Data Governance for Regulated Industries Using Hadoop...*and NoSQL*

Justin Makeig, Director – Product Management, MarkLogic

November 2013

Who am I?

- Product Manager for 6 years at MarkLogic
- Background in consulting 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

Data Governance Considerations

Security

Data Governance Considerations

- 🔒 Security
- Privacy

Data Governance Considerations

- Security
- Privacy
- Provenance

Data Governance Considerations

- Security
- Privacy
- Provenance
- Retention

MarkLogic

Data Governance Considerations

- Security
- Privacy
- Provenance
- Retention
- Continuity

MarkLogic

Data Governance Considerations

- Security
- Privacy
- Provenance
- Retention
- Continuity
- Compliance

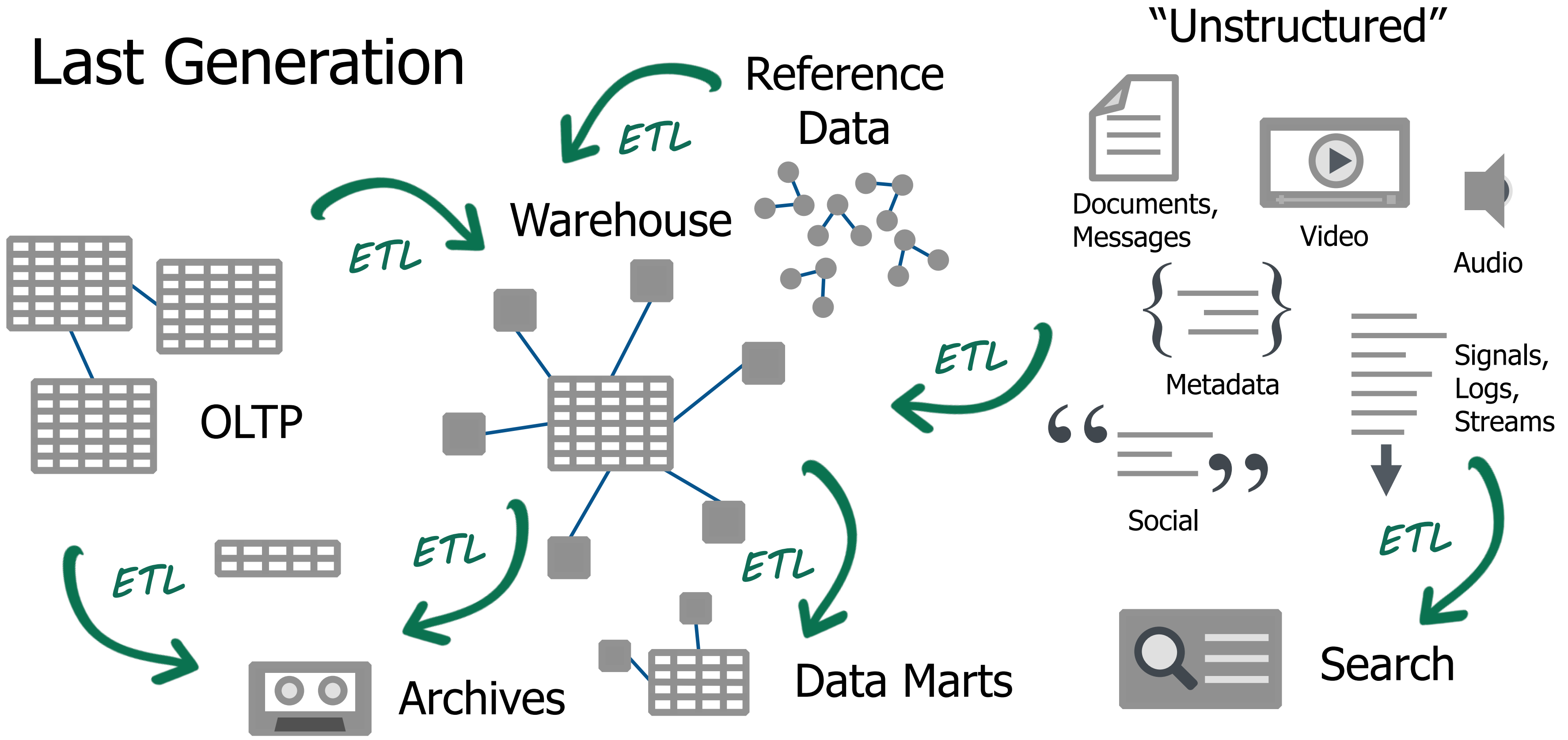
Why is this difficult?

And risky?

And expensive?

And behind schedule?

Last Generation



"Unstructured"

