UNIVERSIDAD PANAMERICANA

Introducción a la Ciencia de Datos

Dr. Leon Felipe Palafox Novack
lpalafox@up.edu.mx

0

Noticias del día

Qué ha pasado en el mundo de Data Science?

RIIAA @MEX

Reunión Internacional de Inteligencia Artificial y sus Aplicaciones

Agosto 24-26, 2018
CDMX, México - UNAM

Con patrocinio de:

UNIVERSIDAD PANAMERICANA

Ingenierías UP
POSGRADOS

UNIVERSIDAD PANAMERICANA

MÉXICO

PRESIDENCIA DE LA REPÚBLICA

CONABIO

COMISIÓN NACIONAL PARA EL CONOCIMIENTO Y USO DE LA BIODIVERSIDAD

IngenieríasUP POSGRADOS

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.

Recursos

- Los tutoriales de Python van a estar disponibles toda la clase
- El machote para el reporte final ya esta en la pagina web.
- La fecha de entrega es la última clase del curso, 18 de Septiembre

2

Gobierno y Administración del Dato

Tenemos que entender nuestra materia prima

Temas

- Ciclo de vida del Dato
- Gobierno de Datos

UNIVERSIDAD PANAMERICANA

Ingenierías UP POSGRADOS

Ciclo de vida del dato

UNIVERSIDAD PANAMERICANA

Cuantos datos se generan?

IngenieríasUP
POSGRADOS

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

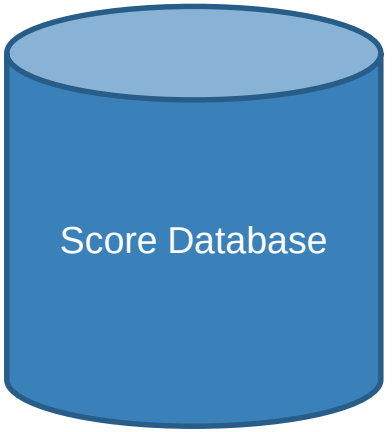
UNIVERSIDAD PANAMERICANA

IngenieríasUP POSGRADOS

UNIVERSIDAD PANAMERICANA

Grinnell

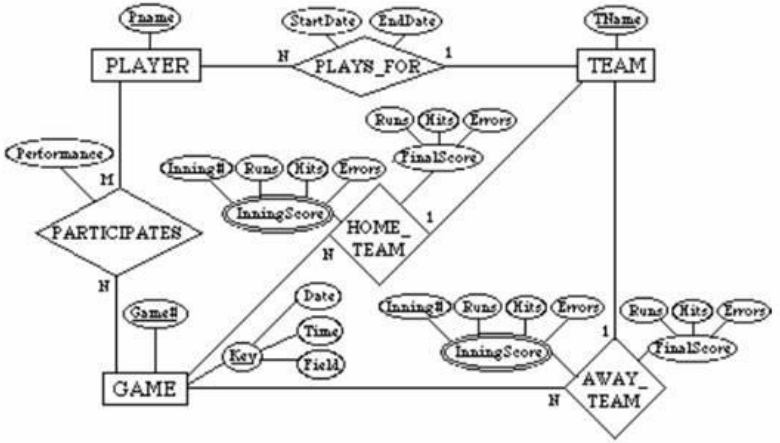
#	STARTERS	MIN	FGM-A	3PM-A	FTM-A	OREB	DREB	REB	AST	STL	BLK	TO	PF	PTS
03	Jack Taylor	36	52-108	27-71	7-10	3	0	3	0	3	0	6	0	138
05	Joe Rogers	13	0-0	0-0	0-0	0	1	1	1	2	0	1	0	0
10	Patrick Maher	15	2-5	0-2	0-0	4	2	6	5	3	0	3	0	4
42	Jesse Ney	14	0-2	0-1	0-0	2	0	2	1	0	0	0	1	0
44	Taque Zachary	14	1-2	1-2	0-0	2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



Que otra información va a la base de datos?

IngenieríasUP POSGRADOS

Score Database



Ciclo de vida del dato

UNIVERSIDAD PANAMERICANA

- 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?

UNIVERSIDAD PANAMERICANA

- 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?

UNIVERSIDAD PANAMERICANA

- 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 módifica.
- En que base de datos se almacena
- Que tipo de transformación se le hace.

UNIVERSIDAD PANAMERICANA

TESLA

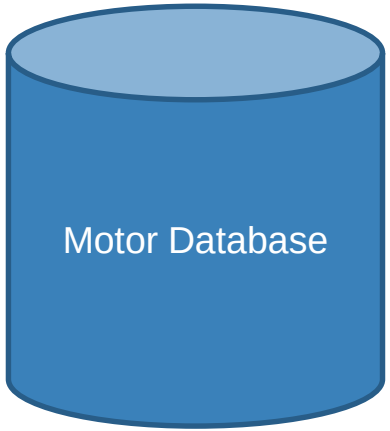
Ingenierías UP
POSGRADOS

Cuantos datos se generan?

