From (big) data mess to data as an innovation enabler

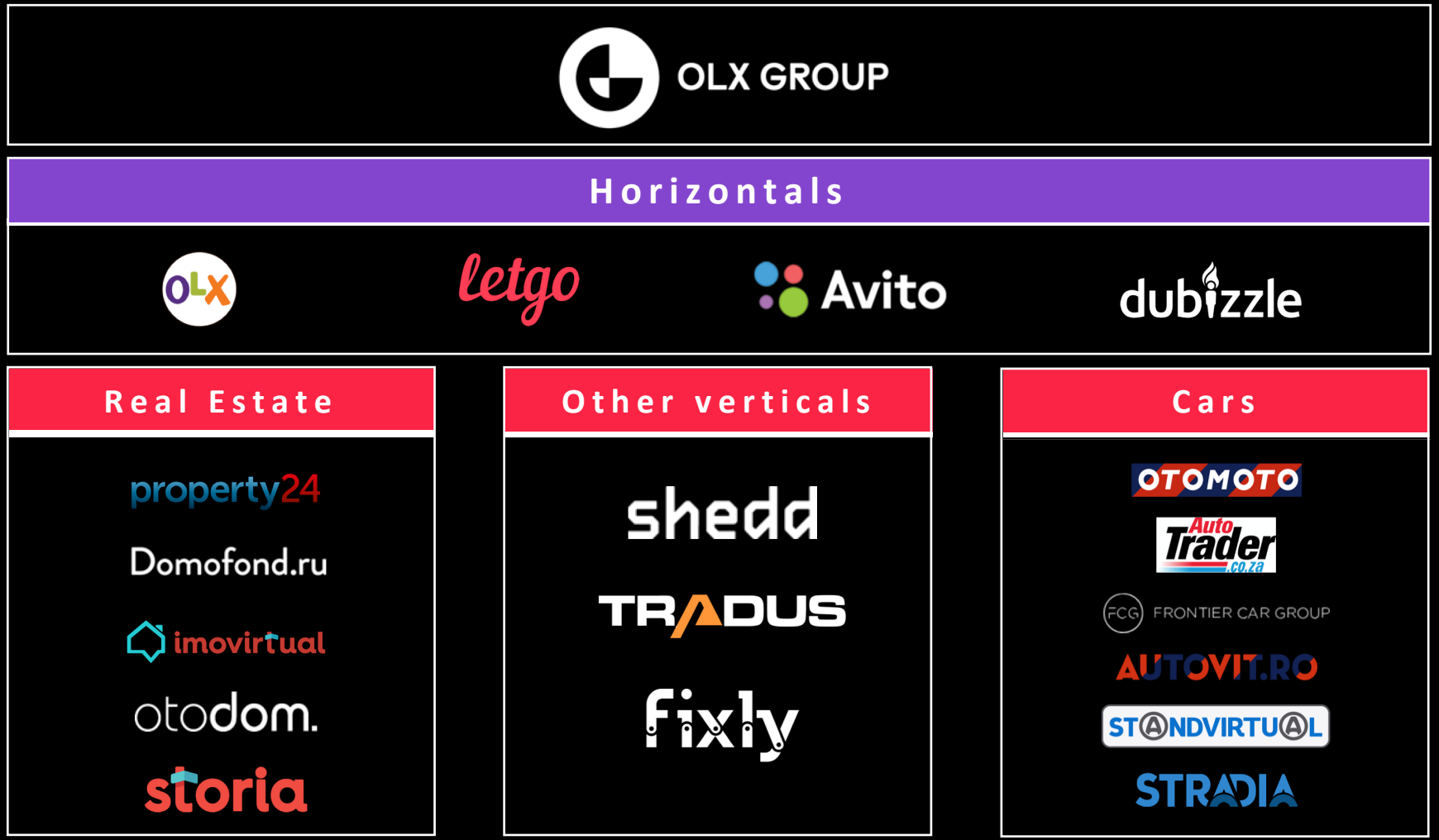
Olaf Zschiedrich – CTO OLX Group – GOTO Berlin 2018

NASPERS

Founded 1915
South-Africa
Market cap: $100B

Delivery Hero


SELLER WINS
BUYER WINS
S
UPGRADE
2ND HAND GOODS
BIG PROGRESS
B
GREAT DEAL
Search in OLX
Vehicles
Property
Electronics & Computers
Home, Garden & Tools
Recommended for you
SIZE US -10, CHUKKA BOOTS
~N~ 17 000
Fiat Panda 1.2l 2006
~N~ 600 000


OLX GROUP
Horizontals
OLX
letgo
Avito
dubizzle
Real Estate
property24
Domofond.ru
imovirtual
otodom.
storia
Other verticals
shedd
TRADUS
fixly
Cars
OTOMOTO
Auto Trader .co.za
FCG FRONTIER CAR GROUP
AUTOVIT.RO
STANDVIRTUAL
STRADIA

43 Countries

35 Offices

+5,000
Employees

+350M
MAU

+4B
Events/Day


Halleluja!
We are a data-driven company!

