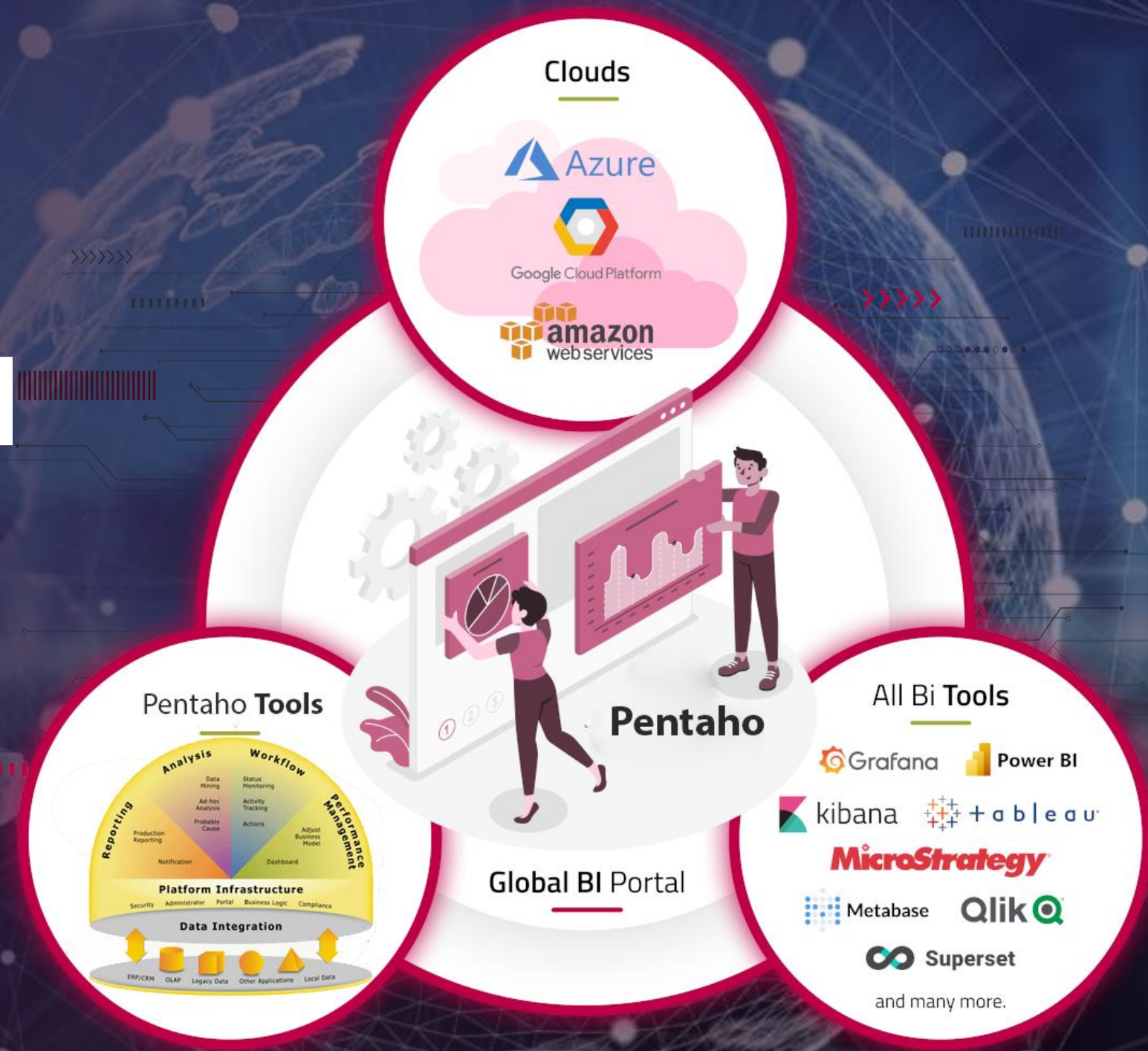




C  
U  
R  
S  
O

# Especialización en Big Data Multi-Cloud & Machine Learning

Curso 100% en español con más de 30 laboratorios de casos reales y doble certificación en Big Data & Machine Learning

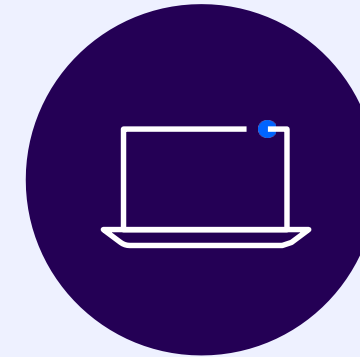




Inicio:  
**18 DE MARZO**



Finalización:  
**15 DE ABRIL**



**60 HORAS**  
académicas



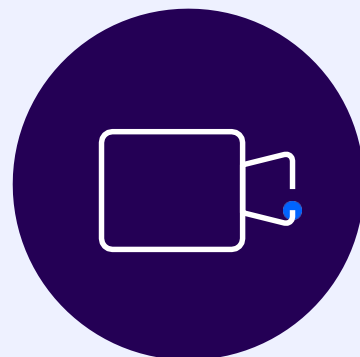
Lunes y viernes  
**De 07:00 pm a  
11:00 pm (GMT-5)**



Soporte  
**TÉCNICO**



Plataforma  
**DIGITAL**



Aplicativo  
**Google Meet**



**CERTIFICACIÓN  
POR NIVEL**

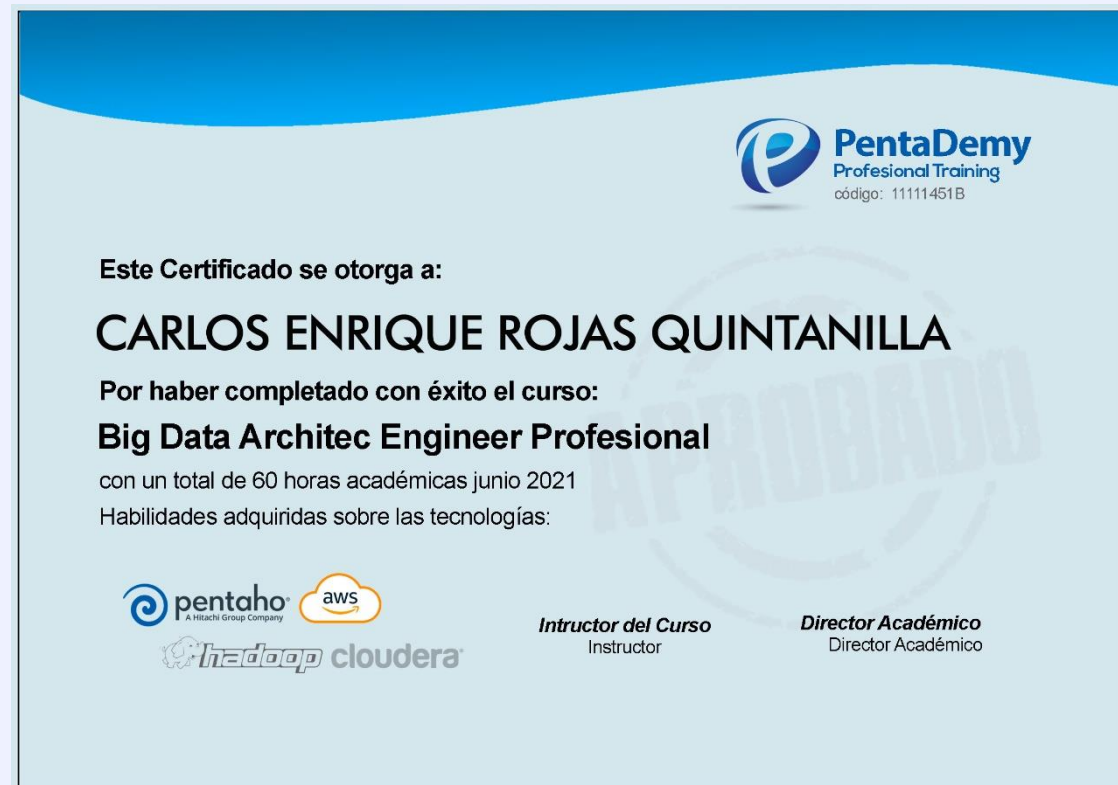


**DIPLOMAS  
DIGITALES DEL  
PROGRAMA**

## Doble Certificación



- ✓ **BIG DATA ARCHITEC ENGINEER PROFESIONAL**
- ✓ **BIG DATA ARCHITEC PROFESIONAL**



## Certificado validez internacional

Nuestros certificados tiene validez en todos los países de Latinoamérica (a excepción de Brasil), código QR y validez en LinkedIn con lo cual podrás compartir tu certificado en





## RESUMEN

El equipo académico de **PentaDemy** cuenta con profesionales que se dedican a desarrollar proyectos TIC en empresas líderes de diversos sectores económicos, enfrentándose a todo tipo de retos. Gracias a esta experiencia, **PentaDemy** propone diferentes cursos que ayudarán a los expertos a incrementar sus skills profesionales permitiéndoles así, abordar proyectos de Big , ya sean **Real Time**, **near Real Time** o procesamiento **Batch**, domina desde la ingestión de datos, el procesamiento, el delivery y el descubrimiento con la analítica.

## OBJETIVO

Formar profesionales que deseen aumentar sus oportunidades laborales y enriquecer su perfil profesional con un elemento diferenciador y de gran demanda actualmente, como lo es el uso de los servicios Cloud y su aplicación al Machine Learning.

## METODOLOGÍA

- Exposición teórica de los temas
- Desarrollo de casos prácticos
- Acceso a las clases grabadas
- Acceso al material exclusivo
- Acceso a una Máquina Virtual con Clusters reales de Big Data

## REQUISITOS

- Conocimientos básicos de SQL
- Portar una laptop personal para las clases de mínimo 1GB de RAM para el uso de Clúster de 128 GB en la nube

## TECNOLOGÍAS

- AWS – Amazon Web Services
- GCP – Google Cloud Platform
- Cloudera Data Platform (CDP)
- Hortonwork Data Platform (HDP)
- Hortonwork Data Flow (HDF)
- Apache Hadoop
- Apache Kafka
- Apache Spark, Hbase y Hive
- Elastic Stack: Kibana, ElasticSearch
- Sqoop. Linux.
- Impala, Python
- MongoDB, Cassandra



# PLATAFORMA MODERNA DE APRENDIZAJE | E-LEARNING

## Big Data & Data Visualization con Pentaho

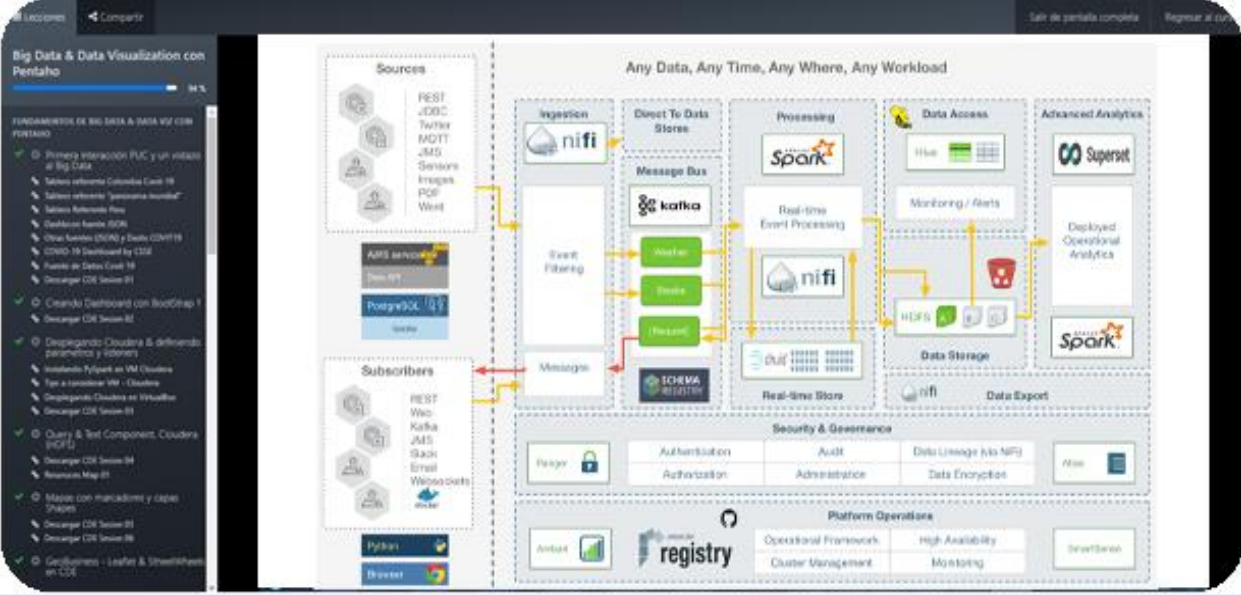
El curso de capacitación "Big Data & Data Viz con Pentaho", presenta un enfoque práctico de los conceptos clave, la arquitectura, herramientas y tecnologías de Big Data y su aplicación en el negocio, potenciando su aprendizaje de forma aplicada con el uso de la suite de analítica Pentaho BA (Business Analytics).

