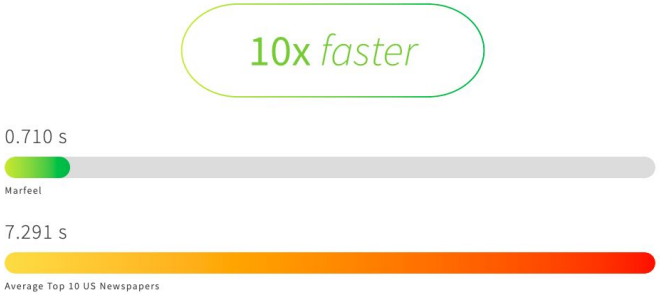
Data at Marfeel

Addressing complexity at scale with the latest technologies

Alessandro Pregnolato
Head of Data

What does **Marfeel** do?

Optimize.
Engage.
Monetize.



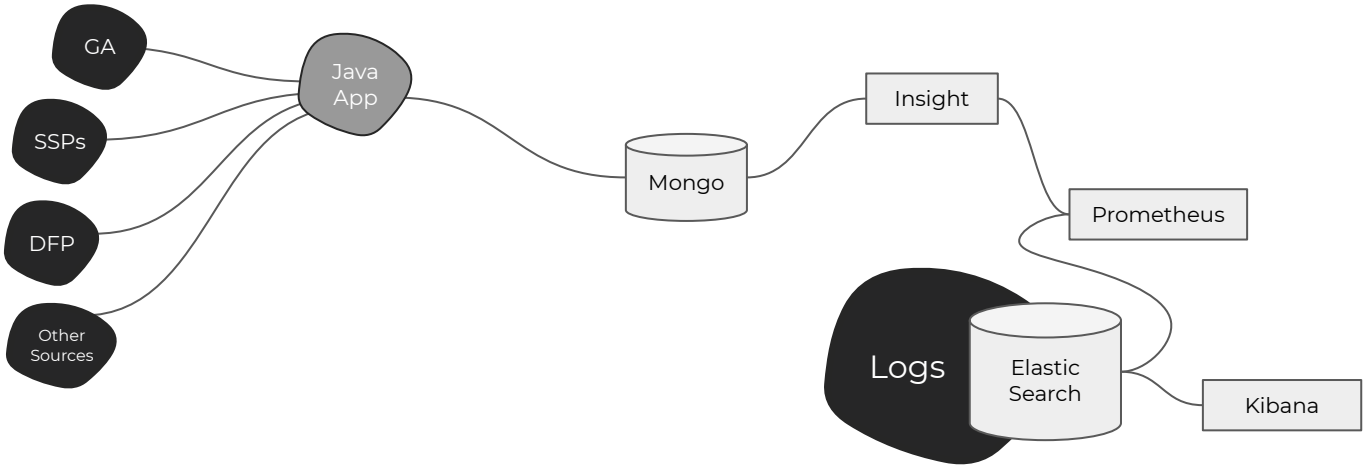
→ A lightning fast, responsive mobile web

