Implementando modelos en Producción

Una revisión de las opciones disponibles

Luis Cajachahua

Senior Data Scientist, Americas Center Of Excelence

Mayo 2021

https://www.linkedin.com/in/lcajachahua/

Agenda

- Motivación
- Pintando la cancha
- ¿Qué modalidades de implementación existen?
  - Piedra, Papel y Tijeras
  - ¿Modelos en Excel?
  - Implementando en SQL
  - Aparecen las suites de Analítica Avanzada
  - PMML y los formatos estándar
  - Empaquetando modelos
  - API/REST
  - Orquestando APIs
- Conclusiones

Los retos más grandes en Advanced Analytics

Muy pocas organizaciones logran implementar modelos de una forma efectiva

Fuente: White A (2019), Gartner Predicts 2019 https://blogs.gartner.com/andrew_white/2019/01/03/our-top-data-and-analytics-predicts-for-2019/

80%

El pre-procesamiento de datos hasta hoy ocupa el 80% del tiempo y costo en los proyectos analíticos, **reduciendo la productividad del equipo de data science y afectando el time-to-market.**

65%

Gartner estimó que **la tasa de fracaso en los proyectos analíticos excede el 80%**; aunque se haya desarrollado un buen modelo predictivo. Las investigaciones sugieren que en el 65% de los casos el modelo no se lleva a producción.

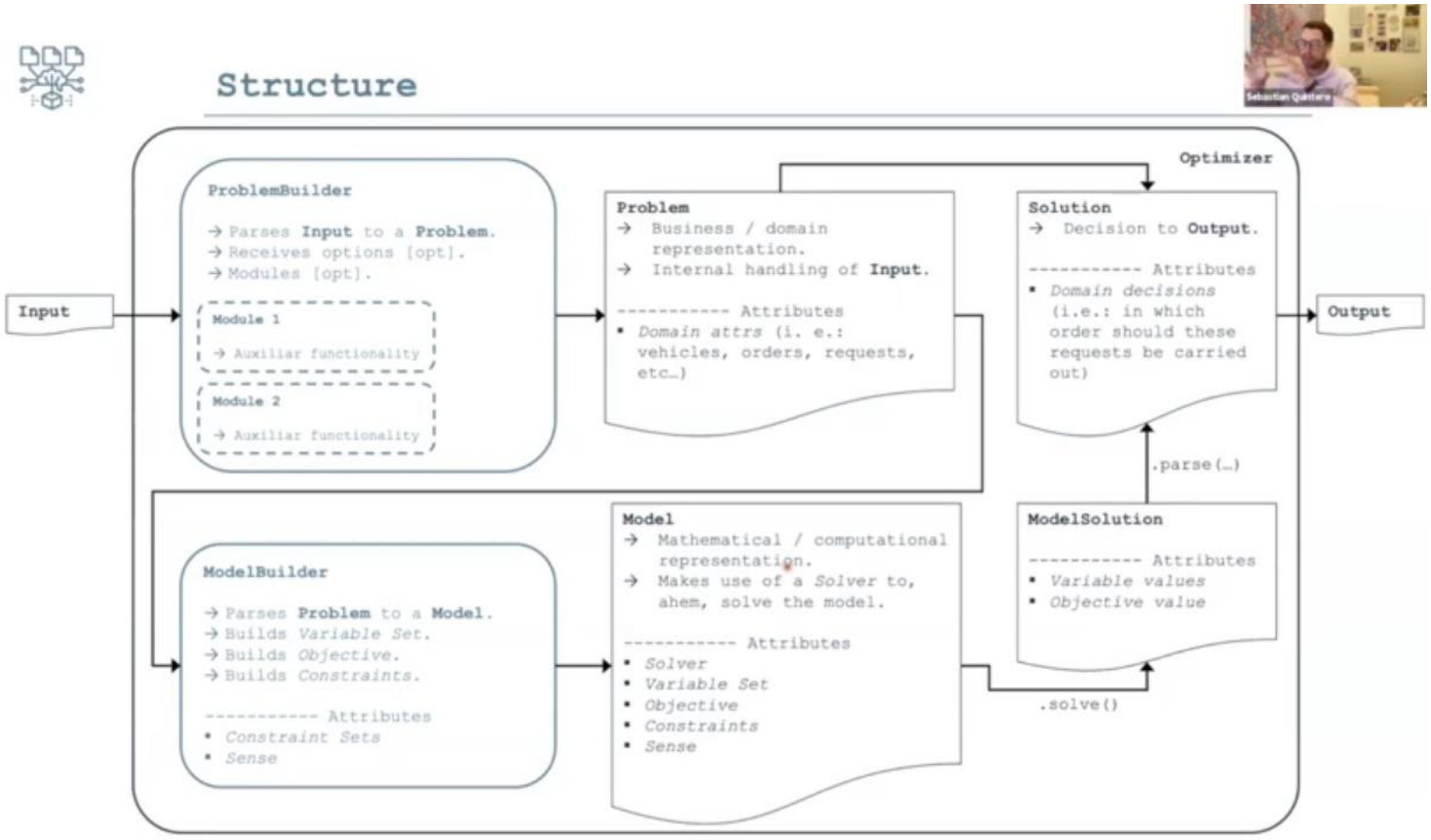
Pintemos la cancha…

En esta charla hablaremos principalmente de Modelos Analíticos de naturaleza probabilística (Supervisados, No Supervisados, Por refuerzo, Pronósticos, etc.). No entran resultados analíticos determinísticos como Reportes, Dashboards, BI, Estadística Descriptiva o Reglas de Negocio.

¿Implementación es lo mismo que Pase a Producción? → **No necesariamente. Un pase de Desarrollo a Producción es una forma de Implementar un Modelo**