Abarca desde los fundamentos de Hadoop, manejo de los lenguajes de programación python, R, uso de spark y streaming, aprenderás los fundamentos en base a laboratorios de machine learning, así como a realizar la ingesta y tratamiento de datos con Pentaho.

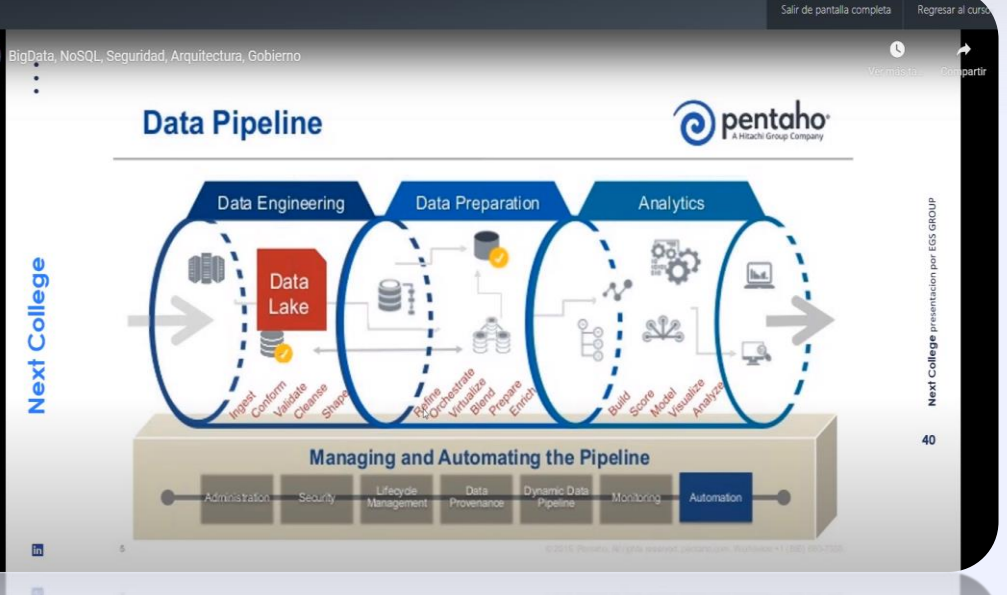
Curso

- Fundamentos de Big Data & Data Viz con Pentaho
- Primera interacción PUC y un vistazo al Big Data
- Creando Dashboard con Bootstrap 1
- Desplegando Cloudera & definiendo parametros y listeners
- Query & Text Component, Cloudera (HDFS)
- Mapas con marcadores y capas Shapes
- GeoBusiness - Leaflet & SteeplWheels en CDE
- Geobusiness - Leaflet & SteeplWheels en CDE
- Mapas con marcadores y capas shapes
- Query & Text Component, Cloudera (HDFS)
- Mapas con marcadores y capas shapes

MUNICIPIO	COUNT	FIRST_IDDP	HECTARES
AMAZONAS	84	01	393046.567
ANCASH	166	02	3596224.6
APURIMAC	80	03	2111415.17
AREQUIPA	109	04	6325588.935
AYACUCHO	111	05	4350381.783
CAJAMARCA	127	06	3304465.549
CALLAO	6	07	14140.954
CUSCO	108	08	7207614.24
HUANCAVELICA	94	09	2206503.876
HUANUCO	76	10	3720052.603
ICA	43	11	2108076.66
JUNIN	123	12	4399729.222
LA LIBERTAD	83	13	2529596.876
LAMBAYEQUE	38	14	1434230.801
LIMA	171	15	3499993.431
LORETO	51	16	37511598.865
MADRE DE DIOS	11	17	8904586.569
MOQUEGUA	20	18	1580730.977
PASCO	28	19	2411394.892



- MySQL Workbench con Windows
- Descargar Instalador de MySQL
- Descargar MySQL Schema Workbench
- Descargar Microsoft Visual C++ 2019
- Postman Restful Dash. & Hue. Hive e impala VM Cloudera
- Compilando e Instalando Hue en Nodo1
- Importación/Exportación con Sqoop (hadoop-nodo1)
- Lab: sqoop, import/export, postgres/mysql/mongodb





Impartido por:

**Ing. Pablo Valdivia**

Chief Data Architect at GIS

Chief Executive Officer in EGS GROUP

<http://www.egs.pe>

## ACERCA DEL EXPOSITOR:

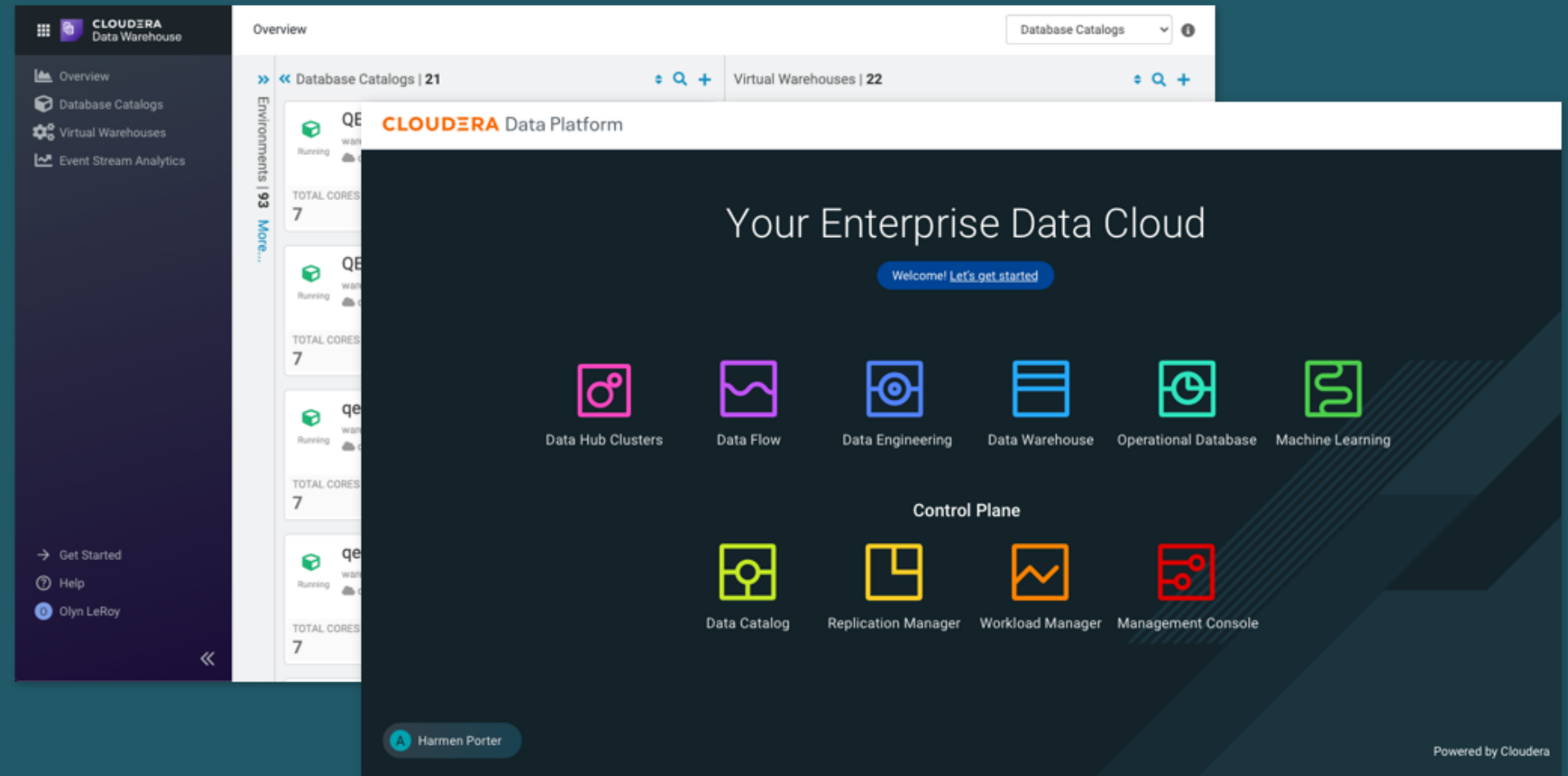
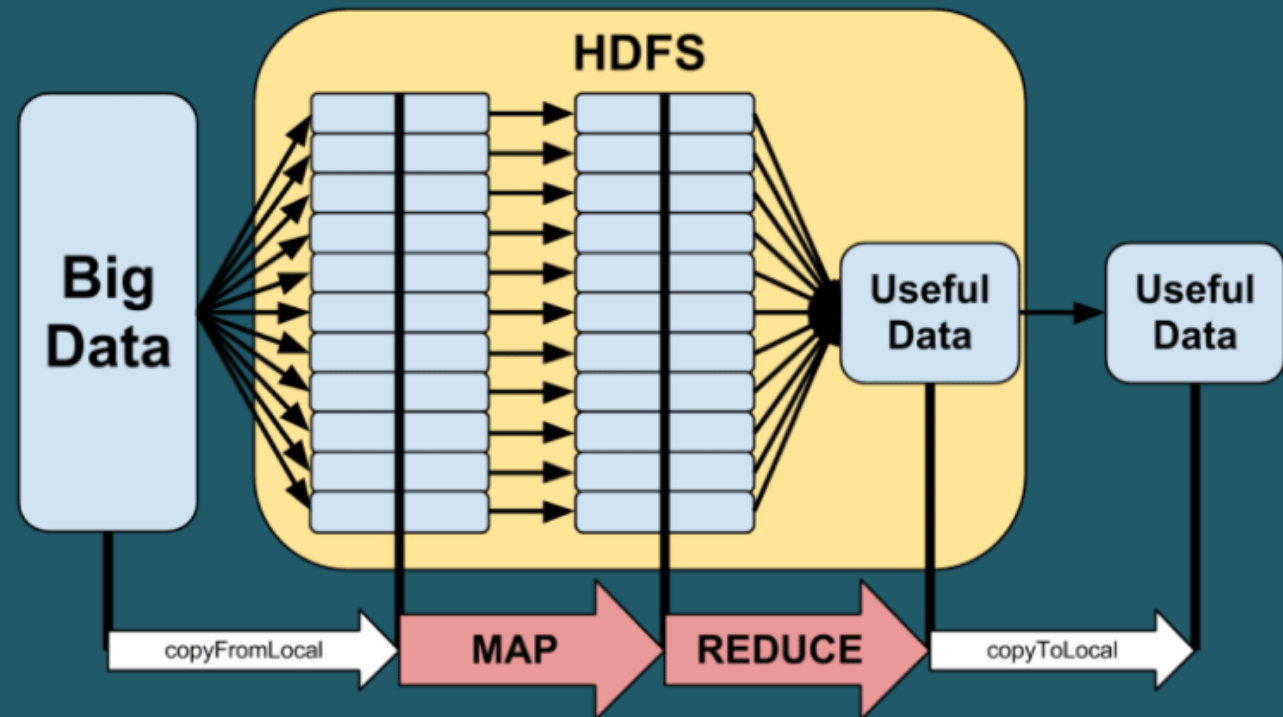
Ingeniero de Sistemas, realizó sus estudios de ingeniería en la UNAC, complementando con estudios en administración empresarial en la PUCP, Pablo es ejecutivo senior en tecnologías de la información, con más de 20 años de experiencia como consultor nacional e internacional en proyectos de Business Analytics y Big Data, así como en la dirección de proyectos & gerencia en tecnologías de la información, es asesor empresarial y especialista en Gobierno Electrónico, con dominio de tecnologías emergentes en Cloud con proveedores tales como AWS, Azure y GCP, es especialista e instructor en soluciones de clase mundial como Pentaho, IDempiere, Odoo, SuiteCRM e instructor en tecnologías privadas como Power BI, Microstrategy, Tableau, con dominio de lenguajes R, Python, Java y dominio de sistemas Linux y Unix, posee diversas especializaciones en seguridad informática, Big Data, DevOps, Pentaho, AWS, Azure y GCP. Desde 1993, es un activista del Software Libre en proyectos como Pentaho, IDempiere, Odoo, entre otros, actualmente se desempeña como **Chief Data Architect at GIS** y **Chief Executive Officer in EGS GROUP**

- Ex-Director de Tecnologías TIC en la empresa transnacional Carvajal S.A.
- Ex-Director de Tecnologías TIC en el Instituto del Mar del Perú – IMARPE
- Fue asesor en la hoy Secretaría de Gobierno Digital de la Presidencia del Consejo de Ministros (ex-ONGEI) – Perú.
- Ha brindado consultorías a diversas empresas nacionales e internacionales, entre las cuales destacan: El Grupo El Comercio, AJE Group, Premier Motors, Rural Telecom, Ministerio de Crédito y Hacienda en Nicaragua, entre otras.

