

DESPLIEGUE DE UN CLÚSTER DE BIG DATA



¿QUÉ ES BIG DATA?

Es un gran volumen de información de diferentes fuentes, con estructuras distintas que son difíciles de procesar y analizar con los sistemas de cómputo tradicionales.



VELOCIDAD



Los datos se generan y almacenan a una velocidad sin precedentes.

VARIABILIDAD

Los datos provienen de diversas fuentes, herramientas y plataformas.



VOLUMEN

Gran cantidad de información difícil de procesar con medios tradicionales.



VALOR

Es necesario saber qué tan pertinente es la información para los objetivos que se busca.

VERACIDAD

Las empresas deben asegurarse que los datos que están recopilando tengan validez.



¿BIG DATA, PARA QUÉ?



PREDICTIVE MAINTENANCE

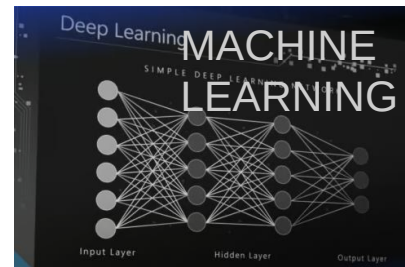
Big data can help predict equipment failure.

OPERATIONAL EFFICIENCY

Operational efficiency is one of the areas in which big data can have the most impact on profitability. With big data, you can analyze and assess production processes, proactively respond to customer feedback, and anticipate future demands.

PRODUCTION OPTIMIZATION

Optimizing production lines can decrease costs and increase revenue.



MACHINE LEARNING

We are now able to teach machines instead of program them.
The availability of big data to train machine learning models makes that possible.



OPTIMIZE NETWORK CAPACITY

Optimal network performance is essential for a telecom's success.

TELECOM CUSTOMER CHURN

By analyzing the data telecoms already have about service factors, telecoms can predict overall customer satisfaction.

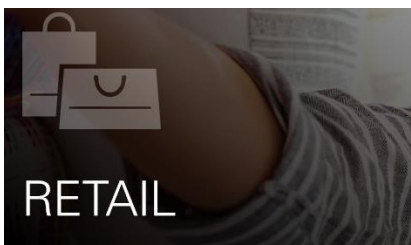
NEW PRODUCT OFFERINGS

Big data provides valuable insights to help companies



GENOMIC RESEARCH

Big data can play in a significant role in genomic research. Using big data, researchers can identify disease genes and biomarkers to help patients pinpoint health issues they may face in the future. The results can even allow healthcare organizations to design personalized treatments.



PRODUCT DEVELOPMENT

Big data can help you anticipate customer demand.

PRICING ANALYTICS AND OPTIMIZATION

Retailers need to know the true profitability of their customers, how markets can be segmented, and the potential of any future opportunities. End-to-end profit and margin analysis can help with identifying pricing improvement opportunities and areas where profits may be leaking.

CUSTOMER LIFETIME VALUE

All customers are valuable. But some are more valuable than others.

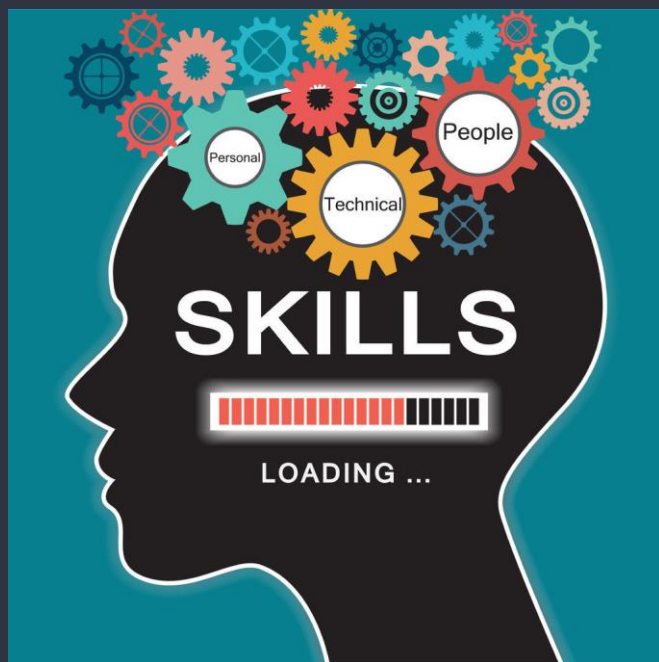
Analizar

Optimizar

Predecir

¿POR QUÉ HACEMOS ESTE PROYECTO?

ADQUIRIR Y TRANSFERIR

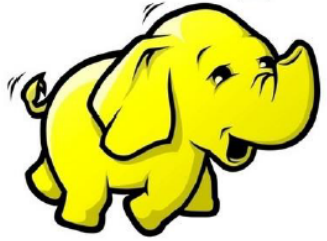


CONOCIMIENTO TÉCNICO

USAR Y APLICAR EL CLÚSTER



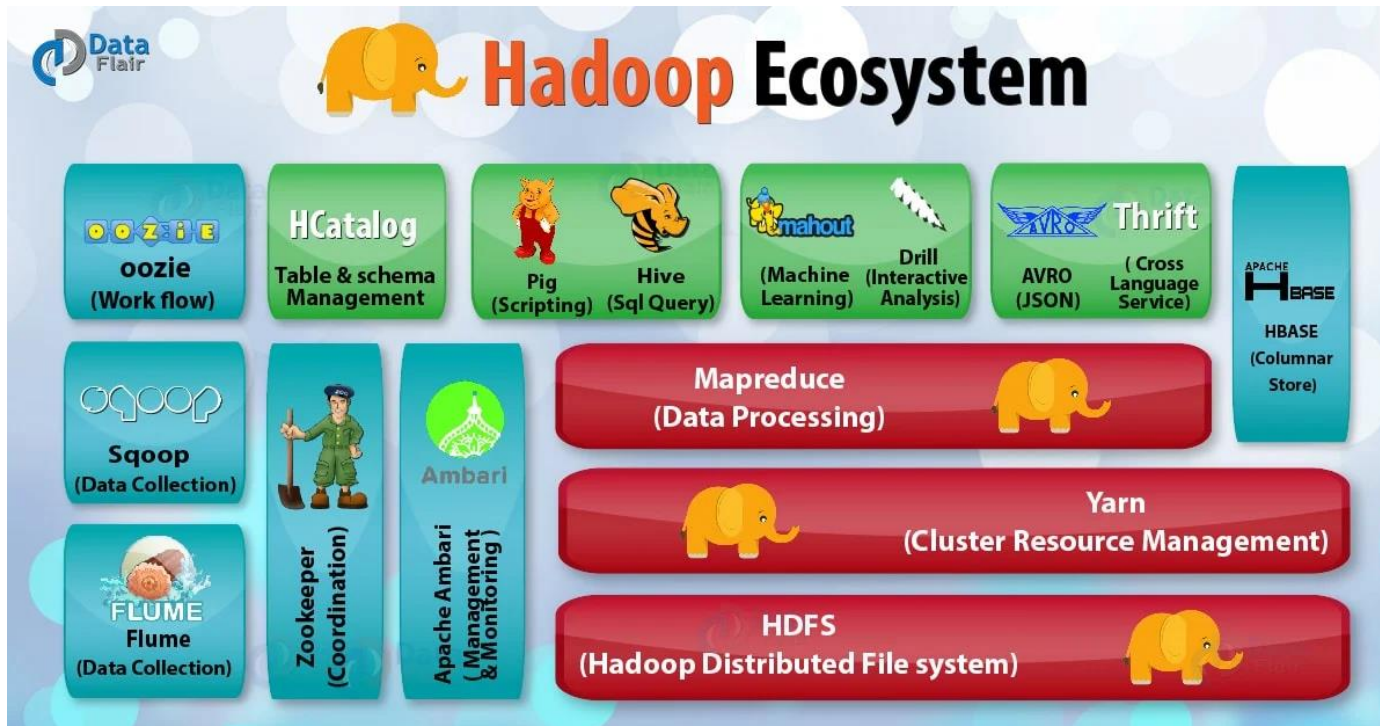
hadoop



QUÉ ES HADOOP

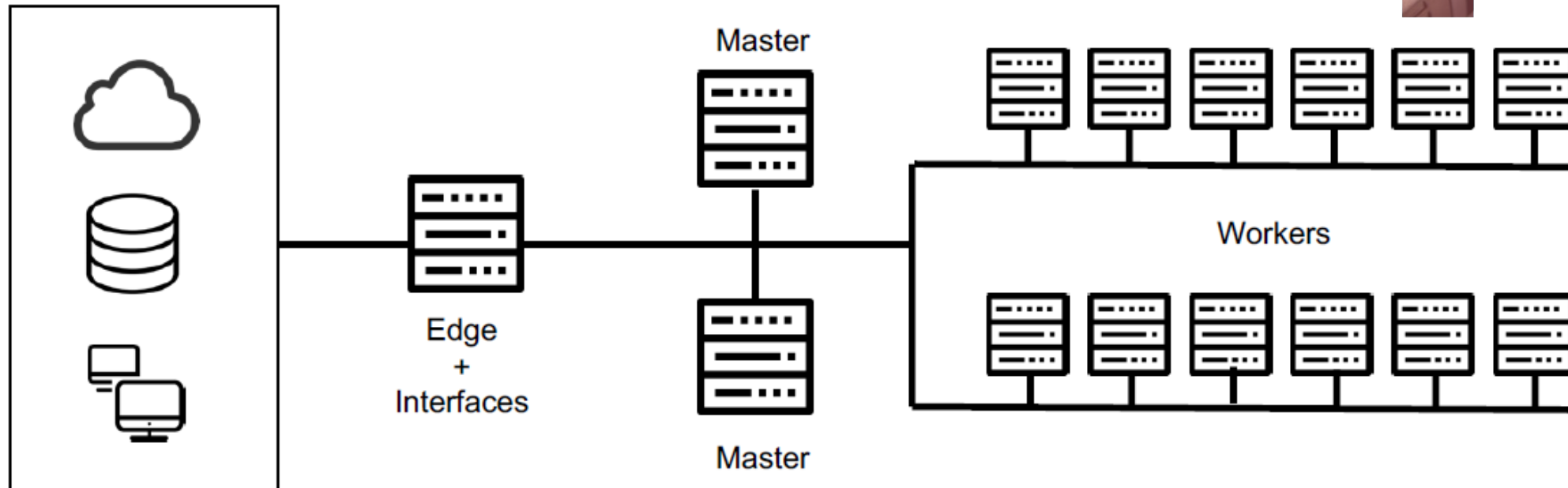
Plataforma opensource que ofrece la capacidad de almacenar y procesar, a “bajo” coste, grandes volúmenes de datos, sin importar su estructura, en un entorno distribuido, escalable y tolerante a fallos, basado en la utilización de **hardware commodity** y en un paradigma acercamiento del procesamiento a los datos.

Se identifica el nombre Hadoop con todo el **ecosistema de componentes** (software) independientes que suelen incluirse para dotar a Hadoop de funcionalidades necesarias en proyectos Big Data empresariales, como puede ser la ingesta de información, el acceso a datos con lenguajes estándar, o las capacidades de administración y monitorización.