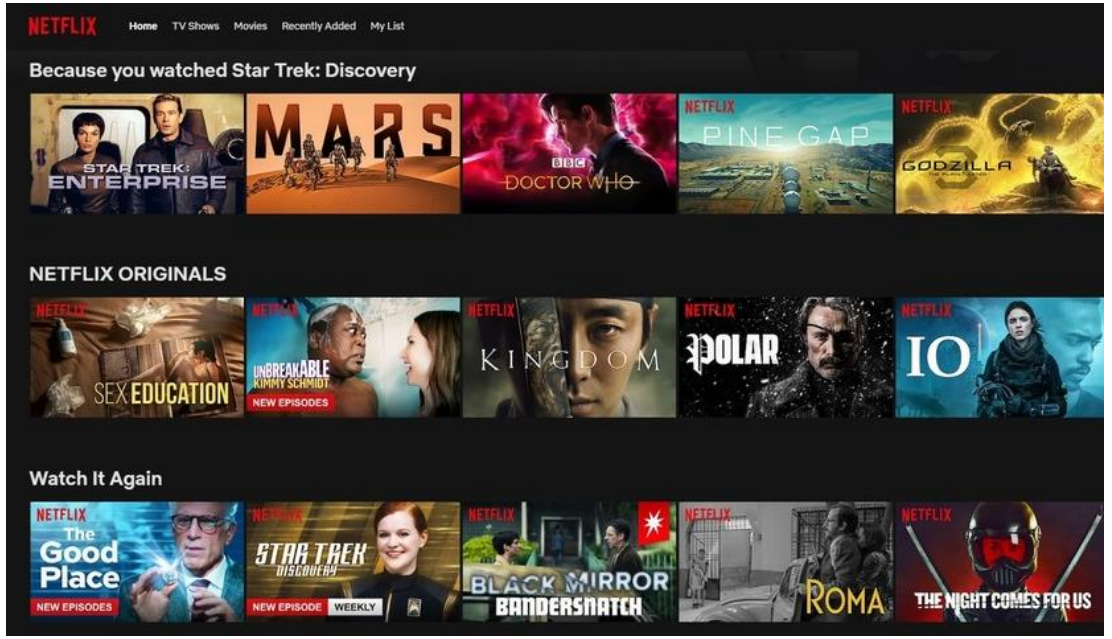
¿Cuál es la relación entre Implementación y Accionamiento? → **La manera de implementar un modelo depende de las acciones que se van a realizar.**

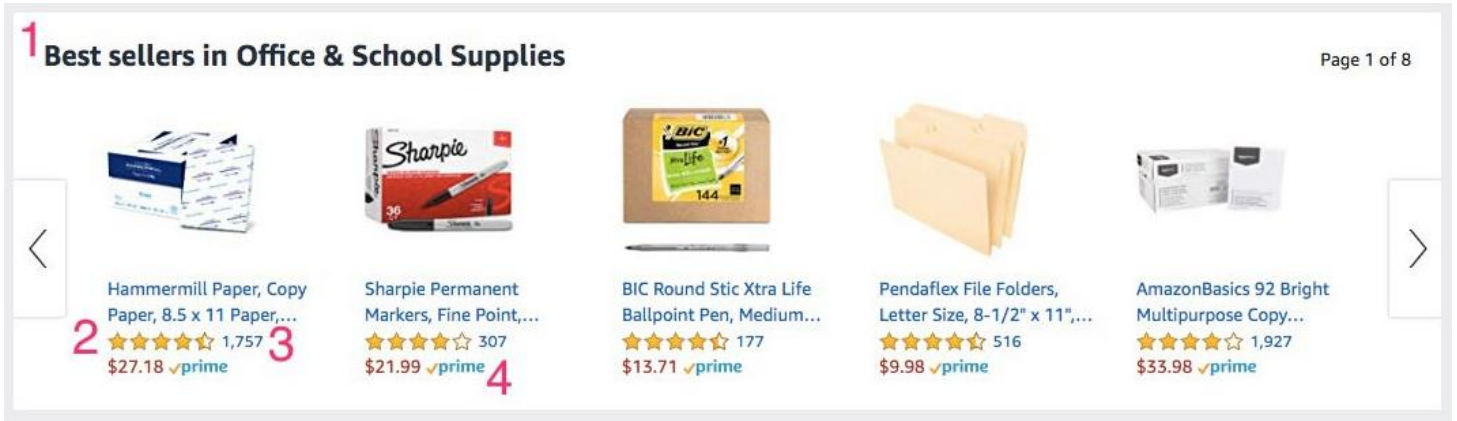
Pintemos la cancha…



teradata.

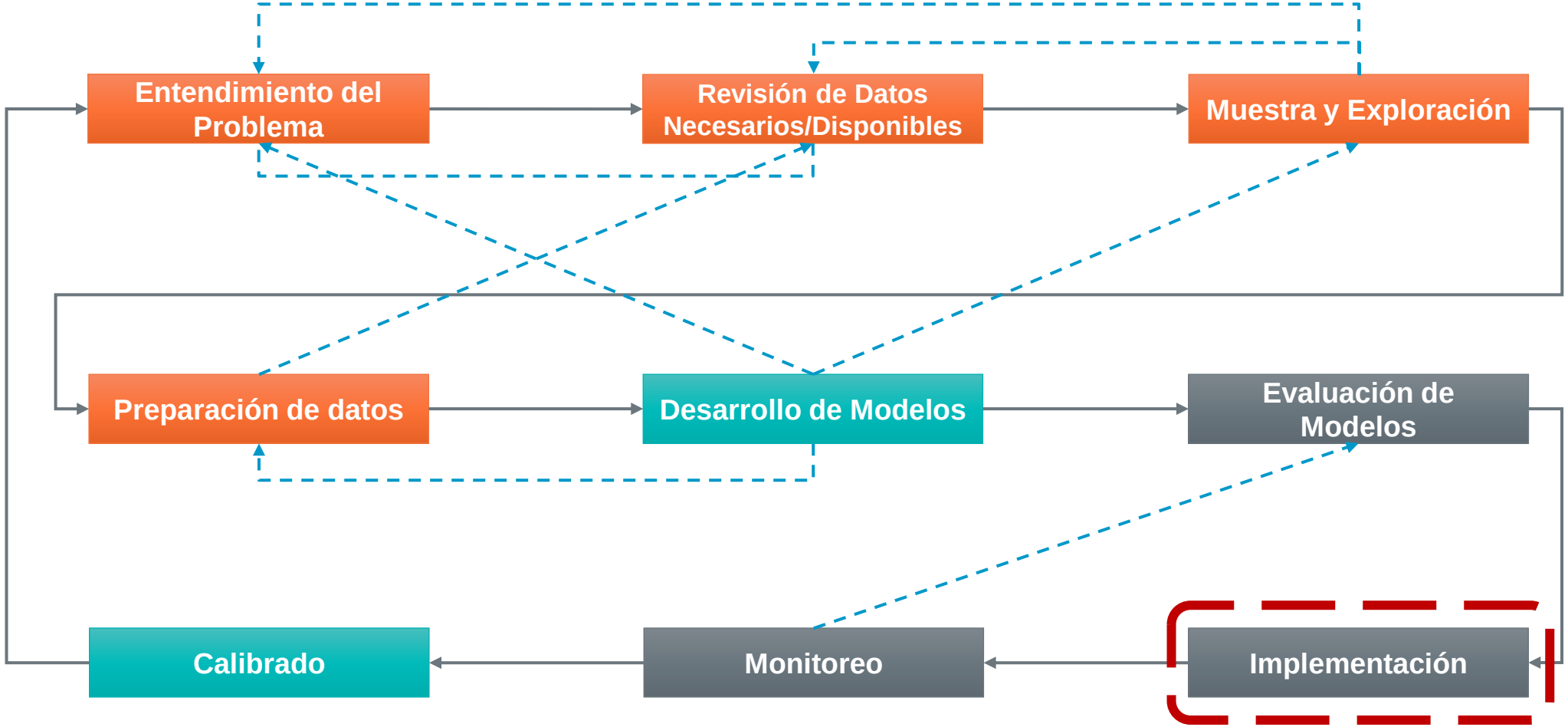
Ejemplos de Accionamiento





teradata.

