Infraestructura de Big Data

Organización- Ecosistemas

- Uso analítico

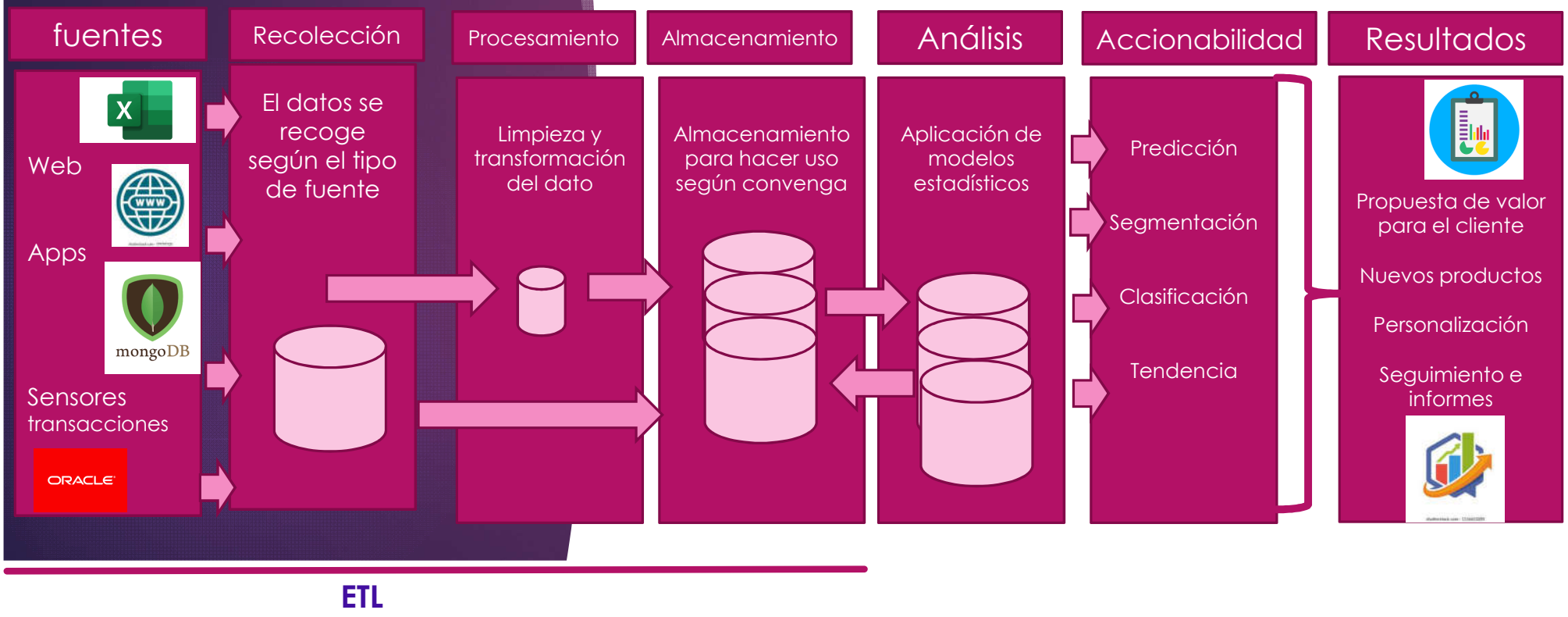
- Nuevos productos

Organización-Ecosistemas

- Funcionamiento de sus elementos
- Relaciones entre ellos
- Que analytics se va a utilizar
- Fases para obtener datos para gestionar la empresa
- Tecnologías de Machine learning y Deep learning

Como funciona Big Data

- A nivel técnico debe seguir una estructura de procesos que pasa por la recolección, almacenamiento y análisis de los datos
- Nivel de análisis y de explotación



Procesos ETL

Extract

- Obtención de datos de un sistema de origen
- Validación, completitud, solidez
- Preparación para iniciar la transformación
- Decisión de que hacer con ellos

Transformation

- Reglas de negocio:
  - claras
  - entendibles
  - independientes
- Finalidad y utilidad para el negocio
- Evitar procesos excesivos,
- Comparaciones innecesarias, operaciones inútiles

Load

- Sistema de destino
- Depende de los requerimientos
- Informe de finalización, errores, tiempo de carga, comparación de tablas, cuantos se insertaron o fallaron.
- Decisión de avanzar

Procesos ETL

Funcionalidad

- Mover datos de múltiples fuentes (internas, externas) y diferentes formatos (txt, Oracle, postgre, mariadb, cvs, etc).
- Permite realizar de manera ágil y controlada migración de datos de una tecnología a otra (ej. Win a web)
- Formatear y depurar los datos antiguos, que no son compatibles con estructuras actuales

