Modulo Tratamiento de Datos

GFT

Pedro Nieto

Esteban Chiner

Roberto López

Miguel Angel Sotomayor

Ruben Sanchís

Curso 2020/2021 - Edición 2

Fecha 26/09/2019

The Team

Pedro Nieto

- **Computer Engineer** from the Universitad de Oviedo

- Working in GFT as a **Senior Data Specialist** specialized in Data Management and IoT

- Part of the technological **Business Development** team

- Leading the **IoT area and Industry sector**

Personal:
- **Freaky Proud**
- Father of **two**
- Owner of **Raspberry Pi** (1,2, 3 and Zero)
- **Technology Evangelist**

@a10pepo

pedronietopelaez

Pedro.nieto@gft.com

Esteban Chiner

- **Computer Engineer** from the Universitat de València & **Master on Software Engineering** from Universitat Politècnica de València

- Working in GFT as a **Senior Architect** specialized in Big Data and Blockchain

- Part of the technological **Business Development** team
- Leading the "**Artificial Intelligence & IoT**" and "**DLT & Blockchain**" domains

- **Lecturer**
  - Big Data & Analytics Master (**UPV**) – "Big Data Architectures in Financial Services"
  - Seminaris d'Empresa (**UPC**) – "Developing decentralized apps with Ethereum"
  - Data Analytics for Enterprises (**EDEM**)

- **Hadoop Certified** in developer, administration and Hbase

- Paper published: "**A Big Data Financial Information Management Architecture for Global Banking**"

@echiner

echiner

esteban.chiner@gft.com

Roberto López

- **Computer Engineer & Master on Distributed Computing** from Polytechnic University of Valencia

- On **GFT since 2009** working as a **Senior Software Engineer** specialized in Big Data and Artificial Intelligence

- Leading **Fast Data Domain**

- **Lecturer**
  - Big Data & Analytics Master (**UPV**) – "Big Data Architectures in Financial Services"
  - Data Analytics for Enterprises (**EDEM**)