Nos concentraremos en un punto, sin olvidar el resto…



Teniendo en cuenta que…

Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

Source: https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

teradata.

¿Qué modalidades de implementación existen?

Piedra, Papel y Tijeras

Ejemplos de Scorecard de Vulnerabilidad, basados en Encuestas de Hogares

Indicator	Response	Points	Score
1. In what province does the household live?	A. Balochistan	0	
	B. Northwest Frontier Province	9	
	C. Sindh	11	
	D. Punjab or Islamabad	12	
2. How many household members are 13-years-old or younger?	A. Five or more	0	
	B. Four	6	
	C. Three	11	
	D. Two	15	
	E. One	22	
	F. None	31	
3. How many children ages 5 to 13 attend school?	A. Not all	0	
	B. All, or no children ages 5 to 13	5	
4. How many household members work in elementary occupations (not senior officials, managers, professionals, technicians or associated professionals, clerks, salespeople, service or shop workers, skilled workers in agriculture or fishery, craft or trade workers, or plant/machinery operators)?	A. Two or more	0	
	B. One	5	
	C. None	12	
5. What is the highest educational level completed by the female head/spouse?	A. Less than Class 1 or no data	0	
	B. No female head/spouse	4	
	C. Class 1 or higher	6	
6. What is the main source of drinking water for the household?	A. Others	0	
	B. Hand pump, covered/closed well, motorized pump/tube well, or piped water	3	
7. What type of toilet is used by your household?	A. None or other	0	
	B. Flush connected to pit/septic tank or open drain	2	
	C. Flush connected to public sewerage	4	
8. Does the household own a refrigerator or freezer?	A. No	0	
	B. Yes	12	
9. Does the household own a television?	A. No	0	
	B. Yes	3	
10. Does the household own a motorcycle, scooter, car, or other vehicle?	A. No	0	
	B. Yes	12	
		Score:	

Pakistan (2005)

Indicador	Respuesta	Puntos	Valor
1. ¿Cuántos miembros tiene el hogar?	A. Siete o más	0	
	B. Seis	7	
	C. Cinco	12	
	D. Cuatro	17	
	E. Tres	22	
	F. Dos	27	
	G. Uno	34	
2. La semana pasada, ¿cuántos miembros del hogar de 14 años y más de edad tuvieron algún trabajo? (sin contar los quehaceres del hogar)	A. Ninguno o uno	0	
	B. Dos	2	
	C. Tres	6	
	D. Cuatro o más	9	
3. ¿Cuál es el último año o grado de estudios y nivel que aprobó la jefa/esposa del hogar?	A. Sin nivel, o educación inicial	0	
	B. Primaria incompleta	3	
	C. Primaria completa, o secundaria incompleta	4	
	D. No hay jefa/esposa del hogar	6	
	E. Secundaria completa, o superior no univ. incompleta	7	
	F. Superior no universitaria completa, o más	13	
4. ¿Cuántas habitaciones se usan exclusivamente para dormir?	A. Ninguno	0	
	B. Uno	2	
	C. Dos	4	
	D. Tres o más	8	
5. ¿El material predominante en las paredes exteriores es . . .?	A. Tapia, estera, quincha (caña con barro), adobe, piedra con barro, u otro	0	
	B. Madera, piedra o sillar con cal o cemento, o ladrillo o bloque de cemento	4	
6. ¿Cuál es el combustible que se usa con mayor frecuencia en el hogar para cocinar sus alimentos?	A. Carbón, kerosene, u otro	0	
	B. Leña	3	
	C. Gas (GLP o natural), electricidad, o no cocinan	7	
7. ¿El hogar tiene una refrigeradora/congeladora?	A. No	0	
	B. Sí	3	
8. ¿El hogar tiene una licuadora?	A. No	0	
	B. Sí	6	
9. ¿Cuántos televisores a colores tiene el hogar?	A. Ninguno	0	
	B. Uno	5	
	C. Dos o más	9	
10. ¿El hogar tiene un teléfono celular?	A. No	0	
	B. Sí	7	
SimplePovertyScorecard.com		Score:	

Perú (2010)

Piedra, Papel y Tijeras

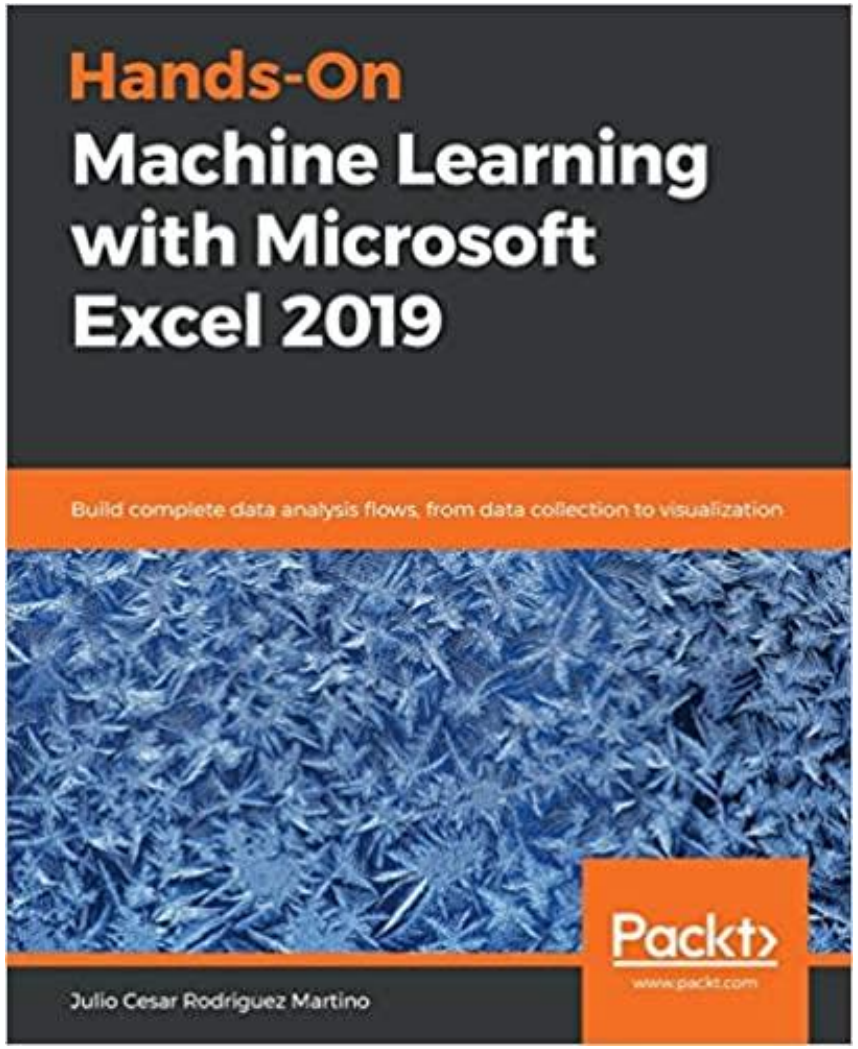
Ejemplos de Scorecard de Vulnerabilidad, basados en Encuestas de Hogares

