Programa AWS Big Data – Analytics Especialist

"Big Data, el nuevo petróleo"

BIGDATA ACADEMY

Reflexión de la noche

« Un esfuerzo más y lo que iba a ser un fracaso se convierte en un éxito; no existe el fracaso si nos esforzamos cada vez mas y mas»

BIGDATA ACADEMY

ANITA QUEVEDO()

Intercorp **Data Engineer Sr;**

BIGDATA ACADEMY ; Universidad Ricardo Palma ; CDA Cloud & Data Analytics

Directora / Docente de Tecnologías Cloud | **Docente de Big Data - Spark** | **Fundadora Comunidad**

Estudiante apasionada, profesora y profesional de TI con 7 años de experiencia en proyectos de BI, BA Big Data y Cloud Computing, y cursando el Master en Data Management en U. de Barcelona

Trabajé en:

Ferreyros CAT UNA EMPRESA FERREYCORP → Interbank → BBVA → everis an NTT DATA Company

Metodología de enseñanza

- ★ **Complementar la teoría con casos reales**
- ★ **Discusiones tecnológicas**
- ★ **Laboratorios con ejemplos útiles y reusables**
- ★ **Material complementario para profundizar**
- ★ **Comunicación 360**

BDA BIGDATA ACADEMY

Módulos del programa

- ★ Overview herramientas
- ★ Collect
- ★ Storage
- ★ Procesamiento
- ★ Analítica
- ★ Seguridad
- ★ Visualización

BIGDATA ACADEMY

Te queremos conocer...

Nos interesa saber tus expectativas, dudas y experiencia.

1.¿Donde trabajas cuáles son tus principales funciones?
2.¿Qué conoces de Big Data?
3.¿Que esperas del programa?

BIGDATA ACADEMY

Cosas a tener en cuenta

Se creará un grupo de wsp para estar todos comunicados

El material estará disponible en **Google Drive**

BIGDATA ACADEMY

Sesión 1:
Introducción

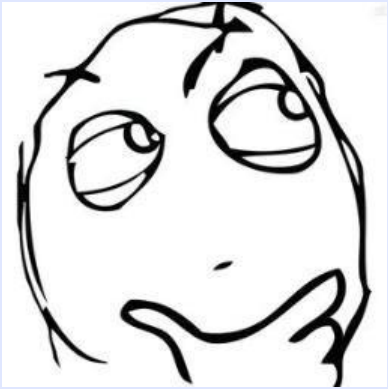
BIGDATA ACADEMY

Sesión 1: Objetivos

- ✓ Definir Big Data
- ✓ Identificar algunas fuentes de Big Data
- ✓ Listar ejemplos de uso de Casos de Big Data
- ✓ Describir el ecosistema de Big Data
- ✓ Review del big data “pipeline” y las herramientas disponibles para cada fase.

BIGDATA ACADEMY

¿Qué es el Big Data?



BIGDATA ACADEMY

¿Qué es el Big Data?

La colección y análisis de grandes volúmenes de data:

Responder preguntas

Crear ventaja competitiva

BIGDATA ACADEMY

Cuando los dataset son grande , resultan difíciles para

- - Almacenar
- - Organizar
- - Analizar
- - Mover
- - Compartir

BIGDATA ACADEMY

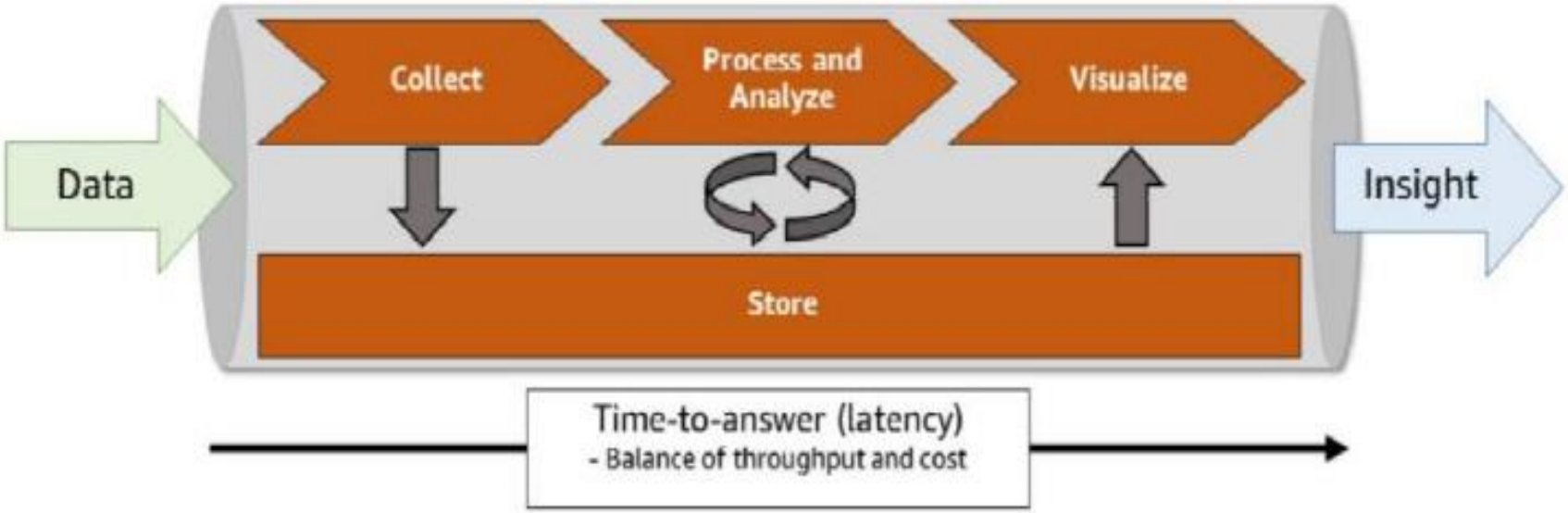
Big Data puede generarse a partir de muchas fuentes diferentes, que incluyen:

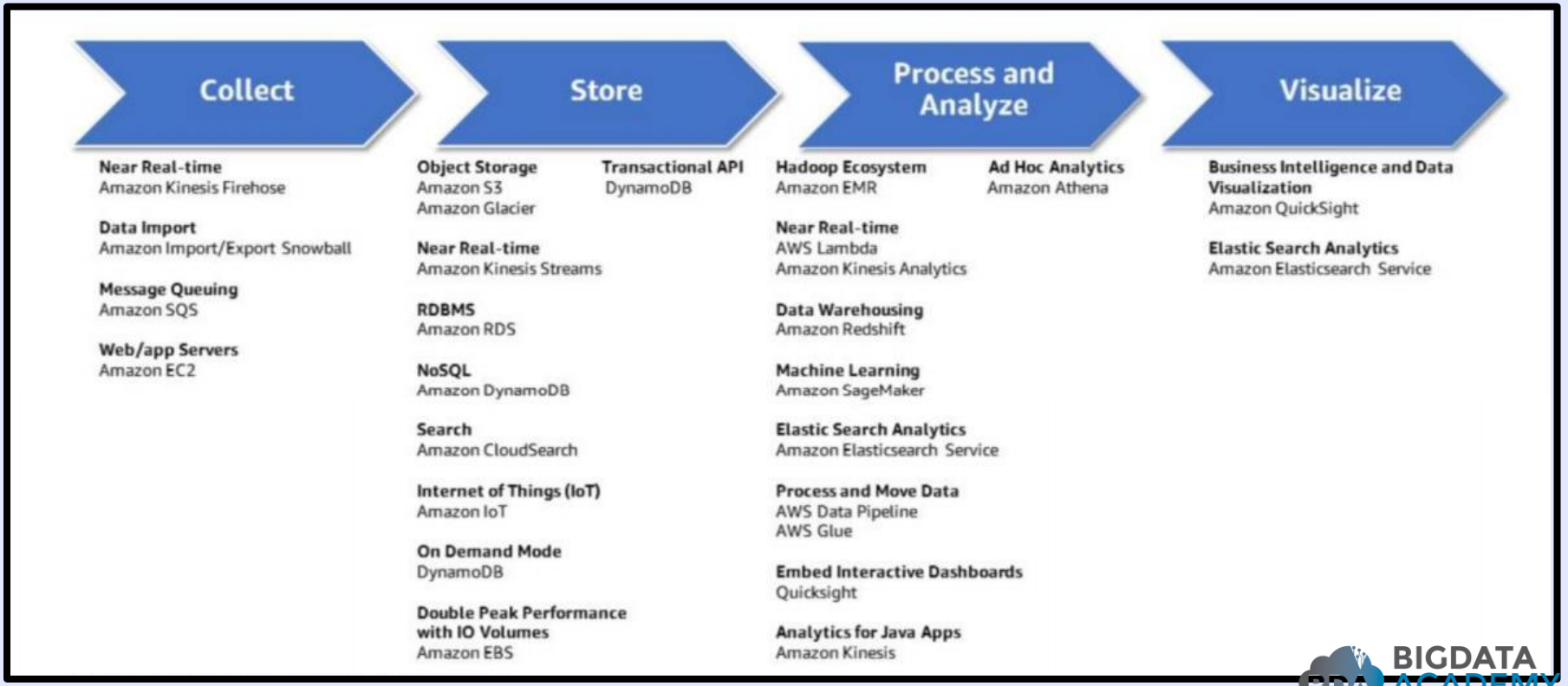
- Redes sociales
- Servicios bancarios y financieros.
- Servicios centrados en la web.
- Búsquedas científicas y de documentos
- Registros médicos
- Registros web

