Programa de Big Data

OBJETIVOS

- ❑ **Ingeniería de Datos**
- ❑ **Terminología**
- ❑ **Business Intelligence**
- ❑ **DataMart**
- ❑ **DataWarehouse**
- ❑ **Procesos ETL**
- ❑ **Big Data**
- ❑ **Conceptos**

¿Qué es la ingeniería de datos?

¿Por qué es tan importante?

¿Cuál es su participación dentro de un proyecto de datos?

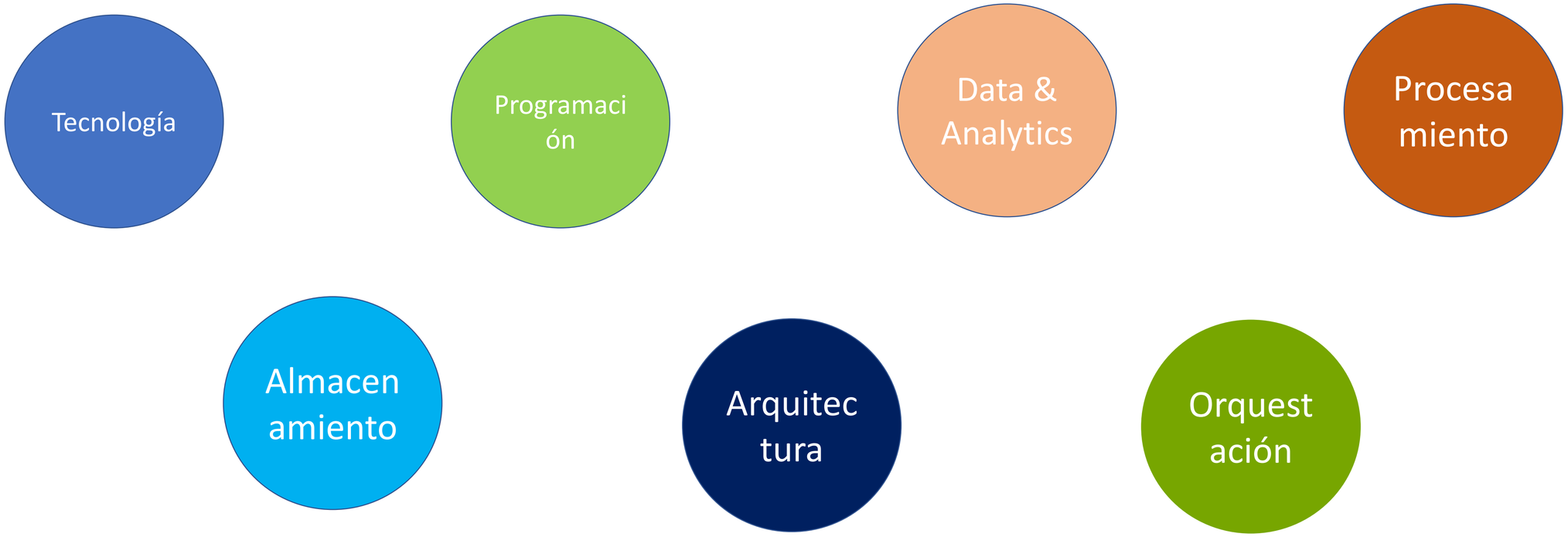
- ✓ Pipelines de carga de datos
- ✓ Extraer fuentes externas

- ✓ Transformación de datos
- ✓ Estandarización de fuentes de información

- ✓ Cálculo de indicadores
- ✓ Flujos de validación

- ✓ Puesta en producción
- ✓ Automatización y Monitoreo

¿Cuál es su participación dentro de un proyecto de datos?