Reference Data

ETL

Warehouse

ETL

OLTP

Documents, Messages

Video

Audio

Metadata

ETL

Signals, Logs, Streams

Social

ETL

ETL

ETL

ETL

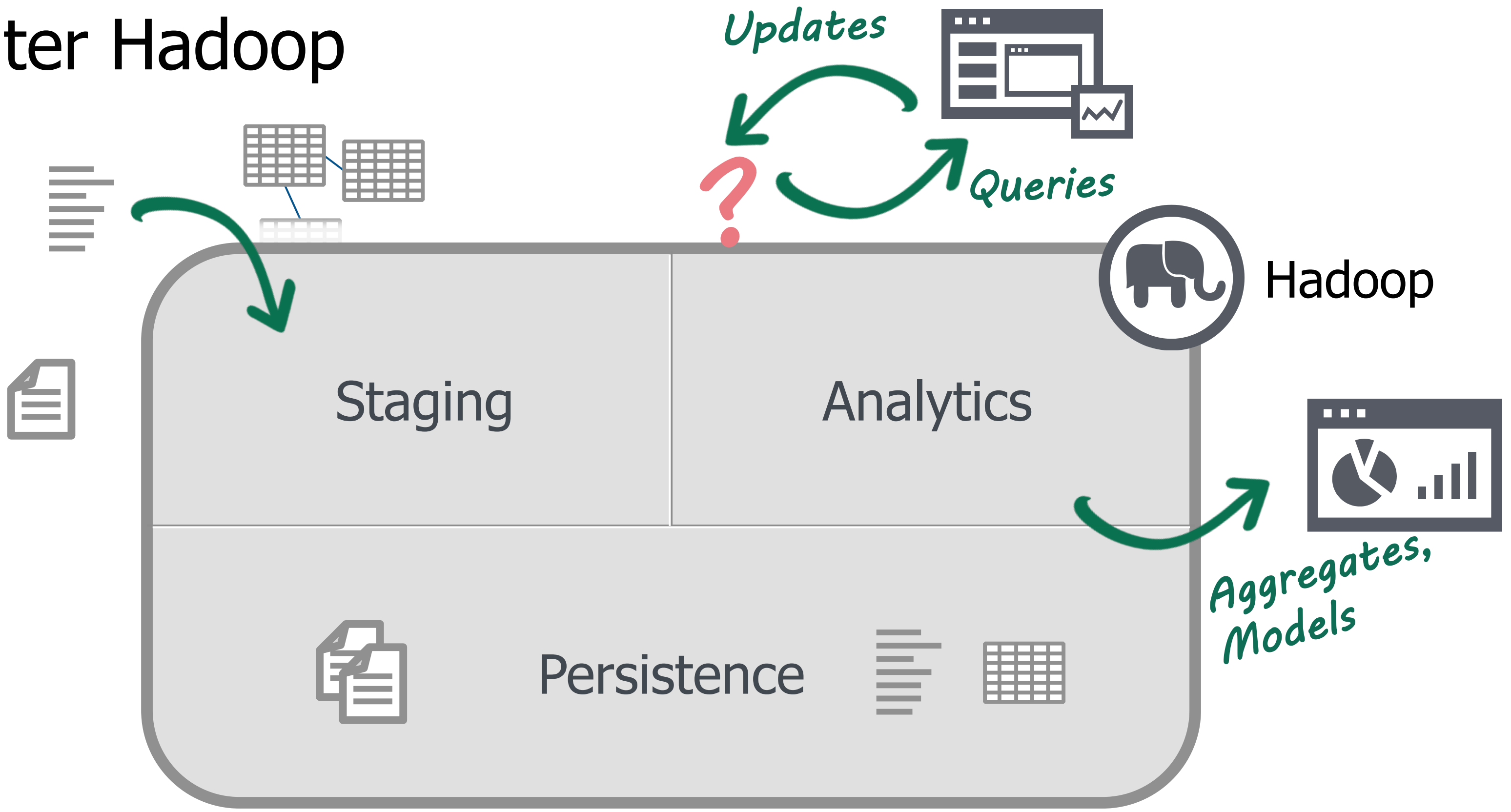
Archives

Data Marts

Search

MarkLogic

Enter Hadoop



Why must we choose?

Legacy RDBMS

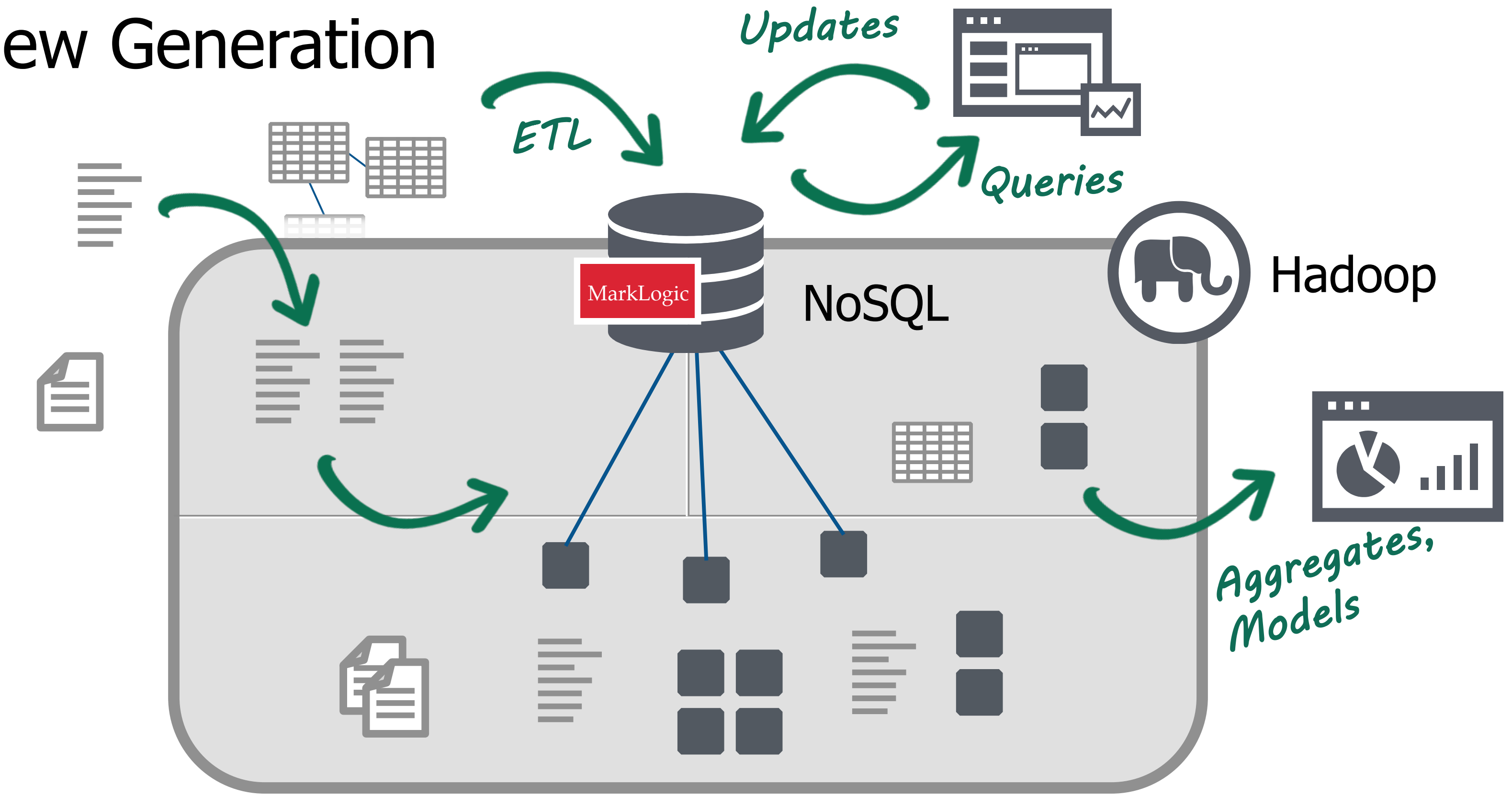
- Indexes
- Transactions
- Security
- Enterprise operations

