

Data Basics: Defining, Sourcing and Harmonizing Data

Masterclass Series (Class 1)
February 17th, 2021

<https://events.ethisphere.com/events/data-analytics-masterclass-series/>



ETHISPHERE[®]
GOOD. SMART. BUSINESS. PROFIT.[®]



Thank you for joining!
Before we get started...

Q&A

Please submit your questions using the Questions feature in your Zoom Experience.

CHAT

Need assistance? The Chat feature will be open throughout the webcast.

RECORDING & PPT

Today's presentation, recording and additional resources will be provided via email after the webcast.

DON'T FORGET

Join us for the extended masterclass series!

March 3rd | Using Data: Generating Compliance

March 17th | Practical Data Analytics: Deep Dives Into Specific Use Cases

Speaker



Andy Miller

Chief Analytics Officer @ Lextegrity

Helping organizations mitigate risk across the enterprise, specifically through employing advanced analytics to expose fraud, bribery, corruption, conflicts of interest and sanctions violations within enterprise data.

Contact me @ amiller@lextegrity.com

Learn more @ <https://www.lextegrity.com/>



The Lextegrity & Ethisphere Partnership

- Knowledge & Best Practices Sharing
<https://ethisphere.com/what-we-do/lextegrity/>
- Data Analytics Working Group
- Integrity Analytics Collective
<https://www.lextegrity.com/collective>



Lextegrity's Mission

Data-Driven + Intuitive + Integrated

FIRST GENERATION



INTEGRITY GATEWAY



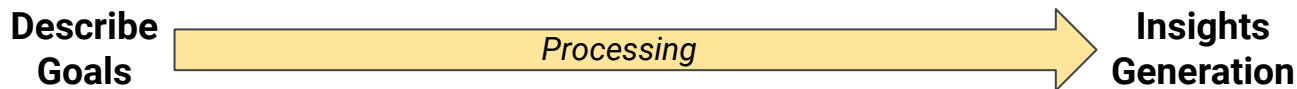
Agenda

- Introductions
- **Raw Data to Information to Insights**
- Q&A

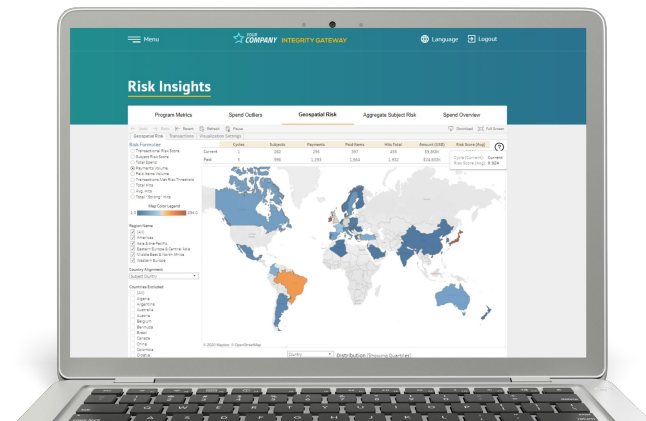
The goal is to turn data into information,
and information into insight.

CARLY FIORINA, FORMER CEO OF HP

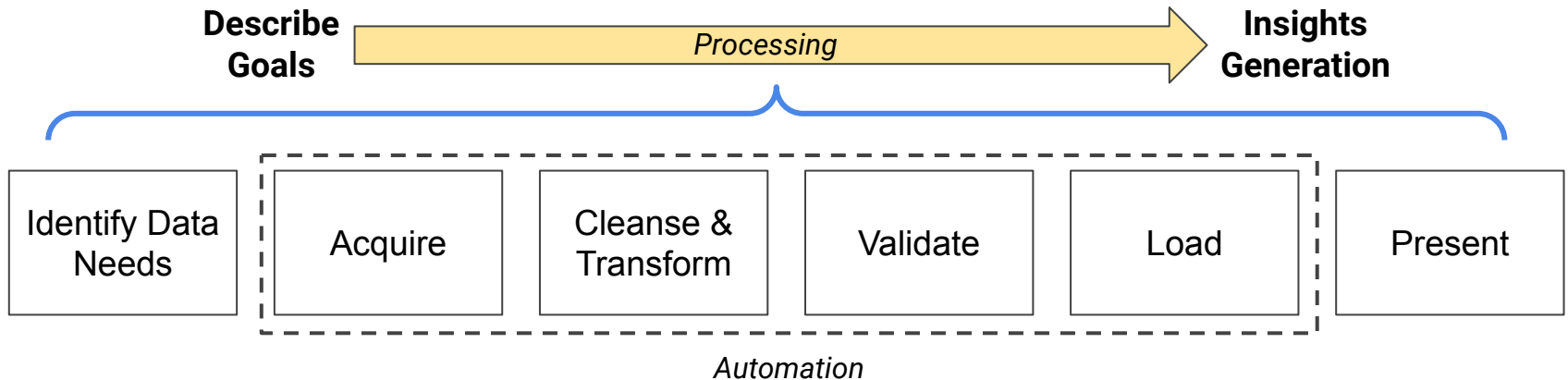
Raw Data to Information to Insights

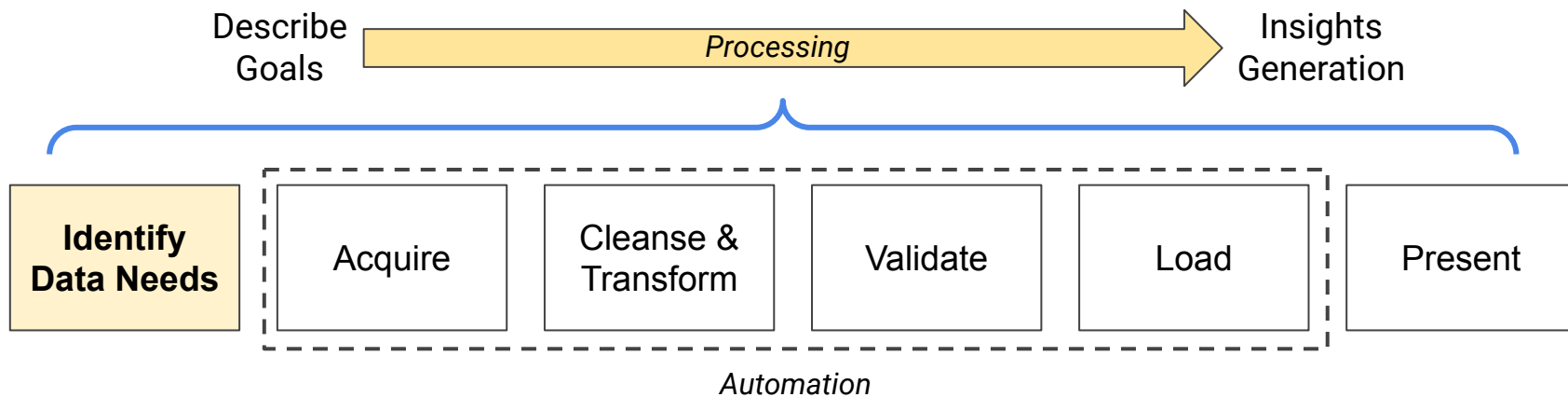


- “Having a pulse on organizational spend”
- “Preventing & detecting fraud, bribery, corruption, conflicts of interest, and sanctions violations”
- “Monitoring higher-risk spend / vendors / employees”
- “Prioritizing spend items for additional review when exhibiting higher-risk attributes”



