Enterprise Data Warehouse Optimization

Dr Barry Devlin
Founder & Principal
9sight Consulting

Piet Loubser
VP Product and Solutions Marketing
Hortonworks

The EDW Lives On

The Beating Heart of the Data Lake

10 August 2017

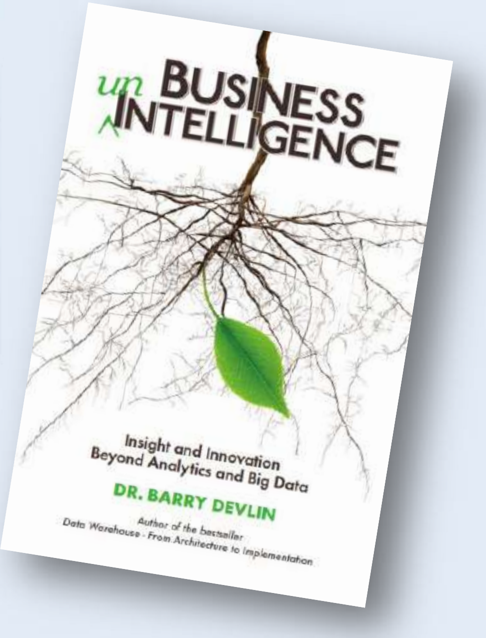
Hortonworks Webinar

Dr Barry Devlin

Founder & Principal
9sight Consulting

Dr. Barry Devlin

Founder and Principal
9sight Consulting, www.9sight.com


unBUSINESS INTELLIGENCE
Insight and Innovation Beyond Analytics and Big Data
DR. BARRY DEVLIN
Author of the bestseller Data Warehouse - From Architecture to Implementation

Dr. Barry Devlin is a founder of the data warehousing industry, defining its first architecture in 1985. A foremost authority on business intelligence (BI), big data and beyond, he is respected worldwide as a visionary and thought-leader in the evolving industry. Barry has authored two ground-breaking books: the classic "Data Warehouse--from Architecture to Implementation" and "Business unIntelligence--Insight and Innovation Beyond Analytics and Big Data" (http://bit.ly/BunI_Book) in 2013.

Barry has over 30 years of experience in the IT industry, previously with IBM, as a consultant, manager and distinguished engineer. As founder and principal of 9sight in 2008, Barry provides strategic consulting and thought-leadership to buyers and vendors of BI and Big Data solutions. He is an associate editor of TDWI's Journal of Business Intelligence, and a regular keynote speaker, teacher and writer on all aspects of information creation and use.

Barry operates worldwide from Cape Town, South Africa.

Email:
barry@9sight.com

Twitter:
@BarryDevlin

Agenda

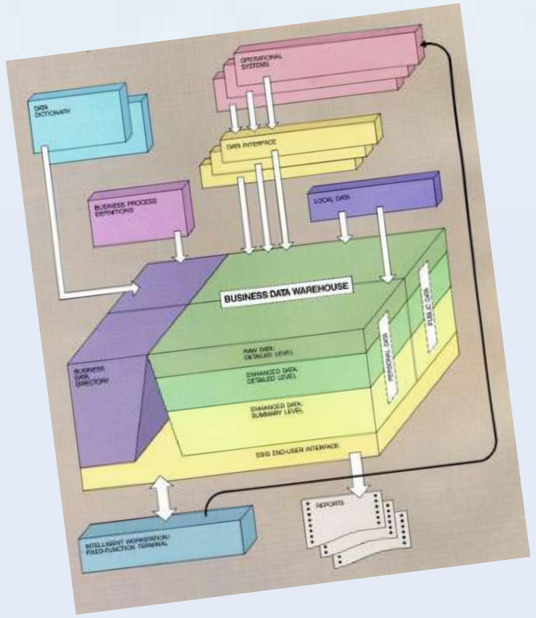
1. Past – from a warehouse to a lake
2. Present – a warehouse *and* a lake
3. Emerging – a warehouse by a lake
4. Conclusions

The data architecture since the mid-’80s

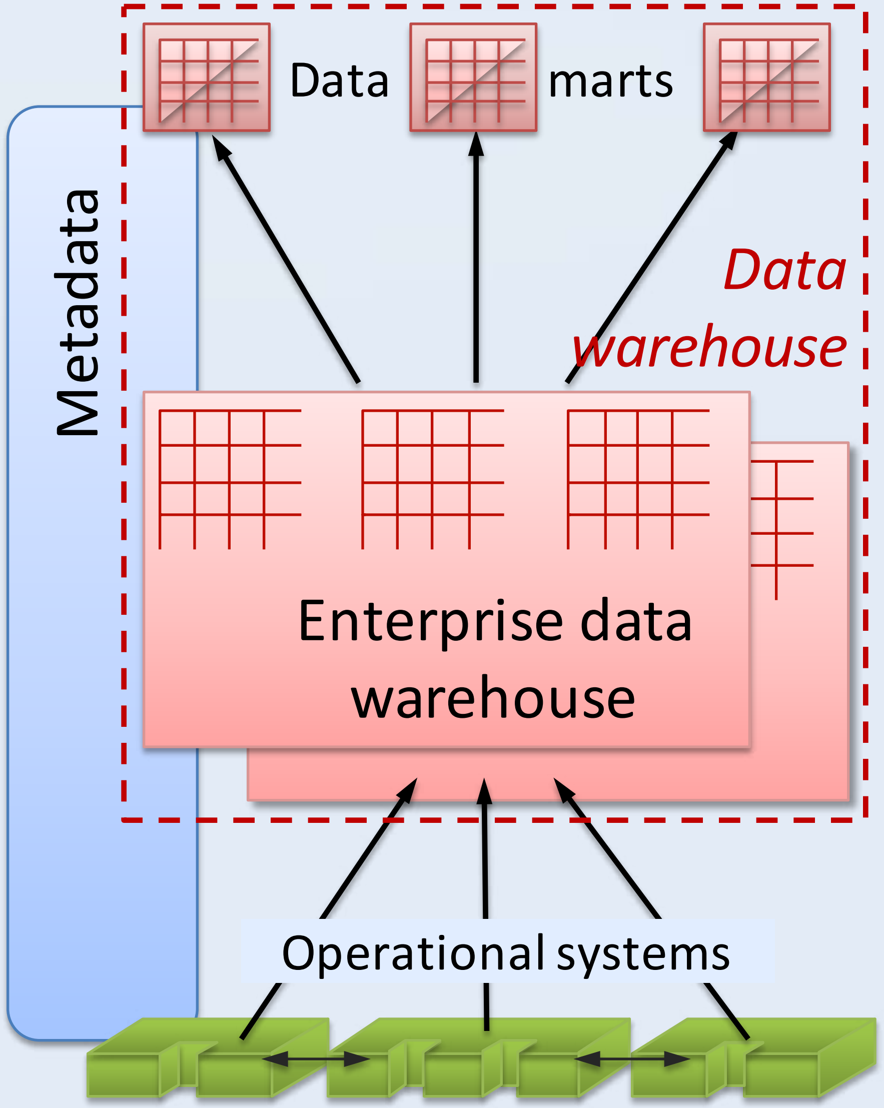
- Two layers within the Data Warehouse…
  - Enterprise data warehouse
    - *Reconciled data*
  - Data marts
    - *What the users need*

- … fed from and separate to operational systems
  - Data to run the business
  - Created by the processes of the business

- All data created within the enterprise (or within partner ecosystem)



“An architecture for a business and information system”,
B. A. Devlin, P. T. Murphy,
IBM Systems Journal, (1988)



The drive toward the data lake since 2010

- Data warehouse architecture “old-fashioned”
  - Linked to (traditional) relational databases
  - Too structured, schema-on-write
  - Too slow / complex to build
  - Lacking support for big data
  - No link to Hadoop

- Data lake proposed as alternative
  - Cheaper, bigger and more flexible
  - Structure-agnostic, schema-on-read (late binding)
  - Supports all data types
  - Agile, flexible, rapid implementation
  - Driven by Hadoop ecosystem
  - Data reservoir – a better(?) architected data lake

