

Evolving Data Pipelines at Scale

January 15, 2024

Copyright © 2024 Tobiko Data, Inc.

About Me

- Co-Founder & Chief Architect at Tobiko Data
- Leading the development of SQLMesh
- Over 10 years of experience in data / ML infra
- Previously Netflix, Apple



• Creating data environments for development is cumbersome



• Populating dev environments with data is inefficient



• Development iterations are slow



• Development iterations are slow

THE DATA OANT DE WRONG



makeameme.org

• Production deployments are ad-hoc and error-prone





Scale



Scale Organizational Complexity





Model: T in ETL

SELECT a, b **FROM** source **WHERE** c > 0

Model



Model Attributes

Impacts data

- Model query
- Type of data object
- Storage format
- Partitioning / clustering

Doesn't impact data (metadata)

- Owner
- Model / column descriptions
- SQL comments
- Formatting

Data hash

Metadata hash

Model Attributes

Impacts data

- Model query
- Type of data object
- Storage format
- Partitioning / clustering

Data hash

Doesn't impact data (metadata)

- Owner
- Model / column descriptions

Metadata hash

- SQL comments
- Formatting

Model Fingerprint

Model Snapshot



Snapshot Properties

- Model frozen in time
- Uniquely identified by the **fingerprint**
- Self-contained
- Immutable

Snapshots in the Data Warehouse



Snapshots in the Data Warehouse



Pillar of Data Versioning

Fast, Safe & Cost-Effective development process



Virtual Data Environments

Virtual Layer

Physical Layer

Virtual Data Environments

Virtual Layer

Physical Layer





Virtual Data Environments



1. Starting production environment



2. Create development as a mirror of production



3. Make a change to Model A



3. Make a change to Model A



4. Deploy model to production



It's nice and all, but what if...

- The table is too large and costly to rebuild
- The model logic is not idempotent
- The historical data is not available upstream

It's nice and all, but what if...



Forward-only Changes



Forward-only Changes



Pillar of Virtual Data Environments

Fast, Safe & Cost-Effective development process



Semantic Understanding

SELECT a, b **FROM** source **WHERE** c > 0;

Semantic Understanding

SELECT a, b **FROM** source **WHERE** c > 0;



Semantic Understanding

SELECT a, b, d **FROM** source **WHERE** c > 0
Analyzing the Change



Analyzing the Change





SELECT a, b FROM source WHERE c > 100





Pillar of Automatic Data Contracts

Fast, Safe & Cost-Effective development process







Fast, Safe & Cost-Effective development process













Putting Ideas into Practice with SQLMesh

About SQLMesh

- Open-source data transformation and modeling tool
- Pure SQL, no Jinja
- Batteries included: interval tracking, unit tests, CI/CD, etc
- Backwards compatible with dbt
- Brings Three Pillars of Data Productivity to life

🔊 sal	Mesh b	y 🔳 tobi	ko																																	Do	cum	ent	atio	n	
-																											Pia	in (D)	F	orod	1 >	ł	1o (Cha	inge	s	N	lo E	rrors	
		target.sql																																							
₽	1 v 2 3 4 5 v 6 7 8 9 10	MODEL (name); SELECT a, b FROM dc WHERE c _b > 0	dc.tar .sourc	get e																																					
	Savec	i: • Forma eage	tted: •	Lang	guage	:: SQL	_ Dia	alect	t: du	ıckd	b S	SQL	.Me	sh	Тур	be: I	nod	el																						1	
	dc.tar	rget													AI	I: 3		Ups	trea	m/l	ow	nstr	rear	n: 2														8	Sho	N N	
																									• •			• •			•				•						
	: : +						0	Bes da		4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	• • • •) es) decta	ant i		•	•		· ·	•	00	 	downst	tream	2				 			•			
	: ::	· · ·								• •									:	() () ()	· ·	• • •									•				•						
	<u> </u>																																								
	. • •																																						i		





