Seminario de Estadística I
Introducción a la Ciencia de Datos y Machine Learning

Jimmy Hernández Morales

Universidad Nacional Autónoma de México
Facultad de Ciencias

Contenido

Objetivos

- Introducción básica a la Ingeniería de Datos, aprender el uso de frameworks y tecnologías para el desarrollo de aplicaciones de Big Data.
- Revisión de los diferentes algoritmos de Machine Learning desde el punto de vista del Aprendizaje Estadístico.
- Aprender a implementar algoritmos de Machine Learning en ambientes Big Data.
- Presentar los nuevos paradigmas y temas de investigación en el área.

Evaluación

Evaluación

- Tareas y Prácticas 50 %
- Proyecto 20 %
- Exámenes conceptuales 30 %

Requisitos

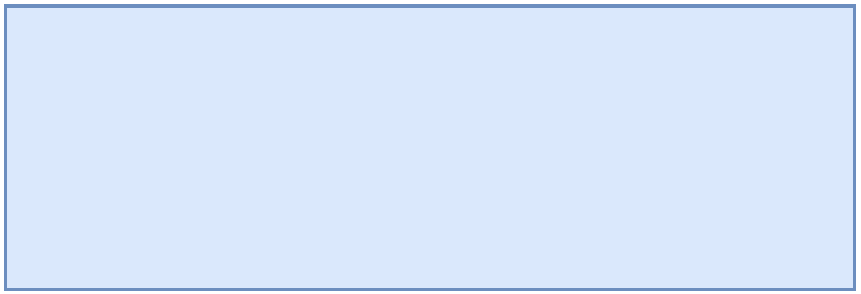
- Programación o ICC
- Manejo de Datos
- Inferencia Estadística
- Modelos no paramétricos y Regresión

Definiciones

Definición (Ciencia de Datos)

- Consiste en el estudio sobre la adquisición de datos, almacenamiento, comunicación, análisis, modelado, y algoritmos escalables para el análisis de los datos.
- Es la extracción de conocimientos usando Matemáticas, Estadística , Machine Learning, Ciencias de la Computación, Ingeniería...
- Big Data potencializa el éxito de la predicción estadística y la inteligencia artificial.

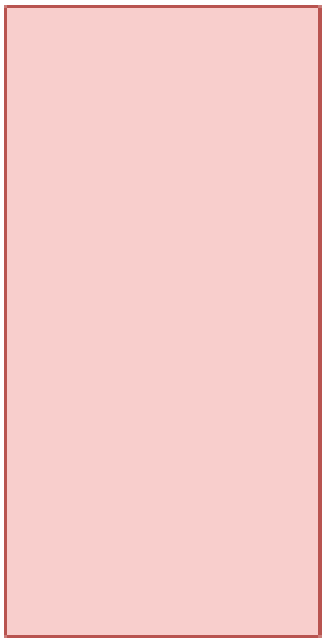
Wide Data



- Miles/ Millones de variables
- Cientos de muestras .

Surgen problemas como overfitting.
Se tiene que eliminar variables o regularizar.
SVM, Lasso, Stepwise

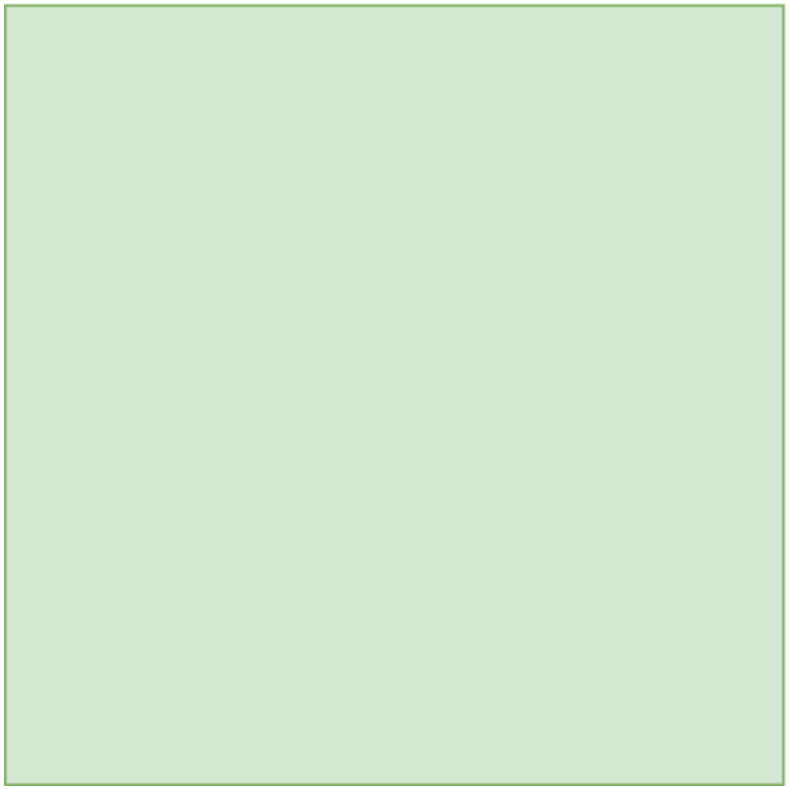
Tall Data



- Decenas/ centenas de variables
- Miles/ millones de muestras.