Sellers

Item Started and Reactivated - Weekday

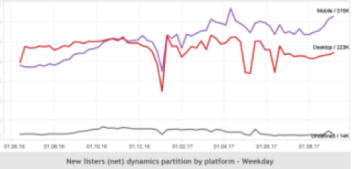
Item Started and Reactivated LF - Weekday

Mobile Item started (net) dynamics - weekday

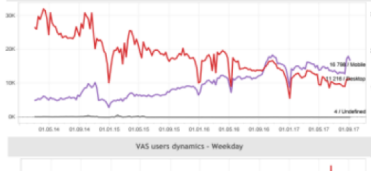
Item started (net) dynamics - weekday

Listers (users who starts or reactivates items) - Weekday

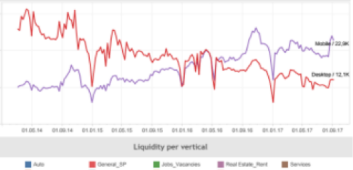
Listers(net) dynamics - Weekday



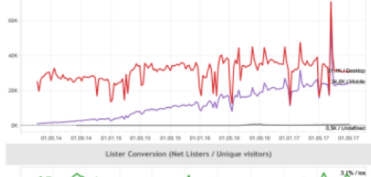
New listers (net) dynamics - Weekday



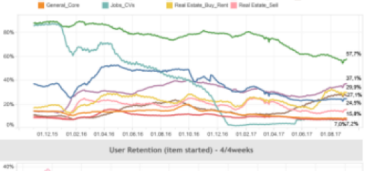
New listers (net) dynamics partition by platform - Weekday



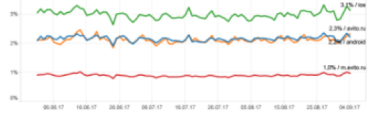
VAS users dynamics - Weekday



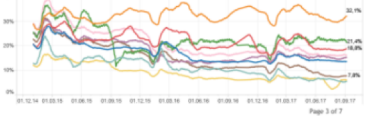
Liquidity per vertical



Lister Conversion (Net Listers / Unique visitors)

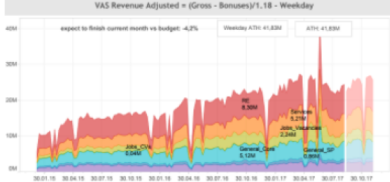


User Retention (item started) - 4/4weeks

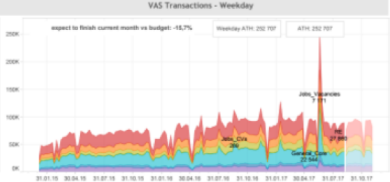


Monetization

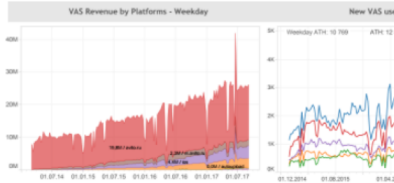
VAS Revenue Adjusted = (Gross - Bonuses)/1.18 - Weekday



