MarkLogic

Data Governance for Regulated Industries Using Hadoop…and NoSQL

Justin Makeig, Director – Product Management, MarkLogic

October 2013

MarkLogic

Who am I?

- Product Manager for 6 years at MarkLogic
- Background in FinServ and web development
- Passionate about data, infrastructure, and user experience

What is MarkLogic?

- Enterprise NoSQL since 2001
- Distributed database + search + app platform
- 250+ paying customers, 500+ production applications

MarkLogic

Agenda

- Data governance considerations
- Legacy approaches: Why it's hard
- New generation: Hadoop + Enterprise NoSQL
- Enterprise NoSQL
- Case studies: FATCA, eDiscovery, Dodd-Frank
- Q&A

MarkLogic

Data Governance Considerations


Security


Retention


Privacy


Continuity


Provenance


Compliance

MarkLogic

Why is this difficult?

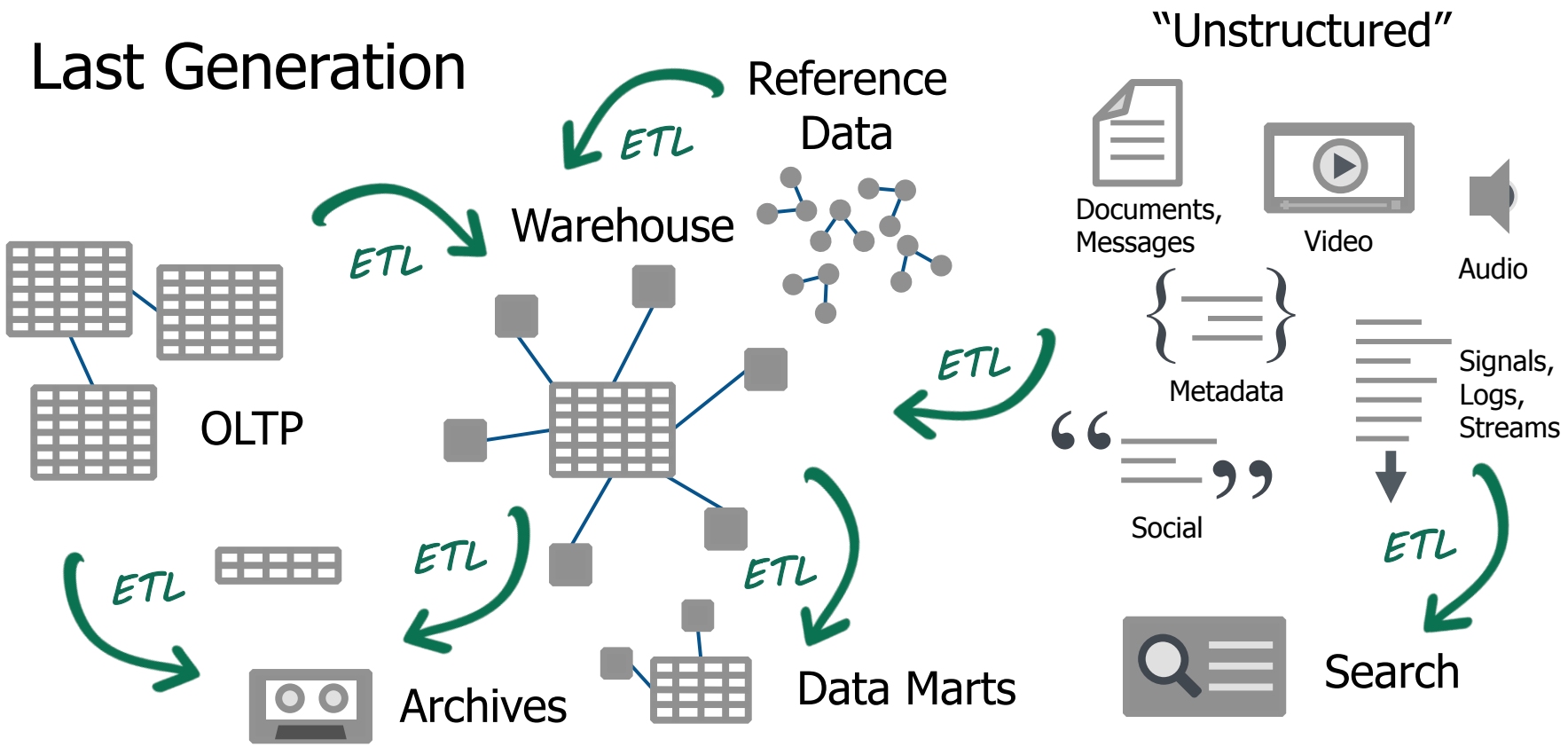
And risky?

