Data Infrastructure:

Use Cases and Architecture

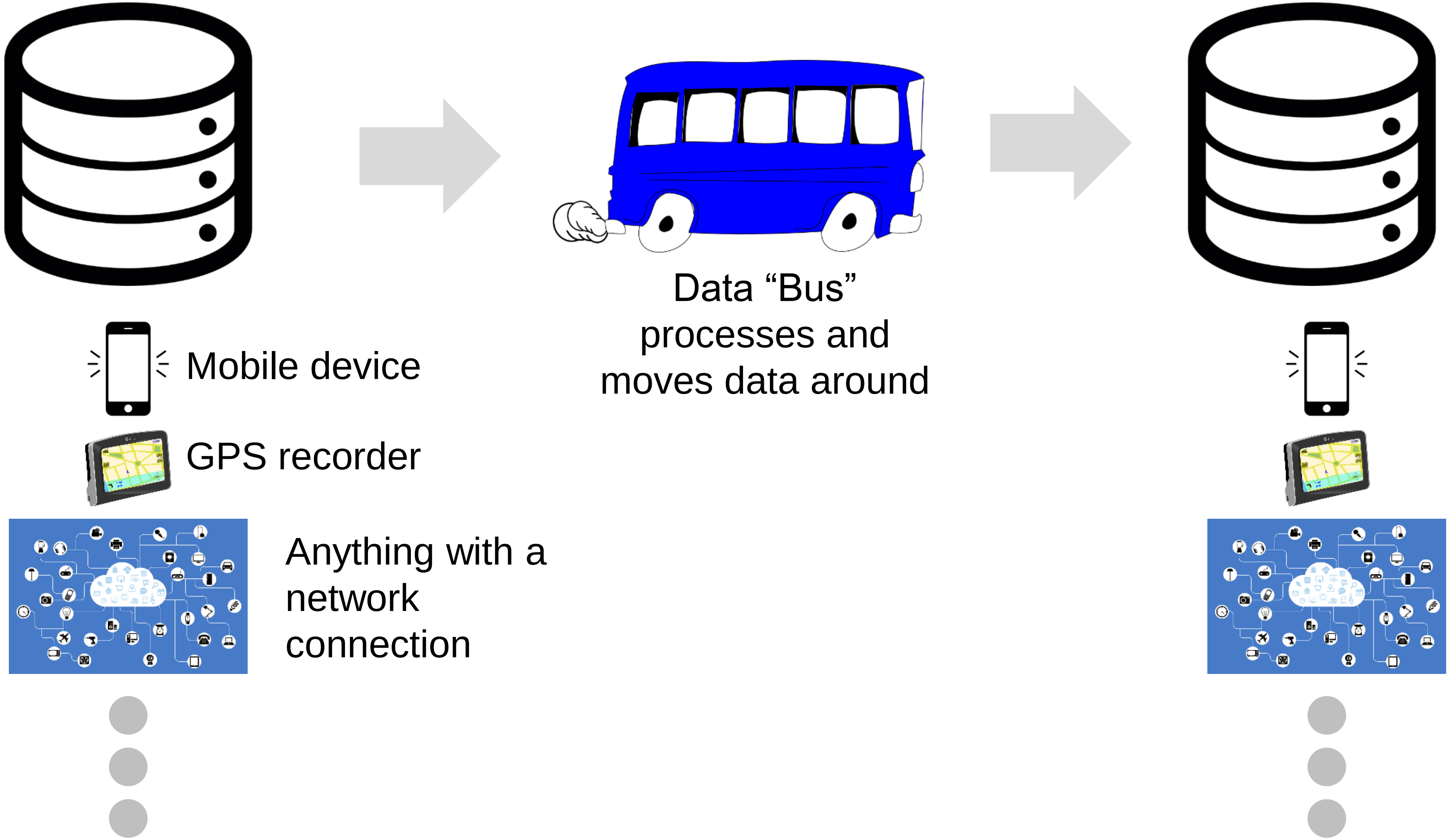
Joy Bonaguro
Chief Data Officer
City and County of San Francisco

COIT Subcommittee, Feb 2, 2018

Agenda

- Data infrastructure use cases
- Architectural choices (within and between)
- Benefits of a strategy

DI Use Case: Move and process data between systems



City database or application

Mobile device

GPS recorder

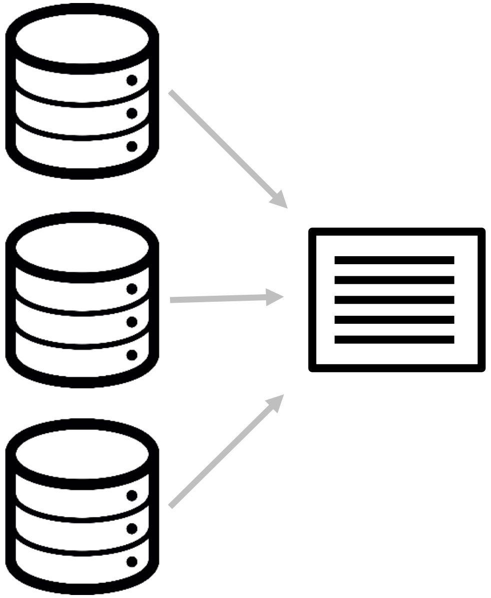
Anything with a network connection

Data “Bus” processes and moves data around

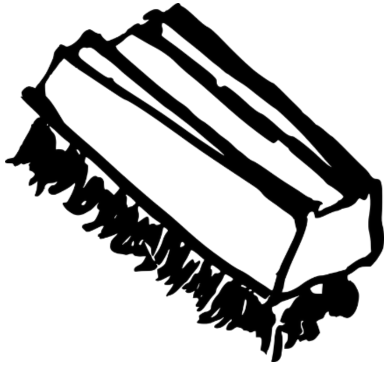
City database or application

DI Use Case: Store data for use

City databases
and applications



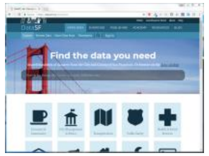
S1 Connect and extract data from source systems