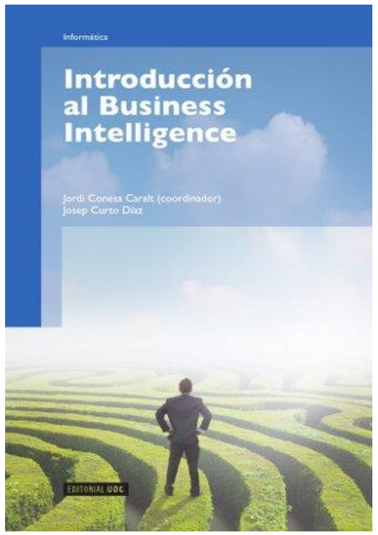
Areas

¿Qué es el Business Intelligence?

En 1989 Howard Dresden, analista de Gartner propone una definición formal del concepto:

Conceptos y métodos para mejorar las decisiones de negocio mediante el uso de sistemas de soporte basado en hechos.

Conjunto de metodologías, aplicaciones, prácticas y capacidades enfocadas a la creación y administración de información, que permite tomar mejores decisiones a los usuarios de una organización



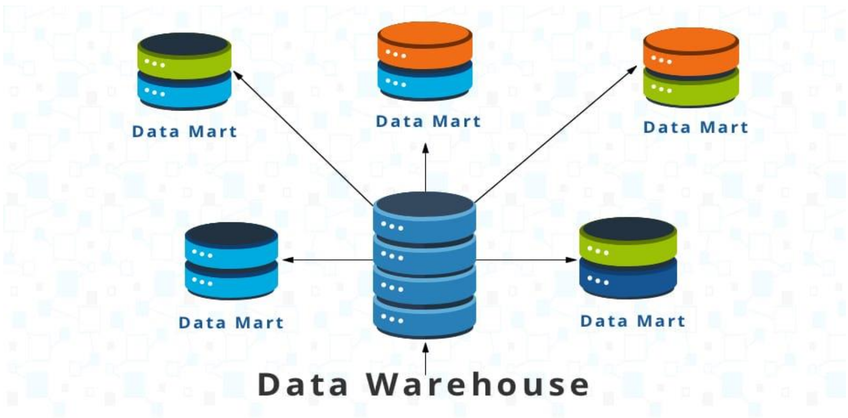
Procesos del BI

Algunas de las tecnologías que forman parte del Business Intelligence son:

- **Data warehouse**
- **Reporting**
- **Análisis OLAP (On-Line Analytics Processing)**
- **Análisis Visual**
- **Cuadro de mando integral**
- **Reglas de negocio**
- **Previsiones**
- **Dashboards**
- **Integración de datos (incluye ETL)**

DataMart

Es un subconjunto de los datos del data warehouse cuyo objetivo es responder a un determinado análisis, función o necesidad, con una población especifica de usuarios.



El datamart está pensando para cubrir las necesidades de un grupo de trabajo o de un determinado departamento dentro de la organización.

DataWareHouse

Un data warehouse es un repositorio de datos que proporciona una visión, global, común e integrada de los datos de la organización, con las siguientes propiedades: estable, coherente, fiable y con información histórica.