“NoSQL”

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

MarkLogic

New Generation

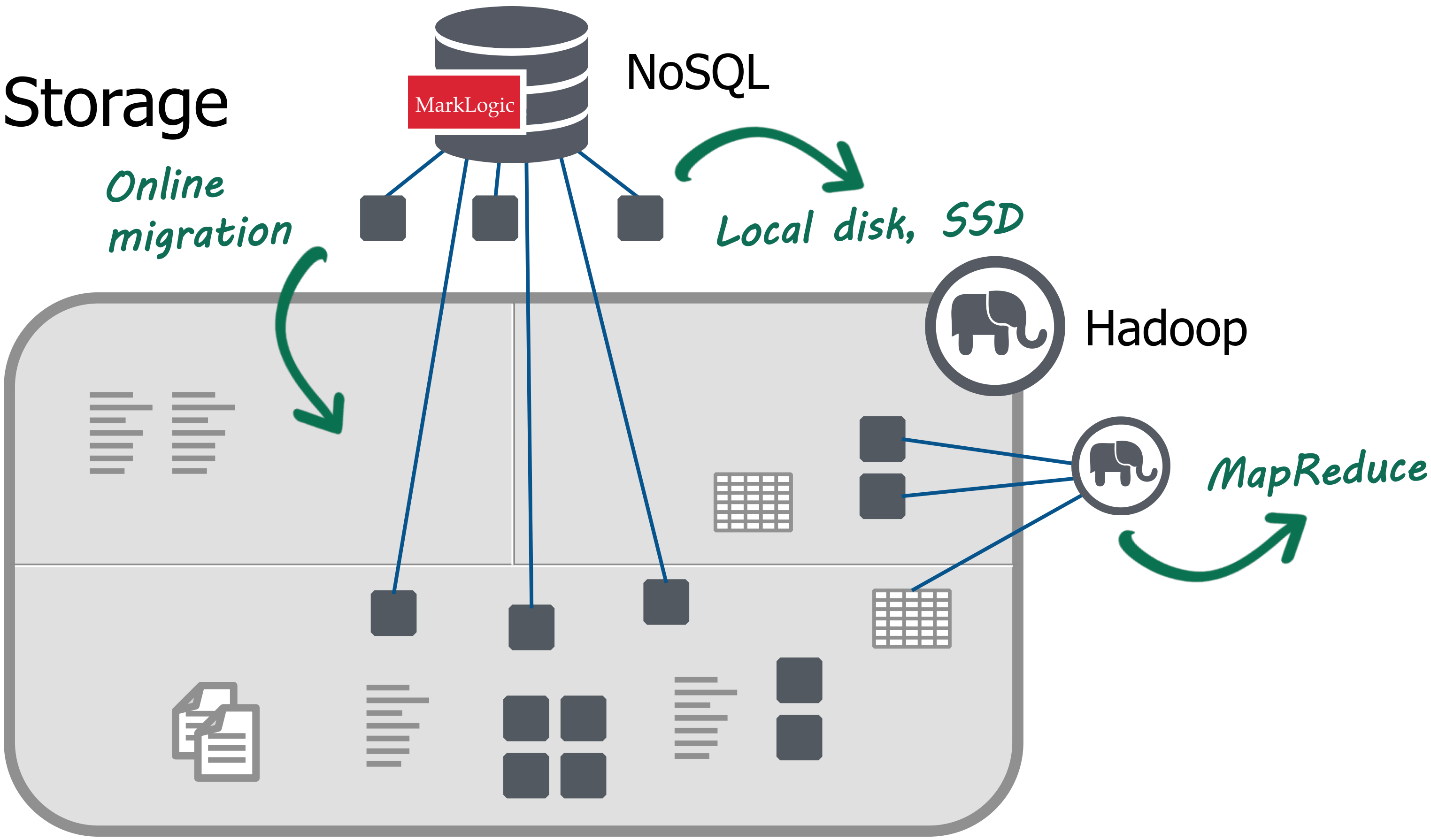


MarkLogic

Tiered Storage

Active
~$25/GB

Historical
~$1/GB



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, security 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

