Amir Jaibaji – ILG Product Management, Program Director

Stop Data Hoarding

Cleaning up your legacy data

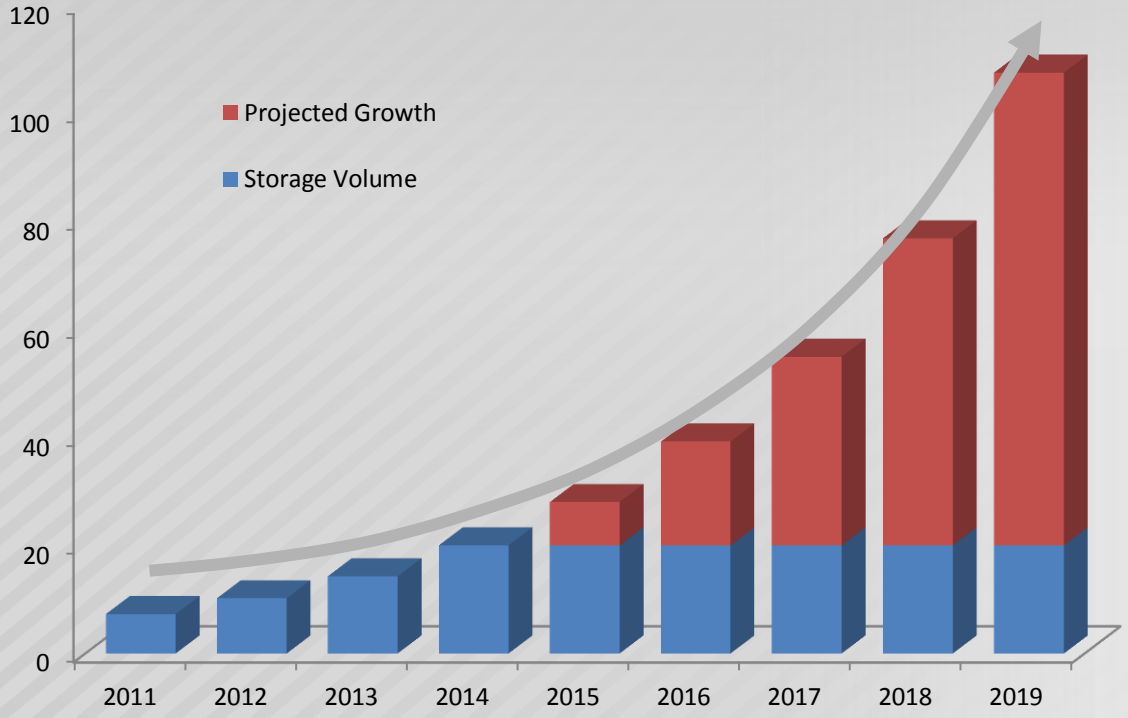
Enterprise big data dilemma

> Data capacity on average in enterprises is **growing at 40 percent to 60 percent year over year** due to a number of factors, including an **explosion in unstructured data**, such as email and documents that have to be stored due to 'regulatory requirements that continue to evolve and change.

SOURCE: Computerworld, "Data growth remains IT's biggest challenge, Gartner says," Lucas Mearian, November 2, 2010.