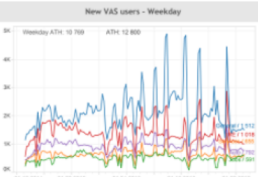
VAS Transactions - Weekday



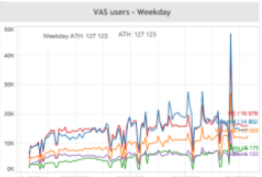
VAS Revenue by Platforms - Weekday



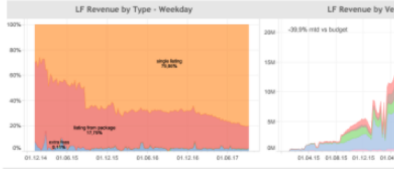
New VAS users - Weekday



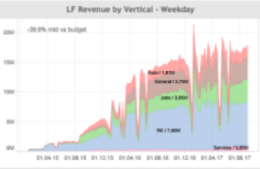
VAS users - Weekday



LF Revenue by Type - Weekday

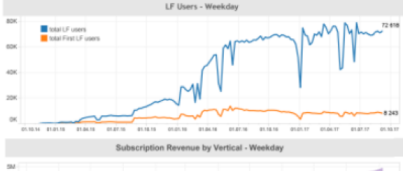


LF Revenue by Vertical - Weekday

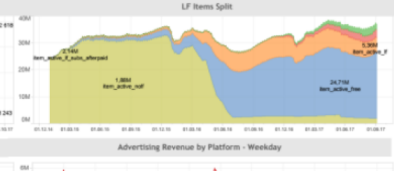


LF Revenue (RUB) - platform distribution - Weekday

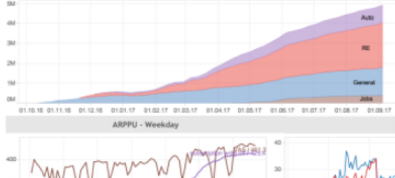
LF Users - Weekday



LF Items Split



Subscription Revenue by Vertical - Weekday



Advertising Revenue by Platform - Weekday

ARPPU - Weekday

RPAI

Native Advertising % - Weekday

"Contact and Promo revenues' share in total advertising revenue.


BenQ
BenQ

Disc

42

is broken

Do I launch a new car portal in Mexico?

CEO

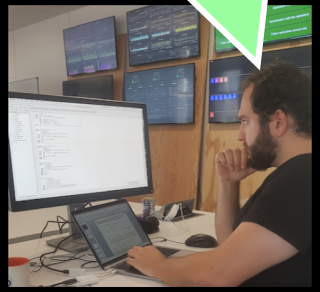
Shall I invest more in online or in offline marketing?

GM

How is CS agent #253 performing?

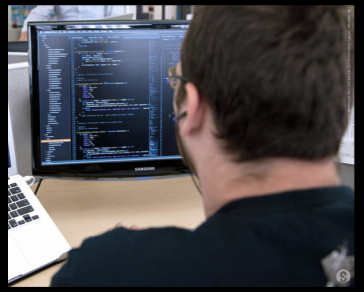
CS Manager

Can I predict which listings have the highest probability to sell?



Business Analyst

Are we still online?



DevOps Engineer

months ← urgency → seconds

Which data points are really influencing your decisions?

‘Give everybody the data that he or she needs’

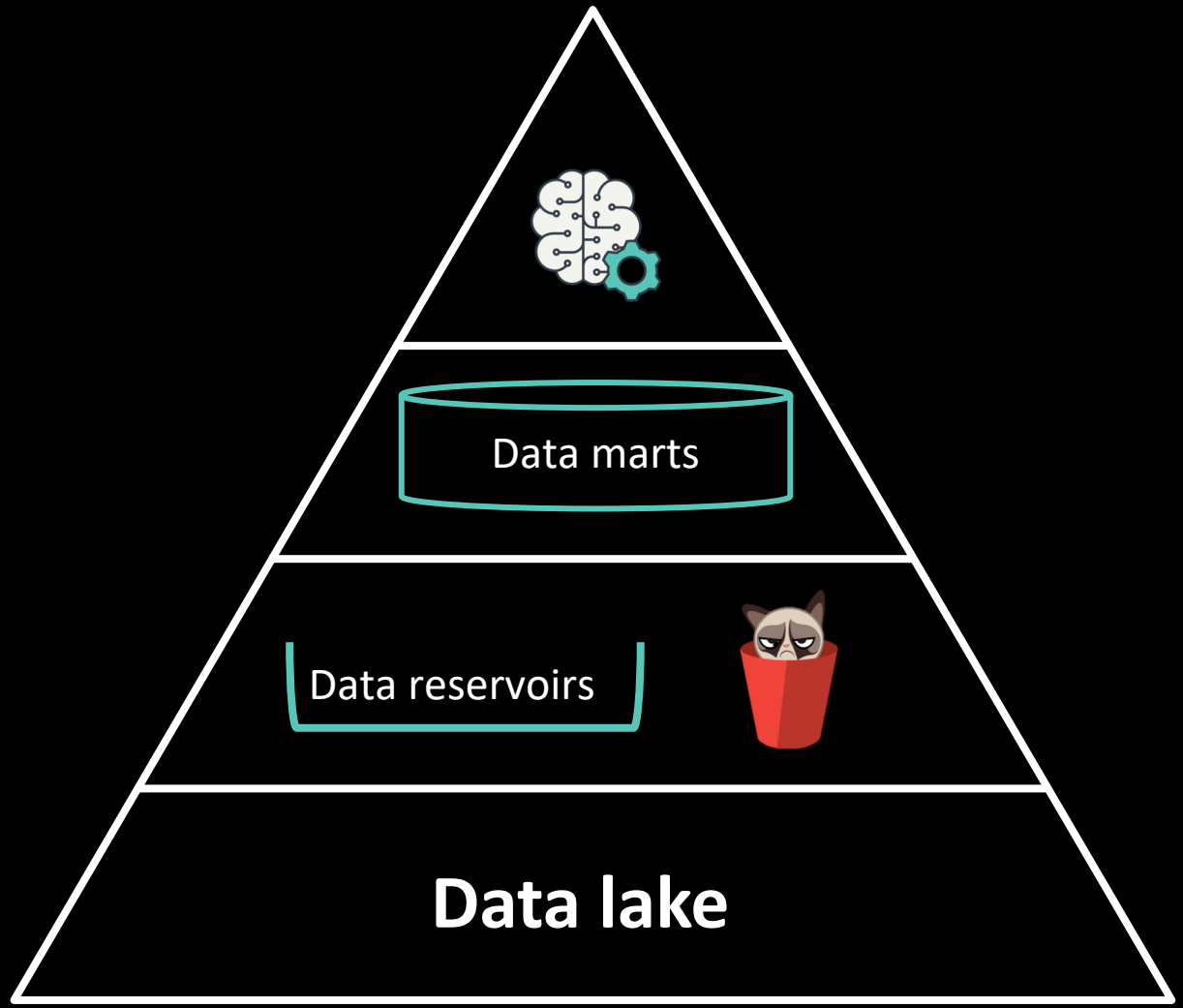
(but also not much more)

