Practica 3 y 4
GRANDES VOLÚMENES DE DATOS

Prof. José Luis Cuadrado García

Universidad de Alcalá

Departamento de Ciencias de la Computación e Inteligencia Artificial

Motivación

- Capturar tweets para análisis científico
- Estudio de la escalabilidad
- Datos Distribuidos
- Fusion de datos

→ Objetivo

Adquisición

Almacenamiento

Consulta de datos

→ ¿Cómo obtener los datos de Twitter?

REST API

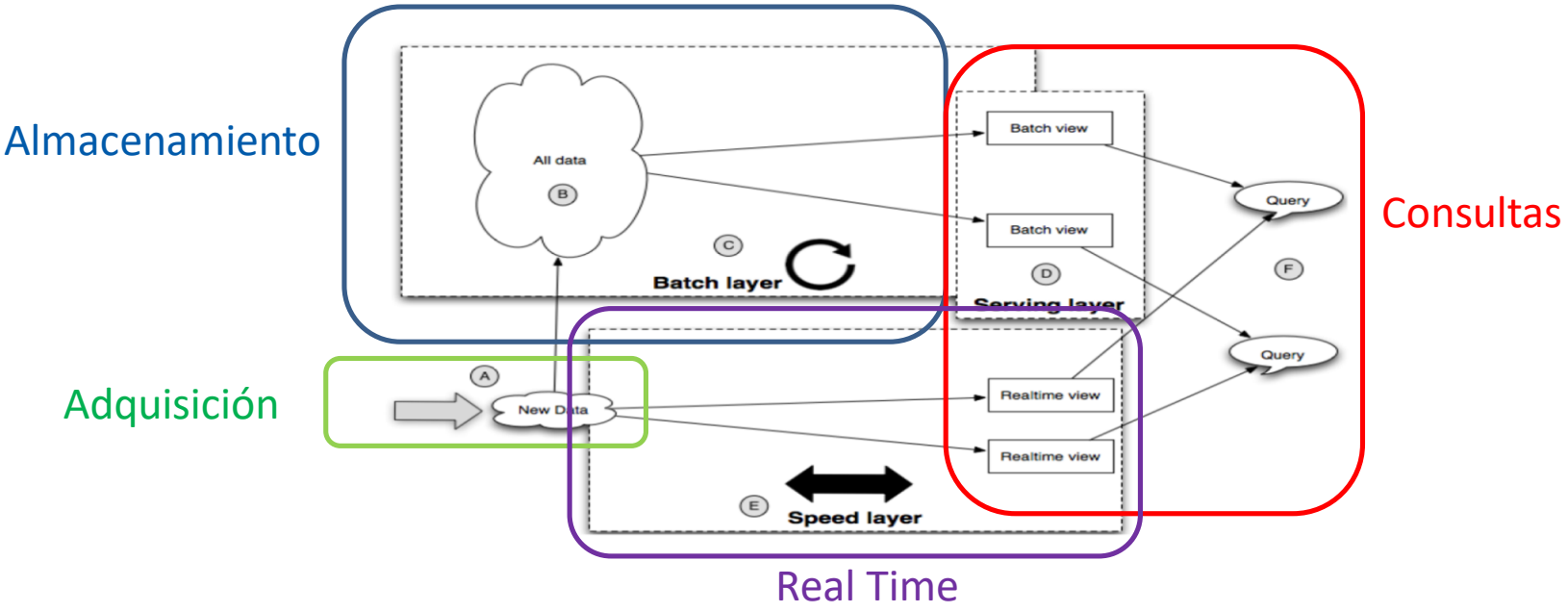
- Consultas de tweets pasados (hasta 10 días)