And expensive?

And behind schedule?

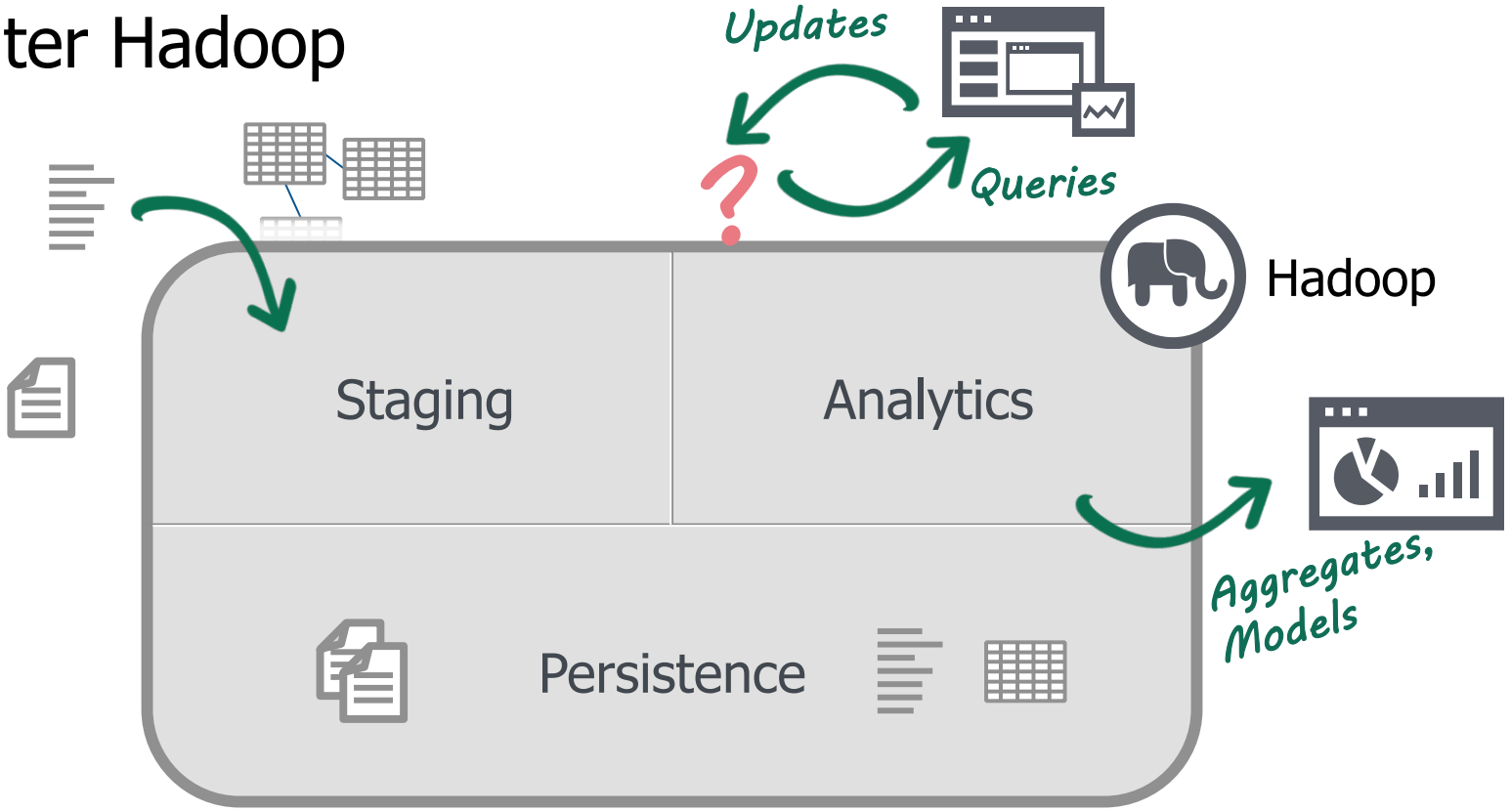
MarkLogic

Last Generation



MarkLogic

Enter Hadoop



MarkLogic

