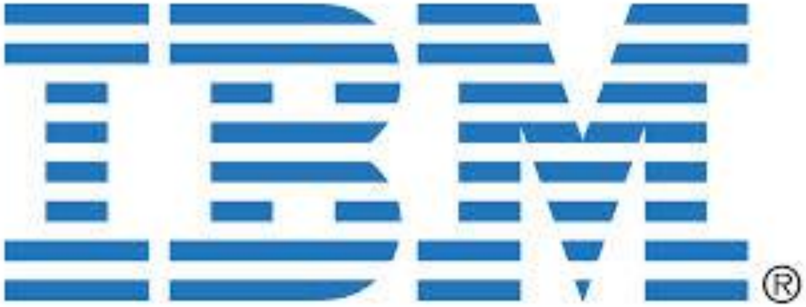
Exploring the Benefits of the Modernized Data Warehouse

Philip Russom

TDWI Research Director for Data Management

October 8, 2014

Sponsor



Speakers

Philip Russom
TDWI Research Director,
Data Management

Gord Sissons
Senior Manager, Product Marketing
IBM

Agenda

PLEASE TWEET
@pRussom, #TDWI, #EDW, #DataWarehouse,
#DataArchitecture, #Analytics, #RealTime

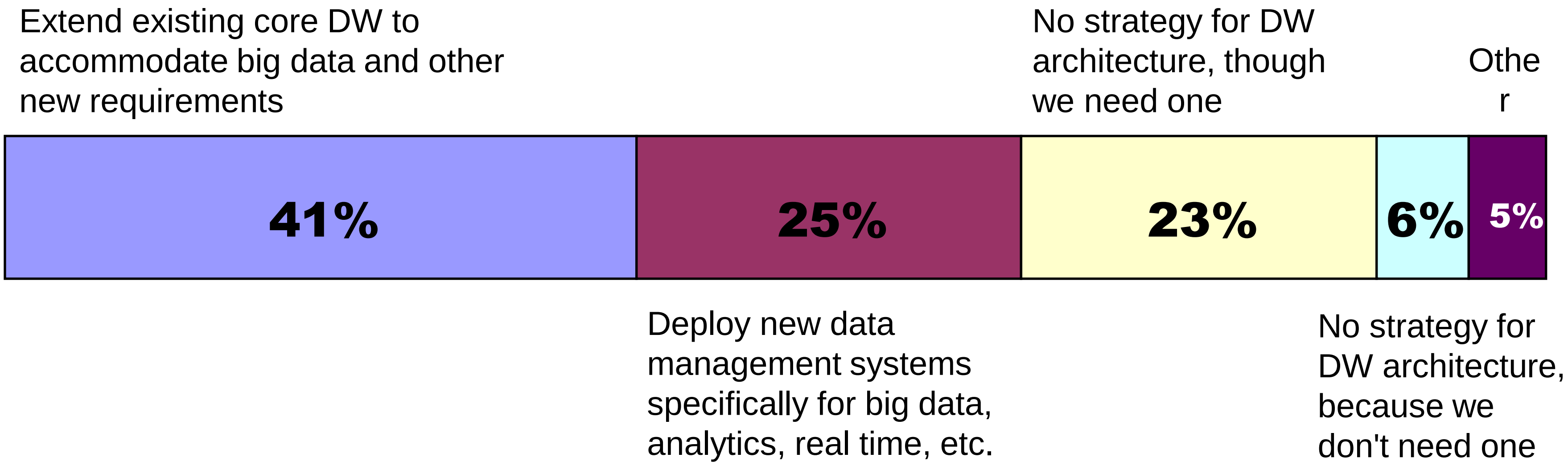
- Background
  - *What is data warehouse modernization?*
  - *What are the business benefits?*
  - *What are the technology reasons?*
- Common measures taken for Data Warehouse Modernization
  - *Multi-platform data warehouse environments*
  - *Logical architectures, data federation, virtualization*
  - *Hadoop issues and opportunities*
  - *Misc: real time, in-memory, streaming data, appliances*
  - *DW economics*
  - *Governing new big data*
- Recommendations

“DW Modernization” takes many forms…

- Additions to existing data warehouse environment (or ecosystem)
  - *New data subjects, sources, tables, dimensions, etc.*
  - *More server instances, nodes, bigger storage*
- More standalone data platforms and tools
  - *Complement DW without replacing it*
  - *Tools for analytics, real time, new data types, new interfaces*
  - *New appliances, columnar databases, Hadoop, NoSQL, etc.*
- Architectural Adjustments
  - *Logical DW design across multiple platforms*
  - *Extending data integration (DI)*
- Upgrades
  - *Newer versions of current database or integration software*
  - *Bigger and faster hardware*
- Rip and Replace
  - *Decommission current DW platform or misc tools; migrate to others*

Which of the following best describes your organization's strategy for evolving [or modernizing] your DW environment and its architecture, relative to big data?