S2 Clean and prepare data

Data warehouse

Data lake



OD portal

TIMESCALE Specialized stores

●
●
●

S3 Store data

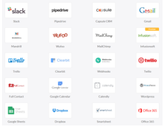
API

Table views

FTP drop

+ableau

BI tools



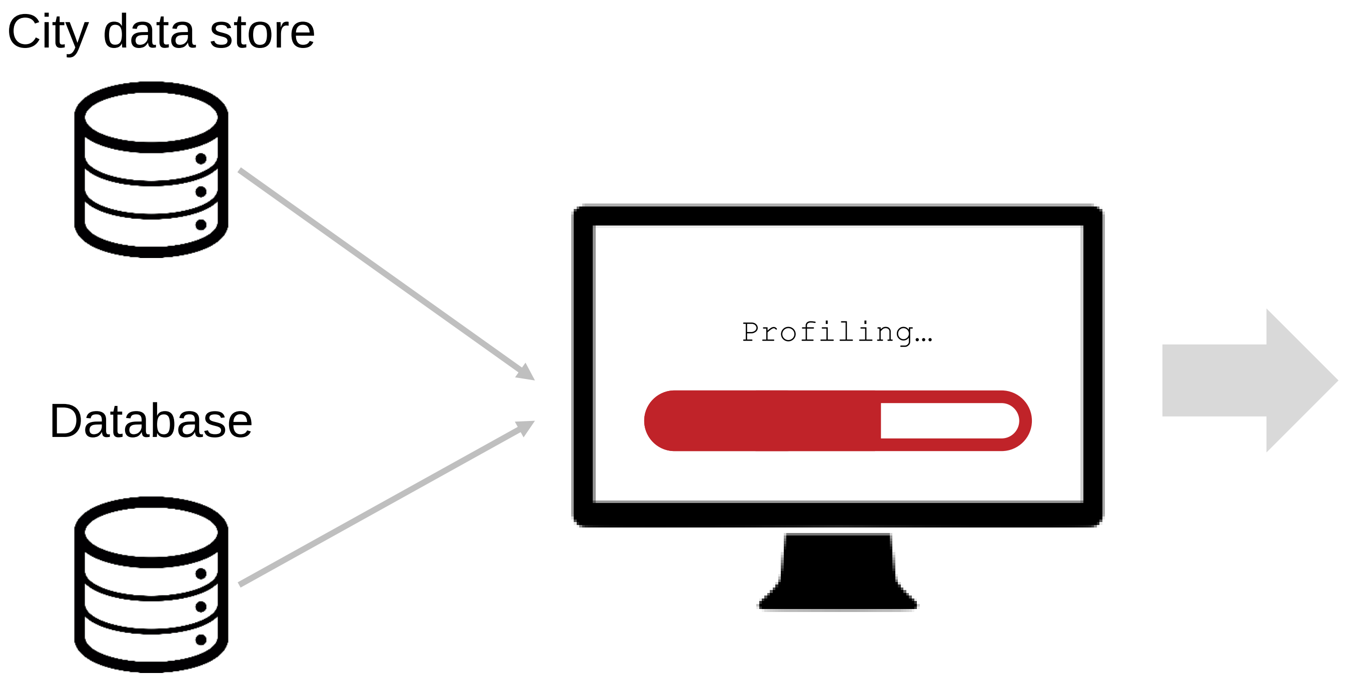
Web connectors

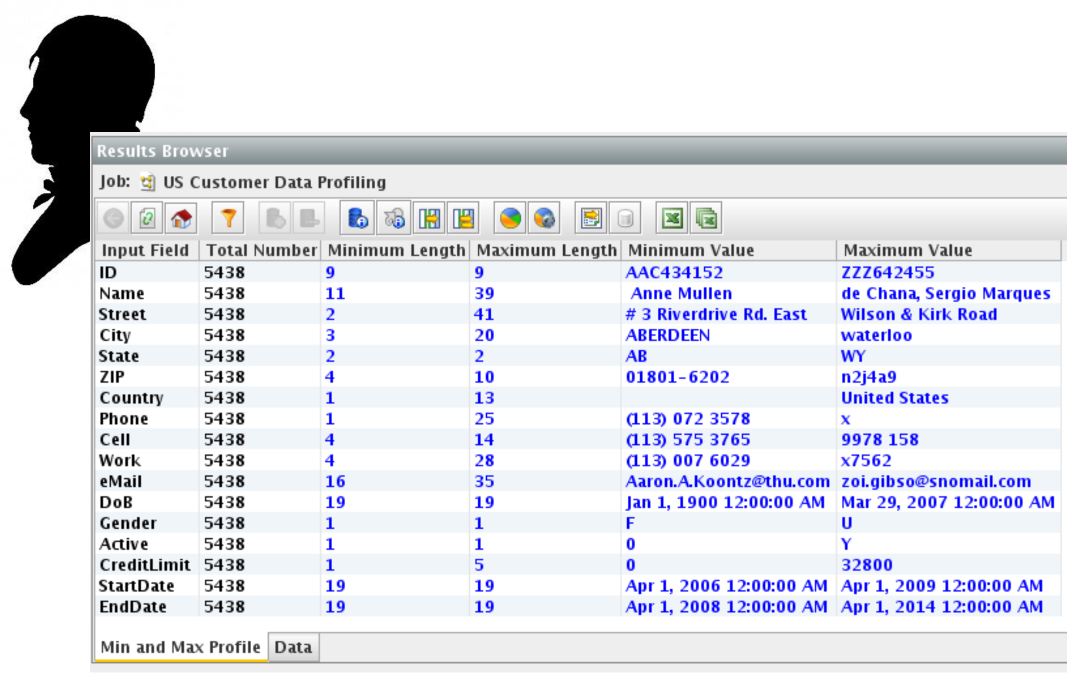
Data cube

●
●
●

S4 Provide access

DI Use Case: Monitor data quality



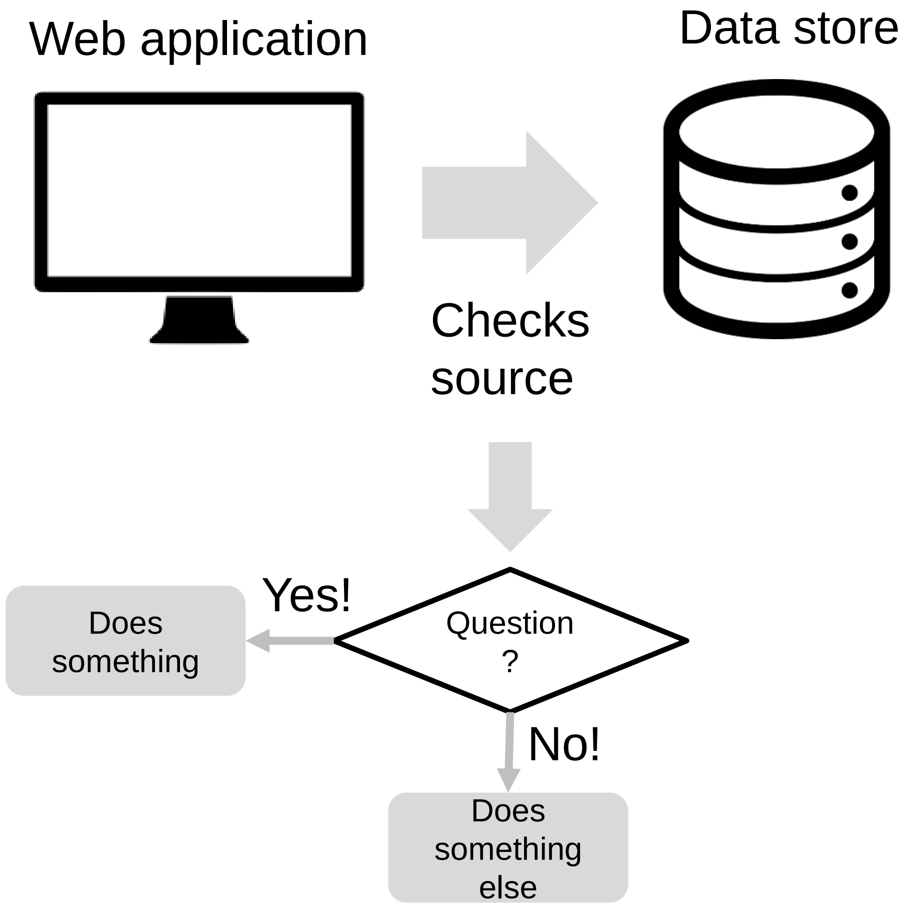


Job: US Customer Data Profiling

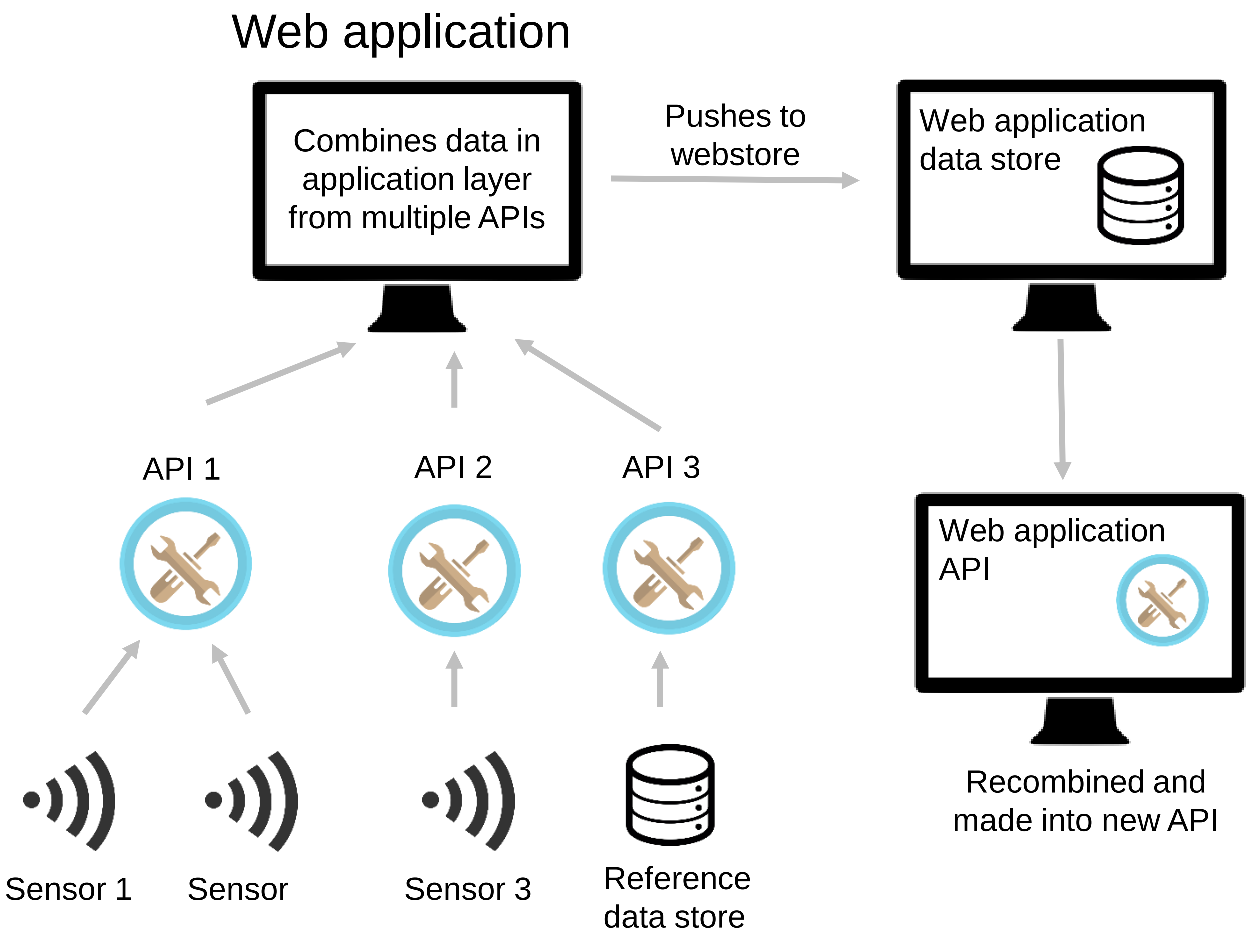
Input Field	Total Number	Minimum Length	Maximum Length	Minimum Value	Maximum Value
ID	5438	9	9	AAC434152	ZZZ642455
Name	5438	11	39	Anne Mullen	de Chana, Sergio Marques
Street	5438	2	41	# 3 Riverdrive Rd. East	Wilson & Kirk Road
City	5438	3	20	ABERDEEN	waterloo
State	5438	2	2	AB	WY
ZIP	5438	4	10	01801-6202	n2j4a9
Country	5438	1	13		United States
Phone	5438	1	25	(113) 072 3578	x
Cell	5438	4	14	(113) 575 3765	9978 158
Work	5438	4	28	(113) 007 6029	x7562
eMail	5438	16	35	Aaron.A.Koontz@thu.com	zoi.gibso@snomail.com
DoB	5438	19	19	Jan 1, 1900 12:00:00 AM	Mar 29, 2007 12:00:00 AM
Gender	5438	1	1	F	U
Active	5438	1	1	0	Y
CreditLimit	5438	1	5	0	32800
StartDate	5438	19	19	Apr 1, 2006 12:00:00 AM	Apr 1, 2009 12:00:00 AM
EndDate	5438	19	19	Apr 1, 2008 12:00:00 AM	Apr 1, 2014 12:00:00 AM