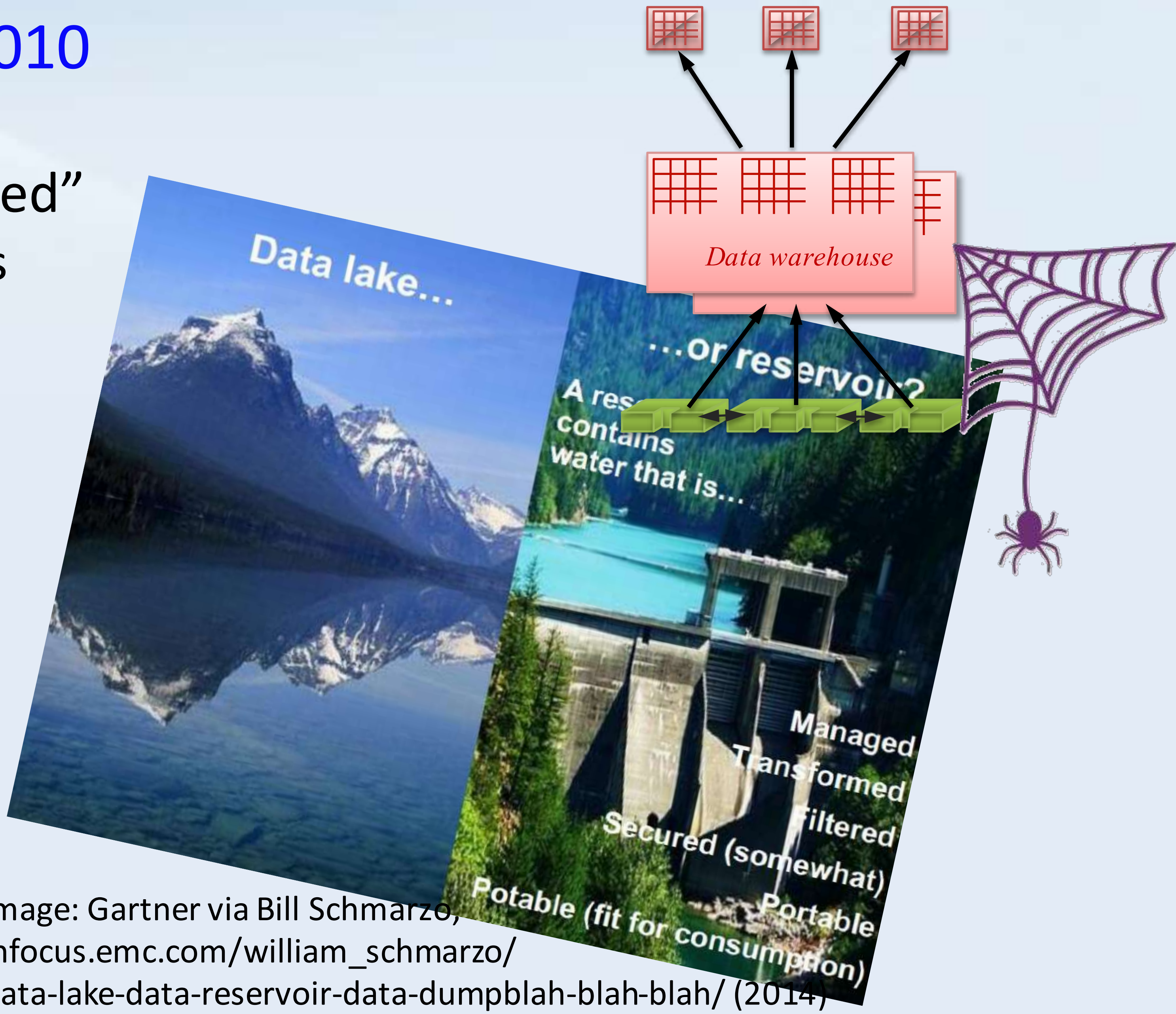
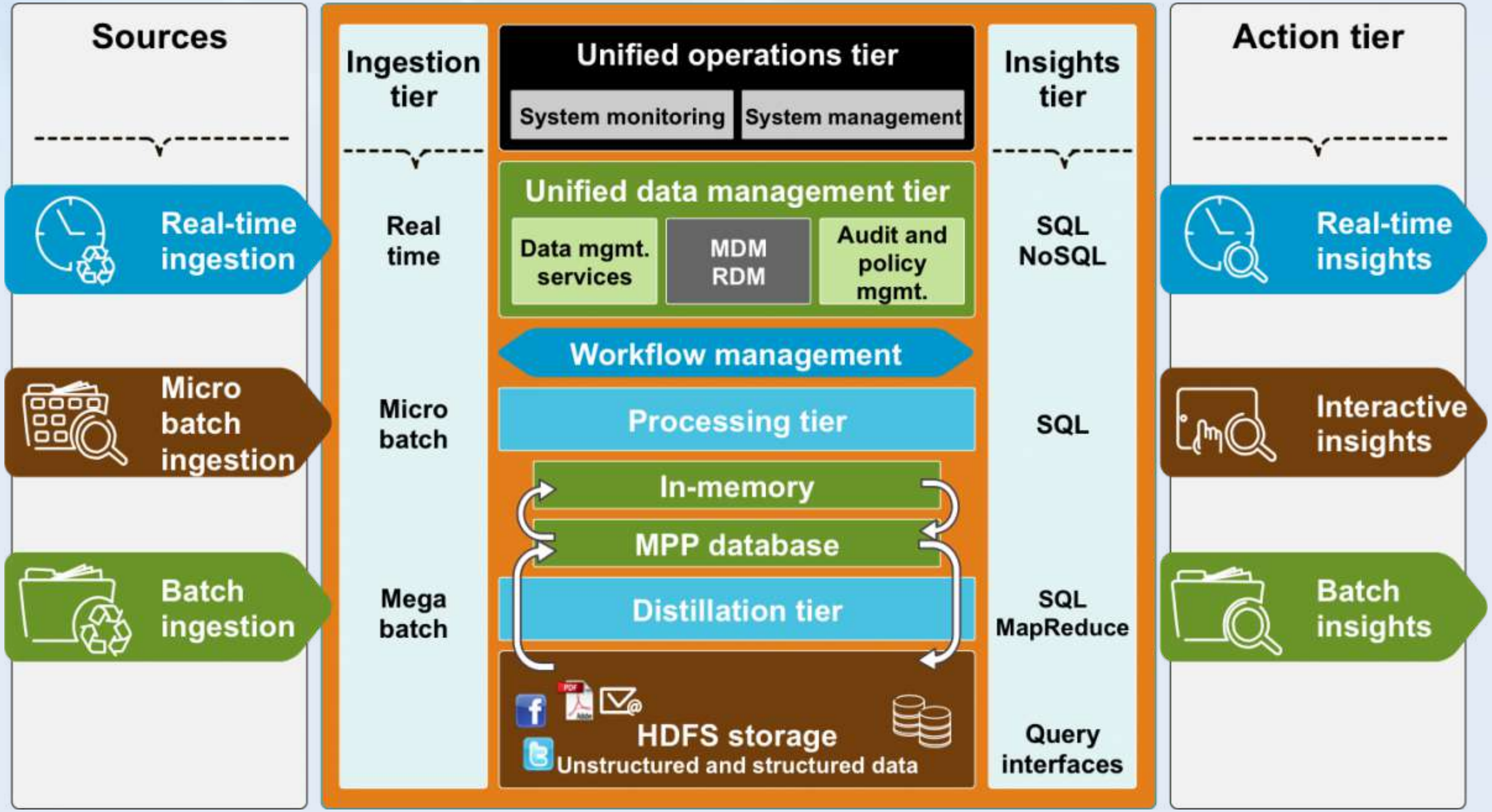


Image: Gartner via Bill Schmarzo, infocus.emc.com/william_schmarzo/ data-lake-data-reservoir-data-dumpblah-blah-blah/ (2014)

Data lake architecture

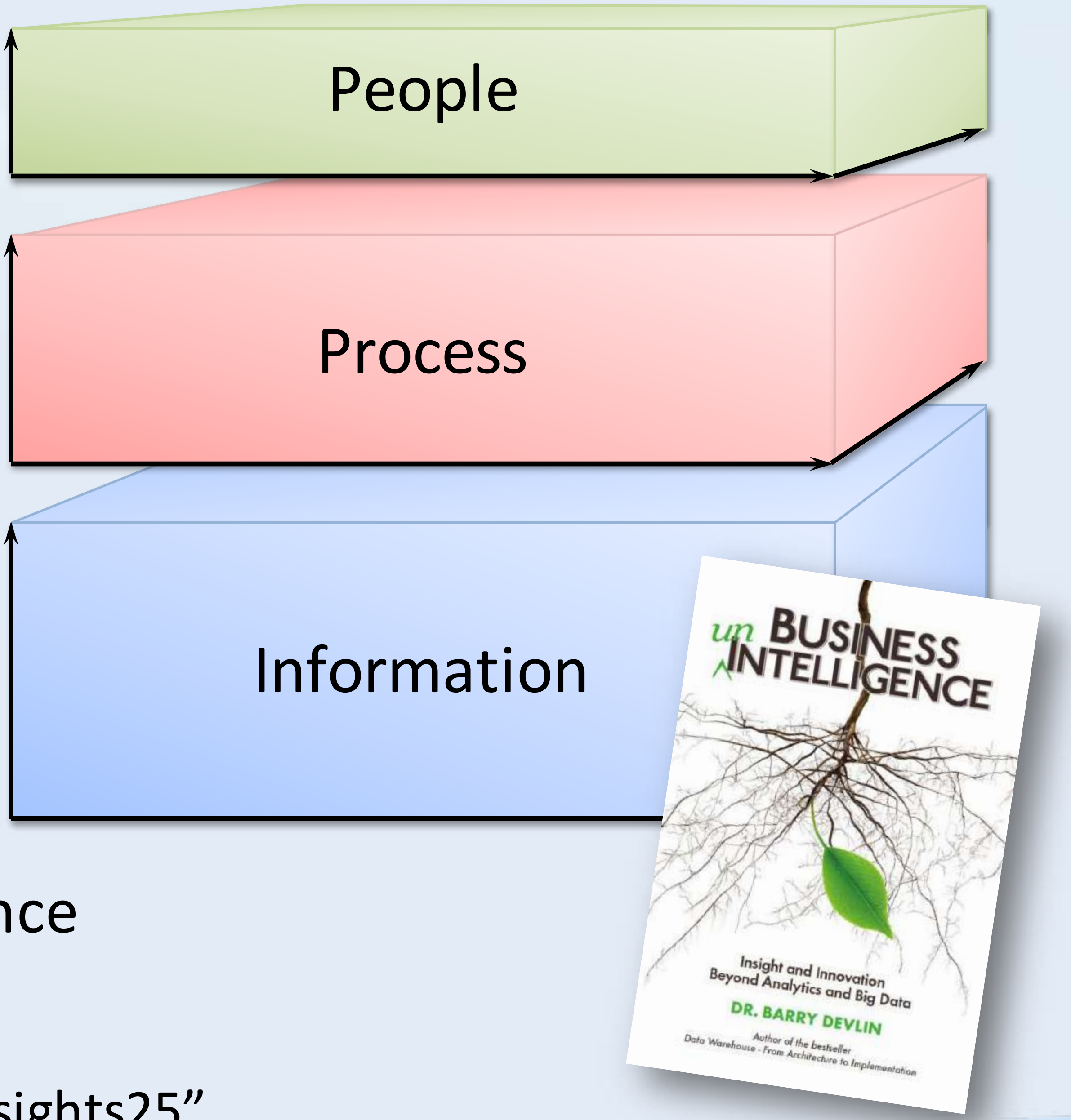


www.capgemini.com/blog/capping-it-off/2014/08/you-have-to-manage-your-data-lake-the-fallacy-of-technology-being-magic

From BI to Business unIntelligence

- People process information

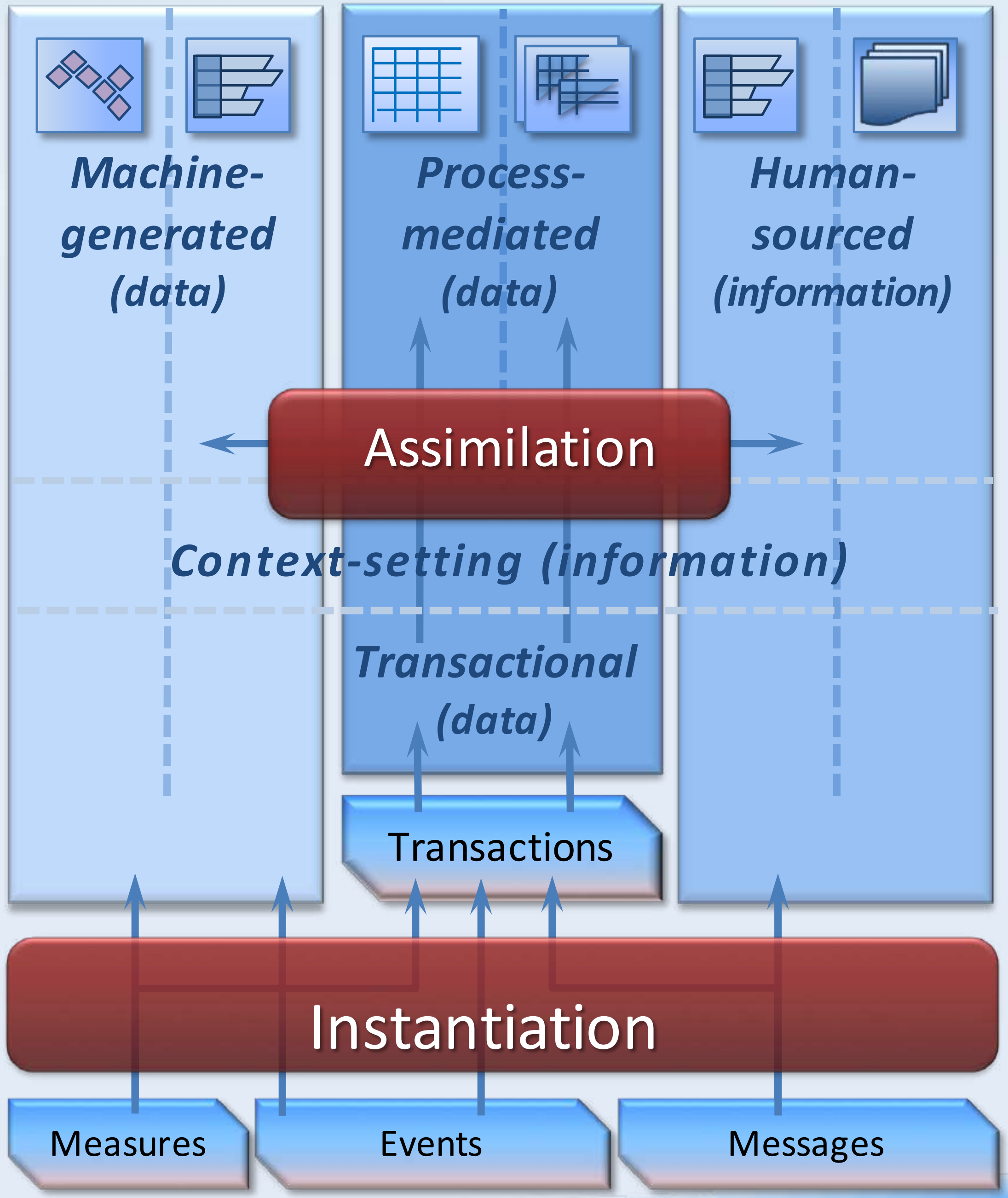
- People: Rational thought and far beyond
  - People make all decisions!
- Process: Logic – predefined, emergent
  - Decision making is a process
- Information: Data, knowledge, meaning
  - Data/information is only the foundation