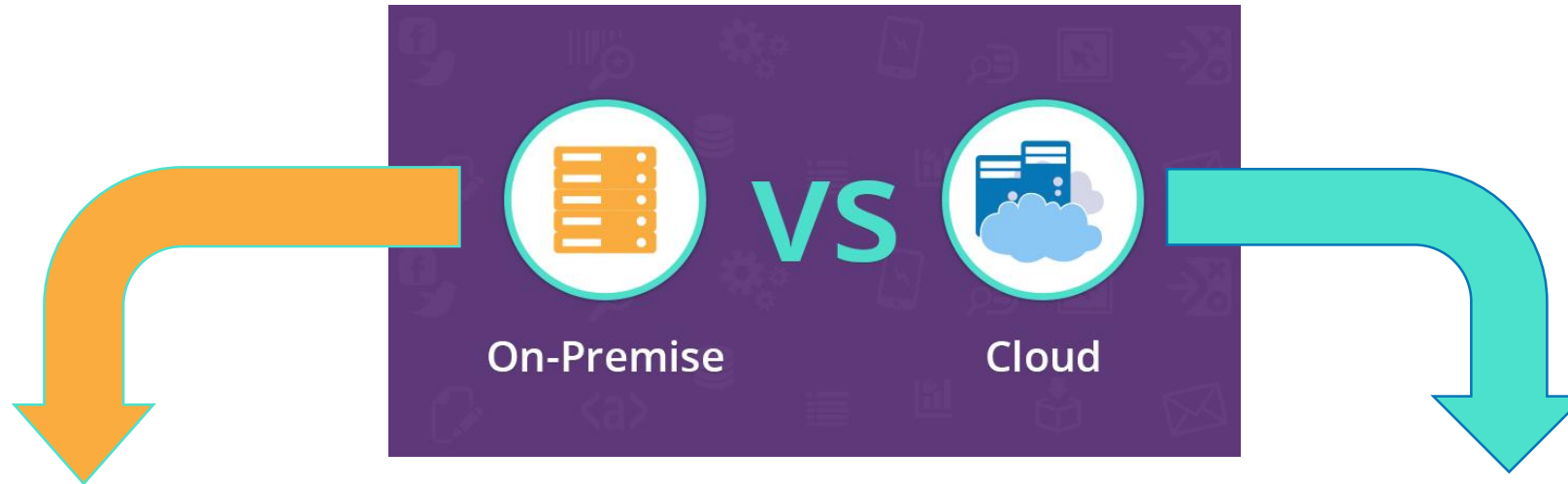


HARDWARE: CLÚSTER DE SERVIDORES

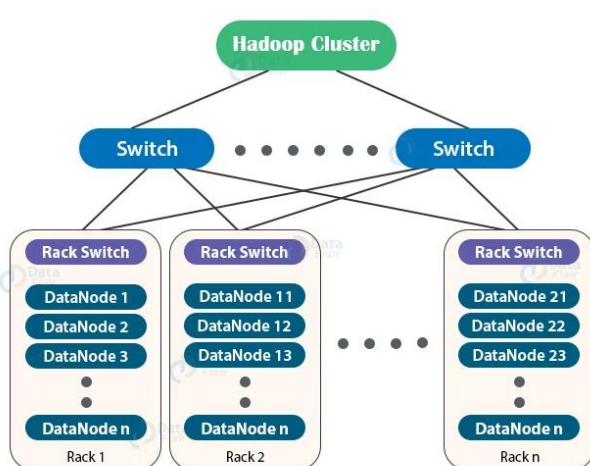
Un clúster es un conjunto de servidores que se gestionan juntos y participan en la gestión de la carga de trabajo.



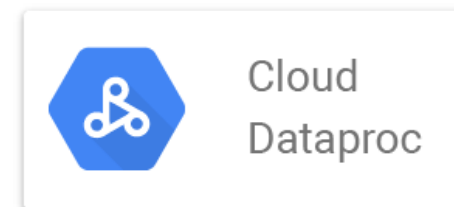
SOLUCIONES CLÚSTER BIG DATA



On Premise Infrastructure



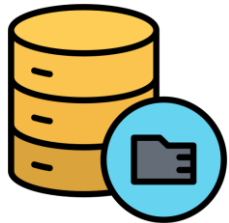
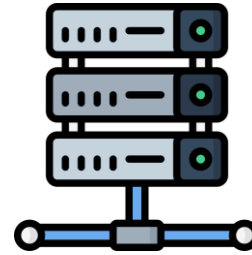
Cloud PaaS (Platform as a Service)



NUESTRO CLÚSTER HW BÁSICO

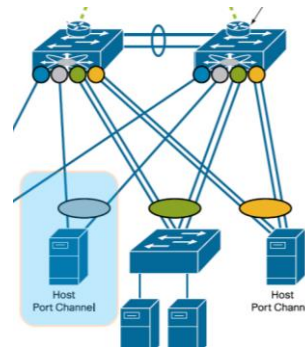
4 SERVIDORES DE 20 NÚCLEOS PROCESADORES Y 256 GB RAM

+ 1 SERVIDOR DE 20 NÚCLEOS Y 64GB RAM



~ 40 TB ALMACENAMIENTO EN DISCO

(13TB útiles en HDFS)



RED DE ALTA VELOCIDAD 20Gbps

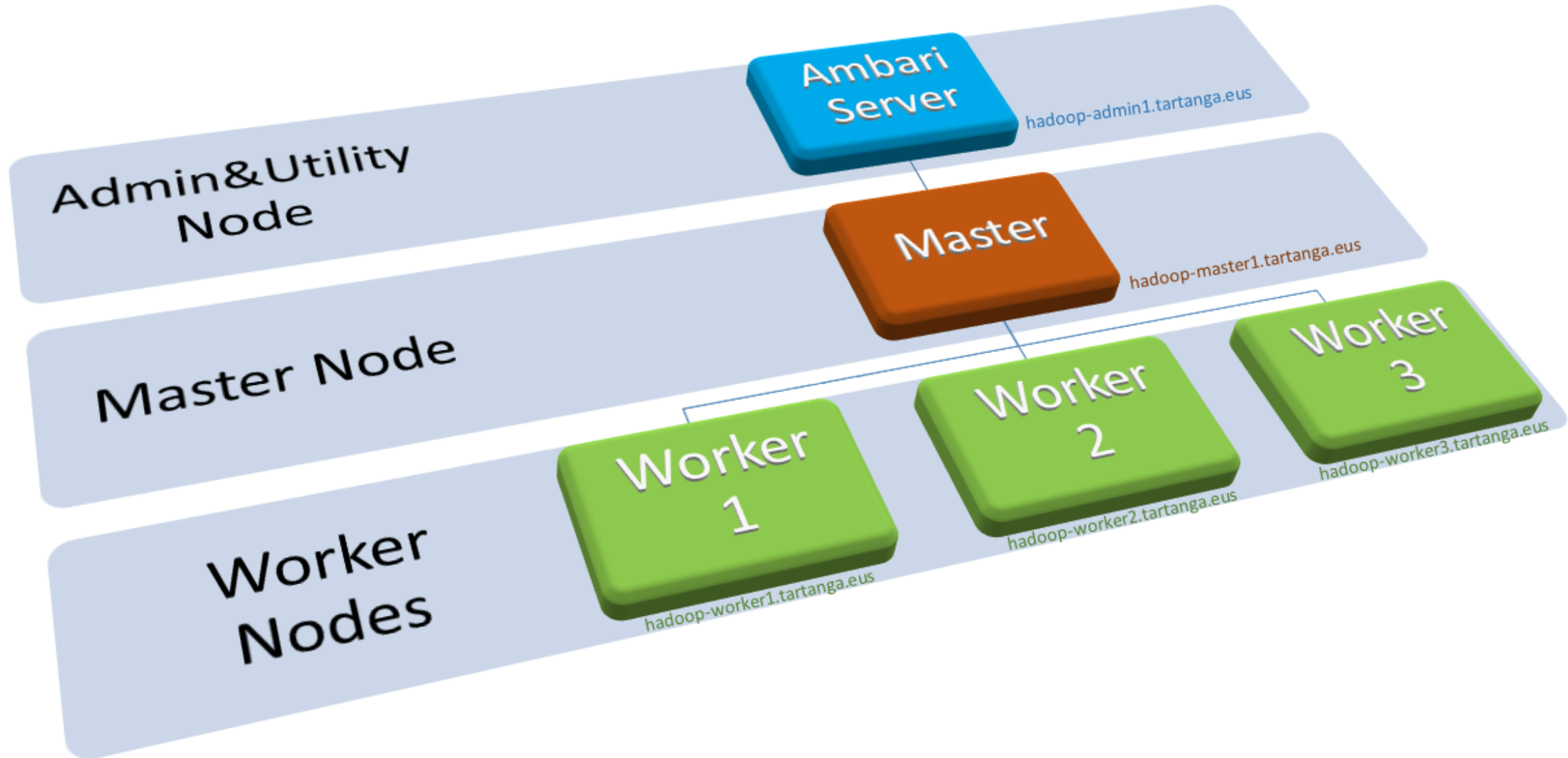
SWITCHES Y ENLACES AGREGADOS de 10Gbps

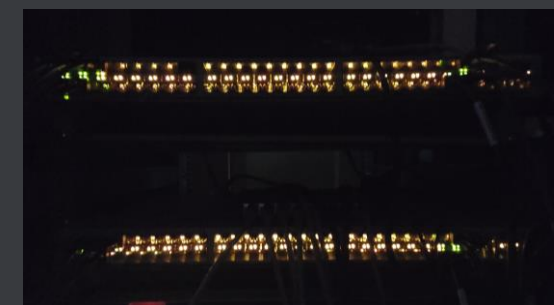
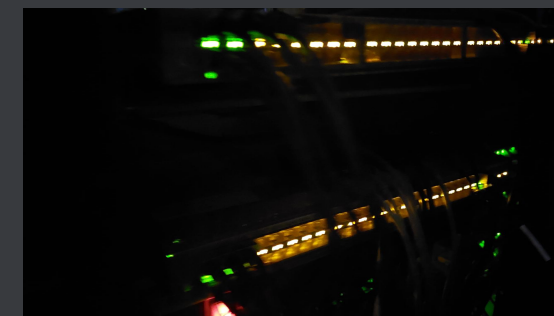
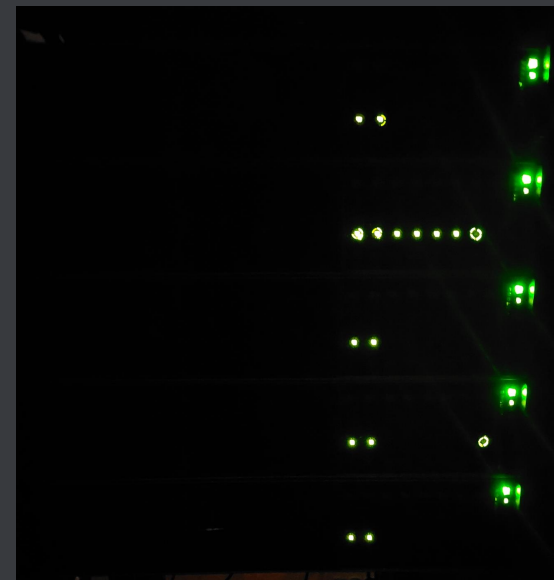
Processing

Storage

Network

Estructura del clúster



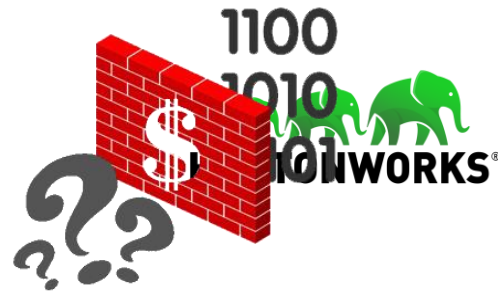




SOFTWARE: EL PROBLEMA DE LAS DISTRIBUCIONES/VERSIONES

- * CARENCIA DE DISTROS NO PROPIETARIAS CON REPOSITORIOS PÚBLICOS
- * COMPATIBILIDAD AMBARI, MPACK Y OS: FALTA DE MANTENIMIENTO CORRECTIVO-EVOLUTIVO DE STACKS PARA DIVERSOS SISTEMAS OPERATIVOS

CLOUDERA



El acceso a los antiguos repositorios públicos de HortonWorks está “paywalled” por Cloudera.

