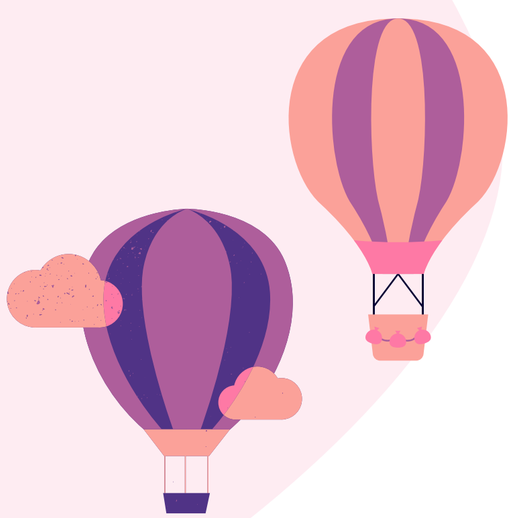
DATA STREAMING Y NLP CON PYSPARK

BY: VALEN ARIZA Y LAURA LÓPEZ



SOBRE NOSOTRAS

VALENTINA ARIZA

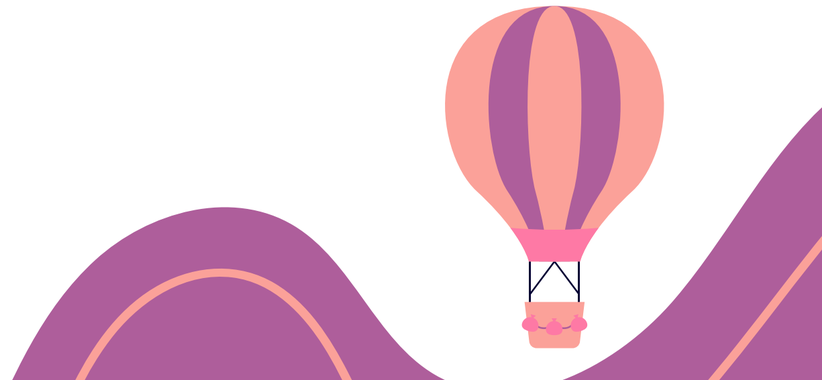
Data Engineer, DataKnow

LAURA LÓPEZ

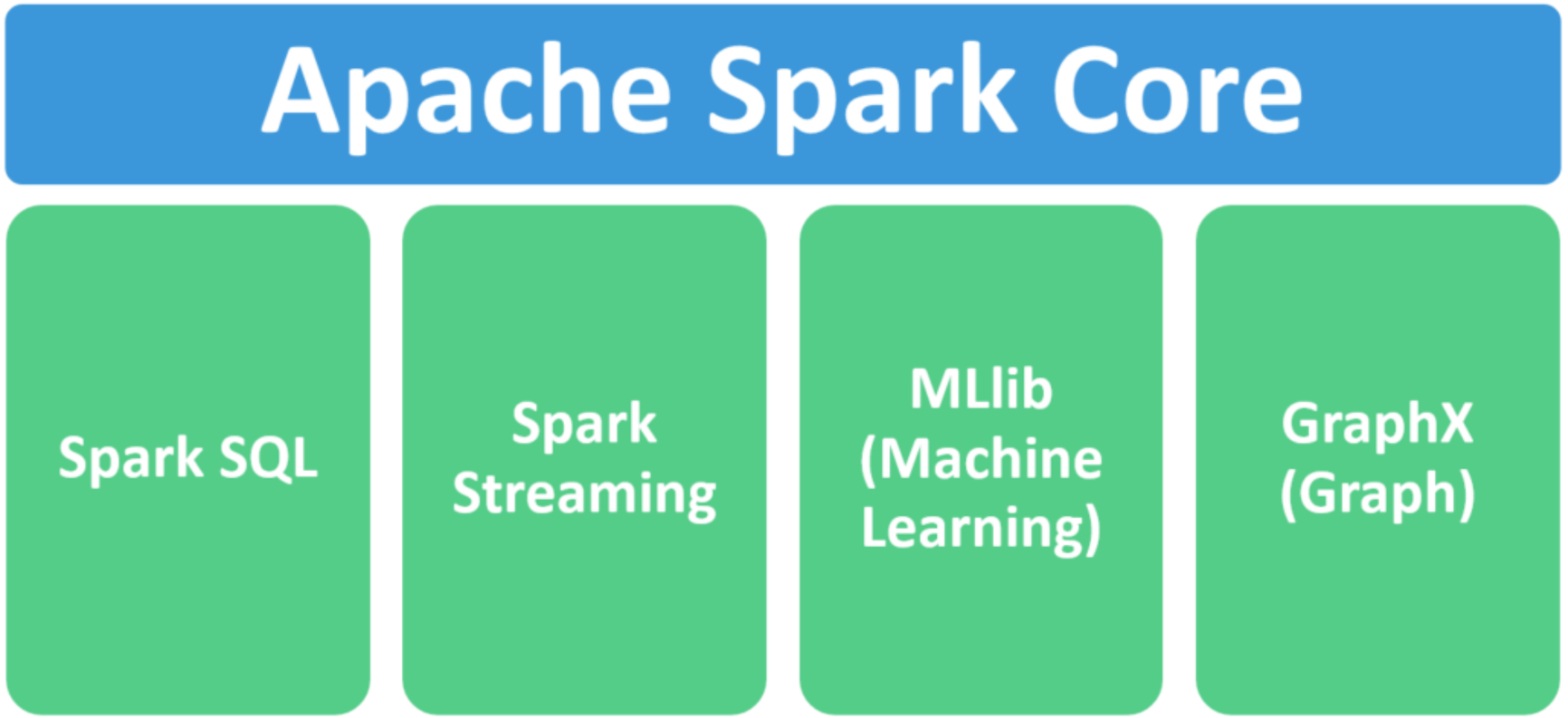
