Bases de dades i Big Data.

Activitat 2

Javier Pedragosa Lozano

Almacenamiento y procesamiento:

Explica el funcionamiento de Hadoop Distributed File System (HDFS). ¿Por qué es una tecnología clave en el almacenamiento de Big Data?

HDFS (Hadoop Distributed File System) divideix els arxius en blocs i emmagatzema cada bloc en un DataNode (node de dades). Múltiples DataNodes s'enllacen al clúster. Després el NameNode (Node de nom) distribueix rèpliques d'aquests blocs de dades per tot el clúster. També dona instruccions a l'usuari o l'aplicacio sobre com localitzar possible informació sol·licitada.

¿Qué es un Data Lake? ¿En qué se diferencia de un Data Warehouse?

Un Data Lake és un repositori centralitzat que permet emmagatzemar tota la informació estructurada i desestructurada a qualsevol escala. Es poden emmagatzemar les dades tal qual, sense haver d'estructurarles abans, i realitzar diferents anàlisis, des de panels de control i visualitzacions fins al processament de les dades, analítiques en temps real, i *machine learning* per facilitar les decisions.
Un Data Warehouse és un tipus de sistema de gestió de dades que està dissenyat per permetre i suportar les activitats de Business Intelligence, especialment anàlisis. Els Data Warehouses estan pensats solament per fer sol·licituds i anàlisis, i normalment contenen grans quantitats de dades històriques.

Compara el procesamiento por lotes (batch processing) y el procesamiento en tiempo real (streaming) en Big Data. ¿Cuando utilizarías cada uno?

Les diferències clau sobre com emmagatzemar totes les dades d'una organització es resumeixen en les següents consideracions:

1. El processament per lots és quan el processament i l'anàlisi passen en un set de dades que ja han sigut emmagatzemats durant un temps. Un exemple són el sistemes de pagament i rebuts que han de ser processats semanal o mensualment.
2. El processament en temps real passa quan les dades flueixen a través d'un sistema. Això dona com a resultat anàlisis i informes d'events tan aviat com passen. Un

exemple podria ser la detecció de frau o d'intrusos. El processament en temps real significa que les dades són analitzades i les accions es prenen sobre les dades donat un temps molt curt o fins i tot en temps real.

Les principals diferències es poden ver de manera molt clara en la següent graella:

	PROCESSAMENT PER LOTS	PROCESSAMENT EN TEMPS REAL	TRANSMISSIÓ DE DADES
Maquinari	És la que més recursos d'emmagatzematge i processament necessita per poder processar grans lots de dades	Menys emmagatzematge necessari per processar el set de dades actual o més recent. Menys requisits computacionals	Menys emmagatzematge necessari per processar el set de dades actual. Més recursos de processament per «continuar desperta» per poder complir amb les garanties del temps real
Rendiment	La latència pot ser de minuts, hores o dies	La latència necessita ser de segons o milisegons	La latència ha de ser si o si de milisegons
Set de dades	Grans lots de dades	El paquet de dades actual o uns quants d'ells	Corrent contínua de dades
Anàlisi	Computació complexa i anàlisi amb un marge de temps major	Computació o informes simples	Computació o informes simples

Tecnologías de Big Data:

¿Qué es MapReduce y cómo facilita el procesamiento de grandes volúmenes de datos?

MapReduce és un framework basat en Java d'execució distribuïda dins de l'ecosistema d'Apache Hadoop. Simplifica la complexitat de la programació distribuïda exposant dos pasos de processament que els desenvolupadors implementen: Mapejar i Reduïr.
Al pas de Mapejament, les dades es divideixen entre tasques de processament paral·leles. Després a cada grup de dades se li pot aplicar una transformació. Una vegada completat, la fase de Reducció pren el relleu per gestionar l'adició de dades del Map set.

Explica brevemente el rol de las siguientes tecnologías en el ecosistema de Big Data:

- **Apache Spark:** és el motor per executar enginyeria de dades, ciència de dades i *machine learning* als màquines mononode o clústers.
- **Apache Kafka:** s'utilitza per construir conductes i aplicacions de transmissió de dades en temps real.
- **MongoDB:** facilitza els desenvolupadors l'emmagatzematge de dades estructurades o desestructurades.

¿Qué es Apache Hive y cómo ayuda en el análisis de Big Data?

Apache hive és un sistema d'emmagatzematge de dades distribuït i tolerant amb els errors que permet la realització d'anàlisis a escala massiva. Apache Hive està dissenyat per gestionar amb rapidesa petabytes de dades mitjançant el processament per lots, proporciona una interfaç familiar (semblant a la de SQL) a la que es pot accedir fins i tot sense ser programador, i és fàcil de distribuïr i escalar en funció de les necessitat.

Bases de Datos y Herramientas:

Explica la diferencia entre las bases de datos Cassandra y HBase. ¿Cuáles son sus principales características y en qué situaciones se utilizaría cada una?

Ambdues són bases de dades basades en NoSQL que emmagatzemen les dades de forma no tabular.
Cassandra proporciona un rendiment de lectura i escriptura superior a HBase, però HBase proporciona una consistència de dades major.
Cassandra és idoni per casos que requereixen escriptura de dades freqüents, per exemple sistemes de missatgeria, dades interactives, solucions de processament i emmagatzematge de dades de sensors en temps real.
HBase és millor per aplicacions que requereixen consistència i processament freqüent, com per exemple en l'àmbit bancari, sanitari i les telecomunicacions.

¿Cuál es el papel de Elasticsearch en la cerca y análisis de grandes volúmenes de datos?

Elasticsearch és un motor d'anàlisi que permet emmagatzemar, buscar i analitzar grans quantitats de dades ràpidament i quasi en temps real i donar resposta en milisegons. És capaç d'aconseguir respostes ràpides en la cerca perquè en comptes de buscar el text directament, busca els índexs. Utilitza una estructura basada en documents en comptes de taules i esquemes i inclou APIs molt extenses per emmagatzemar i cercar les dades.