n sal	Mesh b	by 🚺)to	bik	0																																					Doc	ume	nta	tion	*	
																																(Pla	1 (C		de	v `		١	lo C	har	iges	5	No	Err	ors	
		tar	get.s	sql																																											
₽	1 ~ 2 3 4 5 ~ 6 7 8 9 10	MO); SE FR WH	DEL nami a, b OM ERE c >	(≇d dc.	c.ta	rce	et																																								
	Save	d: 💿	For	nat	ted:	• 1	Lang	juaç	je: S	SQL	Di	alec	t:d	uck	db	sc	QLN	les	h T	ype:	moc	del																									
	Lin	eage																																													
	dc.ta	rget																		All: 3		Ups	stre	am/	Dov	wns	tre	am	: 2															7 SI	now		
	+) 30) 0005 1	• • •	4 80 80 80 80	•						• • • • • • • •	· · ·	dectory		INT						•	00	•	• • •	e e erram	2 NT NT												
SQLMes	h: 0.72	2.2																																	©	202	4 <u>T</u>	obik	to D	ata,	Inc	. AI	l rig	nts	ese	rveo	Ι.

Deploy to Prod

\lambda SQL Mesh	by 🎒 tobiko					Docum	nentation 🔅
- <					Plan 🕞	dev 🦋 No Changes	No Errors
с ÷	target.sql						
E 1 2 2 3 4 6 7 8 9 10	MODEL (name dc.target); SELECT a, b FROM dc.source WHERE c > 100						
Save	ed: • Formatted: • Lar	nguage: SQL Dialec	t: duckdb SQLMesh	Type: model			
Lin	neage						
dc.ta	arget			All: 3 Upstream/Downs	tream: 2 Find		\bigtriangledown Show \checkmark
		. 0 •				M M	
QLMesh: 0.7	2.2				G	2024 Tobiko Data, Inc. All ri	ghts reserved.



Join us on Slack! https://tobikodata.com/slack

Thank you!



Backup Slides

G-OGLANDAN								Docuit	
- <						Plan 🕞	prod 🗸	No Changes	No Errors
+	target.sql								
=0 1.	MODEL (
	name dc.target								
→ ⁴	<i></i>								
5 -	SELECT								
6	a, b								
8	FROM dc.source								
	WHERE								
10	c > 0								
Save	ed: ● Formatted: ● Lang neage	juage: SQL Diale	ect: duckdb SQ	LMesh Type: mo	del				
Save Lin dc.ta	ed: • Formatted: • Lanç neage arget	juage: SQL Diale	ect: duckdb SQ	ILMesh Type: mo All: 3	del Upstream/Downstream: 2				₩ ₩ ₩
Save	ed: • Formatted: • Lang neage arget 	Juage: SQL Diale	ect:duckdb SQ	ILMesh Type: mo All: 3	del Upstream/Downstream: 2 				♥ Show ∨
Save Lin dc.ta	ed: • Formatted: • Lang neage arget 	Juage: SQL Diale	ect:duckdb SQ	LMesh Type: mo All: 3	del Upstream/Downstream: 2				♥ Show ∨ • • • • • •
Save Lin dc.ta	ed: • Formatted: • Lang neage arget 	Juage: SQL Diale	ect:duckdb SQ	All: 3	del Upstream/Downstream: 2	Find			♥ Show ∨
Save	ed: • Formatted: • Lang neage arget	juage: SQL Diale	ect: duckdb SQ	All: 3	del Upstream/Downstream: 2	Find	steam 2		♥ Show ∨
Save	ed: • Formatted: • Lang neage arget	Juage: SQL Diale	ect: duckdb SQ	All: 3	del Upstream/Downstream: 2	Find	steam 2 NT NT		♥ Show ♥
Save Lin dc.ta · · · · · · ·	ed: • Formatted: • Lang neage arget • • • • • • • • • • • • • • • • • • •	Juage: SQL Diale	ect: duckdb SQ	All: 3	del Upstream/Downstream: 2	Find	stream 2 INT NT 		✓ Show ∨
Save	ed: • Formatted: • Lang neage arget	Juage: SQL Diale	ect: duckdb SQ	All: 3	del Upstream/Downstream: 2	Find	stream 2 NT NT NT NT NT NT NT NT NT NT		∑ Show ∨ ∴ ☆
dc.ta	ed: • Formatted: • Lang	Juage: SQL Diale	ect: duckdb SQ	All: 3	del Upstream/Downstream: 2	Find	stream 2 NT NT NT NT NT NT NT NT NT NT		✓ Show ∨

