



Universidad **Ricardo Palma**

RECTORADO

PROGRAMA DE ESPECIALIZACIÓN EN CIENCIA DE DATOS

Formamos seres humanos para una cultura de paz

BIG DATA APLICADO

SESIÓN 1

Expositores:

David Narváez

Eder Pineda

bigdataplicado@gmail.com

ESTRUCTURA DEL CURSO

Sesión 01: Introducción:

- Pipeline Arquitectura Tradicional
- Pipeline Arquitectura Big Data
- Distribuciones Hadoop
- Infraestructura OnPremise – Cloud

Sesión 02: Fundamentos Hadoop

- Que es Apache Hadoop?
- HDFS
- MapReduce
- Yarn
- Ecosistema Hadoop

Sesión 03: Taller de introducción a Python


Sesión 04: Fundamentos Spark

- Introduction to Apache Spark
- Spark InternalsDriver y Workers
- RDDs, Daframes y Datasets
- RDDS operations
- SparkSQL
- Using SparkSQL Datasets
- Spark-submit for execute on the cluster

Sesión 05: Machine Learning - Spark

- Introducción al Machine Learning
- Comandos principales para el manejo de archivos (entrenamiento, prueba, balanceo)
- Generación de modelos Random forest, Decision tree, Gradient boosted tree, Logistic regression, Multilayer perceptron (neural net), Naive Bayes
- Indicadores principales del modelo.

Sesión 06: Implementación de un caso de estudio utilizando Cloudera CDH

- Serialización
 - Ingesta
 - Procesamiento
 - Ejecución y Monitoreo
 - Analítica
- 

EVALUACIONES

Sesión 01: Introducción:

Sesión 02: Fundamentos Hadoop

EVALUACION 01

Sesión 03: Taller de introducción a Python

Sesión 04: Fundamentos Spark

EVALUACION 02


Sesión 05: Machine Learning - Spark

Sesión 06: Implementación de un caso de estudio utilizando Cloudera CDH

EVALUACION 03

Duración de la Evaluación: 40 minutos





Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation (**Gartner**).

AGENDA

- INTRODUCCIÓN:
 - QUE ES BIG DATA
 - PIPELINE ARQUITECTURA TRADICIONAL
 - PIPELINE ARQUITECTURA BIG DATA
 - DISTRIBUCIONES HADOOP
 - INFRAESTRUCTURA ONPREMISE – CLOUD

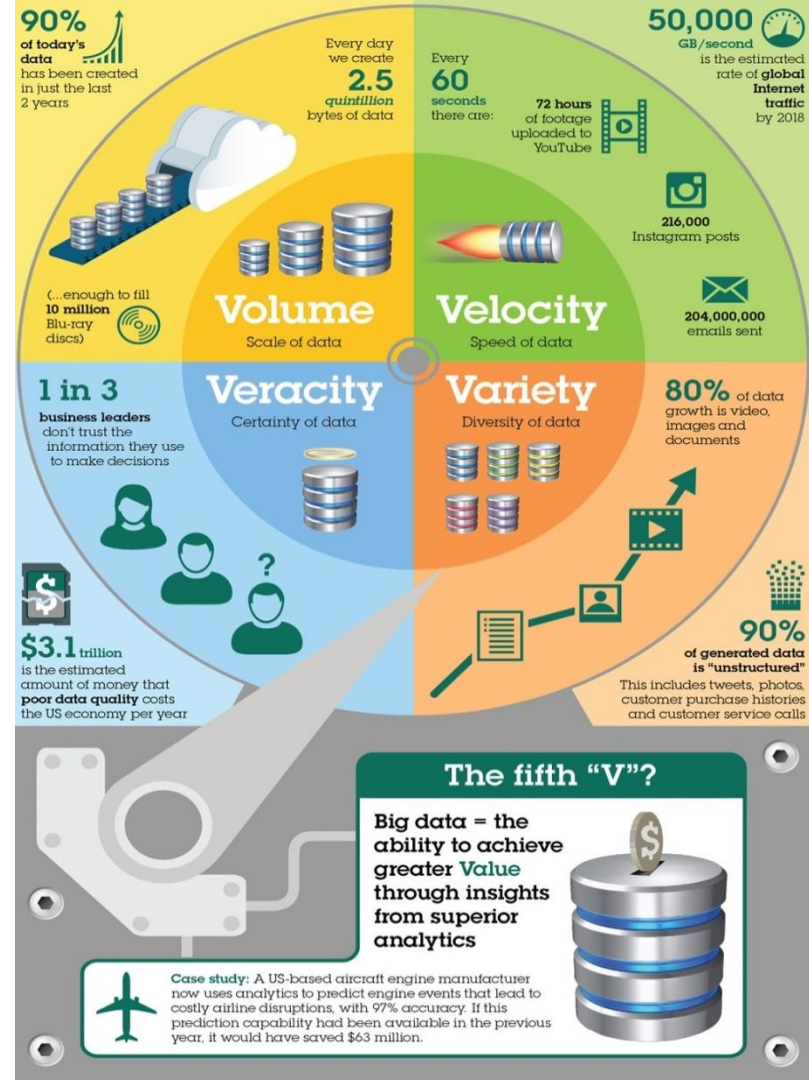
QUE ES BIG DATA?



