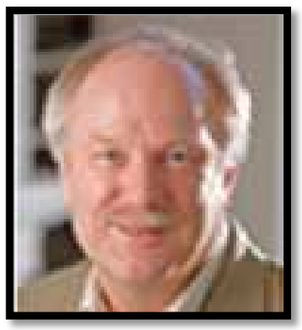
Welcome! We will begin momentarily.

Live Webcast:

Back to the Future – MapReduce, Hadoop and The Data Scientist

Today's event is co-sponsored by:

Featured Speakers:



Colin White is the President of DataBase Associates Inc. and founder of BI Research. As an analyst, educator and writer he is well known for his in-depth knowledge of data management, information integration, and business intelligence technologies and how they can be used for building the smart and agile business. With many years of IT experience, he has consulted for dozens of companies throughout the world and is a frequent speaker at leading IT events.

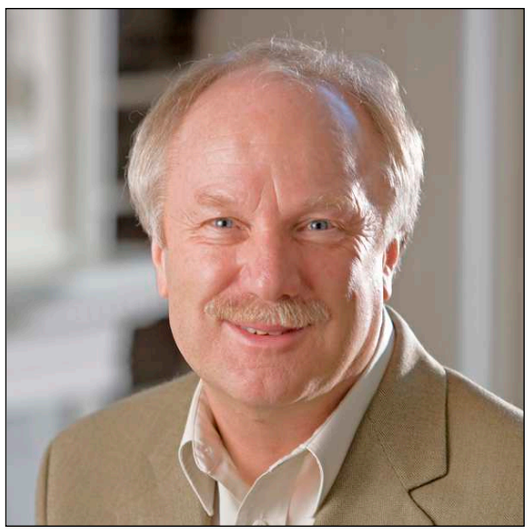
Ari Zilka is the Chief Products Officer at Hortonworks — Ari has more than 20 years of software development expertise and a deep understanding of open source, enterprise software, and the execution required to build successful products. Ari was previously founder and CTO at Terracotta. Previously, Ari was an Entrepreneur-in-Residence at Accel Partners. Before joining Accel, Ari was the Chief Architect at Walmart.com, where he led the innovation and development of the company's new engineering initiatives. Prior to Walmart.com, Ari worked as a consultant at Sapient and at PriceWaterhouseCoopers. Ari holds a B.S. in Electrical Engineering Computer Science as well as in Mechanical Engineering from University of California, Berkeley

Tasso Argyros is Co-President, Teradata Aster, leading the Aster Center of Innovation — Tasso has a background in data management, data mining and large-scale distributed systems. Before founding Aster Data, he was in the Ph.D. program at Stanford University.

Tasso was recognized as one of Bloomberg BusinessWeek's Best Young Tech Entrepreneurs for 2009. He holds a Master's Degree in Computer Science from Stanford University and a Diploma in Computer Engineering from Technical University of Athens.

Back to the Future: MapReduce, Hadoop and the Data Scientist

Colin White
President, BI Research
Teradata Aster Data Webinar
June 2012



Historical Perspective: Application & Data Growth

1960 — 1970 — 1980 — 1990 — 2010

First OLTP systems

Early OLAP products

First commercial RDBMSs

Early DW products

Big data & advanced analytics

Copyright © BI Research, 2012

What is Big Data?

A term that represents *analytic workloads* and *data management solutions* that could not previously be supported because of cost and/or technology limitations

Three important technologies:

- Analytic RDBMSs
- Non-relational systems
- Stream processing systems

It's not just about "big" data volumes!

Copyright © BI Research, 2012

What are Advanced Analytics?

Basic analytics: *report on what happened*

- Canned batch reports and interactive drill down/up reports
- Fixed analytic dashboards (may or may not be event driven)
- BI automation (alerts and recommendations)

Standard analytics: *explore why it happened*

- Interactive analytic dashboards with drill up/down, slice/dice, etc. (OLAP)
- BI automation (decision analysis workflows)
- Customizable data mashups and BI widgets

Advanced analytics: *predict what may happen/investigate new opportunities*

- Data mining, analytic modeling, and statistical/predictive analytic functions
- Advanced visualization

Copyright © BI Research, 2012

Advanced Analytics: Customer Marketing

Internal and External Data
- Retail measurement POS data
- Consumer panel household data
- Customer demographics
- Customer purchase behavior
- Customer billing data
- Customer satisfaction data
  Customer market survey data
- Third-party data
  (ACXIOM, D&B)
- Merchandising sales data
  (SAP,JD Edwards)

Statistical Techniques
- Multiple linear regression
- Non-linear progression
- Factor analysis
- Structural equations model
- Cluster analysis
- Forecasting ARIMA methods
- CHAID
- Logistic regression

Non-Statistical Techniques
- Blog mining
- Neural networks
- Market basket analysis

Market Analytics
- Market volume forecasting
- Market share models
- Marketing/media mix modeling
- Promotion effectiveness
- Market basket analysis
- Price elasticity modeling
- Product portfolio analysis
- Lifestyle segmentation
- Impact analysis
- Demand forecasting

Source: www.dexterity.in

Copyright © BI Research, 2012

New Data Sources: Multi-Structured Data

Data that has unknown, ill-defined, or multiple schemas

- Machine generated event data, e.g., sensor data, system logs
- Internal/external web content including social computing data
- Text/document data
- Multi-media data, e.g., audio, video

May be managed by a variety of different file and database systems - often not integrated into a data warehouse

Increasing number of analytical techniques to extract useful business facts from multi-structured data, e.g., MapReduce

These business facts can be used to extend traditional BI data analytics and models

Copyright © BI Research, 2012

What is a Data Scientist? CITO Research Interviews

"Data scientists turn big data into big value, delivering products that delight users, and insight that informs business decisions. Strong analytical skills are given: above all a data scientist needs to be able to derive robust conclusions from data."

Daniel Tunkelang, Principal Data Scientist, LinkedIn

"A data scientist is someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning."

Hilary Mason, Chief Scientist at bitly

"... someone who has the both the engineering skills to acquire and manage large data sets, and also has the statistician's skills to extract value from the large data sets..."

John Rauser, Principal Engineer, Amazon.com

Source: citoresearch.com/content/growing-your-own-data-scientists

Copyright © BI Research, 2011

Data Science Skills Requirements

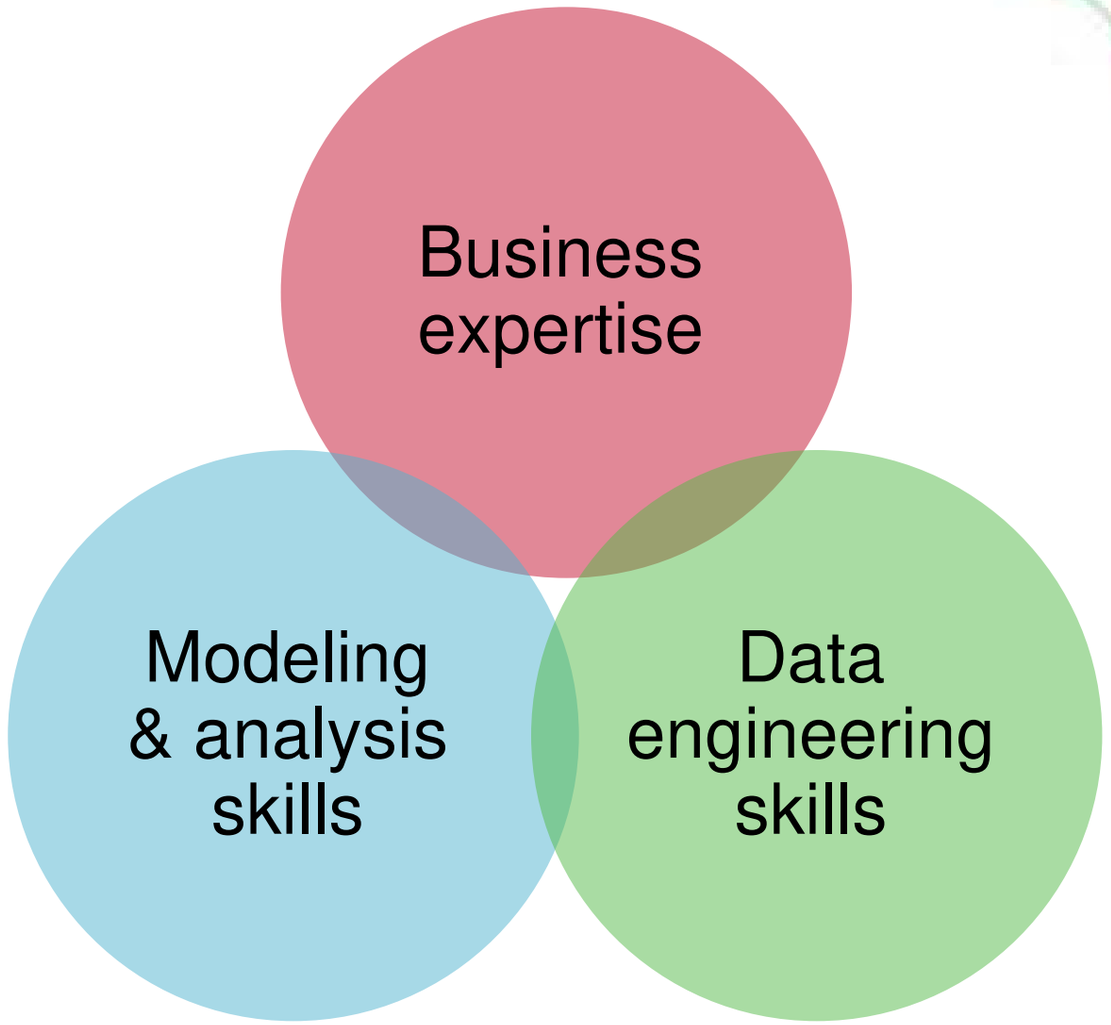
Business domain subject matter expert with strong analytical skills

Creativity and a good communicator

Knowledgeable in statistics, machine learning and data visualization

Able to develop analytical solutions using languages, such as MapReduce, R, SAS, etc.

