

# Making Leaders Successful Every Day



# The Patterns Of Big Data

**A Data Management Playbook Toolkit**

**Forrester Research**

**Brian Hopkins**, Principal Analyst

June 11, 2013

# Table of contents: examples by pattern

Pattern	Firm (industry)/vendor — slide numbers
Enterprise data warehouse augmentation	Pharmaceutical company/Cloudera.....21-23 Wealth management firm (financial services)/Composite Software...24-26
Data refinery plus data warehouse (DW) / business intelligence (BI) database management system (DBMS)	edo interactive/Pentaho.....31-33 Vestas Wind Systems (manufacturing)/IBM.....34-36 NK (social media)/Actian.....37-39 Opera Solutions (IT)/LexisNexis.....40-42 Rubicon (digital marketing)/MapR.....43-45 Razorfish (digital marketing)/Teradata Aster.....45-48
All-in-one	Sears (retail)/Datameer.....54-57 Telecommunications Company/Datameer.....58-59
Hub-and-spoke	Pharmaceutical company/Cloudera.....64-65 Internet analytics firm (telecommunications)/Hortonworks.....66-68

# Table of contents: examples by company and industry

Industry	End user firms	Pattern (vendor) — slides
Digital marketing	edo interactive	Data refinery plus DW / BI DBMS (Pentaho).....31-33
	Razorfish	Data refinery plus DW / BI DBMS (Teradata Aster).....46-48
	Rubicon	Data refinery plus DW / BI DBMS (MapR).....43-45
Financial services	Wealth management firm	EDW augmentation (Composite).....24-26
Pharmaceutical	Pharmaceutical company	EDW augmentation (Cloudera).....21-23 Hub-and-spoke (Cloudera).....64-65
IT	Opera Solutions	Data refinery plus DW / BI DBMS (LexisNexis).....40-42
Manufacturing	Vestas Wind Systems	Data refinery plus DW / BI DBMS (IBM).....34-36
Retail	Sears	All-in-one (Datameer).....54-57
Social media	NK	Data refinery plus DW / BI DBMS (Actian).....37-39
Telecommunications	Telecommunications company	All-in-one (Datameer).....57-59
	Internet analytics firm	Hub-and-spoke (Hortonworks).....66-68

# Table of contents: examples by technology vendor

Vendor — product	Patterns/industry — slide numbers
Action — Vectorwise	Data refinery plus DW / BI DBMS (social media).....37-39
Composite Software — Server	EDW augmentation (financial services).....24-26
Cloudera — Cloudera Hadoop Distribution	EDW augmentation and hub-and-spoke (healthcare).....21-23 .....64-65
Datameer	All-in-one (retail and telecommunications).....54-59
Hortonworks — Hortonworks Data Platform	Hub-and-spoke (telecommunications).....66-68
IBM — InfoSphere BigInsights	Data refinery plus DW / BI DBMS (manufacturing).....34-36
LexisNexis — HPCC Systems	Data refinery plus DW / BI DBMS (IT).....40-42
MapR — M5	Data refinery plus DW / BI DBMS (digital marketing).....43-45
Pentaho	Data refinery plus DW / BI DBMS (digital marketing).....31-33
Teradata — Aster	Data refinery plus DW / BI DBMS (digital marketing).....46-48

# Big data patterns research methodology

- › **This toolkit is a companion to our data management playbook strategic plan report.** See Forrester's June 12, 2013, "Deliver On Big Data Potential With A Hub-And-Spoke Architecture" report to understand how firms are leveraging big data technology to solve problems.
- › **The objective of this research is to see what early adopters have actually done.** Many think big data is synonymous with huge volumes of exotic new external data like mobile, social, machine, and log files. But the reality is that firms are taking a pragmatic approach focused on wringing value from internal data first.
- › **We interviewed 11 firms with production implementations.** We worked with vendors to identify 11 firms we could talk to about their experience with big data implementations. We analyzed 12 examples and present the results here.
- › **We uncovered four patterns in big data production implementations.** The companion research piece identifies a total of seven technology patterns, but some are only now emerging and we did not find examples of clients willing to speak to us. The four patterns we found all lead to a new data management approach that Forrester calls "hub-and-spoke," which delivers on the hyperflexibility your business needs to be successful in the digital age.

# Purpose of this toolkit

## CLARIFY AND ILLUMINATE THE MOST COMMON BIG DATA PATTERNS

- › **Use this research as a basis for business conversations.** Study the problems we identified and the results firms told us about. Use these examples in your business strategy conversation to stimulate discussions about what is really possible.
- › **Use this research to understand technology architecture patterns.** In formulating strategies to provide more flexibility and lower data cost to your business, study these patterns and lessons learned to identify the data management technology building blocks your firm really needs.
- › **Use these patterns as part of your vendor selection and solution design.** We attempted to be very broad in the types of technologies we evaluated as part of the patterns. We want to thank the vendors and users that cooperated in providing this information. We have included one page for product information from each participating vendor; use these to engage in your investigations.

# Key takeaways

- › Big data is about dealing with more data with greater agility and cost-effective performance.
- › None of our examples used social or pure unstructured, external content, despite the hype.
- › We found that production implementations generally follow four of the seven patterns we identified in our report.
- › These patterns illustrate the evolving hub-and-spoke data management architecture with an “extract-load-transform” approach.
- › Improvements in Hadoop, streaming platforms, and in-memory data technology will have a profound impact on the future of big data solutions.

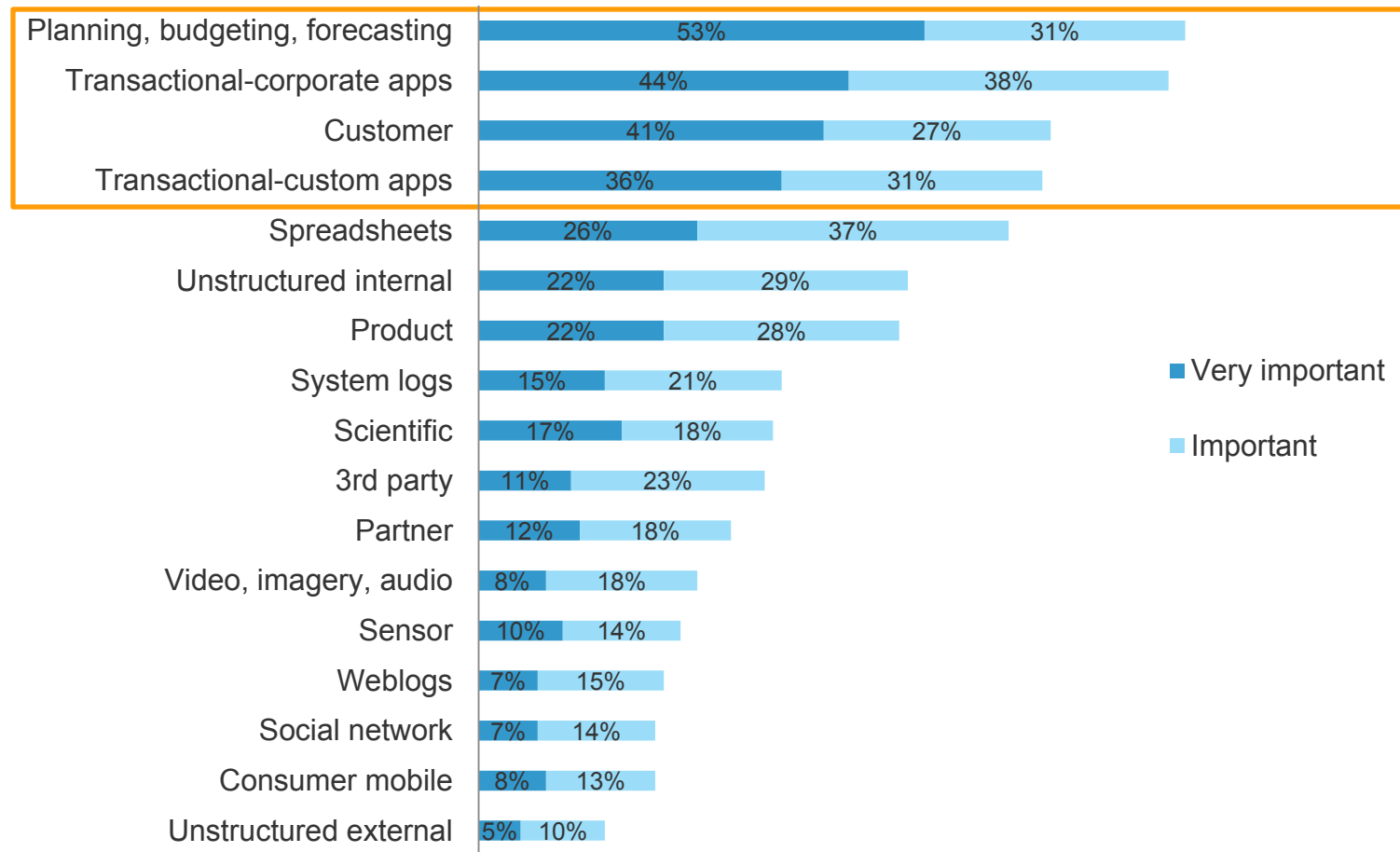


***Forrester defines “big data” as techniques and technologies that make handling data at extreme scale affordable.***

Source: September 30, 2011, “Expand Your Digital Horizon With Big Data” Forrester report

***So what? When the unaffordable becomes affordable, the impossible becomes possible.***

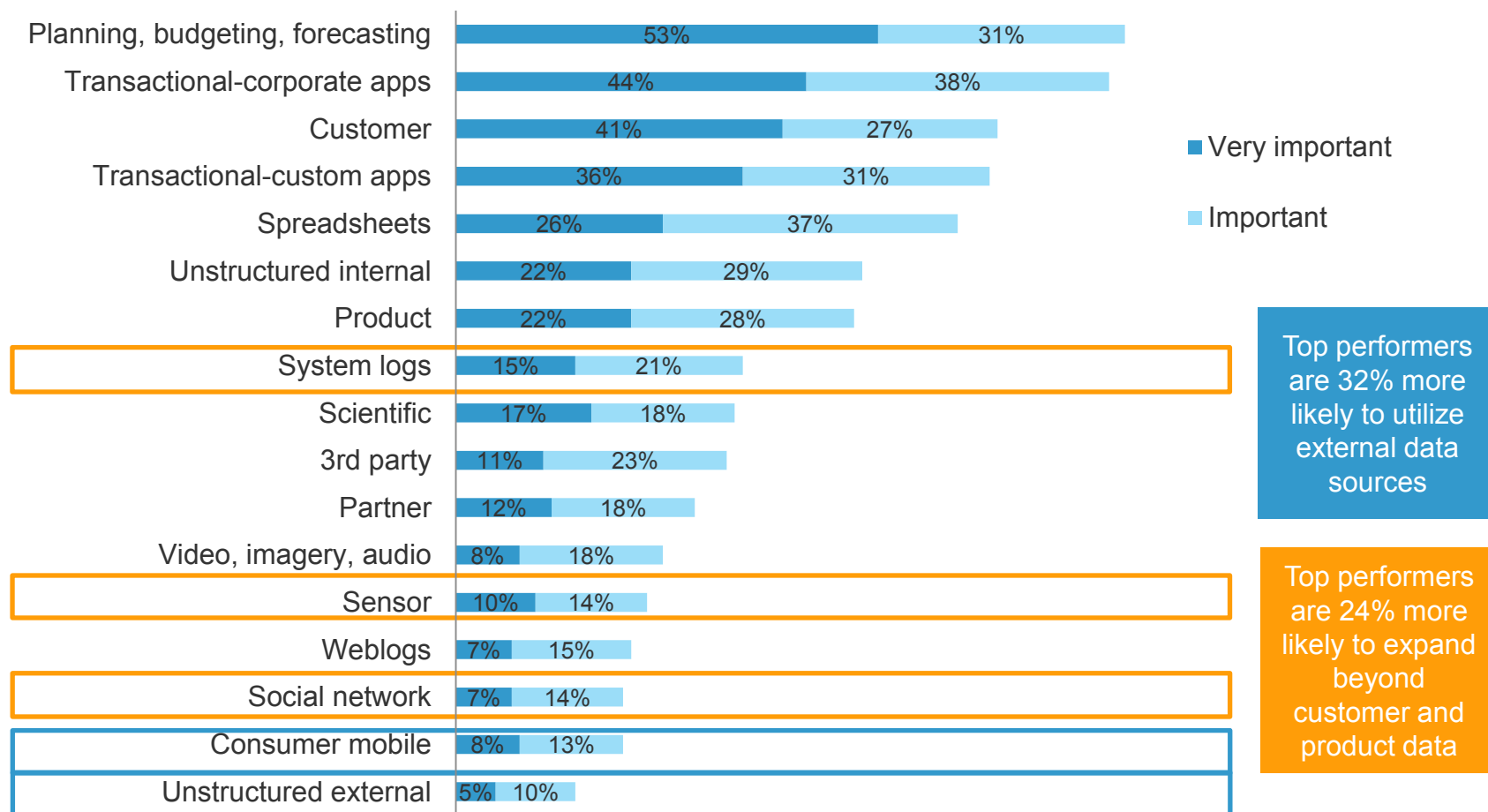
# Financial, customer, and transactional data in core systems is most important to business strategy



Base: 603 global decision-makers involved in business intelligence, data management, and governance initiatives

Source: Forrsights Strategy Spotlight: Business Intelligence And Big Data, Q4 2012

# Top performers (firms with greater than 15% annual growth) utilize more diverse data sources



Base: 603 global decision-makers involved in business intelligence, data management, and governance initiatives

Source: Forrsights Strategy Spotlight: Business Intelligence And Big Data, Q4 2012

# Top performers (greater than 15% annual growth) realize they need more

“What best describes your firm’s current usage/plans to adopt big data technologies and solutions?”

Average performers are thinking about big data

Rest of organizations (<15% growth) (N = 482)



■ Planning to implement in more than 1 year

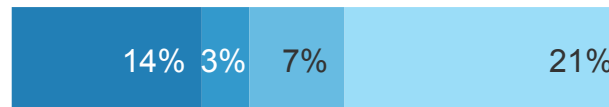
■ Planning to implement in the next 12 months

■ Implemented, not expanding

■ Expanding/upgrading implementation

Top performers are expanding their big data implementations

High performance (>15% growth) (N = 58)



Base: 603 global decision-makers involved in business intelligence, data management, and governance initiatives  
Source: Forrsights Strategy Spotlight: Business Intelligence And Big Data, Q4 2012

# We spoke to early adoption leaders with production big data experience

## OUR INTERVIEWS UNCOVERED MULTIPLE EXAMPLES THAT WE GROUPED INTO FOUR PATTERNS

	Description	Technology considerations
EDW augmentation	An enterprise data warehouse (EDW) remains the locus of analytic data architecture, but cold data is offloaded to hub. High-volume data that is not cost effective to move into a warehouse is added and analyzed using new or existing tools, but primary analytics remain in the data warehouse and marts.	Distributed data hub options: HDFS, HBase Existing BI tools are used against the EDW/marts, data virtualization may be used to integrate NoSQL data with existing BI tools; specialized analytic packages may be added for analytics of hub data directly.
All-in-one	A distributed data system is implemented for long-term, high-detail big data persistence in the hub and analytics without employing a business intelligence database for analytics. Low level code is written or big data packages are added that integrate directly with the distributed data store for extreme-scale operations and analytics.	Distributed data hub options: Hadoop, HBase, Cassandra, MongoDB, LexisNexis BI tools specifically integrated with or designed for distributed data access and manipulation are needed. Data operations either use BI tools that provide NoSQL capability or low level code is required (e.g., MapReduce or Pig script). May use data virtualization technology to integrate other enterprise data and big data data with existing BI tools.
Data refinery plus DW / BI DBMS	The distributed hub is used as a data staging and extreme-scale data transformation platform, but long-term persistence and analytics is performed by a BI DBMS using SQL analytics.	Distributed data hub options: Hadoop, LexisNexis, Cassandra BI database is biggest choice: See Forrester's June 2, 2011, "It's The Dawning Of The Age Of BI DBMS" report. BI tools with Hadoop integration may be used for data manipulation or may write low level scripts (Pig), or code (MapReduce).
Hub-and-spoke	An evolution of the EDW augmentation, all-in-one, and data refinery plus DW / BI DBMS pattern that provides multiple options for both hub-and-spoke technologies. Data may be harmonized and analyzed in the hub or moved out to spokes when more quality and performance is needed, or when users simply want control.	All the options in the previous three patterns, plus the data hub may shift from one physical hub to a logical or distributed one, in which different data platforms work together seamlessly to capture raw data and maintain it in a minimally harmonized and useful stage. For example, EMC, IBM, Microsoft, and Oracle are beginning to provide tightly integrated data warehouse appliances and distributed data store (like Hadoop). If the flow and query of data is seamless, we consider this to be a data hub, even though the hub contains a BI DBMS.

# Emerging patterns we did not find

**WE HAVE SEEN EXAMPLES OF THESE THREE PATTERNS BUT DID NOT FIND PRODUCTION EXAMPLES THAT MET OUR CRITERIA FOR THIS TOOLKIT**




Our criteria for this research was that we could speak with a user that has a production solution. We may update this toolkit in the future with examples of these three immature patterns as we find firms willing to talk with us.

	Description	Technology considerations
Standalone package	Buy a packaged big data analytic tool to meet department needs rapidly. Uses are generally limited to the capabilities of the tools. Most are focused on customer intelligence and marketing use cases.	Examples of packaged big data applications: KXEN, nPario, and NGData
Streaming analytics	A streaming analytics package solution is deployed to capture and analyze high-velocity data as it “streams” through the system.	Distributed data hub options: none initially, but may add later as part of path to hub-and-spoke Streaming package examples: IBM InfoSphere Streams, SQLstream, Apache S4, and Storm
Hub-and-spoke plus in-memory	Very early pattern emerging that extends hub-and-spoke with in-memory and with an elastic caching or data grid technology to provide very high performance, embedded, or interactive analytics without using a BI DBMS	Hub options are the same as for the hub-and-spoke pattern. In-memory data grid platforms examples: Platfora, ScaleOut Software, Tibco Software

Note: The June 12, 2013, “Deliver On Big Data Potential With A Hub-And-Spoke Architecture” Forrester report defines the seven patterns, but three of them are immature or nascent without production examples we could find. The four patterns on the previous slide are covered in detail in this toolkit.

# Basic pattern building blocks


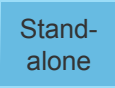

## WE IDENTIFIED SIX BUILDING BLOCKS IN THE PATTERNS

Building block		Description	Considerations
Distributed data hub		The center of the architecture; provides a low-cost data persistence capability that meets minimum requirements for availability, security, and recovery, while exposing data for low-level transformation and analytics	Technology choices: Hadoop, other NoSQL, open source, or vendor supported, use of advanced technologies such as in-memory data grids, incorporation of mainframe data, integrated unstructured and structured data platforms, loading and disposal processes, harmonization standards, metadata, cloud versus dedicated options
Data services		Includes both contextual services such as data quality, master data management, metadata and modeling, and delivery services such as federation/virtualization, transformation, movement, and security services that operate on the hub and on the spokes	Master data management strategy and technology choices, integration approach and technology choices, level of quality, service performance, service availability to hub or spokes, vertical and horizontal scaling, cloud service utilization
Enterprise data warehouse/ departmental BI databases		High-performance BI database appliances and/or homegrown data warehouse solutions that are appropriately spokes. Provides high availability, low latency SQL analytics.	Enterprise versus departmental implementations, data storage cost, analytics requirements for latency and user access, BI tool integrations, loading technology and performance, skills of users, volume and velocity projects versus tool performance characteristics. Examples: Greenplum, Netezza, Teradata Aster, and Exadata.

Note: These building blocks emerged from our assessment of the big data implementations.

# Basic pattern building blocks (cont.)

## WE IDENTIFIED SIX BUILDING BLOCKS IN THE PATTERNS

Building block		Description	Considerations
Big data analytics packages		Packages applications that provide data operations and analytic tools that interact directly with hub data	Integration with NoSQL vendors, needs and skills of users, volume and velocity projects versus tool performance characteristics. Examples: Datameer, Pentaho*
BI and analytics packages		Traditional business intelligence and analytics packages that do not access, analyze, or operate on data in the hub. Instead they access processed data in an operational or structured analytic data store.	What packages to buy for the functionality needed, how it gets supported, how data is sourced into the package. Examples include BusinessObjects, Cognos, Tableau Software, QlikView. As more vendors add Hadoop integration to their capabilities, the distinction between these tools and big data analytics packages will blur.
Data science workbench		Tools used by data scientists to explore, manage hub data, stage for data mining, and model development, management, and deployment	Type of operations, data requirements, departmental versus enterprise team, sandbox and staging area needs, model to application integration, model to BI DB integration, analytic and exploration tool needs, operating procedures, security. Example technology: SPSS, R, SAS, Mahout, MapReduce






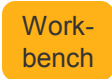
\*Note: Pentaho can function as a big data analytics package or a BI and analytics package, depending on how it's employed.

Note: These building blocks emerged from our assessment of the big data implementations.