People
Process
Information
un BUSINESS INTELLIGENCE
Insight and Innovation Beyond Analytics and Big Data
DR. BARRY DEVLIN
Author of the bestseller Data Warehouse - From Architecture to Implementation

- Not business intelligence… Business unIntelligence
- Amazon: http://bit.ly/BunI_Book
- Or http://bit.ly/BunI-TP2: 25% discount with code “BIInsights25”

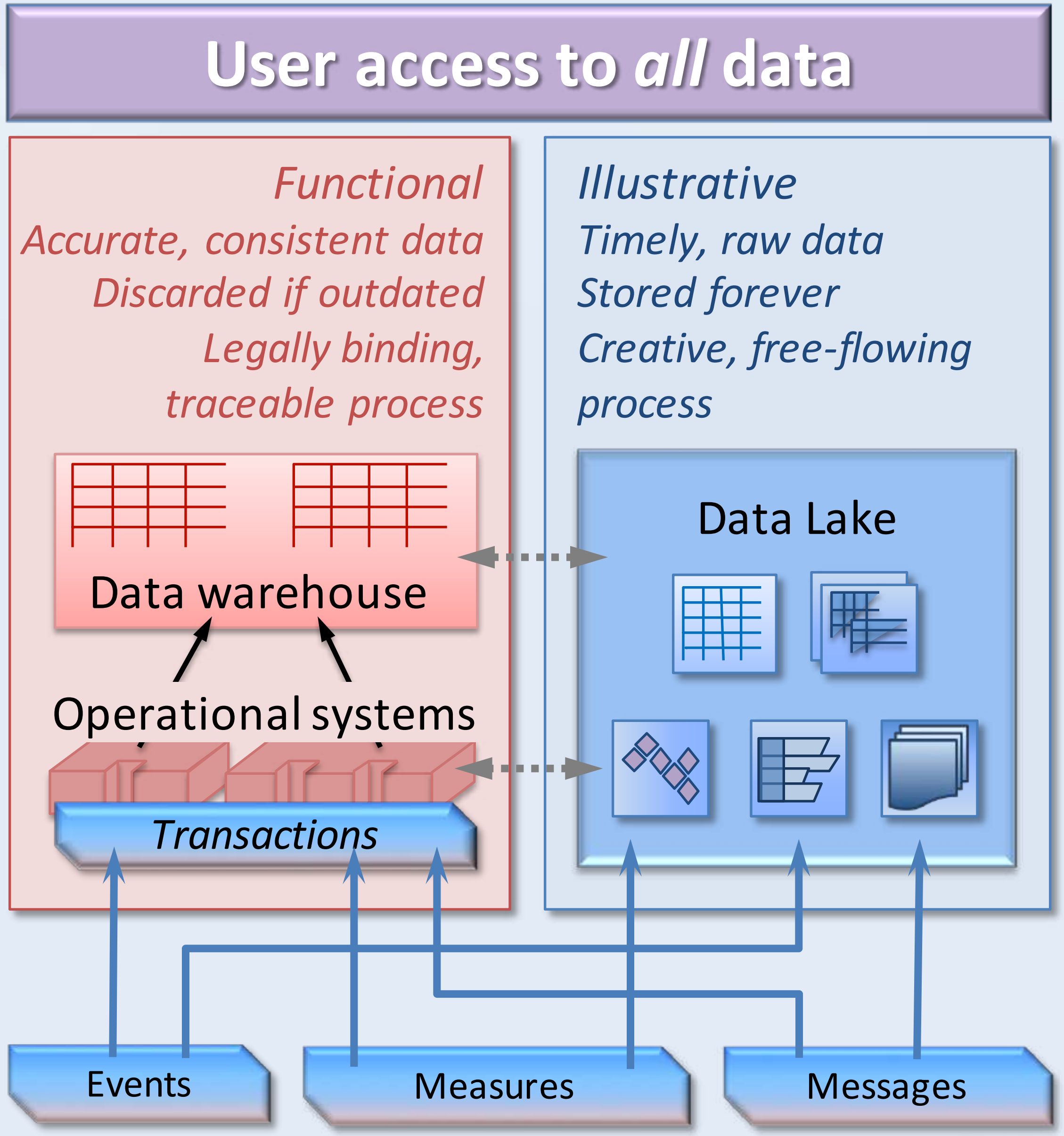
Business unIntelligence – Information pillars

- One architecture for all types of information
  - Mix/match technology as needed
    - *Relational, NoSQL, Hadoop, etc.*
- Integration of sources and stores
  - Instantiation gathers inputs
  - Assimilation integrates stored info.
- Data flows as fast as needed and reconciled when necessary
  - No unnecessary storage or transformations
- Distinct data management / governance approaches as required



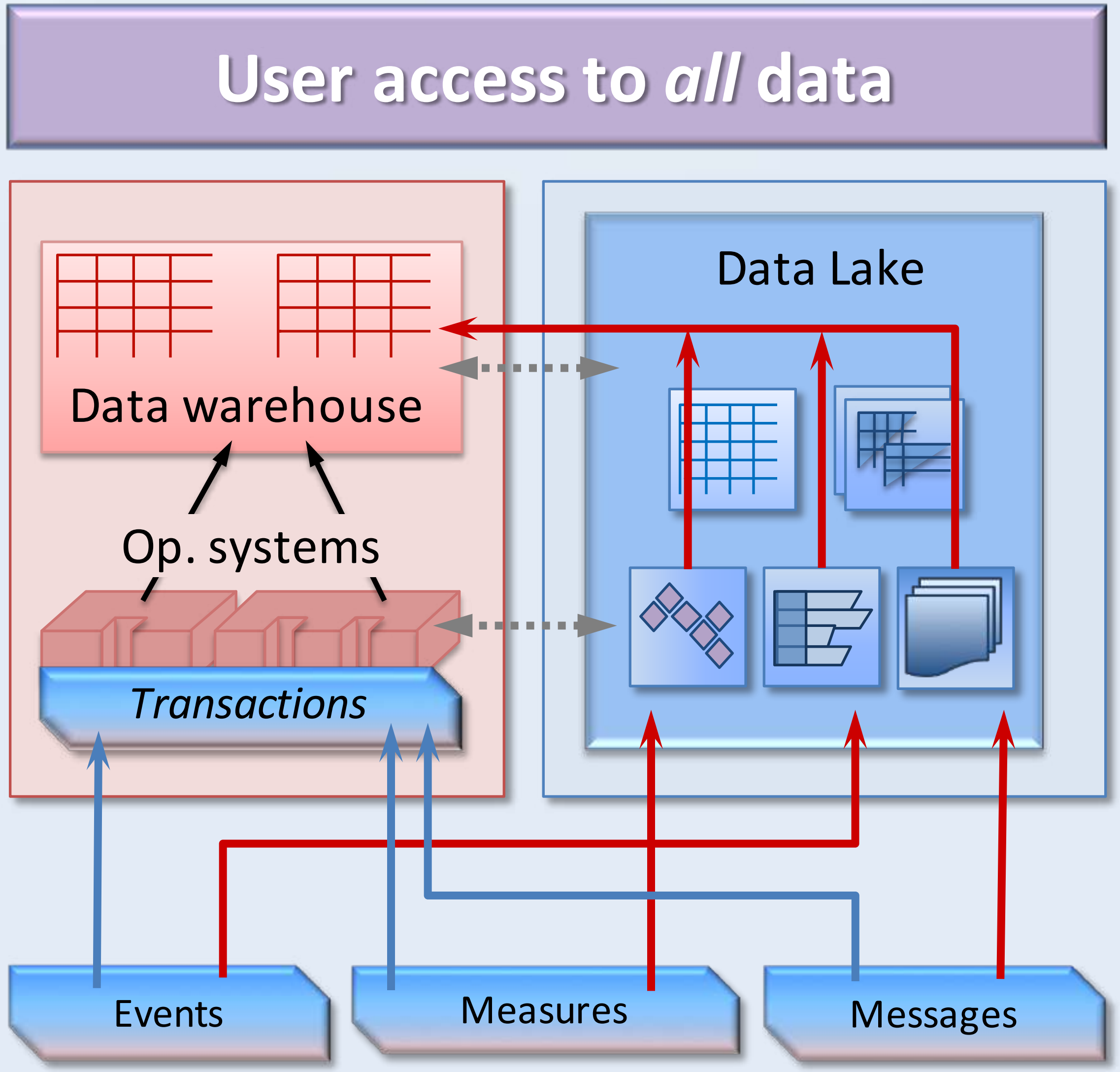
Positioning of data lake and warehouse today

- Serve different purposes
  - Functional – run / manage the business
  - Illustrative – predict / influence the future
- Both required
  - Optimized for different strengths
  - Warehouse = accuracy and consistency
  - Lake = timeliness and rawness
- Links between environments
  - Better than copying everything into one (or both)
- Together – foundation for pervasive analytics



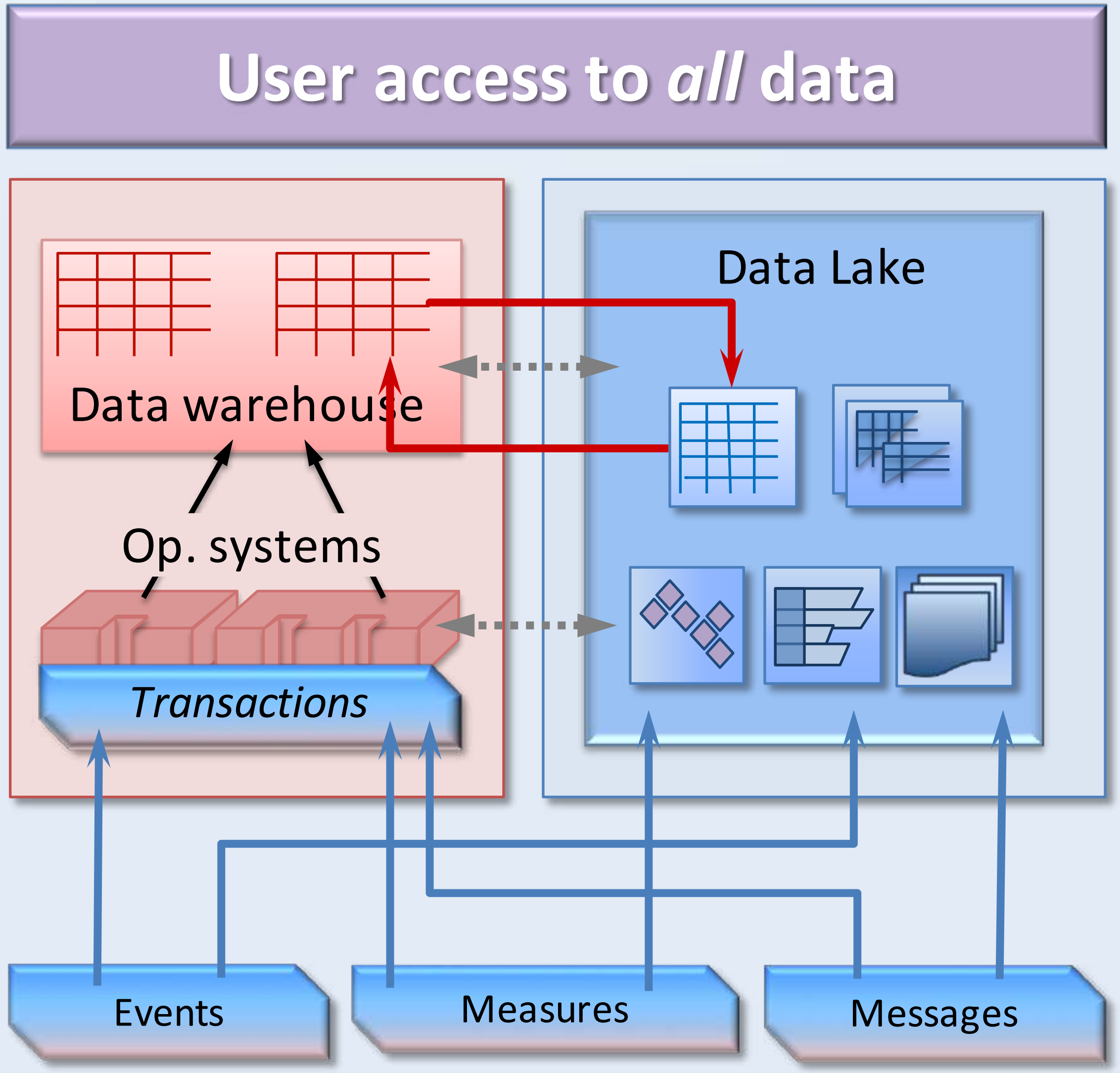
A warehouse by a lake (1) Preparation and enrichment

