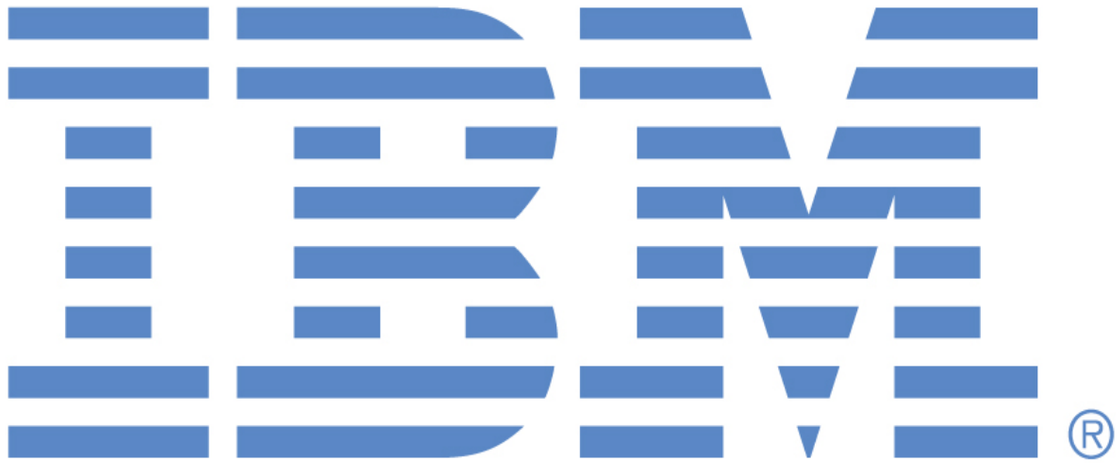
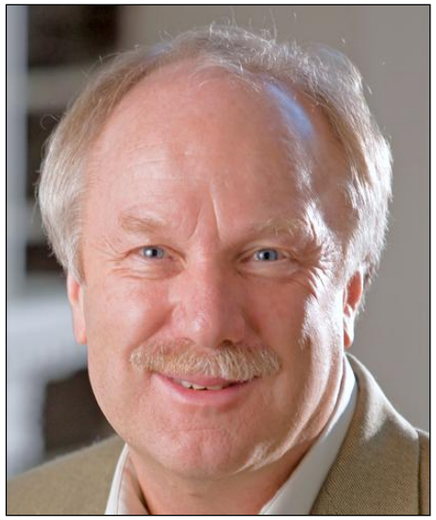
Managing Data Warehouse Growth in the New Era of Big Data

