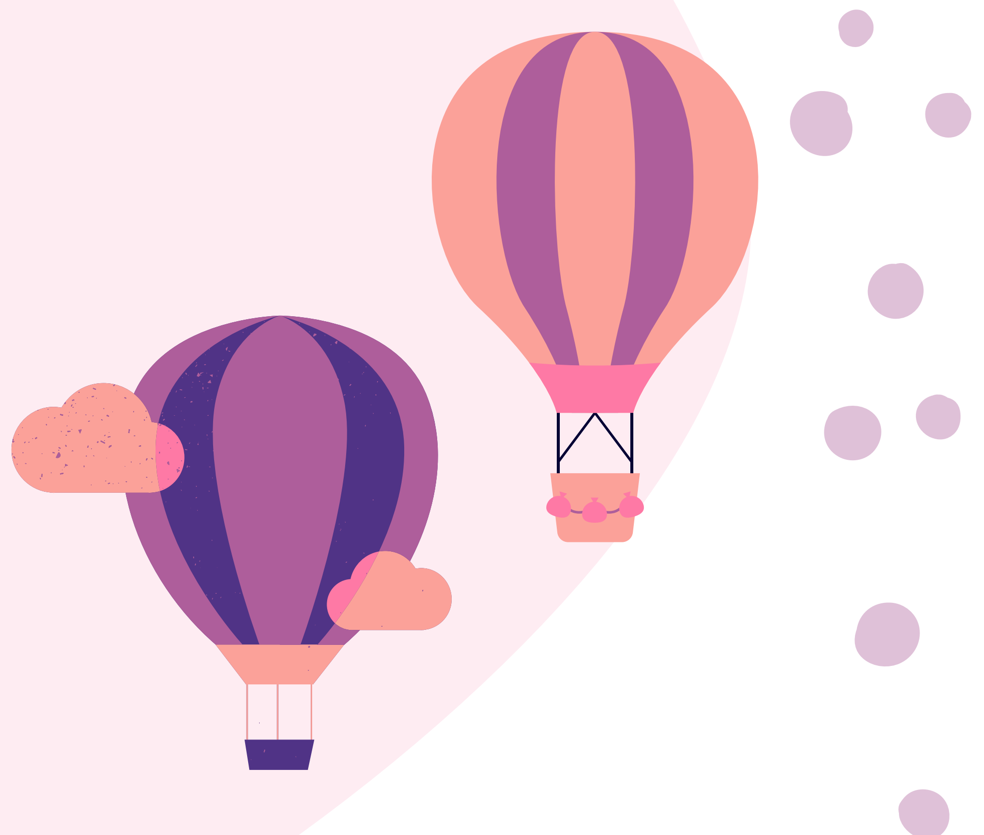
# STRUCTURED DATA STREAMING Y NLP CON PYSPARK