"Storage is cheap" is no longer the answer

- *Growth is out of control*



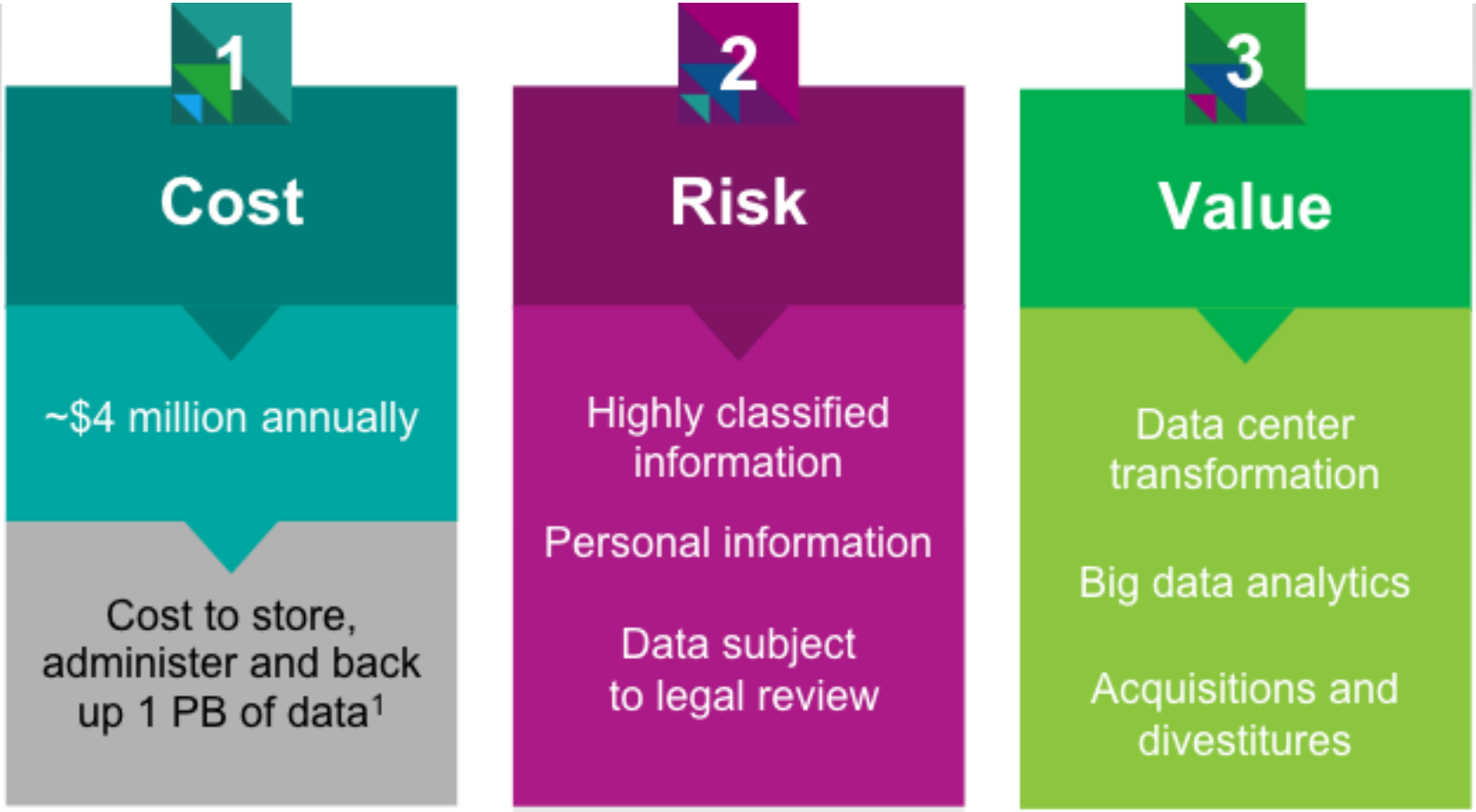
40%-60% Annual Growth

Storage Costs are Consuming IT Budget

Supply-Side initiates are only a temporary stop-gap

- **Virtualization**
- **Over-Allocation**

This growth represents an enormous challenge to IT organizations



Excess information = higher cost and greater risk

Dispose of unnecessary data = **reduce** cost and risk

> The best way to reduce the amount of data—delete it.
>
> —Sheila Childs, Research VP
> Gartner

Where do we start?

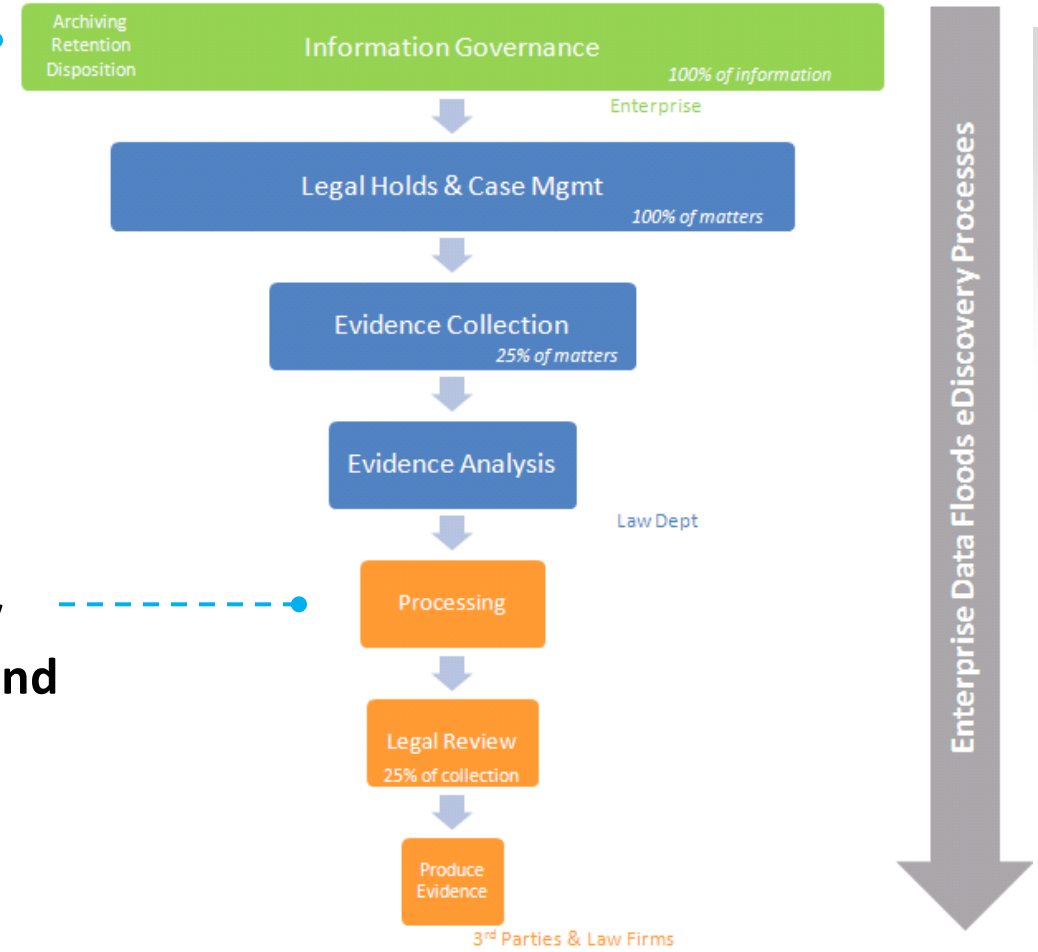
Litigation & Legacy Data Clean Up

Data volume reduction to lower matter costs

Need to cut total information volume and related costs here …

- ~500m discoverable pages per PB
- *Growing 40-60% every year (average)*

… to systemically cut eDiscovery processing and review volume and costs here.



A court vetted solution with numerous client wins

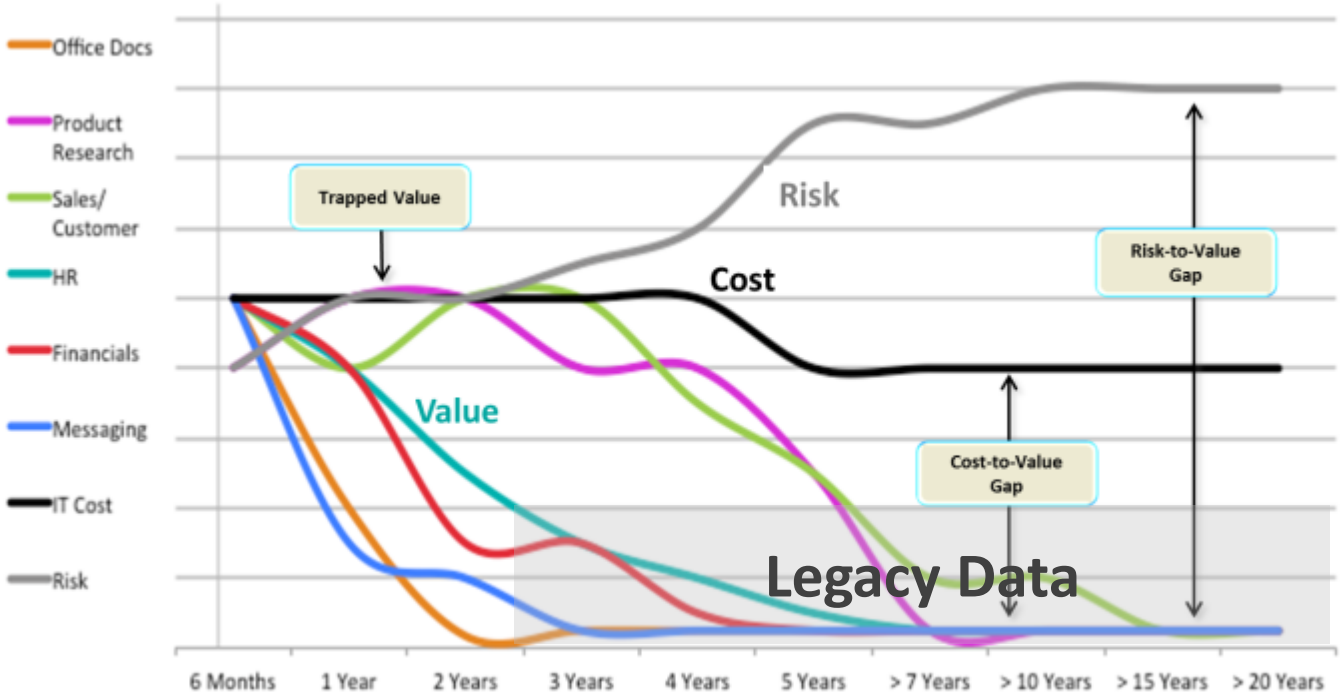
Supported largest litigation case in world by identifying, collecting and analyzing 132 TB of data to produce 200GB of relevant data.

