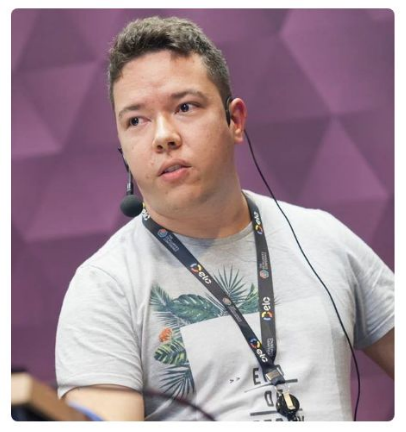
Using cutting-edge open-source technologies to build a World-class Industrial Data Lake

www.datasprints.com

Hello!

My name is Allan Sene



Co-Founder & CTO @Data Sprints

Co-Founder & Podcaster & Instructor @Data Hackers

+10 years in Data & Software, +4 years as Data Engineer

Agenda

- 🤔 The Challenge
- 🤔 What is a Data Lake?
- 🤔 Awesome cutting-edge data tools
- 🤔 Putting everything together do build a IDL
- 🤔 Results and Next Steps
- 🤔 Q&A

THE CHALLENGE

The Challenge

- Multinational Steel Industry, with plants worldwide
- They need to give to global managers the capability of track the production line
- Data columns have BLOBs, Arrays and complex data types
- Migrating from on-prem to cloud
- Data sets with 70 Gb (compressed) and 8 million lines, more than 1600 columns, very complex queries (+200 joins), 15 minutes delay
- Maximum Query Response time of 30s
- Very limited cloud budget - USD 15.000,00/year

Stakeholder's name, can you guess it?

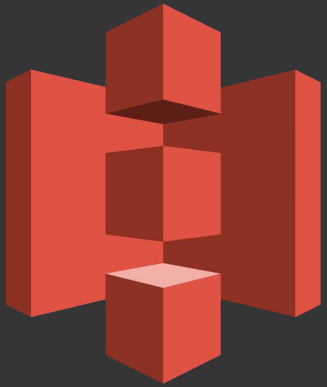
Milagres (Miracles)

What is a Data Lake?

“Is a Data Repository that holds a huge quantity of data on raw state, structured or not. The schema is only defined when is necessary for consumption"

(Anne Buff - Best Practices Leader at SAS)

So, Data Lake is...



Not really...