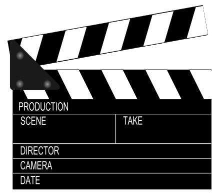
BY: VALEN ARIZA Y LAURA LÓPEZ

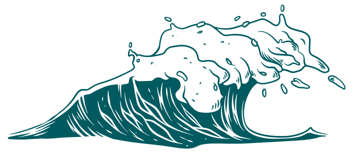


# SOBRE NOSOTRAS

## VALENTINA ARIZA

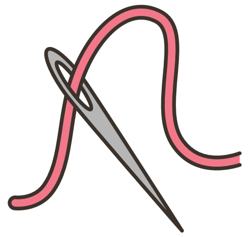
Data Engineer, DataKnow





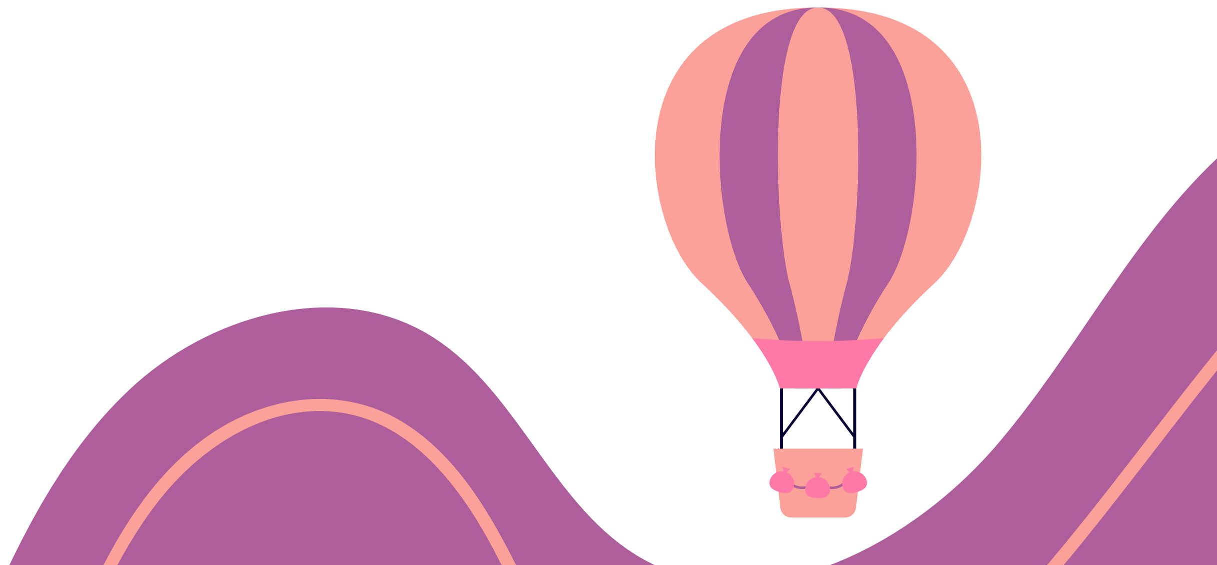
## LAURA LÓPEZ

Data Science Analyst, Accenture



# AGENDA

- Funcionamiento básico de Spark Structured Data Streaming
- Preprocesamiento de texto para NLP
- Proceso en un ambiente productivo
- Práctica

