PASS SUMMIT2018

Cleaning up your Big Data Lakes with Data Quality as a Service

How to “plug in” DQaaS

Paul Bertucci, Principal Architect, Data by Design

Please silence cell phones

Explore everything PASS has to offer

Free Online Resources

Newsletters

PASS.org

24HOURS OF PASS

Free online webinar events

PASS LOCAL GROUPS

Local user groups around the world

PASS SQLSATURDAY

Free 1-day local training events

PASS VIRTUAL GROUPS

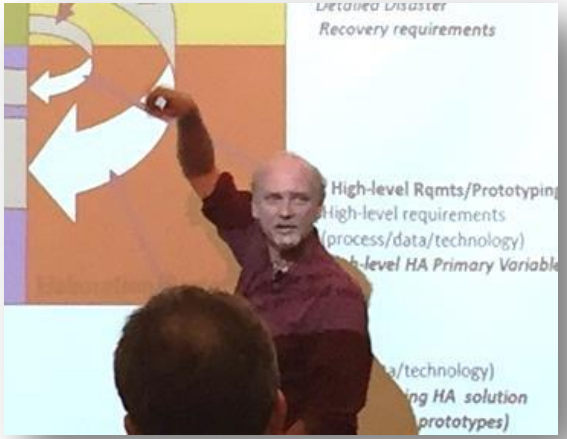
Online special interest user groups

PASS MARATHON

Business analytics training

PASS VOLUNTEERS

Get involved



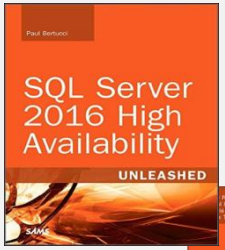
Paul Bertucci

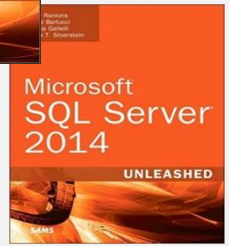
Principal Architect, **Data by Design LLC**

www.dataXdesign.com

linkedin.com/in/paul-bertucci-59738b

@ptbertucci





Author

- **"SQL Server UNLEASHED"** series
- "**High Availability Unleashed**",
- and numerous others.

Architect/CTO

- Former Chief Architect at Autodesk
- Former Chief Data Architect at Symantec
- CTO for Diginome, PointCare, LISI, others

Instructor and Course Author

- Advanced SQL
- Performance & Tuning
- High Availability
- Data Replication
- Data Modeling and Database Design
- Master Data Management

Agenda

- Big data is here to stay and expanding rapidly
- The 4th "V" of big data
- How your data architecture is growing
- Big data, and perhaps a big mess!
- Data quality as a Service in your data pipeline
- Tools of the trade and results you should expect

Big Data's Rapid Expansion

WHAT'S A ZETTABYTE?

1 kilobyte	1,000
1 megabyte	1,000,000
1 gigabyte	1,000,000,000
1 terabyte	1,000,000,000
1 petabyte	1,000,000,000
1 exabyte	1,000,000,000
1 zettabyte	1,000,000,000,000,000,000,000

SOURCES: CISCO

Digital Data (created and replicated)

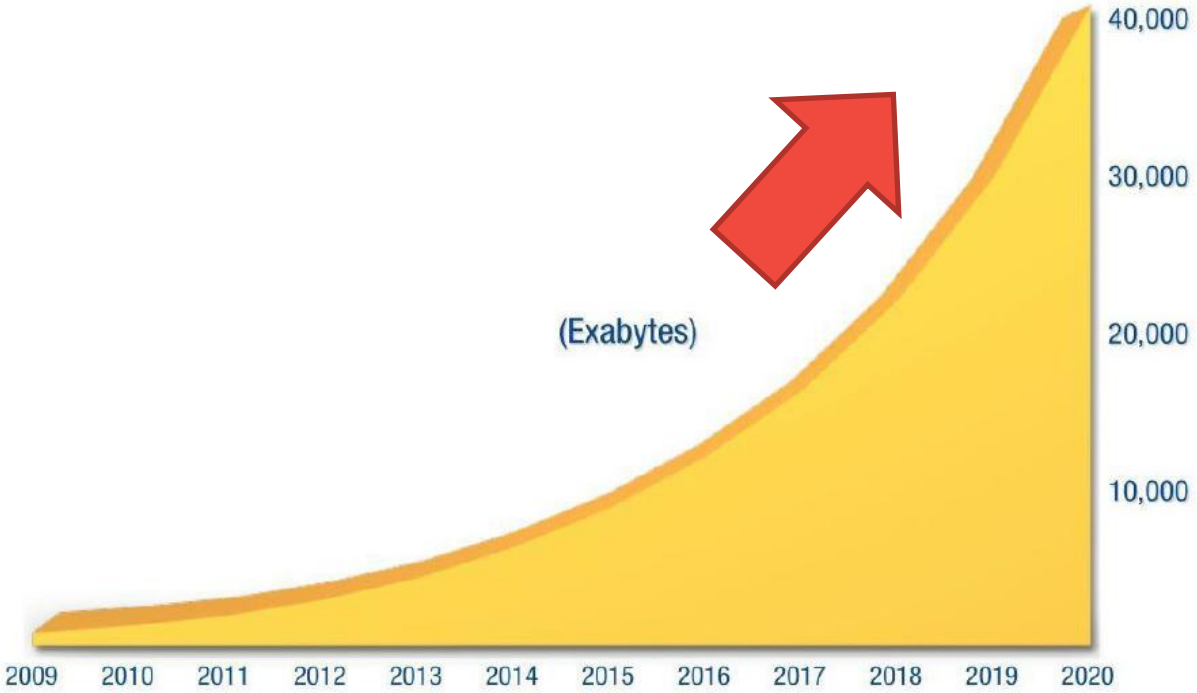
- Reached 4 ZB's at the end of 2013
- That's 50% more than in 2012
- And, 200% more than in 2010
- Will reach 40 ZB's by 2019
- Will approach 500 ZB's by 2025 (and generating 120 ZB's annually)
- This will only continue to grow!

Source: IDC

PASS

Just getting warmed up

(40 Zettabytes)

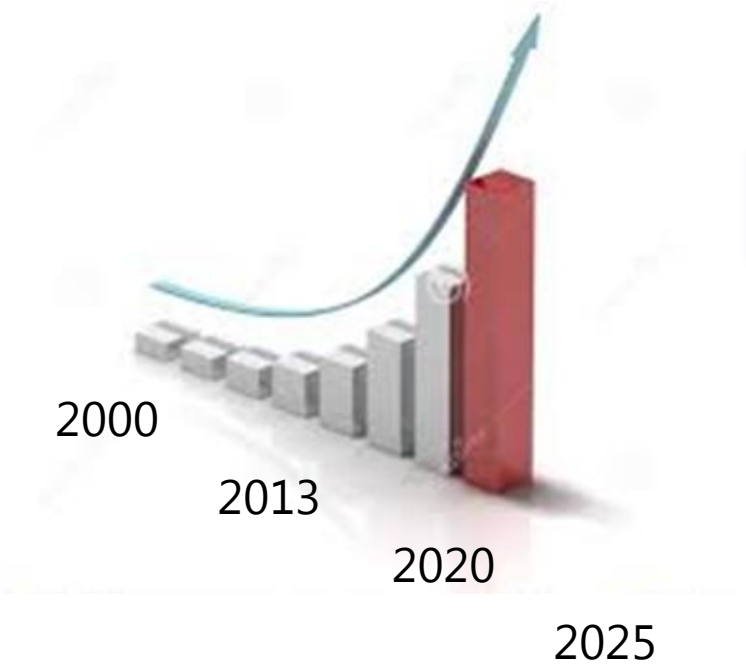


This IDC graph predicts exponential growth of data from around 3 zettabytes in 2013 to approximately 40 zettabytes by 2020. An exabyte equals 1,000,000,000,000,000,000 bytes and 1,000 exabytes equals one zettabyte. Source: IDC's Digital Universe Study, December 2012, http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf.

PASS

The data avalanche is accelerating

Don’t let data quality wipe you out!



Impact of bad data

Increasing
at near the same rate
as data expansion

$3,100,000,000,000

IBM's Estimate of **Annual** Cost of Bad Data to US Economy (IBM BDH)

15%	27%
Surveyed Executives Trusting Overall Data (IDC)	Surveyed Executives Sure of Data Accuracy (IBM)

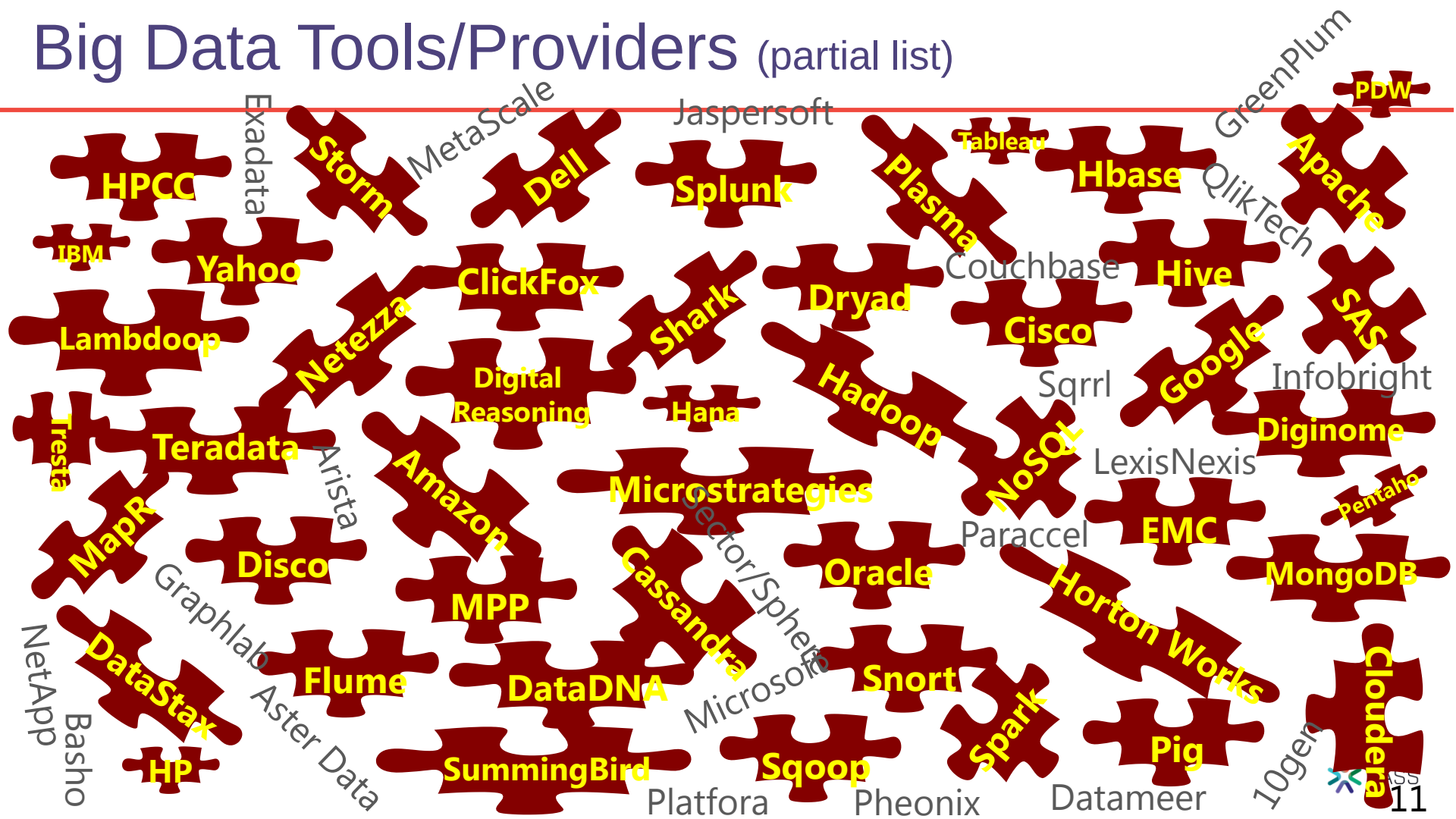
PASS

You will be (or, are already) dealing with..

- ***High-Volumes*** of data you need to access **Volume**
- ***High-Velocity*** of streaming data pouring in **Velocity**
- ***High-Variety*** of information assets (structured, semi-structured, unstructured) **Variety**
- **AND**, you need to surface this data to enable enhanced decision making, insights, discovery and process optimization

Oh, and it better be good data (have **Veracity**) (source: IBM/Diginome) **Veracity**

Big Data Tools/Providers (partial list)

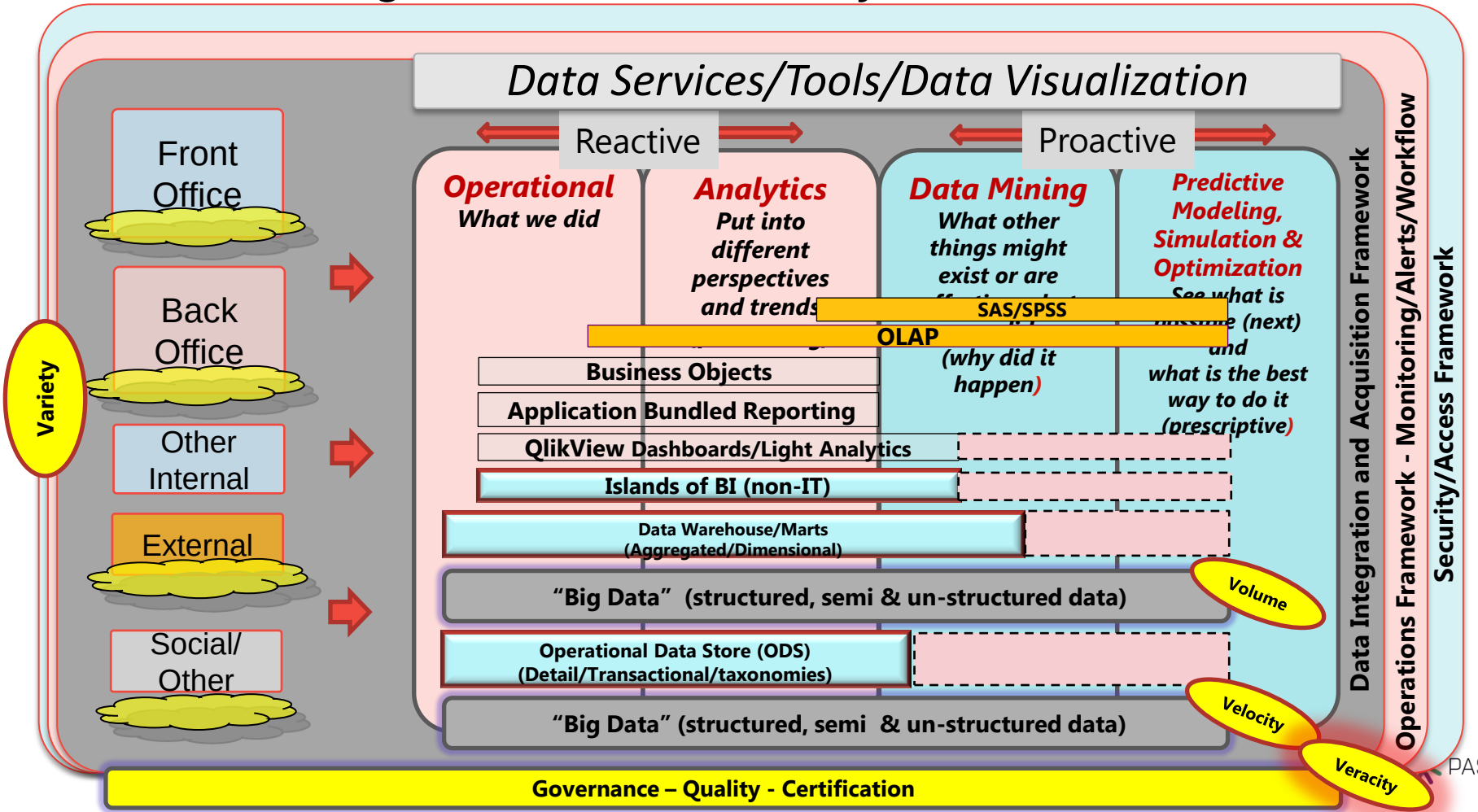


Are you doing the right thing?