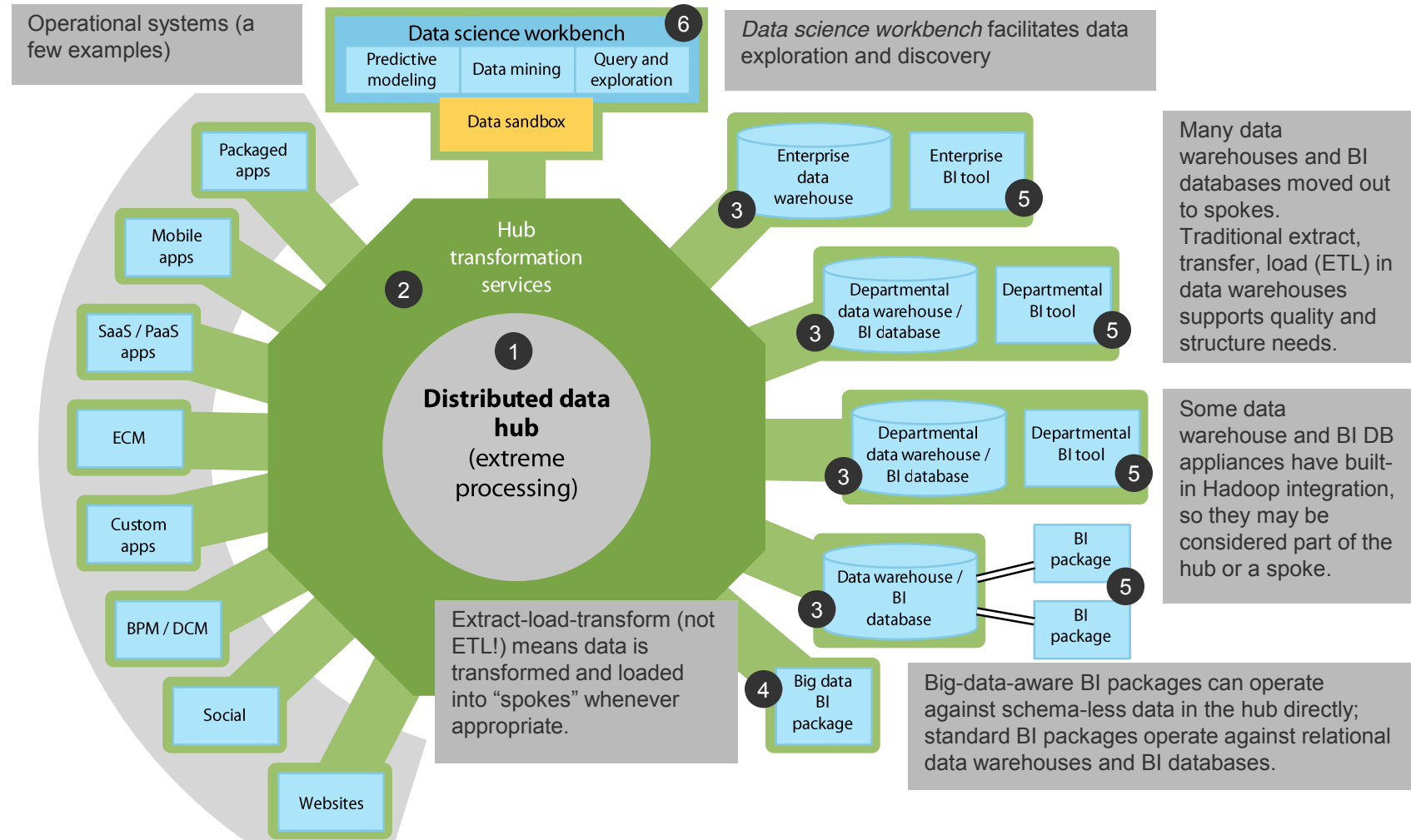
# Key

## WE USE NUMBERED CIRCLES TO MAP HUB-AND-SPOKE COMPONENTS TO PATTERNS AND EXAMPLES

- 1 = data hub 
- 2 = data services 
- 3 = enterprise data warehouse/departmental BI database 
- 4 = big data aware analytics packages 
- 5 = standalone BI and analytics packages 
- 6 = data science workbench 

# Hub-and-spoke architecture

## THE BUILDING BLOCKS CREATE A HUB-AND-SPOKE DATA MANAGEMENT ARCHITECTURE



In our June 12, 2013, "Deliver On Big Data Potential With A Hub-And-Spoke Architecture" report, we present a more abstract picture of the hub-and-spoke. This diagram reduces that picture to a more concrete level.



# Pattern: enterprise data warehouse augmentation

**Primary purpose:** make existing data warehouse environment more cost effective

**Secondary purpose:** add more data and conduct rapid analysis in the hub

Examples:

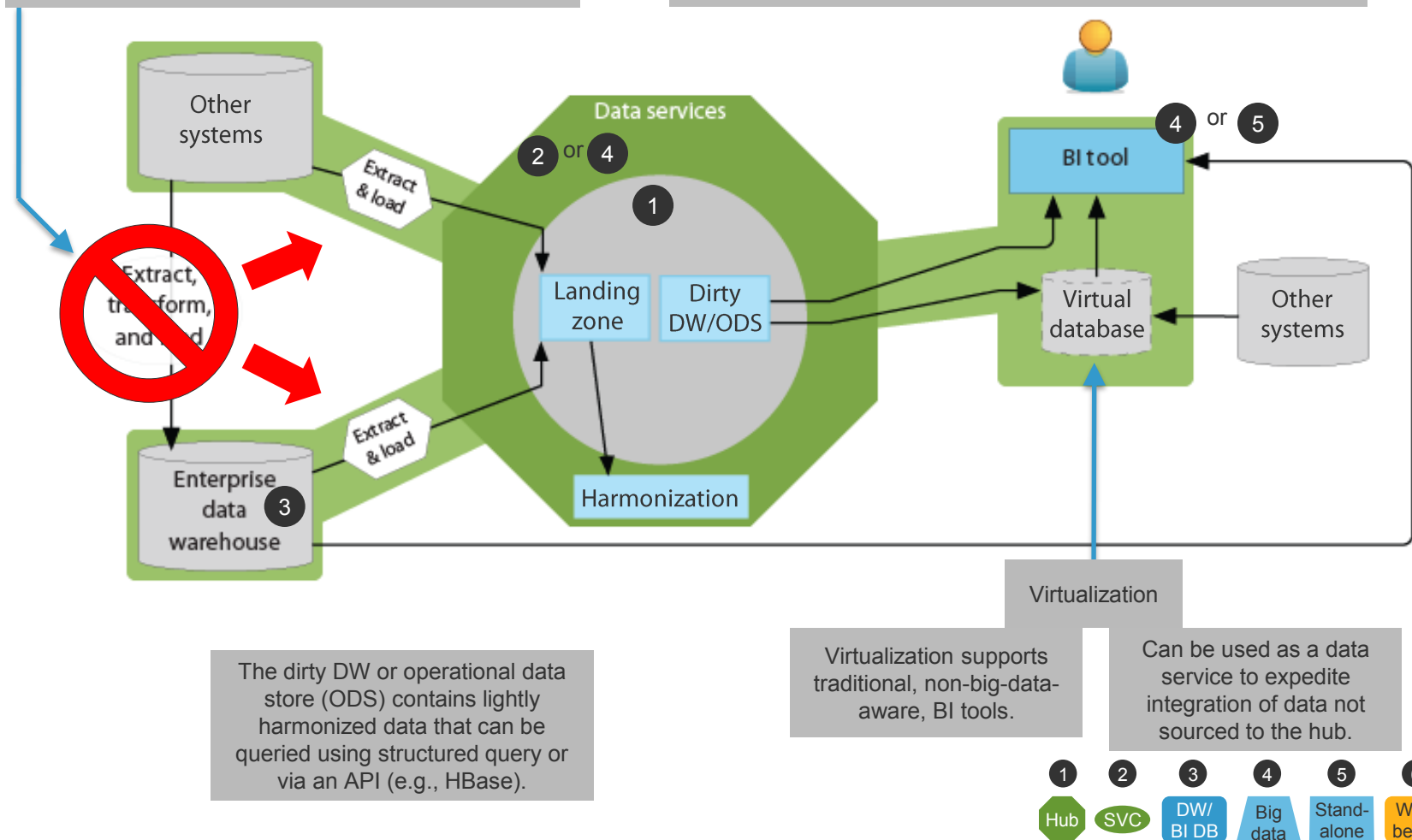
Pharmaceutical company.....21-23

Wealth management firm (financial services).....24-26

# Enterprise data warehouse augmentation pattern

The main feature of this pattern is that some data warehouse loads are rerouted to use the big data hub's data services.

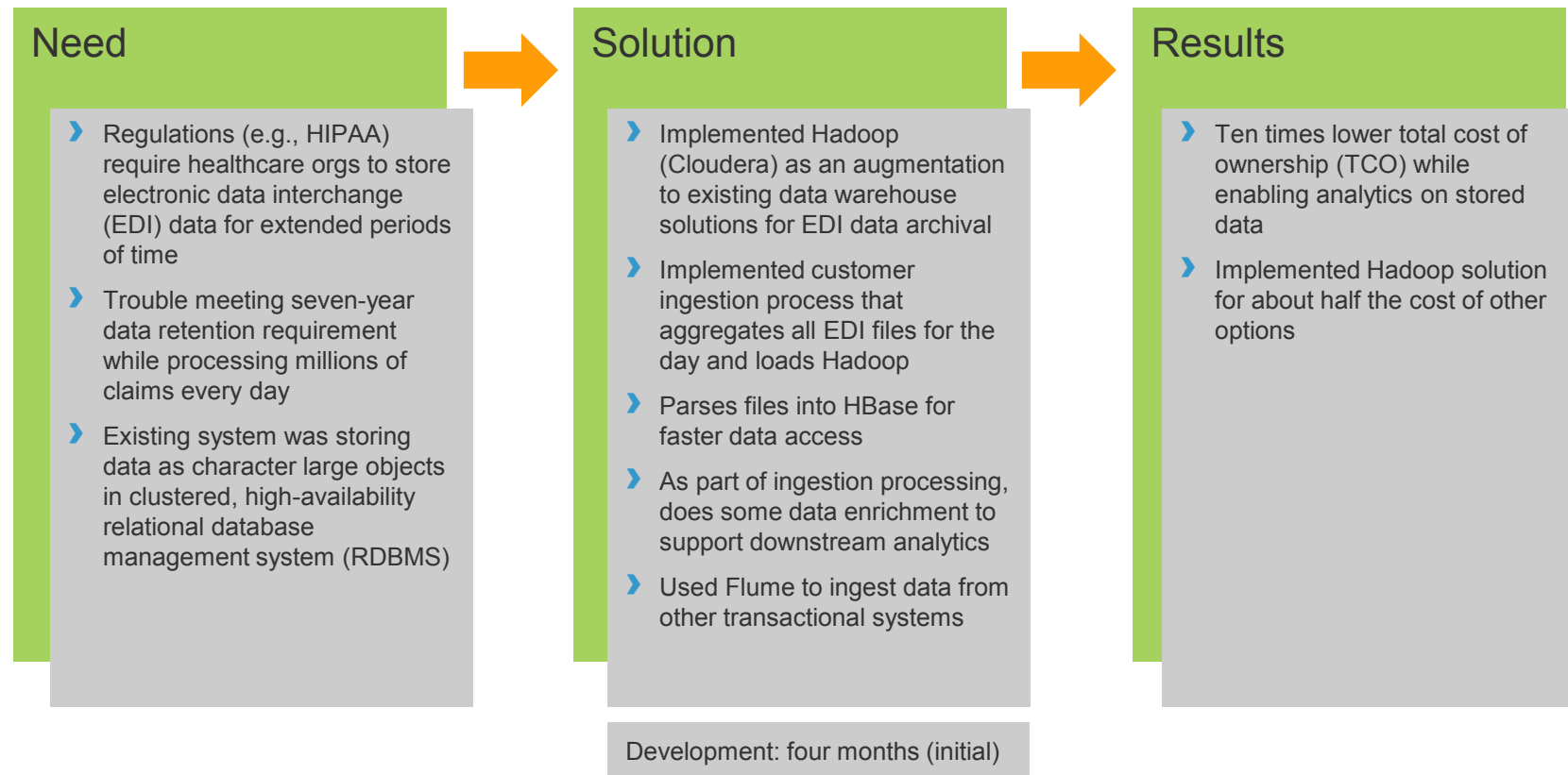
Users employ the same BI tools they are used to. Note: some BI tools have integrations with data hub platforms, others do not and need an intermediary such as a data virtualization layer



The dirty DW or operational data store (ODS) contains lightly harmonized data that can be queried using structured query or via an API (e.g., HBase).

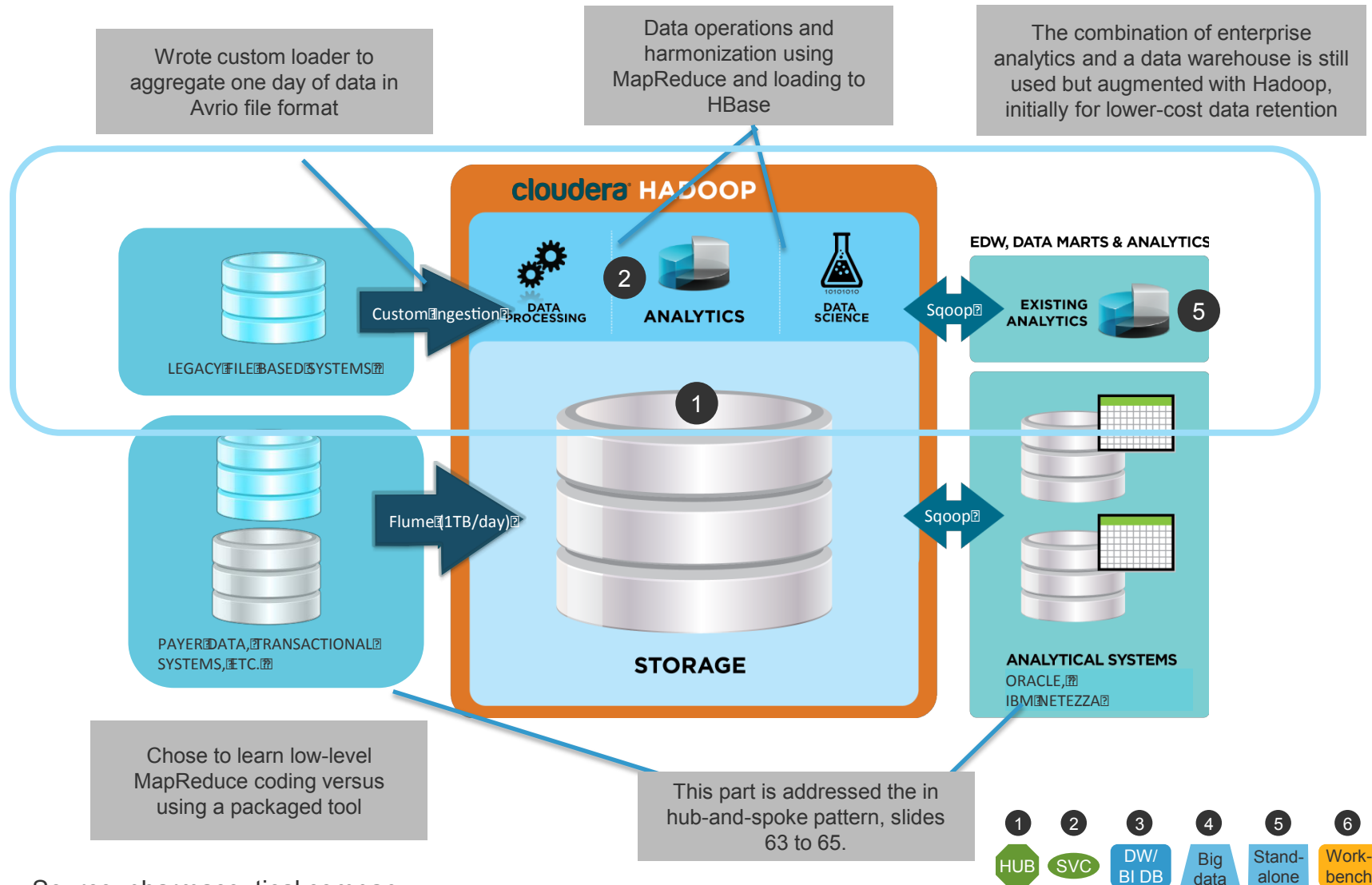
# Example: pharmaceutical company

## EXAMPLE — ENTERPRISE DATA WAREHOUSE AUGMENTATION PATTERN



Note: The pharmaceutical company appears twice, illustrating a firm initially pursuing one pattern then evolving to hub-and-spoke.

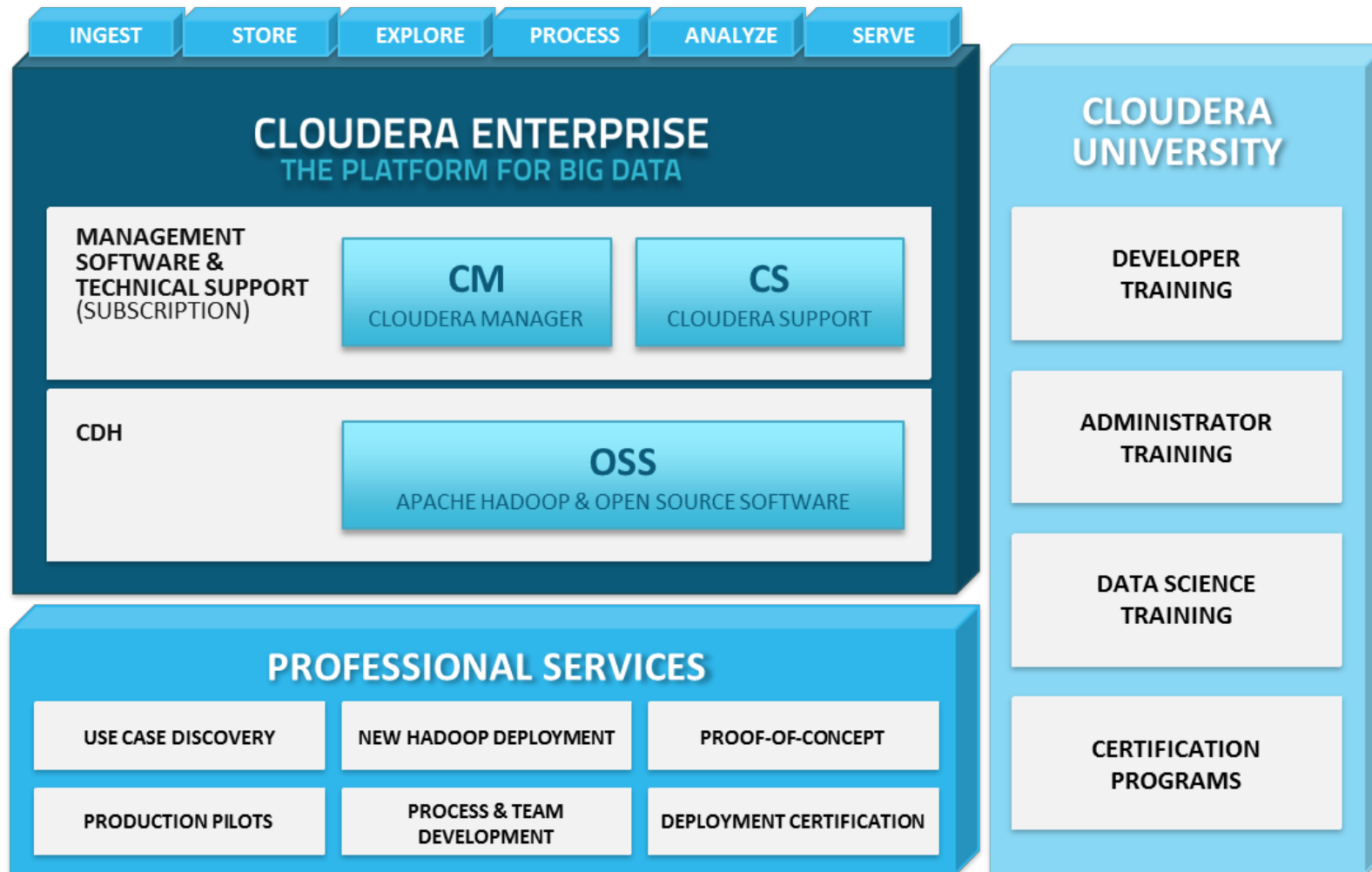
# Pharmaceutical company — conceptual solution architecture