# FUNDAMENTOS DE BIG DATA HADOOP Y EL MULTI-CLOUD



PentaDemy

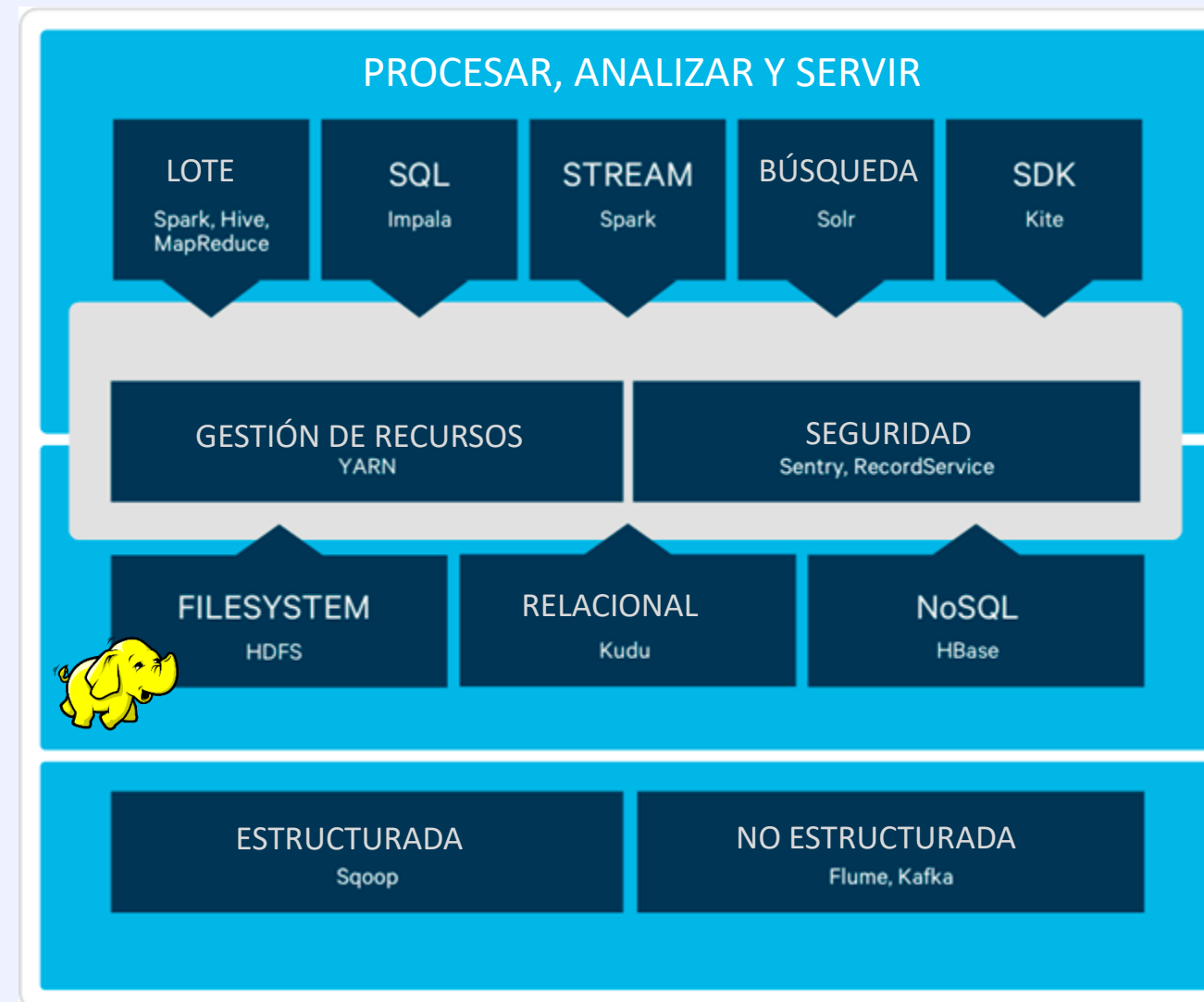


CLUDERA

# MÓDULO 01

## CLOUD COMPUTING BIG DATA & AWS

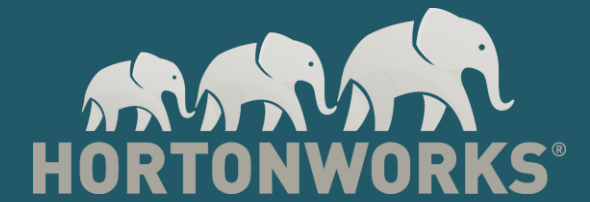
- ¿Qué es el Cloud Computing?
- Creando nuestra cuenta en AWS
- Conceptos de Big Data o Macro Datos
- Big Data en el mundo empresarial
- Las 5 V's del Big Data
- ¿Qué es la Alta paralelización?
- Fuentes de datos y su recolección
- Clúster computacional y alta paralelización.
- Arquitectura de soluciones.
- Pipeline de arquitectura tradicional
- Pipeline de Arquitectura de Big Data
- Almacenamiento y Cloud Computing
- Distribución de carga de trabajo
- Escalabilidad, Alta disponibilidad, Seguridad & Gobierno



- ¿Qué es Hadoop?
- Componentes de Hadoop
  - HDFS
  - Map Reduce
  - YARM
  - Common Utilities
- Distribuciones Hadoop
- Patrones de diseño
- Capas conceptuales
- Arquitectura conceptual
- MapReduce como motor de procesamiento
- Componentes tecnológicos disponibles
- Asegurando el tamaño de bloque
- Arquitectura tecnológica
- Arquetipo de una arquitectura Big Data genérica
- Definición de un Datalake Productivo



# HADOOP ALMACENAMIENTO DISTRIBUIDO EN CLUSTER CON **CLouDERA** DATA PLATFORM (CDP) & **HORTONWORKS** DATA PLATFORM (HDP)



