BIG DATA AND HADOOP

JULITA INCA

Octubre 2017

UNTELS - PERU

EVOLUCIÓN DE LA DATA

EVOLUCIÓN DE LA TECNOLOGIA

DIMENSIONES DE LOS DATOS

- variedad
- volumen
- velocidad
- veracidad
- valor

IOT

EVOLUCIÓN DE LA DATA

GENERALIDADES DE IOT

- IOT conecta un dispositivo con internet y hace el dispositivo “smart”
- Aire acondicionado ejemplo
- 50 billones para el año 2020

SOCIAL MEDIA

EVOLUCIÓN DE LA DATA

FACTOR DEL CRECIMIENTO DE DATOS

- 100 millones de horas de contenido de videos son vistos en Facebook diarios.

- Cada mes 2.5 billones de comentarios son hechos en las páginas de Facebook.

- Instagram tiene 1 millón de publicidad activa desde Marzo 2016.

Cuántos millones de mails se tiene?

Fuente : https://www.bluecorona.com/blog/social-media-statistics-2017

OTROS FACTORES

amazon

NASA

TOYOTA

UBER

Que es Big Data?

Big data es un término evolutivo que describe cualquier cantidad voluminosa de datos estructurados, semiestructurados y no estructurados que tienen el potencial de ser extraídos para obtener información.

¿COMO SABEMOS QUIEN ES BIG DATA?

- es difícil de procesarlo por el alto volumen
- viene de diferentes medios (variedad)
- es alarmante la cantidad generada por segundo
- es necesario tener solamente información útil
- se requiere validar la consistencia de los datos

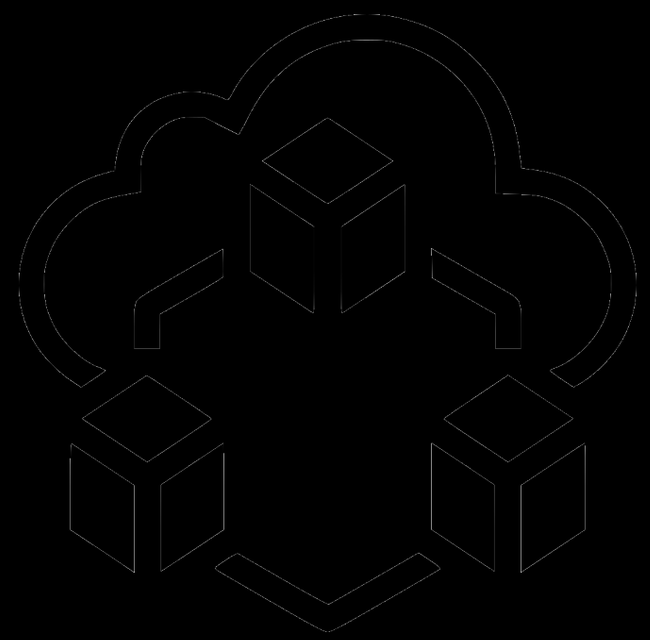
BIG DATA

COMO OPORTUNIDAD

Uso de commodity hardware

Big data analytics -> decisiones

estudio de patrones



BIG DATA CASOS

Spotify está a la altura de Google, Facebook o Amazon.
empresa valorada en 7.500 millones de €
en su día aspiró a comprar Google música.
Spotify predijo el año pasado con un 67% de precisión
a los ganadores de los premios Grammy
Shazan hizo lo propio prediciendo con 33 días de antelación
el top de listas de ventas usando sus propios datos.

https://loogic.com/como-spotify-utiliza-el-big-data-para-cambiar-la-industria-de-la-musica/

http://www.ibmbigdatahub.com/whitepaper/beyond-smart-meters

https://es.slideshare.net/Dell/big-data-use-cases-36019892

Cuales serían los problemas con Big Data?

ALMACENAMIENTO

EXPONENCIAL CRECIMIENTO DE LAS BD

Para 2020 el total de la data crecerá en 44 zetabytes
y cada persona generara 1.7MB por segundo

COMPLEJOS FORMATOS

PROCESAMIENTO DE LA DATA

almacenamiento y procesamiento de data
sin estructura, semi-estructura y estructurada

LECTURA Y ESCRITURA DISCO

PROCESAMIENTO RAPIDO DE LA DATA

capacidad de almacenamiento versus la velocidad de lectura y escritura de disco

Cual sería una solución?