Orientado a un tema

Organiza una colección de información alrededor de un tema central

Integrado

Incluye datos de múltiples orígenes y presenta consistencia de datos

Se realizan fotos de los datos basados en fechas

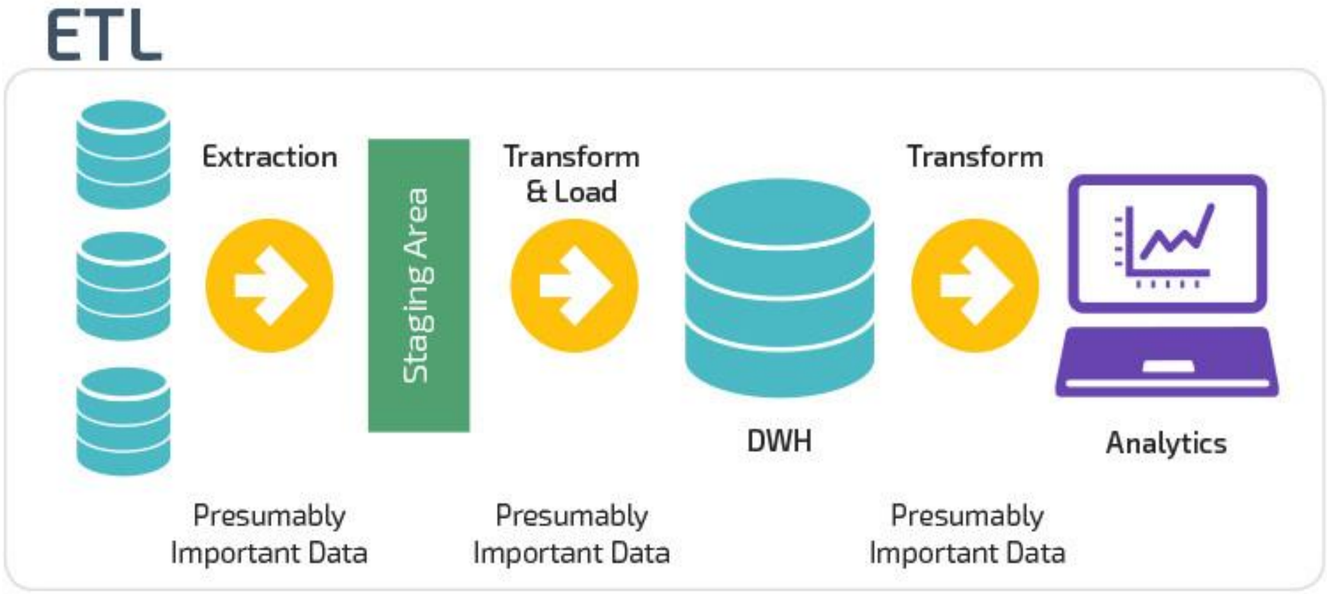
Variable en el tiempo

Solo de lectura para los usuarios finales

No volátil

Procesos ETL

Tecnología de integración de datos basada en la consolidación de datos que se usa tradicionalmente para alimentar data warehouse, datamart y staging área.

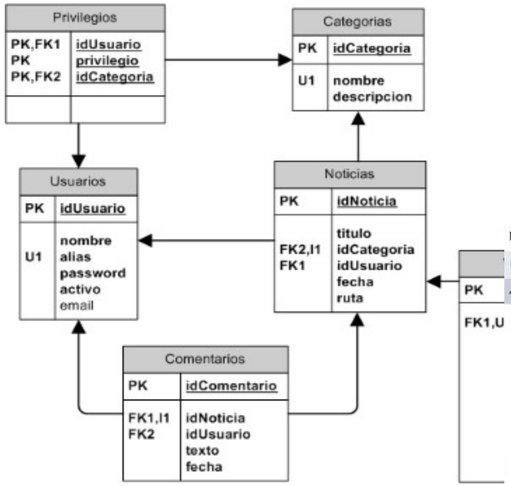


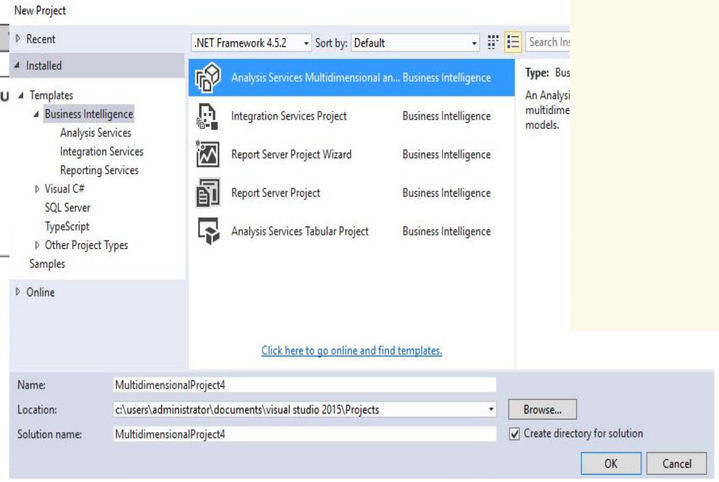
¿Qué procesos ETL conoces?

