

Proyectos de Big Data

Diplomatura en Big Data – ITBA - 2019

ITBA Flowics

¡Hola!
Andrés Moratti
@amoratti



ITBA

Flowics

 Zauber

Agenda

¿Qué es Big Data?

Casos de uso

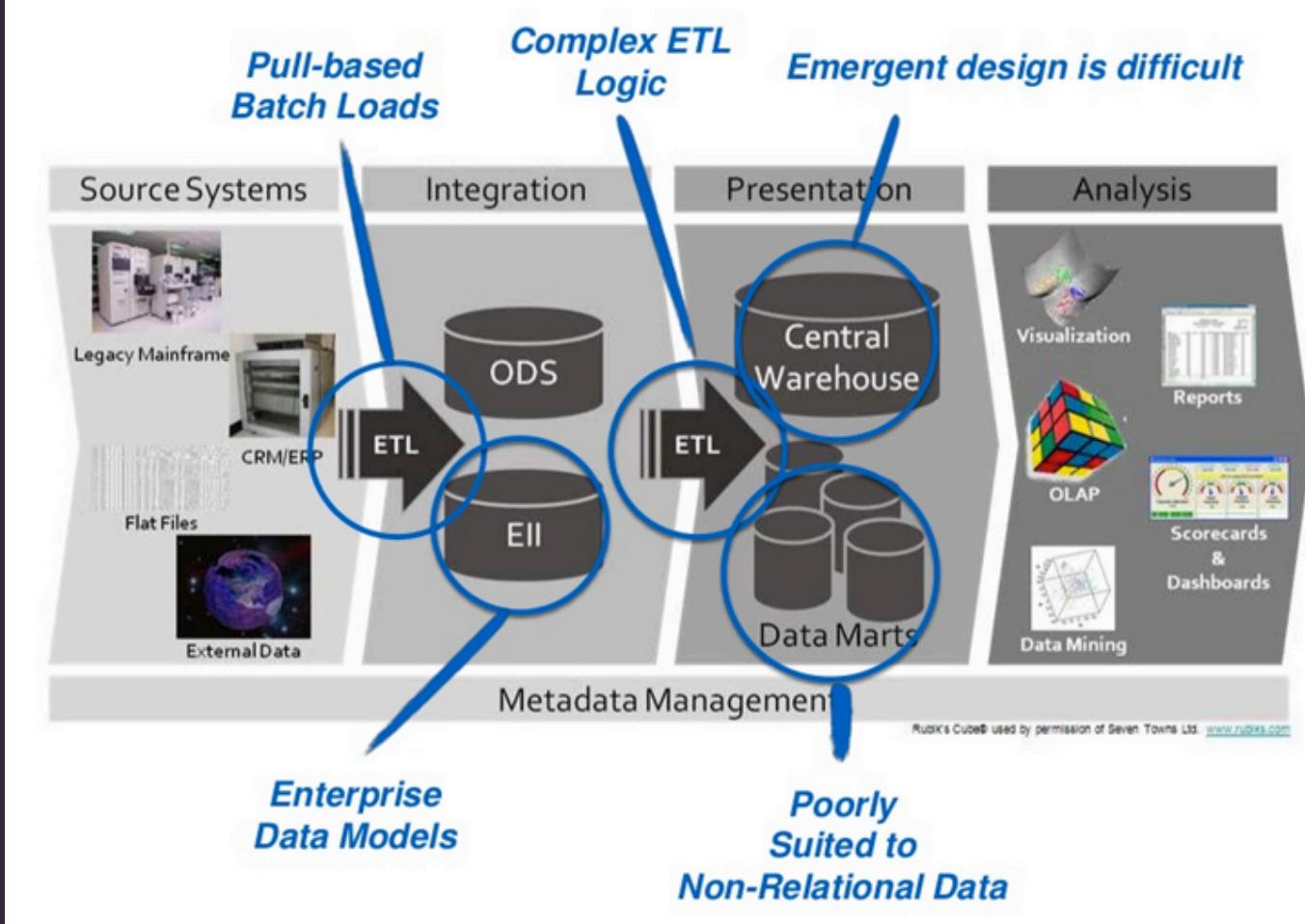
Tecnologías actuales

Implementando un proyecto

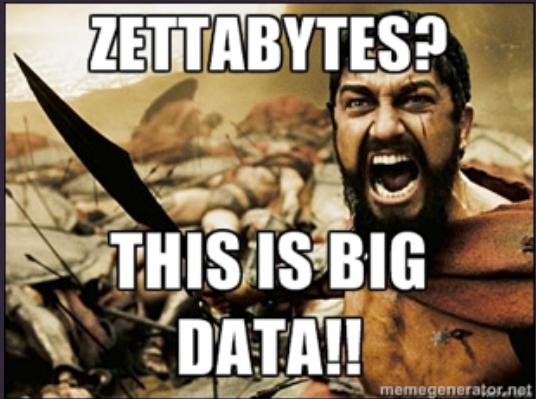
Big Data @ Flowics



¿DE DÓNDE VENIMOS?