Product innovation (AI/ML)

BI and reporting

Data democratization

Data collection



#1 Data Collection


S3
400 TB
after Compression

4B events / day
~ 2T events total

LIVESYNC

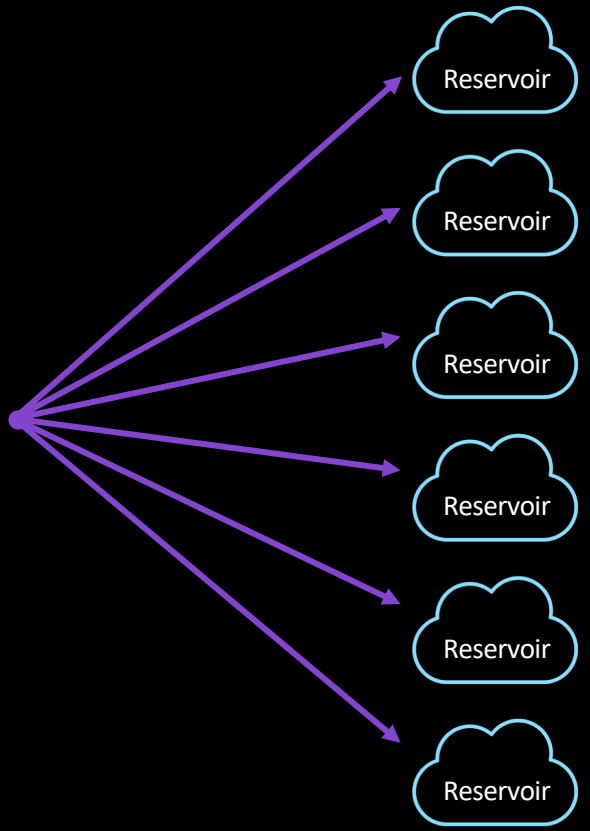
1.3M listings / day
~ 1B listings total

#2 Data Democratization


Amazon S3
Amazon EMR
AWS Glue


Data lake

Catalog


Reservoir
Reservoir
Reservoir
Reservoir
Reservoir
Reservoir

Welcome to Catalog

What do you need to do?

Browse available data sources

There are 59,233 fields in 3,539 sources and 26 reservoirs.

100% of fields are classified

All sources | All reservoirs

Manage your data sources

You are not producing any data.

→ Start publishing data

Overall health

- ✔ No owners with bad sources
- ✔ No subscribers consuming unclassified fields
- ✔ No sources have unclassified fields
- ⚠ 16 of 26 reservoirs have risky subscriptions

(subscriptions using personal fields not anonymized)

Subscription requests

You have 387 requests to review

Review

My reservoirs

Fully anonymised reservoir *2 subscriptions*

- ✔ All subscriptions use only classified fields!
- ⚠ 1 subscription are using personal fields not anonymized

Help

Introduction to Catalog
How to classify data
How to subscribe to data
How to approve subscriptions

