TDWI WEBINAR SERIES

Unifying the Traditional Enterprise Data Warehouse with Hadoop

Colin White
President, BI Research
TDWI and IBM Webinar
June 2015

tdwi
Advancing all things data.

Sponsor

Speakers

Colin White
President,
BI Research

Dwaine R. Snow
Strategy Lead, Analytics and
Big Data Solutions, IBM

Webinar Overview

The three-decade-old enterprise data warehouse is evolving into an enhanced data warehouse architecture where Hadoop acts as a supporting platform for traditional data warehouse activities. The challenge with this enhanced data warehouse approach is how to store and access data transparently regardless where it is located and how it is managed. This presentation reviews why organizations are adding Hadoop to the traditional data warehouse, presents use cases for such an environment, and takes a detailed look at why organizations need a common and transparent interface to both traditional relational and Hadoop data management systems. Topics that will be covered include:

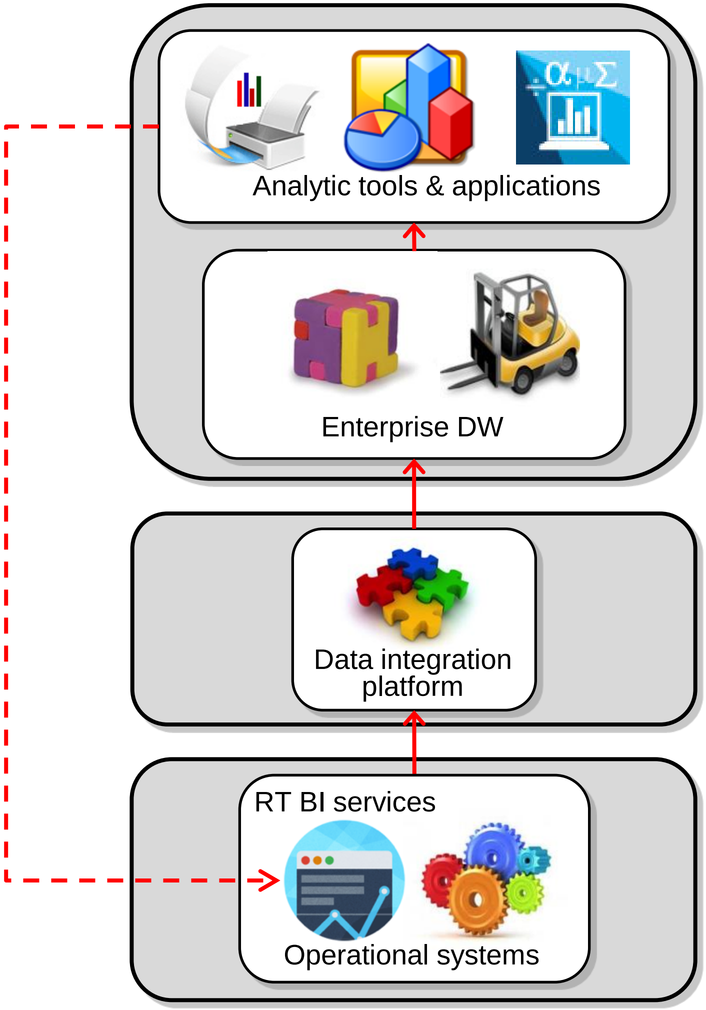
•Extending the traditional data warehouse with Hadoop

•Use cases for Hadoop in a data warehousing environment

•Accessing a mixed data management environment

•Supporting a common and transparent data interface to heterogeneous data and systems

The Traditional Enterprise DW



Data is modeled, acquired, integrated and loaded into the EDW before it is analyzed

The Data Warehouse is Changing

Data Management

- New data sources (big data)
- Analytic relational DBMSs
- Non-relational systems (e.g., Hadoop, HBase, MongoDB, CouchDB)
- Improved price/performance

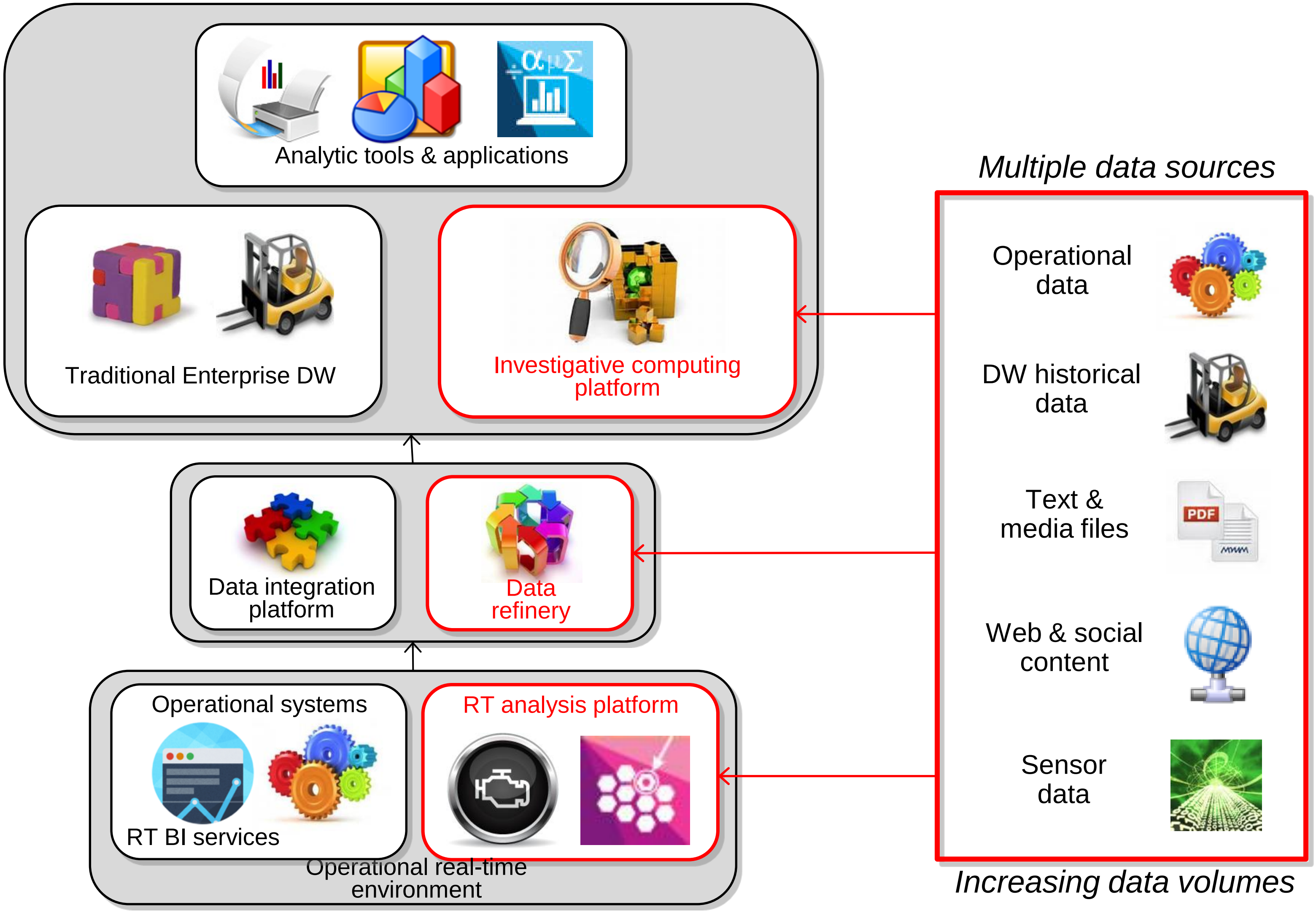
Deployment Options

- Integrated H/W & S/W appliances
- Data hubs (aka data lakes)
- Cloud computing
- Mobile devices (mobile first strategy)