Why must we choose?

Legacy RDBMS

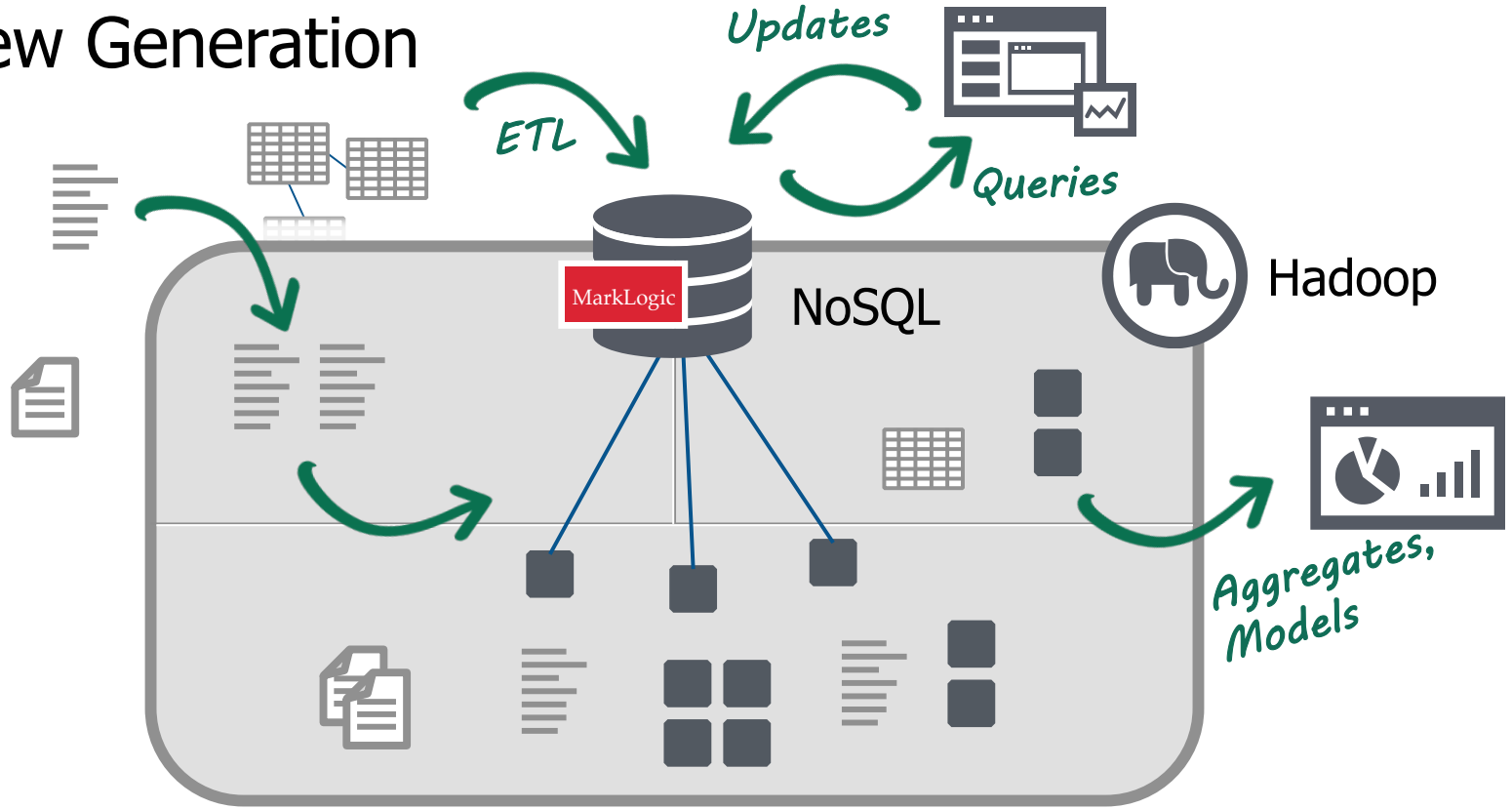
- Indexes
- Transactions
- Security
- Enterprise operations

