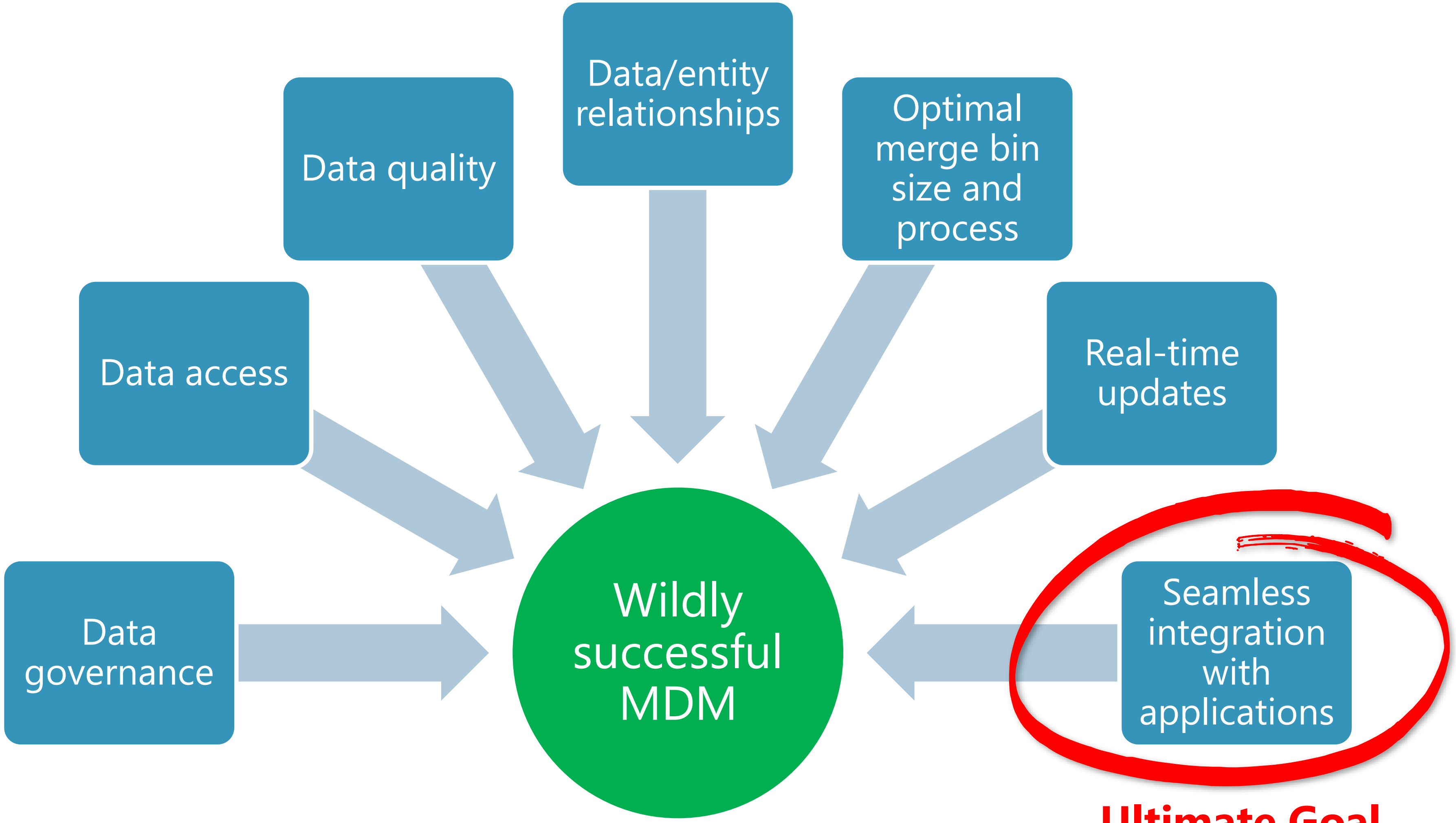
SmartData Fabric® Configuration for Registry*, Repository or Hybrid** Master Data Management (MDM)