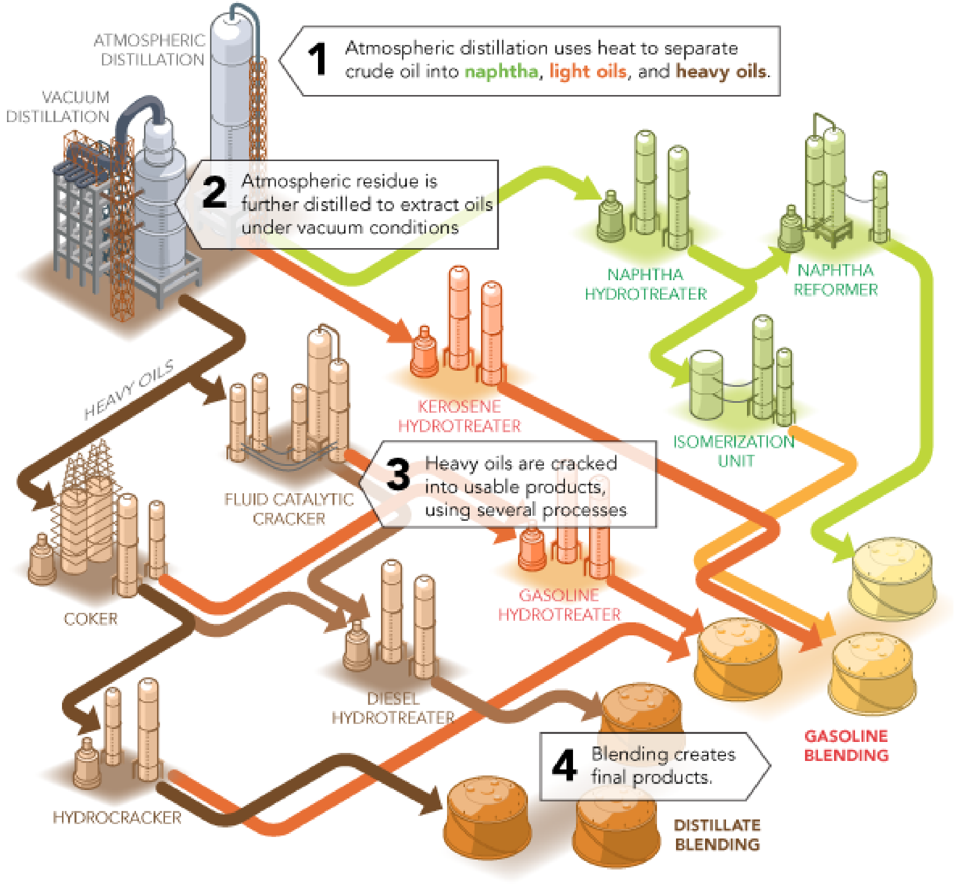
Analytics

- Investigative computing
- Predictive & prescriptive analyses
- Enhanced visualization

Modernizing the EDW



The Data Refinery



1 Atmospheric distillation uses heat to separate crude oil into **naphtha**, **light oils**, and **heavy oils**.

2 Atmospheric residue is further distilled to extract oils under vacuum conditions

3 Heavy oils are cracked into usable products, using several processes

4 Blending creates final products.

Ingests raw data in batch and/or real-time into a managed data store

Distills the data into useful business information and distributes the results to downstream systems

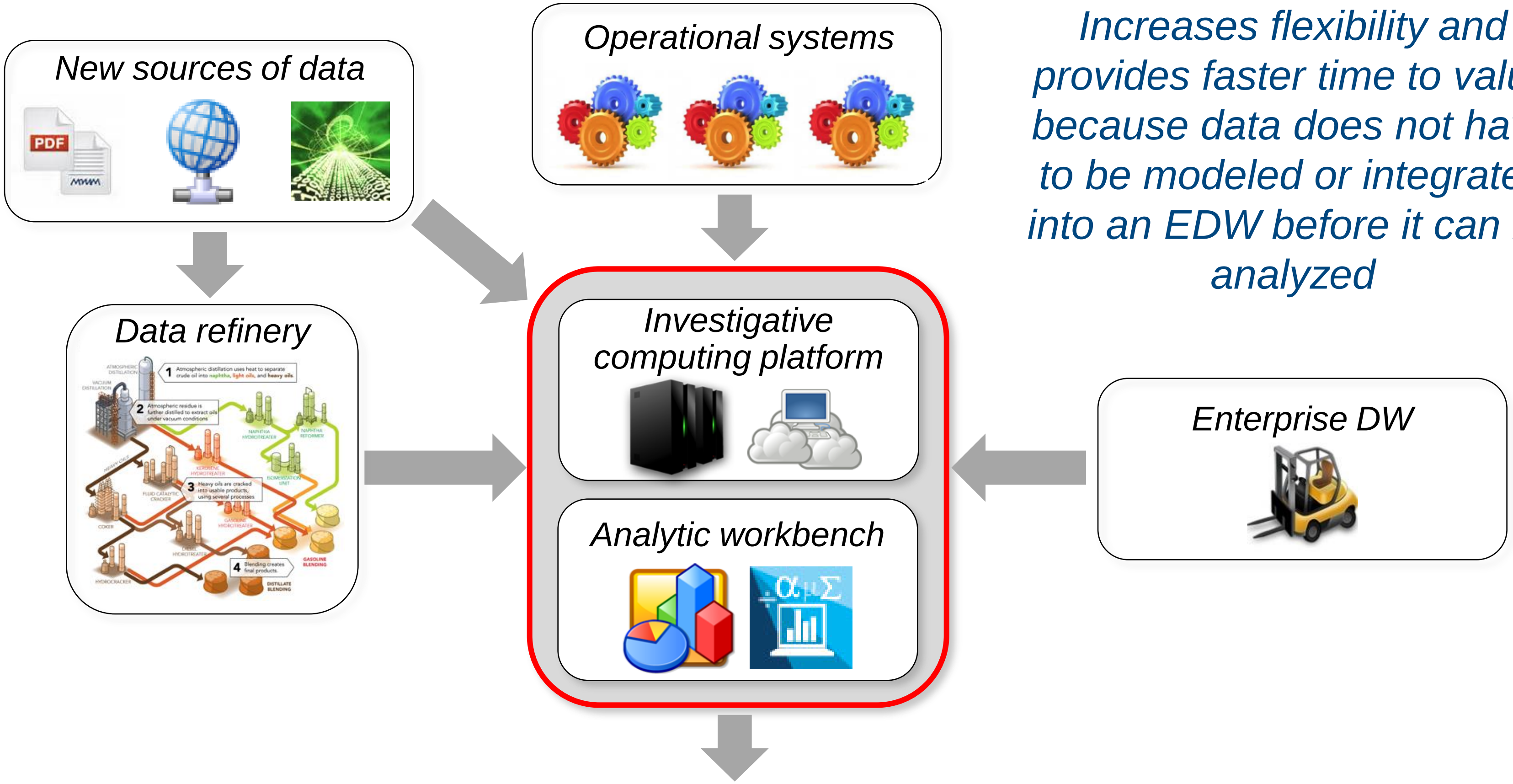
May also be used to directly analyze certain types of data

