Servicios en la Nube para Big Data

Germán Moltó

Departamento de Sistemas Informáticos y Computación - Instituto de Instrumentación para Imagen Molecular

gmolto@dsic.upv.es

http://www.grycap.upv.es/gmolto

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Grid y Computación de Altas Prestaciones GRyCAP

etsinf Escuela Técnica Superior de Ingeniería Informática

i3M

Resultados de Aprendizaje

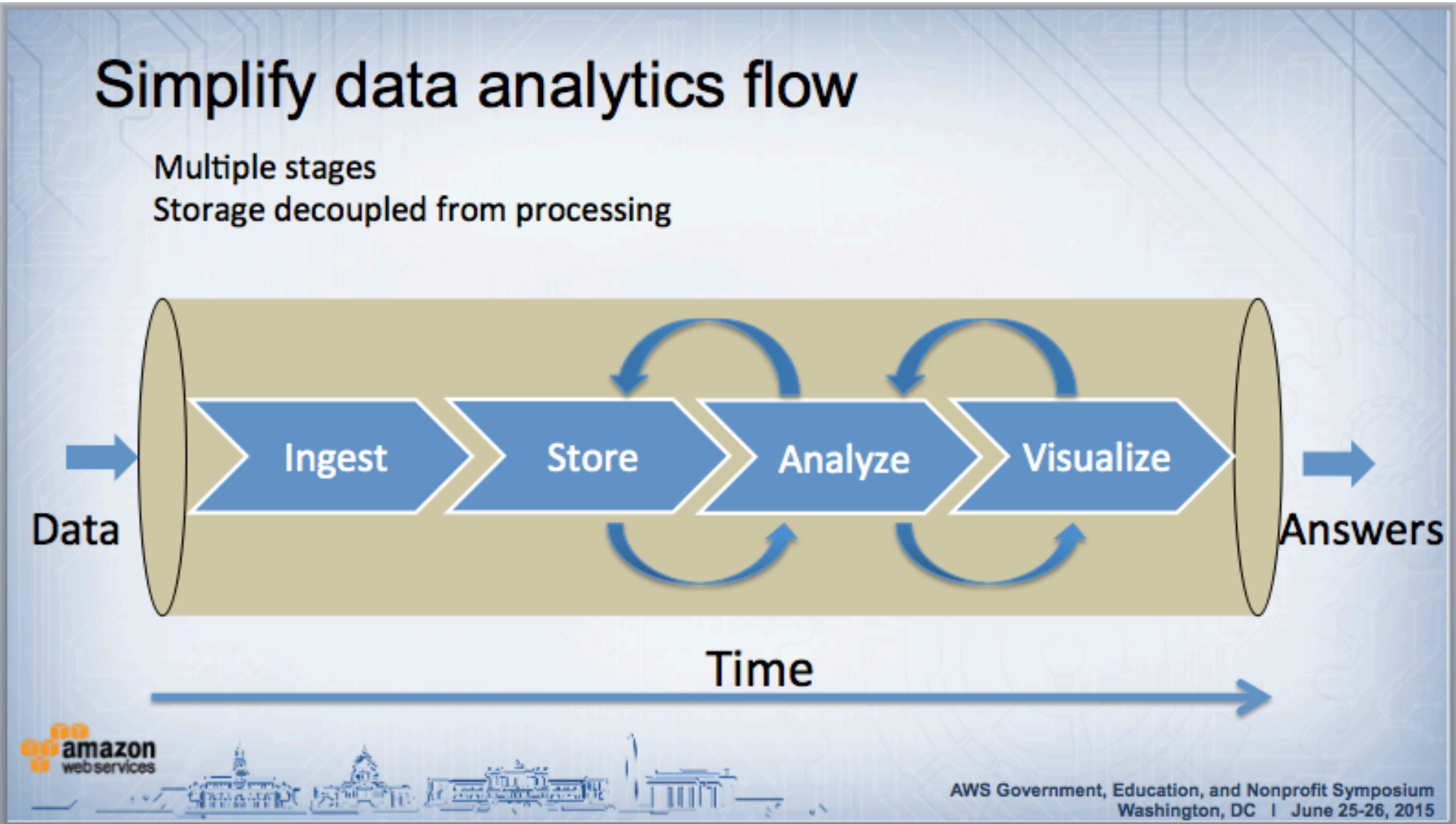
- Se espera que tras este tema seas capaz de:
  - Conocer el portfolio de servicios que ofrece AWS para aplicaciones Big Data.
  - Entender los escenarios de aplicación de estos servicios y su funcionamiento básico.
  - Comprender algunos ejemplos de uso de estos servicios para aplicaciones reales.