Indicator	Value	Points	Score
1. How many members does the household have?	A. Five or more	0	
	B. Four	6	
	C. Three	11	+11
	D. Two	17	
	E. One	20	
2. Do any household members ages 5 to 18 go to private school or private pre-school?	A. No	0	+0
	B. Yes	5	
	C. No members ages 5 to 18	7	
3. How many years of schooling has the female head/spouse completed?	A. Three or less	0	
	B. Four to eleven	2	
	C. Twelve or more	8	+8
	D. No female head/spouse	8	
4. How many household members work as employees with a written contract, as civil servants for the government, or in the military?	A. None	0	
	B. One	4	+4
	C. Two or more	13	
5. In their main occupation, how many household members are managers, administrators, professionals in the arts and sciences, mid-level technicians, or clerks?	A. None	0	
	B. One or more	8	+8
6. How many rooms does the residence have?	A. One to four	0	
	B. Five	2	+2
	C. Six	5	
	D. Seven	7	
	E. Eight or more	11	

Ejemplo: Un hogar de Brasil

teradata.

¿Modelos en Excel?



Ejemplo: Scorecard en Excel

Characteristic	Attribute	Scorecard Points
AGE	<22	100
AGE	22<=AGE<26	120
AGE	26<=AGE<29	185
AGE	29<=AGE<32	200
AGE	32<=AGE<37	210
AGE	37<=AGE<42	225
AGE	>=42	250
HOME	OWN	225
HOME	RENT	110
INCOME	<10000	120
INCOME	10000<=INCOME<17000	140
INCOME	17000<=INCOME<26000	160
INCOME	26000<=INCOME<35000	200
INCOME	35000<=INCOME<42000	225
INCOME	42000<=INCOME<58000	230
INCOME	>=58000	260

Let cutoff=600

So, a new customer applies for credit.....

AGE	35	210 points
INCOME	$38K	225 points
HOME	OWN	225 points
Total		660 points
Decision:		GRANT CREDIT

Note: A scorecard is scaled with the **Odds**, **Scorecard Points** and **Points to Double the Odds** properties.

Ejemplo: Predicción en Excel

=I8/(1+I8)

	D	E	F	G	H	I	J	K
		b_0=	-3.44958					
		b_1=	0.002294					
		b_2=	0.777019					
		b_4=	-0.56003					
	admit	gre	gpa	rank	Logit	e^L	P(X)	LL
	0	380	3.61	3	-1.45292	0.233886656	=I8/(1+I8)	-0.21017
	1	660	3.67	3	-0.76399	0.465805192	0.317781104	-1.14639
	1	800	4	1	0.933639	2.543750227	0.717813069	-0.33155
	1	640	3.19	4	1.74286	0.175018404	0.1489495	1.90415

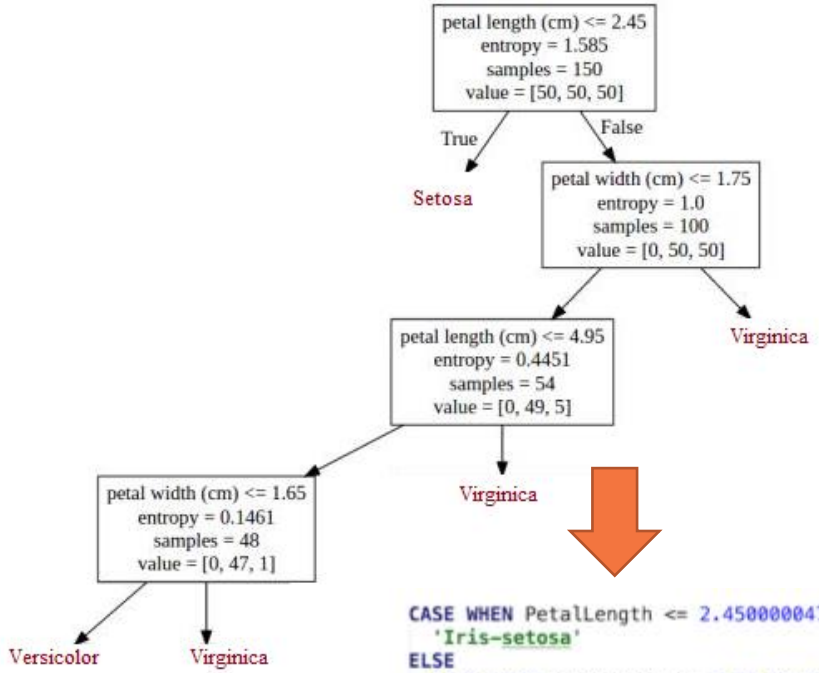
Implementando en SQL (Regresiones y Árboles)

```
Call:
lm(formula = d$failures ~ d$absences + d$internetaccess)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0385 -0.2994 -0.2085 -0.1857  2.8143

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.32151    0.04552   7.063 2.98e-12 ***
d$absences       0.01137    0.00326   3.488 0.000508 ***
d$internetaccess -0.13579    0.04987  -2.723 0.006579 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
INSERT INTO RESULTADOS_MODELOS AS
(SELECT CURRENT_DATE,
    ID,
    0.32151 +
    0.01137 * ABSENCES -
    0.01137 * INTERNETACCESS
    AS FAILURES_PRED
FROM BD.TABLA_MODELO);
```




```
CASE WHEN PetalLength <= 2.45000004768 THEN
  'Iris-setosa'
ELSE
  CASE WHEN PetalWidth <= 1.75 THEN
    CASE WHEN PetalLength <= 4.94999980927 THEN
      CASE WHEN PetalWidth <= 1.65000009537 THEN
        'Iris-versicolor'
      ELSE
        'Iris-virginica'
      END
    ELSE
      'Iris-virginica'
    END
  ELSE
    CASE WHEN PetalLength <= 4.85000038147 THEN
      'Iris-virginica'
    ELSE
      'Iris-virginica'
    END
  END
END
```


teradata.