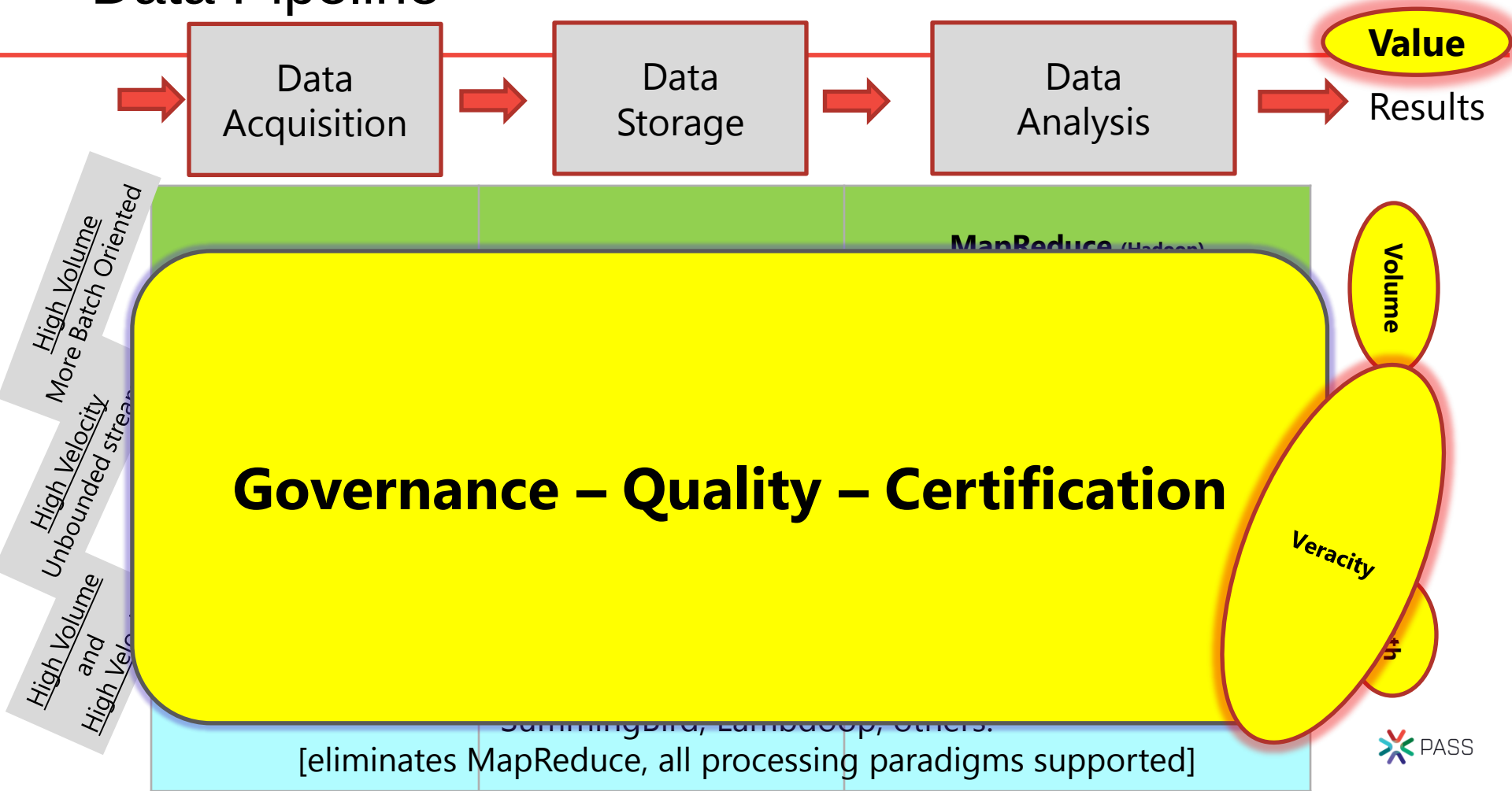
- **Hadoop** (HDFS solutions) lends itself to problems that can be solved through distributed strategies coupled with advanced analytics.
- Ot
wi
PD
- AND, always attack the quality of the data !!!!! **(Veracity)**

Understand the problem FIRST,
NEXT, apply the proper architecture,
and FINALLY, choose the proper tools!

Business Intelligence and Data Analytics



Data Pipeline



PASS

Would you drink this?

NO, but it likely could have been prevented
(or cleaned up during data acquisition or earlier)

PASS

Recent Big Data and Data Quality efforts

Universe of External & Internal Data

100's of sources, dozens of formats, no control of content

High Tech

Computer Components
Sales & Marketing

Health Care

Patient Population Health
Health Insurance/Coverage

Financial Services

Stocks/Portfolio management

PASS

A disturbing pattern has emerged in big data

Universe of External & Internal Data

100's of sources, dozens of formats, no control of content

All new data flows to the big data platform

Big Data Platform

Unidentified Records are just ignored

Nothing is fixing bad data in the data lakes (perhaps on query?)

Zero data governance

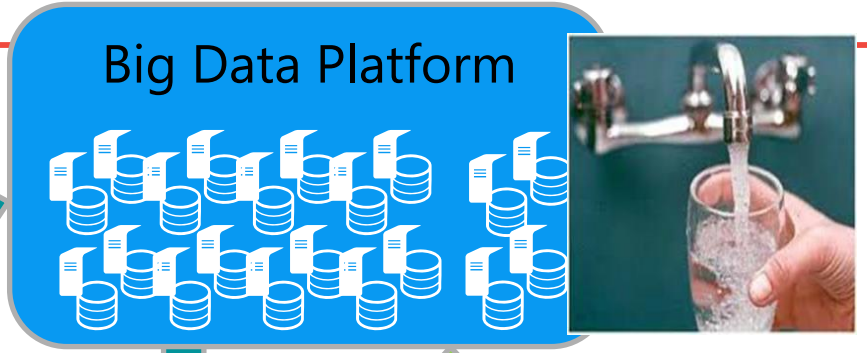
How do you identity what is <u>good</u> data versus <u>bad</u> data?

Note: Not transactional

PASS

So, let's add in data quality as a service!

All new data flows to the big data platform



Data needing cleansing flow to DQaaS

Cleansed/Good data flows back to Big Data Platform

DQaaS Fuzzy Matches, Taxonomies, Workflows, Accumulate Knowledge

General Architecture (Data Pipeline with DQaaS)

External and Internal Transactional Data and Reference Data

Front Office

Back Office

Other Internal

"Enterprise"
Marketing Data
Customer Data
Product Data
Sales Data
Financial Data

EDW
DM's
MDM
Hier's
Taxonomies

External

POS Data
Advertising Data

Social/ Other

Google
FB
Twitter
Other

BIG DATA
Data Staging /Data Acquisition

Raw Data (any new)
- External Data
- Internal Data

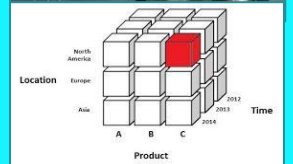
Existing Data
- External Data
- Internal Data

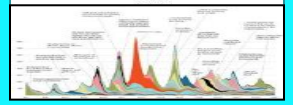
Enrichment Data

Enterprise Data

BIG DATA
Data Delivery
Platform

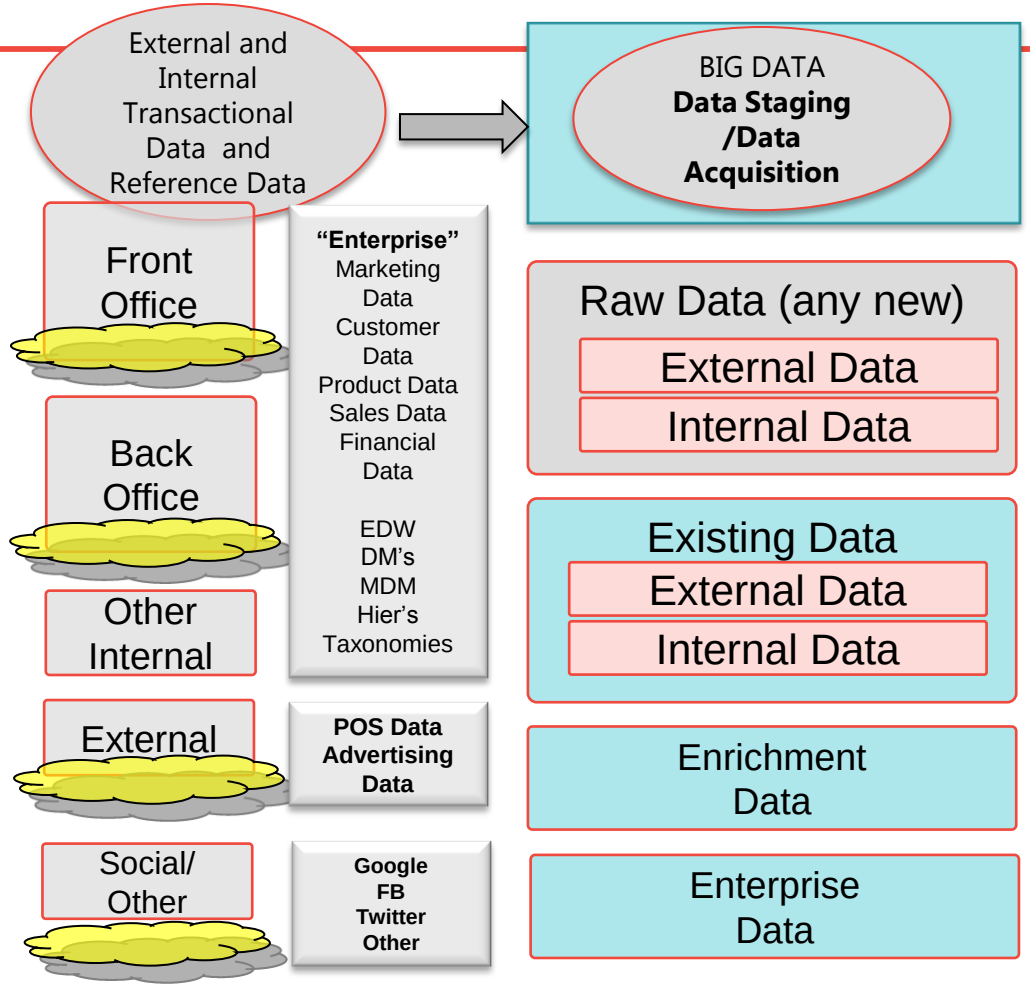
Publishable Data
- External Data
- Internal Data
- Enrichment Data





Data Quality as a Service
- Mastering
- Matching
- Cleansing
- Enriching

General Architecture – Data Staging/Acquisition



Raw = All data starts out as un-mastered, un-reconciled, and un-validated. Then, new data coming in.

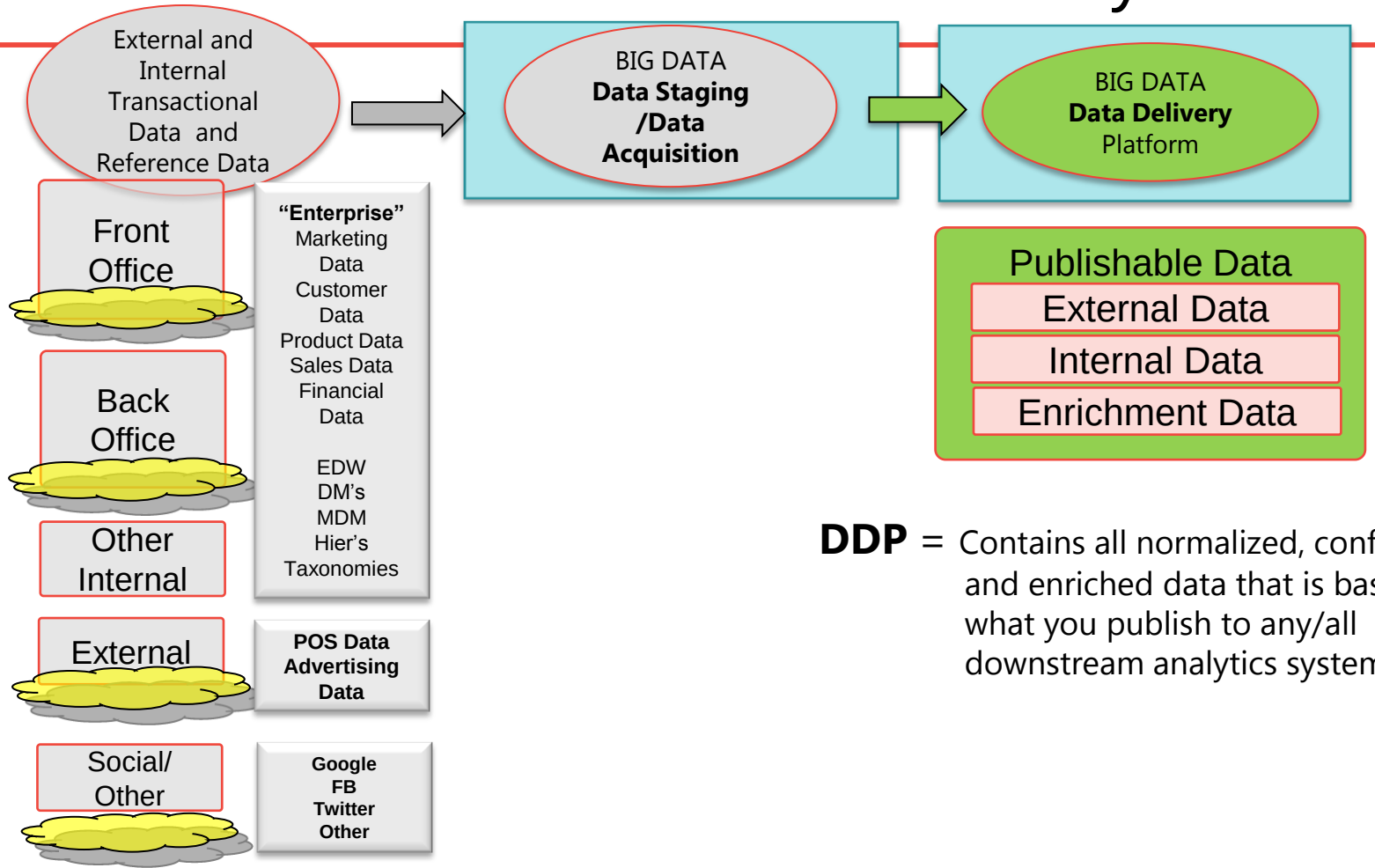
Conformed = matched, deduped, cleansed, hierarchically reconciled, taxonomy adjusted, Standardized and defaulted, selectively merged (survivorship at the attribute level) **(Existing Data)**