May also be used for data archiving and for creating a “queryable” archive

Employs low-cost H/W and S/W to enable large amounts of detailed data to be managed cost effectively

Requires (flexible) governance policies to manage data security, privacy, quality, archiving and destruction

Investigative Computing



New sources of data

Operational systems

Data refinery

Investigative computing platform

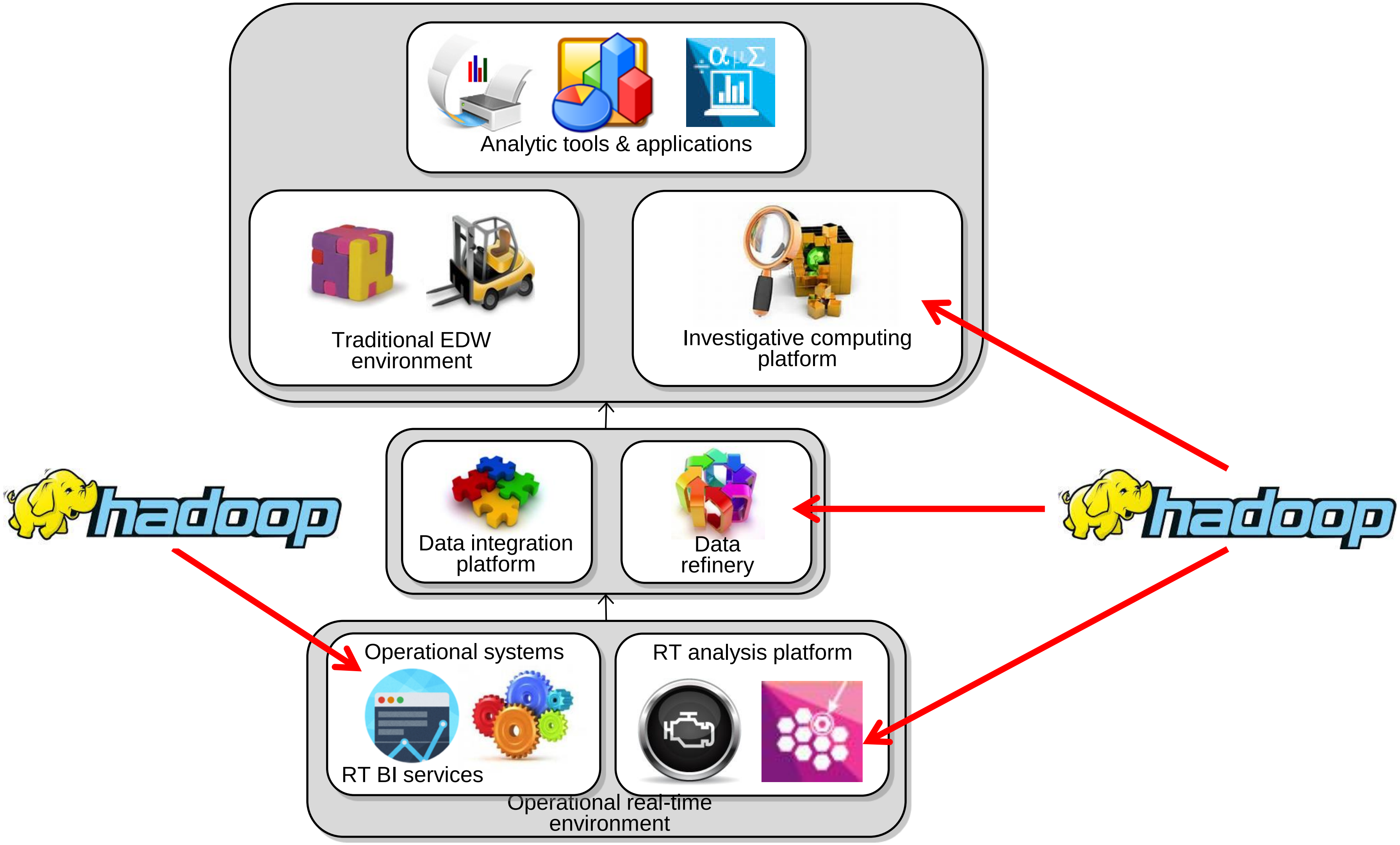
Analytic workbench

Enterprise DW

Increases flexibility and provides faster time to value because data does not have to be modeled or integrated into an EDW before it can be analyzed

Data schemas
Analyses and analytic models
Business rules and decision workflows
Dashboards and analytics-driven LOB applications

Potential Uses for Hadoop



Analytic tools & applications

Traditional EDW environment

Investigative computing platform

Data integration platform

Data refinery

Operational systems

RT BI services

RT analysis platform

Operational real-time environment

TDWI Survey

In your perception, what would be the most useful applications of HDFS if your organization were to implement it? Select four or fewer.

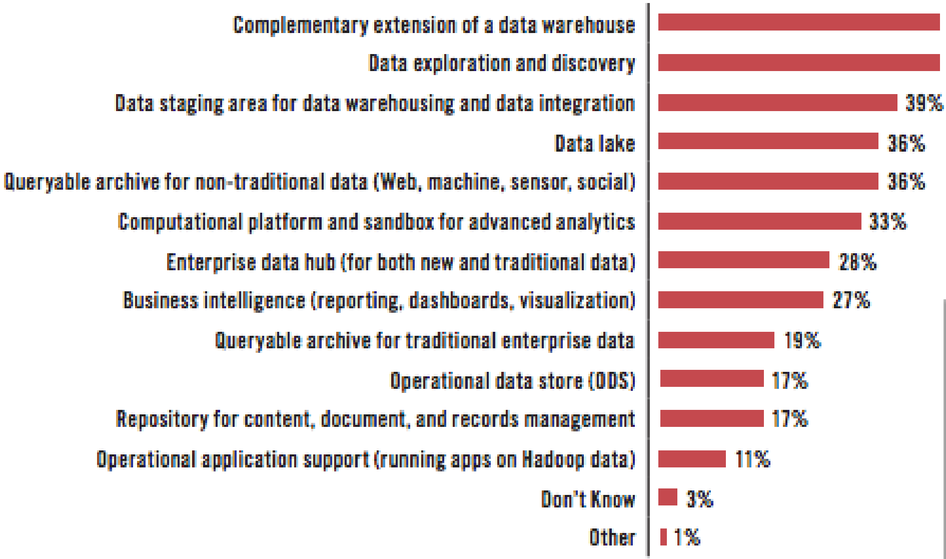
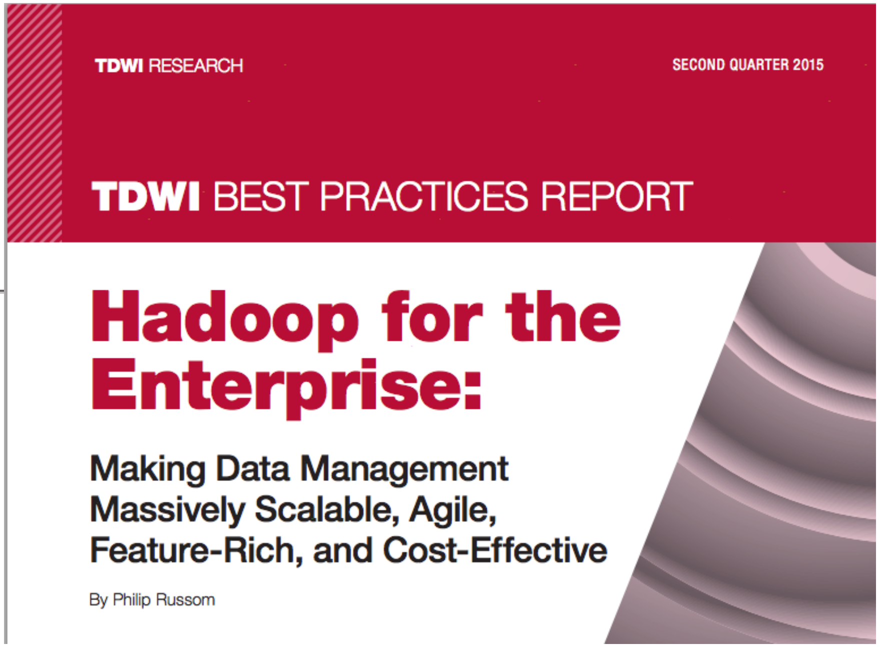