Colin White
President, BI Research
December 5, 2012

Sponsor



Speakers



Colin White
President,
BI Research

Vineet Goel
Product Manager,
IBM InfoSphere Optim

Managing Data Warehouse Growth in the New Era of Big Data

Colin White
President, BI Research
TDWI/IBM Webinar, December 2012

The Evolution of Digital Data

First OLTP systems

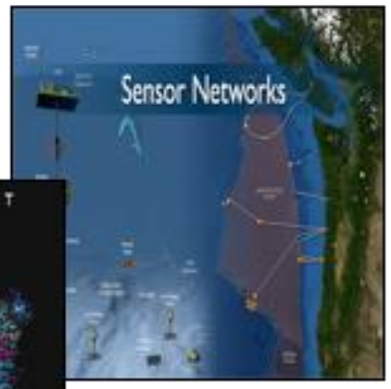
Early decision support products

First commercial RDBMSs

Early data warehousing

Big data & advanced analytics

1960 — 1970 — 1980 — 1990 — 2012

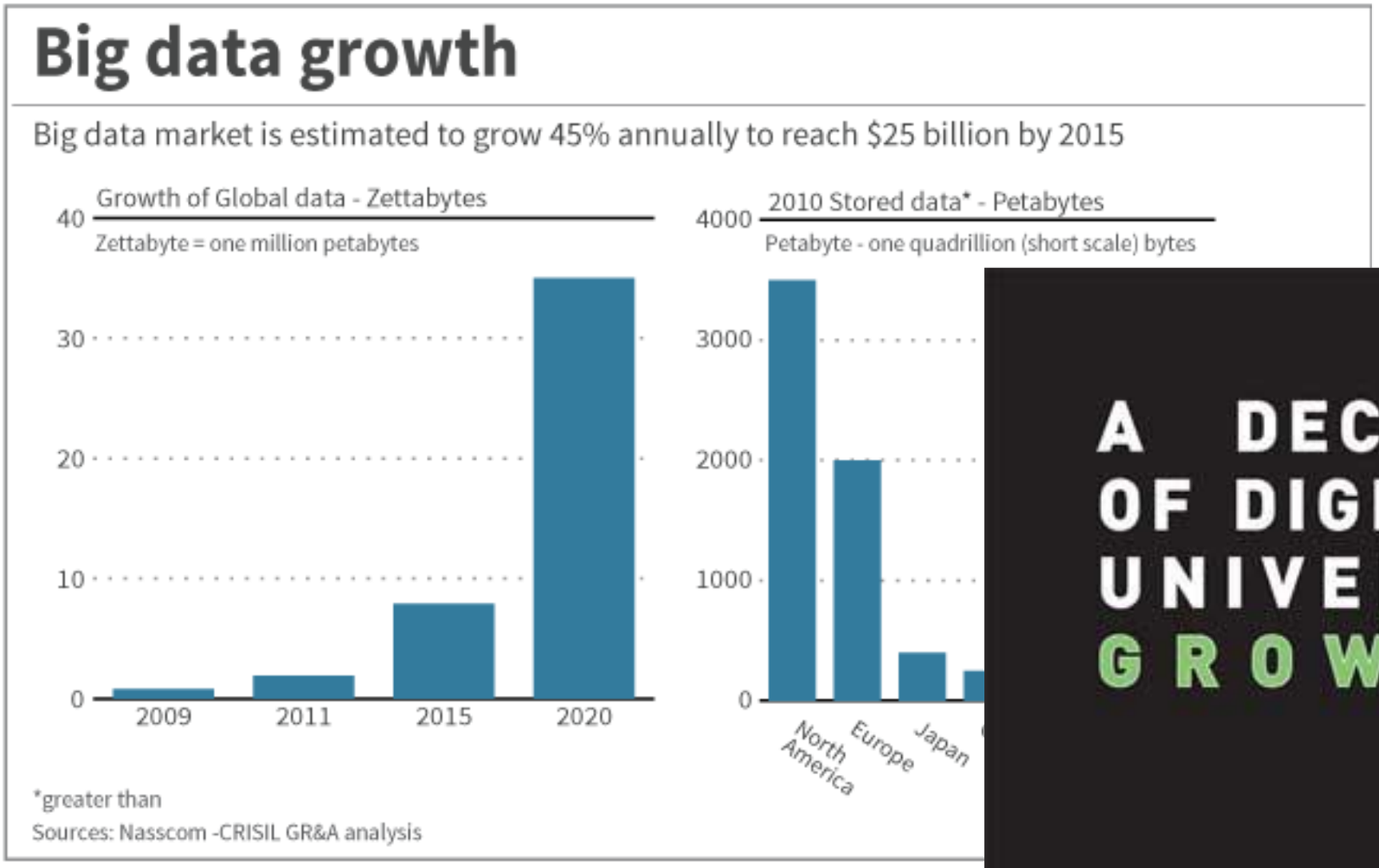


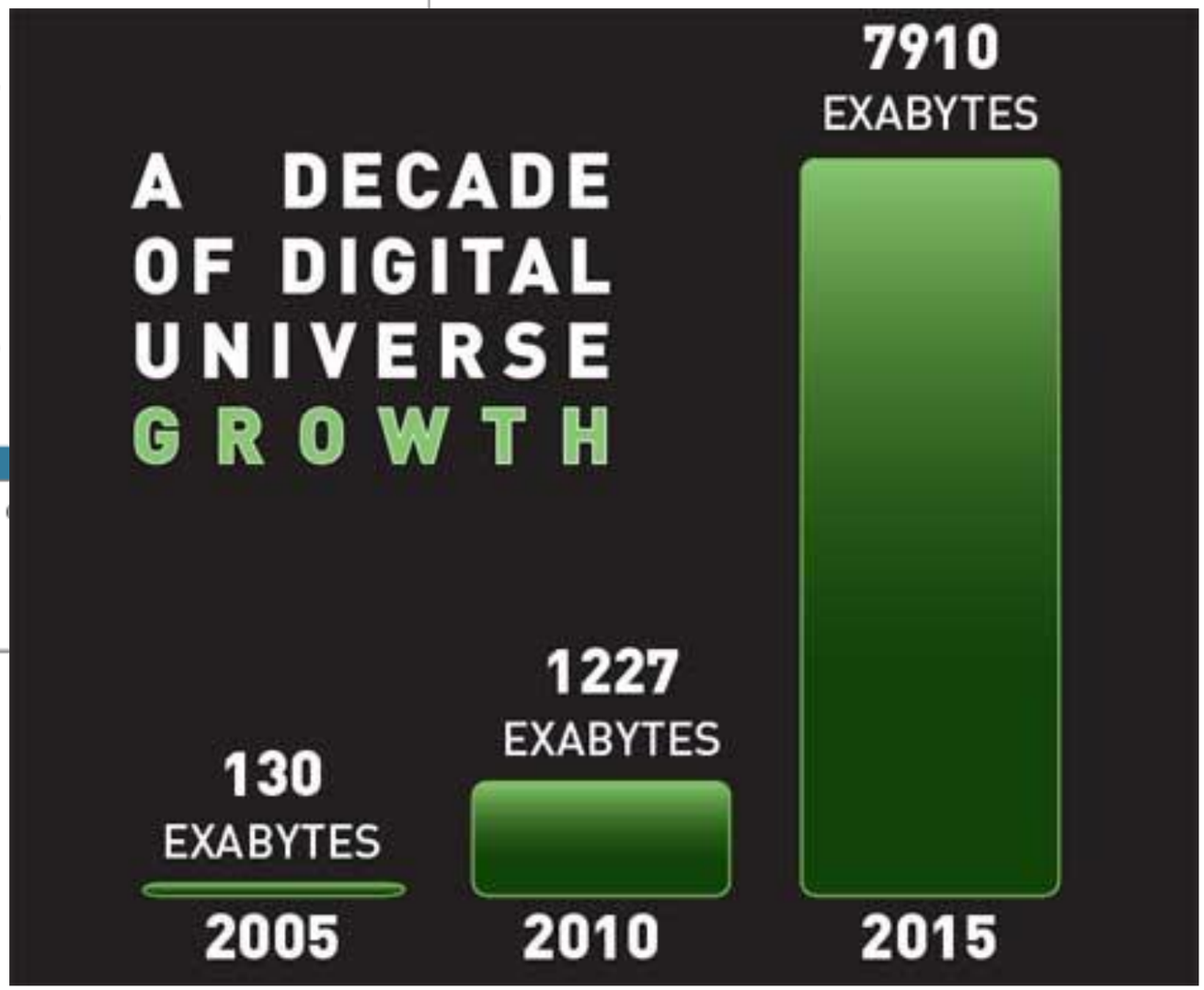
Sabre
84,000 txs/day

Increasing Data Volumes

Sabre
60,000 txs/sec
25 TB EDW

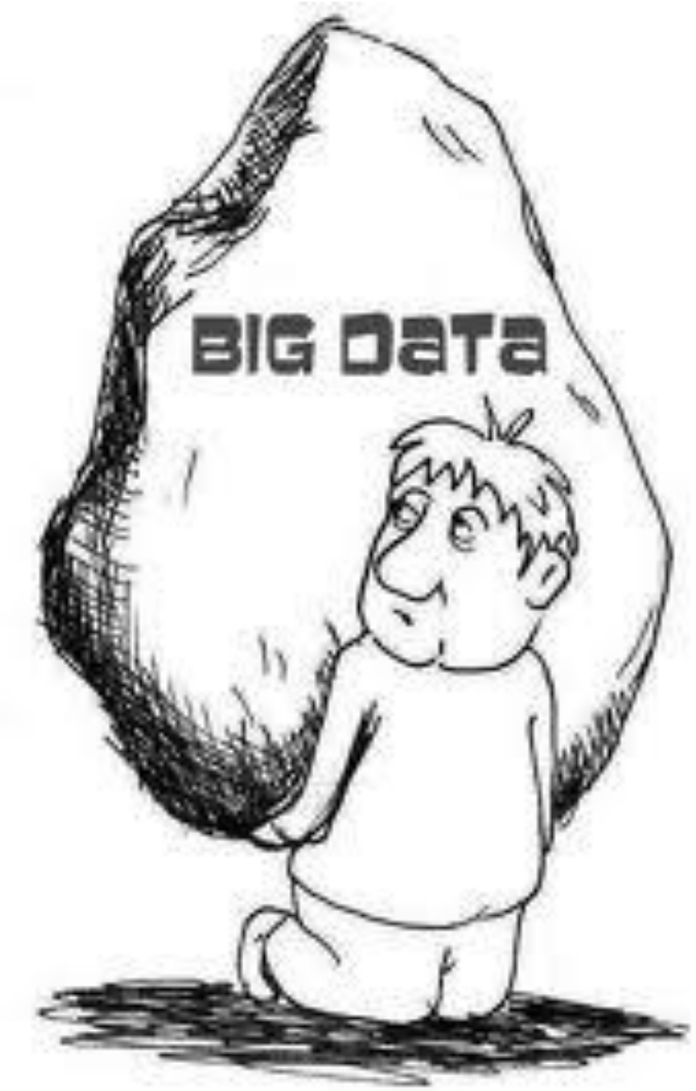
Data Growth – Choose an Analyst!





Source: IDC

Data Growth and Big Data



Big data technologies apply to all types of digital data not just multi-structured data

"Big" is a relative term and is different for each organization and application

What you do with big data and how you use it for business benefit should be the main consideration – business analytics play a key role here

The Value of Data: IBM 2012 Study

IBM Center for Applied Insights

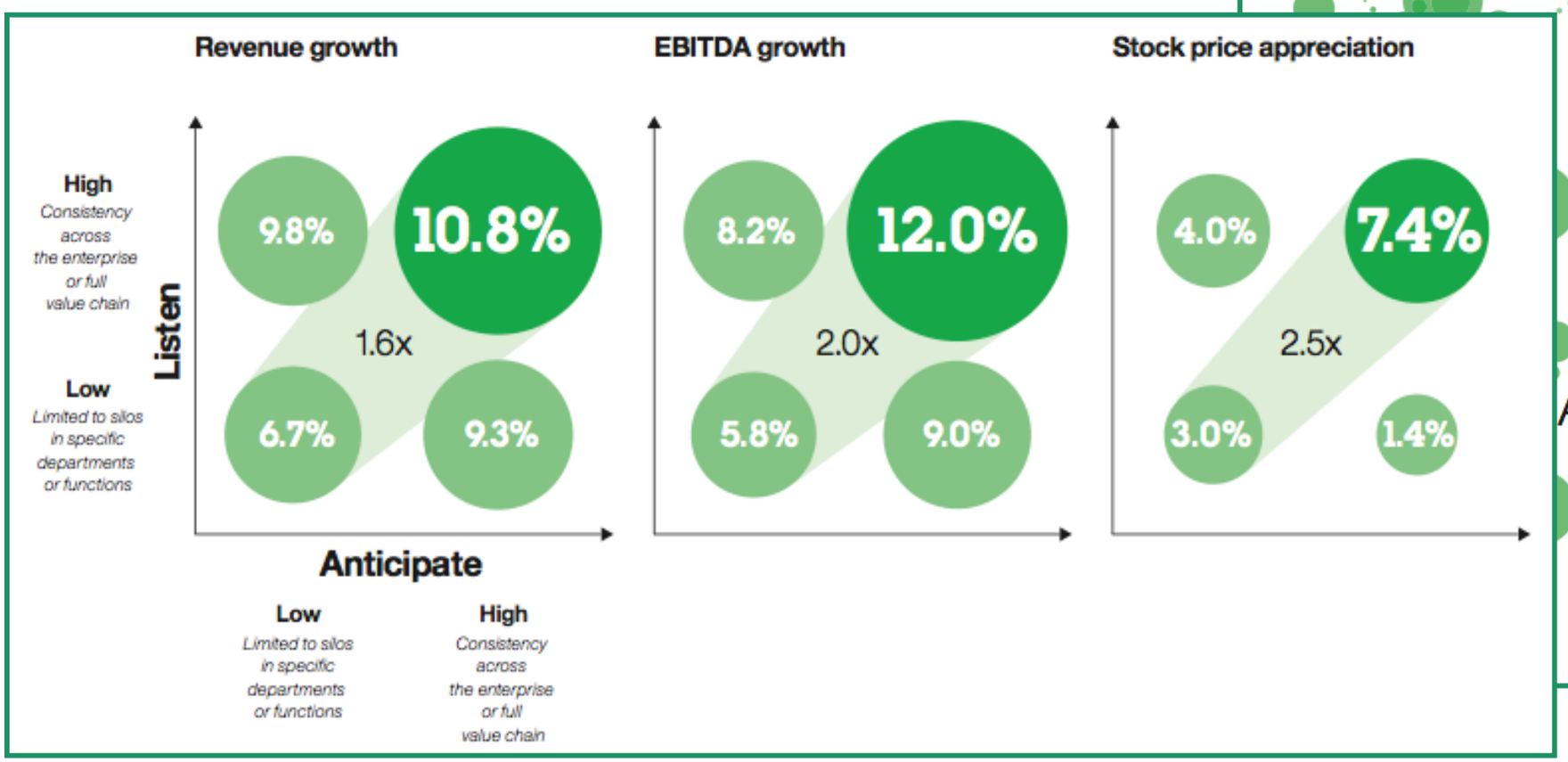
Outperforming in a data-rich, hyper-connected world