→ A sophisticated monetization layer

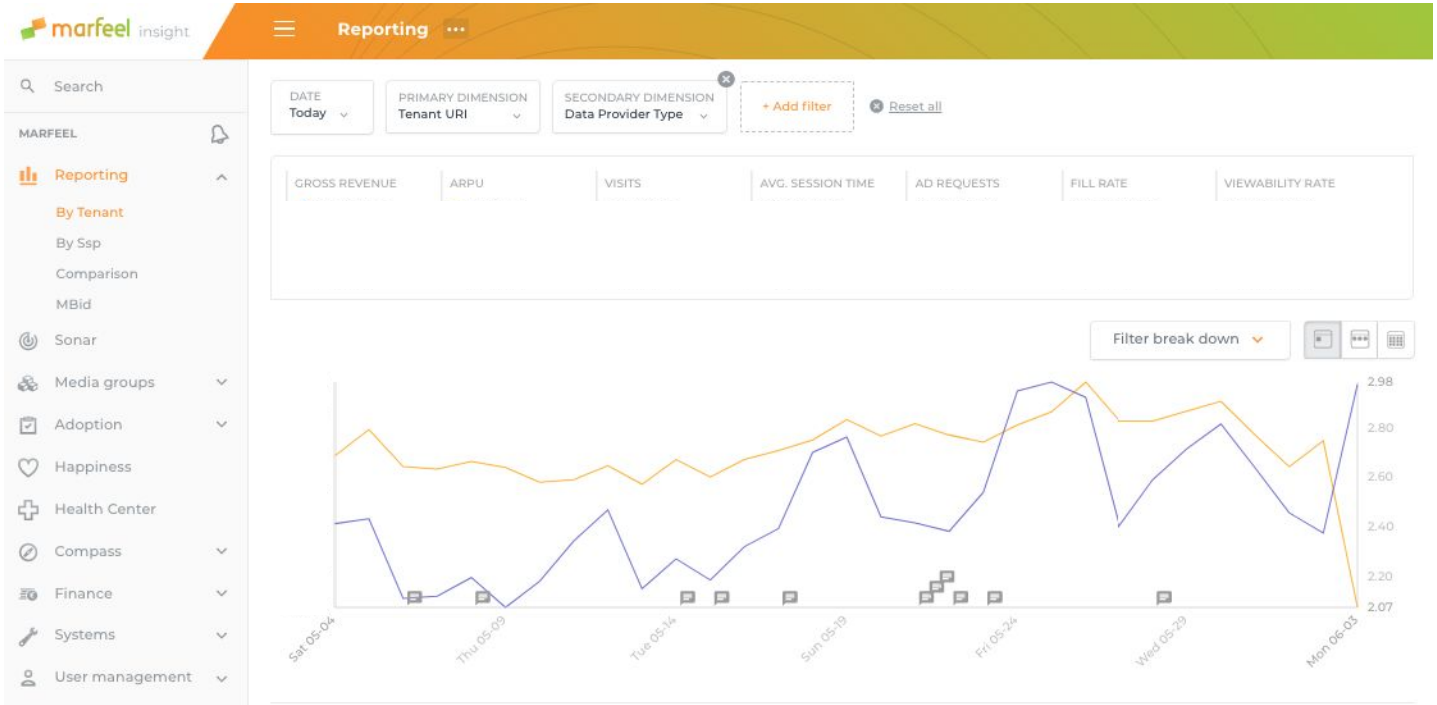
*...& **more** to come*

Tonnes of data, each month.

- **700M** Visits
- **1.75Bn** Page views
- **4 Bn** Ad requests
- **20Tb** of logs data

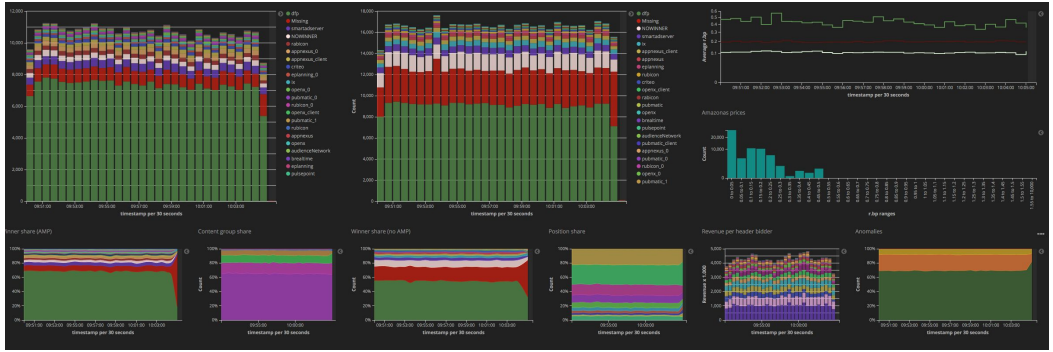


Analytics


marfeel insight
Reporting
Search
MARFEEL
Reporting
By Tenant
By Ssp
Comparison
MBid
Sonar
Media groups
Adoption
Happiness
Health Center
Compass
Finance
Systems
User management
DATE
Today
PRIMARY DIMENSION
Tenant URI
SECONDARY DIMENSION
Data Provider Type
+ Add filter
Reset all
GROSS REVENUE
ARPU
VISITS
AVG. SESSION TIME
AD REQUESTS
FILL RATE
VIEWABILITY RATE
Filter break down
2.98
2.80
2.60
2.40
2.20
2.07
Sat 05-04
Thu 05-09
Tue 05-14
Sun 05-19
Fri 05-24
Wed 05-29
Mon 06-03

Monitoring

SYS_nginx_consumer_tenant_5XX_errors_tooMany (11 active)	
SYS_nginx_consumer_tenant_stale_item_tooMany (21 active)	Tenant m.eldiariodechihuahua.mx has ads.txt but doesn't contain Marfeel correct data ~~CSA-1066~~
SYS_nginx_consumer_tenant_stale_section_home_tooMany (39 active)	
SYS_nginx_consumer_tenant_stale_section_tooMany (15 active)	
DA_cluster_requests_outsideClusters_last4hours (6 active)	
GTB_JS_MSG (2 active)	
HEIS_NOTIFICATIONS_subscribers_not_increasing (2 active)	Tenant economiahoy.mx has ads.txt but doesn't contain Marfeel correct data ~~CSA-1063~~
MPI_CTR_drop_yday_vs_lastWeek (1 active)	
MPI_collector_down (1 active)	
MPI_traffic_not_marfeelized (5 active)	
ALOT_mbidder_timeout_above20Percent_last1hour (0 active)	
CDN_Mobile_First_Detector_Tenant_Above_Transition_Threshold_last_week (0 active)	Tenant prima.fr has ads.txt but doesn't contain Marfeel correct data ~~CSA-1062~~
CDN_Mobile_First_Detector_Tenant_Below_Threshold_last_week (0 active)	
DA_adExchange_executors_enabledForInactiveAdExchange (0 active)	
DA_adExchange_queue_failedActions_isAbove1Percent_last1hour (0 active)	



Some great achievements so far...

