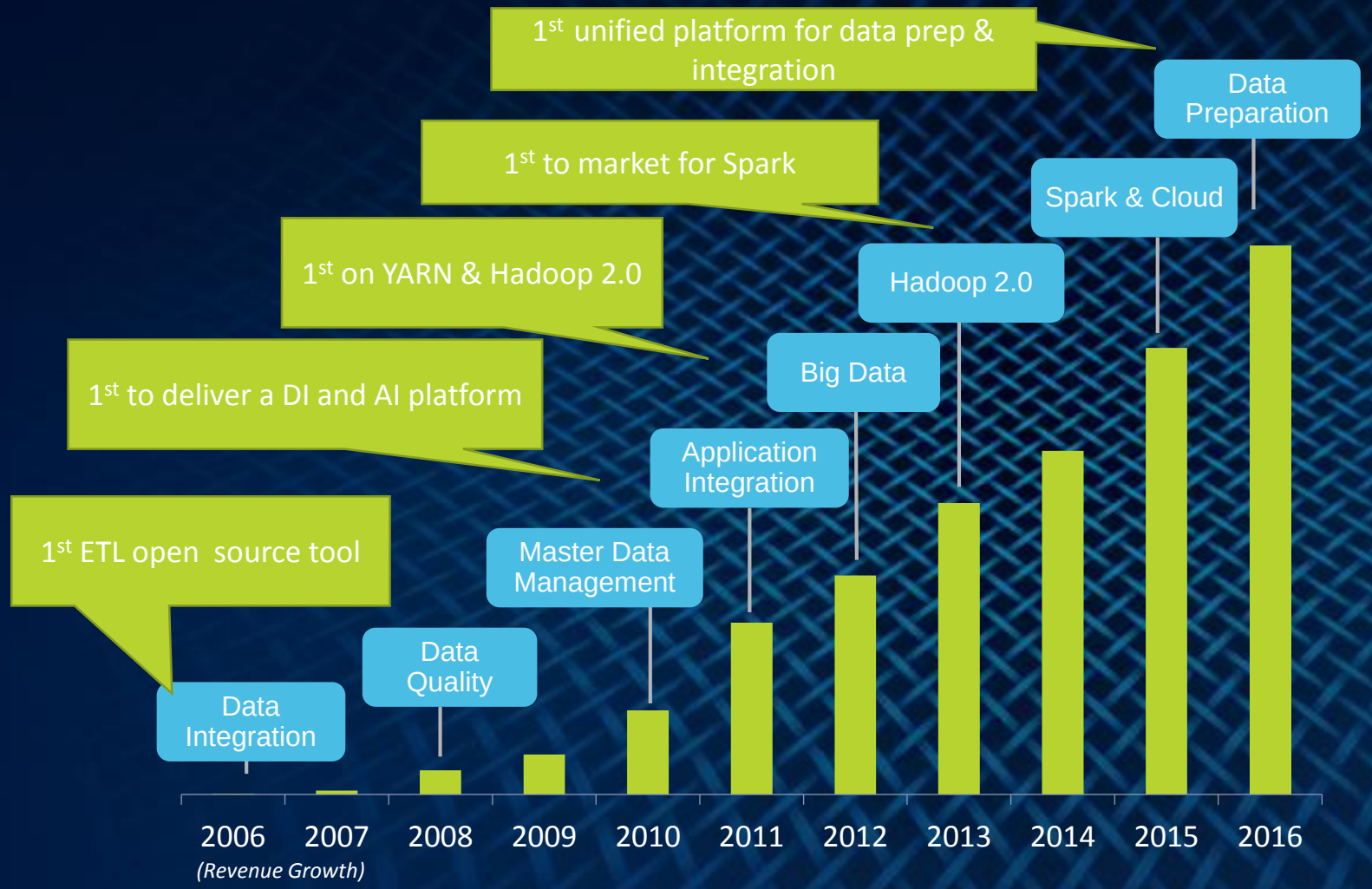
Strata + Hadoop World, 2016

Turning Petabytes of Data into Millions in Cost Recapture for the World's Biggest Retailers