Data Science Analyst, Accenture

AGENDA

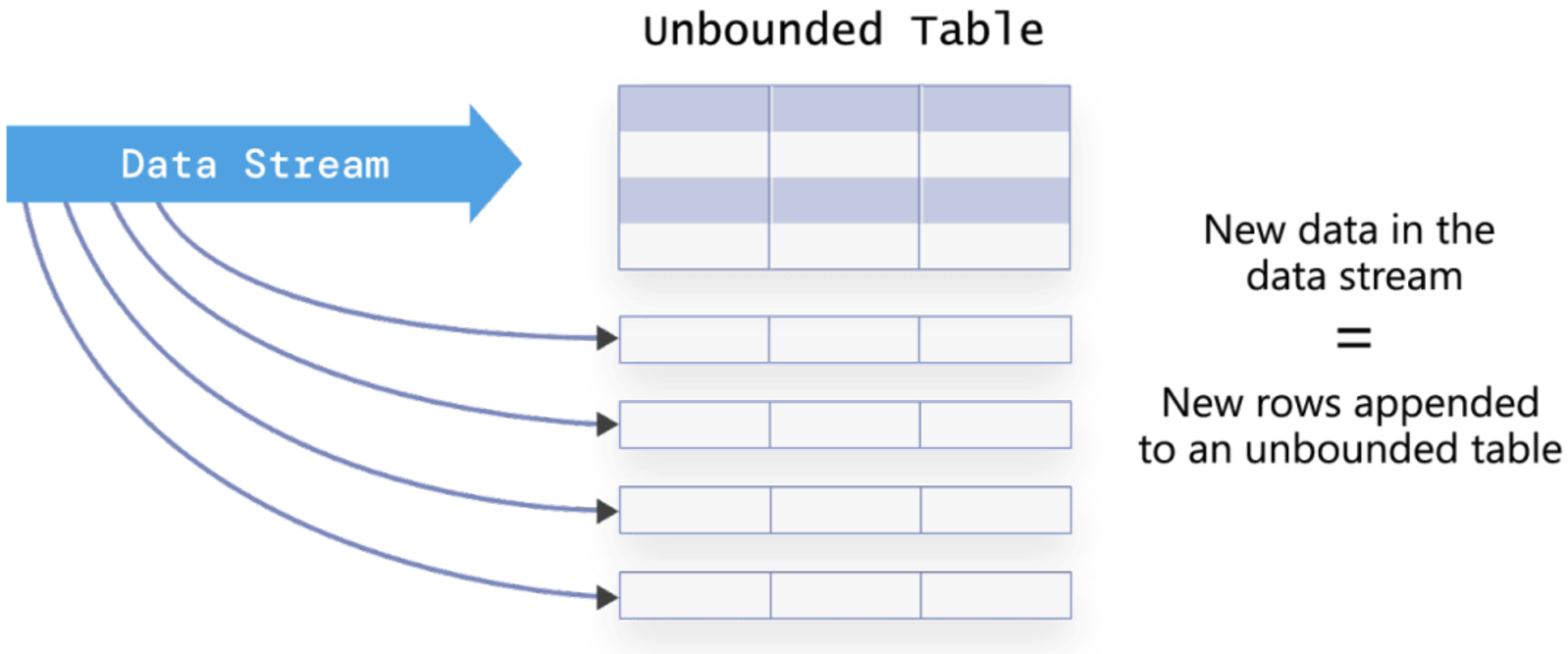
- Funcionamiento básico de Spark Structured Data Streaming
- Preprocesamiento de texto para NLP
- Proceso en un ambiente productivo
- Práctica