Big Data?

When your data sets become **so large that you have to start innovating** around how to collect, store, organize, analyze, and share it

- Velocity
  - Rate of data flow in
- Latency
  - High or Low
- Volume
  - High or Low
- Variety
  - Diversity of source data
- Item Size
  - KB or MB
- Request Rate
  - Access patterns
- Change Rate
  - How much is the data changing?
- Processing Requirements
  - How much computation?
- Durability
  - Preservation of source data?
- Availability
  - Tolerance for downtime?
- Growth Rate
  - Rate of data growth?
- Views
  - The diversity of consumers?

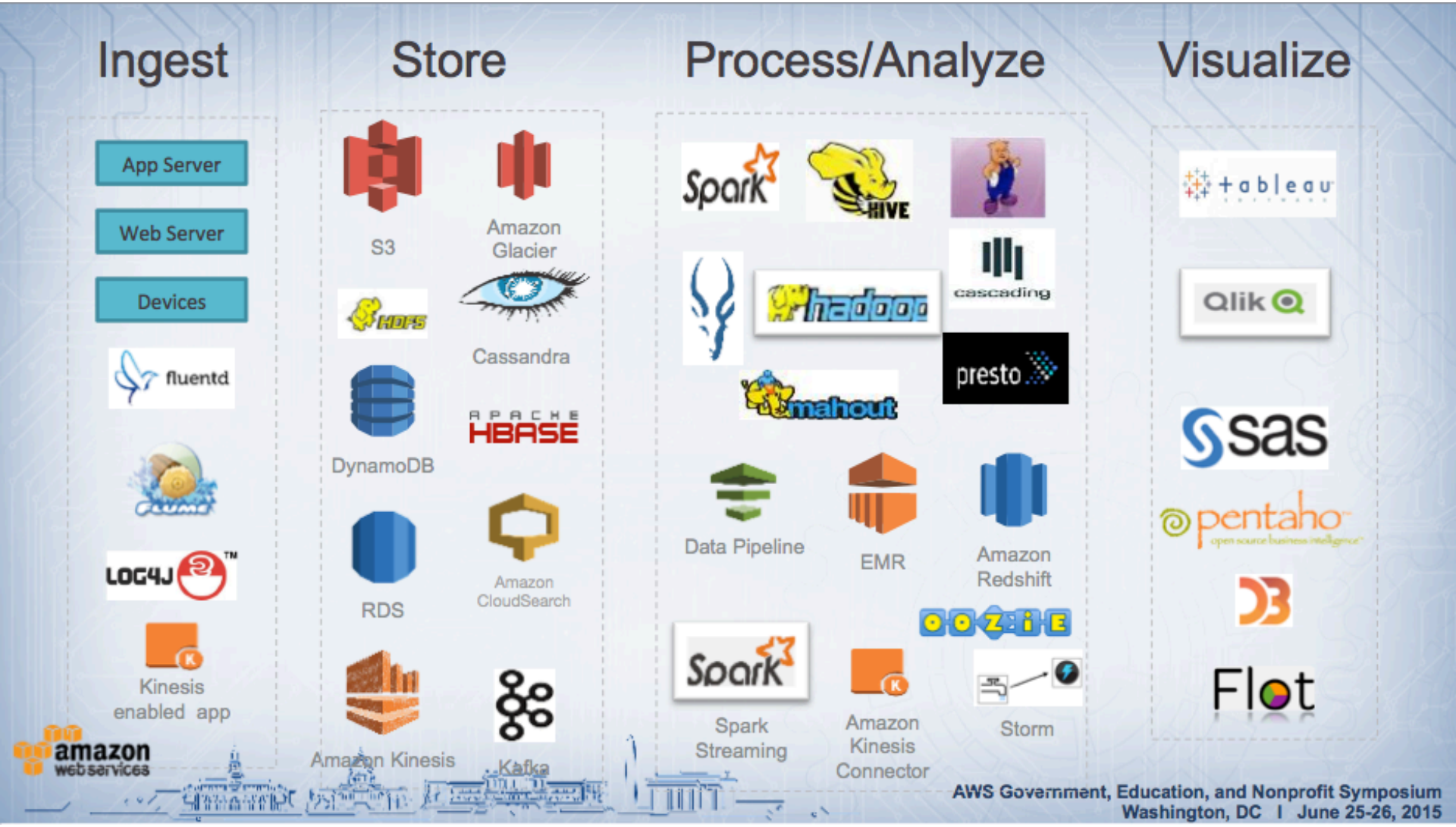
Flujo de Análisis de Datos



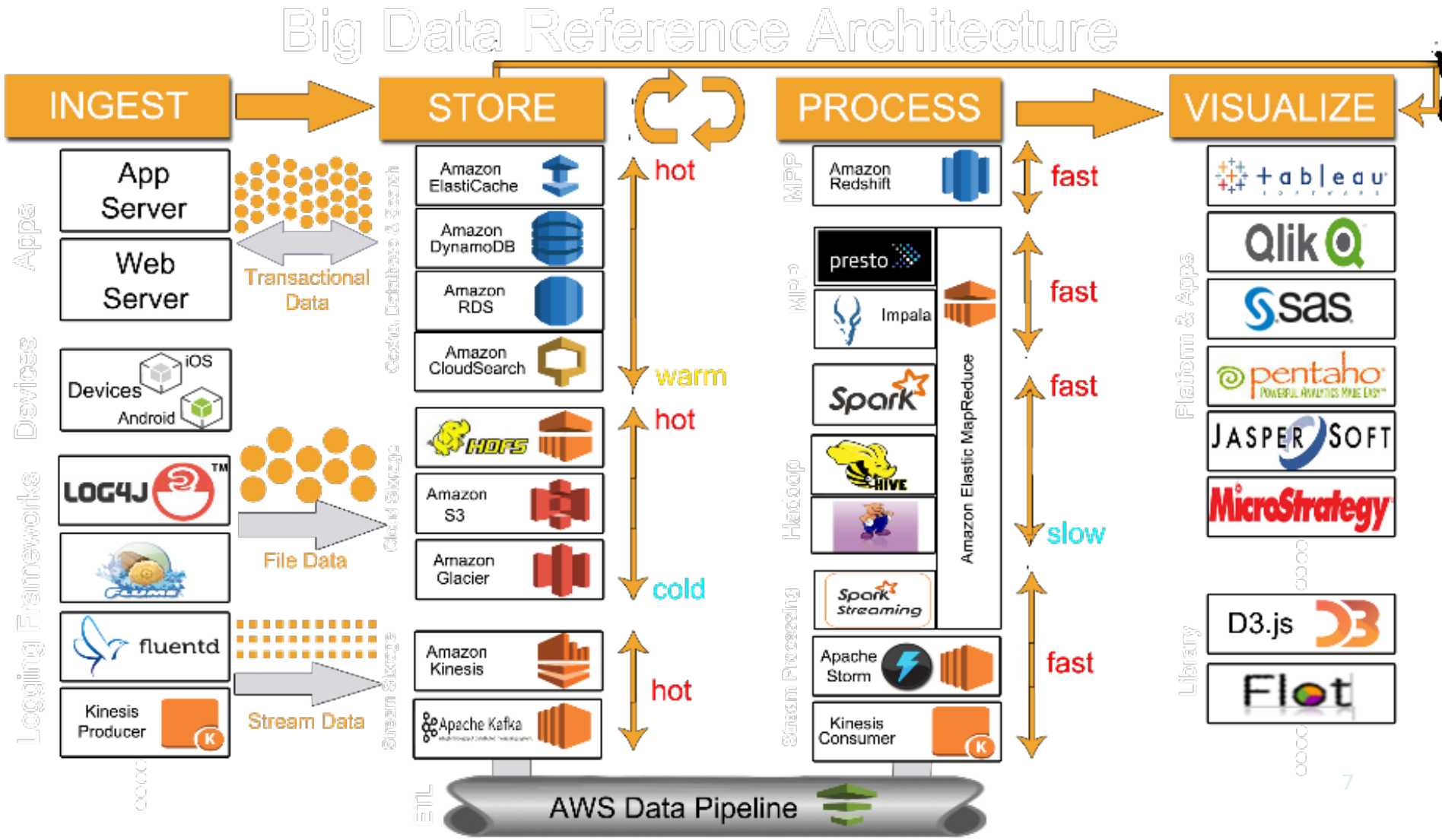
Big Data en AWS

- AWS ofrece un catálogo de servicios apropiado para las necesidades de las aplicaciones Big Data.
  - Analíticas de flujos de datos en tiempo real.
  - Data warehousing
  - NoSQL
  - Bases de datos relacionales
  - Almacenamiento de objetos (ficheros)
  - Herramientas analíticas
  - Servicios de Workflow

Ciclo de Vida en Big Data