"NoSQL"

- Flexible data model
- Commodity scale out
- Distributed, fault-tolerant
- Hadoop sink/source

MarkLogic

New Generation



MarkLogic

Enterprise NoSQL

- Flexible data model, comprehensive indexes
  - Documents: Hierarchy, text, values, tags—schema "on-demand"
  - Scalars: Aggregates and range filters, including geospatial
  - Triples: Linked facts and inferencing
  - Permissions: Users, roles, compartments, and privileges
  - Queries: Reverse indexes for alerting, matching
- Ad hoc dimensions, lock-free reads
- Real-time transformation
- Strict consistency throughout

Preserving Context with Documents

Before

…movement of materials was observed en route to Abattabad some time after 14:30…

Inline Enrichment

After

…movement of materials was observed en route to

```
<place lat="…" long="…" version="2.2.1">
  <original>Abattabad</original>
  <canonical ref="…">Abbottabad</canonical>
  <source>/sources/1234</source>
  <confidence>0.87</confidence>
  …
</place>
```

some time after 14:30…

Transactional updates

MarkLogic

Complementary Approaches

NoSQL

- Online applications
- Delivery
- Decision-making
- Real-time
- Granular updates
- Distributed indexes

Hadoop

- Offline analytics
- Staging
- Model-building
- Long-haul batch
- Write-once, read-many
- Distributed file system

Case Studies

MarkLogic

KPMG: FATCA Compliance for Customer On-Boarding

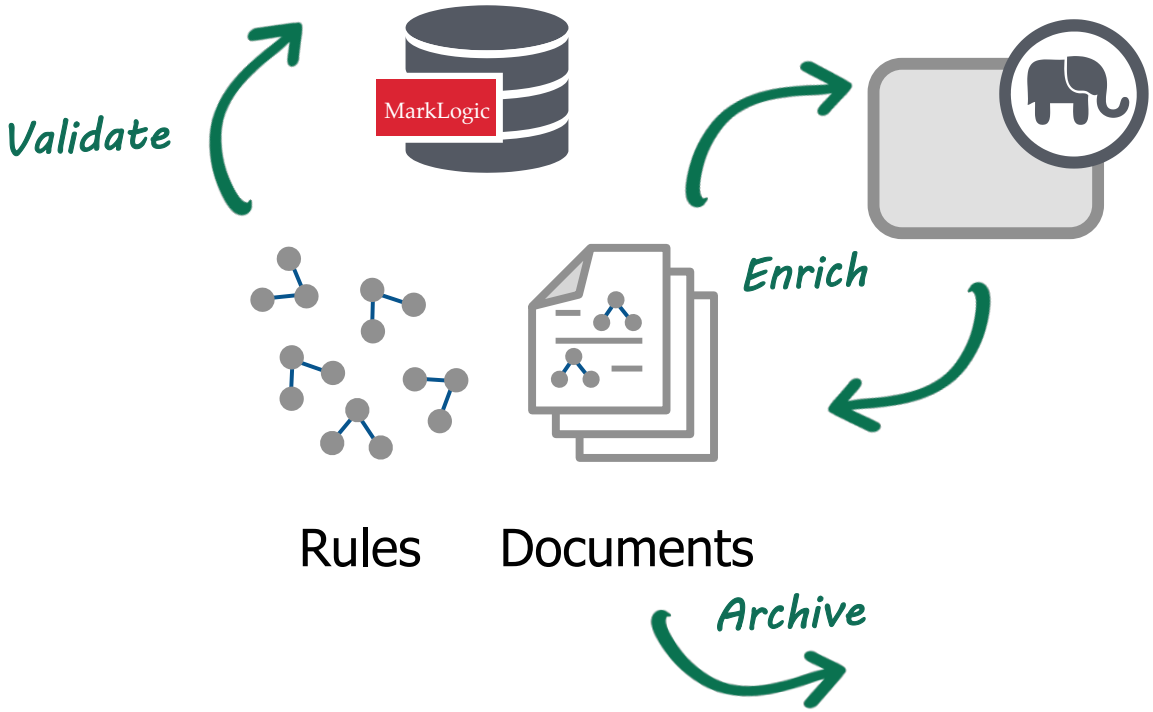
- Thousands of rules, 1–2M accounts, 30–40M documents
- Encoding, adjusting, and matching rules must scale
- Impossible to pre-define dimensions, relationships
- Vet new accounts and "show your work"
- Real-time decision-making