¿QUÉ ES SPARK STRUCTURED DATA STREAMING?

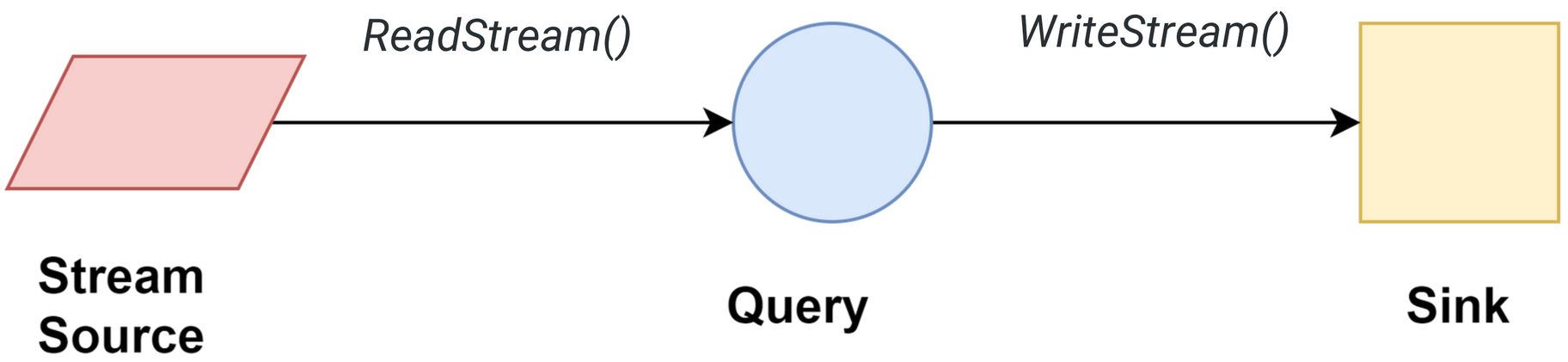


¿QUÉ ES SPARK STRUCTURED DATA STREAMING?