Procesos ETL

ETL

- Extracción
- Transformación
- Carga

ELT

- Extraer datos de las fuentes
- Realizar agregaciones
- Cruces
- Cargar datos al Datawarehouse

ETLT

- Permite ajustar las tecnologías y los tiempos para optimizar las cargas de trabajo.
- Elimina duplicados y valida datos
- Une las fuentes

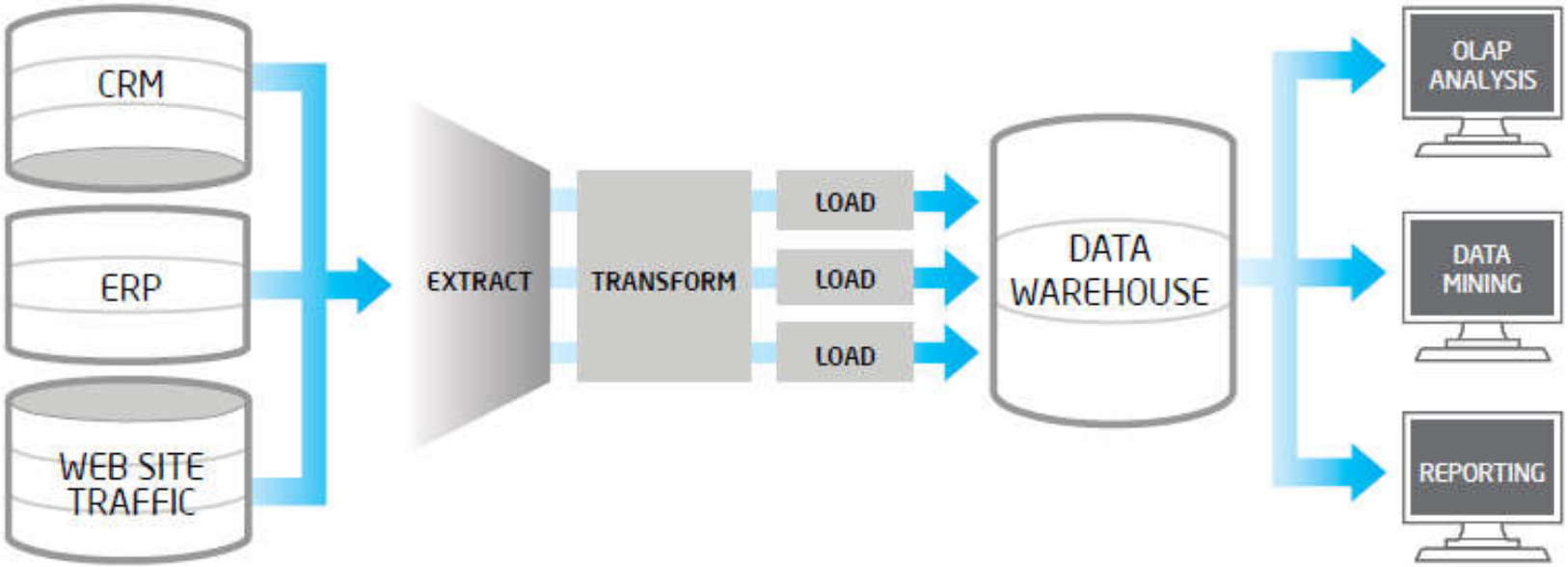
Procesos ETL: ¿cómo elegir una herramienta?

Volumen de datos a gestionar

Naturaleza de los datos

Tareas que se espera que realice la herramienta

Procesos ETL



Data Warehouse

El Data Warehouse concentra y almacena de forma estructurada toda la información obtenida a partir de las múltiples fuentes de datos en nuestra organización, permitiendo así una rápida integración con herramientas de minería de datos, análisis y reportes (dashboards).

La estructura más simple que se encuentra en un Data Warehouse es aquella cuyos datos mantienen su formato bruto (RAW) junto con sus metadatos (datos que describen otros datos). En conjunto están listos para ser explorados y analizados con técnicas de Data Mining.

Una segunda estructura son los **datos procesados**. Previamente se les han aplicado técnicas de limpieza y están diseñados para diferentes grupos de la organización, como el área de inteligencia de negocios, donde las estructuras de datos tienen una relación de dimensiones y tablas de hechos. Una dimensión representa una característica del negocio y los hechos son métricas de interés que se quieren desglosar mediante las dimensiones antes mencionadas.