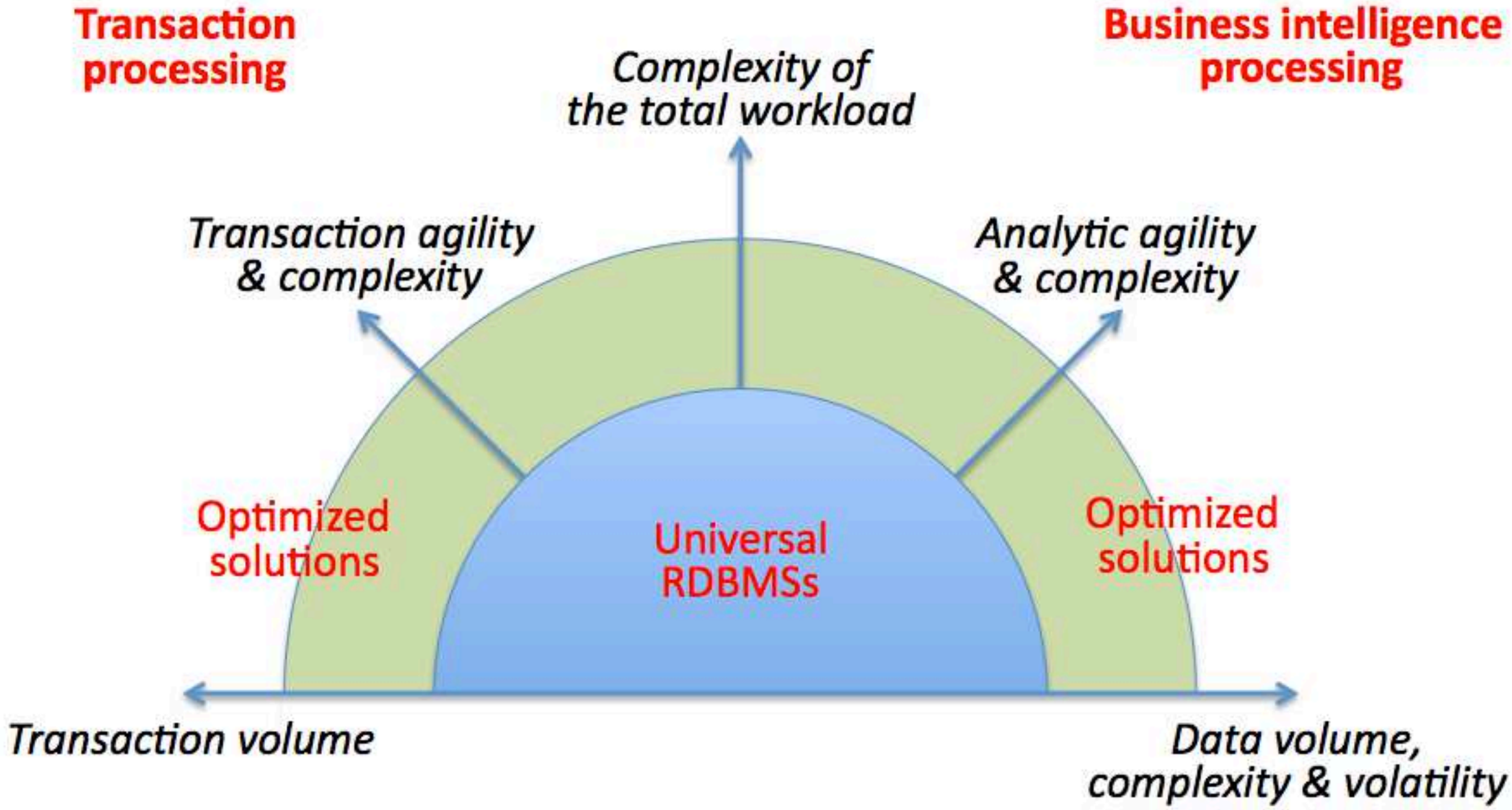
Adept at data engineering, including discovering and blending large amounts of data



Is this one person or a team of specialists?

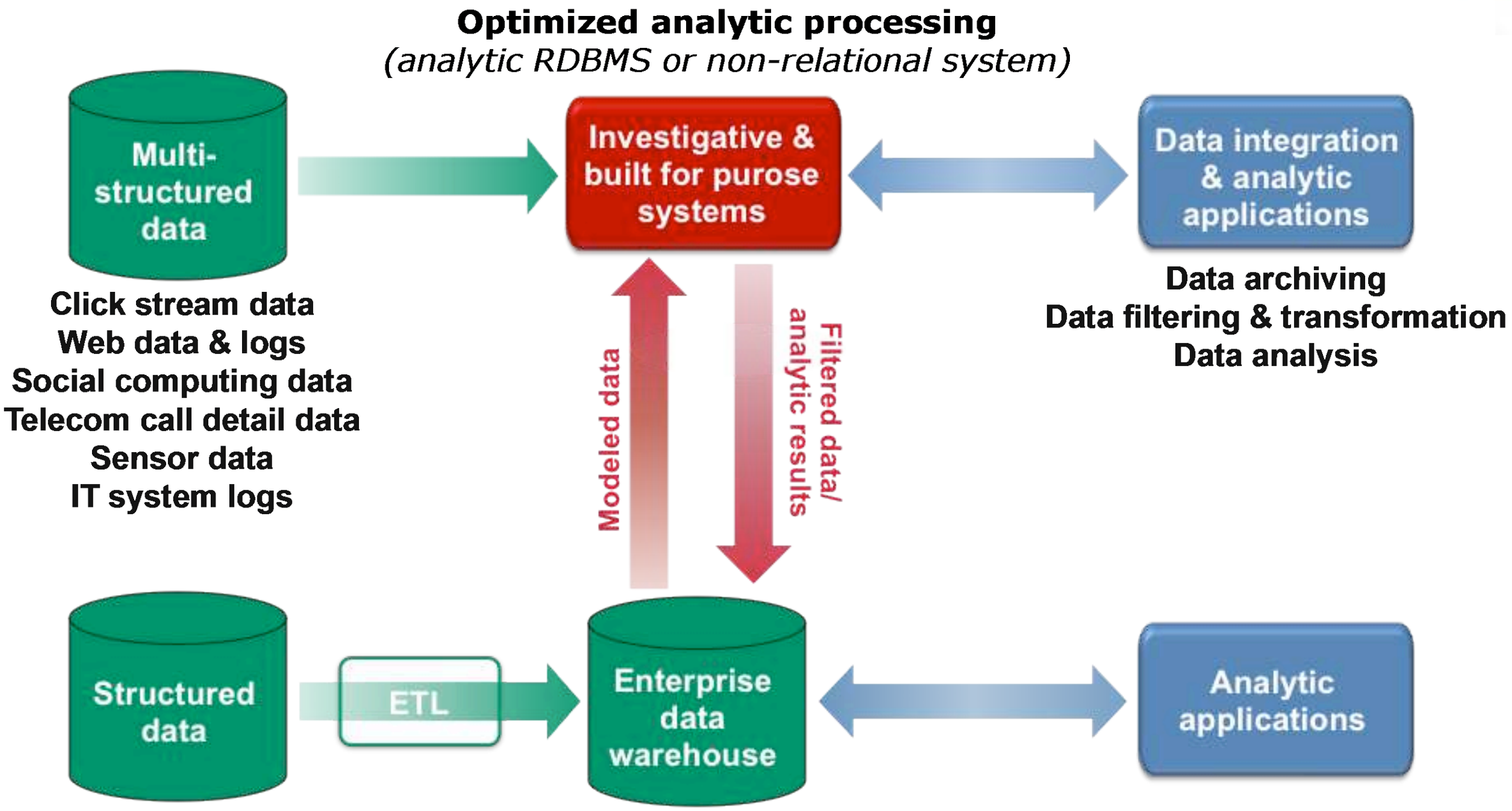
Copyright © BI Research, 2011

Workload Growth: The Need for Optimized Systems



Copyright © BI Research, 2012

Investigative Computing & Built-For-Purpose Systems



Copyright © BI Research, 2012

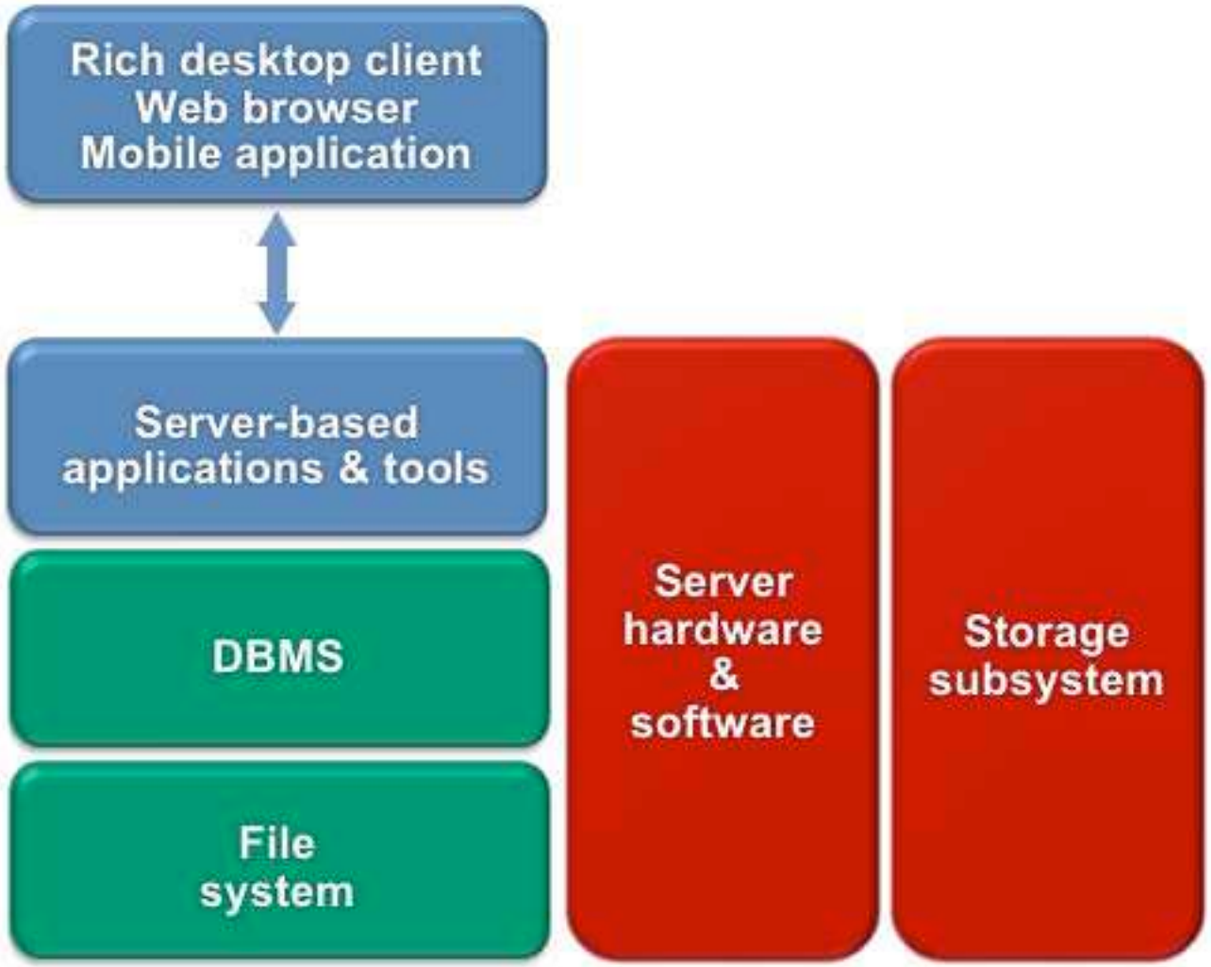
Optimized Analytical Processing

Analytic RDBMSs

advanced analytic techniques
in-database analytic functions
in-memory analytical processing
enhanced storage structures
improved data compression
support for hybrid storage
intelligent workload management
appliances