A Data Lake must have

- Security & Audit
- Catalog & Access
- Data Pipelines & Orchestration

Basically, a **Platform** that ensure safe and efficient data consumption

Awesome cutting-edge data tools

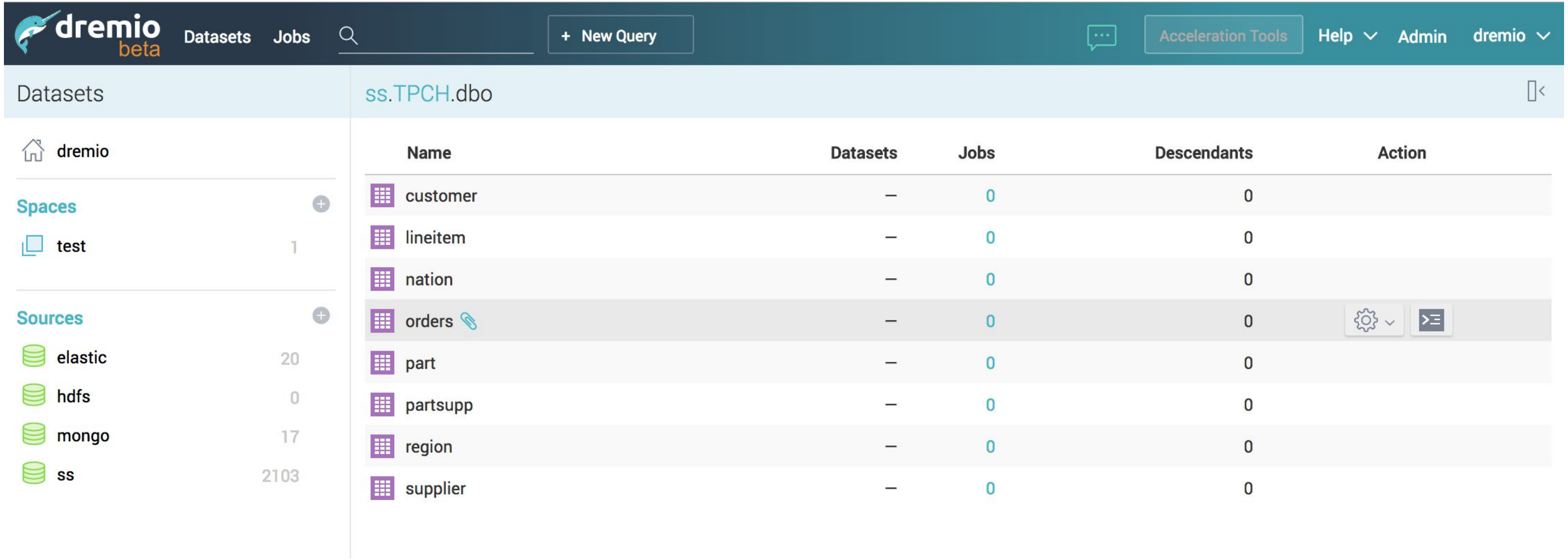
Dremio

- Data Lake Engine
- PMCs of Apache Arrow
- User-Friendly
- Open-source
- Does have an Enterprise Version

Dremio

- SQL & API
- Wiki
- Connections:
  - S3, HDFS, AzureFS
  - Mongo, Elastic,
  - Postgres, MySQL
- Supports JSON, CSV, Parquet, Avro...
- Reflections

Dremio's UI



Datasets

ss.TPCH.dbo

dremio

Spaces

test	1

Sources

elastic	20
hdfs	0
mongo	17
ss	2103

Name	Datasets	Jobs	Descendants	Action
customer	–	0	0	
lineitem	–	0	0	
nation	–	0	0	
orders	–	0	0	
part	–	0	0	
partsupp	–	0	0	
region	–	0	0	
supplier	–	0	0	