## CLouDERA

### Fuente de Datos Telco

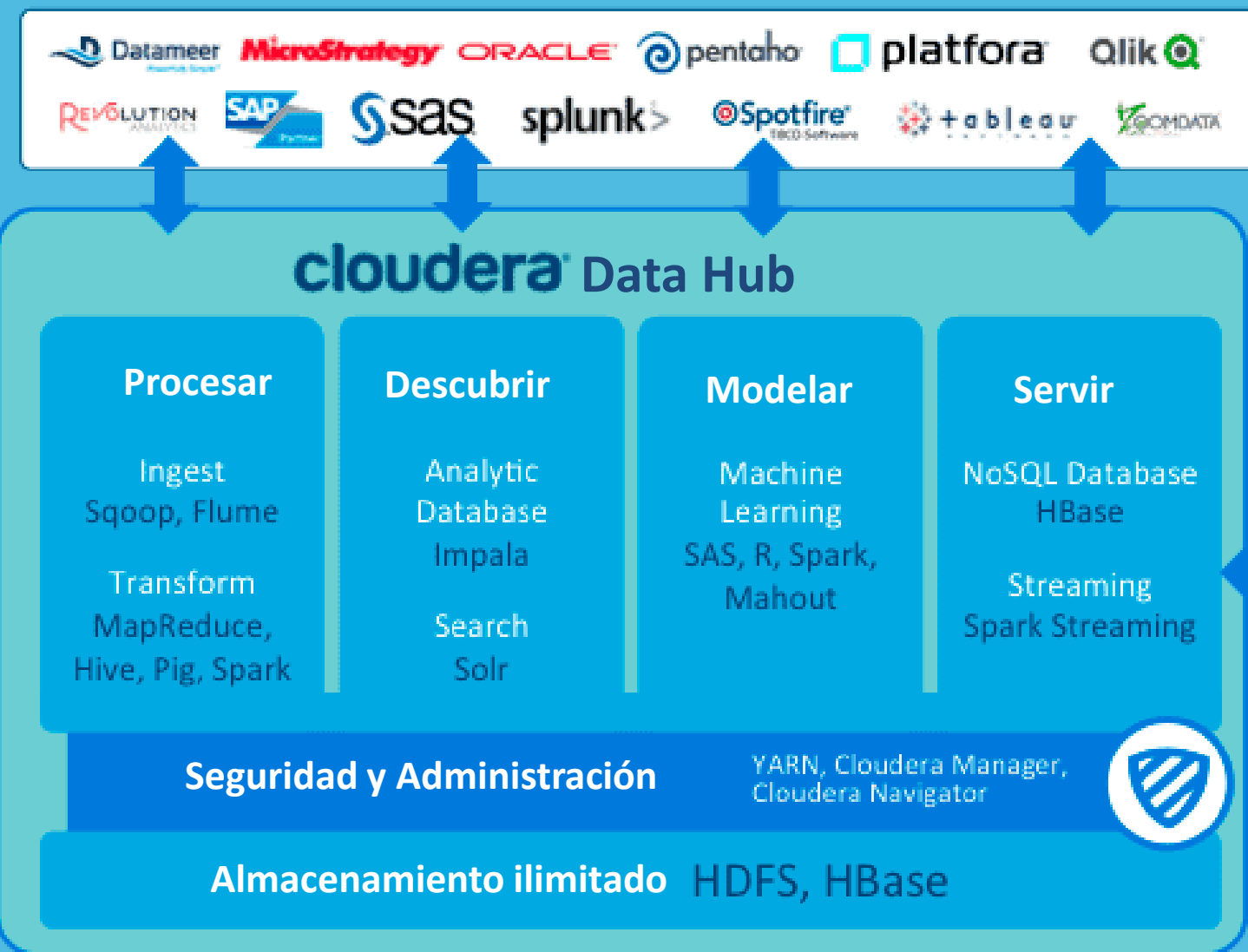
**Estructurada**

- Network
- Billing
- CRM
- Ordering
- Usage
- Inventory

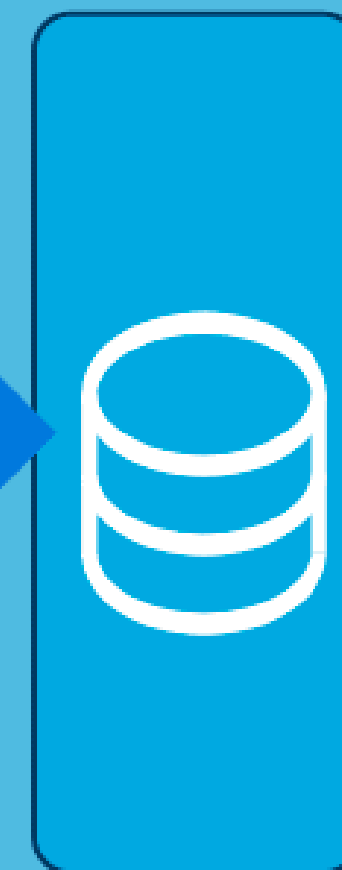
**Unstructured /Semi-Structured**

- Clickstream
- Sensors
- Machine Logs
- Social

Data Ingest -  
Streaming or Batch



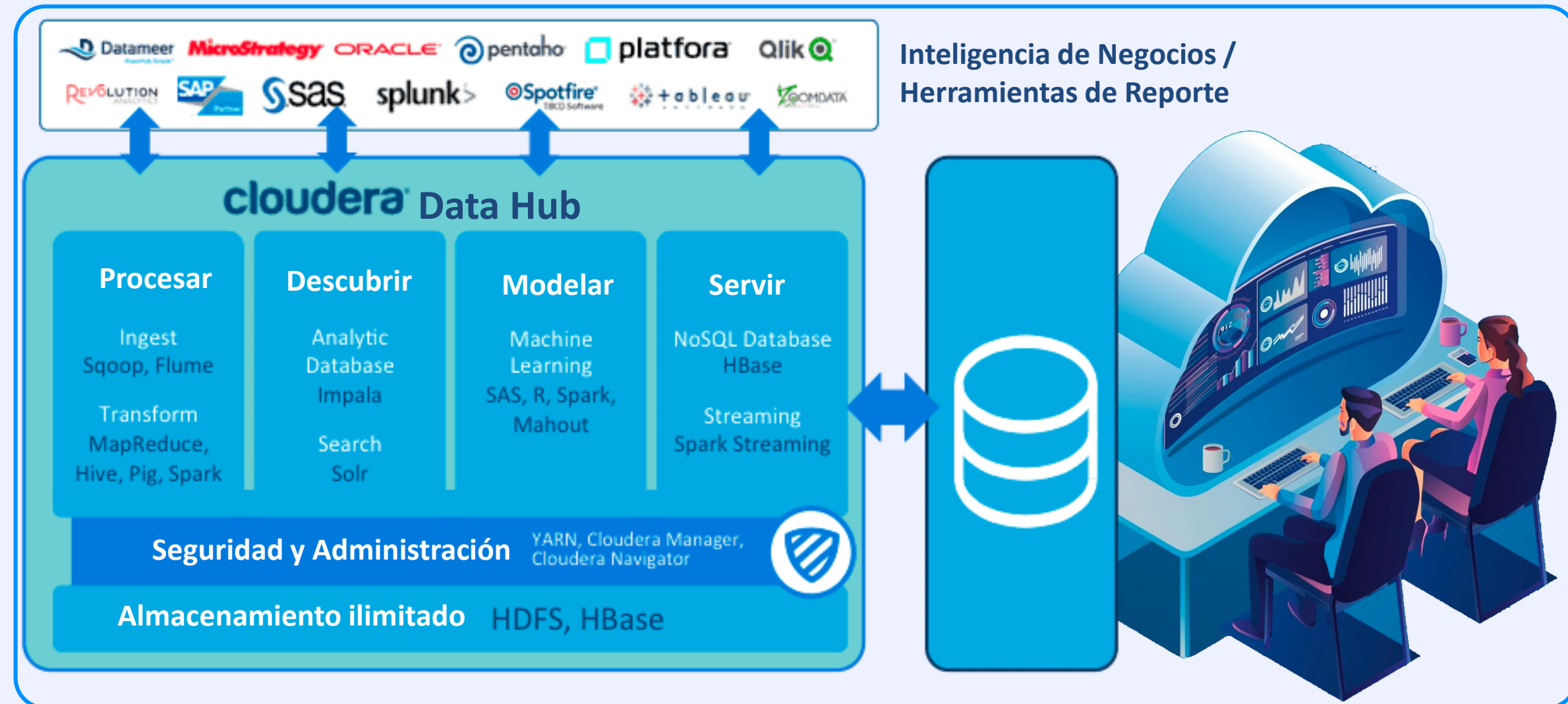
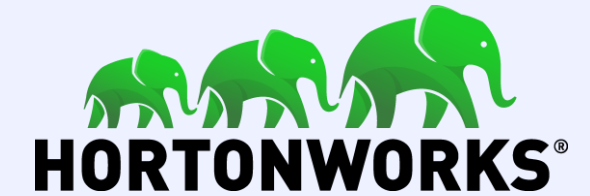
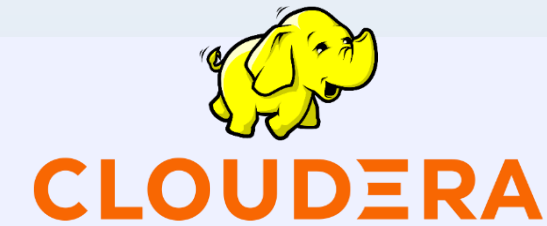
Inteligencia de Negocios /  
Herramientas de Reporte



# MÓDULO 02

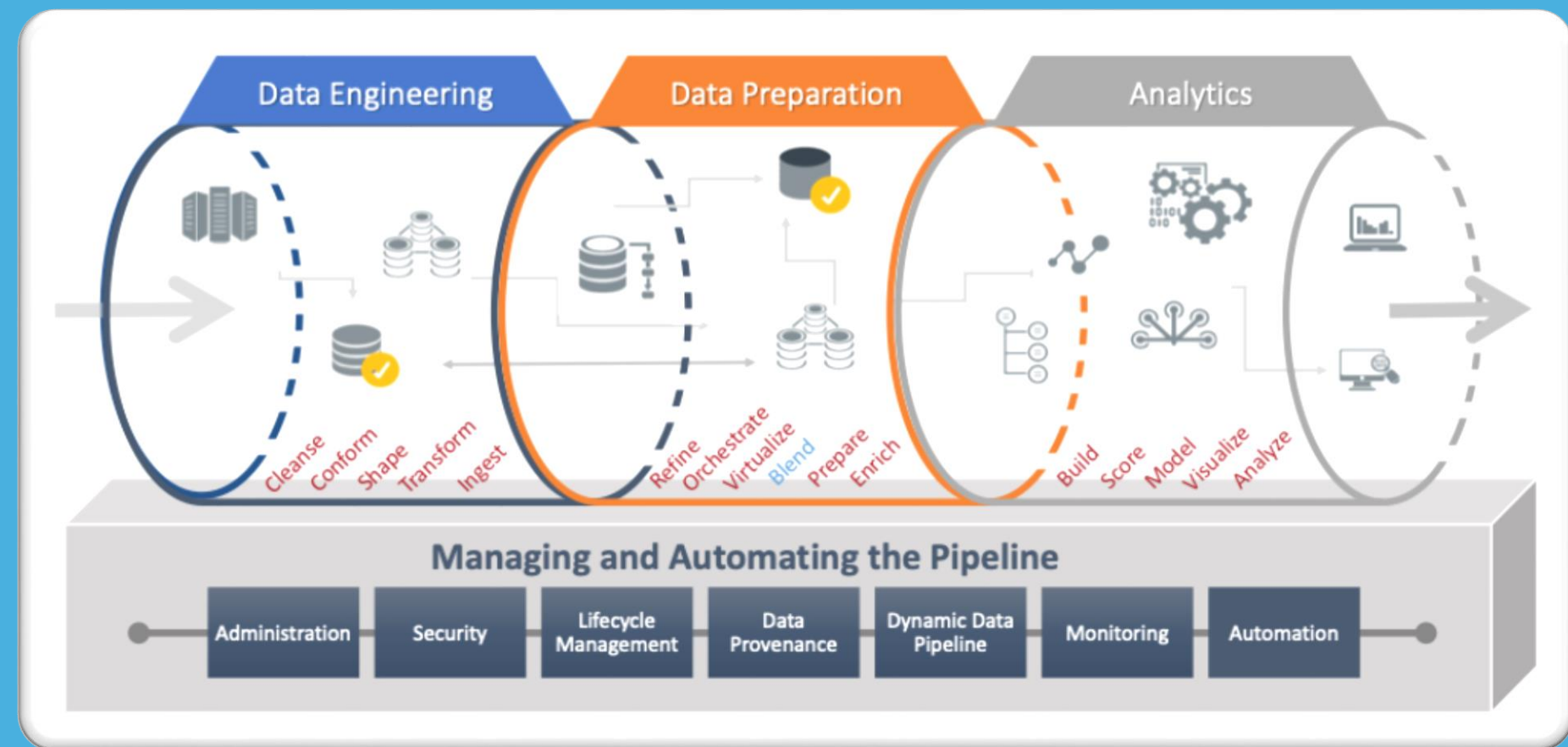
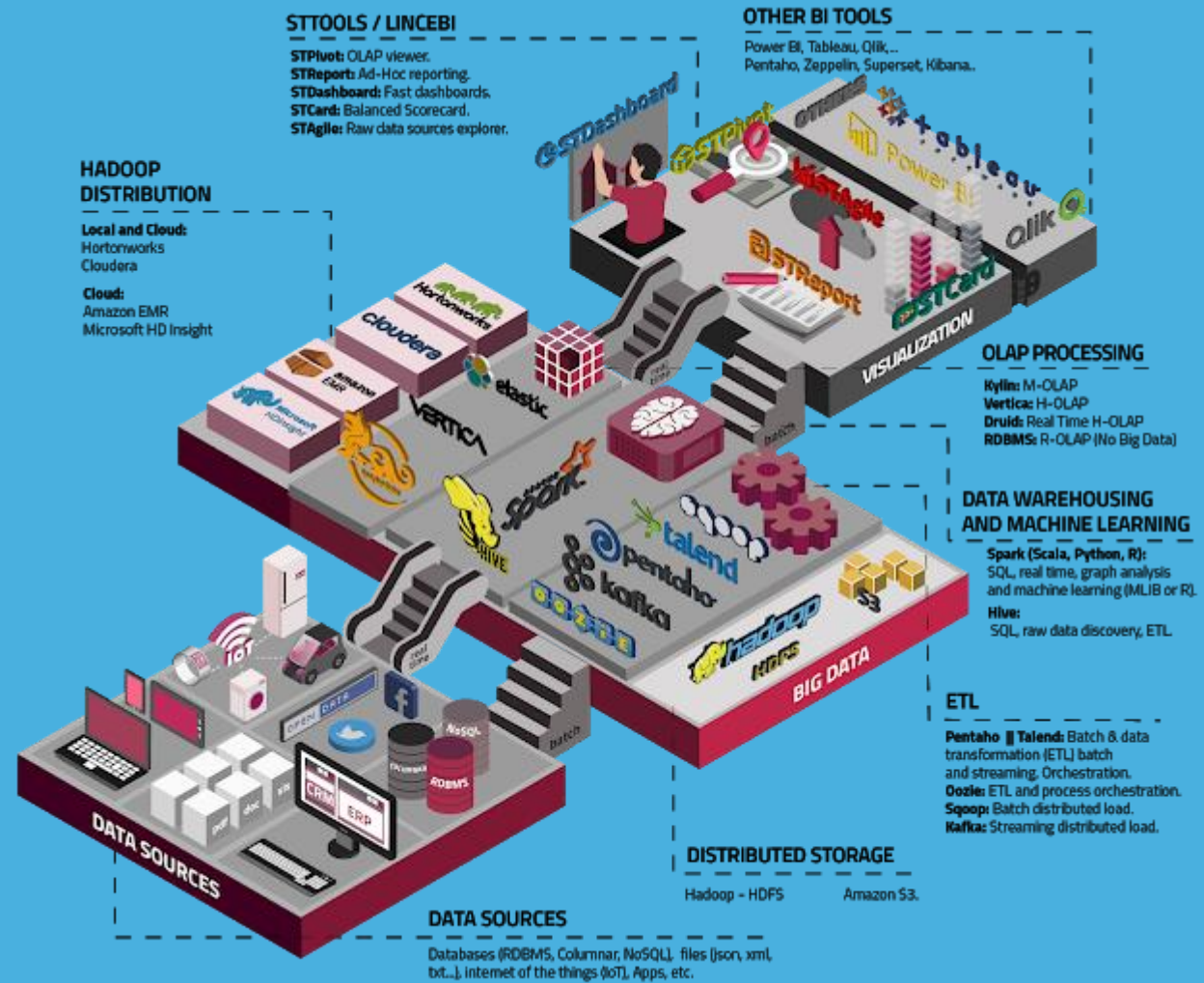
## HADOOP ALMACENAMIENTO DISTRIBUIDO EN CLUSTERS

- Despliegue de Cloudera CDH (VM)
- Despliegue de Hortonworks HDP (Docker)
- Tecnologías batch sobre Big Data.
- Trabajando de manera distribuida sobre un clúster
- Hadoop como estándar en el mundo de Big Data
- HDFS como motor de almacenamiento
- YARM como gestor de recursos.
- MapReduce como motor de procesamiento
- Replicación controlada de datos
- Asegurando el tamaño del bloque
- Capacidad física de un clúster
- Administración de archivos y recursos sobre Hadoop