Hadoop

Framework que permite almacenar data
en ambiente distribuido y procesa
de manera paralela

HDFS

ALMACENAMIENTO

cualquier data de muchos formatos en el cluster

MAP REDUCE

PROCESAMIENTO

procesa paralelamente data almacenada a lo largo de HDFS

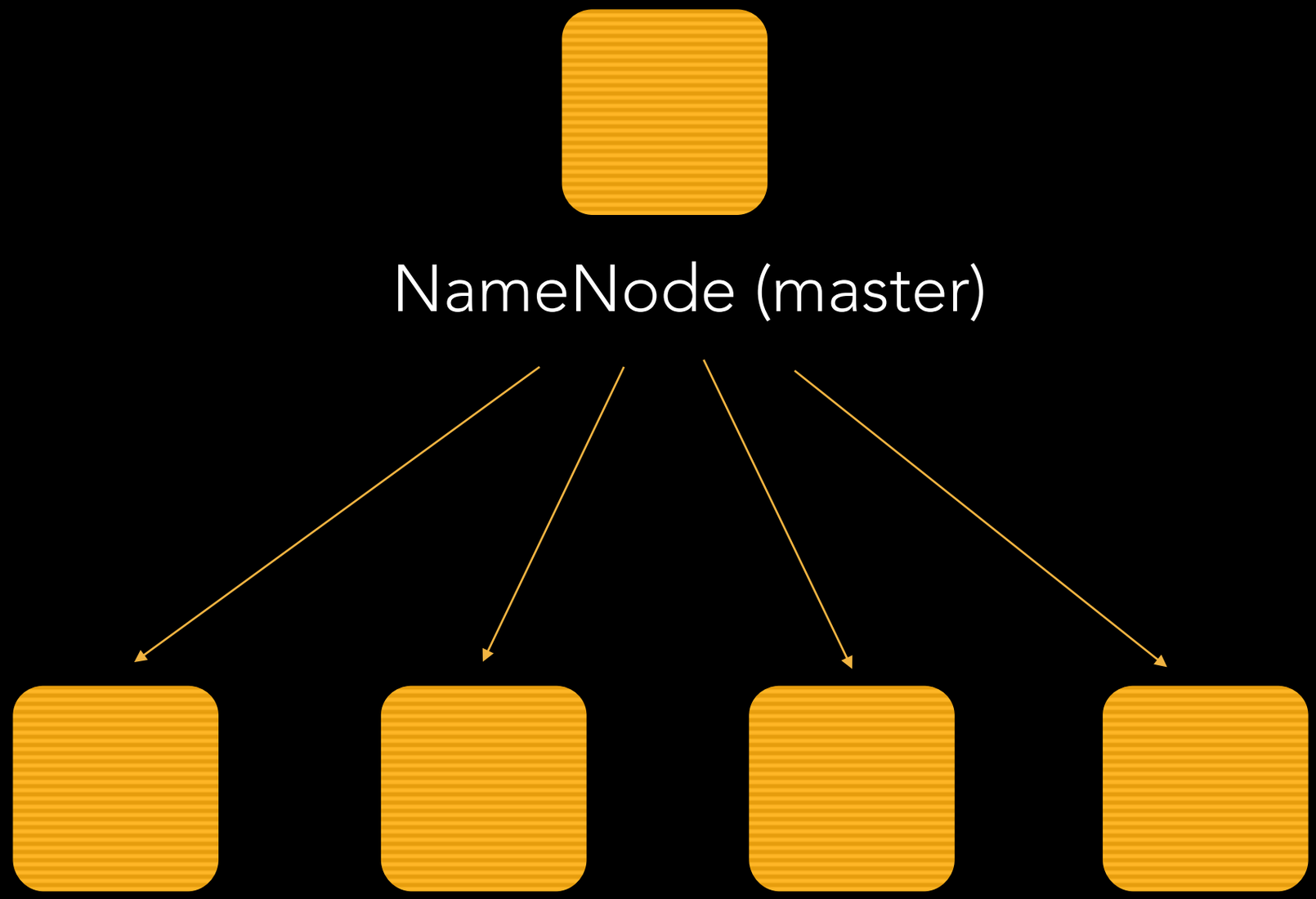
HDFS

Crea nivel de abstracciones de recursos

Unidad lógica para almacenar big data

Se tiene arquitectura distribuida

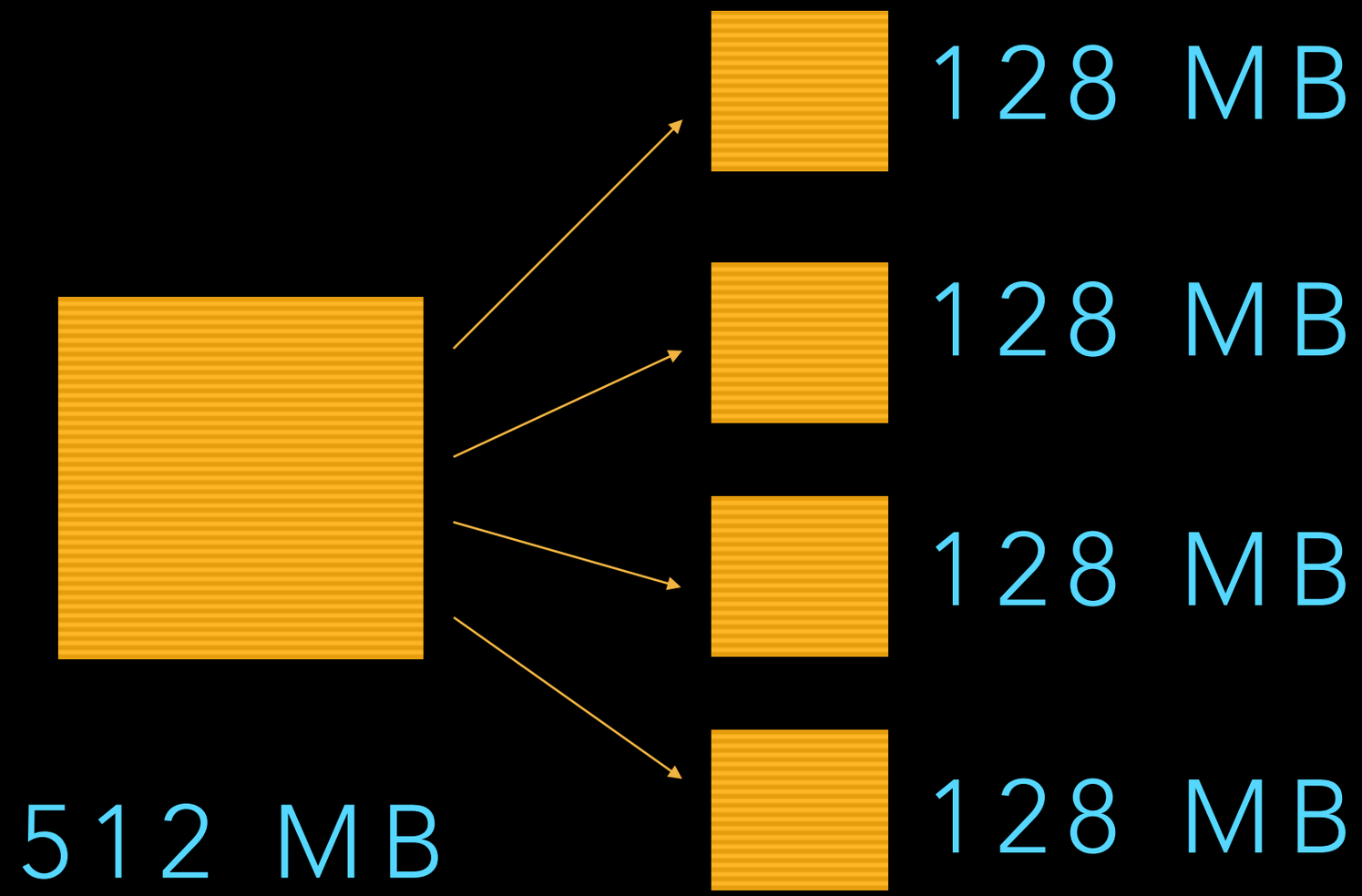
master - slave


NameNode (master)
DataNode (slaves)