Enriched = extend existing data with additional information from other sources

Enterprise = structured/high quality enterprise investment

PASS

General Architecture – Data Delivery Platform



DDP = Contains all normalized, conformed, and enriched data that is basis of what you publish to any/all downstream analytics systems

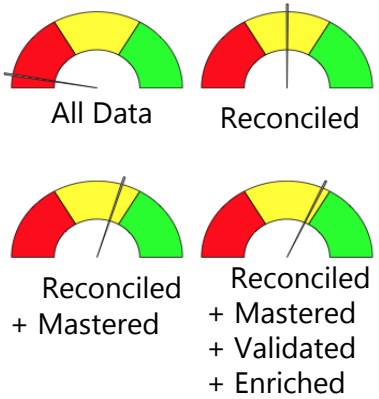
PASS

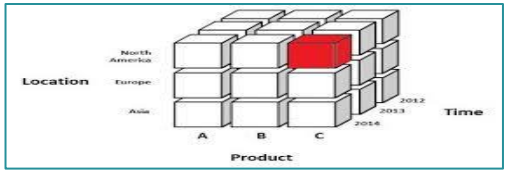
General Architecture (Data Delivery Platform)

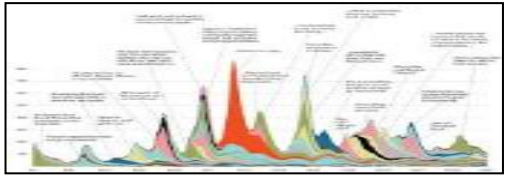
BIG DATA
Data Delivery
Platform

Publishable Data
- External Data
- Internal Data
- Enrichment Data

Turn the dial for what you need





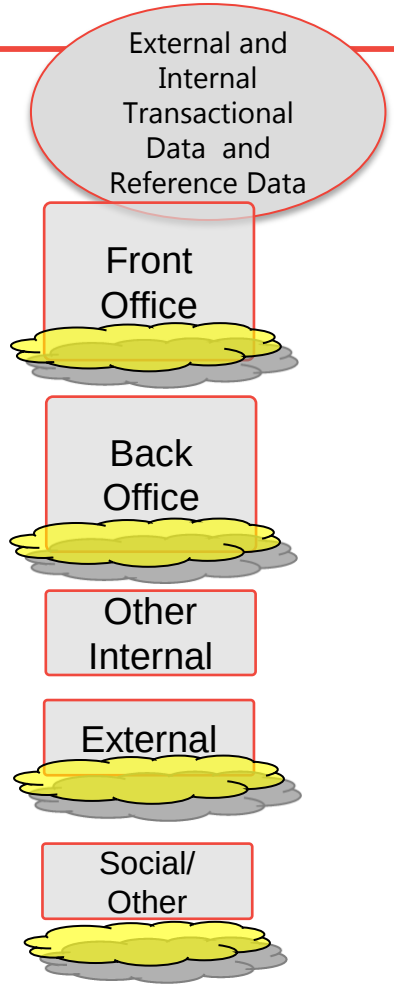


DDP = Contains all normalized, conformed, and enriched data that is basis of what you publish to any/all downstream analytics systems

Original Data + Conformed Data

PASS

Taxonomies (for consistency, accuracy)



SourceProductName (+) InternalProductReference

Product Names correspond to Product reference

"SuperX498", "SuperX490", "SuperX495" → "SuperX4series"

SourceLegalEntity (+) InternalLegalEntity

Legal Entity reference correspond to Legal Entity reference

"Incorporation", "LLC", "PLC", "LTD" → "Corporation"

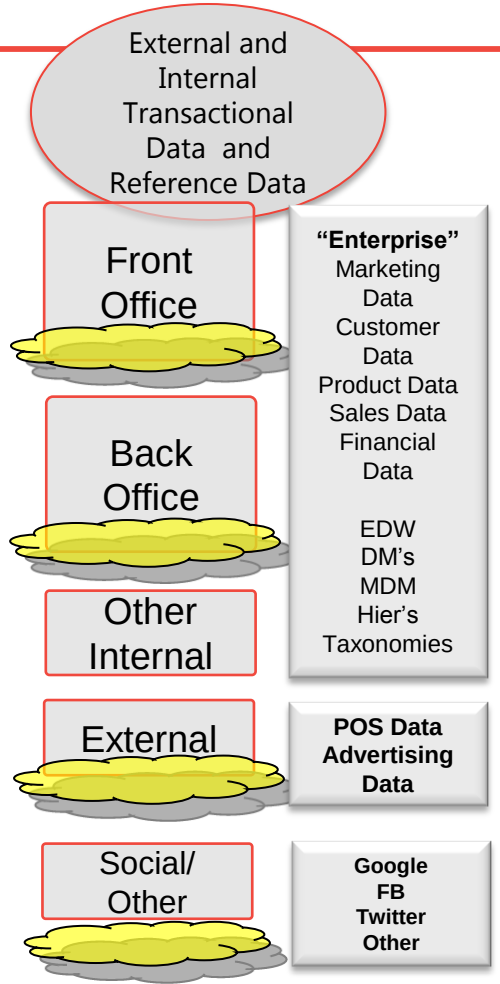
SourceProductFamily (+) InternalProductArchitecture

Product Families corresponds to Processor Architecture

"PC1", "PC2" → "X86"
"PC3", "PC4" → "64"

PASS

Standardized and defaulted



Address parts (for consistency/accuracy/matching)

"Street", "St.", "ST", "STREET" → "St."

"Drive", "Dr.", "Drv.", "D." → "Dr."

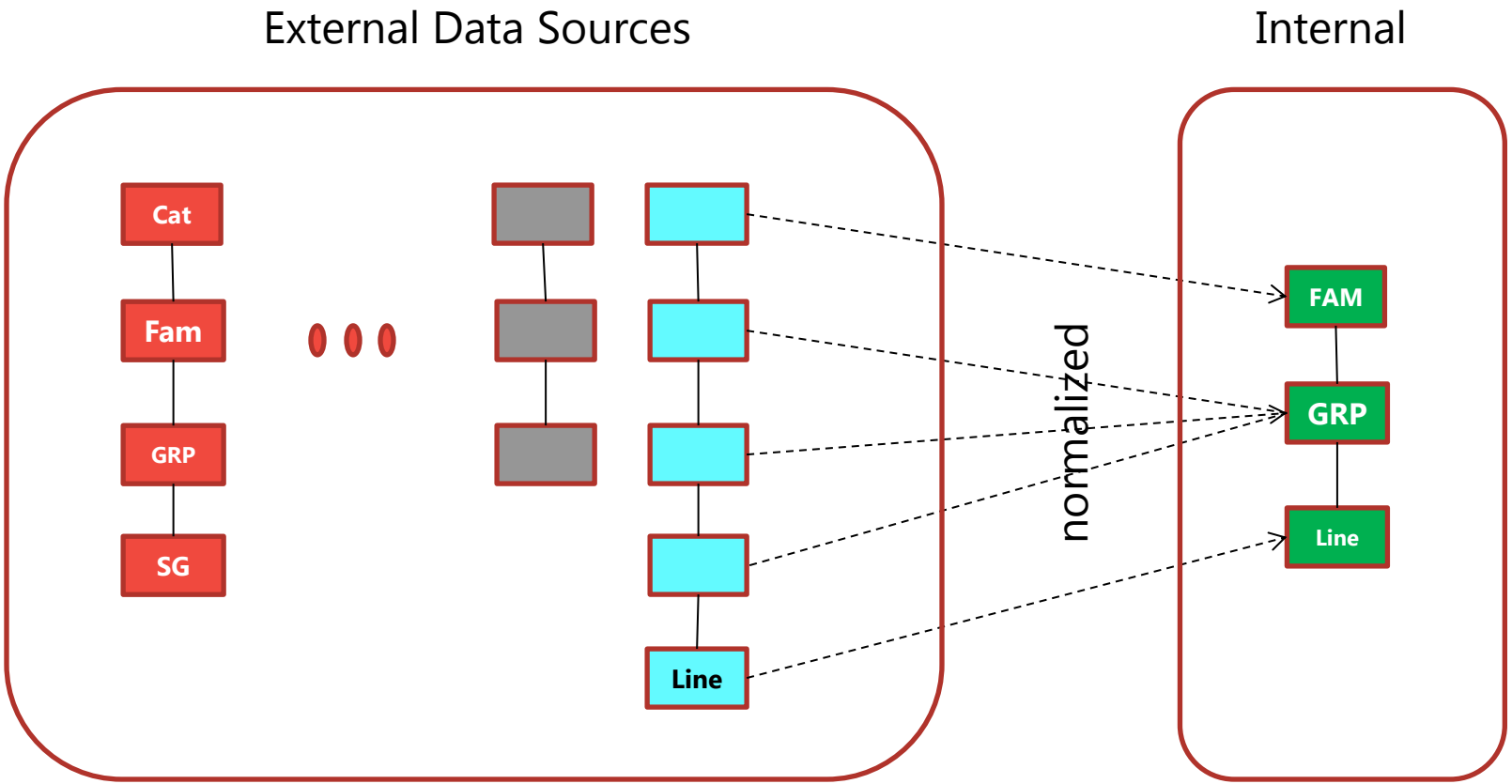
Address parts (for completeness/accuracy/matching)

"1237 NW 23rd Street"
"Portland, OR" → "1237 NW 23rd Street"
"Portland, OR **97035**"

Accuracy also means "completeness"

PASS

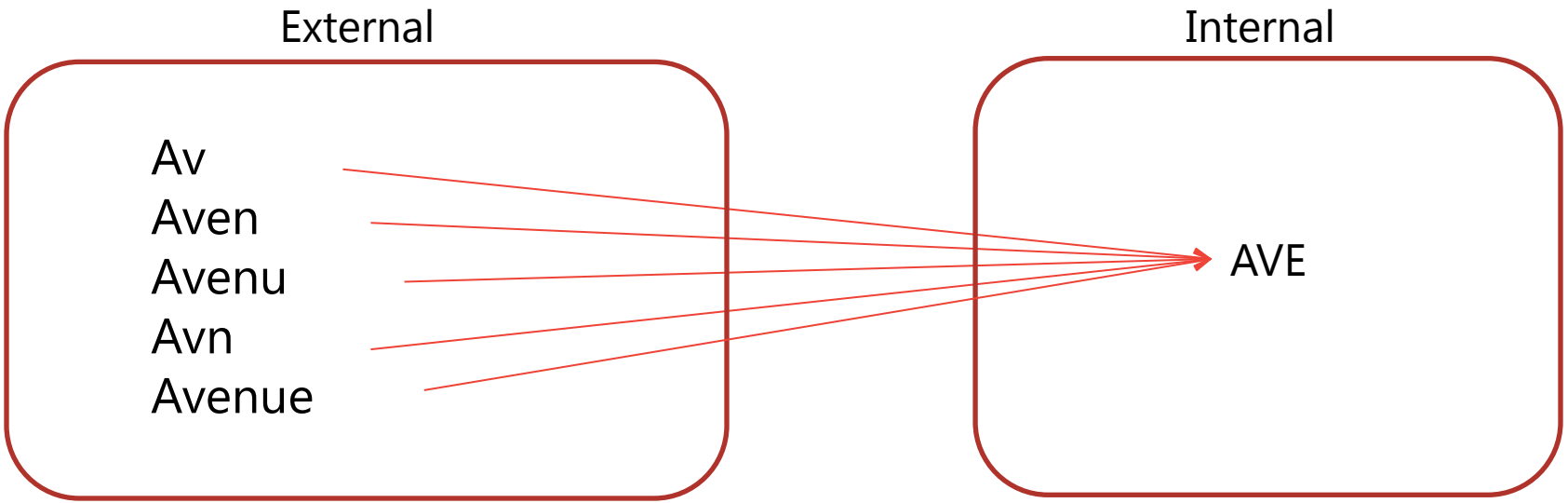
Product Hierarchy reconciliation (for contextual consistency)



PASS

Address Synonyms (AKA's)

Equivalents (for consistency, matching)

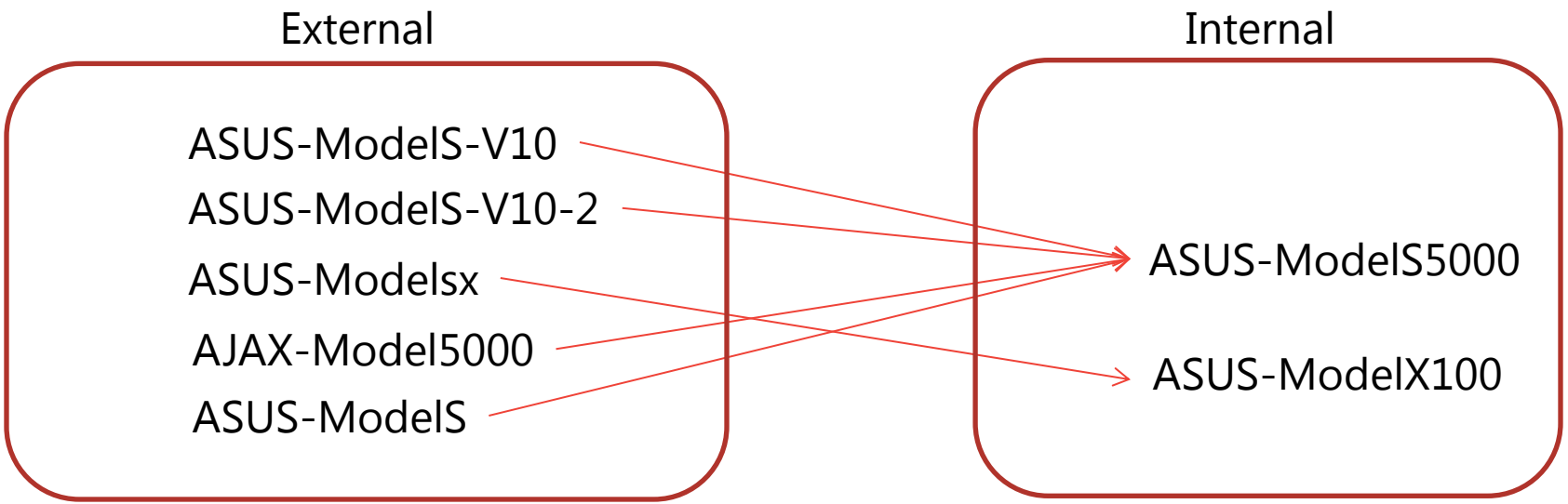


Master		Synonym(AKA's)
AVE	→	Av, Aven, Avenu, Avn, Avenue

PASS

Product Synonyms (AKA's)

