf, stationInputPath,
TextInputFormat.class, JoinStationMapper.class);
          FileOutputFormat.setOutputPath(conf, outputPath);
          conf.setPartitionerClass(KeyPartitioner.class);

     conf.setOutputValueGroupingComparator(TextPair.FirstCompa
rator.class);
          conf.setMapOutputKeyClass(TextPair.class);
          conf.setReducerClass(JoinReducer.class);
          conf.setOutputKeyClass(Text.class);
          JobClient.runJob(conf);
          return 0;
     }
}
```

```
SELECT * FROM Stations JOIN Records ON
       (Stations.StationID = Records.StationID);
```

Some Pros and Cons

Pros

- Jobs easily scale to infinity!
- Built-in resiliency / fault tolerance
- Great for massive full data scans
- Data does not require structuring
- Deep support for complex data structures

Cons

- Batch oriented only
- Map Reduce is hard
- No Metadata
- No Governance
- ***And there are too few developers***

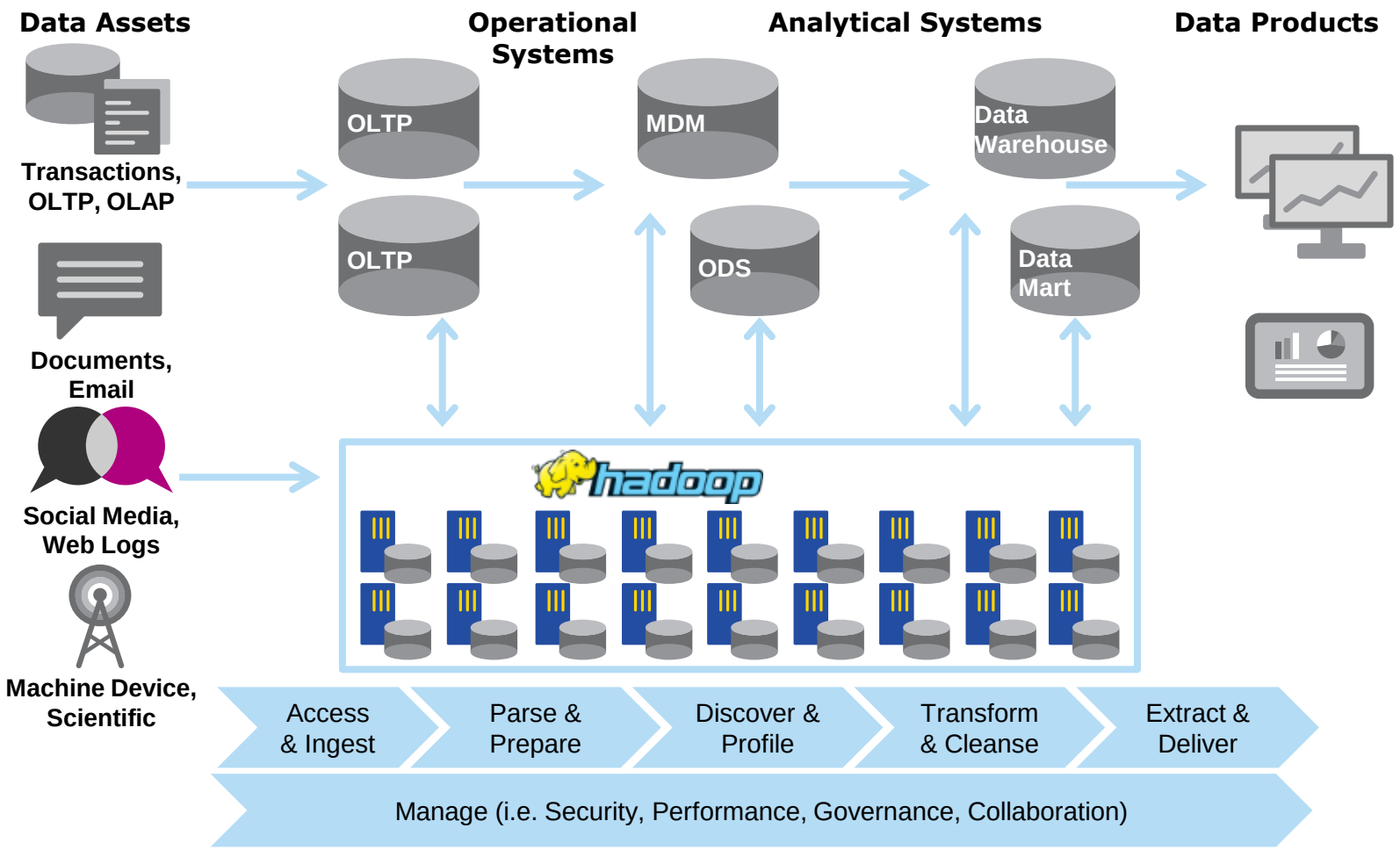
Alternative to Map/Reduce is ***Hive Query Language***

- Faster development
- Interpreted
- Much slower in execution

Affecto

Hadoop – Zoom out

Hadoop Complements Your Existing Infrastructure



Affecto

Big Data is not just about storing and accessing large amount of data

What happened in 2013?

- Big Data “On-the-go”– visualizations on mobile devices
- A plethora of Open Source tools for Big Data
- More tools on top of Hadoop for real time data analysis
- Big Consumer Data – wearable technologies measuring every day life
- Privacy is affected by Big Data

Source: SmartData Collective

Affecto

Big Data is not just about storing and accessing large amount of data

What happens in 2014?

- The Industrial Internet will increase dramatically
  - By 2020, 40% of all data will come from sensor data
- It's going to be cloudy: Big-Data-as-a-Service (BDaaS)
  - "Ready-made analytics in the cloud"
  - Gartner:
    - Cloud computing will become bulk in IT spending by 2016
    - 2014 will be a turning point
- Organizations will start focusing on security
- Personalization will become personal

Source: SmartData Collective

Affecto

Big Data is not just about storing and accessing large amount of data

What happens in 2014?

- Education, education, education…
  - We lack Big Data competencies
  - Steep increase of online and offline Big Data training
- Big Data: mixing and combining data sets
- Proof of Concept
  - You can't just talk the talk. You've got to walk the walk!
  - PoC to better understand Big Data

Source: SmartData Collective

Affecto

Big Data

Henrik W. Andersen
Consulting Manager

Big Data – now ***you*** know what it is! ☺

Affecto