

Amelia Arbisser
Adam Silberstein

Scalable and Incremental
Data Profiling with Spark

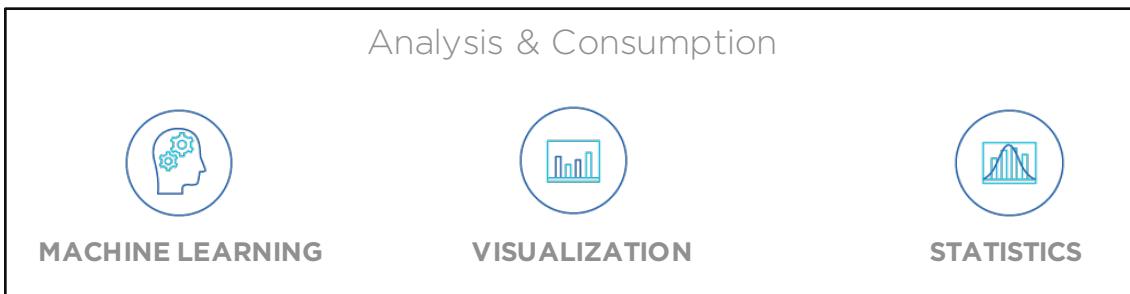
Trifacta: Self-service data preparation

What data analysts hope to achieve in data projects



Trifecta: Self-service data preparation

What data analysts hope to achieve in data projects

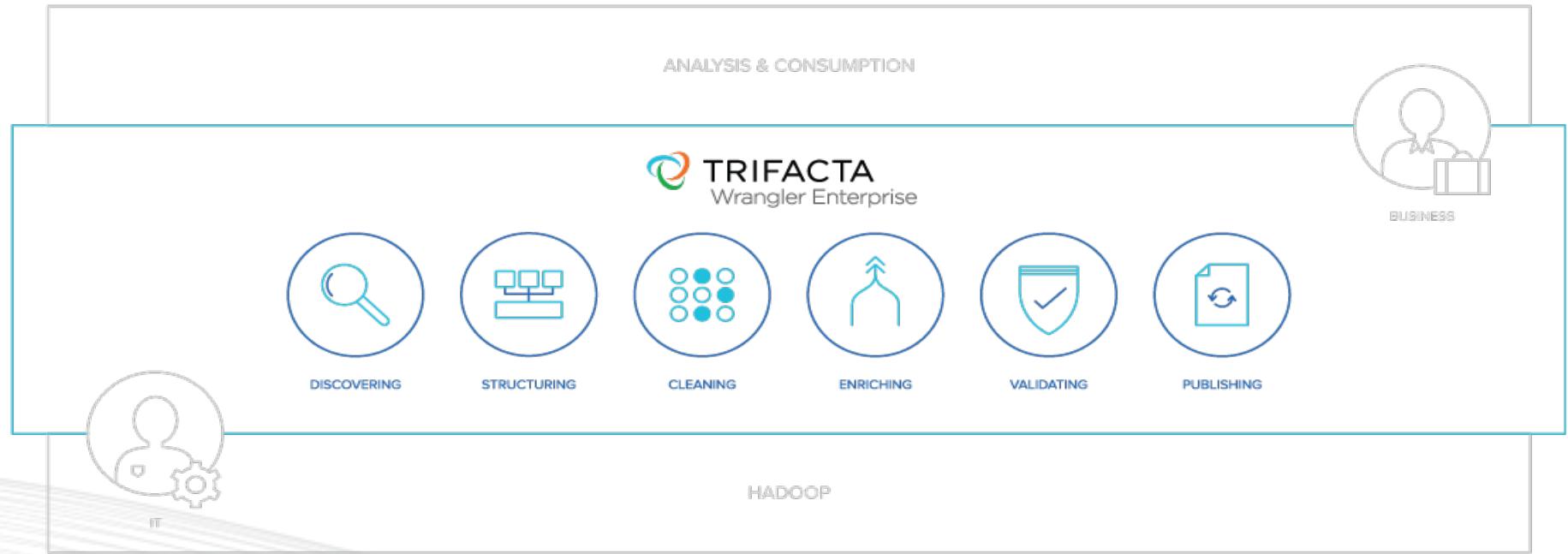


***80% of time spent cleaning and preparing the data
to be analyzed***



Trifacta: Self-service data preparation

We speed up and take the pain out of preparation



Transforming Data in Trifecta

The screenshot displays the Trifecta data transformation interface, which includes three main panels: Data Editor, Preview, and Log.

Data Editor: This panel shows a table with 4,270 rows and 13 columns. The columns are labeled: ID, DATE, SAFETY100, SAFETY90, SAFETY50, DURATION, DISEASE_REASON, GENDER, GENDER_TYPE, GENDER_TYPE_PRIMARY, GENDER_TYPE_SECONDARY, and GENDER_TYPE_TERTIARY. A sidebar on the left lists all columns with their current transformations. A red arrow points from the 'SAFETY100' column to a tooltip that says 'Transformed value'.

Preview: This panel shows a sample of 10 rows from the transformed data. The columns correspond to the ones in the Data Editor. The data includes various numerical values and categorical entries like 'Male' and 'Female'.

Log: This panel shows the history of changes made to the data. It includes a table with columns: ID, DATE, OldValue, NewValue, and Action. The log shows several recent actions, such as changing 'SAFETY100' from '2023-12-0013-00-10' to '2023-12-0013-00-10' at 2023-12-0013-00-10, and changing 'GENDER_TYPE_PRIMARY' from 'Male' to 'Female' at 2023-12-0013-00-10.



Transforming Data in Trifecta

The screenshot illustrates the Trifecta interface for data transformation, showing three panels: the Data Editor, the Transform Editor, and the Preview panel.

Data Editor: This panel displays a table with 4,270 rows and 13 columns. The columns include various metrics such as COUNT, DURATION, and CORRELATION. A sidebar on the left lists the columns with their current types: COUNT (String), DURATION (String), and CORRELATION (String). The table header shows column names like COUNT, DURATION, and CORRELATION.

Transform Editor: This panel contains a series of transformation steps connected by arrows. The steps include:

- Step 1: COUNT → COUNT (String) → COUNT (Int)
- Step 2: DURATION → DURATION (String) → DURATION (Int)
- Step 3: CORRELATION → CORRELATION (String) → CORRELATION (Int)
- Step 4: COUNT → COUNT (Int) → COUNT (String)
- Step 5: DURATION → DURATION (Int) → DURATION (String)
- Step 6: CORRELATION → CORRELATION (Int) → CORRELATION (String)

Preview Panel: This panel shows the state of the data after the transformations. It includes four tables labeled "Input 1", "Input 2", "Output 1", and "Output 2".

