Kafka <3 Dataflow

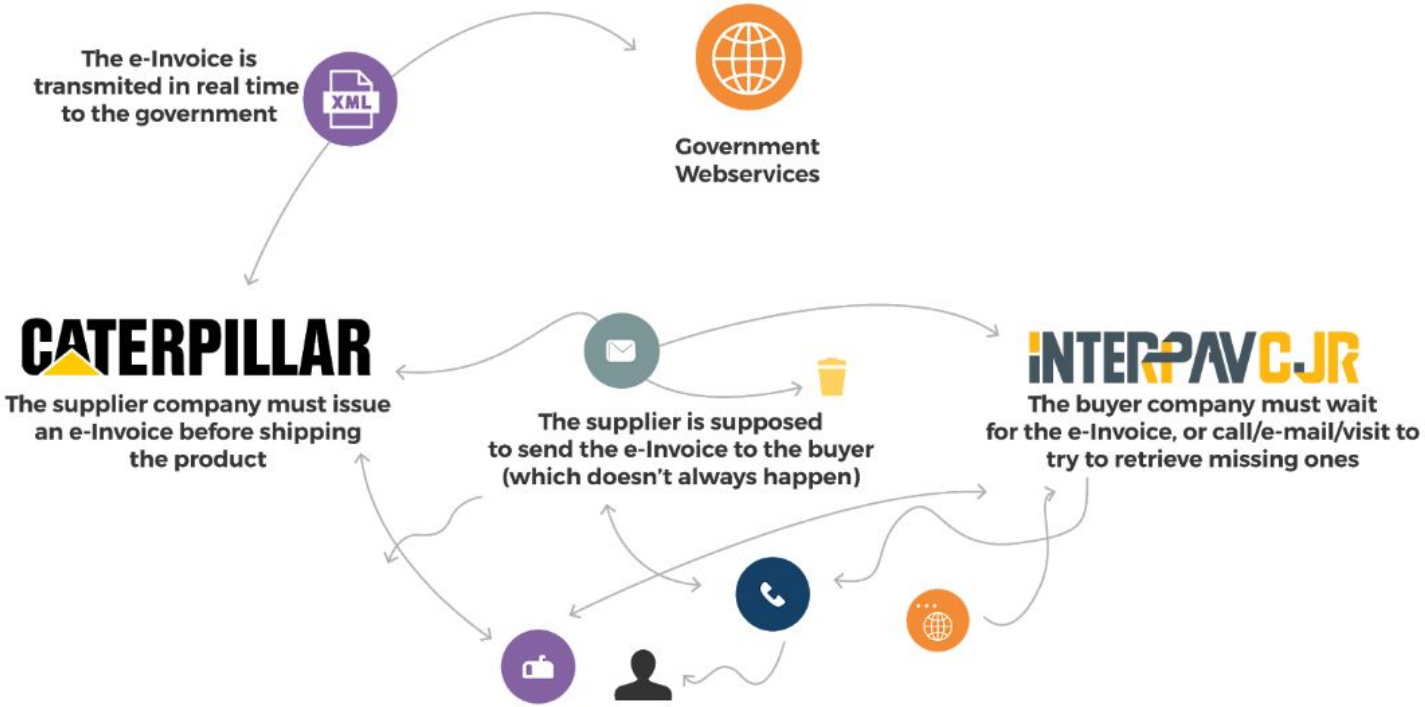
arquivei.com.br

Roadmap

- Dataflow
- Kafka
- Dataflow
- Kafka
- Dataflow e Kafka

Eu

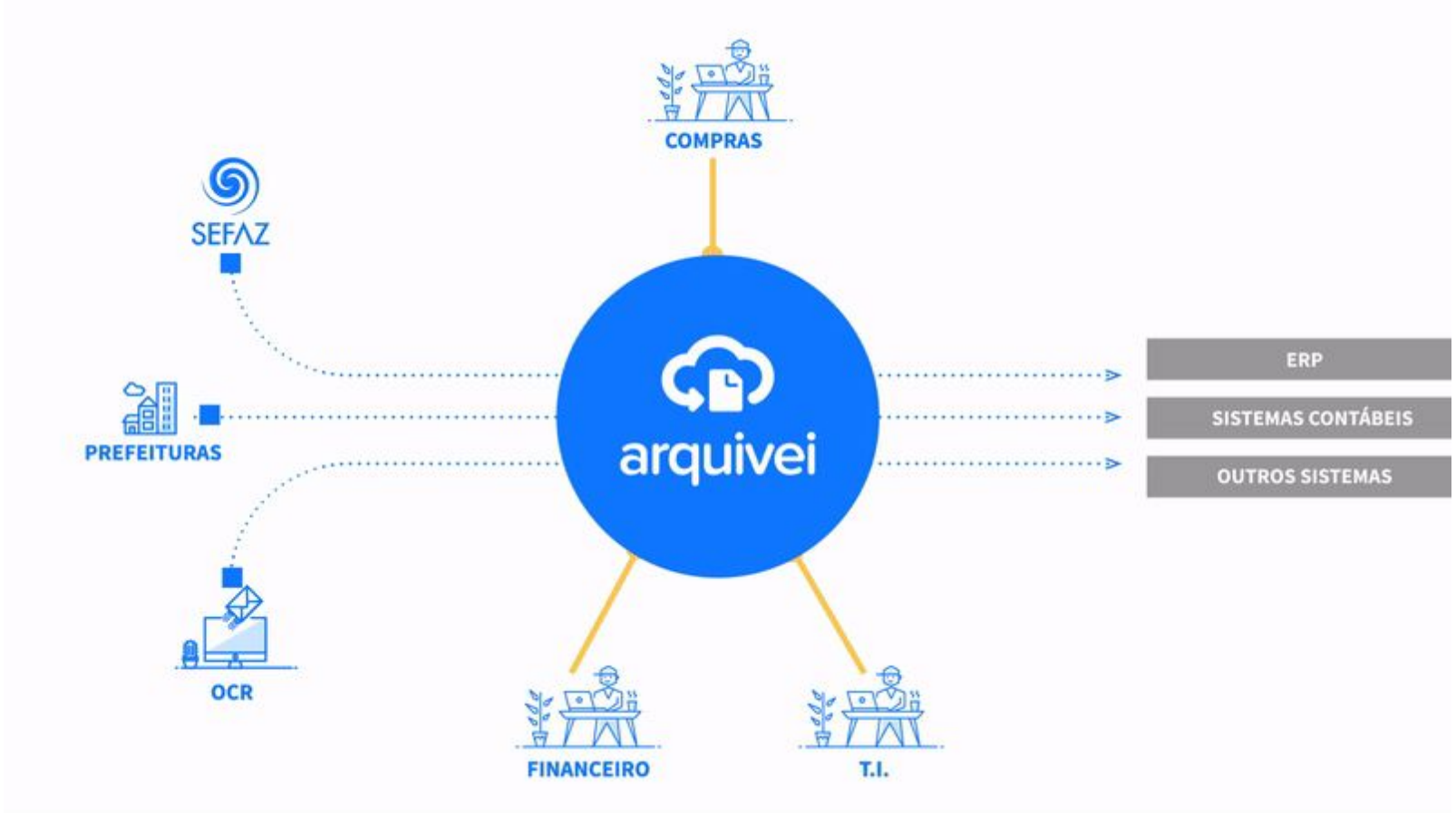
Arquivei

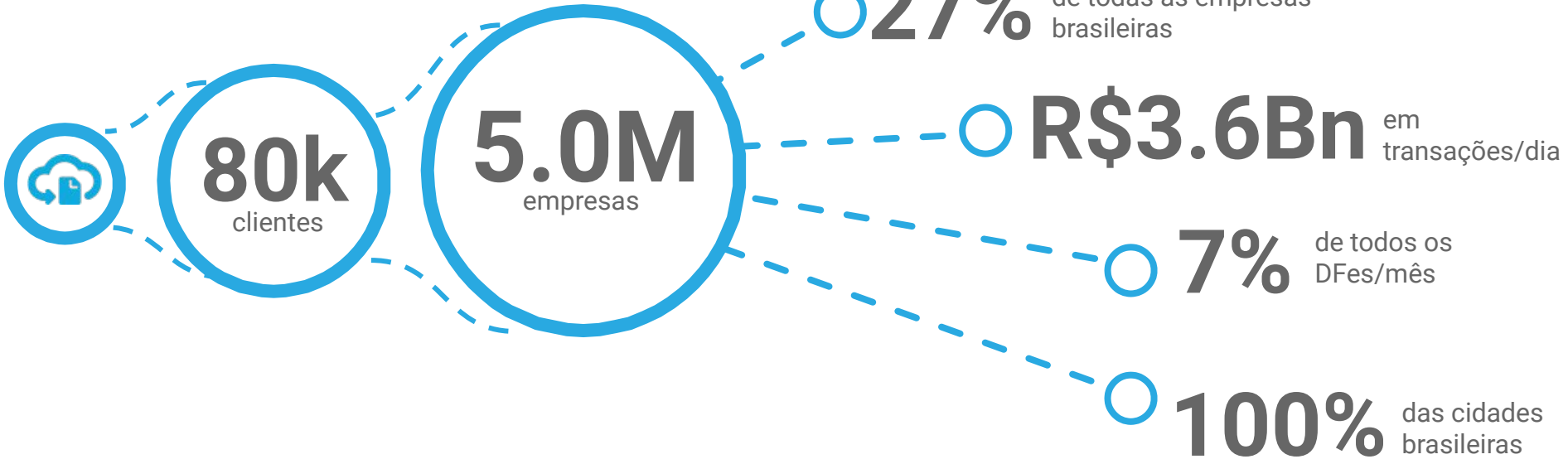

The e-Invoice is transmited in real time to the government
XML
Government Webservices
CATERPILLAR
The supplier company must issue an e-Invoice before shipping the product