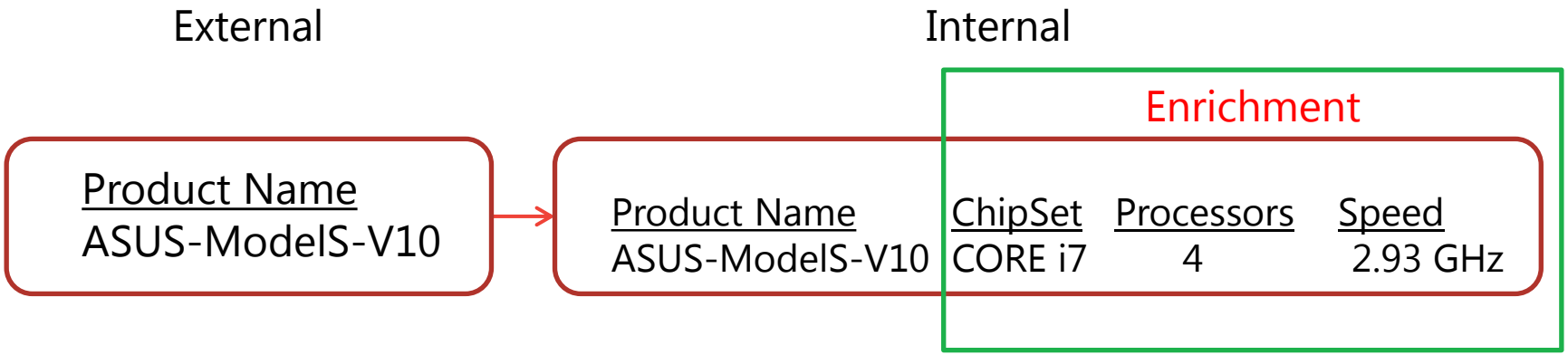
Equivalent (for consistency, matching)



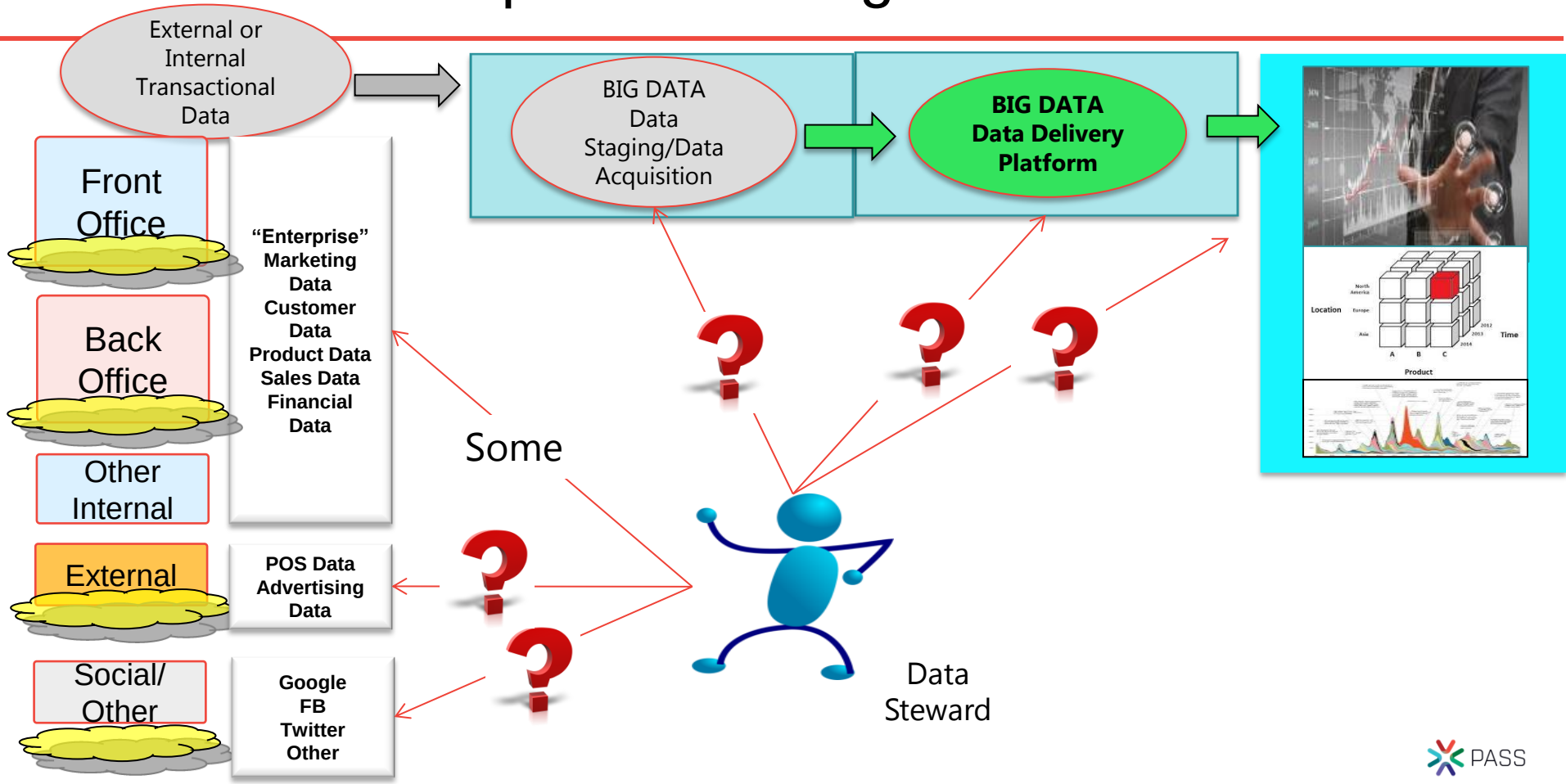
Master		Synonym(AKA's)
ASUS-ModelS5000	→	ASUS-ModelS-V10, AJAX-Model5000, ...

PASS

Product Enrichment

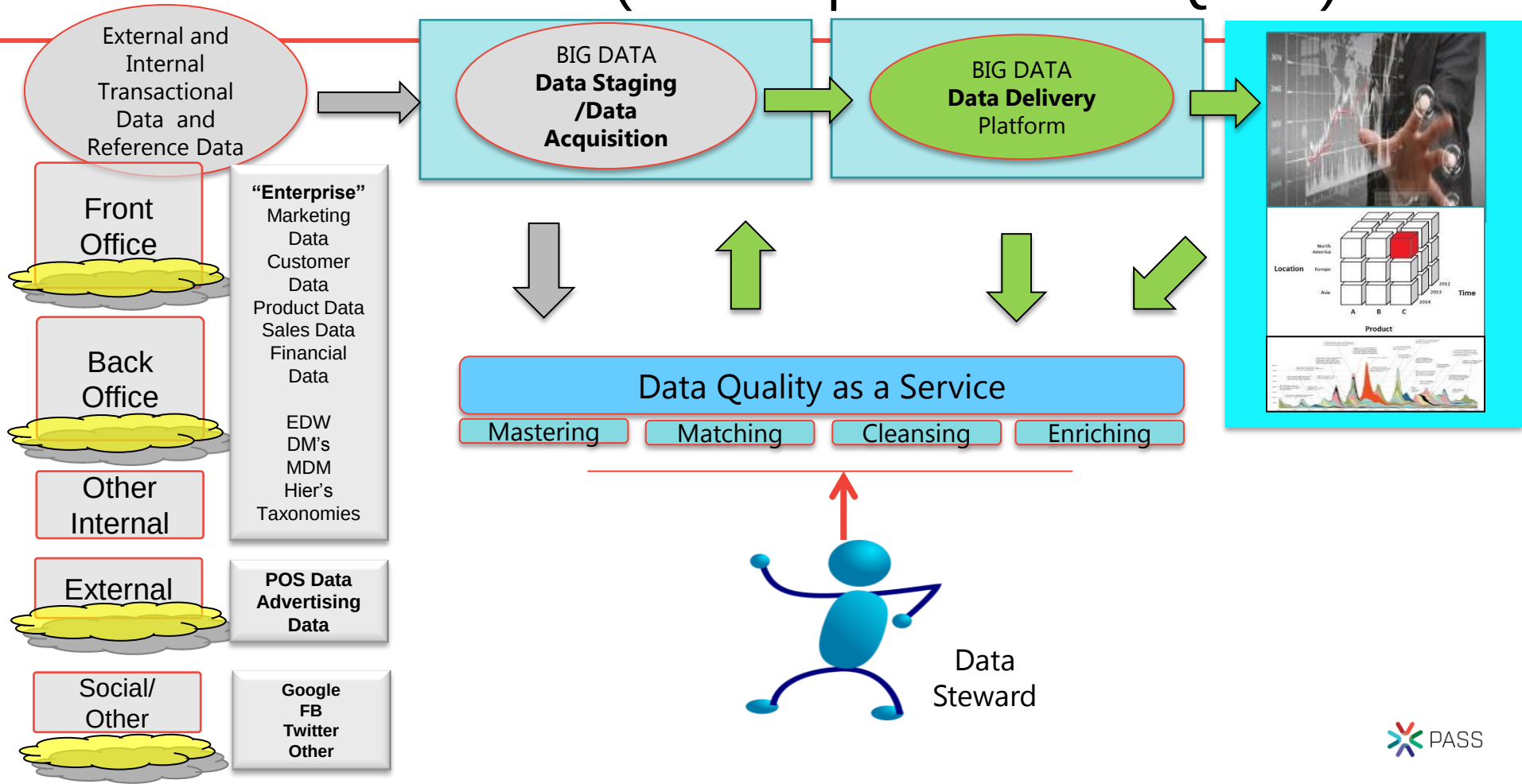


Data Stewardship issues – Big Problem



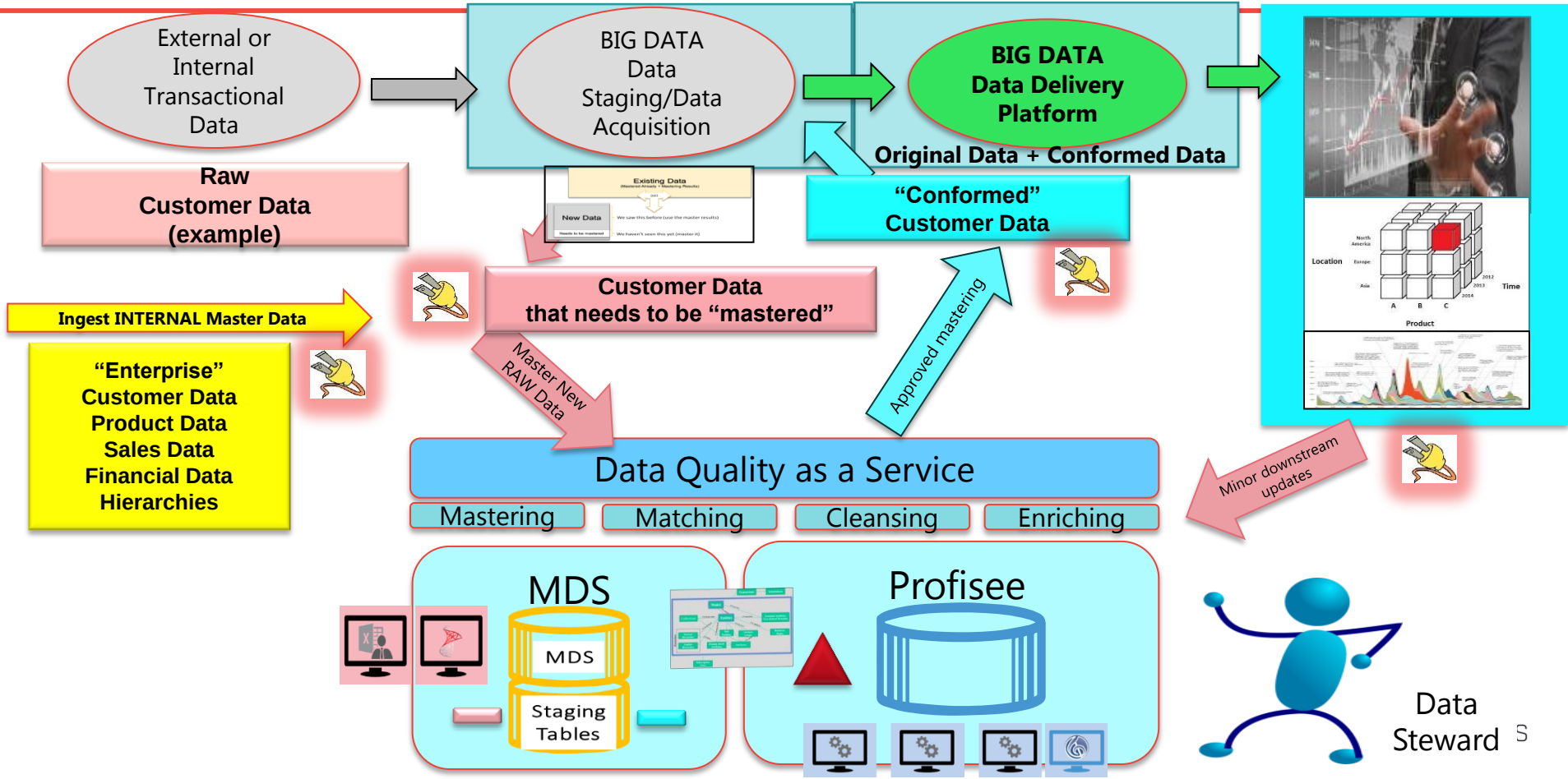
PASS

General Architecture (Data Pipeline with DQaaS)