CARENCIA DE
DISTROS NO
PROPIETARIAS
CON
REPOSITORIOS
PÚBLICOS

CLOUDERA

[Why Cloudera](#)[Products](#)[Solutions](#)[Resources](#)[Support](#)

November 2020

Software Access Update

This announcement pertains to accessing Cloudera's software.

View in [Chinese](#), [Japanese](#) and [Korean](#).

We announced in 2019 that, beginning November 2019, all new Cloudera product releases, including version updates and maintenance releases of current software, would only be accessible by current customers with a valid product subscription to the software.

Effective January 31, 2021, all Cloudera software requires a valid subscription to that software in order to access it. This includes all legacy versions of Cloudera products, including Apache Hadoop (CDH), Hortonworks Data Platform (HDP), Data Flow (HDF/CDF), and Cloudera Data Science Workbench (CDSW). Information regarding software access is available in technical documentation by software type and version. Additional information is available in the Frequently Asked Questions section of this document.



The **Apache Software Foundation**

<http://www.apache.org/>

[Apache](#) / [Bigtop](#) / Apache Bigtop

Version: 3.3.0-SNAPSHOT | Last Published: 2023-08-22

Apache Bigtop

Bigtop is an Apache Foundation project for Infrastructure Engineers and Data Scientists looking for comprehensive packaging, testing, and configuration of the leading open source big data **components**. Bigtop supports a wide range of components/projects, including, but not limited to, Hadoop, HBase and Spark.

Packaging

Bigtop packages Hadoop RPMs and DEBs, so that you can manage and maintain your Hadoop cluster.

Smoke testing

Bigtop provides an integrated smoke testing framework, alongside a suite of over 50 test files.

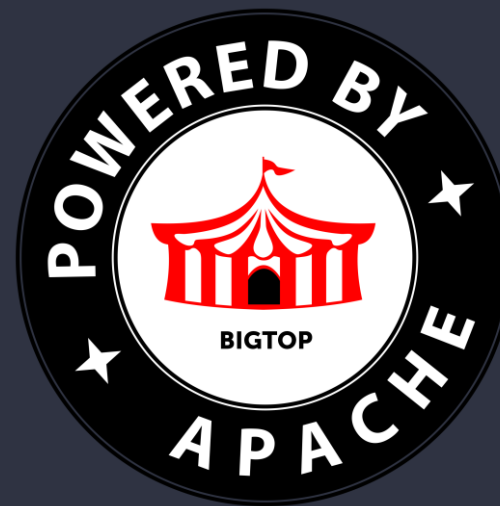
Virtualization

Bigtop provides docker recipes and raw images for deploying Hadoop from zero.

Contenedores y
virtualización:

¿Dificultades en
entornos no
contenidos/
virtualizados?

ALTERNATIVA:



¿CÓMO DESPLEGAR BIGTOP CON AMBARI?

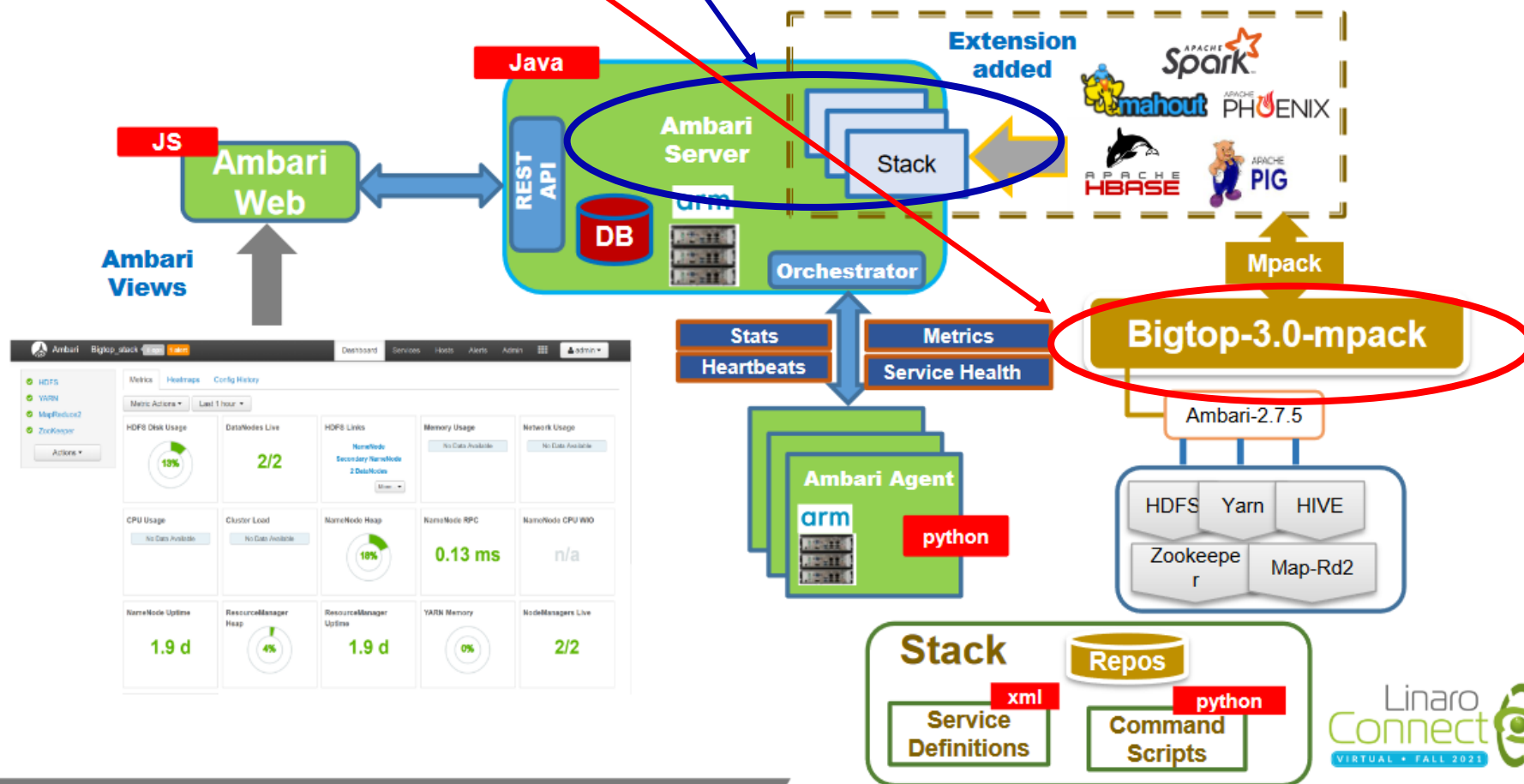
Ambari Management Packs: desacoplan la funcionalidad principal de **Ambari** (administración y monitorización) de la gestión de los stacks de servicios (BigTop, HDP,...).

DESPLIEGUE Y GESTIÓN DEL CLÚSTER CON AMBARI



Apache Ambari

Big Picture of Bigtop-3.0 Mpack



COMPATIBILIDAD DE AMBARI



Apache Ambari

Note: Ambari currently supports the 64-bit v

- RHEL (Redhat Enterprise Linux) 7.4, 7.3,
- CentOS 7.4, 7.3, 7.2
- OEL (Oracle Enterprise Linux) 7.4, 7.3, 7
- Amazon Linux 2
- SLES (SuSE Linux Enterprise Server) 12
- **Ubuntu 14 and 16**
- Debian 9

Installation Guide for Ambari 2.8.0

Creado por Zhiguo Wu, modificado por última vez en mar 31, 2023

Build and install Ambari 2.8.0

Refer [Ambari Development](#) for prerequisites and additional instructions. Maven blocks external HTTP repositories by default so

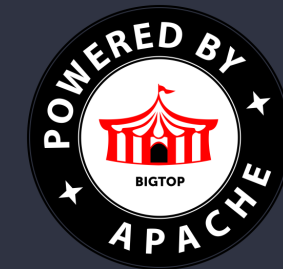
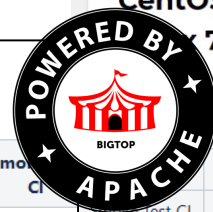
Notice: Ambari 2.8.0 only supports CentOS-7(x86_64) currently

End dates are coming for
CentOS Stream 8 and CentOS

Overview of Bigtop 3.2.1 Support Matrix

Creado por Masatake Iwasaki, modificado por última vez por Kengo Seki el ago 19, 2023

Component	Runtime Dependencies	Package CI	Puppet Deployment	Smoke Test	Smoke Test CI	PPC64LE		
						Smoke Test CI	Package CI	Smoke Test CI
Alluxio	HDFS	✓	✓	✓	✓	✓	-	-
Ambari	-	✓	✓	✓	✓	✓	-	-
						✓	✓	✓
						✓	-	-
						✓	✓	✓
						✓	-	-
						✓	✓	✓
						✓	✓	✓
						✓	✓	✓
						✓	✓	✓



AMBARI: FALTA DE
MANTENIMIENTO
CORRECTIVO-
EVOLUTIVO PARA
DIVERSOS OS



Apache Ambari

ENSAYO 1: INCOMPATIBILIDAD AMBARI 2.7.5 UBUNTU 20.04

Installer

Get Started

Select Version

Install Options

Confirm Hosts

Choose Services

Assign Masters

Assign Slaves and Clients

Customize Services

Select Version

BigTP-3.2

Ambari Metrics

Bigtop+3.2

Review

Please review the configuration before installation

Hive

Metastore : hadoop-master1.tartangalh.eus
HiveServer2 : hadoop-master1.tartangalh.eus
WebHCat Server : hadoop-master1.tartangalh.eus
Database : Existing PostgreSQL Database

HBase

Master : hadoop-master1.tartangalh.eus
RegionServer : 3 hosts

ZooKeeper

Server : hadoop-master1.tartangalh.eus

Ambari Metrics

Metrics Collector : hadoop-master1.tartangalh.eus
Grafana : hadoop-master1.tartangalh.eus