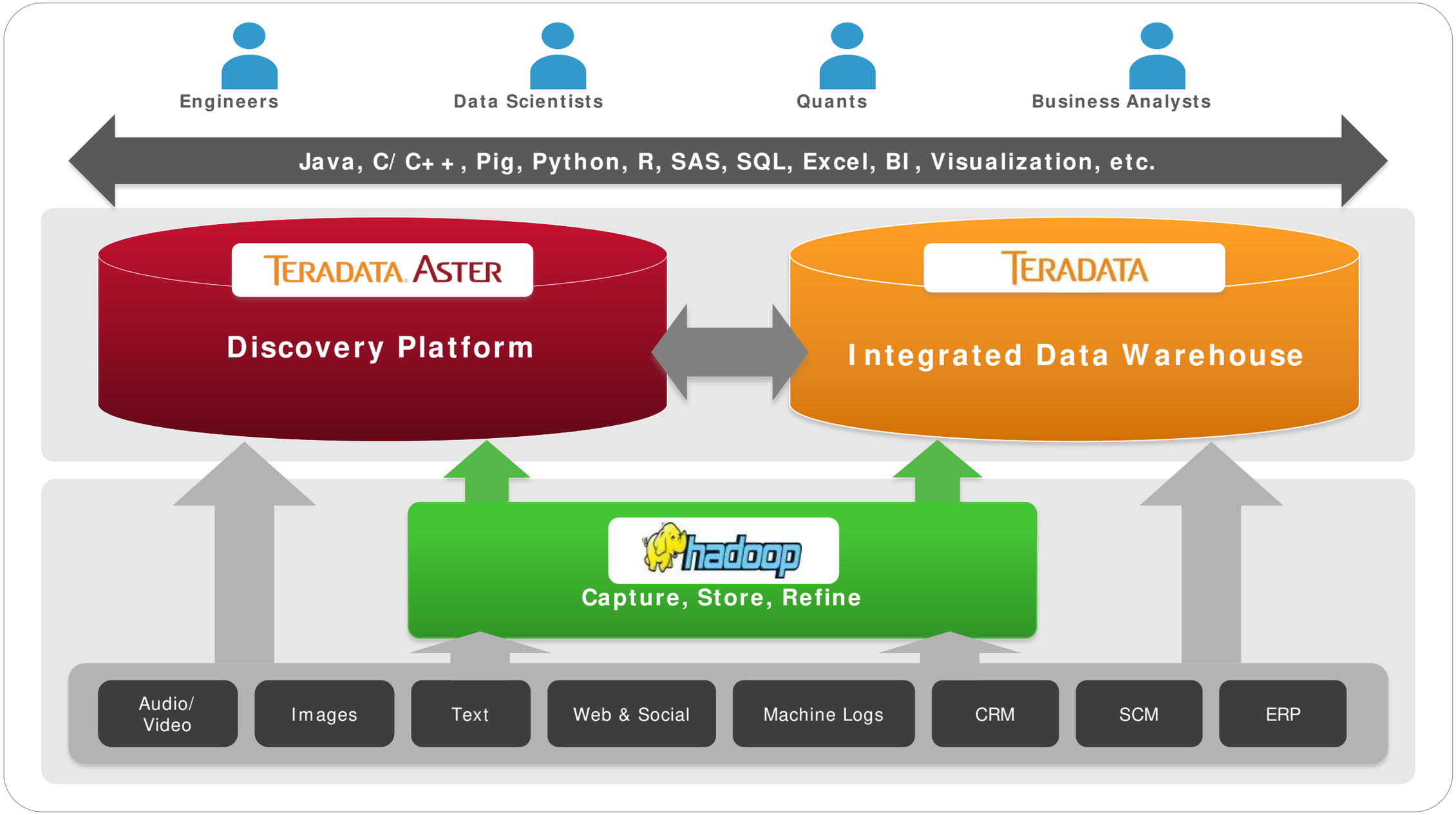
Non-Relational Systems

document stores
graph databases
parallel processing systems, e.g., Hadoop (MapReduce, HDFS, Hive, HBase)



Copyright © BI Research, 2012

The Extended Data Warehouse: Teradata Example



Copyright © BI Research, 2012

Advanced Analytics: Implementing MapReduce

Important to realize that MapReduce (MR) is a programming model for parallel computing not a programming or query language

MR programs can be coded using different languages, e.g., Java, C++, Perl, Python, Ruby, R

MR program libraries may support different file and database systems

The Hadoop distributed computing framework supports MR using the Hadoop Distributed File System (HDFS)

Hadoop also includes Hive, which employs a SQL-like language to generate MR programs

Some analytic RDBMS vendors support MR programs as *in-database* analytic functions that can be used in SQL statements

Copyright © BI Research, 2012

Using Hive: Considerations

> "Hadoop was not easy for users, specially for those who were unfamiliar with MapReduce. Hadoop lacked the expressibility of languages like SQL, and users spent hours/days to write analysis programs. Bringing this data closer to users is what inspired us to build Hive."
>
> *Ashish Thusoo, FaceBook, June 2009, www.facebook.com/note.php?note_id=89508453919*

The benefit of Hive is that it improves the speed of MR development

Hive is immature and for complex queries the user may need to aid the HiveQL optimizer through hints and HiveQL language constructions

The main Hive use case is the sequential processing of large multi-structured data files in batch - it is not suited to low-latency *ad hoc* queries

There are other ways of improving the usability of Hadoop/MapReduce:

- MapReduce IDEs for building Hadoop MR functions and programs
- Analytical tools that run on an Hadoop system
- RDBMS SQL access to HDFS data using a Hadoop connector and HCatalog

Copyright © BI Research, 2011

Conclusions

Organizations need to distinguish between data that can and cannot be consolidated into an enterprise data warehouse (EDW) using an RDBMS such as IBM DB2, Microsoft SQL Server, Oracle or Teradata Database

For data that remains outside of the EDW, organizations should evaluate the best approach to use for archiving, filtering and/or analyzing the data

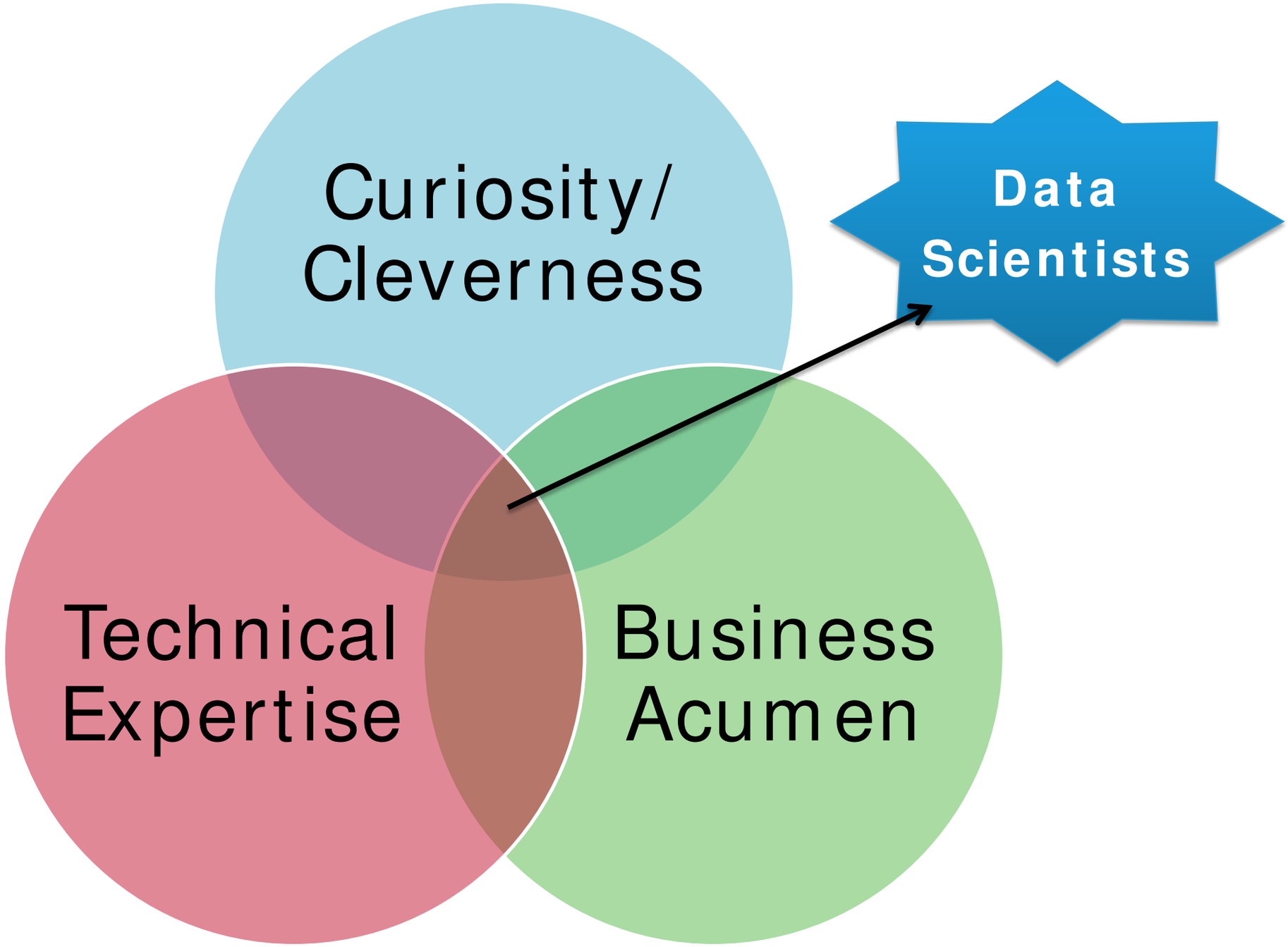
- RDBMS optimized for analytic processing such as the Teradata Aster system
- Non-relational system such as Hadoop with Hive and HCatalog
- Analyze data in-motion as it flows through enterprise systems using data streaming technology

For EDW data it will be necessary to determine the data that needs to be extracted into a data mart, data cube, or investigative sandbox for more detailed analysis and investigation

For analytic RDBMSs and non-relational systems such as Hadoop it will be necessary to integrate these systems with each other and the EDW environment