StreamingAPI

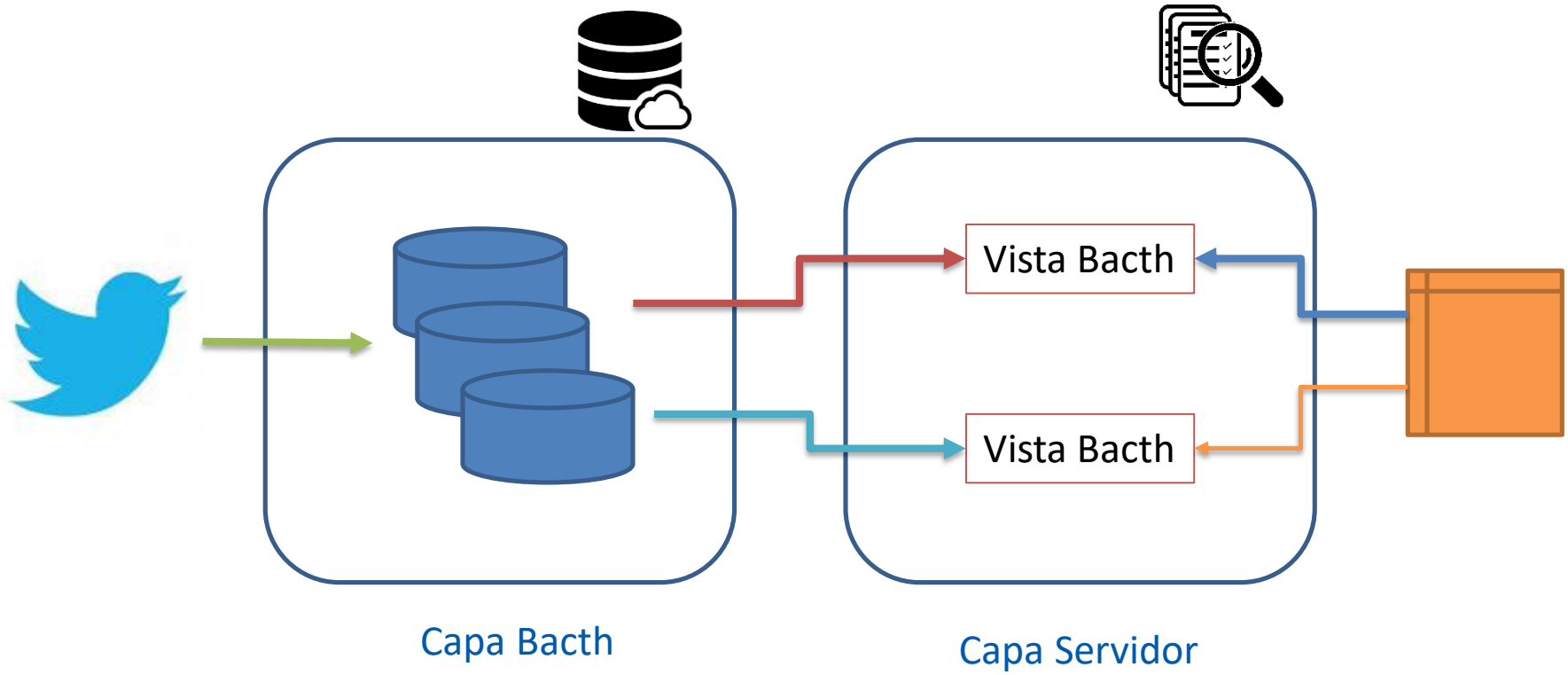
- Mayor rate limiting
- Menos consultas

→ Arquitectura Lambda

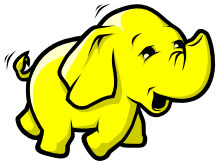
- Marz & Warren, *A new paradigm for Big Data*, 2012
- Grandes volúmenes de datos añadidos continuamente