¿Qué es Big Data?



DevOps Borat
@DEVOPS_BORAT



Big Data is any thing which is crash Excel.

RETWEETS

2,154

LIKES

592



**Big data is like teen sex.
Everybody is talking about it,
everyone thinks everyone else is doing it,
so everyone claims they are doing it.**

Dan Ariely,
Center for Advanced Hindsight
Duke University

¿Qué es Big Data?

Nueva generación de **tecnologías** y
arquitecturas diseñadas para extraer **valor** de
grandes volúmenes de **datos**

Bajos costos de hardware (COTS)
+ software (OSS)
+ escalabilidad horizontal



¿Qué es Big Data?



Casos de uso

Retail

Análisis de compras

Programas de fidelidad

Targeting y Segmentación



Servicios financieros

Detección de fraude

Trading de acciones

Prevenir *churn* de clientes

Casos de uso

Digital

Análisis de *click stream*

Targeting de publicidad

Grafo social

Cyberseguridad



Salud

Análisis clínicos

Patrones de enfermedades

Desarrollo de drogas / ADN

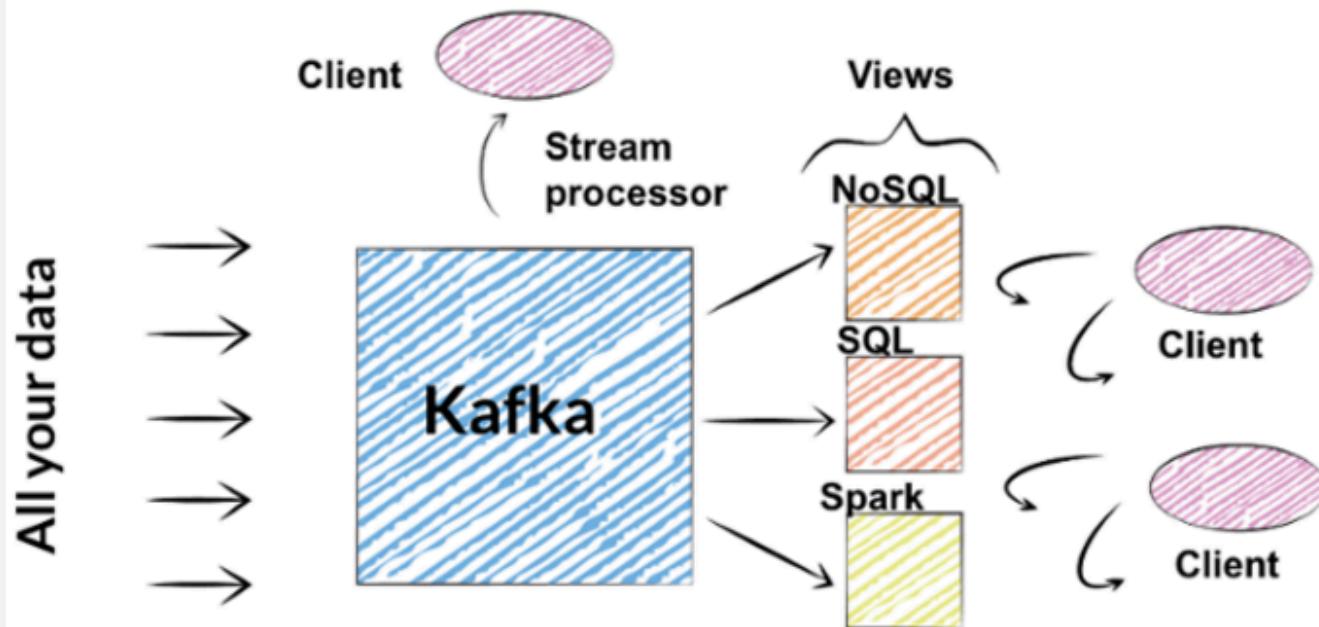
Tecnologías actuales

Map / Reduce & Streaming
Machine Learning
NoSQL
Cloud Computing

Tecnologías actuales

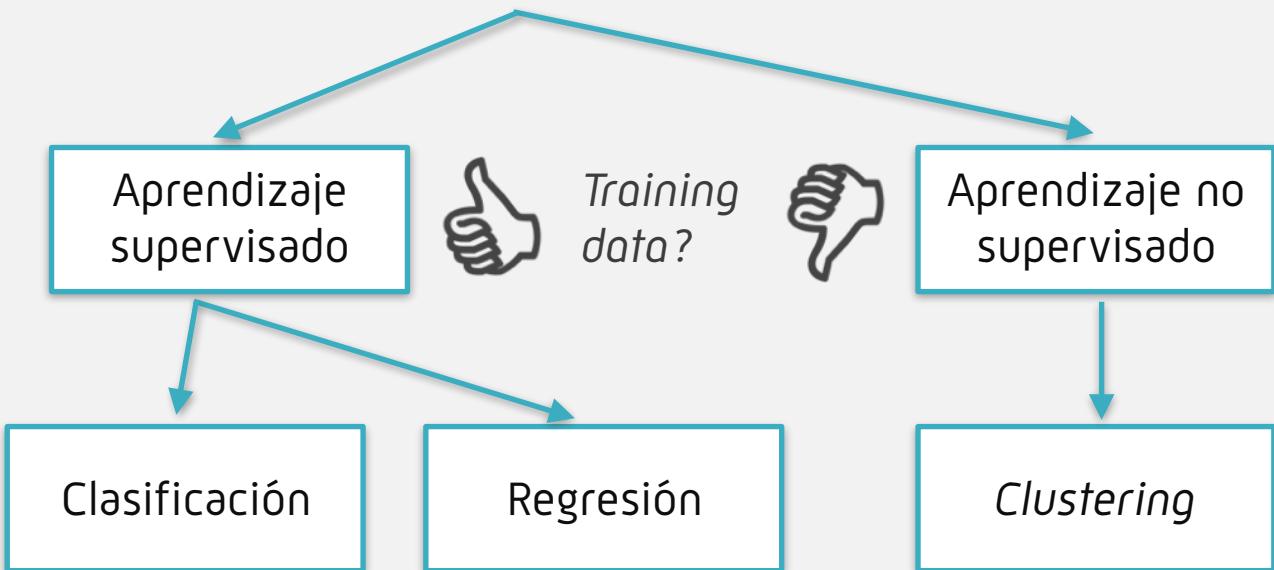
Map / Reduce & Streaming