*aka Federated or Virtual
**aka Coexistence

January 2024

Seven keys to successful MDM



Data governance

Data access

Data quality

Data/entity relationships

Optimal merge bin size and process

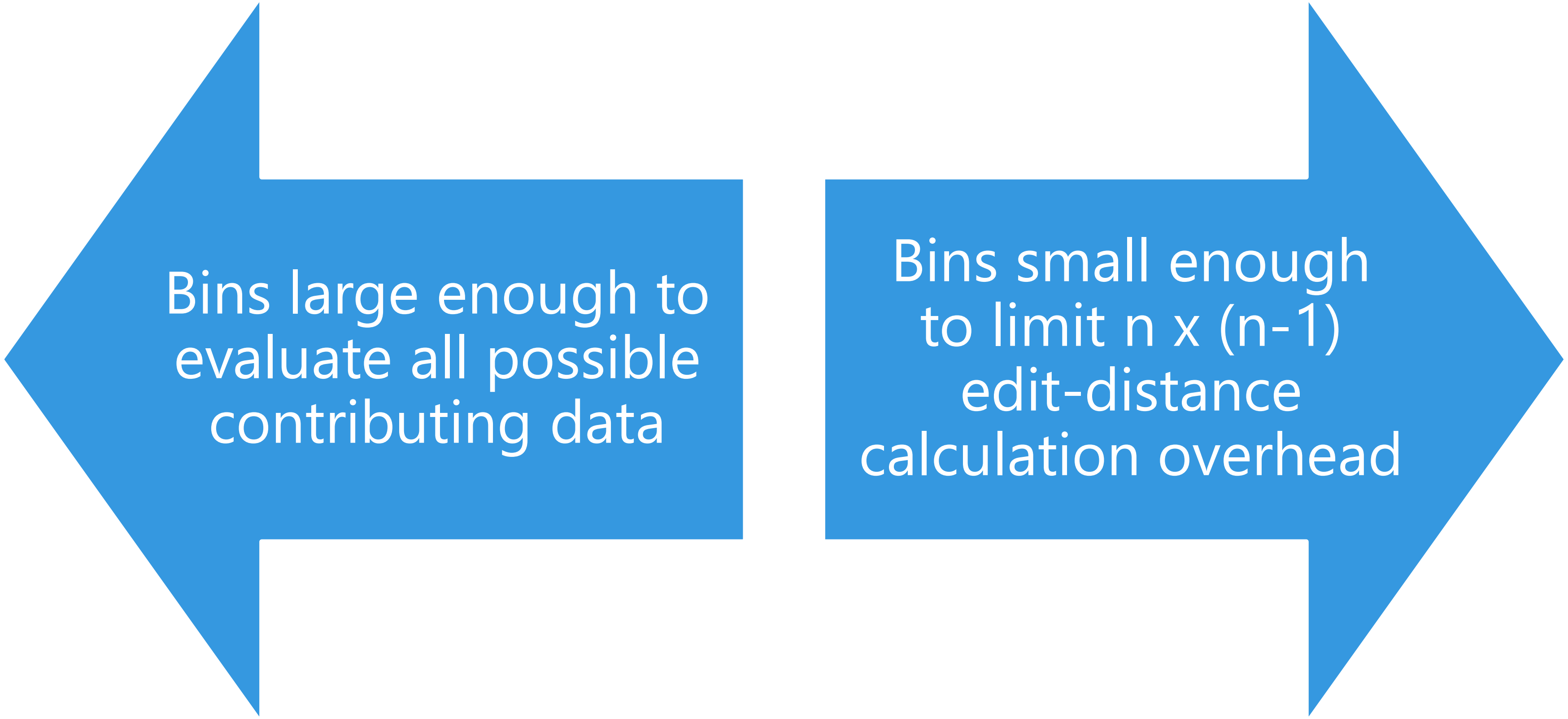
Real-time updates

Seamless integration with applications

Wildly successful MDM

Ultimate Goal

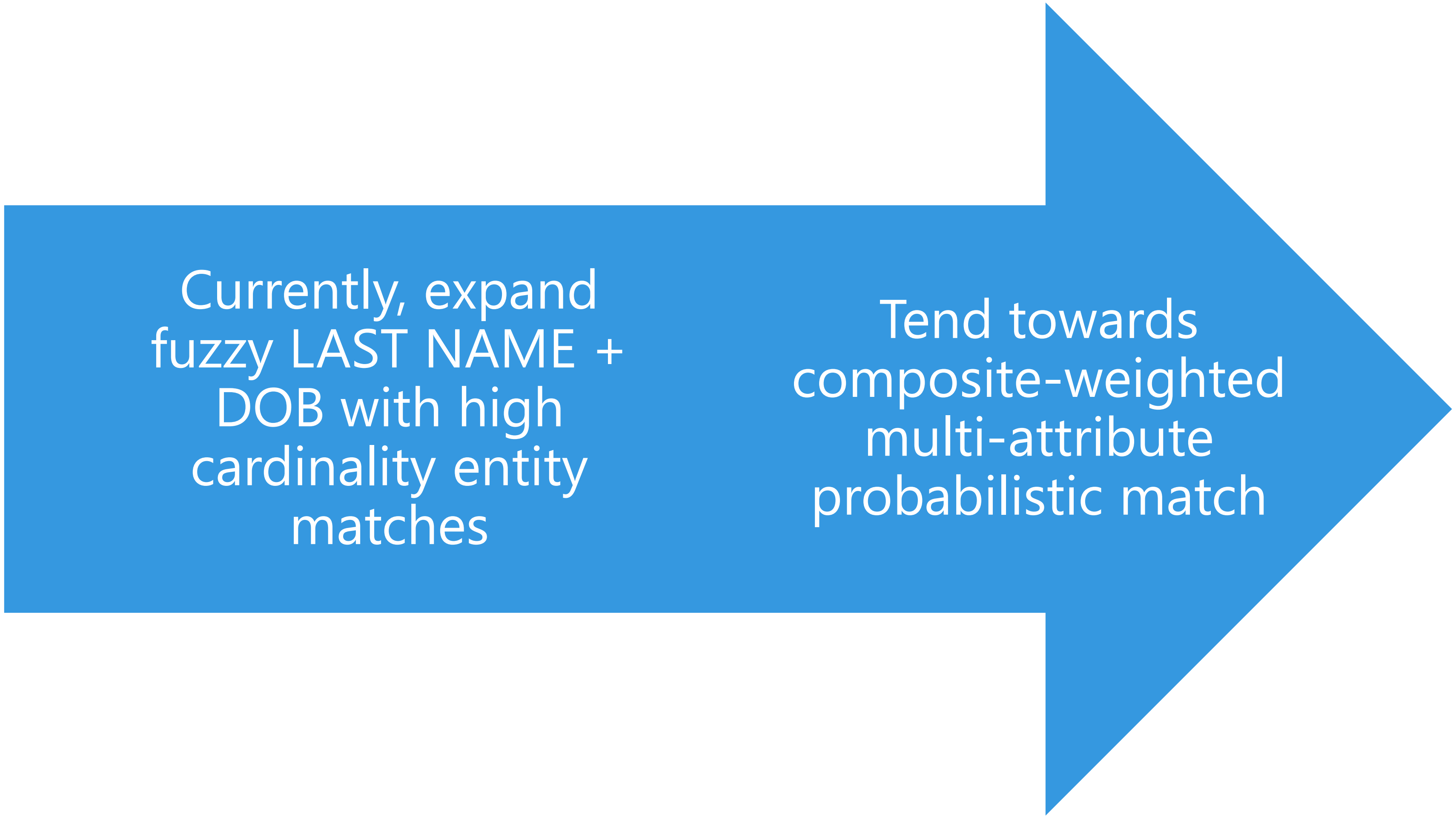
Conflicting goals with bins



Bins large enough to evaluate all possible contributing data

Bins small enough to limit n x (n-1) edit-distance calculation overhead

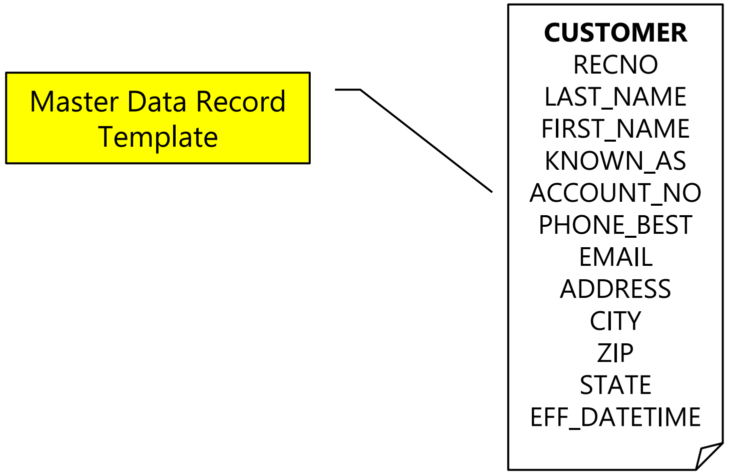
Binning Methods



Currently, expand fuzzy LAST NAME + DOB with high cardinality entity matches

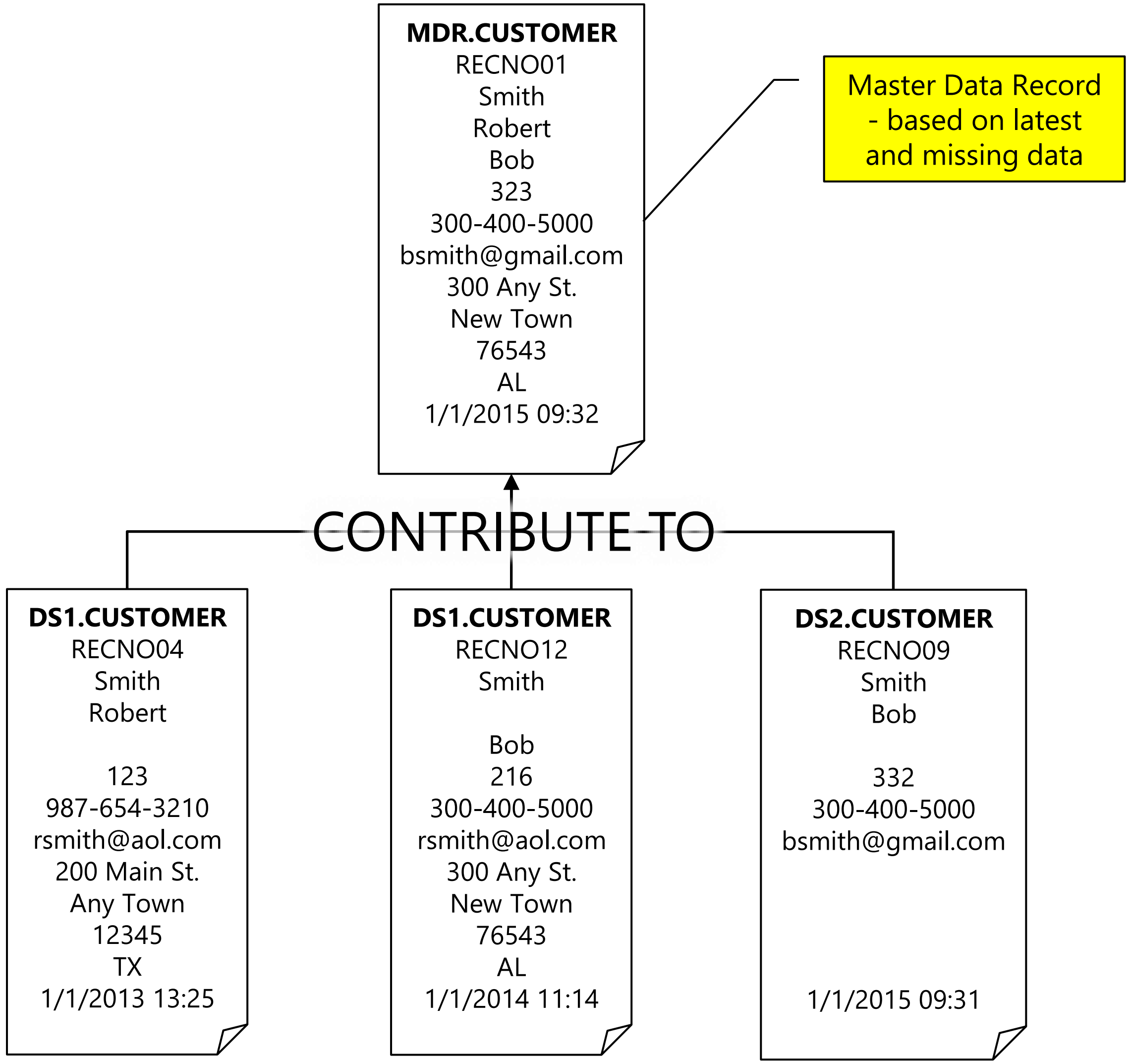
Tend towards composite-weighted multi-attribute probabilistic match

How to retain a master data record (MDR)?



Master Data Record Template

CUSTOMER
RECNO
LAST_NAME
FIRST_NAME
KNOWN_AS
ACCOUNT_NO
PHONE_BEST
EMAIL
ADDRESS
CITY
ZIP
STATE
EFF_DATETIME



MDR.CUSTOMER
RECNO01
Smith
Robert
Bob
323
300-400-5000
bsmith@gmail.com
300 Any St.
New Town
76543
AL
1/1/2015 09:32

Master Data Record - based on latest and missing data

DS1.CUSTOMER
RECNO04
Smith
Robert

123
987-654-3210
rsmith@aol.com
200 Main St.
Any Town
12345
TX
1/1/2013 13:25

DS1.CUSTOMER
RECNO12
Smith

Bob
216
300-400-5000
rsmith@aol.com
300 Any St.
New Town
76543
AL
1/1/2014 11:14

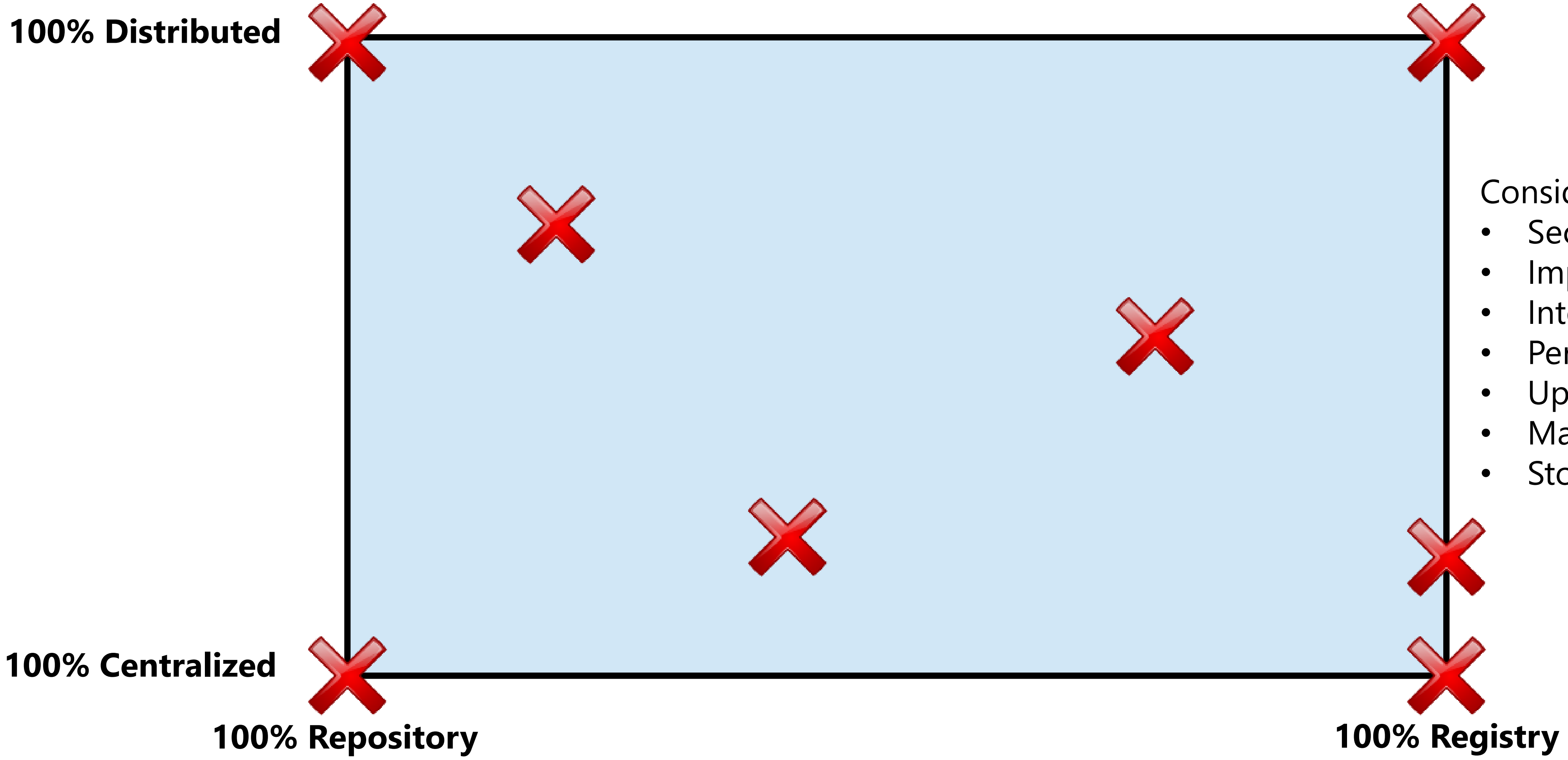
DS2.CUSTOMER
RECNO09
Smith
Bob

332
300-400-5000
bsmith@gmail.com

1/1/2015 09:31

Centralized or distributed?
Repository or registry?

