



UNIVERSIDAD  
PANAMERICANA

# Introducción a la Ciencia de Datos

Dr. Leon Felipe Palafox Novack  
[lpalafox@up.edu.mx](mailto:lpalafox@up.edu.mx)

# 0

## Noticias del día

Qué ha pasado en el mundo de Data  
Science?

## IBM to Buy Red Hat, the Top Linux Distributor, for \$34 Billion



IBM's purchase of Red Hat, the largest distributor of the open-source operating system Linux, is the latest competitive step among large business software companies to gain an edge in the cloud computing market. Pau Barrena/Agence France-Presse — Getty Images

## Cloudera and Hortonworks Announce Merger to Create World's Leading Next Generation Data Platform and Deliver Industry's First Enterprise Data Cloud

Published: October 3, 2018

*Establishes a superior unified platform and clear industry standard from the Edge to AI*

*Strategic combination accelerates market development, fuels innovation and produces substantial benefit for customers, partners and community*

PALO ALTO, Calif. & SANTA CLARA, Calif.--(BUSINESS WIRE)--Oct. 3, 2018--  
[Cloudera, Inc.](#) (NYSE:CLDR) and [Hortonworks, Inc.](#) (Nasdaq:HDP) jointly announced today that they have entered into a definitive agreement under which the companies will combine in an all-stock merger of equals. The transaction, which has been unanimously approved by the Boards of Directors of both companies, will create the world's leading next generation data platform provider, spanning multi-cloud, on-premises and the Edge. The combination establishes the industry standard for hybrid cloud data management, accelerating customer adoption, community development and partner engagement.

1

# Anuncios parroquiales

Proyecto Final



## Proyecto Final

El objetivo es que trabajen en el proyecto a lo largo del curso.

Conforme vayamos aprendiendo las herramientas, se recomienda las practiquen con sus propios datos.

Vayan formando sus equipos oportunamente.

- Los tutoriales de Python van a estar disponibles toda la clase
- El machote para el reporte final ya esta en la pagina web.

# 2

## Gobierno y Administración del Dato

Tenemos que entender nuestra  
materia prima

- Ciclo de vida del Dato

- Gobierno de Datos



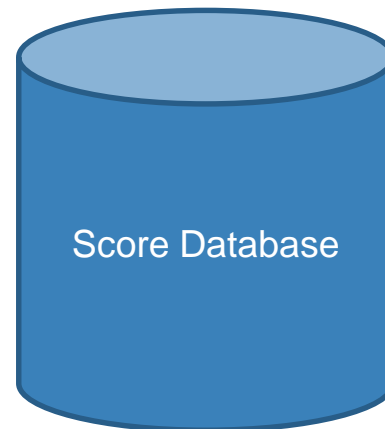
## Ciclo de vida del dato

Cuantos datos se generan?

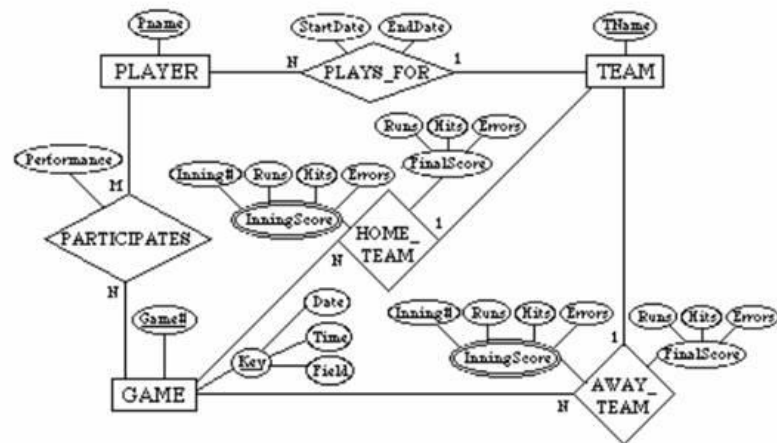
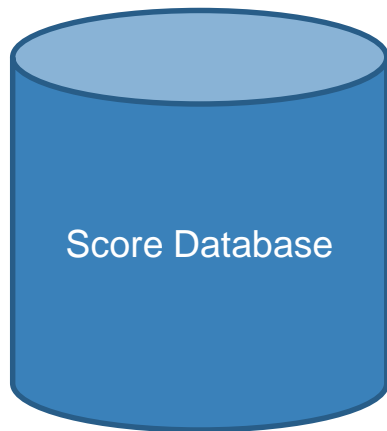