# PROCESAMIENTO DISTRIBUIDO Y PARALELIZADO CON HIVE

## PIPELINE DE DATOS

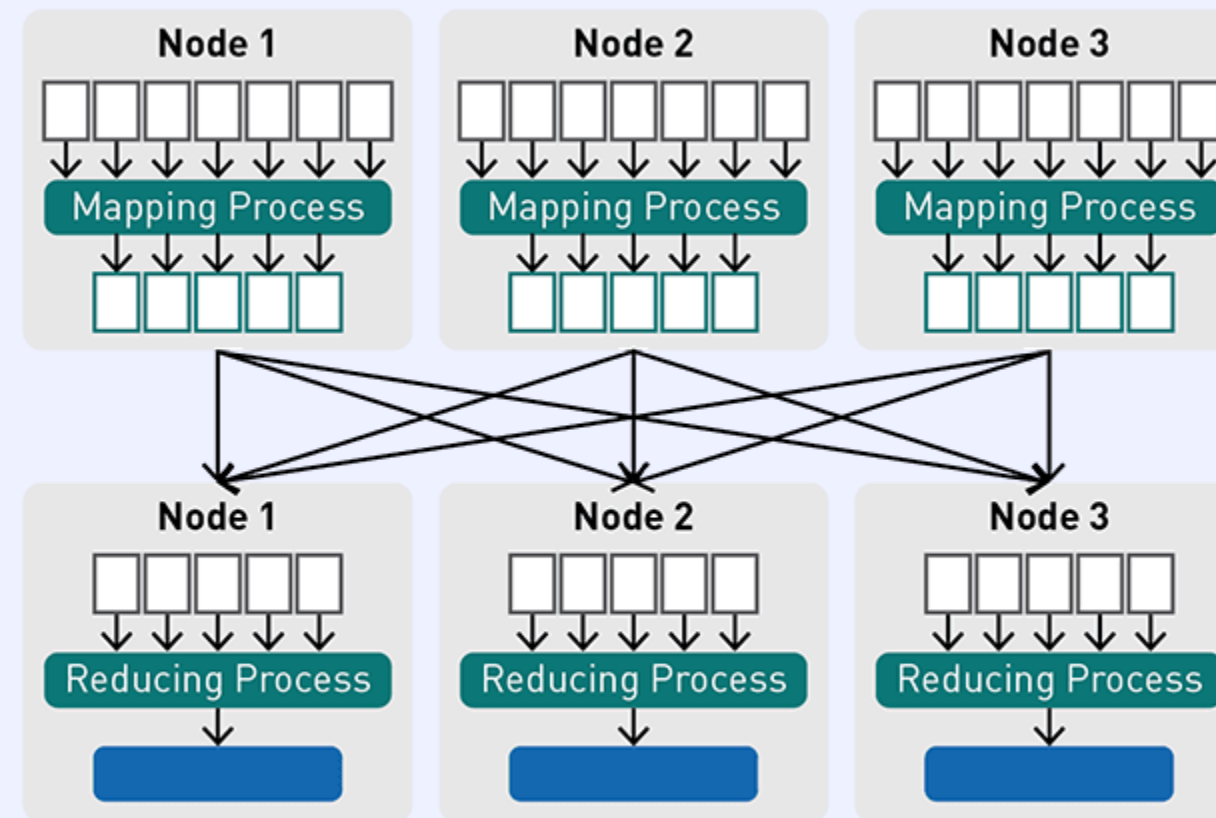


# MÓDULO 03



## PROCESAMIENTO DISTRIBUIDO Y PARALELIZADO

- Arquitectura interna de Hive
- Arquitectura de HBase
- Hive vs HBase
- Hive como infraestructura de almacenamiento
- SQL sobre MapReduce
- Archivos de HDFS como tablas Hive
- Particionamiento estático y dinámico
- Formatos binarios de archivos: Parquet, ORC y Avro
- Compresión optimizada de datos
- Configuración y tuneo de procesos en Hive
- Sqoop como motor de ingesta de datos
- Importando datos a Hadoop de base de datos relacionales.

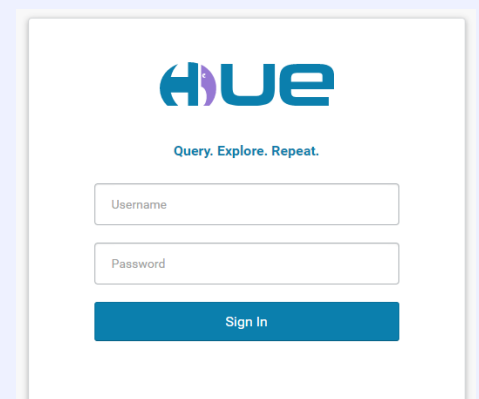


## ARQUETIPOS Y TUNING

- Arquetipo de ingesta de datos batch
- Arquetipo de modelamiento de datos
- Arquetipo de procesamiento de datos
- Hue como interfaz gráfica para los procesos
- Data Analytics Studio como interfaz gráfica para los procesos
- Creando consultas adhoc con Impala
- Tuning de código
- Tuning de paraleización: MapReduce vs Tex vs Spark

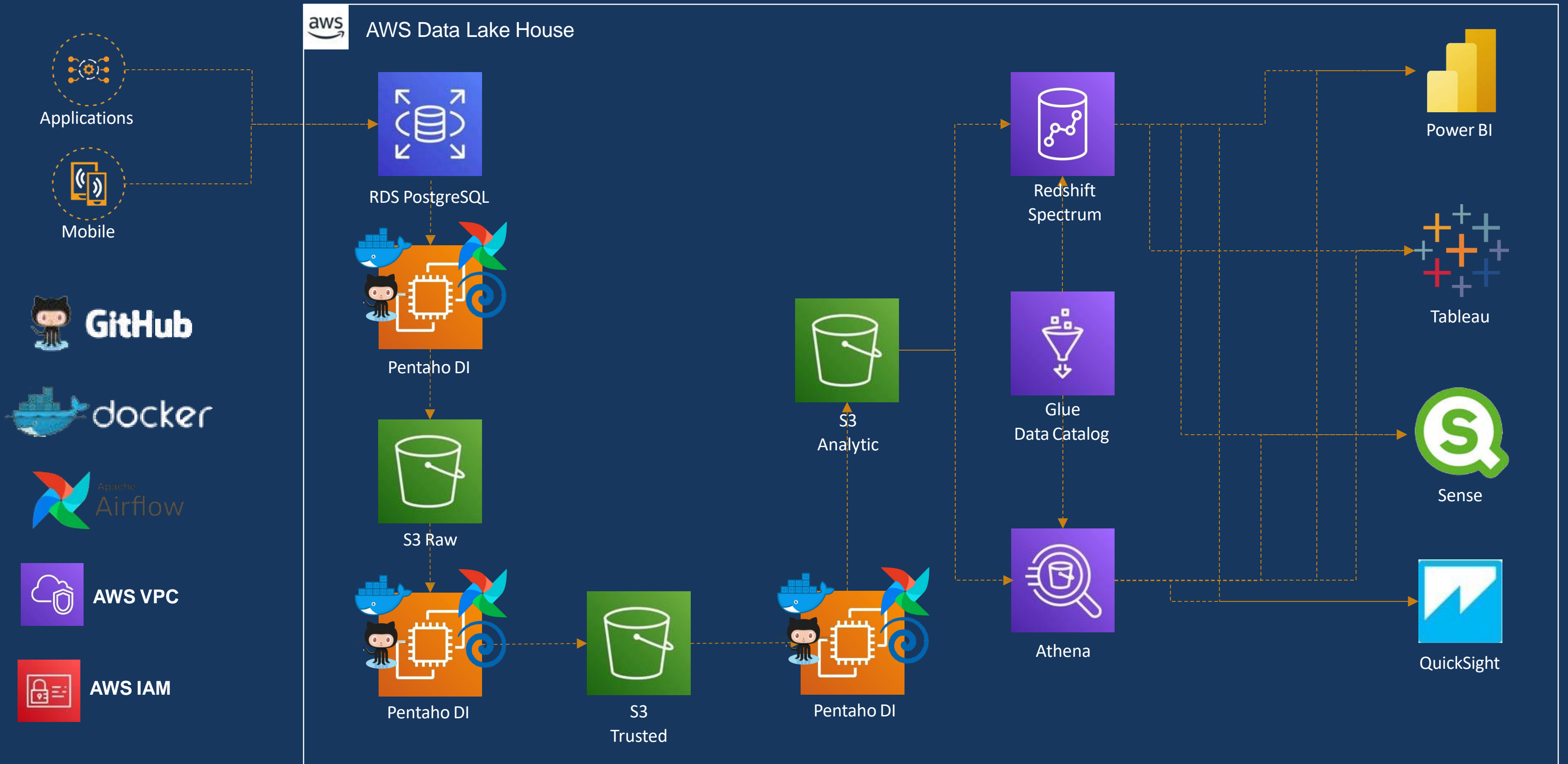


Data Analytics Studio



# IMPLEMENTACIÓN DE UN DATALAKE

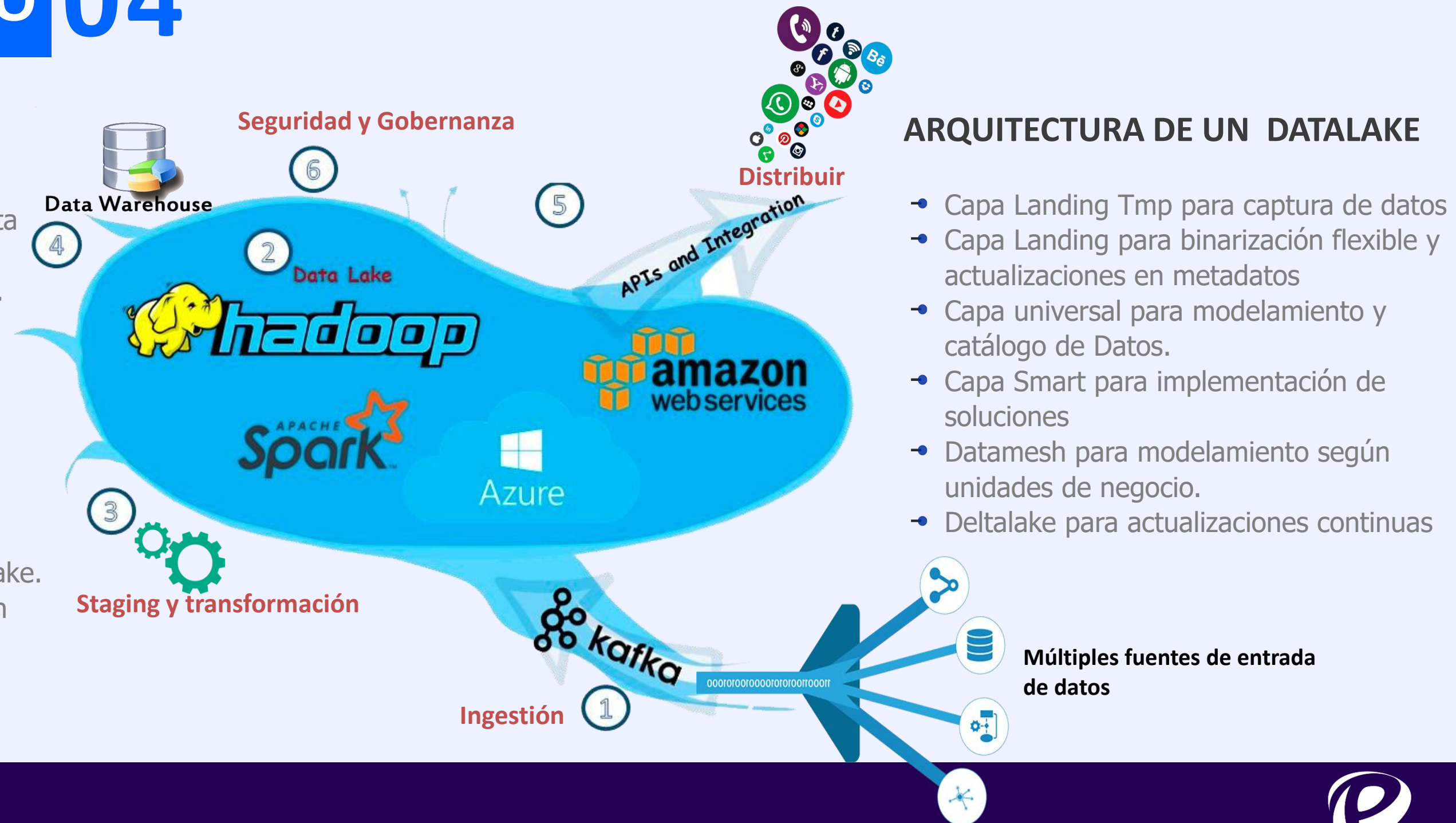
Desmitificando un Datalake, comprende su arquitectura y funcionamiento de forma clara y sin secretos



# MÓDULO 04

## IMPLEMENTACIÓN DE UN DATALAKE

- Arquetipo de una arquitectura Big Data genérica.
- Definición de un DataLake Productivo.
- Datalake como técnica de gobierno de procesos de BigData.
- Flujos ETL sobre un DataLake
- Soluciones de Reporting sobre un DataLake.
- Soluciones Semi-Estructuradas y no-estructuradas sobre un Data Lake.
- Soluciones Real-Time sobre un DataLake.
- Soluciones de Deep Learning sobre un DataLake