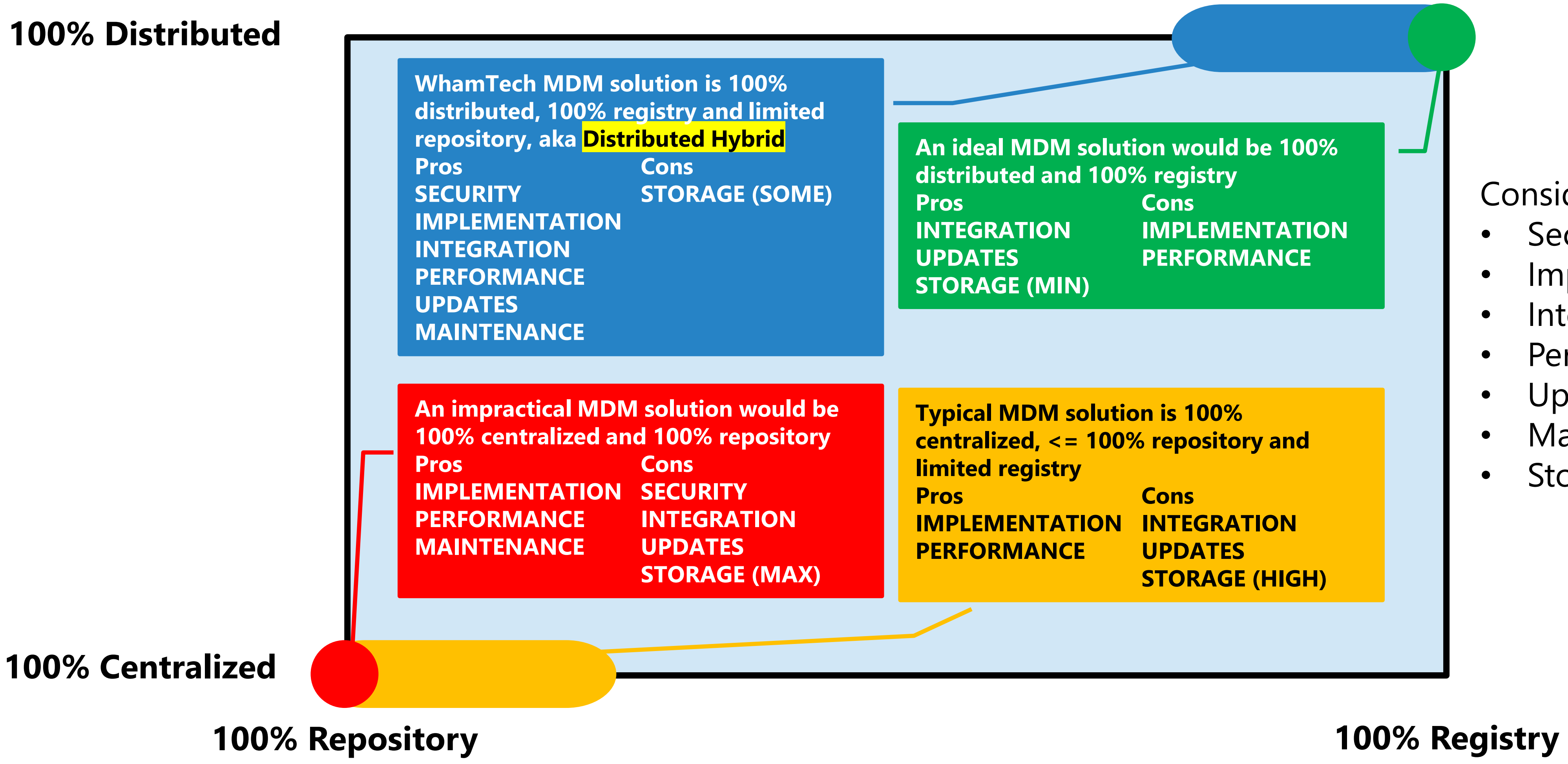
SmartData Fabric® can accommodate any combination



Considerations:

- Security
- Implementation
- Integration
- Performance
- Updates (data)
- Maintenance (system)
- Storage

SmartData Fabric® tends towards a Distributed Hybrid solution



WhamTech MDM solution is 100% distributed, 100% registry and limited repository, aka **Distributed Hybrid**

Pros	Cons
SECURITY	STORAGE (SOME)
IMPLEMENTATION	
INTEGRATION	
PERFORMANCE	
UPDATES	
MAINTENANCE	

An ideal MDM solution would be 100% distributed and 100% registry

Pros	Cons
INTEGRATION	IMPLEMENTATION
UPDATES	PERFORMANCE
STORAGE (MIN)	

An impractical MDM solution would be 100% centralized and 100% repository

Pros	Cons
IMPLEMENTATION	SECURITY
PERFORMANCE	INTEGRATION
MAINTENANCE	UPDATES
	STORAGE (MAX)

Typical MDM solution is 100% centralized, <= 100% repository and limited registry

Pros	Cons
IMPLEMENTATION	INTEGRATION
PERFORMANCE	UPDATES
	STORAGE (HIGH)

Considerations:

- Security
- Implementation
- Integration
- Performance
- Updates (data)
- Maintenance (system)
- Storage

Benefits of a Distributed Hybrid* MDM Approach

Distributed Hybrid*

- Data security and privacy, e.g., PII, personal and master data stays in the same environment as sources
- Adapter-level rules can integrate master data automatically – no application modifications
- Local processing - fast and secure
- Use master data at multiple federation levels
- Add/remove data source adapters with little or no impact
- No central bottleneck or dependencies
- Can still be centrally managed and consolidated

Limited Repository

- Best master data may not exist as such in data sources, so, need to store best data
- Master data immediately available – avoids (1) access to multiple sources and (2) source data transformation
- Allows retention of historic master data, regardless of source data changes

Full Registry

- Updateable
- Traceable
- Avoids repository for all data associated with master data

*Hybrid = Limited Repository and Full Registry

SmartData Fabric®

MDM OPTIONAL ADD-ON

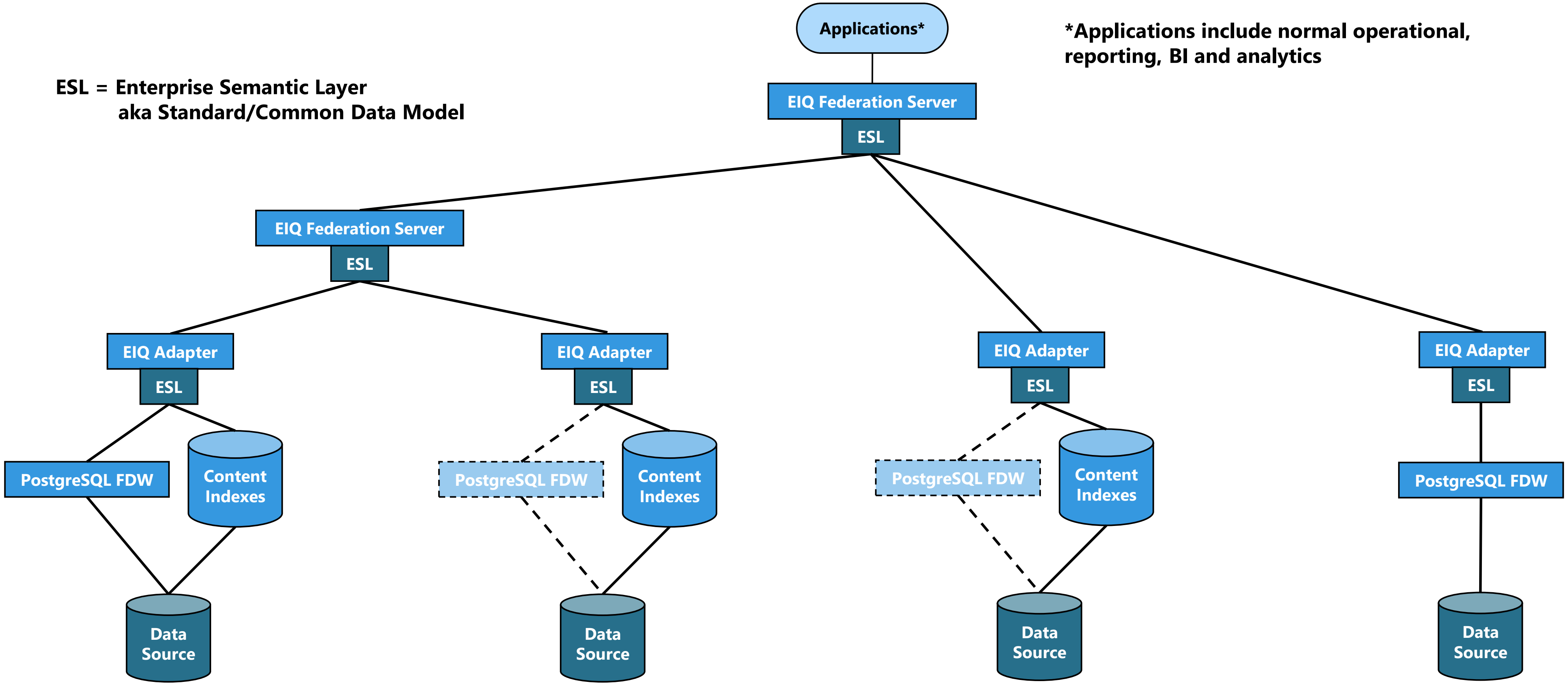
Can also use a third-party MDM system

Increasingly seen as essential to almost every solution, as bridges the gap between operational/transactional data and virtual data warehouse-type views

Basic SmartData Fabric® Configuration

ESL = Enterprise Semantic Layer
aka Standard/Common Data Model

*Applications include normal operational, reporting, BI and analytics



Hybrid adapter queries indexes for some data, e.g., personal, and data source directly for other data, e.g., transactional

Indexed adapter queries indexes for some or all data, and optionally, data source directly for other data and/or drilldown

Indexed adapter

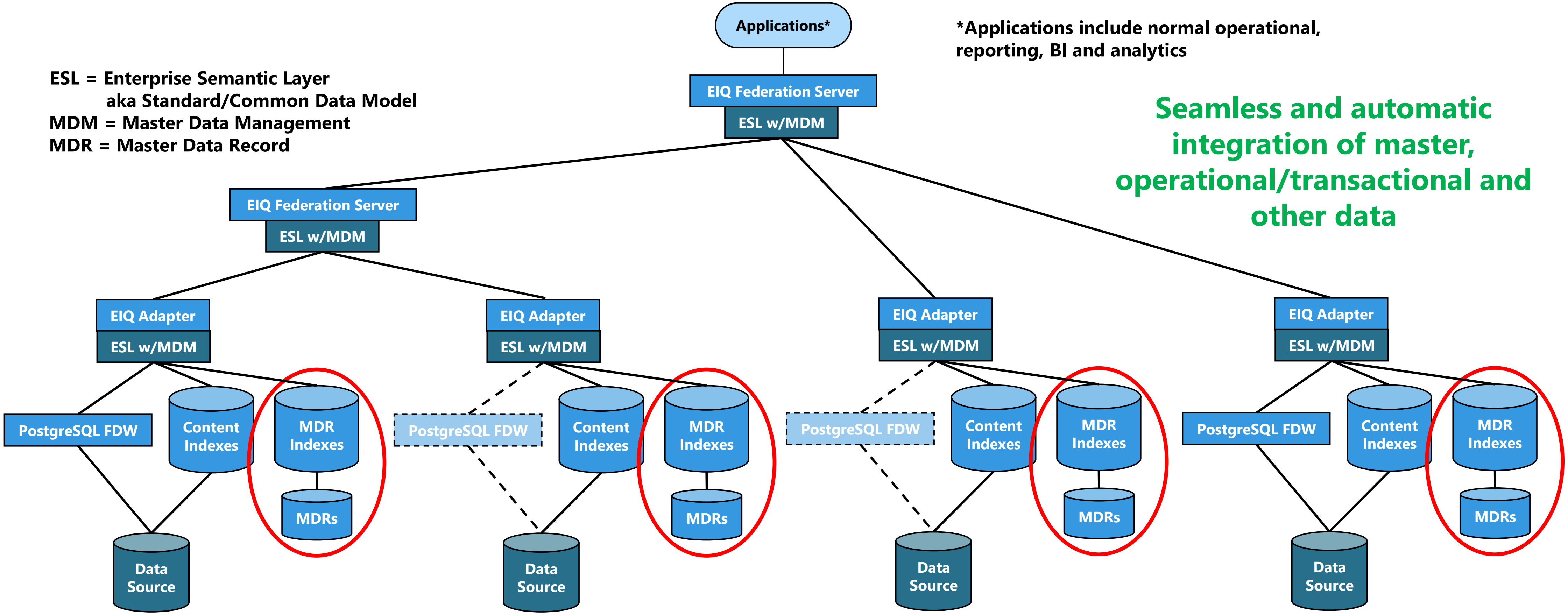
Conventional adapter has no indexes and queries data source directly for all data