The supplier is supposed to send the e-Invoice to the buyer (which doesn't always happen)
INTERPAVC.JR
The buyer company must wait for the e-Invoice, or call/e-mail/visit to try to retrieve missing ones

Arquivei


The e-Invoice is transmited in real time to the government
XML
Government Webservices
The buyer authorizes Arquivei to access e-invoices issued by all of their suppliers
CATERPILLAR
The supplier company simply issues the e-Invoice to the goverment
arquivei
XML
Arquivei automatically fetches, stores in the cloud and analyses the e-invoices in real time
INTERPAVCJR
The buyer company must simply have an Arquivei account

Democratizando a Informação





arquivei .com.br

Temos vagas: arquivei.com.br/vagas

Problema

Problema inicial

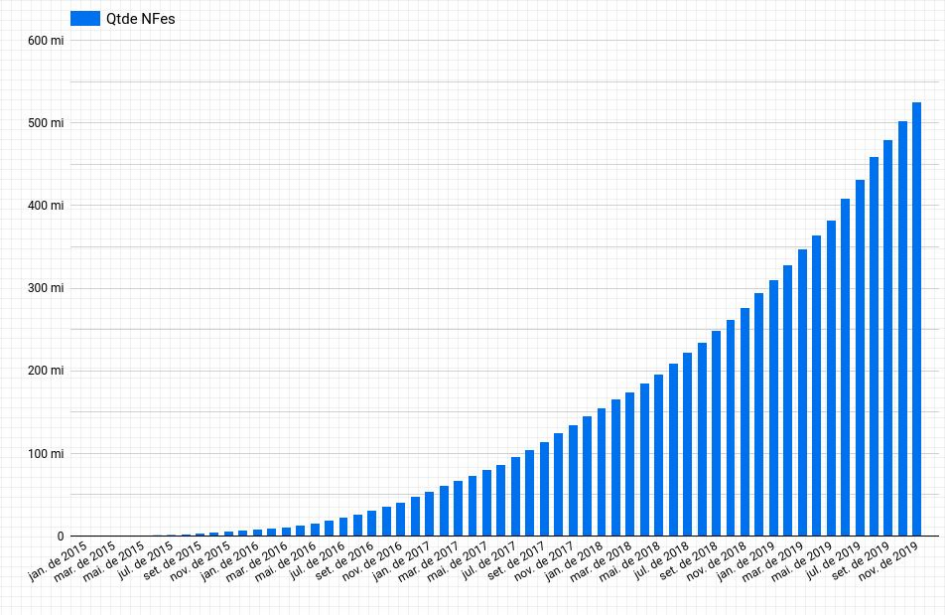
- Extrair 50M de XMLs de um PSQL
  - Contagens
  - Histórico

Problema inicial

- Extrair 50M de XMLs de um PSQL
  - Contagens
  - Histórico
  - Em tempo real...