Raw Data to Information to Insights





The 5 V's of Data

Volume

Scale of Data

Variety

Different Forms
of Data

Value

Potential of Data

Velocity

Fast Moving Data

Veracity

Uncertainty of Data

Relevant Data for Compliance

Transactional

- ERP (e.g. Oracle, SAP)
- T&E Systems (e.g. Concur, Expensify)
- Third Party Due Diligence

Master Data

- HR Systems (e.g. Workday, SuccessFactors)
- Master Data Sources (e.g. vendor/customer lists, HCPs, etc.)

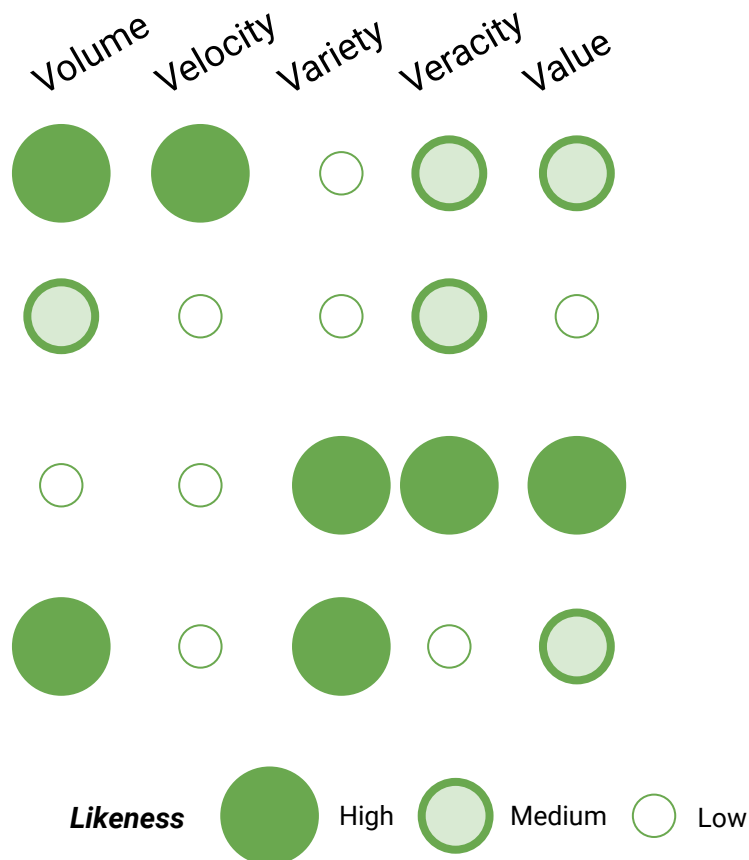
Supportive

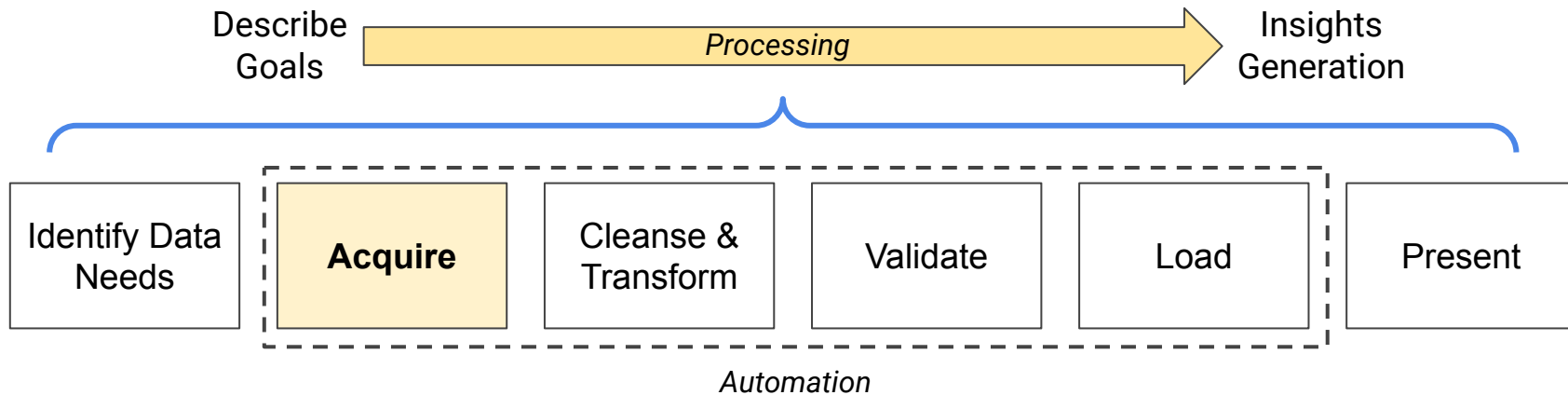
- Risk Program Data (e.g. investigations, training statistics, hotline calls, audit exceptions)
- Aggregated Metrics (e.g., budget to actual financial variances, HR turnover metrics)

Open Source / Publicly Available

- Industry Competitive Sources / Benchmarks
- Transparency Data (Life Sciences)
- Corruption Perception Index (CPI)

The 5 Vs of Data



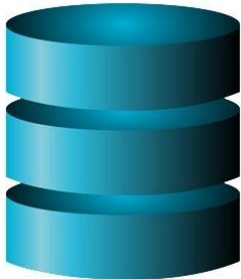


Data Accessibility

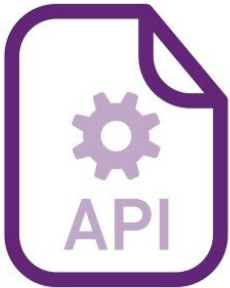
Example Data

Name	Date Modified	Size	Kind
compresed_data.tdsx	Jan 30, 2021 at 5:53 AM	6.8 MB	Tableau...Source
concur_expenses.xlsx	Jan 29, 2021 at 9:23 AM	47 KB	Spreadsheet
CPI2020_GlobalTablesTS_210125.xlsx	Jan 25, 2021 at 3:41 PM	82 KB	Spreadsheet
CPI2020_SignificantChanges_210125.xlsx	Jan 25, 2021 at 3:41 PM	42 KB	Spreadsheet
critical process details.log	Jan 31, 2021 at 3:07 PM	8 KB	Log File
Entity Hierarchies.json	Today at 8:46 AM	262 KB	JSON
Extremely Important Receipt.png	Jan 19, 2021 at 2:03 PM	40 KB	PNG image
Important Contract.docx	Yesterday at 4:28 PM	81 KB	Microso...(docx)
Important x2 Contract.pdf	Feb 8, 2021 at 2:48 PM	148 KB	PDF Document
sap_payments.xlsx	Jan 29, 2021 at 9:23 AM	199 KB	Spreadsheet
transactional_data.csv	Jan 29, 2021 at 9:23 AM	327 KB	comma...d values
very_important_database.sql	Nov 25, 2020 at 1:54 PM	110.59 GB	SQL File