Grinnell														
#	STARTERS	MIN	FGM-A	3PM-A	FTM-A	OREB	DREB	REB	AST	STL	BLK	TO	PF	PTS
03	<a href="#">Jack Taylor</a>	36	52-108	27-71	7-10	3	0	3	0	3	0	6	0	138
05	<a href="#">Joe Rogers</a>	13	0-0	0-0	0-0	0	1	1	1	2	0	1	0	0
10	<a href="#">Patrick Maher</a>	15	2-5	0-2	0-0	4	2	6	5	3	0	3	0	4
42	<a href="#">Jesse Ney</a>	14	0-2	0-1	0-0	2	0	2	1	0	0	0	1	0
44	<a href="#">Tague Zachary</a>	14	1-2	1-2	0-0	2	3	5	0	3	1	0	1	3
#	RESERVES	MIN	FGM-A	3PM-A	FTM-A	OREB	DREB	REB	AST	STL	BLK	TO	PF	PTS
02	<a href="#">Luke Yeager</a>	12	0-0	0-0	0-0	0	0	0	4	3	0	2	0	0
04	<a href="#">Garrett Nitz</a>	9	0-0	0-0	0-0	3	2	5	1	2	0	0	0	0
11	<a href="#">Griffin Lentsch</a>	14	2-3	1-2	2-2	0	1	1	0	0	0	2	0	7
12	<a href="#">Marques Valdez</a>	9	0-2	0-0	2-2	8	1	9	0	1	0	1	3	2
13	<a href="#">Leo Abbe-Schneider</a>	2	1-2	0-0	0-0	0	0	0	0	0	0	0	1	2
15	<a href="#">Dylan Bartuch</a>	4	2-2	0-0	0-0	0	0	0	2	0	0	0	1	4
21	<a href="#">Brent Lemoline</a>	2	1-1	0-0	0-0	1	0	1	0	1	0	0	0	2
22	<a href="#">Evan Johnson</a>	3	0-1	0-1	0-0	1	0	1	0	0	0	0	0	0
23	<a href="#">Aaron Levin</a>	11	5-6	1-1	2-2	4	2	6	2	4	0	0	1	13
24	<a href="#">Dominique Bellamy</a>	12	1-1	0-0	0-0	3	1	4	2	3	1	1	2	2
25	<a href="#">Anthony Lamacchia</a>	2	0-0	0-0	0-0	0	0	0	0	0	0	0	0	0
33	<a href="#">Brian McManamy</a>	11	1-1	0-0	0-0	2	7	9	3	0	0	0	2	2
35	<a href="#">Cody Olson</a>	3	0-0	0-0	0-0	0	0	0	0	0	0	1	1	0
45	<a href="#">Jack Adams</a>	9	0-0	0-0	0-0	2	0	2	1	3	2	0	0	0
54	<a href="#">Ryan Davis</a>	5	0-0	0-0	0-0	0	1	1	0	1	0	0	1	0
TM TEAM						2	1	3						
TOTALS			68-136 50.0%	30-80 37.5%	13-16 81.2%	37	22	59	22	29	4	17	14	179