Kappa architecture



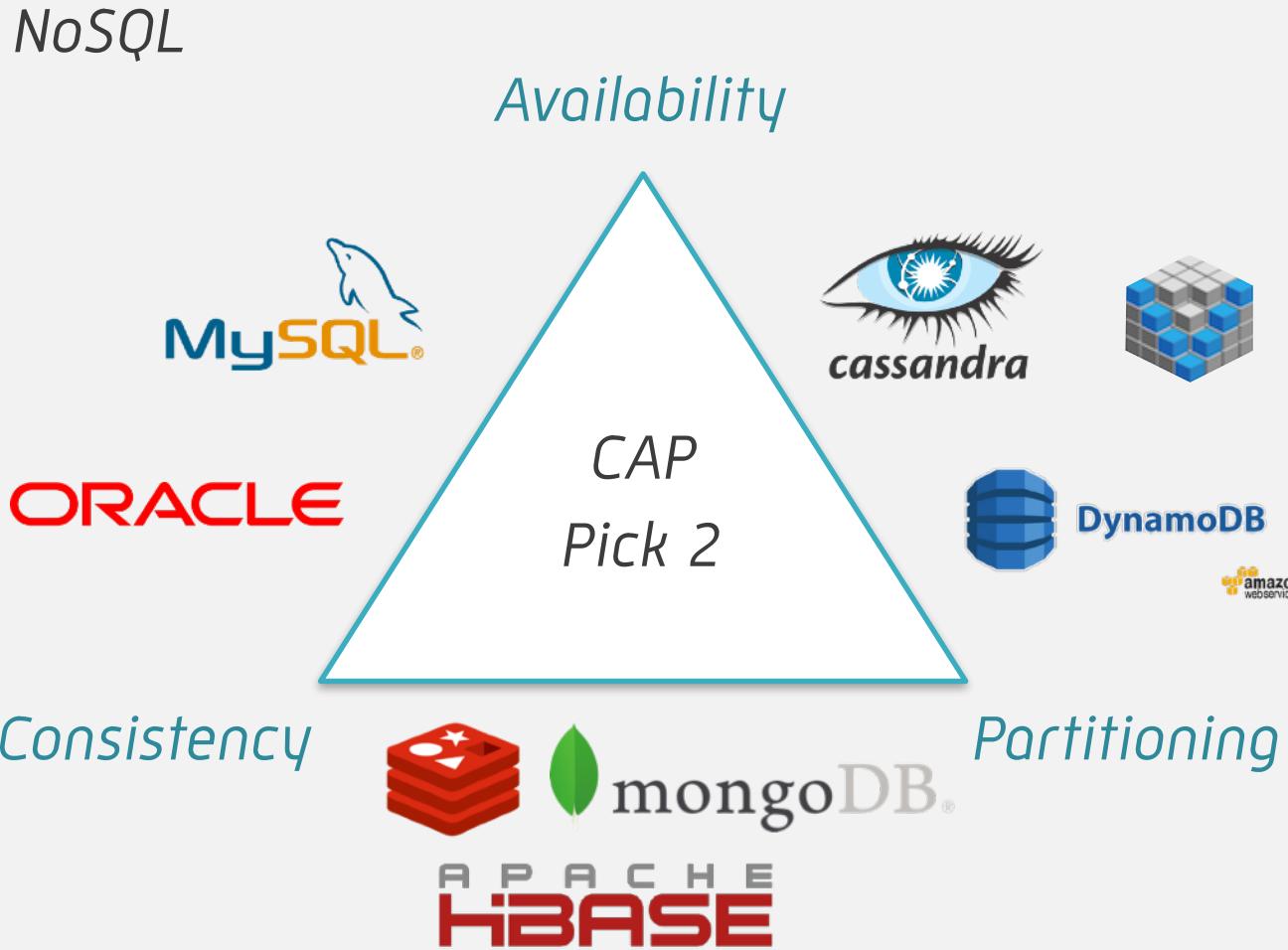
Tecnologías actuales

Machine Learning



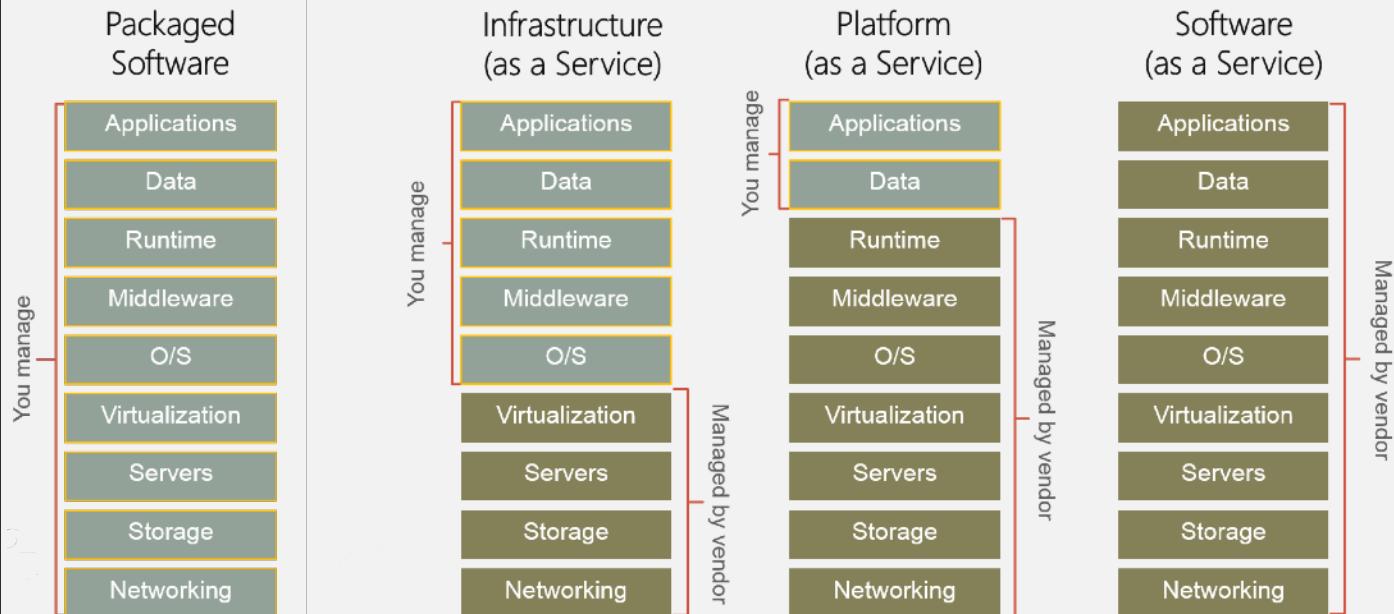
Tecnologías actuales

Modelo datos
Column
Document
Key-Value
Graph



Tecnologías actuales

Cloud Computing



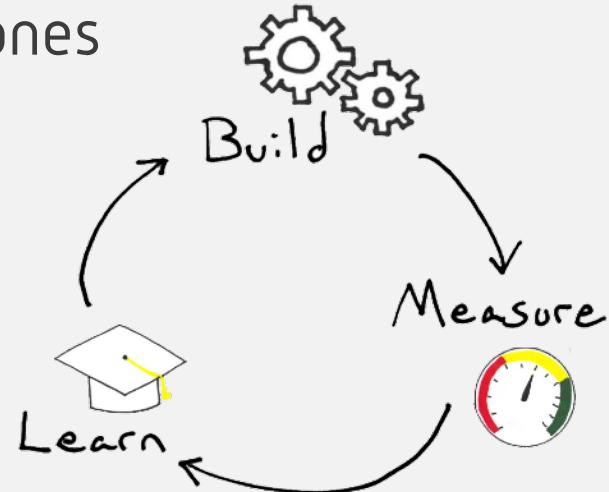
Tecnologías actuales

Cloud Computing

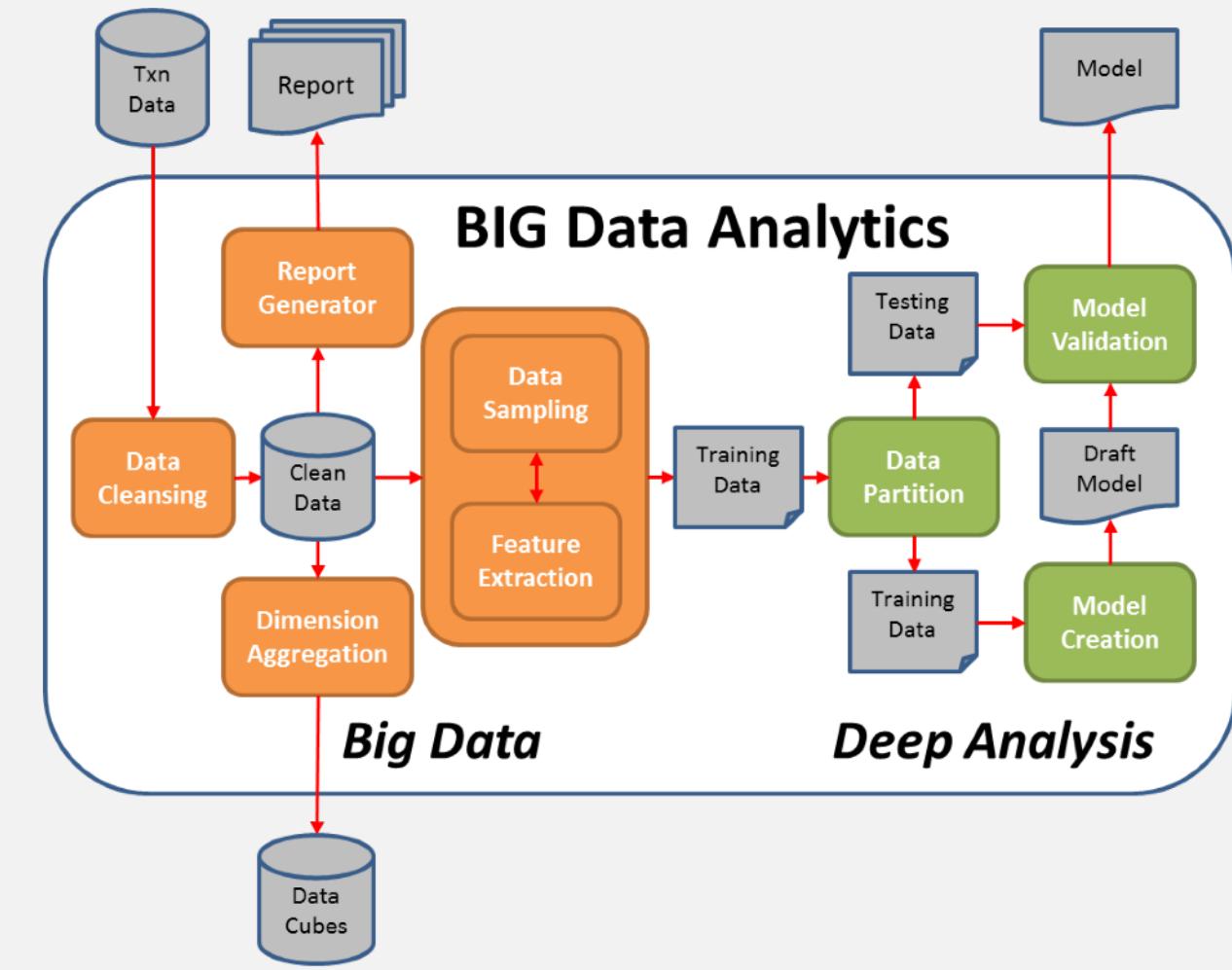
IaaS	PaaS	SaaS
 amazon web services	 heroku	 Google Apps for Work
 Microsoft Azure		 salesforce
 CLOUD FOUNDRY		 GitHub

Implementando un proyecto

1. Adquirir y curar datos
2. Visualizar y entender los datos
3. Generar reportes (OLAP)
4. Respuestas automáticas
5. Análisis predictivo
6. Optimización de decisiones



Implementando un proyecto



Implementando
un proyecto

¿Qué equipo necesito?

Experto de dominio

Data Scientist

UX

Ingenieros

Operaciones



Agenda

~~¿Qué es Big Data?~~
~~Casos de uso~~
~~Tecnologías actuales~~
~~Implementando un proyecto~~
Big Data @ Flowics



Flowics

THE AUDIENCE PARTICIPATION PLATFORM



Flowics es una **plataforma para
participación de audiencias**. Permite a
medios y marcas atraer, retener y monetizar
audiencias en todos sus **entornos digitales**.