Implementando en SQL y Excel (Un caso en PayPal)

Un proyecto heterodoxo de ciencia de datos: detección de fugas en Paypal

152 visualizaciones • 25 abr 2021

11 0 COMPARTIR GUARDAR ...

Gil Bellosta
169 suscriptores

SUSCRIBIRME

Matt Lerner
@matthlerner

After 17 years, we finally "cracked" a $100M churn problem at PayPal. Zero fancy tech. Just a spreadsheet, some simple SQL, and a physicist named Ben. 👇

Traducir Tweet

2:30 a. m. · 30 mar. 2021 · Twitter Web App

2.635 Retweets **554** Tweets citados **17 mil** Me gusta

Matt Lerner @matthlerner · 30 mar.
En respuesta a @matthlerner
@benramsden studied physics @ Cambridge, and I seriously envy his ability to work through a thorny problem from first principles. So I gave it to Ben, and Ben gave it to his intern. 😏

8 23 463

Matt Lerner @matthlerner · 30 mar.
But first... Einstein said "If I had an hour to solve a problem, I'd spend 55 minutes thinking about the problem and 5 minutes on solutions."
PayPal prob loses 1,000,000 merchants per year for various reasons, so step 1 is to narrow down the problem!

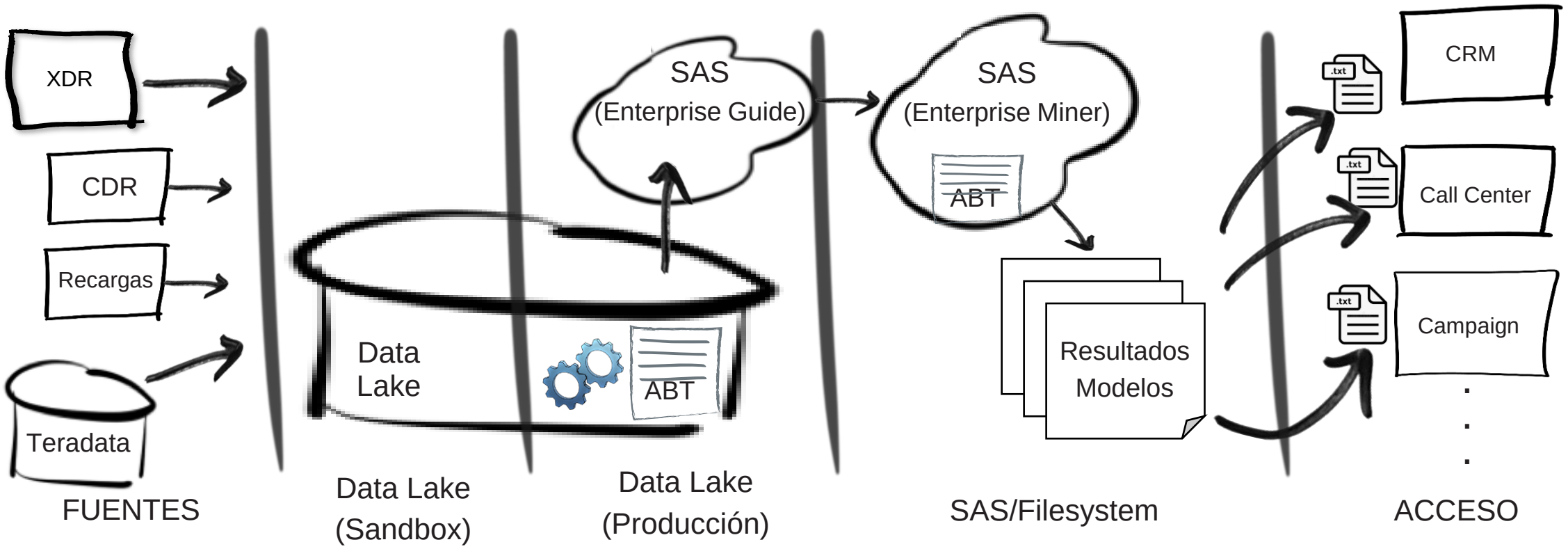
4 47 626

Matt Lerner @matthlerner · 30 mar.
We narrowed it down by applying exclusions, crossing off things that look like churn, but aren't. What can we rule out first? Account closures.

2 3 277

Aparecen las suites de Analítica Avanzada

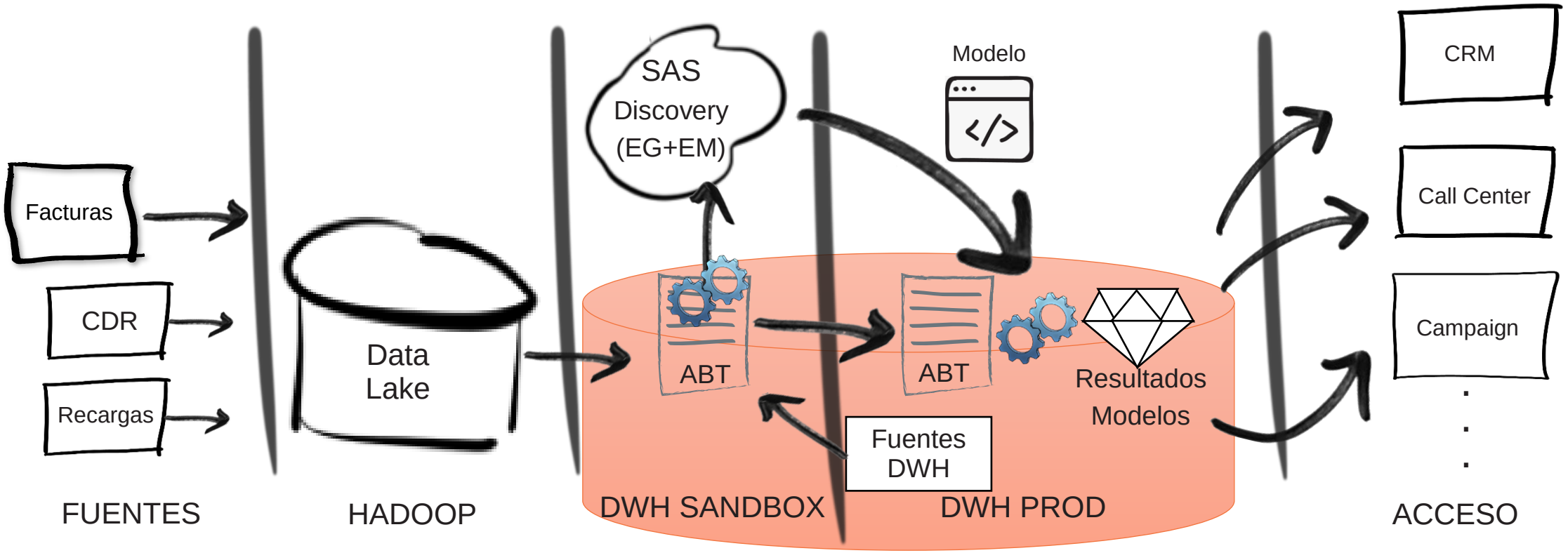
Arquitectura “Accidental”