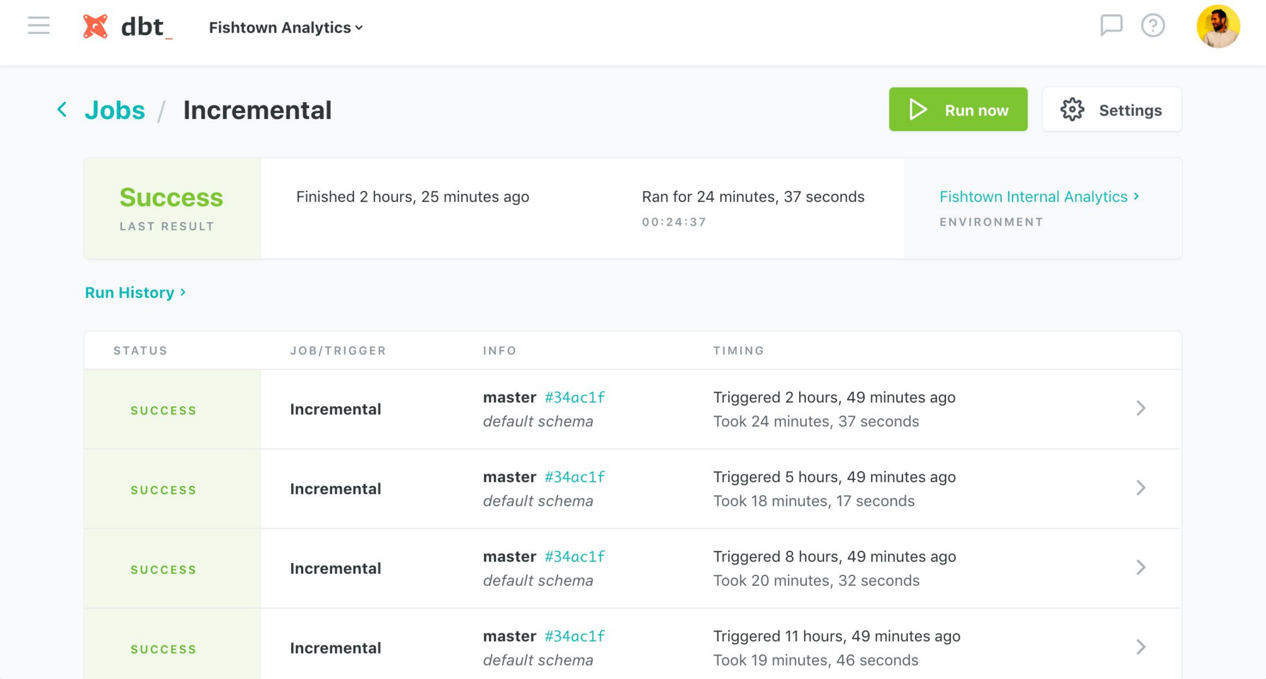
dbt

- Dbt = "Data Build Tool"
- Built by Fisthtown Analytics
- Data Pipeline Orquestration
- "Airflow for Data Analysts"
- Open-source
- SaaS Version

dbt

- ETL over a MPP
- Code Versioning
- Data Lineage & Docs
- Data Validation
- Seed Data

dbt's UI



‹ Jobs / Incremental

▷ Run now

⚙ Settings

Success
LAST RESULT

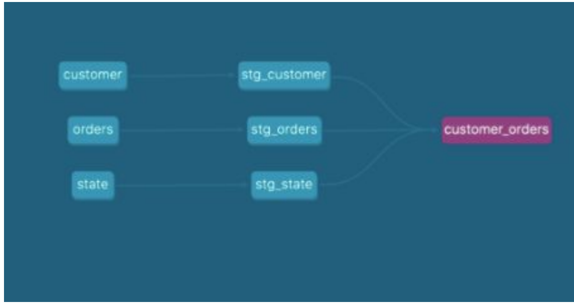
Finished 2 hours, 25 minutes ago

Ran for 24 minutes, 37 seconds
00:24:37

Fishtown Internal Analytics ›
ENVIRONMENT

Run History ›

STATUS	JOB/TRIGGER	INFO	TIMING	
SUCCESS	Incremental	**master** #34ac1f *default schema*	Triggered 2 hours, 49 minutes ago Took 24 minutes, 37 seconds	›
SUCCESS	Incremental	**master** #34ac1f *default schema*	Triggered 5 hours, 49 minutes ago Took 18 minutes, 17 seconds	›
SUCCESS	Incremental	**master** #34ac1f *default schema*	Triggered 8 hours, 49 minutes ago Took 20 minutes, 32 seconds	›
SUCCESS	Incremental	**master** #34ac1f *default schema*	Triggered 11 hours, 49 minutes ago Took 19 minutes, 46 seconds	›