- Challenge: ETL (extract, *transform* and load) to data warehouse complex and computationally expensive
- Transform in:
  - Proprietary ETL server – with high licensing cost
  - Data warehouse server – with impact on analytic tasks
- Solution: Pump some or all data through the data lake
  - Reduced processing cost and/or impact on DW work



A warehouse by a lake (2) Archival

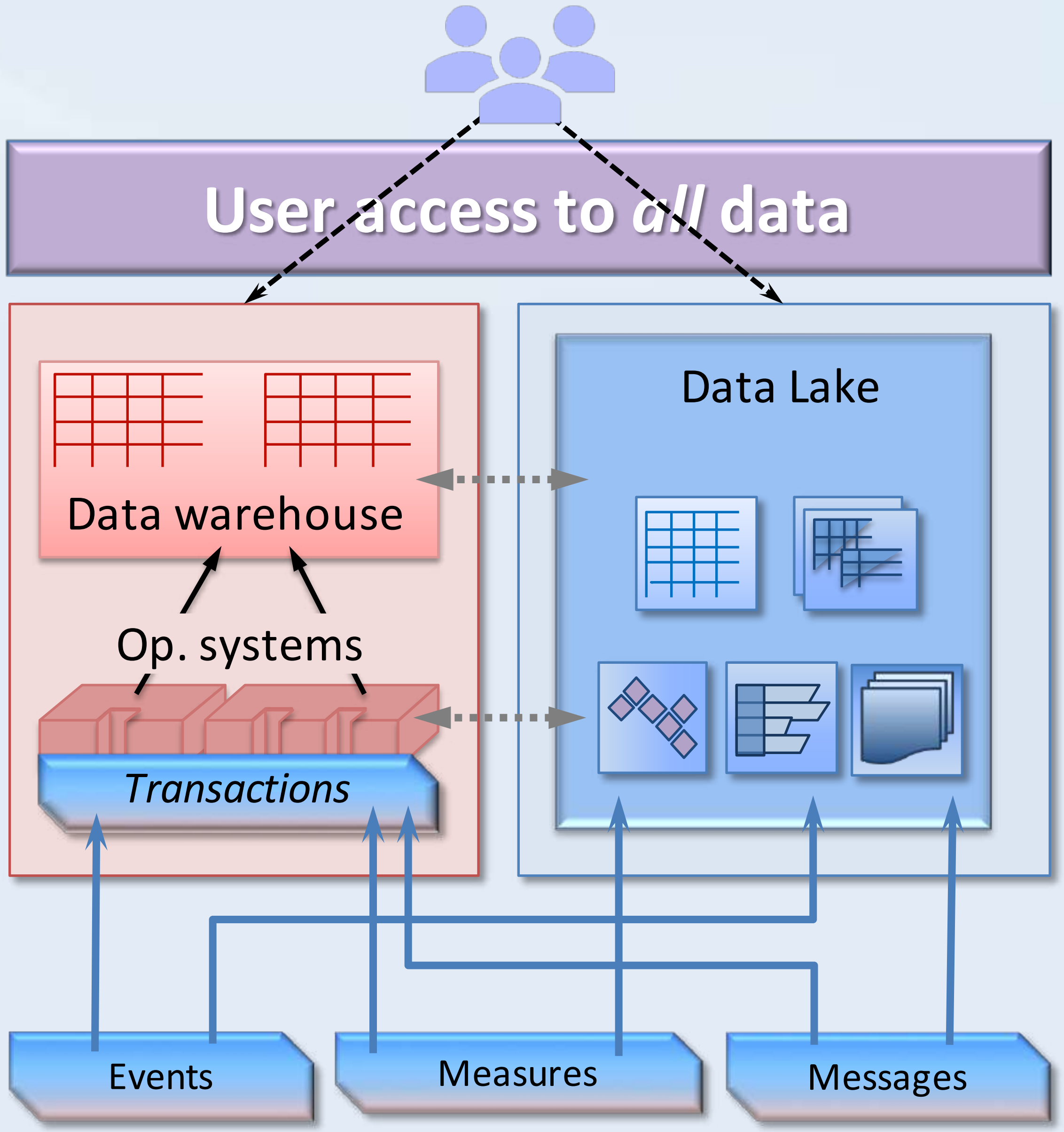
- Challenge: Storing seldom-used (cold) data in a data warehouse is an expensive waste of high-performance hardware
- Archiving to magnetic tape delays and complicates access to off-line data when needed
- Solution: archive to commodity servers and disks in data lake
  - Hadoop – no licensing costs
  - Faster access when needed – almost equal to DW
  - Same tools (SQL-based) for access as DW



A warehouse by a lake (3) Access

- Challenge: Data increasingly resides on disparate platforms
  - Traditional business info in relational
  - Business people familiar with SQL
  - Social media, IoT on Hadoop / NoSQL / etc.
  - Copying back and forth is expensive

- Solution: Virtualize access to data on all platforms
  - SQL-based queries
  - Join data across platforms



Conclusions

1. Enterprise data warehouse lives on
   - Focused on core business information
   - Traditional relational platforms still preferred

2. Data lake complements data warehouse
   - Focused on externally sourced data
   - Linked to data warehouse in multiple ways

3. Data lake can assist / offload data warehouse
   - Use commodity storage and processing power
   - Reduce costs and improve performance

Thank You

Dr Barry Devlin

Founder & Principal
9sight Consulting

Hortonworks: Next-Gen Data Architectures To Power Modern Analytics

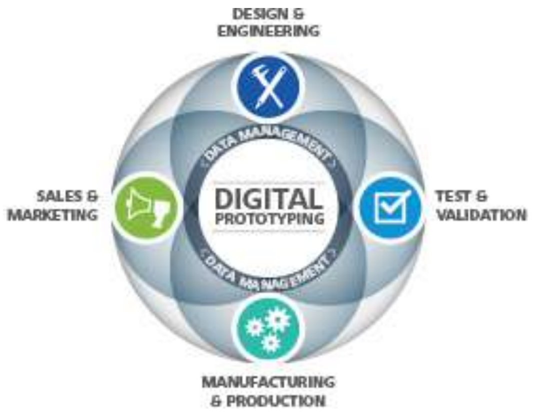
Piet Loubser

VP Product and Solutions Marketing

Hortonworks

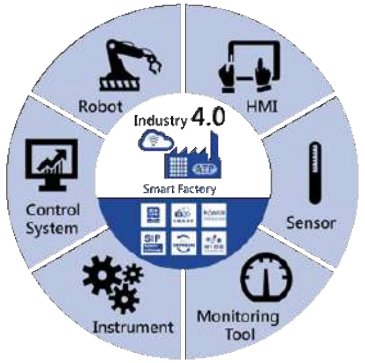
The New Way of Business Is Fueled By Connected Data

Development



- Connected Customers, Vehicles, Devices
- Socially crowd-sourced requirements
- Digital design and analysis
- Digital prototypes and tests (simulations)

Manufacturing



- Connected Factories, Sensors, Devices
- Human-robotic interaction
- 3D-printing on demand

Distribution

- Connected Trucks, Inventory
- Location, traffic, weather-aware distribution
- Real-time inventory visibility
- Dynamic rerouting

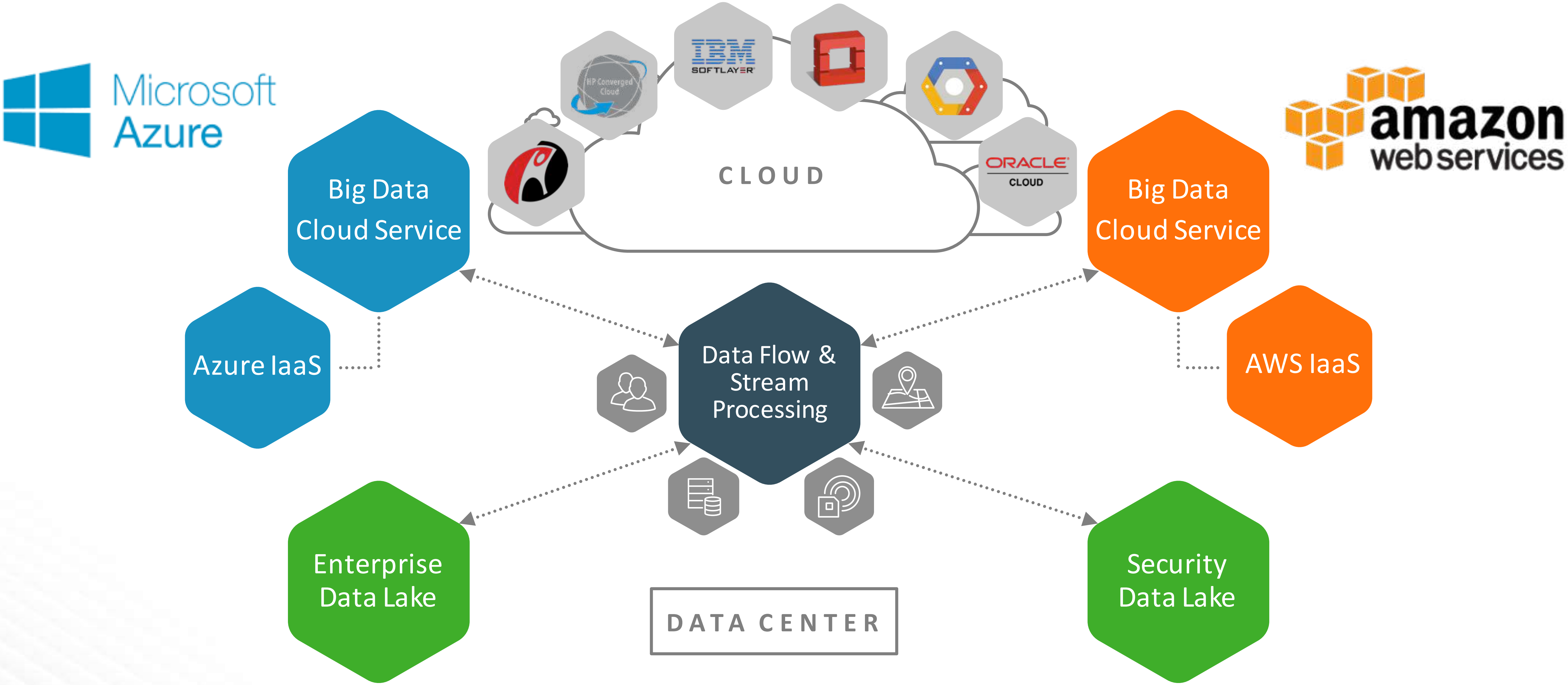
Marketing/Sales

- Connected Customers, Devices
- Omni- channel demand sensing
- Real-Time Recommendations