Grinnell													
#	STARTERS	MIN	FGM-A	3PM-A	FTM-A	OREB	DREB	REB	AST	STL	BLK	TO	PTS
03	Jack Taylor	36	52-108	27-71	7-10	3	0	3	0	3	0	6	138
05	Joe Rogers	13	0-0	0-0	0-0	0	1	1	1	2	0	1	0
10	Patrick Maher	15	2-5	0-2	0-0	4	2	6	5	3	0	3	4
42	Jesse New	14	0-2	0-1	0-0	2	0	2	1	0	0	1	0
44	Taque Zachary	14	1-2	1-2	0-0	2	3	5	0	3	1	0	1
#	RESERVES	MIN	FGM-A	3PM-A	FTM-A	OREB	DREB	REB	AST	STL	BLK	TO	PTS
02	Luke Yeager	12	0-0	0-0	0-0	0	0	0	4	3	0	2	0
04	Garrett Nitz	9	0-0	0-0	0-0	3	2	5	1	2	0	0	0
11	Griffin Lentsch	14	2-3	1-2	2-2	0	1	1	0	0	0	2	7
12	Marques Valdez	9	0-2	0-0	2-2	8	1	9	0	1	0	1	3
13	Leo Abbe-Schneider	2	1-2	0-0	0-0	0	0	0	0	0	0	1	2
15	Dylan Bartuch	4	2-2	0-0	0-0	0	0	0	2	0	0	0	1
21	Brent Lemoine	2	1-1	0-0	0-0	1	0	1	0	1	0	0	2
22	Ryan Johnson	3	0-1	0-1	0-0	1	0	1	0	0	0	0	0
23	Aaron Levin	11	5-6	1-1	2-2	4	2	6	2	4	0	0	1
24	Dominique Bellamy	12	1-1	0-0	0-0	3	1	4	2	3	1	1	2
25	Anthony Lamacchia	2	0-0	0-0	0-0	0	0	0	0	0	0	0	0
33	Brian McManamy	11	1-1	0-0	0-0	2	7	9	3	0	0	0	2
35	Cody Olson	3	0-0	0-0	0-0	0	0	0	0	0	0	1	1
45	Jack Adams	9	0-0	0-0	0-0	2	0	2	1	3	2	0	0
54	Ryan Davis	5	0-0	0-0	0-0	0	1	1	0	1	0	0	1
TM TEAM						2	1	3					0
TOTALS		68-136	30-80	13-16	37	22	59	22	29	4	17	14	179
		50.0%	37.5%	81.2%									



Que otra información va a la base de datos?



# Ciclo de vida del dato



- En que momento se destruye el dato?
- Cuanto tiempo dura el dato?
- Como se va a modificar el dato?

## En que momento se destruye el dato?

- El dato se destruye cuando se decide que no es útil para el negocio.
- El dato se destruye cuando se decide que su integridad ha sido comprometida.
- El dato se destruye cuando su ciclo de vida se ha cumplido.

## En que momento se destruye el dato?

- En el ejemplo de basquetbol, se destruye un dato como un segundo en el que no sucedió nada relevante.
- Se destruyen datos como si se saludo con un amigo.

## ¿Cuánto tiempo dura el dato?



- El dato dura la cantidad de tiempo necesaria para resolver un problema del negocio.



## Cuánto tiempo dura el dato?

- En el ejemplo de basquetbol, la estadística dura hasta que no la necesitemos.
  - ▷ Jugador retirado?
  - ▷ Estadio destruido?

## Como se va a modificar el dato?

- Se debe tener total control de como se modifica.
- En que base de datos se almacena
- Que tipo de transformación se le hace.



TESLA



UNIVERSIDAD  
PANAMERICANA

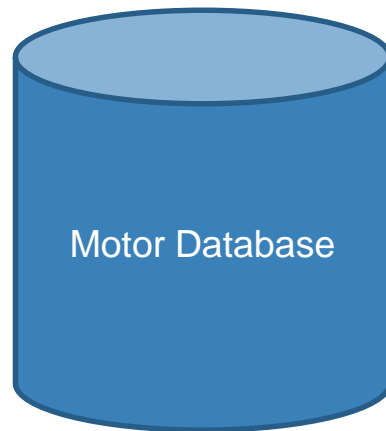


Cuantos datos se  
generan?