6-48 hours to
3 seconds

KPMG

MarkLogic

KPMG: FATCA Compliance for Customer On-Boarding



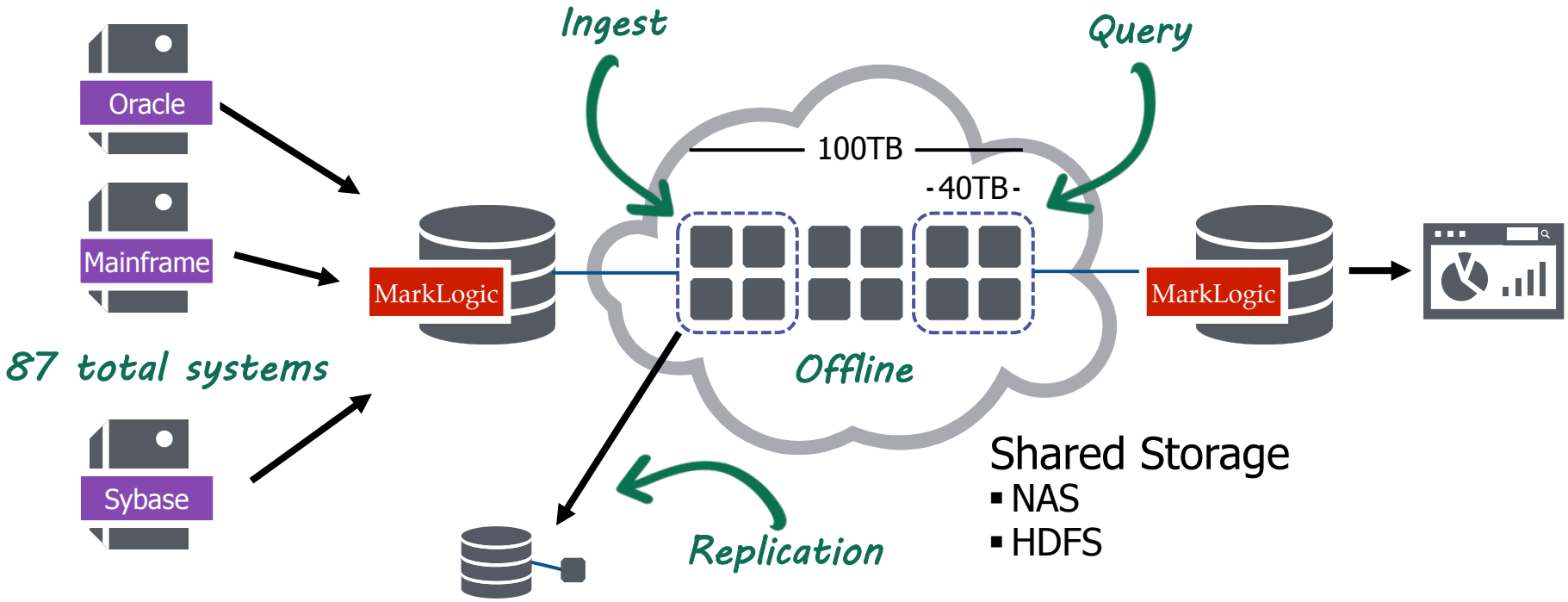
MarkLogic

Tier 1 European Bank: Compliance and Legal Holds

- Accurately respond to discovery as part of litigation
- Hold, review, produce data across current, legacy systems
- Repatriate and reconcile distributed data
- Demonstrate fidelity and audit trail
- Reduce infrastructure and maintenance costs