Summary

← BACK

CANCEL

Summary

Ambari 2.7.5

BigTop 3.2.1

Ubuntu 20.04

.SystemException: Operating System matching ubuntu20 could not be found
grade.RepositoryVersionHelper.getOSEntityForHost(RepositoryVersionHelper.java:367)
grade.RepositoryVersionHelper.getCommandRepository(RepositoryVersionHelper.java:432)
ariManagementControllerImpl.retrieveHostRepositories(AmbariManagementControllerImpl.java:5935)

RESOLUCIÓN DE PROBLEMAS

```
<reposinfo>
  <os family="redhat7">
    <repo>
      <baseurl>http://repos.bigtop.apache.org/releases/3.2.1/centos/7/$basearch</baseurl>
      <repoid>BGTP-3.2.1</repoid>
      <reponame>BGTP</reponame>
    </repo>
  </os>
  <os family="ubuntu18">
    <repo>
      <baseurl>http://repos.bigtop.apache.org/releases/3.1.1/ubuntu/18.04/$(ARCH)</baseurl>
      <repoid>BGTP-3.1.1</repoid>
      <reponame>BGTP</reponame>
    </repo>
  </os>
</reposinfo>
<server/resources/stacks/BGTP/1.0/repos/repoinfo.xml 33L, 1272C 29,22 Bot
```

Adaptación de management pack de BigTop

hadoop-master1.tartangalh.eus

Tasks

hadoop-worker2.tartangalh.eus

✓ Flink History

■ HBase Master

■ HDFS Client

Tasks

✓ DataNode Install

❗ Flink Client Install

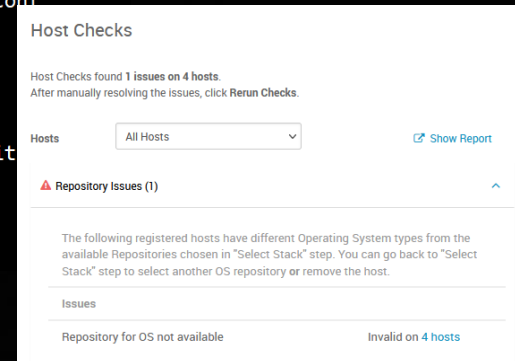
■ HBase Client Install

Resolución de problemas de despliegue

```
<!-- repository-version xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSchemaLocation="version_definition.xsd" -->
<!-- release -->
<release>
  <type>STANDARD</type>
  <stack-id>BGTP-1.0</stack-id>
  <version>1.0</version>
  <build>1</build>
  <!-- compatible-with -->
  <!-- release-notes -->
  <!-- display -->
</release>
<!-- manifest -->
<manifest>
  <service id="AMBARI-METRICS" name="AMBARI-METRICS" version="Bigtop+3.2"/>
  <service id="FLINK-321" name="FLINK" version="Bigtop+3.2"/>
  <service id="HBASE-321" name="HBASE" version="Bigtop+3.2"/>
  <service id="HDFS-321" name="HDFS" version="Bigtop+3.2"/>
  <service id="HIVE-321" name="HIVE" version="Bigtop+3.2"/>
  <service id="KAFKA-321" name="KAFKA" version="Bigtop+3.2"/>
  <service id="SOLR-321" name="SOLR" version="Bigtop+3.2"/>
  <service id="SPARK-321" name="SPARK" version="Bigtop+3.2"/>
  <service id="TEZ-321" name="TEZ" version="Bigtop+3.2"/>
  <service id="YARN-321" name="YARN" version="Bigtop+3.2"/>
  <service id="ZEPPELIN-321" name="ZEPPELIN" version="Bigtop+3.2"/>
  <service id="ZOOKEEPER-321" name="ZOOKEEPER" version="Bigtop+3.2"/>
</manifest>
<!-- available-services -->
<!-- repository-info -->
<!-- os family="redhat7" -->
<os family="ubuntu18">
  <!-- package-version -->
  <!-- repo -->
  <baseurl>http://repos.bigtop.apache.org/releases/3.1.1/ubuntu/18.04/$(ARCH)</baseurl>
  <repoid>BGTP-3.1.1</repoid>
  <reponame>BGTP</reponame>
  <!-- unique -->
</os>
</repository-info>
</repository-version>
```

Creación de Version Definition File (VDF) para BigTop

```
root@hadoop-master1:~# vim /etc/postgresql/12/main/pg_hba.conf
root@hadoop-master1:~# tail /etc/postgresql/12/main/pg_hba.conf
host all all 127.0.0.1/32
# IPv6 local connections:
host all all ::1/128
# Hadoop cluster connections
host all all samenet
# Allow replication connections from localhost, by a user with
# replication privilege.
local replication all
host replication all 127.0.0.1/32
host replication all ::1/128
```



Configuración de Sistemas Operativos y Bases de Datos

NUESTRA COMBINACIÓN SOFTWARE



BIGTOP
3.1.1

Stack de
Hadoop

AMBARI 2.7.5



Gestión del
clúster



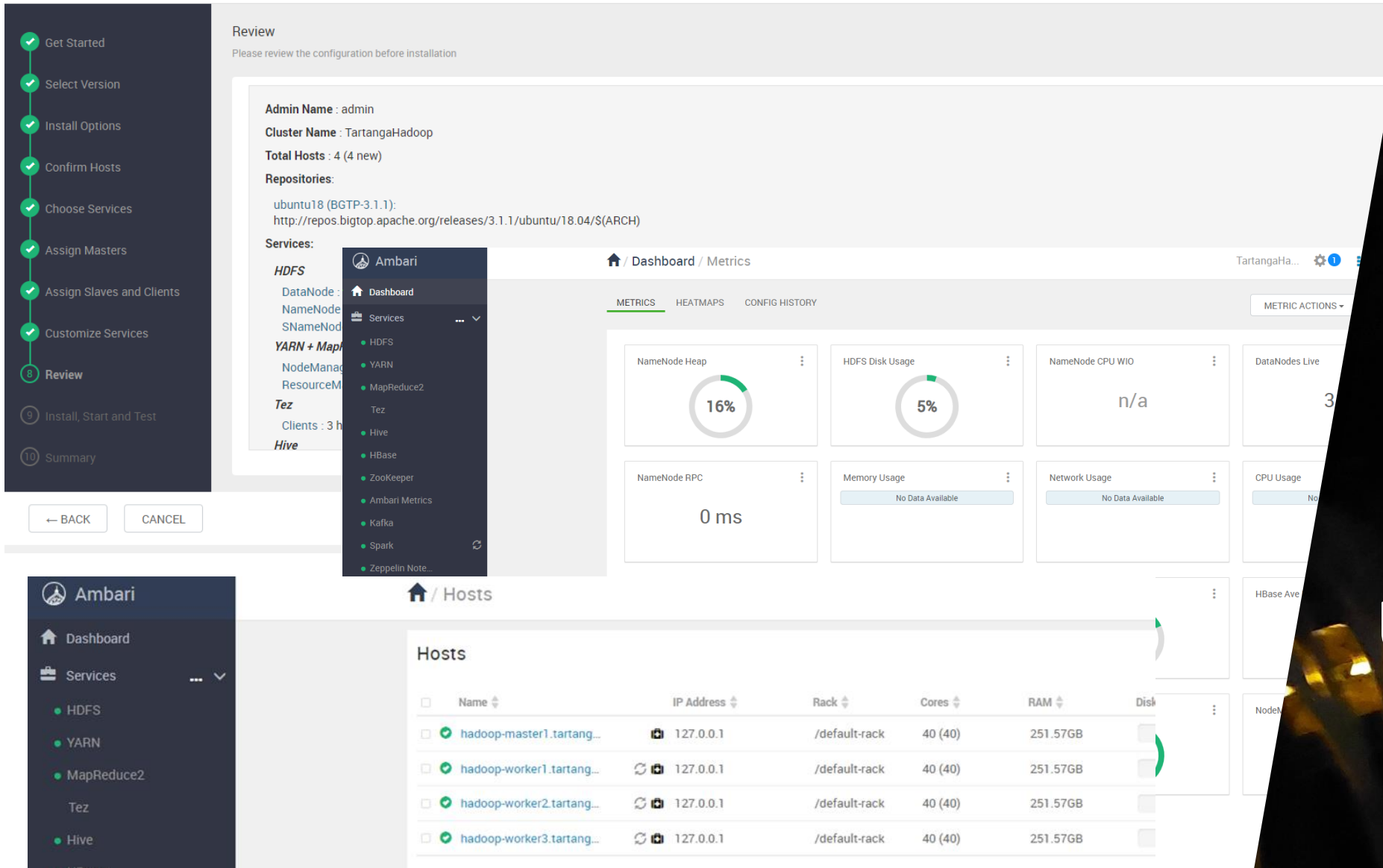
Stack

Manager

Operating System

ENSAYO 2: CLÚSTER DESPLEGADO

Installer



The image displays the Ambari 2.7.5 installer interface during the 'Review' step. On the left, a vertical sidebar shows the installation progress with steps 1 through 10, where step 8 'Review' is highlighted. The main content area is titled 'Review' and includes a warning: 'Please review the configuration before installation'. The configuration details are as follows:

- Admin Name:** admin
- Cluster Name:** TartangaHadoop
- Total Hosts:** 4 (4 new)
- Repositories:** ubuntu18 (BGTP-3.1.1): [http://repos.bigtop.apache.org/releases/3.1.1/ubuntu/18.04/\\$ARCH](http://repos.bigtop.apache.org/releases/3.1.1/ubuntu/18.04/$ARCH)
- Services:**
 - HDFS:** DataNode, NameNode, SNameNode
 - YARN + MapReduce2:** YARN, MapReduce2
 - Tez:** Tez
 - Hive:** Hive

Below the configuration details, there are two screenshots of the Ambari dashboard. The top screenshot shows the 'Metrics' page with various gauges: NameNode Heap (16%), HDFS Disk Usage (5%), NameNode CPU WIO (n/a), DataNodes Live (3), NameNode RPC (0 ms), Memory Usage (No Data Available), Network Usage (No Data Available), and CPU Usage (No Data Available). The bottom screenshot shows the 'Hosts' page with a table of cluster nodes:

Name	IP Address	Rack	Cores	RAM	Disk
hadoop-master1.tartang...	127.0.0.1	/default-rack	40 (40)	251.57GB	
hadoop-worker1.tartang...	127.0.0.1	/default-rack	40 (40)	251.57GB	
hadoop-worker2.tartang...	127.0.0.1	/default-rack	40 (40)	251.57GB	
hadoop-worker3.tartang...	127.0.0.1	/default-rack	40 (40)	251.57GB	

Ambari 2.7.5

BigTop 3.1.1

Ubuntu 18.04

SERVICIOS DEL CLÚSTER



Flink



kafka



Apache Zeppelin



Solr





Apache Hadoop

Hadoop es un “framework” de software.



Permite el procesamiento distribuido de grandes conjuntos de datos en clústers de computadoras utilizando modelos de programación simples.