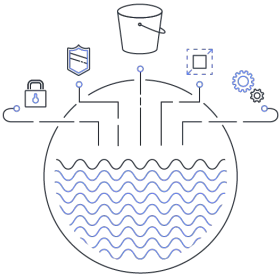
Flat Files



Databases



Application
Programming
Interface



Data Lake

Data Formats: Structured

Database Table

	id	key	value	code2	region	cpi_score
1	6d6b6bb7-0883-4444-a80b-b8e6875d5256	DK	Denmark	DK	Europe	88
2	7396158d-dcff-4c1c-98e9-2ea7fbb4cd4c	NZ	New Zealand	NZ	Western Pacific	87
3	b82b74fe-1a7f-4a28-92c0-3e9d42af7359	CH	Switzerland	CH	Europe	85
4	2d807e73-ba03-49cb-a4b9-e4e2e446c87c	SG	Singapore	SG	Western Pacific	85
5	347bf5cb-22e7-4d7c-a02e-d3b9e7ad84d9	SE	Sweden	SE	Europe	85
6	099ac6ab-742f-4d0f-8abe-38f382a517ec	FI	Finland	FI	Europe	85
7	551d41f2-15b1-4c3a-97e8-0241d1047d1d	NO	Norway	NO	Europe	84
8	1c7c0205-86ed-4d8e-ab70-eac5229e573	NL	Netherlands	NL	Europe	82
9	43d9d6fa-3501-47a4-b6fe-819cbd6bb116	CA	Canada	CA	Americas	81
10	83ab8d9e-8356-4913-9304-28e6591095a4	LU	Luxembourg	LU	Europe	81
11	ab81ea9e-3619-4f04-a9e9-7c96b1d49c4f	DE	Germany	DE	Europe	80
12	0e2a4fc4-9f1f-4843-a184-5278364a3cfc	GB	United Kingdom	GB	Europe	80
13	d9833297-1155-44df-80e0-d58c17418aaa	AU	Australia	AU	Western Pacific	77
14	2dd87ef5-b836-46a0-aa9d-bae689480e60	IS	Iceland	IS	Europe	76
15	bae72bf1-7dba-41cd-b8eb-35130eda90fa	HK	Hong Kong, SAR China	HK	No Region Specified	76
16	0ce4b142-428e-4b7d-8615-65c6539c6d16	AT	Austria	AT	Europe	76
17	213a714a-0179-4f42-baab-e2a6d5c7e9bf	BE	Belgium	BE	Europe	75
18	0d5809ac-af61-4ebe-8b76-95139413d089a	EE	Estonia	EE	Europe	73
19	1b6ed6c9-9da4-44c7-9741-1eb9432a8ab6	JP	Japan	JP	Western Pacific	73
20	25036070-e159-4f29-8437-e610053b97df	IE	Ireland	IE	Europe	73
21	10d5257c-b574-4da3-ab09-ed8c21c17c95	FR	France	FR	Europe	72
22	804f14f3-1fbd-4afe-b46c-5ef2c631097b	US	United States of America	US	Americas	71
23	1529cb16-70ff-423b-8a2f-601eb772e207	AE	United Arab Emirates	AE	Eastern Mediterranean	70
24	8fc07b50-65ba-466e-adbb-ff8b2bcff2ad	UY	Uruguay	UY	Americas	70
25	4cf7a44a-a521-41de-a311-960d1814d2f3	BT	Bhutan	BT	South-East Asia	68
26	df0811d7-3764-4f97-a1a1-c9ce241740bf	BB	Barbados	BB	Americas	68
27	812f382a-dbf6-4c97-845c-c436e4ee5c34	CL	Chile	CL	Americas	67
28	4fae67b2-596f-4ce1-864f-c087bbac1d09	SC	Seychelles	SC	Africa	66
29	9409e504-ea69-41fa-a694-da95c2a24e64	BS	Bahamas	BS	Americas	65
30	b509dc2e-6d55-486d-9c0c-3b65705c5d52	PT	Portugal	PT	Europe	64
31	bffec4f2-099c-4d1b-b41c-3dcea60af682	TW	Taiwan, Republic of China	TW	No Region Specified	63
32	f48fd5b5-4a4d-4195-9af1-82e91ca8607a	BN	Brunei Darussalam	BN	Western Pacific	63
33	0e22de58-15f3-41fb-9199-70c0c34b9286	QA	Qatar	QA	Eastern Mediterranean	62

Data Formats: Semi-structured

XML

```
<?xml version="1.0" encoding="UTF-8"?>
<rdData>
  <dtOrders ShipCountry="France" ShipPostalCode="51100" ShipCity="Reims" ShipAddress="59 rue de
l'Abbaye" ShipName="Vins et alcools Chevalier" Freight="32.38" ShipVia="3" ShippedDate="2010-07-
16T00:00:00.000" RequiredDate="2010-08-01T00:00:00.000" OrderDate="2010-07-04T00:00:00.000"
EmployeeID="5" CustomerID="VINET" OrderID="10248"/>
  <dtOrders ShipCountry="Germany" ShipPostalCode="44087" ShipCity="Münster" ShipAddress="Luisenstr.
48" ShipName="Toms Spezialitäten" Freight="11.61" ShipVia="1" ShippedDate="2010-07-
10T00:00:00.000" RequiredDate="2010-08-16T00:00:00.000" OrderDate="2010-07-05T00:00:00.000"
EmployeeID="6" CustomerID="TOMSP" OrderID="10249"/>
  <dtOrders ShipCountry="Brazil" ShipPostalCode="05454-876" ShipCity="Rio de Janeiro" ShipAddress="Rua
do Paço #67" ShipName="Hanari Carnes" Freight="65.83" ShipVia="2" ShippedDate="2010-07-
12T00:00:00.000" RequiredDate="2010-08-05T00:00:00.000" OrderDate="2010-07-08T00:00:00.000"
EmployeeID="4" CustomerID="HANAR" OrderID="10250" ShipRegion="RJ"/>
  <dtOrders ShipCountry="France" ShipPostalCode="69004" ShipCity="Lyon" ShipAddress="2 rue du
Commerce" ShipName="Victuailles en stock" Freight="41.34" ShipVia="1" ShippedDate="2010-07-
15T00:00:00.000" RequiredDate="2010-08-05T00:00:00.000" OrderDate="2010-07-08T00:00:00.000"
EmployeeID="3" CustomerID="VICTE" OrderID="10251"/>
  <dtOrders ShipCountry="Belgium" ShipPostalCode="B-6000" ShipCity="Charleroi" ShipAddress="Boulevard
Tirou #255" ShipName="Suprêmes délices" Freight="51.3" ShipVia="2" ShippedDate="2010-07-
11T00:00:00.000" RequiredDate="2010-08-06T00:00:00.000" OrderDate="2010-07-09T00:00:00.000"
EmployeeID="4" CustomerID="SUPRD" OrderID="10252"/>
</rdData>
```