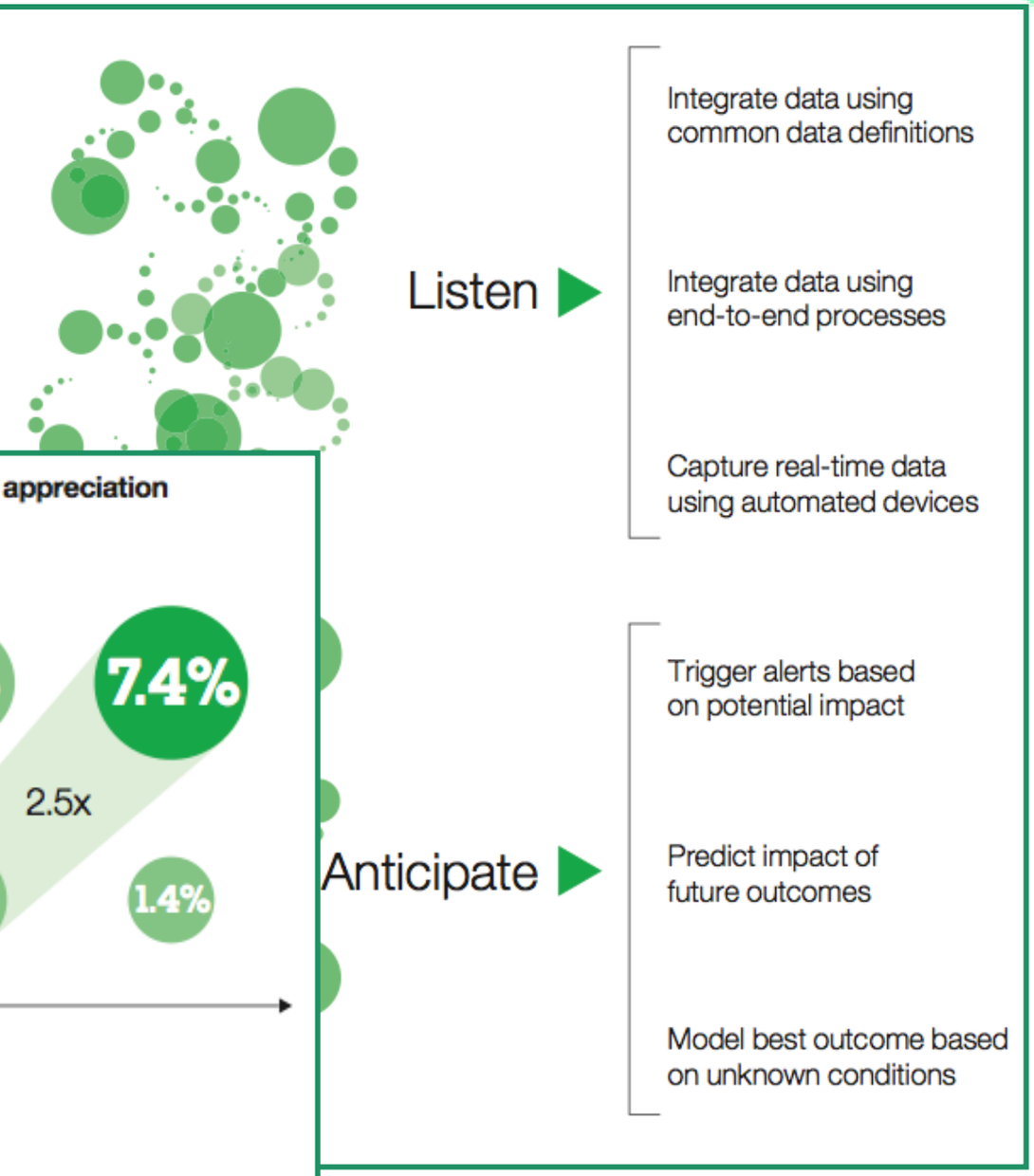
We live in the era of "big data" and connectivity. Collectively, our planet generates **fifteen petabytes** of new information every day[1] – about eight times the information housed in all the academic libraries in the United States.[2] People are connecting and communicating more than at any time in human history. In 2011 alone, an estimated **seven trillion** text messages were sent.[3] There are almost **six billion** mobile phone subscriptions.[4] Today, **two billion** people worldwide are plugged into the Internet,[5] interacting and sharing information on sites like Facebook and Twitter.

About the Study

The IBM Center for Applied Insights, in cooperation with the Economist Intelligence Unit and the IBM Institute of Business Value, recently surveyed 1,168 executives (two thirds of which were CxOs) across nine industries in 64 countries. We investigated their ability to capture data, share insights and take actions based on what they learn, and used a binary logistic regression analysis to understand the statistically significant correlations with financial performance.

The Value of Data: IBM 2012 Study





The Changing World of Business Analytics

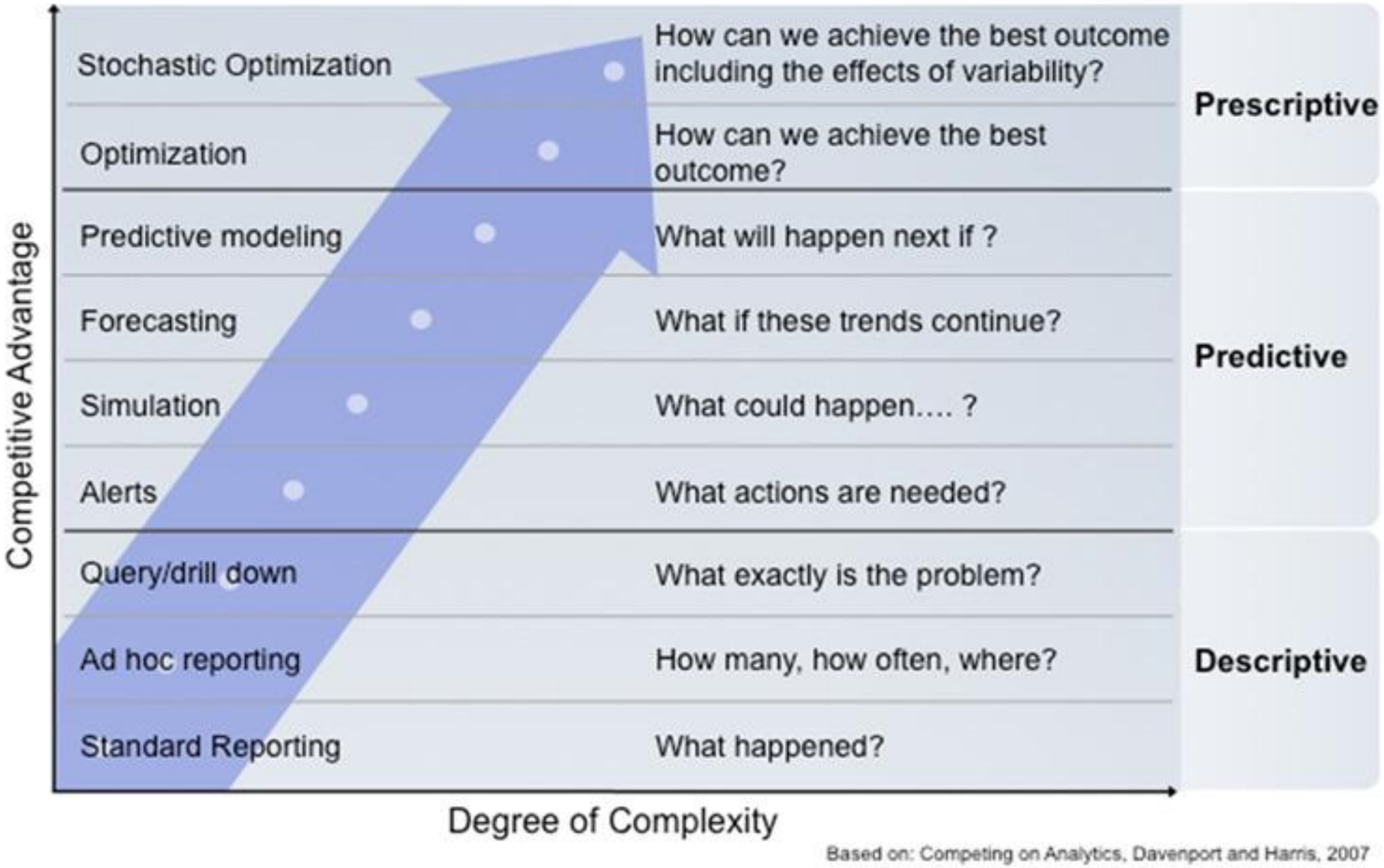
Advanced Analytics

- Improved analytic tools and techniques for statistical and predictive analytics
- New tools for exploring and visualizing new varieties of data
- Operational intelligence with embedded BI services and BI automation

Big Data Management

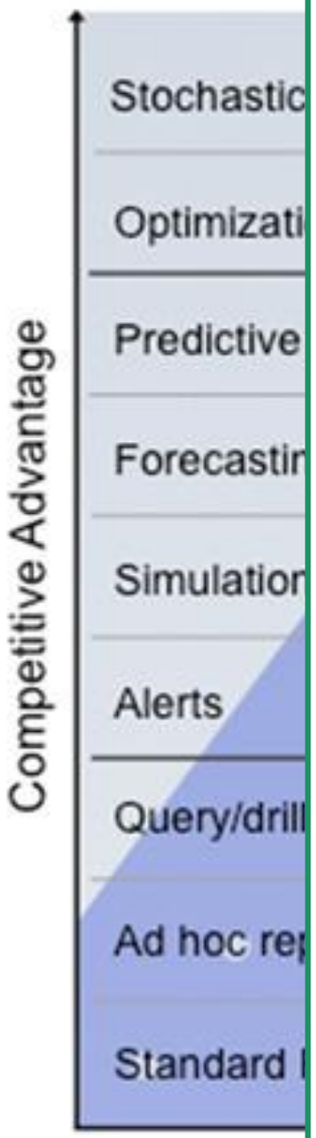
- Analytic relational database systems that offer improved price/performance and libraries of analytic functions
- Non-relational systems such as Hadoop for handling new types of data
- Stream processing/CEP systems for analyzing in-motion data
- In-memory computing for high performance

Advanced Analytics Example: SC Digest 2012



Advanced Analytics in Supply Chain, Dr. Michael Watson, Supply Chain Digest, November 2012

Advanced Analytics Example: SC Digest 2012



1. **Descriptive analytics** – using historical data to describe the business. This is usually associated with Business Intelligence (BI) or visibility systems. In supply chain, you use descriptive analytics to better understand your historical demand patterns, to understand how product flows through your supply chain, and to understand when a shipment might be late.
2. **Predictive analytics** – using data to predict trends and patterns. This is commonly associated with statistics. In the supply chain, you use predictive analytics to forecast future demand or to forecast the price of fuel.
3. **Prescriptive analytics** – using data to suggest the optimal solution. This is commonly associated with optimization. In the supply chain, you use prescriptive analytics to set your inventory levels, schedule your plants, or route your trucks.