→ Arquitectura



Hadoop

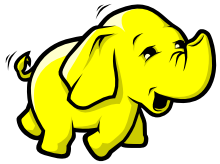


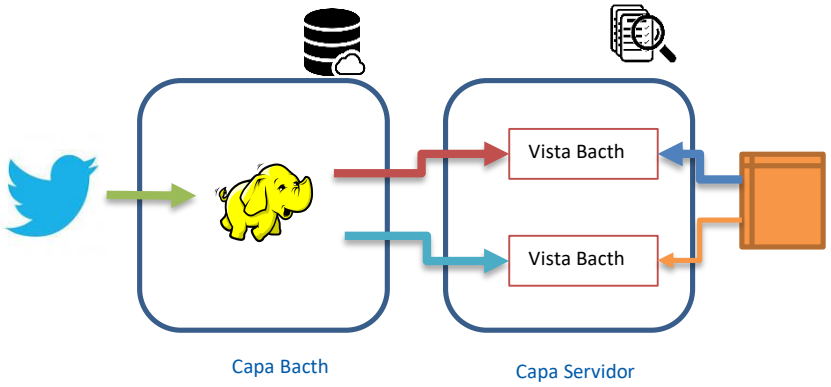
Procesamiento distribuido de grandes cantidades de datos

- Tamaño: Petabytes
- Procesado: por lotes
- Actualizaciones: escribe una vez, lee muchas veces
- Estructura: Base de datos semi-estructurada
- Control del programador: operación en alto nivel
- Escalado lineal: independiente del tamaño de datos y del cluster

→ Componentes

Hadoop





Procesamiento distribuido (**MapReduce**) a través de clusters de computadoras (**HDFS**)

HDFS

Archivos muy grandes
Acceso a datos constante

NameNode(servidor maestro)
DataNodes(trabajadores)

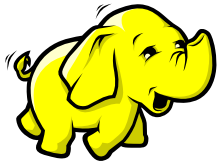
MapReduce

Poca eficiencia con archivos diminutos.

