



Como garantimos qualidade de dados de Tracking sem perder agilidade

The Developer's Conference BH 2019



Tiago Montalvão

Big Data Developer @ OLX



Beatriz Vaz

Data Product Manager @ OLX

A OLX é o maior site de compras e vendas do Brasil

600K



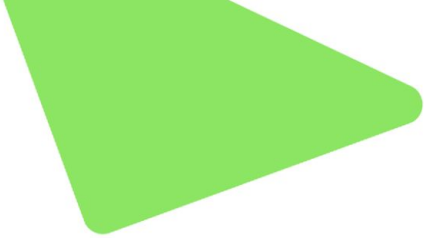
brasileiros
anunciam pela primeira vez todos os meses

+2M

de **vendas** por meio da plataformas por mês, com mais de 50 vendas por minuto

27%

das vendas são realizadas em **menos de 24 horas**



**Grande número de
usuários ativos gera um
grande volume de dados**



+ **600TB** de dados armazenados no Data Lake



+ **3000** queries realizadas por dia, gerando um consumo na ordem de **1PB** por dia



+ **800M** de eventos comportamentais gerados por dia




Dados são fundamentais para a OLX



Cultura data
informed



Estratégia user
centric



Incentivo a entregas
ágeis



Cultura de
experimentação



Dados nos ajudam a
conhecer melhor nossos
usuários e medir seus
comportamentos

TRACKER

Ferramenta de coleta de dados estatísticos sobre tráfego de usuário, nos permitindo entender melhor quem fez o quê em nossas plataformas