Table	Column	Value
Input 1	COUNT	1
	DURATION	1
	CORRELATION	1
	COUNT	1
Input 2	COUNT	1
	DURATION	1
	CORRELATION	1
	COUNT	1
Output 1	COUNT	1
	DURATION	1
	CORRELATION	1
	COUNT	1
Output 2	COUNT	1
	DURATION	1
	CORRELATION	1
	COUNT	1



Predictive Interaction and Immediate Feedback

CDR

New Sample 1 - First 500kB 13 Columns 4,273 Rows 6 Data Types Grid

Sort: Default Edit

#	IMEI	Source	to be dropped	Preview	#	DURATION
	33.02T - 52.03T	Dec 18 2013 00:00 - Jan 1 2014		Dec 18 2013 - Jan 1 2014	00:00 - 23:00	0 - 451
1	35297700840926	2013-12-19T18:49:19		2013-12-19	8:49:19	52
2	35643301870776	2013-12-30T1:55:42		2013-12-30	1:55:42	327
3	35643301870776	2013-12-18T23:29:32		2013-12-18	23:29:32	284
4	35643301870776	2013-12-20T19:52:25		2013-12-20	19:52:25	326
5	35643301870776	2013-12-30T10:02:48		2013-12-30	10:02:48	281
6	35643301870776	2013-12-24T7:50:56		2013-12-24	7:50:56	302
7	35643301870776	2013-12-18T5:57:43		2013-12-18	5:57:43	328
8	35643301870776	2013-12-28T22:42:20		2013-12-28	22:42:20	328
9	35643301870776	2013-12-27T0:41:24		2013-12-27	0:41:24	274
10	35643301870776	2013-12-18T5:33:57		2013-12-18	5:33:57	290
11	35643301870776	2013-12-24T13:40:03		2013-12-24	13:40:03	307
12	35643301870776	2013-12-21T23:36:20		2013-12-21	23:36:20	303
13	35643301870776	2013-12-30T25:50:25		2013-12-30	25:50:25	292
14	35643301870776	2013-12-30T21:08:10		2013-12-30	21:08:10	315

1 mismatched value



Predictive Interaction

CDR

New Sample 1 - First 500kB 13 Columns 4,273 Rows 6 Data Types Grid

Sort: Default Edit

#	IMEI	Source	to be dropped	Preview	#	DURATION
#	IMEI	DATETIME	DATETIME1	DATETIME2	#	DURATION
1	35297700840926	Dec 18 2013 00:00 - Jan 1 2014	2013-12-18T00:49:19	2013-12-19	0:49:19	52
2	35643301870776	2013-12-30T155:42	2013-12-30	1:55:42	327	
3	35643301870776	2013-12-18T13:29:32	2013-12-18	23:29:32	284	
4	35643301870776	2013-12-20T19:52:25	2013-12-20	19:52:25	326	
5	35643301870776	2013-12-30T10:02:48	2013-12-30	10:02:48	281	
6	35643301870776	2013-12-24T7:50:56	2013-12-24	7:50:56	302	
7	35643301870776	2013-12-18T5:57:43	2013-12-18	5:57:43	328	
8	35643301870776	2013-12-28T22:42:20	2013-12-28	22:42:20	328	
9	35643301870776	2013-12-27T0:41:24	2013-12-27	0:41:24	274	
10	35643301870776	2013-12-18T5:33:57	2013-12-18	5:33:57	290	
11	35643301870776	2013-12-24T13:40:03	2013-12-24	13:40:03	307	
12	35643301870776	2013-12-21T23:36:20	2013-12-21	23:36:20	303	
13	35643301870776	2013-12-30T25:50:25	2013-12-30	25:50:25	292	
14	35643301870776	2013-12-30T21:08:10	2013-12-30	21:08:10	315	

1 mismatched value



Immediate Feedback

CDR

New Sample 1 - First 500kB 13 Columns 4,273 Rows 6 Data Types Grid

Sort: Default Edit

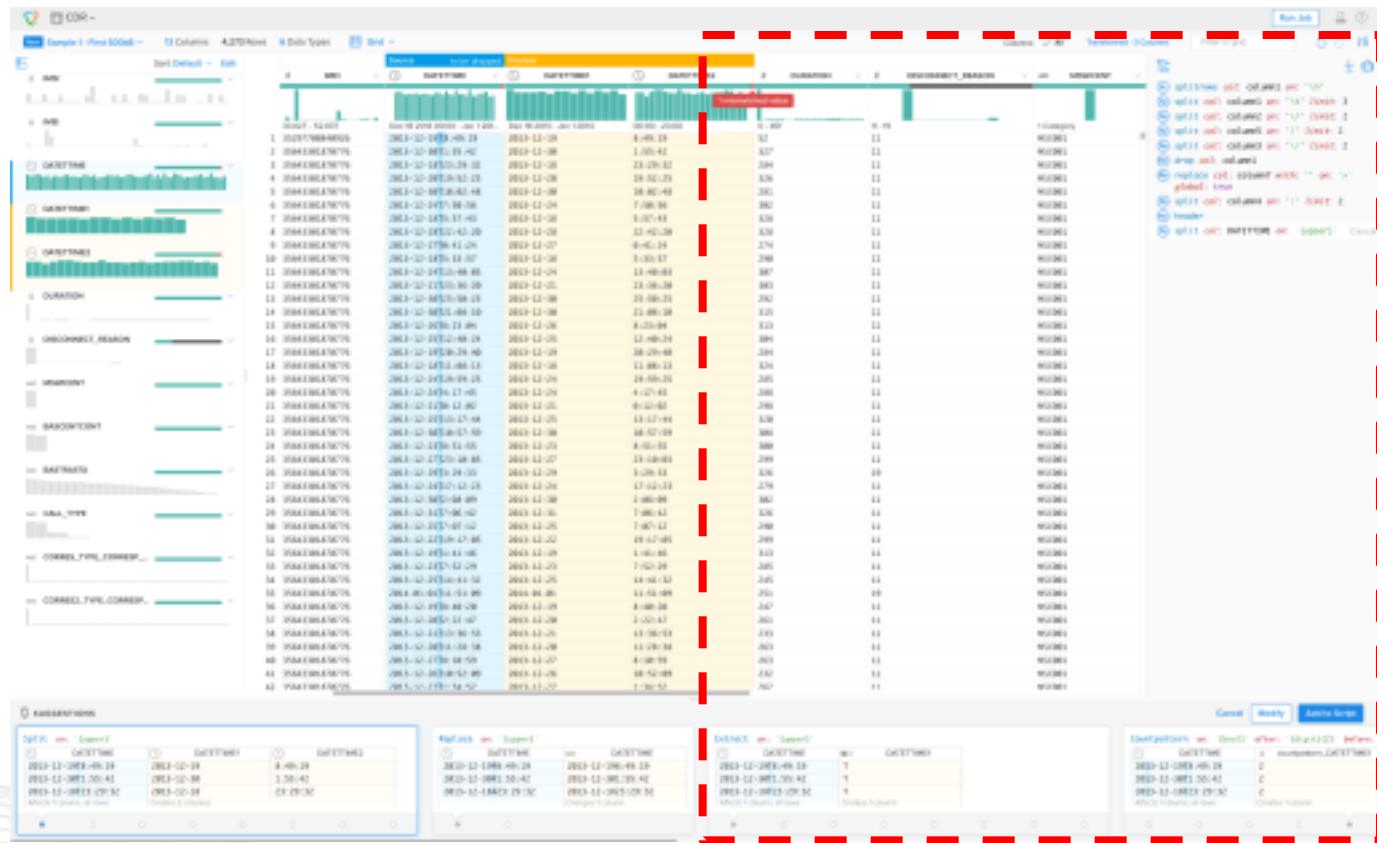
#	IMEI	Source	to be dropped	Preview	DATETIME	DATETIME1	DATETIME2	#	DURATION
	33.02T - 52.03T	Dec 18 2013 00:00 - Jan 1 2014							0 - 451
1	35297700840926	2013-12-19T18:49:19			2013-12-19		00:00 - 23:00		52
2	35643301870776	2013-12-30T1:55:42			2013-12-30		1:55:42		327
3	35643301870776	2013-12-18T23:29:32			2013-12-18		23:29:32		284
4	35643301870776	2013-12-20T19:52:25			2013-12-20		19:52:25		326
5	35643301870776	2013-12-30T10:02:48			2013-12-30		10:02:48		281
6	35643301870776	2013-12-24T7:50:56			2013-12-24		7:50:56		302
7	35643301870776	2013-12-18T5:57:43			2013-12-18		5:57:43		328
8	35643301870776	2013-12-28T22:42:20			2013-12-28		22:42:20		328
9	35643301870776	2013-12-27T0:41:24			2013-12-27		0:41:24		274
10	35643301870776	2013-12-18T5:33:57			2013-12-18		5:33:57		290
11	35643301870776	2013-12-24T13:40:03			2013-12-24		13:40:03		307
12	35643301870776	2013-12-21T23:36:20			2013-12-21		23:36:20		303
13	35643301870776	2013-12-30T25:50:25			2013-12-30		25:50:25		292
14	35643301870776	2013-12-30T21:08:10			2013-12-30		21:08:10		315