- **Hadoop Developer Certified**
- **Google Cloud Data Engineer Certified**
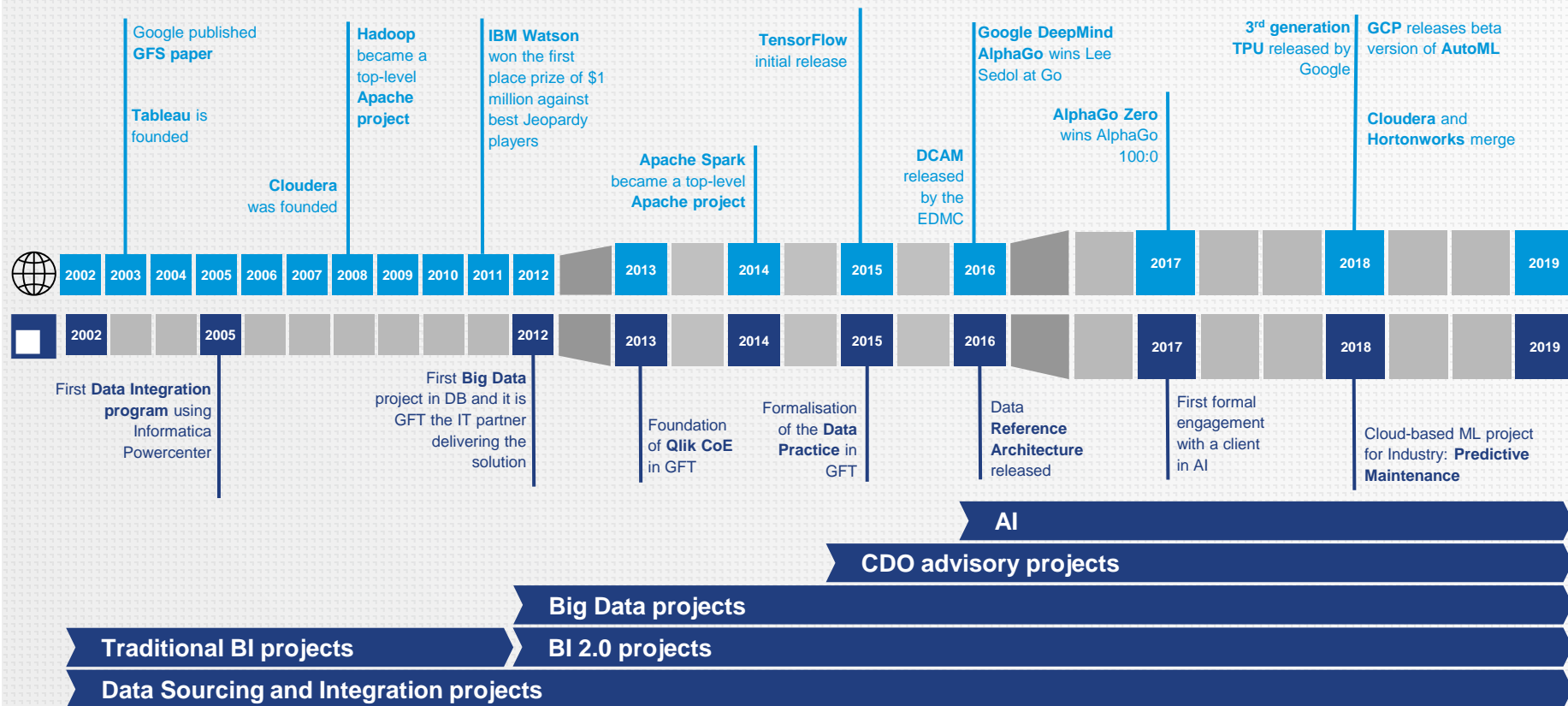
@rlopezherrero
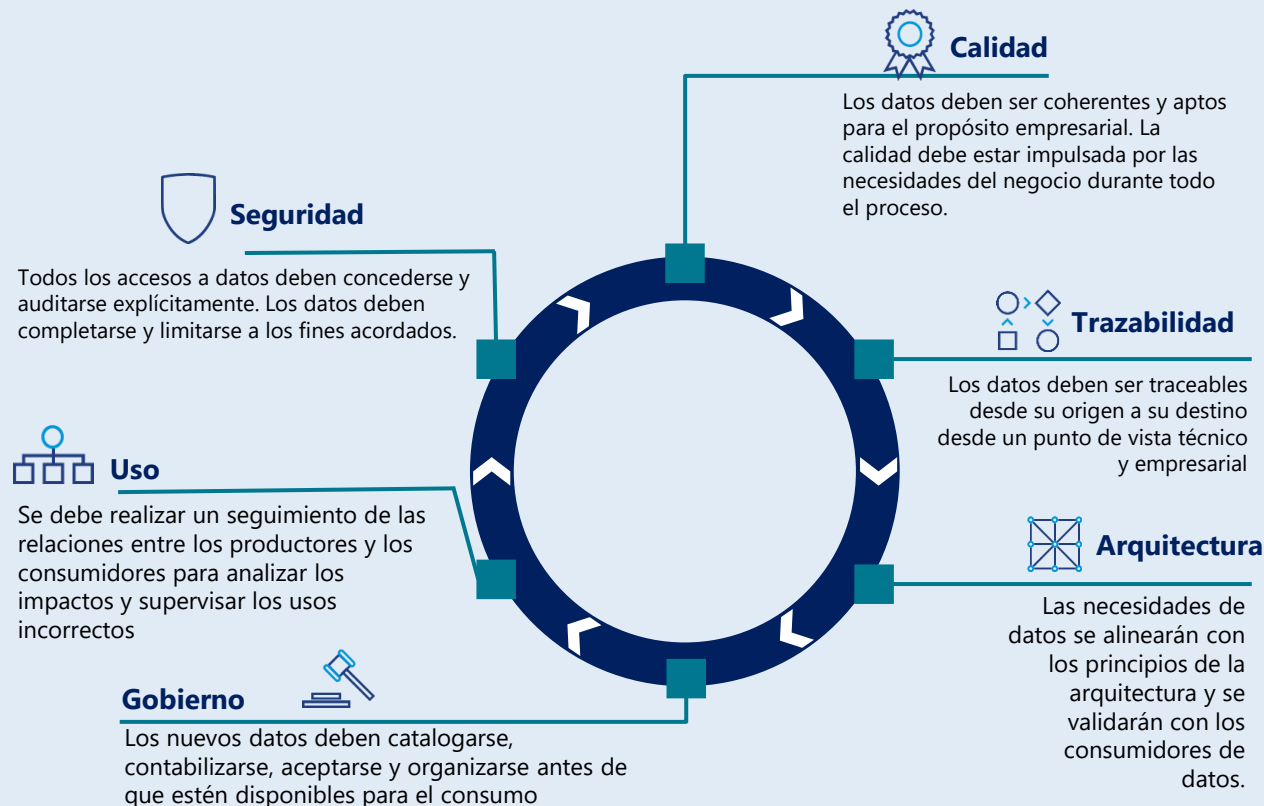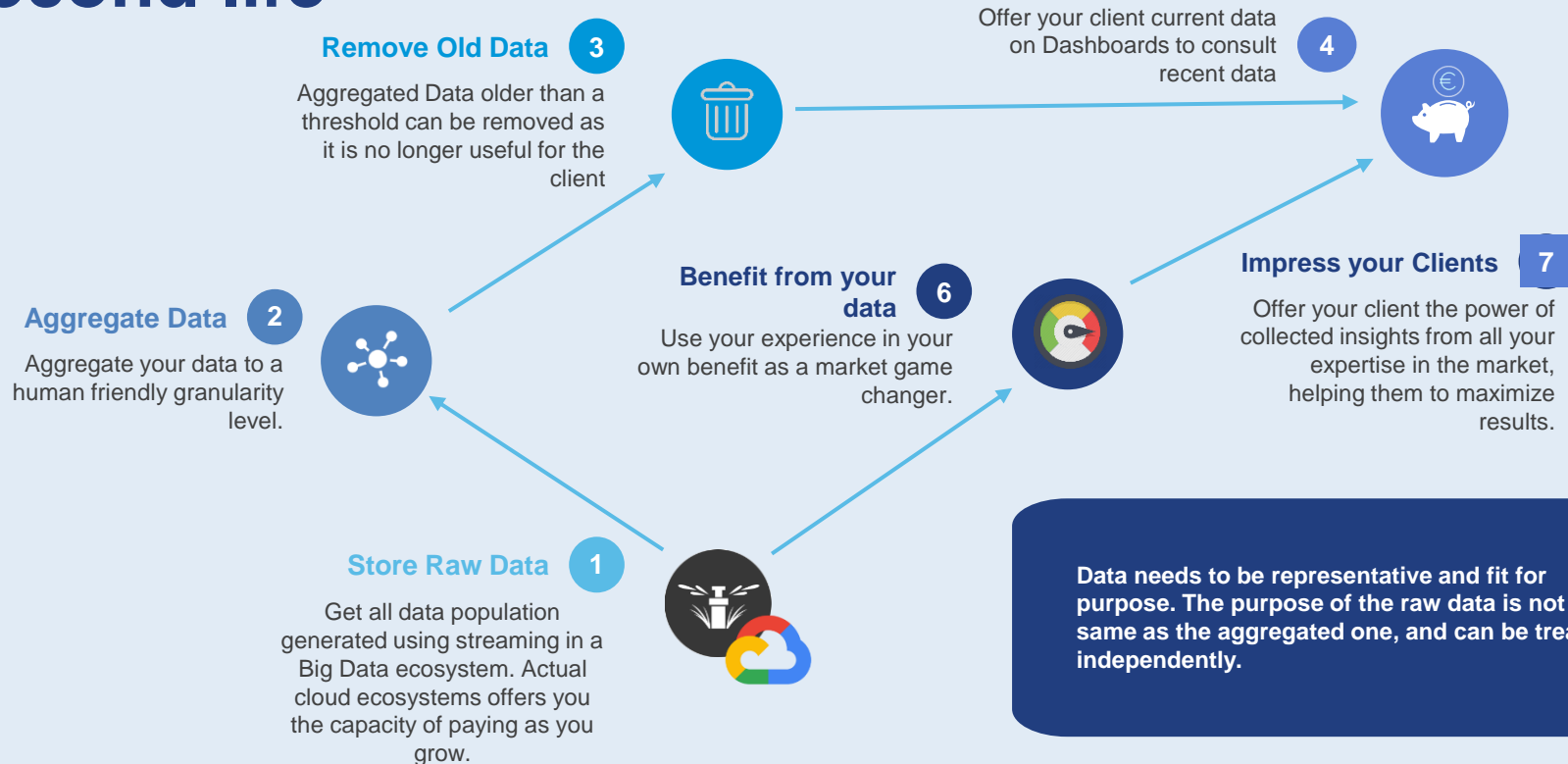
rlopezherrero