Service

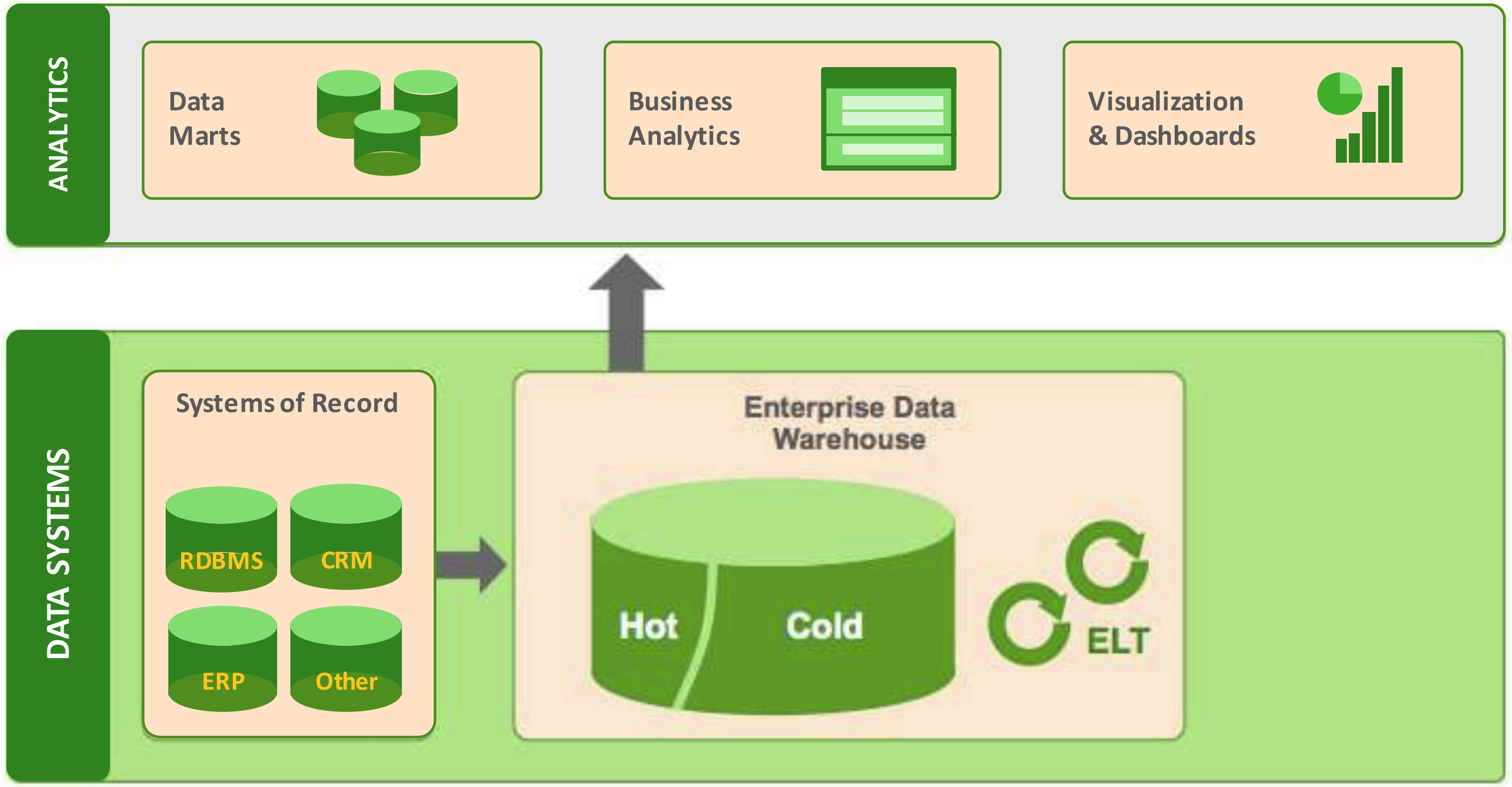
- Connected Assets
- Remote service monitoring & delivery
- Predictive maintenance
- OTA Updates

A Connected Data Strategy Connects Data Center and Cloud



Typical EDW Architecture

Used inefficiently, from $7,500 to $35,000 per TB[1] of data stored and processed



In a typical EDW:

- **50-70%** of data is unused and/or cold
- **45-65%** of CPU capacity is ETL/ELT
- **25-35%** of CPU consumed by ETL is to load unused data
- **30-40%** of CPU is consumed by only **5%** of ETL workloads
- As little as **2.8%** of the data is **Hot**[1]

Source: Hortonworks Innovation and Strategy Team and Appfluent Analysis
1. EY Analysis shows typical range from $10-15k / TB. Hortonworks experience shows a wide range observed in the field, from $35k/TB for massive, in-memory EDW appliances to $7.5k/TB for RDBMS based, home-grown EDW solutions
2. For example, for a client keeping a rolling 36-month window of data for reporting in an EDW, only 1 month of the 36 (2.8%) is new/hot.

HORTONWORKS

Hortonworks Connection: Services and Solutions for Your Success

Hortonworks Solutions

Enterprise Data Warehouse Optimization

Cyber Security and Threat Management

Internet of Things and Streaming Analytics

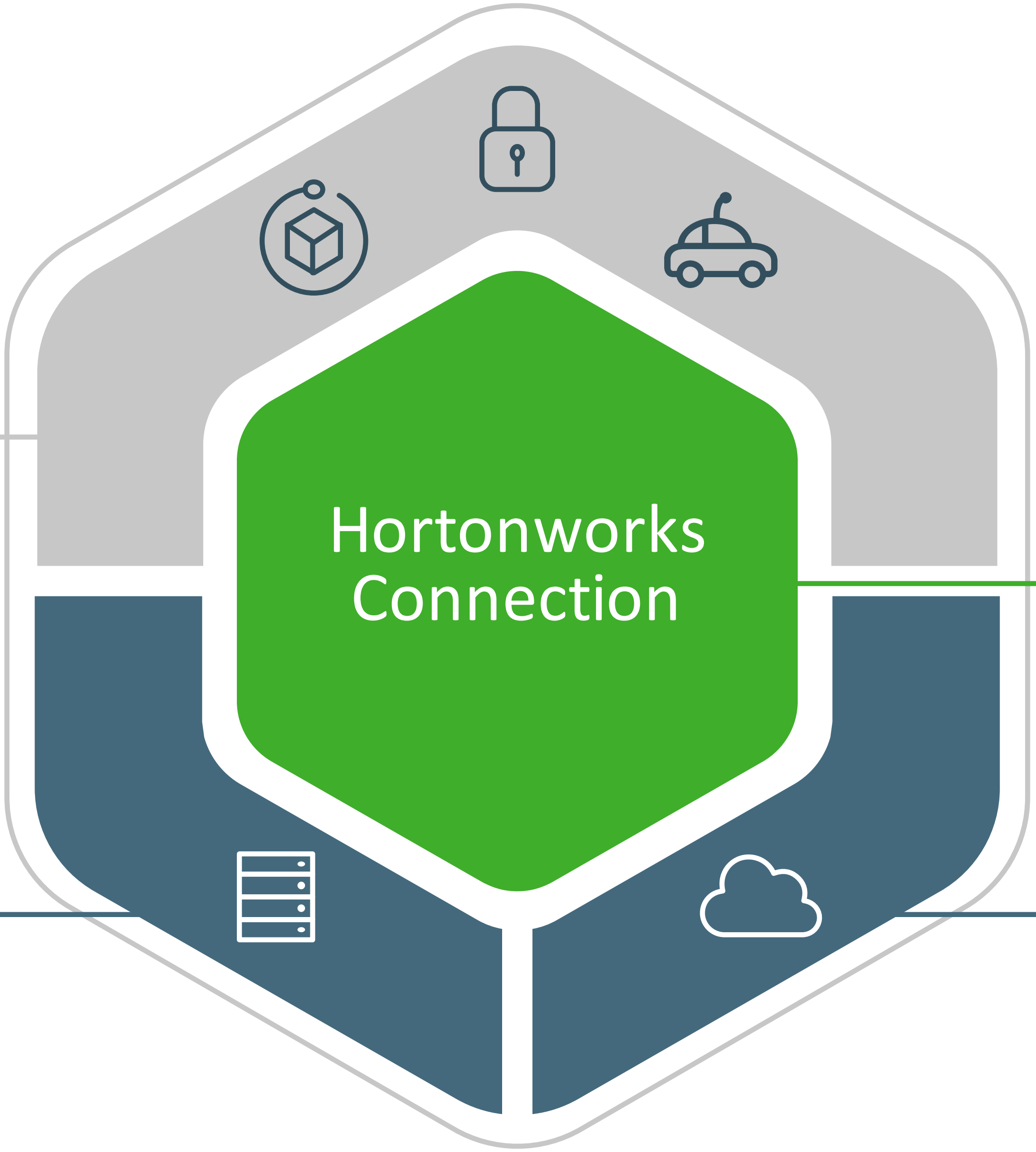
Data Science Experience

Advanced SQL

Data Center

Hortonworks Data Suite

HDP HDF



Data Services

Hortonworks Connection

Enablement Subscription

SmartSense™

Premier Operational Support

Educational Services

Professional Services

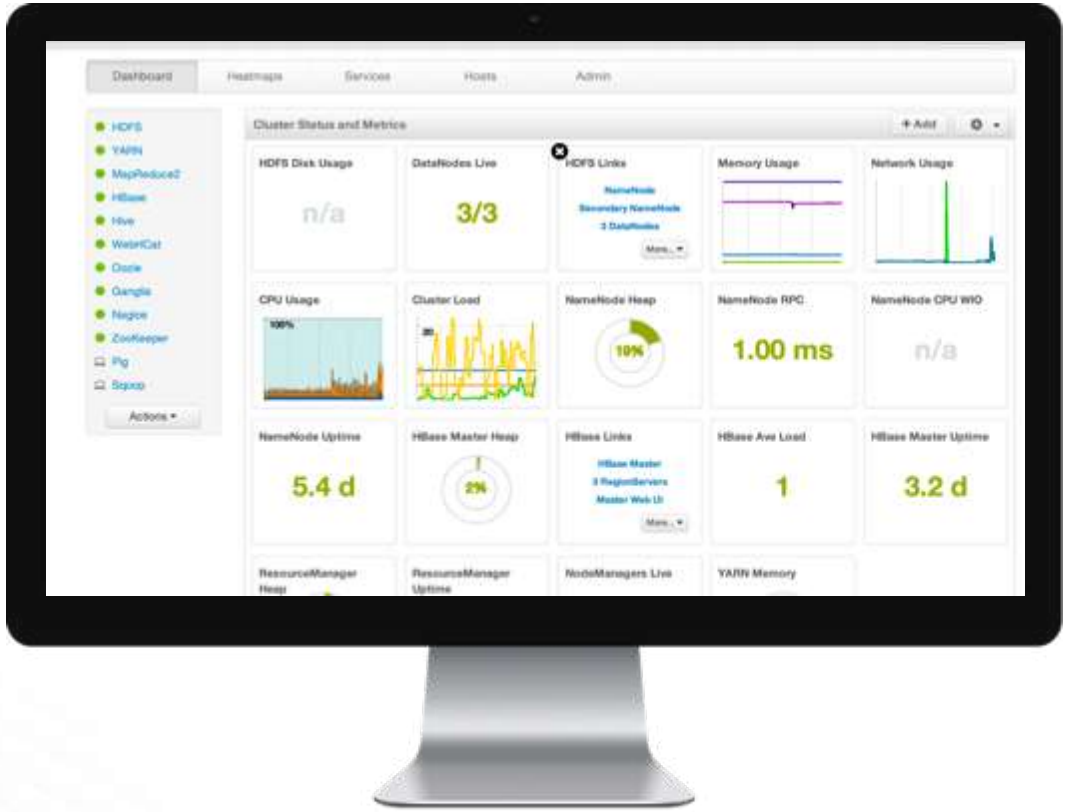
Community Connection

Cloud

Hortonworks Data Cloud

AWS HDInsight

Enterprise Data Warehouse Optimization



Dramatic Cost Reductions

Reduce cost of your EDW Implementation by offloading ETL processes and archiving cold data

