ApuntsTema1.pdf

onafolch

Desenvolupament d'Aplicacions de Dades Massives

3º Grado en Ingeniería de Datos

Escuela de Ingeniería
Universidad Autónoma de Barcelona

BIG DATA

DATA WAREHOUSING

1. OLTP i OLAP

Inicialment les operacions eren transaccions comercials, com per exemple fer una venda, fer una comanda o pagar el sou a un empleat. Però les aplicacions es tornen més complexes i la definició de transacció passa a ser més genèrica, com per exemple un comentari en un blog, una acció en un joc o el funcionament local de NFC.

OLTP (OnLine Transaction Processing) és un sistema de tractament de dades que s'utilitza per gestionar un gran nombre de transaccions curtes en línia. Permet l'execució en temps real d'un gran nombre de transaccions de bases de dades per part d'un gran nombre de persones, normalment a través d'Internet. Admet un processament molt ràpid, amb temps de resposta mesurat en ms. Proporciona conjunt de dades indexats per a una cerca ràpida.

Amb OLTP, les aplicacions cerquen un nombre reduït de registres, i aquests s'insereixen o s'actualitzen. Les dades es carreguen cada cop que es fa una transacció.

Propietats d'una base de dades per garantir la fiabilitat de les transaccions (**Transactional workloads**):

- **Atomicitat**: cada transacció es tracta com una única unitat, la qual triomfa o fracassa completament.
- **Consistència**: les transaccions només poder agafar dades de la base de dades d'un estat vàlid a una altre.
- **Aïllament**: l'execució simultània de transaccions deixa la BD en el mateix estat.
- **Durabilitat**: si es confirma una transacció, el resultat d'aquesta és definitiu.

Exemples: cançons preferides a Spotify, notes acadèmiques, cerca de gens NCBI per nom de gen, últimes 10 operacions del meu compte bancàri.

Si volem obtenir anàlisis de les dades, hem d'utilitzar **OLAP** (OnLine Analytic Processing). Escaneja un gran nombre de registres, llegeix poques columnes per registre, calcula estadístiques d'agregació (ja que les consultes sovint son molt complexes i necessiten agregacions) i no retorna dades no processades a l'usuari. Serveix per fer anàlisis de dades amb grans volums de dades. Normalment aquestes dades provenen d'un data warehouse. Les dades es carreguen periòdicament, s'agreguen i es guarden en un cub.

Analytical workloads: s'utilitzen per a l'anàlisi de dades i la presa de decisions.

- Resums
- Tendències
- Informació empresarial

Exemples: ingressos totals d'un mes, quants productes més es van vendre el passat Black Friday, quantes persones han mort per covid des de la pandèmia.

Propietat	OLTP	OLAP
Patró de lectura	Pocs registres per consulta obtinguts per clau	Agregació d'un gran nombre de registres
Patró d'escriptura	Escriptures d'accés aleatori i de baixa latència	Bulk import (ELT) o flux d'esdeveniments
Utilitzar per	Pel client, mitjançant l'applicació web	Anàlisi intern, per prendre decisions
Dades	Últim estat de les dades	Historial d'events que van passant al llarg del temps
Mida del dataset	Gigabytes to Terabytes	Terabytes to Petabytes
Tipus de dataset	Bases de dades tradicionals (DBMS)	Data warehouse

Ona Folch

El processament de dades (**Data Processing**) és la conversió de dades sense processar a informació significativa mitjançant un procés. Hi ha dos tipus:

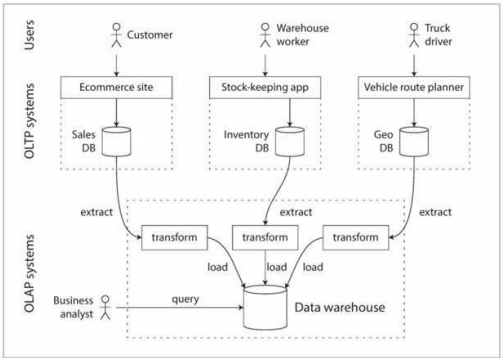
- Batch processing: les dades es recullen en grups, i en un futur es processa tot el grup com un batch.
- Stream processing: cada nova dada es processa quan arriba.