roberto.lopez@gft.com

Hablemos de datos…

Our journey in Data

Google published **GFS paper**

Tableau is founded

Cloudera was founded

Hadoop became a top-level **Apache project**

IBM Watson won the first place prize of $1 million against best Jeopardy players

Apache Spark became a top-level **Apache project**

TensorFlow initial release

DCAM released by the EDMC

Google DeepMind AlphaGo wins Lee Sedol at Go

AlphaGo Zero wins AlphaGo 100:0

3rd generation **TPU** released by Google

GCP releases beta version of **AutoML**

Cloudera and **Hortonworks** merge

🌐 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019

◻ | 2002 | 2005 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019

First **Data Integration program** using Informatica Powercenter

First **Big Data** project in DB and it is GFT the IT partner delivering the solution

Foundation of **Qlik CoE** in GFT

Formalisation of the **Data Practice** in GFT

Data **Reference Architecture** released

First formal engagement with a client in AI

Cloud-based ML project for Industry: **Predictive Maintenance**

AI

CDO advisory projects

Big Data projects

Traditional BI projects

BI 2.0 projects

Data Sourcing and Integration projects

No son solo los datos…

Calidad

Los datos deben ser coherentes y aptos para el propósito empresarial. La calidad debe estar impulsada por las necesidades del negocio durante todo el proceso.

