Data Lineage What is it, and why do we care?

Ross Turk







The principal drivers of **OpenLineage**



The 2022 Airflow User Survey

Open until Friday, June 3

https://bit.ly/AirflowSurvey22





These are individual opinions, I don't speak for anyone else.

I am pretty sure my employer agrees with them (but I didn't ask...)



What creates the need for data lineage?



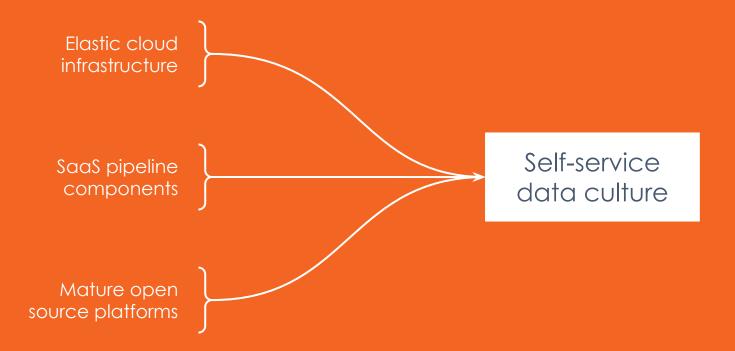
What

Why

How

MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021						
INFRASTRUCTURE	ANALYTICS	MACHINE LEARNING & ARTIFICIAL INTELLIGENCE	APPLICATIONS - ENTERPRISE			
STORAGE HADOOP DATA LAKES DATA STREAMING/	BI PLATFORMS VISUALIZATION	DATA SCIENCE DATA SCIENCE ML PLATFORMS	SALES MARKETING - MARKETING - B2C CUSTOMER EXPERIENCE / SERVICE HUMAN CAPITAL			
MAREHOUSES IN-MEMORY	looker and where the set of the s	NOTEBOOKS PLATFORMS				
Conception Forces Conception F		Spiner Sbinder colob	Cr Coni ghomena / Ap Ania @Rigmant misponel @AmpBlude @HEGALLIA zendesk @AmpBlude @freshdesk Otentio @Bre-			
	T. ThoughtSpot.	Oranitation Contraction Contra	A anto peoplesi Simon intentive Mayania Summan Simon Conter Summan Simon mya d on			
Strim @ sessees Gridaina	ATSCALE OBKO SEPI SCHONA		Change tact.bl S Artis			
A RELEVANT DEPENDENT	SERVER	Adult Martine Lawing and all	collective() have been been been been been been been be			
Construction VAST District Relations OFF	G Ganzoura Monardian Abirst Freyers 1900 Observable	Nextjournal D court @ COCALC ANALONDA	enigma HEADOIP (PERSADOI Orton States Organise verbit's Spiriter y valowal moengage Szakoster			
Occure mit binighere Dice Orgigene binates Barnar	MicroStrategy Aken 10 SRID Preset	Garrener EEREX Contention Arrene Antennas EdetData datmo \$ 2000 abviewsty.ai				
RDBMS - NoSQL DATABASES REALTIME - GRAPH DBs	DATA ANALYST PLATFORMS AUGMENTED ANALYTICS	DATA GENERATION MODEL BUILDING FEATURE STORE DEPLOY MODEL	LEGAL REGTECH & FINANCE AUTOMATION & RPA SECURITY COMPLEX COMP			
DATABASES		& LABELLING MENT & PRO- MONITORING	Commenter Olira Wigo Wigo			
Constitution of the second sec	Microsoft @pentoho alteryx ThoughtSpot.	amazen nedeniedzin dr HTVE 11 Weights & Blases TOCTON @ mereness DUCTION & OBSERVA- BILITY	Training Strendard S			
TANK WARAN ORACLE Improved states Oracle Simply IBM IN	MODE Datameer Children III onod% -outlier	scole Vework sopper The azzart Groups Cloud	SPECIALLY STREAM @ DIGITS III Own ACCOUNT STREAM AND			
P Store Countrie Ontwenties Contractions Contraction	550	Cunty LIDNERIDER Labebox Of Lines Concernent RASCO Conventor				
The second secon	ENDOR Diver B ATTIV/O narrative Science	scortex	- North - North - State - Stat			
	Incorta. GUAVUS @iscoso efacet macheve	\$ secon	🔶 N F O S U M 🔯 SECURITUAL 🥔 Pignarz 20 Winners ALITYTE 👌 barkean 19 Antonin 🥆 Country Co			
Clettery Olectery Proceen		DSuperbAl Steparanda Research Backer Covers. to A Orize				
BEALINGTON OF CARTERS STORES		A reconnect WHYLABS	APPLICATIONS - INDUSTRY			
MPP DBs ETL / ELT / REVERSE ETL DATA INTEGRATION DATA GOVERNANCE DATA TRANSFORMATION & ACCESS	DATA CATALOG METRICS LOG ANALYTICS	COMPUTER VISION SPEECH NLP SYNTHETIC MEDIA	ADVERTISINGEDUCATION REAL ESTATE GOVT & COMMERCE FINANCE -LENDING INSURANCE			
TERADATA	solunk> 🕮	MonsethAuer 🚱 Yaan Ari 🔕 Siri 🔿 amaton aleas 🚳 🎰 👘 🖓 Siri 🖓 amaton aleas	avende Madadata & REDFIN			