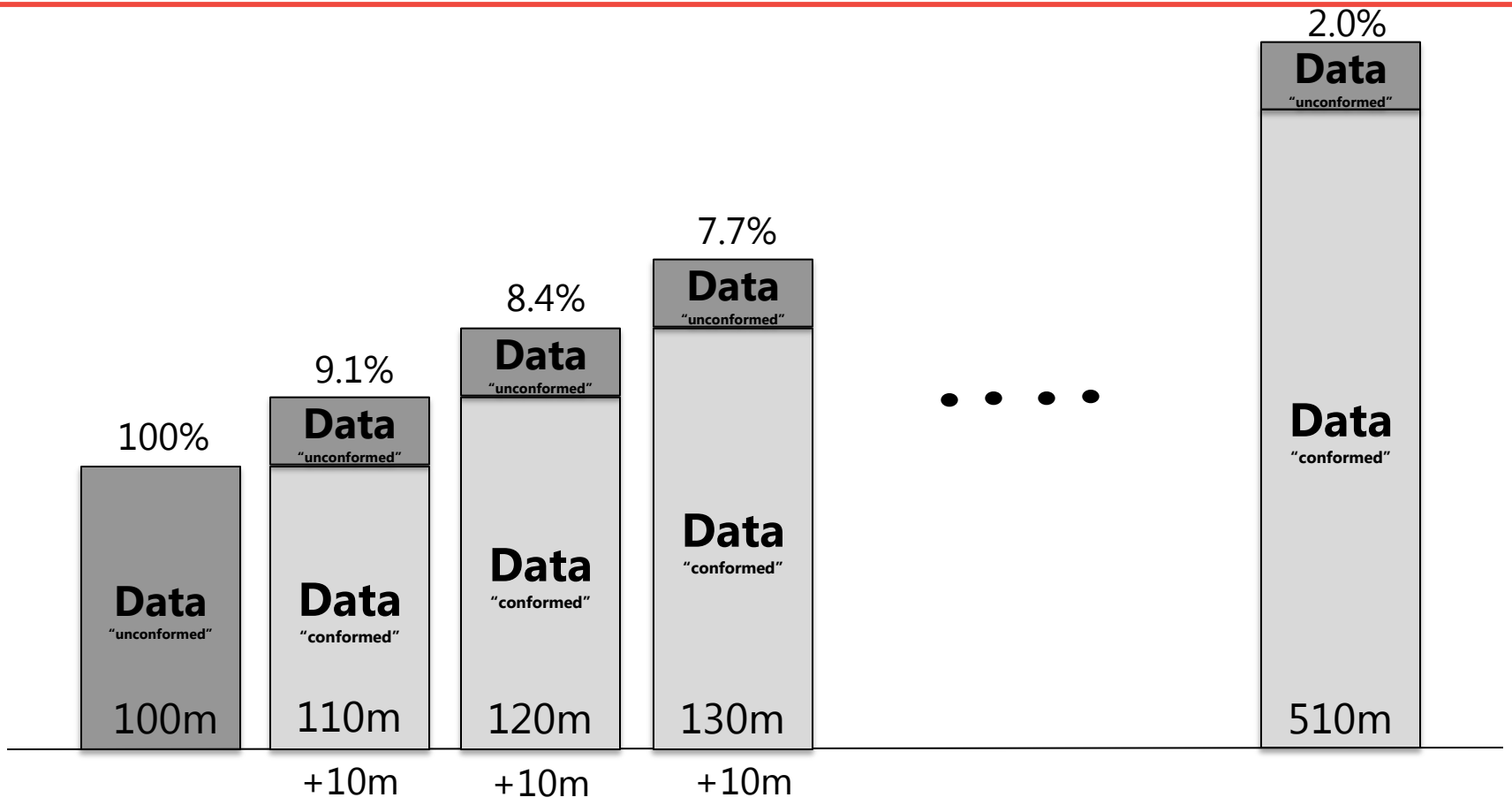


PASS

Plugging in Data Quality as a Service (DQaaS)

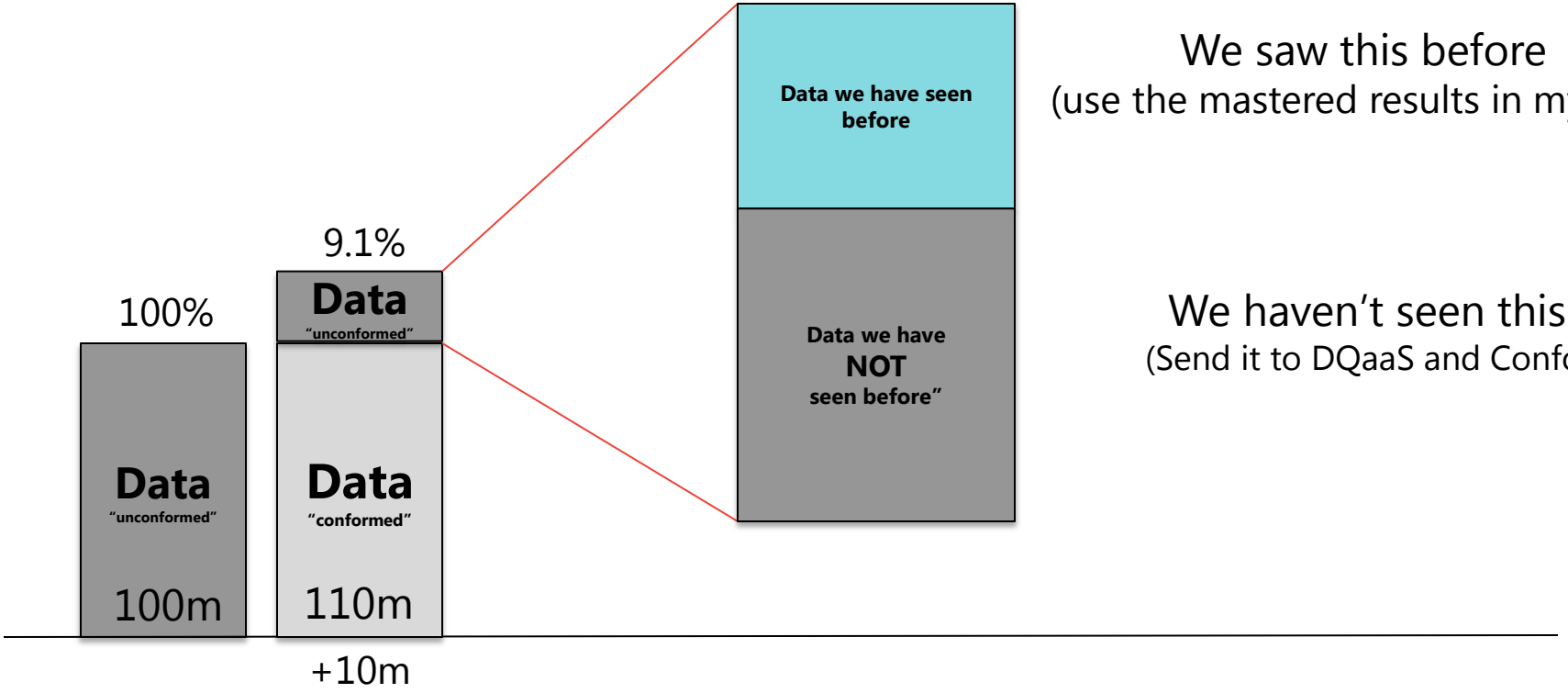


% of data to be "conformed" decreases over time



PASS

% of data to be “cleansed” also decreases

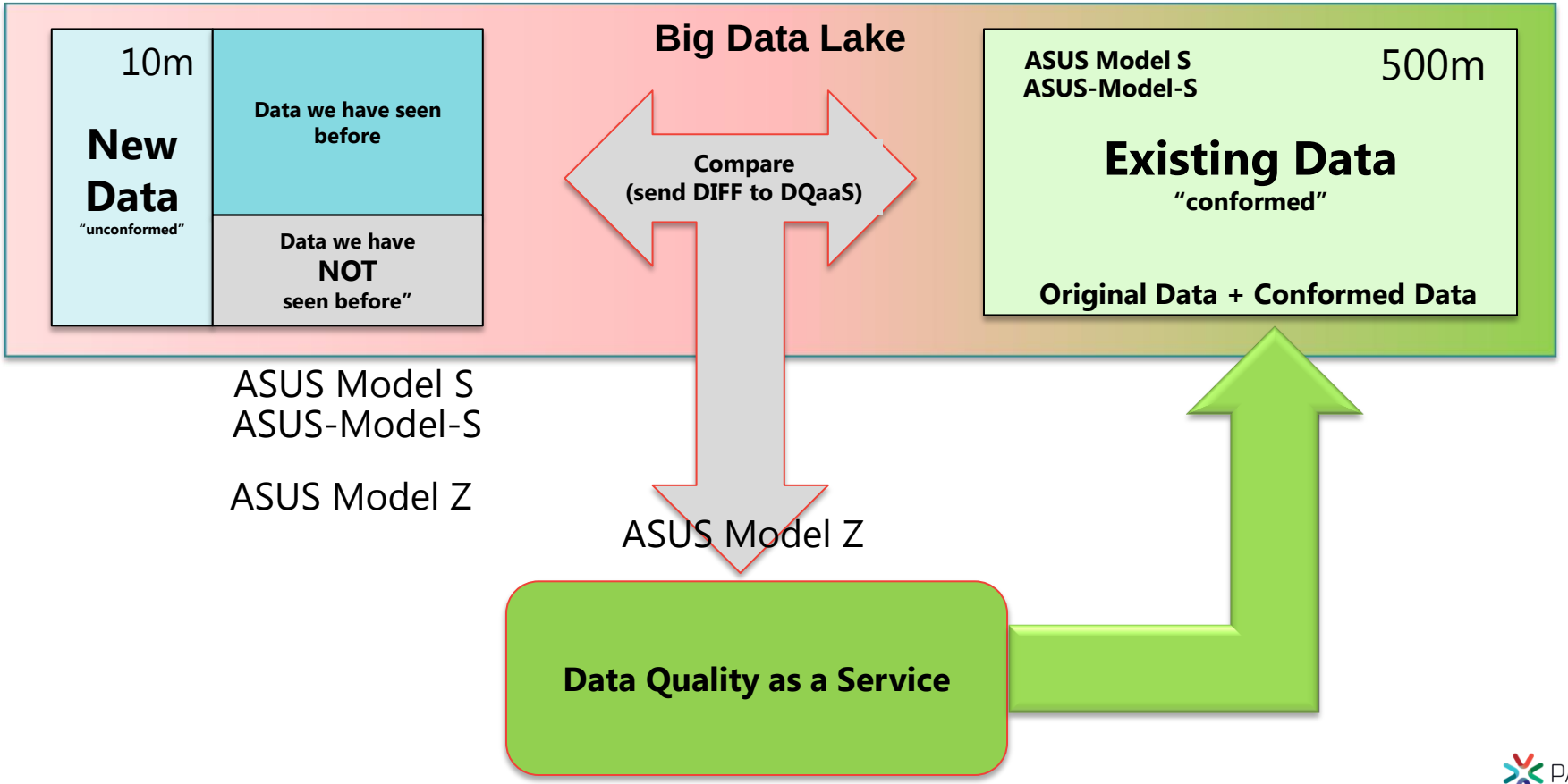


We saw this before
(use the mastered results in my lake)

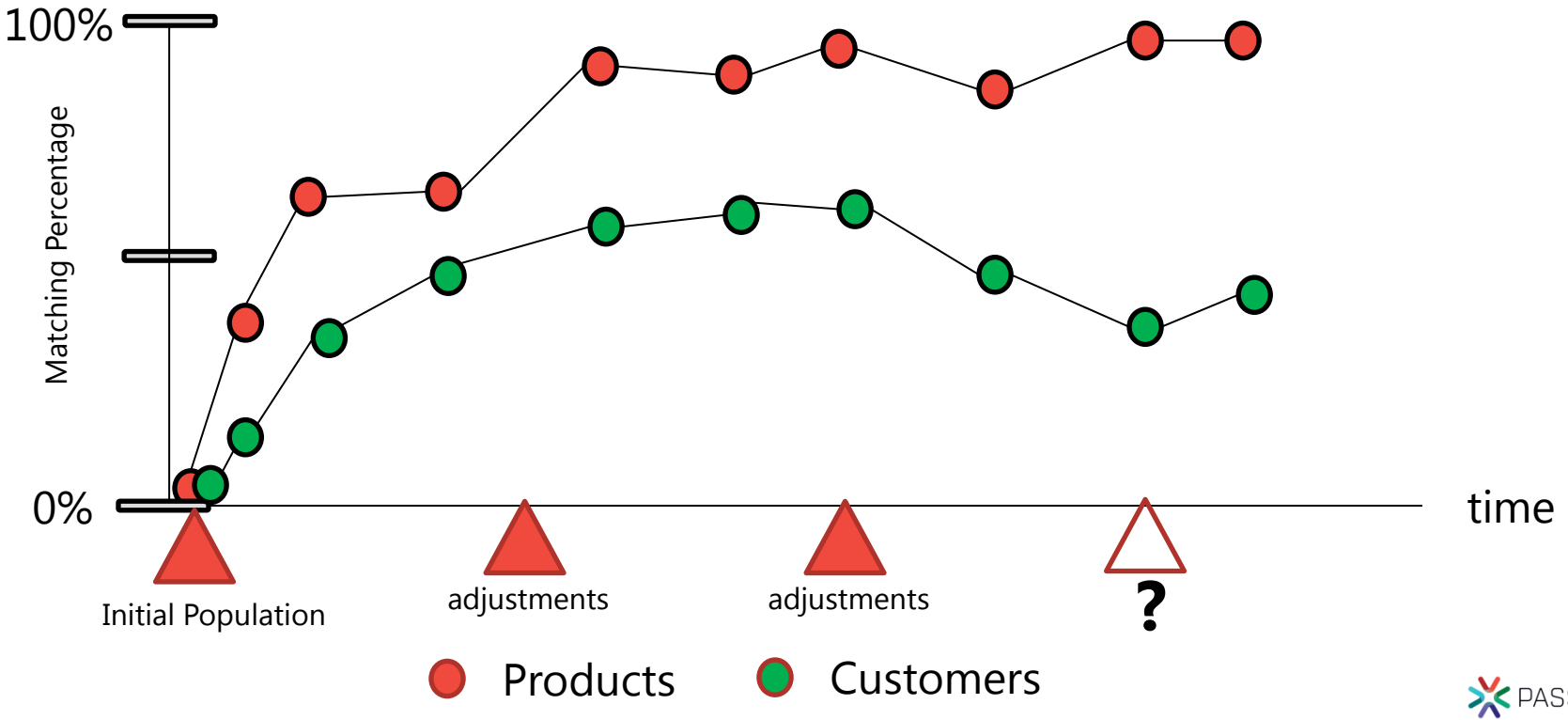
We haven’t seen this yet
(Send it to DQaaS and Conform it!)