Figure 1. *Based on 743 responses from 207 respondents. 3.6 responses per respondent on average.*



TDWI RESEARCH

SECOND QUARTER 2015

TDWI BEST PRACTICES REPORT

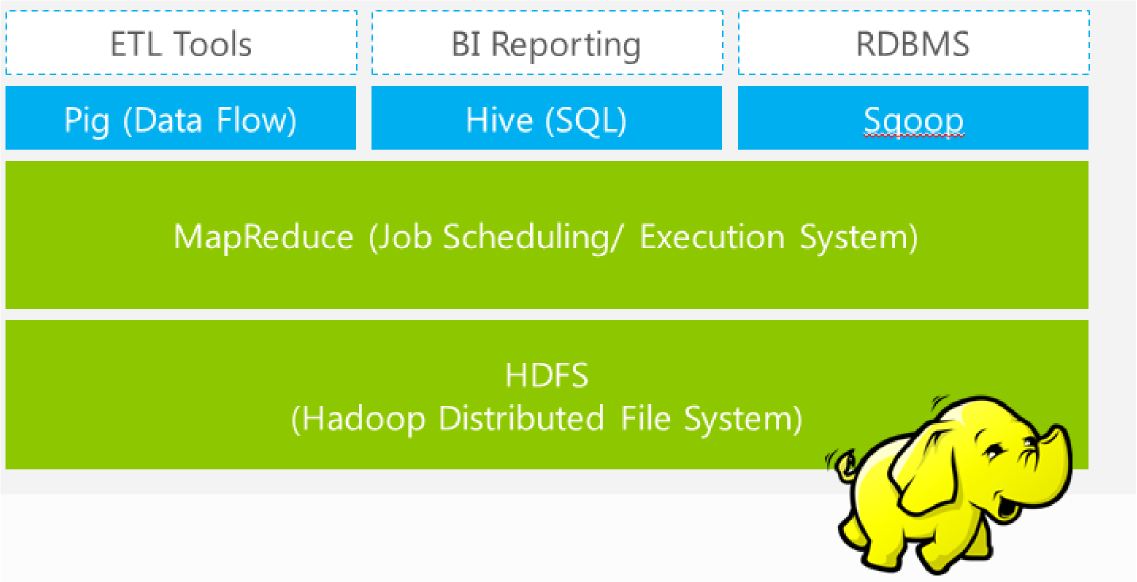
Hadoop for the Enterprise:

Making Data Management Massively Scalable, Agile, Feature-Rich, and Cost-Effective

By Philip Russom

Hadoop Origins

"A framework for running applications on a large hardware cluster built of commodity hardware." wiki.apache.org/hadoop/


ETL Tools
BI Reporting
RDBMS
Pig (Data Flow)
Hive (SQL)
Sqoop
MapReduce (Job Scheduling/ Execution System)
HDFS
(Hadoop Distributed File System)

Evolved from development work done at Internet Archive, Google and Yahoo!

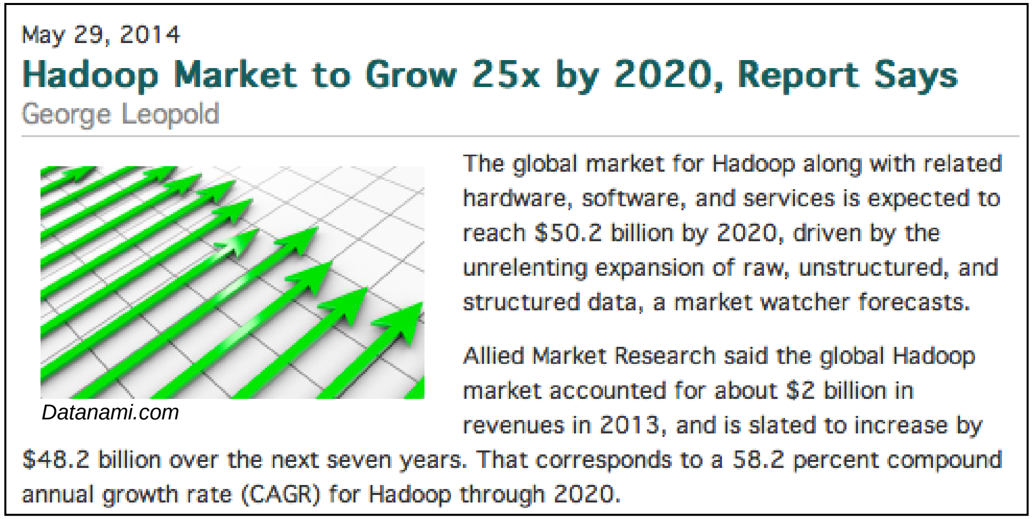
Focused initially on programmatic and batch-oriented applications that processed large amounts of Internet-based multi-structured data (the original "big data")

Systems are often deployed by assembling Apache components or using Hadoop distributions from "open source" providers

Hadoop Today

Has moved beyond batch MapReduce processing to support a range of application use cases

Classic and independent vendors have joined the race to support Hadoop and build applications and services on Hadoop



May 29, 2014

Hadoop Market to Grow 25x by 2020, Report Says

George Leopold

The global market for Hadoop along with related hardware, software, and services is expected to reach $50.2 billion by 2020, driven by the unrelenting expansion of raw, unstructured, and structured data, a market watcher forecasts.

Allied Market Research said the global Hadoop market accounted for about $2 billion in revenues in 2013, and is slated to increase by $48.2 billion over the next seven years. That corresponds to a 58.2 percent compound annual growth rate (CAGR) for Hadoop through 2020.

Datanami.com

Not just for Internet businesses anymore – many traditional businesses have Hadoop projects in evaluation and in production

Hadoop in a Modernized DW: Key Questions

What are the use cases for Hadoop?

What are the TCO considerations for Hadoop?

How mature is the Hadoop ecosystem?

Which Hadoop solution should we use?

What are the skill requirements for Hadoop?

How do we integrate Hadoop applications into the existing EDW environment?