Advanced Analytics in Supply Chain, Dr. Michael Watson, Supply Chain Digest, November 2012

What Then is Big Data?

Represents ***analytic and data management solutions*** that could not previously be supported because of:

- Technology limitations – poor performance, inadequate analytic capabilities, etc.
- High hardware and software costs
- Incomplete or limited data for generating the required solutions

Set of overlapping technologies that enable customers to deploy analytic systems *optimized* to suite specific business needs and workloads

Optimization may involve improving performance, reducing costs, enabling new types of data to be analyzed, etc.

Big Data and Data Life Cycle Management

Data Management and Analytic Performance

- Capacity planning
- Managing data warehouse growth
- Analytic performance management and optimization
- Service level agreements

Data Governance

- Security: user access, encryption, masking, etc.
- Quality: governed/ungoverned data
- Backup and recovery
- Archiving and retention: historical analysis, compliance

The Impact of Big Data on the Data Life Cycle

Need fast time to value to quickly gain business benefits from big data

- Impractical to use traditional EDW approach for all analytic solutions
- Extend existing data warehousing environment to support big data and accommodate data growth

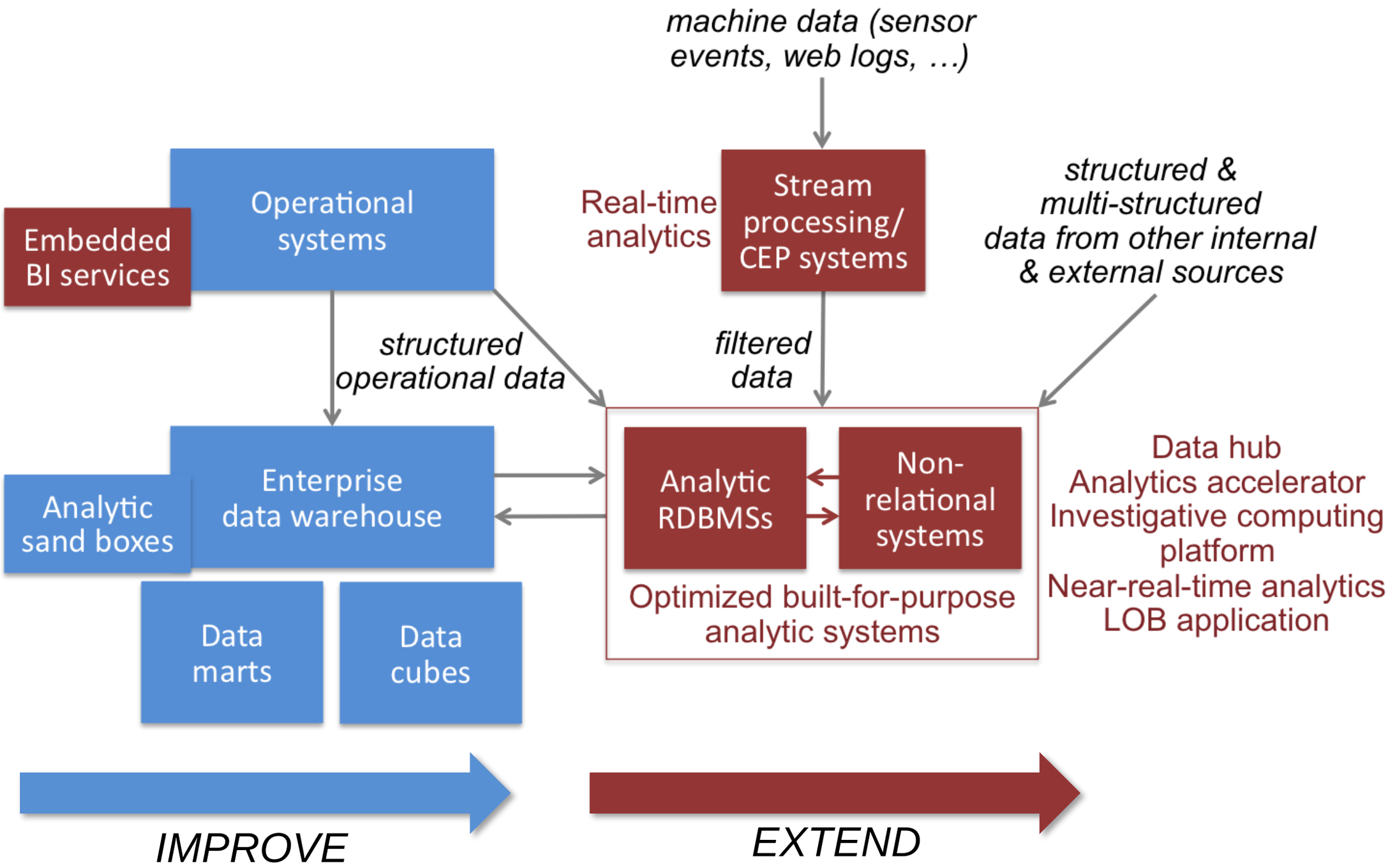
Need high performance solutions for supporting big data analytic workloads

- One-size fits all data management is no longer viable
- Match technologies and costs to business needs and analytic workloads

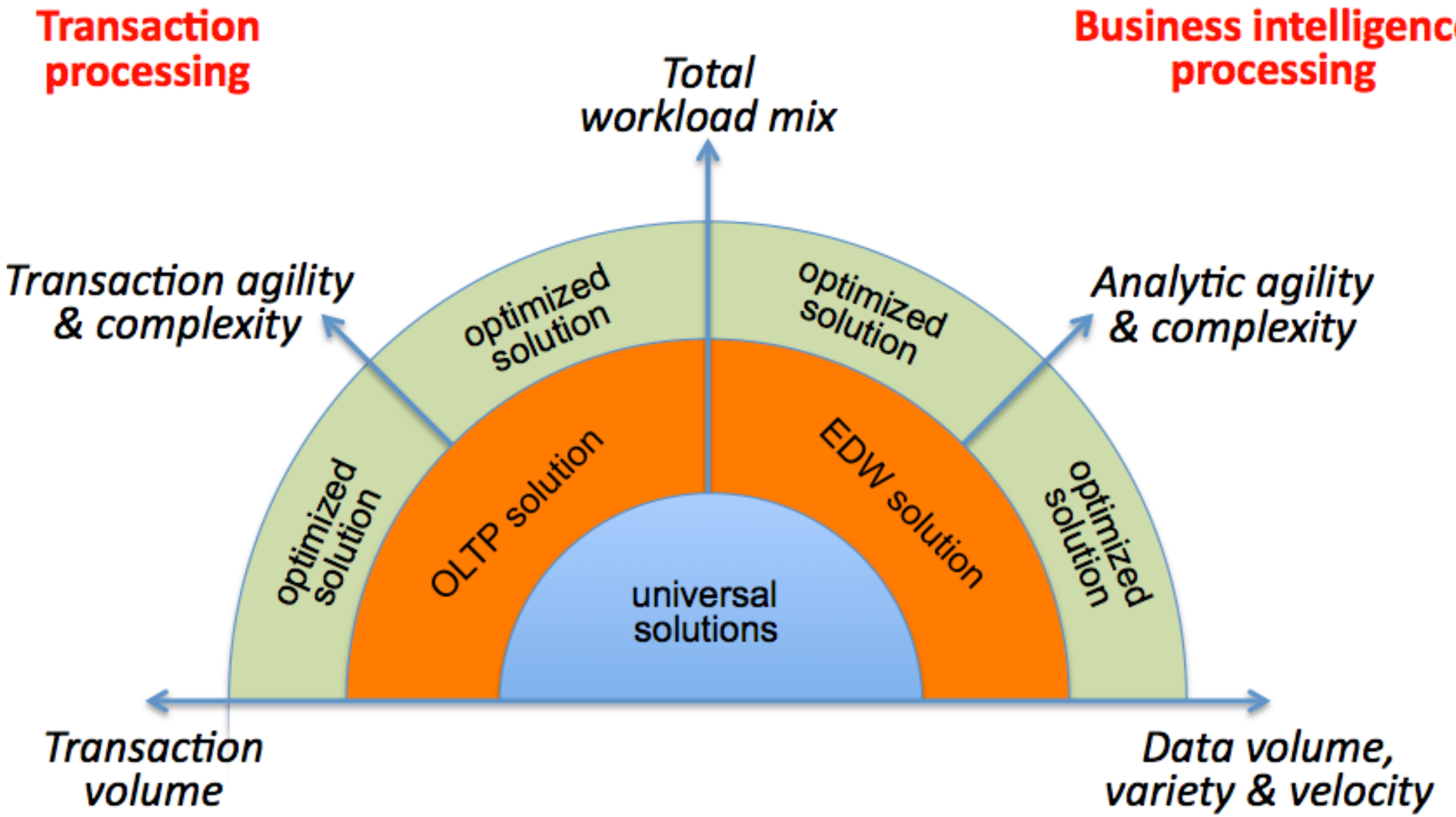
Need improved data governance to handle big data