- **PROBLEM:** For The Deep Water Horizon matter, look across 132TB's, 3 continents and 8 locations. Collect 1TB to a preservation location in Houston. Full text indexing and apply additional terms to reduce to the smallest defensible data set which was sent out for production review by outside counsel. Final data set was approximately 200 GB's."

- **SOLUTION**: Enable a 100:1 reduction in collection process in less than 2 weeks.

- **ROI:** Saved million of dollars, responded to every DOJ request; substantially lowered outsourced review costs and built a defensible audit trial.

Legacy data no longer has value but creates cost and risk



Legacy data is:

- Data that has **aged past its usefulness**
- Data that is **consuming COST without providing value** to the organization
- Data of **low value that still carries RISK** in legal actions

Use cases for legacy data cleanup

Clean-up ROT Data

Redundant, Obsolete and Trivial data

Boolean Search

Remediate Regulated Data

PII, PCI, HIPAA, HR, Financial, Records

Pattern Match & Classify

Secure High Business Value Data

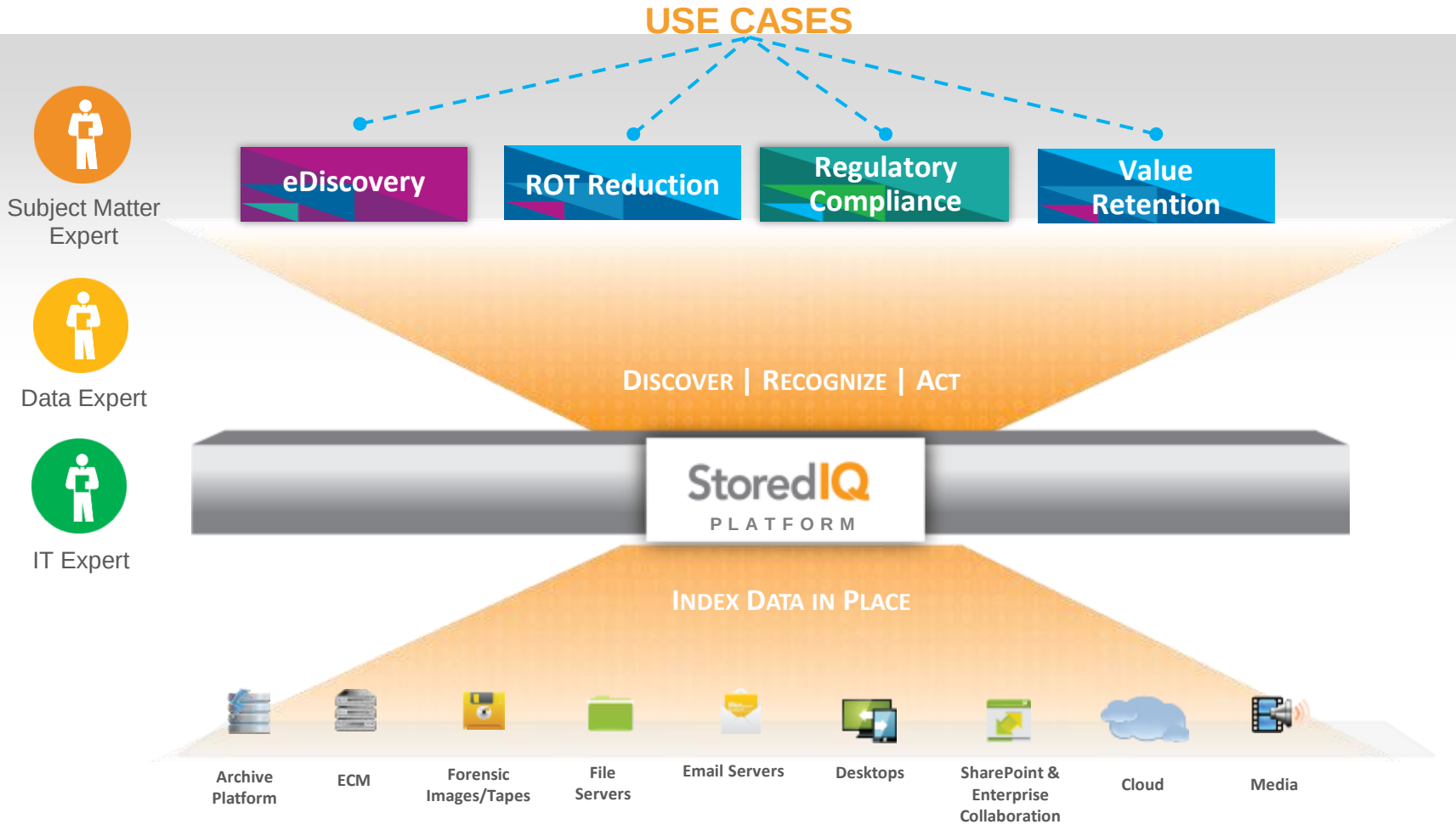
IP, Pricing, Sales and Market, Patent, Planning