Problema inicial

- Extrair 50M de XMLs de um PSQL
  - Contagens
  - Histórico
  - Em tempo real...



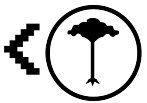
Problema inicial

- Extrair 50M de XMLs de um PSQL
  - Contagens
  - Histórico
  - Em tempo real...

Requisitos

- Sistema de coleta de dados
  - Escalável
  - Pouco intrusivo
  - Consistente

The road to Dataflow



AWS

- Python (Scriptão)
- AWS

AWS

- Python (Scriptão)
- AWS

AWS

- Python (Scriptão)
- AWS
- Airflow
- Amazon EMR

GCP

- Google Dataflow
  - Gerenciado
  - Python *-* (e Java)
  - Modelo de programação fácil
  - Streaming!

The road to Kafka

Primeiros problemas

Primeiro modelo

- SQS
- Pub/Sub
  - Barato
  - Gerenciado
  - Simples

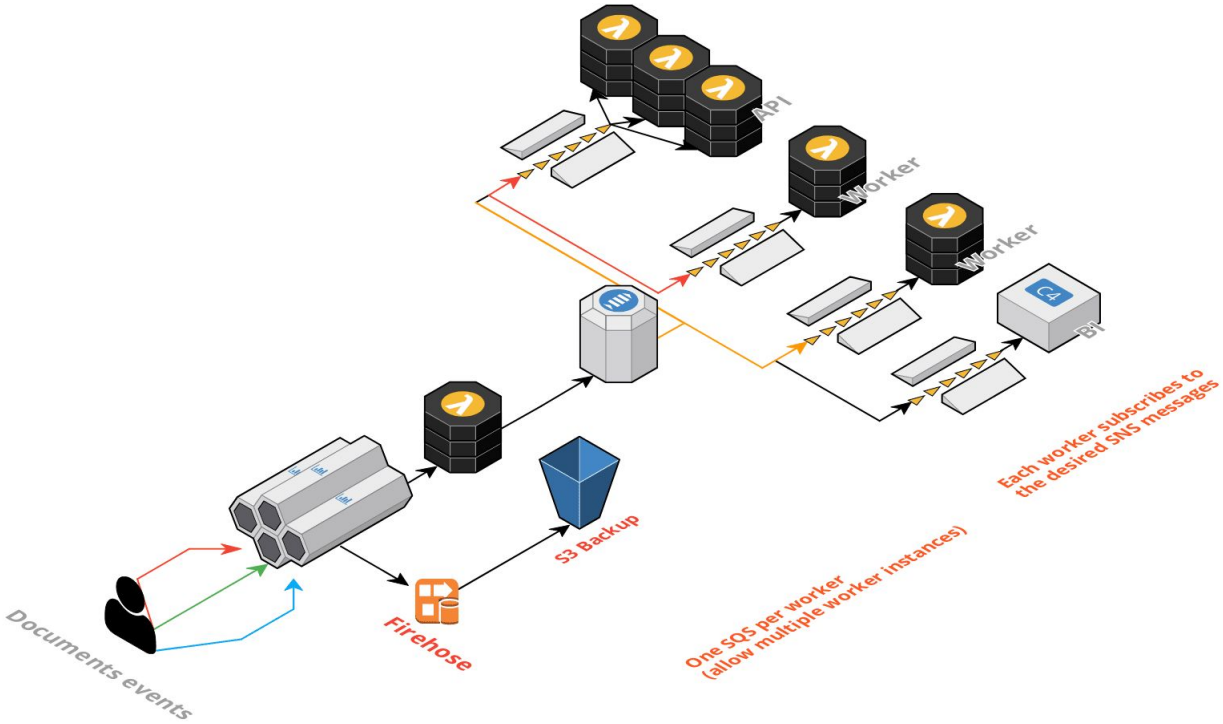
Primeiro modelo

- SQS
- Pub/Sub
  - Barato
  - Gerenciado
  - Simples
  - Problemas:
    - Latência GCP/AWS
    - Ordem
    - Falta de experiência