Data Warehouse

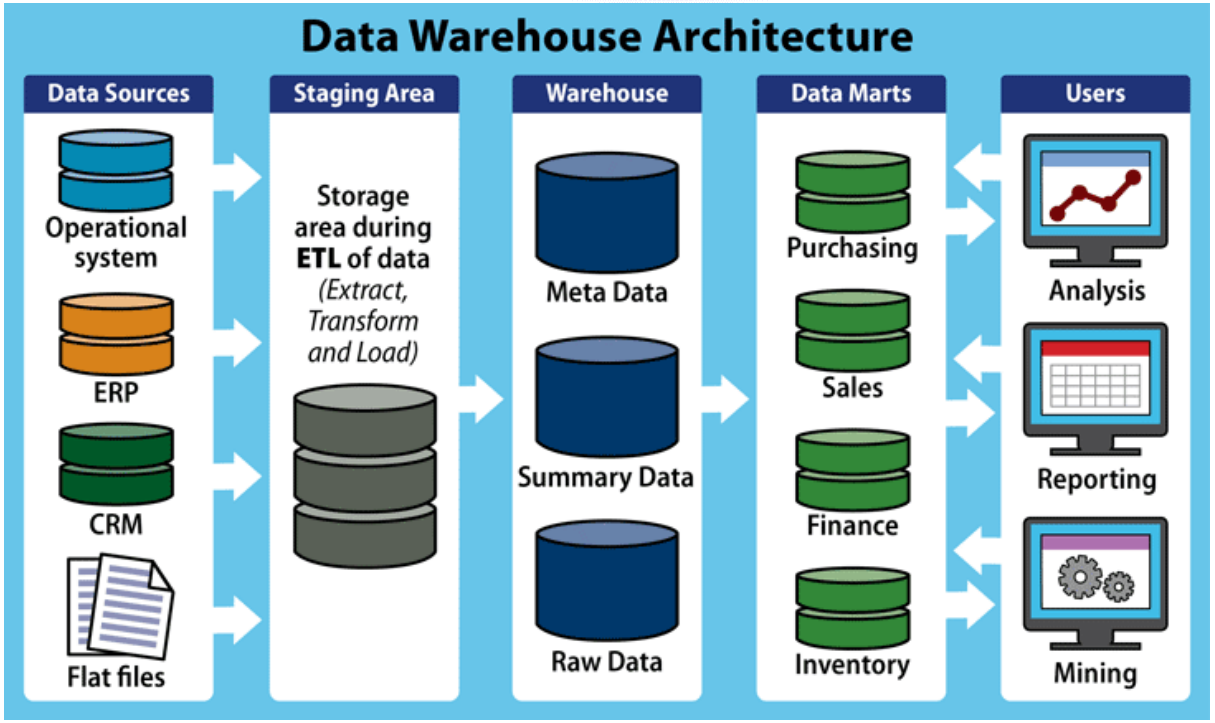
Concentración de datos para explotación

Descomposición en fragmentos llamados Data Marts

No es un SW

No es una marca

No es sólo una base de datos



Algunas arquitecturas más populares en DATAWAREHOUSE

cloudera®

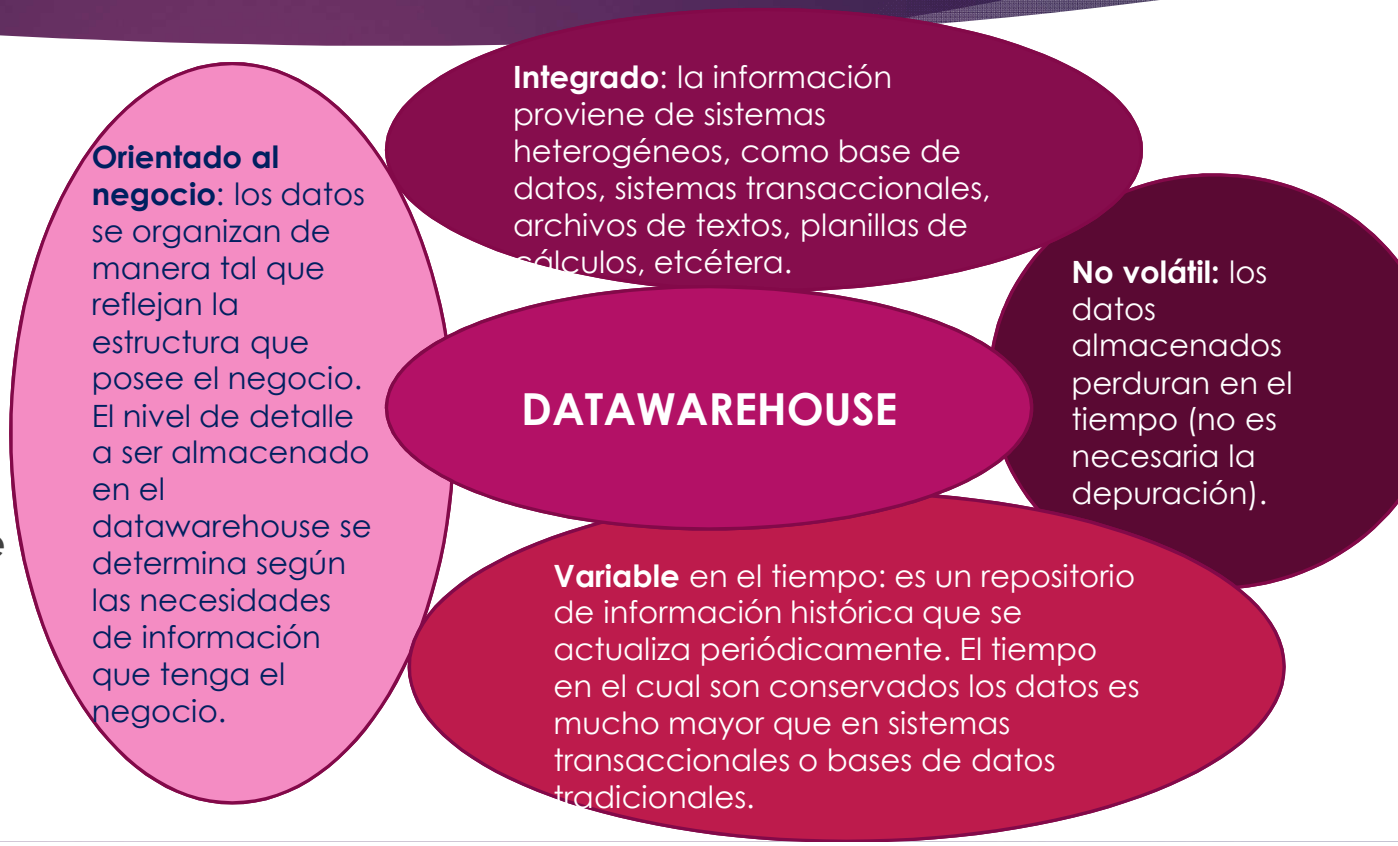
Azure Synapse Analytics

Google BigQuery

amazon REDSHIFT

Datawarehouse: Características Básicas

- **El datawarehouse es uno de los componentes más destacados de la arquitectura de business intelligence.**
- **El datawarehouse es un repositorio de datos, integrado, no volátil, variable en el tiempo y orientado al negocio (Inmon, 1992).**



Requisitos

- Enfocado en toda la empresa para servirla íntegramente
- Diseño flexible
- Preparado para carga masiva de datos en cortos tiempos

- Naturaleza multipropósito
- Estructurados en Data Marts en forma de estrella o copo de nieve


Data Analysis & Platforms
Hadoop
PARACCEL
Storm
HPCC Systems
GridGain
Dremel
Hortonworks
Apache Drill
calpont
ORACLE
HD
Zettaset
Databases / Data warehousing
INFOBRIGHT
Cassandra
APACHE HBASE
Hibari
riak
Infinispan
Bigdata®
OrientDB
Neo4j
HYPERTABLE
redis
Globals
Operational
Versant JPA
MarkLogic
mobject
Multivalue database
Rocket