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

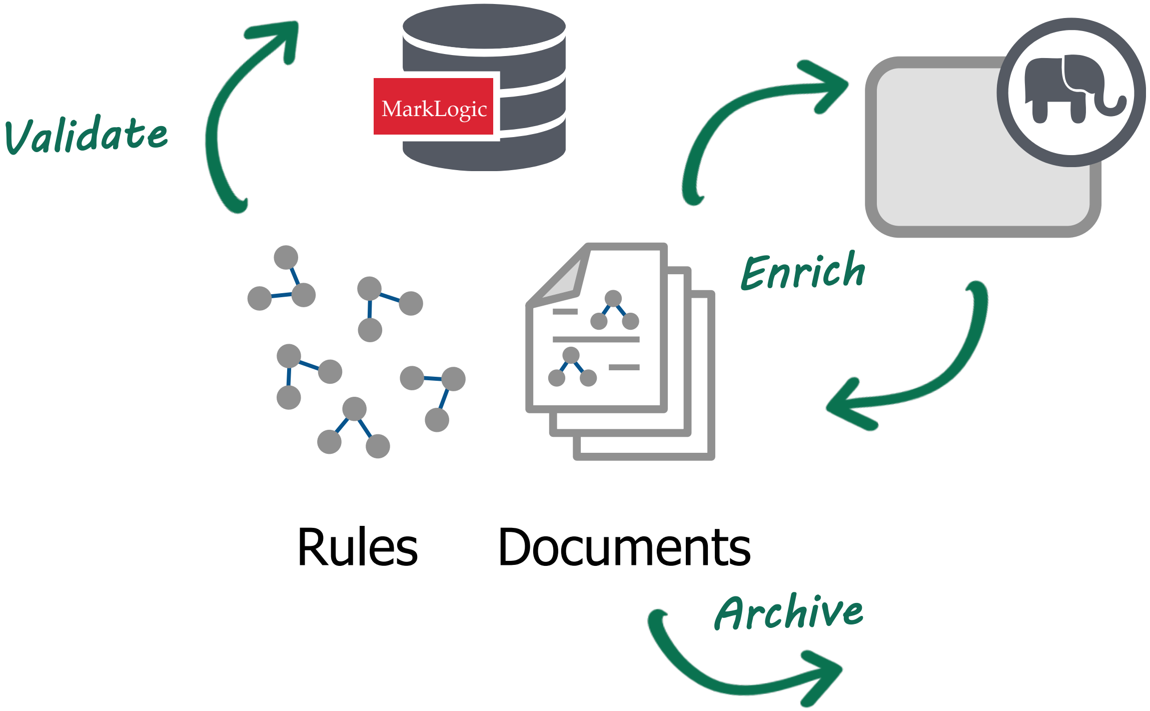
- Thousands of rules, 1–2M accounts, 30–40M documents
- Encoding, adjusting, and matching rules must scale
- Impossible to pre-define dimensions, relationships
- Vet new accounts and “show your work”
- Real-time decision-making

6-48 hours to
3 seconds

KPMG

MarkLogic

KPMG: FATCA Compliance for Customer On-Boarding



MarkLogic

MarkLogic Semantics

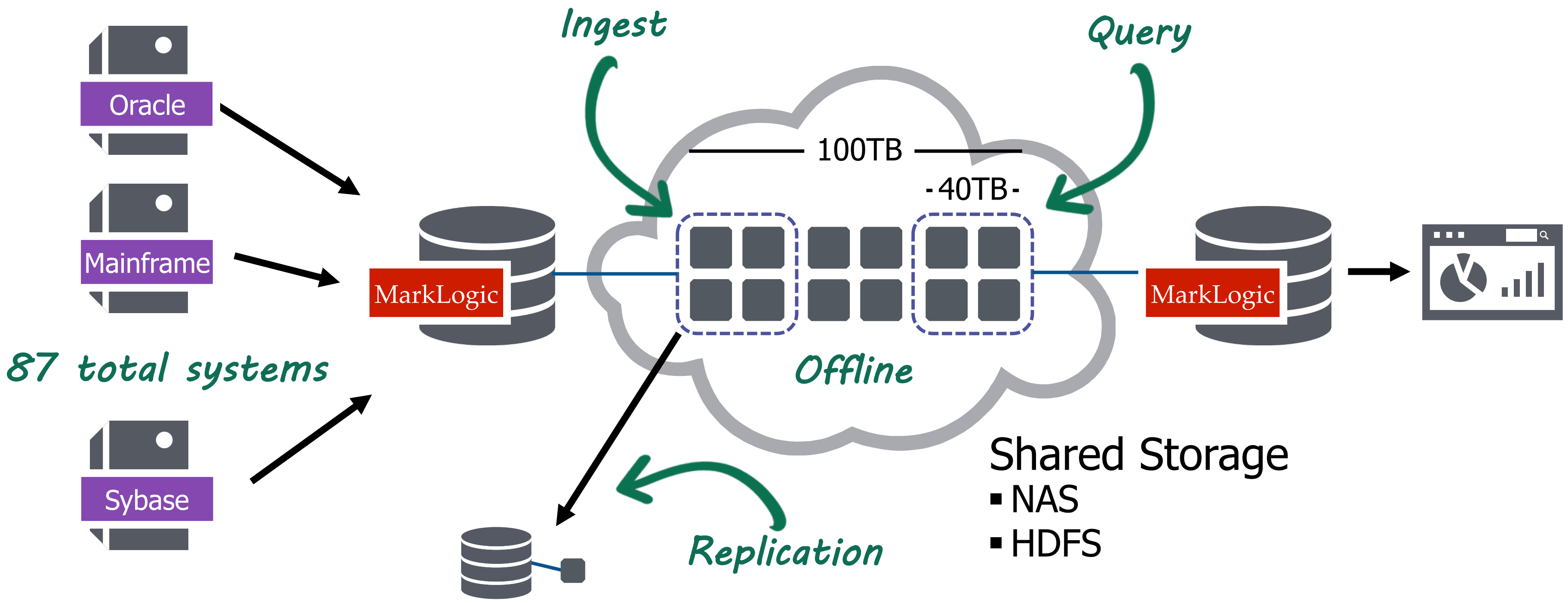
- Combine facts from your documents with linked data
- Improve the search experience with facts about the results
- Enhance recall, improve precision with search *and* semantic queries