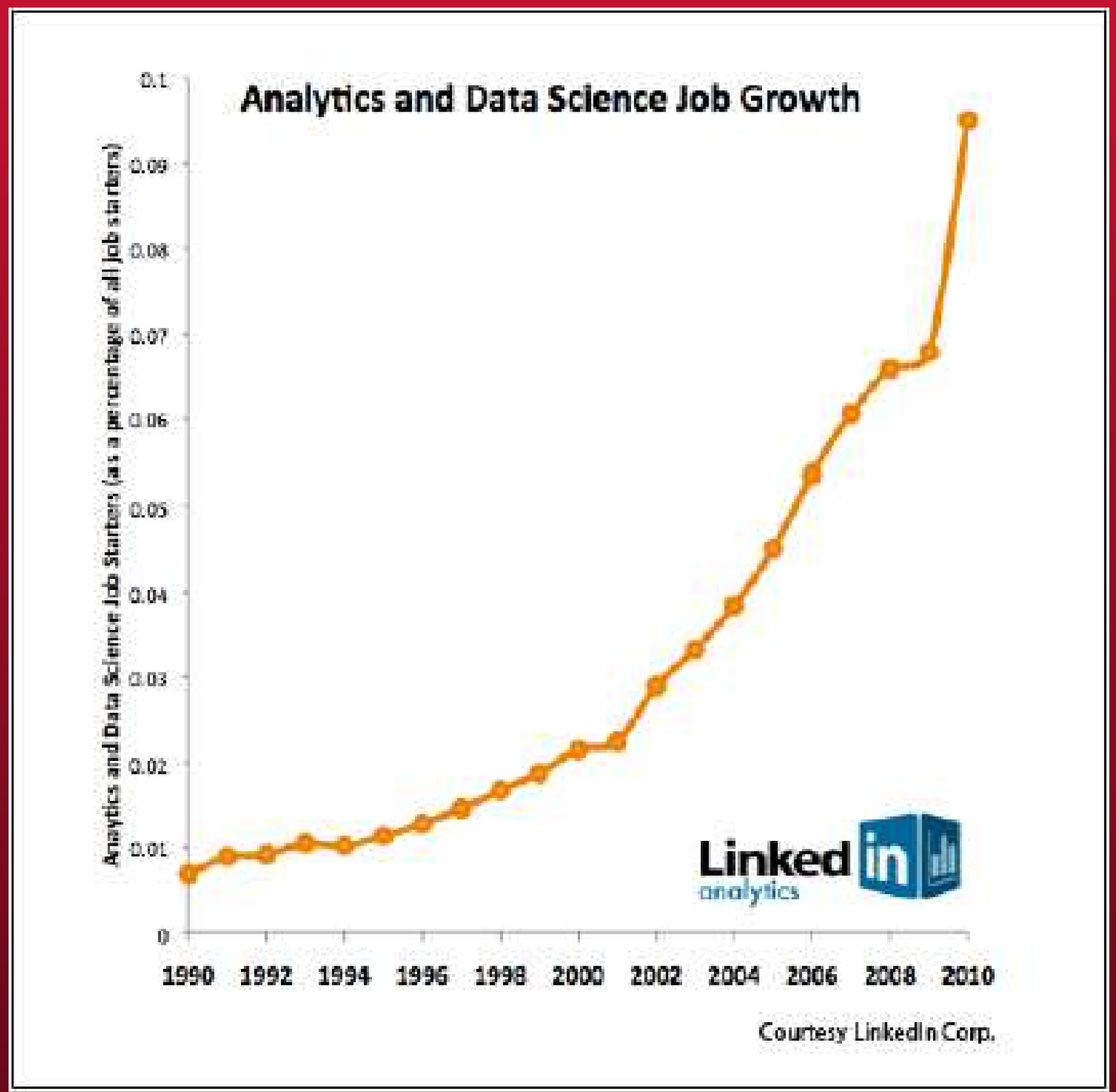
Copyright © BI Research, 2012

What is Data Science?



TERADATA. ASTER

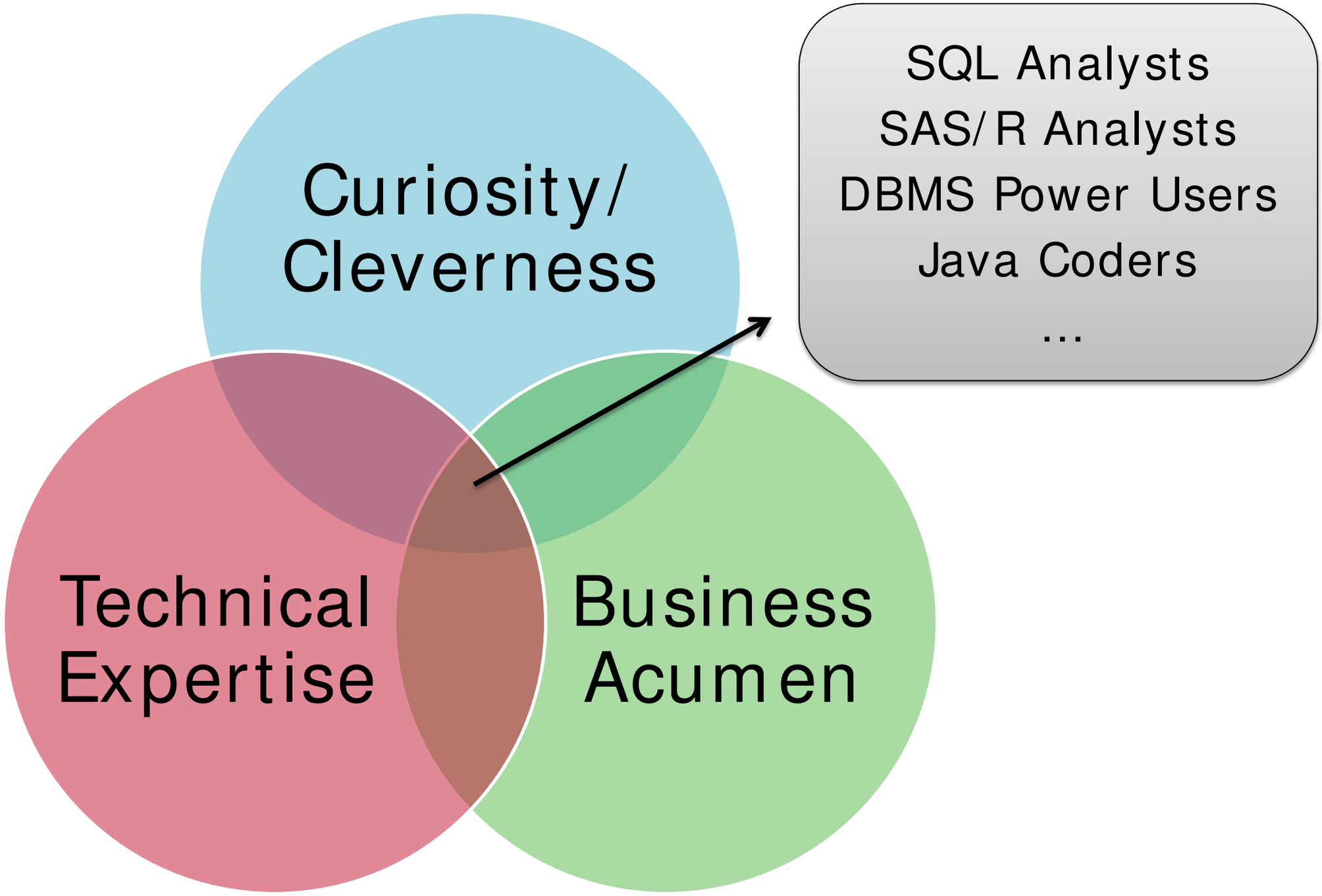
Data Science is Exploding



Become a Data Scientist

Launch your career as an Analytics professional in just 10 months.
analytics.ncsu.edu

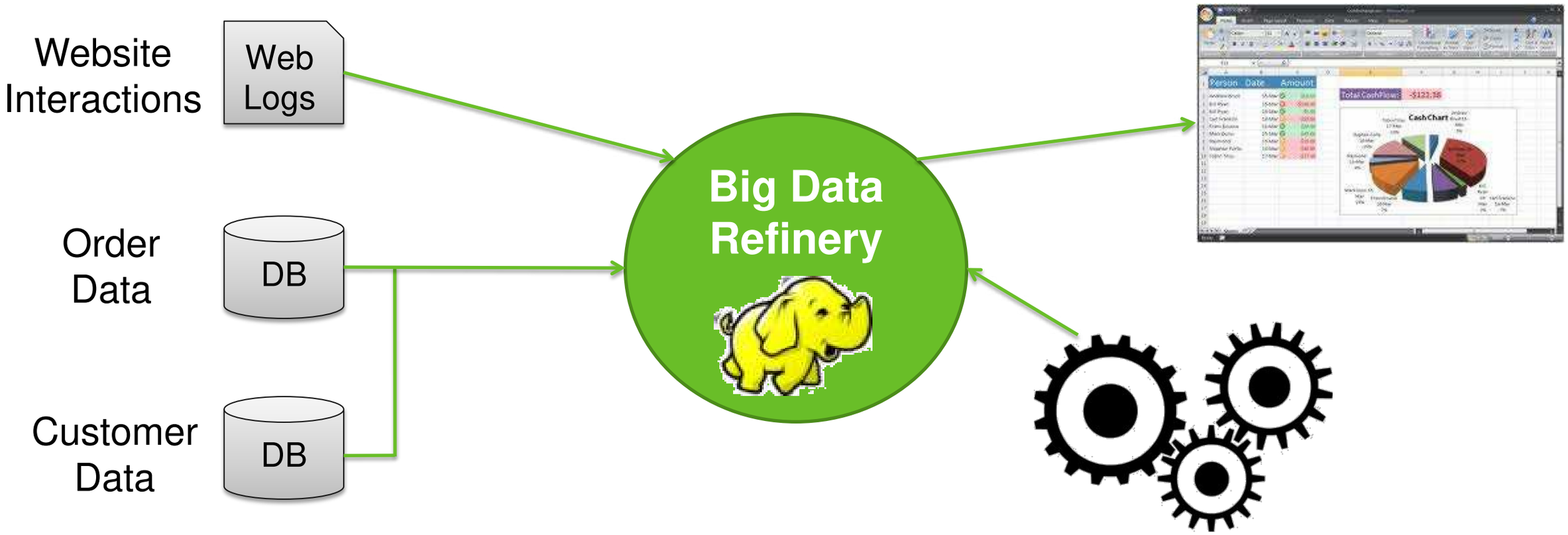
Data Scientists in the Enterprise are **Not Only Developers**



Enabling the Data Scientist

1 Web Log files via WebHDFS APIs

4 Visualize the most effective web offers via BI tool of choice



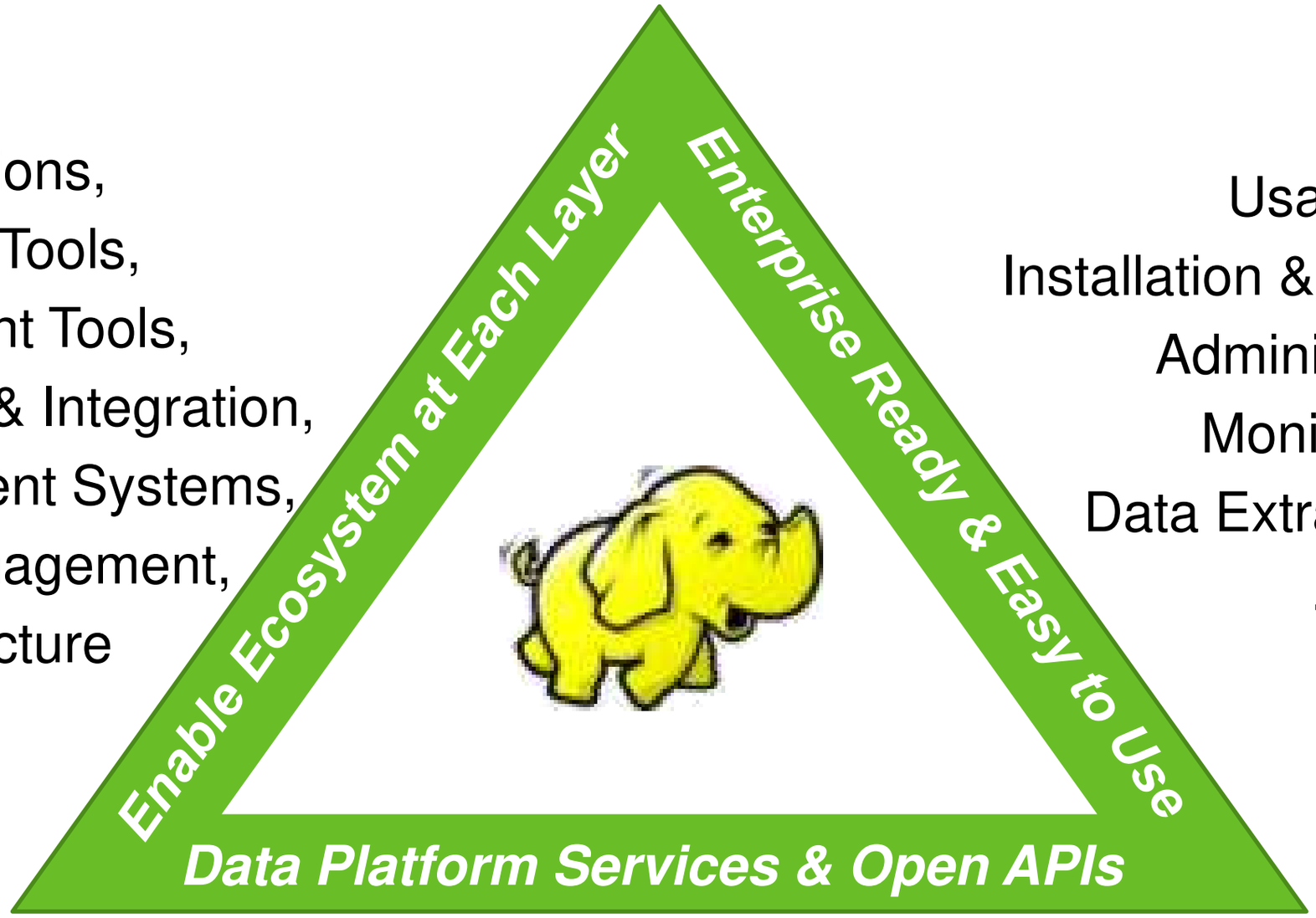
2 Customer & Order data via Talend & HCatalog for schema

3 Process, analyze, and join data via Talend, Pig, & HCatalog

© Hortonworks Inc. 2012

What is needed for traditional enterprises to adopt Hadoop?