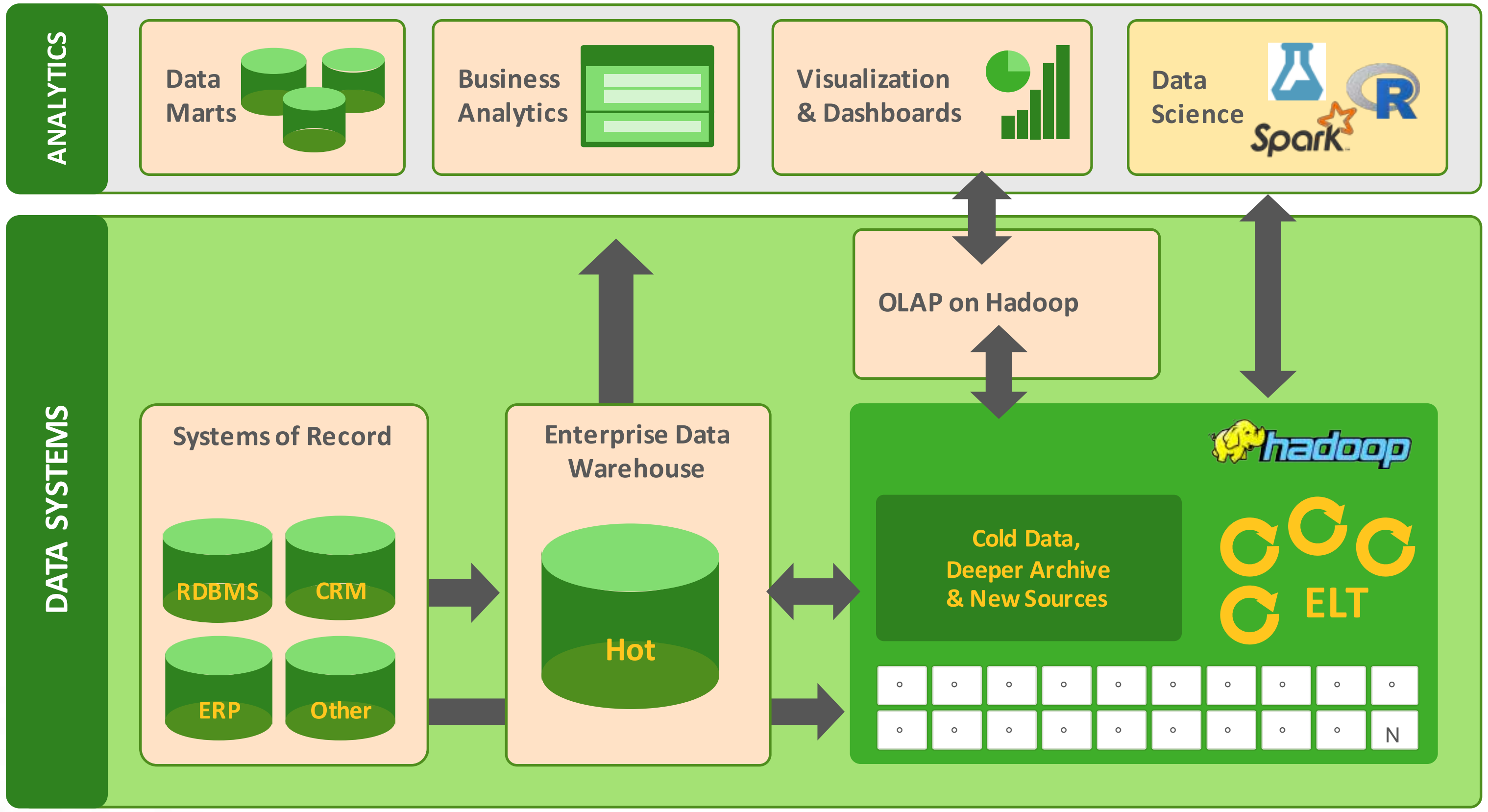
Deploy Business Intelligence on Hadoop

Empower Business users with powerful reporting, new applications, visualization tools, and artificial intelligence

Support More Types of Unstructured Data

Index and search images, videos, text & sound files

EDW Plus Hadoop helps you optimize and reduce costs associated with your EDE



Archive Cold Data away from EDW

- Move cold or rarely used data to Hadoop as active archive
- Store more of your data longer, cheaper

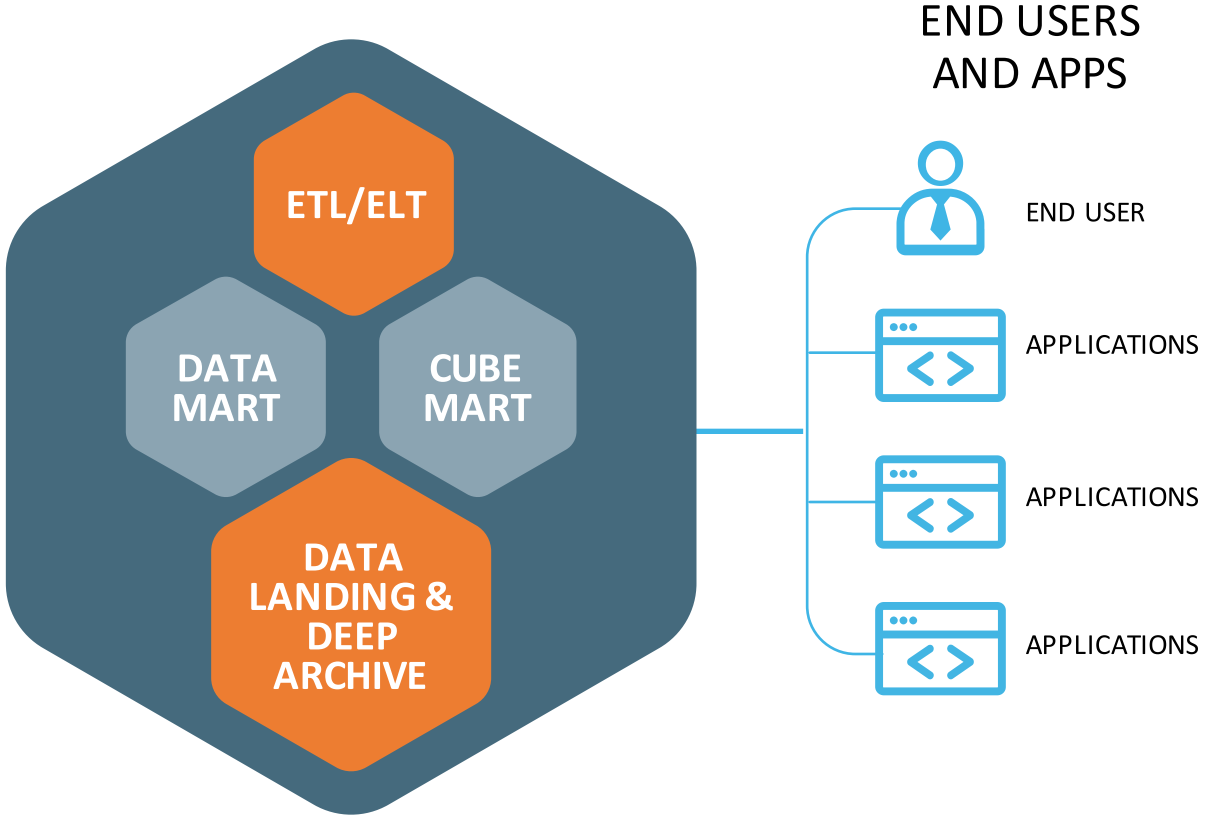
Offload costly ETL process

- Free your EDW to perform high-value functions like analytics & reporting, not ETL
- Use Hadoop for advanced or massive-scale ETL/ELT

EDW Optimization: ETL Offload

- The Problem:
  - EDWs consume between 50% and 90% of CPU just on ETL/ELT tasks.
  - These jobs interfere with more business-critical tasks like BI and advanced analytics.
- The Solution:
  - Hive and HDP deliver ETL that scales to petabytes.
  - Syncsort DMX-h for simple drag-and-drop ETL workflows.
  - Economical scale-out processing on commodity servers.
- The Result:
  - Better SLAs for mission-critical analytics.
  - Limit EDW expansion or retire old systems.

EDW OPTIMIZATION SOLUTION

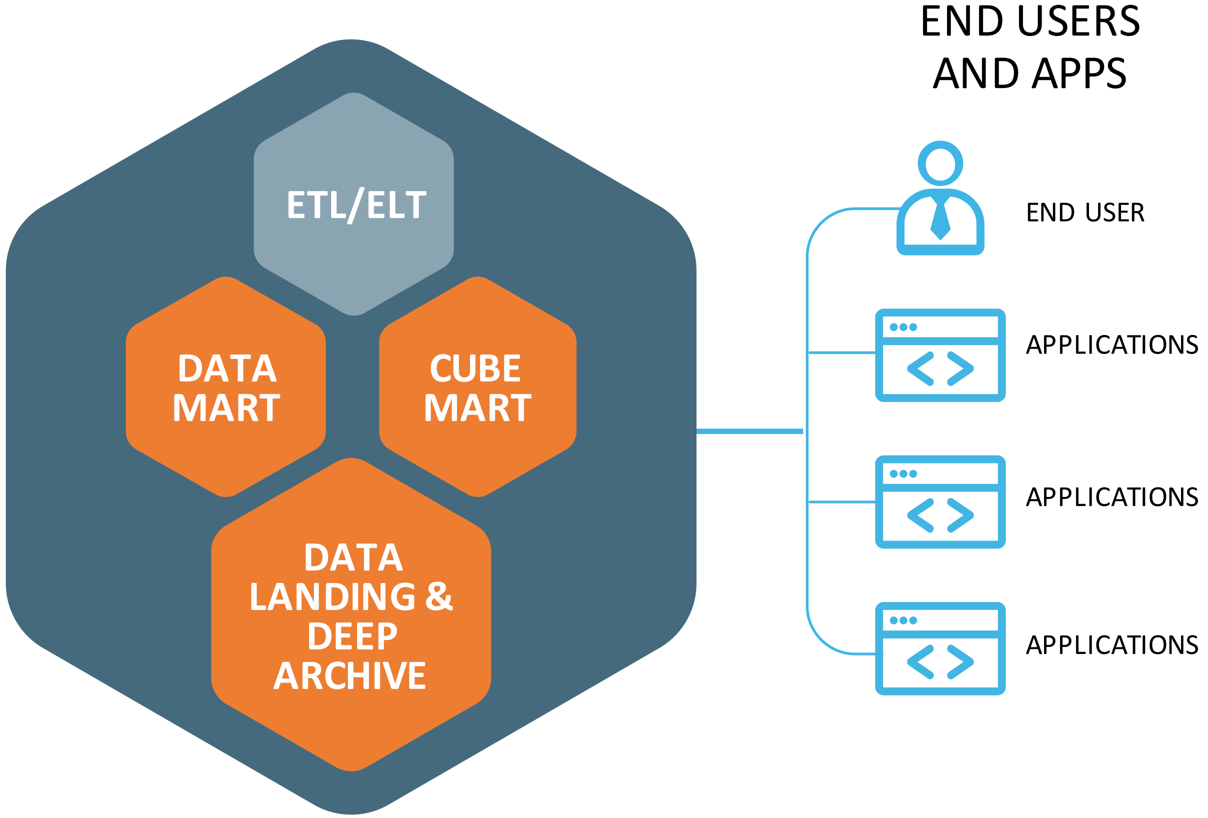


EDW Optimization: Active Archive

- The Problem:
  - Increasing data volumes and cost pressure force data to be archived to tape.
  - Archived data not available for analytics, or must be retrieved at great expense.
- The Solution:
  - Adopting Hadoop delivers cost per terabyte on par with tape backup solutions.
  - Data in Hadoop can be analyzed by all major BI tools, allowing analytics on archive data.
- The Result:
  - Data always available for analytics.
  - Store years of data rather than months.

