Aplicaciones del aprendizaje profundo en el análisis de grandes volúmenes de datos

Especialización en Ciencia de Datos

Mg. Diego Encinas – Ing. Román Bond

Agenda-Clase 4

Spark Streaming

- Flujos de datos
- Tratamiento de flujos de datos
- Algoritmos de streaming
- Funcionamiento de Spark Streaming

MLlib

Flujo de datos

- El flujo de datos es continuo
  - La frecuencia de la llegada de los datos depende del problema
- Los datos son recolectados en tiempo real
- No se almacenan para entrenar el modelo

Flujos de datos - Fuentes

- Redes Sociales: Twitter, Facebook, Instagram.
- Flujos de transacciones: Bancarias o criptomonedas (Bitcoin).
- Monitoreo de redes: Detección de intrusiones en la red, logs de servidores.
- Monitoreo en tiempo real de sensores + Internet de las Cosas (IoT).
- Análisis climático
- Análisis de información generada por dispositivos wearable.

Estrategias para el tratamiento del flujo

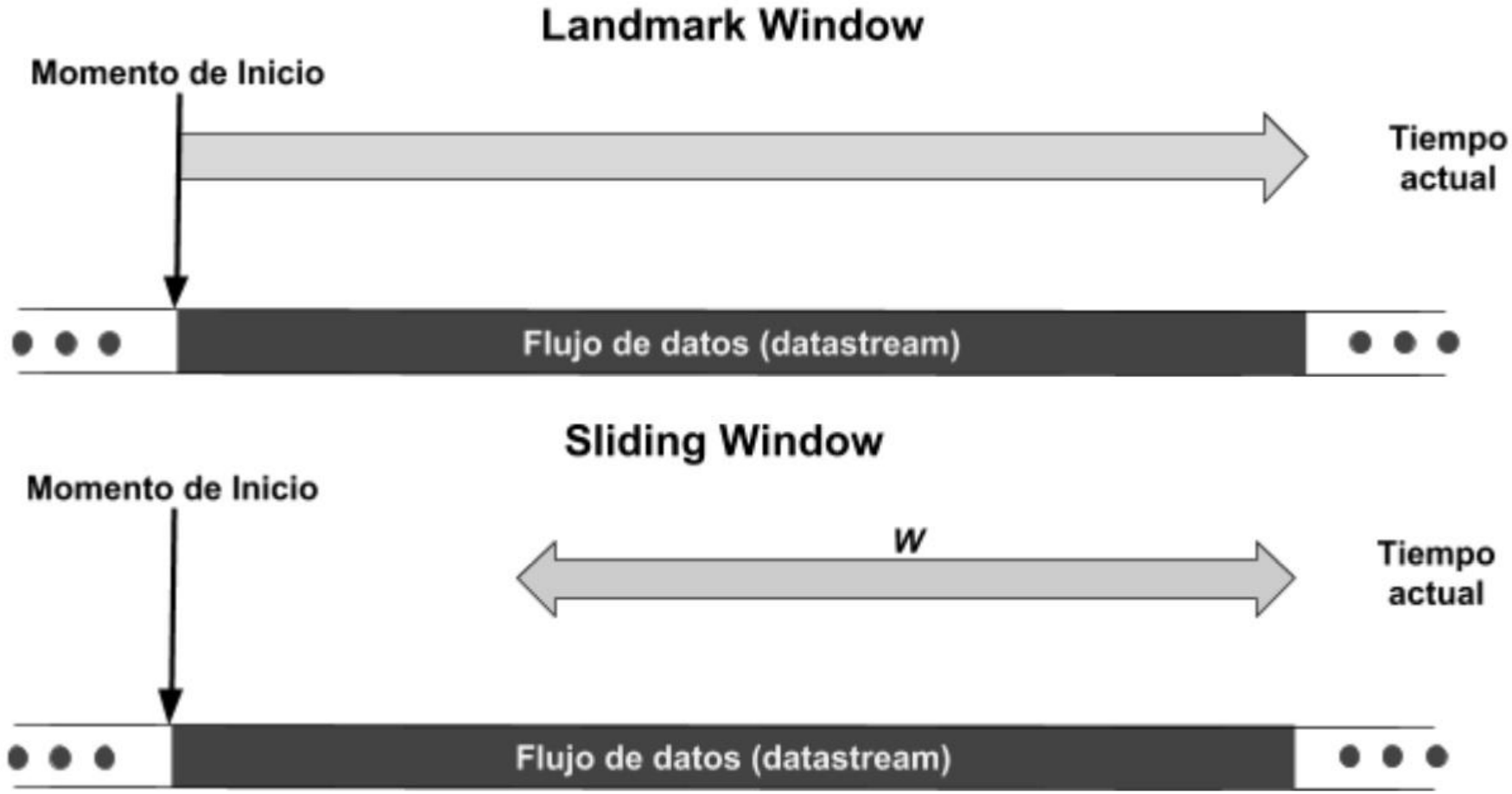
- El dato se recibe, se utiliza y se descarta