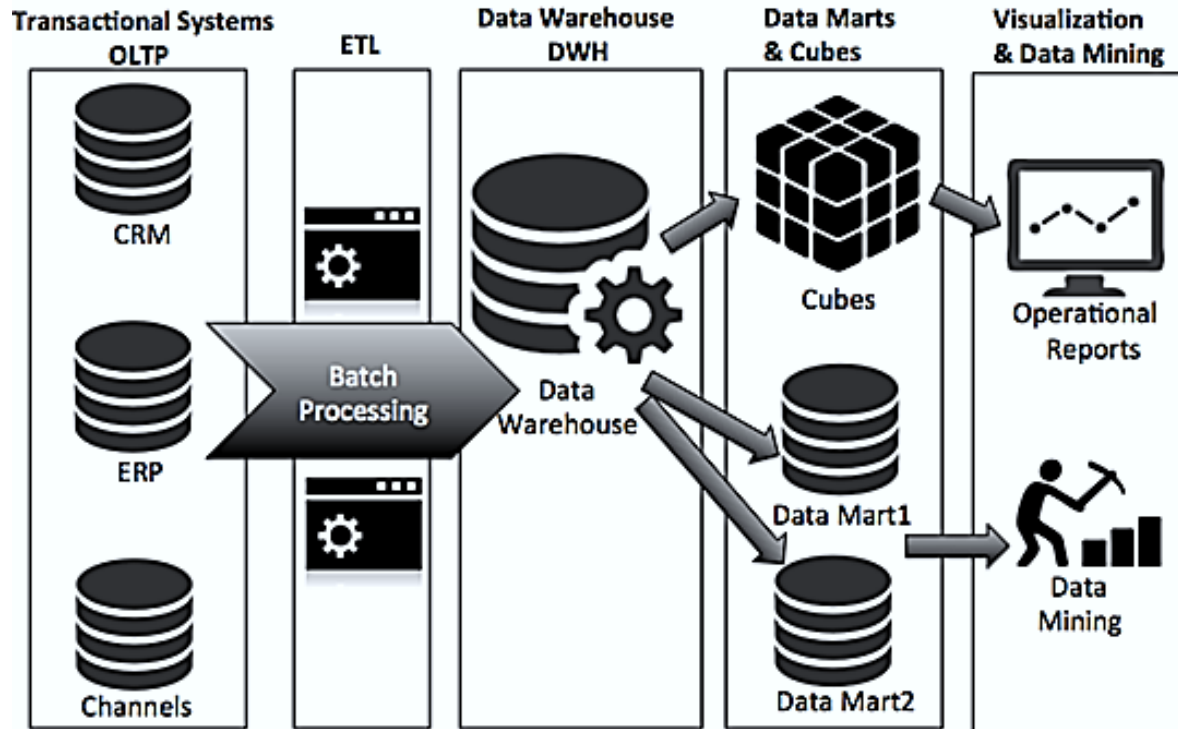
Velocidad
olumen
ariedad
erasidad

Big Data es un conjunto de técnicas y tecnologías para revelar Insights a partir de set de datos diversos, complejos y a gran escala.