Case Study with PRGX Global

Jonathon Whitton
Director Data Services

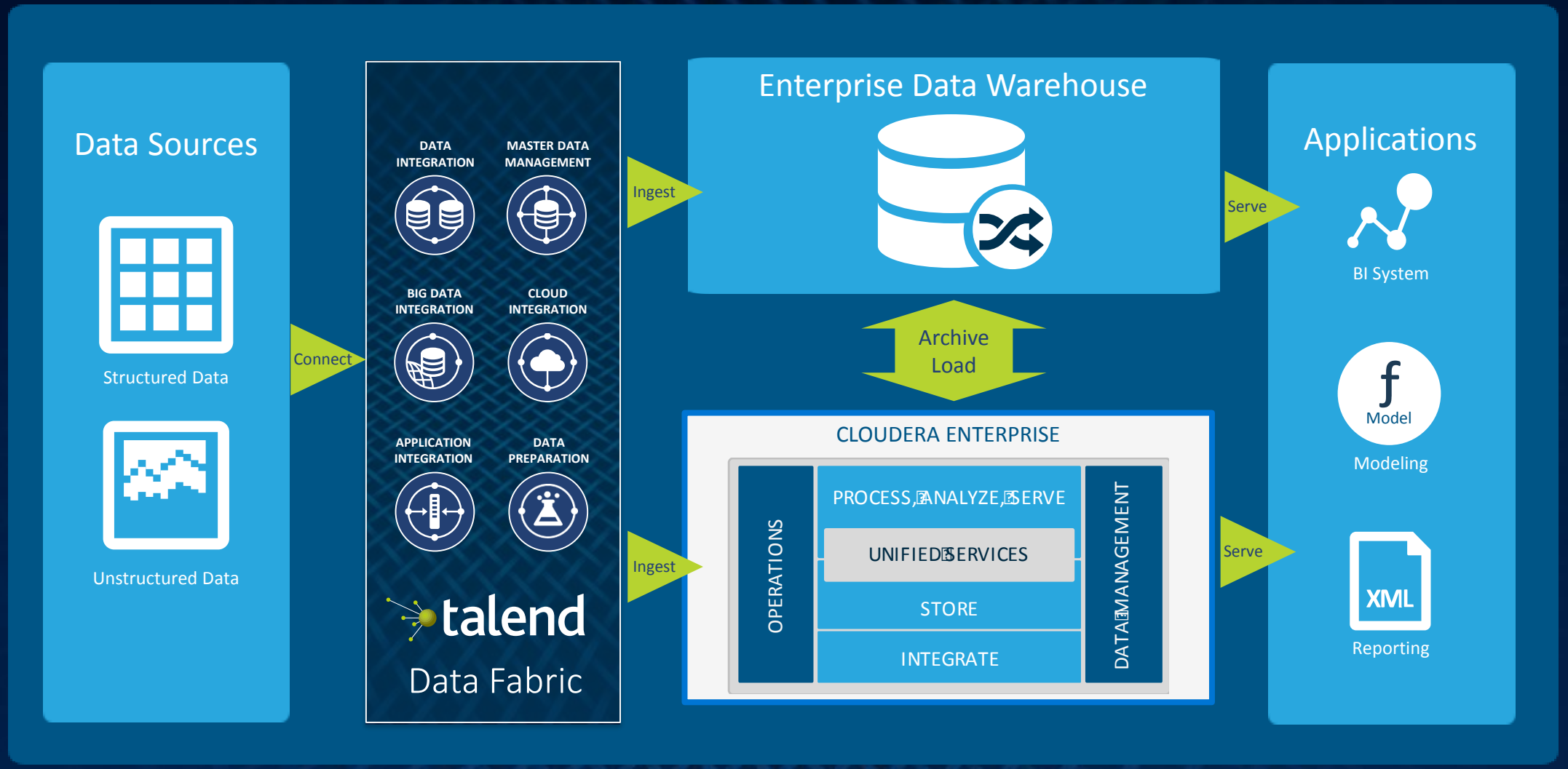
Talend: A History of Delivering Innovation



- Our mission: enable the data driven enterprise
- The most advanced Big Data integration platform
- 10X faster development
- One solution for batch & streaming
- Deploy machine learning in minutes