Applications,
Business Tools,
Development Tools,
Data Movement & Integration,
Data Management Systems,
Systems Management,
Infrastructure



Usability,
Installation & Configuration,
Administration,
Monitoring,
Data Extract & Load,
...

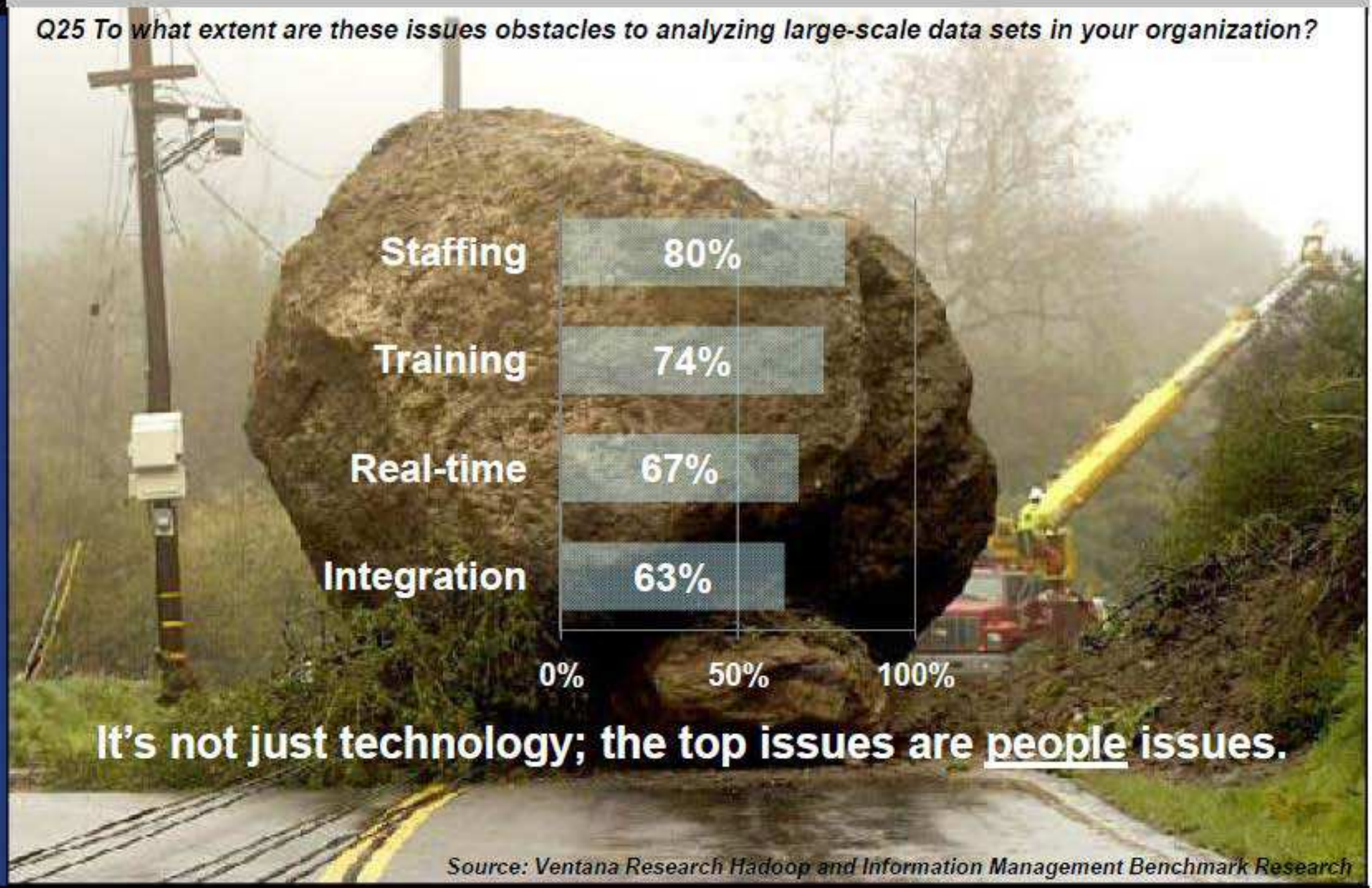
Metadata, Indexing, Search, Security,
Management,
HA, DR, Replication, Multitenancy, ...

© Hortonworks Inc. 2012

Obstacles in Analyzing Big Data Cited by Hadoop Users

Hadoop and Information Management

Q25 To what extent are these issues obstacles to analyzing large-scale data sets in your organization?



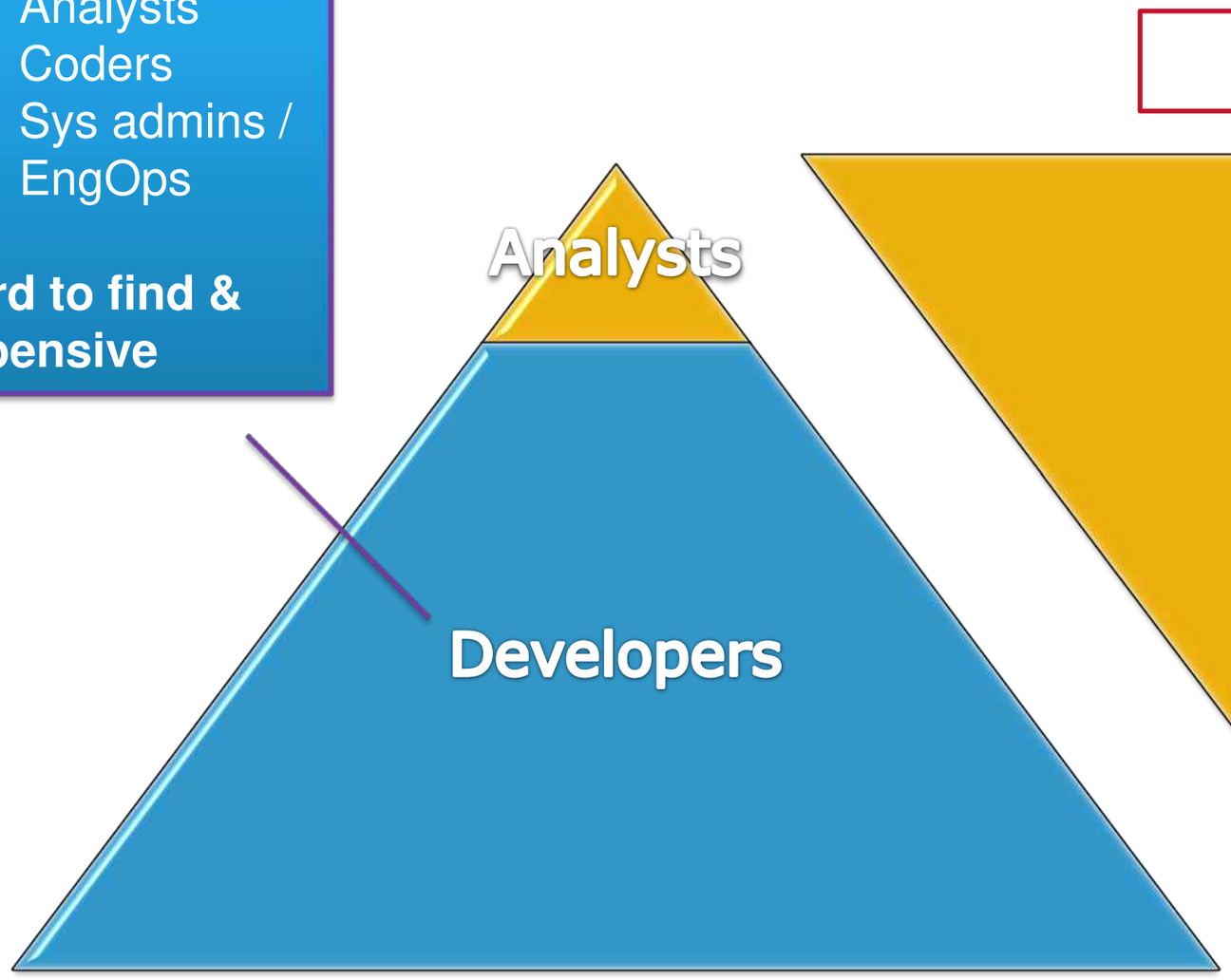
It's not just technology; the top issues are people issues.