<http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data>



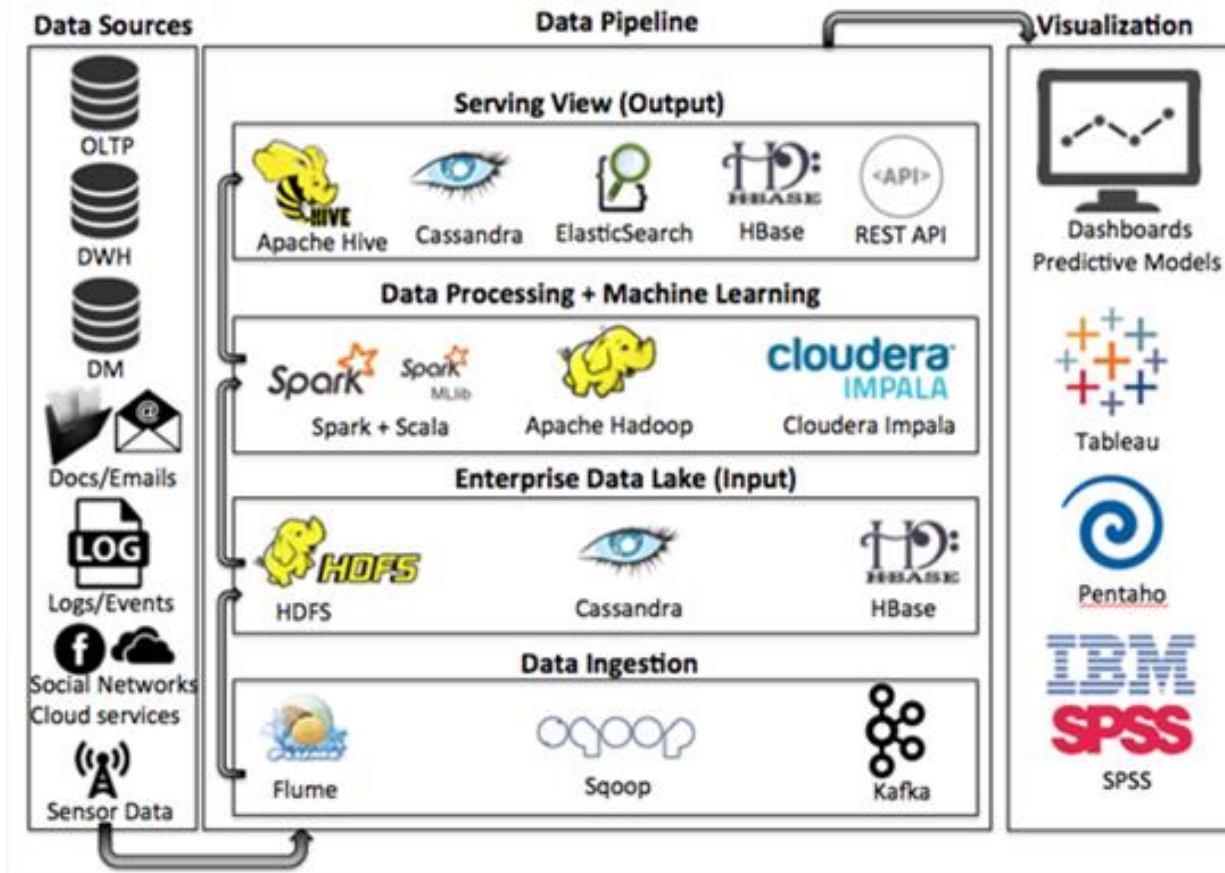
PIPELINE ARQUITECTURA TRADICIONAL



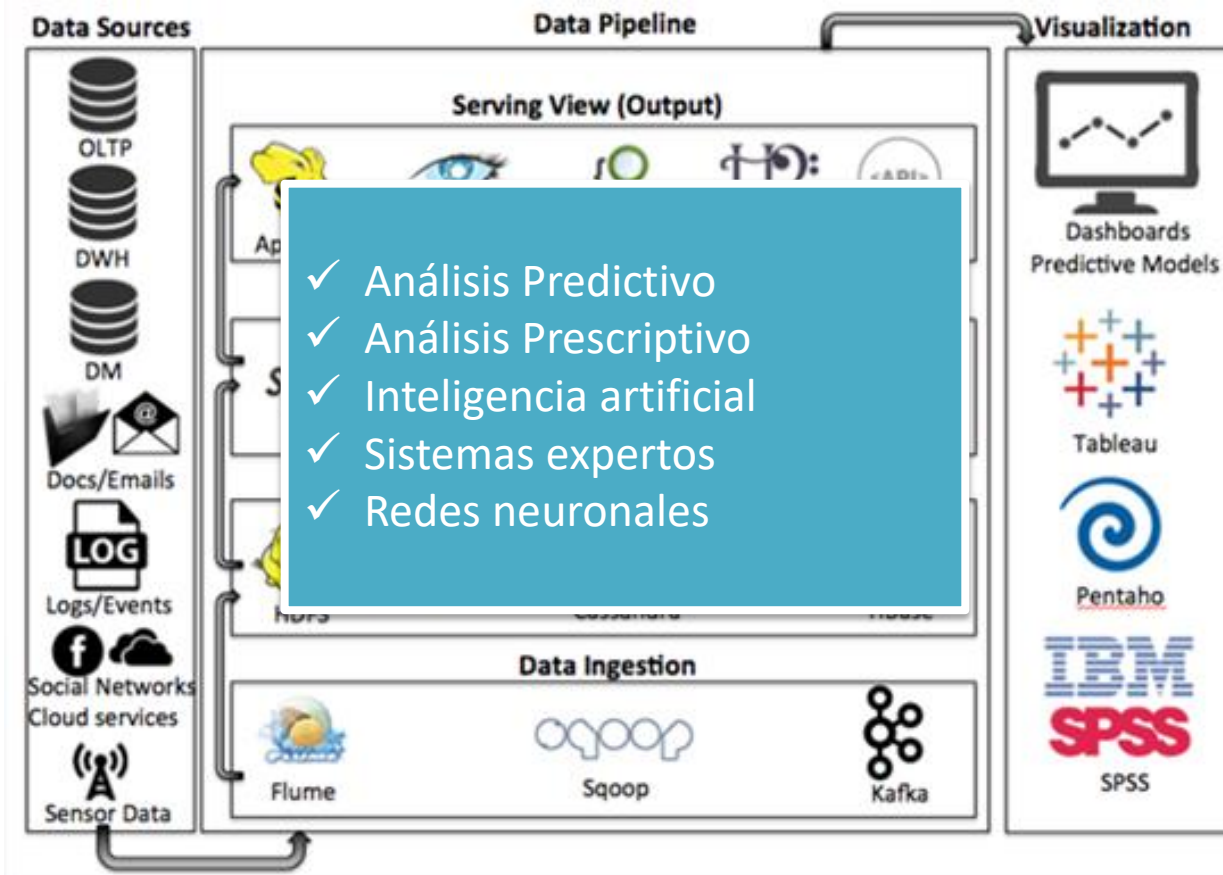
PIPELINE ARQUITECTURA TRADICIONAL



PIPELINE ARQUITECTURA BIG DATA



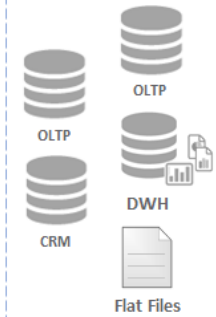
PIPELINE ARQUITECTURA BIG DATA





ARQUITECTURA LAMBDA

DATA SOURCES

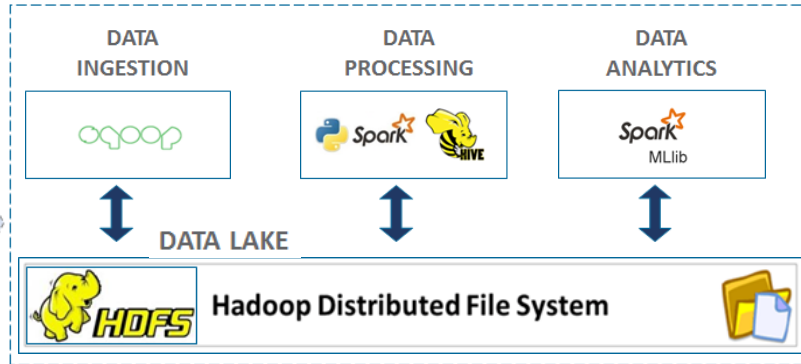


Batch

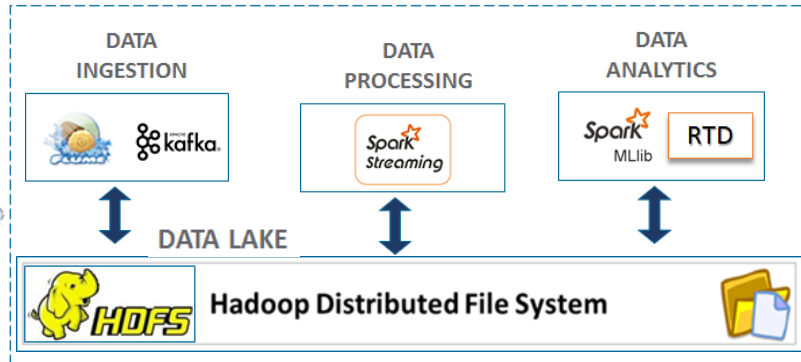


Real Time

BATCH LAYER



SPEED LAYER



SERVING LAYER



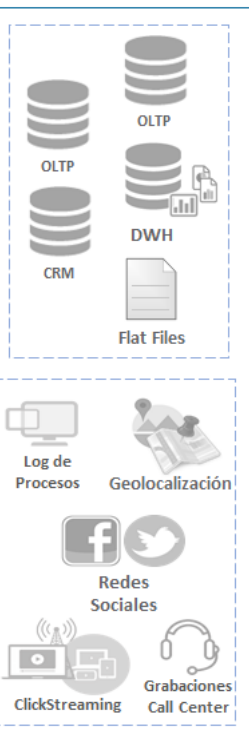
VISUALIZATION



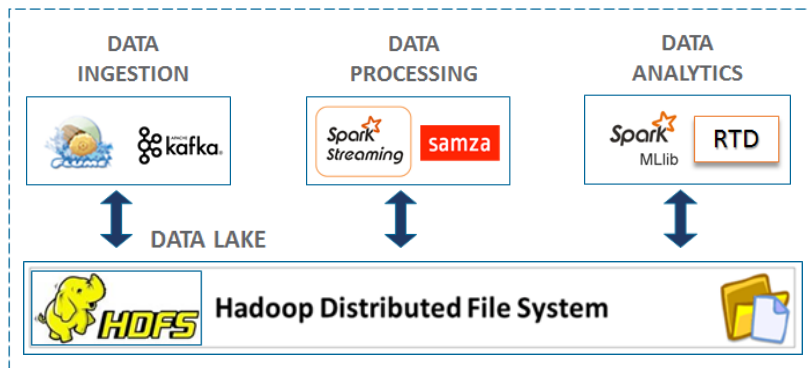


ARQUITECTURA KAPPA

DATA SOURCES



REAL-TIME LAYER

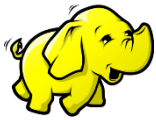


SERVING LAYER



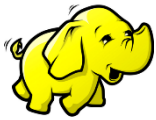
VISUALIZATION





DISTRIBUCIONES HADOOP

Hadoop Distribution	Advantages	Disadvantages
Cloudera Distribution for Hadoop (CDH)	CDH tiene una interfaz fácil de usar con muchas características y herramientas útiles como Cloudera Impala	CDH es comparativamente más lento que MapR Hadoop Distribution
MapR Hadoop Distribution	Es una de las distribuciones de hadoop más rápidas con acceso directo de múltiples nodos	MapR no tiene una buena consola de interfaz como Cloudera
Hortonworks Data Platform (HDP)	Es la única distribución de Hadoop que admite la plataforma de Windows.	La Interfaz de administración de Ambari en HDP es sumamente básica y no tiene muchas funciones avanzadas.