talend | PRGX®

Talend and Cloudera

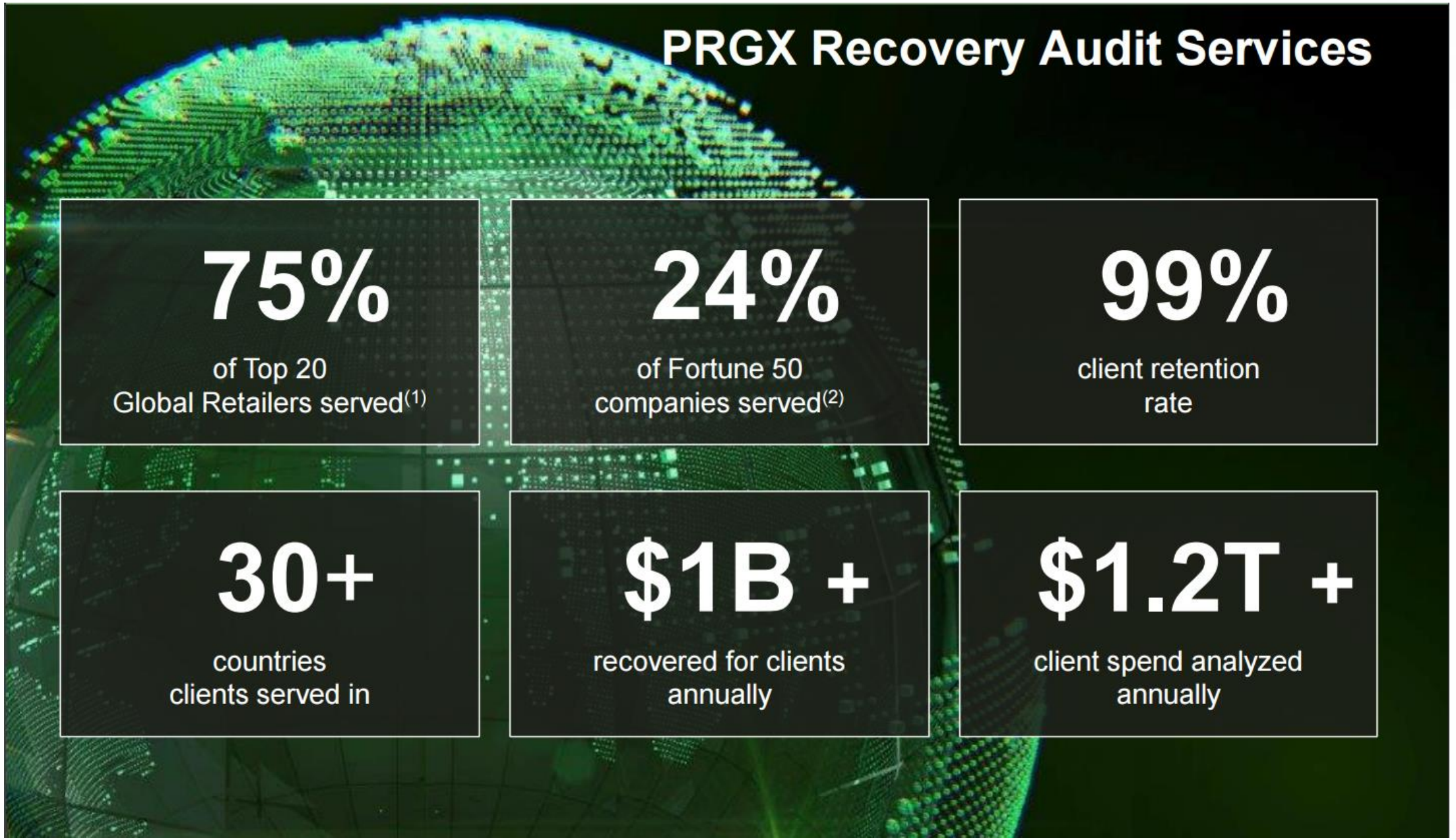


talend | PRGX®

About me

- Jonathon Whitton, Director Data Services at PRGX
- Started working at PRGX in 2000
- BA from Duke University
- MBA from Kennesaw State University

- LinkedIn.com/in/whitton
- Twitter: @JonathonWhitton



talend | PRGX THRIVE IN THE DATA™

About PRGX

- Global leader in accounts payable recovery audit
- Nearly half of recovery audit revenue from outside the U.S.
- In more than 30 countries and across 5 continents
- ~1400 employees

Argentina	Australia	Belgium	Brazil
Canada	Canada French	China	Colombia
Czech Republic	France	Hong Kong	Hungary
India	Malaysia	Mexico	Netherlands
New Zealand	Poland	Portugal	Spain
Taiwan	Thailand	United Kingdom	United States

talend | PRGX THRIVE IN THE DATA™

What Do We Do For a Living?

Aggregate & Manipulate

- Receive over 2 million client files annually
- 2.3 petabytes of data “live” for auditing on average
- Data includes purchasing, payment, receiving, deals, point of sale, and emails

Mine for Overpayments

- Utilize proprietary data mining tools and techniques
- Fully document claim

Recover from Vendors

- Handle majority of vendor communications
- Verify that deductions taken or payments received
- Receive a commission for our services

Data-Driven Business Dealing With Huge Data Sizes…

INBOUND FOR STRUCTURED DATA

5x AFTER PROCESSING

MONTHLY	40 TB	MONTHLY	200 TB
ANNUALLY	480 TB	ANNUALLY	2,375 TB

FILES FOR STRUCTURED DATA

MONTHLY	200,000	ANNUALLY	2,400,000

EMAILS UNSTRUCTURED DATA

MONTHLY	12,500,000	ANNUALLY	150,000,000

talend | PRGX THRIVE IN THE DATA™

… And Huge Data Variety

TYPES OF FILES

EDI
XML
Flat file csv
Flat file delimited
database backups
spreadsheets
Pdfs
Tiff
Jpeg
Png
Prns
Emails
Microfiche
Proprietary formats

% BY FILES

(FOR STRUCTURED DATA)

28% EBCDIC Flat
40% ASCII Flat
25% DB back-ups
7% Proprietary

CHALLENGES

Unexpected
Under- estimated
New schema
Wrong schema

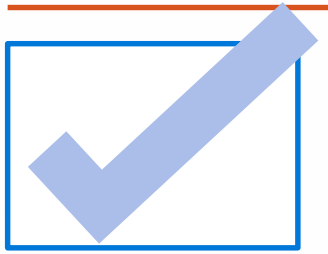
CLIENT DATA ARRIVES

Daily
Weekly
Bi-weekly
Monthly
Quarterly
Semi-annually
Annually

ALSO RANDOMLY

METHOD RECEIVED

Tape
Hard drive
sFTP
Email
Other media – flash drives, DVDs



DAILY WEEKLY BI-WEEKLY MONTHLY

QUARTERLY SEMI ANNUALLY ANNUALLY

talend | PRGX THRIVE IN THE DATA™

Our Legacy Solution Was Designed in The 1990s…

It had all the problems that legacy systems have

- High lead times
- Re-runs take a lot of operations
- Highly labor intensive operations

Tweaked and upgraded over the years

but architecturally remained the same

Transformation process

Significant re-engineering

zero downtime

Our solution had to be

flexible and lightning fast!

HiPer

High Performance Computing

talend | PRGX THRIVE IN THE DATA™

Evaluation on Potential "Architectural Stacks"

19 potential vendors

Short list of 6 vendors

- Functional Requirements
- Information Security Requirements and Privacy Protection
- Performance Requirements
- Technical Requirements
- Vendor Information