Master data available at all levels – update and propagate

ESL = Enterprise Semantic Layer
aka Standard/Common Data Model
MDM = Master Data Management
MDR = Master Data Record

*Applications include normal operational, reporting, BI and analytics

Seamless and automatic integration of master, operational/transactional and other data



Hybrid adapter

Indexed adapter

Indexed adapter

Conventional adapter -> hybrid adapter queries MDR indexes for MDRs, e.g., personal and other entities, content indexes for MDR-associated data and data source directly for all other data

MDM processing scalability

- Challenge #1: Increased number of data sources
    - Response #1:
      This does not impact the federation MDM/query logic and hundreds and more of sources can be accessed through federation
- Challenge #2: Increased number of source records and MDM records
    - Response #2: MDM records are generally small in number and size in source data, as most data is transaction data
    - Response #2 contd.: WhamTech (i) leverages techniques such as smart binning and other MDM strategies, (ii) distributed query processing and storage federated across multiple adapters, and (iii) continuous near real-time incremental processing to update MDM records as source data changes, so less computing resources required than large batch processing
- Challenge #3: Reconciling data source schema differences due to increased number of sources
    - Response #3: Schema differences are harmonized/accommodated upfront at each adapter federation level through mapping to a common data model, e.g., FHIR HL7, regardless of data source schemas, so would not have any significant impact on federation and MDM processing

The End

APPENDICES

What is distributed hybrid master data?

How to manage distributed hybrid master data?

How to use distributed hybrid master data?

How do SmartData Fabric® and MDM processes combine?

Two main forms of MDR – repository and registry

Hybrid master data record creation

Appendix: What is distributed hybrid master data?

Distributed hybrid master data description (1 of 2)

- Master data is either an extension, or part, of the Enterprise Semantic Layer (ESL)
  - Master data can be defined separately from normal content data in ESL, or
  - Normal content data in ESL also tagged as master data, which may also be tagged as link entity data in the future
- One or more master data entities are defined and used to find any and all associated source and external data
  - Example master data entities are PERSON (e.g., patient, doctor, nurse and administrator), ORGANIZATION (e.g., hospital, lab, insurance co., insured co. and healthcare provider), ADDRESS, PHONE, EMAIL, SYMPTOMS, AILMENT, TREATMENT, OUTCOME, etc.
- Master data, e.g., PERSON, can comprise multiple attributes, aka “complex entities”, that may also be entities, e.g., LNAME, FNAME, MNAME, DOB, SSN, PHONE, EMAIL and ADDRESS, aka “simple entities”
- Any source or master data can be masked, tokenized or encrypted, either dynamically or in indexes, including format-preserved encryption (FPE), which is a combination of tokenization and encryption

Distributed hybrid master data description (2 of 2)

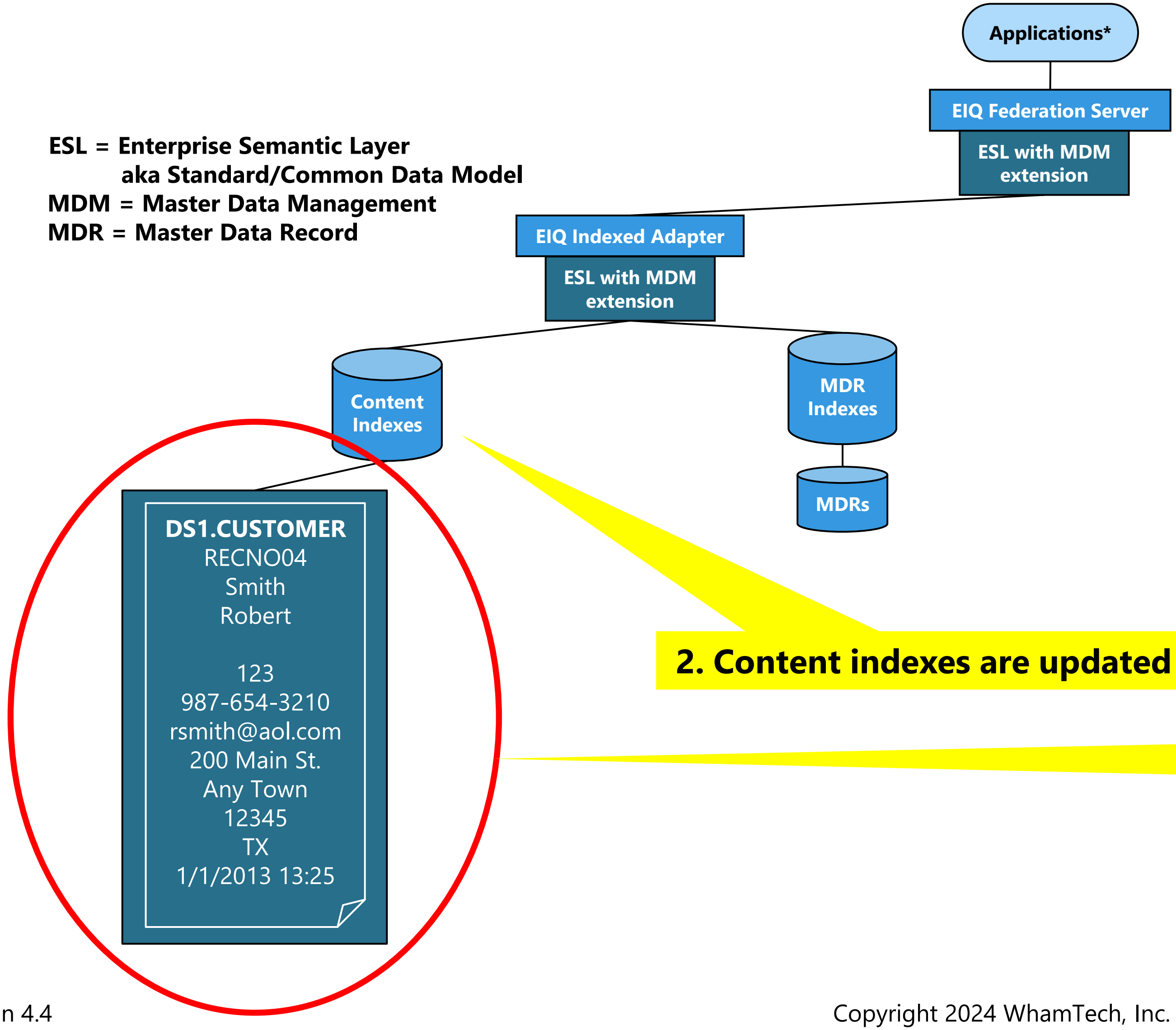
- Access to any data, including master data, should depend on access control permissions
- Dynamic masking, tokenization or encryption may also depend on access control permissions
- Encryption keys can be specific to an individual entity or entities, e.g., PERSON or others, and may be separately passed from result-sets to the recipient(s) for decryption
- Hybrid master data table contains values, any phonetic tokens (for fuzzy match), links to records containing values and date-time of when values were either added to data sources, indexes (assuming near real-time indexing) or master data table
  - Allows for rapid real-time updates as master data tables are all that is needed – no need to query data sources
- All entries in master data tables are indexed, queryable and can be joined to normal content indexes and, in the future, Link Indexes™, for a seamless integration of master data and source data

Appendix: How to maintain distributed hybrid master data?

Managing distributed hybrid master data (1 of 8)

*Applications include normal operational, BI and analytics

ESL = Enterprise Semantic Layer
aka Standard/Common Data Model
MDM = Master Data Management
MDR = Master Data Record



DS1.CUSTOMER
RECNO04
Smith
Robert

123
987-654-3210
rsmith@aol.com
200 Main St.
Any Town
12345
TX
1/1/2013 13:25

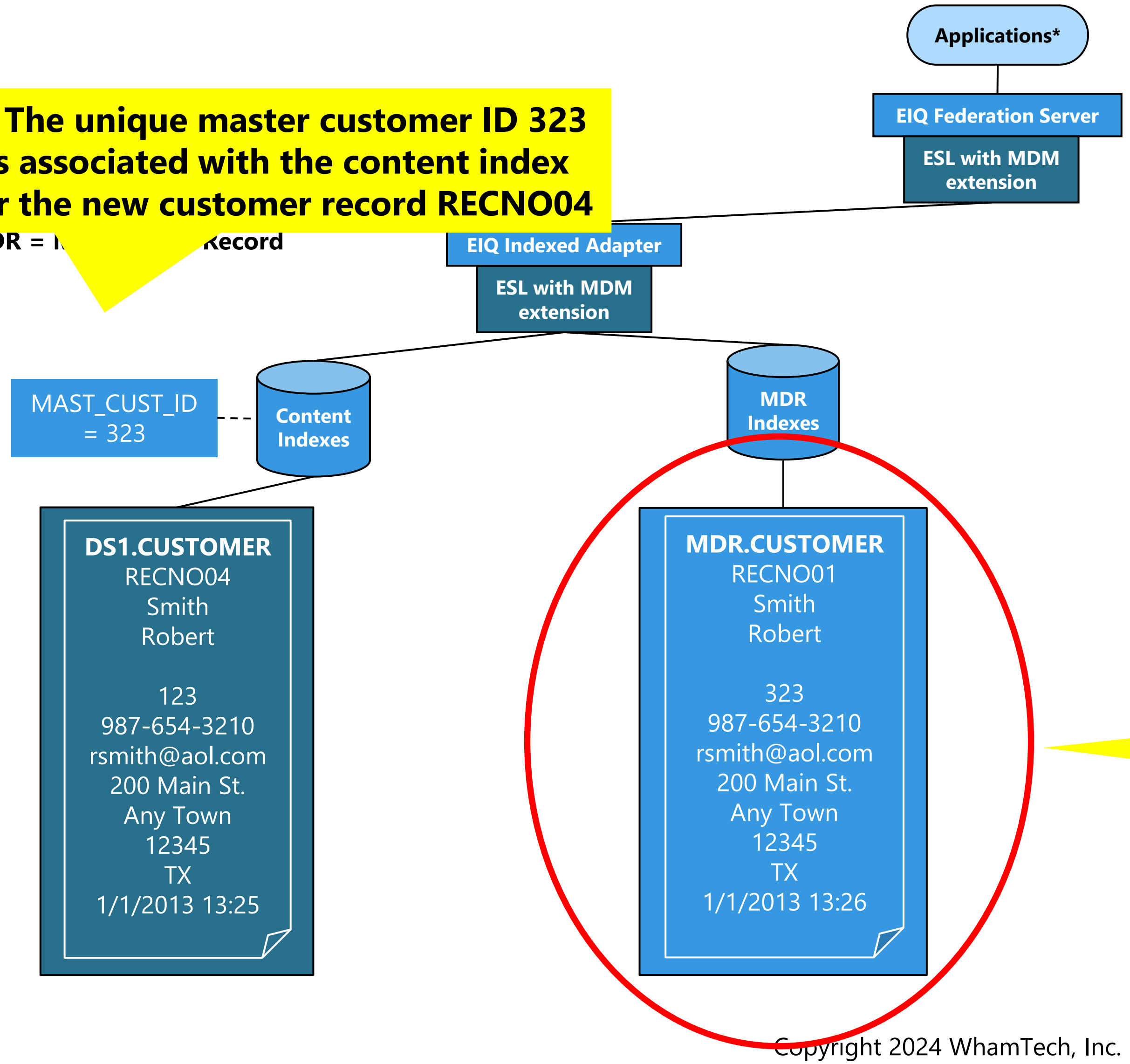
3. A check is made across all data sources, incl. DS1, for a match in master data and if there is no obvious match, i.e., this appears to be a new customer to any data source, then that info is passed back to the DS1 adapter