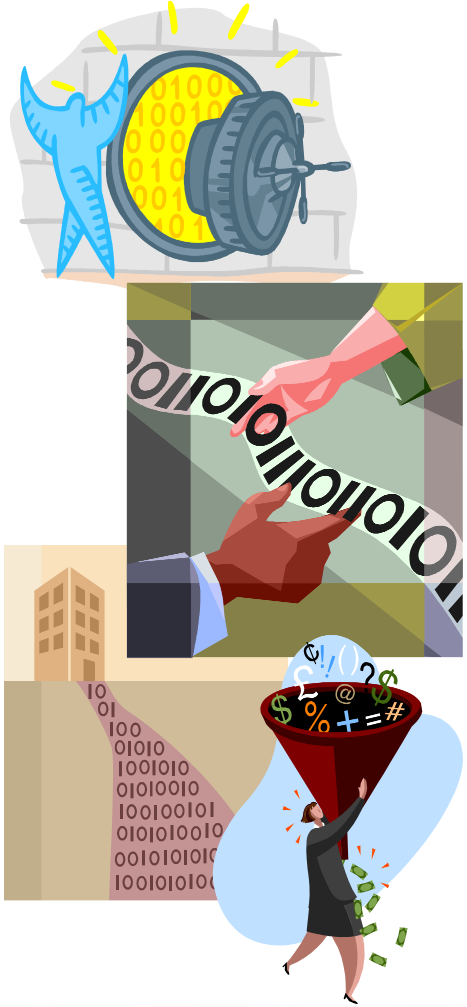
- Most survey respondents plan to extend an existing DW (41%, far left)
- A few will deploy new data platforms (25%)
- 29% have no strategy for DW evolution or addressing big data



Source: TDWI survey run in late 2013. Based on 538 respondents.

DATA WAREHOUSE MODERNIZATION POSITIONS AN ORGANIZATION FOR

Greater Business Benefits via Data



- Run the business by the numbers
  - *Requires fresh data, from the best sources, delivered fast, to key people*
- Complete information
  - *Complete customer views, enterprise-scope data, social media, big data…*
- Trusted data
  - *High quality, governed, audit trail*
  - *For reports, analyses, operations, etc.*
- Real-time information
  - *Enables time-sensitive biz practices*
  - *Streaming data sources*
- Business Analytics
  - *predict the future, correlate diverse entities, understand customers, compete on data, etc.*

Use Cases for Big Data Analytics

- Big Data enables exploratory analytics. Discover new:
  - *Customer base segments*
  - *Customer behaviors and their meaning*
  - *Forms of churn and their root causes*
  - *Relationships among customers and products*
- Analyze big data you've hoarded. Finally understand:
  - *Web site visitor behavior*
  - *Product quality based on manufacturing robotic data*
  - *Product movement via RFID in retail*
- Use tools that handle human language for visibility into:
  - *Claims process in insurance*
  - *Medical records in healthcare*
  - *Sentiment analysis in customer-oriented industries*
  - *Call center applications in any industry*
- Big data improves data samples for older analytic apps:
  - *Fraud detection*
  - *Risk management and actuarial calculations*
  - *Anything involving statistics or data mining*
- Big data adds more granular detail to analytic datasets:
  - *Broaden 360-degree views of customers, etc.*

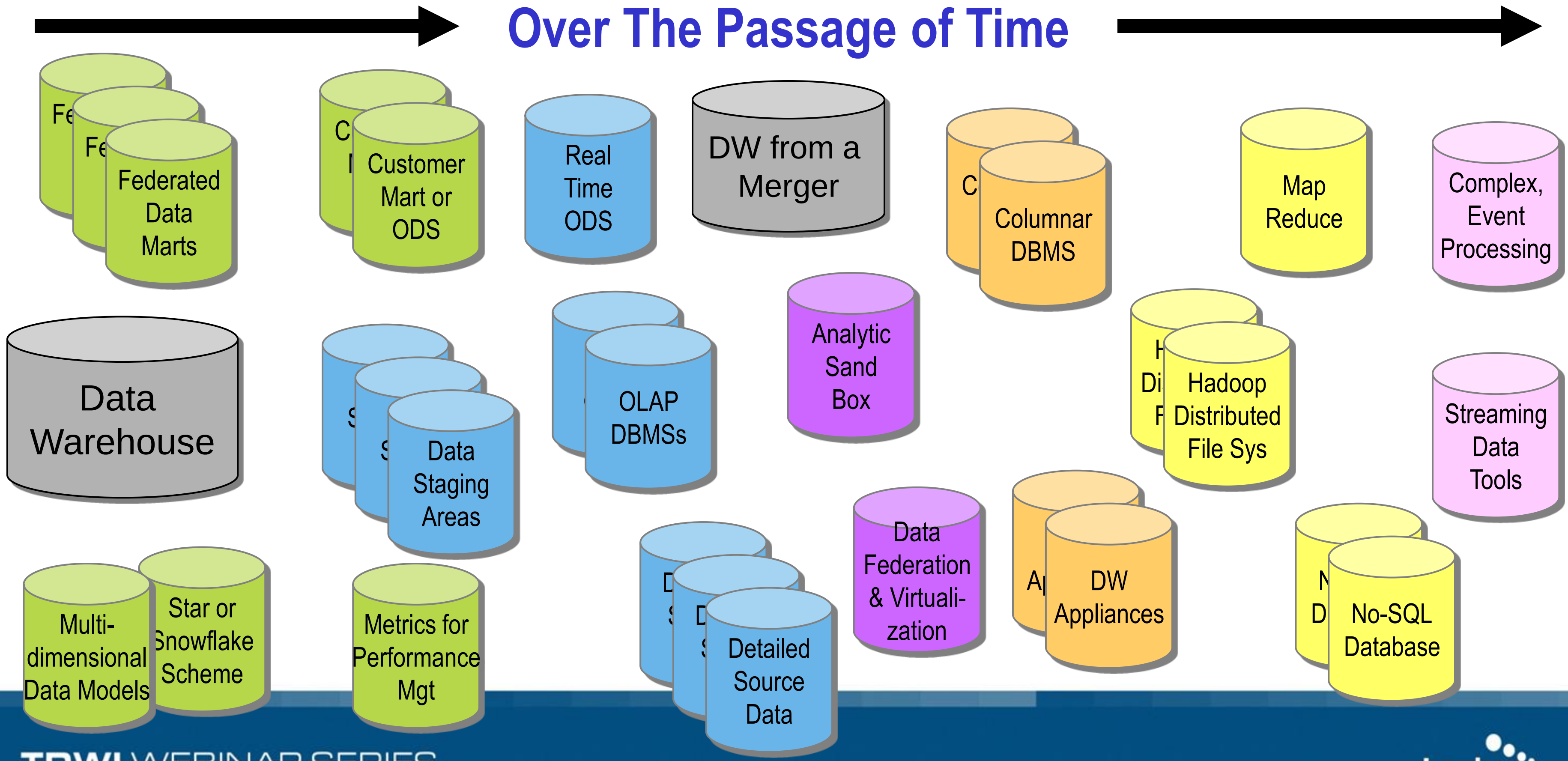
ONE WAY TO MODERNIZE A DW:

Multi-Platform Data Warehouse Environments

- Many enterprise data warehouses (EDWs) are evolving into multi-platform data warehouse environments (DWEs).
  - *Synonyms: data warehouse ecosystem, hybrid data ecosystem*
- Users continue to add additional standalone data platforms to their warehouse tool and platform portfolio.
- The new platforms don't replace the core DW, because it is still the best platform for the data that goes into standards reports, dashboards, performance mgt, and OLAP.
- Instead, the new platforms complement the DW, because they are optimized for workloads that manage, process, and analyze new forms of big data, non-structured data, and real-time data.

Modern DW System Architectures can be Complex

- The technology stack for DW, BI, analytics, and data integration has always been a multi-platform environment.
- What's new? The trend toward many data platforms has accelerated.
- Why? More platform types to serve big data & analytic workload types.



Logical versus Physical Data Architectures

And Other Architectural Components that Coexist

- **Logical architecture** – mostly about data models and their relationships, with a focus on how these represent organizational entities and processes
  - ***Data standards*** – *including standards for data modeling, data quality metrics, interfaces for data integration, programming style, format standards, etc.*
- **Physical architecture** – mostly a plan for deploying data and data structures based on the workload and platform requirements of each
  - ***System architecture*** – *a topology of hardware servers and software servers, plus the interfaces and networks that tie them together*

Data Assumptions for Modern DWEs

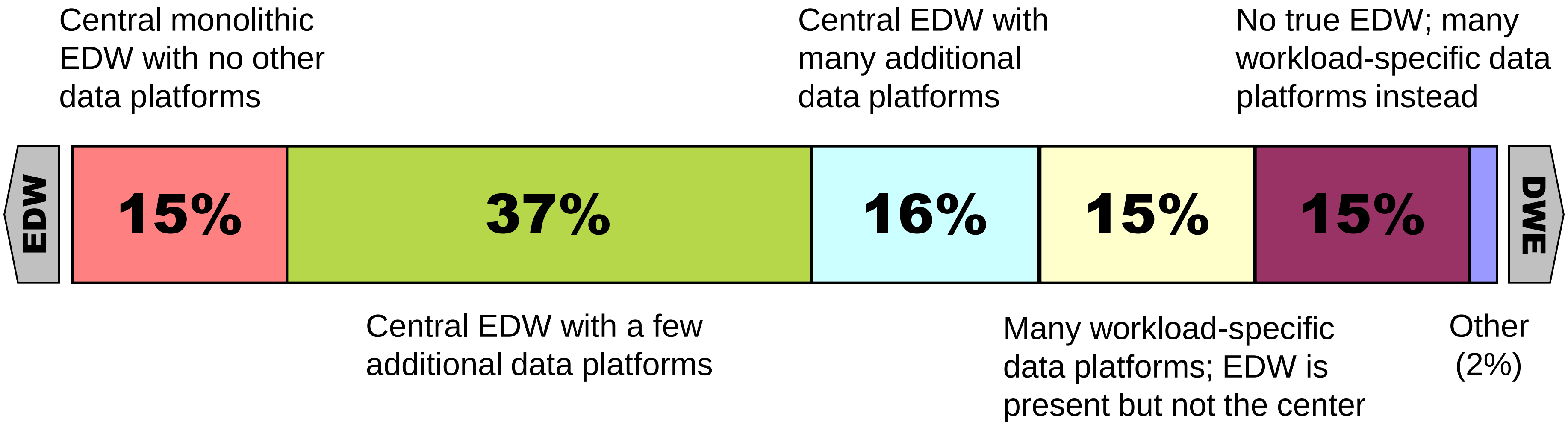
- Data (both big and small) moves around among the diverse data platforms of a modern multi-platform DWE:
  - *Data integration, replication, synchronization, ETL, ELT, etc.*
  - *A DWE needs a well-formed architectural layer for physical data integration (PDI)*
- Minimize moving large data volumes by using more logical approaches to data:
  - *Data federation, virtualization, views, external tables, distributed queries, etc.*
  - *A DWE needs a well-formed architectural layer for logical data integration (but coordinated with the physical data integration layer)*

Ramifications of Multi-Platform Data Environments

- Workload-centric DW architecture
  - *Assumes that some workloads and their data are best offloaded from the core DW and taken to a platform more suited to them*
- Distributed DW architecture
  - *This simply means that data and data structures (as defined in a logical architectural layer) are distributed across multiple physical data platforms*
  - *The logical layer of data architecture can provide the "big picture" across platforms*
- A distributed DW architecture is both good and bad
  - *Good if it serves the unique requirements of multiple workloads & users*
  - *Bad if platforms proliferate like the dreaded data marts of yore; or complexity crushes you*
- This is largely about the system layer of data architecture
  - *The trend is to extend the diversity of the server portfolio*
  - *More options for users, when they map from logical to system layers*
- Other reasons for multiple platforms
  - *Due to departmental funding; sandboxes; mergers and acquisitions*
- Data integration architecture changes, too (not just DW or data arch.)
  - *More data to move & process; sync data across platforms; new data types & sources*

Which of the following best describes your extended data warehouse environment today?