Seguridad

Todos los accesos a datos deben concederse y auditarse explícitamente. Los datos deben completarse y limitarse a los fines acordados.

Trazabilidad

Los datos deben ser traceables desde su origen a su destino desde un punto de vista técnico y empresarial

Uso

Se debe realizar un seguimiento de las relaciones entre los productores y los consumidores para analizar los impactos y supervisar los usos incorrectos

Arquitectura

Las necesidades de datos se alinearán con los principios de la arquitectura y se validarán con los consumidores de datos.

Gobierno

Los nuevos datos deben catalogarse, contabilizarse, aceptarse y organizarse antes de que estén disponibles para el consumo

Give your data a second life

Remove Old Data ③

Aggregated Data older than a threshold can be removed as it is no longer useful for the client

Offer your client current data on Dashboards to consult recent data ④

Aggregate Data ②

Aggregate your data to a human friendly granularity level.

Benefit from your data ⑥

Use your experience in your own benefit as a market game changer.

Impress your Clients ⑦

Offer your client the power of collected insights from all your expertise in the market, helping them to maximize results.

Store Raw Data ①

Get all data population generated using streaming in a Big Data ecosystem. Actual cloud ecosystems offers you the capacity of paying as you grow.

Data needs to be representative and fit for purpose. The purpose of the raw data is not the same as the aggregated one, and can be treated independently.

Everywhere...

Document Processing

Image Processing

Smart Bots

ChatBot Assistants

Anti-Money Laundring

Time Series Analysis

Decission Support

Scalable AI

Connected Devices

AI as a Service

Predictive Maintenance

What all of them have in common?

Data Evolution

Where are you?

Companies are evolving to embrace smart industry, as an IT company we can help you succeed in the process.

Data

Environments usually on premise store small amounts of data for transactional purposes

- **Optimize Transactional Environment**
- **Identify weak points in your technical approach**

Information

Historical Data and innovative data sources start to gain attention in the ecosystem and are stored.

- **Platform Strategy (Cloud)**
- **IoT Integration**
- **Reporting Dashboards**
- **Process Optimization**

Knowledge

Use your experience in your own benefit as a market game changer. Squeeze your data with ML and get valuable insights.

- **Business Rules**
- **Decision Support Systems**
- **Machine Learning Models**
- **Predictive Analysis**

Wisdom

Offer your new view to your clients and predict or suggest new actions.

- **Model Prediction**
- **Process Automation**
- **DAAS (Data as a Service)**

Linked in Learning

The Skills Companies Need Most in 2020

Top 5 Soft Skills
1. Creativity
2. Persuasion
3. Collaboration
4. Adaptability
5. Emotional intelligence

Top 10 Hard Skills
1. Blockchain
2. Cloud computing
3. Analytical reasoning
4. Artificial intelligence
5. UX design
6. Business analysis
7. Affiliate marketing
8. Sales
9. Scientific computing
10. Video production

Business Value Forecast by AI Type

■ Decision Support/Augmentation ■ Agents ■ Decision Automation ■ Smart Products

Millions of Dollars

2,500,000

1,250,000

0

2017 2018 2019 2020 2021 2022 2023 2024 2025

2030: 44% / 13% / 19% / 24%
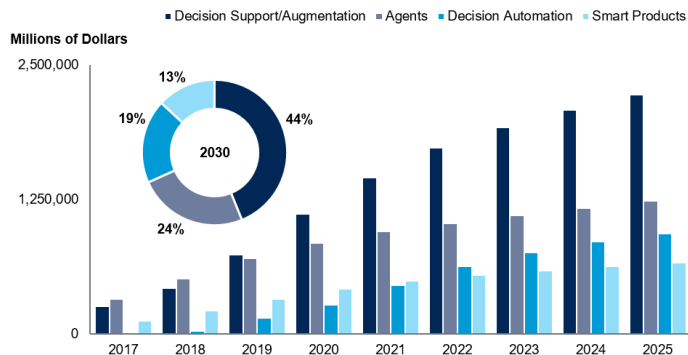
Source: Gartner
ID: 386366

Las profesiones más buscadas en 2020

Sector	El más buscado	
'Telecos'	Arquitecto de computación en nube	> 40.000
Tec. de la información	Desarrollador Java	> 40.000
Marketing	Jefe de producto	35.000-50.000
Comercial	Director de desarrollo empresarial	45.000-65.000
Industria	Ingeniero de automatización	35.000-45.000
Salud	Médico del trabajo	50.000-60.000
Sector financiero	Responsable de control financiero	45.000-70.000

CincoDías

LAS 15 PROFESIONES CON MÁS FUTURO EN 2020 SEGÚN LINKEDIN

Ranking en función del incremento anual de la demanda, según el "Informe Empleos Emergentes 2020" en España, elaborado por LinkedIn.