DISTRIBUCIONES HADOOP

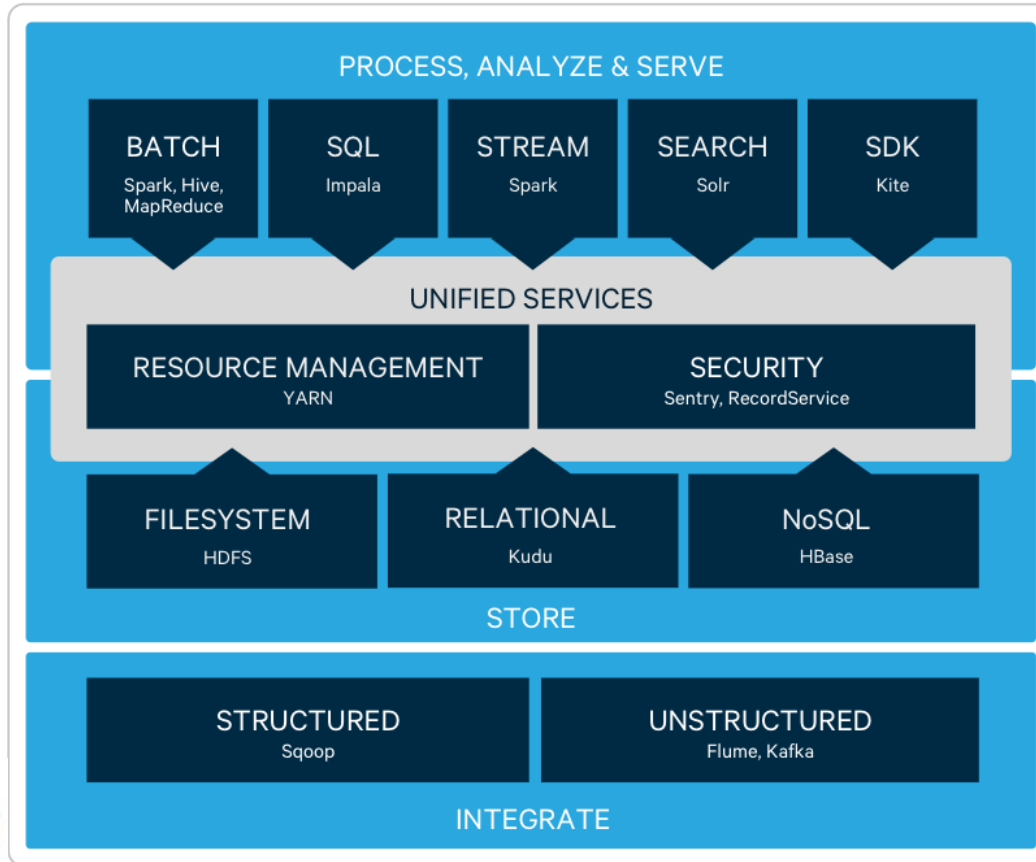
Similitudes

- Los tres -Cloudera, Hortonworks y MapR, están enfocados en Hadoop y todos sus ingresos provienen de ofrecer distribuciones de hadoop listas para la empresa.
- Los tres proveedores ofrecen versiones gratuitas descargables de sus distribuciones, pero MapR y Cloudera también proporcionan adicionalmente distribuciones de hadoop premium pagadas.
- Han |establecido comunidades de apoyo para ayudar a los usuarios con los problemas que enfrentan y también demostraciones, si es necesario.
- Las tres distribuciones de Hadoop han madurado con el paso del tiempo, garantizando estabilidad y seguridad para satisfacer las necesidades del negocio.

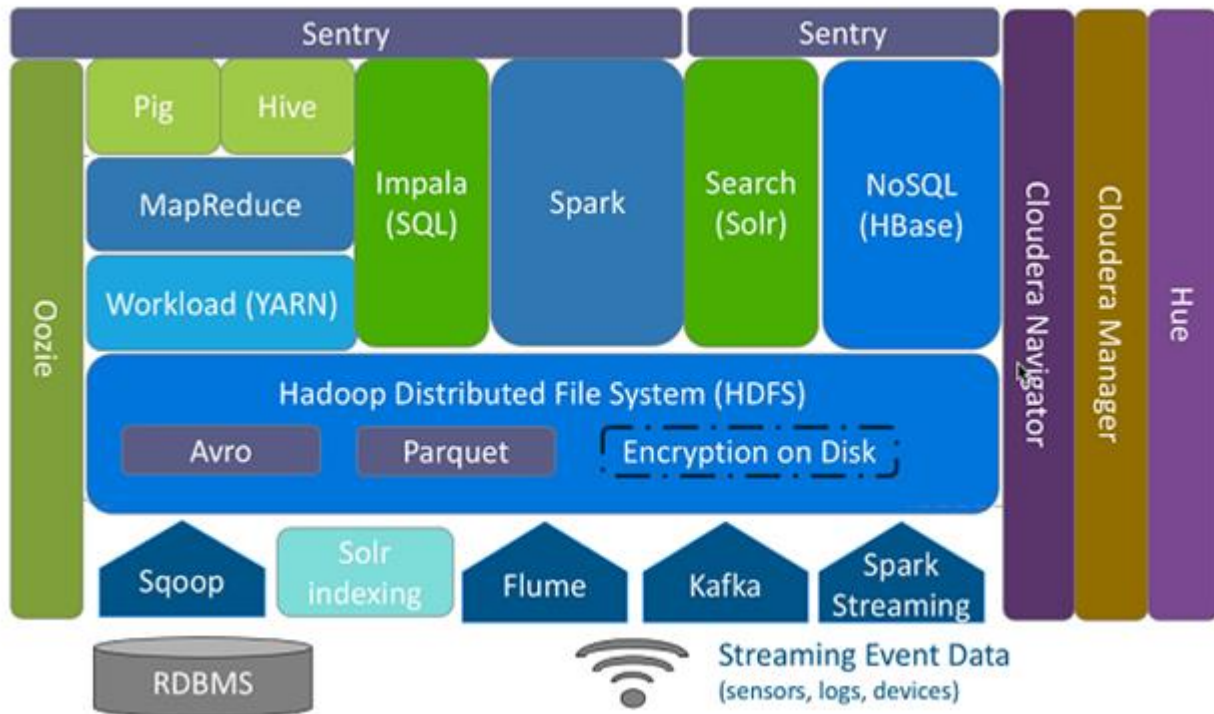
CLOUDERA DISTRIBUTION FOR HADOOP (CDH)

- Cloudera fue la primera distribución comercial de Hadoop.
- La consola de administración – Cloudera Manager, es fácil de usar e implementar. Esta cuenta con una enriquecida interfaz de usuario que muestra toda la información del cluster de una manera clara y organizada.
- La suite de Cloudera Management automatiza el proceso de instalación y también ofrece otros servicios mejorados a los usuarios, mostrando el recuento de nodos en tiempo real, reduciendo el tiempo de implementación, etc.

CLOUDERA DISTRIBUTION FOR HADOOP (CDH)



CLOUDERA DISTRIBUTION FOR HADOOP (CDH)