Algumas
ferramentas
do mercado



Google
Analytics



SNOWPLOW

mixpanel
● ● ●

Ferramentas adotadas na OLX



Google
Analytics



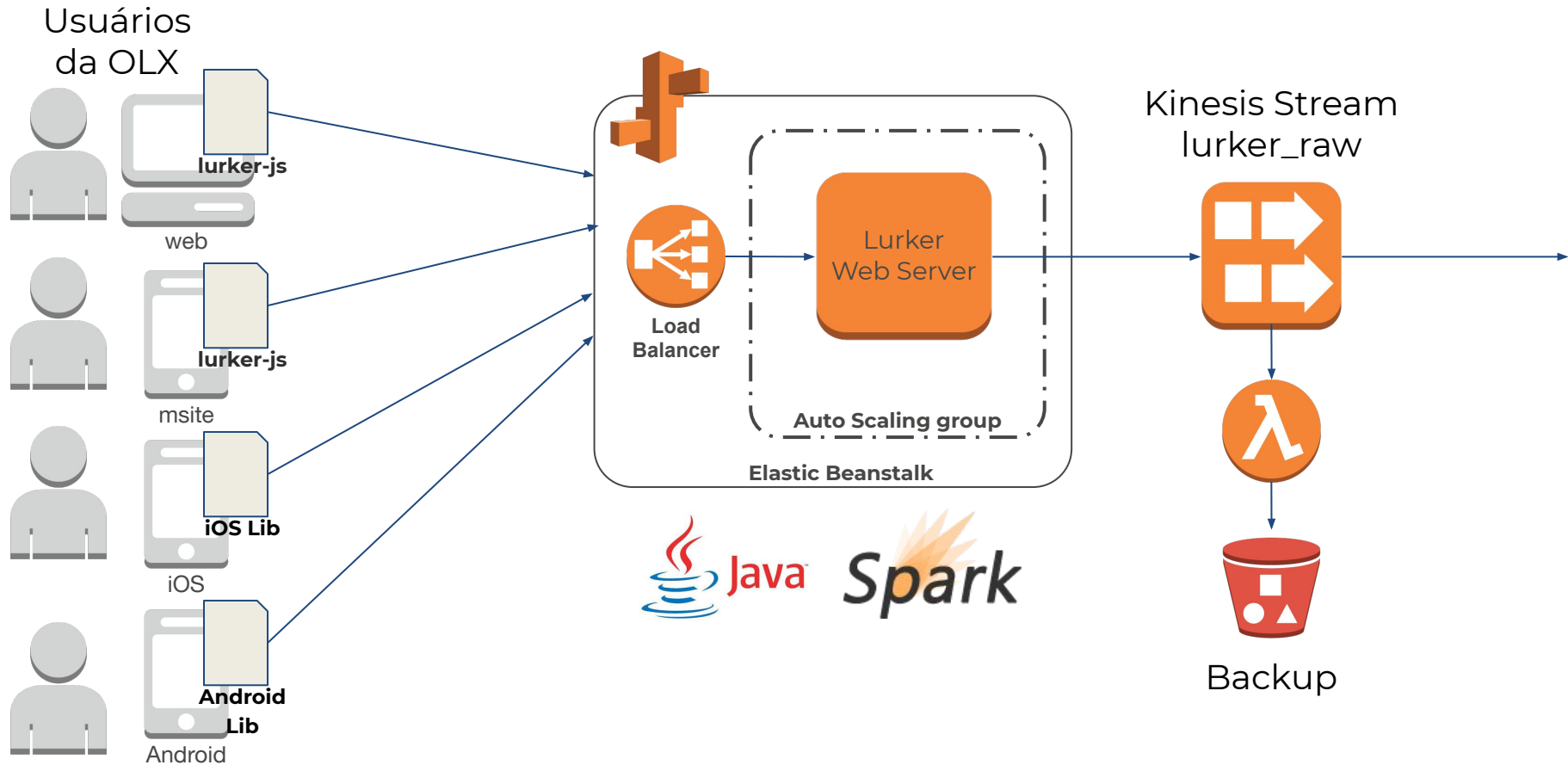
LURKER

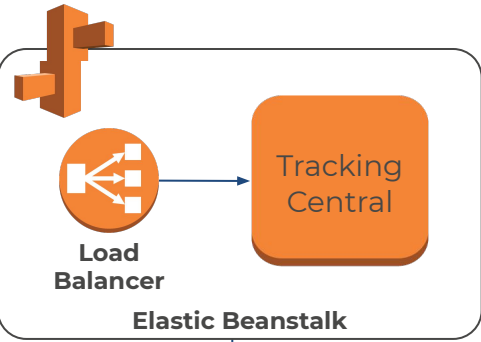
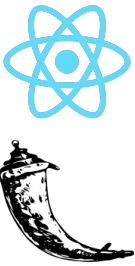