MapReducejob

- *Jobtracker*
- *Tasktrackers*

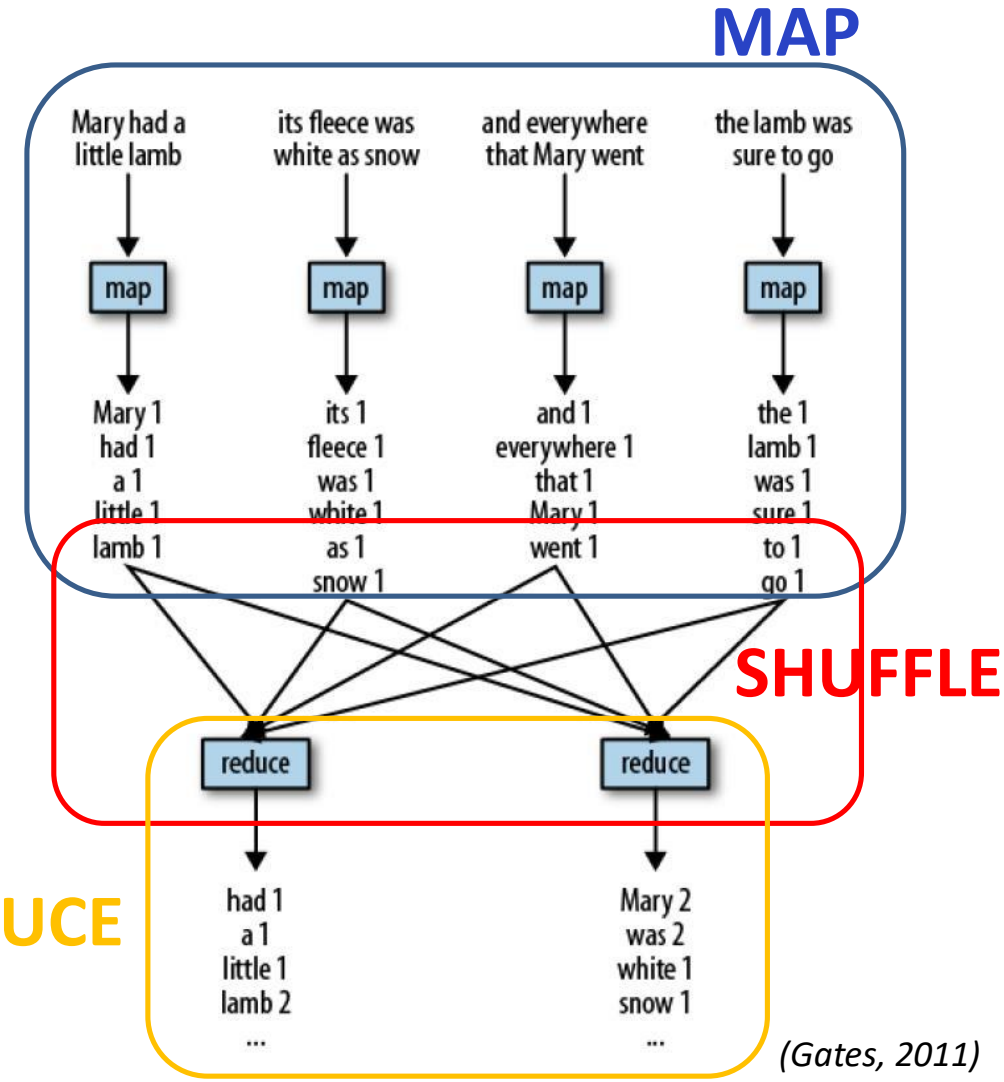
Hadoop



Map-Reduce

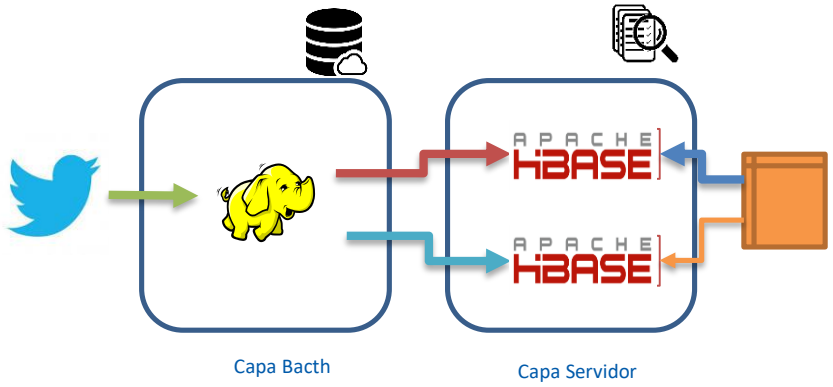
MapReducejob

- Maptasks
- Reduce tasks



(Gates, 2011)

HBase



- Base de datos columnar

Tablas dedicadas para análisis específicos extraídos de Hadoop

Ventajas:

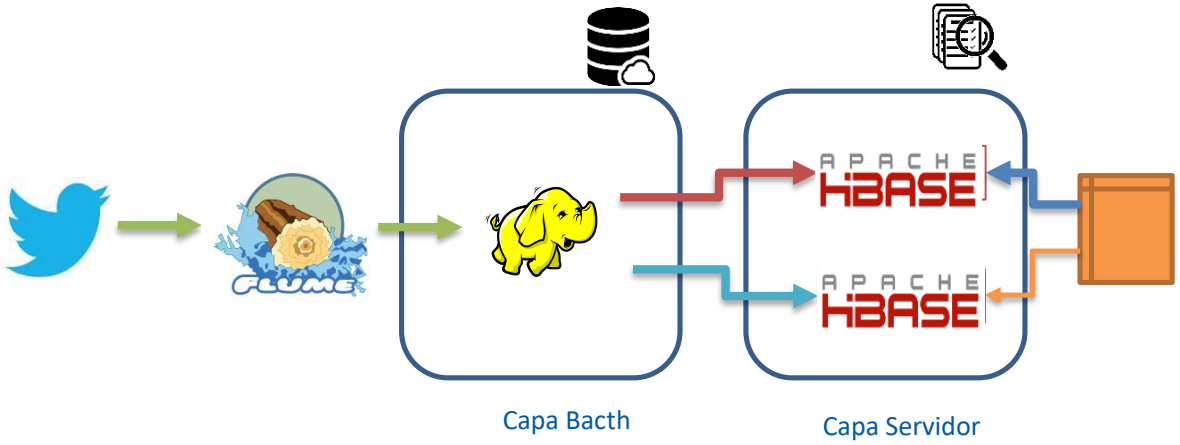
- Lectura aleatoria de datos en tiempo razonable

Inconvenientes:

- **`Joins`** costosos → Denormalización
- Elegir bien las **`rowkeys`** (eliminan duplicados)