 A single source of truth

 Monitoring & alerts on most KPIs

 **A data-driven culture** (to some extent)

“We don’t know much about our tenants”
“I cannot count articles published per day”
“We could segment by tenants’ attributes such as vertical, content type (news/evergreen), keywords/tags, topics (ML), wordcount, images, video, etc.”
“We could create audiences”

“Not enough flexibility”
“I’d like to create my own visualizations and dashboards”
“I need different granularities”
“I’d like to cross this with content and tenant data”
“Cannot compare tenants or YoY”
“Cannot export data”
“Activation and QBR reports are very limited”

“I cannot join collections, nor cross them with other data”
“A big proportion of this data is not being used”
“We cannot look at yearly trends because there’s no historical”
“I cannot perform complex operations (weighted averages, running totals, etc.)”
“The tools are dictating the events modeling”

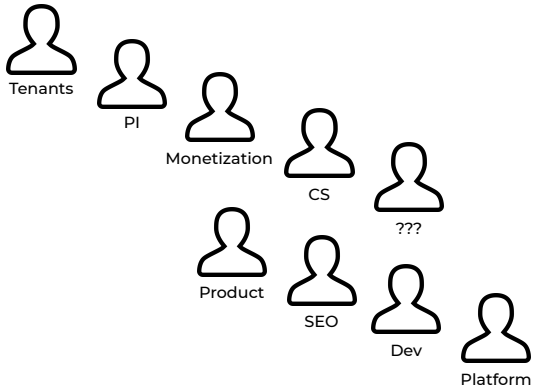
What’s a Data-driven culture?

Five building blocks of a data-driven culture

“Having clean, high-quality data, from a central source, and with clear metadata, is ineffective if staff can’t access it”

Carl Anderson, Michael Li

1. Single source of truth
2. Data dictionary
3. Broad data access
4. Data Literacy
5. Data-driven decision making



… means NO BARRIERS

- Technical (DWH Modeling & SQL)
- Functional (Business knowledge)

Assumptions

- **We can only report to our tenants *Traffic Metrics* consistent with their own data (*Google Analytics*)**
- **We can only report to our tenants *Revenue Metrics* from *SSPs & ADX***

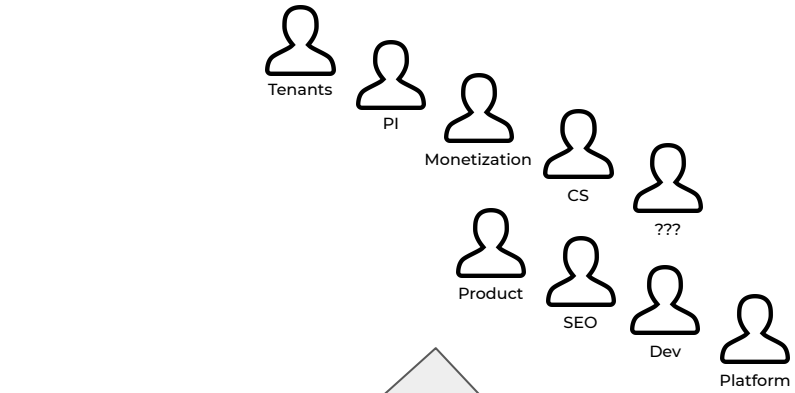
- These source provide limited granularity
- Granular, accurate data requires access to paid tools (such as Google AdManager Premium, GA 360, etc.) whose cost is prohibitive at our scale

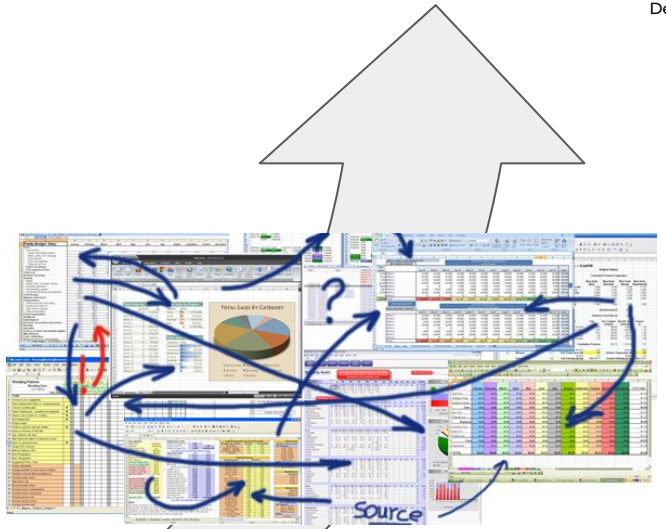
Implications

- **High-level, accurate data** → fit for *reporting* but not for *analytics*
- **Granular, approximated data** → fit for *analytics* but not for *reporting*
- *Exploiting the available data currently requires such a degree of* ***technical and business knowledge*** *that's unreasonable to expect from our stakeholders*