UNIVERSIDAD PANAMERICANA

ANDY GREENBERG SECURITY 07.21.15 06:00 AM

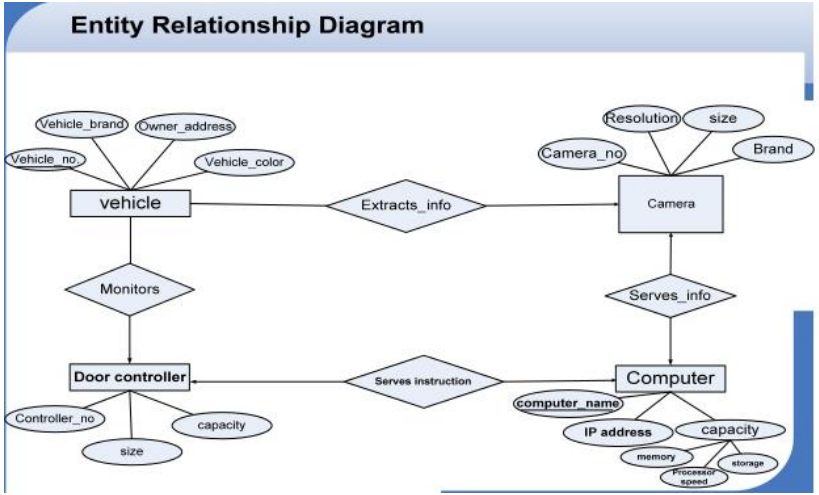
HACKERS REMOTELY KILL A JEEP ON THE HIGHWAY—WITH ME IN IT



Que otra información se almacena en la base de datos

UNIVERSIDAD PANAMERICANA

Motor Database



IngenieríasUP POSGRADOS

Ciclo de vida del dato

UNIVERSIDAD PANAMERICANA

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

Gobierno de Datos (Data Governance)

UNIVERSIDAD PANAMERICANA

IngenieríasUP
POSGRADOS

Gobierno de datos

UNIVERSIDAD PANAMERICANA

- 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

UNIVERSIDAD PANAMERICANA

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...

Big Data

- Se refiere al tratamiento de datos masivos:
  - Almacenamiento
  - Procesamiento
  - Análisis
  - Visualización

UNIVERSIDAD PANAMERICANA

- Almacenamiento:
  - Data Lakes, Bases de Datos, Data Warehouses

UNIVERSIDAD PANAMERICANA

- Procesamiento:
  - Buscadores:
    - Solr y Elasticsearch
  - Hadoop

Hadoop

UNIVERSIDAD PANAMERICANA

- Es almacenamiento/procesamiento en paralelo.