- Pure, central, monolithic EDWs are relatively rare (15%, far left)
- Conversely, environments without a DW are equally rare (15%, far right)
- EDWs coexist well in mixed environments (68%, middle three)



Source: TDWI survey run in late 2013. Based on 538 respondents.

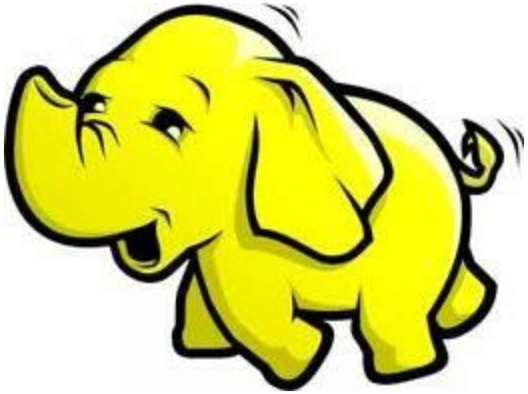
Hadoop integrated with a Relational DBMS

The strengths of one balance the weaknesses of the other

- A Relational DBMS is good at:
  - *Metadata management*
  - *Complex query optimization*
  - *Query federation*
  - *Table joins, views, keys, etc.*
  - *Security, including roles, directories*
  - *Much more mature development tools*
- HDFS & other Hadoop tools are good at:
  - *Massive scalability*
  - *Lower cost than most DW platforms & analytic DBMSs*
  - *Multi-structured data & no-schema data*
  - *Some ETL functions; late binding; custom code for analytics*
- Use HDFS and other Hadoop tools
  - *For scalable ODSs and data staging, to modernize existing DW architecture*
  - *For algorithmic analytics, to extend your DW environment*

Interface Issues

for HDFS & Related Tools

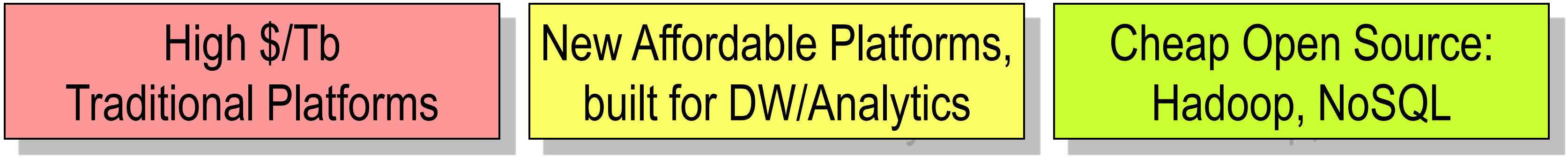


- Look for tools that include interfaces that are appropriate to HDFS and Hadoop tools, for example:
  - *Generate Hive QL; read/write HBase*
  - *Generate Java, C, R, etc. that's optimized for execution by MapReduce or YARN*
- SQL is a special case
  - *Hadoop needs basic SQL support SQL-based analytics, ETL/ELT push-down processing in HDFS, compatibility with SQL-based tools, etc.*
  - *JDBC/ODBS overlay for HDFS*
  - *But highly complex SQL is best on a relational DBMS*

DW MODERNIZATION IS AN OPPORTUNITY TO RE-EXAMINE COSTS

The Economics of Data Platforms

- As you modernize a DW environment, rethink its economics
- Cost continuum of data platforms:



High $/Tb
Traditional Platforms

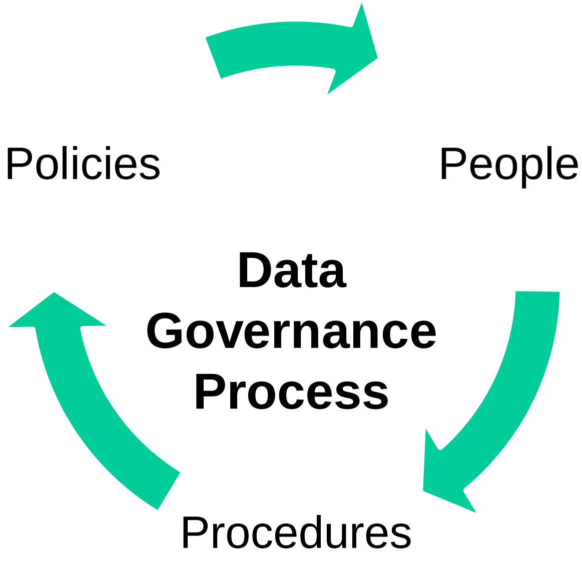
New Affordable Platforms,
built for DW/Analytics