Bayesian Classification

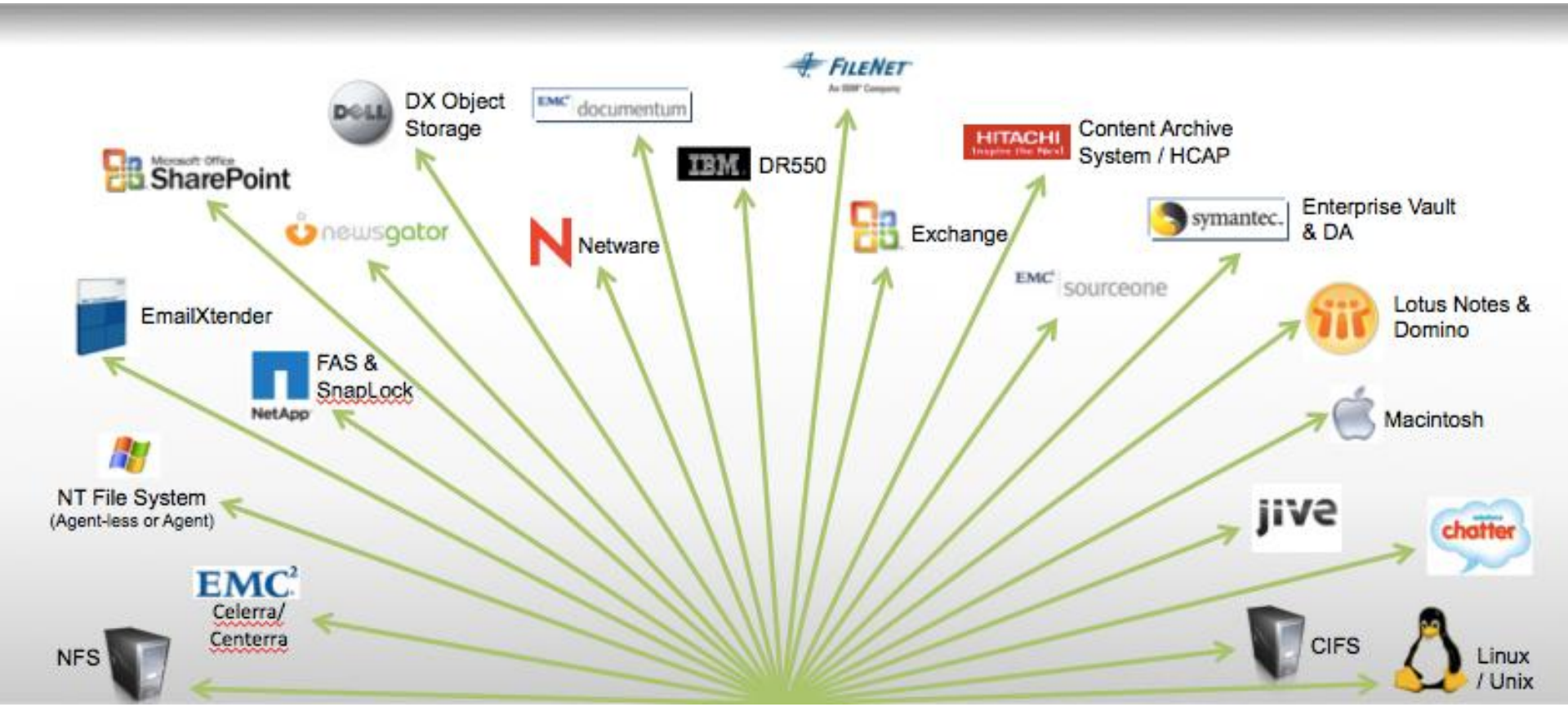
StoredIQ Solution

Clean Up Legacy Data

StoredIQ platform and solutions

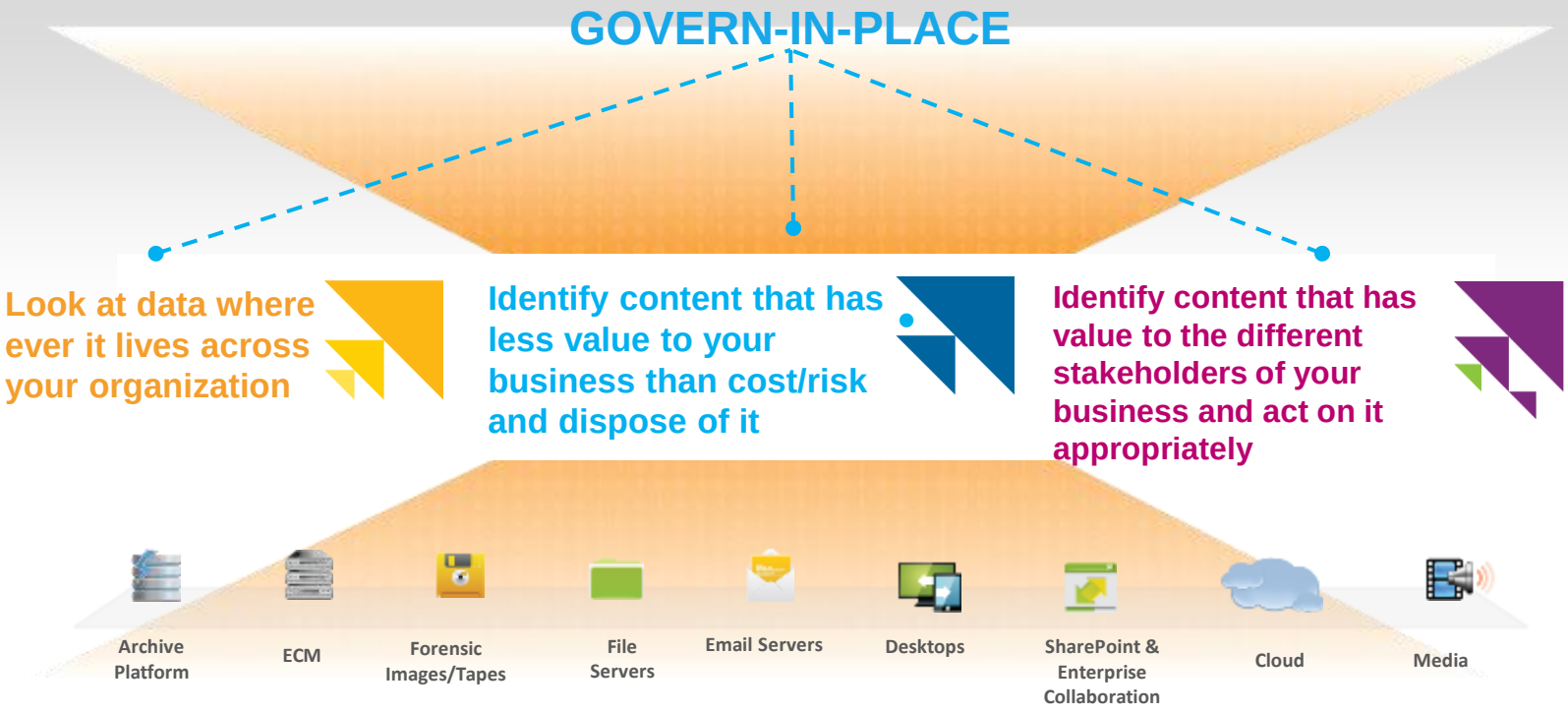


Connect to data in its native location

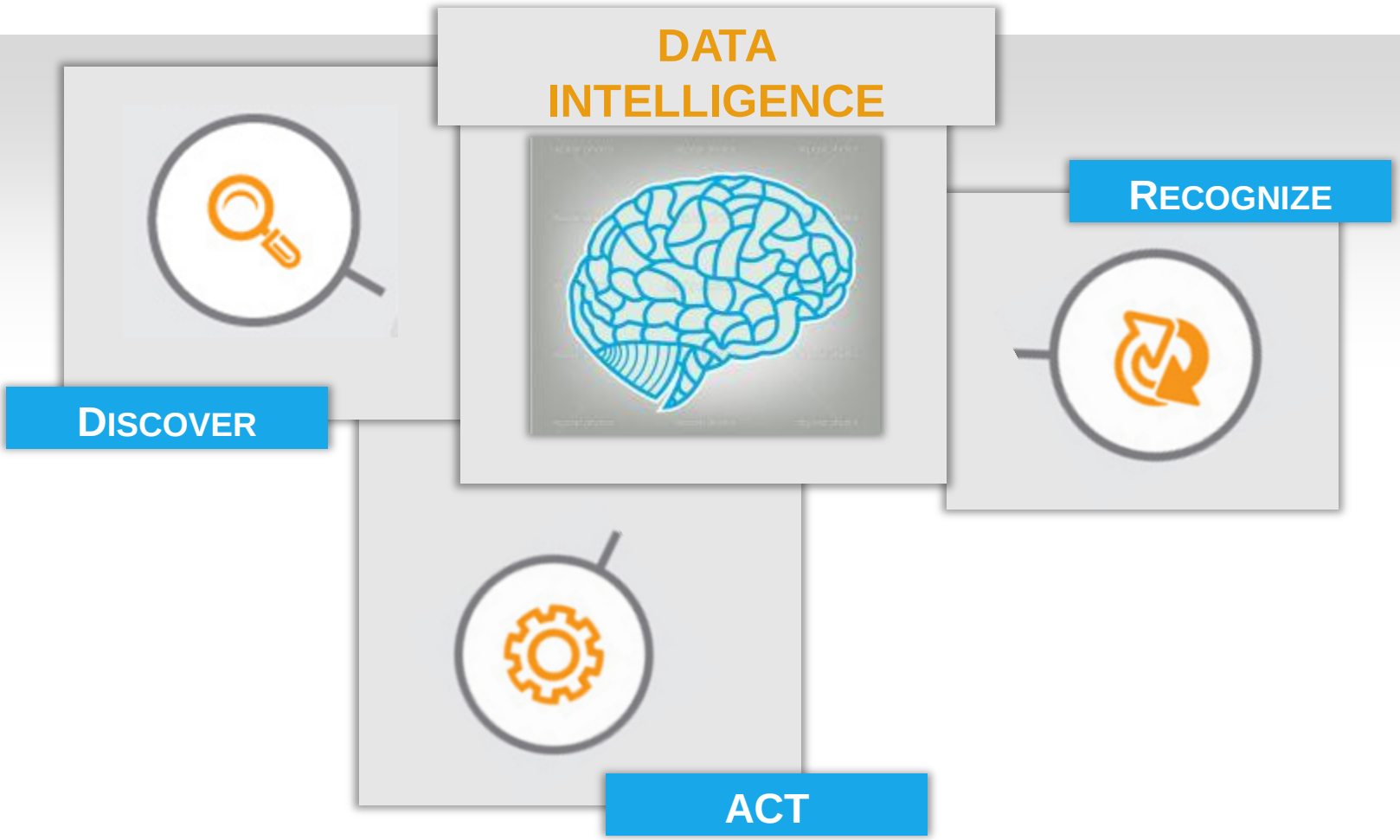


Support for 75+ data sources and 450+ file types

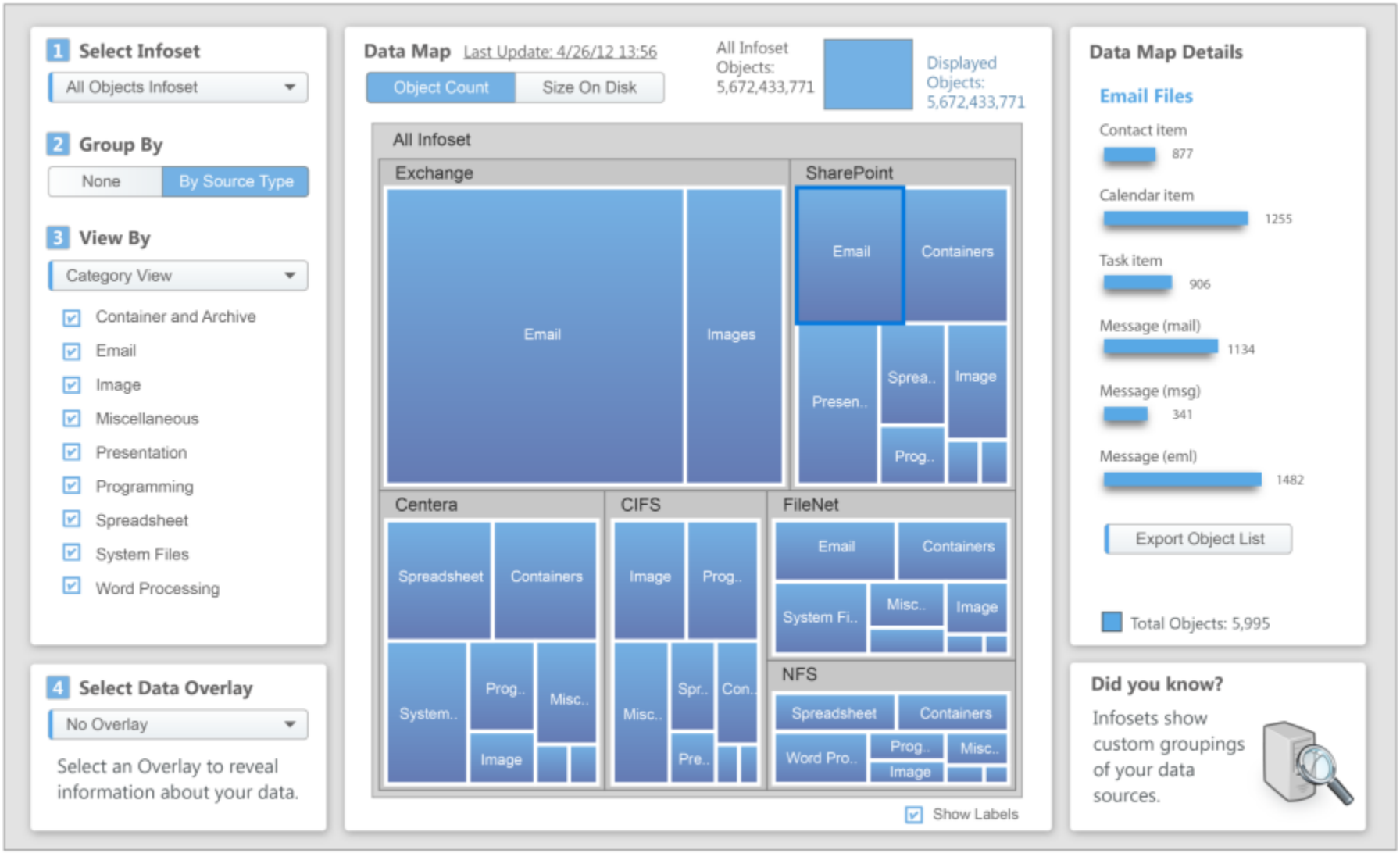