Povo Tuíta nos últimos 30 minutos

- █ UNIÃO GAY
- █ EDUCAÇÃO
- █ COTAS RACIAIS
- █ HABITAÇÃO
- █ PRIVATIZAÇÃO
- █ ECONOMIA
- █ SAÚDE
- █ SEGURANÇA
- █ TRANSPORTE
- █ ABORTO
- █ BOLSA FAMÍLIA
- █ CORRUPÇÃO
- █ DROGAS
- █ EMPREGO
- █ MEIO AMBIENTE
- █ MENSALÃO
- █ PETROBRAS
- █ REFORMA POLÍTICA



VIVO 11:5 mil COP 0) (0 DIM 17:02

Responde la encuesta: ¿Llegará un equipo chileno a la final en Santiago?

¡Sí, llegaré!

Solo hasta cuartos

No tienen posibilidades

Los resultados de la encuesta están ocultos.

Lu Marc, Andrés Barraza y 2 amigos más realizaron esta encuesta.

46% (Sí, llegaré) 39% Solo hasta cuartos 15% No tienen posibilidades

Facebook Watch 4,165 INTERACCIONES



Flowics



AHORA EN VIVO
INTRATABLES

¿CUÁNDO CREE
UD. QUE HUBO
MÁS
INSEGURIDAD?

PERÍODO CRISTINA

PERÍODO MACRI

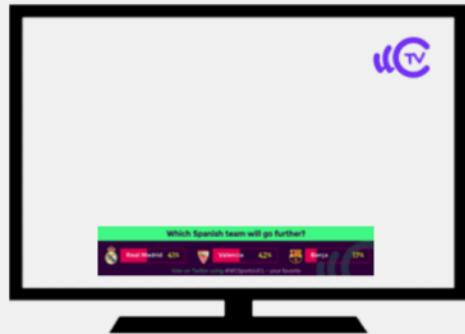


Flowics

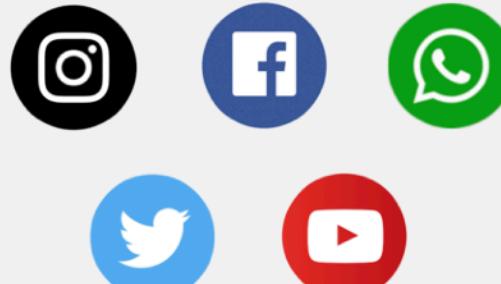


Cómo funciona

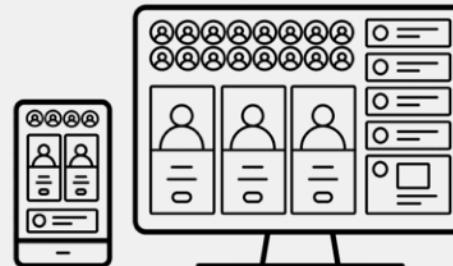
INTERACTIVIDAD
ON AIR + DIGITAL +
REDES SOCIALES



EN REDES SOCIALES



EN DIGITAL (SITIOS O APPS)



Desafíos

Flowics

- Respuesta en **tiempo real**
 - Alto volumen de datos
 - Filtrado
 - Enriquecimiento
 - Alto volumen de *requests*
- Sin *downtime*
- Clasificación con **poca información**



Desafíos

Volumen

Flowics

Alemania vs Brasil ~600.000 tweets / min



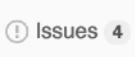
Fiesta inaugural ~800.000 requests / min



FIFA WORLD CUP
Brasil

 [zauberlabs / gnip4j](#)

 Watch 32  Star 42  Fork 41

 Code  Issues 4

Gnip Client Library for Java

Tecnología

 Home APIs ▾ Gnip 2.0 APIs ▾ Enrichments ▾ Sources ▾ More ▾

Other Options

The following utilities are not built or maintained by Gnip, but may provide a jumping-off point for your a

Gnip4j

Gnip4j is an Java Open Source (ASFL 2.0) library to access and process Activities (tweets) from Gnip.



Backend

Captura

Flowics

- Cluster **distribuido**
- **Elástico** dependiendo de la demanda
- Proveedores heterogéneos
 - *push vs pull*
 - *stream vs batch*
- Sin procesamiento
- Conversión a *stream*



Backend

Mensajería

Flowics

- RabbitMQ
 - garantiza *delivery*
 - *broker* mantiene estado
 - baja performance con persistencia
 - no tolera partición de red
- Kafka
 - *at least one delivery*
 - consumidor mantiene estado
 - **alta performance con persistencia**
 - elástico + replicación



Backend Procesamiento

Flowics

- Arquitectura Lambda
 - *Batch*
 - *Real time*
 - Apache Storm
 - *stream processing*
 - elástico y tolerante a fallos
 - procesamiento con estado
 - filtrado + enriquecimiento
 - latencia << 1s



Backend Procesamiento

Flowics

- Técnicas para enriquecer datos
 - *Natural Language Processing*
 - *Machine Learning*
- Detección de lenguaje
- Detección de ubicación
- Detección de rango edades y género
- Análisis de sentimiento