Cheap Open Source:
Hadoop, NoSQL

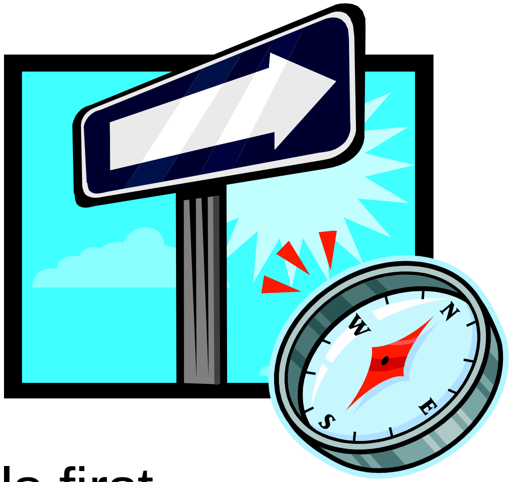
- Choose a platform that fits a given data workload – but also fits the value of data
  - *High-value data on the core EDW*
    - Modeling, cleansing, aggregating, and documenting data (which is required for reports and OLAP) are high value tasks, best done on DW
  - *Analytic datasets in the mid tier*
    - This data is lightly prepared or prepped on the fly; temp sandboxes
  - *Source & archival data on the back tier*
    - This is a “data lake” or archive that preserves data in its original form, so it can be repurposed repeatedly, as analytic projects arise

Data Governance's Role in DW Modernization

- Compliance is a pressing problem, and DG is part of the solution
  - *"Age of accountability" and BI demand high-quality, auditable data*
  - *Security & privacy: DG defines policies for controlled access to and use of data*
- Data Governance is more than compliance
  - *Data standards for data models, metadata, code style, data quality metrics, etc.*
  - *Data stewardship to make DG practical*
- Big Data must be governed, like all data.
  - *Each new data source should be certified per compliance and DG policies prior to use*
  - *DW modernization usually involves new data*
- Data exploration for big data analytics
  - *This is a common goal for DW mods*
  - *It needs boundaries to avoid violations*
- DG improves data and its usability
  - *Big data & adv'd analytics need this, too*
  - *Neither are "enterprise grade" without DG*



Recommendations

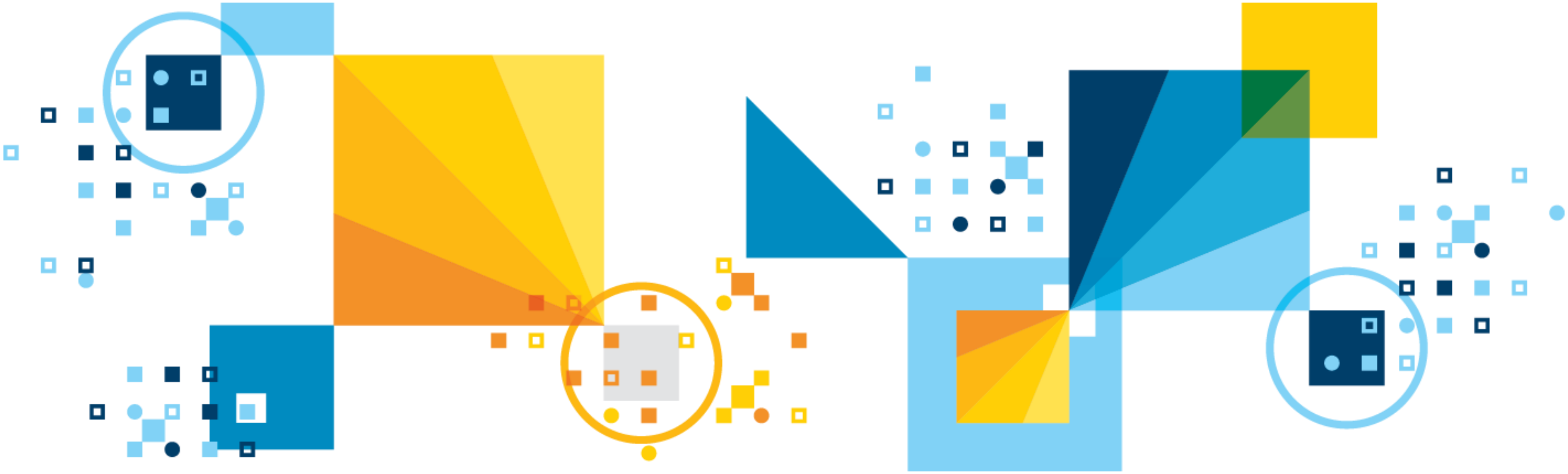


- Revaluate your data warehouse and related systems
  - *There's always room for improvement*
  - *Change is afoot, in both biz & tech*
- Prioritize modernization by putting biz goals first
  - *Biz wants to manage big data and leverage it*
  - *Biz wants to compete on analytics*
  - *Biz needs real-time tech to operate faster*
  - *Biz needs BI/DW solutions sooner, more agile*
- Technology goals are also important, though secondary
  - *Assuring capacity for growth*
  - *Diversifying data platform and tool portfolio to support more types of data, workloads, development methods, analytics, etc.*
  - *Migration to new platforms that are faster, more scalable, tuned for analytics, cost less, etc.*

IBM® InfoSphere® BigInsights

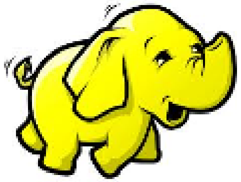
Getting to value faster with a modernized Data Warehouse