BIGDATA ACADEMY

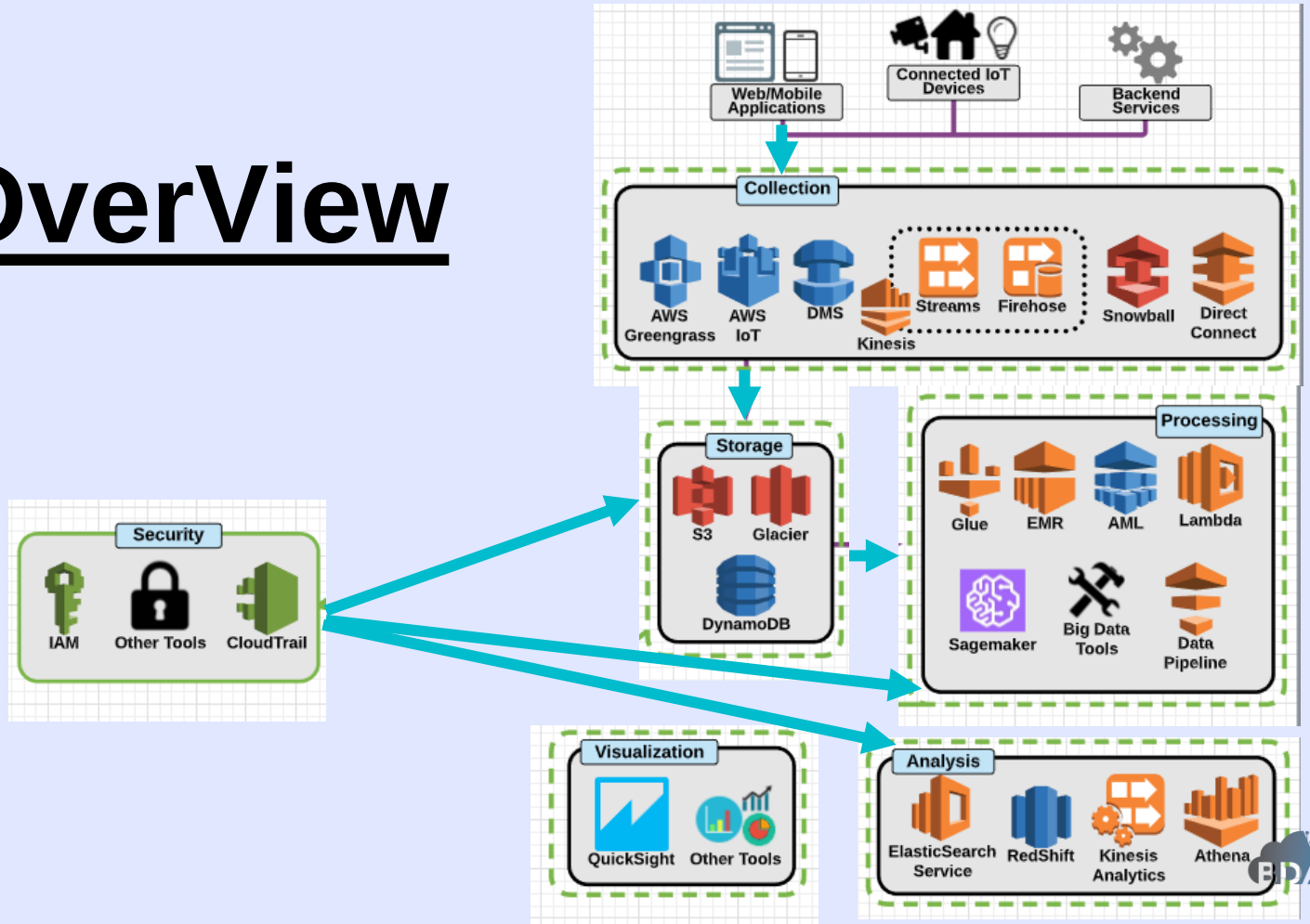
El Big Data “Pipeline”