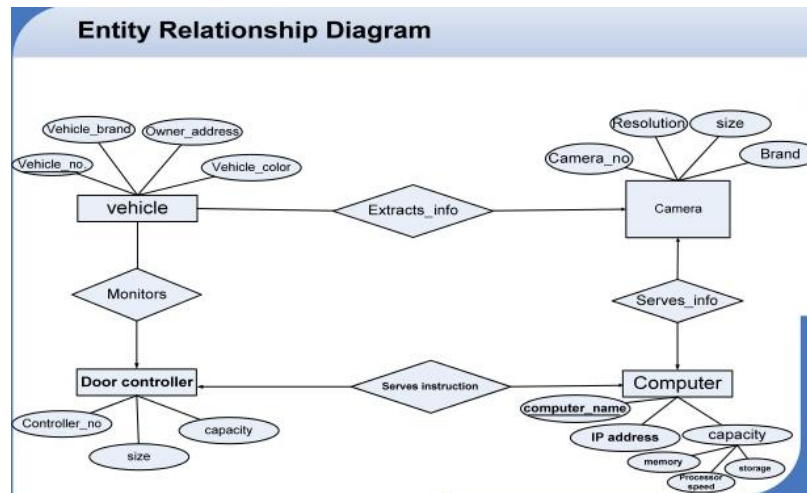
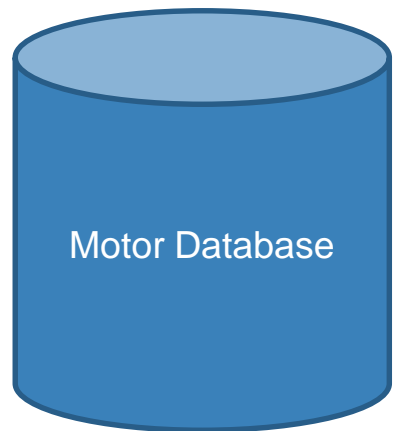
ANDY GREENBERG SECURITY 07.21.15 06:00 AM

# HACKERS REMOTELY KILL A JEEP ON THE HIGHWAY—WITH ME IN IT





Que otra información se almacena en la base de datos



# Ciclo de vida del dato



- En que momento se destruye el dato?
- Cuanto tiempo dura el dato?
- Como se va a modificar el dato?



# Gobierno de Datos (Data Governance)



■ Es una serie de principios para asegurar que:

- ▷ Los datos son de alta calidad
- ▷ Los datos están seguros
  - ▷ Acceso
  - ▷ Escritura
  - ▷ Borrado

# Gobierno de Datos



- ▷ Los datos están correctamente administrados
- ▷ Los datos están correctamente catalogados
  - ▷ Bases de datos
  - ▷ Respaldos
  - ▷ Catalogos
  - ▷ Etc



UNIVERSIDAD  
PANAMERICANA

## Disciplinas para un efectivo Gobierno de Datos





# 3

## Que es Big Data y como se relaciona con la Nube

Con que se come?

“ If we have data, let's look at data. If all we have are opinions, let's go with mine

Jim Barksdale  
Former Netscape CEO



“ Big data is like teenage sex:  
everyone talks about it, nobody  
really knows how to do it,  
everyone thinks everyone else  
is doing it, so everyone claims  
they are doing it...

■ Se refiere al tratamiento de datos masivos:

- ▷ Almacenamiento
- ▷ Procesamiento
- ▷ Análisis
- ▷ Visualización

## ■ Almacenamiento:

- ▶ Data Lakes, Bases de Datos, Data Warehouses

## ■ Procesamiento:

- ▷ Buscadores:
  - ▷ Solr y Elasticsearch
- ▷ Hadoop

# Hadoop

- Es almacenamiento/procesamiento en paralelo.



## ■ Análisis

- ▶ Spark (Hadoop)
  - ▶ Rspark, PySpark
- ▶ SASS
- ▶ SAP Hana

## ■ Visualización

- ▶ PowerBI (Microsoft)
- ▶ Tableau

# Características de Big Data



■ Que características creen que tiene Big Data?