NERTICA hightouch	- · · · ·		CONTRACTOR OFFICE SHEIN FAIRE One ZESTO Amade Generation			
Printes Skitch & MATILLION TO PARA	Trace	🔾 Synthesia 📴 descript	Criteo. IASteries Galesie opensor Botamer owned Transford STANDARD General Standard Standar			
Anteringten Classic Carbon Car	Supergrain solarwinds Habana	(D) and an	GRACHE MUSA			
	stemma @logina	Consequences Conse	Grectora and FinAnce - Investing grand and a second			
Rognitio & Service Wary @ Antyse O Participation ZALENI import.io OKERA Program	(+) SELECT STAR • transform	Second Superior Cili-1 2G Ab semantic nutrices Un usual superiore @ (mbodg/fe	albert gumgum Settian Constant AYASDI KENSHC O Monstant Sofilia			
Excessel Militano Warner Attivery Arounders Information Statilia	Escenda @Castor ZQL OBSERVE	Counter ACE Contractor State VOCAUD				
PRIVACY & DATA OBSERVABILITY MGMT / MONITORING SERVER- CLUSTER SVCS	QUERY SEARCH	HORIZONTAL AI GPU DBS & CLOUD AI HARDWARE	HEALTHCARE INDUSTRIAL OTHER			
🗘 VERV GOOD SECURITY Y dolakin 🔳 MONTE CARLO 👢 BARAN O New Retic 🔝 splank: 📈 O Armontos 📰 IBM	elesticienth anaten CRACLE contacted anaten	Watson @ OpenAI C DeepMind KI∩Ettico Google TPU Orm @@/				
PRIVAGERA	(dremio	③ Voyagerues contribute Vicarious.	Weapon 30Mod @herw TEMPUS			
	Starburst	Torrena Contraction Victorious.	patentary 6yboon" (ACue (2000 ed)ans and any total and a gaptar (2000 ed)ans and a gaptar (2000 ed)ans and a gaptar (2000 ed)ans a g			
Toler Charged DATA DUALITY	🛕 ahaoa 🗳 Lucidworks 👩 swiftype ATTIV/O	A Numenta 🔝 Stiller OGROLogis Pretuum				
sheftar Q salard core from the set		From and	W Genera Lucar Bon instro			
United pageoung exacts inuclio packet contentione	SEXALEAD alphasense omni:us	PG-Strem mellour	ezro cutoritude in a con graden graden graden graden County			
Addematopic Bernatopic Version		Hansoyn R. II NULLER PLOTONO COMPANY Combas () STRAM	Martine Owner Laurenter Beste Supposed			
S Ketch Annie wester Konz Case Gordana Labs Opstand Concentra Case Deprover and		T SATE PARKS STAND	Gruppe Haats a Laterates garrocounty we termine Over Go sermos Country Crisp. Veetmy			
		OPEN SOURCE				
FRAMEWORKS FORMAT QUERY/DATA FLOW DATA DATABASES	ORCHESTRATION INFRA- DATA OPS STREAMING &	STAT TOOLS & ML OPS & INFRA AI / MACHINE LEARNING / DEEP LEARNING	SEARCH LOGGING & MONITORING VISUALIZATION COLLABORATION SECURITY			
Access and Access and Access and Access and Access	taken in the set and		L 🛛 Keras BERT Kabour Caffe 💼 Sol? 💼 Kitani Katan 👔 👀 Korean materiala 👔 🔥 Anale Range Kital			
Cardinak YARN TEZA Okobernetes Auka	al Southern W Flyts States States		weiter DADAN FRANKE WEIS CLEAR + Sphitz Lighten Onester Yarts Park Indian Park Indian Contract of the State			
📓 🖗 MESOS 🖉 MARKAN SECON 🖉 💿 👘 orla						
Andrea Australia MELIX			p 📴 🔤 🚵 👘 👘 👘 👘 👘 👘			
	DATA SOURCES & APIs		DATA RESOURCES			
AIR / SPACE / SEA	PEOPLE / ENTITIES LOCATION INTELLIGENCE	OTHER DATA SERVICES				
O Generative ARGENTICS ASPIRE		IDEX COMPLET LONGING DE COMPLET LONGINE				
xignite times earnest predata MX tink?	Quanteast ASIS SAFEGRAPH					
	emperchen Demyst melissa Zagnal -	VERITORE Comscore DataKind a				
			FIRSTMARK -			
Version 3.0 - November 2021 © Matt Turck (@matttu	ck), John Wu (@john_d_wu) & FirstMark (@firstr	narkcap) mattturck.com/data2021	EARLY STAGE VENTURE CAPITAL			