BIGDATA ACADEMY



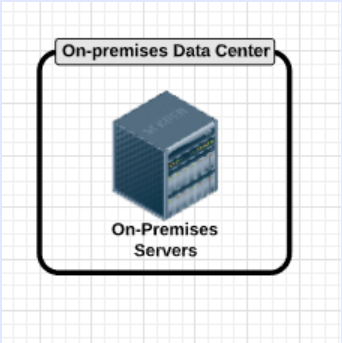


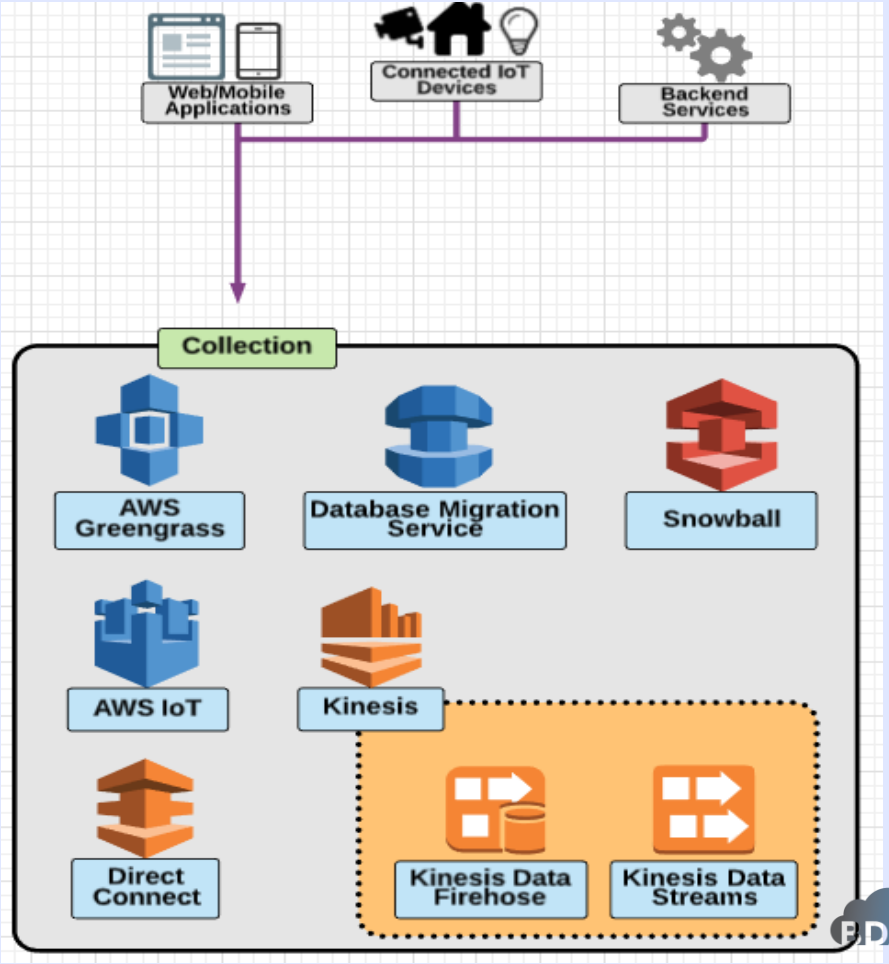
“OverView
”
—



BIGDATA ACADEMY

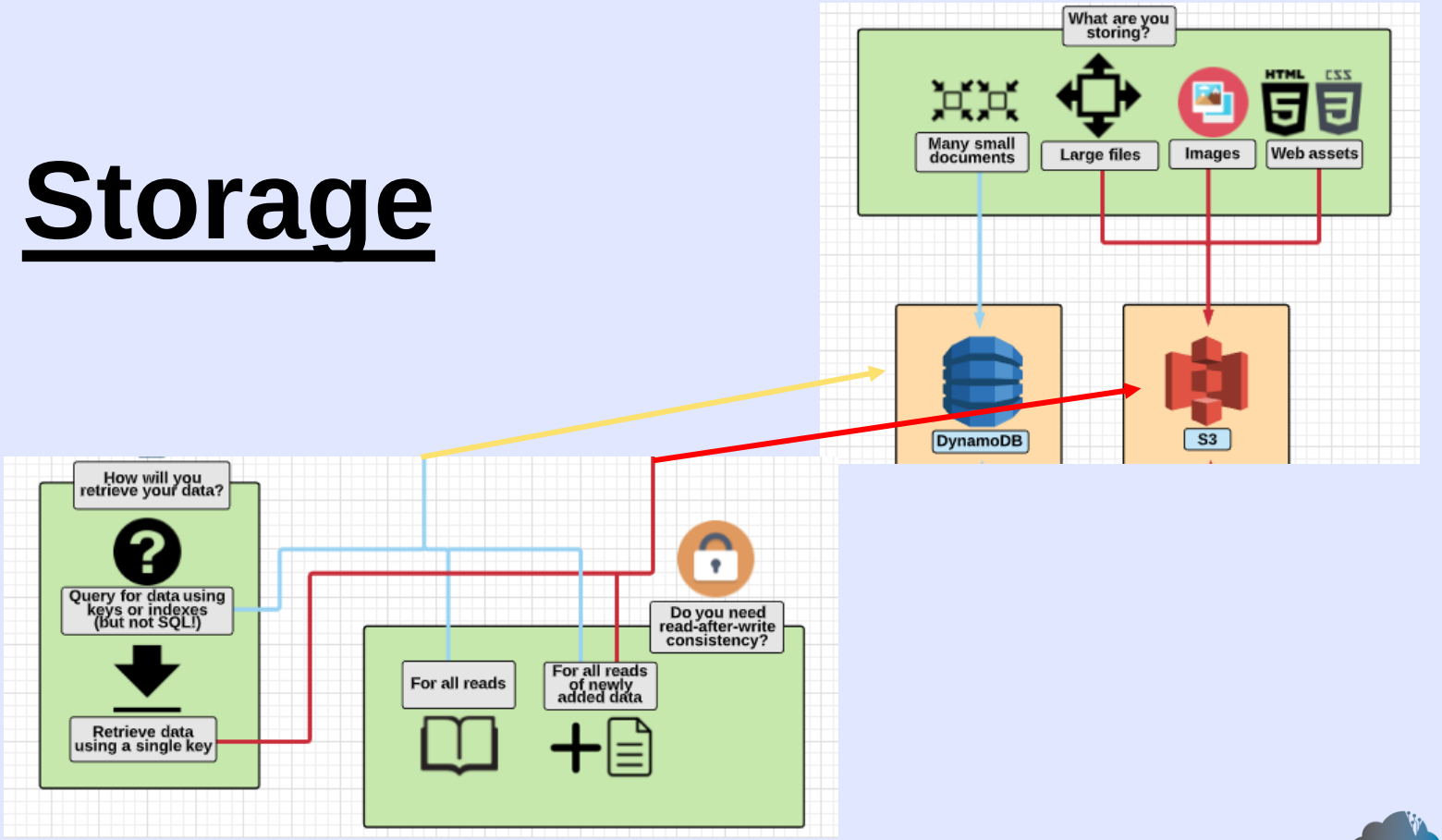
Colección





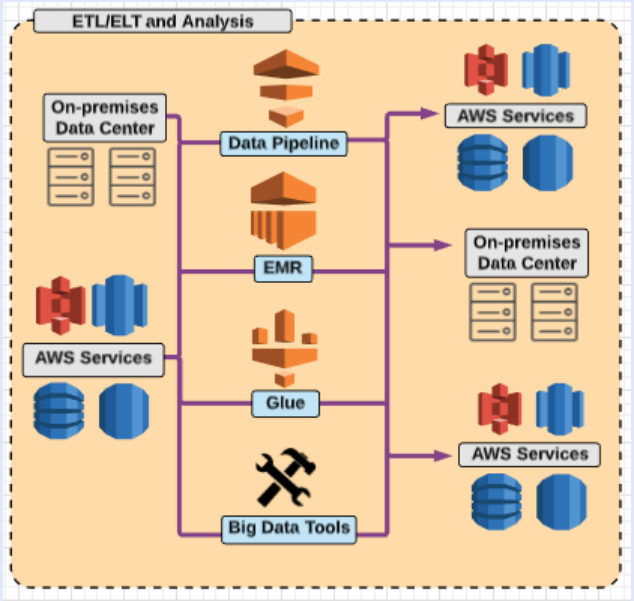
BIGDATA ACADEMY

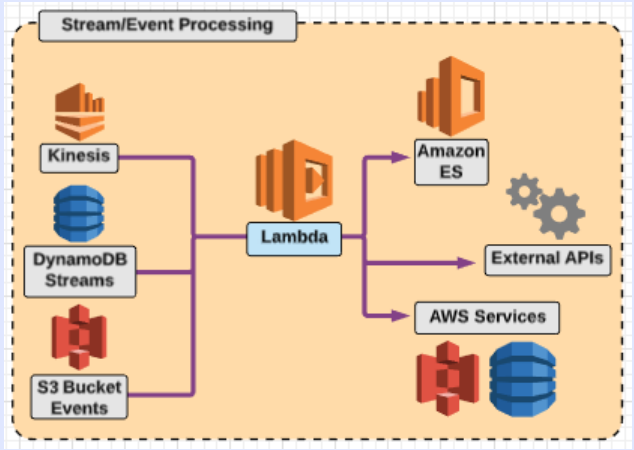