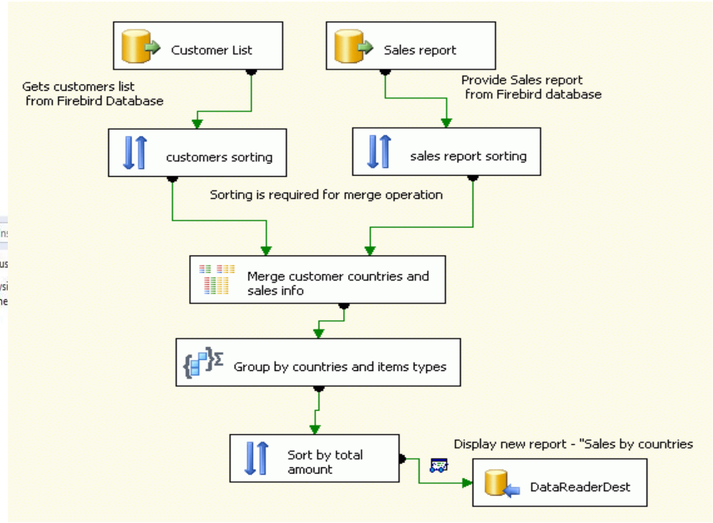
Herramientas ETL

Interfaz de usuario

Ambientes de desarrollo

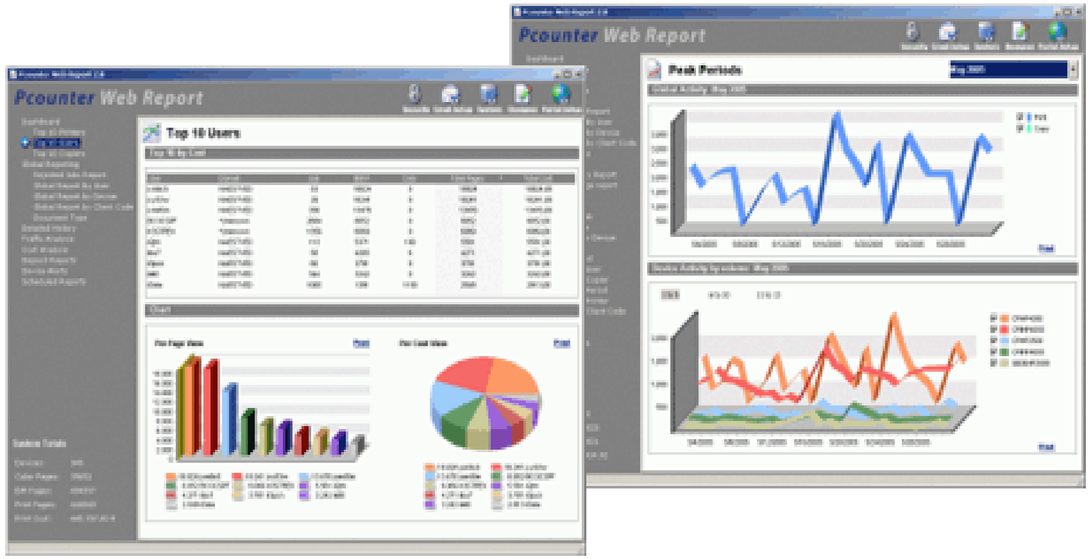






Interfaz de usuario

Areas de Negocio



- Cubos OLAP
- Reporte
- Dashboard

Interfaz de usuario

Areas de Negocio

- “Presentación visual de la información mas importante necesaria para alcanzar uno o mas objetivos los cuales encajan en una sola pantalla de computadora de tal manera que pueda ser monitoreada a una mirada” (Few, 2013 p34)
- Pequeña, concisa, clara, intuitiva y a menudo interactiva presentación
- Foco en monitoreo y medición
- Combinación de Charts(Gráficas) y tablas

¿Qué es el Big Data?

¿Qué es Big Data?

"Big data son datos que contienen una mayor variedad y que se presentan en volúmenes crecientes y a una velocidad superior. Esto se conoce como "las tres V".