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

Estimated $16M savings

Tier 1 European Bank: Compliance and Legal Holds



Shared Storage
- NAS
- HDFS

MarkLogic

Ernst & Young: Dodd-Frank Compliance

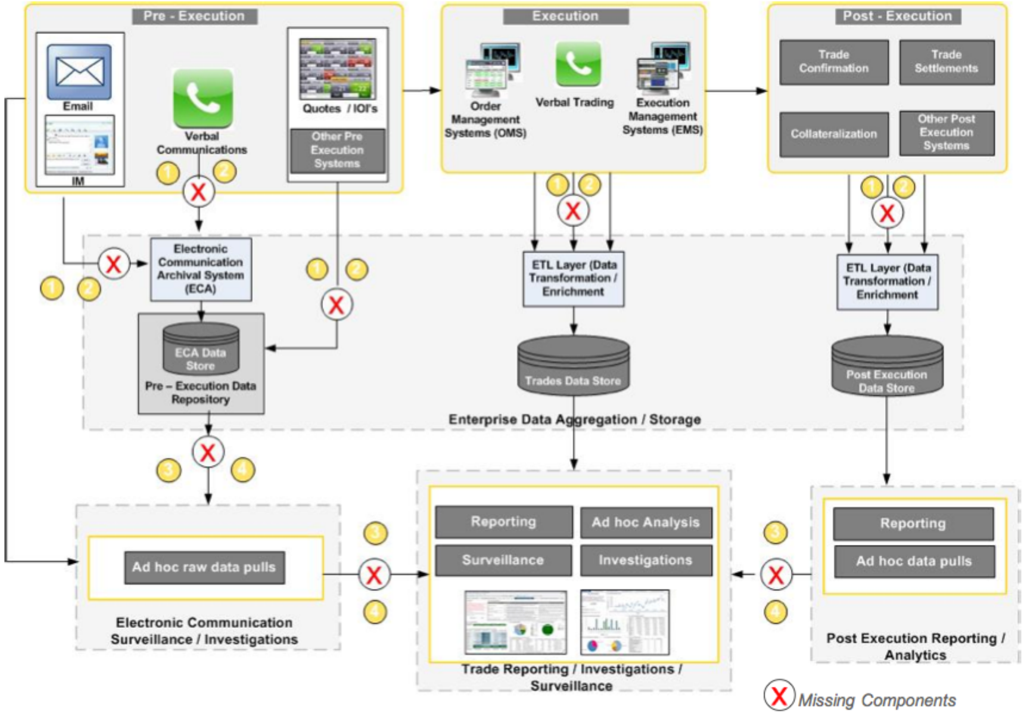
- Trace lineage of order lifecycle for OTC derivatives
- Search, link supporting communications, documents
- Strict reporting and retention rules, response times
- Existing policies, point solutions don’t scale

EY

Building a better working world

MarkLogic

Current State



Pre - Execution

Email

IM

Verbal Communications

Quotes / IOI's

Other Pre Execution Systems

Execution

Order Management Systems (OMS)

Verbal Trading

Execution Management Systems (EMS)

Post - Execution

Trade Confirmation

Trade Settlements

Collateralization

Other Post Execution Systems

Electronic Communication Archival System (ECA)