Segundo modelo

- Requisitos do pipeline de dados
  - Pouca infra
  - Deploys fáceis para produtores e consumidores
    - Desacoplado
    - Independente

Segundo modelo



Terceiro modelo

- Apache Kafka
  - Baixíssima latência
  - Escalável
  - “Barato”

Terceiro modelo

- Apache Kafka
  - Baixíssima latência
  - Escalável
  - “Barato”
  - Problemas
    - Overhead de Infra
    - Curva de aprendizado
      - Escalabilidade
      - Modelos de produção
      - Modelos de consumo
    - Libs limitadas para PHP

Dataflow

Dataflow model

- 2003 - GFS
- 2004 - MapReduce
- 2006 - Hadoop
- 2009 - Spark
- 2010 - Apache Flume
- 2013 - Google MillWheel
- **2015 - Dataflow model**

Dataflow model

- 2003 - GFS
- 2004 - MapReduce
- 2006 - Hadoop
- 2009 - Spark
- 2010 - Apache Flume
- 2013 - Google MillWheel
- **2015 - Dataflow model**
- 2019 - Palmeiras continua sem mundial

Dataflow model

- Akidau
- Bounded: batches
- Unbounded: streaming

Dataflow model

- Akidau
- Flume (batch) + MillWheel (streaming)
- Batches existentes
  - Alta latência
- Streamings existentes
  - Tolerância a falhas, escalabilidade, latência
  - Complexidade (janelamento)

Dataflow model

- Conceitos
  - Event time vs Processing time
  - Windows
  - Triggers
  - Accumulation modes

Apache Beam

- Implementação do modelo
- Linguagens
  - Java
  - Python
  - Go (experimental)
  - Portability Framework
- I/Os prontos
  - GCP
  - AWS
  - Elasticsearch, Kafka, etc

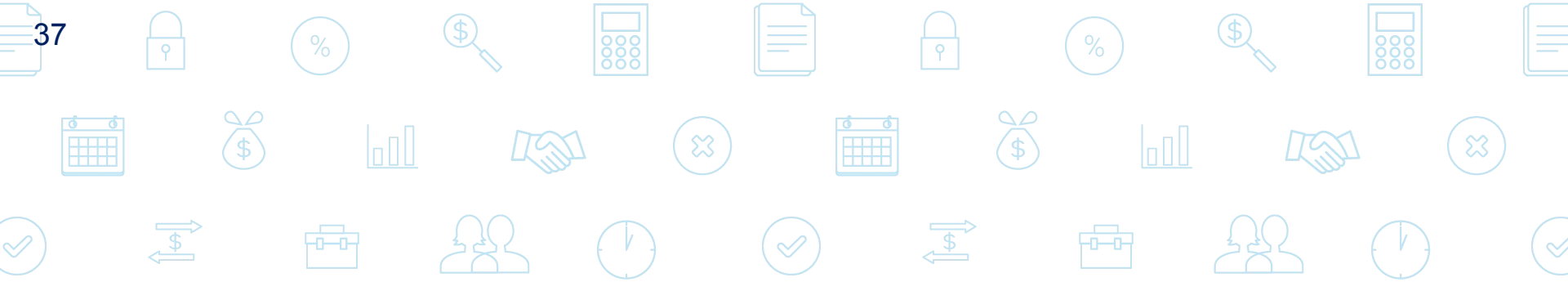
Google Dataflow

- Produto GCP
- Cluster gerenciado
- Autoscaling

Google Dataflow

▸ Produto GCP

Kafka



Apache Kafka

- Kreps, 2011
- Offline: batch
- Online: streaming

Apache Kafka

- Kreps, 2011
- Streamings existentes
  - Projetados para consumo offline
- Sistema proposto pelo LinkedIn