JSON

```
[
  {
    "id": "ab7131e9-b328-443a-ad4a-7d566dac251",
    "key": "AD",
    "value": "Andorra",
    "code2": "AD",
    "region": "Europe"
  },
  {
    "id": "1529cb16-70ff-423b-8a2f-681eb772e207",
    "key": "AE",
    "value": "United Arab Emirates",
    "code2": "AE",
    "cpi_score": 70
  },
  {
    "id": "1e7d02b3-69fd-4111-8d9a-cbda88792097",
    "key": "AF",
    "value": "Afghanistan",
    "code2": "AF",
    "languages": ["Pashto", "Dari Persian"],
    "president": "Ashraf Ghani",
    "population": 3225560,
    "gdp": {
      "amount": 7291000000,
      "currency": "USD"
    },
    "cpi_score": 16
  },
  {
    "id": "18b4a9f7-1e7f-4b74-959e-7b2f211dc0e6",
    "key": "AG",
    "value": "Antigua and Barbuda",
    "code2": "AG",
    "region": "Americas"
  },
  {
    "id": "29d14a7b-1203-4d06-854e-83da0655de3a",
    "key": "AI",
    "value": "Anguilla",
    "code2": "AI",
    "region": "Americas"
  },
  {
    "id": "cfc79608-400f-4d3f-8aef-28a36764a187",
    "key": "AL",
    "value": "Albania",
    "code2": "AL",
    "region": "Europe",
    "cpi_score": 36
  }
]
```

Data Formats: Unstructured

Doctor's Clinical Notes

■ Figure 2. Example of a Clinical Note

Documents Browser - 3 982-540 Hurt, John S.

File Edit Preferences... View Help

Select Documents
☐ My Documents
☐ Specific Patient
☐ Work List
☐ Specific Author

Mayo Clinic # 3-982-540 Get MCR Display Last 3 Years View Note Sections
 Patient Name Hurt, John S. Retrieve

Date/Time	Clinic Number	Patient Name	Provider Name/Pager	Serv	Desc	Loc	Type	Stat	Ltr
18 Nov 2004	3-982-540	Hurt, John S.	By, Mary E, LSNH-6404	SOONRK			CON	Fri	
18 Nov 2004	3-982-540	Hurt, John S.	Langworthy, Valerie LN27-04567	BNDO			ME	Fri	

2837 document(s) found Username: Burstock, James D Retrieve document - 0.25 seconds

Demog Reg Info Revision History Ref/CC HPI Meds Allergy Snr Rx Med Surg Hx Soc Hx Fam Hx Vitals Exam I/RP Disposition Specialty FULL Letter

CHIEF COMPLAINT/PURPOSE OF VISIT:
THIS IS A TEST NOTE. Elevated PSA, GERD, Hearing loss, etc.

HISTORY OF PRESENT ILLNESS:

#1 Elevated PSA
He has had PSA values greater than 7 off and on for a couple of years and prostate volume somewhat proportionately enlarged. Last check here PSA was 7.6 with a prostatic volume of 34 cc which calculated to a 0.22 PSA per volume which is higher than the allowable suggested limit. As best we can tell though, nothing early had changed on exam or even on ultrasound and therefore we did not feel that further intervention was necessary. The patient indicates that his physician in Indiana concurred with that plan and just that we need to keep track of things closely as time goes by. He has no new voiding symptoms.

#2 Visual disturbance
See last year's note. It was suspicious that he did have an amaurosis event. Because of that, we advised that he start taking aspirin on a daily basis. He has had no subsequent symptoms similar to what we described last year. Because of this additional studies had been done which included carotid ultrasound, all of which were fundamentally unremarkable.

#3 Hearing loss
Chronic problem, seems to be fairly stable.

#4 GERD
Periodically he gets fairly typical nocturnal reflux symptoms. This seems to occur most often when he has eaten later than the conventional time for him. It does not really seem to make much difference on food type. Some question was raised about whether this might be some type of food poisoning symptom but the story is rather classic for acid reflux. He gets a very bitter taste in the posterior part of the mouth and throat, washing it out with water clears that. Sometimes he has to cough up "phlegm" that can have mixed mucus.

He may have similar symptoms rarely during the normal upright day but most of these episodes have occurred during the night. We had actually planned an upper GI x-ray a couple of years ago but because of situations could not complete that examination while here.

#5 Exercise leg fatigue
This is a very vigorous man who walks a mile or two a day at a very brisk pace (approximately 4 miles per hour). When he is out doing heavy work in his yard such as pushing a mower, etc. for greater than half an hour he does notice leg fatigue. A very brief rest and he can get back to his activities. He has really no other symptoms suggesting ischemic problems in the lower extremities.

#6 Positional dizziness
For a year or two he has noticed that when he gets up out of a bed or chair rapidly he gets lightheaded and he actually gets some true vertigo-type symptoms for an instant. He has been counseled by his local physician that this is related to aging of the ear and the story is fairly typical for benign positional vertigo.

#7 Health maintenance
Last colon check was in March of 1998. Last PSA checked during the past calendar year. No chest x-ray for a year.