Data stream as an unbounded table

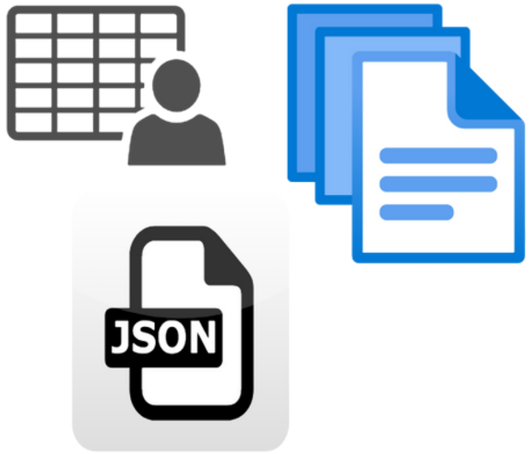
FUNCIONAMIENTO BÁSICO SPARK DATA STREAMING



ORÍGENES Y DESTINOS

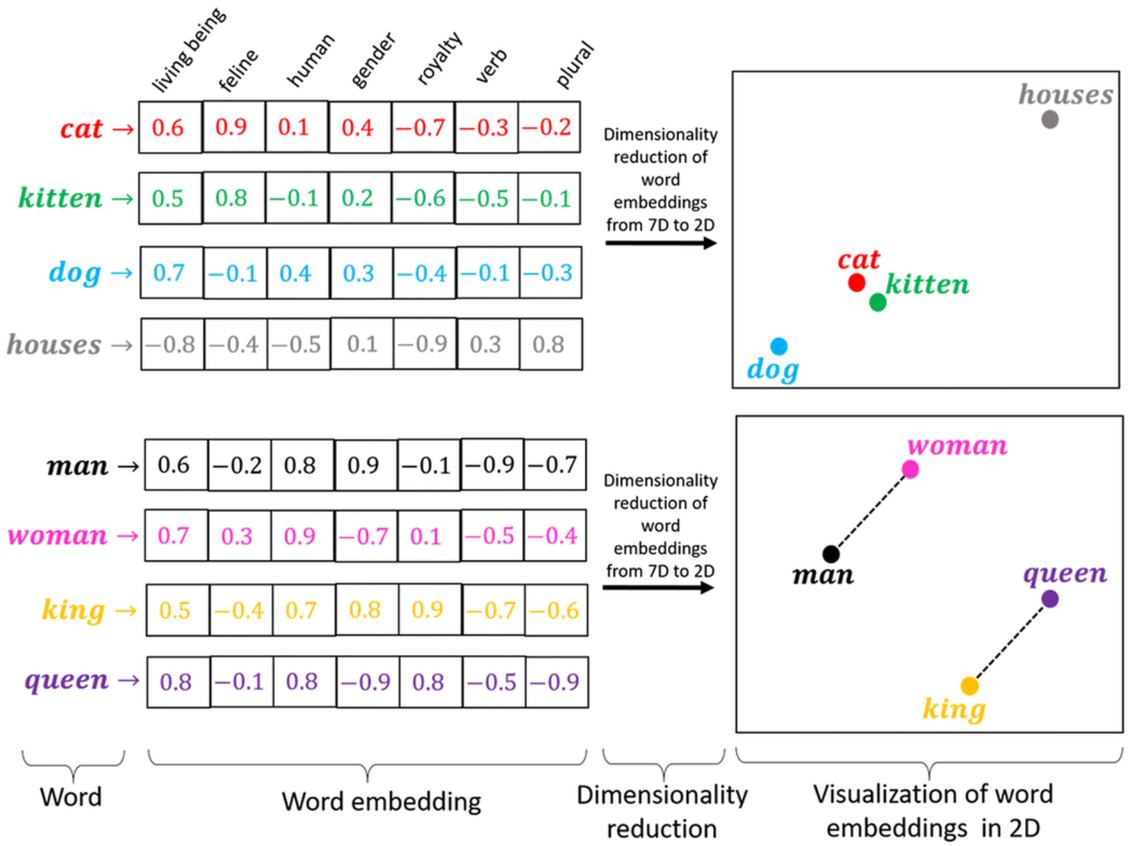
Apache Kafka

Conexiones
Stream Socket



Archivos

CÓMO FUNCIONA NATURAL LANGUAGE PROCESSING (NLP)