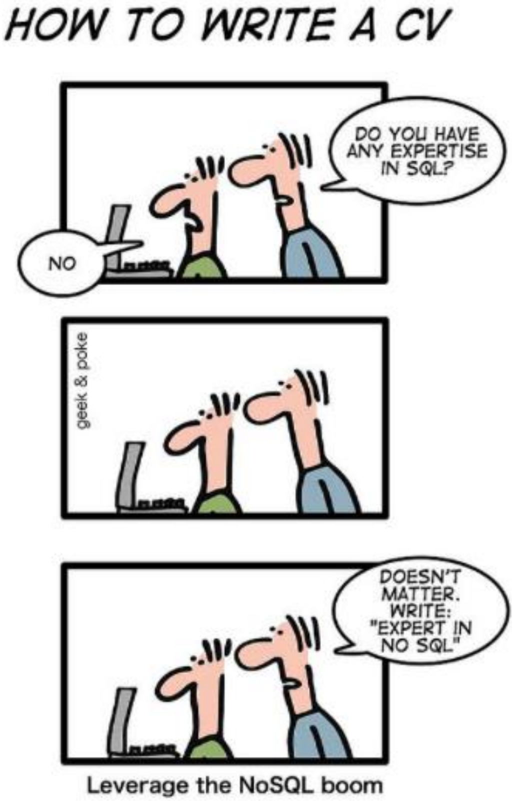
Hadoop Data Management Origins - 1

Several Internet companies developed their own non-relational (NoSQL or NewSQL) systems to support extreme data volumes

- Google example: Google file system, MapReduce, BigTable, BigQuery
- Main goal was the processing of large volumes of multi-structured data
- Several of these developments found their way into the open source community

Non-relational systems are not new, but modern versions are often open source

- Usually deployed on low-cost hardware in a large-scale distributed computing environment
- Support different approaches to data management

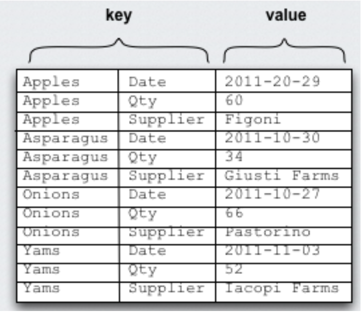

HOW TO WRITE A CV
DO YOU HAVE ANY EXPERTISE IN SQL?
NO
geek & poke
DOESN'T MATTER. WRITE: "EXPERT IN NO SQL"
Leverage the NoSQL boom

Hadoop Data Management Origins - 2

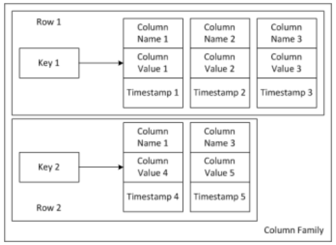
Many types of products, APIs and languages including several different SQL implementations

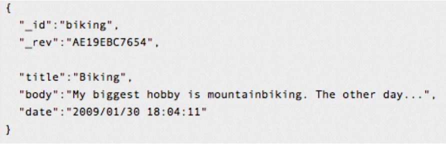
Volume ⟷ *Complexity*

Key/Value Pair | Column Family | Document | Graph

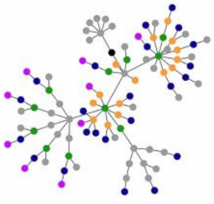


key		value
Apples	Date	2011-20-29
Apples	Qty	60
Apples	Supplier	Figoni
Asparagus	Date	2011-10-30
Asparagus	Qty	34
Asparagus	Supplier	Giusti Farms
Onions	Date	2011-10-27
Onions	Qty	66
Onions	Supplier	Pastorino
Yams	Date	2011-11-03
Yams	Qty	52
Yams	Supplier	Iacopi Farms



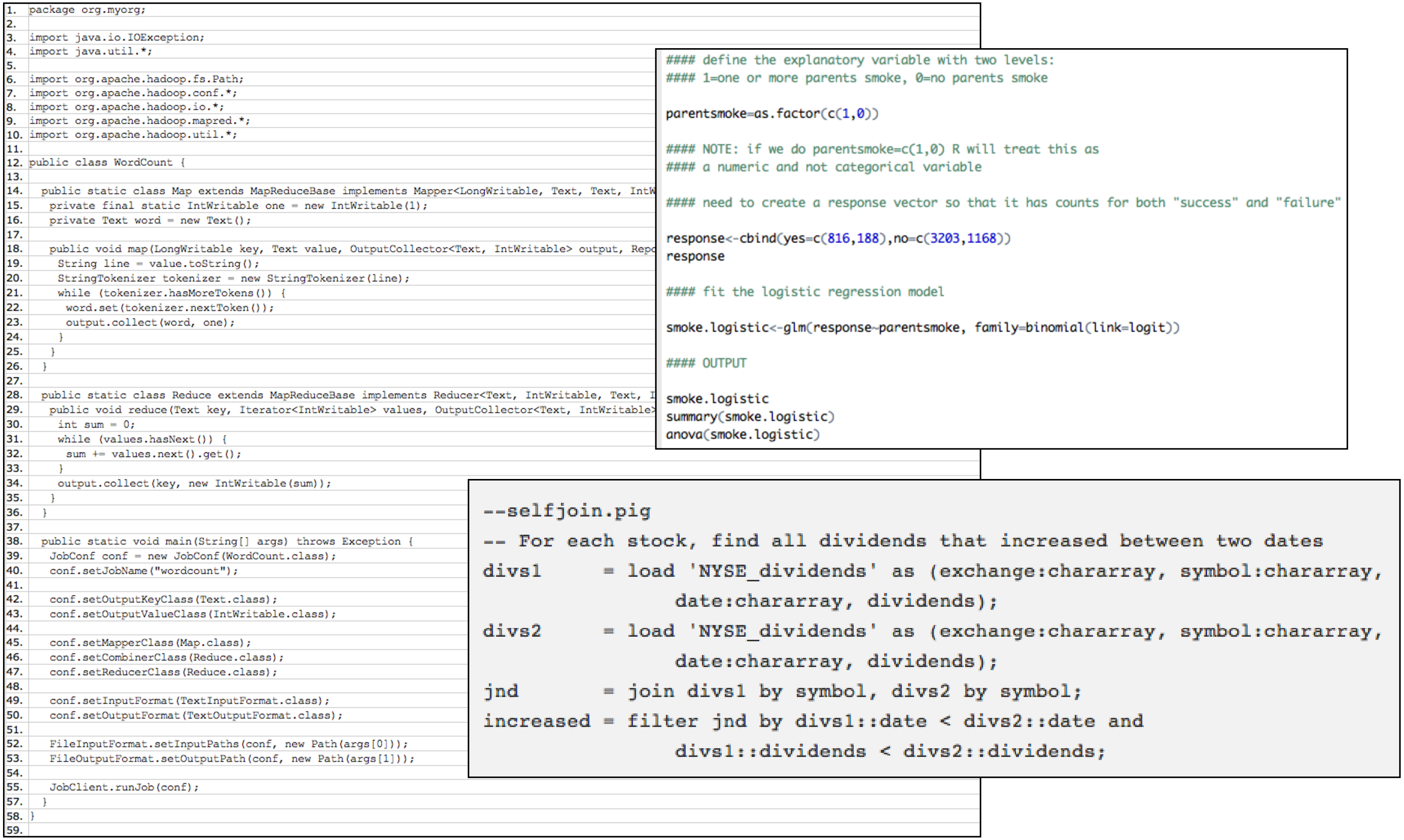


```
{
  "_id":"biking",
  "_rev":"AE19EBC7654",

  "title":"Biking",
  "body":"My biggest hobby is mountainbiking. The other day...",
  "date":"2009/01/30 18:04:11"
}
```



Can handle varieties of data and processing that are difficult to support using a traditional RDBMS