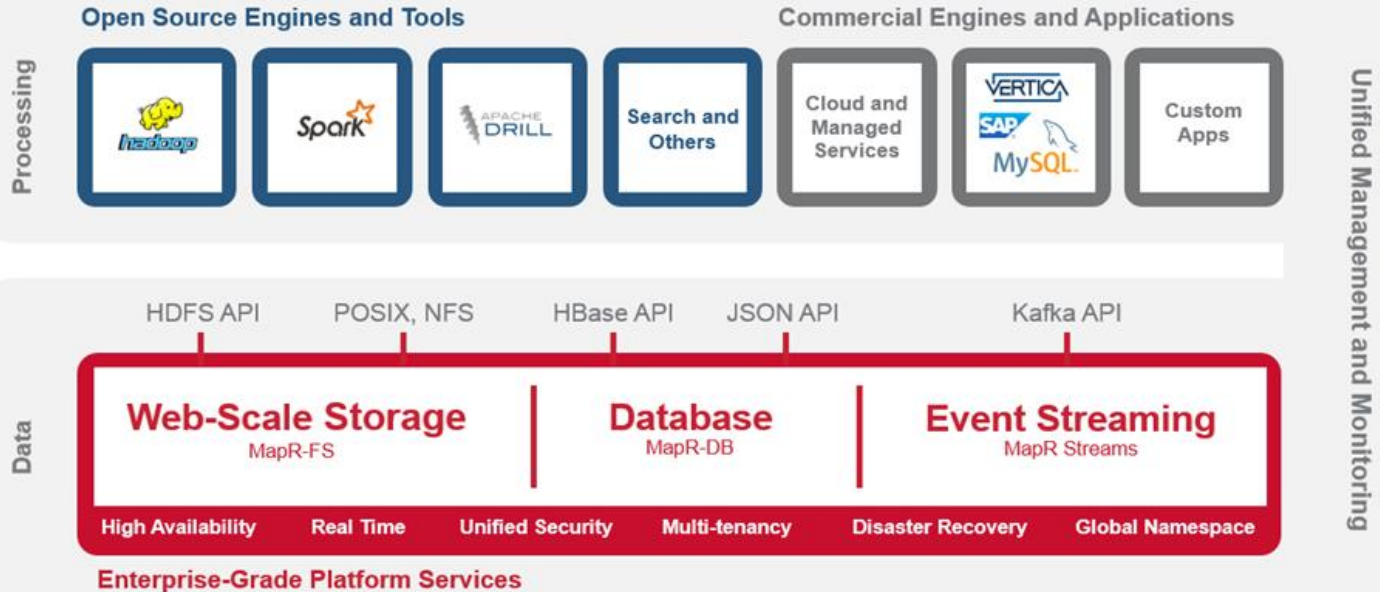


MAPR HADOOP DISTRIBUTION

- Compañías como Cisco, Ancestry.com, Boeing, Google Cloud Platform and Amazon EMR usan MapR Hadoop Distribution para sus servicios de Hadoop.
- A diferencia de Cloudera and Hortonworks, MapR Hadoop Distribution tiene un enfoque más distribuido para almacenar metadatos en los nodos de procesamiento porque depende de un sistema de archivos diferente conocido como MapR File System (MapRFS) y no tiene una arquitectura NameNode.
- MapR hadoop distribution no depende del sistema de archivos de Linux.

MAPR HADOOP DISTRIBUTION

MapR Converged Data Platform

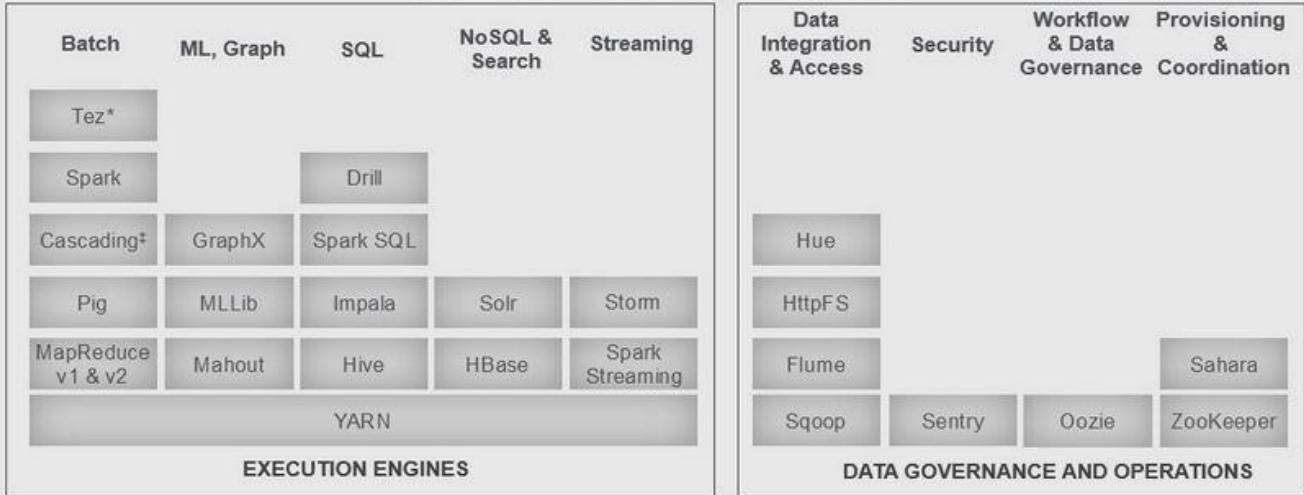


MAPR HADOOP DISTRIBUTION

Management



APACHE HADOOP AND OSS ECOSYSTEM



MapR-FS

Data Platform

MapR-DB

HORTONWORKS DATA PLATFORM (HDP)

- Hortonworks fue fundado por ingenieros de Yahoo.
- Hortonworks es diferente de las otras distribuciones de hadoop, ya que es una plataforma de datos empresariales abierta disponible de forma gratuita.
- Ebay, Samsung Electronics, Bloomberg and Spotify usan HDP.
- Hortonworks fue el primer vendor en proveer una distribución de Hadoop productiva, basada en Hadoop 2.0.
- HDP es la unica distribución de hadoop que soporta plataforma windows. Los usuario pueden desplegar un cluster de hadoop basado en windows en Azure (HDInsight).