1 mismatched value

Transforming Data in Trifacta



Transforming Data in Trifacta



DURATION	#	DISCONNECT_REASON	ABC	MSWICENT	ABC	BASCENTCONT
11-19				1 Category		2 Categories
11				MSC001		BSC001
11				MSC001		BSC001
11				MSC001		BSC002
11				MSC001		BSC002
11				MSC001		BSC002
11				MSC001		BSC002
11				MSC001		BSC001
11				MSC001		BSC002
11				MSC001		BSC001
11				MSC001		BSC001
11				MSC001		BSC001
11				MSC001		RSC001

Split on: '{upper}'

DATETIME	DATETIME1	DATETIME2
2013-12-19T8:49:19	2013-12-19	8:49:19
2013-12-30T1:55:42	2013-12-30	1:55:42
2013-12-18T23:29:32	2013-12-18	23:29:32

Affects 1 column, all rows Creates 2 columns

Extract on: '{upper}'

DATETIME	DATETIME1
2013-12-19T8:49:19	T
2013-12-30T1:55:42	T
2013-12-18T23:29:32	T

Affects 1 column, all rows Creates 1 column

Save Cancel Modify Add to Script

```

Sr splitrows col: column1 on: '\n'
Sp split col: column1 on: '\t' limit: 3
Sp split col: column2 on: '\v' limit: 2
Sp split col: column4 on: ';' limit: 2
Rp replace col: column7 with: '' on: '>' global: true
Rp replace col: column1 with: '' on: '<' global: true
Sp split col: column3 on: '\v' limit: 2
He header
Sp split col: DATETIME on: '{upper}'

```

DURATION	#	DISCONNECT_REASON	ABC	MSWICENT	ABC	BASCENTCONT
11-19			1 Category		2 Categories	
11			MSC001		BSC001	
11			MSC001		BSC001	
11			MSC001		BSC002	
11			MSC001		BSC002	
11			MSC001		BSC002	
11			MSC001		BSC002	
11			MSC001		BSC001	
11			MSC001		BSC002	
11			MSC001		BSC001	
11			MSC001		BSC001	
11			MSC001		BSC001	
11			MSC001		BSC001	
11			MSC001		BSC001	
11			MSC001		RSC001	

Split on: '{upper}'

DATETIME	DATETIME1	DATETIME2
2013-12-19T8:49:19	2013-12-19	8:49:19
2013-12-30T1:55:42	2013-12-30	1:55:42
2013-12-18T23:29:32	2013-12-18	23:29:32

Affects 1 column, all rows

Extract on: '{upper}'

DATETIME	DATETIME1
2013-12-19T8:49:19	T
2013-12-30T1:55:42	T
2013-12-18T23:29:32	T

Affects 1 column, all rows

Cancel Modify Add to Script

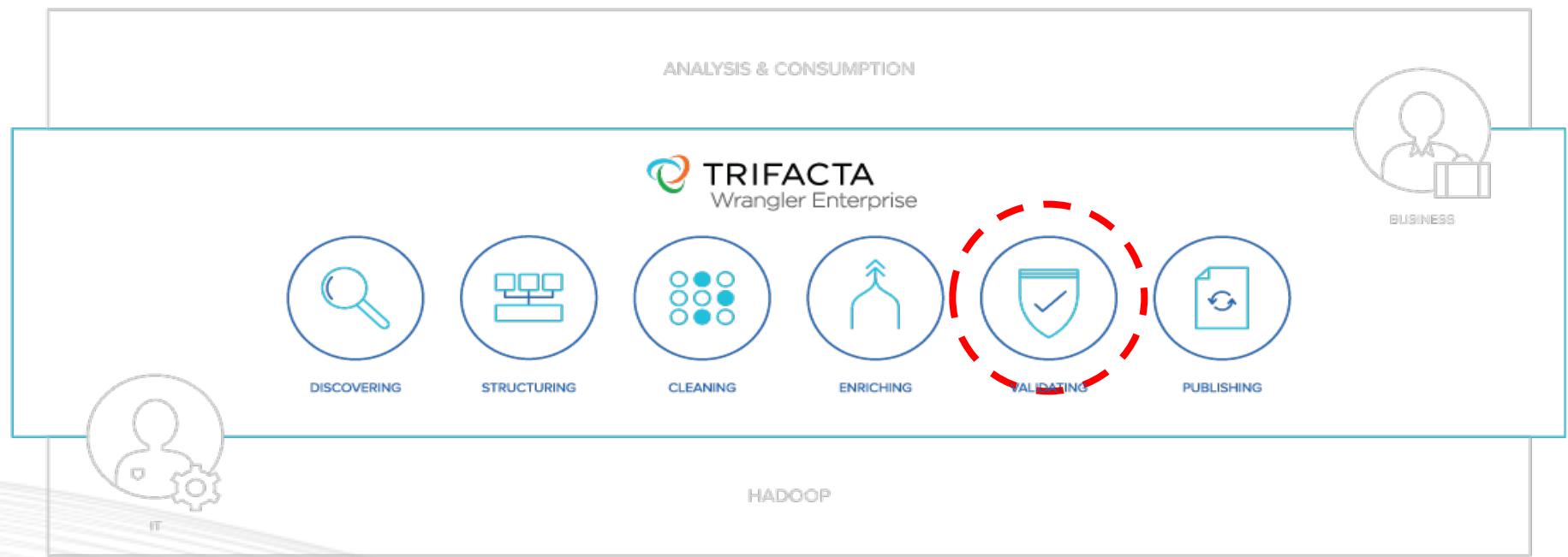
E

```

Sr splitrows col: column1 on: '\n'
Sp split col: column1 on: '\t' limit: 3
Sp split col: column2 on: '\v' limit: 2
Sp split col: column4 on: ';' limit: 2
Rp replace col: column7 with: '' on: '>' global: true
Rp replace col: column1 with: '' on: '<' global: true
Sp split col: column3 on: '\v' limit: 2
He header
Sp split col: DATETIME on: '{upper}'
```