MarkLogic

Tier 1 European Bank: Compliance and Legal Holds



MarkLogic

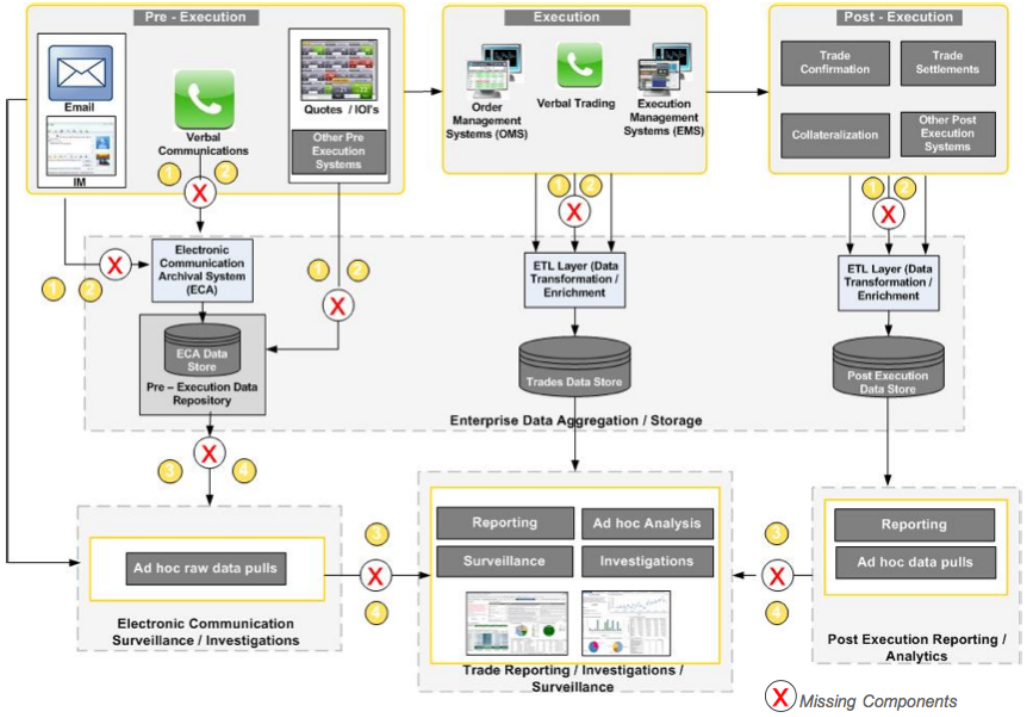
Ernst & Young: Dodd-Frank Compliance

- Trace lineage of order lifecycle for OTC derivatives
- Search, link supporting communications, documents
- Strict reporting and retention rules, response times
- Existing policies, point solutions don't scale