HORTONWORKS DATA PLATFORM (HDP)

Hortonworks Data Platform



GOVERNANCE & INTEGRATION

Data Workflow, Lifecycle & Governance

Falcon
Sqoop
Flume
NFS
WebHDFS

DATA ACCESS

Batch
Map
Reduce

Script
Pig

SQL
Hive/Tez
HCatalog

NoSQL
HBase
Accumulo

Stream
Storm

Others
In-Memory
Analytics
ISV Engines

YARN : Data Operating System

HDFS (Hadoop Distributed File System)

DATA MANAGEMENT

SECURITY

Authentication Authorization Accounting Data Protection

Storage: HDFS
Resources: YARN
Access: Hive, ...
Pipeline: Falcon
Cluster: Knox

OPERATIONS

Provision, Manage & Monitor

Ambari
Zookeeper

Scheduling

Oozie

Linux

Windows

On Premise

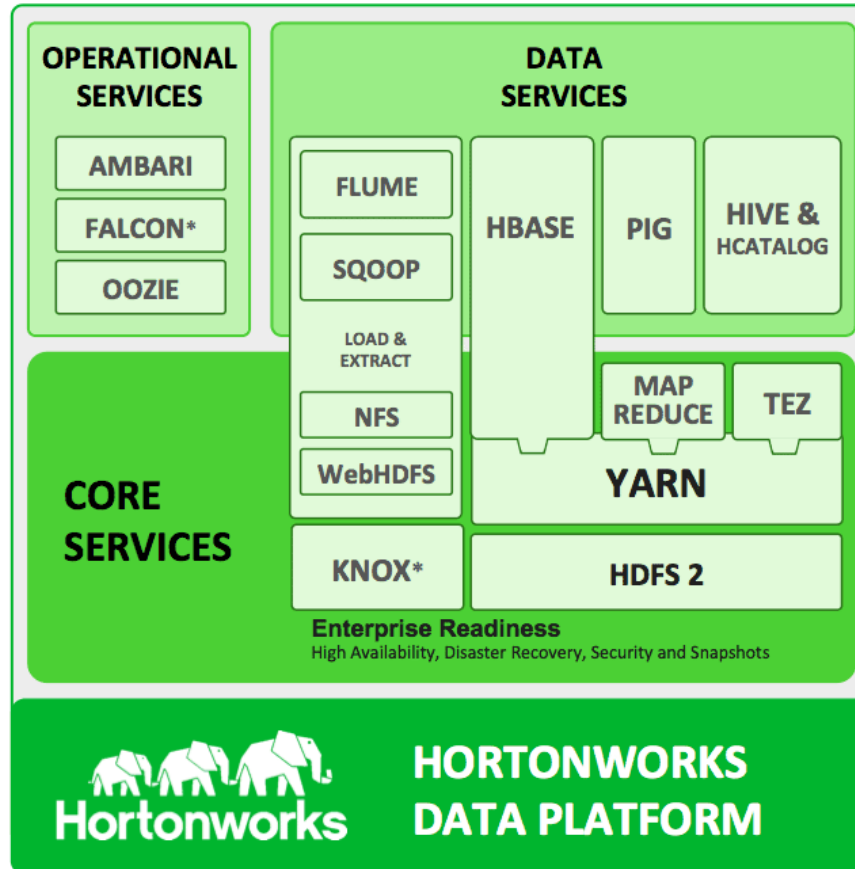
Virtualize

Cloud/Hosted

Appliance

Commodity HW

HORTONWORKS DATA PLATFORM (HDP)



INFRAESTRUCTURA ONPREMISE – CLOUD

Appliance o Sistemas Integrados

Hardware Commodity

Cloud

VENTAJAS

- ✓ El soporte técnico viene de una sola fuente.
- ✓ Interfaces de administración de sistemas Integrado para Gestión de Hardware, Software, Administración de Archivos de Datos, Monitoreo de Clusters y Nodos y Seguridad.

- ✓ Costo de Licencias menores o uso de Open Source.
- ✓ Bajo costo de infraestructura

- ✓ Escalabilidad de la infraestructura
- ✓ Actualizaciones automáticas
- ✓ Elimina la administración y soporte de la infraestructura por parte de la empresa

- ✗ La implementación de la solución suele ser mucho mas costosa.
- ✗ Requiere personal especializado para desplegar, optimizar, y realizar afinamientos a la plataforma.

- ✗ Commodity cluster requiere un tiempo considerable y personal altamente especializado para desplegar, optimizar, y realizar afinamientos a la plataforma.
- ✗ Alto riesgo de presentar defectos por mala configuración.

- ✗ La carga de la data está sujeta al performance de la red, pudiendo demandar mayor tiempo y costo al manejar grandes volúmenes de información.
- ✗ Sujeto a políticas y leyes del país donde se realiza el Hosting.

DESVENTAJAS

INFRAESTRUCTURA ONPREMISE – CLOUD

Appliance o Sistemas Integrados

- ✓ Oracle Big Data Appliance
- ✓ IBM BigInsights
- ✓ Teradata Aster Big Analytics Appliance

Hardware Commodity

- ✓ HPE ProLiant DL380
- ✓ Apollo 4200
- * ***Soporta Virtualización***


Cloud


- ✓ Elastic Map Reduce (AWS)
- ✓ HDInsight (Azure)
- ✓ Cloud Dataproc




INFRASTRUCTURA ONPREMISE – CLOUD

Cluster Big Data - AWS



Services ▾ Resource Groups ▾ 

 BIGDATAPERU ▾ N. Virginia ▾ Support ▾

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group 7. Review

Step 1: Choose an Amazon Machine Image (AMI)

[Cancel and Exit](#)

An AMI is a template that contains the software configuration (operating system, application server, and applications) required to launch your instance. You can select an AMI provided by AWS, our user community, or the AWS Marketplace; or you can select one of your own AMIs.


Quick Start


My AMIs

AWS Marketplace

Community AMIs


☐ Free tier only ⓘ


**Amazon Linux AMI 2017.09.1 (HVM), SSD Volume Type** - ami-1853ac65

**Amazon Linux**
Free tier eligible

The Amazon Linux AMI is an EBS-backed, AWS-supported image. The default image includes AWS command line tools, Python, Ruby, Perl, and Java. The repositories include Docker, PHP, MySQL, PostgreSQL, and other packages.