## ARQUITECTURA DE UN DATALAKE

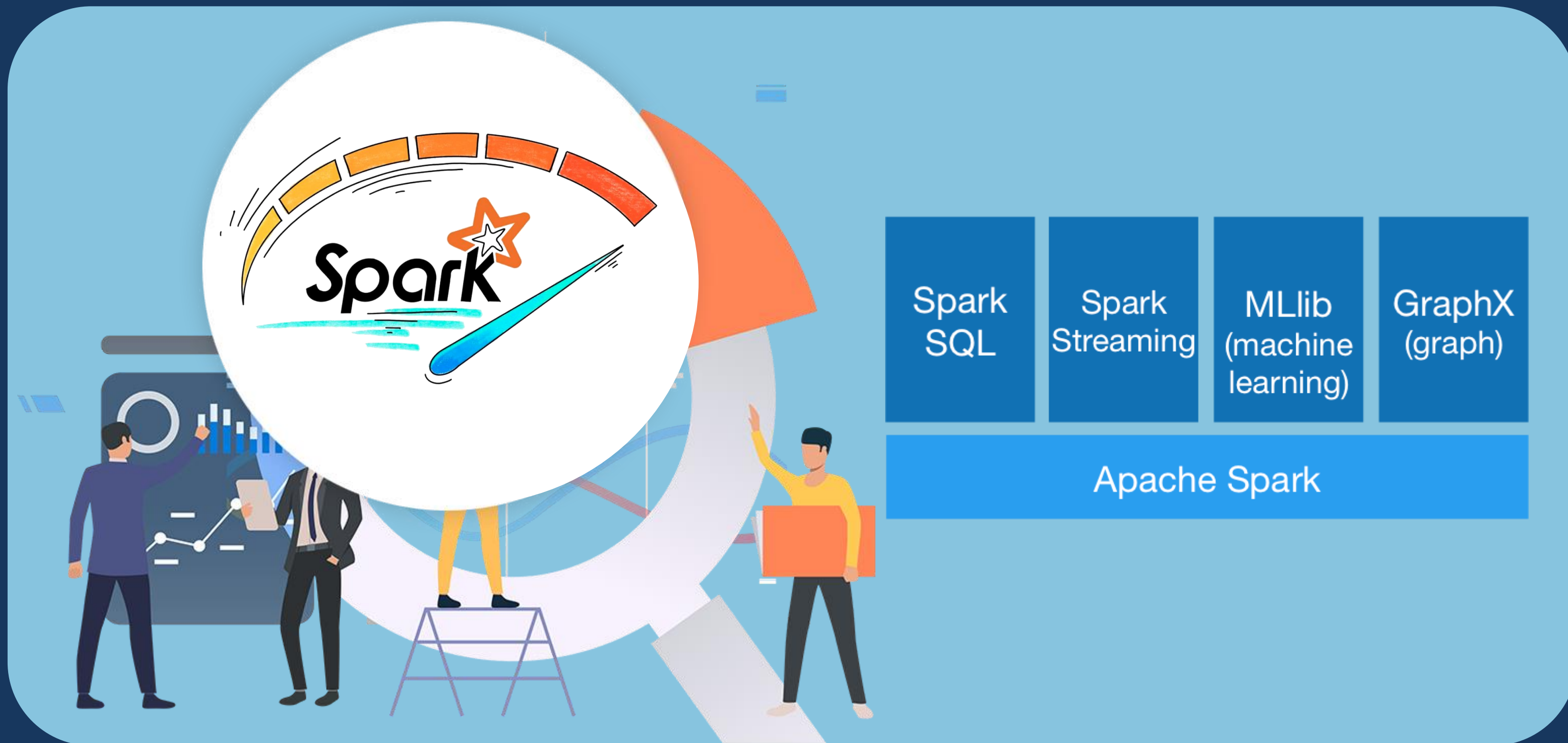
- Capa Landing Tmp para captura de datos
- Capa Landing para binarización flexible y actualizaciones en metadatos
- Capa universal para modelamiento y catálogo de Datos.
- Capa Smart para implementación de soluciones
- Datamesh para modelamiento según unidades de negocio.
- Deltalake para actualizaciones continuas

Múltiples fuentes de entrada de datos

# SPARK PARA PROGRAMACIÓN DISTRIBUIDA



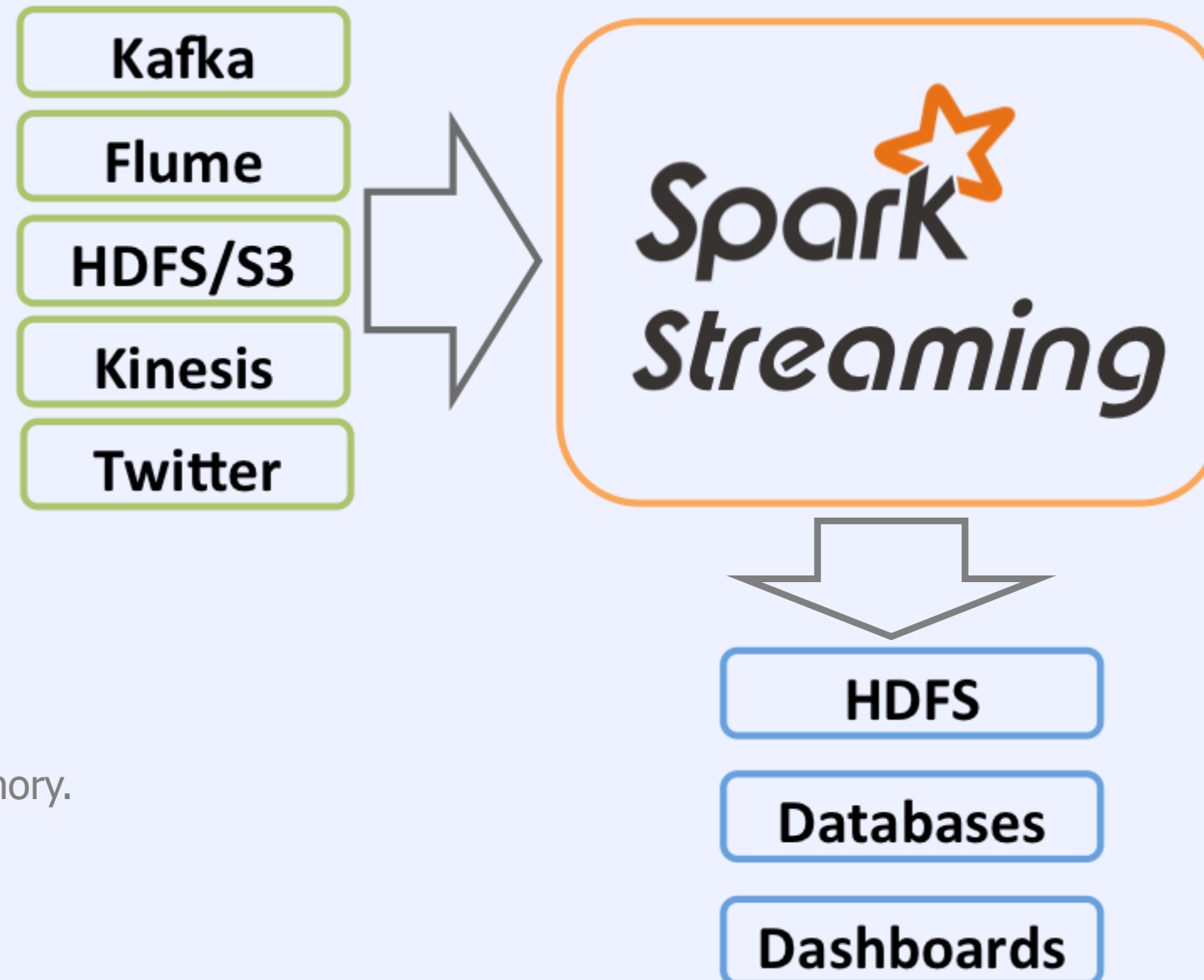
PentaDemy



# MÓDULO 05

## SPARK PARA PROGRAMACIÓN DISTRIBUIDA

- Tecnologías in-memory sobre BigData
- Conociendo Spark
- Spark vs Hadoop
- Hive vs Spark
- Uso intensivo de memoria con Spark
- Módulos Spark
- Spark Core
- Trabajando con Dataframes
- Transformaciones y acciones sobre DataFrames.
- Escritura a disco duro de datos in-memory.



## PATRONES Y TUNING EN SPARK

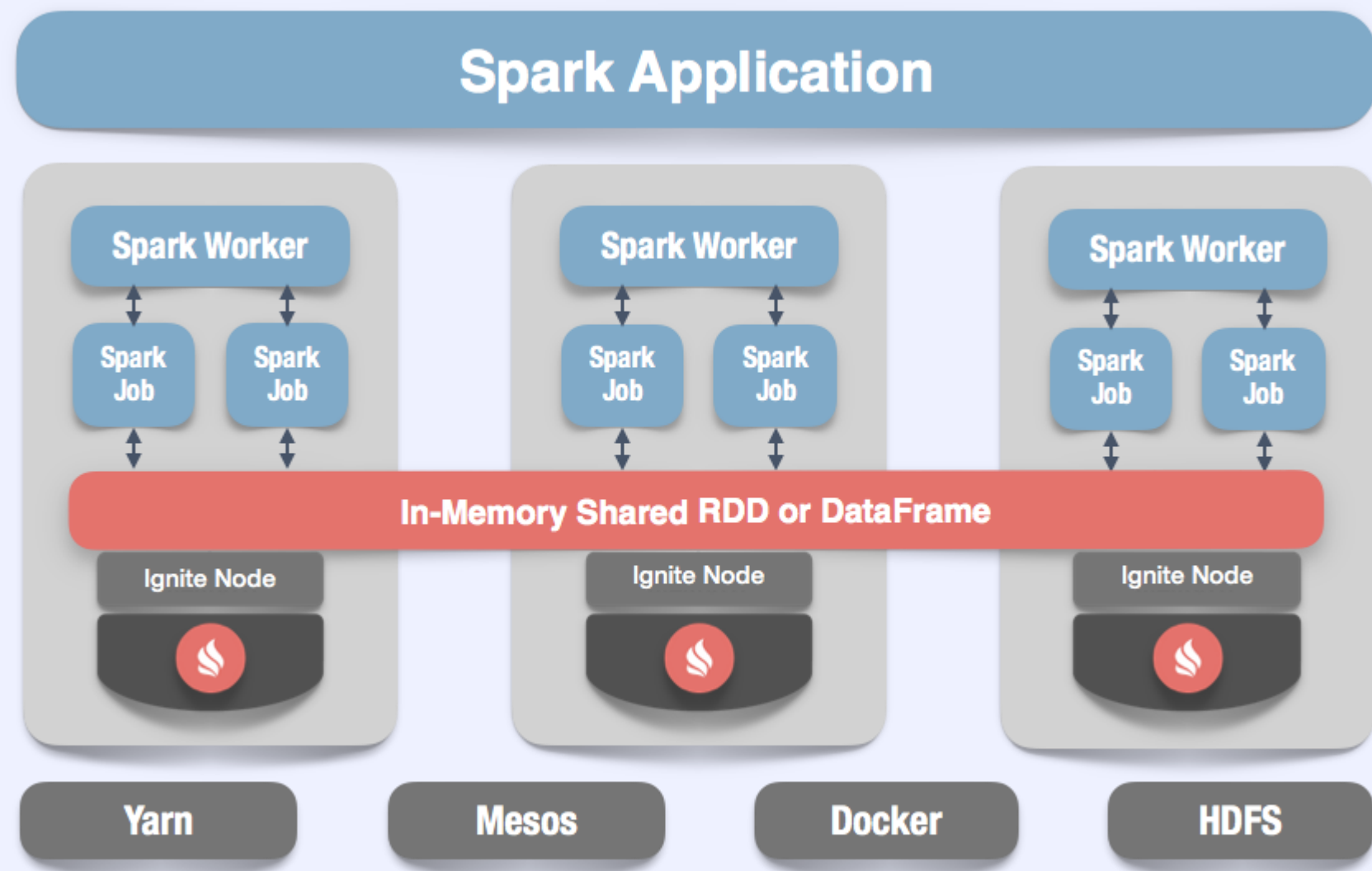
- Spark SQL para procesamiento estructurado
- PySpark para programación y procesamiento funcional
- Creación de funcionales personalizadas con UDFs
- Patrón de diseño caché
- Patrón de diseño checkpoint
- Tuning sobre executors
- Evitando el colapso de memoria RAM.
- Spark con Hortonworks Data Platform
- Spark con Cloudera Data Platform