- Ventana temporal para guardar los últimos n datos recibidos

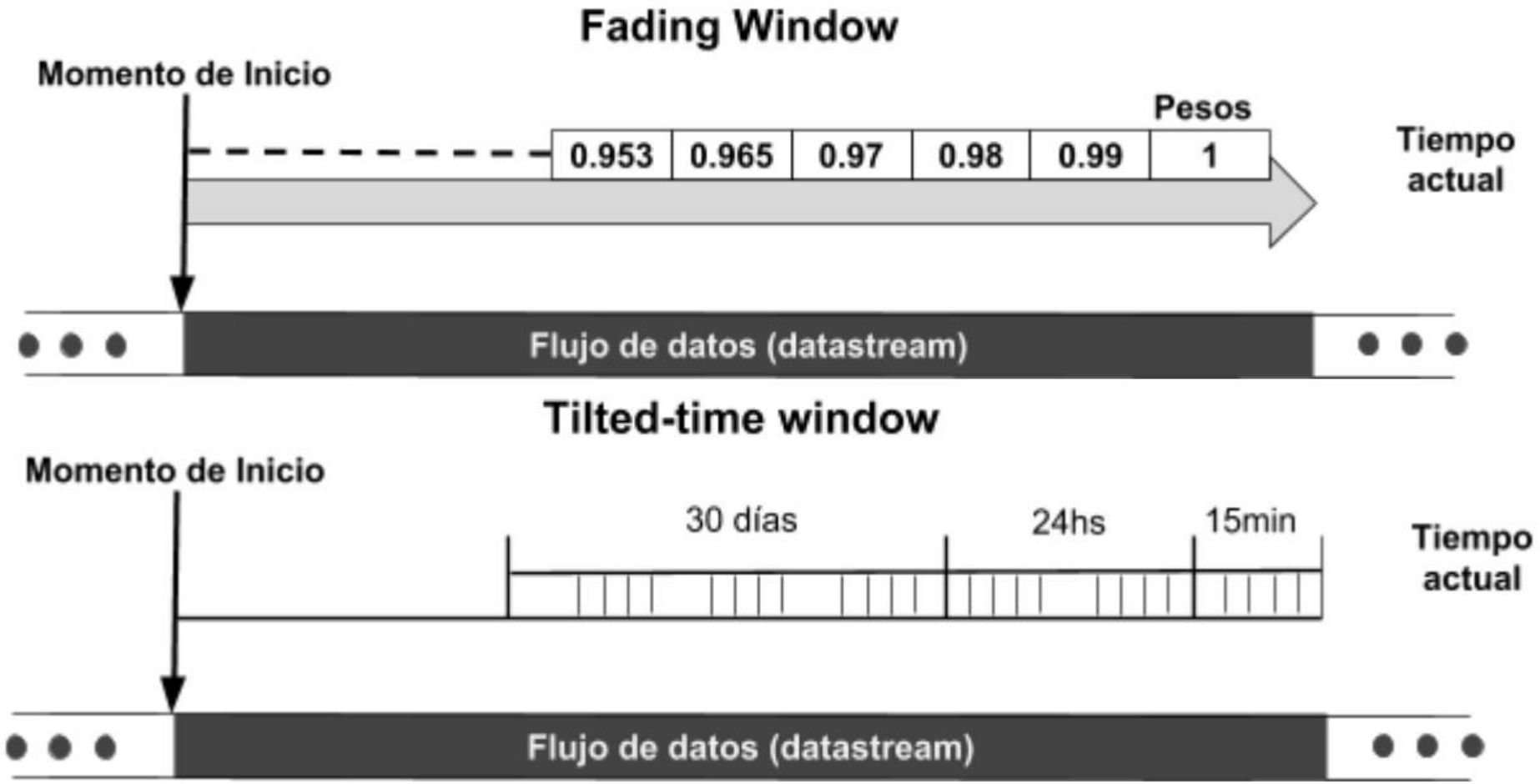
Ventanas de tiempo

- Landmark Window
- Sliding Window
- Fading Window (Damped Window)
- Tilted Time Window

Ventanas de tiempo



Ventanas de tiempo



Uso de algoritmos de streaming

Por lo general se utilizan como clasificadores

- El modelo puede estar entrenado de antemano y usarlo sobre el streaming
- El modelo se entrena con el propio streaming
  - Esta variante puede seguir entrenando y actuar como predictor al mismo tiempo

Stream processing

Un algoritmo de streaming debe cuidar tres aspectos:

- Velocidad. Debe poder operar un nuevo dato en el menor tiempo posible.
- Memoria. Debe ocupar la menor cantidad de memoria RAM.
- Eficacia. Debe poder clasificar nuevos datos con la mayor eficacia posible.