Gord Sissons, IBM



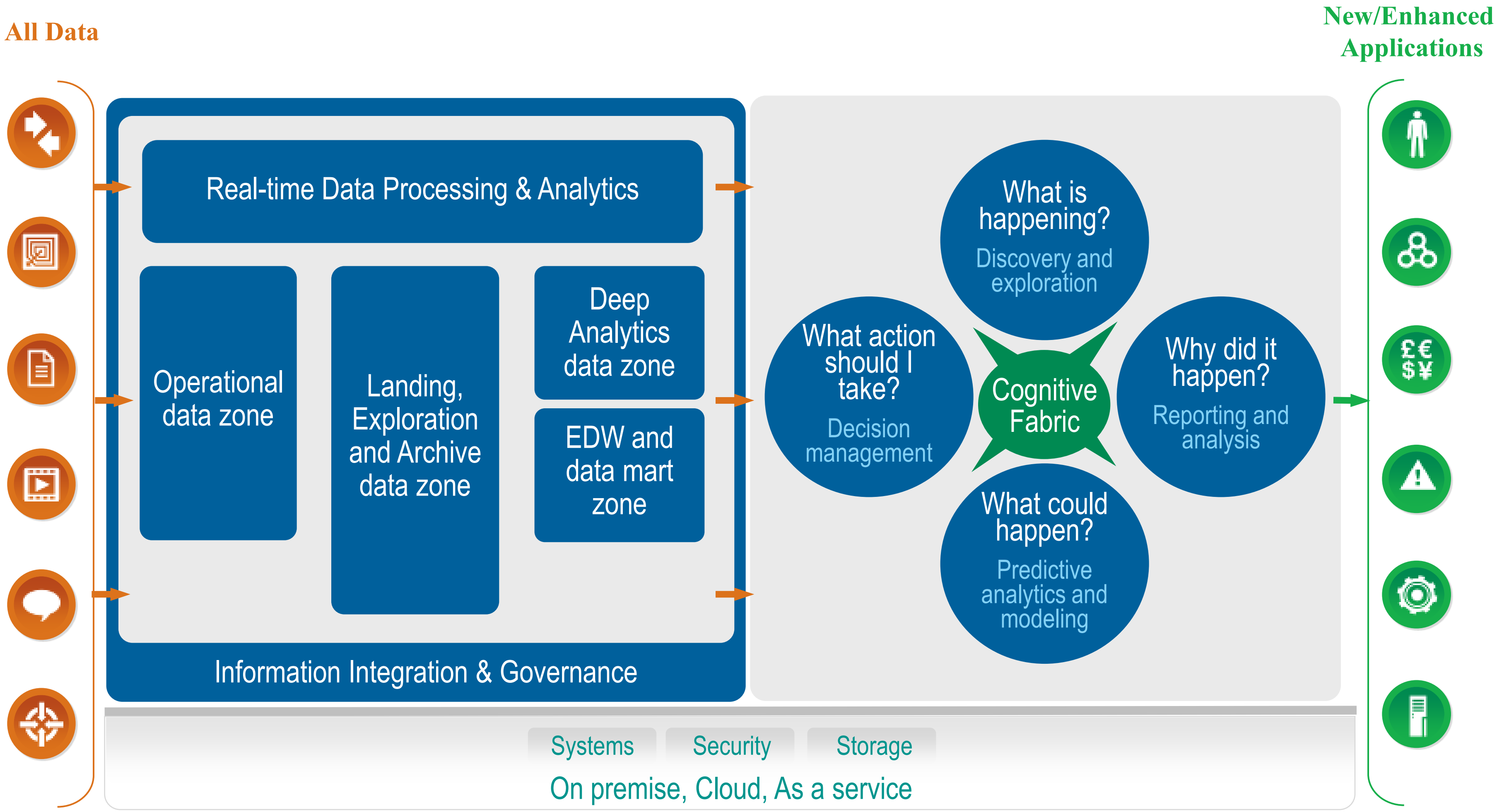
A new way of thinking

Driven by new technological capabilities and transformative economics – Hadoop gets much attention for new workloads