Where Profiling Fits In

We validate preparation via *profiling*

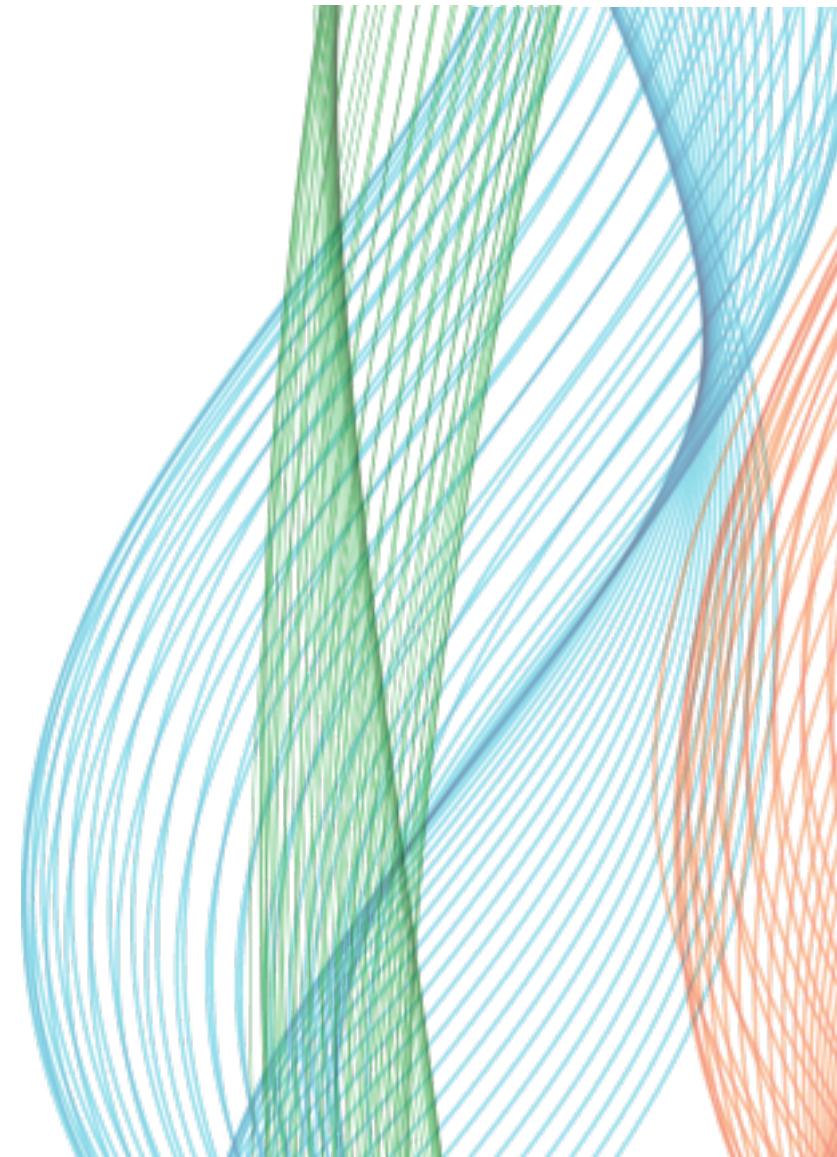


Spark Profiling at Trifecta

- Profiling results of transformation at scale
 - Validation through profiles
- Challenges
 - Scale
 - Automatic job generation
- Our solution
 - Spark profiling job server
 - JSON spec
 - Pay-as-you-go



The Case for Profiling



Transforming Data in Trifacta



Job Results

- Even clean raw data is not informative.
 - Especially when it is too large to inspect manually.
- Need a summary representation.
 - Statistical and visual
- Generate profile programmatically.

HDFS > trifacta > queryResults > SUBSCRIBERS			
	NAME	SIZE	LAST UPDATED
	..		
+	_SUCCESS		Today at 1:51 PM
+	part-00000	104.45MB	Today at 1:51 PM
+	part-00001	104.45MB	Today at 1:51 PM
+	part-00002	104.45MB	Today at 1:51 PM
+	part-00003	104.45MB	Today at 1:51 PM
+	part-00004	104.45MB	Today at 1:51 PM
+	part-00005	104.45MB	Today at 1:51 PM



Visual Result Validation

The screenshot shows a user interface for visual result validation. At the top, there's a header with a bar chart icon and the text "SUBSCRIBERS". Below the header, the text "Last updated: Today at 1:28 PM" and "Created: Today at 1:28 PM" is displayed. Underneath, there's a section labeled "Jobs" with four tabs: "All" (which is selected), "Complete", "Failed", and "Running". A detailed job summary is shown in a box:

- JOB ID: 6
- 1 Datasource
- Completed 01:33pm Jun 6
- Transform Job finished.

At the bottom of this summary box are two small icons: a download arrow and a clipboard.

- Similar to the profiling we saw above the data grid.
- Applied at scale, to the full result.
- Reveals mismatched, missing values.



Visual Result Validation

SUBSCRIBERS

Last updated: Today at 1:28 PM Created: Today at 1:28 PM

Jobs [All](#) [Complete](#) [Failed](#) [Running](#)

JOB ID: 6	1 Datasource
Completed	01:33pm Jun 6
Transform Job finished.	
View Results	

JOB ID: 5	1 Datasource
Completed	01:32pm Jun 6
<div style="background-color: #2e7131; width: 91%; height: 10px;"></div> 91% <div style="background-color: #dc3545; width: 3%; height: 10px;"></div> 3% <div style="background-color: #6c757d; width: 7%; height: 10px;"></div> 7%	
Valid Mismatched Missing	
View Results	

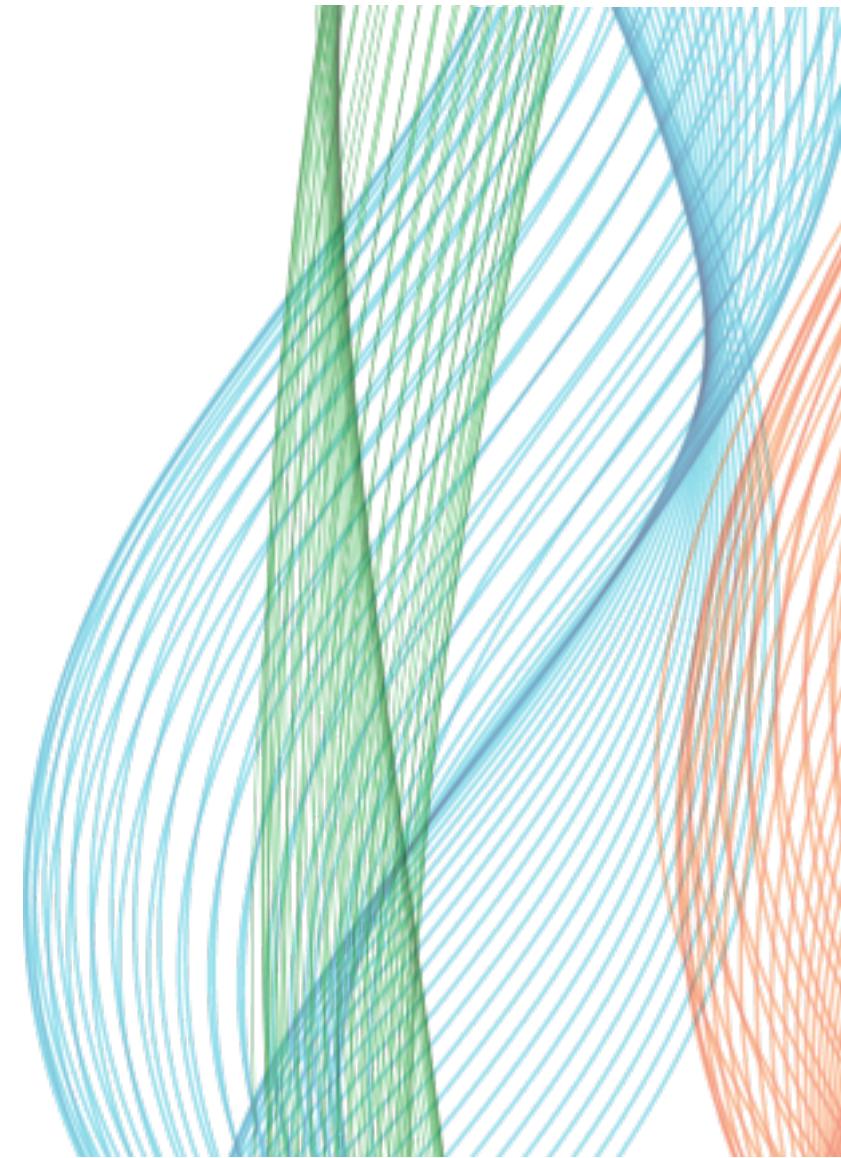
- Similar to the profiling we saw above the data grid.
- Applied at scale, to the full result.
- Reveals mismatched, missing values.