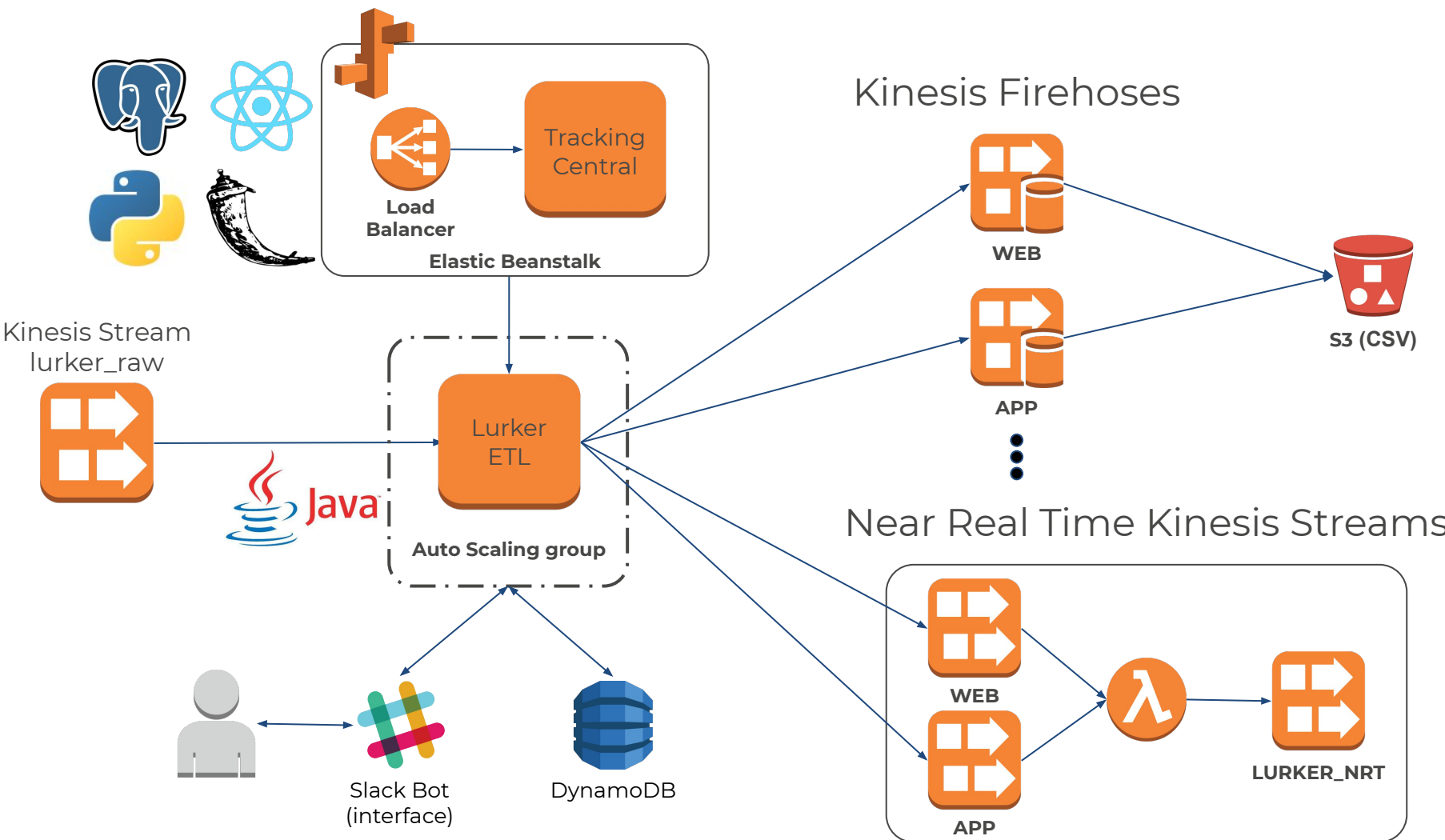
LURKER

- ▲ Ferramenta desenvolvida pela OLX
- Maior flexibilidade na implementação de novas tags
- ▲ Maior controle dos dados capturados
- Integrado com plataforma de experimentação
- ▲ Melhor custo benefício para nosso volume de dados