Big Data Reference Architecture



Ventajas de Utilizar AWS para BigData

- Aprovisionamiento dinámico de capacidad de cómputo y almacenamiento en función de las necesidades de la aplicación Big Data.
- Diferentes regiones geográficas para construir arquitecturas escalables y altamente disponibles.
- Servicios disponibles para Big Data
  - Amazon RedShift
  - Amazon Kinesis
  - Amazon Elastic MapReduce
  - Amazon DynamoDB
  - Aplicaciones sobre Amazon EC2.

Amazon Redshift

- Amazon Redshift es una solución de Data Warehouse totalmente gestionado para almacenar datos en la escala de Petabytes, conectable con herramientas de inteligencia de negocio (*business intelligence)*.
  - Usa almacenamiento columnar (basado en PostgreSQL) y distribución entre múltiples nodos.
  - Apropiado para OLAP (*Online Analytical Processing*).
    - Pocas consultas complejas con agregaciones (*Data Mining*).
  - Consultable con SQL e integrable con herramientas como:
    - Pentaho
    - IBM Cognos, etc: http://aws.amazon.com/es/redshift/partners/

Amazon Kinesis

- Permite procesar flujos de datos en tiempo real.
  - Las prestaciones del flujo (MB/s) son configurables y los datos se mantienen durante una ventanas de 10 segundos para ser analizados (o almacenados en Amazon S3 o Amazon RedShift).
  - Los datos se almacenan temporalmente replicados en múltiples AZ pero no está orientado a almacenamiento permanente.
- Casos de uso
  - Analíticas de datos en tiempo real.
    - Flujos de clicks en webs.
  - Procesado de logs en tiempo real.
  - Informes a partir de métricas en tiempo real.
    - Actualización continua de paneles de control
- Integración para leer de un stream de Kinesis hacia Apache Storm

Amazon EMR

- Amazon Elastic MapReduce permite el despliegue de clusters Hadoop redimensionables para el procesamiento de datos batch.
- Soporte a la distribución MapR
  - No-NameNode Architecture para eliminar SPOFs.
- Herramientas
  - Hive, Pig, Spark, Hbase, Impala, Hunk, Hue, Mahout, Ganglia, R, etc.
- Casos de uso
  - Procesado y análisis de logs
  - ETL (Extract, Transform, Load)
  - *Ad targeting*
  - Analíticas predictivas.