No siempre es suficiente una regresión lineal.
Se pueden aplicar modelos no lineales con muchas interacciones y no tantas varibles.
GLM, Random Forest, Deep Learning

Tall-Wide Data



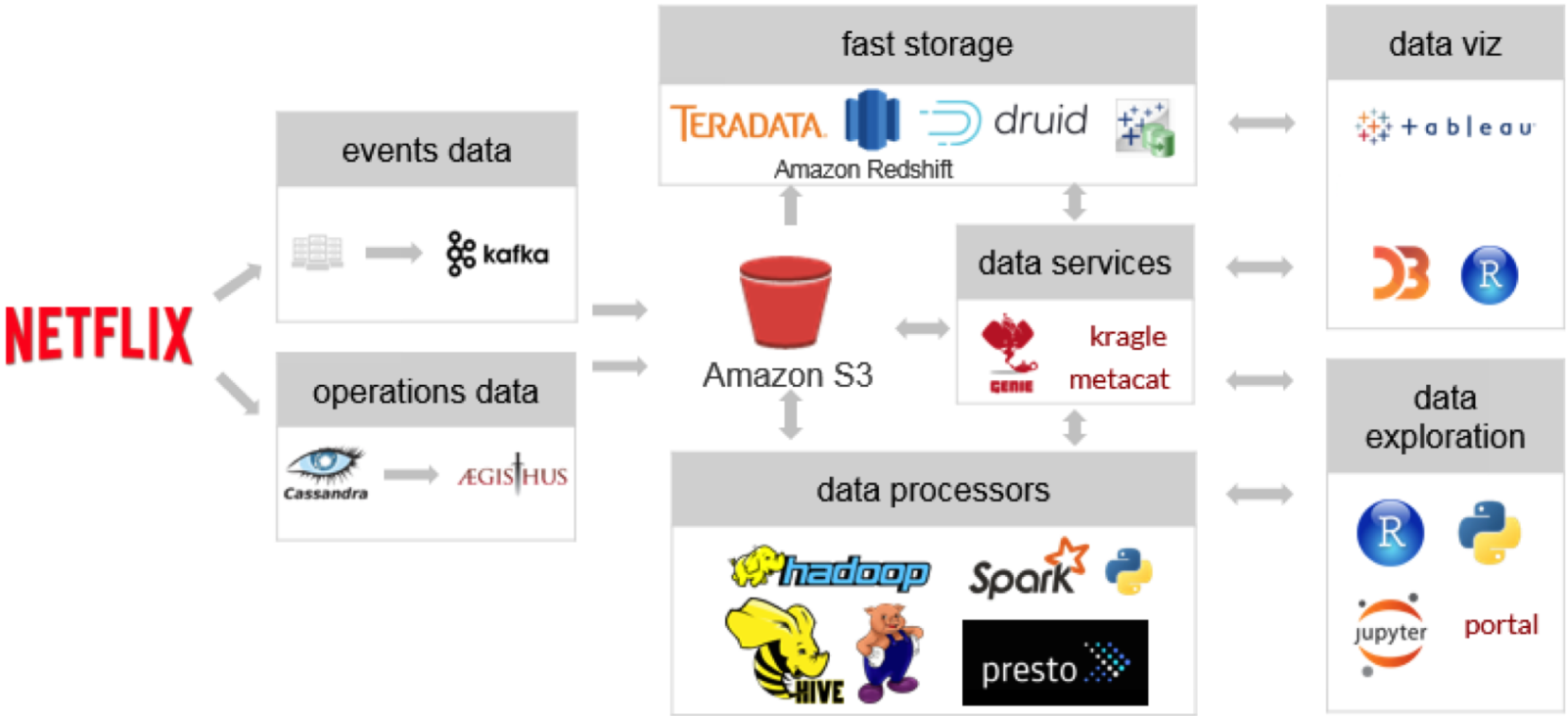
- Miles / Millones de variables
- Millones a Billones de muestras.

Divide y Recombina
MapReduce
(ADMM)Divide y Venceras

Las V de Big Data

Definición (Las 7 V's)

- **Volumen**. Cantidad de información: terabytes, registros, transacciones, tablas y archivos
- **Velocidad**: lote, prácticamente en tiempo real, en tiempo real y transmisiones
- **Variedad**. Son las formas, tipos y fuentes: estructurados, no estructurados y semiestructurados.
- **Visualización**. La manera en la que son presentados los datos.
- **Veracidad**. El grado de fiabilidad de los datos.
- **Viabilidad**.
- **Valor**.


NETFLIX
events data
kafka
operations data
Cassandra
ÆGISTHUS
fast storage
TERADATA.
Amazon Redshift
druid
Amazon S3
data services
kragle
metacat
GENIE
data processors
hadoop
Spark
HIVE
presto
data viz
+ableau
R
data exploration
R
jupyter
portal

Antecedentes

Definición

- Machine Learning es un término acuñado por Arthur Samuel, científico de la computación mientras trabajaba en IBM
- Machine Learning es el estudio de algoritmos que mejoran automáticamente a través de la experiencia.
- "Se dice que un programa de computadora aprende de la Experiencia E con respecto a alguna clase de tarea T y con medida de rendimiento P, si el desempeño sobre la tarea T, medido por P, mejora con la experiencia E."