- Almacenamiento, transformación de datos y armado de la analytical base tables (ABTs) en Hadoop con Hue y SAS Enterprise Guide.
- SAS sobre Hadoop para la ejecución de Modelos en Producción.

Aparecen las suites de Analítica Avanzada

Arquitectura Optimizada



- Almacenamiento de fuentes en el Data Lake.
- Transformación de datos y armado de la analytical base tables (ABTs) en el DWH.
- Desarrollo de Modelos en SAS y Scoring In-Database con C, Java o PMML.

PMML y los formatos estándar (ONNX, NNVM, NNEF, PFA)

Standard Format

PMML
NNVM
ONNX
NNEF

Single stack

Execution

Optimization

Tooling
(Viz, analysis, ...)

teradata.

Empaquetando modelos (RData, Pickle, HDF5, joblib, etc.)

How To Save Trained Machine Learning Models?

Save & Reload Your Trained Machine Learning Models In Python

Farhad Malik Follow

May 23, 2019 · 2 min read ★

How to Save and Load Your Keras Deep Learning Model

by **Jason Brownlee** on May 13, 2019 in **Deep Learning**

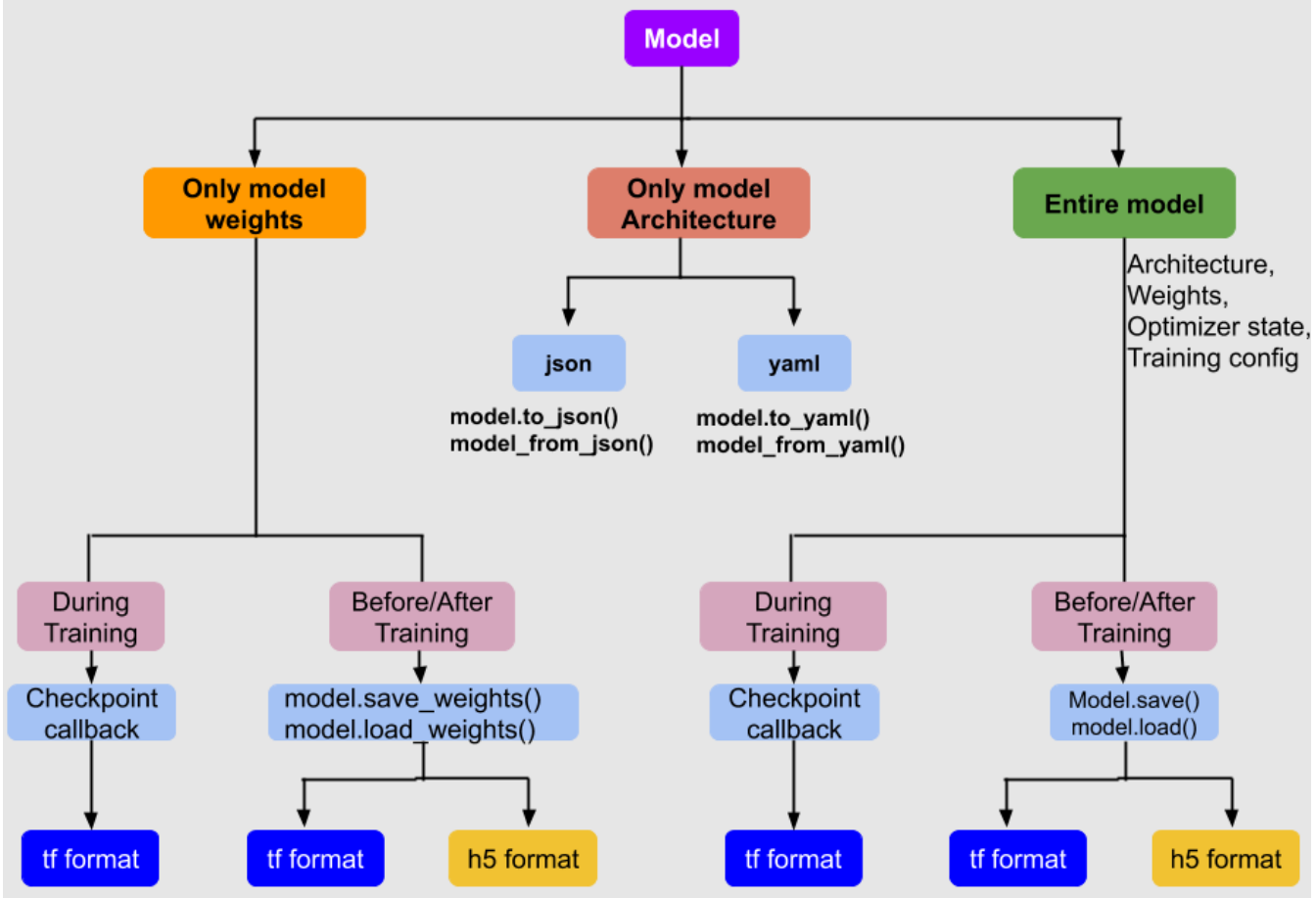
Tweet Share Share

Last Updated on August 27, 2020

Keras is a simple and powerful Python library for deep learning.