What else changed?



The defining dilemma

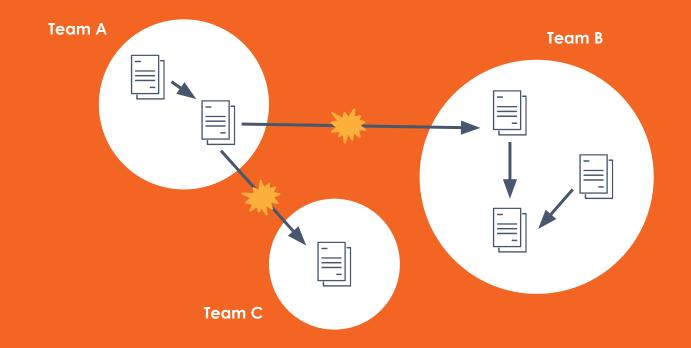


How many pipelines are currently running?

How can we learn about all of them?

How can we know what goes on inside them?

Building a healthy data ecosystem



Ecosystems form around shared understanding



What is the data source? What is the schema? Who is the owner? How often is it updated? Where does it come from? Who is using it? What has changed? What is data lineage?



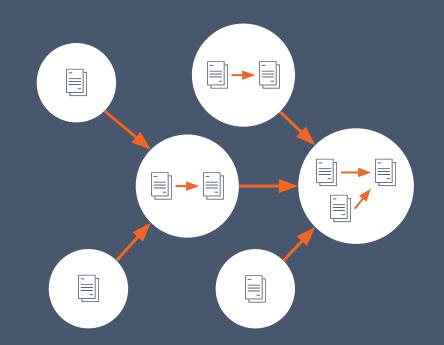


Why



Data lineage is the set of complex relationships between datasets and jobs in a pipeline.

- Producers & consumers of each dataset
- Inputs & outputs of each job





Just know everything, right?

Why do we need data lineage to succeed?



What

Why



Verifying compliance



Optimizing data operations



Establishing context & language







OMG the possibilities are endless

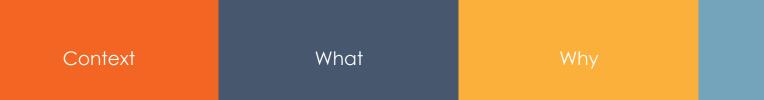
Dependency tracing Root cause identification Issue prioritization Impact mapping Precision backfills Anomaly detection Change management Historical analysis Automated audits



Non-malicious (yet common) lineage lies

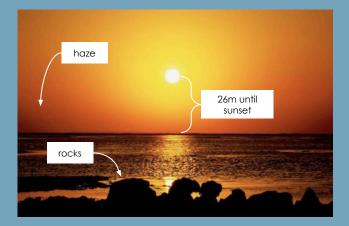
Fully-automated				
Real-time				
End-to-end	360° visibility			
Easy	AI/ML enhanced			

How can we determine data lineage?