Apache Kafka

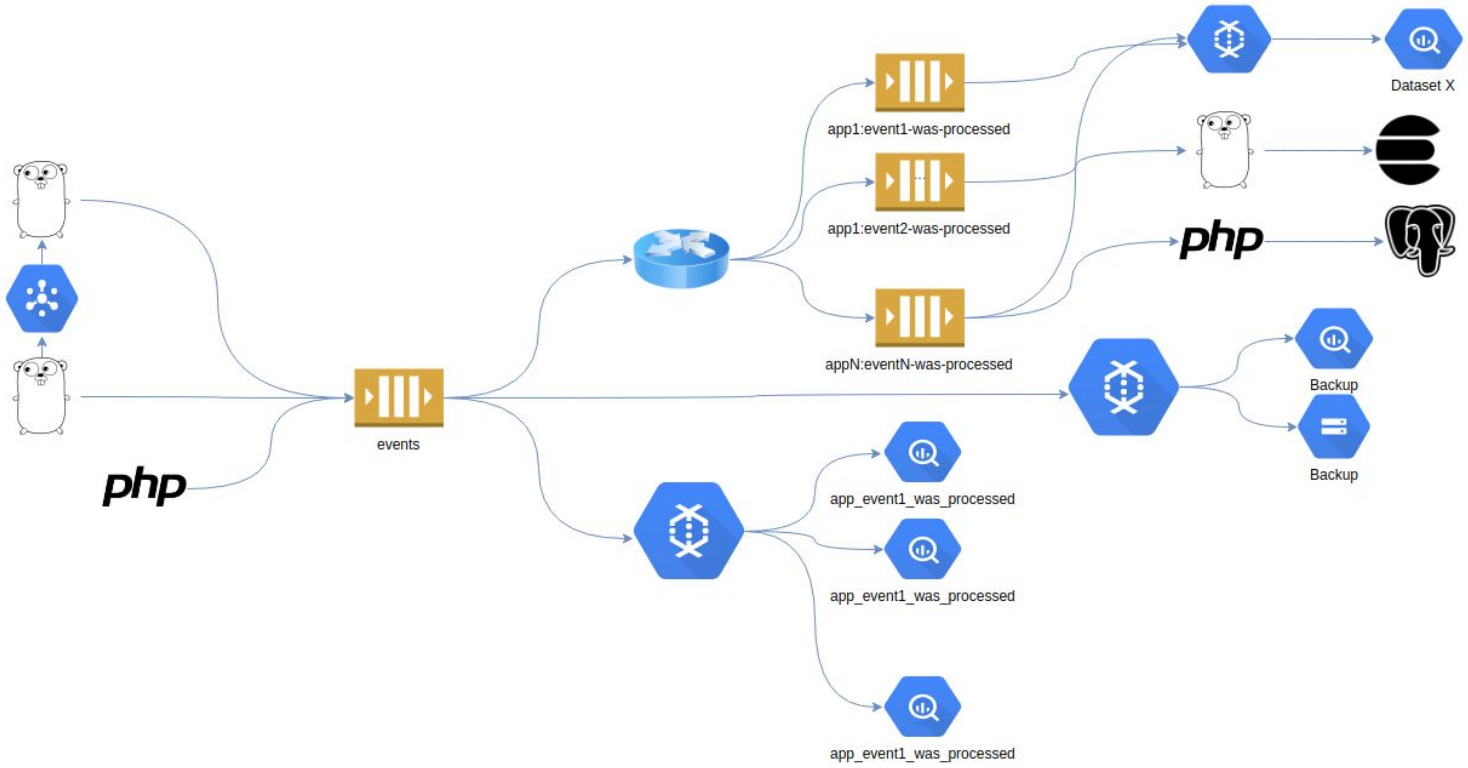
- Kreps, 2011
- Streamings existentes
  - Projetados para consumo offline
- Kafka
  - Distribuído e escalável
  - API consumo realtime
  - Projetado tanto para consumo
    - Offline
    - Online

Apache Kafka

- Conceitos
  - Event time vs Processing time
  - +Log Append Time
  - Streams
    - Ordenados
    - Imutáveis
    - Repetíveis

Kafka + Dataflow

Kafka+Dataflow



Demo

FAQ

- Bibliografia
  - Akidau, 2015
  - Kreps, 2011
  - I <3 Logs
  - Designing Data-Intensive Applications
  - Kafka: The Definitive Guide
- Códigos: github.com/arquivei
- Contatos:
  - @leonardobarbie (twitter)
  - leonardo.miguel@arquivei.com.br