Source: pharmaceutical company

# Vendor information

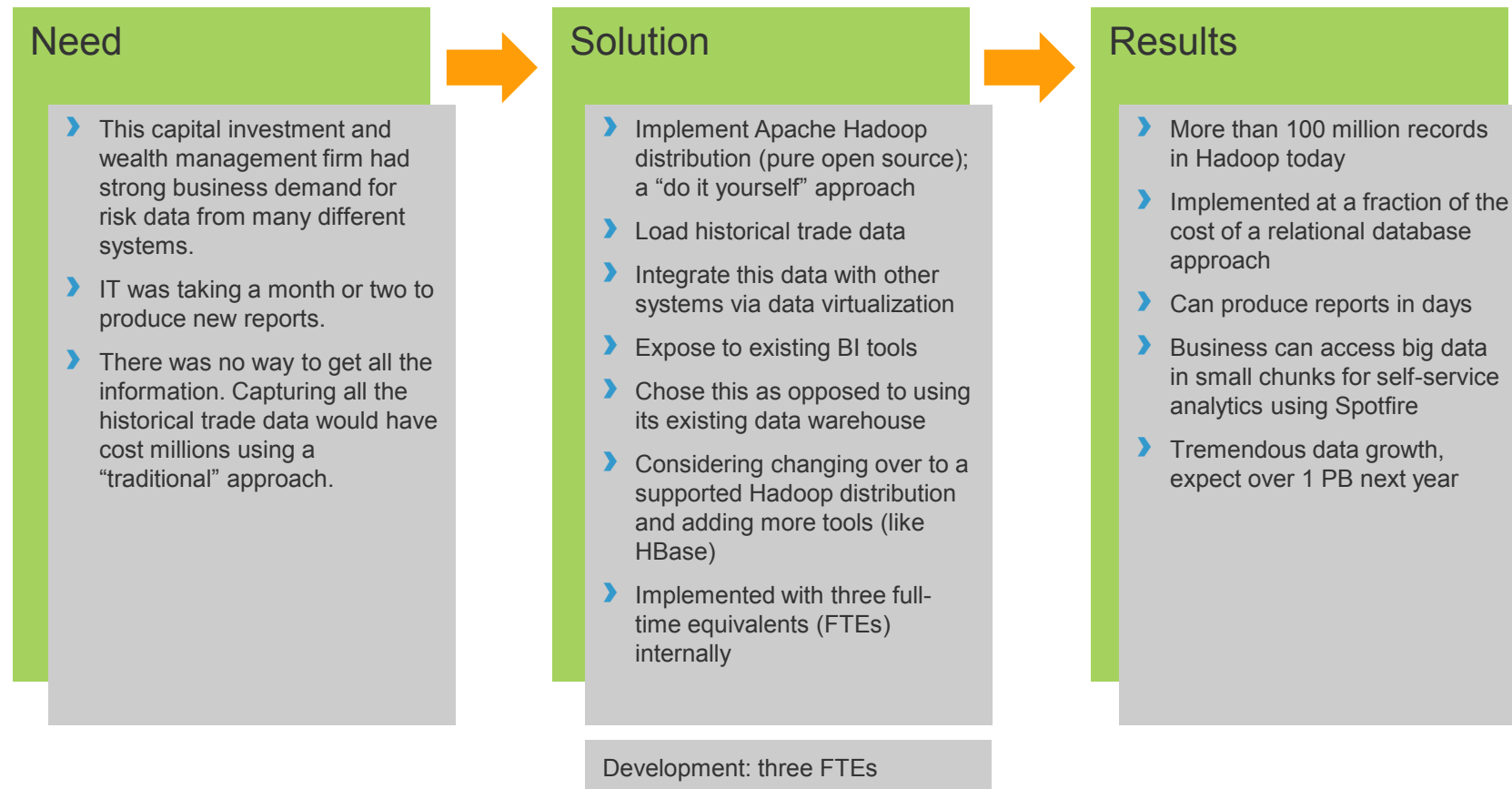
## CLOUDERA



Source: Cloudera

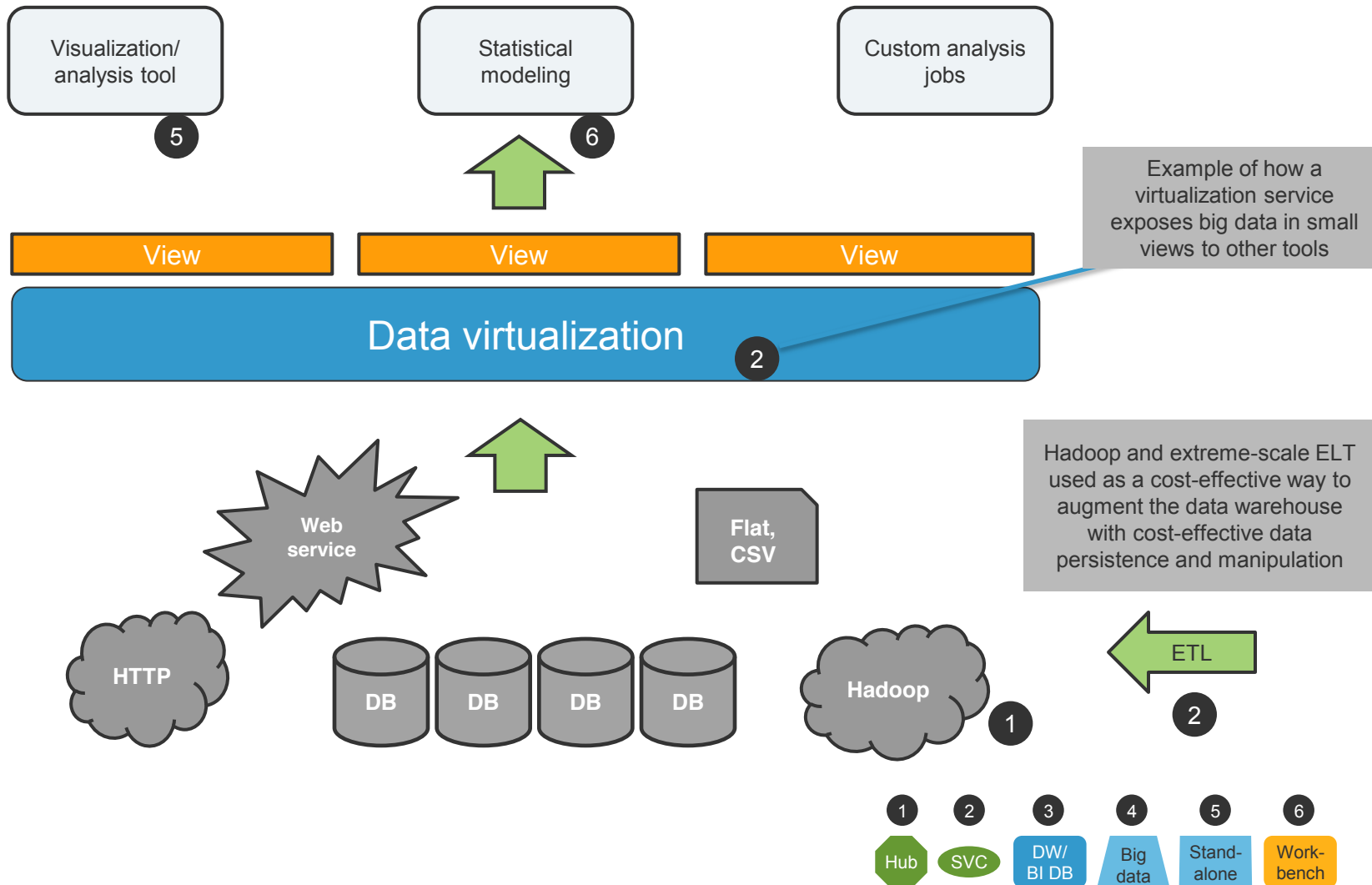
# Example: wealth management firm (financial services)

## EXAMPLE — ENTERPRISE DATA WAREHOUSE AUGMENTATION





# Wealth management firm — conceptual solution architecture



For detailed discussion of the impact data virtualization is having on firm's data architectures, see the June 15, 2011, "Data Virtualization Reaches Critical Mass" Forrester report.

# Vendor information

## COMPOSITE SOFTWARE



### Composite data virtualization platform

#### Development environment

Discovery

Studio

Performance plus adapters

#### Runtime server environment

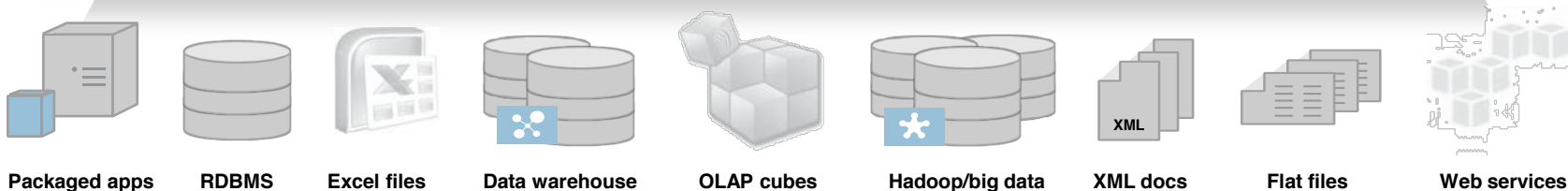
Composite information server

#### Management environment

Manager

Monitor

Active cluster



Source: Composite Software

# Forrester's point of view

- › The enterprise data warehouse augmentation pattern is the easiest to fund when big data is perceived as an “IT thing.”
- › The benefits are tangible and immediately realized, the business impacts are manageable, and the upside is huge.
- › We suggest:
  - Start by doing a five-year TCO calculation on all data in your data warehouse or data mart environments. Include the cost of integrating all that data.
  - Do an analysis of how much data in your data warehouse environment has no or low analytic usage.
  - Determine if any of this cold data has retention requirements that drive its storage.
  - Develop a five-year TCO for an open source distributed data hub. See if a business case can be made.
- › Your biggest technology strategic decisions are:
  - How to enable analytics on data in the hub. Data virtualization and BI tools with big data tool integration capability can help.
  - What distributed data hub technology to choose, and how to acquire the skills.
  - The approach to data movement and the level to which hub data is harmonized.

# Lessons learned from users

- › Dealing with raw data is messy. This is a new way of thinking. You need a data harmonization and integration approach that delivers a minimum quality level. Think minimum viable quality, not completely clean data. An enterprise data model is essential for semantic consistency, but the data doesn't have to conform completely to the model to be useable — this is one reason big data delivers hyperflexibility.
- › Data access will be challenging, both politically and technically. Ensure your strategy and business case is strong enough to overcome these challenges. Be sure you have support from the top to overcome parochial concerns. Define who owns the data once you have sourced and harmonized it.
- › Tool selection makes all the difference. The devil is in the details — understand what “integrates with Hadoop” really means in terms of specific versions of Hadoop components, the specific integration functionality, and quality of community and vendor support available.

# Pattern: data refinery plus DW / BI DBMS

**Primary purpose:** lower the cost of data capture and operations while loading a structured business intelligence database for low latency structured analytics

## Examples:

edo interactive.....	31-33
Vestas (manufacturing) .....	34-36
NK (social media) .....	37-39
Opera Solutions (IT) .....	40-42
Rubicon (digital marketing) .....	43-45
Razorfish (digital marketing) .....	46-48

# Data refinery plus DW / BI DBMS

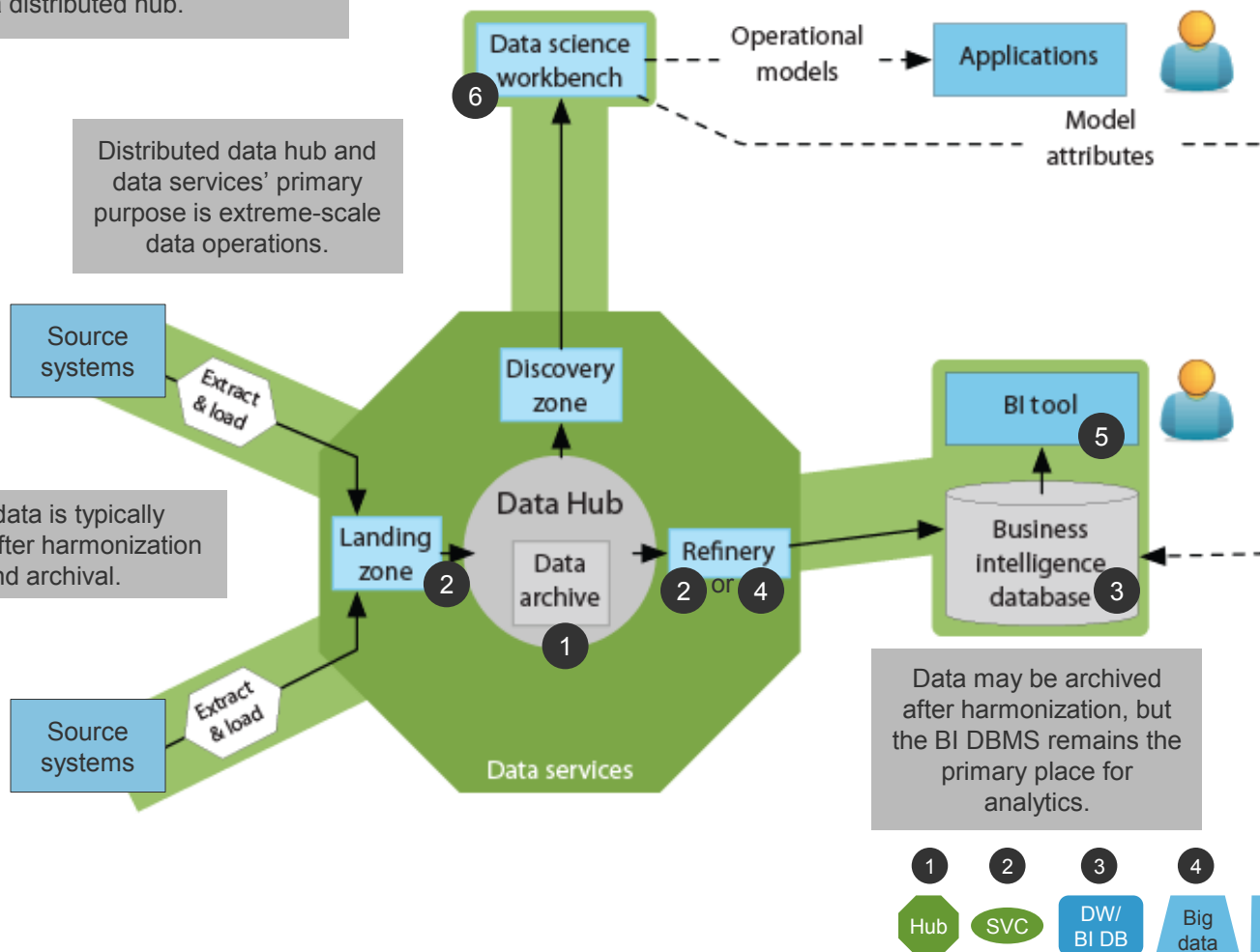
The primary feature of this pattern is the loading of an EDW or other BI DBMS from a distributed hub.

Data science work is primarily developing models and attributes for deployment to embedded analytics and BI DBMS.

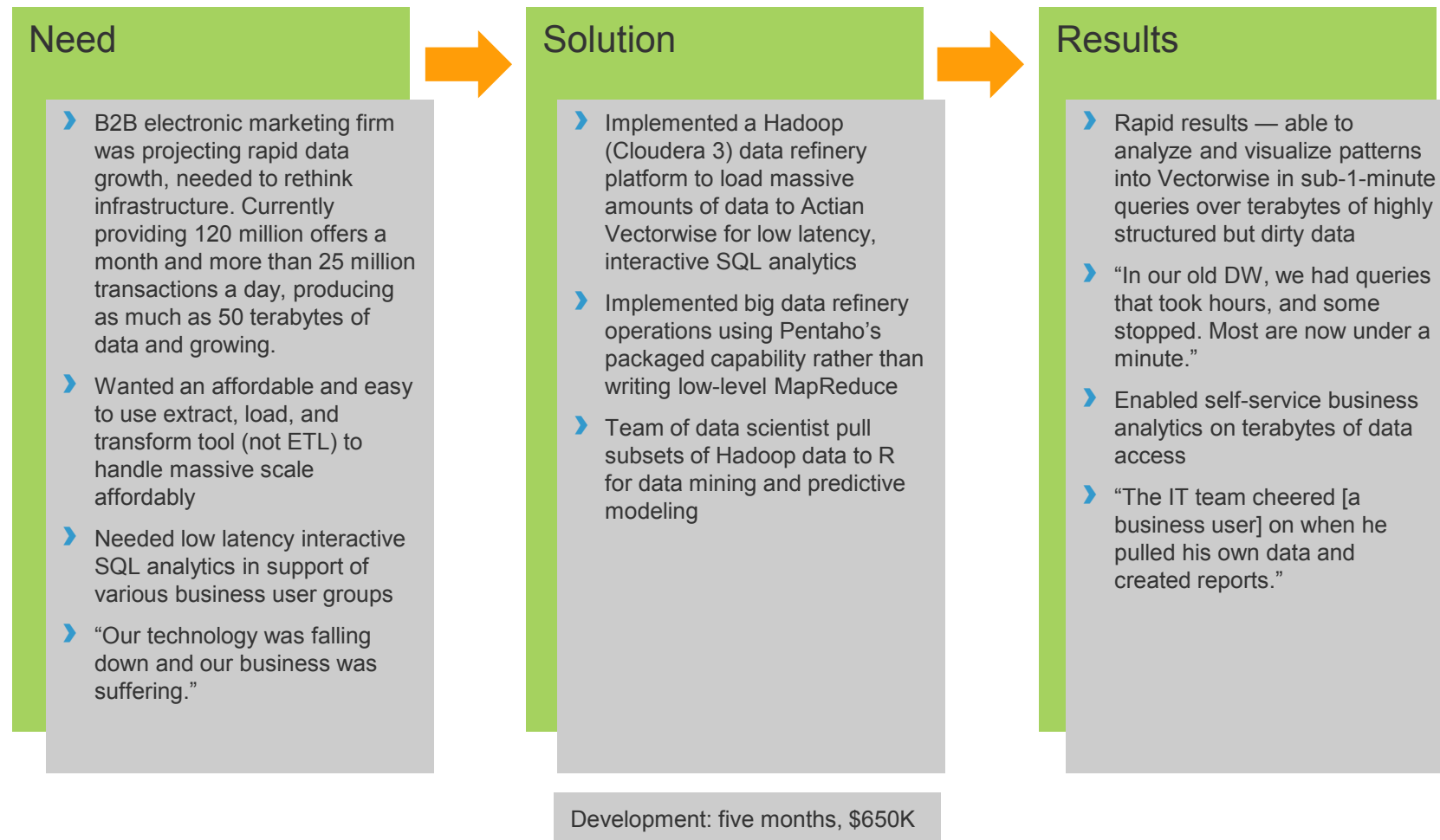
Distributed data hub and data services' primary purpose is extreme-scale data operations.

Raw data is typically purged after harmonization and archival.

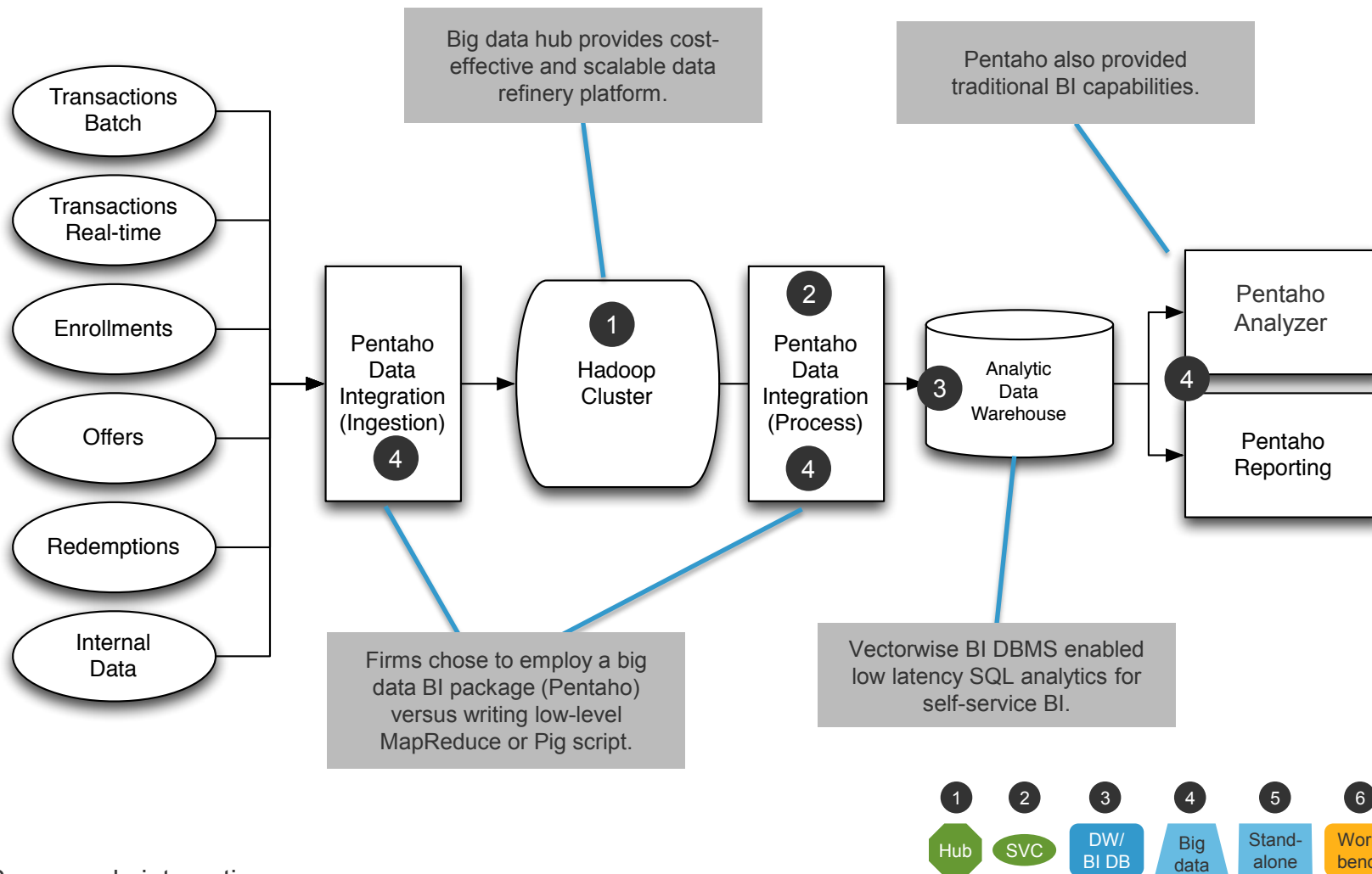
Data may be archived after harmonization, but the BI DBMS remains the primary place for analytics.