EY
Building a better working world

MarkLogic

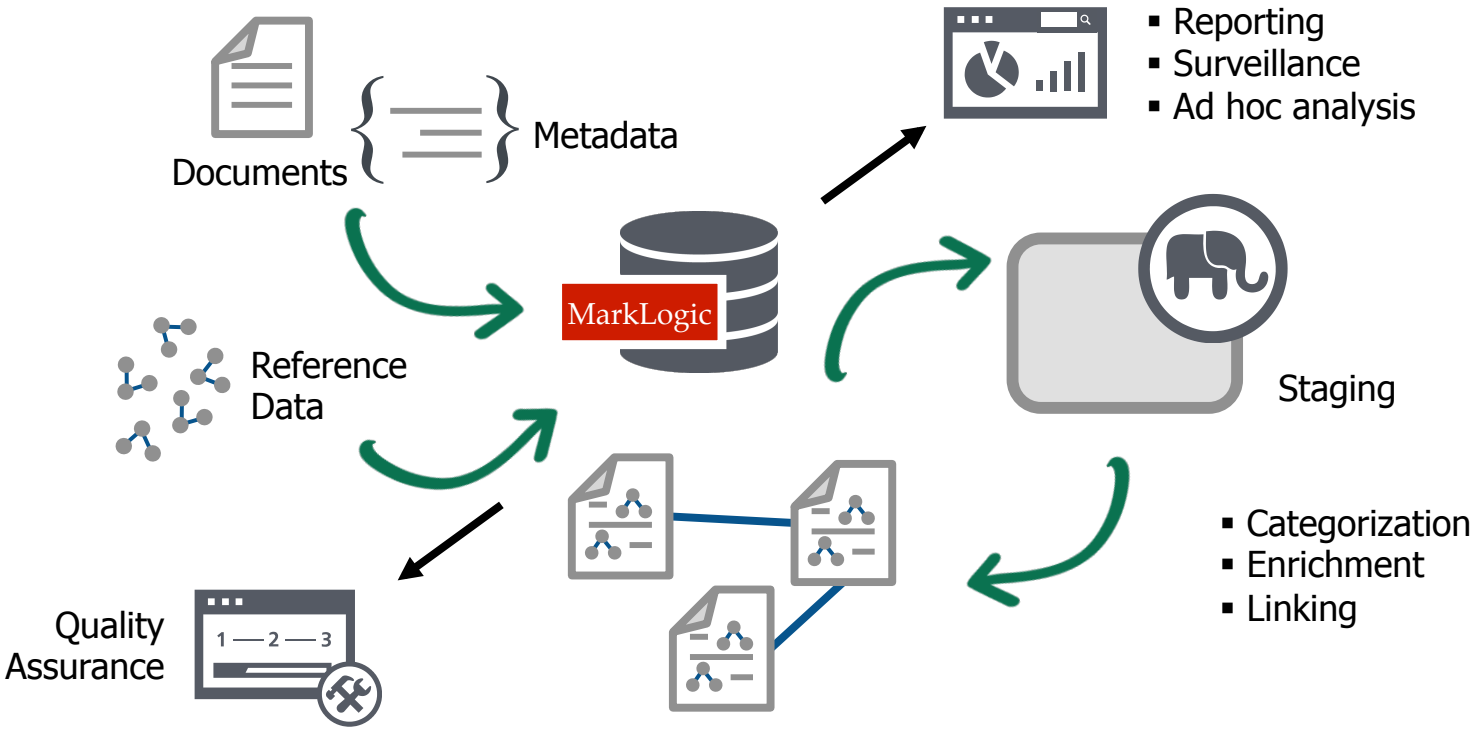
Current State



- Missing key relationships between pre-/post-trade data
- No way to query across silos
- Segregated reporting and surveillance