Source: Ventana Research Hadoop and Information Management Benchmark Research

Data Scientists Have Different Skills

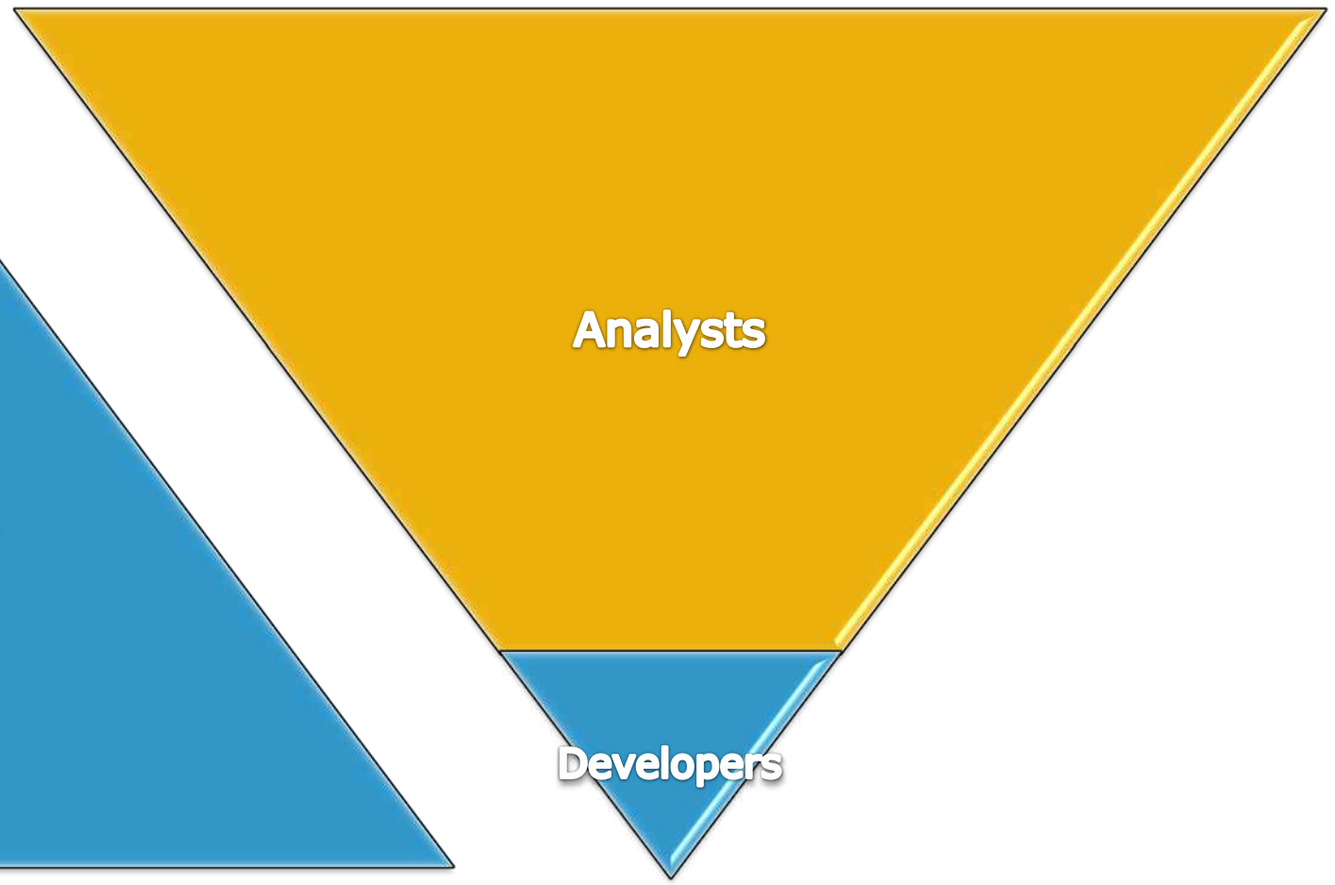
Combination of:
- Analysts
- Coders
- Sys admins / EngOps

Hard to find & expensive



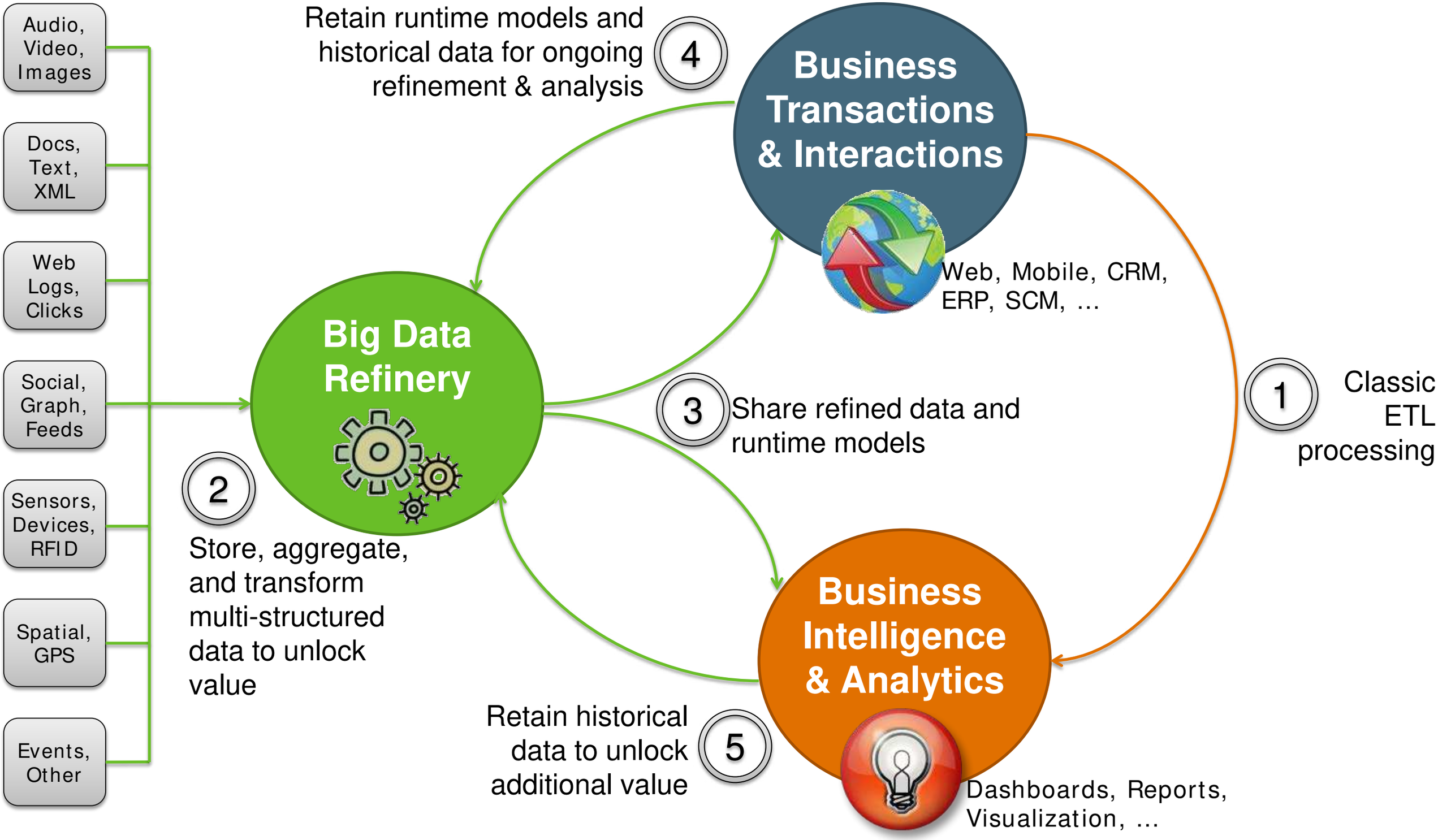
Web Startups

Enterprises



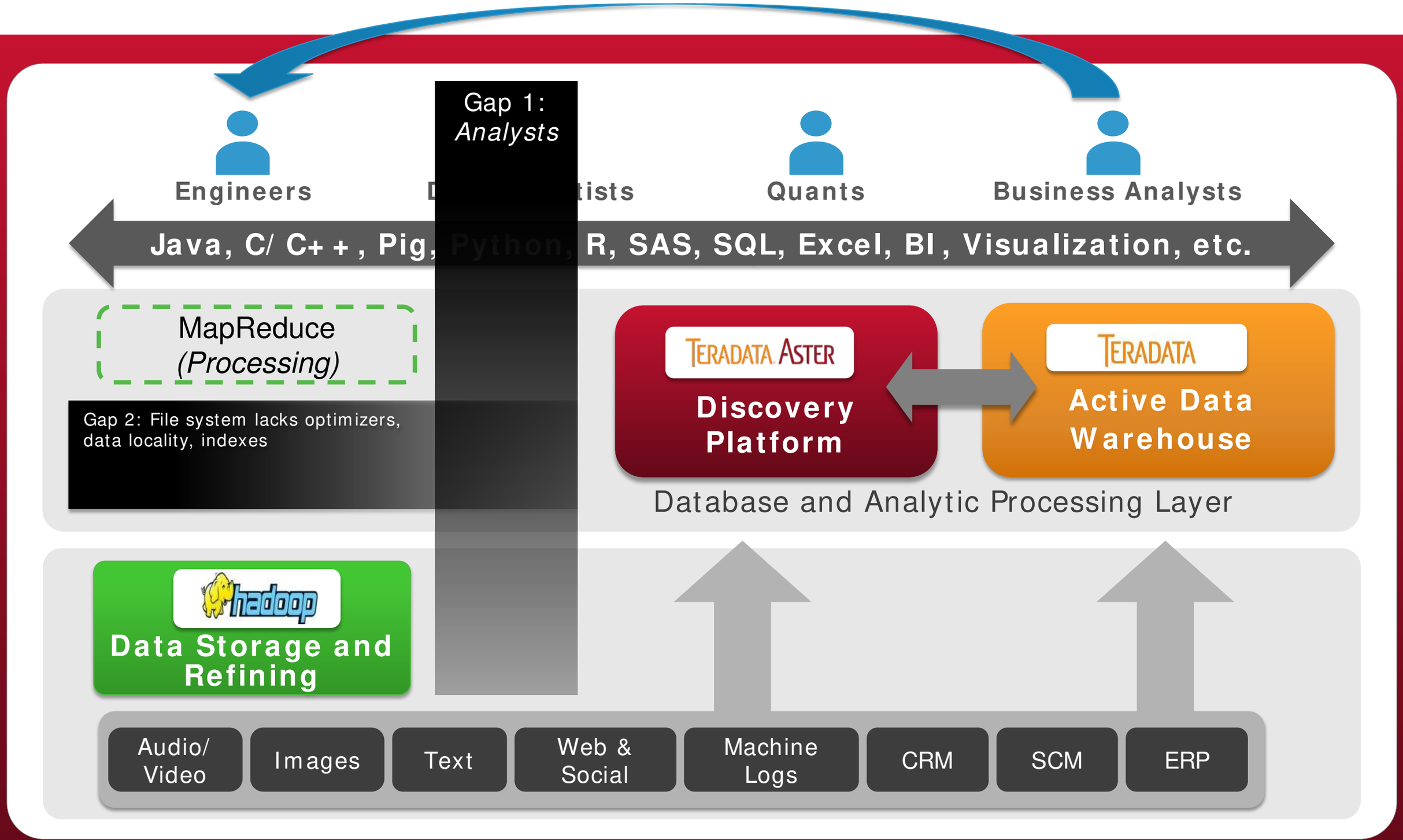
TERADATA. ASTER

Will SQL-H obviate Hadoop?