Storage

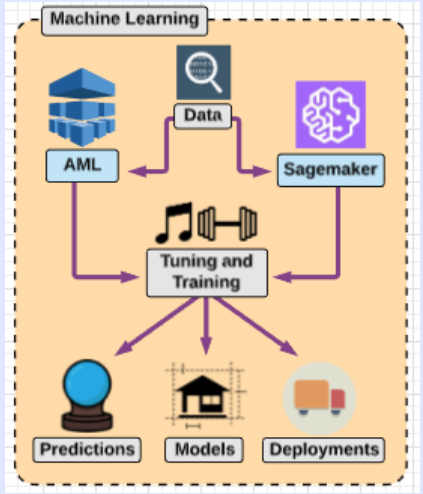


BIGDATA ACADEMY

Procesamiento

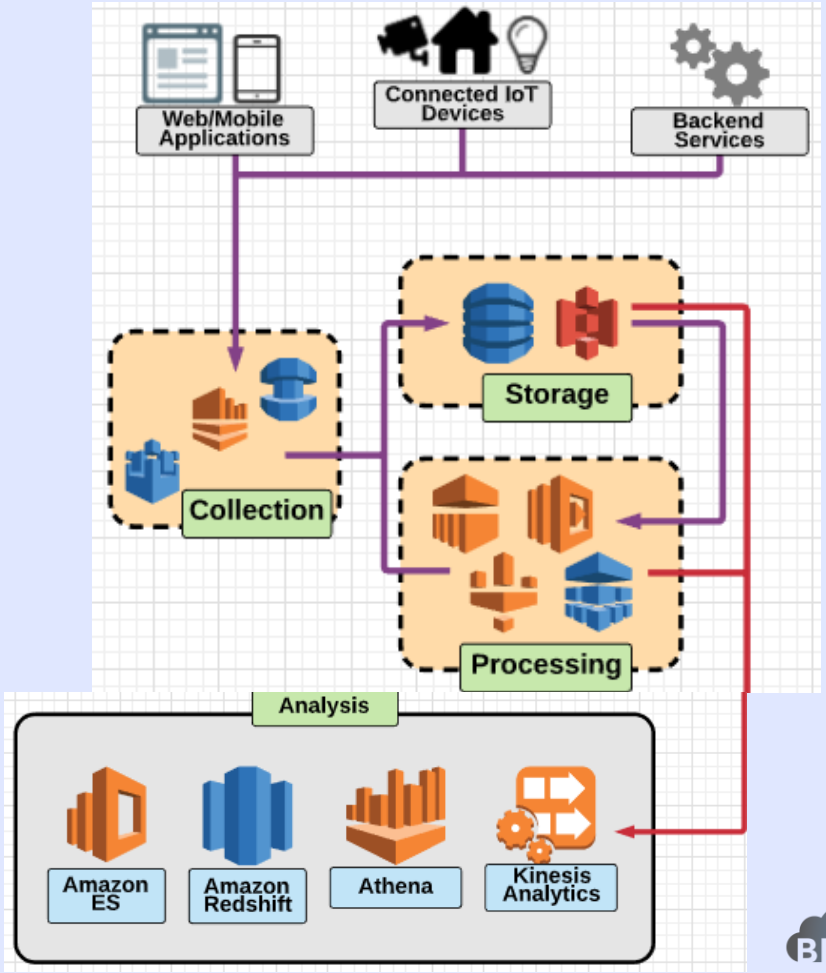






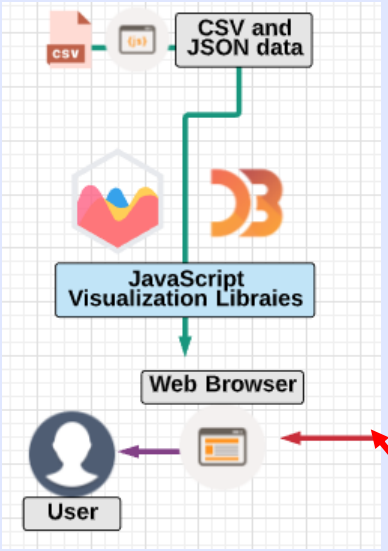
BIGDATA ACADEMY

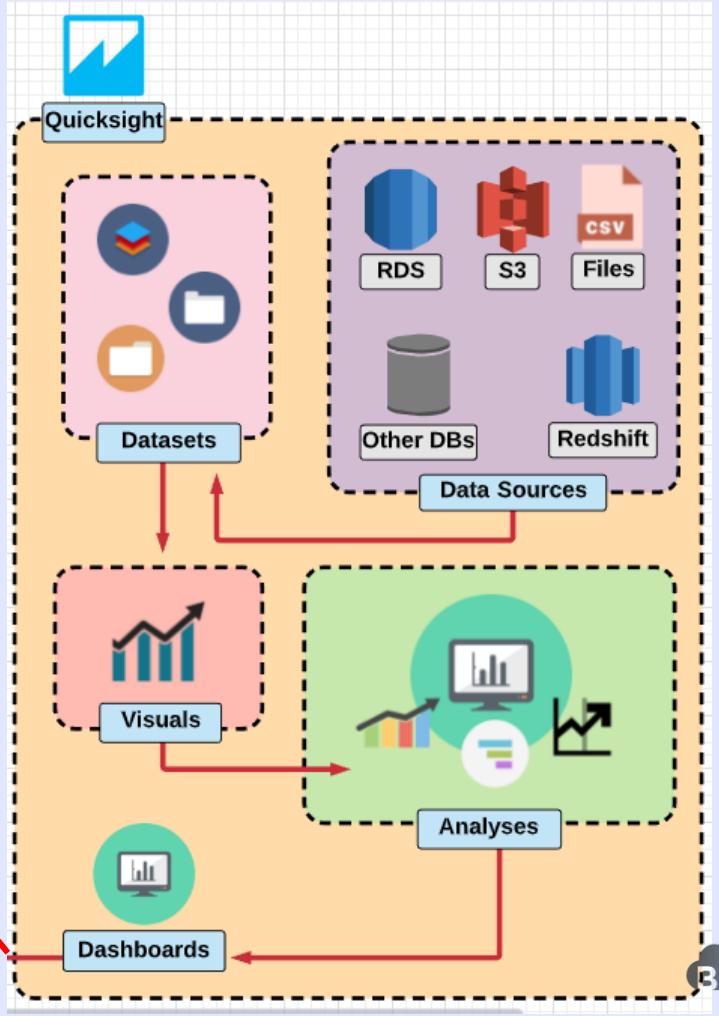
Análisis