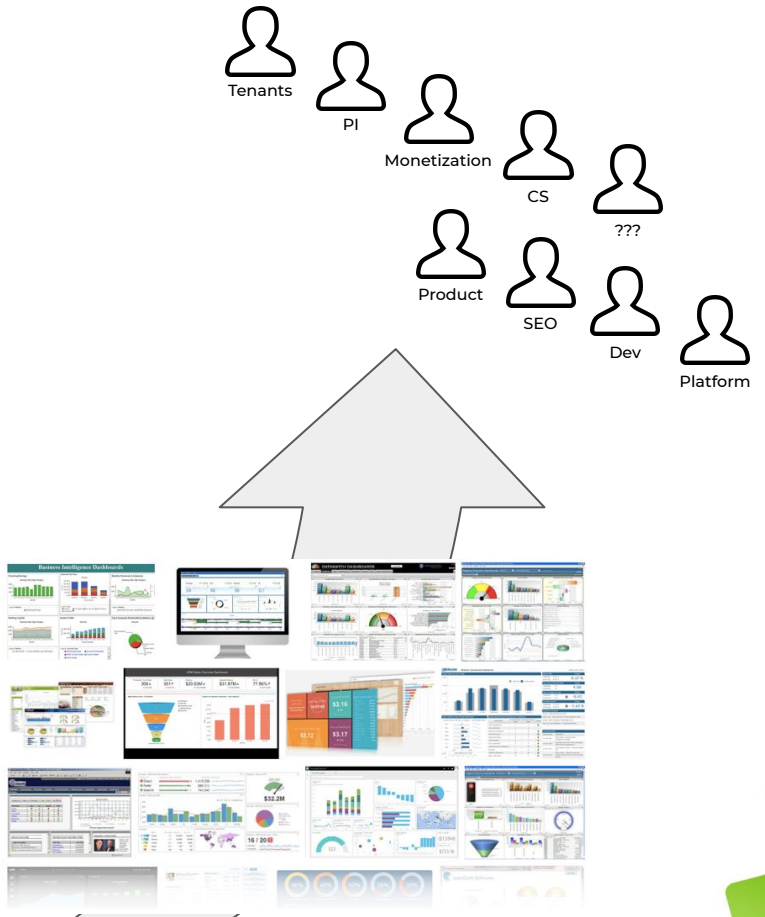
What to do?

Option # 1 A troop of data Monkeys





Option # 2 An army of BI developers



Option # 3 Technology

- A logical layer in which to store table relationships and business rules
- Flexible access to the DWH data (writes SQL for you)

BI Layer - Tech Comparison

	Tableau	Qlikview	Looker
Latency	Low (in-memory)	Low (in-memory)	Mid (from DB)
Loading & Pre-processing	Required	Required	Not Required
Maintenance	Mid (if managed)	Mid (if managed)	Low (centralized)
Development & Deployment time	High (dashboards)	High (dashboards)	Low (data model only)
Logical Model	N	N	Y
SQL Engine	N	N	Y
ETL Layer	Y	Y	Not required
OLAP Layer	Y	Y	Limited
Visualization capability	High	Mid/High	Mid
Mobile Adaptiveness	Low	Mid	High
Learning Curve (Developers)	High	High	Mid
Learning Curve (Business Users)	High	High	Low
Flexibility	Low	Low	High
Price	Mid	Mid	Mid
Pros	Fast (In-memory) OLAP Layer Advanced Visualization	Fast (In-memory) OLAP Layer	**True Self-Service** Embeds business logic Only one model to build Restricts data interactions Unlimited scaling
Cons	**Rigid** **Requires Dashboards Development** **Limited self-exploration** Limited scaling		Not as powerful **Slower** (Relies mostly on DB)

Looker supported DB

- ~~Amazon Aurora~~
- **Amazon Redshift**
- ~~Apache Spark 1.5+~~
- **Apache Spark 2.0**
- ~~Aster Data~~
- ~~Clustrix~~
- Exasol
- ~~Google BigQuery Legacy SQL~~
- **Google BigQuery Standard SQL**
- ~~Google Cloud PostgreSQL~~
- ~~Google Cloud SQL~~
- ~~IBM Netezza~~
- ~~MariaDB~~
- ~~MemSQL~~
- ~~Microsoft Azure PostgreSQL~~
- ~~Microsoft Azure SQL Data Warehouse~~
- ~~Microsoft Azure SQL Database~~
- ~~Microsoft SQL Server 2005~~
- ~~Microsoft SQL Server 2008+~~
- ~~MySQL~~
- ~~Oracle~~
- ~~PostgreSQL~~
- **PrestoDB**
- ~~Qubole Presto~~
- ~~Qubole Presto Service~~
- ~~SAP HANA~~
- **Snowflake**
- ~~Teradata~~
- ~~Vector~~
- Vertica 7.1+