01 Especialista en Inteligencia Artificial
02 Desarrollador de Salesforce
03 Especialista en Customer Success
04 Ingeniero de Robótica
05 Especialista en Ciber Seguridad
06 Agile Coach
07 Consultor de Cloud
08 Desarrollador de Python
09 Científico de Datos
10 Desarrollador de Big Data
11 Ingeniero de Datos
12 Representante de Help Desk
13 Representante de Desarrollo de Negocio
14 Representante de Desarrollo de Ventas
15 Ingeniero de Cloud

Triunfa con
Linked in
Consultoría de Social Selling

What do we expect from you?

Data Scientist
also known as Data Managers, statisticians.

A data scientist will be able to take data science projects from end to end. They can help store large amounts of data, create predictive modelling processes and present the findings.

Skills: Mathematics, Programming, Communication

Will use programmes such as:
SQL, Python, R

Data Engineers
also known as database administrators and data architects.

They are versatile generalists who use computer science to help process large datasets. They typically focus on coding, cleaning up data sets, and implementing requests that come from data scientists.

Skills: Programming, Mathematics, Big data

Will use programmes such as:
Hadoop, NoSQL, and Python

Data Analysts
also known as business Analysts.

They typically help people from across the company understand specific queries with charts.

Skills: Statistics, Communication, Business knowledge

Will use programmes such as:
Excel, Tableau, SQL

Data Engineers

What Will you learn?