- No longer practical to rigidly control and govern all forms of data – implement different levels of governance based on security, compliance and quality needs
- Determine data archiving and policies based on the possible future need to analyze historical data and data compliance requirements

The New Extended Data Warehouse



The Quest for Performance: Optimized Systems



Optimized Systems: Hardware Directions

Faster processors

Multi-core processors

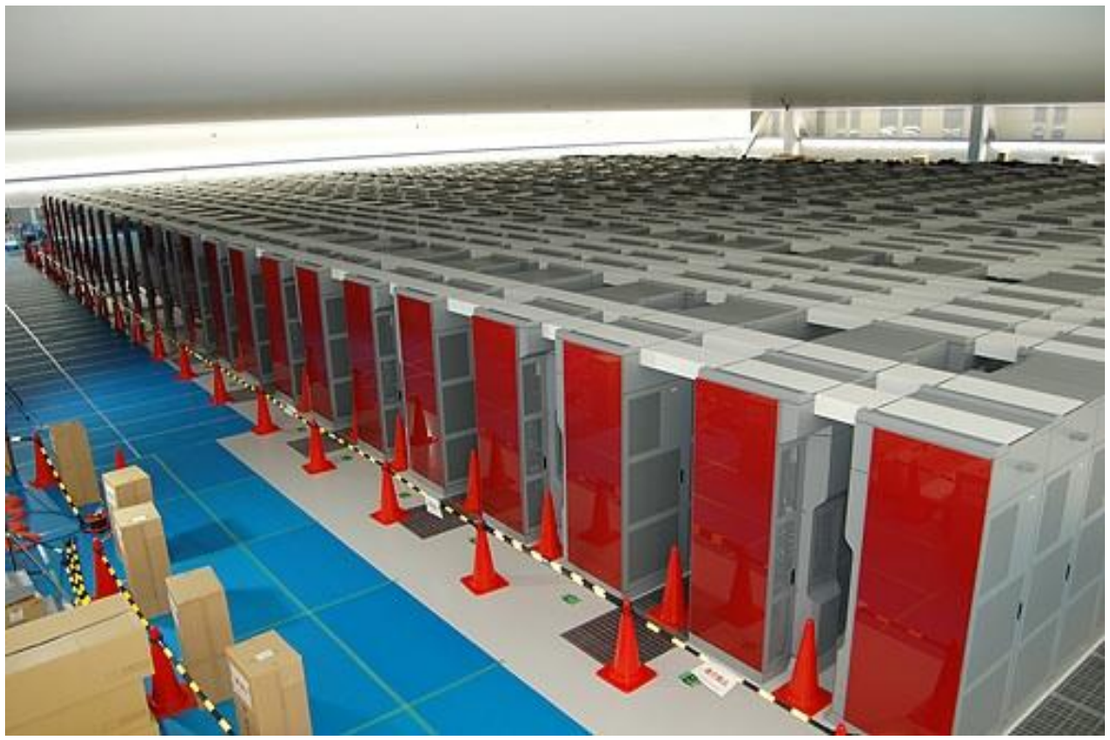
Intelligent hardware

64-bit memory spaces

Large-capacity disk drives

Fast hard-disk and solid-state drives

Hybrid storage configurations



Scale-up/out parallel processing configurations

Lower-cost hardware (blades, clusters)

Reduced power and cooling requirements

Packaged hardware/software appliances

Managing Data Warehouse Growth: Storage Options

Large-capacity hard-disk drives (HDD)

- More economical, less reliable, slower performance, e.g., SATA drives in white-box H/W
- Often “short-stroked” to improve performance

High-performance hard-disk drives (HDD)

- More expensive, more reliable, better performance, e.g., enterprise SAS drives

Solid-state drives (SSDs)

- High and consistent performance
- Better reliability and more energy efficient
- Distinguish between commodity SSDs and enterprise SSDs

Dynamic RAM (DRAM)

- Best performance - eliminates I/O overheads
- Use by in-memory computing systems

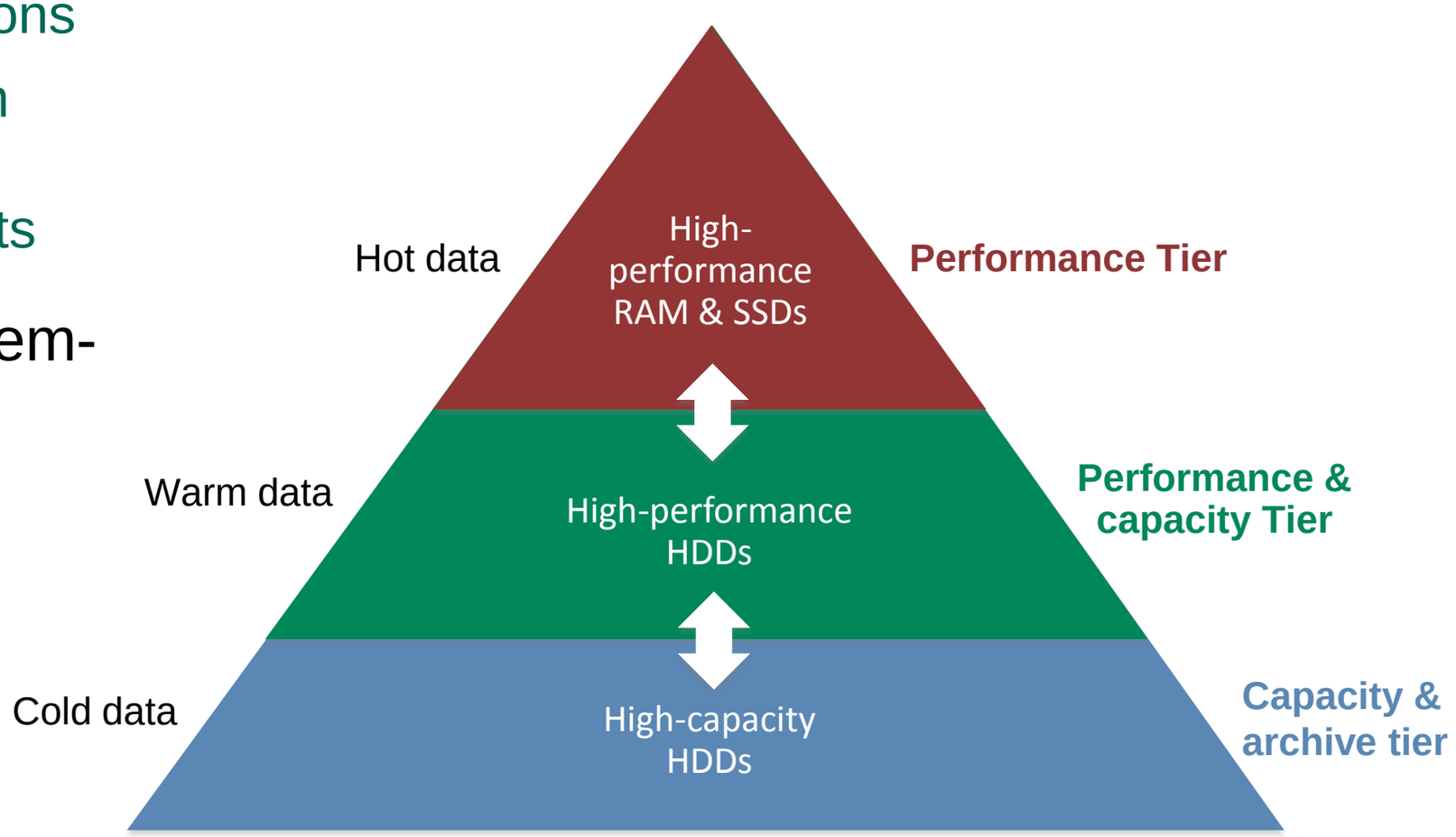
Managing Data Warehouse Growth: Tiered Storage

Hybrid (tiered) storage

- Employs a combination of different storage devices and options
- Data location is based on availability, usage, and performance requirements

Some DBMSs support system-enabled data migration

- Table/partition vs block-level migration
- Administrator-driven vs automatic migration
- Consider relationship to workload management and data archiving & retention strategies



Managing DW Growth: Archiving/Retention - 1

The price/performance of new software and hardware technologies enables more data to be kept online longer

Data warehouse data is likely to be managed on multiple systems for performance, cost, or business reasons

Policies are required to manage the archiving (for possible future analysis) and retention (for compliance) of data warehouse data

These policies must be coordinated with corporate strategy especially in the area of compliance

Managing DW Growth: Archiving/Retention - 2

Operational, analytical and collaborative computing are becoming inextricably linked and this should be considered when developing retention policies

Data security, quality, archiving and retention are related, but may need different levels of governance

Data archiving and retention solutions may be home grown, use best-of-breed products, or employ an enterprise-level software approach

Summary

Organizations are experiencing huge data growth

Big data involves all digital data, but “big” is different for each organization and application

Big data is a set of overlapping technologies that can provide huge business benefits but create some significant governance issues