Gartner's - Doug Laney

Las V’s del Big Data

Es un marco de trabajo **(conceptos + tecnologías)** que permite procesar grandes volúmenes de datos, de diferentes estructuras o con carencia de estas, que pueden variar en el tiempo, a grandes velocidades y que generen valor al negocio.

Principios de un DataLake

Principios de un DataLake

Arquitectura desacoplada

Arquitectura de capa con componentes modulares que aíslan el almacenamiento y el cómputo, minimizando los riesgos de perdida de datos

Persistencia Polygot

Es el principio de una variedad de tecnologías de procesamiento y almacenamiento especialmente diseñadas para resolver distintos casos de uso del negocio **(streaming, nosql)**

Serveless

Aplicación de distintas tecnologías sin servidor (IaaS o PaaS) permitiendo reducir el costo y la carga operativa de la plataforma.

Autoservicio

Un **datalake** debe ser fácil de usar y de autoservicio para todos los interesados. Almacenamiento de datos con SQL, vistas comerciales para BI operativo, así como datos RAW catalogados para exploración y descubrimiento

Organización de un DataLake

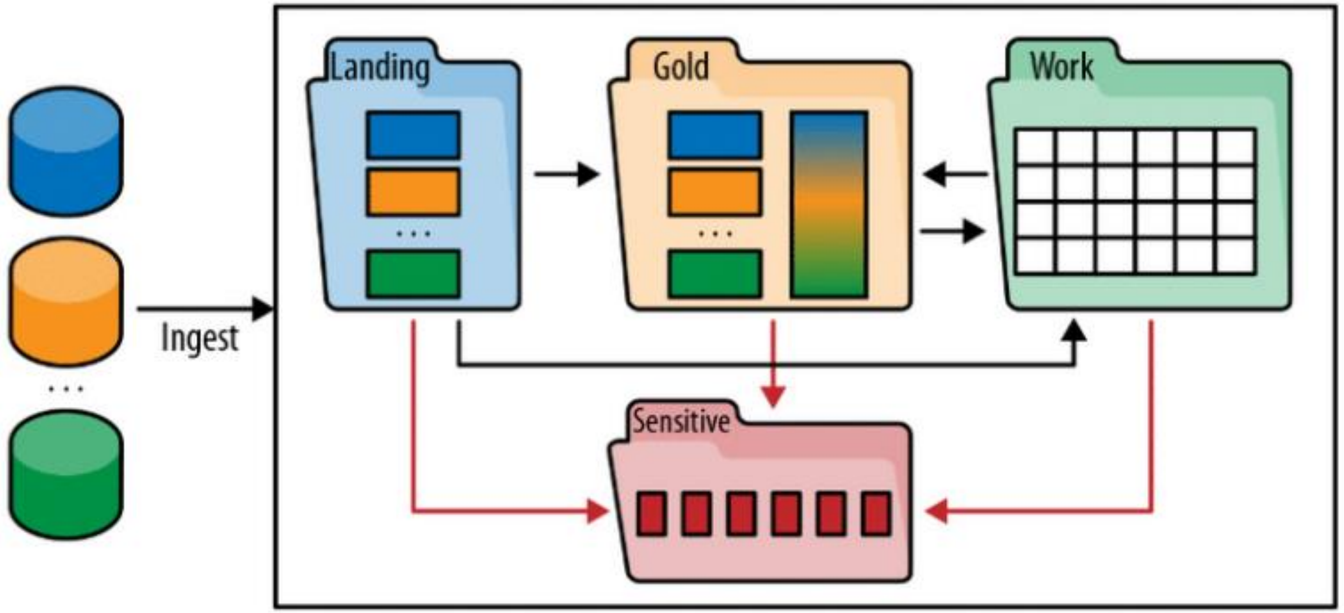


Figure 7-1. Sample breakdown of data lake into workspaces or zones

Zona Landing - Raw

La zona **landing**, algunas veces también llamada raw o zona staging, es usada para almacenar los datos crudos. Por lo general, solo los desarrolladores altamente técnicos, ingenieros de datos y científicos de datos tienen acceso a la zona landing. En general, los usuarios de la zona landing deben tener una razón convincente para realizar su propio tratamiento y procesamiento de datos.