2. Content indexes are updated

1. A customer interacts with one of the organization's contact points and a record RECNO04 is created in DS1

Managing distributed hybrid master data (2 of 8)

*Applications include normal operational, BI and analytics



Applications*

EIQ Federation Server

ESL with MDM extension

EIQ Indexed Adapter

ESL with MDM extension

MDR = [illegible] Record

MAST_CUST_ID = 323

Content Indexes

MDR Indexes

DS1.CUSTOMER
RECNO04
Smith
Robert

123
987-654-3210
rsmith@aol.com
200 Main St.
Any Town
12345
TX
1/1/2013 13:25

MDR.CUSTOMER
RECNO01
Smith
Robert

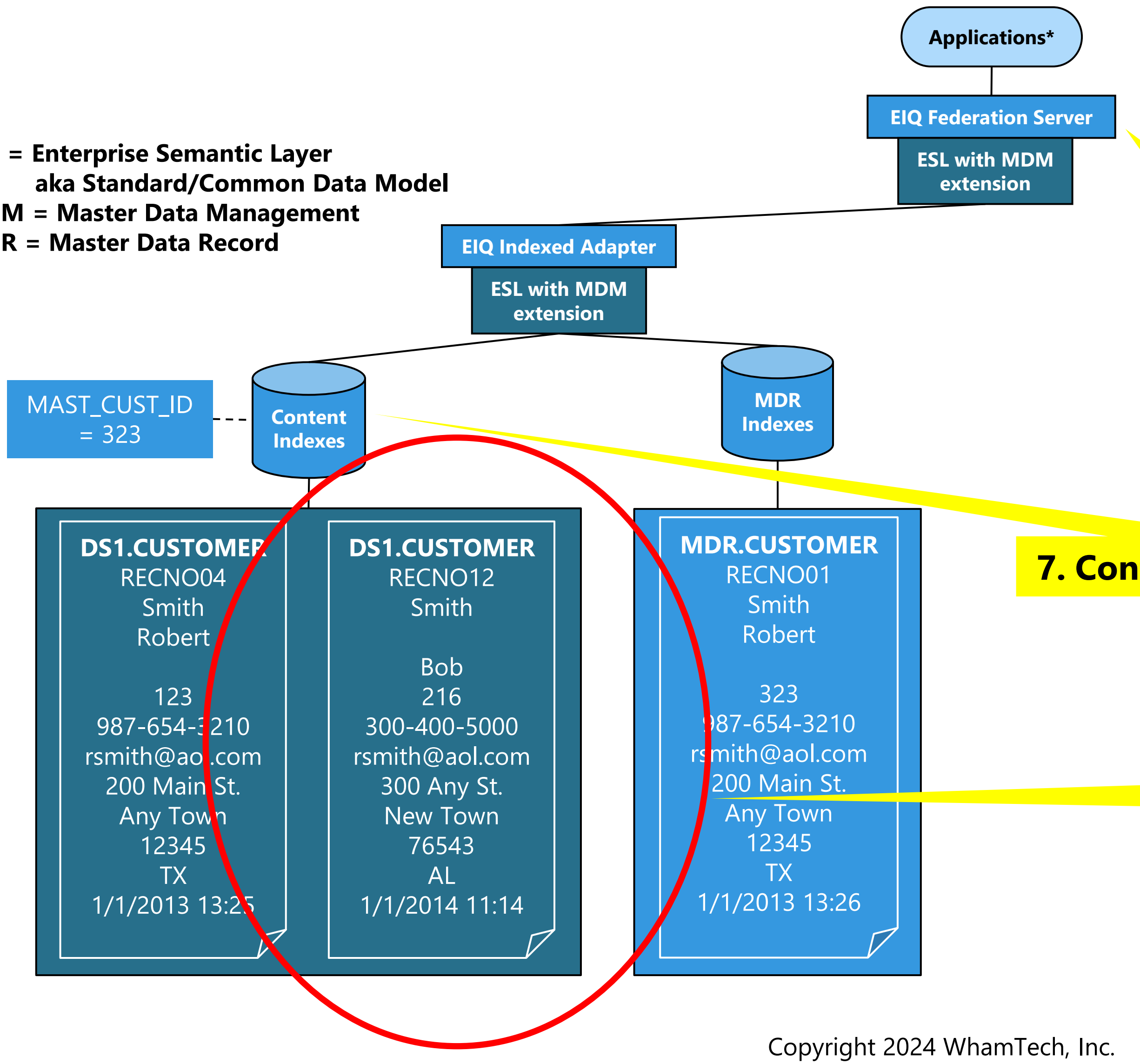
323
987-654-3210
rsmith@aol.com
200 Main St.
Any Town
12345
TX
1/1/2013 13:26

5. The unique master customer ID 323 is associated with the content index for the new customer record RECNO04

4. A MDR with an unique master customer ID 323 is created, stored and indexed for DS1 – no other data sources

Managing distributed hybrid master data (3 of 8)

ESL = Enterprise Semantic Layer
aka Standard/Common Data Model
MDM = Master Data Management
MDR = Master Data Record



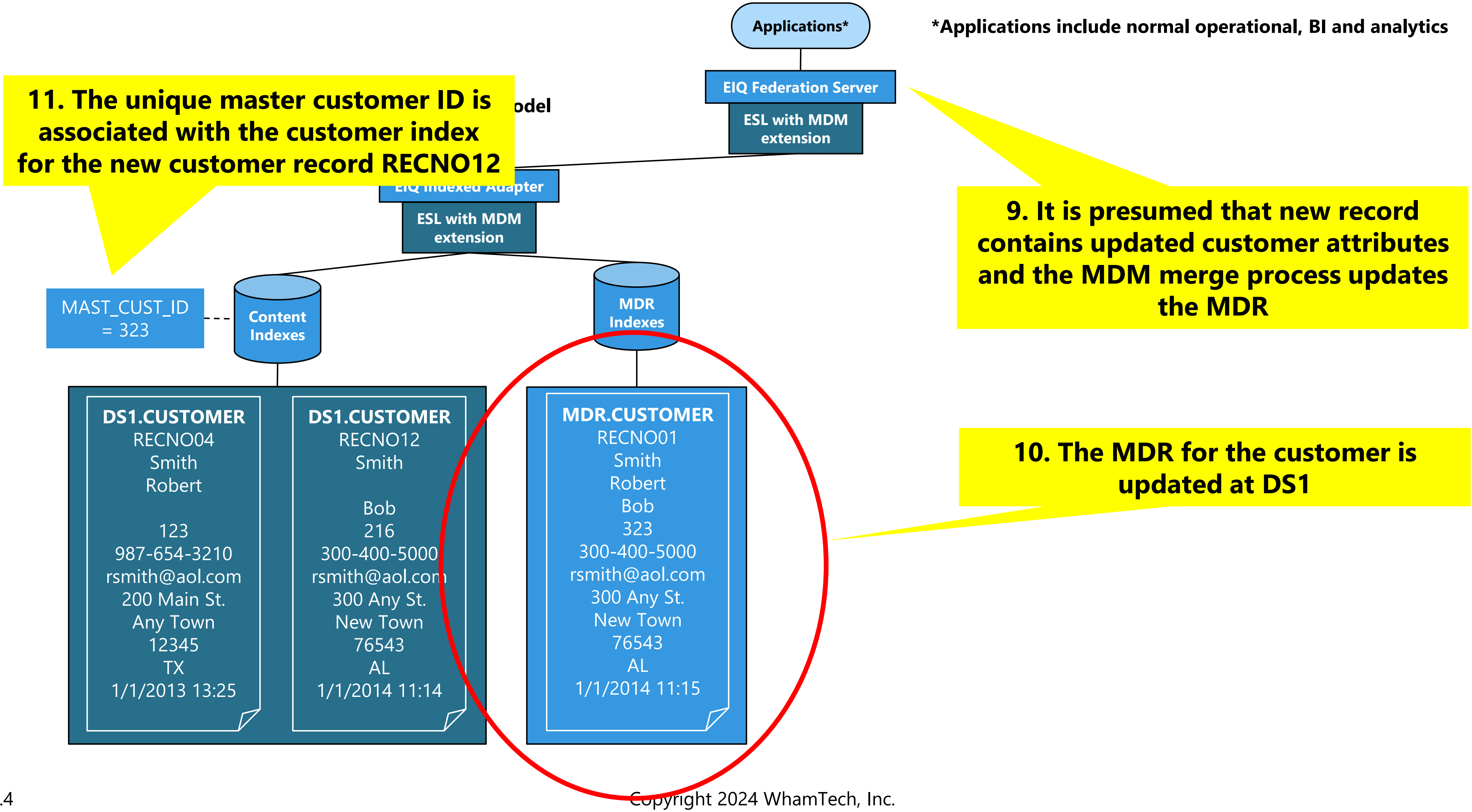
*Applications include normal operational, BI and analytics

8. A check is made across all data sources, incl. DS1, for a match in master data and there is a match, i.e., this appears to be an existing customer in DS1, then all customer data from DS1 is read and

7. Content indexes are updated

6. One year later, the same customer interacts with DS1 again

Managing distributed hybrid master data (4 of 8)