Visual Result Validation

#	IMSI	🕒	CONTRACT_END	@	EMAIL	ABC	OCCUPATION																																																						
Valid	3,531	Valid	3,531	Valid	3,009	Valid	1,648																																																						
Mismatched	0	Mismatched	0	Mismatched	522	Mismatched	0																																																						
Empty	107	Empty	107	Empty	107	Empty	1,990																																																						
 <p>208.10T 208.11T 208.12T</p>		 <p>2009 2012 2016</p>		<p>Top 20 values</p> <table> <tbody> <tr><td>ipsum.leo@felis.co.uk</td><td>10</td></tr> <tr><td>montes@orci.co.uk</td><td>10</td></tr> <tr><td>cubilia.Curae.Donec@N...</td><td>10</td></tr> <tr><td>eros.non.enim@Curabit...</td><td>10</td></tr> <tr><td>enim@magna.com</td><td>10</td></tr> <tr><td>posuere.vulputate.lac...</td><td>10</td></tr> <tr><td>facilisis.facilisis@i...</td><td>9</td></tr> <tr><td>rutrum@tellusfaucibus...</td><td>9</td></tr> <tr><td>Vivamus@quisdiam.net</td><td>9</td></tr> <tr><td>biglia@gmail.com</td><td>9</td></tr> <tr><td>fringilla@condimentum...</td><td>9</td></tr> <tr><td>non@sitametrisus.edu</td><td>9</td></tr> <tr><td>condimentum@Morbisita...</td><td>9</td></tr> <tr><td>In@Donecnibhenim.net</td><td>9</td></tr> <tr><td>eu@nonfeugiat.net</td><td>9</td></tr> <tr><td>eu@nequeseddictum.com</td><td>9</td></tr> <tr><td>sociis.natoque@ipsuml...</td><td>9</td></tr> <tr><td>variis@erat.com</td><td>9</td></tr> <tr><td>purus.sapien@adipisci...</td><td>9</td></tr> <tr><td>adipiscing@lacusUtne...</td><td>9</td></tr> </tbody> </table>		ipsum.leo@felis.co.uk	10	montes@orci.co.uk	10	cubilia.Curae.Donec@N...	10	eros.non.enim@Curabit...	10	enim@magna.com	10	posuere.vulputate.lac...	10	facilisis.facilisis@i...	9	rutrum@tellusfaucibus...	9	Vivamus@quisdiam.net	9	biglia@gmail.com	9	fringilla@condimentum...	9	non@sitametrisus.edu	9	condimentum@Morbisita...	9	In@Donecnibhenim.net	9	eu@nonfeugiat.net	9	eu@nequeseddictum.com	9	sociis.natoque@ipsuml...	9	variis@erat.com	9	purus.sapien@adipisci...	9	adipiscing@lacusUtne...	9	<p>Top 7 values</p> <table> <tbody> <tr><td>Unemployed</td><td>380</td></tr> <tr><td>Director</td><td>229</td></tr> <tr><td>Software Developer</td><td>229</td></tr> <tr><td>Employee</td><td>227</td></tr> <tr><td>Salesman</td><td>227</td></tr> <tr><td>Architect</td><td>203</td></tr> <tr><td>Administrator</td><td>153</td></tr> </tbody> </table>		Unemployed	380	Director	229	Software Developer	229	Employee	227	Salesman	227	Architect	203	Administrator	153
ipsum.leo@felis.co.uk	10																																																												
montes@orci.co.uk	10																																																												
cubilia.Curae.Donec@N...	10																																																												
eros.non.enim@Curabit...	10																																																												
enim@magna.com	10																																																												
posuere.vulputate.lac...	10																																																												
facilisis.facilisis@i...	9																																																												
rutrum@tellusfaucibus...	9																																																												
Vivamus@quisdiam.net	9																																																												
biglia@gmail.com	9																																																												
fringilla@condimentum...	9																																																												
non@sitametrisus.edu	9																																																												
condimentum@Morbisita...	9																																																												
In@Donecnibhenim.net	9																																																												
eu@nonfeugiat.net	9																																																												
eu@nequeseddictum.com	9																																																												
sociis.natoque@ipsuml...	9																																																												
variis@erat.com	9																																																												
purus.sapien@adipisci...	9																																																												
adipiscing@lacusUtne...	9																																																												
Unemployed	380																																																												
Director	229																																																												
Software Developer	229																																																												
Employee	227																																																												
Salesman	227																																																												
Architect	203																																																												
Administrator	153																																																												
Minimum	208,100,112,262,...	Minimum	Nov 17 2009																																																										
Lower quartile	208,104,019,2...	Lower quartile	Feb 19 2014																																																										
Median	208,108,656,229,74...	Median	Oct 23 2014																																																										
Upper quartile	208,114,546,6...	Upper quartile	Apr 28 2015																																																										
Maximum	208,119,614,632,...	Maximum	Dec 29 2015																																																										

Challenges



The Goal: Profiling for Every Job

- We've talked about the value of profiling, but...
- Profiling is not a tablestakes feature, at least not on day 1.
 - Don't want our users to disable it!
- Profiling is potentially more expensive than transformation.