(Discarded all non-distributed & high-end corporate solutions)

DWH Layer - Tech Comparison

	AWS Redshift	Snowflake	Google Big Query	Clickhouse
Speed	Mid/high	Mid/high	Mid/high	**High (?)**
Maintenance	Mid	Low	Mid	Mid/High
Dynamic resizing	Limited	Y	Y	**N**
Concurrency	Low	**High**	Mid	Mid
Indexes	Sort/Dist Key	Self-tuning	Self-tuning	Sort Key (primary only)
Real-time Ingestion	AWS Kinesis Firehose	Snowpipe	**Y**	**Y**
Complex Types	N	**Json/XML**	**Nested Struct Types**	Array
Join on Array/Nested DS	N (UDF?)	**Y (JOIN on Json/XML)**	Hive-like (Explode)	Hive-like (Explode)
Approximated Calculations	N	N	Y	Y
Transactions	Y	Y	Y	**N**
Replication	N	Y	Y	Y
Fault Resistance	**Backup**	Distributed (replication)	Distributed (replication)	Distributed (replication)
Subqueries	Y	Y	Y	N
Window Functions	Y	Y	Y	**N**
UDF (Python, JS, etc.)	Y	Y	Y	**N**
Connectivity	Extensive	Extensive	Extensive	JDBC/ODBC only
Tableau Connectivity	Y	Y	Y	JDBC/ODBC only
Looker Connectivity	Y	Y	Y	**N**
LogStash Output	S3 only	S3 only	**Y**	**N**
Google Analytics integration	**N**	**N**	**Native**	**N**
Cost	Mid	**Mid/High**	**Mid/High**	**Low**
Pros	Highly Tunable, on AWS, widely adopted, previous experience	Self-tuning, fully elastic, high-concurrency, Json/XML support, cheap storage	Fully managed, linear self-scaling, high-concurrency, Json/XML support, Logstash/GA integration	**Very Fast. Open Source**
Cons	**Storage and computing are coupled** **Not great at handling concurrency**	**Speed? Price?**	**Pricing model**	**High maintenance** Non-standard SQL„ No UDF, No Window Functions, No Looker connectivity

Pros

- AWS, established, widely adopted
- Highly Tunable
- It works

Concerns

- Storage and computing are coupled (Spectrum doesn't quite cut it)
- Not great at handling concurrency
- Didn't evolve much since 2013. Outdated (?)

Google BigQuery

Pros

- Google (strong relationship)
- Fully managed, linear self-scaling, high-concurrency
- Json/XML support, GA integration

Concerns

- Awkward pricing model - Pay per query (flat rates start from 10K per month)
- Quite Hadoop-like. More complex to use?

Pros

- Open Source
- Allegedly very fast
- Some prestigious adopters (E.G. CloudFlare)

Concerns

- High-maintenance (concerns about the *Total Cost of Ownership*)
- Even if it was cheaper, do we need to process SO much data?
- Is it worth the trade off?
  - Non-standard SQL
  - No Subqueries
  - No Analytic functions
  - No UDF
  - No Looker connectivity

Pros

- Separates storage and computing. Storage is cheap
- Handles concurrency very well
- Semi-structured data support (Json, XML)
- Virtual warehouses (pay per usage, predictable cost)
- Lots of advanced, handy functionality

Concerns

- Will it be fast enough?
- Cost

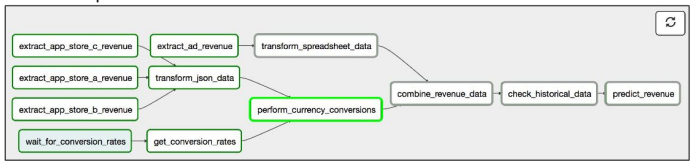
Apache Airflow

- Open Source, great community
- Well established
- Extremely versatile
- Powerful
- Distributed

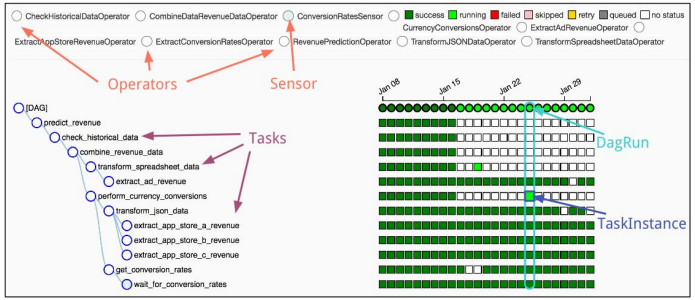
...& it could do some of the heavy-lifting