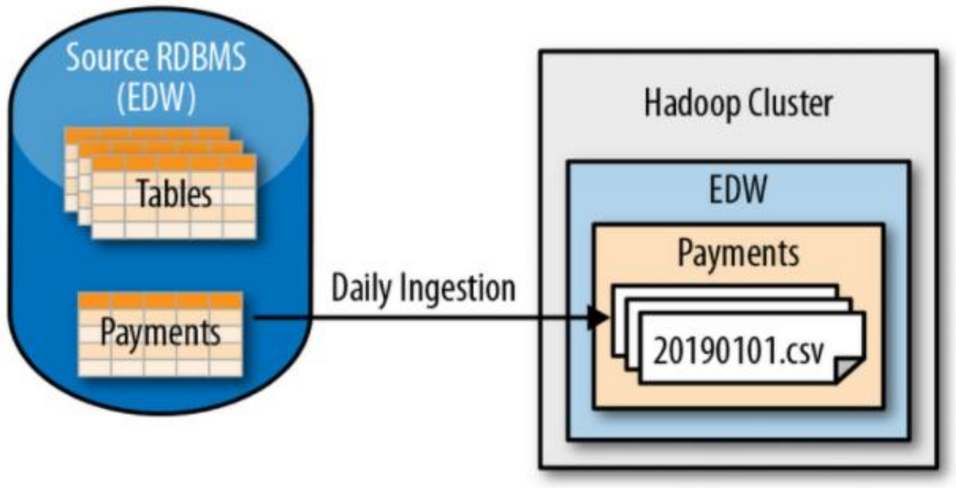
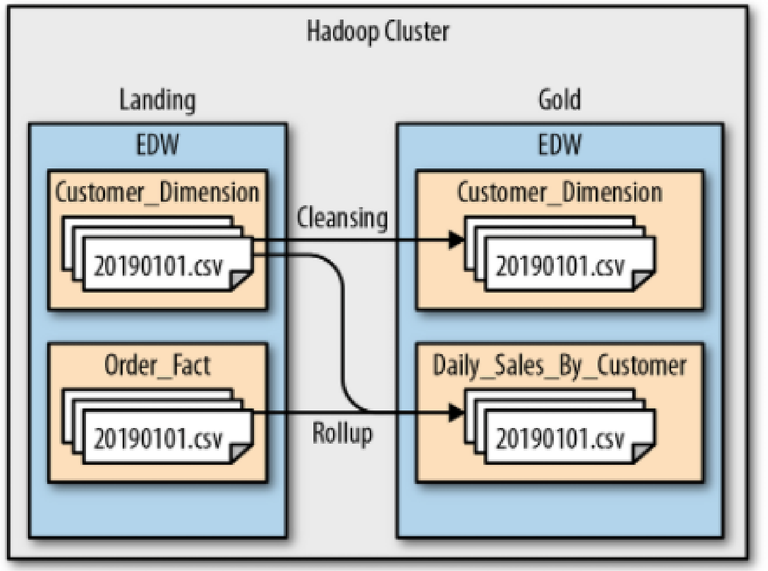