DI Use Case: Consume data from or check against a source

E1 Simple yes/no check



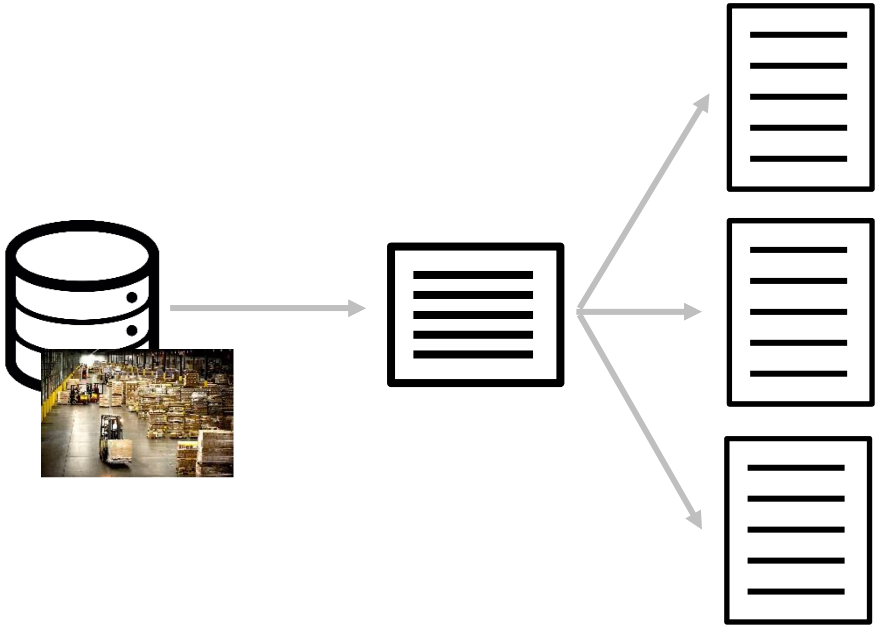
E2 Complex



Web application
Combines data in application layer from multiple APIs
Pushes to webstore
Web application data store
API 1
API 2
API 3
Web application API
Sensor 1
Sensor
Sensor 3
Reference data store
Recombined and made into new API

DI Use Case: Visualize data and KPIs, create dashboards

01 Centrally structured, codified, controlled, slow



Traditional data warehouse

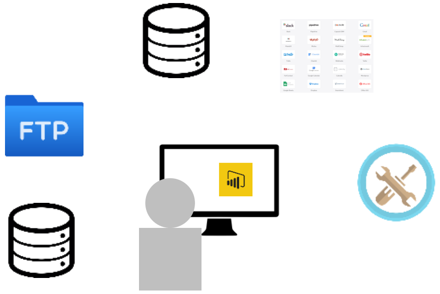
Structured, controlled queries, views

Fixed, standard reports

02 Self-service, dynamic, ad hoc, fast

1 external source

CSV on desktop



Anything and everything

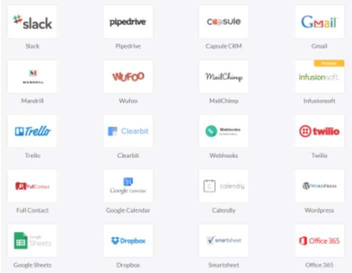
Locally trusted stores

Within each DI use case, there are architectural choices, at a minimum…

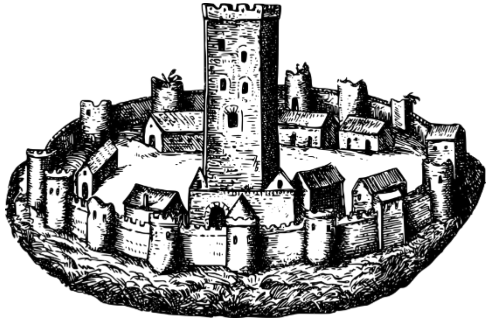
Warehoused vs virtualization

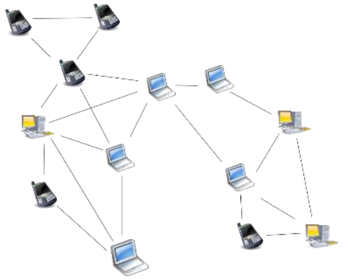
Streaming versus batch

Kitchen sink vs lightly coupled commodity components



Central versus distributed





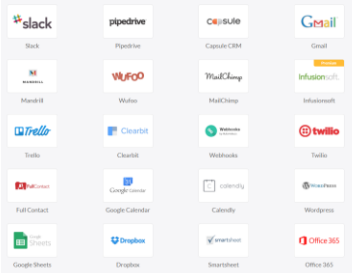
Low versus high volume

Between DI use cases, earlier choices can restrict downstream choices

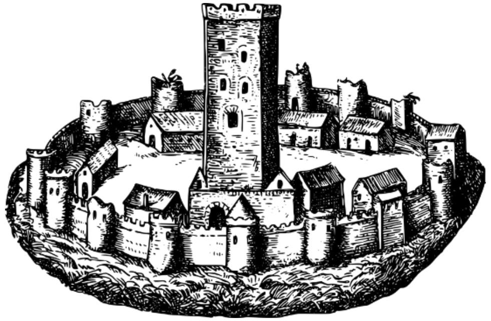
Warehoused vs virtualization

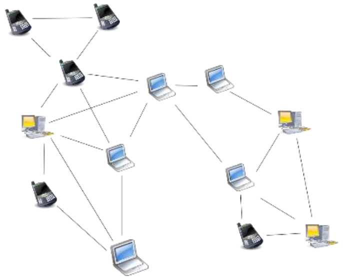
Streaming versus batch

Kitchen sink vs lightly coupled commodity components



Central versus distributed





Low versus high volume

Right now. We are making choices that will affect our downstream capacity and flexibility.