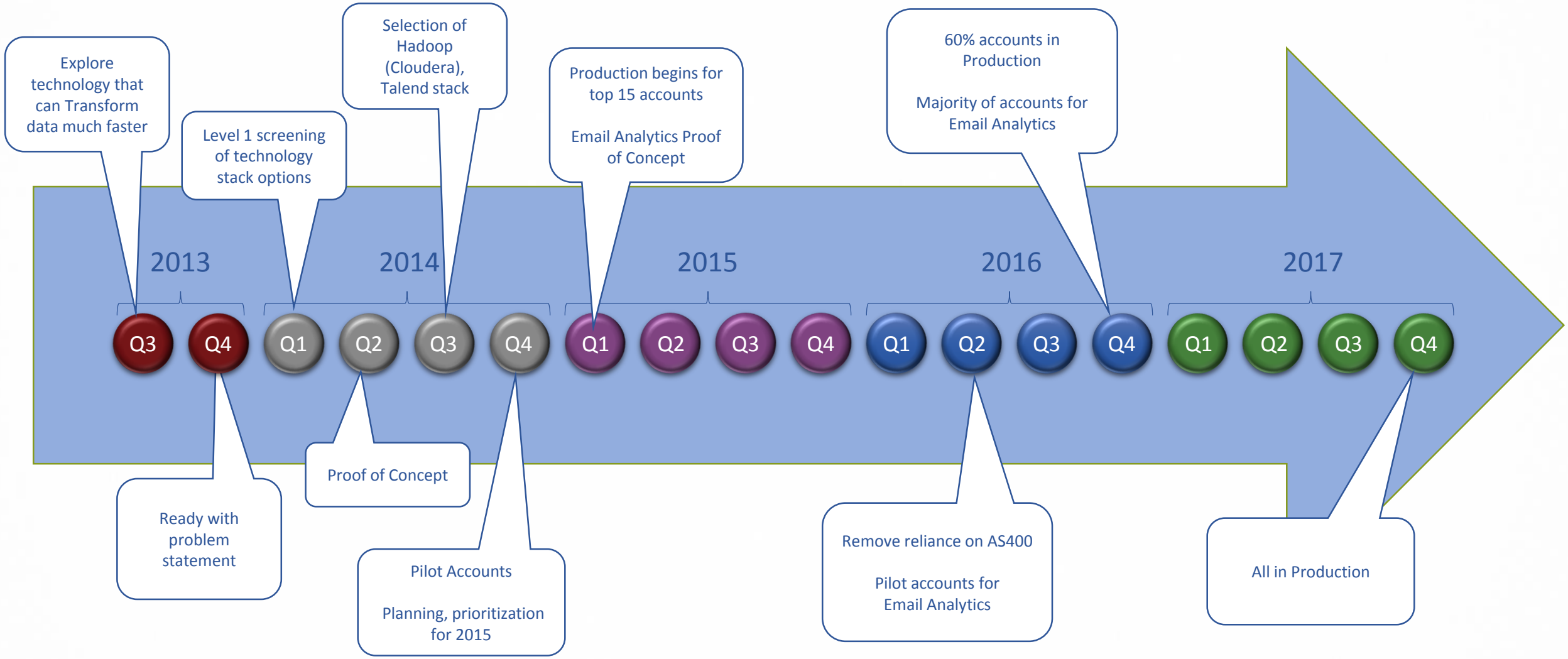
5 POC vendors

Filter 1:
High level assessment focused on vendor's overall company strength, ability for the solution to scale, technology feasibility in PRGX

Filter 2:
Based on willingness of vendor to work with us on a POC and their flexibility