Hadoop Data Management: Language Examples

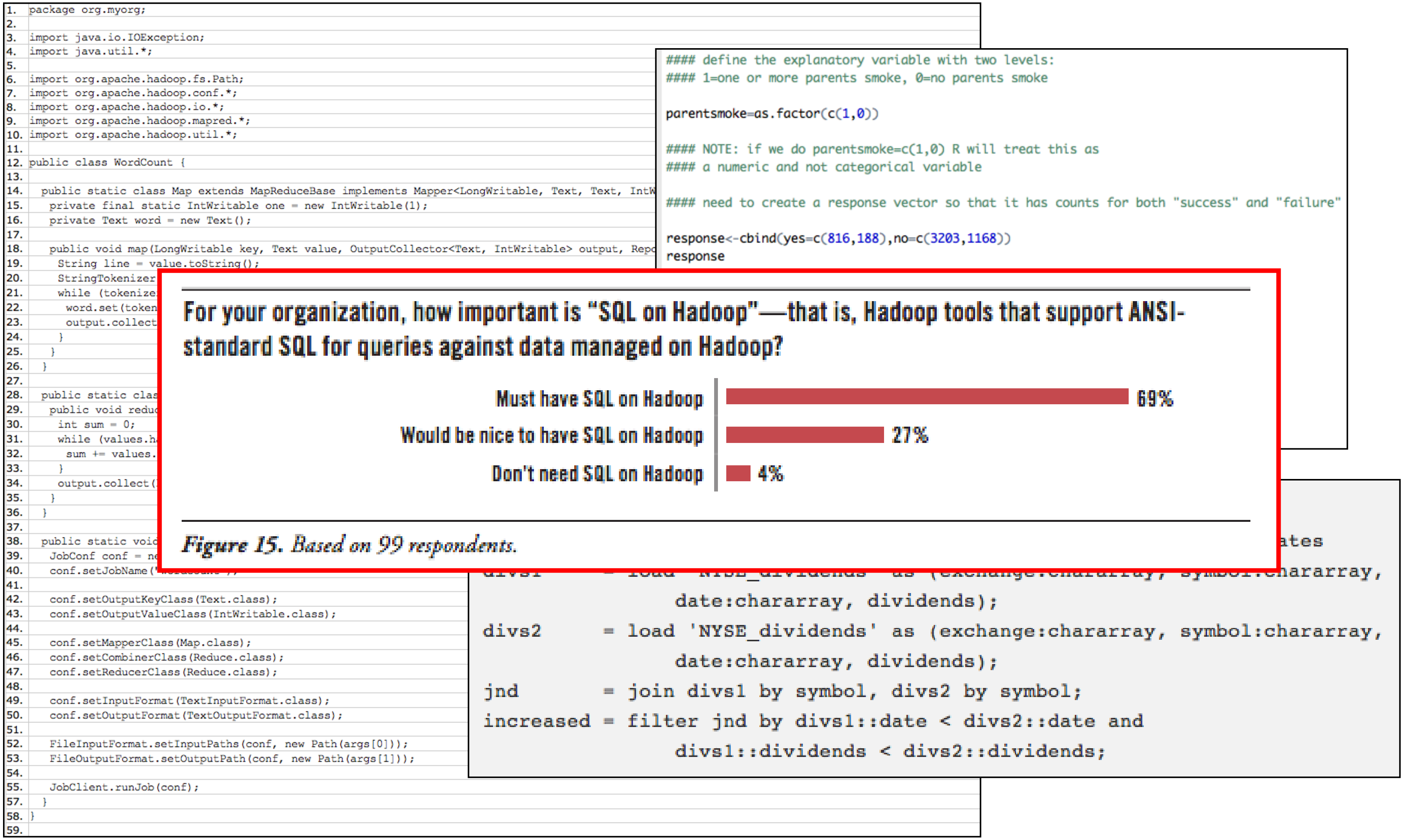


```
package org.myorg;

import java.io.IOException;
import java.util.*;

import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapred.*;
import org.apache.hadoop.util.*;

public class WordCount {

  public static class Map extends MapReduceBase implements Mapper<LongWritable, Text, Text, IntW
   private final static IntWritable one = new IntWritable(1);
   private Text word = new Text();

   public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable> output, Repo
    String line = value.toString();
    StringTokenizer tokenizer = new StringTokenizer(line);
    while (tokenizer.hasMoreTokens()) {
     word.set(tokenizer.nextToken());
     output.collect(word, one);
    }
   }
  }

  public static class Reduce extends MapReduceBase implements Reducer<Text, IntWritable, Text, I
   public void reduce(Text key, Iterator<IntWritable> values, OutputCollector<Text, IntWritable>
    int sum = 0;
    while (values.hasNext()) {
     sum += values.next().get();
    }
    output.collect(key, new IntWritable(sum));
   }
  }

  public static void main(String[] args) throws Exception {
   JobConf conf = new JobConf(WordCount.class);
   conf.setJobName("wordcount");

   conf.setOutputKeyClass(Text.class);
   conf.setOutputValueClass(IntWritable.class);

   conf.setMapperClass(Map.class);
   conf.setCombinerClass(Reduce.class);
   conf.setReducerClass(Reduce.class);

   conf.setInputFormat(TextInputFormat.class);
   conf.setOutputFormat(TextOutputFormat.class);

   FileInputFormat.setInputPaths(conf, new Path(args[0]));
   FileOutputFormat.setOutputPath(conf, new Path(args[1]));

   JobClient.runJob(conf);
  }
}
```

```
#### define the explanatory variable with two levels:
#### 1=one or more parents smoke, 0=no parents smoke

parentsmoke=as.factor(c(1,0))

#### NOTE: if we do parentsmoke=c(1,0) R will treat this as
#### a numeric and not categorical variable

#### need to create a response vector so that it has counts for both "success" and "failure"

response<-cbind(yes=c(816,188),no=c(3203,1168))
response

#### fit the logistic regression model

smoke.logistic<-glm(response~parentsmoke, family=binomial(link=logit))

#### OUTPUT

smoke.logistic
summary(smoke.logistic)
anova(smoke.logistic)
```

```
--selfjoin.pig
-- For each stock, find all dividends that increased between two dates
divs1     = load 'NYSE_dividends' as (exchange:chararray, symbol:chararray,
                date:chararray, dividends);
divs2     = load 'NYSE_dividends' as (exchange:chararray, symbol:chararray,
                date:chararray, dividends);
jnd       = join divs1 by symbol, divs2 by symbol;
increased = filter jnd by divs1::date < divs2::date and
                divs1::dividends < divs2::dividends;
```

Hadoop Data Management: Language Examples



```
package org.myorg;

import java.io.IOException;
import java.util.*;

import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapred.*;
import org.apache.hadoop.util.*;

public class WordCount {

  public static class Map extends MapReduceBase implements Mapper<LongWritable, Text, Text, IntW
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable> output, Repo
      String line = value.toString();
      StringTokenizer
      while (tokenize
        word.set(token
        output.collect
      }
    }
  }

  public static clas
    public void reduc
      int sum = 0;
      while (values.h
        sum += values.
      }
      output.collect(
    }
  }

  public static voi
    JobConf conf = ne
    conf.setJobName("wordcount");

    conf.setOutputKeyClass(Text.class);
    conf.setOutputValueClass(IntWritable.class);

    conf.setMapperClass(Map.class);
    conf.setCombinerClass(Reduce.class);
    conf.setReducerClass(Reduce.class);

    conf.setInputFormat(TextInputFormat.class);
    conf.setOutputFormat(TextOutputFormat.class);

    FileInputFormat.setInputPaths(conf, new Path(args[0]));
    FileOutputFormat.setOutputPath(conf, new Path(args[1]));

    JobClient.runJob(conf);
  }
}
```

```
#### define the explanatory variable with two levels:
#### 1=one or more parents smoke, 0=no parents smoke

parentsmoke=as.factor(c(1,0))

#### NOTE: if we do parentsmoke=c(1,0) R will treat this as
#### a numeric and not categorical variable

#### need to create a response vector so that it has counts for both "success" and "failure"

response<-cbind(yes=c(816,188),no=c(3203,1168))
response
```

For your organization, how important is "SQL on Hadoop"—that is, Hadoop tools that support ANSI-standard SQL for queries against data managed on Hadoop?