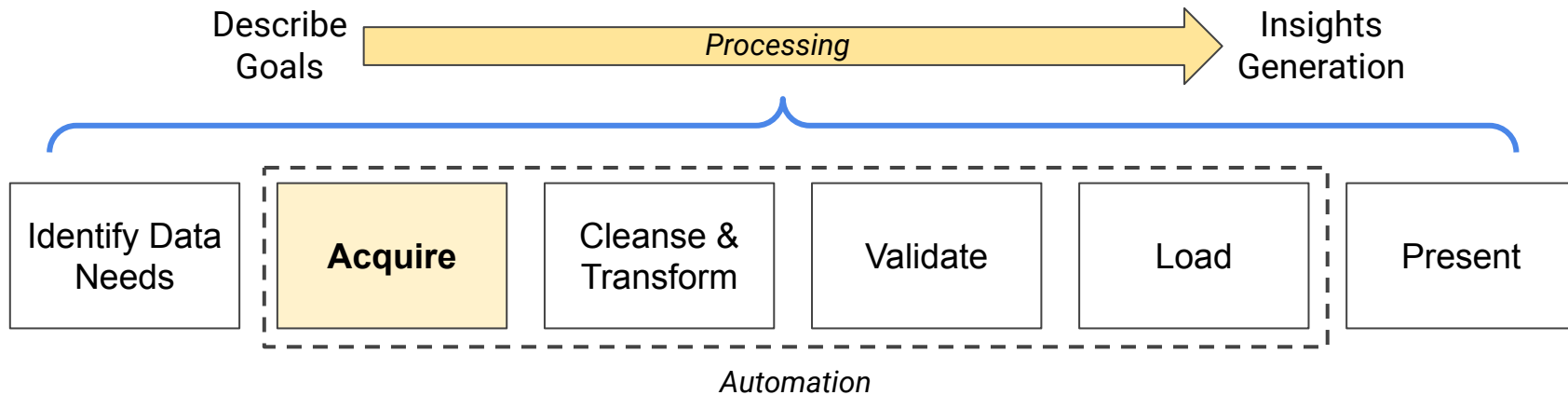
REVIEWED INFORMATION WITH PATIENT AS NOTED ON THE CURRENT VISIT INFORMATION FORM, DATED 22 MAR 1999 AND ON THE PATIENT FAMILY HISTORY FORM, DATED 9 MAR 1998.

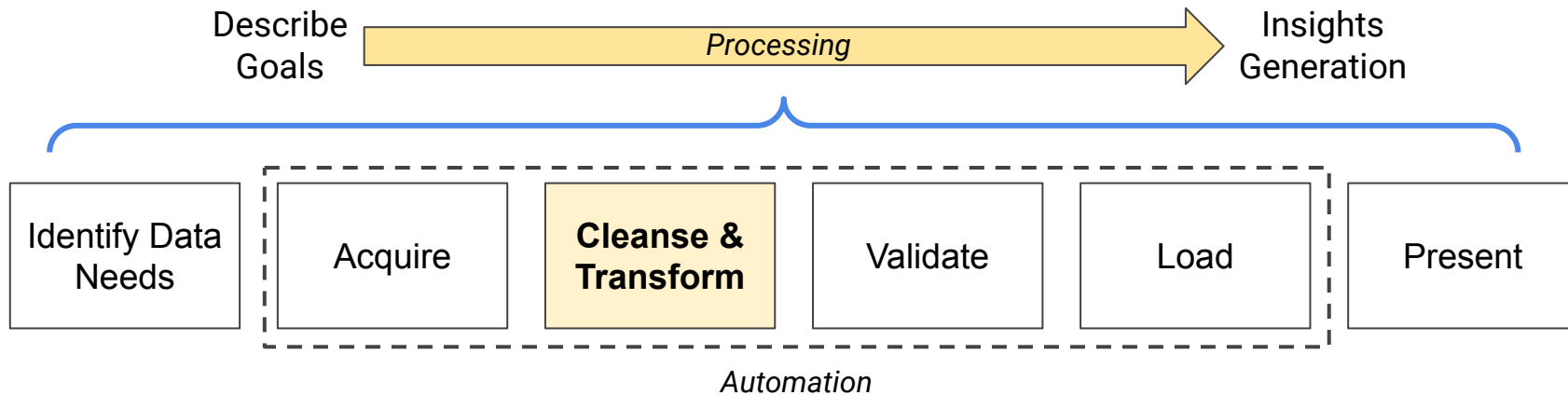
CURRENT MEDICATIONS:
 Lipitor 10mg daily
 ASA 365mg daily
 Meclizine 25mg 1 daily (discontinue)

VITAL SIGNS:
 Height: 176.0 cm, Weight: 76.80 kg, BSA: 1.95 M2, BMI: 24.793 KG/M2