# Example: edo interactive



# Pentaho — conceptual solution architecture

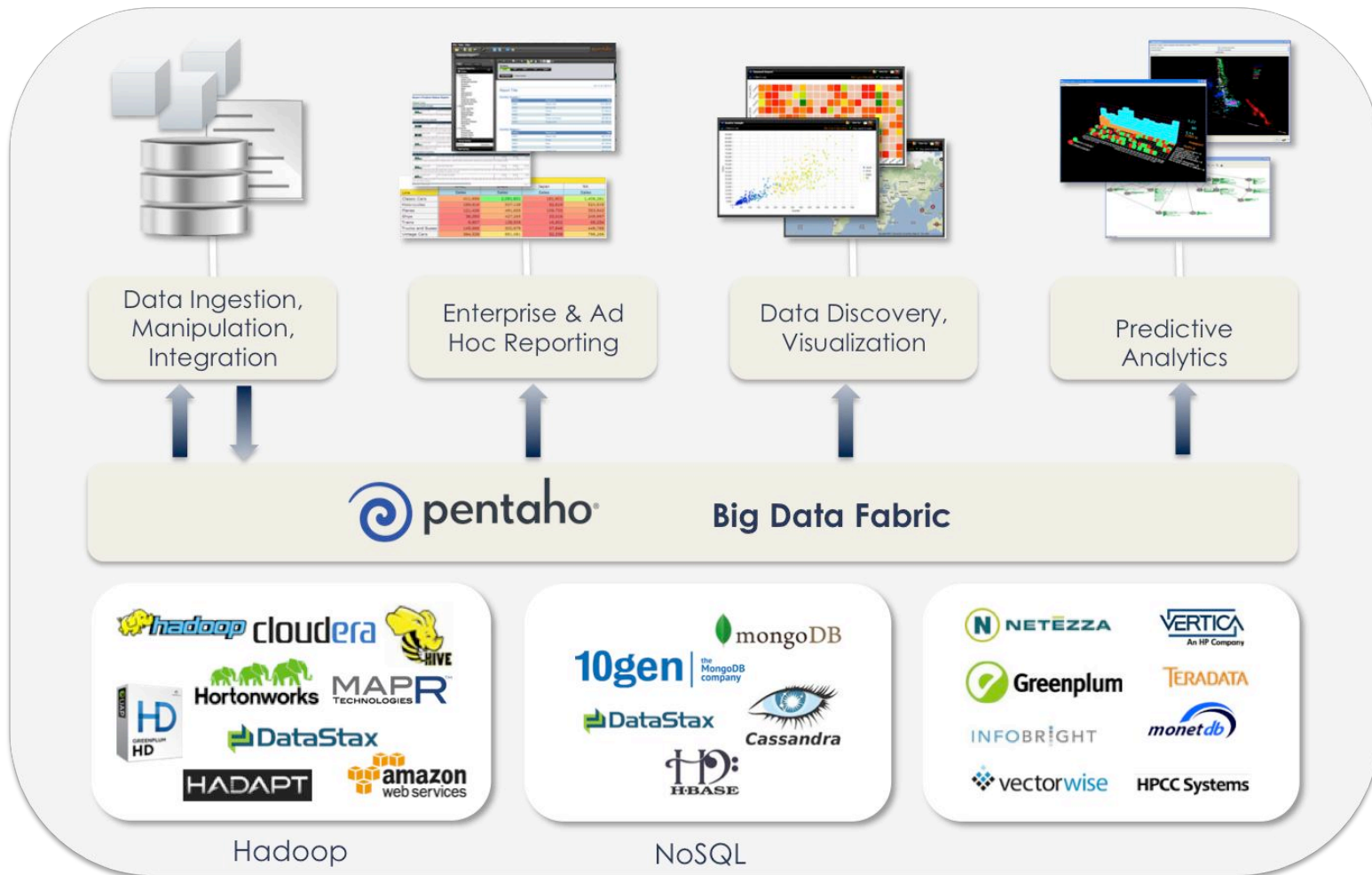


Source: edo interactive



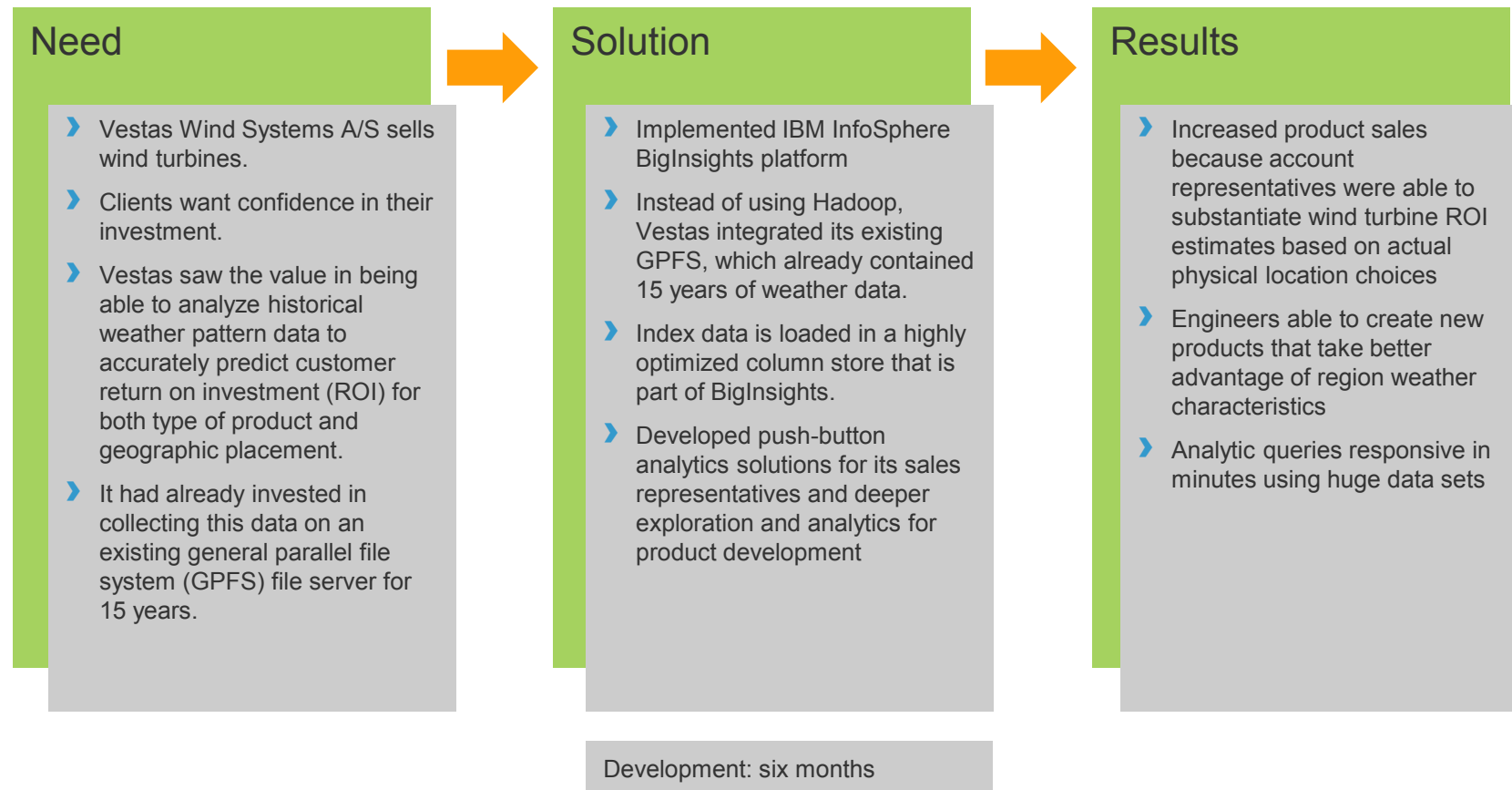
# Vendor information

## PENTAHO

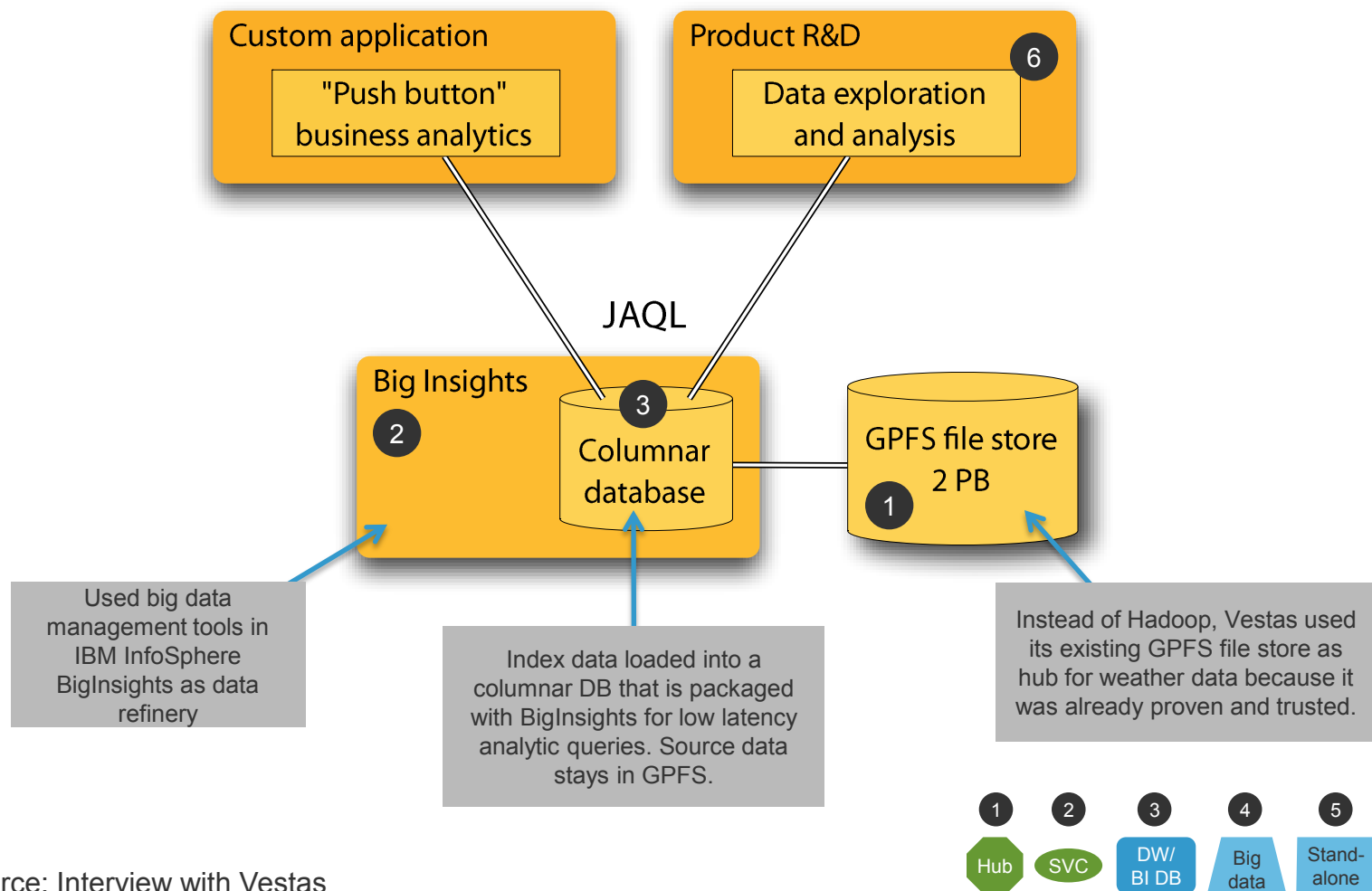


Source: Pentaho

# Example: Vestas Wind Systems (manufacturing)



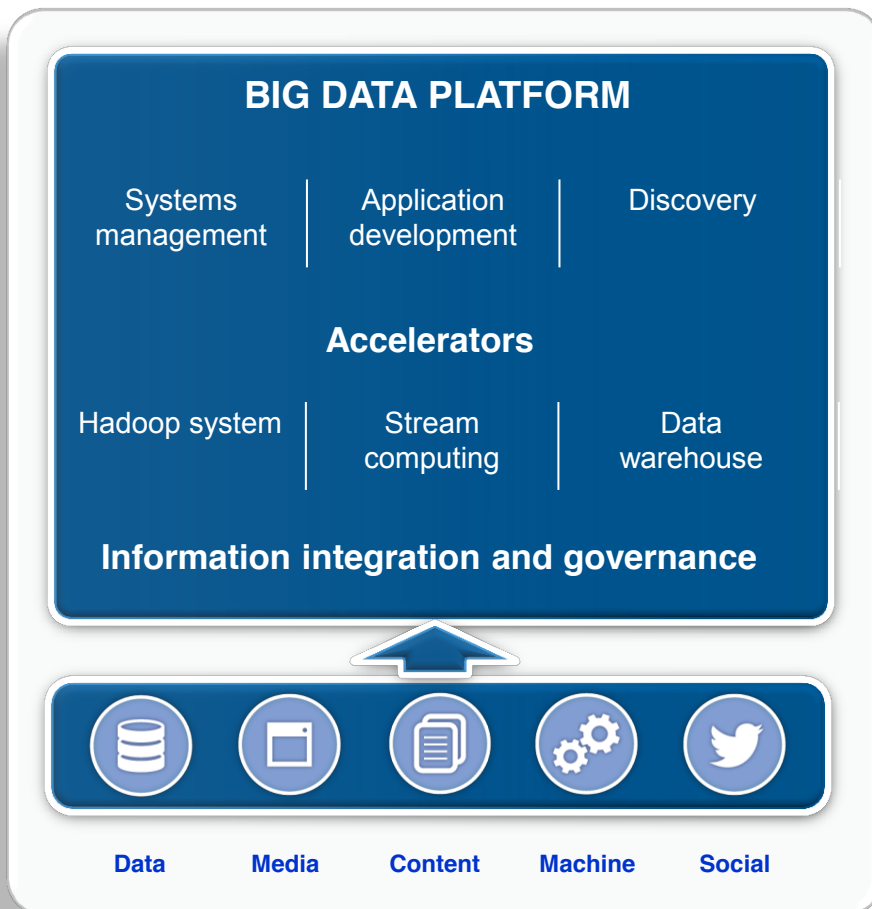
# Vestas' conceptual solution architecture



Source: Interview with Vestas

# Vendor information

## IBM BIG DATA PLATFORM

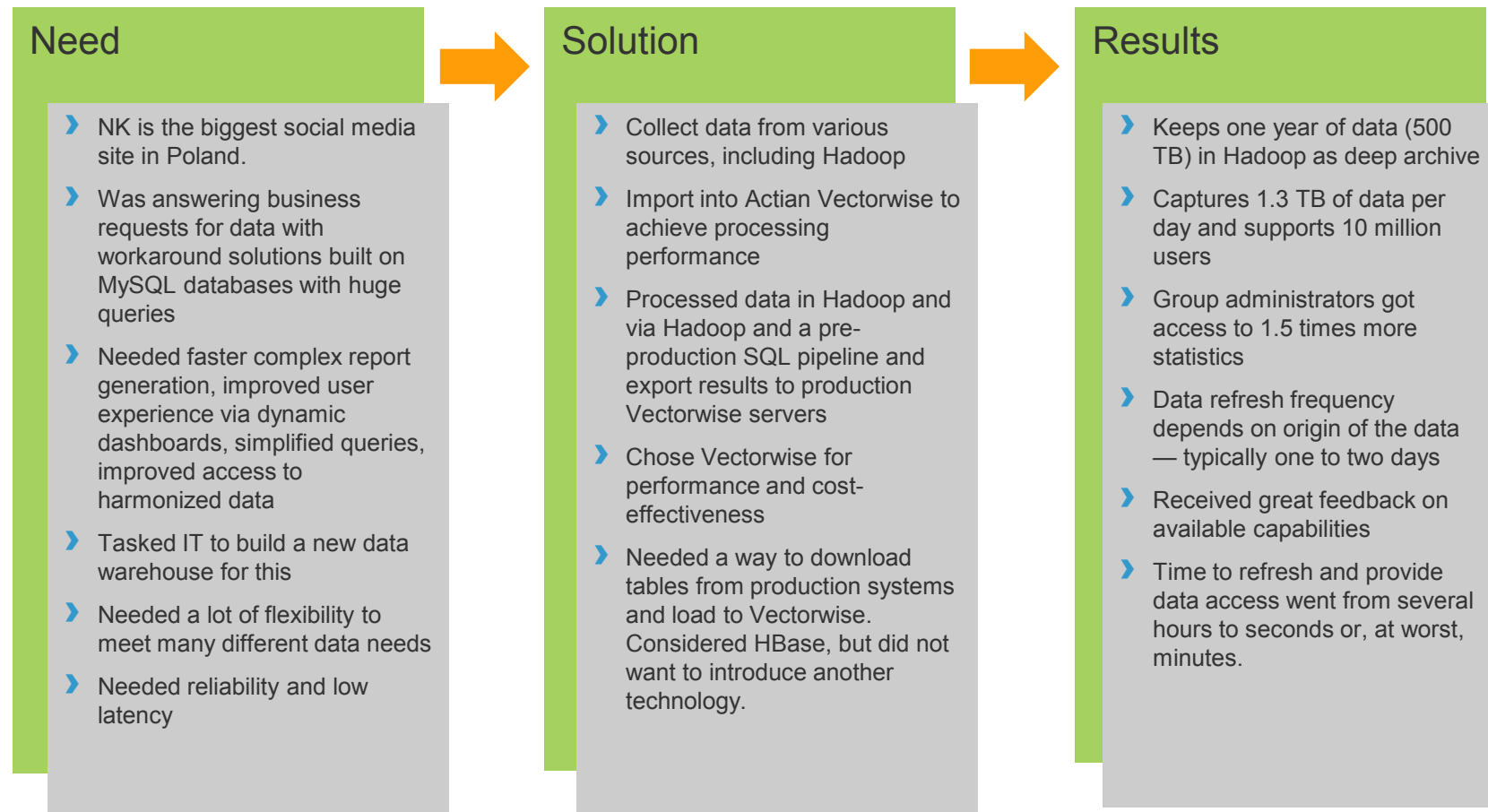


## The whole is greater than the sum of the parts.

- Almost all big data use cases require an integrated set of big data technologies to address the business pain completely
- Reduce time and cost and provide quick ROI by leveraging pre-integrated components
- Provide both out-of-the-box and standards-based services
- Start small with a single project and progress to others over your big data journey

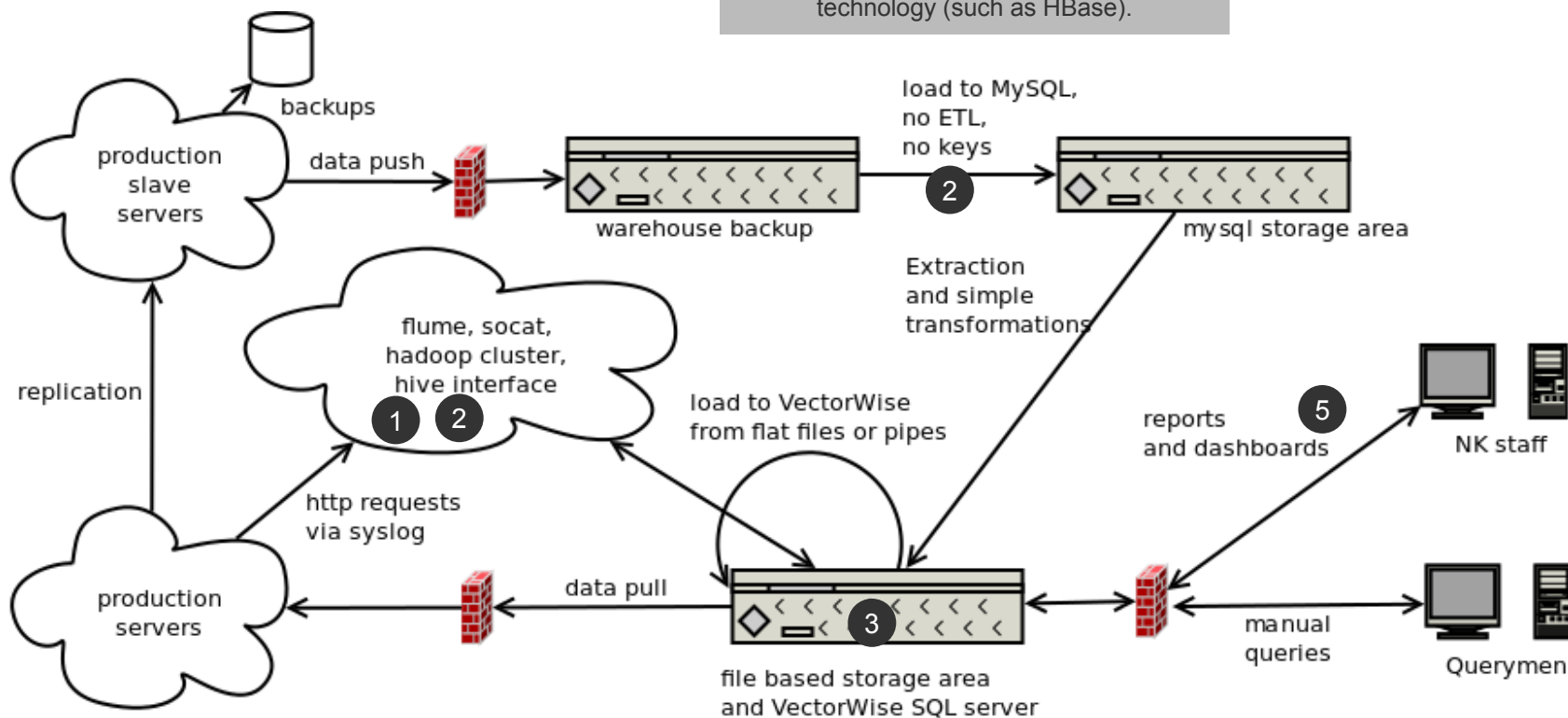
Source: IBM

## Example: NK (social media)



# NK's solution architecture concept

NK created a high-performance straight table loading processing pipeline because it didn't want to introduce another technology (such as HBase).