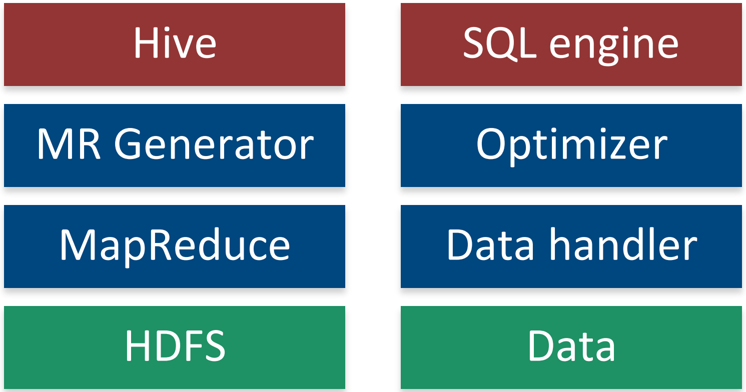
Must have SQL on Hadoop 69%
Would be nice to have SQL on Hadoop 27%
Don't need SQL on Hadoop 4%

Figure 15. Based on 99 respondents.

```
ates
divs1      = load 'NYSE_dividends' as (exchange:chararray, symbol:chararray,
                 date:chararray, dividends);
divs2      = load 'NYSE_dividends' as (exchange:chararray, symbol:chararray,
                 date:chararray, dividends);
jnd        = join divs1 by symbol, divs2 by symbol;
increased  = filter jnd by divs1::date < divs2::date and
                 divs1::dividends < divs2::dividends;
```

Hadoop SQL Support: Several Approaches

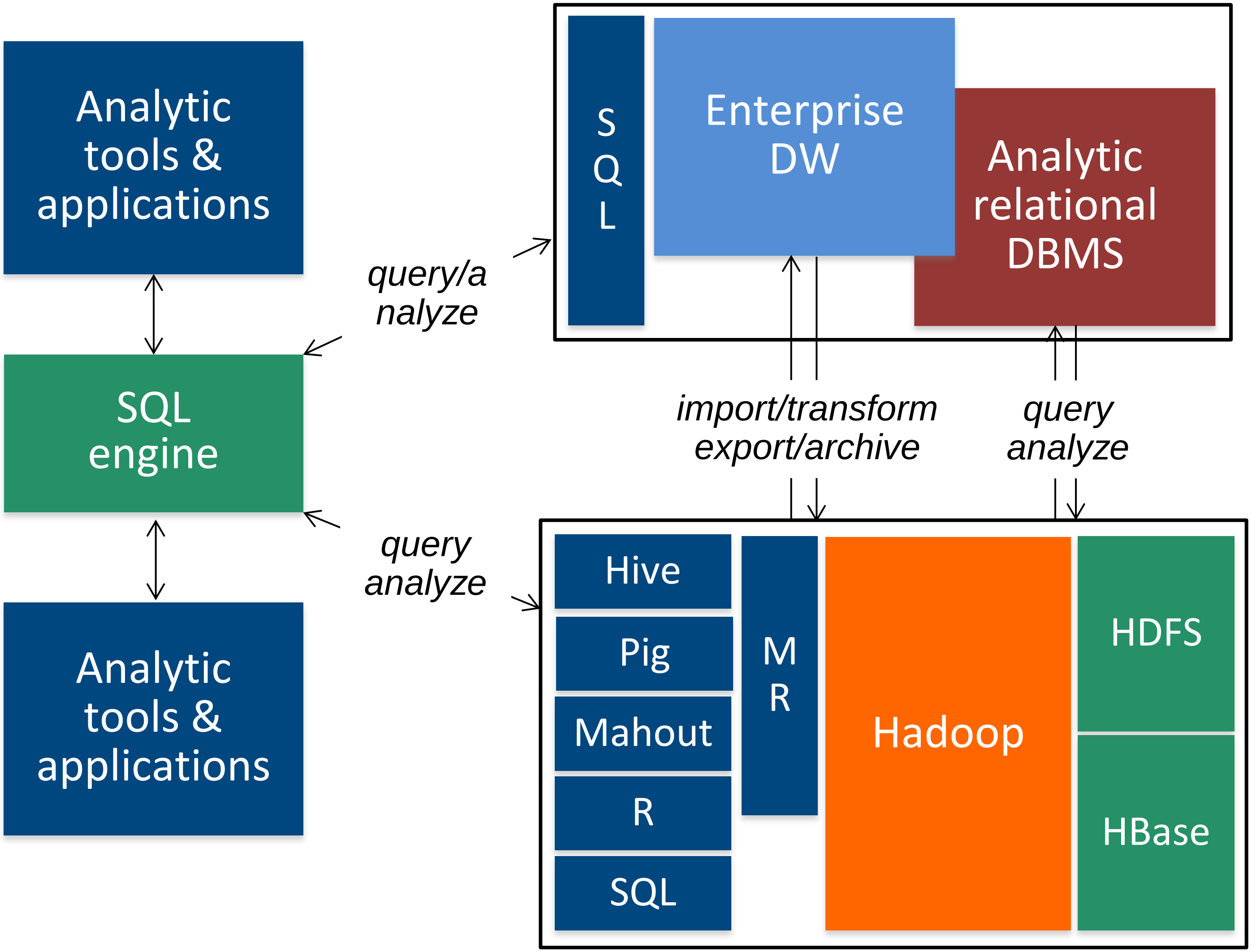
1. Improve the functionality and performance of Hive.
2. Add an SQL layer that bypasses Hive and MapReduce and accesses the Hadoop data directly.
3. Develop new on-disk and/or in-memory Hadoop data handlers and data formats that are more suited to *ad hoc* query processing.
4. Build an SQL query engine that uses a query splitter to route query fragments to one or more underlying data handlers (HDFS, HBase, relational, search index, etc.) to access and process the data.



SQL compatibility?

Performance?

Hadoop Integration Examples



Hadoop Today: Key Questions Revisited - 1

What are the use cases for Hadoop?

- Data refinery (including archiving)
- Investigative computing platform for analyzing large volumes of raw data (especially multi-structured data) for specific LOB solutions

What are the TCO considerations for Hadoop?

- Need to consider more than just hardware and software costs
- Other factors include training, development, administration and support costs, and floor space and utility requirements

How mature is the Hadoop ecosystem?

- Still immature (especially in the areas of governance and systems management), but improving rapidly

Hadoop Today: Key Questions Revisited - 2

What are the skill requirements for Hadoop?

- Despite increasing SQL support, Hadoop still requires highly technical skills in areas such as large-scale Linux and Java

Which Hadoop solution should we use?

- Hadoop is not a single product but a set of different components that satisfy a variety of requirements
- Choice is between traditional and "open source" vendors

How do we integrate Hadoop with existing systems?