C > Plan © prod v No Changes No Errors + target sql • prod (remote) v MODEL (• production Environment 2 name dc.target 3); 5 SELECT 6 a; 7 b 8 FROM dc.source 9 WHERE 10 c > 0 Saved: Formatted: Language: SQL Dialect: duckdb SQLMesh Type: model Lineage dc.target All: 3 Upstream/Downstream: 2 Find ************************************	0	tobiko					Documentation 🔅
<pre> target.sql</pre>						Plan 🖻 prod 🗸	No Changes No Errors
Saved: • Formatted: • Language: SQL Dialect: duckdb SQLMesh Type: model	+ ta 1 v M 2 3) 4 5 v S 6 7 8 7 8 7 9 W 10	arget.sql MODEL (name dc.target); SELECT a, b FROM dc.source WHERE c > 0			⑦ prod (remote) Production Environment dev	☆ Add	
dc.target All: 3 Upstream/Downstream: 2 Find Show + + + + + + + + + + + + + + + + + + +							
	Saved:	● Formatted:● Langua	nge: SQL Dialect: duckdb SQLMe	sh Type: model			
	Saved: CLineag	 Formatted: Langua Langua 	nge: SQL Dialect: duckdb SQLMe	ash Type: model All: 3 Upst	ream/Downstream: 2 Find		⊽ Show ∨



breaking_change.may by 🌐 tobiko		Doc	umentation 🔅
		Plan 🕟 dev 🗸 Changes 1 🌔 1	No Errors
+ target.sql		Indirectly Modified	
↓ MODEL (2 name dc.target 3):		↔ dc_dev.downstream	
b 5 v select			
7 b, 8 d			
9 FROM dc.source 10 WHERE			
11 c > 0			
Saved: • Formatted: • Language: SQL Dialect: d	luckdb SQLMesh Type: model		
Lineage			
dc.target	All: 3 Upstream/Downstream: 2		\bigtriangledown Show \checkmark
· · · · · · · · · · · · · · · · · · ·	ce 4 O	• • • • • • • • • • • • • • • • • • •	
	Poil Bit Bit NT ·	• • • • • • • • • • • • • • • • • • •	
	· · · · · · · · · · · · · · · · · · ·		
· · · · · · · · · · · · · · · · · · ·			

reaking_change.may	by 🏢 tobiko	Documentation
		Plan 🕑 dev 🗸 Changes 🕦 1 No Error
	Start Date (UTC)	End Date (UTC)
		2022-12-13
	The start datetime of the interval	The end datetime of the interval
	Additional Options	0
	Changes	
	Modified Directly ↔ dc_dev.target	Non-Breaking Change 📀
	Modified Indirectly ⊕ dc_dev.downstream	•
	Backfills	
	Models 1	
	C dc_dev.target	
	Apply Changes And Backfill	Start Over
⊕S629Massh: 0.72	2.2	© 2024 <u>Tobiko Data, Inc.</u> All rights reserv