Figure 7-3. Ingesting raw or landing data into folders and files

Zona Gold

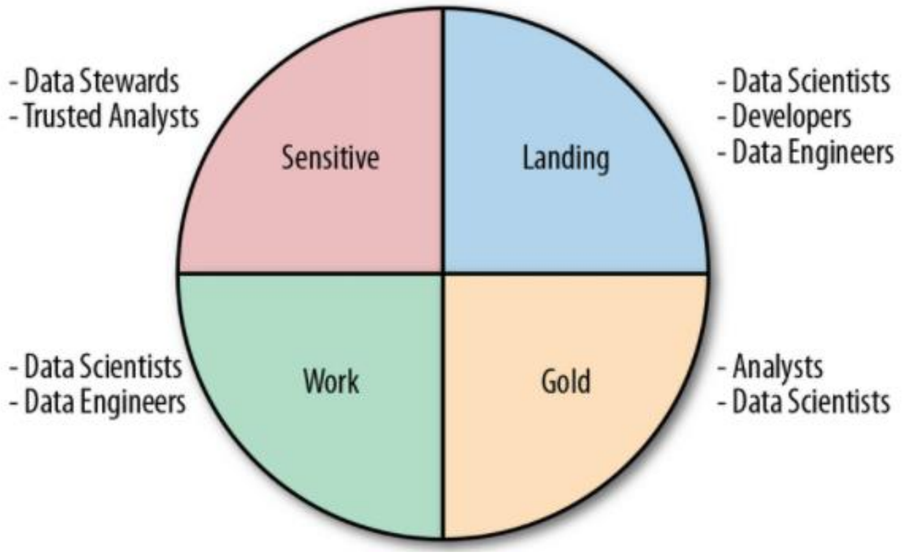
La zona **Gold** frecuentemente es espejo de la zona landing, pero contiene datos limpios y enriquecidos con alguna otra fuente del landing. Esta zona es ocasiones es llamada prod que indica que los datos que contienen son productivos y no tienen problemas de calidad. Usualmente este es la zona más popular. Muchos desarrolladores y data scientist tienen acceso a esta zona ya que tienen los datos mucho más limpios y estandarizados.