EDW OPTIMIZATION SOLUTION

HORTONWORKS syncsort

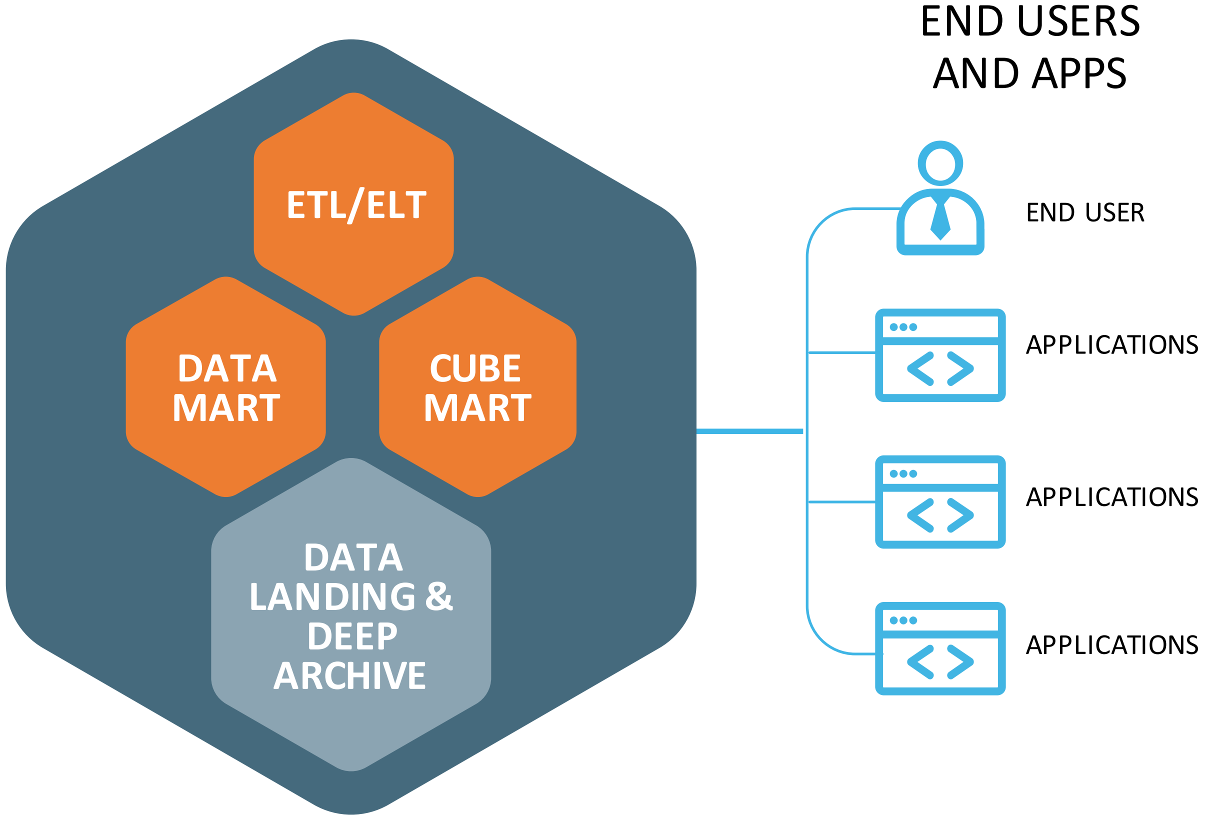


EDW Optimization: Fast BI on Hadoop

- The Problem:
  - Proprietary EDW systems were adopted for Fast BI and deep slice-and-dice analytics, but EDW prices are unsustainably high.
- The Solution:
  - Interactive SQL is a reality on Hadoop today.
  - Partner Solutions (IBM BigSQL, Kyvos, Jethro) adds powerful SQL and OLAP capabilities for deep drilldown at scale.
- The Result:
  - Query terabytes of data in seconds.
  - Connect your favorite BI tools like Tableau and Excel through SQL and MDX interfaces.
  - The EDW Optimization Solution is tailor-made to deliver Fast BI on Hadoop.

EDW OPTIMIZATION SOLUTION

HORTONWORKS IBM kyvos insights jethro



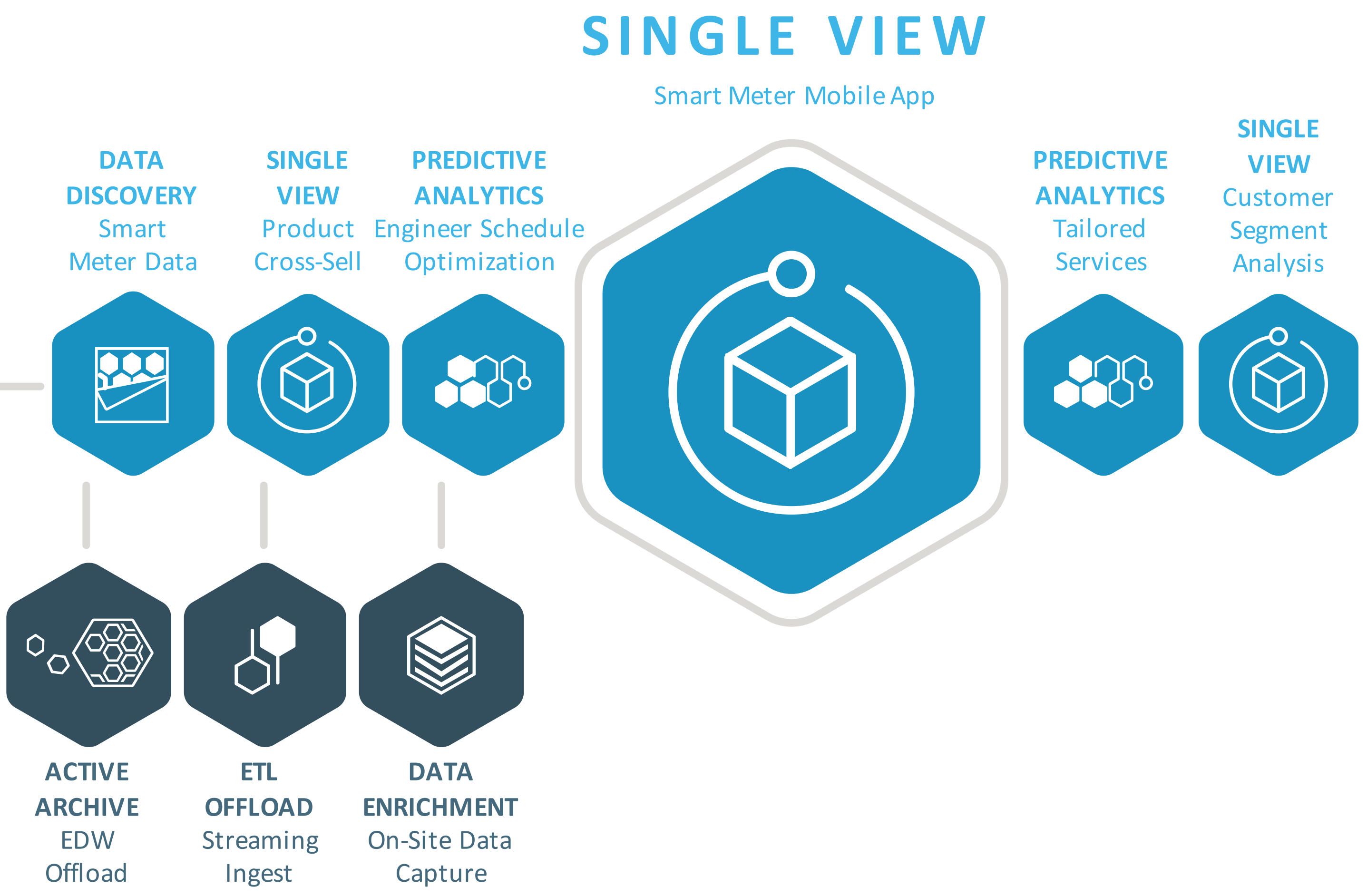
Centrica Transforms Service For Utility Customers

SITUATION

Existing infrastructure made loading data difficult & caused analytic bottlenecks