Adquisición de datos

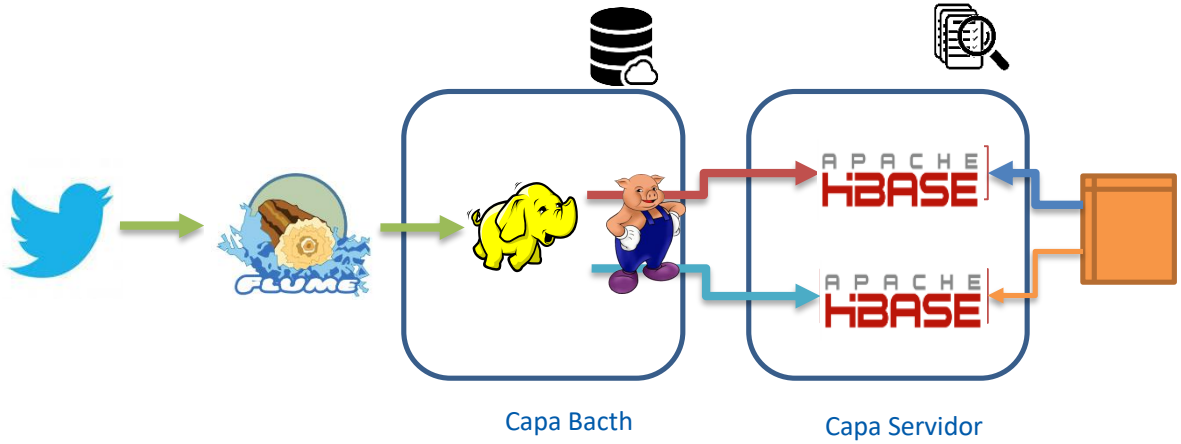
- Existen gran cantidad de APIs.
- Se necesita introducir la respuesta en formato JSON.
- **Apache Flume** es una herramienta para la ingestión de datos en HDFS. Es un servicio seguro, distribuido y de alta disponibilidad. Maneja datos semi-estructurados en JSON



Análisis de datos

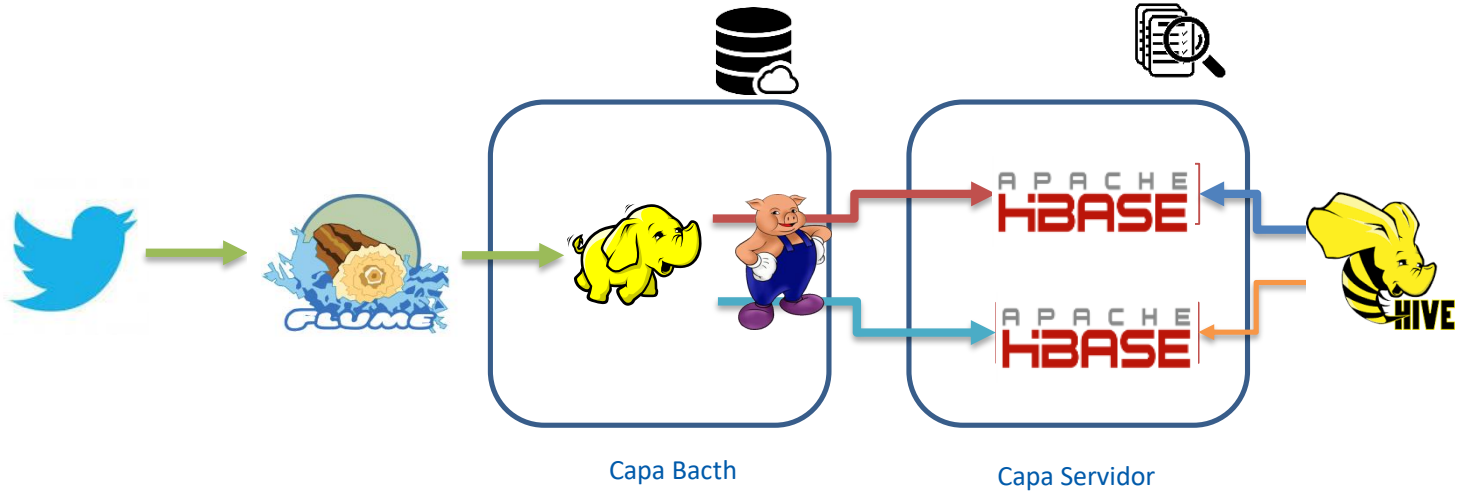
Se necesita analizar parte o todo el conjunto de dato JSON almacenados en bruto que contiene HDFS.