PASS

Things we can do easily on big data platform

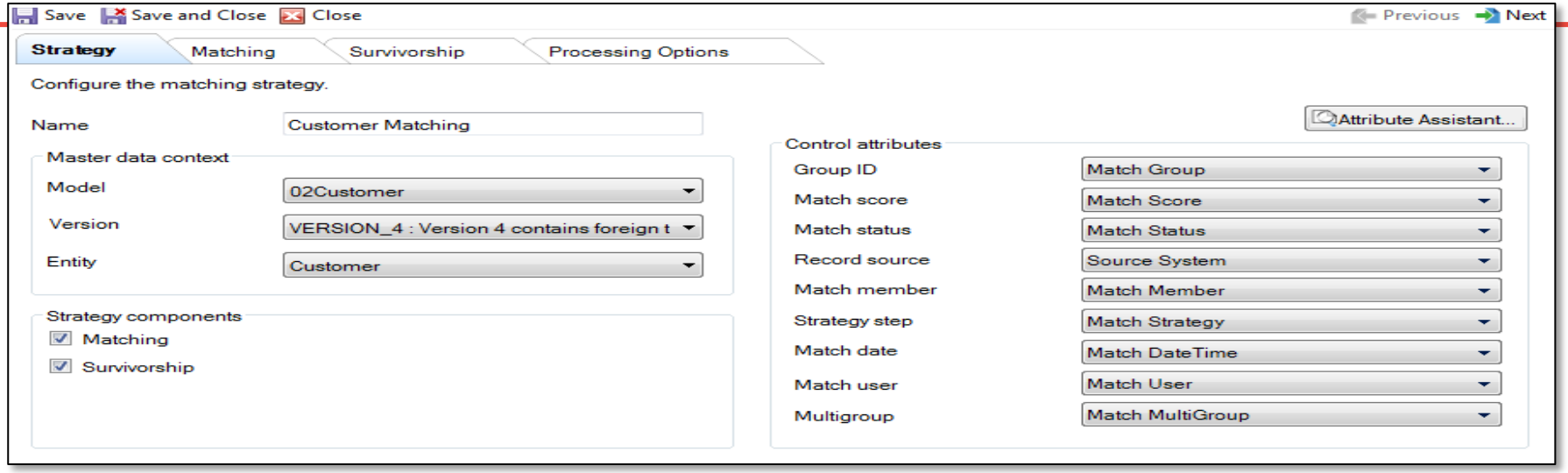


Results (Data we have seen before)



PASS

Key Concept – Strategies



- Defines all settings to execute matching and survivorship
- Configured to match a subset of an entities attributes
- Determines how master records are created and populated

PASS

Key Concept – Match Scores and Thresholds

PASS

Product Matches (grouped)

ASUS-ModelS (Match Group 1)

	ASUS-ModelS	(1.000)	(Match Group 1)
	ASUS-Model S	(.952)	(Match Group 1)
Auto Threshold (.900)	ASUS ModelS	(.950)	(Match Group 1)
Proposed Threshold (.750)	Asus_Model_S	(.885)	(Match Group 1)
	ModelS	(.324)	(Match Group 1)

Master		Grouped
ASUS-ModelS	→	ASUS-Model S, ASUS ModelS, Asus_Model_S

PASS

Key Concept – Match Groups

- Match groups are the basic unit of matching
- Contain similar members
- Mapped together by the same Match Group ID

[Match groups tie records together]

Selected Group: 000142

Approve

Match Group ID	Name	Code	Match Status	Match Score	Record Source
000142	Bike World Incorporated	1988	30 [Proposed]	0.7000	SAP []
000142	Bike World	204	20 [AutoApproved]	1.0000	SAP []
000142	Bike World	2333	20 [AutoApproved]	1.0000	Salesforce []
000142	Bike World Inc.	2335	30 [Proposed]	0.7000	Dynamics []
000142	Bike World	2336	20 [AutoApproved]	1.0000	SAP []
000142	Bike World Inc.	35	30 [Proposed]	0.7000	Salesforce []
000142	Bike World Inc.	96	30 [Proposed]	0.7000	JDE []

Key Concept – Master (Golden) Record

- The mastering portion of survivorship creates master members which hold survived representative data
  - Creates master member to represent the group
  - Preserves the data for each source record
  - Can be used to create the golden record – the best possible representation of the distinct customer, product, etc.

Selected Group: 000142

Approve Reject Previous Next

Proposed Count	Approved Count	Match Group ID	Name	Code	Match Status	Match Score	Address Line 1	City	StateProvince
4	3	000142	Bike World	Master-000142	10 [Master]		1249 Quintilio Dr.	Bear	DE
		000142	Bike World Incorporated	1988	30 [Proposed]	0.7000	1249 Quintilio Dr.	Bear	DE
		000142	Bike World	204	20 [AutoApproved]	1.0000	1249 Quintilio Dr.	Bear	DE
		000142	Bike World	2333	20 [AutoApproved]	1.0000	2100 Ashford Dunwoody Rd	Atlanta	GA
		000142	Bike World Inc.	2335	30 [Proposed]	0.7000	2100 Ashford Dunwoody Road	Atlanta	GA
		000142	Bike World	2336	20 [AutoApproved]	1.0000	2100 Ashford Dunwoody	Atlanta	GA
		000142	Bike World Inc.	35	30 [Proposed]	0.7000	1249 Quintilio Dr.	Bear	DE
		000142	Bike World Inc.	96	30 [Proposed]	0.7000	1249 Quintilio Drive	Bear	DE

Customer Master – Golden Record

Master Customer					
Name	**Code**	**Source**	**Add1..**	**Customer #**	**Master**
XYZ Corporation	Master-6001		329 Main St South	C5321	Master-6001
XYZ c	6001	EXT2	329 Main Street S		Master-6001
XYZ Corporation	6005	EXT9	3229 Main St So		Master-6[illegible]01
Xyz Corp	6009	CRM	329 Main Street So	C5321	Master [illegible]1
[illegible] Design	Master-6003		2520 Northwinds Place	[illegible]5400	Master-60[illegible]
[illegible]XD	[illegible]	[illegible]RM	2520 Northwin[illegible]	C5400	Mas[illegible]3

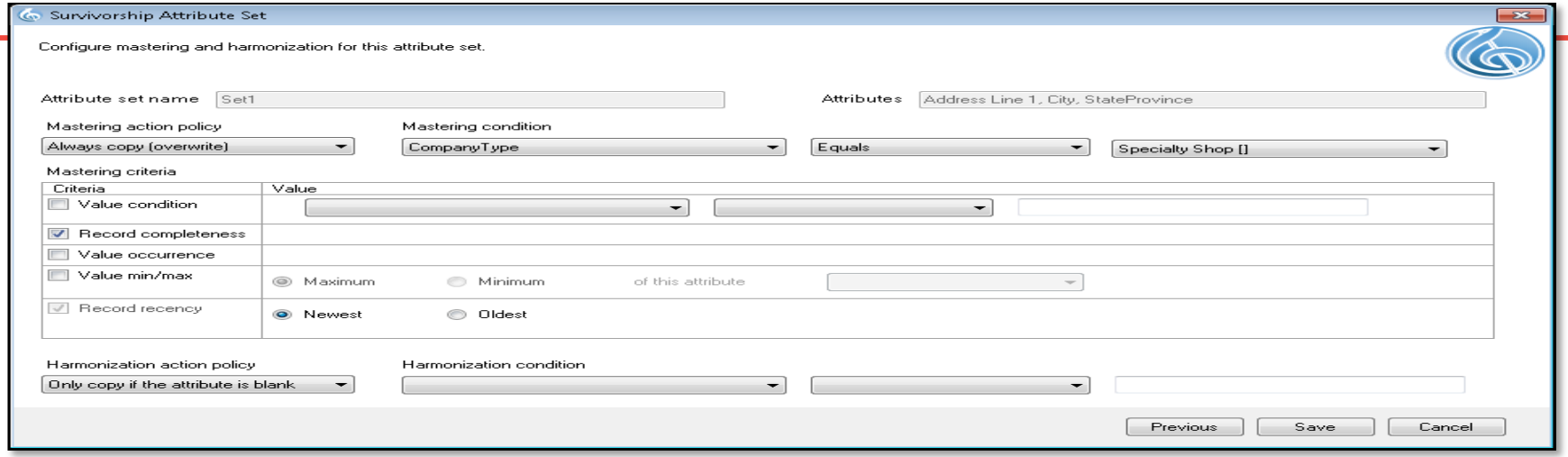
Golden Records Bind other Candidate Records

New Candidate Records are Address Corrected for Sure Matching

Golden Records may have attributes from the candidate records or new attributes altogether

Candidate Records are added to their "Master or Golden" Record Group

Key Concept – Survivorship



- Survivorship includes the concepts of mastering and harmonization
- Survivorship defines how master records get populated, and how master record values are written back to source records
- Profisee supports a logical merging of matched records, leaving source records intact and mapped to a master or golden record
- Survivorship provides automated configuration and execution of both Mastering and Harmonization

PASS

Key Concept – Mastering

- **Mastering**: Creating the best possible representative master record from the source records contained in the match group
- In the example below, the highlighted (blue) source values were used to populate the master record.

Match Group Members : M-172 | Approve | Reject | Previous | Next

Appr...	Propo...	Match Status	Match Score	Name	Code	Record Source	Address Line 1	Phone Number
3	1	10 [Master]		Exemplary Cycles	M-172	MDM [MDM]	1155 Mount Vernon Highway	770-392-1944
		30 [Proposed]	0.8300330	Exemplary Cycle	1788	SAP []	1155 Mount Vernon Hwy	770-392-1944
		20 [AutoApproved]	1.0000000	Exemplary Cycles	1789	Dynamics []	1155 Mount Vernon Highway	770-392-1900
		20 [AutoApproved]	1.0000000	Exemplary Cycles	4288	Salesforce []	1155 Mount Vernon Highway	770-392-1944
		20 [AutoApproved]	1.0000000	Exemplary Cycles	428	JDE []	1155 Mount Vernon Highway	770-392-1944

Key Concept – Matching Algorithm

Algorithm for "fuzzy" matching:

$$\sqrt{\frac{\#CommonTokens^2}{\#ATokens * \#BTokens}}$$

Key Concept – Matching Algorithm

- Compare two strings: “Factor” and “Factorial”
- Specifying the token size to use = **3** for this example

  Factor → [F] [*Fa] [Fac] [act] [cto] [tor] [or*] [r**]**

  Factorial → [F] [*Fa] [Fac] [act] [cto] [tor] [ori] [ria] [ial] [al*] [l**]**

- Number of tokens in Victor (#A Tokens) = 8
  Number of tokens in Vectors (#B Tokens) = 11

- Number of common tokens = **6**

  Factor → [F] [*Fa] [Fac] [act] [cto] [tor] [or*] [r**]**

  Factorial → [F] [*Fa] [Fac] [act] [cto] [tor] [ori] [ria] [ial] [al*] [l**]**

- Similarity of these two strings is **.639**

$$\sqrt{\frac{6^2}{8*11}}$$

PASS

2 token calculation

Factor → **[*F] [Fa] [ac] [ct] [to] [or] [r*]**

Factorial → **[*F] [Fa] [ac] [ct] [to] [or] [ri] [ia] [al] [l*]**

$$\sqrt{\frac{6^2}{7*10}} = .717$$

PASS

Key Concept – Match Scores and Thresholds

Match Group Members : M-428

Master	Match Status	Match Sc...	Name	Code	Record Source
MSTR {}	10 {Master}		Exemplary Cycles	M-428	MDM {MDM}
M-428 {Exemplary Cycles}	20 {AutoApproved}	0.952498	Exemplary Cycle	1788	SAP {}
M-428 {Exemplary Cycles}	30 {Proposed}	0.838799	Exemplary Cycles	1789	Dynamics {}
M-428 {Exemplary Cycles}	20 {AutoApproved}	1	Exemplary Cycles	428	JDE {}
M-428 {Exemplary Cycles}	30 {Proposed}	0.838799	Exemplary Cycles	4288	Salesforce {}

- Score shows the relative sameness of records in a match group (0.010 – 1.000)
- Used to determine whether to auto-approve or propose matches
- Thresholds determine when members will be matched and auto-approved

Note: two other attributes are included in the matching strategy

PASS

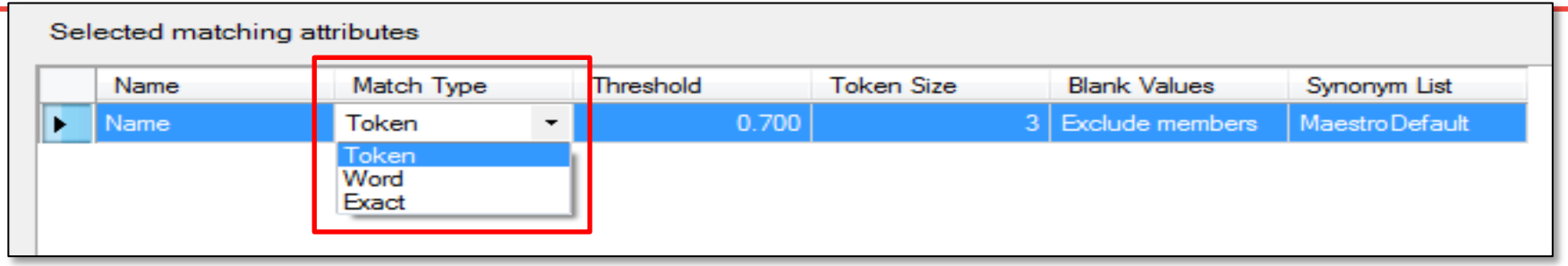
Thresholds – Best Practices

Match Group Members : M-428

Master	Match Status	Match Sc...	Name	Code	Record Source
MSTR {}	10 {Master}		Exemplary Cycles	M-428	MDM {MDM}
M-428 {Exemplary Cycles}	20 {AutoApproved}	0.952498	Exemplary Cycle	1788	SAP {}
M-428 {Exemplary Cycles}	30 {Proposed}	0.838799	Exemplary Cycles	1789	Dynamics {}
M-428 {Exemplary Cycles}	20 {AutoApproved}	1	Exemplary Cycles	428	JDE {}
M-428 {Exemplary Cycles}	30 {Proposed}	0.838799	Exemplary Cycles	4288	Salesforce {}

- Start with:
  - Auto-approval threshold high (.9 or more)
  - Match threshold relatively high (.7 - .8)
- Look for…
  - Over-matches (Proposed members that **should not have** been matched but were)
  - Under-matches (Unique members that **should have** been matched but weren't) and…
  - If too many of either, adjust threshold down and adjust synonyms
- It's better to settle on thresholds that lean towards overmatch
  - The result is more proposed matches which are easy to find and correct in the review process

Key Concept – Match Type Exact or Word



- Performs faster – Breaks attributes into whole words instead of tokens, resulting in fewer comparisons
- Less accurate – With fewer comparisons, result are less matches
- Best Practices
  - Exact is best used for exact match attributes like Social Security Number or Tax-ID

Key Concept – Token Size

Selected matching attributes

	Name	Match Type	Threshold	Token Size	Blank Values	Synonym List
▶	Name	Token	0.700	3	Exclude members	Maestro Default

- Token Size allows for configurability of the “trigram” size on a per Attribute basis
- Smaller Token Size is more accurate, but requires additional processing; larger Token Size is less accurate, but faster
- Token Size Best Practices
  - Raise the token size on large data sets where the additional performance necessitates the reduced accuracy