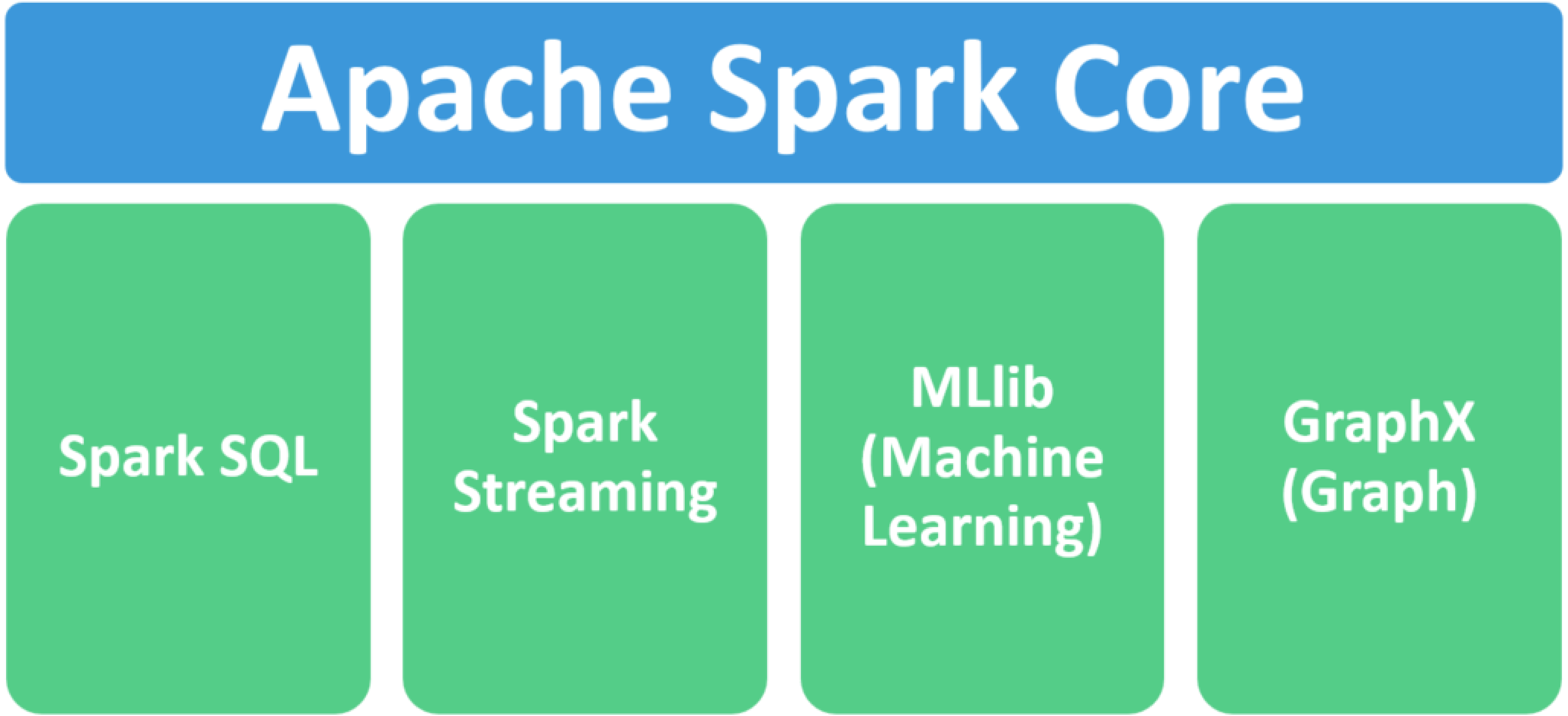


# PREREQUISITOS

- Crea una cuenta de Google Colab
- Crea una cuenta de Reddit https://www.reddit.com  (guarda tu usuario y contraseña)
- Prepárate... 💣

# ¿QUÉ ES  SPARK STRUCTURED DATA STREAMING?

# ¿QUÉ ES  SPARK STRUCTURED DATA STREAMING?



Apache Spark Core

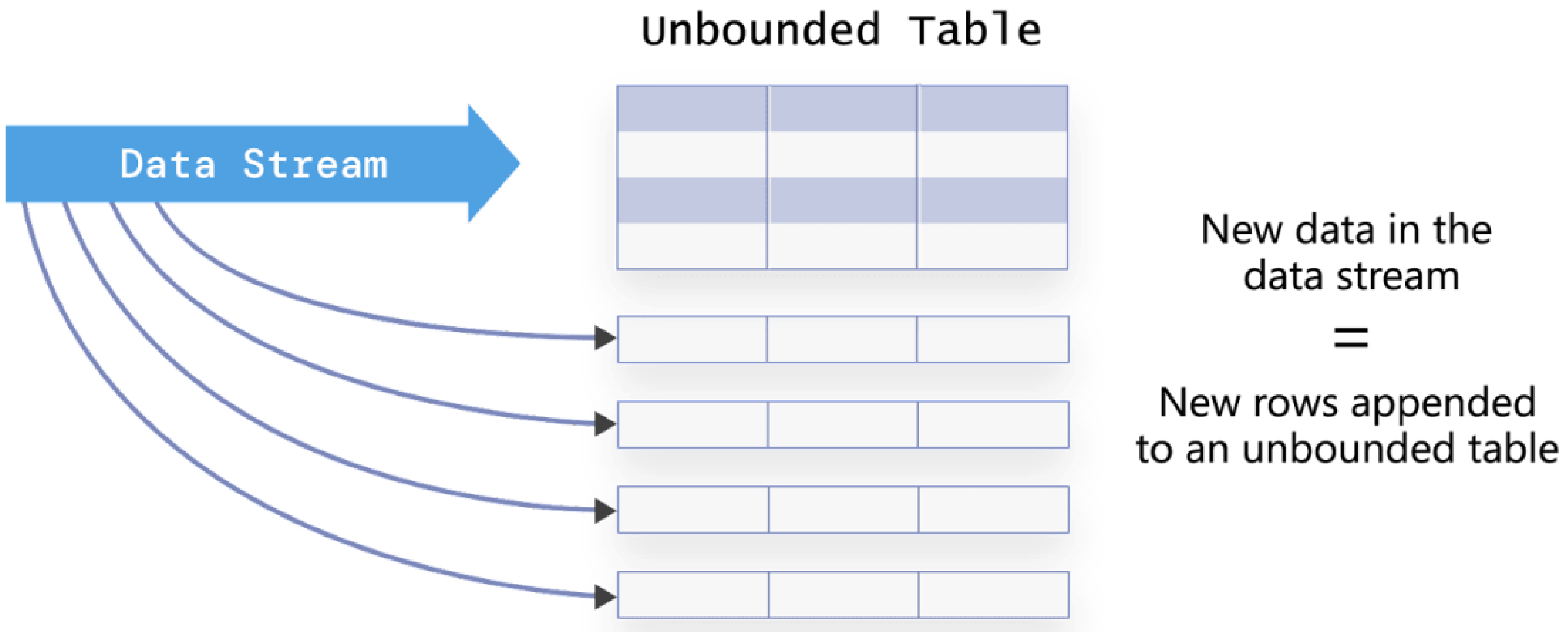
Spark SQL

Spark Streaming

MLlib (Machine Learning)