LANGUAGES, TOOLS AND SOFTWARE
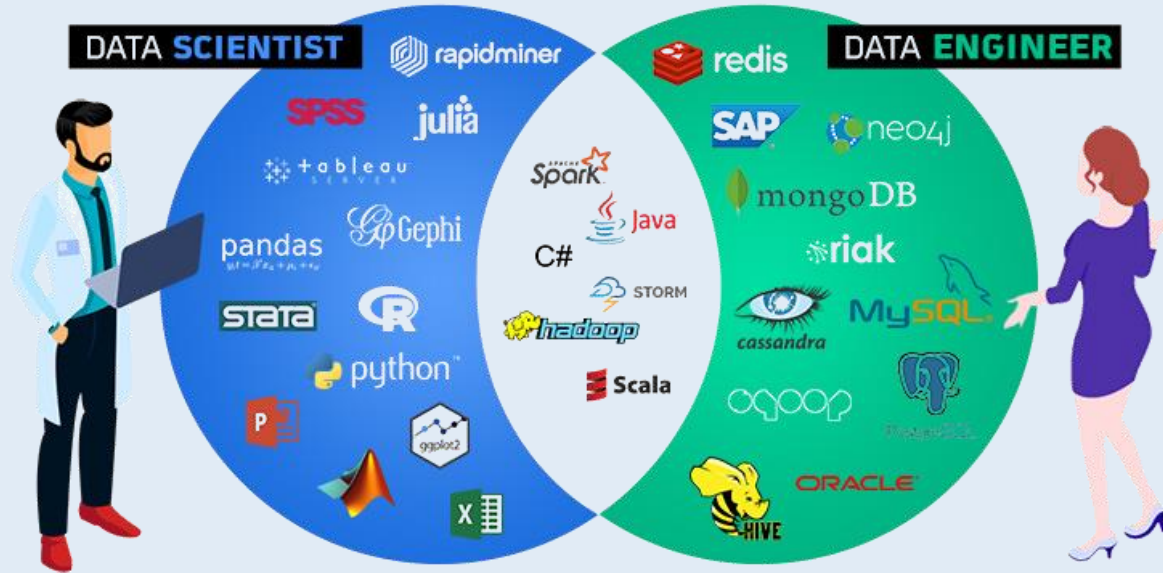
Módulo 1

Data Architecture example
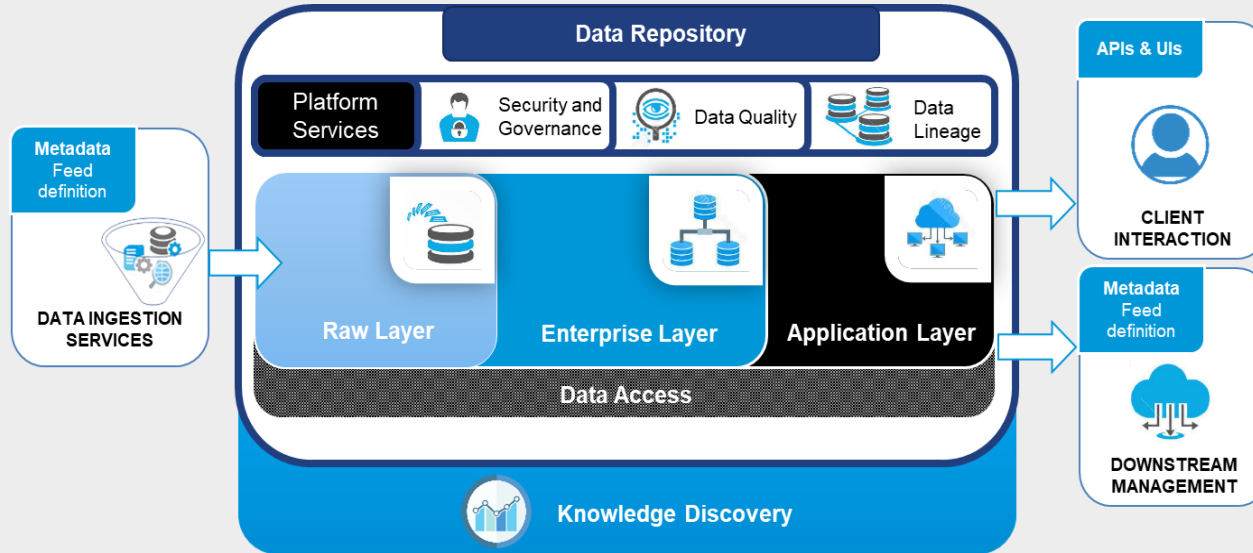
Get the data

- Data should be easily obtained, and frequently updated
- Data can come in any format or shape
- Be ready for all speeds (from batch to real-time)
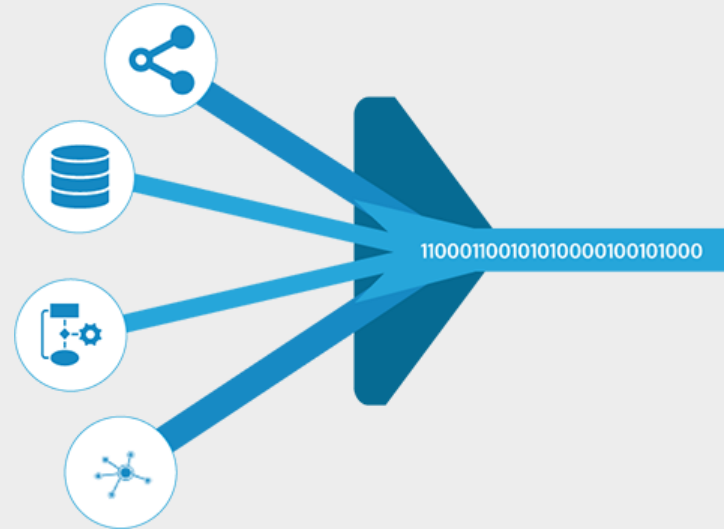- Data as a service
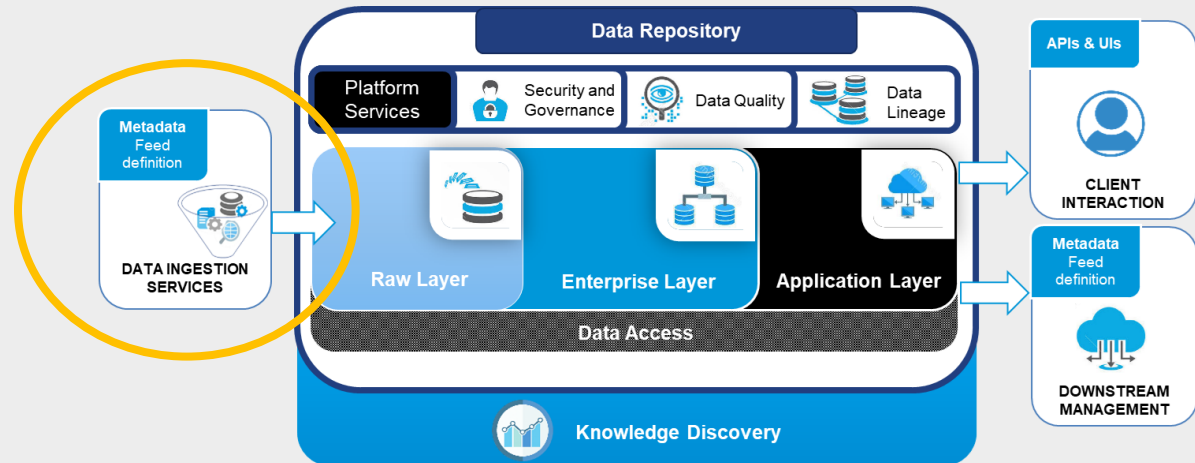
Data discovery