Backend

Bases de Datos



- *NoSQL*
 - Redis
 - MongoDB
 - ElasticSearch
 - Azure Storage
 - Kafka
- *SQL*
 - PostgreSQL HA

Modelo de datos
+ consultas



Backend

APIs

Flowics

- Exponen datos procesados al *frontend*
- RESTful + JSON
- Elástico
- Acceso a bases *NoSQL*



Frontend

Configuración

- Aplicaciones web
- Microservicios
- Permiten configurar
 - Captura
 - *Display*

The screenshot shows a front-end visualization interface. At the top, there's a header with a 'F' icon, the text 'Latin Grammys 2015 > Visualizations > Viz LatinGrammys', and a 'Published 4 months ago' timestamp. To the right are buttons for 'Edit', 'Preview', 'Publish', and a share icon. Below the header, the main content area displays the official website for the '16th ANNUAL LATIN GRAMMY AWARDS'. The website features a dark background with red and white text. It prominently displays the event name '16th ANNUAL LATIN GRAMMY AWARDS' along with a large gramophone award icon. To the right, it says 'LA NOCHE MÁS IMPORTANTE DE LA MÚSICA LATINA' and '19 DE NOV. | LAS VEGAS'. Navigation links at the bottom include 'EN VIVO', 'Premiación', 'Noticias', 'Fotos', 'Videos', 'Prensa', 'Patrocinadores', 'Miembros', and 'Más'. On the left side of the main content, there's a sidebar titled 'ELEMENTS' with a list of components: 'Container' (with a grid icon), 'Title' (with a text icon), 'Paragraph' (with a text icon), 'iFrame' (with a video camera icon), and 'http://' (with a link icon). Each item has a small preview thumbnail.



Frontend

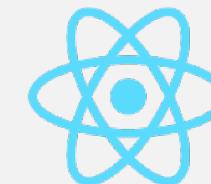
Presentación

Flowics

- Aplicaciones web
- Elástico
- Uso intensivo de *caching*
- *Responsive design*



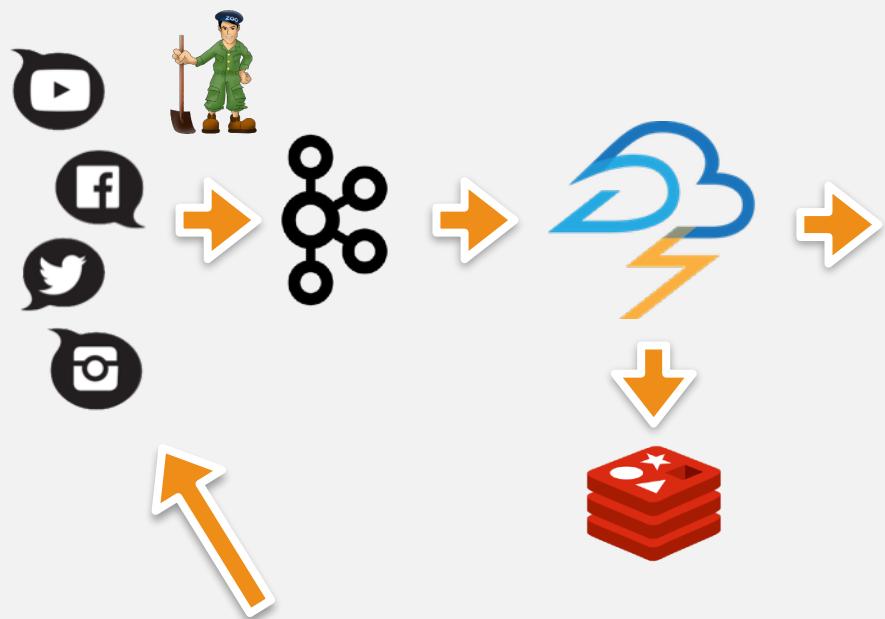
NGINX

The React logo, which is a blue, circular icon composed of three nested, curved segments.

React

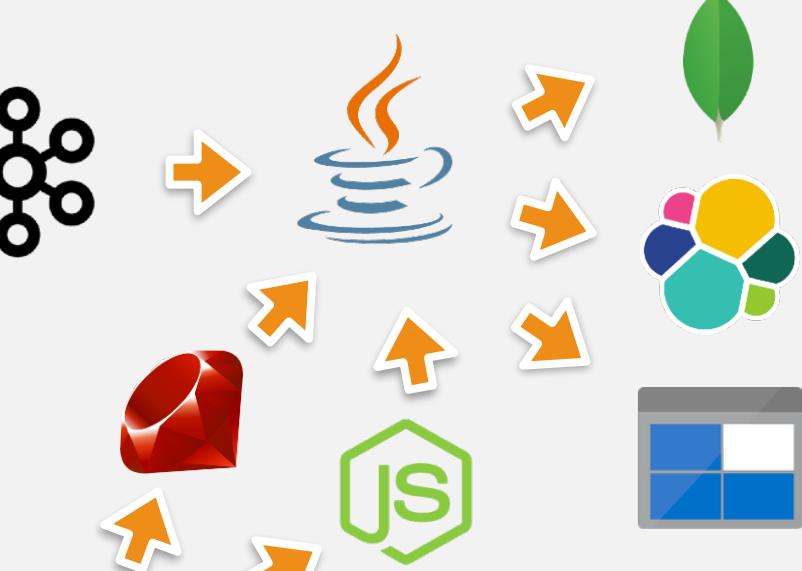
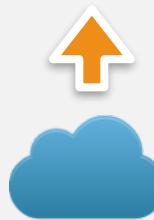
The Redux logo, which is a purple, circular icon featuring three dots connected by lines to form a triangle.