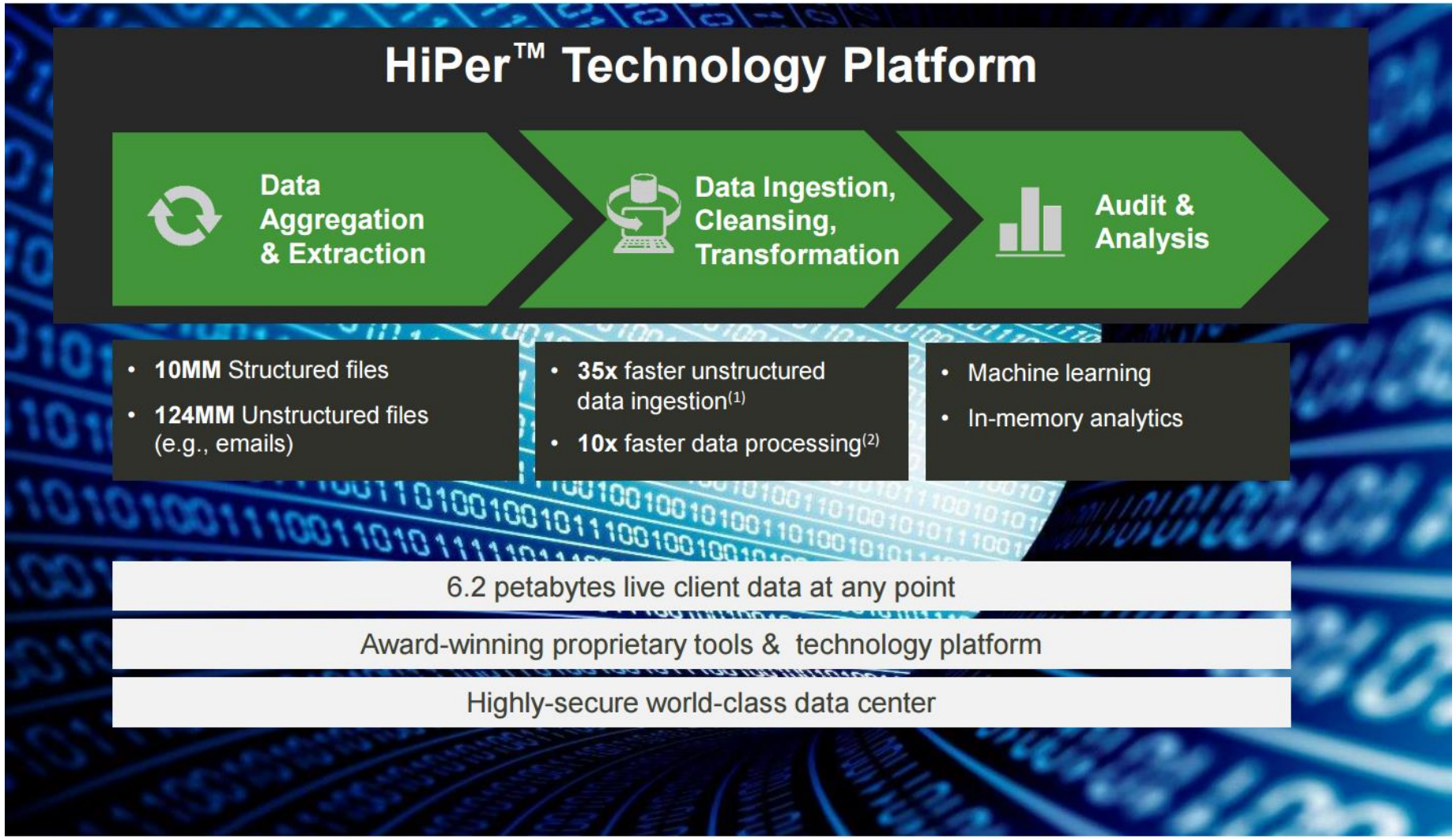
talend | PRGX THRIVE IN THE DATA™

Timeline of HiPer Program



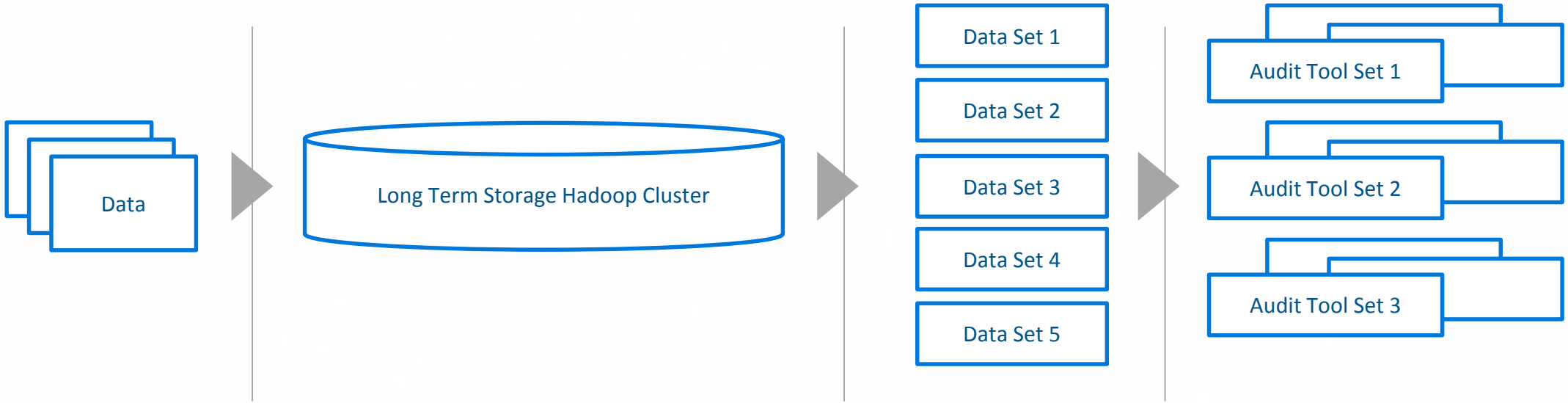
talend | PRGX THRIVE IN THE DATA™

Best-in-Class Technology Foundation



talend | PRGX THRIVE IN THE DATA™

Our Solution For Structured Data Processing



Data Acquisition and Load

Data Transformation

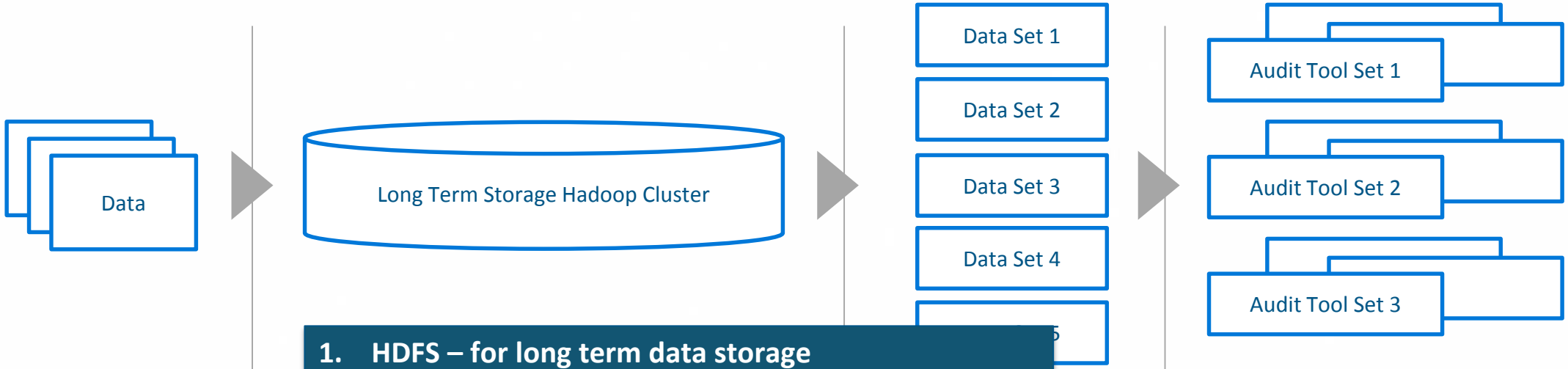
Auditing

Infrastructure management

Data Privacy, Security and Data Lineage

talend | PRGX THRIVE IN THE DATA™

Our Solution For Structured Data Processing



Talend and related automation

1. **HDFS – for long term data storage**
2. **Batch processing – Talend, Hive, and Spark**
3. **Investigative queries (QC) –Impala**
4. **Output to RDBMS –Sqoop and Talend**

Auditing

1. **Continue with legacy technology (MSSQL)**
2. **Explore use of Hadoop BI tools**