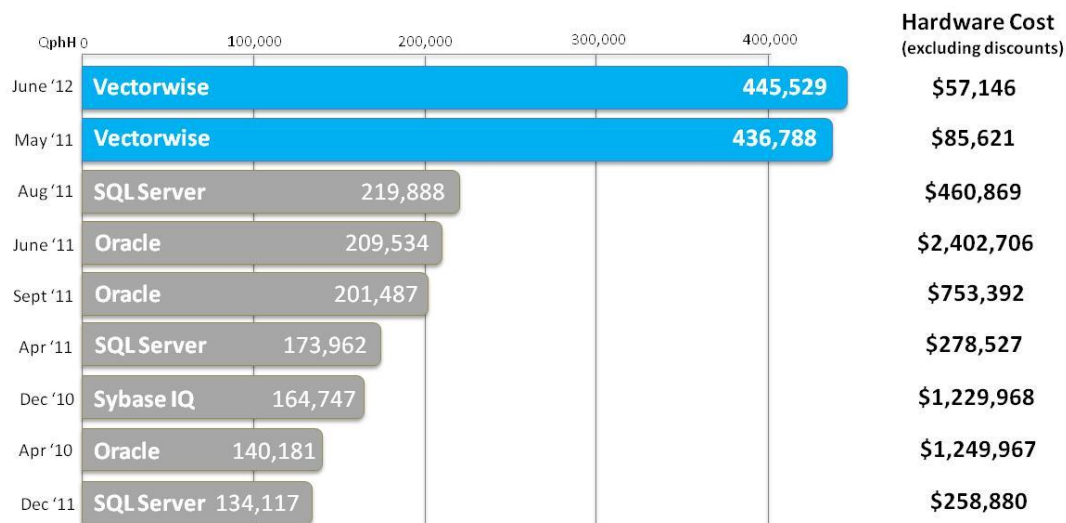


Source: NK

# Vendor information

## ACTION VECTORWISE

- › ANSI-compliant relational database for reporting and data analysis
  - Transparently exploits performance potential in x86-based CPUs
  - Processes large volumes of data incredibly fast
- › Requires no special tuning to achieve fast performance
- › Enables affordable BI for big data analytics and data processing

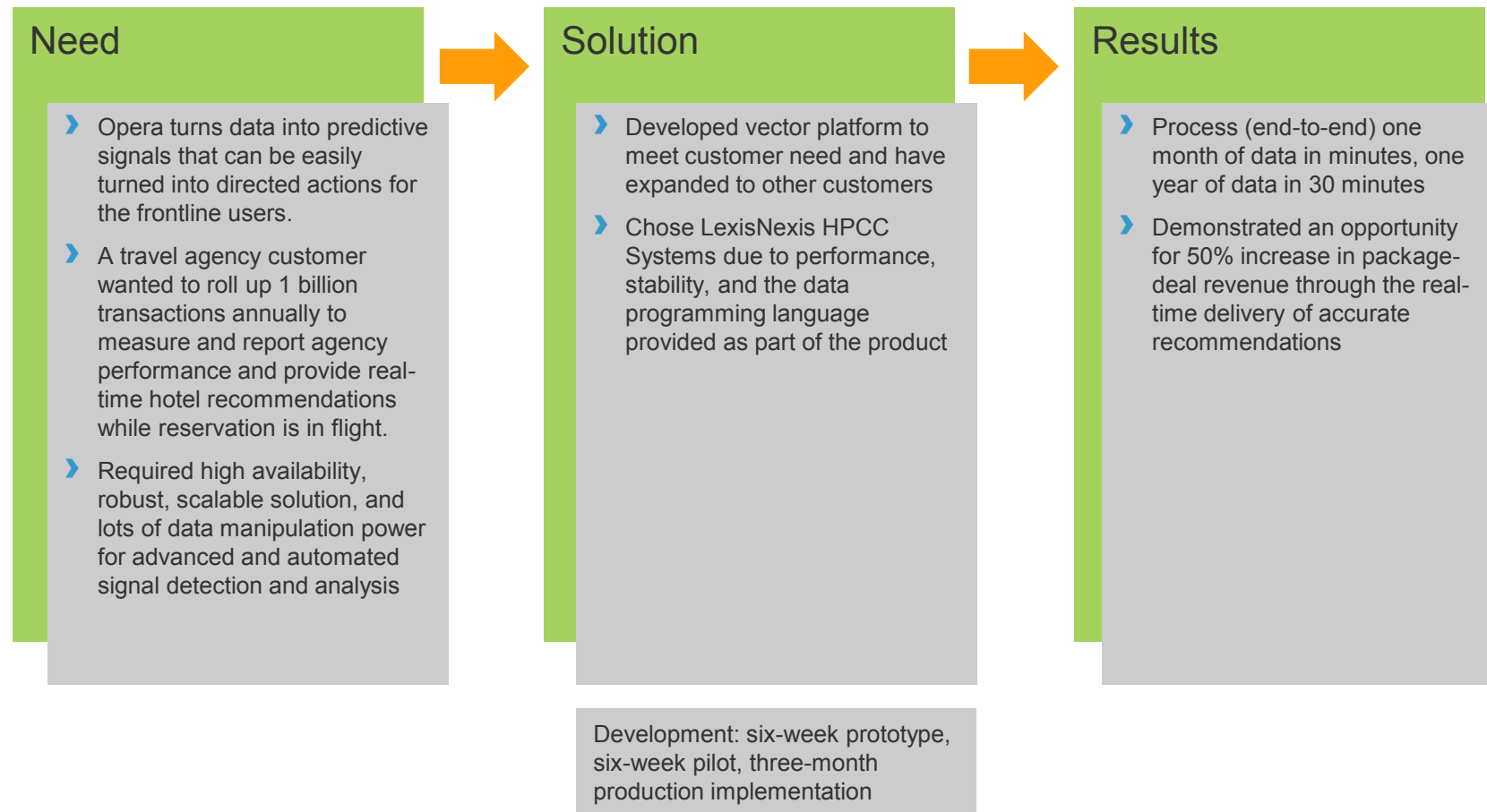


### Fastest TPC-H QphH@1TB Benchmark (non-clustered)

Source: [www.tpc.org](http://www.tpc.org) / December 21, 2012

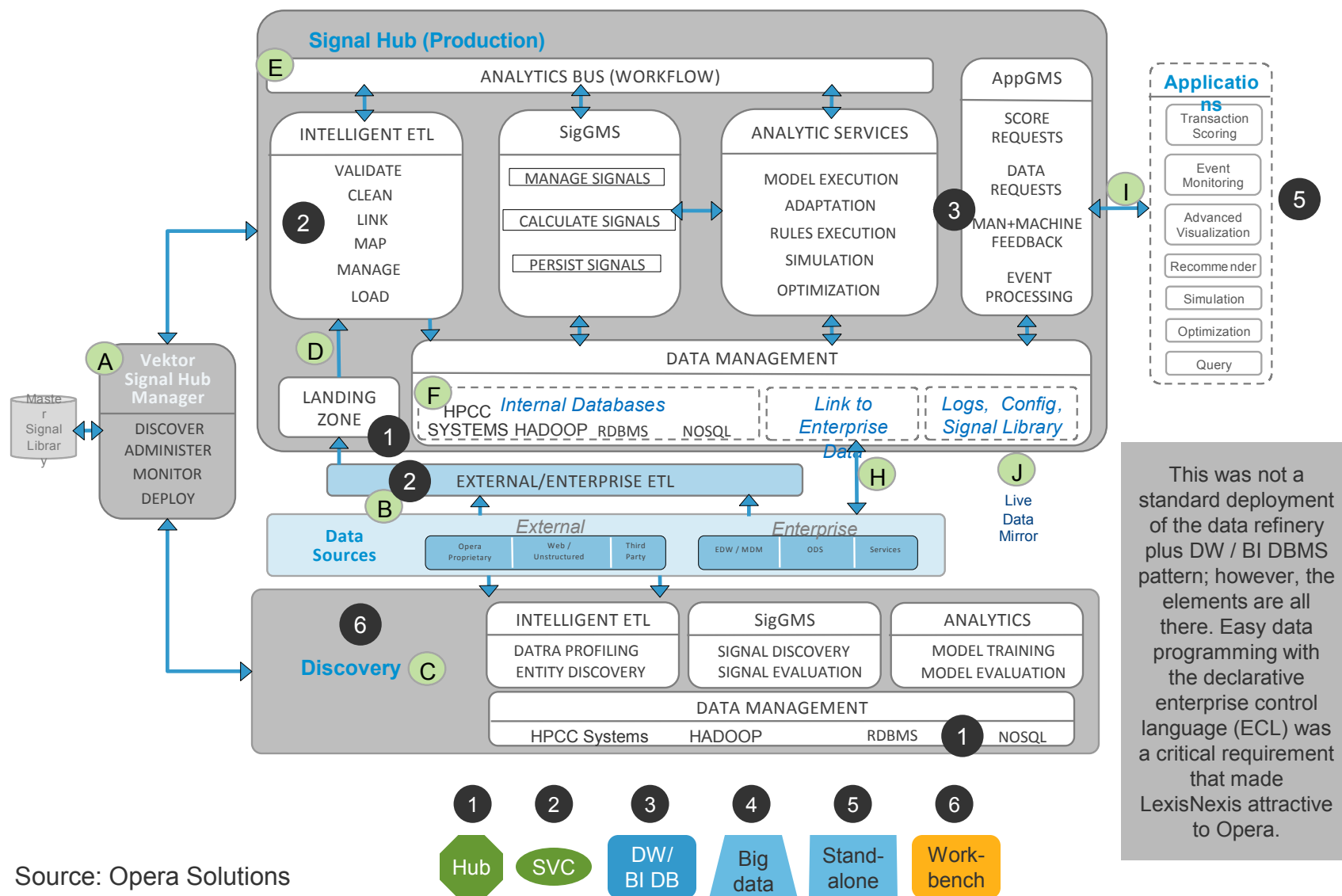
Source: Actian

# Example: Opera Solutions (IT)





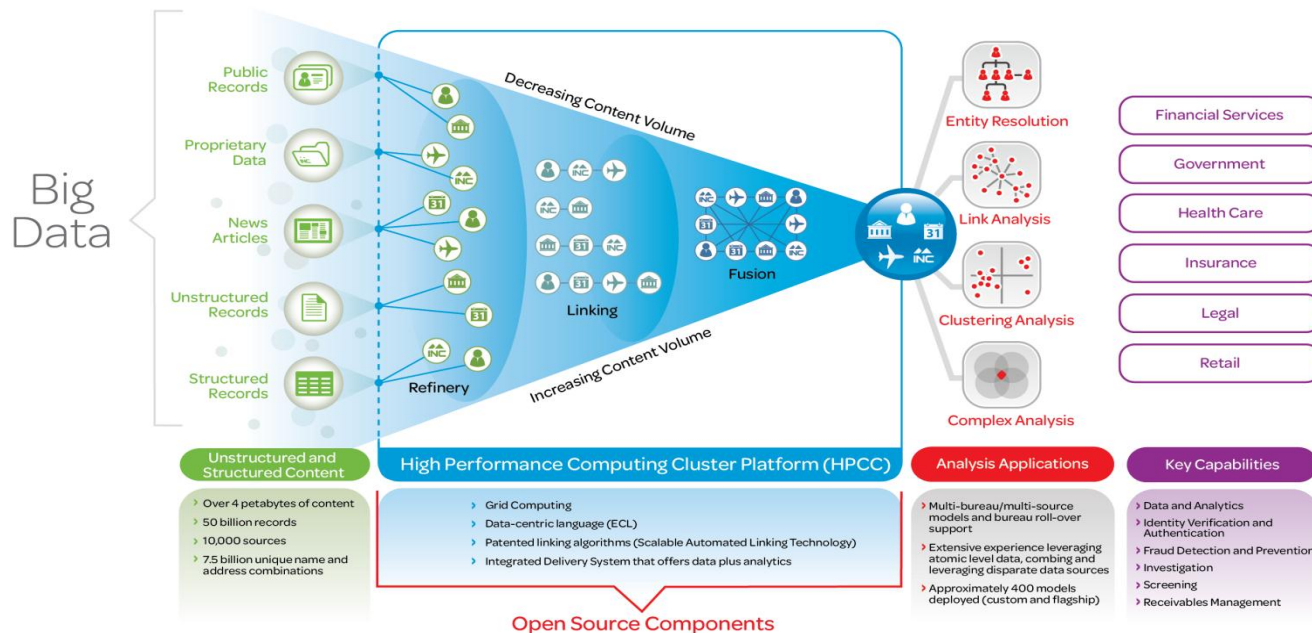
# Opera Solutions' solution architecture concept



Source: Opera Solutions

# Vendor information

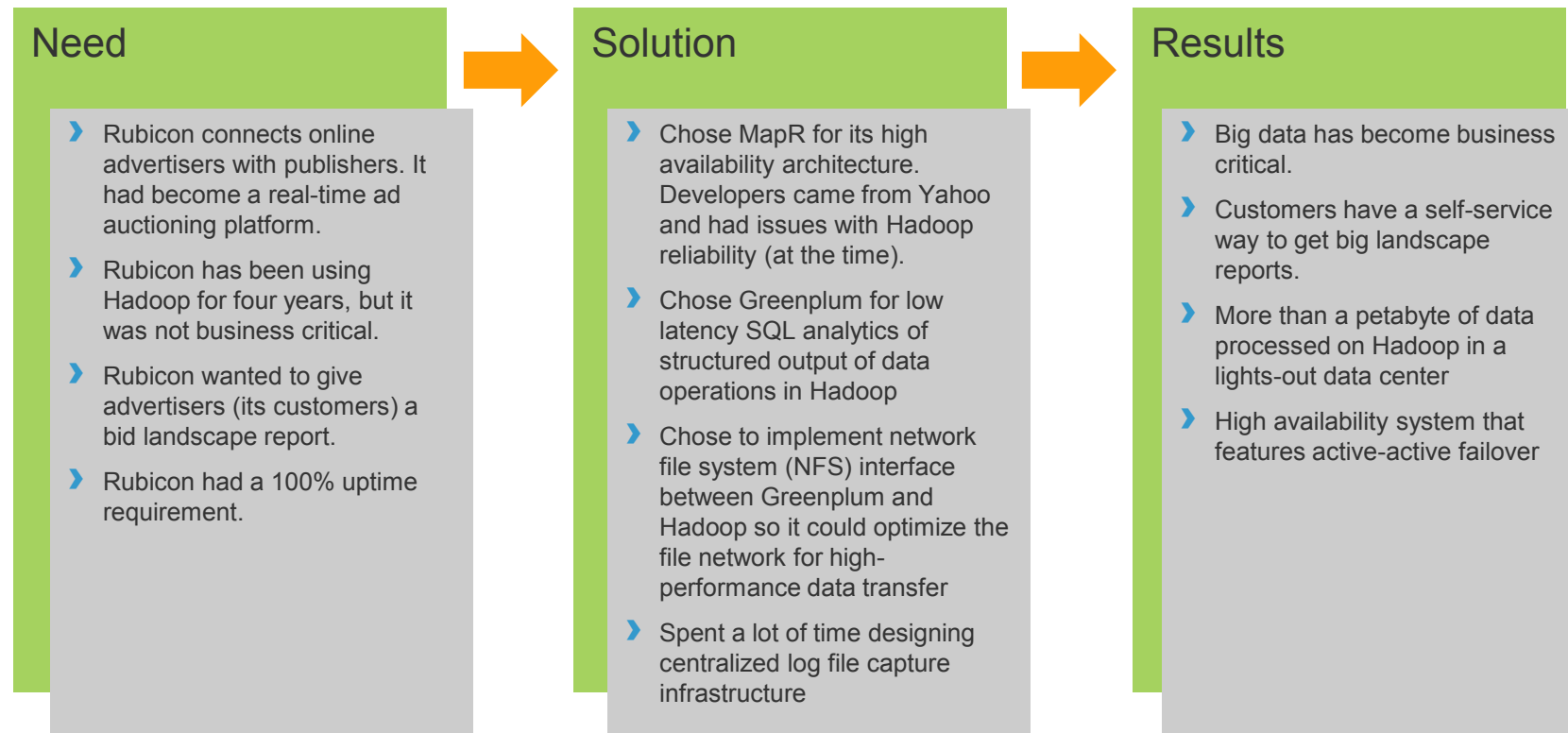
## LEXISNEXIS (HPCC SYSTEMS)



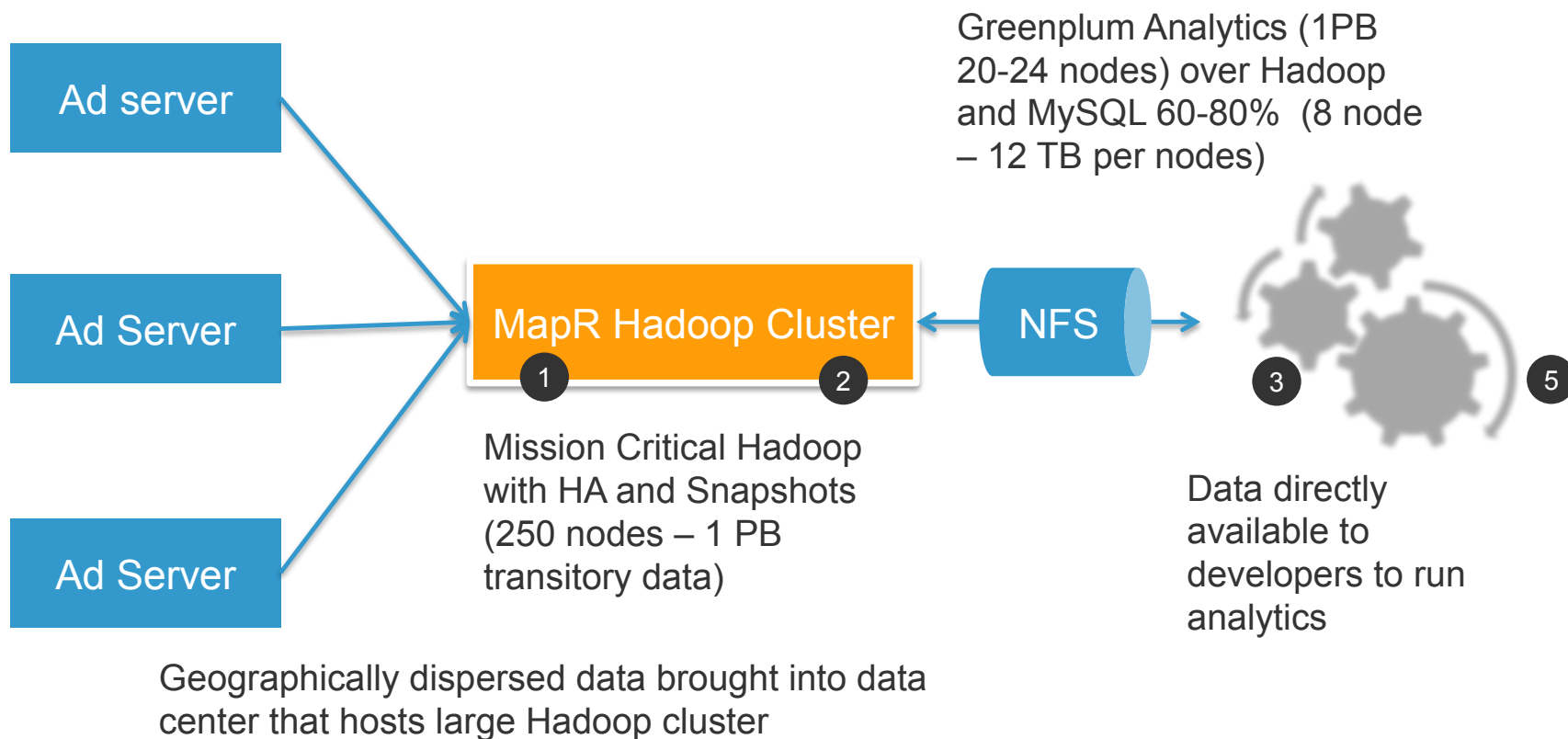
- **High-performance computing cluster (HPCC)** platform enables data integration on a scale not previously available and real-time answers to millions of users. Built for big data and proven for 10 years with enterprise customers.
- **Offers a single architecture**, two data platforms (query and refinery), and a consistent data-intensive programming language (ECL).
- **ECL parallel programming language** is optimized for business differentiating data-intensive applications.

Source: LexisNexis

## Example: Rubicon (digital marketing)



# Rubicon's solution architecture concept



Source: Rubicon



# Vendor information

## MAPR

- › Open, enterprise-grade distribution for Hadoop
  - Easy, dependable and fast
  - Open source with standards-based extensions
- › MapR is deployed at thousands of companies
  - From small Internet startups to the world's largest enterprises
- › MapR customers analyze massive amounts of data:
  - Hundreds of billions of events daily
  - 90% of the world's Internet population monthly
  - \$1 trillion in retail purchases annually



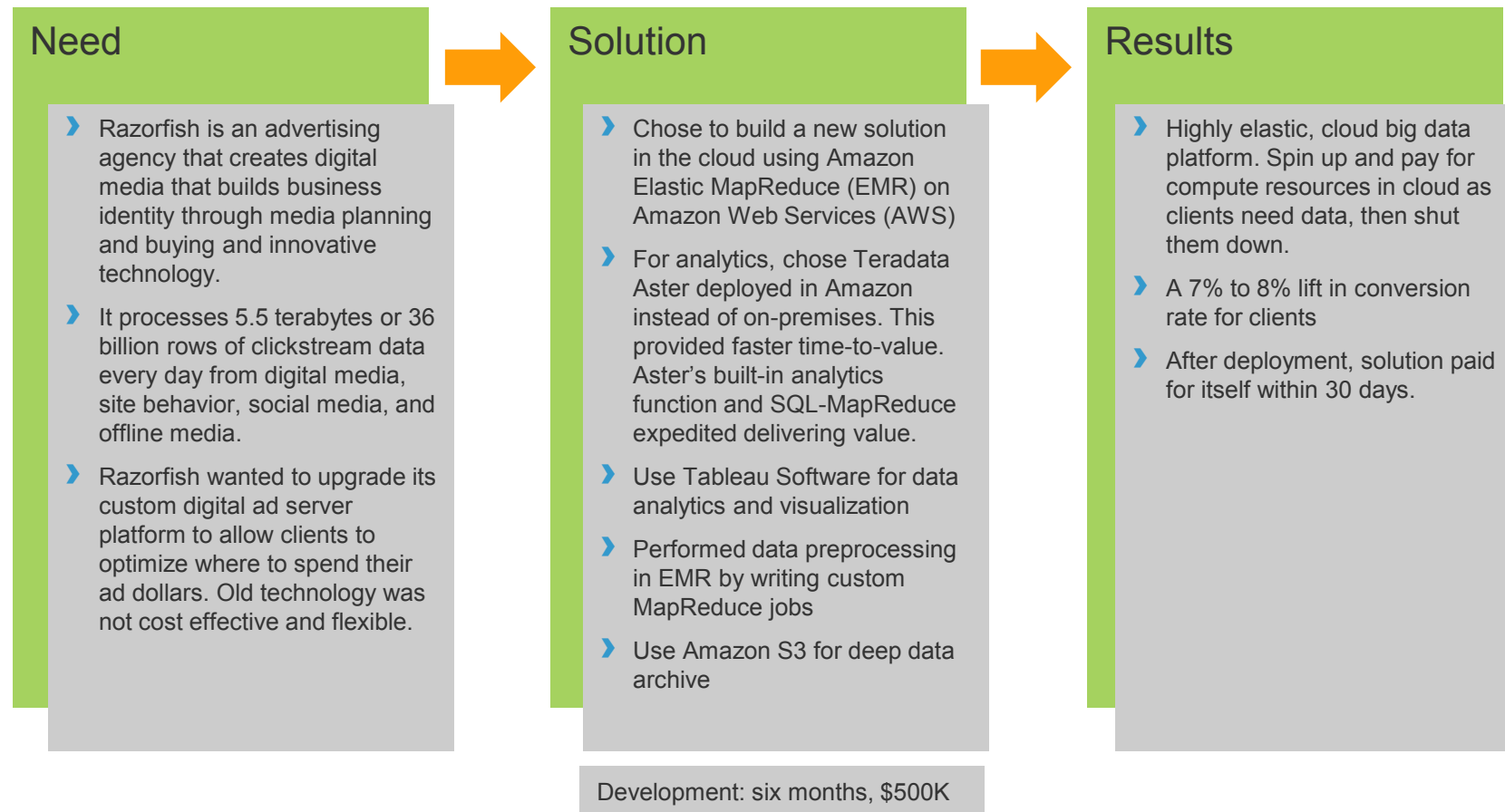
Also available through:



Compute Engine

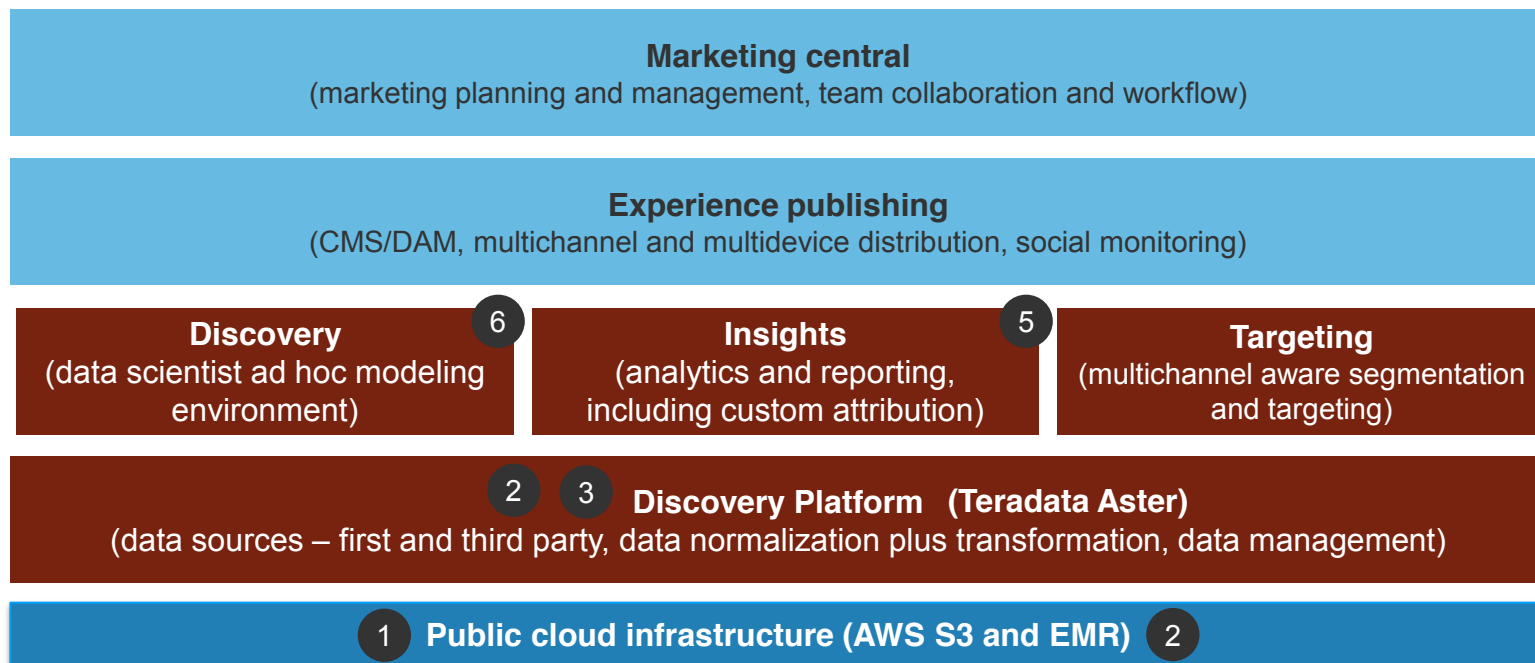
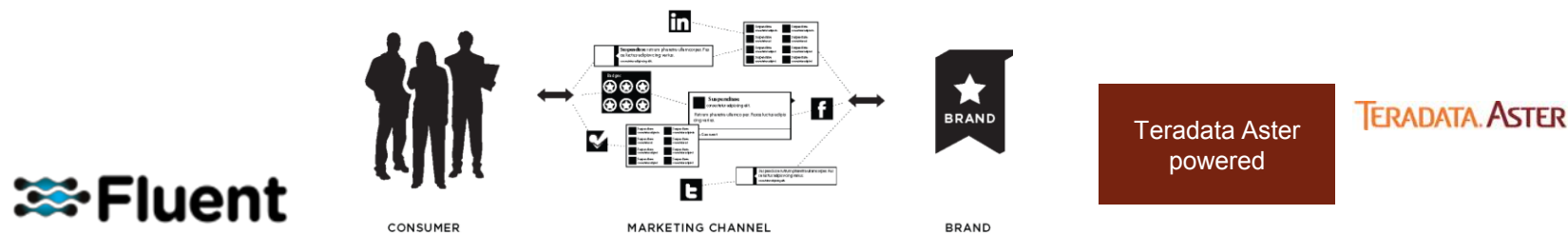
Source: MapR

# Example: Razorfish (digital marketing)



# Razorfish Fluent architecture concept

**Fluent** — a digital marketing technology platform that provides marketers and agencies with a single, integrated software application to target, distribute, and manage multichannel digital campaigns and experiences.



Source: Razorfish



# Vendor information

## TERADATA ASTER

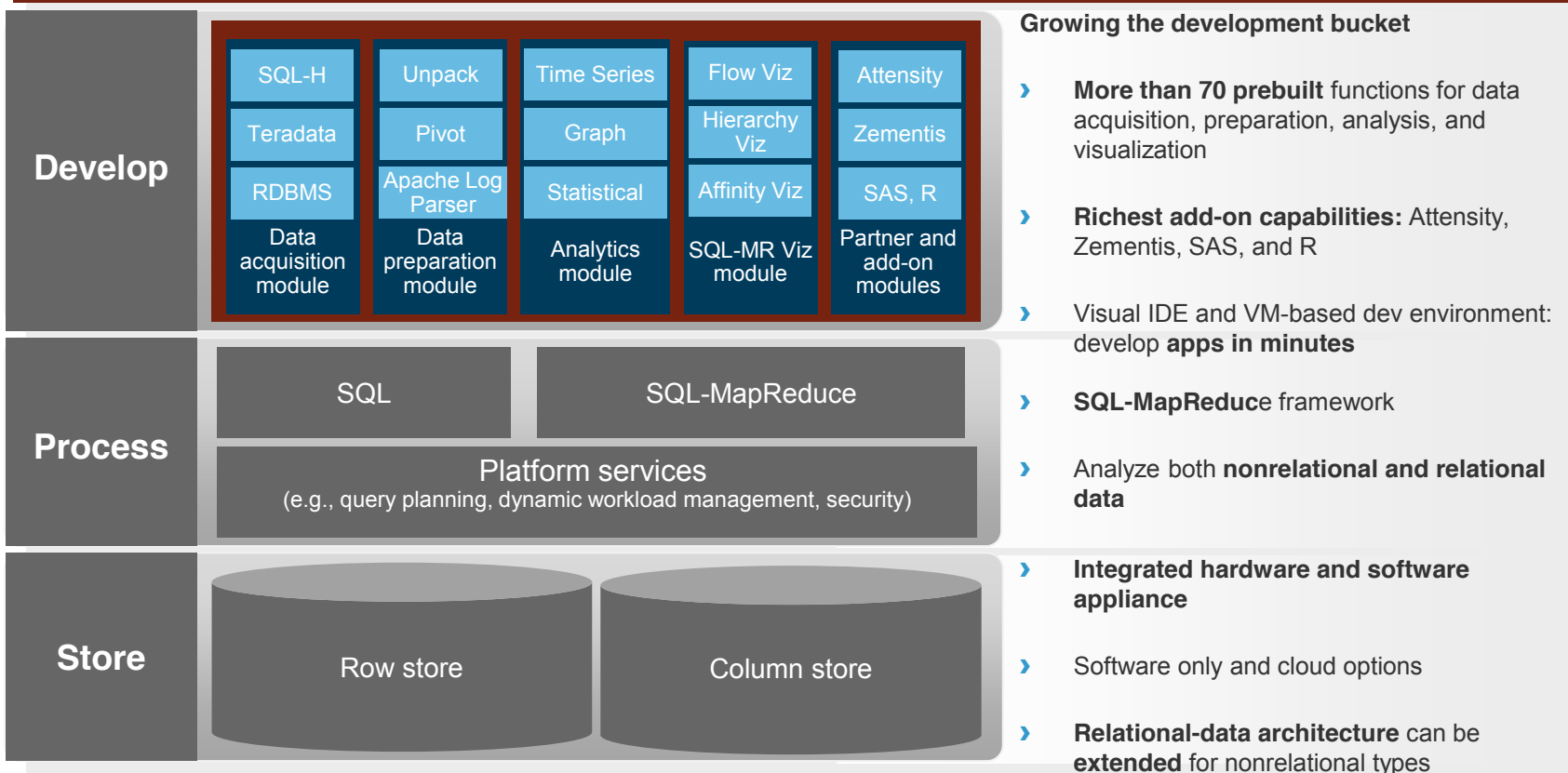
  
Analysts

  
Customers

  
Business

  
Data scientists

### Interactive and visual big data analytic apps



Source: Teradata Aster



## Forrester's point of view

- › Use this pattern when you have well-defined business requirements for high availability, low latency structured data analytics for a specific tactical need, but you also want to source more data for future needs and data science.
  - Future open source advances may challenge BI database vendors on availability and performance versus cost, but not just yet.
- › The benefits are characterized by drastic reductions in the time-to-value for making data availability for analytics in business critical systems. Secondary benefits include a deep archive for data science.
- › We suggest:
  - Start by identifying very tangible potential business benefits that fit this pattern.
  - Plan for time spent in infrastructure engineering and integration challenges.
  - Compare the types of embedded and prepackaged analytics available in your BI database against the needs for value from the data.
- › Your biggest strategic technology decisions are:
  - Which BI database tool to use. One that you already have may be best, since users will already know it.
  - Whether to build a new distributed hub on Hadoop or use a data platform you already have.
  - The selection of infrastructure that can provide needed performance with manageable tuning.
  - Where data integration and harmonization will happen and how much is needed, since BI big data packages, open source tools, and BI databases all provide functionality.

# Lessons learned from users

- ▶ **Open source puts a lot of pressure on packaged vendors.** However, these users felt that BI database vendors like Actian, Greenplum, and Teradata Aster still provided the best cost for performance when low latency SQL analytics was required.
- ▶ **Some workloads can be more cost effective in Hadoop.** Some firms find that a few data crunching workloads in their warehouses consume an extremely high percentage of the total resources. To avoid procuring additional hardware or appliances, firms can move those workloads to Hadoop.
- ▶ **BI/big data package tools accelerate data refinery operation development.** But limitations drove some firms to learn MapReduce. Take time to see if your basic requirements can be met with a BI/big data package, but understand that at some point you'll need MapReduce skills to get the most out of Hadoop.
- ▶ **Data flow design is crucial.** Spend time designing your data flows and think carefully before choosing to design your own analytic jobs. Many BI database packages have good stuff built in that you can leverage.
- ▶ **Public cloud still requires some big tradeoffs.** Public infrastructure-as-a-service (IaaS) VM gives firms limited options for environmental tuning, but that is changing. For now, on-premises is the best way to get cost-effective, high-performance big data. Designing analytics solutions in the public cloud is a big change; developers must get used to it and design for efficient and automatic resources utilization.
- ▶ **Big data can expose network and storage bottlenecks.** MapReduce job design can have big network impacts; you must design for this. High performance can require very low-level storage engineering in Hadoop.

## Lessons learned from users (cont.)

- **Help is essential unless you are an information technology firm.** Since the pattern requires integration between a BI database and a distributed data hub, technology choices can be complex. Unless you are in the business of developing IT, get help.
- **Do not get overly enamored with Hadoop.** It is not your only hub solution. In some cases an existing system, so long as it's cost effective and scales well, may be your best data persistence option. Consider new hardware costs as well as software.

# Pattern: all-in-one

**Primary purpose:** simplify and consolidate data management onto a single platform that provides for cost-effective analytics at scale

Examples:

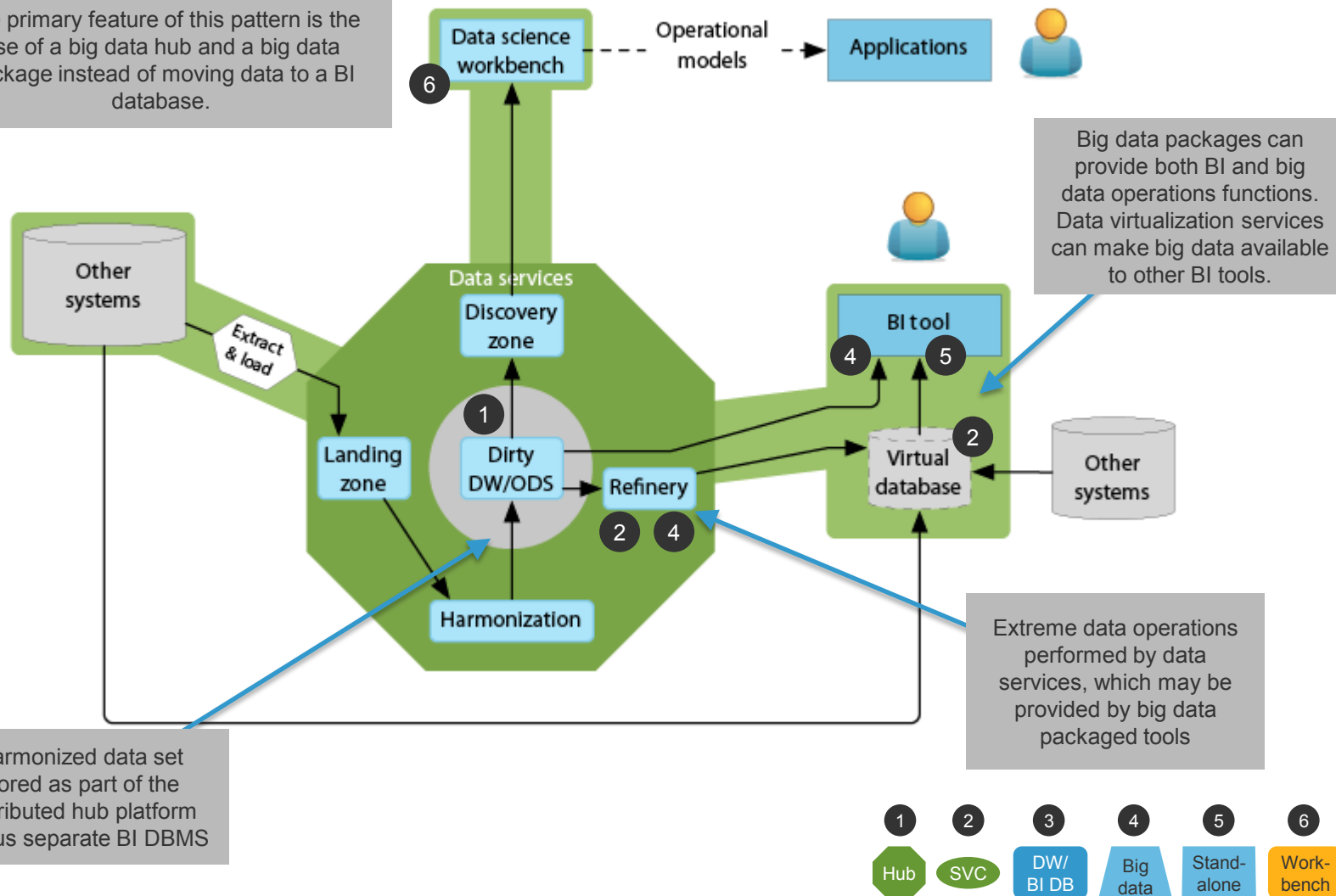
Sears (retail).....54-57

Telecommunications company.....58-59

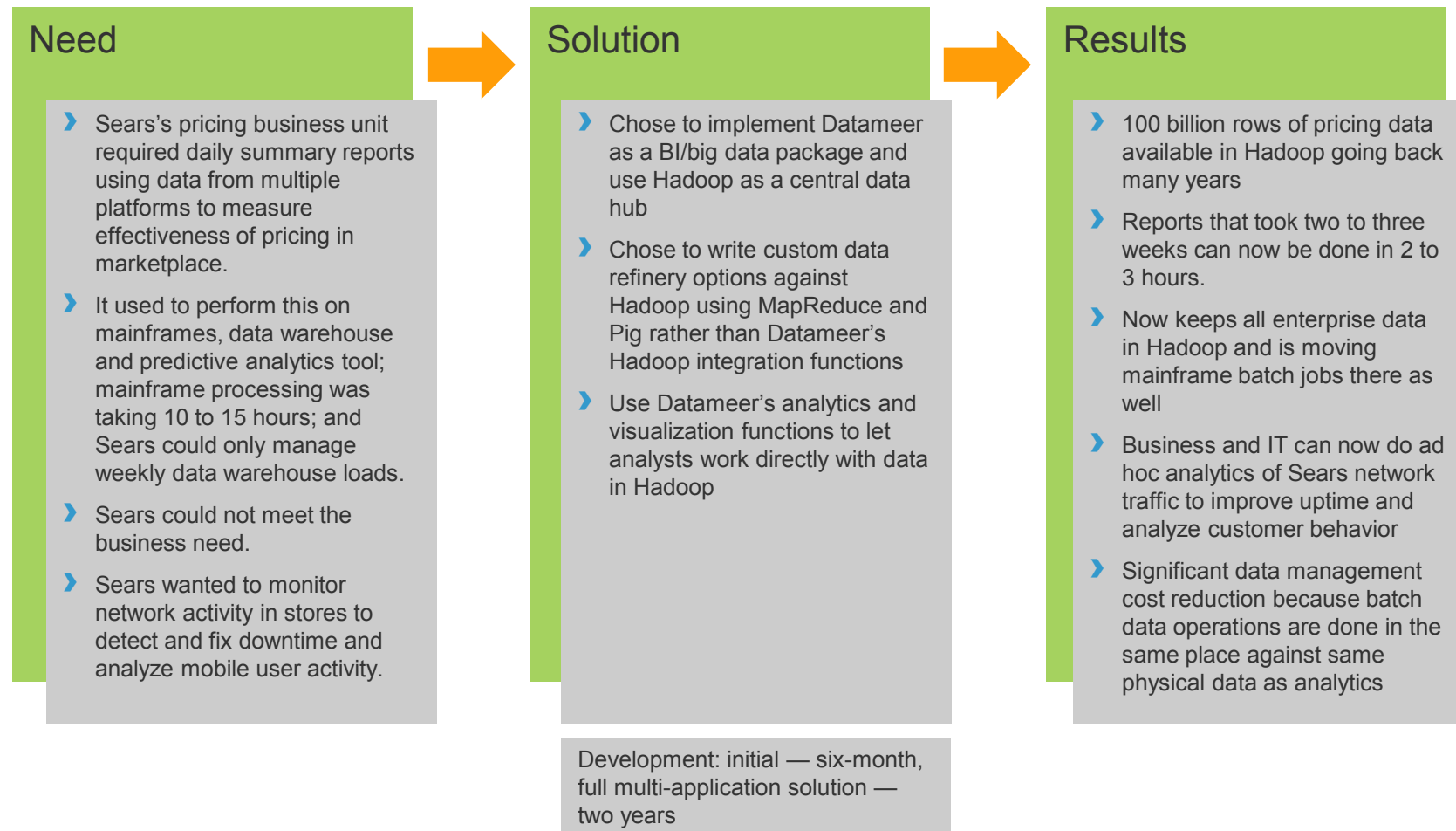
# All-in-one pattern

The primary feature of this pattern is the use of a big data hub and a big data package instead of moving data to a BI database.

Data science teams perform exploration and model, deploying models to applications.