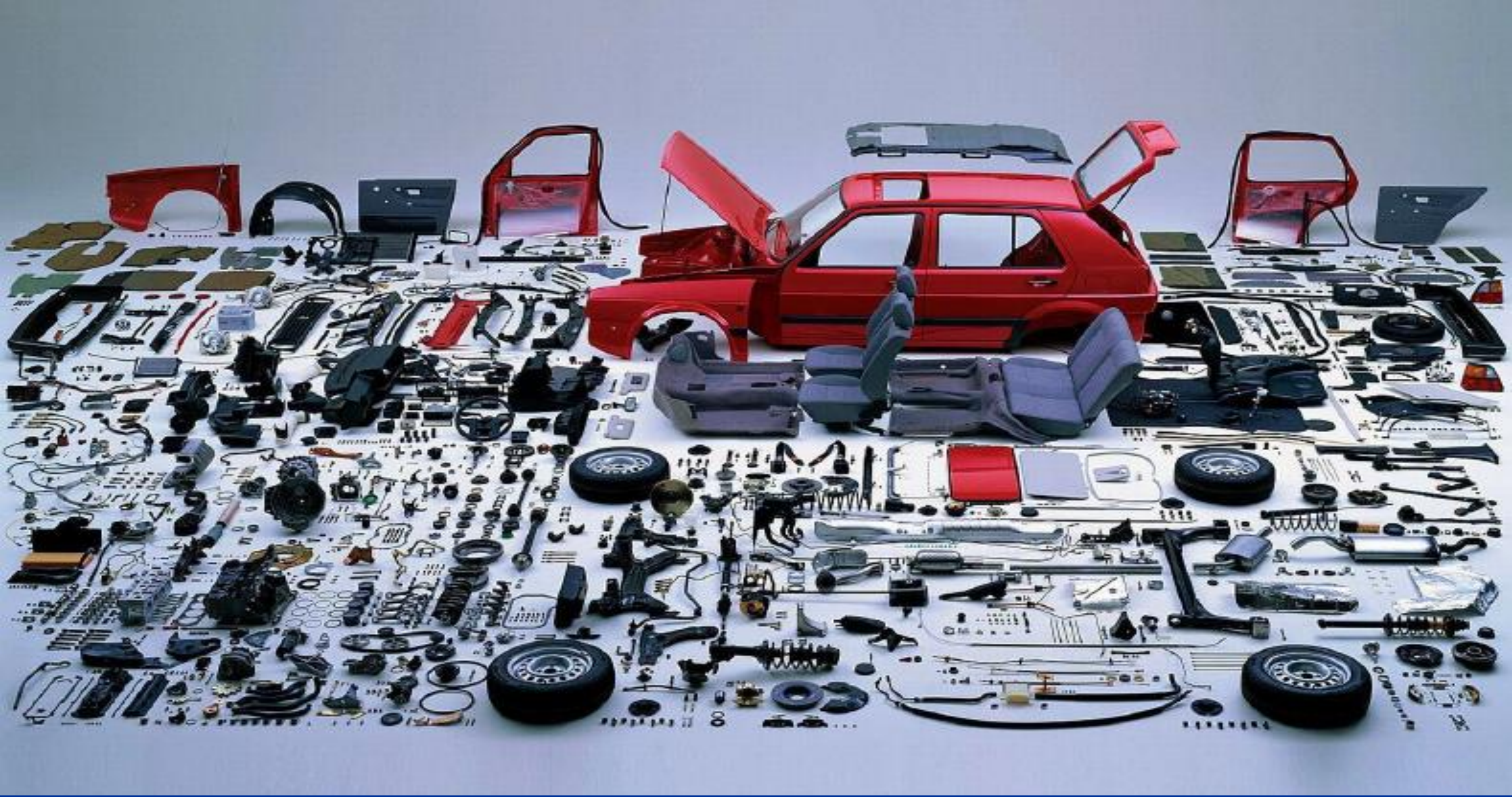


The Old Way	→	The New Way
Vertical infrastructure for DW/BI		Distributed data grids
What data should I keep?		Keep everything just in case
Design schemas in advance		Evolve schemas on the fly
ETL, down-sample, aggregate		Extract knowledge from raw data
What reports do I need?		Test every theory, model “what-ifs” on the fly

A Reference Architecture for Big Data & Analytics


All Data
New/Enhanced Applications
Real-time Data Processing & Analytics
Operational data zone
Landing, Exploration and Archive data zone
Deep Analytics data zone
EDW and data mart zone
Information Integration & Governance
What is happening?
Discovery and exploration
What action should I take?
Decision management
Cognitive Fabric
Why did it happen?
Reporting and analysis
What could happen?
Predictive analytics and modeling
£€ $¥
Systems
Security
Storage
On premise, Cloud, As a service

Rich capabilities translate into less customer risk, and higher chance of project success



Time to value matters

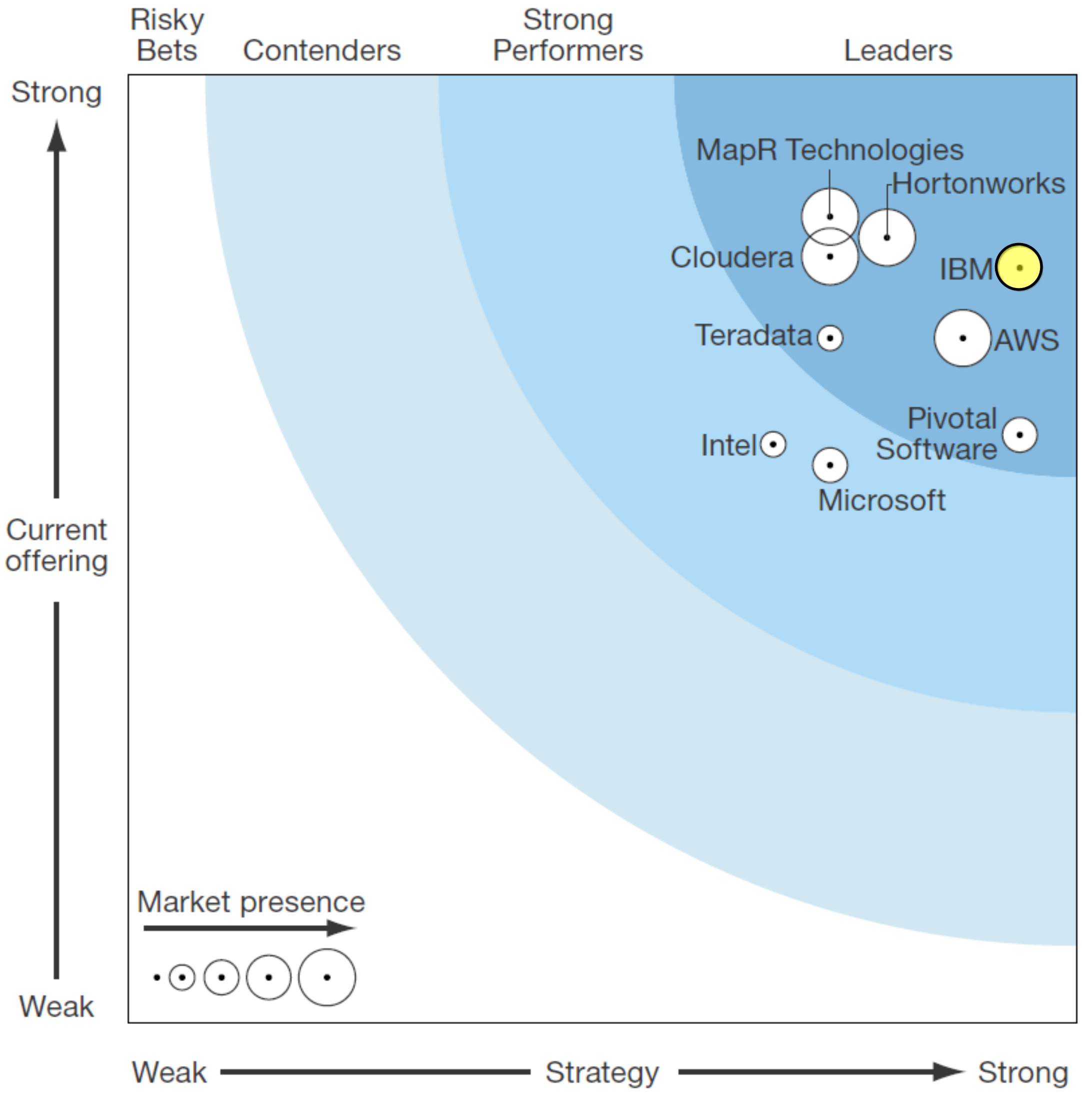
Just because you *can* build a solution from scratch doesn't mean you *should!*