el problema del almacenamiento exponencial de enormes conjuntos de datos

SOLUCIÓN: HDFS

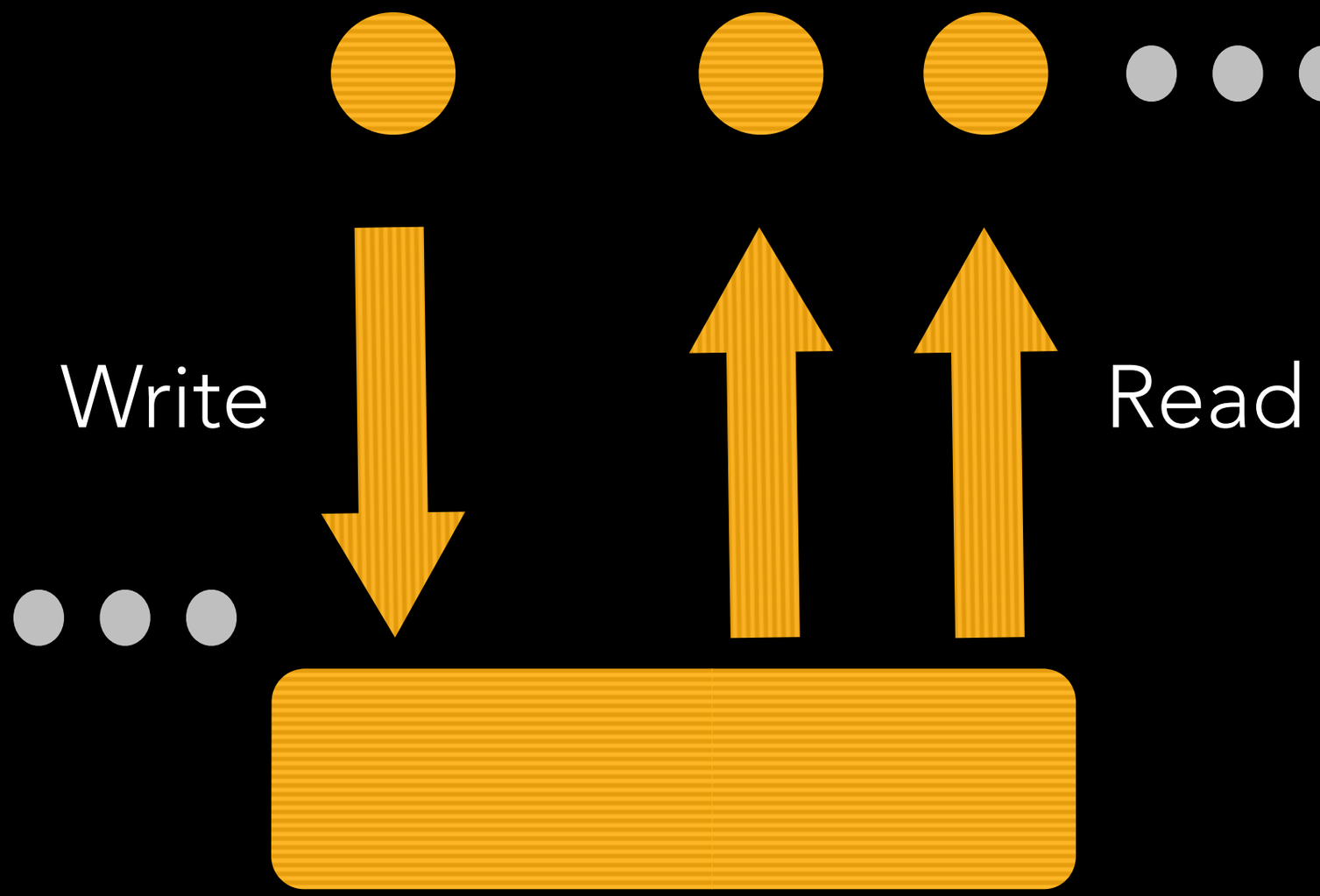
- almacena en unidades de Hadoop
- un sistema de archivo distribuido
- divide los archivos en pequeños y los almacena a lo largo del cluster
- escala según requerimiento


128 MB
128 MB
128 MB
128 MB
512 MB

el problema del almacenamiento de datos no estructurados

SOLUCIÓN: HDFS

- Permite almacenar cualquier tipo de datos, sea estructurado, semi-estructurado o no estructurado
- Cumple WORM (Write Once Read Many)
- Ningún esquema de validación es realizado mientras hay descarga de datos


Write
Read

el problema del procesamiento rápido de datos

SOLUCIÓN: HADOOP MAPREDUCE

- Provee procesamiento paralelo de los datos presentes en HDFS
- Permite procesar data localmente, por ejem. cada nodo trabaja con la data que almacena.

4 horas

1 hora

1 hora

1 hora

1 hora

–TEORIA BASADA EN:

https://www.udemy.com/big-data-and-hadoop-edureka/learn/v4/t/lecture/7296018?start=0

– EXPERIENCIA PRACTICA:

https://lleksah.wordpress.com/2016/07/10/preparing-my-first-paper-related-to-hpc-bigdata/

GRACIAS

jinca· GitHub

@yulwitter