U2. REVELATION
northgate
QM
BASE INTERNATIONAL
Big Data search
Lucene
Apache Solr
elasticsearch.
Data aggregation
oozie
Business Intelligence
talend
open data solutions
JASPERSOFT
SpagoBI
Jedox
pentaho
Palo
BIRT Exchange
ACTUATE
KNIME
Data Analytics Made Easy
Data Mining
RAPID MINER
orange
RAPID ANALYTICS
WEKA
KEEL
togaware
SPMF
Social
Apache Kafka
ThinkUp
Corona
Multidimensional
GT.M
SciDB
rasdaman
raster data manager
Graphs
Gephi
makes graphs handy
InfiniteGraph
FlockDB
AllegroGraph® 4.9
GraphBuilder
Gremlin
INFO GRID
HYPERGRAPH-DB
dex
meronymy
GraphBase
BrightstarDB
Grid Solutions
GIGASPACES
HAZELCAST
Galaxy
KeyValue
EROSPIKE
leveldb
GENIEDB
Chordless
Tokyo Cabinet
Scalaris
SCALIEN
Project Voldemort
A distributed database.
hamsterdb
RAPTORDB
FairCom
STSDB
HyperDex
IQLECT
OpenLDAP
ioremap.net
STORAGE AND BEYOND
Document Store
mongoDB
CouchBase
CouchDB
Raven DB
CLUSTERPOINT
RaptorDB
EJDB
djon DB
JasDB
SchemafreeDB
sisodb
denso db
Object databases
db4objects
ZOPE
NEOPPOD
Picolisp
STARCOUNTER
Magma
Sterling
siaqodb
MORANTEX
INFORMATION SYSTEMS
EyeDB
HSS Database™
FRAMER D
NDatabase
C# Lightweight Object Database
Multimodel
ArangoDB
alchemydatabase
XML Databases
existdb
BASE
Qizx
sedna
xindice
Created by: www.bigdata-startups.com