Govern-in-Place



StoredIQ approach: data about data

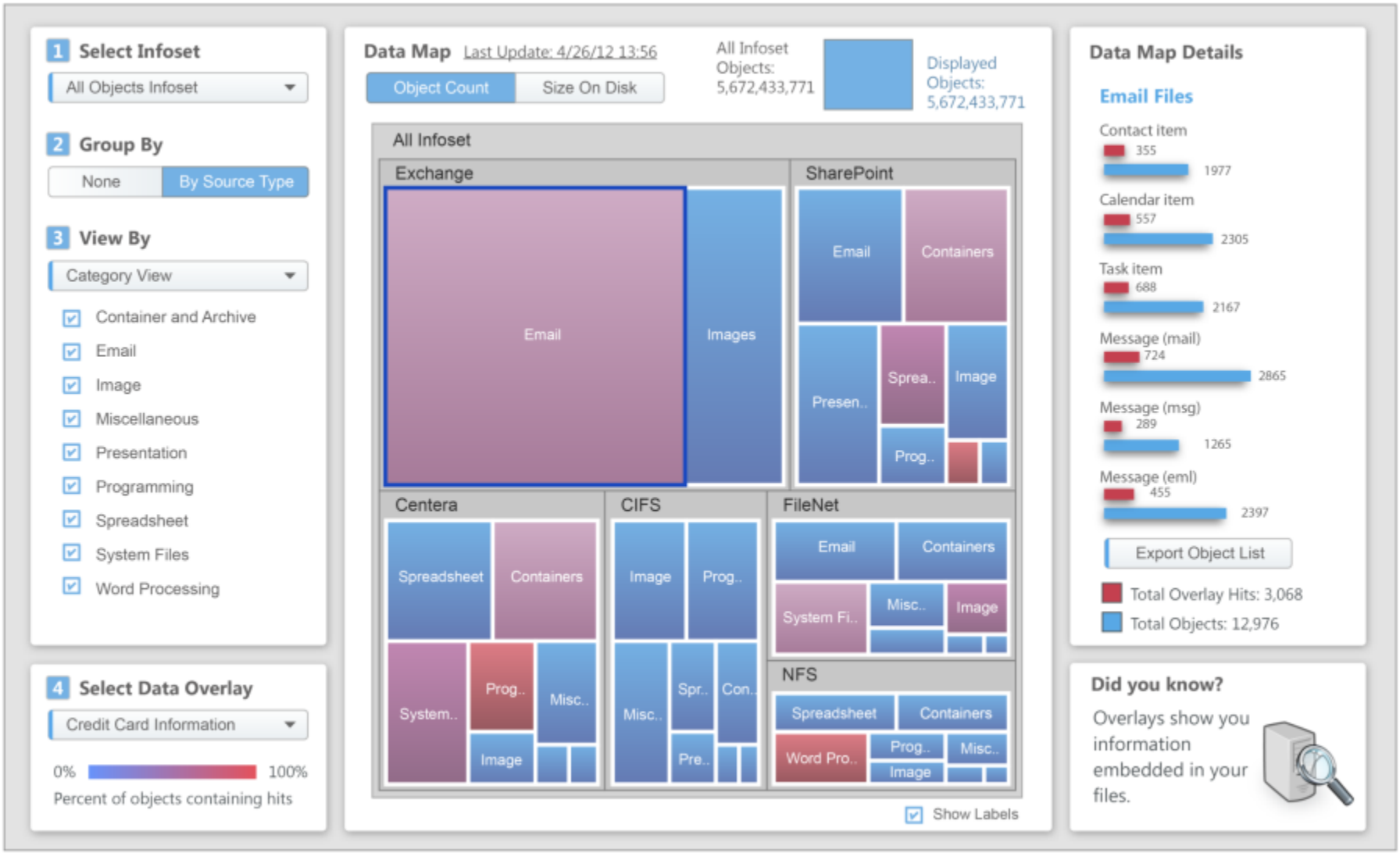


Advanced visualizations show what types of data are stored across your enterprise



Discover where your oldest or least used data resides

Utilize intelligent overlays to spot potential compliance issues



Key benefits

Find the data that matters — Properly discover, classify and manage information according to business value to reduce risk and cost

Get rid of old, obsolete data — Delete nonbusiness, aged and obsolete data to reduce data volume

Identify sensitive and toxic content — Find misplaced client data, PCI and privacy-regulated data

Stratify information to accelerate:

- Cloud migration
- Investigations
- Postacquisition and merger data integration

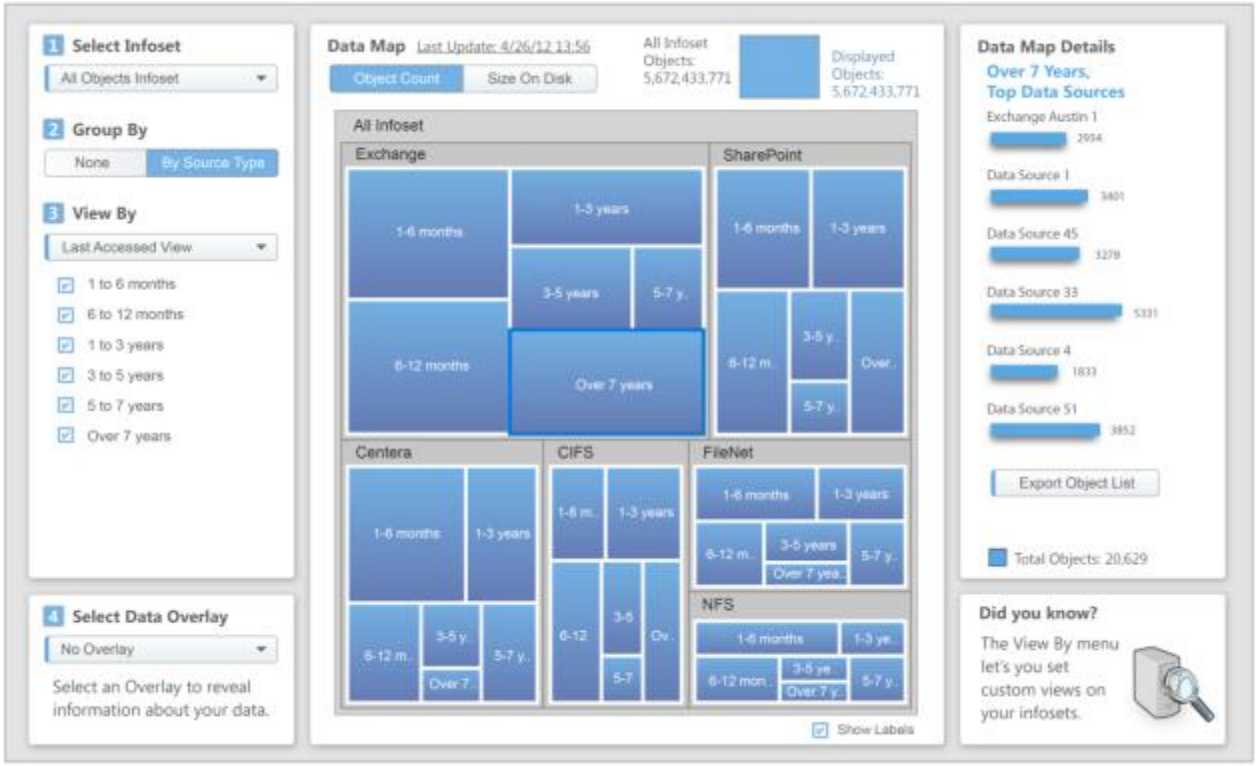
Business process readiness — Practice proactive audit, investigation or disclosure and discovery readiness

ROT Clean Up

Remove Redundant, Obsolete and Trivial Data

Identify data in-place, across multiple repositories AND take action

- **Identify data in 75+ repositories**
- **Categorize data based on metadata and full-text**
- **Create rule sets for data action**
- **Act on data to secure or delete it**



Three-phase reduction of ROT data

Phase 1: Remove trivial data

- Data that never had any value to the organization
- Violates acceptable use policy
- Multimedia files (audio, video and images)
- Temporary files

Phase 2: Move obsolete data for timed disposal

- Orphaned files (employee has left the company)
- Log files
- Files past their longest departmental retention or business value

Phase 3: Remove departmental deduplication

- Create departmental master copy area (ensure that all department employees have access)
- Identify master copies and place in the correct area
- Deduplicate remaining share against the master area

Manage data in a proactive fashion, allowing companies to retain information according to corporate governance and regulatory mandates while disposing of unnecessary data with confidence.

Problem

Risk and cost associated with stale and nonbusiness data in the organization

Solution

StoredIQ identified non-business-related data and enabled the client to create custom classifications for automating file plans and implementing record-retention schedules

Return on Investment

Reduced storage costs through routine data destruction, reduced risk associated with nonbusiness data and ensured overall litigation preparedness.