Spark streaming

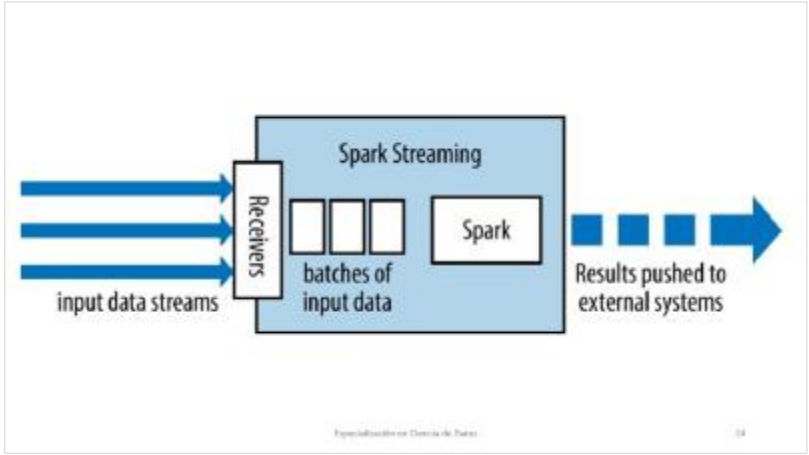
Spark streaming no es 100% streaming.

- Por cuestiones de eficiencia y compatibilidad, Spark streaming guarda el stream en pequeños "chunks" ejecutando pequeños procesos batch (micro-batch)

Spark streaming

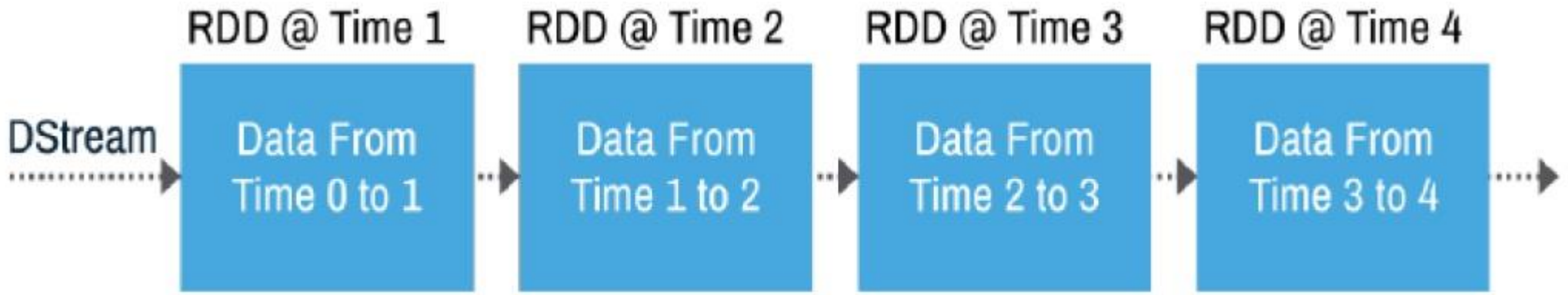
Spark streaming está diseñado para alimentarse desde varias fuentes de datos:

- Apache Kafka
- Apache Flume
- Amazon Kinesis
- Twitter
- Sensores u otros dispositivos via TCP sockets



Spark streaming

- En Spark un stream es representado como un stream discreto (DStream) el cual es una secuencia de RDDs.
- Cada RDD es un snapshot de todos los datos recolectados durante un período de tiempo, el cual luego se procesa como un batch.



Machine Learning

Un sistema inteligente es aquel sistema capaz de resolver problemas complejos y multidisciplinares de una forma automática dando soporte a las decisiones de un experto.

- Algoritmos simbolistas. Razonamiento inductivo
- Redes neuronales artificiales
- Algoritmos genéticos y evolutivos
- Probabilísticos. Teorema de Bayes.
- Algoritmos de "similitudes". K-NN, SVM

Sistemas inteligentes en Big Data

- Los algoritmos que implementan los sistemas inteligentes son algoritmos iterativos, por lo tanto tienen que dar varias "pasadas" a los datos para llevar a cabo su tarea.
- Se los conocen como algoritmos de aprendizaje de máquina (machine learning).
- Deben estar optimizados para un óptimo rendimiento.

MLlib

- MLlib es la librería de algoritmos de machine learning para Spark.
- Los algoritmos están diseñados e implementados para ejecutarse de manera eficiente en un ambiente distribuido

MLlib - Algoritmos

- Logistic regression
- Naive Bayes
- Generalized linear regression
- Survival regression
- Decision trees
- Random forests
- Gradient-boosted trees
- Alternating least squares (ALS)
- K-means
- Gaussian mixtures
- Latent Dirichlet allocation (LDA)
- Frequent itemsets
- Association rules
- Sequential pattern mining

Preguntas? O …