- We want to get not only the data, but the right data, and from the right source
- What data do we have?
- Is it relevant and enough?
- Is external or reference data required?
- Is it updated?
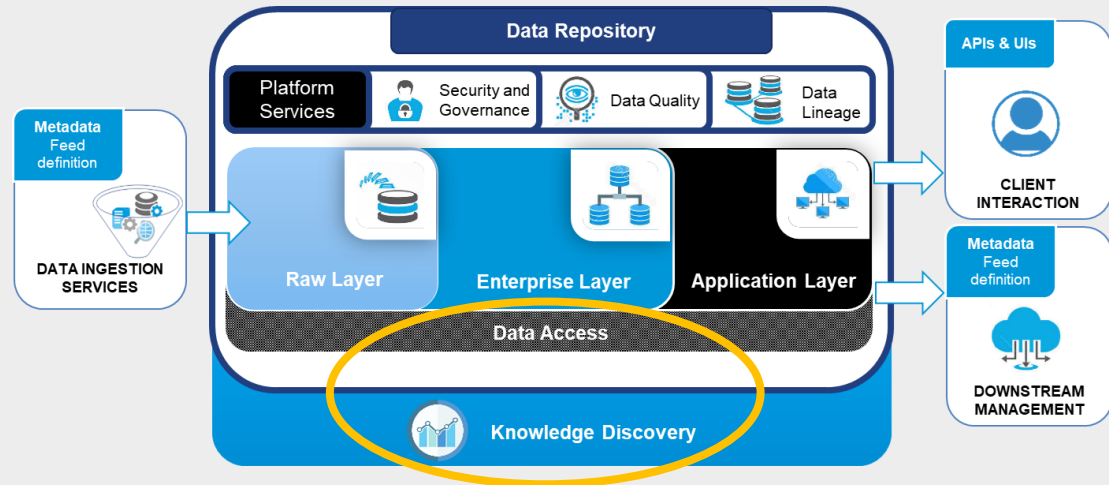
Captura y Recogida Datos

- **Unidad 1: Introducción a los Datos**
- **Unidad 2: Tipos de Datos**
  - Estructurados
  - No Estructurados
  - Semi-estructurados
- **Unidad 3: Integración de Fuentes**
  - Internal Data Sources
  - External Data Sources
  - Data Sources Access
- **Unidad 4: Frecuencia de Creación**
  - Batch Sources
  - Micro-batching
  - Streaming Sources

Captura y Recogida Datos

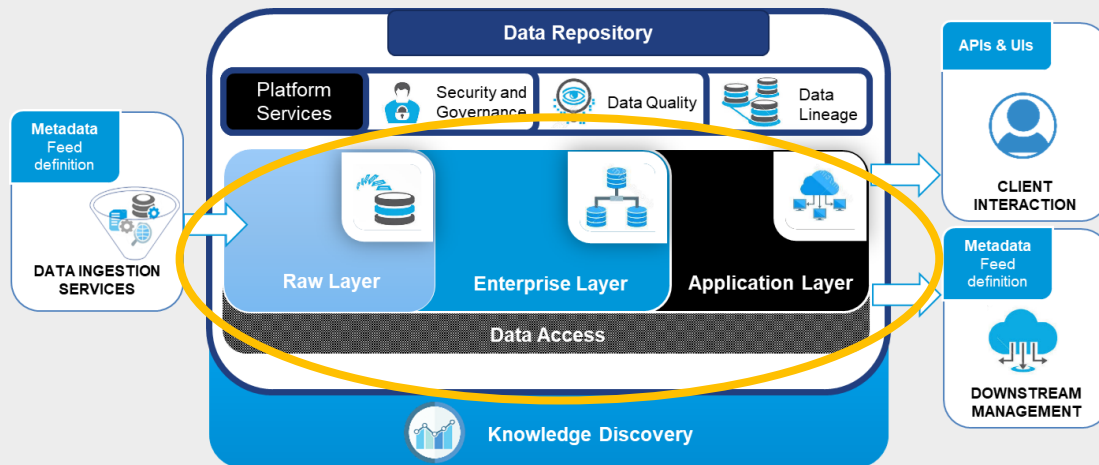- **Unidad 5: Ingestión de Datos**
  - ETL
  - Ingestión Pura
  - Basada en Eventos
- **Unidad 6: Tipos de Fuentes**
  - Open Data
  - Public DataSets
  - Data as a Service

11000110010101000100101000
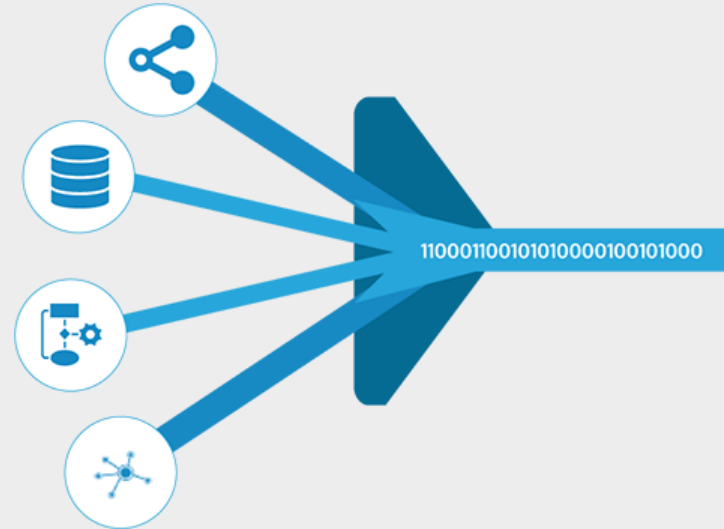
Prepare the data