ECA Data Store

Pre – Execution Data Repository

ETL Layer (Data Transformation / Enrichment

Trades Data Store

ETL Layer (Data Transformation / Enrichment

Post Execution Data Store

Enterprise Data Aggregation / Storage

Ad hoc raw data pulls

Electronic Communication Surveillance / Investigations

Reporting

Ad hoc Analysis

Surveillance

Investigations

Trade Reporting / Investigations / Surveillance

Reporting

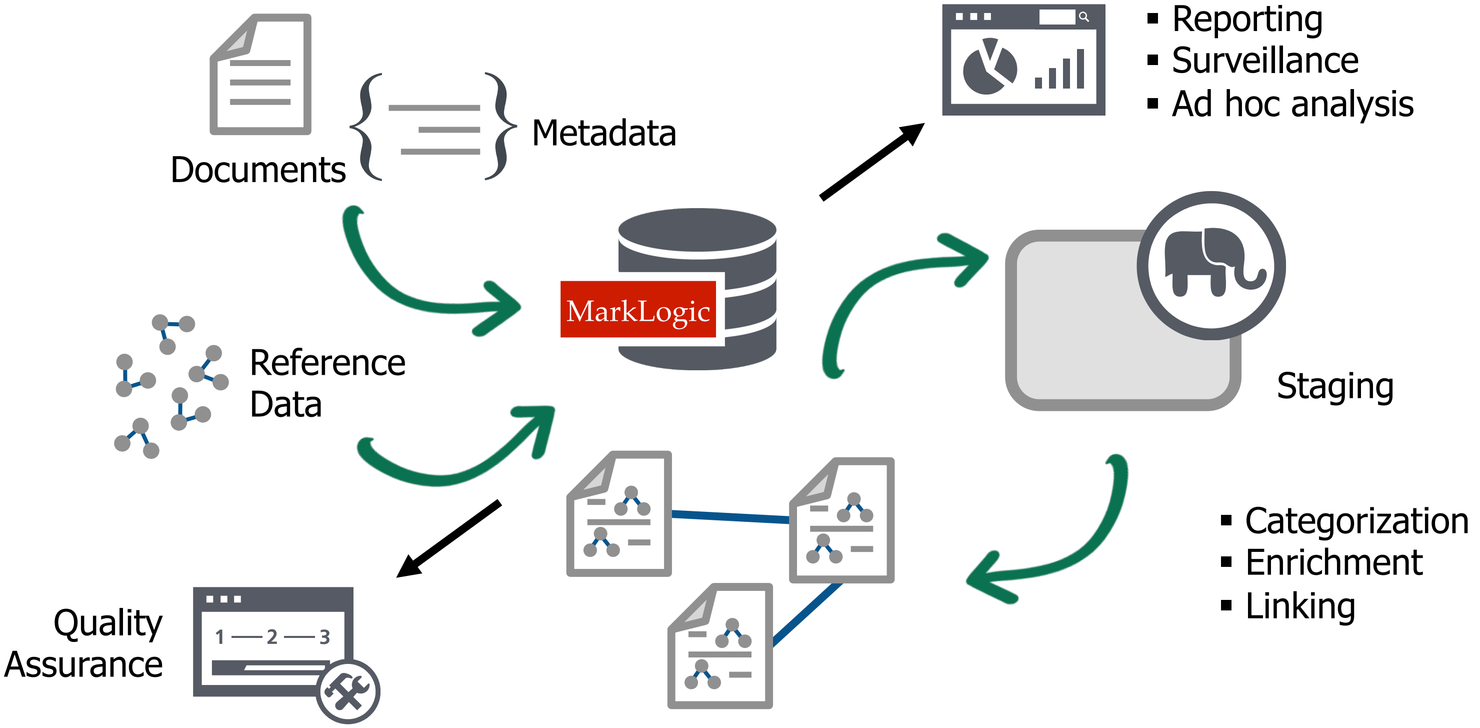
Ad hoc data pulls

Post Execution Reporting / Analytics

X *Missing Components*

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

Ernst & Young: Dodd-Frank Compliance



Documents

Metadata

Reference Data

- Reporting
- Surveillance
- Ad hoc analysis

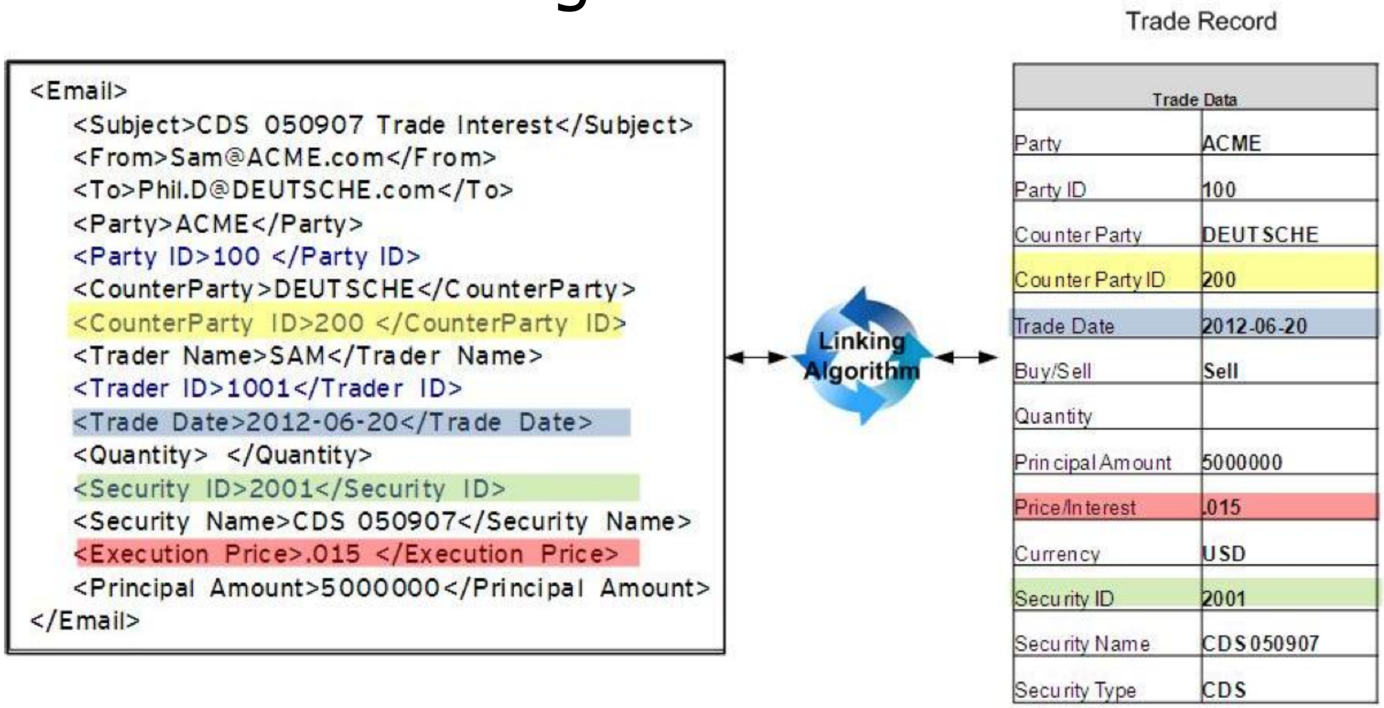
Staging

- Categorization
- Enrichment
- Linking

Quality Assurance

MarkLogic

Enrichment and Linking



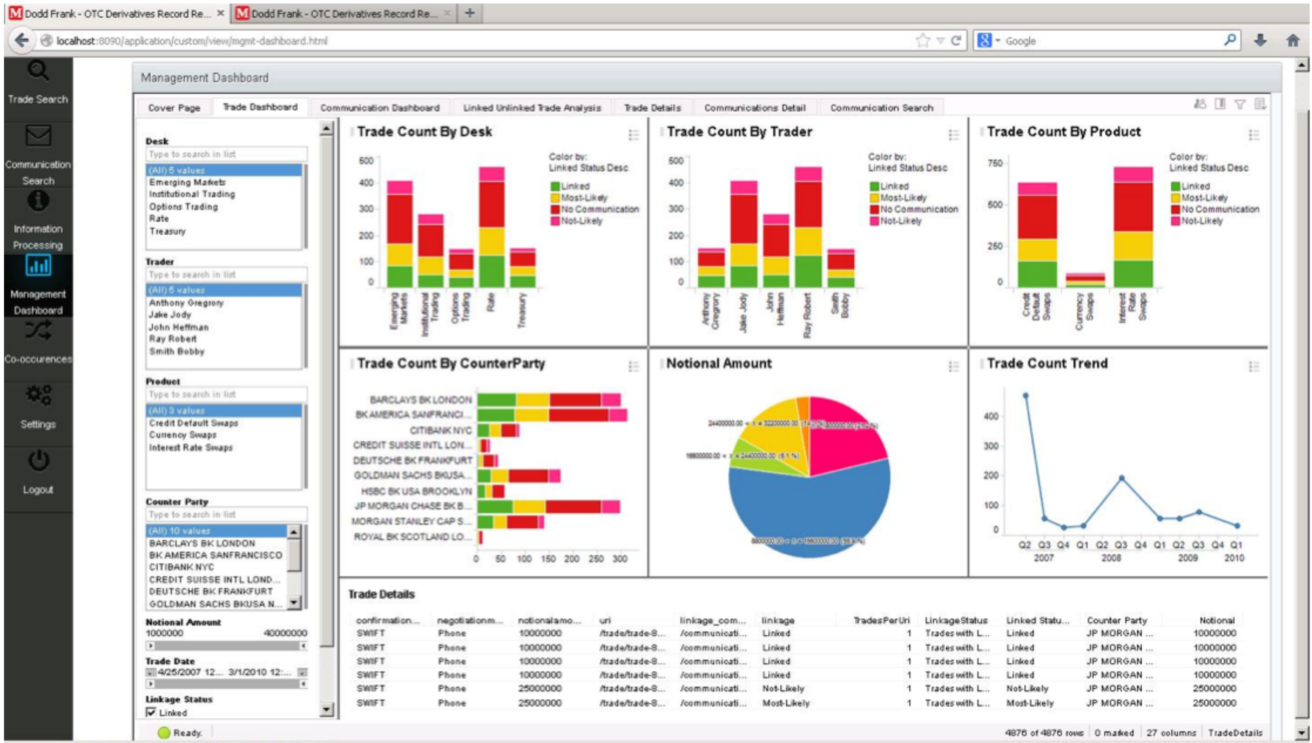
```
<Email>
    <Subject>CDS 050907 Trade Interest</Subject>
    <From>Sam@ACME.com</From>
    <To>Phil.D@DEUTSCHE.com</To>
    <Party>ACME</Party>
    <Party ID>100 </Party ID>
    <CounterParty>DEUTSCHE</CounterParty>