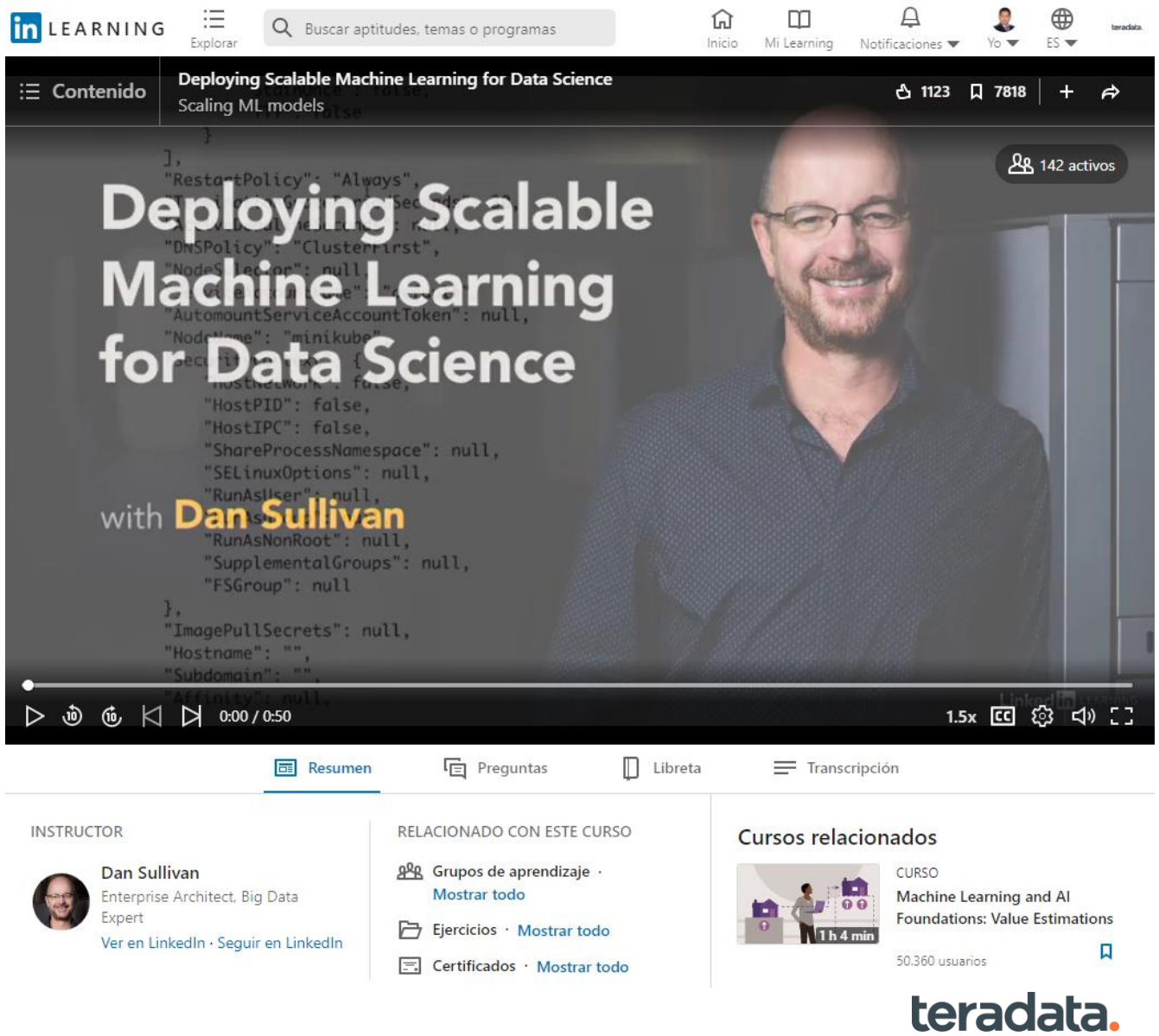
How to save (and load) datasets in R: An overview

Posted on May 27, 2019 by **Rcrastinate** in **R bloggers** | 0 Comments



teradata.

API/REST



teradata.

Orquestando APIs

Carlos de la Torre: Del POC a producción - Infraestructura para machine-learning | PyData Córdoba

269 visualizaciones • 16 dic 2019 4 1 COMPARTIR GUARDAR

PyData
116.000 suscriptores
SUSCRIBIRME

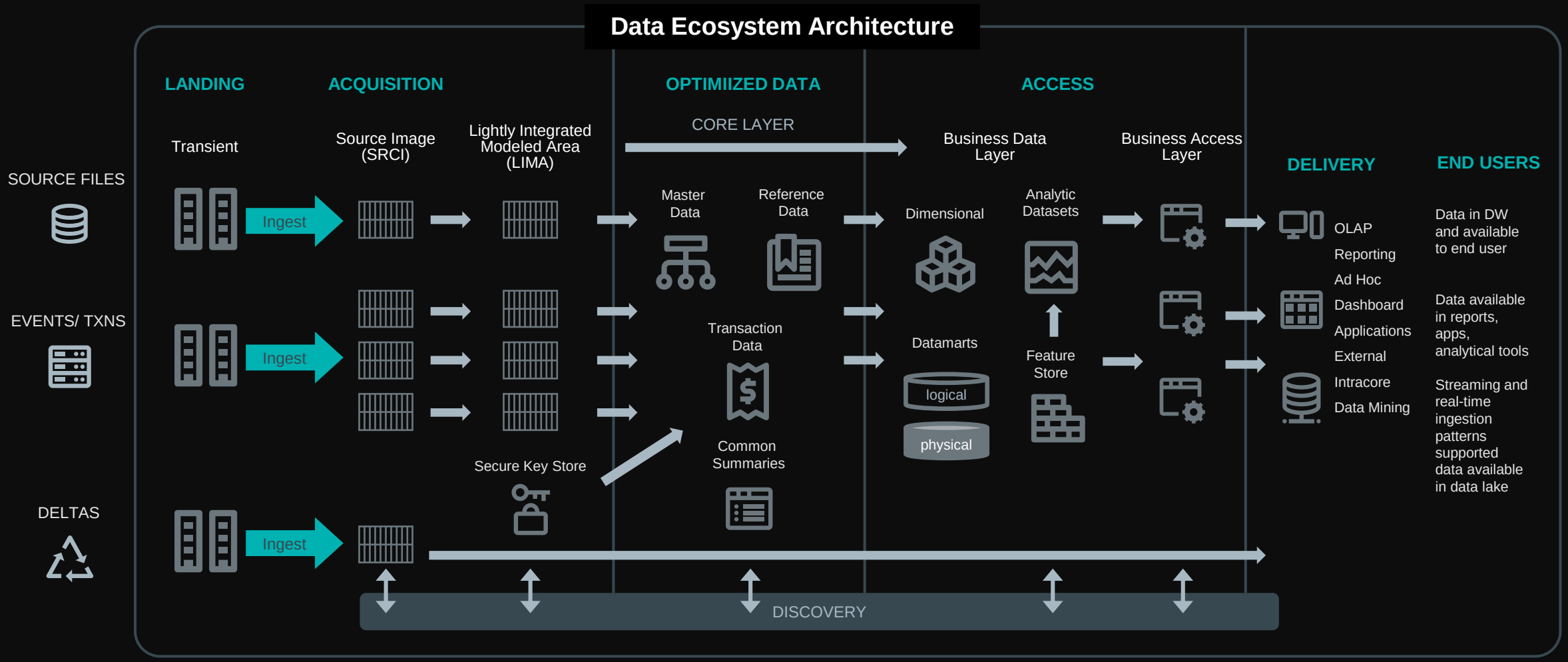
Cada vez que un usuario sube una foto a Mercado Libre, se ejecuta una infinidad de modelos para:

- Detectar texto
- Clasificar el objeto
- Detectar Marca, modelo
- Peso estimado
- Volumen de la caja
- Etc.