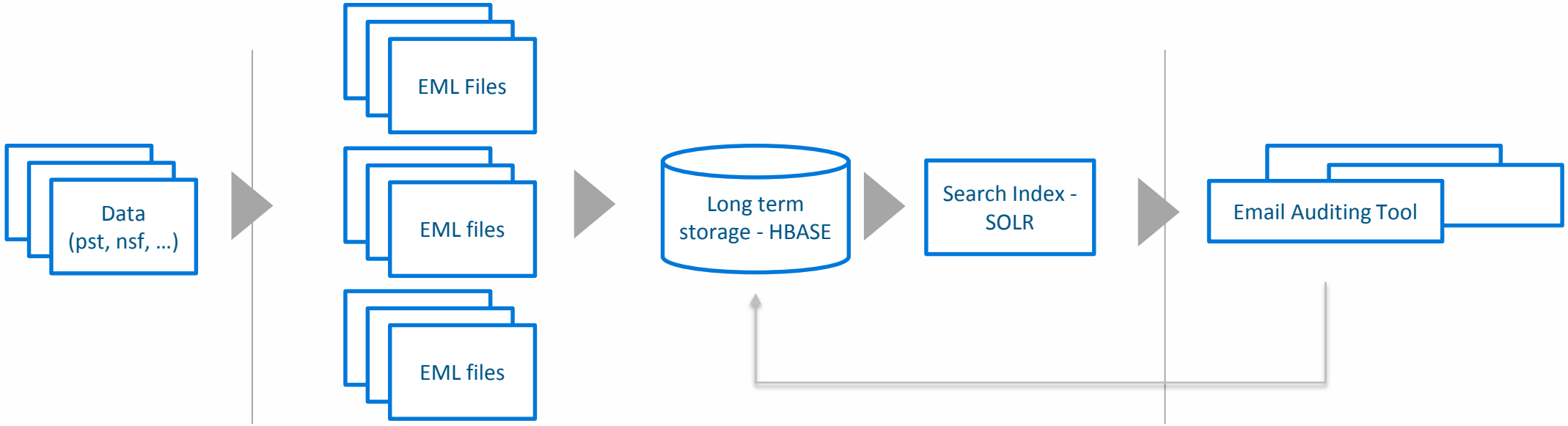
Cloudera Manager

Infrastructure management

1. **Cloudera Navigator**
2. **Kerberos**
3. **Encryption at Rest**

talend | PRGX THRIVE IN THE DATA™

Our Solution For Unstructured Data Processing



Data Acquisition and Load

Data Transformation

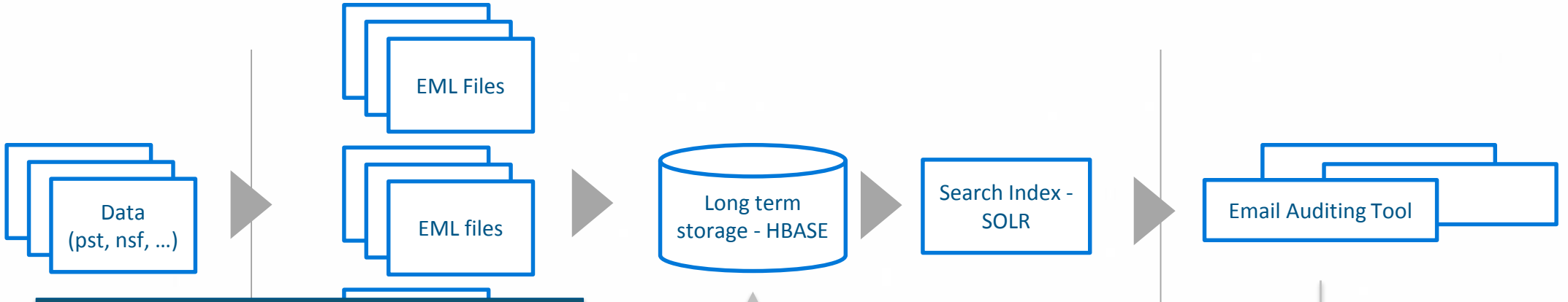
Auditing

Infrastructure management

Data Privacy, Security and Data Lineage

talend | PRGX THRIVE IN THE DATA™

Our Solution For Unstructured Data Processing



1. In-house automation logic in .Net to convert files to eml files
2. Load to Hbase via Apache Thrift
3. Apache Tika to Read Email Data (Headers, Body, Attachments).

1. Hbase – email storage and updates to individual records
2. NGData Keystore indexer (Lily) to update Solr in near real time
3. SOLR – index to query emails
4. SOLRNET - .Net Api to SOLR querying
5. Apache Thrift - Update to HBASE for audit findings