Key Concept – Synonyms

- Synonyms replace parts of attributes used in matching to standardize common abbreviations and common values
- In addresses, you may have the following abbreviations for Avenue:
- Synonyms can be setup to replace all of these with AVE
- Synonyms significantly improve matching accuracy
- Does not replace or change attribute values in MDS

Name: Address Synonyms

Find (Synonym)	Replace With (Term)	Search In String	Priority	Syno
AV	AVE			
AVEN	AVE			
AVENU	AVE			
AVENUE	AVE			
AVN	AVE			
AVNUE	AVE			
BAYOO	BYU			

Synonyms – Best Practices

- Size of the resulting string matters because the matching algorithm works on similarity
  - Replace longer strings with shorter strings when less significant (e.g. change “street” to “st” in addresses)
  - Replace shorter strings with longer strings when significant (e.g. “Bob” to “Robert” in names)
- Remove words altogether that are considered optional because they are just noise to the matching process.
  - For company names, remove all terms (replace with blank) like “company”, “co”, “inc”, etc. because someone could just as likely enter “Microsoft” as they would “Microsoft Corporation”.
- Use different Synonym lists for different attributes
- Large synonym lists or overuse will slow down processing times

Key Concept – Match Status

- Populated based on the results of the matching engine, and the outcome of the review/approval process
- Based on the match score, members will be:
  - **No Status**: The member has not been processed by matching
  - **Master**: The member is a master record created by Maestro
  - **Auto-Approved**: The member has a master and was auto-approved
  - **Proposed**: The member is in a match group awaiting approval/rejection
  - **Approved**: The member was proposed, and subsequently user approved
  - **User Mapped**: A user mapped this member manually
  - **Unique**: No matches found for the member

All

Name	Code
NoStatus	0
Master	10
AutoApproved	20
Proposed	30
Approved	40
UserMapped	50
Unique	60

Key Concept – Review, Approval and Rejection

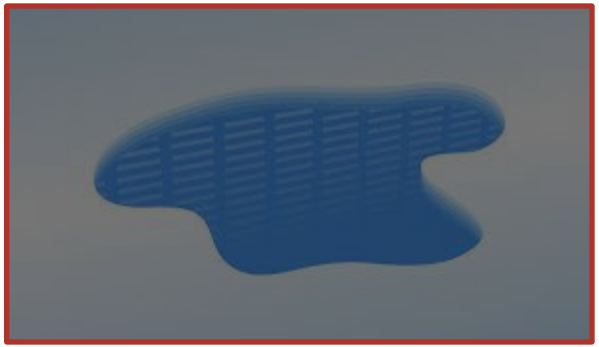
- Review results using the Profesee Desktop Matching Results view
- Users can focus on different subsets of members, including:
  - Unique
  - Proposed
  - Approved
- Approve , reject and create new matching groups

PASS

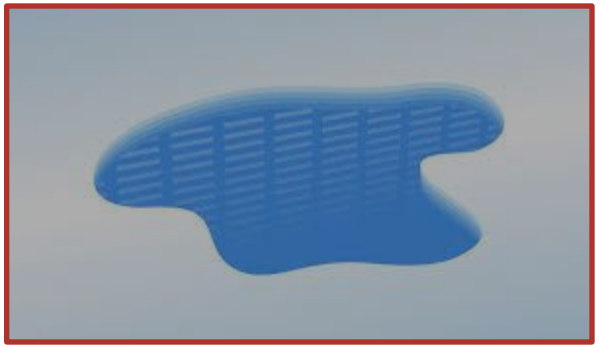
Key Concept – Unique and Proposed

- Review proposed records and approve or reject
  - **Goal**: Have no members in Proposed status
- Master Uniques – In certain scenarios, mastering unique records by creating and populating a master record can be helpful
  - Subscribing systems need consume only master records
  - New records would be matched only to master records

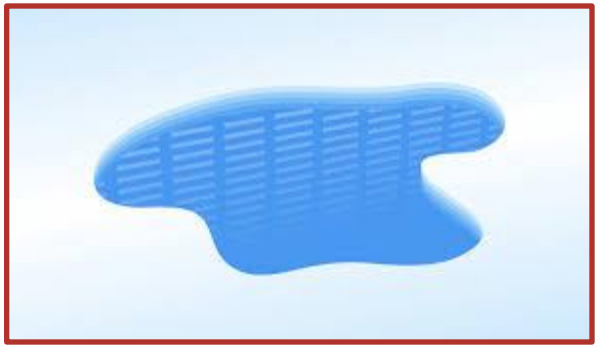
Phases of a dirty data lake (results)



No DQaaS
63.5% good data

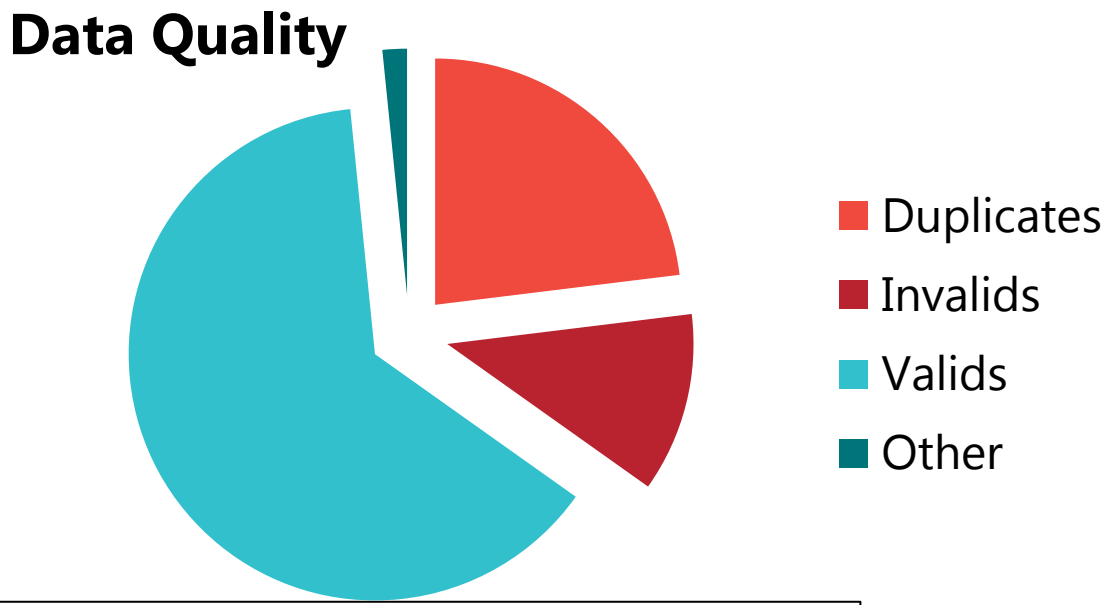


With DQaaS
(first month)
78.1% good data



With DQaaS
89.6% good data

Overall population DQ measurement

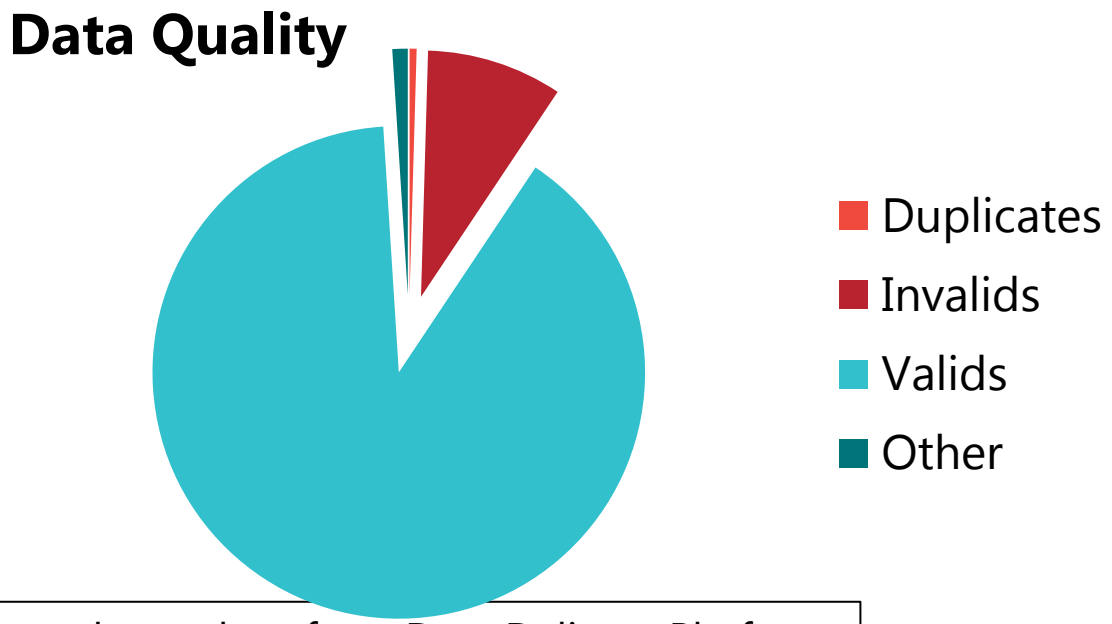


Measure 100,000 random values from Data Delivery Platform

- Number of duplicates 23,056
- Number of Invalids (garbage) 11,771
- Number of Valid's (not garbage) 63,571
- Other (undecided) 1,602

PASS

Overall population DQ measurement

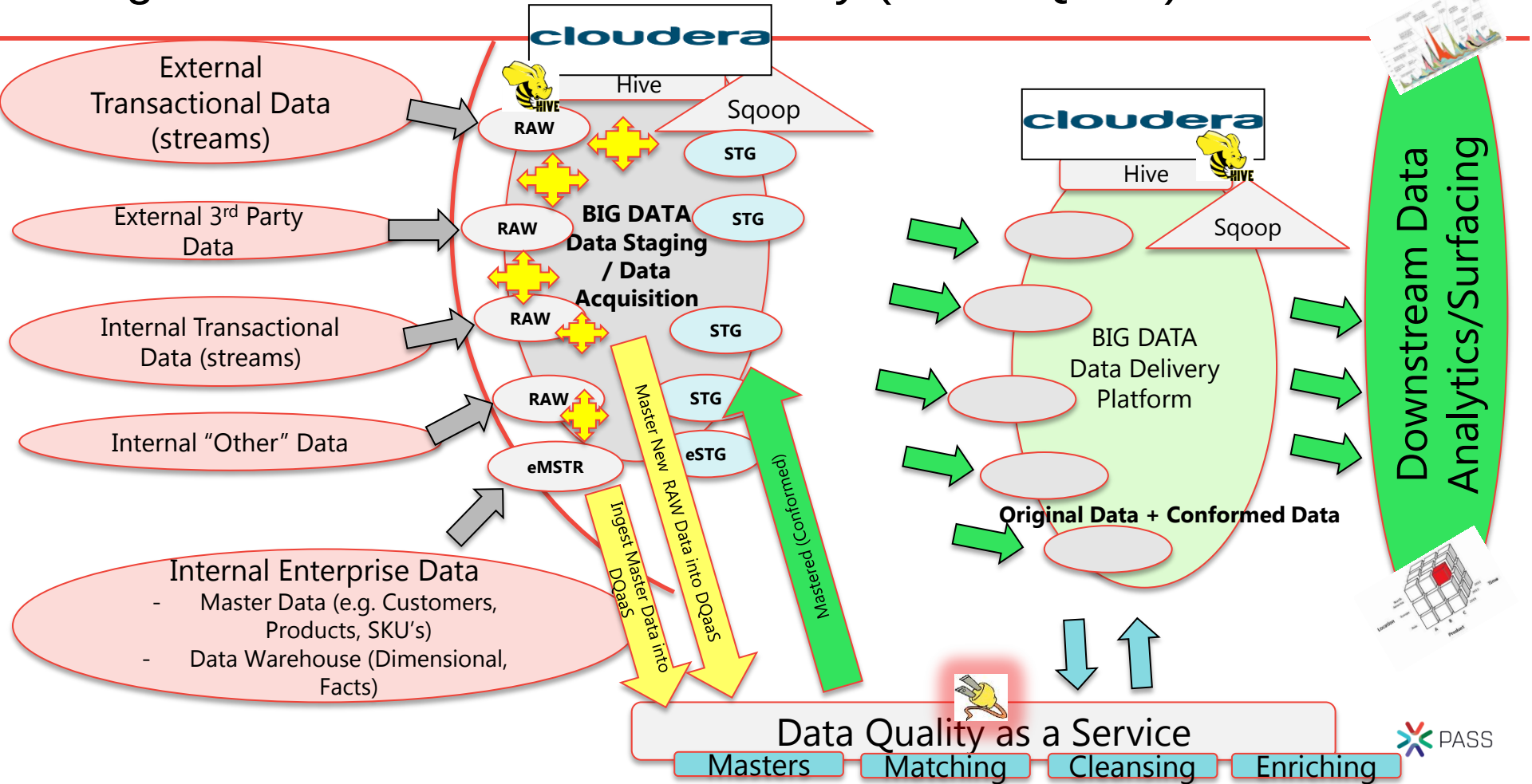


Measure 100,000 random values from Data Delivery Platform

- Number of duplicates	457
- Number of Invalid's (garbage)	8,906
- Number of Valid's (not garbage)	89,631
- Other (undecided)	1,006

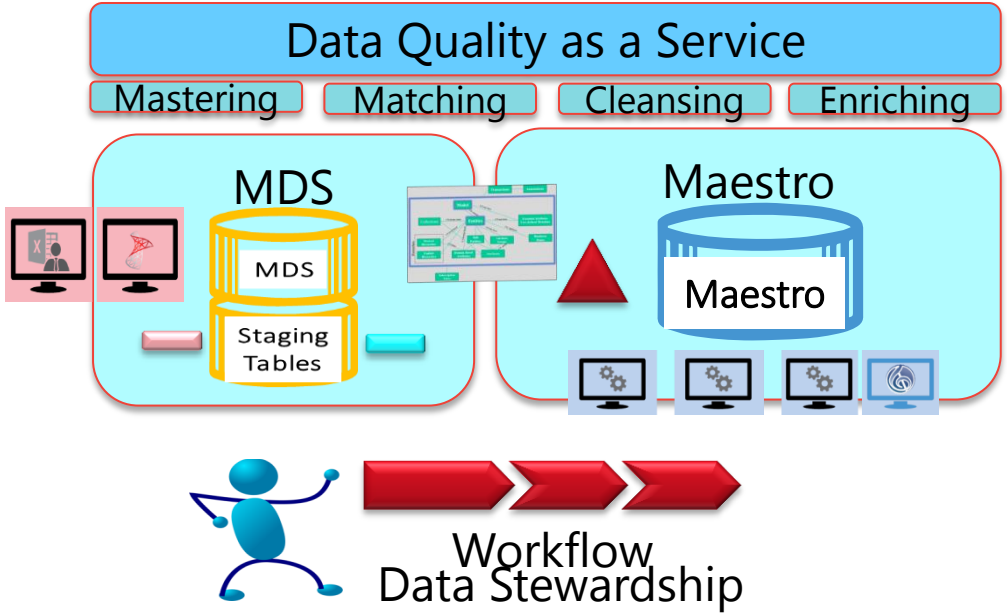
PASS

A big data effort we finished recently (with DQaaS)