Top Owners

- By health
- By amount of sources

Top Reservoirs

- By health
- By amount of subscriptions

Top Subscribers

- By health
- By amount of subscriptions

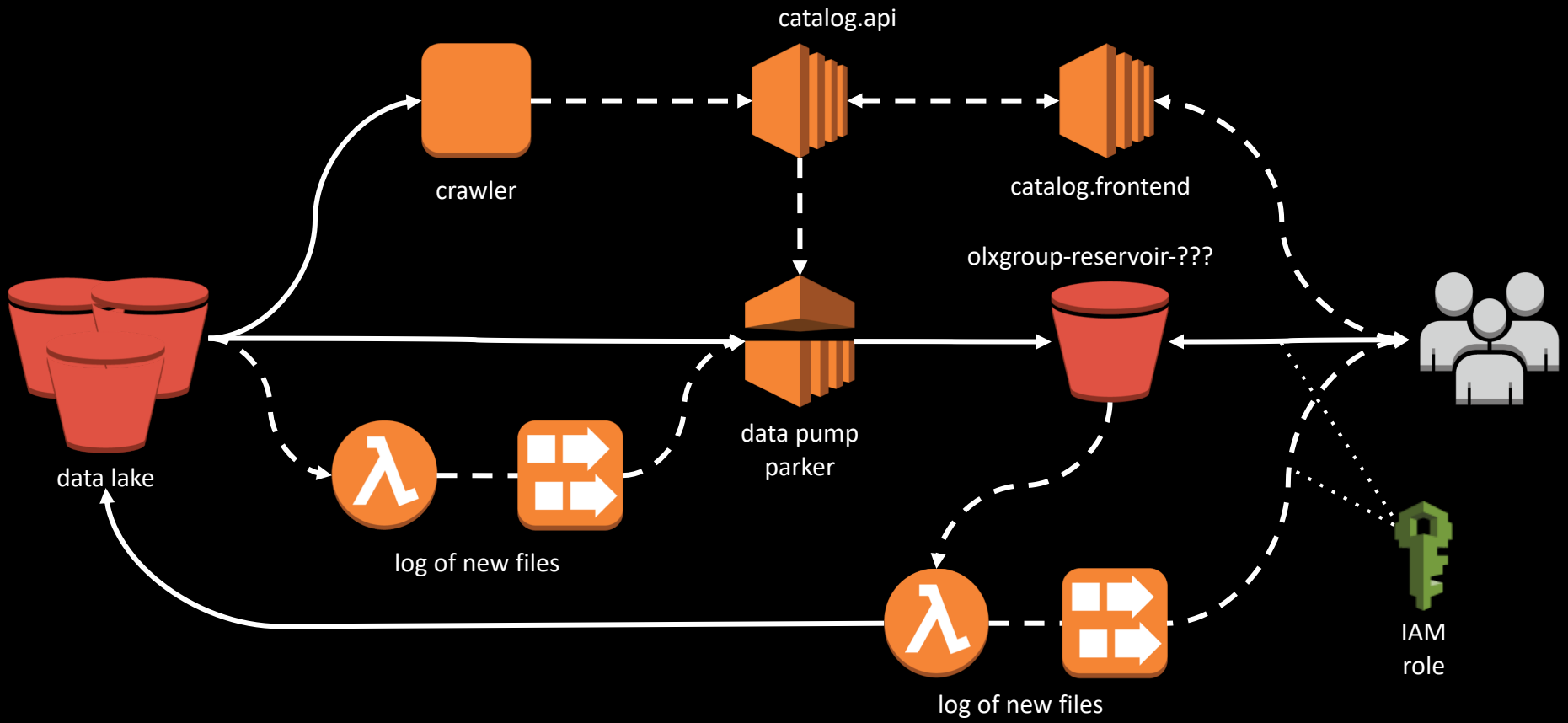
Reservoir Structure – S3 Bucket

/in

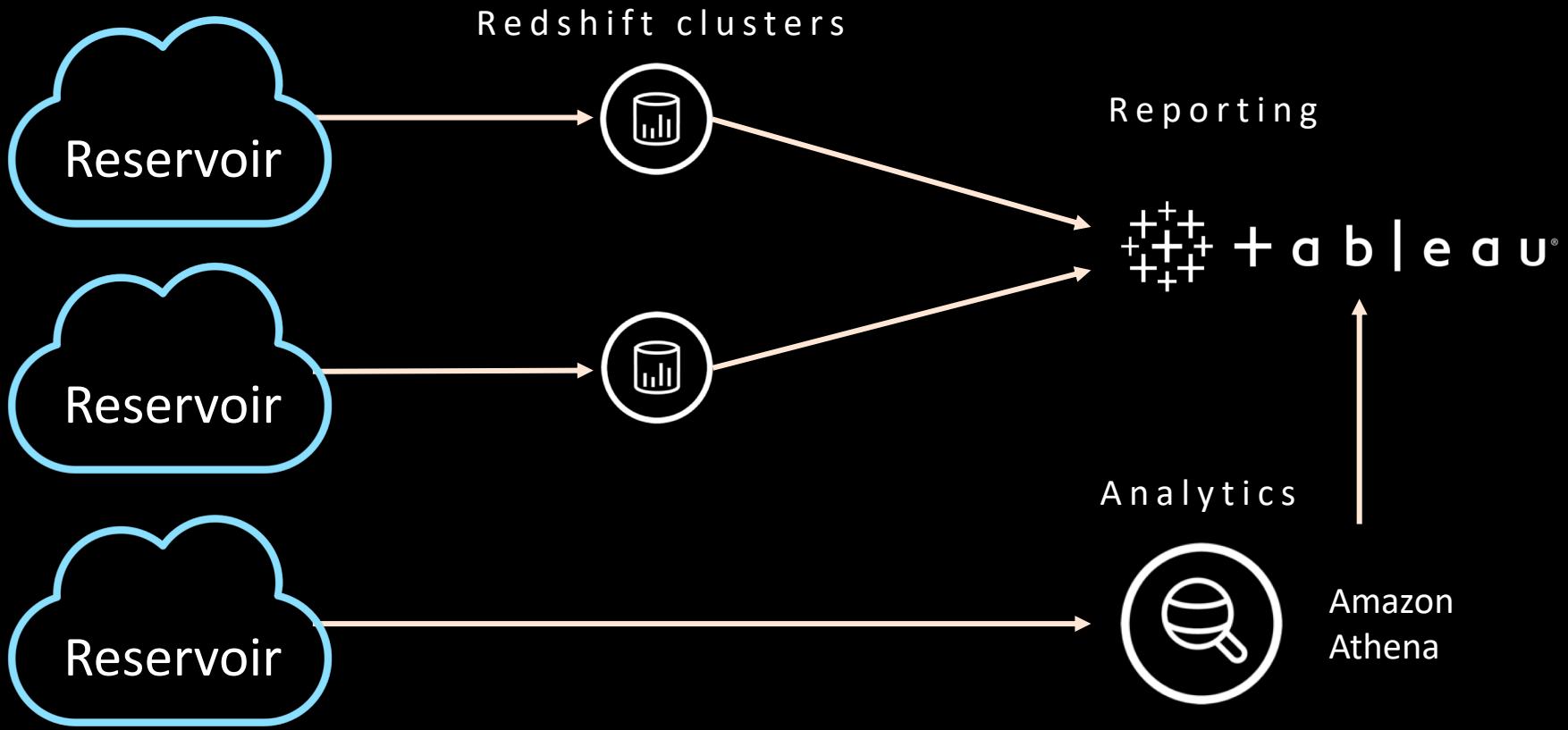
/out

/parquet

/tmp


catalog.api
crawler
catalog.frontend
olxgroup-reservoir-???
data lake
data pump
parker
log of new files
log of new files
IAM
role

#3 BI & Reporting


Redshift clusters
Reservoir
Reservoir
Reservoir
Reporting
+ a b | e a u
Analytics
Amazon
Athena