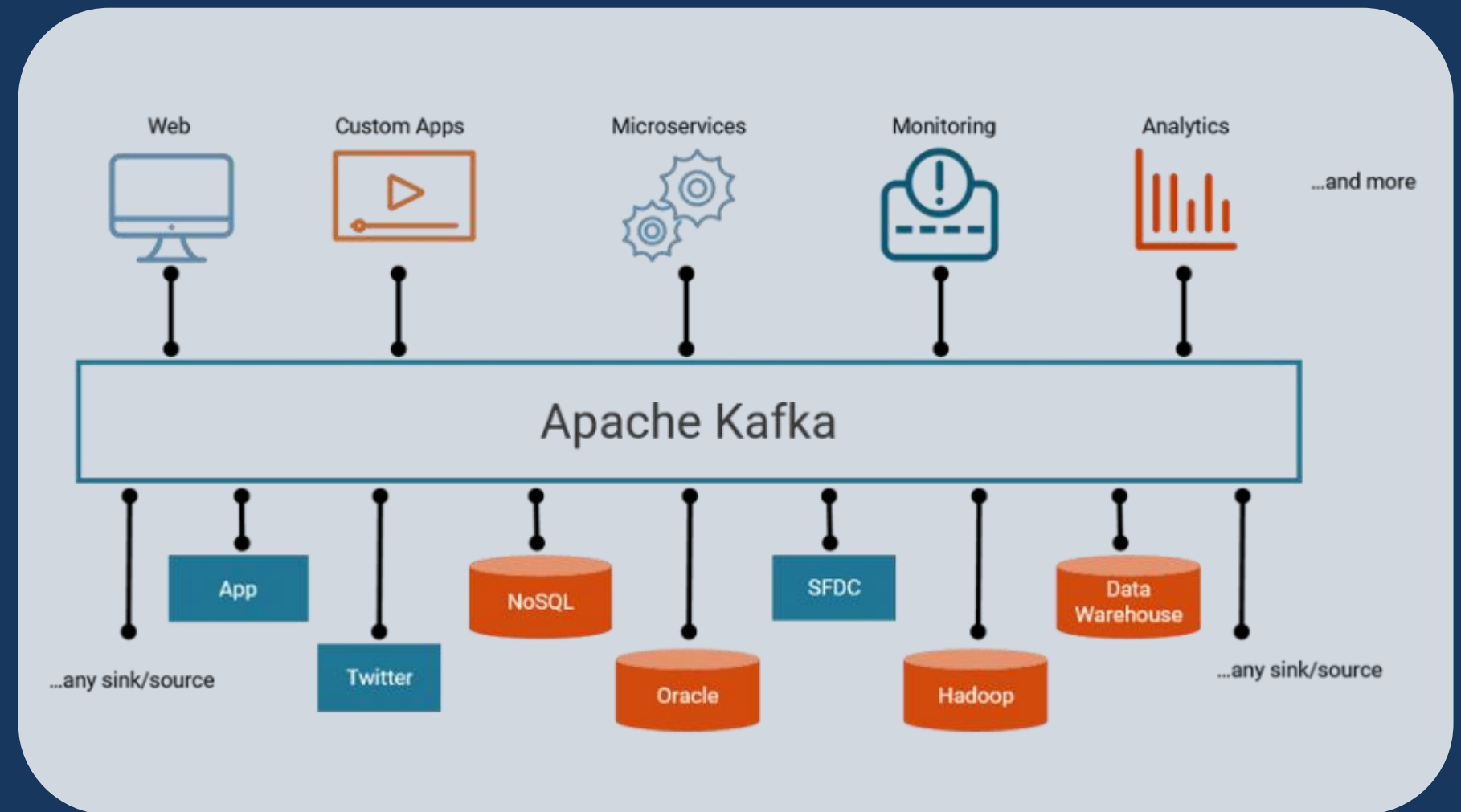
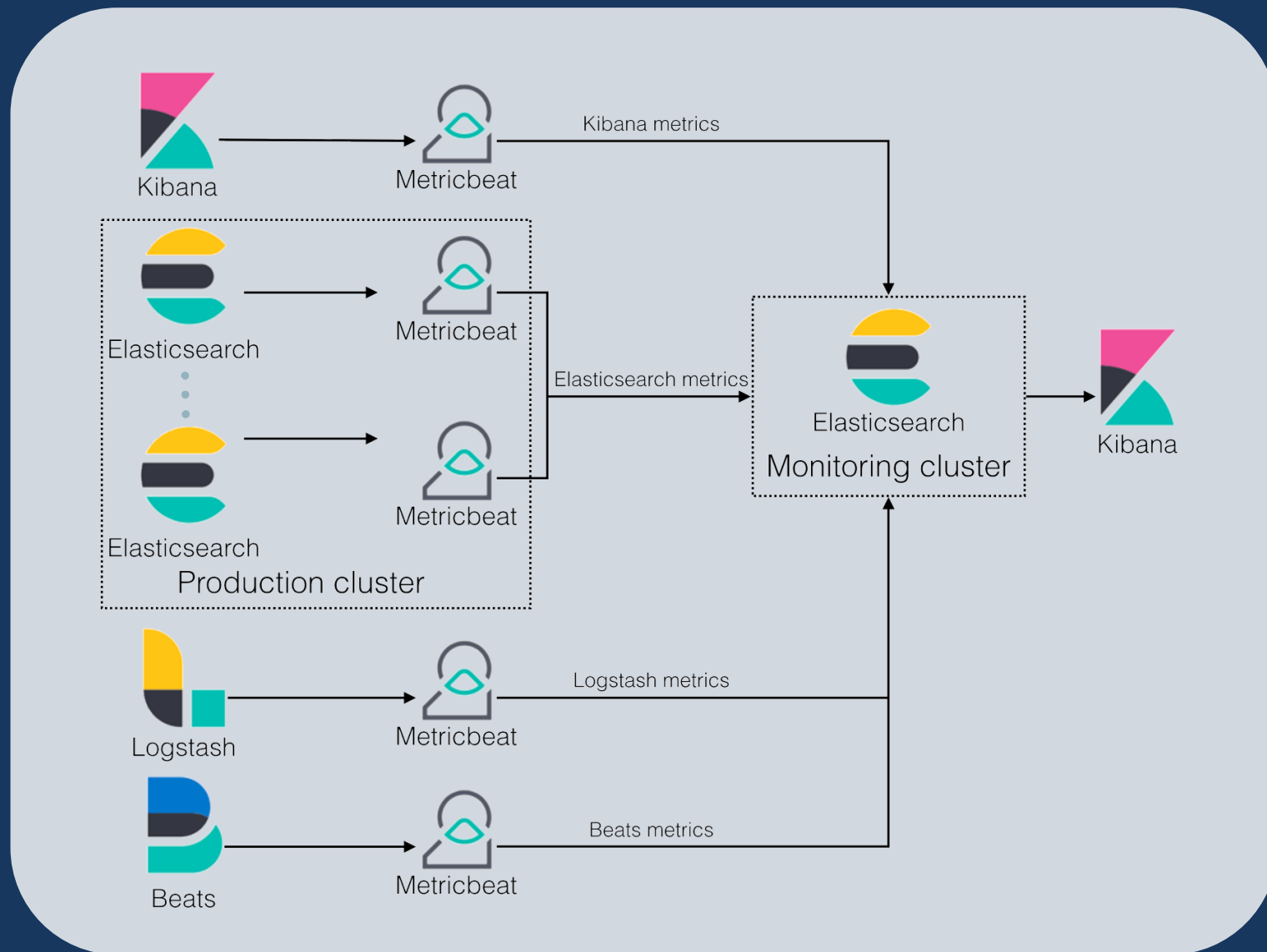
# MÓDULO 06

## PROCESAMIENTO SEMI-ESTRUTURADO CON SPARK

- Fuentes y datos semi-estructurados
- Spark y dataframes semi-estructurados
- Esquemas flexibles
- Lectura de fuentes semi-estructurados
- Campos complejos
- Navegación sobre campos complejos
- Modelamiento y almacenamiento de fuentes semi-estructuradas
- Procesamiento in-memory con spark



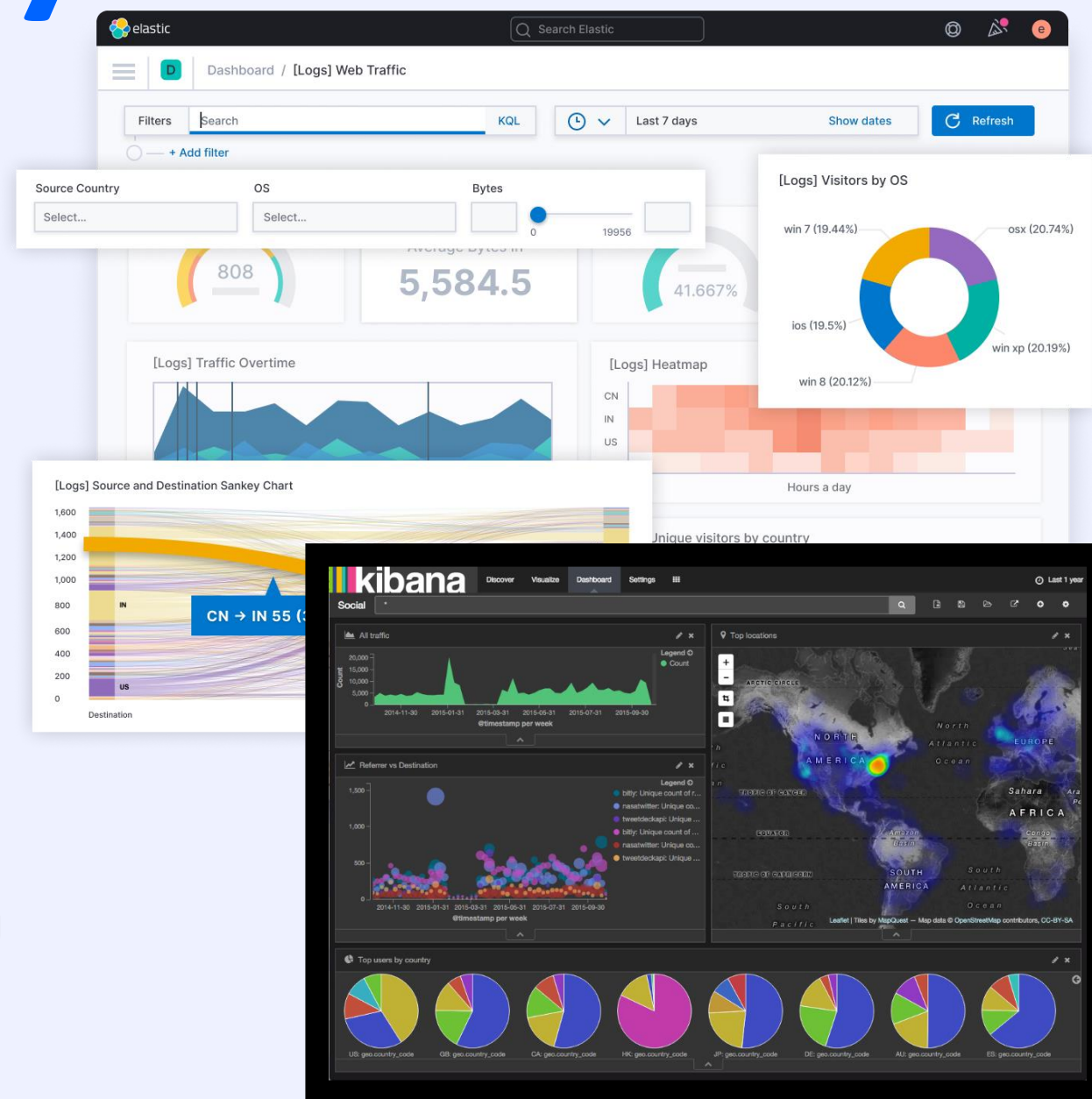
# PROCESAMIENTO REAL-TIME SORBE BIG DATA CON KAKFA & ELASTICKSEARCH



## MÓDULO 07

### PROCESAMIENTO REAL-TIME

- Procesamiento de datos real time
- ¿Streaming, real time, near real time o micro batch?
- Arquitectura general para proyectos real time
- Captura de datos desde fuentes real time: tormenta de datos
- La elasticidad en la capa de captura y procesamiento.
- La importancia de la paralelización elástica.
- Evitando colapso de CPU
- Kafka como repositorio temporal de baja latencia.
- Tópicos, productores y consumidores
- Procesamiento real time con Spark Streaming
- Arquetipo de procesamiento real time
- Arquetipo de enriquecimiento real time
- Limitaciones y cómo superarlas.