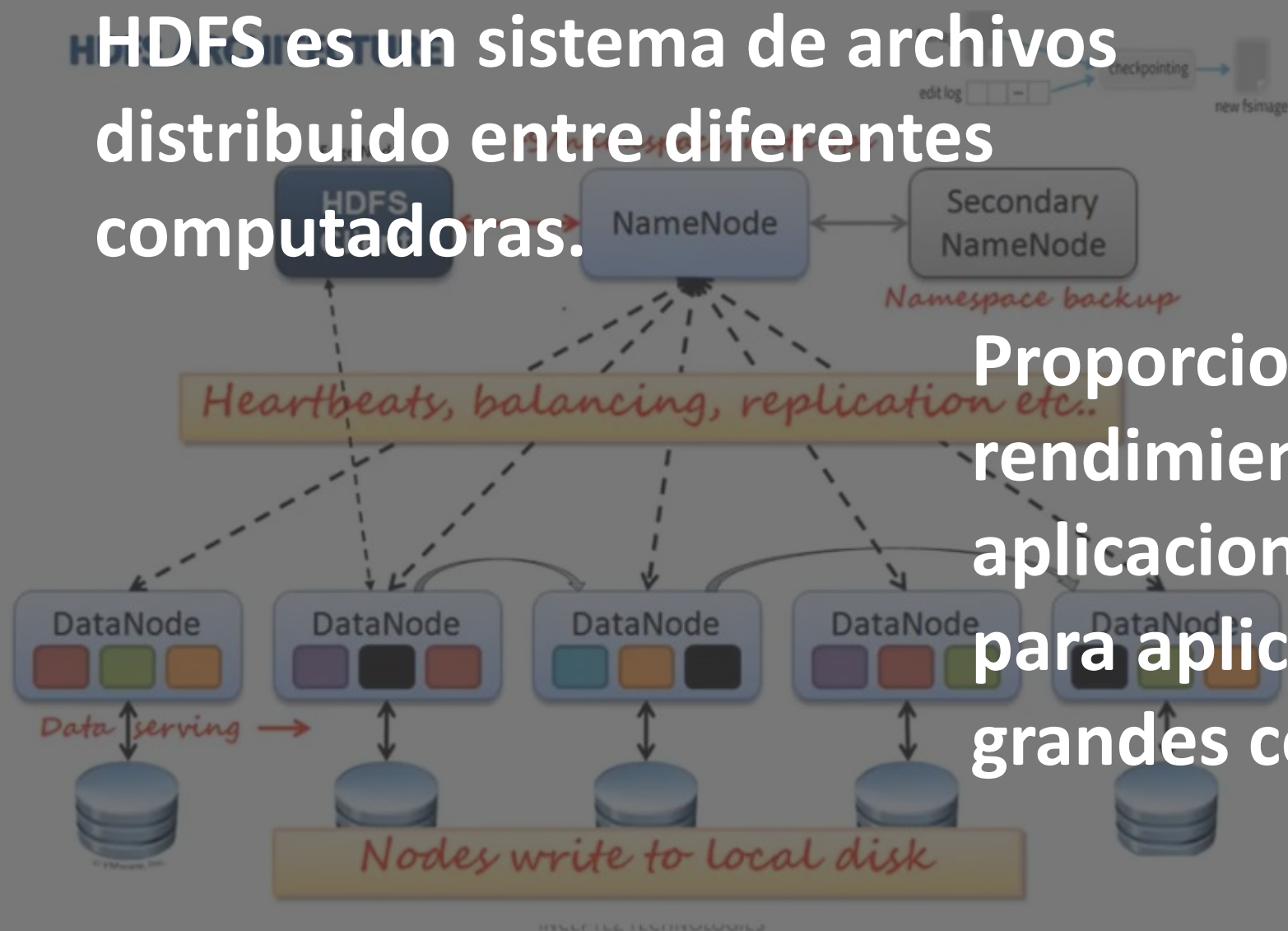
The Apache® Hadoop® project develops open-source, reliable, scalable, distributed computing.

“VANILLA HADOOP”





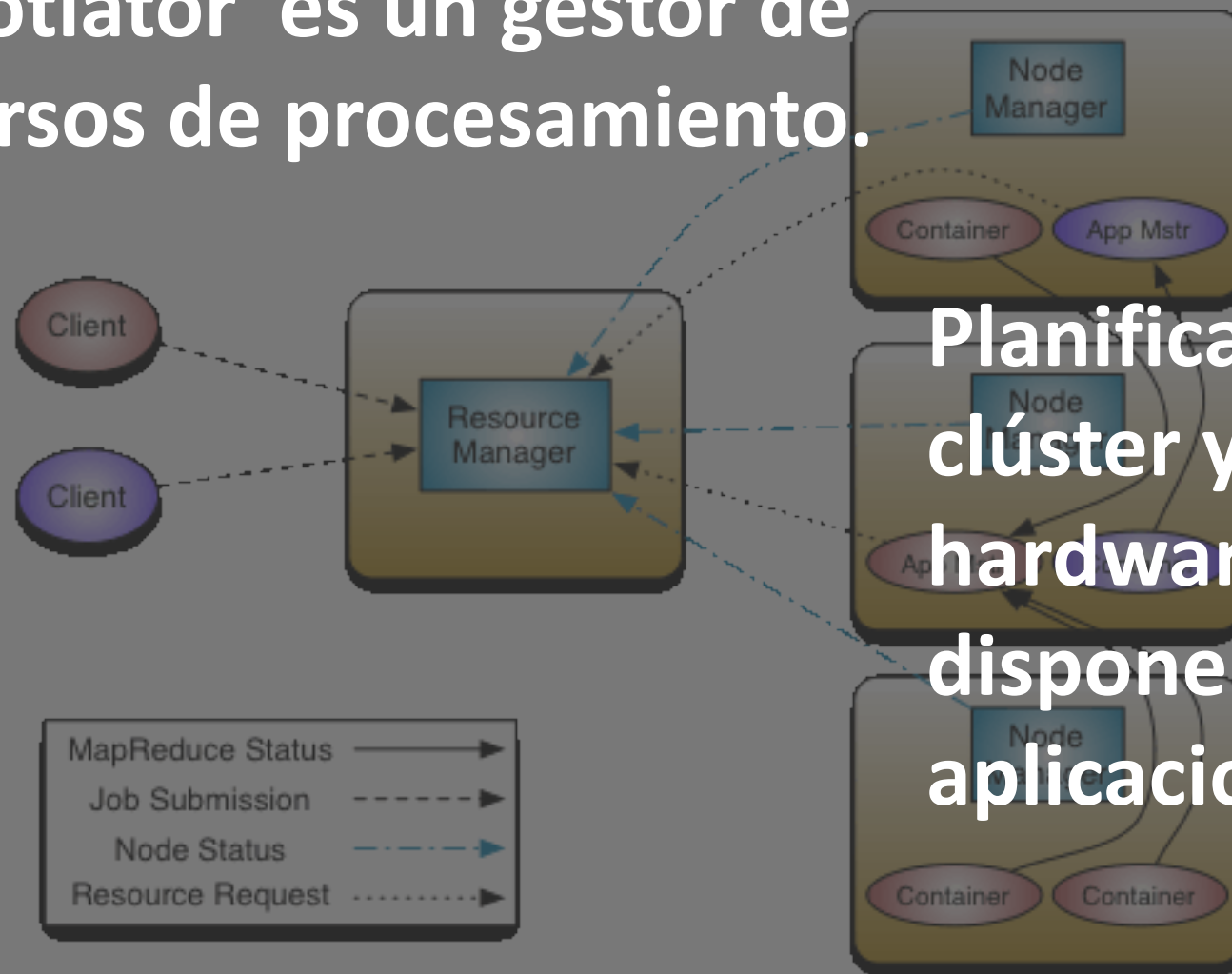
HDFS es un sistema de archivos distribuido entre diferentes computadoras.



Proporciona acceso de alto rendimiento a los datos de las aplicaciones y es adecuado para aplicaciones que tienen grandes conjuntos de datos.

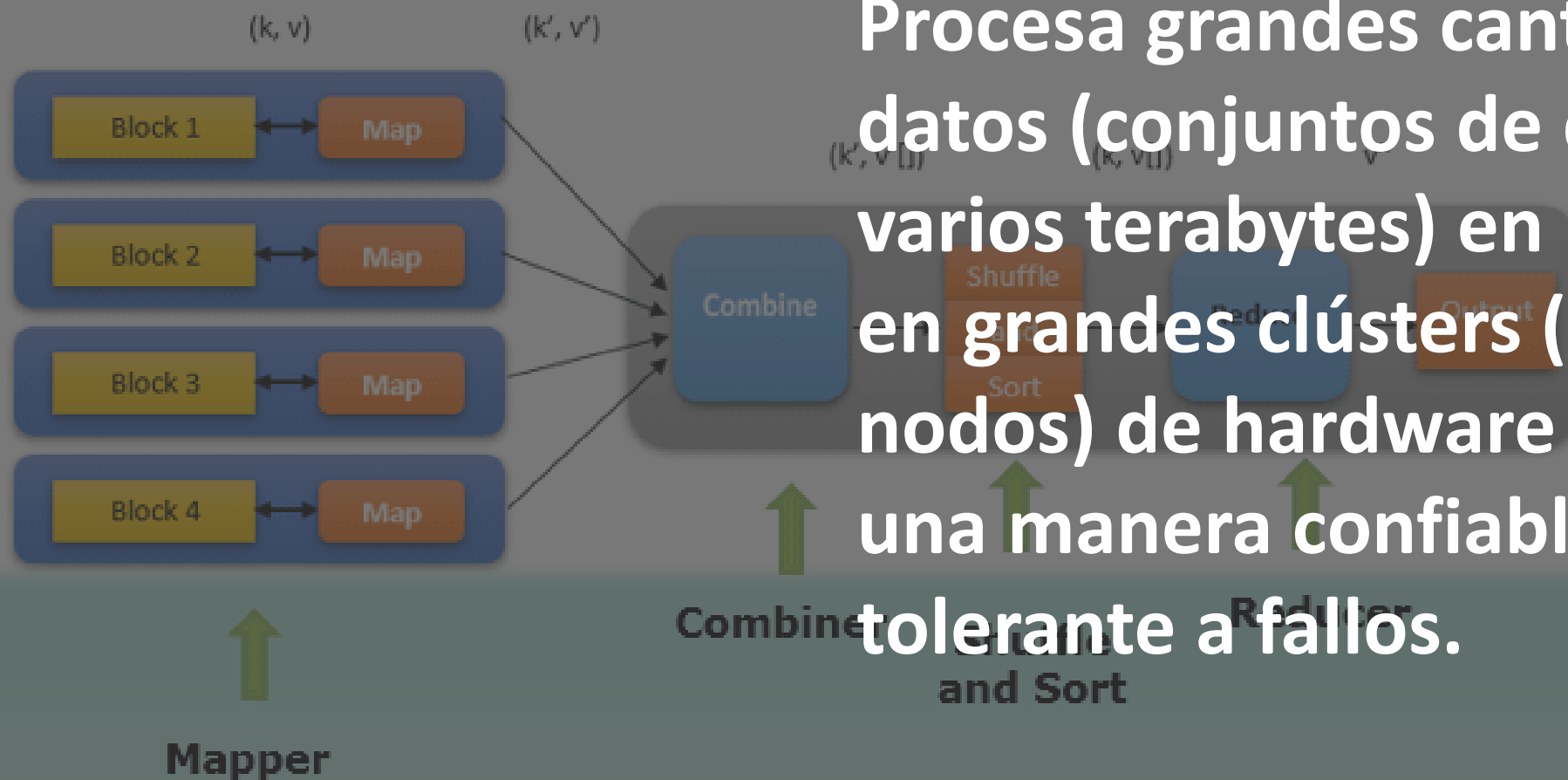
Yet Another Resource

Negotiator es un gestor de recursos de procesamiento.