30%

Reduction in data footprint

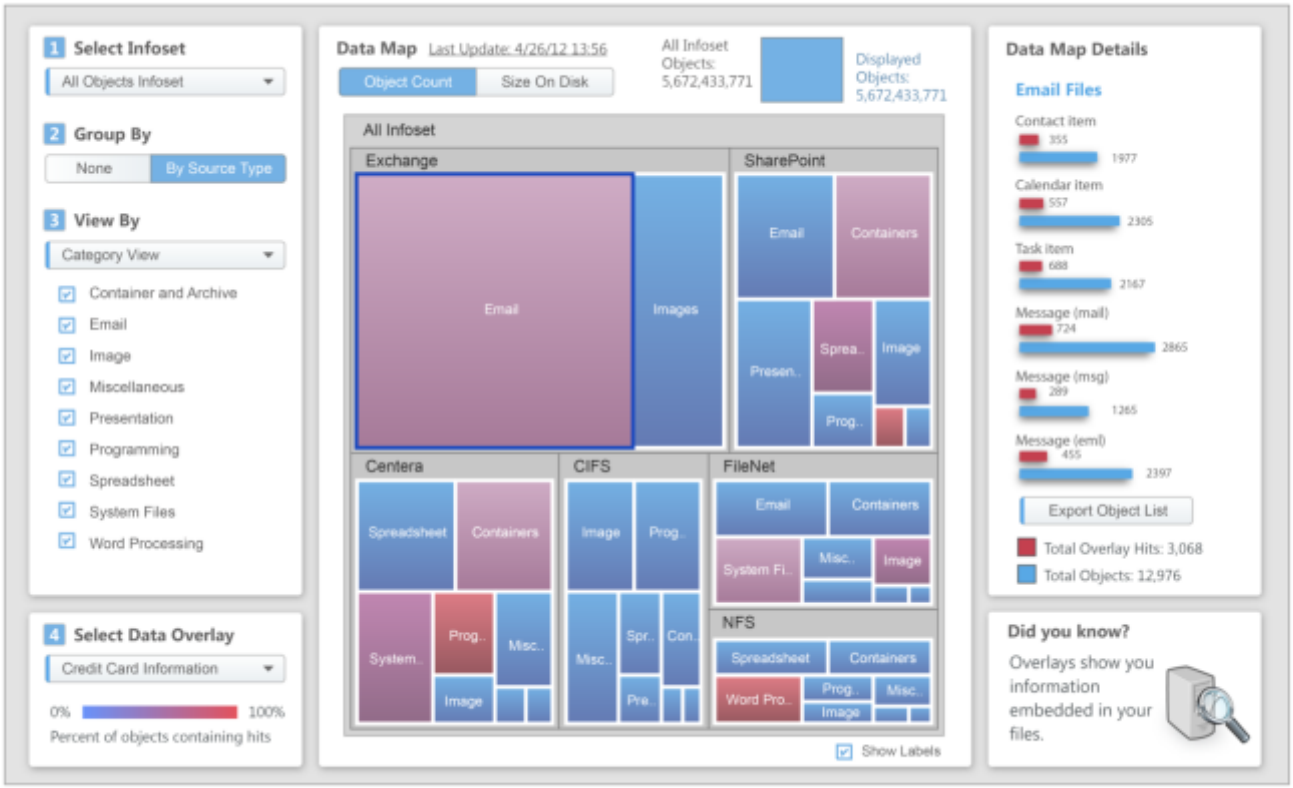
Large financial services organization

Regulated Data

Identify and Remediate Regulated Data

Identify regulated data in-place, across multiple repositories AND take action

- **Identify data in 75+ repositories**
- **Categorize and Classify data based on Boolean search, patterns and machine learning**
- **Create rule sets for data action**
- **Act of data to secure, report, copy, move or delete**



Find and remediate privacy issues

➓ Personally identifiable information (PII)

- Social Security numbers, driver's license numbers, government-issued identification numbers

Are there Social Security numbers on my file shares?

➓ Highly confidential information (HCI)

- Pricing information, engineering, planning, strategy documents

Is customer information being stored inappropriately?

➓ Payment Card Industry (PCI) data

- Credit cards of any type

Is confidential company data at risk?

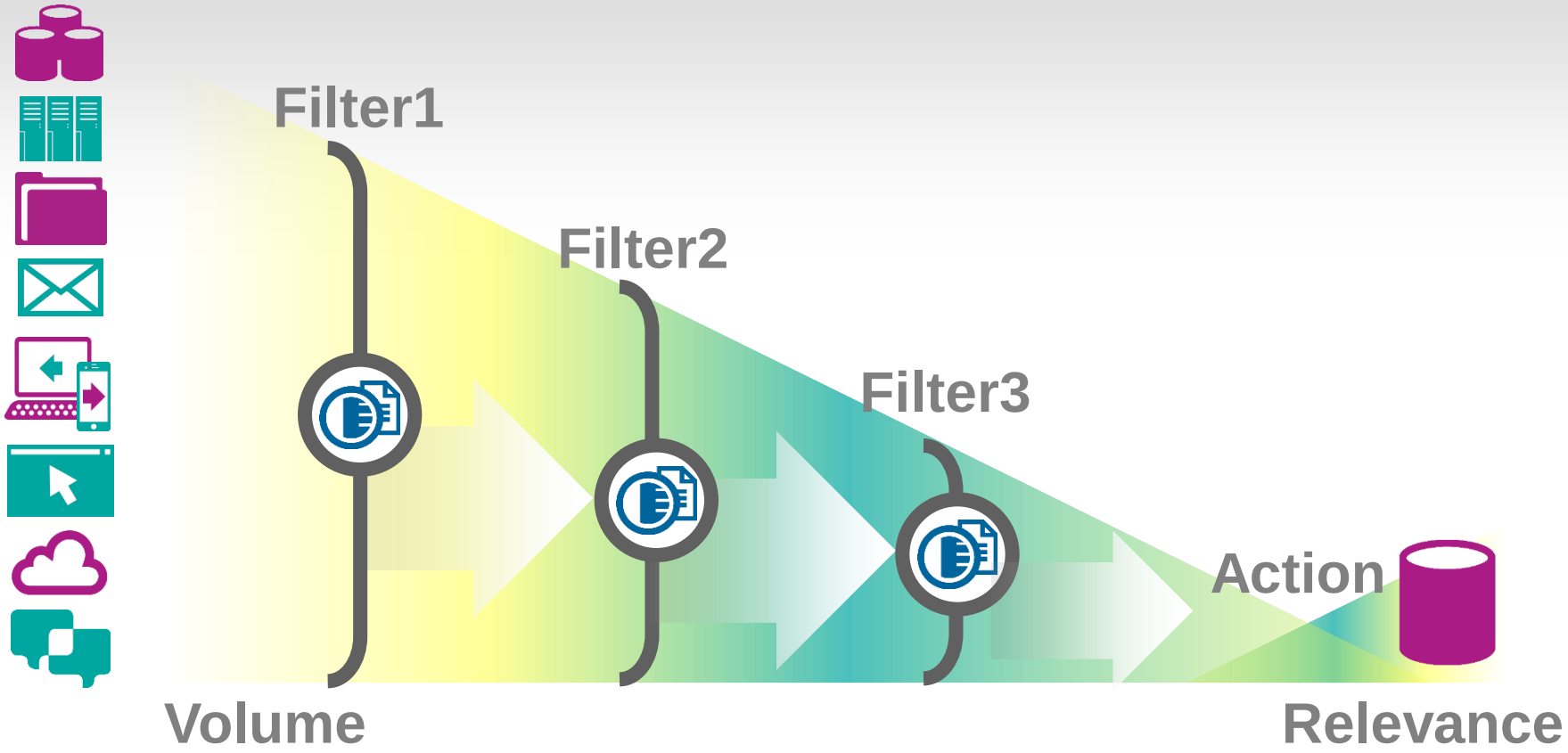
A large energy company identified 21 types of PII, HCI and PCI data in its native location and remediated 17 percent of data.

High Business Value Data

Identify and remediate data with high business value

Classify data to identify business value

❿ *Use a combination of rules and machine learning to identify and classify data of business value, making it readily available for stakeholder needs and data analytics*



Identify, analyze and act on the data that matters

❿ Mergers and acquisitions

- Identify data across more than 75 data source types to be consolidated, protected or remediated

How do I consolidate useful data?

❿ Divestitures

- Identify classified and copied corporate intellectual property prior to divestiture of business units

Where is corporate intellectual property?

❿ Storage migration

- Identify segment data based on type, age and last accessed date prior to storage migration

Which data do employees actually use?

A large global bank consolidated acquired bank data from thousands of desktops to comply with a regulatory mandate.

Please note

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion.

Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

Thank You!