Arquitetura do **LURKER**





Kinesis Firehoses



WEB



APP



S3 (CSV)

Kinesis Stream
lurker_raw



Auto Scaling group

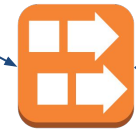


Slack Bot
(interface)



DynamoDB

Near Real Time Kinesis Streams



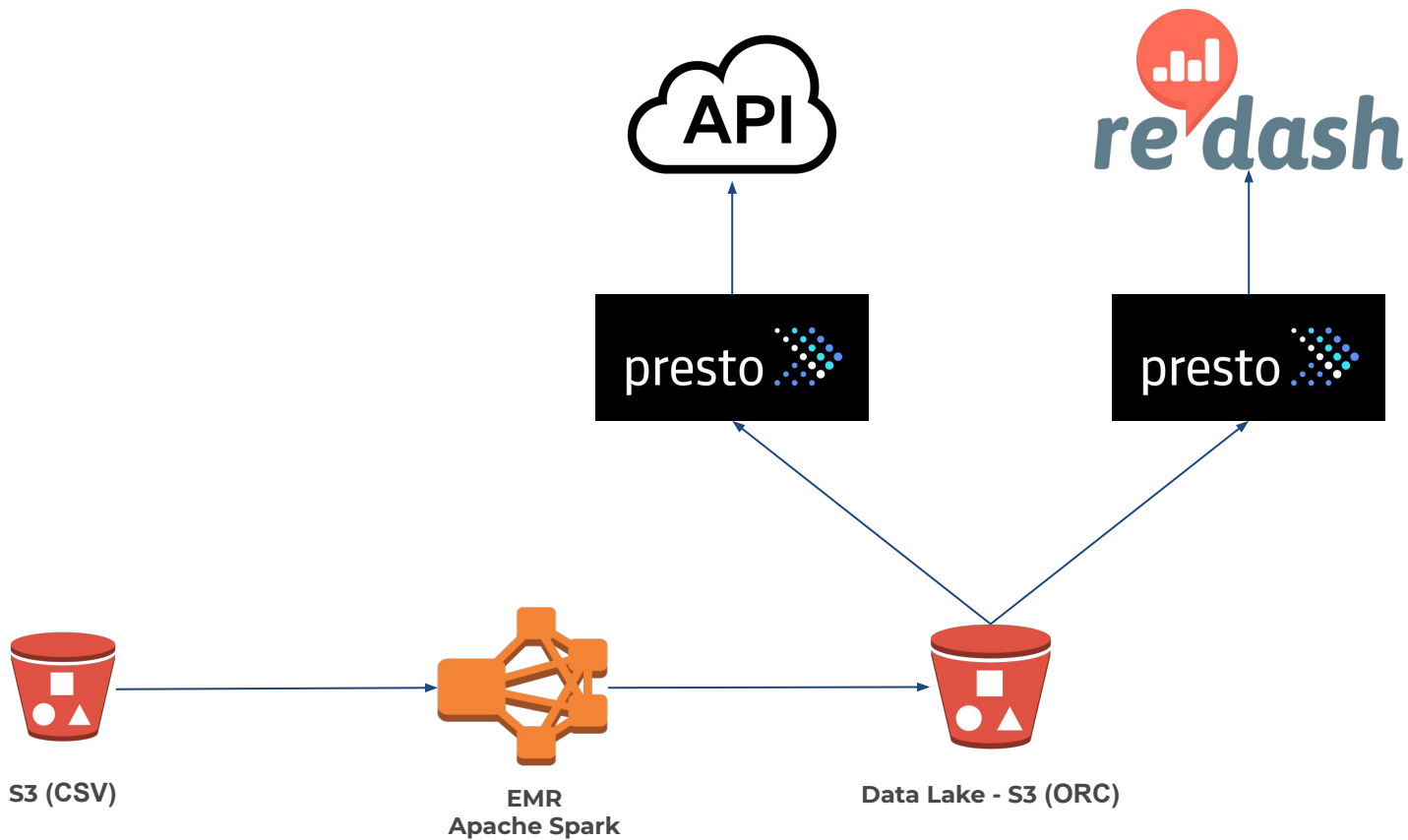
WEB




APP



LURKER_NRT




Como funcionava Tracking na OLX



Cada squad dono do próprio tracking



Maior liberdade e agilidade



Cadastro de eventos constantemente desatualizado



Menor governança

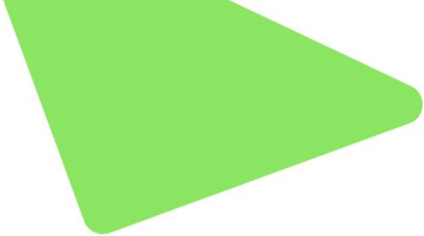


70%



**Eventos do Lurker em
produção não estavam
documentados
corretamente**



**Problemas na
qualidade de dados
eram descobertos até
meses depois,
gerando perda de
informação**



Criamos um **squad
dedicado para garantir a
eficiência e governança
de **tracking****



Proposta 1:

- Centralizar implementação em Tracking
- Tracking pode virar gargalo e perder agilidade
- Squads têm liberdade de escolher as tecnologias que querem implementar