Planifica los trabajos en un clúster y gestionar los recursos hardware de los que se disponen para ejecutar las aplicaciones Big Data

MapReduce es un marco de software para escribir fácilmente aplicaciones.



Procesa grandes cantidades de datos (conjuntos de datos de varios terabytes) en paralelo en grandes clústers (miles de nodos) de hardware básico de una manera confiable y tolerante a fallos.

PROCESAMIENTO ORIGINAL FRAMEWORK HADOOP

1) Fase de **Ingesta**:
almacenar y manejar
grandes archivos de
datos

HDFS

2) Fase de **Mapeo**: los
datos se procesan en
paralelo generando
pares clave-valor

Map 1

Map N

3) Fase de **Reducción**:
Procesan los pares clave-
valor y los agregan para
producir la salida final de
datos

Reduce

HDFS

Map

Reduce

EJEMPLOS DE EJECUCIÓN



EJEMPLO 1: MapReduce Job “Cálculo de decimales de Pi (π)”

El ejemplo pi usa un método estadístico (cuasi Monte Carlo) para calcular el valor de pi. Los puntos se colocan de forma aleatoria en un cuadrado unitario. El cuadrado también contiene un círculo. La probabilidad de que los puntos se encuentren dentro del círculo es igual al área del círculo, $\pi/4$. El valor de pi se puede estimar a partir del valor de $4R$. R es la proporción de la cantidad de puntos dentro del círculo con respecto al número total de puntos que se encuentran dentro del cuadrado. Mientras más grande sea la muestra de puntos usada, mejor resulta el valor calculado.

```
$ yarn jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar pi 128 20000000
```

Entra vía **SSH** en **hadoop-master1.tartangalh.eus** y ejecuta el comando anterior. Este comando usa 128 mapas con 20 millones de muestras cada uno para calcular el valor de pi. El valor devuelto debería asemejarse a **3.14159263125000000000**. Como referencia, las primeras 10 posiciones decimales de pi son 3,1415926535.

USUARIOS PARA EJEMPLOS DE EJECUCIÓN



LOGINS:

iabd-user¹

iabd-user²

iabd-user³

...

...

iabd-user¹⁰

PASSWORD:

abcd*1234

ACCESO SSH: ssh iabd-user^X@hadoop-master1.tartangalh.eus

SERVICIOS DEL CLÚSTER



Apache Zeppelin



Spark es un motor multi-lenguaje.

Unified engine for large-scale
data analytics

GET STARTED

What is Apache SparkTM?

Apache SparkTM is a multi-language engine for executing large-scale data engineering, data science, and machine learning on single-node machines or clusters.

Ejecuta programas de ingeniería de datos, ciencia de datos y aprendizaje automático en clústeres o en máquinas de un solo nodo.



**Zeppelin es un notebook en
forma de aplicación web**

Apache Zeppelin

Web-based notebook that enables data-driven
interactive data analytics and collaborative documents with SQL, S
more.

**Proporciona el uso y
análisis de datos y documentos
de forma interactiva usando
lenguajes como SQL, Scala,
Python, R ...**

[GET STARTED](#)[DOWNLOAD](#)

EJEMPLOS DE EJECUCIÓN



EJEMPLO 2: Zeppelin Notebook usando Spark

Entra en <http://hadoop-master1.tartangalh.eus:9995/> , abre el notebook Matricula.. y ejecuta sus párrafos en orden.

The image shows the Zeppelin Notebook interface. At the top, there's a blue header with the Zeppelin logo and a dropdown menu showing "Notebook" and "Job". Below the header, the notebook title "Matricula" is displayed. The main area contains a Scala script for processing a CSV file. The script defines a case class for "Matricula", reads the file into a text RDD, filters out header lines, and maps the remaining lines into the case class. It then converts the RDD to a DataFrame and registers it as a temporary table. The output shows the execution of the script, including a warning about deprecation and the creation of the temporary table.

```
val matriculaText = sc.textFile("/user/jmarturi/matricula.csv")

case class Matricula(COD_ALU:Integer , FECHA_NACI_ALU:String , COD_SEXO_ALU:Integer, DES_SEXO_ALU:String , NUM_SOLICITUD_
COD_NIVEL_EDUCATIVO:Integer , DES_NIVEL_EDUCATIVO:String , COD_MOD_IMPARTICION:Integer , DES_MOD_IMP
DES_TURNO:String , CURSO:String , DES_MODELO_LINGUIS:String , DES_TIPO_ACCESO:String , COD_MODULO_MA

// split each line, filter out header (starts with "COD_ALU", and map it into Matricula case class, check if the line has
val matriculas = matriculaText.map(s=>s.split(";")).
  filter(s =>(s.size)>38).
  filter(s=>s(0)!="COD_ALU").
  map(
    s=>Matricula(s(0).toInt,s(1),s(2).toInt,s(3),s(8),s(9),
s(13).toInt,s(14),s(15).toInt,s(16),s(17).toInt,s(18),
s(19),s(20),s(21),s(22),s(26).toInt,s(27),s(34).toInt,s(37)
  )
)

// convert to DataFrame and create temporal table
matriculas.toDF().registerTempTable("matriculas")

warning: there was one deprecation warning (since 2.0.0); for details, enable `:setting -deprecation` or `:replay -depreca
import sqlContext.implicits._
matriculaText: org.apache.spark.rdd.RDD[String] = /user/jmarturi/matricula.csv MapPartitionsRDD[36] at textFile at <console>
defined class Matricula
matriculas: org.apache.spark.rdd.RDD[Matricula] = MapPartitionsRDD[40] at map at <console>:31

Took 2 sec. Last updated by anonymous at May 21 2024, 10:32:24 AM.
```

USUARIOS PARA EJEMPLOS DE EJECUCIÓN



LOGINS:

iabd-user**1**

iabd-user**2**

iabd-user**3**

...

...

iabd-user**10**

PASSWORD:

abcd*1234

URL Zeppelin web UI: <http://hadoop-master1.tartangalh.eus:9995/>

SERVICIOS DEL CLÚSTER



Flink



kafka



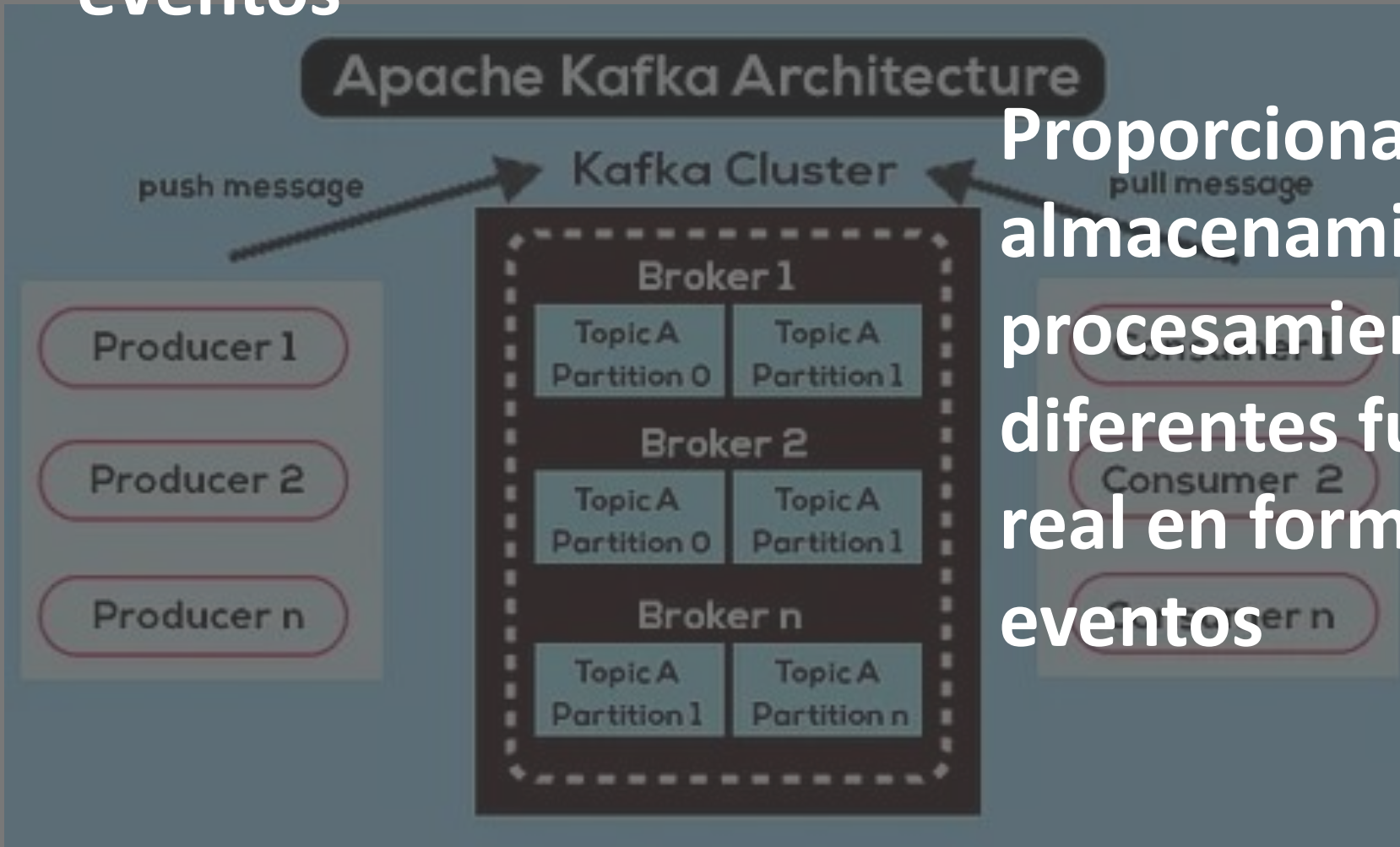
Apache Zeppelin



Solr



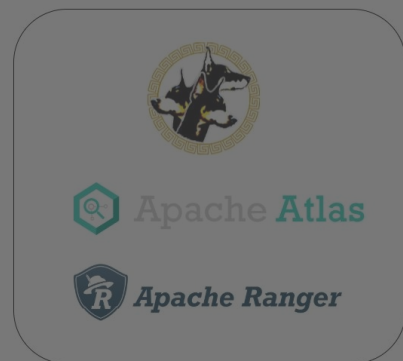
Kafka es una plataforma distribuida de “streaming” de eventos