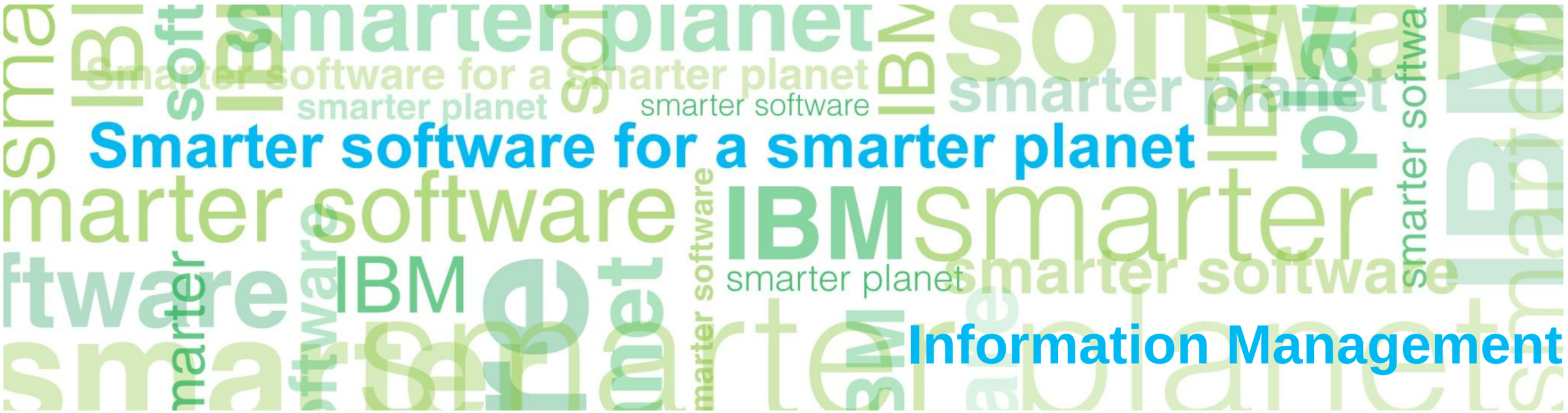
Big data involves the blending of governed and ungoverned data

Impossible to govern all data and organizations need to implement flexible policies that:

- Allow easy but secure access to data
- Provide elastic capacity and good performance
- Efficiently manage data archiving and retention

Managing Data Warehouse Growth

IBM InfoSphere Optim, Information Management



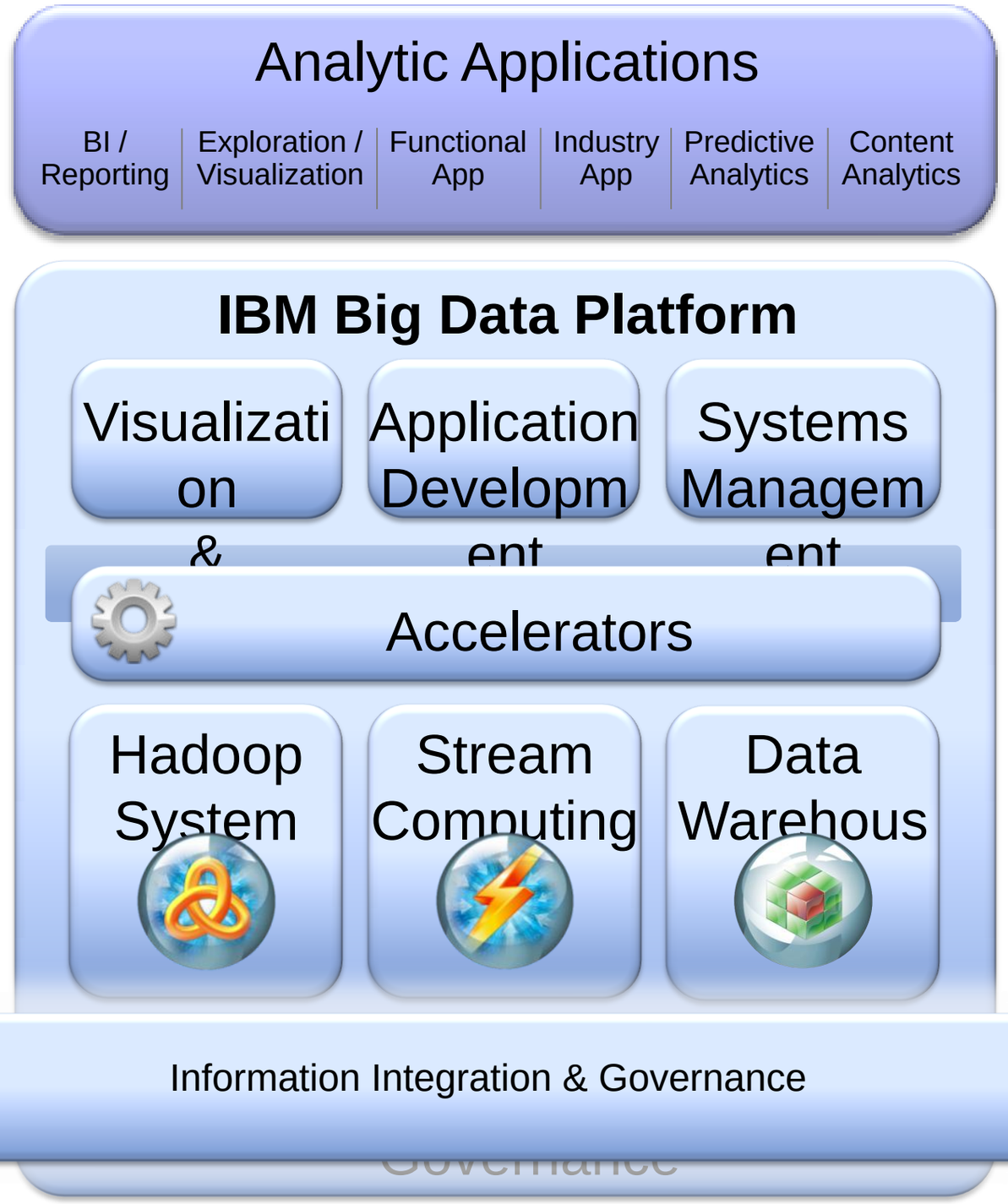
IBM Big Data Strategy

New analytic applications drive the requirements for a **big data platform**

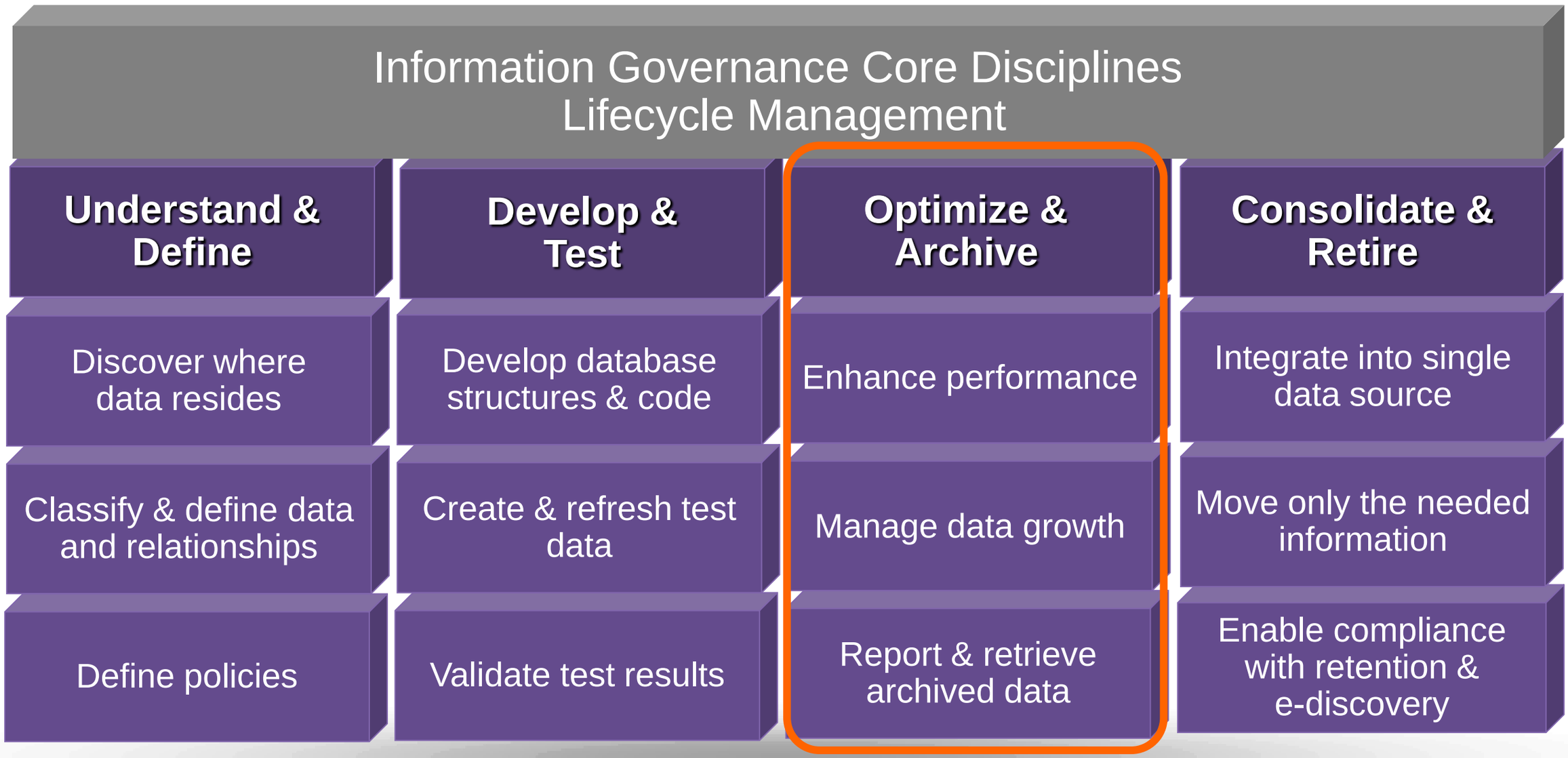
- Integrate and manage the full variety, velocity and volume of data
- Apply advanced analytics to information in its native form
- Visualize all available data for ad-hoc analysis

Need for trusted data drives requirements for an information integration and governance platform

- Ensure the highest quality information
- Master data into a single view
- Secure and Protect sensitive data to minimize risk and exposure
- Govern data throughout its lifecycle



Governing Data Throughout its Lifecycle



Data Warehouse Challenges Related to Data Growth

Increasing Costs and Time to Market

The impact of exponential data growth on infrastructure and operational costs to keep up with data volumes.