InfoSphere BigInsights – Unique capabilities

Reduce time to market, increase customer value

Software Capabilities	Other Hadoop	InfoSphere BigInsights
Open Source Hadoop	✓	✓
Rich SQL on Hadoop – Big SQL	*Varying capabilities*	✓
Tools for business users - BigSheets	-	✓
Advanced Text Analytics	-	✓
In-Hadoop Analytics	-	✓
Rich Developer tools	-	✓
Enterprise-grade workload & storage mgmt.	-	✓

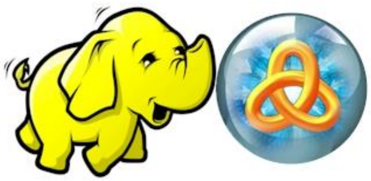
The Forrester Wave™ - Hadoop Solutions Q1 2014



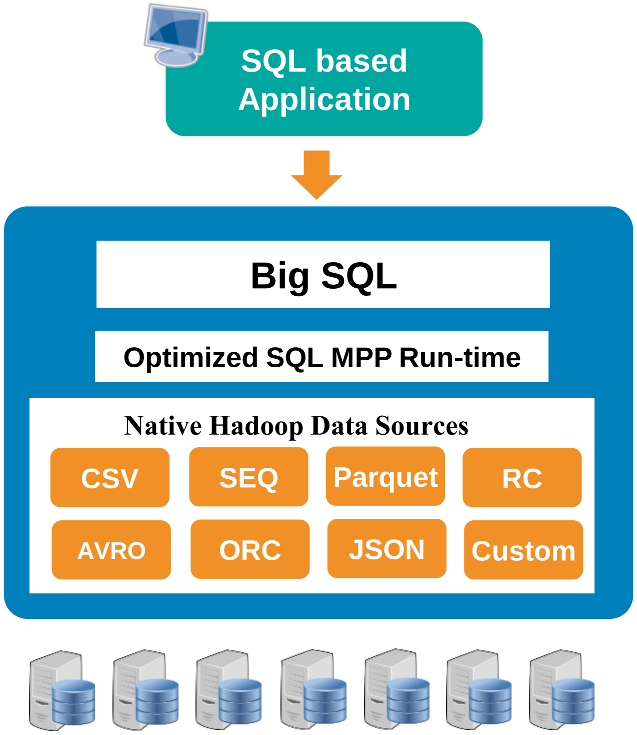
http://www.forrester.com/pimages/rws/reprints/document/112461/oid/1-PBE69P

Time to value - Big SQL

Big SQL = Big Investment Protection



- Rich ANSI SQL support
- Native data sources - No proprietary metadata
- Federate multiple EDW platforms - Teradata, Oracle, DB2 etc..
- Outstanding performance
- Rick analytic functions
- Multi-platform
- Security built-in



Time to value – text analytics

Translating textual information to actionable insights at scale

Customer: I’m calling because I received an incorrect bill. I just paid my bill two days ago, and my payment is not reflected

Agent: Sorry for the inconvenience. May I have your Account Number, please?

Customer: 15635764 – wait – I meant 15365764

Agent: For verification purposes, can I get your name and birth date?

Customer: Marge Simpson, Nov 23, 1975 and the account is under my Husband’s name, Homer

Agent: Thank you for that information. Per our system, you did pay your bill last Aug. 12

```
<call _center_record trans_id=132436>
      <cust_id>15365764</cust_id>
      <account_holder>
                Homer Simpson
      </account_holder>
      <caller_birthdate>
                1975-11-23
      </caller_birthdate>
      <inquiry>balance</inqury>
      <balance>0</balance>
      <pmt_date>2014-08-12</pmt_date>
      <cred_score>3.9</cred_score>
      ..
      ..
</call_center_record>
```

Conclusion

Get to value faster with a modernized data warehouse

- A broad information management portfolio
- Open, Enterprise-grade Hadoop
- Rich information governance
- Analytic tooling
- Unparalleled expertise

Download our free QuickStart Edition – http://ibm.co/quickstart

Questions?

Contact Information

If you have further questions or comments:

Philip Russom, TDWI
prussom@tdwi.org

Gord Sissons, IBM
gsissons@ca.ibm.com