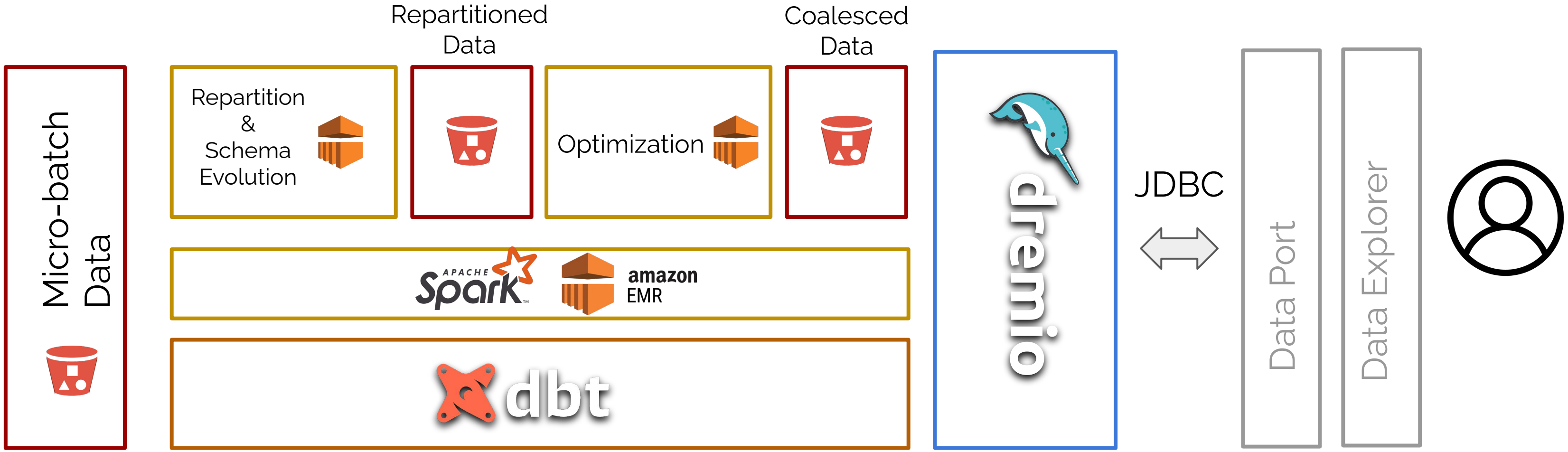


Putting everything together do build a DL

Bringing all together

- **Pipelines & Orchestration => dbt**
- **Catalog & Access => Dremio**
- Processing => Spark
- Storage => Amazon S3