Root device type: ebs Virtualization type: hvm ENA Enabled: Yes


**Amazon Linux 2 LTS Candidate AMI 2017.12.0 (HVM), SSD Volume Type** - ami-428aa838

**Amazon Linux**
Free tier eligible

Amazon Linux 2 is the next generation of Amazon Linux. It includes the latest LTS kernel (4.9) tuned for enhanced performance on Amazon EC2, systemd support, newer versions of glibc, gcc and binutils, and an additional set of core packages for performance and security improvements.

Root device type: ebs Virtualization type: hvm ENA Enabled: Yes

**Red Hat Enterprise Linux 7.4 (HVM), SSD Volume Type** - ami-26ebbc5c



**Red Hat**


Red Hat Enterprise Linux version 7.4 (HVM). EBS General Purpose (SSD) Volume Type

1 to 36 of 36 AMIs

INFRASTRUCTURA ONPREMISE – CLOUD


Cluster Big Data - AWS

 Services ▾ Resource Groups ▾ 


 BIGDATAPERU ▾ N. Virginia ▾ Support ▾


[1. Choose AMI](#) [2. Choose Instance Type](#) [3. Configure Instance](#) [4. Add Storage](#) [5. Add Tags](#) [6. Configure Security Group](#) [7. Review](#)


Step 3: Configure Instance Details

No default VPC found. Select another VPC, or [create a new default VPC](#). 


Configure the instance to suit your requirements. You can launch multiple instances from the same AMI, request Spot instances to take advantage of the lower pricing, assign an access management role to the instance, and more.


Number of instances 

[Launch into Auto Scaling Group](#) 


Purchasing option 

☐ Request Spot instances

Network 


 [Create new VPC](#)


No default VPC found. [Create a new default VPC](#).


Subnet 

[Create new subnet](#)

250 IP Addresses available

Auto-assign Public IP 

IAM role 

 [Create new IAM role](#)

[Cancel](#) [Previous](#) [Review and Launch](#) [Next: Add Storage](#)

INFRASTRUCTURA ONPREMISE – CLOUD

Cluster Big Data - AWS



- 1. Choose AMI
- 2. Choose Instance Type
- 3. Configure Instance
- 4. Add Storage
- 5. Add Tags
- 6. Configure Security Group
- 7. Review

Step 4: Add Storage

Your instance will be launched with the following storage device settings. You can attach additional EBS volumes and instance store volumes to your instance, or edit the settings of the root volume. You can also attach additional EBS volumes after launching an instance, but not instance store volumes. [Learn more](#) about storage options in Amazon EC2.


Volume Type <small>i</small>	Device <small>i</small>	Snapshot <small>i</small>	Size (GiB) <small>i</small>	Volume Type <small>i</small>	IOPS <small>i</small>	Throughput (MB/s) <small>i</small>	Delete on Termination <small>i</small>	Encrypted <small>i</small>
Root	/dev/xvda	snap-01d62e4cbe7d0ddd0	<input type="text" value="8"/>	General Purpose SSD (GP2) <small>v</small>	100 / 3000	N/A	<input checked="" type="checkbox"/>	Not Encrypted
<small>EBS</small> <small>v</small>	<small>/dev/sdb</small> <small>v</small>	<small>Search (case-insensit</small>	<input type="text" value="8"/>	General Purpose SSD (GP2) <small>v</small>	100 / 3000	N/A	<input type="checkbox"/>	<small>Not Encrypt</small> <small>v</small> <small>x</small>
<small>Add New Volume</small>								


Free tier eligible customers can get up to 30 GB of EBS General Purpose (SSD) or Magnetic storage. [Learn more](#) about free usage tier eligibility and usage restrictions.

- Cancel
- Previous
- Review and Launch
- Next: Add Tags

INFRASTRUCTURA ONPREMISE – CLOUD

Elastic Map Reduce - AWS

Services ▾Resource Groups ▾★

BIGDATAPERU ▾N. Virginia ▾Support ▾

Amazon EMR

- Clusters
- Security configurations
- VPC subnets
- Events
- Help

Welcome to Amazon Elastic MapReduce


Amazon Elastic MapReduce (Amazon EMR) is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data.

You do not appear to have any clusters. Create one now:

[Create cluster](#)

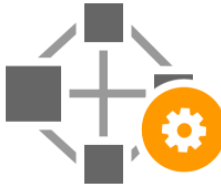
How Elastic MapReduce Works

Upload




Upload your data and processing application to S3

Create



Configure and create your cluster by specifying data inputs, outputs, cluster

Monitor



Monitor the health and progress of your cluster. Retrieve the output in S3

Additional Information

More about Elastic MapReduce

- [EMR overview](#)
- [FAQs](#)
- [Pricing](#)

More Help Using Elastic MapReduce

- [Forum](#)
- [Documentation](#)
- [Developer Guide](#)
- [API Reference](#)
- [EMR on GitHub](#)
- [Help portal](#)

INFRASTRUCTURA ONPREMISE – CLOUD

Elastic Map Reduce - AWS



Services ▾

Resource Groups ▾



BIGDATAPERU ▾

N. Virginia ▾

Support ▾

Create Cluster - Quick Options [Go to advanced options](#)

General Configuration

Cluster name

☒ Logging ⓘ

S3 folder

Launch mode ☒ Cluster ⓘ ☐ Step execution ⓘ

Software configuration

Release ⓘ

- Applications
- ☒ Core Hadoop: Hadoop 2.8.3 with Ganglia 3.7.2, Hive 2.3.2, Hue 4.1.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.8.4
 - ☐ HBase: HBase 1.4.2 with Ganglia 3.7.2, Hadoop 2.8.3, Hive 2.3.2, Hue 4.1.0, Phoenix 4.13.0, and ZooKeeper 3.4.10
 - ☐ Presto: Presto 0.194 with Hadoop 2.8.3 HDFS and Hive 2.3.2 Metastore
 - ☐ Spark: Spark 2.3.0 on Hadoop 2.8.3 YARN with Ganglia 3.7.2 and Zeppelin 0.7.3