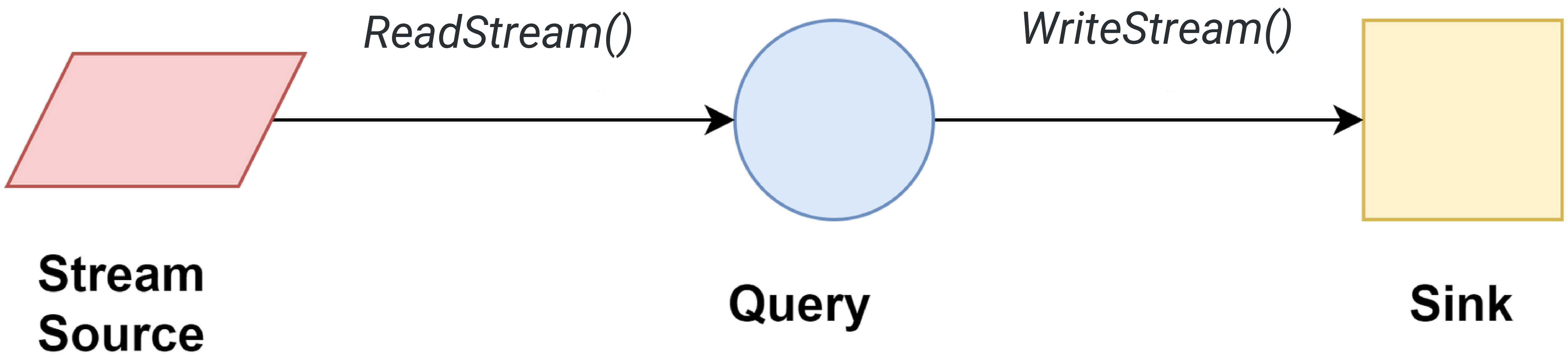
GraphX (Graph)

# ¿QUÉ ES SPARK STRUCTURED DATA STREAMING?


Unbounded Table
Data Stream
New data in the data stream
=
New rows appended to an unbounded table