# Características de Big Data



- No cabe en una sola maquina
- Se necesitan muchas maquinas en paralelo para procesar/almacenar los datos.
- Se necesita mucho personal técnico para administrarlo



- Son maquinas que están en algún lugar del universo:
  - ▷ Otro estado
  - ▷ Otro país
  - ▷ Otro continente

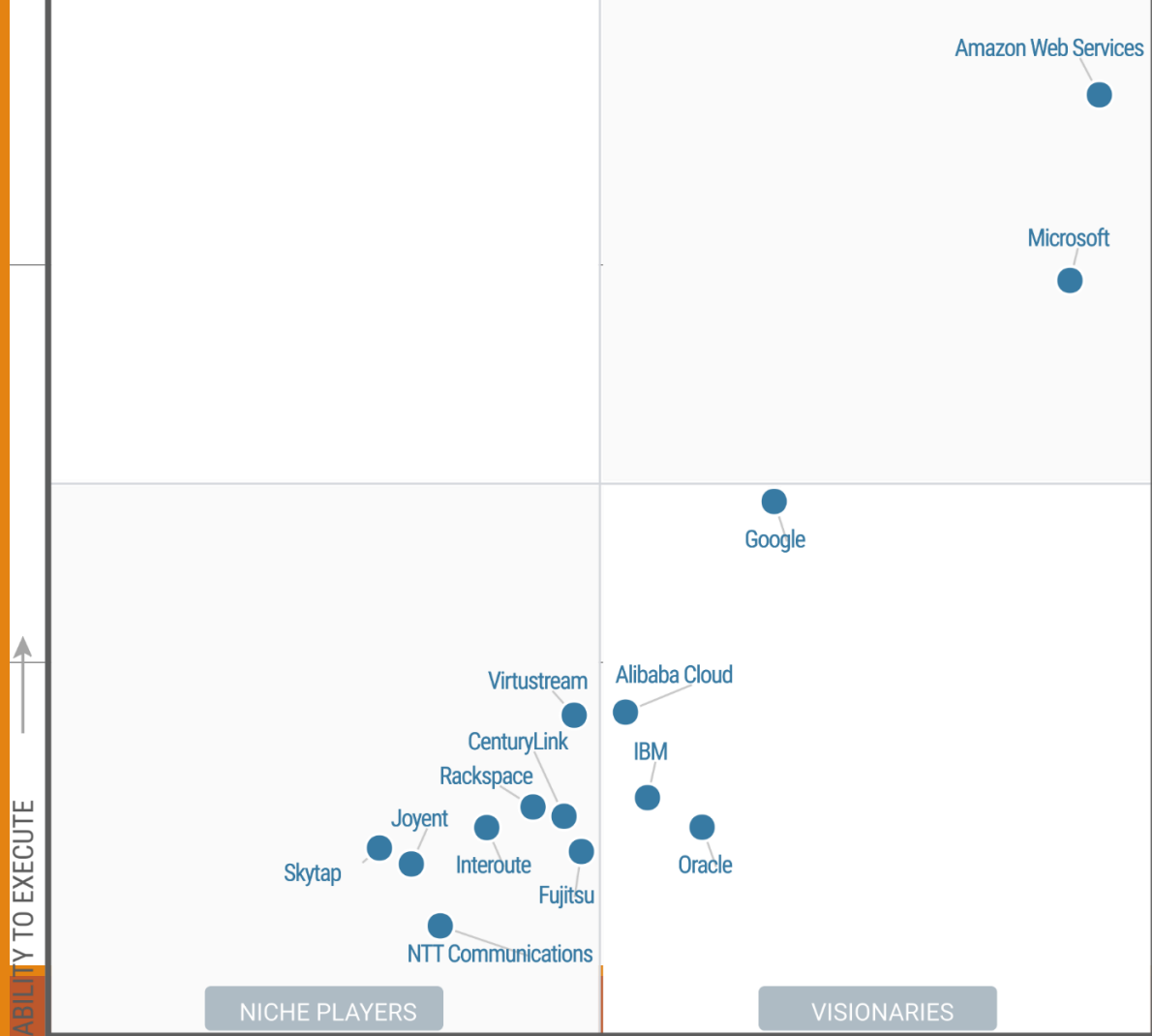
- Hay tres servicios de cloud (bueno, hay más):

**amazon**

**Google**

 **Microsoft**

# Gartner Diagram



# Ventajas de la nube



- Poder de procesamiento infinito
- Capacidad de almacenamiento infinita.
- De los mejores algoritmos y servicios al alcance de las manos.