Poor Data Warehouse Performance

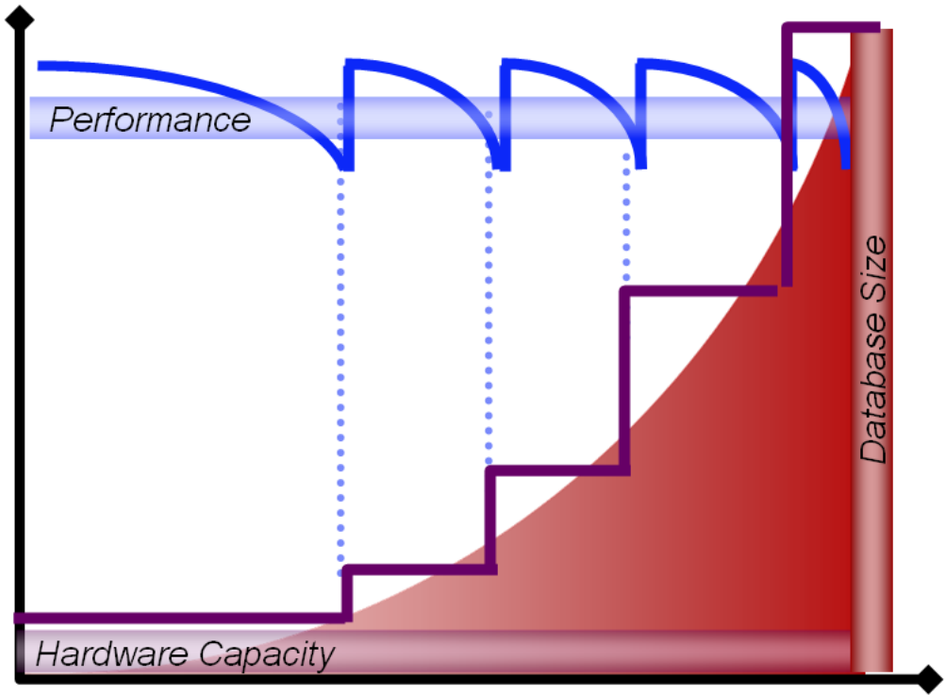
Slow-performing business intelligence (BI) & analytics solutions due to unchecked data growth.

Managing Risk & Compliance

The “keep everything” strategy impacts supporting the complex data retention and disposal compliance.

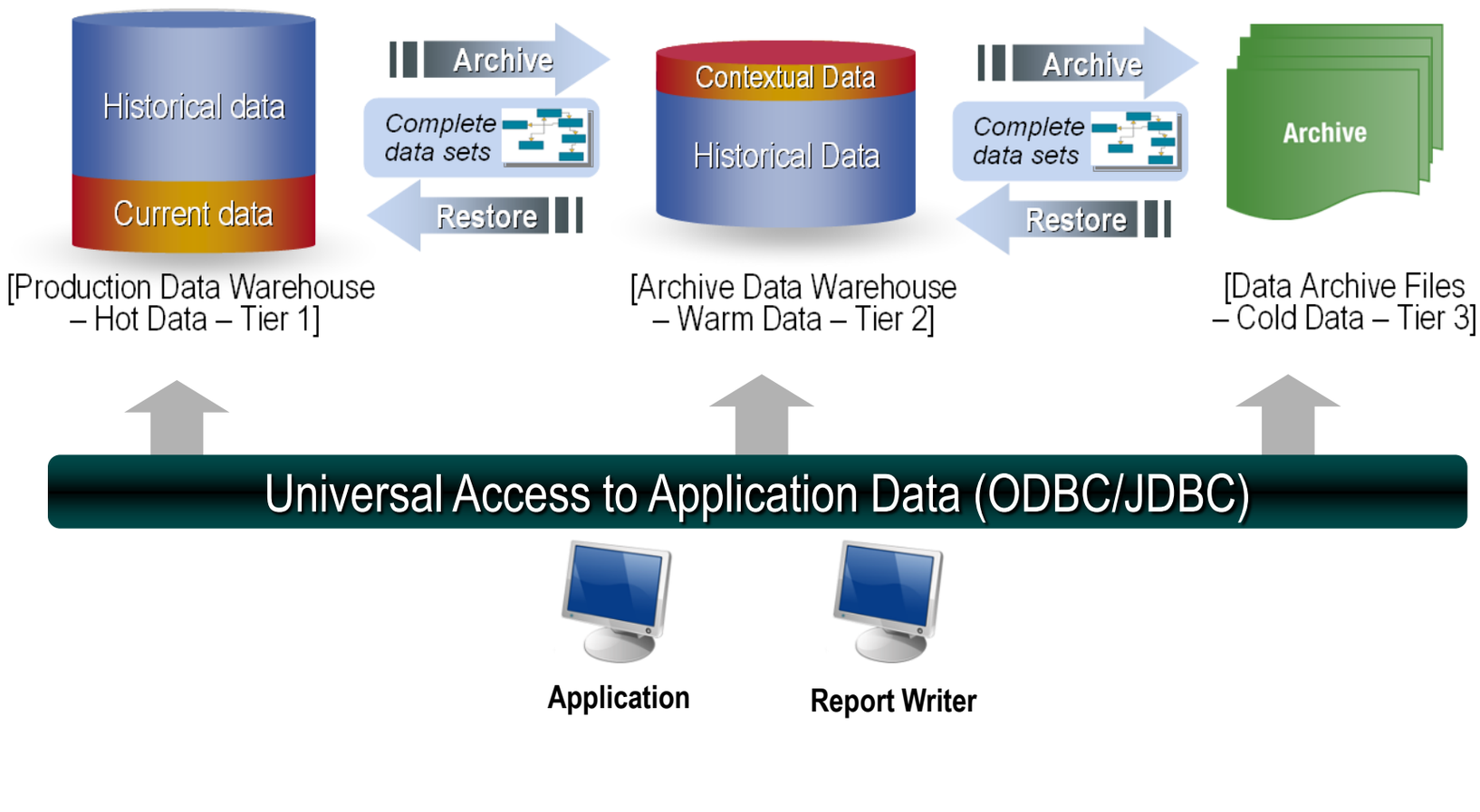
Technical Challenges with High Volume Data Growth

- No policies and processes to manage multi-temperature data cost-effectively
- Transaction & data volume growth drive up infrastructure costs
- Multiple instances of production data – backup, DR, training, test – can compound data growth
- Large production instances affect cost & operations of backups, DR, non-prod
- Archival processes are complex requiring full business context for e-discovery



What is Data Archiving?

Manage data growth and improve performance by intelligently archiving historical data