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”

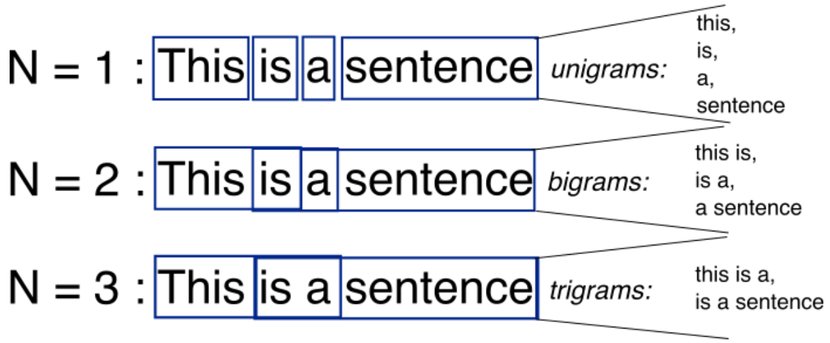
["amazon", "package", "never", "arrived", "fix", "asap”]

PREPROCESAMIENTO

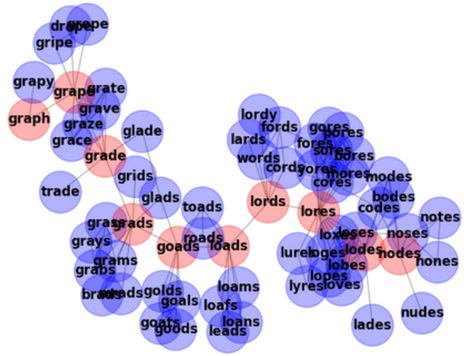
- Normalizar
- Stopwords
- Símbolos y caracteres especiales
- Tokenizar
- Stemming
- Lemmatization

QUÉ PODEMOS HACER CON NLP

Wordcloud



N-Grams

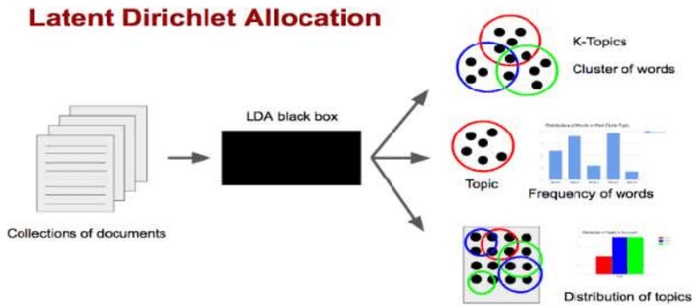


Grafos de palabras

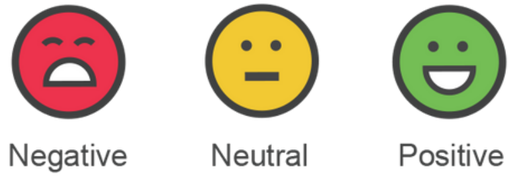
When Sebastian Thrun PERSON started at Google ORG in 2007 DATE, few people outside of the company took him seriously. "I can tell you very senior CEOs of major American NORP car companies would shake my hand and turn away because I wasn't worth talking to," said Thrun PERSON, now the co-founder and CEO 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)