Data Warehouse

S'utilitza menys ús dels sistemes OLTP per l'anàlisi de dades, i com a alternativa s'executen les anàlisis en una base de dades independent (Data Warehouse). No afecta la BD principal, ja que 'aboca' totes les dades a una nova BD per analitzar-les, la qual està preparada per rebre preguntes intensives de dades. Concentra i emmagatzema de forma estructurada la informació obtinguda per poder analitzar-la fàcilment. Són dades que, encara que no estan disponibles a temps real, poden ser analitzades de forma ràpida i massiva sense interrompre els processos de l'usuari.

Construir Datsets per OLAP (ETL, Extract, Transform, Load)

Com crear grans conjunts de dades per l'anàlisi. Operacions principals per aplicar als datasets: extreure les dades de la BD de OLPT, transformar les dades en un esquema per facilitar l'anàlisi, netejar i filtrar errors i carregar les dades al warehouse.



Separem les dades del warehouse amb les de OLTP per optimitzar el procés d'anàlisi, i perquè els algoritmes d'indexeció de OLTP no són adequats per consultes analítiques. Proveïdors de data warehouse: Teradata, Vertica, SAP HANA, ParAccel, AWS Redshift.

Amazon Redshift: Sistema de gestió i consulta de BD relacionals i data warehouse de classe empresarial. Operacions de consulta que recuperen, comparen i avaluen grans quantitats de dades. L'emmagatzematge i el rendiment està optimitzat (processament massiu paral·lel).

The data jouney: Primer de tot s'han d'obtenir les dades i importar-les a una base de dades (**Data Ingestion**), un cop les tenim a la BD les processem i les convertim en un format més significatiu amb ETL o ELT (**Data Processing**), i finalment creem representacions gràfiques de la informació (**Data Visualization**).

Serveis d'azure pel data warehouse:

- Azure Data Factory: és un servei per la integració i transformació de les dades. Recupera dades de més d'una font d i les filtra per treure el soroll i quedar-se amb la informació important. Es va executant a mesura que es reben les dades.
- Azure Data Lake: és un repositori de dades, el qual organitza les dades en directoris per millorar l'accés als fitxers.
- Azure Databricks: és una plataforma basada en apache spark que proporciona processament i transmissió de big data.
- Azure HDInsight: és un servei de processament de big data que ens permet utilitzar biblioteques de codi obert en una plataforma, en un entorn Azure.

Ona Folch

2. DATA SYSTEM ROLES

Administrador de la BD: gestiona la BD, implementa còpies de seguretat de les dades, controla l'accés dels usuaris i supervisa el rendiment.
Eines comunes que utilitza:

- Azure Data Studio: Interfície gràfica per gestionar serveis de dades locals i basats en núvol (Windows, macOS, Linux).
- SQL Server Management Studio: Interfície gràfica per gestionar serveis de dades locals i basats en núvol (Windows).
- Azure Portal/CLI: Eines per a la gestió i el subministrament d'Azure Data Services.

Enginyer de dades: processa les dades, prepara les dades per ser analitzades i du a terme l'emmagatzenament d'ingestió de dades (procés de transport de dades d'una o més fonts a un lloc objectiu per a un posterior processament i anàlisi).
Eines comunes que utilitza:

- Azure Synapse Studio: Azure Portal integrat per gestionar Azure Synapse, ingestió de dades.
- SQL Server Management Studio: Interfície gràfica per gestionar serveis de dades locals i basats en núvol (Windows).
- Azure Portal/CLI: Eines de gestió i provisió de recursos.

Analista de dades: proporciona informació sobre les dades de manera visual, modela les dades per poder-les analitzar i combina dades per visualitzar-les i analitzar-les.
Eines comunes que utilitza: (Power BI és un servei d'analítica empresarial de Microsoft)

- Power BI Desktop: eina per visualitzar dades.
- Power BI Portal/Power BI Service: per elaborar i gestionar informes de Power BI. Per compartir datasets i informes.
- Power BI Report Builder: eina per visualitzar dades, i visualitzar i modelar informes.

Ona Folch