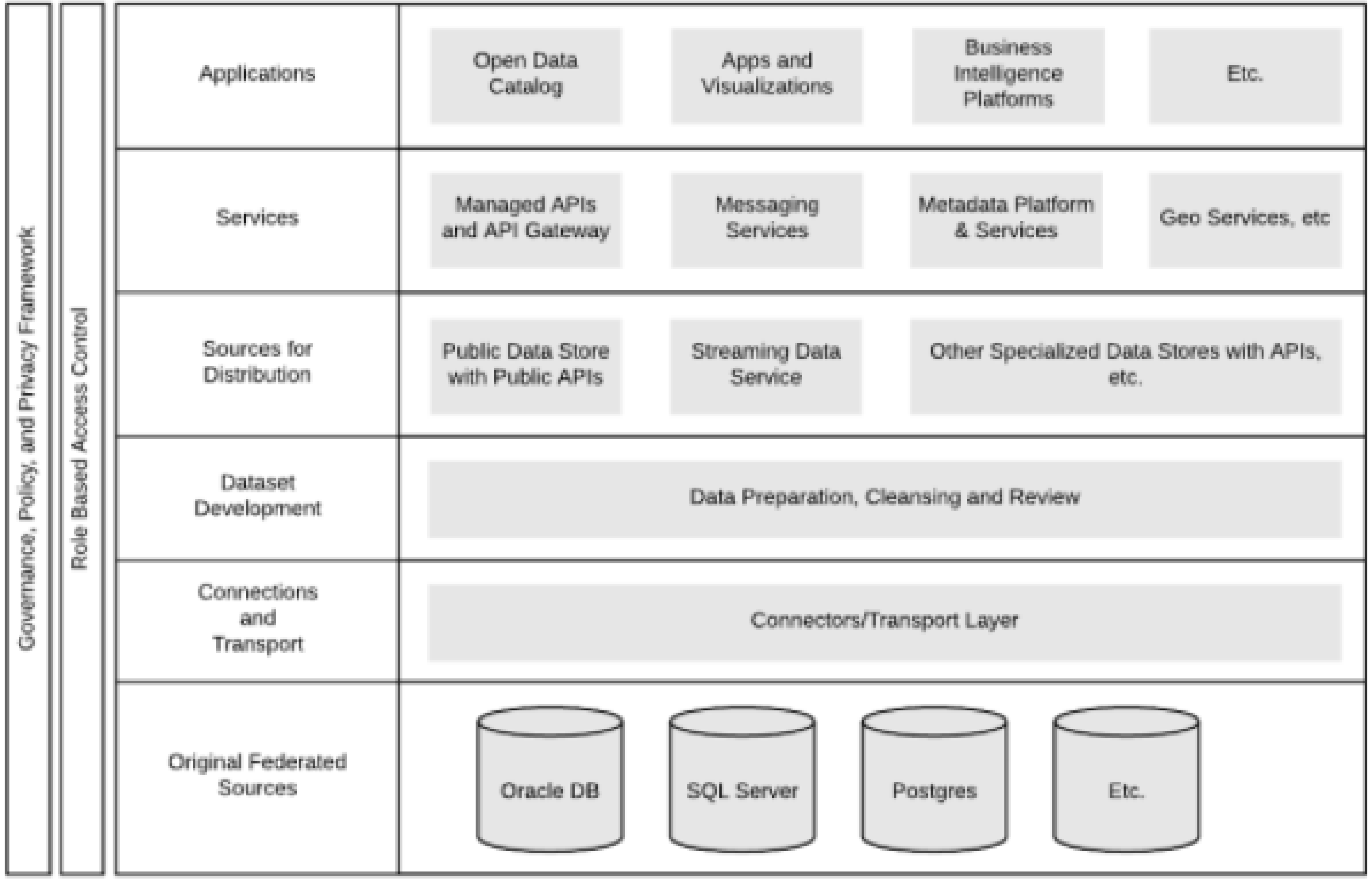
Why we need a conscious data infrastructure strategy

- Improvements in data consistency and quality
- Faster, easier access to data
- Better controls and security
- Data sharing and interoperability between datasets
- Integrated data across departments
- Faster development of digital and web services
- Data analytics and more advanced data science
- New and novel data services

Benefits to different groups

Audience	Benefit
Developers and IT staff	Decrease in technical and development time to create applications, integrations and services
COIT and department budget staff	Decrease in costs for applications and services
Analysts, data users, ShareSF committee	More time and resources for conducting analysis and evaluation → better services and outcomes
Voters, program staff, executives	Better decisions and services

A possible reference architecture



Governance, Policy, and Privacy Framework

Role Based Access Control

Applications	Open Data Catalog	Apps and Visualizations	Business Intelligence Platforms	Etc.
Services	Managed APIs and API Gateway	Messaging Services	Metadata Platform & Services	Geo Services, etc
Sources for Distribution	Public Data Store with Public APIs	Streaming Data Service	Other Specialized Data Stores with APIs, etc.	
Dataset Development	Data Preparation, Cleansing and Review			
Connections and Transport	Connectors/Transport Layer			
Original Federated Sources	Oracle DB	SQL Server	Postgres	Etc.