*Applications include normal operational, BI and analytics

DS1.CUSTOMER
RECNO04
Smith
Robert

123
987-654-3210
rsmith@aol.com
200 Main St.
Any Town
12345
TX
1/1/2013 13:25

DS1.CUSTOMER
RECNO12
Smith

Bob
216
300-400-5000
rsmith@aol.com
300 Any St.
New Town
76543
AL
1/1/2014 11:14

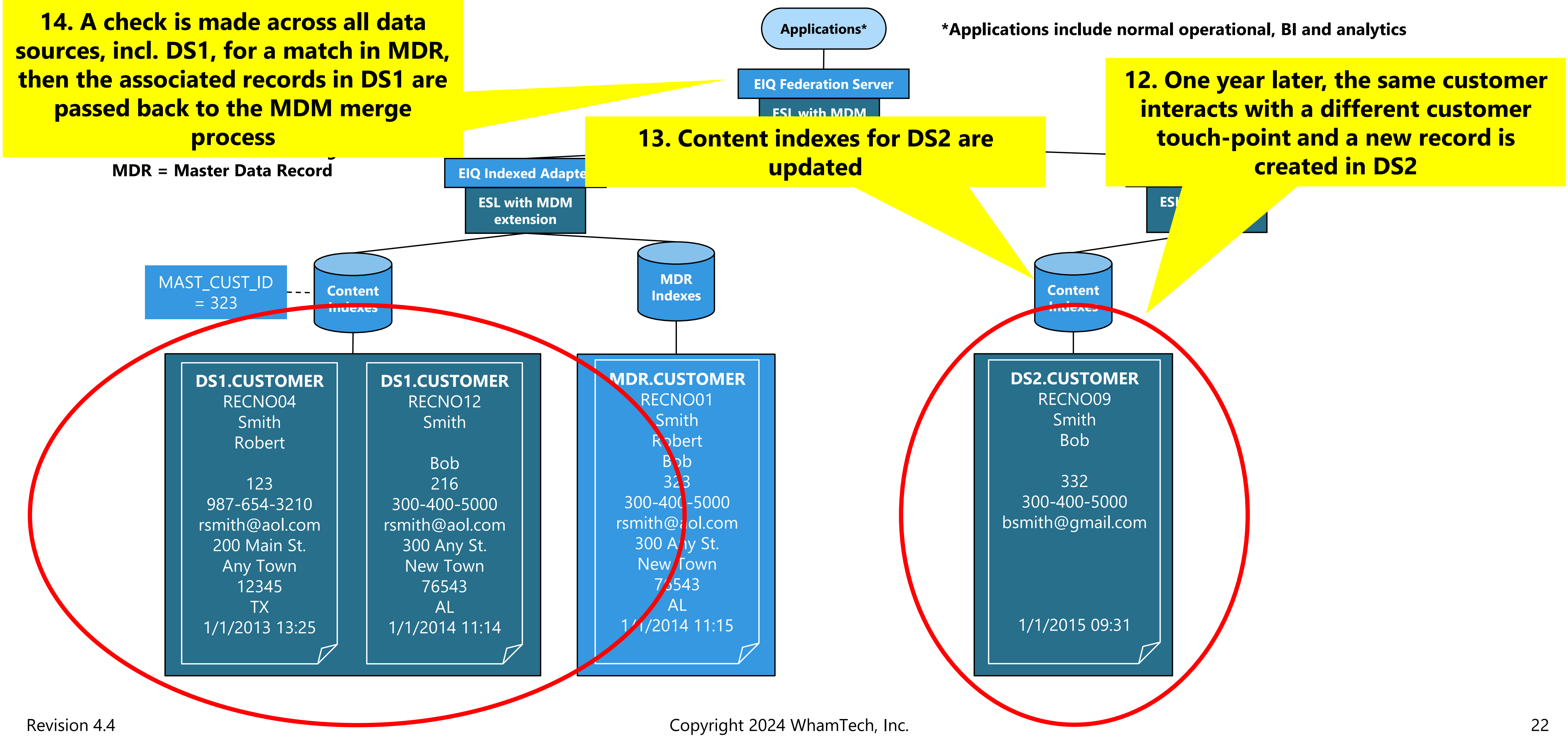
MDR.CUSTOMER
RECNO01
Smith
Robert
Bob
323
300-400-5000
rsmith@aol.com
300 Any St.
New Town
76543
AL
1/1/2014 11:15

9. It is presumed that new record contains updated customer attributes and the MDM merge process updates the MDR

10. The MDR for the customer is updated at DS1

11. The unique master customer ID is associated with the customer index for the new customer record RECNO12

Managing distributed hybrid master data (5 of 8)



14. A check is made across all data sources, incl. DS1, for a match in MDR, then the associated records in DS1 are passed back to the MDM merge process

13. Content indexes for DS2 are updated

12. One year later, the same customer interacts with a different customer touch-point and a new record is created in DS2

*Applications include normal operational, BI and analytics

MDR = Master Data Record

DS1.CUSTOMER
RECNO04
Smith
Robert

123
987-654-3210
rsmith@aol.com
200 Main St.
Any Town
12345
TX
1/1/2013 13:25

DS1.CUSTOMER
RECNO12
Smith

Bob
216
300-400-5000
rsmith@aol.com
300 Any St.
New Town
76543
AL
1/1/2014 11:14

MDR.CUSTOMER
RECNO01
Smith
Robert
Bob
323
300-400-5000
rsmith@aol.com
300 Any St.
New Town
76543
AL
1/1/2014 11:15

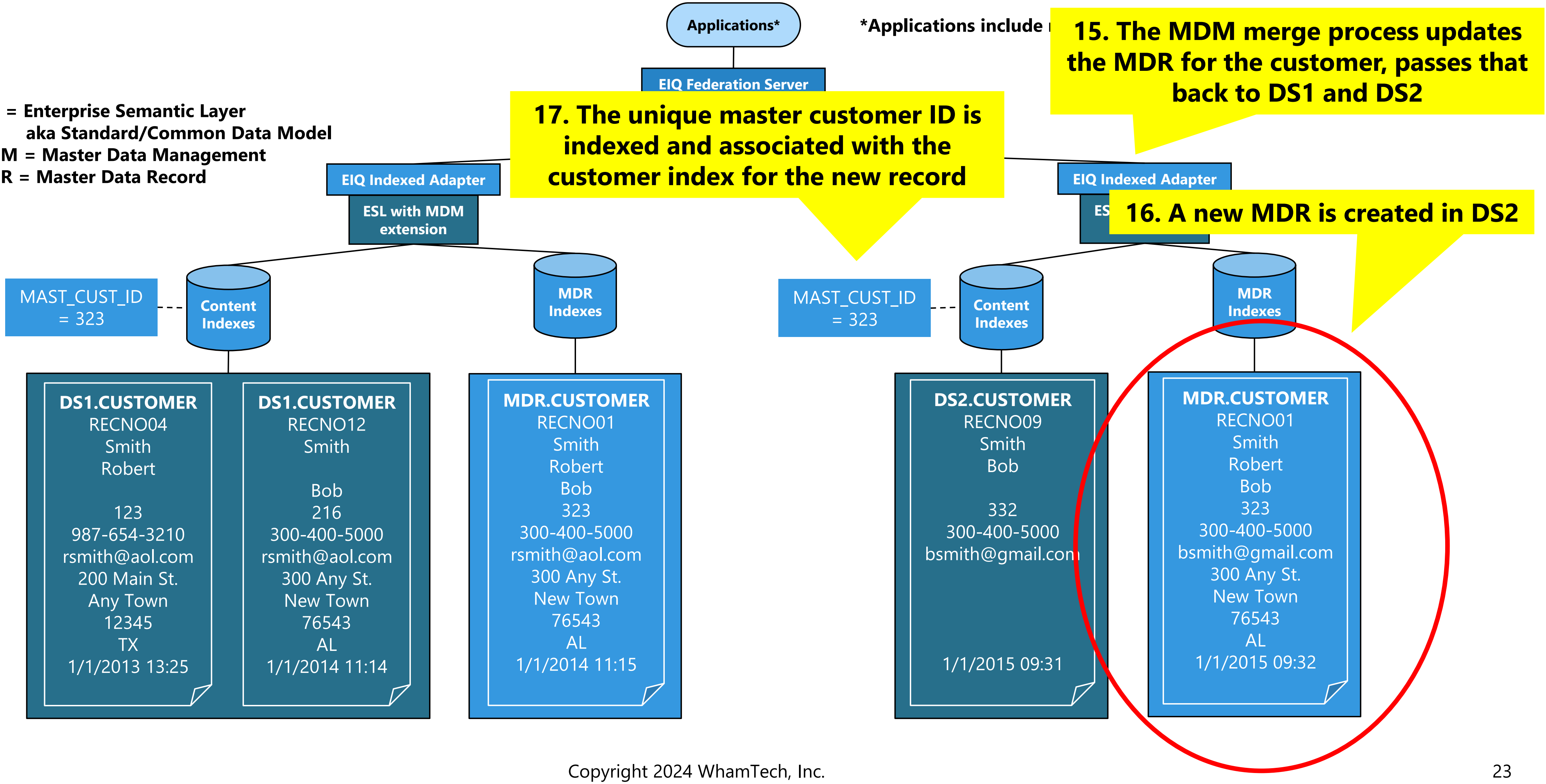
DS2.CUSTOMER
RECNO09
Smith
Bob

332
300-400-5000
bsmith@gmail.com

1/1/2015 09:31

Managing distributed hybrid master data (7 of 8)

ESL = Enterprise Semantic Layer
aka Standard/Common Data Model
MDM = Master Data Management
MDR = Master Data Record



15. The MDM merge process updates the MDR for the customer, passes that back to DS1 and DS2

17. The unique master customer ID is indexed and associated with the customer index for the new record

16. A new MDR is created in DS2

DS1.CUSTOMER
RECNO04
Smith
Robert

123
987-654-3210
rsmith@aol.com
200 Main St.
Any Town
12345
TX
1/1/2013 13:25

DS1.CUSTOMER
RECNO12
Smith

Bob
216
300-400-5000
rsmith@aol.com
300 Any St.
New Town
76543
AL
1/1/2014 11:14

MDR.CUSTOMER
RECNO01
Smith
Robert
Bob
323
300-400-5000
rsmith@aol.com
300 Any St.
New Town
76543
AL
1/1/2014 11:15