Big Data

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

UNIVERSIDAD PANAMERICANA

- Visualización
  - PowerBI (Microsoft)
  - Tableau

Características de Big Data

UNIVERSIDAD PANAMERICANA

- 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


ARE YOU SURE THIS IS HOW WE GET DATA INTO THE CLOUD?
IT

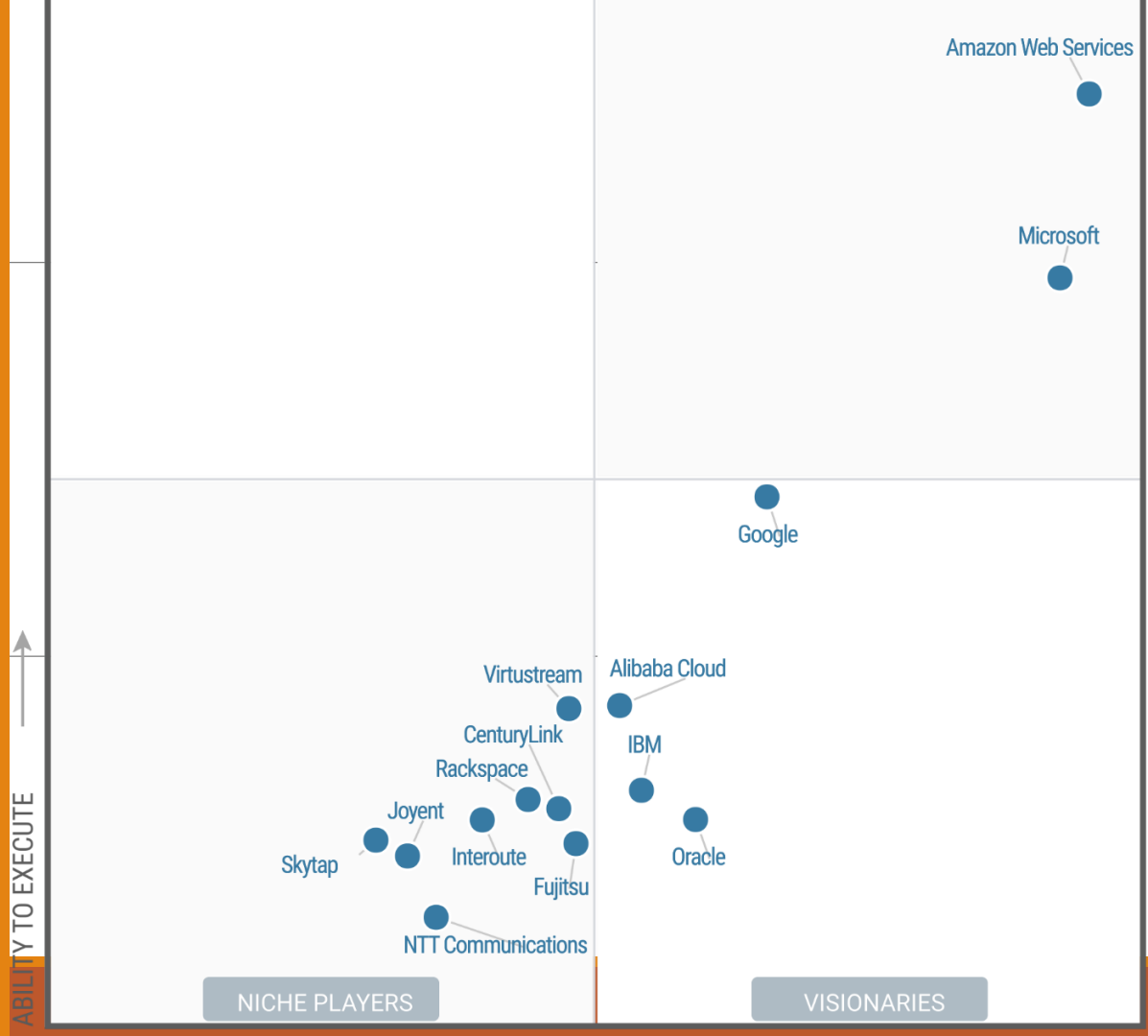
UNIVERSIDAD PANAMERICANA

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

UNIVERSIDAD PANAMERICANA

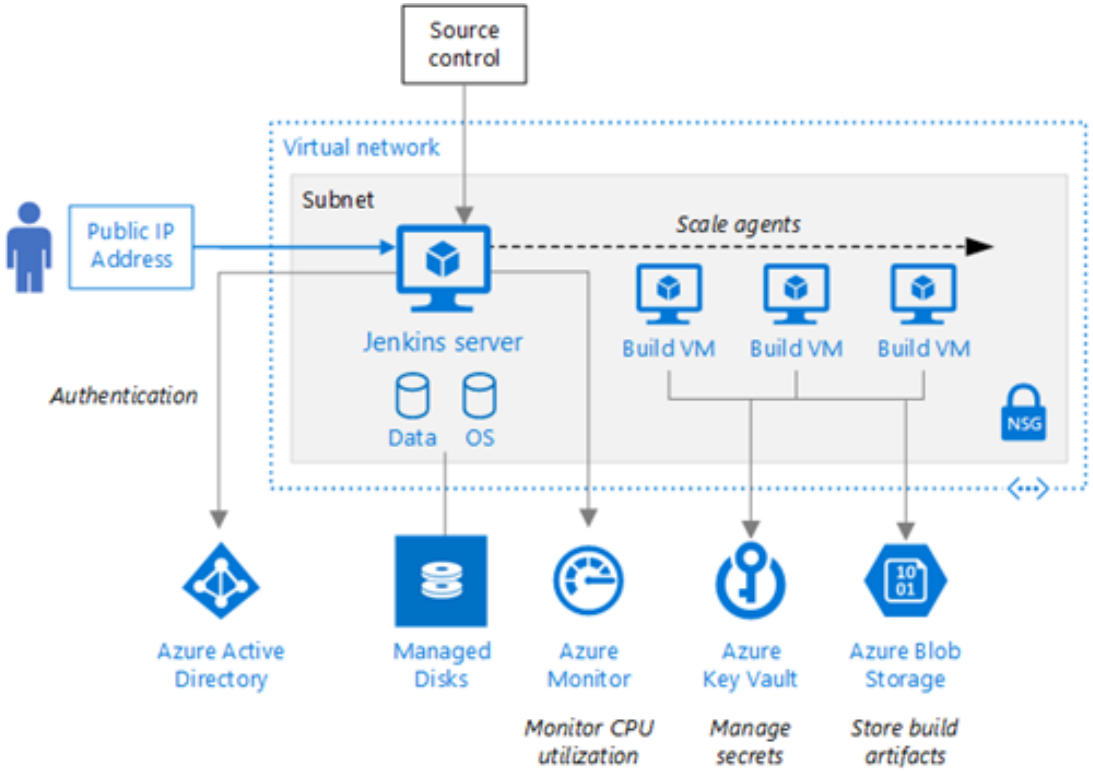
- Hay tres servicios de cloud (bueno, hay más):

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.


Source control
Virtual network
Subnet
Public IP Address
Scale agents
Jenkins server
Build VM
Build VM
Build VM
Authentication
Data
OS
NSG
Azure Active Directory
Managed Disks
Azure Monitor
Azure Key Vault
Azure Blob Storage
Monitor CPU utilization
Manage secrets
Store build artifacts