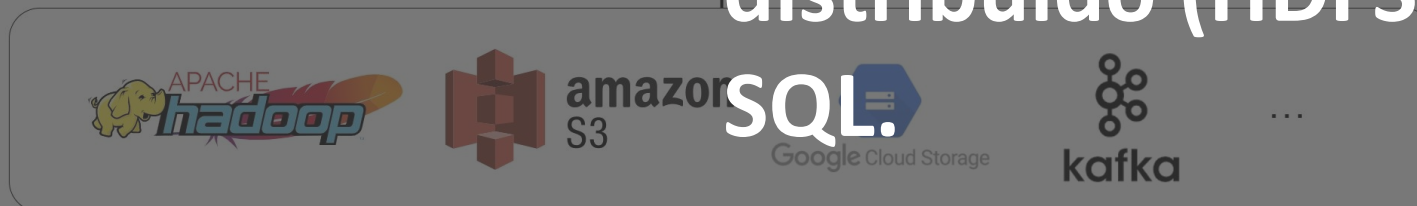
Proporciona captura, almacenamiento y procesamiento de datos de diferentes fuentes en tiempo real en forma de “streams” de eventos

SERVICIOS DEL CLÚSTER

Hive es un sistema de acceso a datos distribuido y tolerante a fallos

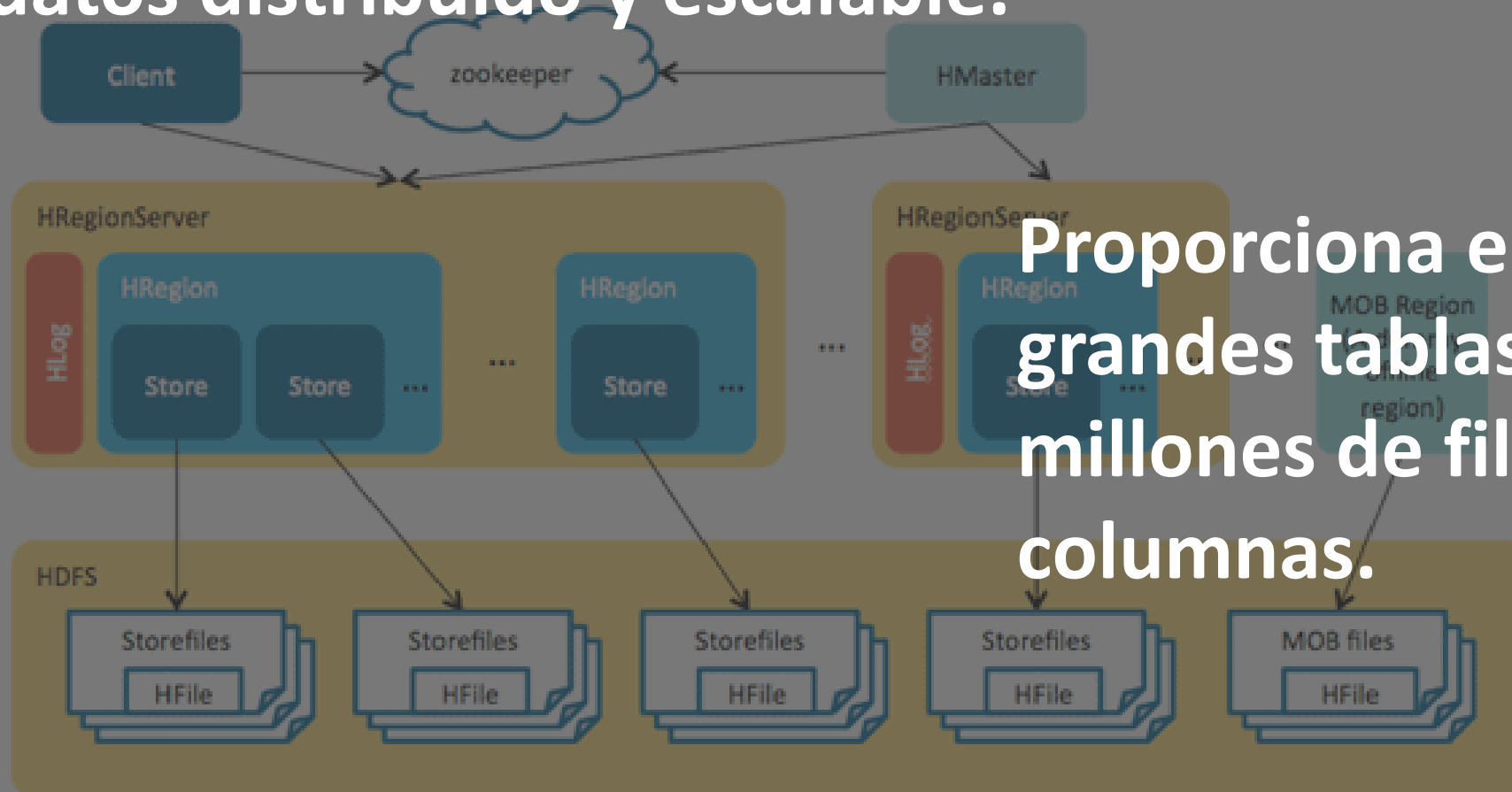


Proporciona análisis a escala masiva y facilita la lectura, escritura y administración de petabytes de datos que residen en el almacenamiento distribuido (HDFS) utilizando SQL.



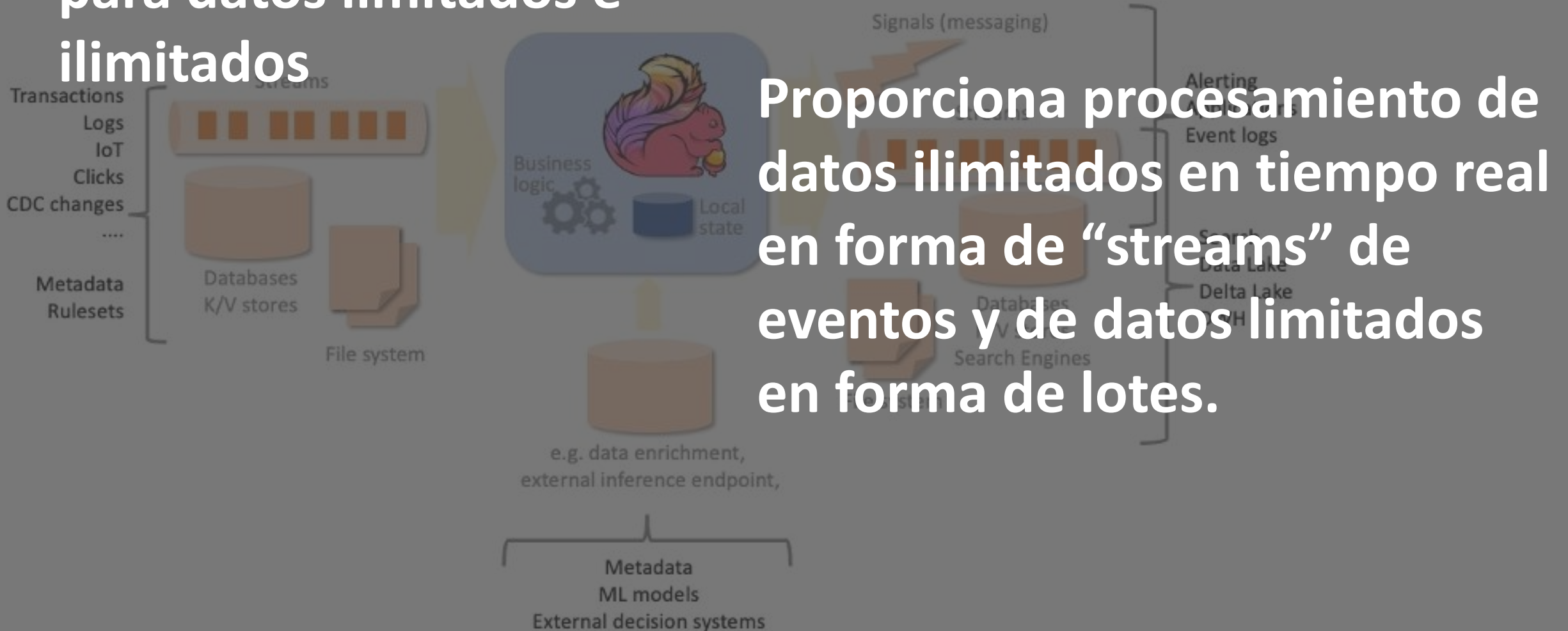


HBase es la base de datos de Hadoop : es un almacén de datos distribuido y escalable.



Proporciona el alojamiento de grandes tablas con miles de millones de filas y millones de columnas.

Flink es un framework y motor de procesamiento distribuido para datos limitados e ilimitados

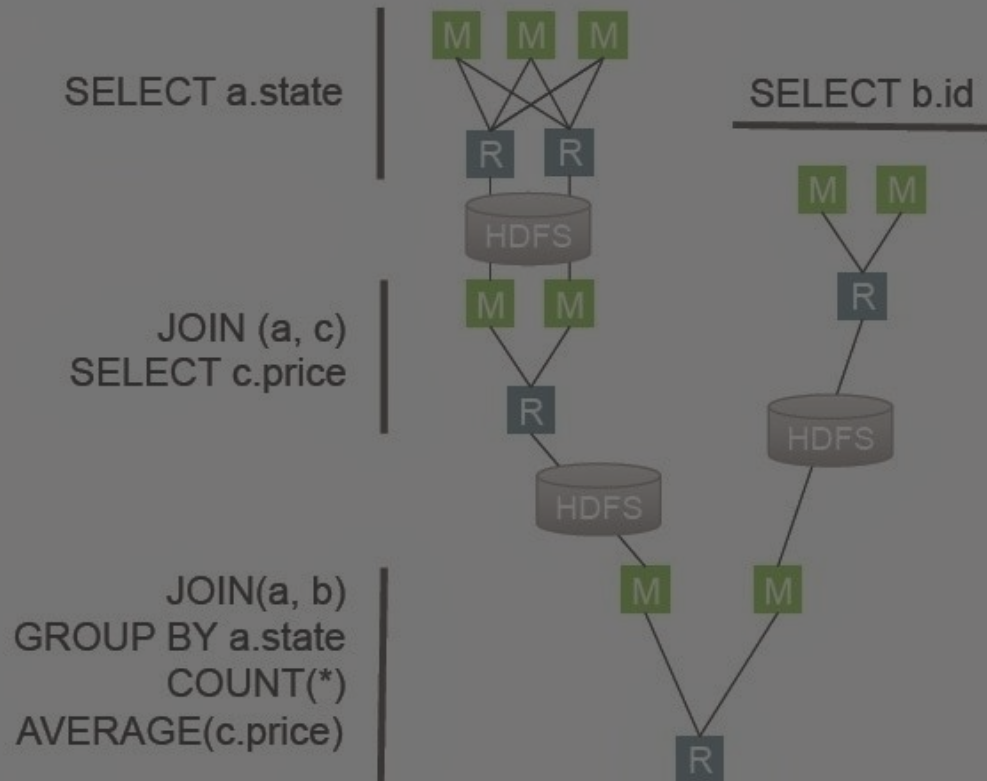


Proporciona procesamiento de datos ilimitados en tiempo real en forma de “streams” de eventos y de datos limitados en forma de lotes.