Proposta 2:

- Tracking oferece ferramentas que garantam a governança dos dados
- Reforçar uma cultura de dados na OLX, mudança de mindset

Proposta 1:

- Centralizar implementação em Tracking
- Tracking pode virar gargalo e perder agilidade
- Squads têm liberdade de escolher as tecnologias que querem implementar

Proposta 2:

- Tracking oferece ferramentas que garantam a governança dos dados
- Reforçar uma cultura de dados na OLX, mudança de mindset

Hoje, como funciona Tracking na OLX



Cada squad dono do próprio tracking



Maior liberdade



Time de Tracking responsável pela visão geral



Mais governança de dados de Tracking

FERRAMENTAS DE GOVERNANÇA DE TRACKING





Solução #1

Tracking Central

Tracking Central

- ▲ Governança dos dados de tracking
 - Documentação dinâmica por meio da validação de eventos cadastrados
 - Autenticação via LDAP
 - Histórico de criação/edição
- Informação do responsável por cada evento
- ▲ Dashboards de monitoramento por evento



NOIS

MP

AUTOMATICO

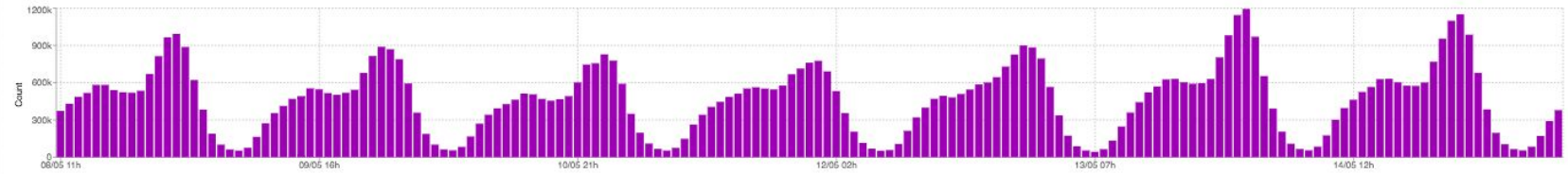
Event Specification Detail

web2_page_view_listing_lead_make

Registered
Before 12/09/2018

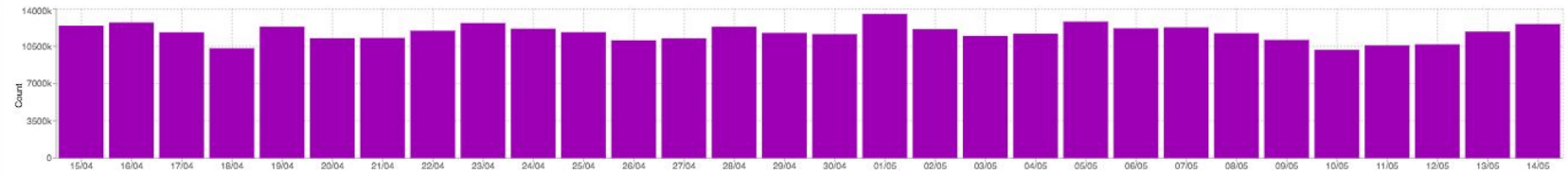
Hourly event count

Last 7 days



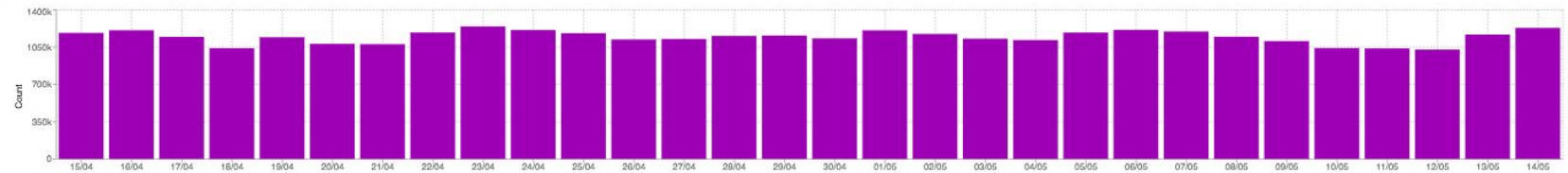
Daily event count

Last 30 days



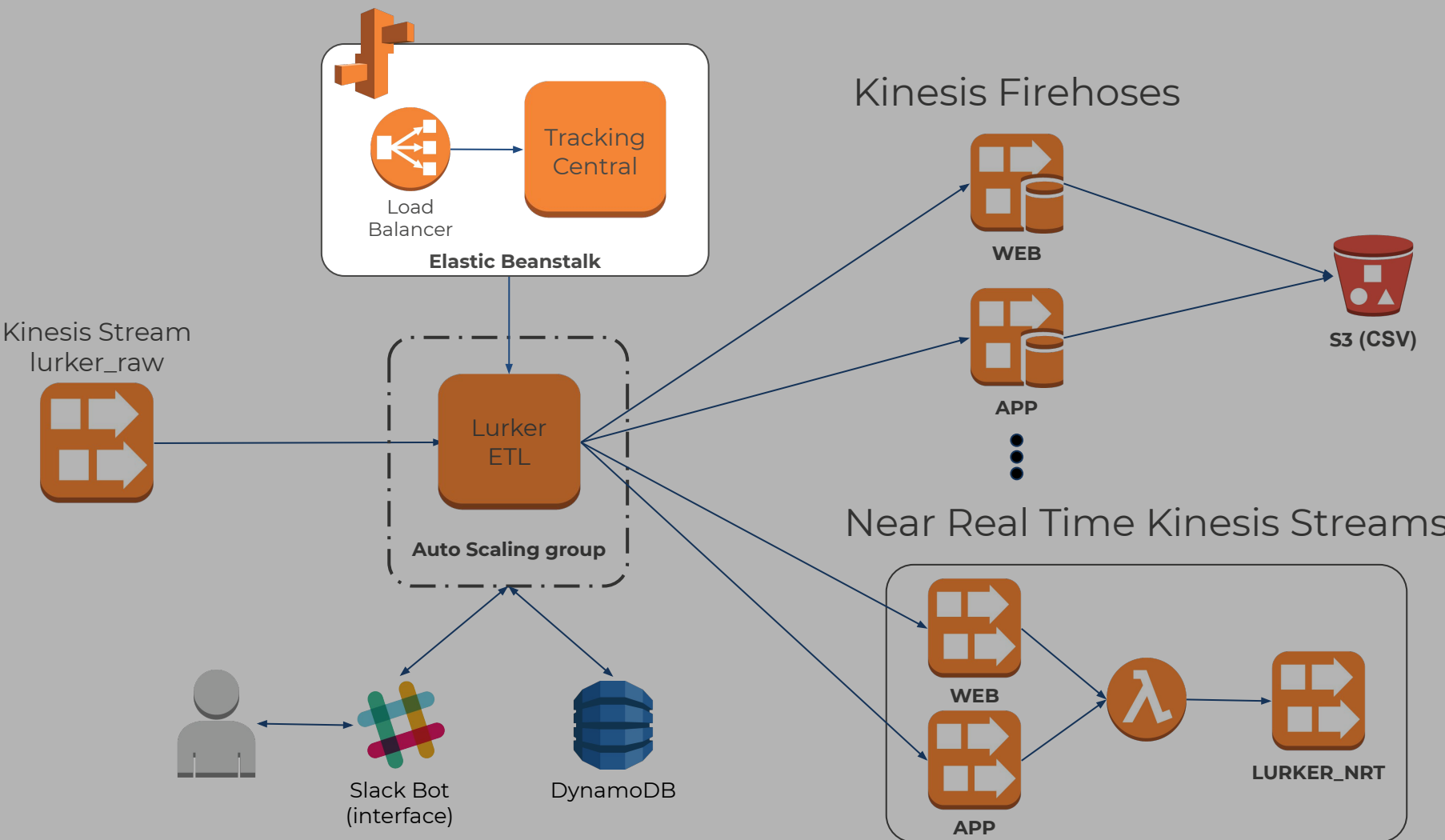
Daily distinct lurker_id count

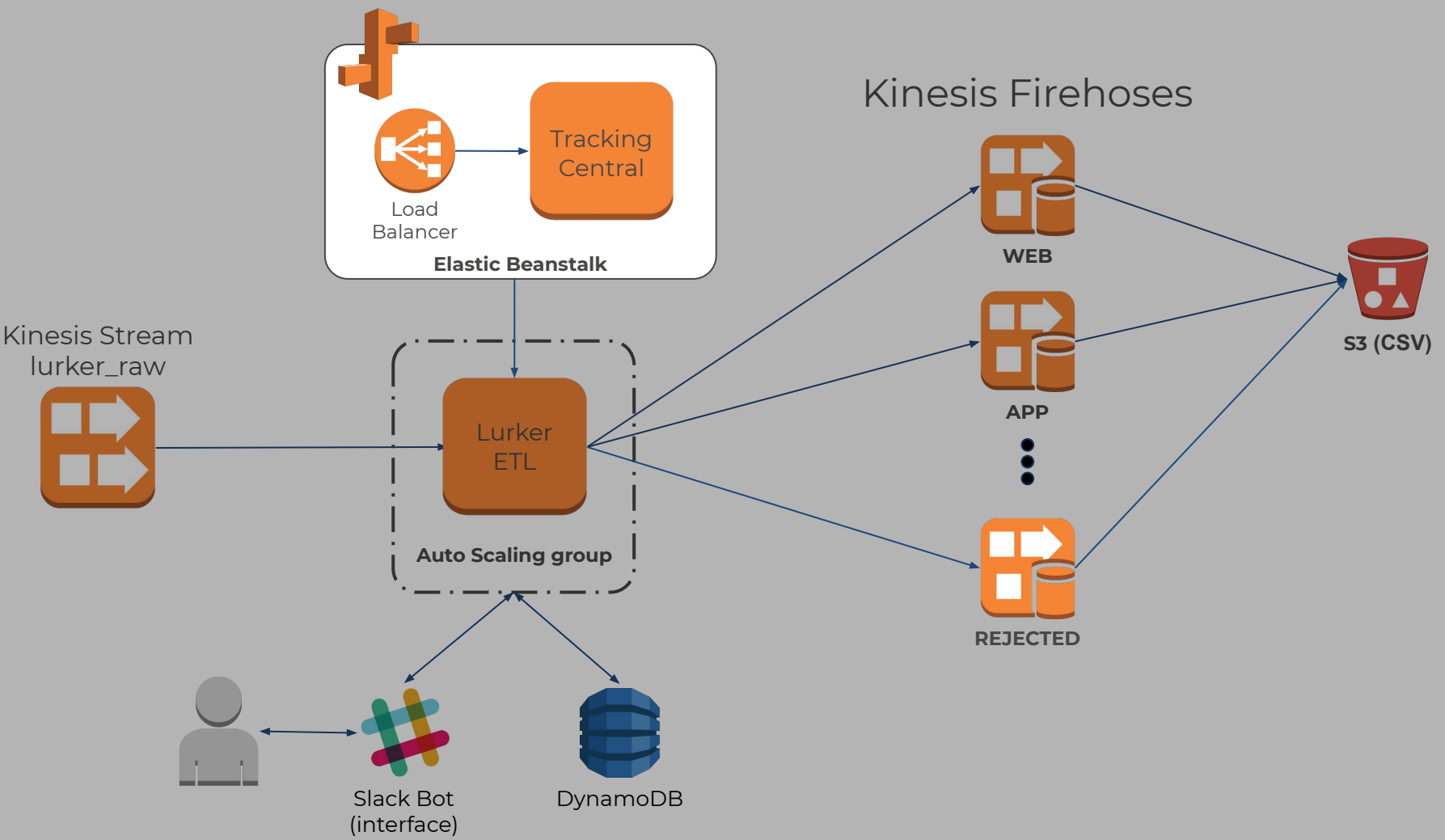
Last 30 days



DELETE

EDIT CLOSE





Números do Tracking Central

~40

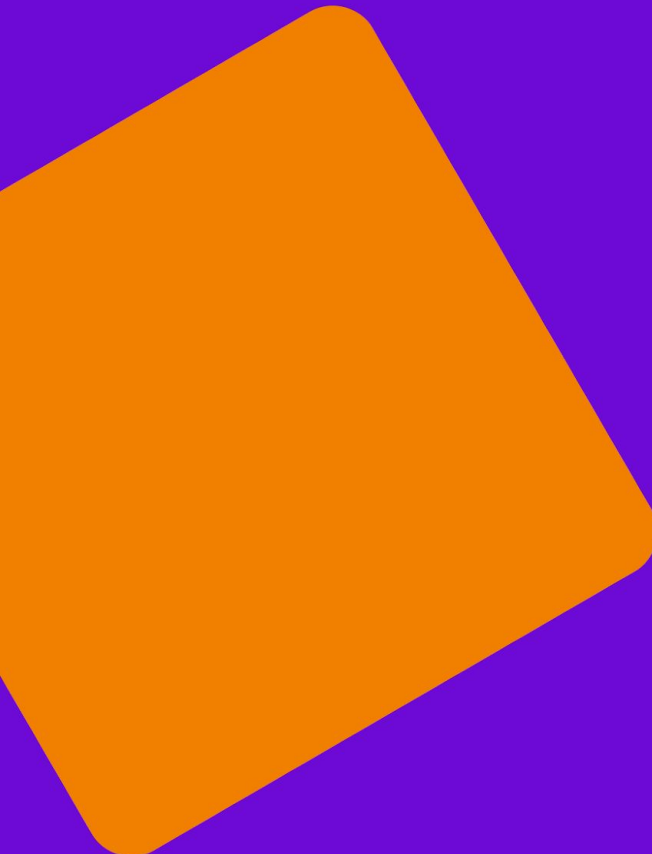
usuários distintos
acessando o
Tracking Central
semanalmente
(WAU)

+2K

eventos cadastrados
no TC referente a
todas plataformas da
OLX, incluindo
produtos internos