Bringing all together

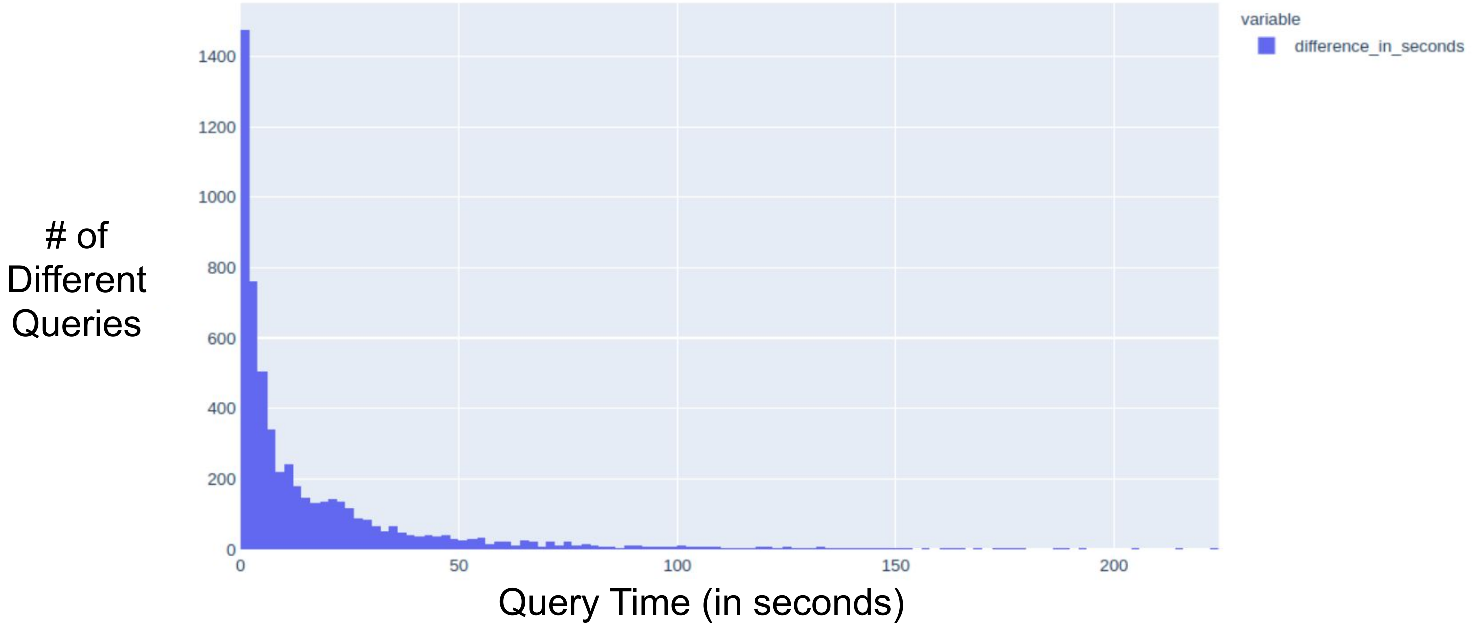


Results and Next Steps

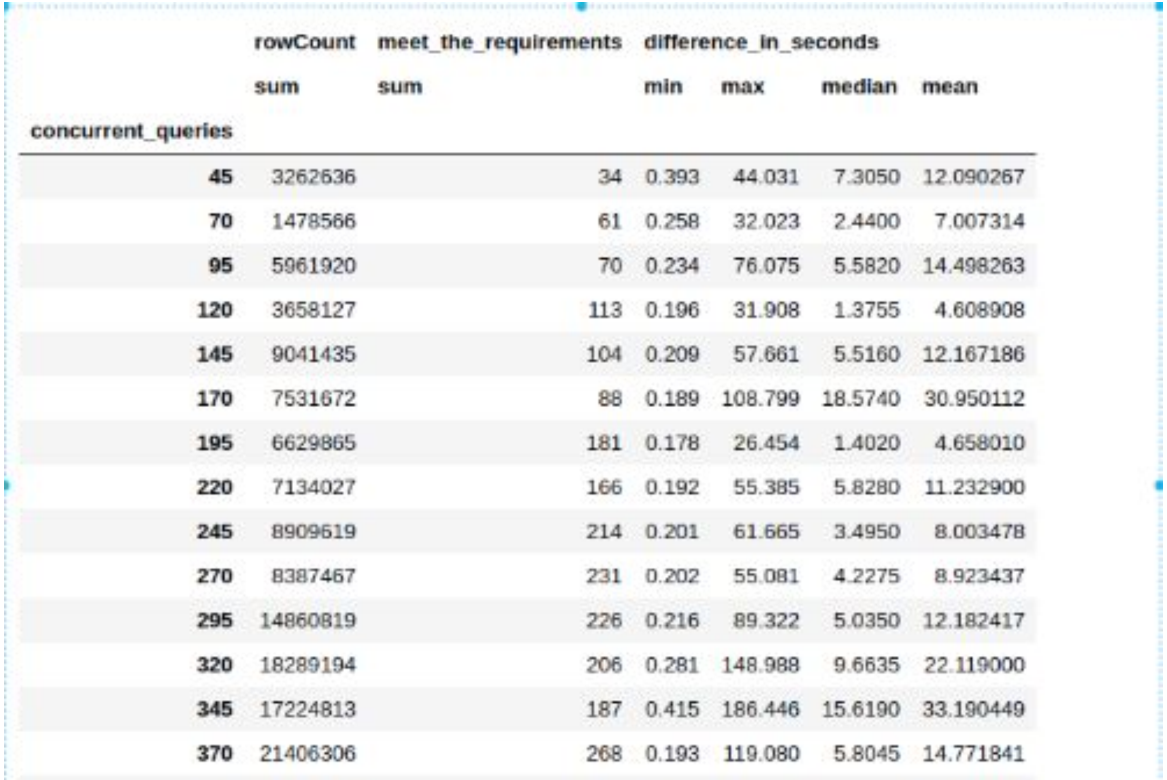
Results, until now

- **Platform built on 6 months by a team of 4 engineers full-time**

Results, until now



Results, until now



	rowCount	meet_the_requirements	difference_in_seconds			
	sum	sum	min	max	median	mean
concurrent_queries						
45	3262636	34	0.393	44.031	7.3050	12.090267
70	1478566	61	0.258	32.023	2.4400	7.007314
95	5961920	70	0.234	76.075	5.5820	14.498263
120	3658127	113	0.196	31.908	1.3755	4.608908
145	9041435	104	0.209	57.661	5.5160	12.167186
170	7531672	88	0.189	108.799	18.5740	30.950112
195	6629865	181	0.178	26.454	1.4020	4.658010
220	7134027	166	0.192	55.385	5.8280	11.232900
245	8909619	214	0.201	61.665	3.4950	8.003478
270	8387467	231	0.202	55.081	4.2275	8.923437
295	14860819	226	0.216	89.322	5.0350	12.182417
320	18289194	206	0.281	148.988	9.6635	22.119000
345	17224813	187	0.415	186.446	15.6190	33.190449
370	21406306	268	0.193	119.080	5.8045	14.771841

Next steps

- Integrating with very common Analytics tools (Power BI, Tableau...), deployed worldwide
- Data Science Models consuming data through the Data Bus
- Data Quality Monitoring

DataHackers

Data Hackers Podcast Spotify

linkedin.com/company/data-hackers/

www.datahackers.com.br

Questions?

www.datasprints.com

Thanks!

allan@datasprints.com
https://www.linkedin.com/in/allansene/

www.datasprints.com