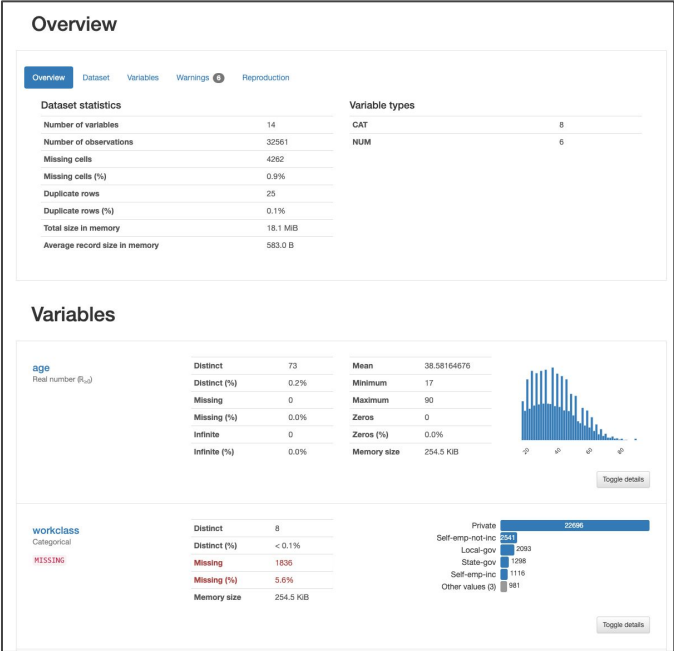
Next Previous Temp Print Edit New

Signs Dict Sign Close



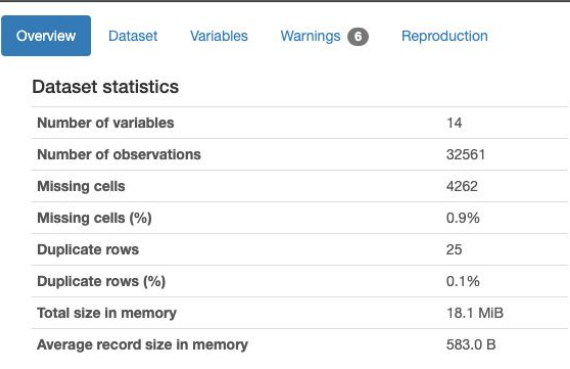


Example of Profiling Report

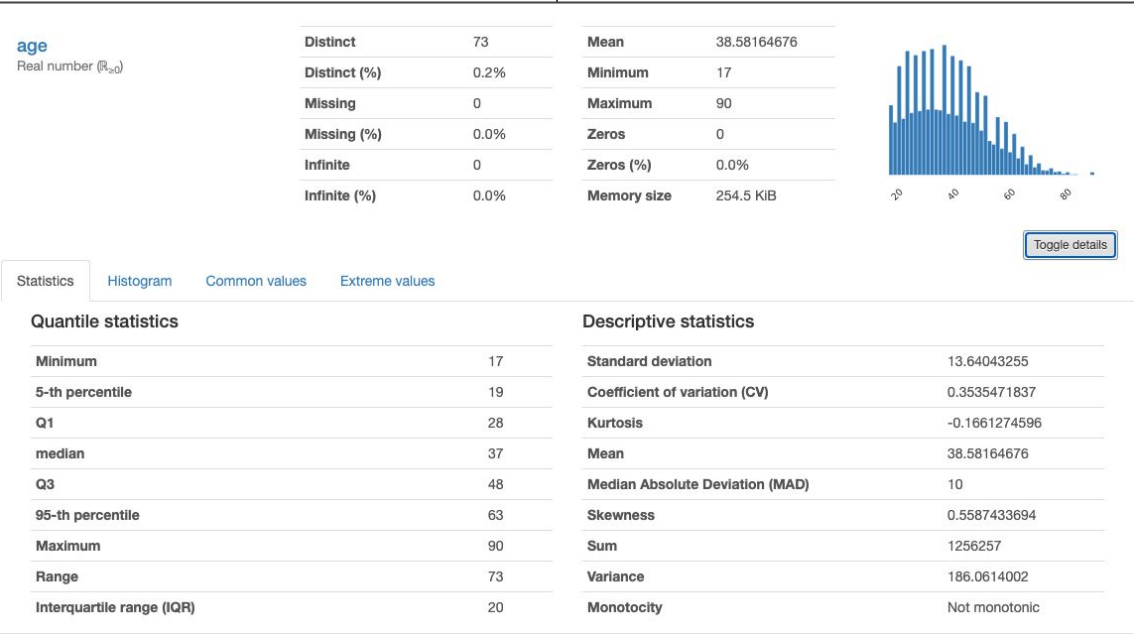


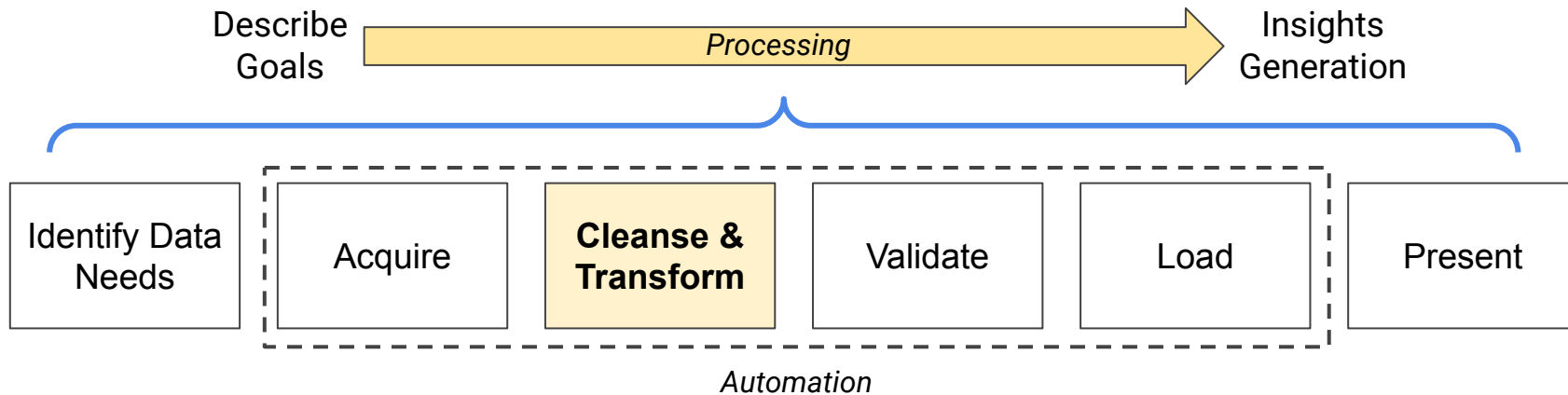
Reference:
<https://github.com/pandas-profiling/pandas-profiling>

Dataset Analysis

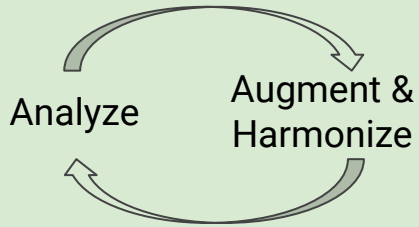


Variable Analysis





Data Transformation



Augmentation

High-Risk Subject Logic

- Vendor Type is "Government Related"
- Vendor Title/Profession contains "Government, Govt, Minister, Official, GO, etc."
- Vendor ID matches Third Party Due Diligence review "Risk Level" = "High"
- Vendor pre-approval request government official question is answered yes ("Is the vendor/individual associated with the the government?")