Data stream as an unbounded table

# FUNCIONAMIENTO BÁSICO SPARK DATA STREAMING



# ORÍGENES Y DESTINOS

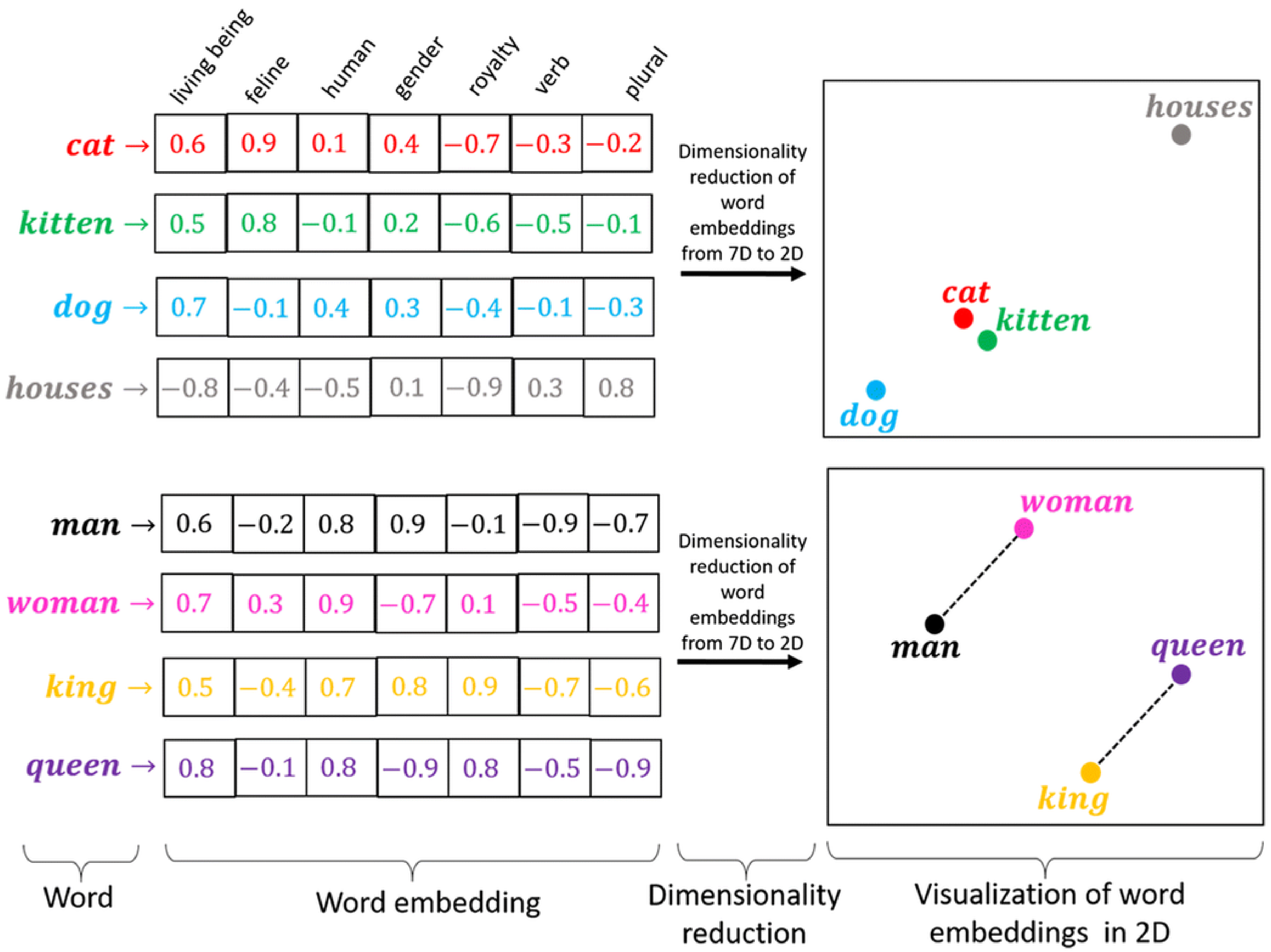
Apache Kafka

Conexiones
Stream Socket

Archivos

# ¿CÓMO FUNCIONA NATURAL LANGUAGE PROCESSING (NLP)?

# ¿CÓMO FUNCIONA NATURAL LANGUAGE PROCESSING (NLP)?



|  | living being | feline | human | gender | royalty | verb | plural |
|---|---|---|---|---|---|---|---|
| cat → | 0.6 | 0.9 | 0.1 | 0.4 | −0.7 | −0.3 | −0.2 |
| kitten → | 0.5 | 0.8 | −0.1 | 0.2 | −0.6 | −0.5 | −0.1 |
| dog → | 0.7 | −0.1 | 0.4 | 0.3 | −0.4 | −0.1 | −0.3 |
| houses → | −0.8 | −0.4 | −0.5 | 0.1 | −0.9 | 0.3 | 0.8 |
| man → | 0.6 | −0.2 | 0.8 | 0.9 | −0.1 | −0.9 | −0.7 |
| woman → | 0.7 | 0.3 | 0.9 | −0.7 | 0.1 | −0.5 | −0.4 |
| king → | 0.5 | −0.4 | 0.7 | 0.8 | 0.9 | −0.7 | −0.6 |
| queen → | 0.8 | −0.1 | 0.8 | −0.9 | 0.8 | −0.5 | −0.9 |

Dimensionality reduction of word embeddings from 7D to 2D
houses
cat
kitten
dog
Dimensionality reduction of word embeddings from 7D to 2D
woman
man
queen
king
Word
Word embedding
Dimensionality reduction
Visualization of word embeddings in 2D

# PREPROCESAMIENTO

*“Hey Amazon - my package never arrived*
*https://www.amazon.com/gp/css/order-history?ref_=nav_orders_first*
*PLEASE FIX ASAP! @amazonhelp”*

# PREPROCESAMIENTO

- Normalizar

*“hey amazon - my package never arrived https://www.amazon.com/gp/css/order-history?ref_=nav_orders_first please fix asap! @amazonhelp”*