DS2.CUSTOMER
RECNO09
Smith
Bob

332
300-400-5000
bsmith@gmail.com

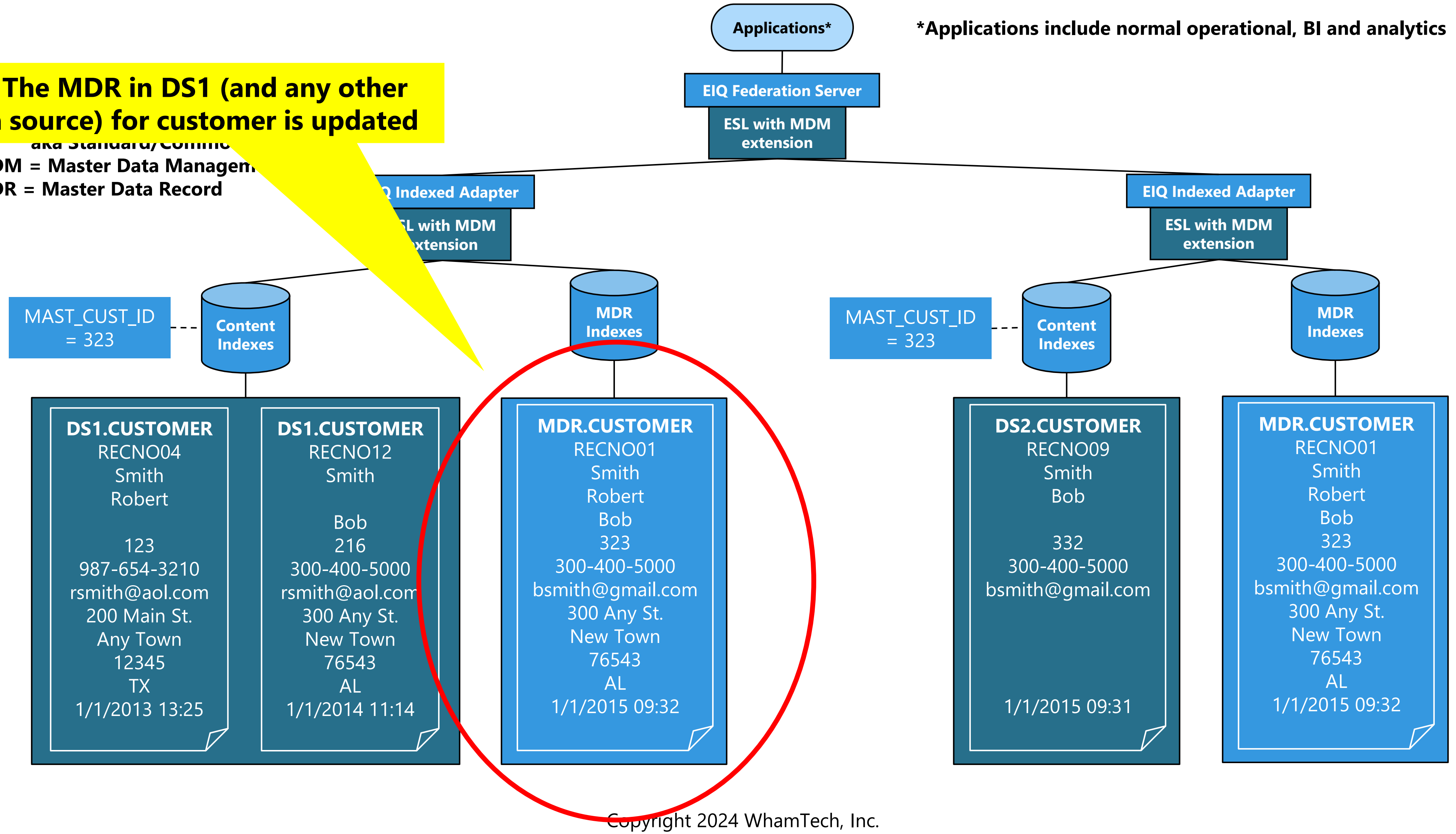
1/1/2015 09:31

MDR.CUSTOMER
RECNO01
Smith
Robert
Bob
323
300-400-5000
bsmith@gmail.com
300 Any St.
New Town
76543
AL
1/1/2015 09:32

Managing distributed hybrid master data (8 of 8)

18. The MDR in DS1 (and any other data source) for customer is updated

aka Standard/Commo
MDM = Master Data Managem
MDR = Master Data Record



DS1.CUSTOMER
RECNO04
Smith
Robert

123
987-654-3210
rsmith@aol.com
200 Main St.
Any Town
12345
TX
1/1/2013 13:25

DS1.CUSTOMER
RECNO12
Smith

Bob
216
300-400-5000
rsmith@aol.com
300 Any St.
New Town
76543
AL
1/1/2014 11:14

MDR.CUSTOMER
RECNO01
Smith
Robert
Bob
323
300-400-5000
bsmith@gmail.com
300 Any St.
New Town
76543
AL
1/1/2015 09:32

DS2.CUSTOMER
RECNO09
Smith
Bob

332
300-400-5000
bsmith@gmail.com

1/1/2015 09:31

MDR.CUSTOMER
RECNO01
Smith
Robert
Bob
323
300-400-5000
bsmith@gmail.com
300 Any St.
New Town
76543
AL
1/1/2015 09:32

***Applications include normal operational, BI and analytics**

Appendix: How to use distributed hybrid master data?

Using distributed hybrid master data (1 of 2)

- Query master data indexes and records to access and read source data
  - Typically, results are a combination of source and a small amount of master data, where raw source data is substituted by its respective master data
  - Optionally, use source data only - do not replace any of it with its respective master data
  - Optionally, use both source and its respective master data
  - Regardless of source/master data combination, consolidate multiple data source results at higher federation levels
  - Master data is essential for almost any meaningful reporting, BI and analytics for virtual single patient index and other entity-centric views, marketing, CRM, etc.
- Query master data indexes to manage master data and/or populate a higher-level, e.g., central, master data repository
- Automatically work with existing applications without application modifications

Using distributed hybrid master data (2 of 2)

- Seamlessly and transparently integrate any and all data, including master, mainframe, operational, warehouse, cloud, partner, SaaS, government, etc.
- Include data sources without EIQ Adapters with indexes through EIQ Adapters without indexes, aka conventional, and other vendor conventional federated adapters and/or query engines
  - Have to potentially integrate or match independent third-party master data
- In the future, use results directly "as is" with Link Indexes™ for degrees of separation and link queries, graph database representation, link analysis and highly interactive graph/link visualization, for all types of analytics
  - Master data is required to represent multiple records of same entities as single nodes, e.g., PERSON, VEHICLE, ADDRESS, ORGANIZATION, etc.
  - Alleviates manual involvement – tends towards automation

Appendix: How do SmartData Fabric® and MDM processes combine?

SmartData Fabric® and MDM Processes (1 of 3)

Automate as a much as possible:

- Network asset/device discovery
- Data source discovery
- Data discovery
  - Optionally, in the future, with raw Link Indexes™ to automatically connect same data in multiple data sources
- Data profiling to develop data transforms for typos, transpositions and non-standard data, e.g., name, address, phone and email correction
  - Lookup dictionaries and thesauri
  - USPS and/or other address correction
  - Regular expressions and/or use open source dbt for data transforms

SmartData Fabric® and MDM Processes (2 of 3)

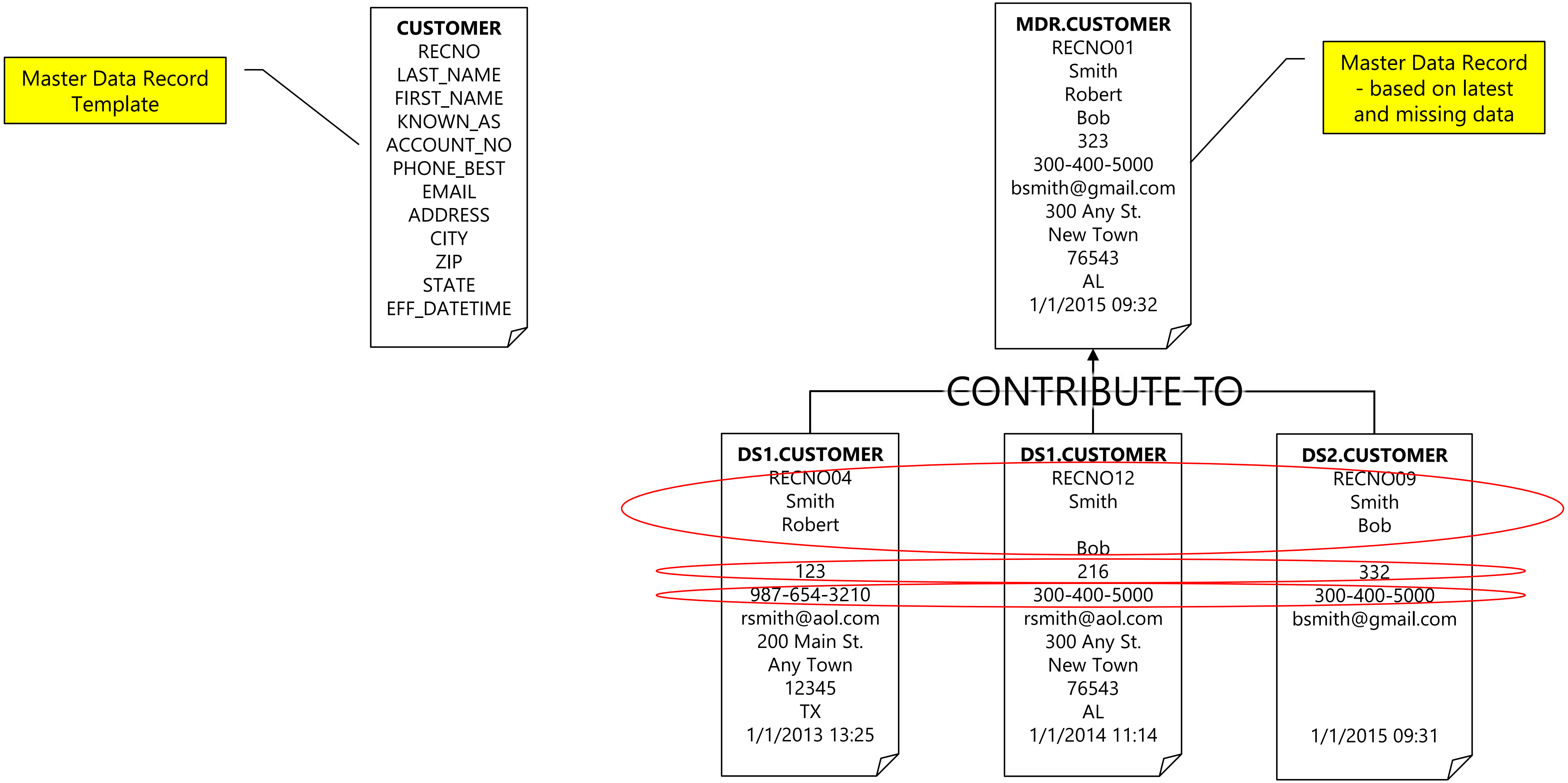
- Multiple indexes and types, e.g., basic content, aggregations, calculations, fuzzy, text, extracted entities, indexed views and, in the future, Link Indexes™
- MDM: Data source-specific tables containing unique indexed primary entity IDs, master data (including phonetic representations), links to source master data and date-time
  - Create with multi-attribute fuzzy match and master data rules, and maybe in the future, Link Indexes™
- Hierarchies honored through joins and/or, in the future, Link Indexes™
  - Inferred ontologies
  - Reasons for hierarchies change depending on application, e.g., view one vendor that has multiple products and/or one product from multiple vendors
- MDM: Master data versioning with access to historic master data

SmartData Fabric® and MDM Processes (3 of 3)

- MDM: Option to replace either data source indexes or source data itself (automatically updates indexes) with master data, e.g., propagate updated phone number, mailing address or email address
- Execute analytics and/or visually represent master data combined with other data and search/query filters, e.g., BI and link analysis/graph database
  - Include aggregations, calculations and other data, e.g., external
- MDM: Write back selective master data updates/corrections to data sources (see above)
- Continuous access to metadata, including latest data profiles
  - Helps identify anomalies/outliers for event processing, alerts and/or data transform modifications

Appendix: Two main forms of Master Data Record (MDR) – repository and registry

How to retain a master data record (MDR)?