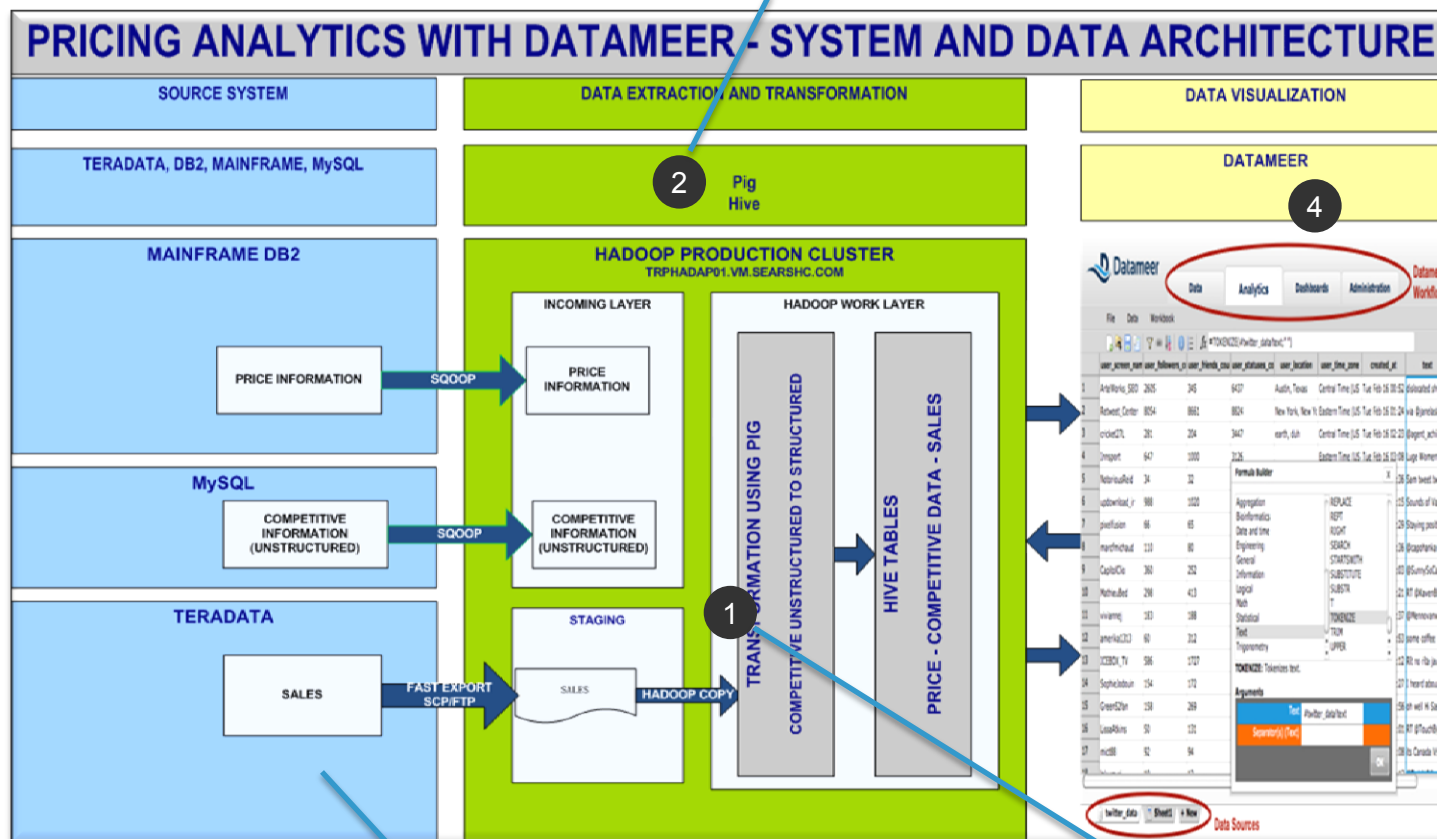


# Example: Sears (retail)



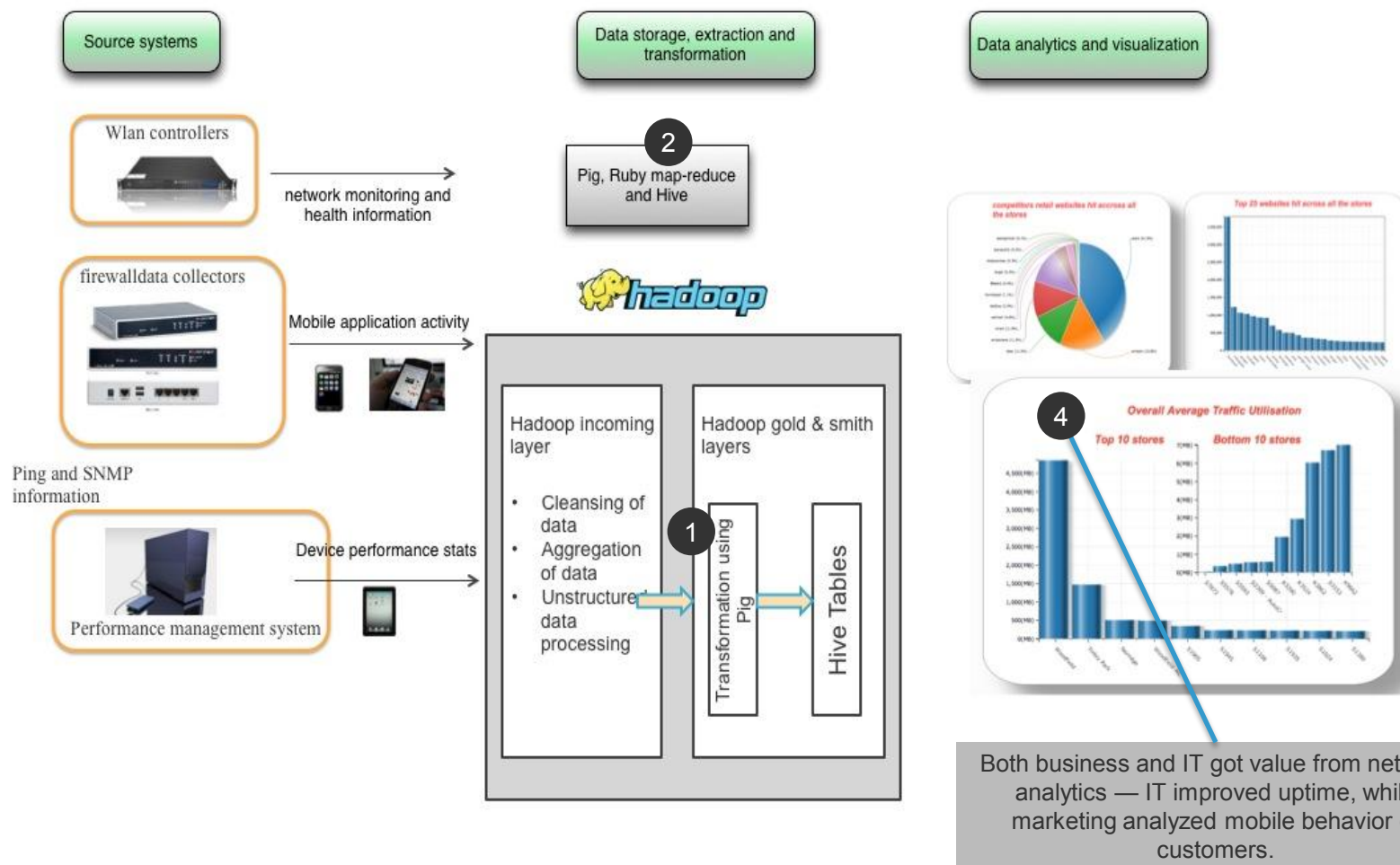
# Sears' solution architecture concept for pricing analytics

Sears chose to skill up for extreme extract, load, transform operations in Hadoop rather than use Datameer's tools



Source: Sears

# Sears' solution architecture concept for network analytics



Source: Sears



# Vendor information

## DATAMEER

### Seamless data integration

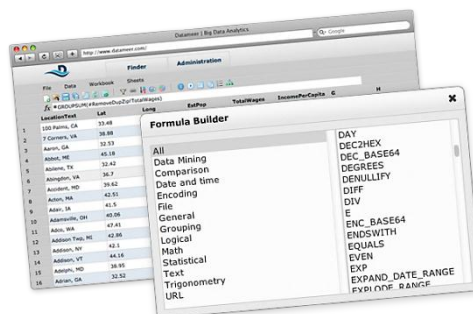
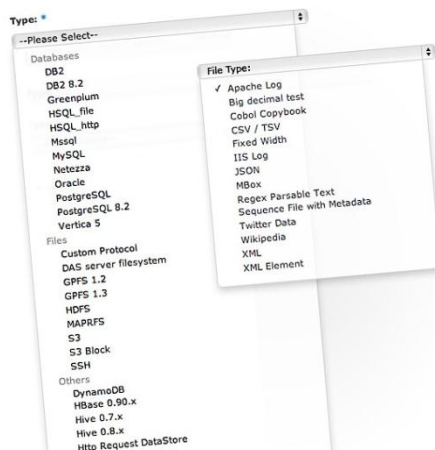
- › Structured, semistructured, and unstructured
- › 25-plus connectors
- › Connector plug-in API

### Powerful analytics

- › Interactive spreadsheet UI
- › More than 200 built-in analytic functions
- › Macros and function plug-in API

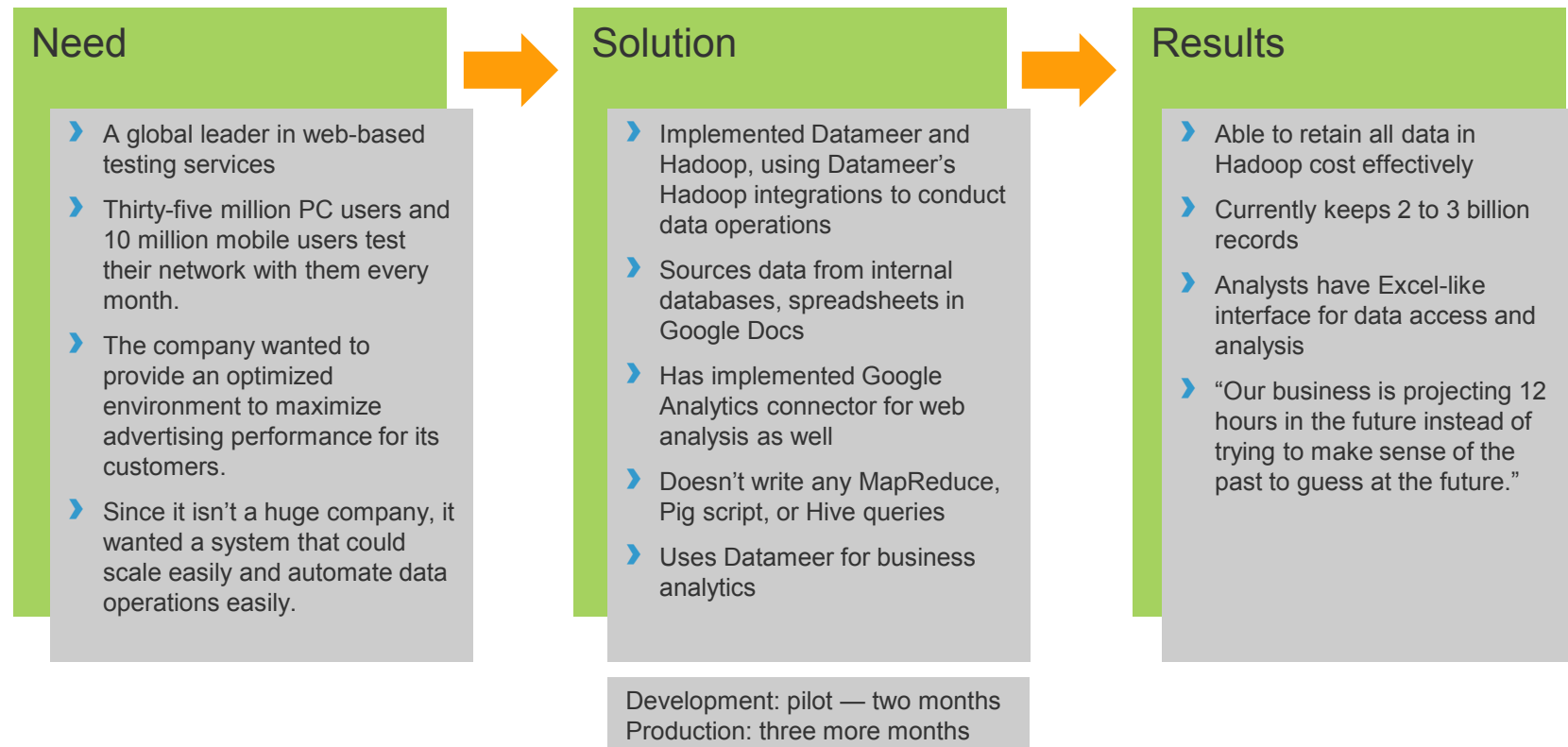
### Business infographics

- › Mashup anything, WYSIWYG
- › Infographics and dashboards
- › Visualization plug-in API

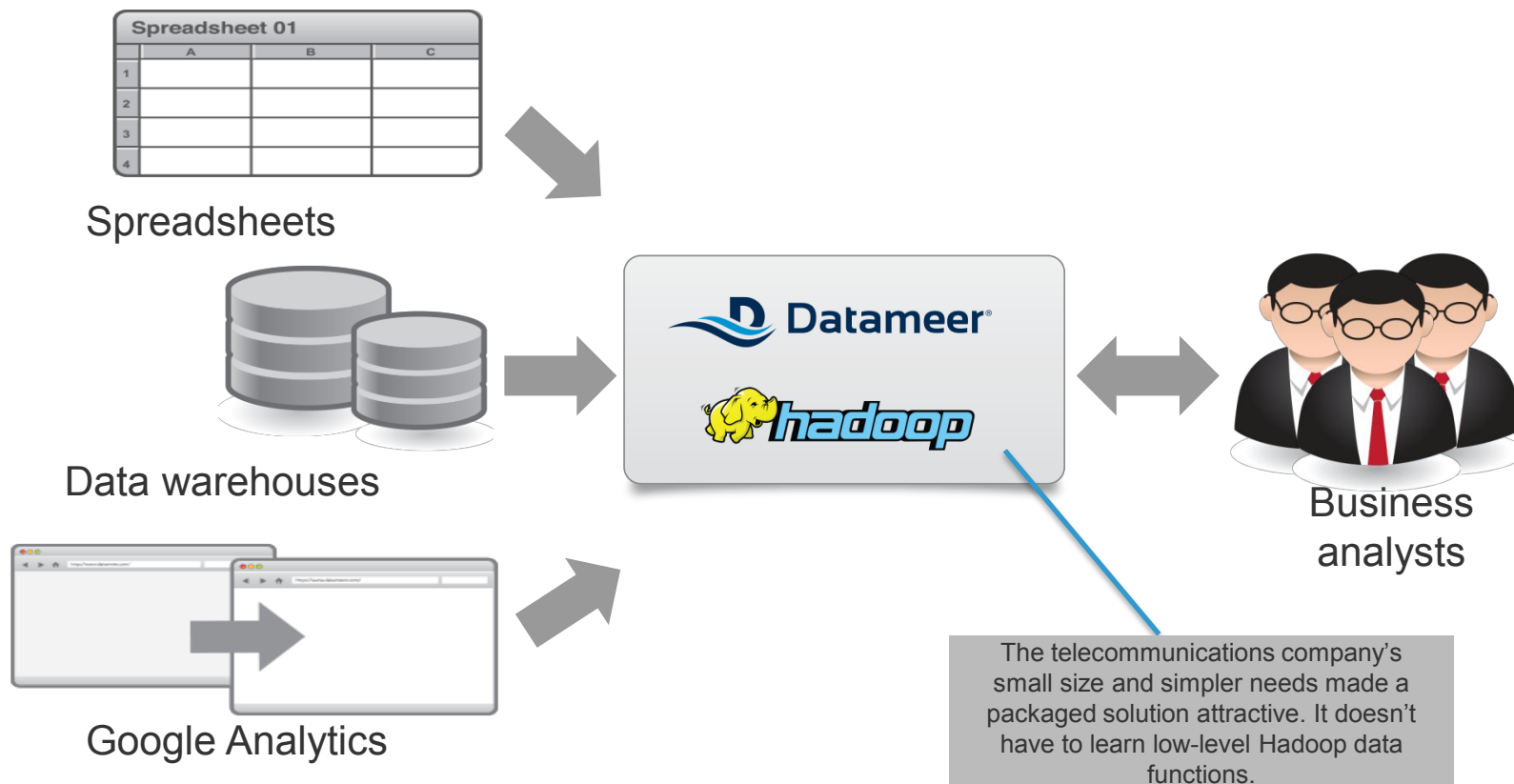


Source: Datameer

# Example: telecommunications company



# Telecommunications company's platform architecture concept



Source: telecommunications company

# Forrester's point of view

- › This pattern is advantageous when:
  - An order of magnitude increase in the volume of data for business analytics is needed.
  - Latency requirements are flexible.
  - Business criticality of solutions does not demand “platinum” service-level agreements (SLAs).
- › This pattern requires the most dramatic change in business use of data. New techniques and technologies are used to access more raw data right in Hadoop, rather than continued use of BI tools and existing data warehouse/BI database technology.
- › Data harmonization to a minimum quality and schema conformance are critical, requiring planning and governance maturity.
- › We suggest comparing this pattern with the data refinery plus DW / BI DBMS pattern to see which one yields a better cost for needed performance, both now and in the future.
- › Your biggest strategic technology decisions are:
  - Which distributed data management system meets enterprise performance requirements.
  - Infrastructure for running distributed data management technology, including the use of public cloud services.
  - The extent to which you leverage a big data BI package for data operations versus learning to “roll your own.”
  - Where you persist harmonized data in your data hub. For example, do you load tables in HBase or keep files in HDFS with Hive tables defined?
  - The extent to which data is made available using existing BI tools and data virtualization (see the financial services example on slide 24)

## Lessons learned from users

- › “This is a completely different way of doing things.” — CTO, Sears. Plan to expend some effort getting your firm’s head around this solution. You need a very influential advocate (CxO level) to spend time with business and IT leaders.
- › Adoption by end users can be challenging. This pattern may require that end users use new big-data-aware tools like Datameer for business analytics; it also requires business users to think inquisitively and understand that more data is available, but it tends to be dirtier. Spend time at the working level on the campaign trail, but remember that if the business doesn’t want to learn new tricks, little can change that.
- › Create a harmonized set of data that is worth persisting in the long-term. Don’t try to keep all the raw junk. Consider doing initial data harmonization in MapReduce before you try Pig for higher-level operations. You will likely want to keep the raw data for a period of time, but how long is based on 1) immediate value opportunities and 2) your confidence that future value can be found in the raw source data.
- › Don’t forget about performance. This pattern generally provides lower-end user performance than the data refinery plus DW / BI DBMS pattern, but it can yield a big cost savings. The key question is: Do the cost savings outweigh the tradeoffs, especially when low latency SQL analytics is the endgame? Infrastructure and Hadoop distribution choices will determine how easy or hard performance is to get.

# Pattern: hub-and-spoke

**Primary purpose:** create an enterprise-class data management architecture capable of a wide range of cost, performance, and quality options while consolidating data into an integrated, scalable platform

Examples:

Pharmaceutical company.....64-65

Analytics platform vendor (IT).....66-67

# Hub-and-spoke pattern

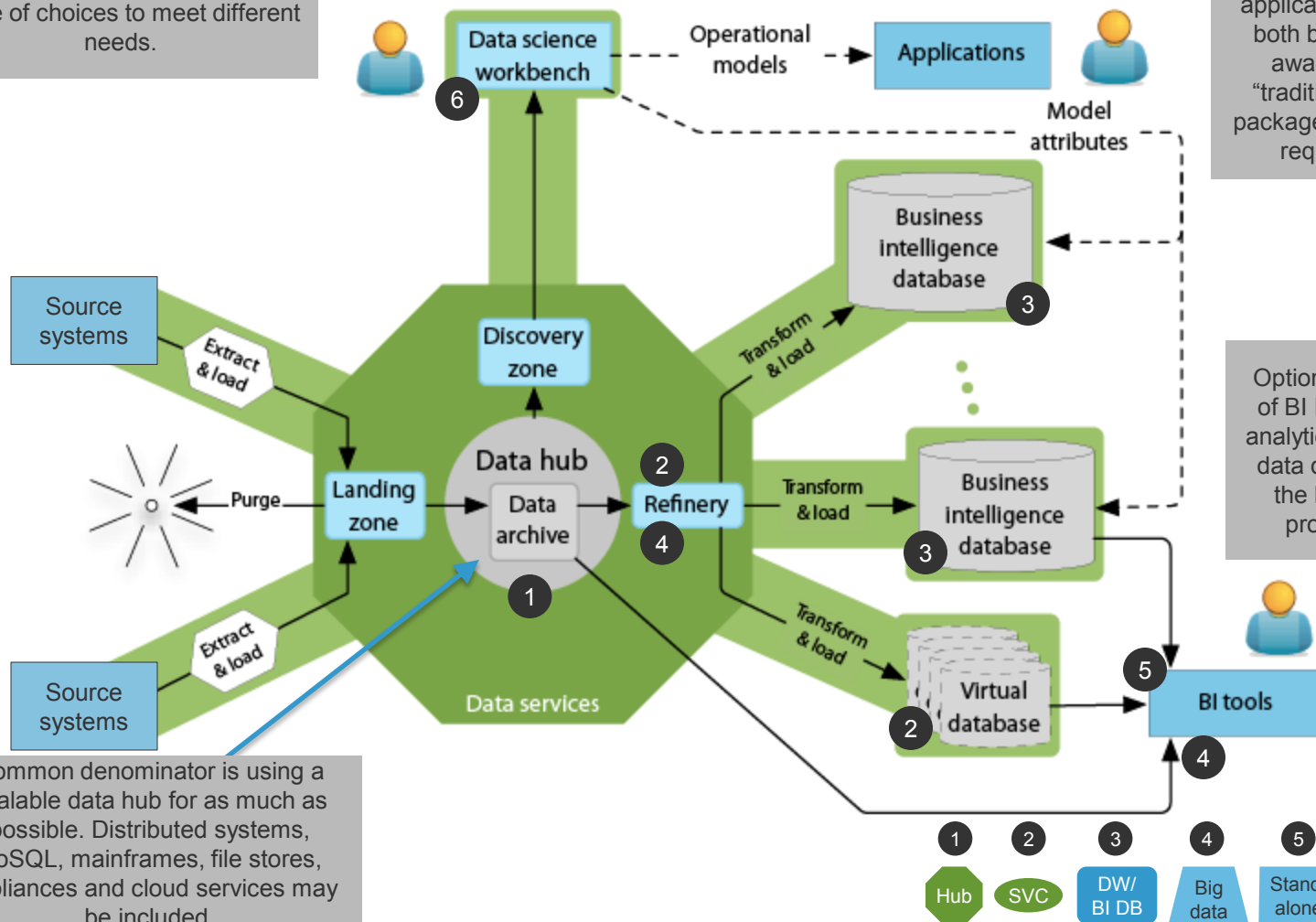
This pattern recognizes that enterprises often need a wide range of choices to meet different needs.

A robust toolset for data science and predictive modeling is provided.