Cuando un usuario navega por la web, determinadas acciones también dispara la ejecución de más modelos, como la estimación de costos de envío, tiempo de llegada, etc.

teradata.

Arquitectura de Referencia para DataOps



Fast, agile access to new data sources via acquisition layer and integration via LIMA. DataOps provides governance, management, and update of the Enterprise Feature Store.

teradata.

Llevar la Analítica a los Datos

Move data samples for model training

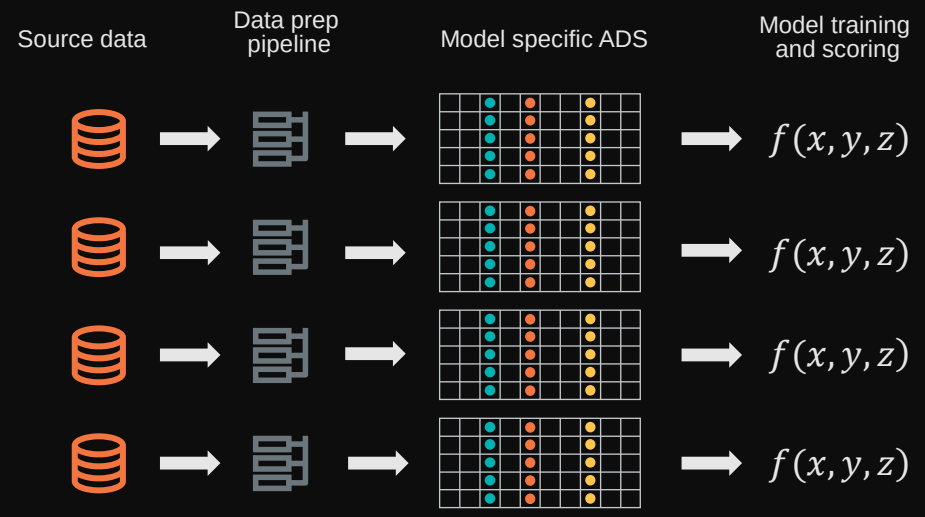
Import the model into Vantage

Data movement is minimized by importing the model back into Vantage for production scoring

Build model with any tool you prefer

Enterprise Feature Store

The one pipeline per model approach



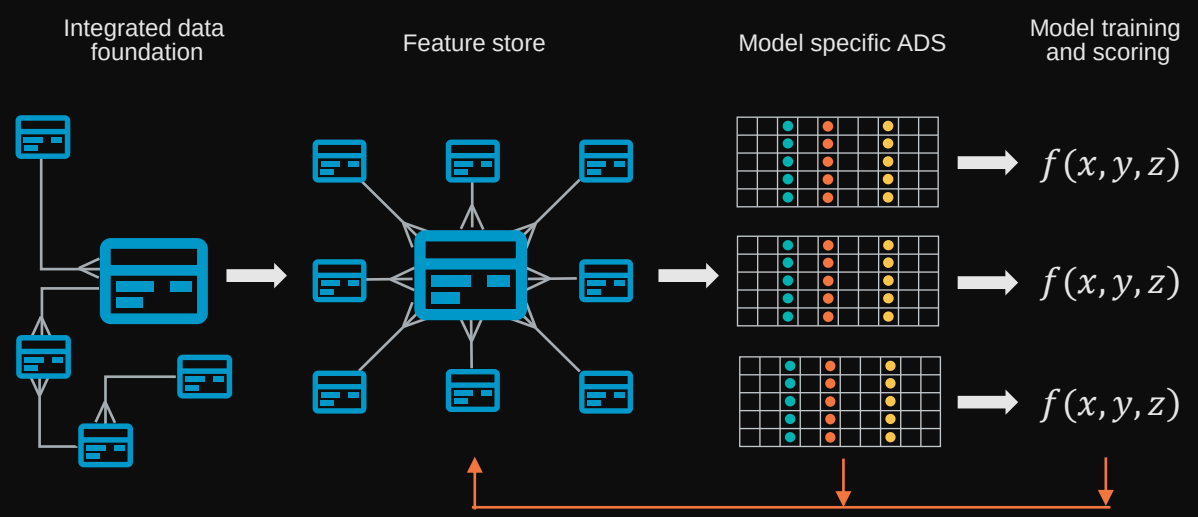
One pipeline per model → Long data prep cycles and poor time to market

Redundant infrastructure, processing, and effort → High TCO

Limited re-use of pipeline or features → Poor productivity and data silos

DSs functioning mostly as data janitors → Inefficient allocation of resources

The feature store approach



"Off the peg" features dramatically improve analytic cycle times and time-to-market

Extensive re-use reduces TCO and improves analytic data quality and predictive model accuracy

ADS layer enables model-specific customization, whilst eliminating analytic data silos

Separation of duties and improved productivity

Champion-Challenger

teradata.

Conclusiones

¿Cómo elegir la mejor modalidad?

Modalidad	Dependencia Tecnológica	¿Qué tan fácil de utilizar?	Gestión de Metadatos y Monitoreo	Escalabilidad a muchos Modelos y Datos
Papel	Ninguna	Muy Fácil	No existe	Muy Difícil
Excel	Cualquier PC	Fácil	Pobre	Difícil
SQL	Usuarios BD	Fácil	Muy Buena	Muy Fácil
Suites	Acceso a la Suite	Fácil	Buena*	Depende
HiperLenguaje	Arq. Simple	Intermedio	Buena*	Fácil
Paquetes	Arq. Intermedia	Complejo	Buena*	Fácil*
API	Arq. Compleja	Complejo	Muy Buena*	Fácil*
Orquestación	Arq. Muy Compleja	Muy Complejo	Muy Buena	Muy Fácil*

* Considerando la implementación de un EFS + Repositorio de Modelos

Para recordar:

1

La implementación de modelos es un paso importante, pero para definir el método, considere la **naturaleza del problema** y las **posibles acciones**.

2

Es importante tener un ecosistema de implementación propio. **Sin metadatos y organización no se pueden mejorar los modelos.**

3

Un modelo que no se acciona no agrega valor. Es importante medir el incremento en revenue, ahorro o eficiencia adicional que produce cada modelo.