1. Refreshed .Net application to read from SOLR/update to HBASE

1. Cloudera Manager

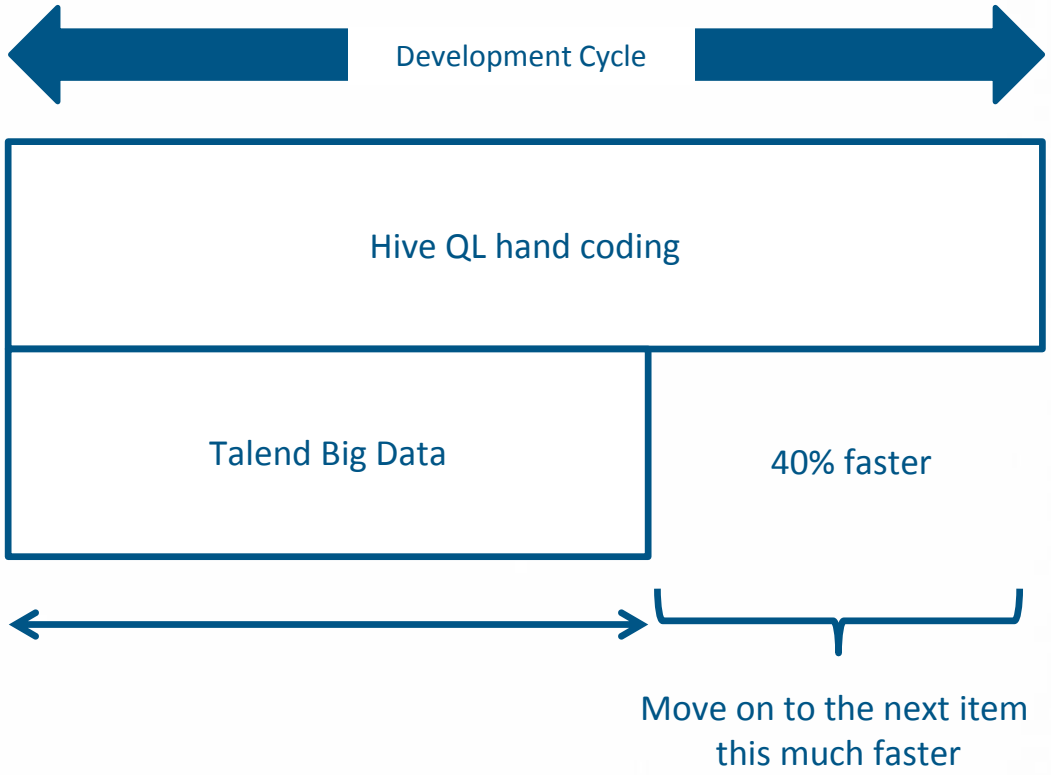
1. Kerberos
2. Mutiple Solr indexes / Hbase Tables

talend | PRGX THRIVE IN THE DATA™

Talend Impact on Our Development

Development Cycle

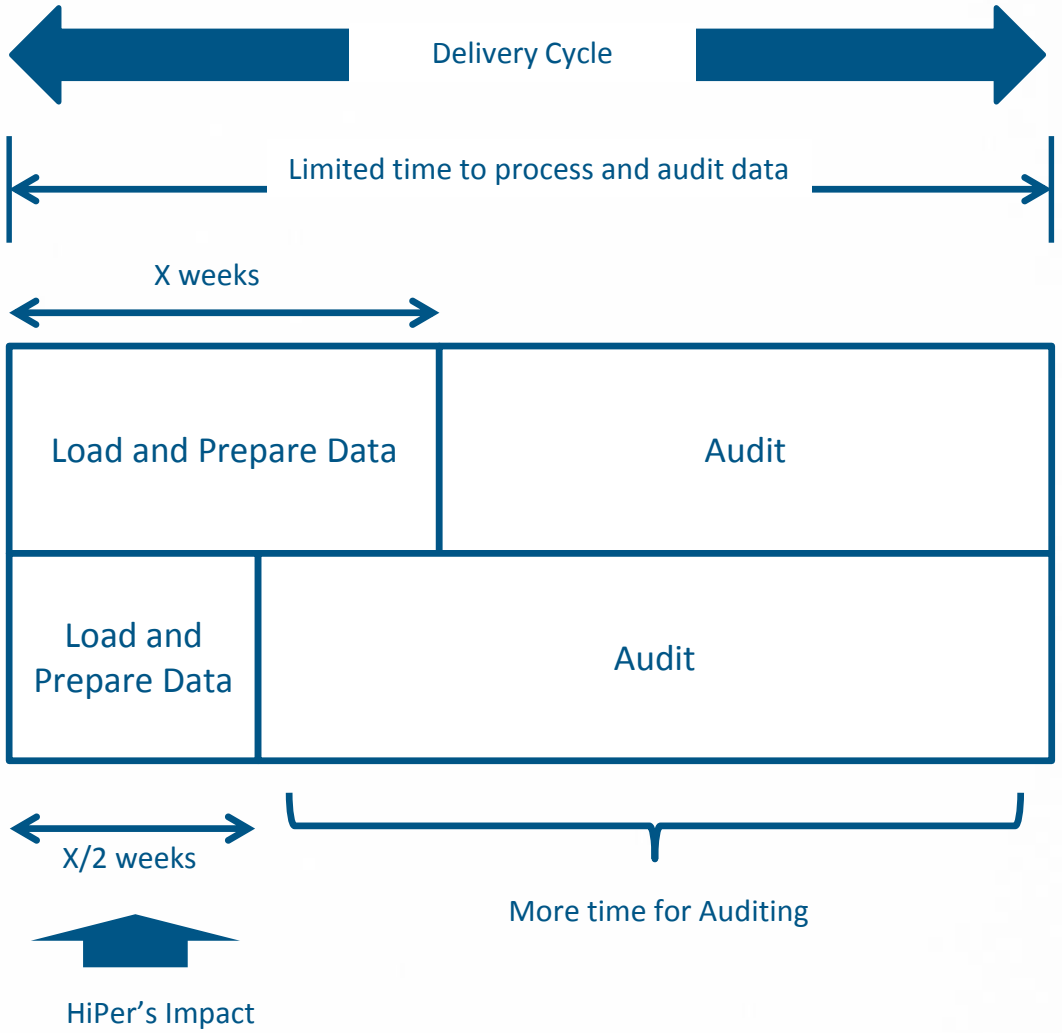
- Reduced developing processes in Hadoop by at least 40%.
- With over 200 distinct process to get through this year, the time savings is very meaningful in achieving our goals
- No more hand coding Hive QL
- Talend has the logic de-coupled from the code that is executing on the Hadoop cluster
- “Upgrade” to Spark via dropdown and then addressing limited component changes that are clearly identified
- “Upgrade” to the next big thing