- Build a modernized data warehouse infrastructure that supports a common and transparent interface to heterogeneous data

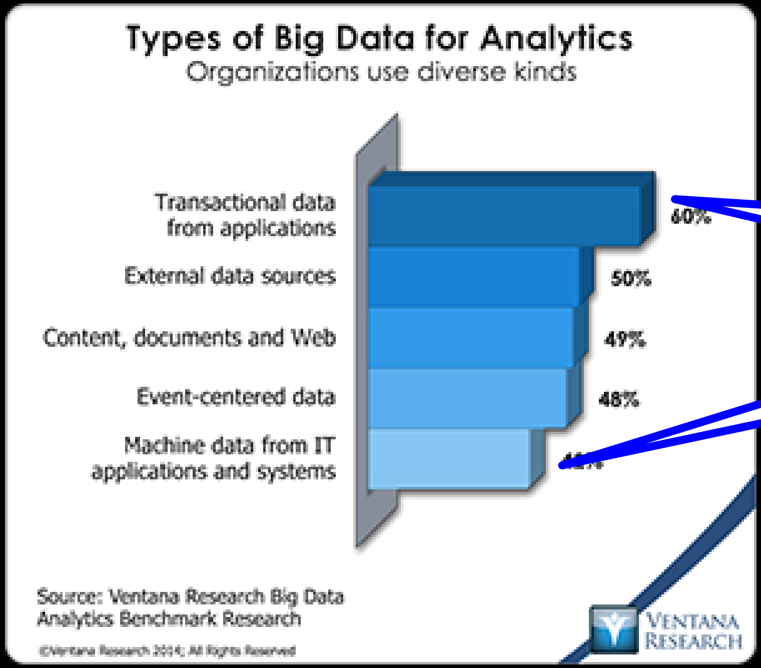
Thanks for Listening

GET ALL THE INFORMATION YOU CAN, WE'LL THINK OF A USE FOR IT LATER.
CHRIS

Hadoop and PureData for Analytics

Integral Parts of the Big Data Ecosystem

Data Comes in Various Types from Various Sources



Source: Ventana Research Big Data Analytics Benchmark Research

©Ventana Research 2014; All Rights Reserved

Some data naturally belongs in Hadoop, while other data could be in Hadoop, the DBMS, or both

Utilize Both Hadoop AND the Data Warehouse

- Optimize Workloads
  - Keep the right, most important data in the data warehouse
  - Move historical reporting, ETL/ELT, exploration to Hadoop
- Free up resources for the workloads that provide the most business value

Big Data and Business Intelligence Ready

Unlocking Data's True Potential

Included with the PureData System for Analytics N3001

Data Warehouse Appliance

The ultimate in simplicity and performance for your Data Warehouse or Data Mart with built-in in-database analytic capability

Built-in, In-Database analytic capability and integration with a variety of 3rd party tools

Exceptional value provided

Business Intelligence
Cognos software, 5 Analytics User licenses, plus 1 Analytics Administrator license

Data Integration & Transformation
InfoSphere DataStage 280 PVUs, 2 concurrent Designer Client licenses and InfoSphere Data Click

Hadoop Data Services
InfoSphere BigInsights Software licenses to manage ~100 TB of Hadoop data

Real-time Analytics
InfoSphere Streams Developer Edition 2 users, non-production licenses

For additional value

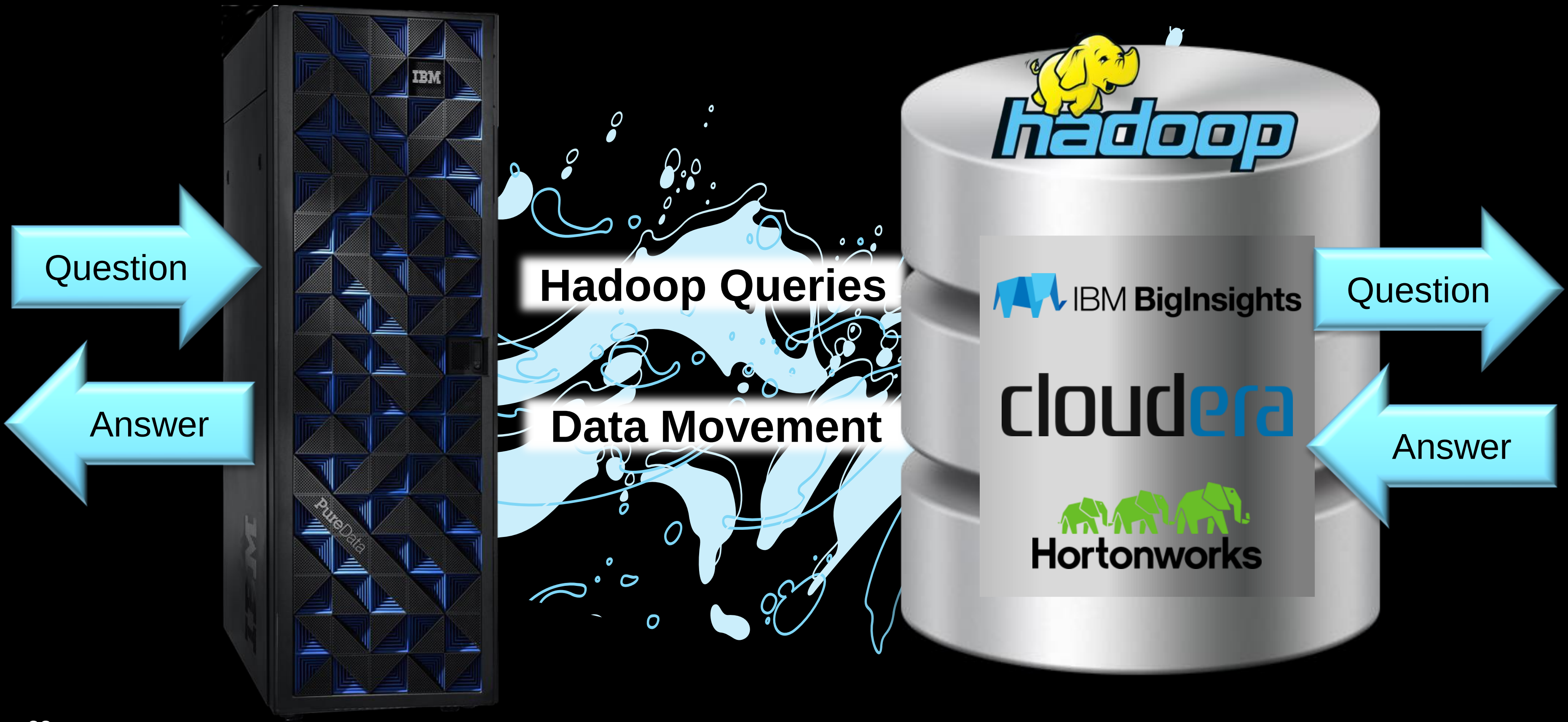
Industry Process & Data Models
Models for Banking, Financial Markets, Healthcare, Insurance, Retail, Telco

IBM InfoSphere Data Privacy and Security for Data Warehousing

IBM Fluid Query

Unifying PureData System for Analytics with Hadoop

Cross platform query & data movement
between PureData System for Analytics and Hadoop



Questions?

Contact Information

If you have further questions or comments:

Colin White, BI Research
info@bi-research.com

Dwaine R. Snow, IBM
dwsnow@us.ibm.com