MarkLogic

Ernst & Young: Dodd-Frank Compliance



MarkLogic

Enrichment and Linking

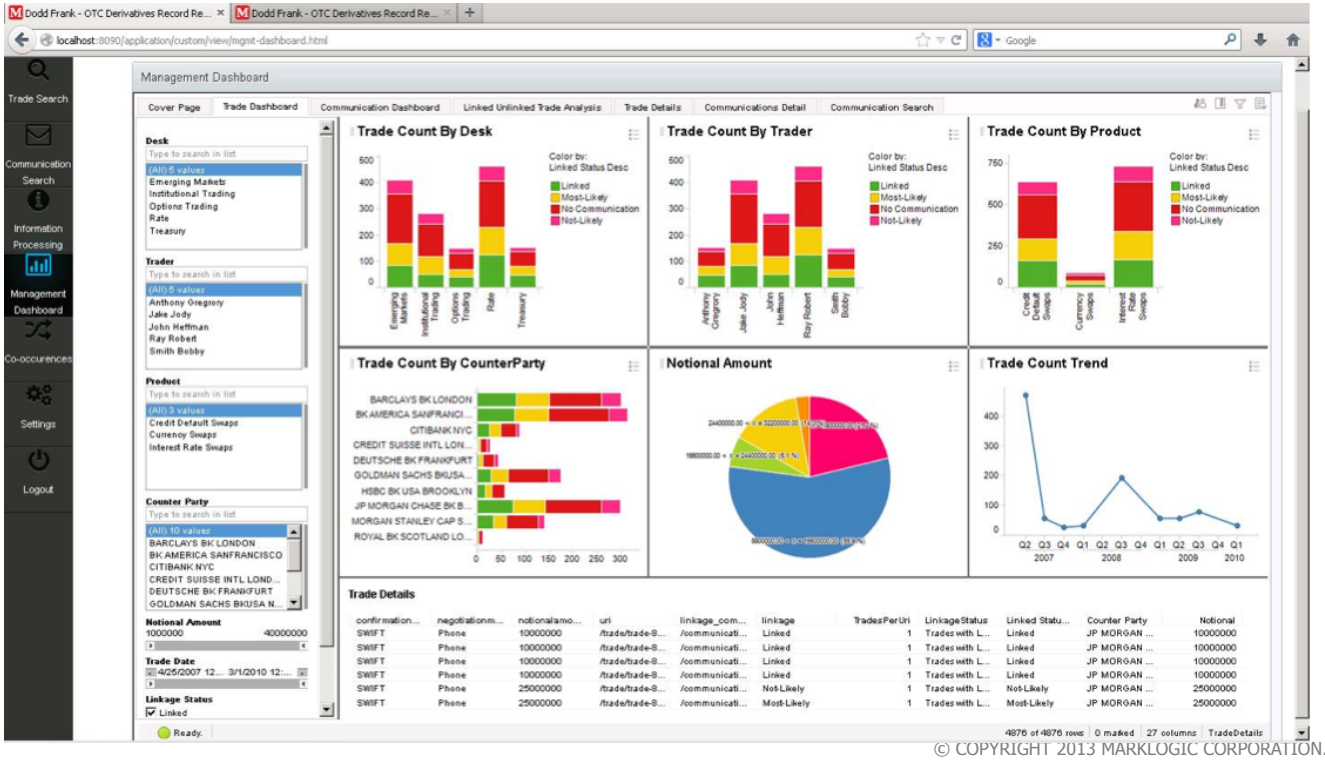
```
<Email>
   <Subject>CDS 050907 Trade Interest</Subject>
   <From>Sam@ACME.com</From>
   <To>Phil.D@DEUTSCHE.com</To>
   <Party>ACME</Party>
   <Party ID>100 </Party ID>
   <CounterParty>DEUTSCHE</CounterParty>
   <CounterParty ID>200 </CounterParty ID>
   <Trader Name>SAM</Trader Name>
   <Trader ID>1001</Trader ID>
   <Trade Date>2012-06-20</Trade Date>
   <Quantity> </Quantity>
   <Security ID>2001</Security ID>
   <Security Name>CDS 050907</Security Name>
   <Execution Price>.015 </Execution Price>
   <Principal Amount>5000000</Principal Amount>
</Email>
```

Trade Record

Trade Data	
Party	ACME
Party ID	100
Counter Party	DEUTSCHE
Counter Party ID	200
Trade Date	2012-06-20
Buy/Sell	Sell
Quantity	
Principal Amount	5000000
Price/Interest	.015
Currency	USD
Security ID	2001
Security Name	CDS 050907
Security Type	CDS

MarkLogic

Management Dashboard



What now?

MarkLogic

Take-Aways

- New and more data is both an opportunity and a threat
- Last generation of data management is not sufficient
- More copies, representations, transformations increase risk
- Index once and reuse across workloads, lifecycle
  - NoSQL: indexing and updates for interactive apps
  - Hadoop: staging, persistence, and analytics

MarkLogic

DO MORE WITH HADOOP

SECURE

Minimize duplication, costly ETL, reduce risk

REAL-TIME

Enterprise-class database for real-time search, delivery & analytics

RUN APPLICATIONS

Run mission critical applications directly on HDFS