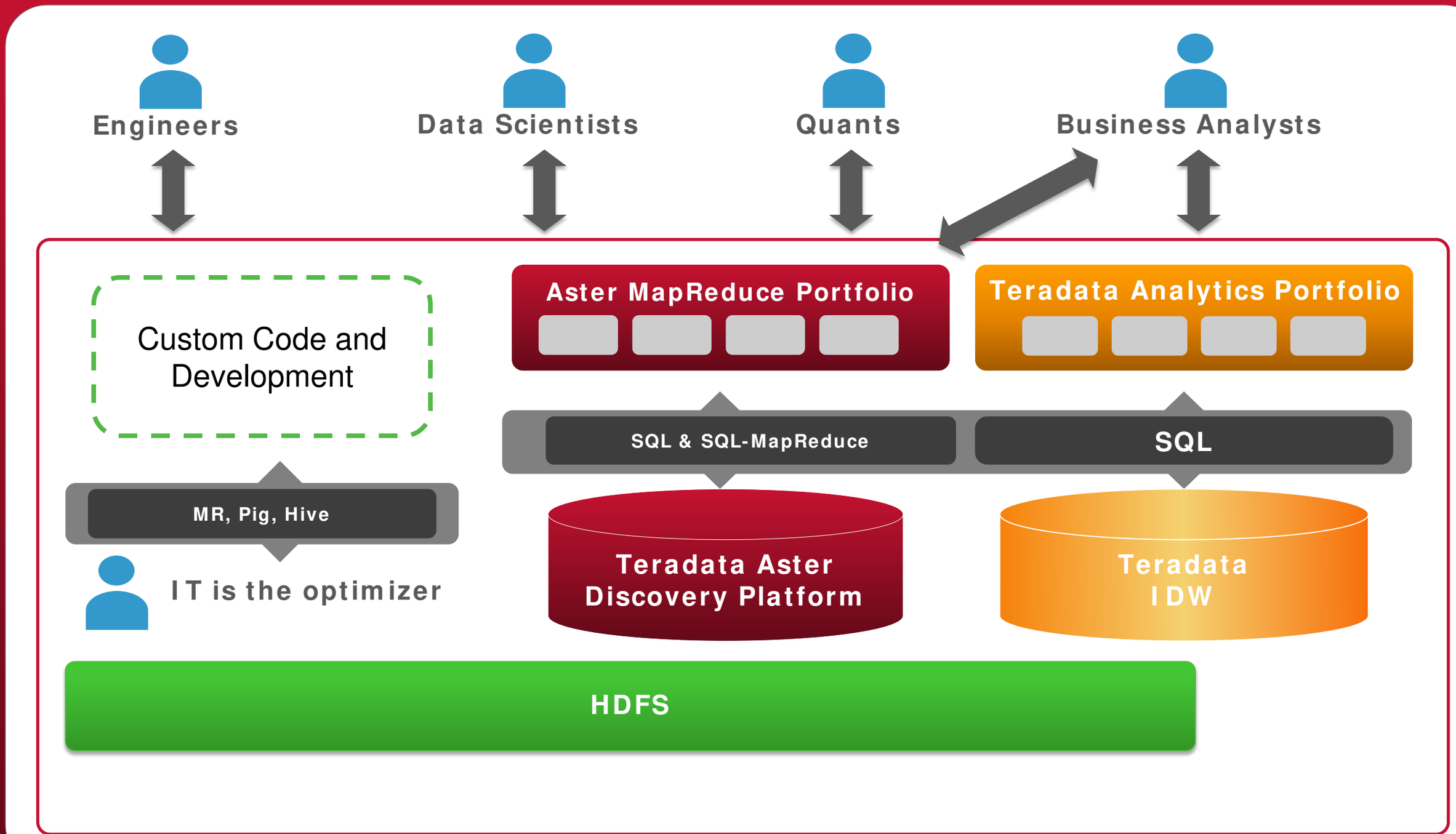
© Hortonworks Inc. 2012

The Big Data Architecture Today Has Gaps



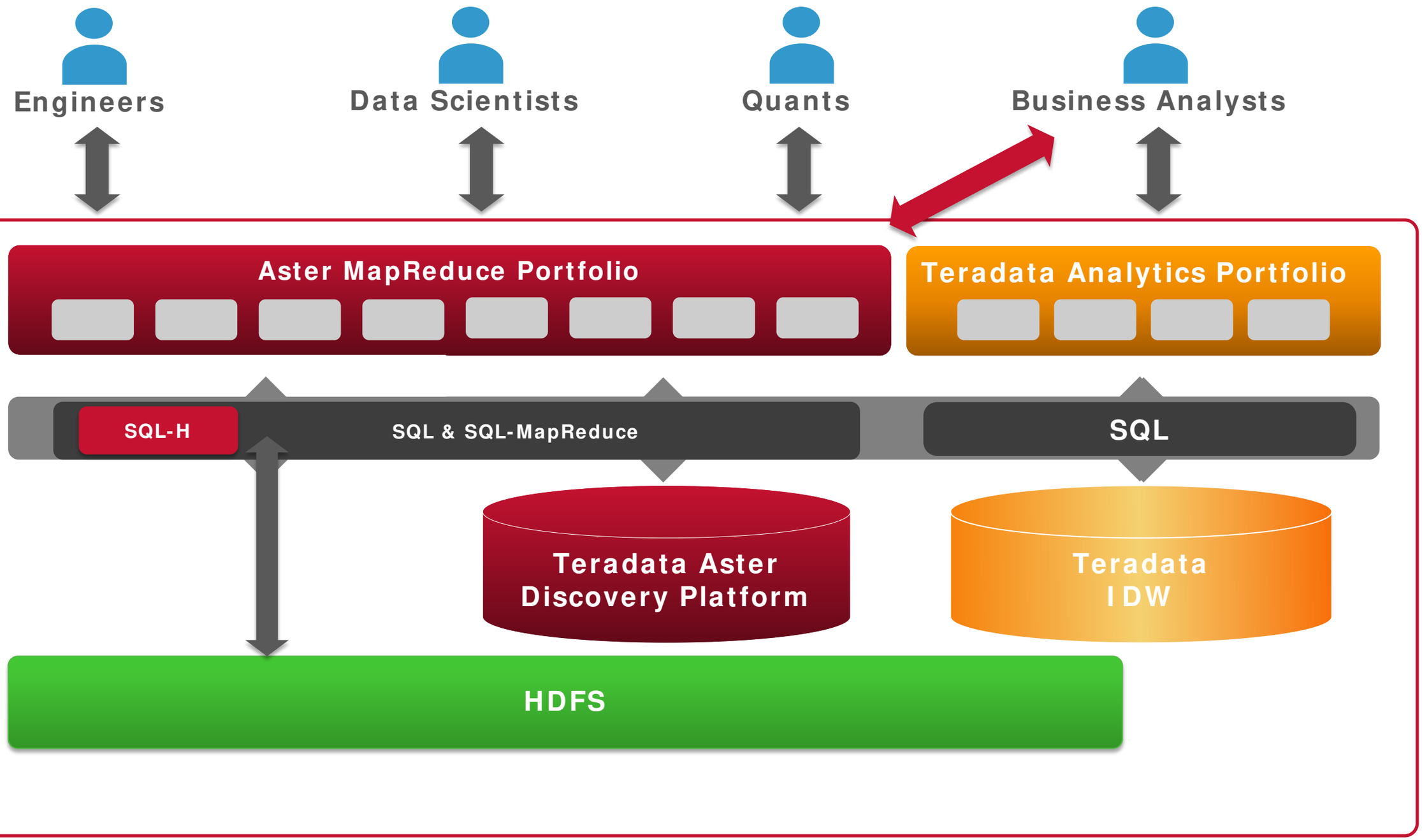
Confidential and proprietary. Copyright © 2012 Teradata Corporation.

Analyst's Goal: Get Insights from Data in Hadoop



Confidential and proprietary. Copyright © 2012 Teradata Corporation.

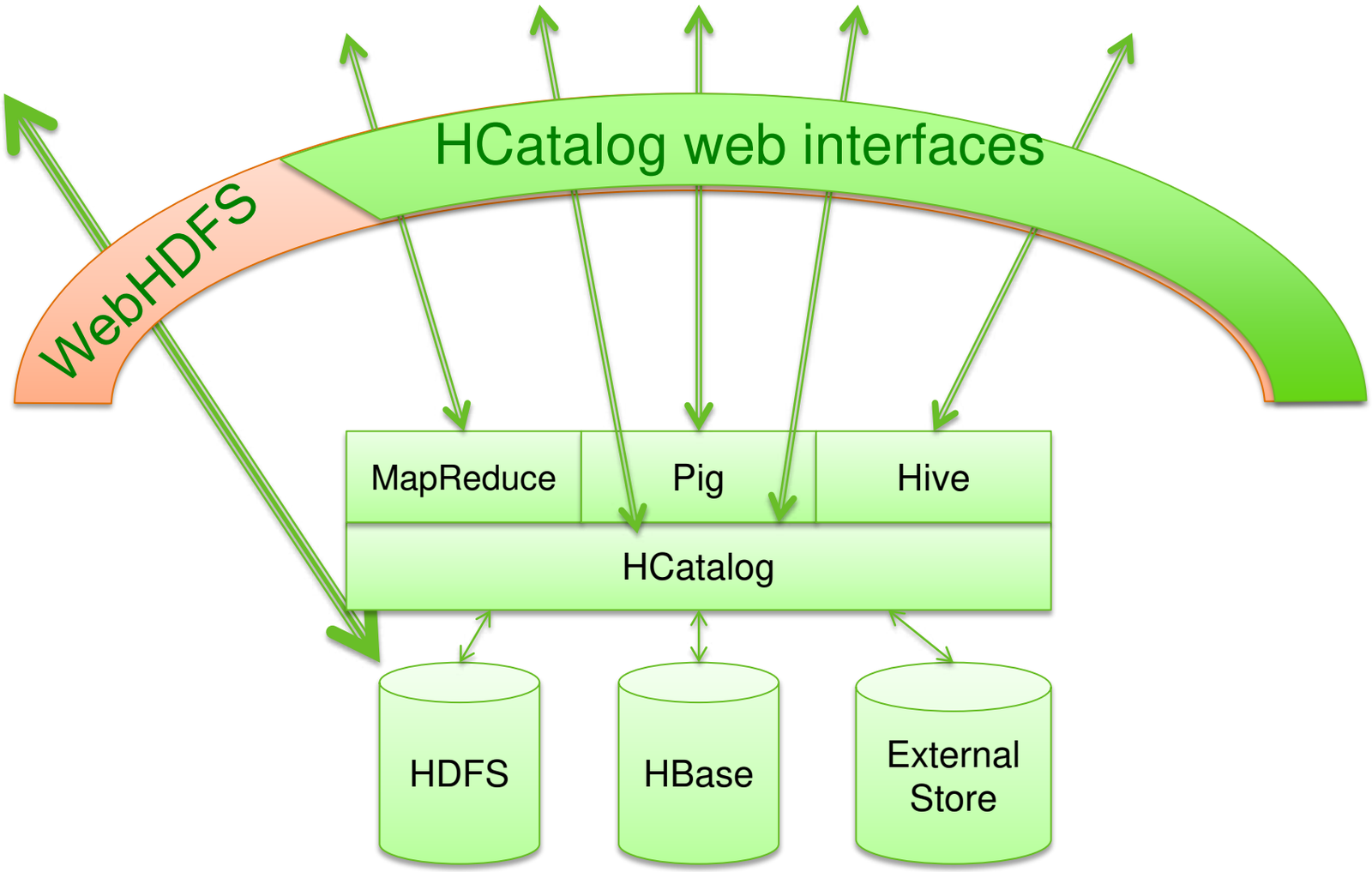
Analytics on Hadoop Data with Aster SQL-H



Confidential and proprietary. Copyright © 2012 Teradata Corporation.