Master Data Record Template

CUSTOMER
RECNO
LAST_NAME
FIRST_NAME
KNOWN_AS
ACCOUNT_NO
PHONE_BEST
EMAIL
ADDRESS
CITY
ZIP
STATE
EFF_DATETIME

MDR.CUSTOMER
RECNO01
Smith
Robert
Bob
323
300-400-5000
bsmith@gmail.com
300 Any St.
New Town
76543
AL
1/1/2015 09:32

Master Data Record - based on latest and missing data

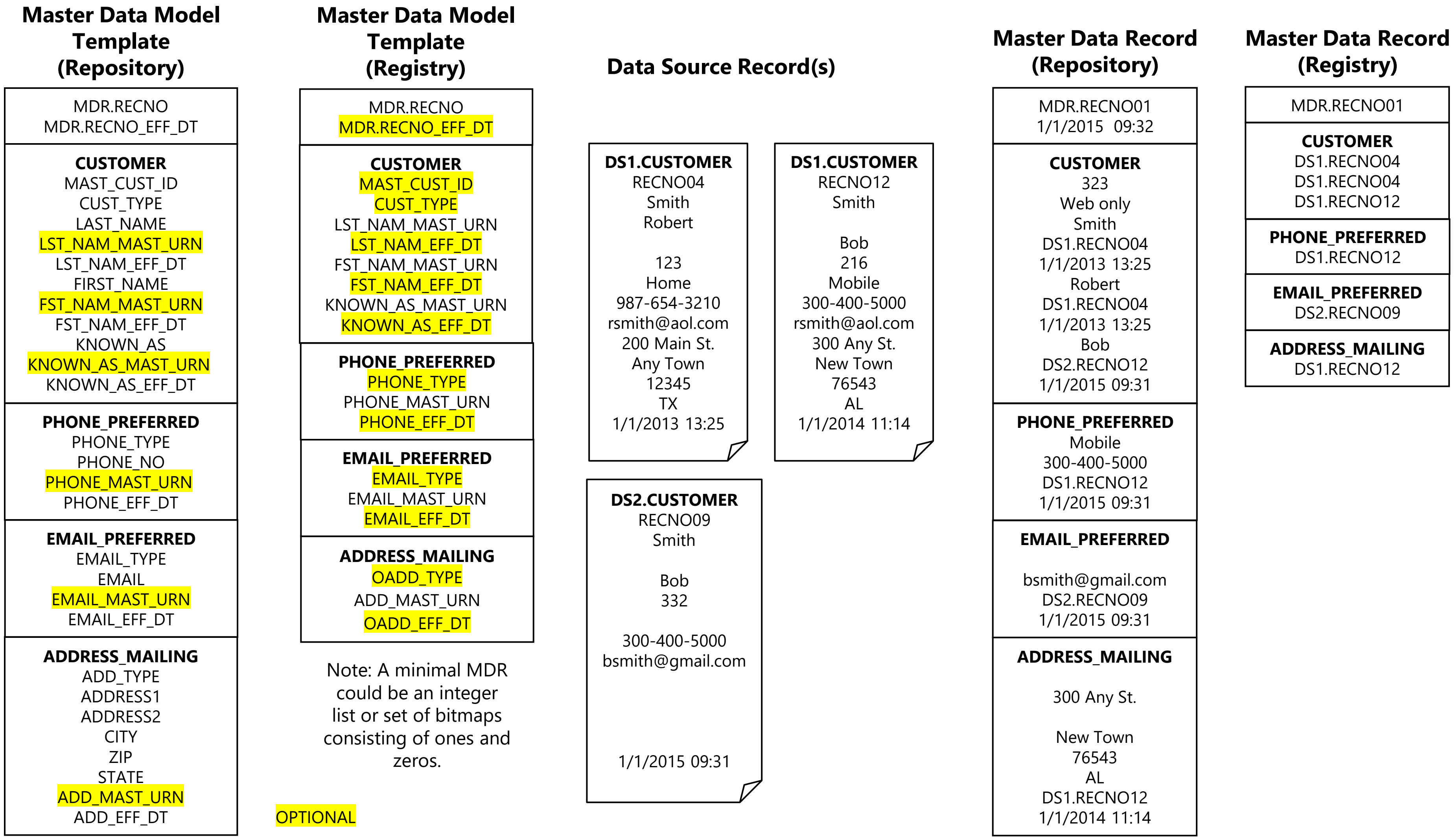
CONTRIBUTE TO

DS1.CUSTOMER
RECNO04
Smith
Robert
123
987-654-3210
rsmith@aol.com
200 Main St.
Any Town
12345
TX
1/1/2013 13:25

DS1.CUSTOMER
RECNO12
Smith
Bob
216
300-400-5000
rsmith@aol.com
300 Any St.
New Town
76543
AL
1/1/2014 11:14

DS2.CUSTOMER
RECNO09
Smith
Bob
332
300-400-5000
bsmith@gmail.com
1/1/2015 09:31

Two main forms of MDR – repository and registry



Master Data Model Template (Repository)

MDR.RECNO
MDR.RECNO_EFF_DT

CUSTOMER
MAST_CUST_ID
CUST_TYPE
LAST_NAME
LST_NAM_MAST_URN
LST_NAM_EFF_DT
FIRST_NAME
FST_NAM_MAST_URN
FST_NAM_EFF_DT
KNOWN_AS
KNOWN_AS_MAST_URN
KNOWN_AS_EFF_DT

PHONE_PREFERRED
PHONE_TYPE
PHONE_NO
PHONE_MAST_URN
PHONE_EFF_DT

EMAIL_PREFERRED
EMAIL_TYPE
EMAIL
EMAIL_MAST_URN
EMAIL_EFF_DT

ADDRESS_MAILING
ADD_TYPE
ADDRESS1
ADDRESS2
CITY
ZIP
STATE
ADD_MAST_URN
ADD_EFF_DT

Master Data Model Template (Registry)

MDR.RECNO
MDR.RECNO_EFF_DT

CUSTOMER
MAST_CUST_ID
CUST_TYPE
LST_NAM_MAST_URN
LST_NAM_EFF_DT
FST_NAM_MAST_URN
FST_NAM_EFF_DT
KNOWN_AS_MAST_URN
KNOWN_AS_EFF_DT

PHONE_PREFERRED
PHONE_TYPE
PHONE_MAST_URN
PHONE_EFF_DT

EMAIL_PREFERRED
EMAIL_TYPE
EMAIL_MAST_URN
EMAIL_EFF_DT

ADDRESS_MAILING
OADD_TYPE
ADD_MAST_URN
OADD_EFF_DT

Note: A minimal MDR could be an integer list or set of bitmaps consisting of ones and zeros.

OPTIONAL

Data Source Record(s)

DS1.CUSTOMER
RECNO04
Smith
Robert

123
Home
987-654-3210
rsmith@aol.com
200 Main St.
Any Town
12345
TX
1/1/2013 13:25

DS1.CUSTOMER
RECNO12
Smith

Bob
216
Mobile
300-400-5000
rsmith@aol.com
300 Any St.
New Town
76543
AL
1/1/2014 11:14

DS2.CUSTOMER
RECNO09
Smith

Bob
332

300-400-5000
bsmith@gmail.com

1/1/2015 09:31

Master Data Record (Repository)

MDR.RECNO01
1/1/2015 09:32

CUSTOMER
323
Web only
Smith
DS1.RECNO04
1/1/2013 13:25
Robert
DS1.RECNO04
1/1/2013 13:25
Bob
DS2.RECNO12
1/1/2015 09:31

PHONE_PREFERRED
Mobile
300-400-5000
DS1.RECNO12
1/1/2015 09:31

EMAIL_PREFERRED

bsmith@gmail.com
DS2.RECNO09
1/1/2015 09:31

ADDRESS_MAILING

300 Any St.

New Town
76543
AL
DS1.RECNO12
1/1/2014 11:14

Master Data Record (Registry)

MDR.RECNO01

CUSTOMER
DS1.RECNO04
DS1.RECNO04
DS1.RECNO12

PHONE_PREFERRED
DS1.RECNO12

EMAIL_PREFERRED
DS2.RECNO09

ADDRESS_MAILING
DS1.RECNO12

Advantages and disadvantages of conventional REPOSITORY MDM

ADVANTAGES

- Single best version of master data
- Queryable

DISADVANTAGES

- Applications need to be modified to use master data – difficult, cost and time – questionable success
- Complex data model – maintenance and modification
- Single (typically) relational data models do not serve many applications – multiple different materialized views required
- Difficult to combine and integrate master data with non-master data, in particular, operational/transactional data

Advantages and disadvantages of conventional REGISTRY MDM

ADVANTAGES

- Use of keys and link tables to point to master data in data sources – avoids ETL and synchronization
- Normal applications indirectly maintain master data – not a separate external application
- Difficult, but possible that existing applications automatically use and substitute master data
- New applications can be written to take advantage of registry MDM easier than with repository MDM

DISADVANTAGES

- Requires access to multiple data sources to retrieve master data
- Query loads on adapters and data sources
- Bandwidth and query performance
- Dirty and potentially unusable non-standard master data requires additional processing
- Pointers tend to be limited to single records instead of multiple attributes in multiple records, leading to compromised data
- Difficult to maintain historic master data

Appendix: Hybrid master data record creation

Hybrid master data record creation (1 of 2)

- Normal content indexes are generated using data quality transforms, including address correction
  - In the future, may include initial binning key index, e.g., for PERSON, LAST_NAME_FUZZY + DOB
  - Fuzzy match indexes are generated using various algorithms, including latest Metaphone 3 for international names
- Exact matches on very high cardinality entities, e.g., ADDRESS, SSN, PHONE and EMAIL – future test for transpositions
  - In the future, exact matches on binning key, e.g., for PERSON, LAST_NAME_FUZZY + DOB
- For “complex entities”, use binning and edit-distance algorithms
  - In the future, may avoid some or all edit-distance algorithms
- See if one or more very high cardinality entities match records across data sources, and if so, pull records into bin and re-bin based on high cardinality entity matches
  - If no very high cardinality records, run edit-distance algorithms without – just more resource intensive

Hybrid master data record creation (2 of 2)

- Run edit-distance algorithms within each bin and composite score for matching probabilities, e.g., PERSON based on names and other entity attributes DOB, SSN, PHONE, EMAIL and ADDRESS
- If the composite score is above a certain threshold:
  - Unique ID is assigned to the primary master entity, e.g., PERSON
  - Unique ID is virtually indexed for each occurrence of the primary master entity in all indexes for all data sources
  - A master data record is created in a separate master table in each adapter containing the master entity, e.g., PERSON, and the unique ID is stored and indexed in the master data table
- Run a process, aka MDM merge process, to determine the best master data values for each master entity, e.g., PERSON ID
- Write best master data values back to every master data table associated with each relevant EIQ Adapter, along with any associated data, e.g., any phonetic token, links to records containing values and date-time
- Automatically index all master data