	Plan 💽 dev 🗸 Changes	s 🕕 🕕 No Error
Start Date (UTC)	End Date (UTC)	
The start datetime of the interval	The end datetime of the interval	
Additional Options		€
Changes		
Backfills		
Models 1 ℃ dc_dev.target		
⊌ ✓ Tests Skipped		
Apply Changes And Backfill		Start Over

reaking_change.may	by 🌐 tobiko	Documentation
		Plan 🕑 dev 🗸 No Changes No Errors
E 3	Backfills	
	✓ No Tests	
	Evaluation started at 2024-02-26 19:27:08	
	Snapshot Tables Created	
	✓ No Models To Restate	
	Backfilled	
	Target Environment dev	1 of 1 task 1 of 1 batch 100%
	2024-02-25 - 2024-02-25 dc_dev.target	1 of 1 batch 0:01
	Environment Promoted *	
	Evaluation stopped at 2024-02-26 19:27:08	
		Go Back
Sol-Mach: 07	2.2	@ 2024 Tabika Data Ing







SQLMesh by	iii tobiko				Documentatio	on 🔅
∼		Plan 📀	dev 🗸	Changes 🚺) 2 No E	rrors
=]]	The start datetime of the interval	The end datetime of the interval				
	Additional Options				Ŧ	
	Changes					l
	Modified Directly					
	⊖ dc_dev.target			Breaking Cl	hange 🔸	
	Modified Indirectly					
	⊖ dc_dev.downstream				•	
	Backfills					
	Models 2					
	᠊᠊ᠿ᠂dc_dev.downstream ᠊ᠿ dc_dev.target					
ſ	Apply Changes And Backfill				Start Over	
QLMesh: 0.7<u>2.2</u>			© 2024	Tobiko Data, Ind	<u>c.</u> All right <u>s re</u>	served

🕖 SQL Mesh b	y 🌐 tobiko			Documentation 🔅
		Plan 🖻	dev 🗸 Change	es 🕦 2 No Errors
	Start Date (UTC)	End Date (UTC)		
	The start datetime of the interval	The end datetime of the interval		
	Additional Options			o
	Changes			
	Backfills ^b			
	Models 2			
	✓ Tests Skipped			
	Apply Changes And Backfill			Start Over
QLMesh: 0.72 .	.2		© 2024 <u>Tobiko D</u>	ata, Inc. All rights reserved.

SQLIMESI	h by 🌐 tobiko	Documentation 🌟
		Plan 💿 dev \vee No Changes No Errors
	✤ Backfills	
	✓ No Tests	
	Evaluation started at 2024-02-26 21:06:24	
	Evaluation started at 2024-02-20 21:00-34	
	Snapshot Tables Created	
	✓ No Models To Restate	
	Backfilled	
	Target Environment dev	2 of 2 tasks 2 of 2 batches 100%
	2024-02-25 - 2024-02-25 dc_dev.downstream	1 of 1 batch 0:01
	2024-02-25 - 2024-02-25 dc_dev.downstream 2024-02-25 - 2024-02-25 dc_dev.target	1 of 1 batch 0:01 1 of 1 batch 0:01
	2024-02-25 - 2024-02-25 dc_dev.downstream 2024-02-25 - 2024-02-25 dc_dev.target	1 of 1 batch 0:01 1 of 1 batch 0:01
	2024-02-25 - 2024-02-25 dc_dev.downstream 2024-02-25 - 2024-02-25 dc_dev.target	1 of 1 batch 0:01 1 of 1 batch 0:01
	2024-02-25 - 2024-02-25 dc_dev.downstream 2024-02-25 - 2024-02-25 dc_dev.target • Environment Promoted Evaluation stopped at 2024-02-26 21:06:34	1 of 1 batch 0:01 1 of 1 batch 0:01 Go Back