Amazon DynamoDB

- Servicio de base de datos NoSQL
  - Tablas sin esquema fijo, almacenamiento en SSD.
  - Latencia de acceso inferior a 9 milisegundos y prestaciones predecibles.
  - Índices primarios y secundarios (para consultas a partir de otros atributos que no sean la clave primaria).
- Casos de uso:
  - Aplicaciones móviles
  - Juegos
  - Votación en vivo
  - Redes de sensores

Aplicaciones Sobre Amazon EC2

- Aparte de los servicios gestionados, puedes optar por desplegar la infraestructura y gestionar la configuración con una herramienta de procesado de Big Data
  - Instancia(s) EC2 + MongoDB
  - Cluster Hadoop en EC2
  - Cluster Storm en EC2
  - …
- Usuario responsable de gestionar y escalar la infraestructura.

Opciones de Big Data en AWS

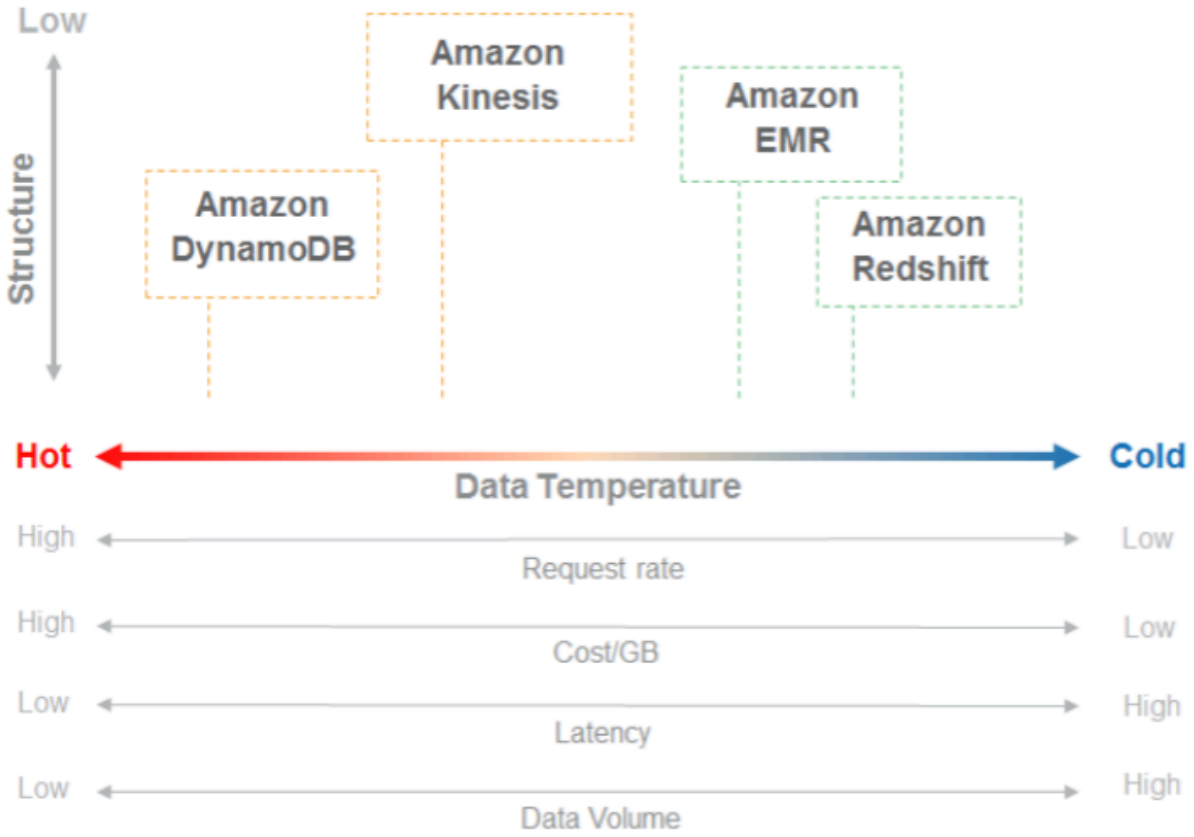


Figure 1: Big Data Analytics Tools on AWS