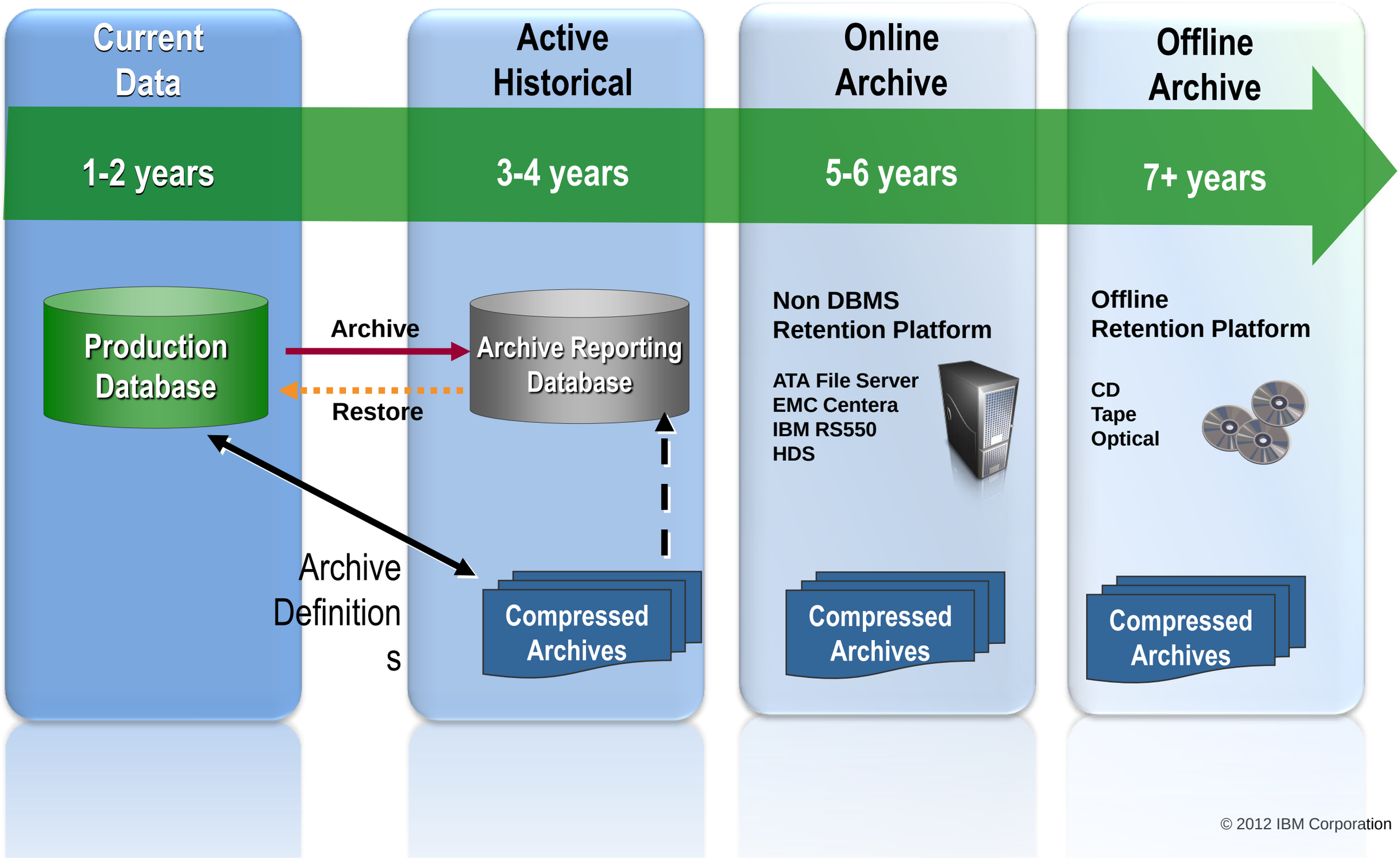
Requirements

- Archive, manage and retain application data according to business policies
- Maintain Referential Integrity
- Support tiered archiving to secondary database or file system

Benefits

- Reduce hardware, storage and maintenance costs
- Improve application performance
- Support data retention or disposal requirements

Tiered Storage for Information Lifecycle Management



Benefits of Archiving for the Data Warehouse

Reduce Costs

Reduce TCO by leveraging a tiered-storage archiving strategy

-Reclaim space & defer hardware upgrades

-Reduce operational and storage costs

Improve Performance

Increase performance by archiving out dormant data.

-Faster analytical processing, ETL and reporting.

-Faster backup, restores and maintenance tasks.

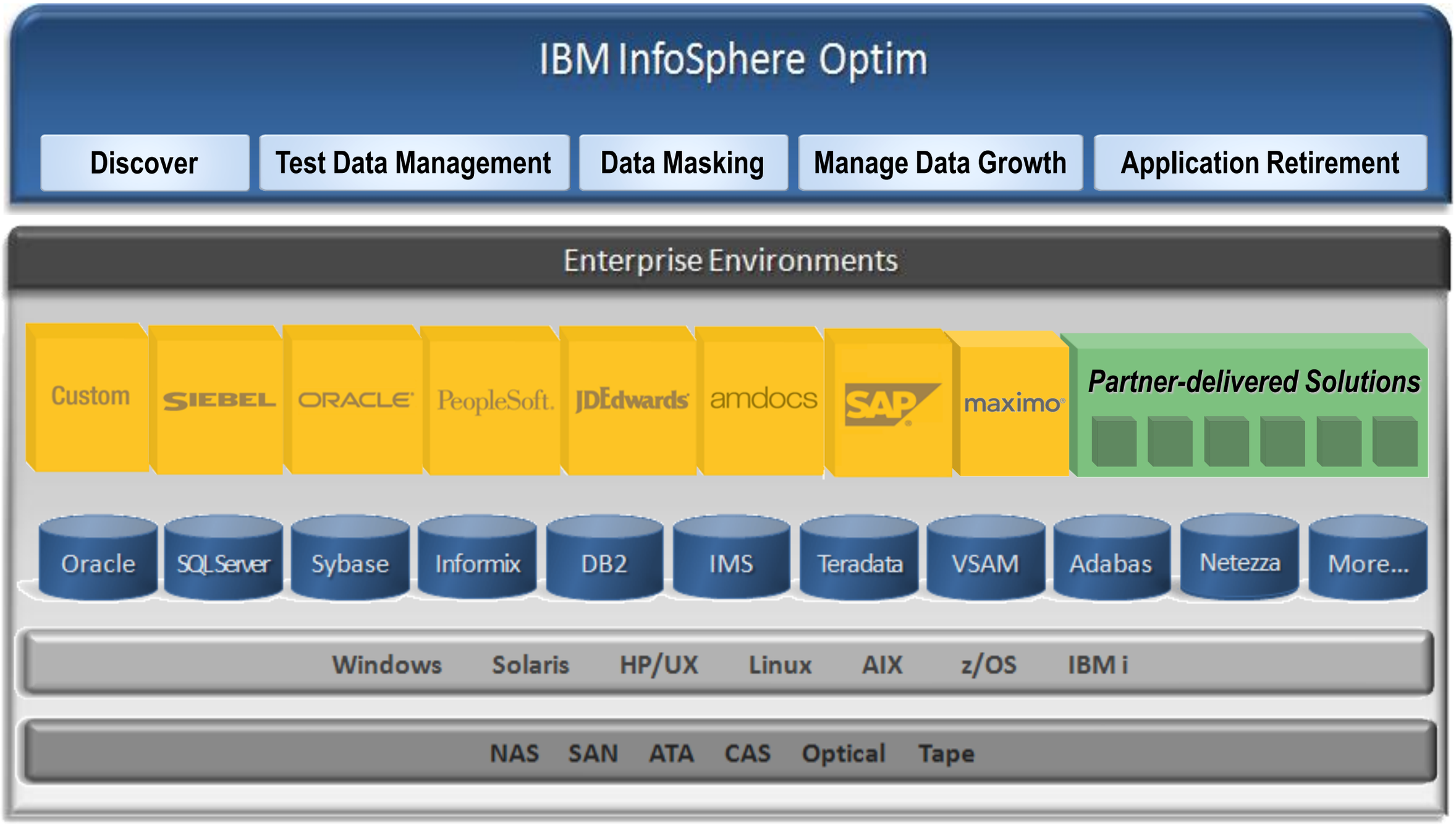
Minimize Risk

Support data retention or legal hold requirement.

-Capture the precise sets of data required based on your audit or e-discovery.

-Support defensible disposal of data

About IBM InfoSphere Optim Solutions

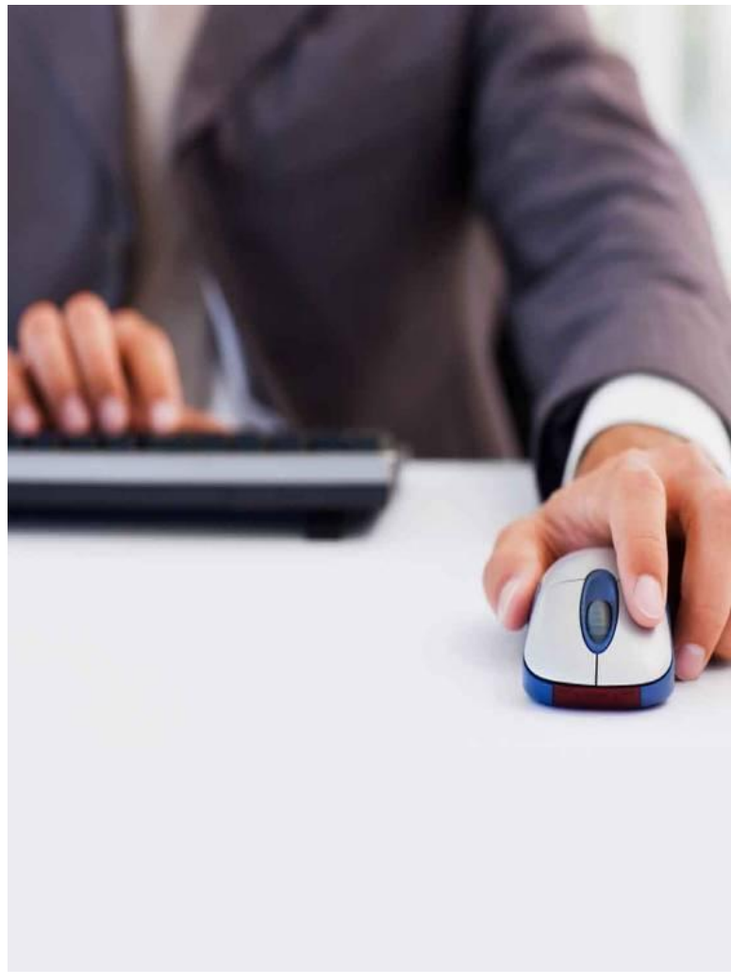


Single, scalable, information lifecycle management solution provides a central point to deploy policies to extract, archive, subset, and protect application data records from creation to deletion

Learn more

- Webpage: InfoSphere Optim
- Solution Sheet: InfoSphere Optim Solutions for the data warehouse
- Demo: InfoSphere Optim archiving
- Whitepaper: Control Application Data Before it Controls Your Business

www.ibm.com/optim



Questions?

Contacting Speakers

- If you have further questions or comments:

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

Vineet Goel
vineetg@us.ibm.com