Profiling at Scale

- Large data volume
 - Long running times
 - Memory constraints



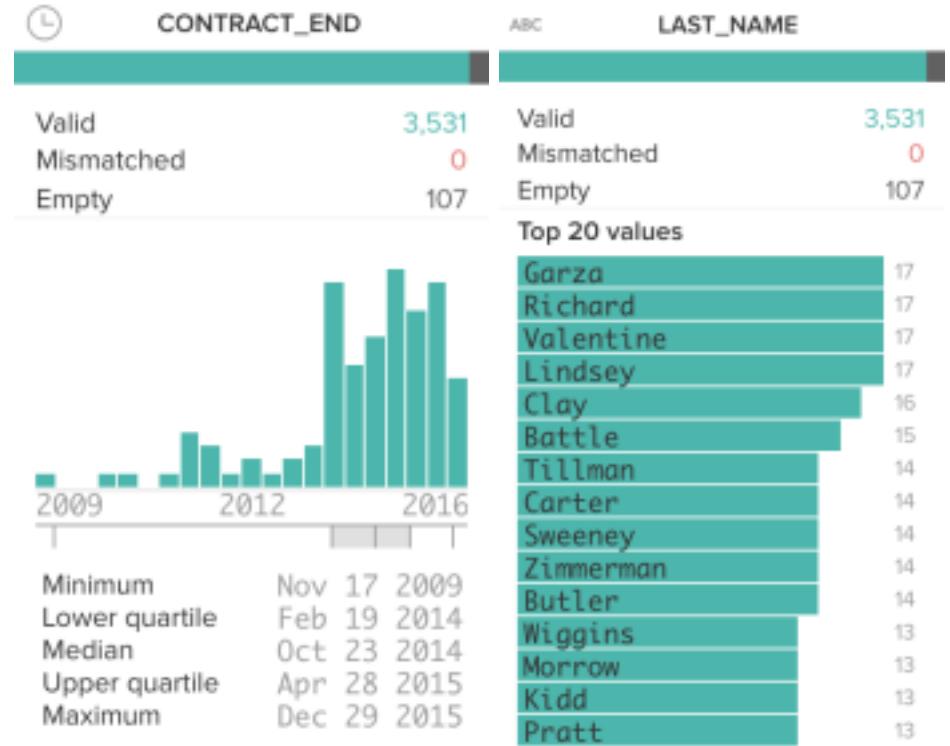
Performance vs. Accuracy

- Approximation brings performance gains while still meeting profiler summarization requirements.
- Off-the-shelf libraries (*count-min-sketch*, *T-digest*) great for approximating counts and non-distributive stats.
- Not a silver bullet though...sometimes approximations are confusing.



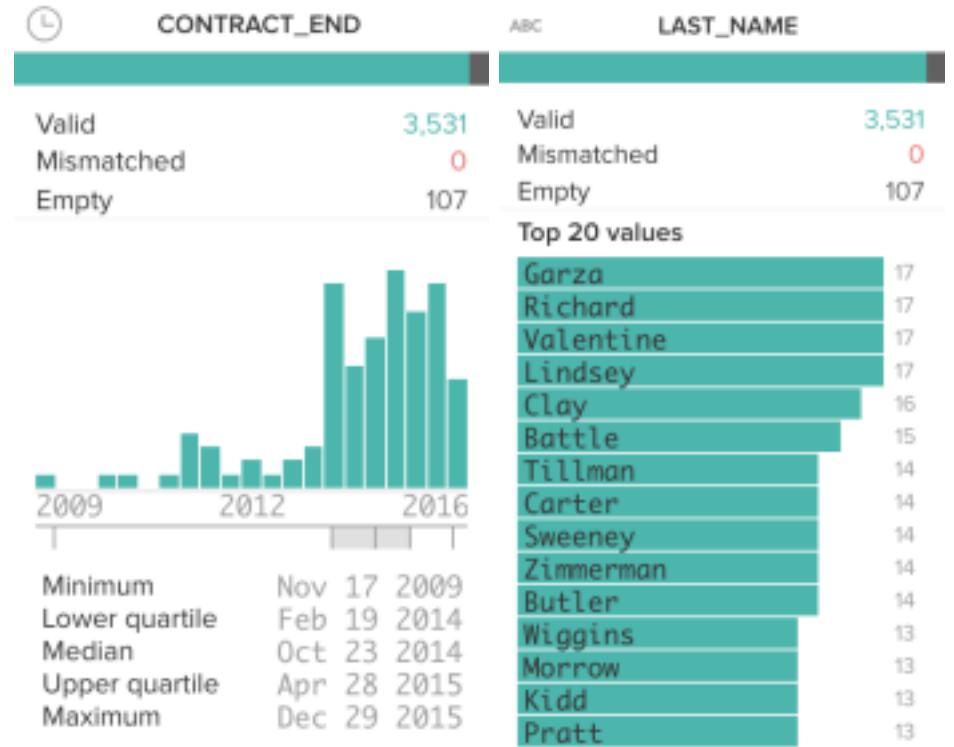
Flexible Job Generation

- Profile for all users, all use cases: **not** a one-off
 - Diverse schemas
 - Any number of columns
- Calculate statistics for all data types
 - Numeric vs. categorical
 - Container types (maps, arrays,...)
 - User-defined



Flexible Job Generation

- Profile for all users, all use cases: **not** a one-off
 - Diverse schemas
 - Any number of columns
- Calculate statistics for all data types
 - Numeric vs. categorical
 - Container types (maps, arrays,...)
 - User-defined

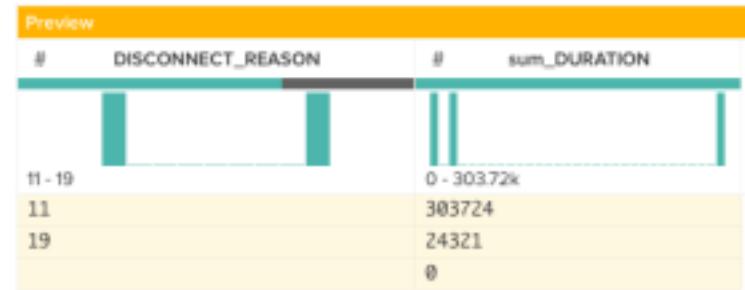
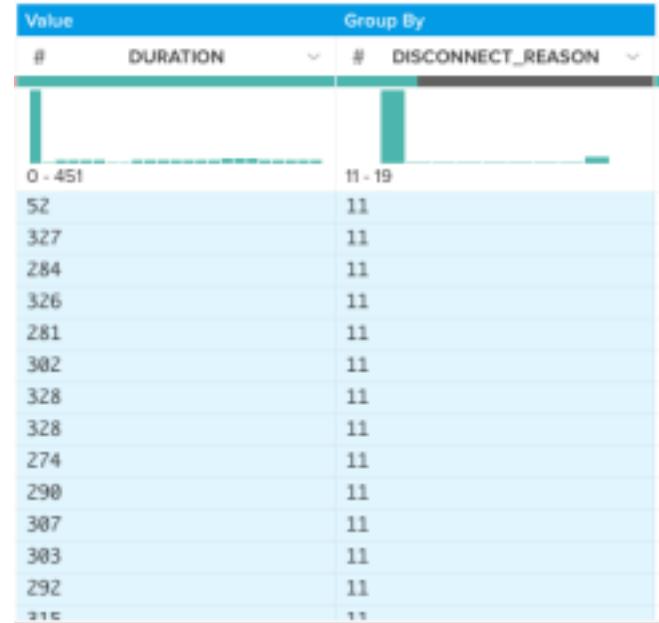