Elimination

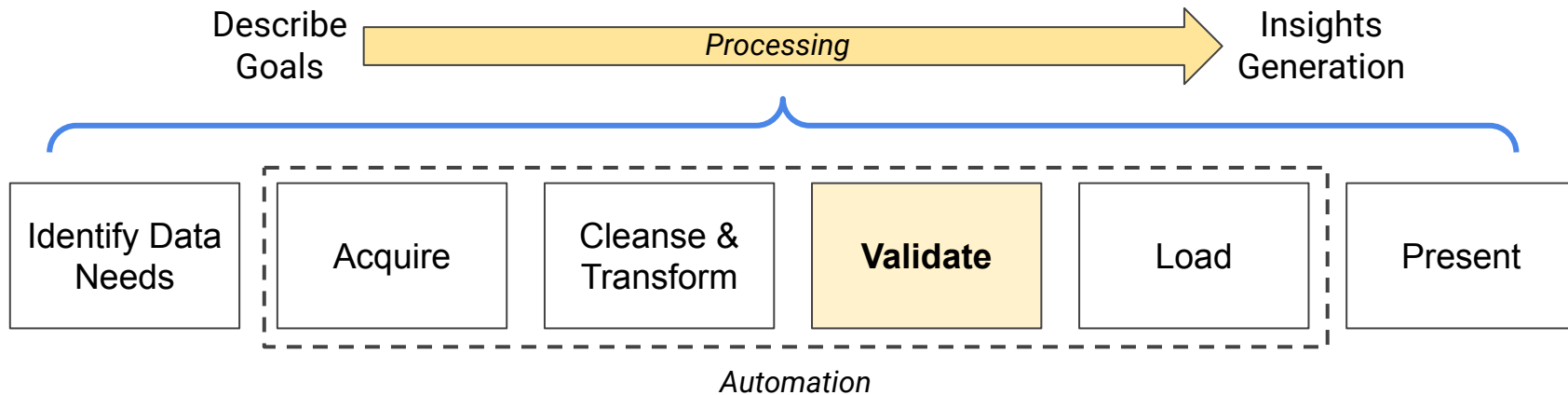
Components

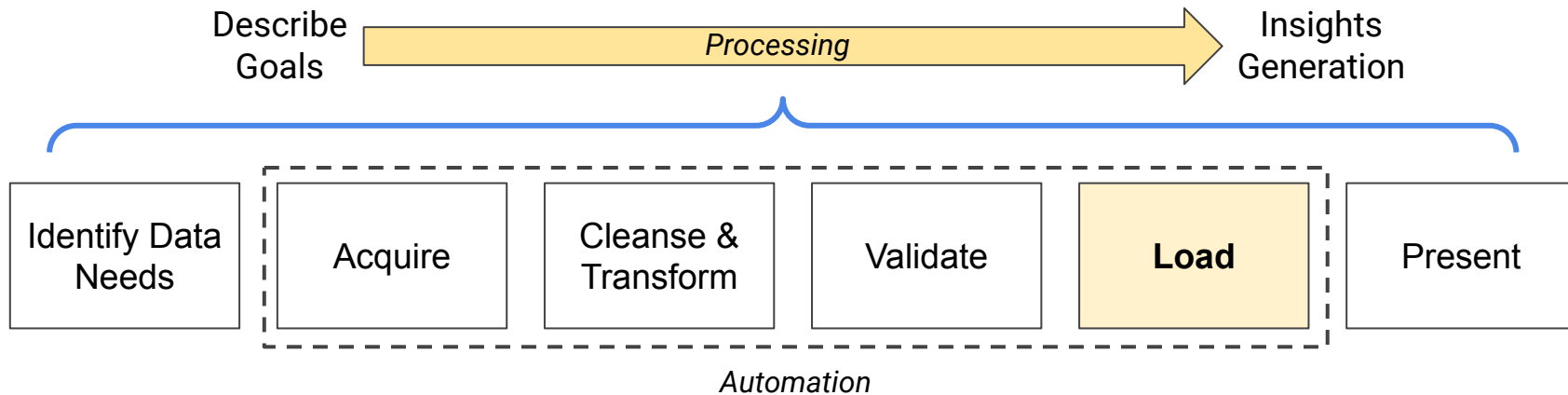
	Field	Key	Data Element	Domain	Data Type	Length	Decimal Places	Short Description
1	MANDT	<input checked="" type="checkbox"/>	MANDT	MANDT	CLNT	3	0	Client
2	LIFNR	<input checked="" type="checkbox"/>	LIFNR	LIFNR	CHAR	10	0	Account Number of Vendor or Creditor
3	ANRED	<input type="checkbox"/>	ANRED	TEXT15	CHAR	15	0	Title
4	BAHNS	<input type="checkbox"/>	BAHNS	BAHNH	CHAR	25	0	Train station
5	BBBNR	<input type="checkbox"/>	BBBNR	NUM07	NUMC	7	0	International location number (part 1)
6	BBSNR	<input type="checkbox"/>	BBSNR	NUM05	NUMC	5	0	International location number (Part 2)
7	BGRU	<input type="checkbox"/>	BRGRU	BRGRU	CHAR	4	0	Authorization Group
8	BRSCH	<input type="checkbox"/>	BRSCH	BRSCH	CHAR	4	0	Industry key
9	BUBKZ	<input type="checkbox"/>	BUBKZ	NUM01	NUMC	1	0	Check digit for the international location number
10	DATLT	<input type="checkbox"/>	DATLT	TEXT14	CHAR	14	0	Data communication line no.
11	DTAMS	<input type="checkbox"/>	DTAMS	DTAMS	CHAR	1	0	Report key for data medium exchange
12	DTAWS	<input type="checkbox"/>	DTAWS	DTAWS	CHAR	2	0	Instruction key for data medium exchange
13	ERDAT	<input type="checkbox"/>	ERDAT_RF	DATUM	DATS	8	0	Date on which the Record Was Created

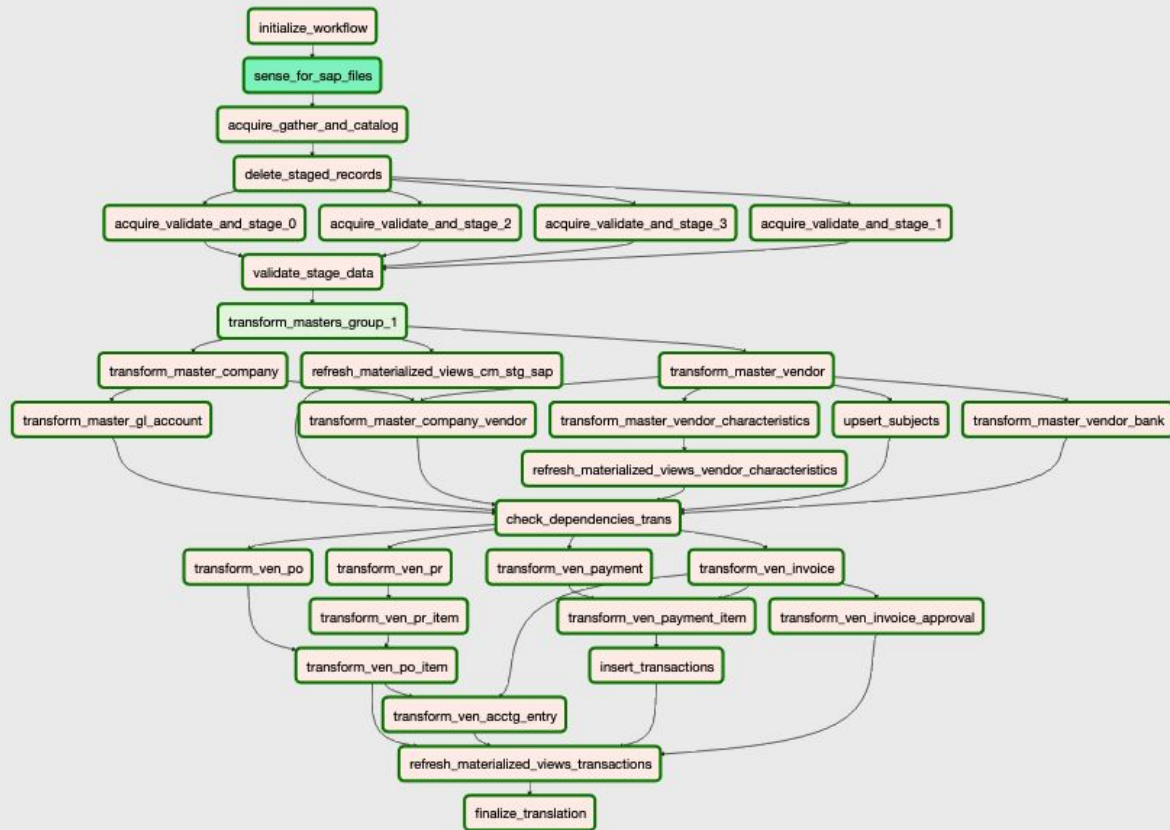
SAP LFA1 table components (Vendor Details)