10M

eventos **rejeitados**
diariamente



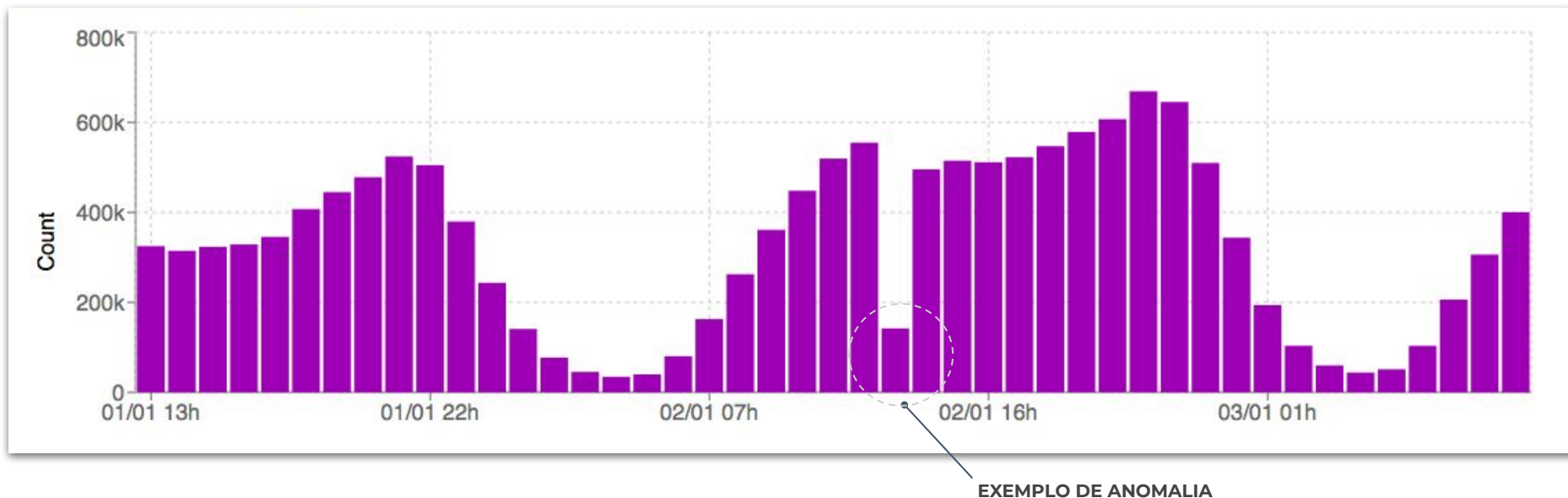
Solução #2

Detecção de

anomalias

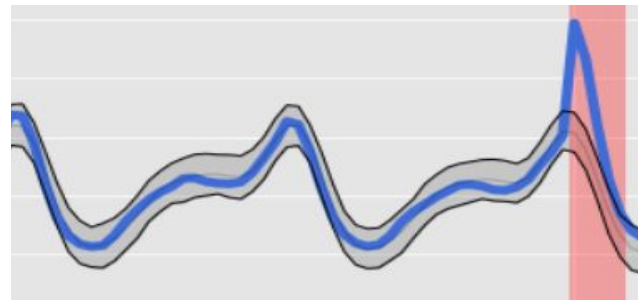
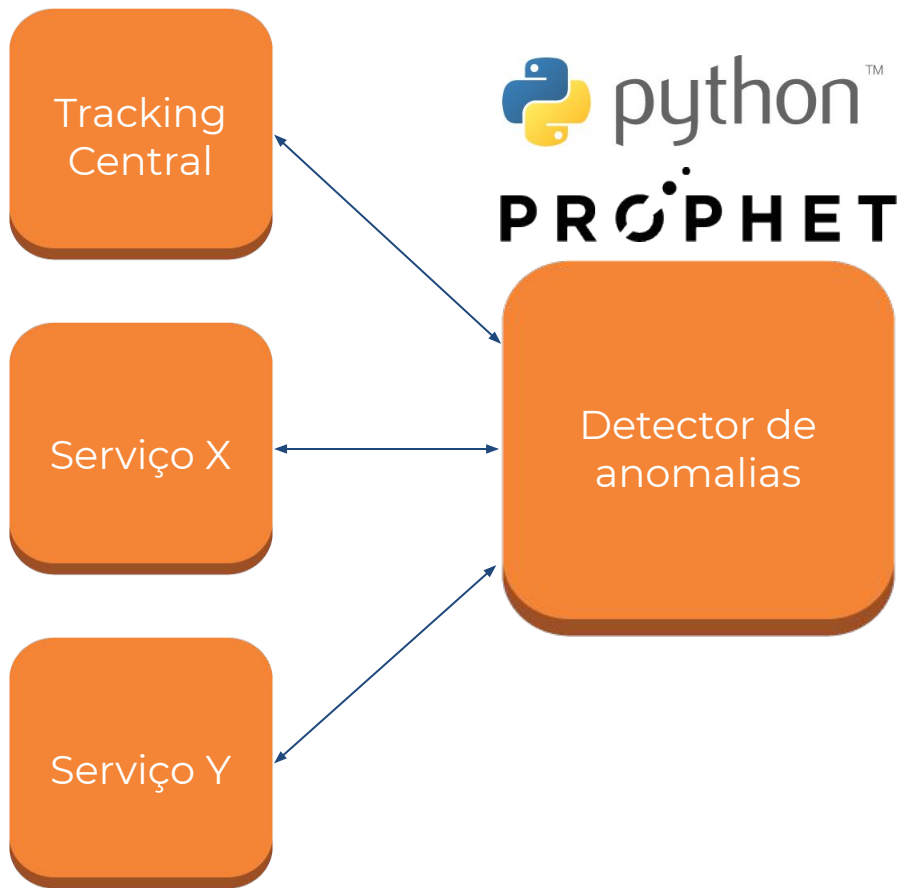
(WIP)