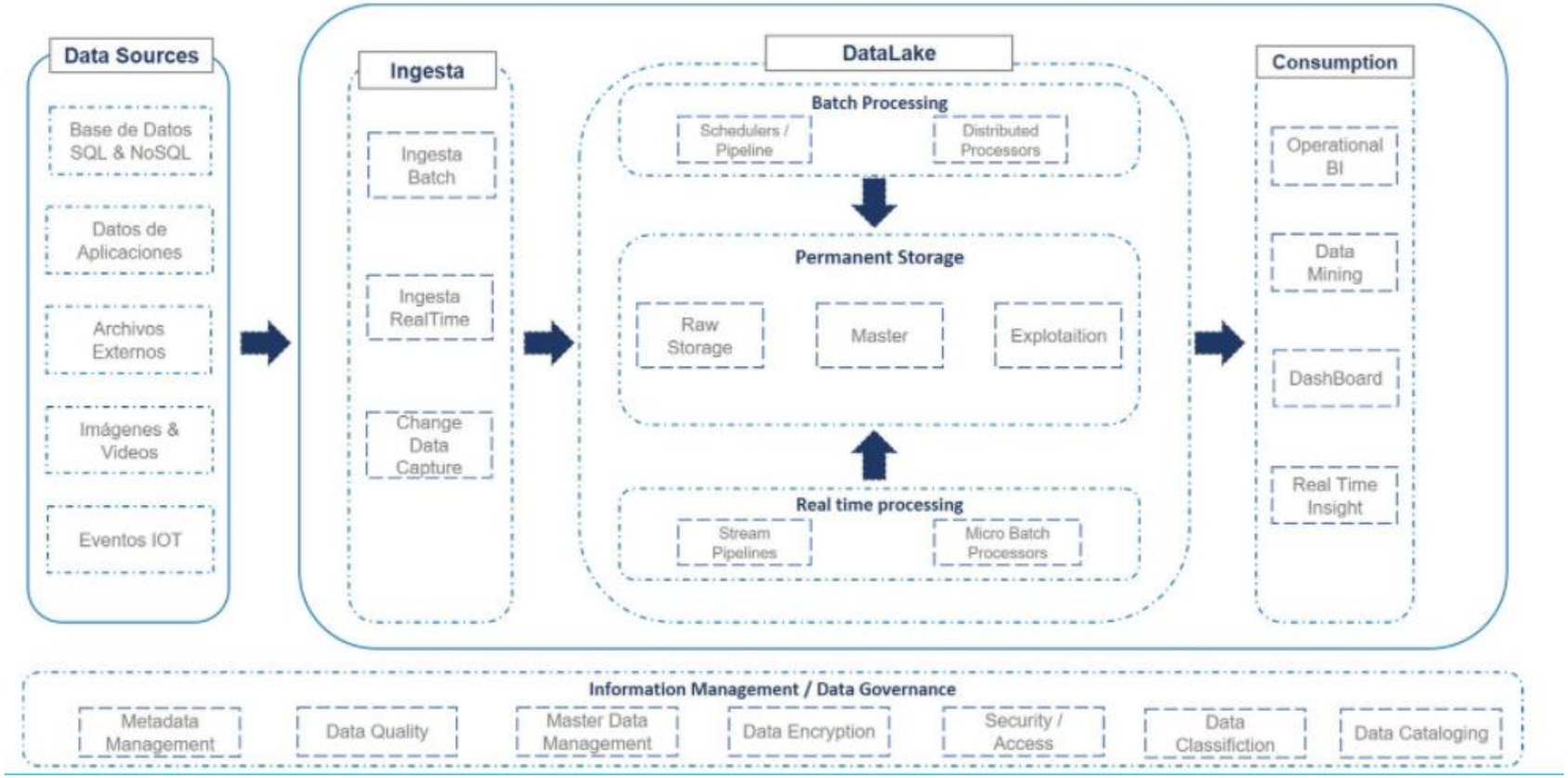
Organización de un DataLake

La mayor parte del análisis ocurre en la zona de **Work**, también conocida como zona de desarrollo o proyectos. Esta zona generalmente está estructurada para reflejar la estructura organizativa de la empresa. Normalmente es el dominio de desarrolladores, científicos de datos e ingenieros de datos, aunque los analistas a menudo lo utilizan para realizar la preparación de datos de autoservicio para sus proyectos.

A veces se crea una zona **Sensible** para mantener archivos que contienen datos que es particularmente importante para proteger de los espectadores no autorizados. va sea debido a los requisitos reglamentarios o las necesidades comerciales general, solo los administradores de datos y otras personas autorizadas obtener acceso a los datos en la zona sensible



Plataforma de datos



Roles de un Equipo de Datos

Data Quality

Data Governance Specialist

Product Owner

Data Engineer

Data Analyst

Data Traslator

Data Enabler

Data Architect

Data Visualization Expert

Project Manager

Data Scientist

Roles de un Equipo de Datos

Chief Data Office

Gobierno de Datos

Advanced analytics
Data engineering

Data Translator
Facilita la comunicación entre perfiles técnicos y de negocio. Fomenta un espíritu de experimentación desde la planificación hasta la ejecución del proyecto. Promueve la adopción de una cultura analítica y de procesos.

Data Quality Specialist
Participa activamente en las iniciativas de gestión de datos de la empresa y co-diseña soluciones eficientes (con foco en calidad) basadas en tecnologías avanzadas de data y analytics.

Data Scientist

Product Owner
Comunica las necesidades y contexto del negocio al equipo; y vela por cumplir los objetivos del proyecto.

Data Visualization Specialist
Facilita la generación de insights con el negocio construyendo visualizaciones de datos que facilitan su análisis y entendimiento.

Data Enabler
Diseña el modelado de datos para facilitar su consumo de forma eficiente y escalable asegurando la calidad de la información.

Machine Learning Engineer

Data Analyst
Extrae, analiza datos y genera reportes para apoyar en la toma de decisiones.

Project Manager
Planifica y monitorea el proyecto. Vela por el cumplimiento de rituales y puntos de control de calidad. Gestiona riesgo y la relación con los sponsors. Impulsa la cultura de autogestión en el equipo.

Data Engineer
Construye el conjunto de variables transformadas que son la base del desarrollo de los modelos analíticos avanzados.

Data Governance Specialist
Participa activamente en las iniciativas de gestión de datos de la empresa y co-diseña soluciones eficientes (con foco en gobierno) basadas en tecnologías avanzadas de data y analytics.

Data Architect
Encargado del diseño e implementación de la Arquitectura de Datos. Lidera y propone nuevas estrategias del manejo de los datos. Diseña nuevas arquitecturas para cada tipo de caso de uso