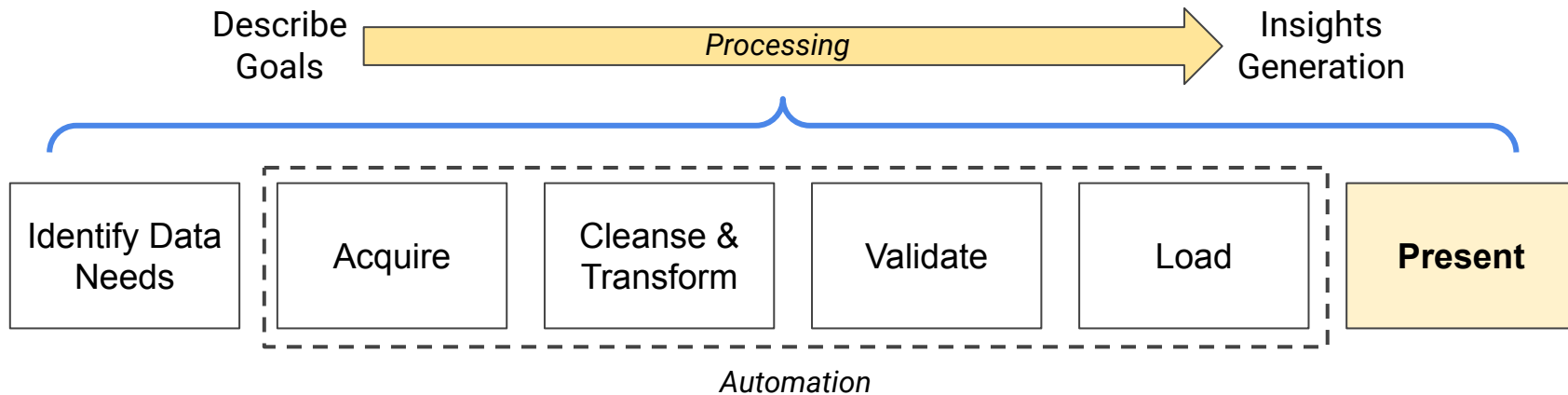
Harmonization

- Airfare - Overseas
 - Airfare - Domestic
 - Airfare - First Class (EU)
 - Airfare - Business Class (EU)
 - Non-Employee Airfare
 - Travel - Airfare Costs
 - Airfare > \$5K
- Airfare









[illegible]

Menu

YOUR COMPANY

INTEGRITY GATEWAY

Groups

Logout

Monitoring Review

STATUS	DOMAIN	SPEND DATE	NAME	COUNTRY	ENTITY	NATURE	AMOUNT	RISK SCORE
(ALL)	(ALL)							
Open	Vendor	2020-01-22	Hill Group	Canada	CC-356	PO Related	\$20,832.50	5.0
Open	HGX	2020-01-02	Wesley Miller	United States of A...	CC-650	Services - Non Co...	\$2,671.98	4.6
Open	Employee	2020-01-02	Sarah Hayes	Germany	CC-079	Meal	\$613.10	4.6
Open	Employee	2020-02-21	Stephen Espinoza	United States of A...	CC-650	Meal	\$1,557.01	4.2
Open	HGX	2019-11-14	Lisa Green	United States of A...	CC-650	Food and Beverage	\$67.73	4.1
Open	Employee	2020-01-19	Zachary Walker	United States of A...	CC-650	Meal	\$1,297.06	3.9
Open	HGX	2020-01-23	Blake Group	United States of A...	CC-650	Consulting Fee	\$138,224	3.8
Open	Vendor	2020-02-26	Donaldson, Larse...	Brazil	CC-350	PO Related	\$8.32	3.7
Open	Vendor	2020-03-26	Thomas Inc	United States of A...	CC-650	Non-PO Related	\$333.75	1.6
Open	Employee	2020-01-10	Joseph Bender	Ireland	CC-272	Transportation Ot...	\$2.95	1.5
Open	Employee	2020-01-20	Paula Wright	Mexico	CC-742	Fees Travel	\$12.05	1.2

Previous

Page: 1 of 1

20 rows

Next

Menu
 YOUR COMPANY INTEGRITY GATEWAY
 Groups Logout

Risk Insights

Geospatial Risk
Aggregate Subject Risk
Spend Outliers
Transparency Review
Ad-Hoc Views
Program Metrics

< Undo → Redo <- Revert Refresh Pause
Subject Profile | Risk Analysis Results | Transaction Detail | Supplemental Information
View: default Download Full Screen

Domain	Spend Date		Days	Items	Subjects	Subject Countries	Amount (USD)	Risk Results	Risk Score (Avg.)
(All)	1/31/2020	Selection Highlights	66	126,014	5,792	52	\$361,618,831	131,522	0.25

Risk Formula

- ☒ Average Risk Score
- ☐ Items > Risk Threshold
- ☐ Subject Risk Score
- ☐ Risk Results
- ☐ Risk Results (Strong Only)
- ☐ Auto Risk Results
- ☐ Average Risk Results
- ☐ Total Spend
- ☐ Spend Item Volume
- ☐ Total Spend With CPI

Toggle Trending View

Map Legend

0.10 ————— 1.36

Compliance Region

(All)

Country Alignment

Company

- ☒ (All)
- ☒ Argentina
- ☒ Australia
- ☒ Austria
- ☒ Belgium
- ☒ Bermuda
- ☒ Brazil
- ☒ Canada
- ☒ China
- ☒ Colombia
- ☒ Czech Republic
- ☒ Denmark
- ☒ France
- ☒ Germany
- ☒ India
- ☒ Ireland
- ☒ Israel
- ☒ Japan
- ☒ Mexico
- ☒ Netherlands
- ☒ Norway
- ☒ Portugal
- ☒ Russia Federation
- ☒ Sweden
- ☒ Switzerland
- ☒ Taiwan
- ☒ Thailand
- ☒ Turkey
- ☒ United Arab Emirates
- ☒ United Kingdom
- ☒ United States
- ☒ Uruguay
- ☒ Venezuela

© 2021 Mapbox © OpenStreetMap

Company Country Distribution by Average Risk Score (Showing Quartiles)

Additional Filters: ▼

Questions?