- Maximize analysts time by having the data ready
- Data transformations
- Filtering
- Data quality checks
- Enrichment
- Use a common repository

Almacenamiento y procesamiento

- Unidad 1: Arquitecturas de Datos
    - Batch Architectures
    - Real Time Architectures
    - Lambda & Kappa Architectures
- Unidad 2: Tipos de almacenamiento
    - BBDD Relacionales
    - BBDD No-SQL
    - Documentos Indexados
    - Sistemas almacenamiento columnar
    - Almacenamiento distribuido
   - Almacenamiento serializado

Almacenamiento y procesamiento

- Unidad 3: Procesamiento de Datos
  - Procesamiento en modo Batch
  - Procesamiento en Streaming
- Unidad 4: Privacidad y protección de Datos
- Unidad 5: Tipos de Infraestructuras
  - Cloud (IAAS, PAAS, SAAS)
  - On premises
  - Hibridas

Organize the data

- Separate and properly catalog the data
- Tag and complete with metadata
- Data stewardship
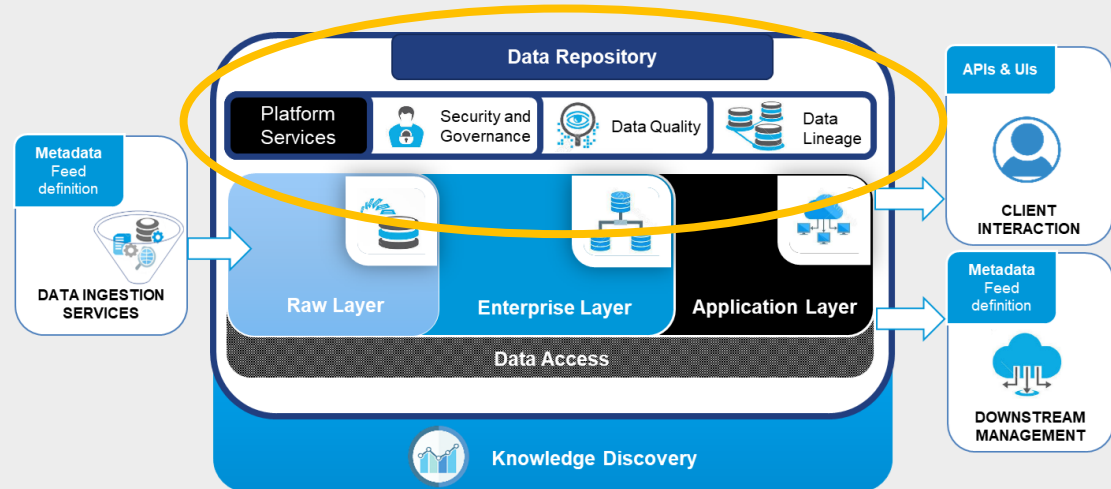
Gobierno del Dato

- Unidad 1: Origen del problema
  - Crecimiento Orgánico
  - Ciclo de vida de los datos
  - Data Oriented Architecture
- Unidad 2: Estándares de Mercado
- Unidad 3: Data Governance
  - Roles y Responsabilidades
  - Data Owner
  - Data Stewardship
  - Data Stakeholders
  - Comités y Funcionamiento
  - Herramientas de Mercado

Gobierno del Dato

- Unidad 4: Data Quality
  - Data profiling
  - Data curation
  - Reglas de calidad
  - Herramientas de mercado
- Unidad 5: Data lineage and Traceability
  - Glosario de términos & diccionario
  - Capa semántica
  - Linaje y trazabilidad
  - Flujos de trabajo
  - Herramientas de mercado
  - Data Stakeholders
  - Comités y Funcionamiento
  - Herramientas de Mercado

Gobierno del Dato

- Unidad 6: Éticas del dato y Seguridad
  - Ecosistema legal (GDPR)
  - Data masking y cifrado de datos
  - Datos sensibles
- Unidad 7: MDM y Reference Data
  - Metadatos
  - Visión 360
  - Reglas de supervivencia
  - Reglas de consolidación
- Unidad 8: Nuevos escenarios de datos
  - Datos como servicio (Daas)
  - Catálogos de datos
  - Casos Reales

Use the data

- Build reports
  - Help on decision support
- Visual dashboards
  - Control and monitor operations
- Build Machine Learning Model
  - Offer new products
  - Improve processes
  - Optimize production