Transformation vs. Profiling

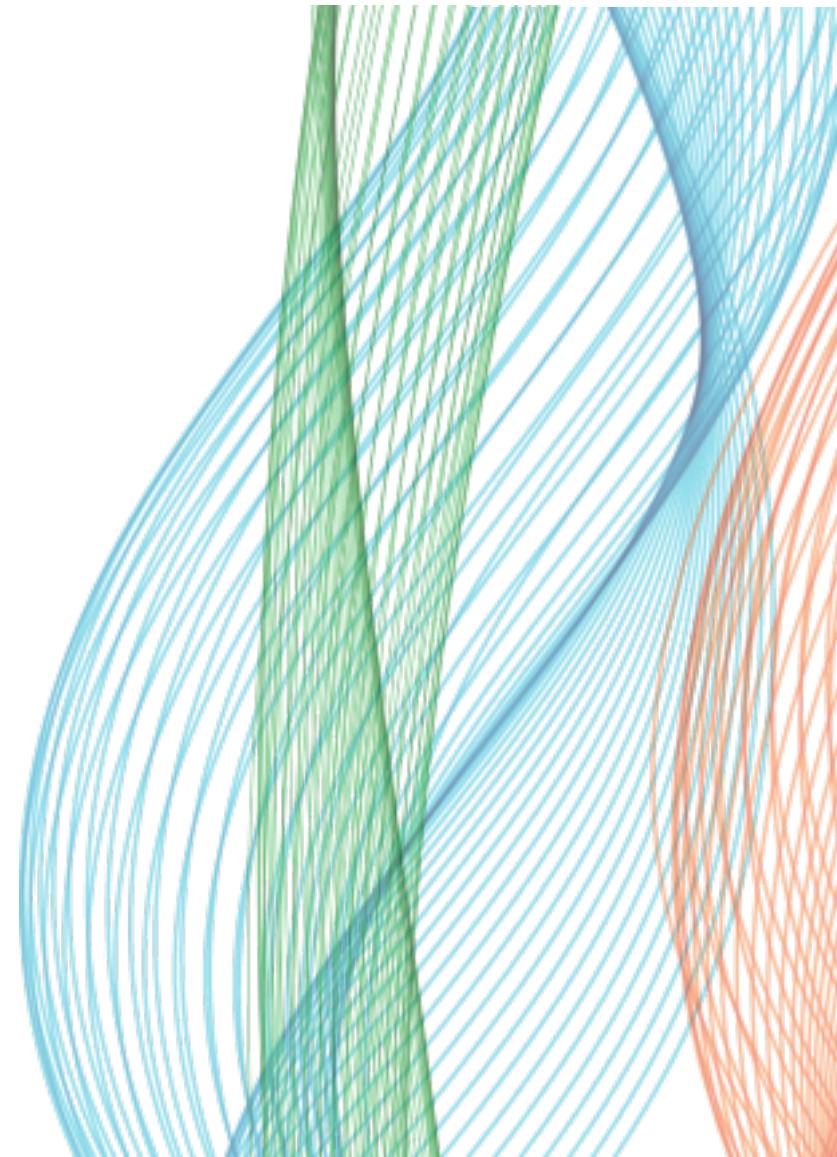
- Transformation is primarily row-based.
- Profiling is column-based.
- Different execution concerns.



Source	to be dropped	Preview
	DATETIME2	# DATETIME3
00:00 - 23:00	0 - 25	0 - 59
8:49:19	8	49
1:55:42	1	55
23:29:32	23	29
19:52:25	19	52
10:02:48	18	02
7:50:56	7	50
5:57:43	5	57
22:42:20	22	42



Solution: Spark in the Trifecta System



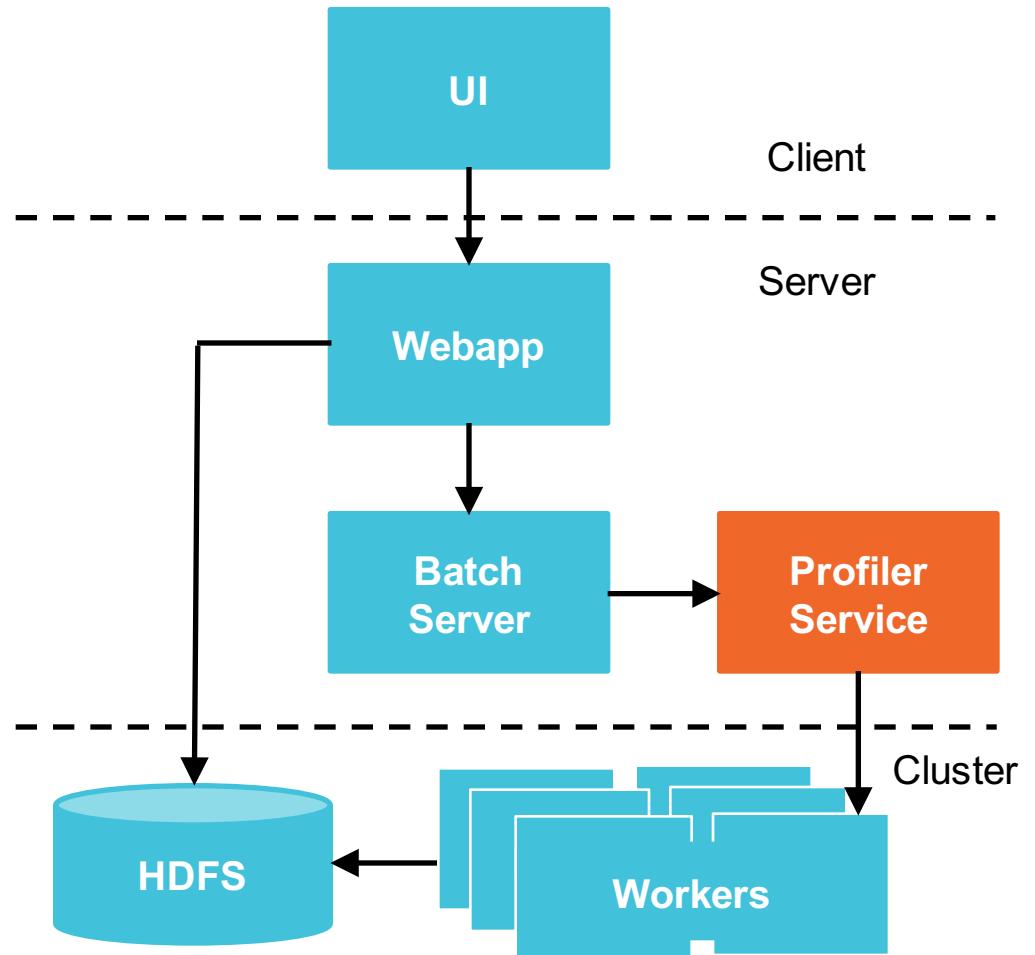
Why Spark

- Effective at OLAP
- Flexible enough for UDFs
- Not tied to distributions
- Mass adoption
- Low latency

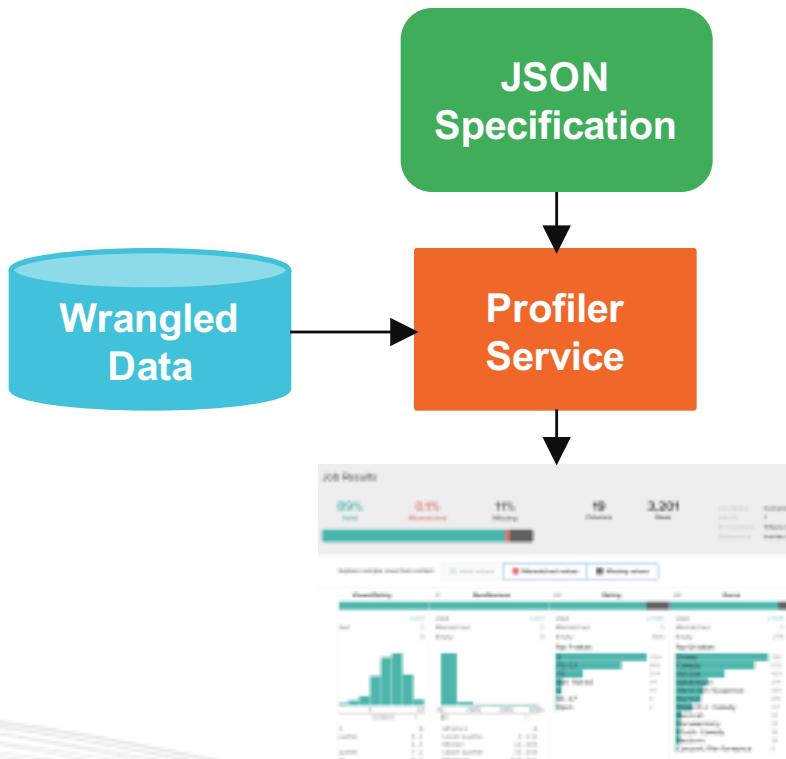


Spark Profile Jobs Server

- We built a Spark job server
- Automatically profiles output of transform jobs
- Outputs profiling results to HDFS
- Renders graphs in Trifecta product