Hcatalog in detail

- **Opens the door to languages other than Java**
- **Thin clients via web services vs. fat-clients in gateway**
- **Insulation from interface changes release to release**

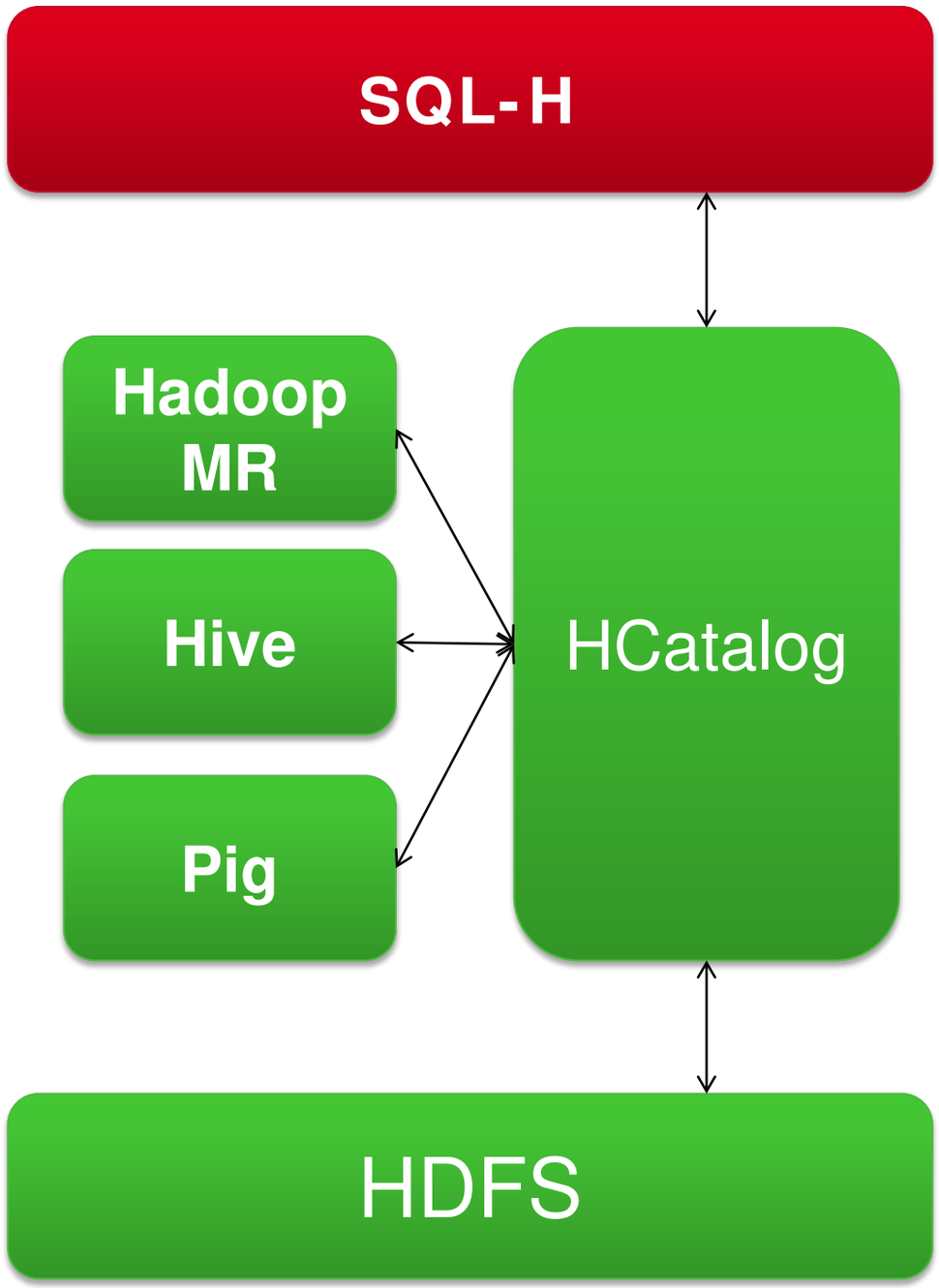


© Hortonworks Inc. 2012

Aster SQL-H Integration with Hadoop Catalog

A Business User's Bridge to Analyzing Data in Hadoop

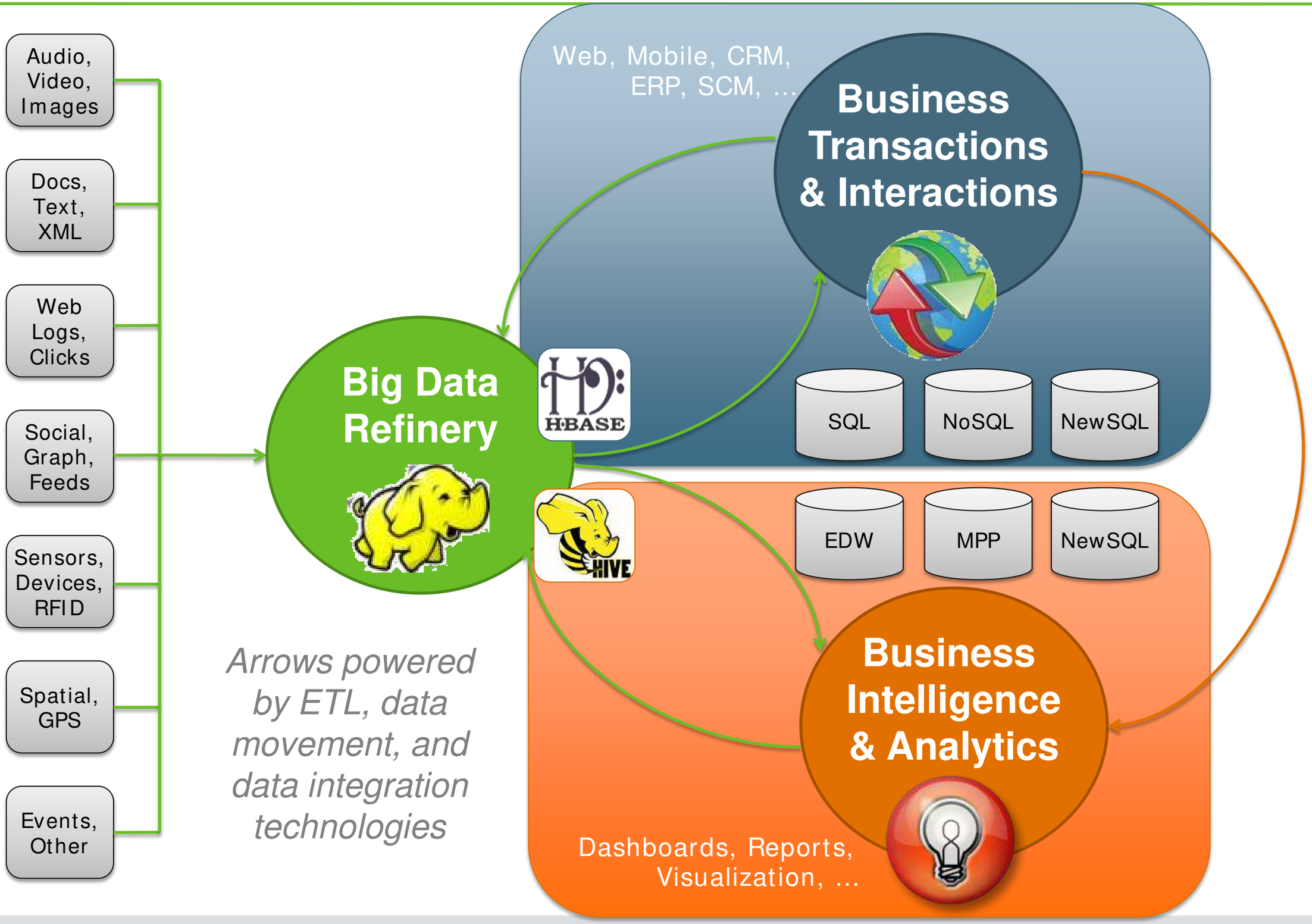
- Industry's First Database Integration with Hadoop's HCatalog
- Abstraction layer to easily and efficiently read structured & multi-structured data stored in HDFS
- Uses Hadoop Catalog (HCatalog) to perform data abstraction functions (e.g. automatically understands tables, data partitions)
- HDFS data presented to users as Aster tables
- Fully accessible within the Aster SQL and SQL-MapReduce processing engines, plus ODBC/JDBC & BI tools



Confidential and proprietary. Copyright © 2012 Teradata Corporation.

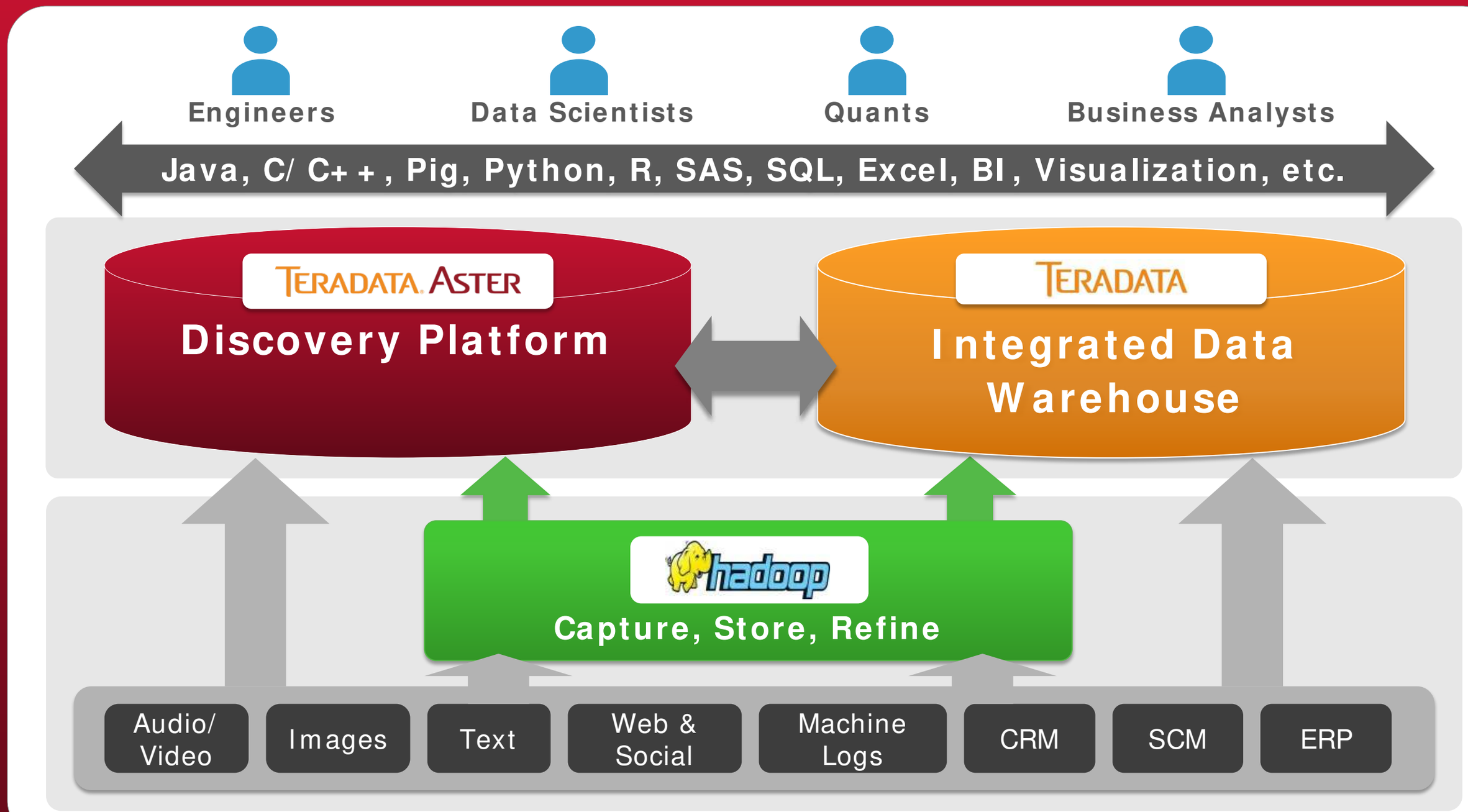
TERADATA. ASTER

The next 12-24 months of Big Data



© Hortonworks Inc. 2012

Unified Big Data Architecture for the Enterprise



Confidential and proprietary. Copyright © 2012 Teradata Corporation.

Thank You! … Questions?

- **Resources:**
  - Download today's presentation
  - Download a new white paper, "Harnessing the Value of Big Data Analytics"
  - MapReduce Resource Center
  - Aster Express downloadable

Confidential and proprietary. Copyright © 2012 Teradata Corporation.