Data, for the love of the City

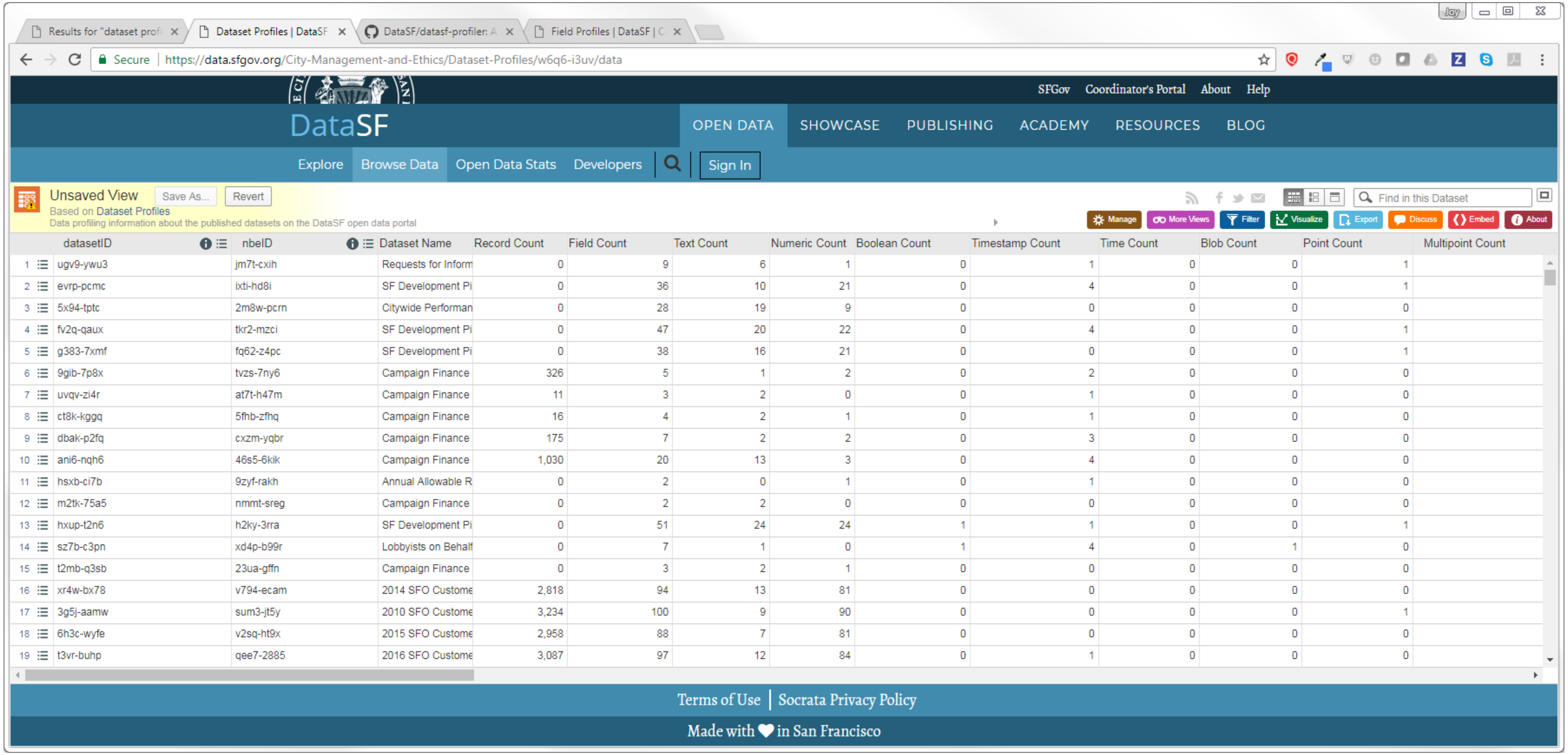
THANK YOU

@datasf | datasf.org |datasf.org/blog

Data profiling

APPENDIX

DataSF aside: We’ve profiled every published dataset



Unsaved View

Based on Dataset Profiles

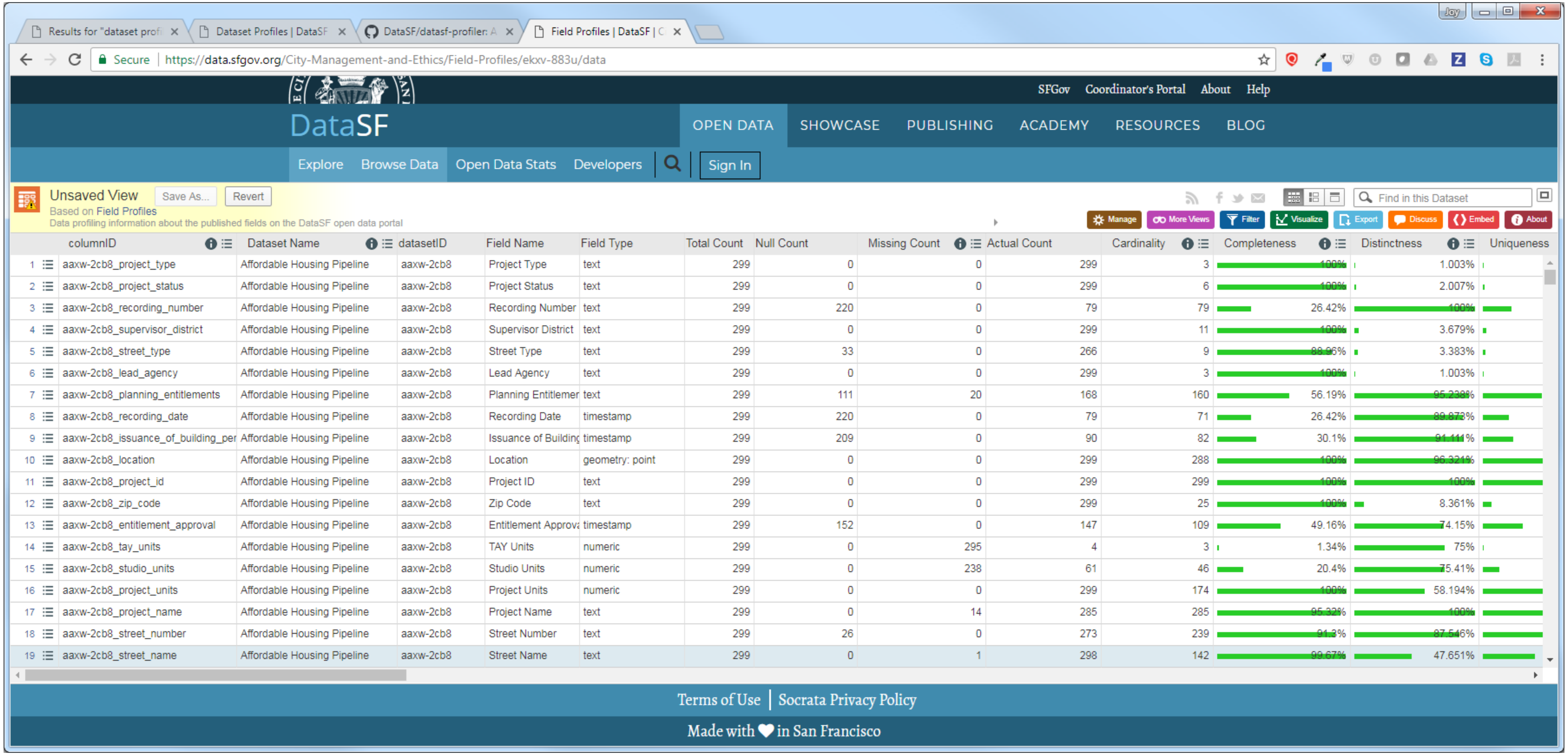
Data profiling information about the published datasets on the DataSF open data portal

	datasetID	nbeID	Dataset Name	Record Count	Field Count	Text Count	Numeric Count	Boolean Count	Timestamp Count	Time Count	Blob Count	Point Count	Multipoint Count
1	ugv9-ywu3	jm7t-cxih	Requests for Inform	0	9	6	1	0	1	0	0	1	
2	evrp-pcmc	ixti-hd8i	SF Development Pi	0	36	10	21	0	4	0	0	1	
3	5x94-tptc	2m8w-pcrn	Citywide Performan	0	28	19	9	0	0	0	0	0	
4	fv2q-qaux	tkr2-mzci	SF Development Pi	0	47	20	22	0	4	0	0	1	
5	g383-7xmf	fq62-z4pc	SF Development Pi	0	38	16	21	0	0	0	0	1	
6	9gib-7p8x	tvzs-7ny6	Campaign Finance	326	5	1	2	0	2	0	0	0	
7	uvqv-zi4r	at7t-h47m	Campaign Finance	11	3	2	0	0	1	0	0	0	
8	ct8k-kggq	5fhb-zfhq	Campaign Finance	16	4	2	1	0	1	0	0	0	
9	dbak-p2fq	cxzm-yqbr	Campaign Finance	175	7	2	2	0	3	0	0	0	
10	ani6-nqh6	46s5-6kik	Campaign Finance	1,030	20	13	3	0	4	0	0	0	
11	hsxb-ci7b	9zyf-rakh	Annual Allowable R	0	2	0	1	0	1	0	0	0	
12	m2tk-75a5	nmmt-sreg	Campaign Finance	0	2	2	0	0	0	0	0	0	
13	hxup-t2n6	h2ky-3rra	SF Development Pi	0	51	24	24	1	1	0	0	1	
14	sz7b-c3pn	xd4p-b99r	Lobbyists on Behalf	0	7	1	0	1	4	0	1	0	
15	t2mb-q3sb	23ua-gffn	Campaign Finance	0	3	2	1	0	0	0	0	0	
16	xr4w-bx78	v794-ecam	2014 SFO Custome	2,818	94	13	81	0	0	0	0	0	
17	3g5j-aamw	sum3-jt5y	2010 SFO Custome	3,234	100	9	90	0	0	0	0	1	
18	6h3c-wyfe	v2sq-ht9x	2015 SFO Custome	2,958	88	7	81	0	0	0	0	0	
19	t3vr-buhp	qee7-2885	2016 SFO Custome	3,087	97	12	84	0	1	0	0	0	