Redux



NGINX

Flowics

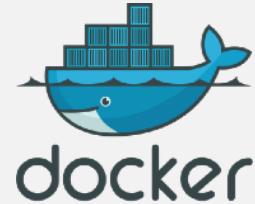


F

Operación Infraestructura

Flowics

- 100 % *cloud*
- *IaaS vs PaaS*
- Tamaño equipo



- Alta disponibilidad

Aplicaciones
Bases de datos

Operación

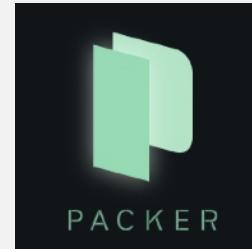
Build & Deploy

Flowics

- Integración continua



- Servidores reproducibles y descartables



Operación

Monitoreo

Flowics

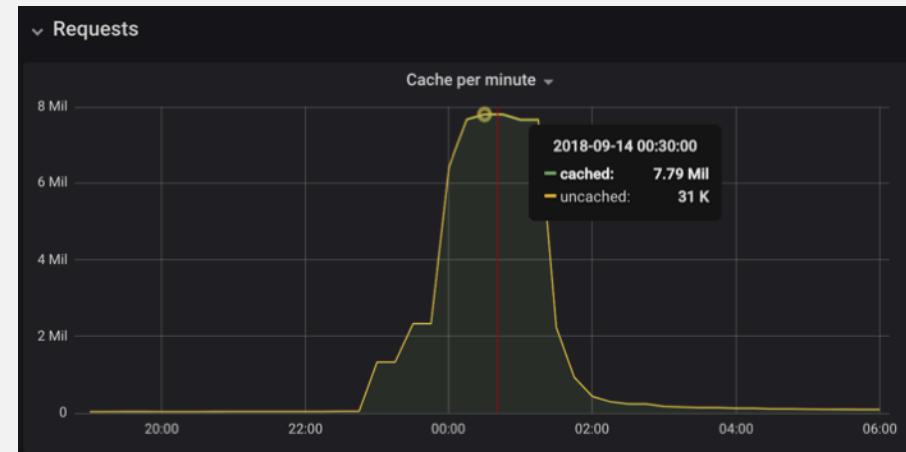
- Métricas de aplicación y de servidores
 - *Thresholds* para alarmas



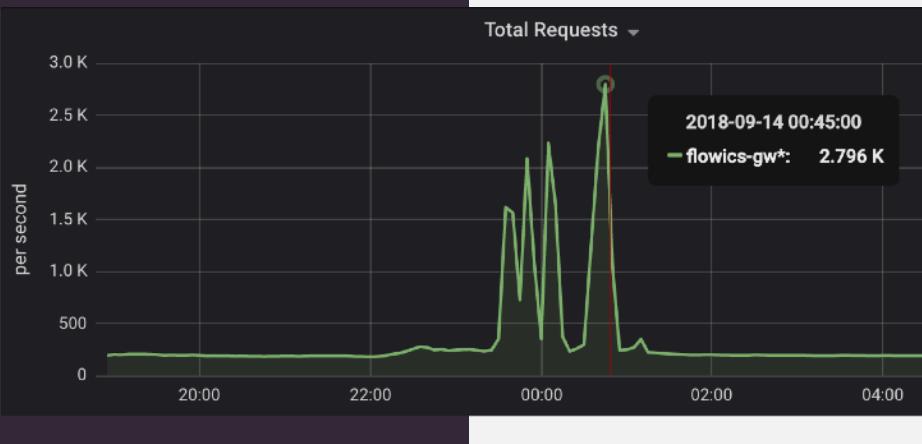
- Logs centralizados



Operación



~130.000
Reqs / seg



~3.000
Reqs / seg

NGINX

Rojo - Final



Takeaways

Flowics

- Plataforma redundante + elástica
- Procesamiento: *batch vs real time*
- Elegir bases de datos
 - modelo de datos
 - consultas
- Desde el día 0
 - servidores reproducibles
 - monitoreo + *logs* centralizados

¿Preguntas?

¡Gracias!



@amoratti

andres@flowics.com

STAY CONNECTED



/flowics



@flowics



/flowicsTV