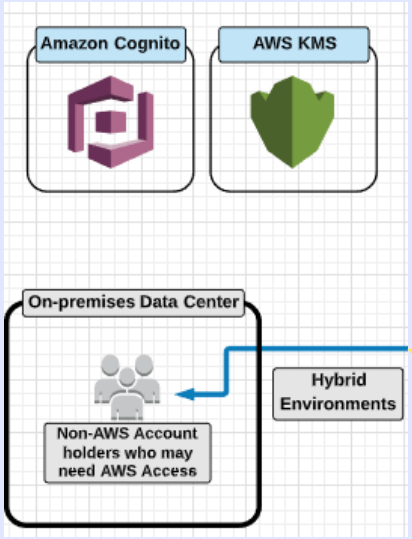
BIGDATA ACADEMY

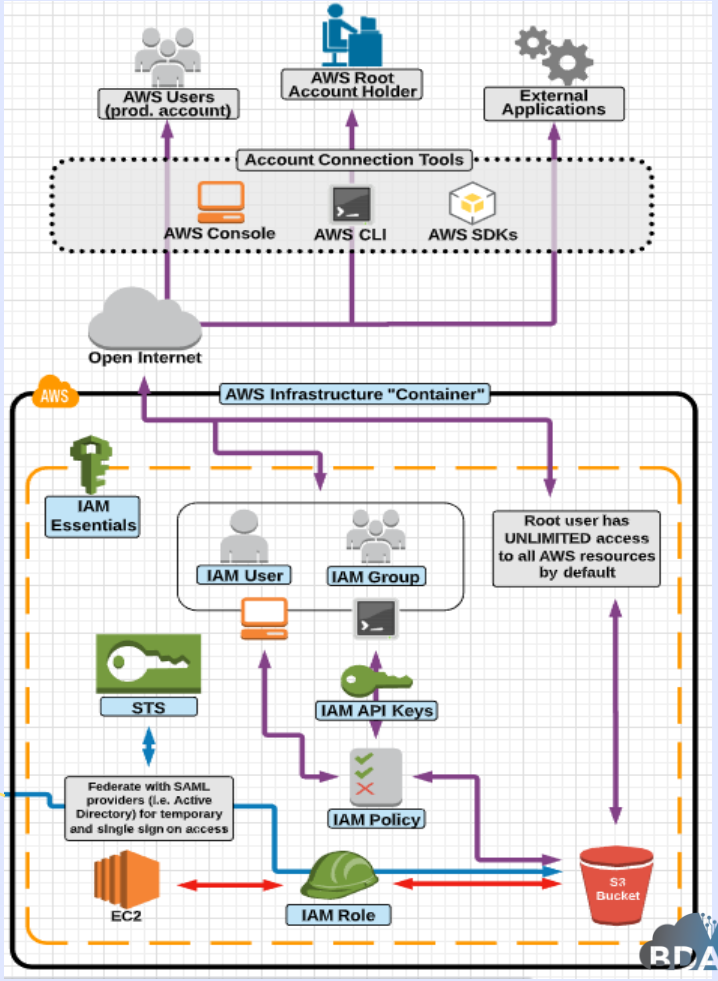
Visualización





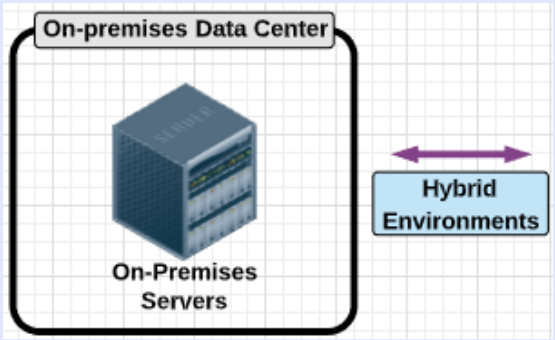
Seguridad

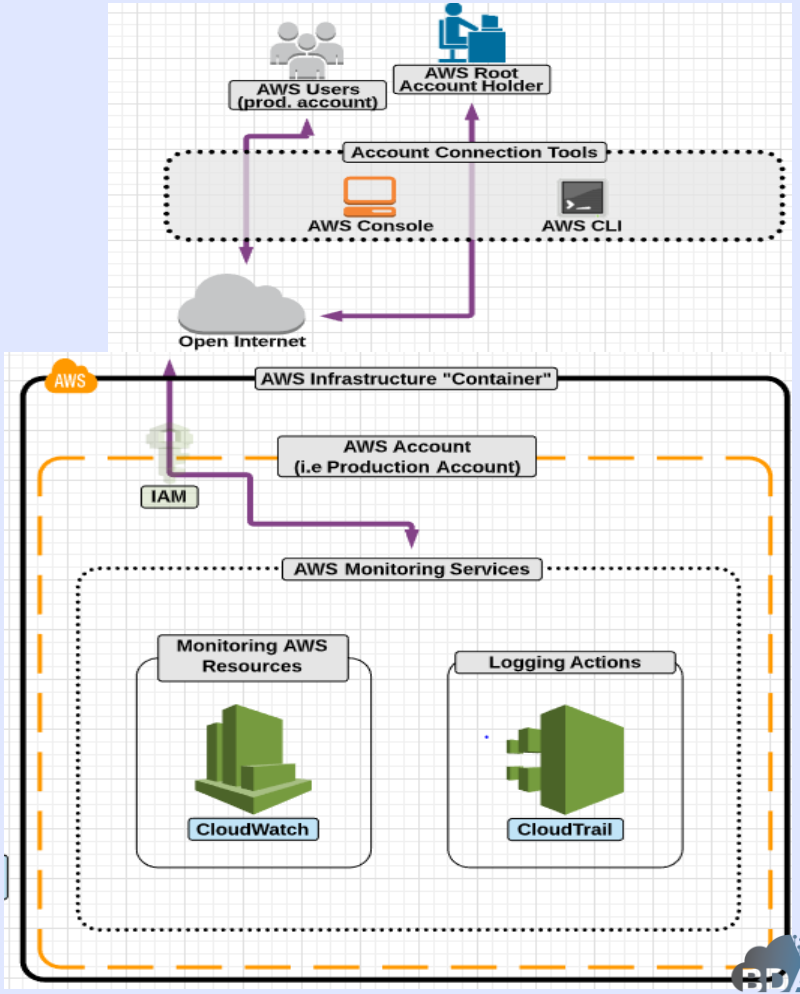




BIGDATA ACADEMY

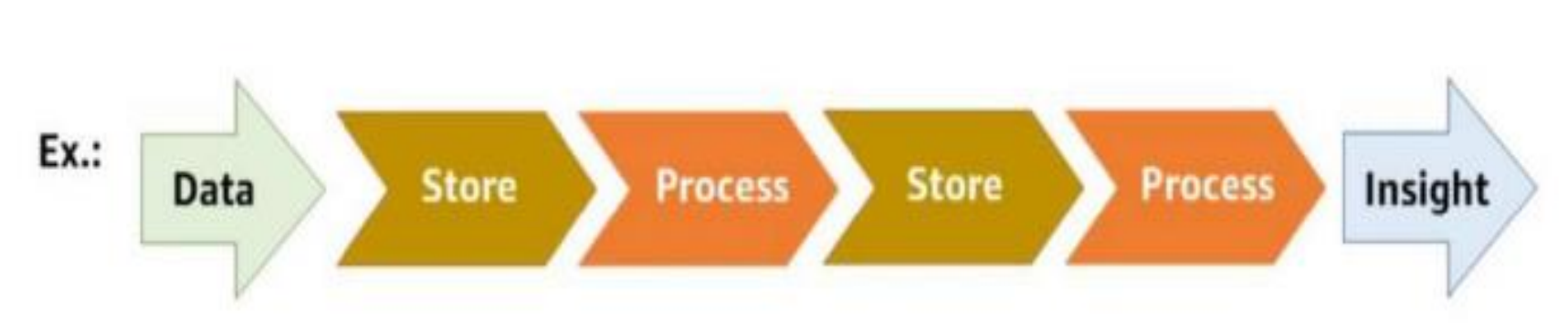
Monitoreo