Pig. Motor para ejecutar flujos de datos en paralelo sobre Hadoop. Usa Map-Reduce para ejecutar todos su procesamiento de datos. Utiliza el lenguaje procedimental.



Consultas

Apache Hive proporciona un lenguaje que facilita el acceso a la información. Permite consultas al estilo SQL sobre HBASE



→ Gestión de los datos

Toma de tweets:

- *1.Geolocalizados*
- *2.Contiene palabras clave*
- *3.(1) OR (2)*

Análisis

- *1.UniformDate*
- *2.Related*
- *3.Coordinates*
- *4.Hashtags*
- *5.UserMentions*
- *6.MD5gen*

Creación de tablas:

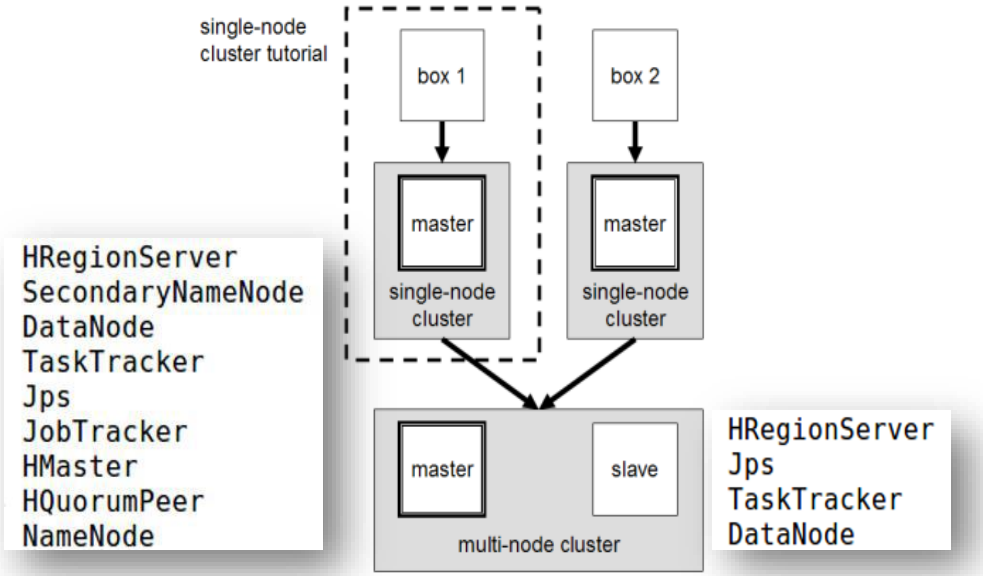
- *1.Tweets*
- *2.Menciones*
- *3.Mencionados por*

*User Defined Functions

→ Cluster Multi-Nodo

Cluster Multi Nodo .- Formado por la unión de varios cluster mono-nodo.

- **Master** : nodo maestro y también esclavo
- **Slave**: nodo esclavo



(Noll, Running Hadoop on Ubuntu Linux -Multi-Node Cluster, 2011)