⬇

*“hey amazon my package never arrived please fix asap”*

# PREPROCESAMIENTO

- Eliminar stopwords

*“hey amazon my package never arrived please fix asap”*

↓

*“amazon package never arrived fix asap”*

# PREPROCESAMIENTO

- Tokenizar

*“amazon package never arrived fix asap”*

↓

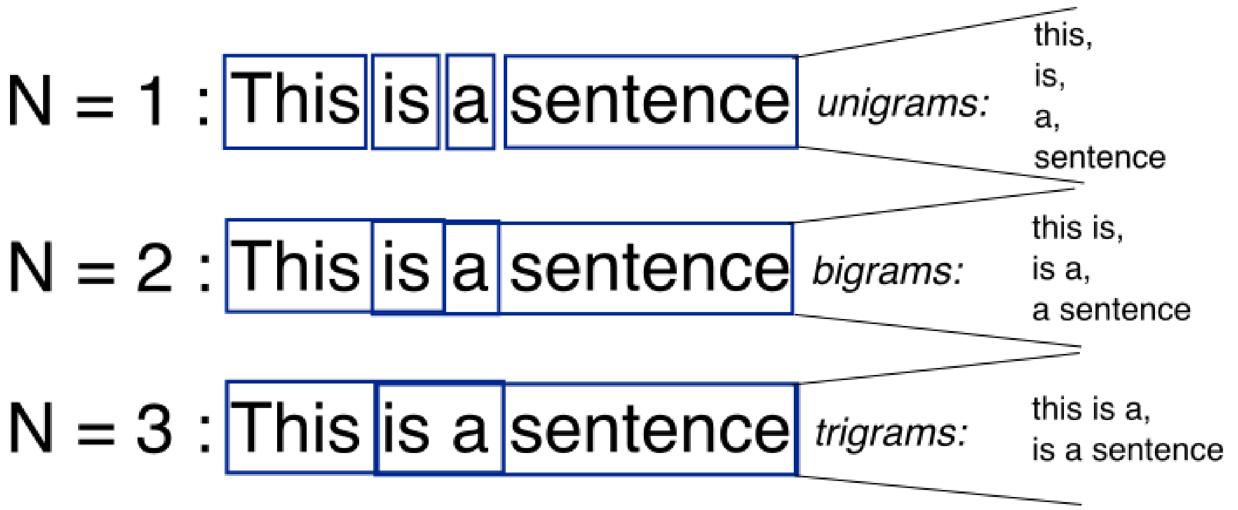
*["amazon", "package", "never", "arrived", "fix", "asap”]*

# PREPROCESAMIENTO

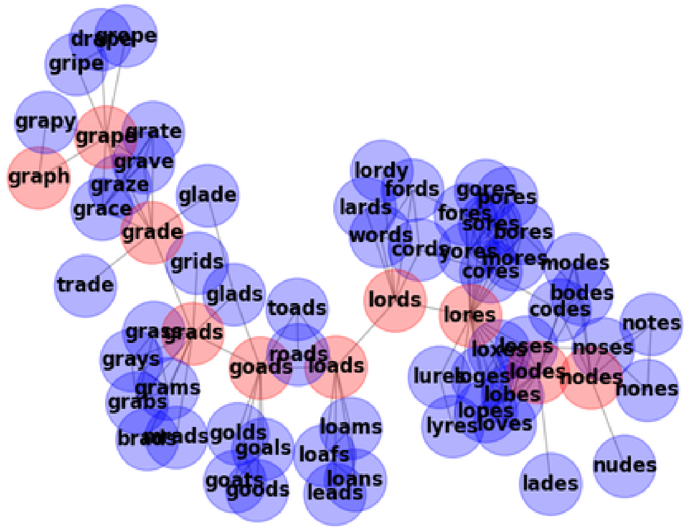
- Normalizar
- Stopwords
- Símbolos y caracteres especiales
- Tokenizar
- Stemming
- Lemmatization