if Snowflake turned out to be too slow, or expensive

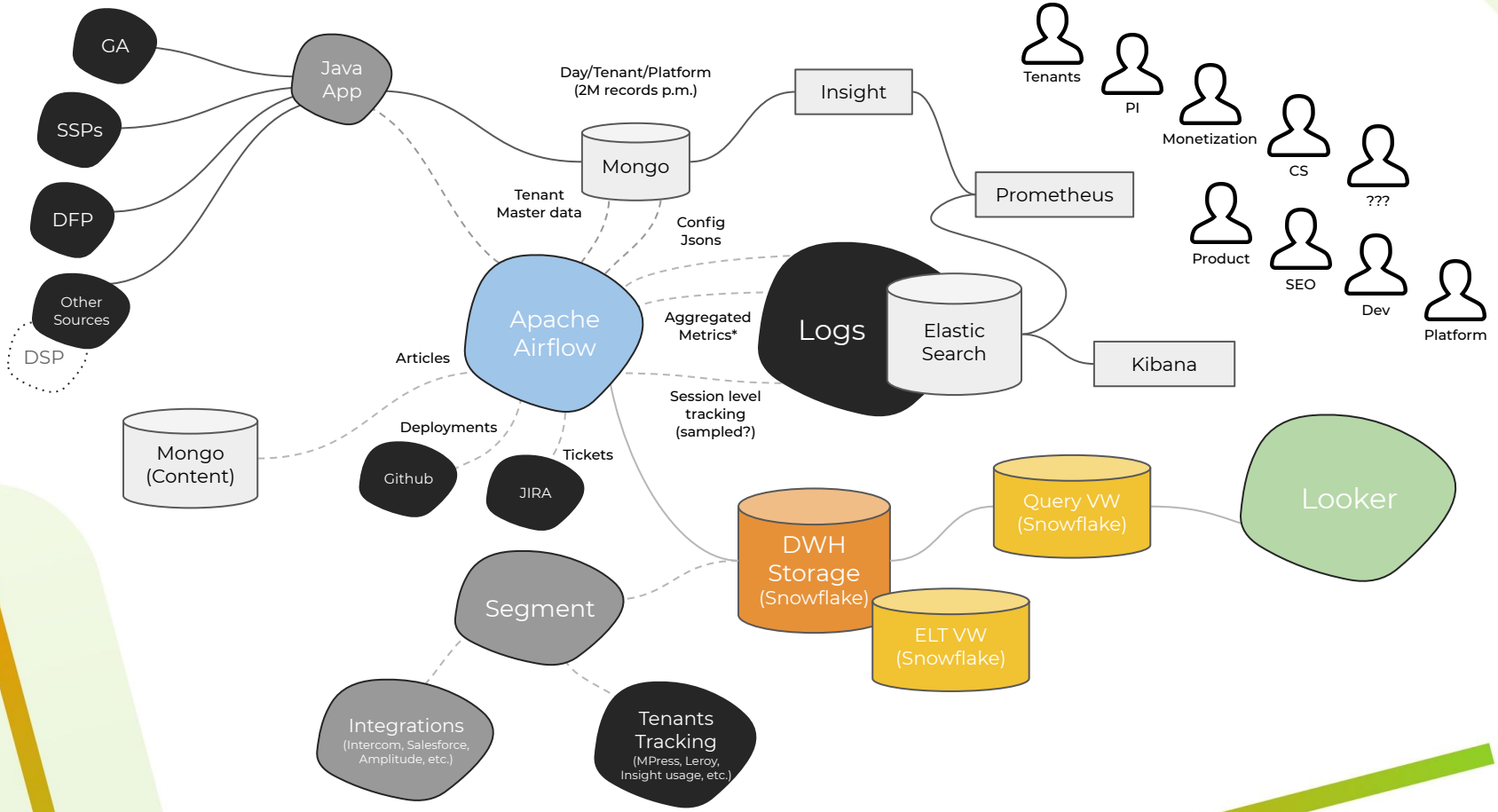
UI **DAG** Graph View



UI **DAG** Tree View



Hi-level data architecture proposal



Proof of Concept (PoC) - Objectives

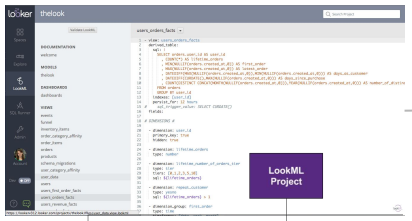
Before proceeding, we must validate that:

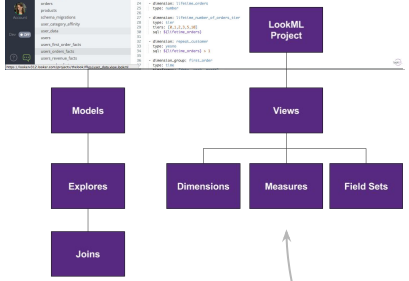
- **Looker** did a good job at:
  - Resolving the complexity and fragmentation of our data sources
  - Removing (most) barriers to Data Accessibility & Literacy by modeling the required technical and business knowledge into its logical layer
- **Snowflake** could scale (both in terms of performance and costs)

Proof of Concept (PoC)