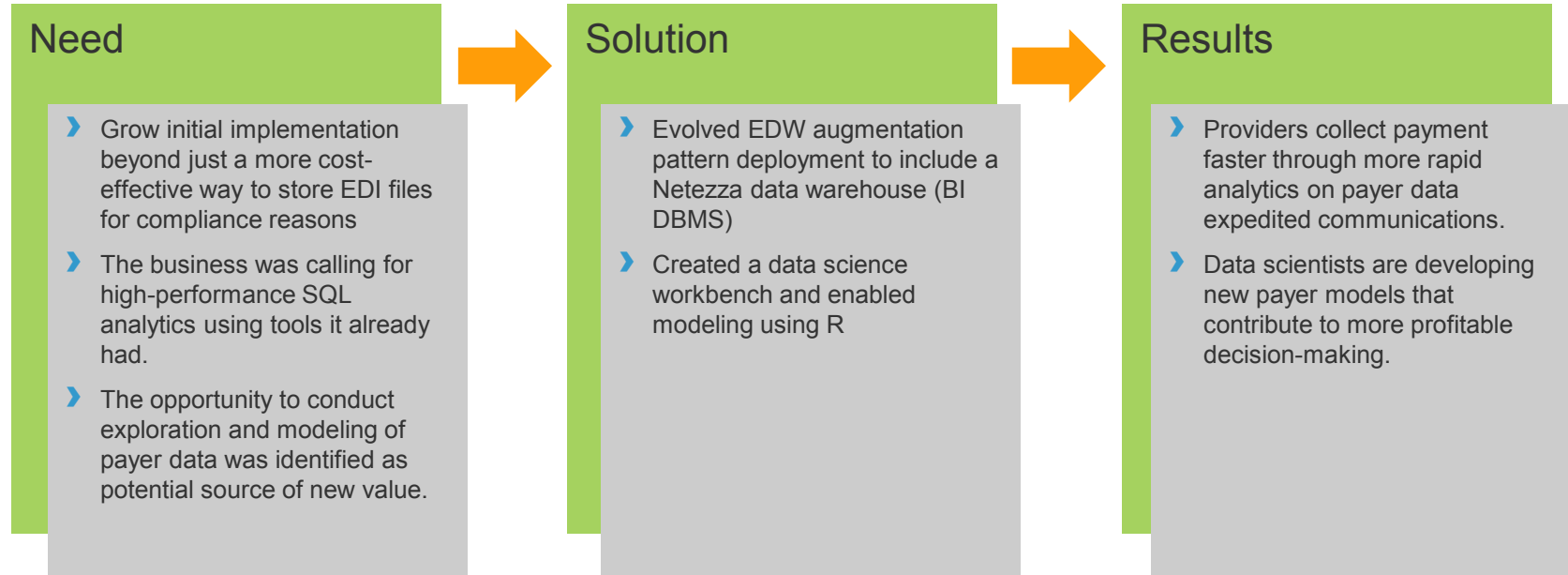
Multiple applications and both big-data-aware and “traditional” BI packages may be required.

Options for use of BI DBMS or analytics against data directly in the hub are provided.

Common denominator is using a scalable data hub for as much as possible. Distributed systems, NoSQL, mainframes, file stores, appliances and cloud services may be included.



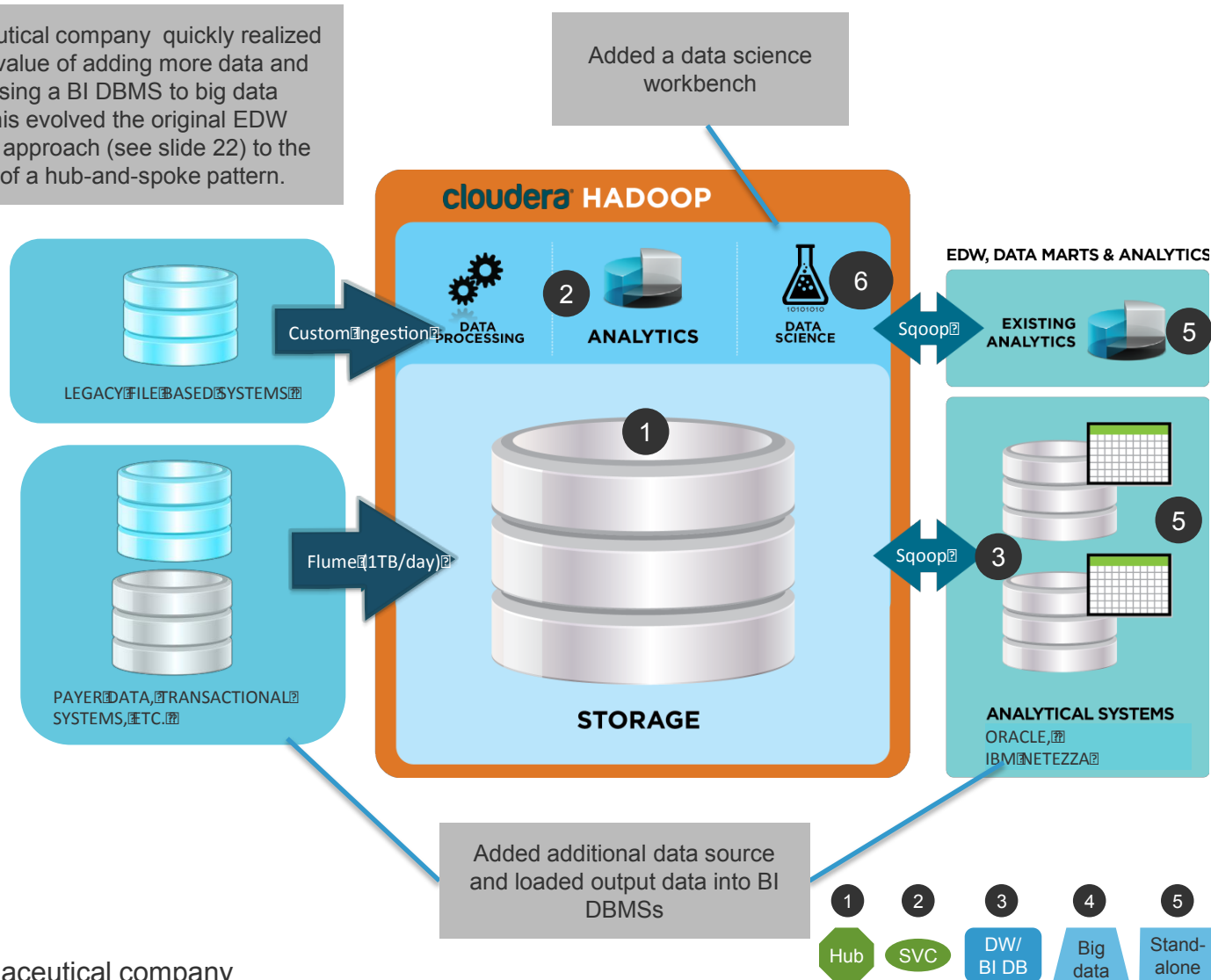
# Example: pharmaceutical company (revisited from slide 21)





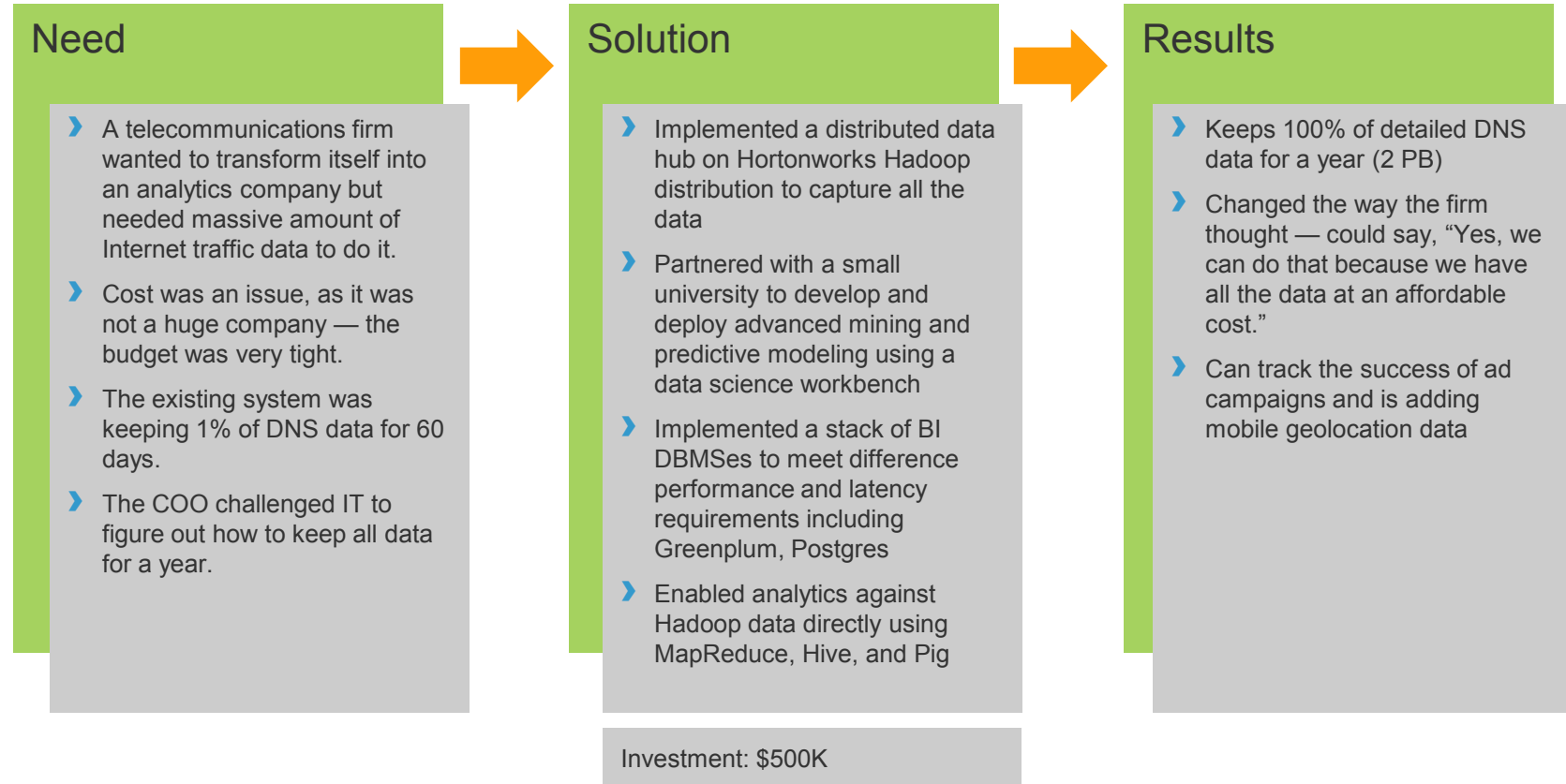
# Pharmaceutical company's architecture is really hub-and-spoke

The pharmaceutical company quickly realized the business value of adding more data and analytics using a BI DBMS to big data platform. This evolved the original EDW augmentation approach (see slide 22) to the beginnings of a hub-and-spoke pattern.

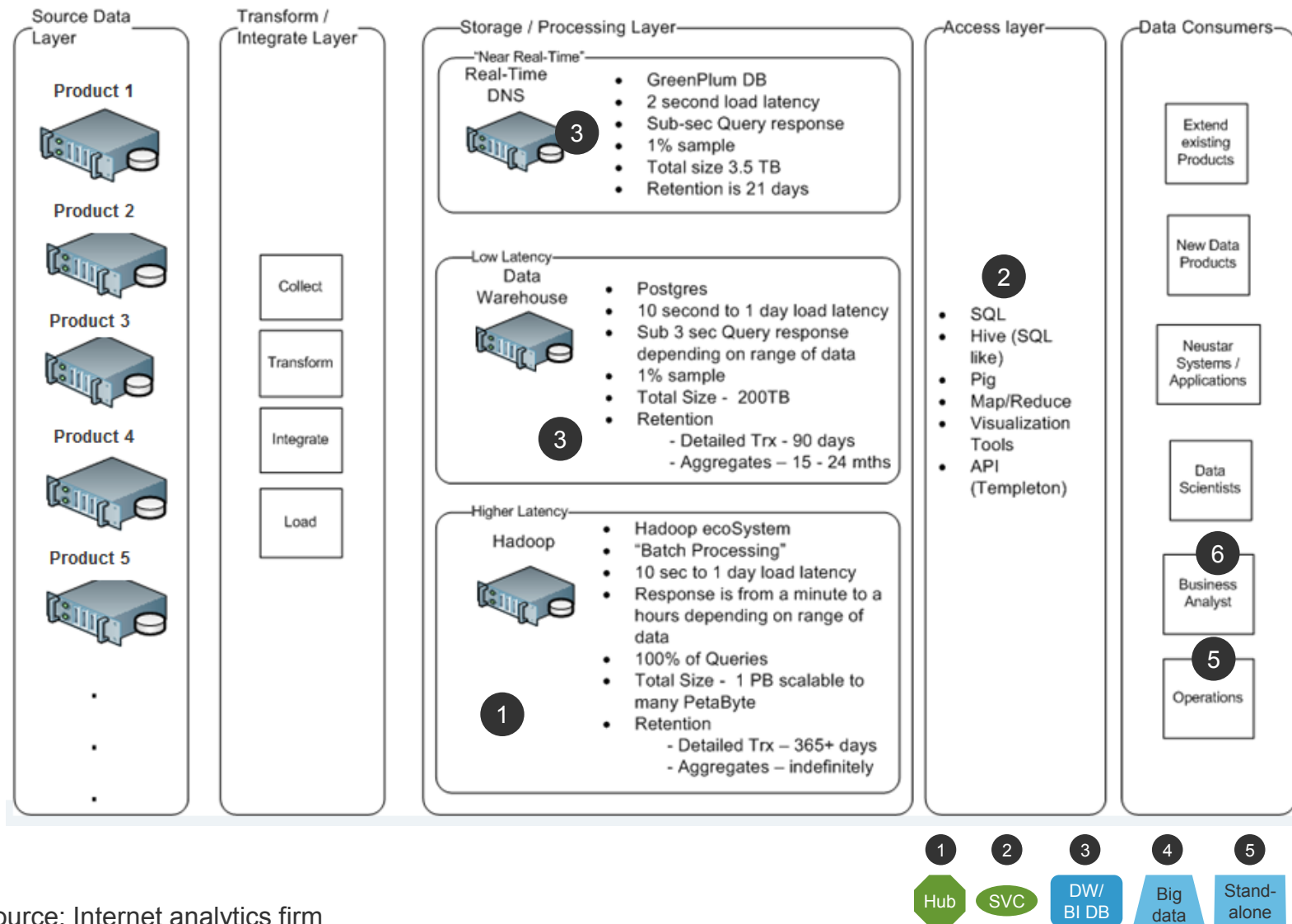


Source: pharmaceutical company

# Example: Internet analytics firm (telecommunications)



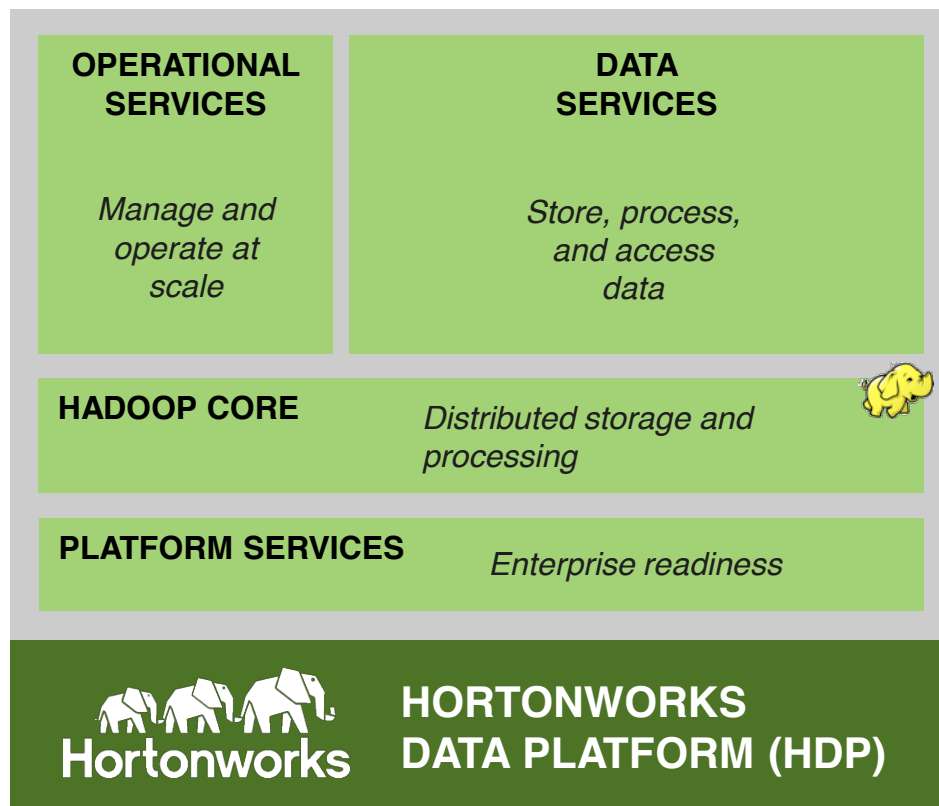
# Internet analytics firm's architecture concept



Source: Internet analytics firm

# Vendor information

## HORTONWORKS



## Hortonworks Data Platform (HDP)

### Enterprise Hadoop

- › The only 100% open source and complete distribution
- › Enterprise grade, proven and tested at scale
- › Ecosystem endorsed to ensure interoperability

Source: Hortonworks

## Forrester's point of view

- › The hub-and-spoke pattern results when firms evolve their implementations to provide multiple options for performance at different price points. Not all your data needs to “fly first class,” but some does. When firms get the advantage of sourcing all their data to one logical platform, it can be revolutionary.
- › The examples we found were Hadoop-based, but the Vestas example (slide 34) could easily evolve to hub-and-spoke if the firm added more data to its GPFS file system or added other scale-out options to capture different types of data in its hub.
- › We suggest considering hub-and-spoke when an existing big data platform requires expansion due to business needs with a broad range of performance requirements.
- › Hub-and-spoke provides the most flexibility to your business, so the business-IT handoffs must be defined. You must answer the question, “Who owns the data when we do X?”
- › Your biggest strategic technology decisions are:
  - How to build the distributed data hub. Hadoop is typically a primary component, but others, such as HBase, other NoSQL systems, and even existing MPP data warehouses and mainframes, can be a part. Data movement and harmonization is highly automated, and a single view of enterprise data is generally the endgame.
  - The selection of tools for extreme-scale data operations. Since this pattern provides a broad range of open source and package tools, the extent to which tools will be used for specific functions must be decided.

# Lessons learned from users

- › **Open source is a commitment, but it creates commitment too.** Open source can be challenging, but it creates a culture where people are committed to the technology.
- › **It takes more time to change attitudes than technologies.** Business reluctance and entrenched business-as-usual thinking is a serious obstacle that must be overcome.
- › **To make an omelet, you have to break a few eggs.** Don't be afraid to fail a few times; just do it fast. A hub-and-spoke architecture makes going back to the well for more data possible.
- › **Hardware isn't free.** In developing the business case for a hub-and-spoke, take into account the hardware costs for any new infrastructure platforms required. Today Hadoop runs only on physical x86 or in the cloud; however, appliance-based and virtualization-optimized options are coming to market that allow more flexibility.

# What it means

- › **The future of data management is a hyperflexible hub-and-spoke.** Diversity of data needs and the rapidly expanding sets of available data will force firms to rethink their data warehouse centered architectures.
- › **The critical change: Move from ETL to “ELT (TL, TL, TL . . .).”** Data will be loaded into cheap distributed storage for extreme-scale operations and transformed many times as it is offloaded to spoke systems for more expensive BI or even operational analytic uses.
- › **Hadoop will have a place in almost all hub-and-spoke architectures.** Especially as it evolves for high availability and low latency analytics. This will challenge the BI DBMS vendors to justify their pricing. Technologies like Impala (Cloudera), M7 (MapR), and Stinger (Hortonworks) address low latency structured analytic problems using different approaches. The wise will follow this technology advancement carefully.
- › **Streaming technology will create a new pattern.** Streaming platforms, such as those from IBM, SQLstream, and Apache, emerge as firms solve static data analytics problems. Streaming allows analytics on extremely high-velocity data because it is not persisted beyond a very short window.
- › **In-memory data grids will enable several orders of magnitude improvement.** Hot vendors like Platfora and ScaleOut Software could change the hub-and-spoke architecture significantly by enabling ultra high performance.
- › **Transactional and analytic technologies will fuse.** As storage technology such as solid state becomes cheaper and compression improves, operational systems in-memory will erase the distinction between analytic and transactional technology infrastructure.

# Research methodology details

- › **Companies we interviewed for this report:** A pharmaceutical company, a wealth management firm, edo interactive, Vestas Wind Systems, NK, Opera Solutions, Rubicon, Razorfish, Sears, a telecommunications company, and an Internet analytics firm
- › **Vendors that helped provide client examples:** Cloudera, Composite Software, Pentaho, Actian, IBM, LexisNexis, MapR, Teradata Aster, Datameer, and Hortonworks
- › **Forrester's Forrsights Strategy Spotlight: Business Intelligence And Big Data, Q4 2012** was fielded to 634 IT executives and technology decision-makers located in Canada, France, Germany, the UK, and the US from small and medium-size business (SMB) and enterprise companies with 100 or more employees. All respondents reported working for companies that were currently using or planning to use business intelligence technologies. This survey is part of Forrester's Forrsights for Business Technology and was fielded during October 2012 and November 2012. Survey respondent incentives included gift certificates and research reports.

Each calendar year, Forrester's Forrsights for Business Technology fields business-to-business technology studies in more than 17 countries spanning North America, Latin America, Europe, and developed and emerging Asia. For quality control, we carefully screen respondents according to job title and function. Forrester's Forrsights for Business Technology ensures that the final survey population contains only those with significant involvement in the planning, funding, and purchasing of IT products and services. Additionally, we set quotas for company size (number of employees) and industry as a means of controlling the data distribution. Forrsights uses only superior data sources and advanced data-cleaning techniques to ensure the highest data quality.



# Selected Forrester research

- › Upcoming “Consumption Diversity Requires Hyperflexible Data Management” Forrester report
- › June 12, 2013, “Deliver On Big Data Potential With A Hub-And-Spoke Architecture” Forrester report
- › January 3, 2013, “The Forrester Wave™: Big Data Predictive Analytics Solutions, Q1 2013” Forrester report
- › September 30, 2011, “Expand Your Digital Horizon With Big Data” Forrester report
- › May 27, 2011, “It’s The Dawning Of The Age Of BI DBMS” Forrester report

# Thank you

**Brian Hopkins**

*[www.forrester.com](http://www.forrester.com)*