#4 Applied Intelligence

REVOLUTION

LOVE

Amazon
Sagemaker

Amazon
Kinesis

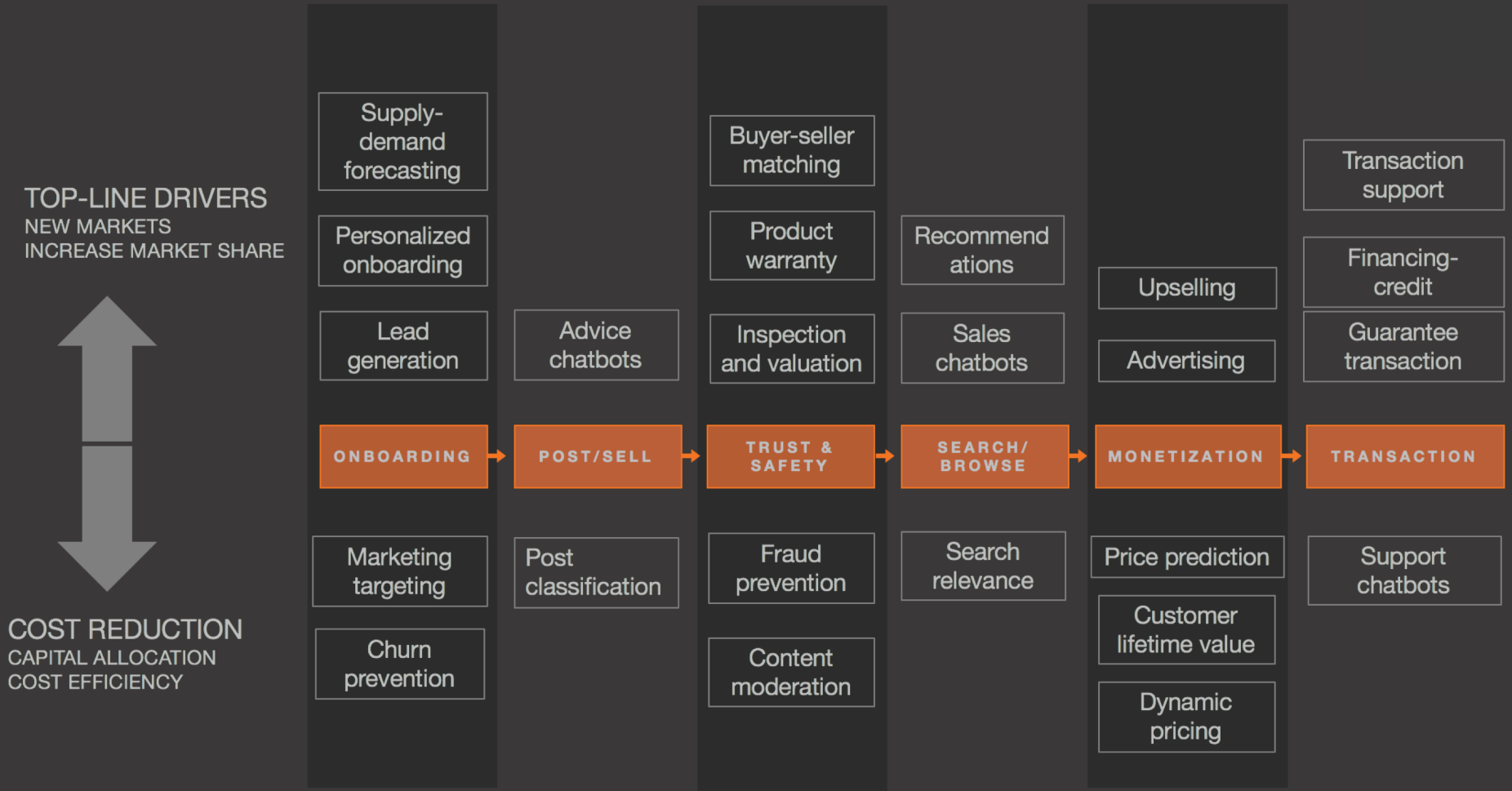
Image recognition

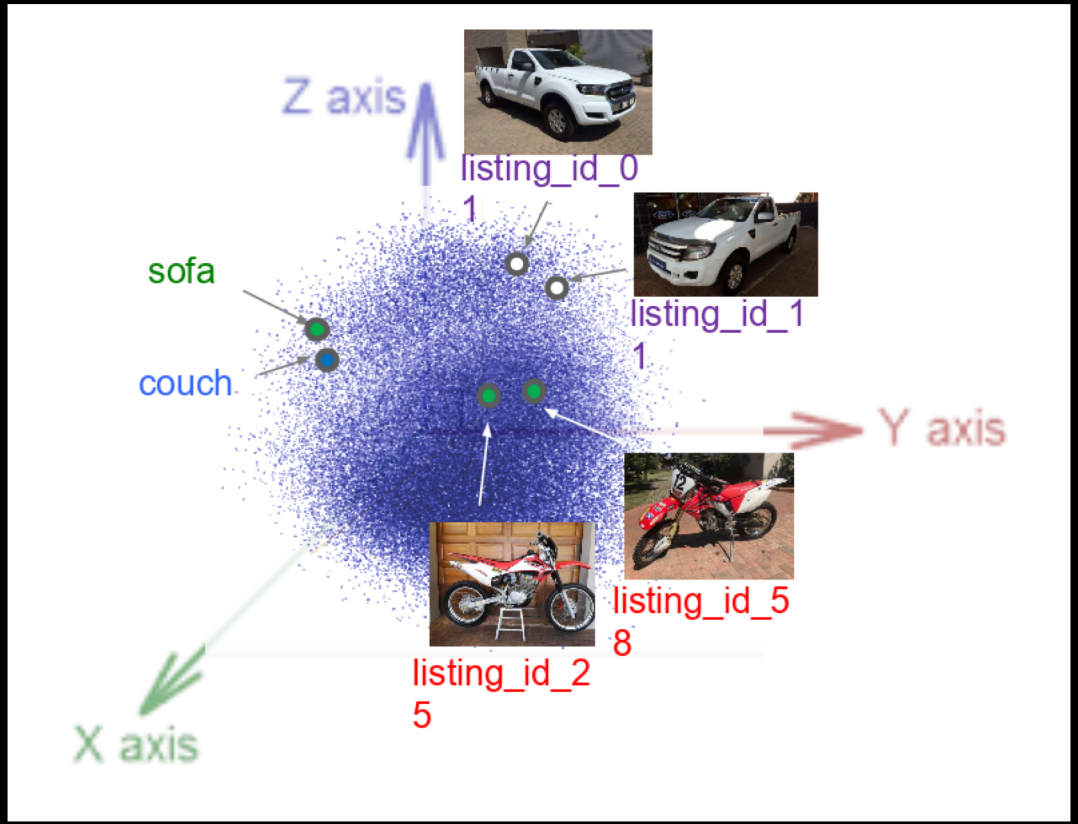
Fraud detection

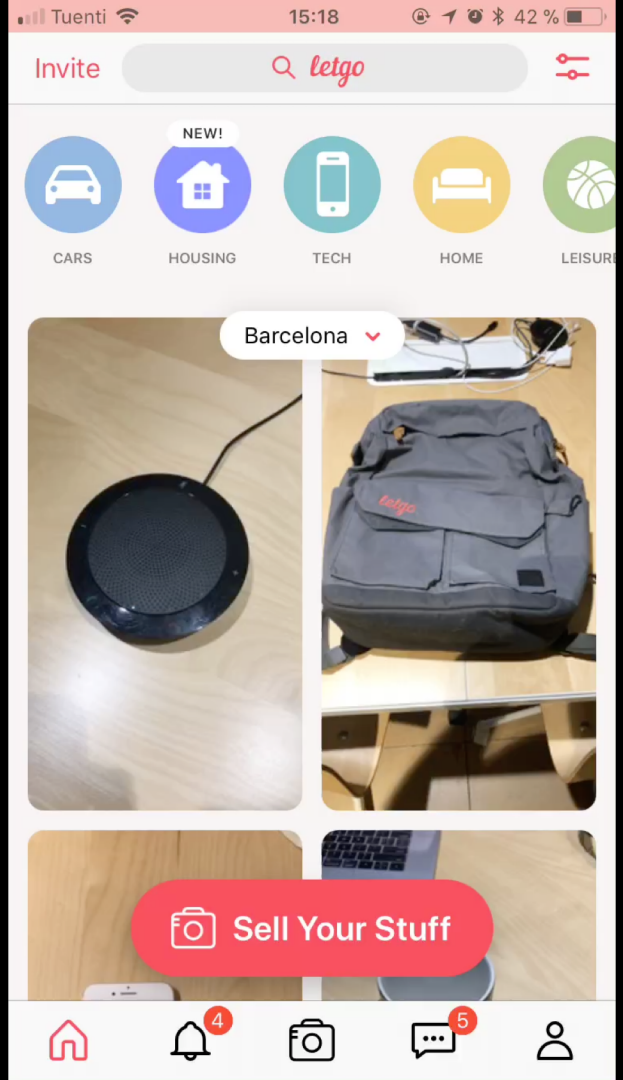
Moderation

Recommendations

And more…


TOP-LINE DRIVERS
NEW MARKETS
INCREASE MARKET SHARE
COST REDUCTION
CAPITAL ALLOCATION
COST EFFICIENCY
Supply-demand forecasting
Personalized onboarding
Lead generation
Marketing targeting
Churn prevention
Advice chatbots
Post classification
Buyer-seller matching
Product warranty
Inspection and valuation
Fraud prevention
Content moderation
Recommend ations
Sales chatbots
Search relevance
Upselling
Advertising
Price prediction
Customer lifetime value
Dynamic pricing
Transaction support
Financing-credit
Guarantee transaction
Support chatbots
ONBOARDING
POST/SELL
TRUST & SAFETY
SEARCH/ BROWSE
MONETIZATION
TRANSACTION


Z axis
listing_id_01
listing_id_11
sofa
couch
Y axis
listing_id_58
listing_id_25
X axis


Tuenti
15:18
42 %
Invite
letgo
NEW!
CARS
HOUSING
TECH
HOME
Barcelona
Sell Your Stuff
4
5

WHAT'S NEXT ?

THANK YOU