BIGDATA ACADEMY

Desacoplar su bus de datos

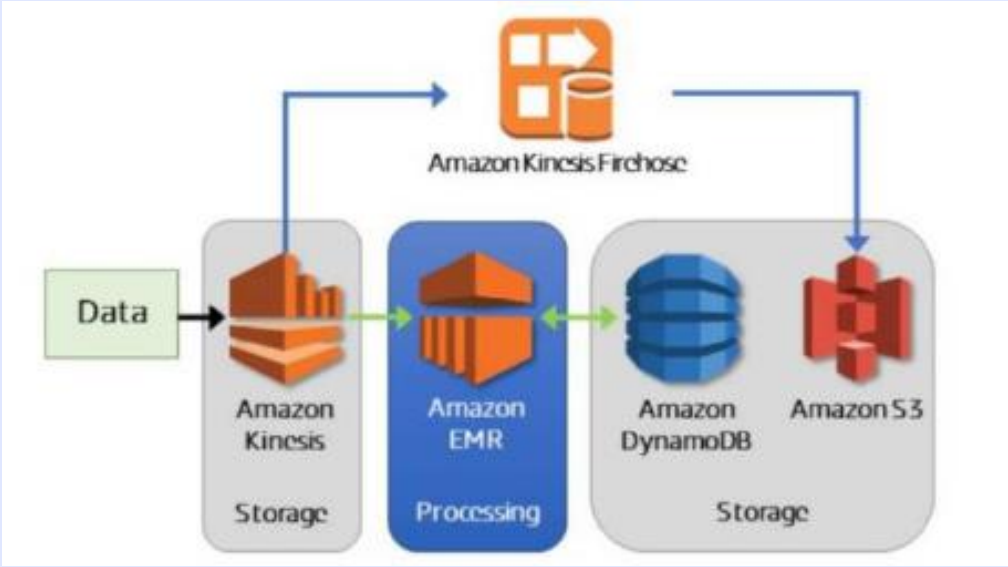


Haga coincidir la herramienta con la tarea requerida.
esto da como resultado una arquitectura más tolerante a fallas y generalmente resulta en un tiempo mejorado para responder