Hadoop & Spark Impact on Our Business

Structured data processing

- Improved performance in data delivery for structured and unstructured data about 9-10 times the original timelines via Hive
- Starting to use Spark



talend | PRGX THRIVE IN THE DATA™

Real World Example

Faster analysis in Hadoop

- Point of Sales (POS) data was aggregated
  - list of stores
  - number of transactions
  - sum of quantity
- Would have run for hours in our AS400 environment
- 2014 would not have been readily available before Hadoop's lower cost of storage made it affordable for us to keep more data online
  - No need to go back to our archive
  - No need to request a restore

2015 data

13 months: 20141201 to 20151231
Row count: 2,048,666,122
Returned: 1385 rows returned (1 per store)

Returned in: 90 seconds

2014 data

13 months: 20131201 to 20141231
Row count: 2,159,626,445
Returned: 1365 rows returned (1 per store)

Returned in: 90 seconds

Field count in source POS data was 14

talend | PRGX THRIVE IN THE DATA™

Impact on The Future of Our Business

Structured data processing

- Support for “machine-learning” techniques in the email and buyer pulls significantly enhancing productivity for audits

Expanded focus

- Key enabler to mine large volumes of data from our customers
- Ability to re-run and experiment on large sets of data
- Framework for industry standard Master Data

Infrastructure

- Backup and archival solution for the future
- Begin standardization and reusability of processes; eliminate person dependency

talend | PRGX THRIVE IN THE DATA™

Key Success Factors

- Create the business case…
- For multiple years, with quick wins
  – no IT programs succeed without business backing
- Plan the program – for multiple years
- Architect it right
  - Architect keeping your current process and tools in mind
  - Architect it for different uses
  - Architect it for time!
- Find the right partners and develop the partnerships
  - Sell your program…you need champions everywhere
  - CEO/Board to your business, to marketing to even HR

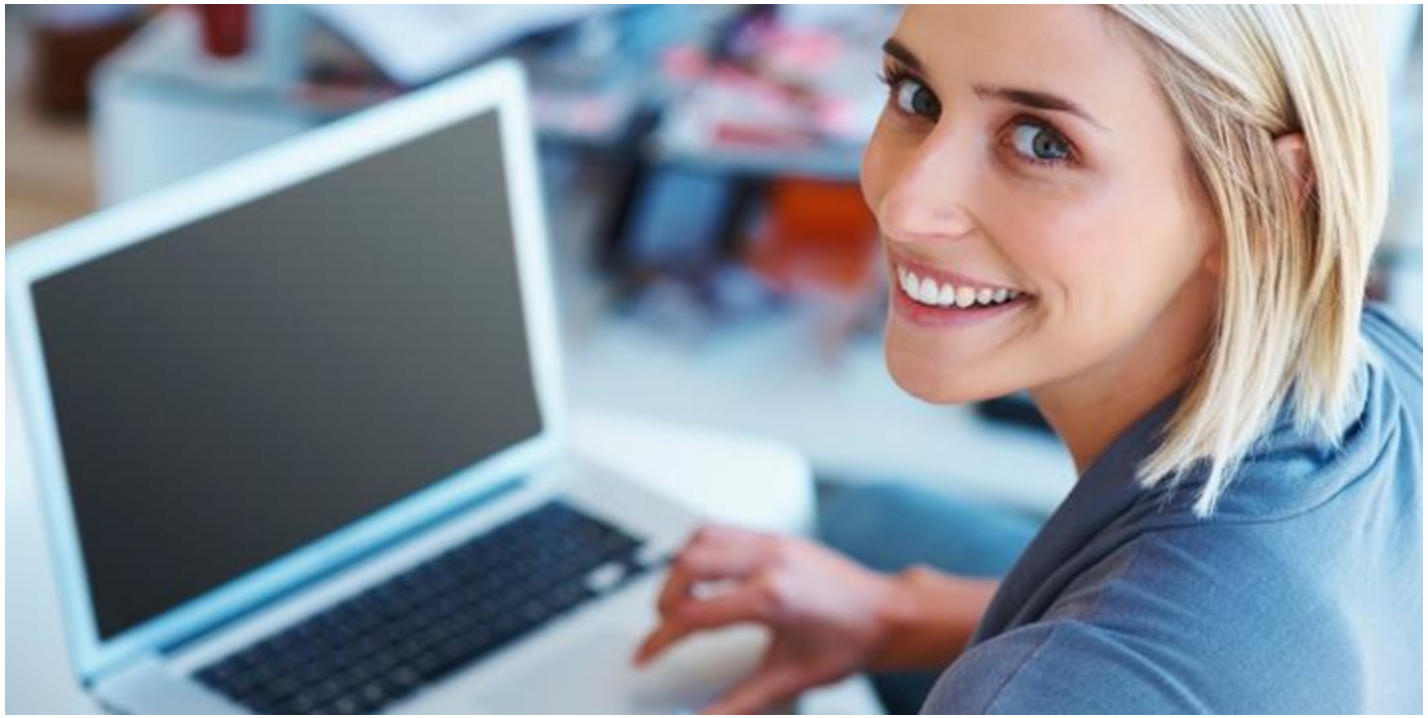
Why Talend and Hadoop + Spark?

- Scalability
  - Cost of storage was reduced by 1 / 5
  - We are now storing three year's worth of our clients data in Hadoop. Data that is ready for immediate analysis and manipulation
- Performance
  - Increased performance 10x on average and 42x faster high water mark, so far
- Support
  - Comprehensive support that covers questions to roadblocks

talend | PRGX THRIVE IN THE DATA™

Our Next Step in The Data Journey

- Self-service data access
- Talend Data Prep



talend | PRGX THRIVE IN THE DATA™

Thank You