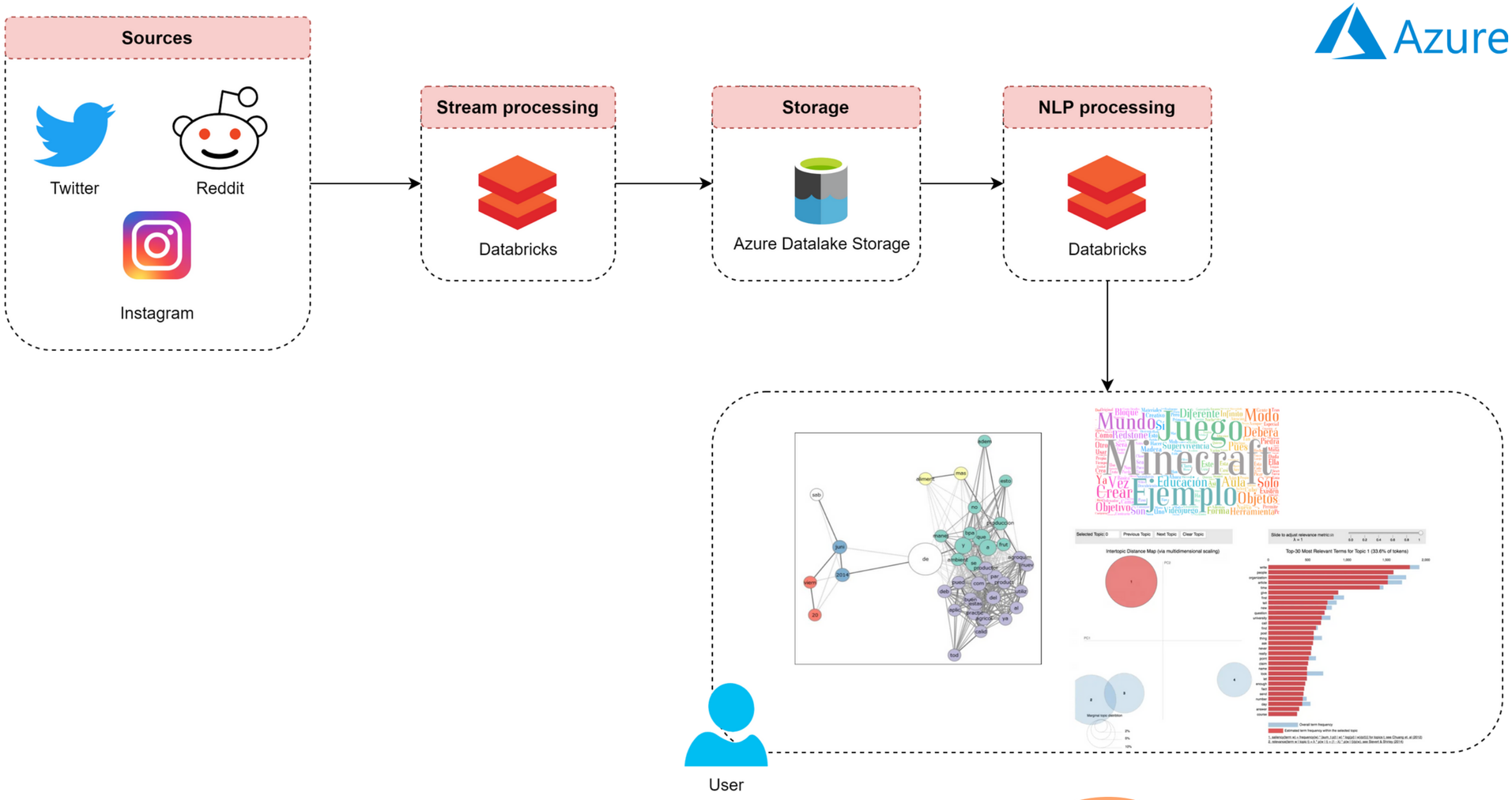


Latent Dirichlet Allocation (LDA)



Análisis de sentimientos

PROCESO EN UN AMBIENTE PRODUCTIVO



MANOS A LA OBRA

RECURSOS NLP

- Documentación oficial de las librerías utilizadas
- Hugging Face: https://huggingface.co/
- Gensim: https://pypi.org/project/gensim/
- NLTK: https://www.nltk.org/

¿TIENES PREGUNTAS?

¡CONTÁCTANOS!

@valearizag

@lauralpezb