BIGDATA ACADEMY

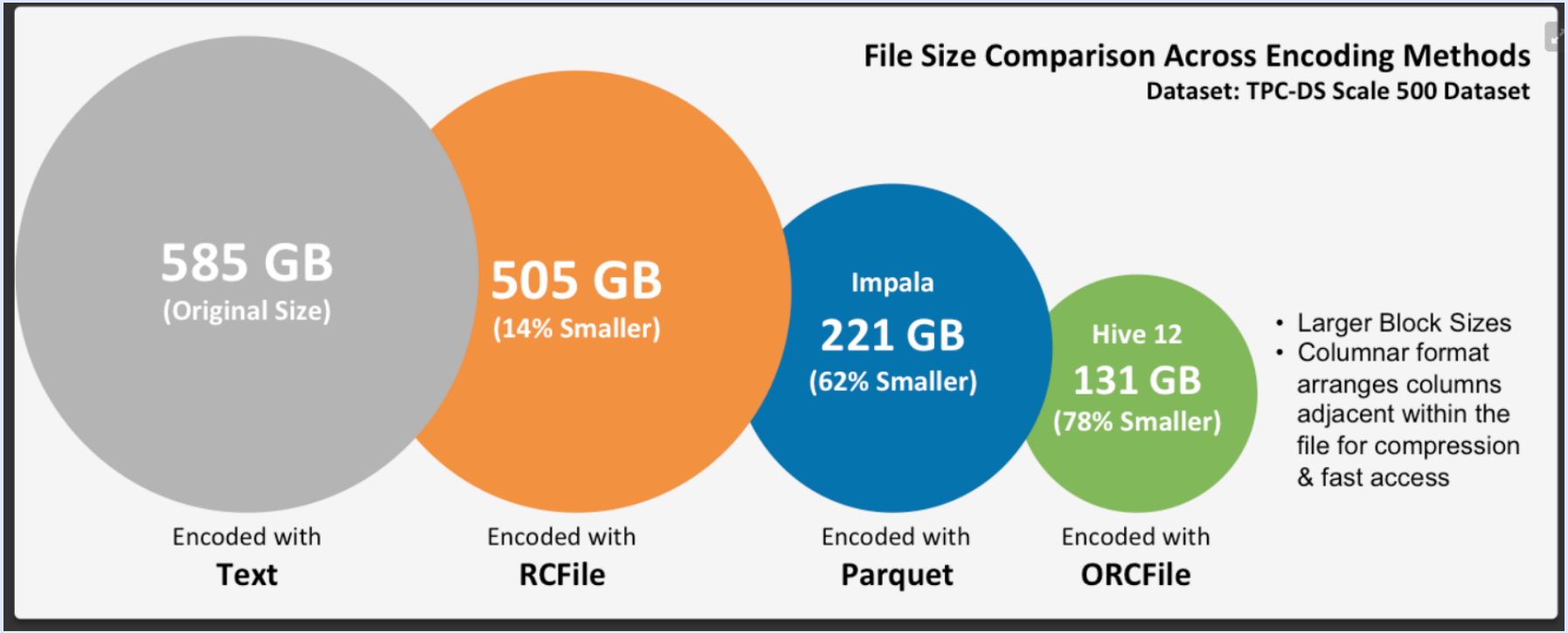
desacoplar su bus de datos: Flexibilidad

con un bus de datos desacoplado, puede hacer que multiples aplicaciones de procesamiento lean o escriban en múltiples almacenes de datos



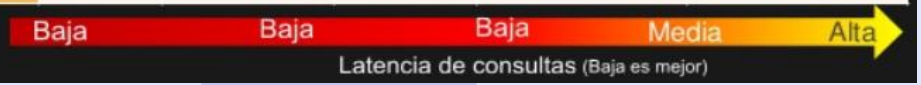
Data Lake options on AWS

Storage	Processing	AWS-Managed Solution	Vendor-Managed
HDFS	Spark	AWS EMR (HDFS+Spark)	EC2 + Vendor Solution
S3	Spark	AWS EMR (Spark Only)	EC2 + Vendor Solution
S3	Serverless	AWS Athena	Serverless + Vendor Solution


File Size Comparison Across Encoding Methods
Dataset: TPC-DS Scale 500 Dataset
585 GB
(Original Size)
505 GB
(14% Smaller)
Impala
221 GB
(62% Smaller)
Hive 12
131 GB
(78% Smaller)
• Larger Block Sizes
• Columnar format arranges columns adjacent within the file for compression & fast access
Encoded with
Text
Encoded with
RCFile
Encoded with
Parquet
Encoded with
ORCFile

¿Qué tecnología de procesamiento de datos debo usar?

	Amazon Redshift	Impala	Presto	Spark	Hive
Latencia de consultas	Baja	Baja	Baja	Baja	Media (Tez) – Alta (MapReduce)
Durabilidad	Alta	Alta	Alta	Alta	Alta
Volumen de datos	1,6 PB máx.	Nodos aprox.	Nodos aprox.	Nodos aprox.	Nodos aprox.
Administrada	Sí	Sí (EMR)	Sí (EMR)	Sí (EMR)	Sí (EMR)
Almacenamiento	Nativo	HDFS/S3A*	HDFS/S3	HDFS/S3	HDFS/S3
Compatibilidad con SQL	Alta	Media	Alta	Baja (SparkSQL)	Media (HQL)



MUCHA TEORIA

A PRACTICAR

Gracias