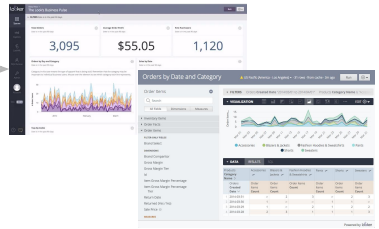
looker

CIS Consulting





Looker Model
Logical layer



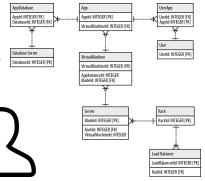
CEO
Product
Monetization
PI

marfeel

Looker dashboard and data exploration POC

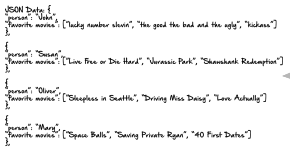
snowflake

Data Warehouse
Analytics data model



Data
PI

Staging and processing (ELT)
Raw Json Data

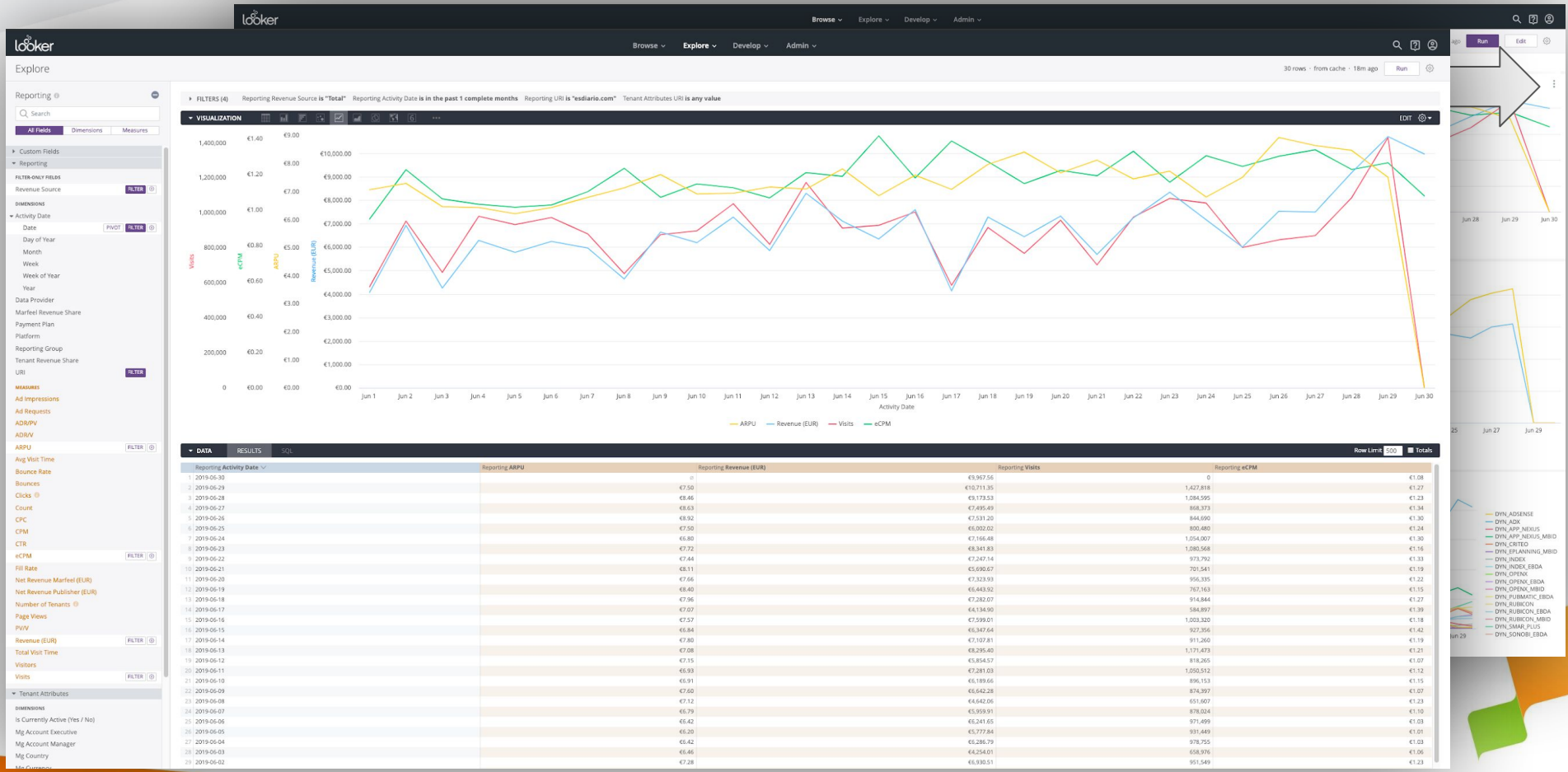


DoubleClick for Publishers

elastic
Pageviews & Visits (varys)

marfeel insight
GA Connect, SSP, ADX, etc.