# QUÉ PODEMOS HACER CON NLP

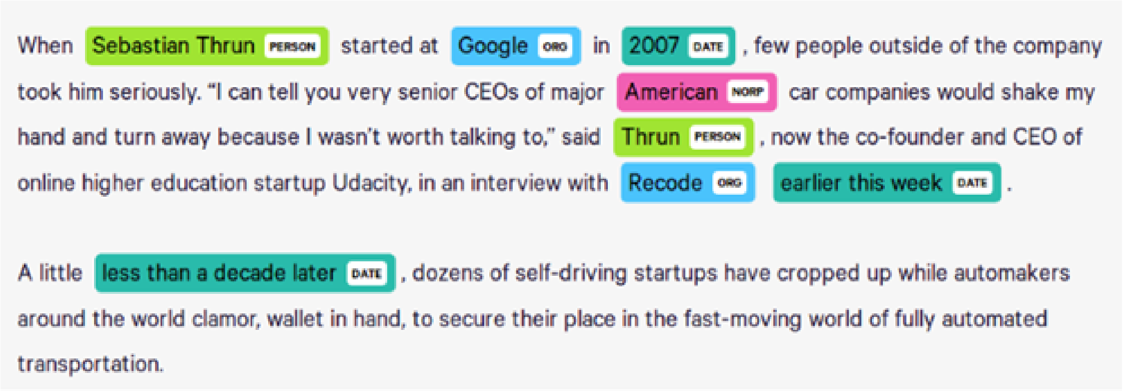
Wordcloud



N-Grams



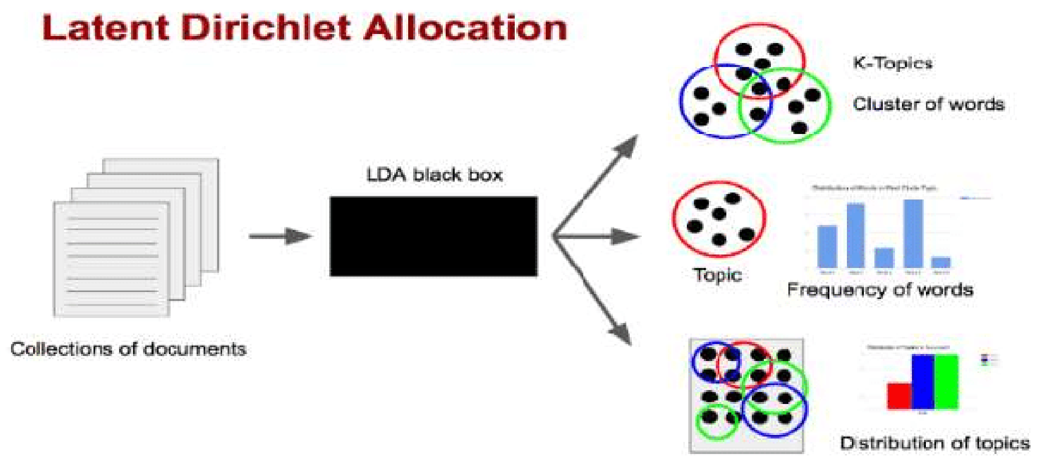
Grafos de palabras



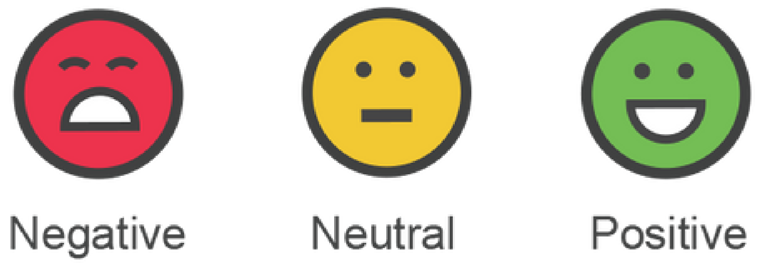
When Sebastian Thrun PERSON started at Google ORG in 2007 DATE , few people outside of the company took him seriously. “I can tell you very senior CEOs of major American NORP car companies would shake my hand and turn away because I wasn’t worth talking to,” said Thrun PERSON , now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode ORG earlier this week DATE .

A little less than a decade later DATE , dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

Named Entity Recognition (NER) y Part Of Speech Tagging (POS)



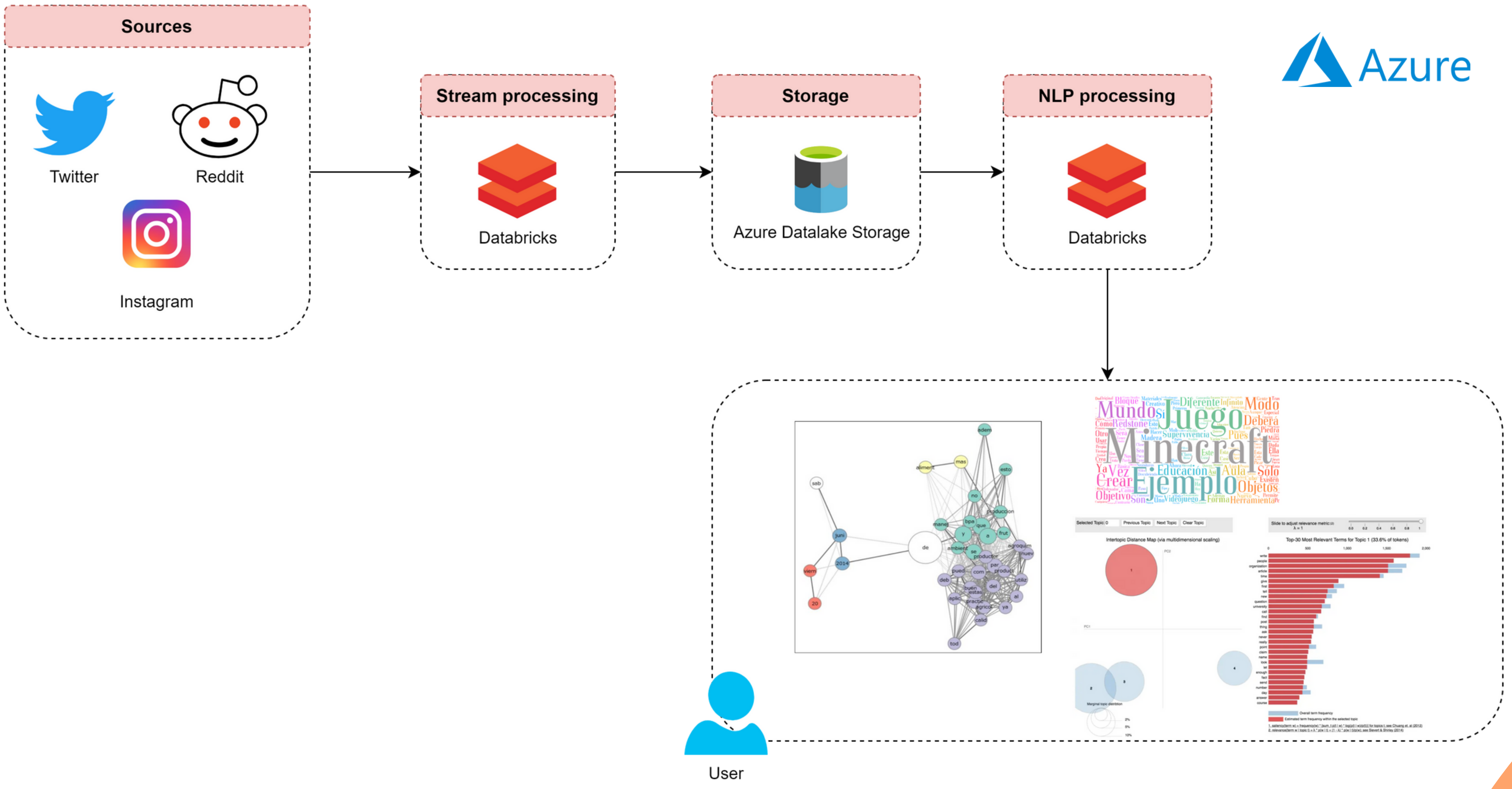
Latent Dirichlet Allocation (LDA)