Profiler Service



- Which columns to profile and their types
- Uses Trifecta type system, eg:
 - Dates
 - Geography
 - User-defined
- Requested profiling statistics
 - Histograms
 - Outliers
 - Empty, invalid, valid
- Extensible
 - Pairwise statistics
 - User-defined functions



Profiling DSL

```
{  
    "input": "hdfs://hadoop.trifecta-dev.net:8020:/trifecta...",  
    "schema": {  
        "order": ["column1", "column2", "column3"],  
        "types": {  
            "column1": ["Datetime", {"regexes": , "groupLocs": {...}}],  
            "column2": [...],  
            "column3": [...]  
        }  
    },  
    "commands": [  
        {  
            "column": "*",  
            "output": "hdfs://hadoop.trifecta-dev.net:8020:/trifecta...",  
            "profiler-type": "histogram",  
            "params": {...},  
            {  
                "column": "column1",  
                "output": "hdfs://hadoop.trifecta-dev.net:8020:/trifecta...",  
                "profiler-type": "type-check",  
                "params": {...}  
            }  
    ]  
}
```



Profiling DSL

```
{  
    "input": "hdfs://hadoop.trifecta-dev.net:8020:/trifecta...",  
    "schema": {  
        "order": ["column1", "column2", "column3"],  
        "types": {  
            "column1": ["Datetime", {"regexes": [...], "groupLocs": {...}}],  
            "column2": [...],  
            "column3": [...]  
        }  
    },  
    "commands": [  
        {  
            "column": "*",  
            "output": "hdfs://hadoop.trifecta-dev.net:8020:/trifecta...",  
            "profiler-type": "histogram",  
            "params": {...},  
            {  
                "column": "column1",  
                "output": "hdfs://hadoop.trifecta-dev.net:8020:/trifecta...",  
                "profiler-type": "type-check",  
                "params": {...}  
            }  
        }  
    ]  
}
```



Profiling DSL

```
{  
    "input": "hdfs://hadoop.trifecta-dev.net:8020:/trifecta...",  
    "schema": {  
        "order": ["column1", "column2", "column3"],  
        "types": {  
            "column1": ["Datetime", {"regexes": , "groupLocs": {...}}],  
            "column2": [...],  
            "column3": [...]  
        }  
    },  
    "commands": [  
        {  
            "column": "*",  
            "output": "hdfs://hadoop.trifecta-dev.net:8020:/trifecta...",  
            "profiler-type": "histogram",  
            "params": {...},  
            {  
                "column": "column1",  
                "output": "hdfs://hadoop.trifecta-dev.net:8020:/trifecta...",  
                "profiler-type": "type-check",  
                "params": {...}  
            }  
        }  
    ]  
}
```



Profiling DSL

```
{  
    "input": "hdfs://hadoop.trifecta-dev.net:8020:/trifecta...",  
    "schema": {  
        "order": ["column1", "column2", "column3"],  
        "types": {  
            "column1": ["Datetime", {"regexes": , "groupLocs": {...}}],  
            "column2": [...],  
            "column3": [...]  
        }  
    },  
    "commands": [  
        {  
            "column": "*",  
            "output": "hdfs://hadoop.trifecta-dev.net:8020:/trifecta...",  
            "profiler-type": "histogram",  
            "params": {...} },  
        {  
            "column": "column1",  
            "output": "hdfs://hadoop.trifecta-dev.net:8020:/trifecta...",  
            "profiler-type": "type-check",  
            "params": {...} }  
    ]  
}
```



Performance Improvements

Spark profiling speed-up vs. MapReduce

	Few Columns	Many Columns
High Cardinality	2X	4X
Low Cardinality	10X	20X

- A few troublesome use cases:
 - High cardinality
 - Non-distributive stats
- Huge gains for large numbers of columns



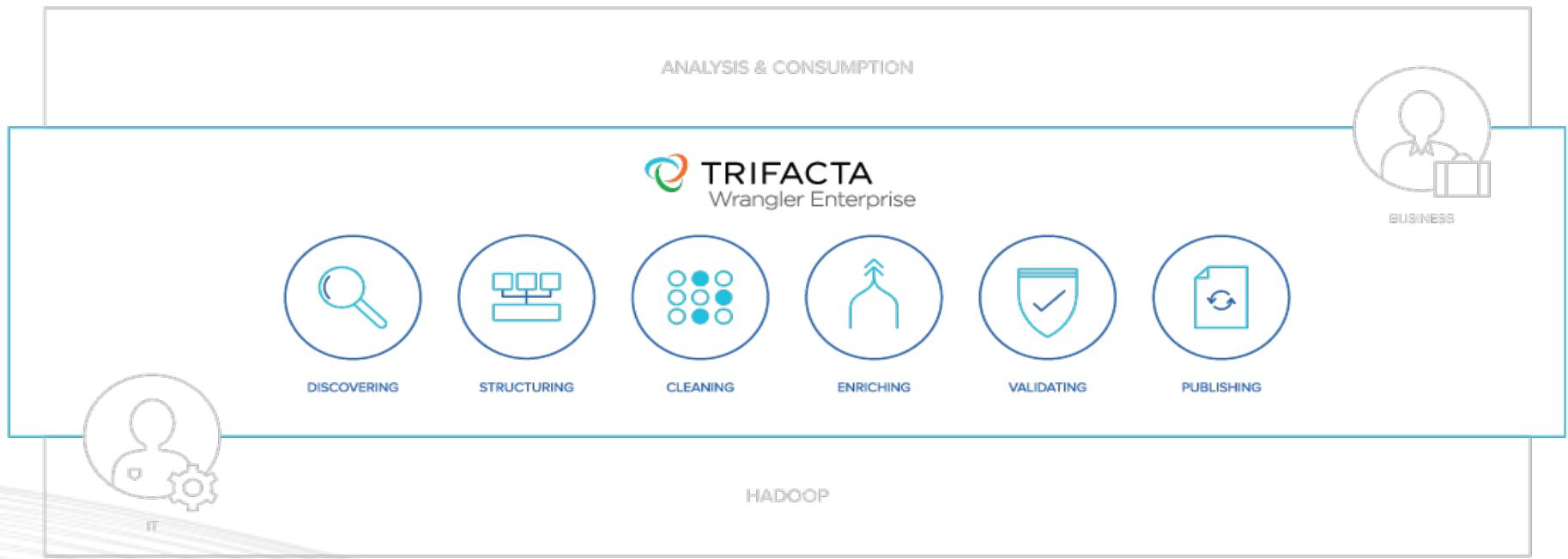
Pay-as-you-go Profiling

- Users do not always need profiles for all columns
- More complex and expensive statistics
- Drilldown
- Iterative exploration



Conclusion

Profiling informs data preparation.



Questions?