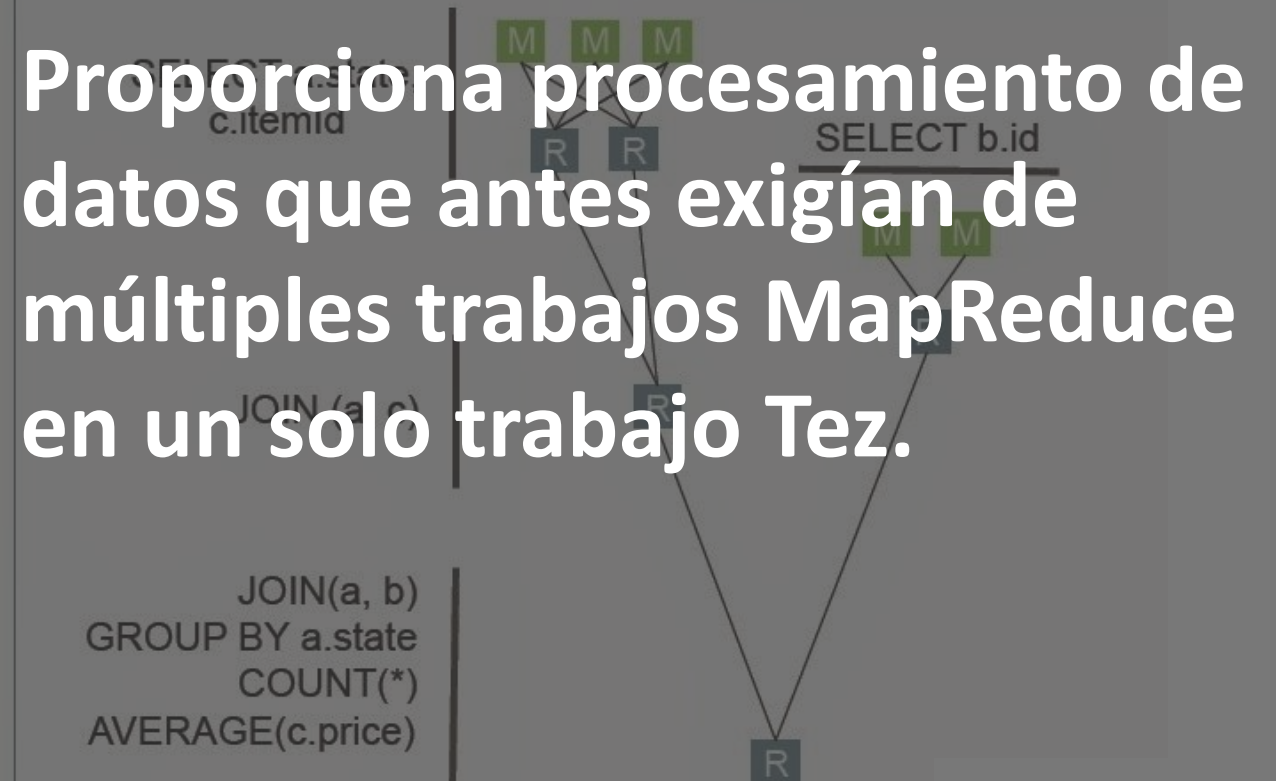
Tez es un framework para ejecución de trabajos distribuidos sobre YARN



Hive – MR



Hive – Tez



Proporciona procesamiento de datos que antes exigían de múltiples trabajos MapReduce en un solo trabajo Tez.

Solr Features

Solr es una plataforma para realizar búsquedas

Proporciona indexación distribuida, replicación y balanceo de carga en las búsquedas



Advanced Full-Text Search Capabilities

Powered by Lucene™, Solr enables powerful matching capabilities including phrases, wildcards, joins, grouping and much more across any data type



Optimized for High Volume Traffic

Solr is proven at extremely large scales the world over



Standards Based Open Interfaces - XML, JSON and HTTP

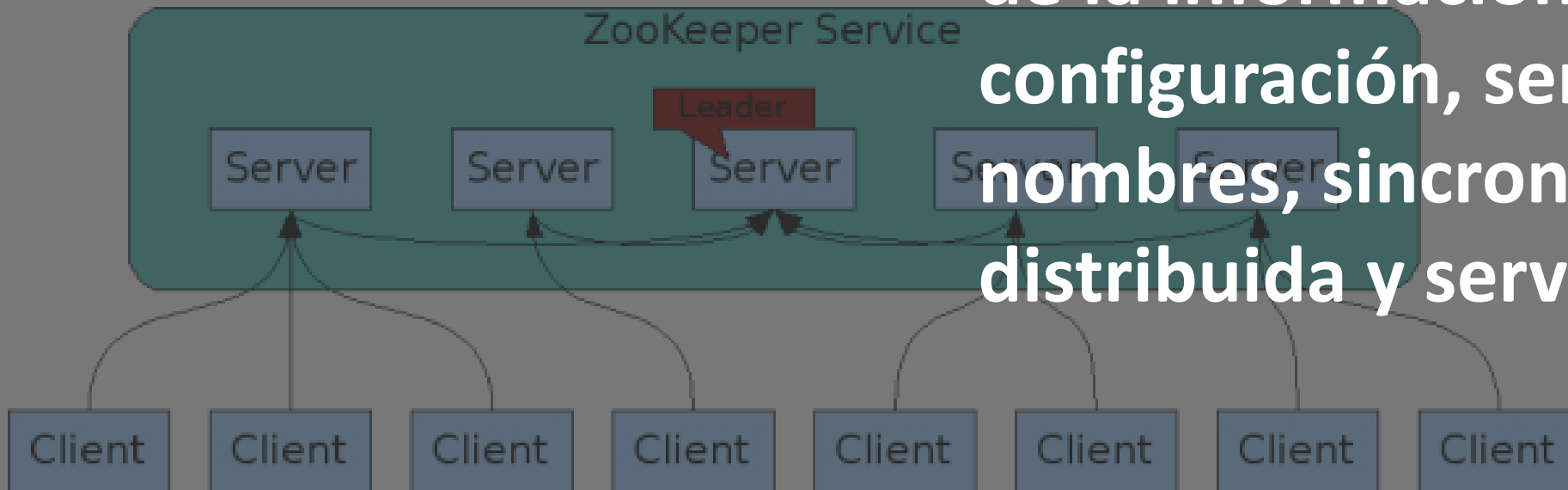
Solr uses the tools you use to make application building a snap



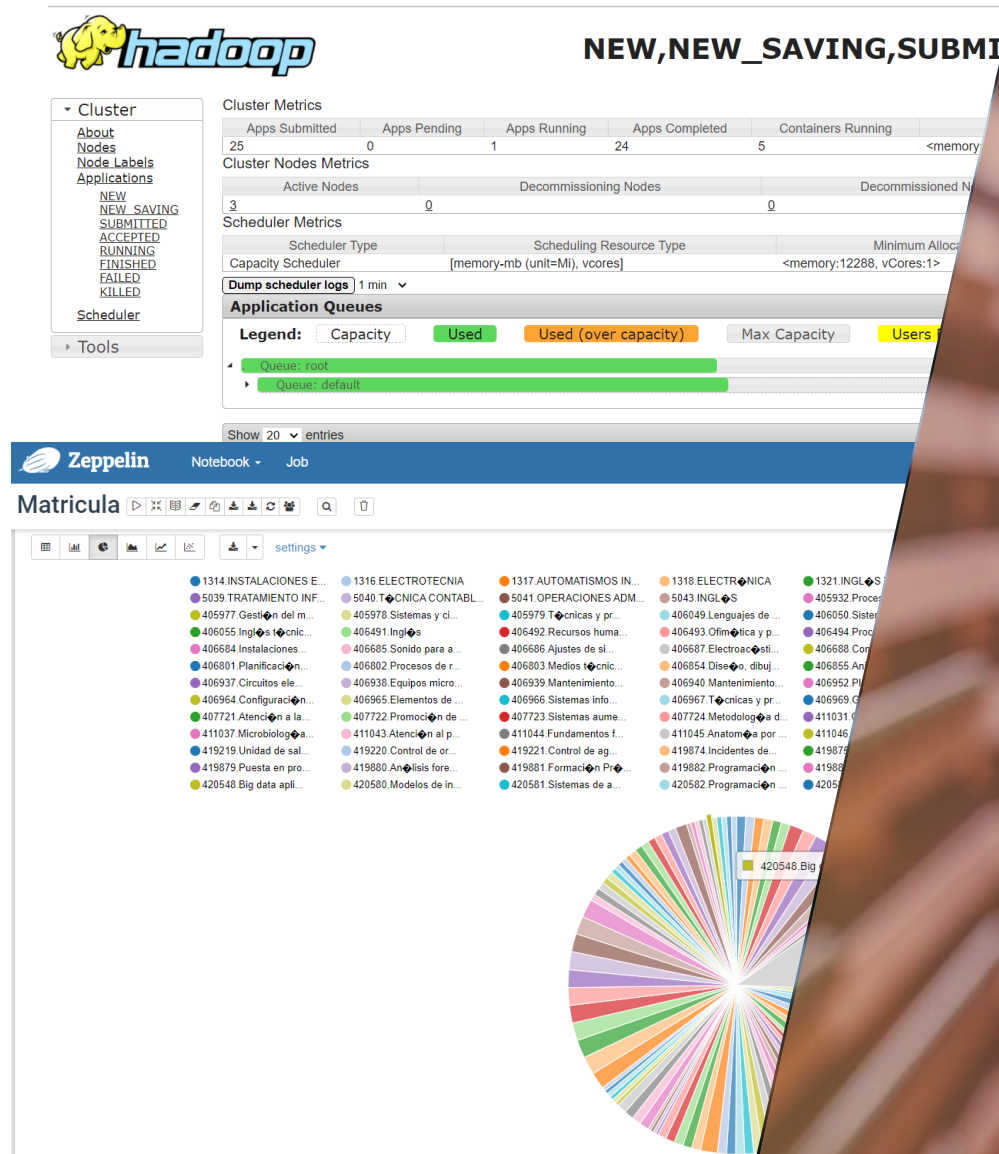
Zookeeper es un servicio de coordinación de aplicaciones distribuidas



Proporciona el mantenimiento de la información de configuración, servicio de nombres, sincronización distribuida y servicios grupales



USOS DEL CLÚSTER



Analizar

Optimizar

Predecir

PROGRESO DEL PROYECTO

TAREAS	ESTADO
Provisionamiento Hardware y Sistema Operativo	
Despliegue del clúster: Instalación del Stack en nodos	
Implementación de la red de alta velocidad para el clúster	
Sintonización del clúster: ajuste de servicios y parámetros	 → 
Validación del clúster: pruebas de ejecución de trabajos	 → 
Diseño de retos/actividades que impliquen el uso del clúster	 → 
Divulgación del proyecto en centros: enlace del blog, charlas, RRSS	 → 

ESKERRIK ASKO GRACIAS



Ainhoa Zugadi Zulueta

Iñigo Pérez Juez

Javier Martín Uría



<https://clusterbigdata.blog.tartanga.eus>



javier.martin@tartanga.eus