Big Data Analytics Options on AWS.
https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf

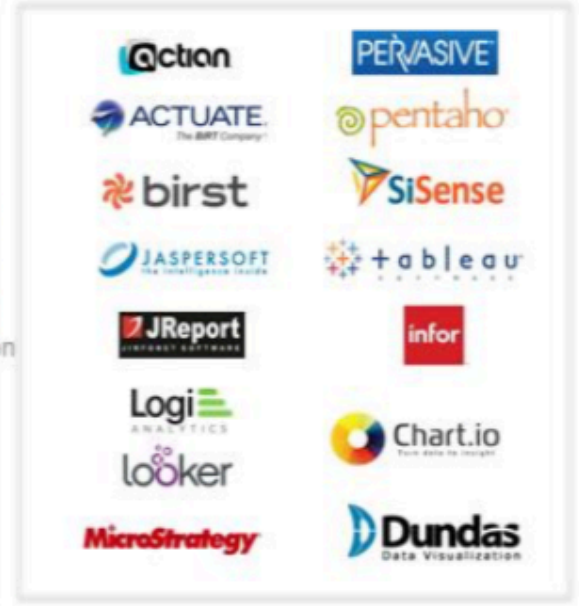
¿Qué almacén de datos utilizar?

	Amazon ElastiCache	Amazon DynamoDB	Amazon RDS	Amazon CloudSearch	Amazon EMR (HDFS)	Amazon S3	Amazon Glacier
Average latency	ms	ms	ms, sec	ms,sec	sec,min,hrs	ms,sec,min (~ size)	hrs
Data volume	GB	GB–TBs (no limit)	GB–TB (3 TB Max)	GB–TB	GB–PB (~nodes)	GB–PB (no limit)	GB–PB (no limit)
Item size	B-KB	KB (64 KB max)	KB (~rowsize)	KB (1 MB max)	MB-GB	KB-GB (5 TB max)	GB (40 TB max)
Request rate	Very High	Very High	High	High	Low – Very High	Low– Very High (no limit)	Very Low (no limit)
Storage cost $/GB/month	$$	¢¢	¢¢	$	¢	¢	¢
Durability	Low - Moderate	Very High	High	High	High	Very High	Very High

Big Data Analytics Options on AWS.
https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf

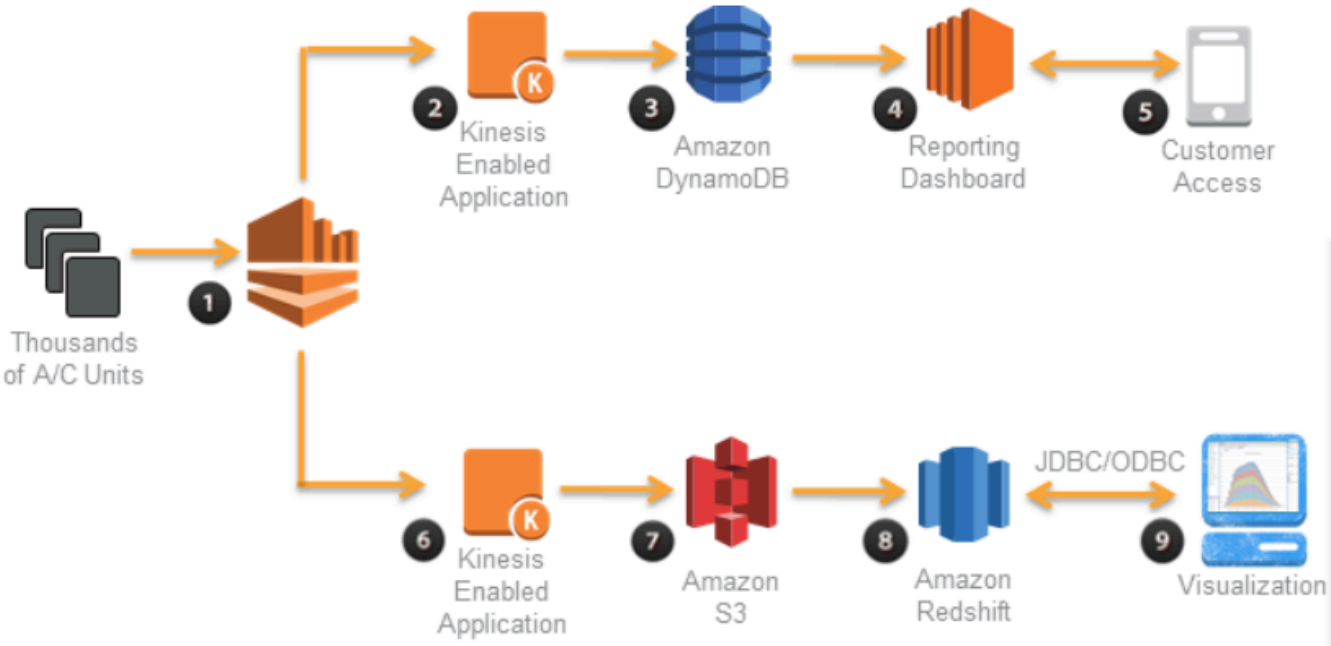
Ejemplo 1: Enterprise Data WareHouse

- Multinacional de venta de ropa donde el CEO desea una visión centralizada de ventas.



1. Carga de datos a Amazon S3 de múltiples fuentes.
2. Amazon EMR para curar los datos (cluster nocturno con Spot Instances)
3. Amazon RedShift para facilitar las consultas analíticas
4. Herramienta de visualización.

Ejemplo 2: Distribuida de A/C Inteligentes



- Usuarios pueden acceder a información de consumo y empresa a datos agregados.

Ejemplo 3: Yelp

- https://aws.amazon.com/es/solutions/case-studies/yelp/
- Inicialmente utilizaban RAIDs y un único cluster Hadoop on-premises.
  - La falta de espacio en disco y de capacidad propició el salto a AWS.
- Yelp utiliza Amazon S3 para almacenar 1.2TB diarios de registros y fotos.
- Utiliza Amazon EMR para ejecutar unos 20 batch scripts para procesar dichos logs, involucrando 250 trabajos Amazon EMR al día, procesando 30TB de datos.
  - People Who Viewed this Also Viewed,
  - Review highlights,
  - Auto complete as you type on search, Search spelling suggestions, Top searches, Ads

Conclusiones

- AWS ofrece una serie de servicios gestionados para el procesado de grandes cantidades de datos.
- También existe la opción de realizar el aprovisionamiento de recursos con Amazon EC2 y desplegar herramientas existentes.

Referencias

1. Big Data Analytics Options on AWS. https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf
2. Big Data. https://aws.amazon.com/es/big-data/
3. Big Data & HPC. Powered by the AWS Cloud. http://aws.amazon.com/es/solutions/case-studies/big-data/
4. Amazon Kinesis Storm Spout. https://github.com/awslabs/kinesis-storm-spout
5. MAPR. No-namenode architecture. https://www.mapr.com/products/m5-features/no-namenode-architecture
6. Big Data and Analytics on AWS. http://es.slideshare.net/AmazonWebServices/big-data-and-analytics-on-aws