Análisis de sentimientos

# ¿QUÉ VAMOS A HACER?

# PROCESO EN UN AMBIENTE PRODUCTIVO



Sources

Twitter

Reddit

Instagram

Stream processing

Databricks

Storage

Azure Datalake Storage

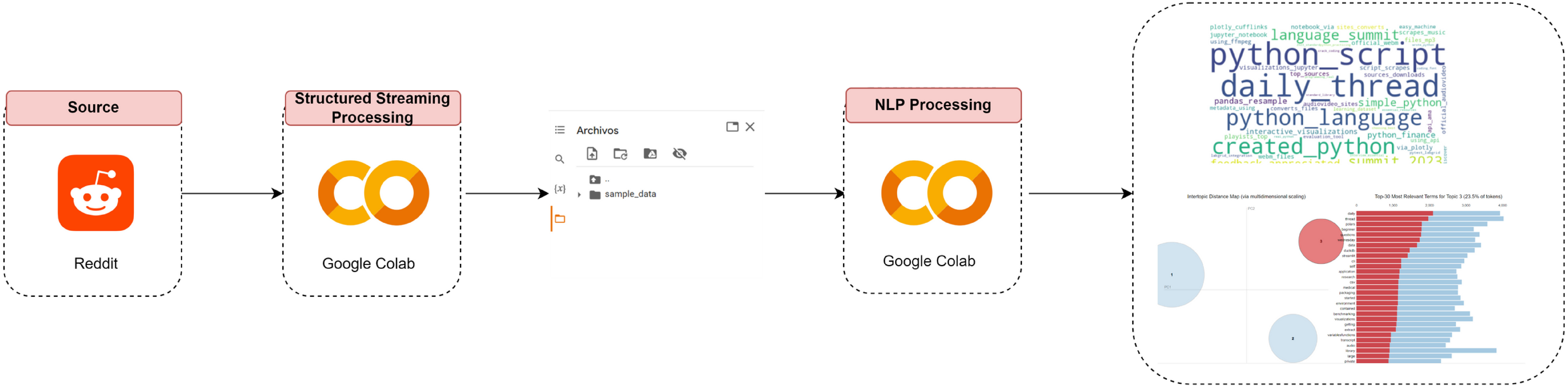
NLP processing

Databricks

Azure

User

# UNA ALTERNATIVA...



Source

Reddit

Structured Streaming Processing

Google Colab

NLP Processing

Google Colab

# MANOS A LA OBRA

# ¿TIENES PREGUNTAS?

¡CONTÁCTANOS!

@valearizag

@lauralpezb