EXIF has the right idea



You can try to infer the date and location of an image after the fact...



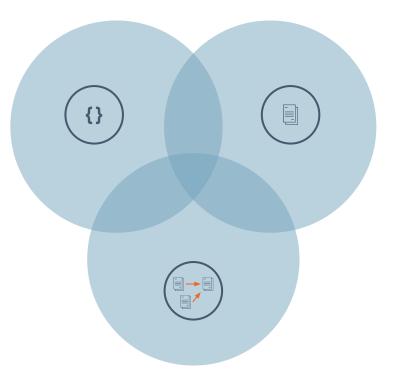
...or you can capture it when the image is originally created!



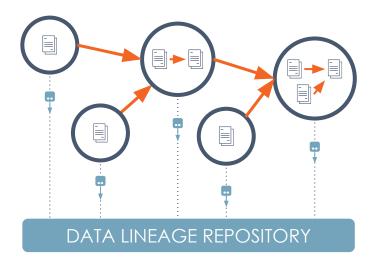
Just collect metadata, right?

MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021					
INFRASTRUCTURE	ANALYTICS	MACHINE LEARNING & ARTIFICIAL INTELLIGENCE	APPLICATIONS – ENTERPRISE		
	BI PLATICRUS VSULUZATION UCSter → → Image: State of the state	DATA SCINCE DATA SCINCE DELEMENT Compared Delement			
MP 2015 Norman Distance of the second seco	Allen and an and an and an and an and an				
PRAMEWORKS FORMAT COURT / DATA FLOW EXCSS MAXABASES Image: sub-2 million Image	ORCHESTINGON MURAL DESCRIPTION MURAL DESCRIPTION DIRECTURE DESCRIPTION DIRECTURE DESCRIPTION DIRECTURE DESCRIPTION <thdirecture description<="" th=""> DIRECTURE DES</thdirecture>	β person of any other basis	College a month of the second se		
International accompany International	Service Strokow Standowich Place Service Constraint	An Ander Ampliery	Desc desc function Oracles Desc desc function Oracles Desc desc function Oracles Orac		

Comparing approaches



Observe the pipeline

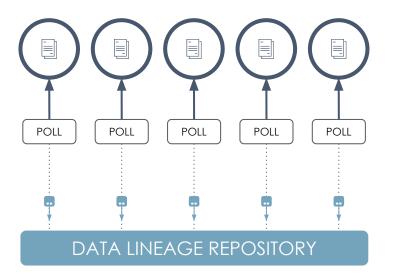


Integrate with data orchestration systems

As jobs run, observe the way they affect data

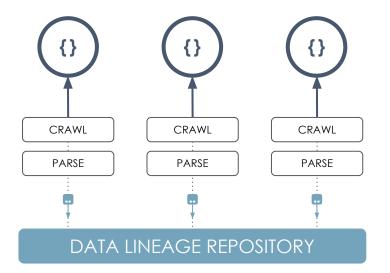
Report to a lineage metadata repository

Process query / activity logs



- Integrate with data stores and warehouses
- Regularly process query logs to trace lineage
- Report to a lineage metadata repository

Analyze source code



Integrate with source code repositories

Look for queries and parse them for lineage

Report to a lineage metadata repository

It's a patchwork



A fable about community



OpenLineage

An **open standard** for the collection of lineage metadata from pipelines **as they are running**.



Standards gain momentum



The future of data lineage

Data lineage will **become standardized** and the means for gathering it will be increasingly commodifized.

Data lineage will become **increasingly multifaceted** as multiple worlds collide.

Data lineage will **work alongside orchestration** to automate operational tasks, e.g. root cause analysis and backfills.

Thanks :)