PASS

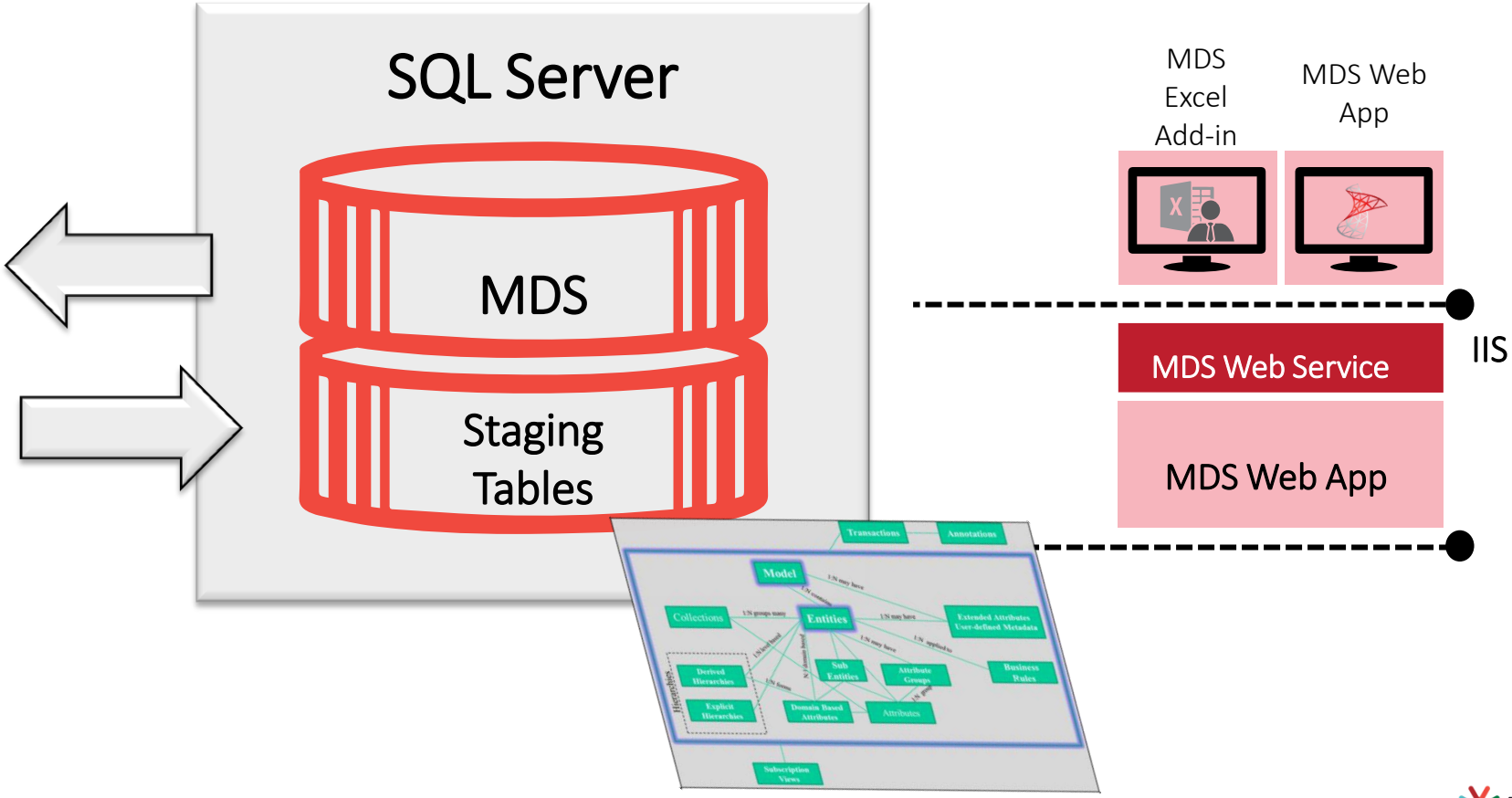
Data Quality as a Service



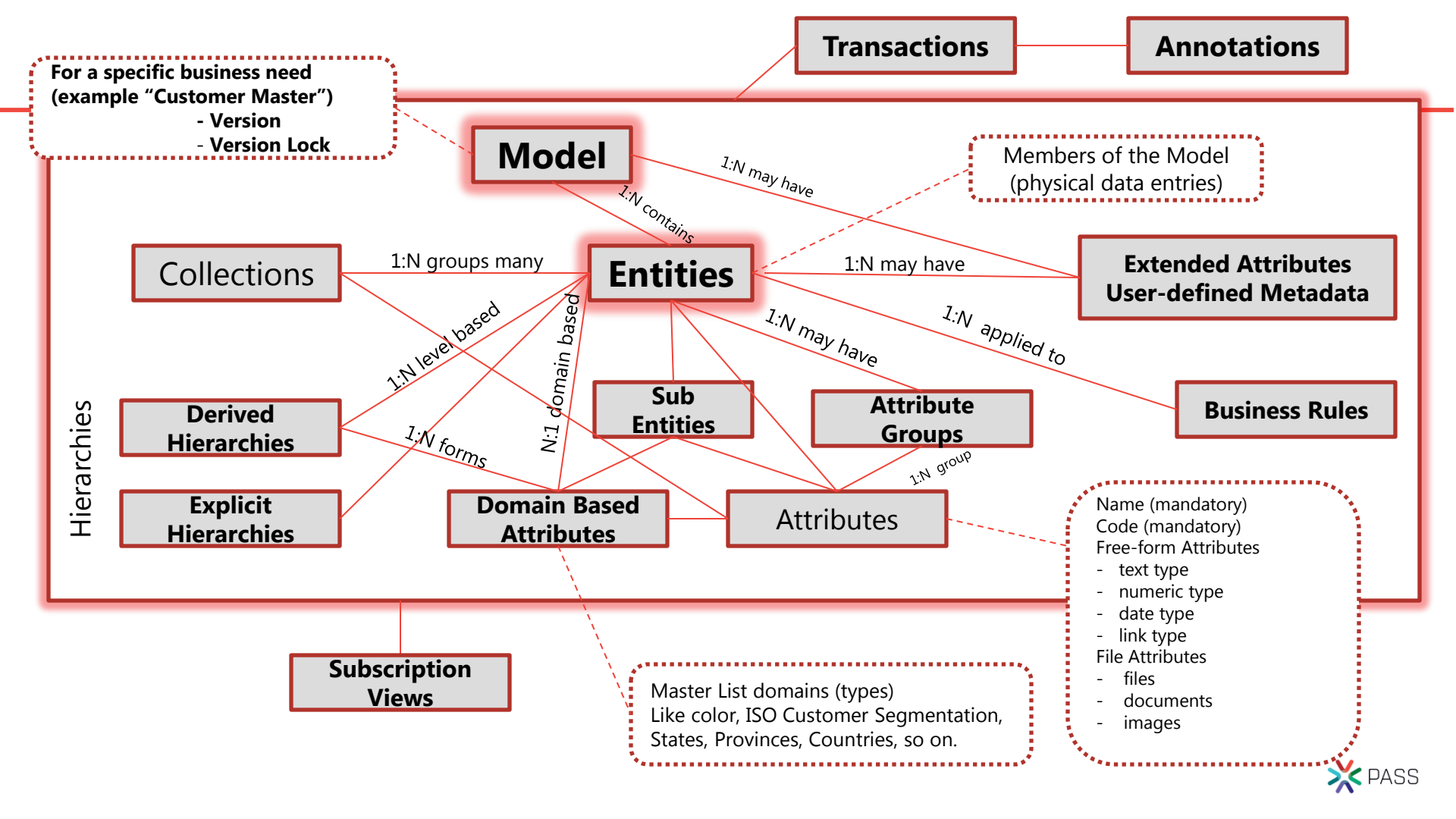
Master Data Services

- Master Data Management Platform on SQL Server
- Model & Rules – Managed Schema
- Security and Access
- Bulk data loads & consumption – table access
- Hierarchy Management
- Deployment, management, versioning
- Application-level transaction management

Master Data Services



PASS


Transactions
Annotations
For a specific business need
(example "Customer Master")
- Version
- Version Lock
Model
1:N may have
1:N contains
Members of the Model
(physical data entries)
Collections
1:N groups many
Entities
1:N may have
Extended Attributes
User-defined Metadata
1:N level based
N:1 domain based
1:N may have
1:N applied to
Hierarchies
Derived Hierarchies
1:N forms
Sub Entities
Attribute Groups
Business Rules
1:N group
Explicit Hierarchies
Domain Based Attributes
Attributes
Name (mandatory)
Code (mandatory)
Free-form Attributes
- text type
- numeric type
- date type
- link type
File Attributes
- files
- documents
- images
Subscription Views
Master List domains (types)
Like color, ISO Customer Segmentation,
States, Provinces, Countries, so on.
PASS

Profisee

- User Experience – Stewardship, Access, Manage
- Workflow – Initiate, Approve, Contribute, Calculate
- Golden Record Management – Matching, Survivorship
- Data Quality – Verification, Address, Person, Email
- Application Integration – MDM, CRM, Federation
- MDM Programmability – Web Objects, Web Services

Filtering the Data Lake

- ❑ Matching Strategies
- ❑ Survivorship
- ❑ Dedupe
- ❑ Normalization (canonical)
- ❑ ~~Harmonization~~
- ❑ Golden Records
- ❑ Taxonomies
- ❑ Cleansing
- ❑ Standardization
- ❑ Defaults
- ❑ Enriching

Data Quality as a Service

Masters | Matching | Cleansing | Enriching

Great options, even better opportunities

- Understand your processing and data requirements!
  - Strive for high quality data that is relevant to your most important business drivers/needs!
- Work within a consistent framework that provides you the needed performance, access, compliance, and quality your company demands!
  - Plug in data quality (DQaaS) as early as you can in the big data food chain (give your data VERACITY) (starting at acquisition (ingest) time)
- Big Data is not only Hadoop!

Session evaluations

Your feedback is important and valuable.

Submit by 5pm Friday, November 16th to win prizes.

3 Ways to Access:

Go to passSummit.com

Download the GuideBook App
and search: PASS Summit 2018

Follow the QR code link displayed on session signage throughout the conference venue and in the program guide

PASS

PASS SUMMIT2018

Thank You

Learn more from Paul Bertucci

@ptbertucci

pbertucci@dataXdesign.com

DXD Operations

- **USA (22 years)**
  Paris, France (15 years)
- **Database/Data Architecture**
  - **RDBMS's:**
    - Oracle, PostGres, MySQL, DB2, ……
    - Microsoft SQL Server & Analysis Services
  - **Master Data Management**
    - MDS/DQS
    - Maestro/Profisee
    - Oracle/CDH
    - IBM (Initiate)
  - **Big Data**
    - Hadoop, ParAccel, NoSQL
  - **Performance and Tuning**
  - **High Availability and DR/BC**
  - **Security/Encryption**
  - **Data Modeling/Database Design**

- **Database Tools:** P&T SQL Shot
  Highly graphical for Sybase, Oracle and MS SQL Server
- **Database Education & Training**
- **Partnerships**
  - **Microsoft**
  - **Profisee**

Contact me → pbertucci@dataXdesign.com

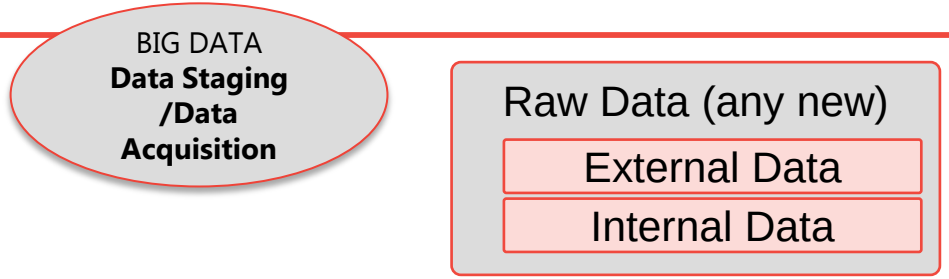
Cleaning up your Big Data Lakes with Data Quality as a Service
Speaker: Paul Bertucci
Duration: 75 minutes
Track: Design
Technology Focus: Big Data and IoT
Audience: Database Developer, Architect, Analyst
Level: 200

Data of poor quality is the single most impactful thing that is affecting the usefulness of both enterprise data environments and Big Data (data lakes). You would not likely drink dirty water (dirty data) let alone try to use it for decision making with any degree of confidence.

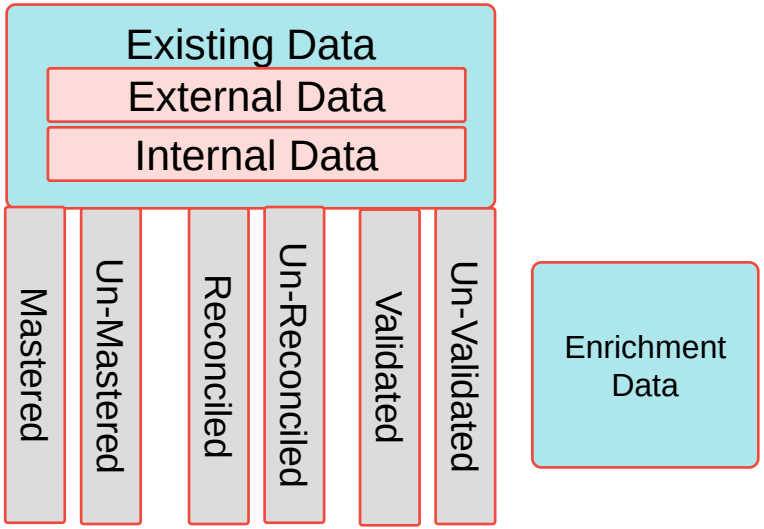
Mr. Bertucci will discuss an emerging strategy around how to plug a Data Quality as a Service (DQaaS) capability into your emerging data lakes (or current enterprise data architecture) which can best be thought of conceptually as a "data filtration" architecture and takes a "be actionable and design in the solution from the beginning" approach.

As a part of this presentation, Mr. Bertucci will also present a high-tech company's large scale use case and business drivers around using this type of Data Quality as a Service approach and share with the audience the result they realized.
Prerequisites: Some exposure to MDS/DQS or other master data management and data quality concepts. Some exposure to Big Data patterns and deployments.

Environments we set up – general architecture



All data starts out as un-mastered, un-reconciled, and un-validated



Mastered = matched, deduped, cleansed, Hierarchically reconciled, taxonomy adjusted, Standardized and defaulted, selectively merged (survivorship at the attribute level)

Enriched = extend existing data with additional Information from other sources

PASS