Terms of Use | Socrata Privacy Policy

Made with ♥ in San Francisco

DataSF aside: And every published field



Unsaved View

Based on Field Profiles

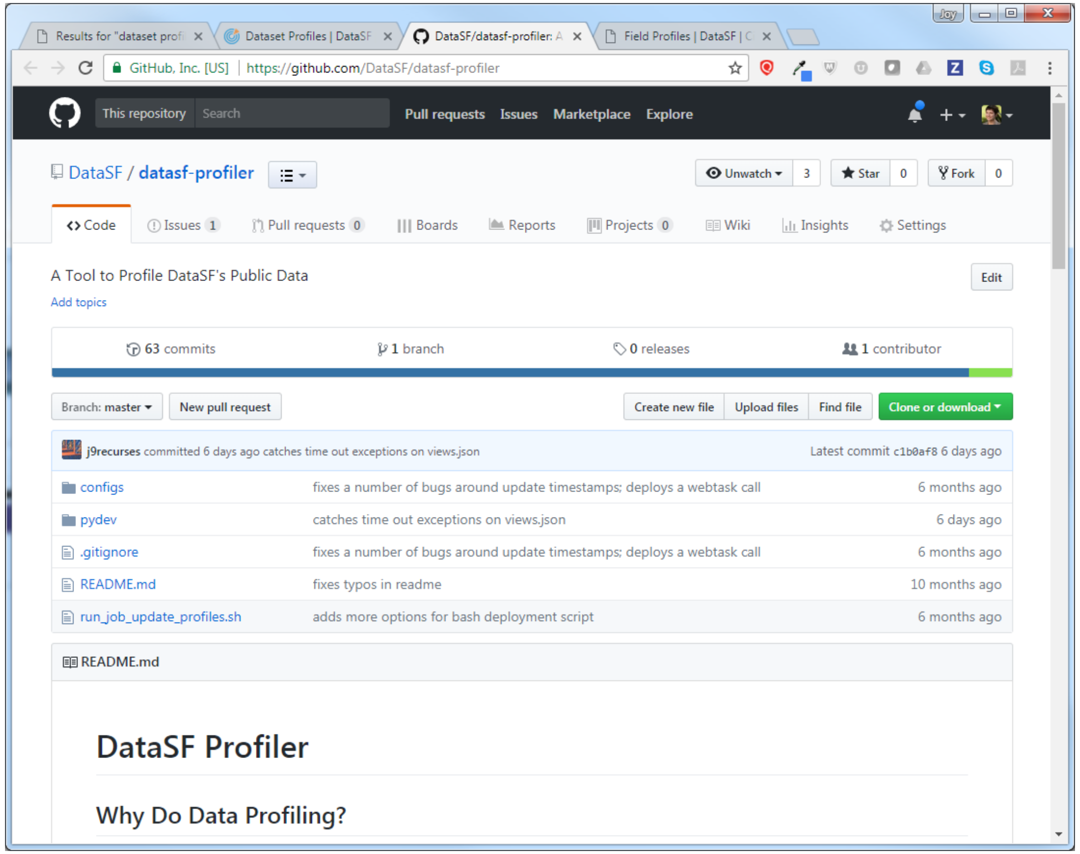
Data profiling information about the published fields on the DataSF open data portal

	columnID	Dataset Name	datasetID	Field Name	Field Type	Total Count	Null Count	Missing Count	Actual Count	Cardinality	Completeness	Distinctness	Uniqueness
1	aaxw-2cb8_project_type	Affordable Housing Pipeline	aaxw-2cb8	Project Type	text	299	0	0	299	3	100%	1.003%	
2	aaxw-2cb8_project_status	Affordable Housing Pipeline	aaxw-2cb8	Project Status	text	299	0	0	299	6	100%	2.007%	
3	aaxw-2cb8_recording_number	Affordable Housing Pipeline	aaxw-2cb8	Recording Number	text	299	220	0	79	79	26.42%	100%	
4	aaxw-2cb8_supervisor_district	Affordable Housing Pipeline	aaxw-2cb8	Supervisor District	text	299	0	0	299	11	100%	3.679%	
5	aaxw-2cb8_street_type	Affordable Housing Pipeline	aaxw-2cb8	Street Type	text	299	33	0	266	9	88.96%	3.383%	
6	aaxw-2cb8_lead_agency	Affordable Housing Pipeline	aaxw-2cb8	Lead Agency	text	299	0	0	299	3	100%	1.003%	
7	aaxw-2cb8_planning_entitlements	Affordable Housing Pipeline	aaxw-2cb8	Planning Entitlemer	text	299	111	20	168	160	56.19%	95.238%	
8	aaxw-2cb8_recording_date	Affordable Housing Pipeline	aaxw-2cb8	Recording Date	timestamp	299	220	0	79	71	26.42%	89.873%	
9	aaxw-2cb8_issuance_of_building_per	Affordable Housing Pipeline	aaxw-2cb8	Issuance of Building	timestamp	299	209	0	90	82	30.1%	91.111%	
10	aaxw-2cb8_location	Affordable Housing Pipeline	aaxw-2cb8	Location	geometry: point	299	0	0	299	288	100%	96.321%	
11	aaxw-2cb8_project_id	Affordable Housing Pipeline	aaxw-2cb8	Project ID	text	299	0	0	299	299	100%	100%	
12	aaxw-2cb8_zip_code	Affordable Housing Pipeline	aaxw-2cb8	Zip Code	text	299	0	0	299	25	100%	8.361%	
13	aaxw-2cb8_entitlement_approval	Affordable Housing Pipeline	aaxw-2cb8	Entitlement Approva	timestamp	299	152	0	147	109	49.16%	74.15%	
14	aaxw-2cb8_tay_units	Affordable Housing Pipeline	aaxw-2cb8	TAY Units	numeric	299	0	295	4	3	1.34%	75%	
15	aaxw-2cb8_studio_units	Affordable Housing Pipeline	aaxw-2cb8	Studio Units	numeric	299	0	238	61	46	20.4%	75.41%	
16	aaxw-2cb8_project_units	Affordable Housing Pipeline	aaxw-2cb8	Project Units	numeric	299	0	0	299	174	100%	58.194%	
17	aaxw-2cb8_project_name	Affordable Housing Pipeline	aaxw-2cb8	Project Name	text	299	0	14	285	285	95.32%	100%	
18	aaxw-2cb8_street_number	Affordable Housing Pipeline	aaxw-2cb8	Street Number	text	299	26	0	273	239	91.3%	87.546%	
19	aaxw-2cb8_street_name	Affordable Housing Pipeline	aaxw-2cb8	Street Name	text	299	0	1	298	142	99.67%	47.651%	