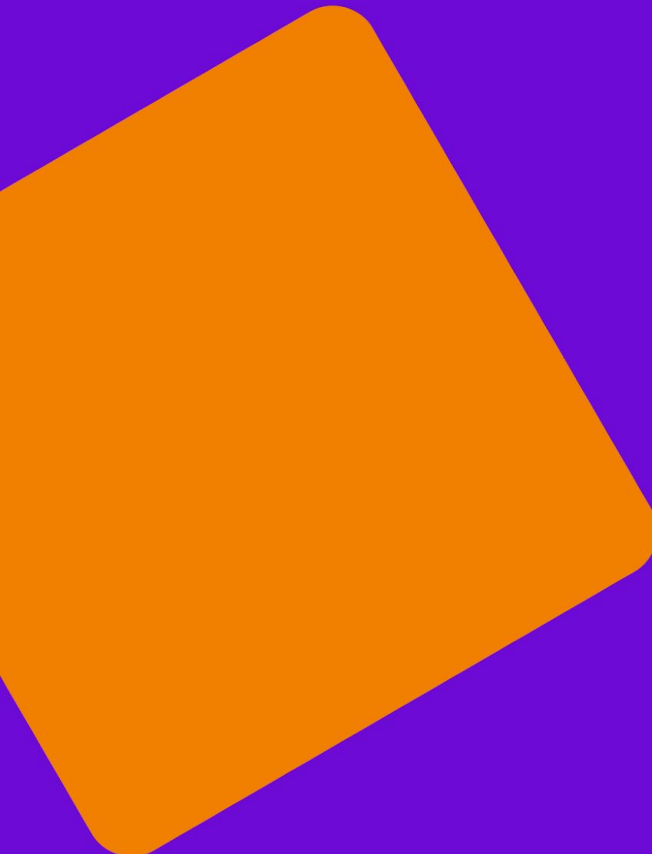
Anomalias em dados temporais



Detecção de anomalias

- ▲ Serviço de detecção de anomalias em séries temporais
- Uso genérico em séries com componente de sazonalidade:
 - eventos de tracking
 - uso de CPU/RAM de instâncias na nuvem
 - acompanhamento de métricas da empresa etc.
- ▲ Biblioteca utilizada: Facebook Prophet para Python





Solução #3
ÍNDICE DE
QUALIDADE DE
DADOS
(Próximos passos)

IQD

- Conhecemos os times responsáveis pela qualidade de cada Tracking
- ▲ Quando uma anomalia for detectada, abrimos um chamado para o time responsável corrigir o problema
- Conseguimos calcular o período de tempo que o dado foi afetado, chegando a um uptime de qualidade de dados

RESUMINDO...





Lurker coleta mais de 1M de eventos de dados de comportamento por minuto



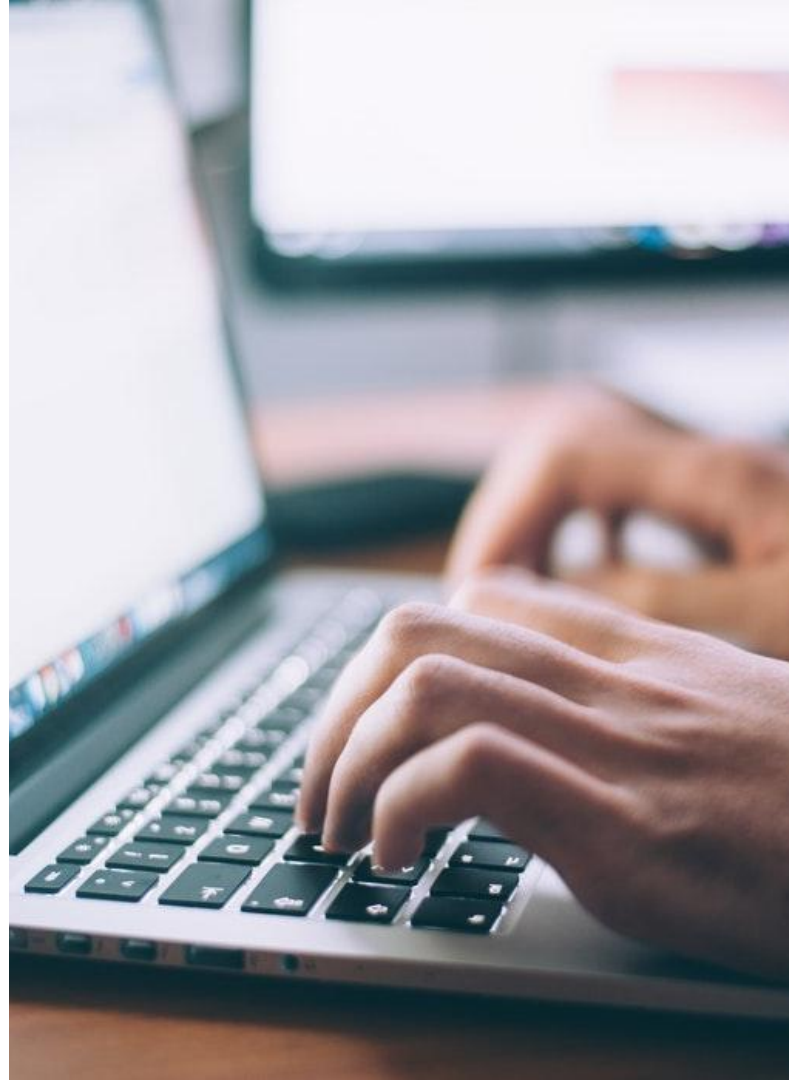
Tracking Central nos oferece uma documentação viva e sempre completa, sem perder agilidade



Detecção de anomalias nos garante a qualidade dos dados de tracking para tomadas de decisão



Índice de Qualidade de Dados fortalecerá a cultura de qualidade de dados nos diferentes squads, mudando o mindset de tech OLX





OBRIgADO

beatriz.vaz@olxbr.com | LinkedIn: mariabeatrizvaz

tiago.montalvao@olxbr.com | LinkedIn: tiagomontalvao