Deploy to Prod

							Plan (I	b) dev 🗸	No Changes	No Erro
+	target.sql				Q	dev (remote)				
1 ,	MODEL (
	name dc.target				S 🚽	prod (remote		\$		
3∠);					Production Envi	ronment			
) 5	SELECT							Add		
6	а,									
	b									
8	FRUM dc.source									
10	c > 100									
Saved	l: ● Formatted: ● La	nguage: SQL Di	alect: duckdb SQLMe	e sh Type: mo	del					
Saved	d:● Formatted:● Lan	nguage: SQL Di	alect: duckdb SQLMe	e sh Type: mo	del					
Saved Line dc.tar	d:● Formatted:● Lan eage	nguage: SQL Di	alect: duckdb SQLMe	sh Type: mo All: 3	del Upstream/	Downstream: 2				♥ Show
Saved Line dc.tar	d:● Formatted:● Lan eage rget	nguage: SQL Di	alect: duckdb SQLMe	esh Type: mo All: 3	del Upstream/	Downstream: 2				♡ Show
Saved Line dc.tar	d: • Formatted: • Lan eage rget	nguage: SQL Di	alect: duckdb SQLMe	All: 3	Upstream/	Downstream: 2 				♥ Show
Saved Line dc.tar	d: • Formatted: • Lai	nguage: SQL Di	alect: duckdb SQLMe	sh Type: mo All: 3	del Upstream/ 	Downstream: 2				♥ Show
Saved Line dc.tar	i: • Formatted: • Lai	nguage: SQL Di	alect: duckdb SQLMe	sh Type: mo All: 3	del Upstream/ 	Downstream: 2	Find 	• • • • • • • • • • • • • • • • • • •		♥ Show
Saved Line dc.tar 	d: • Formatted: • Lai	nguage: SQL Di	alect: duckdb SQLMe	sh Type: mo All: 3	del Upstream/ · · · · · · · · ·	Downstream: 2	Find	winstream 2 0 0		∇ Show
Saved Line dc.tar 	i: • Formatted: • Lai	nguage: SQL Di	alect: duckdb SQLMe	All: 3	Upstream/ · · · · · · · · · ·	Downstream: 2	Find	winstream 2 · · ·		Show
Saved Line dc.tar 	i: • Formatted: • Lan	nguage: SQL Di	alect: duckdb SQLMe	All: 3	del Upstream/	Downstream: 2	Find	winstream 2		♥ Show
Saved Line dc.tar 	d: • Formatted: • Lan	nguage: SQL Di	alect: duckdb SQLMe	All: 3	del Upstream/	Downstream: 2	Find	wenstream 2 NYT		
Saved Line dc.tar 	d: • Formatted: • Lan	nguage: SQL Di	alect: duckdb SQLMe	All: 3	del Upstream/ · · · · · · · · ·	Downstream: 2	Find	winstream 2 NeT NeT		
Saved Line dc.tar 	d: • Formatted: • Lan	nguage: SQL Di	alect: duckdb SQLMe	All: 3	del Upstream/	Downstream: 2	Find	wmstream 2 NYT NYT NYT		♥ Show
oloy_to_prod_movMes	sh by 🌐 tobiko	Documentation 🔅								
---------------------	----------------------	---								
		Plan 🕟 prod 🗸 Changes 🕦 No Errors								
E 3	Prod Environment									
	Additional Options	•								
	Changes									
	Modified Directly	Breaking Change 💿								
	Modified Indirectly	•								
	✓ No Backfills									
	✓ Tests Skipped									
	Virtual Update									
	Apply Virtual Update	Start Over								
wSQdeMash:0	.72.2	© 2024 <u>Tobiko Data, Inc.</u> All rights reserve d .								

Product Mesh	by 🌐 tobiko		Documentation 🔅
		Plan 💽 prod 🗸	Changes 🕦 No Errors
	Prod Environment		
	Start Date (UTC)	End Date (UTC)	
	The start datetime of the interval	The end datetime of the interval	
	Additional Options		•
	Changes		
	✓ No Backfills		
	✓ Tests Skipped		
	Virtual Update		
	Apply Virtual Update		Start Over
₀Sର୍ୗ₄Mash: 0.7	12.2	© 2024 T	obiko Data, Inc. All rights reserv







Join us on Slack! https://tobikodata.com/slack

Thank you!