Terms of Use | Socrata Privacy Policy

Made with ♥ in San Francisco

DataSF aside: Profiling scripts are open source and building a dashboard so publishers can easily track



DataSF / **datasf-profiler**

A Tool to Profile DataSF's Public Data

63 commits | 1 branch | 0 releases | 1 contributor

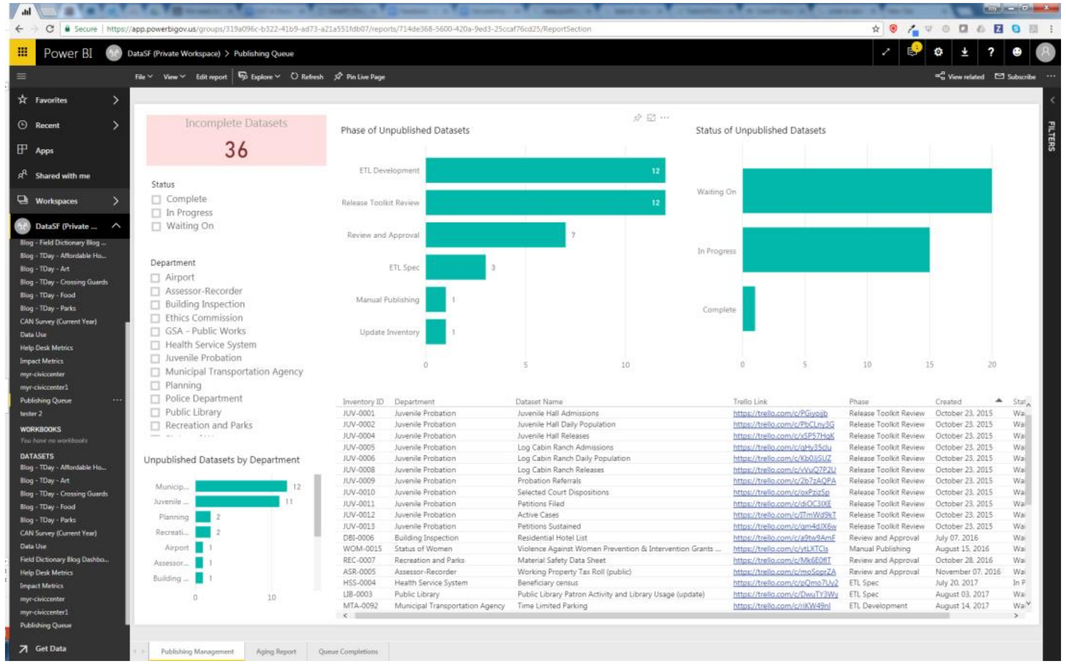
j9recurses committed 6 days ago catches time out exceptions on views.json — Latest commit c1b0af8 6 days ago

configs	fixes a number of bugs around update timestamps; deploys a webtask call	6 months ago
pydev	catches time out exceptions on views.json	6 days ago
.gitignore	fixes a number of bugs around update timestamps; deploys a webtask call	6 months ago
README.md	fixes typos in readme	10 months ago
run_job_update_profiles.sh	adds more options for bash deployment script	6 months ago

README.md

DataSF Profiler

Why Do Data Profiling?



Inventory ID	Department	Dataset Name	Trello Link	Phase	Created	Stat
JUV-0001	Juvenile Probation	Juvenile Hall Admissions	https://trello.com/c/PGiyojjb	Release Toolkit Review	October 23, 2015	Wai
JUV-0002	Juvenile Probation	Juvenile Hall Daily Population	https://trello.com/c/PbCLny3G	Release Toolkit Review	October 23, 2015	Wai
JUV-0004	Juvenile Probation	Juvenile Hall Releases	https://trello.com/c/xSP57HqK	Release Toolkit Review	October 23, 2015	Wai
JUV-0005	Juvenile Probation	Log Cabin Ranch Admissions	https://trello.com/c/qHy35du	Release Toolkit Review	October 23, 2015	Wai
JUV-0006	Juvenile Probation	Log Cabin Ranch Daily Population	https://trello.com/c/Kb0JjSUZ	Release Toolkit Review	October 23, 2015	Wai
JUV-0008	Juvenile Probation	Log Cabin Ranch Releases	https://trello.com/c/vVuQ7P2U	Release Toolkit Review	October 23, 2015	Wai
JUV-0009	Juvenile Probation	Probation Referrals	https://trello.com/c/2b7zAQPA	Release Toolkit Review	October 23, 2015	Wai
JUV-0010	Juvenile Probation	Selected Court Dispositions	https://trello.com/c/oxPzizSp	Release Toolkit Review	October 23, 2015	Wai
JUV-0011	Juvenile Probation	Petitions Filed	https://trello.com/c/diOC3IXE	Release Toolkit Review	October 23, 2015	Wai
JUV-0012	Juvenile Probation	Active Cases	https://trello.com/c/TTmWd9kT	Release Toolkit Review	October 23, 2015	Wai
JUV-0013	Juvenile Probation	Petitions Sustained	https://trello.com/c/gm4dJX6w	Release Toolkit Review	October 23, 2015	Wai
DBI-0006	Building Inspection	Residential Hotel List	https://trello.com/c/a9tw9AmF	Review and Approval	July 07, 2016	Wai
WOM-0015	Status of Women	Violence Against Women Prevention & Intervention Grants ...	https://trello.com/c/ytLXTCls	Manual Publishing	August 15, 2016	Wai
REC-0007	Recreation and Parks	Material Safety Data Sheet	https://trello.com/c/Mk6E0flT	Review and Approval	October 28, 2016	Wai
ASR-0005	Assessor-Recorder	Working Property Tax Roll (public)	https://trello.com/c/moSoprZA	Review and Approval	November 07, 2016	Wai
HSS-0004	Health Service System	Beneficiary census	https://trello.com/c/pQmo7Uy2	ETL Spec	July 20, 2017	In P
LIB-0003	Public Library	Public Library Patron Activity and Library Usage (update)	https://trello.com/c/DwuTY3Wy	ETL Spec	August 03, 2017	Wai
MTA-0092	Municipal Transportation Agency	Time Limited Parking	https://trello.com/c/riKW49nl	ETL Development	August 14, 2017	Wai