Recolección de información

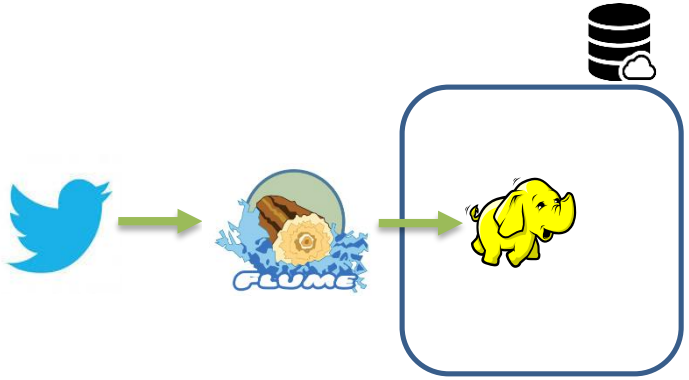
Tweets…

…geolocalizados

swLngLat= -9.299269, 35.999882
neLngLat= 4.327812, 43.79142

…que contienen alguna palabra clave

keywords = @Sanidad, @MUNQU_IMM, @UMEgob, #QuedateEnCasa, #EsteVirusLoParamosUnidos, #COVID19, #coronavirusmadrid, #CoronavirusEspaña #coronavirusEspana, #CoronavirusESP

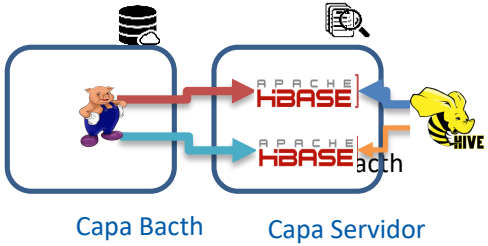


Capa Bacth

→ Casos de uso

Flujo de análisis y consulta

@pepeperez, 2020, 04, 13,@Sanidad, como organiza,#CoronavirusESP



twetts

twetts

twetts

MD5(id_twett)

JSON

mentions

mentions

mentions

@pepeperez,@Sanidad

MD5(idA) MD5(idB)

relationships

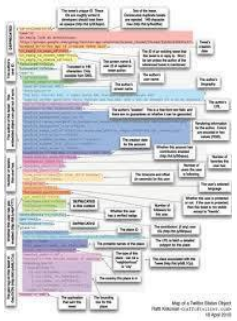
mentioned

mentioned

mentioned

@Sanidad,@pepeperez

MD5(idB) MD5(idA)



→ Referencias

- Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., &Liu, B. (2011). *Predicting Flu Trends using Twitter Data.* Retrieved from IEEE Computer Communications Workshops (INFOCOM WKSHPS), 702–707
- *Admin Manual Metastore Admin*. (n.d.). Retrieved from Apache Hive: https://cwiki.apache.org/confluence/display/Hive/AdminManual+MetastoreAdmin
- ApacheFlume. (n.d.). Retrieved from ApacheFlume: http://flume.apache.org/
- ApacheHadoop. (n.d.). Retrieved from ApacheHadoop: http://hadoop.apache.org/
- ApacheHBase. (n.d.). Retrieved from Apache HBase: https://hbase.apache.org/
- ApacheHive. (n.d.). Retrieved from ApacheHive: http://hive.apache.org/
- Apache Pig. (n.d.). Retrieved from Apache Pig: https://pig.apache.org/
- *Apache Pig Philosophy*. (n.d.). Retrieved from Apache Pig: https://pig.apache.org/philosophy.html
- *Built In Functions*. (n.d.). Retrieved from ApachePig: http://pig.apache.org/docs/r0.12.1/func.html

→ Referencias

- *Configuring the Hive Metastore*. (n.d.). Retrieved from Cloudera docs: http://www.cloudera.com/content/cloudera-content/cloudera-docs/CDH4/4.2.0/CDH4-Installation-Guide/cdh4ig_topic_18_4.html
- Dimiduk, N., & Khurana, A. (2013). *HBase in action.* New York: Manning.
- *Documentation*. (n.d.). Retrieved from Twitter Developers: https://dev.twitter.com/docs/
- *Elephant-bird*. (n.d.). Retrieved from GitHub: https://github.com/kevinweil/elephant-bird/
- *Flume User Guide*. (n.d.). Retrieved from ApacheFlume: https://flume.apache.org/FlumeUserGuide.html
- Gates, A. (2011). *Programming Pig.* Sebastopol, CA: O'Reilly Media, Inc.
- George, L. (2011). *HBase: The Definitive Guide.* Sebastopol, CA: O'ReillyMedia, Inc.
- http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3240992/