THE RESULT


looker
Browse
Explore
Develop
Admin
Explore
Reporting
Search
All Fields
Dimensions
Measures
Custom Fields
Reporting
FILTER-ONLY FIELDS
Revenue Source
DIMENSIONS
Activity Date
Date
Day of Year
Month
Week
Week of Year
Year
Data Provider
Marfeel Revenue Share
Payment Plan
Platform
Reporting Group
Tenant Revenue Share
URI
MEASURES
Ad Impressions
Ad Requests
ADR/PV
ADR/V
ARPU
Avg Visit Time
Bounce Rate
Bounces
Clicks
Count
CPC
CPM
CTR
eCPM
Fill Rate
Net Revenue Marfeel (EUR)
Net Revenue Publisher (EUR)
Number of Tenants
Page Views
PV/V
Revenue (EUR)
Total Visit Time
Visitors
Visits
Tenant Attributes
DIMENSIONS
Is Currently Active (Yes / No)
Mg Account Executive
Mg Account Manager
Mg Country
FILTERS (4)
Reporting Revenue Source is "Total"
Reporting Activity Date is in the past 1 complete months
Reporting URI is "esdiario.com"
Tenant Attributes URI is any value
VISUALIZATION
EDIT
Visits
eCPM
ARPU
Revenue (EUR)
Activity Date
ARPU
Revenue (EUR)
Visits
eCPM
DATA
RESULTS
SQL
Row Limit
Totals
Reporting Activity Date
Reporting ARPU
Reporting Revenue (EUR)
Reporting Visits
Reporting eCPM
1 2019-06-30 € €9,967.56 0 €1.08
2 2019-06-29 €7.50 €10,711.35 1,427,818 €1.27
3 2019-06-28 €8.46 €9,173.53 1,084,595 €1.23
4 2019-06-27 €8.63 €7,495.49 868,373 €1.34
5 2019-06-26 €8.92 €7,531.20 844,690 €1.30
6 2019-06-25 €7.50 €6,002.02 800,480 €1.24
7 2019-06-24 €6.80 €7,166.48 1,054,007 €1.30
8 2019-06-23 €7.72 €8,341.83 1,080,568 €1.16
9 2019-06-22 €7.44 €7,247.14 973,792 €1.33
10 2019-06-21 €8.11 €5,690.67 701,541 €1.19
11 2019-06-20 €7.66 €7,323.93 956,335 €1.22
12 2019-06-19 €8.40 €6,443.92 767,163 €1.15
13 2019-06-18 €7.96 €7,282.07 914,844 €1.27
14 2019-06-17 €7.07 €4,134.90 584,897 €1.39
15 2019-06-16 €7.57 €7,599.01 1,003,320 €1.18
16 2019-06-15 €6.84 €6,347.64 927,356 €1.42
17 2019-06-14 €7.80 €7,107.81 911,260 €1.19
18 2019-06-13 €7.08 €8,295.40 1,171,473 €1.21
19 2019-06-12 €7.15 €5,854.57 818,265 €1.07
20 2019-06-11 €6.93 €7,281.03 1,050,512 €1.12
21 2019-06-10 €6.91 €6,189.66 896,153 €1.15
22 2019-06-09 €7.60 €6,642.28 874,397 €1.07
23 2019-06-08 €7.12 €4,642.06 651,607 €1.23
24 2019-06-07 €6.79 €5,959.91 878,024 €1.10
25 2019-06-06 €6.42 €6,241.65 971,499 €1.03
26 2019-06-05 €6.20 €5,777.84 931,449 €1.01
27 2019-06-04 €6.42 €6,286.79 978,755 €1.03
28 2019-06-03 €6.46 €4,254.01 658,976 €1.06
29 2019-06-02 €7.28 €6,930.51 951,549 €1.23
Run
Edit
DYN_ADSENSE
DYN_ADX
DYN_APP_NEXUS
DYN_APP_NEXUS_MBID
DYN_CRITEO
DYN_EPLANNING_MBID
DYN_INDEX
DYN_INDEX_EBDA
DYN_OPENX
DYN_OPENX_EBDA
DYN_OPENX_MBID
DYN_PUBMATIC_EBDA
DYN_RUBICON
DYN_RUBICON_EBDA
DYN_RUBICON_MBID
DYN_SMAR_PLUS
DYN_SONOBI_EBDA

Looking back...

In just a few years...

When	Where	Data Sources	ETL	DWH	BI Layer
2012	**Softonic**	Mostly backend DB	***Pentaho Data Integration***	SQL Server ***+ Hadoop***	***Qlikview + Tableau***
2014	**King**	Mostly tracking events	Java / Jenkins	***Hadoop*** + Exasol *(then BigQuery)*	Qlikview *(then Looker)*
2016	**Typeform**	Mostly backend DB	Pentaho Data Integration *(then Apache Airflow)*	Redshift	Tableau *(then Looker)*
2019	**Marfeel**	**Multiple APIs & microservices**	**Apache Airflow** (Python)	**Snowflake**	**Looker**
20??	?				

Thank you.

Alessandro Pregnolato
Head of Data