    <CounterParty ID>200 </CounterParty ID>
    <Trader Name>SAM</Trader Name>
    <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



Desk

Type to search in list

- (All) 5 values
- Emerging Markets
- Institutional Trading
- Options Trading
- Rate
- Treasury

Trader

Type to search in list

- (All) 5 values
- Anthony Gregrory
- Jake Jody
- John Heffman
- Ray Robert
- Smith Bobby

Product

Type to search in list

- (All) 3 values
- Credit Default Swaps
- Currency Swaps
- Interest Rate Swaps

Counter Party

Type to search in list

- (All) 10 values
- BARCLAYS BK LONDON
- BK AMERICA SANFRANCISCO
- CITIBANK NYC
- CREDIT SUISSE INTL LOND...
- DEUTSCHE BK FRANKFURT
- GOLDMAN SACHS BKUSA N...

Notional Amount

1000000 40000000

Trade Date

4/25/2007 12... 3/1/2010 12:...

Linkage Status

☑ Linked

Trade Details

confirmation...	negotiationm...	notionalamo...	uri	linkage_com...	linkage	TradesPerUri	LinkageStatus	Linked Statu...	Counter Party	Notional
SWIFT	Phone	10000000	/trade/trade-8...	/communicati...	Linked	1	Trades with L...	Linked	JP MORGAN ...	10000000