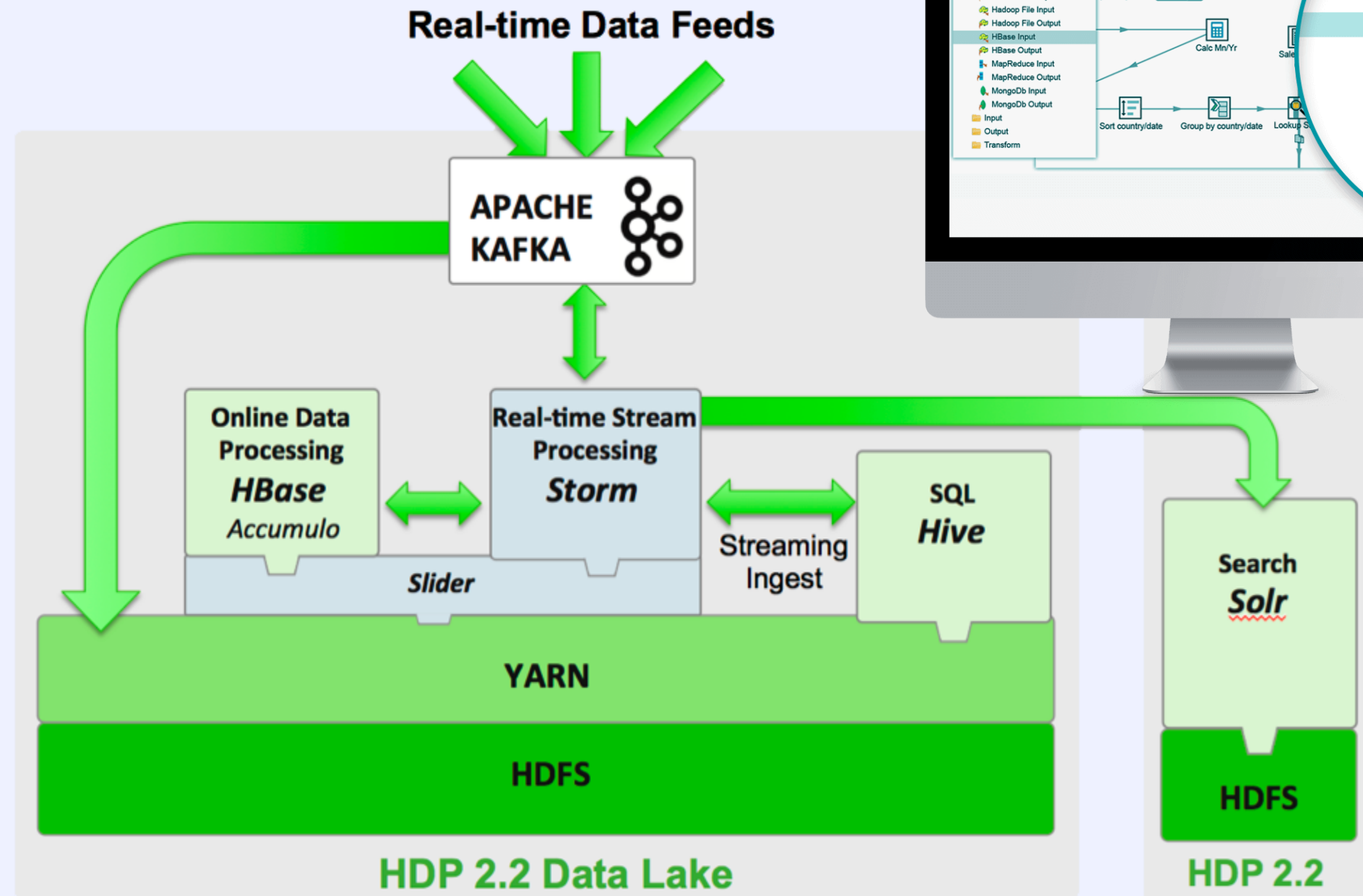


### Elastic Stack: beats, Logstash & Kibana

- Despliegue de Elastic Stack
- Beats: Metricbeat, Packetbeat, FileBeat
- Entradas Logstash: Stdin, File, Beats
- Filtros Logstash: Mutate, Date, Translate
- Elasticsearch: Queries, agregaciones
- Kibana: Mapas, pie, gauge, tablas
- Laboratorio integrado con Elastic Stack: Errores comunes. Login y performance. navegación por los dashboards. recorrido por la aplicación.

## DEMO DE TRANSMISIÓN DE DATOS EN TIEMPO REAL CON PENTAHO & APACHE KAFKA ON HORTOWORKS DATA PLATFORM HDP

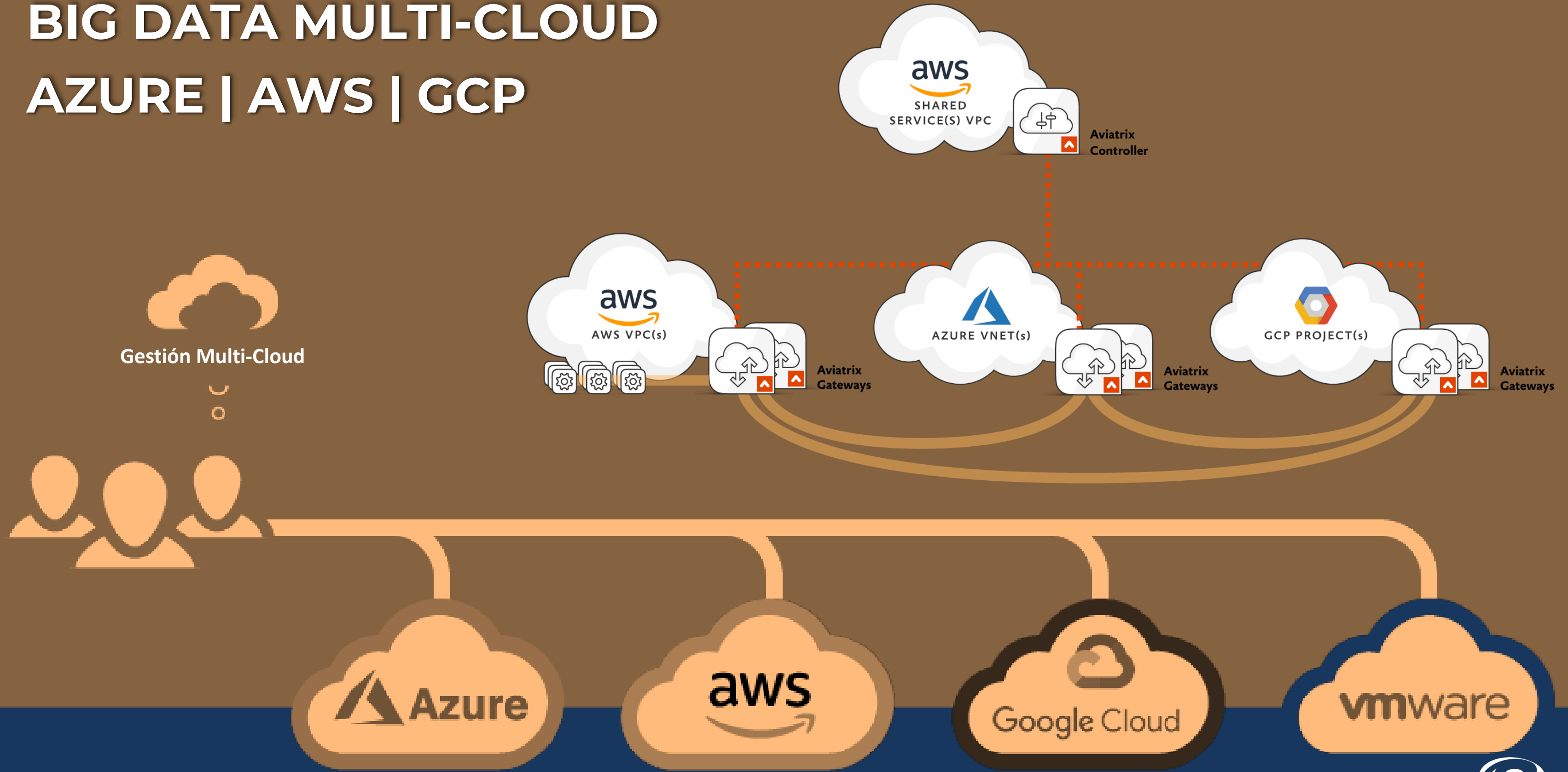
- ¿Qué es Apache Kafka?
- Arquitectura y Despliegue local
- Preparando Pentaho Data Integration (PDI)
- Proyecto Bancario Demo con Kafka
- Acceso a los logs de sitio web bancario
- Productores y consumidores de logs con PDI
- Consumidores Kafka de múltiples Topics
- Demo de procesamiento en tiempo real extremo a extremo



Big Data

- Cassandra Input
- Cassandra Output
- Hadoop File Input
- Hadoop File Output
- HBase Input
- HBase Output
- MapReduce Input
- MapReduce Output
- MongoDb Input
- MongoDb Output
- Input
- Output
- Transform

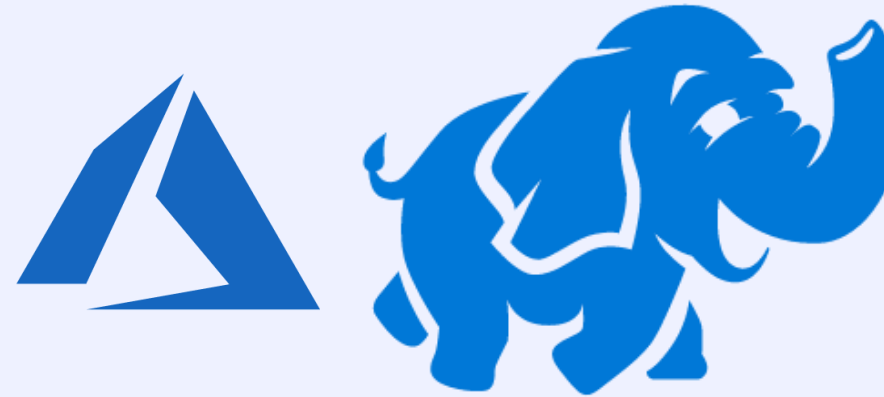
# BIG DATA MULTI-CLOUD AZURE | AWS | GCP



# MÓDULO 08

## BIG DATA ON AZURE

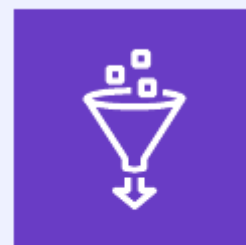
- Servicios de Big Data disponibles en Azure.
- Arquitectura de Big Data sobre Azure.
- Ingesta y almacenamiento de datos sobre Blob Storage.
- Interfaz SQL con Synapse Analytics.
- Implementación de flujos ETL con DataFlow
- Infraestructura para Clústers de Big Data con HDInsight
- Implementación de soluciones con Spark para HDInsight
- Despliegue y Workflows con Data Factory.



Lambda



EMR



Glue



Step Functions

# MÓDULO 09

## BIG DATA ON AWS

- Servicios de Big Data disponibles en AWS.
- Arquitectura de Big Data sobre AWS.
- Ingesta y almacenamiento de datos sobre S3
- Interfaz SQL de AWS con Athena
- Implementación de flujos ETL con Glue
- Infraestructura para Clústers de Big Data con EMR
- Implementación de soluciones con Spark para EMR
- Despliegue y Workflows con Glue.

# MÓDULO 10

## BIG DATA ON GCP

- Servicios de Big Data disponibles en GCP.
- Arquitectura de Big Data sobre GCP.
- Ingesta y almacenamiento de datos sobre Cloud Storage.
- Interfaz SQL con GCP y Big Query.
- Implementación de flujos ETL con Data Fusion
- Infraestructura para Clústers de Big Data con Data Proc
- Implementación de soluciones con Spark para DataProc
- Despliegue y Workflows con Cloud Composer.



# MACHINE LEARNING SOBRE BIG DATA





# MÓDULO 11

## MACHINE LEARNING SOBRE BIG DATA

- Analítica en el mundo empresarial
- Analítica exploratoria
- Creando consultas ad\_hoc con impala
- Compartiendo la metadata del clúster entre herramientas
- Métodos descriptivos sobre Big data
- Algoritmos descriptivos disponibles sobre Spark MLlib
- Implementación de KMeans con Spark
- Métodos predictivos sobre Big Data
- Implementación de un modelo predictivo
- Algoritmos predictivos disponibles sobre Spark MLlib