Goal: reduce costs, streamline processes for a single view of customers

Data fragmentation hid business-wide patterns from analysts



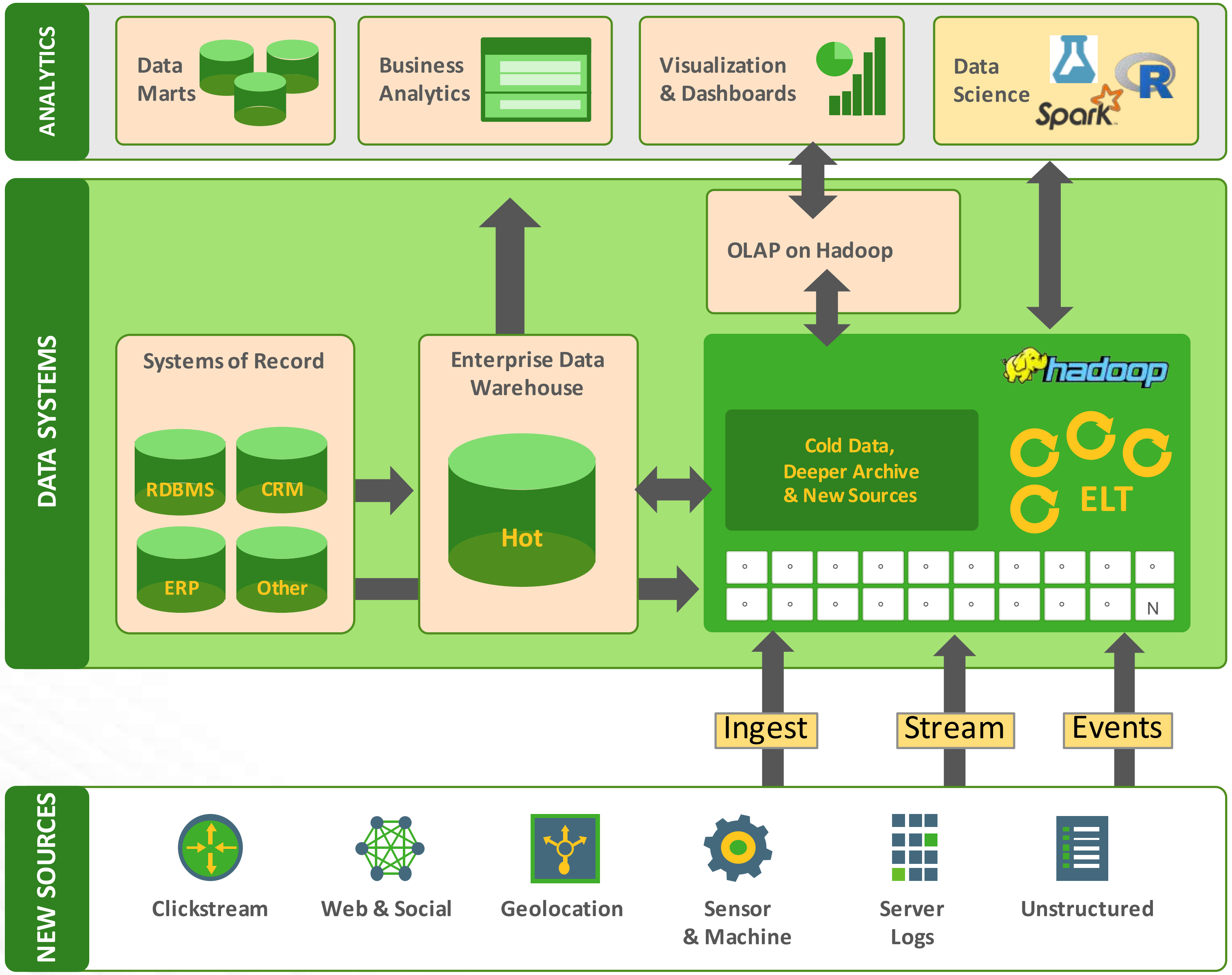
centrica

3 Million Customers	can access “smart energy reports”
ETL efficiency gains	from 11 hours to 45 minutes/job
300 GB/Day Ingest	rationalizes work of field engineers
Decommissioned some EDWs	saving millions annually

“Focusing on innovation, learning to forget traditional legacy ways of working and approaching it in new ways creates unexpected behavioural changes, because people feel freer and they also feel valued.”
Dajit Rehal, Senior Systems Director

HORTONWORKS

EDW Plus Hadoop helps you land and enrich more data to respond faster to new business requests



Archive Cold Data away from EDW

- Move cold or rarely used data to Hadoop as active archive
- Store more of your data longer, cheaper

Offload costly ETL process

- Free your EDW to perform high-value functions like analytics & reporting, not ETL
- Use Hadoop for advanced or massive-scale ETL/ELT

Land & Enrich more data to create more value-add analytics

- Use Hadoop to ingest new data sources, such as web and machine data for new analytical context from **unstructured and semi-structured sources**
- **Create an analytical sandbox** for advanced data science

HORTONWORKS

Prescient Harnesses Machine Learning for Traveler Safety Warnings

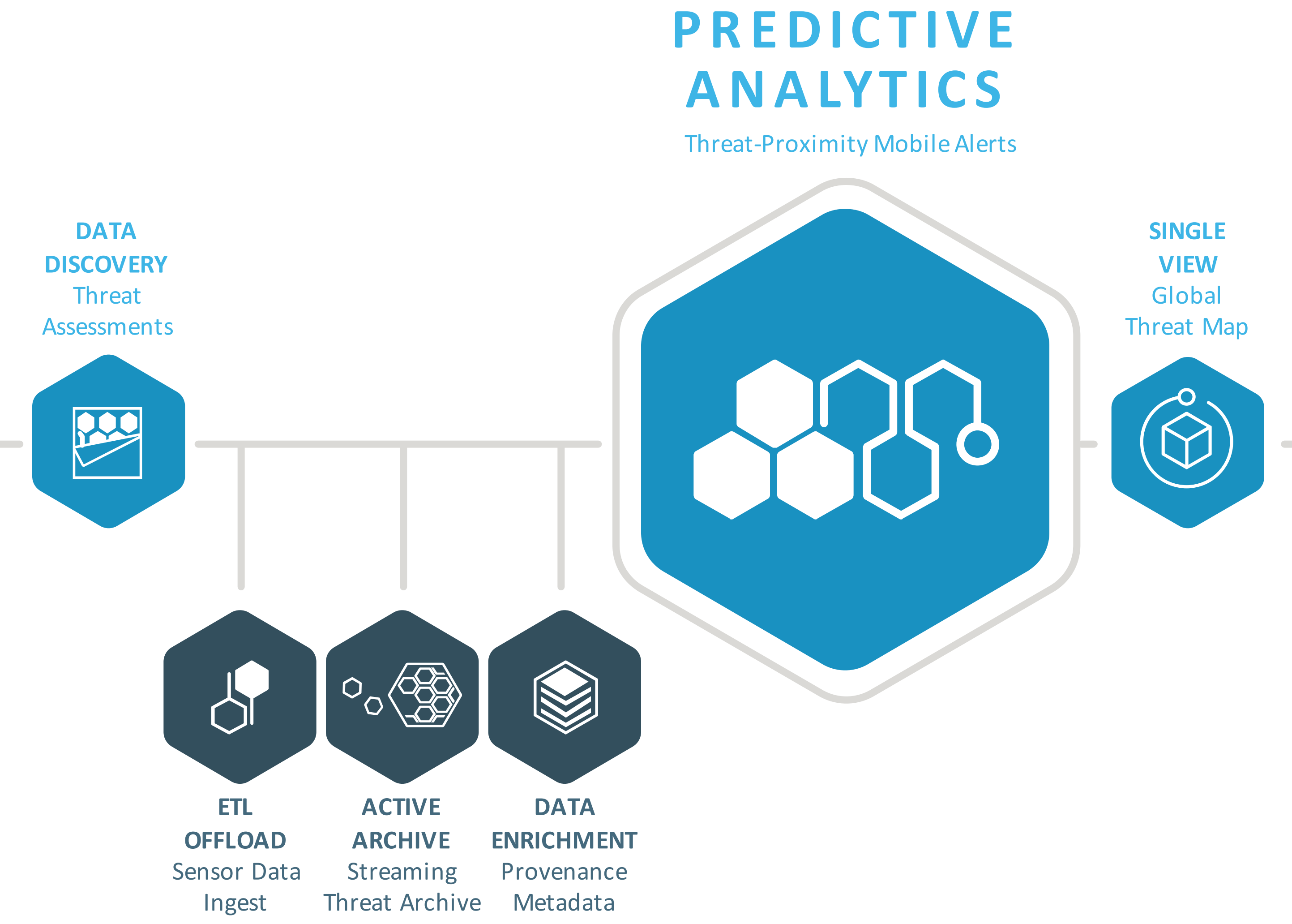
SITUATION

Performs risk management

Uses humans to identify false positives

Could only produce one assessment every 3-4 days

Needed efficient way to store raw data for analytics



PRESCIENT

49,500 Data Sources	ingested by HDF into HDP
700% Productivity Improvement	for geospatial analysts
5 Petabytes of Data	stored in HDP connected EMC
Hybrid Architecture	HDF connects data center to cloud

“We know that when we define a high-threat area in a given area of the world, that it is underpinned by very specific data sources. It’s data-driven, and we can point to those sources—if ever asked—and say, ‘Here’s why.’” Mike Bishop, Chief Systems Architect

Why Hortonworks?

Powered By 100% Open Source

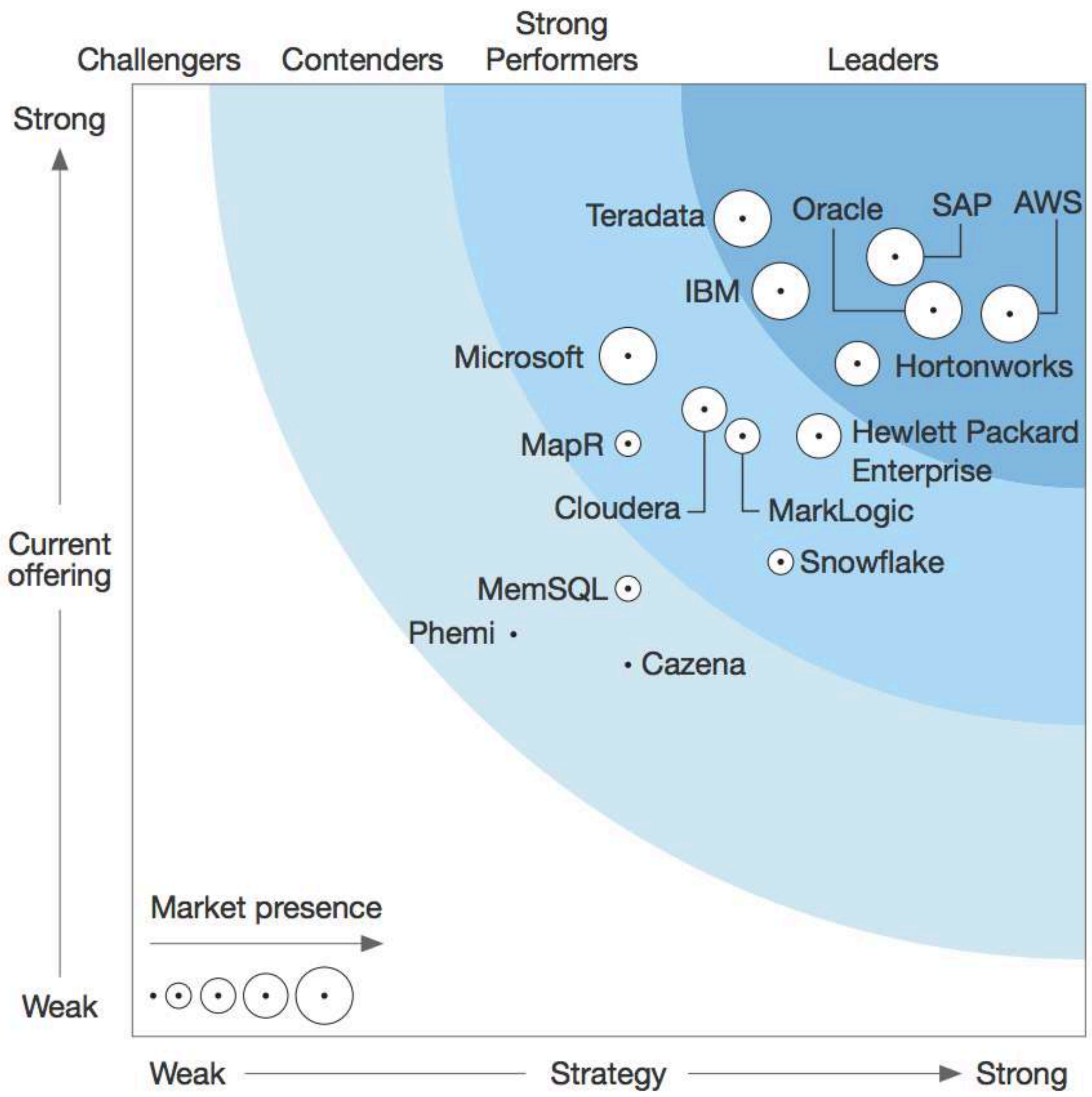
Rapid innovation
Dramatic cost reduction

Enterprise Ready

Governance
Fine grained security
Lineage and data provenance

Powering All Data

Data-at-Rest, Data-in-Motion
Cloud, On-Premises
Structured, unstructured



Forrester Wave: Big Data Warehouse, Q2 2017

hortonworks.com/get-started/big-data-scorecard/

HORTONWORKS

Thank You

HORTONWORKS®
POWERING THE FUTURE OF DATA™