SWIFT	Phone	10000000	/trade/trade-8...	/communicati...	Linked	1	Trades with L...	Linked	JP MORGAN ...	10000000
SWIFT	Phone	10000000	/trade/trade-8...	/communicati...	Linked	1	Trades with L...	Linked	JP MORGAN ...	10000000
SWIFT	Phone	10000000	/trade/trade-8...	/communicati...	Linked	1	Trades with L...	Linked	JP MORGAN ...	10000000
SWIFT	Phone	25000000	/trade/trade-8...	/communicati...	Not-Likely	1	Trades with L...	Not-Likely	JP MORGAN ...	25000000
SWIFT	Phone	25000000	/trade/trade-8...	/communicati...	Most-Likely	1	Trades with L...	Most-Likely	JP MORGAN ...	25000000

4876 of 4876 rows | 0 marked | 27 columns | TradeDetails

What now?

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

DO MORE WITH HADOOP

SECURE

Minimize duplication,
costly ETL, reduce risk

REAL-TIME

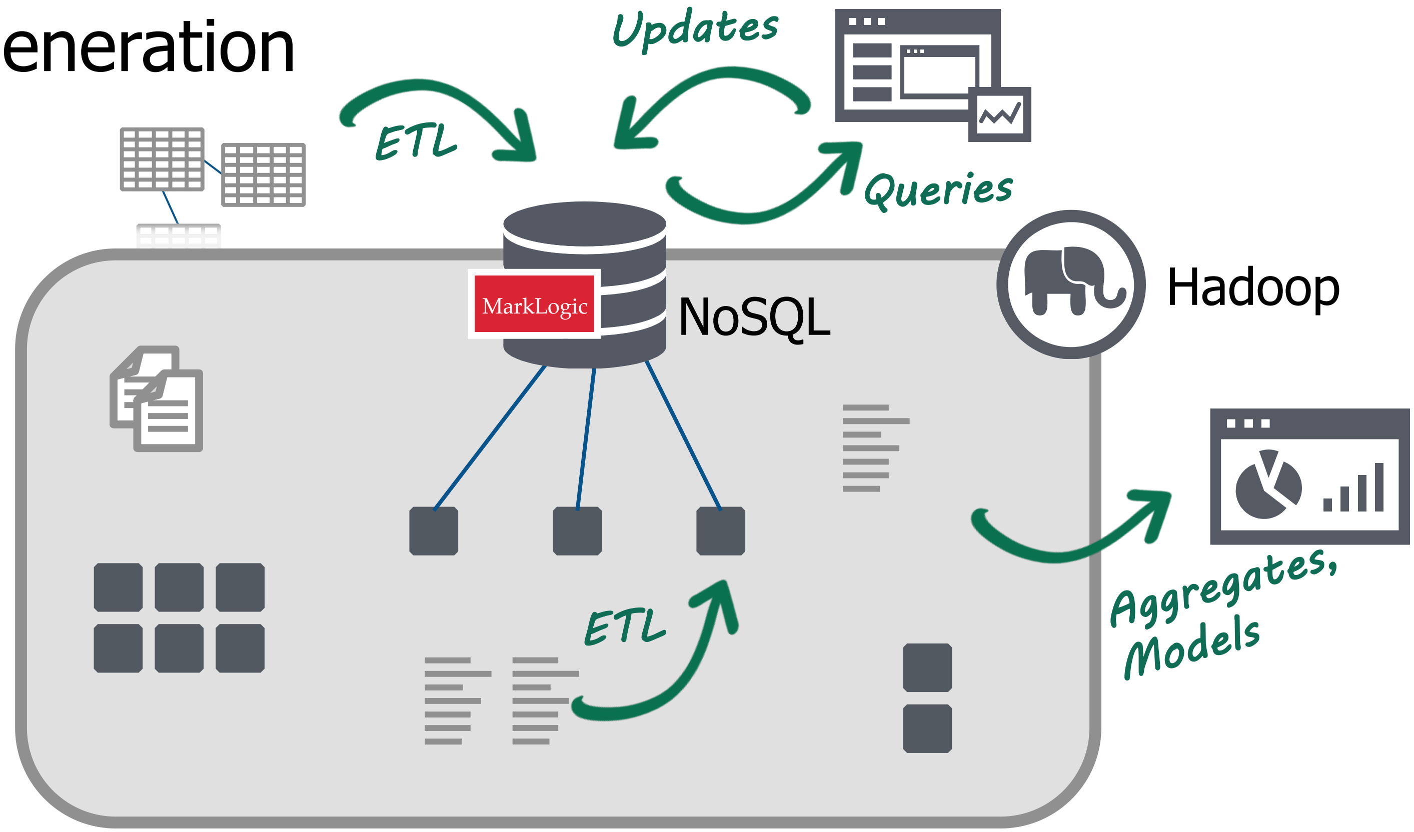
Enterprise-class database for
real-time search, delivery &
analytics

RUN APPLICATIONS

Run mission critical applications
directly on HDFS

MarkLogic

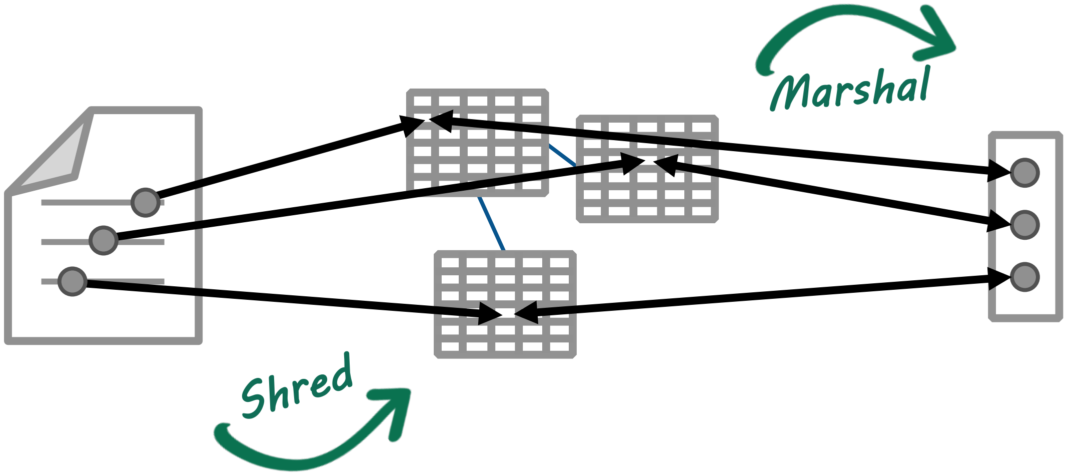
New Generation



MarkLogic

Preserving Context with Documents

- Hierarchy
- Relationships
- Semantics
- Security


Marshal
Shred